Rev Up Your Performance: Three Storage Secrets Every Administrator Should Know

Addison Snell

November 2010

White paper

EXECUTIVE SUMMARY

Storage bandwidth is increasingly a gaiting factor for overall system performance, as big data problems – both in size and in number of files – continue to proliferate. We expect this trend to continue, driven by a persistent demand for higher fidelity and greater detail in scientific and engineering models, to the extent that spending on storage hardware and software will continue to outstrip the overall growth of the high performance computing (HPC) market despite falling prices for storage devices.

Parallel file systems have emerged as a critical technology for achieving I/O performance at scale, and Intersect360 Research predicts their adoption will continue. GPFS and Lustre are currently the two most commonly used, but neither has a dominant share, and various additional competitors and particular market dynamics make it difficult to predict which of these (if any) might become established as a dominant standard.

There are unfortunately several common pitfalls to architecting storage, and even administrators who think they had specified performance and capacity properly may find themselves disappointed by their actual output on real workloads at scale. But by considering the idiosyncrasies of the specific file system and following a few fundamental rules of thumb, there are simple steps that can be taken to maximize I/O performance.

- Communicating needs: Understand how to specify disk capacities without misunderstanding.
- Setting expectations: Run better benchmarks at predict reliability more accurately.
- Power-of-two architectures: Configure systems to take advantage of natural computer mathematics.

New technologies are important but only part of the solution. LSI offers these and other suggestions when architecting its storage solutions for HPC. With these principles in your arsenal, you will be better prepared to architect a storage system that delivers to the specifications you were expecting.

MARKET DYNAMICS

Are you achieving maximum performance? As the supercomputing industry drives into the petascale era, there is an increasing (and appropriate) awareness of issues related to the achievable, sustainable performance of such machines on real-world applications. Across multiple industries, one of the most discussed challenges is the management of big data.

There are multiple, prevalent, compound reasons for this justifiable emphasis on data. As the HPC industry resumes its growth – Intersect360 Research projects the Traditional HPC market will increase from $14.9 billion in product and services spending in 2009 to $21.8 billion in 2014 – an increasing percentage of spending will be devoted to storage hardware and software. (Software and storage are the two highest growth rate spending categories in the forecast period.)[1]

These increases in data spending come despite – or perhaps even because of – the plummeting costs of storage capacity and bandwidth. Multiple industries are producing and devouring data at an exponentially increasing rate, driven partly by the ability to store it, but also in many cases by paradigmatic shifts to data-centric workflows or new generations of technologies. For example, the medical community is well into a broad-scale shift from films to digital images for a variety of scans. The oil and gas industry has begun to capture "wide azimuth" data that represents subsurface formations at significantly higher fidelity. And the entertainment industry has seen the adoption of multiple data-heavy technologies, including digital movies, HDTV, and 3D movies and television.

The scientific research community has a similar rise in big data demands, compounded by the challenge of supporting multiple user types in a wide range of application areas. Large-scale research labs are typically shared resource facilities that must support scientists from many disciplines. This places an increased burden on the resource center to supply an expanding storage pool to support diverse application needs.

The Importance of Data Management in HPC

For users across government, industry, and academia, the most important decision criterion in evaluating and selecting a new HPC system is consistently price/performance – or more accurately, performance/price, maximizing the amount of work that can be accomplished within a specific budget. And even this inverted definition is insufficient unless all aspects of both numerator and denominator are considered. "Performance" should be considered in terms of the true utility of the system, the extent to which engineering achievement or scientific discovery is enabled or accelerated, and "price" should take into account not only the acquisition cost of a system, but also ongoing operational costs, including both facilities and administration.

One frequently underappreciated aspect of the full price/performance evaluation is how the I/O and storage architecture should be architected in order to maximize true utility. Many applications are more I/O-bound than compute-bound; that is, performance is gaited less by the processors' ability to process data quickly than by the storage system's continual ability to feed data to the processors in a timely fashion.

As systems scale, one of the first critical pieces to add for the management of increasing data needs is a parallel file system. In any HPC system, the file system is the interface that connects a user to the data. File names provide pointers to the information in specified locations in storage, and metadata tracks the information about a file, such as its change history. In a traditional, distributed file system (such as NFS, the ubiquitous UNIX and Linux file system), access to data goes through fixed I/O servers that can become choke points as

[1] Intersect360 Research HPC market advisory service, "Traditional HPC Total Market Model and Forecast: 2009 to 2014," July 2010.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [408] 306-2926
www.Intersect360.com | info@Intersect360.com

the computational and data infrastructures scale. Parallel file systems scale by eliminating the central I/O server and allowing all nodes equal access to data.

