UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Arquitectura de Computadors

Energy-oriented Optimizations Towards Sustainable Internet

Sergio Ricciardi

Advisors:
Davide Careglio
Germán Santos Boada

A thesis submitted to the Computer Architecture Department
of the Technical University of Catalonia – Barcelona TECH (UPC)
in partial fulfillment of the requirements for the degree of

PhilosophiæDoctor (PhD)

Barcelona, September 2012

Oficina de Doctorat
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Acta de qualificació de tesi doctoral

Curs acadèmic:

Nom i cognoms

DNI / NIE / Passaport

Programa de doctorat

Unitat estructural responsable del programa

Resolució del Tribunal

Reunit el Tribunal designat a l'efecte, el doctorand / la doctoranda exposa el tema de la seva tesi doctoral titulada ____ ____.

Acabada la lectura i després de donar resposta a les qüestions formulades pels membres titulars del tribunal, aquest atorga la qualificació:

□ APTA/E □ NO APTA/E

(Nom, cognoms i signatura) President/a	(Nom, cognoms i signatura) Secretari/ària

(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal

______________________, _______ d'/de __________________ de _______________

El resultat de l'escrutini dels vots emesos pels membres titulars del tribunal, efectuat per l'Escola de Doctorat, a instància de la Comissió de Doctorat de la UPC, atorga la MENCIÓ CUM LAUDE:

□ SI □ NO

(Nom, cognoms i signatura) Presidenta de la Comissió de Doctorat	(Nom, cognoms i signatura) Secretària de la Comissió de Doctorat

Barcelona, _______ d'/de __________________ de _______________

Ai miei genitori,

che mi hanno insegnato *come* pensare,

lasciandomi sempre la libertá

di scegliere *cosa* pensare.

Abstract

The astonishing development of the Information and Communication Technologies (ICT) of the last decades fosters larger and larger demands in terms of network infrastructures and cloud computing datacenters facilities. The ever increasing data volumes to be processed, stored and accessed every day in the modern cloud infrastructures connected by ultra-high bandwidth networks require huge performance that results in the ICT energy demand to grow at faster and faster pace. Therefore, the energy consumption and the concomitant green house gases (GHG) emissions of the Internet are becoming major issues in the Information and Communication Society (ICS). The Internet infrastructure, comprising both network (routers, switches, line cards, signal regenerators, optical amplifiers, etc.) and cloud facilities (servers, storage systems, racks, power distribution systems, cooling equipment, etc.) have reached huge capacities but their development has not been compensated at the same rate as for their energy consumption. It is estimated that the Internet infrastructure consumes 12,6% of the worldwide electricity production (equivalent to the power output of about 240 modern nuclear power reactors). Furthermore, the overall power consumption of ICT equipment is growing steadily, with a mean rate of 12% per year, further stressing the need for systemic energy-oriented solutions: energy-efficient devices managed by energy-aware protocols and algorithms, and powered by a smart grid power distribution network employing renewable energy sources. The research works leading to this Thesis investigate the current challenges for a sustainable high-performance Internet infrastructure and propose new energy-oriented models, protocols, algorithms and paradigms that, considering energy and GHG as novel constraints, optimize the Internet ecological footprint while not disrupting the performance, towards sustainable society growth and prosperity.

Contents

1

Energy-oriented Optimizations Towards Sustainable Internet

1.1 Introduction

In the last years, the Internet traffic has grown astonishingly, and it is foreseen (1) that by 2016, the total Internet traffic will be three times larger than the one observed in 2012 (equivalent to a monthly traffic of 110 *Exabytes* of data[1], Figure 1.1) and the users connected to the Internet will grow from $2,28$ billions of 2012 to 3,4 billions in 2016 (2) (3) (4).

Besides, higher and higher bandwidth, computing and storage resources are required to cope with the emerging cloud services. Technological advancements in the fields of semiconductors and optics have been able to provide the huge bandwidths and computing/storage resources for satisfying the increasing demands and avoiding the Internet collapse.

However, such a high performance Internet infrastructure requires huge amounts of energy to power network equipment and datacenters facilities, and the power consumption is becoming the main limiting factor even more than bandwidth[2] and computing capacity[3]. Such a growth rate in the performance is not sustainable under the busi-

[1] $1\,Exabyte = 10^3\,Petabytes$;

[2] Advancements in the fiber optic technology can provide almost unlimited bandwidth capacity;

[3] Moore's law seems to encounter a limitation in the energy requirements of the CPUs, a problem known as "dark silicon": a large part of CPUs transistors will be unused because it would require too much energy to power them all – something in between 1 and 10 kW of power per chip;

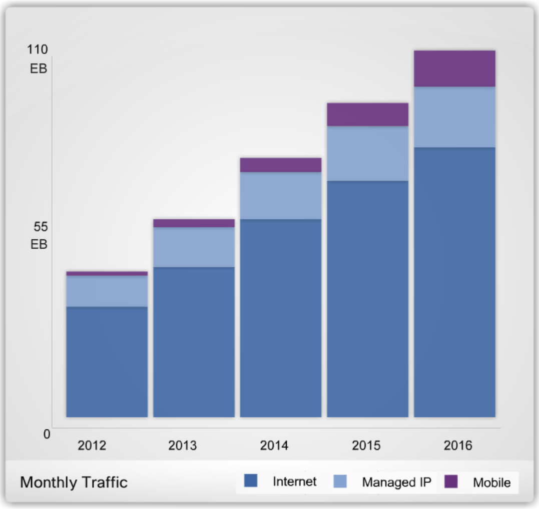


Figure 1.1: Monthly Internet traffic growth forecast (1).

ness as usual (BAU) scenario and proper countermeasures have to be taken towards a sustainable Internet.

Furthermore, even if the ICT sector does not directly emit green house gases (GHG) during the use phase, the traditional power plants feeding the ICT equipment do emit GHG to generate the required energy. Therefore, the power consumption is not the only limiting factor for the Internet growth. Climate changes (mainly, global warming), resource scarcity and pollution are menacing the entire world population if immediate actions will not be taken to drastically reduce the emissions of GHG in the atmosphere.

To give an idea, in Italy and in France, Telecom Italia and France Telecom, are the second largest consumers of electricity after the respective national railway systems, and in the United Kingdom, British Telecom is the largest single consumer of energy (5) (6) (4).

As a consequence, the reduction of energy consumption and GHG emissions through the use of alternative and renewable *green* energy sources are among the most urgent emerging challenges for telecommunications carriers to cope with the ever increasing energy costs, the new rigid environmental standards and compliance rules, and the growing power demand of high-performance networking and datacenters devices. All the above open problems and issues foster the introduction of new energy-efficiency constraints and energy-awareness criteria in operation and management of modern large

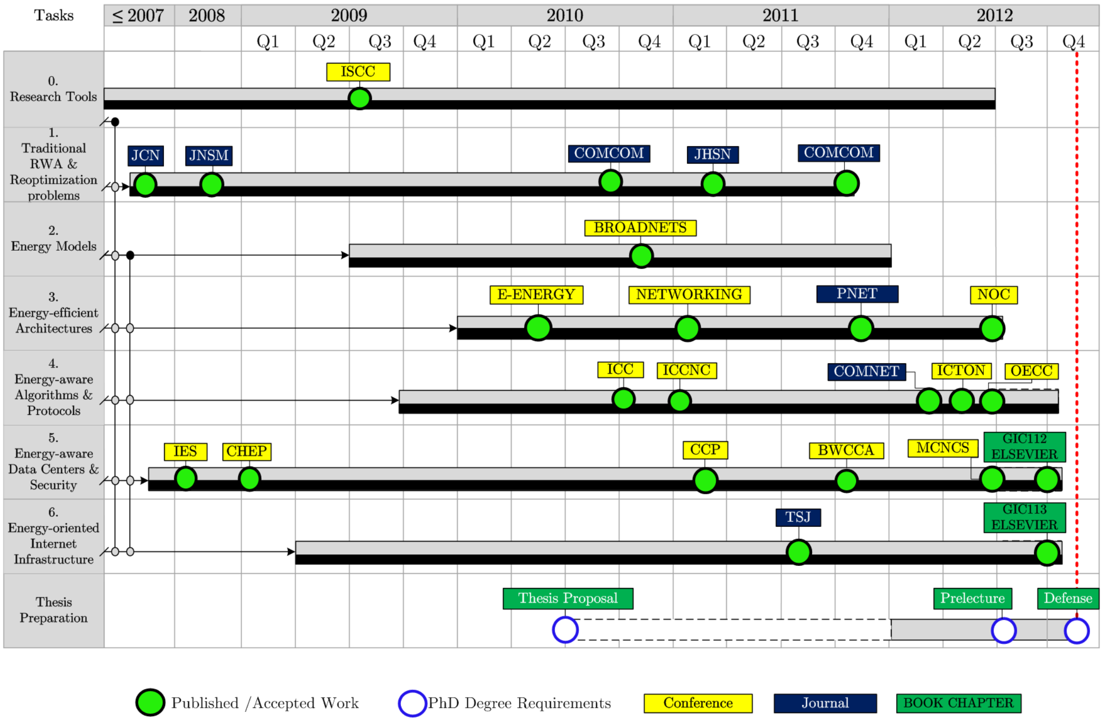


Figure 1.2: Gantt diagram of the Thesis project.

scale communication infrastructures, and specifically in the design and implementation of enhanced energy-oriented control-plane mechanisms to be introduced in next generation Internet.

Towards this goal, this Thesis has been structured in 6 tasks, each addressing a specific issue, and then tied all together in a comprehensive energy-oriented Internet infrastructure, as illustrated in the Figure 1.2. The objectives accomplished in this Thesis are reported in the following sections, together with a discussion of the achieved results.

1.2 The Internet framework

The Internet infrastructure, schematically represented in Figure 1.3 (7), can be logically segmented into a three-level structure, made up of an access network, a metro-edge network and a core network. The access network provides connectivity access to the customers; several technologies are employed, with the most commons ranging from xDSL over copper wire to FTTx over optical fiber.

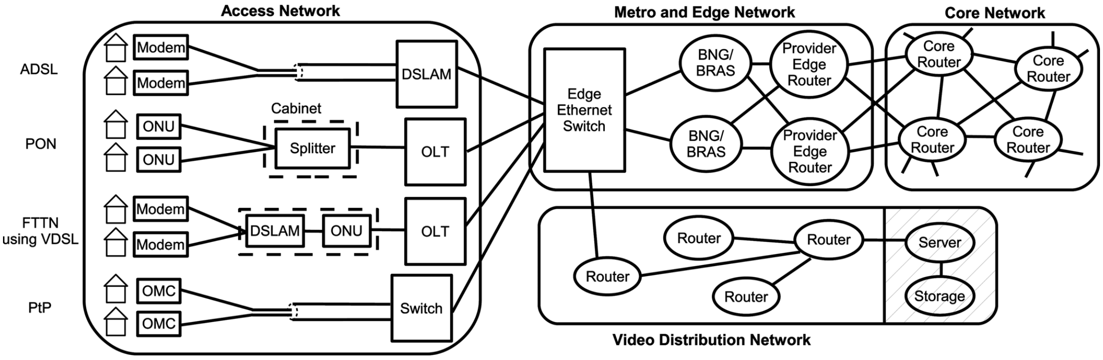


Figure 1.3: Schematic view of a traditional IP network, as used by Internet service providers (ISP) (7).

At the metro-edge network, an edge Ethernet switch aggregates the traffic from the access network and uplinks it to the provider edge router connecting to the core of the network. The BNG and BRAS routers are used by the ISP to provide control, authentication and security services to the customers traffic.

A content distribution network (CDN) is also depicted, providing Video-on-demand (VoD) to the customers. In general, a CDN is made up of a number of datacenters located around the globe, in which the servers provide (replicated) contents to the users, and interconnected by either the same physical metropolitan and core networks above or transported over a dedicated high-capacity network and combined with best effort Internet traffic at the edge Ethernet switch.

In the core network, a relative small number of high-level routers interconnect metro networks, provide the gateway function to neighbor core nodes of different network operators and perform all the necessary routing. The access and the metro networks usually have a tree topology, with the metro segment providing some redundancy. The core network topology is usually highly meshed, with the core routers interconnected by high-bandwidth wavelength division multiplexing (WDM) fiber links, in which several channels are optically multiplexed into a single fiber link using different wavelengths. Optical circuit switched (OCS) networks over WDM technology are usually referred to as wavelength-routed (or WDM-routed) networks , since at the network edge, end-to-end connection requests have to be set-up by creating an appropriate lightpath, i.e. a point to point optical circuit using a specific wavelength that can span more fiber links, typically with a guaranteed bandwidth (Figure 1.4). Traffic flows that share

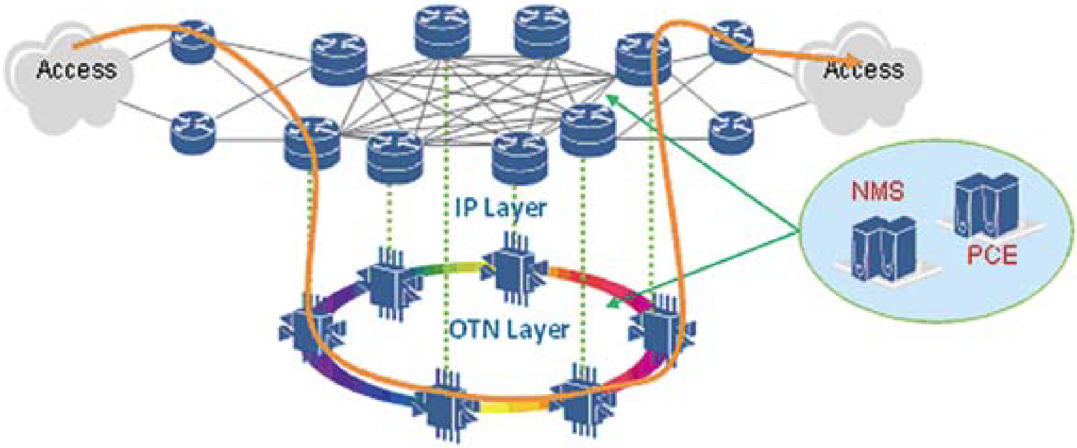


Figure 1.4: An IP-over-WDM core network (8).

the same characteristics in terms of Quality of Service (QoS) requirements can be efficiently multiplexed, or "groomed", onto the same wavelength/lightpath channel in time division or statistical multiplexing.

Therefore, from the logical point of view, a core network can be divided into two layers, an optical transport network (OTN) layer and an electro-optical IP layer built over it, as illustrated in Figure 1.5 (9). At the IP layer, the traffic is electronically processed and then converted to the optical domain to be efficiently transported into the optical layer by optical cross connects (OXC) which perform the optical switching at per-wavelength granularity. A typical control-plane paradigm, like the generalized multiprotocol label switching (GMPLS), operates dynamically setting-up the connection requests arriving at the lambda switching routers (LSR) in the IP layer, and provides an optical connection either over an existing lightpath, by performing electronic traffic grooming in time division multiplexing (TDM), or setting-up one or more new lightpaths in the optical layer, by configuring the appropriate switching in the optical cross connects (OXC) nodes. Such lightpaths are then handled as single IP hop at the electronic layer; individual channels (wavelengths) can be added or dropped into fiber links by means of optical add-and-drop multiplexers (OADM). When a new connection request arrives, the control plane has to choose not only the physical fiber links from the specified source-destination pair satisfying the QoS requirements and providing sufficient residual bandwidth, but it has also to properly select the wavelength on each fiber

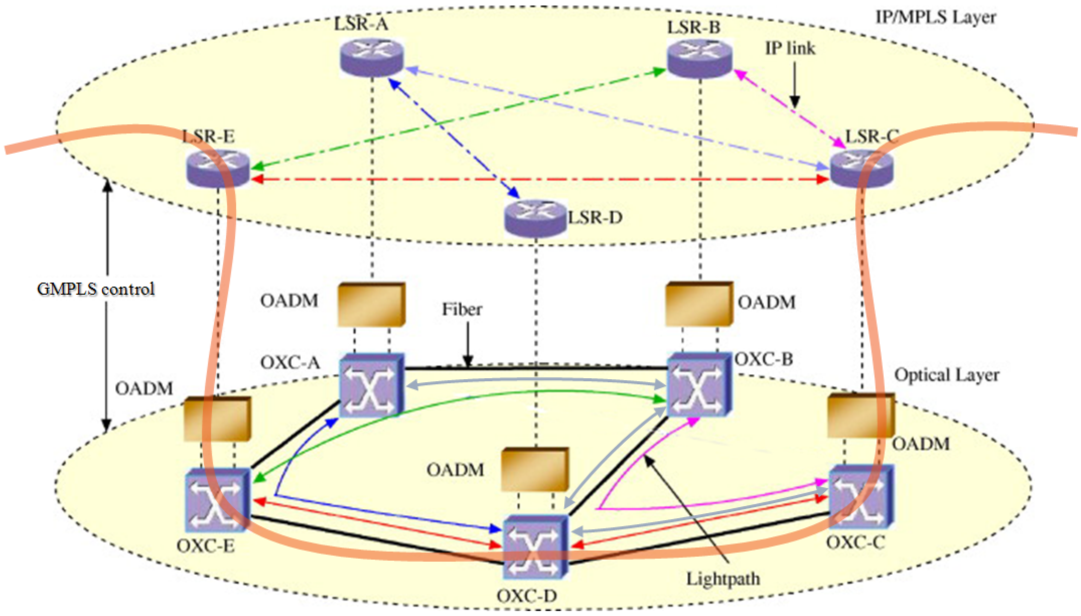


Figure 1.5: GMPLS-controlled connection-oriented multi-layer optical network (9).

link, respecting the wavelength continuity constraint (WCC) in all intermediate optical nodes that are not equipped with wavelength converters (WC). This is referred to as the routing and wavelength assignment (RWA) problem, whose objective is to minimize the blocking probability while satisfying the QoS constraints. The RWA problem, that can be naturally formulated as an integer linear programming (ILP) problem, has been shown to be NP-complete (10). The integrated RWA problem is referred to a combined wavelength routing and grooming optimization paradigm, taking into account the whole topology and resource usage information at both layers. In our model, the GMPLS control plane provides the appropriate routing and signaling protocols, such as the link-state OSPF-TE and RSVP-TE protocol, accurately disseminating correct and up-to-date information about the network state to each node, as well as taking care of resource reservation, allocation and release.

In such a context, we focused on the core network segment, and proposed several integrated RWA schemes working under different assumptions: complete or partial knowledge of network status, global control or individual selfishness of network elements, different requisites of computational and space complexity, etc. (*Task 1.a: Address the (energy-unaware) dynamic RWA problem, in order to accommodate the connection requests minimizing the blocking probability*).

In large-scale communication networks, like the Internet, it is usually unfeasible to globally manage network traffic. Accordingly, when modeling the traffic behavior in absence of global control, it is typically assumed that network users follow the most rational approach, that is, they behave selfishly to optimize their own individual welfare. Such a consideration motivates our RWA approach based on models from the Game Theory in (11) ("Selfish Routing and Wavelength Assignment strategies with advance reservation in inter-domain optical networks"), in which each player is aware of the situation facing all other players and tries to minimize his own cost. We re-formulated the RWA problem in modern connection-oriented all-optical network architectures by considering solution strategies from distributed multi-commodity network congestion games, which are solved by multiple agents operating in a non-cooperative but coordinated manner. The simulation results show that our approach may be particularly attractive for its scalability features and hence useful in large optical networks where many nodes, belonging to different administrative domains, operate selfishly by exchanging only a small amount of information needed for the coordination among them.

In (12) ("Constrained Minimum Lightpath Affinity Routing in multi-layer optical transport networks"), we presented a two-stage wavelength routing algorithm, easily integrable in state-of-the art routing and signaling protocols and technologies, built on an on-line dynamic grooming scheme that finds a set of feasible routes on lightpaths which fulfill some QoS and traffic engineering requirements and bases its final choice on a novel heuristic global path affinity minimization concept. The algorithm demonstrated the capabilities of achieving a better load balance and resulting in a significantly lower blocking probability than the existing methods for both optical networks under the wavelength continuity constraint and with sparse wavelength converters. The ability to guarantee both a low blocking probability and a low computational complexity make the on-line dynamic RWA algorithm very attractive for the modern multi-layer wavelength-switched networks and may truly become part of an effective and flexible control-plane framework.

In (13), a dynamic RWA scheme, called Spark, has been presented, easily integrable in state-of- the art routing and signaling protocols and technologies. Spark is conceived to work on both pure optical and hybrid electro-optical switching and routing devices, transparently handling grooming of lower rate connections, and capable to operate in presence of wavelength conversion devices. The algorithm, despite of its very low computational complexity, significantly lowers the blocking probability as compared with many widely used routing algorithms, thanks to its load balancing cost and scoring functions. These features, together with the high operating flexibility and configurability due to its native parametric design, and flexible network modeling framework, make the Spark algorithm an ideal candidate to be implemented in modern industrial optical control plane frameworks. Spark not only is a parametric – hence *tunable* – algorithm, but it also represents a reusable structure that can be used to obtain different versions of Spark itself. There are several aspects under which Spark may be modified for future works and investigations. The weighting function may be adapted to take into account also other features besides global and residual bandwidths, like, for example, the physical length of fibers, the average signal latency time on the fiber or the charge fee for the use of links, thus minimizing other objectives as well as the existing ones. The path scoring function may be further modified in order to give different importance to critical links (a critical link may be a link whose residual bandwidth dramatically decreases after the route of the connection request on it), even if Spark already implicitly handles critical links through its inherent traffic engineering and resources balancing behavior. In particular, Spark has been extended in GreenSpark (14) (“An Energy-Aware Dynamic RWA Framework for Next-Generation Wavelength-Routed Networks”), which takes into account the energy and GHG constraints (discussed in Section 1.3).

Dynamic demands and topology changes caused by the addition and deletion of new links and/or capacity, together with limited or inaccurate information available for dynamic routing and online routing decisions made on a best-of-now basis, without knowledge of the lightpaths to be set-up in the future, cause the wavelength routing logic to behave sub-optimally, causing inefficiency in the network resources usage and thereby creating opportunities for improvements. Lightpath topology re-optimization seizes on these opportunities and offers network operators the ability to better adapt to the network and user requests dynamics. This is achieved by regularly (or upon a particular event) re-routing the existing demands, temporarily eliminating the drift between the current solution and the optimal one that is achievable under the same conditions. Starting from the above premises, in (15)(“A GRASP-based network re-optimization strategy for improving RWA in multi-constrained optical transport infrastructures”) we formulated a hybrid approach for integrated online routing and offline reconfiguration of optical networks with sub-wavelength traffic (*Task 1.b: Address the connections re-optimization problem in networks*). The key feature of such a scheme is the ability to maintain the network balanced through adaptive on-demand re-optimization by ensuring that a sufficient capacity is kept available between any ingress-egress pair so that the maximum number of connections arriving to the network can be satisfied. The overall focus of this work has been on the balancing between the reconfiguration cost (in terms of disturbance to the users connections already deployed over the network) and a good and simple RWA and grooming solution. We defined a set of suitable goals and strategies for an integrated approach, and provided a formulation of the re-optimization procedure based on an iterative refinement process of multiple local search steps structured as a GRASP meta-heuristic procedure. We also developed a heuristic strategy that attempts to achieve minimal disturbance reconfiguration by performing local reconfiguration and delaying as possible the need for global reconfiguration. Furthermore, re-optimization would only occur when needed (when the rejection ratio become unacceptable and the potential savings from re-optimization exceeds some threshold) or upon certain events such as when new links are added or torn down. Simulation results show the notable margins of re-optimization achievable with our approach as well as the time complexity feasibility in real networks such as NSFNET and GEANT2. Rejection ratios of connection set-up requests decreased, allowing more connections to be successfully routed, and bandwidth gains have been observed in all the simulation runs. Besides, we proposed an efficient parallel implementation of GRASP with path-relinking that showed quite linear speedups in the number of processors and such a strategy has been successfully applied to greatly accelerate the proposed re-optimization scheme. The proposed re-optimization schema achieved prominent improvements in network efficiency, with the consequent cost savings.

In order to support all the research tasks, we developed SimulNet (16) ("SimulNet: a wavelength-routed optical network simulation framework"), a WDM-routed networks simulator, realized for the design and the evaluation of RWA and optimization algorithms (*Task 0: Develop a simulation framework to support the research tasks*). SimulNet has shown good flexibility in managing even complex networks and has exhibited accuracy of simulations results and satisfactory performances, making it a useful tool for the network research community.

In order to achieve the energy-oriented paradigm for the Internet infrastructure, the energy has to be considered as an additional constraint, and the so called “energy problem in the Internet” arises, which is briefly discussed in Section 1.3.

1.2.1 Motivations for the energy problem in the Internet

Nowadays it is becoming mandatory to consider energy-oriented solutions that explicitly take into account the energy and the GHG as additional constraints to operate in the Internet infrastructure. This motivation is supported by a number of factors:

1. the current worldwide energy shortages; the Internet infrastructure absorbs a notable share (12,6% (17)) of the worldwide energy production and it is responsible for the emissions of large quantity of GHG in the atmosphere;
2. the rising costs of energy as fossil fuels become scarcer;
3. the need for new alternative and renewable *green* sources of energy;
4. the growing interests of governments and society into eco-concerns.

In such a context, there is a lack of a comprehensive energy-oriented paradigm for the Internet infrastructure, comprising both energy-efficient architectures and energy-aware algorithms and protocols that take into account the absorbed energy, the emitted GHG and the availability of renewable energy sources. This Thesis is focused on these very issues and tries to address the lack of such a paradigm by proposing energy models for energy-efficient architectures, energy-aware algorithms and protocols conceived to optimize the use of energy and minimize GHG emissions while preserving the traditional criteria such as network and datacenters load balancing in order to serve as many demands as possible and thus maximizing the system availability.

1.3 The energy problem in the Internet

Human activities have severe impacts on the environment: energy-consumption, resources exploitation, GHG emissions, pollution, climate changes, global warming and global dimming form part of the human *ecological footprint*. The ecological footprint measures humanity’s demand on the biosphere in terms of the area of biologically productive land and sea required to provide the resources we use and to absorb our wastes.

The ecological footprint is thus a measure of human demand on the Earth's ecosystem. It compares human's demand with Earth's capacity to regenerate its resources. It is possible to estimate how many natural resources it would take to support humanity if everybody lived a given lifestyle. For 2009[1], human ecological footprint was estimated at 1.5 planet Earths (18) – in other words, humanity uses ecological services 1.5 times faster than Earth's capacity to renew them (by drawing from Earth's reserves)[2]. Simply stated, humanity's demands exceed our planet's capacity to sustain us. Therefore, energy consumption and the GHG emissions have imposed as new constraints for human activities. Furthermore, as fossil-fuels availability is becoming scarcer, the cost of energy is consequently increasing. Therefore, the human ecological footprint represents an energy (Watt), environmental (CO_2) and economic (€) problem that has to be accounted for also when dealing with the ICT sector and the Internet.

In fact, the Internet requires electric energy to work. In order to assess the GHG emissions of Internet equipment, it is necessary to study *where* the electrical energy comes from. Electrical energy is not directly present in nature; it is derived from primary energy sources, i.e. sources directly available in nature, such as oil, solar, nuclear energies, etc.

The conversion from primary to electrical energy is a process that may emit large quantity of GHG gases (*carbon footprint*). About 30% of the world primary energy is used to produce electrical energy (with an average yield of about 40%), and 7% of the worldwide produced electrical energy is absorbed by ICT infrastructures alone (only during the use phase), corresponding to 156 GW of electrical power and to 2-3% of the worldwide GHG emissions, as much as the aviation industry (19). Figure 1.6 illustrates the worldwide energy production and consumption in 2007 and the different shares for producing electrical energy.

When including embodied energy and replacement timespan, the Internet devices and infrastructure (routers, Wi-Fi/LAN, cell towers, telecom switches, fiber optics and copper cables, desktops, laptops, cloud, smartphones and servers) absorbs on average 12,6% of the worldwide produced electrical energy (17), corresponding to the equivalent

[1] Latest available data; every two years, this number is recalculated — with a three year lag due to the time it takes for the UN to collect and publish all the underlying statistics.

[2] If all the world population lived as the United States's does, it would require 4 planet Earths, plus 1 additional planet only for the garbage disposal.

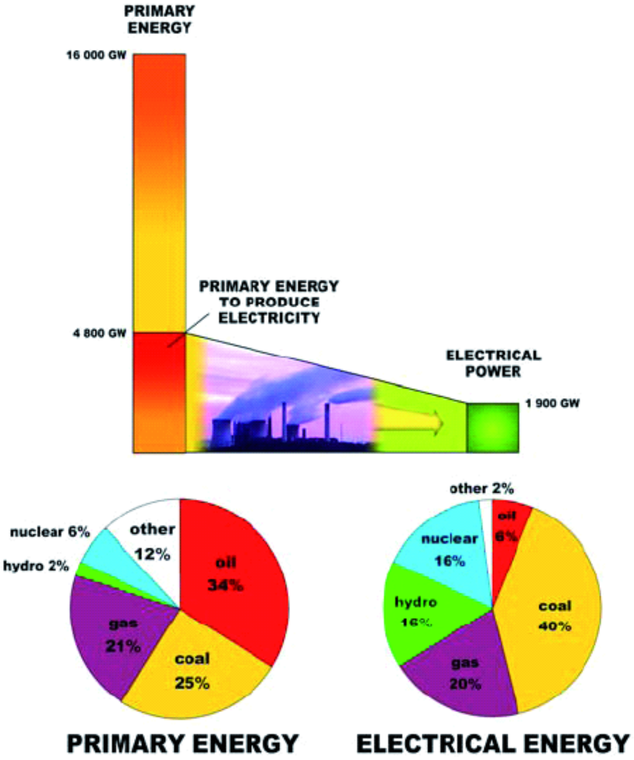


Figure 1.6: Worldwide energy production and consumption (19).

electricity production of 240 modern nuclear power reactors (between 1.1 and 1.9% of the 16 TW used by humanity (19)(20)).

Furthermore, primary energy sources can be divided into two categories: not-renewable and renewable energy sources. Not-renewable energy sources are essentially those coming from Earth's reserves (e.g. fossil fuels such as oil, coal, gas), whilst renewable energy sources are those coming from Earths's flows (e.g. sunlight, winds, tides, etc). Fossil fuels are burned to transform bio-chemical energy into electrical energy, and in this process large quantities of GHG (such as the carbon dioxide, CO_2, and other polluting particles) are emitted in the atmosphere, thus contributing to global warming and pollution. Nuclear energy is a not-renewable source of energy (since uranium and plutonium are available in limited quantities) and, although nuclear plants do not emit considerable amounts of CO_2, they do have other severe impacts on the environment. Nuclear energy is responsible of a massive ecological footprint: the intensive exploitation of natural resources for the extraction of the radioactive materials, the huge amounts of fresh water that are drained and warmed up to cool the reactors[1] and the dangerous radioactive wastes that are produced and for which there is no definitive solution for their disposal yet. As a result, not-renewable energy sources are *dirty* in the sense that they affect the environment in several ways.

Renewable energies (solar, wind, geothermal, hydro-electrical, tidal energies, etc.) represent part of the solution. They are a *limited* resource, since available in limited quantity at a time, but they are virtually *inexhaustible*, since – as the name suggests – they are *renewable*, meaning that they regenerate in a more or less cyclic fashion[2]. Besides, renewable sources are *clean* (usually referred as *green*) as they do not emit GHG during the energy production phase[3], although some drawbacks are still present (lower efficiency when compared to traditional dirty energy sources, visual and noise impact of wind turbine, large surfaces covered by solar panels, etc.).

However, to be effective, any new solution has to be evaluated under its *life-cycle assessment* (LCA). LCA comprises material extraction, transportation, production, use and disposal as the five phases in the life cycle of a product and all phases should be considered. The LCA provides a complete view of the environmental impact of a product (i.e., its ecological footprint) (21). It has been proved that renewable energy sources are beneficial over their entire life-cycle (22).

Since renewable energies are *green* and virtually *inexhaustible*, they are the perfect candidate to support the *eco-sustainable* growth. Nevertheless, renewable energy sources may not be always *available*; sun, wind and tide are cyclic or even almost unpredictable phenomena, though some inertia is guaranteed by energy storage systems (battery packs, potential energy accumulation systems, electric vehicles, etc.). In this sense, a *follow the sun/wind/tide* approach (23) and the knowledge of the *current* energy source and power consumption of the devices may be taken into account and exploited by an energy-aware paradigm to optimize the overall energy consumption and GHG emissions. The Internet infrastructure has inherent capability to exploit renewable energy sources since it mainly needs electrical power to operate (of which about 20% of GHG emissions come from manufacturing, while the remaining 80% comes from equipment use (24)).

[1] In France, 40% of the overall drained water is used to cool nuclear power reactors.

[2] The key benefit of renewable energy sources is that the energy comes from the nature's *flows*, and not from nature's *reserves*.

[3] Except biomasses and geothermal.

Accordingly, the efforts to address the Internet energy problem, which are gaining growing interest by the research community as well as by the governments and industries (25), are focused on three orthogonal dimensions:

- *Energy-efficiency:* refers to a technology designed to reduce the equipment energy consumption without affecting the performance, according to the *do more for less* paradigm. Such solutions are usually referred to as *eco-friendly* solutions;

- *Energy-awareness:* refers to an "intelligent" technology that adapts its behavior or performance based on the current working load and on the quantity and quality of energy that the equipment is expending (*energy-feedback information*). It implies knowledge of the type (green or dirty) of energy sources that supply the equipment thus differentiating *how* it is currently being powered. Energy-aware solutions are usually referred to as *eco-aware* solutions;

- *Smart Grids:* refers to a power distribution network in which consumers and providers have the knowledge of the quantity, quality and cost of energy flowing into the system, and in which the energy is exchanged between neighbors in a dynamic, adaptive fashion. Smart grids promise to change the traditional energy production/consumption paradigm in which one large (dirty) energy plant provides with energy the whole region, towards a configuration in which many small renewable energy plants (e.g. solar panels placed on the top of the buildings, wind turbines in the courtyards, etc.) interchange the energy. Each site becomes an energy consumer/producer, and the excesses are released into the smart grid, which redistributes it to the sites where the energy is needed or the renewable energy is not currently available. Smart grids open a new scenario in which the energy production and consumption can be closely matched avoiding peak power productions, and in which the energy quantity, quality and cost vary in function of the power plant producing it. Therefore, smart grids are foreseen to play a fundamental role in reducing GHG emissions and energy costs since they allow premises, datacenters, storage and computational resources to be interconnected to different energy sources and possibly dislocated near renewable energy plants or where the environmental conditions are favorable (e.g. cold climate can be exploited to efficiently cool down machines).

The three dimensions are orthogonal in the sense that they may and should act in concert. Energy-efficiency shall be applied to lower architectural levels and comprises technological innovations in order to execute a task with lower energy consumption with respect to previous solutions. Energy-awareness acts at higher levels which control the subordinate components in order to modify or adapt their behaviors to achieve lower overall energy consumption and globally lower GHG emissions. In an energy-oriented infrastructure, energy-efficient devices have to be managed by energy-aware technologies. In this sense, energy-efficiency may be seen like the efficient *body* and energy-awareness as the intelligent *mind* of an *energy-oriented* infrastructure.

Therefore, from a high-level perspective, energy-oriented solutions comprise energy-efficient devices, controlled by energy-aware algorithms and protocols, and powered by a smart grid power distribution network employing renewable energies, in a systemic approach encompassing the whole LCA, towards sustainable society growth and prosperity.

Accordingly, to achieve the energy-oriented paradigm for the Internet infrastructure, we proceeded step-wise, addressing specific problems at a time (identified by the tasks of the Thesis), and then tying them all together in a comprehensive energy-oriented framework.

As a first step in order to lower the energy consumption and the concomitant GHG emissions of the Internet infrastructure, it is necessary to assess the power consumption of current and future energy-aware architectures through extensive energy models that characterize the behaviors of the network equipment (*Task 2: Build energy models to represent energy consumption of network nodes, links and circuits*). In (26) ("Energy-oriented Models for WDM Networks"), the main energy models currently employed in the literature have been presented and discussed, providing an overview over the different scenarios that are currently being employed in WDM networks. It has also been presented a comprehensive energy model which accounts for the foreseen energy-aware architectures and the grow rate predictions, including different types of traffic of a WDM networks. The model, based on real energy consumption values, tries to collect the main benefits of the previous models while maintaining low complexity and, thus, high scalability.

Power management strategies that allow network infrastructures to achieve advanced functionalities with limited energy budget are expected to induce significant cost savings and positive effects on the environment, reducing GHG emissions (*Task 3: Analyze energy-efficient architectures in the literature and compare their effectiveness and practicability in routers*). Power consumption can be drastically reduced on individual network elements by temporarily switching off or downclocking unloaded interfaces and line cards. At the state-of-the-art, Adaptive Link Rate (ALR) and Low Power Idle (LPI) are the most effective local-level techniques for lowering power demands during low utilization periods. In (27) ("Analyzing Local Strategies for Energy-efficient Networking"), by modeling and analyzing in detail the aforementioned local strategies, we point out that the energy consumption does not depend on the data being transmitted but only depends on the interface link rate, and hence is throughput-independent. In particular, faster interfaces require lower energy per bit than slower interfaces, although, with ALR, slower interfaces require less energy per throughput than faster interfaces. We also note that for current technologies the energy/bit is the same both at 1 Gbps and 10 Gbps, meaning that the increase in the link rate has not been compensated at the same pace by a decrease in the energy consumption.

However, even though increased energy-efficiency reduces the power requirement without compromising the performance, it does not necessary lead also to reduced *overall* energy consumption and GHG emissions. In fact, any novel solution may present the *rebound effect* (also known as Jevons paradox or Kazzoom-Brookes postulate, depending on the context (23)(21)). An increase in the energy-efficiency may lead to decreased end-user costs, causing a rise in the demand. Such an increase may overcome the offset gained with the energy-efficiency, globally causing higher energy demand and GHG emissions. Therefore, in order to overcome the rebound effect, it is necessary to consider energy-aware solutions in a systemic approach encompassing smart grids with renewable energy sources, that are aware of the quantity and quality (green or dirty) of energy they are requiring and that adapt their behavior in function of such an information.

Towards this goal (*Task 4.a: Propose energy-aware algorithms to work in conjunction within the energy model to reduce the network ecological footprint*), ILP formulations have been presented in (28) ("Energy-Aware RWA for WDM Networks with Dual Power Sources") and (29) ("Towards an energy-aware Internet: modeling a Crosslayer Optimization approach") in order to formally characterize the energy-oriented RWA problem and its complexity. In (28), energy-aware ILP formulations exploiting dual energy sources have been presented along with an energy model in which no sleep mode is available but the optimization relies only on the traffic-variable power consumption of the network elements (NE). Two ILP formulations for the energy-aware RWA problem have been presented: minimum power (MinPower-RWA) and minimum GHG emissions (MinGas-RWA) strategies with the objective to minimize respectively the absorbed energy and the emitted GHG. Results show that the MinPower-RWA strategy may save a considerable amount of energy by routing the lightpaths on minimum consuming NEs and that the GHG emitted may be notably reduced by the MinGas-RWA strategy that prefers NEs powered by green energy sources. In (29), the energy-aware ILP formulations exploiting dual energy sources in (28) have been extended to comprise the connection requirements on the guaranteed bandwidth (lightpath bitrate, thus supporting lightpath with different bitrates), and to prove its effectiveness when the ALR energy-saving technique is employed. Results, obtained in the well-known NSFNET network topology, show that substantial savings are achievable both in terms of energy consumption and GHG emissions. Besides, as drops are observed in the day/night traffic at core network nodes, there is room for some possible optimizations by putting NEs into sleep mode only partially (per-interface sleep mode). In fact, putting into sleep mode single interfaces or line cards may have some sense, saving up to 50% of the total router power (30). The ILP formulations solves at optimum the static RWA problem, in which all the connection requests are known a priori, thus providing a lower bound for the achievable energy and GHG savings. In the dynamic RWA problem, no a priori knowledge is available on the connections requests that will arrive at the network; therefore, heuristic methods that find a high-quality solution in affordable computational time are required.

In (14) ("An Energy-Aware Dynamic RWA Framework for Next-Generation Wavelength Routed Networks"), a novel heuristic-driven dynamic RWA algorithm, called GreenSpark, has been proposed. GreenSpark aims at the minimization of power consumption and GHG emissions in wavelength-routed backbone networks. It operates by progressively routing the dynamically incoming connections on a two-stage basis; in the first stage, a set of k feasible paths is found according to traditional load-balancing objective. Then, in the second stage, the *greenness* of the k paths is evaluated in terms of power consumption (MinPower) and/or GHG emissions (MinGas), and the best path is finally selected to route the connection. Even with low k values (i.e. $k = 3$), and despite its very low computational complexity, GreenSpark achieves significant power savings and carbon footprint reduction together with an increment of the load-balance, resulting in lower blocking probability as compared with several widely used routing algorithms. Apart from defining an energy consumption model for the IP over WDM network, one of the most significant added values of the framework is the incorporation of both physical layer issues, such as power demand of each component, and virtual topology-based energy management with integrated traffic grooming, adversely conditioning the usage of energy hungry links and devices. Moreover, since the above model also takes into account the type of power supply associated with each device, by privileging green sources, the proposed scheme can also be useful for equalizing the carbon footprint of entire areas within a real network scenario in which each device location may be characterized by a differentiated (green or dirty) energy source. Here, multi-objective optimization may help us in finding the appropriate trade-off according to the relative importance of network performance and environmental friendliness.

These energy-aware RWA algorithms require an underlying routing protocol that distributes up-to-date information about the energy consumption and GHG emissions. In (31), (32) and (33), energy-aware OSPF-TE protocol extensions have been presented (*Task 4.b: Propose energy-aware routing protocols to disseminate energy consumption and GHG emissions of the network elements*). In (31), opaque link state advertisements (LSAs) of the OSPF protocol are used to implement the proposed extensions. Considering an agile implementation, new TLVs have been added directly to the TE extensions for OSPF (TE-LSA, Opaque Type 1). Each value in the TLV corresponds to a different type of energy source, where greater values correspond to higher CO_2 emissions (i.e. higher energy level). The proposed TE LSAs are flooded over the whole network on a fixed time-basis, informing the current energy source information per edge. A simple green routing algorithm is also proposed, aiming to route the traffic towards green energy sources. Results show a 16% reduction in the energy level and an increase of 3.3% at most in the blocking rate. In (32), the ESA routing and re-optimization algorithm is proposed to reduce CO_2 emissions in dynamic GMPLS controlled core optical networks. Results show that the ESA routing algorithm can decrease CO_2 emissions, compared to traditional shortest path and pure load balancing algorithms. Employing re-routing strategies together with the ESA algorithm can further bring down CO_2 emissions at the expense of increased blocking probability. By adding load balancing criteria, the algorithm can reach the lowest blocking probability in certain range. A clear trade-off is observed between connection blocking probability and obtained CO_2 savings. In (33), a green-aware routing algorithm, EE, is proposed, supported by proper underlying OSPF-TE protocol extensions. Green-awareness is enabled by flooding energy source information over the network, which is used in OSPF-TE routing decisions to lower the GHG emissions. Observing the behavior of the EE algorithm under different scenarios, it is seen that the proposed algorithm can save up to 27% of the GHG emissions (in terms of cost unit) at the expense of a marginal increase in the path length, compared to traditional shortest path routing algorithm. As a consequence of the higher mean path length and of the created "hot pot" effect, the blocking probability may increase up to 6% when the energy source updating interval is long. On the contrary, when the interval is short, the blocking probability of the EE algorithm reaches values even better than the traditional SP algorithm, due to the better load balancing induced by its energy cost function. However, this option may be limited due to the real dynamics of the green and dirty sources. An operator should also consider extra network overhead, and the possible additional expenses for obtaining the information from Smart grid network. When designing an energy efficient optical network, the proposed approach gives a direct insight into the behavior of a green-aware routing algorithm.

Apart from the network infrastructure, a considerable part of the energy consumed by the Internet is due to datacenters (the equivalent of 26 modern nuclear power reactors are required worldwide to power datacenters (19)). Therefore, a small step has been done in this sector to reduce datacenters ecological footprint (*Task 5.a: Propose energy-aware solutions for optimizing energy consumptions in datacenters and grid sites*).

In (34) ("Saving Energy in Data Center Infrastructures"), we presented EnergyFarm, an energy manager which can be used on the modern and future grid/cloud data center infrastructures to save energy. Current farms are usually over-provisioned and fluctuations in the traffic load are observed at various time periods. To take advantage of such a situation, we developed EnergyFarm which, through the service-demand matching algorithm and the job aggregation capabilities, allows turning off idle servers, while respecting both the demand requirements and the logical and physical dependencies. Results showed that great efficiency in the resource allocation can be achieved (between 20% and 68%), allowing significant energy, cost and emissions savings.

When considering energy-proportional devices, as servers in datacenters or energy-aware network routers, a new form of Denial of Service (DoS) attacks may be put in place: exploiting the computational and storage resources of datacenters with the aim of consuming as much energy as possible, causing detrimental effects, from high costs in the energy bill, to penalization for exceeding the agreed quantity of CO_2 emissions, up to complete denial of service due to power outages (*Task 5.b: Evaluate possible risks in ICT associated with the emerging energy problem*).

In (35) ("Evaluating Network-Based DoS Attacks Under the Energy Consumption Perspective"), we pointed out for the first time the risks related to energy-oriented attacks and, in particular, we evaluated the impacts of network-based distributed DoS (DDoS) attacks under the energy consumption perspective. We analyzed different types of such attacks with their impacts on the energy consumption, and showed that current energy-aware technologies may provide attackers with great opportunities for raising the target facility energy consumption and consequently its GHG emissions and costs: the more energy-proportional, the more vulnerable. DDoS attacks have the potential not only of denying the service of the target facility, but may be carved to explicitly impact its energy consumption. Such attacks may be targeted at several objectives: increment the energy consumption, the GHG emissions and introducing, in the worst cases, power outages. Some of these attacks are relatively easy to perform, e.g. CPU and I/O-bound based ones, whilst others are more difficult to deploy. We also pointed out that, in order to be successful under the energy-consumption perspective, a DDoS attack does not necessary need to penetrate into the target system, but its goal can be accomplished by simply having the intrusion detection/prevention systems (IDS/IPS) to work harder. In any case, the potential of such attacks should not be underestimated and effective power management techniques have to be deployed to prevent detrimental effects.

Finally, in (29) ("Towards an energy-aware Internet: modeling a Crosslayer Optimization approach"), a holistic vision on the energy-oriented Internet is provided in which energy-efficient architectures are powered by a smart grid power distribution system employing renewable energy sources and are controlled by an intelligent energy-aware control plane, which is able to operate the Internet to minimize its ecological footprint (*Task 6: Propose a comprehensive energy-oriented Internet infrastructure* *encompassing energy-efficient technologies, energy-aware algorithms and protocols and smart grid power distribution systems*).

The main ideas of the energy-oriented paradigm proposed in this Thesis have been presented also in the two book chapters (36) ("Towards Energy-Oriented Telecommunication Networks") and (37) ("Green Datacenter Infrastructures in the Cloud Computing Era'), which discuss the energy-oriented Internet paradigm from a high-level, divulging perspective.

1.4 Conclusions and future works

The Information and Communication Society (ICS) is experiencing an astonishing development driven by the possibilities offered by the Information and Communication Technologies (ICT). The Internet infrastructure (both network and cloud facilities) has to support larger and larger demands in terms of bandwidth, computing and storage resources, and one of the main limiting factor for its development is the energy demand and the concomitant GHG emissions. This fact represents the main motivation for this Thesis that, considering energy and GHG emissions as novel constraints, provides a systemic energy-oriented paradigm for the Internet, in which new energy models, protocols and algorithms optimize the Internet ecological footprint while not disrupting the performance, towards sustainable society growth and prosperity. The research carried out in this Thesis opens new perspectives towards sustainable Internet; starting from the works in this Thesis, a number of future works can be outlined.

In (28) and (29), we advised that modifications to current router architecture and routing protocols need to be investigated in order to support per-interface sleep mode of routers. Besides, renewable energy sources may vary their availability with time (e.g. solar panels only generate electricity during the day). While in the current work we handled the availability of green and dirty sources in a static way, in future works statistically variable green energy sources may be considered within a totally dynamic scenario in which the availability of the different types of renewable energies can be associated with the variations of the day time and traffic load (e.g. night/day cycle).

The RWA framework presented in (14) can be adapted to future mixed line rate and flexible-grid networks and the cost/scoring functions can be modified by introducing an omni-comprehensive energy-aware/load-balancing cost function to directly find green paths in a single stage with even lower computational complexity. In addition, new energy-aware traffic engineering strategies and network re-optimization methods can be investigated, aiming at dynamically reducing power demand, GHG emissions and costs on a time basis, by moving data wherever electricity costs are lowest at a particular time.

Further studies in line with (31), (32) and (33) can be focused on building a more detailed energy model, with a wider variety network elements taken into account. A more complex routing schema can be used, utilizing advanced constraint-based RWA algorithms. Furthermore, the extra network overhead for spreading the energy information as soon as a change in the energy source occurs can be studied, and the trade-off between the update frequency and the performance may be further investigated.

In (27), we point out that the different fixed and variable power consumptions of interfaces may studied to exploit circuit over-provisioning techniques as well as load balancing schemes for minimizing the overall energy consumption and, thus, network operational costs.

The network re-optimization framework presented in (15) can be extended in order to comprise energy-oriented criteria that take into account the current availability of renewable energy sources and try to minimize the ecological footprint of the network.

Finally, the ICT sector has the fundamental capability of acting as drawing factor to drive the development of energy-oriented technological innovations for both industry and society. We are confident that the above efforts, together with incrementing the Internet eco-sustainability, will improve the sustainable growth and – in the long run – the society prosperity.

References

[1] **CISCO Visual Networking Index [online]. Available:** http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html. 1, 2

[2] **Internet World Stats [online]. Available:** http://www.internetworldstats.com/emarketing.htm. 1

[3] **ETForecasts [online]. Available:** http://www.etforecasts.com/products/ES_intusersv2.htm. 1

[4] **BT announces major wind power plans [online]. Available:** http://www.btplc.com/News/Articles/Showarticle.cfm?ArticleID=dd615e9c-71ad-4daa-951a-55651baae5bb, 2007. 1, 2

[5] S.Pileri. **Energy and Communication: engine of the human progress**. In *INTELEC 2007, Rome, Italy*, Sept. 2007. 2

[6] L. Souchon Foll. **TIC et Énergétique: Techniques d'estimation de consommation sur la hauteur, la structure et l'évolution de l'impact des TIC en France**. In *Ph.D. dissertation, Orange Labs/Institut National des Télécommunications*, 2009. 2

[7] J. Baliga, R. Ayre, K. Hinton, W.V. Sorin, and R.S. Tucker. **Energy Consumption in Optical IP Networks**. *Journal of Lightwave Technology*, **27**(13):2391 –2403, july1, 2009. 3, 4

[8] **ZTE Green Technology Innovations White Paper, ZTE Technologies**, 2011. 5

[9] Xingwei Wang, Lei Guo, Xuetao Wei, Weigang Hou, Fei Yang, and Lan Pang. **Survivability in waveband switching optical networks: Challenges and new ideas**. *Computer Communications*, **31**(10):2435 – 2442, 2008. 5, 6

[10] I. Chlamtac, A. Ganz, and Karmi G. **Lightpath communications: An approach to high-bandwidth optical WANs**. *IEEE Trans. Commun.*, **40**:11711182, 1992. 6

[11] Francesco Palmieri, Ugo Fiore, and Sergio Ricciardi. **Selfish routing and wavelength assignment strategies with advance reservation in inter-domain optical networks**. *Computer Communications*, **35**(3):366 – 379, 2012. 7

[12] Francesco Palmieri, Ugo Fiore, and Sergio Ricciardi. **Constrained minimum lightpath affinity routing in multi-layer optical transport networks**. *J. High Speed Netw.*, **17**(4):185–205, December 2010. 7

[13] Francesco Palmieri, Ugo Fiore, and Sergio Ricciardi. **SPARK: A smart parametric online RWA algorithm**. *Journal of Communications and Networks*, **9**(4):368–376, 2007. 7

[14] Sergio Ricciardi, Francesco Palmieri, Ugo Fiore, Davide Careglio, Germán Santos-Boada, and Josep Solé-Pareta. **An energy-aware dynamic RWA framework for next-generation wavelength-routed networks**. *Computer Networks*, (0):–, 2012. 8, 17, 21

[15] Francesco Palmieri, Ugo Fiore, and Sergio Ricciardi. **A GRASP-based network re-optimization strategy for improving RWA in multi-constrained optical transport infrastructures**. *Computer Communications*, **33**(15):1809 – 1822, 2010. 8, 22

[16] F. Palmieri, U. Fiore, and S. Ricciardi. **SimulNet: a wavelength-routed optical network simulation framework**. In *Computers and Communications, 2009. ISCC 2009. IEEE Symposium on*, pages 281 –286, july 2009. 9

[17] Barath Raghavan and Justin Ma. **The energy and emergy of the internet**. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, HotNets-X, pages 9:1–9:6, New York, NY, USA, 2011. ACM. 10, 11

[18] WWF Global Footprint Network. **Living Planet Report 2010, The biennial report, 2010**. 11

[19] BONE project. **WP 21 TP Green Optical Networks, D21.2b Report on Y1 and updated plan for activities**, 2009. 11, 12, 19

[20] BP. **Statistical Review of World Energy, Jun. 2011**. 12

[21] W. Vereecken, W. Van Heddeghem, D. Colle, M. Pickavet, and P. Demeester. **Overall ICT footprint and green communication technologies**. In *Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on*, pages 1 –6, march 2010. 13, 16

[22] Christopher J. Koroneos, Yanni Koroneos. **Renewable energy systems: the environmental impact approach**. *International Journal of Global Energy Issues 27(4)*, pages 425–441, 2007. 13

[23] B. St Arnaud. **ICT and Global Warming: Opportunities for Innovation and Economic Growth**. 13, 16

[24] SMART 2020 Report. **Enabling the low carbon economy in the information age**, 2008. 13

[25] Global Action Plan Report. **An inefficient truth**, 2007. 14

[26] Sergio Ricciardi, Davide Careglio, Francesco Palmieri, Ugo Fiore, Germán Santos-Boada, and Josep Solé-Pareta. **Energy-Oriented Models for WDM Networks.** In Ioannis Tomkos, Christos Bouras, Georgios Ellinas, Panagiotis Demestichas, and Prasun Sinha, editors, *BROADNETS*, **66** of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 534–548. Springer, 2010. 15

[27] Sergio Ricciardi, Davide Careglio, Ugo Fiore, Francesco Palmieri, Germán Santos-Boada, and Josep Solé-Pareta. **Analyzing Local Strategies for Energy-Efficient Networking**. In *Networking Workshops*, pages 291–300, 2011. 16, 22

[28] S. Ricciardi, D. Careglio, F. Palmieri, U. Fiore, G. Santos-Boada, and J. Solé-Pareta. **Energy-Aware RWA for WDM Networks with Dual Power Sources**. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1 –6, june 2011. 16, 17, 21

[29] Sergio Ricciardi, Davide Careglio, Germán Santos-Boada, Josep Solé-Pareta, Ugo Fiore, and Francesco Palmieri. **Towards an energy-aware Internet: modeling a cross-layer optimization approach**. *Telecommunication Systems*, pages 1–22. 10.1007/s11235-011-9645-7. 16, 17, 20, 21

[30] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright. **Power Awareness in Network Design and Routing**. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pages 457 –465, april 2008. 17

[31] Jiayuan Wang, Sergio Ricciardi, Anna Manolova, Sarah Ruepp, Davide Careglio, and Lars Dittmann. **OSPF-TE Extensions for Green Routing in Optical Networks**. In *OptoElectronics and Communications Conference (OECC 2012)*, Jul. 2012. 18, 22

[32] Jiayuan Wang, Sergio Ricciardi, Anna Manolova Fagertun, Sarah Ruepp, Davide Careglio, and Lars Dittmann. **Energy-Aware Routing Optimization in Dynamic GMPLS Controlled Optical Networks**. In *International Conference on Transparent Optical Networks (ICTON 2012)*, Jul. 2012. 18, 22

[33] J. Wang, S. Ruepp, A.V. Manolova, L. Dittmann, S. Ricciardi, and D. Careglio. **Green-aware routing in GMPLS networks**. In *Computing, Networking and Communications (ICNC), 2012 International Conference on*, pages 227 –231, 30 2012-feb. 2 2012. 18, 19, 22

[34] S. Ricciardi, D. Careglio, G. Santos-Boada, J. Solé-Pareta, U. Fiore, and F. Palmieri. **Saving Energy in Data Center Infrastructures**. In *Data Compression, Communications and Processing (CCP), 2011 First International Conference on*, pages 265 –270, june 2011. 19

[35] F. Palmieri, S. Ricciardi, and U. Fiore. **Evaluating Network-Based DoS Attacks under the Energy Consumption Perspective: New Security Issues in the Coming Green ICT Area**. In *Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011 International Conference on*, pages 374 –379, oct. 2011. 20

[36] Sergio Ricciardi, Francesco Palmieri, Ugo Fiore, Davide Careglio, Germán Santos-Boada, and Josep Solé-Pareta. **Towards Energy-Oriented Telecommunication Networks**. In *Handbook on Green Information and Communication Systems, Elsevier*, 2012. 21

[37] Sergio Ricciardi, Francesco Palmieri, Jordi Torres-Viñals, Beniamino di Martino, Germán Santos-Boada, and Josep Solé-Pareta. **Green Datacenter Infrastructures in the Cloud Computing Era**. In *Handbook on Green Information and Communication Systems, Elsevier*, 2012. 21

Appendix A

List of Publications

A.1 In this thesis

A.1.1 Journals

1. **Sergio Ricciardi**, Francesco Palmieri, Ugo Fiore, Davide Careglio, Germán Santos-Boada, Josep Solé-Pareta, "An Energy-Aware Dynamic RWA Framework for Next-Generation Wavelength-Routed Networks", Computer Networks (2012), DOI: 10.1016/ j.comnet.2012.03.016. [**IF: 1.176 / Q2: Telecommunications**]

2. Francesco Palmieri, Ugo Fiore, **Sergio Ricciardi**, "Selfish Routing and Wavelength Assignment strategies with advance reservation in inter-domain optical networks", Computer Communications, vol. 35 issue 3, pp.366379, Feb. 2012, DOI: 10.1016/ j.comcom.2011.10.010. [**IF: 0.816 / Q2: Telecommunications**]

3. **Sergio Ricciardi**, Davide Careglio, Germán Santos-Boada, Josep Solé-Pareta, Ugo Fiore, Francesco Palmieri, "Towards an energy-aware Internet: modeling a Crosslayer Optimization approach", Telecommunication Systems Journal special issue on Green Telecommunications, Springer, DOI: 10.1007/s11235-011-9645-7. [**IF: 0.670 / Q3: Telecommunications**]

4. Francesco Palmieri, Ugo Fiore, **Sergio Ricciardi**, "Constrained Minimum Lightpath Affinity Routing in multi-layer optical transport networks", Journal of High Speed Networks, Volume 17, Number 4 2010, pp. 185-205, ISSN 0926-6801 (Print)

1875-8940 (Online), DOI: 10.3233/ JHS-2011-0340, Online Date March 18, 2011. [**IF: 0.379 / Q3: Telecommunications**]

5. Francesco Palmieri, Ugo Fiore, **Sergio Ricciardi**, "A GRASP-based network re-optimization strategy for improving RWA in multi-constrained optical transport infrastructures", Computer Communications (2010), Elsevier Journal, Volume 33 Issue 15, Pages: 1809-1822, September 2010, DOI: 10.1016/ j.comcom.2010.05.003. [**IF: 0.816 / Q2: Telecommunications**]

A.1.2 Conferences

6. Francesco Palmieri, **Sergio Ricciardi**, Ugo Fiore, "Evaluating Network-Based DoS Attacks Under the Energy Consumption Perspective", Broadband and Wireless Computing, Communication and Applications (BWCCA) 2011,vol., no., pp.374-379, 26-28 Oct. 2011, Barcelona, Spain, DOI: 10.1109/BWCCA.2011.66.

7. **Sergio Ricciardi**, Davide Careglio, Germán Santos-Boada, Josep Solé-Pareta, Ugo Fiore, Francesco Palmieri, "Saving Energy in Data Center Infrastructures", in proceedings of 1st International Conference on Data Compression, Communication and Processing (CCP2011), Jun. 21-24 2011, Palinuro, Italy.

8. **Sergio Ricciardi**, Davide Careglio, Ugo Fiore, Francesco Palmieri, Germán Santos-Boada, Josep Solé-Pareta, "Analyzing local strategies for energy-efficient networking", in Proceedings of the IFIP TC 6th international conference on Networking (NETWORKING'11), pp.291-300, LNCS 6827, Valencia, Spain, 9-13 May 2011.

9. **Sergio Ricciardi**, Davide Careglio, Francesco Palmieri, Ugo Fiore, Germán Santos-Boada, Josep Solé-Pareta, "Energy-Aware RWA for WDM Networks with Dual Power Sources", in Proceedings of 2011 IEEE International Conference on Communications (ICC 2011), Kyoto, Japan, June 5-9, 2011. [**UPC notable conference**]

10. **Sergio Ricciardi**, Davide Careglio, Francesco Palmieri, Ugo Fiore, Germán Santos-Boada, Josep Solé-Pareta, "Energy-oriented Models for WDM networks", in Proc. of BROADNETS 2010, LNICST 66, pp. 534547, 2012.

11. Francesco Palmieri, Ugo Fiore, **Sergio Ricciardi**, "SimulNet: a wavelength-routed optical network simulation framework", proceedings of IEEE Symposium on Computers and Communications (ISCC'09) Sousse, Tunisia, July 5 - 8, 2009, Pages: 281-286, IEEE Xplore, DOI: 10.1109/ ISCC.2009.5202259. [**UPC notable conference**]

A.1.3 Book Chapters

12. **Sergio Ricciardi**, Francesco Palmieri, Ugo Fiore, Davide Careglio, Germán Santos-Boada, Josep Solé-Pareta, "Towards Energy-Oriented Telecommunication Networks", accepted for publication in the Handbook on Green Information and Communication Systems, Elsevier, 2012.

13. **Sergio Ricciardi**, Francesco Palmieri, Jordi Torres-Vials, Beniamino di Martino, Germán Santos-Boada, Josep Solé-Pareta, "Green Datacenter Infrastructures in the Cloud Computing Era", accepted for publication in the Handbook on Green Information and Communication Systems, Elsevier, 2012.

A.2 Other publications

A.2.1 Journals

14. Luis Velasco, Oscar González de Dios, **Sergio Ricciardi**, Albert Castro, Fernando Muoz, Davide Careglio, Jaume Comellas, "Value Optimization of Survivable Multilayer IP/MPLS-over-WSON Networks", Photonic Network Communications (PNET), Springer Journal, DOI:10.1007/s11107-011-0354-7. [**IF: 0.600 / Q3: Telecommunications**]

15. Francesco Palmieri, Ugo Fiore, **Sergio Ricciardi**, "A Minimum Cut Interference-based Integrated RWA Algorithm for Multi-constrained Optical Transport Networks", Journal of Network and Systems Management, Springer Journal, vol. 16 No. 4, pp.421-428, ISSN 1064-7570 (Print) 1573-7705 (Online), DOI:10.1007/s10922-008-9097-x, April 04, United States, 2008. [**IF: 0.450 / Q3: Telecommunications**]

16. Francesco Palmieri, Ugo Fiore, **Sergio Ricciardi**, “SPARK: A Smart Parametric Online RWA Algorithm”, Journal of Communications and Networks, Vol. 9, No. 4, December 2007, pp. 368-376, ISSN 1229-2370. [**IF: 0.351 / Q4: Telecommunications**]

A.2.2 Conferences

17. Jiayuan Wang, **Sergio Ricciardi**, Anna Manolova, Sarah Ruepp, Davide Careglio, Lars Dittmann, “OSPF-TE Extensions for Green Routing in Optical Networks”, accepted to OECC, 2-6 Jul. 2012.

18. Jiayuan Wang, **Sergio Ricciardi**, Anna Manolova Fagertun, Sarah Ruepp, Davide Careglio, Lars Dittmann, “Energy-Aware Routing Optimization in Dynamic GMPLS Controlled Optical Networks”, ICTON, 2-5 Jul. 2012. [**UPC notable conference**]

19. **Sergio Ricciardi**, Germán Santos-Boada, Davide Careglio, Jordi Domingo-Pascual, “GPON and EP2P: A Techno-Economic Study”, NOC 2012, June 19-22, 2012, Vilanova i la Geltru, Spain.

20. Benjamin Peterson, **Sergio Ricciardi**, Jordi Nin, “Energy-efficiency and Security Issues in the Cisco Nexus Virtual Distributed Switching”, in proceedings of The Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2012), Palermo, Italy, July 4-6, 2012.

21. Jiayuan Wang, Sarah Ruepp, Anna Vasileva Manolova, Lars Dittmann, **Sergio Ricciardi**, Davide Careglio, “Green-Aware Routing in GMPLS Networks”, 2012 International Conference on Computing, Networking and Communications (ICNC), vol., no., pp.227-231, Jan. 30 2012-Feb. 2 2012, DOI: 10.1109/ ICCNC.2012.6167416.

22. **Sergio Ricciardi**, Davide Careglio, Germán Santos Boada, Josep Solé-Pareta, “Energy Aware Paradigm for Energy Efficient ICT: a Systemic Approach”, proceedings of 1st International Conference on Energy-Efficient Computing and Networking (e-Energy 2010), Apr. 13 - 15 2010, Passau, Germany.

23. Alessandra Doria, Gianpaolo Carlino, Salvatore Iengo, Leonardo Merola, **Sergio Ricciardi**, Maria Carla Staffa, "Powerfarm: a power and emergency management thread-based software tool for the ATLAS Napoli Tier2", proceedings of Computing in High Energy Physics (CHEP) 21 - 27 March 2009, Prague, Czech Republic, Journal of Physics: Conference Series (JPCS), volume 219, part 5, IOP Publishing, DOI: 10.1088/1742-6596/ 219/5/052018.

A.2.3 Conferences (in Italian)

24. **Sergio Ricciardi**, Alessandra Doria, Rosario Esposito, "Progetti SCoPE/ATLAS: Il sistema di monitoring di risorse hardware/software e di servizi OS/Grid", Proceedings of the Conference Italian e-Science (IES08), 27-29 May 2008, Napoli, Italy, p. 170.

25. **Sergio Ricciardi**, Natascia De Bortoli, Salvatore Iengo, Mariacarla Staffa, "Powerfarm: Un software basato sui thread per la gestione dell'energia elettrica e delle emergenze per i siti Grid e centri di calcolo", Proceedings of the Conference Italian e-Science (IES08), 27-29 May 2008, Napoli, Italy, p. 169.

Appendix B

Compendium of Papers

In this section, the papers forming part of this Thesis are reported in their original versions, as they have been published.

The list of publications forming part of this Thesis is reported in page 29, section A.1.

Computer Communications 35 (2012) 366–379

Contents lists available at SciVerse ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Selfish routing and wavelength assignment strategies with advance reservation in inter-domain optical networks

Francesco Palmieri[a,*], Ugo Fiore[b], Sergio Ricciardi[c]

[a] Seconda Università degli Studi di Napoli, Dipartimento di Ingegneria dell'Informazione, Via Roma 29, 81031 Aversa (CE), Italy
[b] Università degli Studi di Napoli Federico II, CSI, Complesso Universitario Monte S. Angelo, Via Cinthia, 80126 Napoli, Italy
[c] Universitat Politècnica de Catalunya (UPC), Departament d'Arquitectura de Computadors (DAC), Carrer Jordi Girona 1–3, 08034 Barcelona, Catalunya, Spain

ARTICLE INFO

Article history:
Received 1 February 2011
Received in revised form 28 July 2011
Accepted 13 October 2011
Available online 20 October 2011

Keywords:
Wavelength-routing
RWA
Multi-commodity network congestion games
Nash equilibrium

ABSTRACT

The main challenge in developing large data network in the wide area is in dealing with the scalability of the underlying routing system. Accordingly, in this work we focus on the design of an effective and scalable routing and wavelength assignment (RWA) framework supporting advance reservation services in wavelength-routed WDM networks crossing multiple administrative domains. Our approach is motivated by the observation that traffic in large optical networks spanning several domains is not controlled by a central authority but rather by a large number of independent entities interacting in a distributed manner and aiming at maximizing their own welfare. Due to the selfish strategic behavior of the involved entities, non-cooperative game theory plays an important role in driving our approach. Here the dominant solution concept is the notion of Nash equilibria, which are states of a system in which no participant can gain by deviating unilaterally its strategy. On this concept, we developed a selfish adaptive RWA model supporting advance reservation in large-scale optical wavelength-routed networks and developed a distributed algorithm to compute approximate equilibria in computationally feasible times. We showed how and under which conditions such approach can give rise to a stable state with satisfactory solutions and analyzed its performance and convergence features.

1. Introduction

The large potential bandwidth available in next generation wavelength-division multiplexed (WDM) optical networks makes this technology of crucial importance for satisfying the ever-increasing capacity requirements in communication networks. Such networks will be based on dynamically configurable switching nodes, connected though a mesh of fiber links and operating transparently at the wavelength layer according to several automatic control plane strategies and protocols. These nodes set up and tear down, on a customer's request basis, pure photonic end-to-end communication channels (lightpaths) that can traverse multiple physical links on a common wavelength and essentially create a virtual topology on top of the physical topology. Information sent via a lightpath does not require to be converted from the optical to electrical form when passing through an intermediate node and converted back to the optical domain for retransmission to the next station, greatly reducing delay and latency phenomena and achieving transfer rates in the order of tens of THz. The efficient allocation of lightpaths on the fiber mesh, given a set of requests between pairs of nodes wishing to communicate through a dedicated end-to-end channel, poses several interesting theoretical problems. Given an optical network and a set of end-to-end communication requests, the routing and wavelength assignment (RWA) problem concerns routing each request on the optical transport network, and assigning wavelengths to these routes so that the same wavelength must be assigned along the entire route (*wavelength continuity constraint*), by realizing a lightpath [1]. Obviously, lightpaths that share a common physical link cannot be assigned the same wavelength (*clash constraint*). The objective of the RWA problem, that has been shown to be NP-complete [2], can be usually associated to the optimization of the overall network resources usage together with the minimization of the number of wavelengths used, or the maximization of the number of lightpaths successfully set up subject to a limited number of available wavelengths. However, if the needed wavelength resources are not immediately available at the request time, the connection setup will be blocked and the associated request refused. This may be intolerable for all the network users that require connection services being set up within a specified time frame and for a specified duration, according to a request/booking schema. To provide such services, it is desirable that the network resource control and management logic support advance reservations, i.e. reserving wavelength resources in advance respect to when they are actually

* Corresponding author. Tel./fax: +39 081503 7042.
E-mail addresses: francesco.palmieri@unina.it (F. Palmieri), ugo.fiore@unina.it (U. Fiore), sergio.ricciardi@ac.upc.edu (S. Ricciardi).

0140-3664/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2011.10.010

needed. This is obviously another useful network service model, which cannot only provide guaranteed services to network users but also allow networks to better plan their wavelength allocations. In fact, advance reservations of network resources are especially useful in environments that require reliable synchronized allocations of various resource types at different locations. Such *co-allocations* are necessary in order to assure that all the resources required are available at a given time. Each request specifies an end-to-end connection between two involved nodes, with a specific duration and a scheduling window, i.e. the time period within which the requestor would accept the connection to be set up. The flexibility of network-aware *advance resources reservation* introduces a new temporal dimension into the overall resource allocation problem. To support advance reservations, an RWA algorithm must take into account not only the network's spatial and topological characteristics (links, wavelengths, traffic matrix) but also their temporal characteristics. This would greatly increase the computational complexity of an RWA algorithm. Furthermore, in a real-world large-scale scenario the switching nodes and fiber links are owned and managed by several independent socio-economic organizations often operating in a non-cooperative fashion. According to the distributed nature of the Internet, in fact, these entities typically prefer to take almost unilateral decisions, such as selecting a path to route a connection request from one of their customers, in order to optimize *their own* resource usage and, of course, maximize their revenue. The lack of a central regulation forcing all the nodes to behave according to a common strategy makes network-wide resource optimization very difficult or even impossible. It should also be noted that end-to-end lightpath selection schemes are selfish by nature in that they allow the providers handling the connection request to greedily select the best available routes to optimize their own performance without considering system-wide criteria. Hence, the understanding of the mechanisms behind the selfish behavior of the involved entities in such non-cooperative network systems is of primary importance in resolving large-scale RWA problems where each organization that has to route a set of end-to-end connection request is driven by completely different and even conflicting measures of performance and optimization criteria. A natural framework in which to study such multi-objective optimization problems is the classic game theory. In such a context, our optimization problem can be modeled as a non-cooperative game of independent entities (*players*). These entities do not operate according to a common strategy and act in a purely selfish manner, aiming to maximize their own objective functions. The algorithmic game theory predicts that selfish behavior in such a system can lead to a Nash equilibrium, that is a state of the system in which no player can gain by unilaterally changing its strategy [3]. This approach can be used to optimize global objective functions taking into account the selfish behavior of the participating entities. That is, in such situations where it is difficult or even impossible to impose optimal routing strategies on network traffic, we exploit some less evident interaction dynamics between all the player's choices so that selfish behavior leads to a socially desirable outcome. Players, according to an advance reservation scheme, selfishly choose their private strategies, which in our environment correspond to best paths from their sources to their destinations, apparently without considering the other players' strategies. In doing this, our schema, ensures that, at the beginning of each reservation, specific additional taxes/marginal costs – associated to the conflicts with the other strategies insisting on the same resources – are bounded to the network resources. These costs can implicitly condition the selection process so that the global game may be kept into an equilibrium state. In other words, although each connection request is handled selfishly, it is deterministically assigned on its minimum-latency path (considering the network "congestion" effect due to the other players' impact), from which the corresponding entity/player has no incentive to deviate unilaterally. Extensive simulations have been carried out to evaluate the performance and scalability of the proposed approach in terms of tolerance to large number of simultaneous advance reservation requests as well as in slow connection rejection rate growth in presence of increasing network congestion. Good performance, limited cooperation between nodes and low computational complexity make the proposed model attractive for the future optical wavelength switched Internet.

2. Background

This section briefly introduces some of the concepts that will be useful to better explain the RWA optimization approach, by presenting the existing related literature together with the underlying architectural scenario, the basic assumptions, building blocks and modeling details as long as the theory behind them.

2.1. Related work

The RWA problem in large-scale all-optical networks has been intensely studied. In recent years, many research efforts have targeted the improvement in the efficiency of the management and control layer, following several directions, among which fuzzy ILP [4], and resource-criticality based heuristics aiming to delay as much as possible the utilization of critical resources, reserving them for future lightpath demands [5,6]. Advance reservation in optical networks has also been extensively studied. Zheng et al. [7] present a basic framework for automated provisioning of advance reservation services based on GMPLS protocol suites. In [8], a simulated annealing based algorithm is proposed to find a solution on predetermined *k*-shortest paths. Lee et al. [9] propose an efficient Lagrangean relaxation approach to resolve advance lightpath reservation for multi-wavelength optical networks. Other works (for example [10]) concentrate on the flexibility in reserving the connections, considering that clients may prefer a moderate delay in the start time of their request rather than having a request blocked. Finally, the exploitation of game theoretic approaches based on the analysis of uncooperative interactions and Nash equilibria in communication networks gave rise to a vast literature [11,12]. The work in [13] analyzed Nash equilibria, by considering their Price of Anarchy (PoA) and Price of Stability (PoS), in selfish routing games on multiple parallel links, where each player desires to minimize his experienced transmission time and seeks to communicate a message by choosing one of the links. In [14] the authors studied atomic routing games on networks, where each player chooses a path to route the traffic from an origin to a destination, with the objective of minimizing the maximum congestion on any edge of the path. Selfish path coloring in single fiber all-optical networks has been studied in [15,16], where the authors investigate the existence and performance of Nash equilibria, considering several information levels of local knowledge that players may have and give bounds for the PoA in chains, rings and trees. Selfish routing games have also been explicitly studied in ring networks [17] by adopting the asymmetric atomic routing model with a load-dependent linear latency on each link. The work in [18] analyzed the existence and complexity properties of pure Nash equilibria and best-response strategies in congestion games with time-dependent costs, in which travel times are fixed but QoS varies with load over time. The complexity of recognizing and computing Nash equilibria under various payment functions has been also studied in [19] where Fanelli et al. analyzed the payment functions in two different settings, both characterized by the objective of minimizing the total number of wavelengths used and minimizing of the number of converters needed. The PoA of selfish routing and path coloring, under payment functions that charge a player only according to his own strategy is discussed in [20,21]. Selfish path multi-coloring games where routing decision are taken in advance and players choose only colors are introduced in [22], providing bounds for the pure price of anarchy and also constant bounds for the PoA in specific topologies.

2.2. Network congestion games

Rosenthal [23] introduced a class of games, called congestion games, in which each player chooses a particular subset of resources out of a family of allowable subsets for him (its *strategy set*), constructed from a basic set of primary resources for all the players. The cost or delay associated with each primary resource e is a non-decreasing function $c_e(x)$ of the number of players x who choose it, and the total cost received by each player is the sum of the costs associated with the primary resources he chooses. In a *multi-commodity network congestion game*, each player is associated to a traffic flow to be routed throughout a network and its strategy set is represented as a set of origin–destination paths in such a network, whose edges play the role of resources. The flow may be *unsplittable*, in which case each player must choose a single path for its entire flow, or *splittable*, if the opposite is true. Furthermore, in the *atomic* case, there are a finite number of players, each with a specific amount of flow to route whereas, in the non-atomic case, there are an infinite number of players, and each one controls only a negligible fraction of the total flow. In addition, a *weighted congestion game* allows users to have different demands for service and, thus, affects the resource delay functions in a different way, depending on their own weights. In modeling the RWA optimization problem we refer to the atomic unsplittable model, where players have to route their connection demands along a single path (as general case, a demand may be split into n flows, but in the optical domain these streams will appear as n unsplittable optical flows). In such a *multi-commodity network congestion game*the strategy set of each player is represented as a set of origin–destination paths in a network, where the adjacencies between nodes and the associated weights/costs play the role of resources. A game with $n \geqslant 2$ players is defined by a finite set of strategies S_i with $i \in [1,n]$ where S_i denote all the possible strategies of the player i, and n cost functions $f_i : S_1 \times \cdots \times S_n \to R$, one for each player, mapping the set of all the possible strategies for each player to the real number set (some of the works present in the literature focus on payoff functions instead of cost functions; clearly, the difference is only a change in sign). The elements of $S_1 \times \cdots \times S_n$ are called states. The possible strategies for each player are implicated by both the topology of the network and the cost associated to each link. A *pure strategy profile*, or simply strategy profile, is a vector $\overrightarrow{S} = (s_1, \ldots, s_n)$ of deterministically chosen strategies, one for each player. Starting from the strategy profiles for all players and given a set of the strategies unilaterally chosen from each player, we say that the game is in an *equilibrium* if no player can decrease its own cost by changing its choices. This equilibrium concept was first introduced by John Nash [24] and it is known as Nash equilibrium. Such equilibrium defines a fundamental point of stability within the system, because no player can unilaterally perform any action to improve its situation. It is very interesting to explore the existence of *pure* Nash equilibria (PNE) in such games: a strategy profile is a *pure Nash equilibrium* if for each player i it holds that:

$$f_i(s_1, \ldots, s_i, \ldots, s_n) \leqslant f_i(s_1, \ldots, s'_i, \ldots, s_n) \tag{1}$$

for any strategy $s'_i \in S_i$.

Although Nash showed that each non-cooperative game can converge to a Nash equilibrium, the existence of a PNE is an open question for many games. Moreover, due to the selfish behavior of the players, such a pure equilibrium does not necessarily optimize a global goal. Such a goal is also known as the *social cost* of a strategy profile $\overrightarrow{S}$, defined as:

$$sc(\overrightarrow{S}) = \max_{i \in [1,n]} f_i(\overrightarrow{S}). \tag{2}$$

Depending on the involved cost function, the players' selfish behaviors might not optimize the social cost. It is also well known that a Nash equilibrium does not necessarily need to minimize the social cost. At the other end, the network management objective is minimizing the *social cost* measured by the total cost incurred by all players. The global performance of Nash equilibria is measured by the so-called *Price of Anarchy* (*PoA*) or *coordination ratio* which is defined as the ratio of the social cost of the worst Nash equilibrium over the optimal solution [25], and reflects the loss in the global performance due to lack of coordination between players:

$$PoA = \frac{\max_{\overrightarrow{S} \text{ is a NE}} sc(\overrightarrow{S})}{opt} \tag{3}$$

where

$$opt = \min_{\overrightarrow{S} \in (S_1 \times \cdots \times S_n)} f_i(\overrightarrow{S}) \tag{4}$$

denotes the optimum social cost for a game.

Clearly, a game with a low Price of Anarchy can be reasonably left almost unconditioned, since the involved selfish players — by virtue of being selfish — are guaranteed to achieve an acceptable performance. On the other hand, in presence of a large Price of Anarchy, it is necessary to introduce some *social control* and *coordination mechanisms* (such as taxes, costs or incentives, etc.) that implicitly force players to collaborate more efficiently. Some congestion games admit a potential function defined over the set of pure strategy profiles, with the property that the gain of a player unilaterally shifting to a new strategy is equal to the corresponding increment in the potential. It has been shown [26] that the existence of such a potential function implies that at least one Nash equilibrium exists. Formally, a real function $\Phi(\overrightarrow{S}) : S_1 \times \cdots \times S_n \to R$ is a β-potential function if it has the property that:

$$\begin{aligned} &f_i(s_1, \ldots, s_i, \ldots, s_n) - f_i(s_1, \ldots, s'_i, \ldots, s_n) \\ &\quad = \beta_i \cdot (\Phi(s_1, \ldots, s_i, \ldots, s_n) - \Phi(s_1, \ldots, s'_i, \ldots, s_n)), \end{aligned} \tag{5}$$

where the β_i are the real-valued components of a vector β. The effect on the cost function f_i of a strategy change by player i will then be the projection, weighted through the vector β, of the variation in potential associated to the change, so that local minima of the potential function will correspond to Nash equilibria. Such equilibria exist, and can be computed in pseudo-polynomial time, in games with linear cost functions $c_e(x)$ associated to the individual resource e [3], where each $c(x)$ function can be defined as:

$$c_e(x) = a_e x + b_e, \tag{6}$$

where a_e, $b_e \geqslant 0$ are constant values conditioning the cost function trend.

2.3. Marginal costs in congestion games

As already sketched in the previous section, to mitigate the performance degradation due to the players' non-cooperative and selfish behavior, we can introduce some incentives that influence the players' selfish choices and hopefully induce an optimal network configuration. These incentives can be naturally modeled by non-negative per-unit-of-traffic *taxes* (or prices) assigned to the resources. Such taxes become an additional cost factor, which the players should take into account. Simply stated, a player's cost for adopting a strategy should be calculated by adding such marginal cost associated to choosing a specific resource to the latency due to the resource's congestion. Although these additional costs increase the players' individual cost, they do not affect the social cost because they are payments inside the system and can be feasibly "refunded" to the players. The goal is to find a set of moderate and efficiently computable *optimal marginal costs*, which make the Nash equilibria of the modified game coincide with the optimal solution. Designing optimal taxes is a central topic in game theory. In general, any traffic equilibrium reached by the selfish players who are conscious of both the resource usage latencies and the taxes will minimize the social cost, that is, will minimize the total latency [27]. According to [28], we can formally define the marginal cost associated to a resource e by:

$$c_e^*(x) = c_e(x) + x \cdot c_e'(x), \tag{7}$$

where $c_e'(x)$ denotes the derivative $\frac{d}{dx}c_e(x)$.

Observe that the function $c_e^*(x)$ describing the marginal cost of increasing traffic on the resource e is composed by a first term capturing the per-unit latency incurred by the additional traffic introduced by the other players' choosing e and a second one accounting for the increased congestion experienced by the traffic already using the resource. Essentially, the only difference between an optimal route assignment and an assignment in the context of a Nash equilibrium is that the former accounts for this "conscious" second term while the latter disregards it.

3. The reference model

We will model our approach to the RWA problem with a game theoretic formulation by working in an atomic unsplittable weighted multi-commodity scenario where the communication resources are booked and allocated according to a time-slotted advance reservation paradigm. This means coping with a "scheduled" traffic model where the setup and teardown times of the demands are known in advance.

It is common in this setting to view the wavelength routed network as a connected graph with its nodes being the optical switching nodes and its edges being the available wavelengths (i.e. different channels) on the optical fibers that provide the actual communication. Since each fiber link can support several WDM channels, there is typically more than one edge connecting the same pair of nodes. The resulting structure is a multi-graph and its construction process is sketched in Fig. 1 below.

To keep the formulation as general as possible, we make no specific assumption on the number of wavelengths per fiber and the number of fiber on each link. All these parameters are fully and independently configurable at the network topology definition time. Nevertheless, we require that all the network nodes operate under a unique control-plane providing a common link-state routing protocol and a signaling facility to handle resource reservations (such as those provided in a multi-domain GMPLS-like framework [29]). Furthermore, we assume that every connection is bidirectional and consists in a single atomic traffic flow that cannot be split between multiple paths (as we have seen, such assumption does not cause any loss of generality). Each connection request, viewed as an independent player in our game-theoretic approach, can be satisfied by establishing a single lightpath between its source and destination nodes. Notice that, to enforce the continuity constraint, this path can only be built on edges associated to the same wavelength. We are given a network (graph) $G = (V,E)$ and a set of end-to-end connection requests $R = (r_1,r_2,\ldots,r_{|R|})$ arriving as an ordered sequence according to a Poisson process with exponentially distributed call-holding time. In our advance reservation model, the time is slotted with a slot size equal to t', where this length depends on the minimum duration of an advance reservation (Fig. 2). Each advance reservation request r_i can only start at the beginning of a timeslot and is described by a 5-tuple, (u,v,d,s,e) where u and v are the nodes in G that are the connection's ingress and egress points, d is the reservation duration expressed in time slots, and s and e are the starting and ending time of the *scheduling window*. The scheduling window defines the acceptable set-up time range of the connection request, so that if the needed connection cannot be established within such time period, the request will be withdrawn. The window size may be fixed if the start and end times of the connection cannot be altered or flexible when those time limits can slide within a larger window. Several integer linear program formulations and algorithms have been proposed to solve these problems [30,31]. In our work we will consider dynamic end-to-end connection requests that belong to a fixed scheduling window.

Despite losing granularity, the above time slotted model allows the reduction in required processing capacity and increases scalability. When applying it to wavelength routed optical networks we obtain a multi-dimensional resource management scenario with hops, wavelengths and time slots. An online instance of the RWA problem is denoted by (G,R) and is defined as the task of finding an assignment of valid single-wavelength paths, at the granularity of a timeslot and for an integer number of timeslots, to a subset of requests $A \subseteq R$ with different wavelengths for overlapping paths, such that $|A|$ is maximal. This is an online scheduling problem because the requests arrive dynamically and, at each time slot, for each request $r_i \in R$, we check if it is inside its validity time range and, if so, compute a path and check whether a common wavelength on each link of this path can be reserved for its duration d. If such a suitable path is not available, the involved connection request will be deferred to the next time-slot, and this process will be iterated until either the request is satisfied or its time window expires. In order to implement this advance reservation mechanism, the RWA logic needs to maintain a schedule of the valid reservations called the reservation table. Also, the network nodes must work in a synchronized way according to a common reference clock. A strategy s_i for player i is a pair (p_i,d_i) where p_i is a simple path connecting the endpoints of r_i and d_i is its requested duration, implicitly associated to all the edges in p_i. Each player's strategy set consists of k different source–destination paths $(s_1,\ldots,s_k)$, corresponding to the first k available minimum cost path choices. For example, these strategies may include the

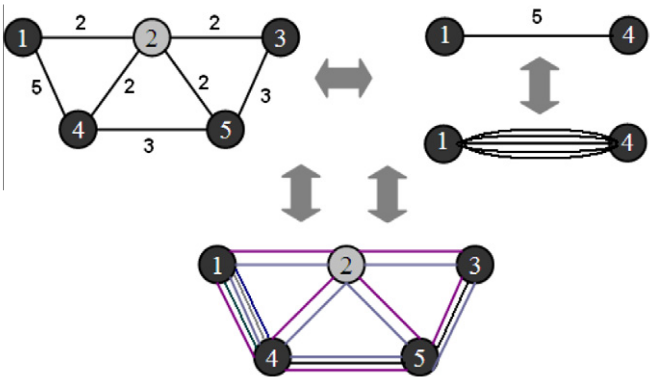


Fig. 1. Generating the working multi-graph.

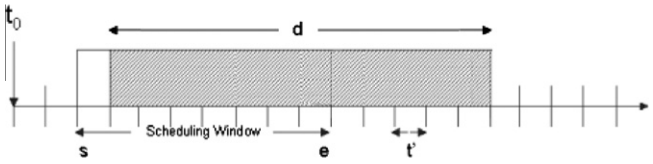


Fig. 2. Connection set-up time range: scheduling window.

first-shortest-path route, the second-shortest-path route, the third-shortest-path route, etc. Hence, the best strategy/path for a connection request can only be chosen from this set. Any feasible path from the source node to the destination node can be a candidate as the actual strategy for satisfying a connection request. The choice of a strategy instead of another one depends on the overall satisfaction of all the players/request and hence on the reachability of an acceptable pure Nash equilibrium status. A routing strategy preferentially choosing the paths with minimal number of hops tends to minimize resources utilization in terms of nodes involved in routing data traffic for the same source–destination pair. On the other hand, paths with low number of hops are expected to be more robust to failures and easy to control/monitor.

We define *social cost* of our problem as the total number of edges needed for routing a given set of requests. Minimizing this quantity is particularly important in cases where fibers are hired or sold *as a whole*. It is straightforward to verify that the social cost of a strategy profile coincides with the maximum player loss of utility in that profile. To quantify the loss in network performance caused by selfish behavior, we investigate the following question: what is the worst-case ratio between the social cost of an uncoordinated outcome and the social cost of the best-coordinated outcome? Hence, the *price of anarchy* of such a game is given by the worst-case number of edges used in a Nash equilibrium (social cost) divided by the optimum achievable social cost, that is, the minimum number of edges that can be used.

4. The two-stage algorithm

In this section we detail our RWA schema, based on a two-stage approach natively conceived to work on large and complex optical transport networks where little or no coordination can be assumed among the participating entities (a common case in presence of multiple independent administrative domains/autonomous systems), properly conceived to cope with the known drawbacks of the state of the art routing algorithms (lack of *global* optimization objectives). Our main goal is to minimize the total blocking probability by optimizing wavelength usage together with the cost and length of designed paths, while keeping the network resource usage fairly balanced, trying to leave on each link sufficient bandwidth to satisfy further requests as much as possible.

While ideally operating in a non-cooperative fashion, all the entities involved in the proposed RWA framework need to be synchronized in some way to share (and manage) a common view of the network topology as long as link resources usage and status. This implies that every node has to run a distributed control-plane providing the necessary link state routing and signaling protocols [29]. An OSPF or ISIS-like protocol can be used to distribute wavelength/label usage and cost information for each link at the optical layer and bandwidth occupation at the IP one. In the case of OSPF, for example, the opaque LSA facility, augmented with new TLVs can support the additional control information to be exchanged among nodes, such as candidate strategies/paths together with marginal costs/allocation-dependent taxes. An extended signaling/reservation protocol, such as RSVP-TE or CR-LDP within the GMPLS framework, can be used to handle all the resource reservation and allocation operations required during the network activity. Also, a common time synchronization is necessary between the network nodes accepting and routing the incoming end-to-end connection requests. Accordingly, a simplified slotted model has been chosen where in each time slot we can distinguish two distinct stages: the reservation and the allocation phase. The reservation phase will start at the beginning of the time slot, with the network state being the result of the allocation phase that happened at the end of the previous time slot. During the reservation phase, each pending connection request within its scheduling window will act as a player. Each player will selfishly choose its own strategy, based on its knowledge of the network state, by looking for the lowest-cost feasible path with a common wavelength. If such a path cannot be found, the connection request will be deferred to the next time slot, if that is still within its scheduling window, otherwise the connection request cannot be honored in the required time range and hence will be discarded. As a consequence of routing connections according to the chosen strategies, players will experience an additional latency (in the game-theoretic sense) caused by the occupation of the available wavelengths on each physical connection between adjacent nodes. This phenomenon can be handled by introducing a marginal cost model properly weighting the proposed strategies by keeping into account the impact of all the proposals and, hence, considering all the available strategy profiles. The principle of marginal cost pricing asserts that on each edge, every player whose strategy is described by a route crossing it, should pay a tax proportional to the additional latency its presence causes for the other players on such edge.

An assignment of edges to paths motivated only by selfish considerations (its associated *Nash equilibria*) does not minimize the total latency; put differently, the result of local optimization by many selfish network users with conflicting interests does not possess any type of global optimality; that is, this lack of regulation carries the cost of decreased network performance. Hence the outcome of selfish behavior can only be improved upon with some form of coordination. The inefficiency of selfish routing (and, more generally, of Nash equilibria) motivates strategies for *coping with selfishness*, that is, introducing methods for ensuring that non-cooperative behavior results in a socially desirable outcome. Accordingly, we have to consider that whenever each player tries to minimize its private cost, expressed in terms of its individual latency, we need a common decision point where each strategy (path) has to be communicated to all other nodes, letting them to build the strategy profile vectors $(s_1,s_2,\ldots,s_{|R|})$ required to construct their final strategies within the congestion game. This information must be made available to all the participating nodes through the aforementioned link state advertisement/update mechanism available at the control plane layer. For each proposed path, the edge costs need to be updated, by computing their associated marginal cost, to account for the candidate reservations that have been proposed within the strategy profiles made available to all the nodes. In simple words, the performance degradation due to the selfish and non-cooperative behavior of the independent players can be mitigated (or even eliminated in the best conditions) by introducing an appropriate set of marginal costs proportionally taxing each connection resource according to the global demand (end hence the degree of conflict on each resource) of all the independent strategies. These marginal costs implicitly charge each network connection/player for the congestion effects caused by its presence. A player, whose ingress node receives a status update, re-computes the next element in its strategy set and uses the strategy/path information obtained by the other players to build an updated strategy profile. The player checks if the costs along the path constituting its original strategy have been updated. If they have not, the player does not change its previous strategy. Otherwise, the player re-adapts the cost to account for the choice it had previously made, by decreasing the costs along its preferred path as if its own reservation had not been made. Note that this step prevents instability: a player would otherwise keep bouncing between its two best choices, if the difference between their total costs were less than the "tax" induced by the reservation. Then, the player computes again a lowest-cost path according to the updated costs, which can be seen as the next choice in its strategy set. If it finds a more satisfactory solution, it makes a strategy change. It computes the necessary adjustments to the costs along the new path and communicates them, along with the updated costs on the old path, to all the participating nodes. The reservation phase terminates at the completion of each time slot; at this point, all the players have the complete knowledge of the network status and of all the proposed strategies. At the end of the reservation phase, each node obtains a strategy profile representing the best desiderata of each player. The use of a common control plane and a link state routing facility implies that all the nodes share a unique synchronized network view and result in the calculation of the same strategy profile. Such profile may be compatible or not with the network resource limitations. In the first case, the strategy profile is a feasible solution of the allocation problems, in the second case it is partially unfeasible and a different solution must be obtained by shifting to the next time-slot the requests that could not be honored because of resource availability conflicts. In detail, players actually trigger resource allocation by issuing a provisional reservation for each resource on the path. If, at the end of this phase, any of the resources in the path is unavailable because it has been requested by other players, the current player will be deferred to the next time slot. The common signaling facility also ensures that all the nodes actually involved in the reservation and allocation of the links/wavelengths resources required in setting up an end-to-end connection (and hence directly involved with a player in our congestion game) have the same view of the resources seizure status independently from their role in the setup process (i.e. if they are originator, destination or transit nodes). The reason for having two phases is that if connection establishment had been allowed as soon as a successful reservation were made, the connection might have needed rerouting many times, since the process of computing a Nash equilibrium involves possibly many strategy changes. While rerouting a connection can be done in a few milliseconds, rerouting of "live" connection carrying user traffic is undesirable, since it is unavoidable, during rerouting, to cause a service disruption that, although momentary, is perceived by the final users.

Note that the order in which players operate in both the reservation and allocation phases plays a crucial role in the outcome of the overall scheme. Players acting later in the reservation phase have a greater probability to achieve their best (original) strategy, because players preceding them could have been forced to abandon their first choice in case of conflicts, and this will decrease the marginal cost of critical resources. Conversely, players acting first in the allocation phase will have an advantage in securing critical resources. Different network management schemas may choose different ordering criteria, depending on their priority objectives. The ordering may:

- be based on an *a priori* weighting of the connection request (maybe for financial reasons or the strategic importance of clients);
- reflect different priorities calculated from the residual request lifetime within the scheduling window, privileging those connection whose setup time range is about to expire, so that the blocking probability will be reduced;
- be conditioned by the connection duration, possibly favoring long-lasting (thus, more remunerative) connections;
- depend on an absolute measure of the impact on network resources, such as the length of the path requested, so that the social cost will tend to be reduced.

4.1. The resource cost and marginal function

To define a reasonable cost function we first have to evidence the required properties and dynamics characterizing such a function. It is intuitive that a *good* cost function should rank each edge proportionally to both the residual and the maximum number of wavelength available on the same pair of nodes. However, the two factors do not need to contribute equally. We have considered the use, for each edge, of the relative load, i.e. the ratio of the number of used wavelength over the total number of available parallel wavelengths. In addition, some provision must be made to appropriately penalize long paths over shorter ones, and to avoid that the cost of an empty link would be zero. Hence, we introduced an additive fixed nonzero cost to each edge. The resulting cost function is therefore the linear function:

$$c_e(x) = a\frac{x}{w_e} + b, \tag{8}$$

where x is the number of used wavelengths on link e, w_e is the total number of wavelength on all edges sharing the same pair of nodes in the multi-graph with edge e, and a and b are adjustable constants ($a > 0,\ b \geqslant 0$), whose value will be tuned by empirical considerations. The ratio between a and b will be determined by the number of hops that an alternative path must have in order to be considered roughly equivalent to the seizing of a single-hop congested link. If, for the sake of simplicity, b is taken to be 1, reasonable values for the number m of hops representing the length of "equivalent" alternative paths yield an estimate of a being near m when the load reaches about 75–90% of the total saturation.

Without introducing any other additional taxation criterion across the edges/resources composing a path, the congestion game players experience only their own traffic delay as their cost. By introducing edge taxation, players are also charged for the right to use edges across a path. This technique has been studied by the traffic community for a long time (e.g. [32] and the references therein), especially in the context of marginal costs [33].

Each selfish player i when using a path p_i will experience a total cost $\Gamma(p_i)$ obtained by combining its initial cost $\gamma(p_i)$ with the influence of the marginal costs $\mu(p_i)$:

$$\Gamma(p_i) = \gamma(p_i) + \tau_i \cdot \mu(p_i), \tag{9}$$

where $\gamma(p_i) = \sum_{e \in p_i} c_e(x_e)$, is the sum of the individual edge costs along the strategy/path p_i, being x_e the occupation of resource e at the beginning of the current time slot. On the other hand, the cumulative marginal cost function $\mu(p_i) = \sum_{e \in p_i} c_e^*(x_e)$ is the sum of the marginal cost taxes $c_e^*(x)$ along the edges of the path p_i. The factor $\tau_i > 0$, denotes the sensitivity of player i to the taxes. In the homogeneous case, all the players can have the same sensitivity to the taxation (i.e. $\tau_i = 1$, for all i), while in the heterogeneous case τ_i can take different positive values for diverse players. Through edge taxation, we aim to force all equilibria on the network to be reached by combining strategy profiles that minimize the social cost. In our approach, we can see the additional marginal cost taxes assigned to every edge as part of the edge latency function itself. Here, instead of taxation, we can speak about artificial delays introduced possibly at the entrance of each edge, in order to minimize the total "congestion" probability due to multiple players that need to traverse the same adjacency between two nodes, and hence the probability to be blocked at the ingress of the edges themselves. Accordingly, we assume that each player's strategy is further charged according to the maximum number of paths that share an edge with it and use the same wavelength. Applying Eq. (7), we can derive the marginal cost function as:

$$c_e^*(x) = 2a\frac{x}{w_e} + b. \tag{10}$$

Marginal cost taxes increase in general the cost for each player, as shown in [34]. The natural question that arises is whether taxes are an efficient mechanism for achieving the desired result. In other words, if the additional "disutility" caused through taxation proportionate to the desired goal, i.e. a routing assignment that minimizes the total latency. Our marginal cost taxes have been conceived as an implicit coordination mechanism obtained through a cost function properly chosen from a family of possible ones, according to a "coordination model" in the sense defined as in [35]. In particular, results presented in [27] suggest that for strictly increasing and differentiable linear latency functions, imposing properly chosen taxes on a selfish routing game not only yields to a game with better coordination ratio, but also that the added disutility for the players is bounded with respect to the original system optimum. That is, with a small decrease in network efficiency, we achieve, at equilibrium, a strategy profile that minimizes the total latency. From Eqs. (8)–(10), we can see that the total cost for an individual resource *e* still has a linear form in the occupation *x*. Hence, according to the results presented in [3], at least one pure Nash equilibrium exists and it can be computed in pseudo-polynomial time.

