



SWITCH FABRICS

Key Building Blocks for Switches and Routers

© DIGITALVISION

The need to move data at super-high speeds across the Internet with minimal delay has led to the development of ultra-fast switching structures capable of switching millions of packets or terabytes of information per second. Switch fabrics play a critical role in networking equipment, such as routers and switches, as they can determine the system's overall capabilities, performance, speed, cost, and lifespan. Thus, they are undoubtedly considered the very heart of the networking system. Switch fabrics are suited for networking equipment and applications, ranging from access networks to metro and core systems and storage switches. A switch fabric has two main functions: to provide connectivity for transferring data between ingress (incoming) and egress (outgoing) ports and to satisfy and support (and enforce) any quality-of-service (QoS) demands. A number of different switching architectures and a variety of technologies exist, offering a wide array of choices for the industry.

There is a plethora of different high-speed switching architectures based on the numerous router and switching systems (optical or electronic) that exist on the market today, often with quite distinctive characteristics. Switching fabrics are

influenced, and, thus, widely differentiated, mainly by the following factors: mode (and configuration) of interconnection between input and output ports, location and amount of memory buffering, internal switching speed, resource contention resolution, performance guarantees, scalability, design and implementation complexity, fault tolerance, management, and control functionality. In this article, we present and discuss a number of the most prominent switching architectures and techniques.

*Christos Koliass
and Ioannis Tomkos*

SWITCHING SYSTEMS AND TECHNIQUES

A size $N \times N$ switch fabric has N inputs and N outputs, with an input port and an output port sharing the same link; thus, input- and output-line speeds are assumed to be identical. Switching architectures are characterized as packet based (switching variable-length packets) or cell based (switching fixed-size packets, called cells). Photonic packet switching employs similar architectures as electronic packet switching, providing circuit-switching-like functionality (semi-permanent connections) at such speeds as 10 Gb/s and higher. A cell-based architecture operates in a completely synchronous

(timeslotted) mode and can present certain advantages with respect to operation and performance issues: guaranteed delay, optimized buffer utilization, and less complexity. We will use the term “packet” to indicate either a variable-length packet or a fixed-size cell.

Based on the interconnection mode (and the buffering strategies implemented), a number of different switching architectures and systems have been proposed to date. There exist the following main types of switching fabrics, according to the mode of interconnect between the input and the output ports of a switch.

Shared Medium (Bus)

Using a shared medium, such as a bus or a ring, to transfer packets from the switch’s inputs to its outputs places a limitation on the switch’s performance since, in essence, this is determined by the available bus bandwidth. As in any bus technology, multicasting and broadcasting become straightforward.

Shared Memory

In memory-based fabrics, shared memory can be accessed concurrently by the inputs and outputs of the fabric. Shared memory fabrics take full advantage of the statistical sharing of buffers, which is beneficial, for instance, in absorbing large traffic bursts as opposed to a (physical) fixed-buffer allocation. Thus, they provide ideal performance (no blocking or scheduling conflicts) and are very simple to build but are limited by the memory access speed achievable (memory bandwidth available). Hence, they are not considered to be very scalable. A typical shared-memory switch fabric is 16×16 .

Crossbar

In crossbar-based architectures, switching is performed in a matrix-like fashion between the inputs and the outputs. Crossbar switches are considerably popular for implementation in routers and switches due to their design simplicity, cost effectiveness, and certain scalability features. Several crossbar-based switching architectures have been proposed in the literature and quite a few are commercially available.

Multistage Interconnection Networks

Multistage fabrics are modular architectures built from smaller switching components. These larger fabrics can support more capacity, thus effectively addressing the scalability issue (e.g., overcoming port and capacity limitations).

