SOLUTION BRIEF

The IT Data Explosion Is Game Changing for Storage Requirements

Sponsored by: NetApp

Steve Conway Benjamin Woo

April 2012

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

MARKET OVERVIEW: THE BUSINESS NEED

Data-intensive computing has been an integral part of high-performance computing (HPC) and other large datacenter workloads for decades, but recent developments have dramatically raised the stakes for system requirements — including storage resiliency.

The storage systems of today's largest HPC systems often reach capacities of 15–30PB, not counting scratch disk, and feature thousands or tens of thousands of disk drives. Even in more mainstream HPC and enterprise datacenters, storage systems today may include hundreds of drives, with capacities often doubling every two to three years. With this many drives, normal failure rates can mean that a disk is failing somewhere in the system often enough to make MTTF a serious concern at the system level.

Rapid recovery from disk failures — resiliency — has become more crucial. A single HPC job may take hours, days, or even weeks to run to completion. For example, restarting a job after an interruption can devastate productivity. This is especially true because large, complex single HPC jobs are often striped across multiple storage systems. Large enterprise datacenters can also encounter this issue.

Why HPC Systems Need Big Storage

The largest HPC systems today contain 250,000–700,000 processor cores, and supercomputers with more than 1 million cores are on the near horizon. When a significant fraction of those cores are employed to attack daunting problems and workloads, they can produce torrents of data. (Utilization rates for HPC systems typically top 90%, far higher than utilization rates for commercial servers.)

Big Data can also accumulate from the multiple results of iterative problem-solving methods in sectors such as manufacturing (parametric modeling) and financial services (stochastic modeling). Therefore, small and medium-sized enterprises (SMEs) are also encountering Big Data challenges.

Data ingested into supercomputers from scientific instruments and sensor networks can also be massive. The Large Hadron Collider at CERN generates 1 petabyte (PB) of data per second when it's running, and the Square Kilometre Array (SKA) telescope is expected to churn out 1 exabyte (EB) of data per day when it begins operating a few years from now. Only a small portion of this data needs to be captured for distribution to the facilities' worldwide clients, but even that portion is very large.

Consider a few more mainstream HPC examples:

- Weather and climate research is one of the most data-intensive application domains. Some sites apply more than 100 years of historical data in their research. Ensemble models have added challenging new elements to the heterogeneous data mix, such as chemical processes involved in the carbon cycle. Weather forecasting is time critical; no matter how much data is involved, tomorrow's forecast is useless if it takes 48 hours to produce. Finally, the global climate community is exploring the efficacy of augmenting existing approaches with new Big Data methods, such as knowledge discovery algorithms, to produce novel insights.

- Petroleum companies rely on HPC systems to process skyrocketing volumes of seismic data involved in oil and gas exploration. The more powerful the company's HPC resources, the farther it can "see" to spot geological formations such as deposits hidden beneath miles-thick subsalt layers in the Gulf of Mexico. Oil and gas majors today are approaching petascale computing capacity, with massive storage systems to accommodate seismic and reservoir modeling data.

- For competitive reasons, large investment banks and other financial services firms may not appear on the Top500 list of the world's most powerful supercomputers (www.top500.org), but some of them doubtless belong there. They use HPC systems for pricing exotics, optimizing investment portfolios, calculating firmwide (global) risk, and other purposes. Large data volumes are involved in this back-office computational work.

- Massive data volumes and storage requirements often accompany genomics, pharmaceutical, and other life sciences research that leverages HPC.

- In the entertainment industry, for the past 25 years, most of the top-grossing films have used state-of-the-art animation and special effects. Leaders in this field, including DreamWorks and Pixar, rely heavily on HPC processing power and massive storage systems for digital content creation — and large telecommunications companies that distribute the films over the Internet as streaming media also depend heavily on large storage configurations.

