Proposal for Teaching High Performance Computing using SMP Clusters

With the cost of commercial off-the-shelf (COTS) high performance interconnects falling and the respective performance of microprocessors increasing, workstation clusters have become an attractive computing platform offering potentially a superior cost effective performance. In recent years, we have seen the maturing of Symmetric Multiprocessors (SMPs) technology, and the heavy reliance upon SMPs as the work-intensive servers for client/server applications. There are already several examples of clusters of SMPs, such as clusters of DEC AlphaServer, SGI Origin, Sun Ultra HPC machines, and the IBM SP system with SMP "High" nodes; moreover, the Department of Energy's Accelerated Strategic Computing Initiative (ASCI) program relies on the success of computational clusters such as Option White, a 512-node IBM SP-2 with 16-way SMP nodes. With the acceptance of message passing standards such as MPI, it has become easier to design portable parallel algorithms making use of these primitives. However, the focus of MPI is a standard for communicating between shared-nothing processors, and although MPI programs run on clusters of SMPs, this is not necessarily the optimal methodology for these platforms. This teaching platform will help develop a hybrid methodology for programming clusters of SMP nodes which aids in the design and implementation of efficient high performance parallel algorithms.

We propose to build an integrated computational cluster of symmetrical multiprocessors (see Figure 1) to use as a teaching tool for advanced computer design (for example, parallel processing and high-speed networks). Several departmental courses would make significant use of this platform:

- **EECE 432: Introduction to Parallel Processing**,
- **EECE 440: Introduction to Computer Networks**,
- **EECE 509: Parallel Algorithms**, and
- **EECE 538: Advanced Computer Architecture**.

For instance, this platform would serve as an architectural learning tool for parallel processing by allowing students to develop practical algorithmic concepts. In conjunction with the networks course, students will get hands-on experience of modifying the interconnection network topology and protocols, measuring the implications each has on overall performance. In addition, the platform would be an excellent departmental teaching platform for high-performance application development using the message passing interface (MPI) parallel

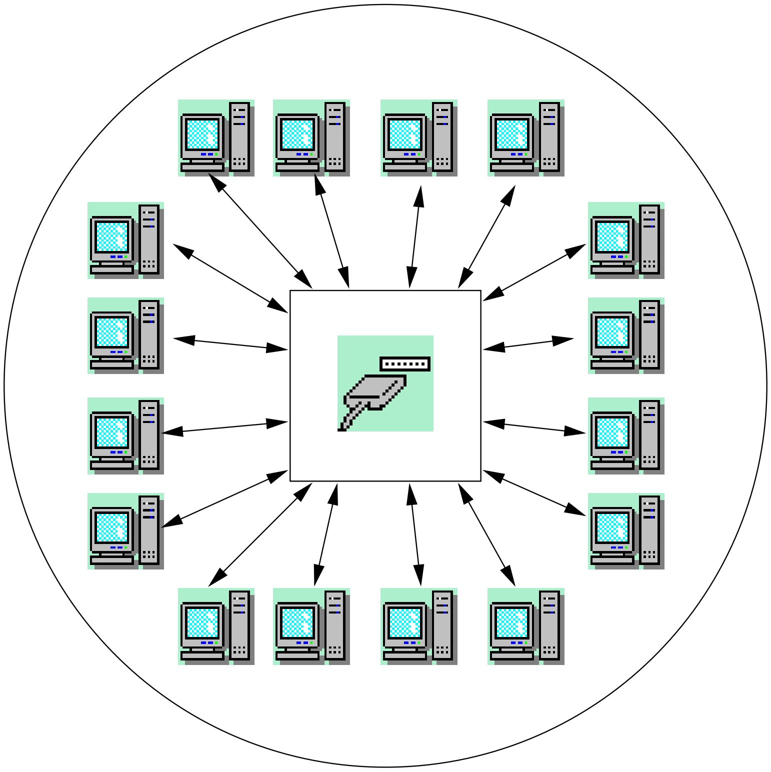

Figure 1: Computational Cluster

programming methodology. Currently, the department has no facilities for these important (and emerging) computer engineering themes.

Two main commodity off-the-shelf components comprise the hardware aspect for this computational platform: 1) SMP nodes and 2) an interconnection network. We envision a system with eight to sixteen nodes and an appropriate network.

Research account funds have already been allocated to finance the cost of the workstation nodes. However, a network is necessarily needed to utilize these workstations as a single computational engine. We are asking the department to pay for the interconnection network as follows.

The department will fund the purchase of a Myrinet system-area network (SAN) including SAN switching technology, and a PCI adaptor cards and locking cable for each computational node.

Myrinet Part	Quantity	Unit Price	Total
Dual 8-port SAN switch (M2M-DUAL-SW8)	1	$ 2,000	$ 2,000
SAN/PCI Interface Card (M2M-PCI32C)	8	1,500	12,000
3 ft SAN Cable, Locking (M2M-CL-03)	8	120	960
TOTAL			$ 14,960

Table 1: Myrinet Pricing

We need an interconnection network called Myrinet from Myricom (`www.myri.com`). We are using the new SAN (system-area-network) configuration which obtains twice the performance of Myrinet/LAN, but only when the switch is within approximately three feet of the nodes. For this reason, the system will be in a star configuration, with the network in the center, and the SMP workstations placed around the switch. The switch is M2M-DUAL-SW8, a dual 8-port SAN switch (that is, two independent 8-port crossbar networks in a single package) and costs $2000. Each PC has a Myrinet-SAN/PCI interface card ($1500/each), and each three foot Myrinet cable connecting the card to the switch costs $120. For example, Myrinet for eight nodes using this technology would cost approximately $14,960. See Table 1.

Myrinet/SAN is about five times faster than fast Ethernet, with a much smaller latency. For example, Myrinet connections have a peak bandwidth of 1.28 Gbps (or a total aggregate bandwidth of over 10 Gbps per switch!), while incurring a latency of only 5 microseconds. This is comparable with the high performance interconnects used in current commercial parallel machines.

The second hardware component is the workstation-class computing nodes. We have chosen each node configuration as a fully-configured workstation-class dual-Pentium II class machine, for instance, a Dell Workstation 400 with dual 333MHz Pentium II processors, 256 MB RAM, 9GB SCSI disk, 12x/24x CD-ROM, 8MB video (Matrox Millennium), 19” Trinitron monitor, integrated 10/100Mbps Ethernet, and sound/speakers. The higher-educational price for this configuration is approximately $5175 (before any quantity discount or negotiation).

The software infrastructure is freely available. Our system runs a free UNIX SMP operating system (for example, Linux/SMP or FreeBSD/SMP), with FreeBSD Myrinet drivers provided by Anne Hutton at ISI/Atomic or Linux Myrinet drivers from Bob Felderman at

Myricom.

Thus, applications can take advantage of a single SMP nodes using POSIX standard threads (pthreads), or utilize the cluster with standard message passing interfaces such as MPI. For example, the public domain MPICH implementation of MPI includes optimizations for Myrinet.

In summary, we are asking the department to fund the purchase of hardware configured to interconnect eight workstations with a high-speed Myrinet system-area network for an amount totaling $14,960. This hardware will be used both for teaching in this new thrust area, as well as to leverage future research grants.

David A. Bader, Ph.D
Assistant Professor
Department of Electrical and Computer Engineering
University of New Mexico
Phone: (505) 277-6724
FAX: (505) 277-1439
email: dbader@eece.unm.edu
http://www.eece.unm.edu/∼dbader