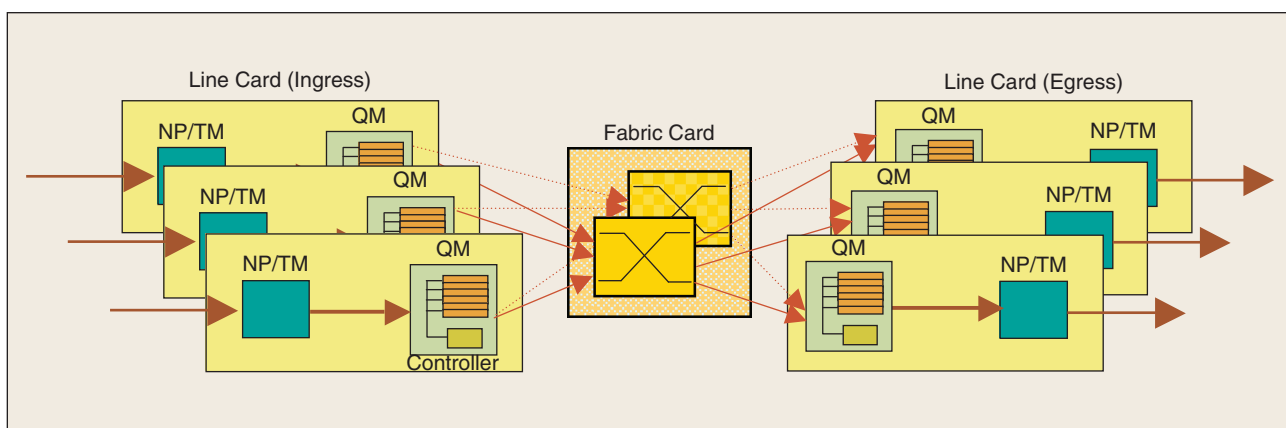
With respect to the packet storage location in a switch, memory buffers can be situated at the ingress side (input queueing), the egress side (output queueing), inside the switch fabric itself (central queueing; e.g., a single memory is shared by all input and output ports), or a combination of the above. Architectures that implement input queueing suffer from a blocking phenomenon known as head-of-line (HOL). HOL blocking occurs when more than one packet, at the front of their respective input queues, attempts to access the same output.

Although the switching fabric and any associated buffers can be placed on the same board or card (as in a single box router or switch), a large switching system configuration may include the queues [e.g., queue manager (QM)] along with any packet processing components, such as a network processor (NP) and/or a traffic manager (TM), placed on a system line-card while the switching element itself is situated on a separate fabric card, as depicted in Figure 1. (Note that a multicard switching system, such as the one shown in Figure 1, is a typical configuration for core and metro network routers and switches.) The fabric card could also contain additional buffering, depending on the architecture implemented. Line cards and fabric cards can be interconnected through a high-speed backplane.

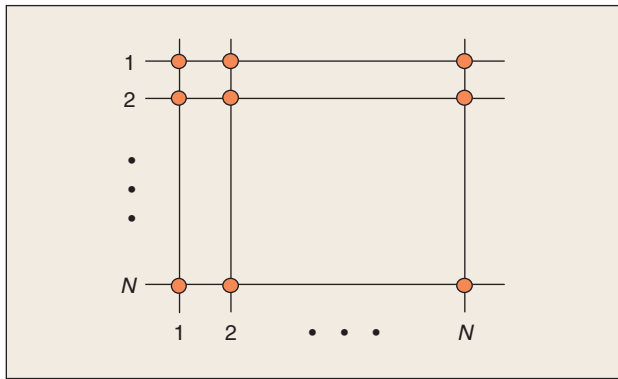
The switch fabric combined with the packet interfaces (buffers) and any processing and control capabilities, i.e., routing, synchronization control, form a routing (switching) node in a packet-switched network. As indicated by the anatomy of a large switching system, as depicted in Figure 1, we often refer to the switching fabric core, which provides the multichannel line-card interconnection, as the system’s backplane.

