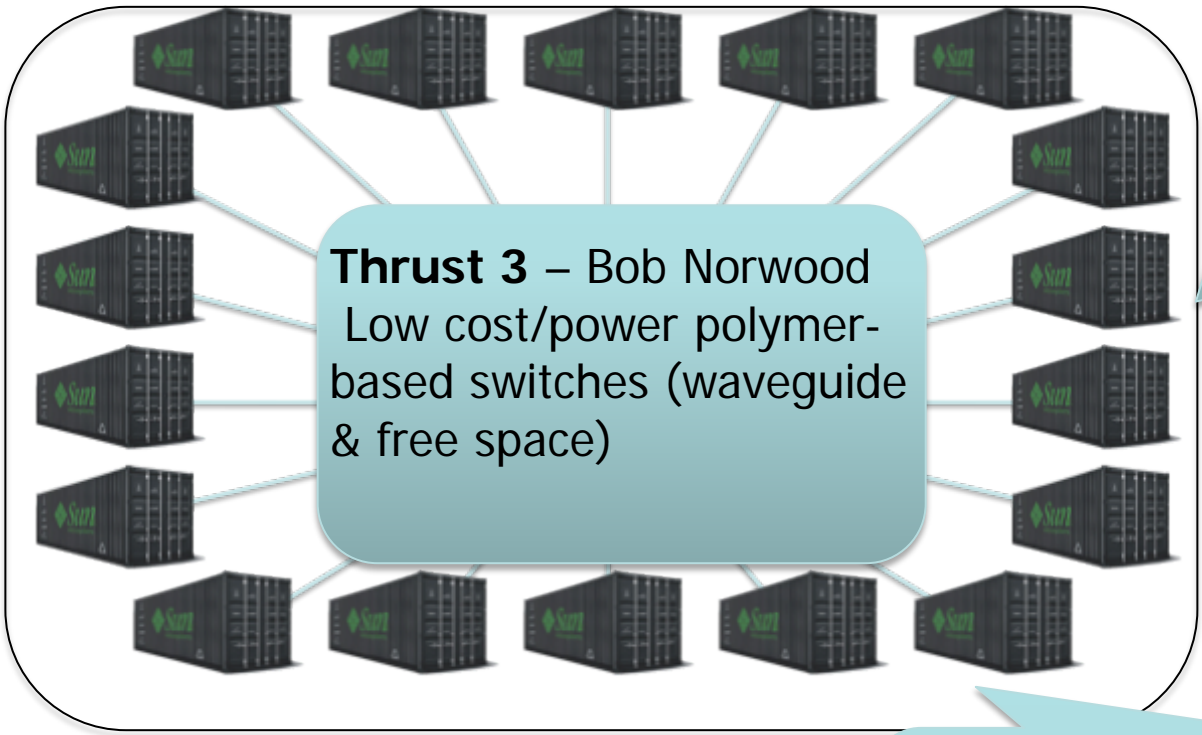


CIAN Projects Supporting Data Centers



Thrust 1 – George Porter

System balance (software/hardware)

Data center topologies & architectures

•Thrust 2 – Connie Chang-Hasnain
Fast tunable VCSELs for wavelength-based architectures



Outline

- System balance (hardware and software)
 - Deployed traffic study: TritonSort
- Data center topologies and architecture
 - Experimental evaluation: Helios interconnect
- Drivers for Thrusts 2 and 3



Back-end data center traffic study: TritonSort

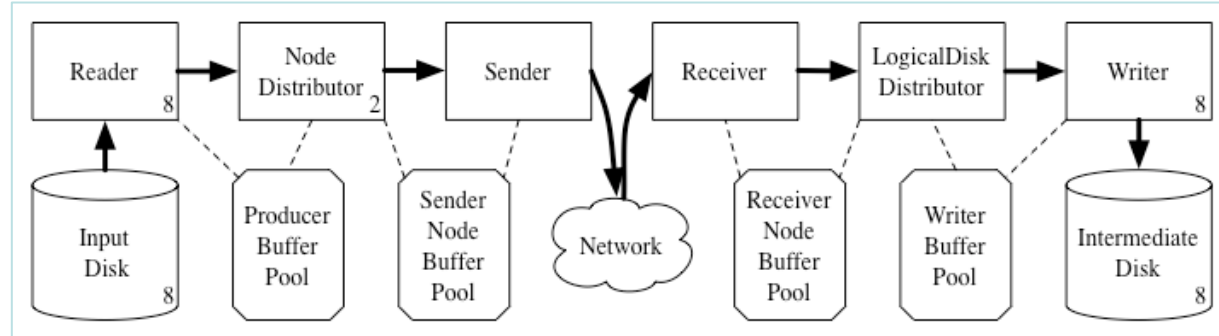
- Intra-datacenter traffic patterns are unique
 - Datacenters need more than just fat pipes
 - All-to-all workloads especially important
- We carried out a traffic study by building a distributed application exhibiting exactly this property: **TritonSort**
- TritonSort:
 - Highly efficient sorting system
 - Sorting forms the basis of many datacenter applications
 - Focus is balancing computation, storage, memory, and especially network



