The Promise of Unified I/O Fabrics

Two trends are challenging the conventional practice of using multiple, specialized I/O fabrics in the data center: server form factors are shrinking and enterprise applications are requiring more connections per server. However, the current practice of using multiple Ethernet and Fibre Channel connections on each server to support fully redundant, cluster-enabled computing environments inhibits scalability. Unified, high-performance I/O fabrics can enhance scalability by providing a single fault-tolerant connection. This approach allows legacy communication technologies to use one I/O "superfabric," and the reduction in physical connections can help achieve better performance, greater flexibility, and lower total cost of ownership—the primary benefits of the scalable enterprise.

BY J. CRAIG LOWERY, PH.D., AND DAVID SCHMIDT

The modern data center is a collection of server and storage components partitioned into various cooperating groups that communicate over specialized networks. In most cases, the technologies behind these networks were conceived decades ago to address particular kinds of traffic, such as user access, file transfer, and high-speed peripheral connections. Over time, data centers incrementally evolved to meet the increasing requirements of their environments, often retaining vestigial characteristics of repeatedly revamped technologies for backward compatibility and interoperability. Although the standardization and stability enabled by backward compatibility and interoperability have paved the way for the proliferation of computer systems, it is becoming increasingly difficult to extend these legacy technologies to meet the fundamentally different requirements imposed by the scalable enterprise.

For example, Ethernet—the de facto standard for local area network (LAN) communication—began as a rather cumbersome bus architecture with performance limitations imposed by the shared nature of its medium access control protocol. Today, Ethernet has become a much faster switched communication standard, evolving from 10 Mbps to 100 Mbps to 1 Gbps. Yet, the remnants of its past—the bus-based architecture and, in particular, the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol—introduce unnecessary overhead for the sake of compatibility, making Ethernet less attractive for protocols such as Remote Direct Memory Access (RDMA) than newer interconnects without the same historical baggage, such as InfiniBand.

Other interconnect technologies, such as SCSI, Peripheral Component Interconnect (PCI), and Fibre Channel,

have followed a similar trajectory. In each case, the technology was created to solve a particular problem and, over time, has been extended to increase both performance and scope of application.

An unfortunate side effect of the proliferation of multiple interconnect technologies is the requirement that they coexist in smaller and smaller spaces. High-density servers—such as rack-dense, 1U servers and blade servers—are required to provide user experiences equivalent to larger systems such as the traditional tower. At the same time, emerging enterprise applications for these high-density systems require an increasing number of connections, or large fan-outs. Some clustering systems, such as high-availability clusters and clustered databases, require multiple LAN connections and two Fibre Channel connections for a fault-tolerant storage area network (SAN). Fitting four or more of these connections into a blade server's form factor can be challenging.

Now that the barriers to mass adoption have been addressed, unified I/O fabrics are set to revolutionize computing infrastructures through their flexible, extensible architectures.

Another drawback of legacy interconnects is that they do not inherently encompass the fabric concept. A fabric functions much like a public utility: it is a multipurpose interconnect that is accessible from virtually anywhere. The vision of the scalable enterprise depends largely on fabric *semantics*—the model of communication that determines how enterprise applications "speak" within the data center that employs the fabric—because next-generation data centers will likely be built using standard, disposable components that plug in to the infrastructure as capacity is needed. Fabrics are the key to this plug-and-play data center. Although some technologies such as Ethernet come closer than others such as SCSI to delivering a fabric-like usage semantic, they still fall short in key areas, primarily by requiring additional unnecessary overhead to support their legacy aspects. For example, using TCP/IP over Ethernet to perform RDMA significantly wastes bandwidth and is unnecessarily slow for a high-performance computing cluster rack interconnect, because TCP's sliding window protocol was designed for the unreliable Internet—not a single, well-controlled rack with a high-speed communication link.

Heterogeneous legacy interconnects are also hindered by the support structure required to maintain data centers that incorporate them. Today, IT support teams must staff skills in each interconnect technology. This redundancy is inefficient when compared to the single-culture support required to maintain a unified fabric. A unified fabric subsumes all communication functions through one fabric connection or—for redundancy—two fabric connections. The fan-out problem can be resolved at the software level by multiplexing multiple virtual interfaces over the single physical interface of a unified fabric. Some of these virtual interfaces may be designed to appear to higher layers of software as legacy technology interfaces to help provide transparency and backward compatibility.

As the deficiencies of heterogeneous interconnects in the scalable enterprise intensify and the need for fabric semantics mounts, a clear gap arises that cannot be adequately filled by additional iterations to refine older technologies or make them suitable and relevant going forward. It is this need that the unified I/O fabric is designed to address.

