Rensselaer

Modeling a Million-Node Slim Fly Network Using Parallel Discrete-Event Simulation

Noah Wolfe, Christopher Carothers
Rensselaer Polytechnic Institute

Misbah Mubarak, Robert Ross, Philip Carns
Argonne National Laboratory

July 12, 2016

Outline

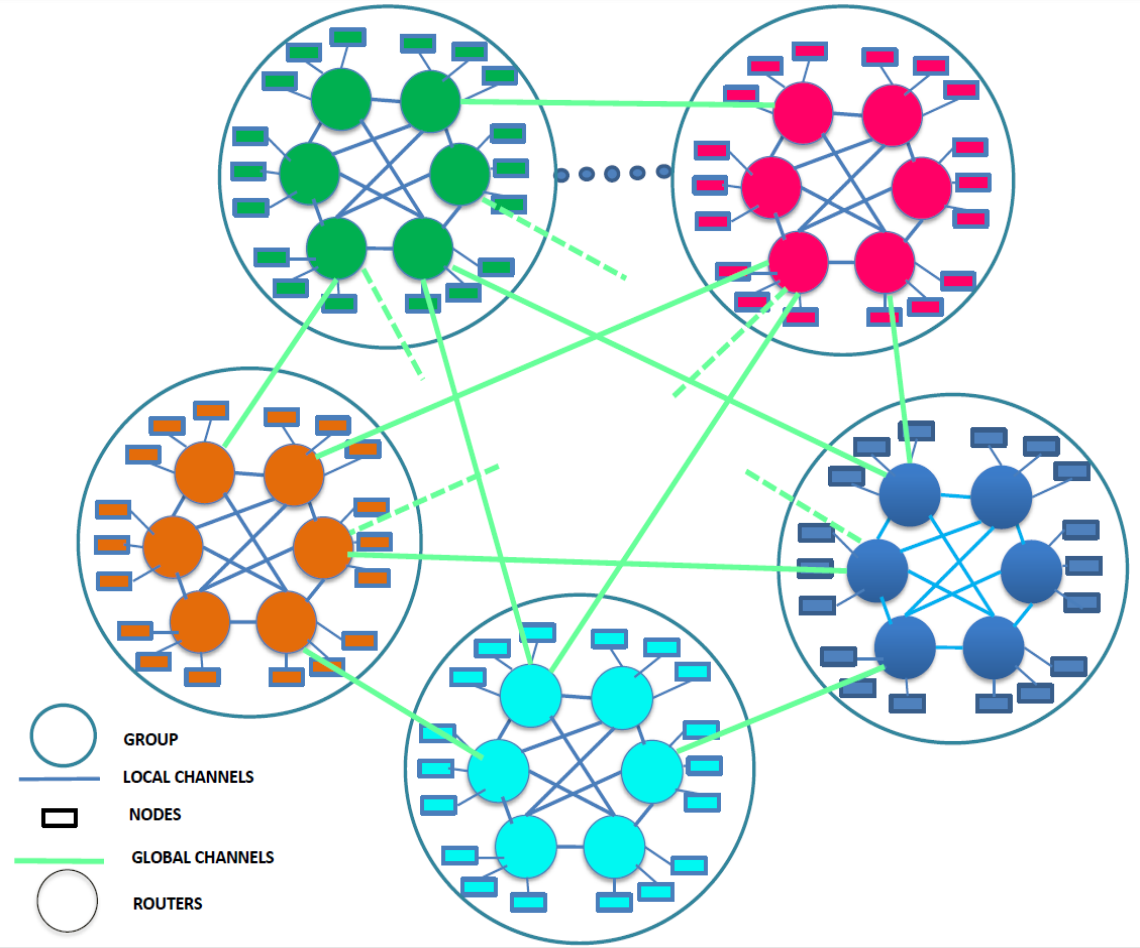
- Background/Motivation
- Slim Fly Network Topology Model
- Discrete-Event Model Implementation
- Model Verification
- Large-Scale Network Performance
- Simulation Performance
- Conclusion and Future Work

Supercomputing Systems

	Current System	Future System
ANL	Mira: 49,152 nodes 5D Torus	Aurora: >50,000 nodes Dragonfly?
ORNL	Titan: 18,688 nodes 3D Torus	Summit: >3,400 nodes Fat Tree
LLNL	Sequoia: 98,304 nodes 5D Torus	Sierra: >3,400 nodes Fat Tree

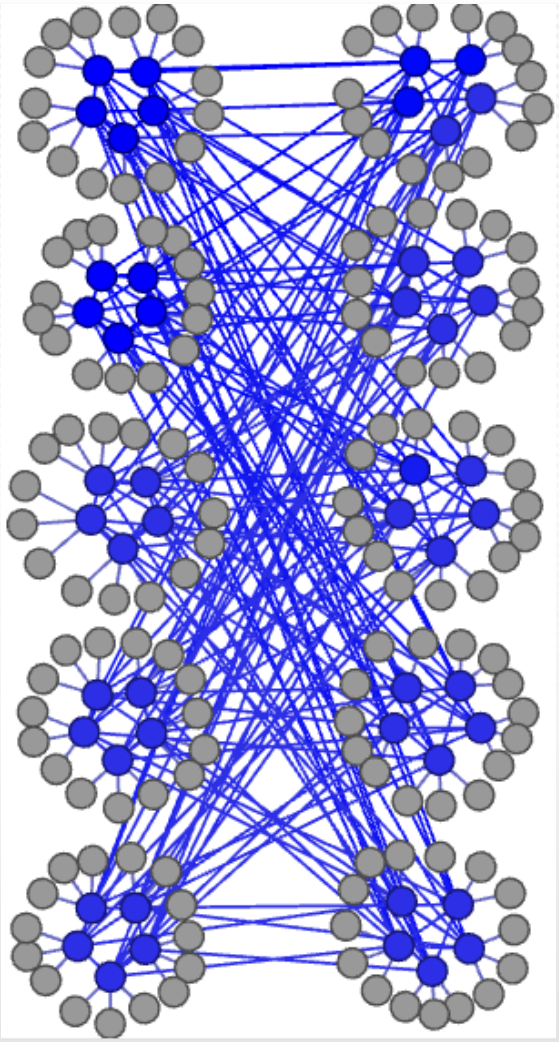
The Dragonfly Network Topology

- A two level directly connected topology
- Uses high-radix routers
  - Large number of ports per router
  - Each port has moderate bandwidth



Slim Fly Network Topology

- Built on MMS graphs
- Uses high-radix routers
- Max network diameter of 2
- Complex layout and connectivity



Network Parameters:

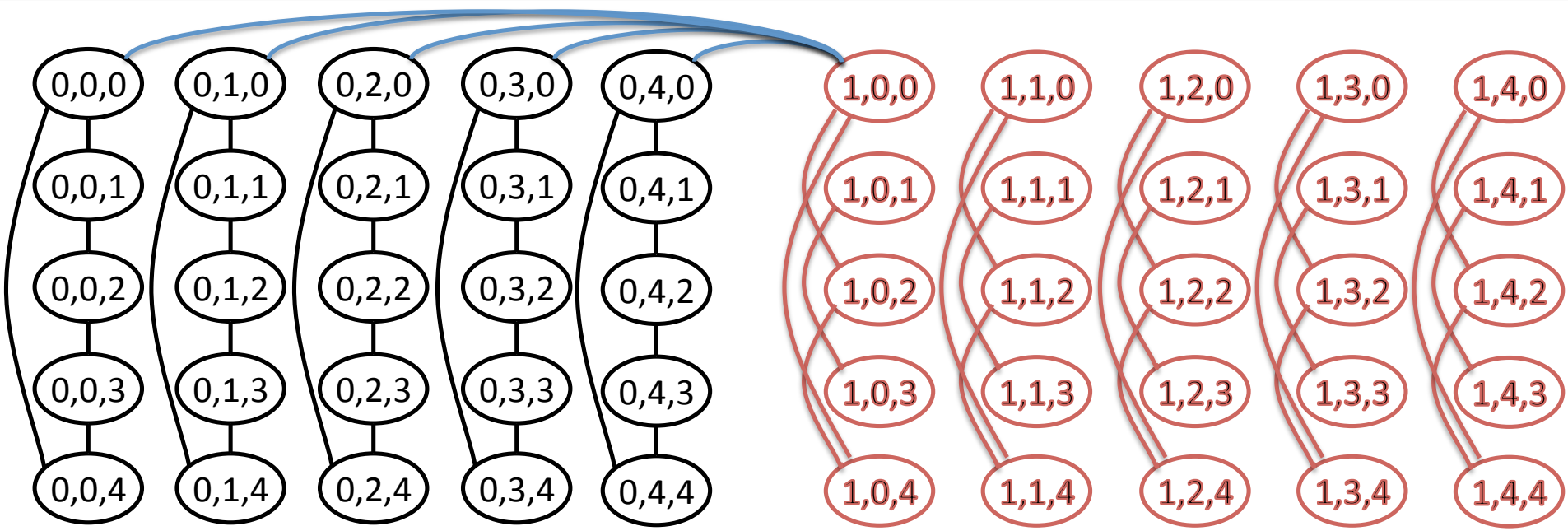
- q: number of routers per group and number of global connections per router
- p: number of terminal connections per router, p=floor(k'/2)
- k: router radix
- k': router network radix

Slim Fly Router Construction

$$\text{router}(0, x, y) \text{ is connected to } (0, x, y') \text{ iff } y - y' \in X; \quad (1)$$

$$\text{router}(1, m, c) \text{ is connected to } (1, m, c') \text{ iff } c - c' \in X'; \quad (2)$$

$$\text{router}(0, x, y) \text{ is connected to } (1, m, c) \text{ iff } y = mx + c; \quad (3)$$



Example MMS graph with q = 5 illustrating the connection of routers within groups and between subgraphs.

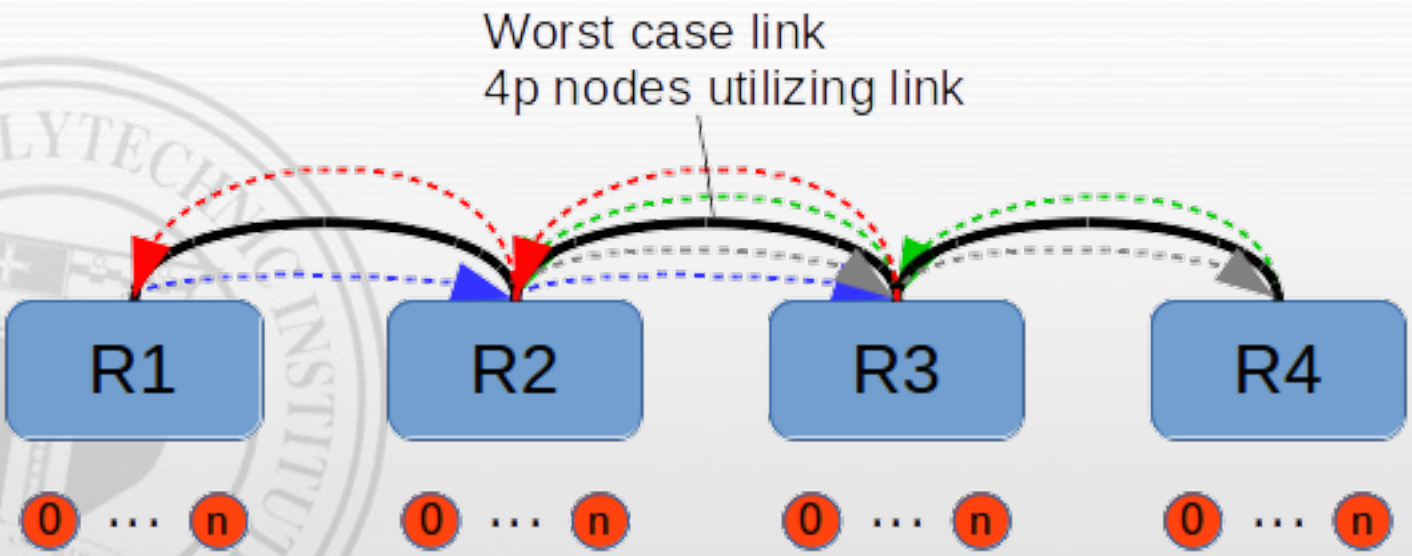
Layout Visualization

Model Configuration

- **Minimal Routing**
  - Messages take path of maximum 2 hops
- **Non-minimal routing**
  - Messages are routed minimally to a random intermediate router and then minimally to the destination
- **Adaptive routing**
  - Chooses between minimal and non-minimal by sensing the traffic conditions on the message's source router
- **Virtual channels**
  - To avoid deadlocks
- **Credit based flow control**
  - Upstream nodes/routers keep track of buffer slots

Synthetic Application Workloads

- **Uniform Random (UR)**
  - Terminals send to random destination terminal
- **Worst-Case (WC)**
  - Simulates an application that is communicating in a manner that fully saturates links in the network creating a bottleneck for minimal routing



