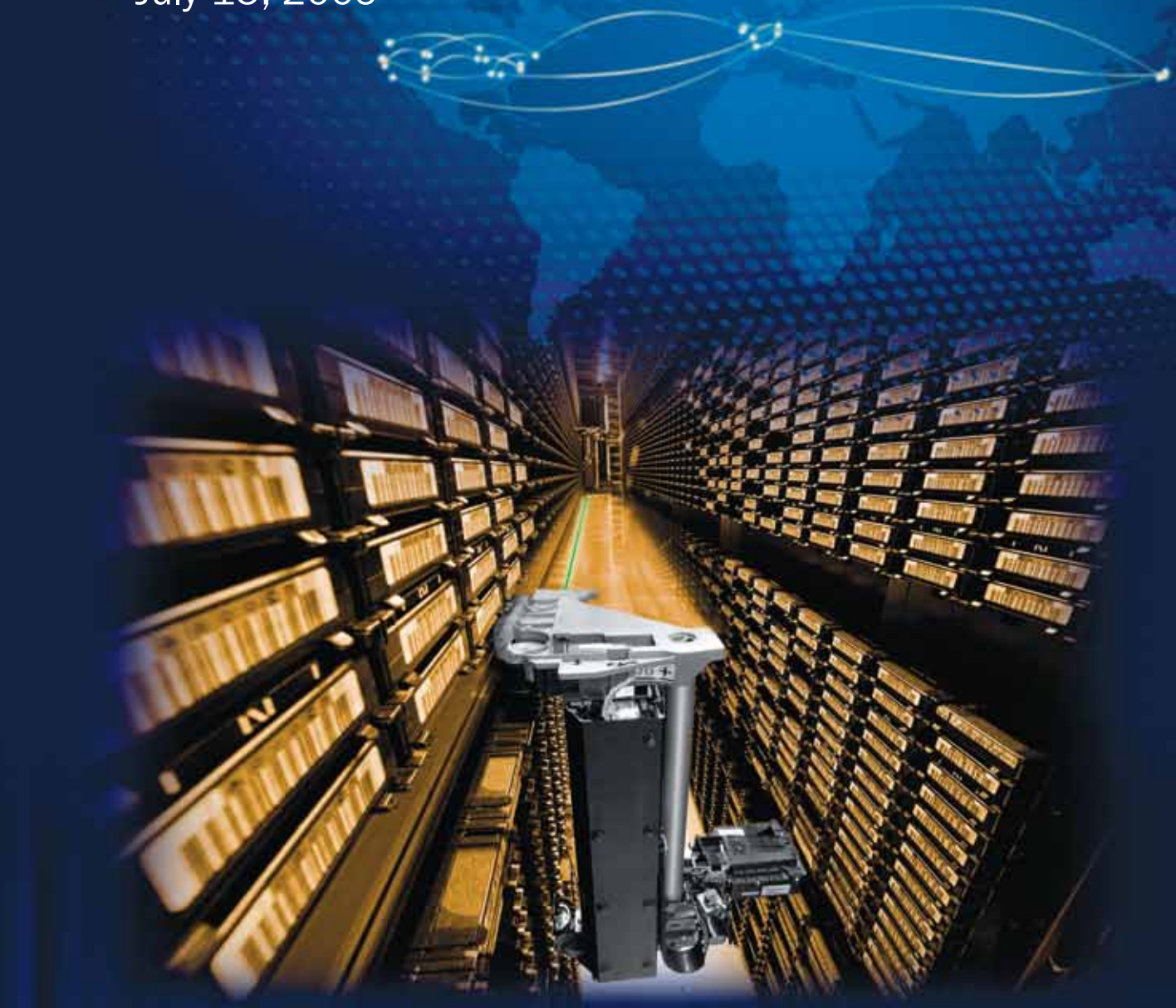
HPSS in the Extreme Scale Era

Report to DOE Office of Science on HPSS in 2018-2022

July 15, 2009



U.S. DEPARTMENT OF ENERGY
Office of Science

NNSA

Lawrence Berkeley National Laboratory

Sandia National Laboratories

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY

IBM

Extreme Scale Workshop

National Energy Research Scientific Computing (NERSC) Facility
Oakland, CA July 14-15, 2009

Abstract

This paper is a product for the Department of Energy's (DOE) Office of Science (OS) reporting on the feasibility of using HPSS into the Extreme Scale era of storage (2018 - 2022). The initial sections provide a summary of the systems environment and expected archival storage requirements extracted from other Extreme Scale workshop reports conducted since 2007 by various applications and programs within the DOE OS. These high level requirements aid in identifying long-term data storage system features that support Extreme Scale science. Participants also separately forecasted data growth in established long-term data storage systems through 2018 - 2022 to get a picture of the amount of data that systems will need to manage. The report concludes that HPSS is well positioned to meet the requirements projected for the Extreme Scale era and provides recommendations from the HPSS Collaboration to the DOE Office of Science for ensuring that HPSS can meet these extreme scale storage requirements of 2018 - 2022.

Workshop Participants:

Danny Cook, LANL
Jason Hick, LBNL (chair)
Jim Minton, LLNL
Henry Newman, Instrumental
Timothy Preston, SNL
Gary Rich, LANL
Cody Scott, LANL
Jerry Shoopman, LLNL
John Noe, SNL
Jack O'Connell, ANL
Galen Shipman, ORNL
Dick Watson, LLNL (co-chair)
Vicky White, ORNL

Introduction and Executive Summary

The Extreme Scale (ES) supercomputing era is expected around 2018 - 2022 when the first exaflop system is delivered to a Department of Energy (DOE) center. The workshop is a result of the Office of Science's interest in the role of HPSS in the Extreme Scale era, in particular, the requirements and development necessary to meet Extreme Scale demands.

This report provides requirements for Extreme Scale archival storage derived from a separate report produced as input to the workshop, an independent market survey, and the workshop conducted by the HPSS collaboration from 14-15 July 2009 at the Oakland Scientific Facility, Lawrence Berkeley National Laboratory. Participants included archival storage executive and technical leaders at ANL, LANL, LBNL, LLNL, ORNL, and SNL. IBM and Instrumental provided input to the workshop on the commercial environment.

Strategic archival storage[1] planning relies on understanding:

- Site projections for systems and storage (Section 1) [2]
- Archival storage requirements as gathered from other workshop reports (Section 2)
- Archival storage industry hardware (Section 3)
- Software (Section 4) projections

The Extreme Scale era of computational science promises to usher in new grand challenge requirements that will require an increase in research and development resources. After more than fifteen years, the HPSS collaboration is very active and remains one of the prime examples of a successful agreement between industry and the government.

The central conclusion of this workshop is that HPSS' scalable architecture is viable for Extreme Scale archival storage. It will, however, require increased development funding to meet new scalability and data management needs of Extreme Scale computing, while presenting a manageable, highly resilient archival storage system to system administrators at sites using industry-leading hardware of the day.

