Experiences deploying a distributed parallel processing environment over a broadband multiservice network

Javier Corbacho-Lozano, Oscar-Iván Lepe-Aldama[1], Josep Solé-Pareta, Jordi Domingo-Pascual

{corbacho, oscar, pareta,jordid}@ac.upc.es
Universitat Politècnica de Catalunya, Spain

Abstract. This paper addresses part of the work that is being carried out within the SABA project. We are deploying and using a virtual parallel machine over a network-of-workstations, which communicates by means of an ATM based broadband multiservice network. The objective is to asses, by experimentation, the effect on the performance of this distributed parallel processing environment produce by the selection of the networking technology and the level of background traffic crossing the network.

1 Introduction

This paper addresses part of the work that is being carried out within the SABA (New Services for the Broadband Academic Network) project. SABA is embedded within the telematic services and applications research program supported by the Spanish government. The project focuses on the development and evaluation of new proposals of technologies, architecture and protocols for communications networks. Moreover, the project considers the use of this technology for various application environments; such as, computer supported collaborative work, videoconferencing and multimedia services, and distributed processing.

Within SABA, we are deploying and using a virtual parallel machine (VPM) over a network-of-workstations (NOW), which communicate by means of an ATM based broadband multiservice network. The objective is to asses, by experimentation, the performance achieved by this distributed parallel processing environment (DPP-E) both, when deployed over a local area and when deployed over a wide area. This is important because, even though the deployment of ATM based NOW's is becoming common ground and there are analytical predictions of its use as a DPP-E under ideal conditions [1], there are no reports[2] of experimental observations using ATM based NOW's for DPP-E under lees idealistic conditions. This paper only describes experiences with the DPP-E over a local area ATM network.

---

[1] Oscar-Iván Lepe is assistant researcher at Centro de Investigación Científica y Educación Superior de Ensenada, México. He is currently at UPC with a Ph.D. grant.

[2] Related reports we found either use a non ATM network to support their PVM based DPP-E [2] or do not use full DPP-E like ours over ATM [3, 4].

It is well know that the performance of a DPP-E is limited by the performance of the underlying networking technology [1]. Consequently, we designed our DPP-E experiments to measure the effect on its performance produced by the selection of the networking technology and by the level of background traffic crossing the network. Moreover, the same literature predicts that it is very likely that the combined power of a NOW based DPP-E may exceed that of a costly supercomputer. Thus, we use experimentally observed performance-measures form a SGI Origin2000 supercomputer as our reference point.

The rest of the paper is organized as follows. In section 2, we describe our DPP-E comprised of the VPM and the transport network. Then, in section 3, we explain our design for the tests and measurements carried out with the DPP-E. Section 4 contains the discussion of observed results. Finally, we conclude and summarize our work in section 5, and section 6 has the bibliography.

2 Distributed parallel processing platform

In this section, we will describe the instrumentation of both our VPM and the underlying transport network. In addition, we will describe the use of a HP Broadband Test system for producing controlled background traffic through the network.

2.1 Virtual parallel machine

We instrumented the virtual parallel machine with PVM (Parallel Virtual Machine) [5]. PVM offers two communication modes, RouteDefault and RouteDirect. In RouteDefault mode, distributed software modules that make up the application interchange messages by means of its local daemon process. In RouteDirect mode, modules communicate directly with each other.

Our VPM has four computing nodes. All are Sun workstations running the Solaris v2.5.1operating system. More specifically, two workstations are Ultra-1 with 128 Mbytes of RAM, another is a Ultra-1 with 64 Mbytes of RAM, and one more is a SparcStation20 with 64 Mbytes of RAM, also.

2.2 Network configuration

SABA project's ATM network was used as the transport network for the experiment. As any other ATM network [6], a virtual private network (VPN) designed to meet the requirements of a particular service is built upon a physical transport fabric designed to meet topological and logistic requirements.

