中国科学技术大学
University of Science and Technology of China

APNet 2026
SINGAPORE

SprayCast: Congestion-Adaptive Native Multicast for Dynamic Sparse All-to-All Communication

Yingying Zeng[1], Ning Xu[1], Xiaobin Tan*[1,2], Shenzhi Yuan[1], XuanZhao[1], Feng Yang [1,2]

[1]University of Science and Technology of China
[2]Institute of Artiffcial Intelligence, Hefei Comprehensive National Science Center

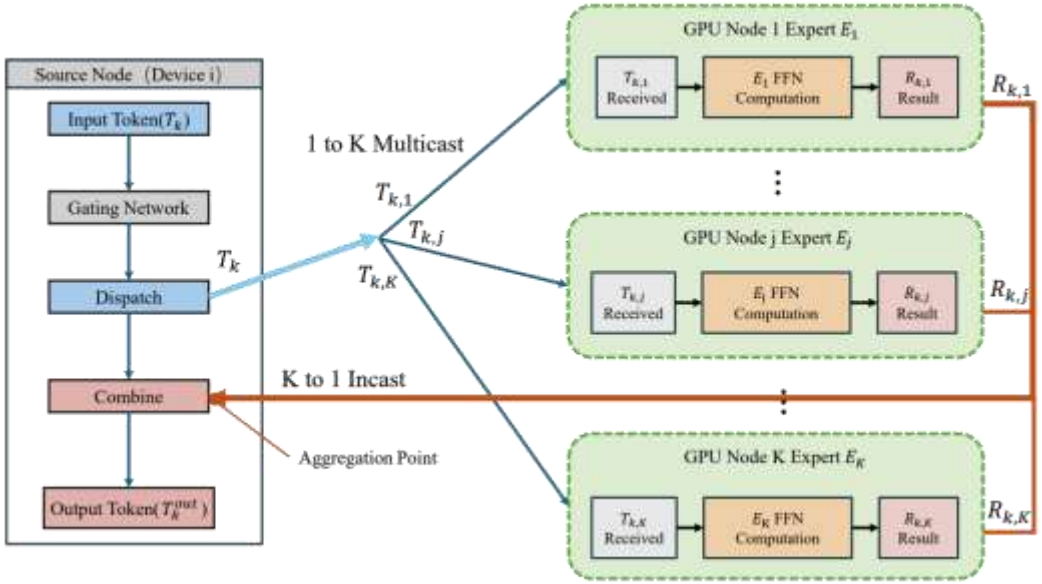
August 2026

01 Background

MoE All-to-All Communication

- Model parameter sizes are growing exponentially, driving the wide adoption of MoE for efficient scaling.

Model capacity is scaling rapidly, and MoE has become a key approach for efficient scaling.



1.MoE Routing

Gating selects top-K experts for each token.

2.Dispatch

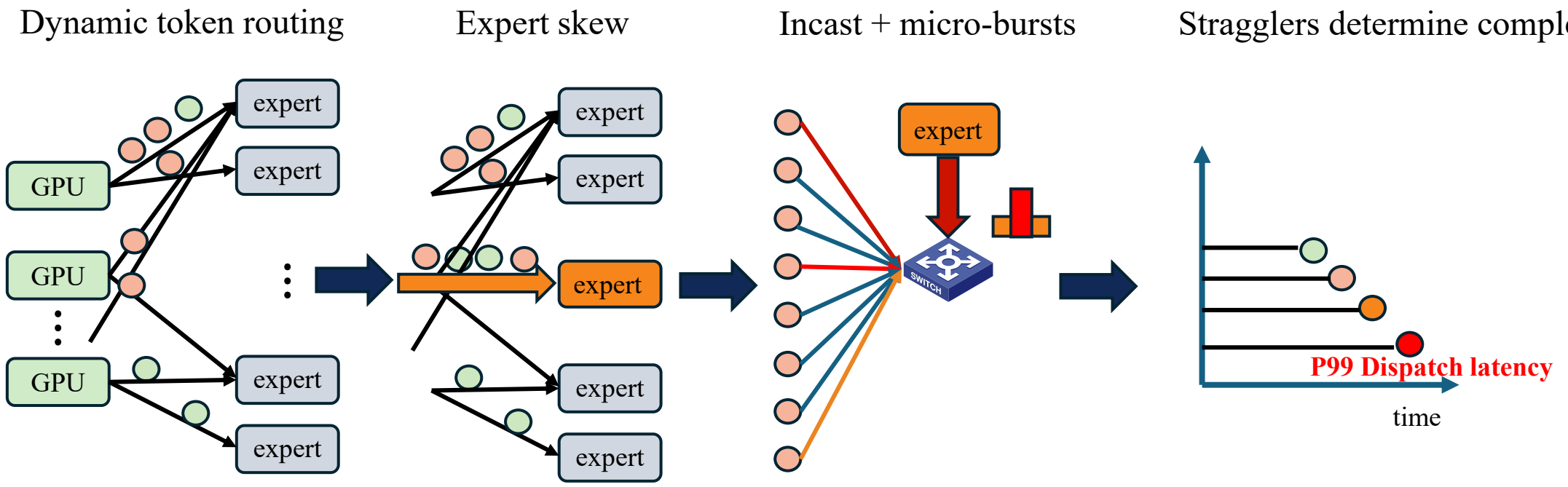
Tokens are sent to selected experts, forming dynamic 1-to-K micro-multicast traffic.

