A Four-Terabit Single-Stage Packet Switch with Large Round-Trip Time Support

F. Abel, C. Minkenberg, R. Luijten, M. Gusat, and I. Iliadis

IBM Research, Zurich Research Laboratory, CH-8803 Ruschlikon, Switzerland

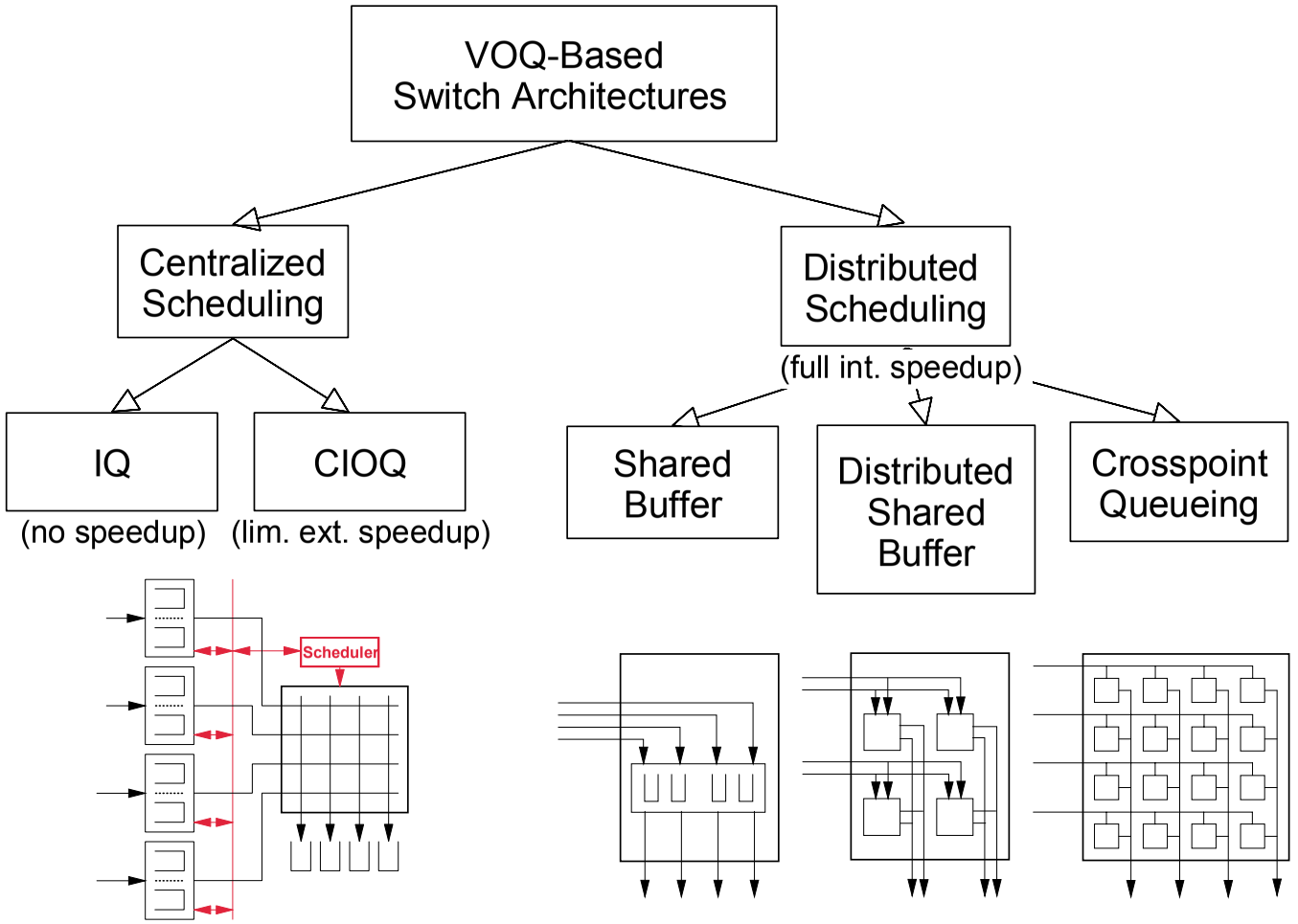
Motivation

- Merchant switch market
  - Achieve coverage of wide application spectrum: MAN/WAN/SAN
- *Can a versatile switch architecture be designed to achieve this?*
- Requires:
  - High performance for different protocols and QoS requirements
  - Allows very little assumptions about traffic properties

