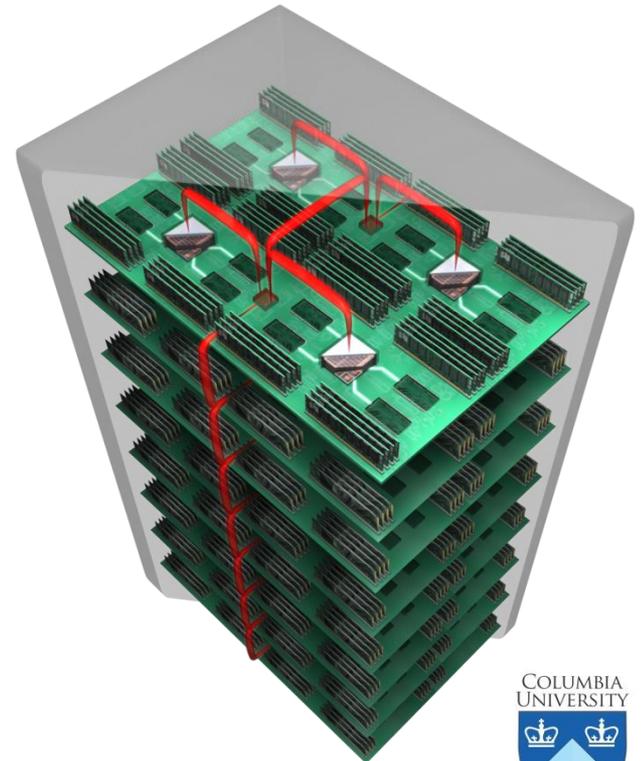


# Energy Consumption and Performance Design Space Trade-Offs for Optical Data Center Networks

Keren Bergman

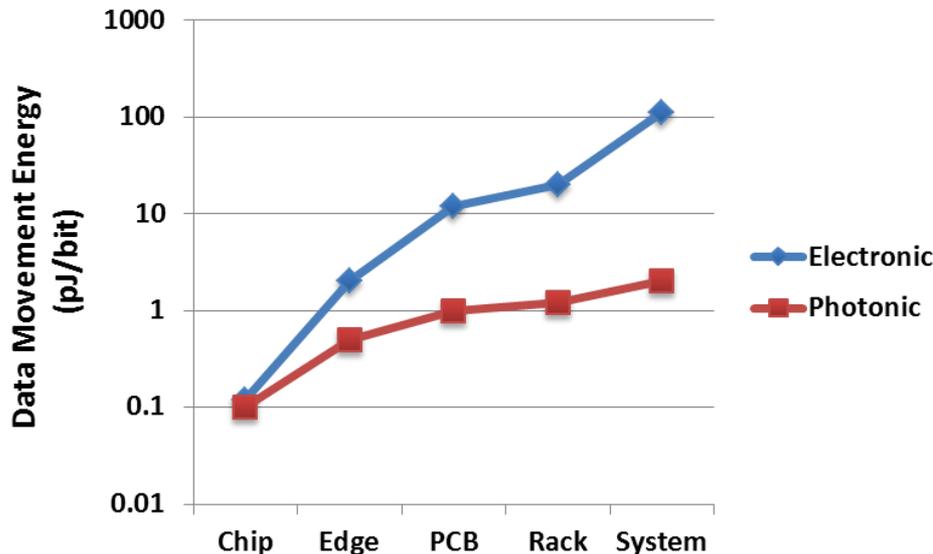
Lightwave Research Laboratory

Columbia University

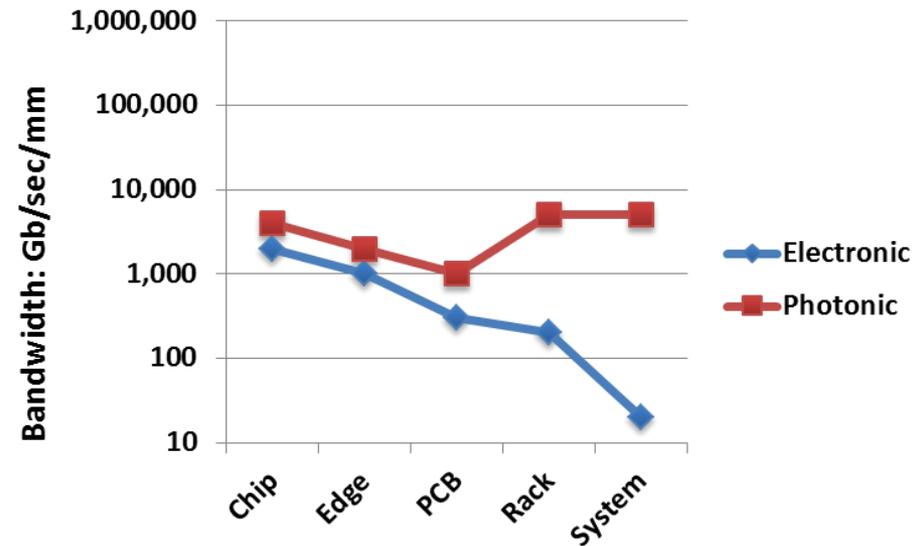


# Data Movement Energy-Bandwidth Challenges

- ❑ Energy efficient, low-latency, high-bandwidth *data interconnectivity* is the core challenge to continued scalability across computing platforms
- ❑ Energy consumption completely dominated by costs of data movement
- ❑ Bandwidth taper from chip to system forces extreme locality



