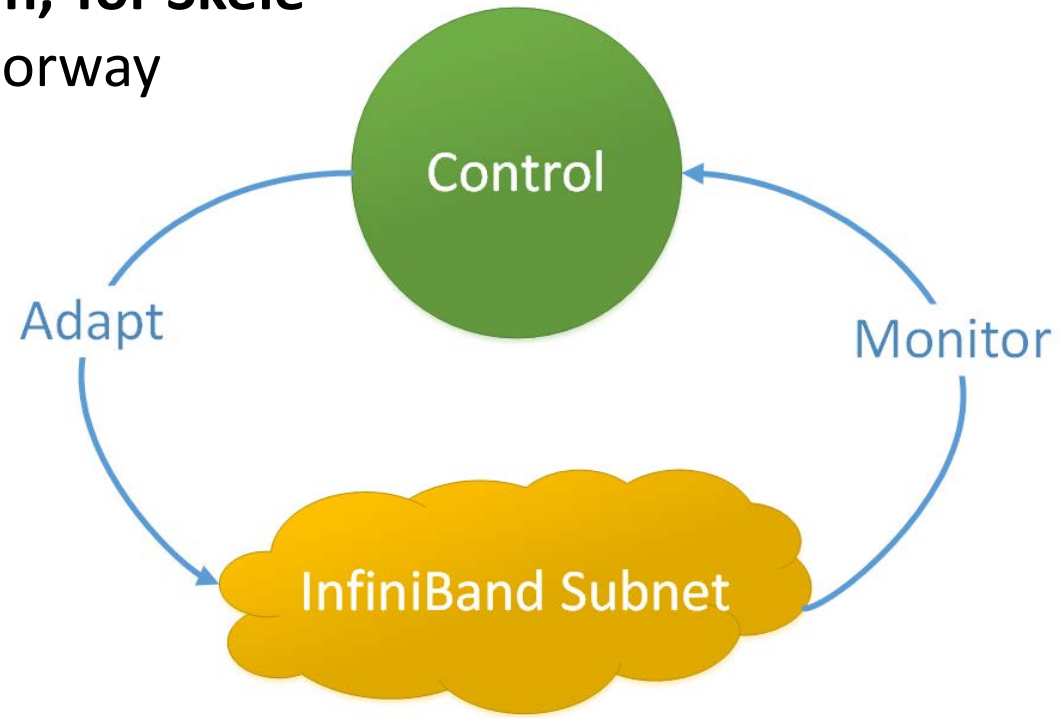
A Self-Adaptive Network Architecture for InfiniBand based HPC Clouds

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



7th Cloud Control Workshop

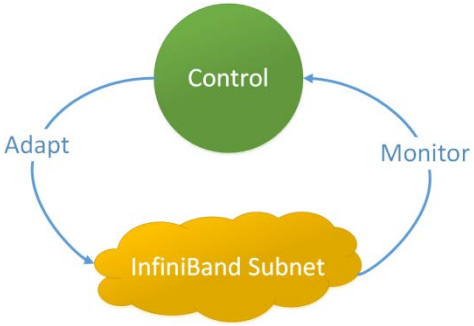
Nässlingen, Sweden

June 9, 2015

[simula . research laboratory]

This presentation leads to the discussion session after briefly introducing our contributions and future goals