SWITCH FABRIC ARCHITECTURES

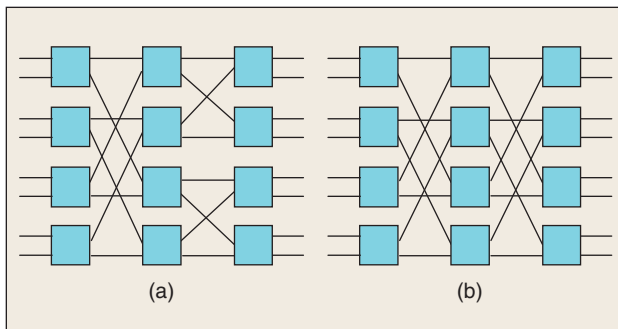
Switch fabrics can be single stage, where packets are switched in a single step between the fabric inputs and output, or multi-stage, where the fabrics are constructed by



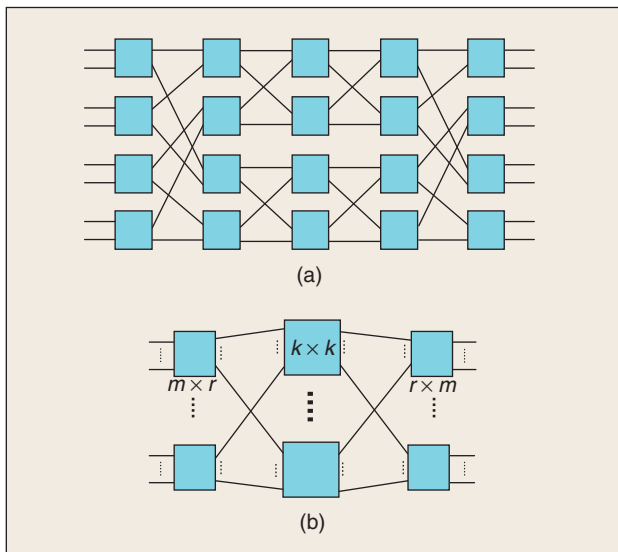
1. A generic switching system architecture.



2. A crossbar.



3. Multistage switches: (a) Banyan switch and (b) Omega switch.



4. Multipath multistage networks: (a) Benes switch and (b) Clos switch.

cascading several single-stage switching modules, called “switching elements” (SEs). Multistage switches are space-division architectures, i.e., switching is performed by interconnecting an input to an output line. Multistage interconnection networks (MINs) constitute a large and special class of packet switches. They have been used extensively in various other fields, such as telephone networks and multiprocessor computer systems, i.e., interconnecting processors to memory modules. Multistage networks are readily and easily scalable, as they are inherently built in a modular fashion,

since they are realized using smaller SEs, and offer much less complexity than the crossbar fabric.

The simplest form (in terms of design) of an interconnection network is the crossbar fabric. A crossbar is based on a single-stage network of crosspoints (intersections) forming a fully interconnected grid or switching matrix, as shown in Figure 2. The crosspoints are implemented using semiconductor switches that activate a path between an input and an output port. It is evident that, since the number of crosspoints grows with the square of the switch’s port-size, crossbar architectures present certain limitations. A crossbar fabric that alternatively includes a buffer at each crosspoint to deal with potential internal blocking is referred to as buffered crossbar.

The SE, which constitutes the basic functional block for an interconnection network, is typically small in port-size (e.g., 2×2) and is capable of switching packets autonomously (no centralized control) and independently, thus introducing a very high degree of parallelism. If SEs have no buffering capability, we then refer to unbuffered interconnection networks. An $N \times N$ network built from $d \times d$ SEs requires a minimum of $n = \log_d N$ stages. Figure 3 depicts examples of multistage switching fabrics based on three-stage interconnection networks ($N = 8$) that are built from 2×2 SEs. A multistage switch can be naturally augmented (in port count) if we consider additional stages.

Depending on how the SEs across the stages of an interconnection network are linked, different types (topologies) of multistage networks can be produced, as shown in Figure 3(a) and (b), where a Banyan and an Omega network are shown, respectively. If a single path exists between an input and an output of a multistage network, packets can be self-routed across disjoint paths (no sharing of interstage links) that are established prior to switching the packets presented to the network. However, Banyan-type networks (such as those depicted in Figure 3) can suffer from blocking due to internal congestion [i.e., a conflict occurs when two separate connections request to access (share) the same interstage link].