We designed our VPN to meet the requirements for building a PVM based VPM; that is, IP connectivity. IP connectivity means that the network configuration has to support the delivery of IP datagrams. This implies support for IP to physical-network address resolution and subnetwork routing. Basically, ATM networks may support IP connectivity in two ways: dynamically setting up switched virtual circuits (SVC) through a signaling protocol, such as, LANE [6]; or statically setting up permanent virtual circuits (PVC) that require no signaling protocol. Because using signaling

protocols increase the number of control variables in the experiments, we avoided using SVC's. While using PVC's, we still had another design decision to take. To use a mesh configuration with 2N(N-1) PVC's, where N is the number of computing nodes and 2 PVC's make up a full-duplex channel; or to use a star configuration with 2(N-1) PVC's. We decided to use a star configuration.

Our local part of SABA's ATM network is composed of two switching nodes and five full-duplex optic channels. Each optic channel runs through a pair of multimode optic fibers, and uses SDH OC-3 physical framing. This means that in each direction optic channels have a 155.52 Mbits per second bit rate, of which 149.76 Mbits per second are available for user data. Fiber optic runs between switching nodes through approximately 50 meters. Then, fiber optic drop cables, between a switch and a workstation, are 3 meters long. This means that the worst case ATM-cell propagation delay is around 0.44µs. ATM switching nodes are realized by one FORE RunnerLE 155 and one FORE ASX-200BX. Each switch is capable of transferring 2.5 Gbits per second [7]. ATM host adapters are from FORE 200 series. Specifically, Ultra workstations use SBA-200E adapters and the SPARCstation uses a PBA-200E. All these adapters are capable to achieve 32-bit Sbus transfers and its embedded logic implements ATM protocols and functions up to AAL5 SAR sublayer [8]. Software support for these host adapters is realized by FORE Throughout tools v4.3.

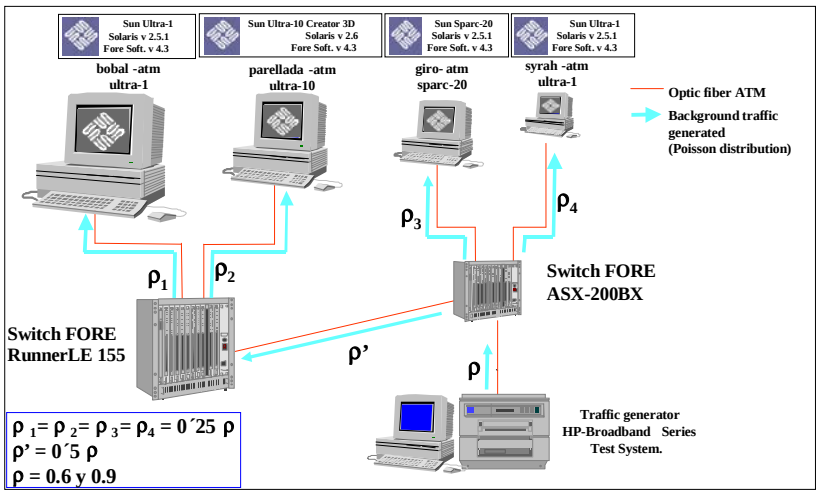

Fig. 1. Background traffic conditions

2.3 Background traffic conditions

In order to assess the effect on the performance of the VPM produced by the selection of the networking technology and by the level of background traffic, we introduced controlled background traffic to the ATM network. Background traffic generation was accomplished by means of a HP Broadband Series Test System (HPBTS). This test system is capable of producing a preprogrammed SDH OC-3 flow of cells, which obeys one of several cell rate and inter-cell time distributions. In order to distribute

the effect produced by the background traffic between the four workstations, we configure the switching nodes as Figure 1 shows. Four simplex PVC's with background traffic flow from the HPBTS to one of each workstation through the ASX switch. This switch forwards two PVC's to the RunnerLE switch and one PVC to each of the two workstations directly attached to it. The RunnerLE switch forwards the received PVC's one to each of the two workstations directly attached to it.

3 Tests and measurements

In order to assess the performance characteristics of the VPM deployed over the SABA ATM network, we designed a set of benchmarking tests. Theses tests, which we will describe shortly, were thought for comparing our DPP-E against a costly supercomputer. In addition, cause the performance of the VPM is limited by the performance of the underlying telecommunications technology, the tests was designed for assessing this. That is, how does the communication protocol selection affect system performance? Moreover, does the communication protocols take the most of the performance features of the underlying transport network?