Outline

- Current single-stage switch architectures
- Preferred architecture
- Physical implementation of a 4 Tb/s switch
- Simulated performance: 256 x 256 system
- Conclusions

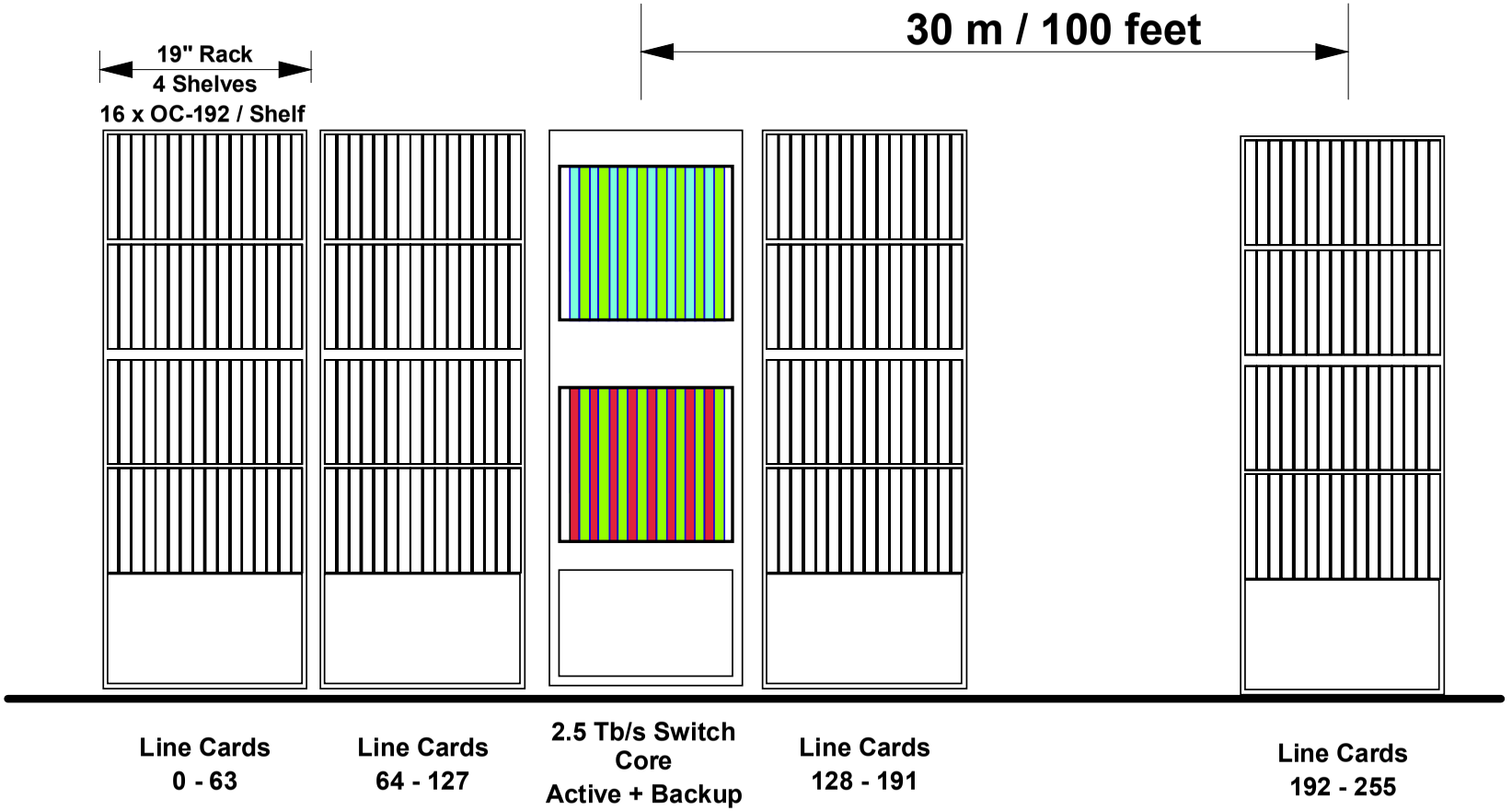
Current Single-Stage Switch Architectures



