Current Issues in Packet Switch Design

Cyriel Minkenberg, Ronald P. Luijten, François Abel, Wolfgang Denzel, Mitchell Gusat
IBM Research, Zurich Research Laboratory
Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland
sil@zurich.ibm.com

ABSTRACT

Addressing the ever growing capacity demand for packet switches, current research focuses on scheduling algorithms or buffer bandwidth reductions. Although these topics remain relevant, our position is that the primary design focus for systems beyond 1 Tb/s must be shifted to aspects resulting from packaging disruptions. Based on trends such as increased link rates and improved CMOS technologies, we derive new design factors for such switch fabrics. For instance, we argue that the packet round-trip transmission time *within* the fabric has become a major design parameter. Furthermore, we observe that high-speed fabrics have become extremely dependent on serial I/O technology that is both high speed *and* high density. Finally, we conclude that in developing the architecture, packaging constraints must be put first and not as an afterthought, which also applies to solving the tremendous power consumption challenges.

1. MOTIVATION

Most research on single-stage, electronic packet switches focuses primarily on high-level architectural issues such as buffering and queuing strategies and scheduling algorithms, but seldom considers all the physical issues that arise when actually building the system, a tendency that has also been noted in [1]. For instance, although many new packet-switch architectures have been proposed in the past few years, the majority of these architectural-level papers reduce memory bandwidth by eliminating the N-fold speed-up required by output queuing and instead optimizing the performance of the centralized scheduling algorithm needed in such architectures. Although reducing memory bandwidth is an important issue, it is not sufficient in itself and may render the resulting system economically infeasible. Good examples are the recent proposals of combined input- and output-queued (CIOQ) switches with limited speed-up, which require speed-up throughout the fabric (i.e., both input and output speed-up), thus multiplying the bandwidth that must be carried across the switch core, whereas the output speed-up implemented in an output-queued switch is purely internal to the switch core.

Designers of practical high-capacity packet switches face challenges on two levels: First, they must choose a design point at the architectural level, in terms of buffering and queuing strategies, scheduling algorithms, and flow-control methods. For example, for the buffering strategy, the choice to be made is between an input-, output-, or combined-queuing structure. Second, they must consider the physical level, i.e., the implementation of the architecture at the system as well as the chip level, in terms of partitioning over racks, cards, chips, and the design of the individual chips comprising the system. Switch designers have little freedom with respect to system packaging issues. On the one hand the technology imposes constraints, on the other hand customers impose their specific requirements, in addition to more general ones such as NEBS (Network Equipment Building System) compliance [2, 3]. NEBS comprises a set of stringent physical (e.g., space planning, temperature, humidity, etc.) and electrical (e.g., EMI, power fault, bonding and grounding, etc.) requirements, originally developed for telephony equipment. Nowadays, NEBS compliance is a prerequisite for networking and computing equipment in general to ensure reliable operation (also under adverse conditions), safety, compatibility, and freedom of interference.

The designer must decide how to distribute functionality over chips, taking into account current technology limitations. At every level, the designer is constrained by requirements and technological feasibility, and has to optimize overall system cost and power. We argue that the new constraints arising from packaging and power invalidate the traditional design approach. Rather than finding a suitable packaging for the architecture, we are forced to find a suitable architecture for the packaging.

Although a multi-Tb/s switch is not (yet) a commodity, for cost reasons we assume the use of "commodity" technology and components in our study, i.e., CMOS technology, standard-cell design (no full custom), commodity chip packaging (< 1000 pins per chip), commercially available connectors, and standard system packaging (e.g., NEBS compliance). We also assume that packet-level QoS and routing are required.

In Sec. 2 we discuss the major trends in switch design in a trend/cause/effect format. Section 3 gives a system-level

HOTNETS '02 10/02 Princeton, NJ, USA

description, introducing some basic assumptions about system structure. In Sec. 4 we discuss the major consequences on switch design in general that result from the trends observed, whereas in Sec. 5 we discuss a number of specific implications for two popular architectures. Finally, we draw our conclusions in Sec. 6. The main purpose of this paper is to draw attention to the issues discussed, rather than to provide solutions.