Background and Contributions



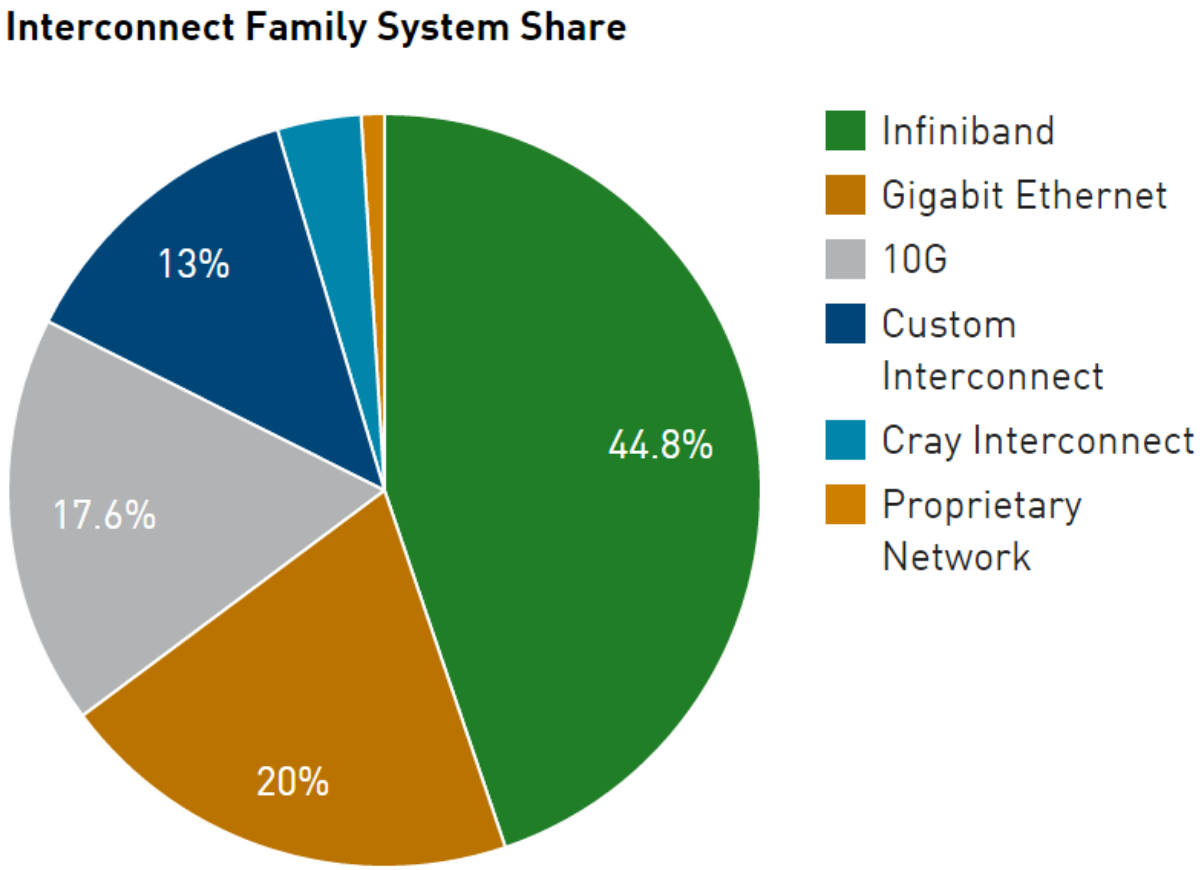
The Goal: Self-Adaptive IB Subnets

Problems and Discussion

InfiniBand (IB) is a popular interconnect for HPC systems

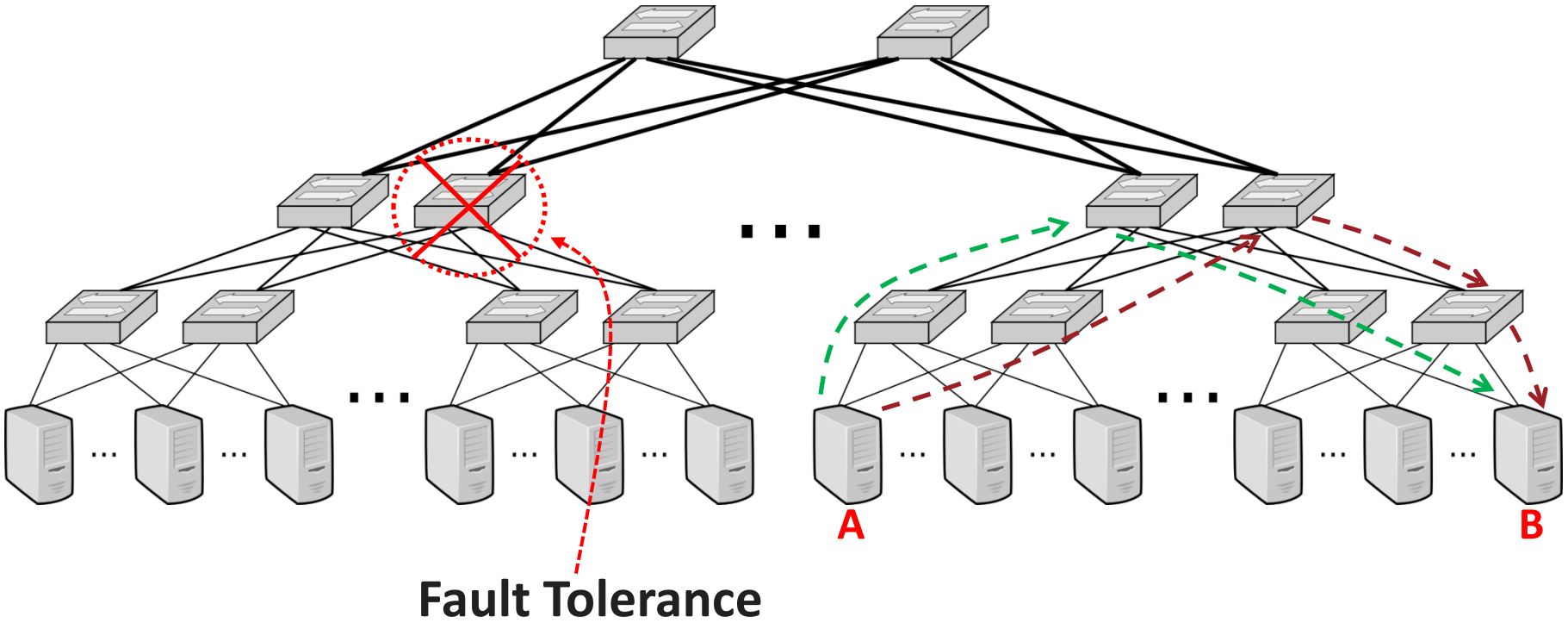
TOP 500 The List.