PDES Implementation

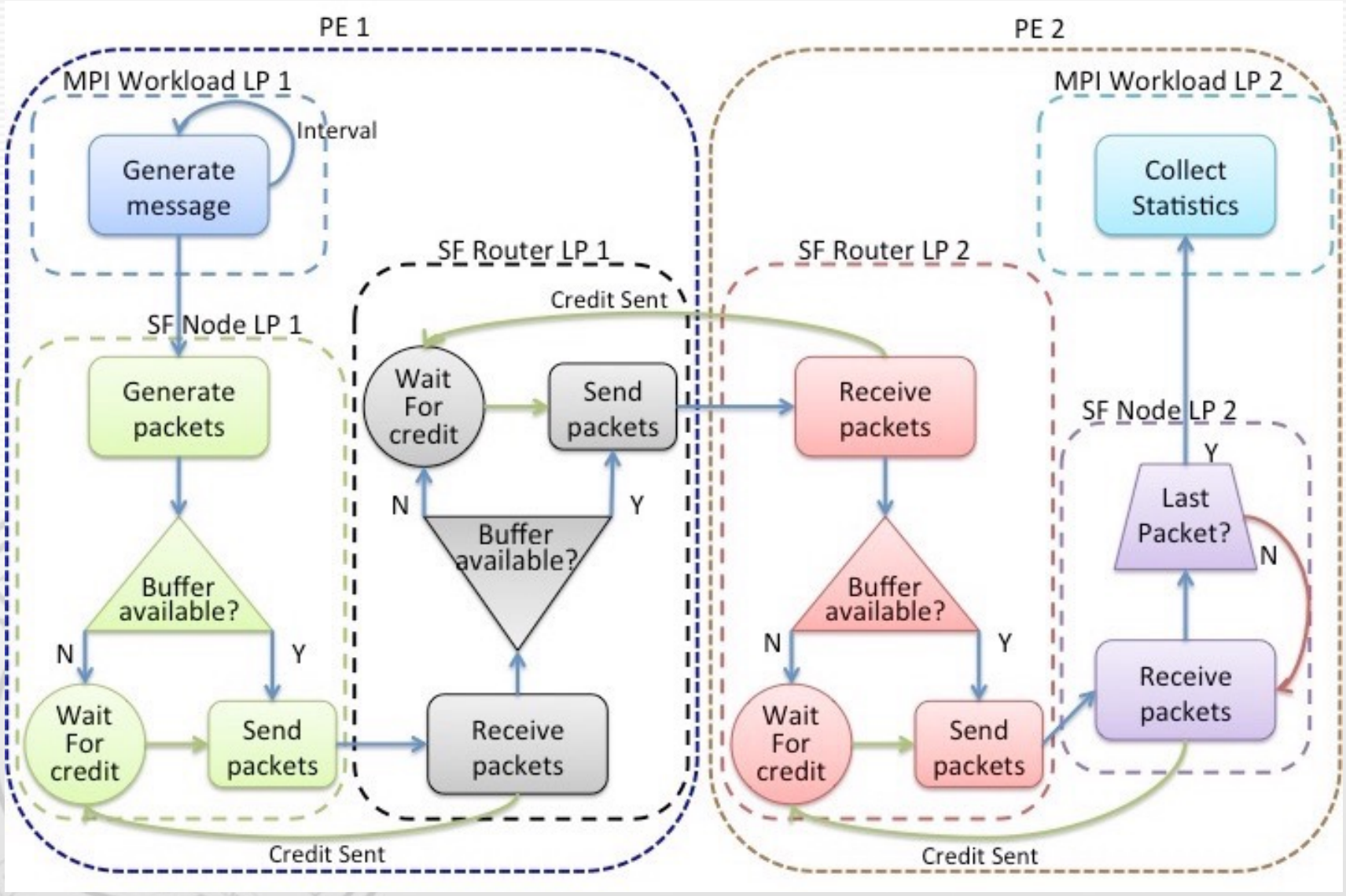
Co-Design of Exascale Storage (CODES)

- Storage systems
- HPC network systems

Rensselaer Optimistic Simulation System (ROSS)

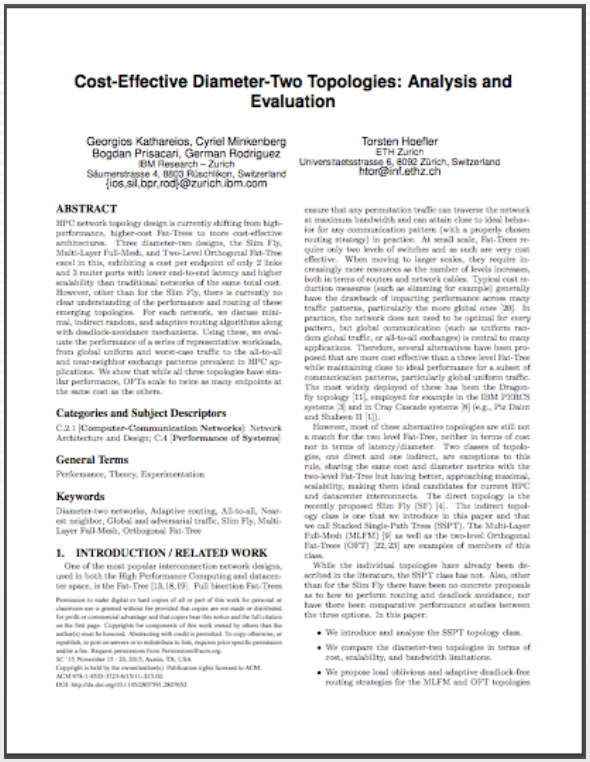
- Event Scheduling
  - Conservative
  - Optimistic
- Logical Processes (LPs)
  - MPI processes
  - Terminals/compute nodes
  - Routers/switches

PDES Implementation



Slim Fly Verification

- Comparison with published Slim Fly network results by Kathareios et al. [1]
- Slim Fly Configuration:
  - 3,042 nodes
  - 338 routers
  - q=13
  - k=28
- Simulation Configuration
  - link bandwidth: 100Gbps
  - link latency: 50ns
  - buffer space: 100KB
  - router delay: 100ns
  - Simulated time: 200us



Cost-Effective Diameter-Two Topologies: Analysis and Evaluation

Georgios Kathareios, Cyriel Minkenberg
Bogdan Prisacari, German Rodriguez
IBM Research – Zurich
Säumerstrasse 4, 8803 Rüschlikon, Switzerland
{ios,sil,bpr,rod}@zurich.ibm.com

Torsten Hoefler
ETH Zurich
Universitaetsstrasse 6, 8092 Zürich, Switzerland
htor@inf.ethz.ch

ABSTRACT

HPC network topology design is currently shifting from high-performance, higher-cost Fat-Trees to more cost-effective architectures. Three diameter-two designs, the Slim Fly, Multi-Layer Full-Mesh, and Two-Level Orthogonal Fat-Tree excel in this, exhibiting a cost per endpoint of only 2 links and 3 router ports with lower end-to-end latency and higher scalability than traditional networks of the same total cost. However, other than for the Slim Fly, there is currently no clear understanding of the performance and routing of these emerging topologies. For each network, we discuss minimal, indirect random, and adaptive routing algorithms along with deadlock-avoidance mechanisms. Using these, we evaluate the performance of a series of representative workloads, from global uniform and worst-case traffic to the all-to-all and near-neighbor exchange patterns prevalent in HPC applications. We show that while all three topologies have similar performance, OFTs scale to twice as many endpoints at the same cost as the others.

