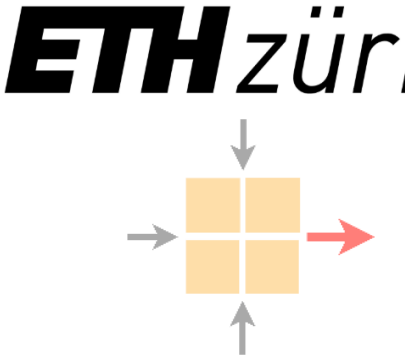
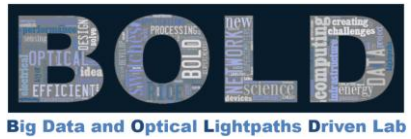




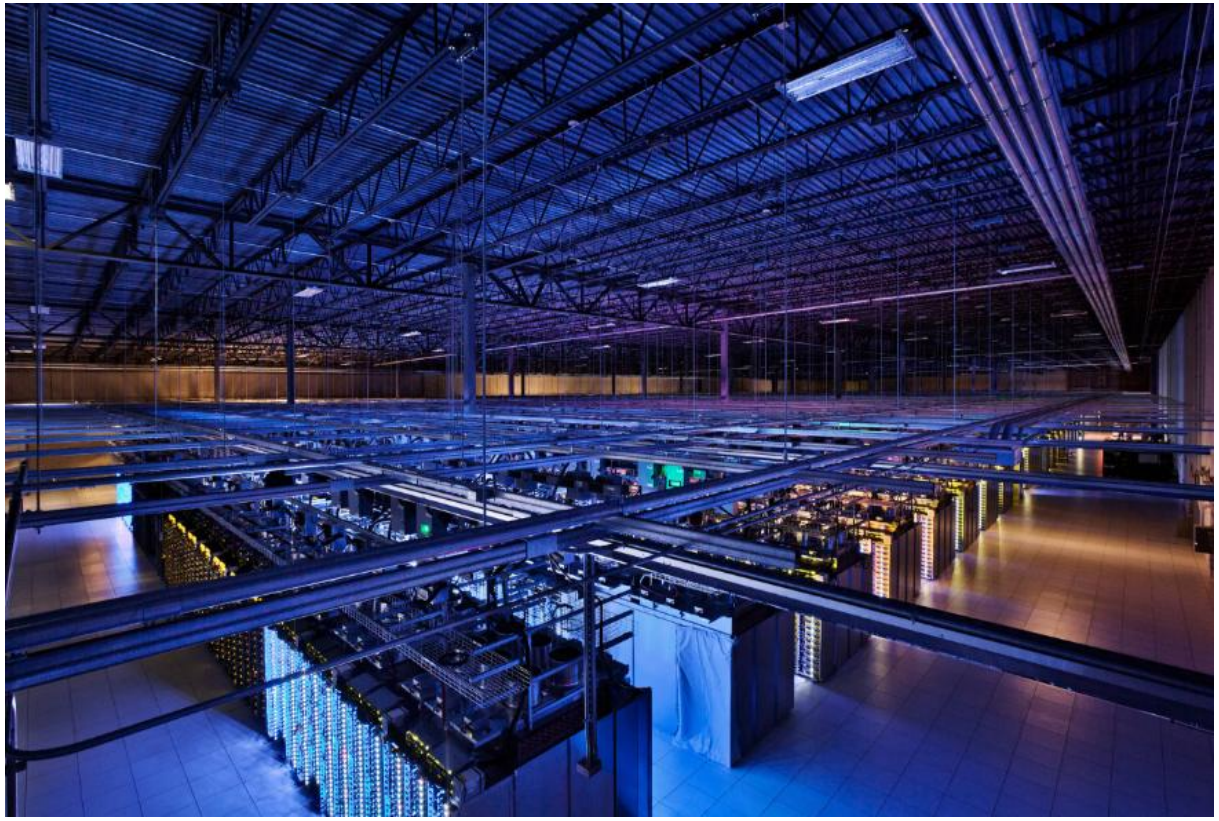
Rearchitecting Datacenter Networks: A New Paradigm with Optical Core and Optical Edge

Dr. Sushovan Das
Postdoctoral Researcher
Networked Systems Group, ETH Zurich

CNI Seminar Series, IISc Bangalore
Jan 5, 2026

Overview of Datacenters

- Infrastructure behind the “Cloud”, large-scale data warehouse
- Massive number of servers (compute and storage nodes)
- Connected via a network of switches: Datacenter Network

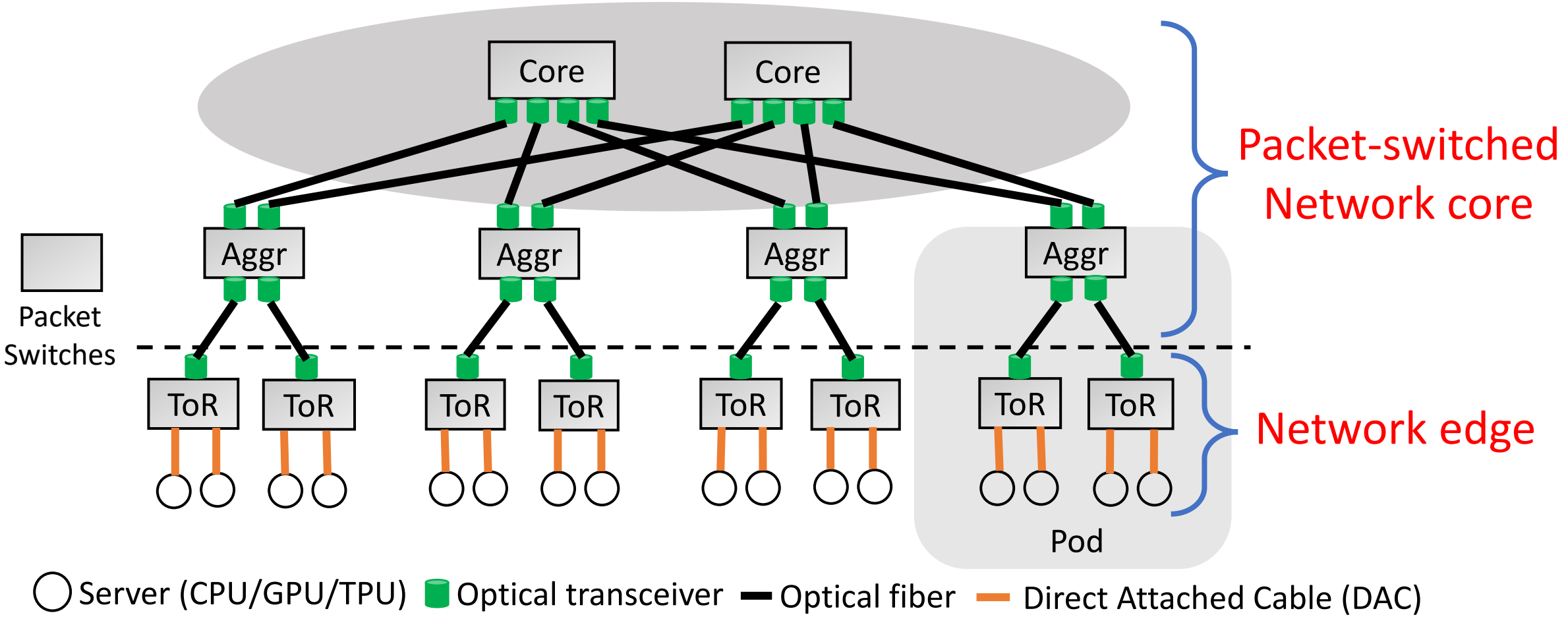


Google’s Datacenter Iowa, US [1]

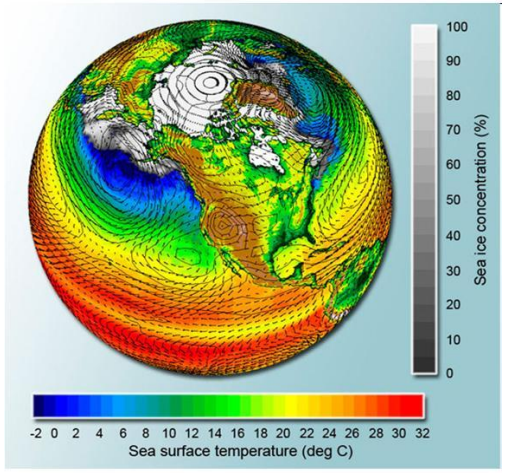
[1] https://datacenters.google/discover-more/photo-gallery/

Datacenter Network (DCN) Architectures

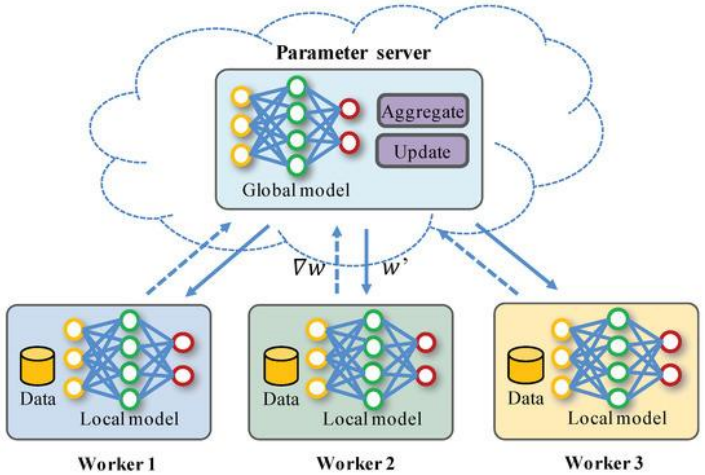
- Hierarchical network topology with Ethernet packet switches
- Top-of-rack switch (ToR), Aggregate switch (Aggr), Core switch (Core)



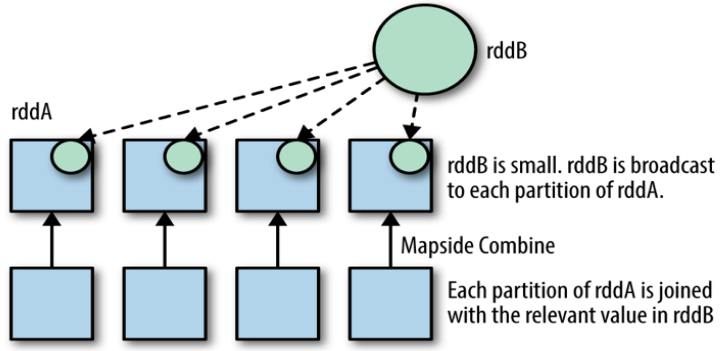
DCN Applications are Diverse



High Performance Computing [1]



Distributed ML/ DNN training [2]



Distributed Database Queries [3]