CIAN WG1 traffic pattern case study: TritonSort—a Balanced Sorting System



CIAN WG1
Testbed



- Moore's law: more cores per node
 - 4, 8, 16, 32, ...
- 1-2 disks / core to keep it busy
 - 16 disks per node (1,136 all)
- Bandwidth requirements:
 - 6.4 Gbps per node
- We used a 52-port electrical switch
 - To determine the requirements without imposing constraints

CIAN WG1 traffic pattern case study: TritonSort—Results

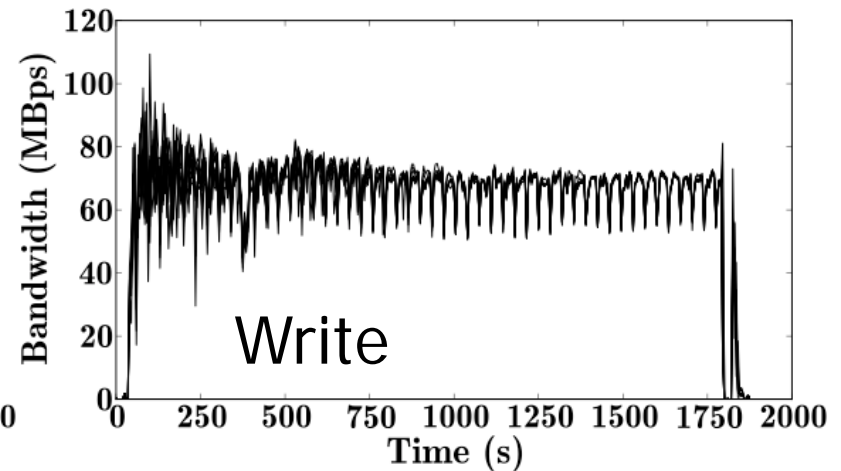
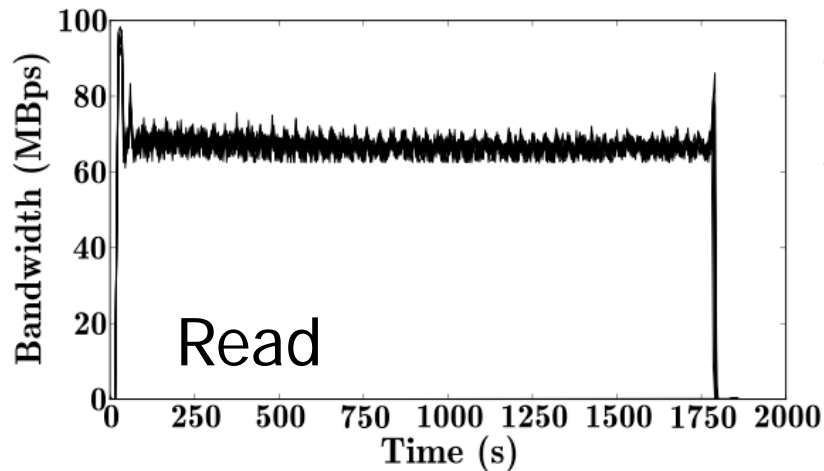
- Sortbenchmark.org: Jim Gray, 1985
- Sort 100 TB as fast as possible
- 5x improvement in efficiency compared to previous winner
 - ✓ Faster with fewer machines
 - ✓ 195 machines → 48 machines

| Category | Previous winner | TritonSort | Per-imp |
|------------|-----------------|-----------------|------------|
| GraySort | 0.564 TB/min | 0.785 TB/min | 565 |
| MinuteSort | 955 GB / 60.0s | 1014 GB / 60.0s | 440 |

World record!