The key recommendation from this workshop is for the Office of Science to increase its funding of HPSS development and to increase collaboration with the large-scale scientific data management community. Archival storage is crucial to Extreme Scale computing. We outline below why we think HPSS is the leading candidate for meeting Extreme Scale archival storage. It will however need development work detailed in this report. Due to its leadership role in Extreme Scale computing, it is critical that the Office

[1] Archival storage refers to systems capable of long-term data retention, with minimal cost, and hardware that protects data at rest for decades. This is not synonymous with Hierarchical Storage Management features or systems.

[2] Primarily the *Infrastructure Plan for Advanced Simulation and Computing (ASC) Petascale Environment* provided by the DOE National Nuclear Security Administration (NNSA), Feb 2009, and the *Defense Advanced Research Projects Agency (DARPA) Exascale Computing Study*, Sep, 2008.

of Science provide leadership in closing the gaps in HPSS to support extreme scale computing.

Section 1. The Extreme Scale System Environment

This section describes how system architectures affect archival storage system requirements. These are derived from proposals of Extreme Scale architectures from the DARPA ExaScale Platform Study, known platform acquisition plans for both DOE OS and NNSA labs, and current archival storage plans.

1.1 Impact of System Architecture on Archival Storage

Based on historical data and experience two key aspects of computational systems affect the amount of data and the sizes and numbers of files sent to archival storage. The first is the amount of system memory and the second is the number of processing cores within the system.

The amount of system memory is the main determinate of the amount of data stored in the archive each year. Through analysis of years of statistics comparing the total system memory and new archival data stored, workshop participant sites have determined that on average every 1 TB of main memory results in about 35 TB of new data stored to the archive each year (See Table 2 – more than 35 TB per 1 TB actually get stored, but 20 – 50% actually gets deleted on average over the year). This ratio has held for the past eight years as we have transitioned from Terascale to Petascale computing, and we expect this trend to continue into the Extreme Scale Era.

The number of processor cores affects the total number and sizes of files stored in the archive each year. As systems continue to increase in both number of cores and processors, the archival storage systems continue to see ever increasing numbers of files. There is no specific planning number that works for each workshop participant site, mostly due to the variability of projects in their use of file aggregation[3] I/O libraries or aggregation capable archival storage clients. We expect this uncertain pattern to continue into the Extreme Scale Era.

1.2 DARPA Exascale System Architectures

Example Extreme Scale systems from the DARPA report[4] show three major systems architectures, which if built and delivered will have significant impact on archival storage requirements: (1) an aggressive silicon-based system that is primarily GPU based, (2) a data center class system that is focused more on

[3] Aggregation is defined as combining some number of files into a single file. The goal is to remove per-file overhead associated with data transfers and instead allow for optimal data transfer rate (e.g. streaming).

[4] *Defense Advanced Research Projects Agency (DARPA) Exascale Computing Study*, Sep, 2008

data analysis rather than computation, and (3) an evolutionary heavy-node system that represents scaling current cluster-based supercomputers to Extreme Scale. There was also a light node proposal, but it was not possible to determine the specifics of the system beyond memory. Table 1 provides the characteristics that impact archival storage of the proposed systems.

Extreme Scale A (Aggressive Silicon System)	
Peak Performance	1 Exaflop
Memory	3.6 Petabyte
Processors	223,872
Cores	166,113,024
Extreme Scale B (Data Center Sized Aggressive Silicon System)	
Peak Performance	300 Petaflops
Memory	1 Petabyte
Processors	67,968
Cores	50,432,256
Extreme Scale C (Evolutionary Light Node Strawman)	
Peak Performance	1 Exaflop
Memory	140 Petabyte
Processors	N/A
Cores	N/A
Extreme Scale D (Evolutionary Heavy Node Strawman Power Unconstrained)	
Peak Performance	1 Exaflop
Memory	300 Petabyte
Processors	297,250
Cores	12,000,000

Table 1. Potential Extreme Scale Systems

The amount of memory is correlated with the amount of data that might be generated by these systems. The numbers of files that one might expect a given architecture to generate cannot be determined because we have no metric that holds across systems, primarily due to differences in applications and the software environment (e.g. using HDF5 to aggregate data). There is however a loose correlation between the numbers of processors and the numbers of files generated. We added these proposed systems to Table 2 below to show the expected impact on archival storage.

1.3 System Architecture Impact on Storage Planning

This section provides information on workshop participant computational system acquisition plans through 2018 – 2022 (as of the workshop date) and how they impact archival storage. The overview of archival storage planning shows how computational systems and file systems impact archival storage. Then using past historical usage, which includes past orders of magnitude improvements in computing and storage, it shows what the centers expect their archival storage systems to manage in terms of amount of data and numbers of files through 2018 – 2022.

1.3.1 Computational System Acquisition Plans

Workshop participants provided their platform acquisition plans through 2018 - 2022, as they were understood at the time of the workshop. See Figure 1 below for planned major computational system acquisitions at the workshop participant sites (ANL, LANL, LBNL, LLNL, ORNL, and SNL). Knowing the planned memory footprints is most critical to forecasting the impact on archival storage systems.

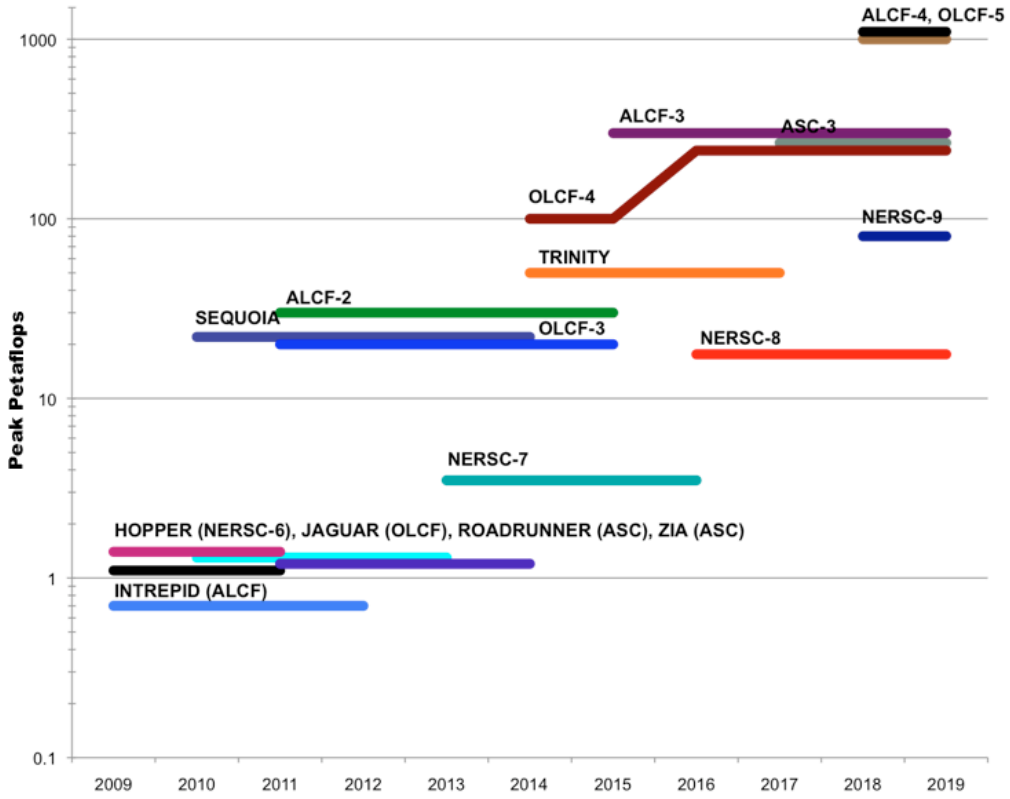


Figure 1. Workshop Participant Computational System Plans

Assuming the ratio of 1 TB of main memory to 35 TB of annual archive growth, Table 2 provides the expected total system memory for planned systems at workshop participant sites and estimates for future systems, based on the DARPA ExaScale Platform Study.