Selection of the Preferred Architecture

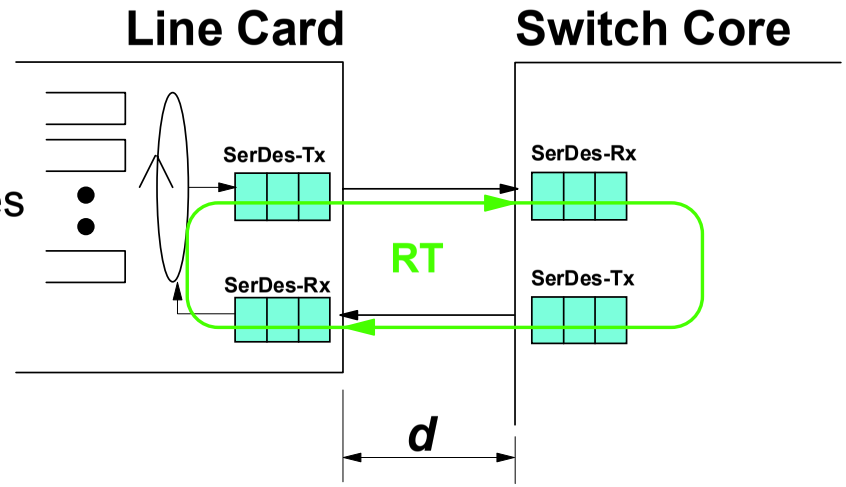
- Initial focus on high-level architecture issues
- Equally significant aspects arise when actually building the system
  - Physical system size
    - Multi-rack packaging, interconnection, clocking and synchronization are required
  - Power has become a tremendous challenge and a major design factor
    - Typically required: 2 kW per rack, 150 W per card, 25 W per chip
  - Switch fabric (SF) internal round trip (RT) has significantly increased
  - Switch core (SC), line cards (LC) and VLSI chip packaging
- Significant consequences for system cost, power and practical implementation

Size of a Terabit-Class System



Switch-Fabric-Internal Round Trip (RT)

- RT = Number of cells in flight:
  - $RT_{total} = RT_{cable} + RT_{logic}$
    - RT_{cable} = cells in flight over backplanes and/or cables
    - RT_{logic} = cells pipelined in arbiter and SerDes (Serializer/Deserializer) logic



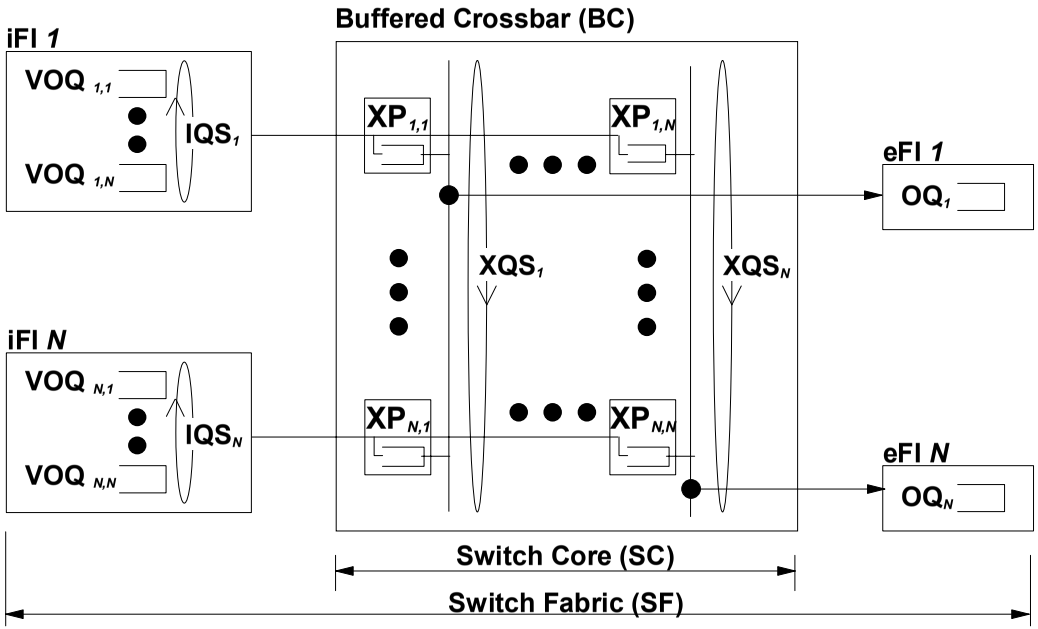
- RT has become an important SF-internal issue because of:
  - Increased physical system size
  - Increased link speed rates
  - SerDes circuits are now widely used to implement high-speed I/Os