2. TRENDS

Trend 1. The aggregate throughput will grow by increasing the *number of ports* rather than port speed.

Cause 1. Although transmission line rates have increased rapidly over the past years [OC-3 (155 Mb/s) to OC-192 (10 Gb/s) and OC-768 (40 Gb/s)] and will most likely continue to do so, it appears that the granularity at the switch level will be OC-192 for the coming years [4]. First, the existing installed base of OC-192 line cards must still be supported. Second, dense wavelength-division multiplexing (DWDM) vastly increases the number of channels available on a single fiber, but not the speed of a single channel. At the switch, this translates into more ports at the same speed. Also, on the electrical side the SRAM clock cycle poses a limit in combination with the minimum packet size; the limit is reached when the line rate times the clock cycle exceeds the *minimum* packet size. This limit can only be overcome by enlarging the minimum packet size, which requires packet aggregation techniques. Right now, a cycle of 2 ns is feasible, which, given a minimum packet size of 64 B, yields a maximum line rate of 256 Gb/s, which is just one generation away (OC-3072 = 160 Gb/s).[1] Finally, it is becoming increasingly difficult and costly to perform line-speed processing in the network processors (NP) at these rates.

Contrary to wavelength density, *port* density in terms of ports per line card is not increasing, for two reasons. First, availability requires that single points of failure be avoided as much as possible. Putting multiple NPs and adapters on one line card would cause service interruption for several ports if one component on the card fails. Second, the increase in density offered by CMOS technology is invested in increased functionality, such as level-4 routing, firewalling, MPLS, and DiffServ, rather than increasing speed, or reducing size or power.

Effect 1. The increase in port count at constant port speed translates directly into an increase in physical line-card space. On the other hand, compliance with NEBS equipment requirements for the physical form of the system imposes strict limitations on the number of cards that fit into a rack. Combined with the above trend, it quickly becomes clear that the switch fabric can no longer be built in a compact, single-rack fashion, and that a multi-rack solution is necessary. In a single-rack system, all fabric-internal communication only crosses the backplane, but once the system becomes multi-rack, line cards and switch core become separated by a much greater physical distance, with (long) cables[2] interconnecting the racks, implying that the rack's backplane is replaced by cables. Rack-spacing requirements and other spatial limitations (adapter racks might even be in other rooms than the switch core rack) determine how far apart the racks have to be—this can easily be tens of meters.

Trend 2. With each new switch generation a higher proportion of the switch chip power is used to transport signals across chip boundaries.

Cause 2. The number of chip pins grows by 5 to 10% with each CMOS generation, whereas density doubles. Rent's rule [5] states that the number of pins, N_p, and number of logic gates, N_g, form a relationship of the form $N_p = K_p N_g^{\beta}$, where K_p is a constant and $\beta < 1$. Therefore, higher per-pin signal rates are needed to scale aggregate bandwidth proportionally, which in turn implies more complex analog macros requiring increased power. Table 1 shows numbers for three generations of a typical switch chip.

Effect 2. Switch chips become increasingly I/O- and power-constrained. At the system level currently more than 50% of the power is spent in transporting rather than switching data.

Table 1: Single-chip I/O Power Consumption Proportion

Switch size	16×16	16×16	32×32
Throughput (Gb/s)	6	32	64
Total power (W)	6	12	25
I/O power (W)	1	4	12
I/O power (%)	**16**	**33**	**50**

Trend 3. The growth in CMOS integration density far outpaces the growth in electrical interconnect bandwidth density.

Cause 3. The bandwidth growth over chip pins as well as card edge is a few percent per year, compared with the doubling of CMOS technology density every 18 months. The card-edge limitations impose even tighter constraints than the chip-pin limit in the area of switching.

