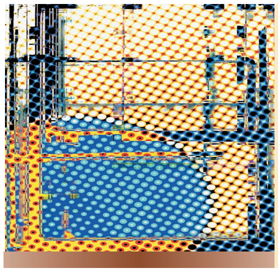
Switch Fabric Interfaces

Itamar Elhanany, University of Tennessee at Knoxville
Kurt Busch, TeraCross
Derek Chiou, Avici Systems

Switch fabrics are fundamental building blocks in a wide range of communication platforms. However, despite the growing need for next-generation switches and routers, semiconductor vendors have been slow to develop switch fabric chipsets. In addition to the many technical challenges associated with the deployment of such fabrics, industry analysts agree that a key factor impeding wide-scale exploitation is a lack of standardization in interconnecting fabric components.

Many of the more than 30 companies that develop switch fabrics offer excellent price and performance, but none guarantee compatibility with other vendor offerings or even with future generations of their own products. The market consists of several point solutions with no reliable and coherent roadmap.

In these uncertain times, assurance of supply is a major issue in selecting a silicon vendor. Standard interfaces will make it possible to replace a discontinued device without requiring a new system design. The guaranteed availability of backup products will reduce the risks associated with each device and let systems designers select newer and more cutting-edge offerings.



Standardizing the interfaces connecting line cards with switch fabrics will facilitate innovation in communication systems.

SWITCH FABRICS

A switch fabric moves incoming data from a set of ingress ports to a single egress port, in the case of unicast devices, or multiple egress ports, in the case of multicast devices. In applications such as video switching, the binding between an ingress and egress port changes infrequently. In IP routers and asynchronous transfer mode switches, however, such fabrics dynamically partition data into fixed-sized or variable-sized cells, frames, packets, and other units. Dynamic fabrics tend to be more complex than static fabrics because they require arbitration between data units that may be simultaneously destined for the same output port.

Much switch fabric research in recent years has focused on Internet packet switching. Due to advances in emerging networking applications, however, the need for high-performance switch fabric solutions has shifted from the pure IP domain. Robust metropolitan area networks, director-class storage area networks, and other emerging switching applications have reached both the capacities and the service requirements to justify the need for advanced off-the-shelf fabric products.

Figure 1 presents a generic overview of a contemporary high-capacity fabric architecture. The *ingress path* connects a line card's network processing subsystem—consisting of network processors and/or traffic managers—to the *switching core*, which dynamically connects ingress ports to egress ports. The *egress path* aggregates traffic from the switching core and forwards it to the line card front end.

Switch fabrics can be implemented in various ways, from shared-memory architectures to fully distributed multistage designs. Regardless of the architecture, however, most are input buffered—they queue incoming cells or packets at the ingress stage until a scheduling mechanism signals them to traverse the switch core. In many implementations, the buffering occurs at the line card, and the switch cards contain little, if any, memory. In addition to payload data, control information also flows from the line cards to the switch fabric.

COMMON SWITCH INTERFACE

Many switch fabrics today support the common switch interface (CSIX), a standard developed by the Network Processing Forum (www.npforum.org), formerly the CSIX Consortium. The CFrame datagrams that traverse CSIX devices are fixed in size. Each CFrame consists of a header, which contains all relevant management information including routing and priority status, and a payload. In the segmentation and reassembly (SAR) process, the switch segments packets arriving from various sources into CFrames and later reassembles them as they depart.

