# ExaComm 2016

## About Management of Exascale Systems

Tor Skeie

Simula Research Laboratory / University of Oslo

[ June 23rd, 2016 ]

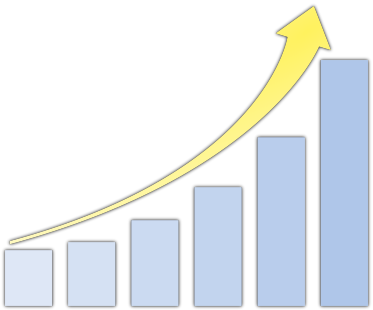
[ simula . research laboratory ]

# ACKNOWLEDGEMENTS


- Vangelis Tasoulas (Simula Research Laboratory)
- Feroz Zahid (University of Oslo / Simula)
- Ernst Gunnar Gran (Simula / University of Oslo )
- Jesús Camacho (Fabriscale Technologies)

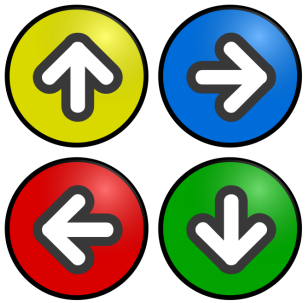
- Mellanox for supporting us with InfiniBand hardware, and in general being very supportive

# THIS PRESENTATION WILL BE ABOUT



## 1. Achieving scalability with InfiniBand

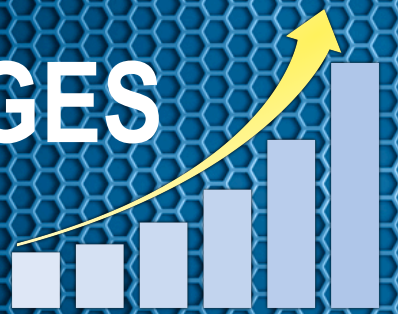
- The limited scalability of InfiniBand
- The Scalable SA solution
- A novel path caching scheme for dynamic environment



## 2. Efficient routing-aware reconfiguration

- Some efficient reconfiguration techniques
- Hierarchical reconfiguration

# THE SCALABILITY CHALLENGES FACED BY OPENSM



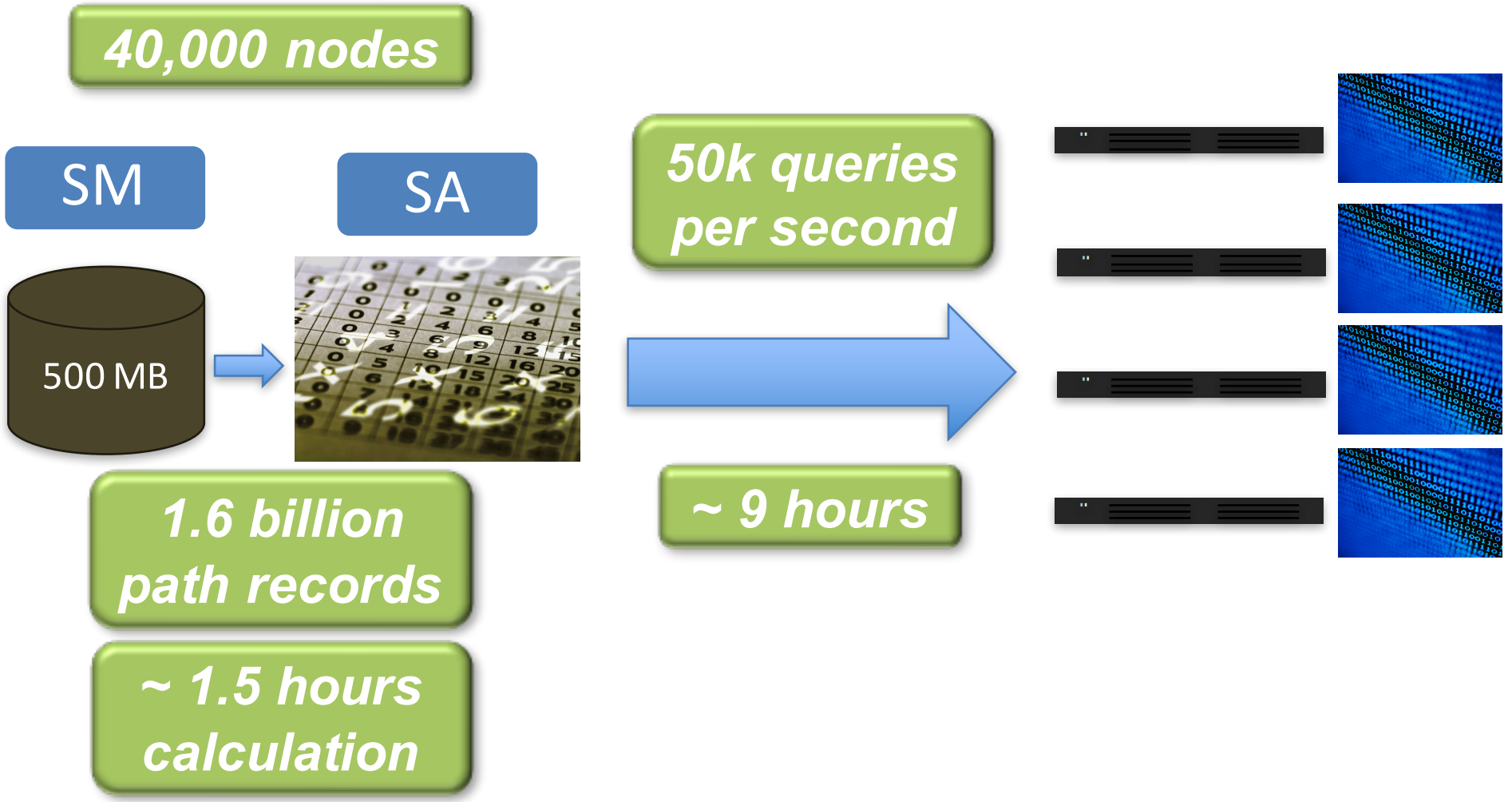
# THE PROBLEM

***$n^2$ SA load***

- **SA queried for every connection**
- **Communication between all nodes creates an $n^2$ load on the SA**
  - In InfiniBand architecture (IBA), SA is a centralized entity
- **Other $n^2$ scalability issues**
  - Name to address (DNS)
    - Mainly solved by a hosts file
  - IP address translation
    - Relies on ARPs