**System Energy Consumption**



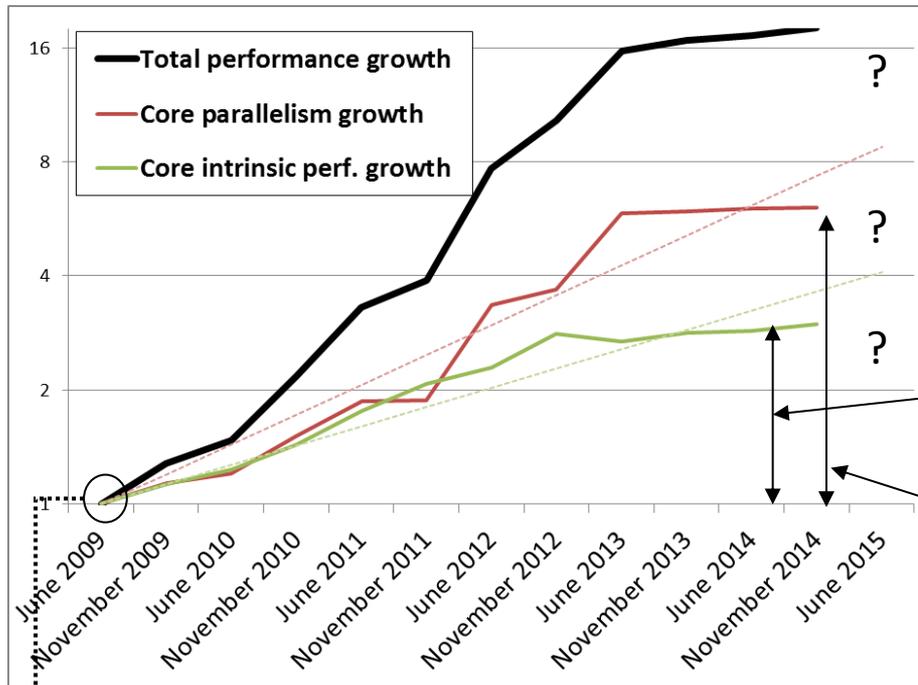
**System Bandwidth Taper**

# Scale Driving Ultra-High Bandwidth

- Data transfers scale with **compute operations**: More Flop/s = More Byte/s

- Data transfers scale with **parallelism**:

- Job division, synchronization...vastly growing parallelism increases the amount of intra data-center traffic
- More “verbose” software, i.e. more network byte per computer operation (more Byte/Flop)



In 5 years, cores (flops) in the top-20 supercomputer increased **2.9 X**

Parallelism increased by **6 X**

Index 1 (June 2009): 374 Teraflop/s, 77k cores (top-20 average)

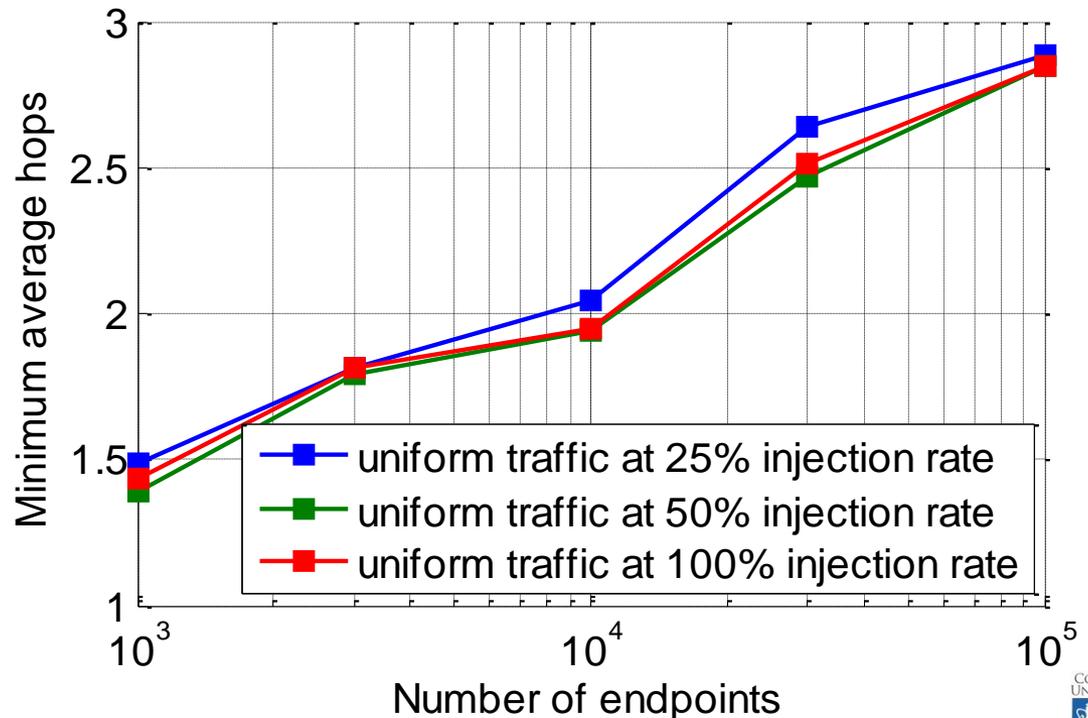
# DC System Size Driving Interconnection Networks

Data transfers scales with data center system size:

- Interconnecting more end-points comes at premium costs...
  - Requires scaling of switch radixes
- As system endpoint nodes scale in (assuming constant switch radix size) minimum number of network hops will increase

Average number of hops in an ideal, optimized topology

- Switches radix 48 ports.
- Uniform traffic



# Summary of Bandwidth Drivers

- Increased aggregated compute power (needed Byte/s)
- Growing parallelism and distributed algorithms (B/F)
- Larger scale systems, vast parallelism = more network hops
- \* *algorithms that reduce communications can help*

→ Clearly, bandwidth needs are growing

- Current numbers:

- Memory interfaces: 100s of Gb/s, soon terabit/s
  - DDR4: 200 Gb/s
  - Hybrid memory cube: 1Tb/s (gen1)
- Network links:
  - 10G widely adopted, 40G emerging
  - 100G already present in HPC
- Router chip envelopes: several Tb/s

} Entering the Tb/s era!



# The Energy Consumption part...

Current systems:

- Sequoia: 2.1 Gigaflop/J; L-CSC (top green500 Nov2014): 5 Gigaflop/J

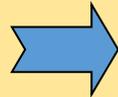
Need for 10-50 Gigaflop/J in the next 5 years (100MW to 20MW at Exascale)

- Challenge for interconnects:

Support increased verbosities within reduced power envelopes

<u>Power envelope</u>	<u>10 Gigaflop/J</u>	<u>50 Gigaflop/J</u>	<u>50 Gigaflop/J</u>
Budget per flop:	100 pJ	20 pJ	20 pJ
Network % of power	20%	20%	20%
Networking budget per flop:	20 pJ	4 pJ	4 pJ
<u>Parallel verbosity</u>	<u>0.1 byte/flop</u>	<u>0.1 byte/flop</u>	<u>1.0 byte/flop</u>
Budget for a 'network' byte	200 pJ/byte	40 pJ/byte	4 pJ/byte
Budget for a 'network' bit	25 pJ/bit	<b>5 pJ/bit</b>	<b>0.5 pJ/bit</b>