Understanding the requirements of unified I/O fabrics

Any technology candidate that puts itself forward as being a unified I/O fabric technology must meet the following suitability requirements:

- **Transparent coexistence:** The fabric must be able to coexist and interoperate with legacy interconnect technologies without placing additional requirements on end users and applications.
- **High performance:** The fabric must be able to accommodate the aggregate I/O that would otherwise have been distributed across legacy interconnects. Nonblocking connectivity, throughput, and latency should be optimized so that the performance of the unified fabric is the same as or better than the performance of multiple legacy networks.
- **Fault tolerance:** The fabric must respond gracefully to component failures at both the legacy interconnect layer and its own unified layer. Furthermore, to meet the requirement of transparent coexistence, the fabric should support legacy fault-tolerance technologies.
- **Standardization:** The fabric must conform to industry standards to ensure competitive pricing, multiple sources for components, longevity of the technology, and the creation of an attendant ecosystem. *Ecosystem* refers to all the companies, services, and ancillary products that must come into existence to make the technology viable and deployable on a large scale.
- **Value pricing:** The fabric must be less expensive to procure and maintain than an equivalent combination of legacy interconnects.

Unified I/O interconnects are not a new idea—proprietary solutions have been developed and deployed with some success in targeted, custom environments. However, most efforts to date have not met all of the preceding requirements, usually failing on transparency, standardization, and pricing. Recently,

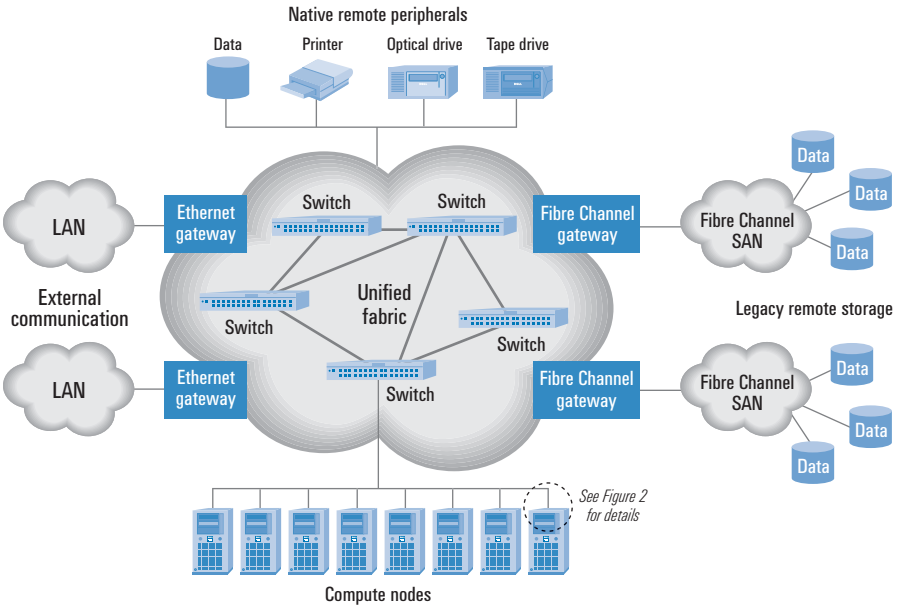

Figure 1. Unified fabric architecture

technologies like InfiniBand have been specifically designed to meet all of these requirements. Now that the barriers to mass adoption have been addressed, unified I/O fabrics are set to revolutionize computing infrastructures through their flexible, extensible architectures.

Examining the unified I/O fabric architecture

Figure 1 shows an overview of a unified I/O fabric architecture. At the center of the figure is the unified fabric, comprising one or more switches. The specific technology used in the switch is not of particular importance to the concept, although InfiniBand is one currently available candidate. Ethernet gateways allow for IP traffic between devices connected to the fabric and external networks. Fibre Channel gateways provide similar connectivity to SANs. Various remote peripherals that have native fabric interfaces are shown at the top of the figure. Such devices can communicate directly on the fabric and do not require a gateway.

The lower portion of Figure 1 shows an array of compute nodes. Although these are depicted as identical in the figure (as would be the case with blade servers), the nodes in the array can consist of various form factors and system models. Each compute node must have at least one fabric interface with which to connect, and each node must host an operating system with a software stack that consists of a native fabric driver for the physical fabric interface, as shown in Figure 2. In addition to transmitting and receiving data, this driver may incorporate or be bundled with additional software to aid in the distributed management of the fabric.

Operating systems on the compute node can communicate over the fabric either by using the native fabric protocols or by using mapped or tunneled legacy protocols, as shown in Figure 3. A mapped protocol is one that can be translated to and from the fabric protocol and requires that the fabric protocol directly support similar functionality. When no direct mapping exists, a protocol must be tunneled through the fabric, meaning that the legacy protocol's messages are embedded, or *wrapped,* in the fabric protocol for transport across the fabric. Mapping is usually more efficient because the fabric comprehends the mapped protocol and can be optimized for it. Both mapped and tunneled protocols require a gateway to connect the fabric to the legacy networks and perform the mapping and tunneling functions.