Effect 3. The maximum switch throughput that can be implemented on a single switch card in a multi-rack system is limited by card connector technology rather than by CMOS density. Given a target system capacity, this limit immediately imposes a lower limit on the number of cards required. Connecting across 5 m of cable plus about 50 cm of FR4 board circuit trace limits the bit rate to 2.5–3.125 Gb/s. Combined with the latest connector technology, this yields a density of about 120 Gb/s/in. With reasonable card sizes

[1]Fabric-internal line-rate escalation (speed-up) is required to compensate for header and segmentation overhead. We assume a typical speed-up of 60%.

[2]By cable we mean electrical and/or optical interconnects spanning a few to tens of meters.

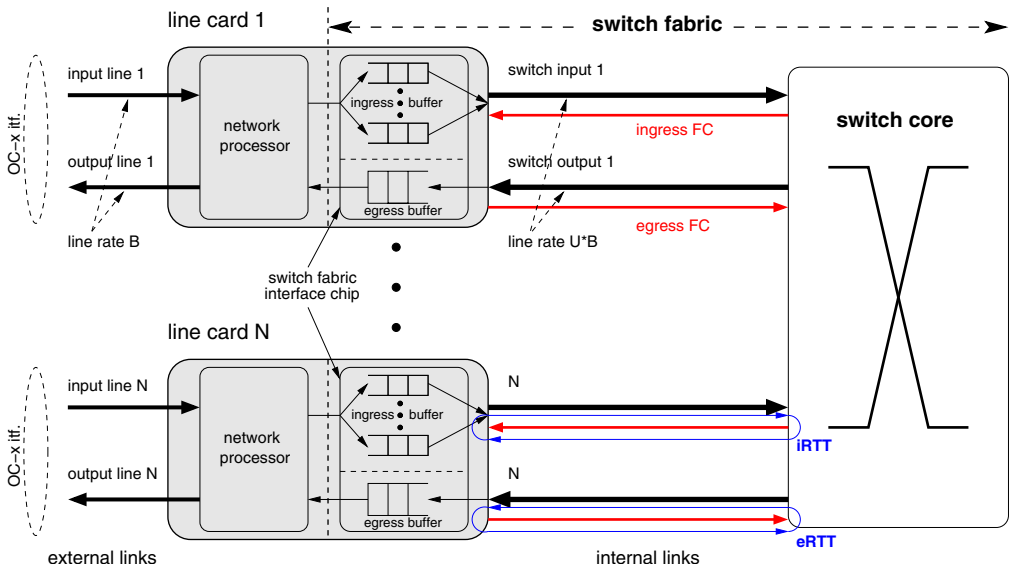

Figure 1: System level architecture.

($\leq$ 50 cm card edge) this results in a maximum card throughput of about 1 Tb/s (bidirectional), although CMOS technology would allow us to build higher throughput on a single card. Consequently, larger fabrics (i.e., multi-Tb/s) have to be split over multiple cards. Optical interconnect technology does not yet provide any improvement here (more space, power, and cost, but not faster), although it may be needed to cover longer distances. To satisfy availability requirements (typically 99.999% uptime is a must, which translates into about 5 min downtime per year), the switch must be implemented with a sufficient degree of redundancy. As the element of redundancy is a card, a small, single-card switch can economically achieve 1+1 redundancy, whereas fabrics larger than 1 Tb/s preferably should employ N+1 redundancy.

Trend 4. The required throughput/power density in terms of Gb/s per Watt per physical volume is increasing.

Cause 4. From one generation to the next, the market expects increased performance at constant or less power consumption. This also applies to performance per physical volume. One of the reasons is that high-end routers are often packaged as telecom racks, which are subject to a fixed power-consumption limit per rack (typically 2 kW/shelf), and must also be NEBS compliant.

Effect 4. The throughput/power density of all system components must be scaled up in proportion to the increase in aggregate bandwidth.

Trend 5. The minimum packet duration has shrunk significantly.

Cause 5. Whether traffic is ATM, IP, or SAN, the minimum packet size is still in the range of 32 to 64 B. The line rate has evolved exponentially to OC-192 currently, and to OC-768 in the near future. Accordingly, the transmission duration of the minimum-sized packet has shrunk from micro- to nanoseconds.

