

Utilizing a Massively Parallel Neural Network and a HPC Data Environment for Cancer Diagnosis

The 2006 International Conference on Machine Learning; Models,
Technologies & Applications

Phil Andrews¹

¹ San Diego Supercomputer Center, University of California, San Diego,
La Jolla, Ca 92014, USA

Tel: 858-534-5000, Fax: 858-534-5117
andrews@sdsc.edu, <http://www.sdsc.edu>

Abstract. Recent improvements disk prices and network speeds, have outstripped even the continuing exponential growth in computational power. Scientists have greatly increased the data created by their applications and can now exceed 10 Terabytes and may approach 100 Terabytes in a single run, with the computational time used to examine the output possibly more than expended in its creation. At the same time, enormous increases in the number of processors available to work on a single problem, is opening up new opportunities for computation to move into new areas. In this paper, we consider such an example, explain its computational environment at the San Diego Supercomputer Center, and describe how we plan to use neural networks. **Keywords: Data, File Systems, Blue Gene, Neural Networks**

1 The proposed Neural Network engine: a Blue Gene/L rack

Today's HPC environment is significantly different from a normal computing center: all aspects must be designed from the beginning for extreme data transfer rates, with a very high level of parallelism essential. Data transfer rates must be in the GB/s ranges and data movement must be minimized. The design of the Neural Network engine and its associated file system is paramount for the efficient utilization of multi-terabyte datasets.

High-performance computing has traditionally emphasized floating-point capability, a tendency reinforced by decades of exponential growth in processor technology. This has allowed computer simulations at higher and higher resolution in all three spatial dimensions plus time. The most demanding of these simulations, such as in cosmology [1], geophysics [2], and engineering fluid dynamics [3], can generate huge amounts of output. To accommodate that output and take full advantage of the resolution provided, I/O capability must grow exponentially too.

Increased I/O capability is also required to process the growing data sets being generated by automated instruments. For example, the National Virtual Observatory [4] currently has a data set of about 50 TB, which allows many investigators to use supercomputers to interact directly with the observed night sky. Meanwhile, new generations of observational instruments are expected to provide data streams in excess of 1 TB/s [5].

Although it would be possible to achieve greater I/O capability by simply scaling existing state-of-the-art systems to greater and greater sizes, problems of power, cooling, machine room space, and, above all, cost come into play. Thus it is attractive to consider systems that have either been designed for truly stellar I/O performance from the beginning or can be easily configured for that purpose.

IBM's new Blue Gene supercomputer [6] offers an interesting possibility. Although originally designed for Lawrence Livermore National Laboratory to provide unprecedented floating-point performance – 360 Tflops in 64 racks – the system can also be configured to provide high I/O performance. Accordingly, the San Diego Supercomputer Center acquired the first Blue Gene system to have 128 I/O nodes in a single rack, which is eight times the standard I/O density. This system is expected to provide exceptional I/O capability.

Named Intimidata, in anticipation of its high I/O performance, the single-rack system at SDSC was installed and accepted in December 2004. In April 2005, about 500 TB of Serial ATA disk was installed and connected to Intimidata. Also in April 2005, a prototype version of IBM's GPFS software for parallel I/O [7] was installed on Intimidata.

