Fit Fly

A CASE STUDY ON INTERCONNECT INNOVATION THROUGH PARALLEL SIMULATION

Neil McGlohon (RPI), Noah Wolfe (AMD), Misbah Mubarak (AWS), Christopher D. Carothers (RPI)

Rob Ross, , Matthieu Dorier and Sudheer Chunduri
Mathematics and Computer Science Division
Argonne National Laboratory
r

Kwan-Liu Ma, Takanori Fujiwara, and Kelvin Li
Computer Science Department
University of California at Davis

Mark Plagge, Caitlin Ross (Kitware), Daniel Yaciuk
Computer Science Department
Rensselaer Polytechnic Institute

Zhiling Lan, Ram Chaulagain, Yao Kang and Xin Wang
Computer Science Department
Illinois Institute of Technology

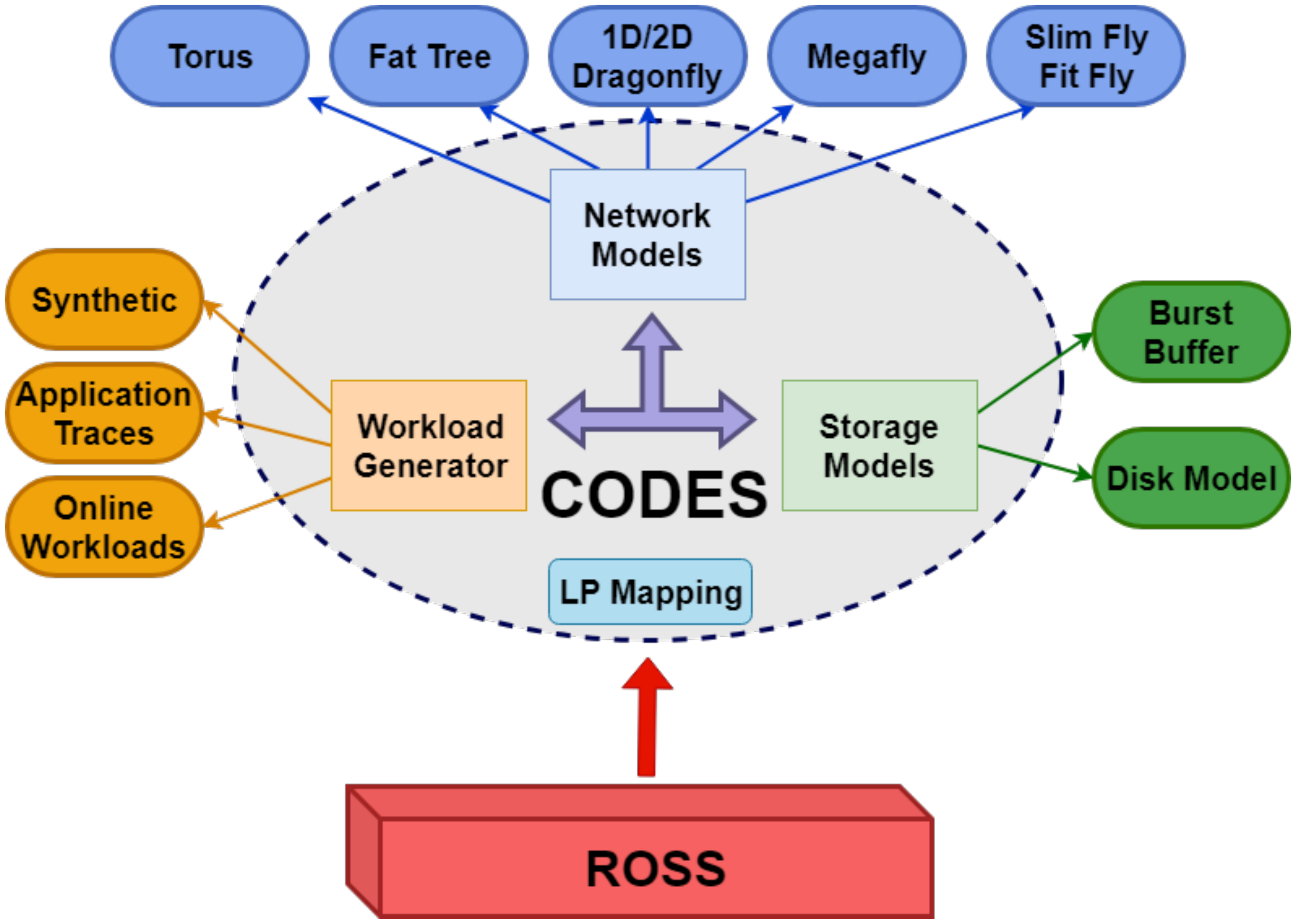
Motivation: Network Design Points for Future Systems

- **Summit and Sierra**: 2x Mellanox EDR NICs @ ~25 GB/s per node in fat-tree topology design.
- **Frontier**: Multiple NICs providing 100 GB/s network bandwidth in a Slingshot dragonfly network topology.
- **Trend**: Systems to use multiple NICs per node to meet bandwidth demands.

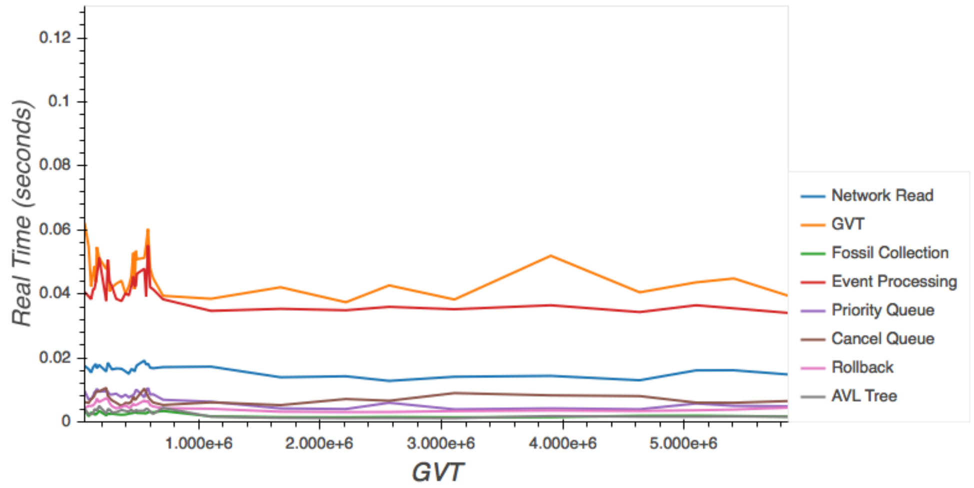
For future systems there appears to be an interesting set of design questions around the number of NICs or "rails" per compute node and network topology ?

