OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

USING HIGH PERFORMANCE NETWORK INTERCONNECTS IN DYNAMIC ENVIRONMENTS

Vangelis Tasoulas

Simula Research Laboratory

[April 7th, 2016]

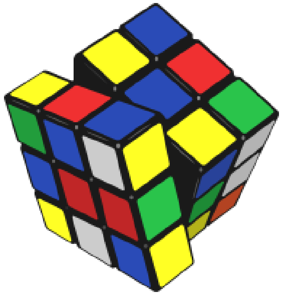
[simula . research laboratory]

ACKNOWLEDGEMENTS

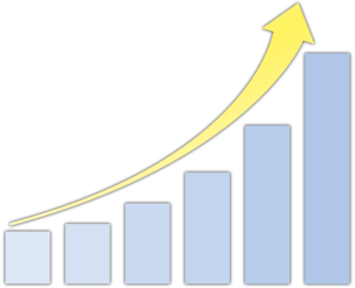
- Feroz Zahid, Ernst Gunnar Gran, Bjørn Dag Johnsen, Wei Lin Guay, Bartosz Bogdanski, Tor Skeie, Kyrre Begnum

- Mellanox for providing InfiniBand hardware for our research

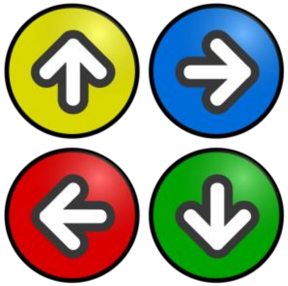
IN THIS PRESENTATION WE WILL GO THROUGH



Challenges



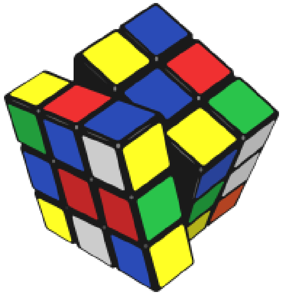
Virtualization and SA scalability



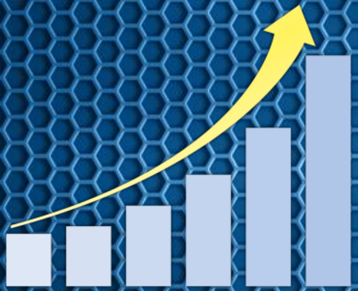
Routing algorithms

CHALLENGES IN DYNAMIC ENVIRONMENTS

- **Cloud environments are typically very dynamic by nature**
  - Pay-as-you-go on-demand service model
  - Multiple tenants
- **Resource fragmentation is very likely**
- **Need for re-optimization and reconfiguration by different means**
  - VM live migrations
  - Rerouting of traffic
- **OpenSM doesn’t scale well for very large subnets**
  - In dynamic environments there is much additional overhead from the different reconfiguration tasks
    - Scalable SA project in the works – our work is not competing, but complements



VIRTUALIZATION AND SA SCALABILITY



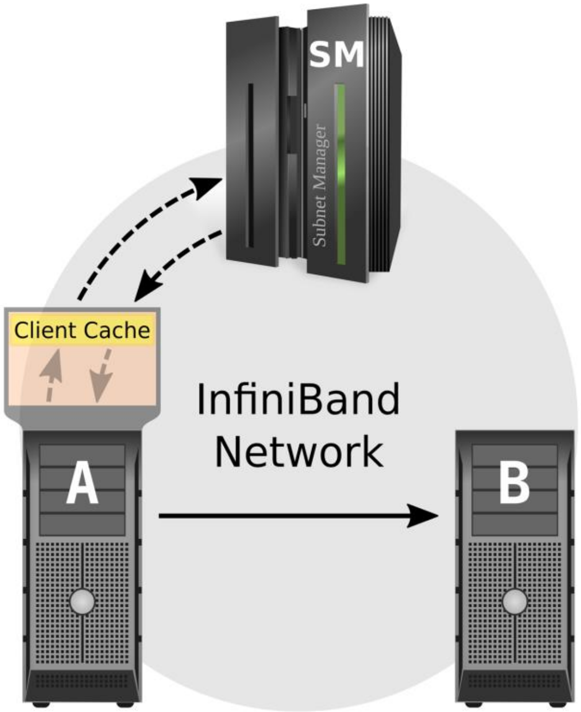
LIVE MIGRATIONS OF VIRTUAL MACHINES AND RECONFIGURATION WITH SIGNALING

- **Live Migrating VMs with the IB SR-IOV Shared-Port architecture**
- **Migrates the Alias-GUID (aGUID) associated with the VM**
- **The path information changes**
  - As a consequence of the LID-aGUID mapping change
  - The LID cannot be migrated in a Shared-Port architecture since it is shared between the hypervisor and the VMs
- **A signaling mechanism that uses the *repath* trap is implemented**
  - One signal is sent per hypervisor by the SM
  - The hypervisor distributes the signal to the rest of the VMs locally
- **This method works, but adds SM overhead**
  - Several signals for each migration are sent