3.Combine

Expert results return to the source, creating K-to-1 incast.

01 Background

The Hazards of Dynamic Sparse Dispatch



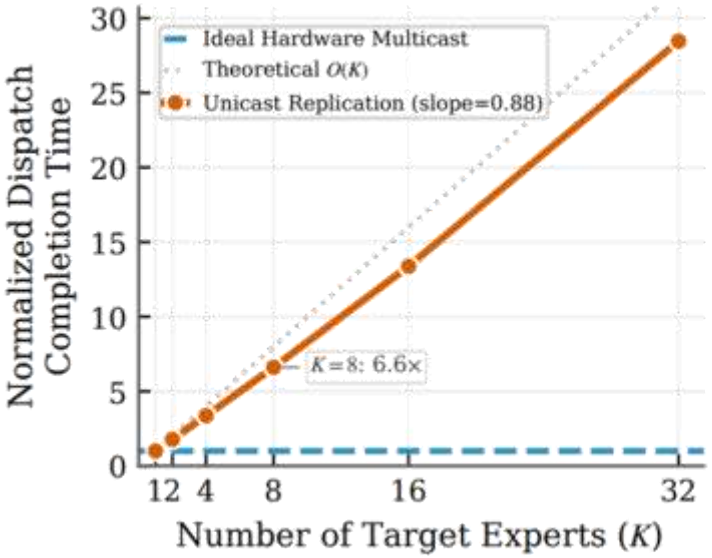
- Expert skew creates micro-bursts and incast
- MoE rounds wait for the slowest receiver
- P99 Dispatch latency becomes the bottleneck

02 Motivation

Problem 1

Problem 1: Unicast replication does not scale.

- Sender injects K copies for each token.
- Dispatch time grows almost linearly with K.
- At K = 8, unicast is 6.6× slower than ideal in-network multicast.



Unicast replication turns Top-K Dispatch into O(K) sender-side injection.

02 Motivation

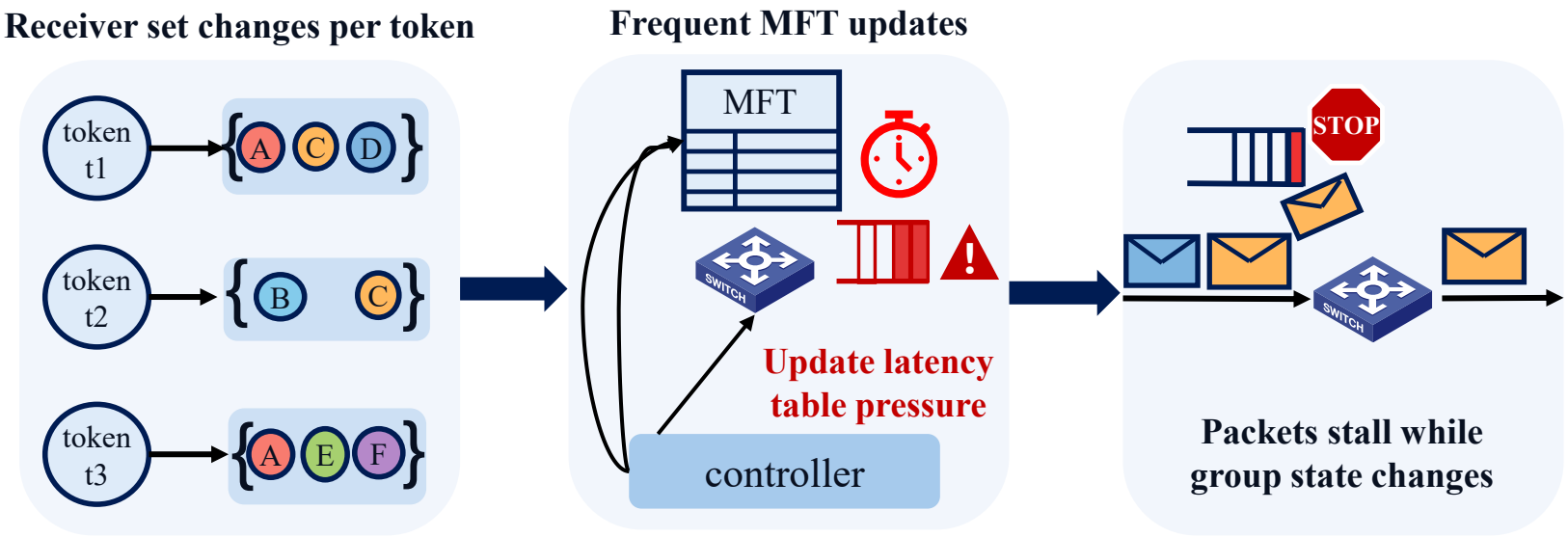
Problem 2

Problem 1: Unicast replication does not scale.