44.8% share in November 2014 top supercomputers list



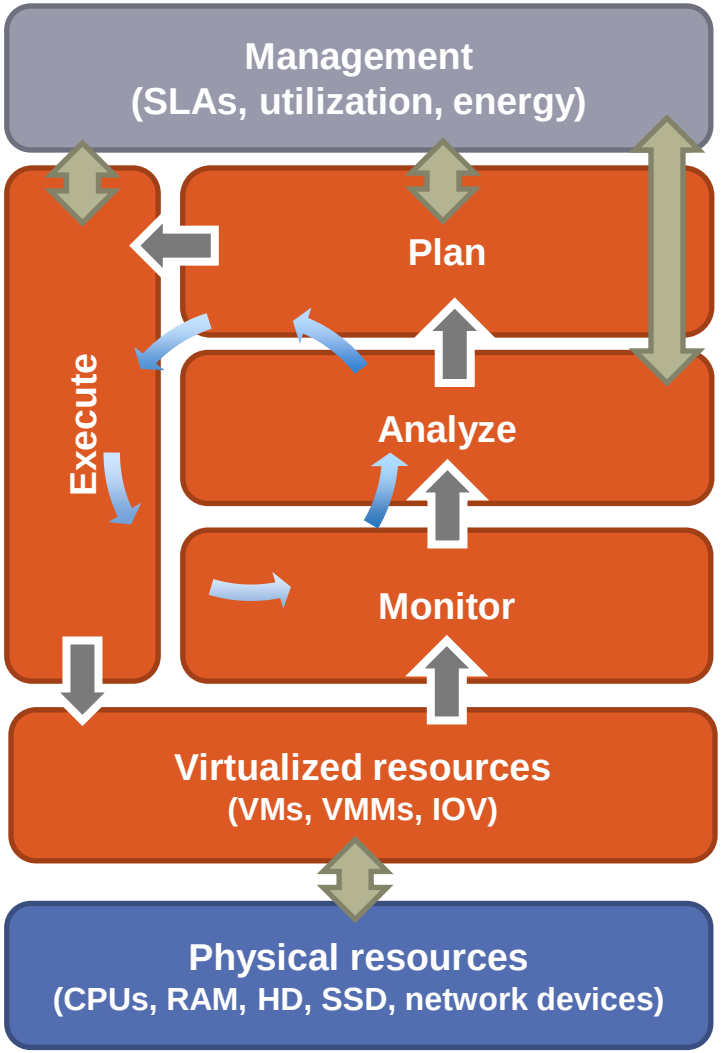
Source: Top500 Supercomputers List, http://top500.org/

In commonly used HPC topologies, the choice of multiple paths is available between a source-destination pair



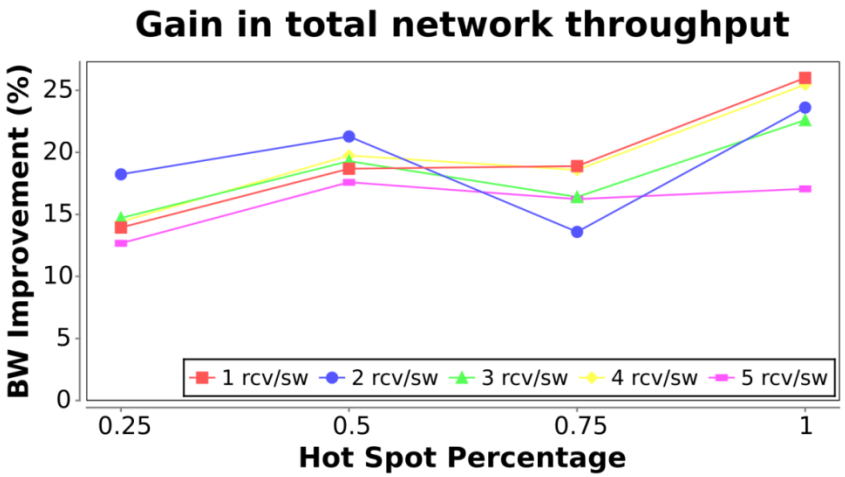
Our aim is to design a holistic self-adaptive architecture for the HPC Clouds based on feedback-control loop