Interconnection networks can be built such that multiple paths exist for any input–output pair, addressing the problem of internal congestion. A Banyan-type network can easily become a multipath network if it is augmented by adding m extra stages, where $m \leq n - 1$. An extended Banyan network, known as a Benes network, is illustrated in Figure 4(a). A Benes network has $2\log_2 N - 1$ stages, where each stage holds $N/2$ (2×2) SEs. Another special class of multipath network is the Clos switch, which is a three-stage network. A Clos switch, shown in Figure 4(b), can be built from k ($m \times r$) switches in the first and third stage and from r ($k \times k$) switches in the middle stage, with $N = mk$.

An interesting multipath architecture is the Spanke-Benes switch, illustrated in Figure 5, which is viewed as a compromise between a Benes and a crossbar switch. An $N \times N$ Spanke-Benes switch requires N stages and can be realized by $N(N - 1)/2$ SEs (2×2). In this type of fabric, not all paths have exactly the same length, and a path can cross between $N/2$ and N SEs.

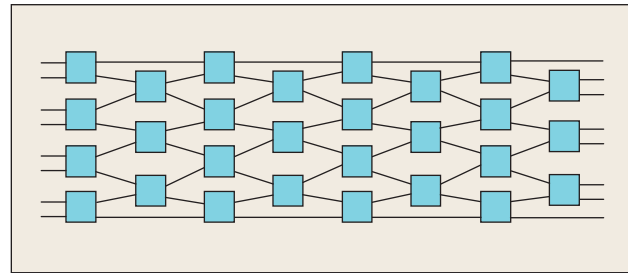
Figure 6 presents a multistage implementation of a crossbar architecture, where a 4×4 crossbar is built using a number of 2×2 SEs. In general, an $N \times N$ multistage crossbar can be realized with N^2 SEs, 2×2 each. We note that this type of network can provide multiple paths for certain input-output connections, while the path length can vary between 1 and $2N - 1$ SEs. Such crossbar-based networks are useful in building large optical crossbar switch fabrics, which are considered scalable due to their regularity, as opposed to the crossbar-type depicted in Figure 2, whose port size is constrained due to pin count, space, and power dissipation requirements.

A switching architecture called the Spanke switch, which uses N SEs (each $1 \times N$), is exhibited in Figure 7. Its internal links operate at $1/N$ th the external link speed. Although not scalable, the Spanke switch is useful for optical switches since any loss is minimized due to the small number of SEs (only two) crossed in any path.

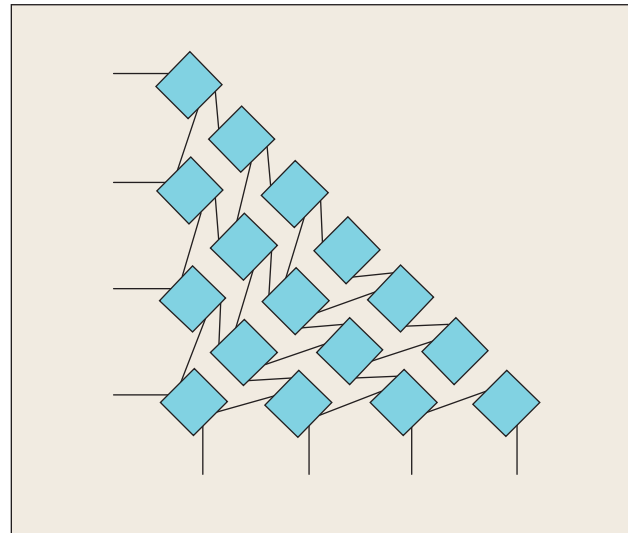
The interconnection networks we have examined can be classified as either blocking or nonblocking. A network is blocking if no connection between an arbitrary pair of unused (free) ingress and egress ports can be established. As we indicated, Banyan-type networks are blocking. Certain multipath networks, such as the Clos networks, can always provide disjoint paths for transferring packets from the switch's ingress ports to egress ports; thus, they are viewed as "strict-sense" nonblocking networks. [Note that a certain condition must exist for a Clos network to be strictly nonblocking; referring to Figure 4(b), it must hold $r \geq 2m - 1$.] Crossbar networks are characterized as "wide-sense" nonblocking since any permutation (interconnection pattern) between the inputs and outputs is possible. Wide-sense architectures usually assume a suitable mechanism or control for assigning the connections. For instance, an arbitration policy can determine which connections are to be established between the switch's ingress and egress ports, taking the maximization of the switch's performance into consideration.

