nemertes RESEARCH
INDEPENDENCE INTEGRITY INSIGHT
issue paper

Moving to the Data Center over Ethernet (DCoE)

By Andreas Antonopoulos, Senior Vice President, and John E. Burke, Principal Research Analyst, Nemertes Research

Executive Summary

Virtualization is changing the makeup of the data center. The old rules, under which each application required a dedicated server and network interface, are gone. Storage, networking, and high-performance computing are converging at a virtual level despite remaining separate at a physical level. Pooling these compute, storage, and network resources facilitates rapid, dynamic service provisioning, enabling IT to repurpose connections on the fly. Moreover, a unified fabric – virtual and physical – reduces the expense and complexity of data centers, with operational savings driven by lower staff costs and capital savings. Ethernet is emerging as the unifying fabric of choice, although standards organizations are still working to address the key challenges of latency, loss, and performance at scale which are required to ensure that a converged infrastructure performs effectively for all data center applications. For most organization, the best approach is evolutionary. As the enterprise needs for agility and lowered TCO converge with standards-based resilience and reliability, we will eventually arrive at the "data center over Ethernet" (DCoE).

The Issue: Data Center Networking Must Go Back to the Future

The desire to consolidate data center networks onto a single platform is as old as data center networks themselves. Data centers of the 1980s and 1990s were characterized by fixed and incompatible networks for storage and high-performance computing (HPC). There were multiple competing networks – Arcnet, Ethernet, Banyan, Novell, SNA – each with their own hardware and incompatible protocols. This inflexibility and incompatibility ultimately diminished with the emergence of Ethernet, and ultimately TCP/IP over Ethernet, as the de facto standard for networking. The acceptance of Ethernet has led to a robust competitive networking environment in which Ethernet's ubiquity has driven down costs and driven out network vendor lock-in. And, Ethernet's continual evolution to higher speed, and more predictable and manageable performance, increases network flexibility and agility — key factors given that the systems Ethernet supports themselves have become far more flexible and agile.

As Ethernet was becoming the de facto network protocol and interface, the 1990s also saw the replacement of several older storage interconnects with Fibre Channel (FC). Successive iterations of HPC fabrics have finally led to broad adoption of InfiniBand. Both FC SAN and HPC fabrics require special adapters and dedicated switching equipment, adding to deployment costs. They cost more indirectly, as well: They add complexity, and staff require broader skill sets to run the data center or compute cluster.

It's clear, then, that the holy grail of networking in the 21st century is convergence to a common switch fabric. Doing so brings down the costs of deploying high-performance storage and compute networks in a couple of ways. First, adopting a common switch fabric drives direct and indirect costs down as component costs drop. Second, managing networks and data centers gets easier, since the infrastructure is more consistent.

That's not all. A converged infrastructure also lays the foundation for exploiting the benefits of next-generation architectures -- virtualization, in particular.

The Virtualization Dynamic

Virtualization is rewriting the rules of computing and storage. For the past 20 years network, server and application infrastructure remained largely unchanged. Networks were designed to support servers – or clusters of servers – each supporting an application. This fixed relationship between network, server and application also led to implementation of multi-tier networks, concentrating data from edge systems to a concentration layer to a core layer, each with incrementally higher bit rates. This static model is changing rapidly as virtualization takes hold of the data center.

More than anything else, IT wants to become more agile in allowing the business to follow new opportunities or respond to new challenges. By virtualizing servers and breaking the one-server/one-system model, IT operations are getting tremendous flexibility in (as well as increased utilization of) their compute resources.

And not just compute resources. Extending virtualization to storage infrastructure increases efficiency and utilization there, as well. Virtualizing compute and storage resources results in higher processor and storage capacity utilization, by concentrating more work on fewer devices and drives. Virtualization also channels more data traffic through fewer network interfaces, increasing utilization of these interfaces and naturally converging networking and storage onto the same virtual network adapter.