A test comprises the execution of a workload hosted by a particular configuration of our DPP-E. We used six configurations grouped in three pairs. Each pair corresponds to a generated background-traffic level, rho (as shown in Figure 1 and described in 2.3), of zero, 0.6 and 0.9. For each level of rho, which comprises two tests, one test is carried out with the VPM configured to use RouteDefault (as described in 2.1) and the other configured to use RouteDirect. For comparing purposes, we executed three more tests. One is for defining a reference measure involving the execution of the workload hosted by a SGI Origin 2000 supercomputer. The other two embraces the execution of the workload by a four-node VPM collapsed within one host computer and configured to use RouteDefault and RouteDirect. It is important to note that although the Origin 2000 we used has 64 MIPS R10000 microprocessors, the workload ran there only used four.

Measurements obtained denote registered computing time spent by either the supercomputer or the **VPM**. We normalized measurement values to the computing time spent by the supercomputer for easing the comparative analysis.

3.1 Benchmarking tests

For benchmarking purposes, we use a set of parallel algorithm realizations extracted from the PVM version of the NAS vLU95 benchmark suite [2]. From the NAS suite we used CG, FT, IS, MG, EP, LU, SP and BT. Table 1 illustrates the communication characteristics of each of these algorithms, accordingly to the distribution of their message lengths. A brief description of these parallel algorithms follows.

Table 1. Distribution of message lengths

Parallel kernels	Parameters used	Number of messages	Length of messages (mean)	Length of messages (median)
CG	1400 matrix size	7116	1963	8
FT	Array of 64^3	186	236776	262144
IS	2^{19} keys in 2^{19} range	599	71364.320	130016
MG	64^3 grid	848	30226.525	2592
EP	2^{23} size	N/A	N/A	N/A
LU	Size 12x12x12	21057	153.006	63
SP	Size 12x12x12	133230	1373	1250
BT	Size 12x12x12	107064	1696	1525

4 Discussion of the results

Table 2 summarizes the registered measurements. We graphically analyzed them in order to asses how does the selection of the networking technology affects the performance of our VPM. In addition, we wanted to see if the communication protocols take the most of the performance features of the underlying transport network. Finally, we wanted to see if our VPM could outperformed a supercomputer as predicted in the literature, and under what conditions this does or does not happened.

Figure 2 graphically shows how does the performance of our DPP-E compares against a supercomputer, a computer simulated DPP-E (showed as ATM Ideal) reported in the literature [1], and against a centralized parallel processing environment. This figure shows one graph for each of the workloads use for benchmarking. Each graph depicts computing time spent by the supercomputer and each of the configurations of our DPP-E. In order to inspect the effect that background traffic produces on the performance of our DPP-E each graph shows three curves, one for each level rho of background traffic. We choose to plot computing time to show that although each workload has different computing requirements the performance of our DPP-E follows similar patterns with respect to the background traffic.

From Figure 2 we can also say that the performance of our DPP-E closely resembles the performance of the computer simulated DPP-E. This is obviously a good thing. Furthermore, the performance of our DPP-E is better than the performance of either of the centralized configurations; that is, the non-parallel (or sequential) and the centralized parallel configurations. By passing, we want to point out that the centralized parallel configuration responds better to increased traffic than the sequential configuration. We think this is due to operating system (OS) scheduling. When we have four processes running over the single CPU, OS scheduling actually pipelines the execution of the problem. This can only happened when computation and communications overlap as in EP, MG, LU, SP and BT. However, in CG, FT and IS this pipelining can be achieved due to communication characteristics of the algorithms.

Comparing the performance between the two configurations of our DPP-E, one using RouteDefault and the other using RouteDirect, we can say that RouteDirect is better than RouteDefault. Because in the first configuration we are avoiding one level

of indirection in the communication path between distributed modules. Moreover, RouteDefault is more sensitive to background traffic, although this is only a slight difference. We think this is because with RouteDirect communication is distributed in N(N-1) TCP connections so congestion in one physical link does not affect the whole communication subsystem.

Figure 3 presents a simplified graphic comparative analysis between the supercomputer and the sequential execution of the workload and the DPP-E. Our DPP-E does not outperform in any case the supercomputer. Nevertheless, it is also true that a supercomputer is far more expensive than an ATM based NOW. In addition, this NOW could be use to solve several other problems not related to high performance computing, such as, computer supported collaborative work or videoconferencing.