Given the criticality of data in these examples, a number of storage-specific issues must be addressed:

- Data loss (While some of the data in these processes can be recovered or recreated, data loss can be very costly. Many of these processes are not only data intensive but also time sensitive.)

- Reduced uptime/productivity (Along similar lines to data intensity, time sensitivity equals productivity and availability. In an increasing number of data-intensive use cases, such as the Large Hadron Collider at CERN, many scientists located around the world are working on the same set of data at the same time. A lack of uptime can severely impact the productivity of these scientists.)

- Reduced ROI for costly HPC compute and storage systems

- Reduced innovation/competitiveness

In HPC applications, where data management and organization is often built into each node within an HPC cluster, high-performance block storage continues to dominate. The NetApp E-Series of solutions continues a very long, rich, and successful history that is focused around performance.

The E-Series has also had an extensive history of implementing the latest and highest-performing interconnects. Given that in most HPC use cases, the focus is on the application itself, NetApp has also focused on simplicity and management in the architecture and design of the E-Series.

Since HPC applications have such high capacity needs, and since storage can often be a significant proportion of the investment, density becomes a very critical consideration. Density, apart from the actual real estate necessary, has an impact on power and cooling. As such, the density consideration has a major bearing on the ongoing operational costs of collecting, storing, protecting, and processing the data.

However, density is not just a physical consideration. The density conversation must be extended to the logical capacity efficiency question and performance density questions. The disk storage system can assist the application environment in accelerating the performance, data placement, and distribution of the data. Creating disk pools can result in consistent, predictable, repeatable, and scalable performance and data protection.

NETAPP E-SERIES WITH SANTRICITY 10.83

NetApp has earned a worldwide reputation as a vendor that addresses the storage and data management requirements of HPC and other large-scale datacenters. The company has built a strong network of major OEMs that serve the HPC and high-end business analytics/business intelligence markets. The modular NetApp E-Series storage with new SANtricity 10.83 was designed from the start to provide improved building blocks for OEMs and extreme scalability for users.

In September 2011, the U.S. Department of Energy (DOE) announced that it had selected NetApp to provide the storage foundation for one of the world's largest HPC systems. Slated for installation later this year at DOE's Lawrence Livermore National Laboratory (LLNL), the "Sequoia" supercomputer will boast 20 petaflops (PF) of peak performance and 1.6 million processor cores.

As part of the Sequoia supercomputer, LLNL will leverage 55PB of NetApp E-Series storage that will provide over 1TBps to the Lustre file system. Dynamic Disk Pools (DDPs) and SANtricity 10.83 are technologies that should appeal to HPC customers such as LLNL.

A detailed technical description of the new NetApp E-Series is beyond the scope of this paper, but some of the product's most salient features are as follows:

- Dynamic Disk Pools are designed to ensure consistent (predictable, uninterrupted) performance and data protection. DDPs dynamically distribute data, spare capacity, and RAID protection information across a pool of drives. The DDP technology is crucial for the consistent performance and resiliency needed for HPC and other Big Data workloads.
- An intelligent algorithm defines which drives should be used for segment placement. This is to help ensure the most efficient and highest-performing use of storage systems.
- Segments are dynamically recreated/redistributed as needed to maintain balanced distribution in case of capacity expansion or drive failure.
- The storage system is designed for significantly faster return to an optimal state following a drive failure. All drives participate in the reconstruction to enable the system to overcome inevitable drive failures while hardly skipping a beat.
- Any stripes experiencing multiple drive failures are given reconstruction priority. This "triage" strategy addresses the highest-risk failures first so that the stripes can be returned to an optimal state before data loss occurs.

