



**Storage Switzerland, LLC**

## **The Case for Scale-Out NAS**

*Prepared by: George Crump, Senior Analyst*

*Prepared on: 7/30/2009*

*<http://www.storage-switzerland.com>*

*Copyright © 2009 Storage Switzerland, Inc. - All rights reserved*

There is a dark cloud looming in storage. Over the last decade, conventional storage platforms have been able to keep up with the demand for ever higher capacity systems at a lower cost per GB, however, the real specter on the horizon is severe and inevitable performance degradation. This degradation of performance is critical because most organizations rely on a scalable facility for servicing I/O to rapidly deliver information to aid in revenue generation. Indeed, resolving the storage I/O performance bottleneck becomes even more critical in a soft economy, when profits are the most elusive.

### **What's Causing the Storage I/O Bottleneck?**

Multi-tenant workloads are at the heart of the problem. Multi-tenant workloads are also known as concurrent/aggregate workloads in which data is shared between multiple users or applications and accessed concurrently by multiple users of the same shared storage resource. These shared workloads and resources are no longer relegated to a few isolated companies whose performance demands are on the fringe of mainstream data center environments. In fact, any organization that is deploying a server virtualization project has by definition a multi-tenant workload demand.

There are three basic elements of performance in a data center; the processing power harnessed by servers, the network harnessed by switches and routers, and the storage which consists of the disks harnessed by SAN and NAS controllers also referred to as "NAS Heads". Each of these elements are under constant strain to keep up with the digital demands of their users. Servers and networks have kept pace through added performance and intelligently utilizing that performance, but storage has not kept pace and has become the bottleneck of the enterprise. Now this storage bottleneck has moved beyond being an IT problem and has created a perilous situation for the organization as a whole.

Of the two elements that have kept pace with the growing digital demand, compute capability has kept pace via increased performance and increased core density, as well as increased intelligence through server virtualization and scale-out clustering or grid infrastructures. Networks similarly have kept pace with increased bandwidth capacity and intelligent use of that capacity through QoS, prioritization, and efficient use of wide area connectivity.

Meanwhile, storage performance has not kept pace. Instead it has remained frozen in the same architectural design for at least a decade; a high performance SAN or NAS controller pair that drives an increasing number of disks. While increasing the number of drives can improve performance, there is a limit to the number of drives these controller pairs can support as well as a limit to the amount of inbound traffic they can sustain. This controller (SAN) or head (NAS) is now the primary bottleneck, limiting improved storage performance.

### **Storage I/O vs. Multi-tenant Workloads**

To compound this problem, the workload is now changing. Workloads are now multi-tenant with multiple shared servers and networks trying to access storage in this out-dated model. Prior to multi-tenant workloads, a single application coming from a single server could only create a limited number of requests. Multi-tenant workloads, running either through multiple virtual machines on a single physical server or through a single application scaled across many physical servers in a cluster or grid, can now generate hundreds if not thousands of requests for storage I/O.

The impact is that these requests saturate the storage controller (or head) and the applications or servers have to wait for it to catch up, which in turn delays processing, eventually costing organizations money and limiting productivity.

A multi-tenant workload is one that typically has multiple owners or users at any given point in time. The presence of these multi-tenant workloads is increasing in quantity and in capacity. They are no longer uniquely restricted to a limited number of enterprises but are in fact very common in some form in almost every enterprise today. Many enterprises now have multiple sources of these workloads.

At a minimum any organization implementing server virtualization today can have multi-tenant workloads; in some cases 20 or 30 virtual servers coming from a single physical server. NAS storage systems have become one of the preferred methods for delivering storage services to the virtual hosts and the access patterns of the virtual machines are inherently random. Storage performance scaling in virtual environments becomes critical as one or more virtual machines begin to consume all the available storage I/O resources which then adversely affects performance across all the other virtual machines on that host, creating a domino effect of lowered performance and lowered confidence in the virtualization project.

Beyond the very common virtual server use case, there is also a rise in the more traditional case of multi-tenant workloads; multiple processing servers grinding through a job. These workloads are not limited to the common example of simulation jobs similar to those found in chip designs or processing SEG-Y data in the energy sector. There are many others; DNA sequencing in bioinformatics, engine and propulsion testing in manufacturing, surveillance image processing in government, high-definition video in Media and Entertainment, and many Web 2.0-driven projects.

Storage I/O performance is critical in these environments because work essentially stops while the processing or simulation job completes. When these jobs stop, so typically does the organization's ability to create revenue. To get around these delays timing of job runs becomes critical to minimize user impact but even with the best planning possible, user productivity will suffer. When that productivity suffers, so does organizational profitability.

