Switch Fabric Basics

References

- Light Reading Report on Switch Fabrics, available online at: http://www.lightreading.com/document.asp?doc_id=25989

- Title: Network Processors Architectures, Protocols, and Platforms
  Author: Panos C. Lekkas
  Publisher: McGraw-Hill

- Multi-Gigabit Serdes: The Cornerstome of High Speed Serial Interconnects, Genesys Logic America, Inc.

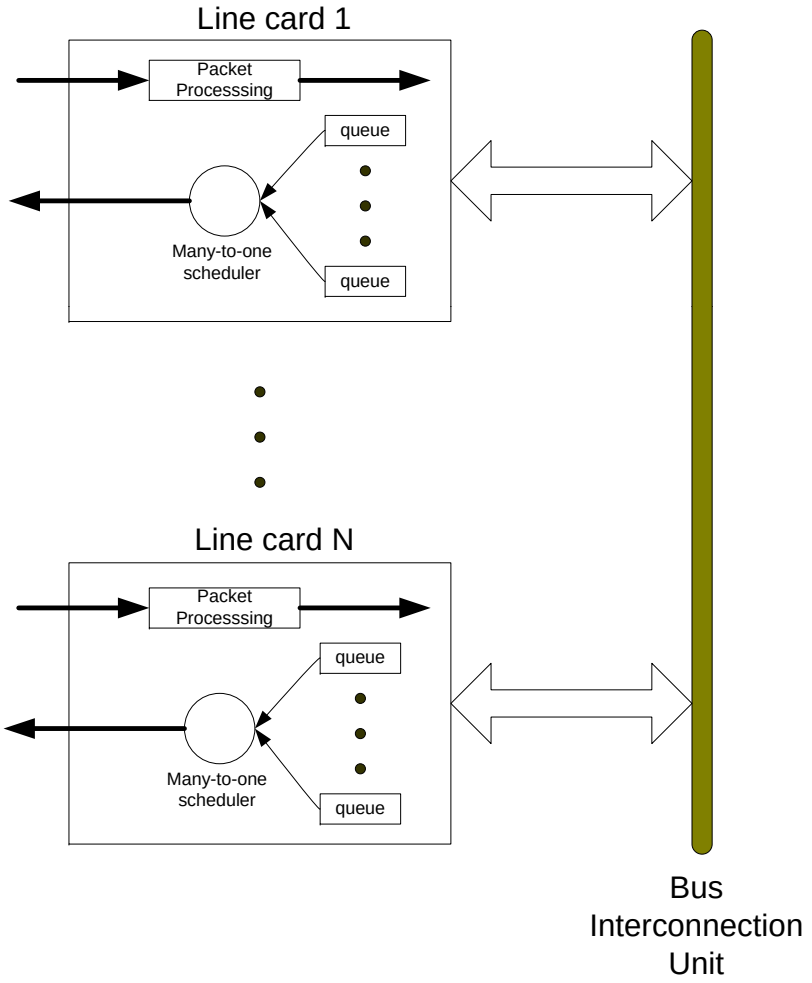
- C. Minkenberg, R. P. Luijten, F. Abel, W. Denzel, M. Gusat, Current issues in packet switch design, ACM SIGCOMM Computer Communication Review, Volume 33 , Issue 1 (January 2003)

Architecture of a Switch

- Data comes in and goes out of the router through line cards.
- Inside the router data should move from the ingress line card to the egress line card.
- How can we do that?

Output Queued Switches

- Every line card can immediately send the arrived data to the egress line card.
- All buffering (queueing) is done at the output side.
- Each line card do the scheduling of its out going data locally and independent of other line cards.
- Scheduling is a many-to-one selection problem.
- We can use well known and studied scheduling algorithms.



Why people like Output Queued Architecture

- It is a very modular and distributed architecture.
- We only need buffering at the output side.
- It is a work conserving architecture and no blocking.
- Scheduling is many-to-one and there are extensively studied (WFQ, WRR, …).

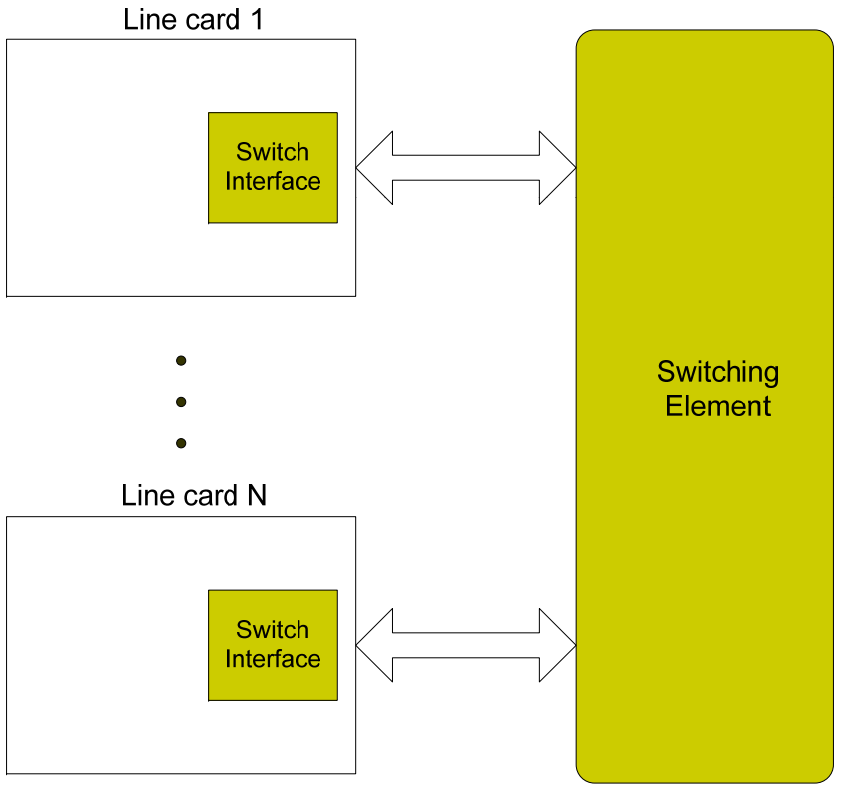
What is the problem with output queued architecture

