Lecture 9: Hybrid Networks

CSE 222A: Computer Communication Networks

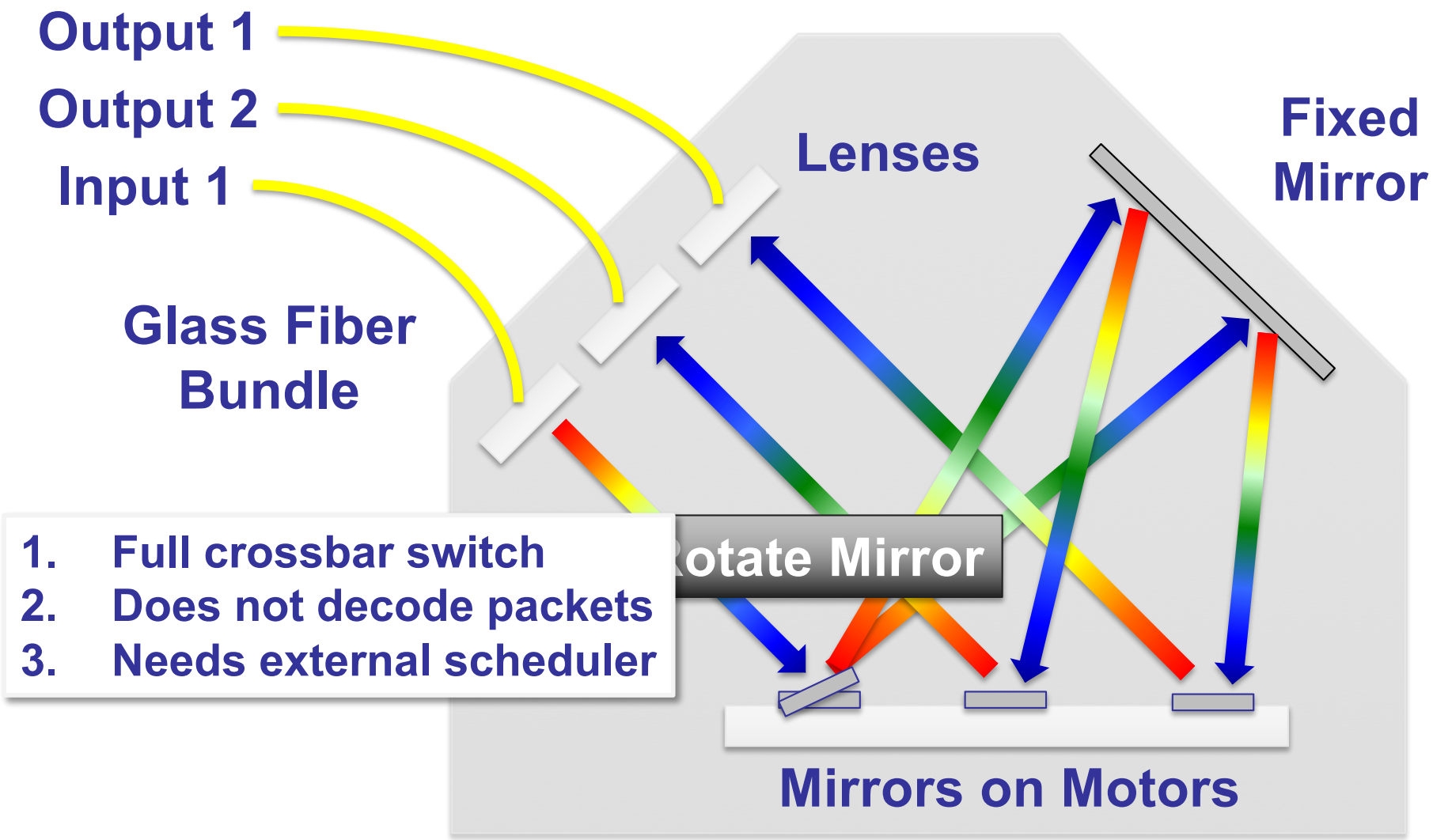
Alex C. Snoeren

Thanks: Nathan Farrington

Lecture 9 Overview

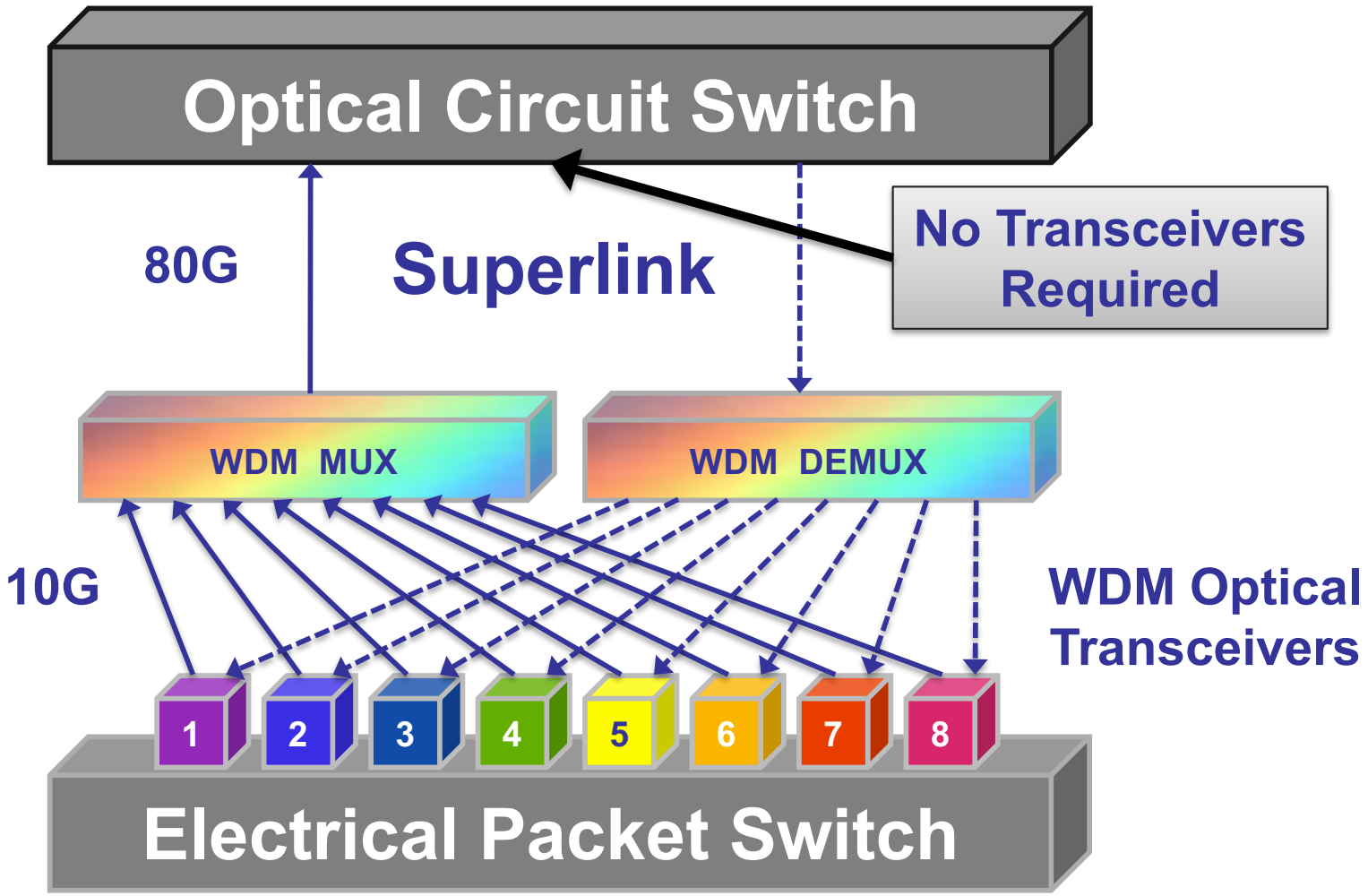
- Project discussion
- Overview optical technology
- Helios paper discussion