Finally, Figure 4 shows the effect produce by background traffic on the performance of the best DPP-E configuration; that is, RouteDirect. There we can see that although performance decreases directly proportional to the level rho of background traffic, communication characteristics of workloads influences the effect produced.

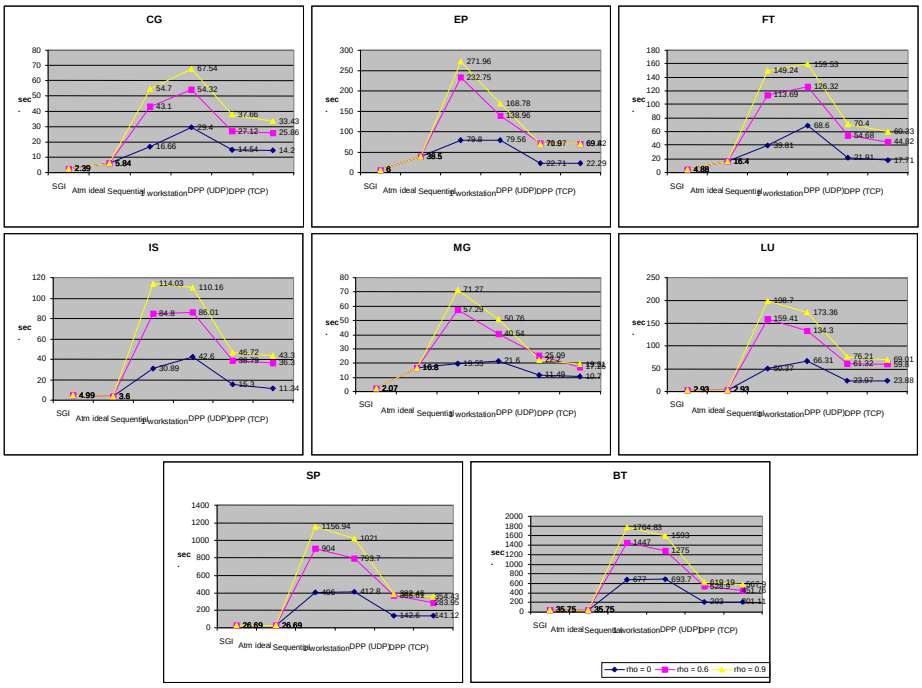

Fig. 2. Computing times for benchmarks hosted by various environments

Table 2. Evolution time of the benchmarks in SGI computer, simulated ATM, and ATM platform with different levels of background traffic.

	SGI	ideal ATM	1 workstation			4 workstations	
			Sequential	Parallel			
				Default	Direct	Default	Direct
rho = 0							
CG	2.39	5.84	16.66	29.4	29.07	14.54	14.2
EP	6	38.5	79.8	79.56	79.95	22.71	22.29
FT	4.88	16.4	39.81	68.6	64.35	21.91	17.71
IS	4.99	3.6	30.89	42.6	39.35	15.3	11.34
MG	2.07	16.8	19.55	21.6	20.73	11.49	10.7
LU	2.93		50.37	66.31	66.66	23.97	23.88
SP	26.69		406	412.8	411.16	142.6	141.12
BT	35.75		677	693.7	690.3	203	201.11
rho = 0.6							
CG			43.1	54.32	53.23	27.12	25.86
EP			232.75	138.96	138.34	71.9	69.42
FT			113.69	126.32	118.19	54.68	44.82
IS			84.8	86.01	79.88	38.79	36.3
MG			57.29	40.54	38.84	25.09	17.26
LU			159.41	134.3	134.15	61.32	59.8
SP			904	793.7	792.61	366.61	283.95
BT			1447	1275	1271.9	528.9	451.76
rho = 0.9							
CG			54.7	67.54	63.38	37.66	33.43
EP			271.96	168.78	167.58	70.07	69.8
FT			149.24	159.53	148.68	70.4	60.33
IS			114.03	110.16	102.1	46.72	43.3
MG			71.27	50.76	48.17	22.2	19.21
LU			198.7	173.36	163.9	76.21	69.01
SP			1156.94	1021	1014	382.46	354.43
BT			1764.83	1593	1483	619.19	567.9