Categories and Subject Descriptors

C.2.1 [Computer-Communication Networks]: Network Architecture and Design; C.4 [Performance of Systems]

General Terms

Performance, Theory, Experimentation

Keywords

Diameter-two networks, Adaptive routing, All-to-all, Nearest neighbor, Global and adversarial traffic, Slim Fly, Multi-Layer Full-Mesh, Orthogonal Fat-Tree

1. INTRODUCTION / RELATED WORK

One of the most popular interconnection network designs, used in both the High Performance Computing and datacenter space, is the Fat-Tree [13, 18, 19]. Full bisection Fat-Trees ensure that any permutation traffic can traverse the network at maximum bandwidth and can attain close to ideal behavior for any communication pattern (with a properly chosen routing strategy) in practice. At small scale, Fat-Trees require only two levels of switches and as such are very cost effective. When moving to larger scales, they require increasingly more resources as the number of levels increases, both in terms of routers and network cables. Typical cost reduction measures (such as slimming for example) generally have the drawback of impacting performance across many traffic patterns, particularly the more global ones [20]. In practice, the network does not need to be optimal for every pattern, but global communication (such as uniform random global traffic, or all-to-all exchanges) is central to many applications. Therefore, several alternatives have been proposed that are more cost effective than a three level Fat-Tree while maintaining close to ideal performance for a subset of communication patterns, particularly global uniform traffic. The most widely deployed of these has been the Dragonfly topology [11], employed for example in the IBM PERCS systems [3] and in Cray Cascade systems [8] (e.g., Piz Daint and Shaheen II [1]).

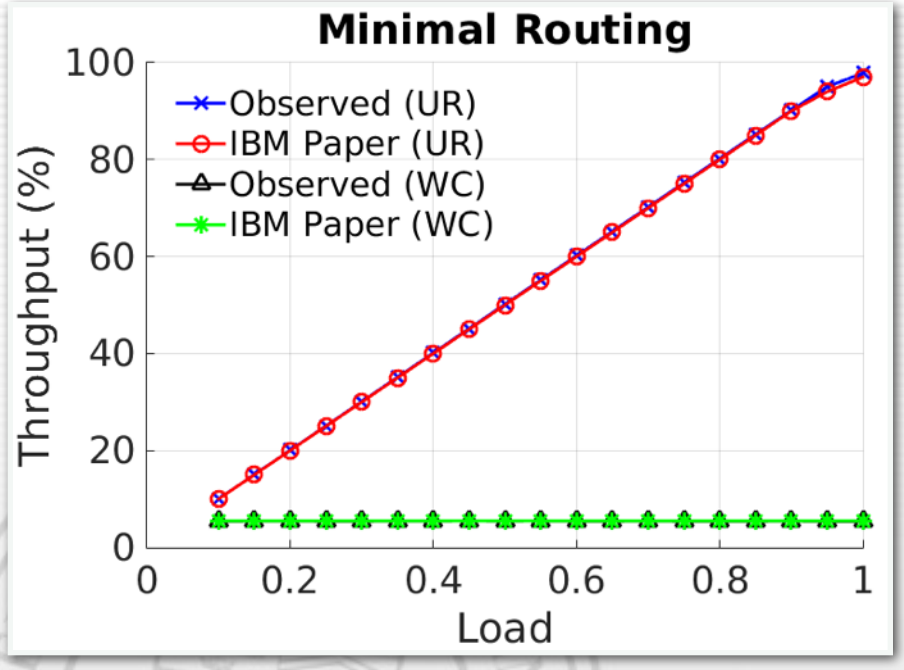
However, most of these alternative topologies are still not a match for the two level Fat-Tree, neither in terms of cost nor in terms of latency/diameter. Two classes of topologies, one direct and one indirect, are exceptions to this rule, sharing the same cost and diameter metrics with the two-level Fat-Tree but having better, approaching maximal, scalability, making them ideal candidates for current HPC and datacenter interconnects. The direct topology is the recently proposed Slim Fly (SF) [4]. The indirect topology class is one that we introduce in this paper and that we call Stacked Single-Path Tree (SSPT). The Multi-Layer Full-Mesh (MLFM) [9] as well as the two-level Orthogonal Fat-Trees (OFT) [22, 23] are examples of members of this class.

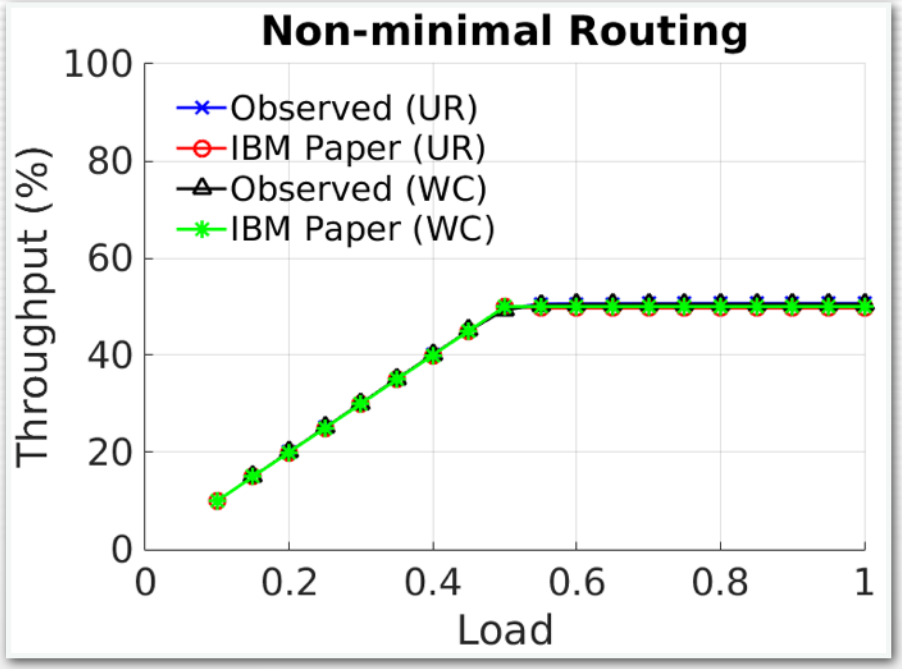
While the individual topologies have already been described in the literature, the SSPT class has not. Also, other than for the Slim Fly there have been no concrete proposals as to how to perform routing and deadlock avoidance, nor have there been comparative performance studies between the three options. In this paper:

- We introduce and analyze the SSPT topology class.
- We compare the diameter-two topologies in terms of cost, scalability, and bandwidth limitations.
- We propose load oblivious and adaptive deadlock-free routing strategies for the MLFM and OFT topologies

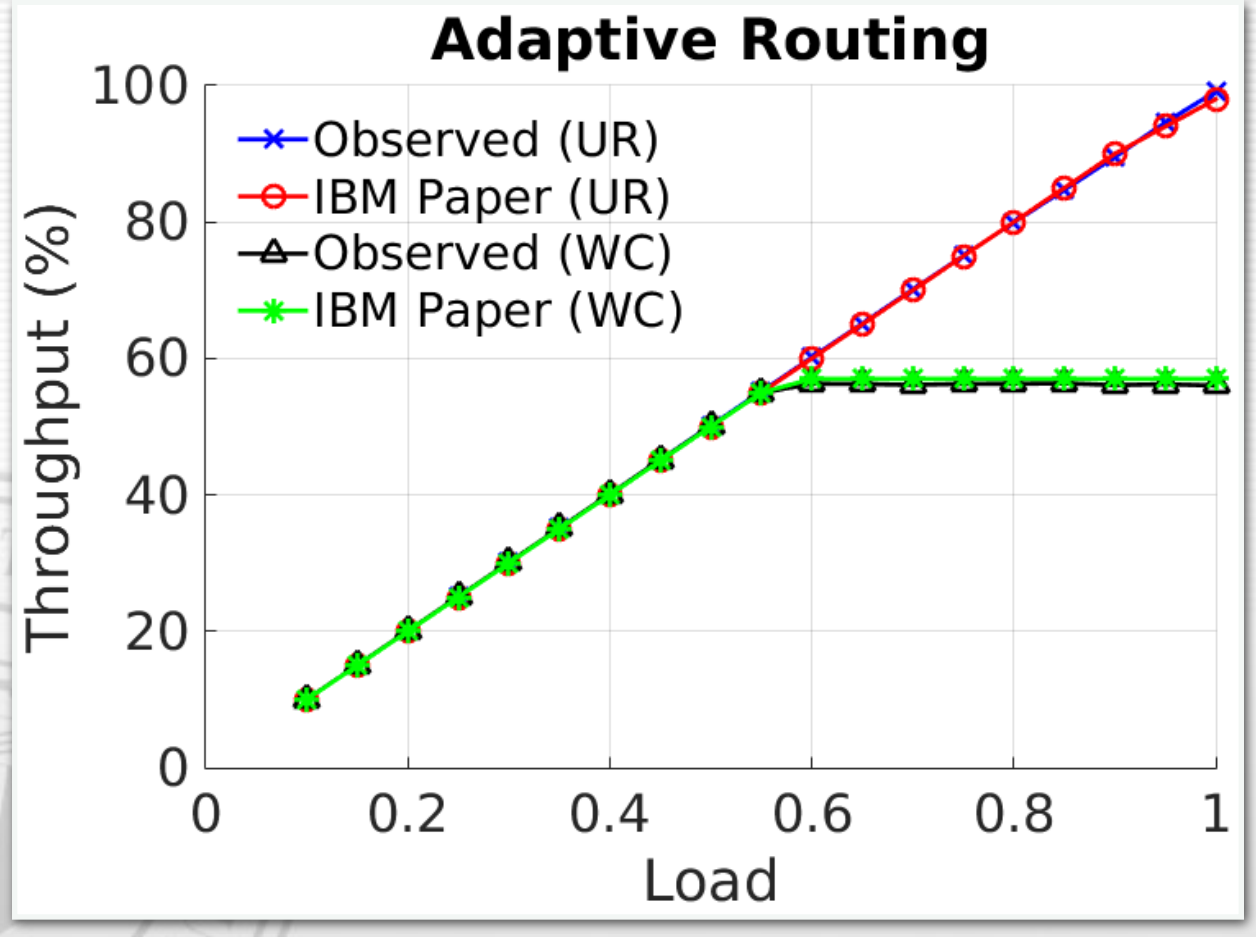
[1] G. Kathareios, C. Minkenberg, B. Prisacari, G. Rodriguez, and T. Hoefler. Cost-Effective Diameter-Two Topologies: Analysis and Evaluation. Nov. 2015. IEEE/ACM ICHPCNSA (SC15).