Optical Circuit Switch

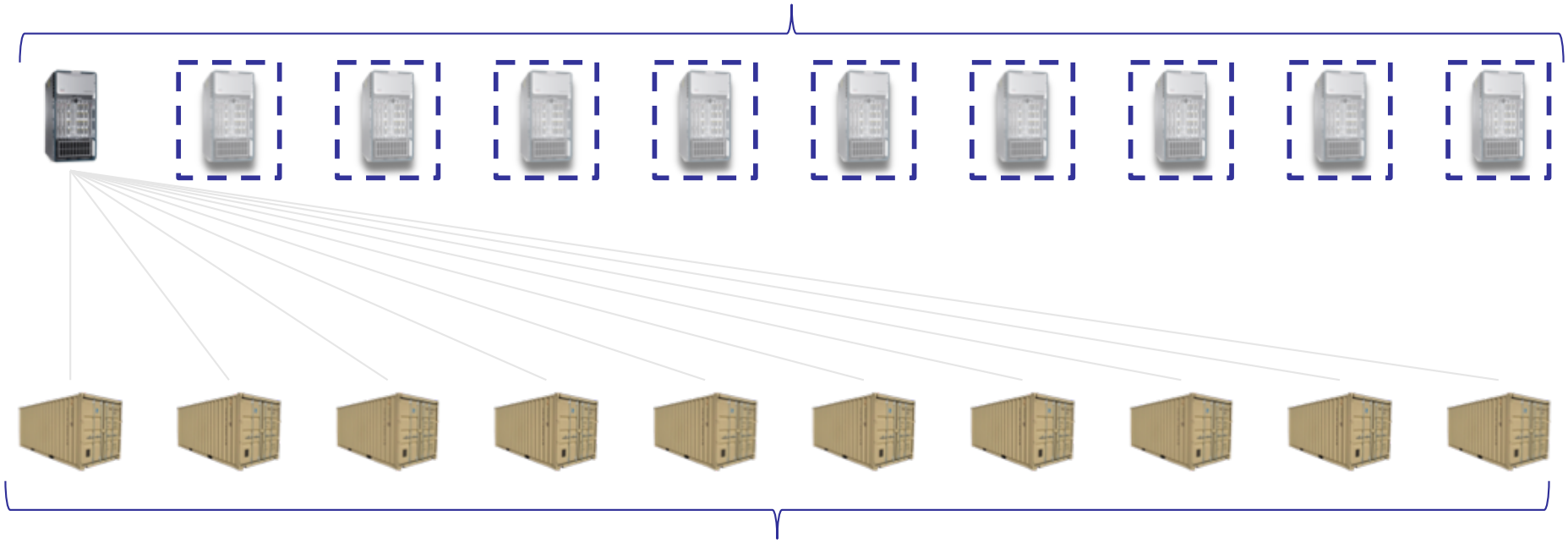


1. Full crossbar switch
2. Does not decode packets
3. Needs external scheduler

Wavelength Division Mux



k switches, N-ports each

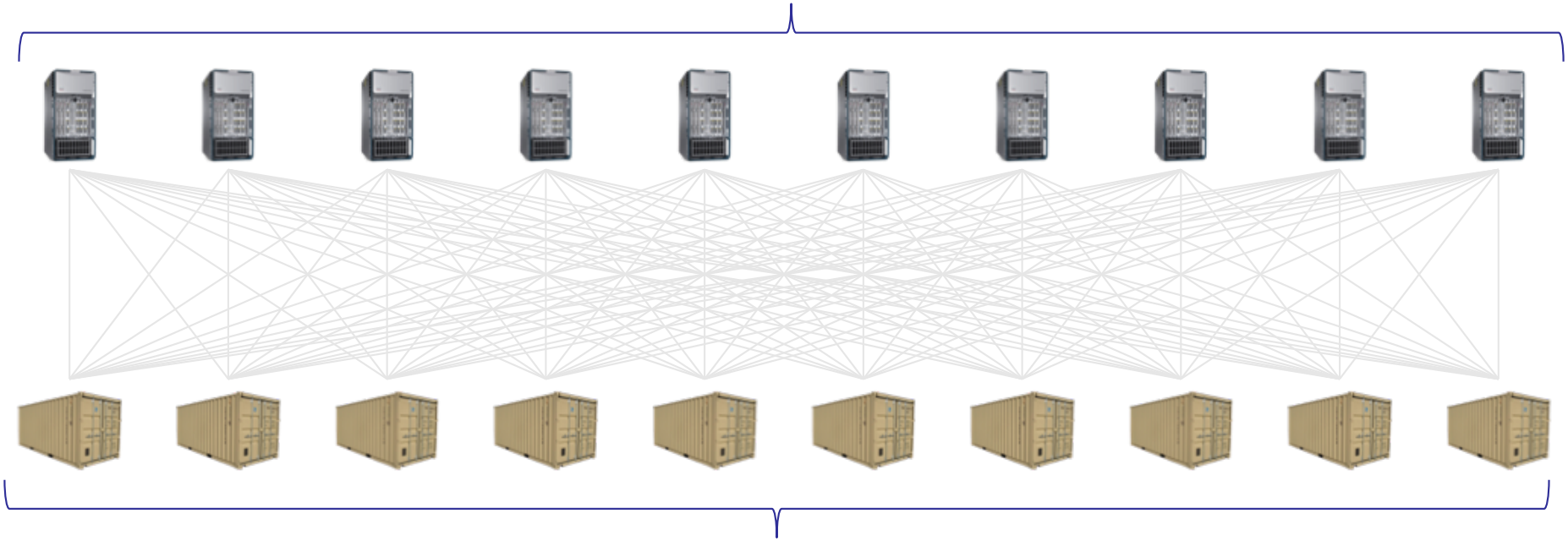


N pods, k-ports each

Example: N=64 pods * k=1024 hosts/pod = 64K hosts total; 8 wavelengths

Bisection Bandwidth	10% Electrical (10:1 Oversubscribed)	100% Electrical	Helios Example 10% Electrical + 90% Optical
Cost	$6.3 M		
Power	96.5 kW		
Cables	6,656		

k switches, N-ports each

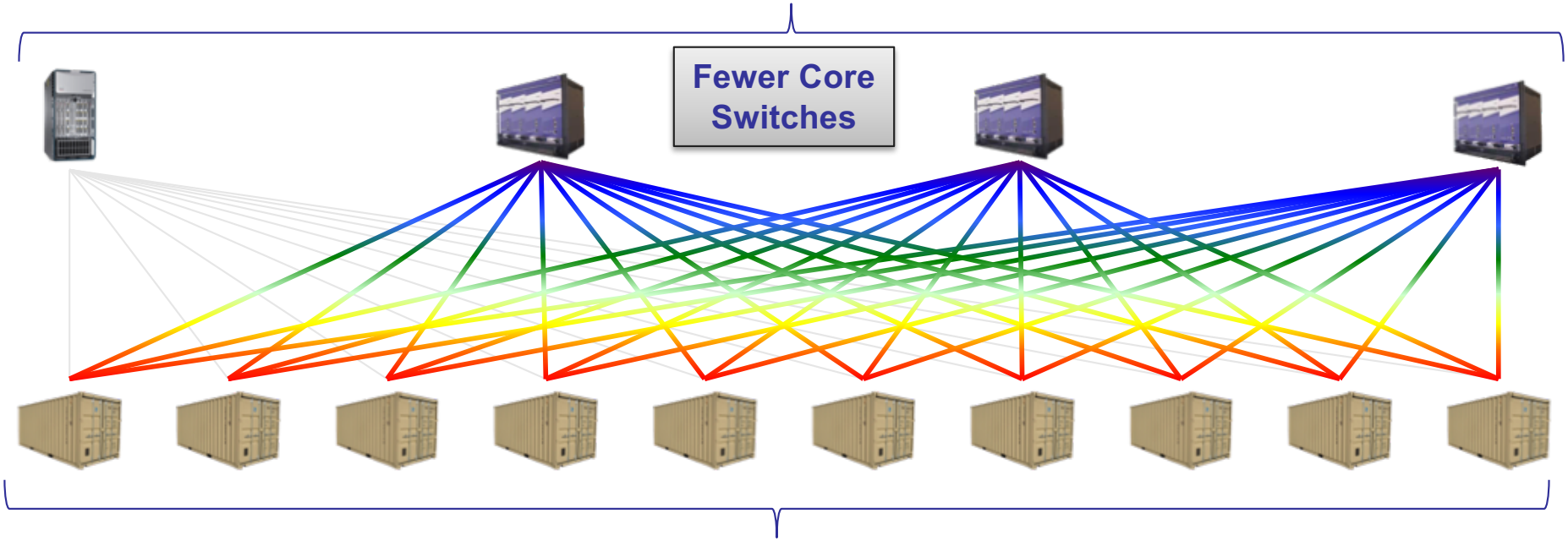


N pods, k-ports each

Example: N=64 pods * k=1024 hosts/pod = 64K hosts total; 8 wavelengths

Bisection Bandwidth	10% Electrical (10:1 Oversubscribed)	100% Electrical	Helios Example 10% Electrical + 90% Optical
Cost	$6.3 M	$62.3 M	
Power	96.5 kW	950.3 kW	
Cables	6,656	65,536	

