Introduction: Converging SANs and Ethernet

The cost of I/O connectivity for servers, as a percentage of overall server cost, is growing rapidly. Server virtualization has dramatically increased the relevance of storage area networks (SANs), as mobile workloads require storage volumes to be accessible by all the physical servers in a server farm. Direct-attached storage (DAS)—the most cost-effective choice for many workloads in the past—simply does not support workload mobility from one physical server to another. For workload mobility, a SAN is required; both Fibre Channel and iSCSI SANs are used for this purpose.

In order to consolidate server I/O connectivity, Ethernet-based SAN protocols such as iSCSI and FCoE have been developed to allow data and storage traffic to share I/O links. iSCSI works today and is growing rapidly in popularity; FCoE is a brand new protocol that is still in development and largely untested. Both have their strengths and weaknesses.

Problem statement

The cost of server CPU cycles and memory is decreasing much faster than the cost of server I/O. Modern servers typically connect separately to multiple shared networks:

- IP networks for production application services
- IP networks for application infrastructure like file servers, DNS, and directories
- Archival networks for backups
- Management networks for system administration
- Storage area networks for shared-access block storage
Typically, each of these networks requires dedicated, isolated infrastructure, requiring separate network interfaces on each server. Why are these networks isolated from one another? Reasons include:

- Different security requirements
- Different traffic pattern requirements
- Different people responsible for management
- Different lifecycles and toolsets
- Ability to isolate and troubleshoot problems
- Ability to minimize the impact of a future change or a problem

As a result, there has been a proliferation of I/O connections on each server, so much so that the I/O cost per server can exceed the remaining cost of the entire server. For example, if the NICs and HBAs in a server cost a total of $3000, and that cost can be reduced by 50%, there would be $1.5M in I/O savings alone for a deployment of 1000 servers. This might be a realistic scenario when replacing five NICs and two HBAs with two converged network adapters (CNAs) that use virtual traffic lanes to consolidate all seven functions. These CNAs are Ethernet NICs that also provide a façade to server applications of both NICs and SAN HBAs.

There are two main classes of networks found in data centers today: IP data networks and storage networks. For the purposes of this discussion, storage networks are defined as those providing a block volume abstraction via initiator/target protocols like SCSI. The pervasive technology used for IP data networks today is Ethernet, and the primary technology used for storage area networks is Fibre Channel (although iSCSI SANs are growing fast). Exceptions exist, of course, notably technologies like Infiniband.

There are three key observations about the value proposition of storage over Ethernet:

**The primary value proposition of storage over Ethernet is I/O consolidation.**

I/O consolidation is the reduction of physical links (and associated cost) required to provide server connectivity to both the storage and data networks. It’s very important to distinguish server I/O consolidation from fully combining these networks—there is actually much less interest in fully combining SANs and IP networks beyond the cost benefits of I/O consolidation. To this end, storage over Ethernet is the set of protocols and technologies for sharing server links between storage and data traffic. The objective of storage over Ethernet is to reduce the cost of data center I/O.

**I/O consolidation is more valuable on the “server side” of the SAN.**

In most SANs, links to storage arrays are more highly utilized than links to servers. Think about it: in a shared storage environment, there will typically be more servers sharing fewer storage systems. Thus, server links are more numerous with more wasted utilization. Similarly, on the IP side of the house, server connectivity to the network typically over-subscribes the core of the network by many factors. Thus, for both networks, consolidating server links is the big payoff.

**SAN and IP/data network administration will not merge anytime soon.**

The requisite skills required for both disciplines are increasing, not decreasing. History has shown again and again that solutions requiring more skills to implement always lose the race to solutions requiring fewer skills to implement. If the benefit of I/O consolidation can be achieved without “re-skilling” infrastructure administration, then that “re-skilling” will not occur.

**Protocols for storage over Ethernet**

**iSCSI**

iSCSI was originally invented to provide consolidation of storage and Ethernet. iSCSI encapsulates a serial SCSI protocol over TCP/IP, which enables iSCSI to re-use the most common transport mechanisms in networks today, allowing iSCSI packets to be carried over commodity Ethernet infrastructure. All modern Ethernet switches are well-suited for the venerable TCP/IP use case—the TCP mechanisms for bandwidth sharing and congestion management allow typical Ethernet switches to be low-cost and relatively simple in design.
Perhaps the biggest imposition of iSCSI on Ethernet switches is the characteristic block size of storage volumes, which at 2 KB is larger than the nominal Ethernet frame size of 1.5 KB. This is one reason why jumbo frame support is so common in Ethernet switches today.