[1] A Scalable Signalling Mechanism for VM Migration with SR-IOV over Infiniband, Guay et al., 2012 IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS)

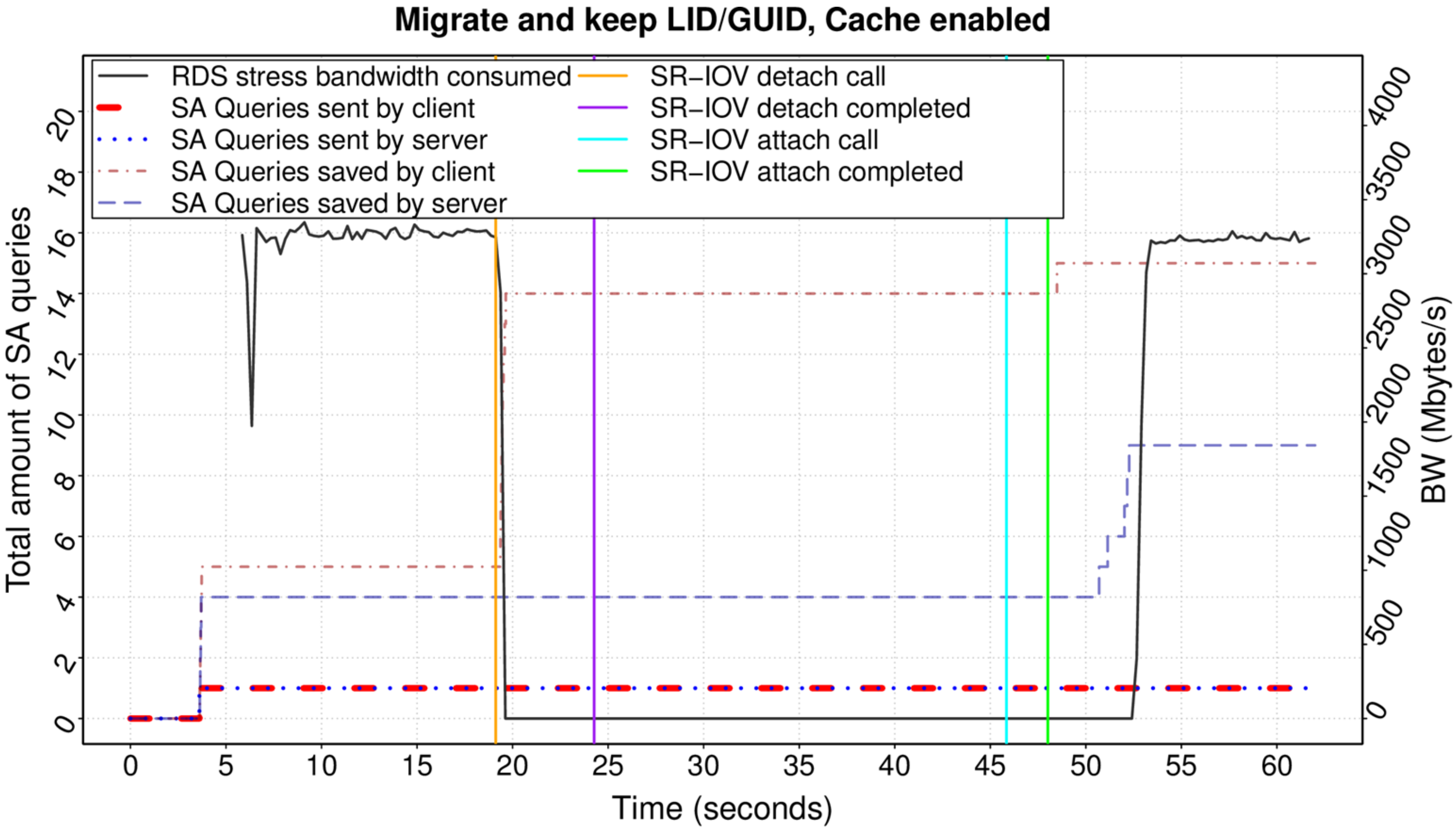
SA QUERY CACHING AND REUSE IN THE CONTEXT OF VM LIVE MIGRATION (1/2)