Typical verbosities supported by current designs



Tianhe-2  
Sequoia  
Standard Xeon server with 10G

At injection

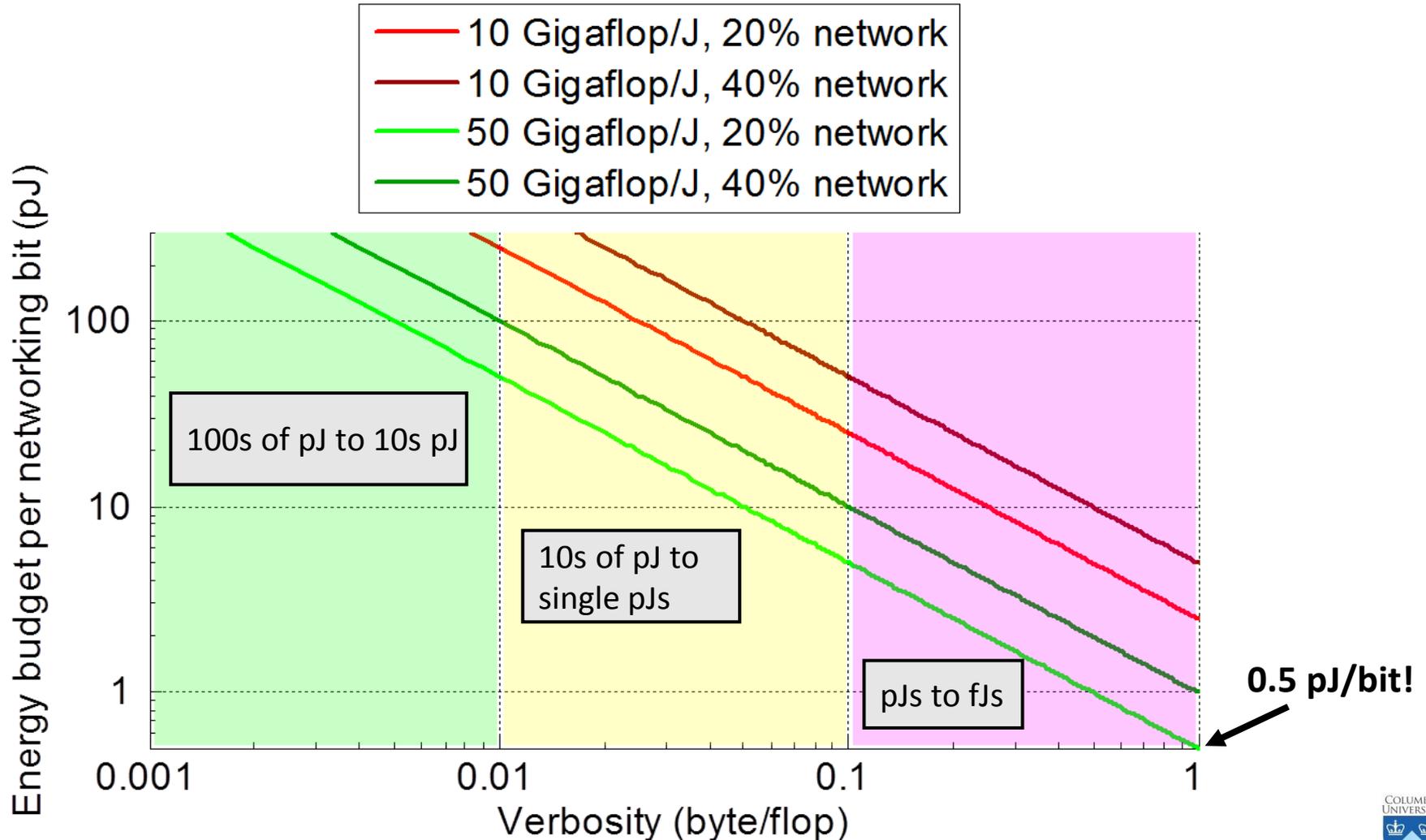
0.001 byte/flop  
0.1 byte/flop  
0.002 byte/flop

Topology wide (uniform traffic)

0.0005 byte/flop  
0.009 byte/flop

# Data movement energy budget vs verbosity (Byte/Flop)

End-to-end network data movement energy budget



# Network energy breakdown

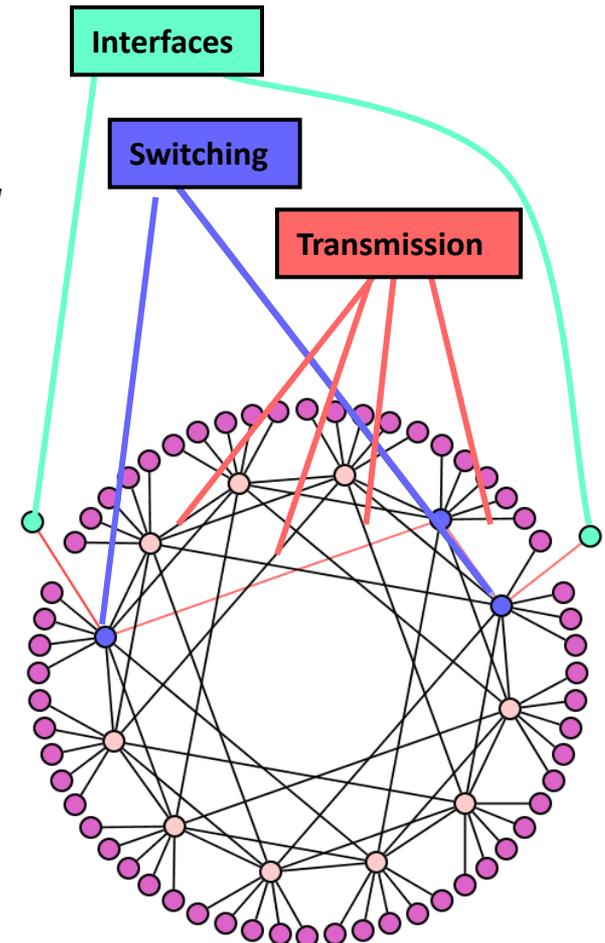
- Decomposing network energy consuming components:

- Interfaces
- Switching
- Transmission
- Number of *internal* network hops: N

*\*assuming 100% network utilization or fully energy proportional*

$$\begin{aligned} \text{Energy}_{\text{network}} = & (N+2) * \text{Energy}_{\text{trans}} \\ & + (N+1) * \text{Energy}_{\text{switch}} \\ & + 2 * \text{Energy}_{\text{interface}} \end{aligned}$$

- Estimating N: (topology independent results)
- N=2
  - For 10,000 endpoints – required switch radix ~48
  - For 100,000 endpoints - required radix of ~96
- N=2.5 – still challenging for 100k endpoints
  - Stress high-locality, low traffic
- N=3 – possible with radix ~48



