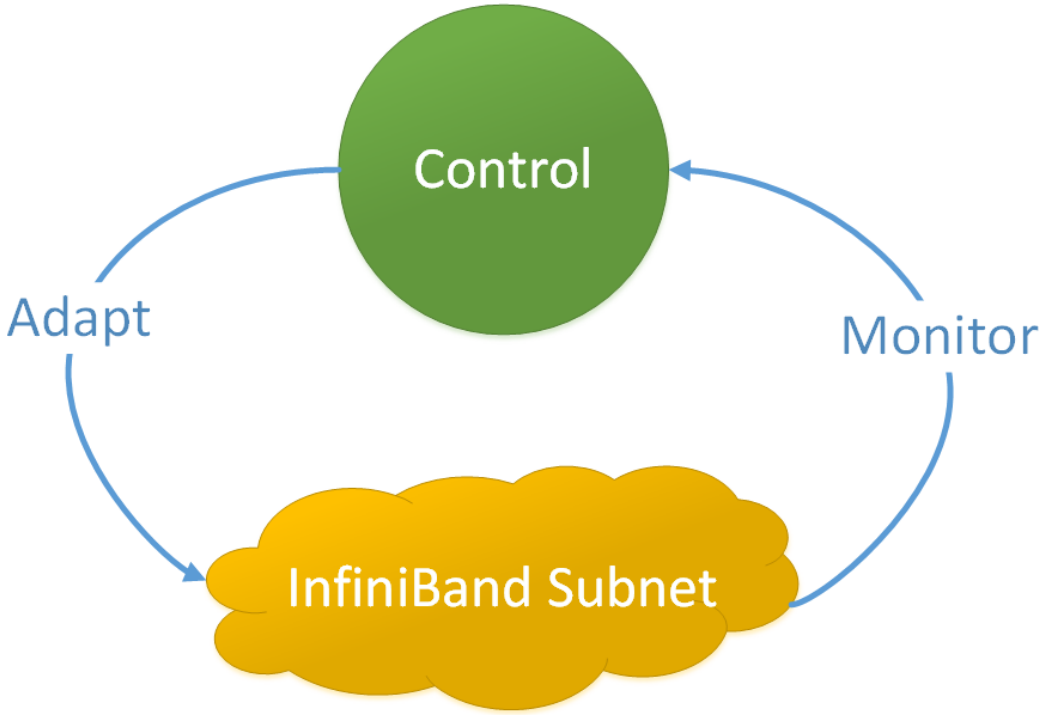
Realizing a Self-Adaptive Network Architecture for HPC Clouds

Feroz Zahid
Simula Research Laboratory

Advisors:
Ernst Gunnar Gran
Tor Skeie



SC '16 Doctoral Showcase

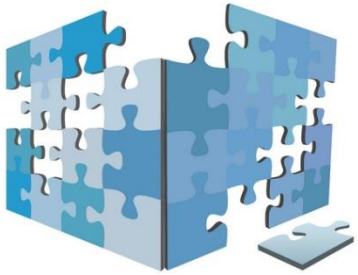
Salt Lake City, UT, USA

November 15, 2016

This presentation will walk through my doctoral work covering our contributions, and 'the big picture' ahead

Motivation and Challenges

Approach and Contributions



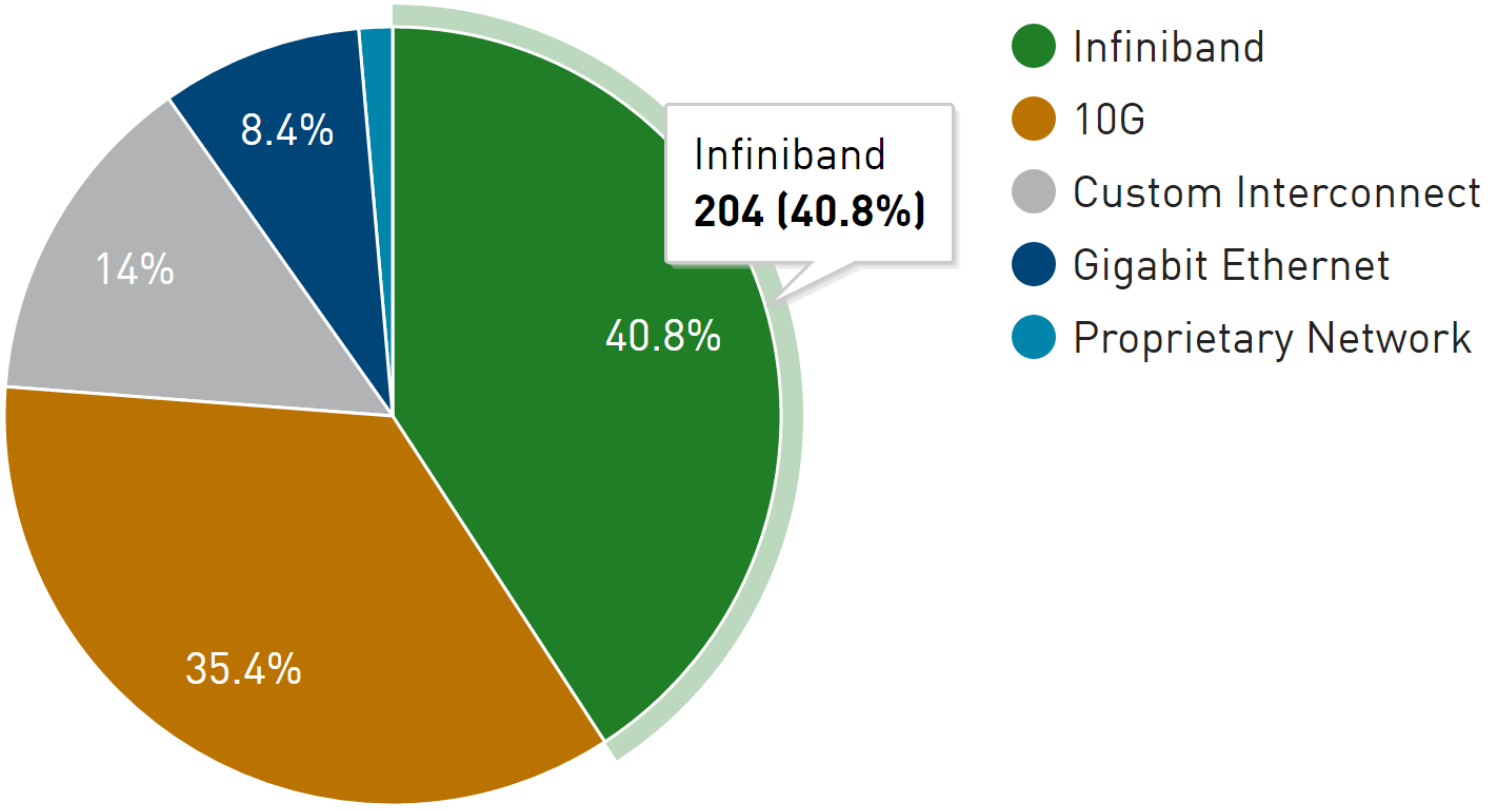
The Big Picture

InfiniBand (IB) is a popular interconnect for HPC systems

TOP 500 The List.