The less an infrastructure is straight-jacketed by complex, special-purpose networking, the more it relies on function, not materials and location, the more agile it becomes. The challenge is to get to there we must learn from the lessons of the past and focus on an evolutionary path to a data center over Ethernet (DCoE), a path that addresses the major critical issues involved in a converged infrastructure.

Latency, Loss and Scale: Three Critical Issues

As noted, there are two major drivers towards a converged switching fabric: Lower costs, and increased agility. A converged switch brings down the costs of deploying high-performance storage and compute networks in two major ways: First, a common switch fabric reduces direct costs by standardizing components, bringing in economies of scale that help drive equipment costs down. Second, managing networks and data centers gets easier (thus cheaper) given a more consistent infrastructure. And the ability to manage a consolidated switching infrastructure ensures the infrastructure's agility can keep up with the demands of the application (and the business).

For these reasons, consolidation is already occurring. Storage has begun to flow over Ethernet via Fibre Channel over IP (FC/IP) and iSCSI, and Fibre Channel over Ethernet (FCoE). Gigabit Ethernet has begun to eat away at the low end of HPC requirements, making steady inroads against Infiniband as a high-speed, low-latency interconnect.

For Ethernet to truly support the DCoE, however, it must first address three critical issues: Latency, scale, and loss.

Latency and Throughput. Data traffic is generally very forgiving if packets don't arrive at a regular pace, in the right order, or if it takes half a millisecond for data to begin to flow. However, the most demanding HPC applications, such as those using shared-memory clusters and massive symmetric multiprocessing, require very low latency and very high reliability. For example, HPC interconnects like InfiniBand (IB) exhibit latencies around 1-3 microseconds. In comparison a typical 1 Gigabit/s Ethernet connection has latency in the 300-400 microsecond range. To take over most HPC uses, Ethernet will have to exhibit sub-5 microsecond latency. HPC fabrics also provide very high throughput rates: Through port aggregation IB achieves speeds of 20 to 120 Gbps. The bottom line is that supplanting IB requires high-bit-rate Ethernet at both switch and network interface card (NIC) to drop latencies low enough and lift throughput rates high enough. This still doesn't address issues of density, scale and loss.

Density and Scale. Many data center functions now scale with virtual machines rather than physical ones, reducing the requirements for physical port counts but increasing the need for massive throughput. Conversely, HPC functions continue to scale physically, not virtually, because they already drive extremely high utilization of processor and I/O resources. Legacy HPC fabric switches have been built to interconnect large numbers of systems, potentially with several trunked ports for each; they provide port densities of 96, 144, even 288-ports per switch. Since they will continue to require large numbers of physical servers with physical NICs, HPC over Ethernet use cases will also continue to require high switch-port counts.

Loss. Ethernet drops packets by design: When a switch has too many packets to deal with, the easiest way to clear the jam is to drop something. 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.

Dropping data is unacceptable for applications like HPC and storage clustering, however. As a result, storage protocols like SCSI and FCP (the FC equivalent of TCP) emphasize flow control, and FC engineering leans heavily on over-provisioning bandwidth and switching capacity to make sure that packets do not get dropped. For Ethernet to replace FC it must do a better job of reducing drops and retransmits. In other words Ethernet must become lossless -- and it is not there yet, today.

10 Gigabit Ethernet: A Unified Fabric for the Data Center… Eventually

The best emerging candidate to be the unifying fabric in the data center is 10 Gbps Ethernet (10GigE). We've already seen 1 Gbps Ethernet become a unifying fabric for voice and data. The spread of Voice over IP (VOIP) has already pushed Ethernet switch vendors to improve the ability of their equipment to move traffic without dropping packets. These improvements nicely support use of Ethernet for storage traffic, too.

Truly lossless Ethernet relies on extensions to the old Ethernet standard, which are currently in development by the converged enhanced Ethernet (CEE) groups in the IEEE. (Please see Figure 1: Emerging Ethernet Standards, Page 4).

Emerging Ethernet Standards

DCB – Data Center Bridging is an IEEE movement to make Ethernet deterministic at high speeds by implementing features at layer 2 rather than at layers 3 and 4 of the OSI stack.