- **Each subnet entity (physical node/VMs) has a local SA path cache**
- **When a VM migrates, all three addresses associated with that VM are migrated as well**
  - For the prototype implementation, the guid2lid file was used to migrate the LID addresses, and the SM was restarted
- **The path information doesn't change after the migration**
- **Peers try to reconnect with the cached path information, and they succeed once the VM is operational after the migration**



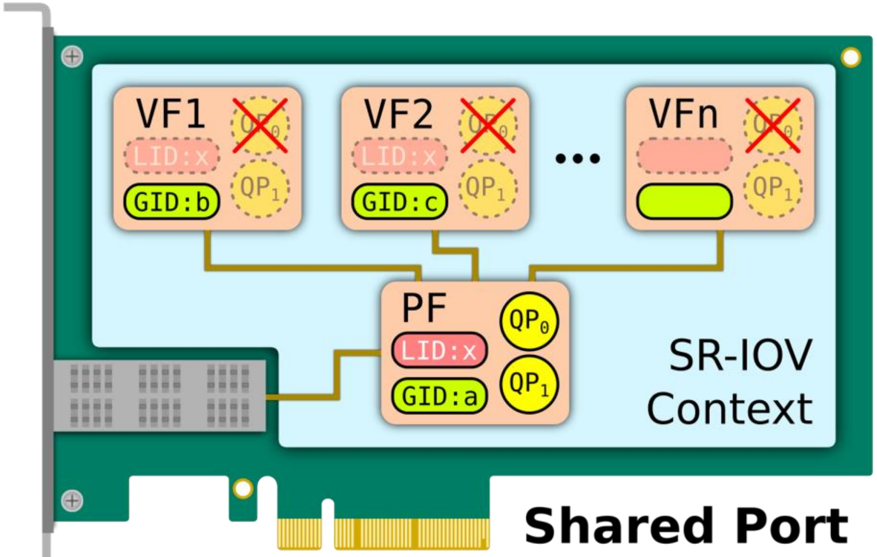
[2] A Novel Query Caching Scheme for Dynamic InfiniBand Subnets, Tasoulas et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid)

SA QUERY CACHING AND REUSE IN THE CONTEXT OF VM LIVE MIGRATION (2/2)

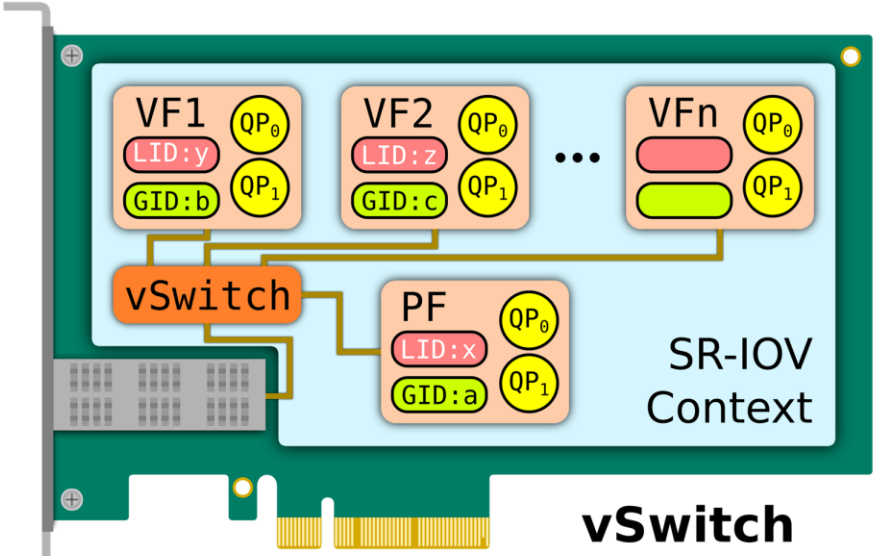


[2] A Novel Query Caching Scheme for Dynamic InfiniBand Subnets, Tasoulas et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid)

TOWARDS AN SR-IOV VSWITCH ARCHITECTURE (1/2)