- **Monitor and Collect**
  - Performance Data
  - Faults / Failures
- **Analyze and Plan**
  - Based on collected data
  - Management Directives
- **Execute and Reconfigure**
  - Reconfigure resources
  - Execute new configuration

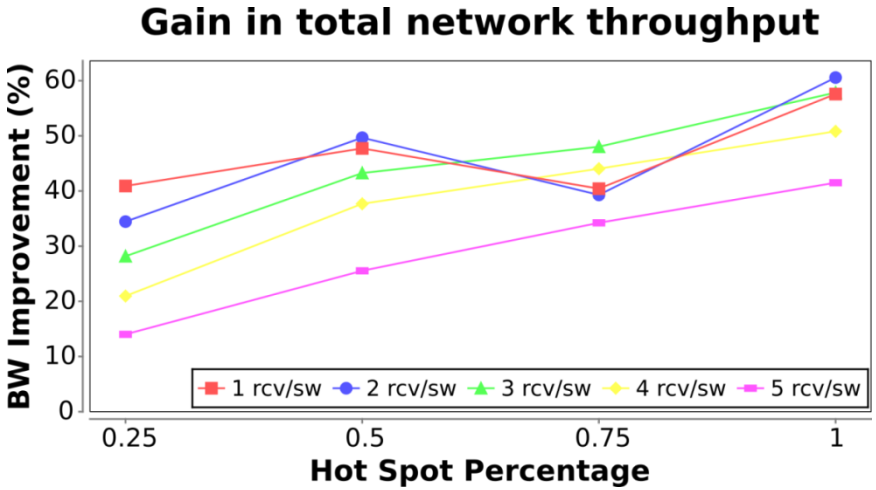


By considering node roles in routing, the overall network utilization can be substantially improved

- In wFatTree routing, nodes are assigned a new parameter - weight
- Weights can be assigned based on
  - Known node roles e.g. storage nodes
  - Known traffic priorities e.g. following QoS levels
  - Traffic profiling
- Nodes are routed in the order of their weights: Predictable
- Port selection is based on both Downward and Upward weight