Minimal/Non-minimal Routing





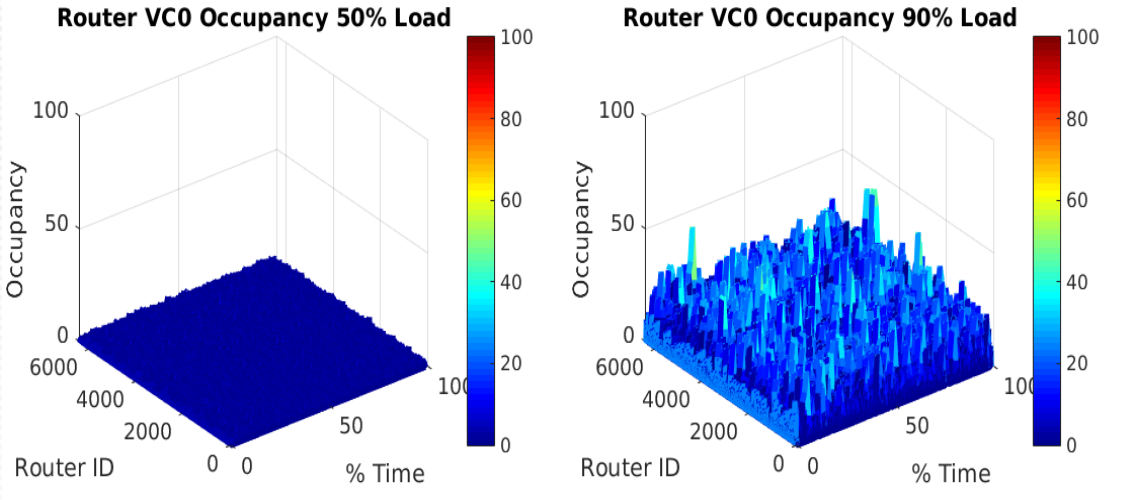
Adaptive Routing



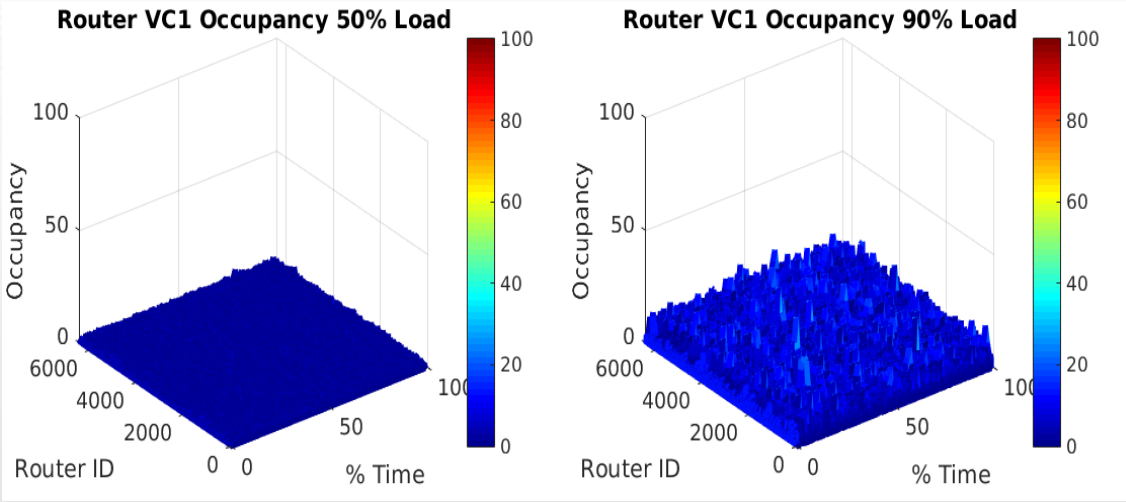
Slim Fly Network Model Results

Virtual Channel Occupancy

- UR traffic
- Minimal routing

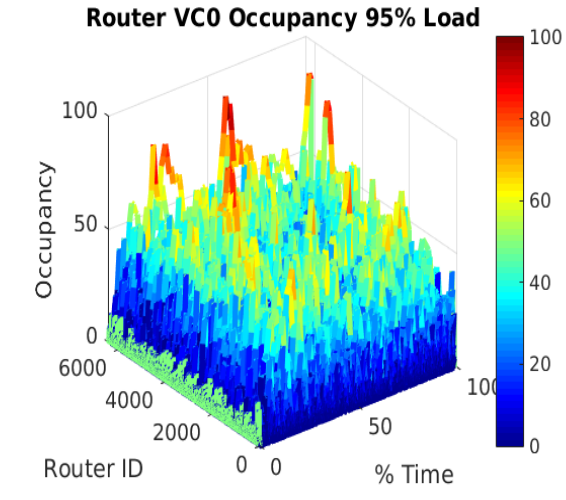


(a) VC0 50% Load (b) VC0 90% Load

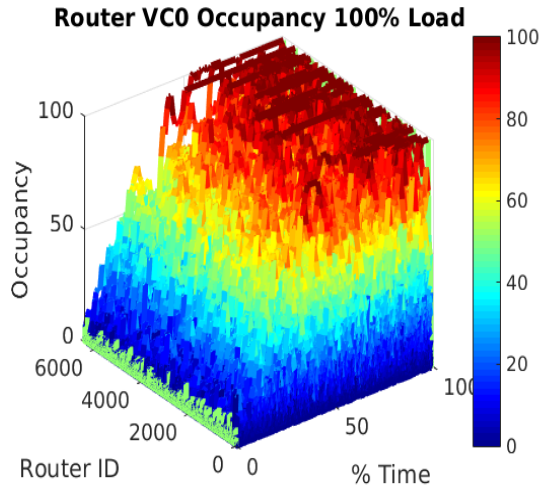


(e) VC1 50% Load (f) VC1 90% Load

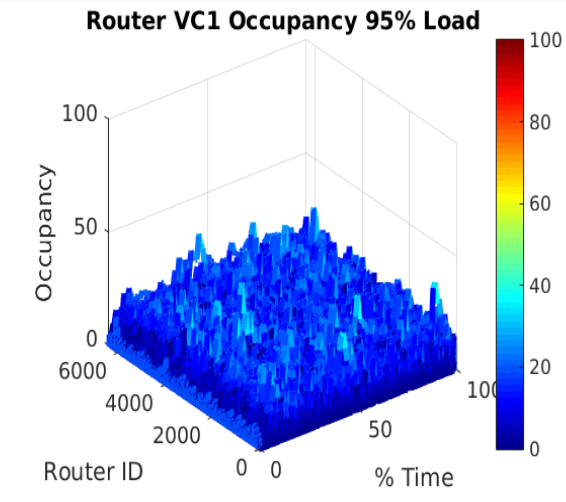
Virtual Channel Occupancy



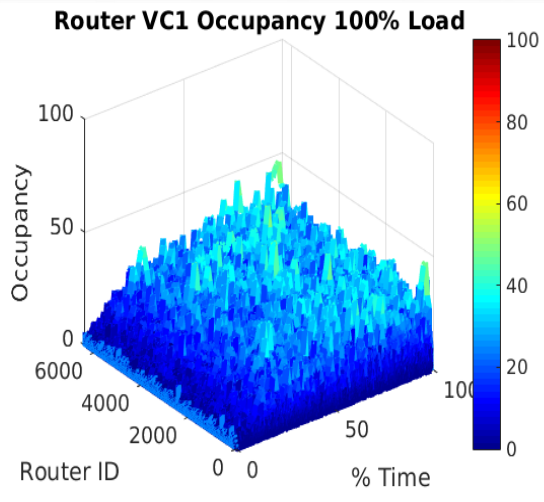
(c) VC0 95% Load



(d) VC0 100% Load



