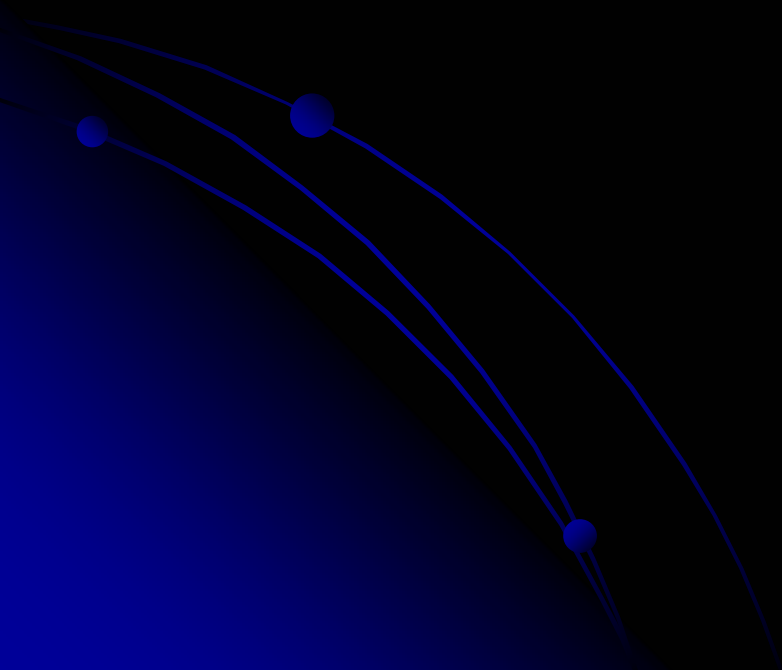


Scaling Internet Routers Using Optics

Producing a 100TB/s Router

Ashley Green and Brad Rosen

February 16, 2004



Presentation Outline

➤ **Motivation**

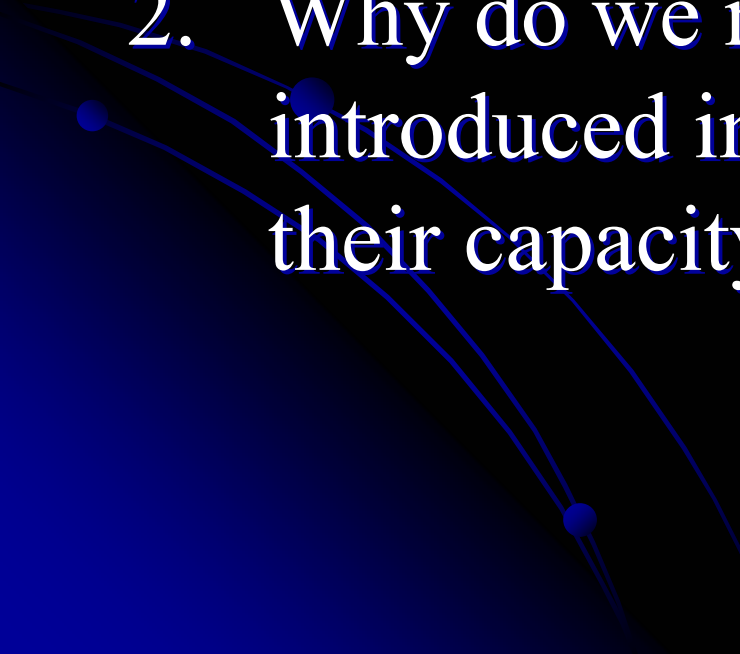
➤ **Avi's "Black Box"**

➤ **Black Box: Load Balance Switch**

➤ **Conclusion**



Motivation

1. How can the capacity of Internet routers scale to keep up with growths in Internet traffic?
 2. Why do we need optical technology to be introduced inside routers to help increase their capacity?
- 

Motivation

Q: How can the capacity of internet routers scale to keep up with growths in Internet traffic?

A: Since the underlying demand for network capacity continues to double every year, we require an increase in router capacity.

My notes: The other option would be for Internet providers to double the number of routers in their network each year, but this would not be practical because the number of central offices would have to be doubled each year and doubling the number of locations would require enormous capital investment and increases in support and maintenance infrastructure.

Motivation

Q: Why do we need optical technology to be introduced inside routers to help increase their capacity?

A:

Each generation of router consumes more power than the last. Have reached the limit for single-rack routers.



Move towards multi-Rack systems. Systems Spread power over Multiple racks, but Have unpredictable Performance and bad Scalability.



Optics?

Motivation

- As we saw with multi-rack routers, throughput is often unpredictable.
 - Predictable throughput limited by
 - switching capacity: 2.5 Tb/s
 - Power consumption

Hence: optics offer virtually zero power consumption and 100Tb/s in a single rack

Presentation Outline

➤ Motivation

➤ Avi's "Black Box"



Black Box: Load Balance Switch



➤ Conclusion

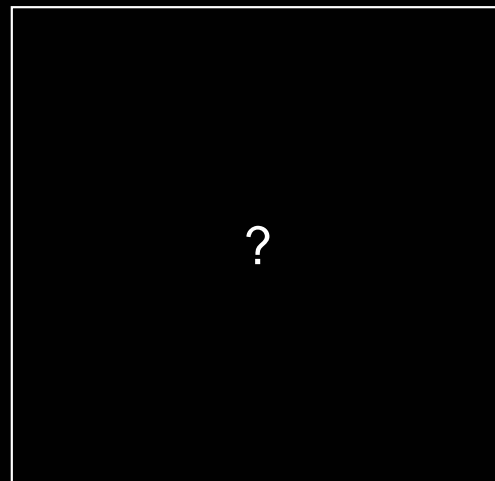


Avi's Black Box

Input

Output

Bits [packets]



