ISSUE PAPER

Data Center I/O Consolidation

From Ethernet in the Data Center to a Data Center over Ethernet

Executive Summary

The "data center network" is a myth. For nearly as long as there have been data centers (DC), there have been several DC networks that interact with and overlap one another, most importantly the data, storage and high-performance compute (HPC) networks. The desire to consolidate these networks onto a single fabric is as old as the networks themselves. As network vendors continue to reengineer and ramp up production of 10G Ethernet equipment, the promise of unifying data, HPC and storage networks onto a common technology—Ethernet—increases. Network vendors have some significant technical and engineering hurdles to clear before they can simultaneously meet the opposing pulls of storage, which demands lossless reliability, and high-performance applications, which demand very high throughput at very low latency.

Consolidation onto a single fabric—a DC over Ethernet—will reduce physical complexity, lower material costs and simplify operations. Ultimately, though, the most important benefit of a unified DC fabric will be increased enterprise IT agility deriving from the ability to rapidly and dynamically reprovision network resources across data, storage and HPC domains.

1 The Issue: Multiple I/O Fabrics

The "data center network" is a myth. In truth there are several DC networks that interact with and overlap one another.

Of course, there is the data network. Servers use this to speak with each other and with clients outside the DC, and that is what most people mean when they speak of the DC network. Even that, though, is not really a single network. It relies on two other networks: the DC-to-branch network and the DC-to-DC network. Enterprise IT services reach the majority of enterprise users over DC-to-branch networks, running on T-1, T-3 or even network-to-network virtual private networks (VPN) over the Internet. IT uses DC-to-DC nets, running over Synchronous Optical Network (SONET) rings, Multi-Protocol Label Switching (MPLS) or direct fiber runs, to do things like distribute loads among resource

pools or maintain business-continuity preparedness via storage replication or even virtual-server state replication.

Within the DC, there are other networks as well. Most prominently and commonly, there is the storage-area network (SAN), running on Fibre Channel (FC). The FC SAN represents a second web of connectivity, linking servers to the DC's central storage, as well as linking storage devices like network-attached storage heads, tape libraries and RAID arrays to one another. It has high bandwidth and high reliability, and one of the primary qualities of the SAN is that it does not drop data packets.

In some DCs, another network overlays or partially replaces the data and FC storage nets: the high-performance fabric. Some use a system like Myricom's Myrinet to interconnect the nodes in a cluster, enabling low-latency, high-bandwidth data sharing among systems. Others use InfiniBand for such interconnects, or to connect servers to peripherals for very high speed I/O.

Lastly, there is one more kind of network found in most DCs: a management net. The management net comes in two flavors, and sometimes both are found in the same DC. The first is the traditional management net, which links servers' serial console ports, as well as keyboard, video and mouse (KVM) ports, to a management appliance. The appliance, in turn provides KVM console access to connected systems and sometimes layers other management features on top of console sharing. The second, newer management net is Ethernet. It is like the data network, but runs parallel to it and is confined to the data center. This separate network provides many of the benefits of traditional KVM access, especially if connecting to a remote-management card in a server. But more often, it works as a complement to KVM by supplying dedicated, out-of-band access to the servers for management, monitoring or provisioning systems.

Some networks are with us for the long haul. DC-to-branch connectivity will have to continue, and will have to use transports suitable to long-haul transmissions. Others, though, like the FC SAN and the high-performance net, may not be so distinct in the future.

2 The Drive to Consolidate

The desire to consolidate DC networks onto a single platform is as old as DC networks themselves. The '80s and '90s saw the replacement of other data networking technologies with Ethernet, and ultimately TCP/IP over Ethernet. The '90s saw the replacement of several, older storage interconnects with FC. Successive iterations of HPC fabrics resulted in broad (though by no means universal) adoption of InfiniBand to replace most older technologies.

Both the FC SAN and the high-performance fabric require special adapters and switching equipment, so they cost more to deploy. They cost more indirectly as well, since they add complexity and increase the number of domains of expertise required to run the DC.

So, within a few years of FC becoming the dominant storage networking technology, there were moves to extend it and to replace it by migrating storage traffic onto data networks. FC over IP (FCIP) was created to allow storage traffic to be routed among SANs over IP data nets. Then, iSCSI was introduced as a way to move primary storage traffic—packets between servers and storage devices—

on a data network as well. And as Ethernet speeds continued to rise, many high-performance applications found that they no longer needed to use a different technology for their interconnects. Although the specialized technologies continue to offer significant performance gains, the fraction of implementations that absolutely require such performance is not increasing. Currently, for example, 42% of the top 500 supercomputing sites use Gigabit Ethernet as their interconnect, nearly double the percentage that used it four years earlier.

As network vendors continue to reengineer and ramp up production of 10G Ethernet equipment, the promise of unifying data, HPC and storage networks onto a common technology increases. Convergence to a common infrastructure technology should bring down the costs of deploying high-performance storage and compute networks, even if data centers continue to deploy them in parallel with the data network, instead of converging onto a single infrastructure. Direct costs decrease as the ubiquity of Ethernet brings component costs down, and indirect costs decrease as DC networking and management get easier.

