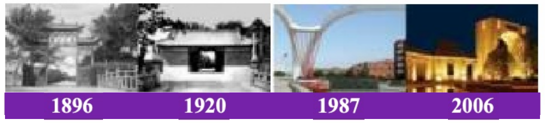
On the Design of High Performance Routing Protocols in AI Clusters

Shizhen Zhao

Associate Professor

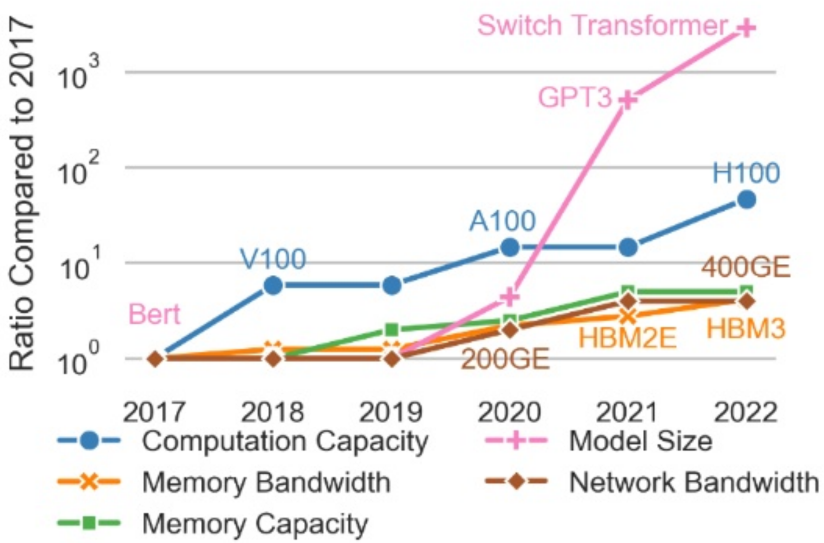
John Hopcroft Center for Computer Science

Shanghai Jiao Tong University



Background

AI traffic is growing faster than network bandwidth, and exhibit distinctive charateristics from traditional DCN traffic. Building a network specific for AI is crucial.



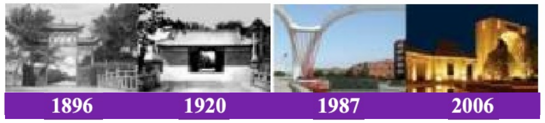
AI demand grows faster than network bandwidth

Collectives	Model Dist. purpose	Traffic pattern	Network congestion	GPU Msg (MB)	NIC Msg (KB)	Flow entropy per NIC	Topology need
`AlltoAll(v)`	Embedding distribution	full mesh with imbalanced traffic	N-to-N with possible incast	1	128	$\log(M * N)$	Full bisection bandwidth
`AllReduce`	DDP	Tree or Ring	2-to-1 incast for Tree	4 (Tree) 1 (Ring)	512	$\log(M)$	Tolerate over-subscription
`AllGather`	FSDP	Ring	1-to-1 low congestion	1	512	$\log(M)$	Tolerate over-subscription
`ReduceScatter`	FSDP	Ring	1-to-1 low congestion	1	512	$\log(M)$	Tolerate over-subscription

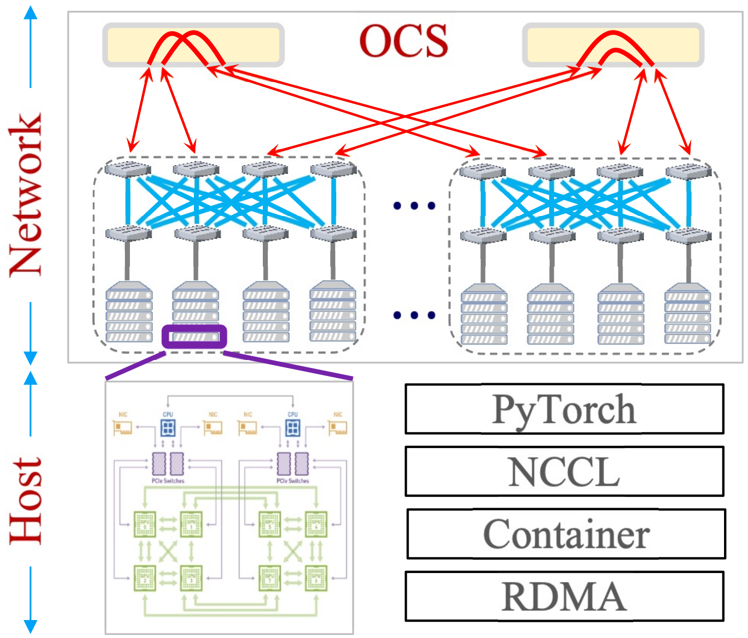
Characteristics of AI Workload:

1. Predictable, alltoallv may exhibit uncertainty.
2. Collective communication, traffic bursts together.
3. Low entropy, ECMP leads to load imbalance.
4. Delay sensitive, prone to long delay tail.

Buiding a Network for AI



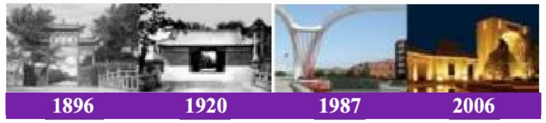
We build hybrid optical & electrical networks for AI, which requires cross-layer optimization for topology, routing, transport protocol, collective communication, etc.



Collective Communication	AI Training Migration	Communication Algorithm Design
Transport Protocol	RDMA NIC Pooling	Fine-grained Load Balancing [1]
Network Routing	Deadlock-free Routing	AI for Traffic Engineering [2]
Network Topology	Hybrid Optical & Electrical Topology	AI-Centric Topology Design

[1] Weihao Jiang, Wenli Xiao, Yuqing Yang, Peirui Cao, **Shizhen Zhao**, "Orderlock: A New Type of Deadlock and its Implications on High Performance Network Protocol Design," in SIGCOMM, 2025.
[2] Ximeng Liu, **Shizhen Zhao**, etc., "FIGRET: Fine-Grained Robustness-Enhanced Traffic Engineering," in SIGCOMM, 2024.