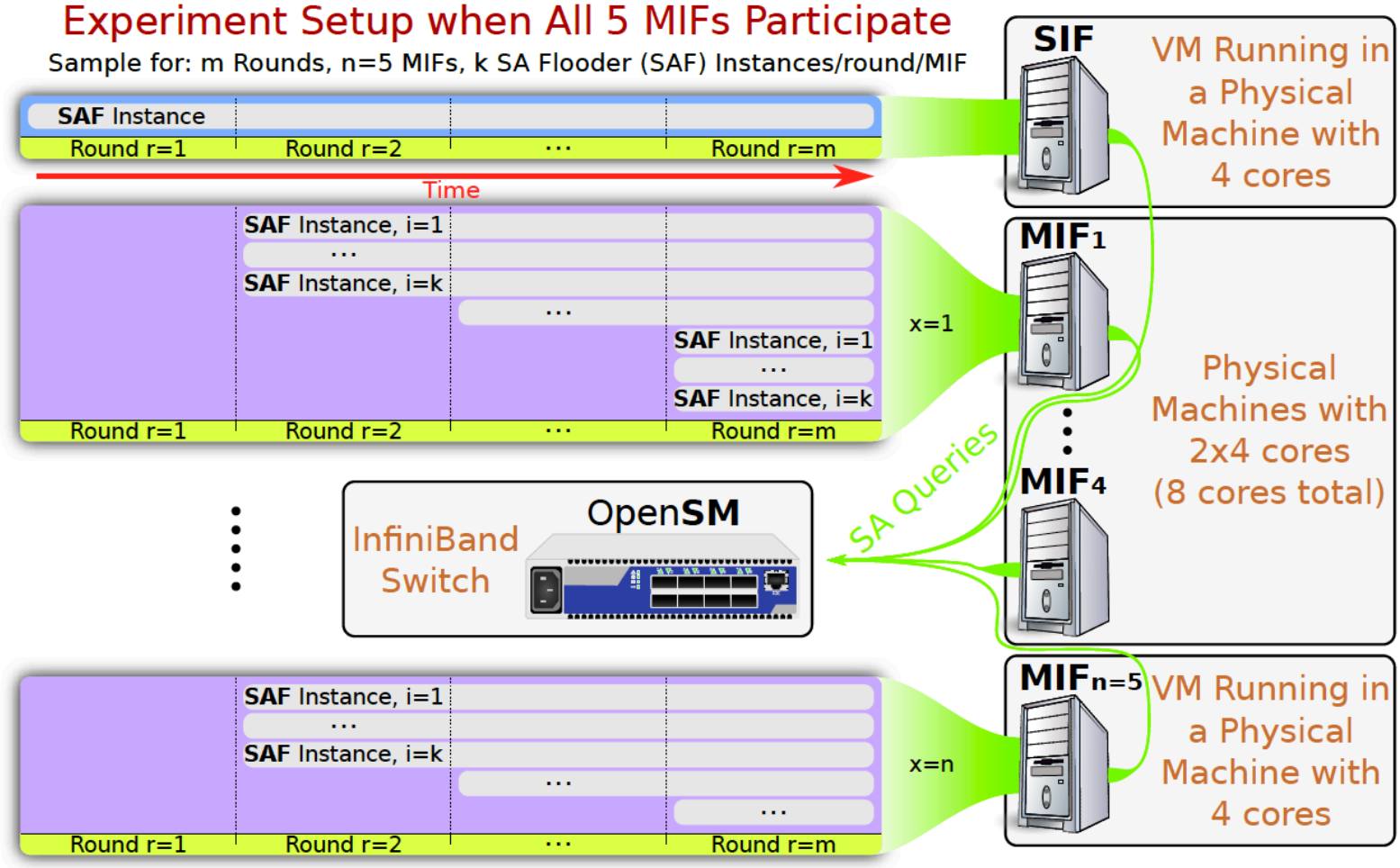
Source: OFA Dev Workshop 2014

# ANALYSIS I



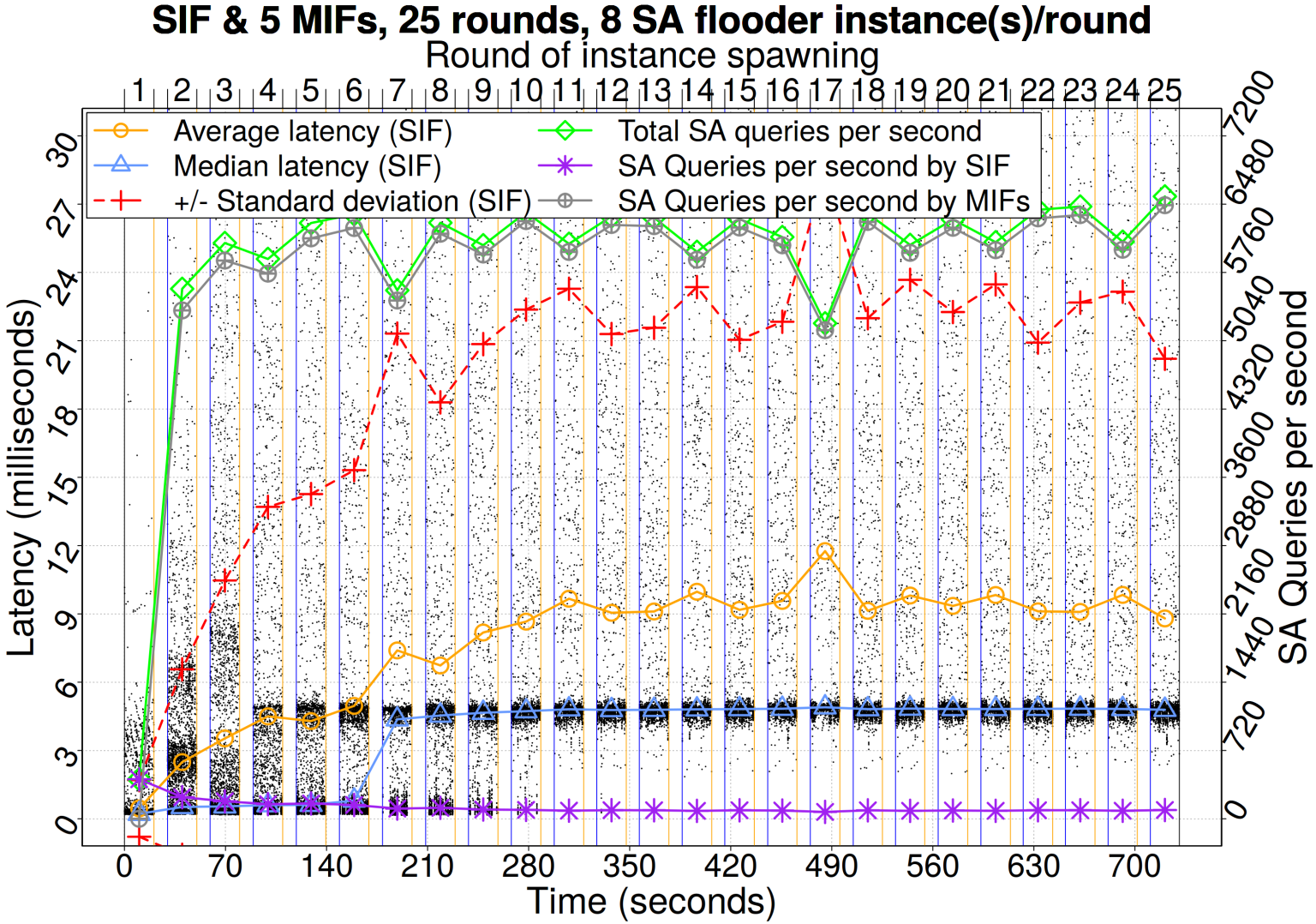

Source: OFA Dev Workshop 2014

# ANALYSIS II: CHALLENGING OPENSM WITH A SMALL TESTBED



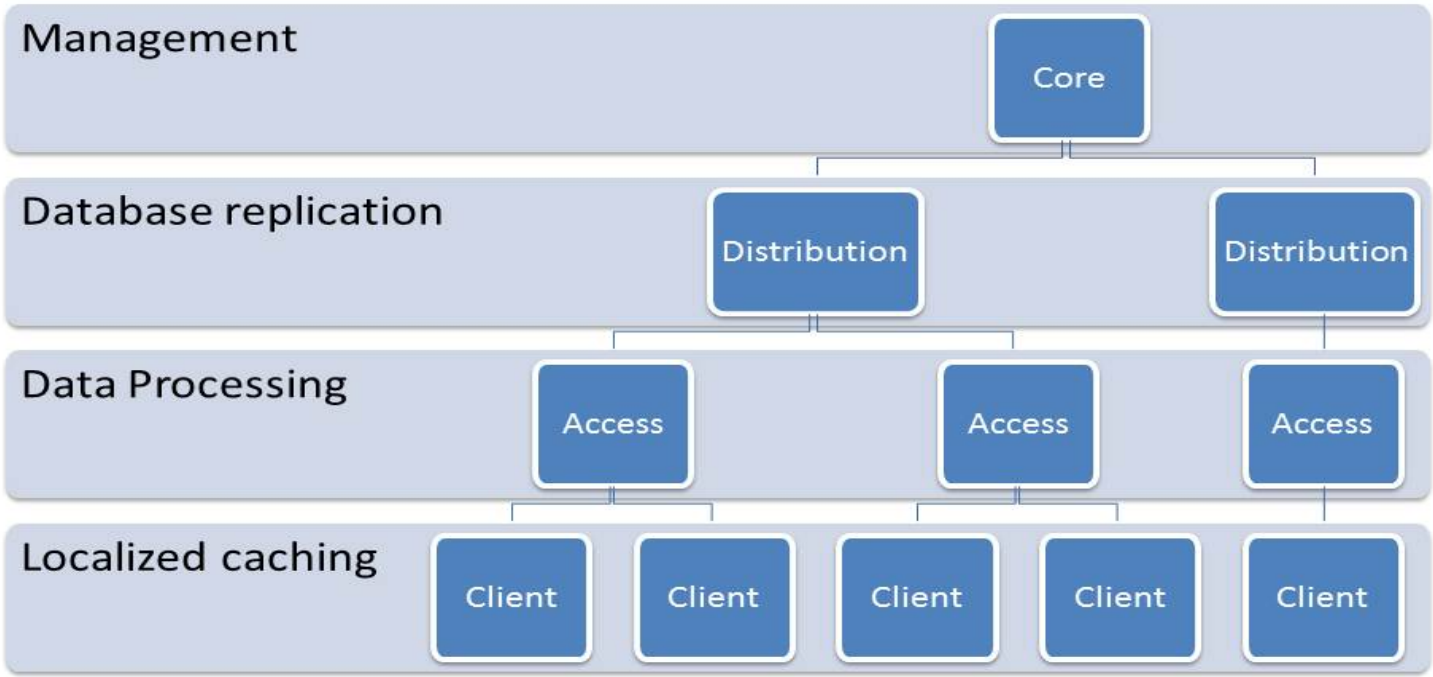

[1] A Novel Query Caching Scheme for Dynamic InfiniBand Subnets, Tasoulas et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid)