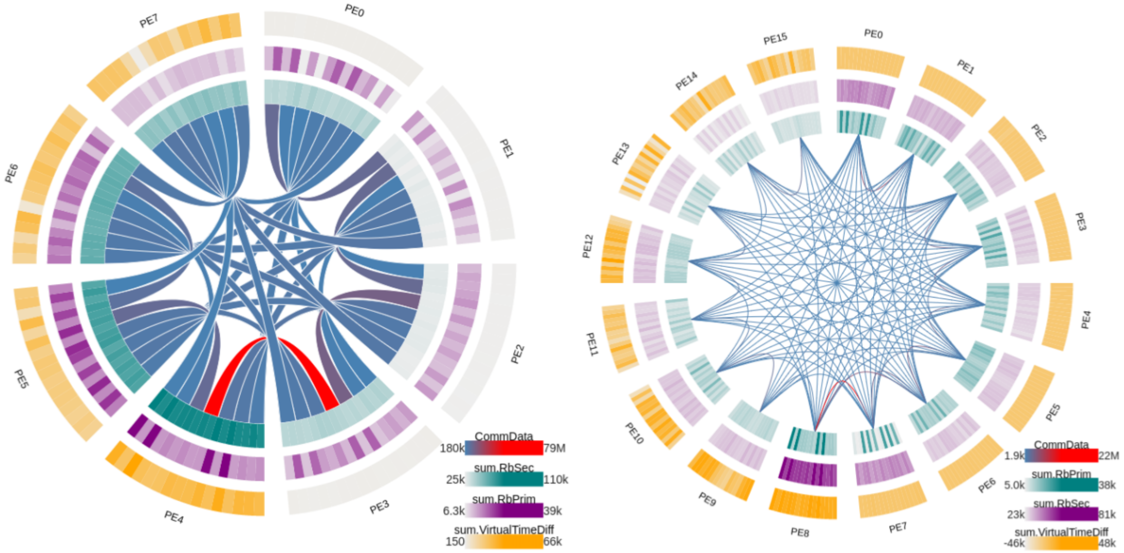
We can address it at full network/system scale using parallel simulation methods and tools

CODES: Co-Design of Exascale Storage Architectures



Parallel Performance Data Collection & Visualization

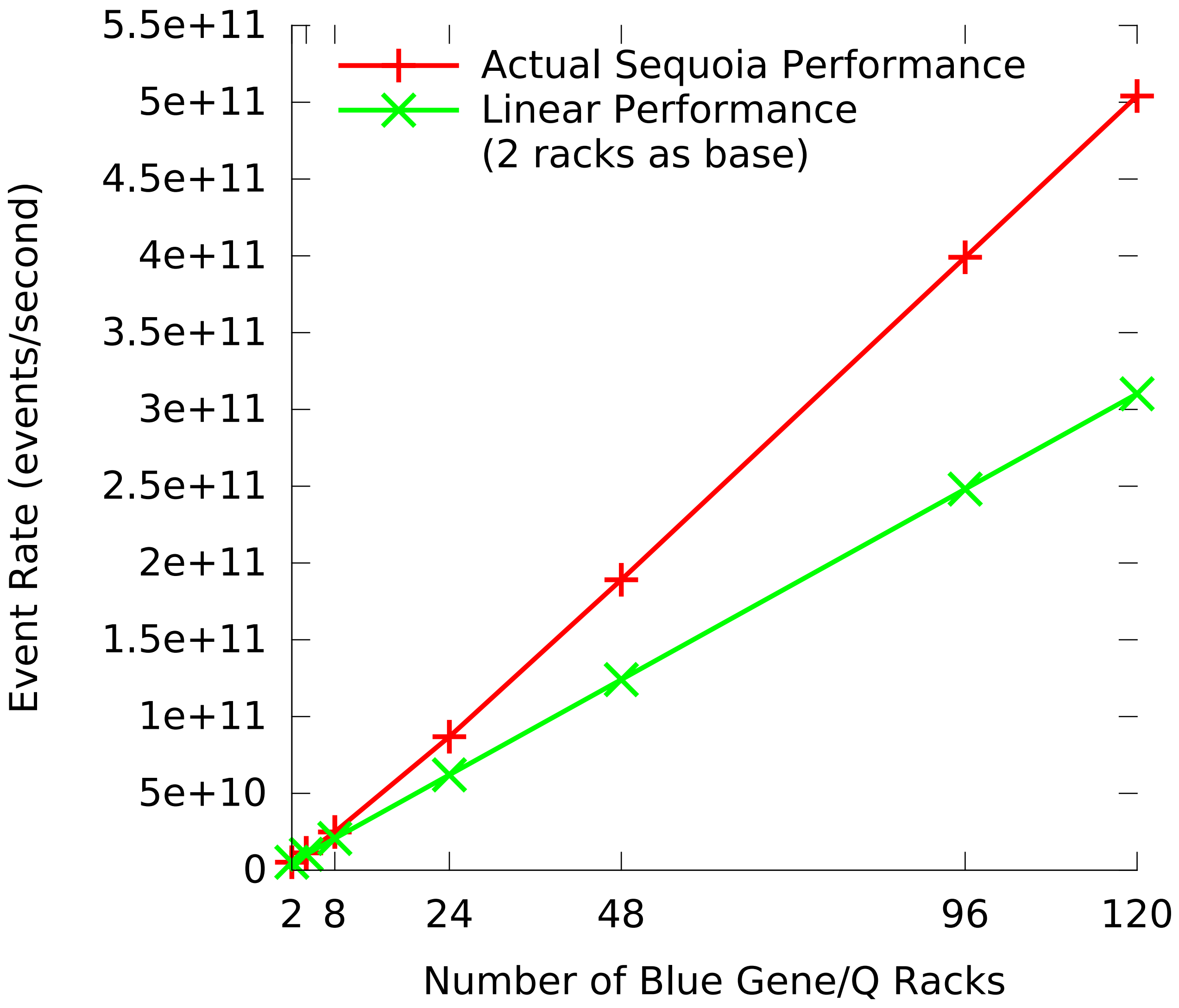




CODES: https://github.com/codes-org/codes
ROSS: https://github.com/ROSS-org/ROSS.git