Another compounding factor is that all of these data sets have increased in complexity in recent years, becoming more granular, shifting to three dimensions, or significantly increasing color depth. This granularity not only increases the physical size required to store this data but also the processing and storage I/O required to create, modify, analyze or test the data.

**In all cases reliable, predictable, scalable storage I/O performance is critical.**

For example, an integrated circuit designer may need to run a simulation on a particular chip design. As with other environments, this data set is becoming significantly more complex and detailed. In the case of chip design, the chips become smaller or the number of functions on the chip increase. There is a tremendous need in these environments for synthesis and regression testing. As a result the time required to process a simulation of the chip takes longer and longer. It is not uncommon for this type of job to take from three days to an entire month to run. There are two bottlenecks in this process; the time required for the CPU to process the data and the time for the storage to read and write the simulation scenarios.

In the virtual server example, the VMs are almost purely random by nature. While the virtual machines don't have to wait for a particular VM to finish its task, if one VM becomes busy it can dramatically impact performance of the other systems. As in the case of simulation-type workloads, the storage I/O pattern on these systems is as large as it is random.

**The Storage I/O Bottleneck**

While it is necessary to address all of the performance bottlenecks, computer, network and storage, often the most challenging for these environments is handling the storage bottlenecks. The compute bottlenecks are well understood and can be dealt with by allocating a higher quantity of faster processors through techniques like clustered and grid computing or simply leveraging Moore's law:

In 1965, Intel co-founder Gordon Moore noticed the number of transistors per square inch on integrated circuits had doubled every year since their invention. Moore predicted the trend would continue for the foreseeable future. Since then the power of microprocessor technology doubles and its costs of production fall in half every 18 months.

In similar fashion, networking has increased bandwidth via techniques like trunking or multi-homing. These techniques will adequately handle the compute and network element of the bottleneck and are also well understood.

Storage on the other hand has not benefited by a similar Moore's law. While capacities have continued to increase, speed of the disk system has not. This leads to the use of more disk per controller or head and results in the storage I/O bottleneck.

What is lacking from most storage manufacturers is a similar scale-out model, as the current dual controller systems quickly become saturated by these workloads, especially many NAS-based systems. Because of the shared nature of these systems, Network Attached Storage (NAS) should be an ideal storage platform for multi-tenant workloads. Unfortunately because of these workloads' highly random data access patterns and very high number of storage I/O requests, either from a single server with multiple requests in the virtual server example, or a single application making requests from multiple physical servers, single or even clustered, NAS heads and ports can quickly become a severe bottleneck.

The result is that many organizations turn to a shared SAN, which is not by its nature shared, nor is it as easy to manage as a single NAS file system. It too will still lead to bottlenecked storage performance, which again not only slows the business down, limits employee productivity, and eventually loses the organization money, but also adds greater complexity to an already complex environment.

In either the SAN or NAS case, job runs tend to produce a significant number of sequential and random writes requiring an equally large amount of very random reads. This is a deadly combination that renders most cache on these storage systems useless because they are too small to have a high degree of cache hits. The result is that in addition to these bottlenecks most if not all requests have to come from the drive mechanism, not the cache that supports it, further lowering performance.

### **Solving the Storage I/O Problem**

As these workloads become more prevalent across the enterprise, the ideal solution is to solve the NAS bottleneck and establish an easy to manage, high-performance NAS infrastructure -- for many organizations it has become an absolute imperative.

One potential solution is to apply the same methodology behind clustered computing to the storage I/O platform. Build a scale-out NAS solution that increases both storage I/O performance and storage I/O bandwidth in parallel to each other. This would allow for the scaling of the environment, as the workload demanded it. Additionally it would also allow coherent use of memory within the NAS solution creating a very large but cost-effective cache. Finally it would keep the inherent simplicity of a NAS environment as opposed to the more complex shared SAN solution.

An increasing number of customers are searching for ways to solve the challenge brought on by multi-tenant workloads and are being challenged first to identify where the performance problem is, and then to solve the problem. In short they are struggling with using legacy storage to address a modern challenge.

### **Using Legacy Storage to Address a Modern Challenge**

While the demand for capacity is being met, although not perfectly, the need for a scalable I/O model for storage is not keeping pace. This is particularly true when compared to the scalable models that are being implemented for compute and network performance. As a result, storage professionals are spending an inordinate amount of time trying to design solutions to address these performance issues but are being severely limited by legacy solutions.

## **Confirming a Storage I/O Problem**

The first step before creating a solution to a storage I/O performance problem is to validate exactly where in the environment that problem exists.