# OPENSM SCALABILITY CHALLENGES DEMONSTRATED



# THE SCALABLE SA (SSA) SOLUTION

## SSA Architecture




**Source: OFA Dev Workshop 2014**

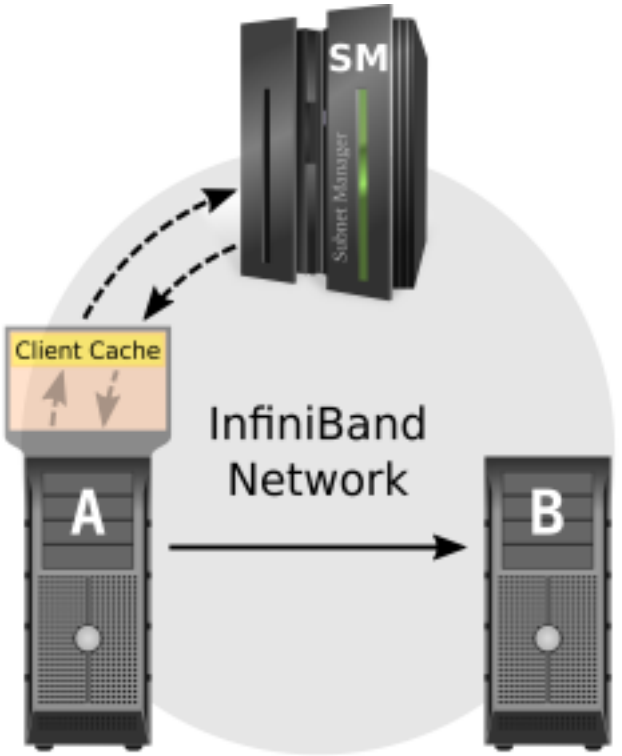
# CHALLENGES IN DYNAMIC ENVIRONMENTS

- **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 does not 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

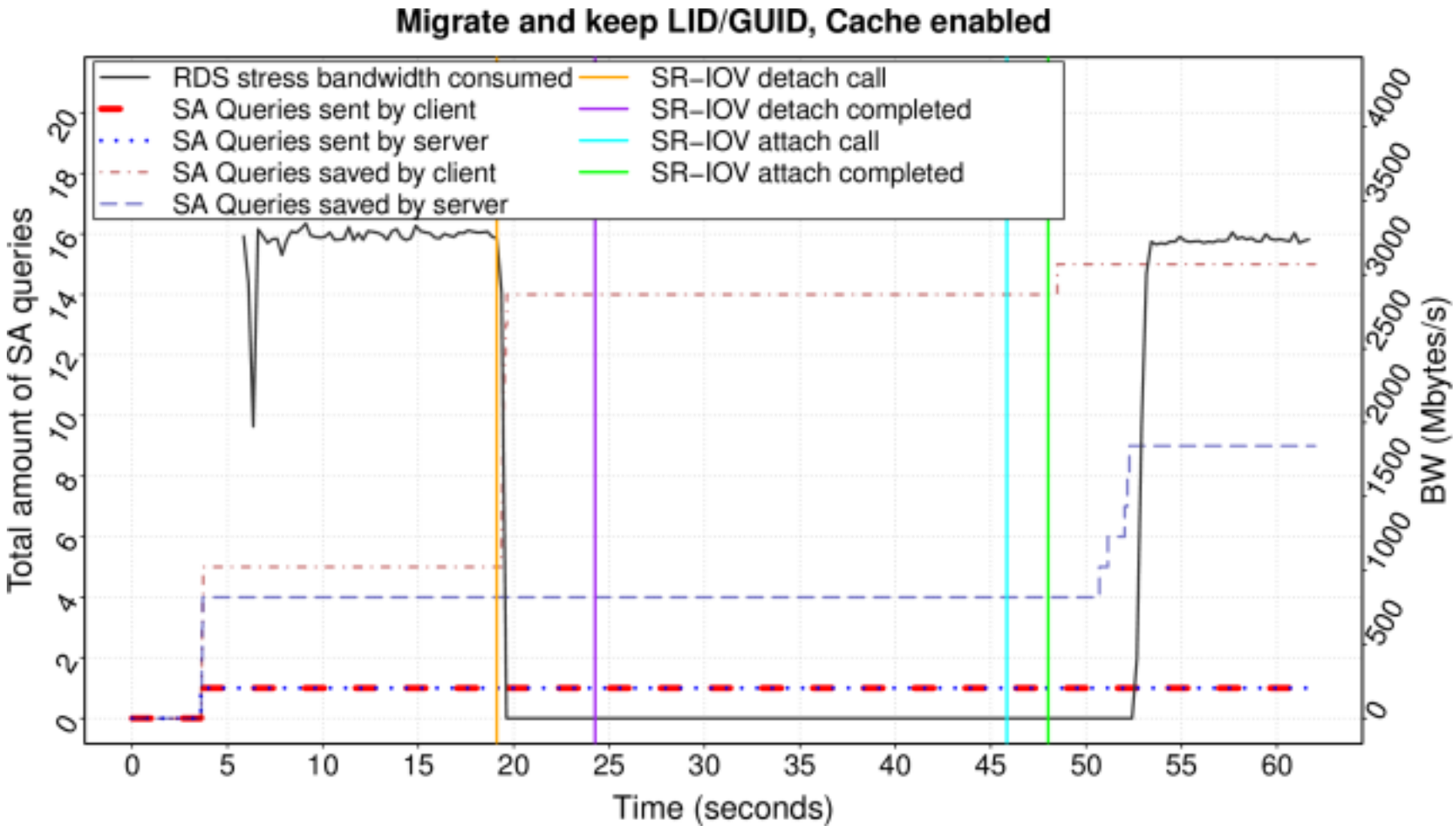
# SA QUERY CACHING AND REUSE IN THE CONTEXT OF VM LIVE MIGRATION (1/2)

