CS-534: Packet Switch Architecture
Spring 2004

Department of Computer Science

0.1 Technology Outlook: Switches and Interconnects

Interconnection networks constitute the *backbone* of all emerging information and communication systems. Interconnection plays a vital role: stand-alone devices are increasingly useless. The switch is the basic building block for wireline interconnections, pretty much like the microprocessor is the building block for processing, and the RAM is the building block for storage. The *switch* is a multi-port device that routes incoming packets each to the proper output port, while resolving conflicts (multiple packets simultaneously desiring to exit through the same port) by temporarily buffering all packets but one, then scheduling their departure at an appropriate later time. The interconnection network and switch technology is in a period of large grouth, radical evolution, and dramatic changes, for the reasons outlined below.

0.1.1 Ubiquitous, Switch-based Interconnection Networks

High-speed interconnects constitute a basic infrastructure for all kinds of electronic systems, because interconnection and communication is becoming ubiquitous. Performance requirements push more and more electronic systems to abandon traditional bus-based architectures, in order to benefit from the performance, reliability, and lower power consumption offered by switched interconnects. Buses are limited by arbitration delay (who will transmit next), turn-around overhead (idle time between two transmissions, due to the bidirectionality of the medium), and lack of parallelism (only one transmission at a time). By contrast, switched interconnections use point-to-point links (which can carry packets back-to-back, with no idle time), and offer massive parallelism (every interconnection link can carry packets, all at the same time). Given this trend, switches will soon become the predominant interconnection component.

Switched architectures gradually proliferate from the long-distance end of the spectrum to short-distance communication: from wide (WAN), metropolitan (MAN), local (LAN) area networks, and cluster interconnects, to storage and system area networks (SAN), computer I/O, processor-memory interconnects, embedded systems, and networks-on-chip (NoC). Many companies are now developing new communication standards and products for different market segments (HyperTransport, PCI Express, PCI Express Advanced Switching, Rapid I/O, InfiniBand, 40Gigabit Ethernet, etc.).

Switches and routers, today, especially at the gigabit level, are much more expensive than computers, but this has to change and will change as switched interconnects enter into new market segments. Optical or electronic switches with a large number of fast ports, appropriate for WDM, are extremely expensive. A fully deployed router with 32 ports at OC-192 (10Gbps) per port will cost around 300 K to 1 M Euro, depending on configuration, options and vendor. By all means, lower-cost interconnection and switching technology is urgently needed, and this must happen while aggregate throughput keeps climbing at a fast pace.

0.1.2 New Market: Next Generation IT Infrastructure

The technology wave, today, is directed at the commoditisation of the technology base that was previously associated with the high-end (super-) computing platforms. Key building blocks have been identified which collectively reduce the burden of the processing engines and create a truly distributed processing environment. The applications of this technology platform extend beyond the traditional data enter, into networking, Grid computing, video server, and wireless applications. The building blocks include:

- Low-cost, high-performance server platforms: Blade Servers;
- Low-cost, high-performance Network Storage Systems;
- Low-cost, high-performance Interconnection Silicon (switching, protocol processing);
- System software support for hardware acceleration and management (system cache, storage management).

Blade servers are already widely available and projected by IDC to be shipping in the millions in 2005. The goal is to provide a truly flexible environment, in which these commodity processors --the blades-- can be repurposed by the revolutionary middleware, so that we can efficiently move processing tasks within the data center distributed processing environments. This will enable loads to be re-balanced on demand during system operation, resulting in better services and lower costs.

In the storage systems marketplace, a similar revolution is taking place. Traditional direct-attached storage systems are being replaced by high-speed, high-availability Networked Storage Systems. The emergence of the commodity SATA disk drive is causing a new market of "Near Line Storage" to emerge, to address the problems of data backup in a 24/7 working environment. Yet another new market is emerging with the deployment of IP Storage Systems. All of this activity is driven by the growing need to manage and have continuous access to digital information; it is at the core of the emerging huge storage system market.

To operate efficiently, blade servers and networked storage will need a high-performance interconnection network that connects them to each other and to I/O and the Grid; latency and quality of service must be equally managed. The mass market, with its need for low cost, will inevitably lead to the commoditisation of these switching, routing, network interfacing, and protocol processing functions, too.

