OSA: An **O**ptical **S**witching **A**rchitecture for Data Center Networks with Unprecedented Flexibility

Kai Chen, Ankit Singla, Atul Singh, Kishore Ramachandran, Lei Xu, Yueping, Zhang, Xitao Wen, Yan Chen

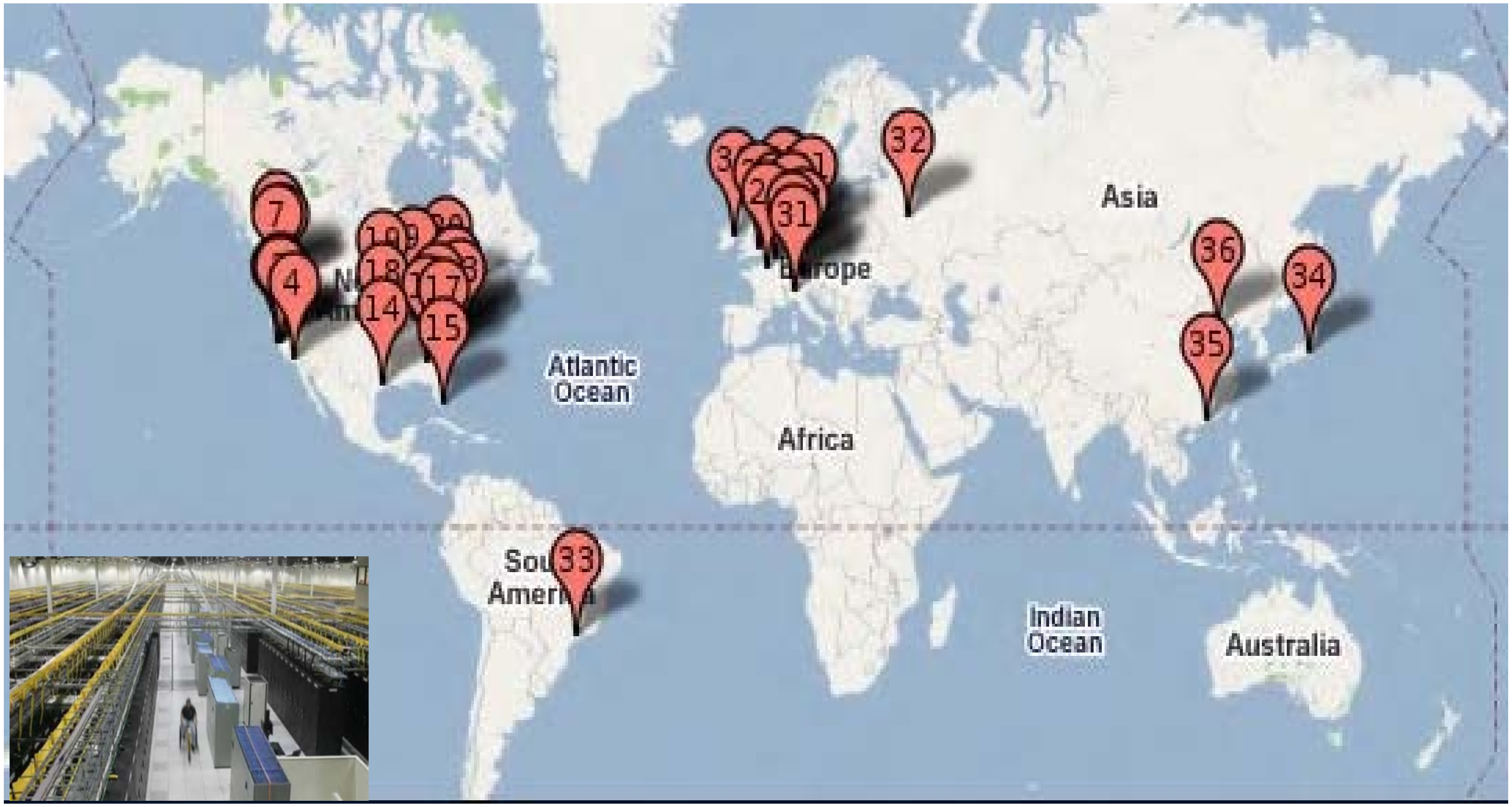
Northwestern University, UIUC, NEC Labs America

USENIX NSDI’12, San Jose, USA

Big Data for Modern Applications

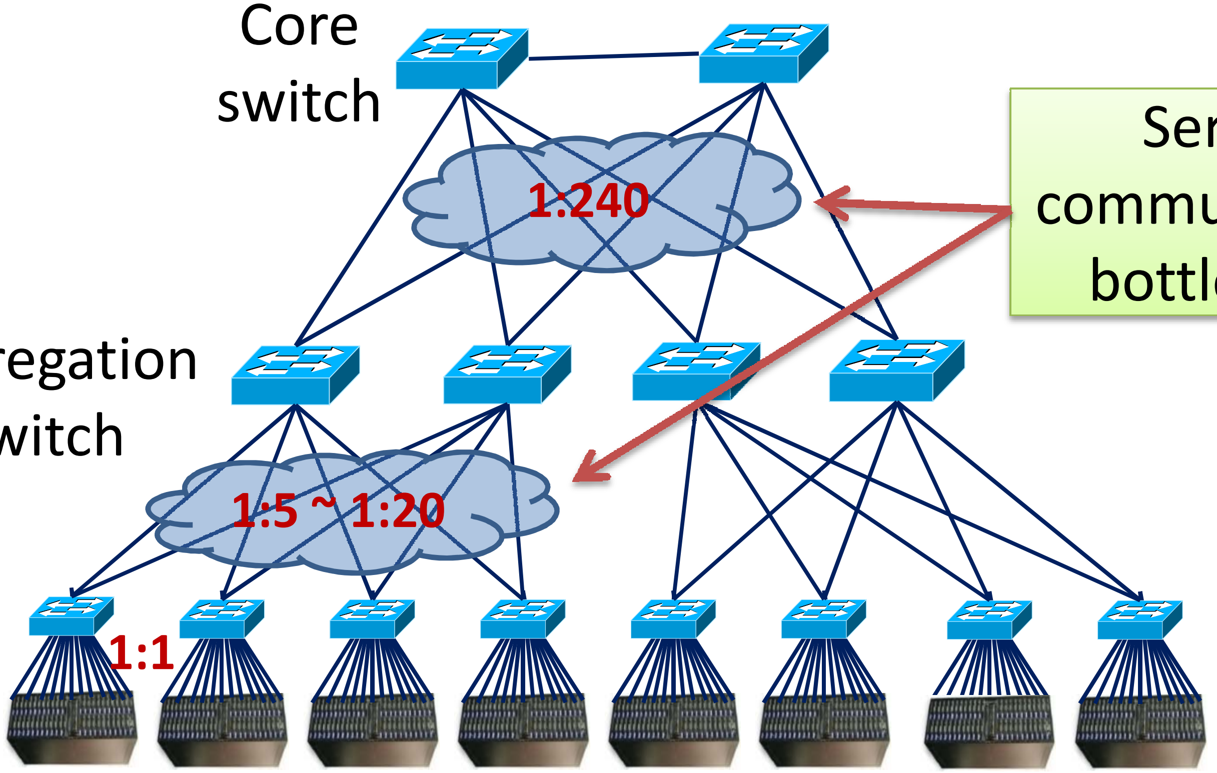
- Scientific: 200GB of astronomy data a night
- Business: 1 million customer transactions, 2.5PB of data per hour
- Social network: 60 billion photos in its user base, 25TB of log data per day
- Web search: 20PB of search data per day

Data Center as Infrastructure



Example of Google’s 36 world wide data centers

Conventional DCN is Problematic



Core switch

Aggregation switch

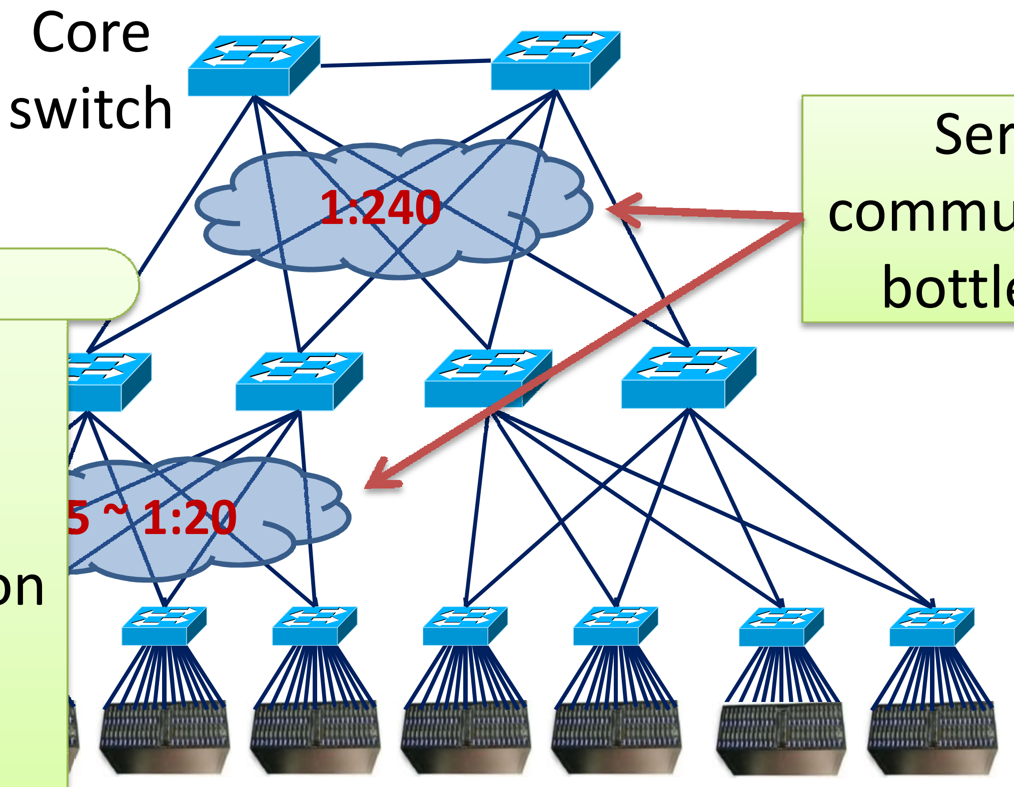
(ToR switch) Top-of-Rack

Serious communication bottleneck!

A DCN structure adapted from Cisco

Efficient DCN architecture is desirable, but challenging

Conventional DCN is Problematic



Serious communication bottleneck!

