

A New Architecture for Switch and Router Design

Overview

This white paper is about the architecture of large ATM switches, Frame Relay switches and IP routers used by network service providers and telecom carriers to build the core of the Internet. We will focus on the high-performance, high-availability systems that are typically located in large telecommunications switching facilities or central offices. These are the *carrier-class* systems, characterized today by: (i) an aggregate bandwidth of 100's of Gb/s; (ii) approximately 8-32 interfaces operating at OC-48c and OC-192c that carry frames, packet-over-SONET (POS) or ATM-over-SONET; and (iii) a system availability in excess of 99.999% (i.e. out of service less than 10 minutes per year).

Although the protocols used by ATM, Frame Relay and IP are quite different, the switches and routers that process cells, frames and datagrams have many similar characteristics. To start with, they all possess two basic components: *Linecards* which terminate each external line (i.e. perform physical layer functions such as framing, clock synchronization and signalling), and determine where each packet is to be sent next; and a *Switch Core* which transfers packets from the ingress Linecard to the egress Linecard (or to multiple egress Linecards if the packet is a multicast packet). Because of their similar structure, we will be referring to ATM switches, Frame Relay switches and IP routers as *packet-switches*. All packet-switches perform three basic functions: (i) when a packet (cell, frame or datagram) arrives at an ingress Linecard, the packet-switch must decide *where* the packet goes next (the next hop towards its destination); (ii) next, the packet-switch must deliver the packet from the ingress Linecard to the egress Linecard(s) that connects to the next hop; and (iii) lastly, the egress Linecard must decide *when* the packet should be sent to its next hop (e.g. the waiting packets could be transmitted in FCFS order, or according to a scheduling discipline that guarantees delay bounds through the packet-switch).

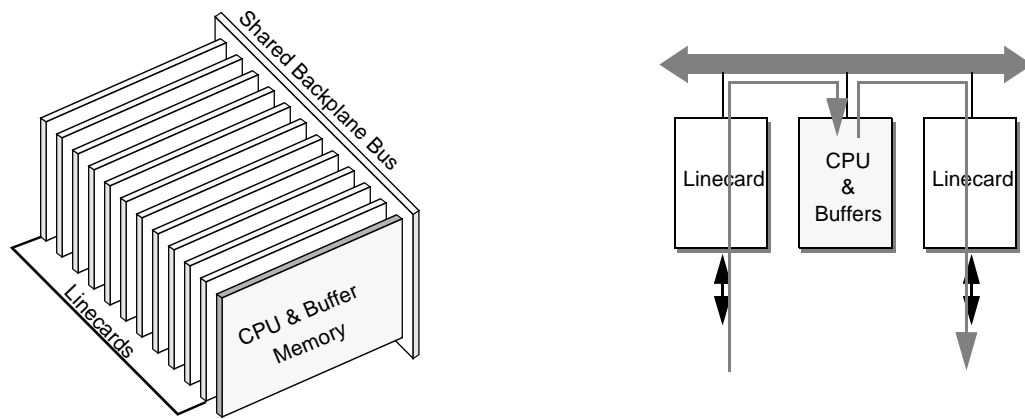
The Evolution of Switches and Routers

Over the years there have been a variety of architectures used for carrier-class packet-switches. While the detailed implementation of different vendors' commercial packet-switches have remained proprietary, they have — in broad terms — evolved in similar ways. Most notably, there has been a trend towards more parallelism to achieve higher performance. Parallelism is exploited in two ways: First, components that were once shared (such as centralized CPUs and shared memory) are now commonly placed on each Linecard where they need only support the requirements of a single interface. Second, parallelism is commonly used to process a stream of packets using multiple, identical elements.