When identifying a storage bottleneck it makes sense to first look at overall CPU utilization within the compute infrastructure. If utilization is relatively low (below 40%), then the compute environment is spending 60% of its time waiting on something. What it is typically waiting on is storage I/O.

To confirm the existence of a storage bottleneck, system utilities like PERFMON provide metrics which offer insight into disk I/O bandwidth. If there is not much variance between peak and average bandwidth, then storage is likely the bottleneck. On the other hand, if there is a significant variance between peak and average disk bandwidth utilization but CPU utilization is still low, as outlined above, then this is a classic sign of a network bottleneck.

In the legacy model of a single application to a single server environment, the first step is to add disk drives to the array and build RAID groups with a high population of drive spindles. The problem is that a lone application on a single server can only generate so many disk I/O requests simultaneously and to perform optimally, each drive needs to be actively servicing a request. In fact, most manufacturers recommend that an application generate two requests per drive in the RAID group to ensure effective spindle utilization. As long as there are more simultaneous requests than there are drives, adding more drives will scale performance until you saturate the storage system controller or NAS head.

The legacy model, in most cases, can't generate enough I/O requests to feed drive mechanisms and saturate the performance capabilities of the controller or NAS head. In fact, the traditional storage manufacturers are counting on the legacy model, because any other scenario exposes a tremendous performance-scaling problem with their systems.

## **Storage Performance Demands of the Modern Data Center**

The modern data center no longer resembles the legacy model. In general, the modern data center consists mainly of high performance virtualization servers that participate in what is effectively a large compute cluster for applications. Each of the hosts within these clusters has the potential to house multiple virtual servers; each with its own storage I/O demands. What initially began as servers merely hosting 5 to 10 virtual machines has led to servers with the potential to hold 20 to 30 virtual machines per server.

Today, 30 very random workloads per host in the virtualized cluster, easily scaling up to 300+ physical machines in a virtual cluster supporting 1,000 plus virtual workloads is not uncommon, driving multiple storage I/O requests. Consequently, current VM environments can easily demand high drive counts within storage arrays. This significantly heightens the risk of saturating the controllers or heads of current storage platforms.

Furthermore, in many environments there are specialized applications that are the inverse of virtualization -- a single application is run across multiple servers in parallel. As stated earlier these applications are not limited to a few commonly cited examples of simulation jobs but cover a broad range of projects. Many, if not most, companies now have one or multiple applications that fall into this category. As is the case with server virtualization, these applications can create hundreds if not thousands of storage I/O requests. Once a high enough concentration of disk drives is configured in an array, the limiting performance factor shifts from drive resource availability to a limit on the number of I/O requests that can be effectively managed at the controller level.

## **The Key Issues**

The modern data center now faces two key issues. First, because of the high storage I/O request nature of these environments, even if adding drive count to the system scales performance, there is often a limitation on the size of the file system. This limitation forces the use of very low capacity drives which considerably increases the expense of the system. Alternatively, if drives that strike a better price for capacity ratio are used, the file system size limitation also limits the population of drives that can be added to a file system assigned to a particular workload, thus limiting the overall performance potential.

Neither option is ideal. The challenge is that storage systems have to continually evolve to provide ever higher drive counts and capacities to achieve efficiencies that make sense. Likewise, file systems must be able to leverage increasingly higher disk drive counts in order to play into the efficiencies of that model. Regardless, higher drive counts will eventually lead to saturation of the storage controller or NAS head. This is the reason that there is typically a performance bell curve on storage system specification. While a given storage system may support 100 drives, it may reach its peak performance at only half its published capacity -- 50 drives.

## **Searching for a High Performance Storage Solution**

As discussed, throwing drive mechanisms at the problem quickly exposes a more difficult bottleneck to address -- the storage controller or NAS head. The traditional solution to this bottleneck was to add increasingly more horsepower in a single chassis. If it was eventually going to require a V12 engine to fix the storage controller, buy the V12. With the equivalent of a V12 storage controller or NAS head engine in place, disk drive count would have a chance to scale to keep up with storage I/O requests. This V12 storage engine often has additional storage I/O network ports connecting the SAN or NAS to the storage infrastructure.

There are several flaws with this technique. First, especially in a new environment or for a new project, there may not be a need for that much processing power upfront. If all that is required is a V4, why pay for a V12 now? This is especially true for technology where the cost of additional compute power will decrease dramatically in price over the next couple of years. In essence purchasing tomorrow's compute demand at today's prices results in a dramatic waste of capital resources.