A DCN structure adapted from Cisco

Considerations:

- Bandwidth
- Wiring complexity
- Power consumption
- Network cost

…

Efficient DCN architecture is desirable, but challenging

Recent Efforts and Their Problems

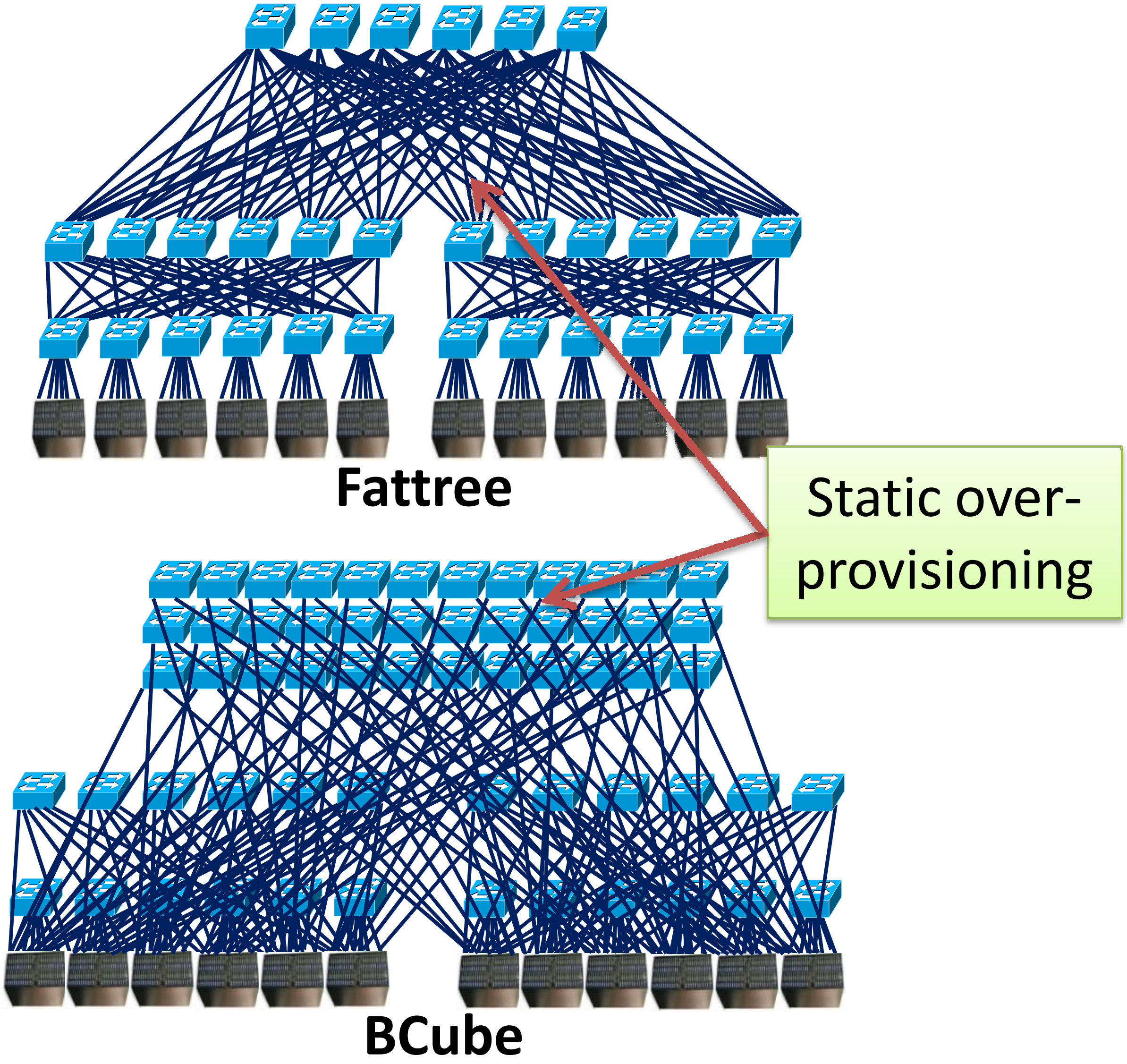
All-electrical (static)

↕

Fattree, BCube, VL2, PortLand [SIGCOMM'08 '09]

↕

High bandwidth, ***but*** high wiring complexity, high power, high cost



Recent Efforts and Their Problems

All-electrical (static)

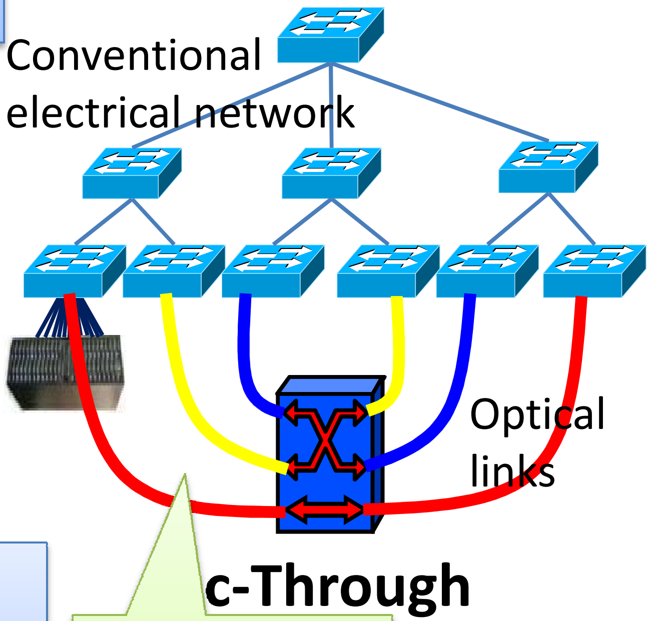
Hybrid electrical/optical (semi-flexible)

Fattree, BCube, VL2, PortLand [SIGCOMM'08 '09]

c-Through, Helios [SIGCOMM'10]

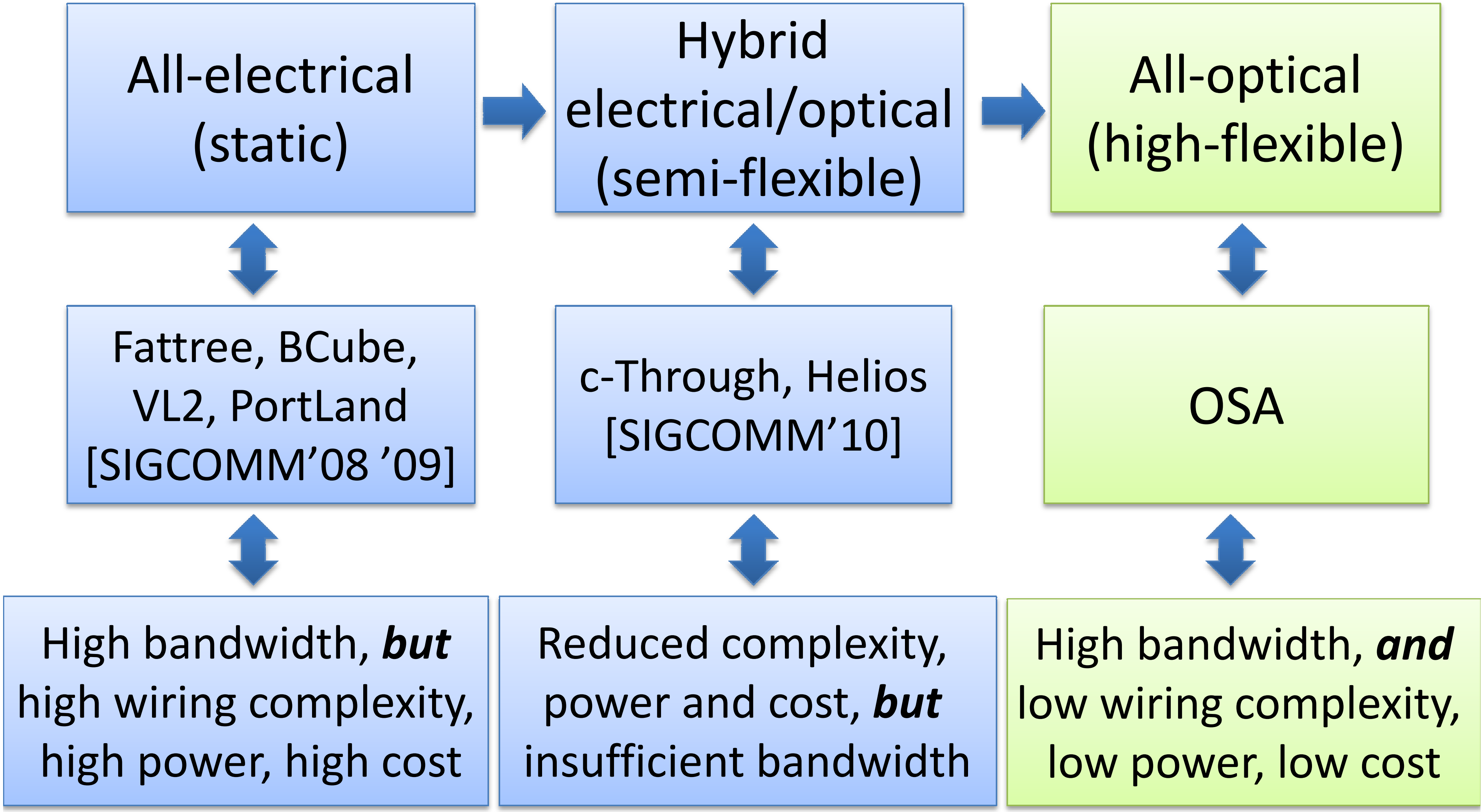
High bandwidth, ***but*** high wiring complexity, high power, high cost

Reduced complexity, power and cost, ***but*** insufficient bandwidth



Limited flexibility

Our Effort: OSA



All-electrical (static)

Hybrid electrical/optical (semi-flexible)

All-optical (high-flexible)

Fattree, BCube, VL2, PortLand [SIGCOMM'08 '09]

c-Through, Helios [SIGCOMM'10]

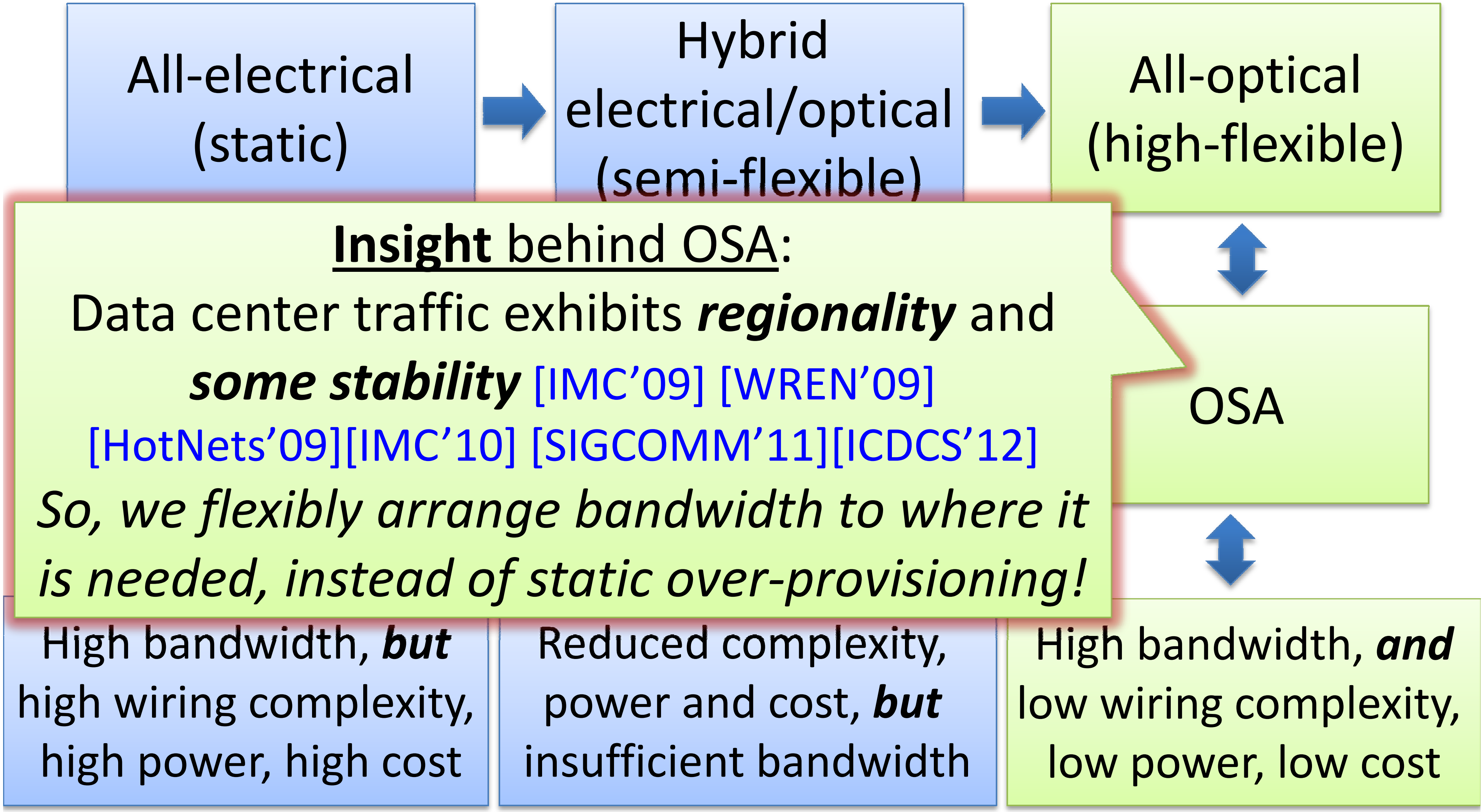
OSA

High bandwidth, ***but*** high wiring complexity, high power, high cost

Reduced complexity, power and cost, ***but*** insufficient bandwidth

High bandwidth, ***and*** low wiring complexity, low power, low cost

Our Effort: OSA



All-electrical (static)

