THE COMMERCIAL HPC STORAGE CHECKLIST

by George Crump

CHAPTER 1

High-Performance Computing (HPC) is no longer limited to the halls of academia and large government labs. Commercial HPC use cases are on the rise. Many organizations now have a large pool of unstructured data that requires analysis. The challenge for most commercial organizations is the HPC project starts as a pilot project that is at some point "thrown over the wall" to IT to store and manage. IT initially tries to leverage its knowledge in traditional enterprise storage systems but quickly learns that HPC is a different animal with different requirements than mission-critical databases or virtual machines.

The next step for the commercial organization is to investigate the various HPC specific storage solutions. IT often finds these solutions, built from a combination of open source file systems and commodity hardware, too time-consuming to assemble and operate.

IT needs an integrated, enterprise-ready HPC specific storage solution. One that meets the HPC IO requirements while at the same time meeting the commercial organization's desire for rapid time to value. The problem is that traditional enterprise vendors try to extend their solution, so it looks more HPC like and traditional HPC vendors try to bundle their solutions to make them look more integrated, leading to confusion among IT personnel.

The goal of this eBook (chapter series) is to provide IT professionals with a checklist that they can use to make sure that their HPC storage selection meets all the requirements of the commercial HPC use case.

Item 1 – Direct Access

The typical answer to most HPC storage problems is a scale-out file system or scale-out NAS. A scale-out NAS consists of multiple servers (nodes), each of which contributes their internal storage to the cluster, creating a centralized pool of storage. Traditional scale-out file systems however, "shard" or stripe data across some number of nodes in a cluster. Scale-out systems claim to eliminate both capacity and performance concerns because each additional node adds to the cluster's potential capacity and performance. The claims of scaling capacity are for the most part accurate but the claims of scalable IO are not.

The typical scale-out cluster creates an IO problem for itself. When a client or application requests a file, each node that has a shard of that file responds with its shard. The requesting client or application though has no idea how to reassemble these shards into the original file. Therefore, most scale-out NAS / file systems have a control node or gateway that does the reassembling of the file from the shards before passing it back to the requesting client or application.

In a read-heavy environment with mixed file sizes, a common attribute of Commercial HPC, these control

nodes or gateways can become overwhelmed with IO requests. Most scale-out file systems have either no, or a limited ability to scale the number of control nodes. As a result, the IO bottleneck becomes worse as the number of nodes increases because the control node has to reassemble data from a higher number of data nodes.

The solution is to provide clients with the intelligence to reassemble the shards themselves and not have to go through a control node. Panasas, for example, provides DirectFlow®, that enables Linux and macOS clients to access the data nodes directly, reducing the control node's role to metadata management and the background operating system tasks. This directs the "heavy lifting" of data to the client.

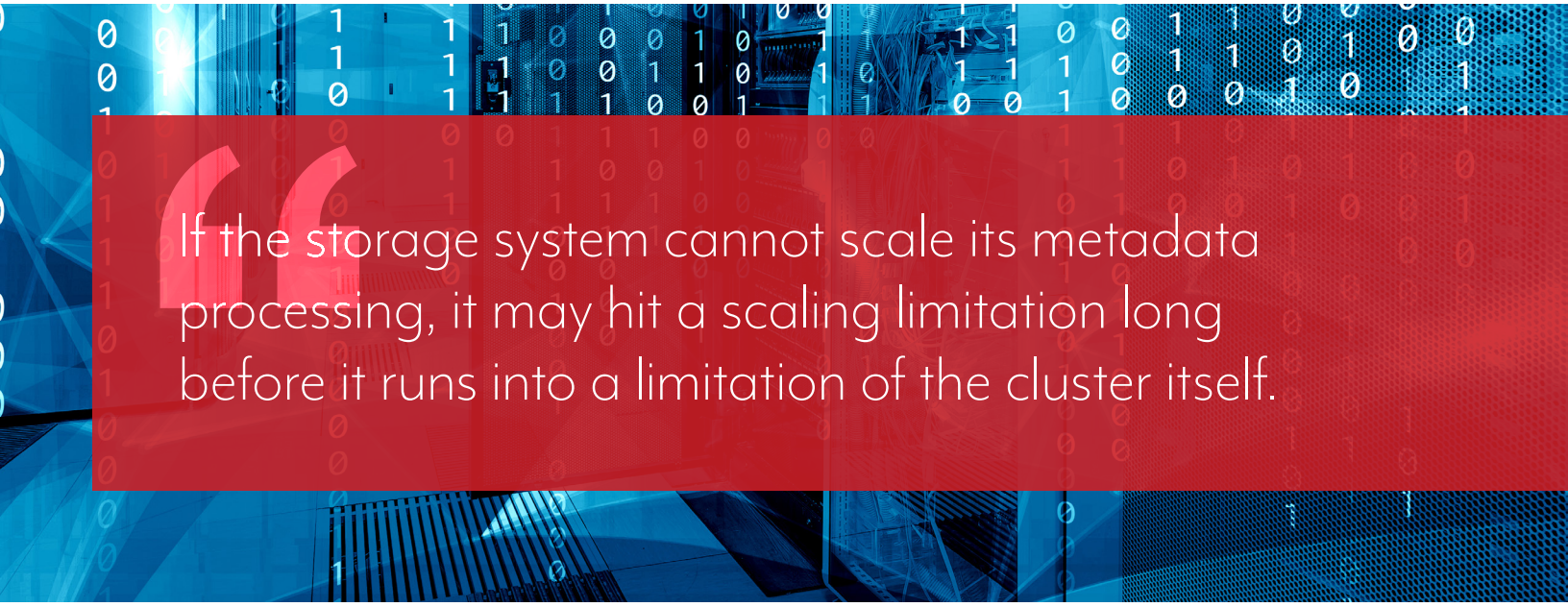
The advantage of direct access to data from the client is two-fold. First, the client or application requesting the data experiences better IO performance because it is directly accessing the data itself instead of going through the single gateway. It is truly parallel. Multiple clients can access data simultaneously, and the nodes can respond to each client in parallel. Second, the overall cluster is more scalable since the control nodes aren't bogged down with IO requests. Their processing can be more dedicated to metadata and cluster management functions.