...where they should be
...in order [mostly]
...on time [mostly]

1. Good scalability
2. 100% throughput
3. Low power consumption
4. Sequenced packets
5. Fault tolerance (missing or failed line cards)

Presentation Outline

➤ Motivation

➤ Avi's "Black Box"



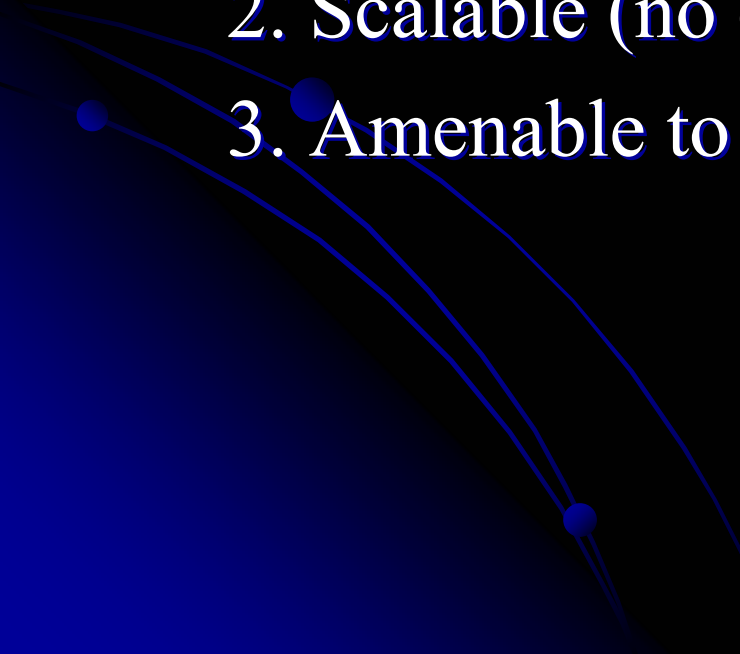
Black Box: Load Balance Switch



➤ Conclusion

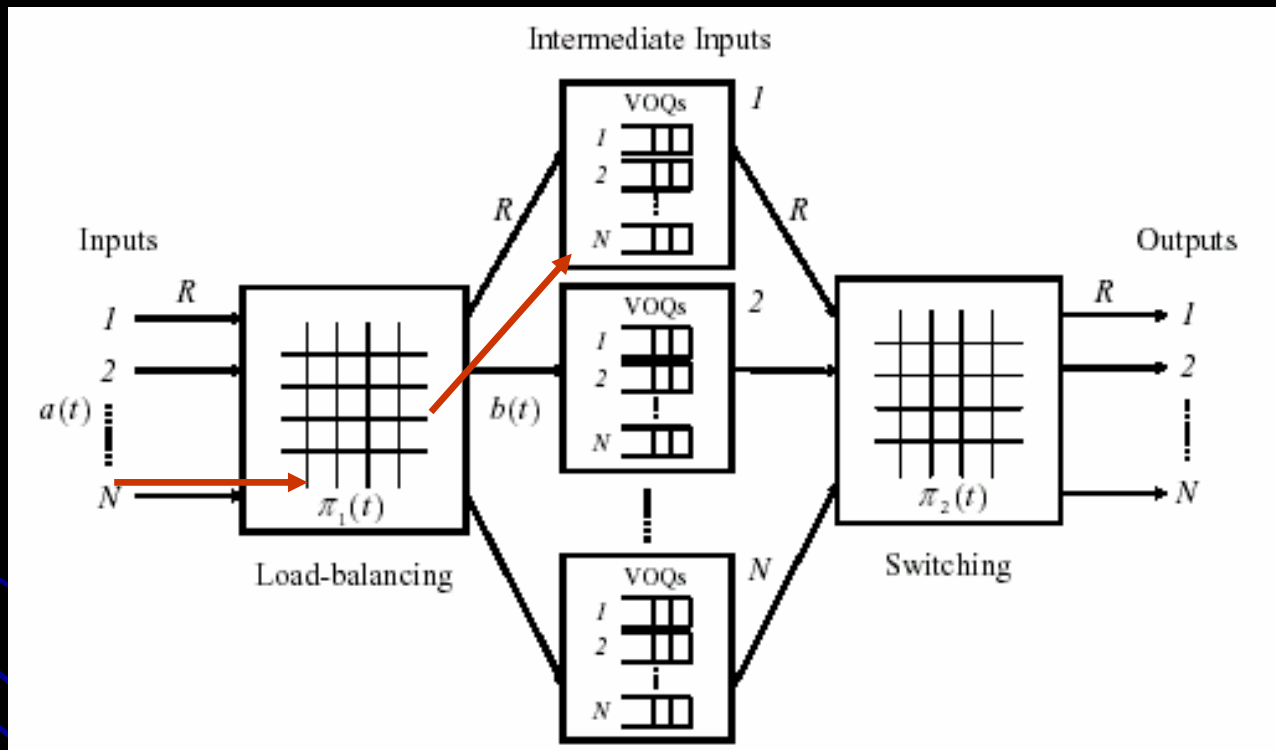


Black Box: Load-Balanced Switch

- Most promising architecture to fulfill the role of the magic black box
 1. Has 100% throughput
 2. Scalable (no central scheduler)
 3. Amenable to optics
- 