4.2. Determining a Nash equilibrium

The distributed algorithm starts on the network nodes with an initial strategy profile $S = (s_1, s_2, \ldots, s_{|R|})$ built on the selfishly chosen minimum cost paths for each request/player r_i in its valid time range. More precisely, on each time slot every node *n* selfishly calculates the strategies s_i for all the requests/players r_i in its locally originated requests set $R_n \subset R$, advertises them on the network and simultaneously learns, from the received advertisements (procedure *Advertise_and_Receive_Strategies*, line 6 in Fig. 3), the strategies proposed by the other nodes so that on each iteration it is able to construct a complete strategy profile *S* containing the strategies associated to all the valid players on the entire network. Then, after a recalculation of the total latencies associated to each path within *S*, performed by adding the marginal costs introduced by the proposed allocations of other players, it iteratively allows each unsatisfied player to recalculate another path, possibly reducing the associated cost. The algorithm iteratively strives to transform a non-equilibrium configuration into a pure Nash equilibrium, performing a sequence of greedy selfish steps, where each player switches to the path that minimizes latency, given the current strategy profile. Each greedy selfish step consists in a player on a node re-computing its minimum-cost path with respect to

```
procedure RWA($R_n$, $S$)
Input:  local connection request set $R_n$
        global strategy profile $S$
1.  $S \leftarrow \emptyset$;
2.  do
3.      for each player $i$ in $R_n$ do
4.          Selfish_Iteration($S$, $C$, $i$) // build local strategy set
5.      endfor
6.      Advertise_and_Receive_Strategies ($S$, $C$) // build global strategy set
7.  while $\exists\ i : \Gamma(s_i^*) \leq \Gamma(s_i)$ OR $\neg$maxIter // until an equilibrium is found or the maximum number
    of iterations has been reached
8.  for each player $i$ in $R_n$ do  // start the resources allocation
9.      if Allocate ($s_i$) is successful then
10.         $R \leftarrow R \setminus \{i\}$ // remove it from the global request set R
11.     endif
12. endfor
13. end procedure RWA
```

Fig. 3. The per-time slot RWA procedure.

```
procedure SELFISH_ITERATION($S$, $C$, $i$)
Input:  current strategy profile $S$
        current costs $C$
        player ID $i$
Output: new strategy profile $S^*$
        new costs $C^*$
$S^* \leftarrow S \setminus \{s_i\}$
$C^* \leftarrow$ UpdateMarginalCosts($S^*$, $C$)
$s_i^* \leftarrow$ MinCostPath($i$, $S^*$, $C^*$)
if $\Gamma(s_i^*) \leq \Gamma(s_i)$ then // if new strategy improves costs
    $S^* \leftarrow S^* \cup \{s_i^*\}$ // add it to the strategy profile
    $C^* \leftarrow$ UpdateMarginalCosts($S^*$, $C^*$)
    if $S^* \neq S$ then
        Advertise($S^*$, $C^*$)
        $S \leftarrow S^*$
        $C \leftarrow C^*$
    endif
endif
end procedure SELFISH_ITERATION
```

Fig. 4. An iteration of the Nashification algorithm.

the choices selfishly made by the other players and possibly changing its best pure strategy and diminishing its latency (cost). In other words, each node dynamically computes, in a stepwise fashion, its local strategy set by indirectly keeping into account (thanks to the marginal costs mechanism) the selfish choices of the players on the other nodes. The process terminates when an equilibrium is reached and no one of the participants is interested in changing its strategies or when a maximum number of iterations (*maxIter*) is reached. However, even without the *maxIter* performance constraint, the linearity of cost functions guarantees [3] the existence of a potential function (Section 2) that, in turn, ensures [26] that a pure Nash equilibrium always exists, so the distributed algorithm will terminate after a finite number of steps into a configuration in which no user has incentive to deviate. When an equilibrium strategy profile is available each node can allocate all the paths associated to its own players (line 8–12 in Fig. 3). Allocation of a path/strategy s_i (procedure *Allocate*, line 9 in Fig. 3) is accomplished by using the traditional two-directions forward provisional resource seizure (i.e. GMPLS RSVP-TE PATH request message) and backward reservation (i.e. RESV message) paradigm. If for a specific request/player this last step is not successful, the associated connection request is shifted (to be served) at the next time slot.

The procedure in Fig. 4 details the selfish iteration step described above. The procedure is run on each node for all the requests/players originating in that node, and it is triggered by the reception of an updated strategy profile. This is, in turn, the result of an invocation of the *Advertise*( ) procedure which, as previously asserted, is implemented through the link-state update mechanism of the control plane layer. After all nodes broadcast their updates with their new strategy, each node knows the entire strategy profile S^*. Procedure *UpdateMarginalCosts*( ) computes the new costs C^* associated with a strategy profile S^*, whereas *MinCostPath*( ) finds a minimum-cost path, by using the traditional Dijkstra algorithm, based on these costs. Finally, in line 7, an advertisement is produced if there is a strategy change.

5. Performance considerations

Let's now examine the computational complexity of the above framework for a network (graph) $G = (V,E)$ with $|V| = N$ nodes and $|E| = M$ edges. The *Selfish_Iteration*( ) procedure shown in Fig. 4 is built up by a number of simple sub-procedures whose complexity is analyzed in the following. Line 1 removes the path from the current strategy profile, while line 2 updates the marginal costs: these operations have both a cost of $O(N)$. The Dijkstra's shortest path algorithm is thus calculated in line 3 requiring $O(M + N \log N)$ [36]. Line 4 requires the computation of the path total costs $\Gamma(s_i^*)$ and $\Gamma(s_i)$ and all the links in the path are to be taken into account during this operation in which both delays and marginal costs are considered. In the worst case, the number of links of a simple path (i.e. a path without loops) in a network with N nodes is $N - 1$; since each of the operation involved in the cost calculation has a constant cost $O(1)$, the computation of the Γ factors costs at most $O(2N) = O(N)$. The strategy is added to the strategy profile at line 5 and the marginal costs of the link associated with the newly proposed path are updated at line 6, both operations requiring $O(N)$. If the strategy profile has changed (line 7), a link state update message is sent to other nodes in order to reflect the change (line 8) and the new values of S^* and C^* are stored in S and C respectively in lines 9 and 10; each of these operations has constant costs $O(1)$. For each node, the overall cost of the *Selfish_Iteration ()* procedure is thus given by $O(N) + O(M + N \log N) + O(N) = O(M + N\log N)$.

The *RWA*( ) procedure shown in Fig. 3 is executed by each network node which, after initializing the global strategy profile S at line 1 with a cost of $O(1)$, repeats (lines 3–5), for each player i in the local set R_n, $|R_n| = k$, the *Selfish_Iteration*( ) procedure to construct the local strategy set; at line 6, the node advertises its local strategies to the other nodes and receives their strategies to construct the global strategy set. These operations (lines 2–7) are repeated until there are no improvements or, in the worst case, the maximum number of iterations has been reached. Thus lines 2–7 have a complexity of $maxIter \cdot k \cdot O(M + N\log N)$. Lines 8–12 repeat the allocation phase for each of the k players and, possibly, remove them from the request set R, with a constant cost $k \cdot O(1)$. Therefore, the total cost of the *RWA*( ) procedure is given by $O(1) + maxIter \cdot k \cdot O(M + N\log N) + k \cdot O(1) = maxIter \cdot k \cdot O(M + N\log N)$. In the worst case scenario, before the call setup may be admitted, each network node repeats the *RWA*( ) procedure at each t' slot size, during a maximum time interval given by $(e - s)$ before the assigned time slot expires, as specified by the connection setup request. In the worst case the *RWA*( ) will be executed exactly $w = (e - s)/t'$ times, thus the total computational complexity is given by $w \cdot maxIter \cdot k \cdot O(M + N\log N)$ for each single node, which, as the results on the average delay show, is an affordable complexity. Nevertheless, these computations will be also done at different times; more precisely, each one will be done exactly after t' time units, for at most a time window of size w so that the parameter w must be carefully tuned in order to let each node compute its *Selfish_Iteration*( ) before a new attempt may be performed. In the next section, we study the choice of the parameter w along with the different results in terms of performance and stability.

6. Experimental evaluation and results analysis

In order to evaluate the functionality of the proposed selfish routing and wavelength assignment strategy, we conducted an extensive simulation study on several network topologies (modeled as undirected graphs in which each link has a non-negative capacity). In the following paragraphs we report the simulation details together with the most interesting results and observations emerged during the experiments.

6.1. The simulation environment

The evaluation of the proposed routing framework has been conducted in an optical network simulation environment [37] that allows the creation of heterogeneous network topologies along with the specification of simulation parameters and configuration options. Simulations have been performed on an HP® DL380 Dual Processor (Intel® Xeon® 2.5 GHz) server running FreeBSD® 4.11 operating system and Sun® Java® 1.4.2 Runtime Environment. In all the experiments, we used a dynamic traffic model in which connection requests, defined by a Poisson process, arrive with a parametric rate of λ requests/s and the call-holding time is exponentially distributed. The connections are distributed on the available network nodes according to a random-generated or predefined traffic matrix.

6.2. Results analysis

The results presented are taken from many simulation runs on several network topologies with various parameter and bandwidth unit request values, as summarized in Table 1.

As can be seen from the Table 1, 20 simulations per topology were executed and, to obtain more confidence in the results, each run has been repeated 10 times and the average performance metric values have been calculated. We considered several values for the parameters and measured the blocked connections and the convergence times of the illustrated "Nashification" process. The length of used time window w assumes values from the set

Table 1
Simulations performed and parameters used.

Parameters	Geant2/Internet2
Number of connections	Varying from 0 to 10,000 (step 100)
Random generated bandwidths (OC-unit)	{1,3,12,24,48,192} with different distribution probability
a, b, τ_i	Varying canonical values: $a = 1$, $b = 0$, $\tau_i = 1\ \forall i$
d, s, e	Varying according to Poisson process
w, $maxIter$	Varying in the range {2,3,4,5,6,7,8,9}
Number of simulations	20 simulations run per topology; each simulation repeated 10 times
Measurements	Blocked connections and experienced delays

specified in Table 1 in order to evaluate the algorithm when players have different time intervals during which they have to choose their strategy profiles. In our lambda-switched optical framework, the resources occupied by the routed connections are counted as sum of the ratio between the free and the busy bandwidths along the edges. Resources are thus represented as the sum of the bandwidths on all the network edges, while the traffic volume is represented by the quantity of the utilized bandwidth in a certain time. We tried out different static, predefined [38,39], or randomly generated traffic demand matrices on several network topologies, both randomly generated and well-known, such as Geant2 [40] and Internet2 [41] (Figs. 5 and 6) with the bandwidths for the links ranging from OC-1 to OC-768 bandwidth units. When we used the traffic matrices defined in [38,39] the traffic volumes have been scaled proportionally to the reported traffic distributions.

In our tests, each connection request was characterized by a bandwidth demand ranging from OC-1 to OC-192 (i.e. up to 10 Gbps) units, and the s, e and d reservation parameters for each connection request (Fig. 2) are generated according to a Poisson process (exponentially negative distribution). As the network load grows, that is, the number of busy connection resources increases more and more respect to the free/released ones, we continuously monitor the network efficiency expressed by the rejection ratio/blocking factor. During the simulations, the performance of the algorithm was tested against different values of the parameters of Table 1: the scheduling window size w, the weight factors a and b of Eqs. (8) and (10), and the order of the connection requests. The first simulation is to test the performance of the algorithm with varying time window sizes. The average blocking probability as function of the network load measured in Erlangs is shown in Figs. 7 and 8 (canonical values assumed for $a = 1$, $b = 0$). Results show that the blocking ratios grow quite regularly, but with some differences according to the time window w that has to be chosen. A time window $w = \{4,5\}$ achieves better performances in terms of blocking ratios with respect to too high ($w = \{6,7\}$) or too low ($w = \{2,3\}$) time windows that may drive to sub-optimal results. In fact, in both simulations the best performances have been achieved with parameter $w = \{4,5\}$, meaning that lower blocking may be achieved by giving only *some* chances to the nodes to change their strategy profile. The results obtained with a low value of the time windows w show that a sub-optimal network equilibrium is reached, but the too little available steps avoid the system to reach optimal configurations in most cases. Similarly, the results obtained with a high value of w indicate that a sub-optimal, but sustainable, network equilibrium is reached within too many steps from the initial strategy profile and that margin of optimizations are possible by decreasing the windows size. Thus, tuning the parameter it is possible to obtain a balance between the performances and the computational times (steps) needed to efficiently find good performances. In any case we can observe that blocking ratio grows quite slowly when the network load increases and that the highest, unacceptable values are physiologically reached only when the available network resources are almost totally saturated.

Average time delays experienced by connection requests during the simulations are plotted in Figs. 9 and 10. In particular, we analyzed how fast the average setup delays grow with the number of simultaneous players requesting in each time slot new end-to-end connections. In this scenario, the network load grows physiologically with the number of simultaneous requests but the connection

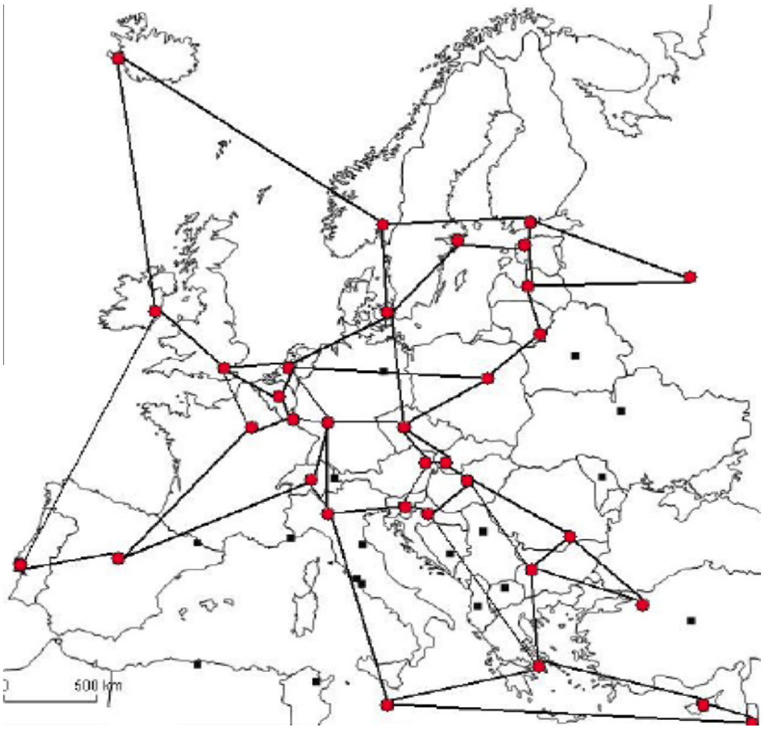


Fig. 5. Geant2.

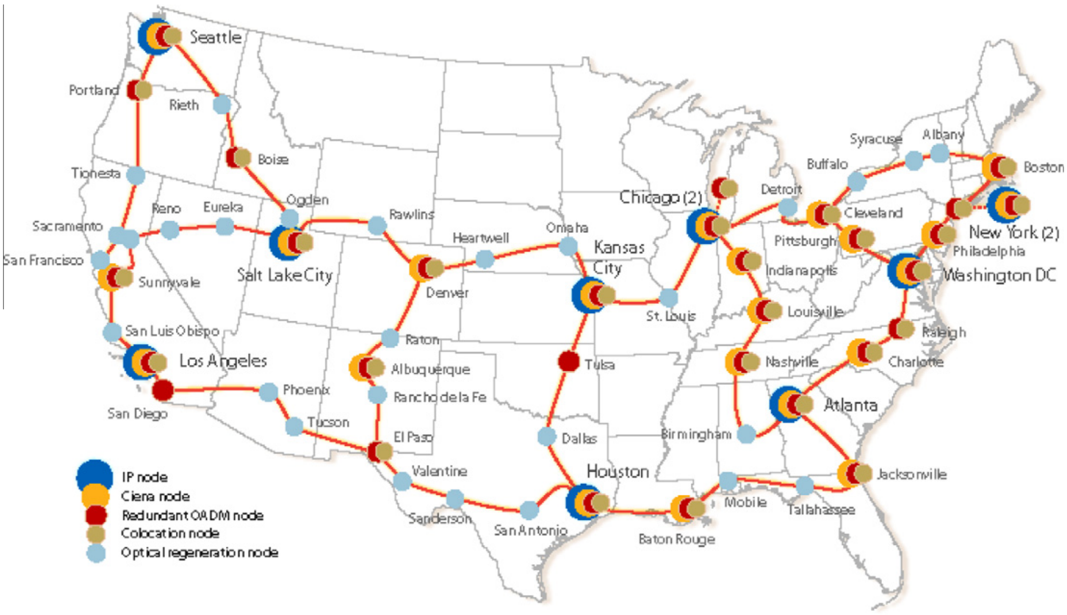


Fig. 6. Internet2.

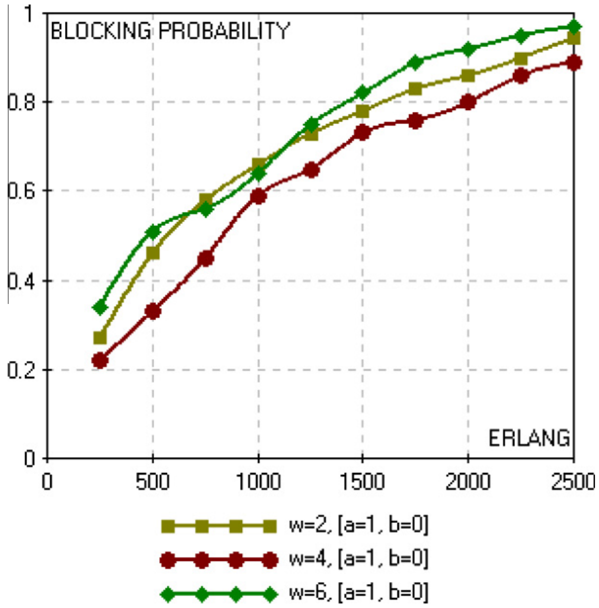


Fig. 7. Geant2 blocking probability, varying window size.

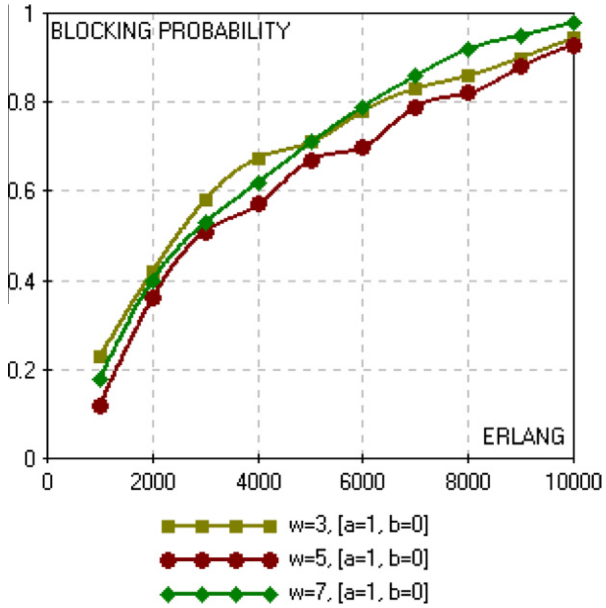


Fig. 8. Internet2 blocking probability, varying window size.

requests and release rates are kept balanced by progressively reducing the connection lifetime. As expected, the greater the time window w is, the higher the delays are. Lower delays have been reported with smaller values of the time window ($w = \{2,3\}$) whilst higher delays have been experienced with greater values of the time window ($w = \{6,7\}$). Results show also that the delays grow faster as w increases and grow at slower rates with low w values. This difference is particularly marked with longer links/lightpaths (higher distances between nodes) as in the case of the Internet2 network (which spreads, in fact, along longer distances). We also observe that, with all the chosen time window values, the delay grows almost linearly with the number of simultaneous players. Also with an high number of simultaneous players/connections, the observed delays always remain under the 1000 ms threshold, which is an affordable time delay for a network [22], thus demonstrating the scalability of the presented approach also in presence of significantly high connection loads.

Now we focus on the behavior of the algorithm when varying the values of the parameters a and b; recall from Eqs. (8) and (10) that the parameter a weights the relative load of links whilst b is a fixed cost for traversing the link. In Figs. 11 and 12 we show the network blocking probability when a is either predominant or negligible with respect to b (medium values of the window size w are assumed). Results for both networks show the same behavior: when choosing values for a greater than b, the cost functions of Eqs. (8) and (10) forces the connections to be spread over the network to avoid paying high costs for traversing loaded links, thus better balancing the load over the available resources, resulting in notable lower blocking probabilities. Anyway, from the Figs. 13 and 14, in which the average experienced time delays are

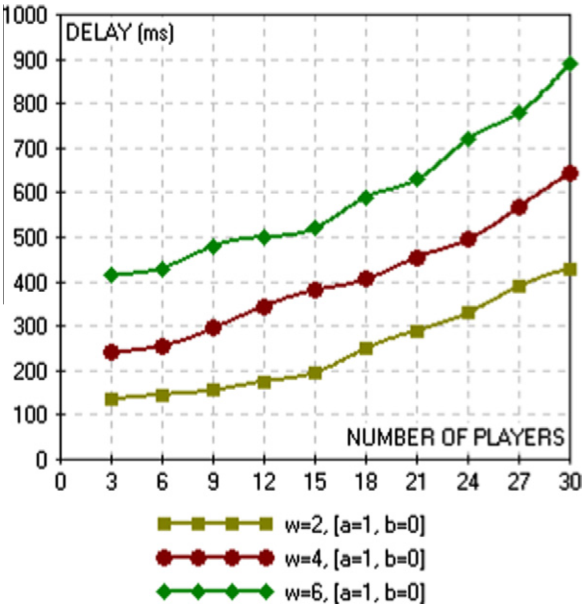


Fig. 9. Geant2 delays, varying window size.

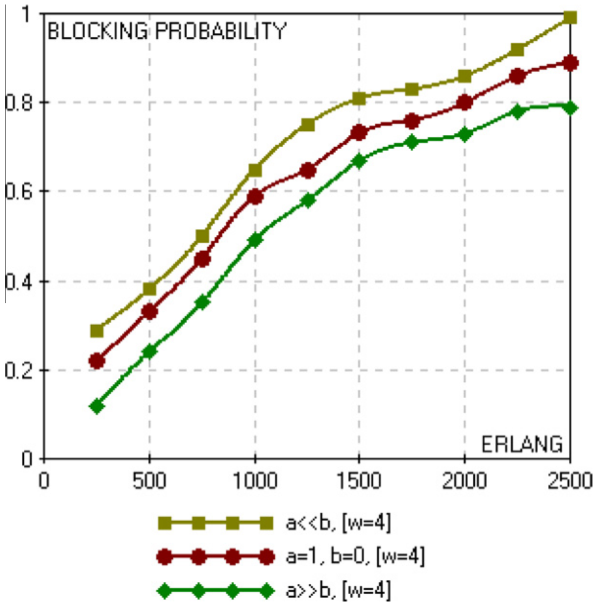


Fig. 11. Geant2 blocking probability, varying parameters *a*, *b*.

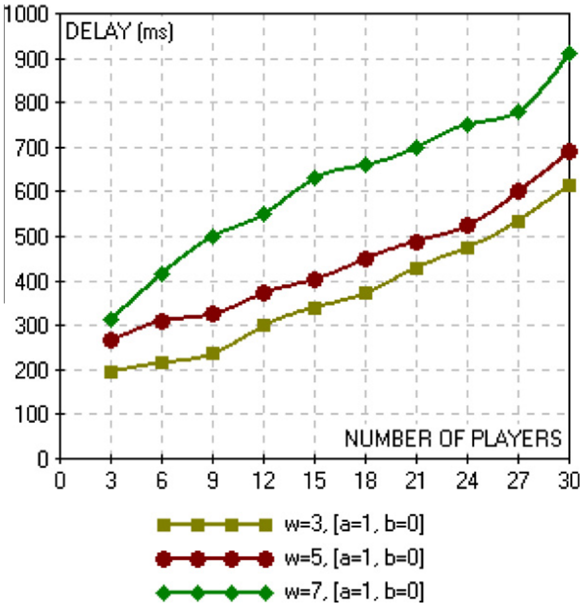


Fig. 10. Internet2 delays, varying window size.

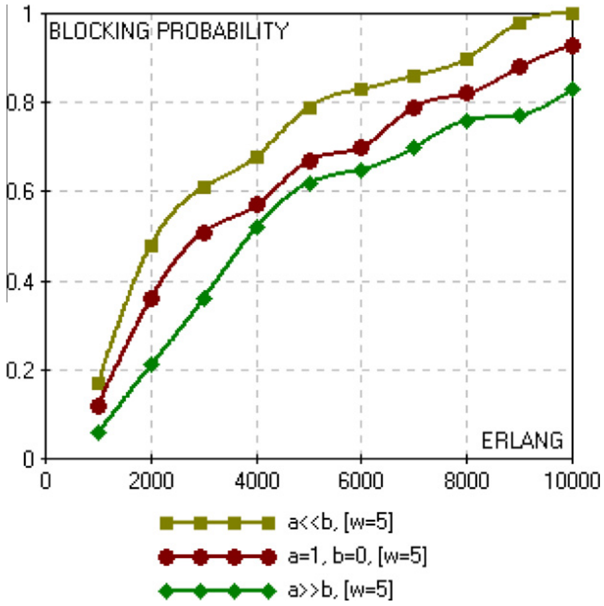


Fig. 12. Internet2 blocking probability, varying parameters *a*, *b*.

shown, we can observe that the better load balancing leads to increased delays, due to the longer paths that will be generally preferred to the shortest ones. Lowest delays have been in fact obtained for values of *a* lower than *b*, since they force shorter paths to be cheaper and, thus, to be chosen more frequently. Anyway, shortest paths mean also greater blocking probability, due to the congestion of critical network links. Therefore, a tradeoff exists between load balancing and delay; if the objective is to maximize the number of served connections, a high value of the ratio *a*/*b* should be preferred, whereas if the objective is to minimize the average delay, low values of *a*/*b* should be chosen.

The performance of the proposed algorithm is compared with three other well-known RWA schemas and the average blocking probabilities are measured and plotted in Figs. 15 and 16. We evaluated our approach against the canonical shortest paths (minimum hop algorithm, MHA [42]), the shortest widest path algorithm (SWP) [43] and the minimum interference routing algorithm (MIRA) [44] transposed into the optical domain [45]. The Dijkstra-based algorithms (MHA and SWP) tend to congest critical links, which results in higher blocking probabilities, more visible in the Internet2 network topology, which is less meshed than Geant2. The proposed algorithm has achieved better performance almost at every load, with MIRA being quite close in terms of rejection ratio. Anyway, even if MIRA performs sometimes better than our algorithm (in some points present at high, medium and low loads), unbalanced network utilization of MIRA and its difficulties on estimating bottlenecks on critical links for cluster nodes make our approach preferable for its more linear behavior achieved in both networks.

Finally, we show the sensitiveness of the algorithm to the order of the connection requests. As we have seen in Section 4, many different ordering criteria are applicable when selecting connection

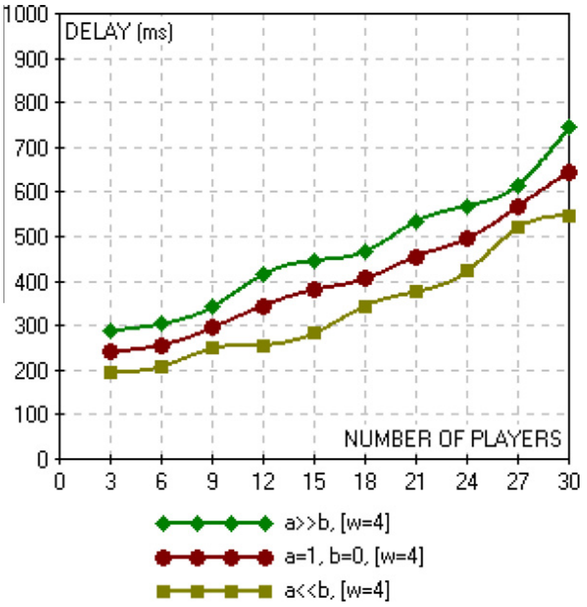


Fig. 13. Geant2 delays, varying parameters *a*, *b*.

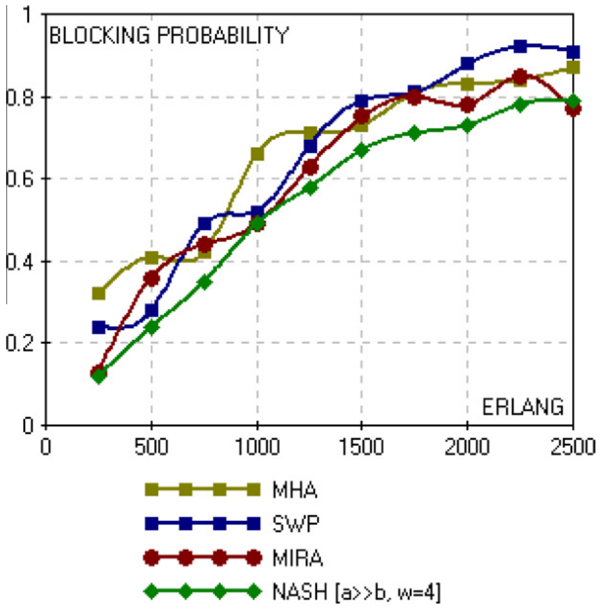


Fig. 15. Geant2 blocking probability comparison with other RWA algorithms.

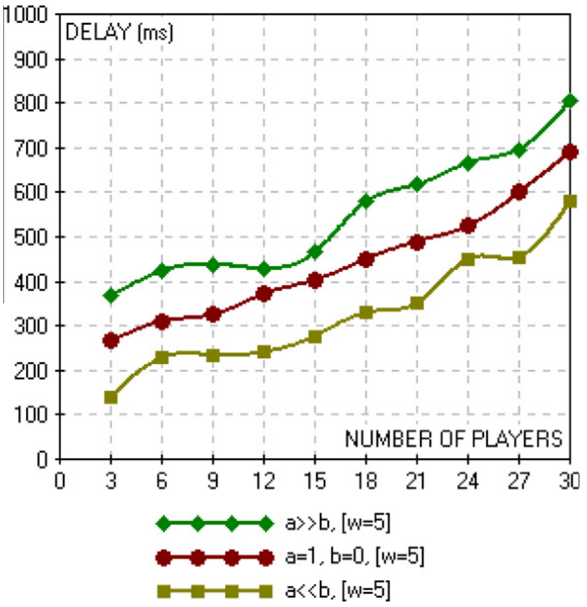


Fig. 14. Internet2 delays, varying parameters *a*, *b*.

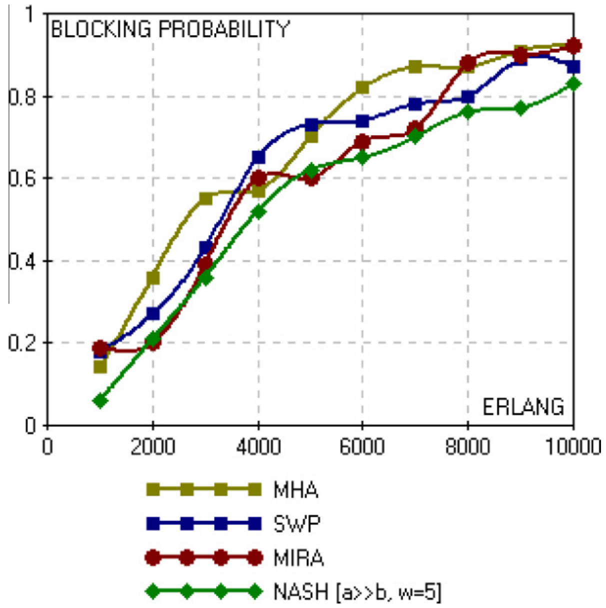


Fig. 16. Internet2 blocking probability comparison with other RWA algorithms.

requests to be served, with the best one depending essentially on the operating scenario and optimization objectives. In order to perform a fair comparison whatever the chosen prioritization is, and to keep the generality of the results, we differentiate between high priority and best effort connection requests. The general framework in which prioritized and best effort connection requests operate is the following. Players (connections requests) choose their strategies by selfishly competing during the reservation phase, so that the Nash equilibrium is preserved, and then, only during the allocation phase, high priority connections are allowed to allocate their resources first. Main results for a medium loaded Geant2 network ($w = 4$, $a >> b$) are shown in Fig. 17 as cumulative distribution function (CDF) of the time in which a given set of connection requests are accepted at or below a given time slot t' in the time window. More than 60% of the prioritized connections are accepted during the first time slot, i.e. their allocation requests have been satisfied and the corresponding resources have been assigned to them, with the acceptance rating growing up to about 95% within the end of the time windows. Connection requests that have not been satisfied at the current time slot move farther to the next time slot, up to the end of the scheduling window. Best effort connections are in general much more delayed toward the end of the time window, with a greater probability of being blocked. Such an high acceptance ratio of the prioritized connections indicate that privileging the connections during the allocation phase is able to differentiate high priority traffic from the best effort one, while keeping intact the properties of the Nashification process.

In conclusions, the results have shown that it is possible to achieve good performances and affordable delays with low/medium values of the time window *w* and by tuning the values for the parameters *a* and *b* in function of the desired optimization criteria (load balancing *vs* delay). The proposed algorithm has often

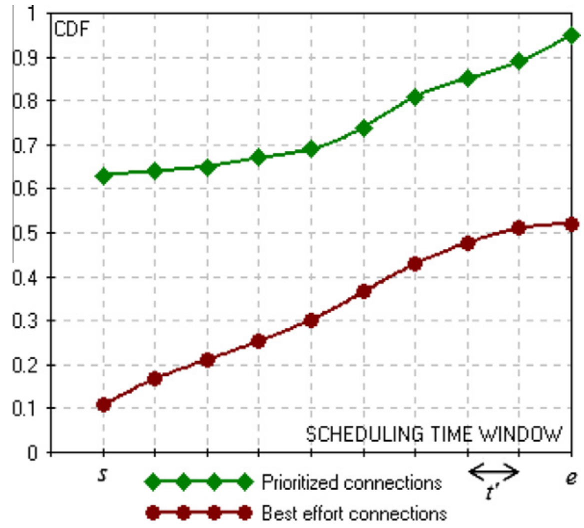


Fig. 17. CDF of the connection requests in the scheduling window.

reached lower blocking ratios with respect to the existing RWA schema with which it has been compared, even in presence of a discrete growing number of simultaneous players and in real network topologies where the geographical distances may be quite long. Finally, we showed that the ordering in which the connections are served in the allocation phase is decisive for privileging high priority requests with respect to best effort traffic, while preserving the presented Nash equilibrium-driven strategy and thus its benefits, especially in presence of highly competitive scenarios with minimum collaboration such as in interconnections of multiple independent autonomous systems networks.

7. Conclusions

In large-scale communication networks, like the Internet, it is usually unfeasible to *globally* manage network traffic. Accordingly, when modeling the traffic behavior in absence of global control, it is typically assumed that network users follow the most rational approach, that is, they behave selfishly to optimize their own individual welfare. Such a consideration motivates our RWA approach based on models from the Game Theory, in which each player is aware of the situation facing all other players and tries to minimize his own cost. We re-formulated the RWA problem in modern connection-oriented all-optical network architectures by considering solution strategies from distributed multi-commodity network congestion games, which are solved by multiple agents operating in a non-cooperative but coordinated manner. The simulation results show that our approach may be particularly attractive for its scalability features and hence useful in large optical networks where many nodes, belonging to different administrative domains, operate selfishly by exchanging only a small amount of information needed for the coordination among them.

Acknowledgments

This work was supported in part by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the Grant FI-201000740.

References

[1] R. Ramaswami, K.N. Sivarajan, Routing and wavelength assignment in all-optical networks, IEEE/ACM Transactions on Networking 3 (5) (1995) 489–500.

[2] I. Chlamtac, A. Ganz, G. Karmi, An approach to high-bandwidth optical wans, IEEE Transactions on Communications 40 (7) (1992) 1171–1182.

[3] D. Fotakis, S. Kontogiannis, P. Spirakis, Selfish unsplittable flows, in: Theoretical Computer Science (TCS) 348, vols. 2–3, pp. 129–366.;
D. Fotakis, S. Kontogiannis, P. Spirakis, Preliminary version in the 31st International Colloquium on Automata, Languages and Programming (ICALP'04), 2004, pp. 593–605.

[4] B. Meliàn, J.L. Verdegay, Fuzzy optimization models for the design of WDM, Networks IEEE Transactions on Fuzzy Systems 16 (2) (2008) 466–478.

[5] T. Matsumoto, T. Takenaka, Overlap degree aware routing in all-optical routing networks, IEICE Transactions on Communications 91 (1) (2008) 212–220.

[6] F. Palmieri, U. Fiore, S. Ricciardi, A minimum cut interference-based integrated RWA algorithm for multiconstrained optical transport networks, Journal of Network and Systems Management 16 (4) (2008) 421–448.

[7] J. Zheng, B. Zhang, H.T. Mouftah, Toward automated provisioning of advance reservation service in next-generation optical Internet, IEEE Communications Magazine 44 (12) (2006) 68–74.

[8] T.D. Wallace, A. Shami, Advanced lightpath reservation in WDM networks, in: Proceedings of IEEE INFOCOM, 2006.

[9] S.S.W. Lee, A. Chen, M.C. Yuang, A Lagrangean relaxation based near-optimal algorithm for advance lightpath reservation in WDM networks, Photonic Network Communications (2009) 1–7, doi:10.1007/s11107-009-0215-9.

[10] E. He, X. Wang, V. Vishwanath, J. Leigh, AR-PIN/PDC: flexible advance reservation of intradomain and interdomain lighpaths, in: Proceedings of Globecom 2006, San Francisco, CA, 2006.

[11] M. Mavronicolas, P. Spirakis, The price of selfish routing, in: Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing (STOC), 2001, pp. 510–519

[12] T. Roughgarden, E. Tardos, How bad is selfish routing?, Journal of ACM 49 (2) (2002) 236–259

[13] M. Hoefer, A. Souza, Tradeoffs and average-case equilibria in selfish routing, ACM Trans. Comput. Theory 2, 1, Article 2 (November 2010), 25 pages. doi:10.1145/1867719.1867721, 2010.

[14] C. Busch, M. Magdon-Ismail, Atomic routing games on maximum congestion, in: S.-W. Cheng, C.K. Poon (Eds.), AAIM 2006, LNCS, vol. 4041, Springer, Heidelberg, 2006, pp. 79–91.

[15] G.F. Georgakopoulos, D.J. Kavvadias, L.G. Sioutis, Nash equilibria in all-optical networks, Discrete Mathematics 309 (13) (2009) 4332–4342.

[16] V. Bilò, M. Flammini, L. Moscardelli, On Nash equilibria in non-cooperative all optical networks, in: V. Diekert, B. Durand (Eds.), STACS 2005, LNCS, vol. 3404, Springer, Heidelberg, 2005, pp. 448–459.

[17] X. Chen, X. Xiaodong Hu, W. Weidong Ma, Reducing the maximum latency of selfish ring routing via pairwise cooperations, Lecture Notes in Computer Science, vol. 6509, Springer, 2010. pp. 31–45.

[18] M. Hoefer, V.S. Mirrokni, H. Roglin, S.-H. Teng, Competitive routing over time, Theoretical Computer Science, in press, corrected proof, Available online 13 June 2011, doi:10.1016/j.tcs.2011.05.055.

[19] A. Fanelli, M. Flammini, G. Melideo, L. Moscardelli, A. Navarra, Game theoretical issues in optical networks, in: Proceedings of the 8th International Conference on Transparent Optical Networks (ICTON), IEE, Nottingham, United Kingdom, 2006.

[20] I. Milis, A. Pagourtzis, K. Potika, Selfish routing and path coloring in all-optical networks, Lecture Notes in Computer Science, vol. 4852, Springer, 2007. pp. 71–84.

[21] E. Bampas, A. Pagourtzis, G. Pierrakos, K. Potika, On a non-cooperative model for wavelength assignment in multifiber optical networks, in: Proceedings of the 19th International Symposium on Algorithms and Computation, Springer, 2008, pp. 159–170.

[22] G. Mohan, C. Siva Ram Murthy, A time optimal wavelength rerouting algorithm for dynamic traffic in WDM networks, Journal of Lightwave Technology 17 (3) (1999).

[23] R.W. Rosenthal, A class of games possessing pure-strategy Nash equilibria, International Journal of Game Theory 2 (1973) 65–67.

[24] J.F. Nash, Equilibrium points in n-person games, Proceedings of the National Academy of Sciences of the United States of America 36 (1) (1950) 48–49.

[25] E. Koutsoupias, C.H. Papadimitriou, Worst-case equilibria, in: C. Meinel, S. Tison (Eds.), STACS 1999, LNCS, vol. 1563, Springer, Heidelberg, 1999, pp. 404–413.

[26] D. Monderer, L. Shapley, Potential games, Games and Economic Behavior 14 (1996) 124–143.

[27] G. Karakostas, S.G. Kolliopoulos, The efficiency of optimal taxes, in: Combinatorial and Algorithmic Aspects of Networking, First Workshop on Combinatorial and Algorithmic Aspects of Networking, CAAN, 2004

[28] T. Roughgarden, E. Tardos, How bad is selfish routing?, Journal of the ACM 49 (2) (2002) 236–259

[29] J. Comellas, R. Martinez, J. Prat, V. Sales, G. Junyent, Integrated IP/WDM routing in GMPLS-based optical networks, IEEE Network 2003 (2003) 22–27.

[30] J. Kuri, N. Puech, M. Gagnaire, E. Dotaro, R. Douville, Routing and wavelength assignment of scheduled lightpath demands, IEEE JSAC 21 (8) (2003) 1231–1240.

[31] A. Jaekel, Lightpath scheduling and allocation under a flexible scheduled traffic model, in: Proceedings of IEEE GLOBECOM 2006, 2006.

[32] R. Cole, Y. Dodis, T. Roughgarden, Pricing network edges for heterogeneous selfish users, in: Proceedings of the 35th Annual ACM Symposium on Theory of Computing, 2003, pp. 521–530

[33] S. Dafermos, F.T. Sparrow, The traffic assignment problem for a general network, Journal of Research of the National Bureau of Standards, Series B 73B (1969) 91–118.

[34] Cole, Y. Dodis, T. Roughgarden, How much can taxes help selfish routing?, in: Proceedings of the 4th ACM Conference on Electronic Commerce, 2003, pp. 98–107.

[35] G. Christodoulou, E. Koutsoupias, A. Nanavati, Coordination mechanisms, in: Proceedings of 31st ICALP, 2004.

[36] H.R. Lewis, L. Denenberg, Data Structures and Their Algorithms, HarperCollins, New York, 1991.

[37] F. Palmieri, U. Fiore, S. Ricciardi, SimulNet: a wavelength-routed optical network simulation framework, in: Proceedings of ISCC 2009, 2009, pp. 281–286.

[38] R. Ramaswami, K.N. Sivarajan, Design of logical topologies for wavelength-routed optical networks, IEEE Journal on Selected Areas in Communications 14 (1996) 840–851.

[39] S. Uhlig, B. Quoitin, I. Lepropre, S. Balon, Providing public intradomain traffic matrices to the research community, ACM SIGCOMM Computer Communication Review 36 (1) (2006) 83–86.

[40] Geant2, <http://www.geant2.net/server/show/nav.00d007009> (accessed 12.09).

[41] Internet2, <http://www.internet2.edu/network/> (accessed 12.09).

[42] D. Awduche, L. Berger, D. Gain, T. Li, V. Srinivasan, G. Swallow, Extensions to RSVP for LSP tunnels, Internet Draft <draft-ietf-mplsrsvp-lsp-tunnel-04.txt>, 1999.

[43] R. Guerin, D. Williams, A. Orda, QoS routing mechanisms and OSPF extensions, in: Proc. IEEE GLOBECOM, 1997.

[44] K. Kar, M. Kodialam, T.V. Lakshman, Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications, IEEE Journal on Selected Areas in Communications 18 (12) (2000) 921–940.

[45] K. Kar, M. Kodialam, T.V. Lakshman, Integrated dynamic IP and wavelength routing in IP over WDM networks, in: Proc. IEEE INFOCOM, 2001.

Journal of High Speed Networks 17 (2008–2010) 185–205
DOI 10.3233/JHS-2011-0340
IOS Press

Constrained minimum lightpath affinity routing in multi-layer optical transport networks

Francesco Palmieri [a,*], Ugo Fiore [b] and Sergio Ricciardi [c]

[a] *Dipartimento di Ingegneria dell'Informazione, Seconda Università degli Studi di Napoli, Aversa (CE), Italy*
[b] *Università degli Studi di Napoli Federico II, Complesso Universitario Monte S. Angelo, Napoli, Italy*
[c] *Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona, Catalunya, Spain*

Abstract. In this paper, a new traffic engineering-capable routing and wavelength assignment scheme is proposed to efficiently handle LSP and lightpath setup requests with different QoS requirements on modern multi-layer (fully optical core and time-division multiplexed edge) transport networks. The objectives of the proposed algorithm are to minimize the rejection probability by maximizing the network load balancing and efficiently handling the grooming of several LSPs on the same lightpath while respecting the constraints of the optical node architecture and considering both traffic engineering and QoS requirements. The proposed solution consists of a two-stage RWA algorithm: each time a new request arrives, an on-line dynamic grooming scheme finds a set of feasible lightpaths which fulfill the QoS and traffic engineering requirements; then, the best feasible lightpath is selected, aiming to keep the network unbalancing and blocking probability as low as possible in the medium and long term, according to a novel global path affinity minimization concept. Extensive simulation experiments have been performed in which our on-line dynamic RWA algorithm demonstrated significant performances. Thanks to its optimal network resource usage and to its reasonable computational space and time complexity, the algorithm can be very attractive for the next-generation optical wavelength-switched networks.

Keywords: Optical multi-layer networks, WDM, lightpath, RWA

1. Introduction

The modern IP-based transport infrastructures can be physically seen as a very complex mesh of variously interconnected optical wavelength-division multiplexed (WDM) and traditional time-division multiplexed (TDM) sub-networks, where each sub-network consists of several heterogeneous routing and switching devices built by the same or different vendor and, ideally, operating according to the same control-plane protocols and policies. With these very different types of devices all the forwarding decisions will be based on a combination of packet/cell, time slot, wavelengths or physical ports depending on the position (edge or core) and role (intermediate or termination/gateway node) of the switching devices in the network layout. In particular, wavelength-switched optical sub-networks are typically used as backbone infrastructures to interconnect a large number of different IP as well as other traditional transport networks such as SDH, ATM and frame relay. In the network core, two adjacent nodes are connected by one or multiple fibers each carrying multiple wavelengths/channels. Each node consists of a dynamically configurable optical switch which supports fiber switching and wavelength switching, that is, the data on a specified input fiber and wavelength can be switched to a specified output fiber on the same wavelength or another wavelength in presence of conversion-capable devices. On the other side, in the network edge we can find

[*] Corresponding author: Francesco Palmieri, Dipartimento di Ingegneria dell'Informazione, Seconda Università degli Studi di Napoli, Via Roma 29, 81031, Aversa (CE), Italy. Fax: +39 081 676 628; E-mail: francesco.palmieri@unina.it.

conventional label-switching routers operating in the TDM environment, usually equipped with optical interfaces for their connection to the transport backbone. Such network elements behave as mediation devices between the WDM-switched and the TDM-routed domains. While several consolidated control-plane technologies exist and are widely deployed to implement automated routing and traffic engineering in the TDM environment, all provisioning and engineering in the optical core often still requires manual planning and configuration, resulting in setup times of hours or even days and a marked reluctance amongst network managers to de-provision resources, to avoid unpredictable impacts with other services. In the last years, several control protocols have been deployed to dynamically provide routing, traffic engineering and provisioning-management assistance in optical networks, but they are essentially proprietary and technology-dependent and thus they greatly suffer from interoperability problems. Consequently, a new fully automated control-plane framework, supporting evolutionary routing and traffic engineering features, is needed to manage dynamically reconfigurable network resources, and this is even more true in multi-vendor environments under distributed control. The fundamental service offered by such a control plane is dynamic end-to-end connection provisioning in the multi-layer hybrid network environment. The operators need only to specify the connection parameters and send them to the ingress node. The network control plane will then determine the optical paths across the network according to the parameters provided and will signal the corresponding nodes to establish the connection. The whole procedure has to be completed within seconds instead of hours. In order to establish a connection that will be used to transfer data between a source–destination node pair, a *lightpath* needs to be established in the all-optical core by allocating the same wavelength throughout the route of the transmitted data (in presence of the *continuity constraint*) or selecting the proper wavelength conversion-capable nodes and wavelengths to be used across the path (if the wavelength continuity constraint is not fully enforced). In fact, some wavelength conversion-capable nodes may be placed in the network to reduce the overall blocking probability in case of wavelength resource exhaustion on some nodes. Lightpaths may span more than one fiber link and must remain entirely optical from end to end. However, according to the mandatory *clash constraint* two lightpath traversing the same fiber link cannot share on that link the same wavelength, that is, each wavelength on a given fiber is not a sharable resource between lightpaths. In general, if there are multiple feasible wavelengths (often also called lambdas) between a source node and a destination node, then a *Wavelength Assignment* algorithm is required to select wavelengths for a given lightpath. The wavelength selection may be performed either after an optical route has been determined (in the so-called decoupled approach), or in parallel while searching a route. In the latter case we refer to the coupled approach, in which the entire job is accomplished by a single *Routing* and *Wavelength Assignment* (RWA) algorithm. When lightpaths are established and taken down dynamically, routing and wavelength assignment decisions must be made as connection requests arrive to the network. It is possible that, for a given connection request, there may be insufficient network resources to set up a lightpath, in which case the connection request will be blocked. The connection may also be blocked if there is no common wavelength available on all of the links and there are no converter nodes along the chosen route. Thus, the objective in the dynamic situation is to choose a route and a wavelength, which maximizes the probability of setting up a (dynamically) given connection, while at the same time attempting to minimize the blocking probability for future connection requests. When a new connection request arrives, the control plane must determine if it can be routed on the current set of lightpaths available on the optical transport network, by time-division (or statistically) multiplexing it together with other already established connections, or if a new lightpath is needed in order to accommodate the request. Different decisions reflect different objectives in term of network resource utilization, and are referred to as grooming and traffic engineering policies [14]. The general criteria for selecting a new traffic engineered path adopted by these algorithms is to consider those paths traversing lightpaths which will result, as possible, in the least impact on the rejection (blocking) of future traffic requests. However, such criteria, with multiple independent bandwidth and QoS requirements, lead to NP-complete problems [5,9]. As a consequence, various heuristic algorithms have been proposed to solve the above problem in computationally feasible time. Although, a heuristic-based approach may not yield the optimal solution, a carefully designed heuristic may produce a near-optimal solution and have low computational complexity. With these premises, we are interested in the problem of dynamically setting up QoS guaranteed paths, needed for dynamic *label switched path* (LSP) and lightpath setup, in an optical network, where the setup requests arrive one by one and future demands are unknown.

The only dynamic information available to the lightpath routing algorithms is the link characteristics and residual capacities, which can be obtained from routing protocol extensions. Accordingly, we developed and studied a new heuristic-based algorithm for optimal LSP and lightpath routing and establishment in presence of explicit QoS constraint, such as minimum required bandwidth. Like Minimum Hop Algorithm (MHA) [3], Shortest Widest Path (SWP) and Widest Shortest Path (WSP) [12], Maximum Open Capacity Routing Algorithm MOCA (the popular Minimum Interference Routing Algorithm MIRA extended to work in the optical domain) [14,15] and other well-known algorithms, our approach operates online and is independent from specific traffic profiles. Nevertheless, it works in a substantially different way, since it has been natively conceived for all-optical wavelength-switched networks with wavelength conversion capability, and is based on a new *path affinity* concept, measuring the degree of resource sharing between the network paths. The algorithm proposed, that we call Minimum Affinity algorithm, aims to minimize the overall affinity between all the existing lightpaths/LSPs, thus ensuring the maximum balance in network resource usage and keeping the blocking probability as low as possible in the medium and long term. It consists of a two-stage process: each time a new request arrives, an on-line dynamic grooming scheme finds a set of feasible routes built on already existing or new lightpaths which fulfill the QoS requirements, then the optimal one between them is selected according to the above global path affinity minimization concept. There are many scenarios in which almost all the existing RWA algorithms fail to satisfactorily route lightpath setup requests and consequently we point out their most significant limitations and try to address them with our traffic engineering and load balancing policies. Our proposed algorithm is able to overcome some of the most common drawbacks by taking into account the overall unbalancing and blocking effects of routing an explicit new path request. It is also based on a totally flexible network model, supporting heterogeneous WDM equipment, with sparse wavelength conversion capability, in which the number and type of lambdas can vary on each link, and provides a fully dynamic path selection scheme in which the grooming policy is not predetermined but may vary along with the evolution of the network traffic. Finally, the algorithm is aware of all the complexity, expensiveness, performance and resource-limitation constrains implicit in the various flavors of optical switching devices, so that it can explicitly and proportionally penalize all the paths that require wavelength conversion. This may be particularly useful if we have a cost associated with wavelength conversion (for example, if wavelength conversion would occur in the electronic domain, the cost may be very high).

Extensive simulations have been carried out to evaluate the performance of the proposed algorithm in terms of total available bandwidth between ingress and egress routers after routing a new request as well as lightpath request rejection probability. Compared to other popular routing algorithms, e.g., MHA, SWP and MOCA, our new algorithm performs considerably well under increasing traffic loads. Good performance together with low computational complexity make our algorithm very attractive for the future multi-layer optical circuit and wavelength-switched networks.

2. Related work

The problem of establishing QoS guaranteed paths, as path setup request arrives one-by-one with no advance about future requests, has been studied elsewhere in [4,7,8,13–15,22,24,25,28]. An interesting grooming approach implementable with extensions of OSPF-TE and RSVP-TE protocols is presented in [7]. The above work essentially reports significant performance improvements over fixed routing strategies. In [8] the authors resort to a local search heuristic in order to optimize the weight setting for a given set of demands in OSPF without having to compute the optimal general routing where the flow for each demand is optimally distributed over all paths between source and destination. However, this technique applies only in a static scenario where the demand matrix is known a priori. In [13], an approach to distributed constraint-based path selection for dynamic RWA in WDM networks is described. It essentially focuses on service-specific path quality attributes, such as physical layer impairments, reliability, policy, traffic conditions and above all the presence of electronic regenerators, which improve flexibility but could induce impairments, such as delays and operational costs. The distributed dynamic routing algorithms proposed in [24] basically provide a common framework for different types of QoS using additive and multiplicative metrics to achieve dedicated and shared protection in the case of single link failure and considers traffic demands at the wavelength level but not at sub-wavelength level that is one of our main drivers to an optimal solution. Also the solution proposed in [25] is not perfectly adequate to fulfill all our tasks. In fact, the profile-based routing algorithm strictly *relies* on the availability of traffic profiles between pairs of ingress and egress routers. Since network traffic is inherently bursty and very dynamic, these traffic profiles may not be easily obtainable and reliable. Analogously, [28] proposes an integrated routing and grooming algorithm for IP over WDM networks based on a blocking island hierarchical network model to abstract network resources. Whereas the main idea of the algorithm is to keep the integrity and load balance of related blocking islands as intact as possible, the paradigm suffers of the same drawbacks described above for the profile-based solution. Other more dynamic solutions will be preferable to meet all the formerly stated requirements, such as [22] in which a parametric adaptive RWA heuristic scheme based on K-shortest paths (SPARK) has been presented. Such grooming-capable dynamic algorithm is transparent with respect to the presence of wavelength converters and achieves very low connection rejection ratios with minimal computational complexity, selecting the best route among multiple feasible paths by evaluating a goodness function on each candidate path. We take advantage of the K-shortest paths as in [22] but we apply this strategy to the idea of other interesting solutions such those based on the Minimum Interference Routing concept, like the native MIRA [14] algorithm itself, and its further refinements DORA [4] and MOCA [15]. In more detail, the minimum interference routing approach is based on the concept of reducing the "interference" of routing current LSP request to potential unknown future requests. The basic observation on which it is based is that routing an LSP along a path can reduce the maximum available bandwidth between some other ingress–egress router pairs. This phenomenon is termed as "interference". We consider the same interference concept from a different point of view and on a more global perspective, that is not only on an end-to-end basis between node pairs but in an hop-by-hop way, by considering all the network connectivity resources, taken both alone and together in the overall network resource economy, trying to minimize stepwise the impact of each new allocation within the context of the already reserved resource. That is to say, if paths that reduce a large amount of available bandwidth between single node pairs or sections of the network are avoided, creation of bottlenecks can also be avoided. On the other side, our path affinity-based algorithm aims to achieve a more balanced network usage achieving better performances in terms of request rejection ratio while exhibiting lower computation complexity. We will now describe our algorithm in the context of the state-of-the-art optical transport networks.

3. The need for new control-plane services

Service providers are facing the challenge of designing and managing their networks to support increasing customer interest in fast, reliable and quality-differentiated services. Stimulated by recent progress in optical networking, there has been a growing interest in designing an unified control plane (i.e., routing and signaling) working for both the pure optical and the traditional time-division multiplexed electro-optical layer, based on reusing and leveraging existing control-plane protocols, to overcome all the existing limitations and scalability problems. These evolutionary control-plane technologies are expected to lower price and improve performance of network layer routing and provide greater flexibility in the delivery of (new) routing services, facilitating traffic engineering. It has been extensively proved that the use of traditional IP routing paradigm, which requires to forward packets based only on destination addresses along the shortest paths computed using mostly static and traffic-independent link metrics, leads to a rapid saturation of some network links on the shortest paths between certain ingress–egress router pairs while links on other alternative paths may be lightly loaded. This is due to the fact that existing routing protocols are largely oblivious to network traffic condition and at the same time incapable of pinning down an explicit route. To address the above shortcomings, new routing practices and features need to be introduced, in most cases integrated in the new control-plane paradigms such that one provided by Generalized Multi-Protocol Label Switching (GMPLS), which is an extension of MPLS, widely known as the strongest enabling technology for traffic engineering. At the state of the art, GMPLS is emerging as the candidate control-plane solution for next-generation optical networking [20], since it integrates in a native and natural way the widely known generalized label swapping/forwarding paradigm with the emerging optical network layer routing practices.

3.1. Routing and signaling

The main control-plane issues in general transport networks centre around selecting the best path (or set of paths) for the transport of traffic across the network. This activity requires neighbour discovery, network resource discovery, topology state information acquisition and dissemination, topology state information management and path selection. The last one, namely path selection, is clearly the most critical consideration in the design of control planes for switched transport networks. Conventional Link-state interior Routing protocols (IGPs) such as OSPF or ISIS in the GMPLS environment, properly modified to transport all the needed traffic engineering information, are specifically responsible for the reliable advertisement of the optical network topology and of the available bandwidth resources within and between network domains. In the case of OSPF, for example [17], the opaque LSA set has been augmented with new TLVs to support additional traffic engineering characteristics of transport networks. Some of the new link characteristics include: incoming and outgoing interface identifiers, link protection type, shared risk link groups and interface switching descriptor. In order to set up a new bandwidth and/or QoS guaranteed path across the network, a signaling protocol is also required to exchange control information among nodes, to distribute labels and to reserve resources along the path. Suitable signaling protocols for the GMPLS control plane include RSVP and CR-LDP. GMPLS has introduced many traffic-engineering enhancements to the above protocols [1]. For example, the concept of MPLS label has been generalized, to support the reconfiguration of various types of switching elements in transport networks. In our case, the signaling protocol is closely integrated with the routing and wavelength assignment protocols.

3.2. Bandwidth on demand and grooming

The other important service is bandwidth-on-demand. It extends the ease of provisioning even further by allowing the client devices that connect to the optical network to request the setup of connection, with specific bandwidth requirements, in real time as needed. In current networks the traffic demands have typically bandwidth by orders of magnitude lower than the capacity of lambda-links and the number of available wavelength per fiber is limited and costly. Hence, it is not worth assigning exclusive end-to-end lightpaths to these demands, so that a better sub-lambda granularity is strongly required. Thus, to increase the throughput of a network with limited number of lambdas per fiber, *traffic grooming* capability is required in certain nodes, typically those on the network edge. A typical control-plane paradigm ensures traffic grooming capability on edge nodes by operating on a two-layer model, i.e., an underlying pure optical Wavelength Routed network and an “opto-electronic” time-division multiplexed layer built over it. In the wavelength routed layer, operating exclusively at lambda granularity, when a transparent lightpath connects two physically adjacent or distant nodes, these nodes will seem adjacent for the upper layer. The upper layer can perform multiplexing of different traffic streams into a single wavelength-based lightpath via simultaneous time and space switching. Similarly it can de-multiplex different traffic streams of a single lambda-path. It can also perform re-multiplexing as well: some of the demands de-multiplexed can be again multiplexed into some other wavelength paths and handled together along it. The electronic layer is clearly required for multiplexing packets coming from different ports and can be a classical or “next-generation” technology, such as IP/MPLS, but it can also be based on any other networking technology (i.e., SDH/SONET, ATM, Ethernet, etc.). However, the upper layer technology must be unique for all traffic streams that have to be de-multiplexed and then multiplexed again.

3.3. Traffic Engineering services

The main goals of almost all service providers are now of optimizing network resource usage and meeting customer service level agreements. Accordingly, Traffic Engineering has become the essential practice to optimize the utilization of existing network resources and to provide quality of services (QoS) as needed. The state-of-the-art Traffic Engineering practices based on GMPLS technologies are essentially based on classic constraint-based routing optimization methods, realized through the provisioning of explicit alternate label/lambda-switched paths, called LSPs. For optimal LSP establishment, dynamic on-line network optimization is required, usually based on multiple and apparently unrelated metrics and constraints with the objective of optimizing the resource utilization according to cost and performance criteria. This optimization is typically applied on a short-term time basis (at least some tenth of seconds) and to a rather coarse level of flow granularity (e.g., aggregation of all traffic flows between specific ingress/egress nodes). When increasing traffic loads or temporary traffic variations cause localized link congestions, routing optimization, based on alternative network path provisioning, can be carried out to resolve – or at least alleviate – potential network performance problems. The idea is to adjust routing to current load situations and, thus, better utilize available network resources, leading to improved Quality of Service (QoS). As a consequence, new alternative on-line routing algorithms are under study to promote a more balanced and conservative utilization of resources in support to traffic engineering through optimal LSP and lightpath establishment.

3.4. A perspective of the existing RWA solutions

Routing and Wavelength Assignment algorithms supporting QoS or Traffic Engineering features can be conventionally categorized into *static* or *dynamic* depending on the type of routing information used for computing LSP routes. Static algorithms only use network information that does not change with time, and both network topology and usage patterns are known in advance, so that the problem, that is known to be NP-complete [6], is simply to find a solution that optimizes a chosen objective function, e.g., minimizing network resource usage (like number of wavelength used to route connection requests set) or maximizing the network flow. This problem can be straightforwardly solved at its optimum off-line with the known Integer Linear Programming (ILP) techniques in exponential time. On the other side, *dynamic algorithms* use the current state of the network, such as link load and wavelength usage. Generally, dynamic algorithms aim to minimize the total blocking probability in the entire network. On the other hand, routing algorithms can be executed either *online* (on demand) or *offline* (pre-computed) depending on when this computation is applied. In online routing algorithms path requests are attended to one by one, while offline routing does not allow new path route computation. Our work only focuses on dynamic online routing since it is the only acceptable paradigm for future intelligent optical networks, although it might be also used to solve the static RWA problem finding a nearly optimal solution in a reasonable computational time. In a dynamic context it is impracticable to solve the RWA at its optimum due to its intrinsic computational complexity; it is instead preferred to solve it with heuristics that find a sub-optimum solution in acceptable computational time. The most popular and widely used routing algorithm in current networks is the shortest-path first algorithm (SPF). SPF selects the shortest path accordingly to a specific metric. One obvious problem with SPF is that it tends to route traffic onto the same set of links until these links' resources are exhausted. This leads to concentration of traffic on certain parts of the network. In addition, SPF typically accepts less path setups into the network than some other more advanced routing algorithms. Another solution is the Min-Hop Algorithm (MHA) [3] that routes an incoming connection along the path, which reaches the destination node using the minimum number of feasible links (hops).[1] This scheme is simple and computationally efficient. However, using MHA can result in heavily loaded bottleneck links in the network too, as it tends to overload some links and to leave others underutilized. Since it only considers the path length that remains the same independently from the current link load, MHA tends to use the same paths until saturation is reached before switching to other paths with underutilized links. The Widest Shortest Path Algorithm (WSP), proposed in [12], is an improvement of the Min-Hop algorithm, as it attempts to load-balance the network traffic. In fact, WSP chooses a feasible path with minimum hop count and, if there are a multiple of such paths, the one with the largest residual bandwidth, thus discouraging the use of already heavily loaded links. However, WSP still has the same drawbacks as MHA since the path selection is performed among the shortest feasible paths, which are used until saturation before switching to other feasible paths. The shortest widest path (SWP) algorithm [12], on the other hand, introduces further improvements in network load balancing

[1] Although MHA is considered a separate and independent algorithm, it can be seen as a particular version of SPF in which the metric function assigns the same cost $c = 1$ to all network links.

by selecting the path with the maximum available bandwidth and if there are more than one such paths, choosing the one with the least number of hops. However, all the above algorithms suffer from bad performance in terms of rejection ratio in a highly-loaded network and the strategies on which they are based are definitely not “future safe” since they do not consider future LSP requests at all. A more sophisticated class of routing algorithm is based on the minimum interference routing paradigm proposed in [14]. Their key idea is to route an incoming connection over a path which least interferes with possible future requests by using the concept of critical links. Critical links have the property that, when their capacity is reduced by 1 bandwidth-unit, the maximum data flow between a given source–destination node is also reduced by 1 bandwidth-unit. Therefore the goal of such algorithms is accomplished by selecting paths that contain as few critical links as possible. However, while the idea of routing over non-interfering links is a logical one, it is known that minimum interference concept has two critical weaknesses: computation complexity (see Section 6) and unbalanced network utilization. Suppose there are two distinct routes with the same residual bandwidth that connect the same source–destination pair. When a path setup request arrives, given sufficient resources, one of the two routes will be chosen to service this request. Afterwards, all the links in the other route become critical links according to the definition of critical links above. This implies that the same route will serve subsequent requests until saturation while the other route remains free. Therefore, given several distinct routes with enough residual bandwidth, minimum interference routing may converge traffic flows onto a single route causing unbalanced network utilization because it does not take into account the current traffic load in routing decisions [14]. Clearly, lack of knowledge about link load and resource availability information may result in poor performance in a lot of topologies. Consequently, a smarter solution that – while still considering the hypothetical future interference between paths – also keeps into account other strategic factors such as the current load and the network balancing and operates at a comparable or lower computational complexity would be highly attractive and challenging.

4. Objectives and driving criteria

In this section, we define the different objectives for effective traffic engineering-aware routing and wavelength assignment mechanisms in optical multi-layer networks. After a deep study of the standardization efforts [2] and current solutions proposed by the scientific community, we found that three main criteria illustrate the relevant tradeoffs involved in a traffic engineering-capable RWA scheme for the above network environment:

- *Reducing blocking probability*: our first goal is to reduce the blocking probability, ensuring that a maximal number of requests are accepted in the network; hence it maximizes operator revenues and enhances client satisfaction.
- *Minimizing network costs*: static metrics, such as hop count or link static costs, have been traditionally incorporated in routing algorithms in order to achieve a minimum network cost objective. Link static costs can be used as a metric, since they usually correspond to the physical link length. Although it is not foreseen that link length will have a big influence in future networking architectures (especially optical ones), it can still be considered as a static way of expressing operator preference to choose some favourite links. However, the use of dynamic link status metrics, such as up-to-date bandwidth availability will ensure much better performance in terms of optimal network resource usage.
- *Load balancing*: load balancing is an important factor for network congestion reduction. The idea is to have some equilibrated load distribution in the network so that bottlenecks congestions are likely to be avoided thus improving the overall situation. However, [19] shows that, in lightly loaded network, load balancing has some undesirable effects such as routing label or lambda-switched tunnels on longer paths. For example, we can consider the simple heuristic way of doing load balancing by routing lightpaths over the least loaded links. We should point out, however, that this strategy is a basic form of load balancing [27] and is better qualified as load minimization.

5. The minimum affinity approach

In this section we propose a novel approach, based on a two-stage routing algorithm natively conceived to work on complex and multi-constrained optical transport networks based on an improved resource and traffic load-aware LSP affinity concept, properly conceived to cope with the known drawbacks of the state of the art routing algorithms in network load balancing and service request rejection. Our main objective is to minimize the total blocking probability by optimizing bandwidth, cost and length of designed paths while keeping the network resource usage fairly balanced, trying to leave on each link sufficient room to satisfy further requests as much as possible. Furthermore, the overall optimization problem is subject to meeting the designated end-to-end demands while not exceeding the edge capacities. The problem is solved in two stages. In stage 1, or pre-selection stage, the goal is to determine all (or, better stated, a parametric number k) the available paths satisfying the above demands, such as QoS and bandwidth. Stage 2, that is the decision stage, computes, for each path found in stage 1, its *affinity* with all the paths already routed in the network and chooses the least affine one, in order to minimize the LSP concentration on links that would otherwise become resource bottlenecks and leave not sufficient room to keep the network usage fairly balanced. The proposed algorithm operates online, running at each request of a dedicated connection with specific QoS requirements (typically bandwidth capacity) between two network nodes. We make the typical assumption that each connection is bidirectional and consists in a specific set of traffic flows that cannot be split between multiple paths. The connection can be routed on one or more (possibly chained) existing lightpaths between s and d with sufficient available capacity or on a new lightpath dynamically built on the network upon the existing optical links. Grooming decisions are taken instantaneously reflecting an highly adaptive strategy that dynamically tries to fulfill the algorithm's network resource utilization and connection serviceability objectives. In detail, as a new request arrives, the control plane on each node, starting from the originating one, runs our source-based localized RWA algorithm and triggers the proper path setup actions:

- it should determine if the request can be routed on one of the available lightpaths, by time-division multiplexing it together with other already established connections, or a new lightpath is needed on the optical transport core to join the terminating (edge) nodes. In presence of multiple options between new feasible and already established lightpaths, the edge weighting and path selection functions, applied on the existing lightpaths and to the wavelength links that can be used to set up new lightpaths, together with the grooming and the lambda conversion costs, dynamically determine the least-cost routing for the request, on the current network status basis. For example, if two lightpaths between s and d exist, both with sufficient available capacity, the tie is resolved in favor of the least-cost lightpath. Such policy guarantees maximum lightpath utilization and automatically achieves, until possible, effective dynamic grooming assuming that the link state database is properly updated;
- in any case the source node sends a request along the existing path, or the determined new lightpath by using an available signaling scheme;
- all nodes in the path, when they receive the request, run the SPF algorithm, calculate the new network topology and actually establish the requested lightpath and reserve bandwidth resources.

The signaling scheme for triggering the new path or lightpath set-up and reserving the needed bandwidth, fiber or wavelength resources along the path is very similar to the TE-RSVP protocol [3] used by MPLS. To make a reservation request, the source needs the path and the bandwidth that it is trying to reserve. The request is sent by the source along with path information. At every hop, the node determines if adequate bandwidth is available in the onward link. If the available bandwidth is inadequate, the node rejects the requests and sends a response back to the source. If the bandwidth is available, it is provisionally reserved, and the request packet is forwarded on to the next hop in the path. If the request packet successfully reaches the destination, the destination acknowledges it by sending a reservation packet back along the same path. As each node in the path sees the reservation packet, it confirms the provisional reservation of bandwidth. In addition, it also performs the required configuration needed to support the incoming traffic such as setting up labels in an MPLS label-switching node, or reconfiguring the lambda switching internal devices (such as MEMS) in a transparent optical wavelength switching system. In order to accept/reject an incoming request, every node must have knowledge of the available and reserved bandwidth and wavelengths on each outgoing link. This implies that every node needs to run a distributed control-plane protocol that keeps up-to-date information about the complete network topology and available resources. More precisely, a periodic link-state advertisement scheme must convey all the link state information to every node in the network, ensuring the complete synchronization between all the nodes' network status views. Since the amount of per-link state information is very small, any appropriate link state scheme like those employed by OSPF [17] can be adequate for this purpose.

The Dijkstra-based path selection scheme, however, should meet certain conditions:

- a link may not reserve more traffic than it has capacity for;
- shorter paths are preferred because they consume fewer network resources;
- critical resources, e.g., residual bandwidth in bottleneck links, should be preserved for future demands.

The last two conditions reflect that what we really seek is to keep the connection blocking probability (or in other words the rejection ratio) as low as possible, or equivalently to increase the network utilization.

5.1. Building the model

Now that we have clearly stated the problem, we can formally define the network model on which our algorithm works. Let us consider an optical network consisting of Q nodes interconnected by L optical links (fibers), where each link can carry up to $\lambda_{\max}$ wavelengths and each node can be a LSR, a lambda-edge router with several WDM interfaces and wavelength conversion capability or a pure OXC without wavelength conversion capability. The multiplexing or demultiplexing of several connections at any granularity, for grooming sake, is possible only on lambda edge nodes, where wavelength conversion is also possible, whereas in the pure optical nodes with or without wavelength conversion capability it is possible to perform traffic switching at wavelength granularity only. We model such a network with a graph $G = (V, E)$, where $|V| = Q$ and $|E| \geqslant L$, in which the nodes in V may have or not the wavelength conversion capability and all wavelengths on an optical link are represented as different edges between the same pair of nodes, thus obtaining a multi-graph (Fig. 1). Each edge, belonging to E, represents an independent communication channel that can be represented by a tuple $(u, v, f_e, \lambda_e, g_e, l_e, t_e, c_e)$ where u and v are the two edge vertexes, f_e is an index associated to the physical link or fiber number, λ_e correspond to the logical channel number or wavelength index on the fiber, f_e, g_e and l_e are respectively the global wavelength capacity and the current load in bandwidth units. To an edge there may be also associated the TE weight t_e that can be used as a metric for traffic engineering and the cost metric c_e which models the signal degradation introduced by the transmission link. Note that our fiber link representation may support a different number of wavelengths and

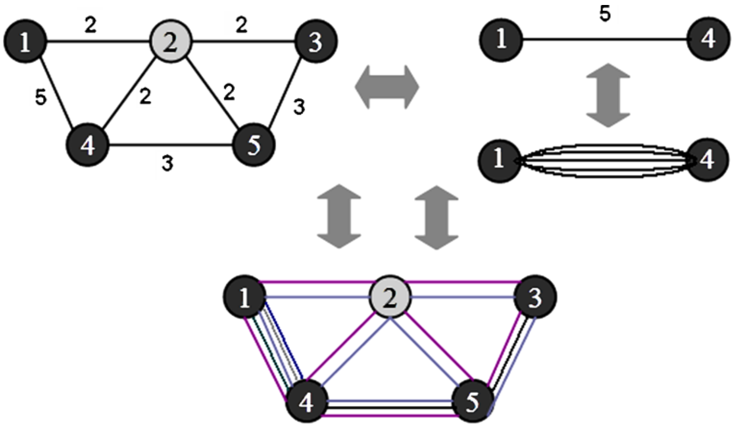


Fig. 1. Generating the working multi-graph. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/JHS-2011-0340.)

that wavelengths on different fibers may have diverse maximum capacities g_x, according to the WDM equipment involved. Optical switching nodes and lambda-edge nodes are completely represented by a couple (v, Λ) by adding to each vertex v in the graph a wavelength conversion capability attribute Λ that indicates the possibility to violate the wavelength continuity constraint by choosing, when building a lightpath traversing v, one of the available connected wavelengths. The conversion capability attribute has been specifically defined in (1) to implicitly represent the complexity and expensiveness, in performance terms, of the conversion devices, so that:

$$\Lambda_x = \begin{cases} 1, & \text{Transparent WXC,} \\ 2, & \text{Opaque WXC,} \\ 4, & \text{Lambda Switching Router,} \\ 8, & \text{Router.} \end{cases} \tag{1}$$

Thus, in our network model, when choosing an output wavelength through a wavelength conversion-capable node, we will be limited only from the *clash* constraint, that is, each converter allows the conversion from each input wavelength to each available output wavelength. We make no specific assumption on the number of wavelengths per fiber, number of fiber on each link and on the presence of wavelength conversion devices on the network. All these parameters are fully and independently configurable on each fiber link or network device at the network topology definition time. Instead, we require that all the network nodes operate under a unique control-plane and share a common network view by relying on a common link-state protocol that is used to distribute resource usage information. Furthermore, we also assume that every connection is bidirectional and consists in a specific set of traffic flows that cannot be split among multiple paths. Our interest focuses on the optimized and fairly balanced setup of traffic engineered paths on the optical transport network, where all the connections both on the network core and edge are typically bi-directional. Demands, in terms of the number of end-to-end connection requests, due to new LSP setup, vary from one source–destination pair to another, but each connection has to be set-up on a lightpath preferably using the same wavelength (i.e., one lambda, with the wavelength continuity constraint). Formally, an LSP request can be characterized by a tuple $(s, d, q_1, q_2, \ldots, q_n)$, where s and d are the source and destination nodes, and $q_1, q_2, \ldots, q_n$ are the QoS requirements of the LSP. Nevertheless, those QoS requirements – such as packet loss rate, packet delay, jitter, etc. – can be incorporated into a single bandwidth requirement as shown in [14]. In this case the LSP request can be represented by a triplet (s, d, b) where b is the LSP required bandwidth. An LSP between s and d is a simplex flow, i.e., packets flow in one direction from s to d along a constrained routed path. For the reverse traffic flow, an additional simplex LSP must be computed and routed from d to s. Clearly, the path from s to d can be different from the path from d to s. Also, the amount of bandwidth reserved on each path can differ. Obviously, after that a LSP P has been routed in the network, the quantity l_e will change for each edge e of P, and the weights of these edges will be updated. More precisely, each time a new lightpath needs to be established between an ingress–egress pair of nodes in our multi-graph we modify it by removing the graph edges traversed by the lightpath (corresponding to the wavelength used) and by adding a direct edge x, called *cut-through* edge, with capacity g_x set to g_{min} and current load l_x set to b, where b is the fraction of the link bandwidth required by the lightpath and g_{min} is the minimum global wavelength capacity between all the edges belonging to the lightpath. A *cut-through* edge x, that can be easily distinguished by having a negative fiber identifier ($f_x < 0$ where $|f_x|$ will be the lightpath identifier), can be used in any path selection operation and thus can participate in one or more LSPs as a single virtual edge (a single hop at the IP layer) terminated on converter nodes. It will not be directly used in affinity computation, but its physical edges will actually do, as explained in Section 5.4. When an established lightpath is torn down because the last connection occupying it is ended, the cut-through arc is removed and the edges in the extended graph corresponding to the underlying physical links are set back with full capacity. Similarly, if an LSP is deleted all the edges belonging to it will have their capacity partially restored by adding the bandwidth required by the LSP to their residual capacity. This schema allows modeling the wavelength availability per link and the residual bandwidth per logical link at the IP layer. The algorithm used to dynamically route the connection requests works on such a multi-graph.

Let us now give a formal definition of the problem to be solved. For each edge $e \in E$, let g_e be the global capacity and l_e the current load on the edge e. Let:

$$w : E \to \mathfrak{R} \tag{2}$$

be the edge weighting function. To simplify notation, let w_e denote the weight associated to the edge $e \in E$, i.e., $w_e \equiv w(e)$.

The edge weighting function w_e can assume the following values:

$$w_e = \begin{cases} 1, & \mathit{MHA}, \\ \mathit{crit}_e, & \mathit{MOCA}, \\ \mathit{load\ balance}, & \mathit{Affinity}, \end{cases} \tag{3}$$

where the specific value, depending on load balance that is used in our schema will be explained in detail in the following Section 5.2. We also introduce $m = |E|$ Boolean decision variables x_e, one for each edge $e \in E$:

$$X = (x_1, x_2, \dots, x_m) \in B^m, \quad B = \{0, 1\}. \tag{4}$$

To formally characterize the problem and its computational complexity, let us model the general problem as an ILP. The problem consists in determining the optimal set of binary variables X that optimizes the objective function (5) in the following mathematical formulation:

$$\min \sum_{e \in E} w_e \cdot x_e \tag{5}$$

subject to the following constraints:

$$\forall v \in V \quad \sum_{e \in \delta_G(\{v\})} x_e = \begin{cases} 0 & \text{if } v \notin P, \\ 1 & \text{if } v = s \vee v = d, \\ 2 & \text{if } v \neq s \wedge v \neq d \wedge v \in P, \end{cases} \tag{6}$$

$$\forall e \in E \quad x_e \cdot b \leqslant g_e - l_e, \tag{7}$$

$$\forall e \in E \quad x_e \in \{0, 1\}, \tag{8}$$

where

$$\delta_G(S) = \{e_{i,j} \in E \mid i \in S, j \in V \backslash S\}, \quad S \subset V \tag{9}$$

is the *cut* of G with respect to S and $e_{i,j} \in E$ is an edge between nodes $i, j \in V$ and P is the path under construction.

The constraint (6) imposes the flow conservation property. Note that the property is guaranteed because the source and the destination elements are the only nodes that have exactly one edge belonging to the path. The intermediate node v has either zero or two edges of the path, meaning that it is, respectively, a node that does not belong or does belong to the path. The constraint (7) imposes the resource availability constraint: an edge $e \in E$ must have enough available residual bandwidth to accommodate the LSP required bandwidth. The constraint (8) formally specifies the Boolean nature of the decision variables x_e. Clearly, such a formulation implies a-priori knowledge of the entire traffic matrix to be solved at optimum through integer linear programming; furthermore, the ILP can be reduced to a single-commodity NP-complete problem. Thus, our fully dynamic operating requirements drives us in searching a more efficient heuristic-based approach, that starting from the same theoretical model, allows us to achieve satisfactory sub-optimal solutions in polynomial times.

5.2. Choosing the weighting function

From (3) it can easily be noted that the edge weighting function w_e, can be properly chosen to reflect different weighting policies and objectives in the overall optimization strategy and strictly influences the behaviour of the path selection algorithm. In order to achieve our load balance objective, we specify the following properties that a *good* weighting function should satisfy:

$$\forall e \in E\colon \quad l_e = g_e, \qquad w_e \approx \infty. \tag{10}$$

That is, the weight w_e associated to an edge e with no available residual bandwidth should have the maximum cost and cannot belong to any new lightpath:

$$\forall e, f \in E\colon \quad g_e = g_f \wedge l_e < l_f, \qquad w_e < w_f. \tag{11}$$

Considering two edges with the same global capacity and different current load values, the weight associated to the highest loaded one must be higher than the other. This property privileges the choice of edges with highest residual capacity:

$$\forall e, f \in E\colon \quad l_e = l_f \wedge g_e < g_f, \qquad w_e > w_f. \tag{12}$$

Given two edges with the same load and different global capacities, the weight associated to the one with highest global capacity must be lower than the other. This is due to the consideration that links with lower global capacities are more prone to saturation and usually have lower chances to recover bandwidth in time from connection teardown. Note that the scaling factors implicit in Eqs (11) and (12) may be different:

$$\forall e, f \in E\colon \quad \left(\frac{l_e}{g_e} = \frac{l_f}{g_f}\right) \wedge (g_e \neq g_f \vee l_e \neq l_f), \qquad w_e \neq w_f. \tag{13}$$

Every two edges with the same load/maximum capacity ratio but different current load or global capacity values must have different associated weights. This avoids a common mistake in assigning the same weight to two edges with the same saturation ratio and bandwidth resources.

Obviously, the weighting function should be proportional to the load and inversely proportional to the global capacity the edge; indeed it should be proportional to the TE parameters:

$$w_e \propto l_e, \qquad w_e \propto c_e, \qquad w_e \propto t_e, \qquad w_e \propto \frac{1}{g_e}. \tag{14}$$

Another important consideration is that the global capacity g_e is less important than the current load l_e in assigning weights; we used the logarithm of g_e instead of g_e, in order to lessen its relative importance in the product weighting function. We chose Eq. (15) as the edge weighting function, which satisfies all the stated properties:

$$w_e = \frac{l_e c_e t_e}{\log g_e}. \tag{15}$$

Thus each edge is weighted according to the statically assigned cost and TE metric, and against its maximum capacity. Different wavelength routing policies can be realized by modifying the t_e components of the edges in E. In fact, these weights may be used to reflect the cost of network elements such as wavelength converters (in LSR routers or lambda-edge nodes) or free wavelengths on some links and to characterize the QoS properties of different wavelength channels (such as delay, capacity, etc.). Also the transmission impairments introduced by the physical layer will be considered in this first path selection stage of our algorithm, which can assign optical paths to incoming requests only if the resulting lightpath is feasible (i.e., the lightpath transmission properties are good enough to guarantee the required transmission quality to the bit-stream). Thus, by modifying the above weighting factors according to the incoming service class request, it is possible to choose a path which minimizes the number of conversions or which maximizes the usage of existing lightpaths. Different decisions reflect different objectives in term of network resource utilization. As a result, a request can be routed over a direct lightpath (a single-hop path at the IP level), modeled as a single cut-through edge in our multi-graph, if it crosses only nodes that cannot perform wavelength conversion between an ingress and an egress router, or over a sequence of lightpaths (a multi-hop path at the IP level, where each hop can be a lightpath), if it crosses nodes that are wavelength conversion capable (lambda-edge or routers as well). Note that a lightpath in the optical domain corresponds to a single wavelength crossing a certain number of nodes, without wavelength conversion, represented as a single cut-through edge. Simply stated, to satisfy a connection request with bandwidth b and class of service q we can first run our constrained SPF algorithm on the above multi-graph G by considering only the edges e with load l_e and global capacity g_e such that $g_e - l_e \geqslant b$ and weight $t_e \leqslant q$, otherwise, if no such a direct lightpath exist and a new path has to be set up in any case, the SPF algorithm should be applied on the graph G considering only those edges in E with sufficient residual capacity $g_e - l_e$. In this way, when all the previous operations do not result in any available direct path, an indirect path, built on multiple lightpaths all with adequate capacity and QoS characteristics (as above) passing through a conversion edge, can be chosen.

5.3. Determining the feasible paths

At the first stage of our algorithm the list of the first k feasible paths in increasing order of cost between the source s and the destination d of the required connection and constrained by the bandwidth requirement of the connection request is obtained by running a constrained K-shortest-path-first (K-SPF) algorithm on the above extended graph. Here k is an external configurable parameter that can be used to limit the number of feasible paths that should be considered in the following steps, thus controlling the depth of the analysis process according to a performance/precision compromise. Simply stated, the main idea is determining the k required paths in a multi-stage process according to the well-known Katoh–Ibaraki–Mine algorithm [16] capable of determining multiple cycle-free shortest paths connecting two terminal nodes in a graph. The above algorithm has been properly crafted to meet the specified constraints and bandwidth requirements of each new request. The SPF algorithm used in each stage has also been modified to cope with the continuity constraint when choosing any edge during the path building process. Accordingly, when traversing origin or converter nodes we are totally free in selecting any outgoing edge (i.e., any wavelength), whereas with all the other nodes we can only select an outgoing edge corresponding to the same wavelength associated to the incoming one. Clearly, the cut-through edges can only be used to build label switched paths multiplexed on already existing lightpaths terminated on converter nodes (through grooming) and not when we need to set-up new lightpaths in the all-optical domain. The above algorithm also tries to optimize path selection by considering each edge's dynamically changing properties such as the available residual bandwidth r_e, in addition to the static ones such as the global capacity, the assigned cost c_e and TE weight t_e.

5.4. Choosing the minimum affinity path

Once the candidate path list for the new LSP has been determined, according to our constraints and weighting function, we have to choose the best path between them to satisfy our more sophisticated medium and long-term network optimization objectives. To do this we need to define a new concept, the affinity between LSPs that will be crucial to our algorithm. The affinity between two paths defines the degree of similarity, in terms of sharing of network resources between them. So, the more two paths are affine, the more they tend to use, and consequently exhaust, some common network resources that will be no more available for future requests. This means overloading some links and leaving others almost unused. This is the exact opposite of our medium and long-term objectives. Consequently, to keep our network usage more balanced, and reduce the blocking probability we have to choose our LSPs in a way that their mutual affinity will be minimal. This can be performed, on-line and stepwise at the time of creation of each new LSP, according to an incremental process aiming to achieve a near-optimum solution.

Let's define formally the affinity concept, starting from the aforementioned multi-graph G with $|V|$ nodes, $|E|$ edges, L fibers and up to $\lambda_{\max} \in N$ (where N is the set of natural numbers) wavelengths available per fiber (the actual number of wavelengths is totally free and may vary from fiber to fiber). The above multi-graph, used for affinity calculation, will be specifically built by discarding all the edges in E with negative fiber index, which are the cut-through edges. Let $F_i, i = 1, 2, \ldots, L$, be a partition of the set of edges E. Let $a \in E$ be an edge. Let $f : E \to N$ be the characteristic function of the partition F_i that returns the fiber index of an edge $a \in E$:

$$f(a) = \begin{cases} f_1 & \text{if } a \in F_1, \\ f_2 & \text{if } a \in F_2, \\ \ldots & \\ f_L & \text{if } a \in F_L, \end{cases} \quad \text{with } f_i \in N, 1 \leqslant i \leqslant L. \tag{16}$$

We can write:

$$F_i = \{a \in E \mid f(a) = f_i\}, \tag{17}$$

that is, F_i is the set of edges belonging to the ith fiber. Now, let $\lambda : E \to N$ be the function that returns the wavelength index on the fiber of an edge $a \in E$:

$$\lambda(a) = \begin{cases} \lambda_1 & \text{if edge } a \text{ has lambda index } \lambda_1, \\ \lambda_2 & \text{if edge } a \text{ has lambda index } \lambda_2, \\ \ldots & \\ \lambda_{\max} & \text{if edge } a \text{ has lambda index } \lambda_{\max}, \end{cases} \quad \text{with } \lambda_i \in N, 1 \leqslant i \leqslant \max. \tag{18}$$

We define *affinity between two edges* $c, a \in E$, and we write $\varphi_{c,a}$, the function:

$$\varphi_{c,a} : E \times E \to N \tag{19}$$

in particular:

$$\varphi_{c,a} = \begin{cases} 0 & \text{if } f(c) \neq f(a), \\ 1 & \text{if } f(c) = f(a) \wedge \lambda(c) \neq \lambda(a), \\ 3 & \text{if } f(c) = f(a) \wedge \lambda(c) = \lambda(a). \end{cases} \tag{20}$$

The weights of the affinity function φ are assigned according to an exponential trend. In the first case we have no *overlap* at all between the two edges as they belong to different fibers; in the second case the two edges belong to the same fiber but they represent different lambdas, so we have a partial overlap; in the last case, we have a complete match, that is the two edges are actually the same one, and the affinity is evaluated with the maximum value. Note that the third case is possible as we support grooming.

Let's see how to use the affinity function between two edges to calculate the affinity between two paths. Let Π be the set of the paths already routed in the network; let C be one of the k candidate paths found in step 1. Let $c \in C$ and $p \in P$ be edges of the respective paths. Let l_P and l_C denote the lengths of the paths P and C, respectively. We want to calculate the affinity of the path C with all the paths in Π. Formally, we define the *affinity* Φ *of a path* C *with all the paths in* Π as the quantity $\Phi_\Pi(C)$ given by:

$$\Phi_\Pi(C) = \sum_{P \in \Pi} \min\{l_P, l_C\} \sum_{p \in P} \sum_{c \in C} \varphi_{p,c}, \tag{21}$$

that is, we sum for each path P in Π the affinity between each edge $p \in P$ with all the edges $c \in C$, multiplied by the minimum length between the lengths of the paths P and C. We chose to multiply by this factor because in this way we can weight the effective overlapping of the two paths.

Equation (21) is theoretically what the affinity is. The above definitions of affinity involve a summation over all the previously established lightpaths. This can be computationally quite inefficient, particularly when connections have a relatively high lifetime and thus the number of active lightpaths grows considerably, as we have to calculate the affinity of each of the k candidate paths with all the paths already routed in the network. Furthermore, the path affinity concept is scarcely dependent from the LSP endpoints and is based on the contribution of each single edge in the path, strongly considering both the lightpath length and the resource sharing between paths. For this reason we have developed the *aggregated affinity* approach. It consists in aggregating together the paths in Π by their length. In fact, note that in Eq. (21) it is not important to keep all the information bound to the path P, but we are only interested in knowing the path length and the path's edges, that is what edges the path crosses.

Formally, we can further develop Eq. (21):

$$\Phi_\Pi(C) = \sum_{P\in\Pi} \min\{l_P, l_C\} \sum_{p\in P} \sum_{c\in C} \varphi_{p,c} = \sum_{c\in C} \sum_{P\in\Pi} \min\{l_P, l_C\} \sum_{p\in P} \varphi_{p,c} \tag{22}$$

obtaining that we can sum on the candidate path's edges before summing on the P's edges. Also note that we are not interested in those edges of P that do not even partial overlap with C's edges (i.e., edges belonging to different fibers). Equation (20) assigns to these edges an affinity value of 0, so they do not contribute in the sum. We do not consider such edges, thus dramatically lowering the computational complexity. We are only interested in knowing how many paths pass through edges belonging to fibers to which C's edges belong.

So, for each edge $e \in E$ of the network we can aggregate paths of a given length passing through the edge e. The number of the possible lengths of a cycle-free point-to-point path in a network with Q nodes is $Q - 1$. We store aggregated information about paths in Π in a matrix $M^{Q-1,L}$, with $Q - 1$ rows and L columns. Let $m_{l,e}$ be the generic element in row l and column e of M. If $m_{l,e} = r$ then it means that there are r paths of Π of length l that pass through the edge e. At this point we are able to calculate the affinity between the candidate path C and all the paths in Π with the efficient *operational* formula reported in (23):

$$\Phi_\Pi(C) = \sum_{c\in C} \sum_{e\in F_{f(c)}} \sum_{l=1}^{Q-1} \min\{l, l_C\} \cdot M_{l,e} \cdot \varphi_{c,e}. \tag{23}$$

Finally, for each new LSP setup request, we have to choose the path C_i, $i = 1, 2, \ldots, k$, from the list of the k constrained shortest paths determined in the previous step that minimizes the affinity with paths in Π; we choose the path C_j, $1 \leqslant j \leqslant k$, that has the minimum affinity with all the paths in Π:

$$C_j = \operatorname*{arg\,min}_{i=1,2,\ldots,k} \{\Phi_\Pi(C_i)\}. \tag{24}$$

The choice of routing the LSP request along the path C_j in Eq. (24) is aimed to keep the resource utilization the more balanced as possible while privileging shortest and less costly paths.

5.5. *Putting all together*: *The minimum affinity routing algorithm*

Finally, the whole minimum affinity routing algorithm whose functional components have been detailed in the previous sections may be described by the following steps:

- *Initialization*: starting from the network layout build the extended graph $G = (V, E)$, as described in Section 5.1, and associate to each edge the corresponding weight and TE constraints according to the weighting function reported in (15). In the network control-plane implementation the weights and resource status information will be made dynamically available online to any network node by the deployed link-state IGP (such as OSPF or IS-IS), properly extended with TE features.

- *Stage* 1: Apply the previously described constrained K-SPF algorithm on G and determine the ordered list of the k shortest feasible paths according to the weighting function (15).
- *Stage* 2: Determine between the k candidate paths found in the previous section which one has the minimum affinity with all the existing LSPs (24), and use it to build the new LSP request, according to the signaling mechanisms provided at the control-plane level. Update the matrix M to take into account the new LSP and update the edges' weights of working graph G.

6. Computational complexity analysis

The first thing to be considered when analyzing the space complexity of the above schema is that our multigraph-based network representation greatly reduces space complexity compared to the layered graph approach conventionally used in solving dynamic online RWA problem, shown in Fig. 2. In a network with $\lambda_{\max} = 2$ wavelengths per fiber the layered graph in (b) is obtained by replicating the original network graph (a) $\lambda_{\max}$ times, one per wavelength, and then by connecting nodes with wavelength conversion capability together between layers. Besides, for each LSP request it is necessary to add two fictitious nodes s and d and connecting them with infinite capacity links to each corresponding source and destination node in all layers. In Fig. 2 is reported an example of the layered graph for a LSP request between source node 1 and destination node 3. For comparison, in Fig. 2(c) is reported the corresponding multi-graph for network graph (a). As it can be seen, the multi-graph approach saves space with respect to the layered graph approach.

Formally, in a network with Q nodes, L links and using up to $\lambda_{\max}$ wavelengths on each link the layered graph representation with Ω converters requires in the worst case $(Q \cdot \lambda_{\max} + 2)$ nodes and $(L \cdot \lambda_{\max} + 2\lambda_{\max} + \Omega \times (\lambda_{\max} - 1))$ edges whereas the equivalent multi-graph requires only Q nodes and $L \cdot \lambda_{\max}$ edges, greatly limiting the overall space complexity. As for the Affinity algorithm, the only space complexity is given by the matrix M which

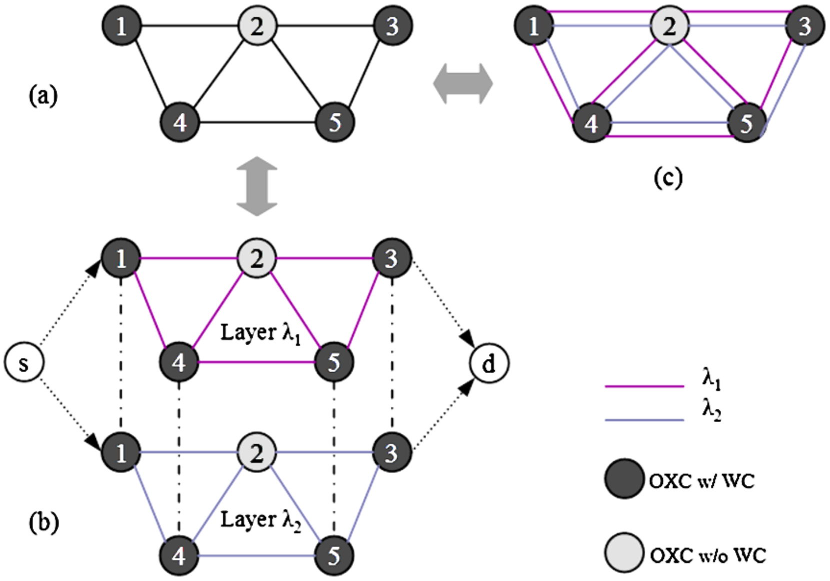


Fig. 2. The original graph (a) with two wavelengths per link. The corresponding layered graph (b) and multi-graph (c). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/JHS-2011-0340.)

requires the memory for a matrix of size $(Q-1)\cdot L$ to be stored only once. Let's now examine the computational complexity of the routing algorithm presented above and compare it with the other widely known algorithm such as MHA/SPF and MIRA/MOCA. Consider our network with Q nodes, L fibers, up to $\lambda_{\max}$ wavelengths on each fiber, E edges. Computing the ordered set of the first k shortest paths for a specified source–destination pair requires in the worst case $O(k\cdot(E+Q\cdot\log Q))$ with the Katoh et al. [16] K-SPF algorithm. The second stage requires for each of the k feasible paths the computation of the affinity with all the already existing LSPs Π. This computation is done by the operational affinity formula reported in (23). This formula involves three sums. The first one is on edges of C: the maximum length of a cycle-free path in a graph with Q nodes is $Q-1$. The second sum is on all edges belonging to a same fiber: this number is upper bound by $\lambda_{\max}$. The third sum repeats $Q-1$ times a calculus that can be done in constant time. Thus, the calculation of the affinity $\Phi_{\Pi}(C)$ takes in the worst case $O((Q-1)\cdot\lambda_{\max}\cdot(Q-1))$. This computation has to be evaluated for each of the k shortest paths found in stage 1. The maintenance of the matrix M can be done for each LSP routed in the network in $O(Q-1)=O(Q)$. Consequently, the overall affinity complexity requires $O(k\cdot(E+Q\cdot\log Q)+k\cdot\lambda_{\max}\cdot Q^2+Q)=O(k\cdot E+k\cdot\lambda_{\max}\cdot Q^2)$. Considering that k and $\lambda_{\max}$ are assigned parameters and their values remains quite low in all cases, the dominating factor in the complexity is $O(E+Q^2)$. This complexity is a little higher than SPF ($O(Q^2)$ with simple Dijkstra shortest path algorithm that can be improved to $O(E\cdot Q\cdot\log Q)$ by using a priority queue with a Fibonacci heap in the implementation [18]), and much lower than that of MIRA/MOCA: let's suppose there are p source–destination pairs in a network with Q nodes and E links. MIRA/MOCA requires p maximum flow calculations to determine the set of critical links each time a path setup request arrives. In the worst case, every node is a source node for every other node, and so p becomes Q^2. Since each maximum flow calculation is $O(Q^3)$ [8], therefore the worst case runtime of MIRA/MOCA is $O(Q^5)+O(E^2)$. However, with the introduction of the Goldberg maxflow algorithm [11] the MIRA/MOCA complexity can be reduced to $O(Q^3\cdot E\cdot\log(Q^2/E))$ that is still much higher than that of the Affinity algorithm.

7. Performance evaluation

In this section, we examine the performance of our new algorithm with an extensive simulation study, by working on several real network topologies, with and without the continuity constraint (i.e., with and without wavelength converters). The simulation details together with the most interesting results and observations emerged from the experiments have been reported in the following paragraphs.

7.1. The simulation environment

In order to evaluate the performance of the proposed algorithm we realized a simple and very flexible ad-hoc optical network simulation environment totally written in Java in order to take advantage of its great extensibility, ease of modifiability, portability and strict math and type definitions. It also supports discrete-event simulations in WDM optical network and fiber/lambda switching for several wavelength routing algorithms, such as our Affinity-based paradigm, MHA, SWP and MOCA, with basic wavelength assignment paradigms such as First Fit and wavelength conversion capability. It supports a very intuitive GUI interface for flexible definition and modification of simulation parameters and a sophisticate configuration from file mode to define complex simulation environments, allowing the creation of new network topologies. Simulations have been performed on several well-known network topologies such as NSFNet [21] and Geant2 [10]. These networks have been modeled as undirected graphs in which each link has a non-negative capacity ranging from OC-1 to OC-768 bandwidth units. In all the experiments, we used a dynamic traffic model in which connection requests arrive according to a Poisson process with a rate λ of requests/seconds. To enhance the effect of connection's load on the network, the session holding time has been set to be infinite, that is each connection lasts through the entire simulation. The connection requests are distributed on all the network nodes according to the probability distribution obtained from the traffic matrices given in [23] for NSFNet and in [26] for Geant, where traffic volumes have been scaled proportionally to the traffic distribution. The LSP bandwidth demands are taken to be uniformly distributed between the values reported in Table 1.

Table 1

Simulations performed and parameters used

	NSFNet/Geant2
Number of connections	Varying from 0 to 10,000 (step 1000)
Random generated bandwidths (OC-unit)	$\{1, 3, 12, 24, 48, 192\}$
k	$\{3, 4, 5, 8\}$
Number of simulations	80 simulations ran per topology; each simulation repeated 10 times

7.2. *Results analysis*

For performance comparison, we ran each simulation based on four routing algorithms: MHA, SWP, MOCA and our Affinity-based algorithm. All the results presented are taken from many simulation runs on the above network with several k parameter values and connection requests varying from 0 to 10,000. Simulations have been performed on an HP® DL380 Dual Processor (Intel® Xeon® 2.5 GHz) server running FreeBSD® 4.10 operating system and Sun® Java® 1.4.2 Runtime Environment. The k parameter and bandwidth unit request values used in our simulations are reported in Table 1.