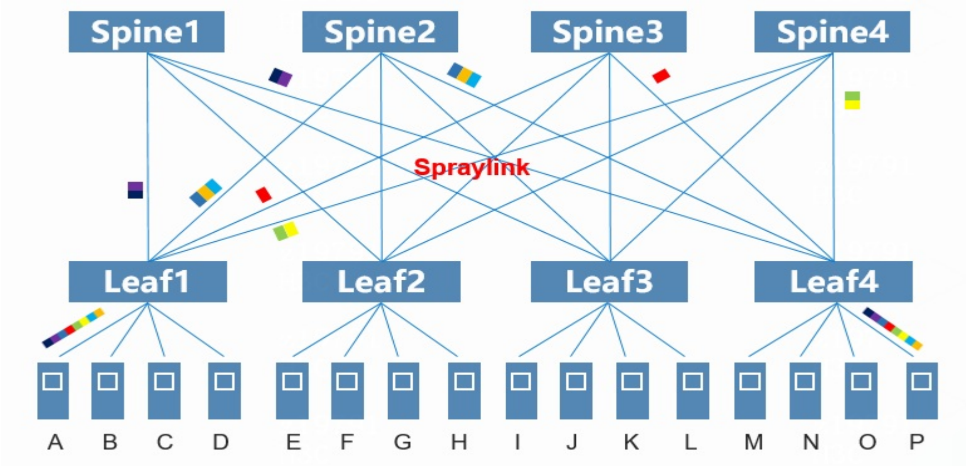
Fine-Grained Load Balancing



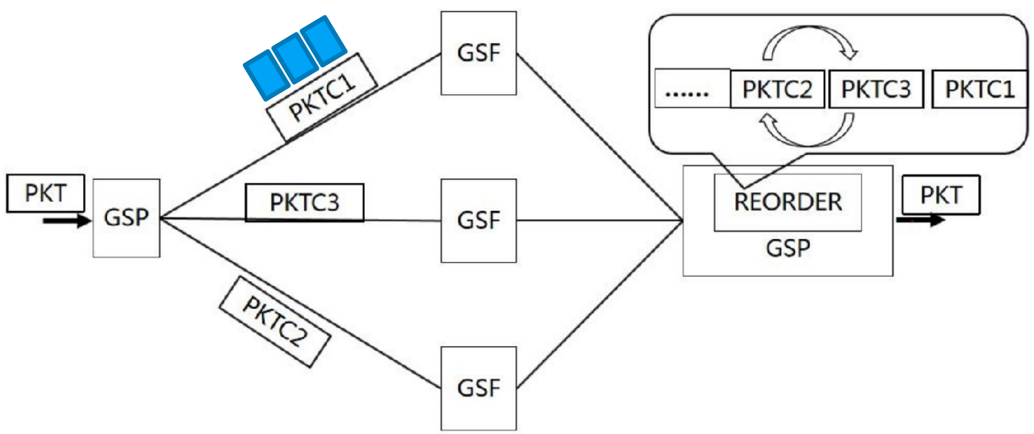
Orderlock: A New Type of Deadlock and its Implications on High Performance Network Protocol Design [SIGCOMM 2025]

Fine-grained LB for AI Workload

AI workload requires fine-grained load balancing to fully utilize network bandwidth, due to its low entropy. Fine-grained load balancing may cause packet disordering.

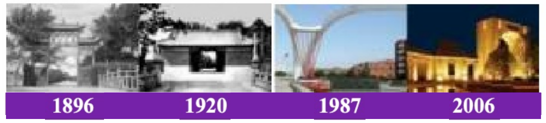


UEC (Ultra Ethernet Consortium) performs packet-level load balancing

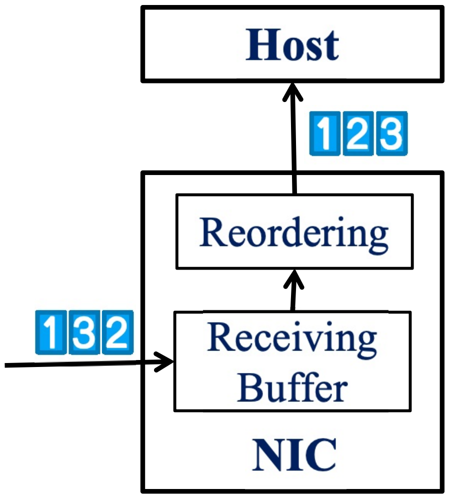


China Mobile's GSE performs packet container level load balancing

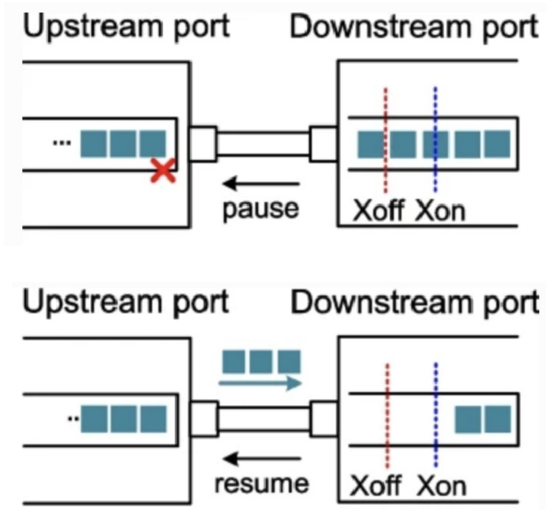
Desirable Features for AI Networks



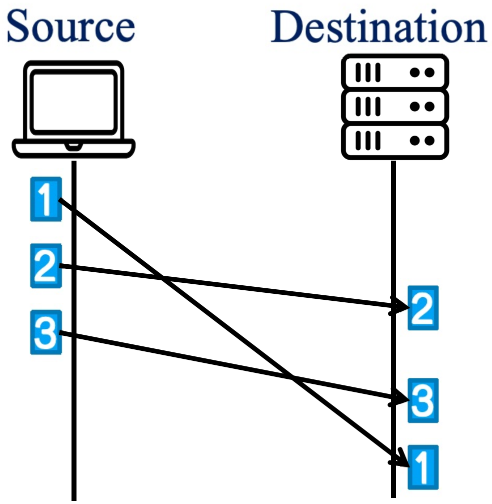
In-order delivery (I), Lossless Transmission (L) and Out-of-order Capability (O) are three important features for AI workloads to fully utilize network bandwidth.



In-order Delivery (I)