Hybrid electrical/optical (semi-flexible)

All-optical (high-flexible)

OSA

High bandwidth, ***but*** high wiring complexity, high power, high cost

Reduced complexity, power and cost, ***but*** insufficient bandwidth

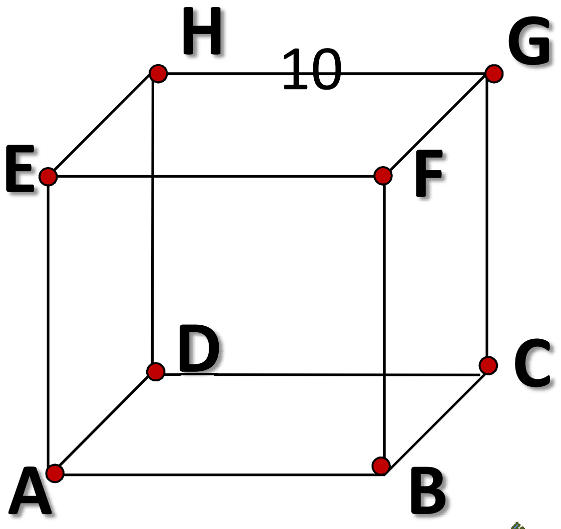
High bandwidth, ***and*** low wiring complexity, low power, low cost

Insight behind OSA:

Data center traffic exhibits ***regionality*** and ***some stability*** [IMC'09] [WREN'09] [HotNets'09][IMC'10] [SIGCOMM'11][ICDCS'12]

So, we flexibly arrange bandwidth to where it is needed, instead of static over-provisioning!

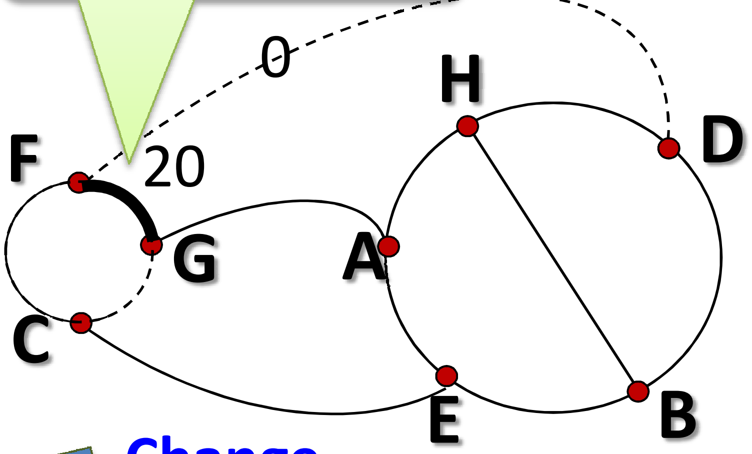
OSA’s Flexibility: An Example



Traffic demand

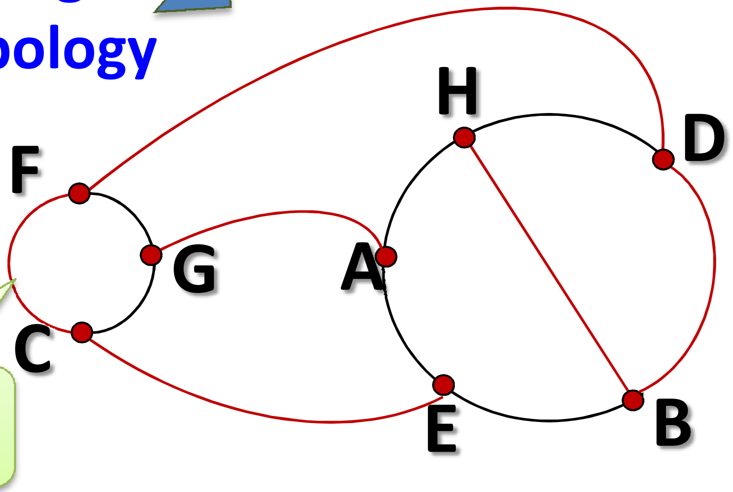
A	G	10
B	H	10
C	E	10
D	F	10
B	D	10
C	F	10

High capacity link for increased demand



Change link capacity

Change topology

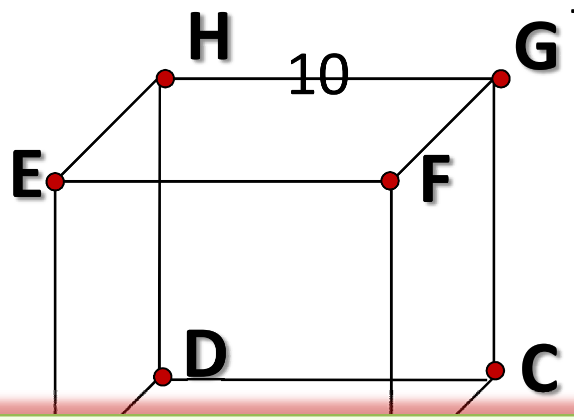


Direct link for real demand

Demand change

A	G	10
B	H	10
C	E	10
F	G	20
B	D	10
C	F	10

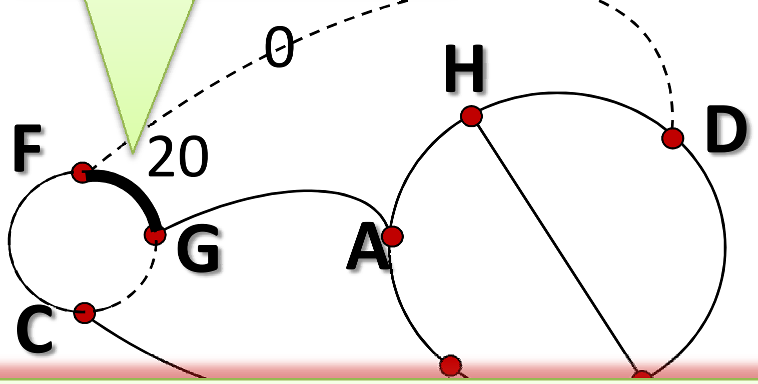
OSA’s Flexibility: An Example



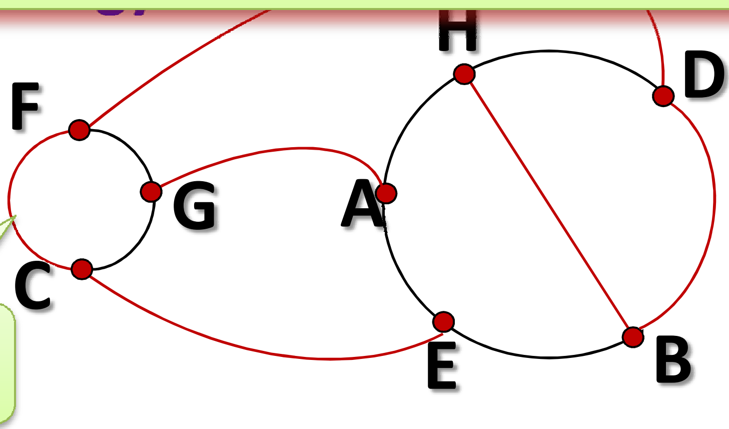
Traffic demand

A	G	10
B	H	10
C	E	10
D	F	10
B	D	10

High capacity link for increased demand



OSA can dynamically change its ToR topology and link capacity to adapt to the real demand, thus delivering high bandwidth without static over-provisioning!



Direct link for real demand

A	G	10
B	H	10
C	E	10
F	G	20
B	D	10
C	F	10

Outline of Presentation

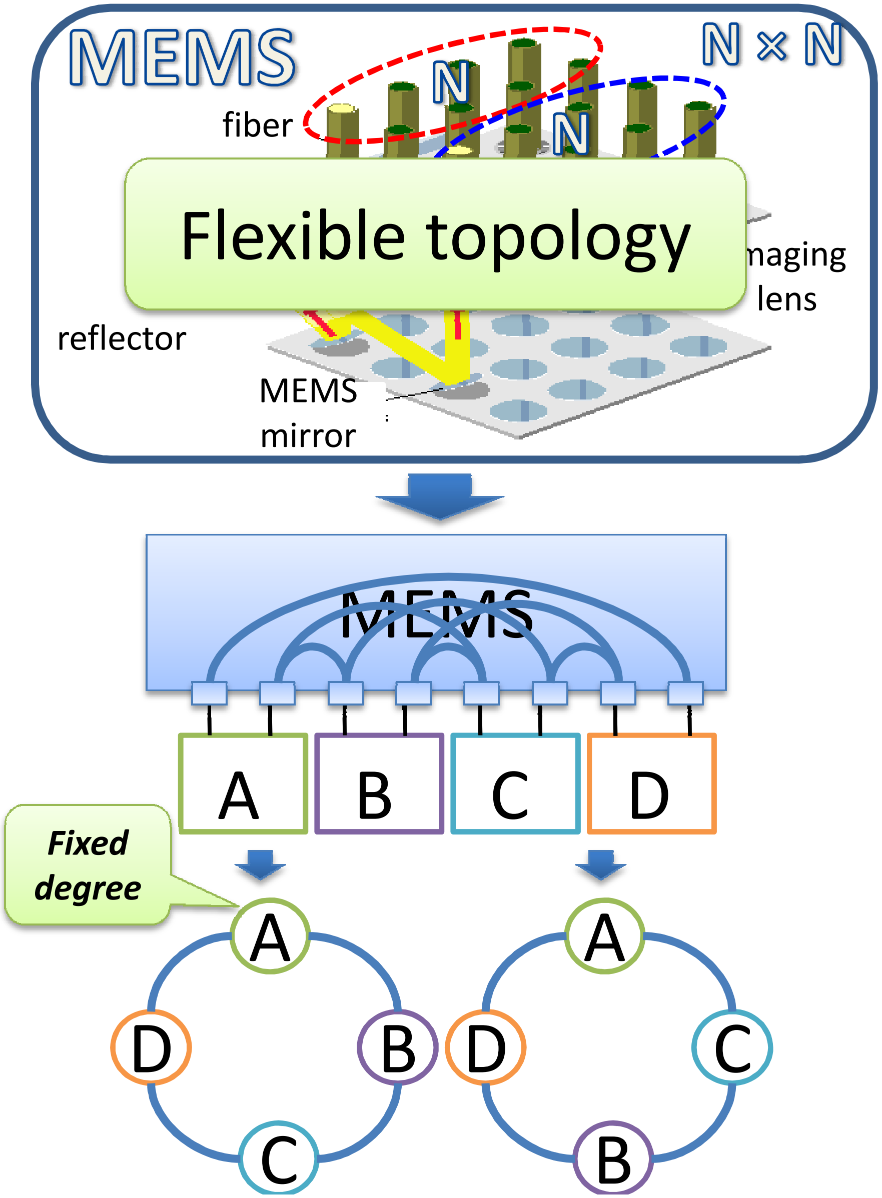
- Background and high-level idea
- How OSA achieves such flexibility?
- OSA architecture and optimization
- Implementation and Evaluation
- Summary

How We Achieve Such Flexibility?

Approach: *identify the advantages of optical network technologies, innovatively apply them in data center networking to design a flexible architecture!*

How We Achieve Such Flexibility?