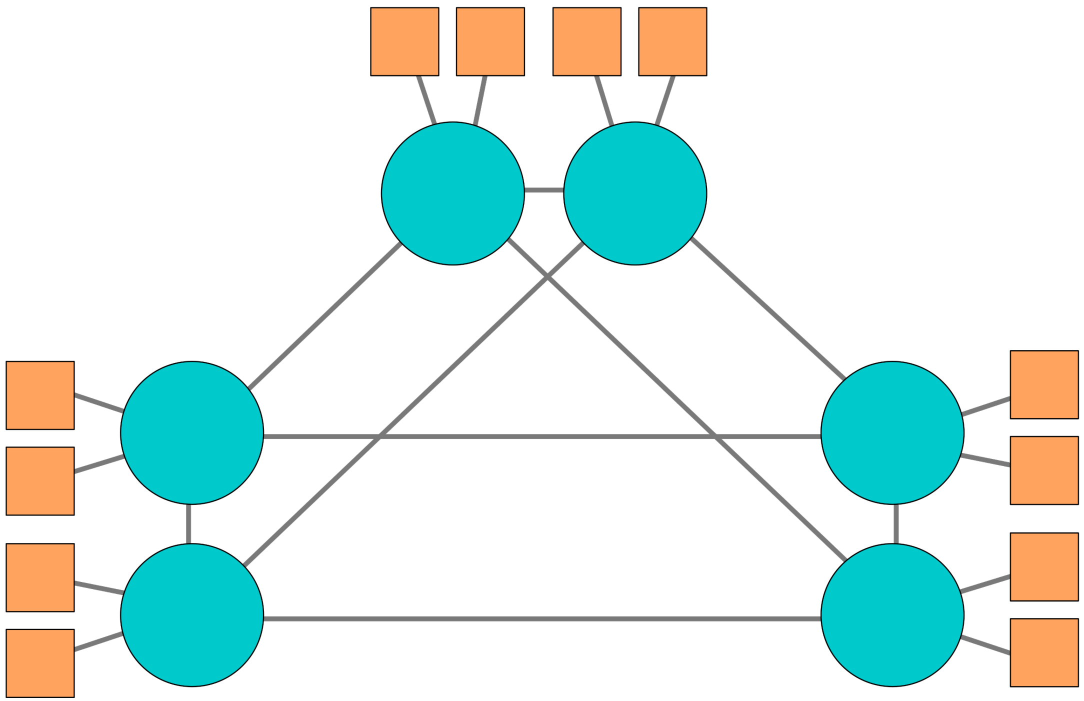
ROSS – Scalable Parallel Simulation Engine

- Discrete-Event approach to modeling
  - Core simulation entity is **Logical Process**
- **Schedulers:** sequential, conservative and optimistic
  - Optimistic rollback supported via “Reverse Computation”
- Opportunity to leverage DOE investments in current supercomputer systems
- **Simulation parallel scaling limited by:**
  - Performance of supercomputer’s or cluster’s network
  - Exploitable parallelism in the DES model.
  - *Increases possible model configurations*
- **Larger network switches & “fat node” designs reduce opportunity to exploit model parallelism.**



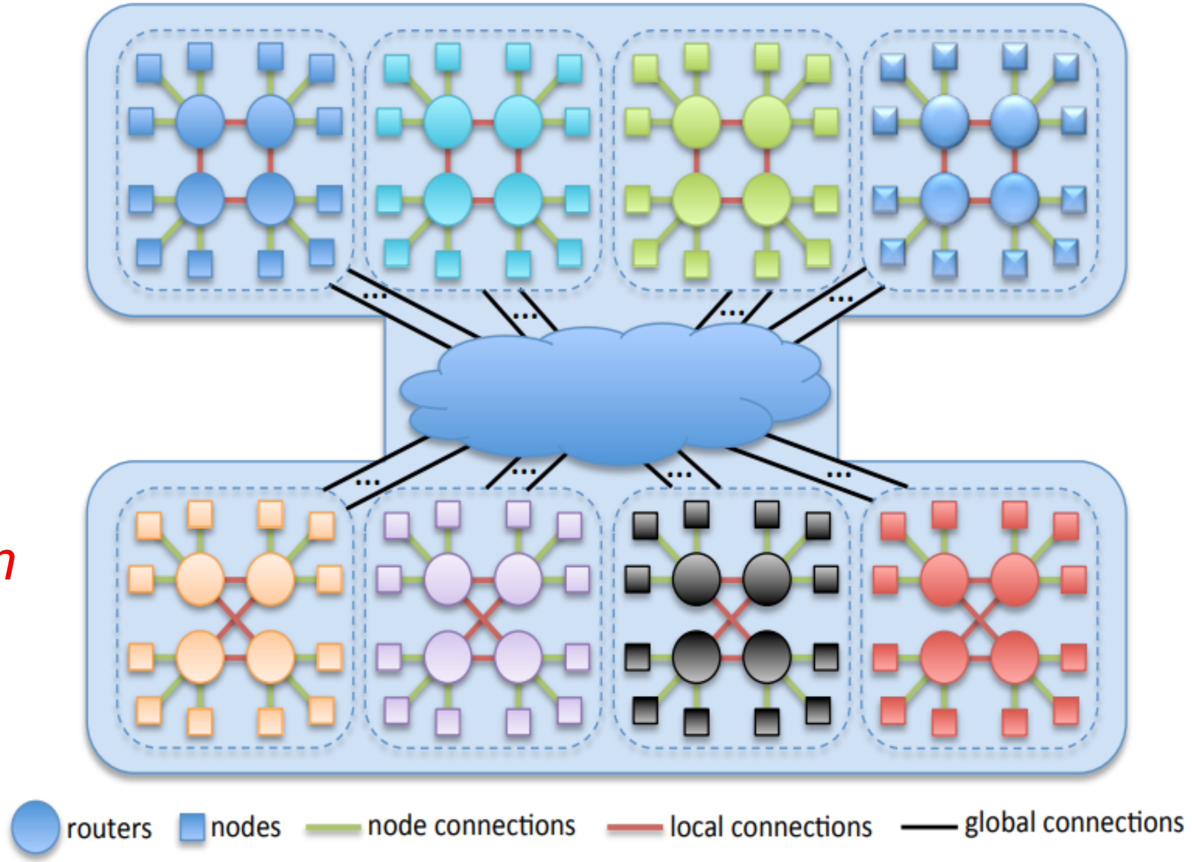
HPC Interconnect Traffic Simulation

- Network of **Switches**
- **Terminals** attached to Switches
- Traffic is **Generated** at Terminals
- Traffic is **Routed** through network of Switches
- Traffic **Terminated** at a **pre-specified** destination Terminal
- All messages/packets realized as events in the discrete-event simulator