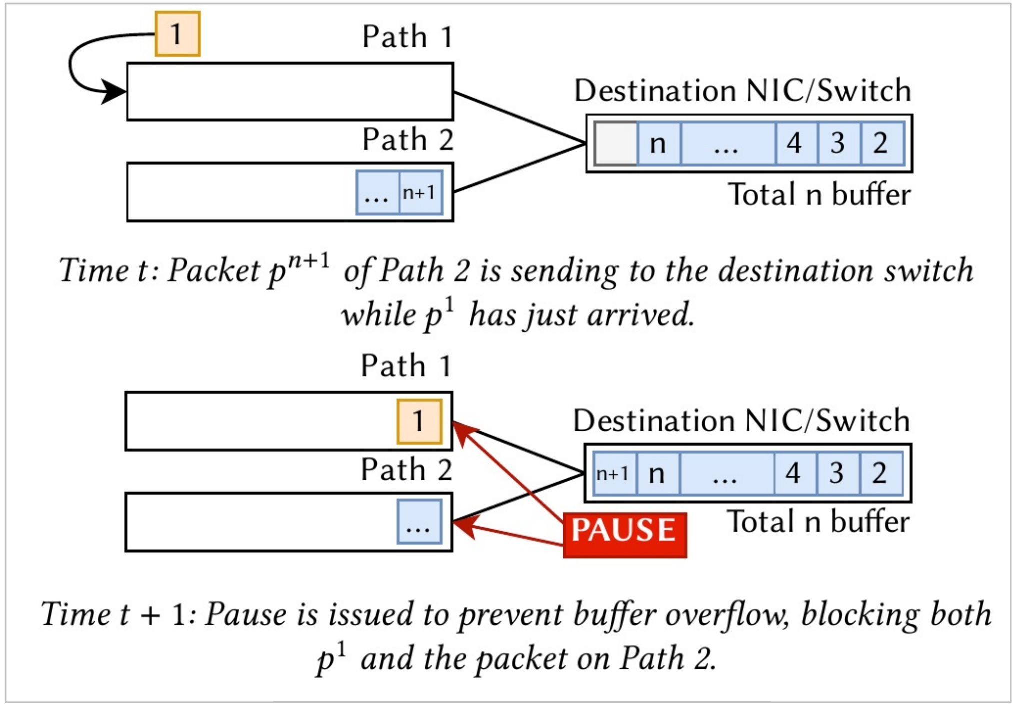
Lossless Transmission (L)



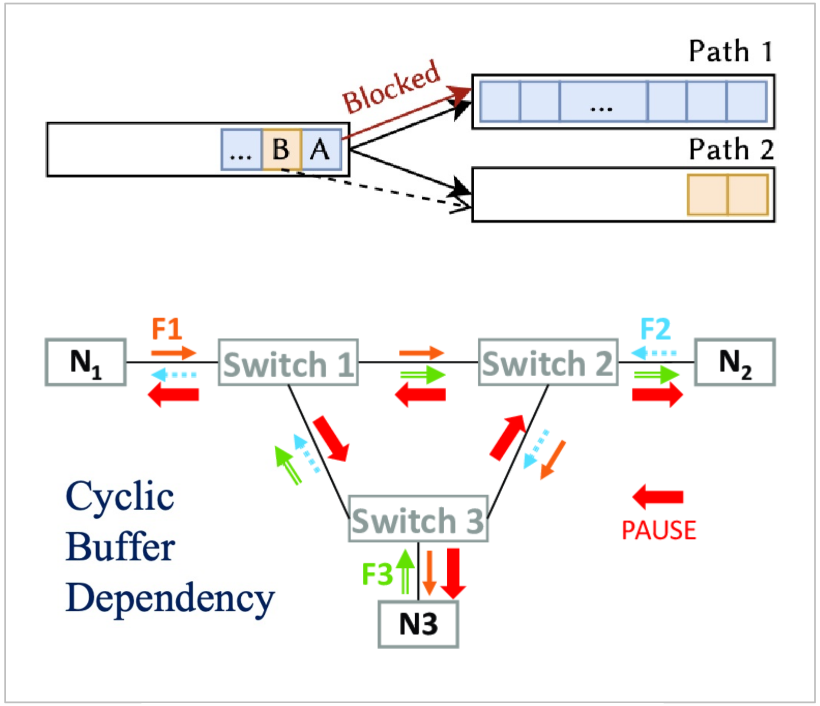
Out-of-order Capability (O)

ILO Triggers Orderlock

Simultaneously achieving In-order delivery, Lossless Transmission and Out-of-order Capability triggers Orderlock, which is different from HLB and PFC deadlock.

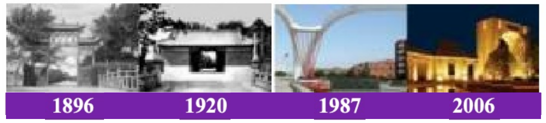


An example of Orderlock

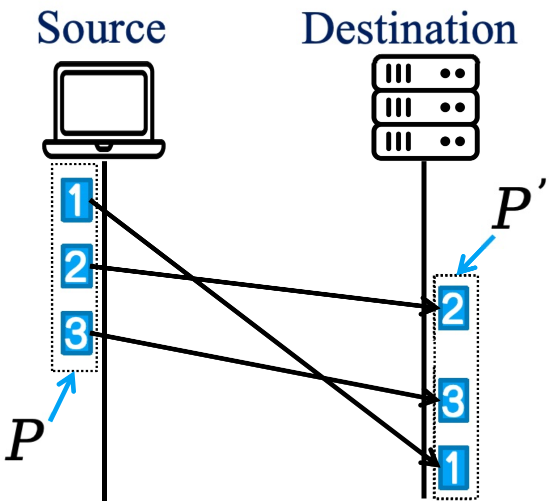


Orderlock is different from HLB and PFC deadlock

Rigorous Definition of O



For each packet p, we define its out-of-order arrival index OA(p) as the number of packets with larger index that arrives earlier than p. Then MOA charaterizes the maximum out-of-order level of the arrival sequence P'. We use MOA to define O.

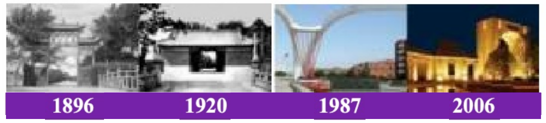


$$OA(p) := \|\{s \in P, \mathrm{idx}_P(s) > \mathrm{idx}_P(p)\ \mathit{and}\ \mathrm{idx}_{P'}(s) < \mathrm{idx}_{P'}(p)\}\|$$

$$MOA := \max\{OA(p), \forall p \in P'\}$$

Out-of-order Capability

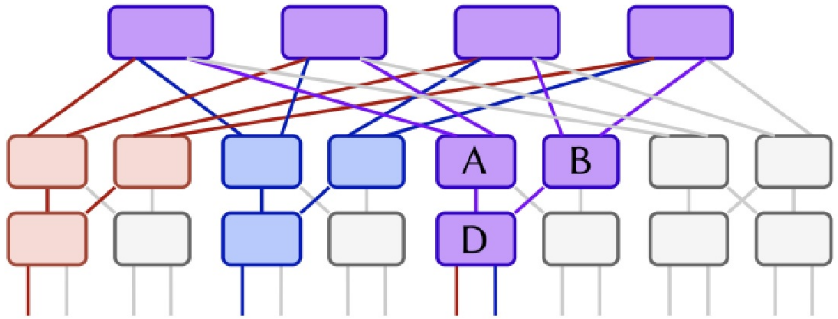
A network routing protocol is O if there exists P' such that its MOA is greater than or equal to the receiver's reordering limit.

