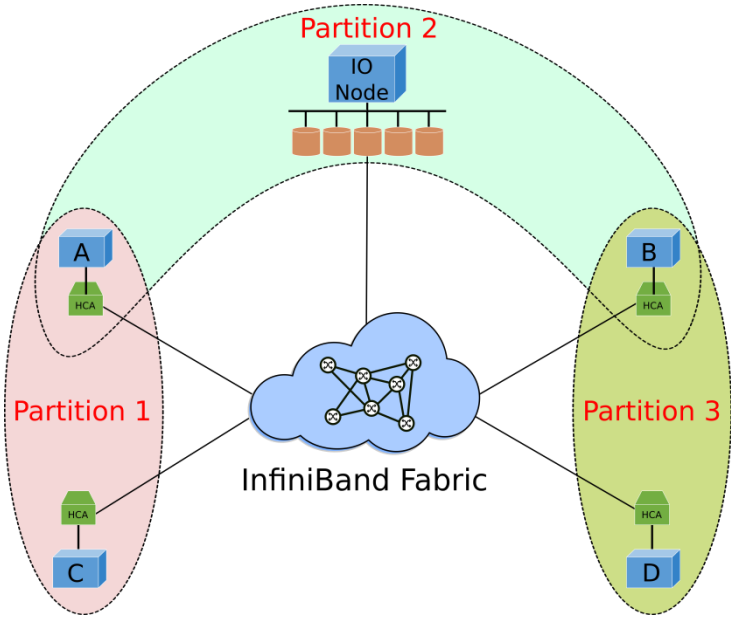
Partition-aware routing to improve network isolation in InfiniBand based multi-tenant clusters

Feroz Zahid, Ernst Gunnar Gran, Tor Skeie
Simula Research Laboratory, Norway

Bartosz Bogdanksi, Bjørn Dag Johnsen
Oracle Corporation



IEEE/ACM CCGrid 2015

Shenzhen, Guangdong, China

May 6, 2015

[simula . research laboratory]

This presentation will walk through the paper discussing three important sections

Background and Problem Statement

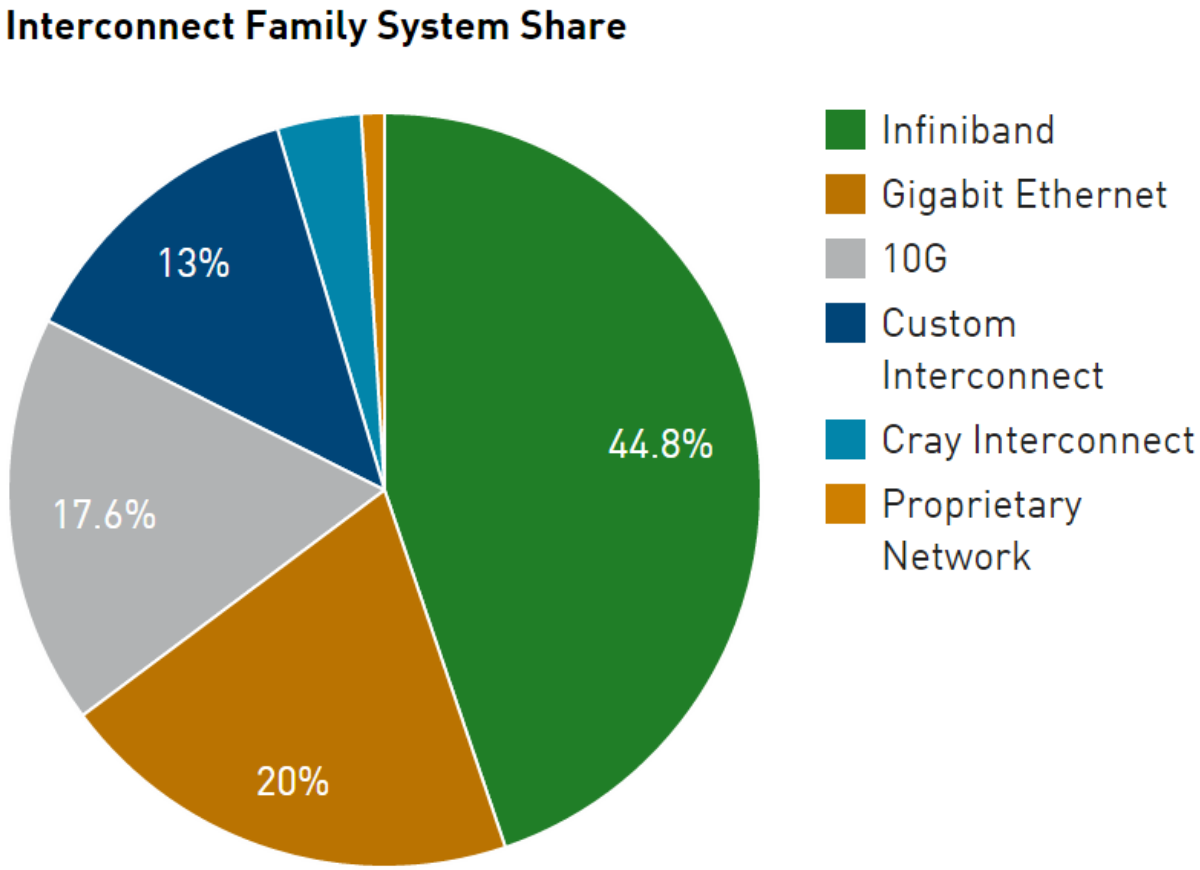
Partition-aware Routing

Evaluation

InfiniBand (IB) is a popular interconnect for HPC systems

TOP 500 The List.