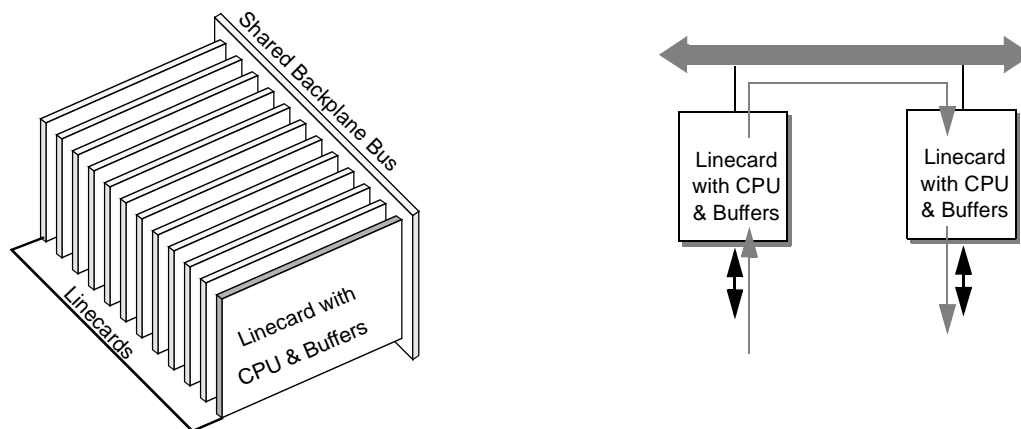
The original packet-switches were built around a conventional computer architecture, as shown in Figure 1(a): a shared central bus, with a central CPU, memory and peripheral Linecards. Each Linecard provides the physical layer framing and interface to the external links. Packets arriving from a link are transferred across the shared bus to the central CPU, where a forwarding decision is made (for example, a VCI translation in an ATM switch, or a longest-prefix matching in an IP router). The packet is buffered in the central shared memory while it waits for its outgoing link to become free. When its turn comes, the packet is transferred once again across the bus to its outgoing Linecard, and out onto the external link. The main limitation of this architecture is that the central CPU and central shared memory must process/buffer every packet, limiting the throughput of the system. This limitation prompted the architecture in Figure 1(b), in which a separate CPU and memory buffer is placed on each Linecard. A local forwarding decision is made in a dedicated CPU and the packet immediately forwarded to its outgoing Linecard. The parallelism of multiple processing elements increases the system performance, and by avoiding a central CPU¹ each packet need only traverse the bus once reducing congestion of the shared interconnect.

1. A shared CPU is still needed for centralized functions, such as maintenance of lookup tables and participation in routing protocols (e.g. PNNI in ATM switches and BGP in IP routers). These functions are performed infrequently and do not limit the throughput of the system.

A New Architecture for Switch and Router Design



(a) Shared backplane packet-switch with a single, central shared CPU and memory buffer.



(b) Shared backplane packet-switch with a separate, CPU and memory buffer on each linecard.

Figure 1 The basic architectures of packet-switches based on shared backplane buses.

The performance of the architecture in Figure 1(b) is ultimately limited by the use of a shared bus — only one packet may traverse the bus at a time between two Linecards. For this reason, recent designs have used crossbar switches in place of a shared bus, as shown in Figure 2. In a crossbar switch, multiple Linecards can communicate with each other simultaneously, increasing the total system throughput. A further advantage is that the electrical connections from each Linecard to the backplane can be short, fast point-to-point links, compared to the long shared and relatively slow multi-drop links in a shared bus.