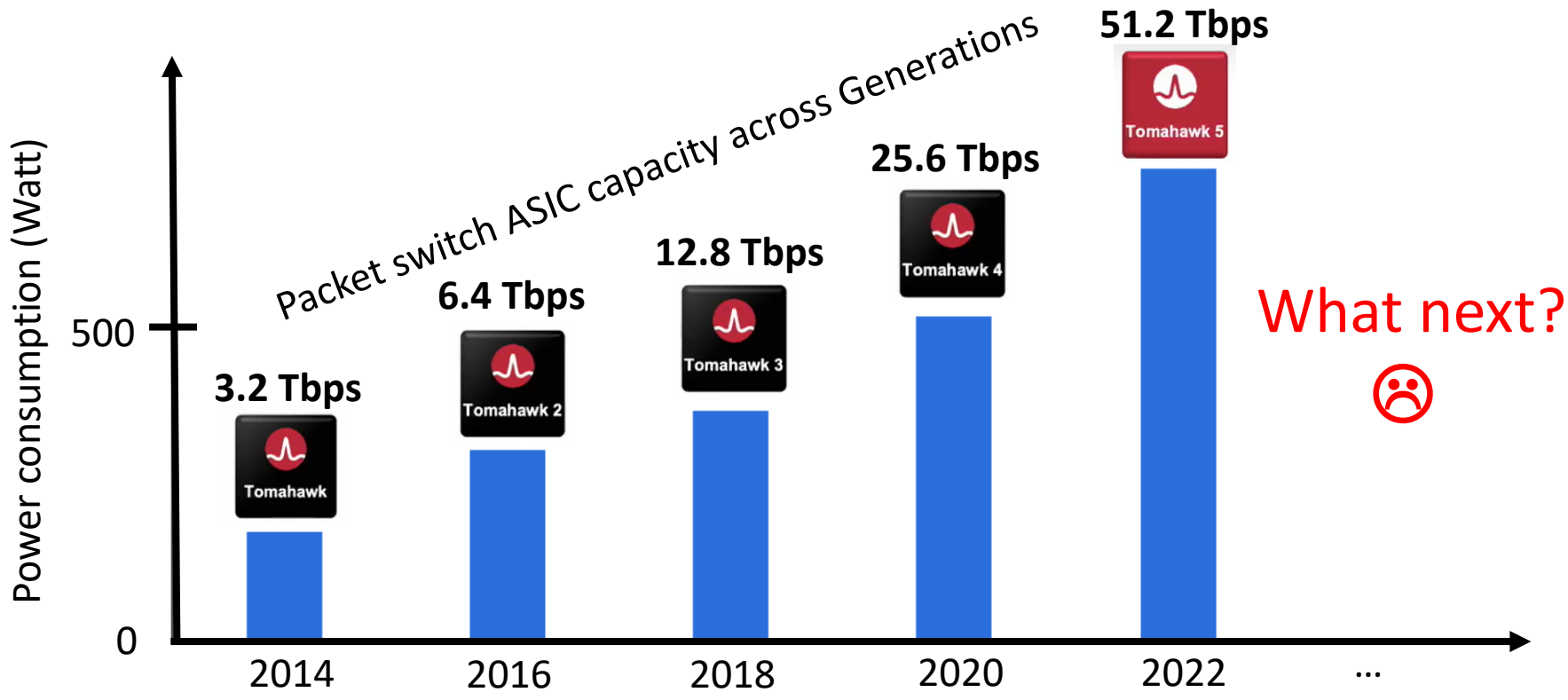
- Stringent performance requirements: High throughput + Low latency
- Advent of domain-specific accelerators, non-volatile memory
- Shifting major bottleneck from computation to network IO
- Super fast scale-up of network capacity is necessary

[1] https://sites.uci.edu/zlabe/arctic-sea-ice-visualizations/
[2] https://www.mdpi.com/2076-3417/12/1/292
[3] https://learning.oreilly.com/library/view/spark-the-definitive/9781491912201/

Challenges of Post Moore's Law Era

Electrical Packet Switches (EPS) are not sustainable

- Increasing gap: Electrical switch capacity vs power/cost/port density
- Broadcom Tomahawk-5 ASIC requires 45% more power w.r.t previous Gen [1]

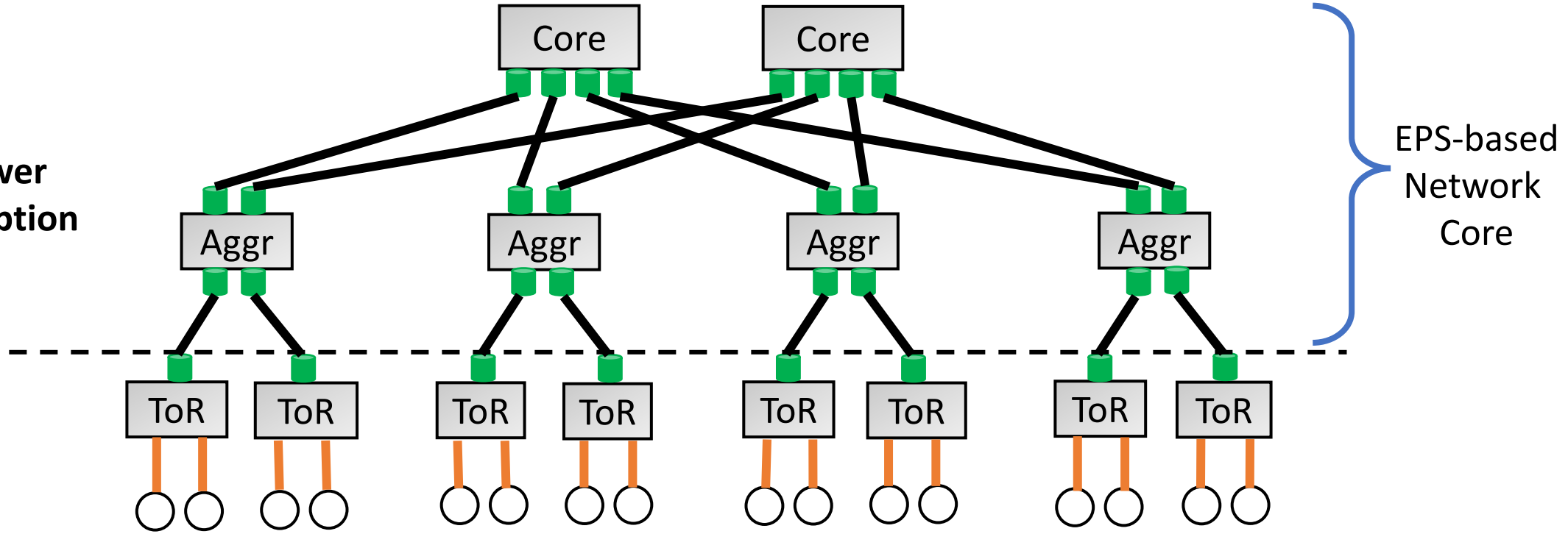


[1] Qian, Kun, et al. "Alibaba HPN: A Data Center Network for Large Language Model Training." ACM SIGCOMM, 2024.

Packet-switched DCN Consumes Excessive Power

- DCN architecture with 65k servers + 64x400G switches
- The network power consumption can shoot up to 48.8 MW [1]
- 52.5% more than the energy budget of the DCN operator
- Critical for AI clusters as network utilization is very high

High Power Consumption

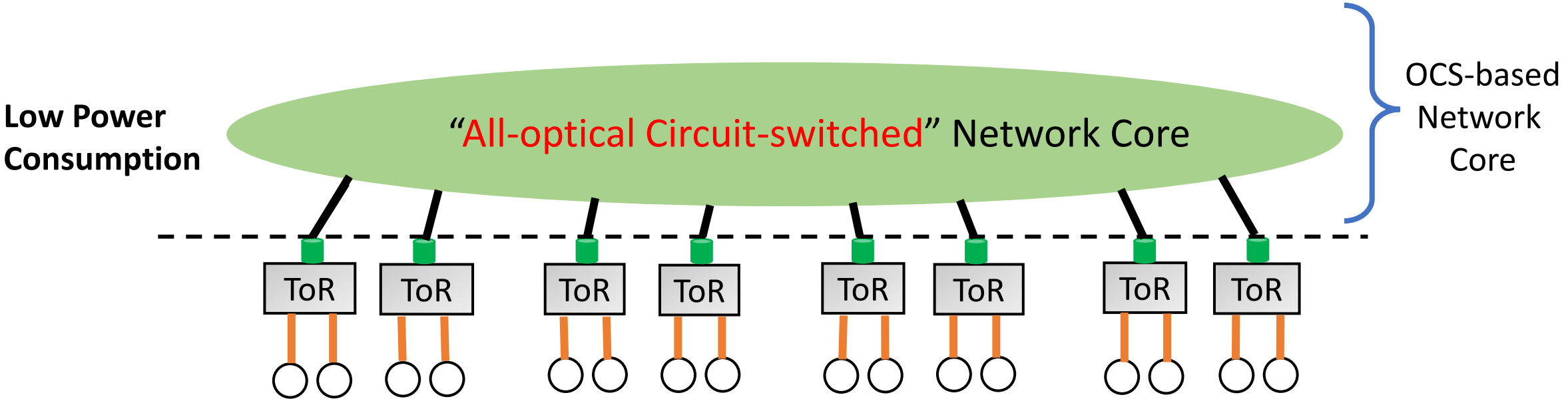


[1] Ballani, Hitesh, et al. "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching." ACM SIGCOMM, 2020.

Packet-switched DCN Consumes Excessive Power

- DCN architecture with 65k servers + 64x400G switches
- The network power consumption can shoot up to 48.8 MW [1]
- 52.5% more than the energy budget of the DCN operator
- Critical for AI clusters as network utilization is very high

Low-Power Optical Circuit Switch (OCS): Promising Alternative

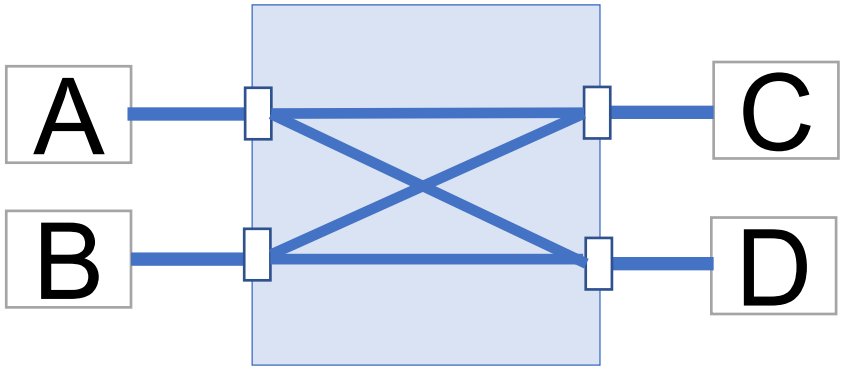


[1] Ballani, Hitesh, et al. "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching." ACM SIGCOMM, 2020.