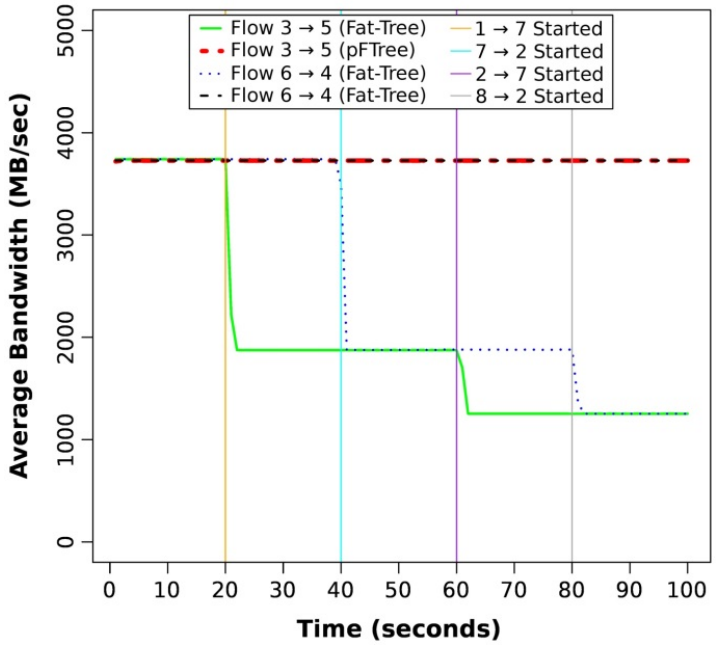
27 Switches with rcv nodes

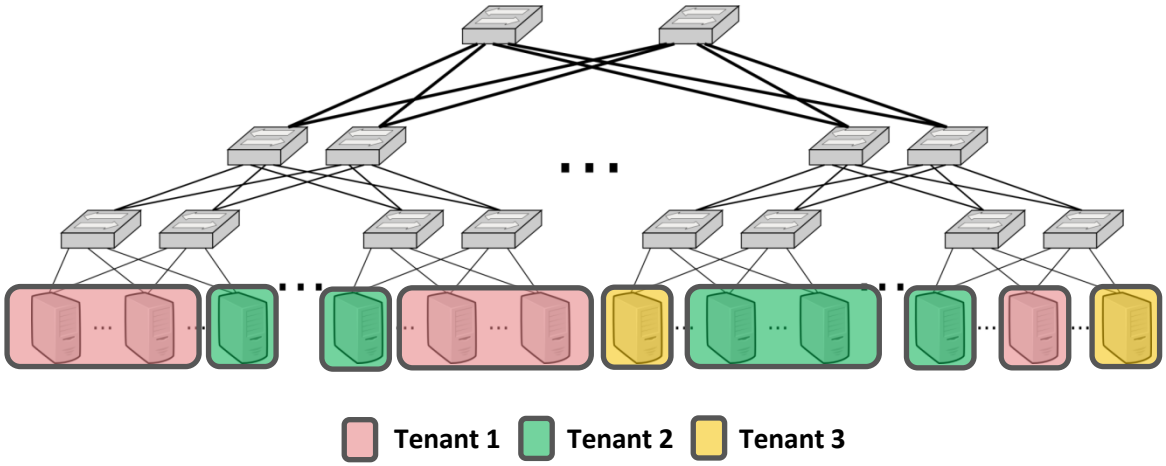


36 Switches with rcv nodes

Tenant-aware routing provides better isolation and also improves performance in cloud systems

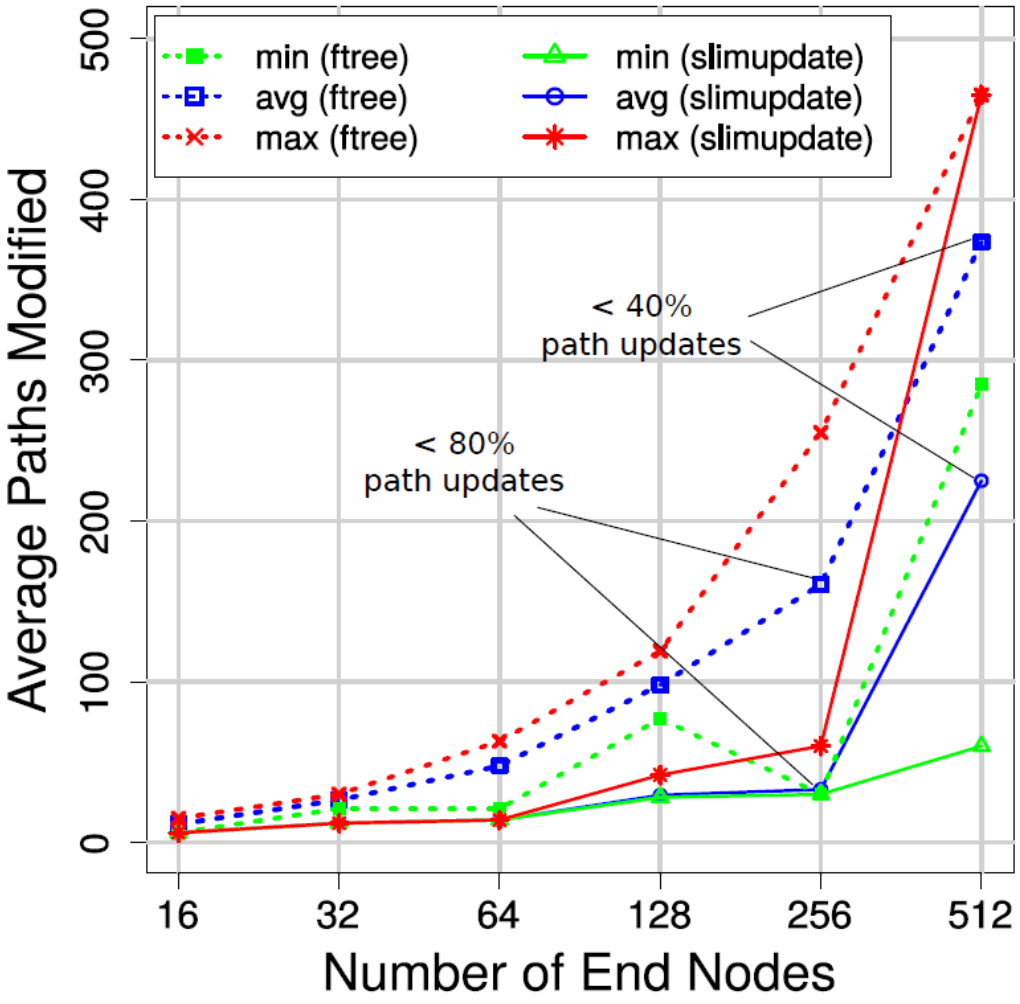
- The pFTree routing has two objectives in the order of priority
  - Well-balanced LFTs, Partition isolation
- Balancing using port counters
- Partition-isolation
  - Physical level, if enough resources available
  - Virtual Lanes