Effect 5. On a given length of cable or backplane trace, more packets are in flight.

Trend 6. Moore's law paradox in the latest CMOS generations.

Cause 6. Although Moore's law doubles performance every 18–24 months at constant cost, this performance is mostly due to increased density rather than increased clock speeds. From one CMOS generation to the next, taking global wiring into account, switch-chip clock speeds can be increased by only 5–10%.

Effect 6. To exploit Moore's law, more parallelism is required, typically at constant clock speeds. In turn, this results in more levels of pipelining in the control path and a higher degree of parallelism in the data path. Similarly, on-chip memory speed does not increase, and memory busses become wider in each new generation.

3. SYSTEM-LEVEL ASSUMPTIONS

To be able to make some more concrete statements, we need to make a few assumptions at the architectural level, see Fig. 1. First, most practical switch systems consist of three components: the ingress line cards, the switch core (routing fabric), and the egress line cards. Typically, the ingress and egress line cards of the same port are implemented together on the same card. The ingress line card comprises a network processor that communicates through a switch-core interface (e.g., SPI-x, CSIX [6]) with an adapter chip that per-

forms the translation to the native packet format. Both network processor and adapter chip contain buffers. Buffers on the egress side are necessary to accommodate downstream blocking and the speed difference between internal and external link rates. These differ to account for the segmentation and header overhead incurred by the switch-internal packet format, which typically does not map directly to the external packet format (e.g., conversion from variable-length TCP/IP packets to short, fixed-length packets). The switch core performs the actual switching function—this can be any type of switch, e.g. a crossbar, a buffered crossbar, a shared-memory switch, etc. We assume that the routing fabric is single-stage and that it contains no further input buffers (all input buffering is performed on the line cards).

To ensure that the switch fabric is internally lossless, flow-control protocols are required between any two stages that contain (limited) buffers. The *round-trip time* (RTT), as illustrated in Fig. 1, is the sum of the latencies of the forward (data) and the reverse (flow control) path, usually expressed in terms of the packet duration. The RTTs at the ingress (iRTT) and egress (eRTT) sides of the switch core may be different. With a buffered switch core, iRTT and eRTT can be decoupled, whereas with a bufferless core they are combined.

4. CONSEQUENCES

We discuss the main consequences that have emerged from the trends observed in Sec. 2. These are typically the result of a combination of several trends.

4.1 Physical system size

Trends 1, 3, and 6 culminate in the following consequence:

Consequence I. Switch fabrics beyond 1 Tb/s aggregate throughput experience a disruption. Single-rack packaging is no longer possible, implying larger systems, cables replacing backplanes, and more transmission power.

This entails significant physical system-packaging issues, see Sec. 4.3, but also immediately gives rise to Consequence II:

Consequence II. Switches have become critically dependent on serial I/O technology that is *both* high-speed *and* high-density and can cross backplanes and cables. These links must be available in the ASIC technology library used by the switching chips and must be very low power because of Trend 4. Owing to transmission line effects combined with power limitations, the I/O technology has not kept pace with the line rates. Therefore, the number of links required to implement a given system capacity has grown substantially. As a result, cabling accounts for a significant part of the total system cost.

4.2 Increased RTT

The physical distance implied by Consequence I combined with the shrinking packet duration due to Trend 5 immediately leads us to Consequence III:

Consequence III. The switch-fabric-internal RTT has significantly increased in terms of the minimum packet duration.