Finally, in some networks, existing path connections can be rearranged, if necessary, to accommodate a new interconnection request. Such networks are characterized as "rearrangeable" nonblocking; e.g., the Benes and the Spanke-Benes switches. In the case of rearrangeable networks it is possible that reorganizing connections is not acceptable as certain applications cannot be interrupted.

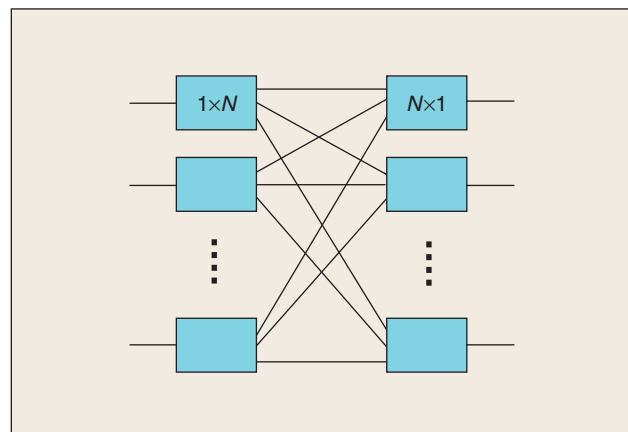
If the switch fabric itself operates at the same speed as the input and output lines of the switch, the fabric is characterized as having no speed-up ($S = 1$), and no output queueing is required. However, if the fabric can operate at a multiple rate of the external links, it can move packets faster from the switch's inputs to its outputs ($S > 1$). In particular, for a speed-up S , up to S packets from a single input can be delivered to the fabric's outputs within the same (arbitration/polling) round. When speed-up is present, output queues are then required to store the switched packets. Internal speed-up drastically improves the switch's performance and can also compensate for overhead (processing or other) introduced by any internal (transport) protocols.



5. The Spanke-Benes switch.

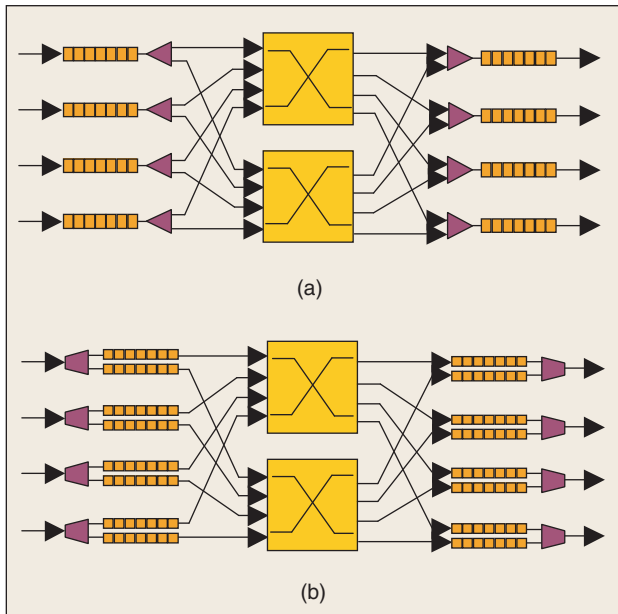


6. A 4×4 crossbar switch implemented with 2×2 SEs.



7. The Spanke switch.

However, achieving a very high speed-up rate is complex and expensive to implement. In [1], the first author has shown that a speed-up of two ($S = 2$), combined with the effect of multiple queues on the switch's inputs and without implementing any particular scheduling mechanism, is sufficient to yield a performance equivalent to that of an ideal switch fabric that employs $S = N$. It is worth mentioning that, indeed, many of the industrial switch fabric implementations today use a speed-up of two (or less than two, if switching variable-length packets).