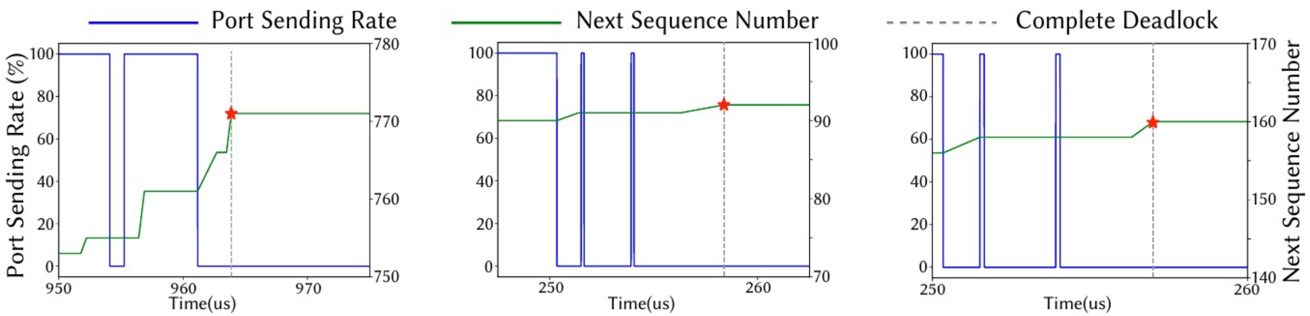


We test three O-policies. When combined with IL, all policies trigger orderlock even for a single flow. The triggering frequency increases as number of flows grows.

O-Policies

1. Random: randomly select a path for each packet at each switch;
2. QDAPS: chooses a port with slightly higher queue length to minimize out-of-order;
3. Adaptive: direct each packet to the port with the least queue length.



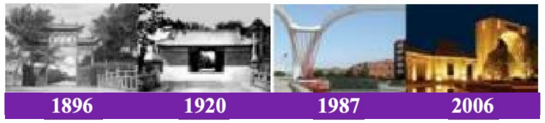


(a) Random (b) QDAPS (c) Adaptive

Table 1: Orderlock occurence rate

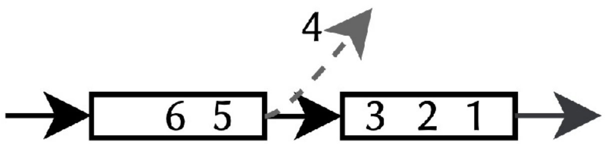
Strategy	Number of Flow 250	500	750
Random	7.1%	58.1%	90.4%
QDAPS	25.3%	67.3%	91.5%
Adaptive	20.1%	60.3%	86.1%
No reorder	0	0	0

Increasing Reordering Buffer?

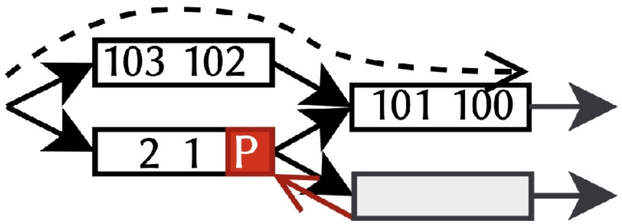


With packet spray and PFC, the on-the-fly packets of a flow can be far more than the bandwidth-delay product. Testbed experiments show that MOA and reordering buffer requirement of a flow grows with flow size.

Analysis

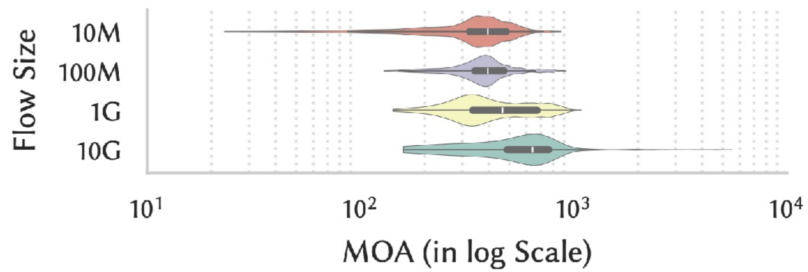


Without packet spray, BDP is enough



With packet spray, PFC increases MOA

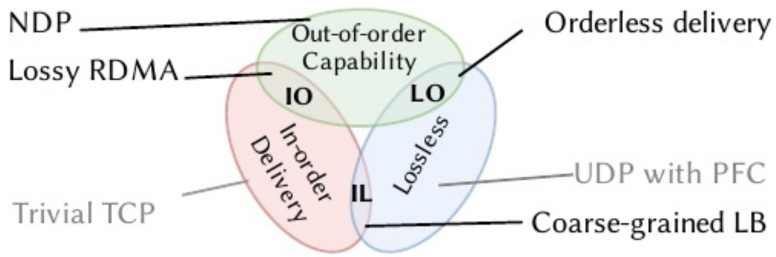
Testbed Measurement



Buffer size (n)	Flow Size				
	1MB	10MB	100MB	1GB	10GB
64	7.1%	99.7%	100%	100%	100%
128	0	97.5%	100%	100%	100%
256	0	87.0%	93.4%	92.8%	92.8%
(Flow ECMP)	0	0	0	0	0

Avoid Orderlock

We test 6 orderlock-free routing policies with AI collective communication workload. LO-policies offer the best performance.



Typical Routing Policies

IL: ECMP, Flowlet, Conweave, Credit-based FC
IO: IRN O: NDP LO: SRNIC

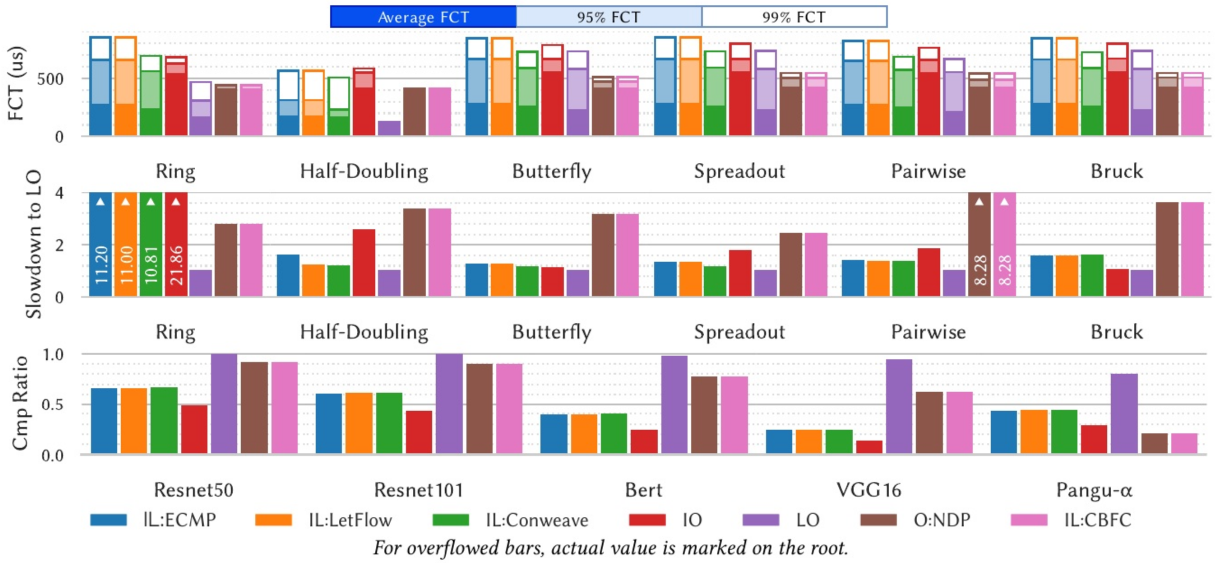
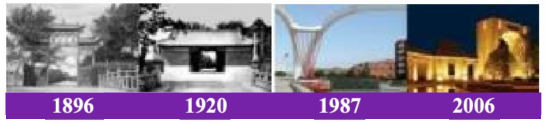