CEE - Converged Enhanced Ethernet to enable protocol convergence over Ethernet is a proposed set of standards to the Internet Engineering Task Force. Proposals include:

- Priority-based Flow Control (PFC) for link level flow control for different classes of service;
- Enhanced Transmission Selection (ETS) is a framework for bandwidth assignment to different classes of service;
- Data Center Bridging eXchange (DCBX) for discovery of configuration and capabilities between neighbors.

Figure 1: Emerging Ethernet Standards

As discussed above, to completely replace legacy interconnects such as Infiniband, 10GigE must have very low latency, very high port density and very high throughput. And, we're already seeing ~1 microsecond 10GigE switches with hundreds of ports on the market. As more follow and standards coalesce, 10GigE will likely replace all commercial HPC fabrics for most uses except for a few niche use cases. Consequently, we expect to see increasing numbers of compute facilities with computationally intensive, data-rich problems such as

meteorology, seismology, and geographic information systems relying on 10GigE as the interconnection framework.

Meanwhile, the natural affinity between server and storage virtualization and iSCSI has driven an increasing number of organizations to shift increasing amounts of block storage traffic to Ethernet. Already, approximately 25% of organizations are adopting 10GigE in the data center for network storage. (Please see Figure 2: Adoption of 10GigE for Network Storage, Page 5).

Likewise, the Fiber Channel over Ethernet (FCoE) standard brings the reliability of FC traffic to the Ethernet infrastructure. FCoE uses Fibre Channel Protocol (FCP)'s packet loss/retransmit mechanisms to address loss and 10 Gbps speeds help ensure that traditional over-provisioning on storage networks remains possible and practical to address latency, throughput and scale.