Line rate	OC-12	OC-48	OC-192	OC-768
Interconnect distance	1 m	1 m	6 m	30 m
Interconnect type	backplane	backplane	cable	fiber
Packet duration	512 ns	128 ns	32 ns	8 ns
Round Trip	<< 1 cell	~ 1 cell	16 cells	64 cells

Evolution of RT

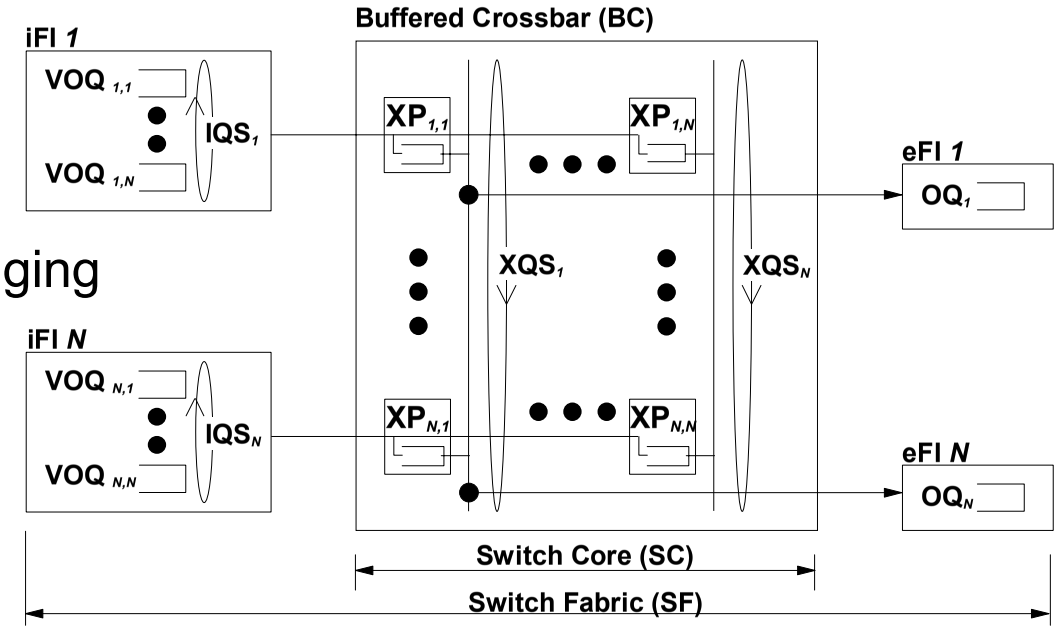
Preferred Architecture (1/2)

- Combined input- and crosspoint- queued (CICQ) architecture
  - Decoupling of the arrival and departure processes
  - Distributed contention resolution over both inputs and outputs
  - Close to ideal performance is achieved without speedup of the SC
  - Memories are operated at the line rate



Preferred Architecture: CICQ (2/2)