For example, the InfiniBand specification incorporates IP over InfiniBand (IPoIB), which allows IP datagrams to be mapped directly to InfiniBand packets for transport and routing over the fabric. InfiniBand also is designed to provide a standard mechanism for mapping SCSI device commands to the fabric, either directly to a SCSI device attached on the fabric or to a Fibre Channel gateway.

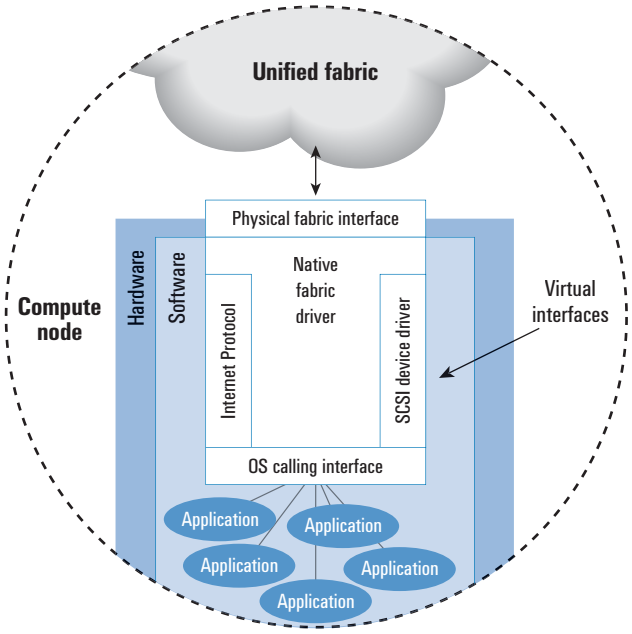

Figure 2. Compute node communication stack

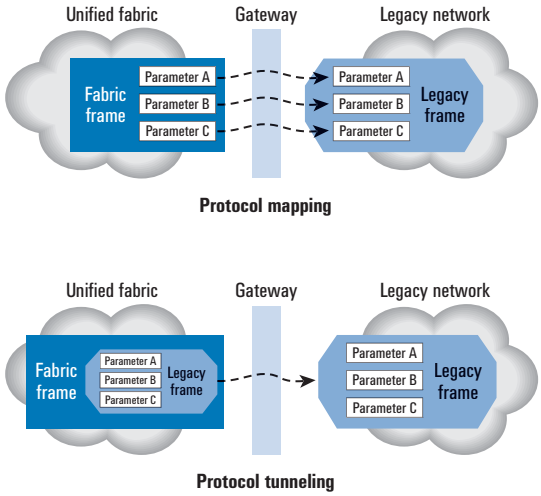

Figure 3. Protocol mapping and tunneling

Considering InfiniBand for the I/O fabric

InfiniBand, designed as a next-generation I/O interconnect for internal server devices, is one viable option for unified I/O fabrics. This switched, point-to-point architecture specifies several types of fabric connections, allowing compute nodes, I/O controllers, peripherals, and traffic management elements to communicate over a single high-speed fabric. Because it can accommodate several types of communication models, like RDMA and channel I/O, InfiniBand is a compelling candidate for unifying legacy communication technologies. InfiniBand meets most of the suitability requirements for a unified I/O fabric primarily because it was conceived and designed to function as a single, ubiquitous interconnect fabric.[1]

Transparent coexistence between legacy technologies and unified fabrics is a crucial requirement for fabric technology. InfiniBand is designed to support connections to legacy networks via target channel adapters (TCAs). TCAs enable legacy I/O protocols to use InfiniBand networks and vice versa. They support both mapping and tunneling operations like those shown in Figure 3. Tunneling requires encapsulating the legacy transport and payload packets within the InfiniBand architecture packet on the unified fabric, allowing the legacy protocol software at each end of the connection to remain unaltered. For example, a TCP/IP connection between a legacy network node and a unified fabric node can take place entirely over the TCP/IP stack. The unified fabric node needs the InfiniBand software only to remove the TCP/IP packet from the InfiniBand packet. Mapping, or *transport offloading*, actually removes the IP information from the InfiniBand packet and transmits it natively on the IP network. The unified fabric node can take full advantage of the InfiniBand transport features, but the TCA must translate between InfiniBand packets and IP packets—a potential bottleneck. Although InfiniBand was designed with transparency in mind, the performance and benefits of these methods will ultimately depend on the design and implementation of TCA solutions.