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

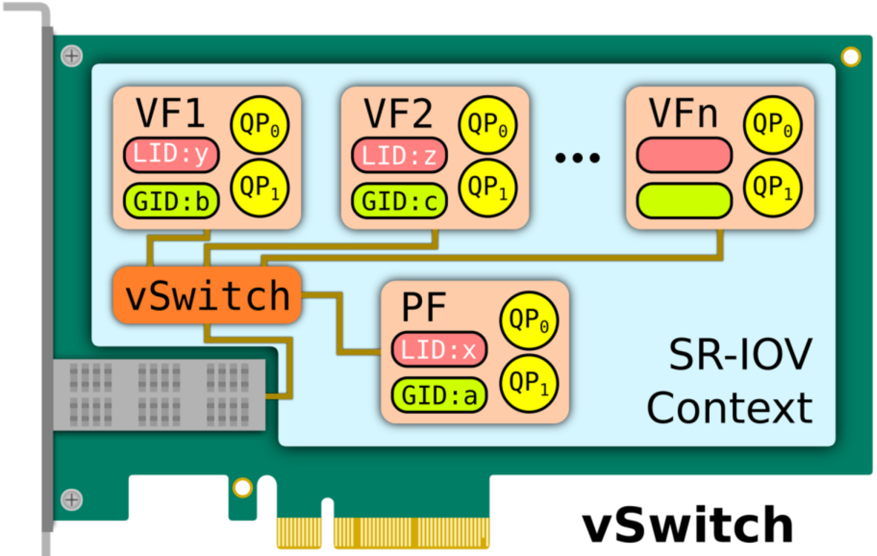


PF: Handled by Hypervisor **VFs**: Assigned on VMs

[3] Towards the InfiniBand SR-IOV vSwitch Architecture, Tasoulas et al., 2015 IEEE International Conference on Cluster Computing (CLUSTER)

TOWARDS AN SR-IOV VSWITCH ARCHITECTURE (2/2)

- **An SR-IOV vSwitch can solve some challenges faced by the Shared-Port:**
  - No need for additional signaling when migrating VMs
  - Each VM is directly visible to the SM and it can even have its own routes in the subnet
- **With one disadvantage:**
  - Bloating of the limited LID space
- **We propose two implementations with different scalability characteristics.**
  - Prepopulated VF LIDs
  - Dynamic VF LID assignment
- **The vPort model was proposed last year in OFA workshop**
  - Improves the shared-port, but still cannot solve the two aforementioned challenges



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

[3] Towards the InfiniBand SR-IOV vSwitch Architecture, Tasoulas et al., 2015 IEEE International Conference on Cluster Computing (CLUSTER)

ROUTING

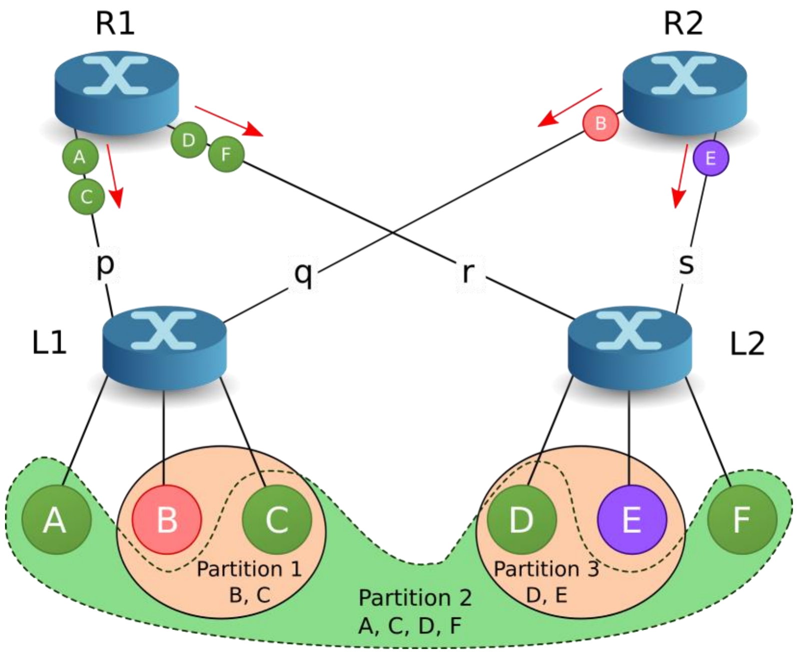
PARTITION-AWARE ROUTING (1/3)

- **In Multi-tenant Infrastructures**
  - Tenants should experience predictable network performance unaffected by the workload of other tenants
- **Network Isolation through Partitioning**
  - Each tenant is assigned a partition
  - Inter-partition communication not allowed
- **But routing is done without considering partitions**
  - Degraded load-balancing
  - Performance interference among partitions
- **Partition-aware Routing**
  - Well-balanced LFTs with partition isolation
  - Physical link level isolation if resources available
  - Use virtual lanes to complement