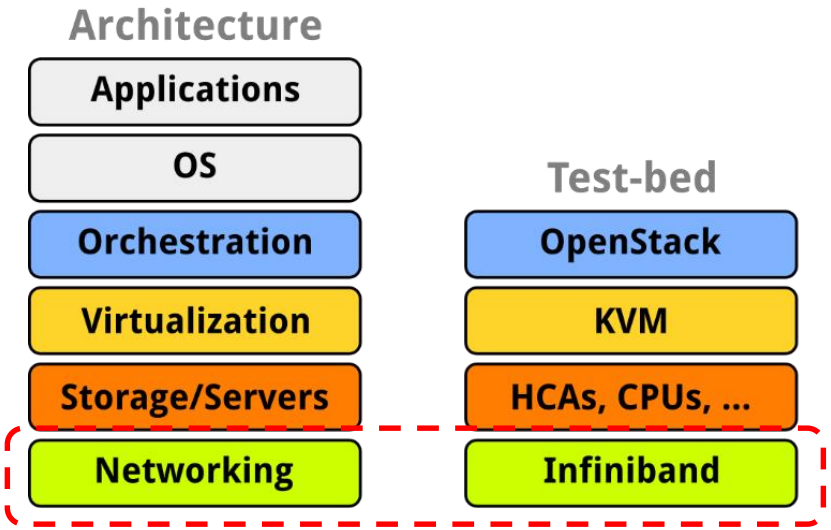
40.8% share in June 2016 top supercomputers list

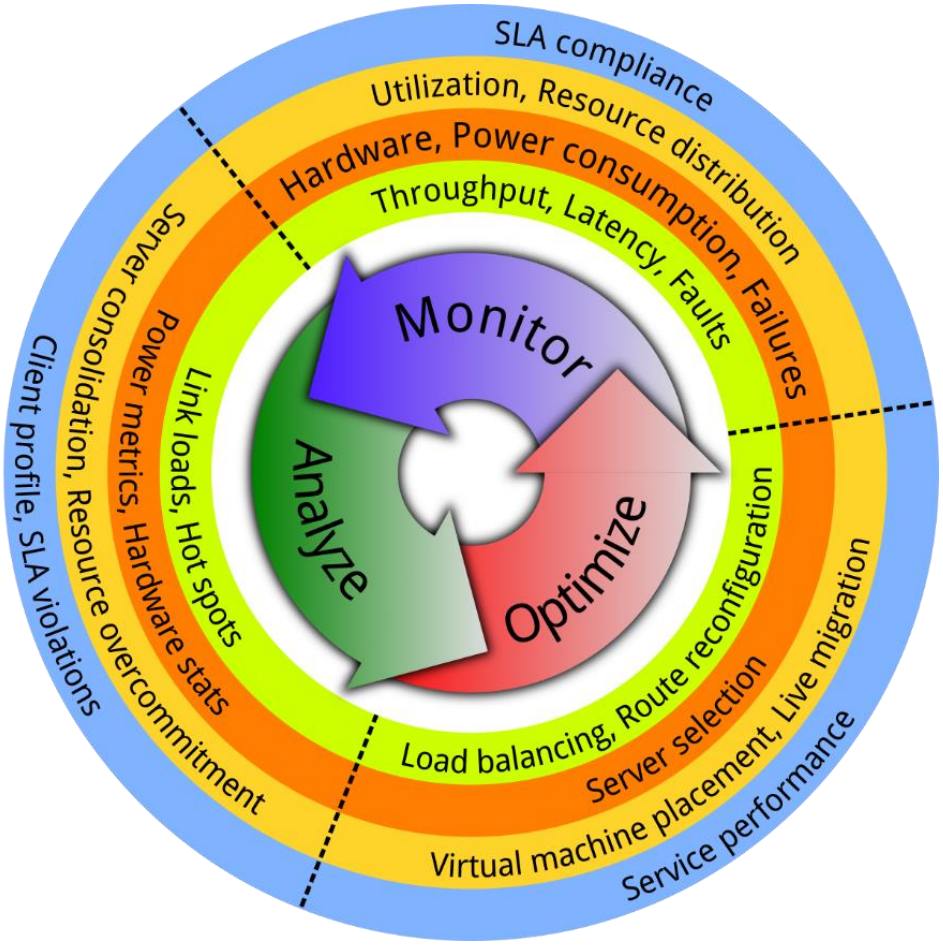
Interconnect Family System Share



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

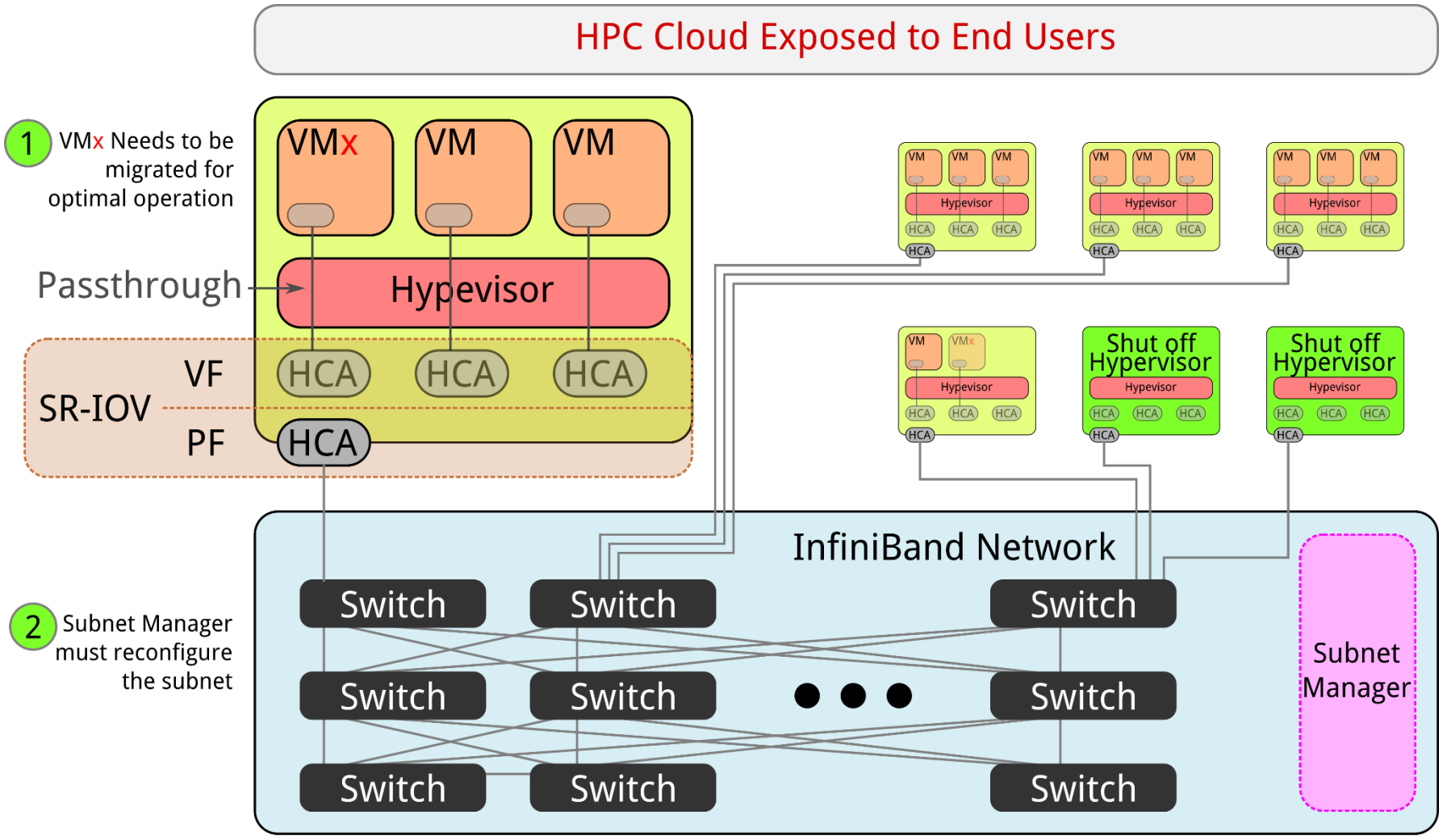
A whole array of challenges need to be addressed to realize a self-adaptive HPC cloud based on feedback-control loop





In this work, the focus has been on the network architecture for HPC clouds

To fully utilize the interconnection network, the network architecture must coordinate with the upper layers of cloud



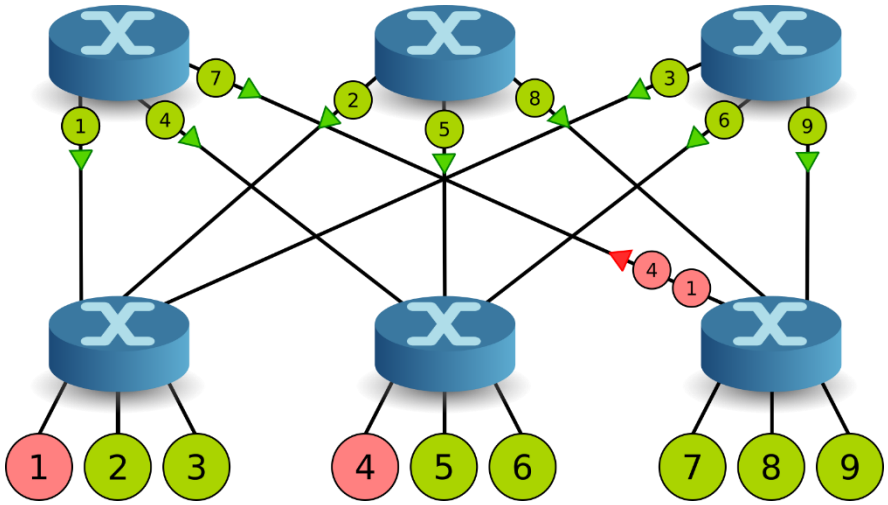
We use a bottom-up approach, and first attack individual research challenges associated with HPC cloud networks