Basic Building Block: Optical Circuit Switch (OCS)

Features

- Physically steer light, programmable
- Circuits can be configured at runtime
- Reconfiguration downtime (δ)



OCS Technology	3D MEMS	2D MEMS, Rotor	AWGR
δ	$\approx$ 10 msec	$\approx$ 10 μsec	$\approx$ 100 nanosec

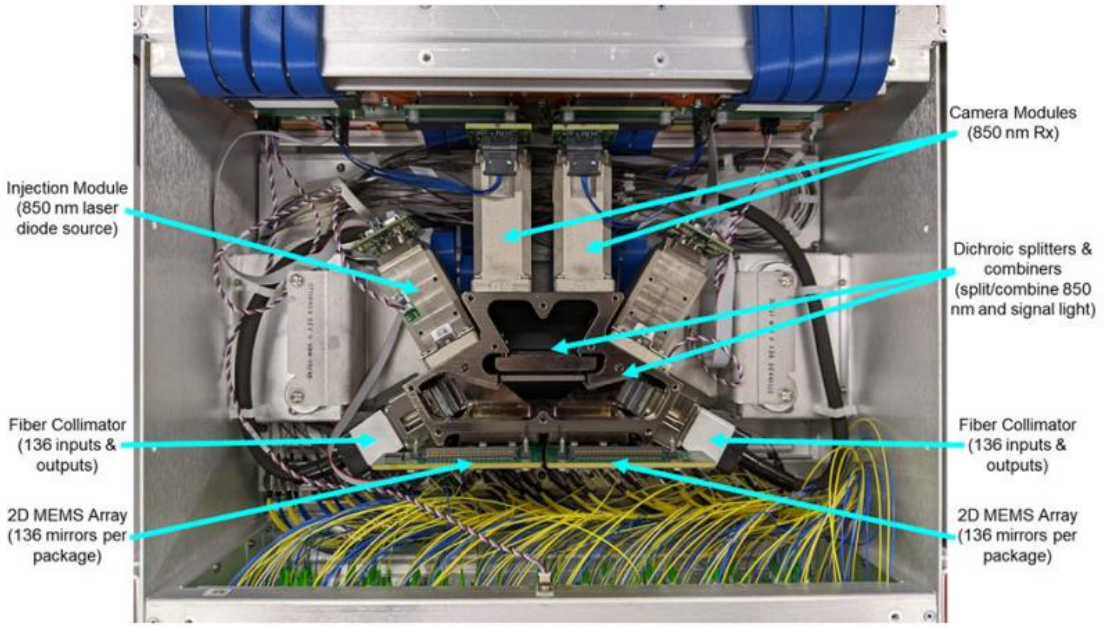
Advantages

- Agnostic to data rate
- No packet processing, negligible forwarding latency
- Negligible/zero power consumption

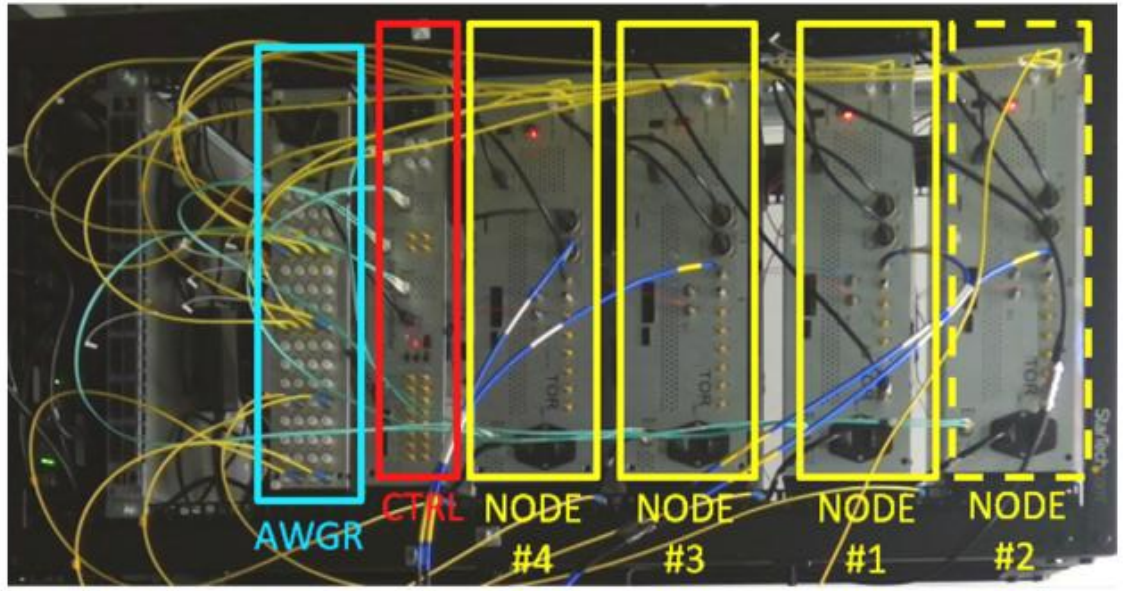
Change of philosophy: cannot reconfigure circuits per-packet

Optical Circuit Switching in DCN is Real

- Industry is adopting OCS in production-scale DCN clusters
- Google's Lightwave [1], Microsoft's Sirius [2] fabrics
- Saving power and CapEx (as OCS can serve multiple generations)



Google's Lightwave [1]
MEMS based (136 x 136)

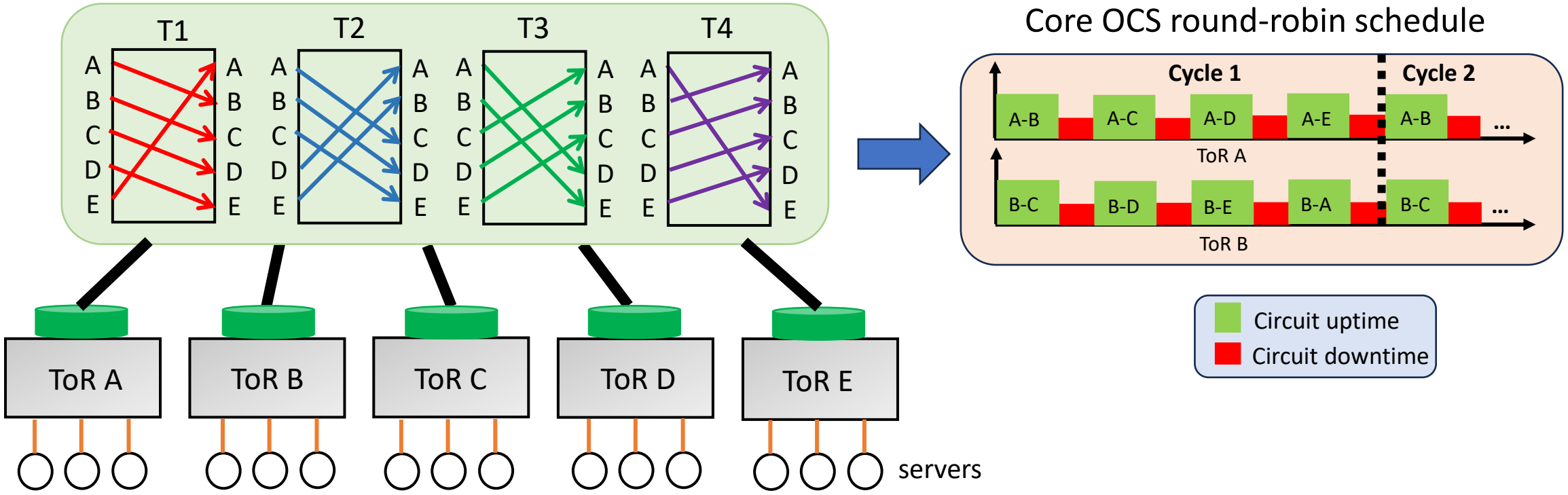


Microsoft's Sirius [2]
AWGR based

[1] Liu, Hong, et al. "Lightwave Fabrics: At-Scale Optical Circuit Switching for Datacenter and Machine Learning Systems", ACM SIGCOMM, 2023.
[2] Ballani, Hitesh, et al. "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching." ACM SIGCOMM, 2020.

Existing Design Paradigm: Traffic Agnostic OCS cores

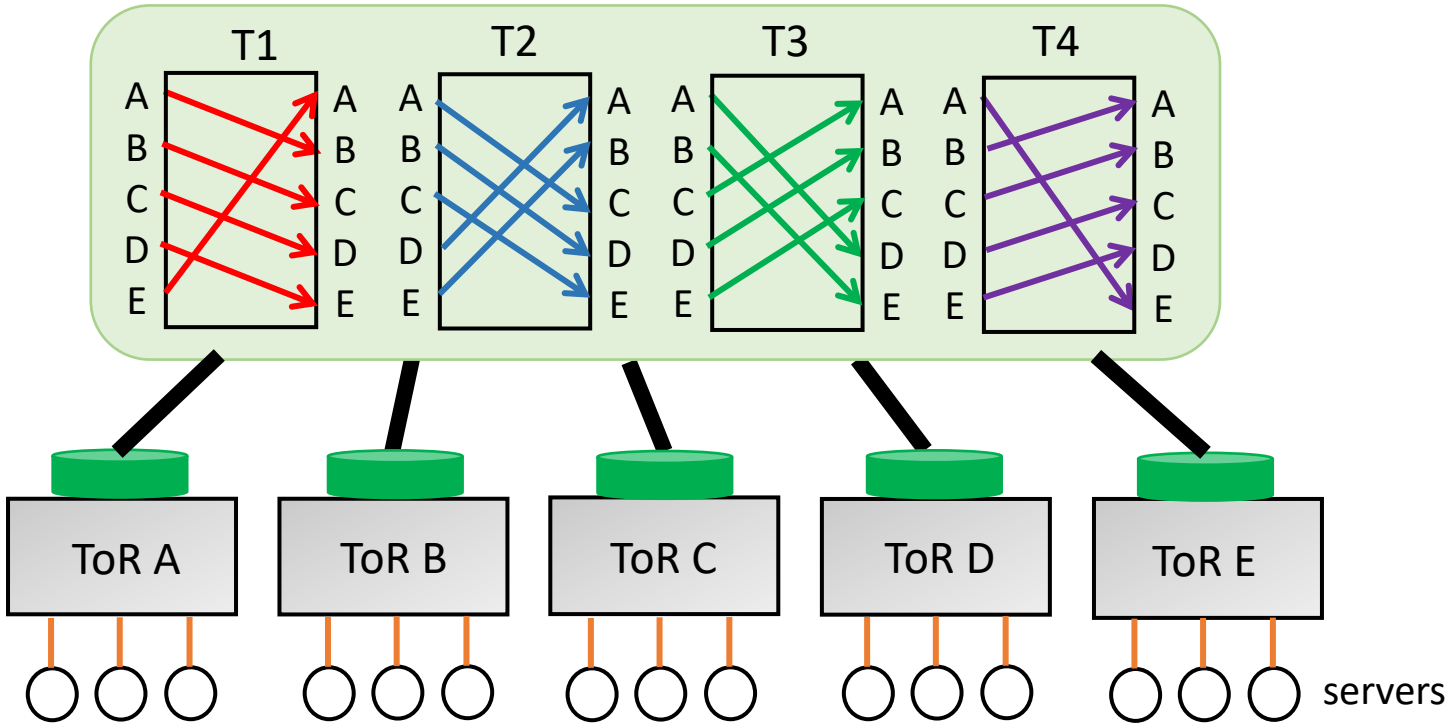
- OCS cycles through that predefined set of circuit configurations [1,2]
- Illusion of any-to-any connectivity among ToRs over time
- Round-Robin Circuit Scheduling: Traffic Agnostic, Open loop control



[1] Mellette, William, et al. "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network." ACM SIGCOMM, 2017.
[2] Ballani, Hitesh, et al. "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching." ACM SIGCOMM, 2020.