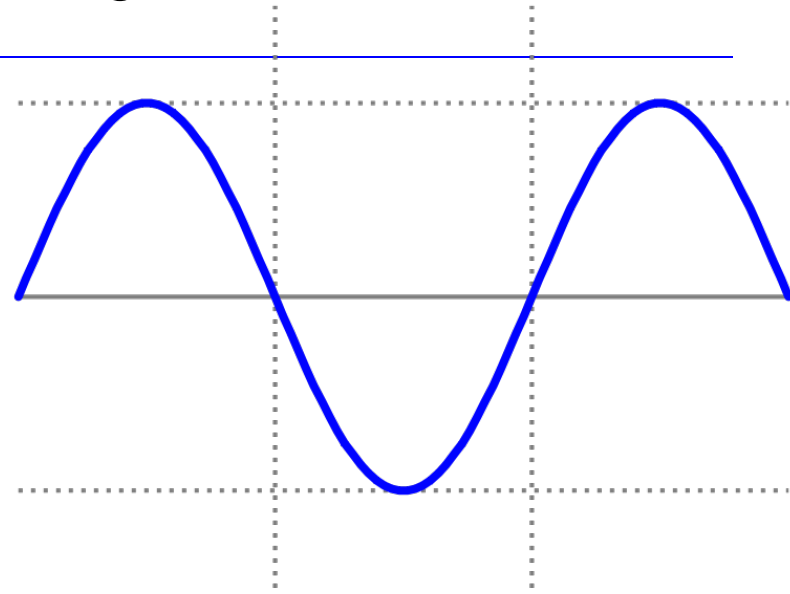
Traffic study: lessons learned for Thrust 3



- All-to-all rate of ~ 6 Gbps to 48 nodes in units of 16KB needed to break world record
- Equivalent optical switching speed: **21.33 μ s**
- Limiting the time spent reconfiguring the circuit to 5%,
we need a 1.07 μ s space switch

Traffic study: Scalability of results

- Amplitude: Data rate per connection
- Frequency: Switching Speed
- Data center requirements for CIAN
 - Fatter pipes growing faster than switching speed → need rate agnostic switches
 - Switching speed increasing, but at slower rate
- Intermediate term: Mix and match space and wavelength
- Long term: large space switch with < 1 microsecond