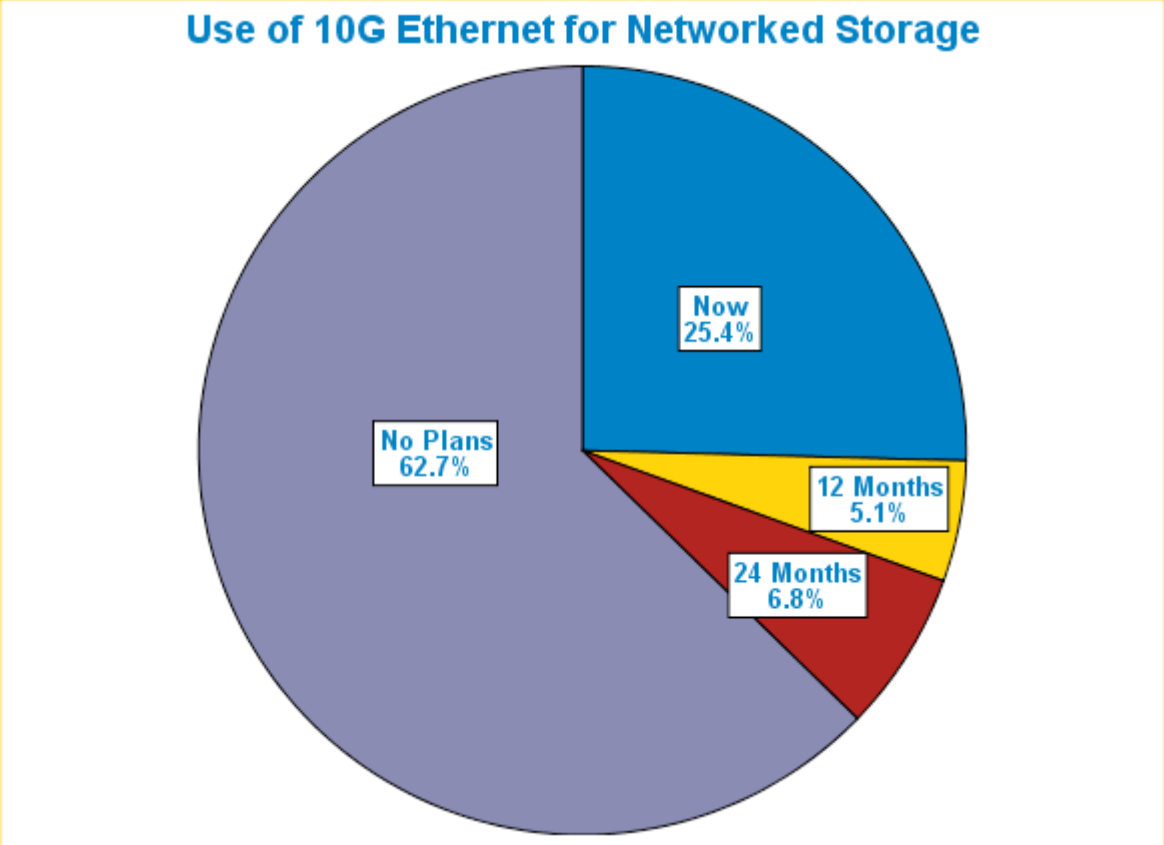

Figure 2: Adoption of 10GigE for Network Storage

The bottom line is that with the advent of affordable 10GigE connectivity, we have seen increasing numbers of even high-load storage access incorporate block storage access over Ethernet, including medical or satellite image processing and high-speed, high-volume database manipulation.

The Fabric is Still a Patchwork

Though 10GigE is very attractive as a unifying fabric -- prices are dropping and adoption is rising -- we're still a long way from a unifying fabric in the data center. Nearly 63% of organizations have no plans for network storage over 10GigE. (Please see Figure 2: Adoption of 10GigE for Network Storage, Page 5). And about 71% of organizations have no plan yet to converge data center switching fabrics into one unified fabric. (Please see Figure 3: Plans for a Converged Fabric, Page 6).