# Network budget breakdown – switches

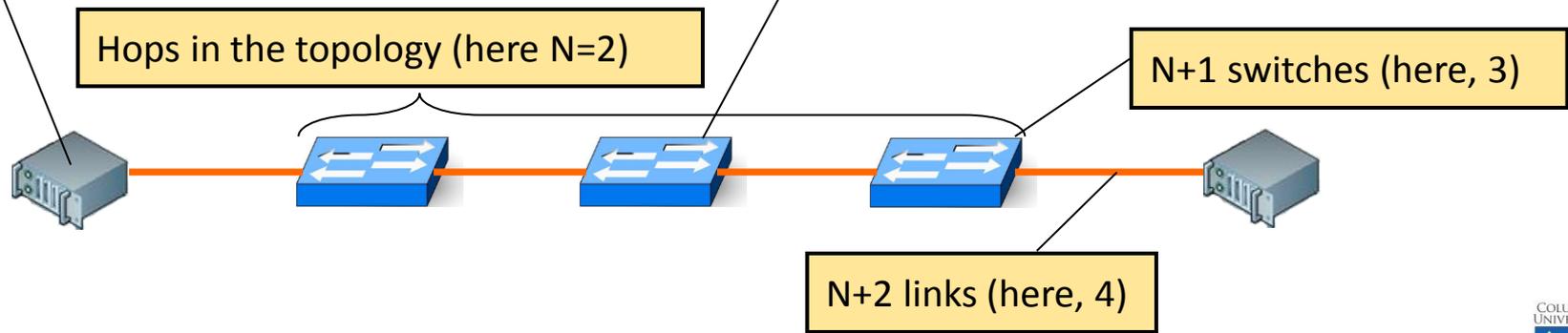
$$\text{Budget}_{\text{network}} = (N+2) * \text{Budget}_{\text{trans}} + (N+1) * \text{Budget}_{\text{switch}} + 2 * \text{Budget}_{\text{interface}}$$

Assuming 200W total chip power and 50% (100W) for switching

**Current switches:**

Cray Aries:	184 lanes @ ~14Gb/s	→ 2.5 Tb/s
	consumption < 100 W	→ < 40 pJ/bit
Upcoming Omnipath:	48 ports @ 100 Gb/s	→ 4.8 Tb/s
	consumption < 100 W	→ < 21 pJ/bit
Exascale switch:	64 ports @ 250 Gb/s	→ 16 Tb/s
		< 6 pJ/bit

Assume  $\text{Budget}_{\text{interface}} = 0$



# Network budget breakdown – links

$$\text{Budget}_{\text{link}} = \frac{\text{Budget}_{\text{network}} - (N+1) * \text{Budget}_{\text{switch}}}{N+2}$$

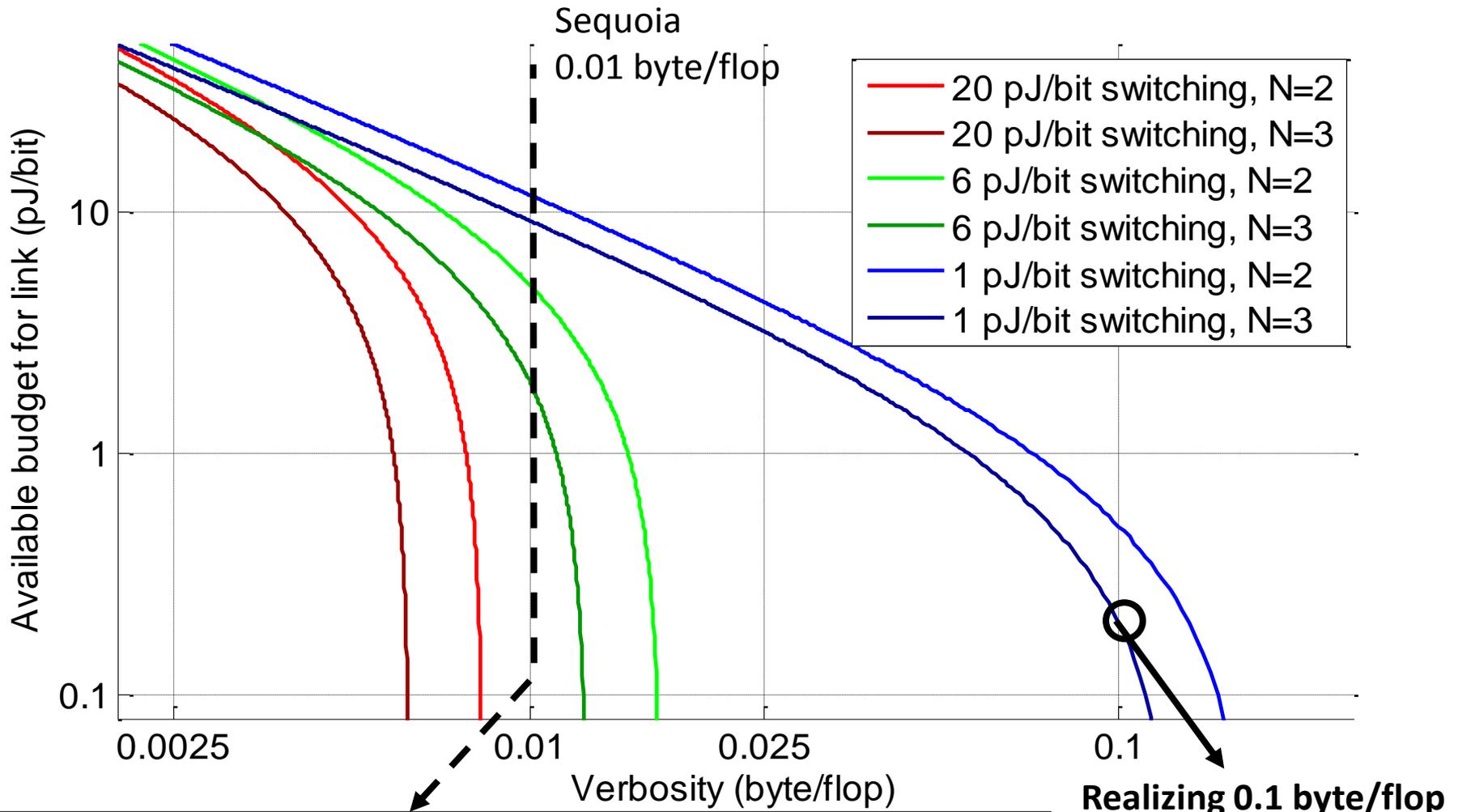
Network portion  
20% in all cases

Verbosity (Byte/Flop)	Energy efficiency (Gigaflop/J)	Total Network Budget <sub>network</sub>	Budget <sub>switch</sub>	N	Budget <sub>link</sub>
0.1	10	25 pJ/bit	6 pJ/bit	2	1.75 pJ/bit
0.1	10	25 pJ/bit	4 pJ/bit	3	1.8 pJ/bit
0.1	50	5 pJ/bit	1 pJ/bit	2	500 fJ/bit
0.1	50	5 pJ/bit	1 pJ/bit	3	200 fJ/bit
1.0	10	2.5 pJ/bit	0.5 pJ/bit	2	250 fJ/bit
1.0	10	2.5 pJ/bit	0.5 pJ/bit	3	100 fJ/bit
1.0	50	0.5 pJ/bit	0.1 pJ/bit	2	50 fJ/bit
1.0	50	0.5 pJ/bit	0.1 pJ/bit	3	20 fJ/bit