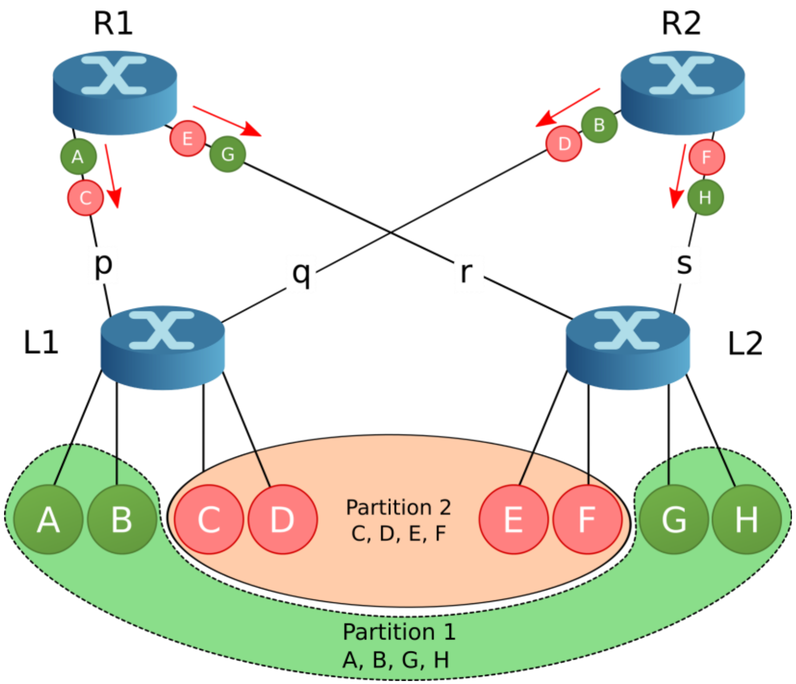
[4] Partition-Aware Routing to Improve Network Isolation in InfiniBand Based Multi-tenant Clusters, Zahid et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid '15).

PARTITION-AWARE ROUTING (2/3)