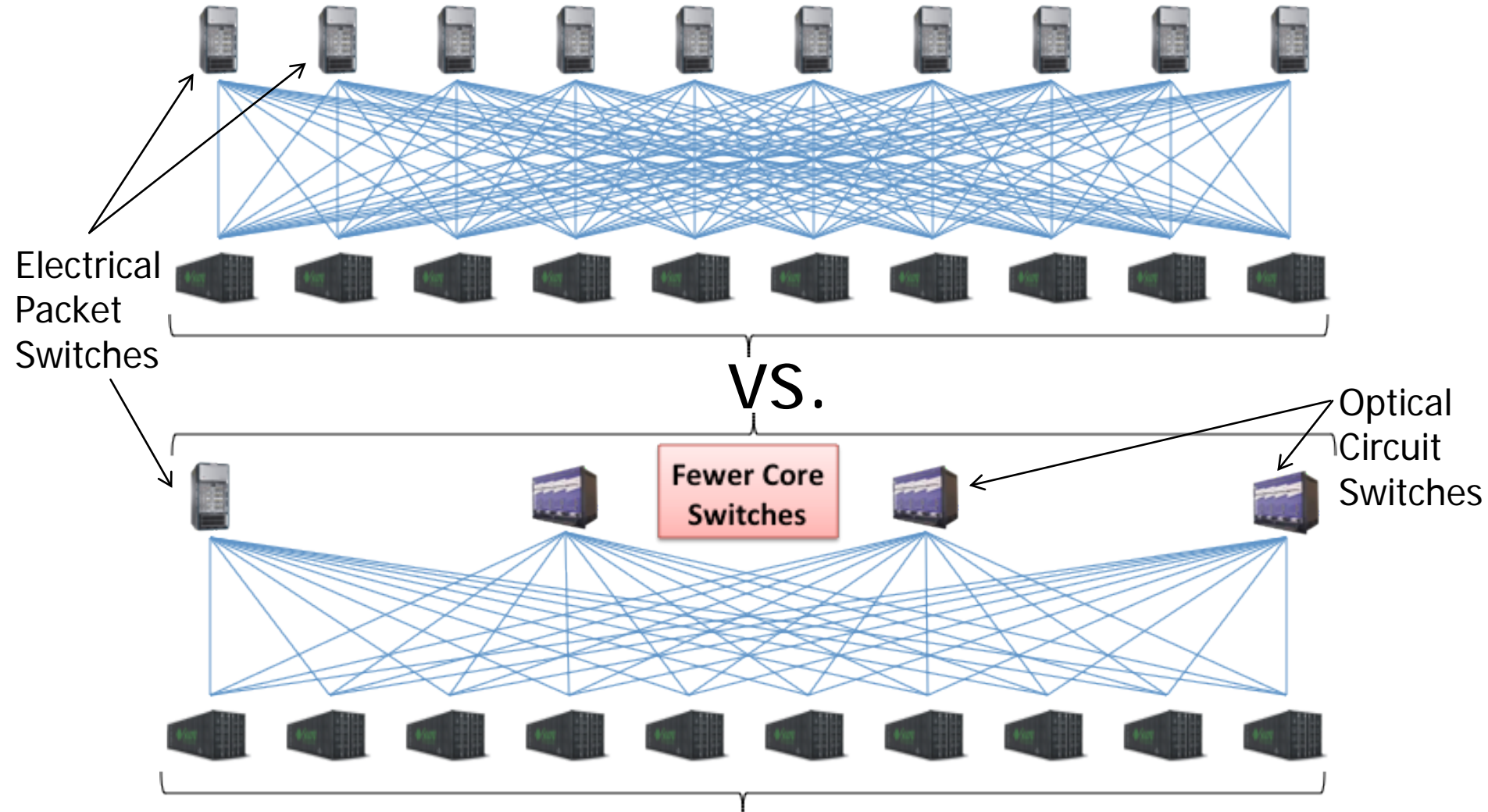


Outline

- System balance (hardware and software)
 - Deployed traffic study: TritonSort
- Data center topologies and architecture
 - Experimental evaluation: Helios interconnect
- Drivers for Thrusts 2 and 3