- **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 does not change after the migration**
- **Peers try to reconnect with the cached path information, and they succeed once the VM is operational after the migration**



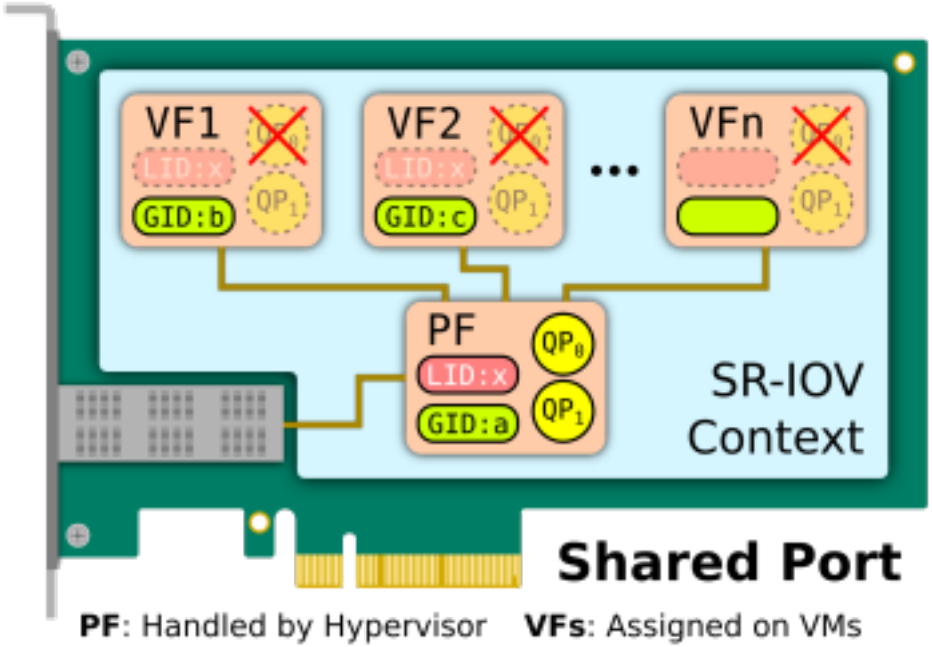

[1] A Novel Query Caching Scheme for Dynamic InfiniBand Subnets, Tasoulas et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid)

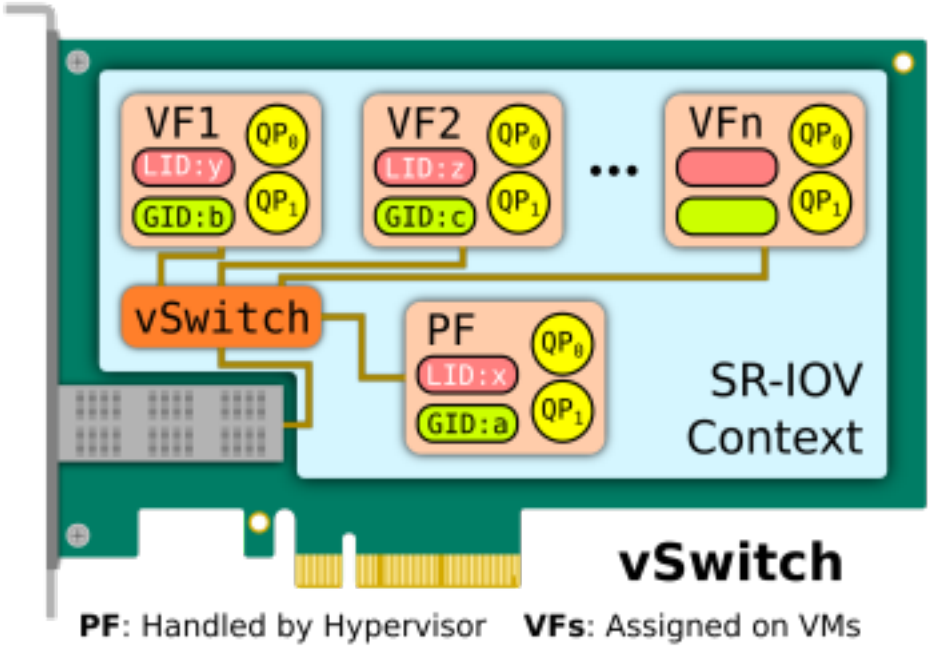
# SA QUERY CACHING AND REUSE IN THE CONTEXT OF VM LIVE MIGRATION (2/2)