Traditional Fat-Tree Routing Issues in Multi-tenant Networks



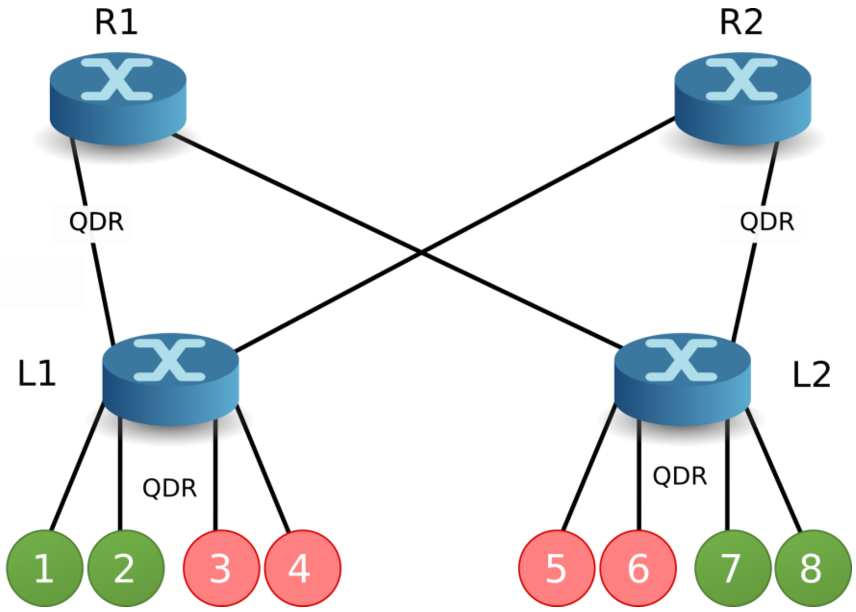
Degraded Load Balancing

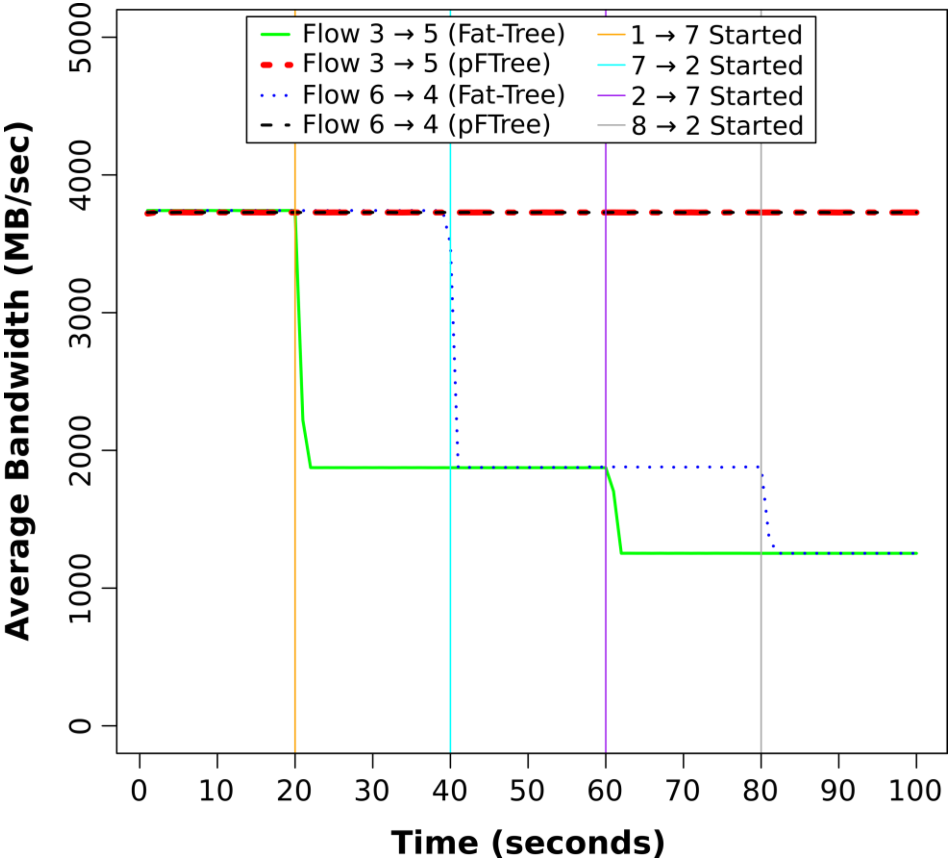


No Isolation Between Partitions

[4] Partition-Aware Routing to Improve Network Isolation in InfiniBand Based Multi-tenant Clusters, Zahid et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid '15).

PARTITION-AWARE ROUTING (3/3)



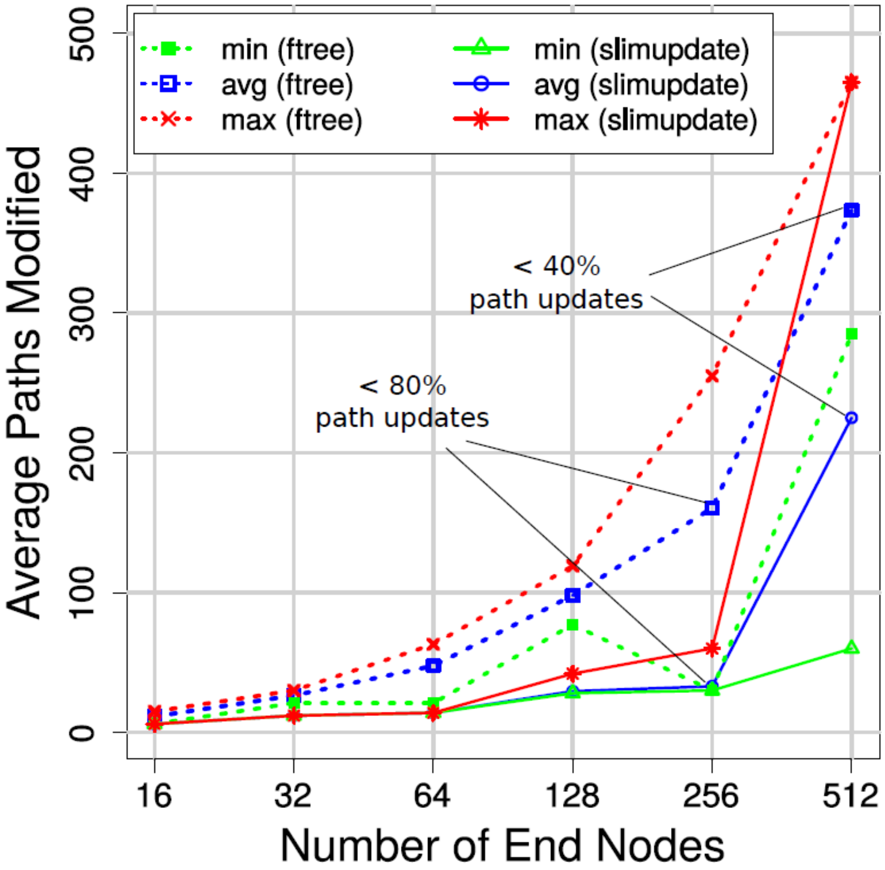


Sample Oversubscribed Topology

[4] Partition-Aware Routing to Improve Network Isolation in InfiniBand Based Multi-tenant Clusters, Zahid et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid '15).