1.3.2 Workshop Participant Archival Storage Plans

The table provides a projection for the amount of total I/O the archival storage system can expect to see per year from each system. It also provides a minimum average sustained bandwidth required for archival storage from those systems in order to handle the I/O expected from them. However, this average bandwidth is not indicative of the peak bandwidth required, as the latter must be significantly higher to offload data from the file systems so that the file systems in

turn are available for offloading main memory to avoid unacceptable idle time. Archival storage peak aggregate bandwidth has been historically planned to be 10% of file systems' peak bandwidth to meet the above needs. Further, the actual growth in the archive on a year-to-year basis does not indicate the total amount of data actually stored during the year, as certain amounts of the data are deleted during that period. Thus, a higher average bandwidth is indicated.

To reflect this operational fact, the last two columns in the table provide expected archival storage system growth relative to total I/O during the year in number of Petabytes. Workshop participant sites typically see archive annual growth between 50% and 80% relative to the total amount of data actually stored during the year.

	System Memory (TB)	Annual Archive I/O by System (PB)	Sustained Minimum Archive BW Needed (GB/sec)	50% of I/O Retained, Annual Archive Growth by System (PB)	80% of I/O Retained, Annual Archive Growth by System (PB)
Roadrunner	98	3.4	0.11	2	3
Jaguar XT5	300	10.5	0.34	5	8
Hopper	217	7.6	0.25	4	6
Zia	200	7.0	0.23	3	5
Sequoia	1,638	57.3	1.86	28	45
Trinity	3,072	107.5	3.49	52	84
Extreme Scale A	3,686	129.0	4.19	63	101
Extreme Scale B	1,024	35.8	1.16	18	29
Extreme Scale C	143,360	4,900.0	162.93	2,450	3,920
Extreme Scale D	307,200	10,752.0	349.13	5,250	8,400

Table 2. Projected Annual New Archive Data Written by System

The only numbers that are a departure from the norm are for the Extreme Scale C and D systems. Since achieving the bandwidth and capacity is mostly determined by hardware, this will be discussed in more detail in Section 4 on Storage Hardware. It is important to note that achieving these capacities and bandwidths is determined by having mature software that is capable of utilizing the underlying hardware at Extreme Scale.

1.3.3 Overview of Archival Storage Planning

Through decades of archival storage planning, the workshop participant sites identified the rule-of-thumb planning conventions illustrated in Figure 1. Most sites have not found a consistent metric for relating system flops to storage bandwidth or capacity. However, historically, capacity is strongly correlated with the amount of main memory. The right side of the figure provides storage capacity rules-of-thumb and the left side provides bandwidth rules-of-thumb. For capacity, workshop participants have found a strong correlation between amount of memory and data generated. The amount of local (file system) storage is

expected to hold a week's worth of full memory dumps. After plotting the amount of main memory and the amount of new archive data generated each year, workshop participants in 2009 averaged about 35 TB of new archive data for each TB of main memory. Capacity is especially important to archiving as it determines the amount of new data generated at the center that may be retained long-term.

Storage bandwidth planning is based on providing the necessary bandwidth and capacity to prevent the file system or computational resources from holding up application execution, that is causing the processing to go idle, while memory is offloaded to persistent storage or the file system is off-loaded to archival storage. The bandwidth for archival storage is similarly needed to ensure that the file system does not fill up and again cause processing to go idle while the file system is off-loaded to the archive. Further, it is important to remember that bandwidth for archival storage must be provisioned to handle more than one client, user or system as the archival storage systems are center-wide assets; it is common to expect data to move between archival storage and all other storage resources at the center (e.g. system memory, local and global file systems).

Most site planning aims for an aggregate bandwidth of local disk that is as fast as possible within bounds of affordability; this report shows 1/5th the memory bandwidth, which is generally achievable and affordable today given the cost of memory and disk. Global file systems predominantly aim to share data amongst systems and users such that bandwidth is not the main design constraint, leading to systems going idle. With affordability as a main factor, achieving a global file system capacity of 20% of the local storage is reasonable to achieve. Archival storage systems aim for an order of magnitude bandwidth decrease from local storage.

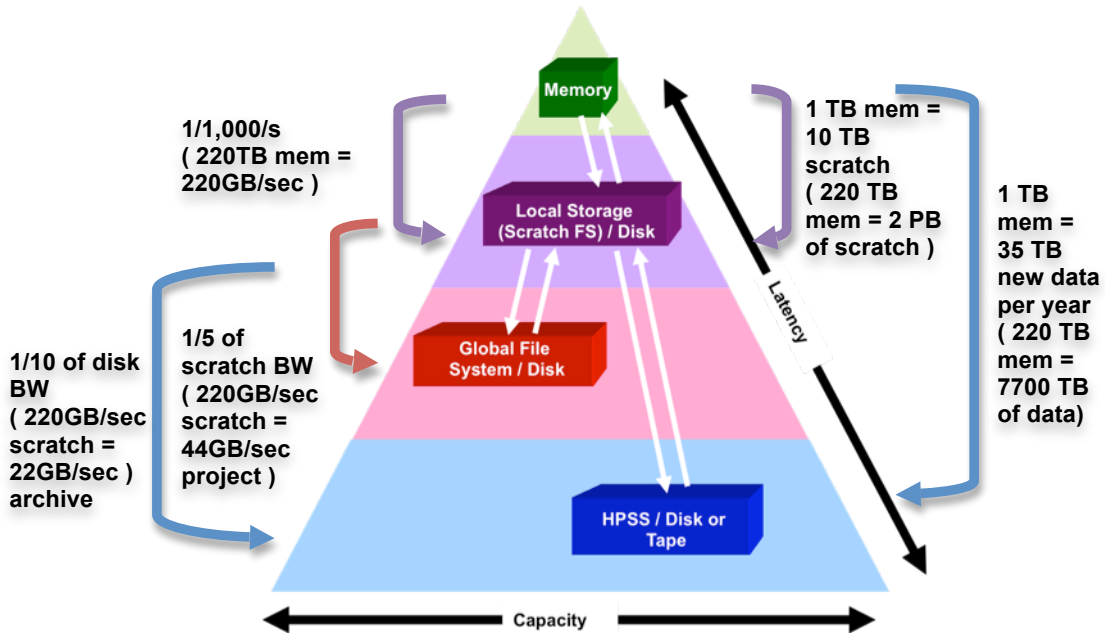


Figure 2. Conventional Storage Planning Guidelines

1.3.4 Workshop Participant Archival Storage Plans

Numbers of files at DOE labs now number in the 100s of millions. A key current problem is that of cost for sustaining growth of the archive. For archives that utilize small disk caches for smaller transfers and direct larger transfers to tape, the problem is buying enough and the right type (e.g. fast access) tape drives to serve all user requests, data migrations from disk, and data copies to new media or formats. For archives with large disk caches favoring quicker access, the cost problem is generally centered on buying enough disks to continually serve the growing number and size of file systems at the site.