The downside to direct access is loading client-side software so the client can perform the shard reassembly function. Since most HPC functions are either Linux based or macOS based in the case of media and entertainment, the storage solution vendor can provide support for these two platforms and have majority of the market covered. The Commercial HPC provider, however, should provide a gateway functionality delivering support for more traditional protocols like SMB and NFS. The gateway is particularly useful when reading in data from devices (IoT and Medical devices for example) that can't have a client installed or aren't running Linux or macOS.

CHAPTER 2

Item 2 – Start Small, Scale Large

Scaling is a capability that almost every vendor claims but few execute well. Most vendors claim to provide scalability via a scale-out storage system, but scale-out by itself is not enough to address all the scaling demands of a Commercial HPC environment. Most Commercial HPC environments are new data sets to the organization and as such can start with relatively small capacities and modest performance expectations. As the environment moves into production, it can scale quickly, and demands for both more capacity and higher performance come suddenly. New nodes and even upgrades to new technology need to happen seamlessly without disrupting ongoing HPC operations.



> If the storage system cannot scale its metadata processing, it may hit a scaling limitation long before it runs into a limitation of the cluster itself.

Start Small

A clustered file system is a standard method for implementing a scale-out storage strategy. The problem is that most turnkey or integrated solutions start too large for the initial commercial HPC deployment and it may take years for the organization to utilize the initial capacity efficiently. That reality leads many commercial organizations to attempt to leverage existing storage solutions in the data center. In an attempt to avoid overbuying computing or storage resources, IT often initially implements Commercial HPC projects on legacy storage systems or homegrown HPC deployments. However, neither of these choices is optimal for the Commercial HPC project, and at some point, requires migration to a purpose-built HPC solution.

Ideally, the Commercial HPC customer should find a solution that can start as small as they need but can also scale as big as is required. Getting the right HPC solution not only enables the organization to start small, it also enables them to move into production with the same technology as they develop and deploy. Also eliminated is the migration to a new system as is the need to retrain personnel.

Scale Large

As the HPC project moves into full-scale production, the organization then faces the opposite problem, making sure the system can scale large enough to continue to meet the capacity demands of the project. Scaling out requires meeting several challenges. First, the system has to integrate new nodes into the cluster successfully, since additional nodes provide the needed capacity and performance. However, adding another node is not always as straightforward as it should be. Many systems require adding the node manually as well as manually rebalancing data from other nodes to the new node.

The Commercial HPC storage customer should look for an HPC storage system that can grow with them as their needs evolve. It should start small during the initial phases of development but scale large as the environment moves into production. The system should make the process of adding nodes as simple as possible; automatically finding available nodes, adding them to the cluster automatically and automatically rebalancing cluster data without impacting storage performance.

Eventually the nodes a customer adds to the storage cluster change. A properly designed, scale-out HPC storage cluster should meet the organization's HPC requirements for years. It is vital then that the new system can incorporate new technology from within the existing nodes.