Micro-Electro-Mechanical Switch

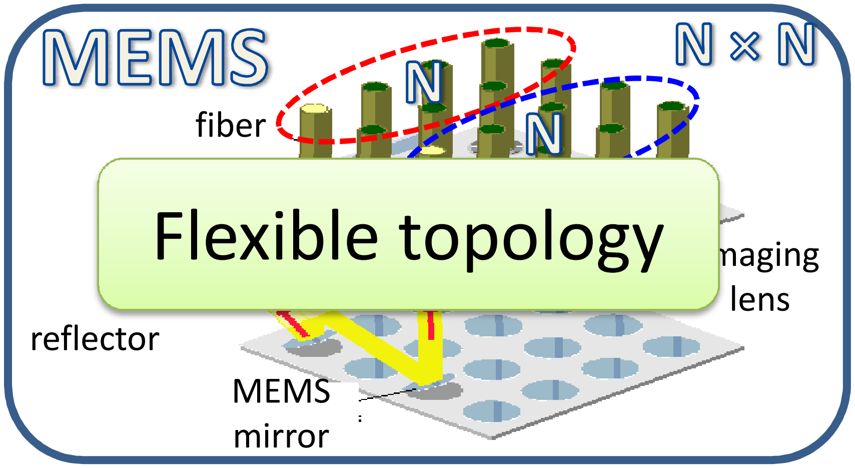


Flexible topology

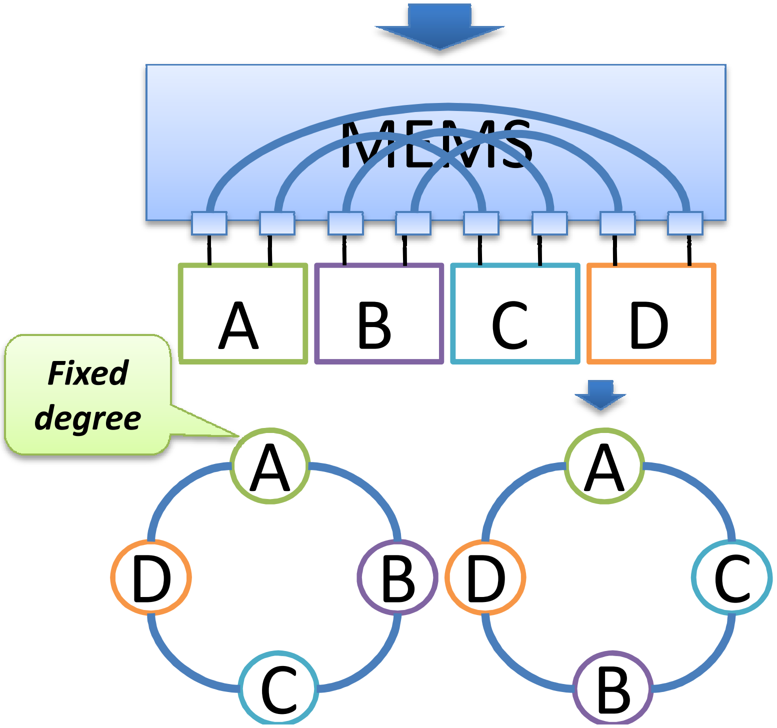
Fixed degree

How We Achieve Such Flexibility?

Micro-Electro-Mechanical Switch

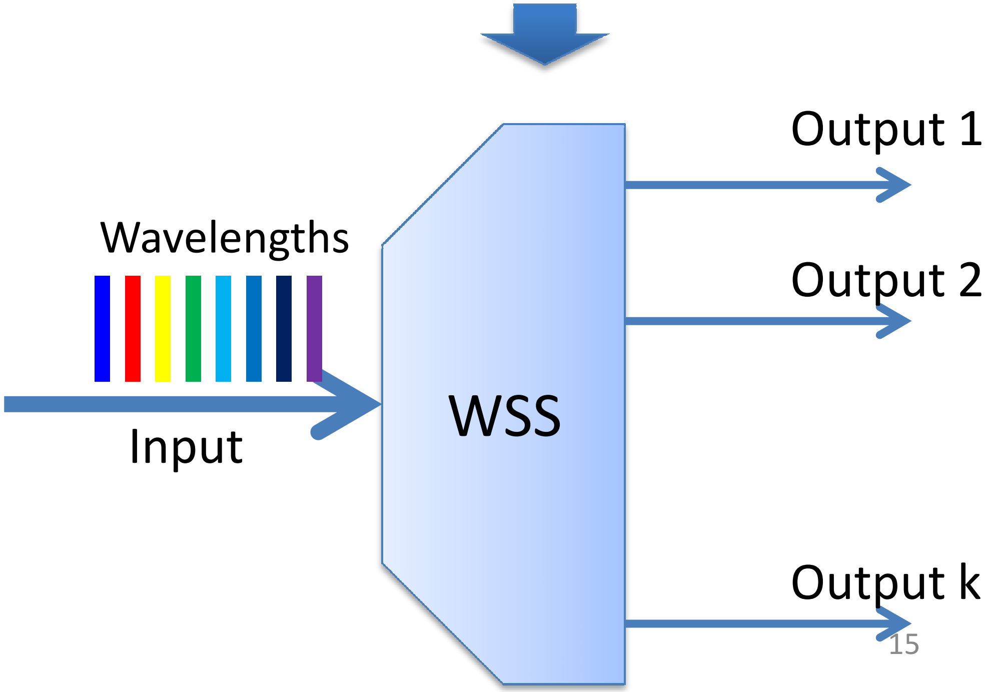


Flexible topology



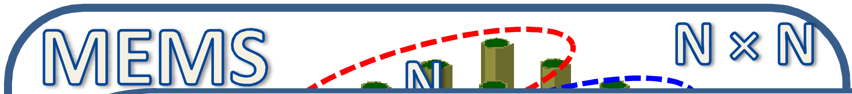
Fixed degree

Wavelength Selective Switch



How We Achieve Such Flexibility?

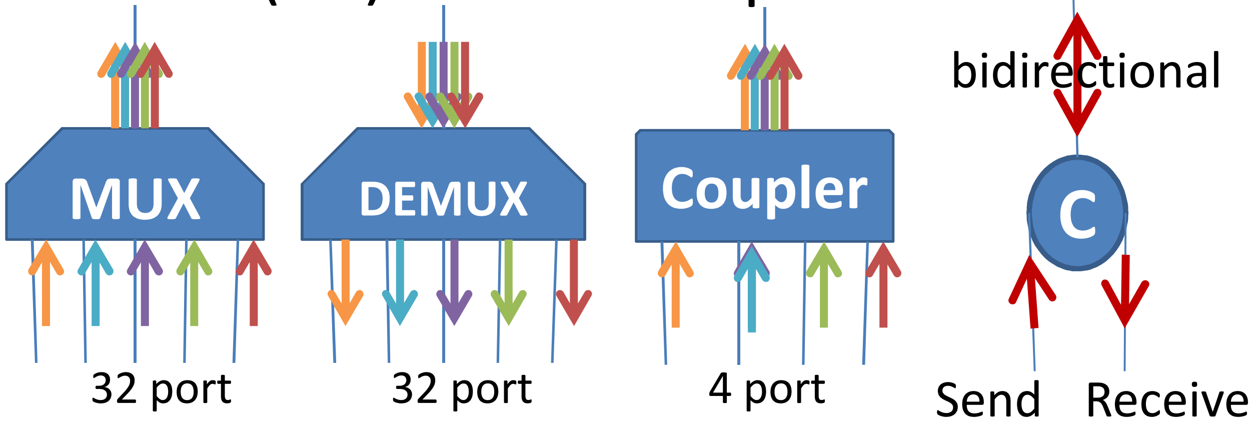
Micro-Electro-Mechanical Switch



Wavelength Selective Switch

Other optical devices:

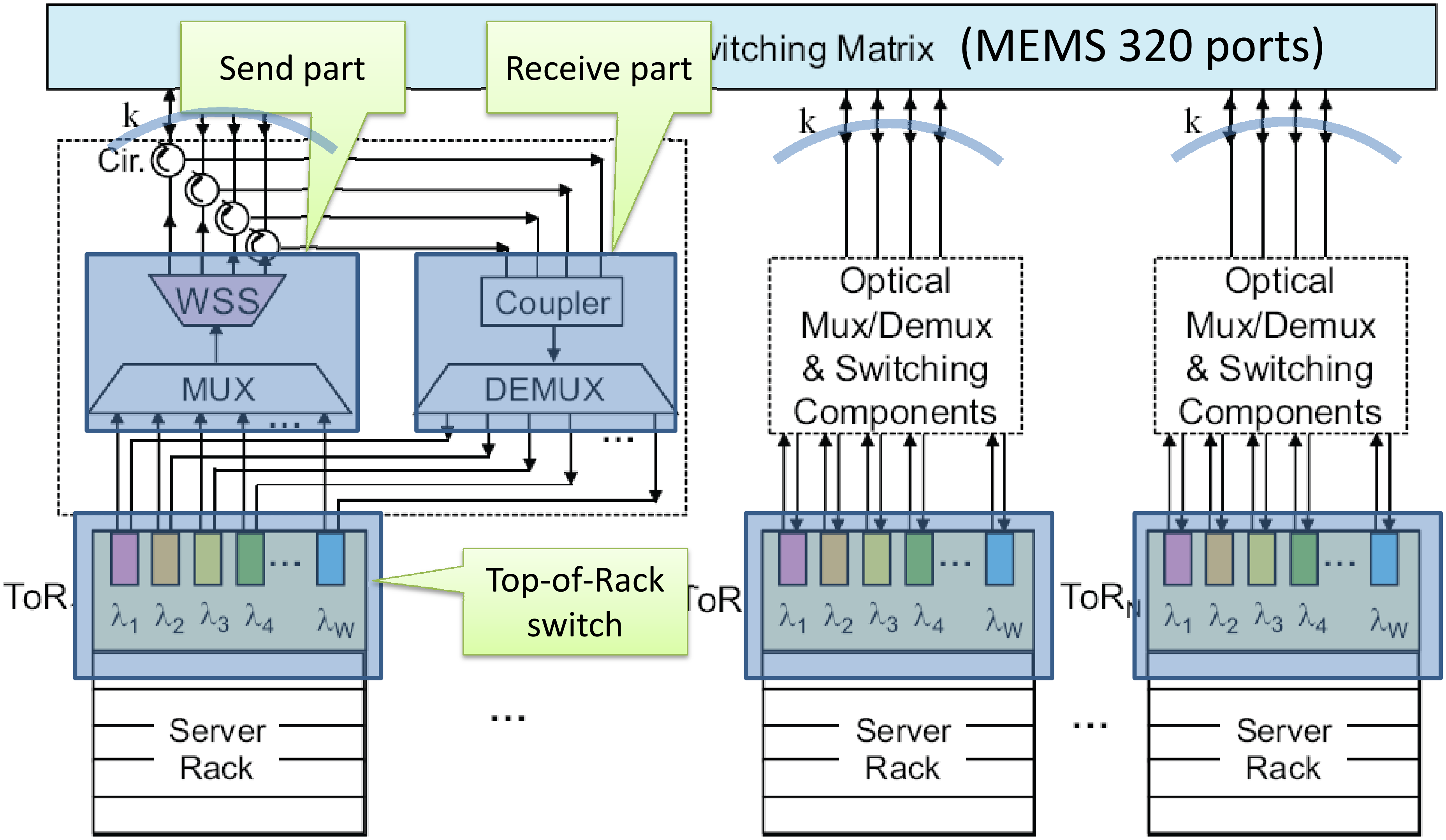
Optical fiber   WDM (DE)MUX   Coupler   Circulator



Common features:

- Support high bit-rate, high capacity
- Power-efficient
- Small and compact (except MEMS)

OSA Architecture Overview


itching Matrix
k
Cir.
WSS
MUX
Coupler
DEMUX
...
k
Optical Mux/Demux & Switching Components
k
Optical Mux/Demux & Switching Components
ToR
ToR
ToR$_N$
λ_1 λ_2 λ_3 λ_4 ... λ_W
λ_1 λ_2 λ_3 λ_4 ... λ_W
λ_1 λ_2 λ_3 λ_4 ... λ_W
Server Rack
...
Server Rack
...
Server Rack