A New Architecture for Switch and Router Design

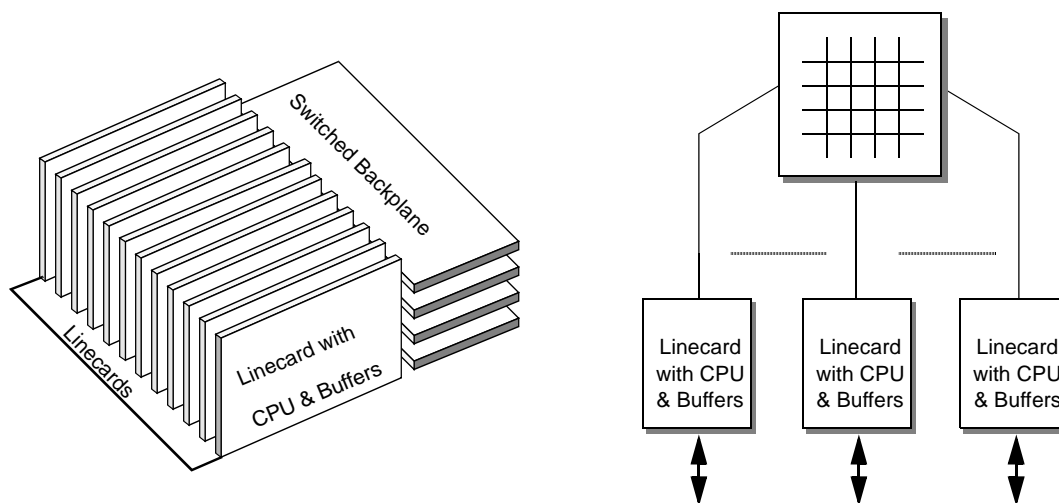


Figure 2 The basic architectures of packet-switches based on a switched backplane.

Increasing the Number of Ports

Most carrier-class packet-switches today support between 8 and 16 Linecards.¹ First, this is because carrier-class systems are — by definition — used at points of aggregation. Prior to reaching the aggregation point, many thousands of low-speed access lines have been multiplexed together using access multiplexers and lower-speed packet-switches. As we progress up the hierarchy and towards the center of the network, the bandwidth per link increases, but the number of links is quite small. A second reason for the small number of Linecards is the packaging density possible within the racks used in central offices. Rack sizes in the U.S.A. are dictated by NEBS² standards, and are currently limited to 19" wide. To allow room for components and air-flow, a spacing of about 1" is needed between Linecards. If arranged in the racks vertically (the preferred orientation), this practically limits the system to 16 or so Linecards. So until now, there has

1. Each Linecard may be further subdivided into multiple sub-ports. For example, an OC-48 linecard may be divided into four OC-12c, sixteen OC-3 external interfaces etc.
2. Network Equipment Building System (NEBS) are generic requirements provided by Telcordia (formerly Bell Communications Research, Inc. or Bellcore) that cover many aspects of network equipment. These are requirements for telephone switch rooms that dictate how equipment should be packaged, and how it should operate under temperature and humidity ranges, as well as earthquake vibration and fires.

been a fortunate coincidence between the number of required Linecards and the practical limit on the number of Linecards that can be placed in a carrier-class packet-switch.

But this is all in the process of changing; and the driving force is WDM (wave division multiplexing). For although the number of fibers entering each central office is not necessarily increasing, each fiber can now contain multiple, independent channels. Systems are available today that multiplex 40 separate OC-48c channels onto a single fiber, and the amount of multiplexing will continue to grow in the future. Currently, the most common data-rate per channel is OC-48c (2.48Gb/s), dictated by the speed of optoelectronic components and the speed of electronics on each Linecard.³ As seen by a packet-switch, the number of channels has now increased forty-fold; after the channels are separated, they need to be terminated by a packet-switch with up to 40 times as many ports, and 40 times the aggregate bandwidth. With an estimated growth rate in the core of the Internet of 10-times per year, the 40-fold improvement brought by WDM will be full utilized in less than two years from now.

Fortunately, due to improvements in semiconductor technology, we can expect new systems to become

3. Over the next few years there will be a move towards systems operating at OC-192c (10Gb/s) per channel.

A New Architecture for Switch and Router Design

available with higher and higher aggregate bandwidth. Can we count on improvements in silicon technology to provide the improvements that are needed? After all, Moore's Law has shown that many aspects of system performance will double every eighteen months. Unfortunately not; at this rate, it would take nine years to see a 40-fold improvement!

Clustering: Separating Linecards From Switch Cores

To increase the number of Linecards and the aggregate system bandwidth, architectural innovations are needed. We describe here one solution to the problem: *clustering* shelves of Linecards around a central Switch Core. This approach physically separates the Linecards from the Switch Core, as shown schematically in Figure 3.