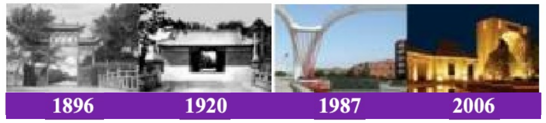
Figure 13: Performance comparison

Traffic Engineering



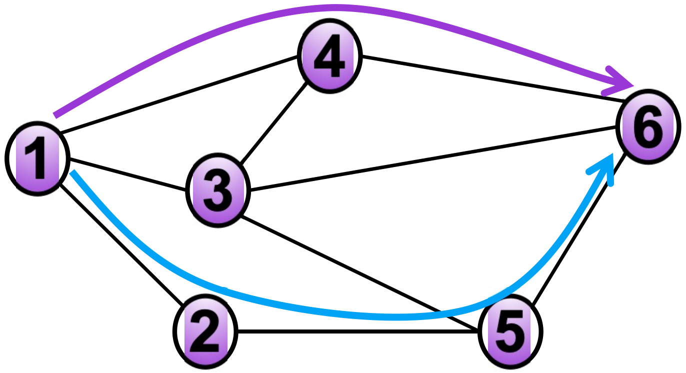
FIGRET: Fine-Grained Robustness-Enhanced Traffic Engineering [SIGCOMM 2024]

AI Workload Needs TE

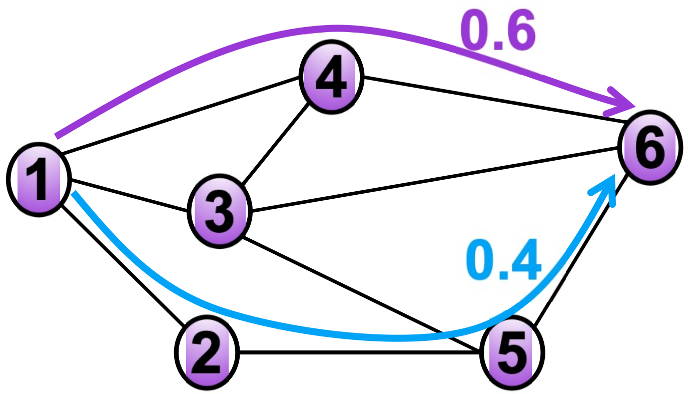


O-Policies performs fine-grained load balancing among multiple paths, but how to split the traffic?

Traffic engineering calculates split ratios among all the paths based on user's traffic demand matrix.

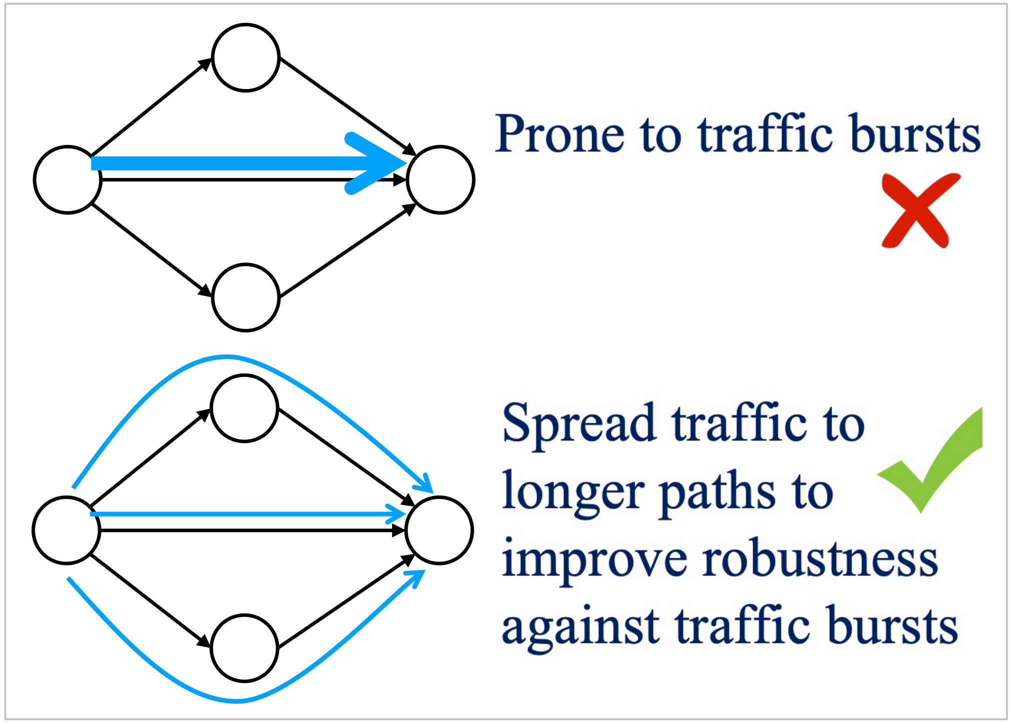


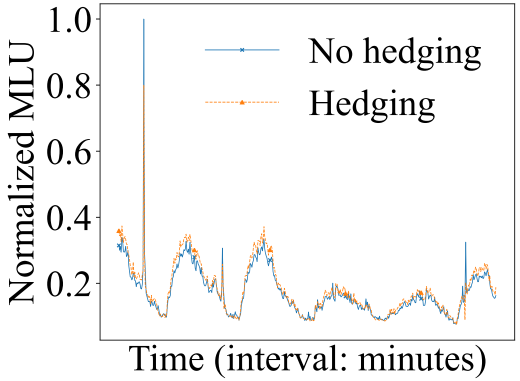
$$\begin{bmatrix} 0 & 34 & \cdots & 8 \\ 9 & 0 & \cdots & 22 \\ \vdots & \vdots & 0 & \vdots \\ 11 & 10 & \cdots & 0 \end{bmatrix}_{6\times 6} \Rightarrow$$

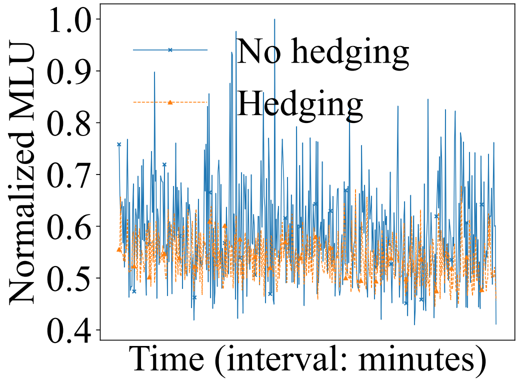


Handling Traffic Uncertainty in TE

CHALLENGE: Network traffic matrix (TM) keeps changing. Even if we compute the optimal TE based on the current TM, TM may change at the next time instance. Note that calculating new TE solutions takes time.