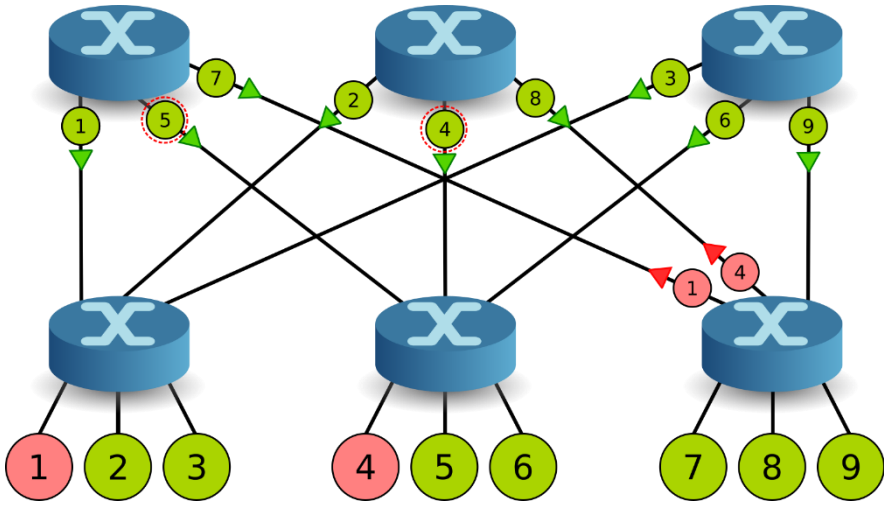
We uses OFED, de-facto standard software stack for IB, and Fat-Tree topology for our prototypes

- High Network Utilization and Better Load-Balancing
  - Weighted fat-tree routing algorithm (wFatTree)
- Multi-tenancy and Network Isolation
  - Partition-aware fat-tree routing (pFTree)
- Fast Network Reconfiguration
  - SlimUpdate routing algorithm (SlimUpdate)
  - Metabase-aided reconfiguration method
- Efficient Virtualization
  - Routing for virtualized subnets

Challenge 1: Efficient Network Utilization

The wFatTree routing algorithm considers node traffic characteristics to balance load across the network links more efficiently





Wt: 100 **100**

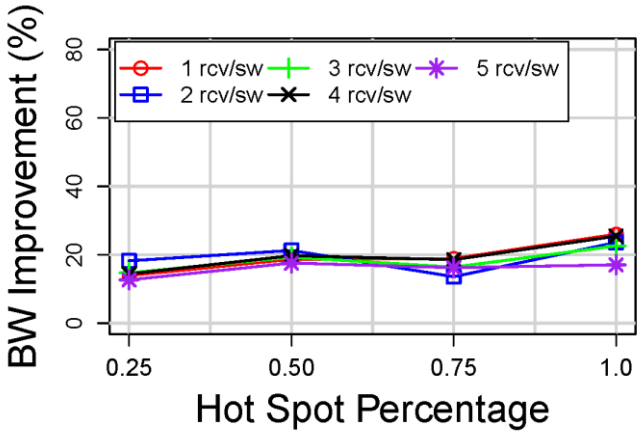
De-facto Fat-Tree Routing

The wFatTree Routing

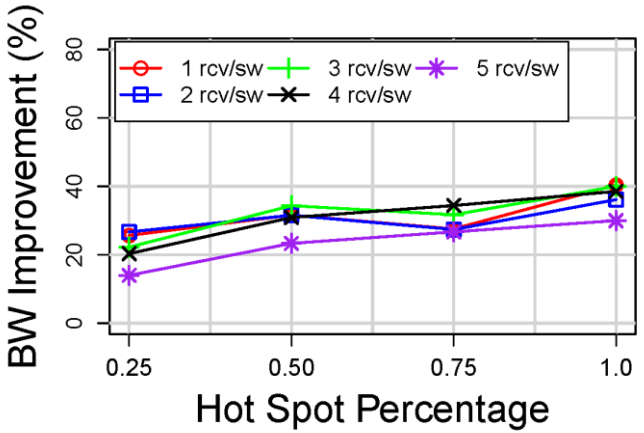
[1] A Weighted Fat-Tree Routing Algorithm for Efficient Load-Balancing in InfiniBand Enterprise Clusters. Zahid, Feroz et al., PDP, 2015.

Challenge 1: Efficient Network Utilization

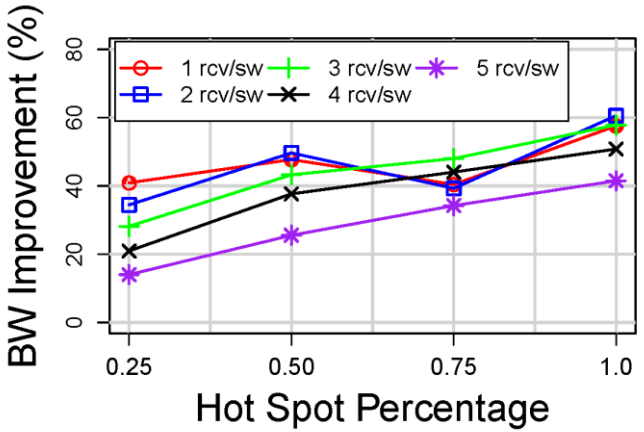
The wFatTree routing algorithm considers node traffic characteristics to balance load across the network links more efficientlv



18 switches with rcv nodes



27 switches with rcv nodes

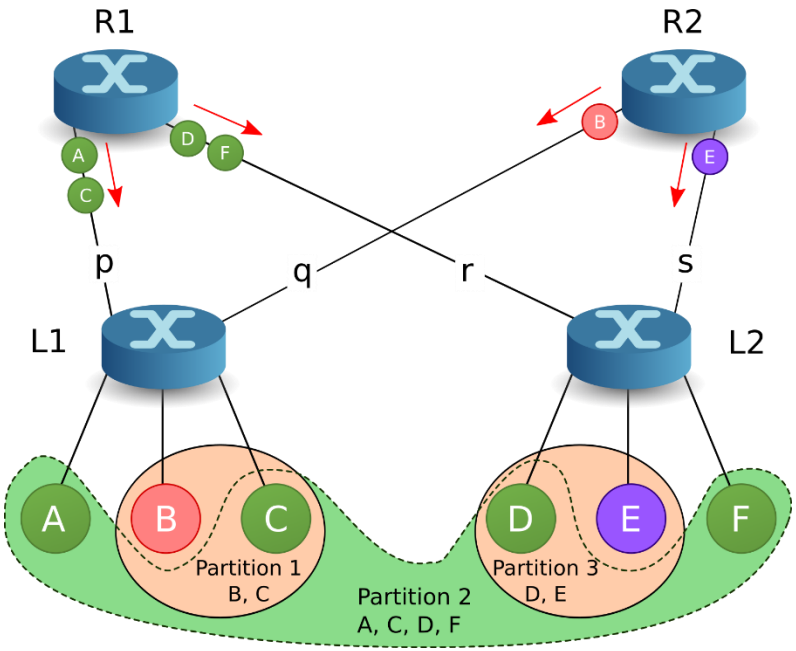


36 switches with rcv nodes

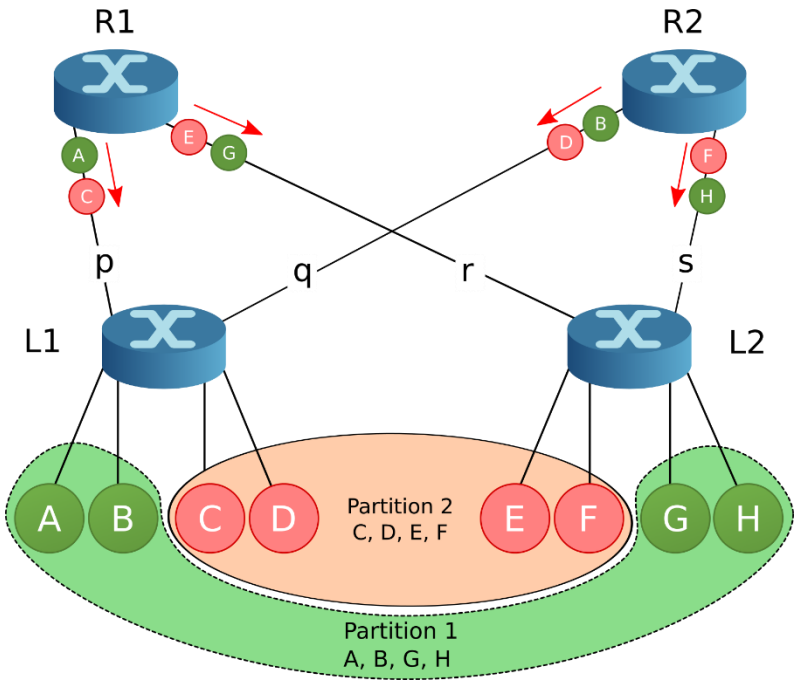
[1] A Weighted Fat-Tree Routing Algorithm for Efficient Load-Balancing in InfiniBand Enterprise Clusters. Zahid, Feroz et al., PDP, 2015.

