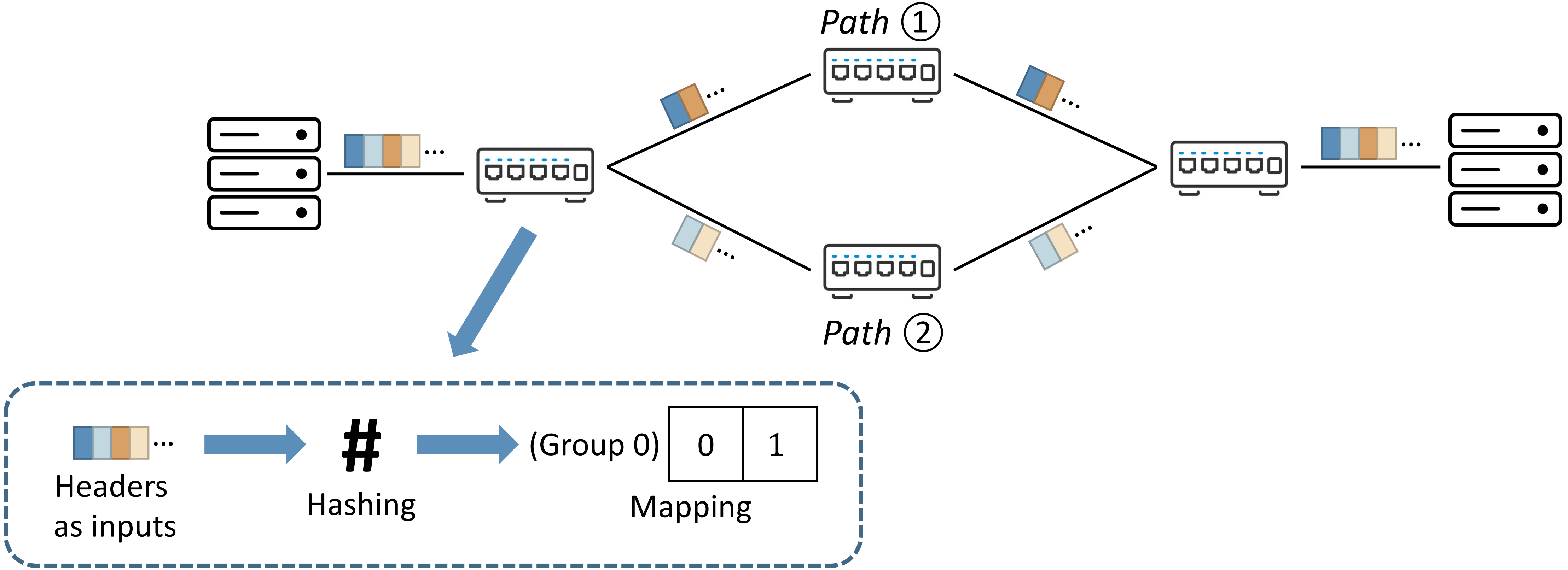
Unlocking ECMP Programmability for Precise Traffic Control

Yadong Liu*, **Yunming Xiao***, Xuan Zhang, Weizhen Dang, Huihui Liu, Xiang Li, Zekun He, Jilong Wang, Aleksandar Kuzmanovic, Ang Chen, **Congcong Miao**