Problem 2: Table-driven multicast cannot track receiver-set churn.

- Each token may have a different receiver set.
- MFT entries must be updated repeatedly.
- Update latency and table pressure stall Dispatch.

Short-lived multicast groups push MFT updates onto the critical path.



02 Motivation

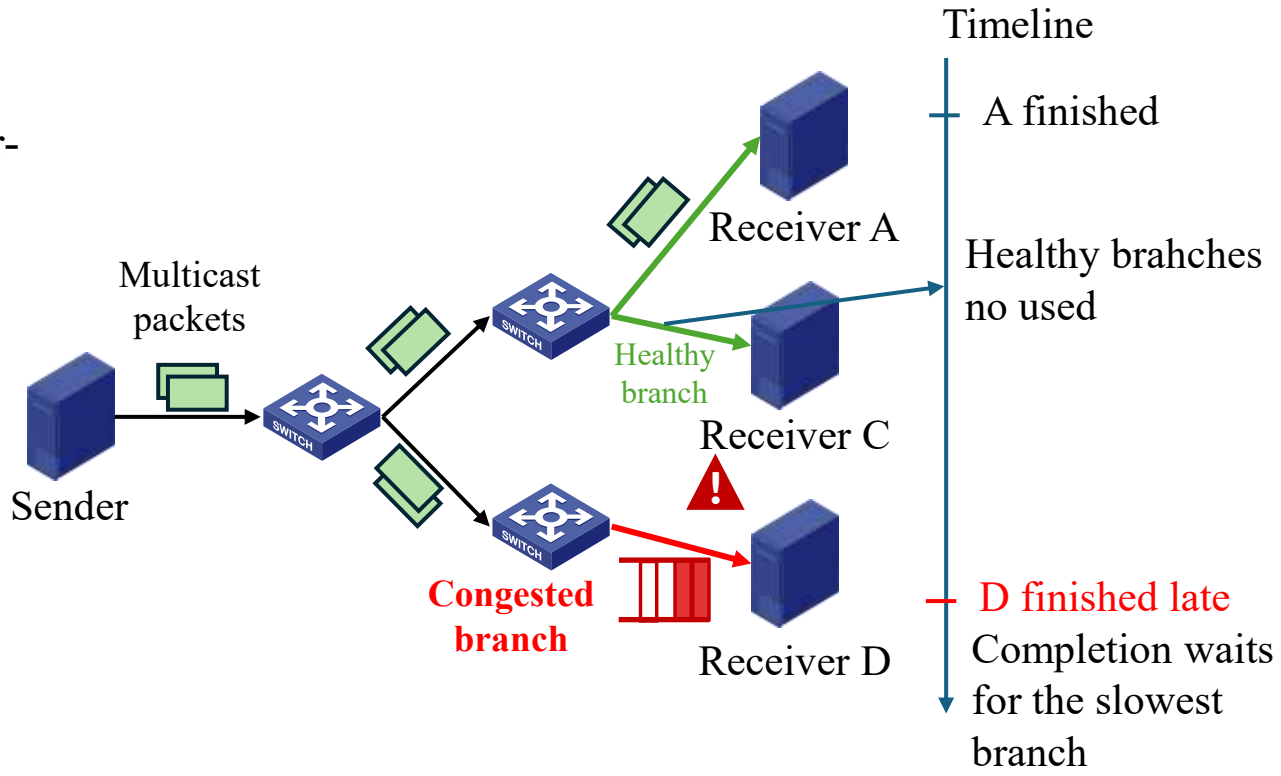
Problem 3

Problem 1: Unicast replication does not scale.

Problem 2: Table-driven multicast cannot track receiver-set churn.

Problem 3: Fixed multicast trees amplify stragglers.

- Skewed experts create transient hotspots.
- A fixed tree cannot steer around congested branches.
- One slow branch dominates Dispatch completion.



Fixed trees cannot adapt to transient hotspots; one congested branch dominates Dispatch completion.

02 Motivation

Problem 3

Problem 1: Unicast replication does not scale.
- Sender injects K copies for each token.