COMPACT NETWORK RECONFIGURATION

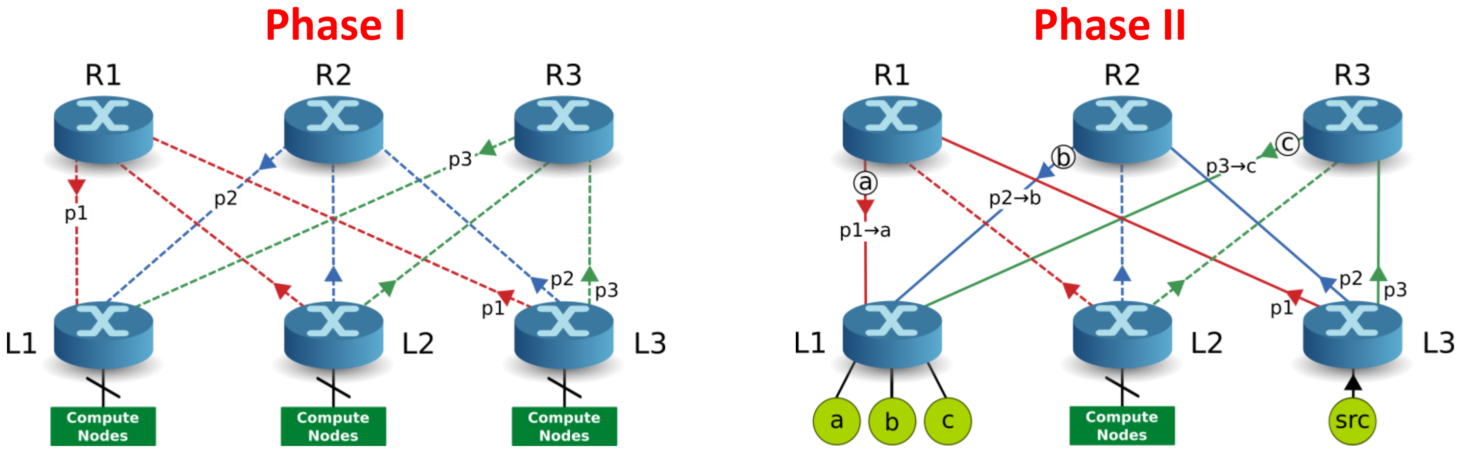
- **Network reconfiguration is required for**
  - Faults and failures
  - Maintaining performance
- **Current network reconfiguration in IB**
  - Static
  - Dynamic
  - Costly, due to large number of path updates
- **Minimal Routing Update**
  - Consider existing paths in the network
  - Minimal number of path updates



[5] Minimal Routing Update for Performance-based Reconfigurations in Fat-Trees, Zahid et al., 2015 1st IEEE International Workshop on High-Performance Interconnection Networks (HiPINEB '15).