For years iSCSI adoption lagged behind aspirations for this technology. Fibre Channel itself has evolved, making it a moving target for iSCSI to replace. Also, the SAN technology ecosystem is clearly subject to network effects, whereby each new solution makes even more solutions possible. These interdependent solutions accumulated over time make Fibre Channel even more indispensable for the use cases it uniquely solves—highly available storage networks that can scale and host large mission-critical databases. The management tools and administrative skills for Fibre Channel reinforce the strength of this ecosystem. These tool sets and skill sets have rendered iSCSI a consensus “non-answer” to replacing Fibre Channel SANs with Ethernet—at least for the mission-critical database use case.

The advent of virtualized servers, however, has driven a significant increase in iSCSI usage. In particular, the need to move virtual servers from one physical server to another (e.g., using tools like VMware’s “vmotion”) has driven more need for low-cost SANs. In particular, a mobile application requires a storage volume that can be accessed by multiple physical servers. iSCSI provides an inexpensive solution to this problem that also minimizes the SAN administration skills required—making iSCSI a good choice for smaller IT teams with a limited budget.

**FCoE**

Fibre Channel over Ethernet (FCoE) is a brand new protocol, still in the process of standardization and initial test deployment. FCoE, unlike iSCSI, does not rely on IP for traffic routing; rather, FCoE sits right on top of Ethernet, as an alternative to IP. Much of the motivation for FCoE was to address shortcomings of iSCSI that have stunted iSCSI adoption. For example:

- Performance: FCoE is intended to relieve initiators and targets from the processing required to be IP endpoints in an iSCSI SAN.
- Interoperability with Fibre Channel: FCoE is intended to preserve the Fibre Channel software stack (albeit layered over Ethernet).
- Management toolset: FCoE allows existing Fibre Channel SANs and management to be retained.

The downside with FCoE, of course, is that it requires modification of the Ethernet standard. Since the Fibre Channel approach to security, flow control, QoS, congestion management, learning, addressing, topology, and others are all different from Ethernet, FCoE requires Ethernet to change and add capabilities to accommodate Fibre Channel.

The changes to Ethernet to support FCoE include mechanisms to separate SAN traffic flow control from IP traffic flow control, and mechanisms to notify senders of congestion more proactively at Layer 2 than via today’s method using TCP. Bottom line: the models for flow control are vastly different between Fibre Channel and Ethernet. Ethernet historically relies on endpoints to manage congestion and flow control, primarily using TCP. Fibre Channel uses buffer credit schemes and does not handle drops very well.

It’s not a foregone conclusion that FCoE will be adopted as pervasively as some vendors hope. Not only are the FCoE standards not yet complete, but the complexity of actually implementing a converged data center network from the customer’s perspective will be daunting. There will be challenges in configuration, troubleshooting, monitoring, and determining which group in IT actually owns what. Two things are clear, however: 1) actual widespread adoption of FCoE will take some time; and 2) convergence of the networks will be limited to as few hops as possible, focused on the server side of the SAN.
**Best practices for I/O consolidation—today**

Since the value proposition of storage over Ethernet is really I/O consolidation—particularly on the server side, where utilization is lower and the payoff for consolidation is much higher—it makes sense to discuss best practices for I/O consolidation that should be considered today.

**Part 1—bladed servers**

For the best I/O consolidation, the obvious server choice is bladed servers. Bladed servers provide significant cable aggregation whenever Virtual Connect modules, Ethernet switches, or Fibre Channel switches are embedded in blade chassis. HP c-Class BladeSystem products offer many choices for cable and I/O aggregation.

**Part 2—Virtual Connect and Flex10**

By aggregating server connectivity into more highly utilized consolidated links, HP BladeSystem Virtual Connect provides significant I/O consolidation without placing a heavy burden on network and SAN administrators. Server traffic for respective networks can be aggregated to very few tagged uplinks, as if presenting to the outside world a set of very few HBAs and NICs that are shared by all the servers inside the chassis. Virtual Connect can also front-end virtualized servers quite well.

Flex10 allows a single NIC to be partitioned into multiple virtual NICs with configurable bandwidth on each; this provides an easy-to-use fair-queuing mechanism for consolidating virtual interfaces.