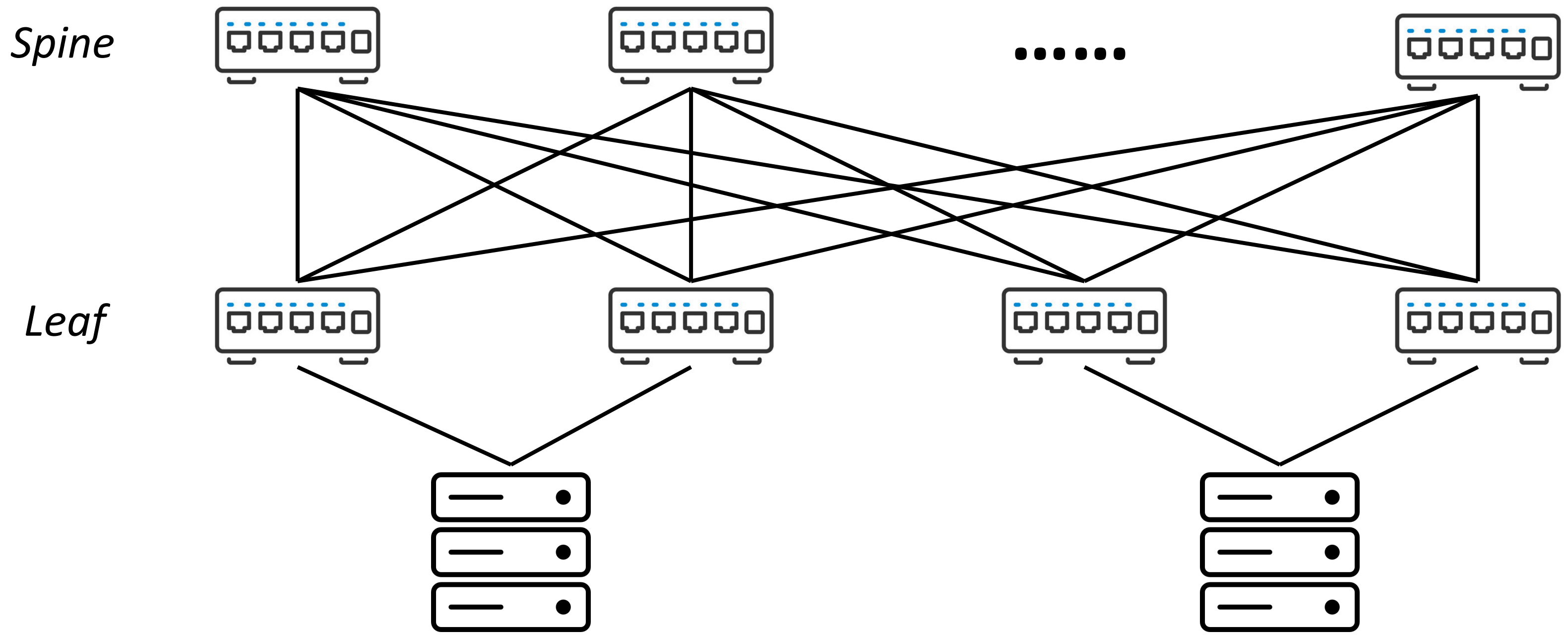
Randomness is vital in modern networking

- ECMP (equal-cost multi-path) routing is prevalent
  - Performance (load balancing), fault tolerance (redundancy), …



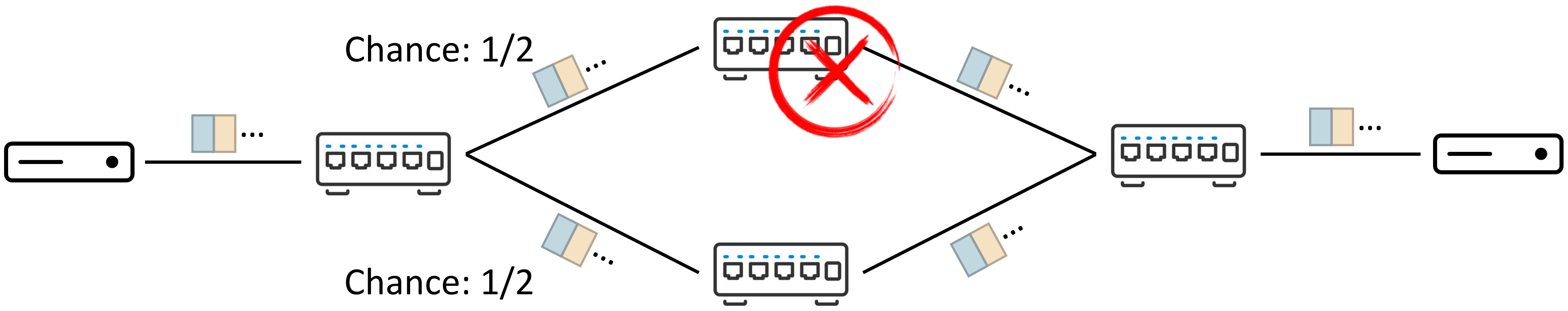
Randomness is vital in modern networking

- ECMP (equal-cost multi-path) routing is prevalent
  - Performance, fault tolerance, …
- Data center networks today are built around ECMP



Yet, pitfalls of randomness exist

- Various critical scenarios demand *Precise Traffic Control* (PTC)
  - Network failover by re-pathing flows