HPC File System Market Dynamics

There are many file systems, including several true parallel file systems, available in the HPC market today. The two most prevalent parallel file systems currently are GPFS and Lustre.[2] The two have similar feature sets but architectural differences, as well as dissimilar organizational outlooks. While GPFS is straightforwardly produced and supported by IBM, Lustre is in an odd limbo between open source and Oracle. (Although Lustre is available under open-source licensing, Oracle owns the design and support of Lustre through its acquisition of Sun Microsystems, which had previously picked up Lustre by buying Cluster File Systems, Inc., the original developer. Oracle has stated that future versions of Lustre will be supported on Oracle hardware only – "Oracle hardware" itself being a relatively new phrase – prompting community initiatives such as the OpenSFS alliance to continue to develop HPC features for Lustre.)

In addition to GPFS and Lustre, other companies provide scalable storage and file system solutions for HPC. Panasas, for example, has a strong HPC position based on its PanFS file system. Companies like BlueArc and Isilon also have offerings.

With so many options for HPC file systems, there is the potential for consolidation. Lustre and GPFS are potential platforms to attract more users, along with pending new development: pNFS. Parallel NFS (pNFS) is a parallel file system under development by a multi-vendor consortium to provide parallel extensions to NFS.

All these file systems options have their own considerations for performance optimization. For scalable HPC performance, system administrators should consider how their storage would be best architected for their particular applications.

THREE SECRETS FOR BOOSTING YOUR STORAGE SATISFACTION

To get the most out of storage performance at scale, there are basic architectural rules of thumb to consider. Paying attention to these guidelines can keep your storage subsystem from becoming an I/O bottleneck that impedes your full computational performance potential. To achieve maximum I/O performance, here are three storage architecture secrets that every administrator should know.

1. Communicate! Be Clear About Your Needs

Are you asking for everything you're expecting? If not, don't be surprised if you get disappointed later. Most storage RFPs specify only two figures, capacity and bandwidth, and even those might be done inaccurately. It takes more than two general figures to build a fulfilling storage architecture, and the more you do to communicate your needs up front, the more satisfied you'll be in the long run.

When specifying capacity and bandwidth, how much do you mean? One basic term, the megabyte, means different things to different people. Is it 1,000,000 bytes? 1,024,000? 1,048,576? The confusion stems from the slight difference between rigid powers of 10, which humans like to count in, and the computer-friendly powers of two that approximate them.

[2] Intersect360 Research HPC Site Census Data, 2010.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [408] 306-2926
www.Intersect360.com | info@Intersect360.com

Since 1998, the U.S. National Institute of Standards and Technology (NIST) has attempted to settle the argument by defining two sets of prefixes, the original metric (kilo, mega, giga, tera, etc.) to signify powers of ten and related variations (kibi, mebi, gibi, tebi, etc.) to reflect their binary counterparts. [See Table.] However, the *bi- prefixes have not made it into the everyday lexicon, and regardless of which word you use to describe your needs, you should make an effort to be clear with your prospective storage partner.

Table: Proposed Binary vs. Metric Prefixes for Computer Capacities

Binary				Metric			
Prefix	Abbr.	Exponent	Value	Prefix	Abbr.	Exponent	Value
Kibi	Ki	(2^10)^1	1,024	Kilo	K	(10^3)^1	1,000
Mebi	Mi	(2^10)^2	1,048,576	Mega	M	(10^3)^2	1,000,000
Gibi	Gi	(2^10)^3	1,073,741,824	Giga	G	(10^3)^3	1,000,000,000
Tebi	Ti	(2^10)^4	1,099,511,627,776	Tera	T	(10^3)^4	1,000,000,000,000
Pebi	Pi	(2^10)^5	1,125,899,906,842,624	Peta	P	(10^3)^5	1,000,000,000,000,000
Exbi	Ei	(2^10)^6	1,152,921,504,606,846,976	Exa	E	(10^3)^6	1,000,000,000,000,000,000

Source: NIST

Note that the greater the prefix value, the more difference there is between the values of the binary and metric expansions. While there is only a 2.4% difference between a kilobyte and a kibibyte, if you are requesting a petabyte of storage, there is potentially a 12.6% difference in expected vs. delivered capacity lurking as a misunderstanding. The exact capacity becomes an issue in designing the eventual architecture to run efficiently.

Even with the units settled, there is often confusion as to what figure constitutes “capacity.” The total number of bytes in the storage system is one figure, but RAID configurations use some of this for parity or redundancy, reducing the amount of usable storage. Many file systems also carry similar overhead. Make sure you agree with your storage vendor as to the notion of available capacity and how it will be measured and checked.

Exact or official definitions aside, it is worthwhile to make sure you are being understood. “Pebibytes” are not likely to be discussed by any other than the most academic diehards, and if there is an argument over how much “peta” means, going to the NIST web site to settle an argument is unlikely to avoid bad feelings over the disagreement. Furthermore, if you are expecting one terabyte of usable storage but find that 20% of it is unavailable, there is likely to be a confrontation. Communicate well, and you can simply avoid these arguments to begin with, and you’ll find your storage partner is better able to satisfy your needs.