[1] 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




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

# EFFICIENT ROUTING-AWARE RECONFIGURATION

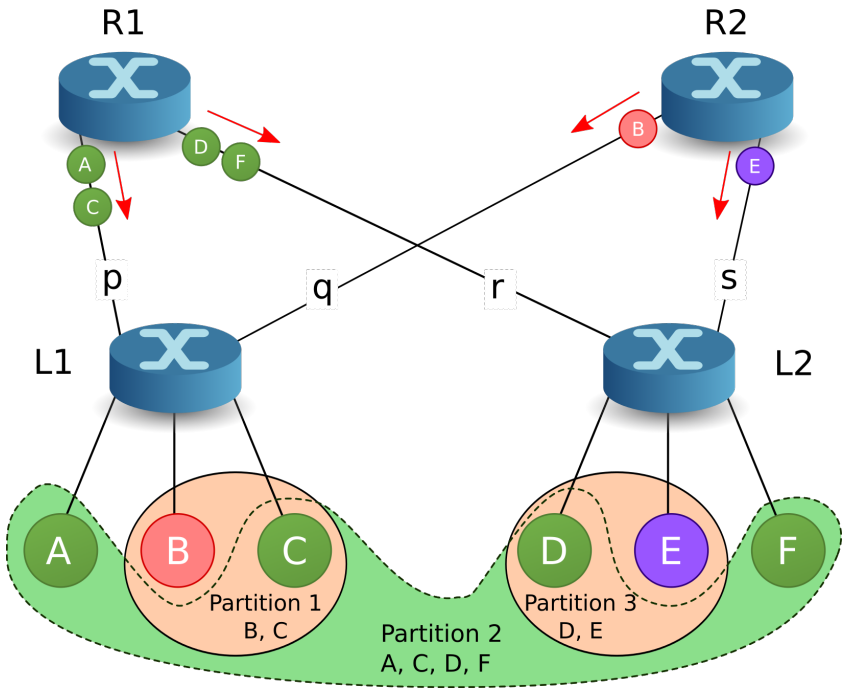
# PARTITION-AWARE ROUTING (1/3)

- **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

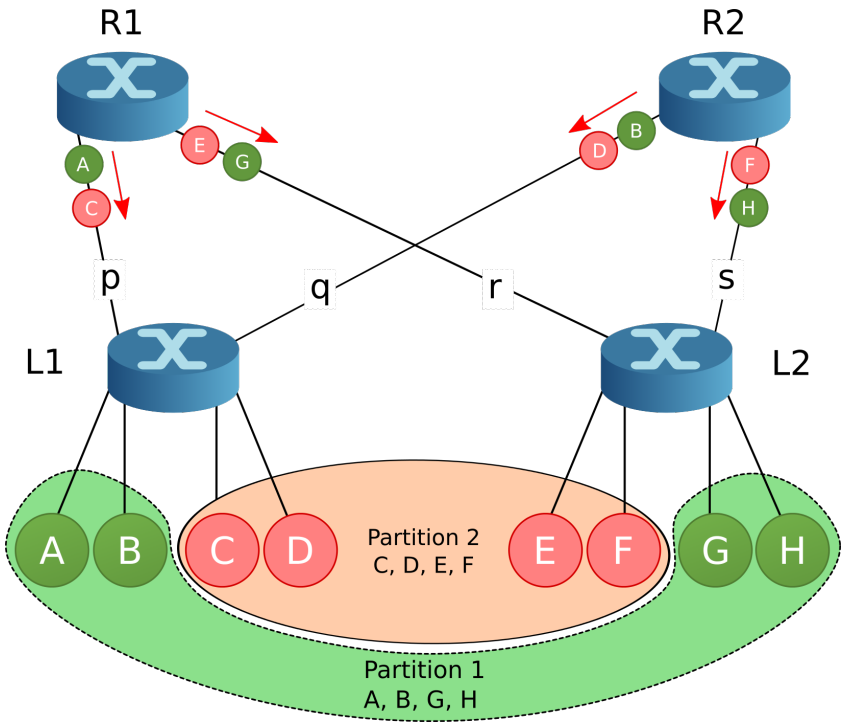
**[3] Partition-Aware Routing to Improve Network Isolation in InfiniBand Based Multi-tenant Clusters, Zahid et al., 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing (CCGrid '15).**