- N=2 requires switch radix ~ 96
- N=3 switch radix ~ 48

- N=2: 3 switches, 4 links
- N=3: 4 switches, 5 links

# Link budgets for 50 GigaFlop/J system with 20% network



To support **0.01 byte/flop** (Sequoia) verbosity at 50 Gigaflop/J:

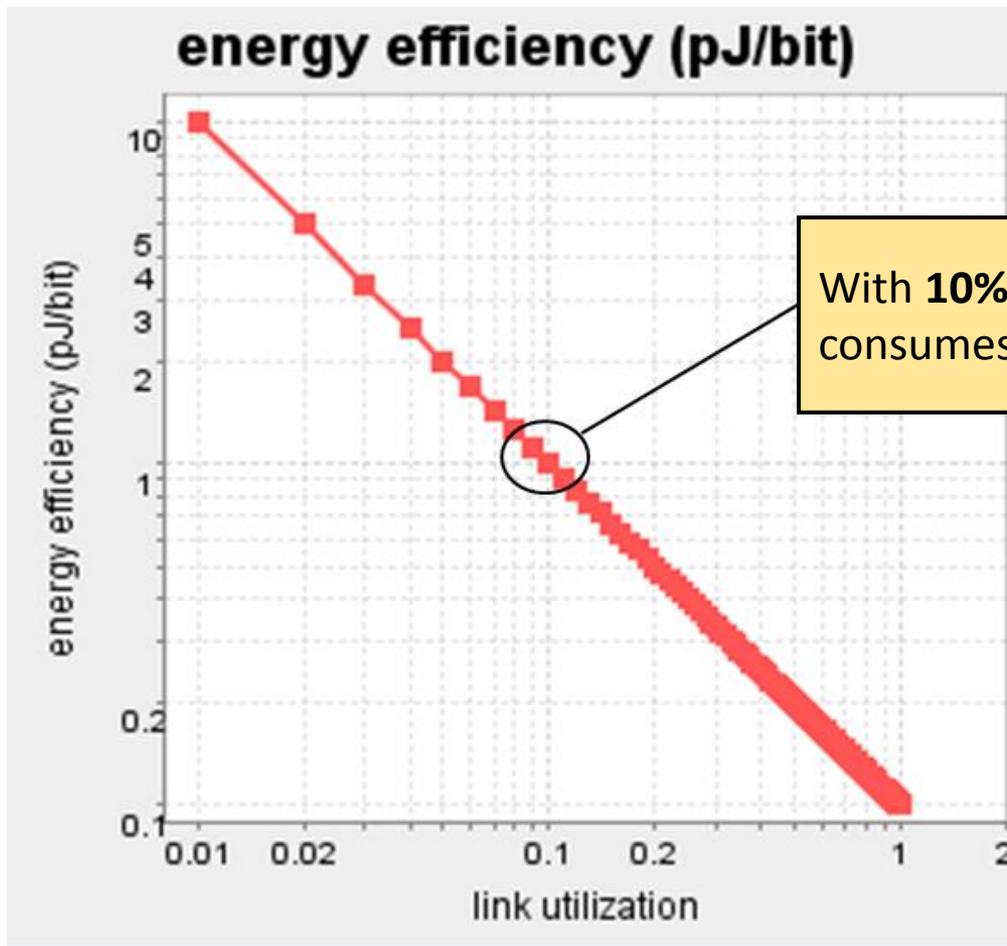
- 1) Switching must consume < 10pJ/bit
- 2) If switches consume 6pJ/bit, link **Energy<sub>trans</sub> ~ 2.5 pJ/bit**

# What about the laser energy consumption...

- Baseline case:
    - 10Gb/s per wavelength
    - Detector sensitivity: -20dBm
    - Link optical budget including modulation: 10dB
    - Launch power -10dBm = 0.1 mW
    - Laser «wall plug» efficiency: 10%
- Laser power: 1mW
- Laser contribution to energy consumption: **0.1 pJ/bit**
- \* assuming no additional power penalties due to WDM

# The role of link utilization in energy consumption...

- Assume laser ON continuously
  - But...link carries real data traffic 10% of the time
  - Energy efficiency inversely proportional to utilization