Second, it is likely that as the business grows and the benefits to revenue creation and CAPEX controls of server virtualization or compute clustering are realized, there will be a need to scale well beyond the current limitations of the V12 example. The problem is there is no way to simply add two more cylinders to a V12. Instead, a whole new engine must be purchased.

This requirement may come from the need to support additional virtualization hosts with even denser virtual machines or increased parallelism from a compute grid. It could also come from the storage controller or NAS head being saturated by the number of drives it has to send and receive requests. Unfortunately, the upgrade for most storage systems is not granular. If all that is needed is more inbound storage connectivity, the whole engine must be thrown out.

"Throwing the engine out" has a dramatic impact well beyond the additional cost of a new engine. Now work must all but stop while decisions are made as to what data to migrate, when to start the migration and then wait for the migration to complete. This, especially in some of the simulation environments where data sets can be measured in the hundreds of TB's, could take weeks if not longer to migrate.

Finally, even if the organization could justify the purchase of a V12 storage engine, rationalize the eventual need to buy a V16 in the future and deal with the associated migration challenges, there is still the underlying problem of file system size. Most file systems are limited to a range of 8 to 16TB's. While some have expanded beyond that they do so at the risk of lower performance expectations.

As stated earlier, the impact of limited file system size is manifested in both the inherent limit itself as well as the limited number of spindles that can be configured per file system. Again if higher capacity but similar performing drives are purchased, it no longer takes many drives to reach the capacity limitations of a file system.

File system limitation also impacts the speed at which new capacity can be brought online and made available to users. While many storage solutions feature live, non-disruptive disk storage capacity upgrades and some even allow for uninterrupted expansion of existing LUNs or file systems, these niceties break when a file system reaches its limit.

When a file system is at its maximum size, the new capacity has to be assigned to a new file system. Then time has to be spent deciding which data to migrate to the new file system. In this instance, down time often is incurred while the data move takes place.

One potential work around for limited file size is virtualized file systems that logically sew two disparate file systems together, even when the components of the file systems are on different storage controller heads. While these solutions work well to help with file system and storage management, they do little to address storage performance challenges. This is because the level of granularity is typically at a folder level or lower. As a result, the individual heads or storage controllers cannot simultaneously provide assistance to a hot area of a particular file system and once again the single head or controller becomes the bottleneck.

As stated earlier NAS is potentially a preferred platform for these applications but many customers look to a SAN to address some of the performance problems. Reality is that both storage types have the similar limitation of being bottlenecked at the data path going into the storage controller or NAS head, or being bottlenecked at the processing capability of the controller/head itself.

### **The Answer is in front of us**

The answer for solving the storage I/O problem is to leverage the same technology that moved the bottleneck to the storage in the first place. Scale-out the storage environment similarly to the infrastructure now common in the compute layer. By developing the clustered approach pioneered by Isilon's Scale-out NAS, infrastructure drive spindles can be added to match the number of requests by the compute platform on a pay as you grow basis without worrying about hitting the performance wall of legacy storage systems.

### **Solving the Storage I/O Performance Bottleneck with Scale-out NAS**

A performance bottleneck caused by multi-tenant workloads and server virtualization, that users have to "live with", can cost companies revenue, customers, or a competitive advantage, all of which may adversely affect profits and long-term viability.

As stated earlier, for many storage engineers the gold standard for improving performance is simply adding more hard disk drive mechanisms to the storage system. This approach only works, however, as long as there are more requests from the storage system than there are drives to service those requests. As a result, storage performance will continue to scale as drives are added. In this scenario, most server applications eventually become their own storage bottleneck because at some point they will not be capable of generating enough requests to the storage system to sustain drive additions. The challenge that multi-tenant workloads introduce is they can easily generate more requests than conventional storage systems can support--regardless of drive count. Essentially the bottleneck moves from a lack of available disk drive mechanisms for servicing I/O to the storage controller or NAS head itself.

The solution for this problem is found within the very same high I/O workloads that created it to begin with -- server virtualization and/or grid compute environments. These environments allow multiple tenant applications to either live on a single physical server or allow a single application to scale across many servers. The same architecture design is now available for storage. In fact, companies like Isilon Systems are providing scale-out NAS built on a clustered architecture that allows for a scalable, high IOPS NAS to address both short term and long term storage I/O performance bottlenecks.

### **Symmetric Architecture**

The first step in designing scale-out NAS storage is to base it on a symmetrical architecture that enables a series of nodes to be grouped together to act as a single entity. In this clustered architecture, multiple industry standard servers can be equipped with SAS drives and network connections to form a node. Each node can be bonded together through either an Infiniband Network or an IP Network.