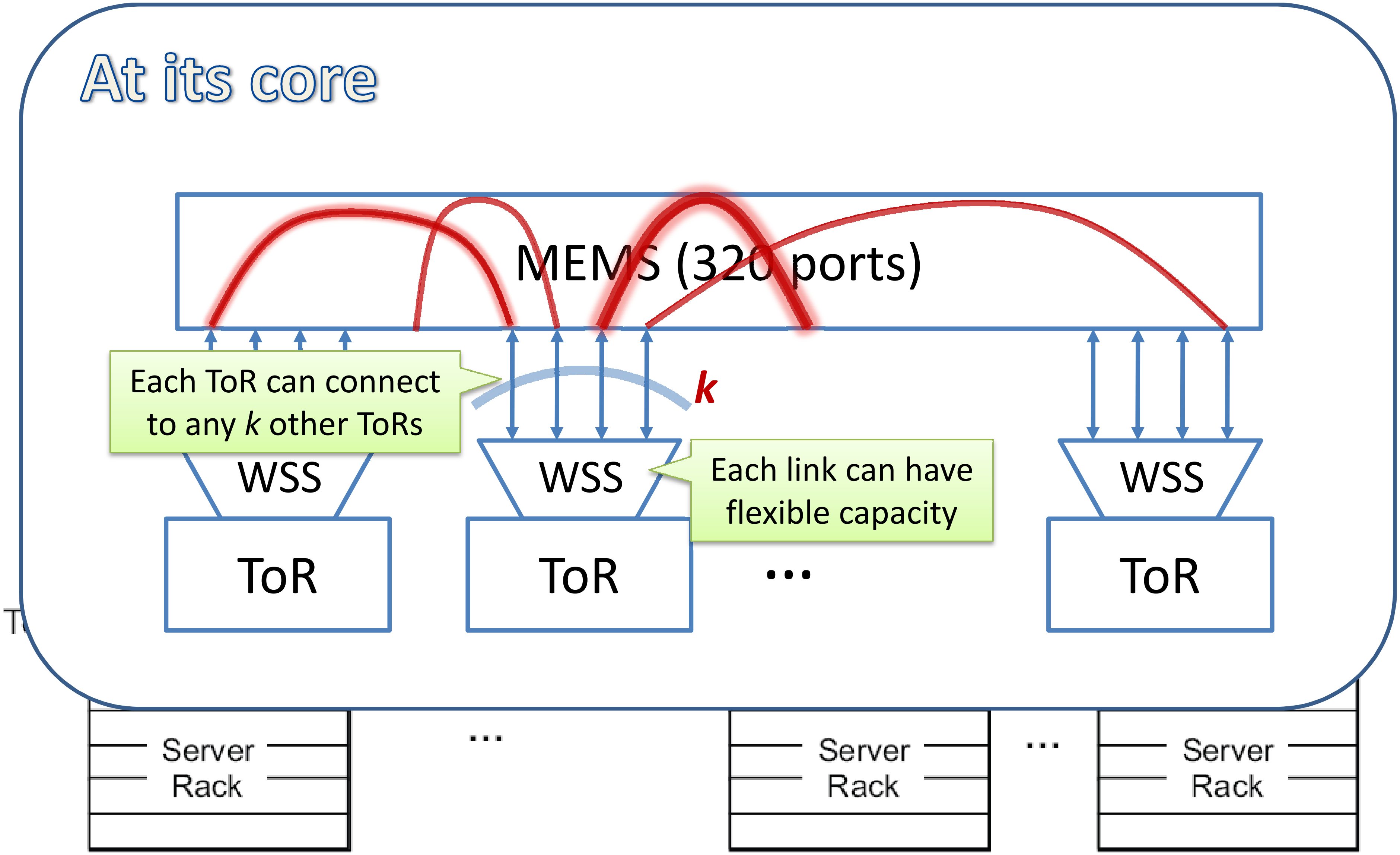
(MEMS 320 ports)

Send part

Receive part

Top-of-Rack switch

OSA Architecture Overview



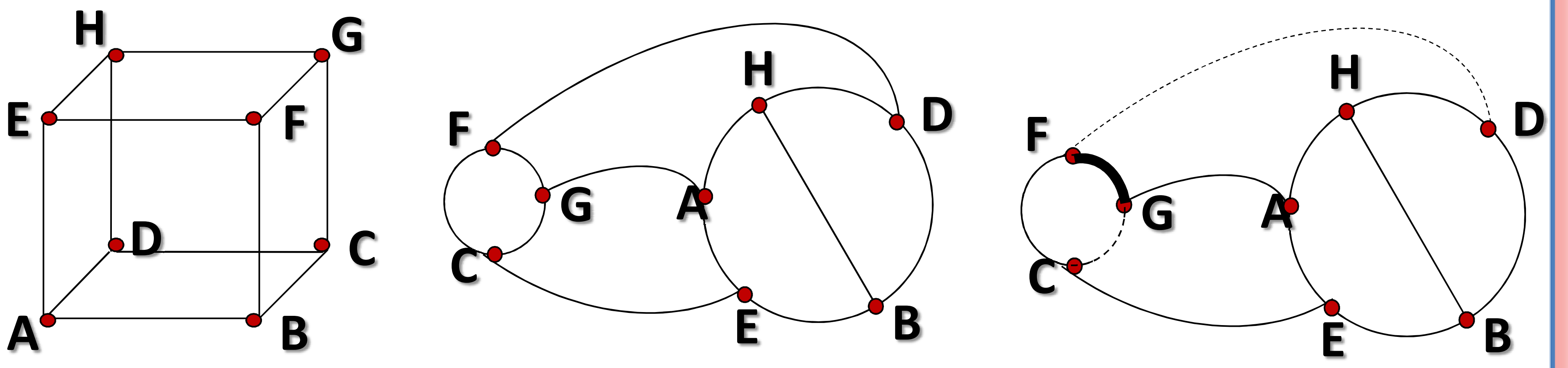
At its core

Each ToR can connect to any *k* other ToRs

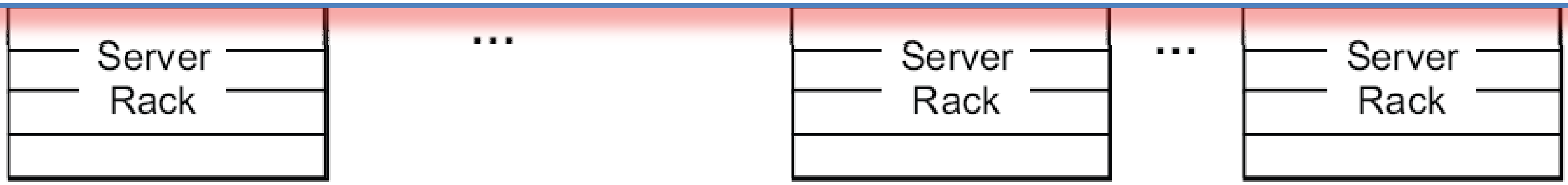
Each link can have flexible capacity

OSA Architecture Overview

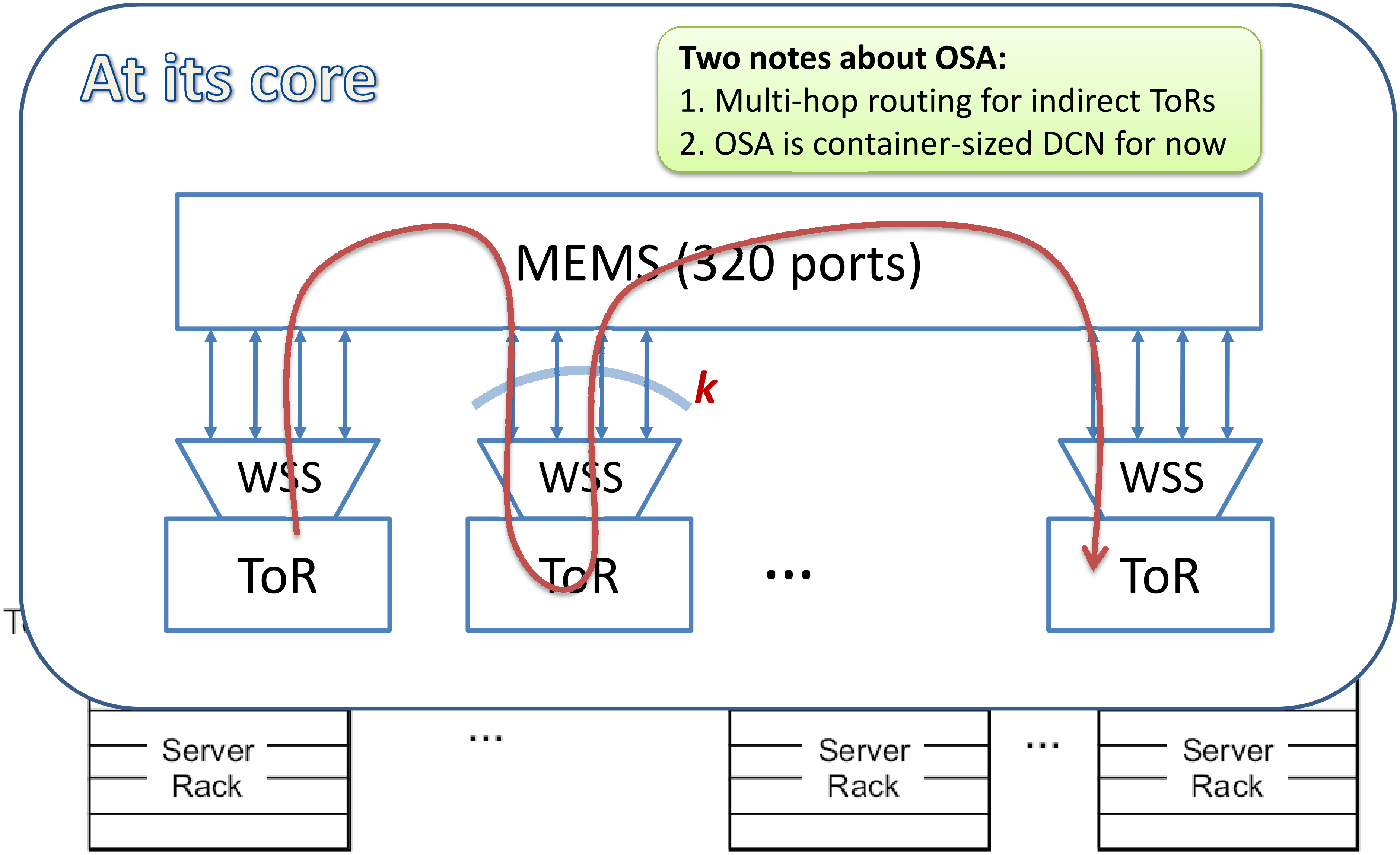
At its core



OSA can arrange *any k-regular topology* with *flexible link capacity* among the ToRs!