- The interface between the line cards should run N times faster than the line cards.
  - The interface could be a bus that works N times faster
  - Alternatively we can have a full mesh connection between the line cards.
  - Neither approach is scalable.

- The output memory should work N+1 time faster than the line card
  - N line cards write into the memory
  - 1 read from the memory.
  - It is not scalable.

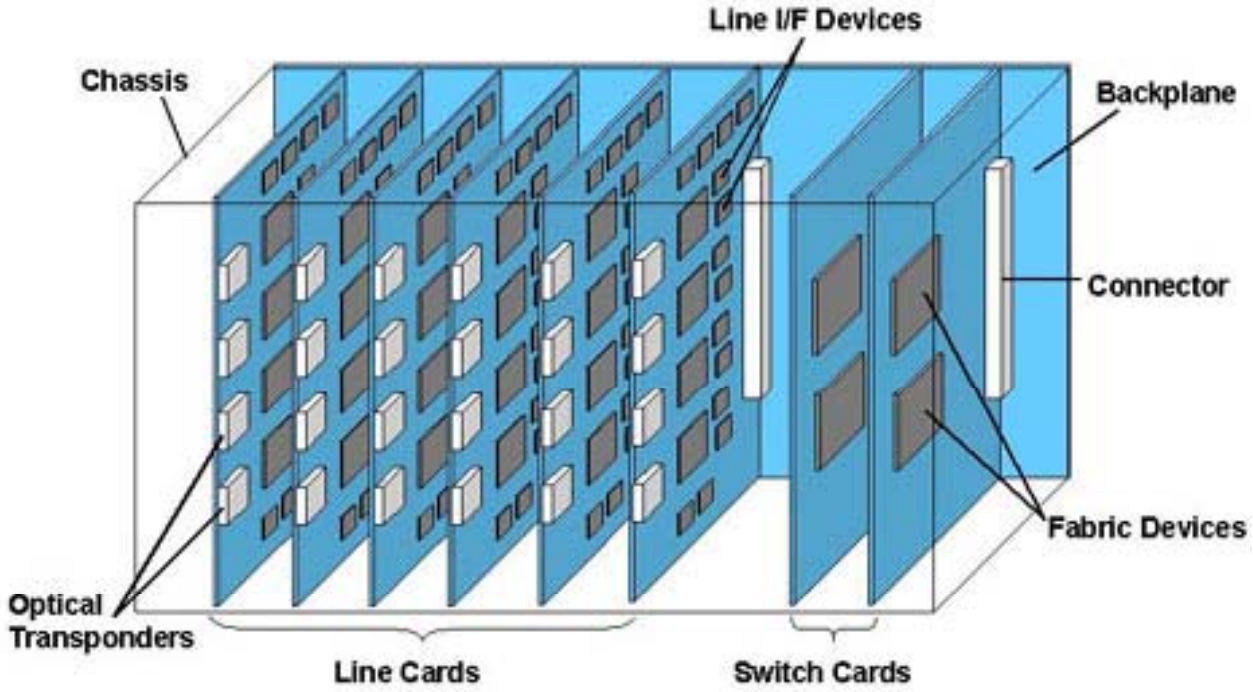
Switch Based Architectures

- There is an intelligent switching element that transfer cells from input side to the output side.
- The interface does not need to work N times faster.
- We may need buffering at both input and output side.
- We usually have an extra switch interface unit element on the line card.
- We need multiple levels of scheduling and buffering
  - Ingress line card
  - Egress line card
  - Switching element