Existing Design Paradigm: Traffic Agnostic OCS cores

- OCS cycles through that predefined set of circuit configurations [1,2]
- Illusion of any-to-any connectivity among ToRs over time
- Round-Robin Circuit Scheduling: Traffic Agnostic, Open loop control



Core OCS round-robin schedule

	A	B	C	D	E
T1	B	C	D	E	A
T2	C	D	E	A	B
T3	D	E	A	B	C
T4	E	A	B	C	D

[1] Mellette, William, et al. "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network." ACM SIGCOMM, 2017.
[2] Ballani, Hitesh, et al. "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching." ACM SIGCOMM, 2020.

Challenges of All-Optical Circuit-Switched Cores

Challenge 1: Lacks native multicast capability

- No point to multipoint circuit: Fundamental to the OCS hardware
- How to enable low-energy high-performance multicast? [TON'22]

Challenge 2: Cannot efficiently handle traffic skewness

- Lack of path diversity: Fundamental to the round-robin abstraction
- How to compensate for the lack of path diversity? [INFOCOM'24, NSDI'22]

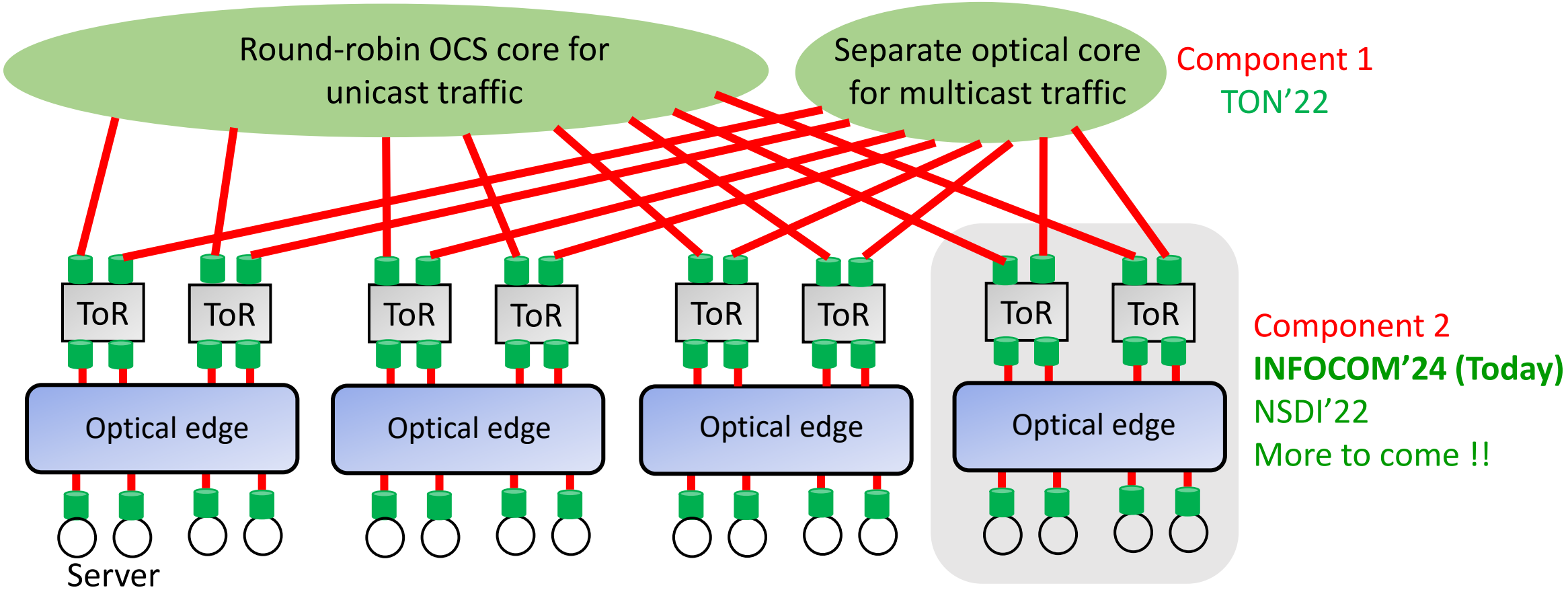
Challenge 3: Terrible tail latency performance

- Physical circuit downtime: Fundamental to the OCS hardware + architecture
- How to minimize the impact of OCS downtime under diverse workloads? [In progress]

How to solve these challenges while preserving most of the energy-saving benefit of OCS cores?

Theme of my Thesis: Holistic DCN Architecture

- Two components: Combination of Enhanced optical core and Optical edge
- Separate optical core enables low-energy and high-performance multicast
- Flexible optical edge handles traffic skewness and improves tail performance



Component 1
TON'22

Component 2
INFOCOM'24 (Today)
NSDI'22
More to come !!

Anatomy of Round-Robin Circuit Scheduling

- Uniform distribution of bandwidth across the node pairs
- For uniform traffic, fair utilization of all circuits

	A	B	C	D	E
T1	B	C	D	E	A
T2	C	D	E	A	B
T3	D	E	A	B	C
T4	E	A	B	C	D

Realistic DCN workloads are not ideal

High Skewness

- Most of the traffic is confined within hot rack pairs
- Microsoft DCN trace: 80% of traffic between 0.03-0.4% of rack-pairs [1]
- Disaggregated workload: 84% of flows between 33% nodes [2]

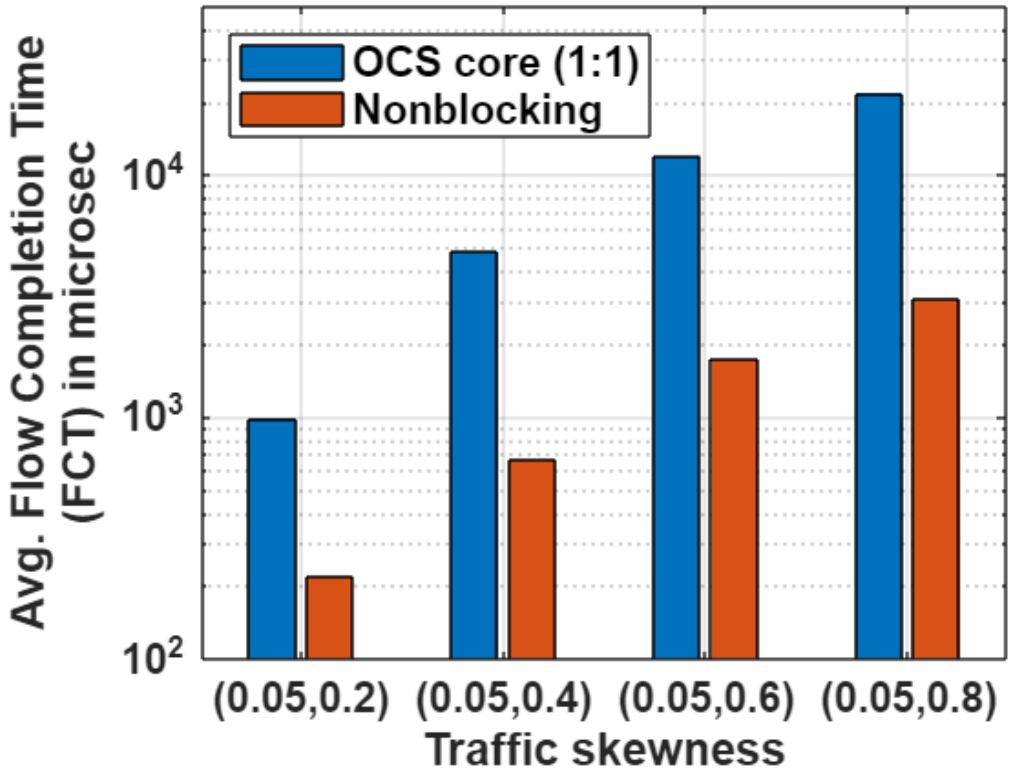
High Inter-rack Traffic Volume

- Most of the traffic crosses the rack boundary
- Facebook Frontend trace: 96.26% Inter-rack traffic [3]
- Facebook Database trace: 92.89% Inter-rack traffic [3]

[1] Ghobadi, Monia, et al. "ProjecTOR: Agile reconfigurable data center interconnect." ACM SIGCOMM, 2016.
[2] Shrivastav, Vishal, et al. "Shoal: A network architecture for disaggregated racks." USENIX NSDI, 2019.
[3] Wang, Weitao, et al. "RDC: Energy-Efficient Data Center Network Congestion Relief with Topological Reconfigurability at the Edge." USENIX NSDI, 2022.