The high bandwidth of InfiniBand makes it possible for several legacy protocols to coexist on the same unified connection. InfiniBand currently supports multiple levels of bandwidth. Single-link widths, also called 1X widths, support a bidirectional transfer rate of up to 500 MB/sec. Two other link widths, 4X and 12X, support rates of up to 2 GB/sec and 6 GB/sec, respectively. Compared with today's speeds of up to 125 MB/sec (1 Gbps) for Ethernet and 250 MB/sec (2 Gbps) for Fibre Channel connections, InfiniBand bandwidth has enough capacity to support both types of traffic through a single connection. However, as legacy network bandwidth improves, unified fabric solutions must scale to meet the enhanced speeds of traditional I/O networks. InfiniBand must increase its bandwidth or unified I/O solutions must use multiple InfiniBand connections to garner a higher aggregate bandwidth.

As with other packet-based protocols, InfiniBand employs safety measures to help ensure proper disassembly, transmission, and reassembly of transported data. This can be considered fault tolerance in its most basic form. Furthermore, just as other protocols specify fault tolerance between connection ports, InfiniBand allows for two ports on the same channel adapter to be used in a fault-tolerant configuration. Therefore, InfiniBand can help enable a unified fabric solution to implement fault tolerance and failover features. Fault tolerance at the legacy technology layer, however, must still be provided by legacy software stacks. If a TCA is used to connect a legacy network to the InfiniBand fabric, the unified fabric solution must provide fault-tolerant links between the two. Again, the performance and benefits of internetwork fault tolerance will depend on the design and implementation of unified fabric solutions.

InfiniBand meets most of the suitability requirements for a unified I/O fabric primarily because it was conceived and designed to function as a single, ubiquitous interconnect fabric.

InfiniBand was designed with a goal of multiple physical computing devices existing on the same network. Different physical connections might be necessary for different elements of the unified network, and different I/O access models might be

[1] For more information, see www.infinibandta.org/ibta.

required for different legacy protocols. Fortunately, the InfiniBand Trade Association (IBTA) has set forth several options for connectivity. Physical connections can use copper or optical interconnects, and connection modules can use channel-based I/O protocols for legacy operations or zero-copy RDMA to reduce latency and CPU overhead. As a variety of unified fabric solutions becomes available, the IBTA plans to help ensure a standardized playing field that will allow interoperability with third-party management and software solutions.

Because the high-performance characteristics of InfiniBand allow for the transport of multiple legacy networks on a single connection, unified InfiniBand fabric solutions can help reduce the number of physical ports.

Because the high-performance characteristics of InfiniBand allow for the transport of multiple legacy networks on a single connection, unified InfiniBand fabric solutions can help reduce the number of physical ports. This reduction can simplify the physical view of the data center configuration, and result in a lower risk of physical port failure and a faster configuration time—which can translate to lower costs for network management. IBTA specifies standards for managing subnets within the InfiniBand network, providing network management solutions with a method of monitoring and controlling InfiniBand network configuration and performance. By utilizing these management standards, organizations can realize the value of unified I/O fabrics with InfiniBand.

Migrating high-speed networks to unified I/O architectures

As other I/O technologies evolve, unified I/O architectures will have more options for the underlying high-speed fabric. The advent of TCP/IP Offload Engines (TOEs) allows traditional Ethernet fabrics to utilize RDMA and hardware-based TCP/IP stacks, thereby reducing CPU overhead. Storage technologies such as Internet SCSI (iSCSI) can then efficiently utilize Ethernet as the unified fabric. InfiniBand, however, is a leading choice for immediate adoption of a unified fabric. Even though some InfiniBand components may not be widely available or competitively priced with legacy networks, the technology itself is viable and proven. Based on previous technology rollouts, mass adoption of InfiniBand over time can help increase its availability and lower its cost. The transparency benefits and legacy support built into this architecture can help drive rapid adoption within data centers, and InfiniBand's high-capacity switched architecture can provide reliable performance for multiple I/O models.

Solutions that utilize the high performance and simplified management of unified I/O are already entering the marketplace, and network architects are planning for data centers in which servers can be dynamically matched with I/O resources via a unified high-performance fabric. With the maturation of high-speed network architectures like InfiniBand, the benefits of unified I/O fabrics cannot be denied. It is only a matter of time before these fabrics become a requirement for the scalable enterprise.

J. Craig Lowery, Ph.D., is a senior engineering development manager on the Enterprise Solutions Engineering team within the Dell Product Group. His team is currently responsible for developing products that realize the Dell vision of the scalable enterprise. Craig has an M.S. and a Ph.D. in Computer Science from Vanderbilt University and a B.S. in Computing Science and Mathematics from Mississippi College.

David Schmidt is a systems engineer on the Enterprise Solutions Engineering team within the Dell Product Group, where he develops solutions for the scalable enterprise. Previously, David worked in the systems management group developing the Dell OpenManage™ Deployment Toolkit. David has a B.S. in Computer Engineering from Texas A&M University.

FOR MORE INFORMATION

InfiniBand Trade Association:
www.infinibandta.org