2. Have Realistic Expectations: Know What You’re Getting Into!

Sometimes your storage system doesn’t live up to the image of the one you thought you were getting. Even if you run a benchmark to assess performance prior to purchase, you may eventually find that the benchmark was not a true representation of reality. Like their computational counterparts, storage systems can be tuned to perform well on particular benchmarks in particular configurations, but once you get them outside their sheltered test environments, they may falter on everyday workloads.

A better approach, when you can, is to run a simulated workload on a representative subset of your target infrastructure, without tuning it to a particular application – unless you really do tend to run a single application environment. Otherwise you may find that while you and the vendor were able to tune the system to run the benchmark fast, your system can’t sizzle quite the same once you get it home.

Another area that can occasionally lead to dissatisfaction is reliability, and again, expectations play a part. Component mean time between failure (MTBF) numbers are only part of a reliability equation. The more

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [408] 306-2926
www.Intersect360.com | info@Intersect360.com

components there are, the more failures there will be. A complete MTBF estimate should include all components at scale, including an estimate of software failures.

Advertised reliability features should be discussed in detail. What specifically will happen if the power fails under 100% workload? You may need to build the cost of an uninterrupted power supply (UPS) into your storage estimate. If the vendor advertises hot-swappable components, test whether you still have visibility to mirrored or striped data if a tray is removed. Knowing your partner before you are committed can help save headaches down the road.

3. The Power of Two

"There are 10 kinds of people in the world: those who understand binary and those who don't." – Old math joke.

As referenced above in the discussion on prefixes, binary is the language of computers, zeroes and ones referring to switches, toggles, or bits that fundamentally can be on or off. Computers are built around powers of two, with round binary numbers like 32, 128, and 512 showing up consistently.[3] You don't tend to see 147MB memory cards or programs that define nine-bit integers. Powers of two rule.

Storage infrastructures are rife with 2^n. For example, there are natural boundaries in most I/O stacks and file systems stacks are aligned to powers of two, such as 16KB or 1,024KB boundaries. Furthermore, it is important to select a disk layout that is aligned to powers of two. Arranging a RAID array as 4+1 (four active disks, one parity disk), 8+1 (eight active, one parity), or 8+2 (eight active, two parity) will help streamline I/O requests.

Data segments also tend to be in powers of two, and efficient I/O requests will be the size that is the product of the segment size times the number of active disks in the base RAID layout. For example, an efficient storage system might have 256KB segments in a 4+1 array. The I/O request length would be 256KB times four, or 1,024KB (usually called 1MB, but we discuss the importance of understanding prefixes above – technically this is 1MiB). With this configuration, I/O requests will naturally align with segment lengths, and you won't wind up with wasted I/O time on each read.

Contrast this with an inefficient architecture in which the same 256KB segments are in a 5+1 array. Now the I/O request is 1,280KB, which will inefficiently miss segment boundaries and result in extra read time. In a mirrored configuration the effect is doubled.

To maximize the performance of a parallel file system, leverage these powers of two wherever possible. This can include:

- The logical unit numbers (LUNs) presented to a Fibre Channel or InfiniBand path
- The LUNs presented to service nodes
- The number of I/O paths to storage
- The number of service nodes
- The total number of LUNs in the configured file system

When you deploy the power of two, you unlock your potential to achieve maximum performance with your storage system. In this case there are 10 types of administrators: those with optimal performance, and those without. Don't be a Zero. With a "bit" of simple planning, you can be the One.

[3] This author once interrupted a marketing briefing upon seeing the figure 65,335 on a PowerPoint slide. My comment – "That's seems wrong. Shouldn't it be 65,535? Two to the sixteenth minus one?" – was met by blank stares from the PR representatives.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [408] 306-2926
www.Intersect360.com | info@Intersect360.com

INTERSECT360 RESEARCH ANALYSIS

Big data is a big topic in HPC, and proposed solutions for data management bottlenecks are a common topic of discussion across the industry. For all the emphasis placed on technological improvements – and parallel file systems are indeed an important industry trend – there are a few basic principles that any storage administrator can follow in order to get more bandwidth and performance out of the storage systems they are already purchasing and deploying.

Parallel file systems will continue to be a critical area to watch in 2011 and beyond. With GPFS, Lustre, pNFS, and other scalable solutions available to HPC users, it is too soon to know if any of these will become a dominant offering. Administrators should consider the particularities of their chosen file system in establishing a storage architecture.

LSI Corporation, one of the leading storage providers for HPC applications, suggests the principles above in optimizing storage configurations for stronger performance. LSI engineers find that while providing high-performance storage solutions is important, it is equally important to offer advice on how those solutions can be streamlined to provide optimal performance for end users' particular environments.

For the uninitiated, storage architecture can be a complicated and confusing topic. It is easy to think that if you have specified a total capacity and aggregate bandwidth figure – and even performed a benchmark! – that you have completely covered your bases with respect to storage optimization. By following the fundamental guidelines presented in this paper, you will have a better chance of finding the storage performance you're looking for.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [408] 306-2926
www.Intersect360.com | info@Intersect360.com