Black Box: Load Balancing Switch

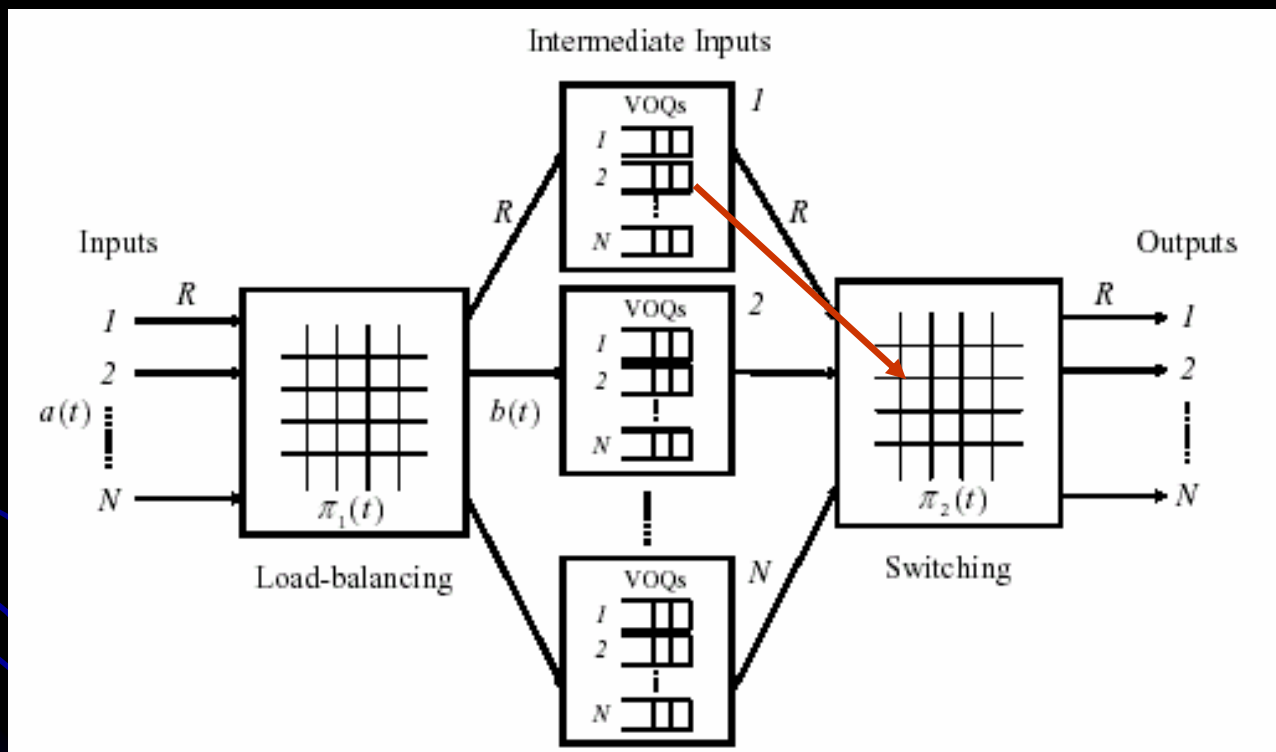
Step 1: when a packet arrives to the first stage, the first switch transfers it to a VOQ.



The intermediate input that packet goes to depends on current configuration of the load-balancer.
The packet is put into the VOQ according to eventual output.

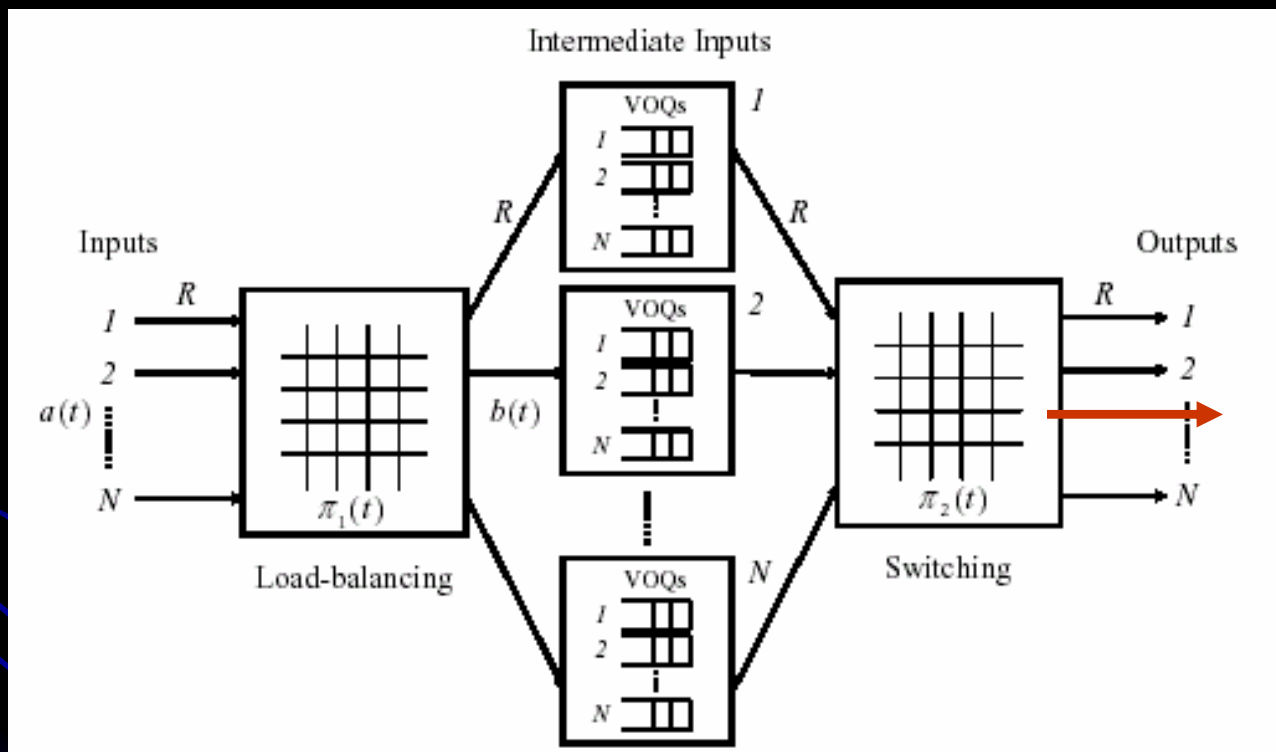
Black Box: Load Balancing Switch

Step 2: VOQ is served by second fixed switch.

