

# A Highly Scalable, Parallel File System for Biological Sciences

Yeo Eng Hee  
(HPC, Computer Centre)

A recent article (Turner, 2009) highlighted the speed in which data is being created in the field of gene sequencing. When the first human genome was fully sequenced, it took the team 13 years to complete. Today, scientists are able to accomplish the same task in less than 10 days! The explosion of data is also seen in the size of the databases in the US National Institute of Health's Genbank: 2 billion base pairs in 1999, 11 billion in 2000 and 86 billion in 2008 (May, 2009). And the size will continue to grow, spurred on by the technological advances that enable the scientists to generate data in the order of Terabytes per day. A good gauge of the size of the data deluge that scientists are facing is given in a white paper by DataDirect Networks (Data Direct Networks, 2009) on their storage solutions for biological science applications, where comparisons were made between the amount of data generated by bio-analytic equipment today and that generated in the past:

Bio-Analytic Equipment Output	
Past	Present
One Run = 13GB 400,000 Bases/Run	One Run = 1TB 1 Billion Bases/Run

Table 1

In anticipation of the demand for more storage by researchers from the biological sciences and other departments in NUS, the High Performance Computing team in Computer Centre will be implementing a total of 120TB of storage to be attached to the computational resources here. As mentioned in the previous issue of HPC@NUS, the storage will comprise three highly scalable, parallel file systems, each having 40TB of disk space (Figure 1). The scalable nature of the parallel file system will allow Computer Centre to grow the file system according to the demands of the users.

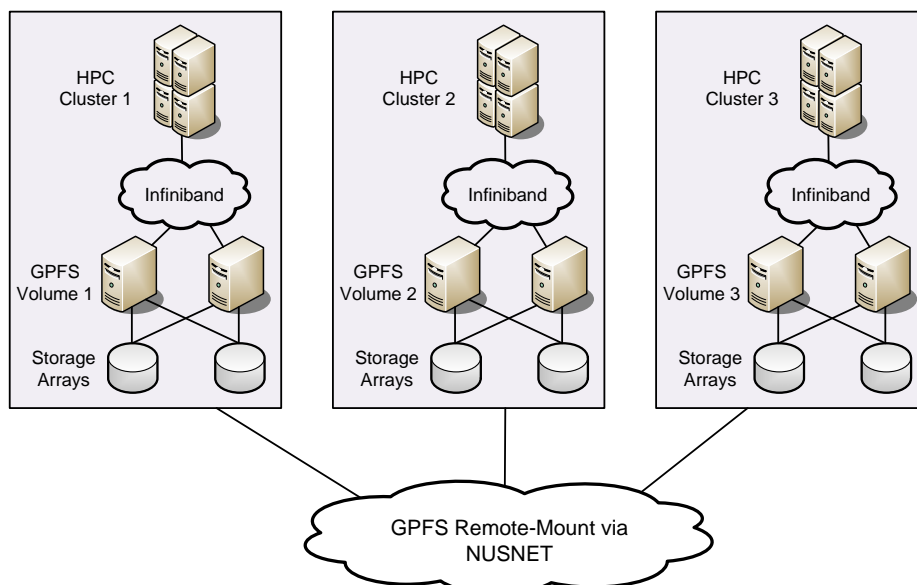


Figure 1

## Works Cited

Data Direct Networks. (2009 31-July). *Optimized Storage for Life Sciences Data (White Paper)*. Retrieved 2009 19-October from <http://www.ddn.com/lifesciences>:  
<http://www.ddn.com/pdfs/Life.Sciences.WP.073009.pdf>

May, M. (2009 24-September). *Scientific Data Lifecycle Management: Preparing for Storage in an Uncertain Future*. Retrieved 2009 24-September from Bio-IT World - White Papers and Special Reports: <http://www.bio-itworld.com/BioIT/WhitePapers.aspx>

Turner, J. (2009 13-July). *Sequencing Genome in a Week*. Retrieved 2009 24-September from O'Reilly Radar: <http://radar.oreilly.com/2009/07/sequencing-a-genome-a-week.html>