As can be seen from the previous table, 80 simulations per topology were run and, to obtain more confidence in the results, each run has been repeated 10 times and the average performance metric values have been calculated. Thanks to the consistency of the results obtained, only the graphs relating to one value of k per topology are shown. The most significant performance metric observed in our experiments is the path-setup rejection ratio. Clearly, a smaller rejection ratio indicates a better resource usage, and hence a more balanced network utilization in the medium and long term. The k factor has to be chosen accordingly to a compromise between execution time and performance. Anyway, in our simulations we found that quite low values of k are sufficient to get the best results (typical values are $k = 3, 4, 5$). The improvement in the results obtained with greater values does not justify the extra computational effort. Anyway, in general, as the meshing degree increases, and thus more solutions are available, higher k values lead to more interesting results.

Compared to other well-known algorithms, such as MHA and SWP (extended to work in a WDM environment), and MOCA our heuristic performs significantly better when the load increases, since it is able to overcome some of their most common drawbacks by taking into account the overall unbalancing and blocking effects (see Figs 3 and 4). We can observe how MHA, SWP and MOCA work better under light loads but their performance drastically reduces when the network load starts to be significant. In our sample simulation results plotted in Figs 3 and 4 tests, MOCA/MHA/SWP show similar behavior while Affinity is fairly different, although it seems to exhibit a closer match to the other algorithms in the Geant2 case. More precisely, Affinity seems to sacrifice some performance under lower loads in order to gain significantly when the load increases. This is due to the Affinity behavior in choosing paths that improve load-balancing and network utilization in the medium term, whereas the other algorithms tend to unbalance the network load, by overloading only the links belonging to the "best" lightpaths. In order to stress the differences between Affinity and the other RWA algorithms, we tested them in low- and high-loaded networks, respectively NSFNet and Geant2. In both scenarios, Affinity achieves to behave substantially better that the other strategies, thanks to its better medium term load balancing technique. In detail, Affinity begins to behave significantly better starting from the intersection point where its blocking factor reaches equality with those of the MOCA, MHA and SWP algorithms. For NSFNet, this happens around 15% blocking probability whereas for Geant2 around 75%. This difference is essentially due to saturation phenomena that occur later in Geant2 because its topology is more complex, having a higher meshing degree and more heterogeneous WDM equipment (sparse wavelength conversion), and therefore exhibiting a better traffic load absorption behavior. The performance gap at higher loads is evident both in presence of converters (Fig. 3) and in networks without conversion capacity (Fig. 4). We can also observe that if only the shortest path heuristic is employed, such as in MHA, the paths composed of multiple active lightpaths are more likely to be picked. The reason is that the cut through arcs which usually bypass several optical links tend to be shorter.

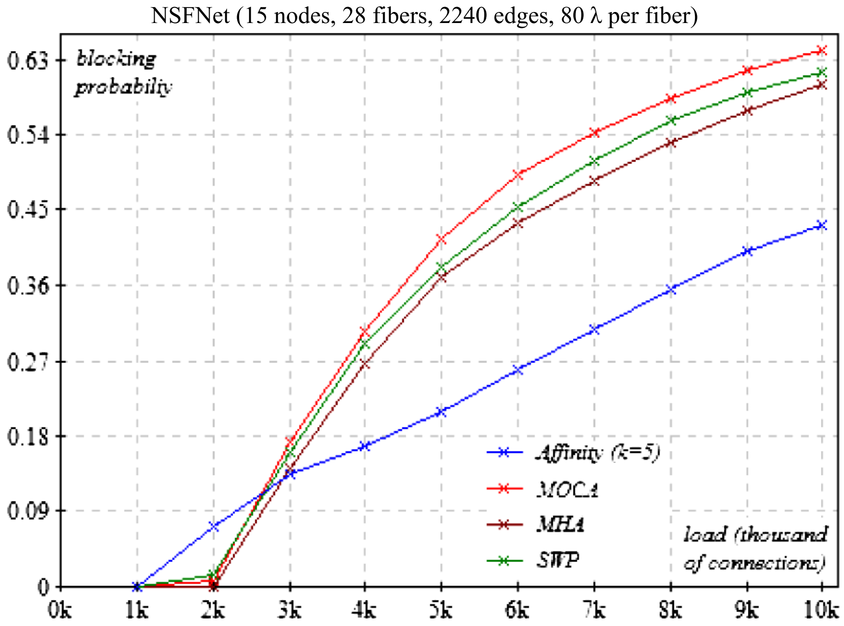


Fig. 3. Average rejection ratio on NSFNet (simulations from Table 1). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/JHS-2011-0340.)

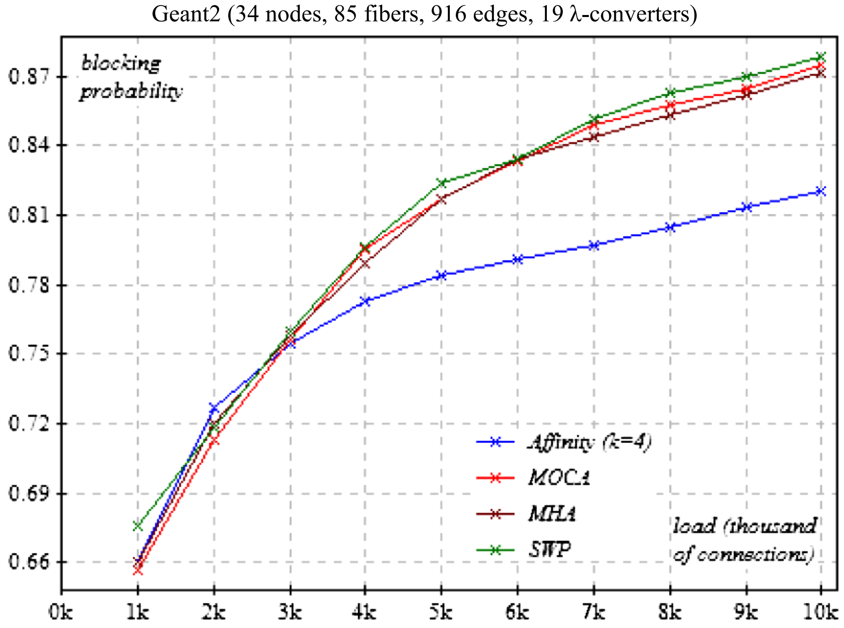


Fig. 4. Average rejection ratio on Geant2 (simulations from Table 1). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/JHS-2011-0340.)

Besides the very interesting results in blocking probability, our RWA schema, as already demonstrated in Section 6, operates significantly faster than MOCA, and moderately slower than MHA and SWP as we can see from Fig. 5 where the average elapsed times on Geant2 are shown. In summary, Affinity is at advantage at high loads

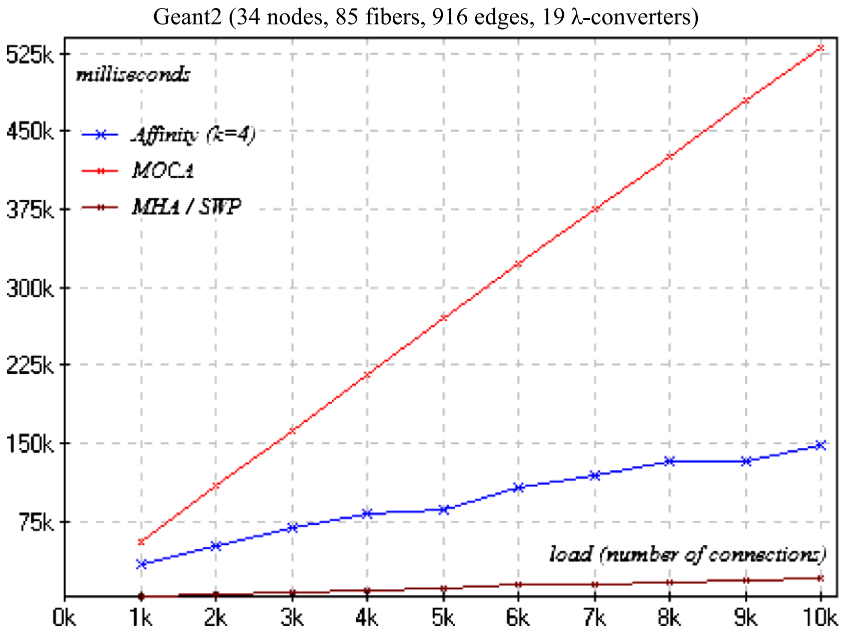


Fig. 5. Average elapsed times on Geant2 (simulations from Table 1). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/JHS-2011-0340.)

while being comparable to other algorithms in terms of computation time, ease to adapt to different network configurations and flexibility to take into account different objectives such as, for example, power consumption.

8. Conclusions

It has been demonstrated [17] that the introduction of wavelength-routed networks not only offers the advantages of higher transmission capacity and routing node throughput, but also satisfies the growing demand for protocol transparency and simplifies operation and management. However, a lot of technical issues need yet to be resolved before a new RWA paradigm truly becomes part of an effective and flexible control-plane framework, and the most important and widely studied one is *dynamic* set-up of *QoS guaranteed* lightpaths or LSPs, performed through an efficient and optimized routing or path design algorithm that determines the “best” paths/routes according to a given objective. To cope with the above challenges our work presents a two-stage wavelength routing algorithm, easily integratable in state-of-the art routing and signaling protocols and technologies, built on an on-line dynamic grooming scheme that finds a set of feasible routes on lightpaths which fulfill some QoS and traffic engineering requirements and bases its final choice on a novel heuristic global path affinity minimization concept. Our new algorithm demonstrated the capabilities of achieving a better load balance and resulting in a significantly lower blocking probability than the existing methods for both optical networks under the wavelength continuity constraint and with sparse wavelength converters, as verified by an extensive simulation study. The ability to guarantee both a low blocking probability and a low computational complexity make the new on-line dynamic RWA algorithm very attractive for the modern multi-layer optical circuit switched and optical wavelength-switched networks.

References

[1] P. Ashwood-Smith et al., Generalized MPLS: signaling functional description, IETF RFC 3471, 2003.
[2] D. Awduche, J. Malcolm, J. Agogbua, M. O’Dell and J. McManus, Requirements for traffic engineering over MPLS, RFC 2702, 1999.
[3] J.D.O. Awduche, L. Berger, D. Gain, T. Li, G. Swallow and V. Srinivasan, Extensions to RSVP for LSP tunnels, Internet draft, 1999.

[4] R. Boutaba, W. Szeto and Y. Iraqi, DORA: efficient routing for MPLS traffic engineering, *Journal of Network and System Management* **10**(3) (2002), 309–325.

[5] S. Chen and K. Nahrstedt, An overview of quality-of-service routing for the next generation high-speed networks: Problems and solutions, *IEEE Network Mag.* **12** (1998), 64–79.

[6] I. Chlamtac, A. Ganz and G. Karmi, Lightpath communications: an approach to high-bandwidth optical WANs, *IEEE Transactions on Communications* **40** (1992), 1171–1182.

[7] J. Comellas, R. Martínez, J. Prat et al., Integrated IP/WDM routing in GMPLS-based optical networks, *IEEE Network* **17**(2) (2003), 22–27.

[8] B. Fortz and M. Thorup, Internet traffic engineering by optimizing OSPF weights, in: *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communication Societies (IEEE INFOCOM)*, Tel Aviv, Israel, 2000.

[9] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Co., New York, NY, USA, 1979.

[10] Géant2 network, http://www.geant2.net/.

[11] A.V. Goldberg and R.E. Tarjan, A new approach to the maximum flow problem, in: *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, and *Journal of ACM* **35**(4) (1988), 921–940.

[12] R. Guerin, D. Williams and A. Orda, QoS routing mechanisms and OSPF extensions, in: *Proceedings of GLOBECOM*, November 1997.

[13] A. Jukan and G. Franzl, Path selection methods with multiple constraints in service-guaranteed WDM networks, *IEEE/ACM Trans. Networking* **12**(1) (2004), 59–72.

[14] K. Kar, M. Kodialam and T.V. Lakshman, Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications, *IEEE Journal on Selected Areas in Communications: Quality of Service in the Internet* **18**(12) (2000), 921–940.

[15] K. Kar, M. Kodialam and T.V. Lakshman, Integrated dynamic IP and wavelength routing in IP over WDM networks, in: *IEEE Infocom*, 2001.

[16] N. Katoh, T. Ibaraki and H. Mine, An efficient algorithm for k shortest simple paths, *Networks* **12** (1982), 411–427.

[17] K. Kompella et al., OSPF extensions in support of generalized MPLS, IETF RFC 4203, 2005.

[18] H.R. Lewis and L. Denenberg, *Data Structures and Their Algorithms*, HarperCollins, New York, 1991.

[19] Q. Ma and P. Steenkiste, On path selection for traffic with bandwidth guarantees, in: *Proceedings of IEEE International Conference on Network Protocols*, Atlanta, GA, October 1997.

[20] E. Mannie et al., Generalized Multi-Protocol Label Switching (GMPLS) architecture, IETF RFC 3945, 2004.

[21] NSFNet, http://www.nsf.gov/about/history/nsf0050/internet/launch.htm.

[22] F. Palmieri, U. Fiore and S. Ricciardi, SPARK: a smart parametric online RWA algorithm, *Journal of Communications and Networks* **9**(4) (2007), 368–376.

[23] R. Ramaswami and K.N. Sivarajan, Design of logical topologies for wavelength-routed optical networks, *IEEE Journal on Selected Areas in Communications* **14** (1996), 840–851.

[24] S.D. Rao and C.S.R. Murthy, Distributed dynamic QoS-aware routing in WDM optical networks, *Computer Networks* **48**(4) (2005), 585–604.

[25] S. Suri, M. Waldvogel and P.R. Warkhede, Profile-based routing: A new framework for mpls traffic engineering, in: *Quality of Future Internet Services*, Lecture Notes in Computer Science, Vol. 2156, Springer-Verlag, 2001.

[26] S. Uhlig, B. Quoitin, J. Lepropre and S. Balon, Providing public intradomain traffic matrices to the research community, *ACM SIGCOMM Computer Communication Review* **36**(1) (2006), 83–86.

[27] Z. Zhang, K. Long and S. Cheng, Load balancing algorithm in MPLS traffic engineering, in: *IEEE Workshop on High Performance Switching and Routing*, 2001.

[28] D. Zhemin and M. Hamdi, Integrated routing and grooming in GMPLS-based optical networks, in: *ICC*, 2004.

Computer Communications 33 (2010) 1809–1822

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

A GRASP-based network re-optimization strategy for improving RWA in multi-constrained optical transport infrastructures

Francesco Palmieri [a,*], Ugo Fiore [a], Sergio Ricciardi [b]

[a] Università degli Studi di Napoli Federico II, CSI, Complesso Universitario Monte S. Angelo, Via Cinthia, 80126 Napoli, Italy
[b] Universitat Politècnica de Catalunya (UPC), Departament d'Arquitectura de Computadors (DAC), Jordi Girona 3, 08034 Barcelona, Catalunya, Spain

ARTICLE INFO

Article history:
Received 14 December 2009
Received in revised form 25 April 2010
Accepted 7 May 2010
Available online 13 May 2010

Keywords:
Network Re-optimization
GRASP meta-heuristic
Wavelength-routing
RWA
Grooming

ABSTRACT

In wavelength-routed optical networks, end-to-end connection demands are dynamically routed according to the current network status. Naïve path selection schemes, the wavelength continuity constraint and the limited or inaccurate information available can cause the virtual topology resulting from the currently allocated lightpaths to become sub-optimal. We propose an efficient re-optimization technique based on a GRASP meta-heuristic. Our work is focused on a hybrid online–offline scenario: connections are ordinarily routed dynamically using one of the available algorithms for online routing, but occasionally, when reorganization of the current virtual topology is desirable, existing paths are re-routed in order to improve load balancing and hence the ability to efficiently accept further connections. Because global changes of the logical topology and/or routing scheme can be disruptive for the provided connection services, we used iterative stepwise approaches based on a sequence of small actions (i.e., single connection re-routing and on local search from a given configuration). Simulation results demonstrate that several network performance metrics – including connection blocking ratios and bandwidth gains – are significantly improved by such approach. In particular, we achieved to accept more connection requests in our re-optimized networks with respect to the same networks without re-optimization, thus lowering the blocking ratio. Besides, in all tests we measured a notable gain in the number of freed bandwidth OC-units thanks to our re-optimization approach.

1. Introduction

While being attractive for their transparent and cost-efficient operation, all-optical networks need complex routing practices and accurate engineering of Wavelength Division Multiplexed (WDM) paths, to match the constraints of the underlying photonic technology with the requirements of the dynamic traffic flows that should be transported. Dynamic routing and wavelength assignment schemes commonly used within these infrastructures tend to lead to network inefficiencies due to the limited or inaccurate information available for online routing [1], to the simple path selection algorithms often used and to the wavelength continuity constraint [2]. Precisely, the paths for the arriving connection requests are calculated starting from the current network state, including all the already routed connections. As the network and traffic evolve, such routing solutions may become sub-optimal [2]. The evolution process may also lead to changes in network topology due to the addition/deletion of new links and/or changes resulting from customers varying demands for different services. Some connection requests may be rejected due to lack of capacity, while a more efficient routing scheme would have allowed successful routing and path set-up. Furthermore, dynamic online routing practices typically tend to unbalance resource usage over time, causing severe congestion on some "critical" links that are most likely needed for satisfying future traffic demands [3]. When this happens, routing optimality may be restored only by periodic offline re-optimization that re-routes some of the existing connections over alternative paths, recovering the stranded capacity and re-balancing the load on the links [2]. Nevertheless, there is a cost directly associated with re-optimization, both in terms of computational complexity and of disruption of active connections [4]. Thus, the re-optimization activity has to be prudently planned to maximize the recovered stranded capacity and an integrated approach for periodic offline re-optimization of optical networks with sub-wavelength traffic is desirable. The focus of this work has been the best balancing of the reconfiguration cost, in terms of both computational complexity and disturbance to the network, within the context of a flexible and effective RWA/grooming solution. To allow a joint consideration of routing and wavelength assignment (RWA) with grooming and reconfiguration/optimization costs, we modeled the problem as a multi-objective optimization problem and solved it heuristically through a greedy randomized adaptive

[*] Corresponding author. Tel.: +39 081676632; fax: +39 081676628.
E-mail addresses: francesco.palmieri@unina.it (F. Palmieri), ugo.fiore@unina.it (U. Fiore), sergio.ricciardi@ac.upc.edu (S. Ricciardi).

0140-3664/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2010.05.003

search procedure (GRASP) [5,6] in conjunction with a path-relinking [7,8] solution refinement procedure. In our approach, the implementation of the GRASP-based re-optimization works at the control-plane layer on each involved network element and includes several novel greedy construction and local search strategies, and a new simplified form of path-relinking. Overall, the heuristic approach is streamlined through the incorporation of advanced network flow re-optimization techniques and is based on a totally flexible network model, supporting heterogeneous WDM equipment, in which the number and type of lambdas can be independently specified for each link. We evaluate the effectiveness of our approach by simulating the proposed re-optimization schema and measuring the freed bandwidth and the percentage of the connections that the network is able to route after the re-optimization process. Results indicate that this implementation may lead to significant improvements of the network in comparison with the existing dynamic RWA solution with an acceptable performance impact due to offline re-optimization. An especially appealing characteristic of this GRASP-based approach–that makes it particularly suitable for the re-optimization of large networks and preferable to other heuristics–is its straightforward implementation. A limited number of parameters need to be assigned and tuned, and consequently development can easily focus on implementing efficient data structures to speed GRASP iterations up. Finally, the GRASP solution-search strategy can be trivially implemented in parallel between the available network nodes. Each processor node has to be initialized with its own instance data and an independent sequence of random numbers needed by the GRASP procedure. All the GRASP iterations are then handled in parallel by using only a single shared global variable, required to store the best solution found over all processors, thus greatly reducing computational complexity and signaling overhead.

2. Background

This section briefly introduces some of the concepts that will be useful to better explain the proposed integrated dynamic RWA/re-optimization paradigm, by presenting the underlying architectural scenario, the basic building blocks, assumptions and modeling details together with the theory behind it.

2.1. Wavelength-routed optical networks

With WDM, a single optical fiber is shared by a number of independent wavelengths (channels), each of which may transparently carry signals in different formats and bit-rates, for example STM-16 and 10G-Ethernet. Over the physical topology composed of optical cross-connection (OXC) devices connected by fiber links, a quasi-static *virtual topology*is superimposed by interconnecting pairs of edge nodes with *lightpaths*, all-optical channels that are never converted into an electrical signal at intermediate nodes across the optical backbone. Edge nodes transform the optical signal in electrical form and route it on client subnetworks. More sophisticated, and costly, OXCs, besides switching specific wavelengths between ports, can also convert input wavelengths into different output wavelengths. Ordinarily, therefore, a lightpath uses the same wavelength on all the links along its route.

2.2. The routing and wavelength assignment problem

Every lightpath must be routed on the physical topology and assigned a wavelength: this process is called routing and wavelength assignment (RWA). In general, the RWA problem is characterized by two constraints specific of an optical network:

- the wavelength continuity constraint, i.e. a lightpath must use the same wavelength on all the links along its route;
- the wavelength clash constraint, i.e. two or more lightpaths using the same fiber link must be allocated distinct wavelengths.

The wavelength continuity constraint may be relaxed if OXCs are equipped with wavelength converters [13]. Different levels of wavelength conversion capability (full or limited) are possible, depending on the number of converter-equipped OXCs and to the number of wavelengths that can be converted in each node. Note that when using full wavelength conversion on each network node, the RWA problem reduces to the classical routing problem in a circuit-switched network. This work is quite general in that we make no assumption on the availability of wavelength converters.

With *static* traffic [9], the entire set of connection requests is known in advance, and so the problem is reduced to setting up permanent lightpaths while minimizing the number of wavelengths or the number of fibers. The RWA problem for static traffic can be formulated as a mixed-integer linear program [10], which is NP-complete [11]. The two sub-problems of routing and wavelength assignment can also be separately faced. A review of these approaches is given in [9]. Here, *Incremental* connection requests arrive in sequence and the lightpaths established to handle these requests remain in the network indefinitely. On the other hand, for *dynamic* traffic, a lightpath is set-up to satisfy each request as it arrives, and such lightpath is released after a finite amount of time (connection lifetime). The objective in both the incremental and dynamic traffic models is to minimize the blocking/rejection probability of each connection (also known as blocking factor), or equivalently maximize the number of connections that are established in the network at any time. The dynamic case is more complex and usually several properly crafted heuristics are used to solve the routing and wavelength assignment sub-problems separately [12]. Here, we have chosen to deal with incremental traffic, as it is a simplified environment in which the effects or the proposed strategy are less dependent on the statistical distribution of connection requests and can therefore be better evaluated.

2.3. Integrating RWA with grooming: from the overlay model to a unified control-plane architecture

Typically, the traffic demand is partitioned into multiple parallel requests (between the same node pairs) with different bandwidth requirements, varying from tens or hundreds of Mbps (e.g. STM-1 or Fast Ethernet) up to the full-wavelength capacity (e.g. 10 Gigabit Ethernet). At the network edge, end-to-end connection requests, sharing the same traffic flow characteristics in terms of termination nodes and Quality of Service (QoS) requirements and involving capacities significantly lower than those of the underlying wavelength channels, can be efficiently multiplexed, or “groomed,” onto the same wavelength/lightpath channel. A typical control-plane paradigm for traffic grooming operates on a two-layer multiple model, i.e. an underlying pure optical wavelength-routed network and an independent “opto-electronic” time-division multiplexed layer built over it. At the optical layer, wavelength routing traditionally sets up an almost static logical topology that is then used at the IP layer for routing, with lightpaths handled as single IP hops. By integrated RWA we instead mean a combined wavelength routing and grooming optimization paradigm, taking into account the whole topology and resource usage information at both layers. We assume that appropriate protocols exist for the unified control-plane, accurately disseminating correct and up-to-date information about the network state to each node, as well as taking care of resource reservation, allocation, and release. Reconfiguration of virtual topology may be carried out for two reasons:

- to re-optimize the virtual topology under a changed traffic pattern or even a different cost metric;
- to create a new topology capable of supporting the current traffic pattern, without using failed or out-of-service network components.

In this work, traffic grooming is accounted by considering each flow as reallocatable. We focus on reconfiguration for re-optimization, when the traffic pattern changes or some part of the network becomes congested. Some of the most relevant issues involved in optical network re-optimization are discussed below.

2.4. Network re-optimization

Sophisticated routing algorithms can keep achieve remarkably low connection reject rates. However, these algorithms do not scale well with the growth in network size. On the other hand, many of the simple and scalable path selection algorithms may cause routing inefficiencies, leading to "stranded" capacity [14]. Whatever the RWA algorithm, the resources dedicated to serve each new connection request are selected according to the current network state, which is, in turn, the result of routing the existing connections. Keeping the network load balanced may lead to effective algorithms achieving good results in terms of blocking probability. This is necessarily the result of some estimation on the distributions of forthcoming requests. That estimation may presume that future requests will adhere to a uniform or Poisson distribution, or that they will repeat the pattern delineated from the currently provisioned demands. However, as the network and traffic evolve, the actual distribution of requests and their sequence of appearance may substantially drift from the estimates, and the network load distribution may become unbalanced. Network re-optimization is usually needed to increase the network utilization and can be performed by re-scheduling the already available connections requests in two ways: either by changing the associated paths only or by changing both the paths and the starting times. The latter solution is not very desirable as it implies re-negotiating the connection set-up times with users, and for this reason we will not consider it in this work. The idea of re-optimization is not new in telecommunications: carriers routinely use reconfiguration to better manage their network and increase utilization, which in turn allows them to defer investments on new infrastructure. Reconfiguration can also be used to provide better service performance, for example, by re-routing services over shortest paths if such paths become available.

2.5. GRASP

GRASP, which first appeared in [5] and was later formalized in [6], is an iterative two-phase meta-heuristic. A meta-heuristic may be defined as an iterative master process that guides and adjusts the operation of subordinate heuristics in order to produce high-quality solutions. When exploring the solution space, sometimes one may get stuck into local optima, i.e. solutions that are good locally but not globally. Meta-heuristics strive to escape such local optima by different strategies: occasionally accept worse solutions, as in simulated annealing [15] or tabu search [8]; combine existing solutions through mutation and crossover following the idea of genetic algorithms [16]; generate new solutions, as in GRASP. Greedy choices are performed and measured by means of an immediate or greedy gain possibly leading to sub-optimal solutions. For overcoming this myopic behavior, a heuristic measure can be introduced to evaluate this gain. At each iteration, the first phase produces a solution through the use of a greedy randomized adaptive construction scheme. In the second phase, local search is applied to this solution, in order to obtain a local optimum in its neighborhood. GRASP meta-heuristic may be customized to solve any problem for which simple construction and local search algorithms are available. Enhanced versions of the basic GRASP meta-heuristic have been applied to a wide range of combinatorial optimization problems [17]. Several new components and techniques have extended the original GRASP scheme (reactive GRASP, parameter variations, bias functions, memory and learning, improved local search, path-relinking, hybrids). These components are presented and discussed in [17]. In particular, path relinking was first introduced as a tool to compound intensification and diversification strategies in the context of tabu search [7,18]. The strategy is formulated on the principles of evolutionary approaches but unlike conventional evolutionary techniques (e.g., genetic algorithms), it does not employ randomization to generate new solutions. Instead, it constructs them through a methodical exploration of trajectories that connect previously generated high-quality solutions. An in-depth description of path-relinking can be found in [19,20]. The first application of GRASP and path-relinking was undertaken by Laguna and Marti [21]. Since then, a few other applications have appeared that combine the two methodologies. Some applications use path-relinking as an intensification strategy within the GRASP procedure [22]; others apply path-relinking as a post-optimization step after the execution of GRASP [23]. Some authors considered both utilizations of the path-relinking strategy [24,25]. According to the survey work on GRASP by Resende and Ribeiro [17], path relinking is more effective when used as an intensification phase. In our implementation, we have chosen to use it at the end of each GRASP iteration in order to intensify the search around local optima.

2.6. Related work

Lightpath re-optimization techniques have been discussed in several works available in literature. The problem of re-routing existing lightpaths in a dynamic routing scenario was addressed in [26,2] by invoking the re-optimization step only when new requests are unable to find a feasible path and it becomes absolutely necessary to re-route some of the existing paths to free up capacity from the most crowded links. Alternatively, [4] models the effect of the reconfiguration phase in terms of packet loss and bases its reconfiguration policy on this penalty criterion. Some different approaches, such as [27,28] reconfigure the underlying virtual topology of the optical network, respectively according to an ILP optimization and a stepwise branch exchange process, to adapt it to changing traffic patterns. Other authors have proposed solutions suitable for static traffic demand and heuristics for long-term on-demand traffic flows [29–32]. In [31] the authors study the problem of re-optimizing lightpaths in resilient mesh optical networks, where connection requests are routed using a pair of disjoint primary and backup paths. In their re-optimization scheme, all paths are re-routed, regardless of their being primary or backup. They also considered the effects of re-routing only the backup paths. An approach to combine dynamic online routing in a connection-oriented network with an offline optimization module, which constantly rebalances the load in the network whenever a certain imbalance threshold is exceeded, has been examined in [32]. In this scenario, the network operator determines a re-balancing benefit indicating the amount of traffic that could additionally be routed if the current traffic were to be redistributed, by computing the gain in "network efficiency" that a potential re-optimization would yield. If a threshold is exceeded, i.e. the benefits of re-optimizing the network exceed the incurred costs of the flow re-routing, then re-optimization is performed. More recently, other formulations of the rearrangement problem have been proposed, differing in the optimization objectives. Notably, in [33], the number of rejected new demands and re-routed lightpaths is minimized through the Lagrangean Relaxation and Subgradient Method, while Din [34] investigated the use of a genetic algorithm and of simulated annealing with the objective of minimizing the average weighted propagation delay.

3. Integrated re-optimization scheme for wavelength-routed networks

We propose a novel dynamic RWA strategy specifically conceived to allow lightwave networks to carry more traffic without adding capacity, through a two-stage scheme based on hybrid online routing and offline re-optimization. Online dynamic routing is used in the first phase: connection requests arriving to the edge nodes over time are immediately routed by using a quick RWA scheme such as min-hop or shortest/minimum-cost path with dynamic weights based on wavelength available capacity. If there are enough resources to accommodate it, the required connection is routed on an already existing lightpath with available capacity and adequate QoS characteristics, or a new lightpath is properly set-up; otherwise the request is rejected. Over time, requests for teardown of existing connections may also arrive, causing the release of the involved resources. However, at a certain time, the residual capacity between certain critical ingress–egress pairs may be insufficient to accommodate new requests, but a different allocation of connection routes would easily permit it. When this happens, the blocking factor may increase indefinitely so that the network seems to be completely saturated even if there are still a lot of available resources. Thus, we continuously monitor the blocking factor, that can be viewed as a good approximated measure of routing efficiency, and when it exceeds a specific threshold β, we invoke re-optimization to restore routing optimality by re-routing some of the already established connections. Alternatively, we can also trigger the procedure when a specific number of connection requests has been received or served starting from the last occurrence of the re-optimization process. In both cases we address the network re-optimization issue as a periodic maintenance measure, activated when a tunable utilization threshold is exceeded, aiming at continuously keeping as much free network resources as possible with minimum total disruptions to the ongoing service. The objective of the second phase of our approach is then to eliminate the blocking or unbalancing generated during the previous quick-and-dirty connection set-up phase. Managing the lightwave network during the reconfiguration phase is a very complex issue, as re-optimization involves path reorganization, which may originate disruption in some critical services carried over the network, and therefore must be implemented carefully. Hence, to keep re-optimization as efficient as possible, all the connection requests arriving during the re-optimization process are queued and served only after its completion. Furthermore, the chosen re-optimization strategy must be conceived in order to combine maximum gains in recovering stranded capacity with minimal impacts on the overall network performance. Re-optimization must support the ability to provide guaranteed fault-tolerance to resilient connections. Finally, in order to preserve the packet arrival sequencing, the re-optimization strategy should not require traffic flow splitting on multiple paths. The sequence of operations through which the virtual topology is reconfigured, and the number of connections/lightpaths affected by such activity, can have a substantial impact on both the performance and capacity that is needed in the process and on the optimality of the obtained solution. The corresponding minimization problem, known as the Reconfiguration Sequencing Problem is indeed NP-hard [35]. Thus a re-optimization solution that re-routes all the connections of the existing virtual topology (without disruption) while keeping the network well balanced, by redistributing load thus freeing sufficient available capacity between all the ingress–egress node pairs, has to be found through the use of some heuristic technique that must ensure an acceptable run-time complexity. A key feature of such heuristic must be the ability of setting up the new re-provisioned paths one-by-one *before* re-routing traffic on them and only releasing the resources on the old paths *after* the new ones are totally established, according to the "make-before-break" principle. In other words, we do not explicitly perform re-routing on predefined backup paths or support specific post-optimization restoration strategies but we propose a new heuristic-based strategy, to be triggered on a maintenance basis, for finding approximate solutions to this problem, starting from a simple greedy approach and improving the quality of the re-optimization performance by using local search, through a combination of GRASP and path-relinking.

3.1. The network model

We denote the network by a graph $G = (V,E)$ where V is the set of nodes and E the set of links. We make no specific assumption on the number of wavelengths per fiber, number of fiber on each link and on the presence of wavelength conversion devices on the network. All these parameters are fully and independently configurable at the network topology definition time. Instead, we require that all the network nodes operate under a unique control-plane and share a common network view by relying on a common link-state protocol that is used to distribute resource usage information. Furthermore, we assume that every connection is bi-directional and consists in a specific set of traffic flows that cannot be split between multiple paths. Each connection can be routed on one or more (possibly chained) existing lightpaths between its source and destination nodes, with sufficient available capacity or on a new lightpath dynamically built on the network upon the existing optical links. Grooming decisions are taken according to adaptive strategy that dynamically tries to fulfill the algorithm's network resource utilization and connection serviceability objectives by determining if the request can be routed on one of the available lightpaths, by time-division multiplexing it together with other already established connections, or, if there are no available resources to satisfy the request, a new lightpath is needed on the optical transport infrastructure. A network with m edge nodes supports bi-directional connection demands only between $m(m-1)/2$ source-sink node pairs (u,v) where source and sink nodes $u,v \in V$ are edge routers. These source-sink pairs can be numbered from 1 to M and for each source-sink pair (u,v) there may be an amount $d(u,v)$ of end-to-end bandwidth demand already provisioned in the network, measured by the aggregate bandwidth of all the connection flows between the source and sink pair. To simplify our model each connection request is only characterized by a QoS commitment on bandwidth, although it can be routed basing the decision on other QoS metrics such as limited latency, error rate, etc. that can be incorporated into Service Level Agreements by converting them into a bandwidth requirement as shown in [36]. In addition we denote by $D(u,v)$ the total desired demand for the source sink pair (u,v). For each link $e \in E$ in the network rb_e and mb_e denote respectively its current residual and total capacity.

Let P be the number of connection requests at re-optimization time, $c_k = (u_k, v_k, b_k)$, $k = 1,\ldots,P$, the generic kth connection request, where $u_k, v_k \in V$ are respectively the origin and destination and b_k the bandwidth required, and p_k the path servicing the connection c_k. A feasible solution to our RWA re-optimization problem is then the set $X = \{X_1,\ldots,X_P\}$, where the generic element X_k is a pair (c_k, p_k) describing the routing choice associated to each connection. The actual routing p_k of a connection c_k is determined by means of a shortest-path computation, with a cost function that only depends on the residual bandwidth on the links. Therefore, routing of a connection only depends on the network state at the moment the connection is considered for routing. Note that, if all the connection requests are serviced sequentially starting from an empty network, the order of arrival uniquely determines the solution.

3.2. Grasp-based re-optimization

In order to find good approximate solutions to the above multi-objective optimization problem, we propose a methodology based on the combined use of GRASP and path-relinking. When implementing a GRASP procedure, several different issues need to be addressed and tailored to the structural characteristics of the problem under study. First, an adaptive greedy function needs to be defined to guide the iterative construction phase, which builds the solution by adding one element at the time. The greedy function is adaptive in the sense that its value must be updated after the insertion of each new element in the partial solution under construction in order to reflect the choice made. Second, a restriction mechanism must be defined to build the restricted candidate list (RCL), that is the list from which to select the next element to be added to the solution. A probabilistic selection strategy (random component) must then be specified to select an element from the RCL. Besides, the essential constituents of the local search procedure (i.e. the neighborhood structure N, the search strategy and the objective function) must be defined. Finally, the objective function for the optimization problem must be defined. The objective function may be aimed at minimizing or maximizing some quantities in order to optimize the problem resolution. We use a minimizing function, i.e. a function whose values must be kept as low as possible while respecting the problem constraints. The whole GRASP procedure is algorithmically sketched in Fig. 1.

Where $f : \mathscr{F} \rightarrow R$ is the objective function of a specific problem $\mathscr{P}$, mapping the set $\mathscr{F}$ of feasible solutions to real values in R. The neighborhood structure $\mathscr{N}$ relates a solution X of the problem to a subset of solutions $\mathscr{N}(X) \in \mathscr{F}$. The procedure consists of *MaxIter* iterations (lines 2–8) in which a new solution is built (line 3), its neighborhood is explored (line 4) and the objective function is evaluated on it looking for an improvement of the current best solution (lines 5–7). The construction phase (line 3) tries to build a new solution X' choosing randomly an element from the RCL. The local search explore the neighborhood $N(X')$ of the construction phase solution X' looking for a local optimum X'' such that $f(X'') \leqslant f(X')$. At the end of each step we compare the value of the objective function*f* evaluated on the solution X" with X^* which is the best solution found till that moment and we eventually keep the better one as the best solution found; if the algorithm has achieved a local optimum X^* such that $f(X^*) \leqslant f(X)$ for all $X \in N(X^*)$, the best solution is updated with the new value. Finally, the best solution X^* found in all iterations is returned as the overall GRASP solution. GRASP may be also viewed as a repetitive sampling technique in which each iteration produces a sample solution taken from an unknown distribution of admissible ones, whose mean and variance depend on the restrictive nature of the RCL. Given an effective greedy function, the mean solution value is expected to be good, but probably sub-optimal. That is, if the RCL is restricted to a single element only, then the same solution will be always produced on all the iterations. Clearly, in this case, the variance will be zero and the mean will exactly match with the value of the greedy solution. If we impose a less restrictive limit on solutions cardinality, i.e. more elements are allowed in the RCL, then many different solutions will be produced, with a larger variance. The size of the RCL controls, then, the tradeoff between the randomness and greediness of the solution. Hence, the value of the parameter α, which regulates the RCL size as explained in the section below, has to be chosen carefully. The lesser the role of greediness as compared to randomness, the worse should the optimality of the average solution be. However, the best solution found outperforms the average and very often is optimal.

```
procedure GRASP($X$, $MaxIter$, $\alpha$, $\mathcal{N}$)
Input:  current network state $X$
        maximum number of GRASP iterations $MaxIter$
        RCL parameter $\alpha$
        neighborhood function $\mathcal{N}$
Output: new network state $X^*$
1.  $X^* \leftarrow \emptyset$
2.  for $i = 1$ to $MaxIter$ do
3.      $X' \leftarrow$ BuildSolution($X$, $\alpha$)
4.      $X'' \leftarrow$ LocalSearch($X'$, $\mathcal{N}$)
5.      if $f(X'') \le f(X^*)$ then
6.          $X^* \leftarrow X'$
7.      end
8.  end
9.  return $X^*$
10. end procedure GRASP
```

Fig. 1. A generic GRASP algorithm.

3.3. The construction phase

In the construction phase, connections are routed one at a time, thus building the solution. The pseudo-code of the Greedy (lines 2 and 4) randomized (line 5) adaptive (line 7) search procedure is illustrated in Fig. 2. First, we sort the connection requests in non-increasing greediness into a list L according to the greedy criterion (line 2); then we start building the solution adding one connection request at a time till the whole candidates are routed (lines 3–8). At each iteration, the list L is restricted into the RCL containing only the first k elements of L (line 4) and a new connection request is randomly selected from the RCL (line 5) and routed in the network (line 6). To drastically reduce the computation times, we combined the strategies commonly used by GRASP and heuristic-biased stochastic sampling [22]. At each iteration, the list is reordered taking into account the choice made at the previous step and the RCL is formed again (line 7). The greedy criterion consists in assigning a highest greediness to the un-routed origin–destination pairs whose source and destination nodes have the largest residual bandwidths on their incident arcs together with a high value of the bandwidth required for the connection. Such strategy allows the requests between nodes that have most residual capacity and that have higher bandwidth demands to be served first.

In detail, we start from an empty solution vector. The P connection requests $c = (u, v, b)$ are ordered according to the greedy adaptive criterion Γ. For each node v, let us denote by $\delta(v)$ the cut separating v from the rest of the graph, i.e. the set of all incident arcs to v

$$\delta(v) = \{u \in V | [u, v] \in E\} \tag{1}$$

where $[u, v]$ denotes an un-directed arc in the graph G and by

$$\gamma(v) = \sum_{a \in \delta(v)} rb_a \tag{2}$$

the sum of the residual capacities rb_a over all the arcs a incident to v. The ordering criterion for a connection $c = (u, v, b)$ will be based on the value:

$$\Gamma(c) = \gamma(u) + \gamma(v) + b \tag{3}$$

The bandwidth term b has the purpose of prioritizing demands that have higher bandwidth requirements and letting smaller ones to be served later as they are easier to be routed. Note that the criterion is adaptive: the sorted list L may be rearranged as a consequence of

```
procedure BuildSolution($X$, $\alpha$)
Input:   current network state $X$
         RCL parameter $\alpha$
Output:  new candidate solution $X$
$X \leftarrow \emptyset$
Order the candidate connection requests $c = (u, v, b)$ in the list $L$ according to the greedy
criterion $\Gamma(c) = \gamma(u) + \gamma(v) + b$
while $L \neq \emptyset$ do
    $RCL \leftarrow$ MakeRCL($\alpha$)
    $v \leftarrow$ RandomSelect($RCL$)
    $X \leftarrow X \cup \{v\}$
    Reorder candidates reflecting the choice made
end
return $X$
end procedure BuildSolution
```

Fig. 2. The solution construction algorithm.

the successfully routing of the chosen connection requestc. In fact, when the requestc is eventually routed – using the same cost function and routing algorithm of the previous phase – the residual bandwidth will decrease by b on all the involved arcs and the values computed by the greedy function will change reflecting the *new* available bandwidth on each arc along the path. For each connection request that is chosen from the list and routed, the greedy value has to be recomputed only on the connections whose extremes are involved in the routes of the previous connection. The Restricted Candidate List *RCL* is then built, selecting only the first $k$elements in the ordered list L, where k is determined by the value of a tuning parameter $\alpha \in [0,1]$, according to the following formula:

$$k = (1 - \alpha) + \alpha \cdot |L| \tag{4}$$

where $|L|$ is the (whole) candidate list size. Among all the elements in the RCL, one is randomly chosen to become the next component of the solution being constructed. This process is iterated until the vector is complete. Note that, as α makes k vary proportionally in $\{1,\ldots,|L|\}$, it allows us to control the amount of greediness and randomness in the choice of the next element to be added in the solution under construction. In particular, when $\alpha = 0$, $k = 1$ corresponding to a completely greedy choice. On the other side, when $\alpha = 1$, $k = |L|$ so that the whole list is selected and the choice is totally random. The randomness factor allows widely different solutions to be constructed at each GRASP iteration, helping us to avoid being trapped into local maxima in the solution space. On the other side such selection strategy does not necessarily compromise the effectiveness of the adaptive greedy component of the method, as only the best-rated elements (on top of the list) can be chosen. It may happen that the construction phase fails, i.e., a state is reached where the current connection cannot be routed. In this case, the construction phase is restarted from scratch in the next iteration.

3.4. Performing local search on the solution space

Since the solutions generated by a GRASP process are not guaranteed to be locally optimal, it is almost always beneficial to attempt at improving each constructed solution by means of a local search in the solution space. A local search algorithm operates according to an iterative scheme, sequentially replacing the current solution with a better one found in the neighborhood of the current solution. The algorithm terminates when no better solution can be found in the neighborhood. A solution X is said to be locally optimal if in its neighborhood $\mathcal{N}(X)$ there are no solutions better than X. A significant limitation of local search is the risk of getting trapped into local optima. To circumvent this limitation, local search has to be driven by general-purpose heuristic strategies aiming at avoiding this phenomenon. The key success factors for such local search strategies are a good starting solution, the suitable choice of a neighborhood structure, and an efficient neighborhood search technique. It should be noted that each solution built in the previous phase might be viewed as a set of routes, one for each end-to-end connection request, where a single lightpath on the optical network must support one or more routes and a single route must use one or more lightpaths. The operation of constructing the solution neighborhood can be expressed as the construction of new "neighbor" network states resulting from the removal and re-routing of a single connection request at a time. For each neighbor, one connection c_i is selected, the resources allocated to c_i are released, and c_i is routed again, possibly along a different path (as the current network state is different from the one at which c_i was originally routed). This strategy ensures a low complexity for the neighborhood generation operation.

The network state is defined by the vector $\vec{S} = (rb_1, \ldots, rb_m)$ of the residual bandwidth on each of the m links. We can represent the route associated to a connection $c = (u,v,b)$ in the network state $\vec{S}$ by the *route* function:

$$route(c, \vec{S}) = (b_1^{(c)}, \ldots, b_m^{(c)}) \tag{5}$$

where $b_i^{(c)}$ is the bandwidth requested by the connection request c on the link i. Note that $b_i^{(c)}$ will be equal to the requested bandwidth b on the links along the route assigned to c and 0 on the other links. We then define a generic allocation function for a connection request c and a state $\vec{S}$,

$$allocate(c, \vec{S}) = \vec{S} - route(c, \vec{S}) \tag{6}$$

Let us consider a fixed order of arrival of the connection requests $c_1 \ldots, c_j$. The progressive routing of the succession of connections leads to a sequence of states. The initial state $\vec{S}_0$ corresponds to the starting condition in which every link is unused (empty network), so that:

$$\vec{S}_0 := (mb_1, \cdots, mb_m) \text{ where } mb_i \text{ is the maximum (initial) bandwidth for each link } i \tag{7}$$

and the generic state $\vec{S}_i$ is given by

$$\vec{S}_i := \vec{S}_{i-1} - route(c_i, \vec{S}_{i-1}) \tag{8}$$

or, equivalently, by

$$\vec{S}_i = \vec{S}_0 - \sum_{k=1}^{i} route(c_k, \vec{S}_{k-1}) \tag{9}$$

The *release* function is the complementary operation to the *allocate* function (6); let us first see what happens if we revert the last allocation in the sequence:

$$release(c_i, \vec{S}_i) = \vec{S}_{i-1} = \vec{S}_i + route(c_i, \vec{S}_{i-1}) \quad \text{by Eq.}(8) \tag{10}$$

In the general case,

$$release(c_i, \vec{S}_P) = \vec{S}_P + route(c_i, \vec{S}_{i-1}) \tag{11}$$

Therefore, we generate a new state with the following sequence:

$$\begin{aligned} \vec{S}_P &= allocate(c_i, release(c_i, \vec{S}_P)) \\ &= release(c_i, \vec{S}_P) - route(c_i, release(c_i, \vec{S}_P)) \end{aligned} \tag{12}$$

We propose two local search procedures: breadth and depth local search. In order to build the solution neighborhood to be explored by breadth local search, we work as follows:

- for each connection request c_k, we start by removing the corresponding units of flow from each edge in its current route p_k;
- next, we calculate the resulting network status by determining the new edge weights. A tentative new shortest path route for the connection c is then computed by using the resulting weights.

Thus, the breadth local search process consists in releasing and reallocating every connection in the network and evaluating the objective function on the new network configurations obtained, keeping from time to time the best solution found. When every connection has been processed, the local search process ends and the best solution found is returned. In the depth local search strategy, whenever a better solution X_b is found in the neighborhood of the previous solution X, the local search process is repeated starting from X_b instead of X; the process is iterated for each connection request until all the candidates have been processed. The local search procedure pseudo-code is illustrated in Fig. 3. Each connection request $c = (u, v, b)$ (line 2) is extracted one at a time from the solution X (line 3) obtained by the construction phase and eventually re-routed in the network (line 4). Then, the objective function is evaluated on the new solution X' (line 5) and the best solution X^* is eventually updated with the new value (line 6). If the depth local search is chosen, the next iteration will start its local search from the new solution X'; otherwise, a breadth local search will be performed, starting from X.

3.5. The objective function

The ultimate objective of the offline re-optimization problem is to minimize the lightpath rejection or delayed creation (due to the duration of the re-optimization process) while balancing the load on the optical links (and hence maximizing network resources utilization). A good balancing can be achieved by minimizing the load on the most utilized fiber trunks. Of course, routing and wavelength assignments with minimum delays may not attain the best load balance within the network and, likewise, RWA algorithm realizing the best load balance may not minimize creation delays or connection rejection at all. We essentially aim at routing the connections in such a way that a desired point in the tradeoff curve between creation delays and network load balancing is achieved. Hence, the objective function we use to compare the solutions found in the previous phase should provide an extremely effective metric for evaluating the degree of network resource load balancing together with an acceptable run-time performance, to avoid, as much as possible, the excessive delay of queued requests that may arrive during the re-optimization phase. The chosen objective function value clearly determines how the virtual topology is best suited for the given traffic demand. When the traffic pattern changes the network state may not remain optimal and the virtual topology needs to be changed to reflect the objective function goals. This change requires reconfiguration of the network components (OXCs and routers) to establish the lightpaths present in the desired new virtual topology but absent in the current one. Similarly, the lightpaths that are not present in the new virtual topology must be torn down. Obviously, such reconfiguration has an operational cost that cannot be ignored. Thus, the best reconfiguration solution is a tradeoff between the improvement in the objective function value and the number of changes to the virtual topology needed to achieve that improvement. A natural choice for our objective function f is the variance of the load vs. capacity ratios for each link (with the minus sign accounting for the fact that we are trying to balance the load as evenly as possible):

$$f(\vec{S}) = Var\left(1 - \frac{rb_i}{mb_i}\right), \quad i \in \{1, \ldots, |E|\} \tag{13}$$

We can note that such choice reflects both effectiveness in describing load balancing and ease of computation so that it contributes to keep the re-optimization delay for pending connections as low as possible. The structure of the objective function is such that as the load balance of the network increases, its maximum utilization rate decreases, providing a useful strategy to achieve the QoS level defined by a desired maximum utilization rate. Finally, once a target solution has been chosen, the current allocation is transformed into the desired one by parallelly signaling (with the RSVP-TE) only the affected elements, so that minimal impact in terms of service disruption is achieved.

```
procedure LocalSearch($X$, $\mathcal{N}$, localSearchType)
Input:  the solution built by the construction phase $X$
        neighborhood function $\mathcal{N}$
        local search type parameter localSearchType
Output: new network state $X^*$
1.  $X^* \leftarrow \emptyset$
2.  for $i = 1$ to $|X|$ do
3.      $X' \leftarrow X \setminus \{c_i\}$
4.      $X' \leftarrow X' \cup \{c_i\}$
5.      if $f(X') \leq f(X^*)$ then
6.          $X^* \leftarrow X'$
7.          if localSearchType = "depthSearch" then
8.              $X \leftarrow X'$
9.          end
10.     end
11. end
12. return $X^*$
13. end procedure LocalSearch
```

Fig. 3. The local search algorithm.

3.6. Path-relinking

Path-relinking may be viewed as an elite selection strategy aiming at adding in new solutions only high quality attributes, by privileging these attributes in the determination of other solution that best improve (or least deteriorate) the initial one. It works on a population of already good solutions by properly combining them to obtain new (better) ones. Such new solutions are then generated by exploring the trajectories connecting high-quality solutions and the name path relinking is used because the involved solutions are linked by a series of transformations, performed during the search process, relinking previous points in ways not already obtained in previous search history. Both a source and a guide solution have to be selected in order to generate these paths in the solution space. We can also start from a set of guiding solutions (multiple parents) generating combinations of elite solutions that link the points in the solution space in several ways. During the process of linking a solution X (initial reference point) to a solution Y (desired or guiding point), a path is constructed by the greedy selection of re-routing actions with respect to the evaluation of the objective function. Simply stated, a transformation is selected if it locally maximizes the objective function value. The main objective of path-relinking is the incorporation of attributes belonging to the guiding solution (or solutions) while recording values of the objective function. The purpose of these actions is to obtain several improved solutions within the neighborhood of the already visited ones. The trajectory from X to Y is generated iteratively, by selecting the greedy X neighbor solution (we will call this solution by Z), from a set of all neighbors that decrease the distance from X to Y. This distance may be determined by calculating the number of differences between the solutions (X and Y). The procedure is restarted, making $X \leftarrow Z$, until the target solution Y is obtained. Path-relinking is applied to pairs of solutions (X_1,X_2), where X_1 is the locally optimal solution, obtained through local search, and X_2 is randomly chosen from a pool of at most *MaxElite* elite solutions found along the search process. Such pool is originally empty. Each locally optimal solution obtained during the local search can be considered as a candidate to be inserted into the above pool only if it is different (by at least one link utilization in one route) from every other solution already present in the pool. If the pool already contains *MaxElite* solutions and the candidate is better than the worst of them, then the former replaces the latter. Otherwise, if the pool is not full, the candidate solution is simply inserted into it. This procedure is iterated until no further change in the pool occurs. Such type of intensification can be done in a post-optimization phase (by using the final elite solutions pool), or periodically, during the optimization process (by using the current elite solutions set). When path-relinking is used in a post-optimization phase, the local search procedure is applied to each elite solution when no further occurs change in the elite set, since the solutions produced by path-relinking are not always local optima. All the local optima found during local search are candidates for insertion into the elite set. The entire post-optimization process is repeated until changes occur within the elite set. In detail, our path-relinking algorithm starts by computing the symmetric difference $d(X_1,X_2)$ between X_1 (the initial solution) and X_2 (the guiding solution), resulting in the set of re-routing actions, which should be applied to the first one to reach the other. Then, by starting from the initial solution, the best topology change still not performed is applied to the actual solution, until the guiding one is reached. The best solution found along this trajectory is also considered as a candidate for insertion in the elite pool. Since the neighborhood of the initial solution is more carefully explored than that of the guiding one, starting from the best of them gives to the algorithm a much better chance of investigating in more details the neighborhood of the most promising solution [22]. Path-relinking can be applied to a pure GRASP procedure in a straightforward manner, as it can be seen in the integrated GRASP/Path Relinking algorithm reported in Fig. 4.

```
procedure GRASPwithPathRlnk($X$, $MaxIter$, $\alpha$, $\mathcal{N}$)
Input:   current network state $X$
         maximum number of GRASP iterations $MaxIter$
         RCL parameter $\alpha$
         neighborhood function $\mathcal{N}$
Output:  new network state $X^*$
1.  $X^* \leftarrow \emptyset$
2.  $\mathcal{E} \leftarrow \emptyset$
3.  for $i = 1$ to $MaxIter$ do
4.      $X' \leftarrow$ BuildSolution($X$, $\alpha$)
5.      $X'' \leftarrow$ LocalSearch($X'$, $\mathcal{N}$)
6.      if $|\mathcal{E}| = MaxElite$ then
7.          $X'' \leftarrow$ PathRelinking($X''$, $\mathcal{E}$)
8.          UpdateElite($X''$, $\mathcal{E}$)
9.      else
10.         $\mathcal{E} \leftarrow \mathcal{E} \cup \{X''\}$
11.     end
12.     if $f(X'') \le f(X^*)$ then
13.         $X^* \leftarrow X''$
14.     end
15. end
16. return $X^*$
17. end procedure GRASPwithPathRlnk
```

Fig. 4. GRASP with path-relinking re-optimization algorithm.

First, the set of elite solutions $\mathcal{E}$ and the best solution X^* are initialized to the empty sets (lines 1–2) and $\mathcal{E}$ is built by including the solutions from the first *MaxElite* iterations (lines 9–10). After that a standard GRASP iteration produces a local optimal solution X'' (lines 3–5), the PathRelinking procedure is called (line 7). Then, a function UpdateElite is called (line 8) in which the elite pool is possibly updated. The solution returned from path-relinking is included in the elite pool if it is better than the best solution in $\mathcal{E}$ or if it better than the worst and is sufficiently different from all elite solutions [37]. Finally, the optimal solution is updated if necessary (lines 12–14).

4. Complexity analysis

Let's now examine the computational complexity of the above GRASP-based optimization framework. We consider a network with n nodes and up to $\lambda_{\max}$ wavelengths on each of the m fiber links in which P connection requests have already been routed on the existing lightpaths. First, we remark that the computational complexity associated to the re-routing of a single connection in the construction phase of each solution is given by the routing algorithm used, in our case the traditional Shortest Path First (SPF) algorithm ($O(m \cdot \lambda_{\max} \cdot \log n)$ by using a priority queue with a Fibonacci heap in the implementation of the Dijkstra algorithm).

The computational complexity of the GRASP procedure can be calculated by summing, for each iteration, the complexity of all its component procedures (in the following indicated as $\langle$*Procedurename*$\rangle_C$) and is given by

$$\begin{aligned}\mathrm{GraspProcedure}_C = O(&MaxIter \cdot (\mathrm{BuildSolution}_C + \mathrm{LocalSearch}_C \\ &+ \mathrm{ObjectiveFunction}_C))\end{aligned} \tag{14}$$

In the *BuildSolution* procedure, the initial sorting of the list L costs $O(P\log P)$, the *while* cycle repeats P times its body that consists of *MakeRCL* and *RandomSelect* that are constant time operations $O(1)$, a SPF routing $O(m \cdot \lambda_{\max} \cdot \log n)$ and a partial reordering that may be reduced to a simple optimized update with a worst case cost of $O(n \cdot \log n)$, because we know exactly what are the elements whose value may only decrease. So, the complexity of the *BuildSolution* procedure is given by

$$\begin{aligned}\mathrm{BuildSolution}_C &= O(\mathrm{Sorting}_C + P \cdot (\mathrm{MakeRCL}_C \\ &\qquad + \mathrm{RandomSelect}_C + \mathrm{Routing}_C + \mathrm{Update}_C)) \\ &= O(P \cdot \log P + P \cdot (2k + m \cdot \lambda_{\max} + n \cdot \log n \\ &\qquad + n \cdot \log n)) \\ &= O(P \cdot \log P + P \cdot (m \cdot \lambda_{\max} + n \cdot \log n))\end{aligned} \tag{15}$$

The *LocalSearch* procedure repeats P times its cycle that consists of a release operation, a route SPF algorithm $O(m \cdot \lambda_{\max} + n \cdot \log n)$ and one evaluation of the objective function. The release operation has to free the bandwidth on all the links crossed by the connection, that may be $n - 1$ in the worst case for a network without cycle paths, so it costs $O(n)$. The objective function has to evaluate the bandwidths on the network's edge, thus it costs $O(m \cdot \lambda_{\max})$ in the worst case. Thus, the computational cost of the *LocalSearch* procedure is

$$\begin{aligned}\text{LocalSearch}_C &= O(P\cdot(\text{Release}_C+\text{Route}_C\\&\quad+\text{ObjectiveFunction}_C))\\&=O(P\cdot(n+m\cdot\lambda_{\max}+n\cdot\log n+m\cdot\lambda_{\max}))\\&=O(P\cdot(m\cdot\lambda_{\max}+n\cdot\log n))\end{aligned}\tag{16}$$

Consequently, the overall *GRASP* procedure costs:

$$\begin{aligned}\text{GraspProcedure}_C &= O(\mathit{MaxIter}\cdot(\text{BuildSolution}_C\\&\quad+\text{LocalSearch}_C+\text{ObjectiveFunction}_C))\\&=O(\mathit{MaxIter}\cdot(P\cdot\log P+P\cdot(m\cdot\lambda_{\max}+n\\&\quad\cdot\log n)+P\cdot(m\cdot\lambda_{\max}+n\cdot\log n)+m\\&\quad\cdot\lambda_{\max}))\\&=O(\mathit{MaxIter}\cdot(P\cdot\log P+2P\cdot(m\cdot\lambda_{\max}+n\\&\quad\cdot\log n)+m\cdot\lambda_{\max}))\\&=O(\mathit{MaxIter}\cdot(P\cdot\log P+P\cdot m\cdot\lambda_{\max}+P\\&\quad\cdot n\cdot\log n))\end{aligned}\tag{17}$$

For typical topologies and traffic values (e.g. $n = 30$, $P = 3000$), $n\cdot\log n$ is the dominant factor with respect to log P as, even in the case in which P is much greater than n, the logarithm function will weight very little its argument while the dominant factor will be the multiplicative n (in the example, $30\cdot\log 30 = 44.3$, $\log 3000 = 3.5$). Thus the worst case complexity of the whole optimization process may be simplified as

$$O(\mathit{MaxIter}\cdot P\cdot(m\cdot\lambda_{\max}+n\cdot\log n)).\tag{18}$$

As we have illustrated, the GRASP re-optimization is based on an iterative improvement process (represented essentially by the *MaxIter* factor) whose computational complexity may be high for large-scale RWA/grooming networks [38]. Parallel local search algorithms, when applicable, are an effective way to cope with this problem. According to an iteration decomposition principle, the search iterations can be partitioned into several threads and the main procedure run in each of them in parallel. In our GRASP-based approach, the *MaxIter* iterations may be easily distributed among the network nodes and run in parallel, thus cooperating to implement the integrated RWA mechanism within the network control-plane logic (following the so called *multiple-walk independent-thread* strategy, based on distributing the GRASP iterations over the available processors, that in our case are the switching nodes). Each processor works on an independent copy of the problem data, and has an independent seed to generate its own pseudorandom sequence number. Clearly, each processor must base its work on a different pseudorandom sequence, to avoid the same solutions to be found by each of them. A single global variable, whose value can be kept synchronized between all the processors through proper message passing, is required for storing the best solution found by all the participating processors. Here, the communication among the different processors running the GRASP iterations in parallel is limited to the random seed and to the current best solution found. One of the processors acts as the master, by generating the random seeds to be used on each processor, reading and distributing the problem data and the iterations, and finally collecting the best solution found by each computing node. Since all the iterations are completely independent and very little information is exchanged between the participants, linear speedups can be easily obtained provided that no major load imbalance problems occur. To further improve load balancing, the iterations may be uniformly distributed over the processors according to their demands. From extensive simulations, it has been observed that a typical value is *MaxIter* = 30, that is in line with the mean number of network nodes in MAN/WAN network. In general, if we assume that each processor core runs one search thread, the computational complexity of the parallel *GRASP* procedure is decreased to

$$O(P\cdot(m\cdot\lambda_{\max}+n\cdot\log n))\tag{19}$$

The time evaluation tests we conducted have showed that it is an affordable time complexity for a single processor thus confirming the feasibility of such approach. Parallel implementations of GRASP may also be used in conjunction with path-relinking. In the multiple-walk independent-thread implementation described by Aiex et al. [24], each processor applies path-relinking to pairs of elite solutions stored in a local pool. The OSPF opaque Link-State Advertisement (LSA) mechanism can be easily used to transport synchronization information between the cooperating nodes.

5. Performance evaluation and considerations

In this section, we examine the performance of our new algorithm with an extensive simulation study, by working on several real network topologies, with and without the continuity constraint (i.e. wavelength converters supported or not). The simulation details together with the most interesting results and observations emerged from the experiments have been reported in the following paragraphs.

5.1. The simulation environment

In order to evaluate the performance of the proposed hybrid on-line/offline routing framework we realized a simple and very flexible ad hoc optical network simulation environment written in Java in order to take advantage of its extensibility, ease of modifiability, portability and strict math and type definitions [39]. To allow us to perform a simple comparative analysis the above environment supports discrete-event simulations in fiber/lambda switched networks for several well-known RWA algorithms, both Dijkstra based, such as Minimum Hop Algorithm (MHA) and Shortest-Widest Paths (SWP), or interference-based, such as Maximum Open Capacity Algorithm (MOCA). It supports flexible definition and modification of simulation parameters and configuration files to define complex simulation test cases, allowing the creation of new network topologies. Simulations have been performed on an HP® DL380 Dual Processor (Intel® Xeon® 2.5 GHz) server running FreeBSD® 4.11 operating system and Sun® Java® 1.4.2 Runtime Environment by using several optical network topologies modeled as un-directed graphs in which each link has a non-negative capacity. In all the experiments, we used an incremental traffic model in which connection requests, defined by a Poisson process, arrive with a rate of δ requests/s and are distributed on the available network node according to a random-generated or predefined traffic matrix. Thus, the session holding time has been set to be infinite to enhance the effect of connection's load on the network, that is, each connection lasts through the entire simulation, letting the network resources saturating more rapidly. This can be done since dynamic connection releases do not adversely affect both the performance and behavior of the whole RWA framework whose periodic re-optimization steps are managed offline. The above choice allowed us to make our tests under the worst-case conditions.

5.2. Results analysis

The results presented are taken from many simulation runs on several network topologies with various GRASP parameter values and an increasing number of connection requests varying from 0 to 1000. The GRASP parameters and bandwidth unit request values used in our simulations are reported in Table 1.

Table 1
Simulations performed and parameters used.

Parameters	NSFNET/GEANT2
Number of connections	Varying from 0 to 1000 (step 100) 50% before and 50% after re-optimization
Random generated bandwidths (OC-unit)	{1,3,12} with different distribution probability
MaxIter	30
α	{0.2, 0.5, 0.8}
Local search	Breadth and depth local searches
Number of simulations	20 simulations run per topology; each simulation repeated 10 times
Measurements	Blocked connections with and without re-optimization Objective function gain Freed OC-units

As can be seen from the previous table, 20 simulations per topology were run. Each run has been repeated 10 times and the average performance metric values have been calculated. We considered several values for the α parameter chosen from the set reported in the previous table and tried out both the breadth and depth search options in local search. Since the main objective of our work is to efficiently address the problem of re-optimization by maximizing the overall network resource usage and hence the medium and long-term carriers' revenues, we were interested in demonstrating the efficiency of the proposed approach on a significant variety of real network topologies through a comparative assessment between our hybrid framework based on Dijkstra SPF for online routing associated to our GRASP-based re-optimization algorithm, and an environment in which no re-optimization was realized. Such assessment focused on their overall effectiveness in term of request rejection ratio/blocking factor, network resource usage optimization and time performances. We did not compare our solution with other re-optimization proposals available in literature due to the peculiarity of our hybrid approach and hence to the lack of comparable results obtained on the same network topologies and traffic distributions. Accordingly, we studied the re-optimization benefits for varying traffic demands and different network layouts. We tried out different static, predefined, or randomly generated traffic demand matrices on several network topologies, both randomly generated and well-known, such as NSFNET and GEANT2 (see Fig. 5) with the bandwidths for the links ranging from OC-1 to OC-768 bandwidth units.

In our tests, each connection request was characterized by a bandwidth demand ranging from OC-1 to OC-12 (622 Mbps) units. We routed these connections using SPF routing. As the network load grows, we continuously monitor the network efficiency expressed by the rejection ratio/blocking factor. When the connections demand exceeds the value of a fixed load threshold we invoke the GRASP re-optimization. We then evaluate the re-optimization gain, comparing the network loads that could be sustained with or without re-optimization. We measured this gain, at varying optimization thresholds, in terms of several quantities. We computed the overall bandwidth gains as the difference (in OC-units) between the total bandwidth available on the network before and after re-optimization and we analyzed the objective function behavior. We also observed the gains in terms of difference between the maximum number of additional end-to-end connections that could be accommodated in the network with and without performing re-optimization. For space limitations and results consistency, we show only the most remarkable results obtained with the well-known NSFNET and GEANT2 networks. In all the presented charts, to make the results more readable and better highlight the evolution trends and properties of the observed performance metrics, the plotted curves have been obtained through polynomial interpolation on the sample observations taken before the beginning and at the end of each GRASP re-optimization step. For the first set of simulations we generated a random demand matrix from all the available source-sink pairs. Next, the network has been loaded by adding end-to-end connections, whose arrival rate is proportional to the values reported for the corresponding pairs in the demand matrix. In Figs. 6 and 7 we show the results in terms of request rejection rate and number of end-to-end connections gain obtained with the first set of simulations on NSFNET in which we used the breadth local search and varying values of $\alpha = \{0.5, 0.2, 0.8\}$ respectively for tests T1, T2 and T3. The simulation without re-optimization is simply indicated with SPF in the figure.

Here, the number of rejected/accepted requests (Y axis) is reported against the number of total generated connection requests (X axis). By looking at the variations in rejection rate as the network was loaded with an increasing number of connections, we can note that, without re-optimization, the network starts rejecting connections much earlier (while the re-optimization approach starts rejecting at about 400 connections) and a substantial and slightly increasing gain can be constantly observed throughout

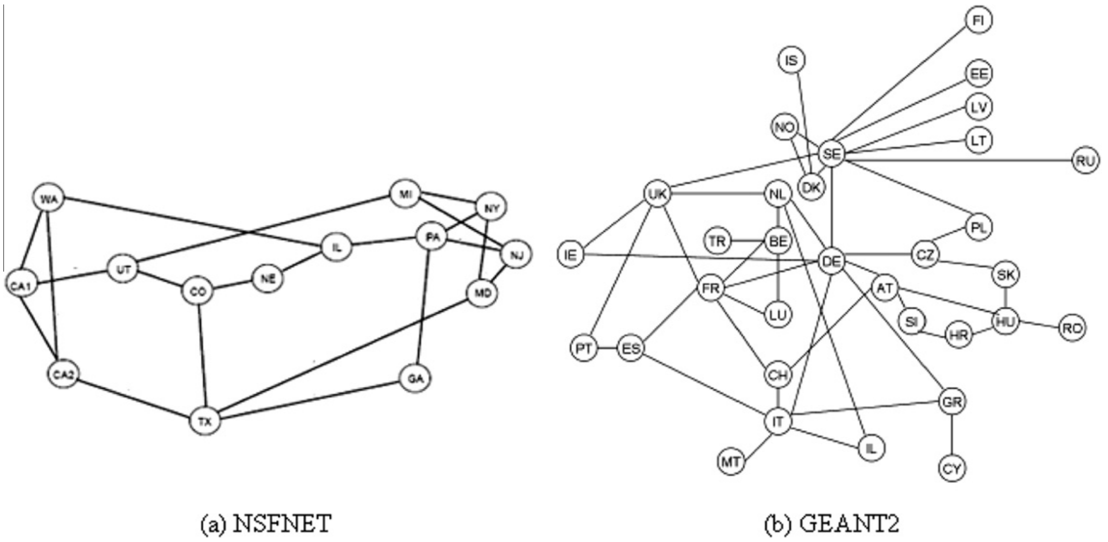


Fig. 5. Sample network topologies used in simulation.

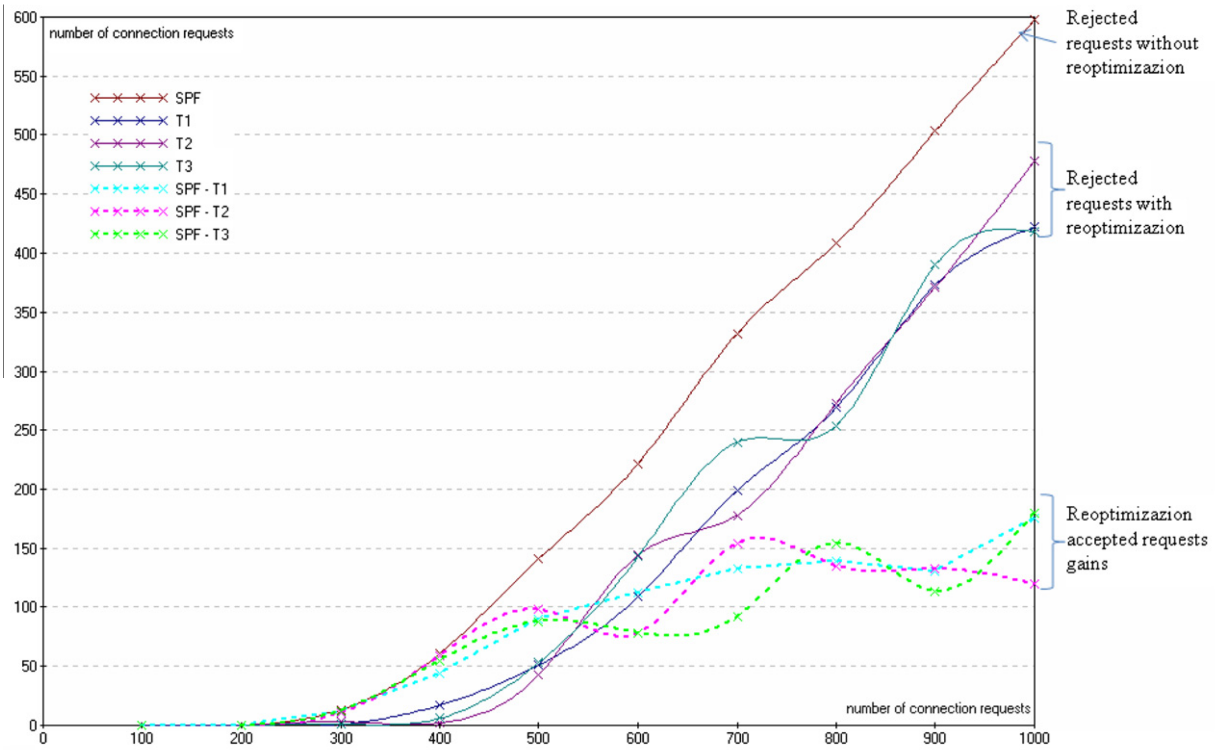


Fig. 6. NSFNET simulation results: blocked connections and re-optimization accepted requests gains.

the experiment also when the load and hence rejection rate increase toward the total network saturation.

Fig. 7 shows the variation in bandwidth gain with increasing network load. The gain starts with a rapid growth and diminishes (dramatically for some α values) after around 50% of the load region; it then restarts increasing and then it progressively decreases, once a local maximum around the 75% of the load is reached, according to an alternate/elastic behavior very common in traffic-related phenomena [40]. The same swinging behavior can also be observed, even if much smoothed, in the above Fig. 6

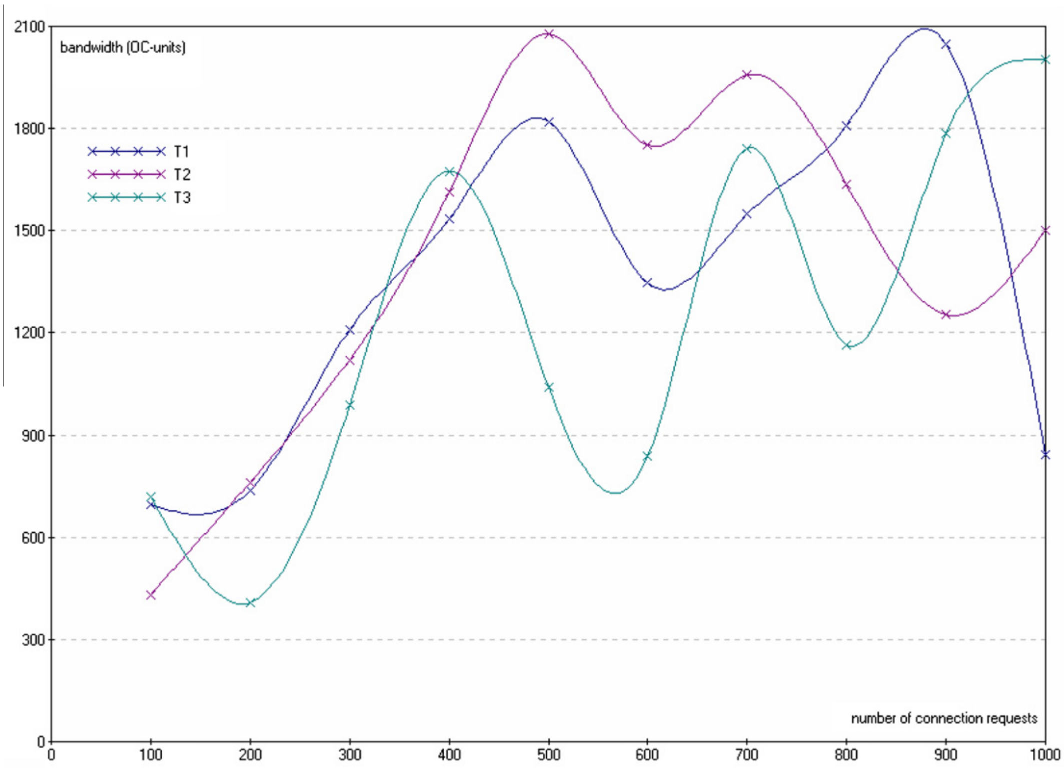


Fig. 7. NSFNET simulation results: bandwidth gains in OC-units.

connection gain curves. The reason for this can be attributed to the reduced flexibility in moving existing paths to other routes because of the periodical reduction in links' spare capacities. This observation suggests that, in a dynamic lightpath routing scenario, the re-optimization procedure should be invoked before the network load reaches 50% (or some other value, dependent on the network specific characteristics). Waiting for the first request to be blocked before attempting re-routing might be too late and reduce the overall efficiency of the re-optimization process. The second set of simulations experiments analyzes the sensitivity of the re-optimization scheme with respect to variations in network topology (and hence link bandwidth) for a fixed demand matrix. Here, the connection requests are distributed on all the network nodes according to the probability distribution obtained from the traffic matrices given in [41] for NSFNET and in [42] for GEANT2, where traffic volumes have been scaled proportionally to the traffic distribution. In Figs. 8 and 9 we show the results obtained with the second set of simulations on GEANT2 in which we used the depth local search and varying values of $\alpha = \{0.5, 0.2, 0.8\}$ respectively for tests T4, T5 and T6. The simulation without re-optimization is simply indicated with SPF in the figure.

When observing the variations in both connections and available bandwidth gain as the network topology changes towards a more connected mesh with higher capacity links (for this sake we can compare the NSFNET and GEANT2 behaviors respectively shown in Figs. 6–9) we can evidence a certain progressive decline in the gain increasing with the network size. This effect highlights that re-optimization algorithms are more efficient in finding good solutions in narrower networks with fewer resources available, since networks with a larger number of links and more capacity have in general less potential for offline reconfiguration gain, as even simple online routing schemes can easily produce acceptable solutions under a physiological load. In fact, in a lightly loaded network, with a lot of available links and capacities, the re-optimization effect allows the admission of only a few additional connection requests while, in an overloaded network, where resource become scarce, the number of additional connections that can be routed on a re-optimized logical topology greatly increases. This behavior is due to the fact that when there is plenty of connectivity resources the number of connections rejected by traditional routing algorithms such as SPF greatly increases, so that any re-optimization strategy – assumed to work on the same traffic matrix – has a much larger potential to satisfy the requests that were previously rejected.

We observe from both Figs. 6 and 9 that the best results in terms of bandwidth and connection gains have been obtained by working with a good balance between greediness and randomness (T1, T4 with $\alpha = 0.5$), in which gains follow a more linear trend, whereas an almost greedy or random selection process (respectively T2, T5 with $\alpha = 0.2$ and T3, T6 with $\alpha = 0.8$) exhibits a sinusoidal-like behavior due to the myopic choices done by the use of a too small or too large RCL. Obviously, a linear trend is much preferable with respect to a non-linear one as it achieves better average results in terms of higher bandwidth gains and lower blocking probability. Similar results have been obtained in breadth and depth local searches, showing that local optima may be reached in both procedures thanks to the robust approach of the Greedy meta-heuristic. Finally, in Fig. 10 we plotted the average re-optimization times for the execution of the illustrated tests referred to the parallel implementation of the Grasp procedure. The results have been obtained by running one search thread on each network node and measuring the starting and ending times. Resulting times have been averaged by the number of nodes and reported in Fig. 10 against the number of connection requests. Communication times among nodes have not been considered in the measurement. As we can see, even in GEANT2, which is a more complex network than NSFNET, the computational times are all below the 1 s threshold which is an affordable delay time for a network [2]. Consequently, the proposed re-optimization strategy is suitable to be implemented on a

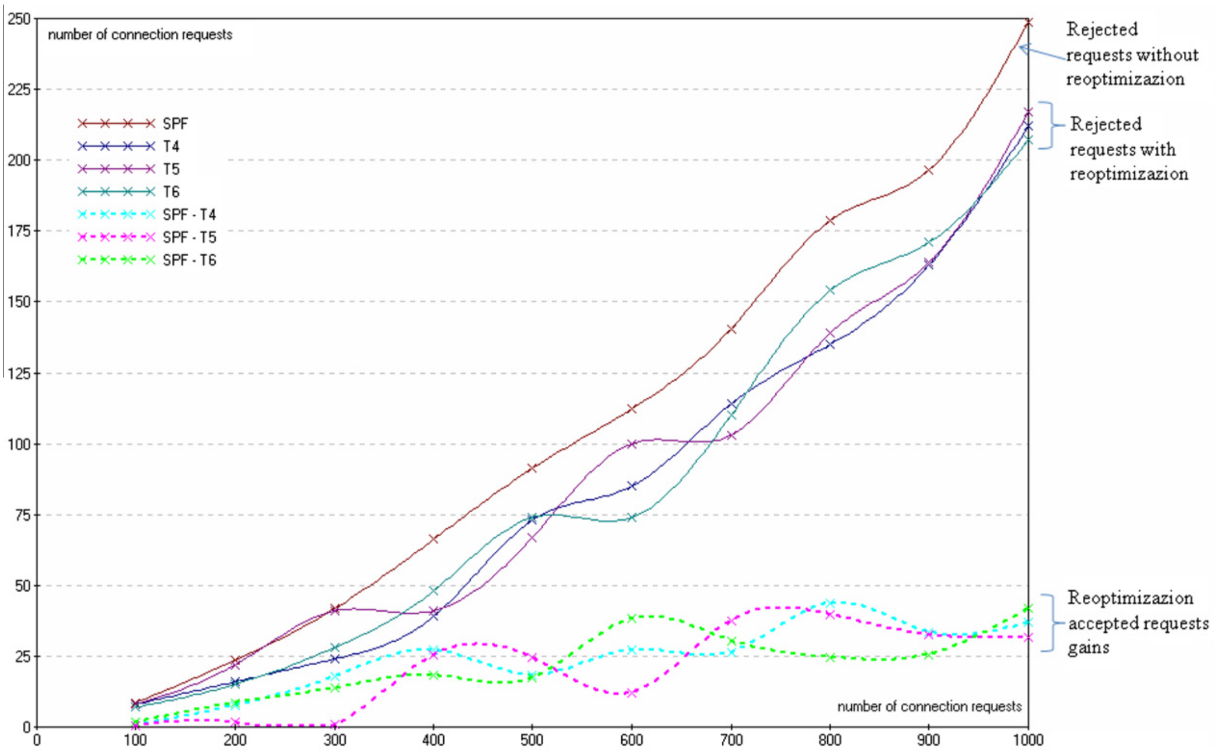


Fig. 8. GEANT2 simulation results: blocked connections and re-optimization accepted requests gains.

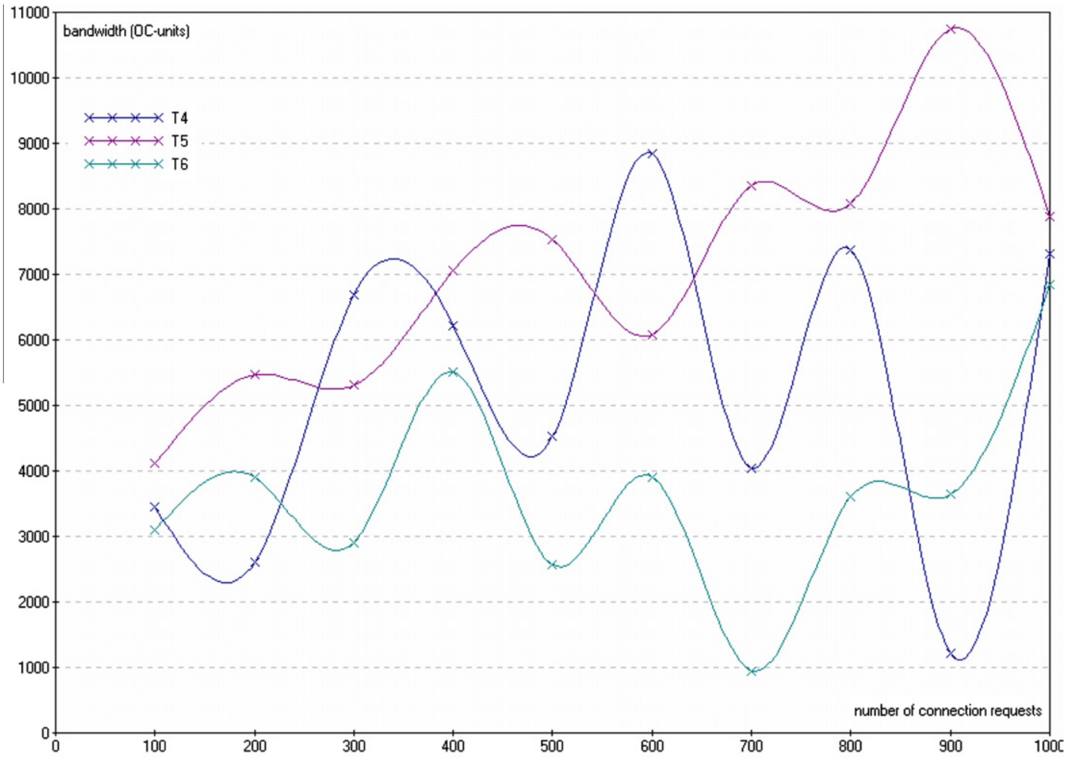


Fig. 9. GEANT2 simulation results: bandwidth gains in OC-units.

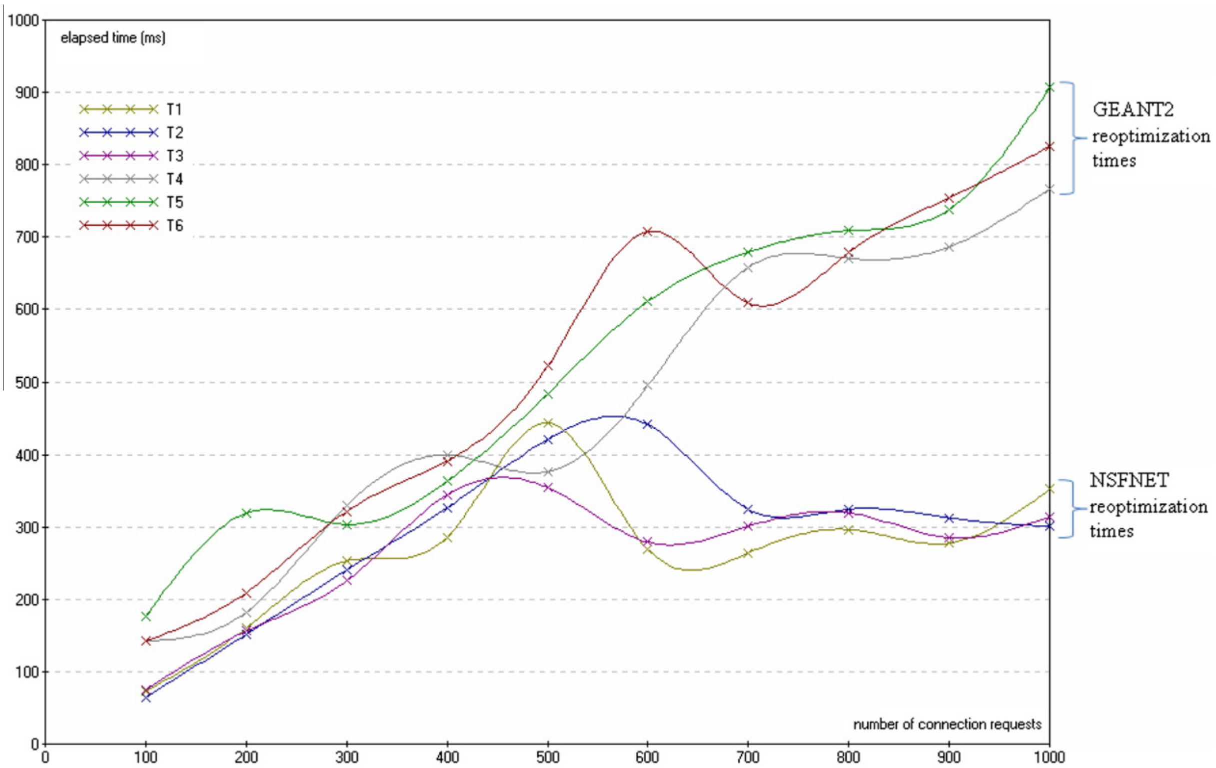


Fig. 10. GEANT2 and NSFNET simulation results: re-optimization times.

frequent basis, also in a production network, before the network gets totally saturated, so that an acceptable degree of efficiency and service continuity is ensured until it is possible, also at higher loads.

6. Conclusions

Dynamic demands and topology changes caused by the addition/deletion of new links and/or capacity, together with online routing decisions made on a best-of-now basis, without knowledge of the lightpaths to be set-up in the future, cause the wavelength routing logic to behave sub-optimally, thereby creating opportunities for improvements in network bandwidth efficiency. Lightpath topology re-optimization seizes on these opportunities and offers network operators the ability to better adapt to the network and user requests dynamics. This is achieved by regularly (or upon a particular event) re-routing the existing demands, temporarily eliminating the drift between the current solution and the optimal one that is achievable under the same conditions. Starting from the above premises, we formulated a hybrid approach for integrated online routing and offline reconfiguration of optical networks with sub-wavelength traffic. The key feature of such a scheme is the ability to maintain the network balanced through adaptive on-demand re-optimization by ensuring that a sufficient capacity is kept available between any ingress–egress pair so that the maximum number of connections arriving to the network can be satisfied. The overall focus of this work has been the balancing between the reconfiguration cost (in terms of disturbance to the users' connections already deployed over the network) and a good and simple RWA and grooming solution. We defined a set suitable goals and strategies for an integrated approach, and provided a formulation of the re-optimization procedure based on an iterative refinement process of multiple local search steps structured as a GRASP meta-heuristic procedure. We also developed a heuristic strategy that attempts to achieve minimal disturbance reconfiguration by performing local reconfiguration and delaying as possible the need for global reconfiguration. Furthermore, re-optimization would only occur when needed (when the rejection ratio become unacceptable and the potential savings from re-optimization exceeds some threshold) or upon certain events such as when new links are added or torn down. Simulation results show the notable margins of re-optimization achievable with our approach as well as the time complexity feasibility in real networks such as NSFNET and GEANT2. Rejection ratios of connection set-up requests decreased, allowing more connections to be successfully routed, and bandwidth gains have been observed in all the simulation runs. Besides, we proposed an efficient parallel implementation of GRASP with path-relinking that showed quite linear speedups in the number of processors and such a strategy has been successfully applied to greatly accelerate the proposed re-optimization scheme. In conclusion, the proposed re-optimization schema achieves prominent improvements in network efficiency, with the consequent cost savings.

References

[1] J. Zhou, X. Yuan, A study of dynamic routing and wavelength assignment with imprecise network state information, in: Proceedings of ICPP Workshop on Optical Networks, 2002.

[2] G. Mohan, C. Siva Ram Murthy, A time optimal wavelength rerouting algorithm for dynamic traffic in WDM networks, Journal of Lightwave Technology 17 (3) (1999).

[3] K. Kar, M. Kodialam, T. Lakshman, Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications, IEEE Journal on Selected Areas in Communications 18 (2000).

[4] G.N. Rouskas, M. Ammar, Dynamic reconfiguration in multihop WDM networks, Journal of High Speed Networks 4 (3) (1995) 221–238.

[5] T.A. Feo, M.G.C. Resende, A probabilistic heuristic for a computationally difficult set covering problem, Operations Research Letters 8 (1989) 67–71.

[6] T.A. Feo, M.G.C. Resende, Greedy randomized adaptive search procedures, Journal of Global Optimization 6 (1995) 109–133.

[7] F. Glover, Tabu search and adaptive memory programming—advances, applications and challenges, in: R.S. Barr, R.V. Helgason, J.L. Kennington (Eds.), Interfaces in Computer Science and Operations Research, Kluwer, Boston, 1996, pp. 1–24.

[8] F. Glover, M. Laguna, Tabu Search, Kluwer Academic Publishers, Boston, 1997.

[9] H. Zang, J. Jue, B. Mukherjee, A review of routing and wavelength assignment approaches for wavelength-routed optical wdm networks, Optical Networks Magazine 1 (2000) 47–60.

[10] R. Ramaswami, K.N. Sivarajan, Routing and wavelength assignment in all-optical networks, IEEE/ACM Transactions on Networking 3 (5) (1995) 489–500.

[11] I. Chlamtac, A. Ganz, G. Karmi, An approach to high-bandwidth optical wans, IEEE Transactions on Communications 40 (7) (1992) 1171–1182.

[12] H. Zang, J. Jue, L. Sahasrabuddhe, R. Ramamurthy, B. Mukherjee, Dynamic lightpath establishment in wavelength routed networks, IEEE Communications Magazine 39 (9) (2001) 100–108.

[13] B. Ramamurthy, B. Mukherjee, Wavelength conversion in WDM networking, IEEE Journal Selected Areas in Communications 16 (7) (1998) 1061–1073.

[14] F. Palmieri, U. Fiore, Dynamic network optimization for effective QoS support in large grid infrastructures, in: Lizhe Wang, Jinjun Chen, Wei Jie (Eds.), Quantitative Quality of Service for Grid Computing: Applications for Heterogeneity, Large-scale Distribution, and Dynamic Environments, IGI Global, 2009, pp. 28–48, ISBN: 978-1-60566-370-8.

[15] S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.

[16] J.H. Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA, USA, 1975.

[17] M.G.C. Resende, C.C. Ribeiro, Greedy randomized adaptive search procedures, in: F. Glover, G. Kochenberger (Eds.), State-of-the-Art Handbook of Metaheuristics, Kluwer Academic Publishers, 2002, pp. 219–249.

[18] F. Glover, M. Laguna, Tabu Search, Kluwer Academic Publishers, Boston, 1997.

[19] F. Glover, Scatter search and path relinking, in: D. Corne, M. Dorigo, F. Glover (Eds.), New Ideas in Optimisation, Wiley, 1999.

[20] F. Glover, M. Laguna, R. Marti, Fundamentals of scatter search and path relinking, Control and Cybernetics 39 (2000) 653–684.

[21] M. Laguna, R. Marti, GRASP and path relinking for the 2-layer straight line crossing minimization, INFORMS Journal on Computing 11 (1999) 44–52.

[22] M.G.C. Resende, C.C. Ribeiro, GRASP with path-relinking for private virtual circuit routing, Networks 41 (2003) 104–114.

[23] C.C. Ribeiro, E. Uchoa, R.F. Werneck, A hybrid GRASP with perturbations for the Steiner problem in graphs, INFORMS Journal on Computing 14 (2002) 228–246.

[24] R.M. Aiex, M.G.C. Resende, P.M. Pardalos, G. Toraldo, GRASP with path relinking for the three index assignment problem, INFORMS Journal on Computing (2003).

[25] M.G.C. Resende, R.F. Werneck, A GRASP with path-relinking for the p-median problem, Technical Report, AT&T Labs Research, Florham Park, NJ, 2002.

[26] K.C. Lee, V.O.K. Li, A circuit rerouting algorithm for all-optical wide area networks, in: Proceedings of IEEE INFOCOM 1994, 1994, pp. 954–961.

[27] D. Banerjee, B. Mukherjee, Wavelength-routed optical networks: linear formulation, resource budgeting tradeoffs, and a reconfiguration study, IEEE/ACM Transactions on Networking 8 (5) (2000) 598–607.

[28] J.F.P. Labourdette, G.W. Hart, A.S. Acampora, Branch-exchange sequences for reconfiguration of lightwave networks, IEEE Transactions on Communications 42 (10) (1994) 2822–2832.

[29] I. Baldine, G.N. Rouskas, Dynamic reconfiguration policies for WDM networks, in: Proceedings of IEEE INFOCOM 1999, 1999.

[30] M. Sridharan, A.K. Somani, M.V. Salapaka, Approaches for capacity and revenue optimization in survivable WDM networks, Journal of High Speed Networks 10 (2) (2001) 109–125.

[31] E. Bouillet, J.F. Labourdette, R. Ramamurthy, S. Chaudhuru, Lightpath re-optimization in mesh optical networks, IEEE/ACM Transactions on Networking 13 (2) (2005) 437–447.

[32] R. Bhatia, M.S. Kodialam, T.V. Lakshman, Fast network reoptimization schemes for mpls and optical networks, Computer Networks 50 (3) (2006) 317–331.

[33] J.Y. Zhang, O.W.W. Yang, J. Wu, M. Savoie, Optimization of semi-dynamic lightpath rearrangements in a WDM network, IEEE Journal on Selected Areas in Communications 25 (9) (2007) 3–17.

[34] D.R. Din, Solving virtual topology reconfiguration problem on survivable WDM networks by using simulated annealing and genetic algorithms, Photonic Network Communications 18 (1) (2009) 1–13.

[35] Y. Xin, M. Shayman, R.J. La, S.I. Marcus, OPNp1-2. Reconfiguration of survivable MPLS/WDM networks, in: Proceedings of IEEE GLOBECOM 2006, 2006.

[36] M. Kodialam, T.V. Lakshman, Minimum interference routing with applications to MPLS traffic engineering, in: Proceedings of IEEE Infocom, 2000.

[37] M.G.C. Resende, C.C. Ribeiro, GRASP and path-relinking: recent advances and applications, in: T. Ibaraki, K. Nonobe (Eds.), Metaheuristics: Progress as Real Problem Solvers, Springer, 2005.

[38] D.S. Johnson, C.H. Papadimitriou, M. Yannakakis, How easy is local search?, Journal of Computer and System Sciences 17 (1) (1998) 79–100

[39] F. Palmieri, U. Fiore, S. Ricciardi, SimulNet: a wavelength-routed optical network simulation framework, in: Proceedings of IEEE ISCC 2009, 2009, pp. 281–286.

[40] C. Casetti, R. Lo Cigno, M. Mellia, M. Munafò, Z. Zsóka, A realistic model to evaluate routing algorithms in the internet, in: Proceedings of IEEE Globecom 2001, 2001, pp. 1882–1885.

[41] R. Ramaswami, K.N. Sivarajan, Design of logical topologies for wavelength-routed optical networks, IEEE Journal on Selected Areas in Communications 14 (1996) 840–851.

[42] S. Uhlig, B. Quoitin, J. Lepropre, S. Balon, Providing public intradomain traffic matrices to the research community, ACM SIGCOMM Computer Communication Review 36 (1) (2006) 83–86.

SimulNet: a Wavelength-Routed Optical Network Simulation Framework

Francesco Palmieri, Ugo Fiore and Sergio Ricciardi*
Federico II University of Napoli
Via Cinthia, 5, Complesso Universitario di Monte S. Angelo, 80126 Napoli, Italy
{francesco.palmieri, ugo.fiore, sergio.ricciardi}@unina.it

Abstract

Simulation seems to be the best available alternative to the deployment of expensive and complex testbed infrastructures for the activities of testing, validating and evaluating optical network control protocols and algorithms. In this paper we present SimulNet, a specialized optical network simulation environment providing the foundation for the study and analysis of the key control plane characteristics of wavelength-routed networks. Such an environment would provide researchers with an open framework for easily exploring the evolving characteristics of WDM-routed technologies which includes developing new protocol suites or performing rapid evaluation and easier comparison of results across research efforts.

Keywords: simulation, WDM-switching, RWA, OXC

1. Introduction

Simulation plays an important role in network protocol design, providing researchers with a cost effective method to analyze and study the behavior of proposed protocol models. Particularly, simulation becomes an indispensable tool in WDM-based transparent optical networking research, since it helps researchers to quickly and inexpensively validate and evaluate the performance of new protocols and algorithms without the need of installing and managing complex network testbeds that require very expensive optical devices and communication infrastructures. In fact, even on a fully featured testbed, not all algorithms, mechanisms and protocols can be readily implemented, tested and evaluated, because most of the available optical network equipments are proprietary vendor products featuring a very low degree of programmability. Implementing new control plane features and algorithms on top of such proprietary equipment can be a very tedious and difficult task, since it requires accessing, controlling and programming the low level capabilities of these devices. Unfortunately only a very limited number of simulation tools are available for conducting research related to different routing and wavelength assignment algorithms (RWA), topology management (e.g. converter placement algorithms), re-optimization, centralized and distributed wavelength reservation schemes, and the effect on RWA algorithms of inaccuracies in network-wide state information or physical-level impairments. Furthermore, most of the available solutions are limited in scope, difficult to use or not totally open. In many cases, they have also been based on simulation models designed specifically to cope only with a specific problem. Since each of these solutions used their own simulation platform, model and assumptions, it is generally difficult to reuse existing protocol modules and compare simulation results under a common assessment environment. Our aim is to address these problems with the introduction of a new simulation framework developed specifically for the testing and performance analysis of distributed dynamic RWA algorithms on optical networks and also for optimization algorithms and protocols validation. The framework we present, called *SimulNet*, has been designed to be easy to use and exhibits satisfactory performance also when simulating large network topologies. It has been implemented according to a modularized, platform-independent, and extensible architecture and provides a useful baseline library of commonly used RWA algorithms and signaling schemes. We based our framework on a totally flexible network model, supporting heterogeneous WDM equipment, with or without wavelength conversion capability, in which the number and type of lambdas can vary on each link. It provides a fully dynamic and configurable path selection scheme supporting sub-wavelength bandwidth allocation (grooming). Furthermore, it allows simulations to be aware of all the complexity, expensiveness, performance and resource-limitation constrains implicit in the various flavors of optical switching devices, for example explicitly and proportionally penalizing, when instructed to do so, all the paths that require wavelength conversion. The simulator has been successfully validated by implementing some well-known algorithms, and comparing the results with those available in literature.

* *Sergio Ricciardi is PhD student at the Department of Computer Architecture of the Technical University of Catalonia (UPC), Jordi Girona 3, 08034 Barcelona, Catalunya, Spain.*

978-1-4244-4671-1/09/$25.00 ©2009 IEEE

2. Related work

At the state of the art, several simulation packages and tools modeling optical network infrastructures are available, but none of them totally provides the necessary support and flexibility needed for the study and the performance assessment of new routing, signaling and wavelength management schemes on modern WDM-empowered networks. GLASS (GMPLS Lightwave Agile Switching Simulator) [1], developed at NIST, is an optical network simulation tool built on the SSFNET Framework [2]. It supports many advanced features, including GMPLS and QoS, and several optical failure recovery schemes. Since GLASS is implemented in Java it is easy to configure, modular and platform independent. Unfortunately, its complexity adversely conditioned its performance and hindered the development of new routing schemes on this platform. OWNS [3] is tool that is being used for research in the optical network domain. It is built on the well known mature foundation of the NS-2 [4] simulation infrastructure with a set of additional optical WDM extensions. Unfortunately, with these benefits it also inherited some inconveniences of NS-2, such as bloated code base, slow execution speed and large memory footprint. Also, configuring and implementing any new routing scheme or protocol requires extensive and careful coding, because of the absence of a modular structure. OWNS also lacks the feature set required for the implementation of converter placement algorithms. JAVOBS [5] is a set of flexible and configurable java libraries dedicated to the simulation of Optical Burst Switched (OBS) networks. It supports simulation of various OBS schemes, reservation protocols and scheduling algorithms, on any topology of reasonable size. SimulNet can be considered quite similar to JAVOBS as for the general simulation framework and the configurability, with the fundamental difference that while JAVOBS operates over OBS networks, SimulNet is expressly designed to operate over wavelength switched networks' control plane layer.