44.8% share in November 2014 top supercomputers list

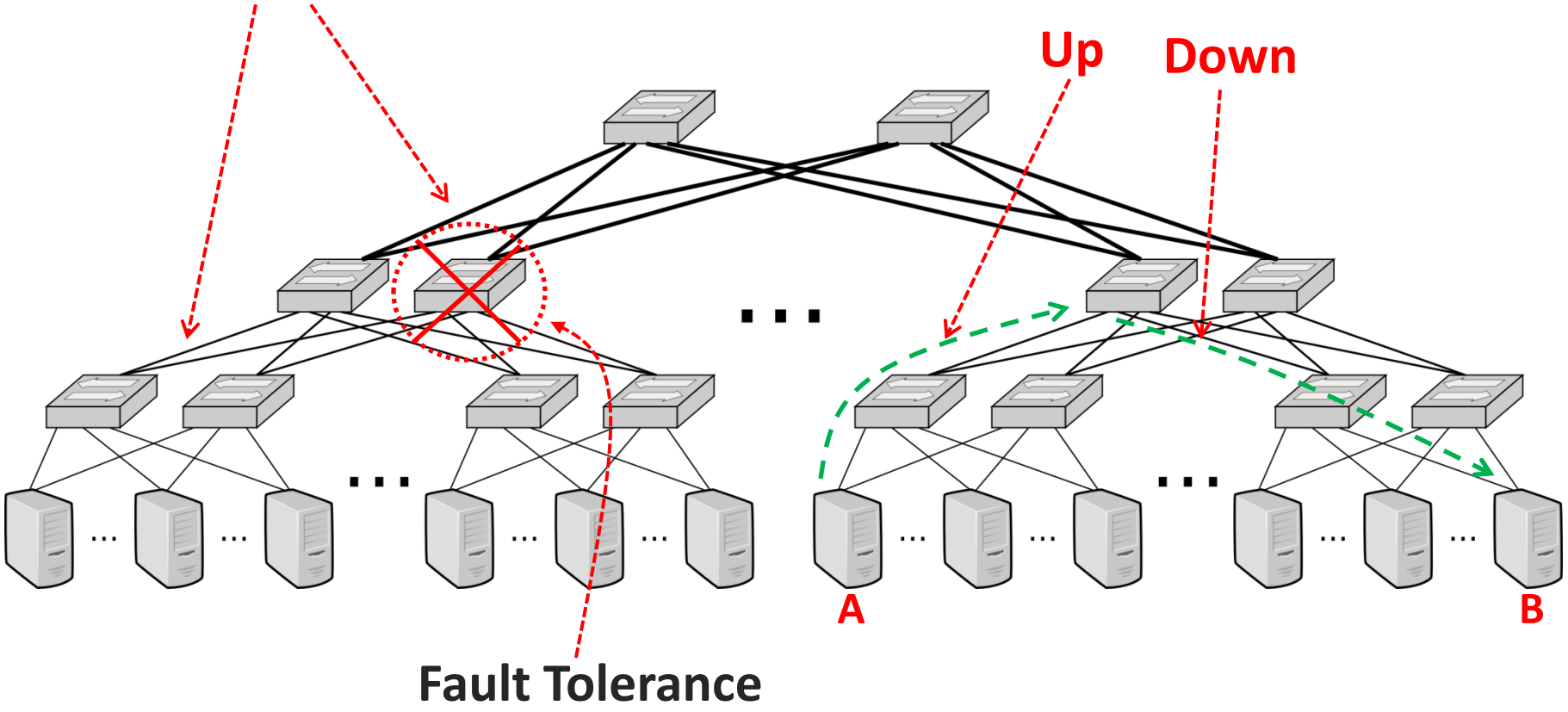


Source: Top500 Supercomputers List, http://top500.org/

Fat-trees have nice properties that make them popular

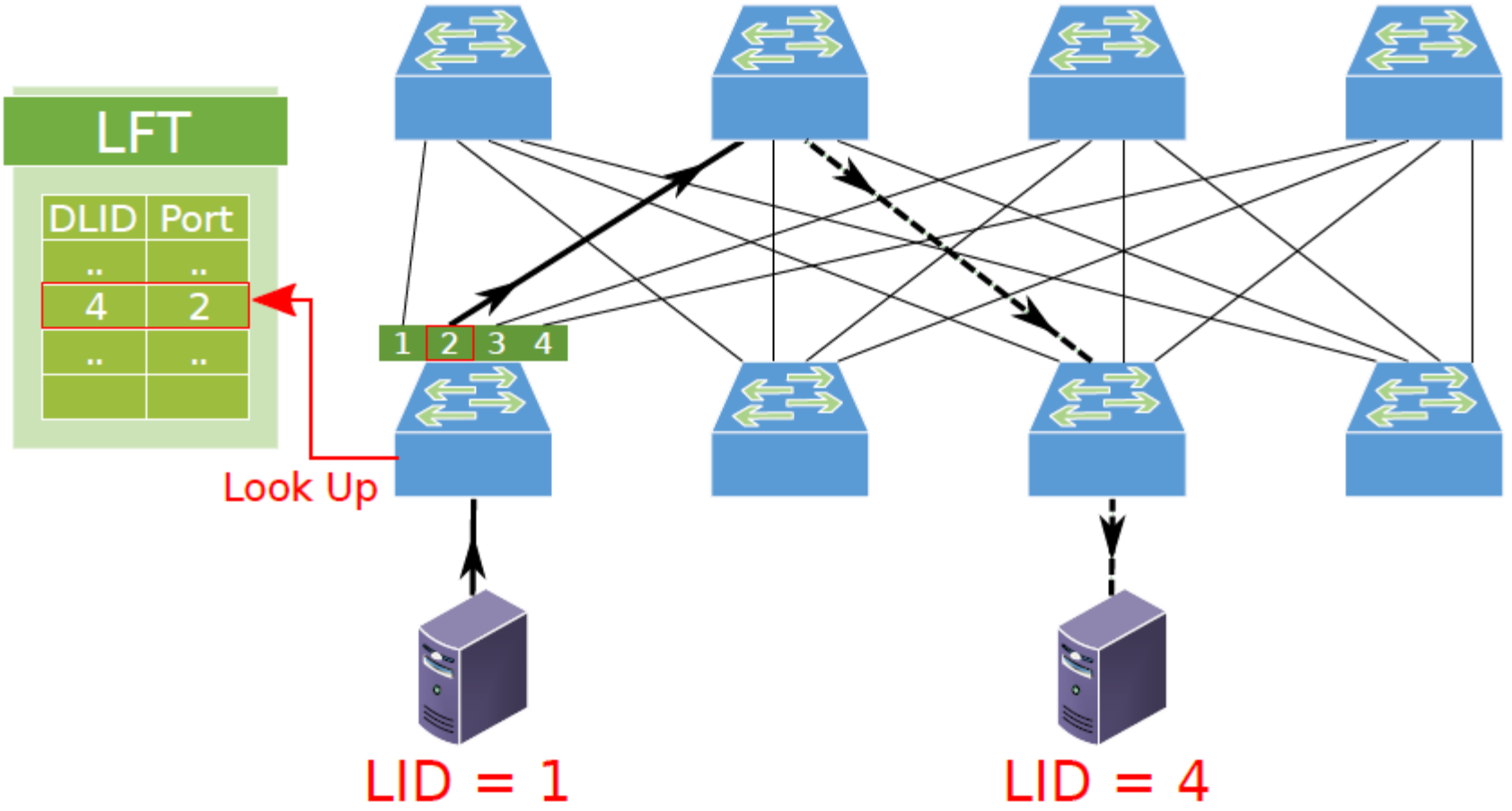
Maintenance of full-bisection bandwidth

Easy deadlock-free Routing

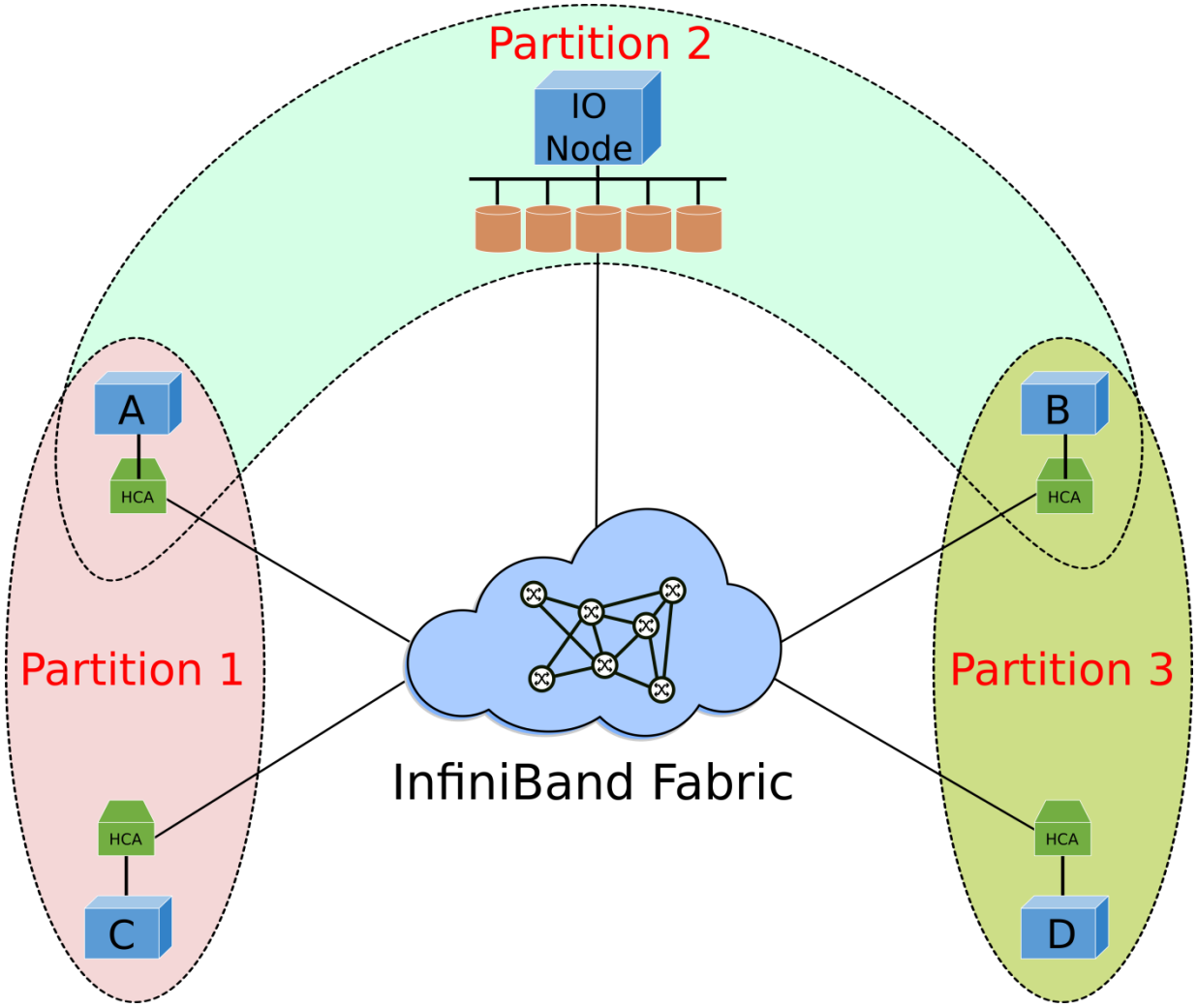


Routing in IB networks is generally deterministic

Based on linear forwarding tables (LFTs) stored in the switches



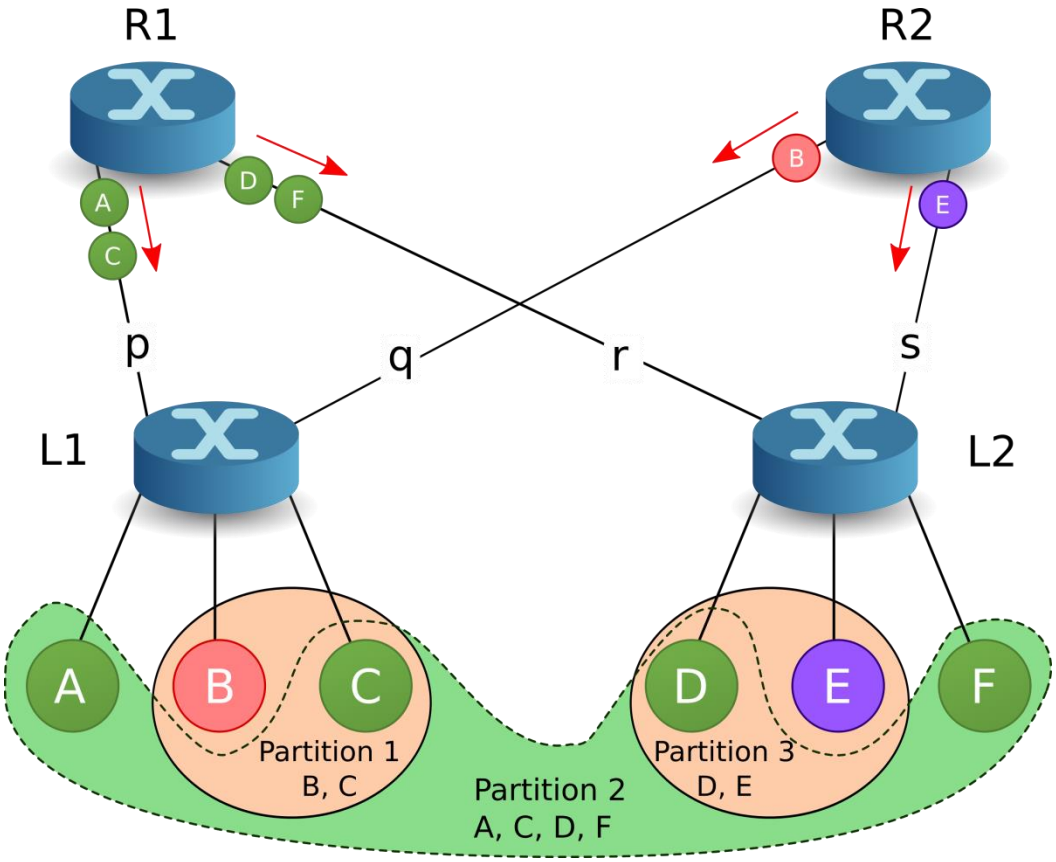
Partitioning is a security mechanism to enforce isolation of logical groups of systems sharing a network fabric



Nodes that do not share a partition are not allowed to communicate!

Routing done without considering partitions results in degraded load-balancing and performance interference