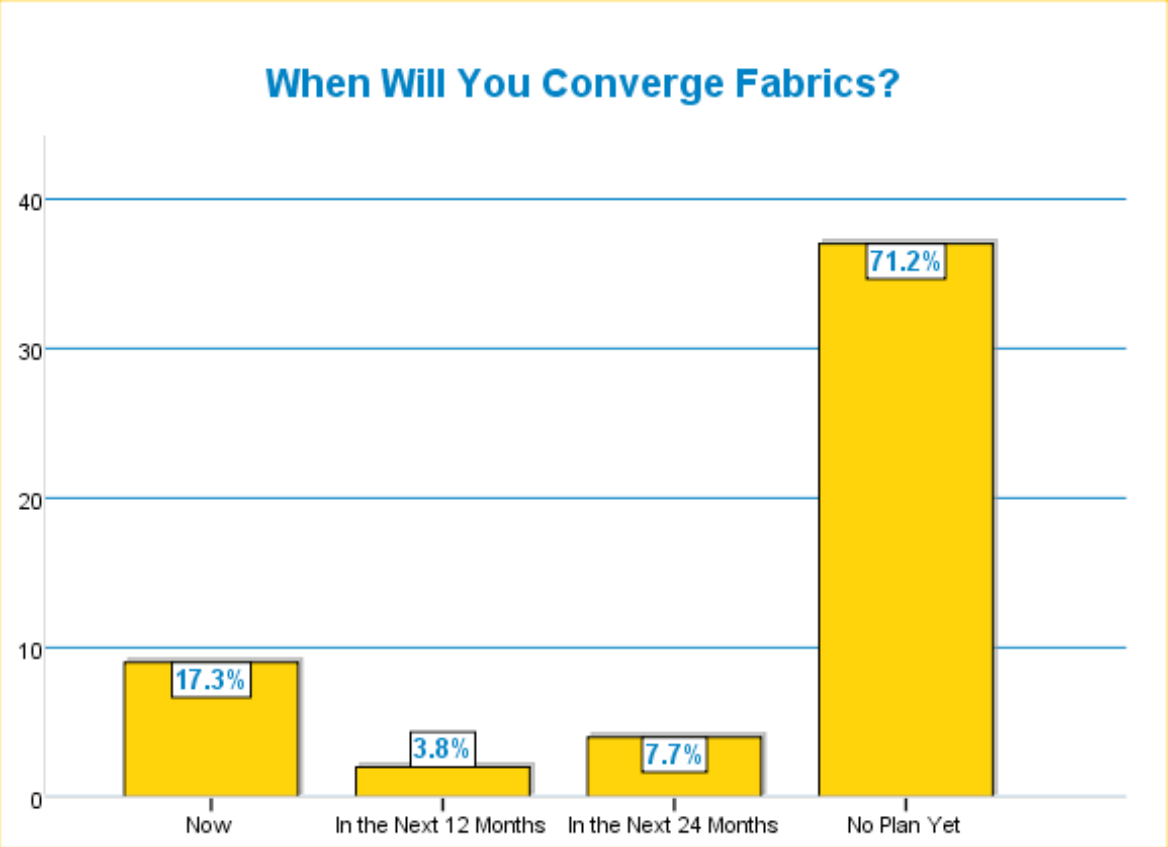

Figure 3: Plans for a Converged Fabric

The main reason that many organizations hold off on implementing a converged fabric is standards are still evolving. Vendors can improve various aspects of Ethernet performance — flow control, packet loss — by adding non-standard extensions to it, but at the cost of interoperability. The only time a non-standard strategy pays off for IT is when a network is going to be single-vendor, end-to-end, such as some small, dedicated compute clusters. When a facility is green field and strategic relationships lock in a vendor commitment over the life of the facility, then a proprietary approach may be both practical and justifiable. However, such cases are rare.

Despite vendors best efforts to implement the equivalent of DCB and CEE, these are still equivalents; resulting in potential vendor lock-in. After all, the key to Ethernet's success over the last three decades has been standards-based interoperability. Network engineers have long been able to count on one Ethernet device talking to any other Ethernet device: Intel NIC to 3COM switch to Cisco router to Juniper firewall – no matter, full function guaranteed by adherence to the standards. The same must be true for 10GigE.

This complex, multi-vendor, multi-generational reality in the data center, more than anything else, makes standards adherence a rock-solid requirement for technology as fundamental as Ethernet.

Multivendor Reality Means Lower TCO Over Time

The reality for almost all IT organizations is a multi-vendor, always-in-flux data center. A compute cluster may last for a decade or more; a data center, for 30 years or more. A typical facility therefore lasts across multiple complete lifecycles of all major technologies in it: compute, storage, and networking, even electrical and HVAC.

Typical network lifecycles run from three to seven years for storage, edge, aggregation, and core switches. IT brings in new network gear as older equipment requires replacement, and new compute and storage systems coming online drive requirements for more ports or newer features and higher speeds. In fact, new system requirements can trigger out-of-cycle replacement of whole tiers of network equipment.

It is Ethernet's ubiquity that has made it easy to extend the useful life of networking equipment. Today's core switch becomes tomorrow's aggregation switch and today's aggregation switch becomes tomorrow's edge switch, regardless of vendor. For this to continue any new investment in switch technology must be 100% interoperable with existing infrastructure. Proprietary extensions may grant temporary performance advantages *if one commits to a vendor end-to-end*. However, the reality for most environments is that each tier and type of equipment has to work with a mixture of older and newer gear from a variety of vendors. The result is lowest-common-denominator (standards driven) performance end to end, regardless of networking vendor and the focus on Ethernet as the common denominator unifying fabric.

This complex, multi-vendor and multi-generational reality in the data center is really an issue of Total Cost of Ownership (TCO), with "total" rightly understood as a function of the whole lifespan of the equipment. Standards

adherence results in the lowest aggregate TCO for a network. Standardization makes equipping each tier of infrastructure an open decision that can always be made based on cost, both immediate and long term. Proprietary gear eliminates the buyer's ability to play vendors against each other to improve a deal, or to substitute higher cost gear for lower. In the short term, per-port costs on any tier may be higher in a standards-based infrastructure, for example for first-generation support of a new capability, like CEE. Even so, over time and across the network infrastructure, TCO will be lower if the focus remains on standardization and interoperability.