Line Card and Switch Cards

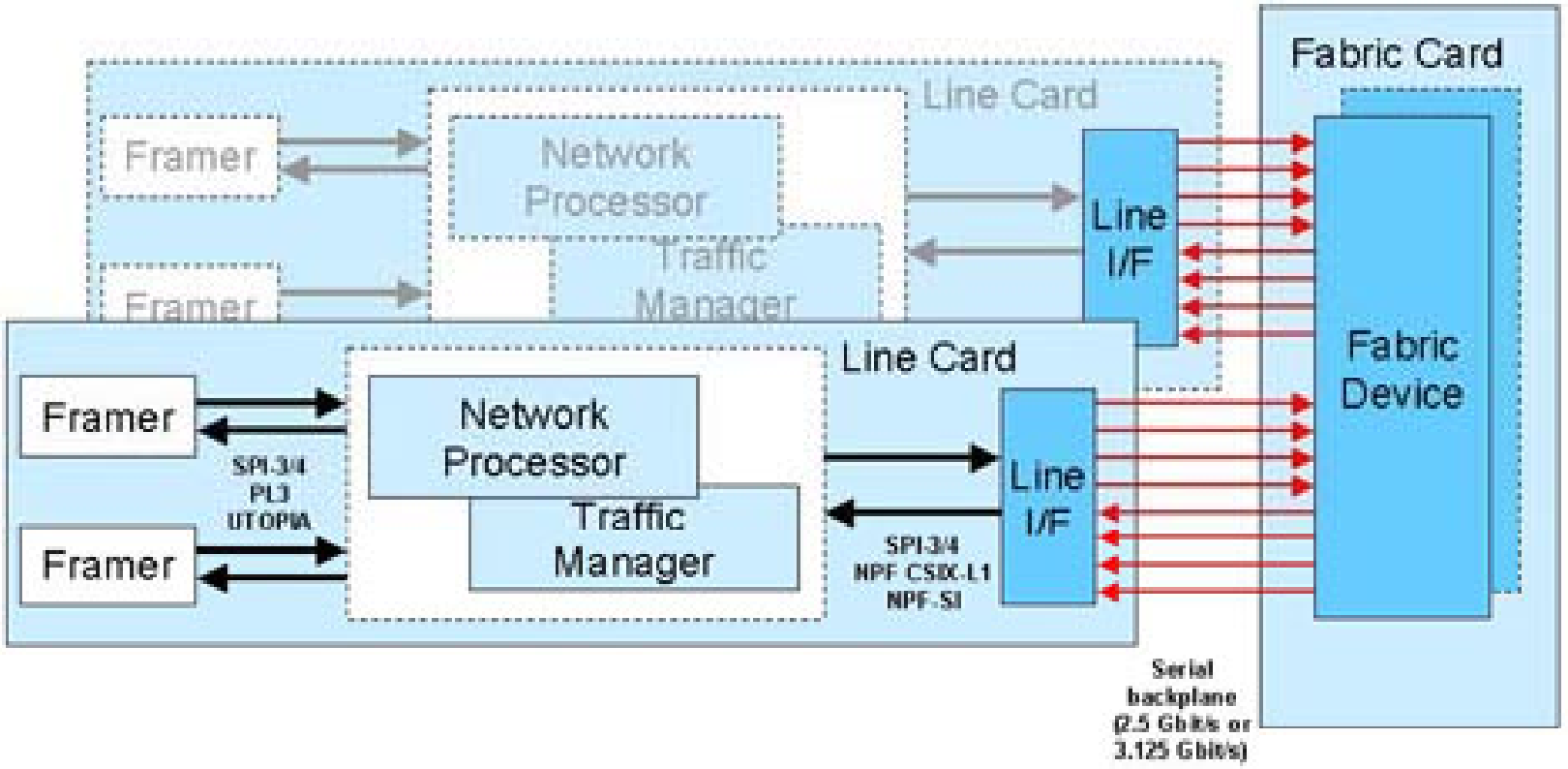
- There are multiple switch cards on the system.
- Connection between line card and switch cards are through backplane traces.



Source: http://www.lightreading.com/document.asp?doc_id=25989

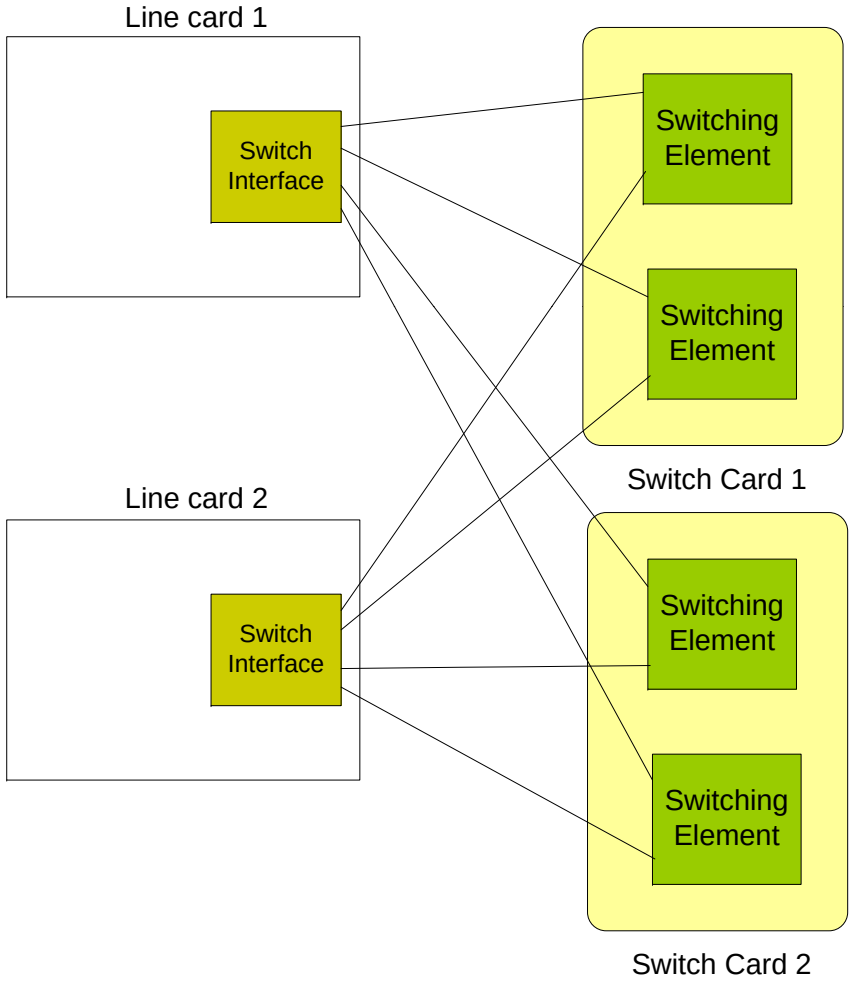
Line Card and Switch Cards

- The data rate over the backplane traces are limited.
- Each line card requires multiple traces to achieve required data rate.