8. Multiplane switch fabrics with buffer queues.

Multiplane fabrics, where switching fabrics operate in parallel, are offered as alternatives to implementing speed-up solutions. [Note that we can view multistage fabrics as two-dimensional (2-D) switching structures in contrast to multiplane fabrics (with parallel switching planes), which can be regarded as three-dimensional (3-D).] In fact, we can view multiplane switches as applying speed-up in space compared to speed-up in time when internal speed-up is used. Moreover, multiplane switching architectures are able to address the scalability and reliability issues; hence, they appear quite attractive as solutions. Depending on the way the switch's inputs and outputs access the parallel crossconnects, different types of multiplane fabrics exist, as displayed in Figure 8. In Figure 8(a), a single first-in, first-out (FIFO) buffer queue feeds all switching planes, while in Figure 8(b), a separate, dedicated queue exists per plane. Multiplane architectures, due to their inherent significant parallelism, can achieve load balancing (traffic is split across the various switching planes).

Multiplane configurations also provide for rapid expansion, i.e., one can add a new line card should the system be able to support it, whereas the switching and processing portions of the system are clearly separated; thus, either can be easily upgraded in the future.

PERFORMANCE ISSUES

The performance of a switch fabric, measured in terms of overall throughput, delay, data loss guarantees, and QoS support, is fundamentally critical to the rest of the system and, hence, to any high-level applications it can support. Requirements can vary considerably; for instance, the delay parameter is crucial to time-sensitive applications, such as real-time packetized voice transfer, while packet loss is of immense significance to data-oriented applications, such as Web applications and e-mail.

Data loss in a network has an adverse impact on the overall performance; hence, such loss inside a switch fabric should be avoided whatsoever. Loss in a fabric can be due to the insufficiency or absence of buffer space (e.g., unbuffered SEs). It is not always possible to provide plentiful memory given that memory modules are expensive. Special flow control techniques, such as backpressure (signal that propagates backwards, which disallows the receipt of additional packets), or routing techniques, such as deflection (e.g., selection of alternate paths), can be applied to prevent the fabric from discarding packets that are in excess. Switch performance can be used as an evaluation and comparison tool when considering different switching systems.

Another important factor, especially for large switches, is the complexity of the fabric and, hence, the cost associated with building it. Using the total number of SEs or, more precisely, the total number of crosspoints in a switching fabric, as a cost index, we can define cost C as a function of the switch's size N , referring to an $N \times N$ switching fabric. For instance, a simple Banyan network, which is comprised of $\log_2 N$ stages, where each stage has $N/2$ (2×2) SEs, has a cost $C = \log_2 N \times N/2 \times (2 \times 2) = 2N \log_2 N$. This cost metric or switch count, while not a comprehensive cost tool, can be used as an index. Based on this cost function, we can produce Table 1, which illustrates a comparison of the switching fabrics and their associated costs for various switch sizes N .

For this table, we assume 2×2 SEs, which, as crossbars, account for four crosspoints; the Spanke switch uses $1 \times N$ switches. For the Clos switch, we assume $m = 2$, $r = 2m - 1 = 3$, and $k = N/m = N/2$, which yields a minimum-cost, strictly nonblocking Clos network. We note that the Benes switch employs far fewer SEs than the other nonblocking architectures.

Fault tolerance (protection) is another vital performance parameter that affects not only the switch's resilience against

Table 1. Switch fabrics and their costs.

Switch Fabric	Type	Stages	Cost	Cost, $N = 8$	Cost, $N = 32$	Cost, $N = 64$
Crossbar	Wide-sense	1	N^2	64	1,024	4,096
Banyan	Blocking	$\log_2 N$	$2N \log_2 N$	48	320	768
Spanke	Strict-sense	2	$2N^2$	128	2,048	8,192
Clos	Strict-sense	3	$4\sqrt{2}N^{3/2} - 4N$	96	896	2,640
Benes	Rearrangeable	$2\log_2 N - 1$	$4N \log_2 N - 2N$	80	576	1,408
Spanke-Benes	Rearrangeable	N	$2N(N - 1)$	112	1,984	8,064

failures, but also the network's survivability, and availability, as a whole. A critical component such as a switch fabric must be protected in the face of operation disruption to ensure the flow of traffic across the switch. (Note that discarding traffic should be avoided at all costs; this is a typical requirement of a service provider or network operator.) It is also required that service should be restored rapidly. Protection in switching fabrics is achieved and implemented in the form of redundancy, i.e., backup fabrics in a multiplane switch configuration. Specifically, protection switching schemes include, in order of additional cost, the following:

- 1) 1:1, where two switch fabrics are present, but only one is active (traffic passing through) while the other one is in a standby mode
- 2) 1 + 1, where both fabrics are used simultaneously to forward traffic and, in the event one fails, the other one will take up its traffic
- 3) $N:1$, a generalization of the 1:1 mode, where more than one standby switch fabric is present, and in the event the active one fails, traffic is automatically switched through one or more back-up fabrics
- 4) $N + N$, $N + M$, where additional, e.g., twice as many, switch fabrics are present, particularly when implemented in large core switches or routers, also offering load-sharing capabilities.

DESIGN CONSIDERATIONS

In order for a switch fabric design to be successful in terms of suitability and market acceptance, it must encompass a number of characteristics that signify its usefulness and value. Such features include:

- ◆ **High performance:** In principle, the ideal switch architecture should yield 100% throughput, regardless of the traffic pattern applied and in conjunction with eliminating any external or internal blocking since blocking can seriously degrade the performance of a fabric; minimal delay (less than a few microseconds); and no data loss (under normal traffic conditions). Actual packet delay and loss are functions of the switch's buffer size. Support for QoS (ability to handle services and applications with differentiated traffic requirements) or prioritized traffic should be also accounted for.
- ◆ **Scalability:** The design must be able to scale and expand not only in terms of the port count, but also in terms of line rate (thus, aggregate capacity, which could affect the internal design of the switch fabric).
- ◆ **Protection:** The appropriate redundancy mode is chosen depending on the switch's, and, overall, the system's, requirements.
- ◆ **Interoperability:** The design should feature compliance with existing standards and the ability to support and interconnect with other interfaces and devices such as network processors and queue controllers.

- ◆ **Physical requirements:** The design should exhibit minimal power dissipation and reduced fabrication board space. Other impairments, more particular to optical switches, include crosstalk and insertion loss across the interstage links.

- ◆ **Low cost:** Given the abundance of switching fabric architectures, cost is of paramount importance. It is certainly affected and determined by the various design parameters, such as memory components, redundancy, switching and processing capabilities, interfaces, and complexity.

Additional features could include support for multicasting, which is the ability to direct copies of a single packet to multiple outputs in concurrent rounds, support for time-division multiplex (TDM) traffic, such as voice circuits, in a multi-service application environment, and traffic control.

CONCLUSION

A switch fabric's main function is to relay data efficiently and rapidly from its inputs to its outputs, preserving any associated QoS characteristics and achieving the best possible performance. The fabric's overall performance, which is ultimately determined by its characteristics and design parameters, generates significant attention.

Clos fabrics have been quite popular in the realm of large, modular switching designs. Banyan networks have also been used extensively, mainly due to their self-routing property, where techniques for removing any blocking conditions are applied, e.g., a sorting network in front of a Banyan network [3]. Crossbar architectures, thanks to their cost efficiency, are also the design choice for many routers and switches that require only a few ports.

We examined a spectrum of (unbuffered) switching architectures and explored various design alternatives, where each is marked by its own merits and drawbacks. A number of trade-offs between the various switching architectures in terms of their characteristics (such as performance, implementation complexity, scalability, and cost) emerge. Hence, the wide variety of available switching fabrics means that no single architecture emerges as the only winning proposition. The ultimate decision is upon the switch architect, based on the overall system design requirements and specifications.

References

- [1] C. Kolias, "Analysis and performance evaluation of new architectures in high-speed packet switching," Ph.D. dissertation, Computer Science Dept., UCLA, 1999.
- [2] R. Ramaswami, and K. Sivarajan, *Optical Communications*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2002.
- [3] A. Pattavina, *Switching Theory*. New York: Wiley, 1998.
- [4] A. Tzanakaki, I. Zacharopoulos, and I. Tomkos, "Broadband building blocks," *IEEE Circuits Devices Mag.*, vol. 20, pp. 32–37, Mar./Apr. 2004.

Christos Kolias and Ioannis Tomkos are with the Athens Information Technology in Greece. E-mail: ckol@ait.edu.gr. CD ■