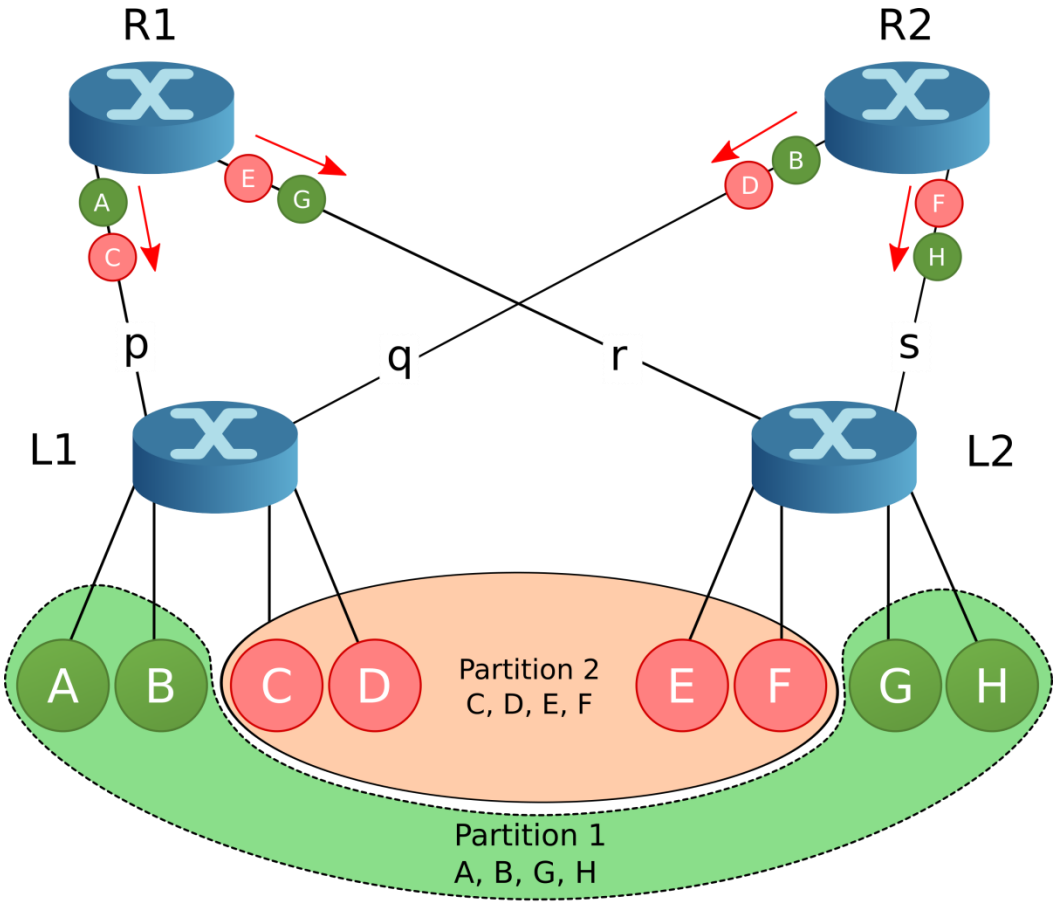
Problem 1 : Degraded Load-Balancing



p and *r* are oversubscribed, while no intra-partition flow from *q* and *s*

Routing done without considering partitions results in degraded load-balancing and performance interference

Problem 2 : Performance interference among partitions



All links are shared by the flows belonging to both Partitions 1 and 2

The partition-aware fat-tree routing algorithm (pFTree) tends to isolate partition flows without compromising on the load balancing

- The pFTree has two objectives in the order of priority
  - Well-balanced LFTs
  - Partition isolation
- Balancing
  - Using port counters
- Partition-isolation
  - Physical level, if enough resources available
  - Virtual Lanes