Multiple Switch Chips and Cards

- Consider that we need to have 4 serdes connection from each line card to get desired data rate.
- This means that we need 4 switching elements.
- If we can put 2 switching elements per switch card, then we need 2 switch cards.
- How many traces over the backplane?
- What if we have 2 more line cards?

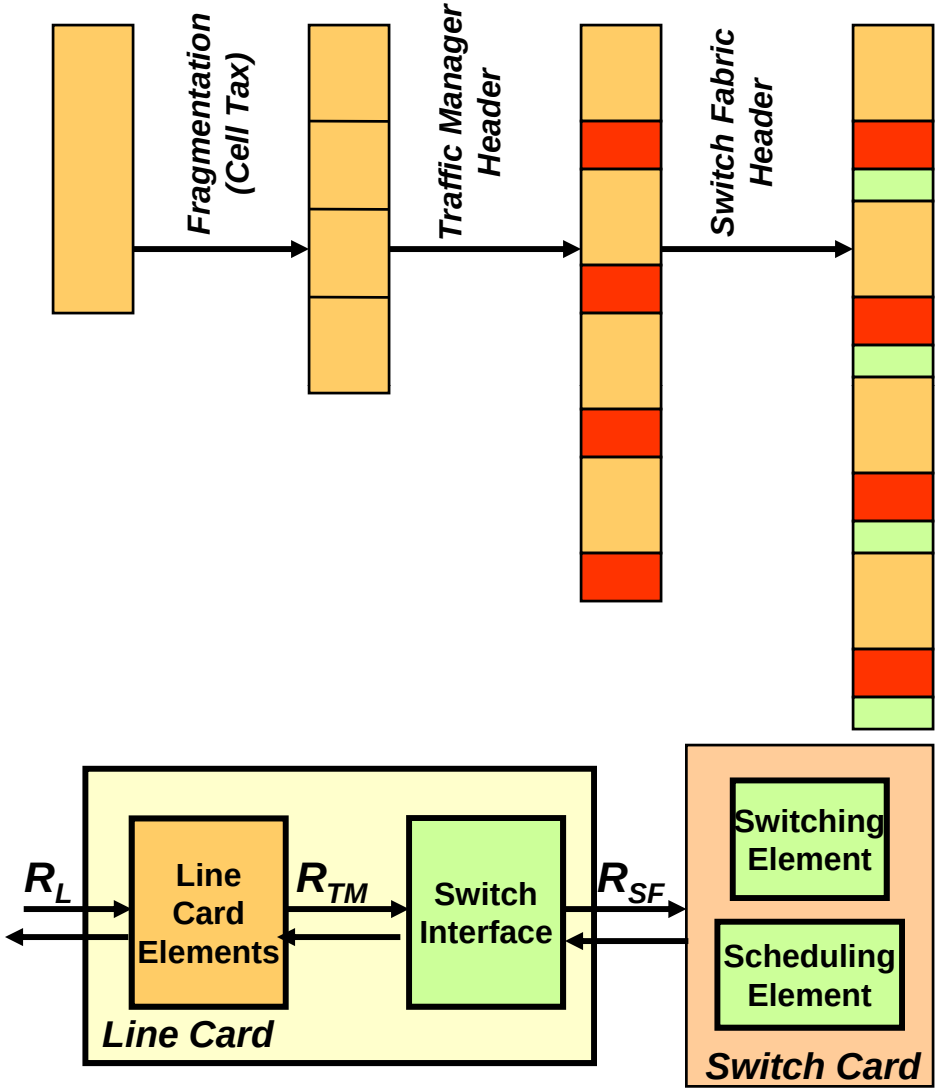


How many serdes do we need?

- How fast should be the connection between switch card and line card?
- The line speed is not enough.
- Switch fabric throughput is less than 100% due to contention.
- Network Processor, Traffic manager and switch fabric add their headers.
- There is also cell tax.