3 Technical Challenges with I/O Consolidation

Of course, the Ethernet and TCP/IP infrastructure that are so well suited to data exchange among systems are not ideally suited to use in storage networking or in HPC applications. The problems boil down mainly to different levels of tolerance for latency and loss.

3.1 Latency for HPC

Data traffic is generally very forgiving if packets don't arrive at a regular pace or if it takes half a millisecond for data to begin to flow. Gigabit Ethernet currently has minimum latencies of around 30 µsec, and 10G Ethernet reduces that to about 10 µsec. However, the most demanding high-performance computing applications expect and require significantly lower latencies; open standards like InfiniBand and proprietary ones like Myrinet provide for latencies around 1 µsec.

Similarly, HPC fabrics provide for significantly higher throughput rates: through aggregation of double- or quad-rate links, speeds of 20G and even 120G bit/sec are possible for InfiniBand, for example.

Lastly, HPC fabric switches, built to interconnect large numbers of systems, currently provide for greater port density at high speeds: 96-, 144- and even 288-port InfiniBand switches are available now, where similar densities in 10G Ethernet are just beginning to ship.

Clearly Ethernet is good enough for many HPC applications—it is the base interconnect for some of the top 500 supercomputing projects. For Ethernet to truly replace HPC interconnects like InfiniBand or Myrinet, however, network vendors will have to reduce latency even further, to well under 10 µsec; scale up port density dramatically without producing equivalent increases in cost; and continue to ramp up speeds, or develop support for link aggregation at the highest speeds. Moreover, for DCs consolidating onto a single fabric, vendors will have to do all this at the same time that they make the fabric responsive to the needs of storage and real-time data applications like VOIP.

Indeed, the rapidly spreading use of VOIP has already helped push network gear to lower and more predictable latencies. Similarly, the continuing growth in demand for denser 10G Ethernet plant is forcing the use of latency-reduction techniques. These include cut-through routing (instead of store and forward), and TCP offload engines, which dump the administrative overhead of TCP onto specialized ASICs, both speeding header processing and freeing the CPU for other work. Such technologies, pioneered in high-performance fabrics, will be critical to the application of Ethernet to high-performance tasks.

3.2 Lossless Layer 2 for Storage

Data traffic is generally capable of handling the occasional dropped packet; one node simply asks the other to retransmit, under TCP, or just ignores the missing data, under UDP. The higher the transmission rate on an Ethernet link, the more likely are dropped packets. When a switch has too many packets to deal with, the easiest way to clear the jam is to drop packets and let them be resent (or ignored). As data rates have increased, the inevitable retransmits have become less and less noticeable on applications using the data network.

Within storage systems, though, there is no tolerance for dropped packets. Consequently, storage protocols such as SCSI and FC go to great lengths to make sure that packets do not get dropped.

So, even as Ethernet must be sped up to meet the needs of HPC applications, it needs to be made more chary of dropping packets if it is to meet the needs of storage systems.

Here again, the spread of VOIP has already pushed network vendors to improve the ability of their equipment to move traffic without dropping packets—dropping pieces of a voice stream is strongly discouraged by users. New standards are also offering a glimpse of the possible future. FC over Ethernet (FCoE) offers a method of moving FC traffic off the specialized equipment it currently requires and onto commodity Ethernet switches and network interface cards (NIC)—as long as the gear supports a critical specification within the 802.3X standard: the PAUSE frame. A PAUSE frame allows one party in an Ethernet packet stream to tell the other to stop sending for a specified time, say, in response to a buffer being close to full. Implementation of PAUSE is optional, and it is also allowable for devices to implement only half of the specification, in which they can either send or understand such packets, but not both. Widespread adoption of FCoE will depend on network vendors and NIC makers implementing the specification fully.

FCoE, or something like it, could be an end state for storage networking over Ethernet, or it may be a step toward using a TCP/IP-based protocol like iSCSI. As the overhead for using TCP/IP decreases (in response to high-performance computing pressure on Ethernet), iSCSI could become more responsive and more broadly adopted. With the infrastructure already converged, switching from Layer 2 use of Ethernet to Layer 3 use of TCP/IP over Ethernet becomes simple.

3.3 Synchronous Replication

An increasingly important issue to the enterprise is zero-downtime storage. One result of this increasing emphasis on truly continuous operation is the spread of synchronous replication of data among DCs. Using traditional FC

connecting over dedicated fiber, the distance limit for such replication is between 50 and 100 miles for most applications. At the root of this limit is latency in the connection; synchronous systems require low latency. The minimum possible latency for a hundred mile separation is 10 msec, due solely to the speed of light limitation within the fiber.

If the DC is truly to move to Ethernet everywhere, then long-distance synchronous replication should not require conversion to good old FC over dedicated, storage-only fiber. If a new, consolidated DC fabric can meet the low-loss demands of storage *and* the low-latency demands of HPC interconnects *simultaneously* on the same traffic stream, it might be possible to squeeze out enough latency at either end of the DC-to-DC links to make synchronous replication practical over business-continuance distances using a shared, Layer 2 MPLS VPN or a full VPLS.