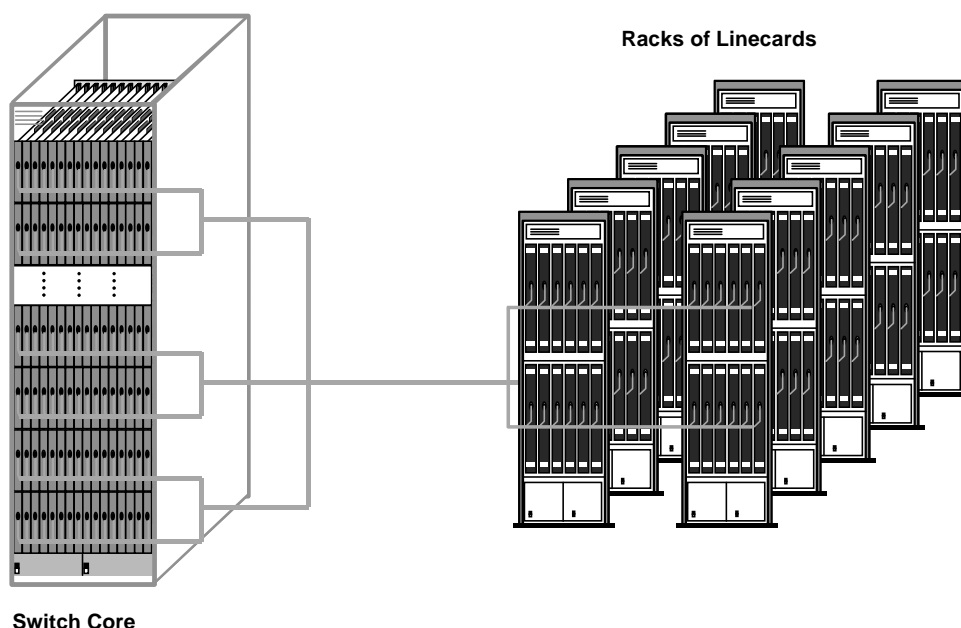


Figure 3 A packet-switch based on *clustering* in which many Linecards are arranged around a central Switch Core.

There are a number of reasons that make this approach attractive:

1. **Large number of Linecards.** By removing the physical packaging constraint of arranging multiple Linecards around a Switch Core on a single rack, the system is easier to package, easier to cool, and most importantly can allow a larger number of Linecards to be interconnected in a single packet-switch.
2. **Simple fault-tolerance of central shared switching sub-system** A single shared Switch Core is a single point of system failure. For high-availability systems, a second Switch Core can be used to provide simple fault-tolerance. This is shown schematically in Figure 4.
3. **Upgrade path with backward compatibility.** A clean separation of Linecards and Switch Core also separates their development. A Linecard developed today, can potentially connect to Switch Cores in the future with larger numbers of ports, or with new features. If the fault-tolerance is implemented carefully, it may even be possible to upgrade a whole Switch Core without interrupting service.

A New Architecture for Switch and Router Design

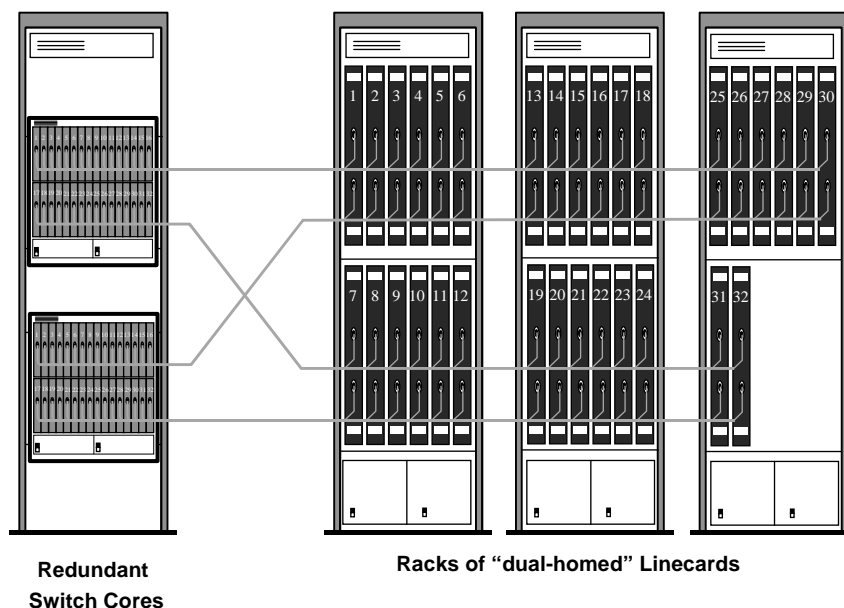


Figure 4 Redundant switch-cores are a simple and manageable way to provide system fault-tolerance.