Take Metadata Seriously

Another scaling problem is specific to HPC, managing metadata as the system scales. HPC applications typically access massive amounts of unstructured data sequentially to perform analysis on that data. The organization continues to add more data to the HPC environment to improve accuracy. The growth in the data set consists of thousands, potentially millions, of individual files. These files can range in size from very small to very large.

The HPC storage system not only has to handle a variety of file sizes, it also has to scale metadata processing. If the storage system cannot scale its metadata processing, it may hit a scaling limitation long before it runs into a limitation of the cluster itself. When the storage cluster has a metadata processing problem, the addition of nodes becomes a case of diminishing returns. The computing resources allocated to managing metadata, do not scale as the cluster itself scales.

All the nodes in a cluster should process metadata instead of just a couple of control nodes. Spreading the workload across all the available cluster nodes allows the system to run at high rates of utilization, which improves performance and dramatically lowers costs. The cluster should also store metadata separately from the data itself, which again should improve metadata scaling and performance as well as increasing metadata durability.

CHAPTER 3

Item 3 – Protection at Scale

Most storage systems provide several layers of protection. At a minimum, there is protection from media failure, typically delivered by some form of parity based protection scheme. In enterprise systems, it is also common to have built-in redundancy in both the hardware and the software. Commercial HPC has these same requirements and expectations but how the traditional HPC storage environment delivers those, causes issues for the enterprise use case.

Commercial HPC, as compared to traditional enterprise workloads, requires much more capacity to store its data. That data is usually unstructured, but file sizes can range from the very small to the very large. Commercial HPC also requires more performance, but its performance tends to be bandwidth driven, consequently the ability to stream data is critical. The combination of high capacity and bandwidth focused high performance means that Commercial HPC storage systems need to implement data protection differently from traditional enterprise storage systems or even traditional HPC storage systems.

Data protection is also more critical on Commercial HPC storage systems. The data that these systems store may be the last known good copy of data or it may be storing a copy of data that is impossible to recreate. Additionally, because of the time required by some of the workloads, some jobs can take hours, days or even weeks to complete. Commercial HPC cannot afford an interruption or a performance loss because a media failure degrades performance.

Requirement #1 – Triple-Parity Protection

The first requirement of a Commercial HPC storage system is triple-parity data protection. Since the capacity requirements of Commercial HPC can dwarf the rest of the enterprise, the protection scheme used must be efficient. If the environment stores 500TBs of data, forcing the organization to store 1PB or even 1.5PB to maintain a protected state may not only exceed the budget it can cause shortages in available floor space.

Many HPC storage solutions provide only replication for data protection. Replication protects against media failure within a node by creating two or three additional copies of data on other nodes in the storage cluster. The problem is a replication only model forces the organization to store two or three full additional copies of data. While replication does maintain performance during a failure, the level of exposure to an additional failure is enormous. Most enterprise storage systems support a single or dual parity protection scheme. While parity does not have the capacity waste of a replicated system, it can hurt storage performance if the design of the storage system cannot maintain performance during a failure/rebuild process.

A Commercial HPC storage system needs to provide a parity-based protection scheme, so they do not waste

> Commercial HPC makes products better, makes customers happier and improves organizational processes.

capacity nor unnecessarily waste data center floor space. Because restarting of workloads is so time-consuming it also needs to have multiple layers of redundancy so that oneor two drive failures don't stop an HPC process from executing.

Additionally, most HPC storage systems don't have distributed drive sparing. If a drive fails, an administrator needs to get involved and replace the failed drive. Given the number of drives common in a Commercial HPC environment, just finding a failed drive is challenging. Even after identifying the failed media, it likely takes core IT quite some time to replace it, which means increasing the level of exposure for the Commercial HPC workload.

Commercial HPC Systems need to have distributed spares that are globally available to the system. With a distributed spare, when a drive fails a spare is automatically assigned to replace it. Distributed spares mean that rebuild can occur immediately upon failure identification, instead of waiting for the administrator to replace it and initiate the rebuild process.

Every data center though, has to deal with a media failure but Commercial HPC may be in a failed state

more often, just because of the number of drives it may deploy to meet the capacity demands. During a failure, the storage system is one step closer to complete data loss, so a rapid rebuild is critical.

Many enterprise storage systems, which use parity, have lengthy rebuilds especially as hard disk drive sizes continue to increase. In most cases, a rebuild requires a complete reading of the drive, so the more massive the drive, the longer the rebuild. The Commercial HPC system has to ensure that its software is efficient enough to read only the portion of the drive that contains data. It also needs to ensure it leverages available processing to perform the rebuild quickly.