# PARTITION-AWARE ROUTING (2/3)

## Traditional Fat-Tree Routing in Multi-tenant Networks



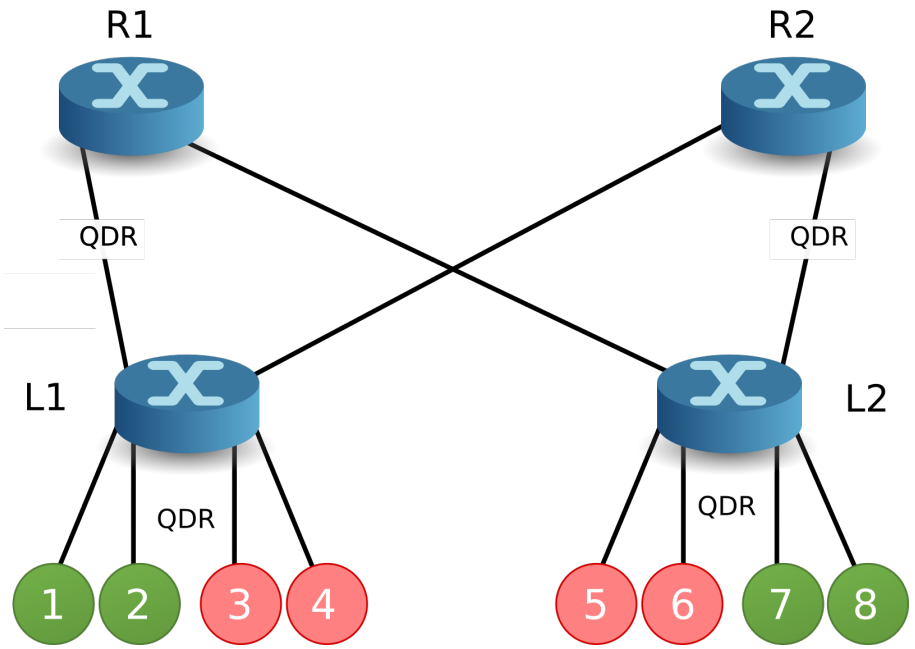

**Degraded Load Balancing**

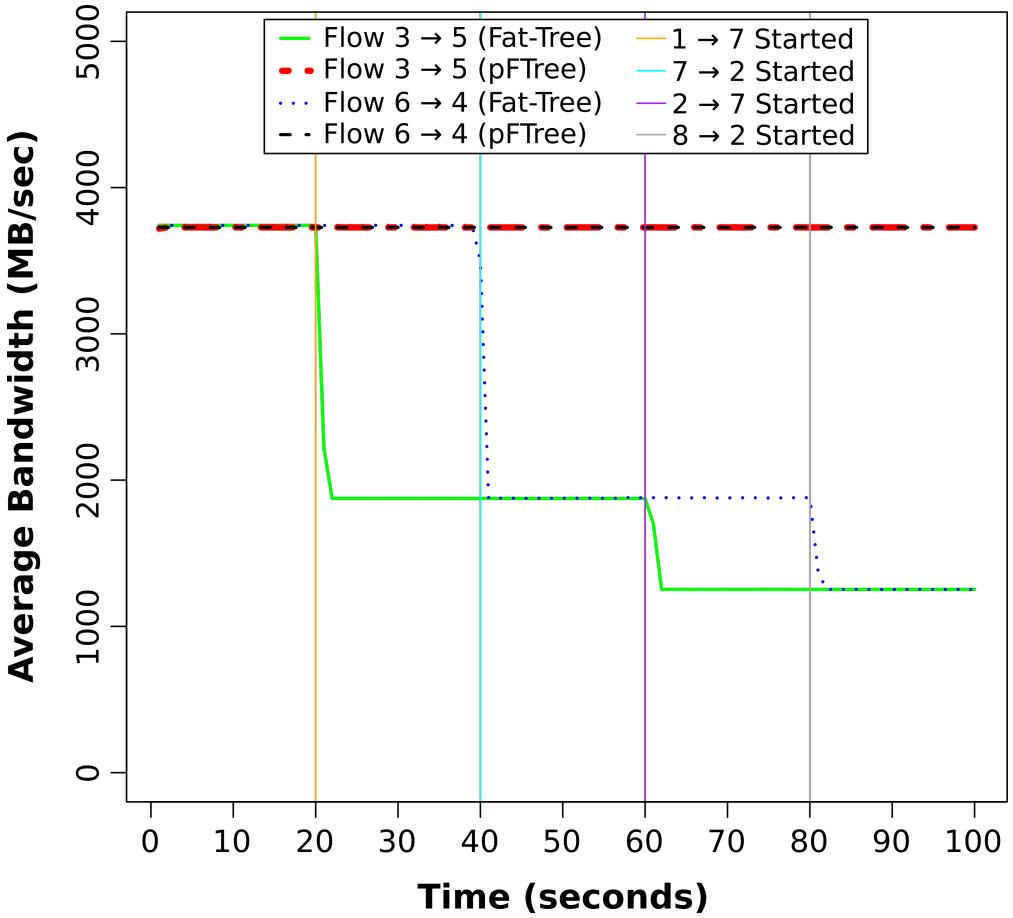



**No Isolation Between Partitions**

[3] 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)



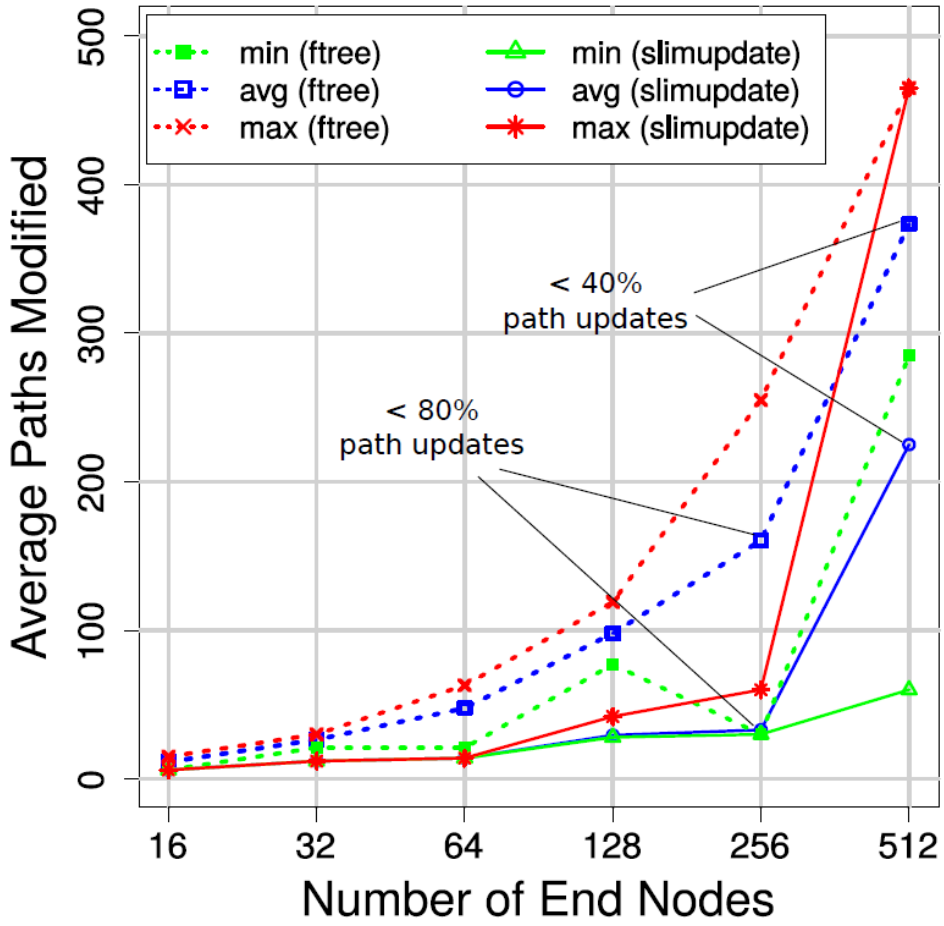




**Sample Oversubscribed (2:1) Topology**

[3] 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

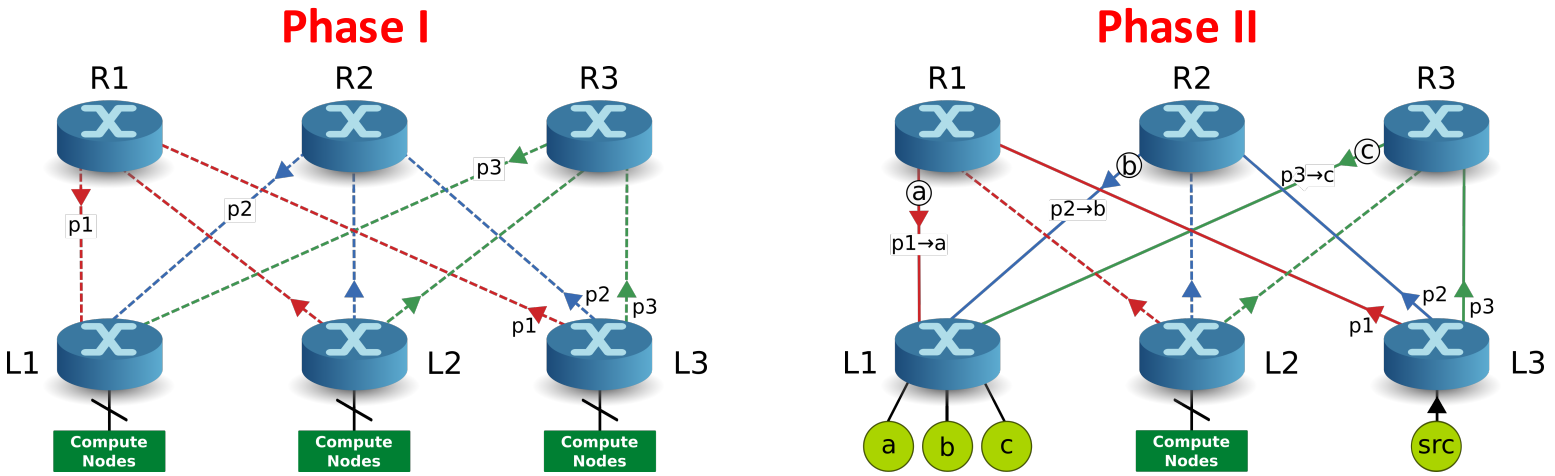
- **Network reconfiguration is required for**
  - Faults and failures
  - Maintaining performance
- **Current network reconfiguration in IB**
  - Static
  - Dynamic
  - Costly, due to large number of path updates
- **Minimal Routing Update**
  - Consider existing paths in the network
  - Minimal number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




[4] 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).