The algorithm is completely contained in the subnet manager

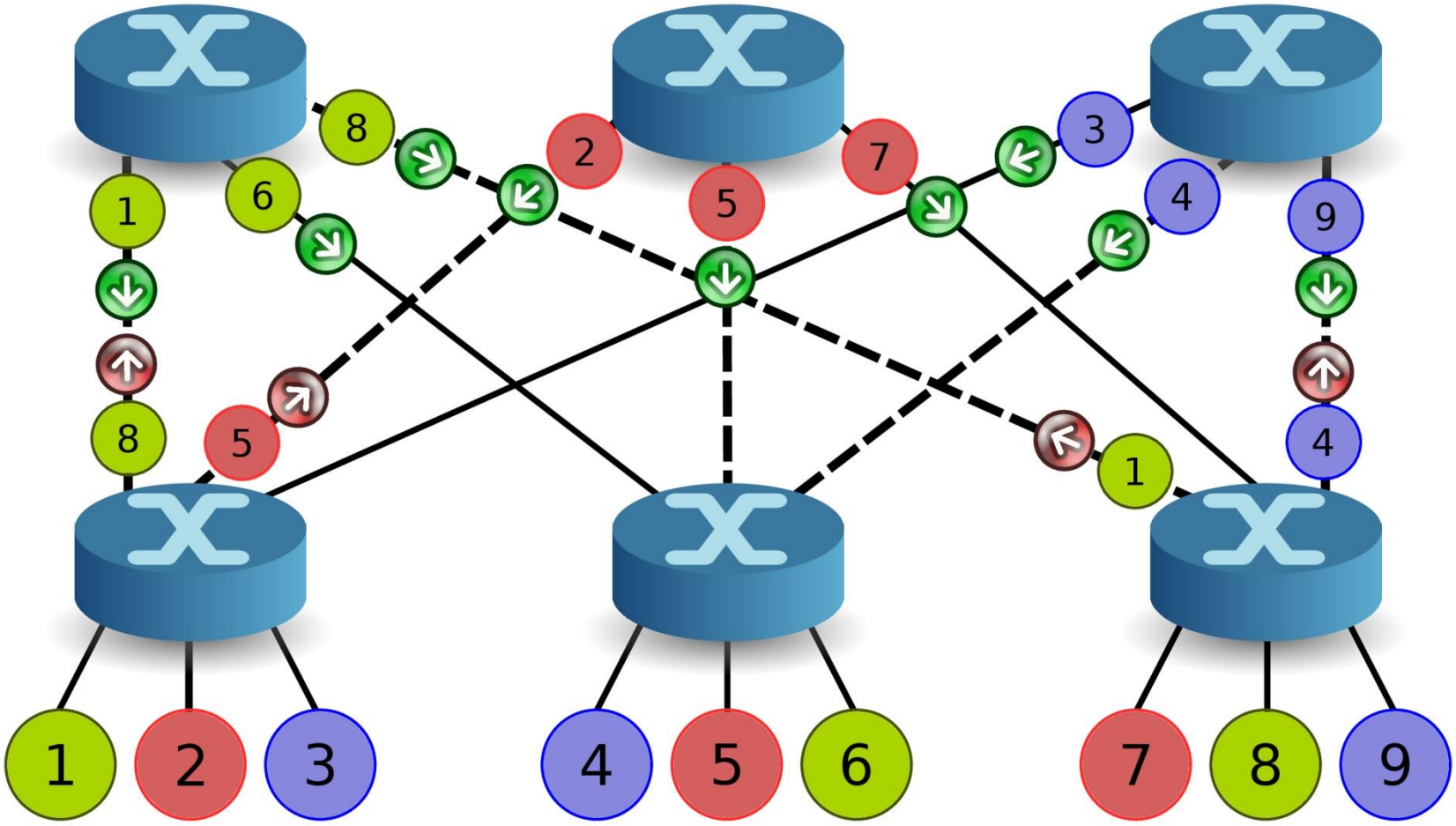
We implemented partition-aware fat-tree routing algorithm (pFTree) in the OFED's subnet manager, OpenSM, for the evaluation

OFED is the de-facto standard software stack for building and deploying IB based applications

- Deterministic
  - High-performance, Avoids out-of-order packet deliveries
- Destination-based
  - Direct realization in IB networks
- Iterative
  - Better routes balancing
- Maintains counters on ports in both DWN and UP directions
  - When a new route is added, +1
- Supports XGFTs, PGFTs, RLFTs

The pFTree routing algorithm works by marking links for the partition nodes, and selecting already marked partitions

The pFTree Routing vs Original Fat-Tree routing

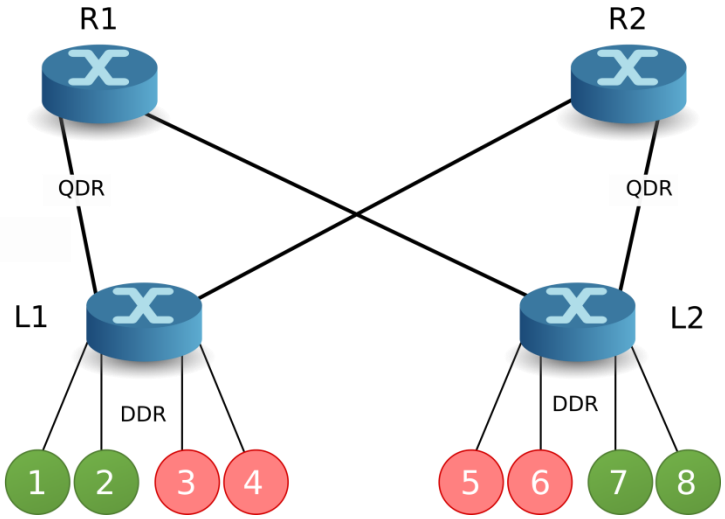


FTree Routing: 9→4, 8→1, 2→5, 1→8

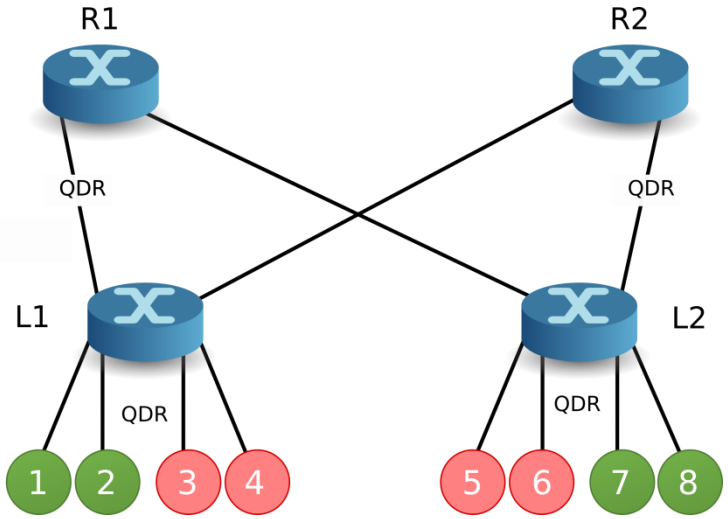
pFTree Routing: 9→4, 8→1, 2→5, 1→8

Evaluation: For real-world experiments, three topologies representing different scenarios are taken

