Emergent high speed network technologies, such as ATM, SCI, Fiber Channel, GigabitEthernet, and high performance Workstations/SMP Servers/Personal Computers are making it feasible to build high performance computing systems with performance equivalent to commercial MPP’s.

However, to exploit the full potential of the system, issues such as fast message passing protocols and shared memory support must be investigated in more depth. This paper presents a new parallel system based on stand-alone workstations and high speed interconnection networks with support for efficient message passing and shared memory.

Several prototypes of the architecture are being and will be implemented, based on interconnection systems composed by network adapters and switches based on ATM and SCI standards. Prototypes with commercial fast Ethernet, ATM and Myrinet are also being implemented.

1 - Introduction

Emergent high speed, low latency interconnection network technologies allow today the implementation of high performance, low cost parallel machines based on commodity processing nodes and networks. These systems, called NOW or COW, constitute currently a hot research topic, both in the academic and industrial areas. The bandwidth and latency of their interconnection systems are comparable to that exhibited by several MPPs, mainly if we take into account the software overheads associated to the network protocol stacks of most MPP operating systems.

In order to make such systems more usable, specially when we consider non scientific programming, it is necessary to provide a more adequate programming environment. Several researchers have shown that shared variable model is more comfortable for both the user and the construction of development tools (compilers).

This paper describes some projects being developed in the INTEGRATED HYPERSYSTEMS PROJECT, in particular the SPADE-II, a system based on commodity workstations and communication networks, with support for efficient message passing and shared memory programming. Following similar research lines, the work is focused on user level network communication protocols and remote memory
access interconnection network. On the other hand, the work is also oriented to the implementation of CC-NUMA and COMA models.

This paper gives the status of the Hypersystems Project. First, we describe the prototypes in development and the associated system and application software. Then, we give the status and future works.

2 - SPADE-2

2.1 General Description

The SPADE-2 is built from commodity high performance SHV PCI SMP servers. We are developing two types of prototypes: 

i) cluster based only on commodity subsystems (processing nodes and network); and 

ii) prototypes based on custom interconnection subsystem.

The first prototype is based only on commodity processing nodes and networks. The idea is to provide both messages passing model, with fast light-weight user level message communication protocols, and shared memory model, implemented in software by means of software DSM employing relaxed memory consistency model.

The other prototypes are based on commodity processing nodes and custom interconnection system. They constitute NUMA, CC-NUMA, COMA architectures.

2.2 - PROTOTYPES

2.2.1 - Spade-II “Aquila”

This is the first prototype built in this Project, and is completely functional today. It is composed of 24 Dual PentiumPro processing nodes interconnected by multiple networks: FastEthernet, ATM, Myrinet and PAPERS+, a extended Purdue PAPERS network version with support for 32 nodes and 32 bits parallel collective data communication.

The software environment includes: 

i) Linux SMP (RedHat 4.2), Solaris, and NT operating systems; 
ii) MPI 1.1, PVM3, Genova Gamma Active Messages, Cornell U-NET, Illinois HPVM (FastMessages, Global Array, MPI, etc.), and LSI-EPUSP FULL communication libraries; 

iii) DQS Cluster Management System; 
iv) Portland Group HPF/F90, etc. FULL is a user level communication library being developed at LSI-EPUSP for FastEthernet, ATM, Myrinet and LSI-IN1 networks with socket API.

2.2.2 - Spade-II “Orion”

It is the second prototype, and is based on LSI-IN1, a custom interconnection network developed by LSI-EPUSP with support for remote memory access, multicast operations, and reliable communication. This network supports remote memory access in order to implement hardware global shared memory (NUMA model) as well as efficient message
passing. The multicasting support is very important to implementations of collective operations. The NIC onboard memory can be used as a local portion of the global shared memory as well as the buffer memory for message passing. The prototype is composed by 24 Quad PentiumPro processing nodes and three interconnection networks: FastEthernet, LSI-IN1, and Myrinet. A small version of this system will be present in the Research Exhibit Booth in SC’97, San Jose, CA, in Nov. 1997.