Challenge 2: Tenant Performance Isolation

Traditional fat-tree routing in multi-tenant clusters suffers with degraded load balancing and no isolation between partitions



Degraded Load Balancing

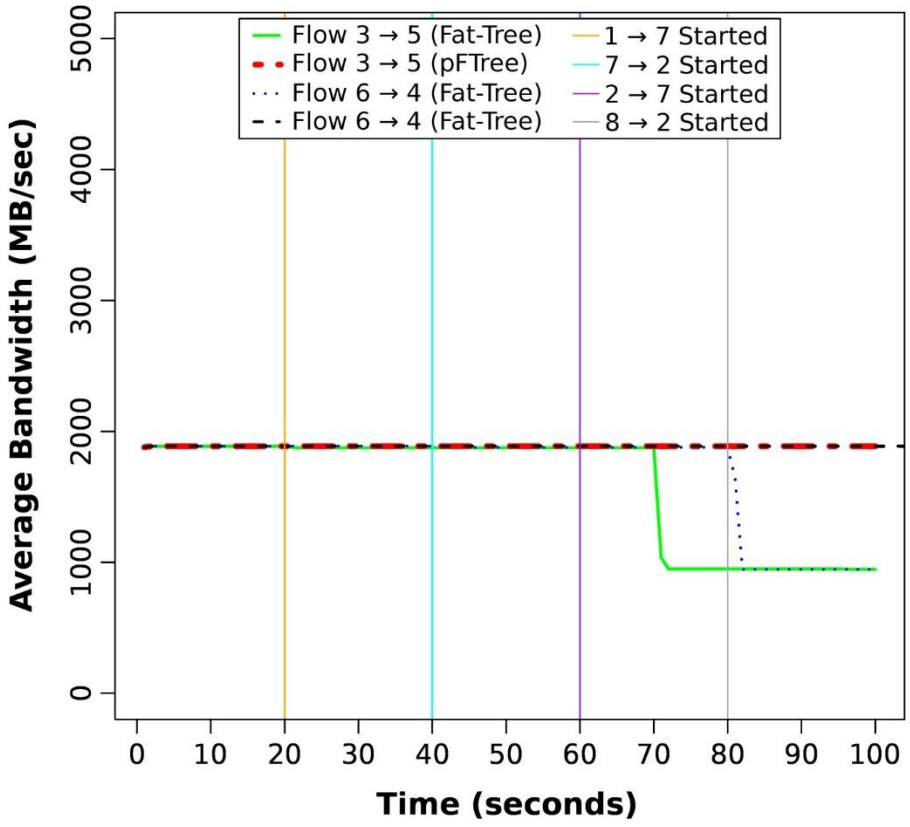


No Isolation Between Partitions

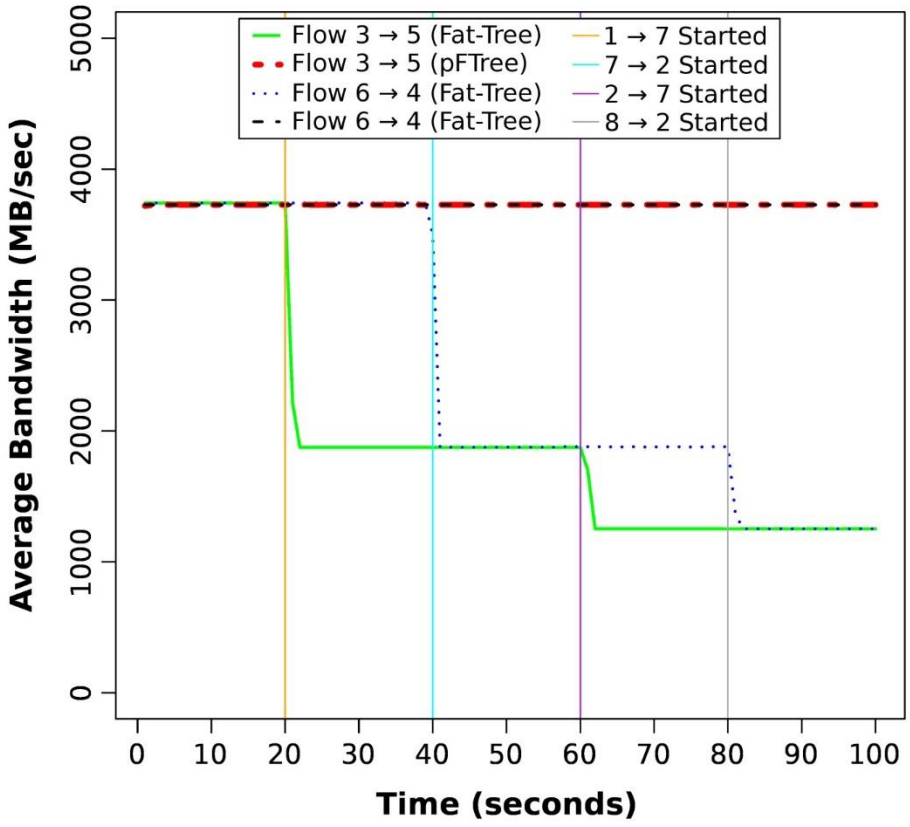
[2] Partition-aware Routing to Improve Network Isolation in Multi-tenant Clusters. Zahid, Feroz et al., CCGrid, 2015.

Challenge 2: Tenant Performance Isolation

The pFTree routing algorithm isolates partitions in a multi-tenant cluster without compromising on the load-balancing



Non-oversubscribed Topology



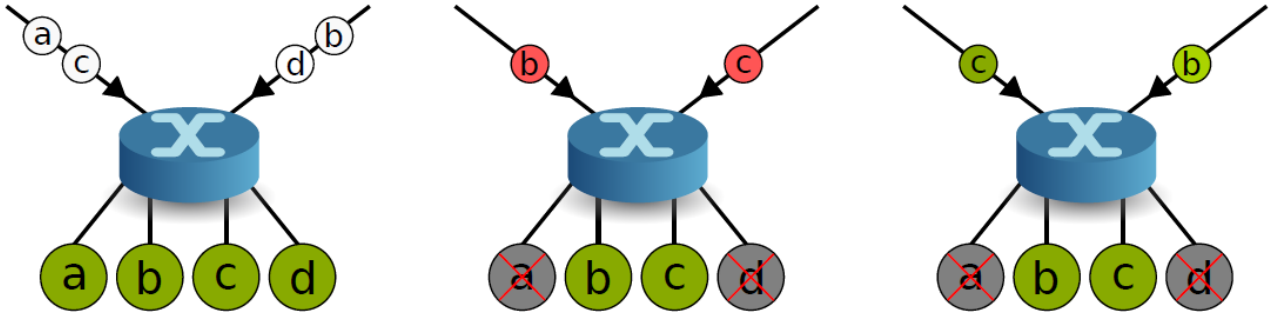
Oversubscribed Topology

[2] Partition-aware Routing to Improve Network Isolation in Multi-tenant Clusters. Zahid, Feroz et al., CCGrid, 2015.