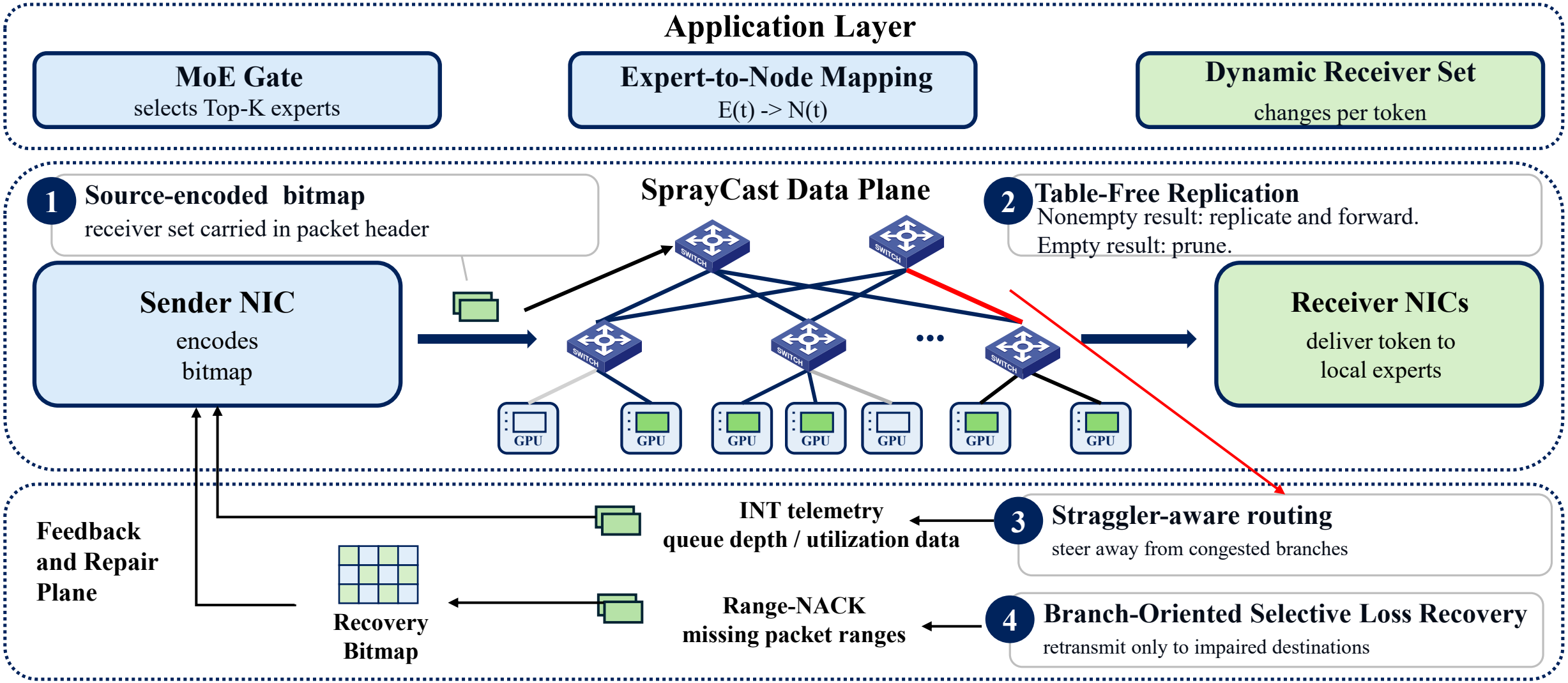
Problem 2: Table-driven multicast cannot track receiver-set churn.
- MFT updates stall short-lived multicast groups.

Problem 3: Fixed multicast trees amplify stragglers.
- One congested branch dominates Dispatch completion.

Limitation: Group state, congestion feedback, and recovery are too coarse-grained.