Less than *k* switches, N-ports each



N pods, k-ports each

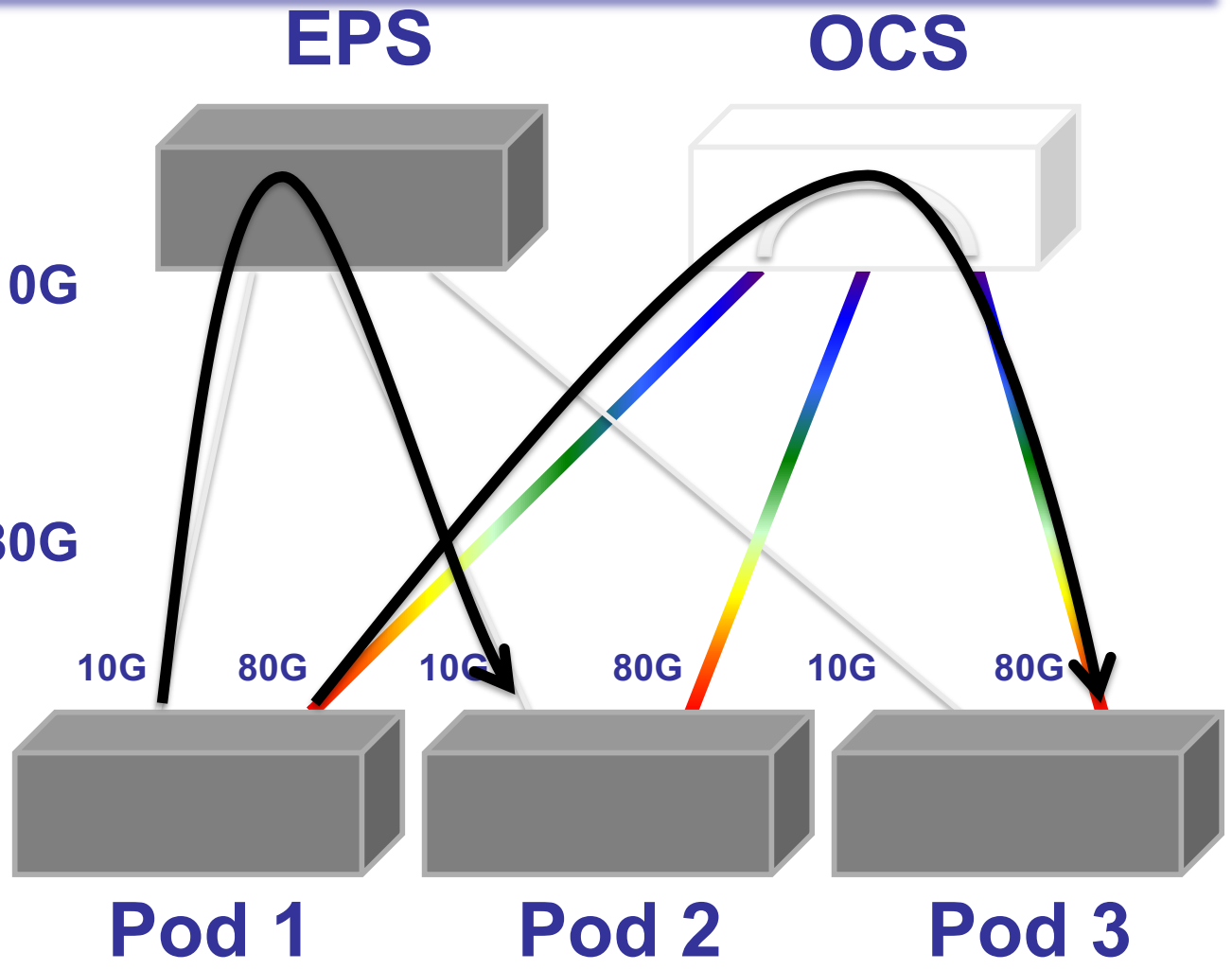
Example: N=64 pods * k=1024 hosts/pod = 64K hosts total; 8 wavelengths

Bisection Bandwidth	10% Electrical (10:1 Oversubscribed)	100% Electrical	Helios Example 10% Electrical + 90% Optical	
Cost	$6.3 M	$62.2 M	$22.1 M	**2.8x Less**
Power	96.5 kW	950.3 kW	157.2 kW	**6.0x Less**
Cables	6,656	65,536	14,016	**4.7x Less**

Setup a Circuit

Pod 1 -> 2:
- Capacity = 10G
- Demand = 10G
- Throughput = 10G

Pod 1 -> 3:
- Capacity = 80G
- Demand = 80G
- Throughput = 80G



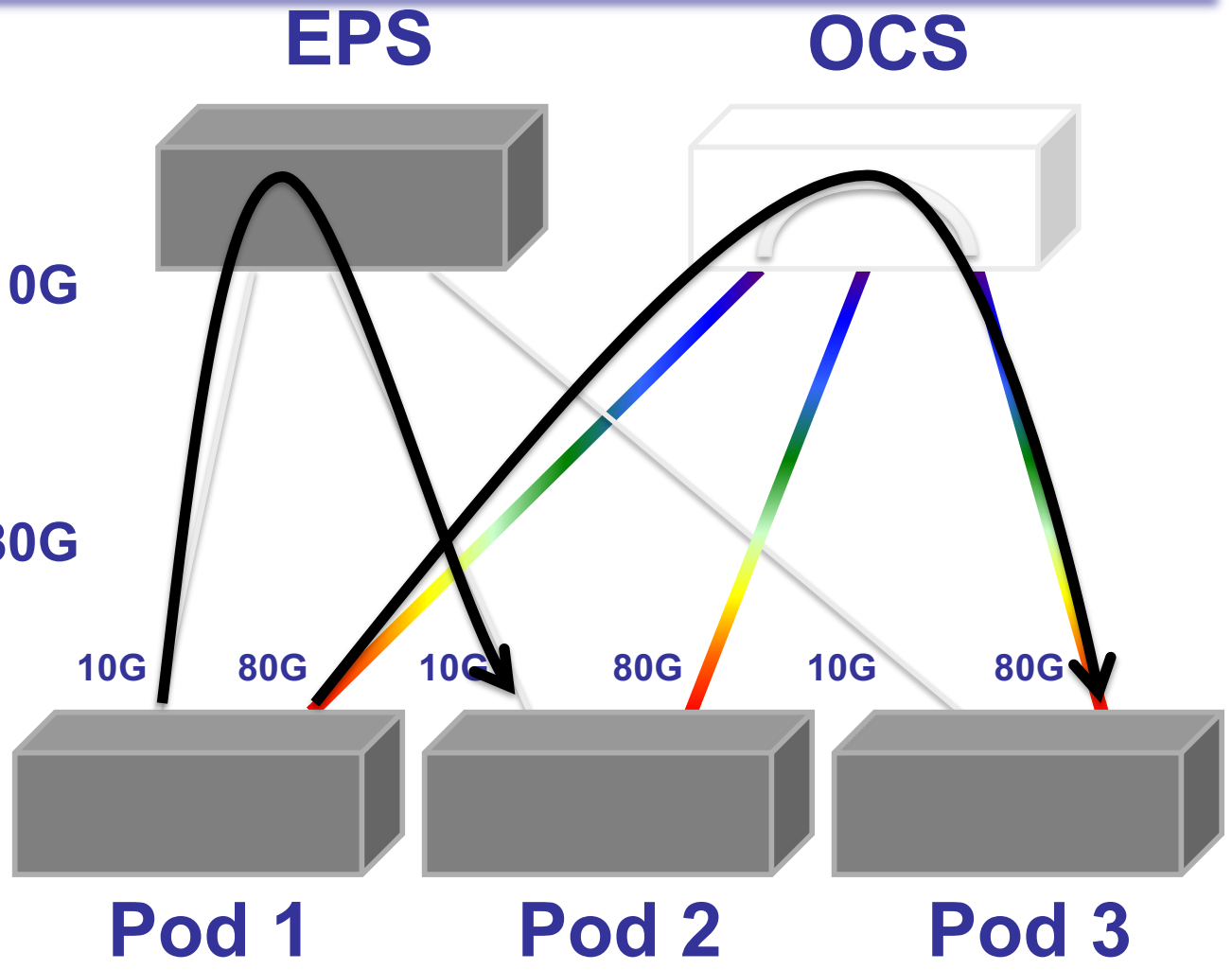
Traffic Patterns Change

Pod 1 -> 2:

- **Capacity = 10G**
- **Demand = 10G**
- **Throughput = 10G**

Pod 1 -> 3:

- **Capacity = 80G**
- **Demand = 80G**
- **Throughput = 80G**



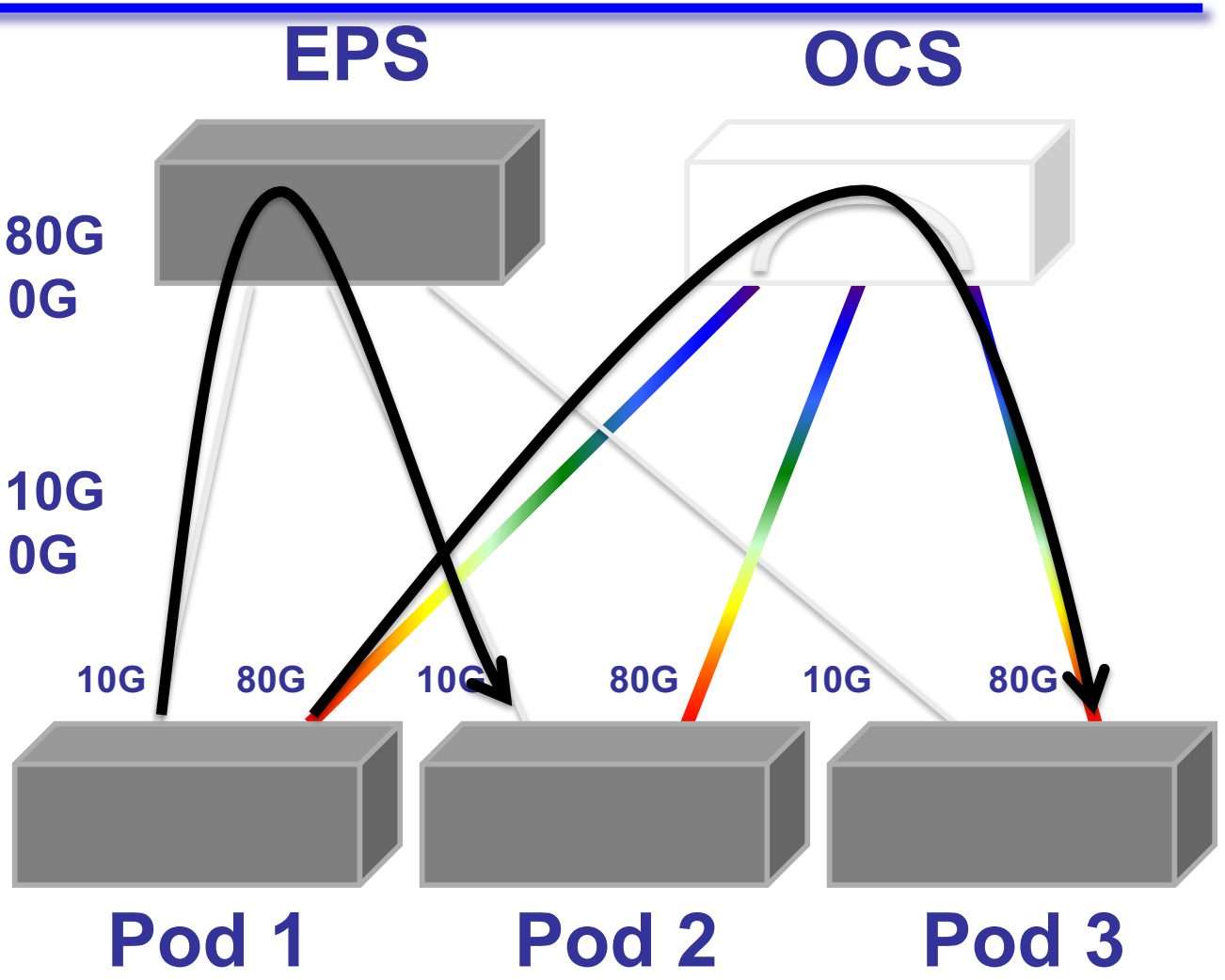
Traffic Patterns Change

Pod 1 -> 2:

- **Capacity = 10G**
- **Demand = ~~10G~~ 80G**
- **Throughput = 10G**

Pod 1 -> 3:

- **Capacity = 80G**
- **Demand = ~~80G~~ 10G**
- **Throughput = 10G**



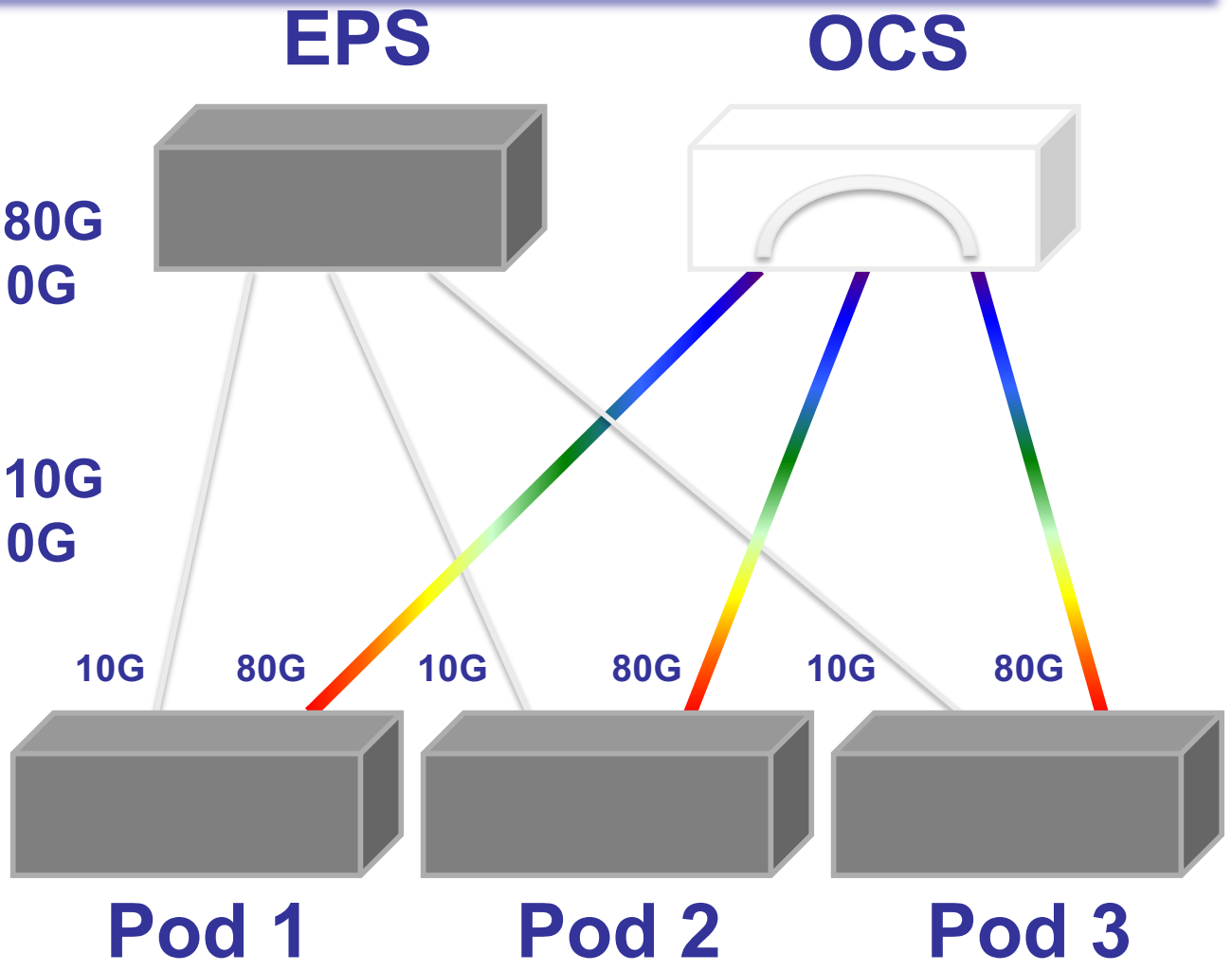
Break a Circuit

Pod 1 -> 2:

- **Capacity = 10G**
- **Demand = ~~10G~~ 80G**
- **Throughput = 10G**

Pod 1 -> 3:

- **Capacity = 80G**
- **Demand = ~~80G~~ 10G**
- **Throughput = 10G**



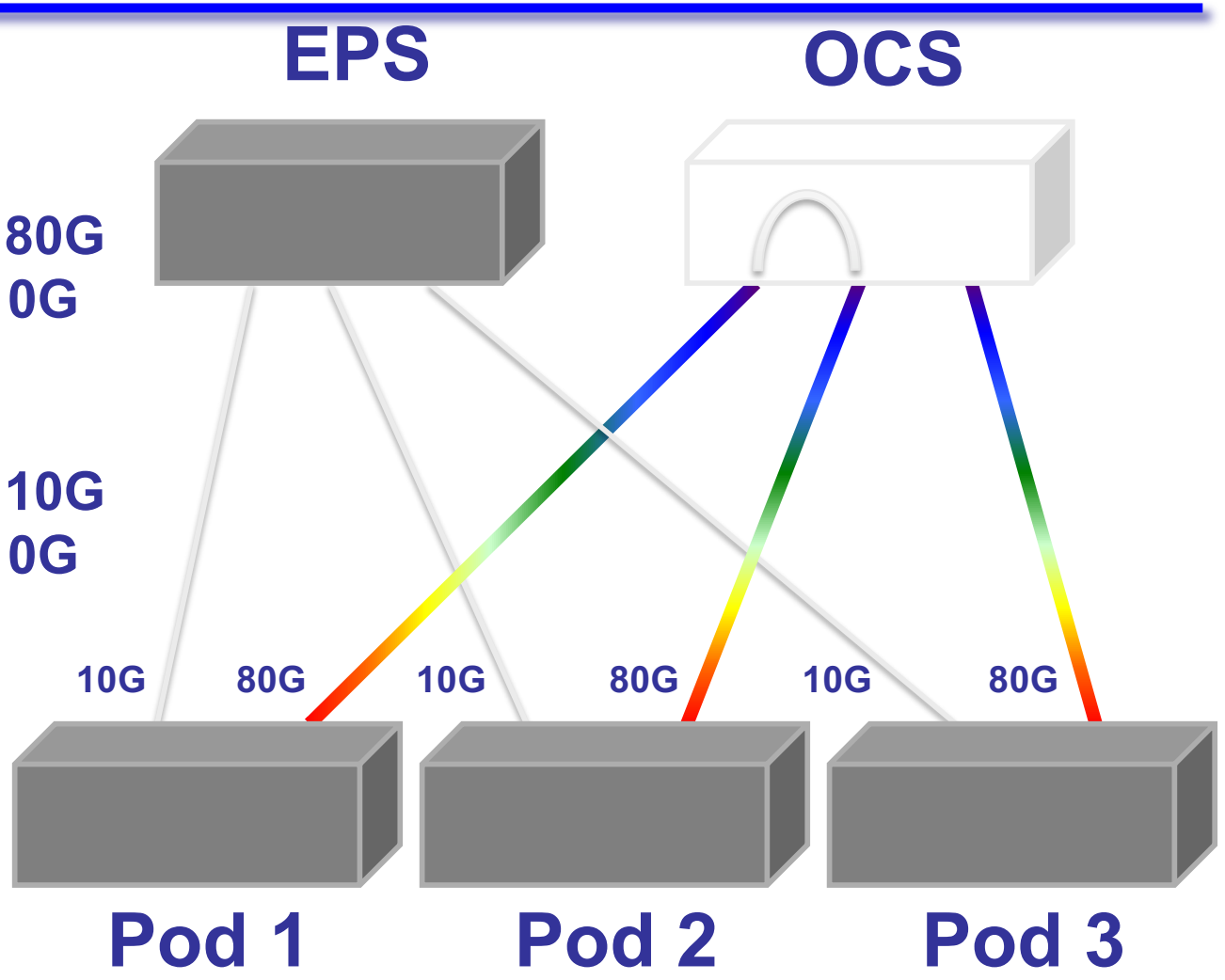
Setup a Circuit

Pod 1 -> 2:

- **Capacity = 10G**
- **Demand = ~~10G~~ 80G**
- **Throughput = 10G**

Pod 1 -> 3:

- **Capacity = 80G**
- **Demand = ~~80G~~ 10G**
- **Throughput = 10G**

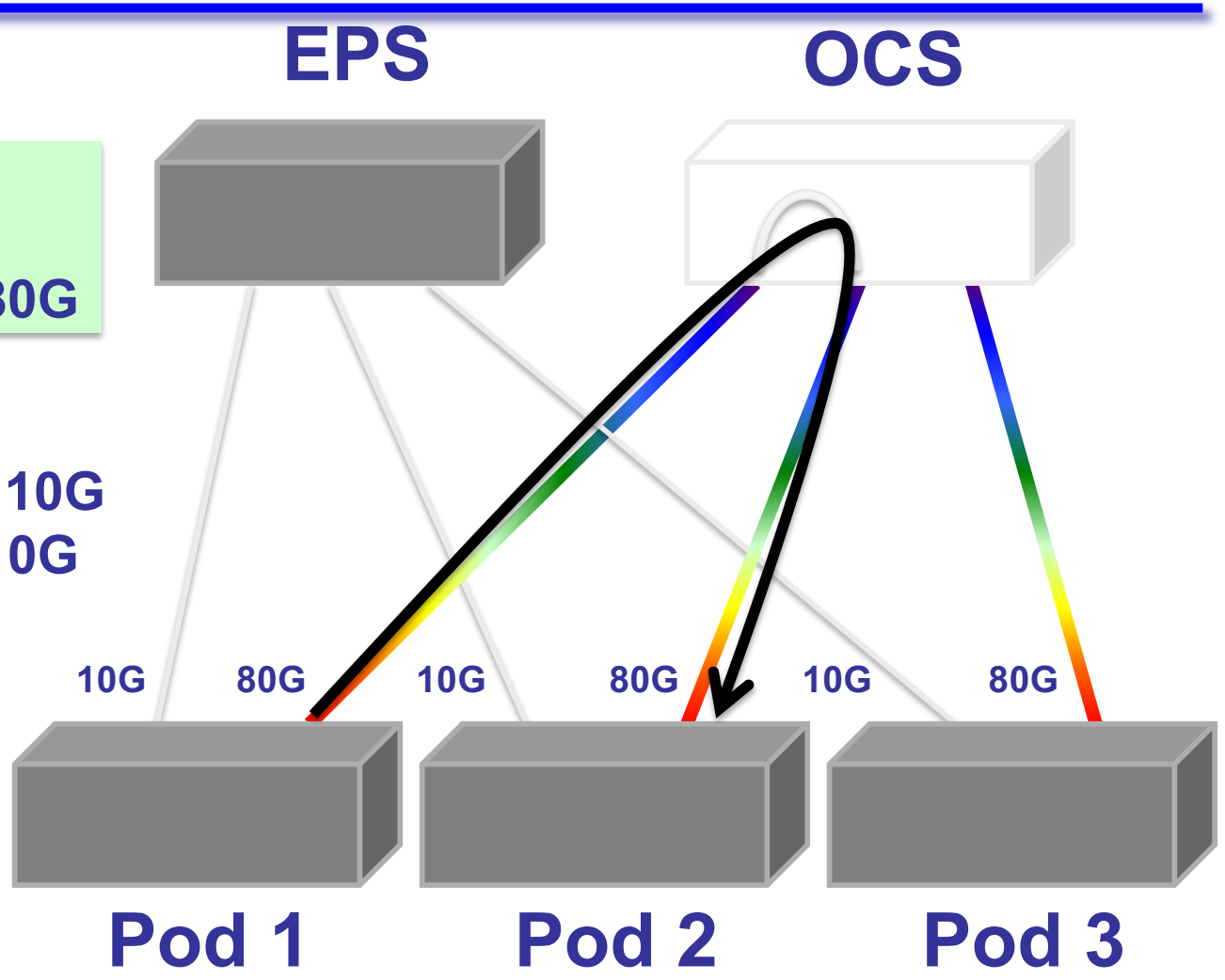


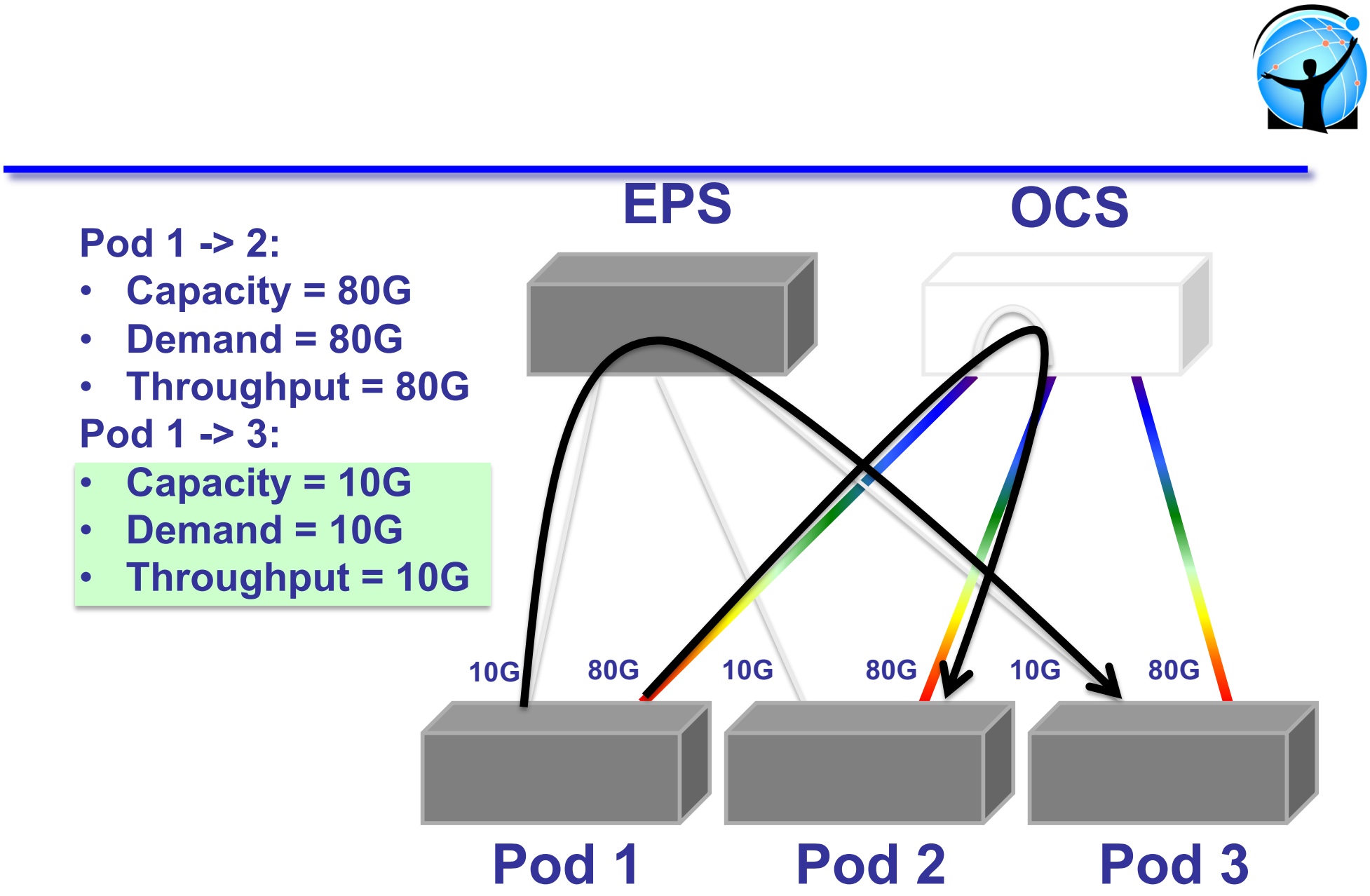
Pod 1 -> 2:

- Capacity = 80G
- Demand = 80G
- Throughput = 80G

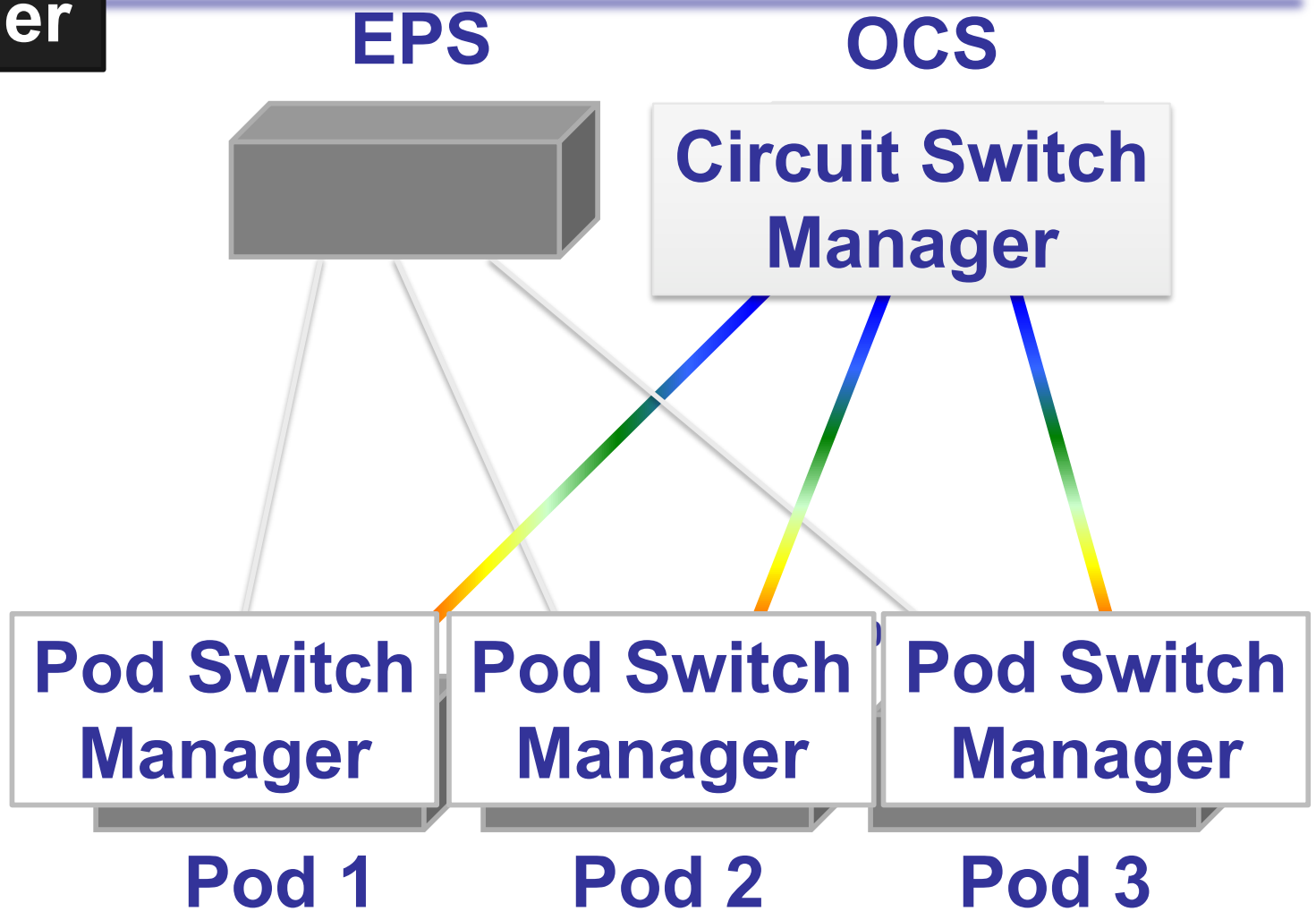
Pod 1 -> 3:

- Capacity = 80G
- Demand = ~~80G~~ 10G
- Throughput = 10G




Pod 1 -> 2:
• Capacity = 80G
• Demand = 80G
• Throughput = 80G
Pod 1 -> 3:
• Capacity = 10G
• Demand = 10G
• Throughput = 10G
EPS
OCS
10G
80G
10G
80G
10G
80G
Pod 1
Pod 2
Pod 3

Topology Manager



Outline of Control Loop

1. Estimate traffic demand
2. Compute optimal topology for maximum throughput
3. Program the pod switches and circuit switches

1. Estimate Traffic Demand

Question: Will this flow use more bandwidth if we give it more capacity?

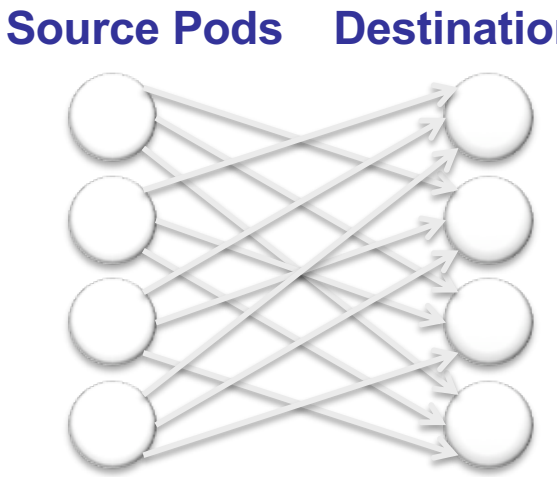
1. Identify elephant flows (mice don't grow)

   Problem: Measurements are biased by current topology
2. Pretend all hosts are connected to an ideal crossbar switch
3. Compute the max-min fair bandwidth fixpoint

Mohammad Al-Fares, Sivasankar Radhakrishnan, Barath Raghavan, Nelson Huang, and Amin Vahdat. *Hedera: Dynamic Flow Scheduling for Data Center Networks*. In NSDI'10.

2. Compute Optimal Topology

1. Formulate as instance of max-weight perfect matching problem on bipartite graph
2. Solve with Edmonds algorithm



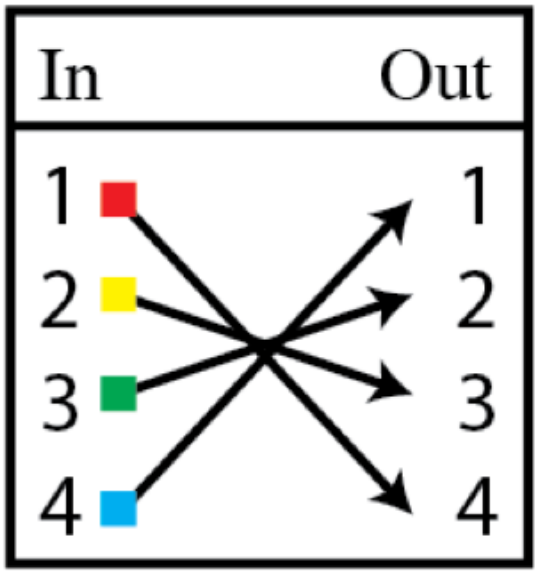
a) **Pods do not send traffic to themselves**
b) **Edge weights represent interpod demand**
c) **Algorithm is run iteratively for each circuit switch, making use of the previous results**