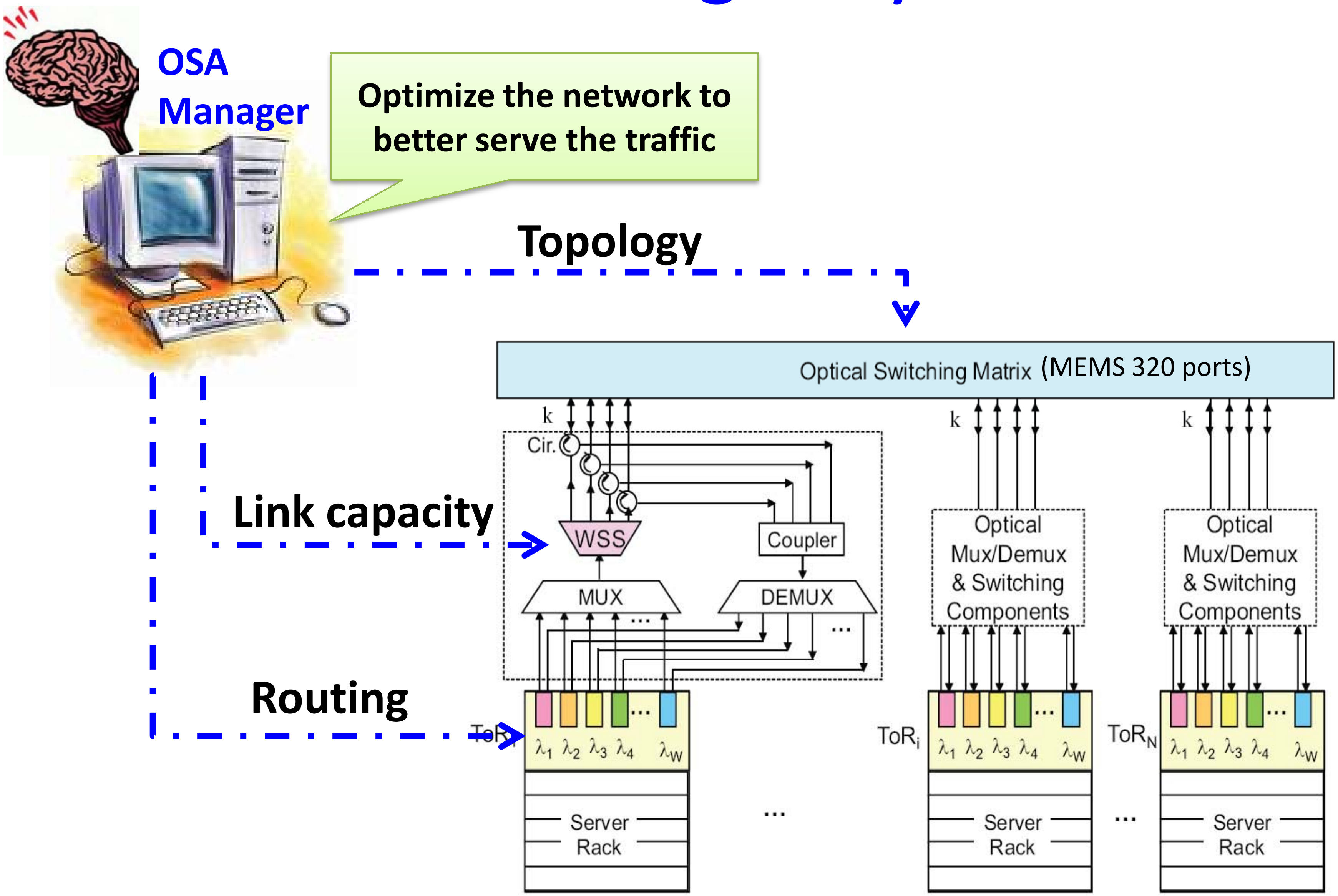
OSA Architecture Overview



Two notes about OSA:

1. Multi-hop routing for indirect ToRs
2. OSA is container-sized DCN for now

Control Plane: Logically Centralized



OSA Manager

Optimize the network to better serve the traffic

Topology

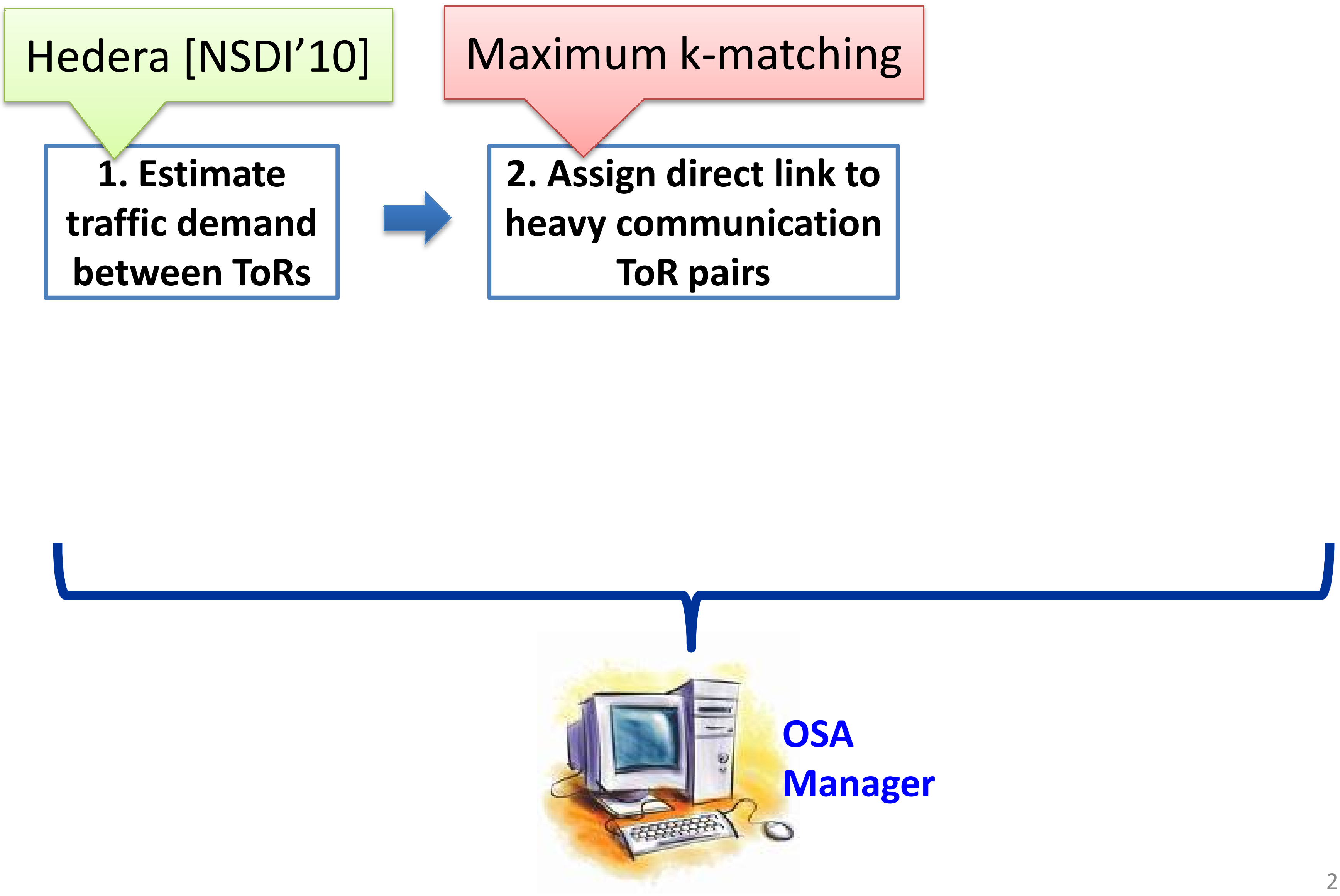
Link capacity

Routing

(MEMS 320 ports)

Optical Switching Matrix
k
Cir.
WSS
Coupler
MUX
DEMUX
...
Optical Mux/Demux & Switching Components
Optical Mux/Demux & Switching Components
ToR$_1$
ToR$_i$
ToR$_N$
λ_1 λ_2 λ_3 λ_4 λ_W
Server Rack
Server Rack
Server Rack

Optimization Procedure in OSA Manager



Hedera [NSDI’10]

1. Estimate traffic demand between ToRs

Maximum k-matching

2. Assign direct link to heavy communication ToR pairs

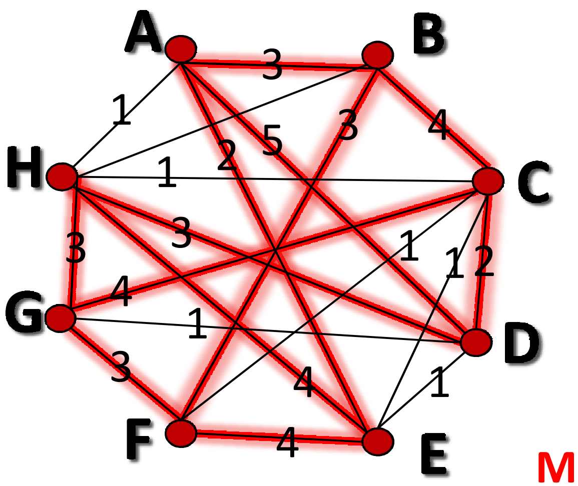
OSA Manager

Maximum *K*-matching for Direct Links Setup

ToR traffic demand

	A	B	C	D	E	F	G	H
A	--	3	0	5	2	0	0	1
B	3	--	4	0	0	3	0	1
C	0	4	--	2	1	1	4	1
D	5	0	2	--	1	0	1	3
E	2	0	1	1	--	4	0	4
F	0	3	1	0	4	--	3	0
G	0	0	4	1	0	3	--	3
H	1	1	1	3	4	0	3	--

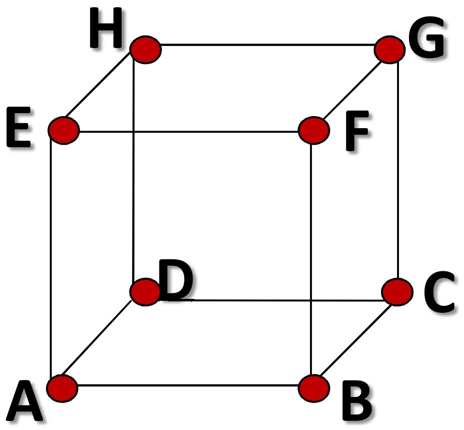
ToR demand graph



Maximum weighted 3-matching

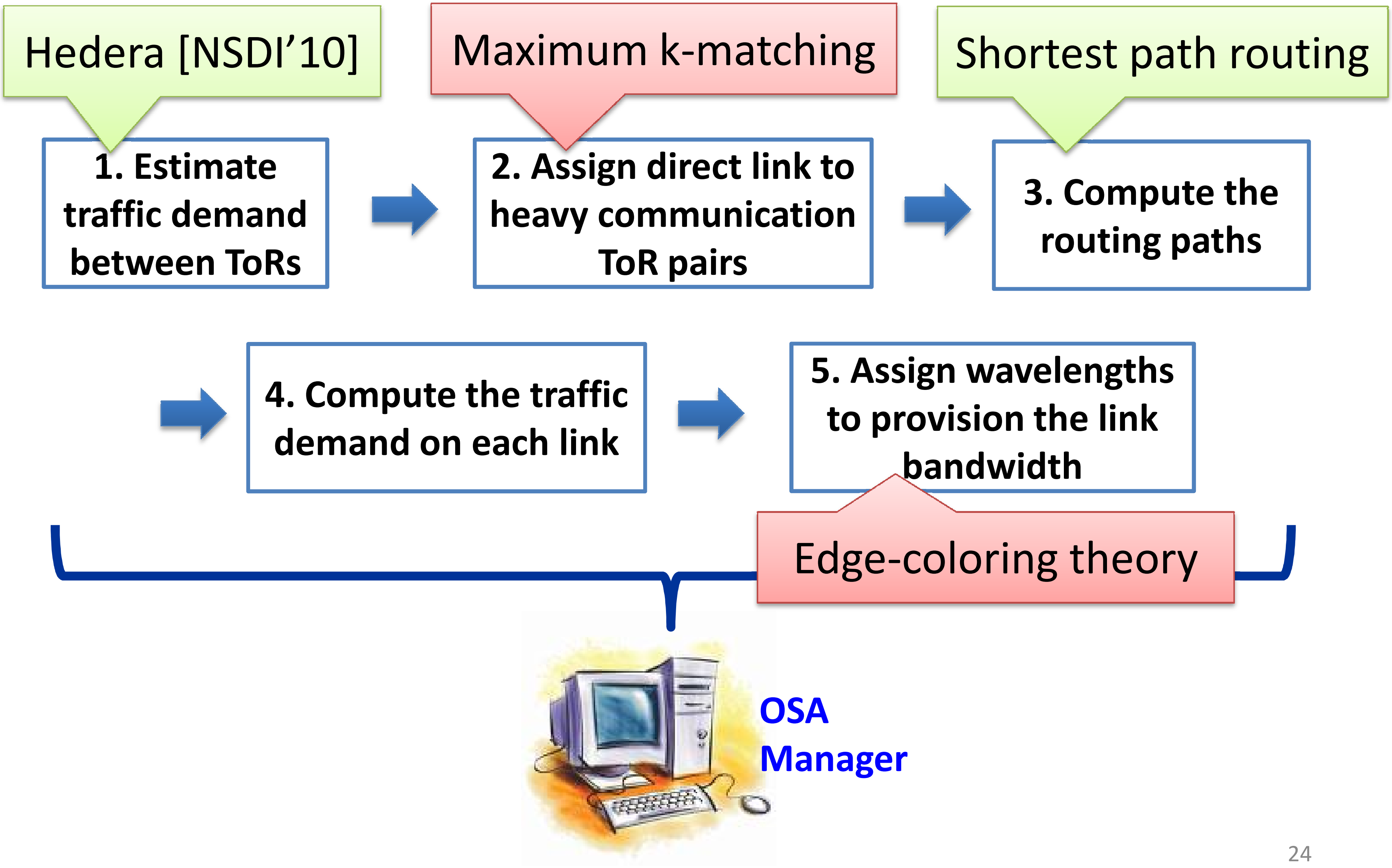
Edmonds' algorithm[1]