Fit Fly Starts w/ Slim Fly Network Topology

- Slim Fly [Besta and Hoefler @SC'14] arranges routers into groups using MMS graph structure
- **Each Router:**
  - Some degree of Local connectivity
  - Some degree of Global connectivity
  - Some degree of Terminal connectivity
- Guaranteed Diameter-2 (MMS graph property)
- **Groups are divided into two subgraphs**
  - *No global connections between two groups within same subgraph*
- Connections are determined via Finite Field generation method.
- **Makes it challenging to physically build**



Slim Fly Network Generation

- Find a prime power $q = 4\omega + \delta$, where $\delta \in \{-1, 0, 1\}$ and $\omega \in \mathrm{N}$, such that $N_r = 2q^2$ is satisfied for the desired number of routers N_r
- Construct a Galois field of order q: F_q
  - Find the primitive element ξ that generates it
  - All nonzero elements of F_q can be written as ξ^{i}, where $i \in N$
- Using ξ, construct generator sets X and X'
- Determine router-router connections using following equations:

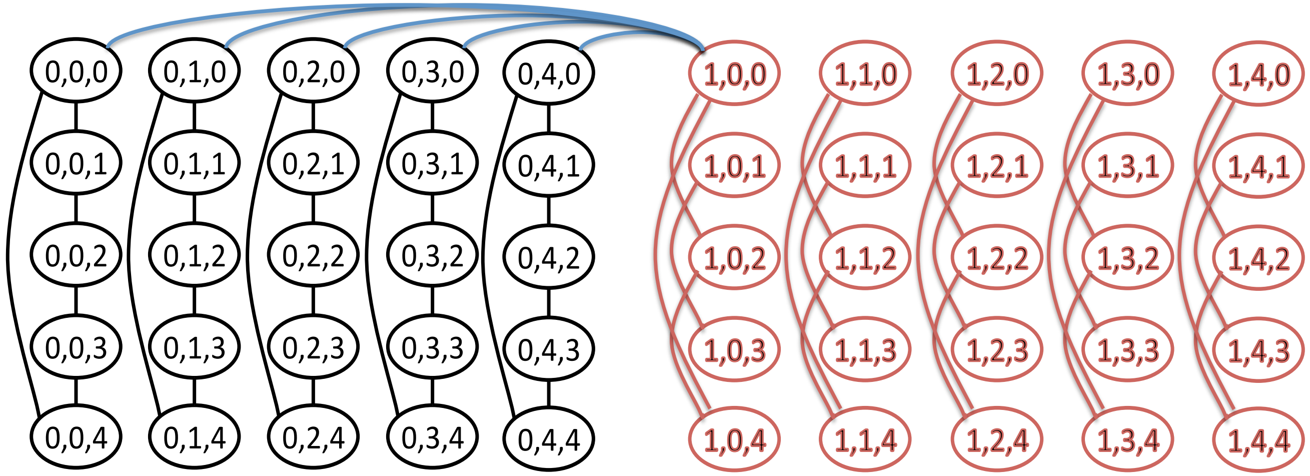


Figure 2: Example MMS graph with $q = 5$ illustrating the connection of routers within groups and between subgraphs.

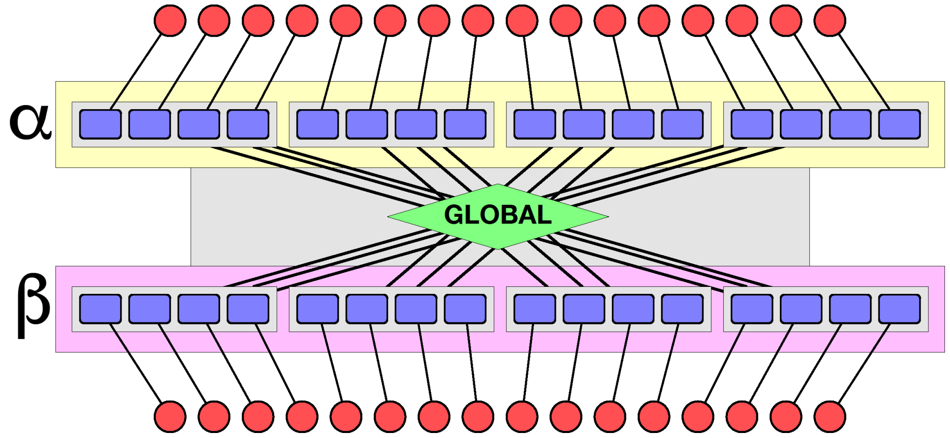
router(α, x, y) connected to (α, x, y') iff $y - \mathrm{y}' \in X$ (1) [intragroup connections alpha]

router(β, m, c) connected to (β, m, c') iff $c - \mathrm{c}' \in X'$ (2) [intragroup connections beta]

router(α, x, y) connected to (β, m, c) iff $y = mx + c$ (3) [connections between alpha, beta]

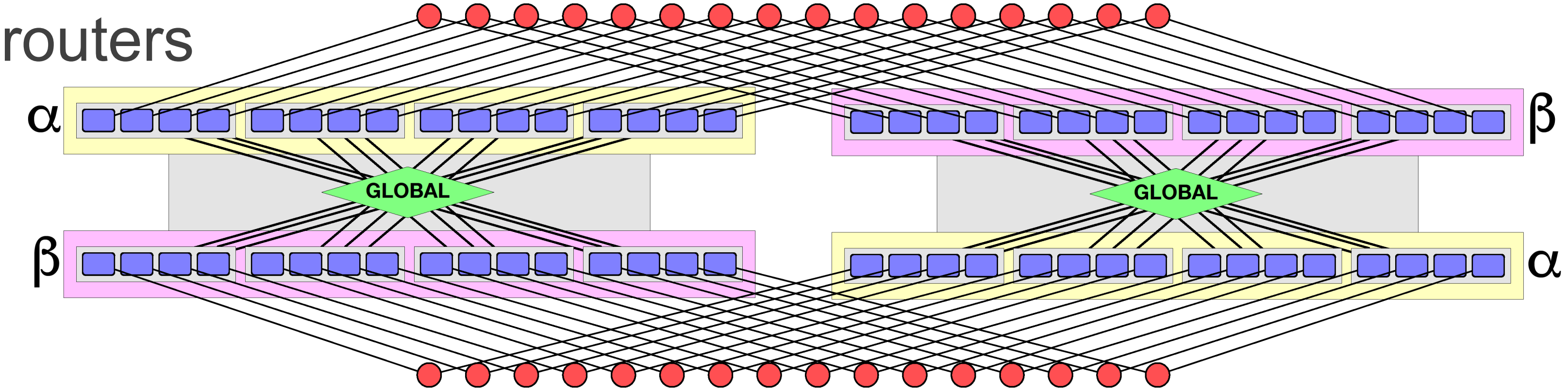
Fit Fly – Multi-Rail, Multi-Plane Slim Fly