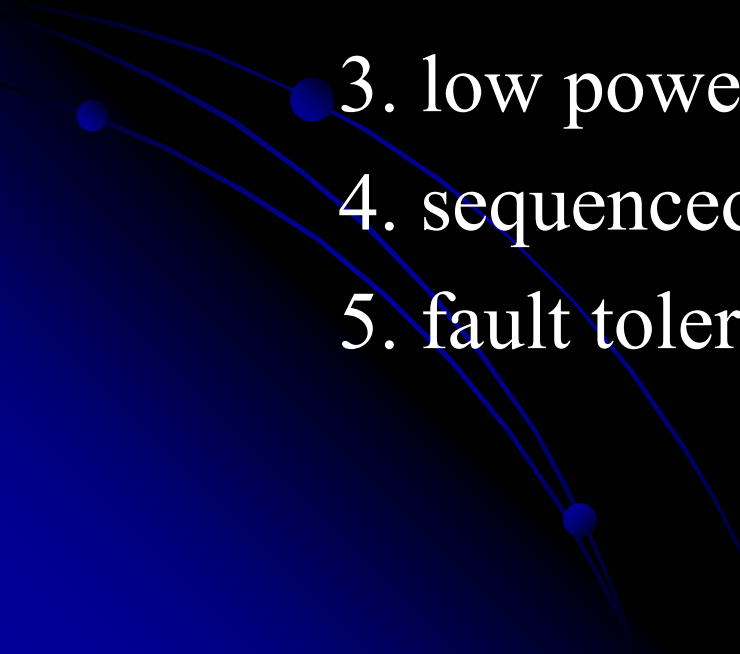


Black Box: Load Balancing Switch


Step 3: the packet is transferred across the second switch to its output and then depart from system.



Black Box: Load Balancing Switch

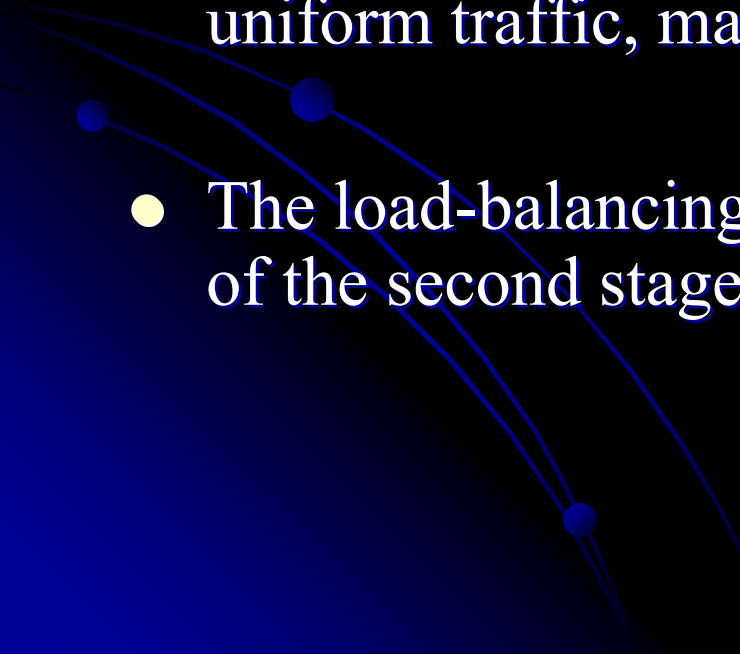
- Why does the architecture guarantee the following expected outputs of the black box scenario?
 1. scalability
 2. 100% throughput
 - 3. low power consumption
 4. sequenced packets
 5. fault tolerance
- 

Black Box: Load Balancing Switch Scalability

- Achieves scalability via quantity
 - 100TB/s Router accomplished via 640 line cards.
 - Pushes the complexity into the mesh
 - [This is where optics are used]
- 

Black Box: Load Balancing Switch

100% Throughput

- If packet arrivals are uniform, a fixed, equal-rate switch with virtual output queues, has a guaranteed throughput of 100%.
 - Real network traffic is not uniform.
 - So, an extra load-balancing stage is needed to spread out non-uniform traffic, making it uniform to achieve 100% throughput
 - The load-balancing device spread packets evenly to the input of the second stage: fixed, equal-rate switching.
- 

Black Box: Load Balancing Switch

100% Throughput [proof outline]