3. Basic Architecture

Simulation environments allow the user to predict the behavior of a set of network devices on a complex network, by using an internal model that is specific to the simulator. Simulators do not necessarily reproduce the same sequence of events that would take place in the real system, but rather apply an internal set of transformation routines that brings the simulated network to a final state that is as close as possible to the one the real system would evolve to. To accomplish this task, simulators work on a model that contains all the relevant information that must be observed, while abstracting irrelevant details, thus simplifying both the simulation and the analysis of the network. This approach typically allows the simulated network to scale well in size and complexity, as irrelevant details may be totally abstracted or represented in a simplified manner. Thus, the simulator can be used to manage also complex networks with hundreds or thousands of network elements and wavelengths, while greatly reducing the probability of introducing programming bugs. The drawback is that the simulated devices may have limited functionalities and their behavior may not closely resemble that of real world devices. This is the reason why both the parameters to be represented or abstracted in the model and the representation detail must be carefully chosen. An excessive degree of sophistication in the simulation environment is often the cause of a limited flexibility in configuration and introduction of new functional modules, and in general heavily taxes both the runtime performance and the ease of use. In fact, the real objective of a simulation is to make available a sustainable and effective model of the involved system that is sufficiently accurate to reliably analyze and evaluate only the specific set of properties of the system that are under observation. Furthermore, simulation results are easier to analyze than experimental ones because important information at critical points can be easily logged to help researchers in diagnosing the network behavior. Starting from these premises, the proposed simulator architecture has been designed according to a fully modular scheme, to accommodate most of the control-plane specific characteristics of wavelength-routed network and provides a useful set of network traffic generators. The simulator models optical fibers and wavelengths individually for each link, thus allowing maximum flexibility in the representation of dissimilar networks. It also provides network utility subroutines to configure, monitor, and gather results and statistics about simulated networks. The resulting framework provides great configuration flexibility in topology definition (each node can or cannot support conversion capability; the number and type of lambdas, the associated maximum bandwidth and propagation delay can be specific for each link) and enables easy extension to introduce new features, not only in terms of RWA and grooming algorithms but also allowing other aspects of the simulator to be customized or expanded if desired in the future. The whole framework has been designed as an object-oriented application in which all functional entities are implemented by individual objects or modular entities interacting with each other. Each object can contain or refer to other objects for the composite identification of an operation. Objects can also be abstracted and encapsulated to facilitate extension. For example, if a new optical node type or new RWA algorithm in the simulator is needed, it can be conveniently added by extending the existing network device object or RWA algorithm object. The key components of the simulator can be divided into physical layer abstractions, such as optical switching devices and multi-wavelength fiber links, and logical layer modules working upon them, such as the RWA simulation engine and the traffic generation module, which together create and maintain the virtual topology.

3.1. Modeling the physical layer

The OXC object models the various types of switching nodes that constitute the Transport Plane of the simulated optical network. OXCs are responsible for switching traffic from an incoming fiber/wavelength pair (or electrical link) to an outgoing one. The necessary mapping information is maintained in a special internal table, called the switching matrix, which the OXC looks up before accomplishing each switching operation. According to their routing, traffic grooming and wavelength conversion capabilities of their interfaces, OXC objects can be used to model both hybrid optical routers, add and drop multiplexers or pure wavelength routers, operating in transparent or opaque mode. They provide wavelength routing & switching, wavelength conversion, fiber/port switching, waveband switching and regeneration. The simulated optical switching functions are supported in either transparent optical switching architecture or in the opaque one with O-E-O conversions. According to the underlying architecture the lambda conversion/switching capabilities may have different limitations, such as the number of converters in an OXC and the range of conversion. Each fiber features an independent WDM capability, i.e. consists of several wavelengths and the actual number of wavelengths is a parameter that can be configured for each fiber. Moreover, within each fiber strand connecting a couple of OXC objects there is one additional virtual link representing the dedicated communication channel needed to convey control plane traffic.

3.2. The RWA engine

The RWA simulation engine that is the heart of our framework is composed of two components: routing module and wavelength assignment module.

3.2.1. Routing. The routing module supports the dynamic creation and deletion of unidirectional lightpaths by determining the routes needed to establish such paths in the current network topology and with the available resources. It maintains two distinct views of the network – the physical and the virtual topology – represented as multi-graphs in which each edge corresponds to an individual channel (wavelengths), each fiber can support multiple channels, and there can be more than one fiber connecting the same pair of nodes.

The physical topology represents the real network infrastructure where every link between adjacent switching nodes and its resources (wavelengths and bandwidths) and conversion capabilities are catalogued. The virtual topology represents the current network status, dynamically modified during the creation of lightpaths. Multiplexing and demultiplexing wavelengths is simulated implicitly in the virtual topology forwarding plane. The simulator switching logic within the RWA engine will ideally bridge the incoming wavelength and outgoing wavelength to set up lightpaths. In a configurable manner, lightpath set-up is realized either first on existing available lightpaths (according to a best fit strategy in case there are several lightpaths satisfying the connection request), or directly executing the RWA algorithm without consulting the lightpath allocation tables. In the current implementation, the RWA engine already provides some basic predefined algorithms, such as Dijkstra shortest path algorithm/Minimum Hop Algorithm (MHA), Shortest Widest Path Algorithm (SWP), MINimum Residual Capacity algorithm (MinRC), k-shortest path algorithm and Minimum Interference Routing Algorithm (MIRA) [6][7][8].

3.2.2. Wavelength Assignment. The wavelength assignment module is responsible for selecting the wavelengths needed for lightpath creation. Our simulator supports optional wavelength conversion capability on each OXC. Also, channel allocation is done separately for each link. Up to 64 wavelengths can be used for transmission between network elements. More than one wavelength channel can be assigned between two network elements depending on traffic demands. Wavelength reuse is employed when possible. Most of the predefined algorithms of our RWA engine treats routing and wavelength assignment as a single joint problem, that is they search paths for connection requests considering both the routing and the wavelength assignment activities at the same time. Nevertheless, our simulator supports also disjoint RWA, in which the above activities are managed as two distinct problems. First the routing algorithm finds a suitable path for the connection in terms of available fibers between source and destination nodes; then, the wavelength assignment procedure provides wavelength assignments according to a specific criterion aimed at optimizing some resources. Different wavelength assignment algorithms have been implemented: first-fit, most-used/pack, least-used/spread, random allocation, round robin. With the first-fit algorithm, the first available wavelength is chosen, causing the wavelength with the smallest index to be chosen more often. The most-used/pack algorithm selects

978-1-4244-4671-1/09/$25.00 ©2009 IEEE

the available wavelength which is currently utilized on the largest number of fibers, thus preserving the maximum number of different wavelengths for future requests that may have the wavelength continuity constraint; analogously, the least-used/spread algorithm selects the available wavelength which is currently utilized on the smallest number of fibers, thus trying to balance the load on the network. The random rule distributes the connection traffic randomly so that average wavelength utilizations are balanced. In the round robin allocation scheme wavelengths are indexed and assigned in a circular manner, thus maximizing the use of all the available wavelengths in the network in order to minimize the blocking probability due to wavelengths unavailability. Finally, the bandwidth on a wavelength can be divided into smaller sub-rate capacities called sub-wavelength units, to be assigned to specific end-to-end connections. A connection request can demand one or more sub-wavelengths to a maximum of the whole available bandwidth supported on a specific wavelength, by performing traffic grooming from properly capable devices. Our simulation environment supports both single-hop and multi-hop grooming capability [9]. Single-hop grooming will be only possible on end-to-end connections realized on a single wavelength. On the other side, multi-hop grooming offers maximum flexibility in allowing sub-wavelength allocation across multiple chained lightpaths or single lightpaths built on multiple wavelengths (when the continuity constraint is relaxed). Clearly our network model allows the RWA algorithms to be aware of the different cost and complexity of the two mechanisms.

3.2.3. Signaling. The simulated control plane signaling logic is based on a reverse path reservation model that can be viewed as a simplified subset of RSVP-TE. The connection setup procedure is divided into two phases: downstream and upstream. During the downstream phase, each of the OXC nodes determines whether the required destination is reachable and there are enough resources to accommodate the new connection. As soon as the RWA engine determines on each node both the outgoing fiber and a wavelength with adequate bandwidth resources, it temporarily allocates them by adding a new record on the switching matrix for the new connection. At this instant, the setup is still incomplete, and therefore the switching matrix record is flagged as provisionally reserved and cannot yet be used by the OXC. At each step, the involved node passes the request to the downstream OXC object. If there is no fiber/wavelength acceptable for routing the current request or the available bandwidth is inadequate, the node rejects the requests and sends a response back to the source. The upstream phase starts as soon as the request reaches the egress node of the connection that acknowledges it along the same path. During this phase, when any node receives an acknowledge from its downstream node, it permanently allocates the resources reserved during the downstream phase, and modifies the switching matrix to activate the switching record for the new connection. As soon as the acknowledge is received by the ingress node, the connection is established and the new lightpath is modeled as a new direct edge in the virtual topology multi-graph, henceforward called cut-through edge (see [9]), with the capacity set to the difference between the minimum capacity on all the edges belonging to the lightpath and the fraction of the link bandwidth required by the involved connection request. A cut-through edge can be used in any path selection operation and thus can participate in one or more groomed paths as a single virtual edge (a single hop at the IP layer). When an established lightpath is torn down because the last connection occupying it is ended, the cut-through edge is removed and the edges in the extended graph corresponding to the underlying physical links are set back with full capacity. This schema allows to flexibly model nearly any network topology and to consider node conversion capabilities, wavelength availability and residual bandwidth per logical link at the IP layer. Each new connection request can be routed over a direct lightpath modeled as a single cut-through edge in our multi-graph, or over a sequence of lightpaths (a multi-hop path at the IP level, where each hop can be a lightpath), if it crosses lambda-edge or routers as well that will link together lightpaths.

3.3. The traffic generation module

For each given WDM network topology, the traffic generator module randomizes source and destination pairs according to a uniform distribution or a user-provided traffic matrix. Thus, traffic, expressed as the arrival of new connection request, may flow between any source–destination pairs. Along its path, each connection will require some amount of the sub-wavelength units according to its bandwidth requirement. If the requirement of a call cannot be fulfilled when it arrives, the request will be blocked. The probability that a new request will be blocked is, thus, an important indicator of the overall efficiency provided by networks. Simulation is a rather indispensable method to estimate the blocking probability, as well as other network performance parameters. In large optical networks, call blocking probabilities may be rare events due to the large capacity of the network or a very low request arrival rate. In such cases, standard simulation may require an extremely long runtime, and usually incurs in large relative errors. Connection requests may arrive or be torn down at any moment during the simulation. We developed a session traffic source object which generates three types of traffic, namely defined by a Poisson, exponential on/off and Pareto distribution with different average bit rate per wavelength. In particular, the Poisson distribution is a very good model for human requested end-to-end connections (i.e. semi-permanent virtual circuits) whereas the Pareto one is very useful for modeling typical Internet traffic-driven activity. In order to get more confidence in the simulation results, we provide the possibility to run multiple simulations on the same network with the same algorithm and parameters and eventually get the average result values.

4. Implementation details

The simulator architecture has been realized in modular form. This design choice stems from the nature of the problem and of the networks to be modeled. Each module is implemented as a collection of cooperating objects and each network component is represented by a separate object characterized by its own attributes and methods. A small number of parameters can be easily determined from simulation for nondestructive measurements. Objects and modules have been used as the building blocks of the whole simulation program. All objects have been designed in unified modeling language (UML), and implemented in Java. UML enabled to show the block diagram of the simulator as well as the simulation flow and the interaction among objects. Java has been chosen in order to take advantage of its object oriented paradigm, great extensibility, ease of modifiability, portability, and object serialization (to checkpoint out simulations). Besides, Java provides also the garbage collector that automatically frees the unused memory, thus simplifying the code and optimizing memory occupation. The simulator can run on workstations with the Java Virtual Machine, e.g. MS-windows or various UNIX machines. The simulator builds a model of the network in the multi-graph object, which is the main object representing the network topology; it is made up of the sets of nodes, fibers, and edges (initially each edge corresponds to a single wavelength). There is also the set of the paths currently routed in the network that grows up as the connection requests are honored. Besides there is a set of support objects build up and maintained for efficiency sake: these objects comprise indexes, vectors and matrices that have proven to speed up the computational time of the simulations. Each object has been designed keeping in mind efficiency, usability and flexibility. Abstract objects are used to uniform interfaces and are extended and instantiated as concrete objects representing real network elements. Object oriented encapsulation, inheritance and polymorphism have been exploited thanks also to the use of design patterns. The simulator also includes an intuitive graphic user interface (GUI) allowing flexible definition and modification of network topology and simulation parameters; a versatile configuration language is used to define complex simulation environments. The configuration file structure mainly consists of the physical topology information, such as nodes and links, and the source/destination enabled nodes and their properties in terms of bandwidth range, its distribution and related parameters. After execution of the simulation a detailed trace file is generated which shows the exact timing of the important events that have occurred during the course of simulation with the relevant details.

5. Functional evaluation

In order to validate the simulator, we ran many simulations based on well-known algorithms on both test network topologies and real networks such as NSFNet and Geant2, that have all been successfully modeled and tested. As performance measures, both throughput and blocking probability have been adopted. Figure 1 shows the blocking probability for NSFNet network topology with various known RWA algorithms.

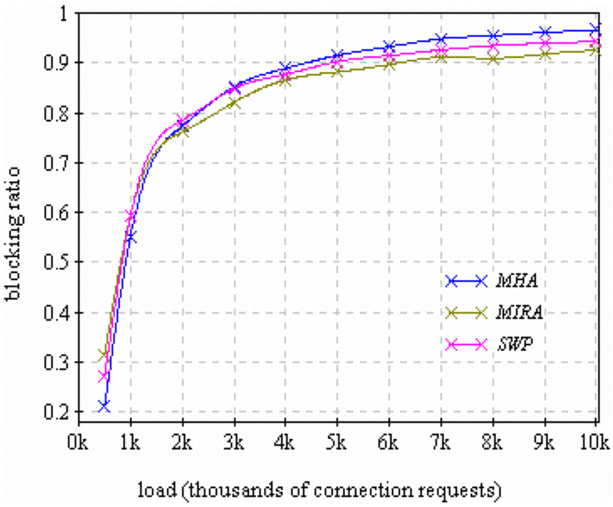


Figure 1. Average rejection ratio on NSFNet.

We can immediately observe from the above Fig. 1 the behavior, in terms of rejection ratio under increasing load, of some test RWA algorithms that closely matches several results available in literature (e.g. [10]). We also tested reliability of our simulator over variable network spacing from ring to full-mesh topology, as in [5] (see Fig. 2). Starting from a simple 8 nodes ring network with 8 fiber links and 32 wavelengths per link, new links are added stepwise up to reach 32 links. At each step, either the number of wavelengths on each link (series #1) or the wavelength transport capacity (series #2) is decreased in order to keep constant the network transport capacity, thus obtaining two series of test networks, as in Table 1.

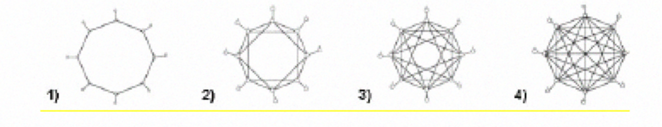


Figure 2. Ring to Full-Mesh test networks topologies.

Test series #1: λ varying, OC-unit constant				
Test network	***1***	***2***	***3***	***4***
Fiber links	8	16	24	32
λ per link	32	16	10	8
OC-unit per λ	24/48	24/48	24/48	24/48
Test series #2: λ constant, OC-unit varying				
Test network	***1***	***2***	***3***	***4***
Fiber links	8	16	24	32
λ per link	4	4	4	4
OC-unit per λ	768	384	256	192

Table 1. Test networks topologies parameters.

The Dijkstra shortest path routing has been used to route connection requests having a bandwidth requirement between 1 and 12 OC-units and uniform distribution between source/destination nodes is assumed. The results are shown in Fig. 3.

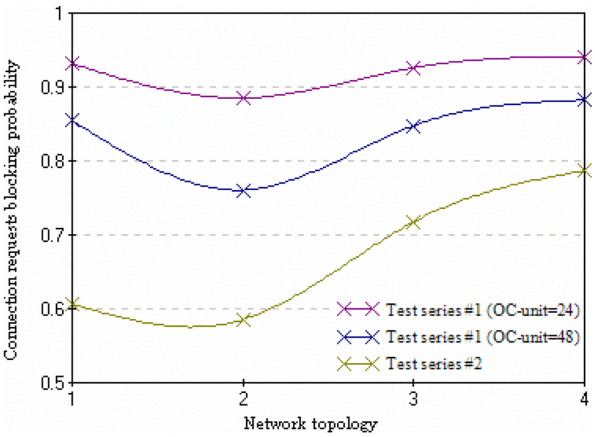


Figure 3. Simulations on Ring to Full-Mesh topologies.

As expected, better performances are achieved with a good balance between the number of channels (wavelengths) and transport capacity (OC-unit) on each channel. The first network is penalized because of the few links between node pairs, while the last because of either the small number of wavelengths per link (series #1) or the low transport capacity (series #2). Comparing the results of tests over known real networks and the effects of varying wavelengths and traffic load for different network topologies, we can conclude that our simulator is implemented properly and could be useful for optical network simulations.

6. Conclusion

Simulators have proven to be an indispensable tool for the study of the rapidly evolving optical networks technologies. In this paper we have presented SimulNet, a flexible simulator for WDM-routed networks. SimulNet has been expressly realized for the design and the evaluation of both RWA and optimization algorithms as well as for the validation of signaling protocols and scheduling schemes. SimulNet has shown good flexibility in managing even complex networks and has exhibited accuracy of simulations results and satisfactory performances, making it an useful tool for the optical network research community.

7. References

[1] Kim,Y., et al., "GLASS (GMPLS Lightwave Agile Switching Simulator) - A Scalable Discrete Event Network Simulator for GMPLS-based Optical Internet", *Proceedings of the JCCI'02* conference, 2002.

[2] Ogielski, A.T., SSFnet. Presentation at the *DARPA Next Generation Internet Conference*, Arlington, VA, 1999.

[3] Wen, B., Bhide, N., Shenai, R., Sivalingam, K., "Optical wavelength division multiplexing (WDM) network simulator (OWns): Architecture and performance studies", *SPIE Optical Networks Magazine*, pp. 16–26, 2001.

[4] The Network Simulator(ns), http://www.isi.edu/nsnam/ns/.

[5] Pedrola, O., Rumley, S., Klinkowski, M., Careglio, D., Gaumier, C., Solé-Pareta, J., "Flexible Simulators for OBS Network Architectures", *Proceedings of the IEEE ICTON 2008* conference, 2008.

[6] Guerin, R., Williams, D., and Orda, A., "QoS Routing Mechanisms and OSPF Extensions", *Proceedings of GLOBECOM*, 1997.

[7] Kar, K., Kodialam, M. and Lakshman, T. V., "Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 12, pp. 921–940, 2000.

[8] Kar, K. Kodialam, M. and Lakshman, T. V., "Integrated Dynamic IP and Wavelength Routing in IP over WDM Networks", Proceedings of *IEEE Infocom* conference, *2001.*

[9] Balasubramanian, S., Somani, A. K., "On Traffic Grooming Choices for IP over WDM networks," Proceedings of *IEEE Broadnets* conference, 2006.

[10] Boutaba, R., Szeto, W., Iraqi, Y., "DORA: Efficient Routing for MPLS Traffic Engineering", J*ournal of Network and Systems Management*, Vol. 10, No. 3, 2002.

Energy-Oriented Models for WDM Networks

Sergio Ricciardi[1], Davide Careglio[1], Francesco Palmieri[2],
Ugo Fiore[3], Germán Santos-Boada[1], and Josep Solé-Pareta[1]

[1] CCABA, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] DII, Seconda Università di Napoli, Aversa (CE), Italy
[3] CSI, Università di Napoli Federico II, Naples, Italy
{sergior,careglio,german,pareta}@ac.upc.edu,
{fpalmier,ufiore}@unina.it

Abstract. A realistic energy-oriented model is necessary to formally characterize the energy consumption and the consequent carbon footprint of actual and future high-capacity WDM networks. The energy model describes the energy consumption of the various network elements (NE) and predicts their energy consumption behavior under different traffic loads and for the diverse traffic types, including all optical and electronic traffic, O/E/O conversions, 3R regenerations, add/drop multiplexing, etc. Besides, it has to be scalable and simple to implement, manage and modify according to the new architecture and technologies advancements. In this paper, we discuss the most relevant energy models present in the literature highlighting possible advantages, drawbacks and utilization scenarios in order to provide the research community with an overview over the different energy characterization frameworks that are currently being employed in WDM networks. We also present a comprehensive energy model which accounts for the foreseen energy-aware architectures and the grow rate predictions which tries to collect the main benefits of the previous models while maintaining low complexity and, thus, high scalability.

Keywords: Energy-oriented models, evolutionary energy-aware WDM networks.

1 Introduction

It is now held as a scientific fact that humans contribute to the global warming of planet Earth through the release of carbon dioxide (CO_2), a Green House Gas (GHG), in the atmosphere. Recently, the carbon footprint of ICT was found to be comparable to that of aviation [1]. It is estimated that 2-3% of the CO_2 produced by human activity comes from ICT [2][3] and a number of studies estimate an energy consumption related to ICT varying from 2% to 10% of the worldwide power consumption [4]. It is worth to mention for example that Telecom Italia and France Telecom are now the second largest consumer of electricity in their country [5][6] and British Telecom is the largest single power consumer in the UK [7].

The reduction and optimization of energy consumption are among the main goals of the European Union (EU). The EU in fact is encouraging the ICT sector to reduce

I. Tomkos et al. (Eds.): BROADNETS 2010, LNICST 66, pp. 534–547, 2012.

its carbon footprint in a drive to drastically reduce Europe's overall carbon emissions by 2020 setting its ambitious 20/20/20 goals: cutting its annual consumption of primary energy by 20% and increase the production of renewable energy to a share of 20% by 2020 [8]. Recent initiatives gathering major IT companies started to explore the energy savings and green energy usage in network infrastructures. For example, Telefonica commits to reducing 30% its network energy consumption by 2015 [9].

In the current telecommunications networks, the vast majority of the energy consumption can be attributed to fixed line access networks. Today, access networks are mainly implemented with copper based technologies such as ADSL and VDSL whose energy consumption is very sensible to increased bitrates. The trend is to replace such technologies with mobile and fiber infrastructure which is expected to increase considerably the energy efficiency in access networks. Such ongoing replacement is moving the problem to the backbone networks where the energy consumption for IP routers is becoming a bottleneck [10][11]. In Japan it is expected that by 2015, IP routers will consume 9% of the nation's electricity [12].

In such a new environment, the development of more accurate cost models which include the energy consumption factor for both the deployment (Capex) and the maintenance (Opex) of network infrastructures is fundamental. In this paper, we discuss the most relevant energy models present in the literature highlighting possible advantages, drawbacks and utilization scenarios in order to provide the research community with an overview over the different energy characterization frameworks that are currently being employed in WDM networks.

This article is structured as follows. Section 2 introduces the energy related problems and the possible energy-efficient and energy-aware solutions. In Section 3, we illustrate the energy-aware architectures on which the energy models are currently based. Section 4 discusses the three main energy models present in the literature. Section 5 illustrates real power consumption models for router architectures with different scaling factors. In Section 6 we present our comprehensive energy model for WDM networks. Finally, Section 7 summarizes the conclusion of this article.

2 Background

Increasing the energy efficiency of the different equipment, operations or processes constituting a network infrastructure is not the ultimate solution, as argued in the Khazzoom-Brookes postulate [13]: “increased energy efficiency paradoxically tends to lead to increased energy consumption” (a phenomenon known as the Jevons Paradox or rebound effect as well). In fact, an improvement of the energy efficiency leads to a reduction of the overall costs, which causes an increase of the demand and consequently of the energy consumption overtaking hence the gained offset.

It is safe to say that a paradigm shift is required in the network in order to sustain the growing traffic rates while limiting and even decreasing the power consumption. In order to overcome the rebound effect, it is necessary to adopt the *carbon neutrality* or, when available, the *zero carbon* approach. In carbon neutrality, GHG emitted by legacy (dirty) energy sources (e.g. fossil-based plants) are compensated – hence, neutrality – by a credit system like the cap and trade or the carbon offset [14]. In the zero carbon approach, renewable (green) energy sources (e.g. sun, wind, tide) are employed and no GHG are emitted at all. Clearly, green energy sources are always preferable with respect to the dirty ones as they limit (or avoid at all) GHG emissions, although renewable sources are variable in nature and their availability may change in time. In order to reduce the energy consumptions and contain the concomitant GHG emissions in the atmosphere, the two following measures have been identified:

- *Energy efficiency:* refers to a technology designed to reduce the equipment energy consumption without affecting the performance, according to the *do more for less* paradigm. It takes into account the environmental impact of the used resources and constraints the computations to be executed taking into account the ecological and potentially the economic impact of the used resources. Such solutions are usually referred to as *eco-friendly solutions*.
- *Energy awareness:* refers to an *intelligent* technology that adapts its behavior or performance based on the current working load and on the quantity and quality of energy that the equipment is expending (*energy-feedback information*). It implies knowledge of the (dirty or green) sources of energy that supply the equipment thus differentiating how it is currently being powered. Energy-aware solutions are usually referred to as *eco-aware solutions*. A direct benefit of energy aware techniques is the removal of the Khazzoom-Brookes postulate.

To become a reality, green Internet must rely on both concepts and a new energy-oriented network architecture is required, i.e. a comprehensive solution encompassing both energy-efficient devices and energy-aware paradigms acting in a systemic approach. The definition of a proper energy model to estimate and characterize the energy consumption of a network infrastructure is hence of primary importance. Nonetheless, due to its distributed character and wide diversity in network equipment types (routers, switches, modems, line cards, etc.), a direct estimation of network equipment power consumption is notoriously difficult. Several energy models have been proposed so far which try to emulate the different network elements (NEs) in an easy and comprehensive manner.

3 Energy-Aware Architectures

Current router architectures are not *energy-aware*, in the sense that their energy consumption does not scale sensibly with the traffic load. In [15] several router architectures have been analyzed and their energy consumptions under different traffic loads have been evaluated. Results show that the energy consumption between an idle and a heavily loaded router (with 75% of offered traffic load) vary only of 3% (about 25 W on 750 W). This happens because the router line cards, which are the most power consuming elements in a router, are always powered on even if they are totally idle. On the contrary, the energy consumption decreases to just 50% if the idle line cards are physically disconnected. Such a scenario suggests that future router architectures will be energy-aware, in the sense that they will be able to automatically switch off or dynamically downclock independent subsystems (e.g. line cards, input/output ports, switching fabrics, buffers, etc.) according to the traffic loads in order to save energy whenever possible. Such energy-aware architectures are advocated both by standardization bodies and governmental programs [16] and have been assumed by various literature sources [15][17][18]. Our study will be therefore focused on such energy-aware architectures that can adapt their behavior, and so, their energy consumption, to the current traffic loads. The energy consumption of such architectures is made up of a fixed part (Φ), needed for the device to be turned on, and a variable part (ε), somehow proportional to the traffic load. It is precisely *how* the variable energy consumption scales with the traffic that differentiates the various energy models. In the following paragraphs, we present them in detail and discuss their major benefits and drawbacks. Note that in each model the power consumption starts from the fixed power consumption value Φ that represents the power necessary for the device to stay up (and idle).

4 Energy Models

Basically, three different types of energy models have been reported in the literature:

1. Analytic energy models
2. Experimental energy models
3. Theoretical energy models.

4.1 Analytic Energy Models

Analytic energy models [18] take into consideration a number of parameters describing the NEs and provide their energy consumption by mean of a mathematical description of the network. The challenge of analytic energy models is to abstract irrelevant details while representing essential aspects in order to obtain a realistic characterization of the network elements energy consumption. Once an analytic model has been set up, it has the ability to describe the energy consumption of NEs in virtually any possible network configuration. Furthermore, as irrelevant hardware, software and configuration details may be totally abstracted or only partially represented, the analytic models have the ability to scale well with the network size. In fact, the abstraction and the generalization are the two key points of this kind of models. Anyway, analytic models have some drawbacks as well. What has to be represented in the model and what should instead kept out is a design choice that has to be carefully planned, as an excessive degree of sophistication may introduce unnecessary complexity and unwanted behaviors. Furthermore, the complexity degree of the modeled devices should resemble the real world devices as far as possible but it is not always possible to know the proprietary internal device architectures and hardware technical specifications.

In [18] the authors propose an analytic energy model in the ILP formulations for energy-efficient planning in WDM networks. They identify three types of traffic: transmitting, receiving and switching traffic, though there is no difference between electronic and optical traffic.

4.2 Experimental Energy Models

Experimental models [19][20][21][22][23] totally rely on energy consumption values of real world devices. They consider the NEs energy consumptions declared by the manufacturers or the experimentally measured values to create a map of well-known off-the-shelf working devices samples. For routers – which are the most studied NEs – the energy consumption is reported against the aggregated throughput and then the mapping is used for interpolating or extrapolating energy consumption data for routers of any size. Anyway, this model has several drawbacks. On the one hand, the declared energy consumptions may not closely resemble the real values especially when the device is working with a specific hardware and/or software configuration. On the other hand, although the experimentally measured energy consumption values may measure the energy consumption under different traffic loads, they only refer to a punctual evaluation under specific assumptions. Furthermore, the interpolation/extrapolation method is not a reliable measure of real devices energy consumption, as the devices energy consumption may vary sensibly with its technology, architecture, features and size (e.g. aggregated throughput, number of line cards, ports, wavelengths, etc.). In fact, in [19] the authors analyze power consumption of core routers based on datasheets found in [20], and conclude that for higher throughputs the routers consume more power. However, smaller routers tend to be located near the edge of the network whereas the larger routers are more central in the network where the traffic is more aggregated. Therefore they consider the power consumption per bit rate. This reveals that the larger routers consume less energy per bit than smaller ones. When aggregating over the entire network, the power consumption will also be the largest at the edge of the network and smaller in the centre. It is also showed how energy consumption depends on the packets size and on the bitrates of the links. Greater packets need less energy than smaller ones, due to the lower number of headers that have to be processed. In [21] it is showed that circuit-based transport layer reduces energy consumption with respect to packet-switched layer, due to the lower processing required for managing connections and to the higher processing needed for analyzing each packets' headers. Nevertheless, it is often difficult to gather real energy consumption values, so it is not always feasible to create a complete mapping of real world devices, and it is practically impossible to measure energy consumption of future NEs architectures before designing and building them. So, an experimental model, though providing some real energy consumption values, is not enough to cope with the requirements of a comprehensive energy model.

In [22] and [23] the authors propose a mixed energy model. Network nodes energy consumption is modeled by averaging experimental data of a real network scenario, whilst the power consumption of links is analytically modeled by a static contribution due to optical transceivers, and by an additional term which takes into account possible (optical) regenerators.

4.3 Theoretical Energy Models

Theoretical models [24] are instead totally based on the theoretical predictions of the energy consumption as functions of the router size and/or the traffic load (in a way similar for the Moore's law [25] for the central processing units and the Gilder's law for the bandwidth of communication systems [26]). Such models have the benefit of being simple and clear, but the predictions may substantially differ on the long run from the real energy consumption values. Besides, it is often difficult to foreseen the NEs energy consumptions and, as they rely only on empirical data, it is not a based on any rigorous scientific model. Furthermore, both experimental and theoretical energy models do not provide detailed energy consumption of each subsystem or component, but they simply describe at high level the energy consumption at the expense of granularity and accuracy. In [24] the author proposes a simple theoretical model in which the router energy consumption grows with a polynomial function of its capacity. This estimation has been proved to be quite similar to the real energy consumption values [23].

5 Power Consumption Models

Power consumption models express the power consumption (*P*) of routers versus the offered traffic load (*L*). In power consumption models, the current absorbed power, i.e. energy per second, is plotted against the traffic load that the router is currently offering. The power consumption may be expressed through a set of concrete models whose growth behaviors are obtained either from analytical, experimental, theoretical energy models or a combination of them. In the following sections, we analyze four different models: linear, theoretical, combined and statistical power consumption models.

5.1 Linear Power Consumption Models

In linear models, the power consumption scales linearly with the traffic load up to the maximum router capacity (its aggregated bandwidth). Here, routers with diverse technology and/or sizes may scale differently with the traffic: three scale factors (ε_1, ε_2, ε_3) are reported in Fig. 1.

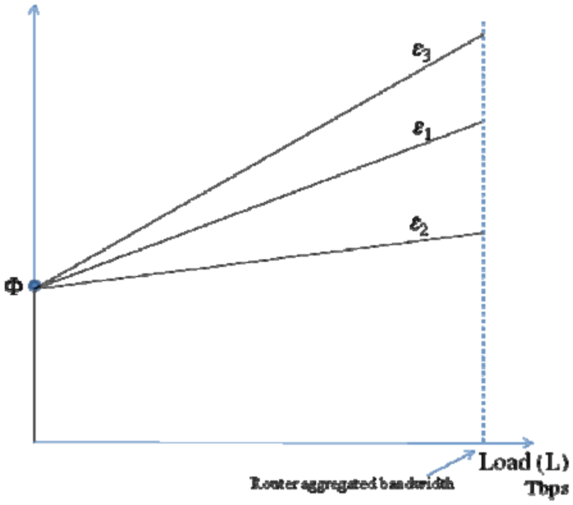


Fig. 1. Power consumption in linear power consumption model

In this model, it holds that:

$$P = \varepsilon_i \cdot L \qquad (1)$$

where ε_i is a scaling factor depending on the technology and size of the router *i*. Alternatively, the diverse slopes (ε_i) may represent different traffic types (see the Section 5.4), as was assumed in [18].

This power consumption model has the benefit of being simple and easy to implement, but it has the drawback that it is not possible to upper bound the power consumption to a desired values (e.g. 2Φ, as the results in [15] suggest).

5.2 Theoretical Power Consumption Models

In theoretical models, the power consumption is expressed as a function of the load that tries to follow the trend of real devices power consumption. Using a high level formula, theoretical models are usually employed to describe in a simple though effective manner the relation between the power consumption and the current traffic load. The theoretical energy model presented in [24] is the following:

$$P = C^{2/3} \qquad (2)$$

which states that the router power consumption grows with a polynomial function of its capacity. Now, if we substitute the router capacity with the load, we obtain a feasible model to represent how the power consumption varies with the traffic load. Such a model has demonstrated to be quite in line with the energy consumption of some real world devices [24], and for this reason has been sometimes used in literature papers [19].

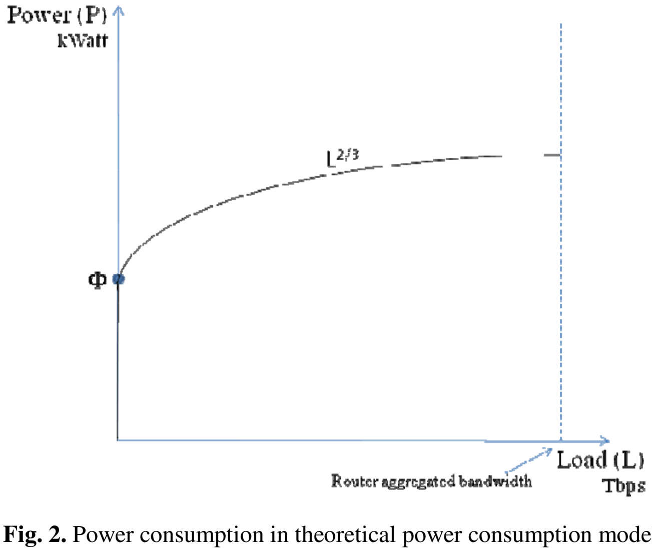


Fig. 2. Power consumption in theoretical power consumption model

Theoretical power consumption models show an easy-of-use advantage as it suffice to substitute the router aggregate bandwidth or current traffic load to immediately get the power consumption value. No tuning of any parameter is needed (such as ε_i) and the power consumption growth rate is always well predictable. Unfortunately, such models have the same drawbacks as the theoretical energy ones (see the section 4.3).

5.3 Combined Power Consumption Models

Combined models are characterized by different power consumption scaling rates at different traffic loads. They are represented by step functions whose domain is partitioned into different traffic load intervals. Each load interval may be characterized by a different function; for example (see Fig. 3), the power consumption may scale linearly (ε) with low loads (lower than t_1), polynomially ($L^{2/3}$) at medium loads (between t_1 and t_2) and exponentially (2^L) at high loads (greater than t_2). Some or all the sub-functions may be derived from other models, as in the example.

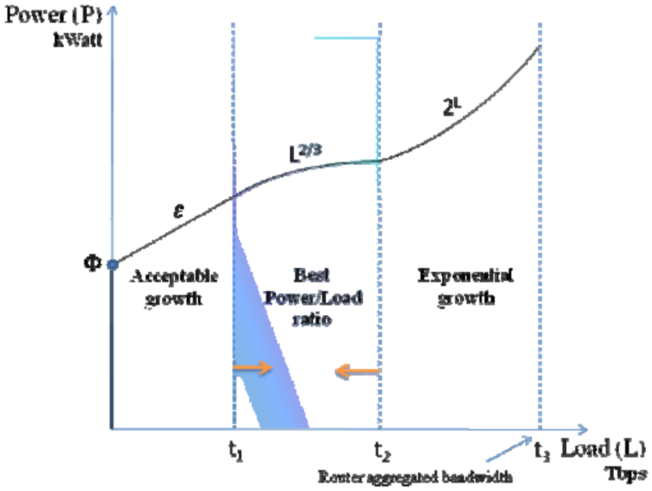


Fig. 3. Power consumption in combined power consumption model

Note that in such a model, it may be convenient to balance the traffic across the network in order to keep the router local traffic inside the acceptable zone where the energy consumption scales polynomially with the traffic load. In fact, it may be worthwhile to keep the traffic above the t_1 threshold, in order to amortize the fixed power consumption Φ, and below the t_2 threshold, to not exceed into the exponential power consumption zone (between t_2 and t_3).

Such power consumption models are pretty complete and may be used to resemble quite complex scenarios in which the network elements have complex architectures and show a known – although not linear – overall power consumption behavior. Note that, thanks to their greater complexity, such models open new perspective on the traditional network load balancing criteria in order to save energy while achieving low connection rejection ratios. Obviously, such added values come at the expense of computational complexity and scalability.

5.4 Statistical Power Consumption Models

Statistical models consider an additional factor contributing to the energy consumption which is the traffic *type*: all optical or electronic traffic, O/E/O conversions, 3R regenerations, optical amplifications, wavelength conversions, are all examples of different traffic types that affect differently the energy consumption inside a given router. In fact, each type of traffic has in principle different power consumptions when traversing a router (either as an optical lightpath or a packet/circuit-switched electronic path), also depending on the technology and the architectural design that the router adopts. The model is defined as *statistical* because the power consumption depends at each moment on the statistical distribution of the overall traffic in the router. The more traffic of kind *i*, the more the energy consumption will depend on the scaling factor ε_i. Furthermore, each router may have its different scaling factors depending on its technology, architecture and size. For example, in Fig. 4 three different types of traffic are represented, each with its own scaling factor: electronic traffic (ε_3), optical traffic without wavelength conversion (WC) capability (ε_2), and optical traffic with WC capability (ε_1). The three types of traffic have different impacts on the overall router energy consumption, but all of them grow linearly. Note that the electronic traffic scales worse than the optical traffic, as reported in [27]. Note also that, in the example reported in Fig. 4, the three traffic types scales all linearly, even if with different slopes. Statistical models may assume that the various types of traffic scale at different growth rates, for example the electronic traffic may scale exponentially while the optical traffic with WC may scale polynomially and the optical traffic without WC may scale linearly. Furthermore, each router may have its own statistical energy model depending on its design choices in order to adapt its energy consumption behavior to different technologies and architectures.

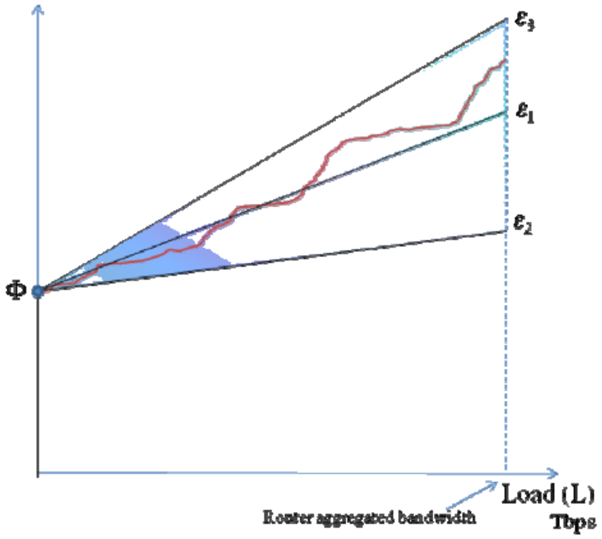


Fig. 4. Power consumption in statistical power consumption model

The statistical model is the most complete one as it allows representing a wide range of devices and power consumption behaviors depending not only on routers technology factors but also on the different traffic types.

6 A Comprehensive Energy Model for WDM Networks

In order to formally characterize the energy consumption of network elements we propose a comprehensive analytic model based on real energy consumption values and in line with the theoretical grow rate predictions encompassing new energy-aware architectures that adapt their behavior with the traffic load in order to minimize the energy consumption.

The energy model comprises three types of traffic of a WDM network:

1. Electronic traffic (with or without add/drop multiplexing, electronic wavelength conversion, 3R regeneration, etc.);
2. Optical traffic with WC;
3. Optical traffic without WC.

These types of traffic are supported by different flavors of optical and electronic network elements (router, switches, transceivers, optical fiber links and amplifiers, 3R regenerators, etc.). Power consumption of real NEs has been obtained by literary sources[15][20][23] [27] [28] and power consumption equations have been derived from these measurements.

Such an energy model characterizes the different components and sub-systems of the network elements involved in energy consumption. It provides the energy consumptions of network nodes and links of whatever typology and size and under any traffic load. The efforts in the developing of such an energy model have been focused on realistic energy consumption values. For this scope, the energy model has been fed with real values and the energy consumption behavior of NEs has been crafted in order to match with the state-of-the-art architectures and technologies. At this extent, future energy-efficient architectures with enhanced sleep mode features have been considered and implemented in the energy model. The energy model is based on a linear combinations of energy consumption functions derived from both experimental results [15][19][20][23][27][28] and theoretical models [22][23][24]. Besides, following the results reported in [15][16][19][28], the power consumption has been divided into a fixed and a variable part; fixed part is always present and is required just for the device to be on; variable part depends on the current traffic load on the device and may vary according to different energy consumption functions. We chose a linear combination of two different functions (logarithmic and line functions) and weighted them with a parameter depending on both the type of traffic and the size of the NE, in order to obtain a complete gamma of values and thus adapting its behavior to the most different scenarios. In particular in our energy model we managed to obtain that larger routers consume less energy per bit than the smaller routers (see Fig. 5), as reported in [19][20], and that electronic traffic consumes more energy per bit that optical traffic (see Fig. 6), as reported in [27][28]. Wavelength conversion and 3R regenerations have a not negligible power consumption which is accounted for in the model. Finally, links have an energy consumption that depends on the length of the fiber strands and thus on the number of optical amplification and regeneration needed by the signal to reach the endpoint with an acceptable optical signal-to-noise ratio (OSNR).

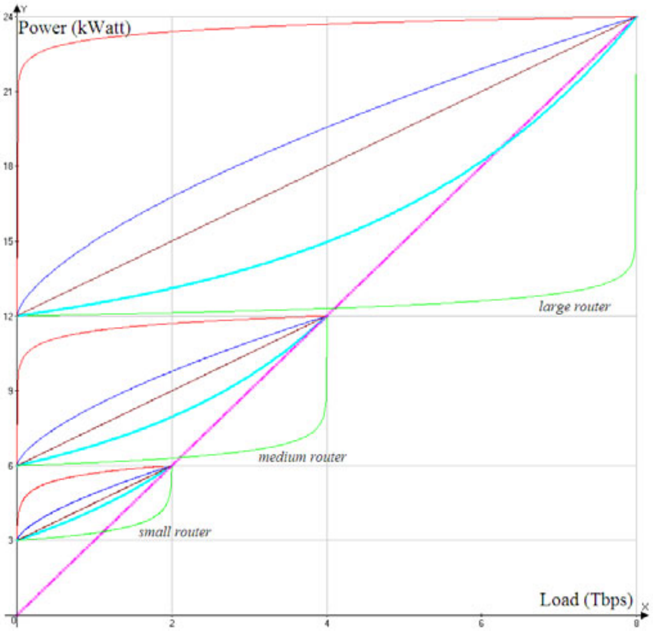


Fig. 5. Power consumption functions for various size electronic routers

The power consumption functions of three routers of different sizes are reported in Fig. 5. Each router may support different types of traffic, each defined by a different curve. In the example in figure, the thicker lines represent the power required by a given type of traffic (e.g. electronic traffic). We can observe that, according to our model, the larger the router, the larger the *total* energy consumption, as the fixed part notably contributes to (half of) the energy consumption. But if we focus only on the variable power consumptions, we observe that, for example, a traffic load of 2 Tbps, requires as much as 3 kW in the smaller router, about 1.5 kW in the medium one and just 1 kW in the larger router. In this way, we managed to obtain that greater routers consume less energy per bit than smaller ones, as reported in [19][20]. Note also that the overall energy consumption scales linearly with the size of the router and that half of the energy consumption is due to the fixed part and the other half to the variable part, according to literature source [15].

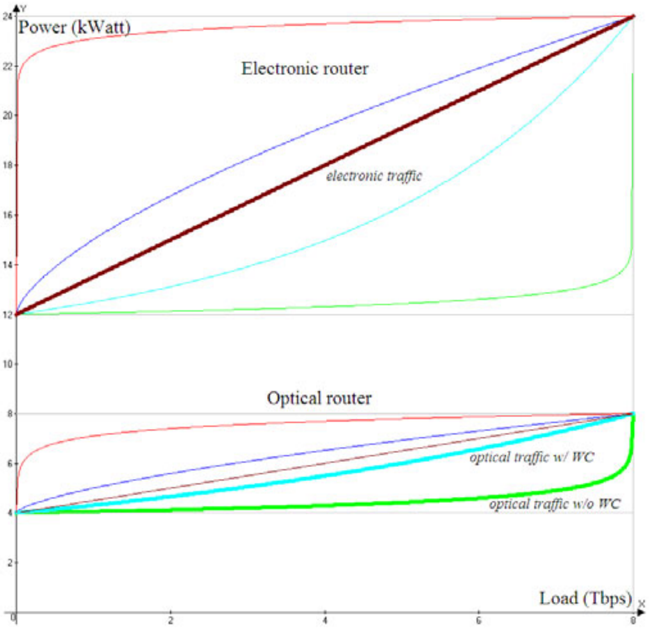


Fig. 6. Power consumption functions for electronic and optical routers

The power consumption functions of an electronic and an optical router are reported in Fig. 6 (optical router values not in scale). Three types of traffic are represented: electronic traffic in the electronic router and optical traffic with and without WC in the optical one. We observe that the electronic traffic grows quickly with respect to the optical traffic and that, among the optical traffic, the WC actually consume a not negligible quantity of energy. As the power consumption functions are obtained by linear combinations of the logarithmic and the line functions, the complete gamma of slopes can be represented by the actual curves.

7 Conclusions

The energy consumption has to be considered as an additional constraint and, given the current ICT energy consumption growth trend, it will likely represent the major constraint in the designing of WDM network infrastructures, even more than the bandwidth capacity. In order to lower the energy consumption and the concomitant GHG emissions of such infrastructures, it is necessary to assess the power consumption of current and future energy-aware architectures through extensive energy models that characterize the behaviors of the network equipment. In this paper we presented and discussed the main energy and power models currently employed in the literature and provided an overview over the different scenarios that are currently being employed in WDM networks. Finally, we presented a comprehensive energy model which accounts for the foreseen energy-aware architectures and the grow rate predictions, including different types of traffic of a WDM networks. The model, based on real energy consumption values, tries to collect the main benefits of the previous models while maintaining low complexity and, thus, high scalability. We believe that such an energy model will help the development of new energy-oriented networks for achieving sustainable society growth and prosperity.

Acknowledgments. This work was supported in part by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.

References

1. Gartner press release 2007), `http://www.gartner.com/it/page.jsp?id=503867`
2. An inefficient Truth by the Global Action Plan, `http://www.globalactionplan.org.uk/upload/resource/Full-report.pdf`
3. SMART 2020: Enabling the low carbon economy in the information age, The climate group (2008)
4. Global Action Plan Report, An inefficient truth (2007), `http://www.globalactionplan.org.uk/`
5. Pileri, S.: Energy and Communication: engine of the human progress. In: INTELEC 2007 keynote, Rome, Italy (September 2007)
6. Souchon Foll, L.: TIC et Énergétique: Techniques d'estimation de consommation sur la hauteur, la structure et l'évolution de l'impact des TIC en France, Ph.D. dissertation, Orange Labs/Institut National des Télécommunications (2009)
7. BT Press, BT announces major wind power plans (October 2007), `http://www.btplc.com/News/Articles/Showarticle.cfm?ArticleID=dd615e9c-71ad-4daa-951a-55651baae5bb`
8. EU Spring Summit, Brussels (March 2007)
9. Telefónica supplement, The environment and climate change, 2008 special report on corporate responsibility (April 2009)
10. Lange, C.: Energy-related Aspects in Backbone Networks. In: Proc. ECOC 2009, Vienna, Austria (September 2009)
11. Tucker, R.S., et al.: Evolution of WDM Optical IP Networks: A Cost and Energy Perspective. IEEE/OSA Journal of Lightwave Technologies 27(3), 243–252 (2009)
12. Nature Photonics Technology Conference 2007, Tokyo, Japan (October 2007)
13. Saunders, H.D.: The Khazzoom-Brookes postulate and neoclassical growth. The Energy Journal (October 1992)
14. St Arnaud, B.: ICT and Global Warming: Opportunities for Innovation and Economic Growth, `http://docs.google.com/Doc?id=dgbgjrct\2767dxpbdvcf`
15. Chabarek, J., Sommers, J., Barford, P., Estan, C., Tsiang, D., Wright, S.: Power awareness in network design and routing. In: Proc. IEEE INFOCOM (2008)
16. Energy Star, Small network equipment, `http://www.energystar.gov/index.cfm?c=new_specs.small_network_equip`
17. Gupta, M., Singh, S.: Greening of the Internet. In: Proc. ACM SIGCOMM 2003, Karlsruhe, Germany (August 2003)

18. Muhammad, A., Monti, P., Cerutti, I., Wosinska, L., Castoldi, P., Tzanakaki, A.: Energy-Efficient WDM Network Planning with Protection Resources in Sleep Mode. Accepted for Globecom 2010, ONS01 (2010)
19. Vereecken, W., Van Heddeghem, W., Colle, D., Pickavet, M., Demeester, P.: Overall ICT footprint and green communication technologies. In: Proc. of ISCCSP 2010, Limassol, Cyprus (March 2010)
20. Juniper, http://www.juniper.net
21. Feng, M.Z., Hilton, K., Ayre, R., Tucker, R.: Reducing NGN Energy Consumption with IP/SDH/WDM. In: Proc. 1st International Conference on Energy-Efficient Computing and Networking, Passau, Germany, pp. 187–190 (2010) ISBN:978-1-4503-0042-1
22. Chiaraviglio, L., Mellia, M., Neri, F.: Energy-aware Backbone Networks: a Case Study. In: GreenComm – First International Workshop on Green Communications, Dresden, Germany (June 2009)
23. Van Heddeghem, W., De Groote, M., Vereecken, W., Colle, D., Pickavet, M., Demeester, P.: Energy-Efficiency in Telecommunications Networks: Link-by-Link versus End-to-End Grooming. In: Proc. of ONDM 2010, Kyoto, Japan, February 1-3 (2010)
24. Tucker, R.S.: Modelling Energy Consumption in IP Networks, http://www.cisco.com/web/about/ac50/ac207/crc_new/events/assets/cgrs_energy_consumption_ip.pdf (retrieved)
25. Moore, G.E.: Cramming more components onto integrated circuits. Electronics 38(8), April 19 (1965)
26. Gilder, G.F.: Telecosm: How Infinite Bandwidth Will Revolutionize Our World. The Free Press, NY (2000)
27. BONE project, WP 21 Topical Project Green Optical Networks: Report on year 1 and updated plan for activities, NoE, FP7-ICT-2007-1 216863, BONE project (December 2009)
28. Aleksic, S.: Analysis of Power Consumption in Future High-Capacity Network Nodes. Journal of Optical Communications and Networking 1(3), 245–258 (2009)

Analyzing Local Strategies for Energy-Efficient Networking

Sergio Ricciardi[1], Davide Careglio[1], Ugo Fiore[2], Francesco Palmieri[3],
Germán Santos-Boada[1], and Josep Solé-Pareta[1]

[1] CCABA, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] CSI, Università degli Studi di Napoli Federico II, Napoli, Italy
[3] DII, Seconda Università degli Studi di Napoli, Aversa, Italy
{sergior,careglio,german,pareta}@ac.upc.edu,
{fpalmier,ufiore}@unina.it

Abstract. Power management strategies that allow network infrastructures to achieve advanced functionalities with limited energy budget are expected to induce significant cost savings and positive effects on the environment, reducing Green House Gases (GHG) emissions. Power consumption can be drastically reduced on individual network elements by temporarily switching off or downclocking unloaded interfaces and line cards. At the state-of-the-art, Adaptive Link Rate (ALR) and Low Power Idle (LPI) are the most effective local-level techniques for lowering power demands during low utilization periods. In this paper, by modeling and analyzing in detail the aforementioned local strategies, we point out that the energy consumption does not depend on the data being transmitted but only depends on the interface link rate, and hence is throughput-independent. In particular, faster interfaces require lower energy per bit than slower interfaces, although, with ALR, slower interfaces require less energy per throughput than faster interfaces. We also note that for current technologies the energy/bit is the same both at 1 Gbps and 10 Gbps, meaning that the increase in the link rate has not been compensated at the same pace by a decrease in the energy consumption.

Keywords: sleep mode, energy-efficiency, power consumption, low power idle, adaptive link rate.

1 Introduction

Most of the currently known non-renewable primary energy sources are becoming scarcer and will get exhausted in only some decades. On the other hand, they are highly polluting as their burning process emits large quantity of GHGs causing climate changes and global warming phenomena. The current growth scenario is not sustainable, and international initiatives are trying to decrease the energy consumption and the GHG emissions by 20% for 2020 [1]. In order to achieve such drastic reductions, it is necessary to adopt a radical change in the current lifestyle and business as usual model. For such a transformation, the key factor is the use of energy-efficient processes together with energy-aware solutions and policies that

V. Casares-Giner et al. (Eds.): NETWORKING 2011 Workshops, LNCS 6827, pp. 291–300, 2011.

increasingly exploit renewable energy sources in providing all the public utility services. In the networking scenario, miniaturization and ICT growing dynamics, effectively described by the Moore's and Gilder's laws [2][3], have not had the expected counterpart in power consumption reduction. Miniaturization has reduced unit-power consumption but has allowed more logic ports to be put into the same space, thus increasing performances and, concomitantly, power utilization (a phenomenon called rebound effect already known as Jevons paradox [4]). As a consequence, the total power required per node is growing faster and faster. Nevertheless, traffic dynamics often result in a significantly different network usage which presents peaks alternated by low load periods, making room for power management techniques that, while satisfying the users' demand, exploit the low load periods for saving as much energy (and, thus, money) as possible. Accordingly, adaptive power management strategies that can be implemented independently and at different levels of granularity on each network device can be introduced at the local equipment-level to decrease power consumption in the operational phase and bring positive effects for the environment and significant cost savings. Power consumption can be drastically reduced by temporarily switching off or downclocking unloaded interfaces. In this work, such local energy containment strategies have been properly modeled and their behavior analyzed through simulations with the goal of better understanding their operating dynamics, strengths and weaknesses.

2 Active Local Strategies for Network Energy Efficiency

Experimental measurements collected from several network devices [5] show that in current architectures half of the energy consumption is associated to the base system and the other half to the number of installed line interface cards (even if *idle*). Furthermore, the power consumption of the actual electronic routing/switching matrix and line cards is, quite surprisingly, almost independent from the network load, so that the energy demand of heavily loaded devices is only about 3% greater than that of idle ones. These results suggest that it is necessary to develop energy-efficient architectures exploiting the ability of temporarily switching off or putting into energy saving mode devices or subsystems (e.g. switching fabrics, line cards, I/O ports, etc.) in order to minimize energy consumption whenever possible. Putting entire nodes into sleep mode (*per node sleep mode*) may be unpractical, especially for large and highly connected ones, since many very expensive transmission links become unused, hence negating significant capital investments (CAPEX) for the entire duration of the sleep interval. Furthermore, per node sleep mode drastically reduces the overall meshing degree, by limiting the network reliability and partially negates the possibility of balancing the load on multiple available links/paths. On the other hand, putting into sleep mode only single interfaces (*per interface sleep mode*) may introduce considerable energy savings in particular when operating at high speeds, since, for example, in a commercial off-the-shelf (COTS) Ethernet switch (Catalyst 2970 24-port LAN switch) a 1000baseT interface adds about 1.8 W to the overall consumption [6] (Table 1). Per-interface sleeping mechanisms (ALR and LPI) have been identified [6][7] as viable and effective solutions. In ALR, the ability to dynamically modify the link rate according to the real traffic needs is used as a technique to reduce the power consumption. Operating a device at a lower frequency can enable reductions in the energy consumption and also allows the use of dynamic voltage scaling (DVS) for reducing the operating voltage. This allows power to scale cubically and hence energy consumption quadratically with operating frequency [8].

Table 1. COTS switch power consumption with varying number of interfaces

# Active Interfaces	10BaseT	100baseTX	1000baseT
0	69.1 W	69.1 W	69.1 W
2	70.2 W	70.1 W	72.9 W
4	71.1 W	70.0 W	76.7 W
6	71.6 W	71.1 W	80.2 W
8	71.9 W	71.9W	83.7 W

For example, the Intel 82541PI Gigabit Ethernet Controller draining about 1 W at 1 Gbps full operation is able to support a smart power down feature by turning off PHY if no signal is present on link and drops the link rate to 10 Mbps when a reduction of energy consumption is required [9]. Also in the last mile, the ADSL2 standard (ITU G.992.3, G.922.4, G.992.5) is able to support multiple data rates corresponding to different link states (L0: full rate, L2: reduced rate, L3: link off) for power management sake [10]. In LPI, transmission on single interface is stopped when there is no data to send and quickly resumed when new packets arrive, in contrast with the continuous IDLE signal used in legacy systems. LPI defines large periods over which no signal is transmitted and small periods during which a signal is transmitted to synchronize the receiver. When operating in low-power mode, the elements in the receiver can be frozen, and then awakened within a few microseconds, as reported in Table 2 [11].

Table 2. Common wake-on-arrival strategy parameters for different interfaces technologies

Technology	**Wakeup Time**	**Sleep Time**	**Average Power savings**
100baseTX	30 μs	100 μs	90%
1000baseT	16 μs	182 μs	90%
10GbaseT	4.16 μs	2.88 μs	90%

Significant energy savings can be obtained when the involved devices spend a considerable fraction of their time in the low power mode. Although the savings vary from device to device, the energy consumption, when the device is in low power mode, can be as low as 10% that the one in active mode. During the transitions back and forth from low power mode there is a considerable increase in energy consumption as many elements in the transceiver have to be active. The actual value will depend on the implementation and possibly ranges from 50% to 100% of the active mode energy consumption. In network environments where packet arrival rates can be highly non-uniform, allowing interface transitions between different operating rates or sleep/active modes can introduce additional packet delay, or even loss, due to the associated transition times. The main issues to be addressed are the coordination among nodes during the transitions from and to a low power consumption state or from a transmission rate to another one. In line of principle, these transitions should be kept as transparent as possible to upper layer protocols and applications. Several solutions can usefully exploit the tradeoff between potential energy savings, performance and transparency. For example, buffering, packet coalescing and coordinated Ethernet strategies may be introduced, to collects packets into small bursts and thereby creating gaps long enough to profitably sleep [6][12][13]. Potential concerns are that buffering will add too much delay across the network and that traffic burstiness will exacerbate the loss.

3 Modeling and Analyzing Local Energy Containment Methods

In this section, we present a model of the aforementioned local techniques built by interpolating realistic data obtained from the available literature and experimental measurement on available state-of-the-art hardware. We exploited and analyzed through simulation some of the most interesting properties and operational features of these techniques when applied to individual non-cooperating network devices. Let *G(V,E)* be a directed graph representing the physical network topology; *V* the set of vertices that represent the network nodes and *E* the set of edges that represent the network links. Note that, as a (unidirectional) link is attached to each interface, the set of links *E* actually coincides with the set of interfaces. Each interface has its own native speed: $\forall i \in E$, $v_i \in R=\{10$ Mbps, 100 Mbps, 1000 Mbps, 10000 Mbps$\}$ represents the *native link rate* of interface *i*. The energy/power consumption of interfaces working at their native link rates [6][14] are illustrated in Table 3.

Table 3. Energy and power consumption of interfaces working at native speeds

Native link rate v_i	**Power per interface**	**Energy Scaling Index (ESI)** - Energy per bit	**Energy Consumption Rate (ECR)** - Power per Gbps
10 Mbps	0.1 W	10 nJ/bit	10 W/Gbps
100 Mbps	0.2 W	2 nJ/bit	2 W/Gbps
1,000 Mbps	0.5 W	0.5 nJ/bit	0.5 W/Gbps
10,000 Mbps	5.0 W	0.5 nJ/bit	0.5 W/Gbps

ESI and ECR are different energy/power consumption metrics that may be reduced to equivalent values, in fact it holds that: W/Gbps = (J/s) / (Gbit/s) = J/Gbit = nJ/bit.

Scaling the energy consumption per bit (ESI metric) reveals that the energy consumption for forwarding one bit is not the same for every interface but depends on its native link rate. In particular, the *energy per bit* is lower for faster interfaces, meaning that forwarding one bit on a slower interface requires more energy than on a faster one (besides occupying the link resource for a longer time). We also note how the energy/bit ratio is the same both at 1 Gbps and 10 Gbps, that is, there is no gain in the energy/bit at 10 Gbps (as instead occurs when switching between 10/100 Mbps and between 100/1000 Mbps). This behavior is due to the current 10 Gbps technology, whose increase in the link rate (achieved through advanced modulation techniques [15]) has not been compensated at the same pace by a decrease in the energy consumption. As a result, 10 Gbps interfaces consume 10 times more energy than 1 Gbps ones, i.e. the power consumption scales linearly from 1 to 10 Gbps. Consequently, the best balance between power consumption and bit rate is reached at 1 Gbps (see Fig. 1). This situation is further stressed when the throughput is not equal to the link rate, which corresponds to an underutilized channel. Our observations confirm that an interface consumes the same power whatever its *current* throughput is: power consumption is *throughput-independent*. For this reason, the link rate can be adapted to the current throughput by using ALR with consequent energy savings. However, the IEEE Energy Efficient Ethernet working group, when analyzing the opportunity to adopt ALR or LPI in the 802.3az standard, decided in favor of LPI [16] since the two strategies have been considered as alternative to be included in the standard. Instead, we evaluate the advantages offered by a combination of them and advocate the complementary use of the two strategies.

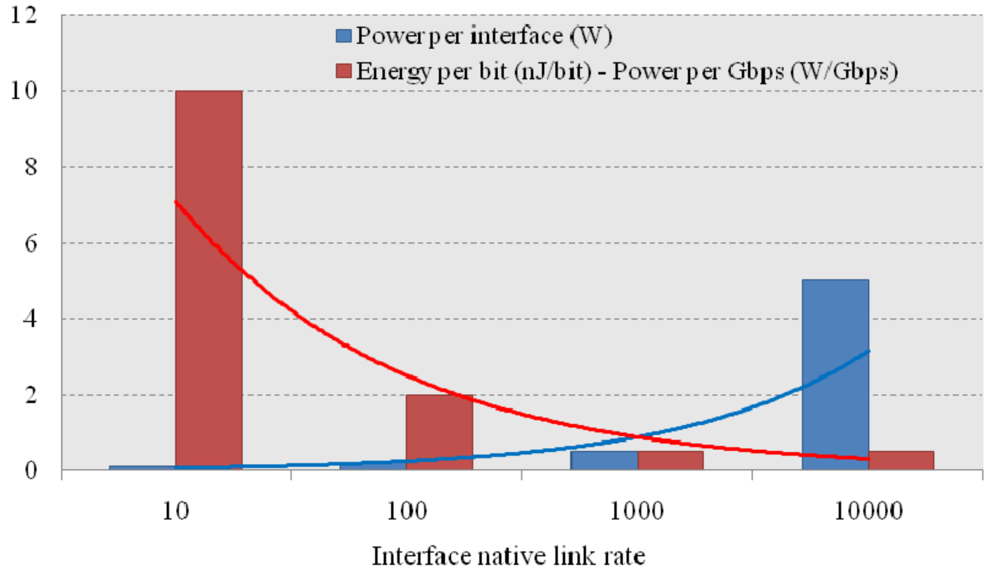


Fig. 1. Energy and power consumption of different interfaces working at their native speeds

To model the ALR, let us consider the set R={r_1, …, r_m} set of available link rates; $\forall i \in E$, $r_i \in R$ represents the *working link rate*, that is the link rate at which the interface i is currently operating; obviously, it holds that $\forall i \in E$, $r_i \leq v_i$. In the actual standard an interface may switch only to the set of existing link rates, thus R={10 Mbps, 100 Mbps, 1000 Mbps, 10000 Mbps}. According to current technologies, we consider three possible operating modes for interfaces: (1) Off, occurs when the interface is *down*; (2) LPI, interface up in low power mode and there is no data to transmit; (3) ALR, there is data to transmit and the interface is up at working rate $r_i \in R$. Let us consider an interface i with native link rate v_i and a constant data throughput t_d; then, with ALR, the interface will switch its current link rate to r_i: $r_{i-1} < t_d \leq r_i$. In our simulations we found that, when using the ALR, the power consumption of an interface i depends not only on the working link rate r_i but also on the native link rate v_i. In other words, transmitting a fixed data throughput t_d has different power consumption depending on the interface native link rate v_i: in this case, slower interfaces consume less power than faster ones for the same throughput t_d, even if they work at the same rate r_i. This result, quite surprising if we consider that slower interfaces consume more energy per bit than faster ones, may be explained considering that the different technologies adopted for reaching higher link rates [15] lead to greater fixed power consumption for faster interfaces. In fact, as routers, also the interfaces have fixed and variable power consumption. The fixed part is always present just for the interface to stay up and accounts for the control circuits, while the variable traffic-proportional power consumption is due to the transceivers. In the following we model such energy consumptions and show a breakdown of the different energy components in a 10 Gbps interface.

Table 4. Power consumptions of interfaces working at different rates in $\{10,10^2,10^3,10^4\}$ Mbps

v_i	r_i	**Mbps**	**Power consumption**
$\forall v_i \in R$	$Off\,/\,r_0$	0/0	$\Psi(v_i, Off\,/\,r_0) \cong 0^*$
v_1: 10	r_1	10	$\Psi(v_1,r_1)$
v_2: 10^2	r_1/r_2	$10/10^2$	$\Psi(v_2,r_1)/\Psi(v_2,r_2)$
v_3: 10^3	$r_1/r_2/r_3$	$10/10^2/10^3$	$\Psi(v_3,r_1)/\Psi(v_3,r_2)/\Psi(v_3,r_3)$
v_4: 10^4	$r_1/r_2/r_3/r_4$	$10/10^2/10^3/10^4$	$\Psi(v_4,r_1)/\Psi(v_4,r_2)/\Psi(v_4,r_3)/\Psi(v_4,r_4)$

* In LPI, the device only sends signals during short refresh intervals and stays quite during large intervals so the power consumption in the LPI mode is almost 0.

In general, to model the fixed and the variable power consumption, we define $\{\Psi(v_i,r_j) \mid j = 1,2,\ldots,m\}$ where $\Psi(v_i, r_j)$ is the power consumption of the interface $i\in E$ with native speed $v_i\in R$ operating at link rate $r_j\in R$ and $\Psi(v_i, r_j)< \Psi(v_i, r_k)$ $\forall j<k$. Also, we define Θ_n as the fixed power consumption of node $n\in V$ accounting for its base system, switching matrix, control circuits, etc. Note that Θ_n does not include the power consumption of the node interfaces, which is given by the Ψ term. Let's see how the term Ψ models the power consumption of the interfaces. Let i, $j\in E$ be two interfaces with respectively native and working link rates (v_i, r_i) and (v_j, r_j). We can observe that Ψ can be characterized from the following properties:

$$\Psi(v_i,r_i): R\times R \to \Re \qquad (v_i,r_i)\mapsto \Psi(v_i,r_i)\in \Re \tag{1}$$

$$\Psi(v_i,r_i) \propto v_i, \forall i\in E \tag{2}$$

$$\Psi(v_i,r_i) \propto r_i, \forall i\in E \tag{3}$$

Where (1) is the functional definition of Ψ, (2) and (3) state the proportionality of Ψ to the native and working link rates respectively;

$$\forall i\in E: (r_i = off \vee r_i = r_0) \Leftrightarrow \Psi(v_i,r_i) \cong 0 \tag{4}$$

the energy consumption of an interface Off or in LPI is nearly 0;

$$\forall i,j\in E:\ (v_i < v_j \wedge r_i = r_j) \Rightarrow \Psi(v_i,r_i) < \Psi(v_j,r_j) \tag{5}$$

two interfaces with the same working link rate but different native link rates have different power consumptions;

$$\forall i, j \in E: (v_i = v_j \wedge r_i < r_j) \Rightarrow \Psi(v_i, r_i) < \Psi(v_j, r_j) \tag{6}$$

two interfaces with the same native link rate but different working link rates have different power consumptions;

$$\forall i, j \in E: \left(\frac{r_i}{v_i} = \frac{r_j}{v_j} \wedge (r_i \neq r_j \vee v_i \neq v_j) \right) \Rightarrow \Psi(v_i, r_i) \neq \Psi(v_j, r_j) \cdot \tag{7}$$

two interfaces with the same *r*/*v* ratio, but with different native/working link rates, have different energy consumptions. In order to model the interfaces power consumption in a realistic case, we first consider a two-level system, that is a system in which interfaces may work only at two link rates: high and low power modes. The total energy consumption is therefore given by the sum of the energy cost spent in low power mode plus the cost in high mode. That is, the sum of the low-power mode instantaneous cost times the total time spent in low-power mode plus the high-power mode instantaneous cost times the total time spent in high-power mode. In the general case, when more than one link rate is available, we divide the time in intervals so that the link rate stays constant during each interval and record the duration of each interval. We indicate as *N* the number of time intervals with unchanging state so that there are *N*-1 link rate transitions with changing states; if t_i is the duration of the *i*-th time interval (in seconds) then the total time considered is given by:

$$T = \sum_{i=1}^{N} t_i. \tag{8}$$

Let r_i be the link rate in the *i*-th time interval; τ the time needed for the link rate transition (assume that every transition requires the same time); c_j the instantaneous power consumption at link rate *j*; ζ a proportionality constant between the instantaneous power demand c_j and the corresponding link rate *j* so that $c_j = \Theta + \zeta \cdot j$; X_{hk} the power consumption when transitioning from link rate *h* to link rate *k.* Then, for a single interface:

$$\begin{aligned} \Psi = \sum_{i=1}^{N} c_{r_i} t_i + \sum_{i=2}^{N} \tau X_{r_{i-1} r_i} = \Theta \sum_{i=1}^{N} t_i + \zeta \sum_{i=1}^{N} r_i t_i + \tau \sum_{i=2}^{N} X_{r_{i-1} r_i} = \\ \Theta T + \zeta \cdot \bar{r} \cdot T + \tau \sum_{i=2}^{N} X_{r_{i-1} r_i} = (\Theta + \zeta \cdot \bar{r}) T + \tau \, (N-1) \bar{X}. \end{aligned} \tag{9}$$

For sake of simplicity, we consider all interfaces to behave in the same way. In the realistic hypothesis [6][17] that $X_{hk} \propto c_{max\{h,k\}} = \zeta \cdot max\{h,k\}$, $\bar{X} \propto \bar{r}$ and eq. (**9**) becomes:

$$\Psi \approx (\Theta + \bar{r}\zeta)(T + N\tau). \tag{10}$$

Starting from the above energy model, combined with several real energy consumption observations available in literature [6][9][14][17][18], we simulated some native speed interfaces working at different link rates. The associated per bit energy consumption values are shown in the chart of Fig. 2. We can see how the energy per bit depends both from the native link rate of the interfaces and on their actual working link rate. Furthermore, we can notice how the energy consumption of native high-speed interfaces does not vary much when switching to lower link rates, whilst the energy consumption of native low-speed interfaces is highly variable, especially when working at low rates.

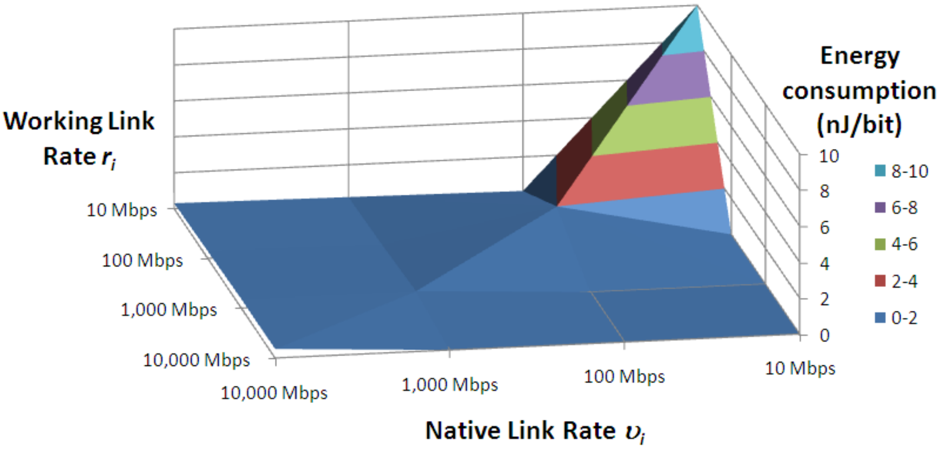


Fig. 2. Energy per bit for native interfaces operating at different link rates (interpolation obtained by putting not defined values to 0)

In Fig. 3 we plotted the power consumption breakdown for a simulated routing device modeled with interfaces at 10 Gbps. The base systems accounts for approximately 50% of the total energy consumption while the interfaces (fixed and variable parts) accounts for the other half.

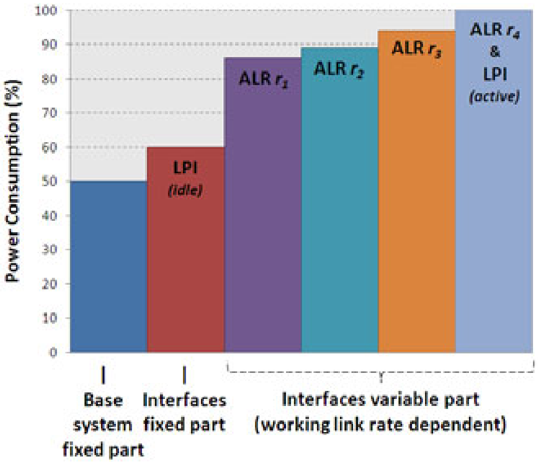


Fig. 3. Power consumption breakdown for a router with interfaces at native link rate v = 10 Gbps and working link rates $r_i = 10^i$ Mbps

As we can see, the larger interfaces energy consumption is due to the fixed part that is independent from the link rate, while only the remaining 15% of energy is due to the rate-proportional energy consumption. This scenario suggests that between LPI and ALR it is preferable to use LPI: when there is no data to transmit, ALR sends the continuous idle signal (which is everything but *idle*), whilst LPI enters the lower power consumption mode, thus consuming lower energy; furthermore, the auto-negotiation mechanism of ALR may be in the order of *ms*, while the sleep/wake-up transitions of LPI only requires some *μs* for 10 Gbps interfaces [17][19] and, consequently, lower delay; finally, transmitting the same amount of data in ALR takes longer time than LPI, since in LPI the transmission is realized at the maximum rate, instead in ALR the transmissions is lowered to best fit the throughput, thus occupying the resources for longer time. The Energy-Efficient Ethernet working group has recently adopted the LPI as power management solution for IEEE 802.3az [16]. Furthermore, the difference in energy consumption between interfaces with different native link rates suggests also the possibility for and advanced ALR with circuit over-provisioning, i.e. a network interface may be provisioned with different circuits, say a low and a high speed one, and may switch between one or other according to the required data throughput. This solution, at the expense of increased capital expenditures (CAPEX) for the additional hardware, may lead to decreased operational costs (OPEX) due to the lower fixed power consumption of slow circuits. Finally, also another possibility is given by the heterogeneity of the equipment in a network. In fact, such heterogeneity may be exploited by a global load-balancing schema, implemented as part of a routing and wavelength assignment algorithm, which tries to distribute the connection requests in such a way that the overall network energy consumption is minimized. A best-fit allocation scheme may be implemented in order to close match the bandwidth demands with the interfaces native link rates, so that the fixed power consumption cost is amortized and the maximum efficiency is reached.

4 Conclusions

By modeling and analyzing the available local energy efficiency strategies, we observed that faster interfaces consume less energy per bit than slower ones, but also that, lowering the interfaces working link rates during low utilization periods, does not lead to the same power savings for all interfaces but depends on the interface *native* link rate. Native slower interfaces (e.g., 100 Mbps) consume less power than native faster interfaces (e.g., 1,000 Mbps) for transmitting the same throughput (e.g., 80 Mbps) due to the higher fixed power consumption that comes with faster interfaces. In other words, while the *energy-per-bit* is lower for faster interfaces, with the ALR the *energy-per-throughput* is lower for slower interfaces. Furthermore, we observed that the energy consumption of native high-speed interfaces does not vary much when switching to lower link rates, whilst the energy consumption of native low-speed interfaces is highly variable, especially when working at low rates. Finally, we point out that the different fixed and variable power consumptions of interfaces may be exploited by circuit over-provisioning techniques as well as load balancing schemes for minimizing the overall energy consumption and, thus, network operational costs.

Acknowledgments. This work was supported in part by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of

Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.

References

[1] EU Spring Summit, Brussels (March 2007)

[2] Moore, G.E.: Cramming more components onto integrated circuits. Electronics 38(8) (April 19, 1965)

[3] Gilder, G.F.: Telecosm: How Infinite Bandwidth Will Revolutionize Our World. The Free Press, NY (2000)

[4] Jevons, W.S.: The Coal Question; An Inquiry concerning the Progress of the Nation, and the Probable Exhaustion of our Coalmines. Macmillan and Co., Basingstoke (1866)

[5] Chabarek, J., Sommers, J., Barford, P., Estan, C., Tsiang, D., Wright, S.: Power awareness in network design and routing. In: Proc. IEEE INFOCOM (2008)

[6] Christensen, K., Nordman, B.: Reducing the Energy Consumption of Networked Devices. In: IEEE 802.3 tutorial (July 19, 2005)

[7] Hays, R.: Active/Idle Toggling with 0BASE-x for Energy Efficient Ethernet. Presentation to IEEE 802.3az Task Force (November 2007)

[8] Zhai, B., Blaauw, D., et al.: Theoretical and Practical Limits of Dynamic Voltage Scaling. In: DAC 2004 (2004)

[9] Intel® 82541PI Gigabit Ethernet Controller, Intel White Paper, http://www.intel.com/design/network/products/lan/controllers/82541pi.htm

[10] Tzannes, M.: ADSL2 Helps Slash Power in Broadband Designs. CommDesign.com (January 30, 2003)

[11] Reviriego, P., et al.: Performance Evaluation of Energy Efficient Ethernet. IEEE Commun. Letters 13(9) (September 9, 2009)

[12] Kubo, R., Kani, J., Fujimoto, Y., Yoshimoto, N., Kumozaki, K.: Sleep and adaptive link rate control for power saving in 10GEPON systems. In: Proc. Globecom (2009)

[13] Nedevschi, S., Popa, L., Iannaccone, G., Ratnasamy, S., Wetherall, D.: Reducing network energy consumption via sleeping and rate adaptation. In: Proc. 5th USENIX Symp. Networked Systems Design and Implementation, NSDI 2008 (April 2008)

[14] BONE project. WP 21 Topical Project Green Optical Networks: Report on year 1 and updated plan for activities. NoE, FP7-ICT-2007-1 216863 BONE project (December 2009)

[15] Nortel. A comparison of next-generation 40-Gbps technologies, white paper, http://www.nortel.com/solutions/collateral/nn122640.pdf

[16] IEEE P802.3az. Energy Efficient Ethernet Task Force (2010), http://grouper.ieee.org/groups/802/3/az

[17] Christensen, K., Reviriego, P., Nordman, B., Bennett, M., Mostowfi, M., Maestro, J.A.: IEEE 802.3az: The Road to Energy Efficient Ethernet. IEEE Communications Magazine (November 2010)

[18] Ricciardi, S., Careglio, D., Palmieri, F., Fiore, U., Santos-Boada, G., Solé-Pareta, J.: Energy-aware RWA for WDM networks with dual power sources. In: Proc. IEEE International Conference on Communications (ICC 2011), Kyoto, Japan, June 5-9 (2011)

[19] Zhang, B., Sabhanatarajan, K., Gordon-Ross, A., George, A.: Real-Time Performance Analysis of Adaptive Link Rate. In: Proc. Conference on Local Computer Networks (October 2008)

Energy-Aware RWA for WDM Networks with Dual Power Sources

Sergio Ricciardi[1], Davide Careglio[1], Francesco Palmieri[2], Ugo Fiore[3], Germán Santos-Boada[1], Josep Solé-Pareta[1]

[1] Advanced Broadband Communications Center, Universitat Politècnica de Catalunya, Barcelona, Spain, sergior@ac.upc.edu
[2] Information Engineering Department, Seconda Università di Napoli, Aversa, Italy, fpalmier@unina.it
[3] University Centre of Computer Science Services, Università di Napoli Federico II, Naples, Italy, ufiore@unina

Abstract—**Energy consumption and the concomitant Green House Gases (GHG) emissions of network infrastructures are becoming major issues in the Information and Communication Society (ICS). Current optical network infrastructures (routers, switches, line cards, signal regenerators, optical amplifiers, etc.) have reached huge bandwidth capacity but the development has not been compensated adequately as for their energy consumption. Renewable energy sources (e.g. solar, wind, tide, etc.) are emerging as a promising solution both to achieve drastically reduction in GHG emissions and to cope with the growing power requirements of network infrastructures.**

The main contribution of this paper is the formulation and the comparison of several energy-aware static routing and wavelength assignment (RWA) strategies for wavelength division multiplexed (WDM) networks where optical devices can be powered either by renewable or legacy energy sources. The objectives of such formulations are the minimization of either the GHG emissions or the overall network power consumption. The solutions of all these formulations, based on integer linear programming (ILP), have been observed to obtain a complete perspective and estimate a lower bound for the energy consumption and the GHG emissions attainable through any feasible dynamic energy-aware RWA strategy and hence can be considered as a reference for evaluating optimal energy consumption and GHG emissions within the RWA context. Optimal results of the ILP formulations show remarkable savings both on the overall power consumption and on the GHG emissions with just 25% of green energy sources.

I. INTRODUCTION

The energy consumption and the concomitant GHG emissions (mainly CO_2) are becoming more and more a sensible issue for the ICS, governments and standardization bodies [1]. The Kyoto protocol imposes on industrialized States to reduce their GHG emissions by 5% from the 1990 level in the 2008-2012 period. It has been estimated [2] that network infrastructures alone consume 22 GW of electrical power corresponding to more than 1% of the worldwide electrical energy demand, with a growth rate of 12% per year, further stressing the need for energy-efficient network devices and energy-aware protocols and algorithms. In fact, the solely deployment of energy-efficient devices is not enough, as their total cost of ownership (TCO) decreases, the demand for using such devices increases and the gained benefits are overcome by greater energy consumption and concomitant GHG emissions. Such a phenomenon is known as *rebound effect* (or, in other contexts, as Jevons paradox or Kazzoom/Brookes postulate [3]). In order to overcome the rebound effect, it is necessary to adopt the *carbon neutrality* or, when available, the *zero carbon* approach. In carbon neutrality, GHGs emitted by legacy (dirty) energy sources (e.g. fossil-based plants) are compensated – hence, neutrality – by a credit system like the cap and trade or the carbon offset [3]. In the zero carbon approach, renewable (green) energy sources (e.g. sun, wind, tide) are employed and no GHGs are emitted at all. Clearly, green energy sources are always preferable with respect to the dirty ones as they limit (or avoid at all) GHG emissions, although renewable sources are variable in nature and their availability may change in time. Therefore, we advocate an energy system in which network elements (NE) are provided with green energy sources alongside the legacy power system and, at the occurrence, they are able to switch to the fossil-based power supply without any energy interruption. Such NEs are *energy-aware* as they adapt their behavior and performance depending not only on the current load but also on the source of energy that is supplying them.

Several approaches to achieve energy efficiency in network infrastructures have been proposed in the literature [4][5][6][7][8][9][10], but, at the state-of-the-art, none of them takes into consideration green and dirty energy sources for all the NEs together with the energy requirements of the different traffic types (optical/electronic, pass-through, add/drop, amplification, 3R regeneration, etc.). Furthermore, all prior works are focused on the reduction of the network energy consumption by switching off network elements. However, the power drawn by NEs is assumed to be given and thus it is not derived from a realistic energy model. In addition, putting into sleep mode entire NEs is not the sole possible solution and has its drawbacks. In [7] several ILP formulations for optical network planning are illustrated, and results show that switching off network nodes is not practically feasible as it was possible only in few experiments for very low loads. Furthermore, putting into sleep mode one big router is economically unviable and technically immature, at least with the architectures and technologies currently employed. A router is a rather costly piece of equipment and it still takes minutes to "wake up"; when returning from sleep mode, a peak in the power consumption is registered and the lifetime of the router will decrease if frequent power up/down cycles occur. Moreover, routing operating systems are not so stable and additional manual configurations may be needed at each power

978-1-61284-233-2/11/$26.00 ©2011 IEEE

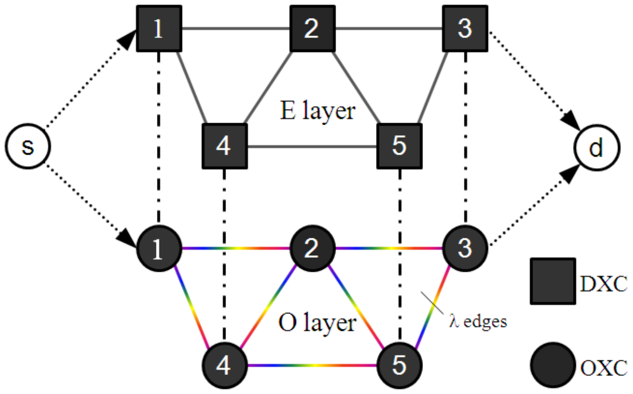


Fig. 1. Schematic representation of the network model with the electronic (E) and optical (O) layers.

up. Finally, powering off routers in a network results in reduced load balancing, as the traffic flows have to be routed only on the subset of active routers.

To the best of our knowledge, this paper represents the first work in which (1) green and dirty energy sources are explicitly considered for all the NE types and (2) sleep mode is assumed to be not available, with energy savings coming exclusively from the optimal routing of the connection requests. Specifically, we tried to combine all the notable features that a comprehensive dual energy-aware network model should have and put them together into a general Routing and Wavelength Assignment (RWA) framework. In doing this, all the energy-related information and concepts associated to devices and links have been abstracted and defined in a formal and concise way within the context of a specific energy-aware network optimization problem. To achieve a more compact and precise representation of such a problem we formally model it as an ILP problem working on the physical network topology comprising routers and links, in which each device is characterized by a known power consumption, varying under different loads, energy source, capacity, and cost. The energy model defining the energy requirements of each NE considers electrical and optical technologies, and differentiates the consumption according to the various flavors of NEs and traffic types. We also assume to have the complete knowledge of the average amount of traffic exchanged by any source/destination node pair. The proposed formulation is applied to a multilayer IP/WDM network with the twofold aim of minimizing the overall GHG emissions and the power consumptions by setup optimal lightpaths through a suitable energy-aware RWA scheme. Solving the above ILP requires knowledge in advance of the entire traffic matrix and hence restricts us to the static, offline, RWA case. In turn, in a dynamic scenario, the optimal solution of the ILP represents a lower bound for the GHG emissions and energy consumptions of energy-aware RWA heuristics.

II. Network & Energy Model

A. The Network Model

We consider wavelength-routed networks in which the traffic unit is the lightpath. Network nodes may be electronic routers (digital cross connects, DXC) or optical switches (optical cross connects, OXC) connected by fiber links with up to λ wavelengths on each. The network is represented as a multigraph with one edge for each wavelength λ in the optical layer (Fig. 1). We assume that the traffic is unsplittable in the optical domain: i.e. a traffic demand is routed over a single lightpath; (in theory, in the electronic domain a demand may be splitted into *n* flows, but in the optical domain these will appear as *n* unsplittable optical flows). The *type* of traffic depends on the NE that is being traversed, thus three types of traffic are possible: (1) opaque electronic traffic; (2) transparent optical traffic with wavelength conversion (WC) and (3) transparent optical traffic without WC. The Table I reports the types of network element and the corresponding supported traffic types. Note that each type of traffic accounts for different power consumption when traversing NEs, as explained in following subsections. We assume that all the nodes have the possibility to convert wavelengths, either in the electronic or in the optical domain, depending on their technology. In the electronic domain, the full range of operations is supported: wavelengths routing/switching, wavelengths add/drop, WC, 3R regeneration; in the optical domain the operations supported are the transparent wavelength switching/pass-through and the WC. NEs may be powered either by green or dirty energy sources statically assigned to each at the network topology definition time. We assume that each node is able to distinguish which power source is currently feeding it through an energy-aware GMPLS-like control plane intelligence. This corresponds to a scenario in which network infrastructure planners consider the construction of new portions of the network or the change of power source for existing parts, and they evaluate the reduction in CO_2 emissions against other issues such as the technical aspects and costs of using green or dirty energy sources.

TABLE I
Traffic Supported by the Devices

Type of Device (NE)	Type of Traffic
Electronic Router [a]	Electronic
Optical Switch [b] w/ WC	Optical (with or without WC)
Optical Switch [b] w/o WC	Optical (without WC)
Fiber Optic	Optical (without WC)
Optical Amplifier	Optical (without WC)
3R regenerator	Electronic

[a] DXC; [b] OXC.

B. The Energy Model

In this model we explicitly considered the influence of traffic on power consumption by using realistic data for traffic demands, network topologies, link costs, and energy requirements of single network elements. Specifically, the amount of power consumed by the NEs depends on the type of device and on the type and load of traffic that it is currently supporting (e.g. an OXC may support transparent optical traffic with or without WC). Even though the energy consumption of current node architectures does not scale with traffic (the energy demand of heavily loaded devices is only 3% greater than that of idle ones [12]), energy-aware architectures that adapt their performances to the traffic load lowering the power requirement under low traffic loads are strongly advised and are being designed [1][12]. Consequently, we assume that the power consumption of the NEs, i.e. both network nodes and links, consists of two factors. When turned *on*, a NE consumes a constant amount of power (*fixed* power Θ) depending on the router size and technology (measured in *J/s=W*) and independent on the traffic load. The second factor (*proportional* power ε) consists of an amount of power proportional to type and quantity of the traffic load (measured in *nJ/bit* or, equivalently, in *W/Gbps*). The overall power drained by the WDM network is thus given by the sum of the fixed and proportional powers of the NEs subject to the current traffic load and varies with the routing of the connection requests. This implies that, as the NEs are always turned on, the routing optimization process works "only" on the proportional power.

Table II reports the mean values of NEs proportional energy consumptions. As we can see, electronic traffic, i.e. traffic that undergoes O/E/O conversion, consumes more power than optical traffic. In the electronic domain the energy necessary for forwarding 1 unit of traffic is 150 times greater than the energy needed in the optical domain w/o WC and 50 times greater if WC is available in the optical domain. Although they are mean values, in our model each NE has its own particular energy consumption factor – resulting from the individual architecture, technology and configuration – that have been obtained by further elaborations of the real measurements in [2] and account for fixed and proportional power consumptions which is a significant portion of the total consumption, according to [4].

III. Energy-aware routing and wavelength assignment

In this section, the problem of energy-aware RWA in WDM networks with dual power sources has been formulated as ILP formulations with different objective functions. In subsection III.A, the problem of minimizing the overall GHG emissions (*MinGas-RWA*) is presented, whilst the problem of minimizing the overall network power consumption (*MinPower-RWA*) is discussed in subsection III.B. To evaluate the power effectiveness and the reduction in GHG emissions of the two previous strategies, minimum cost RWA (*MinCost-RWA*) – i.e. energy-unaware RWA – is presented in subsection III.C.

A. Energy-aware RWA at minimum GHG emissions (MinGas-RWA)

The energy-aware RWA in WDM networks with dual power sources (*MinGas-RWA*) is formalized as an ILP problem. The objective is to route the requested lightpaths so that the overall network GHG emissions are minimized. Only NEs powered by dirty energy sources emit GHGs, whilst NEs powered by green energy sources do not emit any GHG at all, according to the energy model discussed in Section II.B. The ILP problem can be mathematically formulated as follows.

TABLE II
Mean Proportional energy consumption scaling factors of different routing/switching technologies[1]

Router Technology	Energy Consumption Rate (ECR)	Energy Scaling Index (ESI)	*P(B)*
Electronic DXC	1.5 W/Gbps	1.5 nJ/bit	$P = 1.5\cdot B$
Optical OXC w/ WC	0.03 W/Gbps	0.03 nJ/bit	$P = 0.03\cdot B$
Optical OXC w/o WC	0.01 W/Gbps	0.01 nJ/bit	$P = 0.01\cdot B$

P: Power consumption function; *B*: Aggregated bandwidth;

[1] ECR and ESI are different power consumption metrics that may be reduced to equivalent values: W/Gbps = (J/s)/(Gbit/s) = J/Gbit = nJ/bit.

Input parameters (data)

- *G(V,E)*: directed graph representing the physical network topology; *V* set of vertices that represent the network nodes; *E* the set of edges that represent the network links; $|V| = N$, $|E| = M$;
- a_{ij} : number of wavelengths available on link *(i, j)*;
- ℓ_{ij} : length of link *(i,j)* (in *km*);
- Λ: maximum length (in *km*) of links without need of amplification (usually 80/100 *km*);
- t^{sd}: number of lightpaths to be established from *s* to *d*; i.e. $\{t^{s,d}\}_{s,d\in V}$ is the traffic matrix;
- $\pi^{sd,k}$: *k*-th pre-computed route from *s* to *d*;
- $\rho^{sd,k}$: the geographical length of route $\pi^{sd,k}$ (in *km*);
- Θ_n : fixed power of node *n;* depends on node size/type;
- $\varepsilon_n^{t_1}$: proportional power for transporting one lightpath as *transparent pass-through* traffic at node *n*;
- $\varepsilon_n^{t_2}$: proportional power for transporting one lightpath as *opaque pass-through* traffic at node *n* (e.g. 3R regeneration or opaque wavelength conversion);
- $\varepsilon_n^{t_3}$: proportional power for *add/drop* one lightpath at node *n*;
- Ψ_{ij} : fixed power for devices in link *(i, j)*, (e.g. optical amplifiers); among 3 and 15 W;
- δ_{ij} : proportional power for transporting one lightpath through link *(i, j)*; it is assumed that each device (e.g. OA) on the same link *(i, j)* has the same fixed and proportional power consumption;
- $x_n^{sd,k}$ identifies the presence of O/E/O conversion at the node *n*:

$$x_n^{sd,k} = \begin{cases} 1 & \text{if } n \in \pi^{sd,k} \text{ and } \pi^{sd,k} \text{ undergoes O/E/O} \\ & \text{conversion at node } n \\ 0 & \text{if } n \notin \pi^{sd,k} \text{ or } \pi^{sd,k} \text{ transparently passes} \\ & \text{through node } n \end{cases}$$

Note that 3R regeneration and opaque wavelength conversion are implicitly considered in this matrix and this information will be used in the power consumption calculus.

- the following node attributes model the energy source:

$$g_n = \begin{cases} 1 & \text{if node } n \text{ is powered by a green energy source} \\ 0 & \text{if node } n \text{ is powered by a dirty energy source} \end{cases}, \forall n \in V$$

$$h_{ij} = \begin{cases} 1 & \text{if link } \mathit{(i,j)} \text{ is powered by a green energy source} \\ 0 & \text{if link } \mathit{(i,j)} \text{ is powered by a dirty energy source} \end{cases}, \forall (i,j) \in E$$

Variables

- integer $w^{sd,k}$ indicates the number of lightpaths using route $\pi^{sd,k}$ (on the same route there may be several lightpaths using different wavelengths);
- $PC_{G(V,E)}$ indicates the objective function to be minimized;
- $TC_{G(V,E)}$ indicates the overall power consumption of the NEs in *G(V,E)* evaluated in the chosen traffic model;
- $GC_{G(V,E)}$ indicates the power consumption of the NEs in *G(V,E)* due only to green power sources.

Objective function

$$\mathit{Minimize} \quad PC_{G(V,E)} \tag{1}$$

Constraints

$$PC_{G(V,E)} = (TC_{G(V,E)} - GC_{G(V,E)}) + \log TC_{G(V,E)} \tag{2}$$

$$TC_{G(V,E)} = \sum_{n\in V}\left(\begin{array}{l} \Theta_n + \varepsilon_n^{t_1} \cdot \sum\limits_{sd,k:n\in\pi^{sd,k},n\neq s,d} w^{sd,k}\cdot(1-x_n^{sd,k}) + \\ \varepsilon_n^{t_2}\cdot \sum\limits_{sd,k:n\in\pi^{sd,k},n\neq s,d} w^{sd,k}\cdot x_n^{sd,k} + \varepsilon_n^{t_3}\cdot \sum\limits_{sd,k:n=s,d} w^{sd,k} \end{array}\right) \tag{3}$$

$$+ \sum_{(i,j)\in E}\left(\left\lfloor \frac{\ell_{ij}}{\Lambda}\right\rfloor \cdot \left(\Psi_{ij} + \delta_{ij}\cdot \sum_{sd,k:(i,j)\in\pi^{sd,k}} w^{sd,k}\right)\right)$$

$$GC_{G(V,E)} = \sum_{n\in V} g_n \cdot \left(\begin{array}{l} \Theta_n + \varepsilon_n^{t_1} \cdot \sum\limits_{sd,k:n\in\pi^{sd,k},n\neq s,d} w^{sd,k}\cdot(1-x_n^{sd,k}) + \\ \varepsilon_n^{t_2}\cdot \sum\limits_{sd,k:n\in\pi^{sd,k},n\neq s,d} w^{sd,k}\cdot x_n^{sd,k} + \varepsilon_n^{t_3}\cdot \sum\limits_{sd,k:n=s,d} w^{sd,k} \end{array}\right) \tag{4}$$

$$+ \sum_{(i,j)\in E} h_{ij}\cdot\left(\left\lfloor \frac{\ell_{ij}}{\Lambda}\right\rfloor \cdot \left(\Psi_{ij} + \delta_{ij}\cdot \sum_{sd,k:(i,j)\in\pi^{sd,k}} w^{sd,k}\right)\right)$$

$$\sum_k w^{sd,k} = t^{sd} \quad \forall s,d \in V \tag{5}$$

$$\sum_{sd,k:(i,j)\in\pi^{sd,k}} w^{sd,k} \le a_{ij} \quad \forall (i,j)\in E \tag{6}$$

$$w^{sd,k} \in \mathrm{N}, \quad \forall s,d \in V, \forall k \tag{7}$$

The objective (1) is the minimization of the network power consumption due to the network elements powered by dirty energy sources (as we want to minimize GHG emissions) and – among the solutions at minimum power consumption – the minimization of the total power consumption of the network, as reported in Eq. (2). Eq. (3) sets the overall power consumption of the network elements in G(V,E) evaluated in the energy model, whilst Eq. (4) indicates the power consumption of the network elements in G(V,E) due only to green power sources. Constraint (5) selects the routes for the lightpaths among the k pre-computed ones and assures that the whole traffic demand matrix is satisfied. Constraint (6) ensures that the maximum number of lightpaths passing on a link does not exceed the number of available wavelengths on that link. Constraint (7) imposes the integrality of the ILP problem by forcing integer values for the variables $w^{sd,k}$. Note that the fixed power consumptions terms in (3) and (4) are reported only for completeness sake but they are not involved in the optimization process (as sleep mode is not considered, fixed power consumptions are always present and the optimization is realized only on the variable energy consumptions).

B. Energy-aware RWA at minimum power consumption (MinPower-RWA)

The objective of the *MinPower-RWA* problem is to minimize the overall power consumption regardless of the energy sources types and, thus, of the GHG emissions. The set of the input parameters is the same as the *MinGas-RWA* problem except for the g_n and h_{ij} vectors which are no longer necessary; also, an additional constant ξ is considered, where $\xi:\ 0 < \xi \cdot \left(\sum_{n\in V}\Theta_n + \sum_{(i,j)\in E}\Psi_{ij}\right) < 1$. The mathematical formulation of *MinPower-RWA* is the following:

Objective function

$$\mathit{Minimize} \quad TC_{G(V,E)} + \xi \cdot \sum_{sd,k} w^{sd,k}\cdot \rho^{sd,k} \tag{8}$$

Constraints

constraints (3) (5) (6) (7).