The PMC-Sierra LCS Protocol

In order to help system vendors bring clustered packet-switches to market, PMC-Sierra has developed a new clustering protocol called the *Linecard to Switch (LCS) Protocol*. The patented LCS Protocol was developed to allow simple, lossless Switch Cores to be developed that interconnect OC-192 Linecards operating at 10Gb/s. The protocol may also be used at lower rates (such as OC-48) and extended upwards to higher data rates (such as OC-768). The basic LCS Protocol allows Linecards to be placed hundreds of feet away from the Switch Core.

The LCS Protocol is most easily characterized as *label-swapping with flow-control*. Label-swapping is used so that hardware on the Linecard can be developed without explicit knowledge of the number of ports on the Switch Core. Data sent into the Switch Core is tagged with a label that has been pre-agreed between the Linecard and the Switch Core: the label indicates which outputs the data should be sent to and with what priority. Per-queue flow-control is used to allow the Switch Core to operate without loss. A logical representation of the LCS Protocol is shown in Figure 5.

Data is passed from a Linecard to the Switch Core in 64-byte (or optionally, 76-byte) "cells", with a prepended 8-byte LCS Header. The Switch Core is responsible for delivering the 64-byte payload unchanged and without loss from the ingress Linecards to the egress Linecards. No assumptions are made about the contents of the 64-byte payload; it could contain, for example, a fragment from a variable-length IP packet, Ethernet or Frame-Relay frame; or it may contain a complete ATM cell. The independence of the LCS Protocol from the contents of the payload allows LCS to be used in ATM switches, Frame Relay switches, IP Routers and even Gigabit Ethernet systems.

The LCS Protocol implements a simple three-way handshake to transfer data from an ingress Linecard through the Switch Core to one, or more, egress Linecards. The process begins with the ingress Linecard sending an LCS Request into the Switch Core. The Request contains a label to identify the output(s) that the requested cell should be sent to, and the class of service that it is to receive. The reason for using a label (rather than an explicit bitmap indicating which outputs the cell is destined to) is to keep the protocol independent of the number of ports of the switch, and independent from the classes of service that