**Part 3—10G top-of-rack (TOR) switches**

Since most data center networks and SANs are highly over-subscribed anyway, it makes sense to aggregate the network as close as possible to the servers. By containing the maximum number of network and SAN links in the server racks themselves, the fewest possible cables remain to be distributed throughout the rest of the data center. Also, by confining the network edge to server racks, cheaper copper cabling can be used for the short runs, providing a huge savings in I/O when 10Gb links are required. HP ProCurve switches offer excellent 10Gb top-of-rack density with the HP ProCurve Switch 6600 product, and coupled with SFP+ direct-connect copper cables, can provide huge I/O consolidation savings in data center deployments.

**Part 4—iSCSI SANs**

If a working SAN + Ethernet solution is required today, the obvious choice is iSCSI. The vast majority of common storage device products today offer iSCSI as a connectivity option. iSCSI is growing rapidly and is a very popular choice for deploying new SANs. From a network perspective, key capabilities are jumbo frames, buffering, and flow control. HP’s new data center switches are purpose-built to support iSCSI SANs—both 1 Gbps and 10 Gbps connections. With 3 Mbytes of buffering per port, coupled with tuned flow control and jumbo frame support, the 24x10Gb port top-of-rack HP ProCurve Switch 6600 is an ideal I/O consolidation switch for iSCSI. HP StorageWorks products are industry leaders in iSCSI support. From the Modular Smart Array products to the Enterprise Virtual Array and XP systems, the HP StorageWorks portfolio is very popular for iSCSI connectivity and growing more so. HP also offers world-class management tools for iSCSI SANs, including industry-leading capabilities from LeftHand Networks, now part of the HP portfolio.

**Part 5—virtualized network connections**

A significant barrier to I/O consolidation is the common practice of static network plumbing. Why? Two reasons: 1) static plumbing frequently involves error-prone manual or scripted switch configuration; and 2) static plumbing requires either one-size-fits-all port configuration or weighty change processes that place a constraint on which workloads can be assigned to which servers. Both of these characteristics limit I/O consolidation.

The new HP ProCurve Data Center Connection Manager (DCM) enables I/O consolidation by allowing dynamic policy assignment to Ethernet connections—as a workload is being provisioned or moved—and by providing a mechanism for managing, tracking, and validating the compliance of network policies in the process.
By providing this flexibility, DCM reduces the need for dedicated-purpose access layer switch ports and makes sharing of ports and links more practical and efficient.

**Prospects for I/O consolidation—tomorrow**

**Potential future alternatives to FCoE for I/O consolidation**

Since the primary near-term value proposition of FCoE is server I/O consolidation, it stands to reason that server-centric I/O technologies will represent alternatives to FCoE. For example, a purpose-built server I/O network like PCIe—or its evolutionary follow-ons—could be optimized for the use case of providing shared, consolidated NICs and HBAs. While Ethernet takes on more complexity over time, a simple I/O sharing bus architecture need not support the baggage of Ethernet’s installed base and protocols, diversity of endpoint types, traffic patterns, and use models. Rather, a shared I/O bus architecture can target solely the rack consolidation problem and yet still drive high volume due to leverage of server volumes.

Bear in mind that nobody has standards-compliant FCoE wares today. They can’t—the standard is not expected to be completed prior to the end of 2009. Furthermore, Fibre Channel SANs host some of the most precious enterprise data, running on systems that cannot tolerate downtime for migration or retooling. Fibre Channel as we know it today will not be going away any time soon. In fact, upgrades to native Fibre Channel will continue for quite some time. Analysts agree that even in the most aggressive scenarios, FCoE will not reside on “main street” for years to come.

What are the implications of this slow rollout? Well, unlike VoIP’s gradual ascent to replace POTS, FCoE’s benefits are not really in the network itself. Rather, FCoE’s most valuable benefits are in I/O consolidation at the endpoints. It would be as if VoIP allowed you to buy a cheaper phone, but didn’t really make the network itself any better or cheaper. However, FCoE, like VoIP, forces you to change the network itself. Aren’t there ways to get a cheaper phone—i.e., I/O consolidation—without incurring the pain of replacing the network? That’s the argument driving development of shared I/O alternatives to FCoE. If an I/O backplane can be extended to solve the same problem as FCoE, there will be much less pain required to take advantage of it.