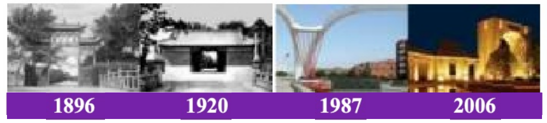




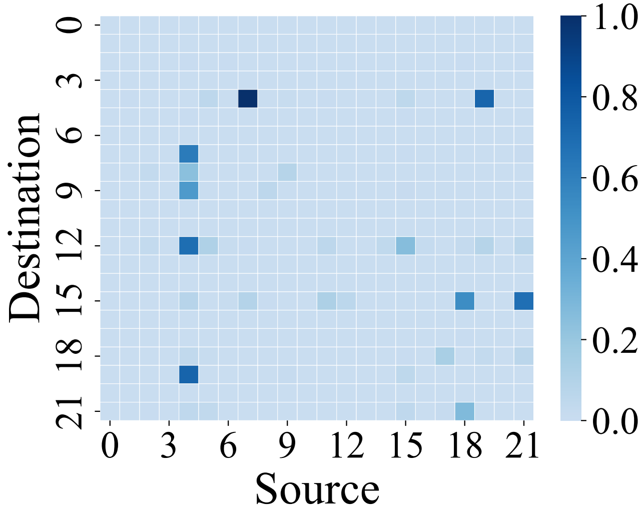


Traffic spreading (called hedging in Google) reduces MLU under traffic bursts, but increases average hop count.

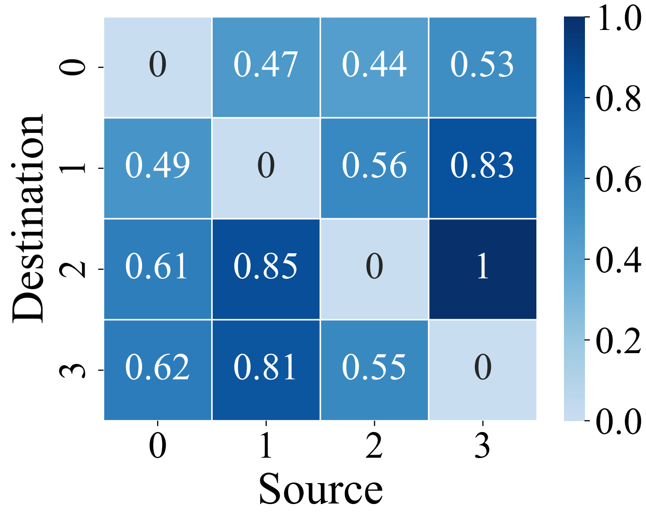
Diversity in Traffic Characteristics



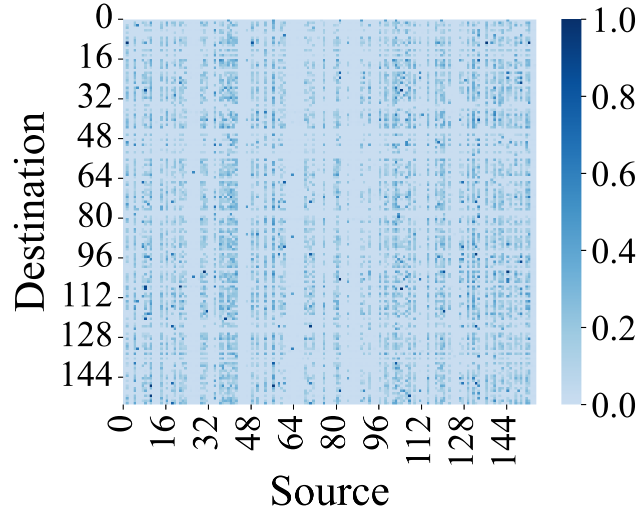
OBSERVATION: Different source-destination pairs have varying burst levels. Treating all SD pairs equally when handling bursts is suboptimal in performance.



GEANT, WAN



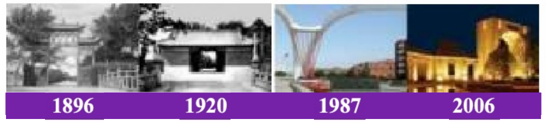
Meta, PoD-level DC



Meta, ToR-level DC

The variance of traffic demand by source and destination

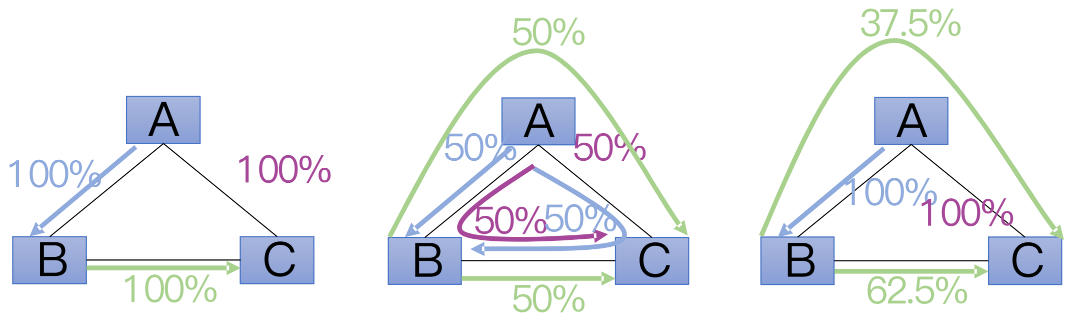
Key Insight



KEY IDEA: Treat different SD-pairs differently based on their likelyhood of burst.

Capacity	**Demand**	
	Normal	Burst
A↔B: 2 A↔C: 2 B↔C: 2	A→B: 1 A→C: 1 B→C: 1	A→B: 1 A→C: 1 B→C: 4

Only SD pair B→C may experience bursts.



Strategy 1 Strategy 2 Strategy 3

	Normal performance	Burst
Strategy 1	0.5	2
Strategy 2	0.75	1.5
Strategy 3	0.6875	1.25

Strategy 3 achieves a better performance tradeoff

Burst-aware TE Formulation

1896 1920 1987 2006

KEY IDEA: Minimize the product of demand variance and maximum split ratio.