Challenge 3: Fast Network Reconfiguration

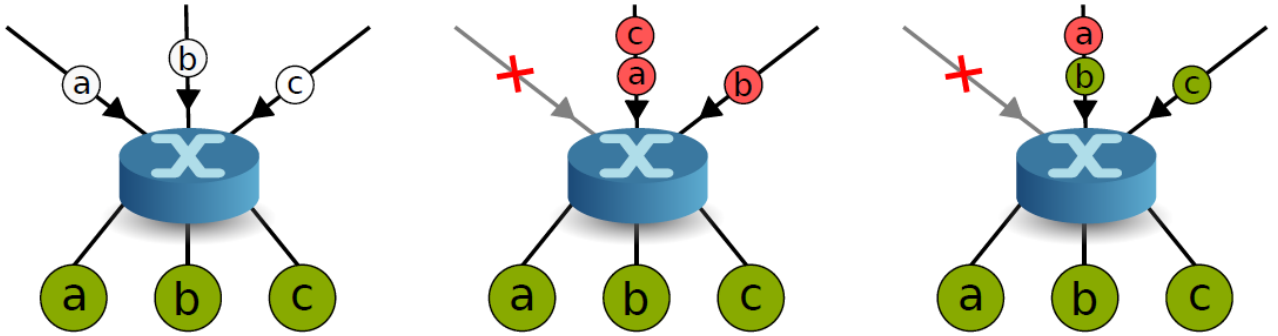
Minimal Routing Update (MRU) technique tends to preserve the configured paths in the network on a reconfiguration event

Nodes Shutdown



(a) Initial Routing (b) Without MRU (c) With MRU

Link Failure



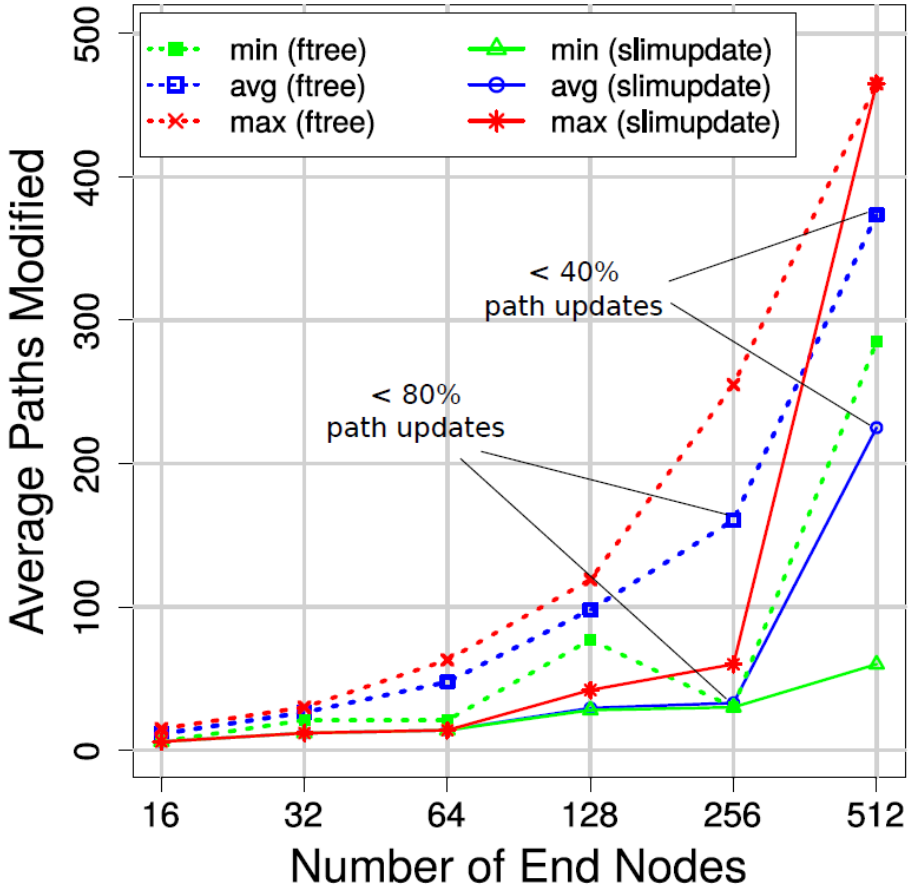
(d) Initial Routing (e) Without MRU (f) With MRU

[3] SlimUpdate: Minimal Routing Update for Performance-Based Recongurations in Fat-Trees, Zahid, Feroz et al., HiPINEB 2015.

Challenge 3: Fast Network Reconfiguration

SlimUpdate Routing algorithm utilizes MRU technique, and saves up to 80% path updates

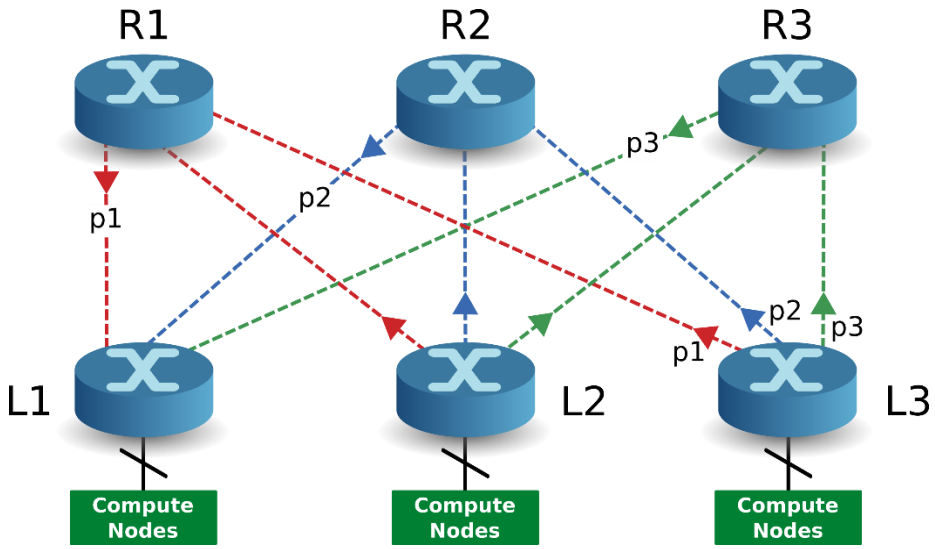
Name	# Nodes	Topology
A	16	4-ary-2-tree
B	32	4-ary-2-tree oversub
C	64	4-ary-3-tree
D	128	4-ary-3-tree oversub
E	64	8-ary-2-tree
F	128	8-ary-2-tree oversub
G	256	16-ary-2-tree
H	512	16-ary-2-tree oversub



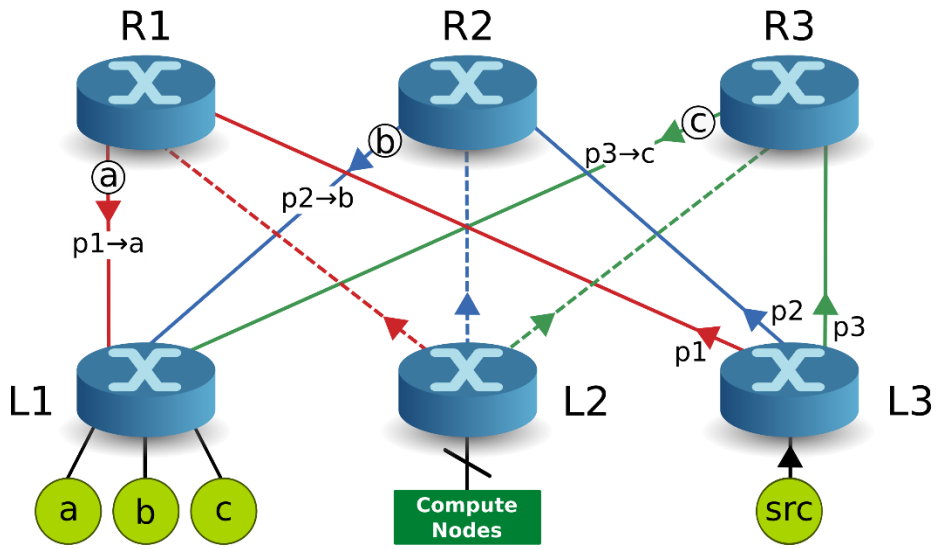
[3] SlimUpdate: Minimal Routing Update for Performance-Based Recongurations in Fat-Trees, Zahid, Feroz et al., HiPINEB 2015.