The objective function (8) is the minimization of the overall network power consumption due to fixed and proportional power consumed by all the devices installed in the network, and – among the solutions at minimum power consumption – the minimization of the installation cost (with the assumption that the installation cost is proportional to the number of wavelengths required and to the length of the chosen lightpaths).

C. Minimum Cost RWA (MinCost-RWA)

The objective of the *MinCost-RWA* problem is the minimization of the installation cost regardless of the NEs energy consumptions and GHG emissions. It will try to aggregate as much lightpaths as possible while minimizing their physical lengths.

Objective function

$$\mathit{Minimize} \quad \sum_{sd,k} w^{sd,k}\cdot\rho^{sd,k} \tag{9}$$

Constraints

constraints (5) (6) (7).

IV. Numerical Results

A. Simulation scenario

In the following we present and analyze the results obtained through ILP optimizations exploiting minimum power consumption and minimum GHG emissions on the well known Geant2 Pan-European core optical network with 16 nodes and 23 fiber links each with 16 wavelengths [13]. Simulations were performed under different power distribution systems, with green energy sources powering 25, 50 and 75% of the NEs and randomly generated traffic matrices. Connection requests are fully satisfied, i.e., the blocking probability is kept strictly *null*. In order to evaluate the amount of emitted CO_2 of the legacy energy plants, we consider the carbon footprints illustrated in Table III. To solve the ILP problems, the CPLEX software tool was used on an Intel® Xeon® 2.5 GHz dual processor Linux server. The available memory (physical RAM + swap area) amounted to 16 GBytes. To reduce the notable requirements in terms of computational and memory resources, we first bound the problem dimension by restricting the paths' alternatives to a static set of *k* pre-computed routes, obtained by using a traditional *k*-shortest paths first (K-SPF) algorithm and hence satisfying the traditional network management objectives without considering any energy-related information. Secondly, we limited the depth of the branch-and-bound/cut algorithms after calculating a pre-definite number of integer solutions. While such simplification techniques are certainly useful to contain the computational burden, the solution they produce is only an approximation of the actual optimal (in terms of power consumption) virtual network topology built on the available physical infrastructure. However in these cases the ILP approach maintains its added value, as far as the approximated solutions can be close to the exact one. Some of the selected paths would probably not be the best ones, but the resulting power savings could be substantial without significant losses on the other optimization objectives.

B. *Results and discussion*

The power consumption resulting from the three ILP RWA strategies with 50% of the NEs powered by green energy sources is reported in Fig. 2. As expected, the *MinCost-RWA* is the most power consuming strategy, whilst the *MinPower-RWA* is the best strategy as for the power consumption, but the best one as GHG emissions is the *MinGas-RWA*. Anyway, the difference in energy consumption between the two latter strategies is lower than 15% in the worst case. This result was somehow expected, as the minimum power RWA strategy saves as much energy as possible regardless of the sources of energy, whereas the minimum GHG emissions may route the lightpaths on longer – thus, more energy consuming – paths but preferring those NEs that are powered by green energy sources. The *MinGas-RWA* energy consumption curve is further decomposed into two parts: the energy consumption resulting from dirty (*MinGas-RWA dirty*) and green (*MinGas-RWA green*) energy sources. At low loads, *MinGas-RWA* attempts to use only green-powered nodes, at the expense of possibly choosing longer paths. The effect of these suboptimal choices is visible at higher loads, when the overall power consumption rises more steeply that that of *MinPower-RWA*. This becomes relevant at network loads as high as 75%, whereas in the 30% - 70% operating range the savings achieved by *MinGas-RWA* with respect to *MinCost-RWA* remain consistently substantial. As for the energy consumption, compared with *MinCost-RWA, MinGas-RWA* saved up to 23% of energy while *MinPower-RWA* reached savings up to 32% of the overall energy consumption.

TABLE III
ENERGY PLANTS CARBON FOOTPRINTS

Type of Energy Plant	Emitted CO_2 per kWh (in grams)*
Natural Gas	880
Fuel	890
Coal	980
Nuclear	6

* Emissions during the use phase only; neither the construction costs nor other environmental effects such as fuel preparation and waste dismissal are accounted for. *Source: ACV-DRD study [2].*

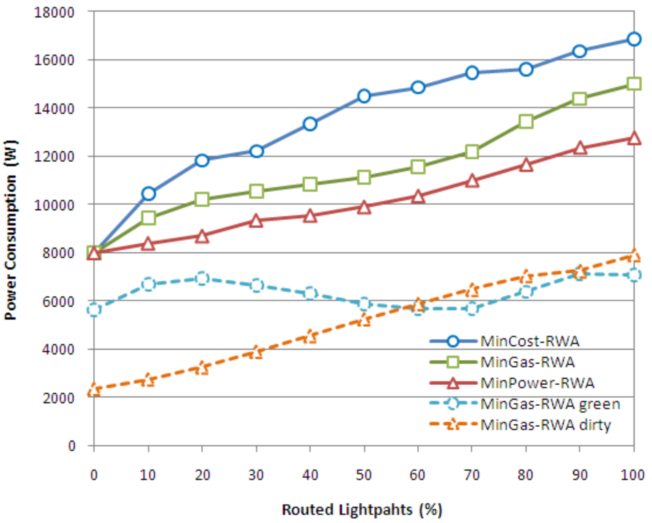


Fig. 2. Power consumption *vs* network load, 50% green power sources.

Besides the saving in power consumption, *MinGas-RWA* achieves to save also considerable quantity of CO_2. For a medium loaded network (50% of routed lightpaths), where one half of the NEs are powered by green power plants and the other half is powered by fuel-based power plants, *MinGas-RWA* strategy saves an average of 40,800 kg of CO_2 per year.

In the Fig. 3, we compared the estimated CO_2 emissions (for one year period) with the three strategies at different network loads, where one half of the NEs are powered by green energy sources and the other half by fuel-based power plants. As can be seen, at low loads the *MinGas-RWA* strategy achieves prominent CO_2 savings (only about 33% of CO_2 were emitted with respect to *MinCost-RWA* and about 50% relative to *MinPower-RWA*), whilst, as the network load increases, the difference between the *MinGas-RWA* and the *MinPower-RWA* strategies decreases, because at higher loads it becomes more and more difficult to satisfy the demand without resorting to dirty-powered nodes. In other words, at high loads, minimizing the overall power consumption implicitly leads to the minimization of the concomitant CO_2 emissions, while at midrange loads the CO_2 savings induced by *MinGas-RWA* are significant. We also explored the power consumptions and CO_2 emissions when the network is powered with different percentages of green energy sources. Results in the Fig. 4 show that, when a high percentage of the NEs (i.e. 75%) are powered by green energy sources, good results in terms of CO_2 emissions are obtained also by RWA strategies that do not take

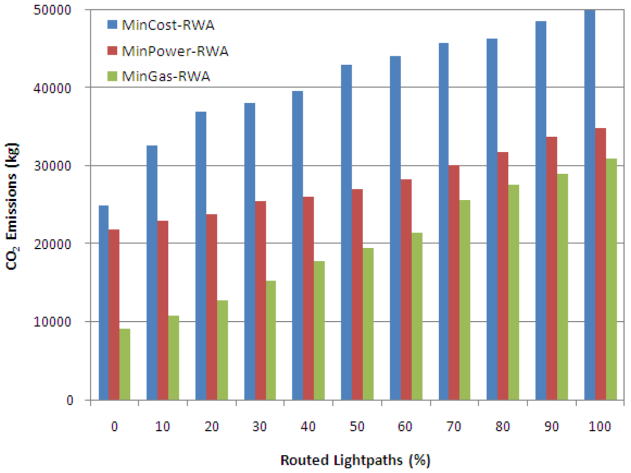


Fig. 3. Emitted CO_2 (during 1 year) *vs* network load; 50% fuel-based power sources; 50% green power sources.

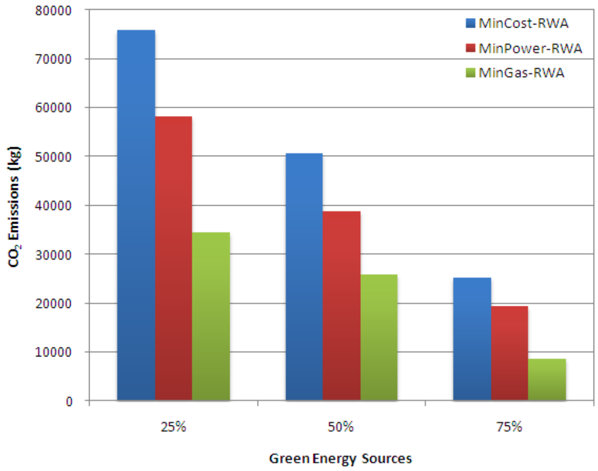


Fig. 4. Average emitted CO_2 at different green power sources percentages (remaining power sources are fuel-based).

into account the type of energy sources, whilst, when green energy sources are scarce (i.e. 25%), a RWA strategy that explicitly optimizes the CO_2 emissions leveraging the green NEs is strongly advised. In the latter case, *MinGas-RWA* would emit only half the CO_2 with respect to *MinCost-RWA*, and one third of the CO_2 with respect to *MinPower-RWA*. These results show that, using the *MinGas-RWA* strategy with as few as 25% of green energy sources, it is possible to considerably reduce the overall network CO_2 emissions. Besides, even without any change in the power sources, with the *MinPower-RWA* strategy it is possible to save up to 25% of the overall network power consumption.

V. Conclusions & Future Works

In this paper, energy-aware ILP formulations exploiting dual energy sources have been presented along with an energy model in which no sleep mode is available but the optimization relies only on the traffic-variable power consumption of the NEs. Two ILP formulations have been presented: minimum power (*MinPower-RWA*) and minimum GHG emissions (*MinGas-RWA*) strategies with the objectives to minimize respectively the absorbed energy and the emitted GHG. Results show that the *MinPower-RWA* strategy may save a considerable amount of energy by routing the lightpaths on minimum consuming NEs and that the GHG emitted may be notably reduced by the *MinGas-RWA* strategy that prefers NEs powered by green energy sources.

The effectiveness of the ILP formulations may be further evaluated according to the network topology and heterogeneity. Renewable energy sources may vary their availability with time (e.g. solar panels only generates electricity during the day). While in the current work we handled the availability of green and dirty sources in a static way, in future works statistically variable green energy sources may be considered within a totally dynamic scenario in which the availability of the different types of renewable energy sources can be associated with the variations of the day time and traffic load (e.g. night/day cycle).

Acknowledgments

This work was supported in part by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.

References

[1] Energy Star, *Small network equipment*, [Online]. Available: http://www.energystar.gov/index.cfm?c=new_specs.small_network_equip.

[2] BONE project, WP 21 Topical Project Green Optical Networks: Report on year 1 and updated plan for activities, NoE, FP7-ICT-2007- 216863, Dec. 2009.

[3] B. St Arnaud, *ICT and Global Warming: Opportunities for Innovation and Economic Growth*, [Online]. Available: http://docs.google.com/Doc?id=dgbgjrct_2767dxpbdvcf.

[4] L. Chiaraviglio, M. Mellia, F. Neri, "Reducing power consumption in backbone networks", in *Proc. IEEE ICC 2009,* Dresden, Germany, June 2009.

[5] Y. Wu, L. Chiaraviglio, F. Neri, "Power-aware routing and wavelength assignment in optical networks", in *Proc. ECOC 2009*, Vienna, Austria, Sep. 2009.

[6] L. Chiaraviglio, M. Mellia, F. Neri, "Energy-aware backbone networks: a case study", in *Proc. GreenComm'09,* Dresden, Germany, June 2009.

[7] A. Muhammad et al., "Energy-efficient WDM network planning with dedicated protection resources in sleep mode", in *Proc. IEEE Globecom* 2010, Miami, FL, USA, Nov. 2010.

[8] F. Idzikowski, S. Orlowski, C. Raack, H. Woesner, A. Wolisz, "Saving energy in IP-over-WDM networks by switching off line cards in low-demand scenarios", in *Proc. ONDM 2010*, Kyoto, Japan, Feb. 2010.

[9] I. Cerutti, A. Fumagalli, M. Tacca, A. Lardies, R. Jagannathan, "The multi-hop multi-rate wavelength division multiplexing ring", *IEEE/OSA J. Lightw. Technol.*, vol. 18, no. 12, pp. 1649-1656, Dec. 2000.

[10] X. Dong, T. El-Gorashi, J. M. H. Elmirghani, "Renewable Energy in IP Over WDM Networks", in *Proc.* ICTON, Munich, Germany, Jun. 2010.

[11] J. Chabarek et al., "Power awareness in network design and routing", in *Proc. IEEE Infocom 2008*, Phoenix, AZ, USA, Apr. 2008.

[12] S. Aleksic, "Analysis of power consumption in future high-capacity network nodes", *OSA/IEEE J. Opt. Commun. Netw.*, vol. 1, no. 3, pp. 245-258, Aug. 2009.

[13] S. De Maesschalck et al., "Pan-European Optical Transport Networks: An Availability-based Comparison", *Phot. Netw. Commun.*, vol.5, no.3, pp. 203-225, May 2003.

Telecommun Syst
DOI 10.1007/s11235-011-9645-7

Towards an energy-aware Internet: modeling a cross-layer optimization approach

Sergio Ricciardi · Davide Careglio · Germán Santos-Boada · Josep Solé-Pareta · Ugo Fiore · Francesco Palmieri

Abstract The containment of power consumption and the use of alternative *green* sources of energy are the new main goals of telecommunication operators, to cope with the rising energy costs, the increasingly rigid environmental standards, and the growing power requirements of modern high-performance networking devices. To address these challenges, we envision the necessity of introducing *energy-efficiency* and *energy-awareness* in the design, configuration and management of networks, and specifically in the design and implementation of enhanced control-plane protocols to be used in next generation networks. Accordingly, we focus on research and industrial challenges that foster new developments to decrease the carbon footprint while leveraging the capacities of highly dynamic, ultra-high-speed, networking. We critically discuss current approaches, research trends and technological innovations for the coming green era and we outline future perspectives towards new energy-oriented network planning, protocols and algorithms. We also combine all the above elements into a comprehensive energy-oriented network model within the context of a general constrained routing and wavelength assignment problem framework, and analyze and quantify through ILP formulations the savings that can be attained on the next generation networks.

Keywords Energy efficiency · Energy-awareness · Energy-oriented network models · Power consumption minimization · Carbon footprint minimization · Integer Linear Programming

S. Ricciardi · D. Careglio · G. Santos-Boada · J. Solé-Pareta
Dept. d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034, Barcelona, Spain

S. Ricciardi
e-mail: sergior@ac.upc.edu

D. Careglio
e-mail: careglio@ac.upc.edu

G. Santos-Boada
e-mail: german@ac.upc.edu

J. Solé-Pareta
e-mail: pareta@ac.upc.edu

U. Fiore
Complesso Universitario Monte S.Angelo, Università di Napoli Federico II, via Cinthia, 80126, Napoli, Italy
e-mail: ufiore@unina.it

F. Palmieri (✉)
Dipartimento di Ingegneria dell'Informazione, Seconda Università di Napoli, via Roma 29, 81031, Aversa, CE, Italy
e-mail: fpalmier@unina.it

1 Introduction

The growing energy requirements for powering and cooling the various devices enabling the up-to-date ICT infrastructures, together with the rising energy costs consequent to the exhaustion of traditional fossil sources, are drawing an increasing attention to the energetic aspects of ICT in the modern world. In addition to the economic motivation, there is also a strong environmental rationale for energetic concerns. Electricity can be obtained by "dirty" primary energy sources (e.g. burning oil, gas), releasing in the atmosphere large quantities of fine particles (aerosols) and green house gases (GHG) contributing to pollution and climate changes. Alternatively, it can be drawn from "clean" renewable sources (e.g. sun, wind, tide) that do not emit GHG at all during the use phase.[1] Both aerosols and GHG are widely recognized as the major cause for global warming.

[1] GHG may be emitted during the construction phase; anyway, renewable energy sources are beneficial over their entire Life-Cycle [34].

Published online: 17 September 2011

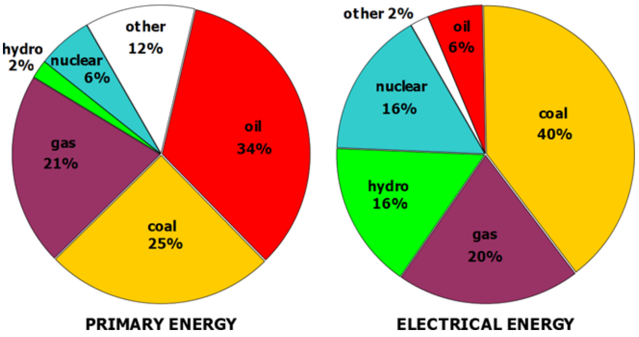


Fig. 1 Primary energy production and electrical energy generation taxonomy [1]

Currently, about 30% of worldwide primary energy is spent for producing electrical energy (with an average yield of 40%), and only a small share comes from renewable sources (Fig. 1) [1].

It has been estimated that ICT worldwide energy consumption amounts to 7% of the global electricity production [1] and the energy requirements of data centers and network equipment are foreseen to grow by 12% per year. Furthermore, with the ever increasing demand for bandwidth, connection quality and end-to-end interactivity, computer networks are requiring more and more sophisticated and power-hungry devices, such as signal regenerators, amplifiers, switches, and routers.

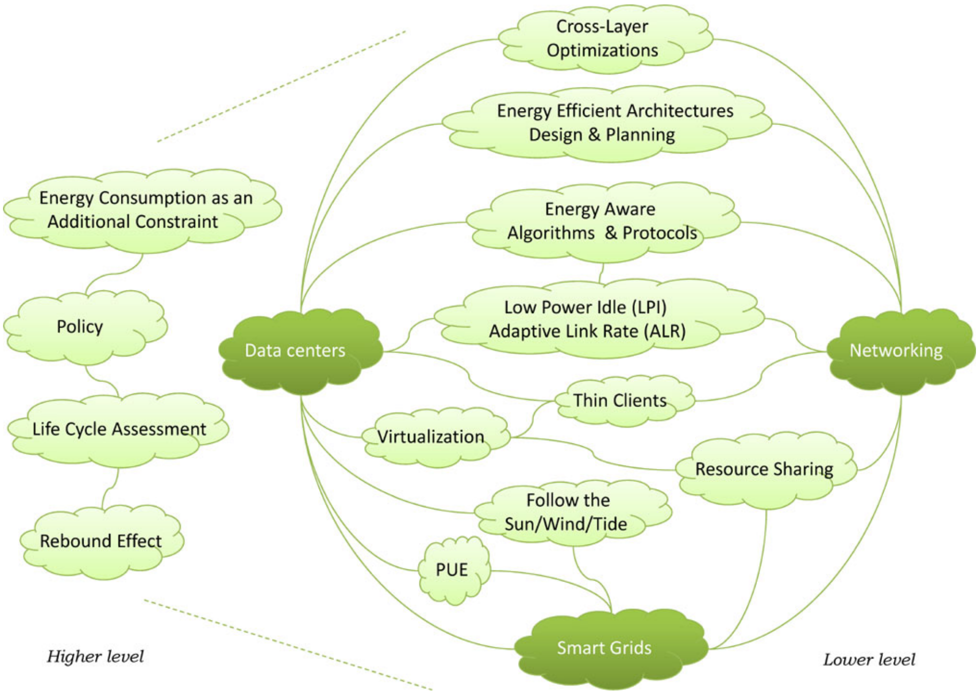
These components tend to increase the energy needs of global communication exponentially. Hence, it can be easily foreseen that in the next years the Internet will be no longer constrained by its transport capacity, but rather by its energy consumption and environmental effects [2]. In this scenario, networking equipment consumes about 1% of the total energy used for ICT, therefore it is important to keep such component into consideration when analyzing sustainable strategies for cutting the energy use. In fact, the amount of power spent worldwide for network infrastructures can be globally quantified in the order of tens of gigawatt [1], and hence limiting power consumption in networks is expected to significantly reduce the overall CO_2 and particle emission, so that the need for a greener, or—better—energy-oriented Internet is rapidly becoming a fundamental political, social and commercial issue.

At the state-of-the-art, miniaturization and ICT growing dynamics (i.e., Moore's and Gilder's laws [3, 4]) have not had the expected counterpart in power consumption reduction in the networking scenario. Miniaturization has reduced unit-power consumption but has allowed more logic ports to be put into the same space, thus increasing performances and, concomitantly, power utilization. Furthermore, the increased energy-efficiency may lead to decreased supply costs which may lead to augmented demand and consequent higher overall consumes that overtake the gained energy savings: a phenomenon called rebound effect (or Jevons paradox/Kazzoom-Brookes postulate [5]). Thus, despite architectural and semiconductor technology improvements, power consumption of network devices is still growing almost linearly with bit-rate and traffic volume [1]. As a consequence, the total power required per node is exploding. It is hence necessary to adopt a systemic approach that comprises state-of-the-art technologies (i.e., *energy-efficiency*) and new operation and management strategies (i.e., *energy-awareness*) exploiting renewable energy sources, acting in a cooperative fashion to achieve *energy-oriented* ICT. Therefore, energy has to be considered as a novel fundamental constraint for design, planning, and operations activities in the networking environment as well as in the whole ICT sector.

The envisioned future technological innovations are presented here together with a holistic view of the research challenges and opportunities that are foreseen to play an essential role in the coming green era towards sustainable (thus scalable) society growth and prosperity. Energy is considered as a novel additional constraint to design, plan and operate in the ICT systems. The semantic network in Fig. 2 illustrates the energy-oriented paradigm where both efficiency and awareness are used to achieve eco-sustainable ICT. It is an effort to visualize a framework in which decrease the energy consumptions and GHG emissions in the green Internet. The main factors that will drive this development are reported, and a visual clue on how these elements are connected with each other is given, helping to identify their relations and co-operations. The paradigm will evolve accordingly as new requirements and technological innovations come into the arena. The energy-oriented paradigm is depicted as an undirected graph, where a number of elements (nodes) and relations (edges) concur to build the complete framework. The connected elements work together and all the elements collectively contribute to achieve a holistic systemic approach. The leftmost part represents the highest-level elements, which *control* the rightmost lower level elements that are part of the global envisioned solution. At higher level, starting from the need to consider the energy consumption as a new constraint, policy should drive the changes both promoting virtuosos practices and discouraging environmental unfriendly approach: cap and trade, carbon offset, carbon taxes and incentives are all viable ways that governments are just starting to explore. In order for any solution to be successful, it is necessary to study its whole life cycle assessment, otherwise it may fall in the rebound effect and get increased energy consumption and concomitant GHG emissions. At lower levels, three main actors are highlighted and discussed: a global distributed energy system (Smart Grids), green data centers and networking.

In our envisioned framework, renewable energy (e.g. solar panels) should be available in every power-consuming site in an effort to provide each Internet service provider (ISP)/data center with its own green energy source. Indeed

Fig. 2 The semantic network for the energy-oriented paradigm in the green ICT era


Cross-Layer Optimizations
Energy Efficient Architectures Design & Planning
Energy Consumption as an Additional Constraint
Energy Aware Algorithms & Protocols
Policy
Low Power Idle (LPI) Adaptive Link Rate (ALR)
Data centers
Networking
Thin Clients
Life Cycle Assessment
Virtualization
Resource Sharing
Follow the Sun/Wind/Tide
Rebound Effect
PUE
Smart Grids
Higher level
Lower level

through an initial capital investment (CAPEX) the organizations that will employ this feature, will get reduced operational costs (OPEX), reduced GHG emissions and reduced energy dependency (which in the near future will mean financial survivability); besides, they may take advantage of the establishing green policies. Following these considerations, we envision a change in the way energy is produced and distributed. Energy production/consumption will shift from a current (insufficient and fossil-based) centralized model in which few big plants give energy to a whole geographic region to a distributed model in which a number of small renewable energy plants are spread over the territory (e.g. solar panels on the buildings roofs, mini-eolic wind mills in the streets, etc.): at every site, energy will be produced, exchanged, released when in excess and acquired when needed. This change will be encouraged by a number of factors: the current worldwide energy shortages, the rising costs of energy as fossil sources become scarcer, the need for new alternative and renewable sources of energy and the growing interests of governments and people into eco-concerns. In the same way as the cap and trade system, energy can be bartered and loaned without fees. Energy supply and demand may encounter reciprocally and exchanges may happen between neighborhoods at different hours of the day according to the different demands. The new model will not replace the existing one, but will come alongside, with the central system operating only when the renewable sources are temporarily exhausted or not available at all. Energy consuming facilities in the ICT sector will have access to several sources of energy, and will be able to switch on demand between the different energy sources dynamically acquiring or releasing excessive produced green energy as needed. In this direction, recent initiatives gathering major energy operators and providers started to explore intelligent and automated management of the future electricity distribution networks (see e.g. smart grid initiative in US [6]). To this extent, energy-awareness will become a fundamental operational feature, which can be also applied to other industrial sectors. We argue that this model can be extended to private houses, business premises, university campuses, ISPs, public buildings providing distributed green energy plants that may produce, release and acquire electrical energy (as well as cold and hot flows) from their neighborhood. Today in fact the great need is not for more energy, but for better energy utilization and wastes avoidance should be our primary objective. Data centers need to be cooled while office rooms need to be warmed (at least for several months, depending on the latitude). This supply/demand situation may be exploited by properly exchanging warm and cold flows between data centers and office rooms with both side advantages and without costs, with the enabling technology being an intelligent GMPLS-like control plane. Energy will change from a perceived cost to a revenue (re)source: saving energy reduces OPEX and produced green energy can be traded making revenues. A case in point is that in southern Germany the energy produced by photovoltaic panels exceeds the demand and a way of storing or exchanging energy is strongly advised [7]. Also, Google's 1.6 MW solar installation is the largest in corporate America at the time and recently Google has received the authorization to trade energy [8]. Another promising initiative was initiated by the administration of Iceland [9] that is promoting data centers placement near renewable energy plants. Industries and governments that will move first will obtain the greatest benefits from the new sustainable economy.

In data centers, energy-awareness can be considered from users' equipment to software and middleware level down to hardware resources. In particular, from the user perspective, the ICT trend is moving towards a network-centric paradigm, in which energy-hungry end-user equipment (e.g. PCs) is being substituted by thin clients with low power consumption and high-speed network connectivity (e.g. smart phones and netbooks), notably incrementing the use of network for connecting them to data centers and content delivery networks (CDN). If well managed, this evolution in the form users connect to the Internet can be properly exploited to significantly decrease the power consumption of both data centers and networking. Virtualization may play an important role in moving computations from the client to the server-side where data centers placed near renewable energy plants may execute the computations with lower carbon footprint with respect to traditionally power consuming (almost idle) PCs. Increasing computing density to sites where green energy is available will be the upcoming challenge for data centers. This trend will further increase the use of the network premises and consequently raise the need for energy-aware ICT network paradigms.

As a very complex combination of heterogeneous equipment, a network infrastructure has to be properly designed and managed in order to achieve advanced functionalities with a limited energy budget. Based on the state-of-the-art technological scenario, to decrease energy consumption in such networking infrastructures it is possible to operate on three different dimensions of the problem, according to a cross-layer approach:

- *Including energy-efficiency as the key requirement for the evaluation of technological advancements* is the first fundamental step towards energy-oriented networking. New devices that improve the performance of their predecessors need to be compared with competing technology also on their power consumption levels.
- Second, all the network *designing, building, and operating actions have to consider energy as an additional constraint* for the success of an energy-aware networking paradigm. Energy-awareness in network design is based on the concept of deploying algorithms and protocols that, taking into consideration the energy requirements, minimize the aggregate power demand while satisfying requirements for coverage, robustness and performance. This implies the adoption of an intelligent control plane together with the deploying of physical network topologies that aim at minimizing the number and the consumption of devices that must always be powered on. Also, dynamic power management strategies designed to decrease power consumption in the operational phase may produce positive effects for the environment and significant cost savings. Accordingly, energy-aware logical network topologies can be dynamically built by making maximum usage of powered-on and highly connected devices exploiting as much as possible resource sharing by cleverly reusing switching nodes and fiber strands.
- The final dimension is the introduction of *energy-oriented control plane protocols* whose goal is to properly accommodate network traffic considering, energy-efficiency, energy-awareness and renewable energy sources. Daily and hourly fluctuations in user demands and green electricity availability may be considered across the involved network infrastructures/areas, in such a way that the overall power consumption and GHG emissions can be optimized. Accordingly, ILP formulations and heuristic methods can be introduced for calculating the routing information subject to power consumption constraints, also by taking into account the specific kind of energy source (dirty or renewable) used for powering the traversed network elements (NE).

Starting from the above operating dimensions, we present a model for the energy consumption aimed at describing and quantifying the savings that can be attained through control plane protocols that affect the route/lightpath choice privileging renewable energy sources and energy efficient links/switching devices. In this endeavor, we tied together into a general constrained Routing and Wavelength Assignment problem framework all the features needed by a comprehensive energy-aware network model. For both contexts where information about the GHG impact of the energy source is available and where it is not, we present Integer Linear Programming formulations to characterize optimization objectives and constraints in a formal and expressive way, and finally discuss some simulation results, analyzing the saving attainable.

2 Current approaches and research trends for energy-oriented networks

In this section we describe the three aforementioned dimensions of the cross-layer approach in network infrastructures, critically discuss the major issues and provide the basic grounds to our approach.

2.1 Evaluating technological advances for energy-efficiency

Technological advances allow having more efficient network devices that consume less and process more bits per second, according to the "do more for less" paradigm. Such solutions are usually referred to as eco-friendly. Currently, the power requirement for electronic network devices is scaling almost linearly with their total aggregate bandwidth.

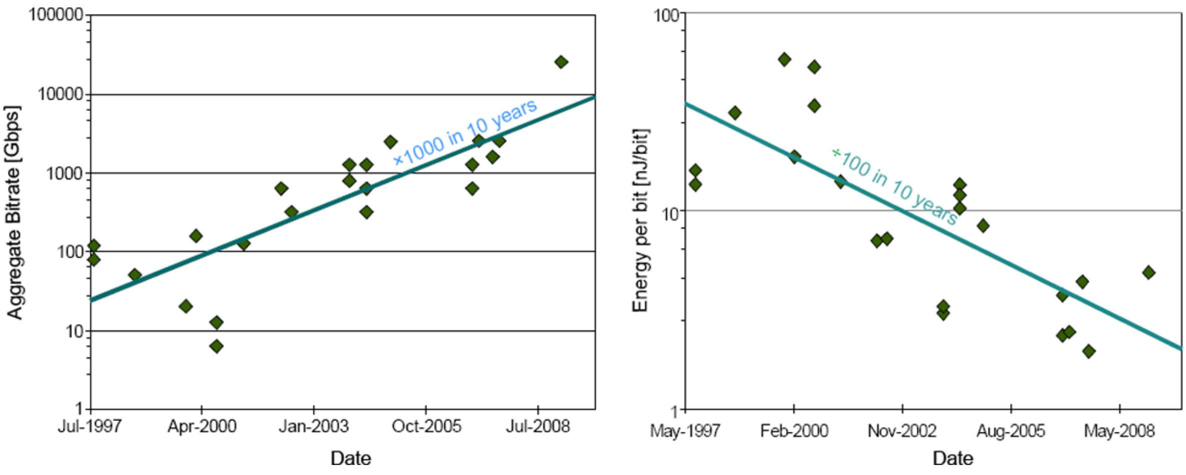


Fig. 3 Evolution of the bandwidth capacity and energy-per-bit consumption [1]

In particular, the energy-per-bit (measured in nJ/bit) decreased approximately by a factor of 100 in the last 10 years while the bandwidth capacity has increased by a factor of 1000 in the same time frame, thus—assuming that the demand almost equals the supply—the energy consumption increment is in the order of 10 times more than 10 years ago (Fig. 3). In other words, technological advancements foreseen by Moore's law have not been fully compensated by the same decrease in the energy-consumption.

The essential reason is that, despite the astonishing performance improvements in terms of transmission and forwarding capabilities observable today in networking devices, the energy dedicated to the primary functions of routing and switching is not exploited in the best possible way. The fundamental cause of energy consumption in networking equipment is the loss effect experienced during the transfer of electric charges, due to imperfect conductors and electrical isolators. Here, the exact consumption rate depends on the transition frequency and the number of gates involved, together with fabrication features (such as architecture, degree of parallelism, operating voltage, etc.). These values can be improved by industrial advancements but only to a certain extent, since there are physical bottlenecks inherent in the electrical switching technology involved. Therefore, traditional electrical networking can be considered as inadequate to support the emerging carbon footprint reduction requirements. Indeed, the power consumption of the actual electronic routing/switching matrix and line interface cards is, quite surprisingly, almost independent from the network load and can reach hundreds of kW for large multi-shelf configurations [10, 11]. Experimental energy consumption measurements [10] on several electronic routers show that in current architectures almost one half of the energy consumption is associated to the base system and up to another half to the active line cards. The traffic load only affects power consumption by a 3%; in other words, the energy demand of heavily loaded devices is only about 3% greater than that of idle ones. These results suggest that it is necessary to develop energy-efficient architectures exploiting the ability of putting into energy saving mode some subsystems (e.g. line cards, input/output ports, switching fabrics, etc.) in order to minimize energy consumption whenever possible. It has been also demonstrated [12] how consumption depends on the packets size and on the bitrates of the links. Traffic flows characterized by bigger packets need less energy than those made of smaller ones, due to the lower number of headers that have to be processed. Analogously, a circuit-based transport layer may reduce energy consumption with respect to a packet-switched one (Fig. 4). Although routers require more power when working at higher throughputs, if the per-bit power consumption is considered, larger routers operating within the core consume less energy per-bit than smaller ones located on the edge [13, 14], thus the power consumption will be greater on the network edge and smaller within the core, due to the higher traffic aggregation in the network core with respect to the edge.

Besides offering huge data rates (theoretically up to 50 terabits per second in a single fiber [15]), limited disturbance, and low cost, optical communication technology requires very low energy for the transmission of signal—light pulses—over the optical fibers. Furthermore, wavelength division multiplexing (WDM) technology (sending several independent optical signals in the same fiber cable using different wavelengths—80 wavelengths devices are commercially available) has dramatically increased the available bandwidth and greatly reduced energy consumption (Fig. 4). For comparison, an Optical Cross-Connect node (OXC) with micro-electro-mechanical system (MEMS) switching logic consumes about 1.2 W per single 10 Gb/s capable interface, whereas a traditional IP router requires about 237 W per port [1].

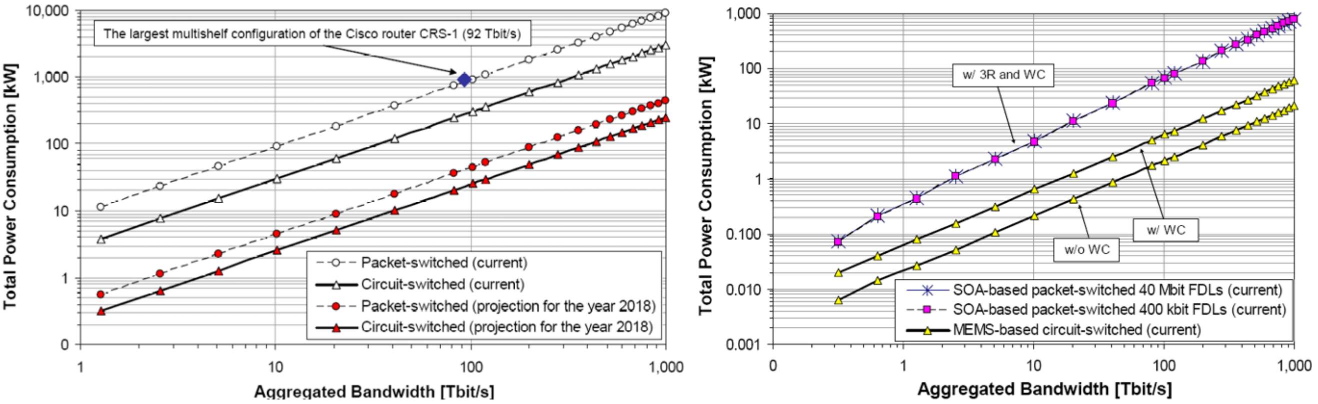


Fig. 4 Comparison of the total power consumption *vs* the aggregated bandwidth for electronic (*left*) and optical (*right*) router/switch technologies [1]

Table 1 Energy/power consumption as function of the aggregated bandwidth

Router technology	Energy Consumption Rate (ECR)	Energy Scaling Index (ESI)	Power consumption (P) as function of the aggregated bandwidth (B)
Electronic DXC	3 W/Gbps	3 nJ/bit	$P = 3 \cdot B$
Optical OXC with conversion	0.062 W/Gbps	0.062 nJ/bit	$P = 0.062 \cdot B$
Optical OXC w/o conversion	0.02 W/Gbps	0.02 nJ/bit	$P = 0.02 \cdot B$

ECR and ESI are different power consumption metrics that may be reduced to equivalent values, in fact it holds that: W/Gbps = (J/s)/(Gbit/s) = J/Gbit = nJ/bit

Many reference metrics can be used when comparing the energy consumption of networking equipment. The Energy Consumption Rate (ECR) [16], targeted towards high-end packet-based network and telecom equipment, defines a testing methodology and expresses the energy consumption per maximum throughput, typically *Watt/Gbps*. ECRW is a similar weighted metric that also takes into account off-peak and idle conditions. The Energy Scaling Index (ESI) is a metric to compare the efficiency of switching devices; the ESI corresponds to the switched aggregate bitrate offered for each *Watt* of energy budget. In Table 1 we report the mean energy/power consumptions for different router technologies under several energy metrics derived from real devices energy consumption values [1]. Electronic routing devices need 150 times more power than optical ones to route the same amount of traffic without wavelength conversion (WC), and nearly 50 times more than optical devices with WC. In electronic routers, an increase of 1 throughput unit corresponds to an increase of 3 units of power consumption; the same increase in an optical node performing WC corresponds only to an increase of 0.062 units of power consumption and to 0.02 units not supporting it.

However, as far as the optical signal still needs to be converted into the electronic domain (such as in current opaque optical network equipment) the power requirement will remain remarkably significant. Therefore, there is much room for improvement towards entirely optical networks, where most of the processing—ideally, all of it—is done at the optical layer, with the associated energy savings. Nevertheless, points of electronic processing are still necessary. At different network levels (access, metro, and core) electronic processing is required in order to aggregate low/medium bandwidth client signals into higher capacity flows—a process known as *traffic grooming*—and achieve high usage percentage of transmission links, reducing the required number of active node ports. At network interface points, the electronic level still seems to be desirable to maintain well distinct and separated Service Level Agreements (SLA) responsibilities between clients and operators. Furthermore, exhaustive network monitoring is currently possible only by converting and analyzing the signal into the electronic domain. Complete optical signal regeneration (“3R” regeneration: re-amplification, re-shaping, re-timing) is commercially available only by means of electronic energy-expensive processing devices (typically 60 W per wavelength/channel). Therefore, 3R regeneration should be avoided as much as possible in planning, designing and managing new paths throughout an optical infrastructure. On the contrary of 3R signal regeneration, optical signal amplification (1R regeneration) may be done entirely in the optical domain and should be used to extend the reach of optical fibers without any electronic conversion. Thus, instead of using energy-hungry 3R regeneration, it is preferable to use optical amplifiers (every 80–100 km) so that the other end of the fiber can be reached without 3R regeneration, with optical add and drop multiplexers (OADM) inserting and extracting client signals when needed. Commercially available optical amplifiers are mainly based on the erbium doped fiber amplifiers (EDFA) technology, while semiconductor optical amplifiers (SOA) are emerging as possible candidates to replace EDFA. Currently, EDFA are more performing (higher gain, lower insertion loss, noise and cross-talk effects) than SOA but have also higher energy consumption (respectively 25 W and 3 W). As SOA technology will evolve and reach EDFA performance, SOA will become the default choice for achieving low-energy optical signal amplification into future long haul optical fibers. In addition to the use of low-power SOA, the use of dispersion compensation fibers (DSF: ITU-T G.653, NZ-DSF ITU-T G.655/656) instead of "simple" single mode fiber (SMF: ITU-T G.652) will reduce the dispersion of the optical signal traversing the fiber and reduce the number of required optical amplifiers.

A research field related to optical networking promising further power savings is optical logic. It consists in incorporating photonic functionality in silicon very-large-scale integrated (VLSI) circuits and it is considered a natural choice for optical networks because it could provide the ability to build optoelectronic systems with integrated control electronics. It is also argued that the energy cost of converting data from the optical to the electronic domain and back is not inherent to the fundamental physics of such conversions, so that a properly designed integrated approach may help reduce this cost. However, optical logic remains challenging and one of the toughest issues is power absorption. Optical logic would indeed lose much of its attractiveness if it would consume more energy than regular silicon. Current silicon CMOS devices operate with energies in the range of femtojoules per operation and future transistors are expected to evolve towards capacitances of tens of attofarads (and therefore energies of tens of attojoules for operation at the expected voltage levels), and matching these values is a demanding target for natively optical devices [17].

New ideas are also emerging in the evolution of core networking and the converged transport and Ethernet for carrier networks. Driven by high-definition video and network computing, the bandwidth requirements are doubling every 18 months and Terabit Ethernet is forecasted to be needed as early as in 2015–2017 [18]. The IEEE 802.3 Working Group and ITU-T Study Group 15 have recently established draft standards for 100 Gigabit Ethernet [19]. These will call for the development of new optical transceivers, whose power requirements have been constrained at an 80 W maximum power consumption and maximum temperature of 70°C. While technology and component reuse is already established as a driver orienting decision about the next leap in speed of Ethernet, power consumption issues does not seem to have yet reached the same importance. The advent of 100 Gb Ethernet will bring many advantages related to the reduction of the required energy. In particular, the price and power consumption of one 100 Gb interface will be significantly lower than those of ten 10 Gb interfaces, and the efficient use of the DWDM links will limit the recourse to parallel links. However, the transition phase must be properly planned and managed. Not every operator will invest into 100 GbE at the same time. Thus, it is important that power awareness will be considered at all levels, according to cross-layer optimization principia, to obtain immediate benefits.

The choice of transport technology in access networks may also be a strong enabler for energy-efficiency. At the state-of-the-art, the vast majority of the energy consumption can be attributed to fixed line access connections. Today, access networks ("the last mile") are mainly implemented with copper based links and technologies such as ADSL and VDSL, whose energy consumption is very sensible to increased bitrates. The trend is to replace such technologies with fiber infrastructure, especially in the emerging countries in which new installations are being deployed from scratch. In access networks, the energy consumption scales basically with the number of subscribers, so that the massive diffusion of fiber to the home (FTTH) in place of old copper xDSL access links would have the dual benefit of dramatically increasing the access bandwidth and decreasing energy consumption. For comparison, a single ADSL link consumes about 2.8 W, while using a gigabit-capable passive optical network (GPON) as the access infrastructure will reduce the consumption to only 0.5 W, an improvement of about 80% for a potentially very high number of users. Such ongoing replacement is moving the problem to the backbone networks, where the energy consumption for IP routers, driven by the ever increasing bandwidth requirement, is becoming a bottleneck [20, 21].

Estimating the global footprint accurately is in many cases highly complex. The specific equipment density and hardware integration, heat dissipation and power supply specifications must also be kept in mind as fundamental parameters for energy efficiency, when considering collateral energy charges such as cooling and power conversion. The Power Usage Effectiveness (PUE), defined by the Green Grid [22], measures the efficiency of an ICT facility as the ratio of total amount of power used by the facility to the power delivered to the equipment, thus assessing the fraction of energy consumption due to, e.g., the HVAC (Heating Ventilation and Air Conditioning), the UPS (Uninterruptible Power Supply) subsystems and the lighting facilities. A PUE value of 2 is the current average, meaning that HVAC and UPS double the energy requirements [13]. In this scenario, overcooling can be considered as the main collateral energy drain and further gains can be obtained by using computational fluid dynamics and introducing contained cooling strategies. The use of cold aisles ducted cooling or in-rack cooling systems help to keep the volume of fluid being cooled at a reasonable minimum. Outlet air can be vented directly to the outside, or preferably reused for space or water heating elsewhere in the building as required with the consequent improvements in energy and carbon footprint.

The above issues become more and more significant when a network is to be built from scratch or network upgrade decisions need to be taken. In these cases, it is necessary to choose equipment and network topology considering not only performance and cost but also the energetic budget: the usual trade-offs in capital (CAPEX) and operational (OPEX) expenditures between design decisions will have to be evaluated under an energy-efficiency perspective. The effort should be twofold: on one side, developing commercially available all-optical devices that perform wavelength conversion and 3R regeneration without the need of energy expensive electronic devices; on the other side, planning the network in such a way that 3R regeneration is not needed at all.

2.2 Designing, building, and operating energy-aware networks

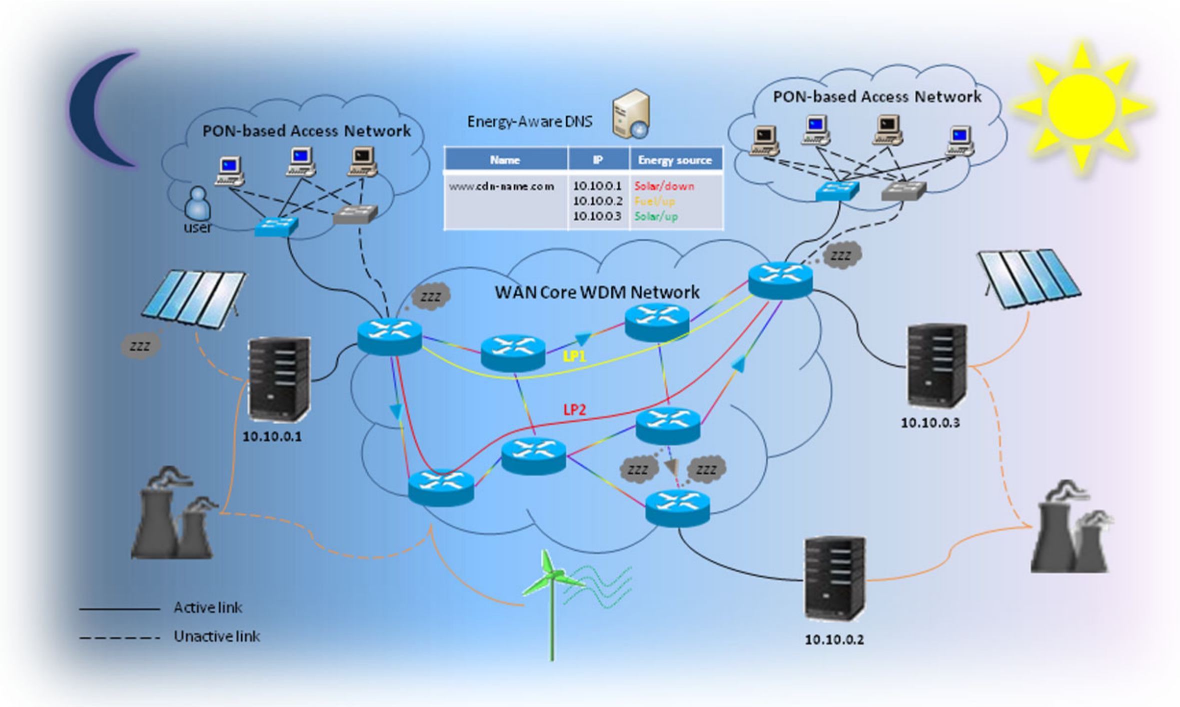
Current network design, configuration and management practices are based on deploying and maintaining infrastructures that are extremely reliable, provide performance that enables competitive SLAs, and offer a set of features and services that are attractive to a broad range of customers. To accomplish these goals, network architects typically conceive network infrastructures that are densely meshed, with many redundant interconnections between nodes, so that many alternative paths can provide multiple reachability options between geographically distant sites. Also, fair load balancing has always been a predilection of network designers and maintainers, because it increases the possibility of putting new traffic into the network. Since the traffic demand may not be known in advance, network operators need to ensure that they have sufficient free capacity for any demand that may reasonably emerge in the operating lifetime of their infrastructure.

When designing the layout of large scale infrastructures, it is desirable to find a good balance between the competing needs to avoid as many electrically-powered hops as possible (to reduce the power consumption at intermediate switching nodes or regenerators) and to not transmit data over excessively long stretches, because it's more energy-expensive to move data farther. Furthermore, traffic dynamics often present notably changes over time, resulting in different network usage between peak hours and the rest of the day. In these cases, the network has to be dimensioned to handle the maximum load, to satisfy the users' demand in peak hours, but the deployed connectivity resources risk to remain under-utilized by a wide margin for most of the time, giving rise to significant energy waste and unnecessary operating costs. It should be considered that it is possible to run only the part of the infrastructure that is really required at any time. This is an opportunity that cannot be missed and hence it is necessary to develop energy-efficient architectures exploiting the ability of selectively shutting down some links or putting into energy saving mode some devices or subsystems (e.g. switching fabrics, line cards, input/output ports, etc.) in order to minimize energy consumption whenever and wherever possible. Accordingly, adaptive power management strategies designed to decrease power consumption in the operational phase may introduce positive effects for the environment and significant cost savings, as a consequence of the reduced energy usage. Maximizing the reuse of energy-conservative transmission links and powered-on, highly connected devices—in contrast to spreading traffic on all the available switching nodes, fibers, and paths—power consumption can be drastically reduced by temporarily switching off unused devices and line cards. Because such "sleep mode" strategy can be implemented at different levels of granularity, the chosen scheme needs to be very flexible and ensure the potential to save energy as soon as few end-customers are disconnected. Nodes may be put into sleep completely (per-node sleep mode) or partially (per-interface sleep mode). However, we deem that a drastic energy containment strategy such as per-node sleep mode is too simplistic and its effectiveness on real world network environments is questionable due to the undeniable impacts on the carrier-level network economy both in terms of capital (making connectivity investments partially useless) and operational expenditures (reducing the meshing degree and hence resiliency and traffic engineering capabilities). Furthermore, state-of-the-art consumer electronics used in broadband infrastructures are typically designed to enable maximum performance in an "always on" mode of operation. By leveraging on hardware equipment similar to those used in laptops, supporting fast "sleep" or "low-power" modes, next generation networking devices will have an outstanding opportunity to efficiently reduce their power consumption when not in use. These may comprise energy proportional computing techniques, meaning slowing down CPU (Central Processing Unit) clock for inactivity periods or "simply" reducing execution speed for energy saving purposes. Fast full-clock return procedures will be needed in order to achieve the desired level of system responsiveness. Introducing these technologies in networking hardware architectures will imply a shift from the "always on" to the "always available" paradigm, where each device can spontaneously enter a sleep or energy saving mode when it is not used for a certain time and quickly wakes up or restores its maximum performance on sensing incoming traffic on its ports. On the other hand, putting into sleep mode at interface level may have some sense, in particular high speed ones, since typical commercial off-the-shelf (COTS) devices drastically improve their consumption when transmitting at their maximum rates. However, the support of sleep mode at the single interface or linecard level also requires modifications to current router architecture and routing protocols. In fact, an interface put into sleep mode may not respond to periodic *hello* messages of the routing protocol and be classified as "down"; consequent link state advertisement messages will spread along the network informing that the interface is down, causing stability problem to the convergence of the routing (or spanning tree) algorithm. For these problems, more than a static sleep mode, a per-interface "wakeup on activity" or "downclocking" mechanism [23] may be more viable and effective solutions. In the first case, the transmission on single interfaces is stopped when there is no data to send and quickly resumed when new packets arrive. To do this, the circuitry that senses packets on a line is left powered on, even in sleep mode. This mechanism can be implemented by introducing the concept of Low Power Idle (LPI) [24], which is used instead of the continuous IDLE signal when there is no data to transmit. LPI defines large periods over which no signal is transmitted and small periods during which a signal is transmitted to refresh the receiver state to align it with current conditions. Alternatively, the ability to dynamically adapt the link rate according to the real traffic needs can be another effective technique to reduce power consumption (Adaptive Link Rate, ALR). Operating a device at a lower frequency can enable reductions in energy consumption for two reasons. First, simply operating more slowly offers some fairly substantial savings. Second, operating at a lower frequency also allows the use of dynamic voltage scaling (DVS) that reduces the operating voltage. This allows power to scale cubically, and hence energy consumption quadratically, with operating frequency [25]. The adaptive link rate speed control mechanisms [26] aims at dynamically adapting the link speeds and interface behavior to the current network load by using some specific thresholds. In [14] it is shown that the energy consumption does not depend on the data being transmitted but only depends on the interface link rate, and hence is *throughput-independent*. In particular, faster interfaces require lower *energy per bit* than slower interfaces, although, with ALR, slower interfaces require less *energy per throughput* than faster interfaces, due to the higher fixed power consumption of faster interfaces circuitry. In such a context, circuit over-provisioning may lead to decreased operational costs (OPEX) at the expense of increased capital expenditures (CAPEX), i.e. a network interface may be provisioned with different circuits, say a low and a high speed one, and may switch between one or the other according to the required data throughput. It is also observed that for current technologies the energy/bit is the same both at 1 Gbps and 10 Gbps, meaning that the increase in the link rate has not been compensated at the same pace by a decrease in the energy consumption. After long discussion about which technology between ALR or LPI should be introduced into the Energy Efficient Ethernet (EEE), the IEEE 802.3 EEE Study Group chose in favor of LPI to reduce the energy consumption of a link when no packets are being sent. In [23] the LPI with packet coaliscing was used to improve the efficiency of EEE while keeping the introduced packet delays under tolerable bounds.

In addition, resiliency to failures can be very energetically inefficient when implemented through the provision of $1+1$ protection, since usually equipment stays always turned on for fast failure recovery, but is used very rarely—only when a failure occurs. Several alternatives to this standard scheme may reduce the energy consumption induced by protection. For example, resiliency can be provided without having redundant equipment stay in the fully operating state all the time, but rather keeping it down-clocked with fast wake-up capability, taking care that it is compatible with industry-standard path/span protection switch times (50 ms).

2.3 Energy-oriented control plane protocols

By considering the problem from the perspective of the topmost layers, it can be envisioned that the future green network will be based on a highly adaptive and reconfigurable transparent optical core. Several optimizations can be performed according to a "cross-layer" approach, whereby issues arising at the physical layer (e.g., energy consumption) can be handled at higher layers (typically within the control plane), through appropriate routing and signaling practices. Introducing energy-awareness in the network control plane is based on the concept of placing network traffic over a specific set of paths (and hence sequences of nodes and communication links) so that the aggregate power demand is minimized while end-to-end connection requirements are still satisfied. Every time a new path is established between any pair of nodes, traffic between these nodes will be routed on it as if in presence of a direct "virtual" connection between them, by creating the abstraction of a logical network topology on top of the physical one. Energy-oriented logical network topologies can be dynamically built by optimizing the choice of energy sources in such a way that renewable sources are preferred when available, possibly with a trade-off between path length and carbon footprint. In fact, network elements may have dual power supply: the always available power coming from dirty energy sources and the not always available power coming from green energy sources. Consider, for instance, the availability of energy produced by solar panels; it is strongly correlated with



Name	IP	Energy source
www.cdn-name.com	10.10.0.1 10.10.0.2 10.10.0.3	Solar/down Fuel/up Solar/up

Fig. 5 The energy-oriented network infrastructure

the time of the day, since it is known that no energy will be produced during the night and that some energy is expected to be produced during the day. Such knowledge should be included in the control plane algorithms for energy-aware routing, signaling and resource allocation, implementing an automatic *follow the sun* paradigm. As another example, we can imagine some equipment powered by wind energy where power supply is a pseudo-random process depending on the availability of wind. Due to the inertia of the power generating mechanisms and batteries, a drop in the wind power does not result immediately in a power generation drop. Hence, if wind stops, it is possible to reconfigure the network dynamically, to consider the new distribution of available clean energy and re-optimize its carbon footprint. Differently from the case of the daylight, whose duration is known in advance, a decrease in wind strength is much more unpredictable and the warning time is shorter. This should be only handled with adaptive and efficient rerouting mechanisms implemented within the network core. For this reason, it is necessary to develop novel routing schemes and resource allocation mechanisms that take advantage of the early notification of the forecast power variation of clean sources with time-varying power output [10]. Furthermore, another interesting perspective in energy-aware networking comes from linking traffic routing to the different available electricity prices, dynamically and continuously moving data to areas/devices when and where electricity costs are lower.

Energy-awareness may be implemented at the application layer, e.g. in an energy-aware DNS (Domain Name System), bringing advantages both to data centers and networking. As an example, a content distribution network (CDN) is made up of several data centers located in several distant sites (e.g. Europe and US). Usually, each data center contains replicated data for security and load balancing purposes. So, large bursts of data have to be transferred between the sites. Among the different possible paths, the most energy-efficient ones may be chosen for transferring the data. Similarly, users' requests to access the CDN contents may be redirected to the current lowest carbon footprint data center by the energy-aware DNS constrained on the current energy supplies. These concepts are illustrated in Fig. 5, in which the following scenario is depicted. The WAN core network is a dynamically reconfigurable transparent WDM network and the access network is based on an energy-efficient passive optical network. The data centers sites 10.10.0.1, 10.10.0.2, 10.10.0.3 are part of the same CDN and data is mirrored among them with high-speed data transfers through the optical core using lightpaths (not represented here) chosen by an energy-aware routing protocol. Some data centers and routers are equipped with dual power supply (in sites where renewable energy source are available): the green energy source and the legacy always-available fossil-based energy source. The control plane is aware of the type of energy source that is *currently* powering routers and servers. When the green energy source is temporarily not available or the accumulated energy in the batteries is exhausted (for example because night has fallen or the wind has stopped), the UPS at the site switches to the fossil-based power supply without any energy interruption. Within data centers, a subset of servers is automatically put into sleep mode when the current load allows it. In the core and access networks, the unused network interfaces and the corresponding links and amplifiers/regenerators are dynamically put into sleep mode using an energy-aware control plane. Following a planning stage, end users premises are connected to two access points (or two line cards of a single access point), such that one access point (or one line card) can go to sleep mode as soon as a suitable number of clients are turned off and require no network activity. When a user (top-left corner) needs to download a file from the CDN, a query is made to a green DNS server that knows how the CDN servers that hold the desired file are currently powered up: server 10.10.0.1, although provided with the dual power supply, is currently powered up by the fossil-based energy plant because it is night; server 10.10.0.2 is using exclusively electricity generated by a coal power plant, while server 10.10.0.3 is currently powered up by clean energy and hence its IP address is returned by the DNS server. In this way, a paradigm shift towards energy-oriented networks and data centers is capable of sustaining the growing traffic rates while limiting and even decreasing the power consumption.

In order to support all the above adaptive behaviors, energy-related information associated to devices, interfaces and links need to be introduced as new constraints (in addition, for instance, to delay, bandwidth, physical impairments, etc.) in the formulations of dynamic routing algorithms and heuristics. Down-clocking or enhanced sleep mode features should be handled as new features in the network element status that need to be accounted for at both the routing and traffic engineering layer, and such information must be conveyed to the various network devices within the same energy-management domain. This clearly requires modifications to the current routing protocols and control plane architecture. For example, the existing routing (OSPF-TE, IS-IS) and signaling (RSVP-TE, CR-LDP) protocols within the GMPLS traffic-engineering framework may be extended to include energy-related information such as the power consumption associated to a specific link and the type of energy source currently used by the entire device. This can be easily accomplished by introducing new specific type-length-value (TLV) fields in IS-IS or opaque Link State Advertisements (LSA) in OSPF. Analogously, the same information has to be handled by signaling protocols such as RSVP-TE and CR-LDP to allow the request and the establishment of power-constrained paths across the network (i.e., path traversing only nodes powered by renewable energy sources or crossing only low-power transmission links).

However, in many cases, the carbon footprint improvements may be achieved at the expense of the overall network performance (e.g. survivability, level of service, stability, etc.), which can in turn be compensated through over-designing (increase of CAPEX) or over-provisioning (increase of OPEX). This implies that the new routing algorithm empowering the energy-aware control plane should be driven by smart heuristics that always take into account the trade-off between network performance and energy savings. In fact, by putting equipment or components that consume energy into a low energy consumption mode (e.g., nodes, line cards, links), or creating traffic diversions driven by reasons different from the network load, we implicitly reduce the network available capacity and hence paths tend to be longer and/or more congested, decreasing the overall transmission quality.

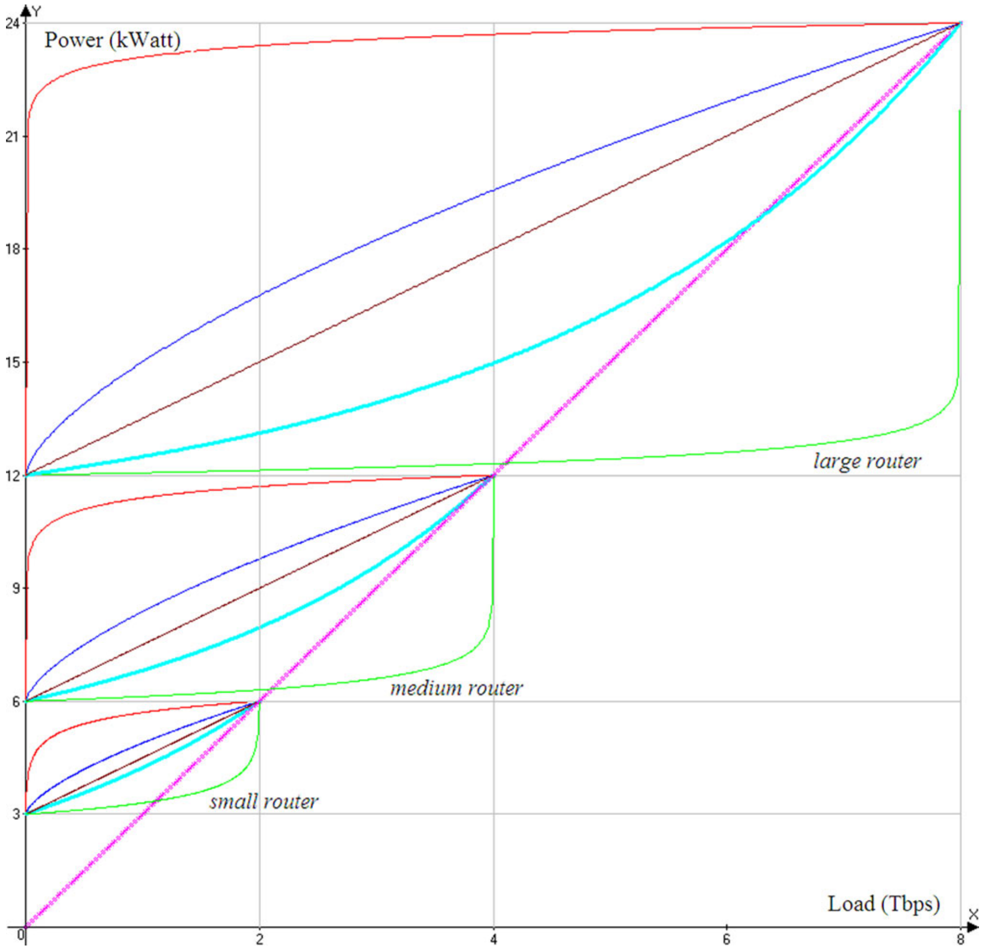
3 Modeling a cross-layer energy optimization framework for wavelength routed optical networks

Starting from the above considerations, we propose the introduction of energy-awareness into control plane protocols whose goal is to properly condition all the route/lightpath selection mechanisms on relatively coarse time scales by privileging the use of renewable energy sources and energy efficient links/switching devices, simultaneously taking advantage from the different users demands across modern wavelength routed network infrastructures, in such a way that the overall power consumption can be optimized. In doing this, we tried to combine, on each involved layer, all the notable features that a comprehensive energy-aware network model should have and put them together into a general constrained routing and wavelength assignment problem framework. Such problem has been modeled through integer linear programming to better characterize its formulation in terms of optimization objectives and constraints. Clearly, the ILP formulation, while effective in its formalism and descriptive power, is inherently static and hence implies a-priori knowledge of the entire traffic matrix to be solved at optimum; furthermore, the RWA ILP can be reduced to a single-commodity NP-complete problem. Anyway, for each real implementation perspective we need to assume that each node is capable to interact and cooperate with its neighbors by using a GMPLS or ASON-like control plane intelligence, enabling the exchange of the aforementioned energy-related information.

3.1 Basic modeling choices and assumptions

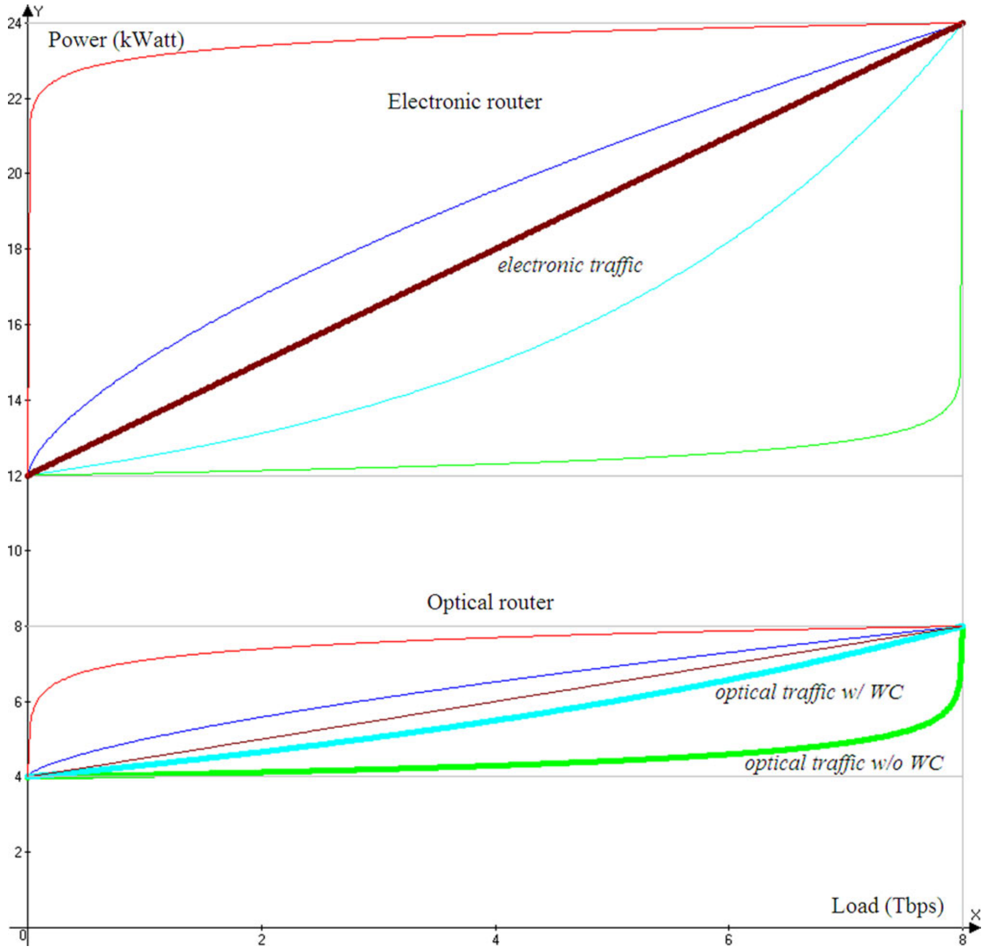
Defining a sustainable and effective model for energy consumption is the essential prerequisite for introducing power awareness within the wavelength routing context. A broad

Fig. 6 Power consumption functions for various size electronic routers



variety of devices contribute to power adsorption in a WDM network: OADM, regenerators, amplifiers, "opto-electronic" and totally optical routers and switches. Each of these devices draws power in a specific way, which may also depend on the relationship between different devices or the components of more complex structures such as switches. In addition, NEs may be powered either by green or dirty energy sources statically assigned to each at the network topology definition time, therefore a differentiation between energy sources is required at the control plane level. In order to formally characterize the energy consumption of network elements we propose a comprehensive analytic model based on real energy consumption values and in line with the theoretical grow rate predictions encompassing new energy-aware architectures that adapt their behavior with the traffic load in order to minimize the energy consumption. Such an energy model characterizes the different components and sub-systems of the network elements involved. It provides the energy consumptions of network nodes and links of whatever typology and size and under any traffic load. The efforts in the developing of such an energy model have been focused on realistic energy consumption values. For this scope, the energy model has been fed with real values and the energy consumption behavior of NEs has been crafted in order to match with the state-of-the-art architectures and technologies. At this extent, future energy-efficient architectures with enhanced sleep mode features have been considered and implemented in the energy model. The energy model is based on a linear combinations of energy consumption functions derived from both experimental results [1, 10, 13, 26, 27] and theoretical models [6, 28]. Besides, following the results reported in [13, 29, 30], the power consumption has been divided into a fixed and a variable part; fixed part is always present and is required just for the device to be on; variable part depends on the current traffic load on the device and may vary according to different energy consumption functions. We chose [31] a linear combination of two different functions (logarithmic and line functions) and weighted them depending on both the type of traffic and the size of the NE, in order to obtain a complete gamma of values and thus adapting its behavior to the most different scenarios. In particular, in our energy model we managed to obtain that larger routers consume less energy per bit than the smaller routers (Fig. 6), as reported in [13, 14] and that electronic traffic consumes more energy per bit that optical traffic (Fig. 7), as reported in [1, 29]. Wavelength conversion

Fig. 7 Power consumption functions for electronic and optical routers



and 3R regenerations have a not negligible power consumption, which is accounted for in the model. Finally, links have an energy consumption that depends on the length of the fiber strands and thus on the number of optical amplification and regeneration needed by the signal to reach the endpoint with an acceptable optical signal-to-noise ratio (OSNR).

The power consumption functions of three routers of different sizes are reported in Fig. 6. Each router may support different types of traffic, each defined by a different curve (Fig. 7). In the example, the thicker lines represent the power required by a given type of traffic (e.g. electronic traffic).

We can observe that, according to our model, the larger the router, the larger the total energy consumption, as the fixed part notably contributes to (half of) the energy consumption. But if we focus only on the variable power consumptions, we observe that, for example, a traffic load of 2 Tbps, requires as much as 3 kW in the smaller router, about 1.5 kW in the medium one and just 1 kW in the larger router. In this way, we managed to obtain that greater routers consume less energy per bit than smaller ones, as reported in [13, 14]. Note also that the overall energy consumption scales linearly with the size of the router and that half of the energy consumption is due to the fixed part and the other half to the variable part, according to literature source [10].

In detail, at the basis of our model we consider wavelength-routed networks and, for the sake of generality, lightpaths that may have different bitrates (i.e., different bandwidth requirements, according to the particular SLA on the QoS of each client). The power required for transporting one lightpath will vary with its bitrate, so we consider as traffic unit the *bps* (bits per second). Network nodes may be electronic routers (digital cross connects, DXC) or optical switches (optical cross connects, OXC) connected by fiber links with up to λ wavelengths on each. The network is represented (Fig. 8) as a multigraph $G = (V, E)$ with $|V| = N$ nodes and $|E| = M$ edges (one for each wavelength λ in the optical layer).

We assume that the traffic is unsplittable, i.e. a traffic demand is routed over a single lightpath. In addition, we explicitly considered the influence of traffic on power consumption by using realistic data for traffic demands, network topologies, link costs, and energy requirements of single network elements. Specifically, the amount of power consumed by the NEs depends on the type of device and on the type and load of traffic that it is currently supporting (e.g. an

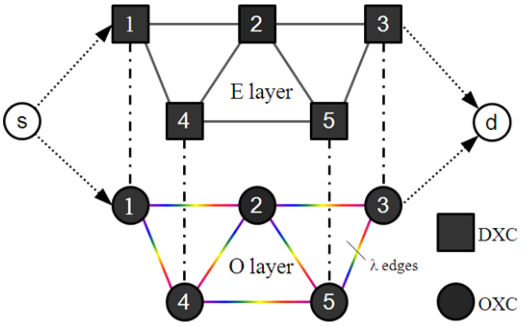


Fig. 8 Schematic representation of the network model with the electronic (E) and optical (O) layers

Table 2 Traffic supported by the devices

Type of device (NE)	Type of traffic
Electronic router[a]	Electronic
Optical switch[b] with WC capability	Optical (with or without WC)
Optical switch[b] without WC capability	Optical (without WC)
Fiber optic	Optical (without WC)
Optical amplifier	Optical (without WC)
3R regenerator	Electronic

[a] DXC

[b] OXC

OXC may support transparent optical traffic with or without WC). Even though the energy consumption of current node architectures does not scale with traffic (the energy demand of heavily loaded devices is only 3% greater than that of idle ones [10]), energy-aware architectures that adapt their performances to the traffic load lowering the power requirement under low traffic loads are strongly advised and are being designed [29]. Consequently, we assume that the power consumption of the NEs, i.e. both network nodes and links, consists of two factors. When turned on, a NE consumes a constant amount of power depending on the router size and technology (measured in $J/s = W$) and independent on the traffic load (fixed power Θ). The second factor (proportional power ε) consists of an amount of power proportional to type and quantity of the traffic load (measured in nJ/bit or, equivalently, in $W/Gbps$). The overall power drained by the WDM network is thus given by the sum of the fixed and proportional powers of the NEs subject to the current traffic load and varies with the routing of the connection requests. This implies that, as the NEs are always turned on, the routing optimization process works "only" on the proportional power. The power consumption functions of an electronic and an optical router are reported in Fig. 7 (optical router values not in scale). Three types of traffic are represented: (1) electronic traffic in the electronic router and (2) optical traffic with and (3) without WC in the optical switches. The Table 2 reports the types of network element and the corresponding supported traffic types. Note that each type of traffic accounts for different power consumption when traversing NEs, as explained in the following. We observed that the electronic traffic grows quickly with respect to the optical traffic and that, among the optical traffic, the WC actually consume a not negligible quantity of energy. As the power consumption functions are obtained by linear combinations of the logarithmic and the line functions, the complete gamma of slopes can be represented by the actual curves.

We also assume that all the nodes have the possibility to convert wavelengths, either in the electronic or in the optical domain, depending on their technology. In the electronic domain, the full range of operations is supported: wavelengths routing/switching, wavelengths add/drop, WC, 3R regeneration; in the optical domain the operations supported are the transparent wavelength switching/pass-through and the WC.

3.2 Energy-aware routing and wavelength assignment

In this section, three ILP formulations of the problem of energy-aware RWA in WDM networks with dual power sources are given, with different objective functions. First, the problem of minimizing the overall GHG emissions (*MinGas-RWA*) is presented. Next, the problem of minimizing the overall network power consumption (*MinPower-RWA*) is discussed. Finally, to obtain a reference for comparison, a minimum cost RWA (*MinCost-RWA*)—i.e., energy-unaware RWA—is derived and used. These formulations extend our previous work [32] to comprise the connection requirements on the guaranteed bandwidth (lightpath bitrate) thus supporting lightpath with different bitrates, and to prove its effectiveness against the well-known NSFNET network topology. In Table 3 we report the problem statements of the three ILP formulations and in Table 4 the associated notation; note that the three ILPs have the same inputs and constraints, but different objective function.

3.2.1 *Energy-aware RWA at minimum GHG emissions (MinGas-RWA)*

The energy-aware RWA in WDM networks with dual power sources (*MinGas-RWA*) is formalized as an ILP problem. The objective is to route the requested lightpaths so that the overall network GHG emissions are minimized. Only NEs powered by dirty energy sources emit GHGs, whilst NEs powered by green energy sources do not emit any GHG at all. The ILP problem can be mathematically formulated as follows.

The objective function for *MinGas-RWA* is:

$$Minimize\ DC_{G(V,E)} + \log TC_{G(V,E)} \qquad (1)$$

Table 3 Problem statements of the three ILP formulations

Schematic view

Given

(1) the physical network topology comprising routers and links, in which links have a known capacity and cost,[a]

(2) the knowledge of the average amount of traffic exchanged by any source/destination node pair,

(3) the maximum link utilization that can be supported (wavelength link capacity),

(4) the energy model (power consumption of each link and node),

(5) a set of k candidate paths (routes) between any source/destination node pair,

Find

the routes that must be used

In order to

[MinGas-RWA]	[MinPower-RWA]	[MinCost-RWA]
minimize the total GHG emissions, and, as secondary objective, the minimization of the total power consumption,	minimize the total power consumption, and, as secondary objective, the minimization of the total installation cost,	minimize the total installation cost,[b]

Subject to

traffic demand volume constraint, maximum link utilization constraint.

[a] The cost, the wavelength conversion and the 3R regeneration are considered in the selection phase of the k candidate paths (routes) between the source/destination node pairs, though they will have different cost impacts on the problem

[b] it is assumed that the installation cost is proportional to the number of wavelengths required and to the length of the chosen lightpaths

Subject to the following constraints:

$$DC_{G(V,E)} = \sum_{n\in V} g_n \cdot \left(\begin{array}{l} \Theta_n + \varepsilon_n^{t1} \cdot \sum_{sd,k,b:n\in\pi^{sd,k,b},n\neq s,d} w^{sd,k,b} \\ \quad \cdot b \cdot (1 - x_n^{sd,k,b}) \\ \quad + \varepsilon_n^{t2} \cdot \sum_{sd,k,b:n\in\pi^{sd,k,b},n\neq s,d} w^{sd,k,b} \\ \quad \cdot b \cdot x_n^{sd,k,b} \\ \quad + \varepsilon_n^{t3} \cdot \sum_{sd,k,b:n=s,d} w^{sd,k,b} \cdot b \end{array} \right) + \sum_{(i,j)\in E} h_{ij} \cdot \left(\left\lfloor \frac{\ell_{ij}}{\Lambda} \right\rfloor \times \left(\Psi_{ij} + \delta_{ij} \cdot \sum_{sd,k,b:(i,j)\in\pi^{sd,k,b}} w^{sd,k,b} \cdot b \right) \right) \tag{2}$$

$$TC_{G(V,E)} = \sum_{n\in V} \left(\begin{array}{l} \Theta_n + \varepsilon_n^{t1} \cdot \sum_{sd,k,b:n\in\pi^{sd,k,b},n\neq s,d} w^{sd,k,b} \\ \quad \times b \cdot (1 - x_n^{sd,k,b}) \\ \quad + \varepsilon_n^{t2} \cdot \sum_{sd,k,b:n\in\pi^{sd,k,b},n\neq s,d} w^{sd,k,b} \\ \quad \times b \cdot x_n^{sd,k,b} + \varepsilon_n^{t3} \cdot \sum_{sd,k,b:n=s,d} w^{sd,k,b} \cdot b \end{array} \right) + \sum_{(i,j)\in E} \left(\left\lfloor \frac{\ell_{ij}}{\Lambda} \right\rfloor \times \left(\Psi_{ij} + \delta_{ij} \cdot \sum_{sd,k,b:(i,j)\in\pi^{sd,k,b}} w^{sd,k,b} \cdot b \right) \right) \tag{3}$$

$$\sum_{k,b} w^{sd,k,b} = t^{sd,b} \quad \forall s, d \in V,\ \forall b \in B \tag{4}$$

$$\sum_{sd,k,b:(i,j)\in\pi^{sd,k,b}} w^{sd,k,b} \le a_{ij} \quad \forall (i,j) \in E, \forall b \in B \tag{5}$$

$$w^{sd,k,b} \in N, \quad \forall s, d \in V, \forall k, \forall b \in B \tag{6}$$

The objective (1) is the minimization of the network power consumption $DC_{G(V,E)}$ due to the network elements powered by dirty energy sources (as we want to minimize GHG emissions) and—among the solutions at minimum power consumption—the minimization of the total power consumption of the network $DC_{G(V,E)}$. Equation (2) sets the power consumption of the network elements in $G(V,E)$ due only to dirty power sources, whilst (3) indicates the total power consumption of the network elements in $G(V,E)$ evaluated in the energy model. Constraint (4) selects the routes for the lightpaths among the k pre-computed ones and assures that the whole traffic demand matrix is satisfied. Constraint (5) ensures that the maximum number of lightpaths passing on a link does not exceed the number of available wavelengths on that link. Constraint (6) imposes the integrality of the ILP problem by forcing integer values for the variables $w^{sd,k,b}$. Note that the fixed power consumptions in (2) and (3) are reported only for completeness sake but they are not involved in the optimization process (as sleep mode is not considered, the optimization is realized only on the variable energy consumptions).

Table 4 Summary of the notation used for the ILP model

Input parameters	Meaning
$G(V, E)$	directed graph representing the physical network topology; V set of vertices that represent the network nodes; E the set of edges that represent the network links; $\|V\| = N$, $\|E\| = M$. Each network link has a different (maximum) bitrate (i.e., bandwidth capacity);
a_{ij}	number of wavelengths available on link (i, j);
b_{ij}	bandwidth capacity of link (i, j);
ℓ_{ij}	length of link (i, j) (in km);
Λ	maximum length of links without need of amplification (in km);
$t^{sd,b}$	number of lightpaths to be established from s to d with required bandwidth $b \in B$; bandwidth ranges from 54 Mbps (1 OC-unit) to 40 Gbps (768 OC-units); *OC-units* $= \{1, 3, 12, 24, 48, 192, 768\}$; $\{t^{s,d}\}_{s,d \in V}$ is the traffic matrix;
$\pi^{sd,k,b}$	k-th pre-computed route[a] from s to d satisfying the bandwidth requirement of b bps;
$\rho^{sd,k,b}$	the geographical length of route $\pi^{sd,k,b}$ (in km);
Θ_n	fixed power (W) of node n;
$\varepsilon_n^{t_1}$	proportional energy (nJ/bit) for transporting one bit of *transparent pass-through* traffic at node n;
$\varepsilon_n^{t_2}$	proportional energy (nJ/bit) for transporting one bit of *opaque pass-through* traffic at node n (e.g. 3R regeneration or opaque wavelength conversion);
$\varepsilon_n^{t_3}$	proportional energy (nJ/bit) for *add/drop* one bit at node n;
Ψ_{ij}	fixed power (W) for devices in link (i, j), (e.g. optical amplifiers);
δ_{ij}	proportional energy (nJ/bit) for transporting one bit through link (i, j); it is assumed that each device (e.g. OA) on the same link (i, j) has the same fixed and proportional consumptions;
$x_n^{sd,k,b}$	identify the presence[b] of O/E/O conversion at the node n: $x_n^{sd,k,b} = \begin{cases} 1 & \text{if } n \in \pi^{sd,k,b} \text{ and } \pi^{sd,k,b} \text{ undergoes O/E/O conversion at node } n \\ 0 & \text{if } n \notin \pi^{sd,k,b} \text{ or } \pi^{sd,k,b} \text{ transparently passes through node } n \end{cases}$
g_n	identifies the presence of dirty energy source at node n: $\forall n \in V, g_n = \begin{cases} 0 & \text{if node } n \text{ is powered by a green energy source} \\ 1 & \text{if node } n \text{ is powered by a dirty energy source} \end{cases}$
h_{ij}	identifies the presence of green energy source at link (i, j): $\forall (i, j) \in E, h_{ij} = \begin{cases} 0 & \text{if link } (i, j) \text{ is powered by a green energy source} \\ 1 & \text{if link } (i, j) \text{ is powered by a dirty energy source} \end{cases}$
Variables	Meaning
$w^{sd,k,b}$	integer, indicates the number of lightpaths using route $\pi^{sd,k,b}$ (on the same route there may be several lightpaths using different wavelengths);
$TC_{G(V,E)}$	indicates the total power consumption of the NEs in $G(V, E)$ evaluated in the chosen traffic model;
$DC_{G(V,E)}$	indicates the power consumption of the NEs in $G(V, E)$ due only to dirty power sources.

[a] In this ILP formulation, a set of pre-computed routes is used for routing the demand lightpath in order to reduce the time complexity, leading to a sub-optimal solution of the ILP. The k paths satisfy the requirement on the bandwidth (b bps) since they are found by the *bandwidth constrained* k-shortest paths algorithm.

[b] Note that 3R regeneration and opaque wavelength conversion are implicitly considered in this matrix and this information will be used in the power consumption calculus

3.2.2 *Energy-aware RWA at minimum power consumption (MinPower-RWA)*

The objective of the *MinPower-RWA* problem is to minimize the overall power consumption regardless of the energy sources types and, thus, of the GHG emissions. The set of the input parameters is the same as the *MinGas-RWA* problem except for the g_n and h_{ij} vectors which are no longer necessary; also, an additional constant ξ is considered,

$$\xi : 0 < \xi \cdot \left(\sum_{n \in V} \Theta_n + \sum_{(i,j) \in E} \Psi_{ij} \right) < 1 \tag{7}$$

The mathematical formulation of *MinPower-RWA* is the same as above, with a different objective function:

$$Minimize\ TC_{G(V,E)} + \xi \cdot \sum_{sd,k,b} w^{sd,k,b} \cdot \rho^{sd,k,b} \tag{8}$$

and taking (3) (4) (5) (6) as constraints.

The objective function (8) is the minimization of the total network power consumption due to fixed and proportional power consumed by all the devices installed in the network, and—among the solutions at minimum power consumption—the minimization of the installation cost.

3.2.3 Minimum Cost RWA (MinCost-RWA)

The objective of the *MinCost-RWA* problem is the minimization of the installation cost regardless of the NEs energy consumptions and GHG emissions. It will try to aggregate as much lightpaths as possible while minimizing their physical lengths. The objective function in this case is:

$$Minimize \sum_{sd,k,b} w^{sd,k,b} \cdot \rho^{sd,k,b} \tag{9}$$

and the constraints are those of (4) (5) (6).

4 Model evaluation

In this section, we analyze the model effectiveness through ILP optimizations exploiting minimum power consumption, minimum GHG emissions and mínimum installation cost through simulation on the well-known NSFNET network topology. The obtained results have been briefly discussed to show the potential benefits achievable through the presented cross-layer optimization approach.

4.1 The proof of concept simulation environment

We used the NSFNET core optical network with 14 nodes and 21 bidirectional fiber links each with 16 wavelengths. Simulations were performed under different power distribution systems, with green energy sources powering 25, 50 and 75% of the NEs and randomly generated traffic matrices. Connection requests are fully satisfied, i.e., the blocking probability is kept strictly null. To solve the ILP problems the CPLEX software tool was used on an Intel® Xeon® 2.5 GHz dual processor Linux server. The available memory (physical RAM + swap area) amounted to 16 GBytes. To reduce the notable requirements in terms of computational and memory resources, we first bound the problem dimension by restricting the paths' alternatives to a static set of k pre-computed routes, obtained by using a traditional K-SPF algorithm and hence satisfying the traditional network management objectives without considering any energy-related information. Secondly, we limited the depth of the branch-and-bound/cut algorithms after calculating a pre-definite number of integer solutions. While such simplification techniques are certainly useful to contain the computational burden, the solution they produce is an approximation of the actual optimal (in terms of power consumption) virtual network topology built on the available physical infrastructure. However in these cases the ILP approach maintains its added value, as far as the approximated solutions can be close to the exact one. Some of the selected paths would probably not be the best ones, but the resulting power savings could be substantial without significant losses on the other optimization objectives.

4.2 Results and discussion

The energy consumption (during 1 year time period) resulting from the three ILP RWA strategies with 50% of the NEs powered by green energy sources is reported in Fig. 9. As expected, the *MinCost-RWA* is the most energy consuming strategy, whilst the *MinPower-RWA* is the best strategy as for the energy consumption, but the best one as GHG emissions is the *MinGas-RWA*. Anyway, the difference in energy consumption between the latter two strategies is lower than 14% in the worst case. This result was somehow expected, as the minimum power RWA strategy attempts to save as much energy as possible regardless of the sources of energy, whereas the minimum GHG emissions may route the lightpaths on longer—thus, more energy consuming—paths but preferring those NEs that are powered by green energy sources. At low loads, *MinGas-RWA* attempts to use only green-powered nodes, at the expense of possibly choosing longer paths. The effect of these suboptimal choices is visible at higher loads, when the overall energy consumption rises more steeply that of *MinPower-RWA*. This becomes relevant at network loads as high as 70%, whereas in the 30%–70% operating range the savings achieved by *MinGas-RWA* with respect to *MinCost-RWA* remain consistently substantial. As for the energy consumption, compared with *MinCost-RWA, MinGas-RWA* saved an average of 18% of energy while *MinPower-RWA* reached savings up to 30% of the overall energy consumption.

Besides the saving in energy consumption, *MinGas-RWA* achieves to save also considerable quantity of CO_2. For a medium loaded network (50% of routed lightpaths), where one half of the NEs are powered by green power plants and the other half are powered by fuel-based power plants, *MinGas-RWA* strategy saves an average of 37,500 kg of CO_2 per year (see [1] as a reference value for the emitted CO_2).

In Fig. 10, we compared the estimated CO_2 emissions with the three strategies at different network loads, where one half of the NEs are powered by green energy sources and

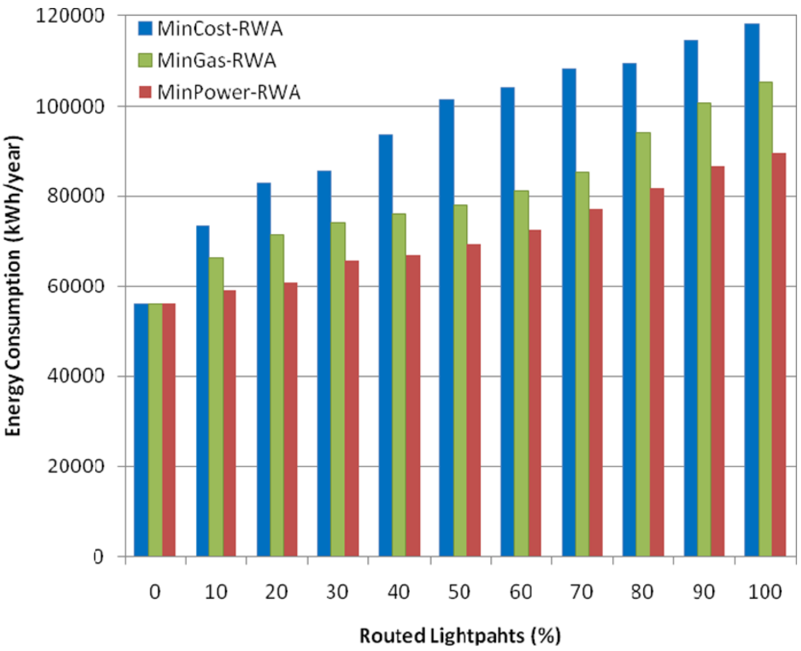


Fig. 9 Network energy consumption *vs* traffic load with the three ILP strategies

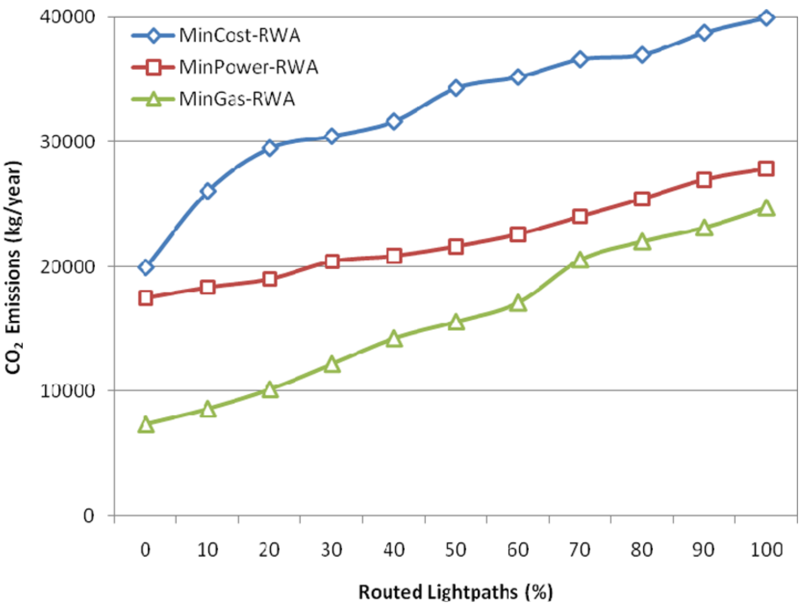
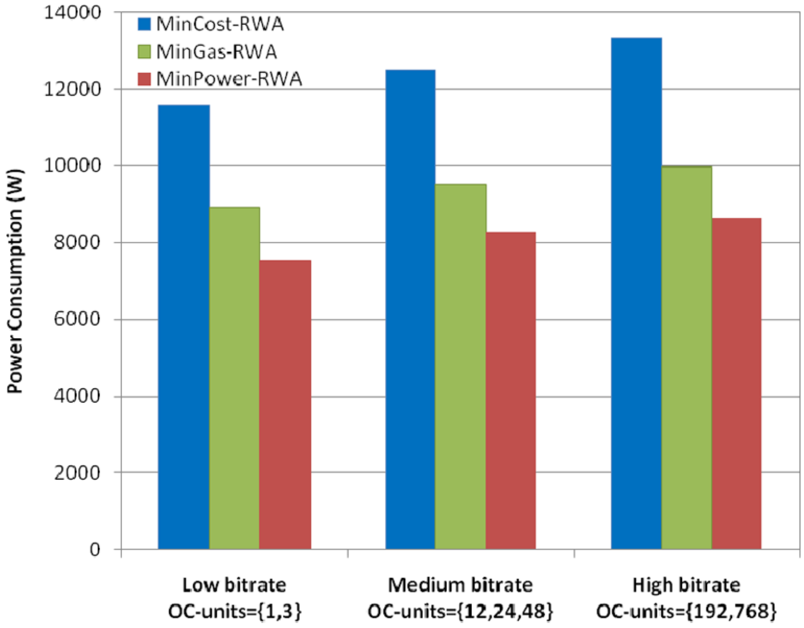


Fig. 10 Network CO_2 emissions *vs* network load with the three ILP strategies

the other half by fuel-based power plants. As can be seen, at low loads the *MinGas-RWA* strategy achieves prominent CO_2 savings (only about 33% of CO_2 were emitted with respect to *MinCost-RWA* and about 50% relative to *MinPower-RWA*), whilst, as the network load increases, the difference between the *MinGas-RWA* and the *MinPower-RWA* strategies decreases, because at higher loads it becomes more and more difficult to satisfy the demand without resorting to dirty-powered nodes. In other words, at high loads, minimizing the overall power consumption implicitly leads to the minimization of the concomitant CO_2 emissions, while at midrange loads the CO_2 savings induced by *MinGas-RWA* are significant.

We also explored the power consumptions when different lightpath bitrates are considered. In the simulation, the network load was kept constant (at 50% of its maximum capacity) while connection requests with different bandwidth requirements (bitrates) are generated, ranging from many low-speed connections to few high-speed ones (Fig. 11). As a general trend, we observe that—though the traffic load is constant—higher bitrates are associated with higher power consumptions. This behavior is due to the fact that smaller connections have more possibilities to be routed over less energy-demanding routes than larger ones which are instead more likely to be routed over high capacity network routes. *MinCost-RWA* power consumption grows quite linearly with

Fig. 11 Network power consumption with different lightpath rates (constant traffic load at 50%)



the increasing of the lightpath bitrates, whilst the *MinGas-RWA* and, above all, the *MinPower-RWA* exhibit a more constant behavior: thanks to the energy-awareness of such ILPs, they are able to accommodate more profitably the connection requests, even if also in these strategies an increase in the power consumption is still observed due to the lower eligible routes. We also note that the *MinCost-RWA* power consumption is always higher than with the other two strategies even at higher bitrates, showing that the energy-awareness may help to substantially compensate the higher energy consumption due to higher bitrates.

Finally, we have analyzed the dependency of the power consumption from the actual values of the fixed and variable components of the power draw by an interface, expressed as a function of the link rate at which the interface operates. Note that, since a (unidirectional) link is attached to each interface, the set of links E in the aforementioned network graph representation $G(V, E)$ actually coincides with the set of interfaces. Each interface has its own native speed: $\forall i \in E$, $v_i \in R = \{10 \text{ Mbps}, 100 \text{ Mbps}, 1000 \text{ Mbps}, 10000 \text{ Mbps}\}$ represents the *native link rate* of interface i. If the link rate is fixed, the power draw of an interface will depend mainly on the link rate, with minor variations due to architecture, circuitry, and components. When using an ALR, instead, the power consumption of an interface i depends not only on the *working* link rate r_i but also on the *native* link rate v_i. In other words, a given throughput t_d results in different power consumption depending on the interface native link rate v_i: in this case, slower interfaces consume less power than faster ones for the same throughput t_d, even if they work at the same rate r_i. This result, quite surprising if we consider that slower interfaces consume more energy per bit than faster ones, may be explained by considering the different technologies adopted for reaching higher link rates (mainly based on advanced modulation techniques [33]) that lead to greater fixed power consumption for faster interfaces. In fact, like routers, also the interfaces have fixed and variable power consumption. The fixed part is always present just for the interface to stay up and accounts for the control circuits, while the variable traffic-proportional power consumption is due to the transceivers. In the following we model such energy consumptions and show a breakdown of the different energy components in a 10 Gbps interface.

In general, to model the fixed and the variable energy consumption, we define $\{\Psi(v_i, r_j) | j = 1, 2, \dots, m\}$ where $\Psi(v_i, r_j)$ (see Table 5) is the power consumption of the interface $i \in E$ with native speed $v_i \in R$ operating at link rate $r_j \in R$ and $\Psi(v_i, r_j) < \Psi(v_i, r_k)\ \forall j < k$.

In Fig. 12 we plotted the CO_2 emissions as a function of the average link rate, assuming a uniform distribution of the working link rates between 0 and the native link rate and a real-world distribution of values for the $\Psi(v_i, r_j)$ as given in [1, 23]. On the horizontal axis there is the percentage of interfaces operating at the native link rate. A notable characteristic is that the savings induced by an extensive use of an adaptive link rate are not as dramatic as one may expect, although sensitivity is slightly higher in the case of *MinPower-RWA* and *MinGas-RWA*.

5 Conclusions

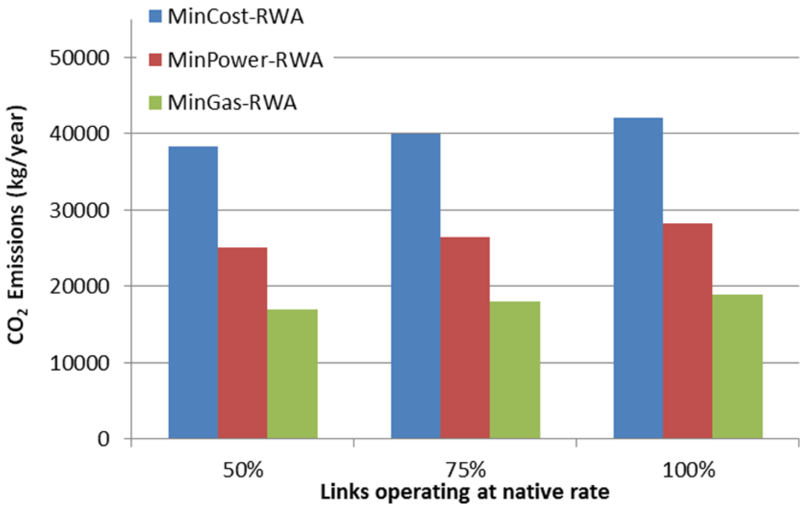
The ICT sector has the fundamental capability of acting as drawing factor to drive the development of energy-efficient technological innovations for both industry and society, although any action in the direction of energy-efficiency may

Table 5 Power consumptions of interfaces working at different rates

v_i	r_i	Mbps	Power consumption
$\forall v_i \in R$	Off $/r_0$	0/0	$\Psi(v_i, Off/r_0) \cong 0$[a]
v_1: 10	r_1	10	$\Psi(v_1, r_1)$
v_2: 10^2	r_1/r_2	$10/10^2$	$\Psi(v_2, r_1)/\Psi(v_2, r_2)$
v_3: 10^3	$r_1/r_2/r_3$	$10/10^2/10^3$	$\Psi(v_3, r_1)/\Psi(v_3, r_2)/\Psi(v_3, r_3)$
v_4: 10^4	$r_1/r_2/r_3/r_4$	$10/10^2/10^3/10^4$	$\Psi(v_4, r_1)/\Psi(v_4, r_2)/\Psi(v_4, r_3)/\Psi(v_4, r_4)$

[a] in LPI, the device only send signals during short refresh intervals and stay quite during large intervals so the power consumption in the LPI mode is almost 0

Fig. 12 Average emitted CO_2 at different ratios of links operating at their native rate



result in direct power and cost savings in the short and medium term, while other indirect effects will be only observed on the long term on both the environment and human health. Anyway, the massive introduction of energy efficiency within the network world requires a coordinated effort of equipment vendors, governments, and service providers to identify technological standards, best practices, and solutions to support the necessary changes in the basic construction and functional requirements for network equipment and control plane algorithms. Accordingly, several energy-aware ILP formulations exploiting dual energy sources have been presented along with an energy model in which no sleep mode is available but the optimization relies only on the traffic-variable power consumption of the NEs. Two ILP formulations have been presented: minimum power (*MinPower-RWA*) and minimum GHG emissions (*MinGas-RWA*) strategies with the objectives to minimize respectively the absorbed energy and the emitted GHG. Results show that the *MinPower-RWA* strategy may save a considerable amount of energy by routing the lightpaths on minimum consuming NEs and that the GHG emitted may be notably reduced by the *MinGas-RWA* strategy that prefers NEs powered by green energy sources. As drops are observed in the day/night traffic at core network nodes, there is room for some possible optimizations by putting NEs into sleep mode only partially (per-interface sleep mode). In fact, putting into sleep mode single interfaces or line cards may have some sense, saving up to 50% of the total router power, but modifications to current router architecture and routing protocols need to be investigated. Renewable energy sources may vary their availability with time (e.g. solar panels only generate electricity during the day). While in the current work we handled the availability of green and dirty sources in a static way, in future works statistically variable green energy sources may be considered within a totally dynamic scenario in which the availability of the different types of renewable energy sources can be associated with the variations of the day time and traffic load (e.g. night/day cycle). We are confident that the above efforts, together with incrementing network eco-sustainability, will improve the sustainable growth and—in the long run—the society prosperity.

Acknowledgements This work was supported in part by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.