METABASE-AIDED ROUTING FOR PERFORMANCE BASED RECONFIGURATIONS

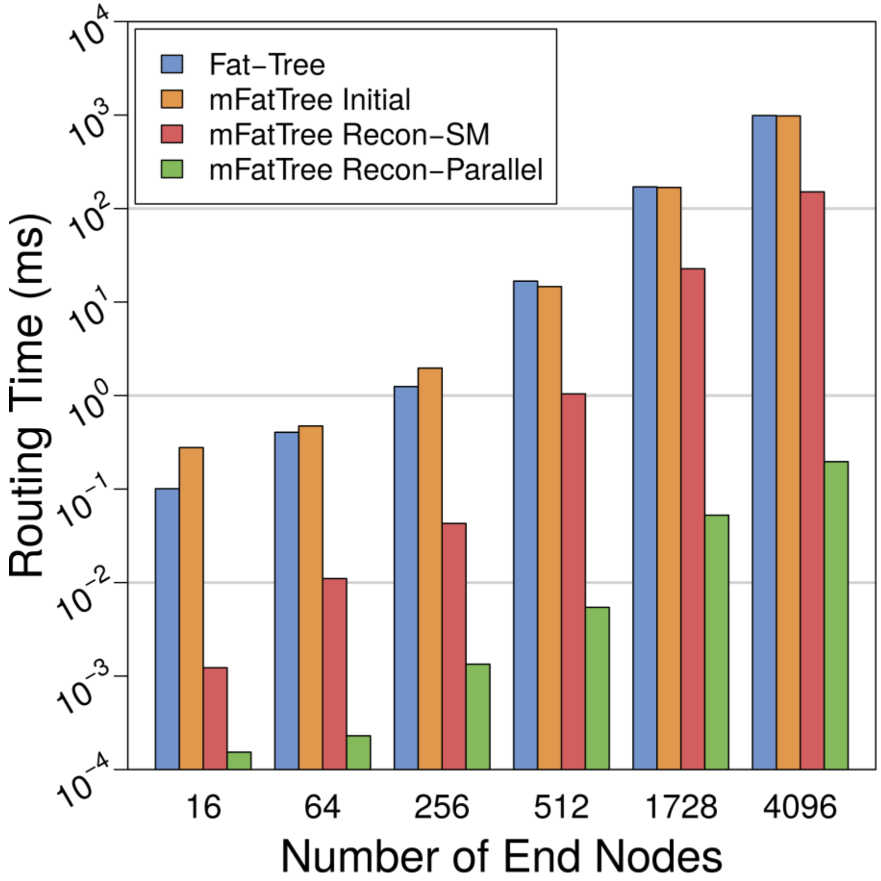
- **Fast network reconfiguration mechanism based on**
  - Two-phase routing
  - Calculation of paths, allocation of calculated paths to actual destinations
- **For performance-based reconfigurations**
  - Routing calculation is avoided
- **For virtualized IB subnets**
  - Quick reconfiguration on VM start/stop/migration


Phase I

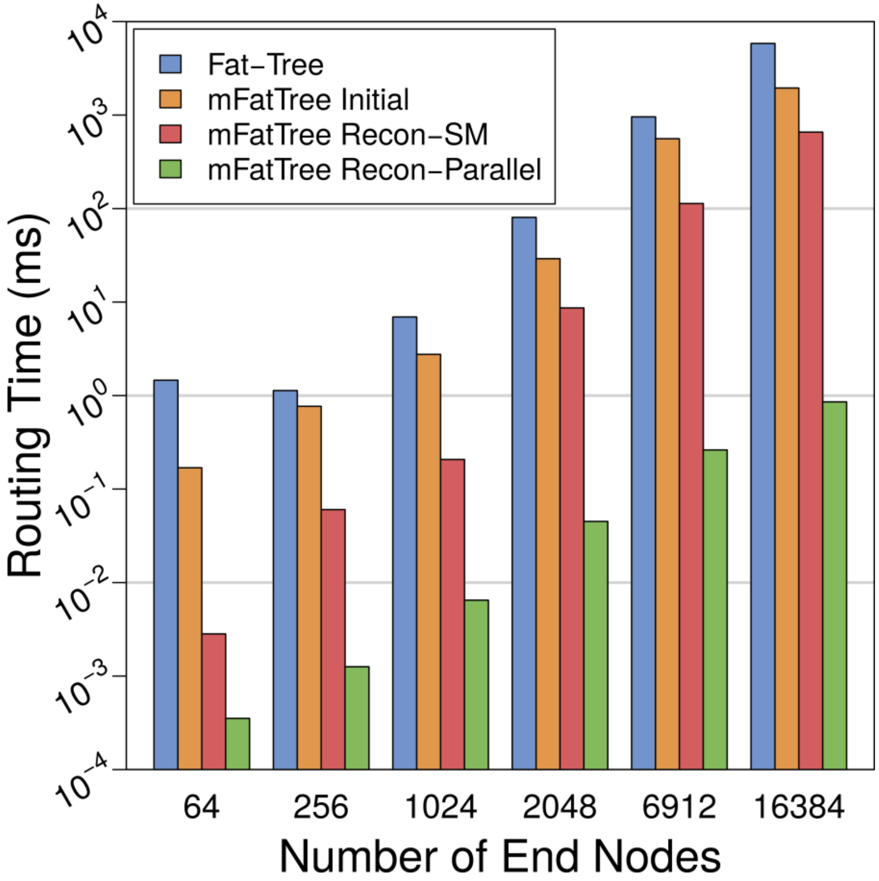
Phase II

[6] Compact Network Reconfiguration in Fat-Trees, Zahid et al., 2016 Under review in an International Journal.

METABASE-AIDED ROUTING FOR PERFORMANCE BASED RECONFIGURATIONS



Non-oversubscribed



Oversubscription = 4

[6] Compact Network Reconfiguration in Fat-Trees, Zahid et al., 2016 Under review in an International Journal.

OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Vangelis Tasoulas

Simula Research Laboratory

[simula . research laboratory]