The most important implication of FCoE’s slow rollout, however, is probably something nobody knows about yet! Since I/O consolidation can be achieved in a number of ways, the industry has plenty of time to build a “better mousetrap” while FCoE endures the long, slow evolutionary adoption cycle.
How might FCoE evolve over time?

If FCoE gets adoption traction, its evolution will probably occur in stages:

Stage 0—typical data center network today

Stage 1—homogenize server I/O connectivity

This stage of FCoE evolution is intended to allow customers to use the same I/O cards and switches for both IP and SAN traffic, while dedicating separate Ethernet domains. Since the Ethernet enhancements required for FCoE—priority-based flow control, transmission selection, and congestion notification—are still pending standardization and ASIC changes, there may be very limited non-proprietary options for FCoE deployment. If 10Gb links are dedicated to FCoE, however, some barriers to adoption are removed. For example, a network designer might dedicate entire ports and a VLAN to a given SAN, much like what occurs today with iSCSI deployments, and avoid tangling with the tougher plumbing and IT governance issues until the technology is more mature. The 10Gb top-of-rack HP ProCurve 6600 Switch is a great fit for this use model, given the large 3 Mbytes of per-port buffering, jumbo frame support, and flow-control tuning.
Stage 1—standardize NIC types
 stil dedicated SANs, but using Ethernet)

Stage 2—consolidate NICs and HBAs

This stage of FCoE evolution involves converged network links between servers and access layer switches. SAN traffic and IP traffic are shared on a 10Gb link. This model exploits the primary current value proposition of FCoE—-I/O consolidation on servers. The assumed model is each physical server has one or two 10Gb CNAs, and the bandwidth of 10Gb is enough to satisfy a dozen or more virtual machines. Traffic is split and uplinked to separate SAN and IP networks in the access layer switch. Since SAN traffic uses different priority queues than IP traffic, congestion in one network does not affect the other so long as per-priority pause is enabled. Note that neither IP nor SAN traffic actually has priority over the other—to do so would create a starvation problem since both networks require significant bandwidth to function. HP intends to support any industry standards that emerge for FCoE. Note also that the top-of-rack switch will typically “look like” a server HBA (i.e., an N_Port or NPIV) to devices on the SAN.
Stage 2—homogenize server I/O—universal NICs with TOR splitting

Stage 3—flexible access domains

The next step beyond I/O consolidation is to enable servers from multiple server farms to access any given LUN, and to enable any given server to access multiple SANs from a single link. This flexibility is not to support workload migration, but rather flexible provisioning and workload assignment. With this model, new servers can be added or assigned anywhere in the data center; new LUNs can be added or assigned anywhere in the data center; and the association of a particular server with a particular LUN can be orchestrated without inter-dependencies. This stage of evolution is taking advantage of the fact that Ethernet subnets are typically larger than Fibre Channel subnets, and there is value to this flexibility.
Stage 3—Ethernet enhances SAN flexibility—managed virtual HBAs “on-a-rope”

Stage 4—single network

If a “single network” ever materializes, very few experts believe it will be based on Fibre Channel like FCoE is. Coercing Ethernet to adopt Fibre Channel’s limitations would be too high of a price to pay. Coercing Fibre Channel to adopt Ethernet’s limitations already has a name—it’s called iSCSI.

Stage 4—fully converged network? 2015? Pipe dream?
Note that much of the hype around FCoE today results from the fact that 10Gb NICs provide more I/O bandwidth than most servers need. Thus, NICs can be shared for multiple purposes. It’s likely that by the time widespread use of FCoE is possible, two 10Gb NICs won’t provide enough bandwidth for a typical server. If a typical server will need four NICs anyway, the benefits of FCoE will be diminished—two NICs could simply be dedicated to SAN, while two NICs would be dedicated to IP.

Conclusion

FCoE will take some time to achieve wide adoption. HP intends to support any industry standards that emerge from FCoE and is even helping with many of these standards activities. HP ProCurve, along with HP servers and StorageWorks products, has solutions tailored for storage over Ethernet use models and will continue to invest in advancing the state of the art. iSCSI is enjoying significant growth today and will tomorrow also, and it should not be overlooked by customers interested in converged data center networks.

For more information

To learn more about HP ProCurve Networking, please visit ProCurve.com

© Copyright 2008 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

4AA2-3826ENW, December 2008