Evolution or Green Field?

As discussed above the only situation that makes sense today for a unified switch fabric (10GigE or any other) is a green field opportunity or a rip-and-replace where the risks and costs of vendor lock-in – because of standards immaturity – may be justifiable to achieve the operational and capital cost advantages of a unified fabric. However, these situations are rare, especially in today's tough economic climate. So, this leaves an evolutionary approach to a unified fabric the best option. After all, this is how Ethernet became a common networking fabric in the first place.

An evolutionary approach means upgrading to 10GigE where it most makes sense while leaving the rest of the infrastructure intact; selective upgrades to 10GigE where the economics and standards are clear. The most logical first step toward convergence is at the adapter level. Converging storage (iSCSI and Fibre channel) and networking (TCP/IP) on one Converged Network Adapter (CNA) eliminates need for separate network connectors: Ethernet and Host Bus Adapter (HBA). This reduces cabling while protecting the investment in fire channel storage. Also, CNAs use the ratified standard for FCoE so there is no risk of vendor lock-in because of standard immaturity. The CNA does require an upstream FCoE-capable switch. A rational approach is to move to a top of rack (TOR) switch architecture with multiple 10GigE uplinks to further simplify rack and data center cabling. Simplification of rack cabling has a number of benefits. First, it reduces complexity, which always reduces the risk of human error. Second, it reduces the adapter count on the servers, potentially from 4-6 down to 1-2. As prices of 10GigE CNAs drop, the economics of this port count reduction becomes more clear. And, third a physical CNA matches the consolidation that is already occurring with virtualization and virtual network adapters that combine storage and networking.

Yet, overall adoption of FCoE is still very low partly because of the cost of rip and replace of the broader FC infrastructure. And, until standards such as CEE and DCB are complete, a lack of end-to-end performance on Ethernet makes widescale FC to FCoE transition technically impractical. For the foreseeable future a logical focal point for this coexistence of FC and FCOE is the TOR switch. As standards are ratified a 10GigE unified fabric can extend from the TOR switch across the data center infrastructure leading to the eventual unification of storage,

networking and high performance computing resulting in the emergence of a data center over Ethernet (DCoE).

Conclusions and Recommendations

The ultimate reward of a unified data center fabric is increased enterprise IT agility. Agility is best enabled by pooling of resources, which allows for rapid, dynamic provisioning of services separate from (orthogonal to) capacity management in the resource pools.

Pooling of resources depends on standardization. Ideally, a single fabric will be used to meet the performance needs of HPC applications, of storage access, and of aggregated network access for virtualized systems. When the *type* of a network link — HPC, storage, data — becomes a matter of how systems use it rather than which infrastructure is used, then it is easier to pool resources to meet all needs. 10GigE is the fabric that can become this single standard ... eventually.

Once CEE and DCB are finalized, IT can manually or through automation reassign how connections are used on the fly, driven by changes in performance, requirements, or priorities. The ability to do so by reprogramming Ethernet switches rather than rearranging physical connections to multiple switch types will increase by orders of magnitude.

Beyond increasing agility, a unified fabric will reduce the expense and complexity of data centers. Operational savings will derive from lower staff costs (fewer skills pools required). Capital savings will come from higher volume purchases of Ethernet switches at lower prices, as increased market volume and competition drive prices down. Using one kind of fabric, and in many places one instead of several actual devices, will result in a simpler and easier to maintain physical layer.

Nemertes recommends that anyone building or managing networks:

- Plans on converging networks over the next ***five to seven*** years.
- Begins converging now, using iSCSI and other storage-over-Ethernet protocols starting with converged network adapters.
- Takes each expansion or upgrade decision as an opportunity to move convergence forward.
- Commits to standards-based networking at all levels.

Data, storage, and HPC networks will eventually converge on Ethernet as a common fabric. Agility, lowered TCO, and standards-based resilience and reliability will drive this convergence, and will lead to a "data center over Ethernet" world.