# Trial (T)	1	2	3	...
Success Chance	½	(½)+(½)^2	(½)+(½)^2+(½)^3	...

Yet, pitfalls of randomness exist

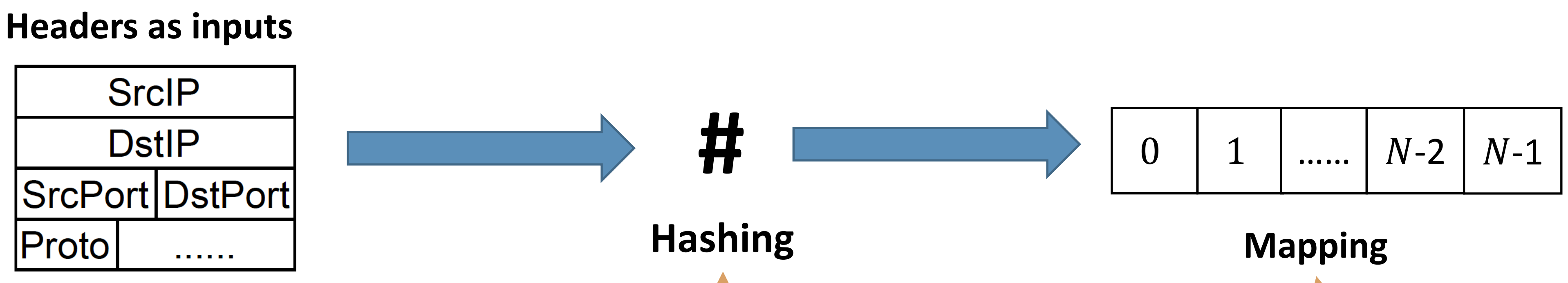
- Various critical scenarios demand *Precise Traffic Control* (PTC)
  - Network failover by re-pathing flows
  - Load balancing
  - Network failure localization (probe all paths)
  - Multi-flow applications
  - Multi-path protocols
  - Packet spraying
  - Segment routing
  - ……

An ideal world: randomness and PTC co-exist

- But there are challenges:
  - Variety of PTC scenarios → we hope to support all of them
  - Compatibility → we hope to deploy it today

Randomness Law and Order

Can we modify ECMP to realize such world?



Solution #1: Update field to change outcome[1]
Cons: Need knowledge of hash algorithm & parameters;
Does not work in heterogenous environment

Solution #2: Change hashing algorithm
Cons : Randomness still exists, no guarantee of PTC