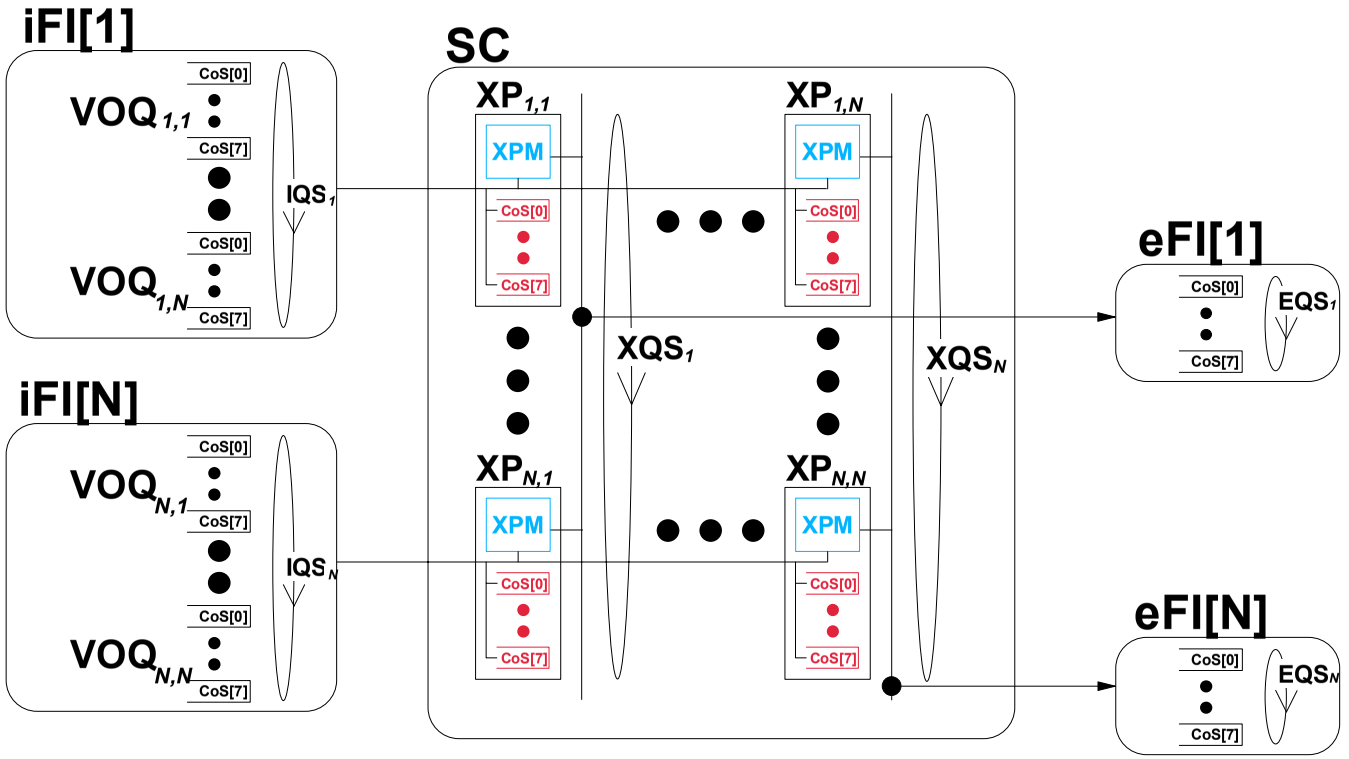
- Advantages:
  - Performance and robust QoS of OQ switches
  - A buffered crossbar is inherently free of buffer hogging
  - A buffered SC enables hop-by-hop FC instead of end-to-end
  - Reduced latency at low utilization



- Distribution of OQs exhibits some of the fair queueing properties
  - Fair bandwidth allocation (e.g. with a simple Round-Robin)
  - Protection and isolation of the sources from each other