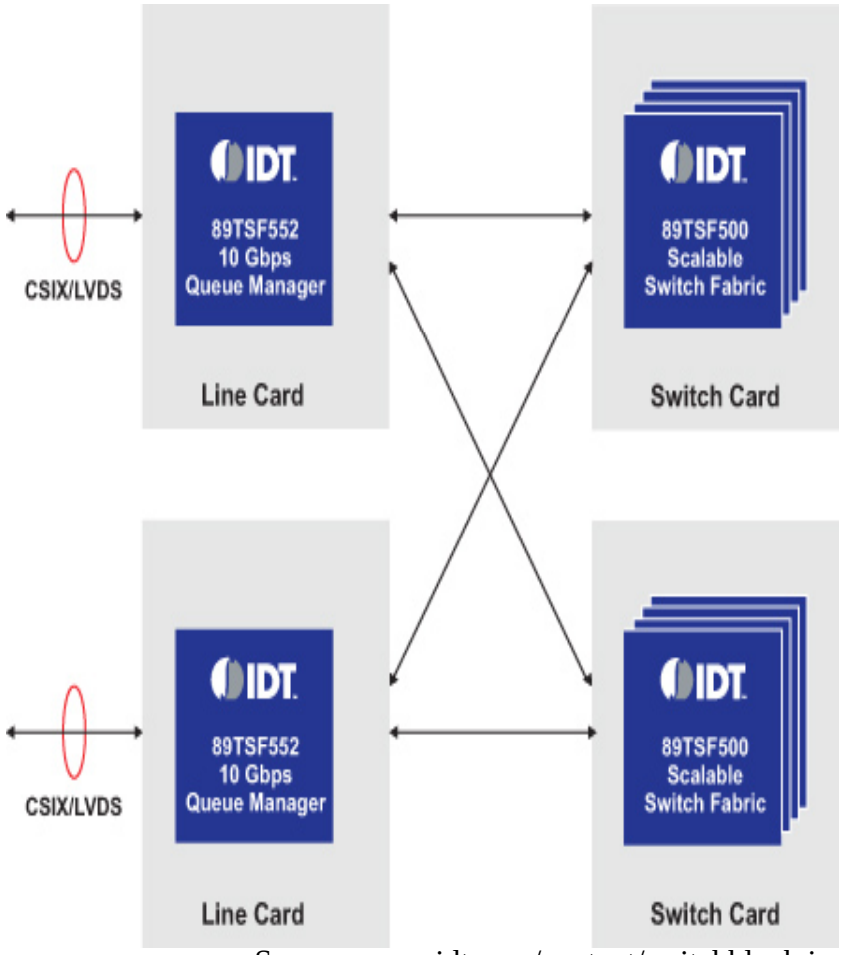
Speedup

- Speedup = R_{SF}/R_{TM}
- In the commercial systems, speedup usually refers to R_{SF}/R_L.
- Higher speedup factor:
  - Increases system design complexity.
  - Increases power consumption.
  - Creates signal integrity issues.
- Required Speedup factor is around 2



Redundancy

- We have spare switch cards and control cards in the system.
- The redundancy models:
  - Passive redundancy (N:1) We have one inactive switch card in the system that starts to work after failure.
  - Passive redundancy (1:1, N:N) for each active switch card, we have one inactive card.
  - Load-Sharing Redundancy (N-1) all cards are active and when a failure happens and the performance will degrade gracefully.
  - Active Redundancy (1+1): Two sets of fabrics carrying the same traffic.



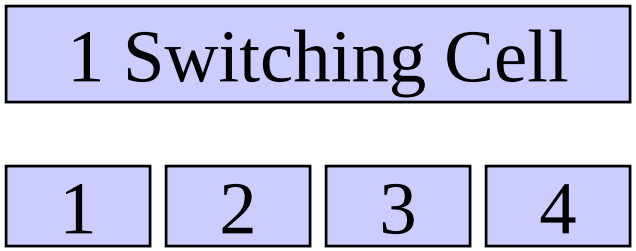
Source: www.idt.com/content/switchblock.jpg