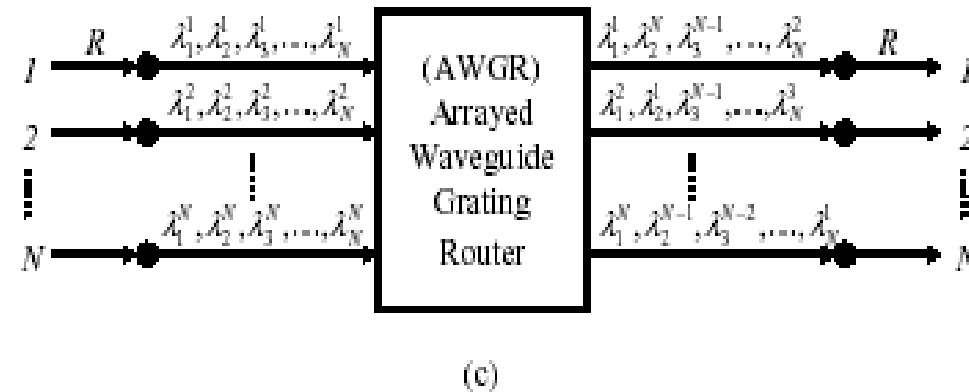
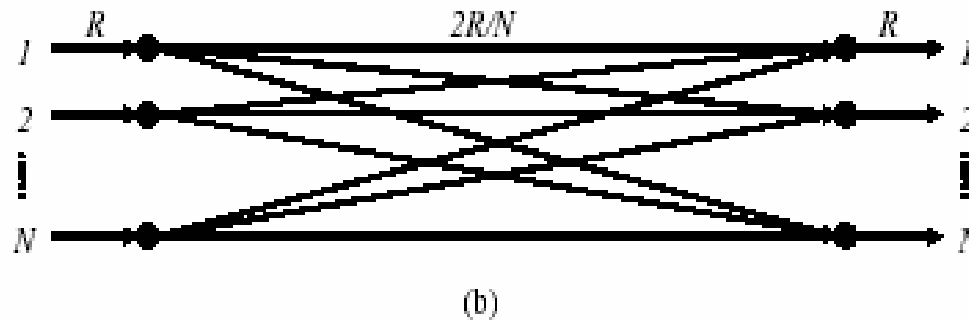
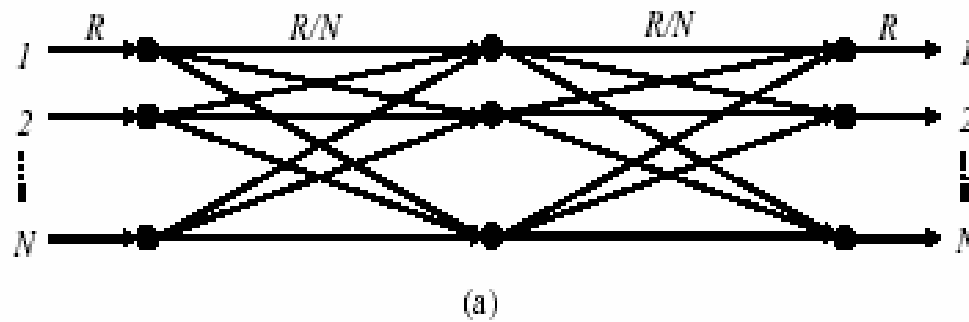
The load balanced switch has 100% throughput for non uniform arrivals for the following reason:

Consider the arrival process $a(t)$ with $N \times N$ traffic Matrix Λ to the switch. This process is transformed by the sequence of permutations in the load balancer $\pi_1(t)$ into the arrival process to the second stage $b(t) = \pi_1(t) * a(t)$. The VOQs are served by the sequence of permutations in the switching stage, $\pi_2(t)$. If the inputs and outputs are not over-subscribed then the long term service opportunities exceed the number of arrivals, and hence the system achieves 100% throughput.

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T (b(t) - \pi_2(t)) = \frac{1}{N} e \Lambda - \frac{1}{N} e < 0,$$

where e is a matrix full of 1's.

Switch Reconfigurations



Black Box: Load Balancing Switch

Low Power Consumption

- Switch fabric of this architecture enables low power consumption because...
 1. It is essentially transparent
 2. consumes no power
 3. eliminates power-hungry conversions between the electrical and optical domain.

Replace fixed, equal rate switch with N^2 fixed channels at rate R/N . Replace two switches with a single switch running twice as fast. [similar logic for meshes]

Black Box: Load Balancing Switch

Sequenced packets

- The load-balancer spreads packets without regard to their final destination or when they depart
 - eg. Two packets arrive at same time, they are spread to different intermediate linecards. It is possible their departure order will be reversed.
- Solution: Full Ordered Frames First (FOFF).
 - Geared towards this 100TB/s routing
 - Bounds the difference in lengths of the second stage VOQs
 - Resequencing in third stage buffer [because number of packets out of sequence is bounded]
 - FOFF is run locally at linecard, using only available info at that linecard.

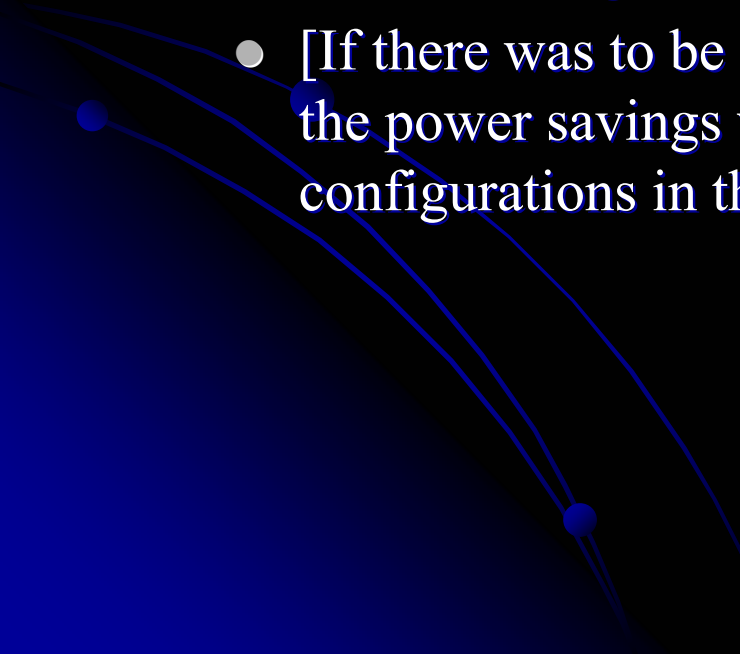
FOFF

- Input I maintains N FIFO queues $Q_1 \rightarrow Q_N$. An arriving packet destined to output J is placed in Q_j .
- Every N time slots, the input selects queue to serve for the next N time slots. First, it picks round robin from among the queues holding more than N packets. If there are no such outputs, then it picks round-robin from among the non-empty queues. Up to N packets from the same queue, and hence destined to the same output, are transferred to different intermediate line-cards in the next N time slots. A pointer keeps track of the last intermediate line card that we sent a packet to for each flow; the next packet is always sent to the next intermediate line card.