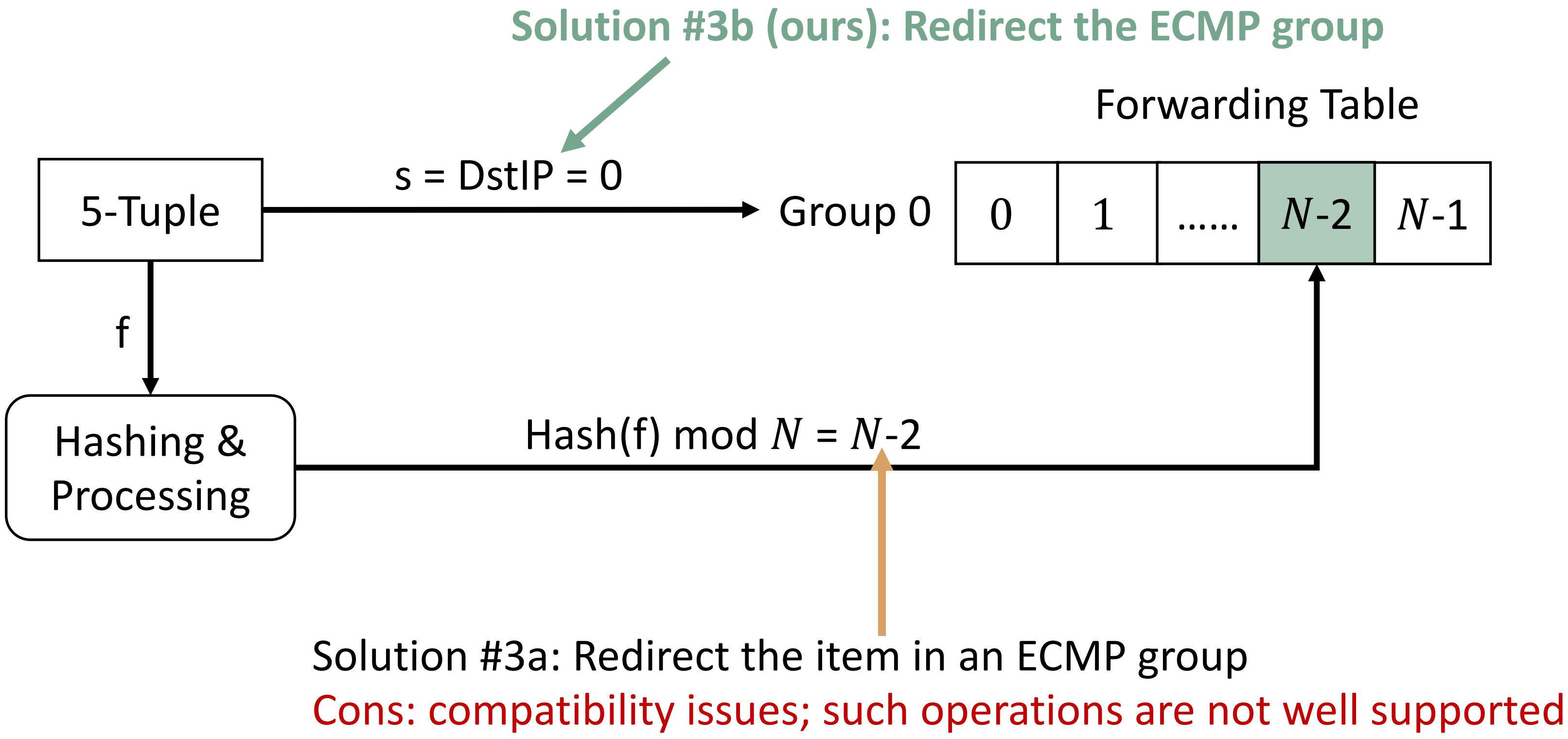
Solution #3: Modify the mapping stage

[1] Zhang et al., “Hashing linearity enables relative path control in data centers”, USENIX ATC’21.