03 Design

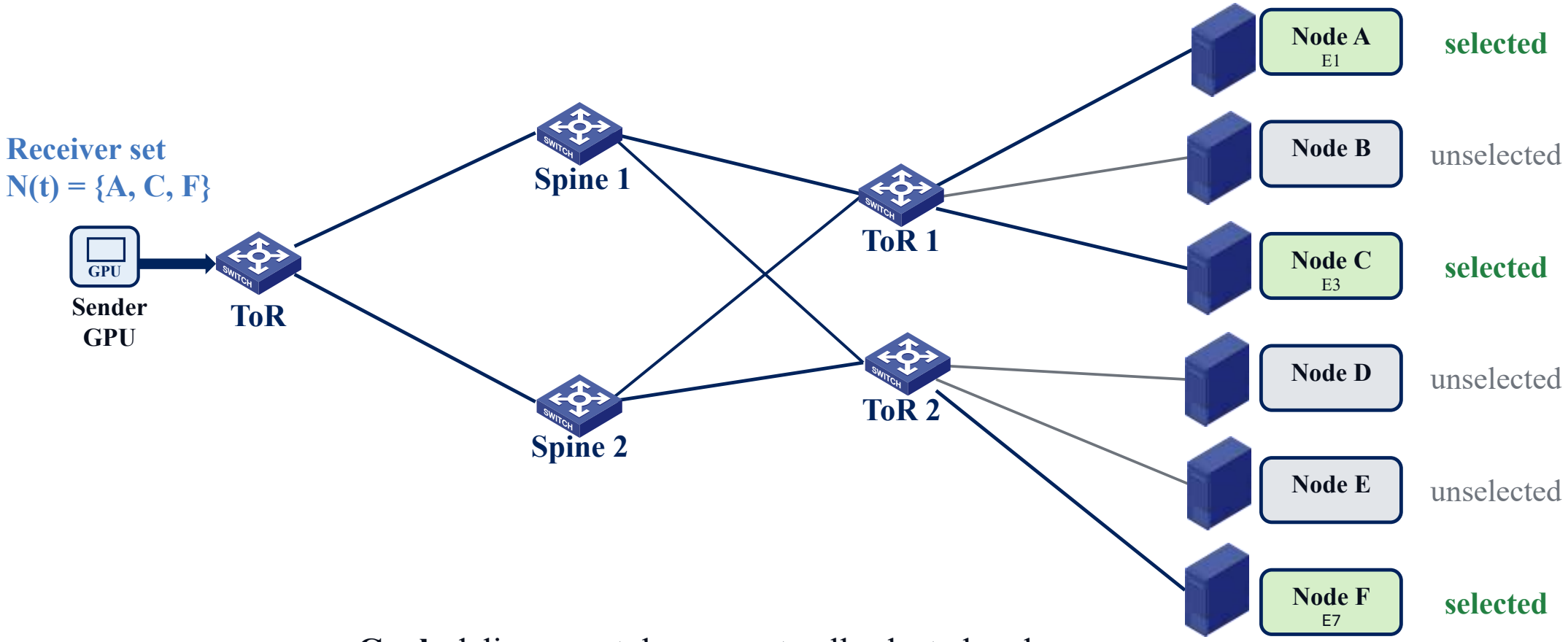
Workflow



03 Design

Example: Dynamic Sparse Dispatch

Token t selects experts on three destination nodes.

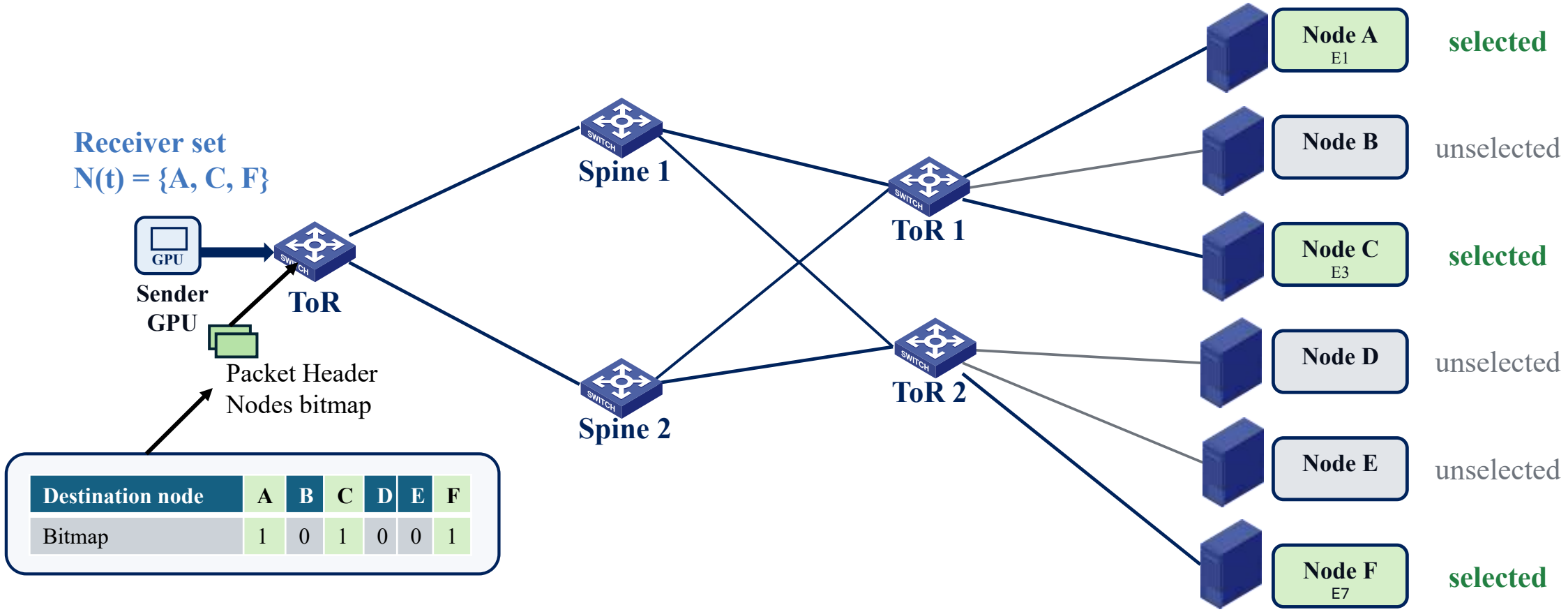


Goal: deliver one token copy to all selected nodes.