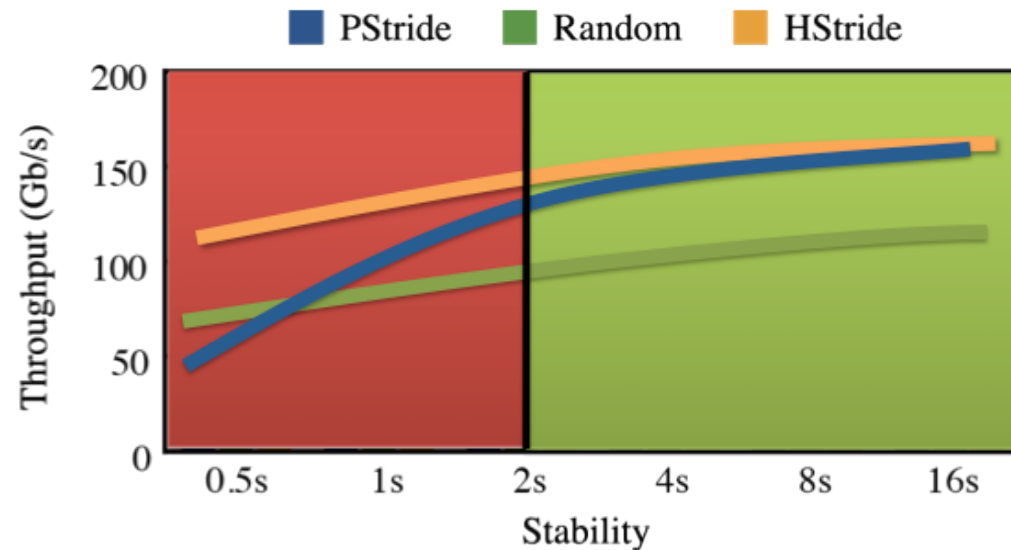


Tradeoff between *Connectivity* and *Latency* with MEMS switches

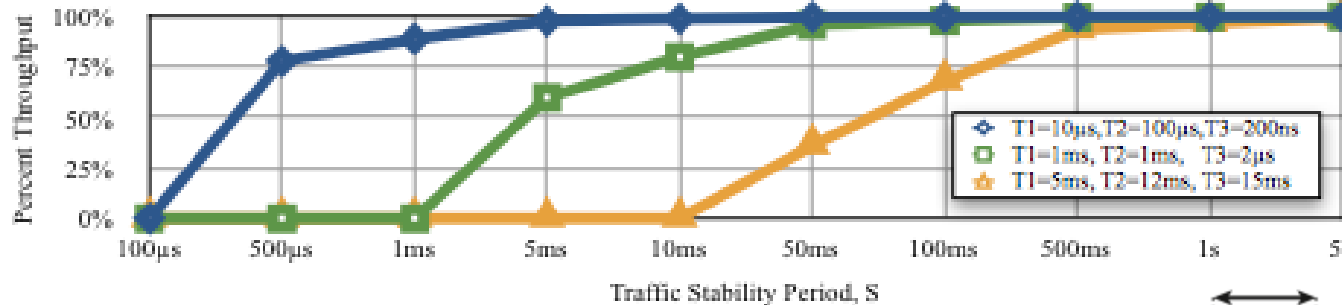
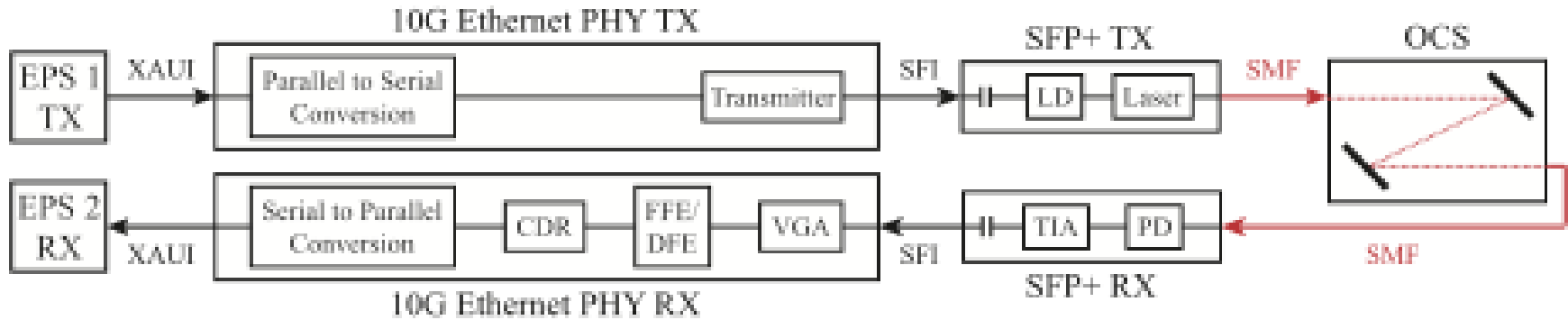


Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers [SIGCOMM 2010]

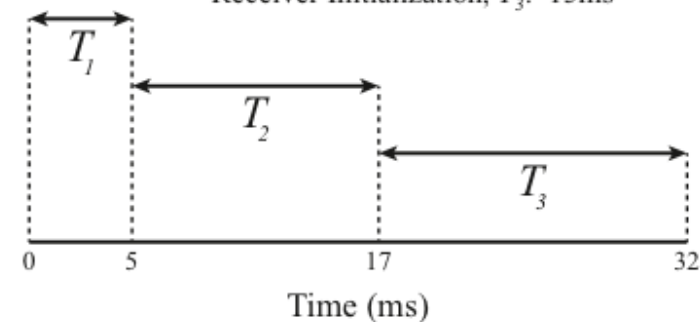
- CIAN WG1 Thrust 1 effort to evaluate tradeoff between connectivity and latency in the datacenter
 - Connectivity: 10-80 Gbps cWDM optical circuits
 - Latency: 25ms Glimmerglass 64-port switch
- Findings:
 - Commodity MEMS switches can support limited datacenter traffic with stability of at least several seconds