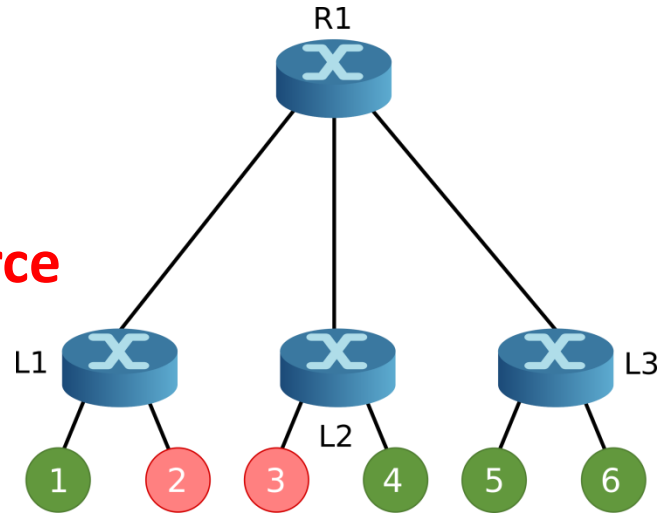
Non-oversubscribed



Oversubscribed

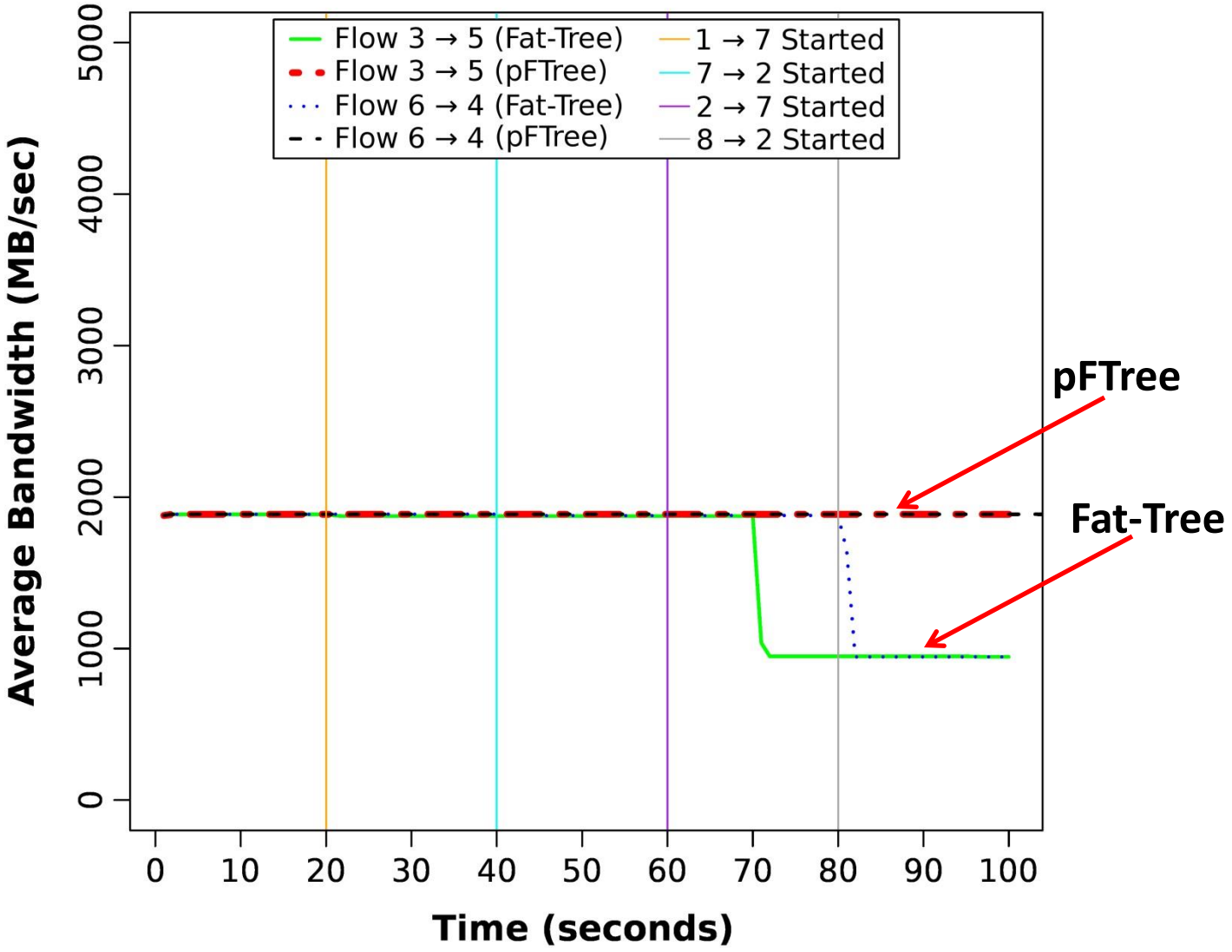


Limited-resource



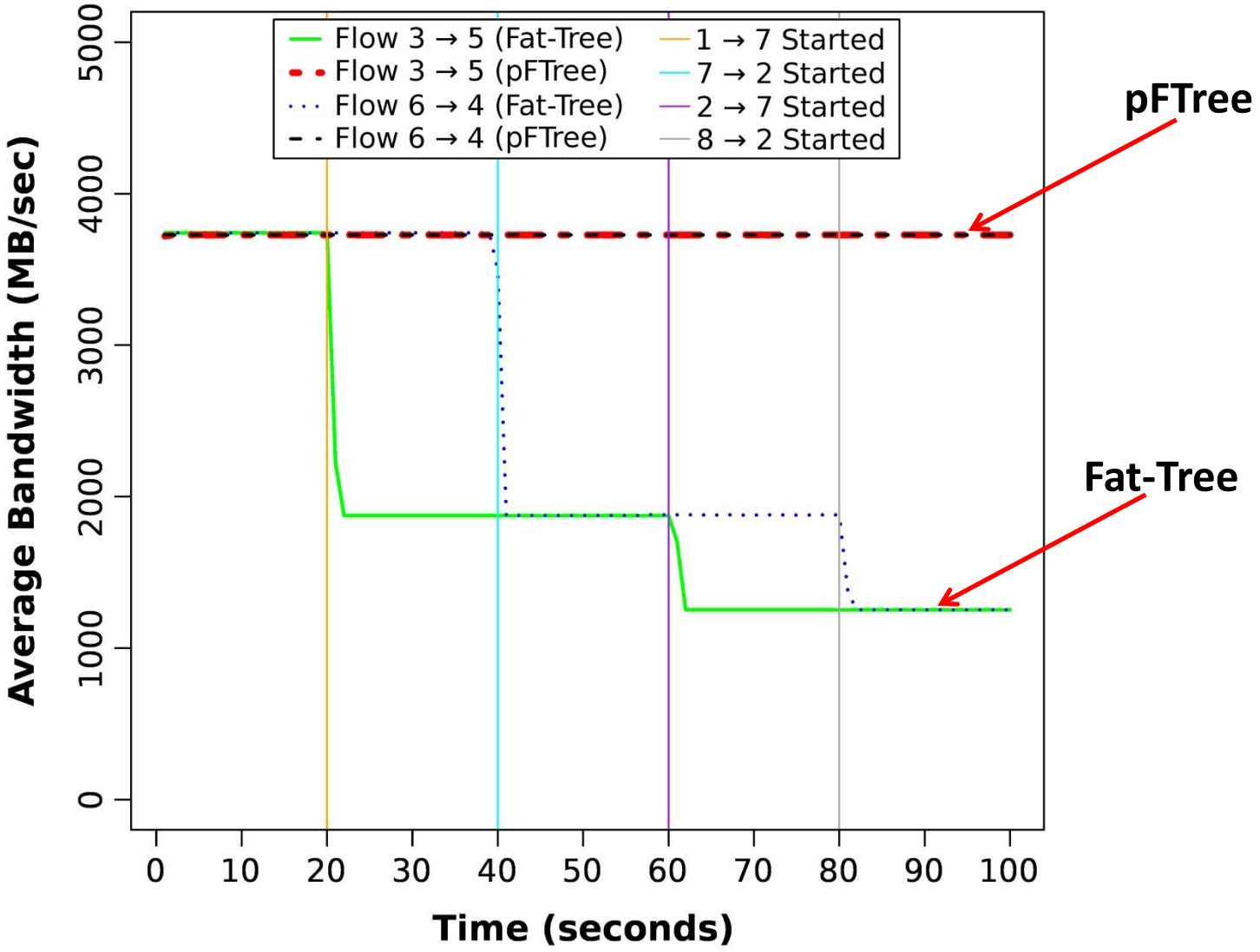
Evaluation: The effect of interference on victim partition is minimized for all three topologies

Non-oversubscribed Topology 1



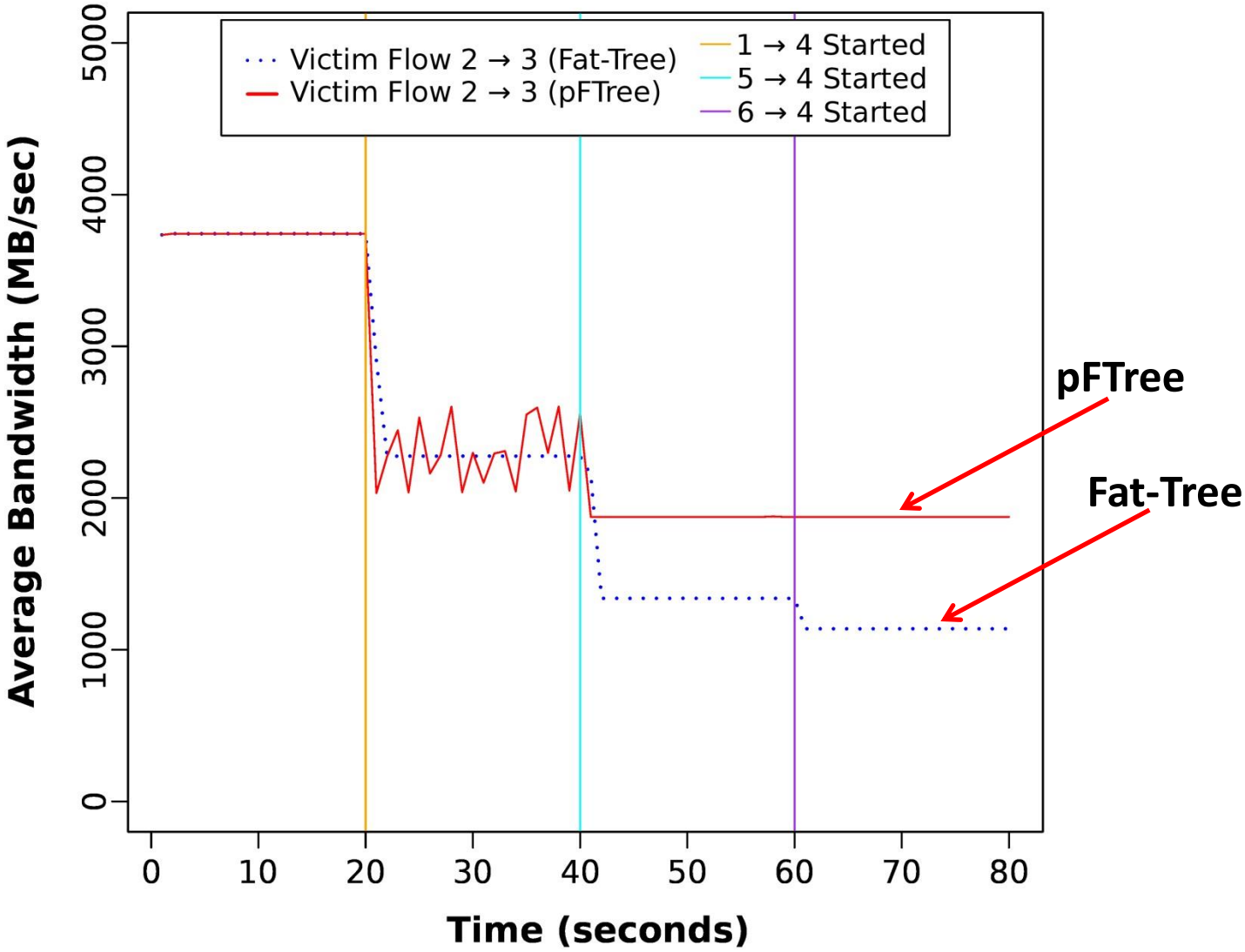
Evaluation: The effect of interference on victim partition is minimized for all three topologies

