CIS 607 Cluster Computing Seminar

- Sameer S. Shende

1. Chapter 1 from High Performance Cluster Computing Vol 1. Architectures and Systems by Rajkumar Buyya: Grade A. As an introductory chapter on cluster computing, the author gives a good overview of technologies and tools available for cluster computing. I found the list of performance tools to be incomplete ; but it made for a good first reading.

2. "An Assessment of Beowulf-class Computing for NASA Requirements" by Sterling, Becker, Warren, Cwik, Salmon, and Nitzberg: Grade B. The paper specifies applications and requirements for a successfully integrating a PC cluster to form a Beowulf. They are more Fortran "application-centric" as opposed to DOE labs that are "C++ centric"; interesting cost comparisons.

3. "Scaling of Beowulf-class Distributed Systems" by John Salmon, Christopher Stein, and Thomas Sterling: Grade B. The paper examines two networking alternatives, one with high bandwidth backplane and the other that uses routed topologies for making a Beowulf cluster; they prefer the latter for building large scale systems. They present a lot of data to prove their point, not all of which is easy to comprehend.

4. "Design and Evaluation of an HPVM-based Windows NT Supercomputer" by A. Chien et. al. : Grade B. They present their experiences in building a 192 processor Windows NT cluster and the use of HPVM software to program these. I am not convinced that Windows NT was the right choice as an OS for assembling the cluster for scientific computation.

5. "Performance Enhancements for HPVM in Multi-Network and Heterogeneous Hardware" by G. Bruno, A. Chien et. al. : Grade A. The proposed enhancements to HPVM include a shared memory transport and an adaptive polling scheme that polls cache lines instead of going through the PCI bus. I liked the ideas of adaptive polling and "pacing" of software to detect the optimal packet size for PCI implementations.

6. "Fast Messages (FM): Efficient, Portable Communication for Workstation Clusters and Mas-

sively-Parallel Processors", IEEE Concurrency, vol. 5, no. 2, April-June 1997, pp. 60-73. Grade B. They present the implementation of a low latency, high throughput software layer that is portable and reaches close to optimal hardware performance. I'd prefer to use an MPI implementation layered on top of FM rather than the core API directly.

7. "Scheduling Parallel Jobs on Clusters" by Dror Feitelson in High Performance Cluster Computing Vol 1. Architectures and Systems by Rajkumar Buyya, pp. 519-533. Grade C. They present job scheduling strategies on a Beowulf style cluster. I didn't find this very interesting.

8. "On the Design and Evaluation of Job Scheduling Algorithms" by J. Krallman, UweSchwiegelsohn, and R. Yahyapour, in Proceedings of Workshop on Job Scheduling Strategies forParallel Processing, Lecture Notes in Computer Science, D. Feitelson and L. Rudolph (Eds.),Springer-Verlag, 1999. Grade B. Different job scheduling algorithms were studied in the paper. Good reading for getting an overview of job scheduling strategies.

9. "Demand-based Coscheduling of Parallel Jobs on Multiprogrammed Multiprocessors" by Patrick G Sabalvarro and William E Weihl. Grade A. They present an algorithm for co-scheduling jobs based on their dynamic communication behavior. Very well written paper; howeverm they did not implement their scheme in a real system.

10. "Distributed Shared Memory" by Alan Judge, Paddy Nixon, Brendan Tangney, Stefan Weber, and Vinny Cahill, in Buyya. Grade A. In conventional RPC model, operations are moved between processes, in DSM, data is moved between the two. An interesting paper; gave indepth information on the choices available for DSM systems, definitely helps us decide the issues involved in choosing whether to go for explicit message passsing based programming, or transparent distributed data access using DSM.

11. Climate Ocean Modeling. P. Wang, B. Cheng, Y. Chao. Grade B. They share their experiences in developing a portable ocean modeling code using MPI. The PC cluster comparison with T3E and T3D was interesting; the PC cluster outperformed the older T3D on their code but was not as good as a T3E made up of newer Alpha clusters.