5 Summary

Current ATM-based networks are potentially capable of satisfactory supporting distributed parallel computing applications. The use of such a complex networking technology makes sense when the support parallel computing applications has to be integrated with other services over a single network. Thus, it is not necessary to adopt a specific network for distributed computing, but rather organizations can take advantage of an existing ATM network to support parallel computing in addition to other networking applications. For this reason, a good network-interface design is vital for the network to provide adequate performance to parallel computing. A good design is one that optimizes both protocol-processing latency and congestion-recovery procedures effectiveness.

The experiments show that latency reductions in the network interface can be sufficient to achieve significant performance improvements, provided that the network load from other applications is sufficiently low. In case of extreme congestion situations, loss recovery mechanisms rapidly degrade performance and, consequently, optimizations are required to cover these circumstances. Although currently hard congestion does not seem to be a very frequent issue in ATM networks, the growing trend of integrating multimedia applications over high-speed networks could lead to significant increases of traffic in the networks. Therefore, the need of efficient loss recovery can become evident in the immediate future.

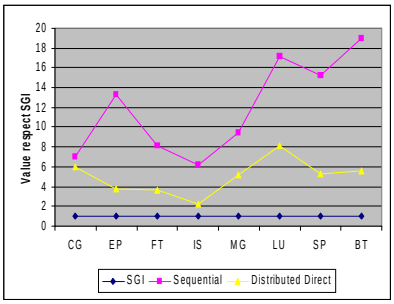

Fig. 3. Comparative analysis between SGI and DPP

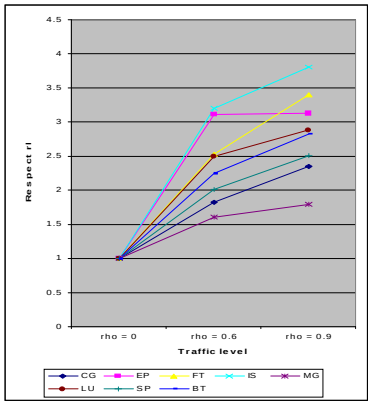

Fig. 4. Traffic effect in the DPP performance

6 Acknowledgments

This work has been supported by CICYT (Spanish Education Ministry) under contract TEL97-1054-C03-03 and by the Mexican Government through CONACyT 66864 grant. The authors want to acknowledge the project SABA for his interest in making this experience possible.

7 Bibliography

1. Joan Vila-Sallent and Josep Solé-Pareta, Potential Capability of ATM to Support Network-Based Parallel Computing, Computer Communications Globecom 97.
2. S. White, A. Alund and V.S. Sunderam, Performance of the NAS Parallel Benchmarks on PVM Based Networks, Journal of Parallel and Distributed Computing, 26, 1994,61-71.
3. Sheue-Ling Chang, David H. C. Du et al., Enhanced PVM Communications over a High-Speed Local Area Network, Distributed Multimedia Center & Computer Science Department, Minnesota, 1995.
4. Mengjou Lin, Jenwei Hsieh, David H.C. Du et al., Distributed Network Computing over Local ATM Networks, IEEE Journal on Communication Special Issue of ATM LAN's,13(4),1995,54-64.
5. A.Geist et al. PVM 3 User's Guide and Reference Manual, Oak Ridge National Laboratory, 1994.
6. Anthony Alles, ATM internetworking, Engineering InterOp, Las Vegas, March 1995.
7. FORE Systems, Inc. – Product Catalog, http://www.fore.com/products/
8. FORE Systems, Inc. Fore Runner SBA-200 ATM Sbus Adapter User's Manual, 1993.

-. P.W. Dowd et al, "Issues in ATM support of high performance geographically distributed computing", proceedings of IEEE workshop on high speed computing ,HiNet'95, pp19-28.

-. H.Zou et al, "Faster message passing in PVM", proceedings of IEEE workshop on high speed computing ,HiNet'95, pp 67-73.

-. P.Papapdodpoulos et al, "Wide-area ATM networking for Large-scale MPPs", Proceedings of the 9th SIAM conference on parallel processing, March 1997.