03 Design

Step1: Source-Encoded Bitmap

The sender sets one bit for each destination node and carries the bitmap in the packet header.



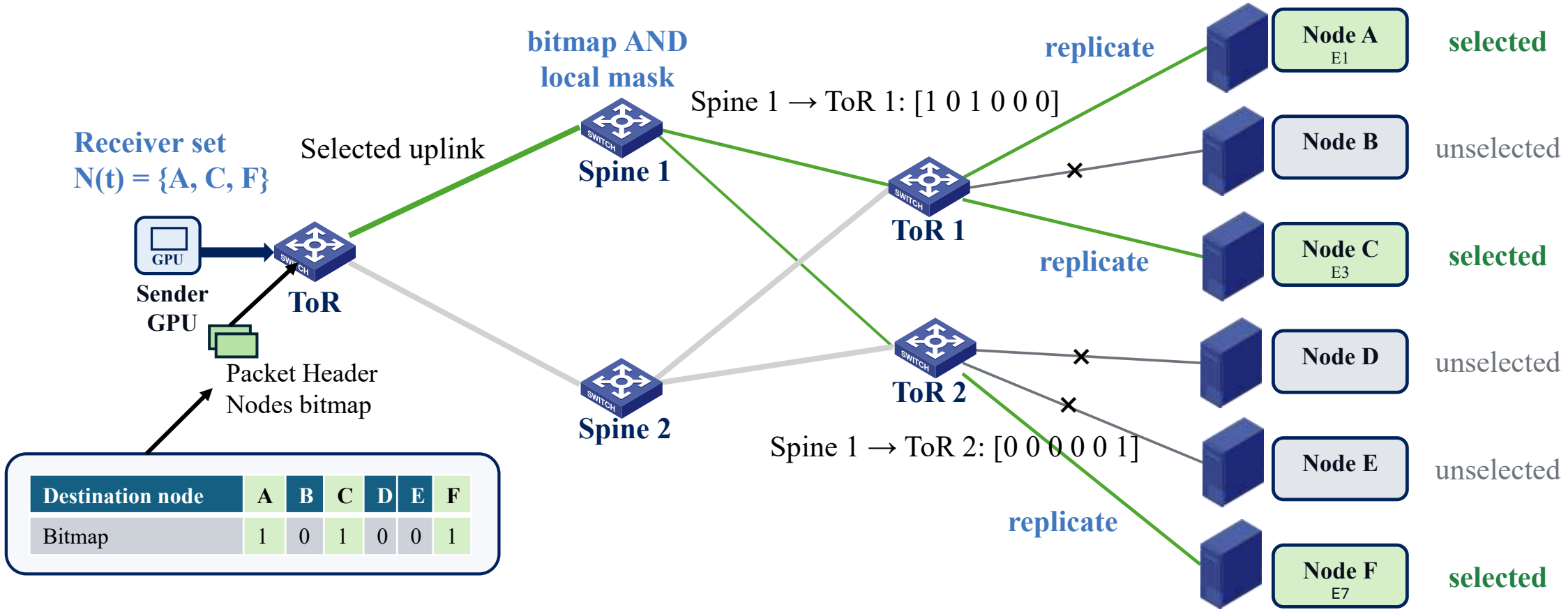
Destination node	A	B	C	D	E	F
Bitmap	1	0	1	0	0	1

Key idea: Dynamic receiver-set state moves from switch tables to packet headers.

03 Design

Step 2: Table-Free Replication

Each switch intersects the destination bitmap with its static per-port reachability masks.