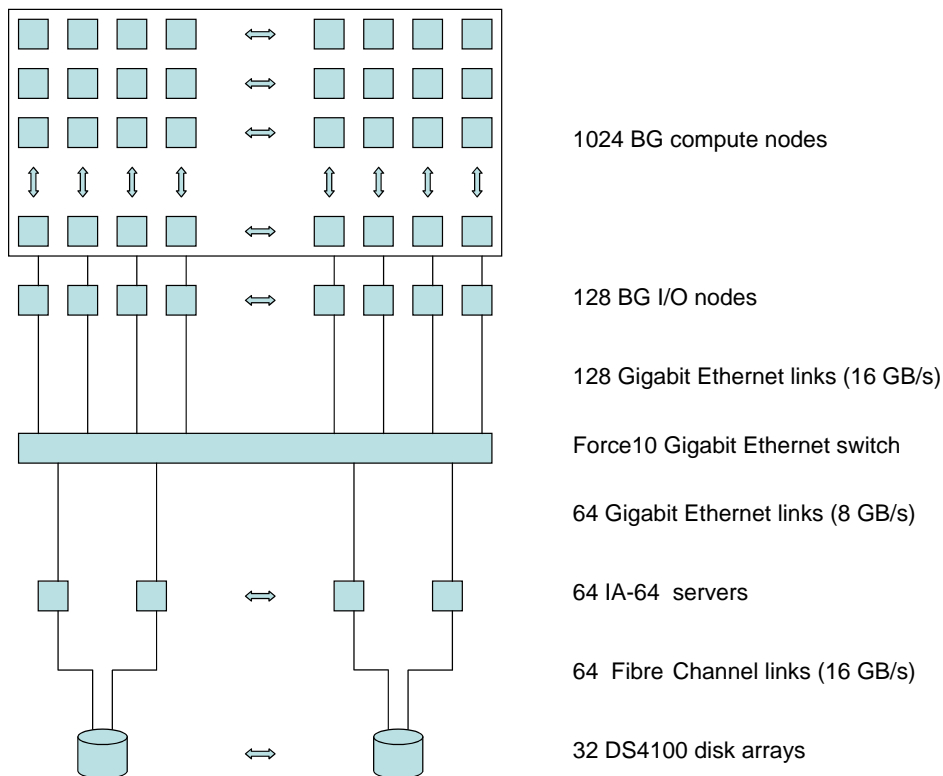


Figure 1. Configuration for parallel I/O on Intimidata, the Blue Gene system at SDSC

The Blue Gene architecture uses relatively slow and low-powered PowerPC processors to achieve high processor density in a compact footprint. Specifically, a single Blue Gene rack has 1024 compute nodes, each consisting of two 700-MHz processors, which gives an aggregate peak speed of 5.7 Gflops. The I/O-rich configuration at SDSC has another 128 I/O nodes, again containing two processors each. Memory per node is 512 MB, which is relatively small.

The two types of nodes are physically identical and differ only in their connectivity and the software they run. Each node has controllers to support five networks:

- a 3-D torus for point-to-point MPI operations,
- a global tree for most collective MPI operations,
- a global interrupt (GI) network for low-latency MPI_Barrier operations,
- a Gigabit Ethernet (GigE) for I/O,
- a JTAG network for machine control.

The compute nodes are connected to the torus, tree, GI, and JTAG networks, while the I/O nodes are connected to the tree, GI, GigE, and JTAG networks. Most communication between the compute and I/O nodes is via the tree.

The overall ratio of compute nodes to I/O nodes is 8:1 in the Blue Gene system at SDSC. However, users can configure the associated connectivity and resulting ratio via a mapfile at run time. This allows ratios as small as 1:1 for runs that use no more than 128 compute nodes.

Jobs running on compute nodes interact with disks through the I/O nodes and their associated Gigabit Ethernet links. Figure 1 shows the specific configuration at SDSC that supports parallel I/O.

Each of the 128 I/O nodes on Intimidata connects to a Force10 Gigabit Ethernet switch via a single (Cu) GigE cable. The 64 servers for GPFS are 4U IBM rack-mounted systems with two Intel IA-64 processors running at 1.3 GHz and 4 GB of memory each. There is a single (Fiber) GigE connection from each server to the Force10 switch. In addition, each server has two Fibre Channel Host Bus Adapters (HBAs).

There are 32 IBM FastT100 DS4100 disk RAID systems, with 67 250 GB drives in each. The total raw storage is 536 TB. The RAID sets within the DS4100 are internally connected via two 2 Gb/s Fibre Channel arbitrated loops, with an independent controller for each loop. The drives within each RAID set are Serial ATA disks. The two DS4100 controllers each have a 2 Gb/s FC connection to the outside world so that each NSD server can connect to two DS4100 systems.