References

1. BONE project (2009). *WP 21 topical project green optical networks* (Report on year 1 and updated plan for activities). NoE, FP7-ICT-2007-1 216863 BONE project, Dec. 2009.
2. Baliga, J., et al. (2009). Energy consumption in optical IP networks. *Journal of Lightwave Technology*, *27*(13), 2391–2403.
3. Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *86*, 82–85.
4. Gilder, G. F. (2000). *Telecosm: how infinite bandwidth will revolutionize our world*. New York: The Free Press.
5. Jevons, W. S. (1866). *The coal question; an inquiry concerning the progress of the nation, and the probable exhaustion of our coalmines*. London: Macmillan.
6. Smart Grid Task Forces (2011). Office of Electricity Delivery & Energy Reliability, Washington, DC. http://www.smartgrid.gov/.
7. Karg, L. (2010). Consult GmbH, *keynote speech*. In *e-Energy 2010*, Passau, Germany Apr. 2010.
8. Lam, W. (2009). Getting the most out of Google's solar panels. Jul. 2009 [online]. Available: https://docs.google.com/present/view?id=dfhw7d9z_0gtk9bsgc.
9. St Arnaud, B. (2011). ICT and Global Warming: Opportunities for Innovation and Economic Growth. http://docs.google.com/Doc?id=dgbgjrct_2767dxpbdvcf.
10. Chabarek, J., Sommers, J., Barford, P., Estan, C., Tsiang, D., & Wright, S. (2008). Power awareness in network design and routing. In *Proc. IEEE INFOCOM*.
11. Gupta, M., & Singh, S. (2003). Greening of the Internet. In *Proc. of the ACM SIGCOMM*, Karlsruhe, Germany
12. Feng, M. Z., Hilton, K., Ayre, R., & Tucker, R. (2010). Reducing NGN energy consumption with IP/SDH/WDM. In *Proceedings of the 1st international conference on energy-efficient computing and networking*, Passau, Germany (pp. 187–190).
13. Vereecken, W., Van Heddeghem, W., Colle, D., Pickavet, M., & Demeester, P. (2010). Overall ICT footprint and green communication technologies. In *Proc. of ISCCSP 2010*, Limassol, Cyprus Mar. 2010.
14. Ricciardi, S., Careglio, D., Fiore, U., Palmieri, F., Santos-Boada, G., & Solé-Pareta, J. (2011). Analyzing local strategies for energy efficient networking. In *Proceedings of sustainable networking SUNSET 2011, IFIP networking 2011*, Valencia, 9–13 May 2011.
15. Saleh, B. E. A., & Teich, M. C. (1991). *Fundamentals of photonics*. New York: Wiley.
16. The Energy Consumption Rating (ECR) initiative (2011). [online]. Available: http://www.ecrinitiative.org/.
17. Miller, D. A. B. (2010). Are optical transistors the logical next step? *Nature Photonics*, *4*(1), 3–5.
18. D'Ambrosia, J. 100 Gigabit Ethernet and beyond. *IEEE Communications Magazine*, March 2010.
19. Anderson, J., & Traverso, M. (2010). Optical transceivers for 100 gigabit Ethernet and its transport. *IEEE Communications Magazine*, *48*(3), S35–S40.
20. Lange, C. (2009). Energy-related aspects in backbone networks. In *Proc. ECOC 2009*, Vienna, Austria, Sep. 2009.
21. Tucker, R. S., et al. (2009). Evolution of WDM optical IP networks: a cost and energy perspective. *IEEE/OSA Journal of Lightwave Technologies*, *27*(3), 243–252.
22. The Green Grid (2008). *The green grid data center power efficiency metrics: PUE and DCiE*. Technical Committee White Paper.
23. Christensen, K., Reviriego, P., Nordman, B., Bennett, M., Mostowfi, M., & Maestro, J. A. (2010). IEEE 802.3az: the road to energy efficient Ethernet. *IEEE Communications Magazine*, *48*(11), 50–56.
24. Hays, R. (2008). Active/idle toggling with low-power idle. In *IEEE 802.3az task force group meeting*. [online]. Available: http://www.ieee802.org/3/az/public/jan08/hays_01_0108.pdf.
25. Zhai, B., Blaauw, D., et al. (2004). Theoretical and practical limits of dynamic voltage scaling. In *DAC*.
26. Christensen, K., & Nordman, B. (2005). Reducing the energy consumption of networked devices, IEEE 802.3 tutorial, July 19, 2005, San Francisco [online]. Available: http://www.csee.usf.edu/~christen/energy/ieee_tutorial.pdf.
27. Van Heddeghem, W., De Groote, M., Vereecken, W., Colle, D., Pickavet, M., & Demeester, P. (2010). Energy-efficiency in telecommunications networks: link-by-link versus end-to-end grooming. In *Proc. of ONDM 2010*, Kyoto, Japan, Feb. 1–3 2010.
28. Tucker, R. S. (2011). Modelling Energy Consumption in IP Networks, [online]. Available: http://www.cisco.com/web/about/ac50/ac207/crc_new/events/assets/cgrs_energy_consumption_ip.pdf.
29. Aleksic, S. (2009). Analysis of power consumption in future high-capacity network nodes. *Journal of Optical Communications and Networking*, *1*(3), 245–258.
30. Energy Star (2011). Small network equipment [online]. Available: http://www.energystar.gov/index.cfm?c=new_specs.small_network_equip.
31. Ricciardi, S., Careglio, D., Palmieri, F., Fiore, U., Santos-Boada, G., & Solé-Pareta, J. (2010). Energy-oriented models for WDM networks. In *Proceedings of 7th international ICST conference on broadband communications, networks, and systems (Broadnets 2010)*, Athens, Greece, 25–27 Oct. 2010 (pp. 1–4).
32. Ricciardi, S., Careglio, D., Palmieri, F., Fiore, U., Santos-Boada, G., & Solé-Pareta, J. (2011). Energy-aware RWA for WDM networks with dual power sources. In *Proceedings of 2011 IEEE international conference on communications (ICC 2011)*, Kyoto, Japan, June 5–9, 2011.
33. Nortel (2011). *A comparison of next-generation 40-Gbps technologies*. White paper [online]. Available: http://www.nortel.com/solutions/collateral/nn122640.pdf.
34. Koroneos, C. J., & Koroneos, Y. (2007). Renewable energy systems: the environmental impact approach. *International Journal of Global Energy Issues*, *27*(4), 425–441.

Sergio Ricciardi received the degree summa cum laude in Computer Science from the University of Naples Federico II, Italy, in 2006 and the M.Sc. degree with honours in Computer Architecture, Networks and Systems from the Technical University of Catalonia (UPC), Spain, in 2010. He worked with the Federico II University and with the Italian National Institute for Nuclear Physics (INFN) within several national and international projects. From 2008 he is research associate in the Advanced Broadband Communications Center (CCABA) at the Department of Computer Architecture of the UPC. His current activities concern energy-efficient architectures and energy-aware RWA algorithms and protocols for optical networks and grid/cloud computing. His research interests are mainly focused on energy-oriented routing, optimization algorithms and topology management for transparent and opaque optical networks.

Davide Careglio (S'05–M'06) received the M.Sc. and Ph.D. degrees in telecommunications engineering both from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2000 and 2005, respectively, and the Laurea degree in electrical engineering from Politecnico di Torino, Turin, Italy, in 2001. He is currently an Associate Professor in the Department of Computer Architecture at UPC. Since 2000, he has been a Staff Member of the Advanced Broadband Communication Center. His research interests include networking protocols with emphasis on optical switching technologies, and algorithms and protocols for traffic engineering and QoS provisioning. He is the coauthor of more than 80 publications in international journals and conferences. He has participated in many European and national projects in the field of optical networking and green communication.

Germán Santos-Boada obtained his M.Sc. degree in Telecom Engineering in 1978, and his Ph.D. in 1993, both from the Technical University of Catalonia (UPC). He worked for Telefónica as manager of engineering from 1984 up to 2007 and simultaneously he joined the Computer Architecture Department of UPC as a partial time Assistant Professor. Currently he is full time Assistant Professor with this department. Dr. Santos current research interests are Quality of Service provisioning in next generation optical access networks and optical energy-aware network modeling. He is currently involved in the COST 804 action. (german@ac.upc.edu)

Josep Solé-Pareta obtained his M.Sc. degree in Telecom Engineering in 1984, and his Ph.D. in Computer Science in 1991, both from the Technical University of Catalonia (UPC). In 1984 he joined the Computer Architecture Department of UPC. Currently he is Full Professor with this department. He did a Postdoc stage (summers of 1993 and 1994) at the Georgia Institute of Technology. He is co-founder of the UPC-CCABA, and UPC-N3cat. His publications include several book chapters and more than 150 papers in relevant research journals (>25), and refereed international conferences. His current research interests are in Nanonetworking Communications, Traffic Monitoring, Analysis and High Speed and Optical Networking and Energy Efficient Transport Networks, with emphasis on traffic engineering, traffic characterization, MAC protocols and QoS provisioning. He has participated in many European projects dealing with Computer Networking topics.

Ugo Fiore leads the Network Operations Center at the Federico II University, in Naples. He began his career with Italian National Council for Research and has also more than 10 years of experience in the industry, developing software support systems for telco operators. His research interests focus on optimization techniques and algorithms aiming at improving the performance of high-speed core networks. He is also actively pursuing two other research directions: the application of nonlinear techniques to the analysis and classification of traffic; security-related algorithms and protocols.

Francesco Palmieri is an assistant professor at the Engineering Faculty of the Second University of Napoli, Italy. His major research interests concern high performance and evolutionary networking protocols and architectures, routing algorithms and network security. Since 1989, he has worked for several international companies on networking-related projects and, starting from 1997, and until 2010 he has been the Director of the telecommunication and networking division of the Federico II University, in Napoli, Italy. He has been closely involved with the development of the Internet in Italy as a senior member of the Technical-Scientific Advisory Committee and of the CSIRT of the Italian NREN GARR. He has published a significant number of papers in leading technical journals and conferences and given many invited talks and keynote speeches.

Computer Networks 56 (2012) 2420–2442

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

An energy-aware dynamic RWA framework for next-generation wavelength-routed networks

Sergio Ricciardi [a,*], Francesco Palmieri [b], Ugo Fiore [c], Davide Careglio [a], Germán Santos-Boada [a], Josep Solé-Pareta [a]

[a] Department of Computer Architecture, Technical University of Catalonia, c. Jordi Girona 1-3, 08034 Barcelona, Spain
[b] Department of Information Engineering, Second University of Naples, v. Roma 29, 81031 Aversa, Italy
[c] Center of Information Services, University of Naples Federico II, v. Cinthia, 80126 Naples, Italy

ARTICLE INFO

Article history:
Received 30 September 2011
Received in revised form 23 February 2012
Accepted 20 March 2012
Available online 29 March 2012

Keywords:
Green networks
Cross-layer optimizations
Energy consumption
GHG emissions

ABSTRACT

Power demand in networking equipment is expected to become a main limiting factor and hence a fundamental challenge to ensure bandwidth scaling in the next generation Internet. Environmental effects of human activities, such as CO_2 emissions and the consequent global warming have risen as one of the major issue for the ICT sector and for the society. Therefore, it is not surprising that telecom operators are devoting much of their efforts to the reduction of energy consumption and of the related CO_2 emissions of their network infrastructures. In this work, we present a novel integrated routing and wavelength assignment framework that, while addressing the traditional network management objectives, introduces energy-awareness in its decision process to contain the power consumption of the underlying network infrastructure and make use of green energy sources wherever possible. This approach results in direct power, cost and CO_2 emissions savings in the short term, as demonstrated by our extensive simulation studies.

1. Introduction

The containment of power consumption and the reduction of the associated green house gases (GHG, mainly CO_2) emissions are emerging as new challenges for telecommunication operators. In fact, the rising energy costs due to the scarcity of fossil fuels, the increasingly rigid environmental standards and the growing power requirements of modern high-performance networking devices are imposing new constraints, further stressing the requirements towards an energy-aware business model in which the ecological footprint of the network elements (NEs) is explicitly taken into account. In this scenario, governments and society are endorsing the development of "green" renewable energy sources (such as solar panels, wind turbines, and geothermal plants) for powering NEs. Green energy sources are preferable with respect to the traditional "dirty" ones (e.g. coal, fuel, gas) since they do not emit GHG in the atmosphere while producing electrical energy. Nonetheless, green energy sources (e.g. wind, sun, tide) are not always available at all sites and are variable with time; for this reason, the NEs that are powered by green energy sources are also provided with the legacy, dirty sources. At the occurrence, the smart grid power distribution system switches to the dirty power supply without any energy interruption.

Recent studies [1] confirm that the use of optical technology in high-capacity switches and routers is more energy-efficient than electronic technology and that circuit-switched architectures consume significantly less than their packet-switched counterparts. However, despite the recent efforts in improving the energy-efficiency of the involved technological components [2], the amount of power to be spent worldwide for powering network

[*] Corresponding author.
E-mail addresses: sergior@ac.upc.edu (S. Ricciardi), fpalmier@unina.it (F. Palmieri), ufiore@unina.it (U. Fiore).

1389-1286/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2012.03.016

infrastructures can be globally quantified in the order of tens of gigawatts, corresponding to more than 1% of the worldwide electricity consumption [2] (to give an idea, the equivalent of 22 nuclear reactors are needed to generate such a huge power demand). Thus, limiting power consumption in network infrastructures can bring great benefits and reduce their overall ecological footprint, so that the need for a greener, energy-aware Internet is rapidly becoming a fundamental political, social and commercial issue. Furthermore, with the ever increasing demand for bandwidth, connection quality and end-to-end interactivity, computer networks are requiring more and more sophisticated and power-hungry devices, such as high-end routers, signal regenerators, optical amplifiers, reconfigurable add-and-drop multiplexers and very fast (digital signal) processing units. These components tend to increase the energy needs of global communication exponentially so that power consumption is becoming a significant limiting factor for the overall scalability of next-generation high-capacity telecommunication networks. In the next years, large-scale optical transport infrastructures will no longer be constrained mainly by their capacity, but rather by their energy consumption costs and environmental effects [3].

As a consequence, it is necessary to envisage how the next-generation network architectures and protocols can be modified to meet the purpose of energy-efficiency. Unfortunately, the rush for achieving energy-efficiency resulted in the fact that many of the solutions proposed to-date (e.g. [4,5]) tend to minimize *only* the energy consumption of the networks while disregarding the traditional network management goals such as the overall network load-balancing. It is instead mandatory to guarantee that the above modifications will not adversely affect the fundamental operators' optimization objectives of keeping the resource usage fairly balanced, to save on each available link sufficient free capacity for demands that may reasonably emerge in the infrastructure operating lifetime, and minimizing the network usage costs, considered as a static way of expressing operator preference to choose some favorite link resources. In the ideal case, new solutions should not only lower the ecological footprint, but also increase the offered quality-of-service such as the connection blocking probability.

Starting from the above considerations, we introduce [6] energy-awareness into control plane protocols whose goal is to properly condition the route/path selection mechanisms on relatively coarse time scales by privileging the use of green energy sources and energy-efficient links/switching devices, simultaneously taking advantage from the different users' demands across the interested network infrastructures. The selected paths are likely not to be the shortest or best ones, but the resulting power and GHG savings are substantial, and possible losses on the other optimization objectives (i.e. number of blocked connection requests) are taken into account and kept as low as possible. In such a way, the overall power consumption and GHG emissions can be minimized while the traditional optimization objectives (such as load-balancing) are not disrupted. In doing this, we combine all the notable features that a comprehensive energy-aware network model should have and put them together into a general routing and wavelength assignment (RWA) framework.

The RWA problem is known to be NP-complete [7] and in the dynamic case no optimality is possible since there is no previous knowledge of the connection requests that will be handled by the network. Therefore, we introduce a new heuristic method for efficiently calculating (in polynomial time) the routing information subject to power consumption constraints, taking into account also the specific kind of energy source (dirty or green) used for powering the traversed NEs. In order to evaluate the performance of our approach, we compared our approach with well-known RWA algorithms in the literature. The proposed approach, that in the following will be referred to as *GreenSpark*, introduces energy-awareness into the Spark framework [8], which is a two-stage integrated RWA scheme structured in a pre-selection phase where a number of k candidate paths satisfying the connection constraints are found and a final selection stage where the optimum path among the candidates determined in the previous phase is chosen according to a properly crafted heuristic. GreenSpark is a simple and effective two-stage on-line RWA scheme providing wavelength routing as well as grooming capabilities in the state-of-the-art hybrid electric-optical network infrastructure. In its first stage, this enhanced RWA scheme finds, for each new connection request, a set of feasible lightpaths satisfying both the users' specific end-to-end demands (QoS, bandwidth, etc.) and traditional optimization objectives, while in the second stage it bases its final choice on the aforementioned power and GHG containment requirements. In the end, it finally achieves an optimal trade-off between energy optimization and network/users requirements in an affordable computational time. GreenSpark differs from Spark for the second stage, which has been here introduced to meet the energy-related criteria. Furthermore, Spark used a special parameter (*kHop*) to explicitly limit the length of lightpaths, whilst in GreenSpark this is not needed anymore due to the additive nature of the energy consumption function: longer paths will have higher energy consumption and, thus, will have lower probability to be chosen for the routing of the connections.

GreenSpark is based on a totally flexible network model supporting heterogeneous equipment, in which the number and type of lambdas can vary on each link or node, together with the associated power consumption, and provides a fully dynamic path selection scheme in which the grooming policy is not predetermined but may vary, along with the evolution of the network traffic. We explicitly consider the influence of traffic on power consumption by using realistic data for traffic demands, network topologies, link costs, and energy requirements of the NEs.

This approach is also based on deeper network engineering considerations that make it behave very differently from the other already existing energy-aware networking approaches, mainly based on the concept of temporarily switching off entire devices or subsystems (the least used ones) in order to minimize energy consumption by rerouting the involved traffic. Such approaches, often referred as sleep mode [9], may be unpractical, especially for large and highly connected switching nodes, since many very expensive transmission links become unused, hence leaving significant capital investments (CAPEX) unproductive for the entire duration of the sleep interval. Furthermore, sleep mode drastically reduces the overall meshing degree, by limiting the network reliability, and partially negates the possibility of balancing the load on multiple available links/paths [10]. Finally, results in [11] show that sleep mode is achievable just for very few nodes and only at very low loads. Conversely, in our model, energy-aware architectures allow the NEs power consumption to scale with traffic load, as in [10–14]; such architectures are strongly advocated by current efforts from standardization bodies and governmental programs [15] and can be made up using off-the-shelf standard technologies [16,17].

2. Related work

”Greening the network” is an active subject of recent research. Several papers have concentrated on the reduction of power consumption. In [13] Gupta and Singh were among the first researchers to envision the idea of energy conservation in Internet-based infrastructures. Shen and Tucker [4] developed mixed integer linear programming (MILP) methods and heuristics to optimize the energy consumption of a IP over WDM transport network. In detail, their objective was minimizing power consumption of the network by switching off router ports, transponders, and optical amplifiers; they proposed two heuristics (“direct bypass” and “multi-hop bypass”). Another approach focusing on a MILP-based formalization of the power-aware routing and wavelength assignment has been also presented by Wu et al. [18]. In their work, energy savings can be achieved by switching off optical cross connects (OXC) and optical amplifiers according to three different algorithms and criteria. In [19,10], ILP mathematical formulations are presented with the double objective of reducing both the energy consumption and the GHG emissions of network infrastructure. Since the ILP solves at the optimum the offline RWA problem, these works give an upper bound for energy and GHG savings. However, using ILP for real-world networks with dynamic traffic is unpractical due to its intractable computational complexity. Energy saving by dynamically switching off idle IP router line cards in low-demand hours was also the approach presented by Idzikowski et al. [5]. They analyzed the effects of reconfiguring routing, at the different layers, by assuming complete wavelength conversion capability in each node. In [20], Chiaraviglio et al. have proposed and evaluated some greedy heuristics based on the ranking of nodes and links with respect to the amount of traffic that they would carry in the context of an energy-agnostic configuration. Silvestri et al. [21] combined traffic grooming and transmission optimization techniques to limit energy consumption in the WDM layer. Traffic grooming shifts traffic from some links to other ones in order to switch empty ones off, and transmission optimization adjusts dispersion management and pulse duration, which decreases the need for using in-line 3R regenerators. The power savings that can be achieved by dynamically adapting the network topology to the traffic volume are investigated in [22], where a linear programming approach is proposed that is able to identify optimal topologies for given traffic loads and generic network topology. In [23], various power-efficient grooming strategies, combined with lightpath extension and lightpath dropping, are evaluated in WDM networks where nodes have the tap-or-pass capability.

Most of these approaches are characterized by a limited dynamism, and hence are not easily applicable in a fully adaptive online scenario or use power containment techniques based on switching off of inactive elements. In our fully dynamic on-line approach, no switching off is assumed to be feasible (as explained in the previous section) and so, in this, it is completely different and not directly comparable, in term of both performance and effectiveness, with all the previous ones.

3. Backgrounds

3.1. Wavelength routed networks

A wavelength-routed network, sometime also referred to as an optical circuit switched (OCS) network, is basically composed of several OXC devices and opto-electronic edge routers connected by a set of fiber links. The WDM technology is used to carve up the huge bandwidth available on the optical fibers into lower-capacity wavelengths (optical channels), which may be independently used to carry information across the same physical links. Circuit switched connections, usually with high bandwidth and QoS-on-demand, are typically implemented by dynamically creating and tearing down multi-hop optical channels between client sub-networks according to a specific RWA strategy. The above connections, called lightpaths, “transparently” traverse the fiber network without being converted into an electrical signal. In some cases, they may pass through an optical/electrical/optical (O/E/O) conversion for regeneration, wavelength conversion or add/drop purposes. At the state of the art, there is still a large gap between the available capacity of an optical channel and the much lower bandwidth requirements of a typical connection, but, on the other side, the number of wavelength channels (lambdas) available in most of the networks of practical size is much lower than the number of source–destination connections that need be made. Hence, traffic grooming capability is required on opto-electronic routers operating on the network edge. Accordingly, all the connection requests, which share the same traffic flow characteristics and involve significantly lower capacities than those of the underlying wavelength channels, can be efficiently multiplexed or “groomed” onto the same wavelength/lightpath via simultaneous time and space switching. Similarly, different traffic streams can be demultiplexed from a single lambda-path.

3.2. Power requirements in network devices

The fundamental cause of energy consumption in electronic equipment is the effect of loss during the transfer of electric charges, which in turn is caused by imperfect conductors and electrical isolators. Here, the consumption rate depends on the transition frequency and the number of gates involved, together with fabrication features (such as architecture, degree of parallelism, and operating voltage). In pure transparent optical equipment, the main energy-hungry devices are the lasers, since the optical signal has to reach the other end of the fiber with sufficient "quality" in spite of the signal attenuation, dispersion and non-linear optical phenomena. Besides, the power consumption is also conditioned by sophisticated electronic devices for coping with the technological complexity of the photonic environment. For example, when the involved fiber strands need to cover long distances, several intermediate electrical signal re-generators (3R) or optical amplifiers (OA) are necessary (typically, an OA is needed every 80–100 km and a 3R every 500–1000 km) to ensure that the signal power and quality will be sufficient to reach the other end of the fiber with acceptable optical signal to noise ratio (OSNR). Such OA and 3R have a not negligible energy cost that has to be taken into account when setting up the lightpath requests.

3.3. Energy-aware RWA

Introducing energy-awareness in RWA is based on the concept of placing network traffic over a specific set of paths (sequences of nodes and communication links) so that the overall network power demand and/or GHG emissions are minimized, while end-to-end connection requirements are still satisfied. Typical infrastructures are densely meshed, with many redundant interconnections among nodes, so that many available paths can provide multiple reachability options between geographically distant sites. On such a mesh, wavelength routing is used to set-up a logical topology, which is then used at the IP layer for routing. Every time a lightpath is established between any two nodes, the traffic of the lightpath will be handled as a single IP hop by creating the abstraction of a "virtual" network topology on top of the physical one. This "overlay" approach is based on the full separation of the routing functions at each layer, i.e. the connection routing/grooming at the IP layer is independent from the routing of wavelengths at the optical layer. One of the key features of the above model is rearrangeability, i.e. the ability to dynamically optimize the network as a consequence of the independence between the virtual and the physical topology. The above architectural flexibility in building logical topologies, together with physical connection redundancy and over-provisioning, provide fertile grounds for saving energy, since a large number of available traffic routing and device management options can be exploited to optimize energy and carbon footprints network wide. Hence, energy-aware logical network topologies, explicitly conceived to decrease power consumption in the operational phase, can be dynamically built by minimizing the number of energy-hungry devices traversed by the existing lightpaths. In doing this, it is desirable to find a good balance between the competing needs to avoid as many electrically powered hops as possible (to reduce the power consumption at intermediate switching nodes, optical amplifiers and regenerators) and avoid data transmission over excessively long stretches, since moving data is quite energy-expensive. Energy consumption can be drastically reduced by maximizing the reuse of low-power transmission links and highly connected devices, especially when powered by green sources, instead of obliviously spreading the traffic on the available routing/switching devices and communication resources. In other words, since a logical network topology is described by its constituent lightpaths, a logical topology that minimizes the overall energy requirement and the associated carbon footprint is one in which the choice of each individual lightpath, while satisfying the traditional RWA objectives and constraints, is driven by the above energy-efficiency optimization criteria.

In order to support all the above behaviors, energy-related information associated with devices, interfaces and links need to be introduced as additional constraints (together with delay, bandwidth, physical impairments, etc.) in the formulations of dynamic RWA algorithms. Such information must also be handled as new status features in each network element that have to be considered in all the routing and traffic engineering decisions, and conveyed to all the various network devices within the same energy-management domain. This clearly requires modifications to the current routing protocols by properly extending them to include energy-related information in their information exchange messages, such as the power demand associated with a specific end-to-end circuit or the type of energy source currently used by a network element [6]. Analogously, the same information has to be handled by control plane signaling protocols used for the reservation and establishment of paths minimizing the use of dirty power sources, as well as the overall energy consumption, across the network.

4. The energy-aware network model

Defining a sustainable and effective network model taking into account power consumption as well as energy source considerations is the essential prerequisite for introducing energy-awareness within the wavelength routing context. A broad variety of NEs contribute to power adsorption in a network: regenerators, amplifiers, opto-electronic and totally optical routers and switches. Each of these devices draws power in a specific way, which depends on their internal components and structures, on the traffic load and on the relationship between the devices. In addition, some nodes may be powered by green energy sources, while others may use traditional dirty energy plants; therefore, a differentiation between energy sources is required. NEs powered by green energy sources will not contribute to the CO_2 emissions but only to increase the overall network energy consumption. In the dynamic RWA problem, the routing of connection requests is done on a local optimality basis, i.e. considering the current information available at the connection setup time. The potential of such an approach should not be underestimated; the conditions that determined such optimality may change, but the RWA strategies keeps its effectiveness as far as it is able to foresee the future network evolutions, both in terms of resources and energy utilization.

The above energy-related information and concepts associated with devices and links must be abstracted and defined in a formal and concise way into a comprehensive model that needs not to delve into unneeded details, but should only describe the essential aspects needed to drive in an energy-conscious way the RWA algorithms and strategies developed upon it. Therefore, we modeled the network from a high-level perspective in an attempt to keep the reference scenario as general as possible focusing on the effectiveness and energy-efficiency of our approach; the issues raised by modulation techniques, spectrum-sliced elastic networks, and other technological breakthroughs, although interesting, fall outside the scope of this paper, which is to provide an energy-aware dynamic RWA schema to route as many connections as possible.

In detail, at the basis of our model we consider a multigraph $G = (V,E)$ representing the network (Fig. 1), where V is the set of nodes and E the set of edges, $|V| = n$, $|E| = m$. No specific assumption is made on the number of wavelengths per fiber link and on the number of fibers on each link: any two nodes u, $v \in V$ may be connected by several edges (thus, *multigraph*). Each fiber link $(u,v) \in E$ is characterized by its physical length $l_{u,v}$, together with the number of available wavelengths $w_{u,v}$. There can be more than one fiber connecting the same pair of nodes and, for simplicity sake, we assume that all the fibers are of the same type (e.g. NZ-DSF ITU-T G.655/656), requiring an intermediate amplification or regeneration stage every Λ units of distance. Typically Λ can assume the values $\Lambda_{OA} = 80$ km for native optical amplification systems and $\Lambda_{3R} = 500$–1000 km for *3R* electric regeneration devices. On each fiber link (u,v) there can be multiple wavelength links $(u,v)_\lambda$, modeled on the graph G as an additional "virtual" tagged links, where the tag λ can be any of the wavelengths available on the physical circuit. Each tagged link $(u,v)_\lambda$, is characterized by its static global capacity $a_{(u,v)_\lambda}$ and dynamic residual capacity $r_{(u,v)_\lambda}$. Clearly, for each link $(u,v)_\lambda$, its current load is given by $a_{(u,v)_\lambda} - r_{(u,v)_\lambda}$. Provided that a single established lightpath or a chain of lightpaths between the source and destination nodes has sufficient available capacity, each connection request can be routed onto that lightpath or chain. Also, a new lightpath may be dynamically established, as the result of grooming decisions.

The nodes of the graph model the routing and switching devices deployed in the network. We consider two types of nodes: LERs (Lambda Edge Routers) and LSRs (Lambda Switching Routers). LER nodes have both the electronic and optical interfaces, and have the capability to insert/extract traditional electronic traffic into/from the network. LSR nodes are OXC or reconfigurable optical add and drop multiplexers (ROADMs) that are capable of switching the traffic at wavelength level (since we model optical circuit switched networks) and may be equipped or not with wavelength converters. Whenever an optical signal is converted into the electronic domain, it is implicitly assumed that it is possible to apply 3R regeneration as well as wavelength conversion and add/drop at sub-wavelength granularity (grooming). Consequently, network traffic may be of two types: electronic time division-multiplexed (TDM) traffic (i.e. traffic that undergoes electronic processing) and pure optical traffic (i.e. WDM traffic entirely managed in the optical domain) with or without optical wavelength conversion. Electronic routers have the ability to add/drop traffic into/from the network, to make electronic WC (Wavelength Conversion) and to regenerate the signal in the electronic domain (3R regeneration). Optical routers support optical traffic with or without all-optical WC. That is, they may deflect wavelengths through an electronic 3R

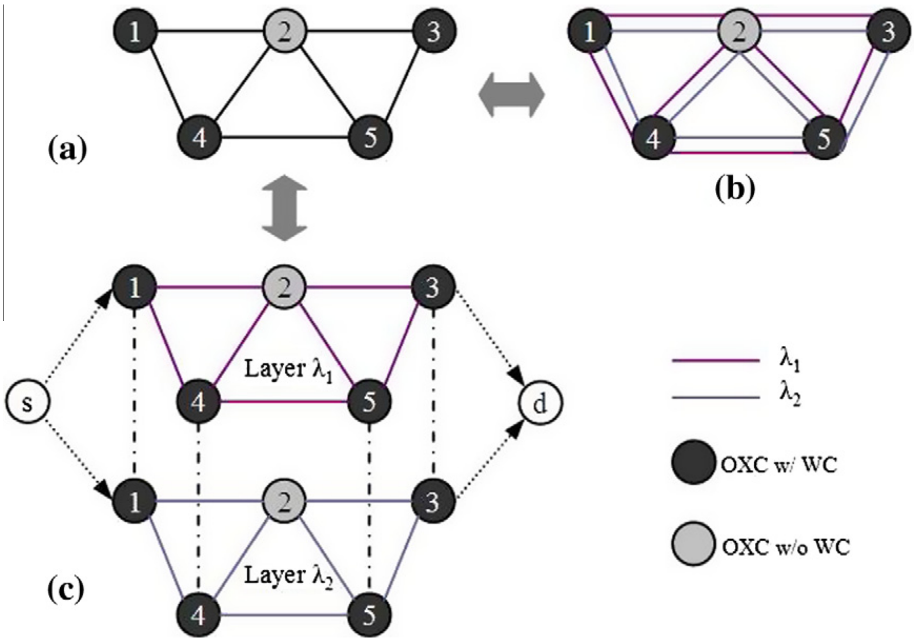


Fig. 1. A network topology with two wavelengths per link (a) and its representations as multigraph (b) and as layered graph (assuming a connection request from node 1 to node 3) (c).

regenerator if the OSNR is too degraded and then switch the wavelength through the corresponding output port [2]. In our network model, connections are bidirectional and unsplittable, i.e. a traffic demand is routed over a single lightpath, and LER nodes can be source or destination of a connection.

As for the energy, we derived a properly crafted per-node, per-link and per-lightpath energy model and power cost function, basing our estimation on the literature [1–3,14,15] and on the manufacturers technical sheets [24,25], with the aim of fitting with the future energy-aware technologies that will adapt their power-consumption with their load [10–14].

We distinguish between green and dirty energy sources, i.e. carbon-emitting and zero-carbon plants. Each node $n \in V$ has a statically associated attribute s_n representing the type of energy source (green or dirty) powering the corresponding device, and we assume that green and dirty energy sources are heterogeneously distributed in the network. This attribute has been kept static to avoid unwanted fluctuation in the "green-biased" node selection function due to the temporary unavailability of the specific source (e.g. sun, wind or tide), since the hosting sites are usually equipped with battery systems, ensuring the availability of the accumulated green energy also during the source off-times (e.g. the night hours for solar panels). Anyway, the devices powered by green energy should be always preferred also for cost containment reasons, since the energy costs in the hosting sites/installations are regulated by significantly advantageous contractual conditions. In fact most of the electricity providers and supplying utilities apply some balancing policies for sites producing their own energy from renewable source and placing the energy produced in excess in the "public" electric grid (by reducing the carbon footprint on a more global scale), so that, even in the case in which the required energy would be currently extracted from a dirty source, its cost will be significantly lower when compared with other dirty-only powered sites.

4.1. Per-node power requirements

In order to characterize in a realistic and quantifiable way the energy requirements of a specific network path (needed to accomplish our optimization goals within the RWA context), we need to estimate the power consumption of all the traversed NE (devices on the nodes and transmission links) as a function of the involved traffic type. In doing this, we essentially consider two main traffic types:

1. Electronic traffic, comprising add/drop, electronic WC, 3R regeneration.
2. Optical traffic, with or without optical WC.

The above traffic types are characterized by a considerably different power consumption when traversing a NE: electronic traffic requires more power than optical traffic with WC and the latter consumes more power than optical traffic without WC, due to the different devices involved [1,2].

Therefore, the power consumption of a specific lightpath depends on:

1. The type of devices traversed along its route from source to destination node, e.g. router, switch, signal amplifier/regenerator, etc.
2. The "device class", in terms of hardware architecture and aggregated switching performance of the network element itself. More precisely, modular switching nodes capable to handle higher throughputs consume less energy per bit that smaller ones [26,27] since they are more optimized and tend to be located in the center of the network where the traffic is more aggregated, and "opaque" nodes equipped with electronic switching matrices are more energy-hungry that their transparent photonic counterparts.
3. The type of traffic that it transports through each network element, i.e. electronic, optical with WC and optical without WC.

The power consumption of real electronic and optical switching nodes with and without WC are reported in [1,2], where it can be observed that the electronic traffic grows quickly with respect to the optical one and that, within the optical traffic context, the WC is the main factor internal to the switching device requiring a not negligible quantity of energy. In [14] it is shown that the base system of an idle network device consumes approximately half of the total power drained by the device, while the other half is consumed when the router is in its maximum configuration, i.e. maximum number of line cards/modules installed and operating at their full load. These power consumptions refer to commercially available devices whose architectures are not energy-aware: their power consumptions only slightly depend (2–3%) on the current traffic load, but strongly depend on the number of line cards installed [14,28]. Next-generation energy-aware routing/switching nodes, designed with energy-efficiency in mind and allowing dynamical adjustment of their power consumption with the variation of the traffic load by selectively putting into sleep or low-power mode some interfaces, line cards, and subsystems, will be characterized by a significantly dominating load-dependent energy consumption component. However, by estimating the power demands used in our model from the available quantitative data gathered from the current devices implicitly forces the model to operate in a worst-case situation making the achieved results more comforting (since they will be greatly improved with the introduction of next generation energy-aware devices). Therefore, even if actual router architectures are not energy-aware, in the sense that they consume the same amount of power regardless of the traffic load, here we consider future energy-aware architectures that scale their power consumption with their current traffic load, thus giving rise to optimization [1,10,27].

Consequently to [14,28], we assume that the power consumption of a network element, modeled by a node or link, can be divided into two equal parts: "fixed" and "variable" power absorption. The fixed power consumption is always present and is needed just for the device to stay "on", while the variable power consumption depends on the actual traffic load that the device is currently supporting. The case in which the fixed power consumption is significantly greater or lower than the variable part would affect more or less proportionally the optimization gains. In particular, for the current energy-unaware devices, the fixed part is much greater than the variable part (which only represents 2–3% of the total power consumption), leaving almost no space for optimization. In the opposite situation, if the fixed power consumption was much lower than the variable part, the likely outcome would be that the optimization margins will increase a lot; in this sense, our approach of assuming 50% for fixed and 50% for variable power consumption can be considered as conservative.

The previous considerations can be used to build a sufficiently general per-node power consumption model. Starting from the power consumptions of the network routing devices (in Watts) as function of their aggregated bandwidth (in Gbps) [1,2], we obtained the linear power consumption equations [10] reported in Table 1, which we used to calculate the real *maximum* power consumption of any kind of network node given its aggregated bandwidth. In such a linear model, a slope of m means that for each unit of traffic (Gbps) the router consumes m units of power (W). For example, an electronic router with an aggregated bandwidth of 10 Tbps is characterized by a maximum power absorption of 30 kW. An optical switch with the same aggregated bandwidth consumes 0.62 kW with WC and 0.2 kW without WC, which totally agree with the values reported in [2,14].

Starting from such maximum power consumption values, we obtain the curves in Fig. 2, in which the minimum power consumption associated with the network device n in the idle state is given only by the fixed power consumption ϕ_n of its base. The maximum power consumption $2\phi_n$ is achieved when the node is fully loaded, i.e. when the current load x is equal to the maximum aggregated bandwidth B_n of the node n. How the power consumption scales between these two values has been studied carefully in [10]. In this work we observed that the power consumption associated with electronic traffic is higher than the one associated with optical traffic (Fig. 3a – optical node power consumption not in scale; see the peak power consumption of optical nodes in Fig. 3b for in-scale values). Furthermore, we also observed that the power consumption of smaller nodes follows a worse trend with respect to bigger ones, in which the per bit energy consumption is lower (Fig. 3b). Therefore, we can delineate two extremes: big nodes transporting optical traffic as the least power consumers, and small nodes transporting electronic traffic as the most power hungry devices. To this end, we studied different power consumption functions, both present in literature and analytically conceived, to describe and carefully balance the power consumption of all the possible combinations between these two extremes.

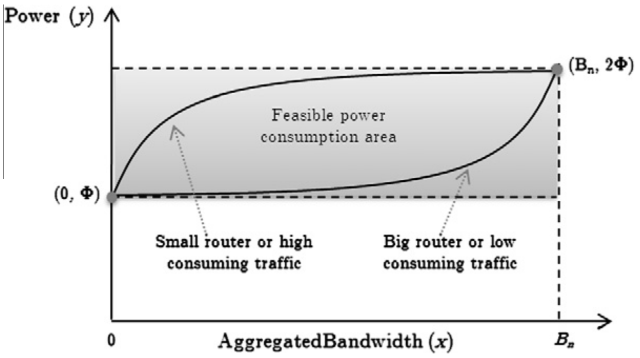


Fig. 2. Minimum and maximum power consumption of a network node.

In more formal terms, we define for all the routing and switching nodes, a power consumption function $\Psi_n(x)$ expressing the power requirements of a node n characterized by device-specific static consumption ϕ_n and performance class (aggregated bandwidth) B_n, variably conditioned by a traversing traffic load x. The function $\Psi_n(x)$ can be viewed as a linear combination of the logarithmic function $\theta_n(x)$ and the line function $\vartheta_n(x)$ weighted by the parameter $\alpha_n(x)$:

$$\Psi_n(x) = \alpha_n(x) \cdot \theta_n(x) + (1 - \alpha_n(x)) \cdot \vartheta_n(x) \tag{1}$$

where

$$\begin{aligned}\theta_n(x) &= \underbrace{\left[\phi_n - \ln\left(\frac{e^{\phi_n}}{B_n} \cdot (B_n - x) + \frac{x}{B_n}\right)\right]}_{\text{variable power consumption}} + \underbrace{\phi_n}_{\substack{\text{fixed power}\\ \text{consumption}}} \\ &= 2\phi_n - \ln\left(\frac{e^{\phi_n}}{B_n} \cdot (B_n - x) + \frac{x}{B_n}\right),\end{aligned} \tag{2}$$

is the equation of the logarithmic function passing through the points $(0, \phi)$ and $(B_n, 2\phi)$, modeling the best per-bit energy consumption (i.e. optical traffic w/o WC in Fig. 3a) and:

Table 1
Power consumption (in Watts) dependency laws on aggregated bandwidth and load (in Gbps) for different types of nodes (linear case).

Node type	Power consumption (y) as function of the aggregated bandwidth (x)	Power consumption (y) as function of the load (x) assuming half fixed (ϕ) and half variable ($m \cdot x$)
Electronic	$y = 3x$	$y = 1.5x + \phi$
Optical w/ WC	$y = 0.062x$	$y = 0.031x + \phi$
Optical w/o WC	$y = 0.02x$	$y = 0.01x + \phi$

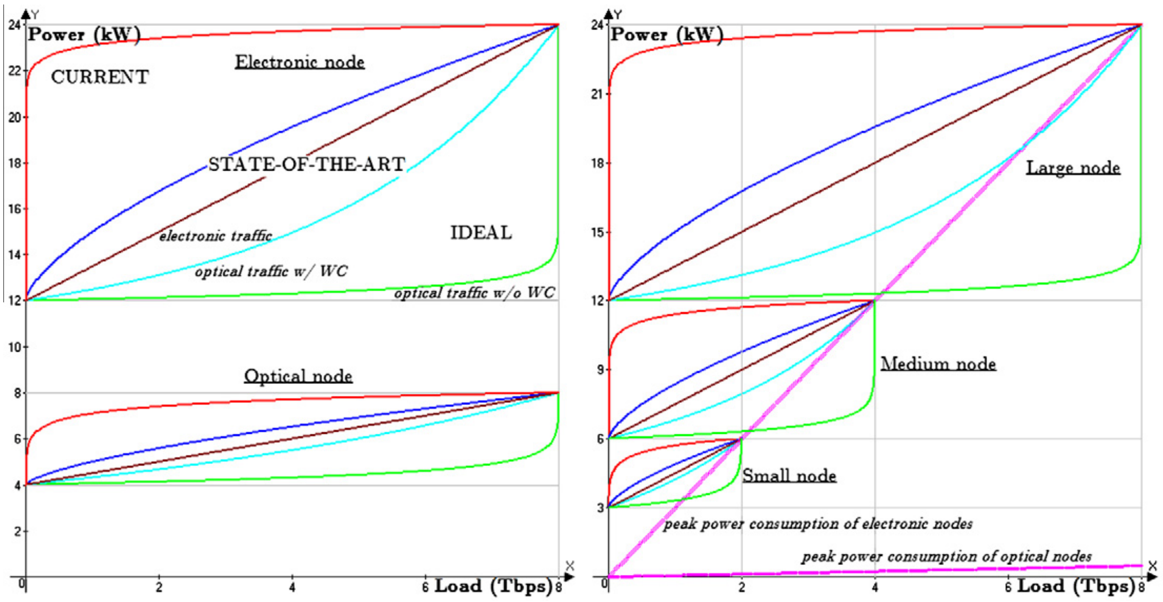


Fig. 3. Power consumption functions of current, ideal and state-of-the-art for electronic and optical nodes (a) and for different sizes of nodes (b).

$$\vartheta_n(x) = \underbrace{\frac{\phi_n}{B_n}x}_{\substack{\text{variable power}\\ \text{consumption}}} + \underbrace{\phi_n}_{\substack{\text{fixed power}\\ \text{consumption}}}, \tag{3}$$

is the equation of the line function passing through the points $(0,\phi)$ and $(B_n,2\phi)$, modeling the worst per-bit energy consumption (i.e. electronic traffic in Fig. 3a).

B_n is the capacity (aggregated bandwidth) of the router n (its performance class), and $\alpha_n(x)$ is the weighting parameter between $\theta_n(x)$ and $\vartheta_n(x)$ depending on the class/performance of the router n and on the parameter $\beta(x)$ which accounts for the specific type of traffic associated with the load x that is actually passing through the node n:

$$\alpha_n(x) = \frac{B_n}{\max\{B_n, \forall n \in \mathrm{V}\}} \cdot \beta(x), \quad 0 \leqslant \alpha \leqslant 1, \tag{4}$$

$$\beta(x) = \begin{cases} 1 & \text{if the traffic } x \text{ is optical w/o WC} \\ 0.323 & \text{if the traffic } x \text{ is optical w/WC} \\ 0.00\bar{6} & \text{if the traffic } x \text{ is electronic} \end{cases}$$
$$0 \leqslant \beta \in \{1, 0.323, 0.00\bar{6}\} \leqslant 1. \tag{5}$$

Note that:

1. The values of $\beta(x)$ have been obtained using the values of real routers from Table 1, taken in such a way to penalize the more power consuming devices and traffic types; e.g. for the electronic traffic, $\beta(x) = \frac{m_{optic\ w/o\ wc}}{m_{electronic}} = \frac{0.01}{1.5} = 0.00\bar{6}$.
2. $\alpha_n(x)$ weights one Eq. (2) or the other Eq. (3) function according to the device class and traffic characteristics of the involved NE.

Table 2
Notation used in the energy model.

Parameter	Energy model
ϕ_n	Fixed power consumption of node n
$\Psi_n(x)$	Overall power consumption (fixed + variable) of node n with traffic load x
$\theta_n(x)$	Logarithmic function
$\vartheta_n(x)$	Line function
$\alpha_n(x)$	Linear combination weighting function
e	Euler's number (base of the natural logarithms)
B_n	Performance class of node n (aggregated bandwidth of all interfaces)
$\beta(x)$	Weighting function on the type of traffic
m	Slope of the power consumption functions (linear case)
$w_{u,v}$	Number of wavelengths/channels crossing fiber (u,v)
$\Psi_{(u,v)_\lambda}(x)$	Power consumption of the link (u,v) on the wavelength λ with traffic load x
$l_{u,v}$	Length of the fiber (u,v) (km)
$\eta^u_{(u,v)_\lambda}(x)$	Power consumption of the interface on the node u associated with the wavelength λ on the fiber (u,v) supporting the traffic load x
$Q_{u,v}$	Power consumption associated with the individual amplification device on link (u,v)
Λ_{OA}	Maximum allowed length of a link without need of optical amplification (km)
Λ_{3R}	Maximum allowed length of a link without need of 3R regeneration (km)
$R(x)$	Power consumption associated with the individual 3R regeneration of the traffic load x (one for each wavelength)
$\Psi_\pi(x)$	Power consumption of lightpath π with traffic load x
l_π	Cumulative length of path π (km)

3. The fixed power consumptions ϕ_n of nodes are obtained from [1,2,14].

For an explanation of the symbols used in the notation refer to Table 2.

4.2. Per-link power requirements

End-to-end transmission links are characterized by a power consumption depending not only on the specific demand associated with the hardware interfaces located in both the endpoints, but also on the impact introduced by the optical amplification and regeneration devices needed by the signal to reach the endpoints with an acceptable quality, and thus, on the length of the traversed fiber strands.

Accordingly, the power absorption of a transmission link realized on the wavelength λ between the nodes u and v can be entirely described by the specific power demand characterizing the involved end-to-end interfaces plus the power required for powering all the possible intermediate regenerators or optical amplifiers, if any. If the fiber between u and v is currently crossed by $w_{u,v}$ wavelengths, we consider that the power requirements due to the intermediate devices will be shared between the $w_{u,v}$ channels simultaneously; typically, as for the OA, the entire frequency band (e.g. C-band) is amplified as a whole, without per wavelength granularity, whilst as for 3R regeneration, per wavelength 3R is required. Thus the power consumption $\Psi_{(u,v)_\lambda}(x)$, associated with a link on the wavelength λ with load x traversing the fiber (u,v), whose length is $l_{u,v}$, can be described as:

$$\Psi_{(u,v)_\lambda}(x) = \eta^u_{(u,v)_\lambda}(x) + \eta^v_{(u,v)_\lambda}(x) + \left\lfloor \frac{l_{u,v}}{\Lambda_{OA}} \right\rfloor \cdot \frac{Q_{u,v}}{w_{u,v}} + \left\lfloor \frac{l_{u,v}}{\Lambda_{3R}} \right\rfloor \cdot R(x), \tag{6}$$

where

1. $\eta^u_{(u,v)_\lambda}(x)$ is the power consumption of the interface on the node u associated with the wavelength λ on the fiber (u,v) when operating at the minimum speed allowing to support a traffic load x without any loss or delay increase. Here we do not take into account the variable effect on power consumption of the dynamic traffic traversing the interface, whose impact is already considered in the per-node consumption, and only model the specific per-interface power demand according to its specific static hardware features (type of laser, its power, etc.) and to a multiple threshold scale characterizing its consumption depending on current operating speed (implicitly dependent from the load x). The above $\eta^u_{(u,v)_\lambda}(x)$ function can be modeled as in [27].
2. $Q_{u,v}$ is the power consumption associated with the individual amplification device on link (u,v) (between 3 and 15 W) operating throughout the fiber link (one for the entire frequency band); for simplicity, we assume that all the amplification devices operating on the same link have the same power consumption.
3. $R(x)$, defined in the same way as $\vartheta_n(x)$ (i.e. electronic traffic in a node), is the power consumption associated with the individual 3R regeneration device of the traffic load x (one for each wavelength), when present; for simplicity, we assume that all the regeneration devices operating on the same link have the same power consumption.

4.3. Per-lightpath power requirements

Given a lightpath π as a sequence of nodes and tagged links $(u,v)_\lambda$ with a traffic demand x, the power consumption $\Psi_\pi(x)$ of π is given by the sum of the individual power absorption of all the traversed nodes and links plus the required regenerations:

$$\Psi_\pi(x) = \sum_{n\in\pi} \Psi_n(x) + \sum_{(u,v)_\lambda\in\pi} \Psi_{(u,v)_\lambda}(x) + \left\lfloor \frac{l_\pi}{\Lambda_{3R}} \right\rfloor \cdot R(x), \tag{7}$$

where l_π is the cumulative path length. The third component in the Eq. (7) sum is needed to cope with fully transparent lightpaths whose length exceeds the maximum length Λ_{3R} that a signal can travel without need of 3R regeneration. In this case, since all the intermediate devices do not convert the signal back and forth into electrical/optical form and only introduce impairments, a 3R regeneration stage is required in correspondence with at least $\left\lfloor \frac{l_\pi}{\Lambda_{3R}} \right\rfloor$ intermediate nodes. If the lightpath π is fully transparent (without wavelength conversion), the tag λ is the same on all the wavelength links.

5. The two-stage RWA scheme

The proposed energy-aware RWA scheme operates online, running at each request of a dedicated connection with specific service requirements (typically QoS on the bandwidth capacity) between two network nodes. In such a dynamic scenario, connection requests have to be served as soon as possible when they arrive; thus, we designed GreenSpark with simplicity in mind, which was considered as a necessary requisite when developing the dynamic RWA scheme to keep as low as possible the computational complexity. According to the typical assumptions in OCS networks, each connection is considered to be bidirectional and consists of a specific set of traffic flows that cannot be split between multiple paths. A connection can be routed on one or more (possibly chained) existing lightpaths between the source and the destination nodes with sufficient available capacity or on a new lightpath dynamically built on the network upon the existing optical links. Connection routing and grooming decisions are taken instantaneously reflecting an highly adaptive strategy that dynamically tries to fulfill the network resource utilization and connection serviceability objectives together with minimizing the overall power consumption by privileging cheaper (in terms of power demands) chains of nodes/links and, between them, trying to maximize the usage of devices powered by green energy sources.

Without loss of generality, we route connection requests with only a constraint on the required bandwidth, and rely on the incorporation of other policies within the bandwidth-routing framework to perform routing based on several QoS and impairment metrics such as limited latency, error rate, hop-count, delay, and losses. Such constraints can be incorporated into SLAs by converting these requirements into a bandwidth requirement as shown in [29]. Impairments are accounted for by modeling 3R regeneration into the framework, supported by the study in [30] in which impairment-awareness is included into the regeneration placement for WDM networks, and optoelectronic signal regeneration is employed to address the signal quality of lightpaths that are found to be impaired without compromising the signal quality of any of the lightpaths.

The apparently conflicting goals of minimizing cost and length of designed paths while keeping the network resource usage fairly balanced, and optimizing the overall power consumption by reusing, as possible, energy-efficient paths across the network, give origin to a multivariate and multi-objective optimization problem that can be solved according to a *divide-et-impera* strategy, set up of two-stages in which each stage separately handles a specific objective by using properly crafted heuristics. Specifically, in the first stage (pre-selection phase), the goal is to determine an ordered list (whose length is defined by a parametric value k) of feasible minimum cost paths that fully satisfy the connection demands, trying to leave on each link of these paths sufficient room to satisfy further requests as much as possible. Such strategy clearly implies balancing the load on all the available network resources. In the second stage (energy-aware decision phase), the proposed scheme analyzes, for each path found in the stage one, its power requirement as given by the aforementioned energy model, and then selects the best available solution according to several heuristic criteria based on finding a good compromise between the traditional carrier's objectives and the new green requirements (i.e. limiting power consumption or using green energy sources to reduce GHG emissions).

Towards this goal, we explicitly defined and studied two different green optimization objectives: the first one, aiming at reducing the power consumption throughout the network, and hence its operating expenditures; the second one, oriented to minimize the network carbon footprint on the environment by minimizing the GHG emissions.

5.1. Prerequisite control plane facilities

The proposed scheme also requires several forms of cooperation between the nodes concurring to the RWA problem solution. This implies that every node needs to run distributed control-plane services (such as those provided by the GMPLS framework) keeping up-to-date information about the complete network topology, resource usage and power demand attributes, as well as taking care of resource reservation, allocation, and release.

More precisely, a periodic link-state advertisement (LSA) scheme must convey all the link and node state information (including energy related ones) to every node in the network, ensuring the complete synchronization between all the nodes' network status views. Since the amount of per-link state information is very small, any appropriate enhanced link state scheme like those employed by OSPF can be adequate for this purpose, like the one developed in [6].

The Dijkstra-based path selection scheme of stage one should meet certain conditions:

1. A link may not reserve more traffic than it has capacity for.
2. Shorter paths should be preferred when they consume fewer network and energy resources.
3. Critical resources, e.g. residual bandwidth in bottleneck links, should be preserved for future demands.

The last two conditions reflect that what we really seek is to keep the connection blocking probability (or, in other words, the rejection ratio) as low as possible, or equivalently to increase as much as possible the network utilization.

In addition, an extended signaling/reservation protocol, such as RSVP-TE, can be used to setup and release paths and lightpaths and handle all the bandwidth, fiber or wavelength resources reservation and allocation/deallocation operations required during such activities. In detail, as a new request arrives, the control plane on each node, starting from the originating one, runs our source-based localized RWA algorithm, calculates the new overlay network topology and triggers the proper path setup actions by sending a reservation request toward the destination and provisionally reserving bandwidth resources. The RWA scheme, operating according to a two-layer model (i.e. an underlying pure optical wavelength routed network core and an opto-electronic time division multiplexed layer built over it) should determine if the request can be routed on one of the already available lightpaths, by time-division multiplexing it together with other already established connections, or a new lightpath is needed on the optical transport core to join the terminating (edge) nodes. In presence of multiple options between new feasible and already established lightpaths, the link weighting and path selection functions of the two stages, applied on the existing lightpaths and to the wavelength links that can be used to set up new lightpaths, together with the energy costs, dynamically determine the best compromise (between network and energy costs) routing solution for the request, starting from the current network status. For example, if two lightpaths between source and destination exist, both with sufficient available capacity, if the difference in network cost between them falls below a specific acceptability threshold, the tie is resolved in favor of the greener lightpath. Such policy guarantees maximum lightpath utilization and automatically achieves, as long as possible, effective dynamic grooming and power usage, assuming that the topology (link state) database is properly updated.

The signaling scheme for triggering the new lightpath set-up and reserving the required bandwidth, fiber or wavelength resources along the path is very similar to the RSVP-TE protocol used by GMPLS. To make a reservation request, the source node needs the path and the bandwidth that it is trying to reserve. The request is sent by the source along with path information. At every hop, the node determines if adequate bandwidth is available in the onward link. If the available bandwidth is inadequate, the node rejects the requests and sends a response back to the source. If the bandwidth is available, it is provisionally reserved, and the request packet is forwarded onto the next hop in the path. If the request packet successfully reaches the destination, the destination acknowledges it by sending a reservation packet back along the same path. As each node in the path sees the reservation packet, it confirms the provisional reservation of bandwidth. In addition, it also performs the required configuration needed to support the incoming traffic such as setting up labels in a GMPLS label switching node, or reconfiguring the lambda switching internal devices (such as MEMS) in a transparent optical wavelength switching system.

5.2. The first stage: selecting the candidate paths

The first stage of the GreenSpark RWA schema computes a list of k feasible cycle-free paths, in increasing order of cost, between the source and the destination nodes of the connection to be routed, constrained by its QoS requirements. Here k is a configurable parameter that can be used to limit the number of feasible paths that should be considered in the following step, thus controlling the depth and granularity of the analysis process according to a performance/precision compromise. The K-SPF (k-shortest paths first) algorithm used has been explicitly modified to meet the specified bandwidth requirements of each new request and to enforce the wavelength continuity constraint so that, when traversing converter nodes, we are totally free in selecting any outgoing link of the multigraph (i.e. any wavelength), whereas with all the other nodes we can only select an outgoing link corresponding to the same wavelength associated with the incoming one. The above pre-selection process is driven by a link weighting function $\omega((u,v)_\lambda)$ taking into account, for each link $(u,v)_\lambda$, the (static) global capacity $a_{(u,v)_\lambda}$ and the current (dynamic) residual capacity $r_{(u,v)_\lambda}$ still available on the link. Intuitively, a *good* weighting function should be inversely proportional to both the residual and the maximum capacities, but the contribution of these two factors need not be the same. Following the analysis in [8], the link weighting function is defined as:

$$\omega : E \rightarrow \Re, \omega((u,v)_\lambda) = (r_{(u,v)_\lambda} \cdot \mathrm{loga}_{(u,v)\lambda)^{-1}} \tag{8}$$

Such a function exhibits the desirable property of leading to a good load-balancing over the network, since it tries to avoid bottleneck link by assigning higher costs to smaller (low global capacity $a_{(u,v)_\lambda}$) and more congested (low residual capacity $r_{(u,v)_\lambda}$) links. Note that every two links with the same residual/maximum capacity ratio but different residual or maximum capacity values will have different associated weights. This avoids assigning the same weight to two links with the same saturation ratio but with different residual or global capacity. Therefore, we choose the weighting function (8), which satisfies all the desired properties discussed before. Note also that the first stage is exclusively based on load-balancing criteria, and no energy consideration is present at all; this guarantees that the k paths selected in this stage are the best balanced ones, thus giving priority to the traditional network optimization criteria of minimizing the connections blocking ratio. Energy-awareness is introduced only in the second stage, where the greenest path among the k best-balanced candidate paths is finally selected.

5.3. Second stage: choosing the best path

The k minimum cost paths found by the K-SPF algorithm in the first stage are the k best paths as for network's blocking probability (the percentage of rejected connection requests), since the weighting function $\omega((u,v)_\lambda)$ tends to balance as much as possible the use of the network resources. Among these k best-balanced paths, we now have to choose the *optimal* path among them according to our energy-aware selection criteria, aiming at minimizing the power consumption or the carbon footprint. For this purpose we need to introduce a properly crafted heuristic working as a path scoring function, to differentiate among the available preselected paths and choose the most energy-efficient one. The scoring function f_S is defined on the set of all the possible paths Π and will evaluate the power consumption and carbon footprint of the k paths $K = \{\pi_i, i = 1,2,\ldots,k\}$ obtained from the first step:

$$f_S(\pi) : \Pi \rightarrow \Re. \tag{9}$$

The (total) power consumption $\Psi_\pi(x)$ of a path π defined in eq. (7) can be decomposed as the sum of the power consumption of the traversed devices that are powered by green $\Psi^G_\pi(x)$ and dirty $\Psi^D_\pi(x)$ energy sources:

$$\Psi_\pi(x) = \Psi^G_\pi(x) + \Psi^D_\pi(x). \tag{10}$$

Note that the carbon footprint of a lightpath is only given by the power consumption of the involved NEs that are powered by dirty energy sources, as the NEs powered by green energy sources do not contribute to GHG emissions.

Therefore, if our primary objective is to minimize the GHG emissions (GreenSpark MinGas), we have to choose the path π which has the lowest carbon footprint $\Psi^D_\pi(x)$ (primary objective) and, among paths with the same minimum carbon footprint (if any), we choose the path that minimizes the total power consumption $\Psi_\pi(x)$ (secondary objective):

$$f_S(\pi) = \Psi^D_\pi(x) + \log \Psi_\pi(x). \tag{11}$$

Analogously, if our main goal is reducing the overall power consumption and, thus, the network operating energy costs (GreenSpark MinPower), we need to use an objective function privileging the paths with minimal total power consumption $\Psi_\pi(x)$ and, among them, choosing the one with the minimum carbon footprint $\Psi^D_\pi(x)$:

$$f_S(\pi) = \Psi_\pi(x) + \log \Psi^D_\pi(x). \tag{12}$$

The computation of the scoring function is done for each of the k minimum cost paths, and the path π^* eventually chosen is the one with the lowest $f_S(\pi)$ value:

$$\pi^* = \arg\min\{f_S(\pi) | \pi \in K\}. \tag{13}$$

If more than one such lightpaths exist (i.e. with the lowest f_S (π) value), the one with the minimum i index in the set of lightpaths K is selected (to maximize load-balancing).

The path π^* is the "best" path between the best load-balanced paths that minimizes the carbon footprint or overall power consumption according to the proposed energy model, and it will be used to route the connection request.

Note that the first stage cost function is defined over the set of edges (8), whereas the energy-aware scoring function is defined over paths (9) to reflect our intent of achieving an acceptable compromise between the traditional network optimization objectives, typically based only on specific link properties, and the energy-related ones that need to take into account more complex considerations to be done on the higher layer concepts of lightpath/channel, interface and node role and wavelength processing practice such as optical amplification and 3R regeneration. That is why we structure the decision process into two independent phases and select among the k candidate paths the one with the minimum carbon footprint or power consumption according respectively to the functions (11) and (12), instead of simply selecting *the* minimum cost path based only on the traditional cost function (8).

The generic GreenSpark algorithm is sketched in Fig. 4. The algorithm takes as input the current network state G, the connection request $\rho = (s,d,b)$ between node s and d with QoS bandwidth requirement b, the k parameter of the K-SPF and the objective function f_S. In the first stage (lines 1–2), the K-SPF algorithm finds the k minimum cost paths with sufficient free bandwidth connecting the source and destination nodes. The k paths are the best ones according to the load-balancing cost function $\omega((u,v)_\lambda)$ of Eq. (8). In the second stage (lines 3–7), the chosen objective function f_S is evaluated for the k minimum cost paths and the best path π^* is eventually chosen to route the connection request ρ. Finally, the new $\omega((u,v)_\lambda)$ costs are updated only for the edges of the path π^*, and the chosen path and new network state are returned.

6. Time and space complexity analysis

In the first stage, the computing of the K-SPF for finding the k feasible paths for a specified source–destination pair requires in the worst case a time complexity of $O(k \cdot (m + n \cdot \log n))$ [31]. The second stage computes the objective function f_S for each of the k paths found. The function calculation requires the computation of the power consumption or GHG emissions for each network element in the path. The maximum length of a cycle-free path in a graph with n nodes is $n - 1$, thus the second stage requires $O(k \cdot n)$. Hence, since the K-SPF complexity is the dominating factor between the two stages, the worst case runtime is given by the polynomial time complexity $O(k \cdot (m + n \cdot \log n))$. Therefore, GreenSpark belongs to the same polynomial complexity class of the fastest SPF improved by using a priority queue with a Fibonacci heap in the implementation, $O(m + n \cdot \log n)$ [32]. GreenSpark complexity is also lower than the quadratic complexity of the original SPF algorithm, $O(n^2)$, and significantly lower than the cubic complexity of naïve MIRA $O(n^3 m \cdot \log (n^2/m))$ optimized with the Goldberg max-flow algorithm [33].

As for space complexity, our multigraph network representation requires less space with respect to the layered graph approach conventionally used in dynamic RWA algorithms (Fig. 1). Using up to λ wavelengths on each edge, the layered representation with C converter nodes will require $\lambda n + 2$ nodes (λ layers, each dedicated to an individual wavelength, plus two additional nodes to serve as ingress and egress) and $\lambda m + 2\lambda + C \cdot (\lambda - 1)$ edges (converters can be modeled by cross-layer edges that connect each layer to the λ adjacent layer – a wavelength conversion spanning multiple frequencies will thus entail many such edges in sequence), whilst the equivalent multigraph representation will require only n nodes and λm edges, thus notably reducing the space complexity. Besides, in the layered graph, the ingress and egress nodes as well as the edges

```
Algorithm 1 GreenSpark($G, \rho, k, f_S$)
Input:
$G$: current network state
$\rho = (s, d, b)$: connection request; s, d: source, destination nodes; b: required bandwidth
$k$: parameter of K-SPF
$f_S$: objective function (MinPower / MinGas)

Output:
$G^*$: new network state
$\pi^*$: selected route and wavelength assignment

1: Label network edges with cost function $\omega((u,v)_\lambda)$
2: $K \leftarrow$ K-SPF($G, \rho, k$)
3: for each path $\pi \in K$ do
4:     evaluate $f_S(\pi)$
5: end for
6: $\pi^* = \arg\min \{f_S(\pi) | \pi \in K\}$
7: $G^* \leftarrow$ Update the $\omega((u,v)_\lambda)$ costs of network edges $(u,v)_\lambda$ along the chosen path $\pi^*$
8: return $(\pi^*, G^*)$
```

Fig. 4. The GreenSpark algorithm.

connecting them to the network have to be built each time a new connection arrives, whilst in the multigraph approach this preprocessing phase is not necessary thanks to its compact representation. Note that, even in absence of wavelength conversion, all the layers of the layered graph have to be explored, since the (first) wavelength of the lightpath may be any, which compensates the additional check needed in the multigraph approach to enforce the wavelength continuity constraint. Furthermore, the higher number of nodes and edges required by the layered graph with respect to the multigraph approach increases the time complexity which strictly depends on the *n* and *m* parameters.

The low computational and space complexity required by the GreenSpark algorithm with the multigraph network representation helps lightening the computational burden of path computing elements and serving the connections with lower delay with respect to more complex approaches.

7. Performance evaluation and results analysis

In order to evaluate the effectiveness of the GreenSpark energy-aware RWA framework and its impact on the power consumption and carbon footprint of telecommunication networks, we conducted an extensive simulation study on the network topology modeled as undirected graphs in which each link has a non-negative capacity and a specific power demand depending on both its physical and technological features. All the nodes in the graph are characterized, apart from the traditional network-level capabilities such as wavelength conversion and add-and-drop capability, by their power absorption and type of energy source (i.e. green or dirty), as defined in the energy model of Section 4.

To improve the significance of the obtained results and make them more easily comparable with the other experiences available in literature, we spent a significant effort on the use of realistic data in all our experiments (network topology, traffic demands, costs, and power consumption models). Accordingly, we used in our simulations the well-known network topology Geant2 [34] of Fig. 5 with the bandwidths for the links ranging from OC-1 to OC-768 bandwidth units. Here, traffic demands have been modeled by using different randomly generated or static predefined [35,36] traffic matrices. In the latter case, the traffic volumes have been scaled proportionally to the reported traffic distributions. The energy model has been fed with the realistic power consumption values associated with nodes and links taken from [2,10,37]. Recall that, since no per-node sleep mode is assumed to be possible, the network elements are always powered on and therefore the GreenSpark algorithm bases its decisions exclusively

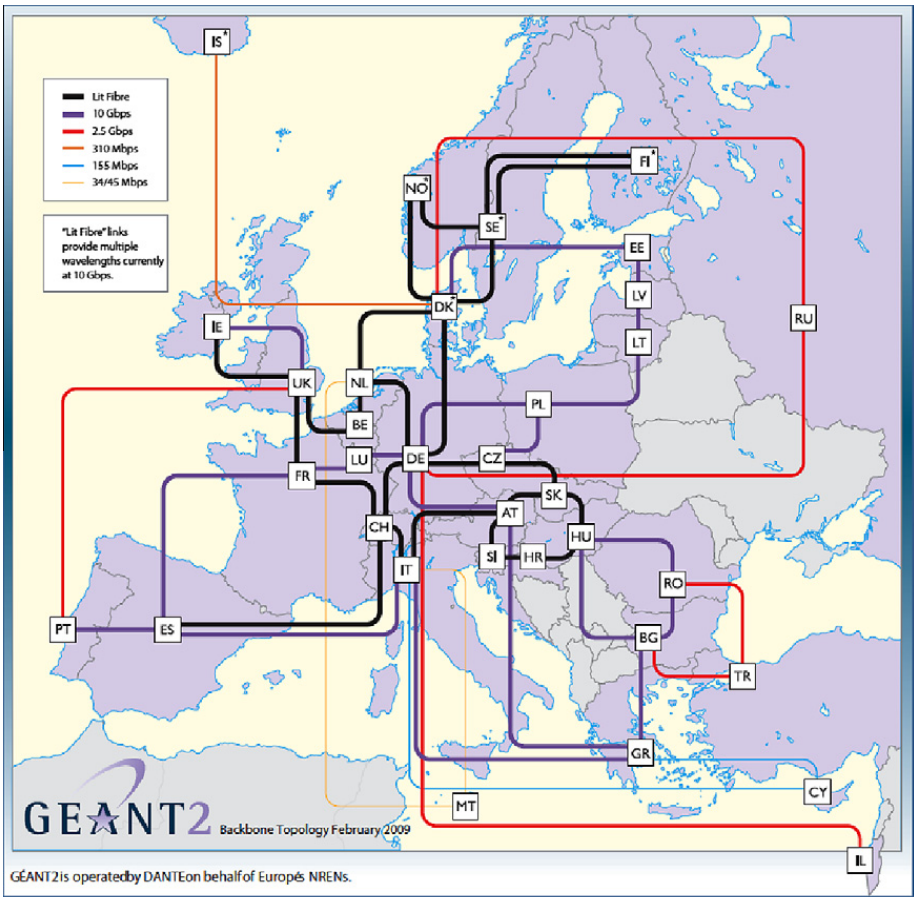

Lit Fibre
10 Gbps
2.5 Gbps
310 Mbps
155 Mbps
34/45 Mbps
"Lit Fibre" links provide multiple wavelengths currently at 10 Gbps.
IS
NO
FI
SE
EE
LV
LT
RU
DK
IE
UK
NL
BE
LU
DE
PL
CZ
FR
SK
AT
CH
HU
SI
HR
IT
RO
PT
ES
BG
TR
GR
CY
MT
IL
GÉANT2 Backbone Topology February 2009
GÉANT2 is operated by DANTE on behalf of Europe's NRENs.

Fig. 5. Geant2: real network topology used in simulations.

on the variable power consumption part, which is the only one that can vary and thus be optimized. Power consumption results refer thus only to the variable power consumption.

Each connection request was characterized by a bandwidth demand ranging from OC-1 to OC-192 units (i.e. from 50 Mbps up to 10 Gbps). As the network load grows, that is, the number of busy connection resources increases more and more with respect to the free/released ones, we continuously monitored the overall network power demand, the percentage of green energy used compared to the maximum available, and the network efficiency expressed by the rejection ratio/blocking factor. All the simulation experience has been conducted in a properly crafted optical network simulation environment [38] that allows the creation of network topologies along with the specification of simulation parameters and configuration files. All the results have been determined with a 95% confidence interval not exceeding 6% of the indicated values, estimated by using the batch means method with at least 25 batches. All the runs have been performed on an Intel® Core™ i7-950 CPU @ 3.07 GHz with 16 GB RAM and 64 bit operating system server running Sun® Java® Runtime Environment v.1.6. In all the experiments, we used a dynamic traffic model in which connection requests, defined by a Poisson process, arrive with a parametric rate of γ requests/s and the session-holding time is exponentially distributed. The connections are distributed on the available network nodes according to the above random-generated or predefined traffic matrices, as summarized in Table 3.

In our lambda-switched optical framework, the resources occupied by the routed connections are counted as the sum of the ratio between the free and the busy bandwidths along the edges. Resources are thus represented as the sum of the bandwidths on all the network edges, while the traffic volume is represented by the quantity of the utilized bandwidth in a certain time.

In all the experiments, GreenSpark has not been compared with other analogous power-containment solutions known in literature because, at the state-of-the-art, almost all the available schemes achieve their savings by powering off interfaces or entire nodes (practice avoided in real network as already mentioned in the introduction), so that the comparison would be misleading since shutting down an entire device would cancel its fixed power consumption which, in our always-on approach, is present all the time. Conversely, the use of provably efficient and publicly available algorithms such as min hop algorithm (MHA) [39] and minimum interference routing algorithm (MIRA) [40], already implemented in several commercial solutions, gives us a real portrait of the power and GHG savings that will be consequent to the introduction of the proposed schema within real world infrastructures, and, at the same time, demonstrates the absence of significant performance burdens in traditional network management objectives (increasing blocking probability, reduced load-balancing, etc.) due to the new energy optimization goals.

The behavior of the algorithm varying the k parameter has been extensively studied (Section 7.2) and, for the considered network topology, an optimal value of $k = 3$ was chosen as the best compromise between the different optimization objectives of the two stages (load-balancing and greenness) and time performance (recall from Section 6 that the complexity depends on k). However, for clearness sake, we first show the results of the comparative simulations with the other RWA algorithms (Section 7.1) and, then, show how the biasing of the k parameter affects the performance (in terms of the two stage objective) and the time complexity of GreenSpark for the given network topology.

7.1. Comparative simulation study

In the first set of simulation shows, we report the comparison of the GreenSpark framework with other well-known RWA algorithms. In these tests, the k parameter of GreenSpark has been tuned to an optimal value ($k = 3$) for the considered network topology, as a result of the extensive simulation study reported in Section 7.2.

In Fig. 6 we plotted the connection blocking probability versus the generated connection requests. We can observe how MHA exhibits the highest blocking probability, essentially due to the congestion of the communication links associated with the shortest paths. MIRA [40] improves the performance of MHA, and achieves lower blocking probability. However, starting from 500 connection requests, its blocking probability grows at quite a fast pace. All the algorithms belonging to the Spark family (Spark, GreenSpark MinPower and GreenSparkMinGas) perform sensibly better than the other ones, and all their versions achieve similar and very satisfactory results in terms of the connection blocking probability.

In Fig. 7 we compared the total power consumption (green and dirty) obtained by the different algorithms versus the connection requests. MIRA reveals to be the highest power consumer, followed by MHA which, in contrast with

Table 3
Parameters used in the simulations.

Simulation parameters	Dante Geant2 network
Number of connections	Varying from 0 to 3000 with different resolutions
Random generated bandwidths	{1, 3, 12, 24, 48, 192} OC-units with different distribution probability
GreenSpark k	1, 3, 5
Spark k, $kHop$	3, 20
Λ_{OA},Λ_{3R}	80 km, 1000 km
Source, destination	Varying according to a Poisson process, duration times exponentially distributed
RWA algorithms	MHA, MIRA, Spark, GreenSpark MinPower, GreenSpark MinGas
Measurements	Blocked connections, power consumptions, green energy percentages

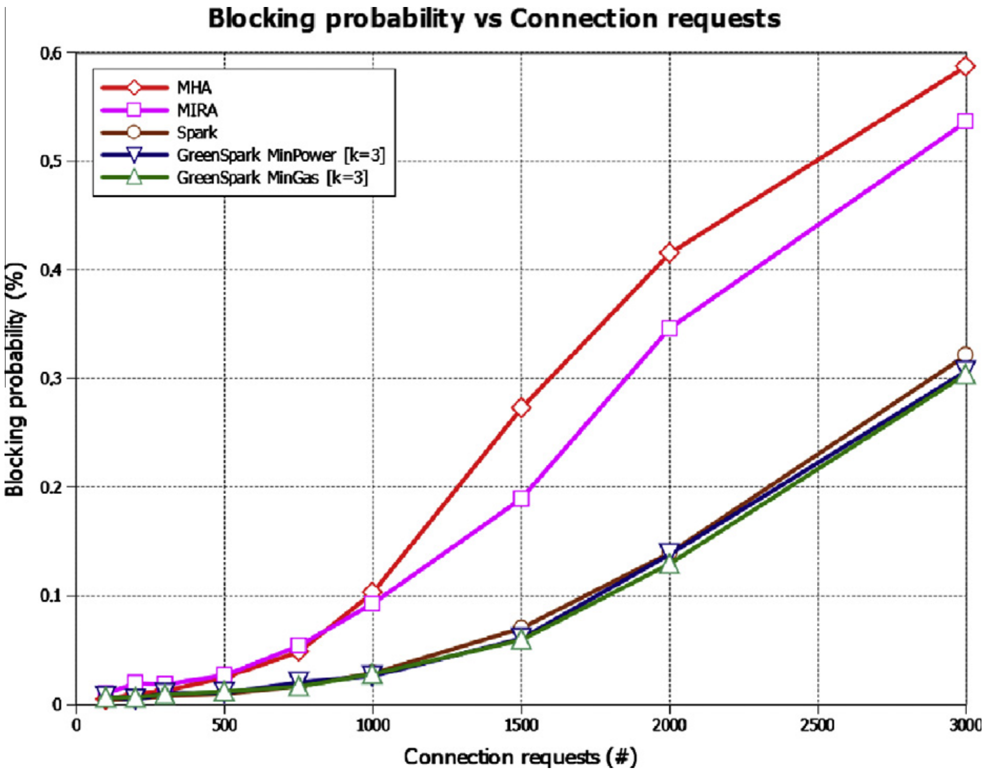


Fig. 6. Connection blocking probability versus connection requests.

the previous graph, performs better than MIRA. This is due to the longer paths chosen by MIRA with respect to MHA that, in turn, always chooses the shortest paths to route the connections (and, thus, statistically introduces less power consumption). In this graphic, we can also observe the first big difference inside the Spark family: the

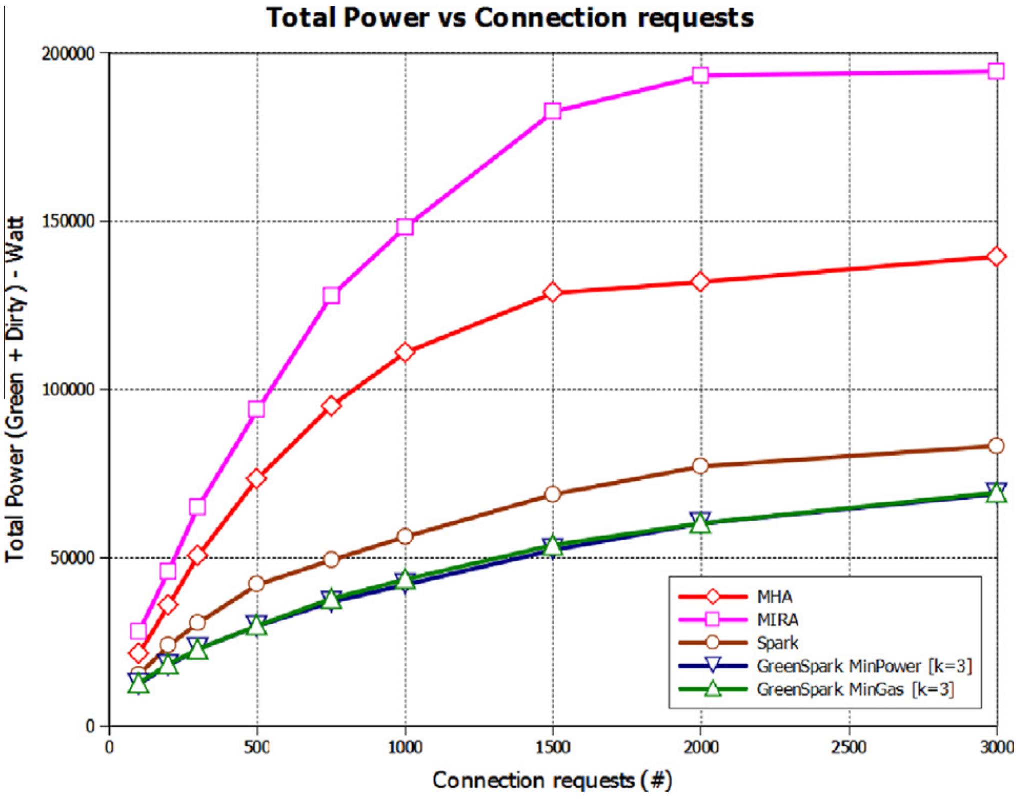


Fig. 7. Total power versus connection requests.

GreenSpark algorithms have lower total power consumption with respect to the energy-unaware Spark. Besides, we note that the two GreenSpark algorithms, MinPower and MinGas, perform almost the same as for the total power consumption, with MinPower doing slightly better, as expected. Anyway, the fact that the two GreenSpark

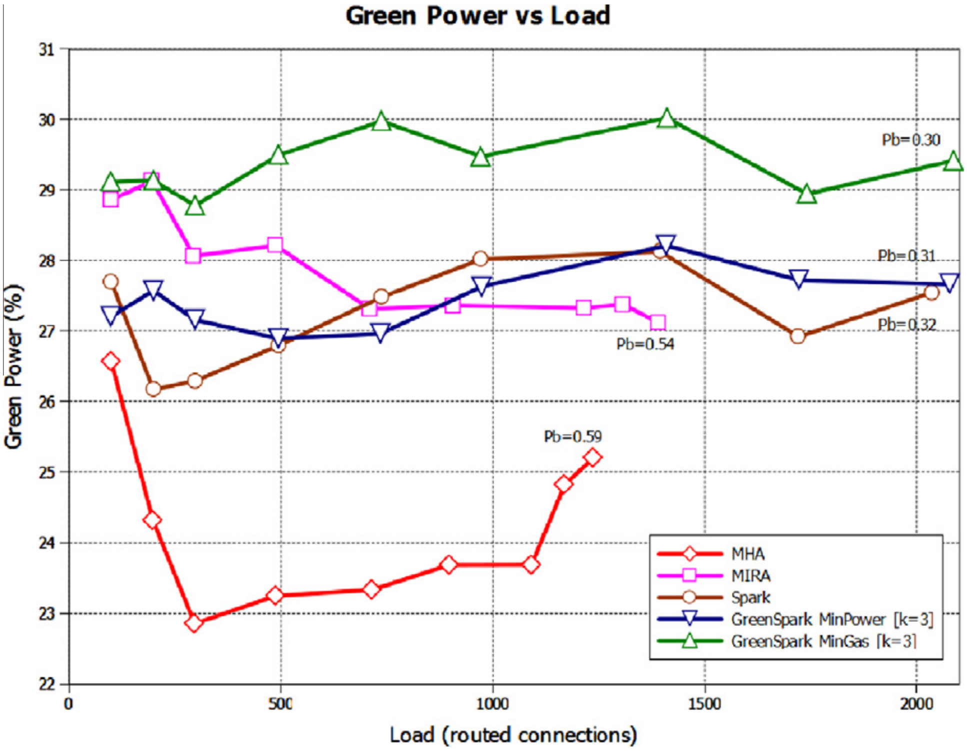


Fig. 8. Green power (percentage) versus load (routed connections).

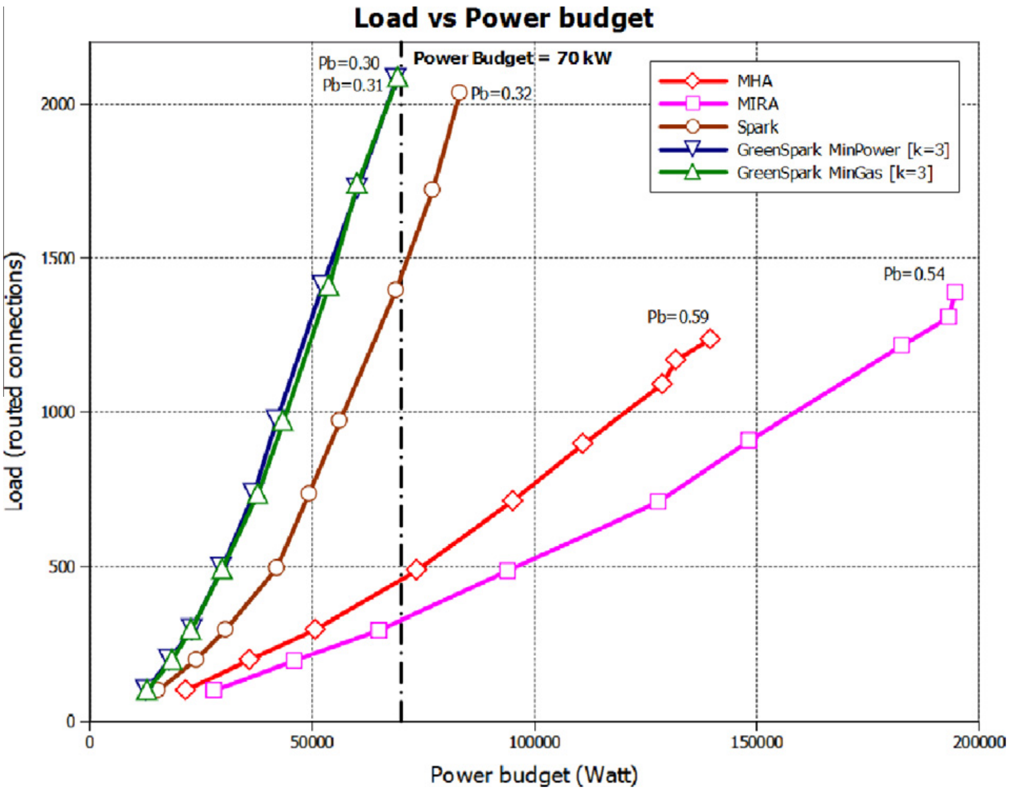


Fig. 9. Load (routed connections) versus power budget.

algorithms have almost the same total power consumption does not mean that their carbon footprint is the same. This leads us to the next graphic of Fig. 8, in which the green component of the power consumption has been reported.

The results, highlight that there is big difference in the use of green energy depending on the chosen GreenSpark optimization goal, as well as for the other algorithms. GreenSpark MinGas exhibits the topmost green power usage percentage, i.e. it prefers lightpaths passing through green-powered NEs and avoids sites powered by dirty sources as much as possible. More than 29% of the total power used by GreenSpark MinGas comes from green energy sources, thus saving considerable quantity of CO_2 from being emitted by the network during its operations. The other two algorithms of the Spark family are characterized by a lower green power usage, as expected, while keeping the blocking probability unaffected. Although being energy-unaware, MIRA performed quite well in our tests in terms of green power usage percentage, basically due to its minimum interference driving criteria which tends to balance the usage of network resources (whose energy is equally distributed among green and dirty energy sources), even if it exhibits a very high connection blocking probability, reaching values of 54% starting from a load of just 1450 connections. MHA exhibits the worst performance both for the green power usage and for the connection blocking probability, showing its limitations in complex network scenarios where a number of constraints, comprising the energy-efficiency ones, have to be taken into account. A particularly interesting issue comes from the observation of the pseudo-sinusoidal trend in the use of the green resources characterizing all the algorithms of the Spark family. This behavior is due to the specific cost and scoring functions associated with this family, in which less costly/greener paths will be chosen first, making the green energy percentage rise. As the usage of green paths raises, however, also the dynamic cost assigned to such paths increases as a consequence of their increased load (according to the load-balancing criteria of Eq. (8), until alternative non-green paths will be cheaper than the green ones and thus will be preferred for connections routing. This will make the green energy percentage decrease, but, at the same time, increase the cost of these alternative paths, until it will be again more convenient to route the incoming connections on green paths, and so on. Therefore, the pseudo-sinusoidal trend of the Spark family is somehow a visual proof of the efficiency of the two phase selection scheme which, at first, tries to balance the network load and, then, to minimize the specific scoring function, such as the total power (GreenSpark MinPower), the total GHG emissions (GreenSpark MinGas) or the total cost (Spark).

Starting from the consideration that it is a common practice that network operators contract a fixed power budget with their energy supplier and then strive to remain within that budget since surpassing the threshold will result in high penalty rates on the overall energy costs, in Fig. 9 we plotted the load versus the power budget required to route the connections. The energy-aware GreenSpark algorithms exhibit an almost optimal growth trend,

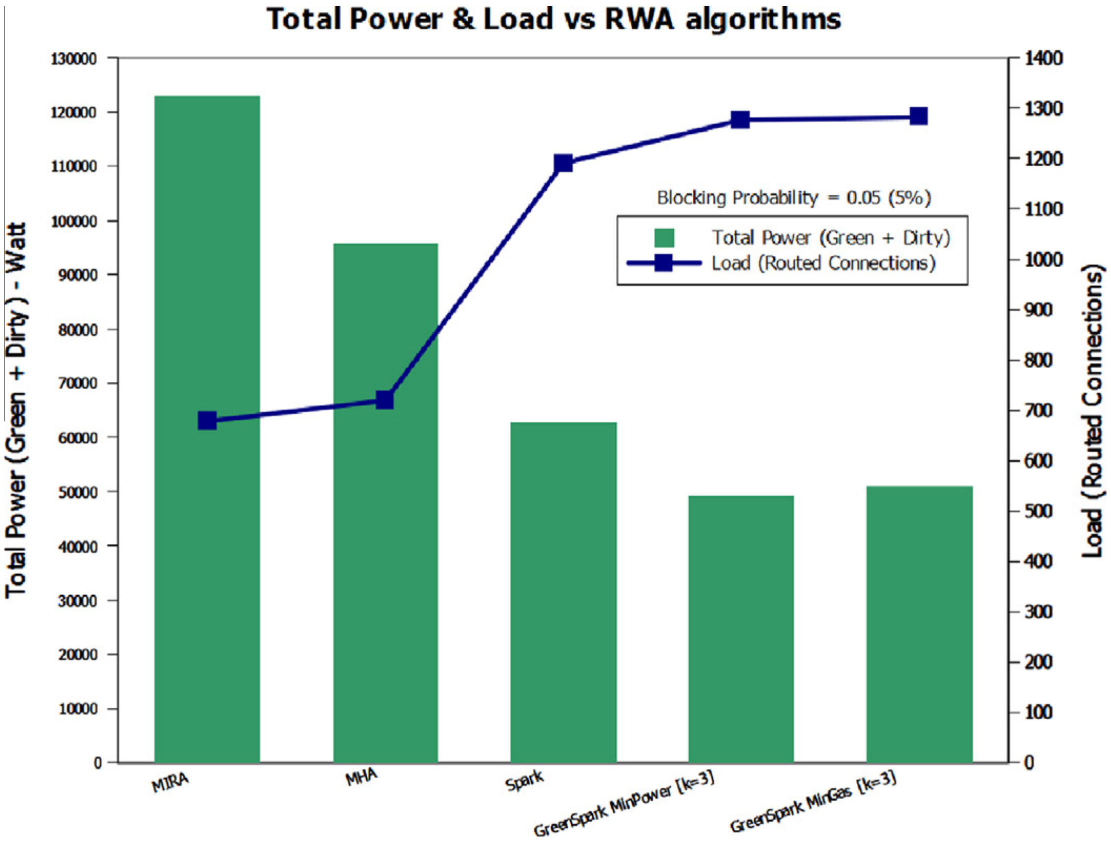


Fig. 10. Total power (green + dirty) and load (routed connections) versus different algorithms at a load causing a blocking probability of 0.05 (5%).

showing the highest increase in the load against a fixed increase in the power budget with respect to the other algorithms. We can observe that the entire set of connections can be routed in the network keeping its power budget below the 70 kW threshold (and the blocking probability at the lowest observed values). From this point of view, Spark performs notably well, considering that it is energy-unaware: its power budget is only 83 kW. It is

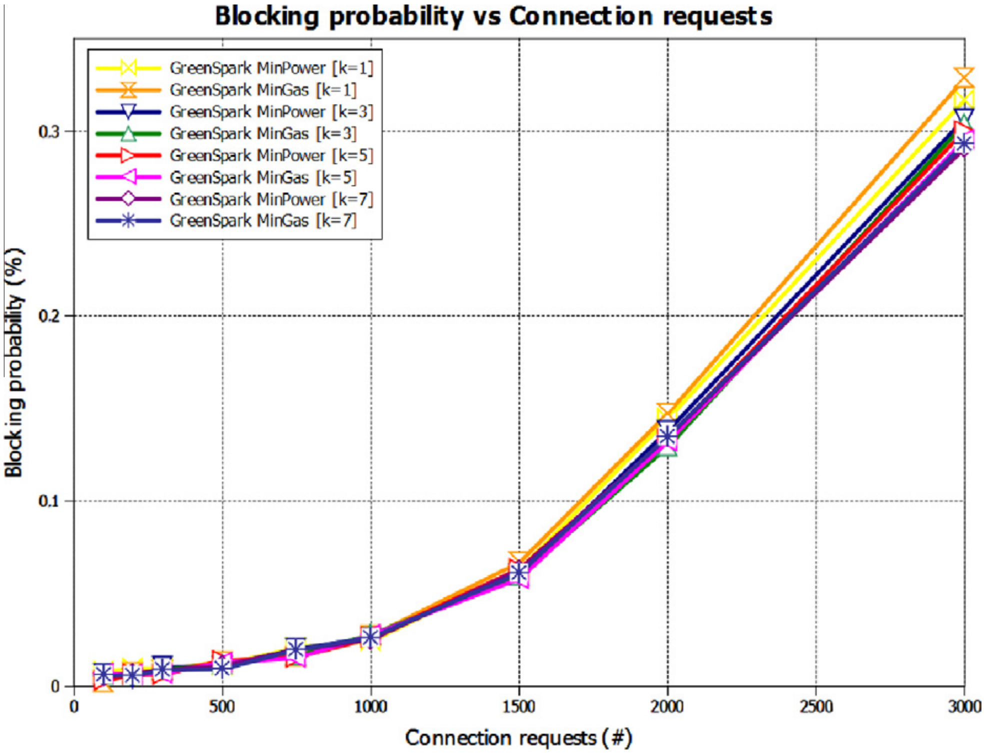


Fig. 11. Connection blocking probability versus connection requests.

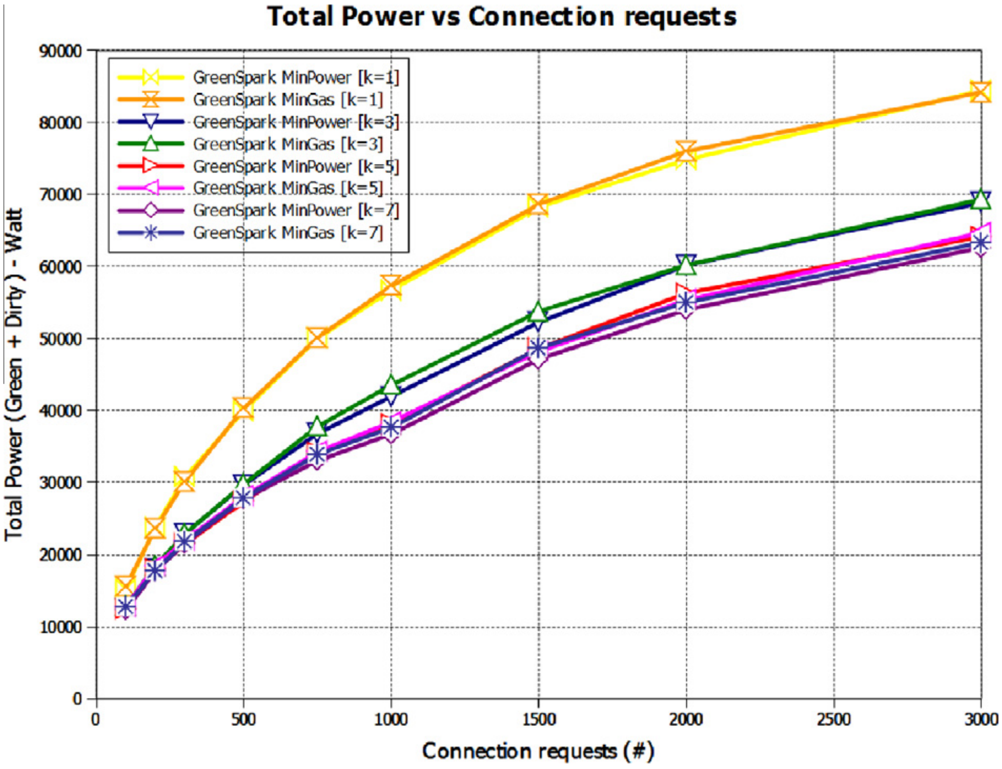


Fig. 12. Total power versus connection requests.

worthwhile to note that inside the power budget of the Spark algorithms, there are much many connections (more than 2000) with respect to the MHA and MIRA algorithms, which only route between 1200 and 1400 connections.

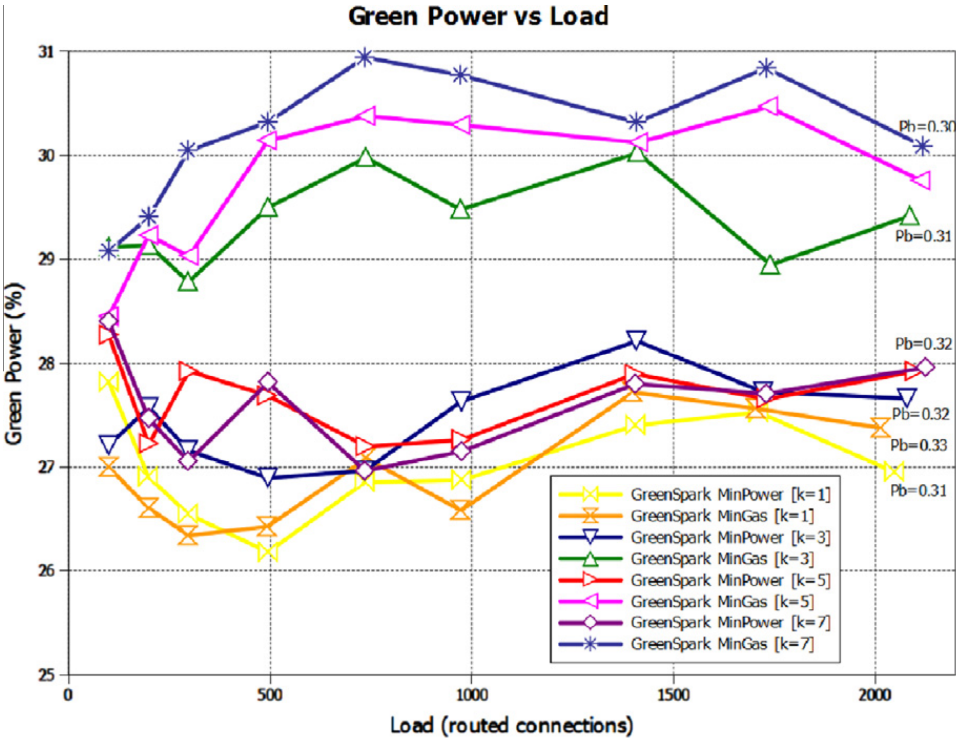


Fig. 13. Green power (percentage) versus load (routed connections).

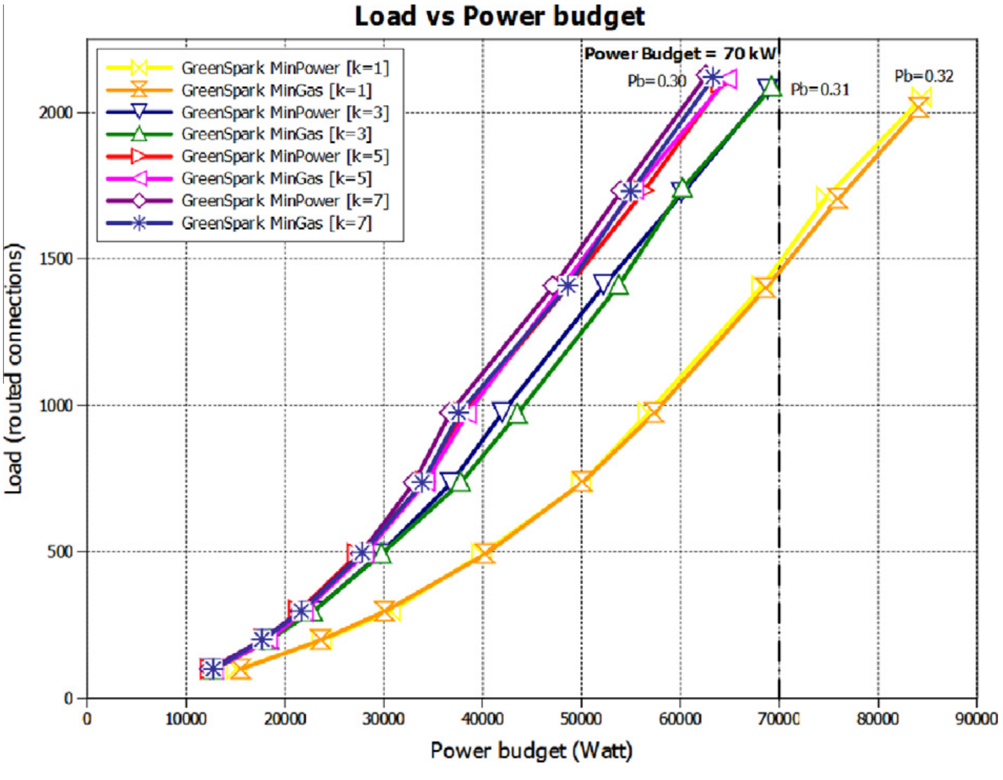


Fig. 14. Load (routed connections) versus power budget.

Furthermore, their power budgets are sensibly higher, being 140 kW and 195 kW respectively, between two and three times more than GreenSpark.

The last test of this series was conducted by keeping the QoS on the blocking probability at the constant value of 5%. We measured both the total required power and the load offered by the algorithms. The graphic in Fig. 10 clearly shows the great improvements introduced by the Spark family both for the power consumption and the routed connections. In particular, we observe that the GreenSpark algorithms need as low as half of the power required by MIRA and MHA and route double the number of their connections, making them very attractive also for QoS constrained networks with strict requirements on the connection blocking probability.

As a conclusion, we observe that there is a generation gap between the Spark family and the traditional MIRA and MHA algorithms, both in terms of power consumption and blocking probability. In particular, Spark performances are quite satisfactory, but GreenSpark algorithms, thanks to the two stages load-balancing and green objectives balanced by the k parameter, perform much better in terms of power and GHG, with MinGas even superior than MinPower, since it considerably lowers the GHG emissions while keeping almost the same total power requirements than MinPower. Results showed that GreenSpark algorithms not only significantly lower the required power and GHG emissions but also increase the connections acceptance ratio, showing that properly crafted RWA algorithms can enable greener networks with even better performance than before.