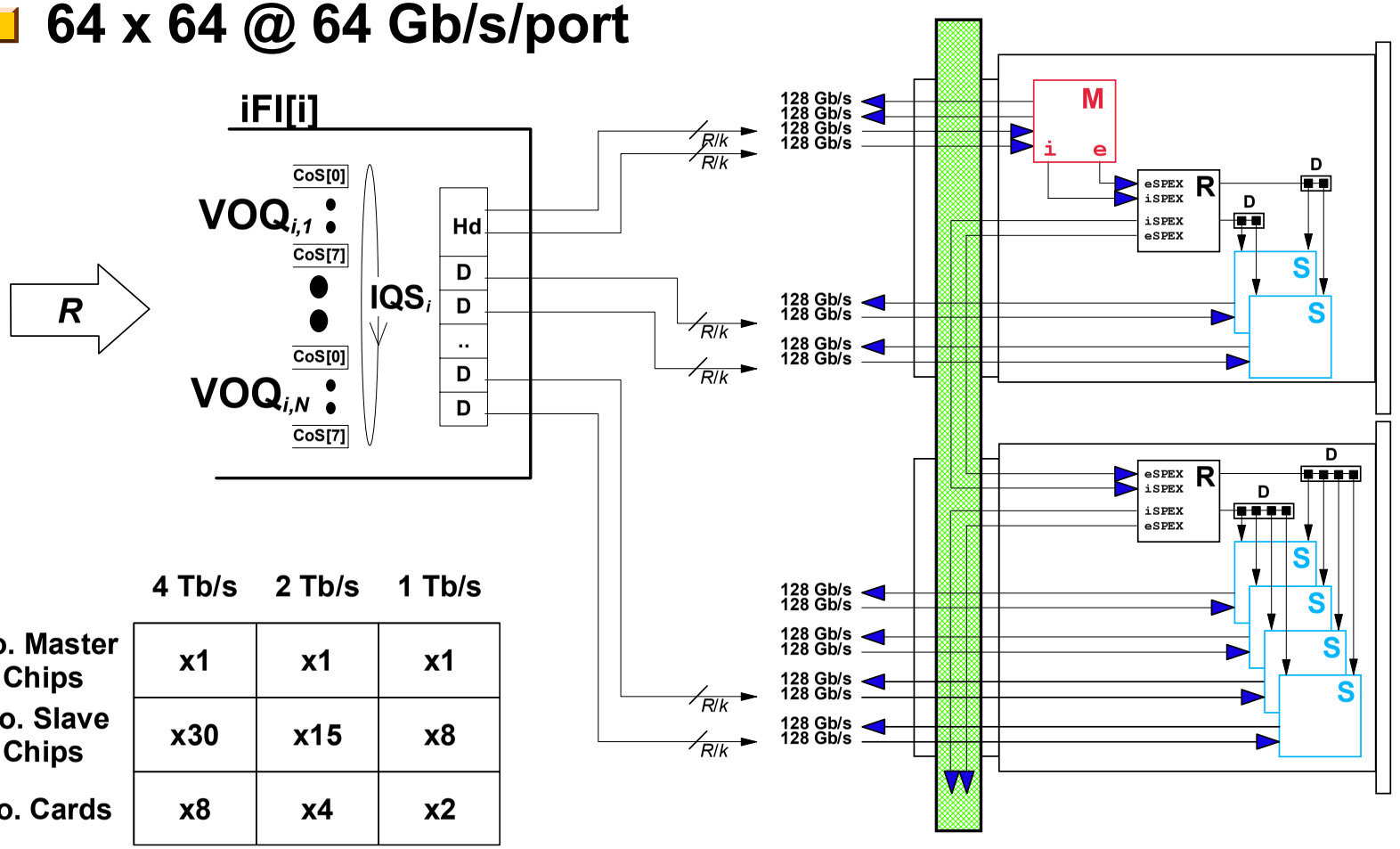
CICQ and CoS Support

- Selective queueing at each queueing point (iFI, SC, eFI)
- Service scheduling in addition to contention resolution (IQS, XQS)
- Additional scheduler at the egress (EQS)



CICQ and Parallel Sliced Switching

- 64 x 64 @ 64 Gb/s/port



	4 Tb/s	2 Tb/s	1 Tb/s
No. Master Chips	x1	x1	x1
No. Slave Chips	x30	x15	x8
No. Cards	x8	x4	x2

Crosspoint Buffer Dimensioning (1/2)

- Bandwidth (on the links) is becoming the scarce resource
  - Hence
    - Utilization must be maximized
    - Link speedup should be avoided as much as possible
  - Assuming a credit-based FC and a commun. channel with an RT of τ cells

    ☛ τ credits are required to keep link busy

- *Do we also need τ credits per XP ?*
  - Traffic agnostic principle:
    - The bandwidth of each flow can vary on an instantaneous basis
  - Link utilization principle:
    - Full utilization of the link bandwidth must be achieved in the absence of other flows
  - To provide 100% throughout under any traffic condition:
    - A minimum of τ cells are required per XP to ensure that any input can transmit to any output at any instant and at full rate.

      (e.g. in the case of fully unbalanced traffic or in absence of output contention)

Crosspoint Buffer Dimensioning (2/2)

- RT evaluation
  - $RT_{cable} = 2dR / S_{light}C_{size} \simeq$ 30 cells
    (with R = 64 Gb/s, d = 30 m, S_{light} = 250 Mm/s (over the dielectric), C_{size} = 512 bits)
  - $RT_{logic} \simeq$ 30 cells (estimated by design)
  - $RT_{total} \simeq$ 60 cells
- Buffer requirement (assuming RT_{total} = 64 cells, C_{size} = 64 B)
  - Per logical XP: $XPM_{size} = \tau = (64 \times 64)$ = 4 kB
  - Total for the switch core: $N^2 \times \tau$ = 16 MB
- $XPM_{size} = \tau$ = 64 cells provides:
  - 100% throughput under contentionless traffic and d = 30 m / 100 feet
  - 100% throughput under uniform traffic and $d \simeq$ 3 km / 10.000 feet