The peak I/O rate possible is determined by the 64 GigE links between the switch and the IA-64 servers. This rate is 8 GB/s. In practice, other components may impose limitations that reduce performance substantially. For small payload messages, processing the TCP packets will consume much of the compute capability of the I/O nodes, while if any RAID disk rebuilding is going on, disk performance will be severely compromised.

For the tests reported here, the full configuration shown in Figure 1 was not available. Specifically two or three of the DS4100 systems were offline along with the corresponding IA-64 servers.

Significant upgrades are under consideration. One would double the bandwidth from each IA-64 server to the Force10 switch with another GigE connection. This would double the peak I/O rate to 16 GB/s. Another upgrade would add a second HBA to each IA-64 server. This would allow redundant connectivity and protect against any single DS4100 server failure.

The software for parallel I/O installed on Intimidata is a prototype version of IBM's GPFS [7], a parallel, shared-disk file system for cluster computers. GPFS allows programs in the cluster to share a file system with full POSIX semantics as if they were on the same physical machine.

SDSC has run GPFS on production supercomputers for several years and so has considerable experience in its use. Moreover, a recent study [8] showed favorable performance comparisons between GPFS and alternate parallel file systems.

Even though the I/O nodes operate as a cluster, GPFS works differently on Blue Gene than on a conventional cluster. On Blue Gene, user programs run on the compute nodes, which run a lightweight kernel that does not locally handle many operating system calls, in particular file I/O. Instead, the compute node kernel forwards user program file I/O over the tree network to the compute node's associated I/O node. There, the CIOD process, acting as a surrogate for the user program, forwards the call to GPFS, which runs as part of the Linux kernel on the I/O node. GPFS, in turn, does disk I/O over the I/O node's gigabit Ethernet interface to a GPFS NSD server running on an external Linux server. This server does the actual disk I/O to its Fibre Channel-attached storage.

Implementing a prototype version of GPFS for Blue Gene required several changes from the existing PowerPC version of GPFS. Most of the changes arose because GPFS stores a variety of data (configuration files, logs, etc.) in each node's local file system, and the Blue Gene I/O nodes do not have local disks. This required changing GPFS to use NFS instead of local disk.

Additionally, the lightweight version of Linux running on the I/O nodes does not contain many of the services required by GPFS. A special Linux kernel had to be built for GPFS. GPFS administration commands must be run on one of the nodes in the GPFS cluster. Since administrators cannot run commands on the I/O nodes, an external node is added to the cluster for the purpose of administration. In theory, the NSD servers could be used to administer the Blue Gene GPFS cluster, but it was decided to use a separate node and thereby keep the Blue Gene GPFS cluster self-contained. This permits using the GPFS Multi-Cluster facility [9], which allows the Blue Gene to share data with the several other SDSC clusters that also run GPFS.

The IA-64 server nodes run SuSE Linux Enterprise Server 8 with a 2.4.21 kernel. The GPFS software is a custom IA-64 build of the GA version 2.3 with the latest fixpack (PTF 2) applied. The file system consists of 192 approximately 2-TB disk arrays defined as Network Shared Disks (NSDs), with three NSDs served by each of the 64 nodes.

Since each pair of server nodes is attached to one set of seven disk arrays through two separate controllers, and there are performance issues with having more

than one node/controller access the same LUN, the NSD failover mechanism was not enabled, providing only a single primary server for each NSD. Only 6 of the 7 total available arrays from each DS4100 were used so that I/O would be evenly balanced across the 64 NSD servers. The file system being exported includes all 192 NSDs for a total capacity of just under 360 TB, and the block size for GPFS writes is set to 512k to match the stripe size on the underlying RAID arrays.

The Linux kernel TCP parameters have been tuned for increased performance in a GPFS configuration, notably by increasing the maximum read and write TCP window sizes as well as the maximum number of outstanding requests that can be handled by the TCP stack. These tunings have been previously tested in similar GPFS configurations and found to be near-optimal.