With **10% utilization**, laser consumes the full **1pJ/bit** budget

# Typical (low) utilization in Data Centers

*“Given the large number of unused links (40% are never used)...”*

**Links are highly utilized (more than 95%) only 10-30% of the time**

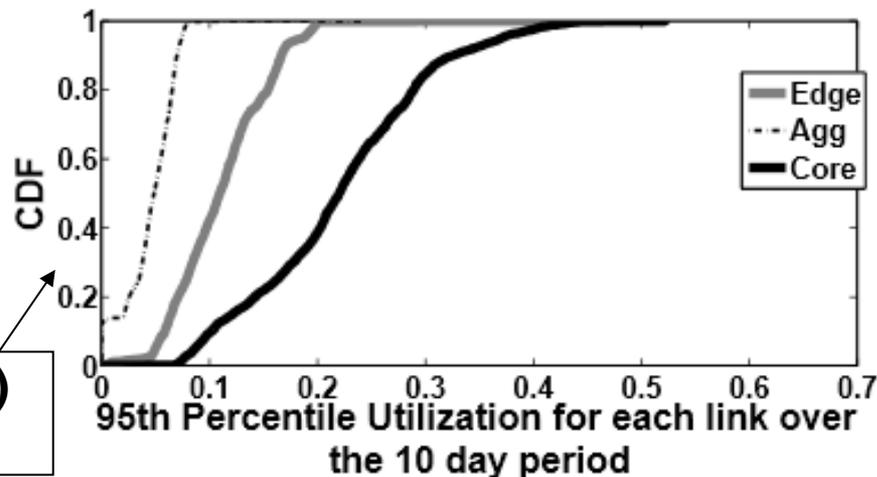


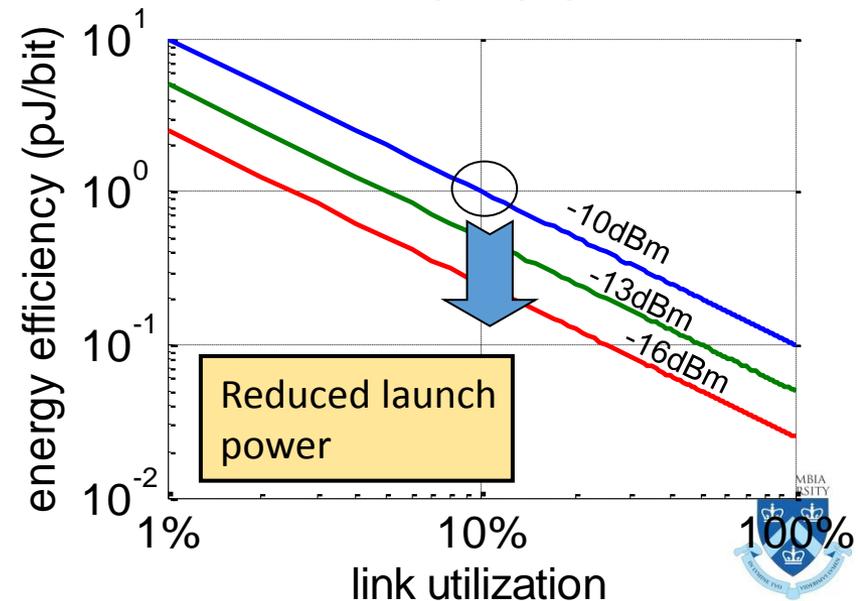
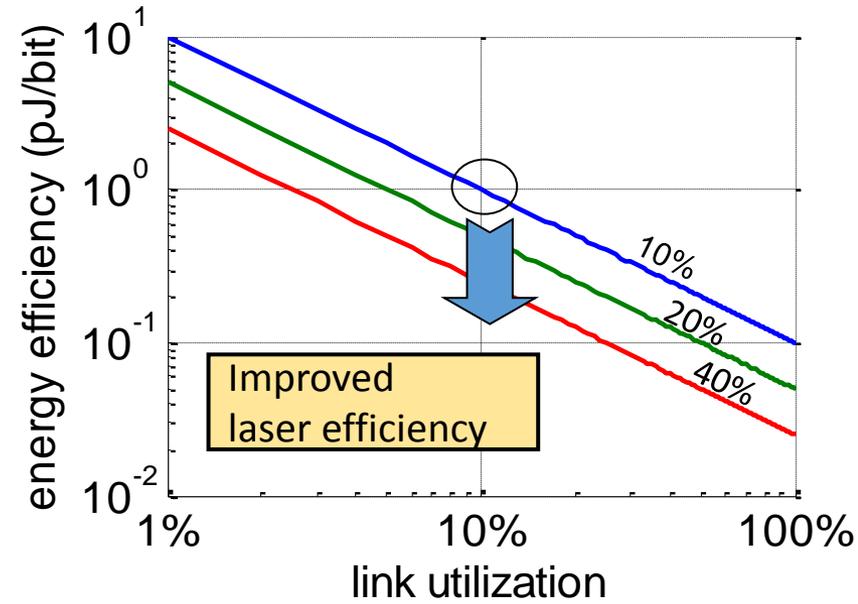
Figure 2: A CDF of the 95th link utilization at the various layers in the Data Centers Studied

## Understanding Data Center Traffic Characteristics

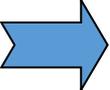
Theophilus Benson\*, Ashok Anand\*, Aditya Akella\* and Ming Zhang†  
\*UW-Madison, †Microsoft Research

# Laser energy consumption VS utilization trade-off

- 10% utilization “adds” 10dB
- Increase energy efficiency by:
  - Improved laser efficiency
  - Reduced launch power
    - Better receiver sensitivity
    - Reduced link power penalties
- Need combined factor of 10X improvement to achieve 0.1pJ/bit at 10% network utilization



# Low average utilization is desirable for performance

- Why is low utilization advantageous?
  - A close to 100% utilization case. 



- Low utilization needed to guarantee low queuing
  - In particular, queuing synchronization messages threatens parallel efficiency

# Another factor: optical circuit switching...

- Optical circuit switching: *inherently* low average utilization
- Low utilization as the result of circuit switching:

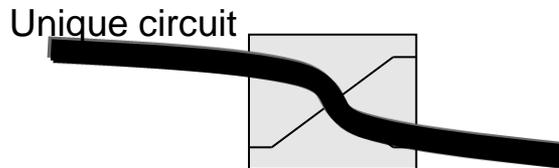


- Streaming circuit data cannot be slowed when in motion

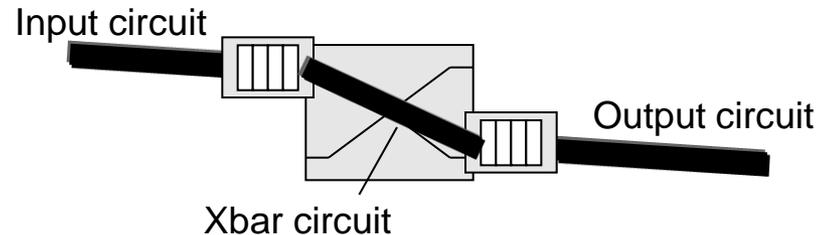
# OCS – why low average utilizations

- The optical ‘circuit’ is the transmission link
- When a switch “turns,” no transmission can occur
  - Turning the switch = breaking circuits
  - No active circuits over a turning switch
- Unless the circuit is never reconfigured...circuit switch cannot be 100% fully utilized
  - Utilization can be high if reconfiguration  $\ll$  circuit ON time
  - Poor utilization if reconfiguration  $\geq$  circuit ON time

## Optical switching



## Packet (electrical) switching



# Packet duration shrink with increased bandwidth

- Packet durations will trend to ~1-10ns

		Packet sizes			
		100B	1KB	10KB	100KB
Aggregate Line rates	100Gb/s	8ns	80ns	800ns	8 $\mu$ s
	400Gb/s	2ns	20ns	200ns	2 $\mu$ s
	1Tb/s	800ps	8ns	80ns	800ns
	2.5Tb/s	320ps	3.2ns	32ns	320ns

# Impact of optical circuit switching on utilization

- Link **unavailability time** composed of:
  - Switch configuration (optical path)
  - Link re-establishment (equilibrate, preamble, etc.)