A large RTT only used to be an issue from node to node, but it has now become important also within the switch fabric. Table 2 lists RTT values for four switch-fabric generations (from OC-12 to OC-768), assuming a minimum packet length of 64 B. The RTT is expressed in packet duration, and includes the contribution of both the time of flight (over either backplane or cable) and the pipeline logic and serialization/deserialization (serdes) delay. Note that the latter contribution is at least as important as the former.[3]

Table 2: RTT Expressed in Packet Duration

	Switch generation			
Line rate	OC-12	OC-48	OC-192	OC-768
Conn. dist	1 m, backpl.	1 m, backpl.	6 m, cable	15 m, cable
Pkt. dur.	512 ns	128 ns	32 ns	8 ns
RTT	$\ll 1$	~ 1	16	64

The increased RTT results in more packets in flight, and, independently of the switch architecture chosen, this needs to be accounted for with buffers in the system to ensure that the system is both work-conserving and lossless. In general, the amount of buffering required to ensure both losslessness and work-conservingness must be scaled proportionally to the RTT. However, because of Trend 6 we cannot simply add buffers at leisure. Moreover, the flow-control mechanism selected has a large impact on buffer sizes. As a result, RTT has become a major design parameter for switch architectures.

4.3 Packaging impacts

A decade ago, we could simply build switch chips with the maximum density permitted by CMOS, and a suitable packaging could always be found. Consequences IV and V express the different approach that is required nowadays:

Consequence IV. Now, because of Trend 3, the physical packaging of the systems must be established first, and the switch architecture must be designed to fit these constraints. Owing to Trend 1, long cables are now needed, but because of Trends 2 and 4, adding external cable drivers at leisure is not desirable, for power and cost reasons.

Consequence V. Because of Trend 4, the objective must be to avoid intermediate drivers and drive cables directly from the switch chip.

Packaging examples

Figure 2 shows packaging for a generic 4 Tb/s system, using two racks each with two shelves of 18 cards, with cables connecting the racks. In general, to prevent link blocking, the bandwidth between two racks must equal the maximum bandwidth that can be sourced and sunk by all the line cards

[3]For short and medium range cables.

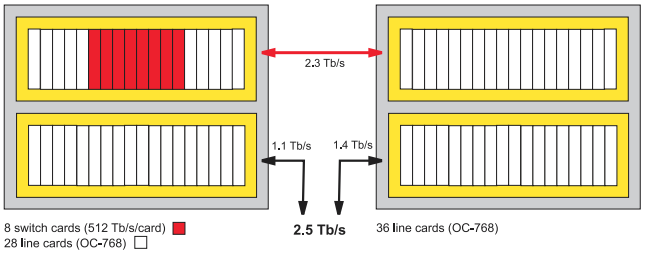

Figure 2: Generic 4 Tb/s system packaging (1.6x speed-up), fitting 64 OC-768 line cards and eight 512 Gb/s switch cards into two racks with two shelves per rack and 18 slots per shelf.

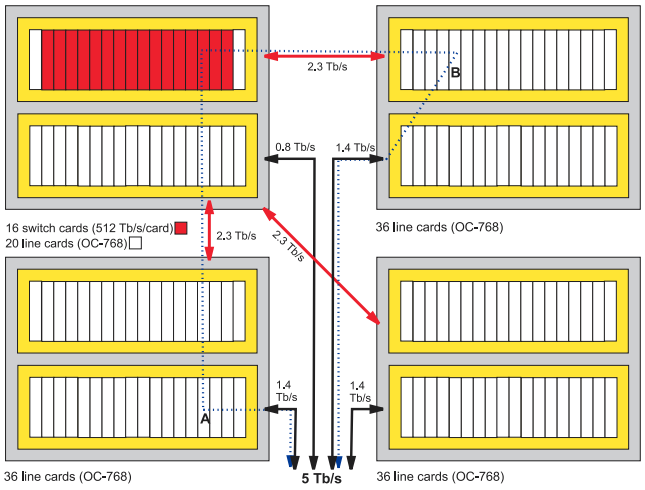

Figure 3: Generic 8 Tb/s system packaging (1.6x speed-up), fitting 128 OC-768 line cards and 16 512 Gb/s switch cards into four racks with two shelves per rack and 18 slots per shelf.

together in one rack, in this case 36 × OC-768 bidirectional. Figure 3 shows an 8 Tb/s system packaged in a similar fashion in four racks. The dotted line shows the path taken by a packet from line card A to line card B.