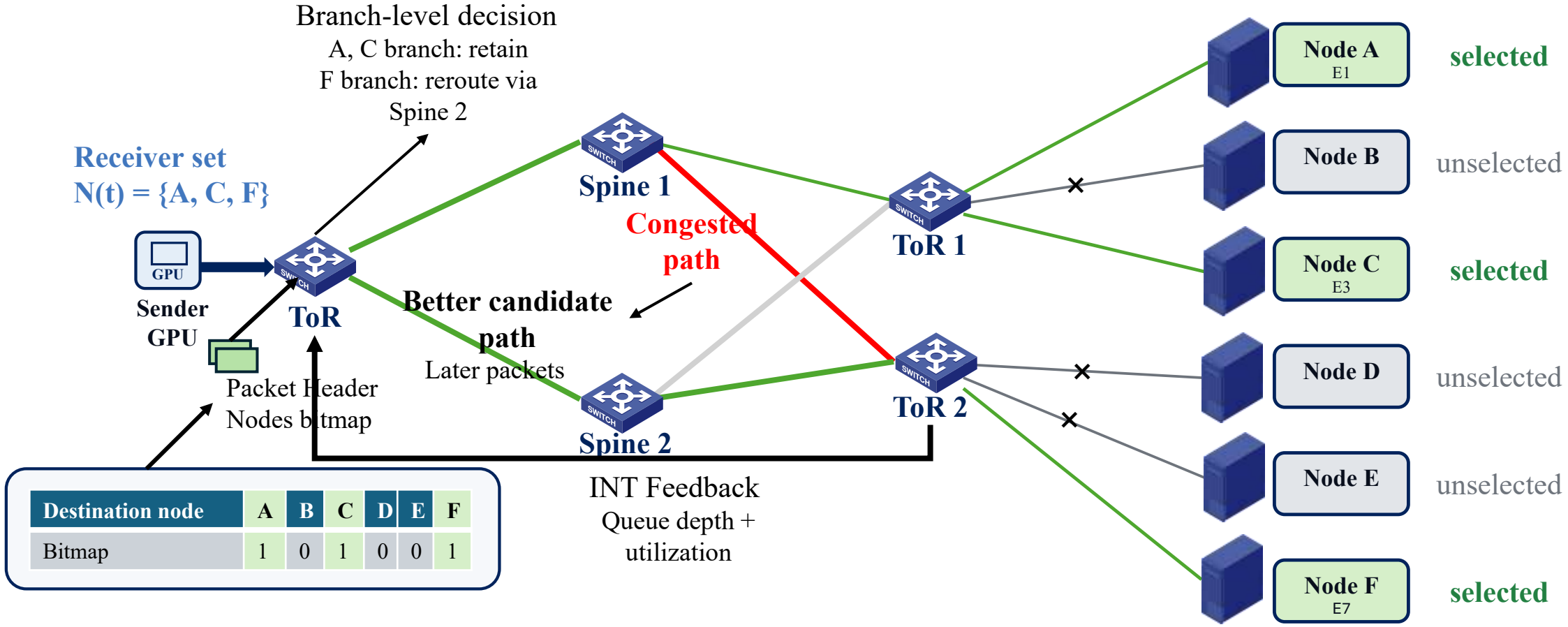


Key idea: Nonempty intersection → forward; empty intersection → prune.

03 Design

Step 3: Straggler-Aware Routing

INT reports queue depth and link utilization. The sender steers later packets to a better candidate path.

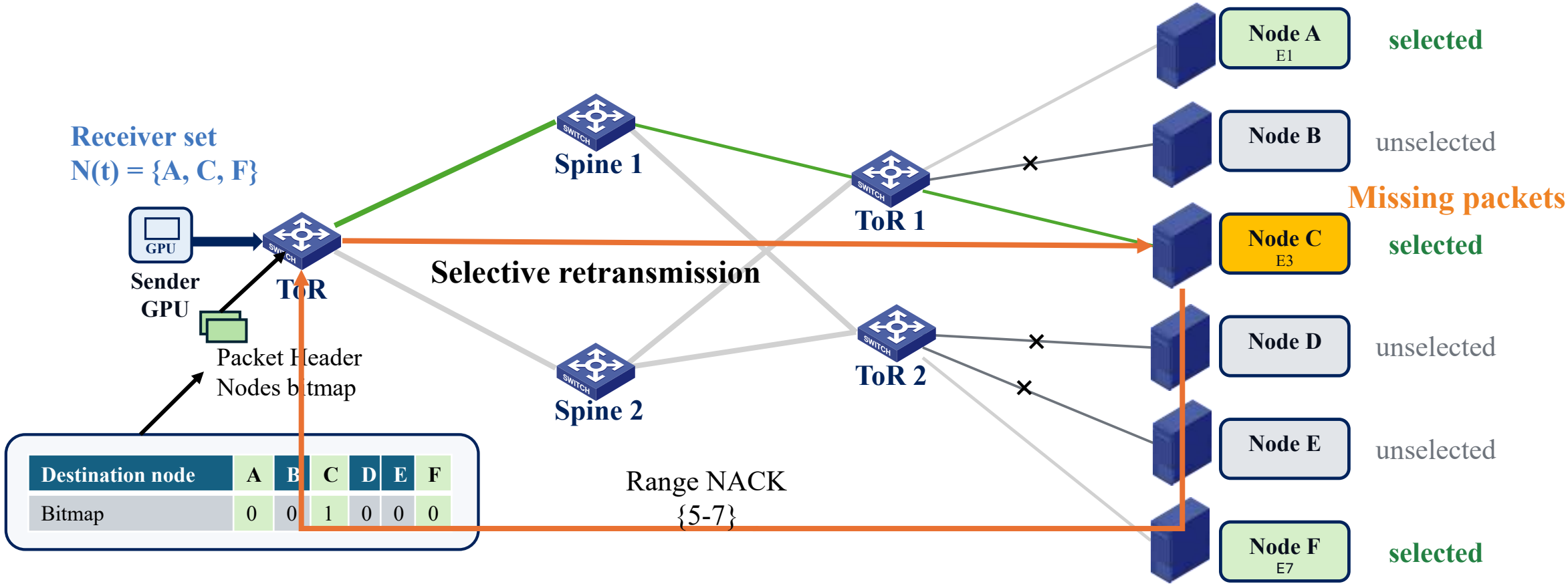


Key idea: Later packets avoid the congested branch by selecting a better candidate path.