System Design

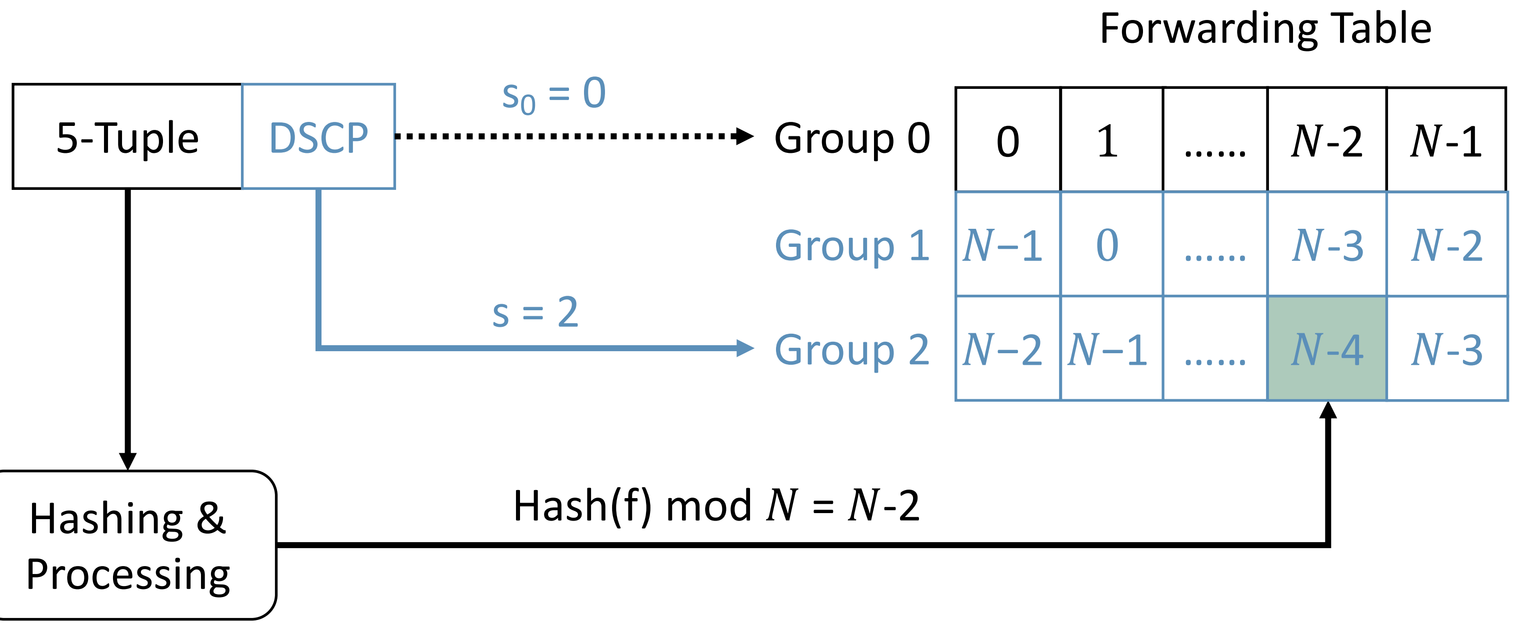
ECMP Mapping Stage Details

- f → flow's 5 tuple, s → group selector, C → Forwarding table (control matrix)
- Output port p = C[s, Hash(f)]



Programmable ECMP (P-ECMP)

- Use an additional field as the group selector *s*



Programming models and use cases (I)

- PTC of path offset:
  - deterministically deviates the selected next hop/path
  - $C[s, Hash(f)] - C[s_0, Hash(f)] = s - s_0$
- Use cases:
  - Network failover
  - Load balancing
  - Multi-path protocols
  - ……

Forwarding Table

$s_0 = 0$	Group 0	0	1	……	N-2	N-1
	Group 1	$N-1$	0	……	N-3	N-2
s	Group 2	$N-2$	$N-1$	……	N-4	N-3
				……		
	Group $N-1$	1	2	……	N-1	0

Programming models and use cases (II)

- PTC of the exact next hops:
  - C[s_i, Hash(f)] = i (i^{th} next hop/path)
- Use cases:
  - Network failure localization
  - Packet spraying
  - Segment routing
  - ……

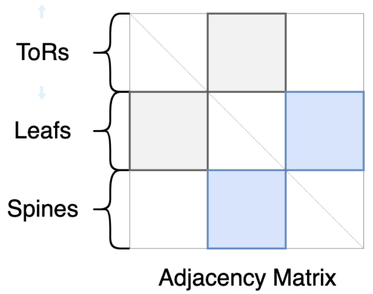
Forwarding Table

s_i →

Group 0	0	1	……	N-2	N-1
Group 1	0	0	……	0	0
Group 2	1	1	……	1	1
			……		
Group M	M	M	……	M	M

Two models can co-exist by concatenating the groups

P-ECMP compilation



Topology

PTC type

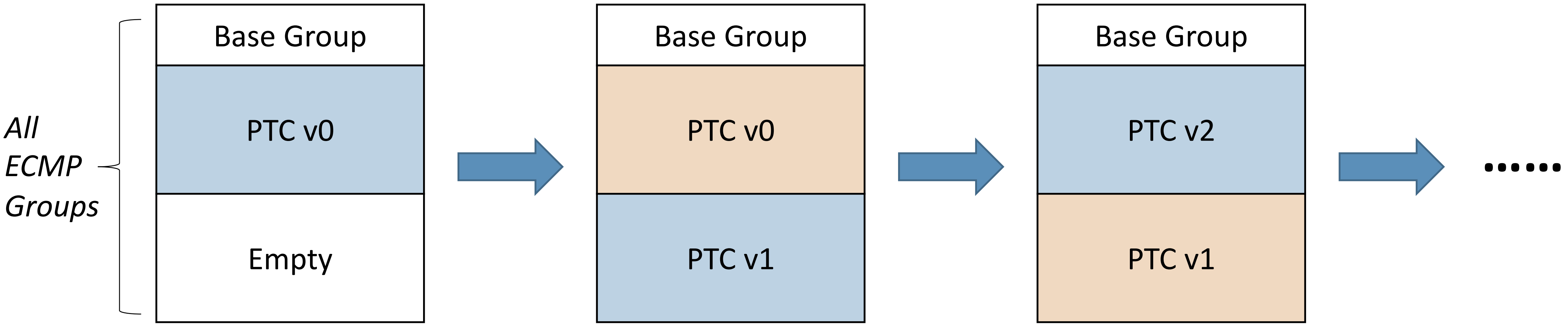
Constraints
(SRAM size)