Original TE

$$\min_{w} \mu$$

$$\begin{cases} w_{ij}^1 + w_{ij}^2 + \dots + w_{ij}^K = 1 \\ \forall \text{edge } l,\ \sum_{l \in P_{ij}^k} D_{ij} w_{ij}^k \le \mu C_l \end{cases}$$

Many w_{ij}^k's are 1.
Prone to traffic bursts

Traffic of an SD pair on path p:
$(D_{ij} + \delta_{ij}) \times \boldsymbol{w_{ij}^k}$

Utilization of edge e on path p affected by the burst:
$\delta_{ij} \times \boldsymbol{w_{ij}^k / C_l}$

Burst-aware TE

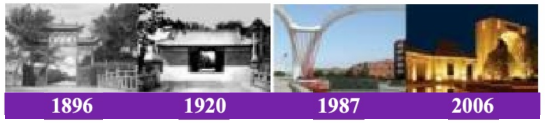
$$\min_{w} \{ \mu + \alpha \sum_{ij} Var(D_{ij}) \times \max_{k} \{w_{ij}^k\} \}$$

$$\begin{cases} w_{ij}^1 + w_{ij}^2 + \dots + w_{ij}^K = 1 \\ \forall \text{edge } l,\ \sum_{l \in P_{ij}^k} D_{ij} w_{ij}^k \le \mu C_l \end{cases}$$

Flow burstiness
$Var(D_{ij})$ increase

Weight bound
$\max_{k} \{w_{ij}^k\}$ decrease

Computational complexity of burst-aware TE is high!



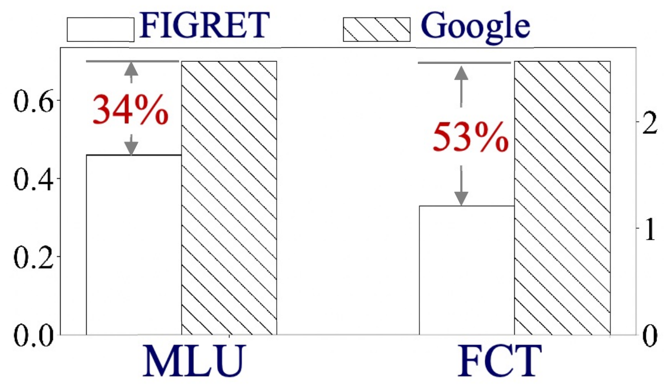
AI-aided TE Solver

SOLUTION: Use deep learning to solve TE problems, speed up over 1000×.

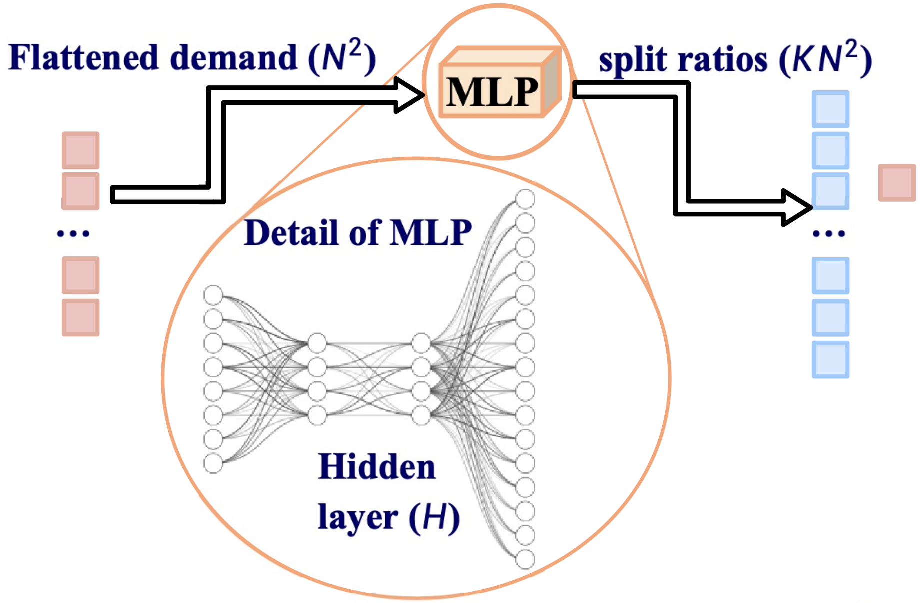
Burst-aware TE

$$\min_{w} \{ \mu + \alpha \sum_{ij} Var(D_{ij}) \times \max_{k} \{w_{ij}^k\} \}$$

$$\begin{cases} w_{ij}^1 + w_{ij}^2 + \ldots + w_{ij}^K = 1 \\ \forall \text{edge } l,\ \sum_{l \in P_{ij}^k} D_{ij} w_{ij}^k \le \mu C_l \end{cases}$$