### **Symmetrically Aware File System**

Individual nodes are united through software intelligence to create a symmetrically aware file system capable of leveraging disparate components into a single entity. When this file system is applied to the hardware nodes, it creates a high IOPS NAS cluster that can address the challenges of today's -- and tomorrow's -- multi-tenant workloads.

### **Eliminating the Storage Compute Bottleneck**

As stated earlier, with the emergence of multi-tenant workloads, no matter how large a traditional storage system is scaled, no matter how many drives are used, eventually the I/O capabilities of the storage compute engine become the bottleneck. The value of a scale-out NAS configuration is that the storage compute engine is no longer confined to a single system and a single set of controllers or heads, as is the case in a traditional NAS.

With a symmetrically aware file system, such as OneFS® from Isilon, each node in the cluster provides storage compute resources. In fact, it can also ensure that all the nodes in the cluster actively participate. By comparison, some clustered storage solutions must designate a primary set of nodes, typically two, for each request. While these systems benefit from the redundancy of a cluster, they often have the same performance bottleneck of a traditional NAS.

With a cluster-aware file system, each file is broken down into small blocks and those blocks are distributed throughout nodes on the cluster. As a result, when a file is needed, multiple nodes in the cluster are able to deliver the data back to the requesting user or application. This dramatically improves overall performance, especially when hundreds, if not thousands, of these requests are made simultaneously from a multi-tenant application.

Compared to traditional storage solutions where performance flattens out long before the system reaches its theoretical maximum drive count, the symmetrical design of a scale-out NAS system allows performance to scale linearly as nodes are added. Each node in the cluster delivers additional storage capacity in the form of drives, additional cache memory, storage network I/O in the form of inter-cluster connections, and additional connections out to the users. What's more, each node contains additional processing power capable of addressing both external and internal (such as replication, backup, snapshot management and data protection) requests.

## **Beyond Enterprise Reliability**

Since multi-tenant environments often support hundreds of applications, it is critical that they actually provide higher levels of reliability beyond the standard “five 9’s” offered by enterprise-class storage systems. A failure in storage can affect hundreds of applications or the performance of a mission-critical, revenue-generating compute cluster. These workloads also can't be subject to a one size fits all RAID protection scheme. Some applications or even specific files may demand specialized data protection so they can remain operational beyond multiple drive or node failures.

The symmetrical nature of a clustered high IOPS NAS typically delivers beyond enterprise-class reliability. First, there is the inherent value of any clustered environment due to the redundant nodes. When coupled with a file system, like Isilon's OneFS that is fully storage cluster-aware, the platform can deliver granular levels of protection at an application or even file level. This allows for not only data availability and accessibility, in the case of multiple drive or node failures, but also rapid recovery.

Conventional storage systems that are limited to two controllers or heads must process the rebuilding of a failed drive in conjunction with its other tasks. As a result, these systems can take 10+ hours to recover today's high capacity drives. Furthermore under a full load, the time to recover can increase to 20 hours or more. Multi-tenant workloads by their very nature are almost always under a full load and as a result, incur the worst case rebuild times.

## **Cost-Effective - Pay as you Grow Performance**

Finally, performance has to be cost-effective and justifiable. Theoretical high-end storage systems provide a top level of performance by utilizing specialized and expensive processors. Until the environment's storage performance demands scale to match the capabilities of these processors, the investment in them represents wasted capital. Ironically, soon after the environment's performance demands have matched the capability of the specialized and expensive processor, they quickly scale right past the capabilities of that processor. What is needed is a solution that can start with a small footprint and scale modularly to keep pace with the growth of the environment.

Scale-out NAS is the embodiment of a pay as you grow model. The cluster can start at the precise size required to meet the performance and capacity demands of the environment, allowing the upfront investment to match the current need. Then as the environment grows, nodes can be added which improve each aspect of the storage cluster -- capacity, storage performance and storage network performance.

Most importantly, like their compute cluster brethren, NAS storage clusters can take advantage of industry-standard hardware to keep costs down and enable the NAS cluster vendor to focus on the abilities of the storage system, not on designing new chips.

The challenge of multi-tenant workloads can ideally be addressed by a highly scalable but cost-effective NAS. A scale-out NAS system allows you to leverage the simplicity of NAS – maximizing IT efficiency while at the same time outpacing the performance capabilities of most legacy storage architectures. Any organization planning to deploy a multi-tenant workload, whether it is as common as a virtualized server environment or a more specialized revenue-generating compute cluster or application, should closely examine a scale-out NAS solution to fulfill their storage needs.