- Multi-Rail, Multi-Planar Slim Fly Network
- Planes share single set of terminals
- Each plane follows same Slim Fly network generation method
- Terminal to Router mapping is alternating mirrored on each new plane
  - Increase path diversity
  - Increases number of 1-hop routers



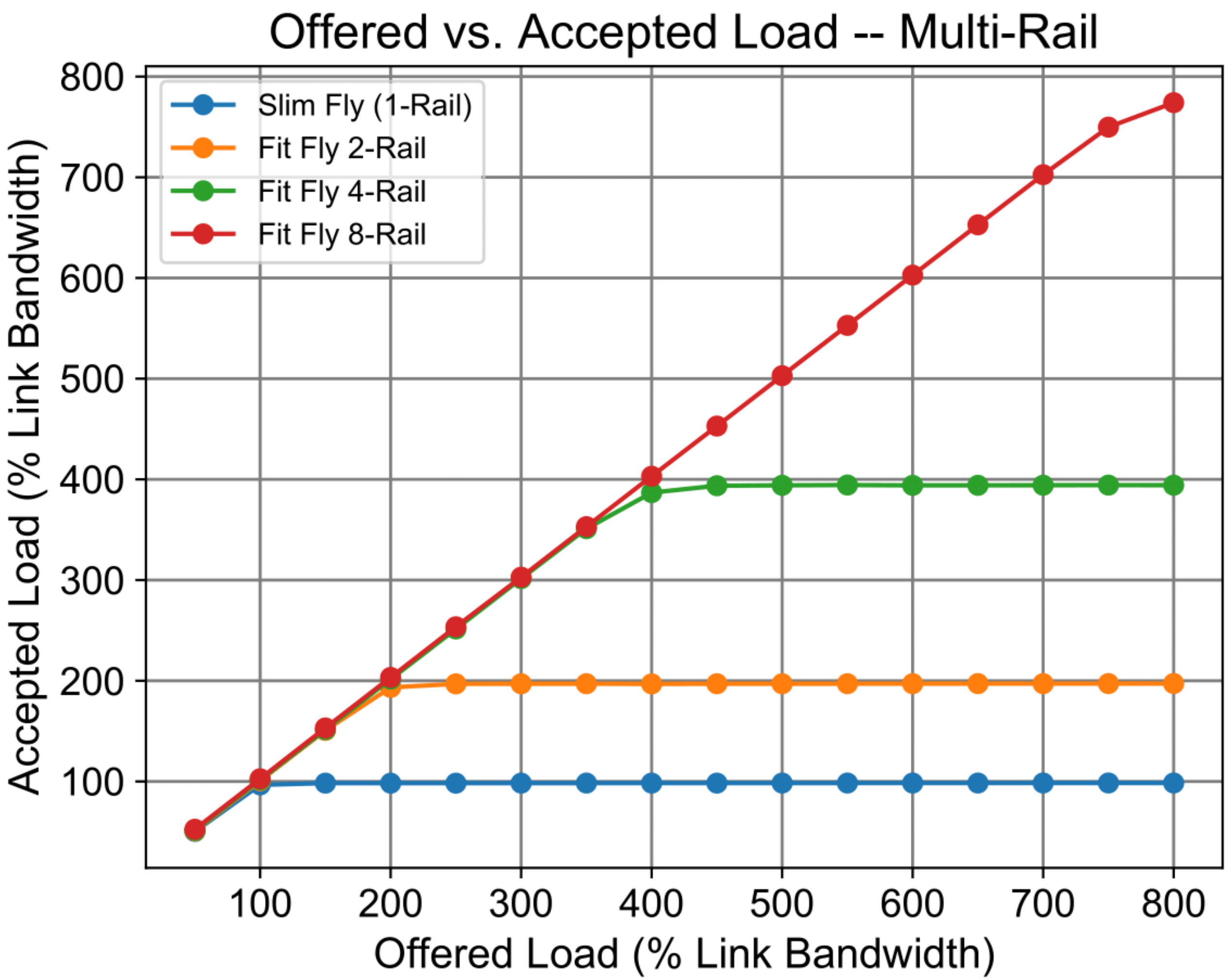
Slim Fly

Fit Fly



Fit Fly Validation

- Based on previously validated Slim Fly Model
- Additional planes bring additional throughput
- Observed expected increase in throughput with synthetic uniform random traffic
- Conducted visualization tests to make sure all links are used as expected.



Workloads

- **DOE Design Forward Application Traces (dumpi format)**
  - Algebraic MutliGrid Solver (AMG) @1728 ranks – mini-app for unstructured mesh physics that spends over 50% of it's time in comms.
  - MultiGrid (MG) @1000 ranks –- mini-app for adaptive mesh refinement that spends near 4% of time in comms.
- **Synthetic Background**
  - 1000 ranks
  - Uniform Random w/ mean Interval 100μs
  - Varied payload size for different levels of intensity
- **Compute Cluster**
  - All runs used upto 128 MPI ranks across 8 nodes of Intel/Xeon cluster.
  - Runtimes: 18 mins wall-clock worst case for up to 35ms of simulated network traffic.