Figure 3 below shows the local archival storage projections for amount of data through 2018 - 2022. The numbers are based on actual storage up through May 2009. The rate of growth is determined from examining all data previous to May 2009 for each system and then projected forward to 2018 - 2022.

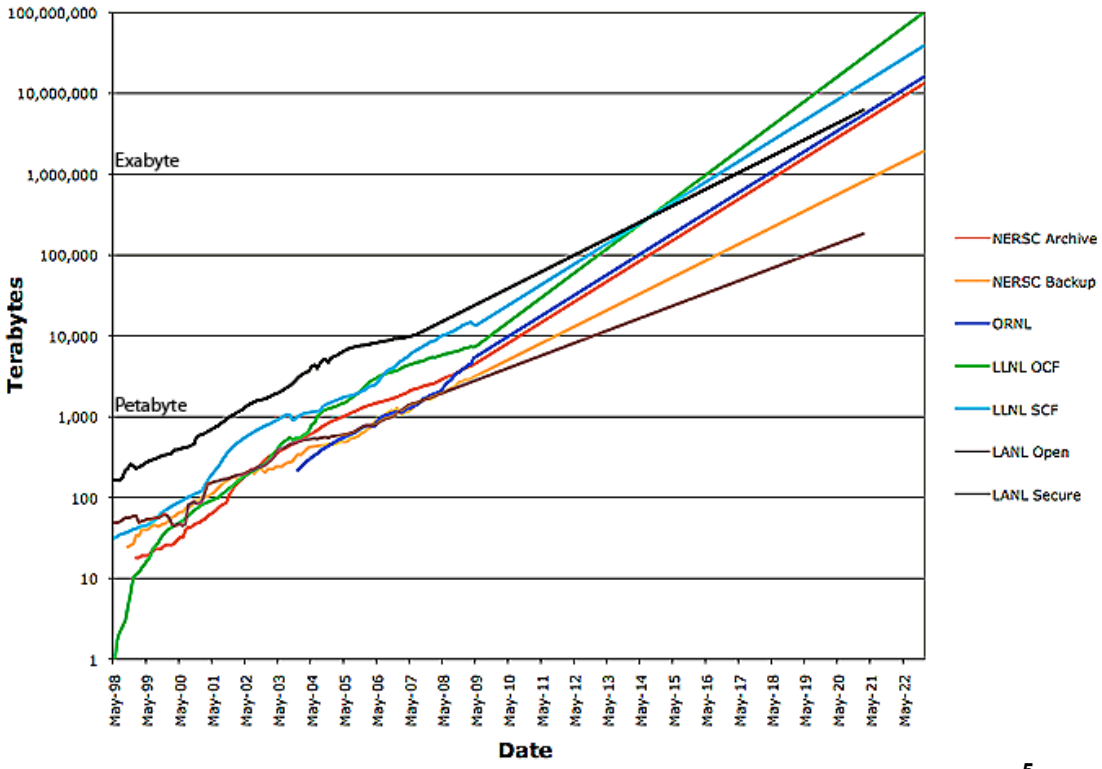


Figure 3. Archive Data Stored by DOE Lab through 2018 - 2022[5]

Figure 3 shows that if data growth in the next decade holds with data growth from the previous decade, workshop participant sites will have single archive systems that retain between 2 and 50 Exabytes of data.

[5] Figures 3 and 4 generated by input from workshop participants on total data stored and total numbers of files in each archive system by month from 1998 or inception to May 2009. The average growth rate for each determined projections from Jun 2009 through Dec 2018 - 2022.

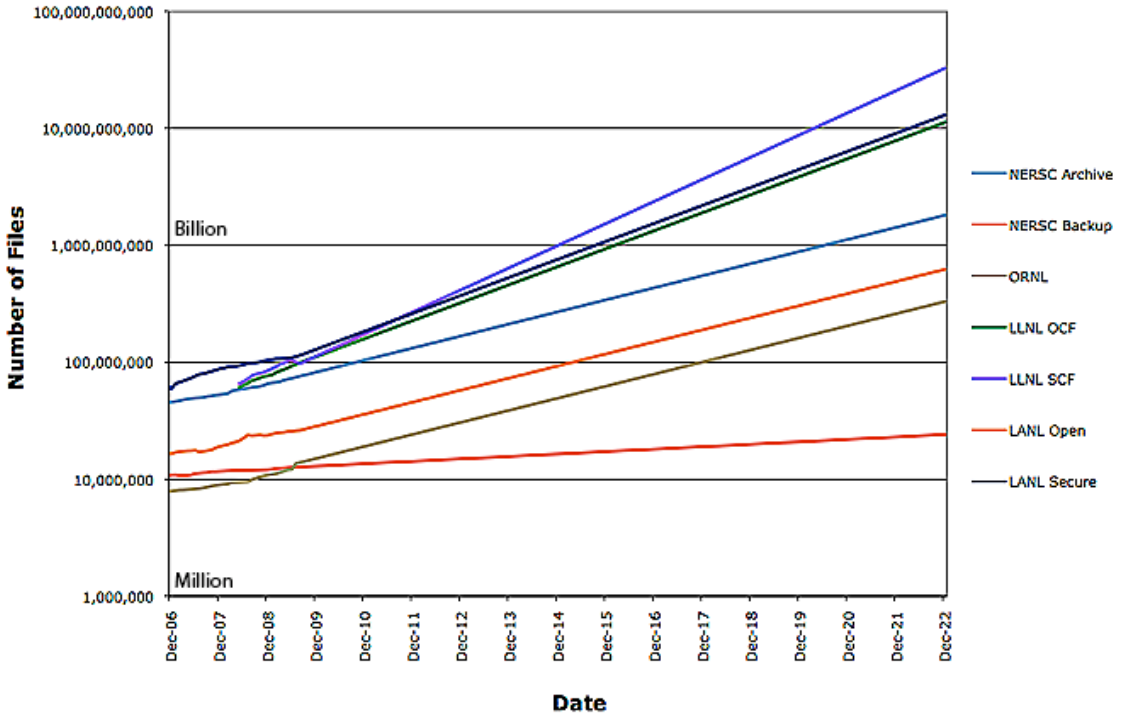


Figure 4. Archive Number of Files Stored by DOE Lab through 2018 - 2022

Over the next decade we expect to see archives with a total of 10s to 100s of billions of files at workshop participant sites.