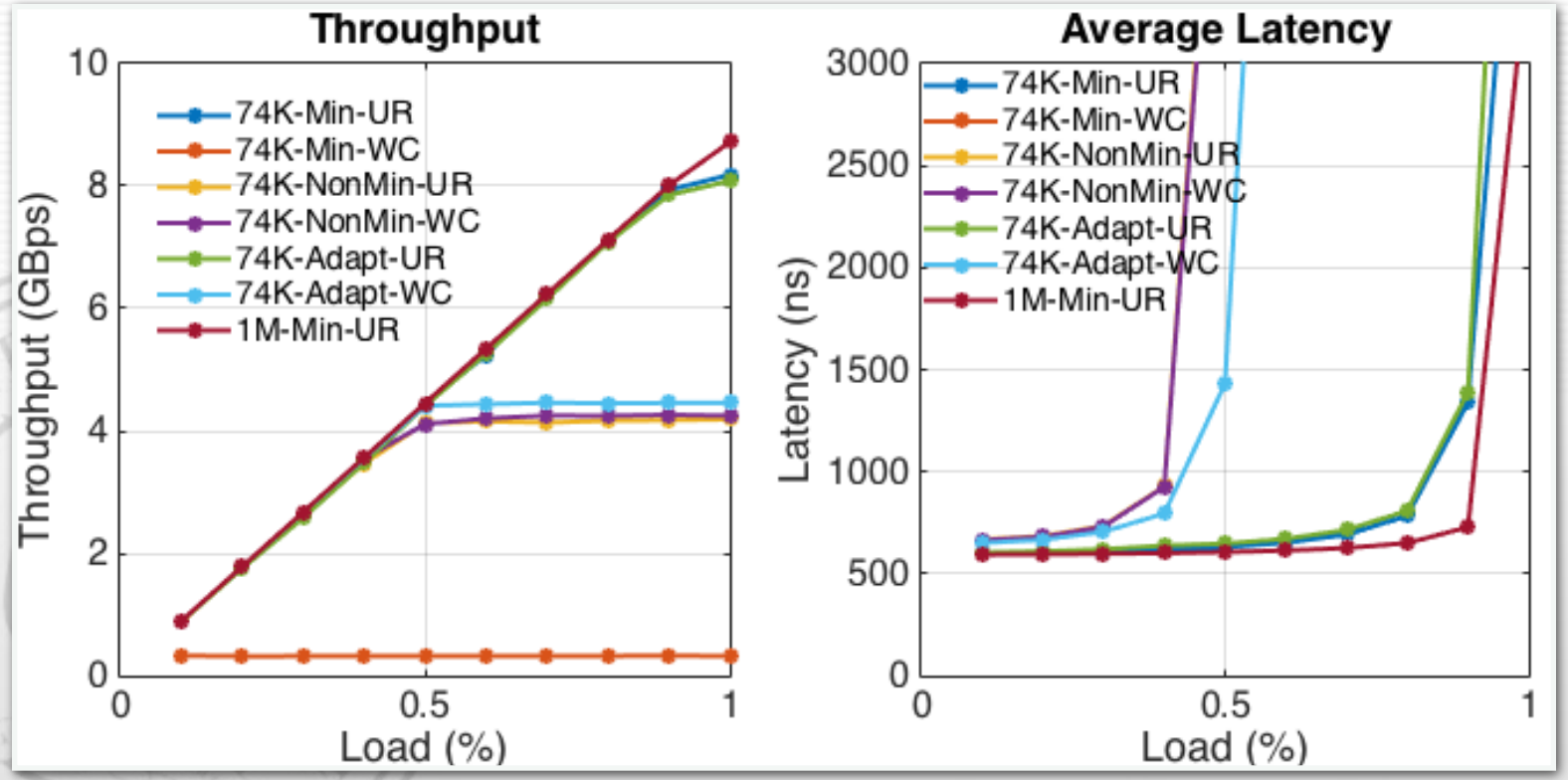
(g) VC1 95% Load



(h) VC1 100% Load

Large-Scale Performance

- 74K node (Aurora) system
  - 2,738 routers
  - q=37, k=82
- 1M node system
  - 53,178 routers
  - q=163, k=255



ROSS PDES Results

Performance Parameters

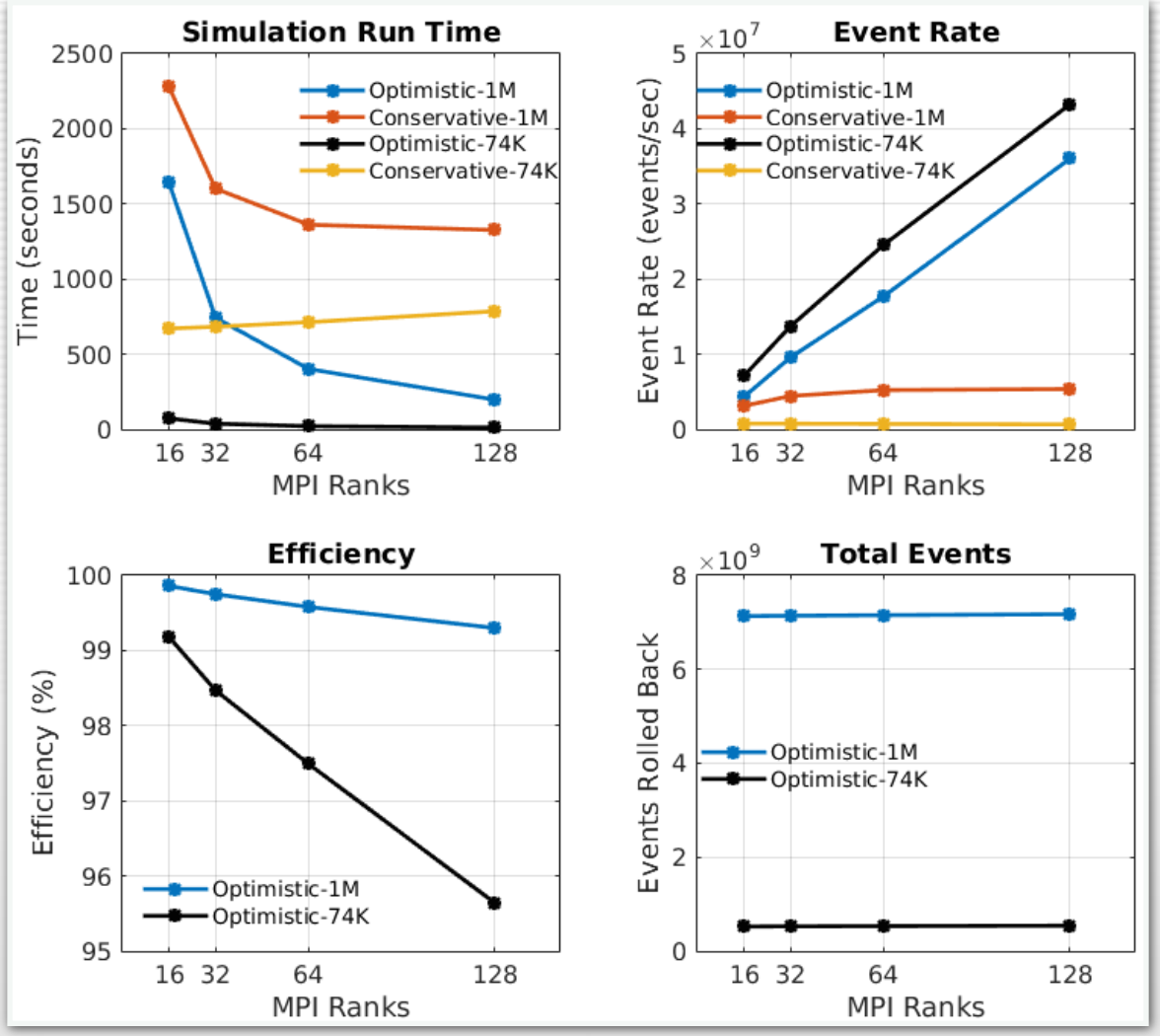
- Evaluation System
  - Center for Computational Innovations (CCI) RSA Intel cluster
  - 34 nodes, 2x 4-core Intel Xeon E5-2643 3.3 GHz processors
- Execution Parameters
  - 1M-node model: 4 MPI tasks per node
  - 74K-node model: 8 MPI tasks per node