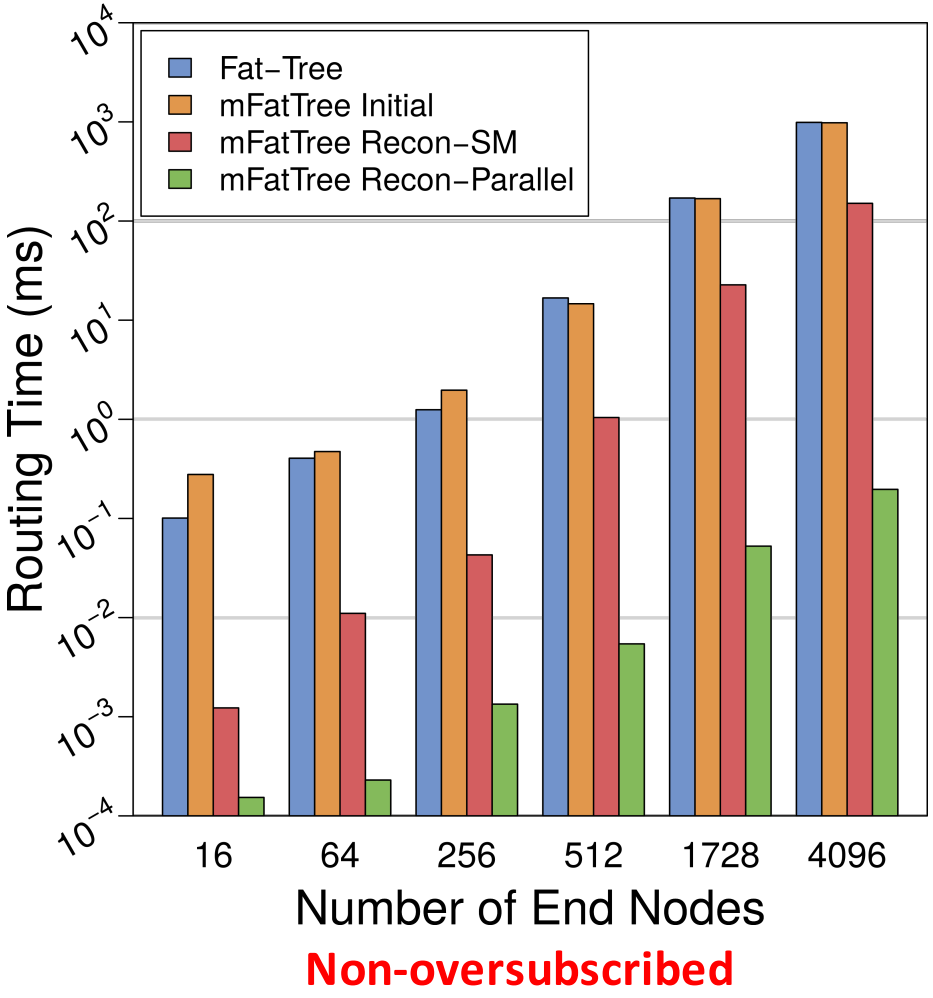
# METABASE-AIDED ROUTING FOR PERFORMANCE DRIVEN RECONFIGURATIONS

- **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

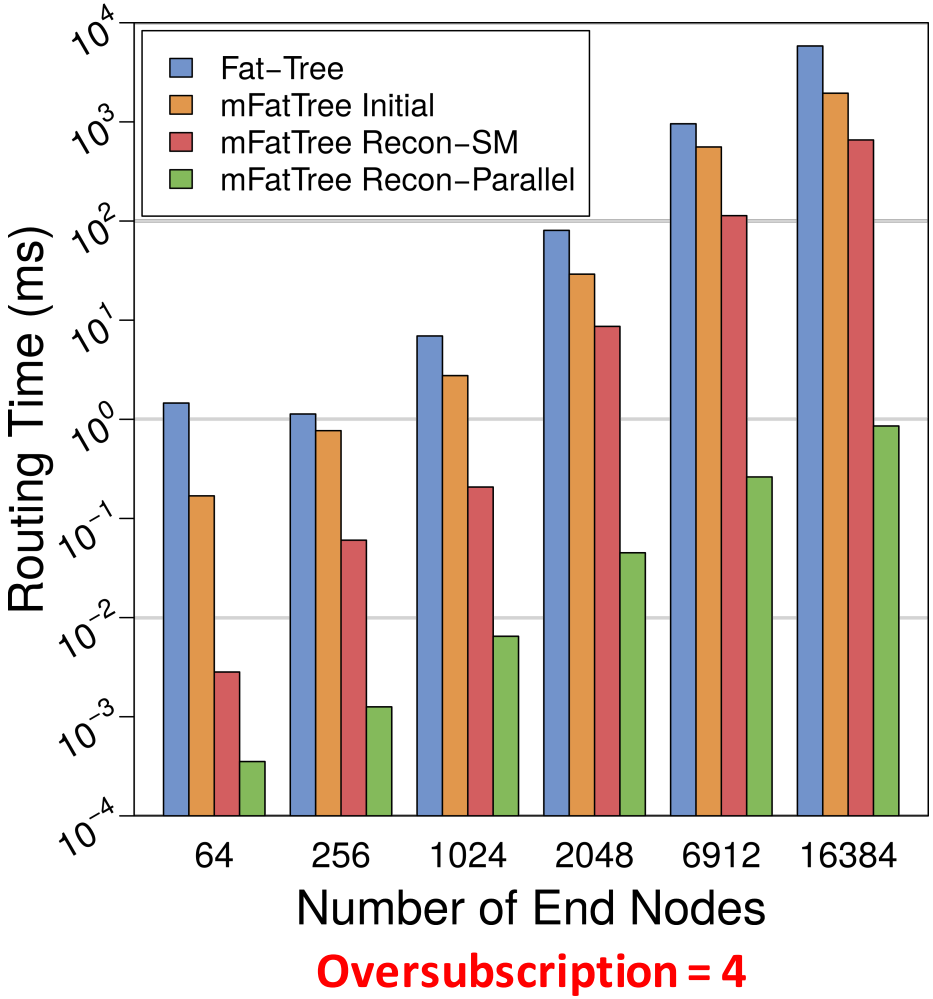



[5] Compact Network Reconfiguration in Fat-Trees, Zahid et al. Accepted to Journal of Supercomputing, 2016.