Realistic Workloads: Poor Performance

- Realistic workloads (Cache): a) heavily skewed, b) high inter-rack volume
- Round-robin OCS-based core (Sirius) vs Non-blocking network
- Skewness (x, y): x fraction of hot-rack pairs exchange y fraction of traffic

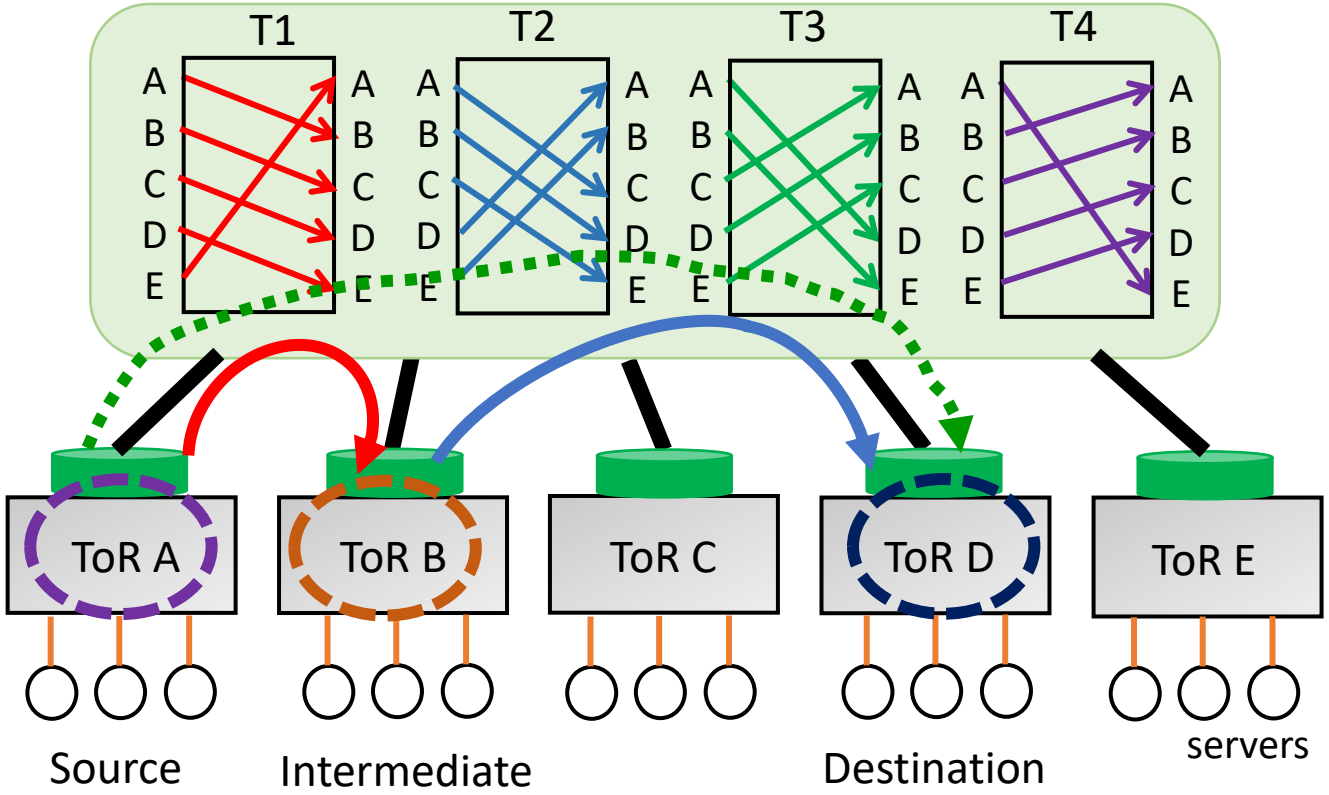


Impact of skewness and high inter-rack volume

- Average FCT Slowdown up to 7.2x
- Hot rack circuit pairs heavily utilized
- Cold rack circuit pairs remain underutilized
- Cannot leverage the full core bandwidth

State-of-the-art Technique to Improve Performance

- Valiant Load Balancing (VLB): Each ToR sends packets via an intermediate ToR
- Phase 1: ToR A sends traffic to an intermediate ToR X now (current slot)
- Phase 2: ToR X forwards traffic to destination ToR B later (future slot)



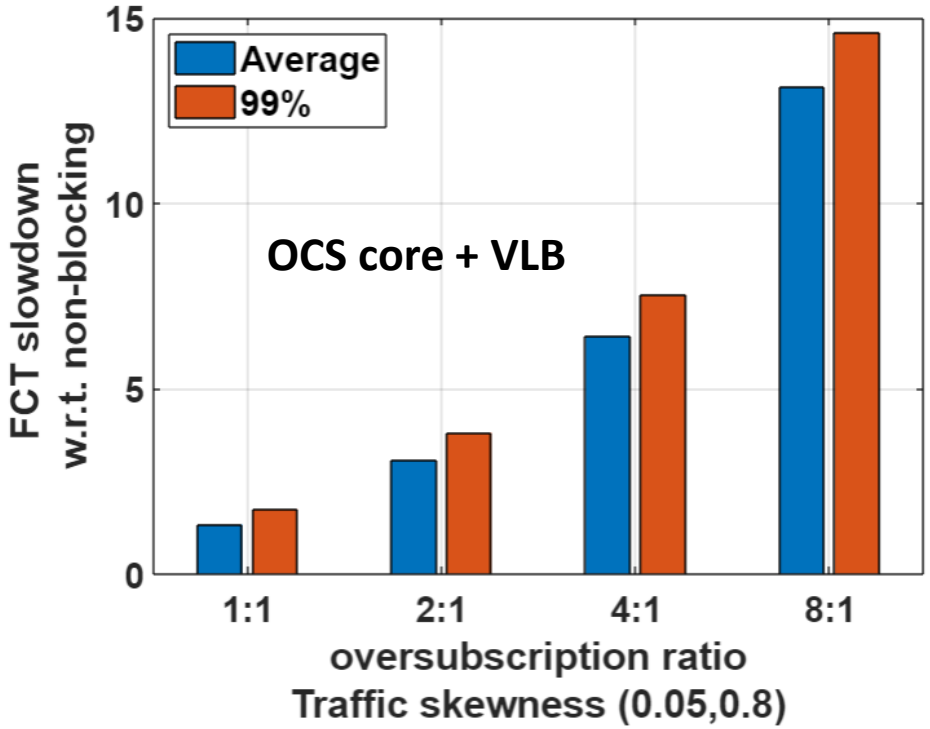
	A	B	C	D	E
T1	B	C	D	E	A
T2	C	D	E	A	B
T3	D	E	A	B	C
T4	E	A	B	C	D

- Path: A->B->D
- Leverage indirect hop to reduce latency

VLB Helps but Not Enough to Unlock Full Potential

Packet-level simulation

- VLB atop round-robin OCS core (Sirius) vs Non-blocking network
- High skewness (0.05,0.8) and vary oversubscription ratio (os)



os 1:1: Impact of traffic skewness

- Average FCT slowdown by 33%
- 99% FCT slowdown by 74.1%

Higher os: Impact of inter-rack traffic

- Performance degrades rapidly
- 8:1 os: Average FCT slowdown by 13.6x

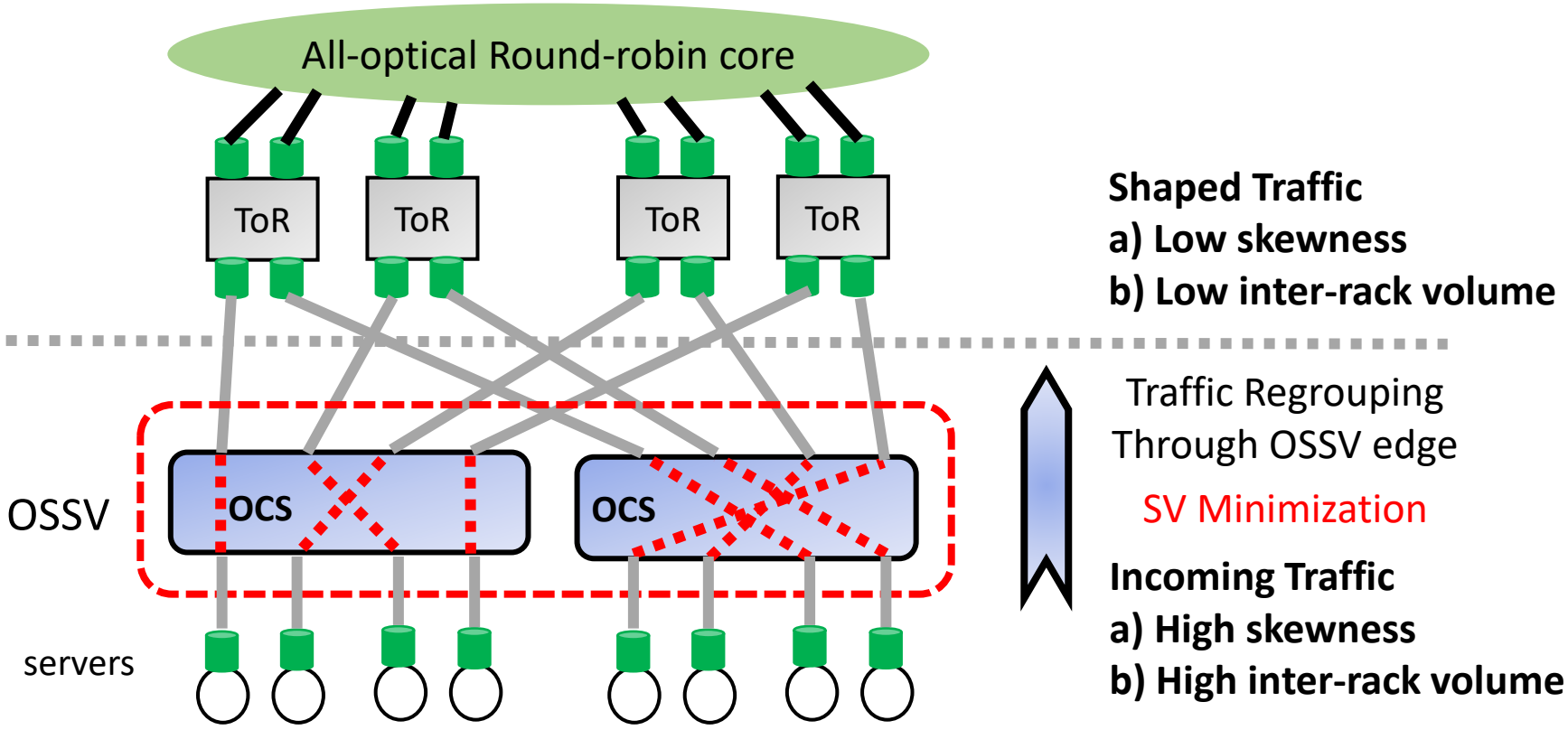
Root Cause

- Congestion at the intermediate hop
- Cannot reduce inter-rack traffic

How to efficiently support highly skewed + inter-rack traffic?

Our Proposal: Intelligent Traffic Regrouping at Edge

- OSSV [1]: Optical Substrate for Skewness and Volume Minimization
- Traffic agnostic core + Traffic adaptive reconfigurable edge



- Localize most of the traffic within a rack: Minimize inter-rack volume
- Remaining inter-rack traffic close to uniform: Minimize inter-rack skewness

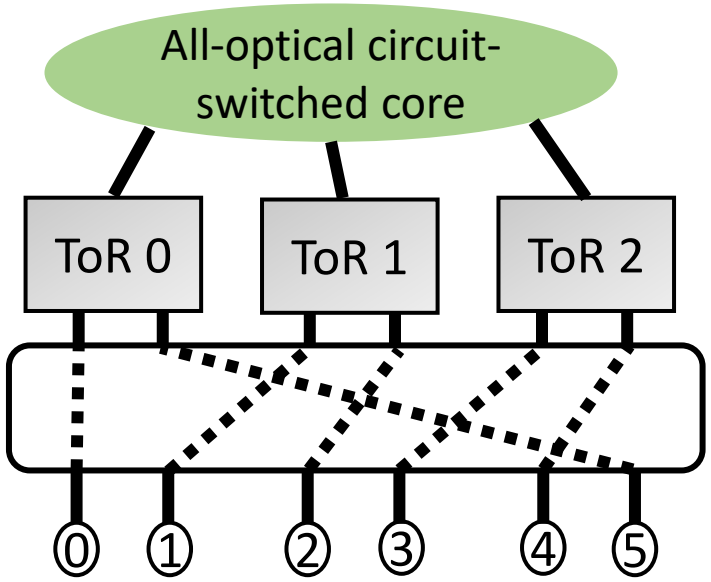
[1] Das, Sushovan, et al. “Rearchitecting Datacenter Networks: A New Paradigm with Optical Core and Optical Edge." IEEE INFOCOM, 2024.