Of course, in the wake of regional disasters like Hurricane Katrina, DC planners have come to understand "minimum safe distance" a bit differently. Getting well past 50- or even 100-mile distances between DCs is now highly desirable. To meet this challenge using Ethernet everywhere, the enterprise must either give up on synchrony—asynchronous replication has no similar limitation—or find ways to drive down infrastructure latencies, for both Ethernet and MPLS, even further, since the latency due solely to distance is not likely to decline.

4 I/O Consolidation for Business Agility

The ultimate reward of the drive to get to a simpler, unified DC fabric is increased enterprise IT agility. Agility is the ability of the enterprise to rapidly adapt infrastructure to pursue new or modified goals and to accommodate new modes of computing. It is best supported by the standardization-driven pooling of resources that allows for the rapid, dynamic provisioning of resources, allocated flexibly to meet new and changing demands of the enterprise.

Standardization of server platforms is what allows IT to create compute pools, with the ability to devote to each application the amount of processing power dictated by demand and priority. Similarly, standardization on a single network fabric will make it easier to devote network resources to the task for which they are most needed, by making the shifting of resources as simple as modifying switch settings and/or swapping patch-panel connections.

Pooling of Resources depends on standardization. When the use of a network link—for HPC interconnects, data or storage—becomes a matter of *how* systems use the infrastructure and not *which* infrastructure is used, then it is easier to pool the resources available to meet all those needs.

Dynamic and Rapid Provisioning is the practical result of the creation of resource pools. Once fabrics are standardized and can be pooled, IT can manually or through automation redirect resources to meet new needs, driven by changes in performance, business continuity requirements or application priority. Being able to do so by reprogramming Ethernet switches and (if necessary) rearranging patch-panel connections increases by orders of magnitude the speed with which such changes can be made.

Flexible Allocation is the corollary of dynamic provisioning of pooled resources. With the ability to reallocate fabric from one domain to another, IT can increase bandwidth for one use, such as storage, by repurposing redundant capacity in another area, such as HPC interconnects, rather than being forced to automatically build out a special-purpose network to meet a growing need in its area.

5 I/O Consolidation: Operational Advantages

The benefits of converging on a single I/O infrastructure go beyond the important benefits accompanying increased IT agility. Reducing the expense and complexity of DCs through this convergence is a sufficient reason to pursue the goal.

5.1 Skills and Training

The skills base of IT represents a significant form of capital for the enterprise. As demands on IT increase—for new applications and for zero downtime—but IT staffing does not, IT staff get spread more thinly. By reducing the number of domains of expertise that IT must master to run the DC, an enterprise decreases both the skills burden on its staff and the training and certification burden on the enterprise, in the long run. It also increases the size of the potential labor pool from which the enterprise can draw skilled engineers. In the short run, though, there will be the need for new training for nearly everyone. Consolidation will create an incentive to combine storage, data and HPC network staff, as well as to better integrate the work of the networking, systems and storage staff.

5.2 Simplified Cabling

One of the most straightforward benefits of convergence is the savings in the DC cabling plant. The use of one kind of switch—Ethernet—connected with already ubiquitous media such as fiber or Cat-6 twisted pair, using standard patch bays, means higher volume purchases and lower prices for enterprises individually, as well as declining prices as overall market volume and competition increase. It also means simpler DCs physically, with (potentially, at least) one switch and one cable run replacing several.

5.3 Ease of Troubleshooting

Consolidation to a common fabric increases the pool of available talent within IT that can address a problem when one arises. For example, organizations may staff three Ethernet experts, instead of one Ethernet, one InfiniBand and one FC expert. At the same time, consolidation decreases the number of potential transport problem domains, while making a common set of tools applicable to resolving issues. Above, the transport layer protocols will still diverge, of course, but consolidating the fabric will pull IT toward tighter integration of staff across the higher level functional domains.

6 Conclusions

Should data, storage and HPC networks truly converge on Ethernet as a common platform, and should we arrive at a DCoE world, then it may begin to make sense to talk about a "DC network" in the singular. The enterprise will reap the benefits of that singular nature in many ways. Operationally, the benefits flow from reduced physical complexity, decreasing material costs and simplified operations. Strategically, they center on improved agility flowing from standards-based resource pooling and the resulting dynamic, flexible allocation of resources. Such convergence will also support and help drive the continued break-down of silos within IT—server, storage and network—forcing closer and better collaboration among the three.

Network vendors have some significant technical and engineering hurdles to clear in simultaneously meeting the opposing pulls of storage, demanding lossless reliability, and high-performance applications, demanding high throughput at low-latency. Meeting those challenges, though, will open new markets to them and create new opportunities for the enterprise.

About Nemertes Research: Founded in 2002, Nemertes Research specializes in analyzing the business value of emerging technologies for IT executives, vendors and venture capitalists. Recent and upcoming research includes Web services, security, IP telephony, collaboration technologies and bandwidth optimization.