# TX-2500 – An Interactive, On-Demand Rapid-Prototyping HPC System

Albert Reuther, Bill Arcand, Tim Currie, Andy Funk, Jeremy Kepner,
Matthew Hubbell, Andrew McCabe, and Peter Michaleas
{reuther, warcand, currie, afunk, kepner, mhubbell, amccabe, pmichaleas}@ll.mit.edu
MIT Lincoln Laboratory, Lexington, MA 02420

May 21, 2007

## Introduction

Rapid prototyping of advanced signal processing algorithms is critical to addressing current and emerging threats. Core algorithm development requires interactive high performance computing using high level programming languages and tools. Lincoln has designed, implemented and deployed the LLGrid technology to meet this need.

MIT Lincoln Laboratory (http://www.ll.mit.edu/) has just enabled its next generation LLGrid interactive/on-demand parallel computing system [1]. This system was provided to Lincoln Laboratory via a DoD High Performance Computing Modernization Program (HPCMP) Distributed High Performance Investment in collaboration with Dell Computer, Inc. The LLGrid system consists of ~1500 InfiniBand connected processors, ~6 Terabytes of RAM and ~0.8 Petabytes of local disk/virtual memory.

Lincoln Laboratory will use this system to develop, prototype, and transition next-generation signal and image processing algorithms for DoD applications. A critical element of the algorithm development process is interactive test and refinement, which require interactive/on-demand access using high-level programming environments. LLGrid supports 200+ users at Lincoln Laboratory, ~85% of whom run parallel MATLAB® codes using the Lincoln Laboratory developed pMatlab library (http://www.ll.mit.edu/pMatlab) [1] or The Mathworks developed Distributed Computing Toolbox® (DCT) (http://www.mathworks.com/products/distribtb/).

An important LLGrid subsystem is the TX-2500, which consists of over 400 servers each with a full 6-drive RAID storage system. This provides ~0.8 Petabytes of local high bandwidth storage for sensor data. It also provides a unique experimental platform for testing next-generation parallel file system technology. Using the Lincoln Laboratory pMatlab XVM (eXtreme Virtual Memory) software [2], the entire storage can be treated as a single large global array of data enabling data sets as large as 50,000 x 50,000 x 50,000 grid elements to be processed. [Note: The TX-2500 is named in honor of the TX-0 (http://en.wikipedia.org/wiki/TX-0) developed at Lincoln in the 1950s, the world;s first interactive high performance computing system.]

This work is sponsored by the United States Air Force under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

## The TX-2500 System

The TX-2500 cluster system was designed to address interactive algorithm development for DoD sensor processing systems. Regarding the hardware system components, each compute node of TX-2500 consists of a Dell PowerEdge 2850 with dual 3.2 GHz Xeon processors, 8 GB of RAM, dual gigabit Ethernet interfaces, and a 4 Gbps InfiniBand interface. The processors were chosen for computational performance and cost effectiveness, while the RAM and networks were chosen to keep large datasets near the processors and transport large datasets, respectively. In addition, each compute node also has a 6-disk RAID with a hardware RAID controller implementing a RAID-5 configuration for a total of 2,592 disks. This design decision was made to temporarily store data sets from the central file server and to take advantage of Lincoln-developed pMatlab eXtreme Virtual Memory (XVM) toolbox [2], as mentioned above. In terms of TX-2500 service nodes, the scheduler uses LSF-HPC version 6.2 for resource management and scheduling; the cluster provisioning and management is handled by NPACI Rocks 4.2; and the central shared file storage uses the IBRIX fusionfs 2.0 parallel file system serving 36 TB of high-speed storage from a dual-headed Data Direct Networks S2A-8500 storage array with 160 data disks.

The software configuration is based on the Lincoln LLGrid software infrastructure, which consists of most of the HPCMP Baseline Configuration (http://asc.hpc.mil/consolidated/bc/), MPIch, LAM MPI, OpenMPI for Infiniband, several compilers, etc., plus additional software to support interactive algorithm development including the three Lincoln-developed MATLAB toolboxes: pMatlab, MatlabMPI, and gridMatlab [1].

For rapid prototyping in MATLAB these technologies have combined to create a unique interactive, on-demand grid computing experience, whereby running a parallel MATLAB job on LLGrid is identical to running MATLAB on the desktop. Users can use LLGrid from Windows, Linux, Solaris, and Mac OS X computers with their desktop computer becoming a personal node in the LLGrid thereby establishing a transparent interface between the user's computer and the grid resources. LLGrid is enabling faster algorithm development, prototyping, and validation cycles for Lincoln staff.