Maintaining performance during the failed state is also critical. The Commercial HPC team likely has jobs running all the time, and it can't afford to restart them but it also can't have their processing slow down during a failure. Any parity-based system is going to impact performance during both a parity calculation and during a rebuild process. The Commercial HPC system has to take special precautions to make sure that the workloads counting on the system don't see a loss in performance.

Finally, more can fail on a storage system than just the media. The Commercial HPC system needs to maintain availability 100% of the time. Again, the length of job time means thatrestarting a job from the beginning can be very time-consuming. The systems need to have complete redundancy in both hardware and software. While many systems have redundant hardware, they don't offer redundant software. Software redundancy allows a second copy of the file system to take over if the primary software fails.

CHAPTER 4

Item 4 – Transparent Operation

Commercial HPC makes products better, makes customers happier and improves organizational processes. The infrastructures that support commercial HPC need to accomplish those objectives transparently. The commercial HPC storage system, being one component of those systems, needs to remove the mundane tasks often associated with traditional HPC storage systems.

Transparent Data Placement

Traditional storage systems typically manage data placement to lower cost, but commercial HPC systems place data to improve access to it. The key file variable with which HPC systems concern themselves is file or data size, not access dates. Commercial HPC data can reactivate at almost any time so HPC needs to store data for easy accessibility regardless of access dates.

Commercial HPC file systems deal with content (files, data or objects) of different sizes. The commercial HPC system needs to place that content logically, based on their size. A key concern is how the HPC storage system deals with the metadata of that content as well as the accesses to that metadata. Certain commercial HPC applications can generate a high number of metadata accesses to generate results. Storing metadata on flash media allows rapid access to the corresponding content which speeds time to results.

Organizations' commercial HPC sequential and metadata intensive workloads are routinely run on the same commercial HPC storage. The streaming performance of a hard disk system, tuned for sequential access, usually provides adequate performance for those workloads. Flash is often unnecessary and of course expensive. For these sequential file workloads, equipping the HPC storage system with hard disks enables the organization to keep costs down while meeting performance expectations.

The IT team's time is too limited to make sure that the appropriate data types are stored on the media to which they are best suited. Additionally, how does IT deal with a workload that requires both streaming and metadata intensive operations? The manual monitoring required cripples IT's ability to properly manage the process.

Automatic Protocol Management

Many commercial HPC workloads use a parallel file protocol so that they can directly interact with the storage system components storing the data that the workloads needs. It is well worth it for the organization to fine tune the applications to take advantage of a parallel protocol. The problem is the systems feeding the HPC storage system or the applications that need only occasional access may not support the parallel protocol, especially in the commercial market.

The commercial HPC storage system needs to interact seamlessly with both parallel and traditional file protocols like SMB and NFS without having to setup separate volumes for each. Seamless protocol access enables the commercial HPC system to ingest data from IoT devices or log files from legacy systems via NFS or SMB while enabling simultaneous analysis from modern applications via the parallel protocol.

Automatic Operation

The commercial HPC system needs to appear, to IT, like any other storage system. Turnkey implementation, especially in the commercial space, means that IT doesn't have to separately source software and hardware, assembling its own solution from scratch. The special HPC capabilities of the storage system like those we've discussed in the previous chapters need to operate transparently so the administrator doesn't require special training.

ABOUT US

Storage Switzerland, LLC

Storage Switzerland is an analyst firm focused on the storage, virtualization and cloud marketplaces. Our goal is to educate IT Professionals on the various technologies and techniques available to help their applications scale further, perform better and be better protected. The results of this research can be found in the articles, videos, webinars, product analysis and case studies on our website storageswiss.com

PANASAS®

Panasas is the performance scale-out NAS leader for unstructured data, driving industry and research innovation by accelerating workflows and simplifying data management. Panasas ActiveStor appliances leverage the patented PanFS storage operating system and DirectFlow protocol to deliver performance and reliability at scale from an appliance that is as easy to manage as it is fast to deploy. Panasas storage is optimized for the most demanding workloads in life sciences, manufacturing, media and entertainment, energy, government as well as education environments, and has been deployed in more than 50 countries worldwide. For more information, visit www.panasas.com.

George Crump is President and Founder of Storage Switzerland. With over 25 years of experience designing storage solutions for data centers across the US, he has seen the birth of such technologies as RAID, NAS and SAN. Prior to founding Storage Switzerland he was CTO at one the nation's largest storage integrators where he was in charge of technology testing, integration and product selection.