4.4 Other considerations

The above consequences reflect new requirements for the design of switch fabrics, but most existing requirements also still apply, such as in-order delivery, multicast, losslessness, availability, and the ability to perform nondisruptive repair actions. Of great importance is also the ability to follow an evolutionary migration path that allows as much of the existing investment as possible to be preserved, which requires backward compatibility and the flexibility in the switch core to support both lower and higher line rates.

A robust switch fabric must support any mix of traffic types, which imposes the worst-case requirements of each individual type of traffic simultaneously. Moreover, QoS must be provided for any given number of traffic classes (with widely varying characteristics). First, this has significant implications on buffer dimensioning, and second, as a result of Consequence III, this poses new challenges for intra-fabric flow control. Existing work in the area of link-level flow control, e.g. [7, 8], can most likely be adapted for this purpose.

5. DISCUSSION

The consequences discussed in Sec. 4 have a number of important architectural implications for practical packet-switch designs. Here, we will point out some of these for multi-Tb/s implementations of two popular architectures, namely, the input-queued architecture with virtual output queuing (VOQ), e.g. [9], and no speed-up, and the combined-input-and-output-queued (CIOQ) architecture with limited (2 to 4 times) speed-up, e.g. [10]–[12], both with centralized scheduling. We do not aim for completeness in the coverage of either all currently popular architectures or of all the implications for each architecture, but want to point out the more salient issues that need to be addressed.

5.1 Input-queued (VOQ) with centralized arbitration

In a purely input-queued architecture with VOQ a centralized arbitration unit (scheduler) is required. The architecture consists of three basic parts, which can straightforwardly be mapped to the generic architecture shown in Fig. 1: the input line cards containing the VOQs, the routing fabric (usually a crossbar or parallel crossbars), and the scheduler. We assume that the switch core comprises both the routing fabric and the scheduler. Consequence I implies that multiple racks are required; therefore, at least some, possibly all, of the line cards are at a significant distance from the switch core. The RTT resulting from this distance has some interesting implications. There are two basic approaches to address this issue:

First, the VOQs can be brought as close as possible to the core, either on the same card as the scheduler or on other cards in the same rack. Both options suffer from two drawbacks: First, N extra chips with VOQ, buffering, routing lookup, and flow control are required, thus inefficiently duplicating functionality already implemented in the adapters on the line cards, and, second, Consequence V is violated because the extra chips add to the overall power consumption. Moreover, the first option is clearly not practical except for very small systems because of the space, power, and bandwidth limitations of the card, whereas the second option is very costly because many extra cards (and possibly extra racks) are required.

In the second approach, the VOQs and their associated buffers remain on the line cards, and the scheduler maintains VOQ state information for all N^2 VOQs in the system. The VOQs do not send requests to the scheduler, but rather communicate the arrivals (encoded as a tuple consisting of packet source, destination, and traffic class).[4] The scheduler performs the bookkeeping, computes the matchings based on its local image of the VOQ state information, and sends the corresponding grants.

However, the scheduler is not aware of the arrivals in the last RTT/2 packet cycles, and is therefore likely to compute a

[4]Such an incremental state-update scheme raises the issue of error robustness.

sub-optimal matching, even if the actual matching algorithm is optimal. The impact of the RTT on performance in such a system under various types of traffic requires further study.

Furthermore, in an architecture with a centralized scheduler, communication with all line cards must be tightly coupled because every matching is one-to-one and the core is bufferless, i.e., the request-grant-accept process must be synchronized on a system-wide level. This implies that the RTT between card and switch core must be equal for all line cards. This can be achieved either by using equally long cables or by compensating for cable-length differences by means of delay lines, where all links must be matched to the *longest* cable.

This architecture also suffers from increased latency because packets at the ingress cannot proceed before the request-grant-accept process has been completed. As a result, even packets arriving at an empty input without any contention for their destination have to wait for at least one RTT until the grant arrives.

5.2 CIOQ with limited speed-up

CIOQ architectures with a limited speed-up typically comprise VOQs at the ingress, a bufferless switch core with a centralized scheduler, and output queues at the egress. The switch core runs at a speed-up S times faster than the external line rate, where S is typically between 2 and 4.