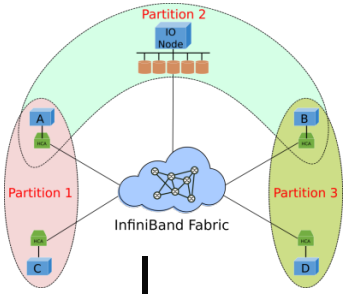
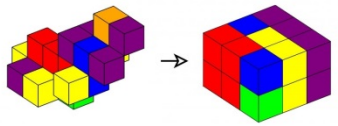


By employing intelligent techniques in the routing algorithm, cost of reconfiguration can be reduced significantly

- Reconfigured needed for:
  - Faults and Failures
  - Performance Maintenance
- Network Reconfiguration in IB
  - Static
  - Dynamic
  - Costly!
- Minimal Routing Update
  - Preserve existing paths
- SlimUpdate Routing algorithm
  - Fat-tree topologies
  - Low path updates

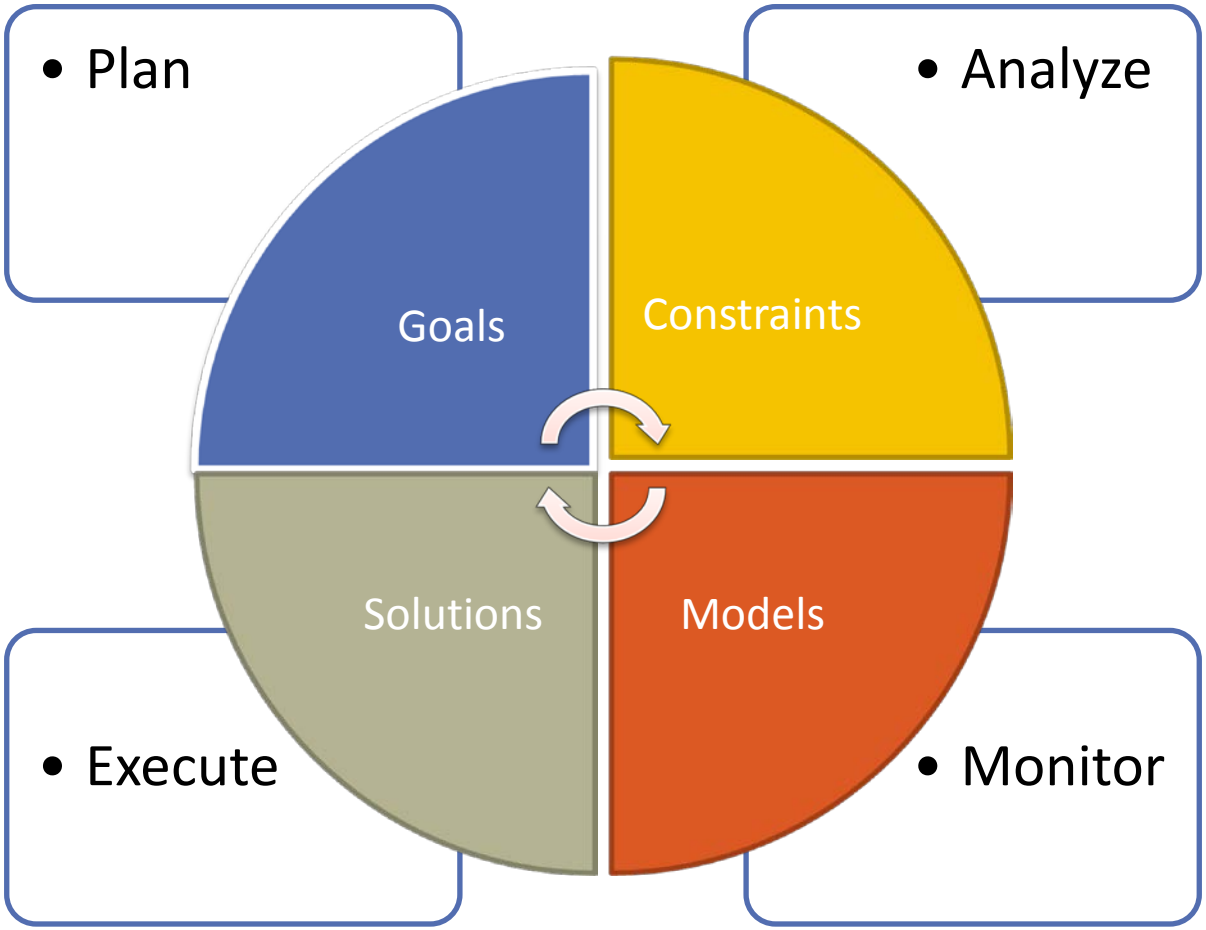


Big Picture: Enable smart network provisioning for the HPC clouds – We focus on four important components

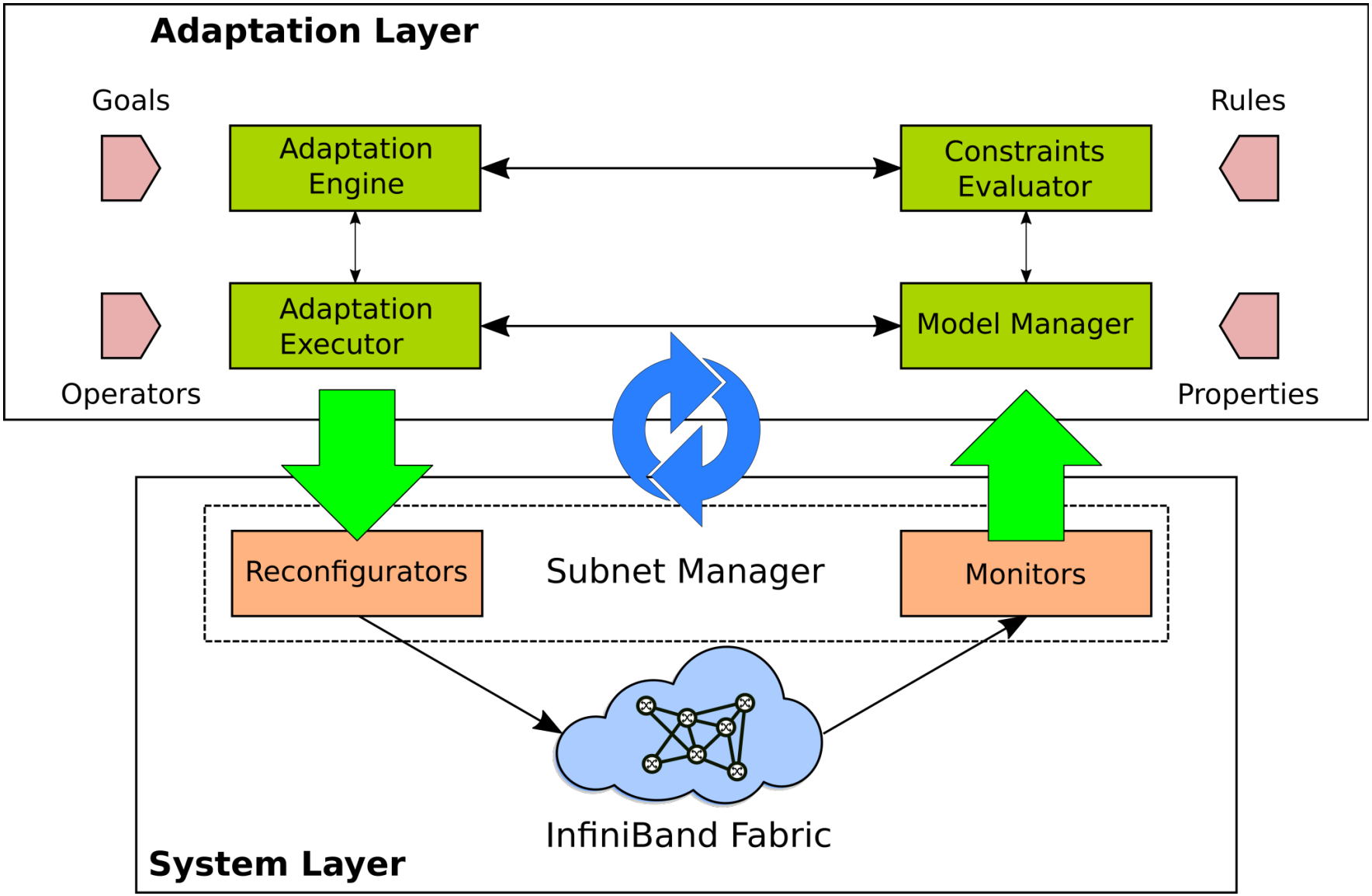
Smart Routing	Network Utilization	Network Isolation	Reconfiguration
			