The file system is exported to the Blue Gene I/O nodes using the multi-cluster feature added in GPFS 2.3 and tested by SDSC and IBM in the StorCloud Challenge demonstration at SC2004. This feature allows physically and logically separate GPFS clusters to share file systems over TCP/IP networks in the same way that GPFS file systems are shared within a single cluster when all nodes are not physically attached.

Performance is now sustaining ~4.5 GB/s reads and writes; impressive for a single rack system.

2 The Global File System

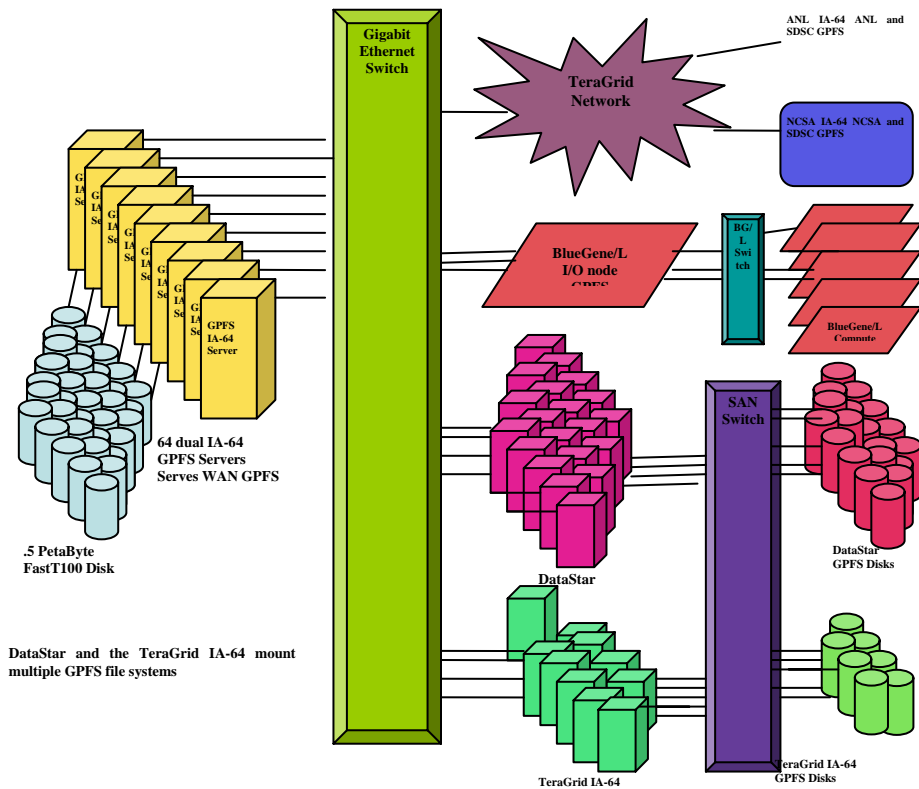


Figure 2. SDSC Local and WAN GPFS Configuration

The local file system for Intimidata, also forms the global file system[10] for both the SDSC compute systems and the TeraGrid itself with many compute systems writing directly to these disks, placing Intimidata in the perfect position for data mining the output of numerous data-intensive applications.

3 Prostate Cancer Diagnosis from Biopsy Images

We now have an environment where data of up to 100's of TeraBytes can be written at multiple GB/s, and immediately read at a similar rate into 2,048 processor system oriented towards data exploration. We are in the process of using this system for the automatic diagnosis of Prostate Cancer from Biopsy images.



Figure 3. Artist's rendering of Prostate Cancer progression thru Gleason Grades 1-5

In a previous paper[11] we showed how algorithmic techniques could be used to provide a diagnosis of Prostate Cancer level from Biopsy images. This procedure was made simpler by the existence of a scalar measurement typifying the severity of the disease: the Gleason Grade, which in this approach ranges from 1 thru 5. An idealized representation of how this relates to a microscopic Biopsy image is shown in Fig. 3.