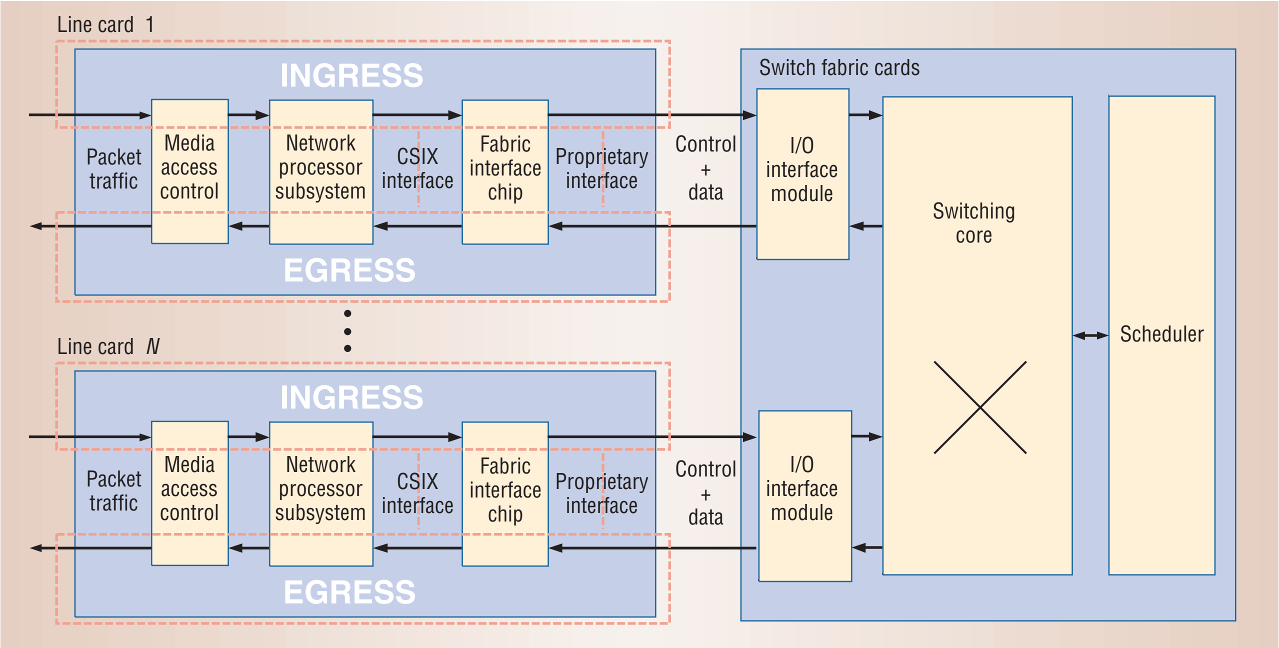


Figure 1. Common switch fabric architecture. The architecture allows multiple input ports to transmit data packets to multiple output ports simultaneously.

A fabric interface chip (FIC) resides on each line card and interfaces to either a network processor or a traffic manager; the fabric boundaries thus logically encompass part of the line cards.

While the CSIX standard is invaluable in guaranteeing interoperability between different networking devices—namely switch fabrics and traffic managers/network processors—it intentionally does not specify the interface between the FIC and the fabric cards. Consequently, a network equipment manufacturer that designs a switch fabric chipset is confined to the developer's proprietary interface. Changing interfaces would require redesigning high-speed boards that comprise the line cards and fabric, an intricate task at best.

RECENT EFFORTS

Researchers are exploring various ways to standardize the FIC-to-fabric interface. The three primary technologies currently driving the effort are PCI Express Advanced Switching (PCI ExAS), Rapid IO, and Ethernet.

PCI ExAS

The PCI Express physical layer (www.us.design-reuse.com/articles/article5306.html) is developed around a building block that includes two point-to-point unidirectional paths, each of which consists of a low-voltage differential signaling (LVDS) pair operating at 2.5 GHz. This yields an effective bandwidth of 2 Gbps in each direction, or a full duplex bandwidth of 4 Gbps. Lanes can be combined to provide even higher bandwidth, as necessary.

The PCI ExAS architecture is at its root a serialized, packetized version of PCI. It incorporates features such as inherent multicast capabilities and error detection and correction at the protocol level to eliminate the need to implement them elsewhere.

PCI ExAS's primary goal is to encapsulate all the information traversing between the different line cards and the fabric core over a standardized interface that is defined both on the physical and logical layers. Such unification will allow different vendors to add their variations, while still adhering to an industry standard.

In addition, PCI ExAS's ability to handle packet traffic eliminates the need for an external SAR device, thereby greatly reducing system cost.

Rapid IO

The RapidIO initiative (www.rapidio.org) is attempting to address the same problem in similar ways. Like PCI ExAS, RapidIO is a layered architecture; it also uses LVDS signaling with up to 1 GHz on both edges of the clock. RapidIO includes specifications for both parallel and serial signaling and provides error recovery and reporting.

The serial interface particularly appeals to switch designers and uses 3.125-GHz signals with 8b/10b encoding for an effective bandwidth of 2.5 Gbps per line. Multiple lines can be used in parallel for greater bandwidth per connection.

Unlike PCI ExAS, Rapid IO is a more traditional cell-based architecture that requires the use of an external SAR device.

Ethernet

Some vendors are promoting standard 1-Gbps and 10-Gbps Ethernet as an interconnection standard. The clear advantage it offers is availability of a wide variety of inexpensive silicon building-block devices. However, because Ethernet was not inherently designed for this function, it will likely be most useful in areas where cost rather than quality of service, latency, or any other performance-centric parameter is the dominant decision-making factor.

A version of the CSIX streaming interface over high-speed serial links, preferably 2.5 Gbps and above, offers an alternative path toward efficient convergence of fabric-related interconnect technologies. Such an interface would allow using a CSIX-based interface standard to connect FIC devices to fabric cards as well as to connect traffic managers and network processors to the FIC. In fact, such an interface would let the FIC reside on the fabric card or even be integrated into the switch fabric. ■

***Itamar Elhanany** is an assistant professor in the Department of Electrical and Computer Engineering at the University of Tennessee at Knoxville. Contact him at itamar@ieee.org.*

***Kurt Busch** is vice president of marketing at TeraCross, Inc., based in Campbell, CA. Contact him at kurt.busch@teracross.com.*

***Derek Chiou** is a principal engineer at Avici Systems Inc., based in North Billerica, Mass. Contact him at dchiou@avici.com.*

Editor: Gary Robinson, 85 East India Row, Apt. 18C, Boston, MA 02110; robinson_gary@emc.com