

STORAGE SWITZERLAND

THE SCALABILITY DEMANDS OF VIRTUAL WORKLOADS



George Crump, Senior Analyst

As the number of hosts in the virtual environment grows and the density of virtual machines per host increases new levels of scalability are required so that storage can match the flexibility of server virtualization. Without the ability to scale storage the expandability and eventual return on investment (ROI) in the virtual infrastructure is greatly hindered. Scaling is a key requirement not only for the storage system but also the storage network that the system uses.

What Needs To Scale? - Capacity

Scalability in terms of storage is most often thought of as the ability to add storage capacity quickly and non-disruptively when the storage system begins to run out of disk space. Adding disk capacity when it's needed is critical for a storage system. The cost of storage is continually decreasing and the longer the storage manager can wait before adding that storage the better. This 'just in time' treatment of storage expansion also requires that it be done during the middle of the day, without fear of an outage. And implementation has to be completed and available to the connected environment quickly. While

most storage systems can physically add storage capacity without disrupting the storage system itself, getting that storage assigned to the attached servers is not so easy. Most of these systems have no concept of storage virtualization technologies that abstract the individual hard drives from the assignment process. Each LUN has to be manually created and then manually provisioned to physically attaching servers. The unfortunate reality is that many systems don't have what should now be considered basic, functionality of storage spindle virtualization and thin provisioning. Legacy systems suffer from the complexity of both adding storage to the system itself and integrating that new storage through the storage network so that attached hosts can take advantage of it. Capacity scaling needs to change so that it can match the dynamic nature of the virtual environment.

What Needs To Scale? - Bandwidth

Scalability is not limited to capacity. Another area where scalability is important is in the area of storage I/O bandwidth, essentially the connection between the storage systems and the physical servers or hosts that are using it.

As more I/O resource consuming hosts are deployed the storage infrastructure needs to have scalability of bandwidth performance. Bandwidth performance is not only needed on the storage system but also on the physical host itself. But most storage systems or solutions only concentrate on increasing storage bandwidth. As more virtual machines are stacked onto a single host that host needs greater and greater storage I/O bandwidth so it can handle the requests of those virtual machines. In other words, bandwidth between the virtual hosts and their storage systems needs to keep up with the increasing density of virtual machines on these hosts. This is going to require cost effective upgrades to the latest in network technology, like 10GbE, and better scaling of that bandwidth.

Once the network interface card has been upgraded to the fastest and most realistically affordable bandwidth, the storage manager has only one option left, add more cards and combine them via software. This is called interface or “link aggregation”. The speed of the individual network card can only go so far and most customers will look to some form of link aggregation to scale bandwidth performance to a given host. The advantage of aggregation is that it uses currently available interface cards and switch technology. Although switch ports are consumed more quickly and more interface cards need to be purchased, this is often far less expensive than upgrading to the next generation of higher speed bandwidth, if there is even an upgrade available.

For many protocols though, aggregating multiple network interface cards is a case of diminishing returns. This is especially true with the IP based protocols, as most IP link aggregation techniques require additional cards and do not effectively double performance. This is partly due to inefficiency in the aggregation process and because the IP to SCSI storage conversation is often done by the host CPU, not the interface card itself. As a result the more cards, the more conversions that need to take place, the greater the performance impact. FC aggregation on the other hand is more efficient, it doesn't put more overhead on the host CPU. The problem is that each additional FC

card adds cost and consumes additional, more expensive ports. In both the IP and FC based protocols adding interface cards may also cause some downtime as the aggregation of cards is made.

For bandwidth to *not* be an inhibitor in scaling a virtual environment it needs to cost effectively scale-up to take advantage of the fastest available bandwidth, typically 10GbE. Then, as the performance limits of a single card are reached by a host, it should scale-out bandwidth via an efficient interface aggregation technique that provides near linear performance as more cards are added.

What Needs To Scale? - Performance

The final area of scale is processing capability. As storage systems are asked to process more and more storage requests and perform more functions beyond basic RAID, the speed of the storage system's processors becomes increasingly important. Many storage systems are simple, one-box configurations with redundant processors, power supplies and network connections. They are simple to configure, implement and initially manage. The challenge is that legacy systems, known as “scale-up” systems, are typically limited to dual processors, which are fine for many small to medium sized data centers but can become a limitation as the environment grows. The other challenge is that all the processing power and, for the most part, storage I/O bandwidth is fixed within the initial system. Each time a drive shelf of capacity is added or an additional host, performance degrades. With scale-up storage the above two resources (capacity and I/O bandwidth) can be quickly exceeded by a virtual workload.