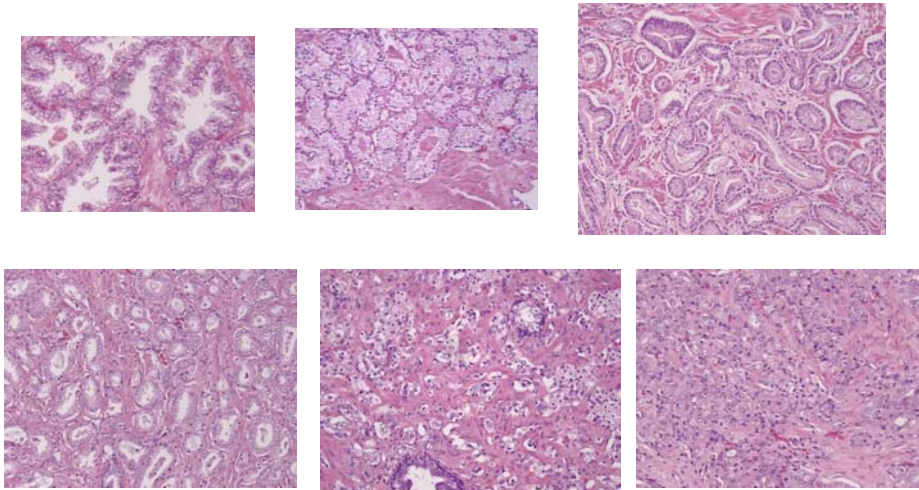


Figure 4. Actual microscope images of Prostate Biopsies

Although this approach is both useful and conceptually simple, in reality the actual images are not as easily evaluated as Fig. 3 might suggest. In Fig. 4 we show six actual images taken directly from a microscopic image of Prostate biopsies.

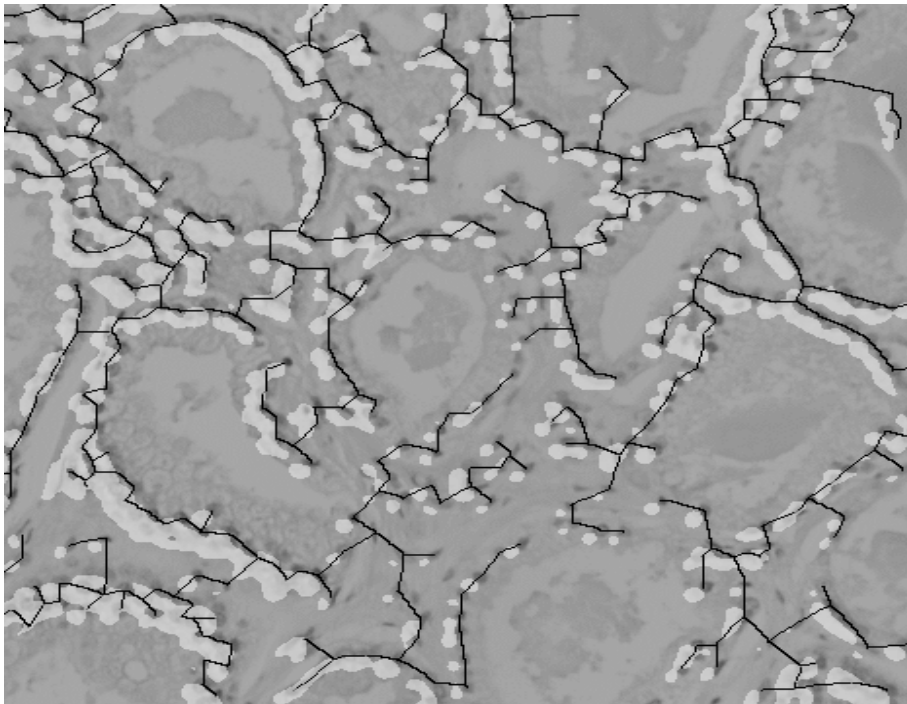


Figure 5. Identification of a distinguishing feature of one Biopsy

In order to progress, it was necessary to decide what distinguishing features of Prostate Biopsy images determined the Gleason Grade (as decided upon by experienced medical professionals). E.g., in Fig. 5 we show the length of coherent structures which was one important factor in the algorithmic determination of the Gleason grade[11].