Oversubscribed Topology 2

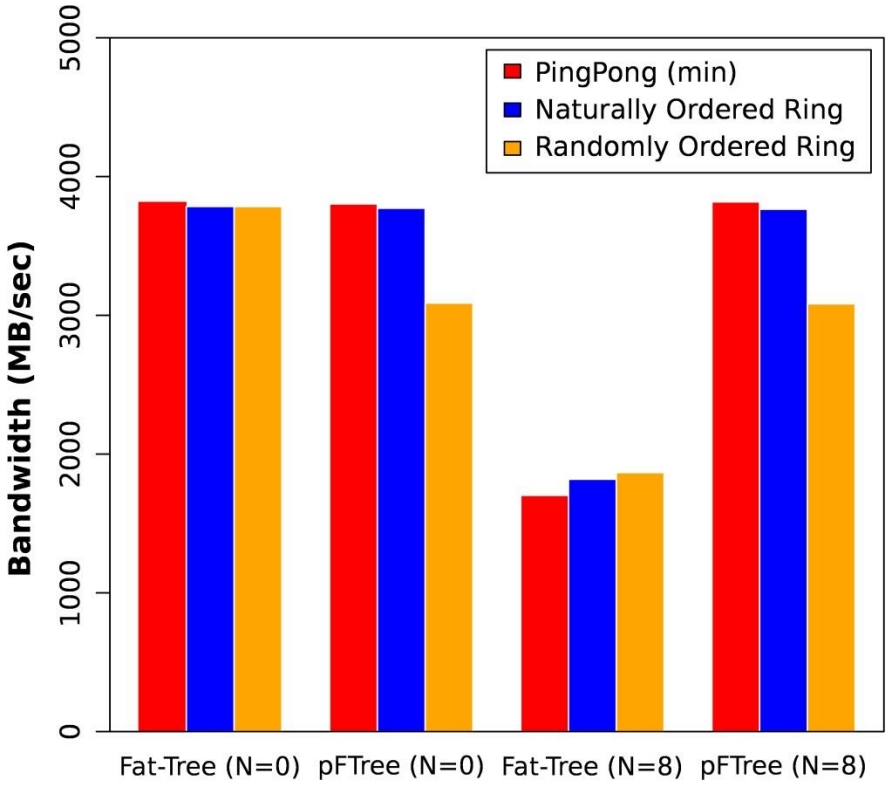


Evaluation: The effect of interference on victim partition is minimized for all three topologies

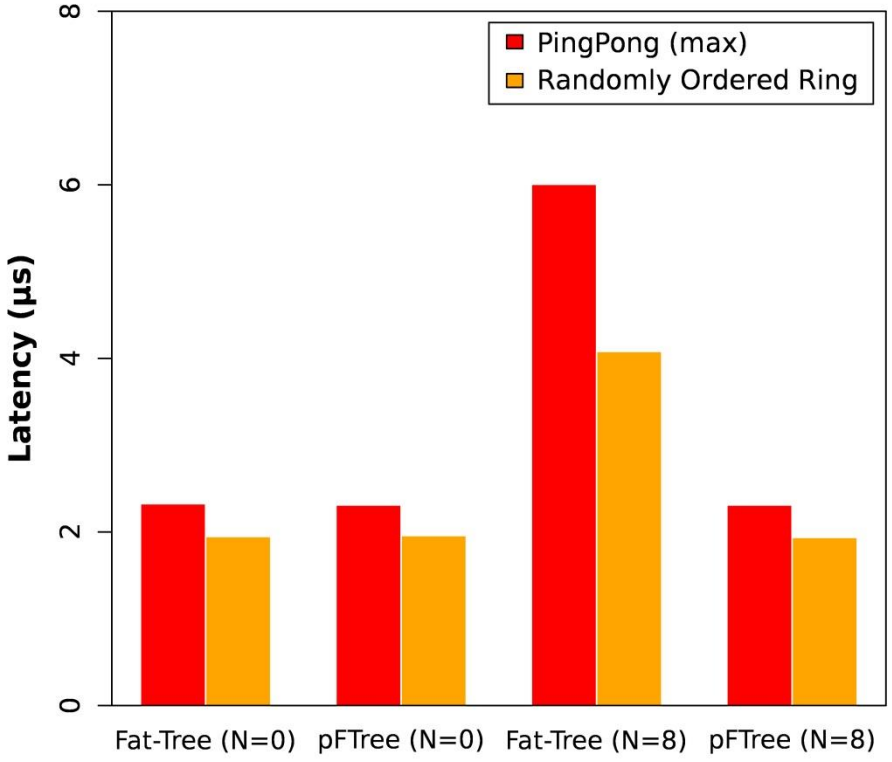
Limited-resource Topology 3



Evaluation: HPCC challenge benchmark shows 109% increase in throughput for randomly ordered ring