- Performance metrics
  - Committed event rate
  - Total events
  - ROSS event efficiency
  - Simulation run time

- Event Schedulers
  - Optimisitic
  - Conservative

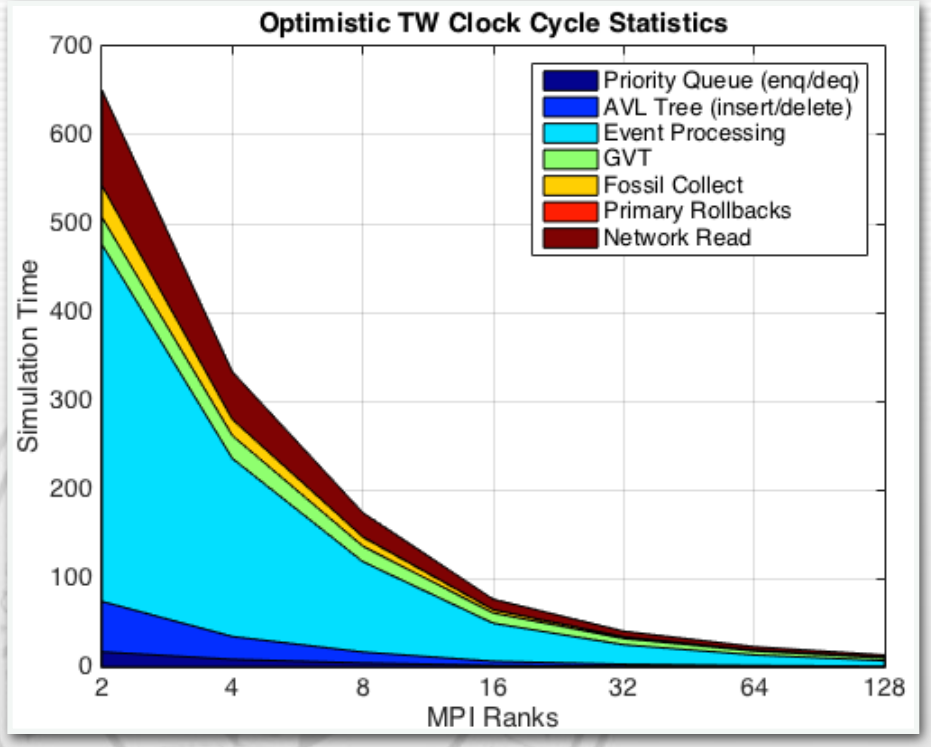
Scaling Analysis

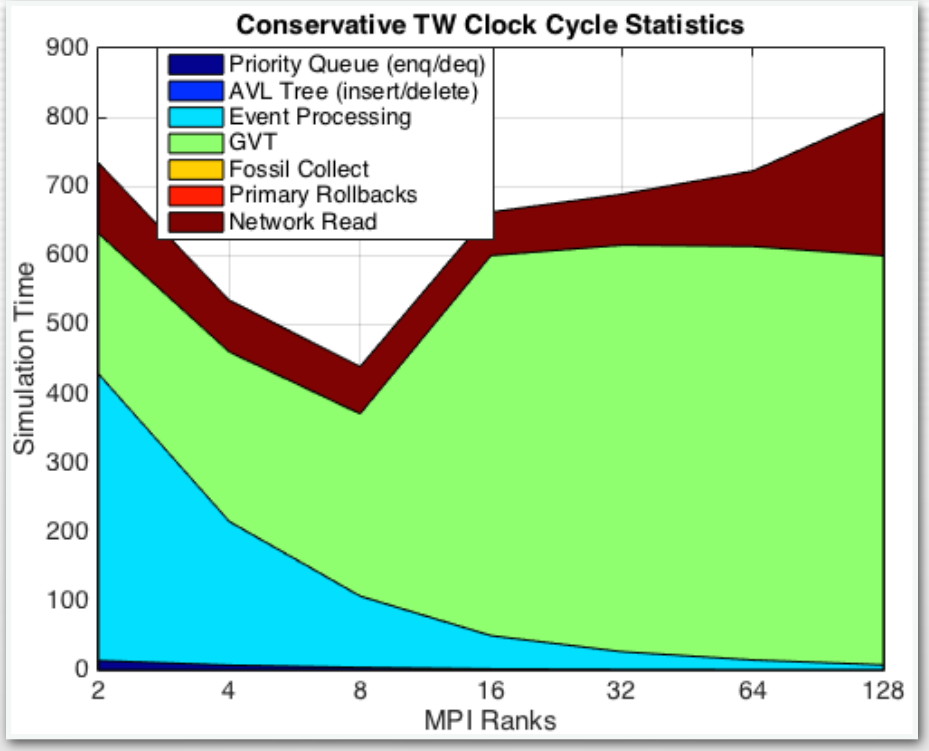
- **74K Node Model:**
  - 43 million events per second
  - 543 million events processed
- **1M Node Model:**
  - 36 million events per second
  - 7 billion events processed



PDES Analysis

- Distribution of simulation time scales linearly in Optimistic event scheduling indicating a uniform distribution of work among all processing elements (MPI ranks).





Future Work

- Future work
  - Compare the slim fly network model with other candidate topology models for exascale computing
  - Integrate slim fly simulator with NeMo simulator to investigate the possibility of a hybrid supercomputing system incorporating neuromorphic hardware such as IBM's TrueNorth processor

Questions?

Rensselaer