7.2. Tuning the GreenSpark k parameter

The k parameter value biases the load-balancing criteria of stage one (Eq. (8)) and the greenness criteria of stage two (Eq. (11) and Eq. (12)). Stage one restricts the set of possible paths to the best balanced k paths between ingress and egress nodes according to its cost function $\omega((u,v)_\lambda)$; stage two selects, among such paths, the greenest one according to its scoring function (MinGas or MinPower). At the extreme cases, a k value of 1 would restrict the stage one to always select the minimum cost path (thus, the best load-balanced path according to Eq. (8)) and the stage two to always select the only path available from stage one, making the algorithm totally energy-unaware, and therefore reducing it to a "simple" Dijkstra-based weighted shortest path (that is a minimum cost one); from the other side, a large enough k value (greater than the maximum number of possible paths between any two nodes) would make the algorithm totally "green", completely discarding the load-balancing effect of stage one. Therefore, smaller values of the k parameter bias the solution by privileging well balanced paths, while larger values of the k parameter privilege energy related objectives rather than the traditional network

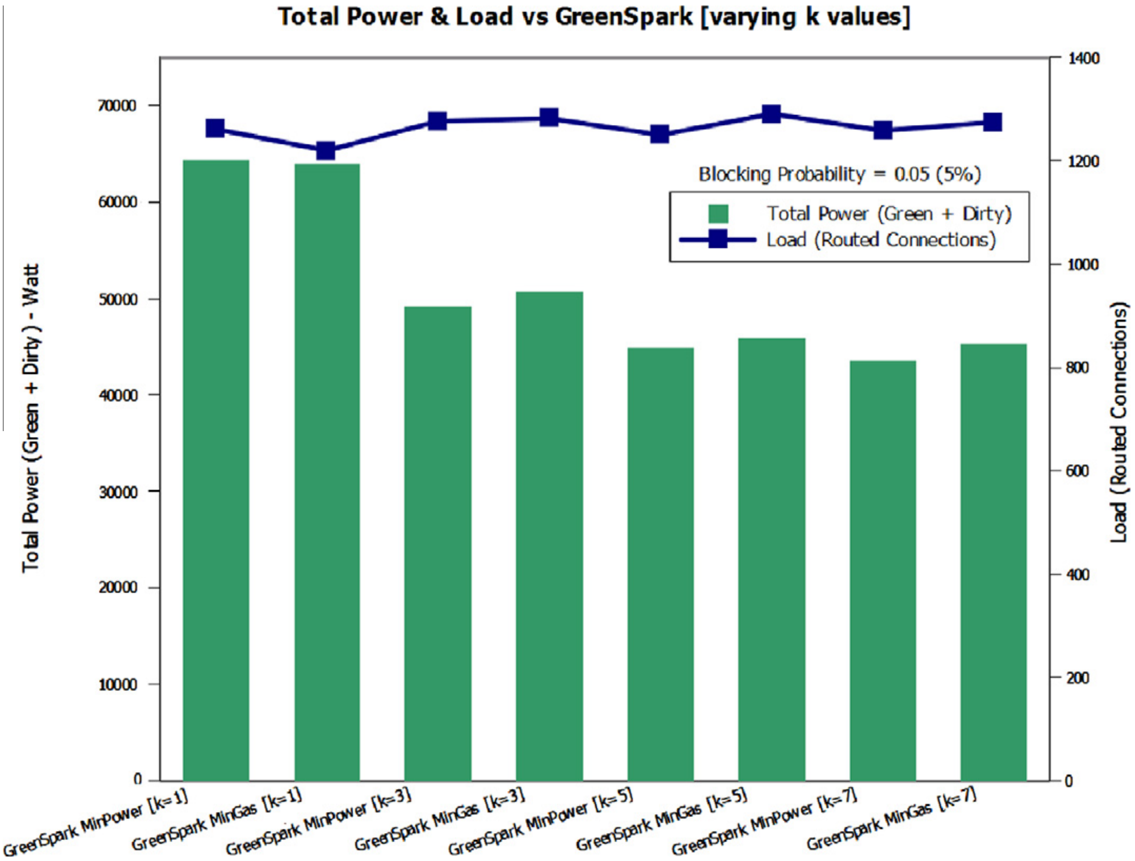


Fig. 15. Total power (green + dirty) & Load (routed connections) versus GreenSpark with varying k values at a load causing a blocking probability of 0.05 (5%).

management ones. A in depth study of the k parameter is therefore interesting and it is reported here for the Geant2 network topology.

In Fig. 11 we show the blocking probability of GreenSpark with varying values of the k parameter versus the connection requests. As expected, the k value affects very little the blocking ratio since the actual load-balancing is done in the first stage, whose output are always the k best balanced paths; in this sense, load-balancing is assured by stage one.

On the contrary, if we look at the total (green + dirty) power consumption versus the connection requests reported in Fig. 12, we observe that the higher the k value, the lower the total power consumption, as expected. In fact, with higher k values, the second stage will have the possibility to choose among a greater number of alternative paths to minimize ecological footprint of the network, both for MinPower and MinGas, which indeed perform very similar; in this sense, the green aspect is committed to stage two.

However, if we take a look inside the total power consumption at the green power percentage versus the load reported in Fig. 13, we see that there is a notable difference in the use of green energy sources at high k values. With $k = 5$ and $k = 7$, the green power percentage of GreenSpark MinGas markedly increases (whilst the MinPower is in the average as expected being GHG-unaware), showing that there is still room for green optimization at the expense of additional computational complexity due to the calculation of the higher number of alternative paths in the stage one. This result suggested the basis of a future work of ours, in which we will try to reach such greener paths through a one-stage algorithm with an omni-comprehensive energy-aware/load-balancing cost function employed to directly achieve such paths. We also note that the green power percentage for $k = 1$ is quite high, showing that the load-balancing may have positive effects on the GHG emissions when the energy sources are heterogeneously distributed in the network.

In Fig. 14 we plotted the load versus the required power budget for different k values. As seen in Fig. 12, with the same k value, MinPower and MinGas perform quite similarly in terms of total power consumption, but GreenSpark will require different power budgets depending on the k value. The higher the k, the lower the power budget but also the higher the computational complexity required for the path calculation at each connection request set-up time. Anyway, it is worthwhile to note that there is great improvement between $k = 1$ and $k = 3$, and only limited gain for greater values, meaning that already with $k = 3$ alternative paths the GreenSpark framework is able to sensibly reduce the power budget in an optimal balance between greenness and performance.

Finally, the results of the test on the QoS on the blocking probability at the constant value of 5% is shown in Fig. 15. The total power required decreases with the increase of the k parameter value but again, while from $k = 1$ to $k = 3$ there is a great reduction of the total power, when passing from $k = 3$ to $k = 5$ and $k = 7$ there is no such a great benefit. Note also that, at such low load (5% of blocking probability), there is no great difference in varying the k values as for the number routed connections since most of the connections will have sufficient resources to be routed, even if a slightly better performance is observed in correspondence of the $k = 3$, i.e. when both the load-balancing and the greenness objectives are fairly weighted.

8. Conclusions and future work

In this work, we focused our research effort on GreenSpark, a novel heuristic-driven dynamic RWA framework aiming at the minimization of power consumption and GHG emissions in wavelength routed backbone networks. GreenSpark operates by progressively routing the dynamically incoming connections on a two-stage basis; in the first stage, a set of k feasible paths is found according to traditional load-balancing objective. Then, in the second stage, the greenness of the k paths is evaluated both in terms of power consumption (MinPower) and GHG emissions (MinGas), and the greenest path is finally selected to route the connection. Even with low k values (i.e. $k = 3$), and despite its very low computational complexity, GreenSpark achieves significant power savings and carbon footprint reduction together with an increment of the load-balance, resulting in lower blocking probability as compared with several widely used routing algorithms, as verified by the extensive simulation study.

Apart from defining an energy consumption model for the IP over WDM network, one of the most significant added values of the framework is the incorporation of both physical layer issues, such as power demand of each component, and virtual topology-based energy management with integrated traffic grooming, adversely conditioning the usage of energy hungry links and devices. Moreover, since the above model also takes into account the type of power supply associated with each device, by privileging green sources, the proposed scheme can also be useful for equalizing the carbon footprint of entire areas within a real network scenario in which each device location may be characterized by a differentiated (green or dirty) energy source. Here, multi-objective optimization may help us in finding the appropriate trade-off according to the relative importance of network performance and environmental friendliness.

As future work, we are studying an omni-comprehensive energy-aware/load-balancing cost function to directly find green paths in a single stage with even lower computational complexity. In addition, we are investigating new energy-aware traffic engineering strategies and network re-optimization methods, aiming at dynamically reducing power demand, GHG emissions and costs on a time basis, by moving data wherever electricity costs are lowest at a particular time.

References

[1] S. Aleksic, Analysis of power consumption in future high-capacity network nodes, IEEE/OSA Journal of Optical Communications and Networking 1 (3) (2009) 245–258, http://dx.doi.org/10.1364/JOCN.1.000245.

[2] B. Project, WP 21 TP green optical networks, D21.2b report on Y1 and updated plan for activities, 2009.

[3] J. Baliga, R. Ayre, K. Hinton, W. Sorin, R. Tucker, Energy consumption in optical IP networks, Journal of Lightwave Technology 27 (13) (2009) 2391–2403, http://dx.doi.org/10.1109/JLT.2008.2010142.

[4] G. Shen, R. Tucker, Energy-minimized design for IP over WDM networks, IEEE/OSA Journal of Optical Communications and Networking 1 (1) (2009) 176–186, http://dx.doi.org/10.1364/JOCN.1.000176.

[5] F. Idzikowski, S. Orlowski, C. Raack, H. Woesner, A. Wolisz, Saving energy in IP-over-WDM networks by switching off line cards in low-demand scenarios, in: Optical Network Design and Modeling (ONDM), 2010 14th Conference on, 2010, pp. 1–6. doi:http://dx.doi.org/10.1109/ONDM.2010.5431569.

[6] J. Wang, S. Ruepp, A.V. Manolova, L. Dittmann, S. Ricciardi, D. Careglio, Green-aware routing in GMPLS networks, in: Computing, Networking and Communications (ICNC), 2012 International Conference on, 2012, pp. 227–231. http://dx.doi.org/10.1109/ICCNC.2012.6167416.

[7] I. Chlamtac, A. Ganz, G. Karmi, Lightpath communications: an approach to high bandwidth optical wan's, IEEE Transactions on Communications 40 (7) (1992) 1171–1182, http://dx.doi.org/10.1109/26.153361.

[8] F. Palmieri, U. Fiore, S. Ricciardi, SPARK: a smart parametric online RWA algorithm, Journal of Communications and Networks 9 (4) (2007) 368–376.

[9] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, D. Wetherall, Reducing network energy consumption via sleeping and rate-adaptation, in: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, NSDI'08, 2008, pp. 323–336.

[10] S. Ricciardi, D. Careglio, G. Santos-Boada, J. Solé-Pareta, U. Fiore, F. Palmieri, Towards an energy-aware internet: modeling a cross-layer optimization approach, Telecommunication Systems (2011) 21–22.

[11] A. Muhammad, P. Monti, I. Cerutti, L. Wosinska, P. Castoldi, A. Tzanakaki, Energy-efficient WDM network planning with dedicated protection resources in sleep mode, in: GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference, 2010, pp. 1–5. doi:http://dx.doi.org/10.1109/GLOCOM.2010.5683205.

[12] S. Ricciardi, D. Careglio, F. Palmieri, U. Fiore, G. Santos-Boada, J. Solé-Pareta, Energy-oriented models for WDM networks, in: Green Networking 2010 Workshop (GN2010), Co-located with the 7th International ICST Conference on Broadband Communications, Networks, and Systems (Broadnets), 2010.

[13] M. Gupta, S. Singh, Greening of the internet, in: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM '03, 2003, pp. 19–26.

[14] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, S. Wright, Power awareness in network design and routing, in: INFOCOM 2008, The 27th Conference on Computer Communications, IEEE, 2008, pp. 457–465. doi:http://dx.doi.org/10.1109/INFOCOM.2008.93.

[15] E. Star, Small network equipment, June 2011.

[16] C. Gunaratne, K. Christensen, B. Nordman, S. Suen, Reducing the energy consumption of ethernet with adaptive link rate (alr), IEEE Transactions on Computers 57 (4) (2008) 448–461, http://dx.doi.org/10.1109/TC.2007.70836.

[17] R. Hays, Active/idle toggling with 0BASE-x for energy efficient ethernet, in: IEEE 802.3az Task Force, 2007.

[18] Y. Wu, L. Chiaraviglio, M. Mellia, F. Neri, Power-aware routing and wavelength assignment in optical networks, in: 35th European Conference on Optical Communication, ECOC'09, IEEE, 2009, pp. 1–2.

[19] S. Ricciardi, D. Careglio, F. Palmieri, U. Fiore, G. Santos-Boada, J. Solé-Pareta, Energy-aware RWA for WDM networks with dual power sources, in: 2011 IEEE International Conference on Communications (ICC), 2011, pp. 1–6. doi:http://dx.doi.org/10.1109/icc.2011.5962432.

[20] L. Chiaraviglio, M. Mellia, F. Neri, Reducing power consumption in backbone networks, in: Proceedings of the 2009 IEEE International Conference on Communications, ICC'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 2298–2303. doi:http://dx.doi.org/10.1109/ICC.2009.5199404.

[21] A. Silvestri, A. Valenti, S. Pompei, F. Matera, A. Cianfrani, Wavelength path optimization in optical transport networks for energy saving, in: 11th International Conference on Transparent Optical Networks. ICTON '09, 2009, pp. 1–5. doi:http://dx.doi.org/10.1109/ICTON.2009.5185212.

[22] Kist, Alexander A, Aldraho, Abdelnour, Dynamic topologies for sustainable and energy efficient traffic routing, Computer Networks, 55 (9) (2011) 2271–2288, http://dx.doi.org/10.1016/j.comnet.2011.03.008.

[23] F. Farahmand, M. Hasan, I. Cerutti, J. Jue, J. Rodrigues, Power-efficient lightpath-based grooming strategies in WDM mesh networks, in: 2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN), 2011, pp. 1–6. doi:http://dx.doi.org/10.1109/ICCCN.2011.6006064.

[24] C. Systems, Cisco IP router CR1, September 2011.

[25] J. Networks, T series routing platforms, September 2011.

[26] W. Vereecken, W. Van Heddeghem, D. Colle, M. Pickavet, P. Demeester, Overall ICT footprint and green communication technologies, in: 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP), 2010, pp. 1–6. doi:http://dx.doi.org/10.1109/ISCCSP.2010.5463327.

[27] S. Ricciardi, D. Careglio, U. Fiore, F. Palmieri, G. Santos-Boada, J. Solé-Pareta, Analyzing local strategies for energy-efficient networking, in: V. Casares-Giner, P. Manzoni, A. Pont (Eds.), NETWORKING 2011 Workshops, Lecture Notes in Computer Science, vol. 6827, Springer, Berlin/Heidelberg, 2011, pp. 291–300.

[28] A. Adelin, P. Owezarski, T. Gayraud, On the impact of monitoring router energy consumption for greening the internet, in: 2010 11th IEEE/ACM International Conference on Grid Computing (GRID), 2010, pp. 298–304. doi:http://dx.doi.org/10.1109/GRID.2010.5697988.

[29] M. Kodialam, T. Lakshman, Minimum interference routing with applications to MPLS traffic engineering, in: INFOCOM 2000, Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, vol. 2, 2000, pp. 884–893. doi:http://dx.doi.org/10.1109/INFCOM.2000.832263.

[30] K. Katrinis, A. Tzanakaki, On the dimensioning of WDM optical networks with impairment-aware regeneration, IEEE/ACM Transactions on Networking 19 (3) (2011) 735–746, http://dx.doi.org/10.1109/TNET.2010.2090540.

[31] N. Katoh, T. Ibaraki, H. Mine, An efficient algorithm for K shortest simple paths, Networks 12 (4) (1982) 411–427, article id 8506112, http://dx.doi.org/10.1002/net.3230120406.

[32] R. Ramamurthy, B. Mukherjee, Fixed-alternate routing and wavelength conversion in wavelength-routed optical networks, IEEE/ACM Transactions on Networking 10 (3) (2002) 351–367, http://dx.doi.org/10.1109/TNET.2002.1012367.

[33] A.V. Goldberg, R.E. Tarjan, A new approach to the maximum flow problem, Journal of the ACM 35 (1988) 921–940.

[34] Geant, The Geant2 network, June 2011.

[35] R. Ramaswami, K. Sivarajan, Design of logical topologies for wavelength-routed optical networks, IEEE Journal on Selected Areas in Communications 14 (5) (1996) 840–851, http://dx.doi.org/10.1109/49.510907.

[36] S. Uhlig, B. Quoitin, J. Lepropre, S. Balon, Providing public intradomain traffic matrices to the research community, SIGCOMM Computer Communication Review 36 (2006) 83–86.

[37] F.Idzikowski, Power consumption of network elements in IP over WDM networks, TKN Technical Report Series TKN-09-006, Telecommunication Networks Group, Technical University Berlin, July 2009.

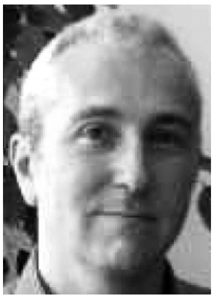
[38] F. Palmieri, U. Fiore, S. Ricciardi, Simulnet: a wavelength-routed optical network simulation framework, in: IEEE Symposium on Computers and Communications. ISCC 2009, pp. 281–286. doi:http://dx.doi.org/10.1109/ISCC.2009.5202259.

[39] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, RSVP-TE: Extensions to RSVP for LSP tunnels, Internet Draft draft-ietf-mplsrsvp-lsp-tunnel-04.txt, 1999.

[40] K. Kar, M. Kodialam, T. Lakshman, Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications, IEEE Journal on Selected Areas in Communications 18 (12) (2000) 2566–2579, http://dx.doi.org/10.1109/49.898737.

Sergio Ricciardi is a research associate in the Advanced Broadband Communications Center (CCABA) at the Department of Computer Architecture of the Technical University of Catalonia (UPC). He holds two Masters of Science in Computer Science (University of Naples Federico II, Italy, in 2006 and Technical University of Catalonia, Spain, in 2010, both with honors). He worked with the Federico II University and with the Italian National Institute for Nuclear Physics (INFN) within several national and international projects. His research interests are mainly focused on energy-aware RWA algorithms and protocols for telecommunication networks and energy-oriented optimizations for grid/cloud computing.

Francesco Palmieri is an assistant professor at the Engineering Faculty of the Second University of Napoli, Italy. His major research interests concern high performance and evolutionary networking protocols and architectures, routing algorithms and network security. Since 1989, he has worked for several international companies on networking-related projects and, starting from 1997, and until 2010 he has been the Director of the telecommunication and networking division of the Federico II University, in Napoli, Italy. He has been closely involved with the development of the Internet in Italy as a senior member of the Technical-Scientific Advisory Committee and of the CSIRT of the Italian NREN GARR. He has published a significant number of papers in leading technical journals and conferences and given many invited talks and keynote speeches.



Ugo Fiore leads the Network Operations Center at the Federico II University, in Naples. He began his career with Italian National Council for Research and has also more than 10 years of experience in the industry, developing software support systems for telco operators. His research interests focus on optimization techniques and algorithms aiming at improving the performance of high-speed core networks. He is also actively pursuing two other research directions: the application of non-linear techniques to the analysis and classification of traffic; security-related algorithms and protocols.

Davide Careglio (S'05–M'06) received the M.Sc. and Ph.D. degrees in telecommunications engineering both from Universitat Politcnica de Catalunya (UPC), Barcelona, Spain, in 2000 and 2005, respectively, and the Laurea degree in electrical engineering from Politecnico di Torino, Turin, Italy, in 2001. He is currently an Associate Professor in the Department of Computer Architecture at UPC. Since 2000, he has been a Staff Member of the Advanced Broadband Communication Center. His research interests include networking protocols with emphasis on optical switching technologies, and algorithms and protocols for traffic engineering and QoS provisioning. He is the coauthor of more than 80 publications in international journals and conferences. He has participated in many European and national projects in the field of optical networking and green communication.

Germán Santos-Boada obtained his M.Sc. degree in Telecom Engineering in 1978, and his Ph.D. in 1993, both from the Technical University of Catalonia (UPC). He worked for Telefónica as manager of engineering from 1984 up to 2007 and simultaneously he joined the Computer Architecture Department of UPC as a partial time Assistant Professor. Currently he is full time Assistant Professor with this department. Dr. Santos current research interests are Quality of Service provisioning in next generation optical access networks and optical energy-aware network modeling. He is currently involved in the COST 804 action.

Josep Solé-Pareta obtained his M.Sc. degree in Telecom Engineering in 1984, and his Ph.D. in Computer Science in 1991, both from the Technical University of Catalonia (UPC). In 1984 he joined the Computer Architecture Department of UPC. Currently he is Full Professor with this department. He did a Postdoc stage (summers of 1993 and 1994) at the Georgia Institute of Technology. He is cofounder of the UPC-CCABA, and UPC-N3cat. His publications include several book chapters and more than 150 papers in relevant research journals (>25), and refereed international conferences. His current research interests are in Nanonetworking Communications, Traffic Monitoring, Analysis and High Speed and Optical Networking and Energy Efficient Transport Networks, with emphasis on traffic engineering, traffic characterization, MAC protocols and QoS provisioning. He has participated in many European projects dealing with Computer Networking topics.

2011 First International Conference on Data Compression, Communications and Processing

Saving Energy in Data Center Infrastructures

Sergio Ricciardi
Davide Careglio
Germán Santos-Boada
Josep Solé-Pareta
Dept. d'Arquitectura de Computadors
Universitat Politècnica de Catalunya
Barcelona, Spain
e-mail: sergior@ac.upc.edu

Ugo Fiore
Centro Servizi Informativi
Università di Napoli Federico II
Naples, Italy
e-mail: ufiore@unina.it

Francesco Palmieri
Dipartimento di Ingegneria
dell'Informazione
Seconda Università di Napoli
Aversa, Italy
e-mail: fpalmier@unina.it

Abstract— **At present, data centers consume a considerable percentage of the worldwide produced electrical energy, equivalent to the electrical production of 26 nuclear power plants, and such energy demand is growing at fast pace due to the ever increasing data volumes to be processed, stored and accessed every day in the modern grid and cloud infrastructures. Such energy consumption growth scenario is clearly not sustainable and it is necessary to limit the data center power budget by controlling the absorbed energy while keeping the desired level of service. In this paper, we describe EnergyFarm, a data center energy manager that exploits load fluctuations to save as much energy as possible while satisfying quality of service requirements. EnergyFarm achieves energy savings by aggregating traffic during low load periods and temporary turning off a subset of computing resources. EnergyFarm respects the logical and physical dependencies of the interconnected devices in the data center and performs automatic shut down even in emergency cases such as temperature peaks and power leakages. Results show that high resource utilization efficiency is possible in data center infrastructures and that huge savings in terms of energy (MWh), emissions (tons of CO$_2$) and costs (k€) are achievable.**

Energy-efficiency, power management, sleep mode, green data centers, grid computing, cloud computing.

I. INTRODUCTION

It is estimated that worldwide data centers alone consume 26 GW of electrical power corresponding to about 1.4% of the worldwide electrical energy consumption, with a growth rate of 12% per year [1][2]. To give an idea, the Barcelona Supercomputing Center (a medium-size data center) pays every year more than € 1 million just for the energy bill and consumes 1.2 MW [3], as much power as a town of 1,200 houses [4]. The power consumption in data centers originates from the involved computing, storage and interconnection equipment, together with the associated HVAC (heating, ventilation and air conditioning), UPS (uninterruptible power supply) systems and lighting facilities, with the servers being the most energy-hungry devices. The power usage effectiveness (PUE) index, defined by the Green Grid [5], measures the efficiency of an ICT facility as the ratio of the total amount of power used by the facility to the power delivered to the computing equipment alone. While larger data centers tend to be able to implement more efficient cooling, high availability needs may require the use of expensive UPS and more redundancy, which then result in a higher PUE. A PUE value of 2 is the current average [6], meaning that HVAC and UPS double the energy requirements. In data centers, a wide variety of computing resources are usually available, ranging from small servers with computational capabilities comparable to personal computers, to large supercomputers. Furthermore, there are different types of servers optimized for specific tasks such as web and database servers. One of the largest problems of this equipment is the relative independence of the power consumption with their real operating load [6] and the consequent need for energy-proportional architectures [7]. This, combined with the fact that many servers are being operated far below their actual capacity [7][8], leads to a lot of wasted energy in data centers, thus enabling great potential energy savings. Next to using optimized components, a second level of optimization lies in power management. In such a scenario, three energy-saving approaches are available: "do less work", "slow down" and "turn off idle elements". In the "do less work" strategy, the processes are optimized so that the load to be executed becomes minimal, resulting in lower power consumption. The "slow down" strategy considers that the faster a process runs, the more resource intensive it becomes. In complex processes, the speeds of several sub-processes don't match and thus resources are used without being absolutely required. There are two ways of slowing down processes. They can be run with adaptive speeds, by selecting the minimal required speed to complete the process in time. Alternatively, buffering can be introduced so that instead of running a process immediately upon arrival, one can collect new tasks until the buffer is full and then execute them in bulk. This allows for components to be temporarily switched off resulting in lower power consumption. The "turn off idle elements" strategy refers to the possibilities offered by exploiting a low-consumption state (sleep mode). Basically, the sleep mode aims at switching into an idle mode the devices during periods of inactivity. Unloaded servers can be dynamically put into sleep mode during low-load periods, contributing to great power savings. For data centers and grid/cloud infrastructures, if properly employed, the sleep mode may represent a very useful mean for limiting power

978-0-7695-4528-8/11 $26.00 © 2011 IEEE
DOI 10.1109/CCP.2011.9

consumption of lightly loaded sites. Data centers are in fact inherently modular, as they are built up by a number of logical-equivalent elements (bulks of servers). A grid site, for example, is basically composed by a disk pool manager (DPM) that controls data storage (storage elements SE, disk servers DS, storage systems SS), and a computing element (CE) that sends jobs to working nodes (WN). While the DPM and the CE are usually hosted on individual servers (for a grid of medium size), the SEs and especially the WNs functions may be distributed over a very large number of nodes. All these servers are up and running even if the farm is scarcely loaded or idle. Aggregating the jobs on a subset of SEs and WNs allows putting into sleep mode all the remaining nodes greatly reducing the energy consumption (assuming replicated data on SEs). In this direction, we started from the PowerFarm software [9] developed to manage power losses and temperature/humidity peaks in grid sites, that can be used to automatically shut down devices in case of emergency (such as temperature peaks, smoke or fire alerts, etc). It works by processing the SNMP trap alerts and taking the corresponding preconfigured actions, but it lacks the intelligence to take any energy saving action. Thus, we extended the PowerFarm framework to monitor current loads and server power consumptions and to turn on/off servers as needed while respecting physical and logical dependencies among them. Accordingly, we developed EnergyFarm, a simple and effective energy control system, which, through a service-demand matching algorithm, determines the subset of servers that may be powered off while satisfying the data center computing and storage demand.

II. Related Work

In order to reduce the energy consumption of data centers, a number of directions have been highlighted in the literature. In [1] it is argued that significant power savings can be realized through virtual server configurations, allowing to switch off most servers during night hours and only using the full capacity of servers during peak hours. In [10], a number of measures are identified: legacy equipment requiring appropriate software may undergo hardware upgrades (such as modified power supply modules) and their network presence may be transferred to a proxy or agents allowing the end device to be put in low consuming mode during inactivity periods while being virtually connected to the Internet. The authors also plead the need to enable renewable energy sources, such as solar, wind or hydro power, to supply power to ICT systems. This approach seems specifically applicable to data centers, which can be located at renewable energy production sites. However, since renewable energy sources tend to be unpredictable (e.g., wind), or vary during day and night (e.g., sun), this would imply that the data itself need to be migrated from one data center to the other, according a so-called follow-the-sun or chase-the-wind scenario [11]. As a consequence, energy-efficient high bandwidth networks and routing architectures will be required. Along this line of thought, a study performed in [12] investigates cost-aware and energy-aware load distribution across multiple data centers. The study evaluates the potential cost and carbon savings for data centers located in different time zones and partly powered by green energy and founds that, when optimizing for green energy use, green data centers can decrease CO_2 emission by 35% by leveraging the green data centers at only a 3% cost increase. Several sources [1][6][13][14] in literature have pointed out the sleep mode as a solution for achieving energy-efficiency. In [6] the authors focus on component level where more efficient technologies should be used. In [15] a network power manager is presented, which dynamically adjusts the set of active network elements (links and switches) to satisfy changing data center traffic loads; it is focused on the network infrastructure of the data centers. Our work is instead focused on improving the operating energy efficiency of the computing resources (servers), which are responsible for the greatest part of data centers energy consumption.

III. An Energy-Aware Data Center Control Plane

A data center is composed by a number of servers running jobs (or tasks) that come from the Internet. Every server has a processing capacity, depending essentially on the number of cores and/or processors. The data center *workload* is thus represented by the jobs that the data center has to process in each moment. Typical data centers are strongly over-provisioned to work well under peak workloads [7][8]. However, idle servers are normally kept turned on even if there are no jobs to process. This clearly represents a waste in the power utilization and a cost in the energy bill. Our goal is to reduce the set of active servers to a subset of servers and turn off the idle ones, according to the "turn off idle elements" approach. In this scenario, EnergyFarm will be the high-level energy-aware control plane logic that complements and extends the low-level PowerFarm actuator facilities, which physically manage the power distribution in the data center. In order to exploit load fluctuations by turning off inactive servers and saving energy, we defined a specific operating *policy* within EnergyFarm that establishes *what* should be done, and implemented the corresponding *mechanisms* in PowerFarm which specify *how* it should be realized. Thus, as the job traffic load changes, the policy indicates which servers have to be turned off and the PowerFarm facilities implement the correct procedures to accomplish the task while respecting the physical and logical dependencies among the data center devices. In particular, the main EnergyFarm operating policy has been implemented through a proactive algorithm – running within the farm resource broker – that constantly monitors the traffic load and dynamically decides the subset of servers that may be turned on/off, while the PowerFarm actuator functions have the task to correctly power on/off such servers.

A. Modeling Resource Allocation and Traffic Fluctuations

In our model, we follow the usual scenario in which each job is assigned to one CPU core, so that a server with multiple CPUs making available *n* total cores may run *n* jobs without experiencing any performance slowdown. For multi-core CPUs, we take advantage of such characteristic by aggregating jobs on a subset of servers in order have more idle servers to turn off. For servers with *n* CPU cores, several aggregation strategies are possible: among the active servers, first-fit assigns a new job to the first server with one CPU core available. Best-fit tries to compact the jobs as much as possible; the new job is assigned to a server with just 1 core free (and, thus, $n - 1$ busy) if any such server exists. Otherwise, it looks for a server with 2 free cores, then with 3, and so on, up to *n*. Clearly, first-fit is faster but it leaves a great number of servers not fully loaded; best-fit gives the best results since it compacts the jobs as much as possible and frees the maximum number of servers that may be turned off. Besides compacting as much as possible, best-fit is also profitable since a multi-core server with a high number of busy cores is less probable to get free of all his jobs (and, thus, of being put into sleep mode) than a server with a low number of jobs. Best-fit computational complexity may be improved to work in constant amortized time by implementing the server priority queue with a Fibonacci heap. If servers are single-core devices, no aggregation is possible and all the energy savings come exclusively from the shutdown of the idle servers.

Typical data centers traffic demands are not constant over time: on the contrary, they are characterized by high utilization periods (e.g., during some hours of the day) followed by low utilization periods (e.g., during the night). In particular, it has been observed that the traffic load fluctuations are almost predictable within certain fixed time periods (e.g., day-night, months or years) and resemble a pseudo-sinusoidal trend [15][16]. In Fig. 1 it is reported the theoretical daily traffic variation of a typical energy-unaware production site [15]: the traffic load (demand) follows a pseudo-sinusoidal trend whilst the power keeps constant during high and low load periods. This behavior is due to data center resources that are always on and consume energy even during low load periods. The idea is to introduce *elasticity* in the demand-capacity provisioning, by dynamically varying the capacity with the demand, like depicted in Fig. 2. In theory, the capacity should resemble the demand as closely as possible, but two main problems have to be addressed. First, the capacity is not a continuous curve but is instead a step function in which each step corresponds to a computing resource (e.g., a server in the farm) turned on/off. Thus, the demand curve has to be approximated with a step service curve that serves the demand while minimizing the energy consumption (Fig. 3). Second, the provisioned capacity should have a *safety margin* (i.e., a distance *d* between the demand and the capacity curves) to cope with peak loads. The margin represents the number of servers that are preventively turned on for serving new jobs to come. The smaller the *d*, the lower the energy consumption, but also the lower the number of jobs that will be served without delay. The higher the *d* value, the more the jobs that will be served as they arrive, but also the higher the energy consumption (since a greater number of servers have to be powered on). The safety margin *d* has to be large enough to avoid oscillating between states for little variations of the load. During the start up of a server, in fact, peaks in the power absorption are experienced, due to the server bootstrap procedure and the OS loading process. Therefore, *d* is upper-bounded by the energy consumption and lower-bounded by the peak load absorption capacity and oscillation minimization requirement. At any instant, the absorbed power is directly proportional to the number of active servers; so the closer the service curve resembles the demand curve, the lower the required power will be. With the safety margin *d*, a bulk of $k \leq d$ incoming jobs will not have to wait. Thus, the *d* parameter sets the size of the zero-waiting queue of jobs that are immediately served as they arrive. If $k > d$, there will be $k - d$ jobs that will have to wait a time *t* before they can get served, where *t* is the start-up time of the servers (obviously, if the load reaches the site maximum capacity, all new jobs will have to wait for new resources to become available). The start-up time *t* may sensibly vary with the available technology. For agile servers equipped with enhanced sleep mode capabilities, *t* may be in the order of *ms*, whilst for legacy equipment a complete bootstrap procedure will be required and *t* may grow up to some minutes (see Table I). In general, the higher the *t* value, the higher the *d*, and thus the lower the energy saving margin, while with low values of *t*, greater energy savings are possible. In Table I we reported the (software and hardware) turn off and wake-up times measured in the INFN[1] Tier2 Grid Site of the CERN[2] LHC[3] experiments. Legacy servers, not equipped with the sleep mode, need several tens or even hundreds of seconds to switch state. Such high times indicate that the enhanced sleep mode feature is strongly advised and may bring great benefits in terms of energy savings, as the results in Section IV confirms.

TABLE I. COMPLETE TURN ON/OFF TIMES (SECONDS) FOR DIFFERENT DEVICES.

Server type	*Power on (hardware)*	*Power off (software)*	*Power off (hardware)*
Computing Element (CE)	120	20	5
Storage Element (SE)	180	10	5
Home Location Register (HLR)	120	60	5
Pizzabox form factor Servers	120	10	5
Blade Servers (Dell® DRAC)	160	45	45
Storage Server (IBM® DS400 Storage System)	60	10	10

B. Energy Savings Potential

In order to evaluate the maximum potentialities of our energy saving approach, we consider instantaneous transitions among the sleep and the active states ($t = 0$) and theoretical sinusoidal traffic, like the one depicted in Fig. 3. The demand curve represents the traffic load during the day, while the service curve represents the servers that must be

[1] Italian National Institute for Nuclear Physics, Naples, Italy.

[2] European Organization for Nuclear Research, Geneve, France-Switzerland.

[3] Large Hadron Collider.

active to process the job requests. Without any energy saving management, the power consumption of the data center stays constant [15], and the energy is the integral of power over time:

$$\int_{t_1}^{t_2} p(t)dt \qquad (1)$$

where $p(t)$ is the power consumption function and t_1 and t_2 are the considered time extremes. Ideally, the lower bound for the data center energy consumption is given by:

$$\int_{t_1}^{t_2} l(t)dt \qquad (2)$$

where the $l(t)$ function describes the load curve. EnergyFarm approximates such curve with the service curve $s(t)$, which is the step function that establishes the minimum set of resources that have to stay powered on to serve the current demand. Therefore, with our energy saving schema the theoretical energy consumption is given by:

$$\int_{t_1}^{t_2} s(t)dt \,. \qquad (3)$$

Clearly, it holds that (2) < (3) << (1), and the bigger the difference between (1) and (3) the greater the energy saving. Theoretically, the energy saving is upper-bounded by:

$$\int_{t_1}^{t_2} (p(t) - l(t))dt \qquad (4)$$

while the actual energy saving is given by:

$$\sum_{i=1}^{n} (p(i) - s(i)) \cdot \Delta i \qquad (5)$$

where n is the number of intervals in which the time interval $[t_1, t_2]$ is divided and Δi is the duration of the i-th time interval; note that n sets the time-basis on which the EnergyFarm is executed. Therefore, eq. (5) represents the energy saving of our EnergyFarm approach which will be evaluated in detail in Section 6.

C. *The service-demand matching algorithm*

Given a demand curve, the EnergyFarm service-demand matching algorithm determines the service curve that satisfies the demand while limiting the number of active server and, thus, the power consumption. As an example, let's consider the scenario depicted in Fig. 4 and Fig. 5.

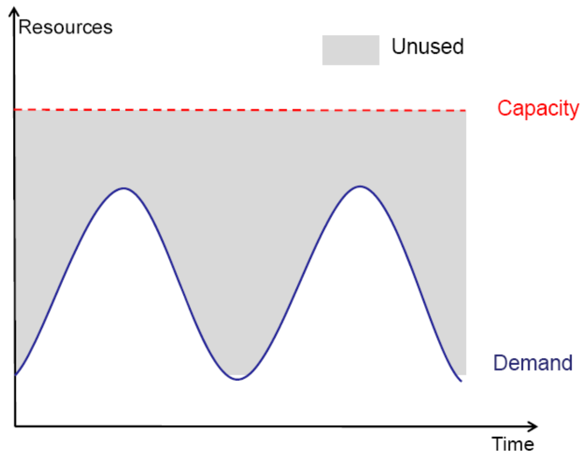


Figure 1. Capacity-demand mismatch leads to resource and energy wastes.

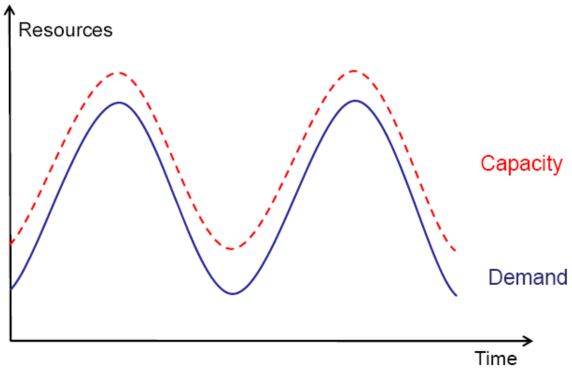


Figure 2. Theoretical provisioning elasticity concept.

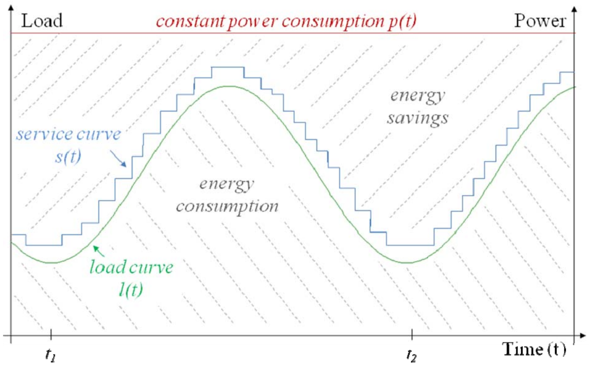


Figure 3. Service-demand matching.

In Fig. 4 the demand curve increases between t_i and t_{i+1}, consequently the distance from the service curve decreases from d_i to d_{i+1}. Since $d_{i+1} < d$, the algorithm detects the increase in the demand (totally absorbed by the guard band d, thus no delay is added in this case) and consequently increases the number of active servers by turning on $s_{i+1} - s_i$ servers.

Figure 4. Incrementing the service curve in response to demand increase.

The opposite situation is depicted in Fig. 5, where a decrement $d_{i+1} > d$ in the demand curve causes the algorithm to decrease the service curve from s_{i+1} to s_i.

Figure 5. Decrementing the service curve in response to demand decrease.

IV. Performance Analysis Results

We evaluated the performances of EnergyFarm through simulations against several available data referring to two different-sized data centers. To model a large data center we used the Google farm [7][8] composed by more than 5,000 blade servers monitored over a six-month period; for the small data center we used the Naples LHC Tier2 Grid site of the INFN section [9] composed by more than 100 servers in the pizzabox form factor. First, we evaluated the impact of the safety margin *d* against the potential savings, in terms of energy (MWh), emissions (tons of CO_2) and economical cost (k€). For a commercial/industrial facility like a data center, the average cost of energy is € 0.12 per kWh [17]. We considered fossil-fueled energy plants powering the data centers, which emits 890 grams of CO_2 per kWh (ACV-DRD study [1]). Several simulations have been conducted for different values of the safety margin *d*. Results show that, for the small data center, the maximum cost savings is more than 35 k€ per year, while for the large data center the cost saving may reach € 1.5 millions, with a reduction of more than 13 GWh in the energy consumption and more than 11 kTons of CO_2 in the emissions (see Table II). Such results should not surprise: servers are rarely utilized at their full capacity and most of the time operate at between 10% and 50% of their maximum utilization levels [7]. As expected, the *d* value affects the energy savings and the consequent CO_2 emissions and bill costs. Best results have been achieved with low values of the safety margin.

TABLE II. Per Year Savings With EnergyFarm (Single-Core Servers) And Variable Safety Margins *d*.

Safety margin	***Energy (MWh)***	***Emissions (Tons of CO2)***	***Cost (k€)***
Small data center			
d = 1%	299.2	266.2	35.9
d = 10%	259.9	231.3	31.2
d = 50%	92.2	82.1	11.1
Large data center			
d = 1%	13184.9	11735.0	1582.2
d = 10%	11455.3	10195.3	1374.6
d = 50%	4065.3	3618.1	487.8

Note that, since our goal is to provide a lower bound for the energy savings of the modern and future data center, the transition time *t* between the on and off states have been put to 0, thus there is no delay in the powering on/off the servers (i.e. all agile servers). As a consequence, the frequency of the load variations (i.e., how and how often the traffic load varies in time) only affects the number of transitions between on/off states, but it does not influence the energy savings at all, as each variation is immediately followed by the corresponding on/off action on the involved servers. In our tests, the efficiency in the resource utilization has reached similar values for the small and the large data centers, varying from 20% to 68%, meaning that a good percentage of the servers has been put into sleep mode for considerable time (see Table III).

TABLE III. EnergyFarm Efficiency In The Resource Utilization.

Average resource utilization in EnergyFarm (%)		
d = 1%	***d = 10%***	***d = 50%***
68,0282	59,1045	20,975

The EnergyFarm saving margins decrease almost linearly as the *d* values increases (Fig. 12). In fact, while the load is far from the actual data center capacity, savings and *d* vary linearly but, as load approaches higher values, the threshold *d* will exclude a higher number of devices from being turned off, leading to relatively lower savings. When considering multi-core devices, job aggregation is possible. Two aggregation strategies have been studied: first-fit and best-fit. In our tests first-fit has always performed worse than best-fit (up to 50%), so here we only focus on the best-fit strategy. Results, both for the small and large data centers, show a common behavior, even if with different rates: the more the cores in the data center, the more the energy consumption. This is due to the fact that the data centers work far from their actual maximum utilization capacity and causes multi-core server to operate with only few jobs even with the best-fit strategy, i.e. multi-core servers present internal fragmentation (not all the cores are always busy).

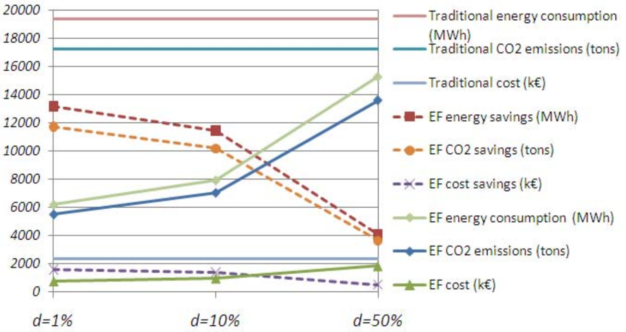


Figure 6. Energy, CO_2 and costs with varying *d* values (large data center).

Thus, at low loads, assigning one job to a single-core server costs less than executing it on a 8-core server (due to the greater energy consumption of the latter), whilst, at higher loads, the greater computing density of multi-core servers may be exploited by the best-fit strategy to lower the overall data center energy consumption.

TABLE IV. ENERGYFARM PERFORMANCES WITH VARYING NUMBER OF CORES PER SERVERS ($d = 1\%$).

Cores per server	*1*	*2*	*4*	*8*
Aggregation	No	Best-fit		
Small Data Center				
Energy (MWh)	see Table II	138.36	142.79	152.05
CO2 (Tons)		123.14	127.08	135.32
Cost (k€)		16.60	17.13	18.25
Large Data Center				
Energy (MWh)	see Table II	6003.57	6007.36	6015.16
CO2 (Tons)		5343.18	5346.55	5353.49
Cost (k€)		720.43	720.88	721.82

V. CONCLUSIONS

In this work, we presented EnergyFarm, an energy manager which can be used on the modern and future grid/cloud data center infrastructures to save energy. Current farms are usually over-provisioned and fluctuations in the traffic load are observed at various time periods. To take advantage of such a situation, we developed EnergyFarm which, through the service-demand matching algorithm and the job aggregation capabilities, allows turning off idle servers, while respecting both the demand requirements and the logical and physical dependencies. Results showed that great efficiency in the resource allocation can be achieved (between 20% and 68%), allowing significant energy, cost and emissions savings. In the optic of the future ICT developments, EnergyFarm may become an indispensable instrument towards sustainable society growth and prosperity.

ACKNOWLEDGMENTS

This work was in part supported by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.

REFERENCES

[1] BONE project, “WP 21 Topical Project Green Optical Networks: Report on year 1 and updated plan for activities”, NoE , FP7-ICT-2007-1 216863 BONE project, Dec. 2009.

[2] J. Koomey, “Estimating Total Power Consumption by Servers in the U.S. and the World”, February 2007, http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf.

[3] Jordi Torres, “Green Computing: the next wave in computing”, Ed. UPCommons, Technical University of Catalonia (UPC), February 2010, Ref. http://hdl.handle.net/2099.3/33669.

[4] Peter Kogge, “The tops in flops”, pp. 49-54, IEEE Spectrum, Feb. 2011.

[5] The Green Grid, “The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE”, Technical Committee White Paper, 2008.

[6] W. Vereecken, W. Van Heddeghem, D. Colle, M. Pickavet, P. Demeester, “Overall ICT footprint and green communication technologies”, in Proc. of ISCCSP 2010, Limassol, Cyprus, Mar. 2010.

[7] L.A. Barroso, L. A., Hölzle, U., “The Case for Energy-Proportional Computing”, IEEE Computer, vol. 40, 33-37, 2007.

[8] X. Fan, W.-D. Weber, and L.A. Barroso, “Power Provisioning for a Warehouse-Sized Computer”, http://research.google.com/archive/power_provisioning.pdf.

[9] A. Doria, G. Carlino, S. Iengo, L. Merola, S. Ricciardi, M. C. Staffa, “PowerFarm: a power and emergency management thread-based software tool for the ATLAS Napoli Tier2”, proceedings of Computing in High Energy Physics (CHEP) 21 - 27 March 2009, Prague, Czech Republic.

[10] W. Van Heddeghem, W. Vereecken, M. Pickavet, P. Demeester, “Energy in ICT - Trends and Research Directions”, in Proc. IEEE ANTS 2009, New Delhi, India, Dec. 2010.

[11] B. St Arnaud, “ICT and Global Warming: Opportunities for Innovation and Economic Growth”, http://docs.google.com/Doc?id=dgbgjrct_2767dxpbdvcf.

[12] K. Ley, R. Bianchiniy, M. Martonosiz, T. D. Nguyen, “Cost- and Energy-Aware Load Distribution Across Data Centers”, SOSP Workshop on Power Aware Computing and Systems (HotPower '09), Big Sky Montana (USA), 2009.

[13] M. Gupta, S. Singh, “Greening of the internet”, in Proc. of the ACM SIGCOMM, Karlsruhe, Germany, 2003.

[14] K. Christensen and B. Nordman, “Reducing the energy consumption of networked devices”, in IEEE 802.3 tutorial, 2005.

[15] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, N. McKeown, “Elastictree: Saving energy in data center networks”, in Proceedings of the 7th USENIX Symposium on Networked System Design and Implementation (NSDI), pages 249--264. ACM, 2010.

[16] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, “Above the Clouds: A Berkeley View of Cloud computing”, Technical Report No. UCB/EECS-2009-28, University of California at Berkley, USA, Feb. 10, 2009.

[17] U.S. Energy Information Administration, State Electricity Prices, 2006, http://www.eia.doe.gov/neic/rankings/stateelectricityprice.htm.

2011 International Conference on Broadband and Wireless Computing, Communication and Applications

Evaluating Network-Based DoS Attacks Under the Energy Consumption Perspective

New security issues in the coming green ICT area

Francesco Palmieri
Dipartimento di Ingegneria
dell'Informazione
Seconda Università di Napoli
Aversa, Italy
francesco.palmieri@unina.it

Sergio Ricciardi
Departament d'Arquitectura
de Computadors
Universitat Politècnica de Catalunya
Barcelona, Spain
sergior@ac.upc.edu

Ugo Fiore
Centro Servizi Informativi
Università degli Studi
di Napoli Federico II
Naples, Italy
ufiore@unina.it

Abstract *— In the green Information and Communication Society (ICS), new form of Denial of Service (DoS) attacks may be put in place: exploiting the computational and storage resources of datacenters with the aim of consuming as much energy as possible, causing detrimental effects, from high costs in the energy bill, to penalization for exceeding the agreed quantity of CO_2 emissions, up to complete denial of service due to power outages. To the best of our knowledge, this is the first paper which investigates the impacts of network-based DoS attacks under the energy consumption perspective. We analyzed different types of such attacks with their impacts on the energy consumption, and showed that current energy-aware technologies may provide attackers with great opportunities for raising the target facility energy consumption and consequently its green house gases (GHG) emissions and costs.*

Keywords: denial of services DoS; energy consumption; networking; datacenters; green house gasses emissions GHG.

I. Introduction

In the last years, the emergence of the new green, energy-sustainable computing paradigms has gained a lot of attention in both the research and industrial arenas. Consequently, the development of modern ICT architectures with the additional requirement of keeping the energy consumption under control while maintaining the services offered at a satisfactory level, according to the new concepts of energy-efficiency and energy-awareness, has become central. By observing the electrical power demands of the largest networked computing farms, such those empowering the modern distributed computing and cloud infrastructures, it has been estimated [1][2] that ICT worldwide energy consumption amounts to more than 8% of the global electricity production and the energy requirements of datacenters, storage and network equipment are foreseen to grow by 12% per year. Clearly, such a huge electricity demand will result in environmental and engineering issues and bottlenecks, seriously conditioning the evolution of the whole ICT sector. For example, we can consider that the number of transistors integrated within the recent Intel Itanium processors reaches to nearly 1 billion of elements. If this growth rate continues, the heat (per square centimeter) produced by next-generation CPUs would exceed that of the sun's surface [3], by reaching a critical technological limit and energy demand threshold. Furthermore, together with the growth of the energy required by the above infrastructures, there is an alarming rise in their correct usage involving thousands of concurrent e-commerce transactions and millions of Web queries per day, handled through large-scale distributed datacenters, which consolidate hundreds or thousands of servers with other auxiliary systems such as cooling, storage and network communication ones. In this scenario, there has been an equally dramatic evolution in security. The need for efficient ways of detecting and attempting to prevent intrusions, as well as of mitigating attacks, has led to the elaboration of sophisticated analysis techniques and countermeasures. This has brought a corresponding advance in the cleverness of attack strategies and tools, also affecting the attack objectives that can become different from the traditional ones (confidentiality, integrity, availability or performance of the computing elements offering the service). The developments in the areas of energy-awareness/efficiency and network/site security have been considerable but separate. This paper underlines that there are areas in common between these two fields, and addresses a new perspective, which might become commonplace over the next years: attacks could change in their main aims, either exploiting weaknesses in power-saving and management mechanisms to disrupt services, or even attempting to increase the energy consumption of an entire farm, by causing financial damages. Therefore, it becomes clear that the energy-efficiency and security challenges can be better addressed in a combined way if the energy requirements and the bottlenecks of the underlying security technologies and protocols are better understood and coped accordingly. For this sake, we evaluated the efficiency of common attacks with respect to their troublemaking potential in terms of the impact on the energy consumption of the target infrastructure. To the best of our knowledge, this is the first paper that evaluates the DoS attacks under the energy consumption perspective.

II. Energy-Oriented Denial Of Service Attacks

Denial of Service attacks are becoming a more and more important disturbance factor on all the sites that are connected to the Internet. Any defense against these menaces is very difficult because they strive at consuming all the

978-0-7695-4532-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BWCCA.2011.66

available resources at both the computing service and network transport layers, where it is very hard to distinguish whether an access or service request is genuine or malicious. By affecting the server systems or the network connection on the target sites, the attacker may be able to prevent any access to e-mail relays, websites, online accounts (banking, e-commerce, etc.) or other services that rely on them [4]. Attacks to the network connection take place by exhausting the available bandwidth through the generation of a very a large number of packets directed to the target site. Typically, these packets are ICMP ECHO packets but in principle they may be anything [5]. On the other hand, the computing resources on the service nodes within the target site can be saturated by overwhelming them with a huge quantity of CPU intensive service requests, such repeated transaction attempts on an HTTPS or any kind of SSL-empowered server. In order to increase the attack power, many remotely controlled computers can be simultaneously used as the source. This kind of menace is also widely known as Distributed Denial of Service (DDoS) attack. We advise the possibility that the above menaces can be made more effective, by introducing new and more subtle objectives and attack scenarios, based on power management and energy-efficiency considerations. A fundamental property of such kind of attacks is that they exploit the hardware components and subsystems that experience the heaviest difference in power consumption between their busy and idle/sleeping states. The attacks perform their offending activity by keeping the target component as busy as possible, and preventing it from going into low power usage modes, and thus forcing it to work at its near-maximum speed, frequency, voltage or temperature. In modern ICT infrastructures, the most critical components from the power consumption perspective are the server systems, whose energy demand is tightly related to their load: a fully loaded server absorbs about two times the power of an idle one, with a linear increment of the power consumption with respect to the server load. The CPU alone contribution to the server power consumption goes from 25% to 55%, depending on the server, followed by memory and networks interfaces [6][7]; disks, motherboard and fans consume less energy (Table 1). Anyway, it should be pointed out that, in order to assess the potential of an energy-oriented attack, it is not a priority to focus on the major power hungry device, but rather on the most energy sensible devices, i.e. components whose energy consumption strongly varies with the traffic load. These components are mainly CPU, disks, network interface cards (NIC); in the Section III we focus on such components and analyze their impact on the energy consumption. The first and most critical component that exhibits these operating characteristics is the CPU/Memory subsystem whose energy consumption is known to scale linearly with its utilization [6][8]. Since the goal of such energy-oriented attacks is to maximize the power consumption by keeping the CPU and memory on the target systems as busy as possible, they try to add additional load on the servers by introducing a large number of service request which subtract most of the resources to the legitimate ones and let the CPUs working at their maximum operating frequency. This can be achieved by overwhelming the CPU/Memory runtime subsystem with fake SSH or SSL/TLS-based transactions and service requests or forcing the continuous execution of a huge number of random read and write operations on very large arrays located in memory to generate a large quantity of cache misses.

TABLE I. ENERGY CONSUMPTION BREAKDOWN OF A LOW-END SERVER

Component	*Peak Power*
CPU [19]	80 W
Memory [20]	36 W
Disk subsystem [21]	12 W
Network Interface [22]	2 W
Motherboard [6]	25 W
Fans [6]	10 W

CPU and memory dominate power absorption; disks are relevant only if there are many.

Another effective way of draining more and more system energy is overloading the device's hard disks with millions of read or write operations by forcing them to constantly operate at their maximum sustained transfer rate or to continuously spin up and down the hard disks spindle engines. This kind of attack is very common in the offending strategies of several computer viruses and Trojans that are typically able to directly run malicious codes on the target nodes. In the worst cases, the malicious agents can alter the operating system kernel or some application binary code so that more energy is needed for their execution. However, the binaries altered in such a way may or not continue to behave correctly from the users' point of view. Finally, the last device/component that can be solicited is the network interface, when its energy consumption depends on the actual connection rate, that is, in implementations supporting adaptive link rate technologies (ALR), low power idle (LPI) and dynamic voltage scaling (DVS) mechanisms. Clearly, the disruption of these attack schemes is dependent on how much power the device consumes in maximum speed mode respect to the one required in lower power modes; such a gap may be as high as 90% between idle and full load states for higher speed interfaces [9]. Perhaps unexpectedly, security systems themselves also offer a wealth of opportunities for energy-oriented attacks. Firstly, it must be remarked that security systems, while being essential to the correct operations of networked systems, also have an impact on the power expenditure. Such security systems strive to monitor the behavior of the device under control as non-obtrusively as possible. However, they consume a not negligible amount of energy [10]. For example, keeping in mind that, (a) in reasonably well-managed organizations, end-user PCs are ordinarily equipped with properly configured and updated antivirus software, and (b) this software will scan on-the-fly some or all the content trying to reach the computer local storage, and (c) there are conditions under which the antiviral scanning causes long periods of full CPU load, a disruptive energy-aware attack can be orchestrated in the following way. Firstly, the attacker selects a (ideally innocent) content, which will trigger the antivirus reaction, consuming a great amount of CPU in the process. Secondly, the attacker sets up a Web site with a content appealing to the individuals belonging to the target organization (alternatively, he/she may set up a spam campaign) and have the malicious content delivered to the target. Spam, that is a nuisance to email users, since it eats up remarkable resources spent to deal with it and prevents it from reaching user mailboxes, can also be exploited for energy-oriented attacks. Spam messages are usually cheap to send, because they are normally originated from compromised computers belonging to a botnet. Mail servers run anti-spam software, which has the purpose of identifying and filtering out unwanted messages, and this software consumes CPU, disk and network resources [11]. An energy-oriented attack could then increase the footprint on a target mail server by simply increasing the amount of spam addressed to it. In any case, a successful attack will maximize power consumption and excessively solicit hardware components while presenting to the user the appearance that the system is behaving normally, with the possible exception of an increased CPU, disk or network activity. Some of the side effects that one would expect to observe in presence of these menaces, if they are not implemented more subtly, include the legitimate user requests being served slowly, the CPU fan turning on while the user is performing some action that does not normally cause the fan to come on, the system becoming less interactive than usual, the network loosing part of its speed/responsiveness and the hard drive spinning up immediately after a spin down.

A. *Affecting the energy costs*

Incrementing the power usage has direct and immediate consequences on the energy expenses. If well designed, attacks may exploit the different energy costs (e.g. during the the night-day cycle) or the energy budget threshold that the facility agreed with the power supplier, resulting in very high energy bills. That's worse, traditional power provisioning strategies, aiming at keeping as much computing and storage equipment as possible within a given power budget in order to maximize the utilization of the deployed datacenter power capacity, may present the drawback of offering more subtle vulnerabilities to possible attackers. More precisely, such strategies try to fill the gap between achieved and theoretical peak power usage in order to deploy additional equipment within the power budget [6]; the full utilization of the datacenter is offset by the risk of exceeding its maximum capacity, resulting in power outages or costly SLA violations due to the fact that the maximum drained power of a datacenter may be conditioned by a physical and/or contractual limit. The contractual enforcement exceeding will result in economic penalties (that can be exploited by a malicious competitor), or even overcoming the physical power limits resulting in power outages.

B. *Neutralizing energy saving systems*

If attackers know that some energy-saving mechanisms operate in the target system/network, and if they know the details about these system, they can devise attacks aimed at neutralizing them. This is a subtle issue, because the amount of extra work to be "injected" into the system does not need to bring the processor or storage to full load, but is limited to the amount necessary to avoid the triggering of the energy-saving mechanisms, which are, in general, threshold-based. This means that detecting such attacks can be significantly harder. Furthermore, energy saving techniques are the more vulnerable to energy-oriented attacks, since they offer to the attackers greater opportunities to rise the energy consumption. It is quite common, in fact, that an infrastructure, like a datacenter, buys a given amount of energy to be used into an agreed period of time, according to the mean energy consumption of the site; exceeding such threshold may result in additional costs. An attacker may exploit such situation by raising the computational needs of the site and, thus, its energy consumption, above the threshold, therefore causing an economical damage or, even worse, an energy outage resulting in a complete denial of service. A very simple example of the above concepts can be observed when per-server sleep mode is deployed in the datacenters. An attack that simply generates continuous fake demands/traffic for all the servers may prevent machines to go into sleep mode during low load periods, thus having large impacts on the medium and long-term power consumption and hence conditioning the overall energy containment strategy.

C. *Incrementing the operating temperature*

Even if harder to put into practice, since it requires attacker to gain access to computing resources, thermal-based attacks are another potential menace that has to be taken into account. Such offensive strategies aim at executing a particular piece of code whose objective is not to saturate the computing or storage resources, but instead to subtly execute a relatively small cycle-loop that heats the CPU and the memory banks. The current CMOS technology provides modern microprocessors with not only transistors but also capacitors and resistors. Under normal circumstances, the CPU is not always active at 100%, but instead enters and exits from low power periods (HLT machine code instruction) in which the clock is halted and the circuitry enters a suspend mode until an interrupt or reset happens. Also, low power states (C-states) are available in the latest Intel® CPUs. Malicious codes may prevent the CPU to enter such low power states and continuously executes loops that charge resistors, notably increasing the temperature. Current datacenter infrastructures, in fact, have a power usage effectiveness (PUE) of 2, meaning that the heat, ventilation and air conditioning systems (HVAC) consume as much energy as the computing and storage resources. Therefore, the quantity of power absorbed by the HVAC system is not negligible: the potential of thermal-based attacks is as high as the energy-oriented one. The result is that the cooling infrastructure will work harder consuming a considerably higher quantity of energy. Detrimental effects of such attacks include the increase of the CPU and memory temperatures, with the consequent stability problems, reduced component life (an increase by 10°C halves the chip life span), and increased cooling power consumption.

D. *Exhausting the power budget*

As we have seen, attacks may result not only in high energy costs, but, in the worst cases, also in complete power outages. It has been observed that the power consumption declared by manufacturers (nameplate value) is usually an overestimated conservative value [12], and thus it is of limited usefulness when predicting the total power budget of the datacenter, giving the idea that "there will be enough power" if the nameplate values are considered when dimensioning the power facilities. This scenario, together with the periodical updates of new components (additional memory banks, disks, network interface cards, etc.) in an effort to accommodate the growing business demand and the wear of the devices as long as the substitution of old components with newer ones, which are more efficient but also more power-hungry (Moore's law has not been compensated at the same pace by energy efficiency), exposes the datacenter facility to the risk of exceeding its maximum power budget, in particular under energy-oriented attacks. Accordingly, we point out the security related risks of over-subscribing the datacenter under the energy consumption perspective. In fact, a sustained energy-oriented attack may put an entire datacenter out of service by totally blocking the underlying electrical distribution system (by exhausting its capacity). Such kind of attacks may be hard to detect, unless a constant fine-grained on-line monitoring and data-collection systems are deployed directly on the power distribution sub-system (i.e. UPS, PDU, RACKS, etc).

E. *Incrementing dirty emissions*

Energy-oriented attacks may also be exploited under an additional dimension: the green house gases emissions (GHG). Several practices have been adopted by the industry and the governments to reduce the GHG emissions [13]: carbon taxes, cap & trade, and carbon offset are all susceptible of being exploited by attackers to increase the GHG emissions of a facility and thus its costs. In a carbon tax approach, industries pay taxes according to the amount of emitted GHG (mainly CO_2); in this context, an attacker may obtain a double objective: raising both the energy consumption and the costs associated with the increased GHG emissions. In cap & trade containment strategy, a limit (cap) is imposed on the maximum allowed emissions and a market (trade) is created in which additional emission permissions may be bought by virtuous industries that do not reach the cap. In the carbon offset approach, industries are committed to compensate their emissions by buying in "green", such as tree reforestation, etc. Both the cap & trade and the carbon offset policies may attract unsavory practices from organizations that take advantages of third party emissions induced by the aforementioned attacks.

III. Modeling Power consumption in Energy-oriented DoSes

To illustrate the potential of energy-oriented attacks and analyze their dynamics and behaviors, we modeled the additional power consumption associated to each one of them. When exploiting the CPU/Memory subsystem, we consider that a modern CPU dynamically adapts its operating frequency to the current load so that its instantaneous power demand at the frequency f can be estimated as:

$$P(f) = \frac{1}{2} C V_f^2 A f \, . \tag{1}$$

In the above theoretical formulation [8], f can assume values within the range [f_{min}, f_{max}], C (aggregated load capacity) and A (activity factor) are fixed constant parameters (depending on the involved CPU characteristics), and V_f is the CPU voltage scaling linearly with the frequency f, that is:

$$V_f = V_{\max} \frac{f}{f_{\max}} \, , \tag{2}$$

where V_{max} is the maximum operating voltage required at the frequency f_{max}. Since the goal of all the CPU-based attacks is overloading the CPU by forcing it to work at its maximum operating frequency f_{max} for the longest possible time, we can estimate the worst case and best case power demands P_{max} and P_{min} as:

$$P_{\max} = \frac{1}{2} C V_{\max}^2 A f_{\max}, \quad P_{\min} = \frac{1}{2} C \left(V_{\max} \frac{f_{\min}}{f_{\max}} \right)^2 A f_{\min} \cdot \tag{3}$$

Consequently, if we consider that the average server utilization of datacenters is very low, often below 30% of its CPU capacity [6][14][15], we can assume that the average CPU power consumption approximates to P_{min} and hence the additional energy consumption introduced by a CPU based DoS attack can be estimated as:

$$E_C = (P_{\max} - P_{\min}) t_d = \frac{1}{2} C V_{\max}^2 A \left(\frac{f_{\max}^3 - f_{\min}^3}{f_{\max}^2} \right) t_d, \tag{4}$$

where t_d is the duration of the attack. Thus, the energy increase is proportional to the difference of the cubes of the maximum and minimum frequencies, and depends only linearly on the attack duration. This means that attack intensity is more critical than attack duration. Many bursty, strong attacks can achieve the same objective as one single sustained attack with lower intensity and, while the latter may be harder to detect but easier to prevent, the former will be easier to detect but harder to prevent. Analogously, the additional energy demand E_D for a typical attack based on repeated disk operations can be calculated by referring to the involved transfer rate r and considering the maximum sustainable drive transfer rate r_{max} as a worst case metric to calculate the amount of time spent in read mode when transferring data at speed r. We focus on attacks based on read operations since large-block-size reads consume more energy than writes (approximately P_{read}=13.3$\mu W/Kbyte$ against P_{write}=6.67$\mu W/Kbyte$ [16]), and the fact that reads occur 4-5 times more than writes becomes significantly important when considering that read operations may be used much more easily also on a partially compromised host. Let P_D be the power required by a disk (read) operation, as sum of engine-dependent mechanical power consumption [17], with the operation-dependent (read-write) electronic power consumption:

$$P_D = \frac{K^2 \omega^2}{R} + D_r w_r P_{read}\,, \qquad (5)$$

where K is a motor voltage constant, R is the motor resistance, D_r is the (read) demand (*kByte*) and w_r is a weighting factor depending on the current rate r [18]:

$$w_r = \frac{r}{r_{\max}}\,. \qquad (6)$$

The first factor of eq. (5), referring to mechanical movement, quadratically depends on the angular velocity ω, and is the most critical part from the energy consumption perspective: an attack generating randomly sparse and bursty block read operations, forcing the disk hardware to continuously spinning up immediately after a spin down, can introduce a near maximum burden to the overall disk energy demand. Then, by considering that the energy required by the drive in low power P_l is already included in the default system power consumption, we can argue that the additional energy required during a disk-based DoS attack is upper-bounded by the above activities that can be expressed by:

$$E_D = (P_r - P_l) \cdot t_d\,, \qquad (7)$$

where t_d is the duration of the attack. Note that eq. (7) is a function of the involved transfer rate r, so that the higher the sustained transfer rate during the attack is, the greater the impact on the overall power consumption will be. Finally, also if a more limited quantity of energy is required for the network interface, in presence of modern NICs supporting dynamic link rate adaptation or low power idle mechanisms (P_{min}), and hence reducing their speed and energy requirements in case of limited or no traffic, significant increments in power usage can be achieved by forcing the interfaces to work at their maximum throughput by flooding the target hosts with typical DDoS-generated traffic. Also in this case, if P_{max} is the power demand in active/maximum speed mode, the additional energy absorption introduced by an attack of duration t_d can be expressed by:

$$E_N = (P_{\max} - P_{\text{avg}}) \cdot t_d\,, \qquad (8)$$

where P_{avg} is the average power consumption of the NICs with normal traffic load during the time interval t_d.

IV. Tests Results and Discussion

Using the power consumption model of Section III, we estimated the potential of the CPU-based attack of eq. (4), for the reference processor AMD® Athlon® 2,4 GHz. Following the approach used in [8], we imposed a lower bound for $f_{min} = f_{max} / 2.4$ to prevent asymptotic trend during low utilization periods, where the constant 2.4 and the V_{max} = 1.4 V values are based on the frequency range and maximum voltage provided by the specification sheet of the reference CPU. As we can see from Fig. 1, the CPU-bound has great potentiality to exploit the energy consumption with a maximum intensity attack. The energy consumption surplus reached during the maximum intensity is up to 13.8 times the minimum energy consumption under low utilization periods and up to 4.1 times the energy consumption with medium load (absolute values scaled by constant factor $K = AC$).

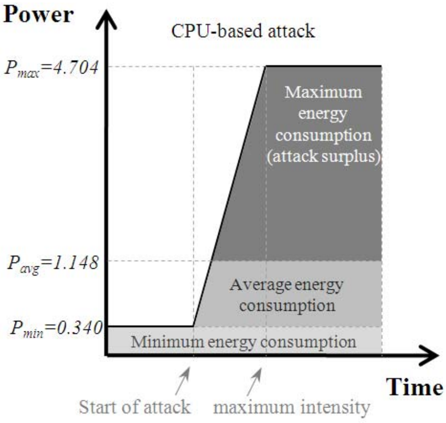


Figure 1. Power consumption upper bound for aCPU-based attack.

Power consumptions of the I/O-based attack are plotted in Fig. 2. The maximum read rate has been assumed in all the cases (i.e., r=r_{max}) and the variation of the power consumption has been reported for different values of the angular velocity $\omega \in$ {5400, 7200, 10000, 15000} *rpm* (absolute values scaled by constant factors K_1=K^2/R, K_2=$D_r P_{read}$). As we can see, the angular velocity strongly influences the disk power consumption, with the highest intensity attack that may reach peaks of 7.6 times the low power consumption mode.

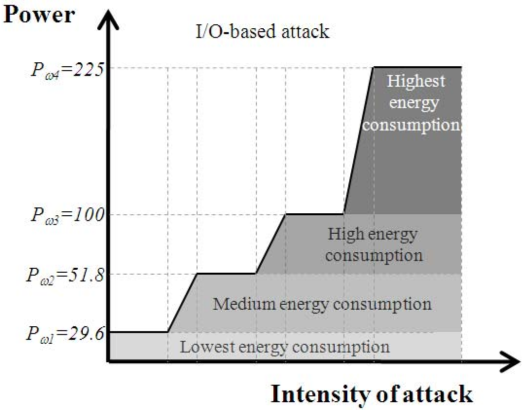


Figure 2. Power consumption upper bond for anI/O-based attack.

The CPU-bound based attack achieves the higher power consumption, while the I/O-bound one is less sensible to the power consumption, even if the latter may slow down the datacenter responsiveness even more than the former. The energy consumption potentiality of an energy-oriented attack on the NICs can be found in [23]. Nevertheless, the potential of the attacks should be contrasted also with the difficulties of being deployed. CPU and I/O-bound attacks are much easier to commit, since a DDOS may easily generate a huge number of web searches or mail requests, whilst the offline job processing requires access to back office facilities, like in the service as a service (SaaS) cloud computing paradigm.

V. Conclusions

DDoS attacks have the potential not only of denying the service of the target facility, but may be carved to explicitly impact its energy consumption. Such attacks may be targeted at several objectives: increment the energy consumption, the GHG emissions and introducing, in the worst cases, power outages. Some of these attacks are relatively easy to perform, e.g. CPU and I/O-bound based ones, whilst others are more difficult to deploy. In any case, the potential of such attacks should not be underestimated. Effective power management techniques have to be deployed to prevent detrimental effects. The most effective technique is the power capping scheme that set a maximum power consumption threshold and operate the facility always below that value. A power monitoring system constantly monitors the power absorption, and if an increment is detected, takes the corresponding actions to decrease the power, from job de-scheduling/ migrating to using any available component-level strategy to decrease the energy consumption, e.g. CPU voltage/ frequency scaling (DVS), downclocking devices, forcing sleep mode, etc., i.e. implementing an energy proportional computing system which has proved to be an effective way to reduce peak power usage. Anyway, it should be pointed out that power capping alone, although is an immediate measure to prevent facility detrimental, is not enough to detect attacks. Network based DoS attacks have to be recognized and isolated from the allowed traffic through a comprehensive security system.

Acknowledgment

This work was supported in part by the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.

References

[1] BONE project, 2009, “WP 21 Topical Project Green Optical Networks: Report on year 1 and updated plan for activities”, NoE , FP7-ICT-2007-1 216863 BONE project, Dec. 2009.

[2] J. G. Koomey, “Estimating total power consumption by servers in the U.S. and the world”, Lawrence Berkeley National Laboratory, Stanford University, 2007.

[3] G. Koch, “Discovering multi-core: Extending the benefits of Moore’s law”, Technology@Intel Magazine, 2005, http://www.intel.com/technology/magazine/computing/multi-core-0705.pdf.

[4] M. McDowell, “Understanding Denial-of-Service Attacks”, National Cyber Alert System, Cyber Security Tip ST04-015.2004, 2004.

[5] Denial of Service Attacks, 1999. Online, [2010.11.10] http://www.cert.org/tech_tips/denial_of_service .html.

[6] X. Fan, X-D. Weber, L.A. Barroso, “Power provisioning for a warehouse-sized computer”, in Proc. 34th annual international symposium on computer architecture (ISCA ’07), pp 13–23, 2007.

[7] L.A. Barroso, U. Hölzle, “The Case for Energy-Proportional Computing", IEEE Computer, vol. 40, pp. 33-37, 2007.

[8] D. Meisner, B.T. Gold, T.F., Wenisch, “PowerNap: eliminating server idle power”, In Proc. of ASPLOS ’09, pp 205–216, 2009.

[9] K. Christensen, P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi, J. A. Maestro, “IEEE 802.3az: The Road to Energy Efficient Ethernet”, *IEEE Comm. Magazine*, Nov. 2010.

[10] J. Bickford, H. A. Lagar-Cavilla, A. Varshavsky, V. Ganapathy, L. Iftode, “Security versus Energy Tradeoffs in Host-Based Mobile Malware Detection”, MobySis 11, Bethesda, Maryland, USA, 2011.

[11] McAfee and ICF International, “The Carbon Footprint of Email Spam Report”, 2009.

[12] J. Mitchell-Jackson, J. G. Koomey, B. Nordman, and M. Blazek, “Data center power requirements: measurements from silicon valley”, Energy ISSN 0360-5442, 837–850, 2003.

[13] B. St Arnaud, “ICT and Global Warming: Opportunities for Innovation and Economic Growth”, http://docs.google.com/Doc?id=dgbgjrct_2767dxpbdvcf.

[14] C. Bash and G. Forman, “Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations in the data center,” in Proc. of the 2007 USENIX Annual Technical Conference, 2007.

[15] P. Bohrer, E. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, and R. Rajamony, “The case for power management in web servers,” Power Aware Computing, 2002.

[16] A. Lewis, S. Ghosh, and N.-F. Tzeng, “Run-time energy consumption estimation based on workload in server systems. In HotPower”, 2008.

[17] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, “Reducing Disk Power Consumption in Servers with DRPM”, Computer 36, 12, pp. 59-66, 2003.

[18] W. West, E. Agu, “Experimental Evaluation of Energy-Based Denial-of Service Attacks in Wireless Networks”, IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.6, Jun. 2007.L.A. Barroso, U. Hölzle,, “The Case for Energy-Proportional Computing", IEEE Computer, vol. 40, pp. 33-37, 2007.

[19] Intel Corporation, “Intel Xeon Processor with 512KB L2 Cache at 1.80 GHz to 3 GHz Datasheet”, http://download.intel.com/design/Xeon/datashts/29864206.pdf, Mar. 2003.

[20] Micron Technology Inc., “Calculating Memory System Power for DDR”, http://download.micron.com/pdf/technotes/ddr/TN4603.pdf, 2001.

[21] Seagate Technology LLC. Product manual Barracuda 7200.7. http://www.seagate.com/support/disc/manuals/ata/cuda7200pm.pdf, Sep. 2005.

[22] R. Sohan, A. Rice, A. W. Moore, K. Mansley, “Characterizing 10 Gbps network interface energy consumption”, LCN 2010, 268-271, 2010.

[23] S. Ricciardi, D. Careglio, U. Fiore, F. Palmieri, G. Santos-Boada, J. Solé-Pareta, “Analyzing Local Strategies for Energy-Efficient Networking”, in Proc. of SUNSET 2011, IFIP NETWORKING, Valencia, Spain, LNCS 6827, pp. 291-300, 2011.

Towards Energy-Oriented Telecommunication Networks

Sergio Ricciardi[a], Francesco Palmieri[b], Ugo Fiore[c], Davide Careglio[a], Germán Santos-Boada[a], Josep Solé-Pareta[a]

[a] Technical University of Catalonia - BarcelonaTech
Department of Computer Architecture
C. Jordi Girona 1-3, E-08034, Barcelona, Spain, sergior@ac.upc.edu
[b] Second University of Naples
Department of Information Engineering
V. Roma 29, I-81031, Aversa (CE), Italy, fpalmier@unina.it
[c] University of Naples Federico II
Centre of Computer Science Services
V. Cinthia, 5, I-80126, Napoli, Italy, ufiore@unina.it

Abstract

Latest developments in Information and Communication Technology (ICT) have led to a remarkable bandwidth increase in telecommunication networks. As result, the energy consumption of network infrastructures is rising both in the metro/backbone and in the access segments. The Internet growth is going to be no longer conditioned by the overall bandwidth demand, but rather by its operational cost and environmental effects, mainly associated to the energy consumption. The increasing interest on environmental problems, as fossil-based fuels are becoming scarcer and renewable energy sources are arising, may play an important role for a greener Internet, with reduced CO_2 emissions, energy consumptions and network operating costs. Energy models are being provided to characterize the energy consumption of the various networking devices under different traffic loads for both optical and electronic network layers. We present several ideas about new emerging paradigms, energy-efficient architectures and energy-aware algorithms and protocols exploiting traffic fluctuations and variations in energy prices to save energy and reduce the operating costs. These ideas can foster the development of new energy-oriented design and operating solutions/models that, if introduced in the next generation networks, will lead to a comprehensive energy optimization framework for the future green telecommunication infrastructures.

Keywords: Green networks, energy costs, energy consumption, GHG emissions.

1. Introduction

In the last years, we have assisted to an uncontrollable growth of the Internet. By 2015 the total Internet traffic will be three times larger than the one observed in 2011, equivalent to a monthly traffic of 60 *Exabytes* of data (1 *Exabyte* $= 10^3$ *Petabytes*) (Figure 1, [1]).

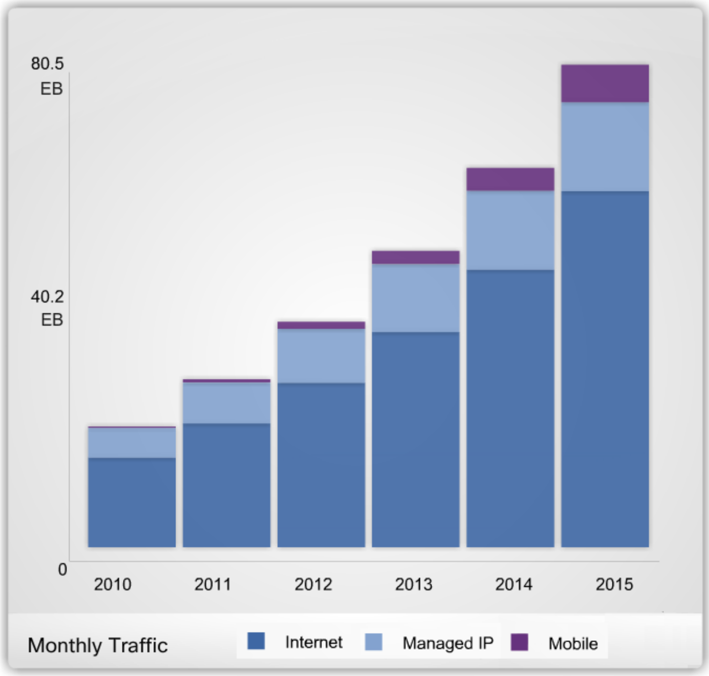


Figure 1: Monthly Internet traffic growth forecast.

The number of user connected to the Internet will pass from $2,1$ billions of 2011 to 3 billions in 2015 [2] [3] [4]. Furthermore, Internet users are asking for higher and higher bandwidth to enable video-on-demand and interactive contents, high quality of service (QoS), online gaming, conferencing tools, voice-over-IP, etc. Fortunately, the advent of optical technologies in the networking arena has been able to provide huge bandwidths for satisfying the increasing traffic demand and avoiding the Internet collapse, besides being characterized by a very low energy consumption when compared to the electronic technologies. Anyway, all-optical networks are still far from being the *only* solution, since they do not provide all the features required by the modern network infrastructures: processing, buffering, monitoring and grooming are at the moment fully provided only by the electronic layers, which in turn do consume a lot of energy. Therefore, the growth of the Information and Communication Society (ICS) is dragging on the energy demand of the whole Information and Communication Technology (ICT) sector to values reaching $7-8\%$ of the worldwide energy production, foreseen to double by 2020 [5]. Such growth rate is not sustainable and proper countermeasures have to be taken to change the business as usual (BAU) scenario to a *greener* one. In order to give an idea of the above numbers, in Italy, like in France, Telecom Italia and France Telecom respectively, are the second largest consumers of electricity after the National Railway systems: 2 TWh per year, and in the UK, British Telecom is the largest single power consumer [6] [7] [4]. In both data centers and networking plants, the energy consumption is further increased by the impact of HVAC system (Heating Ventilation and Air Conditioning), UPS (Uninterruptible Power Supply) systems and lighting facilities. The power usage efficiency (PUE) [8] is a metric used to measure the impact of these system on the overall energy consumption of the site; at the moment, a PUE value of 2 is the standard, meaning that for each Watt of power spent to do the work, another one is spent to keep the site cool and assure the energy supply. Many efforts are being dedicated to lowering the PUE value in large telecommunication or computing plants, ranging from implementing more efficient cooling systems to moving the whole facilities where the external temperature may be used to naturally cool the site.

That's not all; apart from the energy consumption, another problem is menacing not only the ICS growth but the entire world population: climate changes (mainly, global warming) will drastically change the aspect of the world, as we know it, if immediate actions will not be taken to drastically reduce the emissions of green house gases (GHG) in the atmosphere. Even if the ICT is perhaps the only industry sector that does not directly emit GHG during the use phase, the traditional power plants feeding the ICT equipment do emit GHG to generate the required energy. For ICT, the greater shares of GHG are emitted during the use phase (80%), while the construction phase is responsible only for a small percentage (20%) of the emissions.

As a consequence, the reduction of energy consumption and the use of alternative renewable energy sources to limit GHG emissions as possible, are the most urgent emerging challenges for telecommunication carriers, to cope with the ever increasing energy costs, the new rigid environmental standards and compliance rules, and the growing power demand of high-performance networking devices. All the above open problems and issues, foster the introduction of new energy-efficiency constraints and energy-awareness criteria in operation and management of modern large scale communication infrastructures, and specifically in the design and implementation of enhanced energy-conscious control-plane mechanisms to be introduced in next generation transport networks.

Accordingly, this chapter analyzes the available state-of-the-art approaches, the novel research trends and the incoming technological innovations for the new green ICT era and outlines the perspectives towards future energy-oriented network architectures, operation and management paradigms.

1.1. The energy problem

Human activities have severe impacts on the environment. The *human ecological footprint* measures the human demand on the biosphere which, in 2007 was 1,5 planet Earths[1] [9], meaning that we are consuming resources at rate faster than the one characterizing the natural capacity of the Earth to regenerate them. Consequently, an immediate change in our energy consumption habits is needed, if we don't want to exhaust the Earth's natural resources. Therefore, when talking about the energy problem, we usually refer to two concepts: the scarcity (and the consequent higher and higher costs) of the traditional fossil-based fuels like oil, coal and gas, and the effect that burning such fuels has on the biosphere, mainly resource exploitation, pollution and GHG emissions which are responsible for climate changes (especially global warming and global dimming).

The *carbon footprint*, part of the human ecological footprint, measures the total set of GHG emissions of a human activity. The ICT emissions correspond to the $2-3\%$ of the worldwide produced GHGs [5], as much as the ones characterizing the aviation industry, but with a fundamental difference: airplanes burn large quantity of fossil fuels to travel, whilst ICT only needs electrical energy to work, which can be produced by green renewable energy sources, not emitting GHGs during the energy production process

[1] Latest available data; every year this number is recalculated, with a three year lag due to the time it takes for the UN to collect and publish all the underlying statistics.

Energy source	**Renewable**	**Availability**	**GHG**	**Environmental impacts**
Fossil-based	no	24 h	yes	high
Solar, wind, tide	yes	limited	no	low
Hydro-electrical	yes	24 h	no	medium
Biomass	yes	24 h	limited	medium
Geothermal	no	24 h	no	low
Nuclear	no	24 h	no	high

Table 1: Energy sources schematic comparison.

(e.g. solar panels, wind mills, hydro-electrical, etc.). For this reason, we say that ICT is *undirectly* responsible for the GHG emissions.

Therefore, we can identify three dimensions in the energy problem: energy consumption (Watt per hour, Wh), GHG emissions (kg of CO_2) and costs (Euros). Fortunately, there is a greater and greater attention by society, industries, academia and governments to *alternative* energy sources, which promise to help solving both the energy shortage and the GHG emission problems at the same time. The alternative energy sources are those ones which are not based on burning fossil fuels (like carbon, oil and gas), but the energy they produce comes from other, often renewable, sources such as the sun, wind, geothermal and nuclear. Anyway, even if alternative energy sources are part of the energy solution, they have also drawbacks that should be taken into account. First, not all the alternative energy sources are *green*: some of them emit GHG or have other polluting effects that have great impact on the environment (such as biomass and nuclear). Second, alternative energy sources may not be always-available, like the *legacy* ones, but their availability may vary with natural phenomena (like the sunlight or the wind), and it may depend on the geographical location of the plant: not all the sites have the same *potential* of generating renewable energy (see Table 1).

2. Network Infrastructure

The energy consumption and the GHG emissions associated to the operation of network infrastructures are becoming major issues in the ICS. Current network infrastructures have reached huge bandwidth capacity but their technological development and growth has not been accompanied by an equivalent evolution in energy efficiency. In 2008, the network infrastructures alone consumed a mean of 22 GW of power corresponding to $1,16\%$ of the worldwide produced electrical energy, with a growth rate of 12 % per year [5], further stressing the demand for effective energy optimization strategies affecting network devices, communication links and control plane protocols.

Optimizations performed onto a network during either the design or operation phase are generally aimed at maximizing performance, flexibility, and resilience while at the same time containing operating costs. The previous considerations foster the inclusion of global power consumption and carbon footprint (GHG emission) reduction in these objectives. Thus, the minimization of the above energy-related metrics will be the first containment goals, while the capability to fulfill service requests, as well as meeting budgetary limitations, will act as constraints. The savings achieved, in comparison with traditional routing approaches, will be proportional to the relative weight assigned to energy saving or carbon footprint reduction with respect to the other competing objectives.

However, the heterogeneous features and complex energy production processes associated the available energy sources make these ranking choices often difficult and contradictory, particularly when considering their environmental impacts. For example, several carbon footprint measurement criteria consider GHG emissions during the use phase only; neither the construction costs nor other environmental impacts taking place during fuel preparation and waste dismissal are considered at all; nuclear energy, although does not emit considerable quantities of CO_2 has other sever impacts on the environments and is not renewable as its fuel (mainly uranium and plutonium) is available only in limited quantity; the continuous exploitation of a geothermal source may induce a reduction of its efficiency and hence its attractiveness as a reliable energy source.

Other contradictory issues come from two prominent principles that have driven traditional network design, namely over-provisioning and redundancy, which, if taken in their native form, i.e. in an *energy-agnostic* way, may conflict with energy-saving efforts. More specifically, networks have been traditionally provisioned for worst-case estimated peak loads that typically exceed their long-term utilization by a significant margin. Thus, the energy demand of network equipment remains considerably high even when the network is idle, so that it is straightforward to observe that most of the energy consumed in networks is wasted. This presents several opportunities for substantial reductions in the energy consumption of existing networks, all based on the idea of limiting the energy-related impact of the part of network infrastructure that is not currently in use or is not used at its maximum possible load.

All the above considerations suggest that strategies oriented at power saving alone may have adverse effects on other metrics, such as path lengths, environmental friendliness and energy costs. For example, the most power-efficient routing solution may involve the choice of longer routes for paths than are found in conventional shortest-path routing, or dirty energy sources, adversely affecting the carbon footprint or the overall energy costs.

It is clear that a more sustainable scenario has to be configured in which energy must be considered as a fundamental constraint in designing, operating and managing telecommunication networks under multiple, sometimes conflicting, optimization objectives. The main components of such a scenario are *energy efficiency* and *energy awareness* that work together in an *energy-oriented paradigm* encompassing renewable energy sources in a systemic approach that considers the whole life-cycle assessment (LCA) of the new solutions.

3. Energy Efficiency

Te simplest strategy for reducing the power impact of network infrastructures is to improve the energy efficiency of the involved devices, so that more services can be delivered with no increase on the energy input, or the same services can be delivered for a reduced energy input. At the state-of-the-art, the most energy-efficient network devices are capable of operating on at least two different power levels. These levels will be henceforth referred to as the high power and the low power ones. The service offered by the device will be, generally, proportional to the power required: when running on a low power level, devices will be able to provide only limited throughput. If there are more than two power levels, a discrete set of power levels can be modeled, or power can be assumed to take values in a continuous interval. It is crucial to determine when to perform a transition from a power level to the other, so that the maximum amount of power can be saved without impairing the provided service. In presence of an increase in demand, the corresponding increase in power level can be started when the demand approaches a given threshold. A conservative choice of this threshold, i.e. at a relatively low value, increases the likelihood that a high power level is used when it wasn't really needed. At the other extreme, if the threshold is set too high, the system may fail to adapt to the required power level as quickly as needed, and service may suffer. Similarly, the transition from a high power level to a low one can be triggered when the demand falls (and stays) under another threshold. The performance of algorithms that govern such transitions will be determined by how closely these algorithms can match the demand. Hence, the key to performance is the ability to accurately *forecast* the evolution in the demand. The most sophisticated algorithms take advantage of statistical properties of network traffic, in order to make realistic predictions about upcoming service requests. In addition, the most comprehensive models also quantify a cost associated with the transition itself and take this cost into account. In the absence of a forecast technique, buffers can be used, usually at the expense of a controlled increase in the delay, to gain a low power time interval on low loaded interfaces.

When planning and designing a network infrastructure, budgetary considerations will also affect the choice of the devices to deploy. Energy-efficient devices must be bought and deployed, and this carries a cost. It is then important that the acquisition cost of energy-efficient devices be kept low as compared with "traditional" ones. The energy-efficiency capabilities must in fact be weighted against the additional costs that the network operators will incur if they select expensive hardware. If a network has to be built from scratch, including all of its components, and the initial design includes energy-efficiency considerations, then the network may be expected to require less energy for the same throughput. Alternatively, adapting the performance of systems or subsystems to the different operating conditions and workloads is a mechanism to bring the power draw down.

Any strategy intended to increase power efficiency should rely on accurate and detailed measurements of the network power requirements, taking into account all the subsystems and components that underlie the complex structure of a modern communication infrastructure. In fact, not all parts of a network will draw power in the same way: at any given time and performance level, some components will be drawing much higher fractions of the total energy than others. Although improvements are requested at all levels and on all components, it is evident that targeting energy-saving efforts to the more "thirsty" parts has the potential to yield considerably sized savings. In addition, measuring the per-component power drawn alone will not suffice within the context of a broader analysis. Since efficiency can be defined as performance/power ratio, to compute an accurate estimate of efficiency, a good measurement of performance is also called for. Another question arising in this context relates to what is the most appropriate timescale at which power-saving decisions and actions should happen. While shorter timescales have the potential to larger savings, they are most likely confined to single devices or even components, and thus they are likely to require newly designed hardware, with an increase in expenditure. On the other hand, solutions working at larger timescales could also apply to existing hardware, but they will likely span multiple elements across the network and thus require a strict coordination between these elements, that is often a challenging and still open issue. Whatever the mechanism advocated to achieve a reduction in the power requirements, a real-world implementation may not ignore the huge investments made in the existing infrastructure: strategies that squeeze the maximum performance at the minimum energetic cost from existing hardware (or with marginal additions) are likely to be very welcome to operators' headquarters.

These considerations assume a greater significance in presence of a new network infrastructure to be built from scratch or when massive network upgrade activities have to be started. In these situations, it is highly desirable to drive technological, architectural and topological choices by keeping into account not only the traditional performance and cost parameters but also considering the energetic budget as a first-class objective. Consequently, the well-known trade-offs between design and management decisions concerning capital (CAPEX) and operational (OPEX) expenditures must always be evaluated under an energy-efficiency perspective. Thus, each new device that improves the performance of its predecessors needs to be technologically compared with the competing ones also on its power requirement features.

As for the access networks, the currently specific deployed technology is the strongest enabler for energy-efficiency. Most of the current energy demand in carrier's infrastructures is associated to wired access connections. Unfortunately, the current access networks are implemented for their most part by using legacy copper-based lines ("the last mile") and transmission technologies such as ADSL and VDSL, whose power consumption is extremely sensible to their operating bit rates. The current trend is completely replacing in a few years such older technologies with fiber-only access infrastructures, which have the potential for improving significantly the overall energy efficiency. Thus, since energy consumption in access networks scales together with the number of end-users, the massive diffusion of fiber

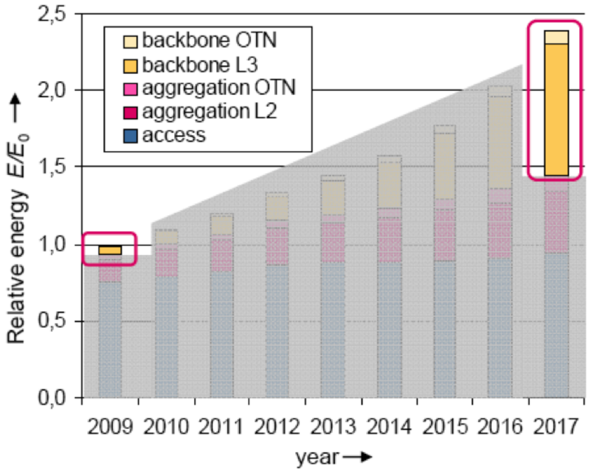


Figure 2: Energy consumption trend in communication infrastructures.

to the home (FTTH) local loop solutions, replacing the old copper xDSL access connections, would bring the dual advantage of radically enhancing the access bandwidth and simultaneously reducing the associated energy consumption. Only as a reference for comparison, it can be considered that an individual ADSL connection requires about 2.8 W, whereas an access infrastructure based on the passive optical network (GPON) paradigm will reduce the per-link consumption to only 0.5 W (for a giga-speed connection), with an improvement of about 80% in presence of very large number of users. The ongoing replacement of legacy last-mile copper infrastructure with fiber-based one is shifting the problem to the backbone component, essentially affecting the internet highways of the telecommunication world, where the energy requirements of high-end IP routers is becoming a bottleneck [10], since with the rising traffic volume, the major consumption is expected to shift from access to core networks (from less than 10% in 2009 to about 40% in 2017, see Figure 2 [11].

At the backbone network level, Optical Wavelength-Division Multiplexing (WDM) communication infrastructures are an ideal field of application of energy-efficiency. These networks are characterized by considerable bandwidth, arbitrary topologies (fully virtualizable through the flexible creation of wavelength-based end-to-end connections or lightpaths) and are highly reconfigurable.

While in traditional electronic or opto-electronic equipment the energy consumption is mainly due to effect of loss during the transfer of electric charges, and thus it directly relies on the specific operating voltage, frequency and on number of gates involved (often becoming unavoidable physical bottlenecks associated to the electrical technology), in transparent optical equipment the only factor conditioning power requirements is the technological complexity of coping with the fundamental physics of photons [10] (e.g., zero rest mass, weak photon-photon interaction, and 10^6 times larger size than electrons), resulting in a power demand that is often ten times lower, or more.

Minimizing energy consumption of optical networks can be generically addressed at four levels: component, transmission, network, and application [12]. Technological advances in optical devices such as optical add/drop MUXes (OADM), optical cross connects (OXC) and dynamic gain equalizer (DGE) can be complemented with power-saving solutions.

At the transmission level, low-attenuation and low-dispersion fibers, energy-efficient optical transmitters and receivers improve the energy efficiency of transmission.

In addition, when configuring and tuning a chain of optical amplifiers, the input power at the transmitter is adjusted to match the noise caused by intermediate components. This tuning is independent of the actual usage of the optical channel. The power values are thus chosen to support the maximum load that the associated channel can provide.

Finally, to really support energy efficiency, network solutions should easily adapt their power demand to the network topology (through the support of redundancy, multi-paths, etc.), to the traffic trend (bursts or always low) and to the specific operating scenario. They should also be realistic in terms of technology (providing compatibility and interoperability) and sufficiently reliable and scalable.

4. Energy Awareness

While energy-efficiency is the basic fundamental step towards energy-oriented networking, considering this aspect alone is not sufficient to achieve really satisfactory results in the medium and long term. Dynamic and flexible power management strategies, specifically conceived to decrease power consumption in the operational phase, are needed to substantially improve the positive effects on the environment and introduce more significant cost savings.

Such strategies originate from several studies [13][14] demonstrating that in a typical communication infrastructure, designed to ensure a satisfactory degree of reliability and availability trough link redundancy/meshing and to seamlessly support the maximum load also during traffic peak hours, a non-negligible portion of network links are, on average, scarcely used. These consideration, together with the necessity of introducing the global power demand as an additional constraint in the routing decision process, suggest several possible approaches to take advantage of link under-utilization in order to save energy. Such approaches imply the adaptive choice of energy-efficient devices and communication technologies, together with the adoption of dynamic network topologies whose task is minimizing the number and the overall energy requirements of devices and links that must always be powered on. Based on their specific energy containment choices, the aforementioned approaches can be classified in three main categories:

- the selective shutdown approach, taking advantage of the idle periods to switch off entire network devices or some of their ports;
- the adaptive slowdown approach, putting the network components (switching/routing devices or their interfaces) in low power mode or reducing their speed during under-utilization periods and immediately re-enabling their full operation capability/speed when needed;
- the coordinated energy-management approach, which advocates global solutions for network-wide power management based on energy-efficient routing, and more generally is based on new energy aware control plane services.

4.1. Selectively Turning Off Network Elements

Energy savings can be achieved by keeping the number of active network element to the minimum level that suffices to provide the services requested. This kind of optimization can be done both statically and dynamically, depending on whether the distribution of upcoming traffic demands is known (or it can be accurately estimated) or it is not. In the static case, which often involves re-provisioning of active connections, minimization of the total consumption is done via switching off the set of active elements accounting for the greatest amount of power, while maintaining sufficient capacity to support traffic. Dynamic scenarios are instead centered on the definition of traffic thresholds that trigger the shutting down or the starting up of network elements, together with the needed re-routing. Selective shutdown strategies can be implemented at different levels of granularity provided that the involved operating scheme has to be extremely flexible and should offer

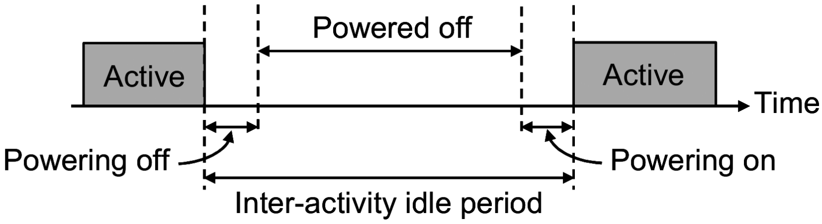


Figure 3: Selective shutdown timeline scenario.

the opportunity of proportionally saving energy as the number of active end-customers decreases. That is, when the inactivity phenomenon only affects some specific interfaces, instead of shutting down an entire network device, only those interfaces or their associated line-cards, when there is no activity on all the interfaces present on them, can be temporarily powered down. Clearly, if in a switching device all the line cards are put into down status, then the entire node can also be powered down safely, by achieving further significant energy savings since the chassis and its control logic (switching matrices, routing supervisor elements and timing cards) can consume about one half of the device's maximum energy budget [15].

Essentially, the major drawbacks of the aforementioned strategies result from a dichotomy: either a device is fully operational or it is not. This means that when the required throughput is very low as compared to the maximum, the device will be wasting most of its processing power and hence it should be powered off, reducing the overall operational costs (OPEX). On the other side when an entire switching device is powered down many expensive transmission resources become completely unused, hence negating significant capital investments (CAPEX) for the whole duration of the sleeping time. Furthermore, shutting down entire nodes can significantly reduce the meshing degree, by affecting the network reliability and partially negating the possibility of balancing the network load on multiple alternate paths. Unfortunately, with today's technology, putting a big router to sleep may be unpractical and even the selective shutdown of links and nodes is not gracefully supported. Also a router costing several hundred thousand Euros or more may take tens of minutes to get up again. Finally, when returning on from a previous down status, a peak in the power consumption is registered and the lifetime of the device will decrease a lot if frequent power up and down occurs.