03 Design

Step 4: Branch-Oriented Selective Loss Recovery

The receiver reports missing ranges with a Range-NACK.
The sender retransmits only to the impaired destinations.



Key idea: Selective Loss Recovery.

04 Evaluation

Experiment Setup

Simulator: htsim

- Topology: 128-server three-tier fat-tree
- Link bandwidth: 100 Gbps
- Link delay: 500 ns

Baselines:

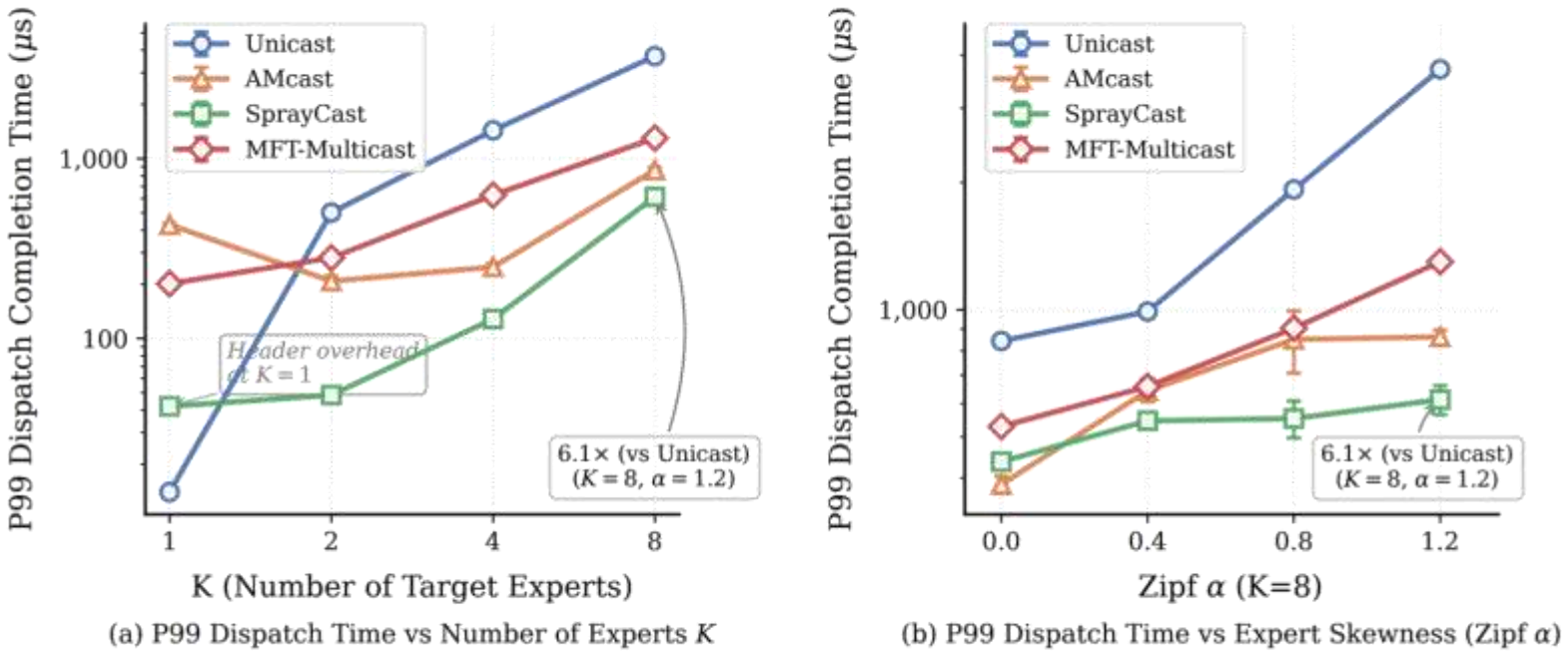
- Unicast: K point-to-point copies
- AMcast: host-based relay multicast
- MFT-Multicast: table-driven multicast

Primary Metric:

P99 Dispatch Completion Time per MoE round

04 Evaluation

End-to-End P99 Dispatch Latency



SprayCast reduces P99 Dispatch latency by 6.1× over Unicast at K = 8 and α = 1.2.

05 Summary

Conclusion

Contribution:

- Packet-carried bitmaps enable native multicast without K-way unicast replication or per-group MFT updates.
- Worst-branch-aware routing steers traffic away from hotspots.
- Range-NACKs selectively repair only impaired destinations.
- SprayCast reduces P99 Dispatch latency by up to 6.1× at K = 8.

Future work:

- Scale beyond 10K nodes through bitmap compression and segmented delivery.
- Use SRv6 or direct source routing for finer-grained path control.
- Prototype recovery on programmable NICs/DPUs.

Thanks!