This was a straightforward (if not necessarily simple) algorithmic approach which we employed 10 years ago. In the meantime, computational hardware has changed significantly, with, we expect, the greatest changes to come in the next few years. We refer to the multi-core[12] processor technologies likely to dominate chip design in the near future. Although offering enormous peak performance, the applications likely to be able to take advantage of many (8-32) cores per chip are small in number. One approach that we feel has an excellent chance of doing so, is artificial neural networks[13]. Indeed, artificial neural networks have already been used profitably in Cancer Diagnosis[14, but we propose for the first time, to use a truly massively parallel systems (Intimidata) capable of enormous data movement for both near term experiments and as preparation for future massively parallel multi-core systems which we expect to dominate computing by the end of this decade.

References

- [1] Enzo – AMR Cosmological Simulation Code, cos-mos.ucsd.edu/enzo.
- [2] D. Komatitsch, J. Ritsema, and J. Tromp, “The Spectral-Element Method, Beowulf Computing and Global Seismology,” *Science*, vol. 298, 2002, pp. 1737-1742.
- [3] M. Vysokhid and K. Mahesh, “Large-eddy simulation of propeller crashback,” RTO AVT-123, Proc. Symposium on Flow Induced Unsteady Loads and the Impact on Military Applications, Budapest, 2005.
- [4] A.S. Szalay, The National Virtual Observatory, in ASP Conf. Ser., Vol. 238, *Astronomical Data Analysis Software and Systems X*, 2001.
- [5] C.M. de Vos, K. Van der Schaaf, J.D. Bregman, “Cluster Computers and Grid Processing in the First Radio-Telescope of a New Generation,” Proc. 1st International Symposium on Cluster Computing and the Grid, Brisbane, Australia, 2001.
- [6] N.R. Adiga, et al., “An Overview of the BlueGene/L Supercomputer,” Proc. SC2002 on High Performance Networking & Computing Conference, Baltimore, MD, 2002.
- [7] F. Schmuck and R. Haskin, “GPFS: A Shared-Disk File System for Large Computing Clusters,” Proc. Conference on File and Storage Technologies (FAST’02), Monterey, CA, 2002, pp. 231–244.
- [8] M.W. Margo, P.A. Kovatch, P. Andrews, and B. Banister, “An Analysis of State-of-the-Art Parallel File Systems for Linux,” Proc. 5th International Conference on Linux Clusters: The HPC Revolution 2004, Austin, TX, 2004.
- [9] General Parallel File System Administration and Programming Reference Version 2.3, SA22-7967-02, IBM Corp., 2004, pp. 27-30. S. Haykin, editor. *Unsupervised Adaptive Filtering vol.1 : Blind Source Separation*, John Wiley and Sons, New York, 2000.
- [10] Massive High-Performance Global File Systems for Grid computing, Phil Andrews, Patricia Kovatch, Christopher Jordan. SuperComputing 2005, Seattle, Washington, Nov. ‘05.

- [11] Computational Support for Pathology Content Based Image Retrieval, A.W. Wetzel, P.L. Andrews, M.J. Becich and J. Gilbertson, Journal of Supercomputing, 1997
- [12] Single-ISA Heterogeneous Multi-Core Architectures for Multithreaded Workload Performance, Kumar, R. Tullsen, D. Ranganathan, P. Jouppi, N. Farkas, K., ANNUAL INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE , Bibliographic details 2004, VOL 31, pages 64-75
- [13] Neural Networks: A Comprehensive Foundation, Simon Haykin, Prentice Hall
- [14] Artificial Neural Networks in Cancer Diagnosis, Prognosis, and Patient Management, Naguib and Sherbert, Editors, CRP