Extreme Scale archival storage systems need to handle both very large numbers of small (Megabyte and below) files as well as handle extremely large files (Gigabyte and above). Current composition of archives at workshop participant sites is shown in figures 5 and 6.

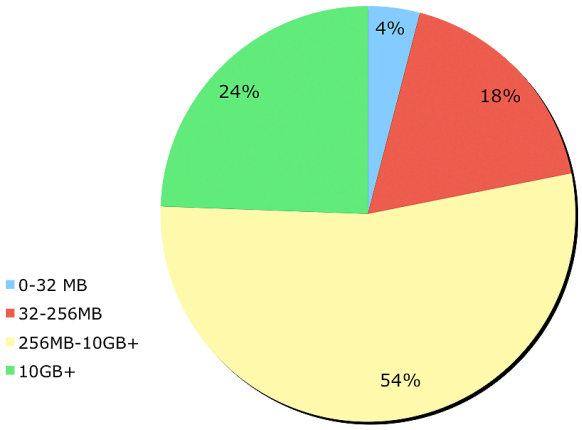


Figure 5. Archive Composition by File Size

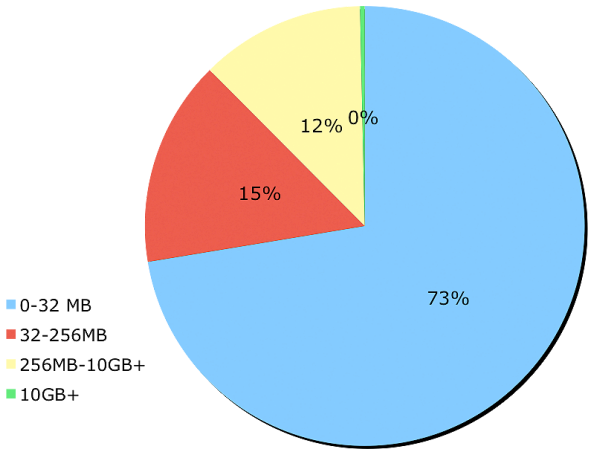


Figure 6. Archive Composition by Number of Files

Figures 5 and 6 provide LANL, LLNL, ORNL and LBNL archive statistics

combined, so that amount data and number of files from each site were added together to provide the distribution shown. The statistics differ only slightly across sites. They show that 78% of data at the sites is retained in files greater than 256 MB in size; that 87% of all files in the archives are less than 256 MB in size and 72% of all files are less than 32 MBs in size. The average file size will continue to increase as memory per node increases, but the ratio of small to large files is expected to remain the same. This is primarily due to the fact that the size of the systems in terms of memory and computational units is expected to continue to increase as well. I/O libraries and interfaces that perform aggregation will continue to be used but by less than a majority number of users.

Section 2. Extreme Scale Archival Storage Requirements

The workshop identified high-level requirements for meeting Extreme Scale data demands for archival storage. These requirements are used later in Section 4 of this report as criterion to evaluate archival storage software.

Through careful analysis of other available Extreme Scale workshop reports and site-specific long-range storage planning, the workshop determined the following top requirements for archival storage in the 2018 – 2022 timeframe:

Requirement Category	Determined From
Scalability of: • Storage capacity • Database and metadata speed • Performance between systems (Network)	Extreme Scale Workshops
Data Management • Data discovery • Data mining • Data set operations	Extreme Scale Workshops
System resiliency • Usability of GUI for very large systems • Logging • Monitoring	Site-specific Planning
Storage Hardware • Affordability at scale • Performance at scale	Market Trend Analysis

Table 3. Archival Storage High-Level Requirements for 2018 - 2022

2.1 Scalability Requirements

2.1.1 Scalability of Total Storage Capacity and Bandwidth

Archival storage peak aggregate bandwidth is planned to 10% of file systems' peak bandwidth, as mentioned earlier. We are estimating that aggregate peak bandwidth for archival storage required will be 100's of Gigabytes/second given that file systems will be at Terabytes/second speeds. Single client bandwidth to

archival storage is determined by user file size and numbers of data connections available to each transfer for the given file size. User file size is trending upwards but over fifteen years has changed from several Megabytes (MB) to several hundred MBs. Because file size is trending modestly upward relative to the large increases in numbers of files and bandwidth of disk and tape devices, Extreme Scale sites will need to continue sending aggregated files to archival storage to maintain acceptable bandwidth. As well, parallelizing operations on small files will help.

2.1.2 Scalability of Numbers of Files and Metadata Performance

Extreme scale archival storage systems will need to manage Terabytes of metadata effectively. The metadata storage requirement will be nearly as large as the entire archive is today. Exascale storage will likely require one to two more decimal orders of magnitude improvement in metadata size and metadata operations per unit of time. Any additional metadata, likely to be needed to accelerate data discovery and other data management features provided below, will add significantly to the overall size and necessary performance enhancements from current projections.

2.1.3 Scalability of Administration, Backup and Recovery

As the amount of metadata scales with the amount of data stored, full backups and restores of metadata should not exceed 2-4 hours to meet operational requirements. From a production perspective, the smaller the backup window the better because often databases are at increased risk of experiencing hardware failures during the backup or restore operation. This is due primarily to the additional strain that backup and restore operations put on the hardware. For the integrity required in archival storage systems, point-in-time recovery[6] is essential.

2.2 Data Management Requirements

An important requirement is that data management and archival storage need tighter integration.

2.2.1 Data Discovery

Workshop reports called for providing methods beyond traditional file and directory naming and location to manage data. Data will be more widely distributed among different resources and centers due to specialization, budget limitations, and scientific need. Exascale system cost will be such that the number of such systems is likely to be limited, thus requiring their sharing between sites, thus increasing data distribution.

[6] Point-in-time recovery is meant to denote the ability of the database to provide recovery up to any past instant in time.

Several reports called for software to help users better manage their own geographically dispersed data.

One report stated that "current mass storage systems that use tape storage prefer certain fairly large size files, a fact that forces scientists to either aggregate smaller files into larger files or partition very large datasets into multiple files. This approach does not scale and is irrelevant to the scientists. Dealing with the volume of data generated by exascale machines will require support for datasets regardless of their size."[7]

Our interpretation of the above requirement is that scientists do not want to be affected by the limitations or constraints of the hardware that archival storage systems utilize. This is an ideal requirement for data management experts to embrace and work in collaboration with archival storage system providers to resolve.

2.2.2 Data Stewardship

In addition, data sets are expected to grow in number and size. Software that simplifies and optimizes data set management (e.g. moving, deleting, creating, copying, sharing) would enable scientists to avoid spending time managing the data and focus their attention more on analyzing the data for scientific results.