If there is always at least one queue with N packets, the packets will be uniformly spread over the second stage and there will be no mis-sequencing. ALL the VOQ's that receive packets belonging to a flow receive the same number of packets, so they will all face the same delay and won't be missequenced. Missequencing only arises when no queue has N packets; but the amount of mis-sequencing is bounded and is corrected in the third stage using a FIXED LENGTH resequencing buffer.

Black Box: Load Balancing Switch

Fault Tolerance

- No centralized scheduler (no single point of failure)
 - Flexible line-card placement (failure of one linecard will not make the whole system fail)
 - Due to the switching fabric not relying on hard coded linecard placement, MEMS devices can accommodate dynamic linecard configurations.
 - [If there was to be a constant “churn” of linecards, we would not expect the power savings we normally achieve from having relatively static configurations in the MEMS]
- 

Partitioned Switch Fabric

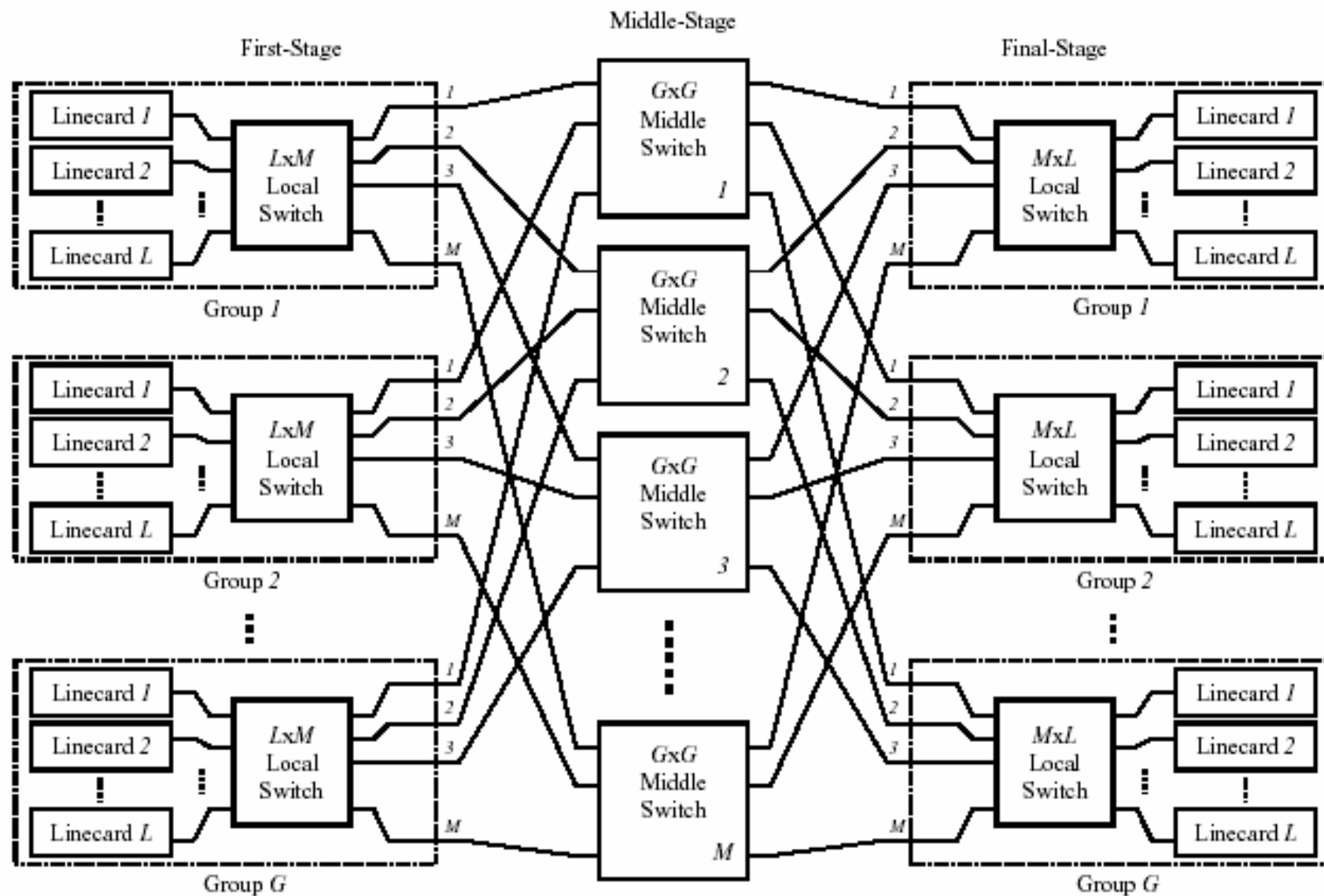
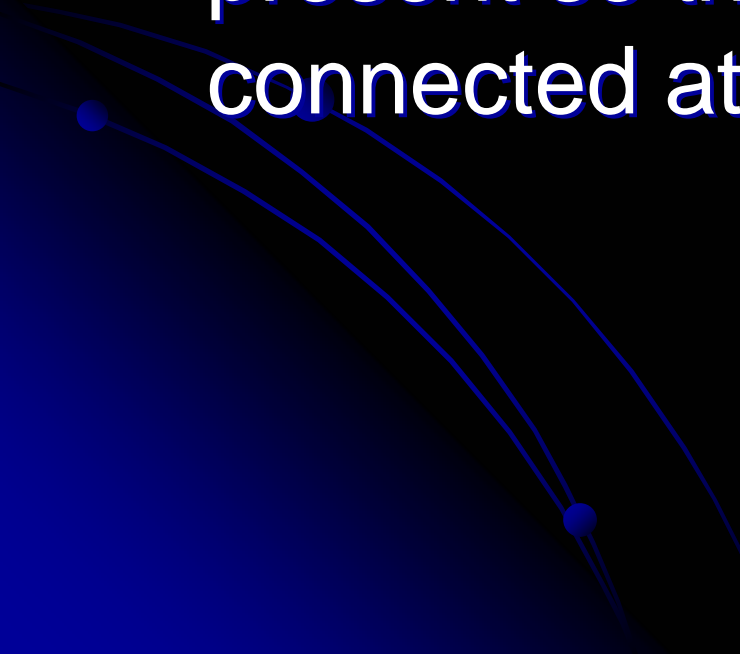