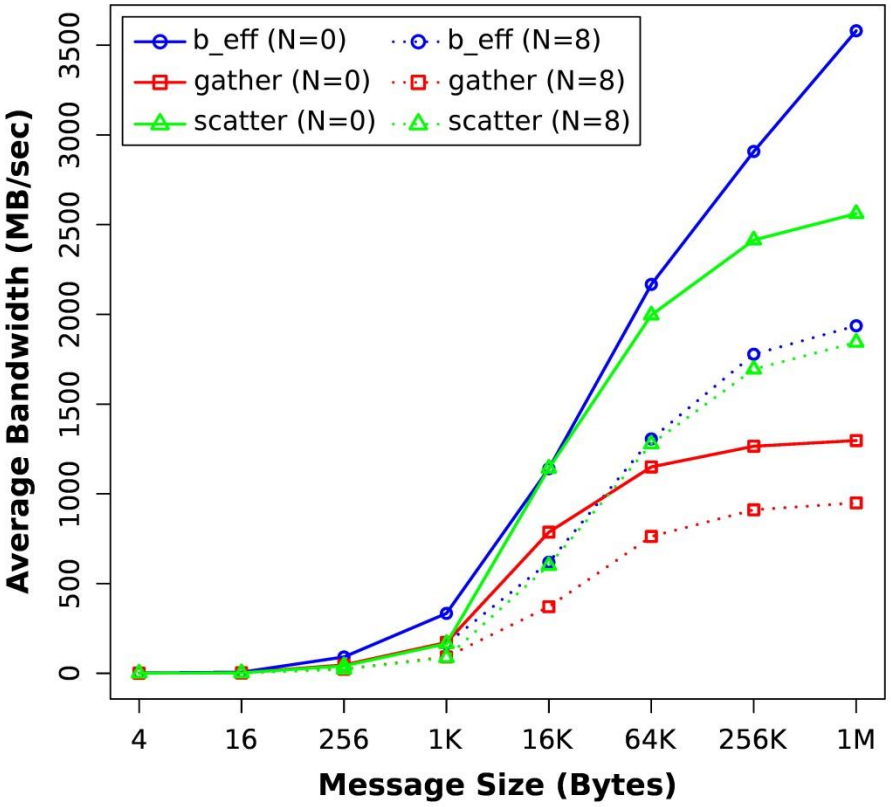


Bandwidth Tests

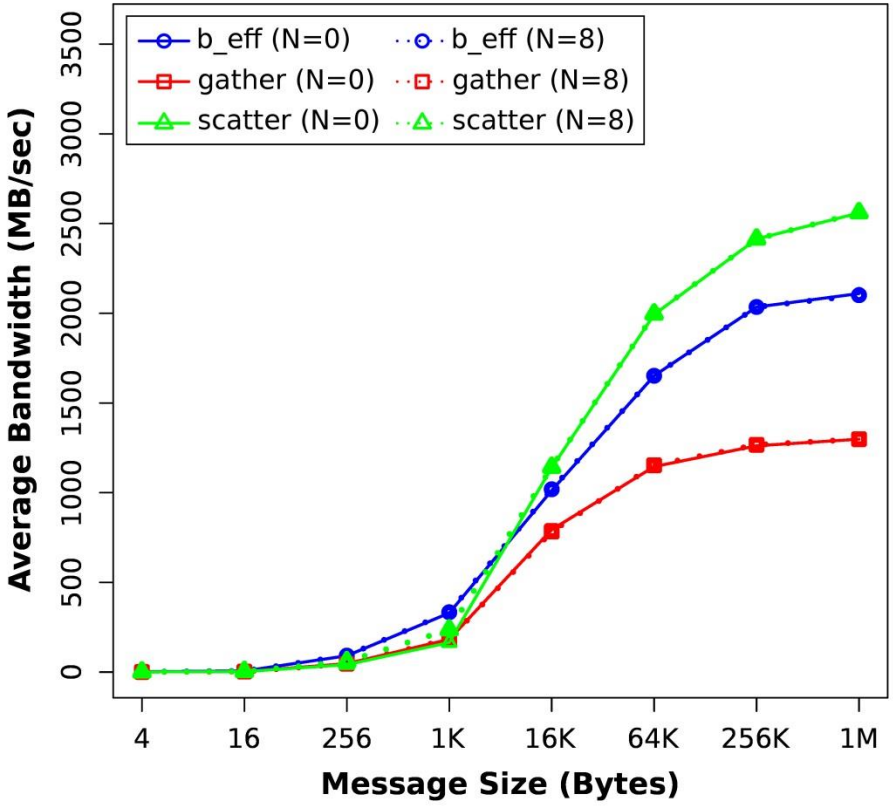


Latency Tests

Evaluation: We also run different communication patterns: Effective bisection bandwidth improves by 46%

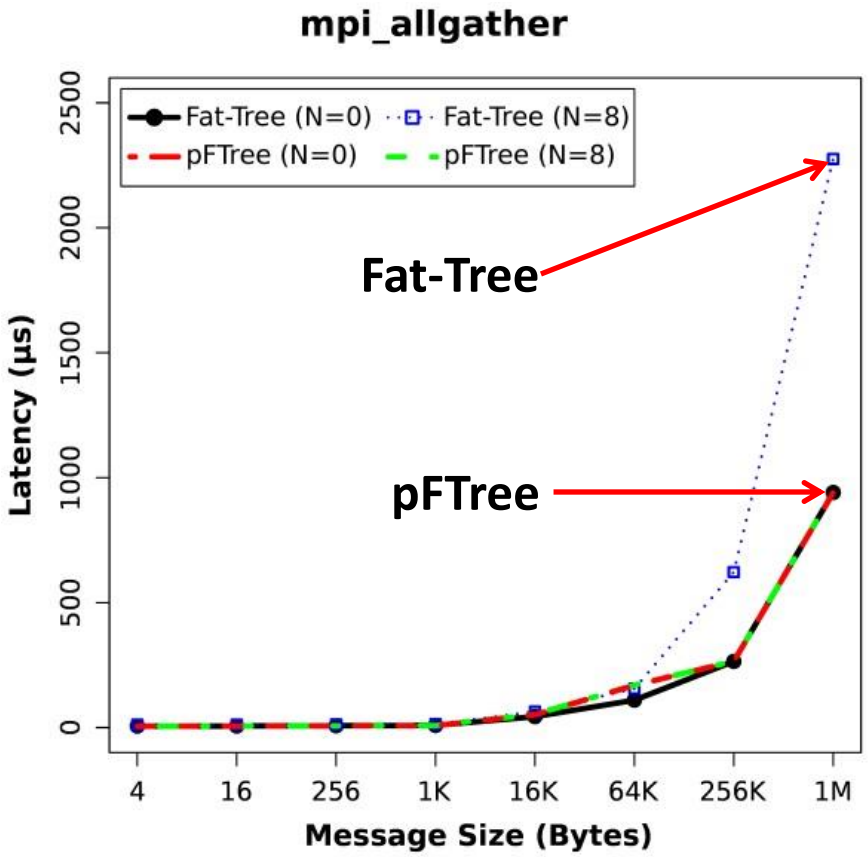


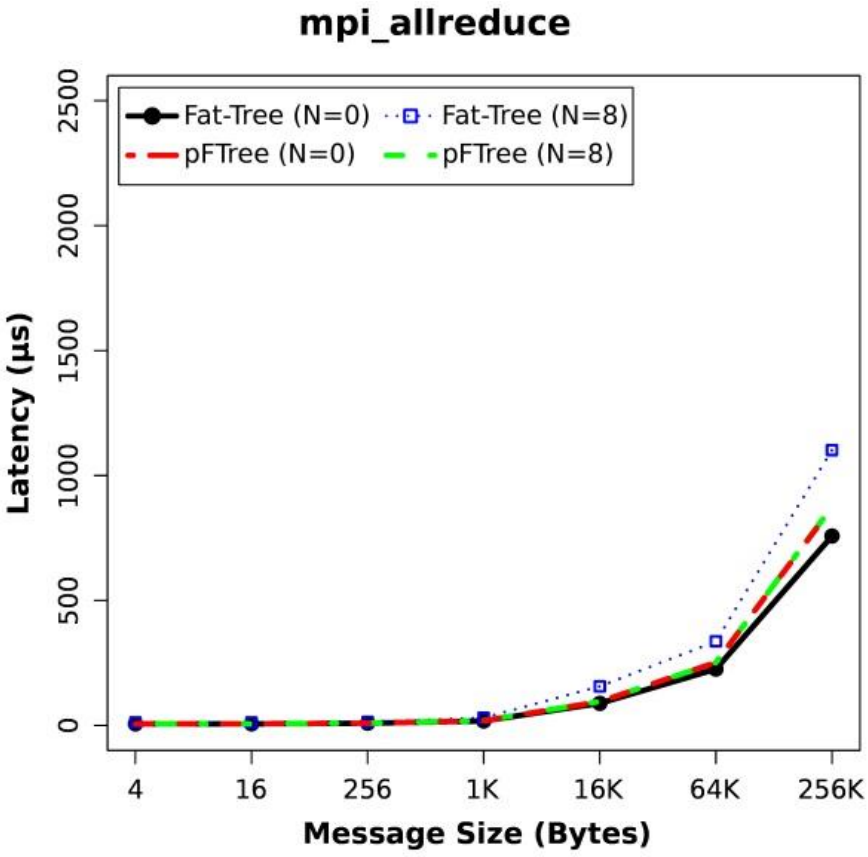
Fat-tree Routing



pFTree Routing

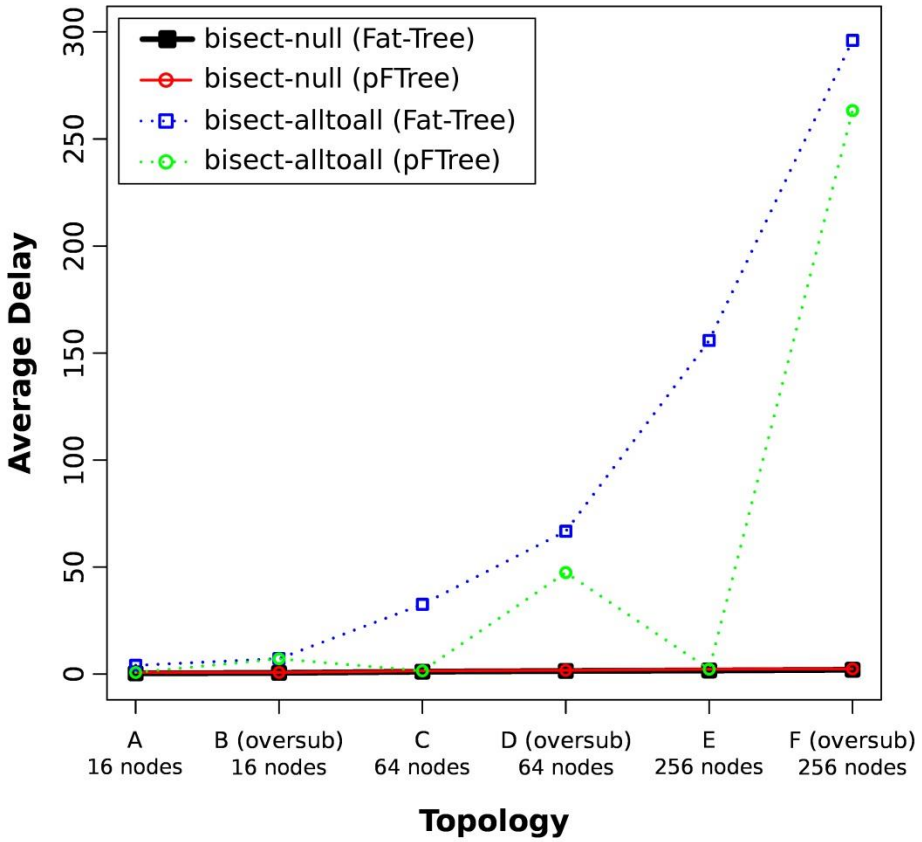
Evaluation: Many MPI collective operations also improves by more than 50% by eliminating partition interference