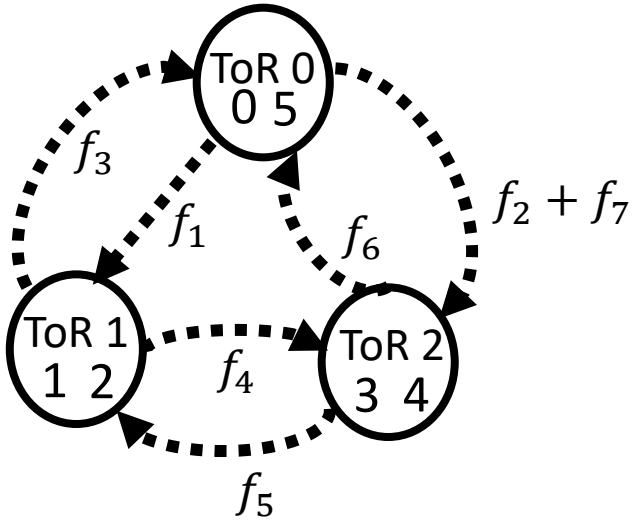
Skewness and Volume (SV) Minimization: Intuition

- Skewness minimization and traffic localization are inter-twined
- Naïve traffic localization can make skewness worse

f_1: 0->1 100000
f_2: 0->4 1000
f_3: 1->5 1000
f_4: 2->3 500
f_5: 3->2 500
f_6: 4->5 500
f_7: 5->4 500

Traffic demand





Config (a): Initial server-to-ToR mapping

Localized traffic: 0
JFI of inter-rack: 0.18

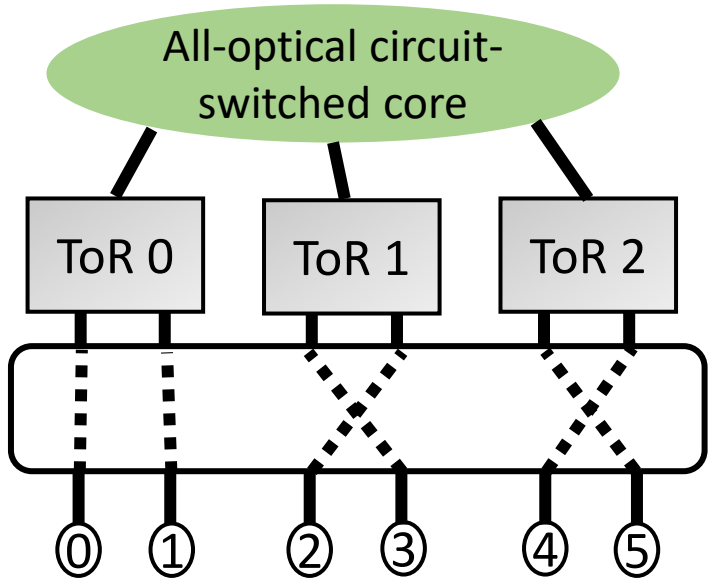
JFI: Jain's Fairness Index
Higher is better

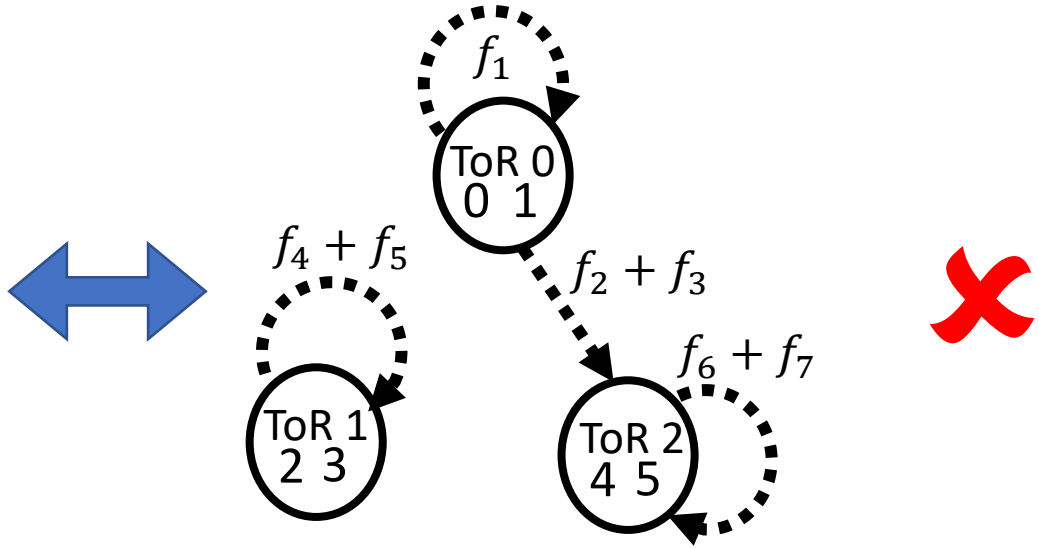
Skewness and Volume (SV) Minimization: Intuition

- Skewness minimization and traffic localization are inter-twined
- Naïve traffic localization can make skewness worse

f_1: 0->1 100000
f_2: 0->4 1000
f_3: 1->5 1000
f_4: 2->3 500
f_5: 3->2 500
f_6: 4->5 500
f_7: 5->4 500

Traffic demand





Config (b): Naïve traffic localization

Localized traffic: 102000
JFI of inter-rack: 0.17

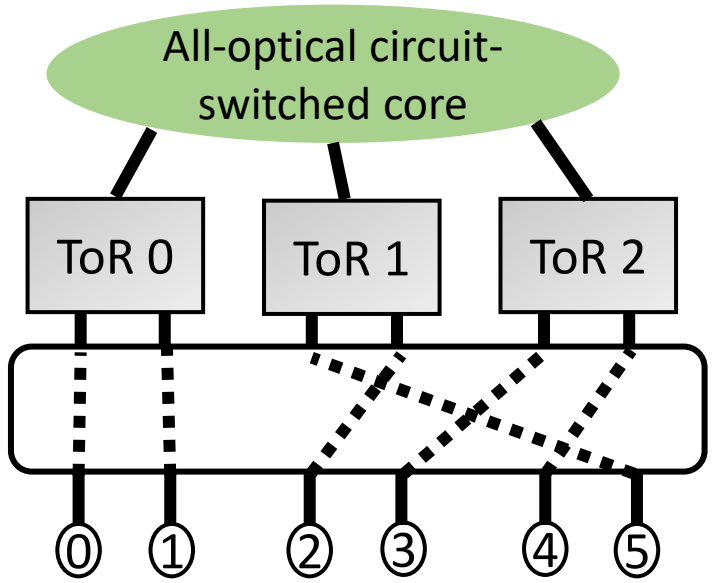
JFI: Jain's Fairness Index
Higher is better

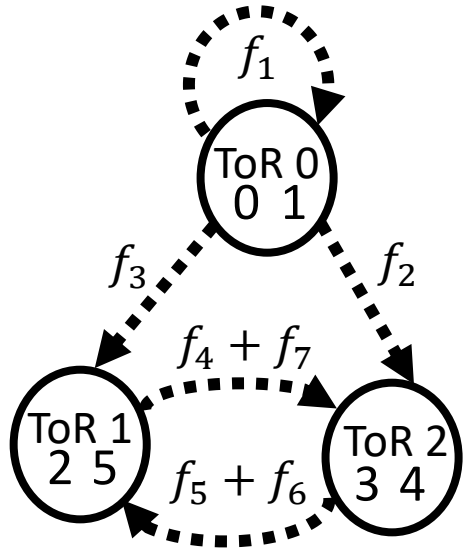
Skewness and Volume (SV) Minimization: Intuition

- Skewness minimization and traffic localization are inter-twined
- Naïve traffic localization can make skewness worse

f_1: 0->1 100000
f_2: 0->4 1000
f_3: 1->5 1000
f_4: 2->3 500
f_5: 3->2 500
f_6: 4->5 500
f_7: 5->4 500

Traffic demand





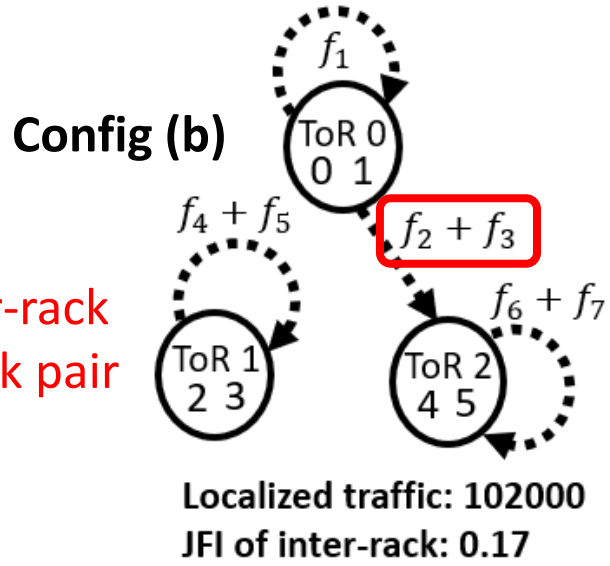
Config (c): SV minimization

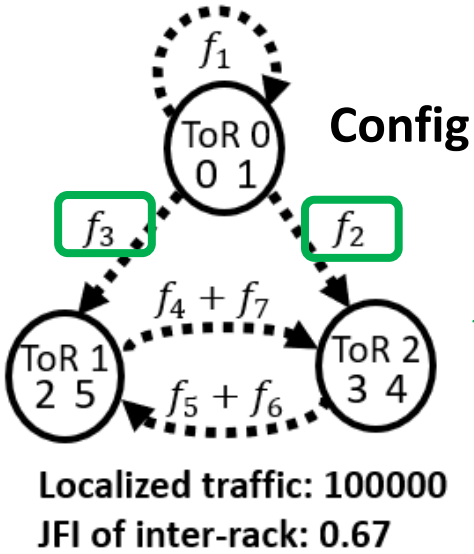
Localized traffic: 100000
JFI of inter-rack: 0.67
JFI: Jain's Fairness Index
Higher is better