A New Architecture for Switch and Router Design

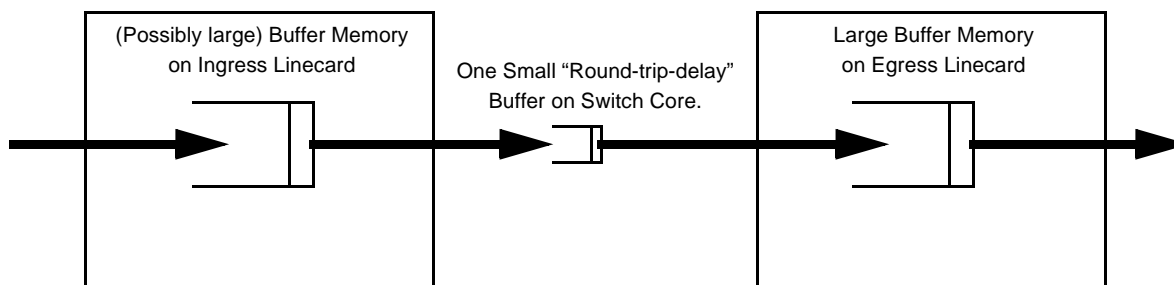
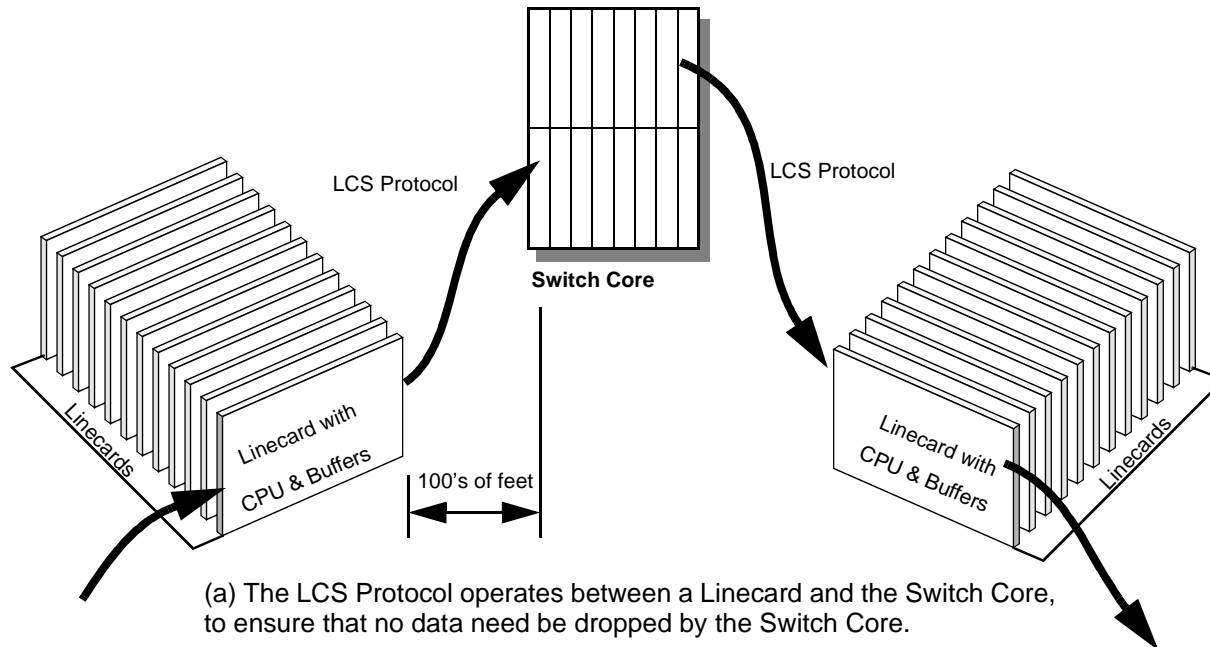


Figure 5 Schematic representation of the cluster architecture made possible by the LCS Protocol.

the Switch Core provides. Hence, LCS can work with a variety of generations of Switch Core.

The second stage of the handshake occurs when the Switch Core has processed the LCS Request and is ready to accept the cell. Essentially, the Switch Core schedules a time, in the future, when the cell is to pass through the Switch Core. Once the time has been picked, the Switch Core retrieves the cell from the ingress Linecard by sending an LCS Grant/Credit. The Linecard responds by delivering the cell into the Switch Core, where it is ready to

be passed to the egress Linecard(s). Because of variations in distance between the Switch Core and different Linecards, a cell may arrive before the time it was scheduled. To ensure that the cell is sent at the correct time, each ingress port of the Switch Core maintains a single "delay-matching" FIFO that holds cells until their time to move on. The buffer holds enough cells for one round-trip time between the Switch Core and Linecard: approximately 128 cells if they are separated by up to 2000ft.

A New Architecture for Switch and Router Design

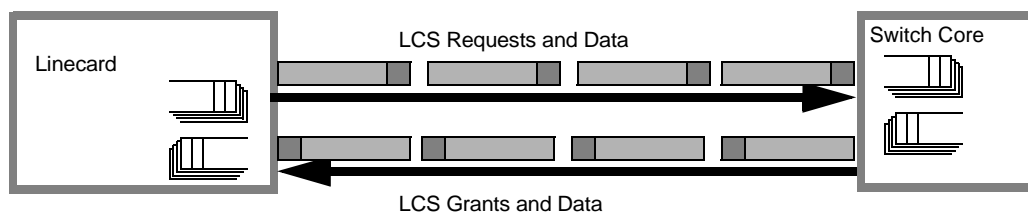


Figure 6 The main function of the LCS Protocol is to provide a per-queue flow-control between an ingress Linecard and the Switch Core. Flow-control ensures that no data loss need occur in the Switch Core, allowing a Linecard to determine if and when data should be dropped.