- Resulting utilization:  
(worse-case)

		Link unavailability		
		1ns	10ns	100ns
Packet duration	100ns	99%	91%	50%
	10ns	91%	50%	9%
	1ns	50%	9%	1%

- Resulting utilizations:  
(switch turns after every second packet)

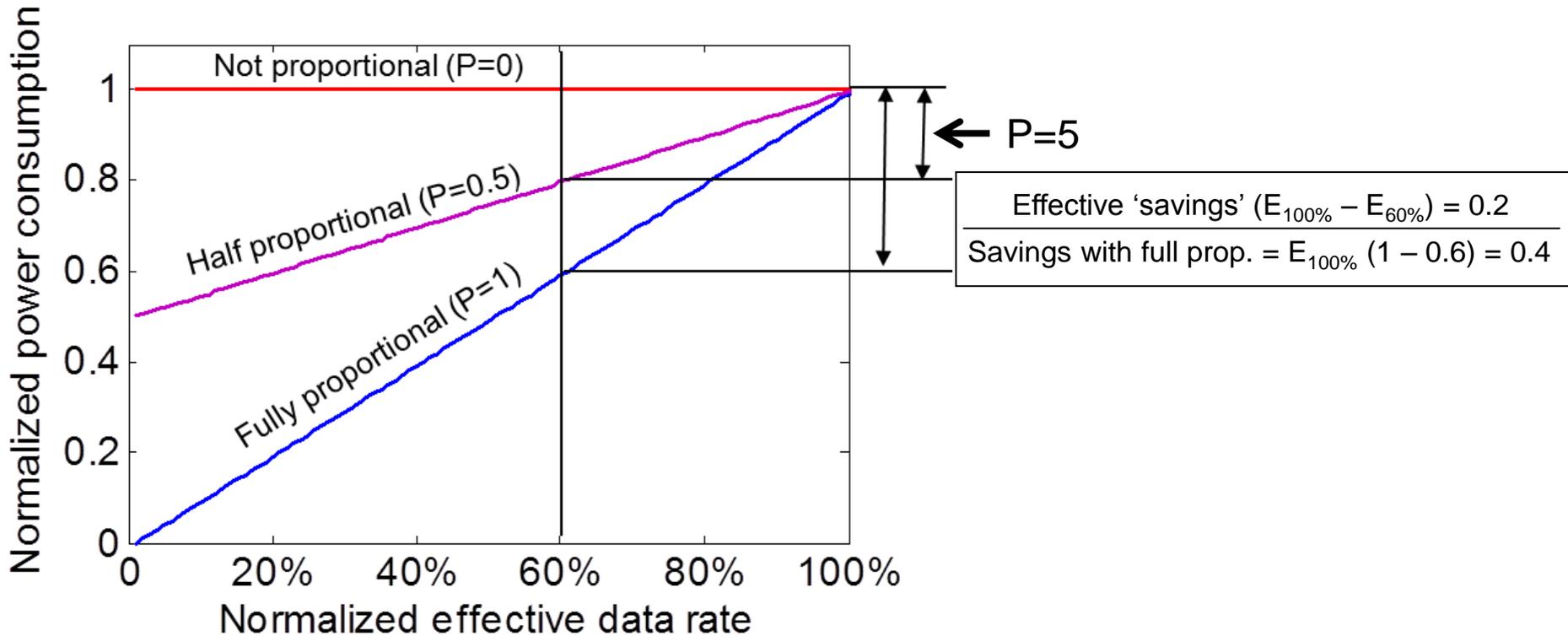
		Link unavailability		
		1ns	10ns	100ns
Packet duration	100ns	99%	95%	66%
	10ns	95%	66%	16%
	1ns	66%	16%	2%

- **Need circuit 'down' time no more than ~1ns!**

# Energy *proportional* links

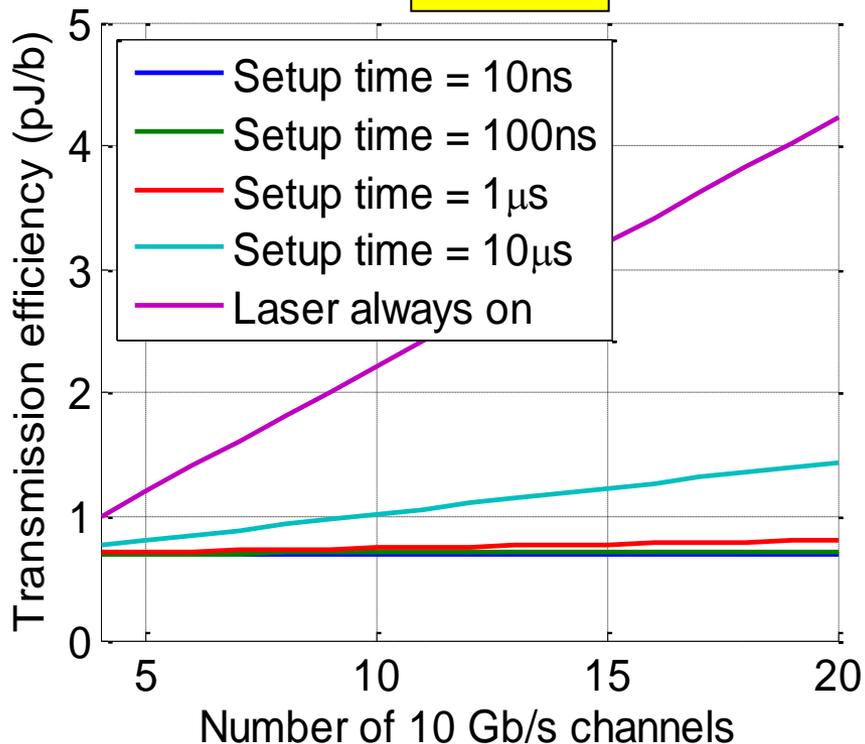
- Energy proportionality factor P:

$$P = \frac{\text{Energy savings compared to 100\% utilization case}}{\text{Energy savings with full proportionality}} = \frac{E_{100\%} - E_{\text{util}}}{E_{100\%} (1 - \text{utilization})}$$

