Building I/O Fabric Super Highways in the Data Center

By David Lytle

Simply described, a super highway is a large, wide highway used for traveling at high speeds over long distances. When all goes well, you can quickly travel from point A to point B. However, even the best engineered highway system can become congested, leading to snarled and slow traffic.

Computer systems are similar to highway systems, as they're designed for data to travel at high speeds by using more than one lane (bandwidth) for each direction of traffic. When all goes well, the data can easily make a smooth and fast trip from the point of origin to its destination. This is exactly the kind of I/O fabric we all try to deploy within our storage networks.

But even the best engineered and created storage area network (SAN) can lead to snarls and slow I/O traffic as the SAN grows and changes and congestion creates backpressure that begins to overwhelm the ability of the SAN to get data from one place to another easily and quickly. To ensure a low and consistent response time and that data flow within a storage network node, or between nodes, is fast, smooth and as uncongested as possible, SAN architects need to minimize path latency and control and balance I/O flow within a fabric and especially across interswitch links (ISLs). Latency, very simply, is I/O wait time.

Reducing I/O path latency to preserve an acceptable system response time has always been important but with the advent of solid state disk (SSD), reducing latency becomes mandatory to overall systems performance. Not only will SSD provide faster I/O response times, the system in general should also "feel" faster as well.

Traditional disk, available since the '50s, has rotating components and fantastically complex control mechanisms and operates thousands of times slower than other tiers of computer storage such as main memory and cache on the computer system. In terms of time, rotating disk is often characterized as having glacier-like speed compared to other system components. And that's even before the additional latency of the data path is taken into account.

Since SSD doesn't have rotating platters to contend with, each and every piece of data is available at the same, predictable, very high speed, which means read/write operations happen more quickly. Faster response time (lower latency) plus faster data transfers (more bandwidth) means that SSD can move more data faster, which means its throughput is higher.

Although much more important when used with SSD devices, data I/O paths to SSD and traditional disk need to run at the lowest latency possible to achieve the highest possible value from these resources. So, outside of the computer node and the SSD or disk device nodes, customers must understand the technology that connects host and storage to minimize wait times (latency) due to path design.

When a data path consists of a direct-attached link between host and storage, then the only potential delays are the speed of the optics (2GB, 4GB, 8GB, 16GB) and the speed of light as it traverses the cabling, which is approximately 5 microseconds (µs) per kilometer.

Since the vast majority of users today deploy switch-based fabrics upon which to place their I/O, there are a number of other considerations that will come into play.

Any I/O infrastructure must consider the optics being used and the speed of light through the cables. But uniquely, a switch allows one port to multiplex with many other ports, which is known as fan-in, fan-out. Minimizing path latency in a SAN begins with making sure that fan-in, fan-out ratios on a data path don't overwhelm the bandwidth capability of a link (over-subscription). For example, if six 4Gbps ports (e.g., storage) can all multiplex into a single 8Gbps port (e.g., channel path identifier [CHPID]), then it's possible that too much data can be simultaneously routed to that single port and overwhelm its bandwidth capabilities. This, of course, assumes that each port is being utilized at 100 percent, which usually isn't the case. The reason that fan-in, fan-out is so useful is that many I/O ports are typically underutilized and can be multiplexed in with other ports to a single source or target port to help drive its utilization at optimum levels. Each storage vendor has guidelines they will provide about optimizing fan-in, fan-out.

Switching devices are designed to use either "cut-through" or "store and forward" frame routing. Cut-through frame routing is a technology where the switch starts forwarding a frame along the path before the whole frame has been received, normally as soon as the destination address is processed. Functioning at near wire speed, Fibre Channel (FC) switches forward an I/O frame almost as fast as the FC switch receives the frame. Compared to store and forward frame routing, cut-through frame routing significantly reduces switch latency and, as per the Fibre Channel (FC) standards, relies on the destination devices for error handling. Store and forward frame routing requires the entire frame be gathered into each receiver before it can be moved to the next station within a switch. Each of those whole-frame moves takes additional time. So there's a significant difference in switch latency time between cut-through (2 to 3 µs) technologies and store and forward (8 to 100s µs) technologies.

Building Bridges for the I/O Super Highway

Potentially, the most significant obstacle to building an I/O super highway is the "bridge" that connects one switching device to another. If the bridge isn't well-architected, then I/O flow bottlenecks can occur. So, let's concentrate on good bridge building techniques that allow these links to be optimized for I/O traffic.

FC bridges are known as Inter Switch Links (ISLs), which are simply physical cable connections between two switching devices. The number of ISLs needed to build the bridge is a function of the amount of bandwidth that needs to traverse those links.

Potentially, the most significant obstacle to building an I/O super highway is the "bridge" that connects one switching device to another. If the bridge isn't well-architected, then I/O flow bottlenecks can occur.

The trick to optimizing these bridges to avoid congestion is how data gets placed onto them.

The FC protocol has a standard way of determining which "bridges" are available to be used by the fabrics. Avoiding all the technical jargon and hidden internal mechanisms, basically, as a port in an FC fabric logs into the fabric, it's told about the ISLs that are available to be used from point A to point B. This process is known as Fabric Shortest Path First (FSPF). It isn't FSPF's task to determine which ISL a given ingress port is assigned to. Assigning ISL paths is accomplished by a vendor firmware algorithm, such as port-based routing, exchange-based routing or device-based routing. These algorithms make an attempt to fairly allocate and share all ISLs with all the ports in the fabric.

A vendor can provide both hardware- and software-oriented capabilities to optimize ISL bridges. Hardware trunking allows circuitry within the switch to help optimize ISLs. Since the hardware is handling the I/O flow, it can make real-time decisions about how best to optimize those ISLs. Device-based routing (DBR) is a new and useful switch software-oriented capability to better distribute ingress ports across ISLs. Only when a source port actually requires the use of an ISL for data transport will an ISL be selected for it to use. Therefore, only ports that actively need to use an ISL will ever get an ISL assigned to them. This typically eliminates unused ISLs and helps reduce overutilized ISLs.

When data is sent across a fabric bridge via an ISL, that data can be encrypted and/or compressed. This provides greater security for the data that's in transit and reduces the size of the frame. This technology reduces the bandwidth used by the frame on the ISL, freeing it up to handle even more work, thus optimizing the ISL bandwidth.

Summary

Whether used together or separately, these technologies provide powerful tools for maximizing efficiency, optimizing performance and increasing reliability for FICON and open systems storage networks. They help you build your own FC fabric super highway! **ETJ**

David Lytle is a principal engineer/global solutions architect with Brocade Communications, Inc. in San Jose, CA. Based just outside of Atlanta, GA, he specializes in Gen 4 (8Gb) and Gen 5 (16Gb) Fibre Channel solutions and specifically in System z mainframe technologies, and provides sales enablement assistance to Brocade OEMs, partners and customers. David, a 45-year veteran of the IT industry, speaks at user conferences and is a co-creator and instructor for the Brocade Certified Architect for FICON (BCAF) professional certification class.
Email: david.lytle@brocade.com