U.S. Department of Energy Best Practices Workshop on File Systems & Archives San Francisco, CA September 26-27, 2011 Position Paper

David Cowley
Pacific Northwest National Laboratory
david.cowley@pnnl.gov

ABSTRACT / SUMMARY

The EMSL facility, located at Pacific Northwest National Laboratory (PNNL), operates terascale HPC and petascale storage systems to support experimental and computational researchers in molecular sciences. This position paper addresses the Workshop's Business of Storage Systems track and describes EMSL's approach to operating file systems and data archives.

INTRODUCTION

The Environmental Molecular Science Laboratory (EMSL) is a scientific user facility located at PNNL. EMSL houses PNNL's largest concentration of high performance computing systems and data storage systems. While other organizations within the Laboratory are working on obtaining their own significant HPC and data storage resources, the center of mass has not shifted yet. We will be careful in this document to distinguish between PNNL and other sub-organizations within PNNL, including EMSL.

EMSL operates a suite of cutting-edge scientific instruments, capable of generating terabytes of data per week. EMSL has operated HPC systems ranked in the top 20 of the Top500 list since 2003, in addition to a multi-petabyte archive for scientific and computational data. EMSL has been working since 2010 on a scientific data and metadata management system known as MyEMSL.

GENERAL APPROACH TO STORAGE SYSTEMS

EMSL HPC systems have had more types of filesystems than at most HPC sites. Each is intended to meet different levels of capacity, performance, and accessibility. So far each of EMSL's HPC systems has been procured with its own filesystems of 3 types:

Filesystem Type	Capacity	Nominal Bandwidth
Global Home	20 TB	1 GByte/sec
Global Scratch	277 TiB	30 GiByte/sec
Node Scratch	350 GiB/node 808 TiB Aggregate	400 MiByte/sec/node 924 GiByte/sec Aggregate

The global home filesystem is available to all nodes in the HPC cluster. its capacity is determined loosely by a "not too big to be backed up" rule of thumb, its performance is determined loosely by a "'cd' and 'ls' commands have to not be slow" rule of thumb.

The global scratch filesystem is a parallel filesystem both larger and higher performance than the home filesystem. It is available to all nodes in the HPC cluster, and its performance and capacity requirements have been derived from a formula based on the theoretical peak

performance of the system. EMSL does not have a requirement to checkpoint whole-system jobs as some other sites do, so this eases some of the requirements on this filesystem.

The node scratch filesystems provide high disk bandwidth per Flop to each node. For these filesystems, performance in terms of write bandwidth and Ops/second are again derived from theoretical peak performance on the compute node. Capacity has been a side effect of the need to provision enough disk spindles to meet the required performance. This may change as magnetic disk and solid-state disk technologies evolve. While providing a scratch filesystem on each compute node does involve considerable cost and added maintenance, the aggregate performance has been more scalable and better in absolute terms than shared parallel filesystems. EMSL has found this to be a differentiating and enabling capability, and will carefully consider it in its upcoming system procurements.

EMSL is planning to move to a "two systems" approach where rather than procuring one large system every 3 to 4 years, we will procure smaller systems every two years and overlap their lifecycles. We will switch to having the home filesystem shared between compute clusters. We expect that each cluster will have its own high performance parallel global scratch filesystem. We will consider critically whether new systems require node scratch filesystems.

MANAGING ARCHIVE GROWTH

EMSL's growth in archive capacity is driven by two factors, the output of scientific instruments and the output of its HPC systems. In effect, the scientific instruments are computers themselves, as is the HPC system, so Moore's law drives data growth rates in both cases. Fortunately magnetic media growth rates (sometimes cited in "Kryder's Law") are on a similar trajectory so storage systems likewise exhibit the behavior of offering twice the capacity for roughly the same cost year over year. This behavior is expected to continue through 2020 [1,2].

This allows us to provide space for exponential data growth as long as a relatively consistent storage budget is available year-to-year. Successive generations of storage have so far had the sheer capacity to swallow up data from earlier generations of technology, provided there is a bridge between the technologies. Ensuring that there is such a bridge between generations is feasible provided there is sufficient planning and investment both in time and dollars to execute it.

Exponential growth rates *are* sustainable with proper planning and funding, but this only provides for storage *space*. By itself, this does not address the problems of managing, understanding, or using the accumulation of data. To that end, EMSL is investing in creating a new scientific data and metadata system known internally as MyEMSL. MyEMSL is addressed in the PNNL position paper for the Usability of Storage Systems track.