Goal: Jointly optimize to find the right balance

SV Minimization: Designing Suitable Cost Function

f_2 and f_3 both inter-rack but under same rack pair





f_2 and f_3 both inter-rack but under different rack pairs

Linear cost function is not suitable

- Can't distinguish between Config (b) and Config (c) in terms of f_2 and f_3
- In both cases, cost is $f_2 + f_3 = 1000 + 1000 = 2000$
- Only penalizes sum of inter-rack traffic, NOT the variance

Intuitively, cost function (Φ) should satisfy:

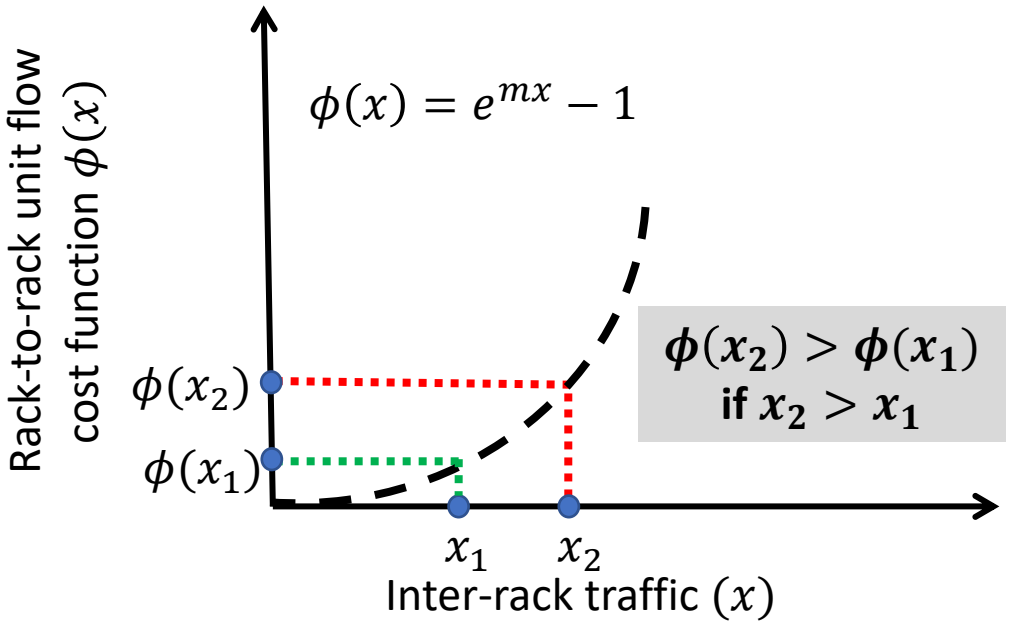
- $\Phi(1000) + \Phi(1000) < \Phi(2000)$

We need super-linear and strictly increasing cost function

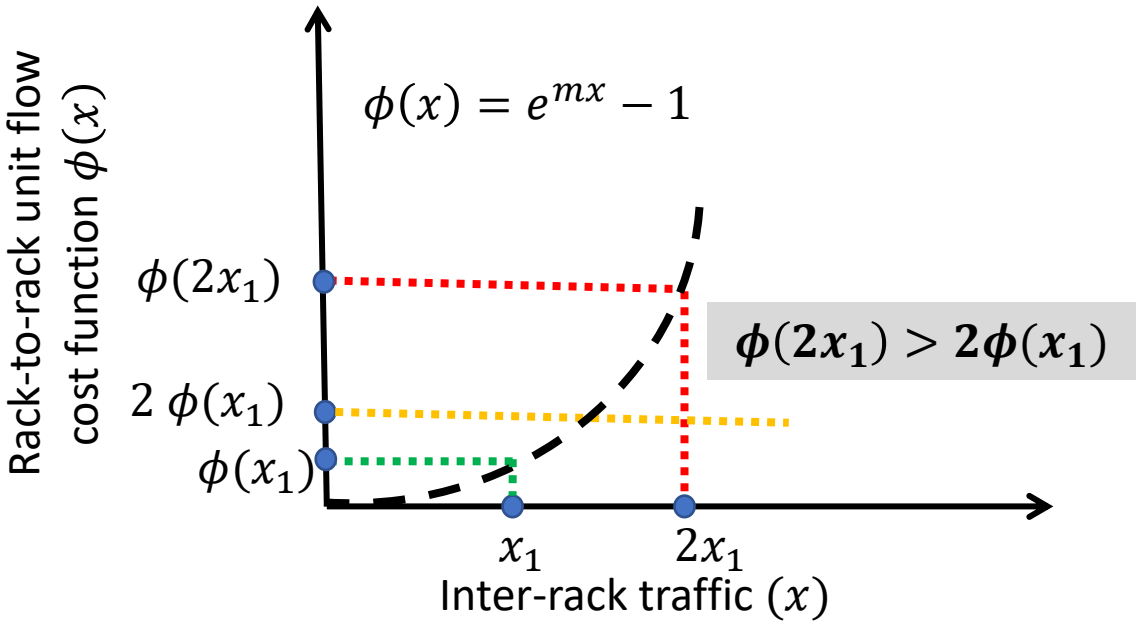
Exponential Cost Function: Achieve the Right Balance

$$\Phi(x) = e^{mx} - 1$$

x: traffic load between a rack pair
m: Tunable parameter controlling skewness-volume trade-off



Higher penalty for high inter-rack traffic value
(Minimize inter-rack Volume)



Higher penalty for high inter-rack variance
(Minimize Skewness)

Find optimal server-to-ToR configuration to minimize cost

SV Minimization: Problem Definition

Input: N: Number of ToRs , M: servers per ToR, f: server level traffic matrix

Find optimal server-to-ToR configuration (I) to minimize cost

$$Cost\ (f, B, \Phi, I) = \sum_{j=1}^{N} \sum_{i=1}^{N} \Phi_{ij}\left(F_{ij}, B_{ij}\right) F_{ij}$$

F: rack-level traffic matrix

$$F_{ij} = \sum_{s=1}^{MN} \sum_{s'=1}^{MN} I_{si}\, I_{s'j}\, f_{ss'}$$

$I_{si} = 1$, if server s belongs to ToR i

$I_{s'j} = 1$, if server s' belongs to ToR j

Φ: cost function per unit flow size between rack i to rack j

$$\Phi_{ij} = e^{\frac{mF_{ij}}{B_{ij}}} - 1$$

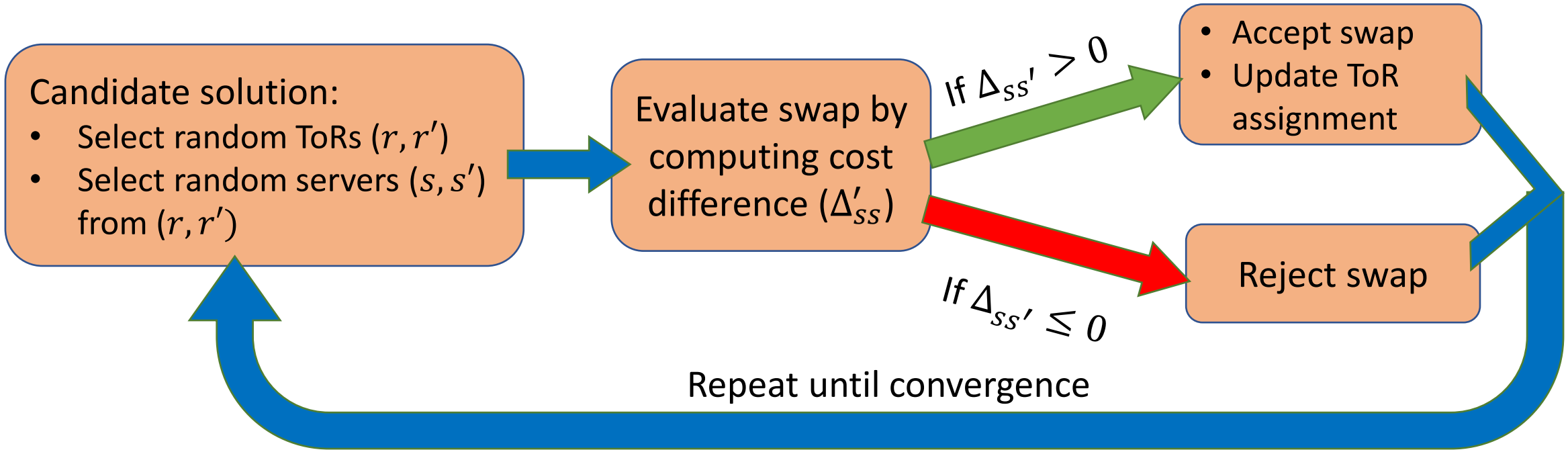
B_{ij} = Inter-rack bandwidth

$m = 2$

A variant of the Balanced Graph Partitioning (BGP) problem

NP-Hard !!!

Hill-climbing Based Randomized Heuristic



Naïve Evaluation of $\Delta_{ss'}$

- Computing F from scratch
- Complexity $O(M^2N^2)$
- Total servers = MN

Efficient Evaluation of $\Delta_{ss'}$

- Avoid Redundant Computation
- Complexity $O(MN)$
- Inspired by Kernighan-Lin Algorithm

Evaluation in Detail

Baseline Architectures
- OCS round-robin core
- OCS core + Valiant Load Balancing (VLB)
- Packet-switched cores

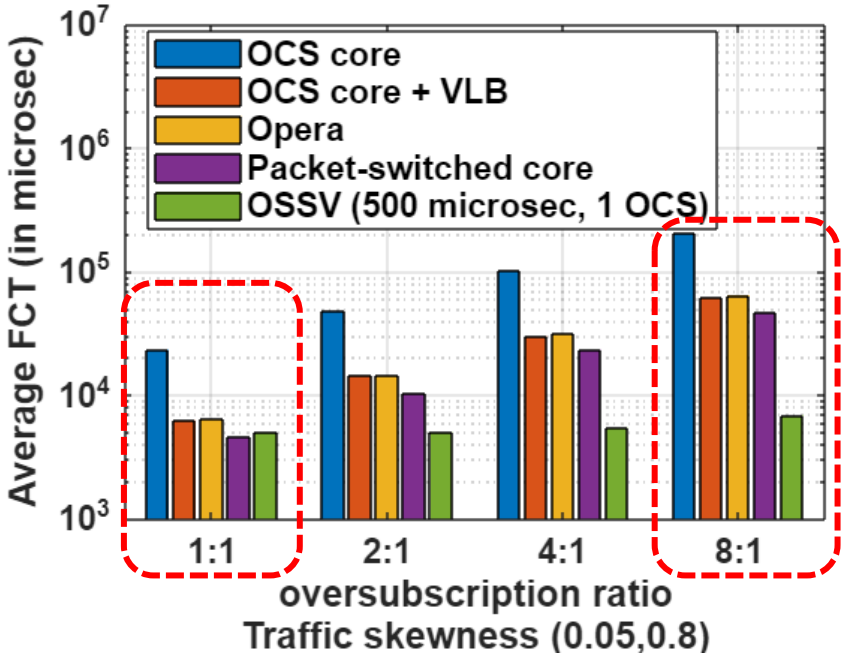
Performance
- Impact of skewness
- Impact of network load
- Impact of network oversubscription