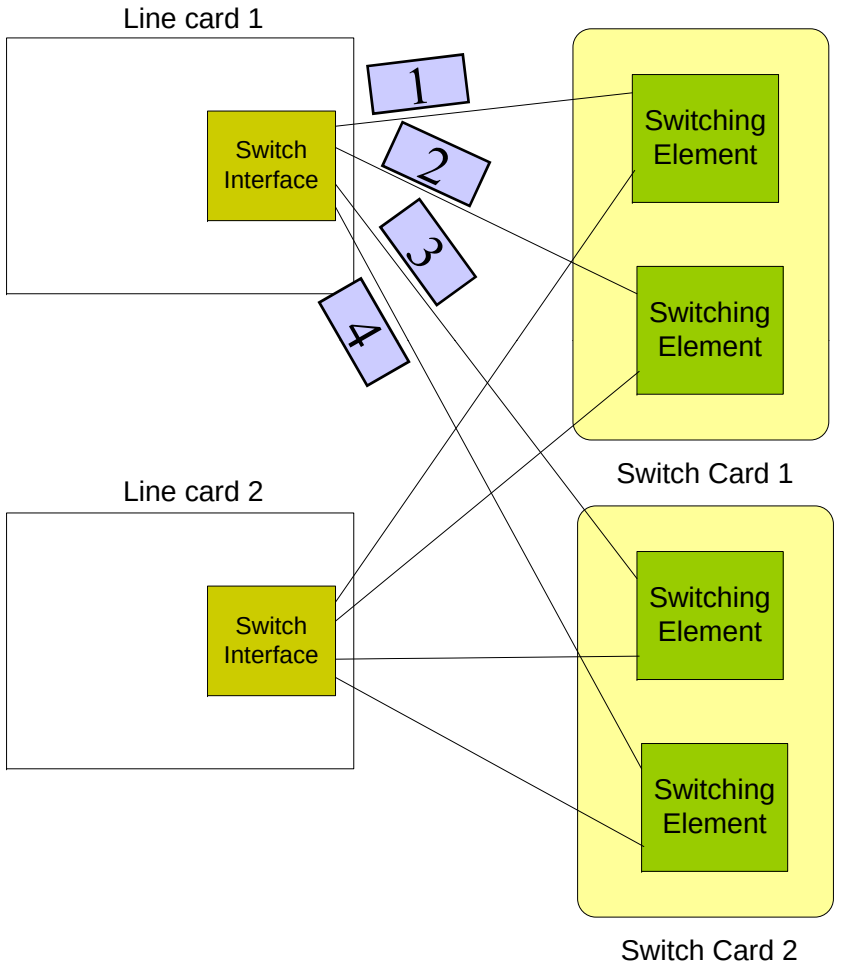
Switch Card Redundancy

- Note that redundancy must be switch card based.
- If we need two switch cards and 4 switch elements for normal operation.
- In N+1 redundancy model we need 3 switch cards and 6 switching elements.

Byte Slice Parallelism

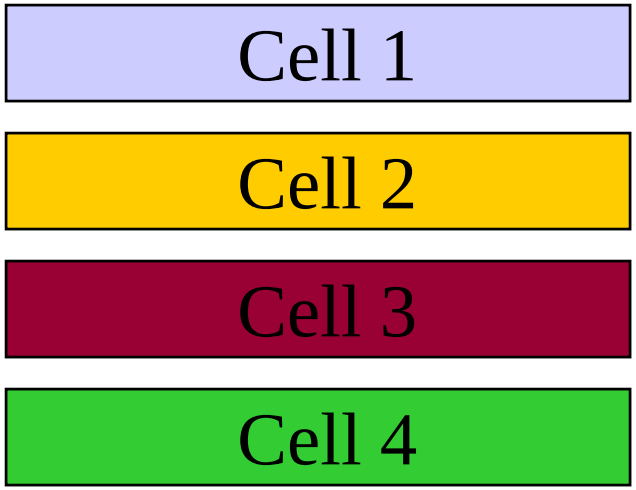
- In the byte slice parallelism switching elements carry different segments of the same cell in parallel.
- All switching elements should work synchronously.

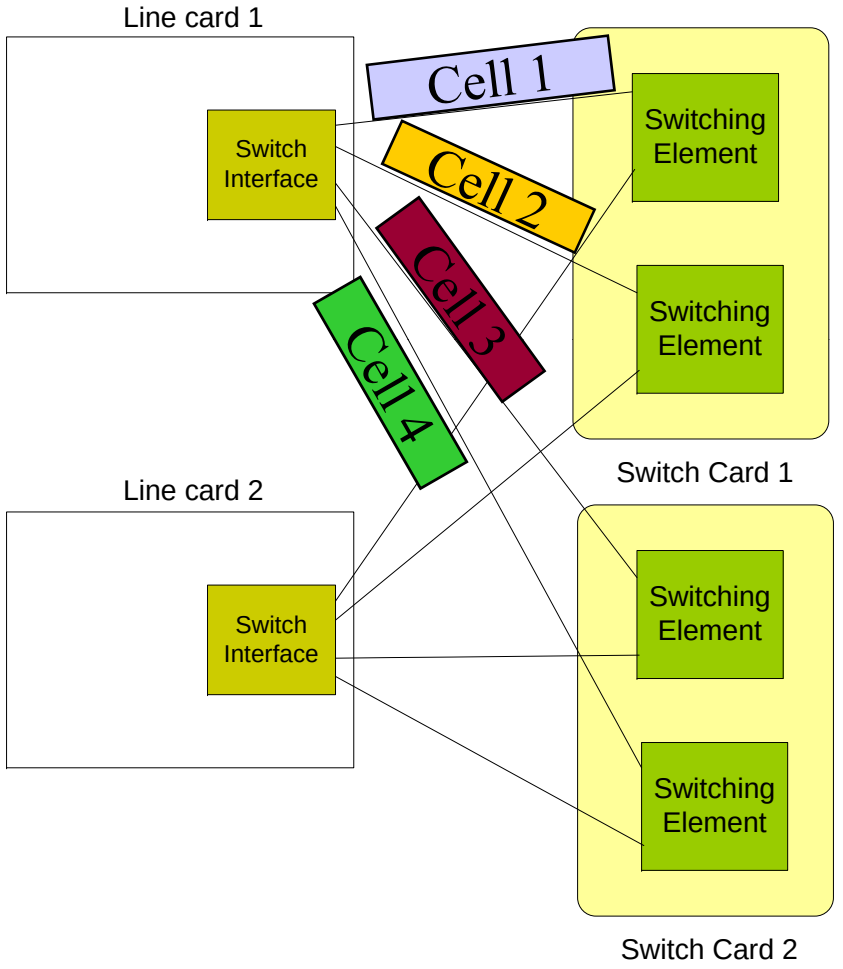




Cell Slice Parallelism

- In the Cell slice parallelism switching elements carry separate cells in parallel.
- Switching elements can work independently.





Cell vs. Byte Slice

- Can we have N-1 Redundancy with byte slice?
- Which architecture have more time for scheduling?
- Which architecture needs cell reordering at the egress side?
- How can we do load balancing in the cell slice model?
- Do we need load balancing in byte slice model?
- Which architecture requires coordination and synchronization among switch cards?
- If there is a failure in one switching element how many cells we loose in
  - Cell slice model?
  - Byte slice model?