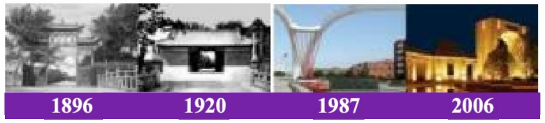


FIGRET

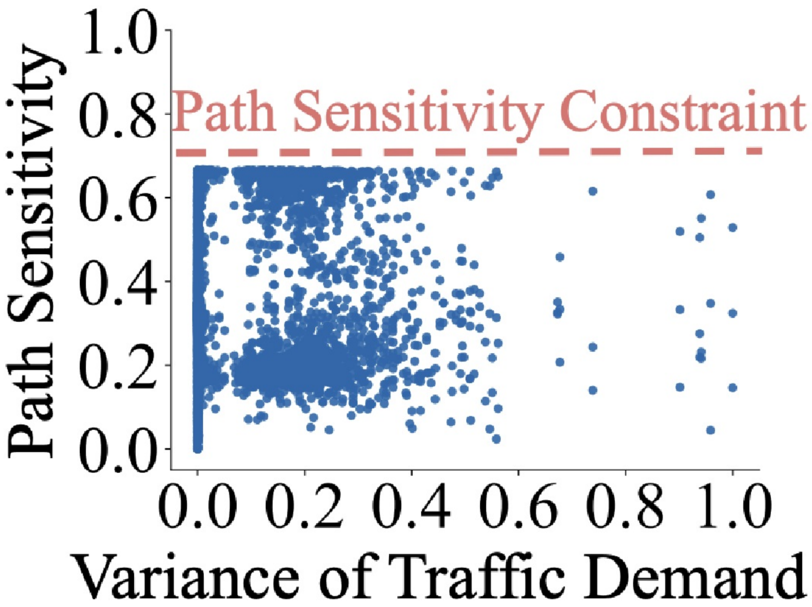


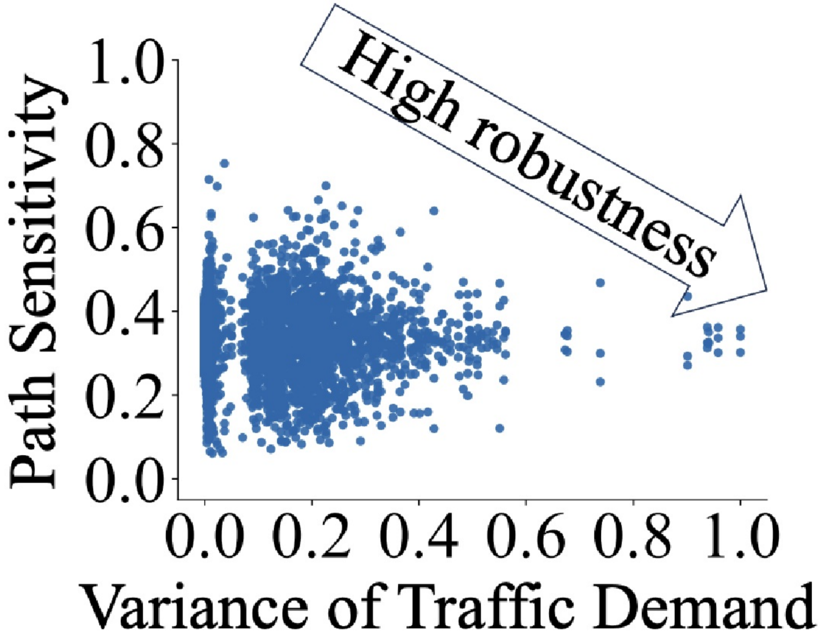
Loss function: $\mu + \alpha \sum_{ij} Var(D_{ij}) \times \max_{k} \{w_{ij}^k\}$

Why FIGRET Performs Better?

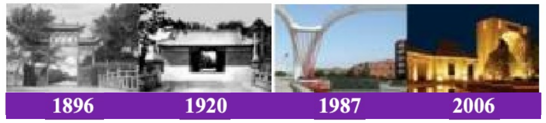


UNDERSTANDING: We plot (demand variance, demand split ratio) pairs, and find that FIGRET enforces lower demand split ratios for high variance demand.





New Opportunities of AI for TE



Handle Network Failures (APNet'25 [1]):

- Traditional failure-resilient TE has poor compatibility

Handle Network Dynamics (Arxiv'25 [2]):

- Retraining is costly when network changes

Joint Optimization of Collective Communication and Routing (Ongoing):

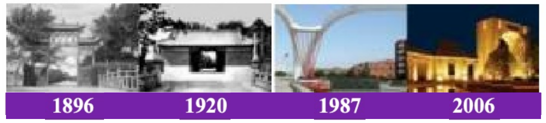
- Alltoallv is bottleneck; existing CCL optimization algorithms are too slow

Close-loop Traffic Matrix Prediction for TE:

- Minimizing MSE may not be a good choice for TM prediction in TE

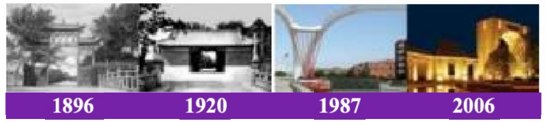
[1] Xiyuan Liu, Yang Liu, Jingyi Cheng, Ximeng Liu, Shizhen Zhao, "FauTE: Fault-tolerant Traffic Engineering in Data Center Network," in APNet, Shanghai, China, August, 2025

[2] Ximeng Liu, Shizhen Zhao, Xinbing Wang, "Geminet: Learning the Duality-based Iterative Process for Lightweight Traffic Engineering in Changing Topologies," available online: https://arxiv.org/pdf/2506.23640v1

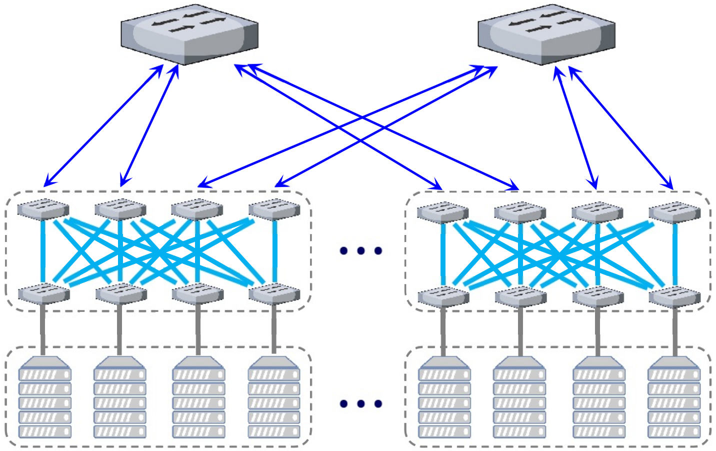


Questions?

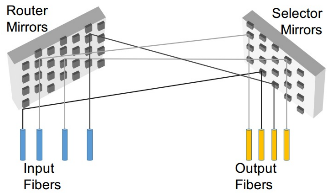
Electrical Core vs Optical Core



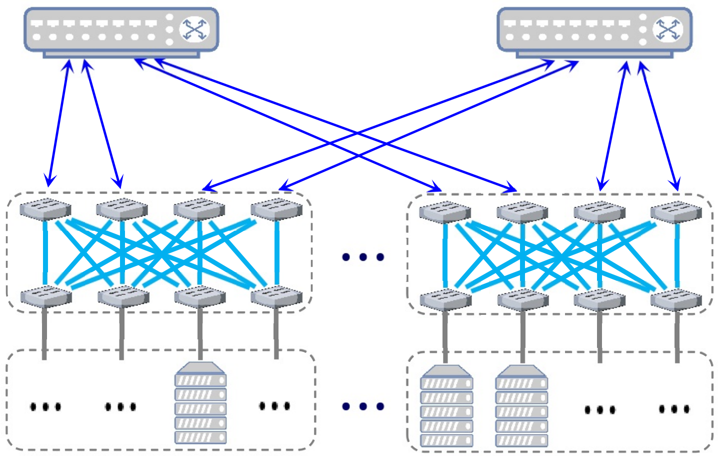
Electrical Packet Switches (EPS)



Replace the core layer EPSes by OCSes



Optical Circuit Switches (OCS)



Broadcom Tomahawk 5 chip bandwidth 51.2T
Configuration: 800Gbps x 64 ports
3-layered FatTree: 2*32^3=65536 nodes
Configuration: 1.6Tbps x 32 ports
3-layered FatTree: 2*16^3=8192 nodes

OCS size: 256 speed agnostic ports
Broadcom Tomahawk 5 chip bandwidth 51.2T
Configuration: 800Gbps x 64 ports
Hybrid network: 256*32^3=262144 nodes
Configuration: 1.6Tbps x 32 ports
Hybrid network: 256*16^2=65536 nodes