SOFTWARE FOR FILE SYSTEMS AND ARCHIVES

EMSL uses the software technologies that best fit its needs and budget, whether open source or proprietary. As much as possible, we wrap proprietary solutions so that they play well in an open-source environment. We were an early adopter of the Lustre filesystem, having used it since the implementation of our MPP2 system in 2003. We have built low-cost filesystems out of commodity hardware up to 1.2 petabytes (the "NWfs" storage system in 2008), and PNNL is building a similar institutional Lustre storage system that will have a 4-petabyte capacity by the end of fiscal year 2011. In 2008, EMSL identified a need to implement a hierarchical storage system, and in 2009 retired NWfs in favor of a new HPSS system.

HPSS provides the right mix of capacity, expandability, and scalable performance for EMSL's needs. The EMSL HPSS system provides archive storage capacity, and we have implemented open source filesystem-like interfaces to it, in addition to the traditional native HPSS interfaces.

EMSL and the rest of PNNL continue to make use of Lustre and will continue to do so until it is clearly dead or orphaned. At this point, PNNL

has enough experience and expertise to not require Lustre support. Even if advanced and long-promised features (e.g. multi-way clustered metadata) are never delivered, Lustre's cost, performance, scalability, and "good enough for us" reliability meet our needs very well.

The difficult to control costs are in additional work scope, i.e. supporting more systems or more users without attendant increases in budgets. Inflation alone causes increased labor costs over time, creating difficulty in operating with flat or declining budgets. Additional work scope compounds this problem if not very carefully managed.

HARDWARE FOR FILE SYSTEMS AND ARCHIVES

Being at the upper-mid range of HPC in terms of system sizes and performance, EMSL is not using and does not expect to use custom hardware in the foreseeable future. We take the best advantage we can of common off the shelf hardware and the economies of scale that come with it.

EMSL does expect to continue to take advantage of commodity storage technologies for the foreseeable future, mostly in conjunction with the Lustre filesystem. Selected high-value storage systems may be constructed of enterprise-grade storage for serviceability features. While we may apply creative engineering approaches to commodity or enterprise-grade building blocks, we do not foresee significant use of custom storage hardware.

I/O capacity and bandwidth requirements for filesystems on EMSL HPC systems are established as a function of peak performance ratings. We have not carefully specified metadata operation or operations/second requirements on our filesystems, though we have re-engineered metadata servers to improve performance when there is a need to do so. MTTI requirements have not been rigorously specified either, though we do specify that common failures (e.g. single disk failure on a node) must not interrupt computation or I/O. During technical review, we assess whether the I/O system is robust enough to remain serviceable with good maintenance procedures. It has been said, "we don't need five nines, we just need two or three!"

Most of the barriers we see to adoption of commodity storage have to do either with low performance or lack of Reliability, Availability and Serviceability (RAS) features. In our experience, neither of these has presented insurmountable difficulties. The engineering approaches and software tools we apply allow performance to be scaled linearly (or nearly so) by adding more components. The essential RAS features we need are typically available in mid-grade commodity or enterprise hardware. At the other end of the spectrum, many higher end RAS features such as active-active failover/failback prove to cause as much downtime as they are advertised to prevent!

SYSTEM EVOLUTION

EMSL plans a three to four year lifetime for its HPC systems, and has recently decided to switch from operating one large HPC system to two smaller systems with overlapping lifecycles. With this change, we will pull the persistent "home" filesystem out of the cluster and place it where it can be shared between systems and provide continuity between HPC systems as they age out and are replaced.

EMSL procured a new HPSS storage system for archive purposes in 2009, and plans to operate it through at least 2017, with planned lifecycle replacements and technology refreshes for the storage (disk and tape) components.

CONCLUSIONS

EMSL employs multiple tiers of data storage systems with different capacity and performance characteristics to satisfy various needs. Storage system capacities are planned based upon projected output from the facility's scientific instruments and from HPC system performance. All storage systems have a planned lifecycle with expansions, technology refreshes, and retirement as appropriate. EMSL generally uses commodity or enterprise-class components as building blocks, in concert with a mixture of open source and proprietary software.

REFERENCES

1. Kryder, H, Kim, C. *After Hard Drives – What Comes Next?*. IEEE Transactions on Magnetics, Vol. 45, No. 10, October 2009 *http://www.dssc.ece.cmu.edu/research/pdfs/After_Hard_Drives.pdf* .
2. Zyga, L. *What Comes After hard Drives?* http://www.physorg.com/news175505861.html .