VLSI Implementation

- CMOS 0.11-μm, Std. cell design, 2.5 Gb/s SerDes
- Slave chip (64x64@2Gb/s/port)
  - 200 mm^2, 20 W, 750 SIOs
- Split master chip (48x48@4Gb/s/port)
  - 225 mm^2, 28 W, 825 SIOs

Simulated Performance

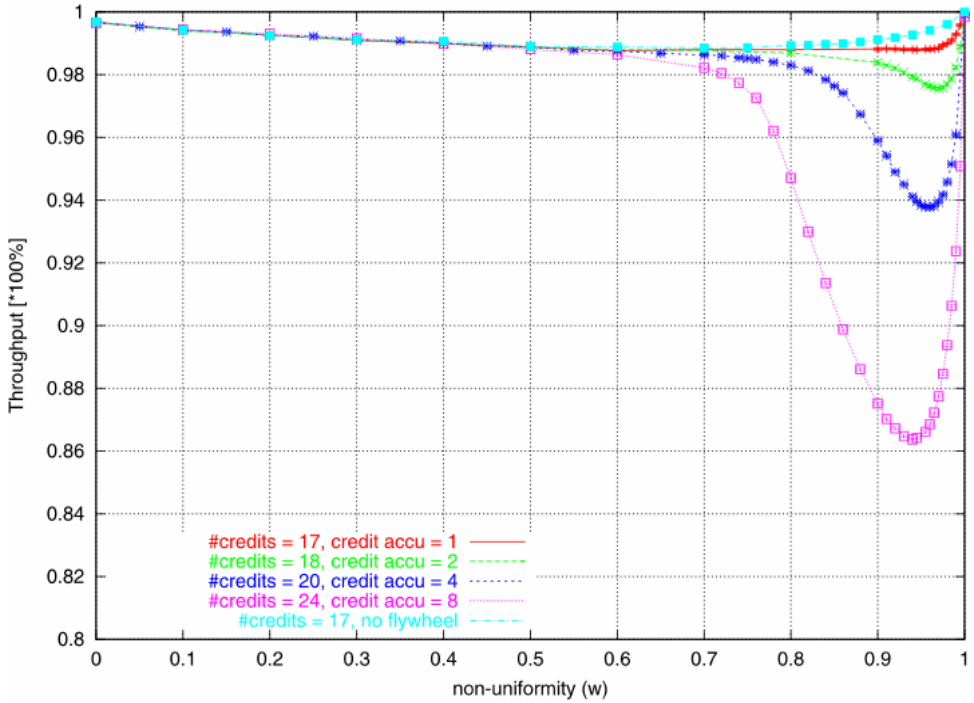
- **Parameters:**
  - 256 x 256 CICQ switch fabric
    - 64 x 64 SC with 4 external ports (OC-192) per SC port (OC-768)
    - 64/128 cells per XP partitioned into 4 areas of 16/32 cells
    - Ingress and egress link RT = 64 cells (at the OC-768 level)
    - Line card egress buffer = 4 x 256 cells
  - CoS
    - 8 classes of service (C_0 is the highest, C_7 is the lowest priority)
    - Uniform distribution, i.e. 12.5% of offered traffic per class
    - Strict priority scheduling throughout the system (iFI, SC, eFI)

Non-Uniform Traffic

- Non-uniform traffic: We adopt the distribution used by *Rojas-Cessa et al.* (*HPSR 2001*) where:

  $\lambda_{i,j} = \lambda(w + (1-w)/N)$ **if** $i = j$, $\lambda(1-w)/N$ **otherwise**