A report on scaling software for the Exascale system states, "another fundamental requirement is the automatic allocation, use, and release of storage space. Replicated data cannot be left in storage devices unchecked, or storage systems will fill and become clogged. A new paradigm of attaching a lifetime to replicated datasets, and the automatic management of data whose lifetime expires, will be essential."[8]

When collections contain millions of files, operations based upon POSIX semantics are too inefficient. An 'ls' command no longer makes sense for discovering a relevant file. Embedding descriptive metadata in a file name also becomes impractical.

Optimizing data set management, automating file lifetime management, and enhancing metadata are all areas that would directly benefit from increased research and development. If left unaddressed, these areas are poised to be looming barriers to scientific advancement.

2.3 System Resiliency Requirements

Archival storage systems today need to be capable of handling 100's of devices storing 10's of petabytes of data. In the Extreme Scale era, archival storage systems will need to provide a highly resilient service utilizing thousands to tens of thousands of devices managing up to 10 exabytes of data. The storage

[7] DOE Town Hall summary report from 3 town hall meetings (ANL, ORNL and LBNL) of the E3SG group, Apr-Jun 2007, http://www.er.doe.gov/ascr/ProgramDocuments/Docs/TownHall.pdf

[8] Major Computer Science Challenges at Exascale, Geist and Lucas, February 2009, http://www.exascale.org/mediawiki/images/8/87/ExascaleSWChallenges-Geist_Lucas.pdf

system administrator(s) will need to manage 10's of metadata servers and 1,000's of data mover systems in a simple and efficient manner.[9] The future storage system administrator will evolve to include the skills of an HPC system administrator and database administrator who also requires unique skills, knowledge, and abilities with high-performance storage hardware devices and specialized software.

System lifetime resiliency also involves ongoing management of data migration from older tape technologies to newer technologies, along with typical large-scale namespace management operations. The extreme scale storage system will require a process to perform operations like re-mastering data to new tape technology such that hundreds or thousands of tape volumes may be handled in a single operation. Likewise, the namespace may need to enable millions of files having their ownership changed in a single scalable operation.

Section 3. Storage Hardware

Today tape and disk are the key hardware technologies enabling archival storage. Workshop participants expect these to remain central to archival storage through 2018 - 2022, but flash hardware will be important as well.

3.1 Flash

The primary new technology emerging as a viable HPC long-term storage technology is Flash storage. Some sites and vendors believe that Flash storage will define a new tier in HPC centers between memory and disk. It is affordable for select uses, and its performance characteristics (mean-time between failure, bandwidth, random access/seek time, I/Os per second) are similar to or exceed other mid-range enterprise disk solutions. Due to its cost today, it is still not widely used, but if the trends continue it is likely to find usage in HPC centers for a variety of purposes.

For archival storage systems, Flash could improve metadata performance. It also has potential use as a low latency cache for user data to be used in the hierarchy of storage that most archival storage systems offer today with disk and tape alone.

The emergence of an affordable solid-state storage device, Flash, has put pressure on the rotational media (disk) market. Although, the disk market continues to approximately double capacity each year while reducing costs and it is unclear whether flash will compete with disk in this regard. Recently,

[9] Metadata server refers to the physical server that manages file information or the name space for the archive. Data mover refers to the physical server that presents tape or disk that maintain user data in the archive.

commodity disk capacities and cost became nearly equal with enterprise tape solutions ignoring significant operational costs such as energy, support and maintenance. This has put increased pressure on the removable media (tape) market; see discussion in the next section.

3.2 Tape

The tape market has seen a 20% decrease in media sales each year since 2007.[10] The tape market has two traditional uses: backups and archival storage. Backup applications have been moving away from tape and onto disk solutions due to the competitive cost of disk and the desire to have online access to backup data for quicker restore. The emergence of de-duplication solutions fuels the switch from tape to disk for backups by making them require even less disk. However, archival storage use of tape is increasing each year as the number of archives increases as well as the size of the archives themselves. Our market survey states that archival storage growth will simply replace the decline in backup usage of tape.[11]

Sites participating in this workshop use tape to store ever-increasing amounts of data economically. Tape is critical to providing adequate archival properties over other currently available storage technologies; most notably a 30-year lifetime, media reuse across tape technology improvements, and one or more orders of magnitude reduced bit error rates. Tape drives double in capacity approximately every two years and maintain relatively flat pricing on new tapes that can store twice as much data. Tape capacity projections are physically achievable due to the smaller densities (bits/square inch) that tape provides, but will involve at least two media formulation changes and one or two new tape head designs. The price of media falls over the life of the tape drive, normally five to ten years. One major benefit of tape is media reuse, which is generally possible for at least two generations of tape (e.g. model A and B of a particular tape drive). Media reuse saves the DOE sites millions of dollars in operating costs for archival storage.

Tape library and drive costs although expensive at the onset, are reduced as a single drive can handle an arbitrarily large number of tape cartridges. As long as enough tape drives are available for concurrent user requests and migration from disk there is no fixed number of tape drives necessary. Those drives that have high utilization are also making maximal efficiency of the Fiber Channel or other storage network that attaches them to their hosts, minimizing “dark fiber” from being a waste of resources.

For sites that stay at or below 50% growth per year in total amount of data stored, the tape technology roadmap will allow them to stay within fairly stable operational costs despite the increase in data stored; due to the ability to simply replace the old tape drive

[10] *The Role of Future Magnetic Tape Technology for Digital Archive, Preservation and Sustainability*, a presentation by Barry Schechtman, INSIC, Sep 2008.

[11] *What is wrong with the HSM business model*, a presentation by Henry Newman, Instrumental Inc., July 14, 2009.

with the new one and to copy data from two or more old tapes to one new tape, thus reclaiming tape slots in libraries. Sites exceeding 50% growth cannot keep up with growth with tape drive technology alone and require additional tape libraries, tape drives, and data movers to keep pace with data ingest.

On the downside for tape, transfer rates for smaller files to tape do not see optimal performance. Generally clients cannot deliver small amounts of data at optimal speed and the tape drive being a mechanical device needs to spin up to reach full speed. With small files there just isn't enough data and its often not delivered quick enough for the tape drive. File aggregation to tape is essential if trends in large numbers of small files will continue in the Extreme Scale era. With file aggregation, files are bundled as they are sent to tape to enable the tape drive to reach optimal transfer rates. The application supporting the file aggregation must then be responsible for knowing where the file is located within the aggregate, as well as other metadata related to each file.

3.3 Disk

The price of disk is falling to where commodity disk solutions are now beginning to be competitive with enterprise tape solutions. However, disk densities (bits/square inch) are also approaching the super-paramagnetic limit and must continue to develop novel methods of increasing capacity. This may slow disk capacity increase and affect the price of disks depending on the solution. Figure 7 below shows the densities of various storage technologies over time and specifically that disk is expected to approach physical limitations in density in the near future.