Switch Fabric Requirements

- Support for QoS
  - Throughput
  - Delay
  - Jitter
- Support for multicast and broadcast
- Support for TDM traffic
  - Dynamically adjust the capacity mix in small increments
  - Low and very strict delay
  - ITU standard restrict delay to less than 150ms and OEMs want less than 10us delay through the switch fabric.
- High reliability
  - Graceful degradation: Failure reduces throughout but not the switching capability.
  - Lossless controlled switchover to redundant path.
  - Continuous monitoring of the data path integrity
- Backward compatibility
  - The interface between line-card and switch-card should be the same.
- Space: Fabric chip must fit in the switch cards (around 400 square inches)
- Power Dissipation of a fabric card can be around 250W.

Back Plane

- High-speed backplane connects line-cards and switch-cards.
- Back-plane consists of serial links providing point-to-point connection between the line-card and switch-cards.
- The back-plane carries
  - Packet Data
  - Flow-control messages
  - System management messages
  - Synchronizing clock signal
- We can have limited number of traces on the backplane.
- We need to use high-speed serial links to achieve the required speed.

Why serial and not parallel backplane connections?

- We need to limit number of traces.
- Large buses operating at relatively higher frequencies over long interconnect PCB causes problems:
  - Signal noise (cross talk and reflection)
  - Power
- Serial connection results in:
  - Area reduction (fewer traces and connections)
  - Noise reduction by using differential signals.
  - Better migration path to higher speeds

Back-plane high-speed serial connection

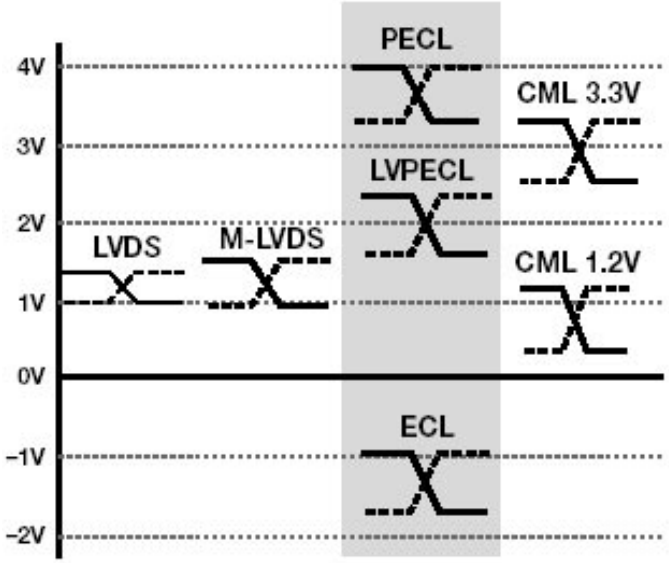
- This connection should pass through the Backplane.
- Serdes (Serializer-Deserializer) is used for this connection.
  - Each Serdes signal run over two wires and two pins (differential mode signal).
  - The speed is usually around 3.125 Gbps.
  - They usually run some sort of coding (8b/10b encoding)
    - Adds two bit at the start and end of each byte to assist clock recovery and maintain a DC balance.
  - The actual data rate would be around 2.5 Gbps.
  - There are attempts to provide 5-10 Gbps serdes.
- Serial link drivers:
  - PECL
  - LVDS (Low Voltage Differential Swing) 155Mbps-1.25Gbps
  - CML (Current Mode Logic) 600Mbps- 10Gbps

Serial Link Drivers

- There are three main differential signaling technologies:
  - PECL (Positive Emitter Coupled Logic)
  - LVDS (Low Voltage Differential Swing) 155Mbps-1.25Gbps
  - CML (Current Mode Logic) 600Mbps- 10Gbps

Parameter	LVDS	PECL (5V)	LVPECL (3.3V)	CML*
TX VOH	1.425V	4.0V	2.3V	VCC
TX VOL	1.075V	3.2V	1.6V	VCC—0.8V
TX VOD	350mV	800mV	0.7V	800mV
TX VOS	1.25V	3.6V	1.95V	VCC—0.4V
TX RT	100 Ohm	50 Ohm	50 Ohm	50 Ohm
RX VTH	±100mV	±100–200mV	±100mV	±50mV
RX VIN	GND to 2.4V	Depends	Depends	Limited

*CML numbers are shown for an 800mV output example; 400mV is also common.



Source: http://www.national.com/nationaledge/may03/article.html

Serial Link Drivers

	ECL	LVDS	CML
Bus Structure	Point-to-Point, Multidrop, Multipoint	Point-to-Point, Multidrop, Multipoint*	Point-to-Point
Power Dissipation	high	low	med
Speed	DC to >10Gbps	DC to >2Gbps	DC to >10Gbps
Coupling	DC or AC	DC	DC or AC
Process	Bipolar	CMOS, BiCMOS	Bipolar, CMOS