Forwarding Table Plan

Group 0	0	1	……	N-2	N-1
Group 1	N−1	0	……	N-3	N-2
Group 2	N−2	N−1	……	N-4	N-3
			……		

Forwarding table update

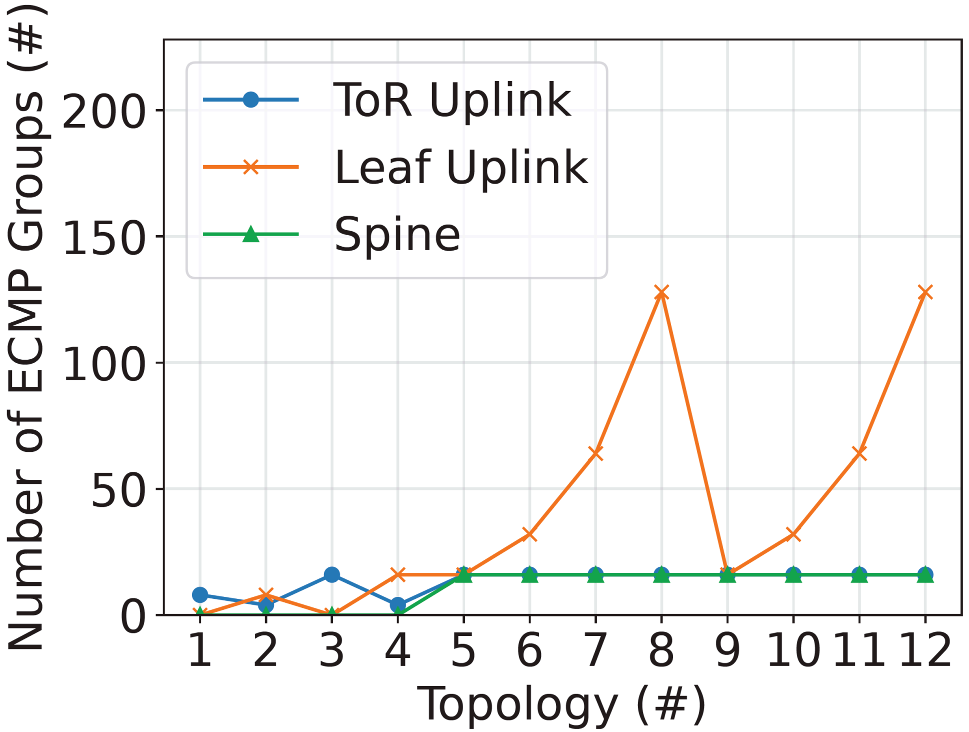
- Keeping the consistency when updating the forwarding table is hard
  - All end servers need to be updated as well
  - Packets in flight need to be considered
- Our solution: transactional update



Evaluation

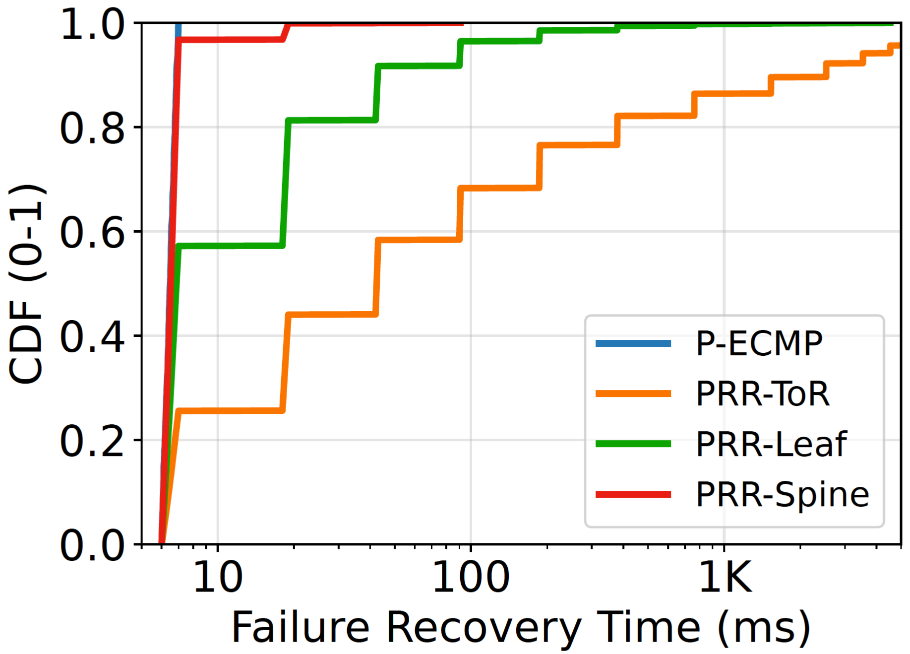
ECMP group consumption

- P-ECMP occupies up to 128 groups out of 8K available (<1.5%)