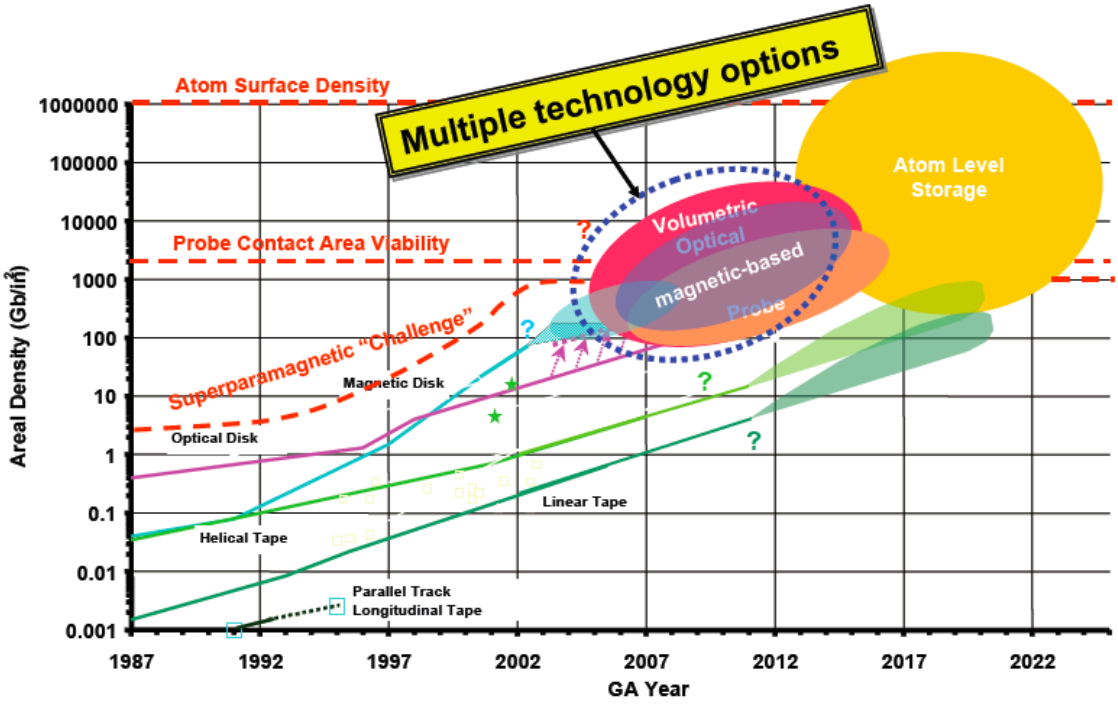


Figure 7. Storage Technology Physical Limitations in Aerial Density[12]

[12] Large Tape Users Group conference, Bob Raymond, Sun Microsystems, Apr 2007.

Disk has poor archival properties and is not expected to replace tape in DOE archival storage systems. Disk also has significant other costs such as power, cooling and management. Multiple hardware failure analysis studies have concluded that the practical life of disk in the HPC Center is three to five years. This would require data migration from old storage media to new on a frequency at least twice as often as it currently occurs with tape archives within DOE. Data would frequently be in motion within the archive, thus not taking advantage of any power saving features of advanced disk systems today. The market has voted that MAID is not a viable technology given that the power saving and other promises was not fully realized. It takes years to migrate Petabytes of data off old technology onto new simply because the performance is constrained to the peak read rates of the older devices. With a three to five year technology refresh cycle, data would nearly constantly be in motion or data would be retained on media with an exceptionally high failure rate.

3.4 Other Storage Hardware

Optical disk, various persistent memory-based devices, holographic storage, and numerous other storage technologies are in some state of research or development. Though the hardware has different characteristics, it still needs software that presents the storage to the user. It should also be pointed out that these technologies have been “emerging” for many years and have yet to make a significant appearance in the market place. If one of these technologies becomes available by 2018 - 2022 and proves cost effective and meets archival storage needs, it would simply be folded into the archival storage system.

3.5 Storage Hardware Analysis

Tape is and will remain the dominant archival storage medium. Workshop participant sites will expect to have 20 to 100 Exabytes of total data in each of the largest archival storage systems by the time the first Extreme Scale system arrives at a workshop participant site. Table 5 provides the impact of the sample Extreme Scale systems on archival storage. For each Extreme Scale system, the table assumes that tape cartridge capacity is up to 128 TB by 2018 - 2022 and provides the total number of tapes required to handle the I/O and expected data retained in the archive at the end of the year. It also provides some cost estimates based on current costs of tape cartridges, drives and libraries. The main point of providing the costs are to clearly state that archival storage costs will likely need to be included in the procurements for certain types of Extreme Scale systems as the archival storage systems at the site will likely not have the space, regular media budget, and numbers of tape drives required to meet these requirements.

	Sustained Minimum Archive BW	Num Tape Drives for Sustained	Num Tapes for I/O	Num Tapes for 50% Retained	Num Tapes for 80% Retained	Tape Cost Estimate ($ Millions)	Tape Drive Cost Estimate ($ Millions)	Tape Library Cost Estimate

	Needed GB/sec	BW						($ Millions)
Extreme Scale A	4.19	6	672	336	538	Negligible	Negligible	Negligible
Extreme Scale B	1.16	2	187	94	150	Negligible	Negligible	Negligible
Extreme Scale C	162.93	203	25,521	12,761	20,417	$3	$5.5	$1.2
Extreme Scale D	349.13	437	56,000	28,000	44,800	$6.5	$11	$2.4

Table 5. Extreme Scale Tape Counts and Cost[13]

Archival storage systems will need to make use of state-of-the-art tape drive, tape library automation, disk, solid-state and other storage hardware. This hardware will need to enable the performance Exascale Systems require and maintain a similar cost savings over file systems to remain relevant. The emergence of solid-state disk as a new storage medium is promising to archival storage in that it offers a boost to metadata performance as well as an option for new user data cache within the archive.

In archiving the 100-8,400 Petabytes of new data generated by one of the proposed Extreme Scale systems, a site will need more than 3,100 tapes (uncompressed data) to store 100 PB of new data on 32 TB cartridges. However, at the same 32 TB capacity a site will require between 175,000 tapes (compressed data) and 250,000 tapes (uncompressed) to store 8,400 PB of data. The lowest estimates of new archive data generated by Extreme Scale systems is feasible given current tape technology plans, however, upper estimates would mean filling 18-25 tape libraries per year. If centers plan to generate and archive Exabytes of data each year in the 2018 - 2022 timeframe then tape drive and tape library research and development needs drastic acceleration.

Section 4. Storage Software

It is the conclusion of the market survey that HPSS is the leading archival storage software system to fulfill Extreme Scale requirements.[14]

With the help of an independent market survey of archival storage software conducted by Henry Newman of Instrumental, workshop participants analyzed the archival storage marketplace to assess viability of the product for satisfying high performance storage requirements; six candidates were identified.

Today, storage software consists of both file systems and archival storage systems. There are numerous efforts underway to integrate file systems with archival storage

[13] Costs provided with 2010 pricing on tape drives and libraries and with 2018 - 2022 projected media costs.

[14] *Large Archives, Requirements and Trends*, by Henry Newman, Instrumental Inc., Jul 2009.

systems. Combined with current market trends, this means that the archival storage system marketplace is contracting and consolidating. New storage software solutions take a minimum of five to eight years and major investments of $50-100 million to arrive to market. There is unlikely to be a new solution other than what is currently in the marketplace coming to market in the 2018 - 2022 timeframe that would immediately satisfy extreme scale storage requirements.