Optimized Algorithms	Weighted Routing	Partition-aware Routing	Adjust for Load/Faults
Better Routes	Balanced Traffic	Multi-tenancy	Dynamic Optimizations

Monitor → Analyze → Plan → Execute

We aim to combine our pieces together to build a self-adaptive architecture for HPC clouds based on MAPE



We aim to combine our pieces together to build a self-adaptive architecture for HPC clouds based on MAPE



Reference: Garlan, David, et al. "Rainbow: Architecture-based self-adaptation with reusable infrastructure." *Computer* 37.10 (2004): 46-54.

Discussion Time

Questions / Challenges

Monitoring

- Monitoring mechanisms available in InfiniBand
  - SMAs notifications, probes
- Challenges
  - Dynamic workload
  - Virtualized environments
    - Shared-port, vSwitch, vPorts
  - Quick fault detection

Analysis

- Two dimensions
  - Is current configuration satisfying constraints?
  - Can we do better – optimze?
- Model-based approach enough?
  - Dynamic workload
- QoS, SLA violations
- Multi-tenancy, fairness

Planning

- Decision Making
  - Multi-criterion decision making (MCDM) problem
  - Define quality of routing?
    - Multiple same-quality routings available
- How to predict performance for the new reconfiguration
  - Test runs, simulations?

Execution

- Network reconfiguration in InfiniBand
  - Static reconfiguration
  - Dynamic reconfiguration
    - Avoid deadlocks!
    - For large subnets, interim routes
- Minimal routing update is desirable

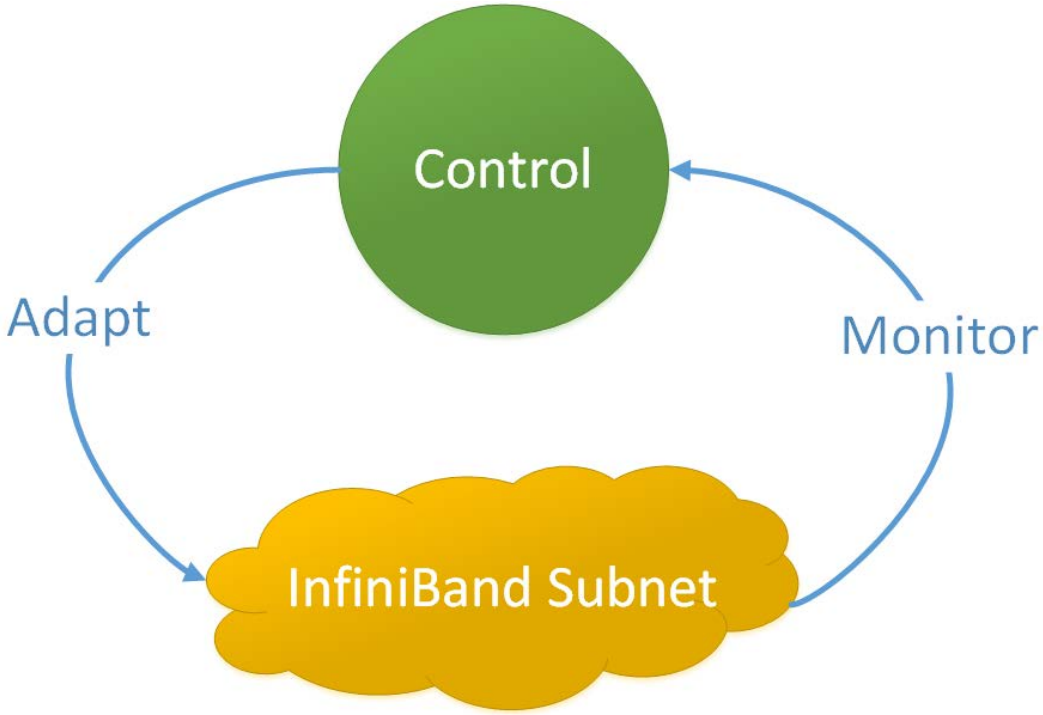
Putting all together

- Is this problem realistic?
- Are feedback-control systems feasible?
- What existing solutions we should look at?
- What policies can be applied in the available solutions?

In summary, a self-adaptive network architecture is needed to efficiently support dynamic HPC clouds based on IB

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

A Self-adaptive network architecture enabling dynamic HPC clouds



Thanks for your attention!

[simula . research laboratory]