PTC scenario: network failover

- PRR[2] proposes to use TCP RTO as the indicator for network failure
- P-ECMP outperforms PRR by guaranteeing successful failover on the first attempt



[2] Wetherall et al., “Improving network availability with protective reroute”, SIGCOMM’23.

PTC scenario: load balancing

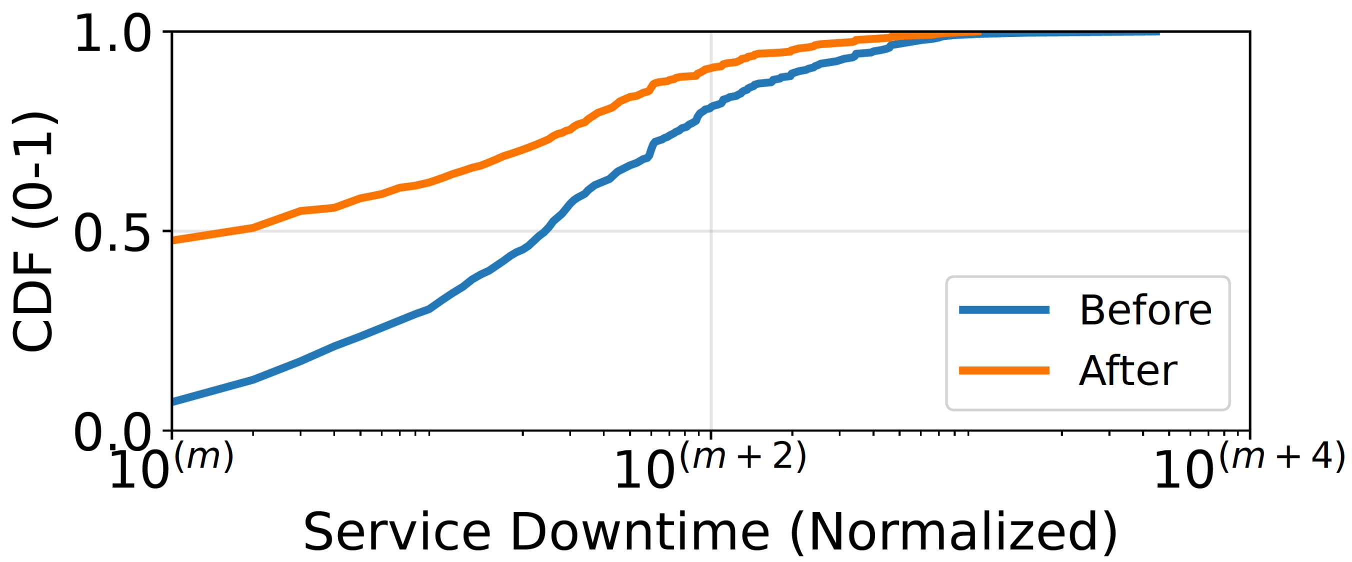
- PLB[3] proposes to re-path flows randomly upon detecting congestion
- P-ECMP effectively improves the load balancing compared to random re-pathing

Load balancing scheme	Tail latency (30% load)	Tail latency (50% load)	Tail latency (80% load)
PLB	33.0	214.3	366.7
P-ECMP	**13.6**	**182.4**	**189.0**

[3] Qureshi et al., "PLB: congestion signals are simple and effective for network load balancing", SIGCOMM'22.

Production deployment

- P-ECMP has been deployed in Tencent Cloud for supporting fast network failover



Conclusion

- Randomness and Precise Traffic Control are both important in data center networking
- We propose Programmable ECMP (P-ECMP) that leverages extra ECMP group to realize PTC
- P-ECMP enables many critical scenarios such as network failover and load balancing
- P-ECMP has been deployed in production network