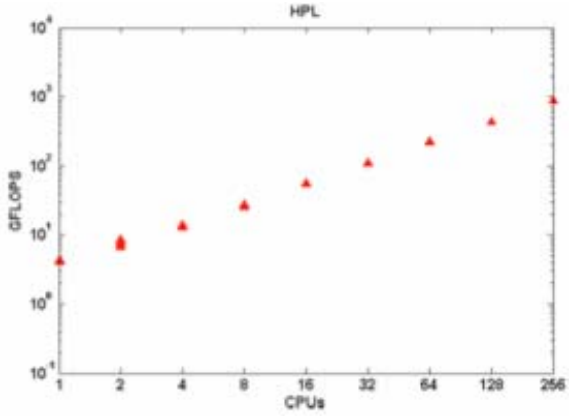


**Figure 1: HPL benchmark results for varying number of CPUs.**

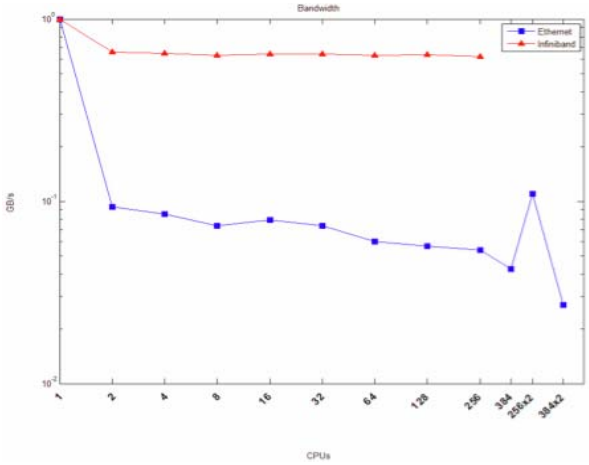


**Figure 2: InfiniBand throughput during Random Access benchmark for varying number of CPUs.**

## Results

The DoD HPCMP-mandated acceptance test was conducted using the HPC Challenge benchmark suite (http://www.HPCchallenge.org). This benchmark suite is designed to stress a wide variety of parallel components: processor, memory, network bandwidth, and network latency. As part of the acceptance test, Lincoln Laboratory ran 170 variations using different processors and memory sizes. This baseline performance data is now publicly available at the HPC Challenge website.

The results of the system acceptance test give us a clear quantitative picture of the capabilities of the system. The system acceptance test for the TX-2500 hardware involved an array of benchmarks that exercised each of the components of the memory hierarchy, the Infiniband network, and the disk arrays. These benchmarks isolated the components to verify the functionality and performance of the system. The memory hierarchy was evaluated using the HPC Challenge benchmark suite, which is comprised of the following benchmarks: High Performance Linpack (HPL, also known as Top 500), FFT, STREAM, and RandomAccess [3]. The Infiniband network was evaluated using b_eff, the effective bandwidth benchmark, which is also part of the HPC Challenge suite. The compute nodes' RAID arrays were evaluated with the iozone benchmark (http://www.iozone.org/).

Testing the memory hierarchy using the HPC Challenge benchmark suite on 416 processors yielded the following results: 1.42 TFlops (HPL), 34.7 GFlops (FFT), 1.24 TBytes/sec (STREAM Triad), and 0.16 GUPS (RandomAccess). We also were able to collect detailed data on performance as a function of number of processors and problem size as well as statistical data on the performance consistency under load. A plot of the HPL benchmark over a variety of CPUs is shown in Figure 1. For the numbers of CPUs shown, the performance of HPL scales linearly. The results of testing the InfiniBand network with b_eff (also part of HPC Challenge) [3] are shown in Figure 2. Here we see that the effective bandwidth in GB/s remains steady as we add processors, while the bandwidth over gigabit Ethernet deteriorates as processors are added. Further, we launched the iozone benchmark onto each node to exercise the 6-disk RAID arrays; on over 98% of the RAID arrays of the nodes, the performance baseline of 50 MB/s throughput for the read, reread, write, and rewrite tests was exceeded. For those nodes that did not pass, the RAID system went through troubleshooting and failure assessment. Overall, the battery of benchmarks that comprised our system acceptance test enabled us to not only determine the performance of our system, but it also revealed a number of hardware and configuration issues.

## References

[1] N. T. Bliss, R. Bond, J. Kepner, H. Kim and A. Reuther, "Interactive Grid Computing at Lincoln Laboratory", *Lincoln Laboratory Journal*, Volume 16, Number 1, 2006.

[2] H. Kim, J. Kepner, M. Vai, and C. Kahn, "Advanced Hardware and Software Technologies for Ultra-long FFTs," High Performance Embedded Computing (HPEC) Workshop 2005.

[3] J. Dongarra and P. Luszczek, "Introduction to the HPCChallenge Benchmark Suite," *ICL Technical Report*, ICL-UT-05-01, http://www.hpcchallenge.org/pubs/, 2005.