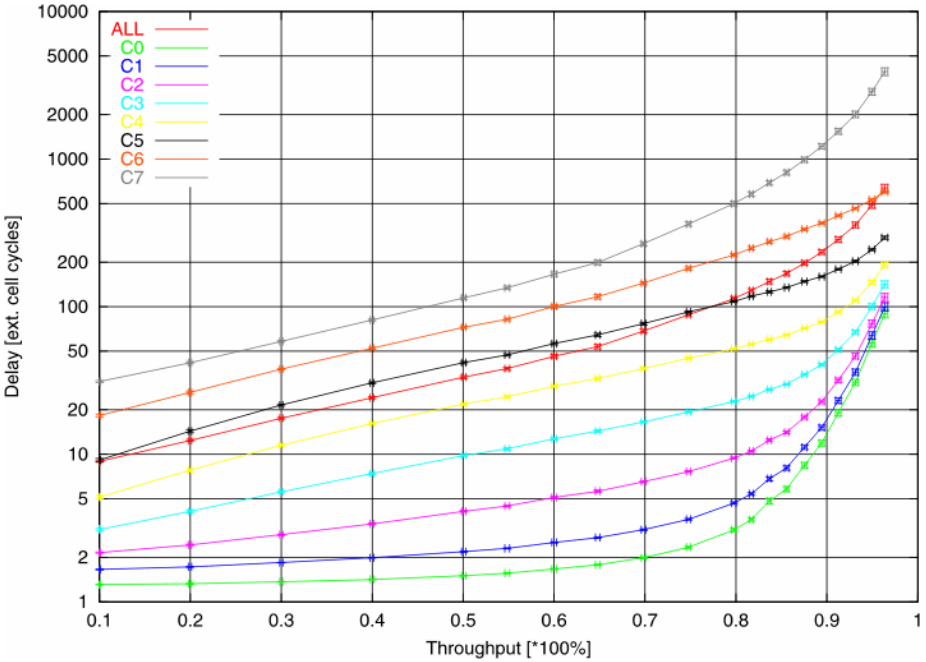
  - N is the number of ports (256), $\lambda_{i,j}$ is the traffic intensity from input i to output j
  - λ is the aggregate offered load (100%), w is the non-uniformity factor



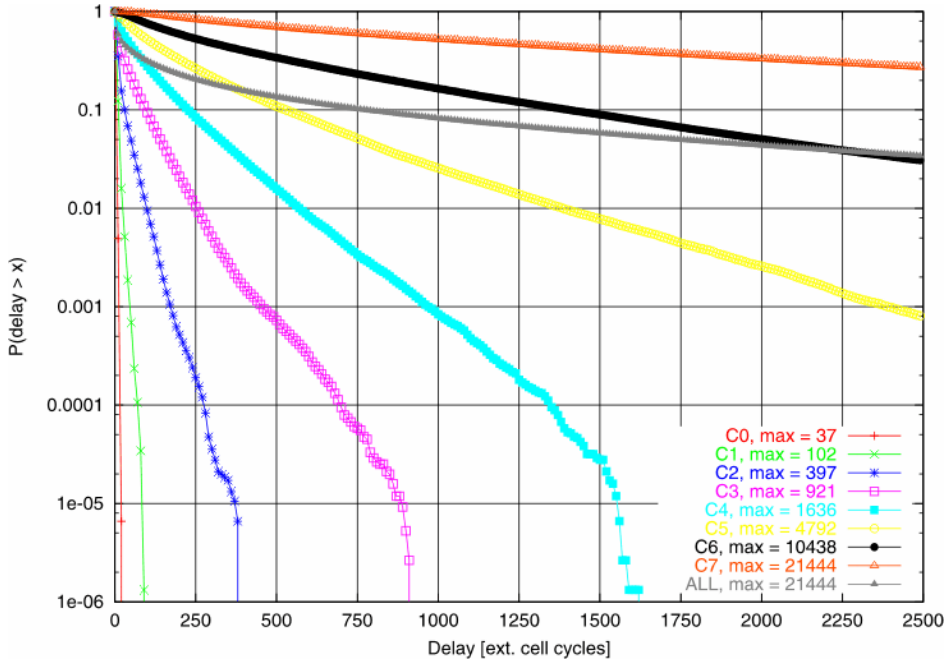
XPM = 68/72/80/ cells

Uniform Traffic

- Uniform traffic
  - Uniformly distributed bursts over all 256 destinations
  - Geometrically distributed bursts



XPM = 64 cells / Bursts = 30 cells



XPM = 128 cells / Bursts = 30 cells

Conclusions

- System design and implementation are equally important as performance considerations
  - Impact of power, packaging, links, RT
- Traffic agnosticness requirement in OEM
  - CoS support
- CICQ architecture is a viable solution
  - Scalable
- Demonstrated sizing
  - VLSI implementation of a single-stage 4 Tb/s switch
  - Excellent performance

Contacts

- IBM Prizma research team
  - http://www.zurich.ibm.com/cs/powerprs.html
- IBM PowerPRS™: Switch fabric products
  - http://www-3.ibm.com/chips/products/wired/products/switch_fabric.html