Figure 6: Partitioned switch fabric.

Partitioned Switch

- Thm: We need at most $M = L+G-1$ static paths, where each path can support up to $2R$, to spread traffic uniformly over any set of $n \leq N = G \times L$ linecards that are present so that each pair of linecards are connected at rate $2R/n$.
- 

Hybrid Optical and Electrical

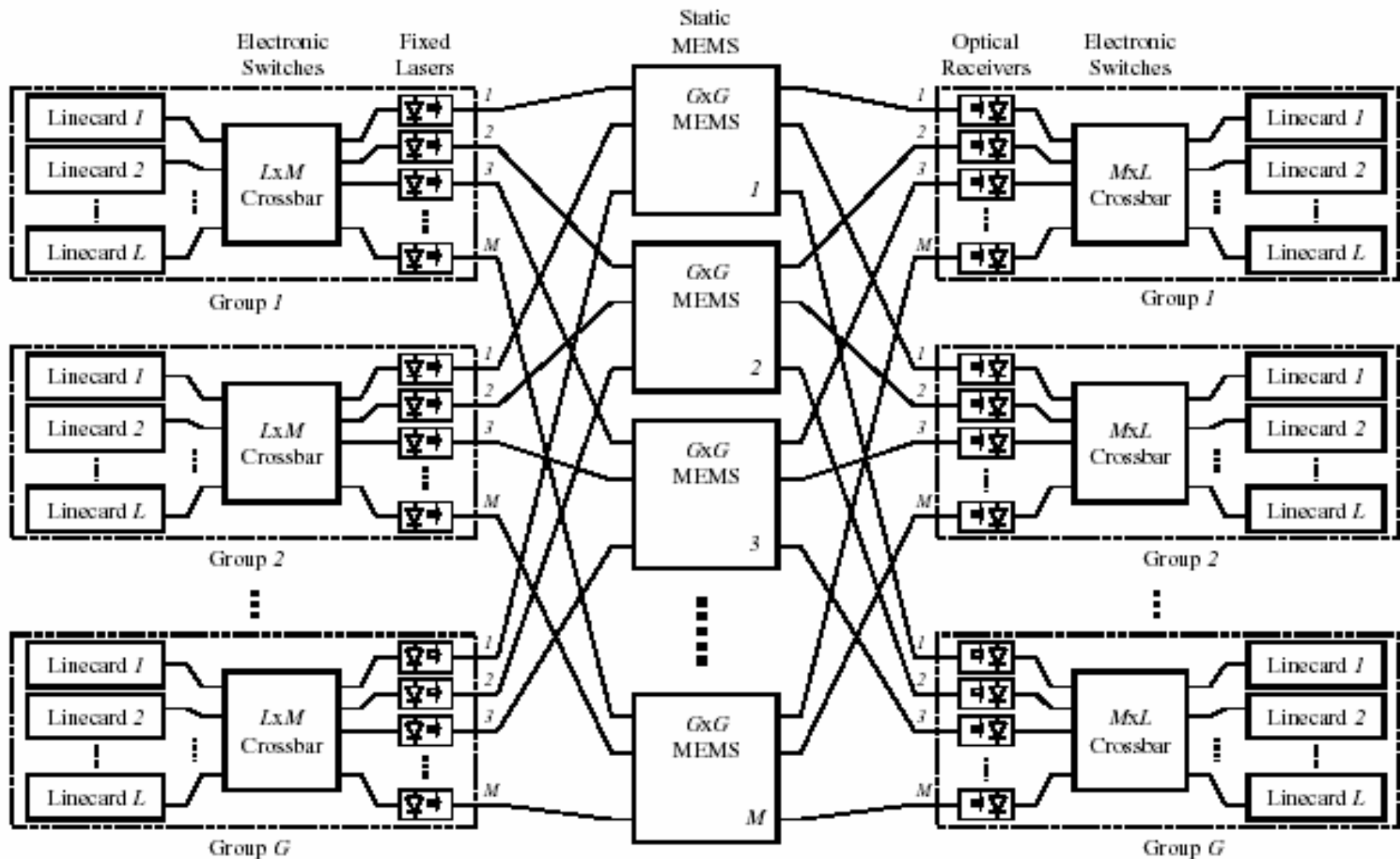
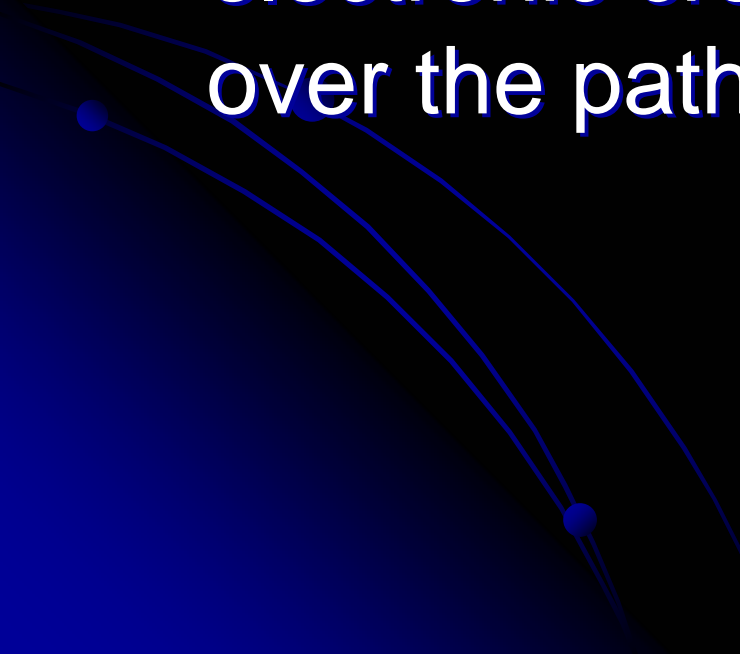