0.1.3 Commodity Switches: Next Generation Interconnects

The new markets for switches --SAN and cluster interconnects, or short-distance communication in general-- are much larger than the existing markets in long-distance networks --WAN and LAN. Larger markets need and bring lower cost. There is a positive feedback effect in a product becoming a commodity: in order to be sold by the millions --as a commodity-- it has to be inexpensive; and when it sold by the millions, economy-of-scale effects make it even less expensive. The PC market evolved in this way, and switches will inevitably evolve in the same manner.

It is predicted that, in the next couple of years, *commodity switches* will emerge as a key commercial product that will become a basic building block --pretty much in the same way that the microprocessor is today a basic building block-- for the future networked computing and communication systems, across all above areas. Distributed processing environments are enabled by low cost high bandwidth switch technology, integrated with hardware acceleration technology, to remove latencies from the new compute intensive applications.

Once commodity switches become a reality, they are likely to drastically affect the entire router and switch market, effectively becoming a disruptive technology. Let us use an analogy: in the mid-90's, supercomputers disappeared from the market, because they were expensive machines sold to a small market; development time and cost were too high. They were replaced by clusters of workstations --systems built out of commodity components. Even though a cluster performs less than a hypothetical supercomputer with the same number of processors and built out of the same technology, *(i)* clusters ended up always using newer technology --because new processors are first applied to commodity products-- and *(ii)* commodity microprocessors were much less expensive, so customers could have many more of them for the same price. The same phenomenon is likely to be repeated with switches: we expect high-performance switches and routers to be based on switch fabrics built out of commodity switches (e.g., switches for PCI-Express), thus taking advantage of sales volume to drastically reduce costs.

0.1.4 Switch and Interconnects Architecture

Switch architecture is a relatively young field --15 to 20 years old-- in computer engineering, as compared e.g. to computer (processor) architecture, which is 40 to 50 years old. Because of this, *(i)* there is not much awareness of switch architecture in the education and research community (e.g. few books, courses, conferences), hence similarly among graduating engineers; and *(ii)* current switch architectures are not yet mature.

Contemporary interconnection and switch architectures have not matured yet: they vary widely, evolve rapidly, and do not yet meet a number of objective goals, especially if one considers the wide spectrum of their application domains. While throughput and feature requirements increase, complexity and cost grow at an alarming pace; thus, new techniques and architectures are needed in order for the cost of switches and routers to be kept down. If we use an analogy to the wide spectrum of processor architectures before the mid-eighties, contemporary switch architectures are still in their "pre-RISC" stage: the *"RISC architecture" for switches* still remains to be found.

Contemporary switch architectures also differ significantly in their various application domains --WAN, MAN, LAN, SAN, etc-- thus resulting in large required design effort, hence increased cost. As discussed above, in the same way that supercomputers moved from expensive proprietary vector architectures to massive parallelism based on commodity microprocessors, in the future we expect high-performance switches and routers to be based on switching fabrics built out of commodity switches, thus taking advantage of sales volume to drastically reduce costs. We must search to discover the *unifying concepts* for the switches at all of the above scales, so as to allow reuse of design and cost savings.

Another reason why switch architecture has not yet matured is that a number of other central issues remain as *open questions*. How to build a single stage switch with limited ports (up to 64) is fairly well understood today. However, for telco applications higher numbers of ports will be needed, and bus replacement applications will typically require switching fabrics (networks of switches) with limited bandwidth links. For large numbers of ports, fabrics is the only solution. Today, switching fabric architecture is still an unsolved problem when compounded with high utilization, throughput, and Quality of Service (QoS). Another open question is how to optimally mix electronic and optical switching, dedicating each to the tasks where it is best suited, and making them both cooperate in the best possible way. Other open research topics include chip-to-chip interconnects, power consumption, network processor architecture, hardware support for monitoring and for security, how to handle multicast traffic, fixed size cells versus variable length packets, reserved capacity versus bandwidth-on-demand, routing algorithms, flow and congestion control, flow isolation versus aggregation, queueing structures, buffer management, packet dropping versus backpressure, scheduling algorithms and QoS.