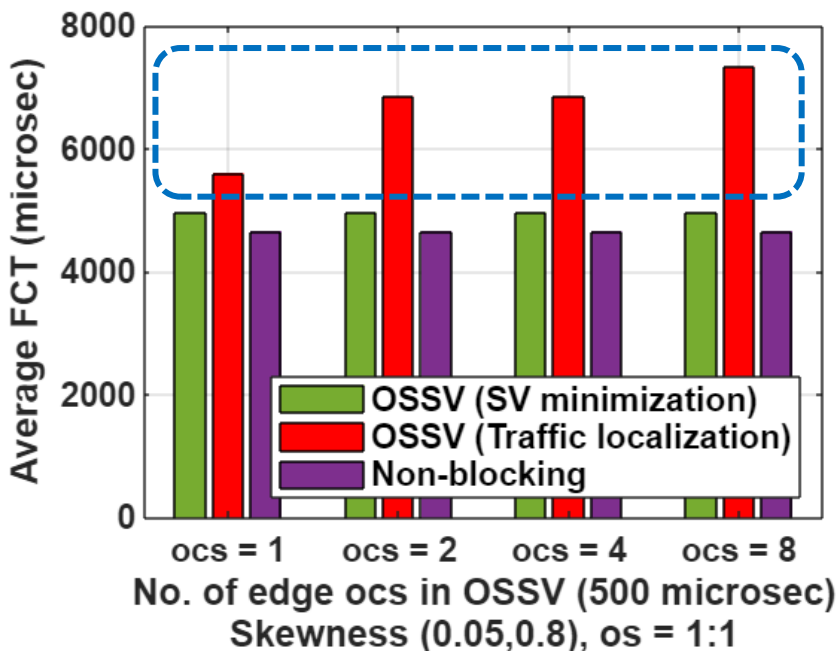
Ablation Study
- Impact of edge reconfiguration epoch
- SV-minimization benchmark
- Impact of multiple edge OCS with different cost function

Practical Viability
- Prototype implementation on testbed
- Power and cost analysis

Simulation Framework
- 32 servers per ToR, 16 ToRs, 512 servers, 100 Gbps
- Core OCS δ = 100 nanosec (AWGR), duty cycle 99%
- Edge OCS δ = 10 μsec (2D MEMS)
- Trace-based traffic (Cache) synthesized for different skewness and network load

Results: High Performance at Scale





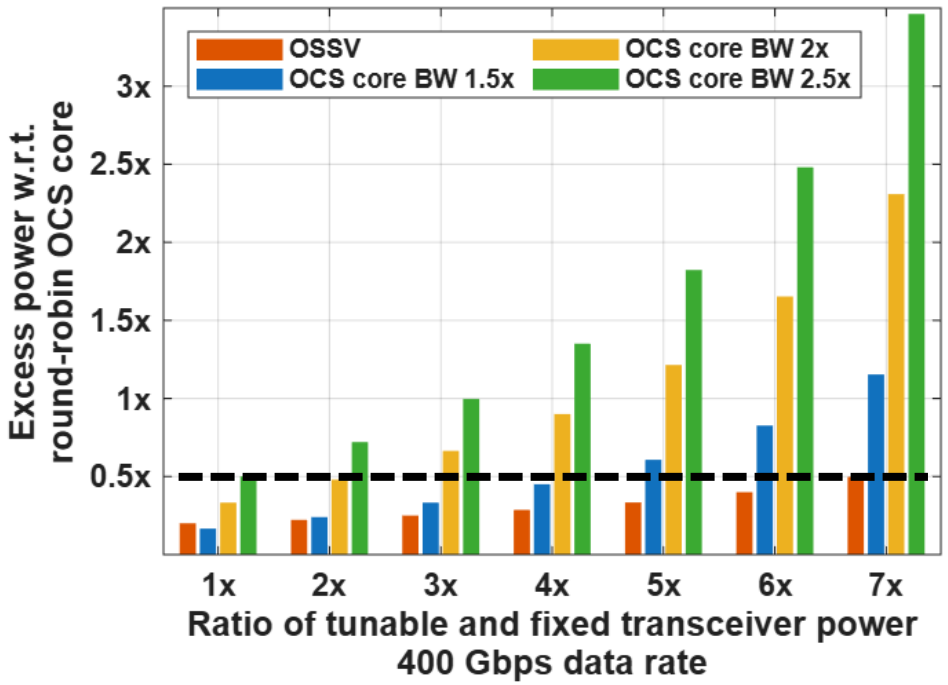
OSSV (green bar) vs. Packet-switched core (purple bar)

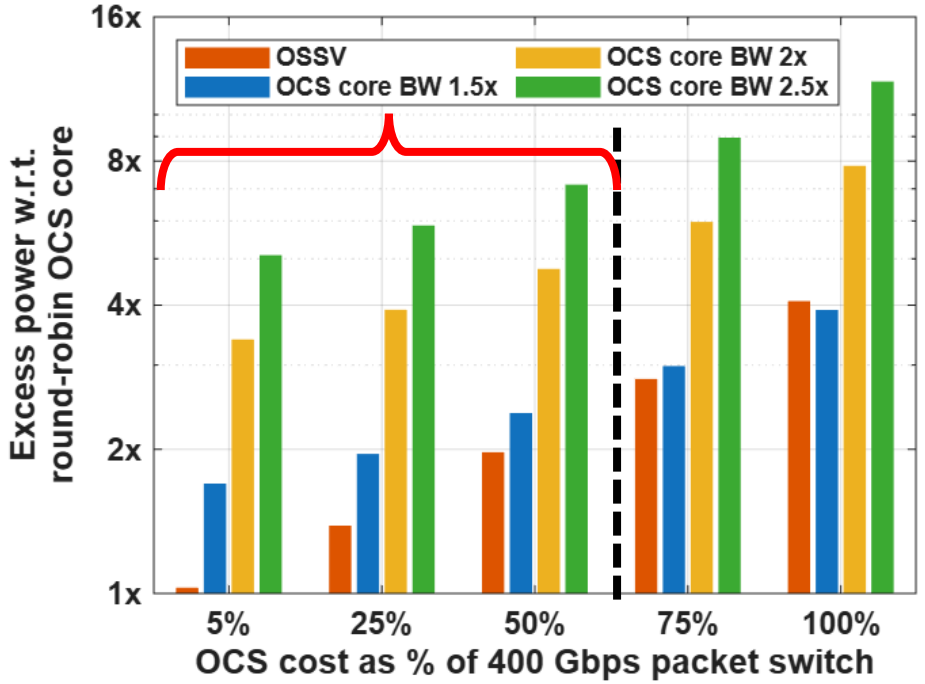
- At os 1:1: Avg FCT gap is within 6.4% of non-blocking
- At os 8:1: Avg FCT gap is within 47% of non-blocking

OSSV (one edge OCS) vs. OSSV (multiple edge OCS)

- OSSV (SV Min) 8-OCS vs. 1-OCS: degrades by 2.19% (green bar)
- Aggressive traffic localization harms inter-rack traffic fairness

Results: Power and Capital Cost Analysis





Power analysis: OSSV (orange bar) vs. OCS core

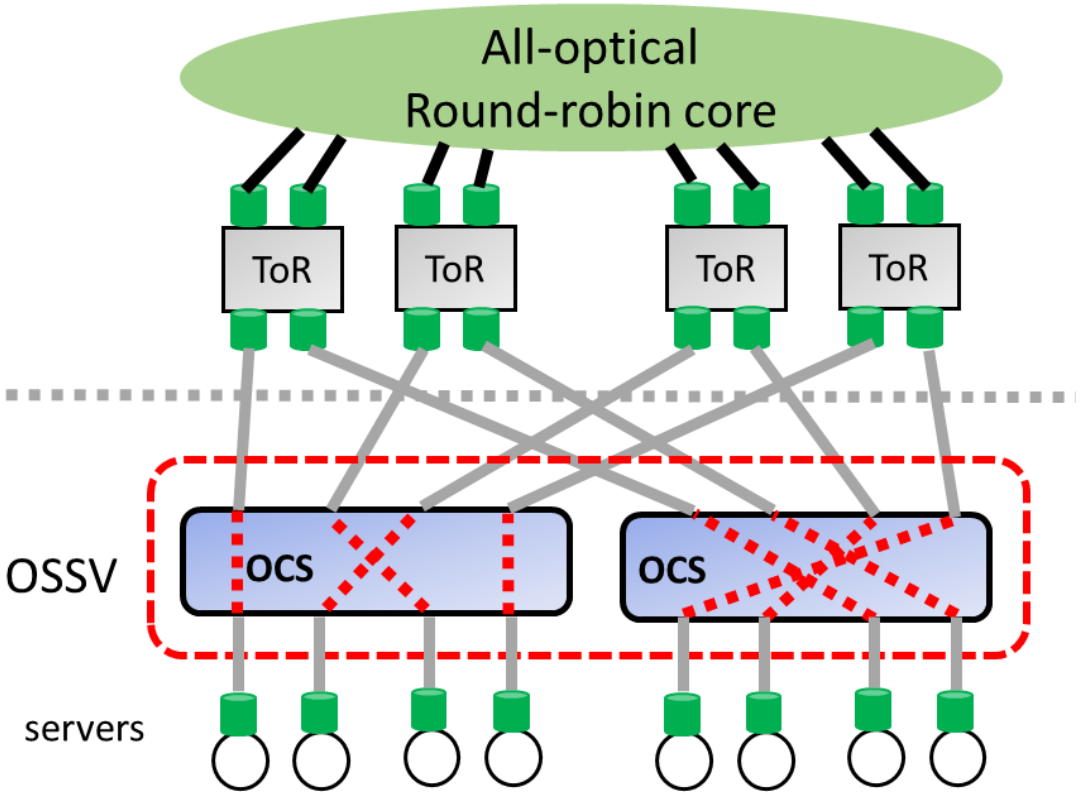
- Packet switch power = 23.5 Watt/port vs. MEMS OCS power = 0.14 Watt/port
- Keep most of core energy-saving benefit (20-50% excess)

Cost analysis: OSSV (orange bar) vs. OCS core

- Less expensive than overprovisioned OCS cores for realistic OCS cost range
- In longer term, OCS will cost will be amortized across generations

Summary of OSSV

- All-optical DCN core: Saves power but cannot support diverse network traffic
- OSSV: Traffic agnostic core + Traffic adaptive reconfigurable edge
- Novel SV Minimization: jointly optimize skewness and inter-rack traffic volume
- High Performance: Enables OCS core to perform close to non-blocking network
- Power and Cost Saving: deployable as next-gen cloud infrastructure



Q&A