MARKET OPPORTUNITIES AND CHALLENGES

Market Opportunities

- **The HPC storage market is growing robustly.** The HPC server market exhibited healthy growth in 2011, with revenue expanding by 8.4% to an all-time record of $10.3 billion. The HPC storage market has been growing 2–3% faster than the HPC server market and attained a record level of more than $3.6 billion in 2011. IDC forecasts that the HPC storage market will exceed $5 billion in 2015. The anticipated ramp-up of newer, data-intensive ("Big Data") methods in the HPC market will increasingly contribute to HPC storage growth. This robust growth presents attractive opportunities for storage vendors that are positioned to take advantage of it.
- **NetApp has earned a strong, positive reputation as an HPC storage vendor.** NetApp is one of the few storage/data management vendors that HPC users closely associate with having a direct focus on the storage needs of the worldwide HPC community. NetApp can leverage this hard-earned reputation to exploit the anticipated growth of the HPC storage market.
- **The HPC storage market is fragmented today.** A recent IDC worldwide study showed that seven vendors (including NetApp) have double-digit shares in the HPC storage market. Some of the largest IT storage vendors have not focused sharply on the HPC market yet, providing a continued window of time for HPC-focused storage vendors to gain market share.

- **Innovative solutions can substantially benefit storage vendors.** In general, HPC-focused storage vendors — more so than the largest storage vendors — have the opportunity to innovate more specifically for HPC users.
- **Storage resiliency is quickly gaining in importance.** NetApp's increased focus on storage resiliency, especially rapid recovery from inevitable disk failures in large storage systems, matches important market trends. These trends include the fast growth in the average size of HPC and other large datacenter storage systems and the related fast growth in the average size of single jobs and their attendant data sets.

Market Challenges

- **The largest, most experienced HPC buyers have become sophisticated about specifying storage systems, but other buyers have not.** Historically, HPC buyers have focused much more on server systems than on storage. But as storage has become more important in recent years, larger HPC buyers have grown savvier about specifying and purchasing storage systems. But other buyers need to become more knowledgeable about storage purchasing, and vendors such as NetApp may need to invest more in this market education process.
- **The largest IT storage vendors are starting to pay more attention to the HPC market.** As the HPC storage market becomes a larger revenue opportunity, it is motivating the largest storage vendors to start sharpening their focus on this market. NetApp and other storage vendors that have successfully focused for some time on HPC are likely to remain more agile than their larger counterparts, but IDC expects storage competition within the HPC market to heat up.

CONCLUSION

Recent developments in the HPC market and data-intensive commercial markets have dramatically raised the stakes for storage system requirements, including resiliency. Data volumes have skyrocketed in the HPC market, fueled by data input from massive sensor networks and scientific instruments (e.g., the Large Hadron Collider) and by data output from supercomputers of record-breaking size and power.

Storage environments with capacities of 15–30PB, and with thousands or tens of thousands of disk drives, are becoming more common. With this many drives, normal failure rates make rapid recovery from disk failures — resiliency — a crucial concern. This is especially true for large, complex HPC jobs that are often striped across multiple storage systems. Inadequate resiliency can lead to serious data loss, reduced uptime/productivity, lower ROI for costly HPC compute and storage systems, and — most important — reduced innovation and competitiveness.

The new, modular NetApp E-Series storage with SANtricity 10.83 was designed from the start to provide improved resiliency, better building blocks for OEMs, and extreme scalability for users. LLNL will leverage 55PB of NetApp E-Series storage as part of the 20PF Sequoia supercomputer slated for deployment later this year.

IDC believes that the new NetApp E-Series, featuring NetApp's DDP and SANtricity 10.83 technologies, should appeal to HPC customers in data-intensive domains, such as government, weather and climate research, energy, life sciences, and financial services. The NetApp E-Series may also attract tier 1 players in the digital content creation and distribution industry.

In sum, NetApp is well positioned to exploit the robust growth expected in the HPC storage and data management market, which IDC projects will grow to more than $5 billion in 2015, as well as analogous opportunities in data-intensive commercial markets. To do this effectively, NetApp will need to advance its standing before other storage companies sharpen their focus on HPC and related market opportunities. As a company that has already demonstrated its agility and commitment to HPC, NetApp seems ready to meet this challenge.

Copyright Notice