Application	Background Intensity (% Link Bandwidth)						
AMG1728	0	2	4	7.5	15	36.25	72.5*
MG1000	0	2	4	7.5	15	36.25	72.5*

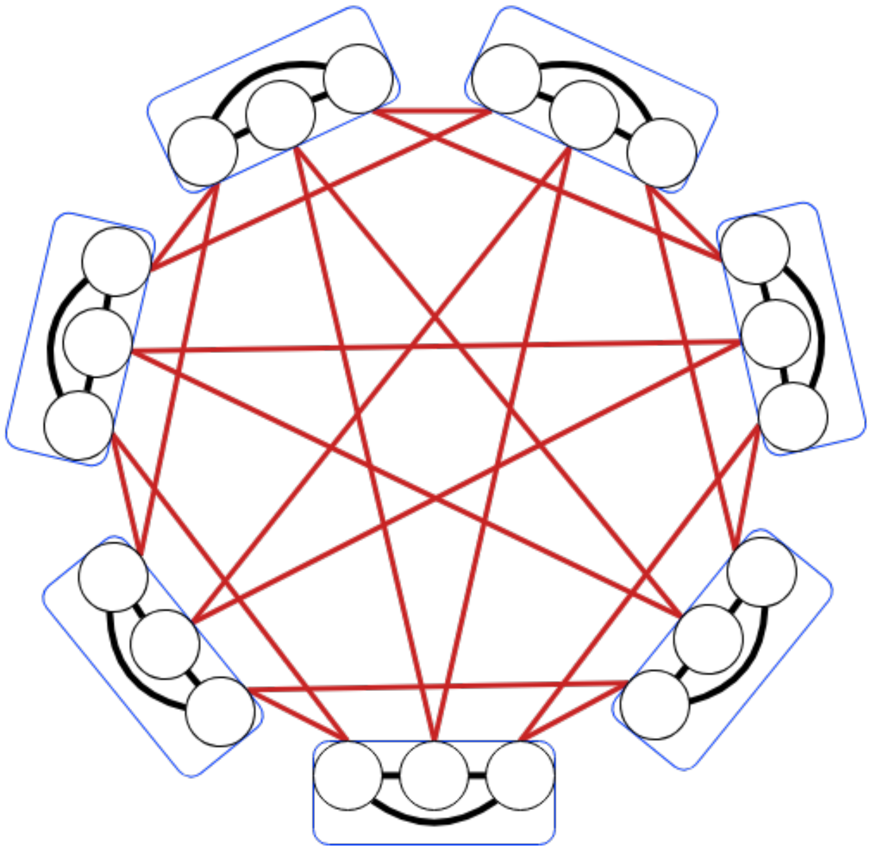
Evaluated Metrics

- **Maximum Communication Time**
  - Total amount of time spent by any one rank of the primary workload from the first MPI message it sends to the final message
- **Average Packet Latency**
  - Measures how long on average packets spent in transit
  - Correlated with communication time but application agnostic
- **Average Hop Count**
  - Measures mean number of routers visited by packets in route to destination

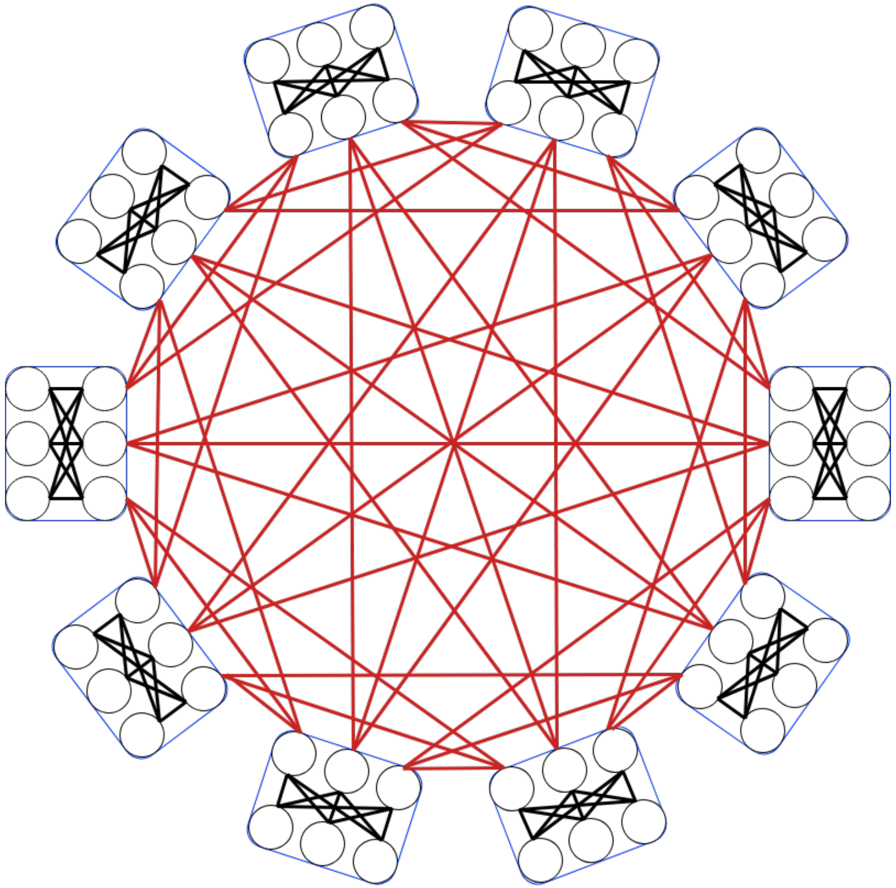
These three metrics give insight into induced congestion in the networks and their ability to manage increasing levels of interference traffic

Compared Networks

- Slim Fly, Fit Fly, **1D Dragonfly and Megafly**



1D Dragonfly



Megafly
(Dragonfly Plus)

Compared Networks

	Slim Fly	Fit Fly	Dragonfly	Megafly
Router Radix	28	28	36	36
Planes	1	2	1	1
Rails	1	2	1	1
Groups	26	52	19	10
Node Count	3042	3042	3078	3240
Router Count	338	676	342	360
Global Connections	4732	9464	3078	3240
Nodes/Group/Rail	117	117	162	324
Global Connections / Group	169	169	162	324
Link Bandwidth	12.5 GiB/s	12.5 GiB/s	12.5 GiB/s	12.5 GiB/s
Nodes per Router	9	9	9	18 (Leaves only)
Routing Algorithm	Adaptive (UGAL)	Adaptive (UGAL)	Adaptive (PAR) [35]	Adaptive (PAR)
Planar Selection Scheme	n/a	CONGESTION	n/a	n/a