Source: http://www.national.com/nationaledge/may03/article.html

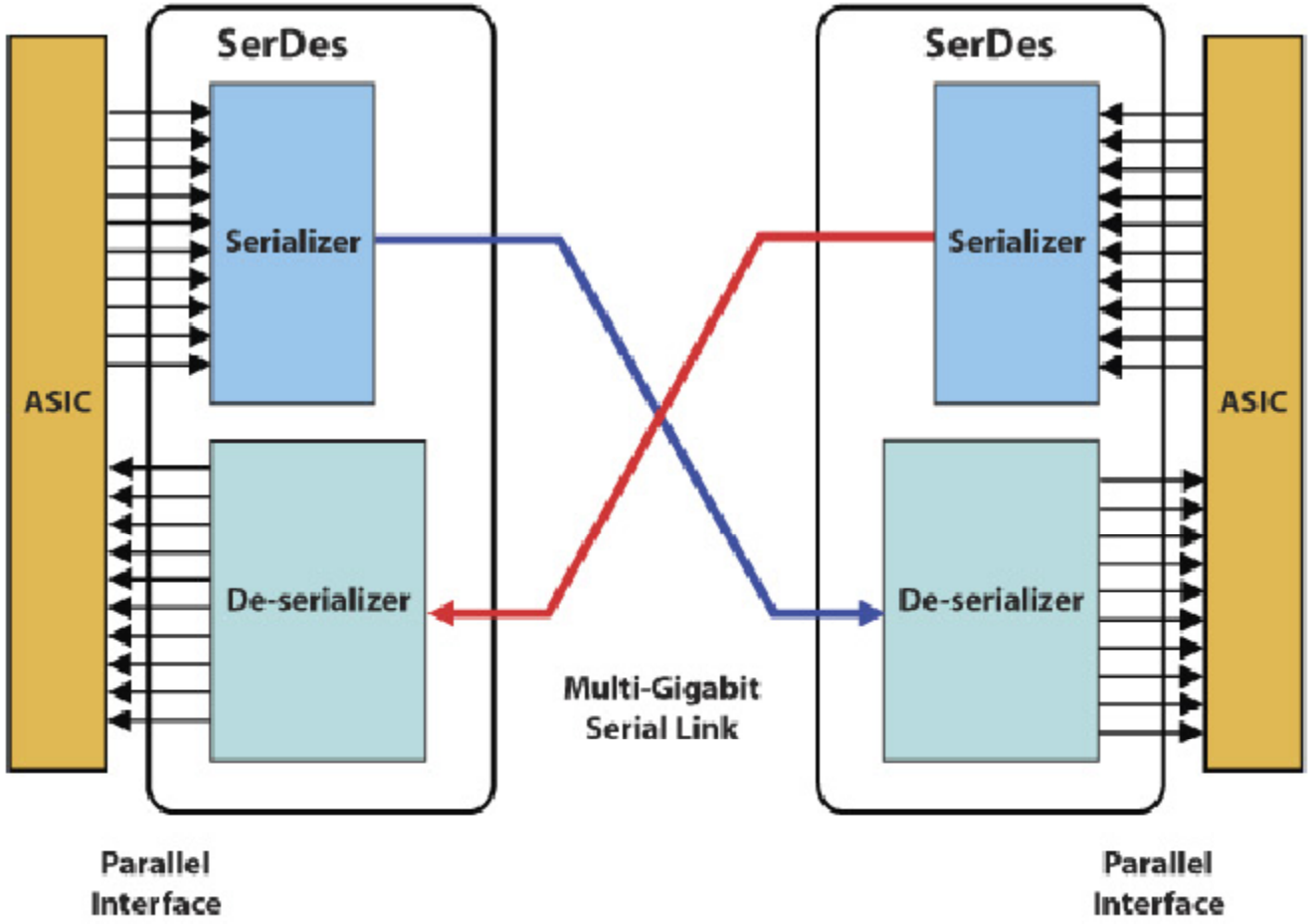
Back plane

- There is not enough space in one shelf for high speed in 19-21 inch shelves.
- Usually we can have 16 line cards, two switch card and one control card in one shelf.
- In multi-shelf systems, shelves are connected using optical fiber.
- Back plane can be designed for synchronous or asynchronous operation.
- Synchronous operation distributes a central clock across the backplane.
- Asynchronous operation requires a precise clock generator on each card (100 ppm).
- We can use idle cycles or cells to compensate for clock drifts.
- We use FIFO buffers before data passes boards clock domains.
- FIFO buffers also compensate for variable distance between the line and switch cards (specially in multi-shelf systems).

How Many Traces do we need?

- Typical LVDS speed is 1.25 Gbps, for 2.5 Gbps we need 2 channels.
- LVDS is differential, so we need 2 traces per channel
- LVDS is unidirectional, so we need 2 channels for full duplex
- Therefore, full duplex 2.5 Gbps, using LVDS requires 8 traces.
- We have to take care of channel alignment too.
- For an OC-48 line-card with 1:1 redundancy and 2X speedup we need 2.5x4 = 10 Gbps data rate.
- This translates into 8x4=32 traces per line-card.
- For 16 OC-48 line cards we need 32x16 =512 traces.
- For 16 OC-192 line cards we need 2048 traces.

Serdes



Source: http://www.genesyslogic.com/images_product/gbeserdes.pdf

Serdes Quality

- Jitter (affects the bit error rate)
  - PCI express with 400 ps bit time
    - Max. serialize output jitter 120 ps
    - Min. deserializer input jitter 240 ps
- Smaller size and lower power
  - Use same PLL for multiple SerDes cores
  - Distributing multi-gigahertz clock consumes a lot of power and causes signal integrity concerns.
- Testability
  - Serdes should have built-in self test (BIST) functions.
  - Serdes usually offer Pseudo Random Bit Sequence (PRBS) pattern generator in the serializer and pattern checker in deserializer.
  - Jitter injection filter