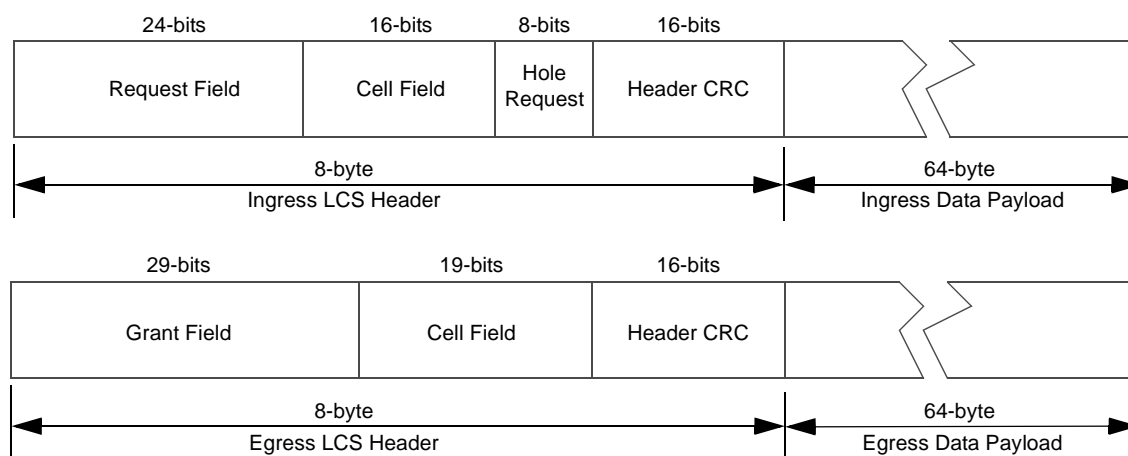


Figure 7 The main fields in the 8-byte LCS Header. The Cell Field is used to identify the flow to which the 64-byte Data Payload belongs. The Request Field of the Ingress LCS Header requests that a pending cell in a queue from the ingress Linecard be scheduled for transfer at a future time. The Grant Field of the Egress LCS Header grants permission to transfer a previously requested ingress cell to the Switch Core within a bounded round-trip time. The Hole Request requests that the transfer of a cell to a specified egress flow be delayed a single cell time. A 16-bit CRC protects the whole LCS Header.

The mechanism is shown schematically in Figure 6, and a simplified version of the LCS Header format is shown in Figure 7.

In addition to the attributes described above, the LCS protocol provides a number of other features, including:

1. LCS allows necessary housekeeping functions (such as fault isolation and self-test) to temporarily suspend the flow of packets without losing data.
2. LCS provides an in-band mechanism for control of the Switch Core from one or more Linecards.
3. LCS provides a simple mechanism for the Switch Core to synchronize scheduling to a system-wide Stratum 1, 2 or 3 clock.
4. The LCS Protocol allows recovery from bit-errors on the link.

A New Architecture for Switch and Router Design

Conclusion

Continued growth in demand has forced telecommunication carriers to seek cost-effective ways to carry more data over existing optical fibers. This has led to the widespread installation of WDM systems, with upwards of forty separate OC-48 channels on each optical fiber. With the bottleneck removed from the links, attention is focussed once again on the switches and routers that process cells and packets at data exchange-points. As we have seen in this paper, the advent of WDM does not simply require systems with higher aggregate bandwidth; we now require systems that can accommodate a much larger number of linecards. This presents a problem — packaging a system with 16 OC-48 Linecards into a single 19" rack is only just possible today. Fitting hundreds of Linecards into one rack is unthinkable.

But by physically separating Linecards from the central Switch Core to form a clustered system, allows much larger systems to be built. In particular, the LCS Protocol was developed to enable clustered packet-switches to be built. Using this protocol, scalable, fault-tolerant systems can be built with hundreds of feature-rich Linecards. Furthermore, this approach allows protection of the end-user investment through re-use of today's Linecards in future carrier-class systems.

With this approach, we can expect to see systems with aggregate bandwidths of multi-terabits per second over the next few years.