ToR connection graph



[1] J. Edmonds, “Paths, trees and flowers”, Canad. J. of Math., 1965

Optimization Procedure in OSA Manager



Hedera [NSDI'10]

Maximum k-matching

Shortest path routing

1. Estimate traffic demand between ToRs

2. Assign direct link to heavy communication ToR pairs

3. Compute the routing paths

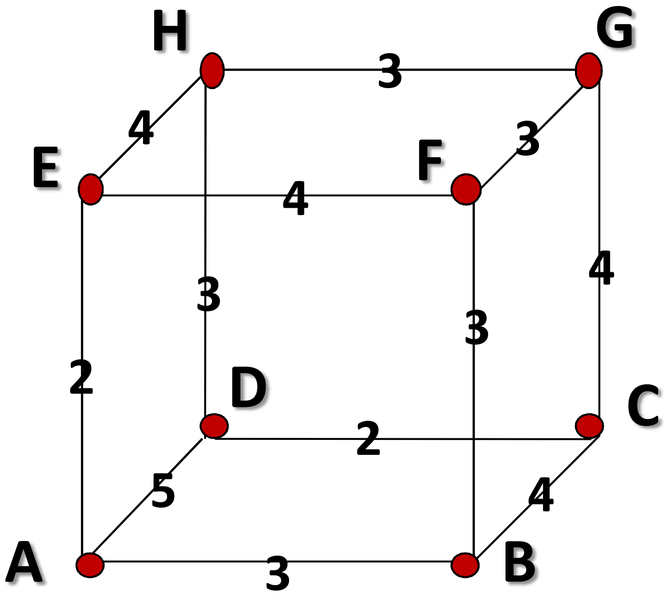
4. Compute the traffic demand on each link

5. Assign wavelengths to provision the link bandwidth

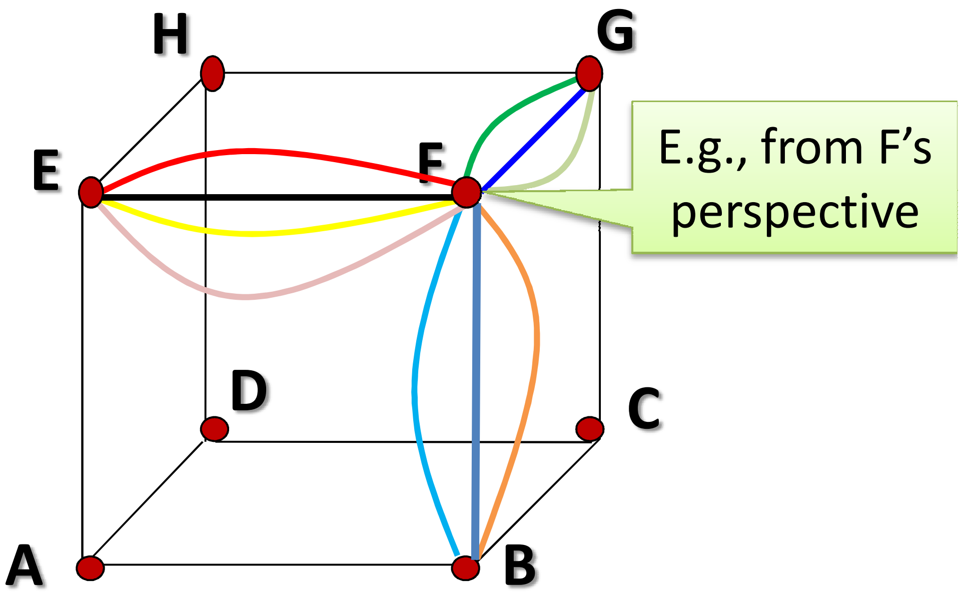
Edge-coloring theory

OSA Manager

Edge-coloring for Wavelength Assignment



Expected wavelength graph



E.g., from F’s perspective

Multigraph based on # of wavelengths

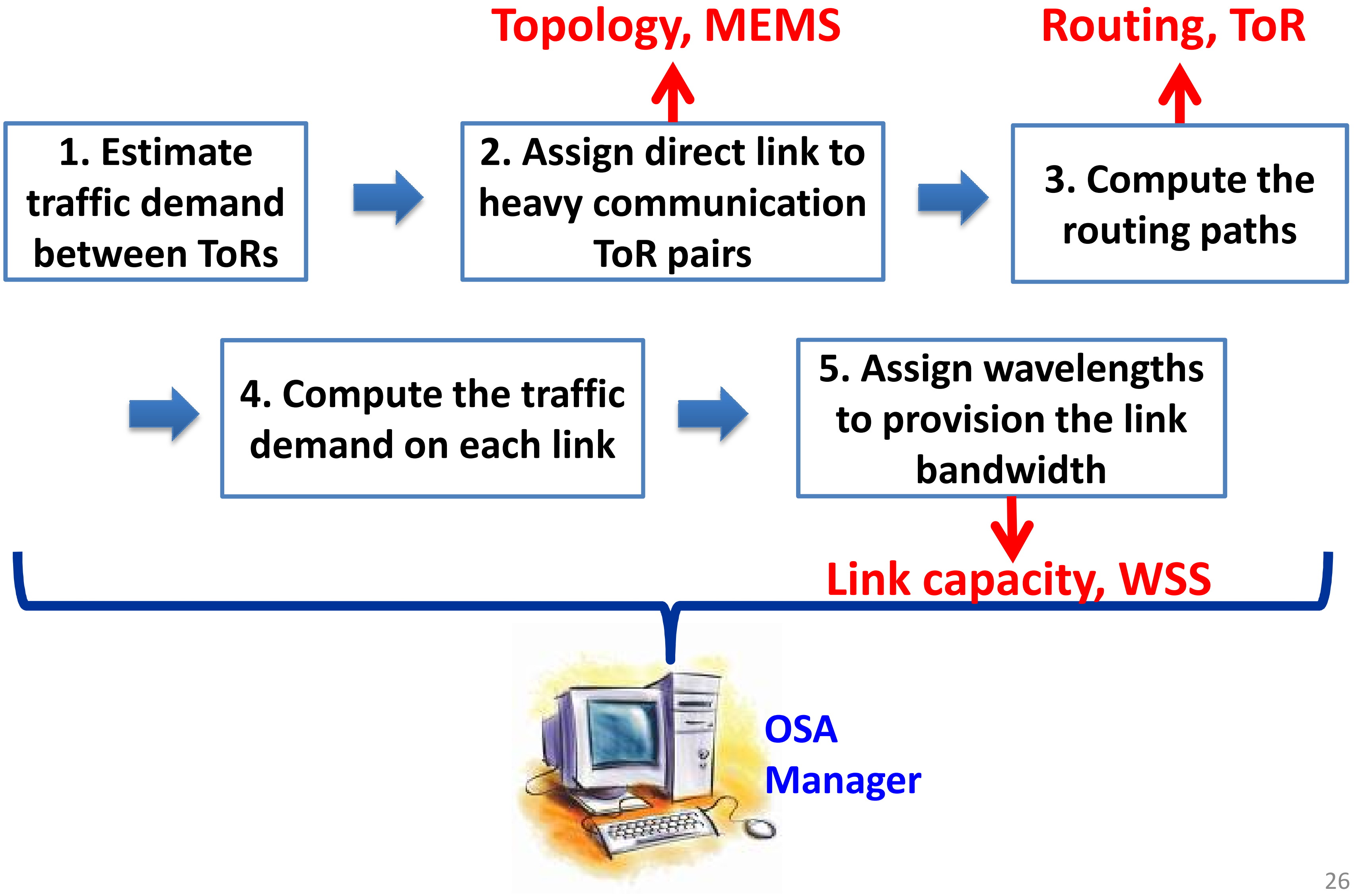
Wavelength assignment:
A wavelength cannot be associated with a ToR twice

Edge-coloring:
A color cannot be associated with a node twice

Vizing’s theorem[2]

[2] J. Misra, et. al., “A constructive proof of Vizing’s Theorem,” *Inf. Process. Lett., 1992.*

Optimization Procedure in OSA Manager



Topology, MEMS

Routing, ToR

1. Estimate traffic demand between ToRs

2. Assign direct link to heavy communication ToR pairs

3. Compute the routing paths

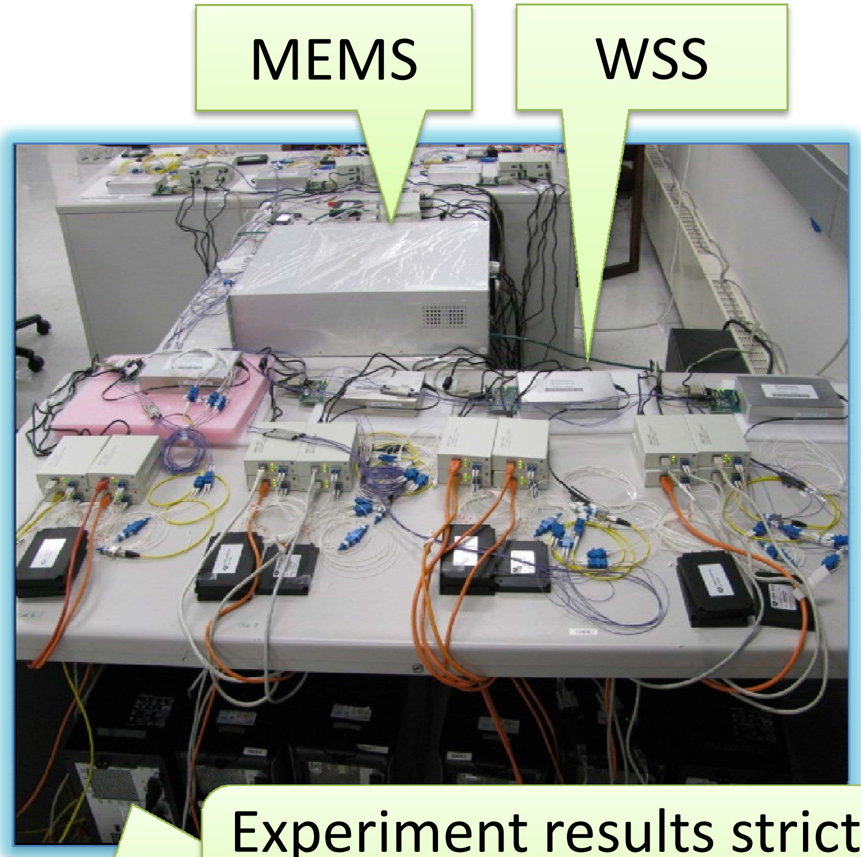
4. Compute the traffic demand on each link