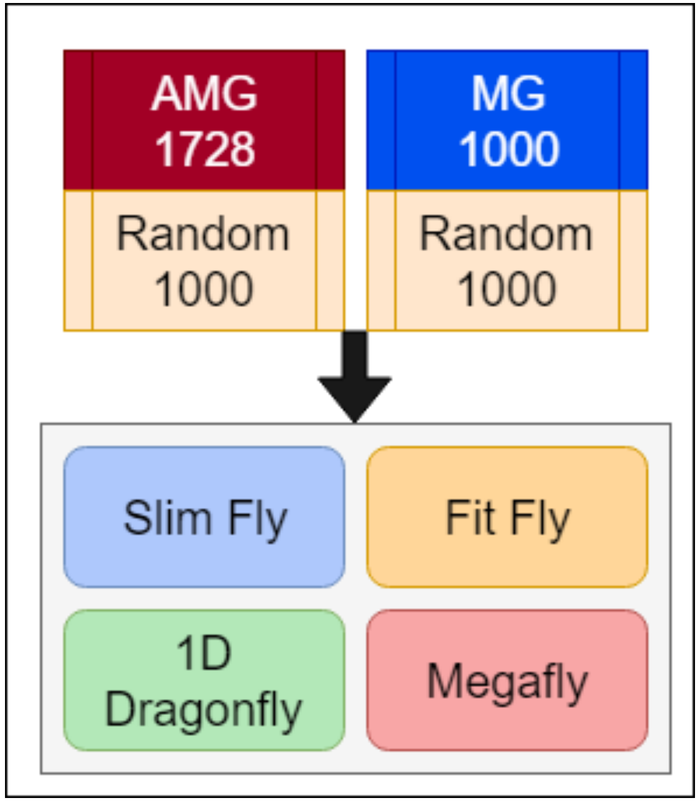
Fit Fly Bandwidth Considerations

- Fit Fly has distinct advantage due to its increased bandwidth and routers
  - Not super fair comparison
- Configure Slim Fly and Fit Fly so that they are of comparable throughput

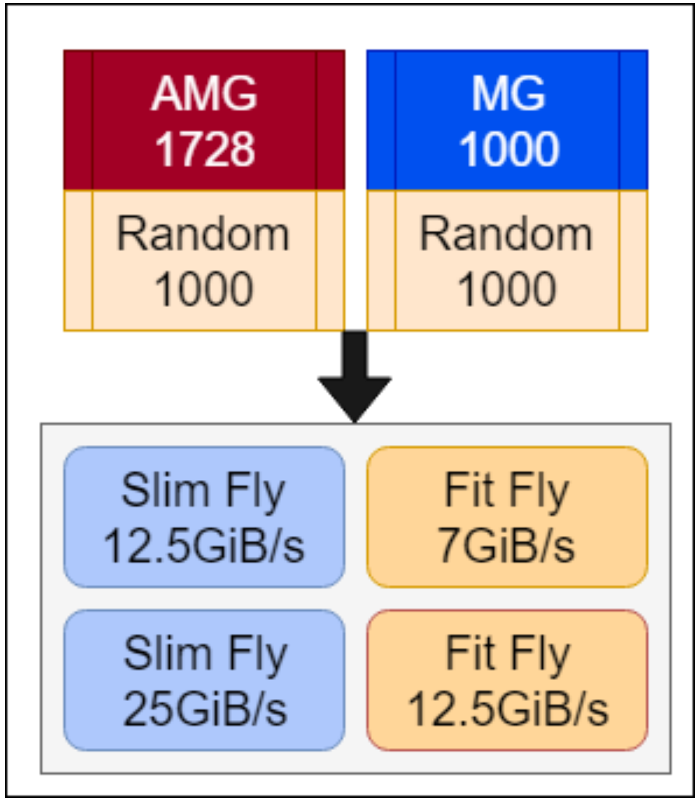
Additionally Tested Configuration Pairs	
Slim Fly (12.5GiB/s) (EDR)	**Fit Fly (7GiB/s) (FDR)**
Slim Fly (25GiB/s) (HDR)	Fit Fly (12.5GiB/s) (EDR)