A single unit, scale-out storage system is ideal in its initial configuration. If multiple units can be added, integrated and managed well by the environment, then for some storage managers this is ideal. For larger environments though, the cost to connect and, more importantly manage, multiple systems as the number of units grows becomes a limiting factor to scaling, especially with virtual workloads.

The Scale-Out Workaround

To get around these scaling shortcomings many data centers are beginning to consider scale-out storage. These systems allow storage to be added like bricks to a wall, each brick or storage “node” being a small server with its own processing power, storage capacity and network bandwidth. Each additional node provides an upgrade to all three scaling components. While it sounds ideal, the legacy implementation of these systems creates its own storage scaling challenges.

Scale-out systems were typically designed with file sharing in mind, not enterprise virtual workloads. As a result they may also suffer from limited per-box performance and per-box capacity, which can lead to an extremely high node count when these systems are used to support virtual infrastructures. The problem is that the performance and scaling demands of virtual workloads often lead to nodes being added very rapidly to meet a particular resource demand, causing one of the three components (capacity, bandwidth or processing performance) to get out of balance.

In larger enterprises there is also a latency concern. This may lower bandwidth efficiency caused by high node count because of the amount of inter-node communication that has to occur. These systems can also be a challenge for small to medium sized businesses as well, since they often require that the initial purchase be of three or more nodes. For many small to medium businesses that may be too much capacity and performance to start with; another example of spending money up front for resources that may not be needed for years.

The larger challenge for legacy scale-out storage systems is that they do nothing to fix the scaling of bandwidth at the host, the source of these virtual workloads. These hosts are limited to the same connections, often IP, that legacy scale-up storage systems are. A solution is needed that addresses both the storage system and the storage connection scaling issues so that potentially large virtual workloads can be fully and efficiently supported.

AoE Based Systems Scaling Both Sides of the Problem

An example of a system that addresses the limitations of traditional storage and legacy scale-out storage are those offered by [Coraid](#). Coraid’s EtherDrive storage systems leverage ATA over Ethernet (AoE) to address both the host-side and storage-system side scaling issues. The use of AoE gets around the network complexity that makes adding capacity difficult in a fibre channel environment, but it does so without the overhead of the other Ethernet protocols, iSCSI and NAS. To add capacity, one needs to simply plug an additional storage shelf into the network, make host assignments and hosts can start writing data to it.

As discussed in a recent article by Storage Switzerland "[Storage Evolution: FC SAN, IP SAN, Ethernet SAN](#)" these systems use raw Ethernet as the storage transport, and don't try to convert back and forth between the two. This allows for the use of nearly 100% of the available Ethernet bandwidth and each storage system is independently addressable. To improve host side bandwidth scaling, AoE, unlike the other traditional Ethernet protocols, can scale linearly as cards and ports are added to the host with a technique called “port flooding”. In other words all ports are instantly available to the server and the full, aggregate bandwidth is achieved. Incidentally, if there is a port failure the traffic simply throttles down to the capacity of the remaining available ports, no special configurations are required. The result is that the bandwidth resource can scale in terms of performance, reliability and cost.

Finally, EtherDrive can start as standalone systems, but then be virtualized to create a scale-out storage solution. Unlike scale-out storage, which sometimes is a cluster of limited nodes behind a control head, each Coraid EtherDrive system retains its per-unit high performance and capacity. While scaling with additional systems is virtualized and seamless, the amount of nodes required is more finite than in traditional scale-out systems.

Yet, with the systems virtualized, they benefit from the processing scaling of scale-out storage. This means that data written to the virtualized systems is spread out across all the systems. When a storage I/O request is made all the systems have a part in responding to it. When compared to other IP based systems, both scale-up and scale-out, AoE may provide a more efficiently scalable alternative.

AoE based systems may be the ideal solution for meeting the scalability demands of virtual workloads. Adding capacity is as fast and simple as plugging in an Ethernet cable. Bandwidth automatically uses all available ports and does not require host CPU cycles for complex protocol conversions. Processing power is readily available through either independently addressable system or by virtualizing all the systems for a single management point.

About Storage Switzerland

Storage Switzerland is an analyst firm focused on the virtualization and storage marketplaces. For more information please visit our web site: <http://www.storage-switzerland.com>

Copyright © 2011 Storage Switzerland, Inc. - All rights reserved