Challenge 3: Fast Network Reconfiguration

In metabase-aided reconfiguration method, routing is divided into two distinct phases: calculation of paths, and assignment of paths to the actual destinations



Phase I: Calculation of paths

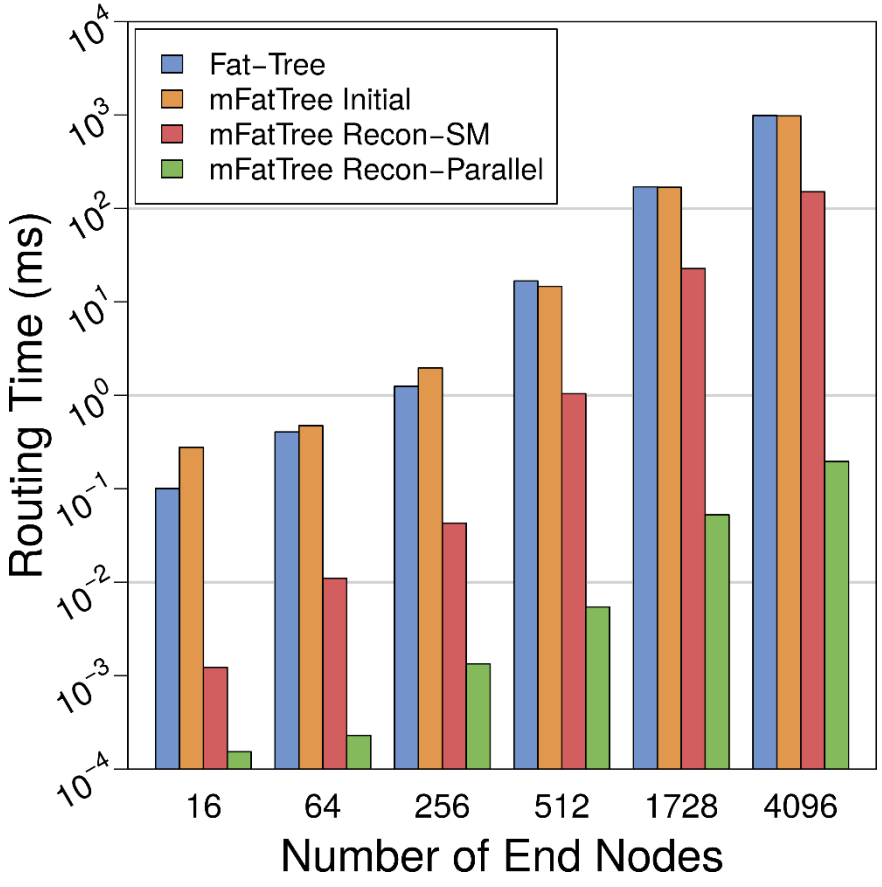


Phase II: Assignment of Paths

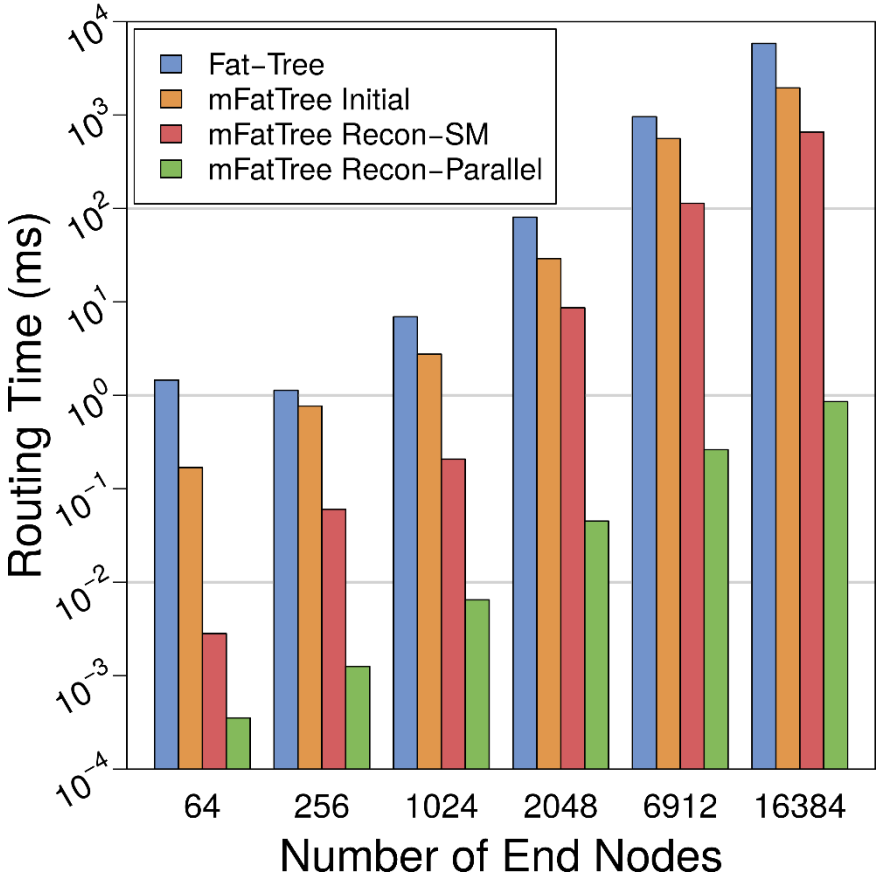
[4] Compact Network Reconfiguration in Fat-Trees, Zahid, Feroz et al., The Journal of Supercomputing, 2016.

Challenge 3: Fast Network Reconfiguration

Metabase-aided routing substantially reduces network reconfiguration time on performance-based reconfigurations



Non-oversubscribed Topologies

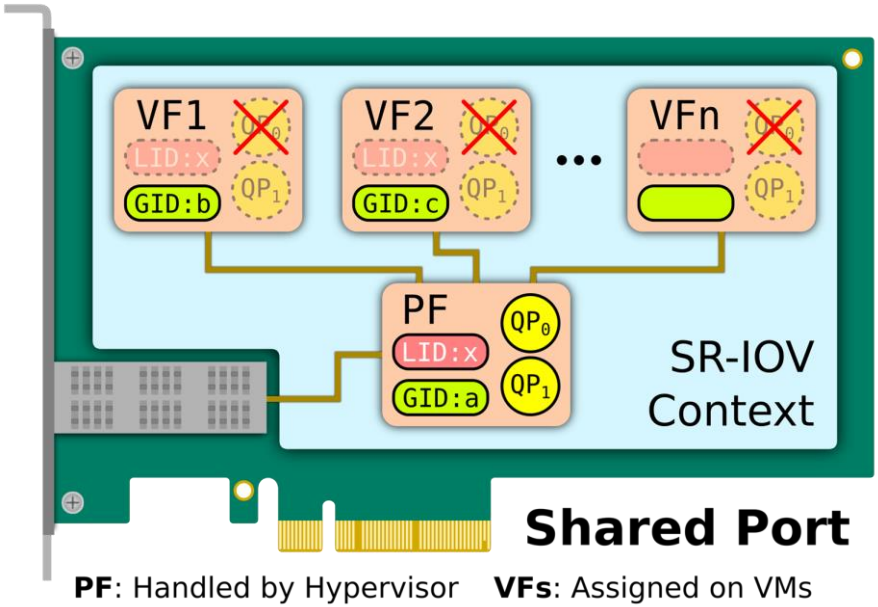


Oversubscribed Topologies

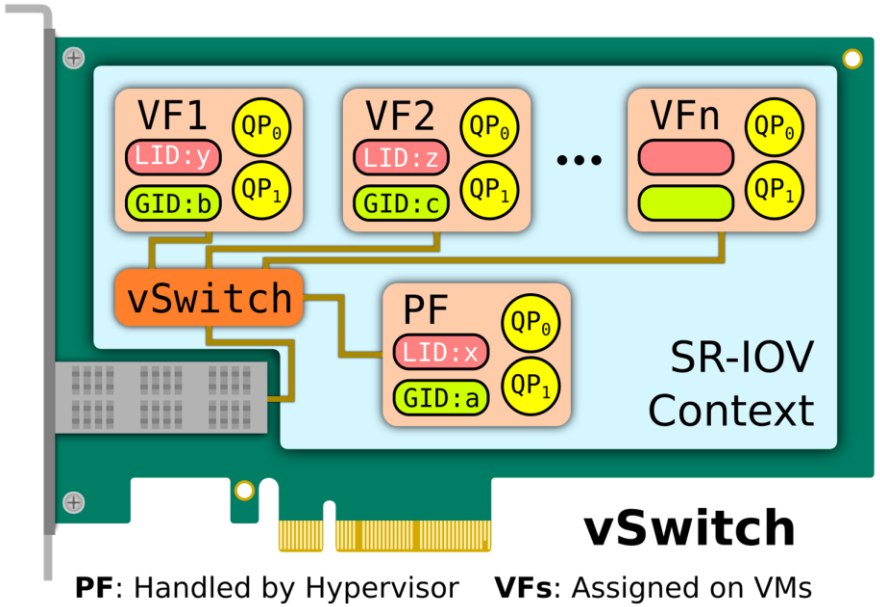
[4] Compact Network Reconfiguration in Fat-Trees, Zahid, Feroz et al., The Journal of Supercomputing, 2016.