Example: Compute Optimal Topology

Demand Matrix i

	1	2	3	4
1	0	1	1	**3**
2	$7_{(4)}$	0	**3**	1
3	1	**3**	0	$9_{(4)}$
4	**3**	2	1	0

Circuit Switch i

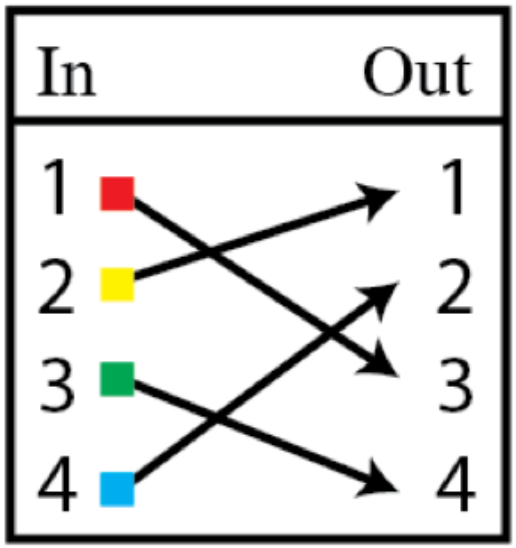


Example: Compute Optimal Topology

Demand Matrix i+1

	1	2	3	4
1	0	1	**1**	0
2	**7**(4)	0	0	1
3	1	0	0	**9**(4)
4	0	**2**	1	0

Circuit Switch i+1

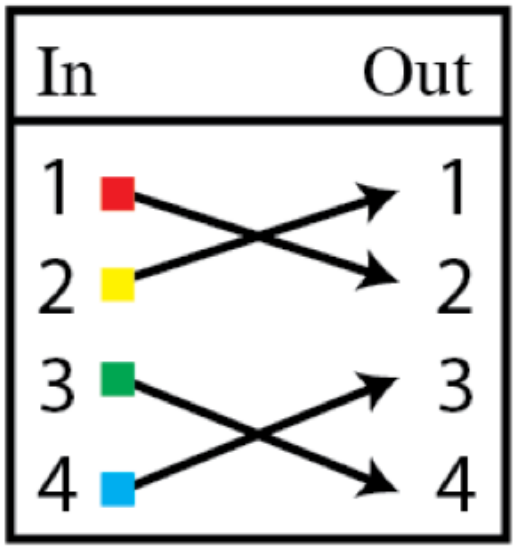


Example: Compute Optimal Topology

Demand Matrix i+2

	1	2	3	4
1	0	**1**	0	0
2	**3**	0	0	1
3	1	0	0	**5** (4)
4	0	0	**1**	0

Circuit Switch i+2



Traditional Network

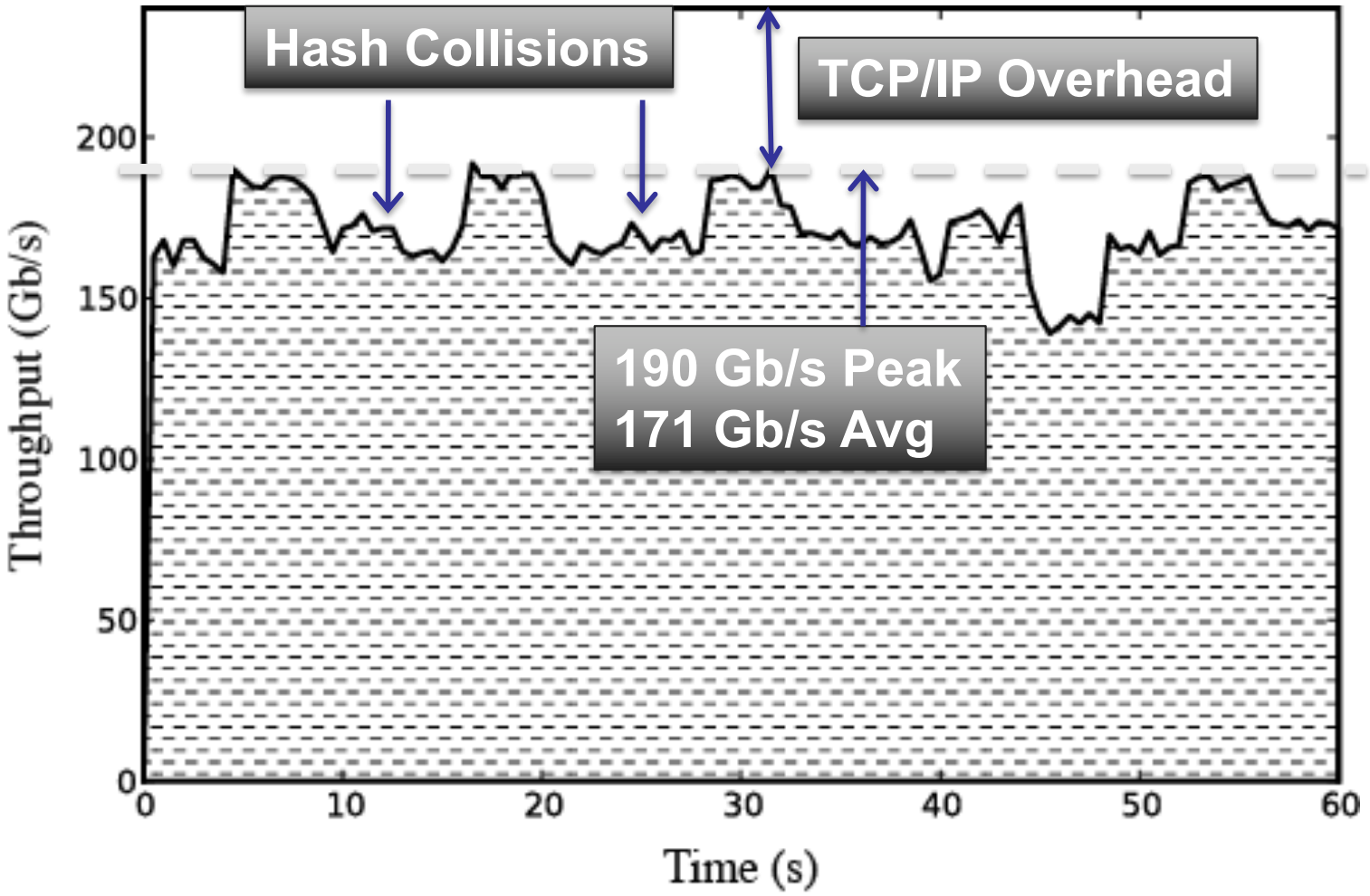
Helios Network

100% bisection bandwidth (240 Gb/s)

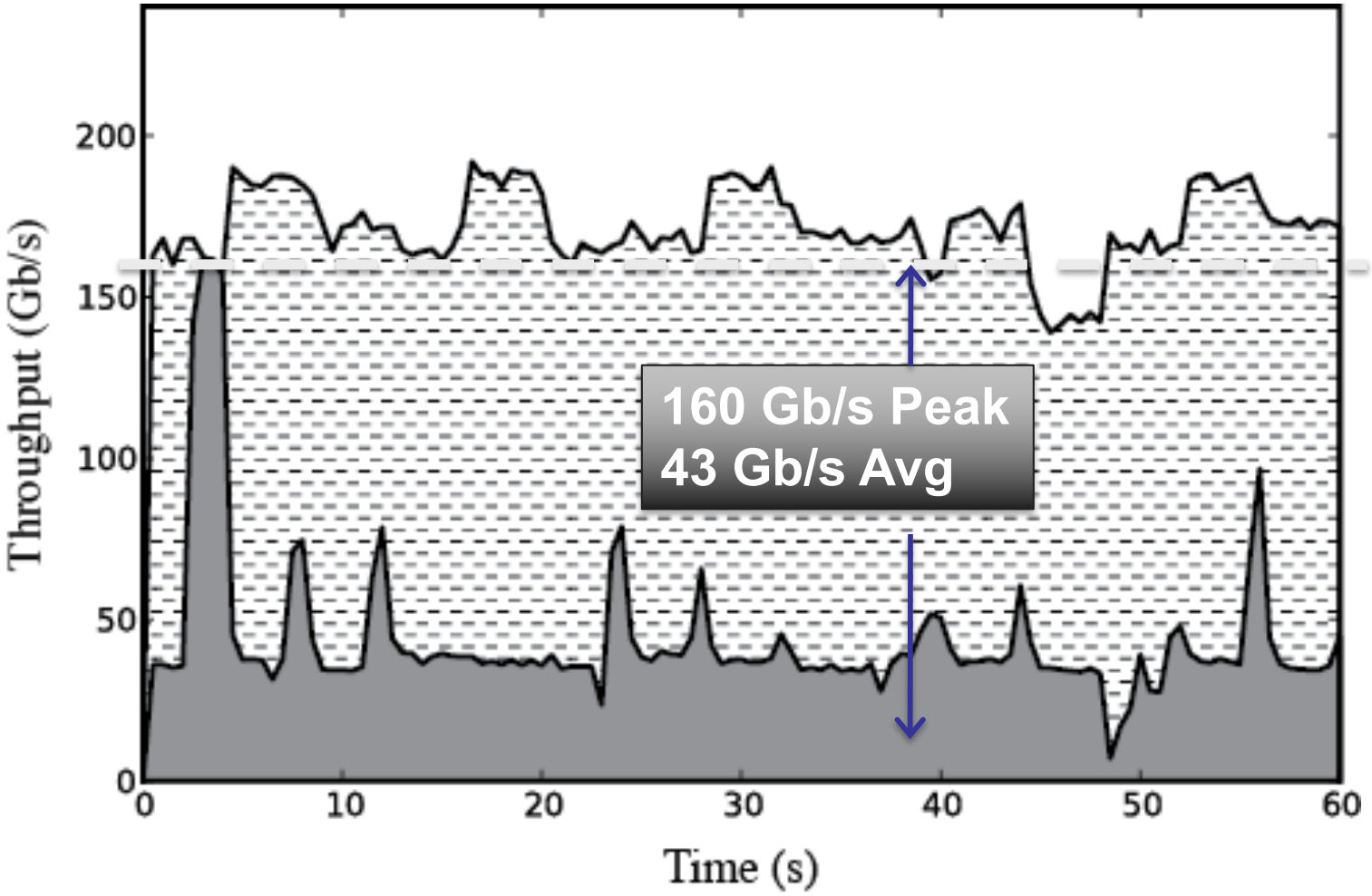
Hardware

- 24 servers
  - HP DL380
  - 2 socket (E5520) Nehalem
  - Dual Myricom 10G NICs
- 7 switches
  - One Dell 1G 48-port
  - Three Fulcrum 10G 24-port
  - One Glimmerglass 64-port optical circuit switch
  - Two Cisco Nexus 5020 10G 52-port