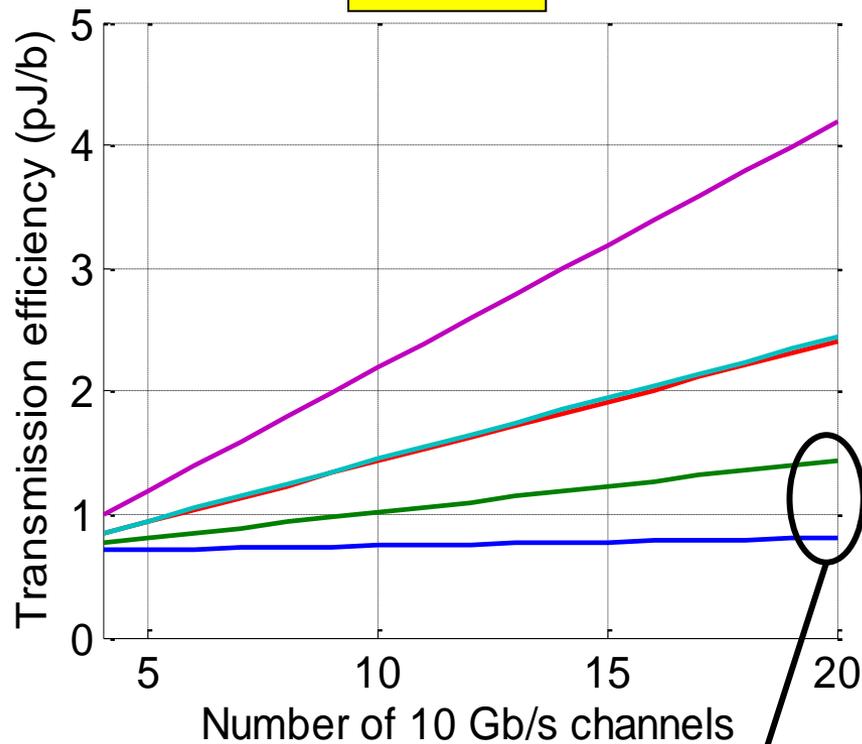


# Need for ns-scale energy proportionality

100KB

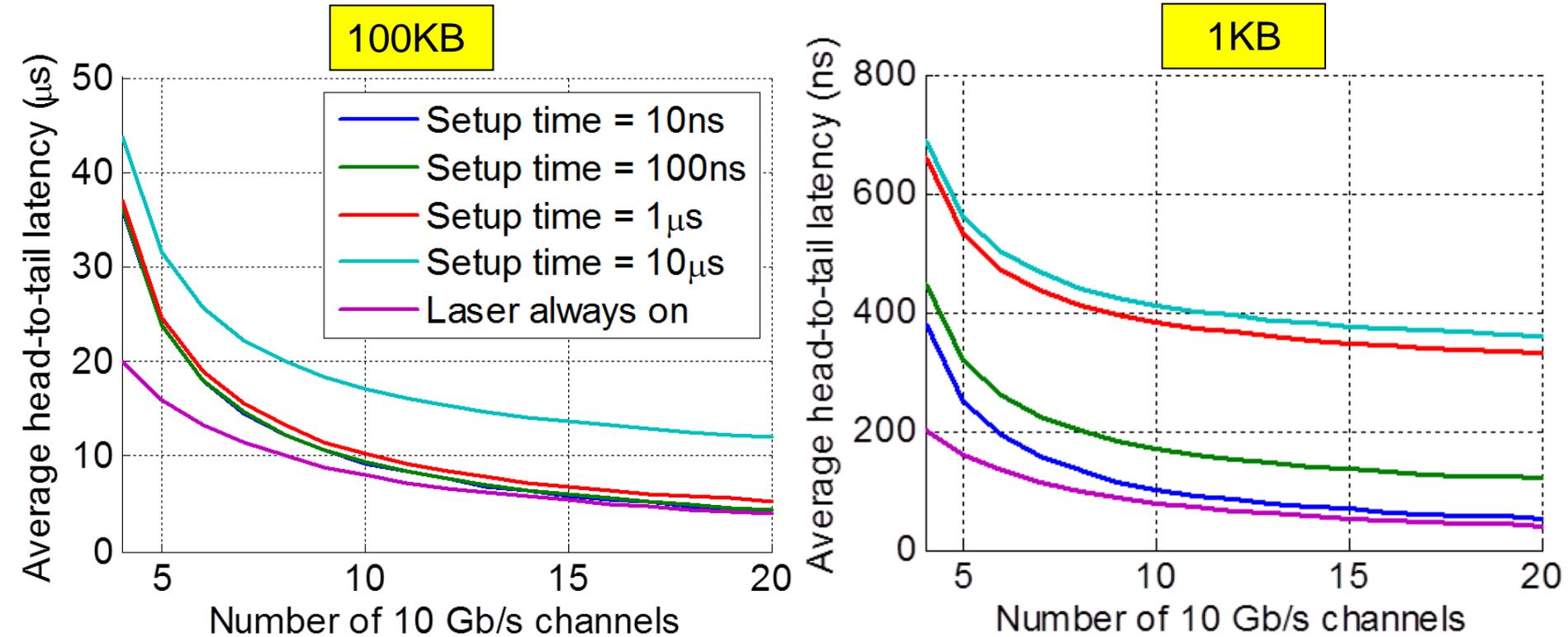


1KB



1KB packets require at least 100ns and **~10ns** dynamic data optimal proportionality

# Latency performance impact



- Head-to-tail latency includes both queuing and serialization times
- Keeping the laser ON yields the best performances – but highest energy cost
- Adding channels improve performance (reduces serialization times)
- Laser setup time >100ns inflicts a substantial penalty

# summary

24

- Data center scalability drives increased interconnectivity bandwidth:
  - Aggregated compute power (needed Byte/s)
  - Growing parallelism and distributed algorithms (B/F)
- System wide connectivity and data movement bandwidth
  - key to performance and scalability
- Energy consumption interconnection network total budget:
  - 0.1B/F and 50GigaFlop/J → 5.0pJ/bit
  - 1.0B/F and 50GigaFlop/J → 0.5pJ/bit
- Laser power:
  - At 1mW and 10% wall-plug efficiency: consumes 0.1pJ/bit with 100% utilization
  - 10% network utilization “adds” 10dB, to 1pJ/bit
  - Need combined 10X improvement to regain 0.1pJ/bit at 10% network utilization
- Unless the circuit is never reconfigured...cannot be 100% utilized
  - Utilization can be high if reconfiguration  $\ll$  circuit ON time
  - Poor utilization if reconfiguration  $\geq$  circuit ON time
- Packets 1ns-10ns for 1KB and ~Tbit/sec scale
- Need circuit ‘down’ time no more than ~1ns