Figure 7: Hybrid optical and electrical switch fabric.

Hybrid Electro-Optical Switch

- Thm: There is a polynomial time algorithm that finds a static configuration for each MEMS switch, and a fixed-length sequence of permutations for the electronic crossbars to spread packets over the paths.
- 

Optical Switch Fabric

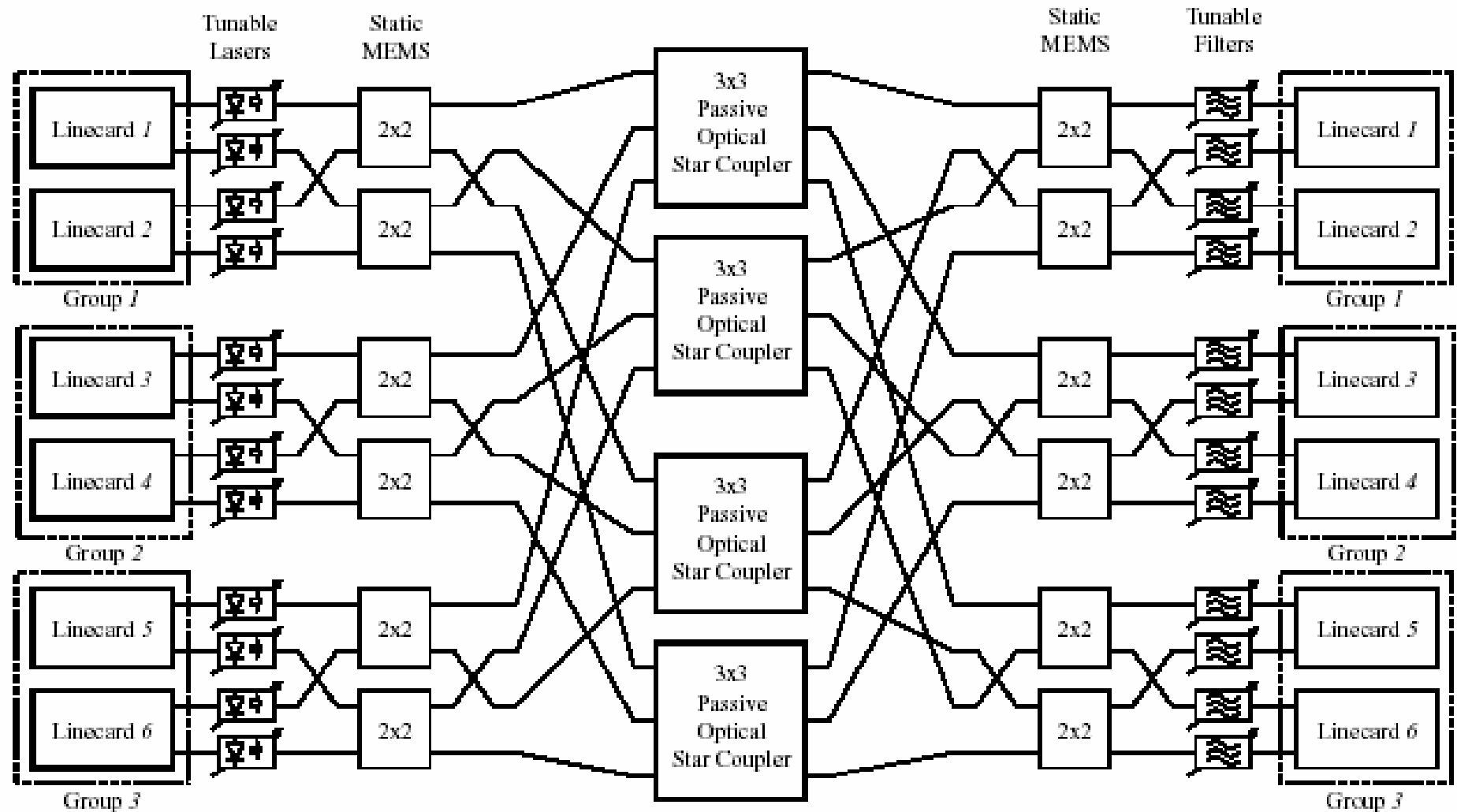


Figure 8: An optical switch fabric for $G = 3$ groups with $L = 2$ linecards per group.

Presentation Outline

➤ Motivation

➤ Avi's "Black Box"

➤ Black Box: Load Balance Switch

➤ Conclusion



Conclusion

- Given optical technology, we can implement a router
 1. that scales
 2. has 100% throughput
 3. consumes low power
 4. delivers ordered packets
 5. is fault tolerant.
- Future considerations include further power reduction:
the replacement of hybrid elctro-optical switch with an
all-optical fabric