Traditional Network

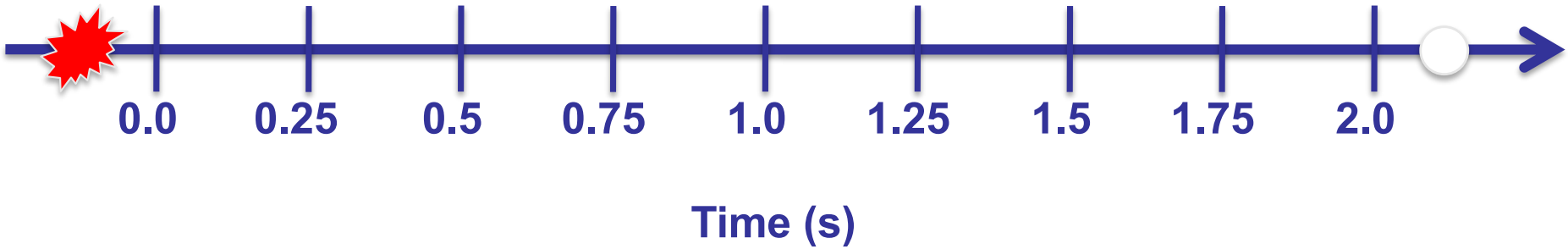


Helios Network (Baseline)

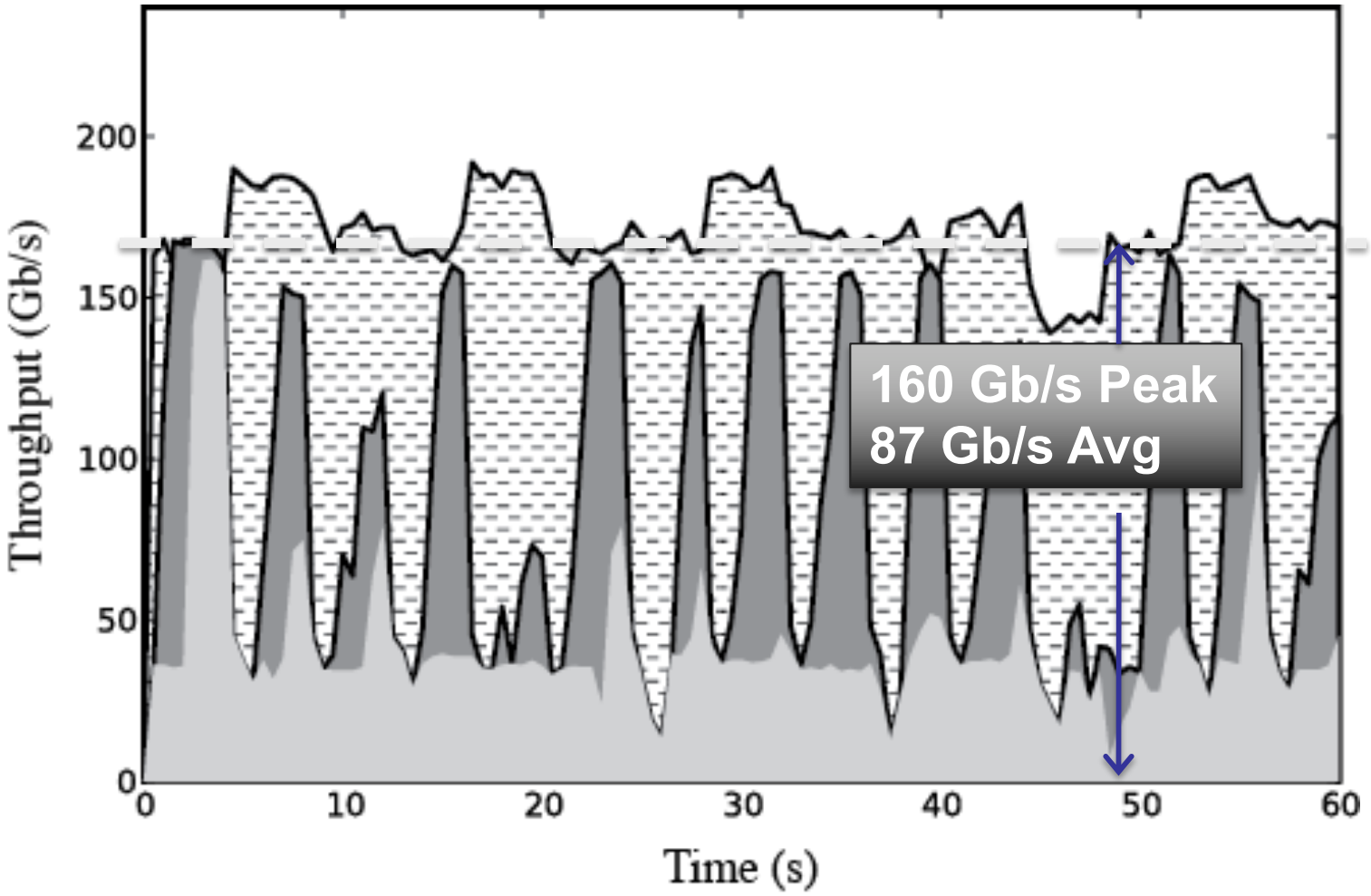


Port Debouncing

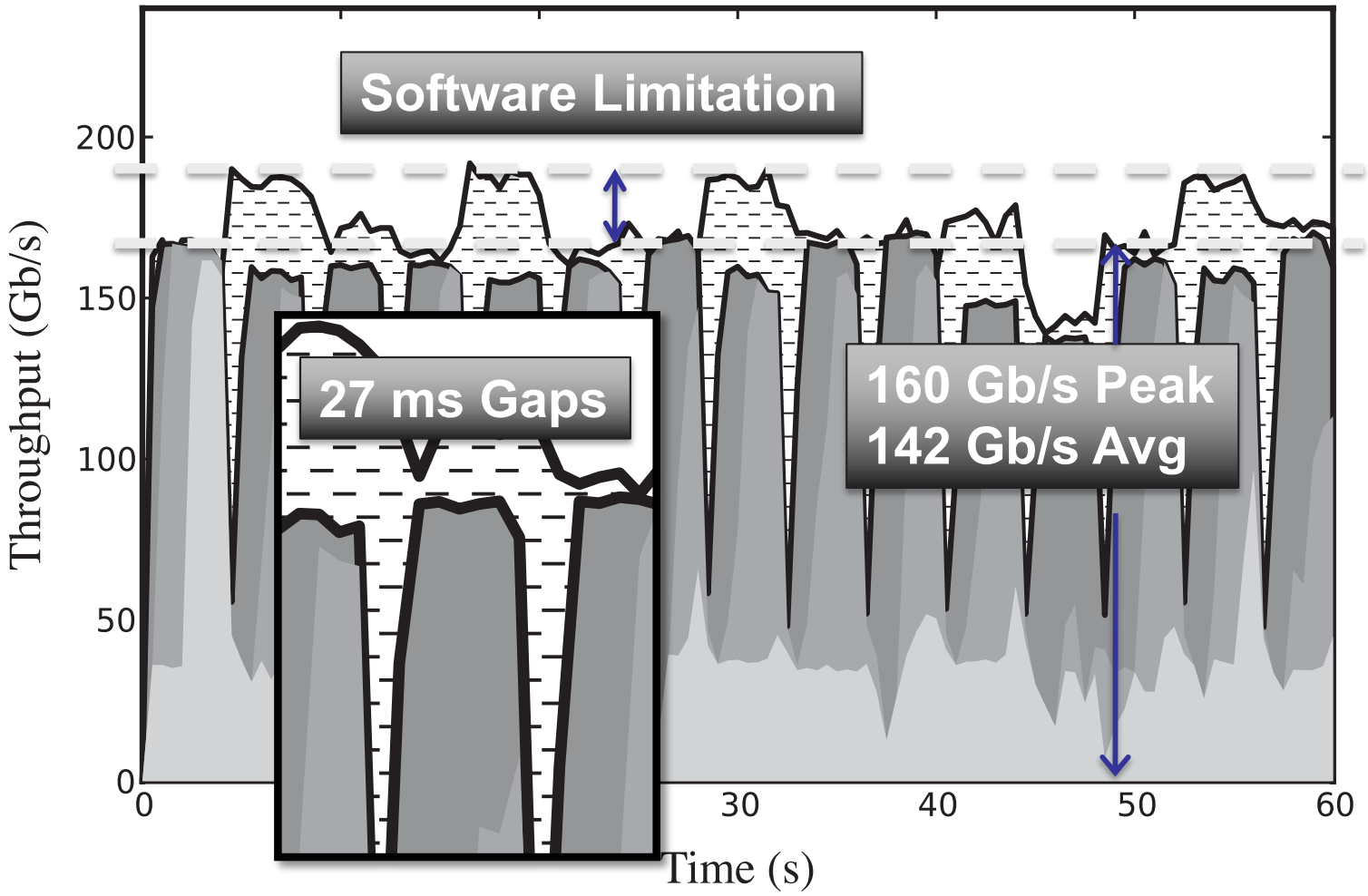
1. Layer 1 PHY signal locked (bits are detected)
2. Switch thread wakes up and polls for PHY status
   - Makes note to enable link after 2 seconds
3. Switch thread enables Layer 2 link



Without Debouncing



Without EDC


Software Limitation
27 ms Gaps
160 Gb/s Peak
142 Gb/s Avg
Throughput (Gb/s)
Time (s)
0
50
100
150
200
0
30
40
50
60

For Next Class…

- Read and review FairCloud paper
- Keep moving on term projects!