Challenge 4: Efficient Virtualization

The vSwitch Architecture has an advantage over shared-port architecture that it allows configuring routes for the individual VMs in the subnet (but bloats LID space); hybrid models can save LIDs



PF: Handled by Hypervisor **VFs**: Assigned on VMs



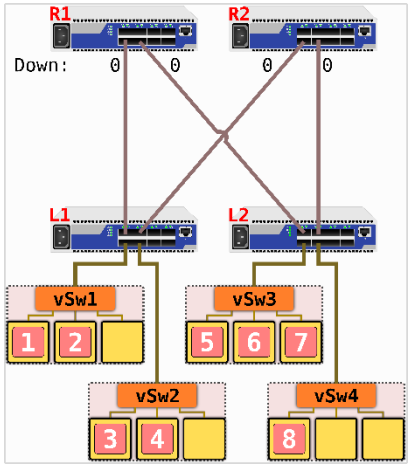
PF: Handled by Hypervisor **VFs**: Assigned on VMs

[5] Towards InfiniBand SR-IOV vSwitch Architecture, Tasoulas, Evangelos et al., IEEE Cluster, 2015.

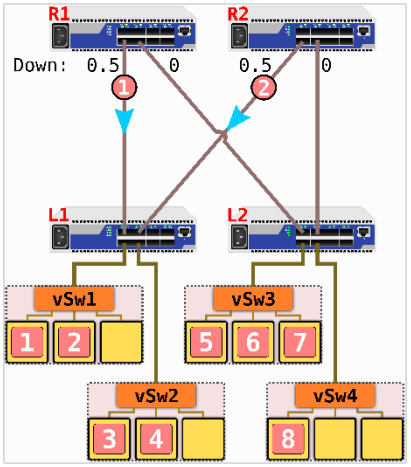
Challenge 4: Efficient Virtualization

The vSwitchFatTree routing considers VMs in the subnet

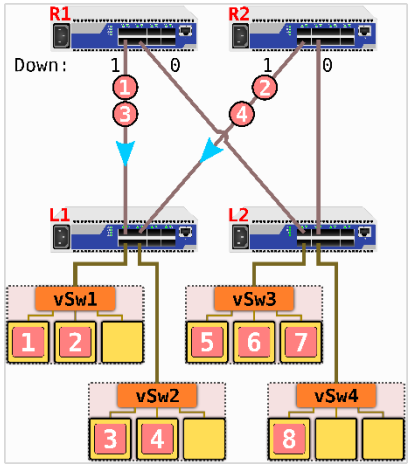
(a)



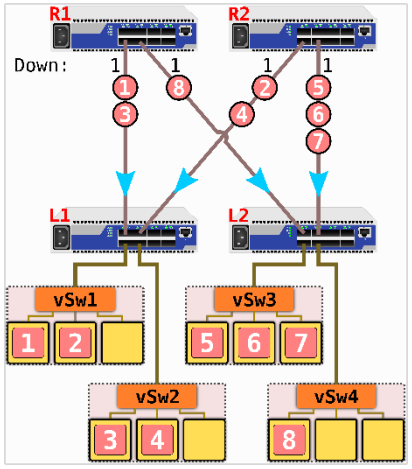
(b)



(c)



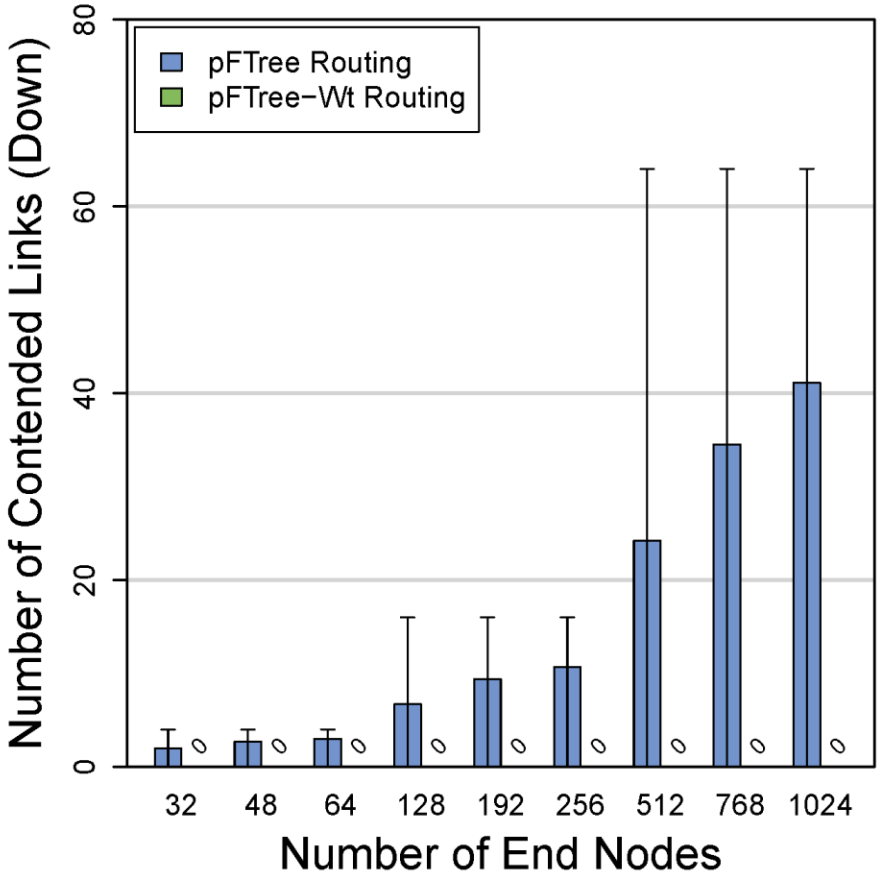
(d)

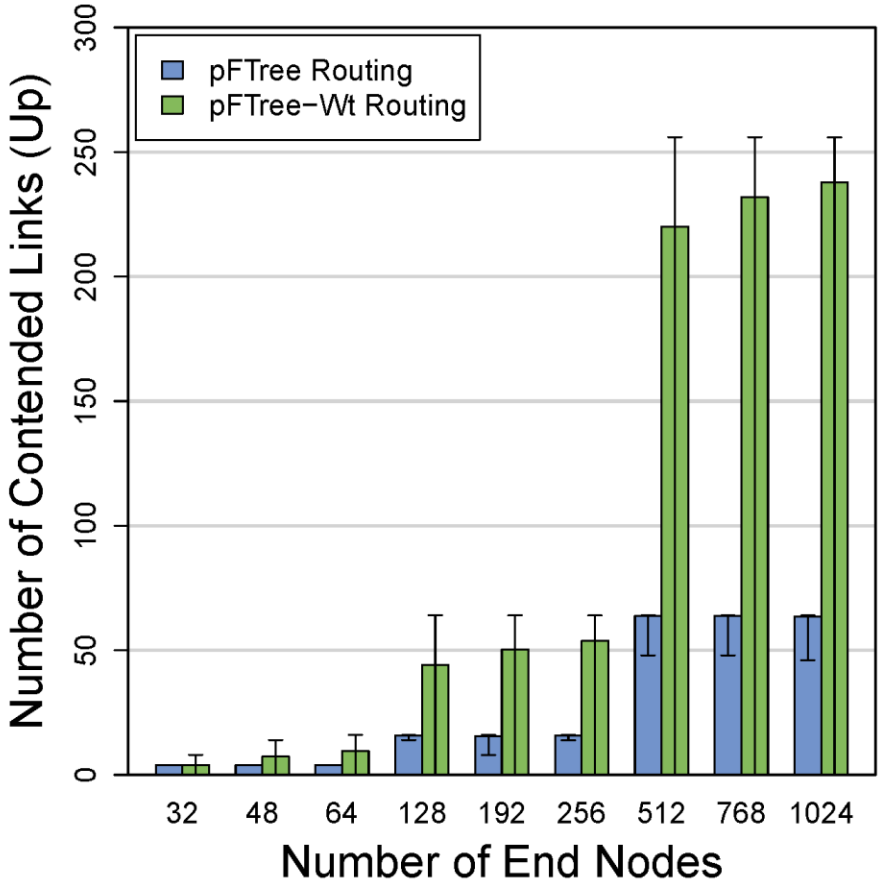


[6] Towards Efficient Virtualization in HPC Environments. Tasoulas, Evangelos, Zahid, Feroz et al., Submitted to an Internatioal Journal.