# METABASE-AIDED ROUTING FOR PERFORMANCE DRIVEN RECONFIGURATIONS



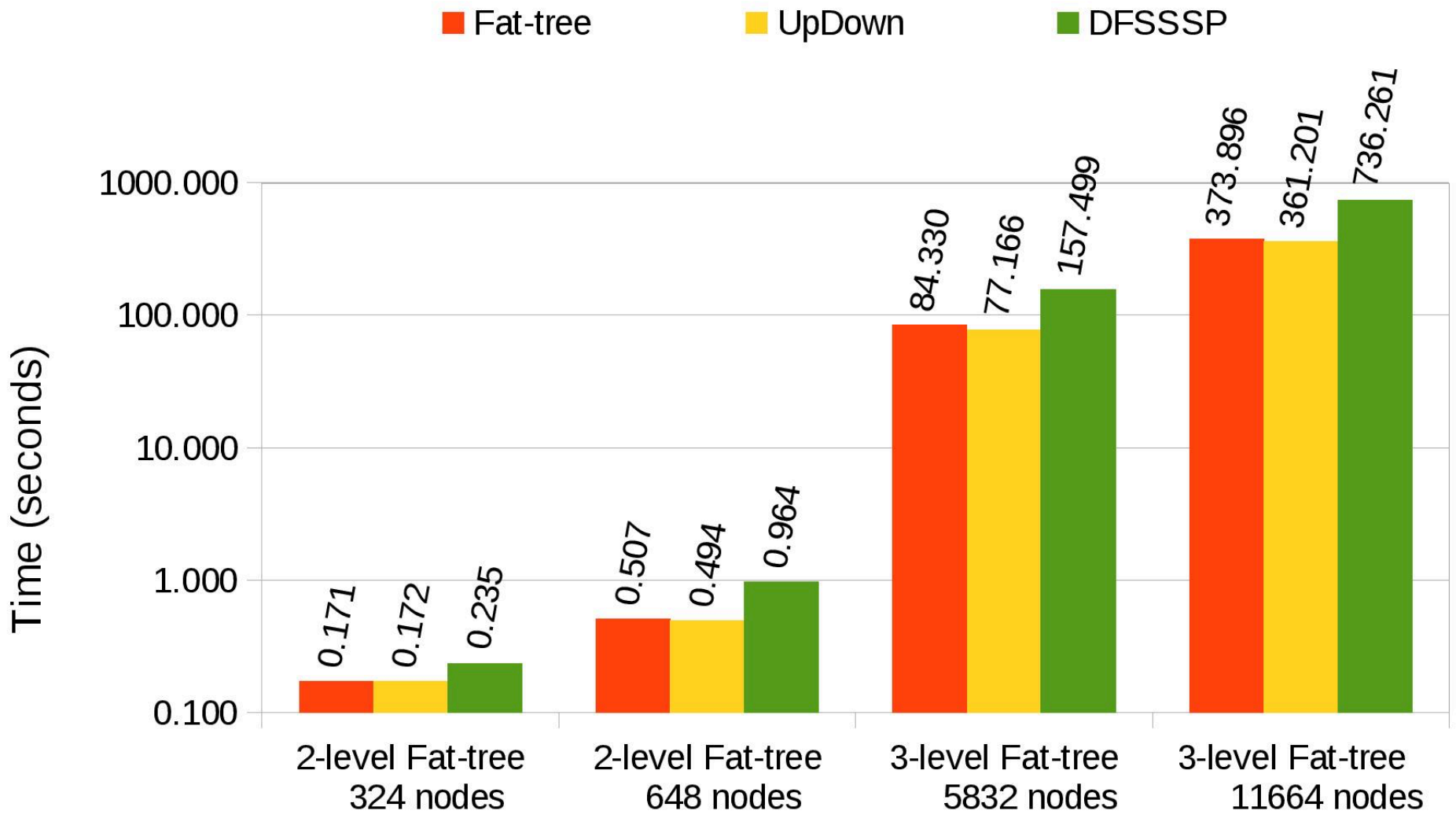

Non-oversubscribed



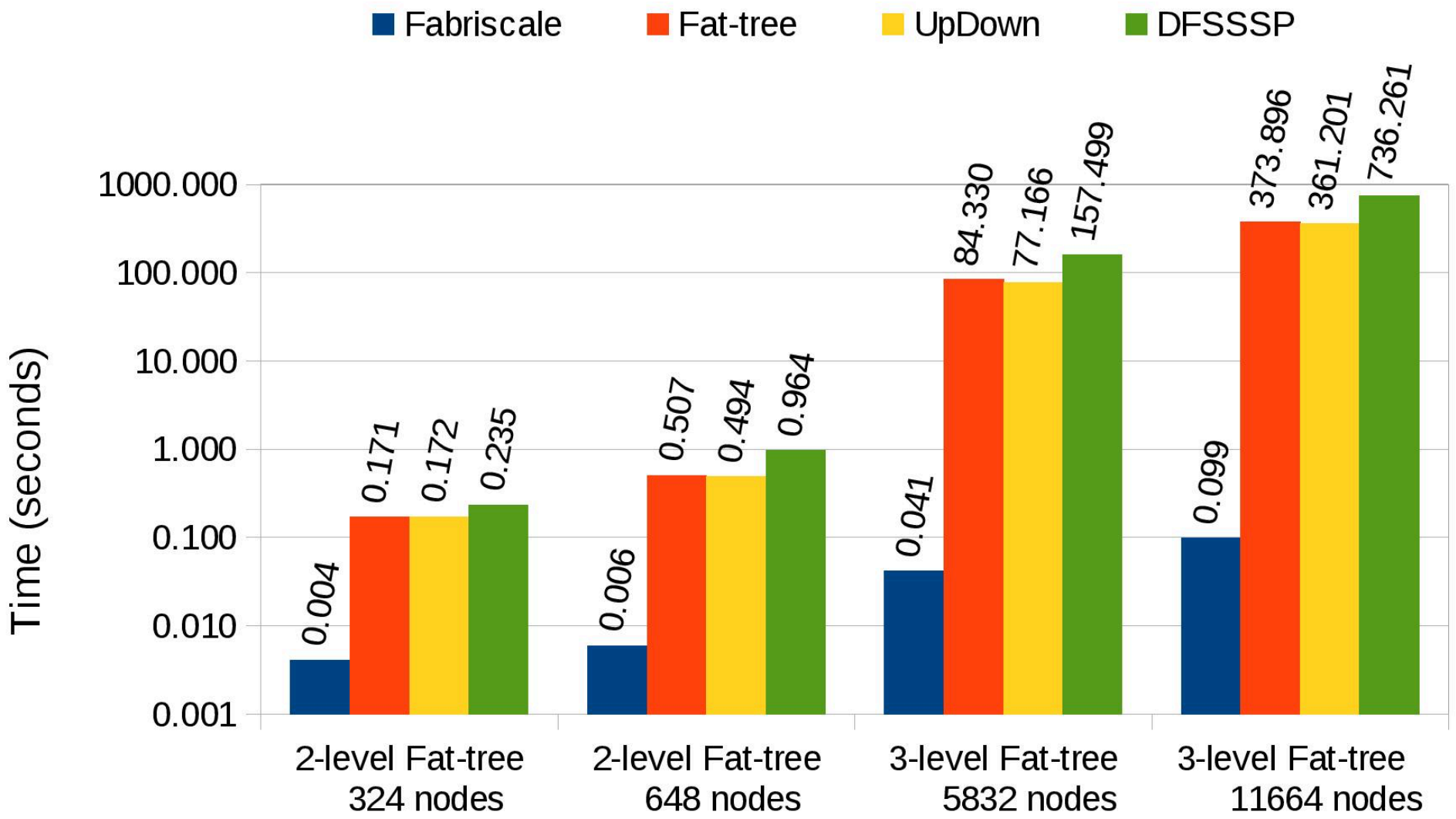

Oversubscription = 4

[5] Compact Network Reconfiguration in Fat-Trees, Zahid et al. Accepted to Journal of Supercomputing, 2016.

# RECONFIGURATION-/FAILOVER TIMES FOR OPENSM AS THE FAT-TREE SCALES



# FABRISCALE IS REDUCING RECONFIGURATION-/FAILOVER TIMES






fabriscale

# THANK YOU

Tor Skeie

University of Oslo / Simula Research Laboratory

[ simula . research laboratory ]