Some of the candidates are integrated solutions presenting both a file system and archive as a single solution. These are still two distinct, although integrated products, a file system and archive, much like workshop participant sites already have. The key difference is the elimination of user directed transfers to the archive. The transfers still occur and still involve client software and separate file system and archive storage devices.

We believe HPSS is the leading competitor for archival storage in HPC going into the Exascale era; thus, we believe that the HPSS infrastructure should be leveraged by the DOE OS to minimize cost and risk in achieving extreme scale archival storage.

Section 5. HPSS Plans

The HPSS architecture can address and handle Extreme Scale archival storage requirements. HPSS has many architectural features supporting scalability[15]. However, to adapt to meet the requirements outlined in this report additional resources are required. This Section provides recommendations on specific additional resource requirements needed to make HPSS viable for the Extreme Scale era.

5.1 Scalability

The original HPSS architecture and development for the past 15 years of its deployment has been focused on scalability of data moving in and out of the storage system and between storage levels within the storage system as well as capacity. For 15 years, HPSS sites have continued to demonstrate that they could transfer data at significantly faster rates each year achieving three orders of magnitude performance improvements from Megabytes/second to Gigabytes/second and capacity improvements from 10s of TBs to 10s of PBs. While all HPSS sites continue to see file sizes increase over time, the actual increase is less dramatic than originally predicted. The HPSS network centric architecture and parallel I/O capabilities, coupled with improvements in tape drive performance and data movement protocols, HPSS server and client software and network capabilities, enable HPSS deployments to lead the industry in providing

[15] Richard W. Watson, "High Performance Storage System Scalability: Architecture, Implementation and Experience," msst, pp.145-159, 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST'05), 2005

exceptional performance to users with large files. A key HPSS asset is that its metadata storage and operations are built on top of the commodity scalable IBM DB2 Database Management System (DBMS) that has many capabilities to allow HPSS to grow its capabilities into the ES era. The HPSS scalability network-centric architecture is uniquely positioned to exploit the extreme scalability features of DB2 DBMS without significant changes to the current software design. We think there is no roadblock to scalability in these areas to continue to exascale archival storage.

HPSS excels in both single client data rate and in aggregate data rate. HPSS cluster architecture, with adequate provisioning, is capable of saturating both the network connecting HPSS to a client and even the storage bandwidth of the client system. Exceptional aggregate data rate is due to the cluster architecture of HPSS in which many HPSS movers operating in parallel can sustain a much higher aggregate data rate than conventional non-cluster architectures. Exceptional single client data rate is due to HPSS' ability to stripe a single file as wide as required to achieve the desired bandwidth to disk or tape. A new capability called Redundant Array of Independent Tapes (RAIT) is also in development with NCSA and IBM and expected to be available before HPSS v8.

A challenge for Extreme Scale HPSS scalability continues to be handling greater numbers of small files. This capability improved by orders of magnitude in HPSS v7 with the introduction of the ability to aggregate large numbers of small files into a very large tape record, enabling small files to be written to tape without the previous performance limitation of synchronizing after each file. Going forward, HPSS will focus on the metadata aspect of small files, enabling a smaller metadata footprint and orders of magnitude more file entities in the metadata.

HPSS 8 is currently prototyped and being designed with a target availability of 2012. The primary focus of HPSS 8 is providing multiple metadata servers to increase the total number of metadata operations per second and number of concurrent file transfers. The goal is to design a solution that scales linearly with the number of metadata servers in the HPSS system. This architecture will be able to handle the scalability requirements of the ES era as it will enable HPSS to scale metadata operations as required.

5.2 Data Management

One example of a data management feature in HPSS is User Defined Attributes (UDAs). This feature enables HPSS users to define valuable metadata that will be indexed to enable fast search (e.g. enabling them to find data meeting custom specifications stored in the UDAs). HPSS contains other data management features, but none are broadly used.

Data management research and development to date has been limited to middleware. There is a push to embed more data management features directly into HPSS and

demonstrable progress has been made integrating HPSS with open source and commercial content management middleware. HPSS requirements needed to support exascale data exploitation are not specific enough to act upon. It is here that work with large-scale scientific data management researchers would be particularly useful both to determine how to best use such capabilities and to determine what others might be required.

5.3 System Resiliency

Currently HPSS uses a custom built Java graphical user interface called Storage System Manager (SSM) to manage the HPSS software at sites. The current interface works well for managing the 100's of devices typical of most HPSS systems today. With HPSS 8 and beyond, it is possible for sites to have 10's of 1000's of devices and this increase in scale will need to be managed in a fundamentally different way from today's systems in order to provide the high level of system resiliency required in this time frame.

We need to rethink the issues around system management and resiliency in developing a new approach to managing and monitoring the HPSS system as it prepares to scale several orders of magnitude in numbers of devices. There aren't any obvious examples or like products to leverage. This is an area that will require significant resources to develop in future versions of HPSS.

5.4 Hardware

The emergence of Flash storage as a viable storage medium for use by HPSS is being investigated currently. The most obvious use for it in the near term is as storage for our metadata; this will occur soon. The second use for Flash in HPSS is as a tier of storage for user data. HPSS is already capable of using a very broad range of different disk and tape technologies. Adopting Flash for user data in HPSS is fairly easy and will require slight design modifications for HPSS to achieve maximum benefit from the new storage devices.

Section 6. Workshop Recommendations

Workshop participants reached a consensus on recommending that DOE OS take a lead role in funding HPSS into the Extreme Scale era commensurate with its leading exascale computing role. The requirements are sufficient to justify a new program within DOE OS. Significant funding is going to be invested in building the computational systems for Extreme-scale computing and it is crucial that the total environment in which these systems will operate have balanced development investments, including archival storage. Establishing stable funding for a decade-long development effort is necessary to enable the HPSS collaboration research and development resources to focus specifically on addressing the archival storage requirements for the first Extreme Scale system.

Specific recommendations on the components of that funding follow:

1. Data Management - Fund people in the data management community to research, develop or gather data management related archival storage requirements for Extreme Scale systems. Fund people in the HPSS collaboration to focus on receiving those data management requirements and working with the data management experts researching, designing and implementing solutions for HPSS.

2. Storage System Resiliency - Provide additional developers to research, design and develop a new approach to storage system management for HPSS capable of managing the new Extreme Scale distributed HPSS system, with its requirement to handle an order of magnitude or more servers and devices. Specific features will need to be developed to enhance overall system resiliency and availability from both a user and administrator perspective.

3. Ongoing HPSS Version 8.X evaluation and development – Provide additional developers to maintain viability of HPSS collaboration for specific DOE OS requirements.

HPSS is a specific HPC solution applicable to meeting both DOE NNSA and OS requirements and those of other commercial and government agencies with Extreme Scale archival storage demands. It is important that the work towards Extreme Scale storage begins now and includes a strong contingent of DOE OS funding.