Also the support of selective shutdown at the individual interface or line card level presents substantial drawbacks, mainly due to the needed changes to the traditional routing/switching device architecture and protocols. Hence, an interface switched into sleep mode stops responding to the periodic *hello* solicitations used for neighbor discovery in all the common routing protocol and thus can be classified as “down” or “faulty” at the control plane layer. As a consequence, the link state advertisement facility immediately floods such information throughout the network, flagging the interface operational status change. Thus, sleeping links are detected as link failures, potentially introducing severe stability problem that affect the convergence of the involved routing (or spanning tree) algorithm. To cope with this problem, the *hello* facility should be restricted to awake interfaces only, loosing a lot of protocol flexibility. In addition, network devices need to acquire the ability to predict low load periods and hence know in advance when performing shutdown or wake-up operations. Such knowledge should also be incorporated into routing protocols and generally into the entire set of control plane facilities.

4.2. Enabling Low-power Modes

Notable energy savings can be obtained when a significant number of devices spend a major fraction of their idle time in an operating mode characterized by reduced power draw (i.e. the aforementioned “low power” mode). Although the potential savings vary from device to device, the energy demand when the device is in low power status can be as low as 10 percent than the one in active mode. However, during the transitions from or to low power mode the device may experience a considerable increase in energy consumption since many elements in the transceiver have to be kept active. The experienced value will be strongly dependent on the device implementation details and may possibly range from 50 percent to 100 percent of the active mode energy demand.

Unfortunately, current network equipment can only be either in the “on” or the “off” state, and the transition between these two states can be lengthy (minutes) and usually requires manual intervention.

To cope with this problem, future devices must be capable of quickly entering into and exiting from the low-power status or supporting some type of down-clocking feature to adapt to extemporaneous changes in traffic demand. State-of-the-art electronic devices used in the realization of broadband infrastructures are usually designed to achieve their maximum performance when operating according to an “always on” paradigm. The development of next generation networking devices based on hardware architectures supporting fast “sleep” or “low-power” modes will introduce

Technology	**Wakeup Time**	**Sleep Time**	**Average Power savings**
100baseTX	30 μs	100 μs	90%
1000baseT	16 μs	182 μs	90%
10GbaseT	4.16 μs	2.88 μs	90%

Table 2: Common wake-on-arrival parameters

new opportunities for efficiently reducing their power consumption when idle or partially idle. Energy proportional computing techniques, reducing the involved microprocessors' clock during inactivity periods can reveal to be really effective for energy saving purposes only if effective full-speed clock return procedures are available, in order to ensure an acceptable degree of responsiveness and avoid perceivable status switching delays.

For example, modern Intel processors such as the Core Duo [16] implement a sequence of sleep states (called C-states) that offer reduced power states at the cost of increasingly high latencies to enter and exit these states. The massive introduction of these energy-control technologies in networking hardware design will imply an epochal shift from the "always on" to "always available" paradigm in which each network device can spontaneously enter a "sleep" or "energy saving" mode when it is not used for a certain time as long as be able to wake up very quickly, by restoring its maximum performance, when detecting new incoming traffic on its ports. However, to achieve this "wake-on-arrival" (WoA) behavior, the circuitry sensing packets on a line must be left powered on, even in sleep mode.

By putting in low-power mode a single interface, any transmission activity on it is interrupted in presence of no associated traffic to be forwarded and quickly resumed when new packets arrive from it or are directed to it. The use of such technique is based on the definition of significantly large time intervals in which no signal is transmitted and smaller time slices during which a brief signal is transmitted to synchronize the receiver. When operating in such way, also known as Low Power Idle (LPI), the elements in the receiver can be frozen [17] and then awakened within a few microseconds, as reported in Table 2.

On the other hand, the ability of dynamically modifying the link rate according to the real traffic volume is used as a method for reducing the power consumption. Letting a device work at a reduced frequency can significantly lower its energy consumption and also enable the use of dynamic voltage scaling (DVS) technologies to reduce its operating voltage. Such technique

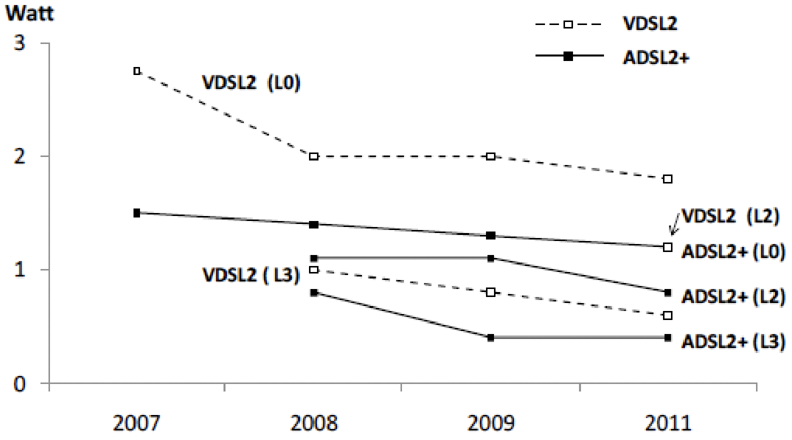


Figure 4: The evolution of rate-dependent power demand in traditional access infrastructures.

allows the needed power to scale cubically, and hence the associated energy demand to grow quadratically with the operating frequency [18]. For example the Intel 82541PI Gigabit Ethernet Controller consuming about 1 W at 1 Gbps full operation supports a smart power down facility turning off the PHY interface layer if no signal is present on the link and dropping the link rate to 10 Mbps when a reduction of energy consumption is required. Analogously, in the last mile scenario, the ADSL2 standard (ITU G.992.3, G.922.4, G.992.5) has significantly evolved to support multiple data speeds corresponding to different link states (L0: full rate, L2: reduced rate, L3: link off see Figure 4) for power management purposes [19].

The IEEE Energy Efficient Ethernet working group, when evaluating ALR as an alternative to LPI in the 802.3az standard, decided in favor of LPI.

4.2.1. Building energy models

Defining a sustainable and effective model for energy consumption is the fundamental prerequisite for introducing power awareness within the routing context. A great variety of networking devices may contribute to the overall power absorption: ranging from "opto-electronic" regenerators, optical amplifiers, to opaque routers and totally optical switches. Each device draws the needed power in a specific way, also depending on the relationship occurring between the different components of more complex structures such as switching systems or end-to-end communication links. In addition, some nodes may be powered by renewable sources, while others may use traditional, “dirty” energy, therefore a differentiation between energy sources is required. An energy model has the main task of characterizing the different components of the network involved in energy consumption. It provides the energy demand profile of network devices and communication links of any typological layout and under any traffic load.

Modern communication networks are usually modeled as a graph $G(V, E)$, where V is the set of nodes (a node is an optical router or an OXC equipped with optical transceivers) and E is the set of optical links. Since real-world networks have several characteristics that can impact on power draw, models of increasing complexity will embed larger sets of these characteristics. A review of some dimensions along which models differ is the following:

- *Direction of links.* Typically a link is considered as bidirectional, i.e. as having two optical fibers, each in one direction. Finer-grained models may consider unidirectional links. The resulting network graph will then be a directed graph.

- *Available wavelengths.* A simplifying assumption is that the transceivers deployed across the network are all identical, or at least interchangeable. The same set of wavelengths will then be available on all links. If this assumption is relaxed, a specific set of available wavelengths is to be added to the description of each link. These models will be more useful when evaluating networks where achievement of savings in energy is set as an objective, but the evolution towards this objective is done in an incremental way.

- *Wavelength conversion capability.* A constraint in optical routing is that, in absence of wavelength conversion capability at the nodes, a lightpath has to use the same wavelength over all its span. This greatly increases the difficulty of routing a lightpath, as it is not enough to find a feasible path with available resources, but that path must share a single wavelength on all the links. If, conversely, some or all the nodes can convert wavelengths, i.e. set up a lightpath using different wavelengths, the problem becomes simpler. Wavelength conversion capability can be implemented either in the electronic domain or in the optical one. Whatever the case, usually, conversion-capable hardware is more expensive than regular hardware. A model allowing the coexistence of conversion-capable equipment with nodes not having that capacity will describe more closely those real-world scenarios where an operator is progressively upgrading its infrastructure and wishes to determine the nodes where deploying conversion-capable equipment will yield the most benefit. These nodes could then be given precedence in the upgrade plans.

- *Wavelength capacity.* The capacity that a single fiber is capable of carrying may vary, again depending on the equipment used. Models describing the different capacities are focused on grooming smaller-sized electronic channels onto the optical ones.

- *End-to-end connections.* The power drawn by a single routed connection is given by:

  - the power absorbed by the transceivers
  - the power required by intermediate optical switches
  - the power consumed by optical amplifiers along each fiber link on the path

  These components will drain a given amount of power Θ for just being turned on. The value of Θ will depend on equipment technology and size. Additional power Φ will then needed for operation. If this *variable* power Φ is assumed to be proportional to the provided level of service/load ℓ (e.g. considering the current transmission rate and the maximum bandwidth), a realistic power consumption model can be sketched as in Figure 5. Accordingly, the energy consumption $P_{i,j}$ associated to the individual connection (i, j) can be expressed as:

  $$P_{i,j} = \sum_{k \in (i,j)} \Theta_k + \sum_{k \in (i,j)} \Phi_k(\ell) \tag{1}$$

  where, for each intermediate device k occurring on the path serving the connection (i, j), the individual Θ_k and Φ_k correspond respectively to the fixed and load-dependent consumption associated to that device.

- *Pre-computed paths.* The full problem specification for the establishment of a virtual topology and for the routing of connection requests onto it involves determining the most appropriate path in the network to be used for routing. While, in principle, very long paths can be selected when routing a request, shorter paths will evidently tend to occupy minor amounts of resources, and shortest paths will require

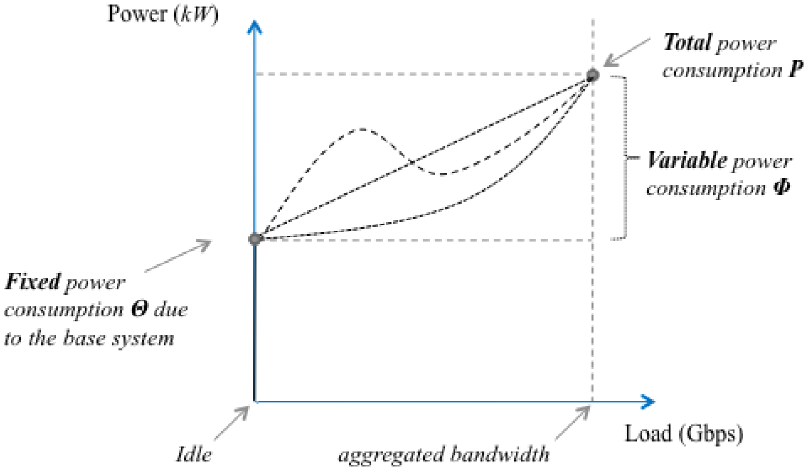

Power (kW)
Total power consumption P
Variable power consumption Φ
Fixed power consumption Θ due to the base system
Idle
aggregated bandwidth
Load (Gbps)

Figure 5: Modeling per-connection energy consumption.

the minimum. Only when a single link is present on the shortest path between many node pairs, that link will become critical and its usage will need to be regulated wisely. The risk is that some node pairs will become disconnected. Even in these cases, inordinately long paths are hardly of benefit: paths that are only slightly longer than the shortest ones are instead likely to provide the most appealing route for the incoming connection. This explains why many models avoid computing and exploring all the possible paths. They rely instead on a pre-computed set of paths between all node pairs, and restrict the routing choice to those paths only. An important parameter in these models is the number of alternate paths that is pre-computed. Alternatively, path pre-computation may be restricted to paths whose length would not exceed the length of the shortest paths of more than a specified amount.

- *Node structure.* The majority of models consider optical switches or ROADMs (Reconfigurable Optical Add-Drop Multiplexers) as single, atomic, units. These elements are, instead, composed of many parts and components and the trend in technological evolution of this hardware is to alleviate the central CPUs from as many computing chores as possible, delegating them to “satellite” CPUs such as those found, for example, on line cards. The single cards could, therefore, have a separate behavior as compared to the other cards and even to the central unit. The behavior of the unit as a whole will be in this case the average of the behavior of the single cards. Models aiming at exploring these aspects should describe the single cards, their parameters, and the association with the chassis backplane.

- *Traffic matrix.* Advance knowledge of the entire traffic matrix is a simplifying assumptions often used when traffic reconfiguration, as opposed to initial handling, is examined. There are in fact times when, by re-computing routes for all the active connections and by rerouting these onto the new routes, dramatic gains in efficiency can be achieved. The objective is then to best provide for the current traffic demand: an "optimized" network is anyway likely to have enough room to accommodate forthcoming requests, too. If the upcoming requests are assumed to be unknown, instead, the model will be aimed at finding the best adaptiveness, that it at having a well balanced network able to satisfy as many arbitrary requests as possible. An intermediate scenario is given when future requests are not known, but they are assumed to be drawn from a known distribution, possibly repeating past behavior.

- *Power levels.* Different technologies will have different power states and associated levels (see for examples Table 2). A detailed description of these levels is needed when try to evaluate the viability or effectiveness of a power-saving strategy. The saving achieved should be assessed with respect to the constraints, for example the capability to satisfy a certain level of demand.

- *Transition times.* The ability to adapt the power level to the characteristics of service demand is a critical factor. This ability can be taken for achieved when the transition times required to shift from a power level to another are so short that power transitions may happen with a null or minimal or impact on service. Models, intended as a validation for energy-saving strategies, should deal with this issue in great detail.

4.3. Control plane protocol extensions

Decisions taken at the control plane level to reflect power demand characteristics of the network devices, their power supplying source, its associated energy cost and the way they change over time, will need the availability of accurate and up-to-date information about the state of network elements. The state should be described in terms of services offered and of power requirements, including both the value of power draw and the type of supply used. To this end, energy-oriented extensions to the routing protocols used to gather and disseminate information are needed. These extensions clearly require modifications to the current routing protocols and control plane architecture. That is, the existing routing protocols (e.g. OSPF or IS-IS) within the GMPLS framework should be extended to associate new energy-related information, such as the power consumption, to each link or the type of energy source to each network element.

Different energy cost values can be dynamically associated to network nodes and links depending on the preference degree associated to green and dirty energy source and their specific types. That is, they can be powered by different green or dirty energy sources, updated on prefixed time basis, where the green sources represent energy from wind mills, solar panels or hydro-electrical plants, with their respective preference degree, whilst dirty energy sources models energy from coal, fuel or gas.

Clearly, these costs, together with additional ones associated to specific device-dependent power demand information associated to network elements and links, have to be used to influence the routing decision. Accordingly, to achieve the goal of green-aware routing, the routing decision can be based on the “lowest GHG emissions”, or “minimum power consumption” rather than on the “shortest/minimum cost path”, where the energy costs are defined within the specific energy model used.

In addition new node and sleep attributes are needed to support energy efficient traffic engineering features by modeling new link/path status and link or device power on/off capability, aiming at distinguishing between down and sleeping network elements.

The transport of the aforementioned energy-related capabilities and costs and their diffusion/exchange throughout the network can be easily managed by introducing new specific type-length-value (TLV) objects in IS-IS or opaque Link State Advertisements in OSPF.

Of course, the same information has to be fully handled by signaling protocols such as RSVP-TE and CR-LDP to allow the request and the establishment of power-constrained paths across the network (i.e., path traversing only nodes powered by renewable energy sources or crossing only low-power transmission links).

5. Energy-oriented network infrastructure

Whereas the selective shutdown and slowdown approaches typically work at the specific network components level, further energy-efficiency improvements may be achieved by moving the focus at a wider scale. Accordingly, control plane protocols and routing algorithms can be improved to save energy by introducing a new dimension whose goal is to properly accommodate network traffic by considering both energy-efficiency and energy-awareness.

These improvements essentially consist in properly conditioning the route/path selection mechanisms on a relatively coarse time scale by promoting the use of renewable "green" energy sources as long as energy efficient links and switching devices, simultaneously taking advantage from the different users demands/current load across the interested communication infrastructures, so that the global power consumption can be optimized (Figure 6).

The likely outcome would be that some of the "best" paths, selected according to traditional network management criteria, would not be the better ones in an absolute sense, but the power savings and environmental benefits consequent to such an apparently sub-optimal choice, could be substantial without excessive losses on the other concurring optimization goals. Furthermore, it has been observed [20] that the increments in path lengths do not increase energy consumption in a perceivable way, since routers and switches are not designed to be energy proportional and the power absorbed by a packet when crossing a network node is several orders of magnitude below the energy requested at the terminal point of the path.

Coordinated power-management schemes may also benefit from the previously cited selective shutdown or rate adaptation mechanisms, by taking the associate power off/on and mode change decisions on a network-wide perspective, and hence basing them on a more complete awareness of the overall network economy, so that greater energy savings can be achieved.

Providing energy-awareness at the network control plane level also implies the necessity of periodic re-optimization campaigns with the aim of placing the already "in production" network traffic over a new set of paths so that, in presence of substantial modifications in traffic load/distribution, device power consumption, energy costs and/or specific energy source availability, the aggregate power consumption is minimized whereas all the end-to-end connection requirements are still satisfied. Thus, in highly redundant network scenarios, entire network paths can be switched off by re-routing the associated connections on other already existing paths or on newly created ones. In particular, energy-aware routing implies just-in-time optimization

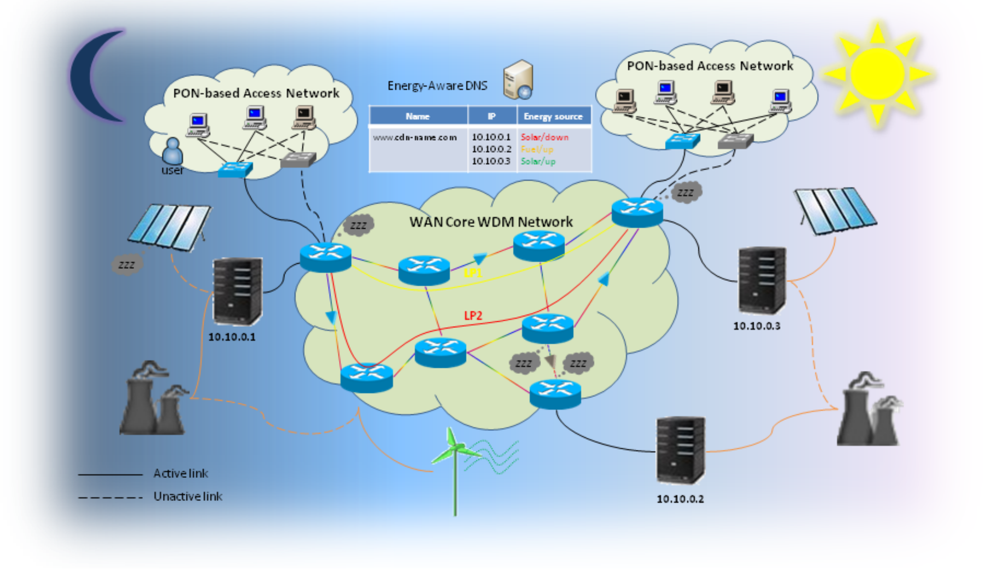


Figure 6: Moving the traffic on energy cost or consumption basis.

of the energy sources choice in such a way that renewable sources are always preferred when available. This new requirement originates from the consideration that network elements may be served by multiple power supplies, specifically providing an always available power source coming from traditional dirty energy as long as an intermittent power source associated to green energy, when available. Consider, as an example, the availability of the green energy produced by wind or solar panels; it is strongly correlated with the weather status or time of the day. Prediction of weather conditions, such as the presence of wind implies some degree of uncertainty whereas, it is easier to assume that no solar energy will be produced during the night hours and that a certain amount of energy is expected to be produced when the sun shines.

Energy aware routing has the additional goal of distributing the network-wide traffic so that, when the wind is blowing in a specific geographic area, all the viable network traffic should be forced to pass through the routing infrastructures located there, and when the wind stops blowing, the above traffic should be dynamically shifted back elsewhere, e.g. where the sun is shining, in such a way that the use of green energy sources is maximized.

Alternatively, by working on the energy cost dimension, significant re-

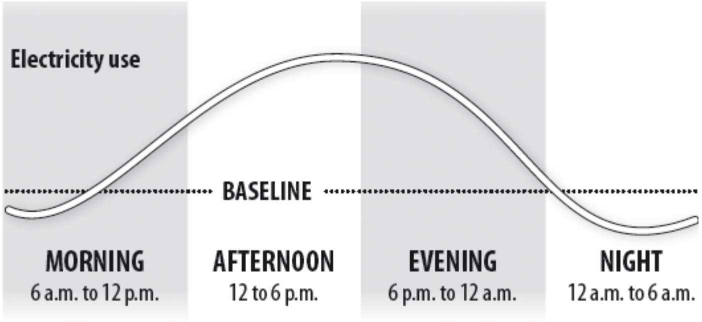


Figure 7: The typical daily electricity demand cycle.

ductions could be achieved by adaptively rerouting the network traffic to locations where electricity prices and associated taxes are lowest on particular hours of the day. Electricity prices present both geographic and temporal variations, due to differences in regional demand (often based on lifestyle and weather conditions), power grid-related transmission inefficiencies, and generation technology diversity.

Since electricity cannot be stored and has to be consumed instantly, the electric system typically has to keep spare “peaking” generation capacity online for times when demand may surge on short notice. Often, these “peaking” power plants are only run for a few hours during the day, during the maximum demand period (Figure 7) adding to the average cost of providing electricity. Dynamic pricing encourages electricity consumers to reduce their usage during peak times, especially during the critical hours of the day by substantially lowering the energy costs during the off-peak hours and hence limiting the need of “peaking” power plants.

Thus, the intuition is that well-known time-based fluctuations in electricity costs and tax incentives for reduced carbon-emission across the covered geographic area may offer opportunities for reducing OPEX, provided that, by considering the involved physical distances and delays, the cost of moving the traffic is sufficiently lower than the likely cost savings from reduced energy prices. Of course, the more rapid is the reaction to price changes, the greater are the overall savings.

Hence, to ensure the necessary flexibility, the aforementioned energy cost and absorption knowledge should be introduced at the control plane level for energy-aware/price conscious routing and communication resource allocation, to implement automatic and adaptive follow-the-sun, chase-the-wind or follow-the-minimum-price paradigms.

Furthermore, in order to support all the above adaptive routing behaviors, energy-related information associated to devices, interfaces and links need to be easily and reliably determined, depending on the involved technology and current traffic load, according to a comprehensive energy models, and introduced as additional constraints (together with delay, bandwidth, physical impairment etc.) in the formulations of dynamic routing algorithms and heuristics. In this scenario, down-clocking or selective shutdown facilities should be handled as new features in the network element status that need to be considered at the routing and traffic engineering layer. Clearly, to support the aforementioned energy-aware behaviors, such information must be conveyed to all the network devices operating within the same energy-management domain.

6. Conclusions

Until now, the design and development of new network equipment and solutions was fundamentally dominated by performance-related objectives. However, due to the astonishing growth of network infrastructures worldwide, their electric power demand has reached extraordinary peaks, strongly affecting the carriers, their operating costs and the overall model scalability and sustainability. As a direct consequence, it is now strictly necessary to consider energy efficiency as a basic constraint and a fundamental priority in the design and development of next generation networks. In addition, Governments and telecommunication operators are endorsing the development of green renewable energy sources (such as solar panels and wind turbines) for powering network elements so that the choice of a specific energy source, when possible, becomes a fundamental parameter not only for reducing energy costs but also for deploying more sophisticated and environmentally friendly energy-aware network management strategies.

In fact, while energy-efficiency efforts may yield considerable savings, if the problem is considered in a wider perspective, including the evolution in the demand for advanced services and the availability of renewable energy sources, other consideration will emerge. Energy-aware network elements adapt their performance and behavior depending not only on the actual load, but also on the currently used energy source.

Hence, coordinated energy-aware control plane strategies, driven by multi-objective optimization, may be more helpful in finding the appropriate point on the Pareto front according to the relative importance of network performance, energy consumption, cost containment and environmental friendliness. Its most significant added values originate from the definition of comprehensive energy consumption models for the modern networks, incorporating both physical layer issues such as energy demand of each component and virtual topology-based energy management information. The resulting strategies will limit the usage of energy-hungry links and devices, privileging instead energy-efficient equipment and solutions, giving attention to the type of sources. Moreover, intelligent grooming mechanisms that reuse the same switching nodes, fiber strands and interfaces along the same path as much as possible, will also optimize resource usage by concentrating the traffic load. Thus, energy-aware decisions, taken on larger aggregates and hence on smaller amounts of data, will be simpler from the operating point of view and will yield amplified energy-related benefits.

References

[1] Cisco visual networking index [online]. Available: `http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html`.

[2] Internet world stats [online]. Available: `http://www.internetworldstats.com/emarketing.htm`.

[3] Etforecasts [online]. Available: `http://www.etforecasts.com/products/ES\_intusersv2.htm`.

[4] Bt announces major wind power plans [online]. Available: `http://www.btplc.com/News/Articles/Showarticle.cfm?ArticleID=dd615e9c-71ad-4daa-951a-55651baae5bb` (2007).

[5] BONE, Bone project, 2009, wp 21 tp green optical networks d21.2b report on y1 and updated plan for activities. (2009).

[6] S.Pileri, Energy and communication: engine of the human progress, in: INTELEC 2007, Rome, Italy, Sept. 2007.

[7] L. S. Foll, Tic et énergétique: Techniques d'estimation de consommation sur la hauteur, la structure et l'évolution de l'impact des tic en france, in: Ph.D. dissertation, Orange Labs/Institut National des Télécommunications, 2009.

[8] T. G. Grid, The green grid data center power efficiency metrics: Pue and dcie, technical committee white paper (2008).

[9] Living planet report 2010, the biennial report, wwf, global footprint network, zoological society of london (2010).

[10] R. Tucker, R. Parthiban, J. Baliga, K. Hinton, R. Ayre, W. Sorin, Evolution of wdm optical ip networks: A cost and energy perspective, Lightwave Technology, Journal of 27 (3) (2009) 243 –252. doi:10.1109/JLT.2008.2005424.

[11] C. Lange, D. Kosiankowski, R. Weidmann, A. Gladisch, Energy consumption of telecommunication networks and related improvement options, Selected Topics in Quantum Electronics, IEEE Journal of 17 (2) (2011) 285 –295. doi:10.1109/JSTQE.2010.2053522.

[12] Y. Zhang, P. Chowdhury, M. Tornatore, B. Mukherjee, Energy efficiency in telecom optical networks, Communications Surveys Tutorials, IEEE 12 (4) (2010) 441 –458.

[13] K. J. Christensen, C. Gunaratne, B. Nordman, A. D. George, The next frontier for communications networks: power management, Computer Communications 27 (18) (2004) 1758 – 1770. doi:10.1016/j.comcom.2004.06.012.

[14] A. Odlyzko, Data networks are lightly utilized, and will stay that way, Review of Network Economics 2 (3) (2003) 210–237.

[15] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, S. Wright, Power awareness in network design and routing, in: INFOCOM 2008. The 27th Conference on Computer Communications. IEEE, 2008, pp. 457 –465. doi:10.1109/INFOCOM.2008.93.

[16] A. Naveh, E. Rotem, A. Mendelson, S. Gochman, R. Chabukswar, K. Krishnan, A. Kumar, Power and thermal management in the intel core duo processor, Intel Technology Journal 10 (2) (2006) 109–122.

[17] P. Reviriego, J. Hernández, D. Larrabeiti, J. Maestro, Performance evaluation of energy efficient ethernet, Communications Letters, IEEE 13 (9) (2009) 697–699.

[18] B. Zhai, D. Blaauw, D. Sylvester, K. Flautner, Theoretical and practical limits of dynamic voltage scaling, in: Proceedings of the 41st annual Design Automation Conference, DAC '04, ACM, New York, NY, USA, 2004, pp. 868–873. doi:10.1145/996566.996798.

[19] M. T. [online]. Available: `http://www.eetimes.com/electronics-news/4136967/ADSL2-Helps-Slash-Power-in-Broadband-Designs`, ADSL2 Helps Slash Power in Broadband Designs (2003).

[20] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, B. Maggs, Cutting the electric bill for internet-scale systems, in: Proceedings of the ACM SIGCOMM 2009 conference on Data communication, SIGCOMM '09, ACM, New York, NY, USA, 2009, pp. 123–134. doi:10.1145/1592568.1592584.

Green Datacenter Infrastructures in the Cloud Computing Era

Sergio Ricciardi[a], Francesco Palmieri[b], Jordi Torres-Viñals[a,c], Beniamino Di Martino[b], Germán Santos-Boada[a], Josep Solé-Pareta[a]

[a]*Technical University of Catalonia - BarcelonaTech*
Department of Computer Architecture
C. Jordi Girona 1-3, E-08034, Barcelona, Spain
{sergior, german, pareta, torres}@ac.upc.edu
[b]*Second University of Naples*
Information Engineering Department
V. Roma, 29, I-81031, Aversa (CE), Italy
{francesco.palmieri, beniamino.dimartino}@unina.it
[c]*Barcelona Supercomputing Center*
Autonomic Systems and eBusiness
C. Jordi Girona, 29, E-08034 Barcelona, Spain
jordi.torres@bsc.es

Abstract

Energy consumption and the related green house gases (GHG) emission of datacenters are becoming a major issue in the information and communication society (ICS). Datacenters and vast computer warehouses consume a considerable percentage of the worldwide produced electrical energy, and the energy required to cope with the ever increasing demand for runtime and storage power grows faster and faster together with the widespread diffusion of cloud-based services delivered over the Internet. In this scenario, the costs of delivering power, cooling, network connectivity, storage space, and real estate to underutilized servers continue to increase. Green computing paradigms are emerging for reducing the energy consumption, the consequent GHG emissions and the operational costs, for example by letting data and computational load to follow-the-sun/wind/tide in distributed and cooperative service infrastructure, such as the clouds, to exploit renewable energy sources and variable energy prices offered by Smart Grids. In this chapter, we present the current challenges and research trends for datacenter infrastructures and discuss a new energy-oriented model based on server consolidation that, considering the energy as an additional constraint, minimizes the energy consumption and the associated operational costs and environmental effects towards an eco-sustainable society growth and prosperity.

Keywords: Green datacenters, cloud computing, energy consumption, GHG emissions, energy cost.

1. Introduction

In the last decades, the society has experienced drastic changes in the way the information is accessed, stored, transmitted and processed. Data has been digitalized to allow electronic processing, by migrating from physical to digital supports, and the Internet has made it accessible from whatever device connected to the global network. This has produced deep changes in the society, industries and communications, giving rise to the so-called *Information and Communication Society* (ICS).

Several milestones can be identified in the history of the ICS, approximately every fifteen years [1]. At the beginning of the ICS, there were the mainframes: big machines that processed jobs in batch, i.e. sequentially and offline. With the advent of the personal computers (PC), the users could have their own small computation and storage resources on their desktops. This was a significant change in the paradigm, since each user could process its jobs on its own. The interconnection of the PCs was possible with the advent of the Internet, mainly based on the client-server model. The servers grew in size and functionalities, and federated datacenters and farms distributed throughout the world started cooperating to offer more and more services to the users, evolving into the actual Internet-scale computing paradigm and service facilities commonly known as Grid and Cloud infrastructures.

Such a paradigm represents a technological revolution in the way the information is accessed, stored and shared from whatever device connected to the Internet which, in turns, evolves from a "simple" connection infrastructure to an integrated platform offering services to millions of smart, always-connected terminal devices. The fundamental advantage of the aforementioned cloud paradigm is the abstraction between the physical and logical resources, needed to provide services, and the users, which can simply *use* the services they need (according to the so-called *software, platform and infrastructure as a service* models) without having to worry on *how* they are actually implemented.

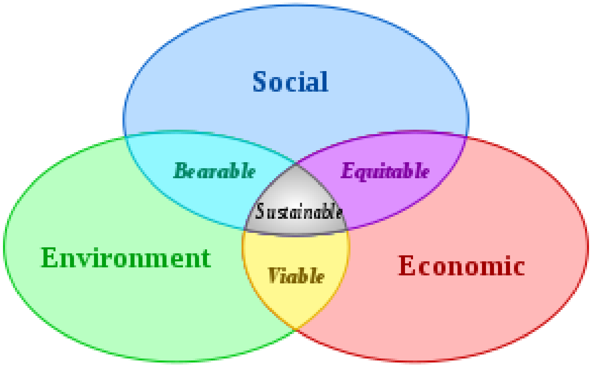


Figure 1: Eco-sustainability as equilibrium among Society, Environmental and Economics.

In addition to the large number of technological and architectural issues originated by the above paradigm change, the ICS is now facing another major challenge: *eco-sustainability*. The eco-sustainability has become a major issue in the ICS since resources exploitation, climate changes and pollution are strongly affecting our planet. In particular, eco-sustainability refers to the equilibrium among the society, the environment and the economics (Figure 1), by meeting the needs of the present generation without compromising the ability of future ones to meet their own needs [2].

Today, the ICS has the opportunity to put together the above evolutionary distributed computing technologies and eco-sustainability into an integrated ICT paradigm, encompassing both the ever increasing need of the development and the preservation of the world's natural resources heritage. The ICT sector has in fact a unique ability to reduce its ecological footprint (basically, energy consumption and GHG emissions) as well as the ones of other industry sectors through the use of technological innovative solutions [3]. The first step is therefore the reduction of the ICT ecological footprint that will drive the change towards society growth and prosperity.

The key factors of such an energy-oriented paradigm are (1) *energy-efficiency*, that lowers the energy requirement of the involved service provisioning devices without affecting their performance, and (2) *energy-awareness*, which adapts the overall energy consumption to the current load, introducing energy-proportionality, and exploiting the use of renewable energy sources and emerging *Smart Grid* power distribution networks. Energy-efficiency and energy-awareness goals can be achieved together into an (3) *energy-oriented* systemic approach considering the whole life cycle assessment (LCA) of any new solution to ensure that it will be effective and avoid a rebound effect[1].

2. Background

2.1. The energy problem

The humans demand on the biosphere, known as the human ecological footprint[2], has been estimated to be equivalent to 1.5 planet Earths [5], meaning that humanity uses ecological resources 1.5 times faster than the capacity of Earth to regenerate them. Simply stated, humanity's demands exceed the planet's capacity to sustain us. The scarcity of the traditional fossil energy sources with the consequent rising energy costs have become one of the major challenges for the ICS. Therefore, as part of the ecological footprint, the energy consumption and the concomitant GHG emissions of cloud computing infrastructures (datacenters and ultra-high speed interconnection networks) have imposed as new constraints for ICT.

The ever increasing data volumes to be processed, stored and accessed every day through the modern Internet infrastructure result in the datacenters energy demand to grow at faster and faster pace. For this reason, energy-oriented datacenters are being investigated in order to lower their ecological footprint. However, since the electrical energy needed to power the ICT is not directly present in nature, it has to be derived from primary energy sources, i.e. from sources directly available in nature, such as oil, sun and nuclear, that may be renewable or not-renewable. Not-renewable energy sources, like fossil fuels, are burned emitting large quantities of GHG in the atmosphere (usually measured in carbon dioxide equivalent[3], CO_2e), thus contributing to global warming and pollution. Renewable sources may be also exploited to

[1] There is a logical problem related to have more efficient components: higher efficiency may lead, in fact, to decreased costs and to increased demand, possibly overcoming the gains obtained with efficiency, a phenomenon known as *rebound effect* or, in other contexts, as *Jevons Paradox* or *Kazzom-Brokes postulate* [4].

[2] Resources exploitation, pollution, climate changes, global warming and global dimming form all part of the human ecological footprint.

[3] Carbon dioxide equivalent is a quantity that describes, for a given mixture and amount of greenhouse gas, the equivalent amount of CO_2 that would have the same global warming potential (GWP), when measured over a specified timescale (generally, 100 years).

generate electricity. Most of the renewable energy sources are clean (usually referred as *green*) as they do not emit GHG during their use (with the exception of biomasses[4]) although some other drawbacks may be still present, like visual and noise impact in the case of wind turbines or large surfaces covered in the case of solar panels. Nevertheless, from an ecological point of view, renewable energy sources are beneficial over their entire life-cycle [6], even if their cost and efficiency may be still low when compared to fossil-based energy sources [7], which is the main reason for which their employment has to be carefully considered against a trade-off assessment between advantages and drawbacks. Besides being clean, renewable energy sources – as the name suggests – are virtually *inexhaustible*, thus they are the perfect candidate to support eco-sustainable growth. Nonetheless, renewable energy sources may be not always available; sun, wind and tide are cyclic or even almost unpredictable phenomena, though some inertia is guaranteed by battery packs and potential energy storing techniques.

2.2. Smart Grids

Both fossil-based fuels and renewable energy sources can be employed together in a dynamic adaptive fashion within the intelligent power distribution system provided by the Smart Grids. Smart grids are therefore emerging as promising solution both to achieve drastically reduction in GHG emissions and to cope with the growing power requirements of ICT network infrastructures. They promise to change the traditional energy production/consumption paradigm in which one large (legacy) energy plant provides with energy the whole region, towards a configuration in which many small renewable energy plants (e.g. solar panels placed on the top of the buildings, wind turbines in the courtyards, etc.) interchange the energy with the power distribution grid, by producing their own energy and releasing the excesses to the smart grid which redistribute it together with the energy produced by the legacy power plants to the sites where the energy is needed or the renewable energy is not currently available. Smart grids open a new scenario in which the energy production and consumption can be closely matched avoiding peak power productions, and in which the energy quantity, quality and cost vary in function of the power plant producing it. Therefore,

[4] When burned to generate electrical energy, biomasses emit large quantity of CO_2, but the growing of the plants partially compensates the emissions. Biomasses, however, have other impacts on crops since they take away soil to the agriculture.

smart grids are foreseen to play a fundamental role in reducing GHGs emissions and energy costs since they allows premises, datacenters, storage and computational power to be interconnected to different energy sources and possibly dislocated near renewable energy plants or where the environmental conditions are favorable (e.g. cold climate can be exploited to efficiently cool down machines).

2.3. Follow-the-energy and follow-the-data

Therefore, an energy-aware paradigm relying on smart grid infrastructure will be able to either choose to direct the computing tasks or the data toward a site which is currently green powered (thus, in a *follow-the-energy* manner), or to request to the smart grid the quantity and quality of energy (e.g. from an available renewable energy production site) to the facility (in a *follow-the-data* manner). Such energy-aware paradigm unveils totally new potentials for the ICT which had not been explored before, not only for the cloud computing infrastructure but for the entire ICT, industrial and transportation sectors.

In this sense, a follow-the-energy (e.g. follow-the-sun, follow-the-wind, follow-the-tide, etc.) approach and the knowledge of the current power consumption of the devices may be taken into account into an integrated approach to optimize the overall energy consumption, GHG emissions, energy costs and performance. In the follow-the-energy approach, the preferred sites to which data or tasks are retrieved, stored or transmitted are the ones currently powered by green renewable energy (e.g. the current energy source is solar and it is daytime). In this case, it is possible to “make light from light”, in the sense that the energy coming from the solar panels can power the fiber optics and transmit the required data. This approach will be commonly used to load the facilities which are already powered by green energy sources.

In the follow-the-data approach, instead, the smart grid will be able to fulfill an energy provisioning request specifying both the quantity and quality of the required energy. The smart grid will thus reply in the same way as a typical telecommunication network, operating under the control of an engineering-capable control plane, i.e. by fulfilling the request and establishing the appropriate energy path or rejecting the request according to profitability/availability criteria. If the smart grid control plane decides to fulfill the request, an appropriate energy path has to be established between one of the energy sources available that satisfying the request. The path will be created, like in the telecom network domain, by establishing the correct circuits in the smart energy switches distributing the power from the energy sources to the grid. This way, the energy is driven from a preferably renewable energy source towards the desired site, and the information of the current energy source will be forwarded to the site through the deployed smart meters. This approach will be commonly employed to switch the energy source powering the facility with a greener source when available.

The two approaches are not exclusive and can be used both at the same time. Anyway, in some cases one approach will be preferable to the other. For example, when it would be profitable to use some specific datacenter sites, the follow-the-data approach will be preferable (i.e., forward an energy provisioning request to the smart grid), whereas, if the data or the computing power can be obtained by different equally-cost sites (e.g. as in the case of content distribution networks with replicated data in all the sites), the follow-the-energy approach will be preferable (i.e., routing the data request thorough the greenest sites).

2.4. Energy containment strategies

In such a dynamic and heterogeneous context, it is essential for the cloud computing infrastructure (several datacenters widespread around the world) to be *aware* what is the current energy source that is powering its equipment and possibly request to the smart grid (energy provider) a specific power provisioning (quantity and quality of energy) in order to exploit the energy sources and lower its ecological footprint. Such information is critical to manage and operate the cloud in the greenest way and it will be a requirement in the CO_2 containment strategies that are being approved by the governments, such as Cap&Trade, Carbon Offset, Carbon Taxes and Green Incentive frameworks [4]. In this scenario, two main approaches have been developed to reduce the carbon footprint: carbon neutrality and zero carbon. In the carbon neutrality approach, the industries GHG emissions are partially or totally compensated (hence, *neutrality*) by a credit system (e.g. cap and trade or carbon offset); besides, incentive or tax models are also possible to encourage the use of green sources and limit industry carbon footprints. In the zero carbon approach, green renewable energy sources are employed and no GHG emissions are released at all. Zero carbon is considered to be the only long term viable solution as it does not suffer from the rebound effect: even with increased demand no GHG are emitted at all. Thus, to achieve eco-sustainability, energy-efficiency and energy-awareness should be both exploited in a systemic energy-oriented approach leveraging smart grids employing green renewable energy sources and techniques that range from high level policies to low level technological improvements cooperating with each other. This is a complex task that represents one fundamental step of the challenge that humanity has to face in the XXI century: not only inverting the global warming trend but also achieve sustainable solutions for the decades to come. Here, *sustainability* represents the key word in order to successfully address all these problems.

2.5. ICT energy-efficiency metrics

The huge energy demand originated by datacenters worldwide (approximately quantified in about 1.5% to 2% of global electricity, growing at a rate of 12% annually [8] [9]) is strongly conditioned not only by the power required by the individual runtime, storage and networking facilities that constitute their basic building blocks, but also by cooling (referred to HVAC, heat ventilation and air conditioning), uninterruptible power supply systems (UPS), lighting and other auxiliary facilities. The *power usage effectiveness* (PUE) index, proposed together with the *datacenters infrastructure efficiency* (DCiE) by the GreenGrid [10], are universally recognized metric, used to estimate the energy efficiency of a datacenter by considering the impact of the auxiliary component respect to the basic ones. PUE is defined as the ratio between the total amount of power required by the whole infrastructure and the power directly delivered to the ICT (computing, storage and networking) facilities:

$$PUE = \frac{total\ datacenter\ power}{ICT\ equipment\ power}. \tag{1}$$

DCiE is expressed as the percentage of the ICT equipment power by total facility power:

$$DCiE = \frac{ICT\ equipment\ power}{total\ datacenter\ power} \cdot 100\%. \tag{2}$$

A PUE value of 2 or, equivalently, a DCiE of 50%, can be typically observed by examining most of the current installations [11], demonstrating that HVAC and UPS systems approximately double the datacenter energy needs. Furthermore, the need for redundancy and the use of more sophisticated and expensive energy supply systems, make the things worse in largest and mission critical installations. In this scenario, the cooling facilities represent the most significant *collateral energy drain*. However, their energy efficiency is improving thanks to new cost-containing cooling strategies based on the use of computational fluid dynamics and air flow reuse concepts (free-cooling, cold aisles ducted cooling etc.).

2.6. Energy and the Cloud

As for the ICT equipment, the computing and storage facilities can be considered the most energy-hungry components. As an example, the Barcelona Supercomputing Center (a medium-size datacenter, hosting about 10,000 processors) has the same yearly energy-demand of a small (1,200 houses) town, with a power absorption of 1.2 MW, resulting in an energy bill of more than 1 million Euros [12] [13].

A significant variety of computing equipment populates the state-of-the-art datacenters, ranging from small-sized high-density server systems (in single unit arrangements or blade enclosures) with computational capabilities limited to 2-4 multicore CPUs, to large supercomputers with hundreds of symmetric CPUs/thousands of cores or more complex parallel (e.g. systolic array or vector-based) computing architectures.

Different server architectures can be properly customized for specific computing or service provisioning tasks such as deploying network info-services (HTTP, FTP, DNS or E-Mail servers) or managing large databases. In addition, servers may assume specialized roles within a cloud computing organization by behaving as general-purpose “worker” nodes or as control devices running specialized resource brokering/scheduling systems that manage the dynamic allocation of jobs/applications or virtual machines on the available worker nodes and/or assign, with the role of storage pool managers, the required storage space to them.

Considering that, also within a fairly dimensioned farm, with a limited degree of over-provisioning to handle peak loads, most of the servers operate far below their maximum capacity for most of the time [14] [15], a lot of energy is usually wasted, leaving great space for potential savings to energy-proportional architectures [14] and strategies consolidating underutilized servers. The devices belonging to a server farm are, in fact, usually always powered on also when the farm is currently solicited by a very limited computational/service burden or is totally idle. This consideration can be immediately exploited by a service-demand matching approach, consolidating the current load on a minimum size subset of the available resources, and putting into sleep-mode all the remaining ones, with the immediate effect of greatly reducing the energy consumption, as the one that will be presented in the Section 7. Such a subset can dynamically expand or shrink its dimensions by powering up or down some servers, when necessary, to provide more computing or storage capacity, or reduce the current energy consumption when the load falls under a specific threshold.

2.7. Energy-saving approaches

The energy-saving approaches currently available can be described by the three different "do less work", "slow down" and "turn off idle elements" strategies. In the first one, the applications/jobs requiring services to the cloud are optimized in time and space in order to keep the execution load at a minimum level, resulting in reduced power consumption. The second strategy starts from the consideration that the faster a process runs the more resource intensive it gets. In complex runtime application, the speed of some component processes does not perfectly match and thus several resources remain locked for some time without being really useful. By lowering the speed of the faster activities such unnecessary waiting or resource locks can be avoided by not affecting the overall application completion time. Processes can be slowed down in two ways: they can be run with adaptive speeds, by selecting the minimal required speed to complete the process in time or, alternatively, intermediate buffering techniques can be introduced so that instead of running a process immediately upon arrival, one can collect new tasks until the buffer is full and then execute them in a bulk. This allows for some runtime components to be temporarily switched off resulting in a significantly lower power consumption. Finally, the last strategy refers to the opportunity of exploiting the availability of a low-power consumption status, the *sleep-mode*, of the involved device. That is, the "turn off idle elements" approach aims at switching the devices into sleep-mode during their inactivity periods and restoring them to their fully operational status when more power is needed. If properly used, all the strategies based on the use of sleep-mode, may represent a very useful mean for achieving great power savings in large datacenters, when they are lightly loaded. Such infrastructures are usually designed according to a modular approach, since they are built up by a number of logical-equivalent elements (bulks of servers or computing aisles), so that unloaded bulks can be dynamically put into sleep-mode during low-load periods. Such approaches may be employed in orthogonal dimensions in the sense that they may and should act in concert and simultaneously, multiplying their benefits with respect to a one-dimensional optimization.

3. State of the Art

Several approaches have been proposed in literature in order to contain the energy demand of modern datacenters.

The works presented in [8] [11] [16] [17] focus on the use of sleep-mode to attain energy-efficiency in different ways. In particular, [8] asserts that a really effective strategy to achieve significant power savings can be based on switching off most of the available servers during the night or in presence of a limited load and using the full servers' capacity only during peak hours. On the other hand, between the other approaches proposed, the work in [11] presents a power containment technology based on the use of sleeping mode at the single component level where specific technological features can be exploited to achieve a significant degree of optimization. Another perspective is discussed in [16] analyzing the impact on network protocols for saving energy by putting network interfaces and other components to sleep. The tutorial [17] explains several network-driven power management strategies to dynamically switch computers back and forth from sleep/power-idle mode. In [18], an energy manager for datacenter network infrastructure is presented, dynamically adjusting the set of active network elements (interfaces and/or switches) to accommodate new capacity to support increased datacenter traffic loads. In [19] several ideas are presented: legacy equipment may undergo hardware upgrades (such as modified power supply modules) and their network presence may be transferred to a proxy or agent allowing the end device to be put in low power mode during inactivity periods while being virtually connected to the Internet. The authors also promote the use of renewable energy sources, such as solar, wind or hydro-power, as a valid alternative for supplying power to ICT equipment. Such strategy seems to be well suited for datacenters, which can be located near to renewable energy production sites. However, since renewable energy sources tend to be unpredictable and not always available (e.g., wind), or may present significant variations during day and night (e.g., sun), to really benefit from their usage the involved jobs/applications and the associated data, in case of supply variations, should be migrated from one datacenter to the other, where the energy is currently available, according to the follow-the-sun or chase-the-wind scenario [4]. This implies the presence of an energy-efficient and high capacity communication network always available to support the above facility. Based on similar concepts, a study presented in [20] investigates cost and energy-aware load distribution strategies across multiple datacenters, using a high performance underlying network. The study evaluates the potential energy cost and carbon savings for datacenters located in different time zones and partially powered by green energy and determined that, when striving at optimizing the overall green energy usage, green datacenters can decrease CO_2 emission by 35% by leveraging on the green energy sources at only a 3% cost increase.

This chapter focuses on improving the operating energy efficiency of the computing resources (mainly the available servers) within a complex distributed infrastructure, which are responsible for the greatest part of datacenters energy consumption.

4. Energy-aware datacenter model

A distributed service-provisioning infrastructure, such as a cloud, is composed by datacenters spread throughout the world, eventually belonging to different facilities, acting together as a federated entity.

Each of these cooperating datacenters contains a large number of servers whose main task is running applications, virtual machines or processes submitted by the cloud clients, typically by using the Internet. Each individual server is characterized by a processing capacity, depending essentially on the number of CPUs/cores and on the quantity of random access memory (RAM) available and by a storage capacity dynamically assigned to it by some centralized or distributed storage management system. Thus, the datacenter workload at the time t is measured by considering the applications/jobs that need to be processed or are still running at this time. Usually, datacenters are designed with a significant degree of resource over-provisioning to always reserve some residual capacity needed to operate under peak workload conditions [14] [15] and ensure a certain scalability margin over time. However, also in presence of a limited workload, the servers that are totally idle, since they have no processes to run, are normally kept turned on by adversely affecting the overall power consumption and introducing additional and unnecessary costs in the energy bill.

In such a context, most of the effort should be oriented to the reduction of the active/running servers to a minimal subset and turning off the idle ones, according to the previously presented "turn off idle elements" strategy. For this sake, a high-level energy-aware control logic is needed to dynamically control the datacenter power distribution by exploiting load fluctuations and turning off inactive servers to save energy. Thus, as the current load decreases under a specific attention threshold, a properly designed policy should clearly identify (1) *how many* servers and (2) *which* of them have to be powered down, and the correct procedures to perform this task have to be implemented, ensuring that the physical and logical dependencies among the devices within the datacenter will be always respected (as described in Section 4.1).

In particular, such operating policy has to be implemented by using a proactive algorithm, running within the datacenter resource broker/scheduling logic, that constantly monitors the current runtime load and adaptively determines the subset of servers that can be turned on or off, by using properly crafted actuator functions with the specific task of correctly switch the operating status between on and off and vice-versa.

4.1. Physical and logical dependencies

In modern datacenter farms participating to grid or cloud-based distributed computing infrastructures, a large number of devices are interconnected together into cooperating clusters, which are made up with heterogeneous computing, storage or communication nodes, each with its own role in the farm. Each of these nodes, often known as computing resource broker or Control Element (CE), storage manager/element (SE), disk server (DS), gateway, router, or whatever else – has its own hardware and software features that must be considered when operating in the datacenter. Furthermore, nodes interact among them according to their logical role in the farm and, secondary but not least, to their physical placing, as depicted in Figure 2.

Such logical and physical dependencies must be evaluated, especially in power management operations where devices are switched on and off.

An example of a *power on* procedure executed on the node SE_1 is illustrated in Figure 3, in which the highlighted nodes are turned up, starting from the node UPS_1 and going down to nodes $RACK_1$, $FARM_1$ and eventually SE_1.

Analogously, a *power off* procedure executed on the node SE_1 is depicted in Figure 4, in which the highlighted nodes are turned off, starting from the nodes SS_1 and SS_2 and going up to the nodes $DS_1, \ldots, DS_n$ and eventually to the node SE_1. In this way, the portion of the cluster related to the storage subsystem will be completely turned off and while the remaining part of the cluster will be still working.

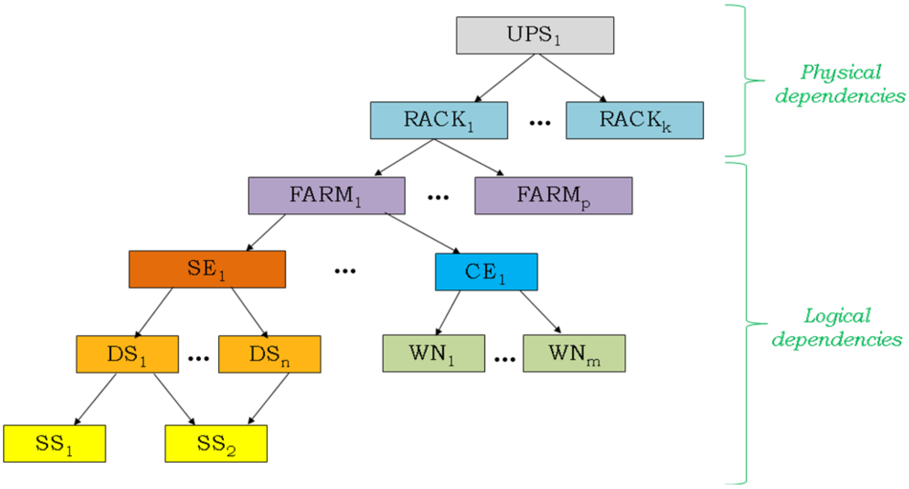


Figure 2: Dependency graph for a grid farm.

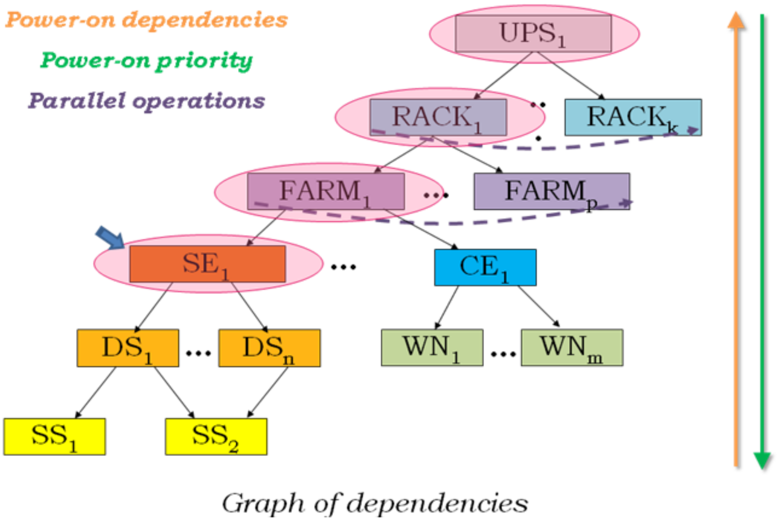


Figure 3: Power on procedure executed on device SE_1.

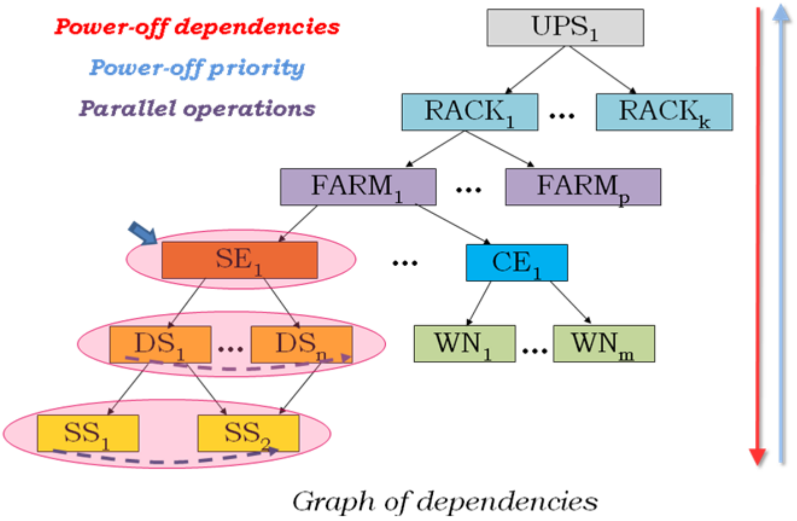


Figure 4: Power off procedure executed on device SE_1.

4.2. Job aggregation strategies

The above model, implicitly assumes that each job/process or virtual machine is assigned to a single computing core, in such a way that each multi-processor server, with n total cores available, is able to run n independent tasks without experiencing any performance slowdown. Accordingly, in presence of multicore systems, a runtime task consolidation strategy can be applied, by moving the current tasks throughout the datacenter farm to aggregate them on a minimal subset of servers, in order to have a greater number of idle servers to be turned off. In presence of multicore CPUs with n cores per server, several consolidation strategies are possible. Such strategies can also be progressively applied as new tasks require service to the datacenter, to avoid expensive re-compaction procedures and complex job migration processes affecting already running applications. For example, in a datacenter using a traditional *first-fit* scheduling strategy, a new task is assigned to the first server, among the available ones, with at least a single core available. As a valid alternative, a *best-fit* strategy aims at compacting the available tasks as much as possible; thus each task is assigned to a server with just one core free (and, thus, $n-1$ already busy) if any such server exists. Otherwise, it looks for a server with only two free cores, then with three, and so on, up to n, according to the principle of distributing the load

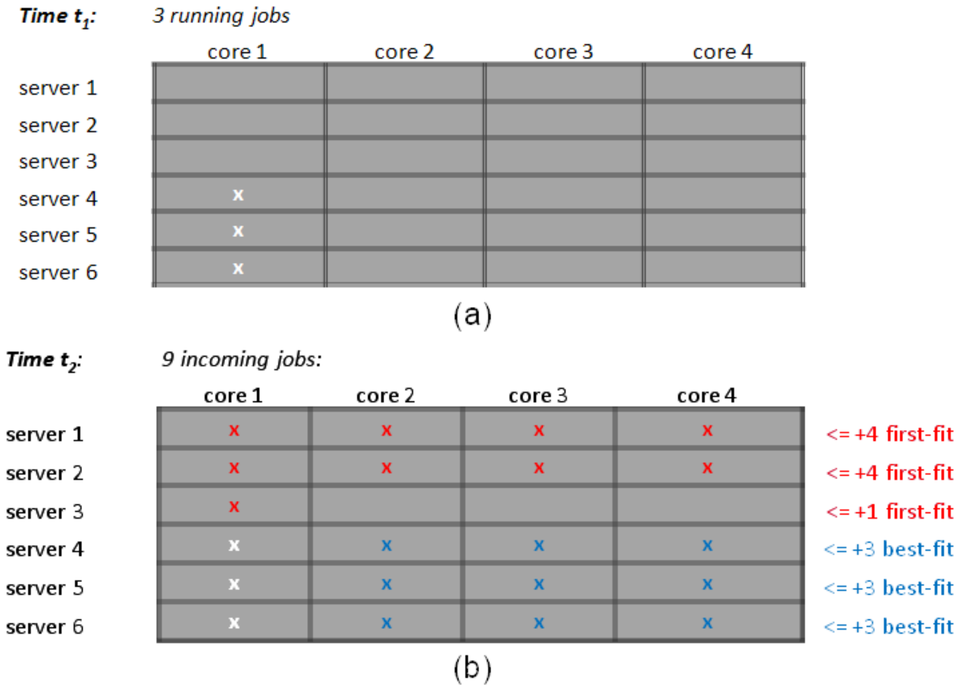


Figure 5: Visualization of first-fit and best-fit allocation strategies for a subset of 6 servers with 4 cores each. At time t_1, there are three running jobs arranged (as result of previous allocations) as in (a); at time t_2, nine new jobs arrive (b). First-fit will use up to 2x more servers than best-fit.

on the servers that are already the most loaded ones in an effort to totally saturate their available capacity. Note that, in this aspect, the optimization of datacenters differs from the one of telecommunication networks, in which the load balancing criteria has to try to *not* saturate the available resources (e.g., fiber links) in order to leave enough “space” (e.g., bandwidth) for future connection requests to come.

Clearly, the first-fit scheme is faster, but it has the disadvantage of leaving a large number of servers only partially loaded. On the other hand, best-fit gives the more satisfactory results according to the aforementioned consolidation strategy, since it compacts the jobs as much as possible on a few servers and leaves the maximum possible number of servers totally unloaded so that they can be immediately powered down, with a significant reduction of the wasted energy (see Figure 5). Besides achieving optimal compaction, the best-fit strategy is also more profitable since a multicore server with a significant number of busy cores is statistically less likely to get free of all his runtime duties (and, thus, of being put into sleep-mode) than a server with a low number of jobs. The inherent computational complexity of the best-fit strategy may be improved to work in a constant amortized time by implementing per-server priority queues with Fibonacci heaps, so that it will not introduce additional burden to the overall computing facility. Unfortunately, in presence of single-core devices, no further aggregation is possible and hence energy savings can be achieved only by powering down the idle servers.

5. Traffic fluctuation

The datacenter workload is usually characterized by recurring fluctuation phenomena where higher utilization periods (e.g., during some hours of the day) are followed by lower utilization ones (e.g., during the night) and so on. Due to the regularity of these recurrence phenomena, driven by the 24 hours or weekly rhythm of human activities, the aforementioned fluctuations are typically predictable within certain fixed time periods (e.g., day/night or working day/weekend cycles, months of the years, etc.) and they can be described by a pseudo-sinusoidal trend [18] [21]. The theoretical daily workload variation for a typical energy-unaware datacenter [18] is shown in Figure 6.

It can be immediately observed that, while the demand load follows a pseudo-sinusoidal trend, the power drained remains almost constant during both the high and low usage periods. This is essentially due to the impact of computing resources that are always kept up and running also when they are underutilized or not utilized at all, thus wasting a large amount of energy during the low load periods. The fundamental idea behind a more energy-conscious datacenter resource management is introducing *elasticity* in computing power provisioning under the effect of a variable demand, by adaptively changing the datacenter capacity so to follow the current demand/load, as shown in Figure 7. This can be accomplished by dynamically managing, at the resource scheduling level, the allocation of tasks to the available servers, and putting the unused ones in low-power sleep-mode and rapidly resuming a specific block of sleeping servers when more capacity is required and the currently operating servers pool is not able to accommodate it.

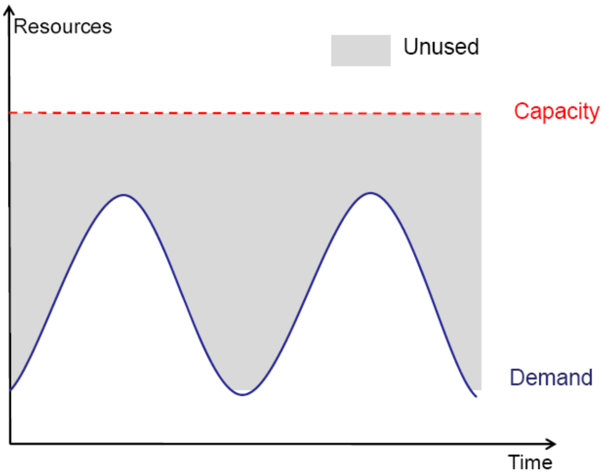


Figure 6: Capacity-demand mismatch leads to resource and energy wastes.

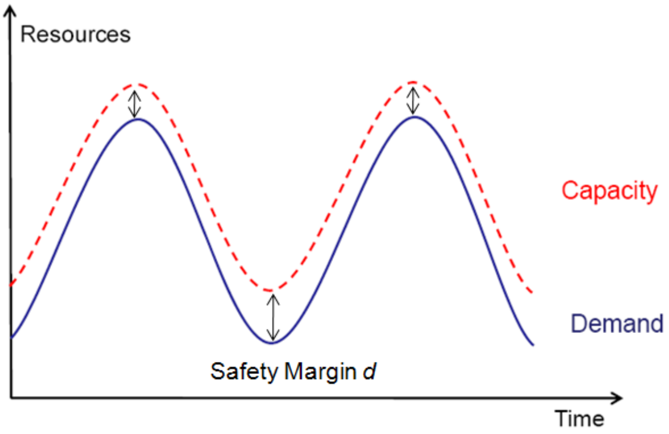


Figure 7: Theoretical provisioning elasticity concept. The capacity curve should resemble the demand curve as close as possible, leaving a safety margin to serve possible peak loads.

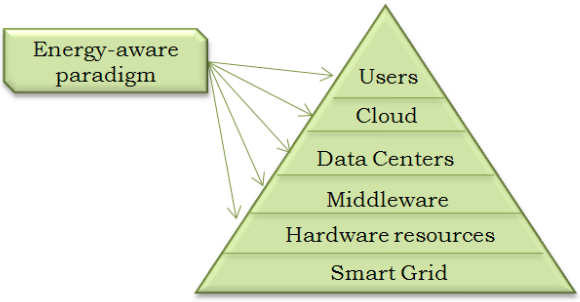


Figure 8: Energy-awareness multilevel approach overview.

6. Energy-oriented optimization

Since a distributed computing infrastructure, such as a cloud, can be seen both as a single federated entity but also as a set of autonomous datacenters working together, energy-oriented optimization can be introduced at different levels. Energy-efficiency is the first step towards reducing the ecological footprint of cloud infrastructures. It is related to the "do more for less" paradigm, and aims at using more efficient computing or manufacturing techniques that lowers the components power requirements. At higher level of abstraction, it is possible to introduce the energy-awareness at all levels, meaning that the device (from the hardware resources to individual servers up to the datacenter and the whole grid/cloud) is *aware* of its power requirements, and adapts its behavior to its current load and even to the current energy source that is feeding it with energy (as sketched in Figure 8). Energy-awareness is that it can be applied to all levels, from a local datacenter basis (*intra-site optimization*) to a wider cloud scope (*inter-site optimization*).

6.1. Energy-efficiency

From the lowest granularity, energy-efficiency can be implemented at the devices level, designing and building more energy-efficient components that decrease the power consumption while not disrupting the offered service level, or even increasing it. As an example of technological innovation and energy efficiency, consider the ASCI Red supercomputer of Sandia National Laboratories (in Albuquerque, New Mexico, US), built in 1997. With its peak performance of 1.8 teraflops, it has been the most powerful supercomputer in the world (according to TOP500) between June 1997 and June 2000, and occupied a physical size of 150 square meters, with a power absorption of 800,000 Watts. Just nine years later, in 2006, the Sony Playstation 3 achieved the same peak performance with a power absorption lower than 200 watts and a physical size of 0.08 square meters [13].

As another example, consider the Apple iMac in the year 2000: it had the dimensions of a CRT monitor, the weight of 15.8 kg and its technical specification were: 500 MHz CPU, 128 MB of RAM and 30 GB of storage, with a power consumption of about 150 Watts. In 2010, ten years after, Apple released the iPhone 4, which, despite of its handy-size and 137 g of weight, had a 1 GHz CPU, 512 MB of RAM and 32 GB of storage capacity, with a power absorption of just 2 Watts. Furthermore, it added also advanced functionalities, like GPS, Wi-Fi, Compass, mobile phone, etc., in line also with the“do more for less” paradigm.

The energy-efficiency in the design of the new hardware has more or less been always present in the industry processes, and it remains one of the most important issues in reducing the energy requirements. Nonetheless, it is not sufficient to have more and more energy-efficient components, for two main reasons. From one side, the energy-efficiency alone cannot cope with the ever increasing needs of energy and the consequent carbon footprint. From the other side, increased efficiency can lead to the the rebound effect, limit or even making worse the ecological footprint.

6.2. Virtualization and thin clients

From the user point-of-view, the ICT market is evolving towards a new network-centric model, where the traditional energy-hungry end-user devices (e.g. PCs, workstations, etc.) are being progressively replaced by mobile lightweight clients characterized by moderate processing capabilities, limited energy requirements and high-speed ubiquitous network connectivity (e.g. smart phones, handhelds, netbooks, etc.), significantly incrementing the strategic role of the network for connecting them to the cloud infrastructures. These devices typically do not directly run complex and expensive applications but rely on *virtual machines* and remote storage resources residing on distributed cloud organizations, accessible from the Internet, to accomplish their more challenging computing tasks, thus limiting their roles to very flexible and high level interfaces to the cloud services. Such an evolution in users habits, if properly managed, can be exploited to significantly reduce the power consumption of both datacenters and user-level computing equipment.

Modern Virtualization [16][19] technologies may play an fundamental role by dynamically moving, in a seamless way, entire virtual machines and hence their associated computations, between different datacenters across the network, in such a way that the datacenters placed near renewable energy plants may execute the involved computational demands with a lower energy cost and a reduced carbon footprint with respect to traditionally power consuming (and almost idle) PCs statically located in the users' premises. Increasing the computing/runtime density and power in sites where green and/or less expensive energy is available, will be the upcoming challenge for next generation datacenters.

Virtualization can be applied on two levels. Firstly, multiple virtual machine instances can be shared on a single physical server, by reducing the number of servers and thus the power consumption. Secondly, scalable systems using multiple physical servers can be built, allowing most servers to be switched off during low usage periods, and only using the full capacity of the computing farm during peak hours.

Resource sharing may also play an important role both for datacenters and network equipment. In datacenters, virtualization may be exploited by an energy-aware middleware that dynamically moves individual tasks or virtual machines to the most energy-convenient high-density sites (those characterized by the lowest carbon footprint or energy costs) in order to increase, as much as possible, their computational burden. From the networking perspective, an energy-aware control plane may properly route the traffic/connections associated to the above task/VMs by privileging intermediate nodes currently fed with renewable energy while simultaneously grouping the above connections on the same path instead of spreading them over the whole network. This management practice will maximize the usage of the already active devices/paths and consequently save the energy resulting from temporarily powering down the networking devices that should serve the alternate/secondary and no more useful paths.

7. Intra-site optimization

In line of principle, the instantaneously available capacity in a datacenter should closely follow the current load (Figure 7). However, the instantaneous capacity function cannot be described by a *continuous* curve, but is instead a step function in which each step is associated to a variation in the quantity of *discrete* resources (e.g., single servers or blocks of servers in the farm)

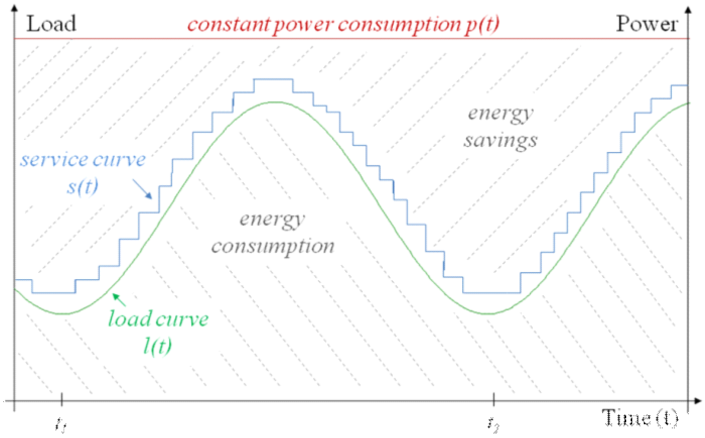


Figure 9: Service-demand matching.

being turned on or off. Hence, the demand curve must be approximated with a step-shaped service curve that is always able to support the current demand/load while minimizing the overall energy consumption (Figure 9). Since the power drained is directly proportional to the number of active servers, the closer the service curve approximates the demand one, the lower will be the power wasted.

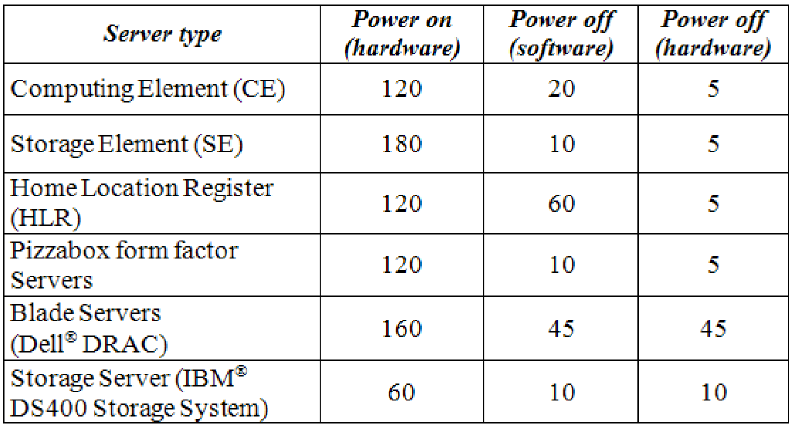
Furthermore, the instantaneously available capacity should be dimensioned by ensuring the presence of a safety margin (i.e., a minimal distance d to be always maintained between the values assumed by the demand and the capacity curves, expressed as number of servers or cores) to properly handle bursty traffic phenomena or peak loads. This margin is directly associated to a certain number of servers that are preventively kept turned on to serve new incoming tasks whose number cannot be easily foreseen by using statistical observations. Reducing the distance margin d, directly implies decreasing the energy consumption, but also introduces, as a side effect, an increment in the average service delay for incoming tasks that, since the number of new jobs that can be immediately served without reactivating previously sleeping devices decreases proportionally with the safety margin. Conversely, larger values for the distance d, introduce more tolerance to extemporaneous load variations, thus a greater number of task can be immediately served as they arrive (with no delay), but, obviously, the energy consumption increases proportionally (since more servers have to be kept up and running). Thus, the safety margin d has to be large enough to avoid unnecessary oscillations of a certain group of servers back and forth from the sleep-mode, in correspondence to limited variations of the load. This is also required to avoid, as possible, the peaks in the power absorption that are usually experienced during the start up phase or when switching from the sleep-mode status to the fully operating one. Therefore, the safety margin d can be considered as *upper-bounded* by the energy consumption and *lower-bounded* by the peak load absorption capacity together with the minimization of the aforementioned power fluctuation requirement.

Choosing a safety margin value of d, ensures that a bulk of $k \leq d$ incoming task can be immediately served without waiting. Thus, the d parameter also identifies the size of the zero-waiting queue of the tasks that are served as they arrive. If $k > d$, there will be $k - d$ jobs that will have to wait a time t before they can get served, where t is the start-up time of the servers (obviously, if the load reaches the site maximum capacity, all the new tasks will have to wait for new resources to become available). The start-up time t may sensibly vary with the available technologies. In presence of *agile* servers equipped with enhanced sleep-mode capabilities, the value of t may range in the order of a few ms, whilst for less sophisticated legacy equipment a complete bootstrap procedure will be required and the start-up time t may grow up to some minutes. As a general rule, the higher is the t value, the higher has to be the safety margin d, and consequently the lower will be the energy saving, whereas lower values of t, allow significantly greater energy savings margins. In Figure 10 we reported the (software and hardware) power off and power on times measured in the INFN[5] Tier2 site of the CERN[6] LHC[7] experiment. When enhanced sleep-mode is not employed, servers need several tens or even hundreds of seconds to switch their state from turned down to up and running, under the control of specific *wake-on-LAN* or external power management facilities (e.g. “intelligent” power distribution units, PDU). Such high times clearly indicate that the enhanced sleep-mode feature is strongly required to make modern datacenters agile and may bring great benefits in terms of reduced energy waste and consequent electrical bills.

[5] Italian National Institute for Nuclear Physics.

[6] European Organization for Nuclear Research.

[7] Large Hadron Collider.



Server type	*Power on (hardware)*	*Power off (software)*	*Power off (hardware)*
Computing Element (CE)	120	20	5
Storage Element (SE)	180	10	5
Home Location Register (HLR)	120	60	5
Pizzabox form factor Servers	120	10	5
Blade Servers (Dell® DRAC)	160	45	45
Storage Server (IBM® DS400 Storage System)	60	10	10

Figure 10: Complete power ON/OFF times (seconds) for different legacy devices.

7.1. Analytically evaluating the energy saving potential

In order to evaluate, from the analytical point of view, the upper-bound for the energy saving potential of the discussed service-demand elasticity approach, we can consider instantaneous transitions among the sleep and the active states ($t = 0$) and theoretical sinusoidal traffic, like the one depicted in Figure 6. The demand curve represents the request for service load experienced during the day, while the service curve represents the servers that need to be fully operating to process the job requests. Without the introduction of any energy saving policy, the power consumption of the datacenter remains constant [18] over the entire observation interval, and the energy required is represented by the integral of the power drained over time:

$$\int_{t_1}^{t_2} p(t)dt, \tag{3}$$

where $p(t)$ is the power consumption function and t_1 and t_2 are the considered the extremes of the observation time interval. Ideally, the lower bound for the datacenter energy consumption is defined as:

$$\int_{t_1}^{t_2} l(t)dt, \tag{4}$$

where the $l(t)$ function describes the demand/load curve. Such a curve can be closely approximated by the adaptive service curve $s(t)$, which is the step function that establishes the minimum set of runtime resources that have to be powered on to serve the current demand. Therefore, with the proposed energy saving schema, the theoretical energy consumption is defined as:

$$\int_{t_1}^{t_2} s(t)dt. \tag{5}$$

Clearly, between the above equations it holds that (4) < (5) << (3), and the bigger the difference between the values assumed by eq. (3) and eq. (5), the greater the energy saving. Theoretically, the energy saving is upper-bounded by:

$$\int_{t_1}^{t_2} (p(t) - l(t))\, dt, \tag{6}$$

while the actual energy saving is defined as:

$$\sum_{i=1}^{n} (p(i) - s(i)) \cdot \Delta_i, \tag{7}$$

where n is the number of intervals in which the time interval $[t_1, t_2]$ is partitioned and Δ_i is the duration of the i-th time interval; note that the value n defines the time-basis on which the optimization process is executed; thus eq. (7) represents the potential energy savings that can be achieved.

7.2. The service-demand matching algorithm

Given a specific demand/load curve, the service-demand matching algorithm determines the service curve that is able to always satisfy the current demand while minimizing the number of active servers and hence the power needed. As an example, consider a scenario in which the demand curve increases between the times t_i and t_{i+1}. As a consequence, the distance from the service curve decreases from d_i to d_{i+1}. Since $d_{i+1} < d$, the algorithm detects the increment in the demand (totally absorbed by the safety margin d, thus no service delay occurs in this case) and consequently increases the number of active servers by turning on $s_{i+1} - s_i$ servers.

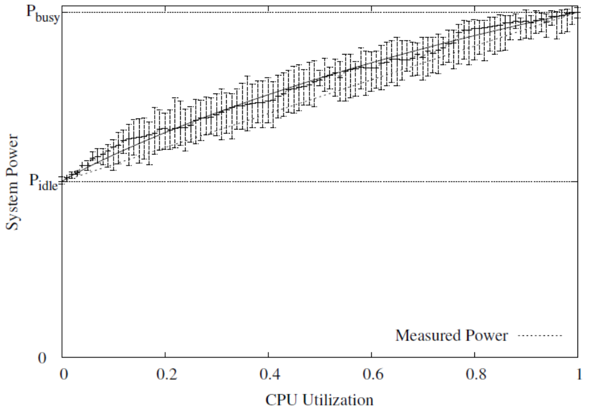


Figure 11: Server energy model: the power consumption varies linearly with the CPU load.

At the opposite situation, when a decrement $d_{i+1} > d$ of the demand curve occurs, it causes the algorithm to decrease the service curve from s_{i+1} to s_i, thus allowing more devices to be put into sleep-mode.

7.3. Experimental Evaluation

To evaluate the effectiveness of the service-demand matching algorithm (best-fit aggregation with sleep-mode during low load periods) in terms of energy saving potential, a large datacenter simulation model, composed by more than 5,000 heterogeneous single-core and multicore servers, has been built starting from data available in literature [14] [15] concerning the observation over a six-month period of a large number of Google servers.

The servers power consumption was modeled as depicted in Figure 11, in which the server always consumes a fixed power needed for the device to stay on, and a load-dependent variable power, equal at maximum to the fixed part, scaling linearly with the CPU load [15].

The jobs arrive with a pseudo-sinusoidal trend as illustrated in Figure 6. The duration of the jobs was taken to be exponentially distributed [22], extrapolated from a duration of 1 to a maximum of 24 hours, according to the distribution reported in Figure 12.

First, we evaluated the effectiveness of the approach considering non-zero transition times (t) between the on and the sleep-mode states (taking as a reference the value of Figure 10). Day one of simulation is depicted in Figure 13. At the beginning, no job is present in the datacenter. As time

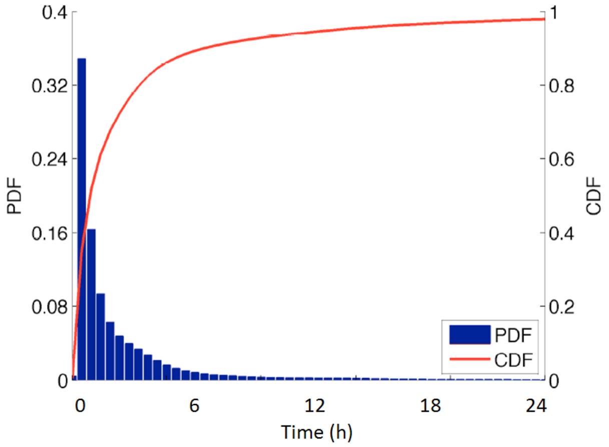


Figure 12: Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs) of jobs duration.

passes, we monitor the energy consumption with all devices on (i.e., without our approach, red area) with the energy consumption of the approach (blue area), and plot the obtained energy savings (green area). Note that, when the servers are turned on but idle (during the first 12 hours approximately), they consume about half of their maximum power consumption, which is in turn achieved when the CPU works at 100% (Figure 11), according to the energy model of Figure 11.

In Figure 14 we reported the generic day n of simulation. We can observe that a number of CPU (set by the safety margin d) is kept on even if idle to serve future jobs that may arrive at the datacenter. Anyway, a peak of traffic can still cause that a number of jobs will have to wait, as observed in the Figure 15.

Therefore, we evaluated the impact of variations in the safety margin d against the potential savings, in terms of reduced energy consumption (MWh), GHG emissions (tons of CO_2) and economical cost (Euros).

For a commercial/industrial facility like a datacenter, we assumed that the average cost of energy is about 0.12 Euros per kWh [23], and we considered fossil-fueled energy plants powering the datacenters, which emits 890 grams of CO_2 per kWh [9].

Several simulation experiments have been run with different values of the safety margin d. The observed results show that the maximum achievable cost savings may reach about 1.5 millions Euros, with a reduction of more

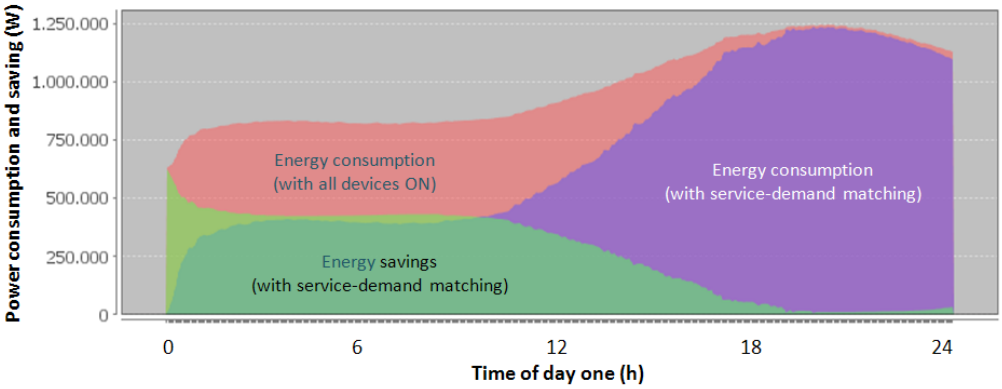


Figure 13: Energy consumption of day one with and without the service-demand matching algorithm.

than 13 GWh in the energy consumption and more than 11 kTons of CO_2 in the GHG emissions. These results are not surprising since the servers are rarely utilized at their maximum capacity and operate, for most of their time, at an average load ranging between the 10% and 50% of their maximum utilization levels [14].

As expected, the d value significantly affects the overall energy savings as long as the consequent CO_2 emissions and electric bill costs. The best results have been achieved with lower values of the safety margin. When evaluating the impact of the safety margin, since the basic goal of the experiments was to provide a lower bound for the energy savings of modern and future datacenter, the transition time t between the powered on and off states have been put to 0, so that switching the servers between the sleep and operating mode introduces no delay (i.e. all the considered servers are agile). As a consequence, the frequency of the load variations (i.e., how and how often the traffic load varies in time) only affects the number of transitions between the on and off (or sleeping and operating) states, but it does not influence the energy savings at all, as each variation is immediately followed by the corresponding power mode switching action on the involved servers.

The efficiency that can be achieved in resource utilization may reach values between 20% and 68%, meaning that an high percentage of the servers can be put into sleep-mode for a considerable time. The saving margins decrease almost linearly as the d values increases (see Figure 16). In fact, while the load is far from the actual datacenter capacity, both the achieved energy

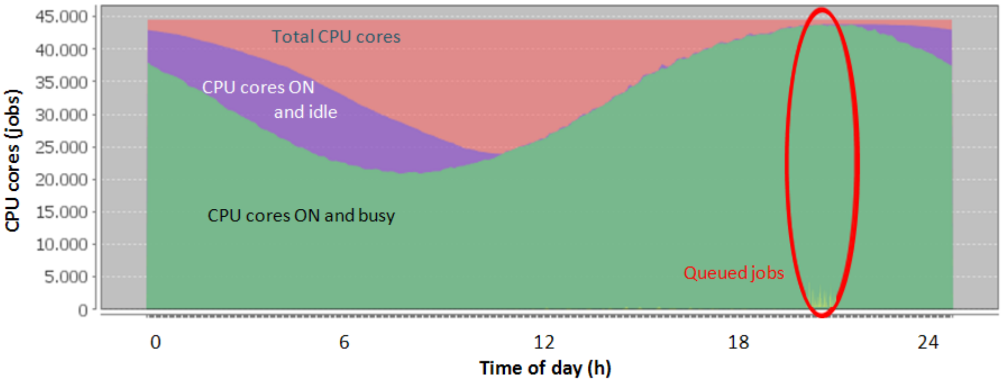


Figure 14: Energy consumption of day n with and without the service-demand matching algorithm.

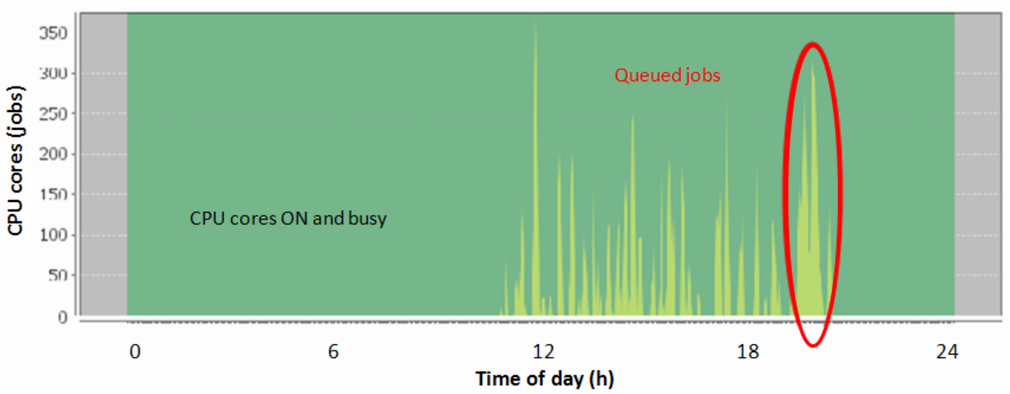


Figure 15: Queued jobs that have to wait due to a peak in the traffic load.

savings and the safety margin d vary linearly but, as the load approaches higher values, the d threshold will exclude a higher number of devices from being switched down, leading to relatively lower savings. When considering multicore devices, job aggregation is possible. When comparing the two aggregation/consolidation strategies based on first-fit and best-fit scheduling of new tasks in Section 4, we observed that first-fit performed significantly worse than best-fit (up to 50%), so that in evaluating the consolidation strategy we only focused on the best-fit task scheduling scheme.

Finally, varying the number of cores per servers shows a common behavior: the more the cores available in the datacenter, the higher the energy

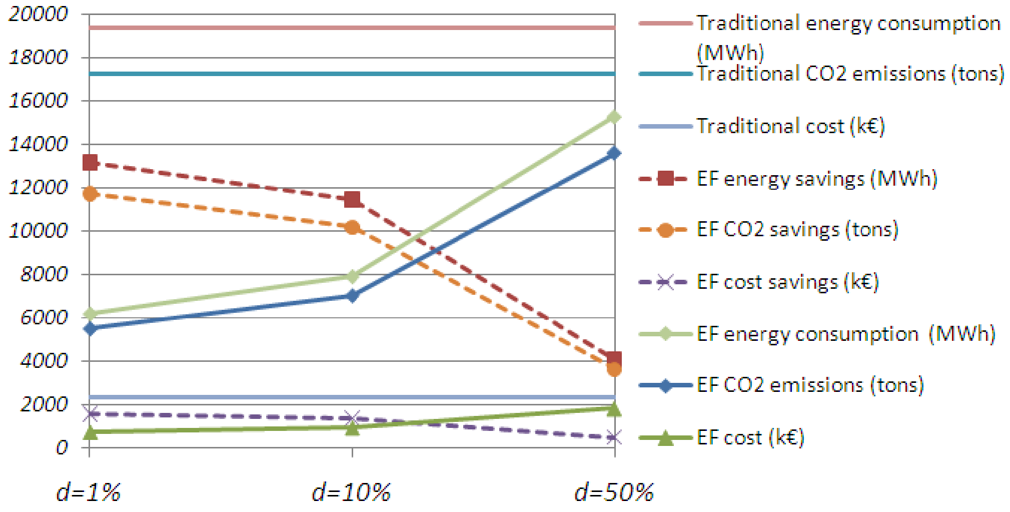


Figure 16: Energy, CO_2 and costs with varying safety margins (d) for a large datacenter (5,000 servers).

consumption. This is due to the fact that the datacenters usually operates very far from their actual maximum utilization capacity and this causes multicore servers to run with only few jobs even when using the best-fit scheduling strategy. That is, multicore servers are more affected by internal fragmentation phenomena (not all the cores are always busy). Thus, in presence of a limited load, assigning a task to a single-core server costs less than executing it on a multicore server (due to the greater energy consumption of the latter device), whilst, at higher loads, the greater computing density of multicore servers may be exploited by the best-fit strategy to lower the overall datacenter energy consumption.

8. Inter-site optimization

When dealing with inter-site optimization, new opportunities are added to the discussed intra-site optimization. Being the cloud distributed over the territory, even with large distances (two sites of the same confederation may be in different continents), new perspectives are possible for reducing the energy, the GHG emissions and the energy costs, especially in conjunction with the discussed smart grid infrastructure. In this sense, a *follow-the-energy* approach [4] and the knowledge of the *current* electricity prices in the different regions on which the cloud infrastructures spans may be taken into account and exploited by an energy-aware task scheduling paradigm to optimize the overall energy consumption (see Figure 17) and/or the use of renewable energy sources. This requires the presence of a high performance communication infrastructure, joining the cloud sites and supporting the capability of dynamically moving the cloud tasks from a site to another one with minimum latency, according to a fully distributed scheduling paradigm.

For example, since electricity prices may present significant geographic and temporal variations, due to differences in the actual regional demand or in the use of a cheaper energy source, by working on the energy cost dimension, significant reductions can be achieved by adaptively moving and rescheduling the running tasks from a cloud site to other locations, offering runtime services to the cloud, where the electricity prices and associated taxes are the lowest on particular hours of the day.

Energy-oriented scheduling criteria may also be introduced in multi-site clouds by optimizing the choice of energy sources in such a way that runtime resources located within sites powered by renewable sources are always preferred when available. In fact, datacenters may be equipped with a dual power supply: the always-available power coming from dirty energy sources and the not-always-available power coming from renewable energy sources. Consider, for instance, the availability of energy produced by solar panels; it is strongly correlated with the time of the day, since it is known that no energy will be produced during the night and that some energy is expected to be produced during the day. Such knowledge should be included in the distributed scheduling logic, by implementing automatic follow-the-sun or chase-the-wind paradigms. Accordingly, the aforementioned scheduling system should have the additional goal of distributing the available power to the incoming new tasks or to the already running ones to dynamically follow the 24-hours daylight cycle. Thus, when the sun is shining in a specific geographic area, all the incoming new jobs should be scheduled on the sites located in such area that are entirely powered by the solar energy. Analogously, the tasks already running in sites that are powered by dirty energy sources should be eventually forced to pass through the underlying network infrastructure to be relocated on sited powered by the solar energy. Of course, when the day light is no more available, the above tasks should be dynamically moved back elsewhere, e.g. where the sun is now shining, in such a way that the use of green energy sources is maximized and the carbon footprint minimized.

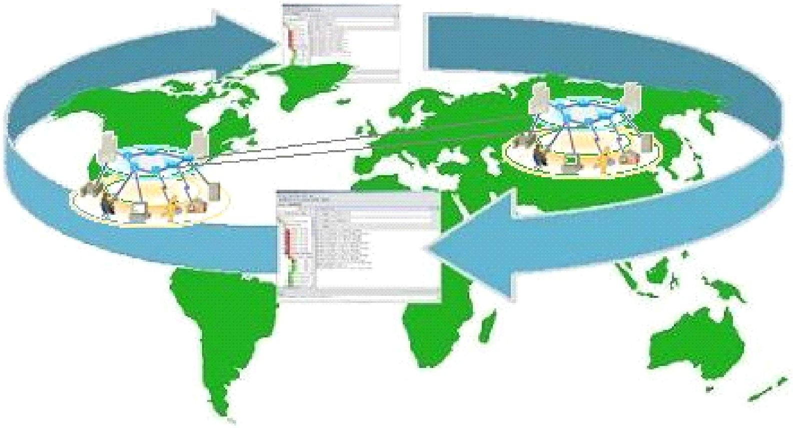


Figure 17: Follow-the-whatever paradigm ad inter-site optimization schema in the cloud infrastructure.

As another example, we can imagine some sites powered by wind energy where power supply is a pseudo-random process depending on the availability of wind. Due to the inertia of the power generating mechanisms and batteries, a drop in the wind power does not result immediately in a power generation drop. Hence, if wind stops, it is possible to dynamically reconfigure the load over the cloud, to consider the new distribution of available clean energy and re-optimize its carbon footprint. Differently from the case of the daylight, whose duration is known in advance, a decrease in wind strength is much more unpredictable and the warning time is shorter. This should be only handled with adaptive and efficient task migration mechanisms implemented within the cloud middleware. For this reason, it is necessary to develop novel migrating schemes and resource allocation mechanisms that take advantage of the early notification of the forecast power variation of clean sources with time-varying power output. Furthermore, another interesting perspective in energy-aware job allocation comes from linking job scheduling/dispatching to the different available electricity prices, dynamically and continuously moving data to areas/devices where electricity costs are lower.

9. Conclusions

In order to support all the above adaptive behaviors, energy-related information associated to the datacenters belonging to the cloud need to be introduced as new constraints (in addition to the traditional ones, e.g. computing and storage resources details) in the formulations of dynamic job allocation algorithms. Down-clocking or sleep-mode should be handled as new capabilities of the datacenter equipment that need to be considered at the cloud traffic engineering layer, and the associated information must be conveyed to the various devices within the same energy-management domain. This clearly requires modifications to the current protocols and middleware architecture. However, in many cases, the carbon footprint improvements will be achieved at the expense of the overall performance (e.g. survivability, level of service, stability, etc.), which can in turn be compensated through over-designing (increase of CAPEX) or over-provisioning (increase of OPEX). In fact, by putting energy-hungry equipment or components into low power mode, or creating traffic diversions driven by reasons different from the traditional load balancing ones, we implicitly reduce the cloud available capacity and hence the experienced delays tend to be longer and/or datacenters more congested, decreasing the overall service quality. This implies that the new algorithms empowering the energy-aware middleware should be driven by smart heuristics that always take into account the trade-off between performance and energy savings.

References

[1] J. Torres, Empreses en el nuvol, Libros de cabecera, ISBN: 978-84-938303-9-7, 2011.

[2] United Nations General Assembly Report of the World Commission on Environment and Development: Our Common Future. Transmitted to the General Assembly as an Annex to document A/42/427 - Development and International Co-operation, 1987.

[3] SMART 2020: Enabling the low carbon economy in the information age, The climate group, 2008.

[4] B. St Arnaud, "ICT and Global Warming: Opportunities for Innovation and Economic Growth", [online]. Available: http://docs.google.com/Doc?id=dgbgjrct_2767dxpbdvcf.

[5] Living Planet Report 2010, The biennial report, WWF, Global Footprint Network, Zoological Society of London, 2010.

[6] Christopher J. Koroneos, Yanni Koroneos, Renewable energy systems: the environmental impact approach, International Journal of Global Energy Issues 27(4), pp. 425-441, 2007.

[7] John O. Blackburn, Sam Cunningham, Solar and nuclear costs – the historic crossover, NC WARN: Waste Awareness and Reduction Network, www.ncwarn.org, July 2010.

[8] J. Koomey, "Estimating Total Power Consumption by Servers in the U.S. and the World", [online]. Available: http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf.

[9] BONE project, "WP 21 Topical Project Green Optical Networks: Report on year 1 and updated plan for activities", NoE , FP7-ICT-2007-1 216863 BONE project, Dec. 2009.

[10] The Green Grid, "The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE", Technical Committee White Paper, 2008.

[11] W. Vereecken, W. Van Heddeghem, D. Colle, M. Pickavet, P. Demeester, "Overall ICT footprint and green communication technologies", in Proc. of ISCCSP 2010, Limassol, Cyprus, Mar. 2010.

[12] Jordi Torres, "Green Computing: the next wave in computing", Ed. UPCommons, Technical University of Catalonia (UPC), Feb. 2010.

[13] Peter Kogge, "The tops in flops", IEEE Spectrum, pp. 49-54, Feb. 2011.

[14] L.A. Barroso, L. A., Hlzle, U., "The Case for Energy-Proportional Computing", IEEE Computer, vol. 40, 33-37, 2007.

[15] X. Fan, W. dietrich Weber, L. A. Barroso, Power provisioning for a warehouse-sized computer, in: Proceedings of ISCA, 2007.

[16] M. Gupta, S. Singh, "Greening of the internet", in Proc. of the ACM SIGCOMM, Karlsruhe, Germany, 2003.

[17] K. Christensen and B. Nordman, "Reducing the energy consumption of networked devices", in IEEE 802.3 tutorial, 2005.

[18] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, N. McKeown, “Elastictree: Saving energy in data center networks”, in Proceedings of the 7th USENIX Symposium on Networked System Design and Implementation (NSDI), San Jose, California, USA, pp. 249–264, ACM, 2010.

[19] W. Van Heddeghem, W. Vereecken, M. Pickavet, P. Demeester, “Energy in ICT - Trends and Research Directions”, in Proc. IEEE ANTS 2009, New Delhi, India, Dec. 2010.

[20] K. Ley, R. Bianchiniy, M. Martonosiz, T. D. Nguyen, “Cost- and Energy-Aware Load Distribution Across Data Centers”, SOSP Workshop on Power Aware Computing and Systems (HotPower ’09), Big Sky Montana (USA), 2009.

[21] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, “Above the Clouds: A Berkeley View of Cloud computing”, Technical Report No. UCB/EECS-2009-28, University of California at Berkley, USA, Feb. 2009.

[22] D. Meisner, T. F. Wenisch, Stochastic queuing simulation for data center workloads, in: EXERT: Exascale Evaluation and Research Techniques, workshop held in conjunction with ASPLOS, Mar. 2010.

[23] U.S. Energy Information Administration, State Electricity Prices, [online]. Available: http://www.eia.gov/electricity/state/.

Appendix C

List of acronyms

ADSL Asymmetric Digital Subscriber Line

AWG Arrayed Waveguide Grating

CAPEX Capital Expenditure

CDN Content Delivery Network

CE Computing Element

CERN European Organization for Nuclear Energy

CMOS Complementary Metal-Oxide-Semiconductor

CO_2 Carbon Dioxide

CR-LDP Constraint-based Routing Label Distribution Protocol

DNS Domain Name System

DPM Disk Pool Manager

DSF Dispersion Shifted Fiber

DWDM Dense WDM

DXC Digital Cross Connect

ECR Energy Consumption Rating

ECRW energy Consumption Rating Weighted

EDFA Erbium Doped Fiber Amplifier

FDL Fiber Delay Line

FTTH Fiber To The Home

GHG Green House Gases

GLASS GMPLS Lightwave Agile Switching Simulator

GMPLS Generalized MultiProtocol Label Switching

GPON Gigabit-capable Passive Optical Network

GRASP Greedy Randomized Adaptive Search Procedure

HVAC Heating Ventilation and Air Conditioning

ICS Information and Communication Society

ICT Information and Communication Technologies

INFN Italian National Institute for Nuclear Physics

IEEE Institute of Electrical and Electronics Engineers

ILP Integer Linear Programming

IP Internet Protocol

IS-IS Intermediate System to Intermediate System

ISO International Standard Organization

ITU International Telecommunication Union

LCA Life Cycle Assessment

LSA Link State Advertisement

MEMS Micro-Electro-Mechanical Systems

MILP Mixed Integer Linear Programming

MPLS Multi-protocol Label Switching

NE Network Element

NZ-DSF Not-Zero Dispersion Shifted Fiber

O-E-O Optical-Electrical-Optical

OADM Optical Add and Drop Multiplexer

OBS Optical Burst Switching

OCS Optical Circuit Switching

OPEX Operational Expenditure

OPS Optical Packet Switching

OSI Open Systems Interconnection

OSNR Optical Signal-to-Noise Ratio

OSPF Open Shortest Path First

OSPF-TE Open Shortest Path First with Traffic Engineering

OTN Optical Transport Network

OXC Oprical Cross Connect

PUE Power Usage Effectiveness

PLI Physical Layer Impairment

QoS Quality of Service

RSVP ReSource reserVation Protocol

RSVP-TE ReSource reserVation Protocol with Traffic Engineering

RWA Routing and Wavelength Assignment

SE Storage Element

SOA Semiconductor Optical Amplifier

TDM Time Division Multiplexing

TLV Type-Length-Value

UPS Uninterruptible Power Source

VLSI Very Large Scale Integrated

WC Wavelength Converter

WCC Wavelength Continuity Constraint

WDM Wavelength Division Multiplexing

WN Working Node

WSS Wavelength Selective Switch

XML eXtensible Markup Language

Acknowledgements

I would like to acknowledge the COST Action IC0804 on Energy Efficiency in Large Scale Distributed Systems, the Spanish Ministry of Science and Innovation under the DOMINO project (TEC2010-18522), the Catalan Government under the contract SGR 1140 and the DIUE/ESF under the grant FI-201000740.