2.2.3 - Spade-II "Taurus"

This system fully implements CC-NUMA and COMA architectures, using SHV SMP Quad PentiumPro processing nodes. We are doing several simulation runs in order to evaluate the best organization and physical parameters of the third level cache, as well as the best place for it in the processing node (processor bus, memory bus or PCI bus). As the main goal is to implement a cache coherent multiprocessor systems based on commodity SHV SMP processing nodes, we do not consider solutions that imply processing board hardware modifications, such as adopted by Sequent and Data General, in order to attach the tertiary cache. We are considering to use high speed low latency links, such as SCI, ATM with Credit Flow Control, Gigabit Ethernet (Fibre Channel), etc. in the implementation of the interconnection system, called LSI-IN2.

3 - SOFTWARE

3.1 - Software Distributed Shared Memory

Software DSM systems constitute a hot research topic, providing a shared memory abstraction over a network of workstations. We developed a software DSM system based on the relaxed memory consistency model to provide a shared memory over conventional NOWs. This system, Pulsar, is available today for SunOS, Solaris, FreeBSD and Linux. Currently, we are implementing optimizations in order to exploit user level communication libraries, collective communication, and remote memory access provided by the LSI-IN1 and LSI-IN2 networks.

3.2 - Message Passing Libraries

The Project includes the porting of standard message passing libraries, such as PVM and MPI, and development of non-standard light-weight user level communication protocols. Currently we are evaluating the performance and studying several user level communication protocols implemented by other universities, namely, Cornell UNET (over ATM and FastEthernet), Illinois HPVM (over Myrinet), and Genova GAMMA, using a profiling tool with a CPU clock cycle resolution.

3.3 - Operating Systems, Parallel Programming Tools and Environments

The Aquila and Orion Spade-II prototypes by default employ the RedHat 4.2 Linux Operating System. Each node runs a copy of Linux SMP, and the communication is implemented through message passing or SW DSM libraries. In the Taurus Spade-II prototype, we intend to run only one copy of operating system - the Linux CC-NUMA.
This operating system, a version of Linux for CC-NUMA architectures, is currently under study.

Other important development in system software is the CPAR programming environment. CPAR is an extension of the "C" language for parallel programming, with support for multiprocessors systems with complex memory hierarchies, such as provided by the CC-NUMA Spade-II prototypes. Currently we have CPAR ported to SMP nodes. We expect to have in few months versions for network of workstations with PVM or SW DSMs.

3.4 - Applications

Besides the scientific and engineering applications, we intend to investigate the use of the described architectures in other fields such as distributed data bases, Web servers, video on demand servers, image rendering engine, etc, mainly exploiting the shared memory model implemented by SPADE-2.

Currently we have several applications running or being ported to Spade-II: RTP (ray tracing tool), Python (medical image processing program), PVV (portable volume visualization library), neural network simulation programs, etc.

4 - Related Works

In this work we presented SPADE-II, a high performance system using, as far as possible, standard hardware and software, while also offering the possibility of the effective implementation of shared memory as well as message passing models. Unlike other proposals, this one offers innovative solutions that don't rely on exotic, not open technologies. Most of the features discussed here about the custom interconnection system LSI-IN1 is provided by the Washington University in Saint Louis's Gigaswitch, namely, credit-flow control, reliable multicasting, etc. Other related projects are FORTH's Telegraphos which is also developing credit flow control based ATM switches, Princeton SCHRAMS, Cornell UNET, etc. In order to implement the CC-NUMA and COMA versions (based on LSI-IN2), we intend to work with IEEE-ANSI SCI standard compliant circuits and protocols, mainly because of its 1Gbytes/s per link bandwidth.

One prototype - the Aquila Spade-II system - is concluded today and is fully operational, with 24 DualPentiumPro processing nodes. The Orion Spade-II prototype will be in few weeks operational with 8 to 24 QuadPentiumPro processing nodes. The Taurus Spade-II is in the design phase.

Acknowledgments

This work is supported by FINEP, through grant n. 56.94.0260.00., CNPq, CAPES and FAPESP.