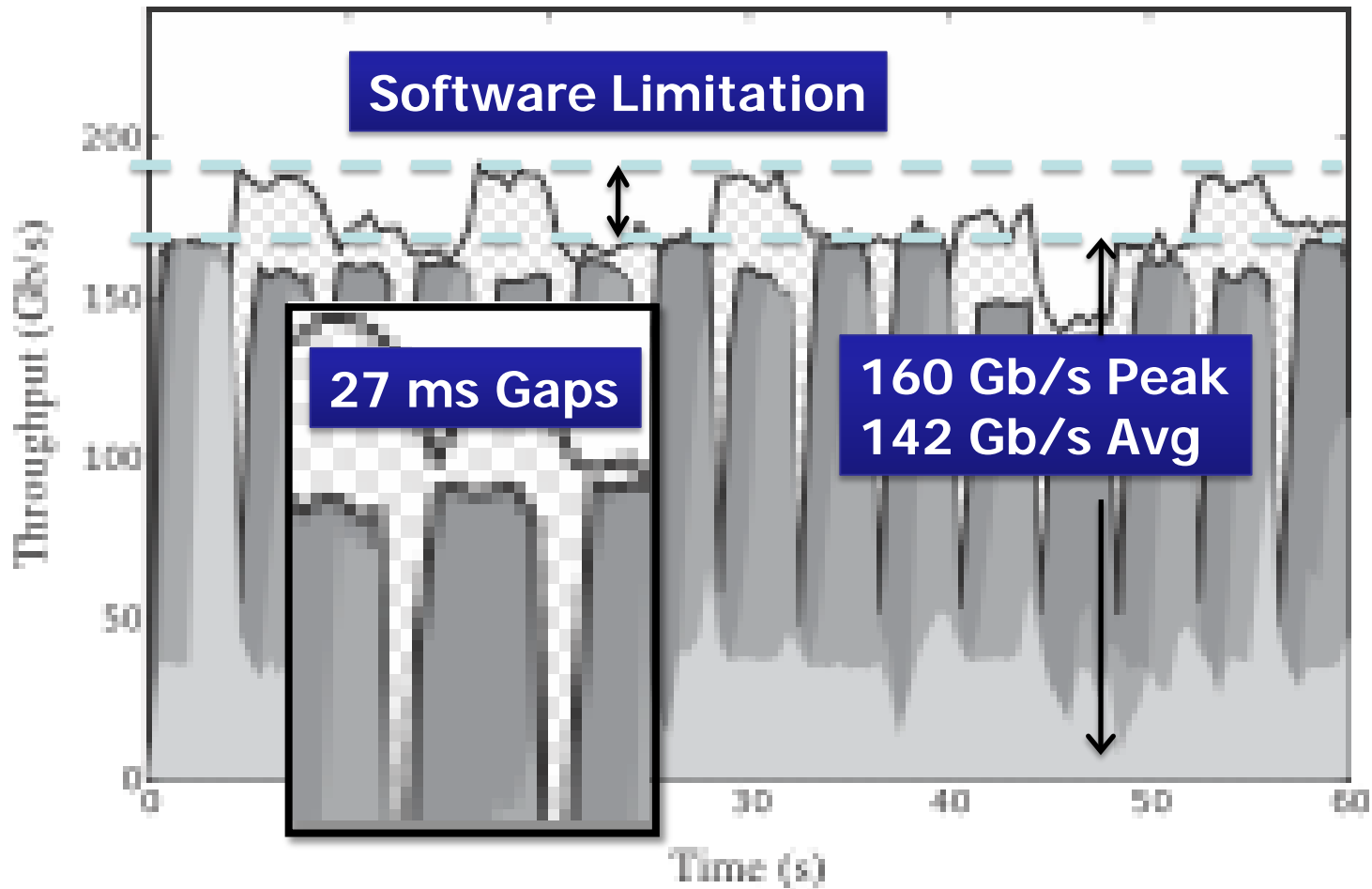
Hardware Requirements for Optically Switched Datacenter Networks



Command Processing, T_1 : 5ms
 Mirror Reconfiguration, T_2 : 12ms
 Receiver Initialization, T_3 : 15ms



Motivation for wavelength switching: reducing “gaps” during switching



Connectivity vs. Latency

- Where to go from here in terms of wavelengths?
 - Intermediate design point between large port, but slow space switch and all-electrical packet switch
 - Which level to apply wavelength?
 - Pod → wavelength switch → space switch
 - Pod → space switch → wavelength switch
- Takeaway from traffic study and Helios prototype evaluation:
 - Space switch switching speed: 1.07 μ s
 - Mixture of space and wavelength



Discussion

- Thank you