5. Assign wavelengths to provision the link bandwidth

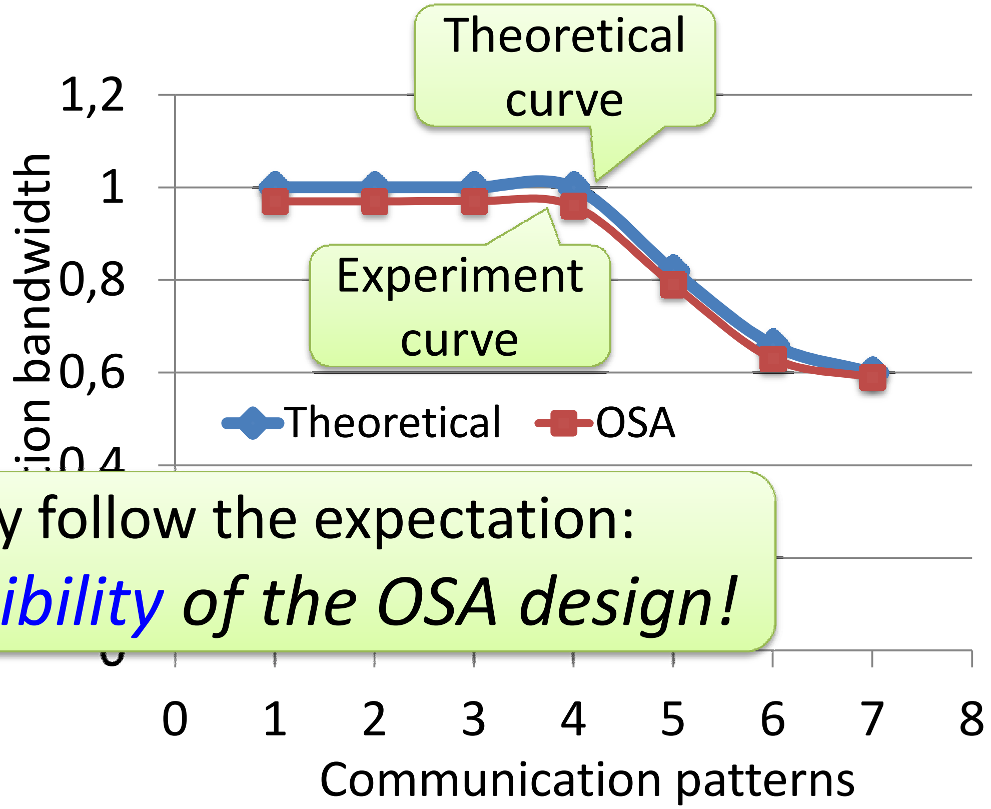
Link capacity, WSS

OSA Manager

Prototype Implementation



- 1 MEMS (32 ports: 16×16)
- 8 WSS units (1×4 ports)
- 8 ToRs* and 32 servers

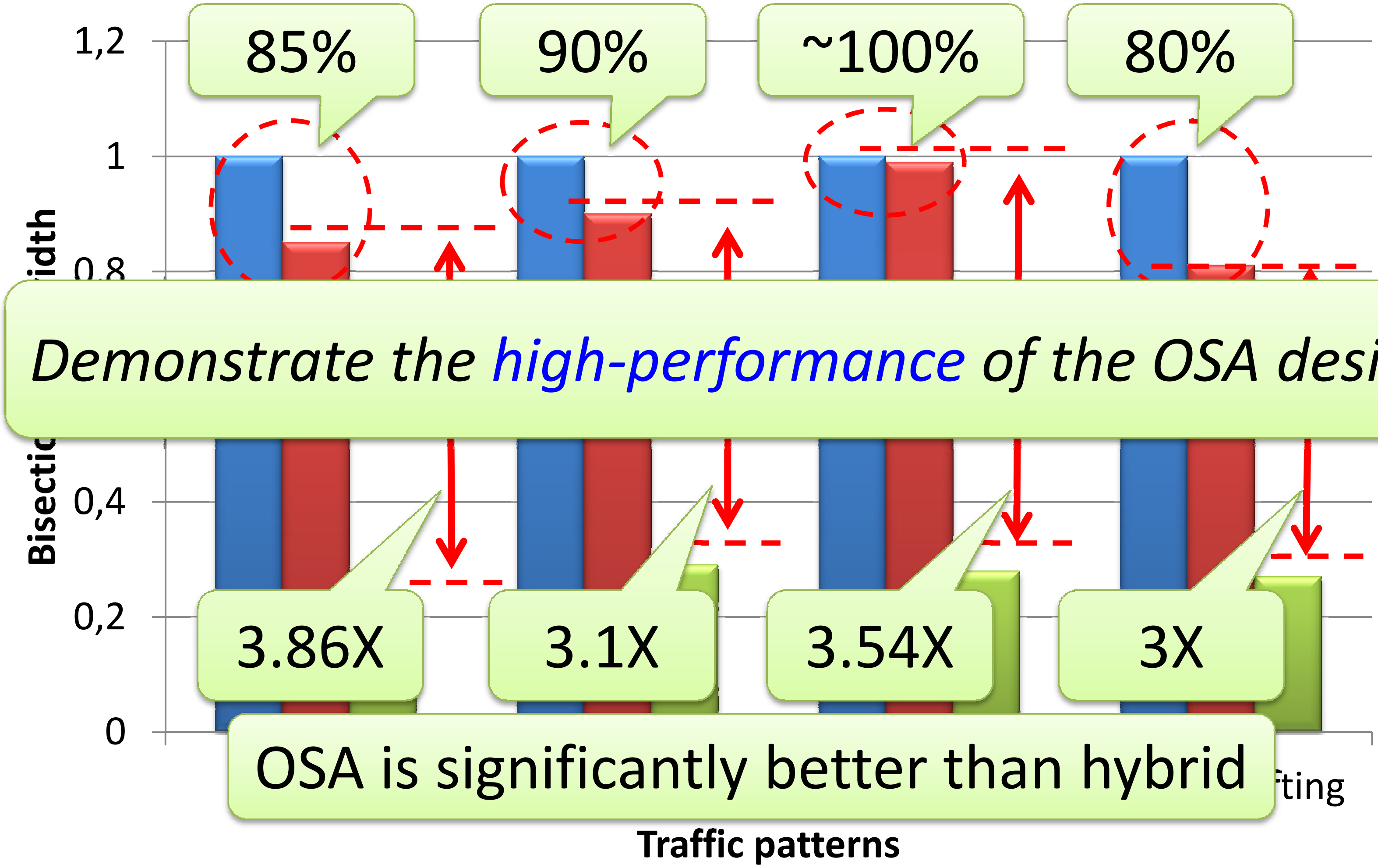


Experiment results strictly follow the expectation:
Demonstrate the feasibility of the OSA design!

*Serve

Simulation Results (2560 servers*)

OSA can be close to non-blocking

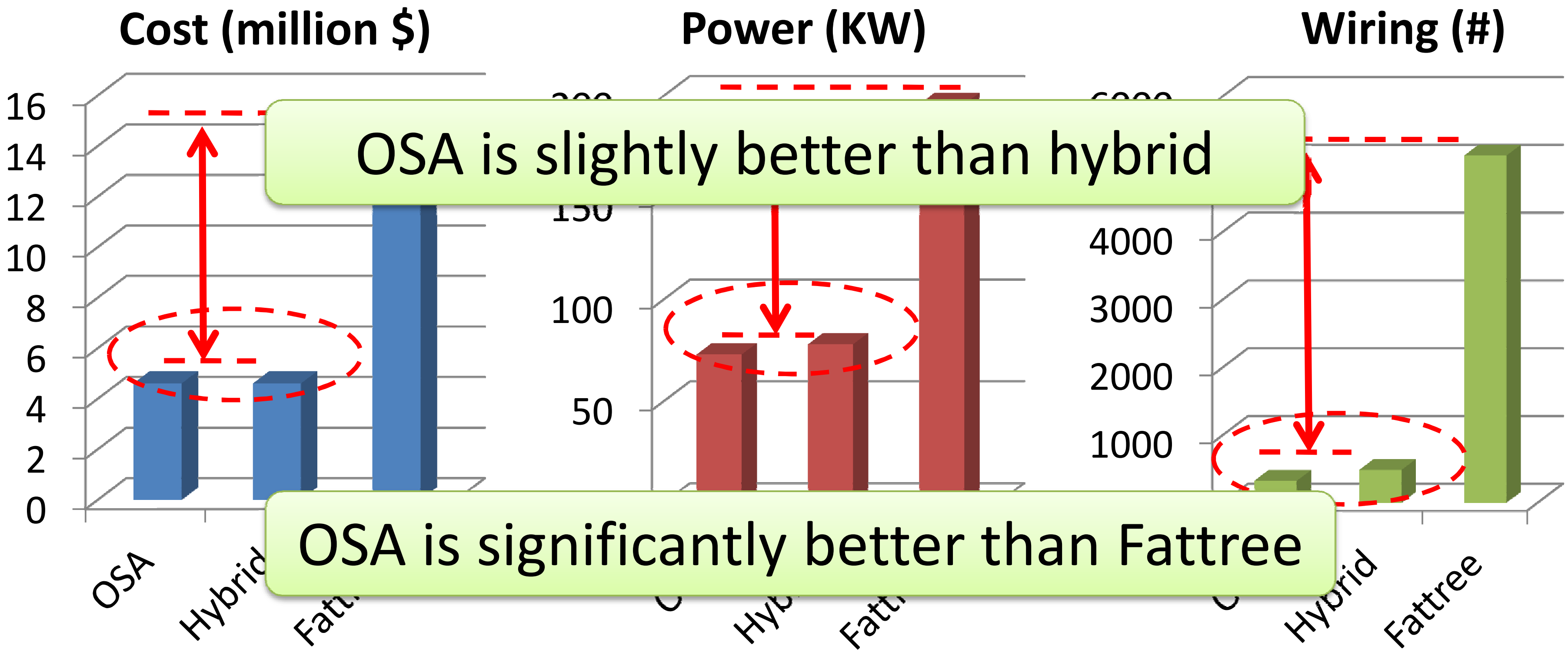


Demonstrate the ***high-performance*** *of the OSA design!*

OSA is significantly better than hybrid

*80 ToRs (each with 32 servers) form a 4-regular graph for OSA.

Cost, Power & Wiring (2560 Servers)

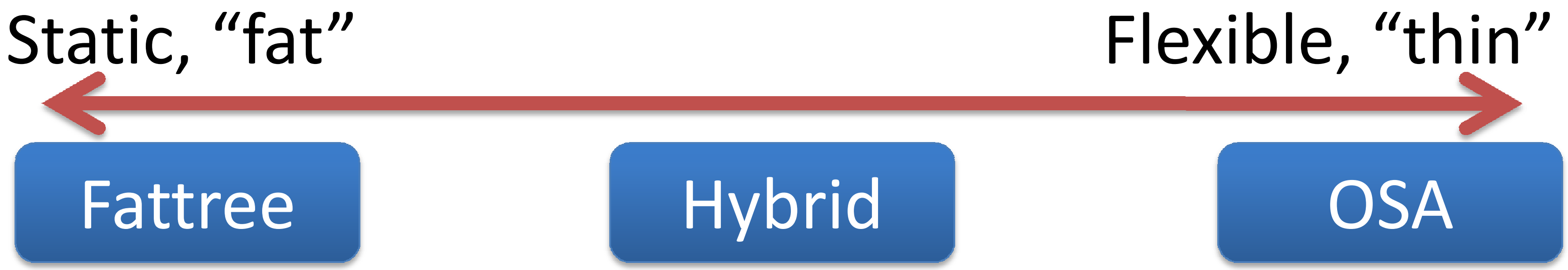


OSA is slightly better than hybrid

OSA is significantly better than Fattree

Demonstrate OSA can potentially deliver high bandwidth in a simple, power-efficient and cost-effective way!

Summary



	Performance	Complexity	Power	Cost
Fattree	√	X	X	X
Hybrid	X	√	√	√
OSA	√	√	√	√

- OSA is inspired by traffic regionality and stability
- Sweet spot for performance, cost, power, and wiring complexity
- Caveats: not intended for all-to-all, non-stable traffic
- Acknowledgement: CoAdna Photonics (WSS) and Polatis (MEMS)

Thanks!

Data Center Traffic Characteristics

[IMC’09][HotNets’09]: *only a few ToRs are hot and most of their traffic goes to a few other ToRs*

[IMC’10]: *traffic at ToRs exhibits an ON/OFF pattern*

[SIGCOMM’09]: *over 90% bytes flow in elephant flows*

[WREN’10]: *60% ToRs see less than 20% change in traffic volume for between 1.6-2.2 seconds*

[ICDCS’12]: *a production DCN traffic shows stability even on a hourly time scale*

Static full bisection bandwidth between all servers at all the time is a waste of resource!