Big Picture: Enable smart network provisioning for the HPC clouds – combine individual contributions

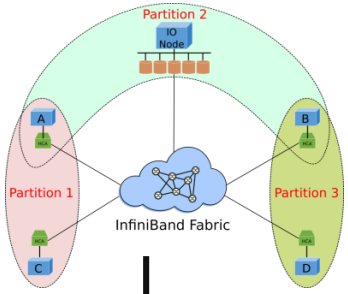
Weighted pFTree routing (pFTree-Wt) can substantially reduce contention in a partitioned subnet

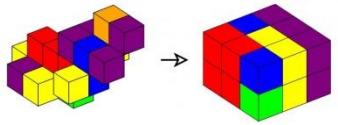




[7] Efficient Network Isolation and Load-balancing in Multi-tenant HPC Cluster, Zahid, Feroz et al., Future Generation Computer Sys, 2016.

Big Picture: Enable smart network provisioning for the HPC clouds – combine individual contributions





Optimized Algorithms

Weighted Routing

Partition-aware Routing

Adjust for Load/Faults

Better Routes

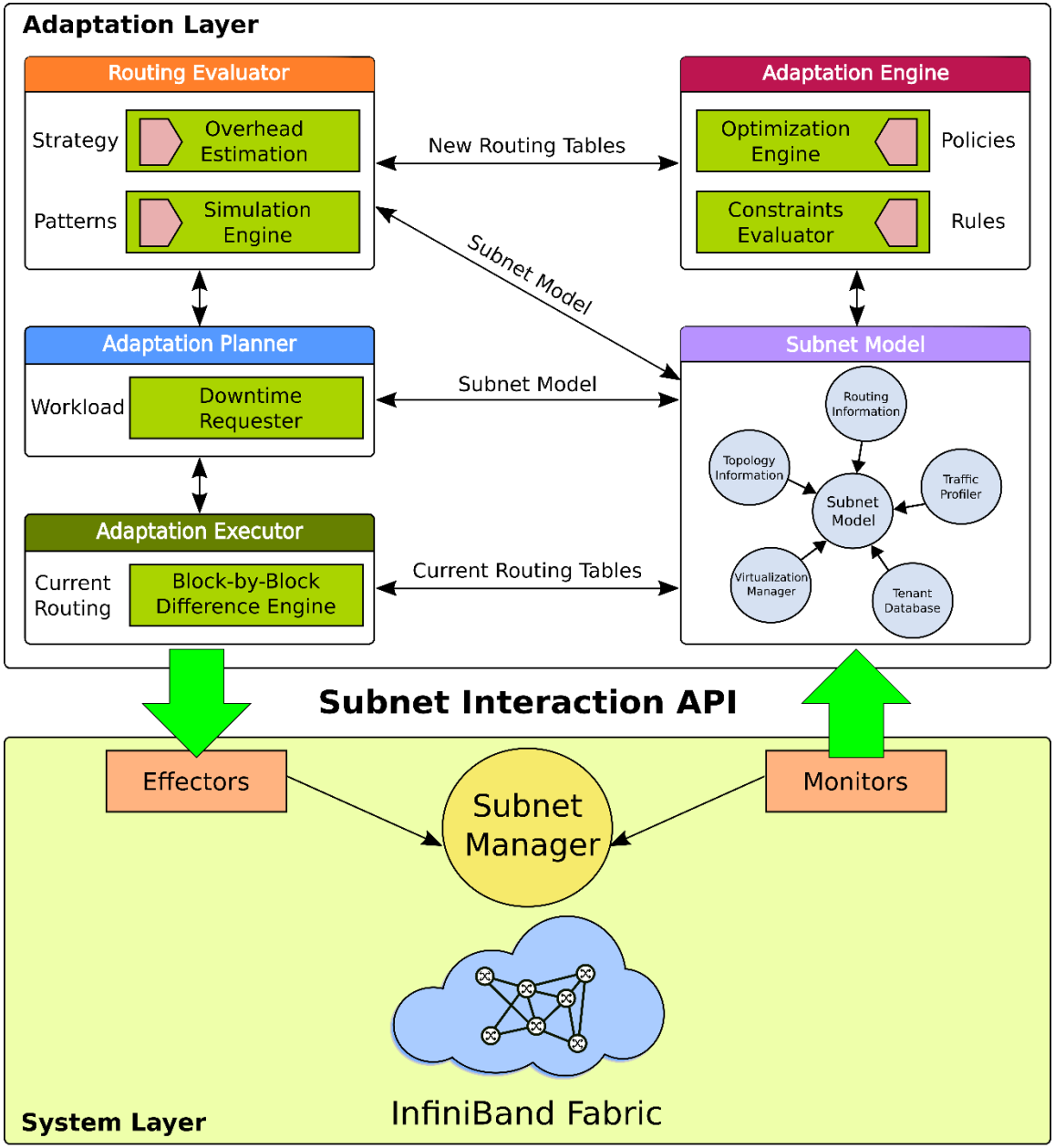
Balanced Traffic

Multi-tenancy

Dynamic Optimizations

Monitor->Optimize->Execute Loop

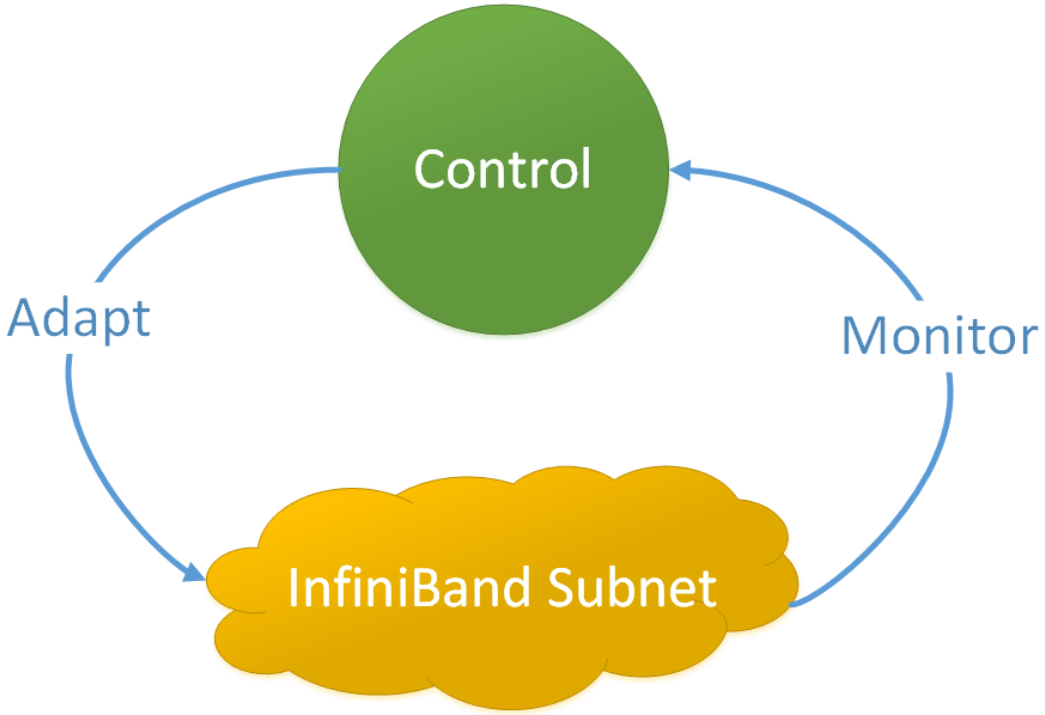
Big Picture: A Self-Adaptive Network Architecture



In summary, a self-adaptive network architecture can make HPC clouds fully utilize underlying interconnection network

State-of-the-art network architecture with static configurations

A Self-adaptive network architecture enabling dynamic HPC clouds



Thanks for your attention!