Redundant Data Storage and Data Processing Computer Hardware Solution for Mass Spectrometry Laboratories on a Budget

James West, Weiwei Tong, Yang Su, Catherine E. Costello, Mark E. McComb

Cardiovascular Proteomics Center, Boston University School of Medicine, Boston, MA.

Novel Aspects: Hardware, software and infrastructure considerations for deploying high-performance, cost-effective computing data processing and storage for proteomics.

Introduction: Mass spectrometry based proteomics yields large volumes of data in raw and processed forms. Data handling, processing and storage quickly overwhelms the typical capacities of modern PC workstations in a university environment. Recent advances in computer technology have made high performance computing available to individual laboratories. Here we describe infrastructure, hardware and software requirements for the design and implementation of an affordable and easily deployable high performance computing network for proteomics and bioinformatics analyses.

Methods: We explored hardware and software technologies available with respect to the needs of a proteomics laboratory. These included symmetric-multiprocessing (SMP) vs. blade and clustering solutions, AMD vs. Intel chip design, 32- vs. 64-bit design, storage technologies, operating system (OS) interoperability, local and network security policies, secure and virtual LAN, and integration within existing infrastructure.

Results: The resultant high performance compute solution consists of a RAID based data storage array server and two compute servers. The compute servers consist of 4-way AMD Opteron 880 dual-core chips at 2.4 GHz on a Tyan motherboard with 32 GB RAM operating in full SMP mode with 2 TB RAID 5 storage and dual gigabit data ports. BU Linux 4.5 Server Edition (Zodiac) and Windows 2003 Server Enterprise R2 were chosen to fully support software deployment and programming development. The servers host MASCOT (Matrix Science), ProteinLynx Global Server 2.2 (Waters Corporation) and BUPID for data processing. The redundant data storage array server utilized two AMD Opteron 2216 processors (2.4 GHz), 4GB RAM, dual gigabit data ports, with 4 TB of local disk space; RAID 6 double parity, and a separate 40 GB RAID 1 array for Windows 2003 Server R2. Laboratory integration with instrumentation is via virtual LAN and integration via the BUSM infrastructure is via Active Directory. To evaluate system performance we analyzed existing LC-MS/MS data sets. We processed data in identical runs and compared results to those obtained using a Dell Optiplex GX620. Additional benchmarking was by Sandra Lite 2007. Data processing tests showed a 10-fold increase in floating point operations per second (FLOPS) and a 6-fold increase in overall performance for the 4-way system. System and network stability were obtained.

Conclusions: High powered compute solutions with large redundant storage capacities are no longer financially and technically out of reach of most scientific laboratories.

Acknowledgement: This project was funded by NIH NHLBI contract N01-HV-28178.