Evaluation: Simulations show substantial improvements in achieved throughput for several patterns

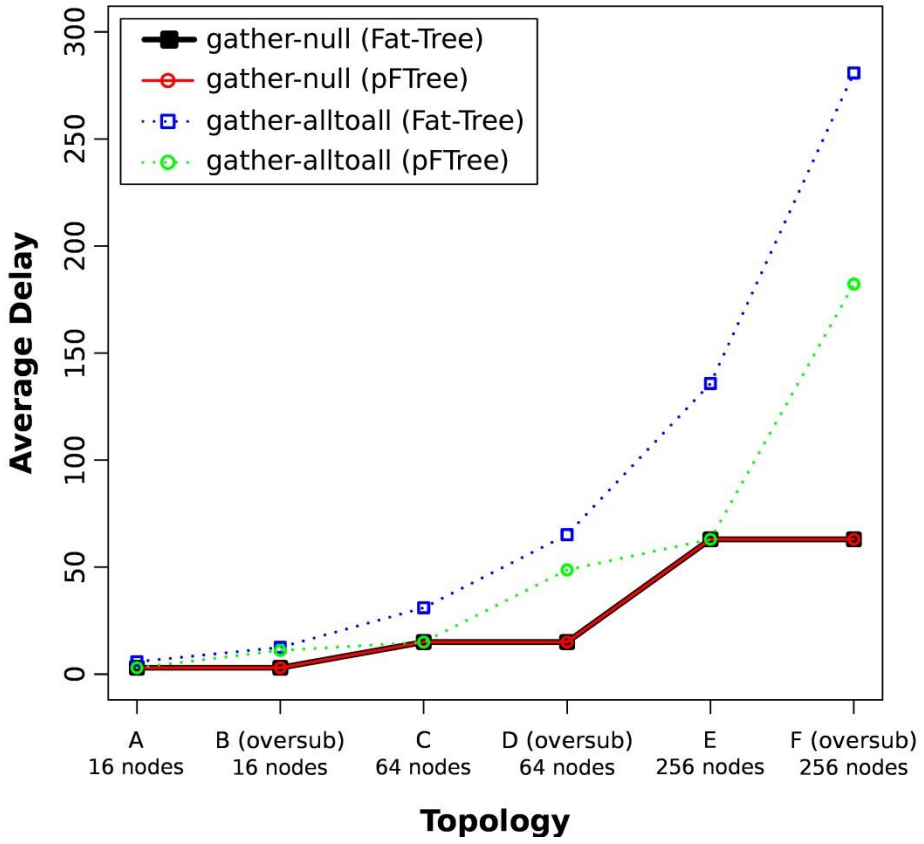
Topology	# Nodes	Subscription	Victims
A	16	1:1	4
B	16	2:1	4
C	64	1:1	16
D	64	2:1	16
E	256	1:1	64
F	256	2:1	64



Bisect Pattern

Evaluation: Simulations show substantial improvements in achieved throughput for several patterns

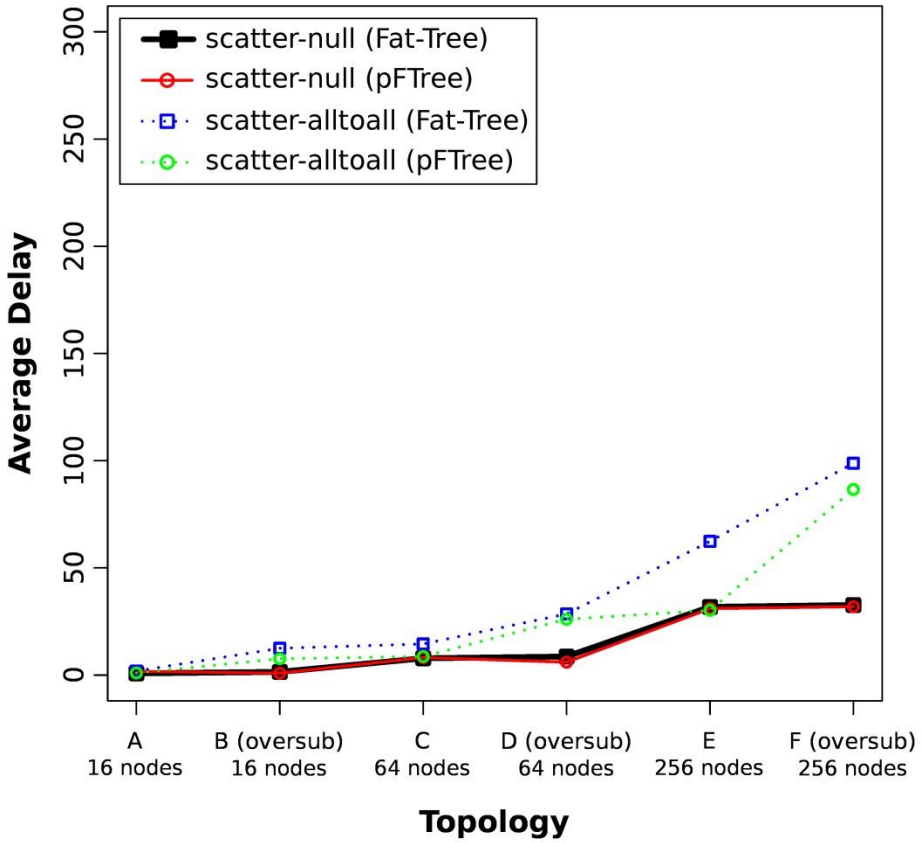
Topology	# Nodes	Subscription	Victims
A	16	1:1	4
B	16	2:1	4
C	64	1:1	16
D	64	2:1	16
E	256	1:1	64
F	256	2:1	64



Gather Pattern

Evaluation: Simulations show substantial improvements in achieved throughput for several patterns

Topology	# Nodes	Subscription	Victims
A	16	1:1	4
B	16	2:1	4
C	64	1:1	16
D	64	2:1	16
E	256	1:1	64
F	256	2:1	64

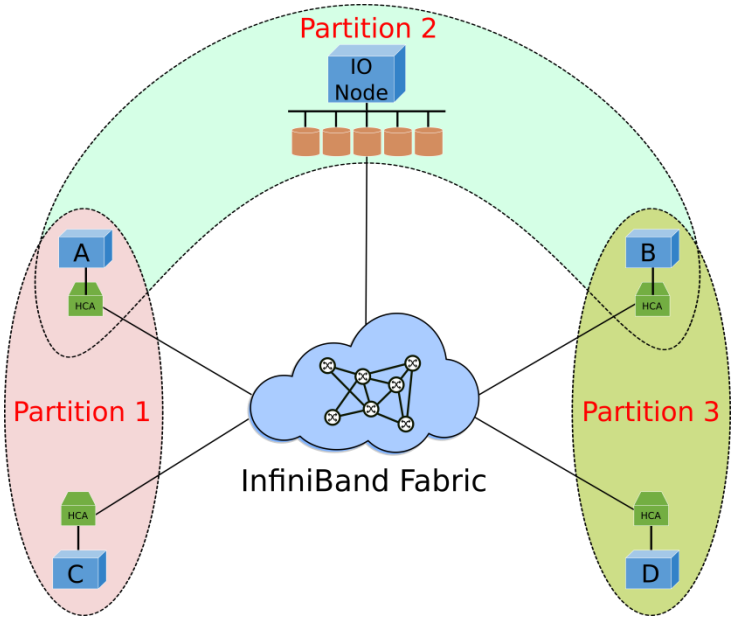


Scatter Pattern

In summary, partition-aware routing improves network isolation and performance in IB based multi-tenant clusters

State-of-the-art partition-oblivious fat-tree routing

The partition-aware fat-tree routing with better isolation



Questions?