The main implication for this type of architecture is that the chip and card bandwidth bottleneck problem is exacerbated by a factor of S because the entire core, including the internal links to the line cards, must be run S times faster (see Fig. 1). Given Consequences II and V this is prohibitively expensive in both hardware (TX/RX macros and cables) and power. Roughly speaking, beyond 1 Tb/s the physical size of the core must grow by a factor of S to accommodate the additional bandwidth.

A possible solution is integrate the VOQs (ingress side) and/or the output queues (egress side) and the switch core, but again this is only feasible for small systems because of space and power limitations.

This architecture also suffers from a possible performance penalty because of the RTT between scheduler and line cards, and faces the same system-wide synchronization challenge as mentioned in Sec. 5.1.

Finally, the introduction of a significant RTT has implications for the egress buffer sizing. Even if there is no internal speed-up (as in Sec. 5.1), egress buffers usually are still necessary to accommodate packets when the downstream links are blocked (e.g., because of congestion in the downstream node). However, if the downstream link is 100% available, any input should be allowed to utilize this bandwidth fully, without the possibility of packets being lost, i.e., under completely unbalanced traffic both work-conservation and losslessness must be ensured. The egress buffer size per output must be scaled proportionally to RTT, speed-up, and number of traffic classes to satisfy these requirements.

6. CONCLUSION

The traditional approach to switch design puts the high-level architecture first and optimizes for performance at this level. The resulting architecture is then mapped to a physical packaging. Although this approach has worked well so far, we have identified a number of trends that, when combined, present a severe disruption at about 1 Tb/s. Beyond that throughput, switch fabrics must be distributed, which introduces a critical dependency on link technology and causes the round-trip time to become an intra-switch issue. However, most importantly, we argue that the switch architecture must be developed starting from the packaging and power requirements, to which the desired architecture must then be suited, instead of the other way around.

7. REFERENCES

[1] A.G. Wassal and M.A. Hasan, "Low-Power System-Level Design of VLSI Packet Switching Fabrics," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 20, no. 6, Jun. 2001, pp. 723-738.

[2] http://www.telcordia.com, "Network Equipment-Building System (NEBS) Requirements: Physical Protection," GR-63-CORE.

[3] http://www.telcordia.com, "Electromagnetic Compatibility & Electrical Safety," GR-1089-CORE.

[4] J. Bolaria and B. Wheeler, "A Guide To Switch Fabrics," Feb. 2002, The Linley Group.

[5] H.B. Bakoglu, "Circuits, Interconnections, and Packaging for VLSI," Addison-Wesley Pub. Co., Reading, MA, 1990.

[6] http://www.npforum.org

[7] H.T. Kung and A. Chapman, "The FCVC (Flow Controlled Virtual Channel) Proposal for ATM Networks," in *Proc. Int. Conf. Network Protocols*, San Francisco, CA, Oct. 1993, pp. 116-127.

[8] C.M. Özveren, R. Simcoe, and G. Varghese, "Reliable and Efficient Hop-by-Hop Flow Control," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 4, May 1993, pp. 642-650.

[9] McKeown, N., "The iSLIP Scheduling Algorithm for Input-Queued Switches," *IEEE/ACM Trans. Networking*, vol. 7, no. 2, Apr. 1999, pp. 188-201.

[10] I. Stoica and H. Zhang, "Exact Emulation of an Output Queueing Switch by a Combined Input Output Queueing Switch," in *Proc. 6th IEEE/IFIP IWQoS '98*, Napa Valley, CA, May 1998, pp. 218-224.

[11] S.-T. Chuang, A. Goel, N. McKeown and B. Prabhakar, "Matching Output Queueing with a Combined Input Output Queued Switch," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 6, Jun. 1999, pp. 1030-1039.

[12] J.G. Dai and B. Prabhakar, "The Throughput of Data Switches with and without Speedup," in *Proc. INFOCOM 2000*, Tel Aviv, Israel, Mar. 2000, vol. 2, pp. 556-564.