Experiments Overview

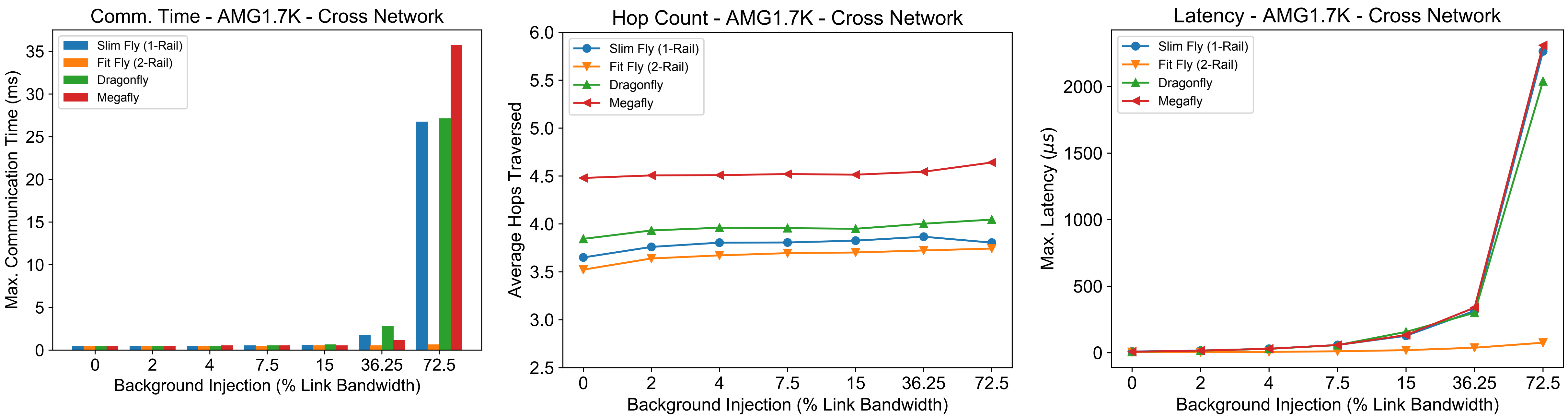
Experiment Set 1
Cross Network



Experiment Set 2
Equalized Bandwidth



Cross Network – AMG1728

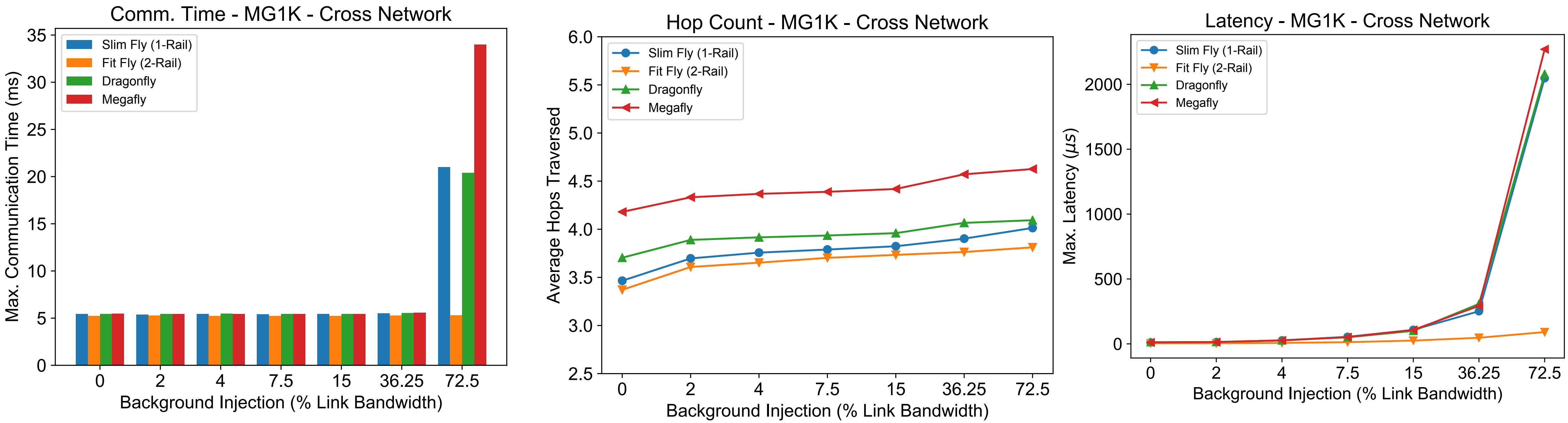


Note: link bandwidth is normalized per node/terminal
Fit Fly has distinct advantage here!

Max. Communication Time @ 72.5% Background Injection

Slim Fly	Fit Fly	Dragonfly	Megafly
26.4ms	**0.67ms**	27.1ms	35.7ms

Cross Network – MG1000



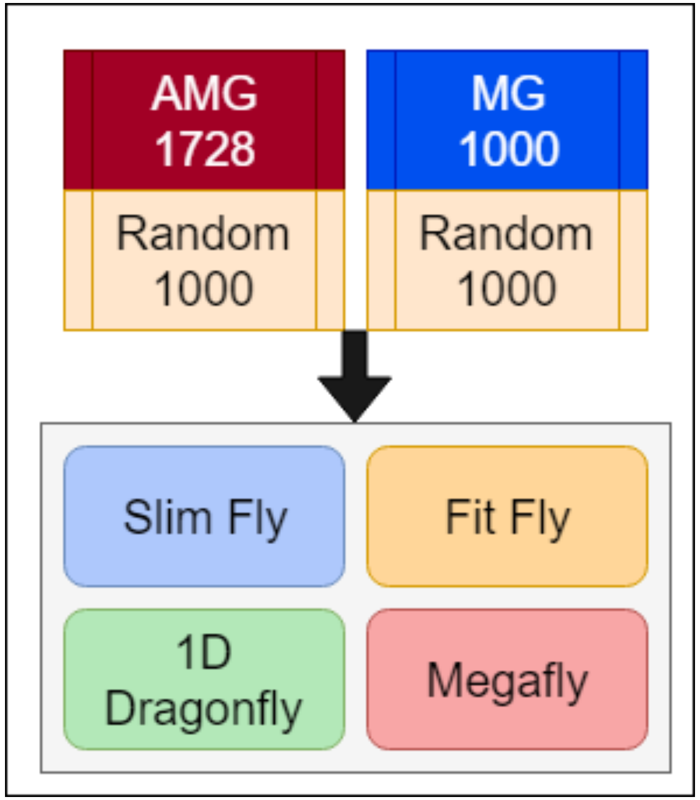
Max. Communication Time @ 72.5% Background Injection

Slim Fly	Fit Fly	Dragonfly	Megafly
21.0ms	**5.3ms**	20.3ms	34.0ms

Discussion: Cross Network

- Slim Fly performed well against state of the art Dragonfly and Megafly networks
  - Possible candidate for future networks?
- Fit Fly showed great resilience to high levels of interference traffic
  - Beat Slim Fly by an order of magnitude
  - Remember Fit Fly finished in < 1 millisecond even at high interference
- **Slim Fly and Fit Fly networks show great promise**
  - Low-diameter-high-path-diversity

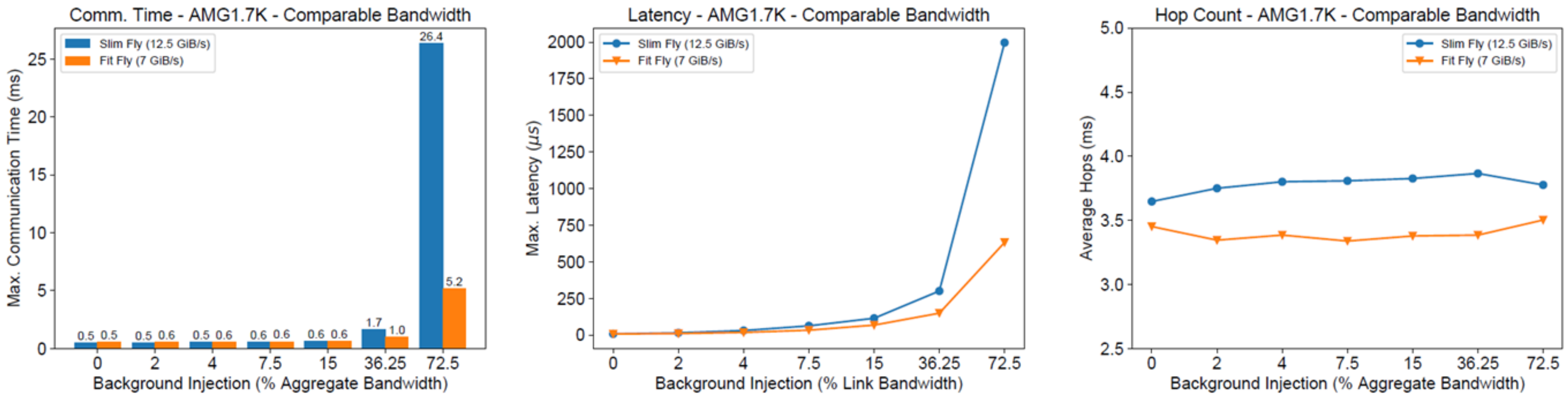
Experiment Set 1
Cross Network



Equalized Bandwidth (12.5GiB/s) – AMG1728

Note: agg. bandwith is used to make FF and SF more comparable

Tested Configuration Pair	
Slim Fly (12.5GiB/s)	**Fit Fly (7GiB/s)**

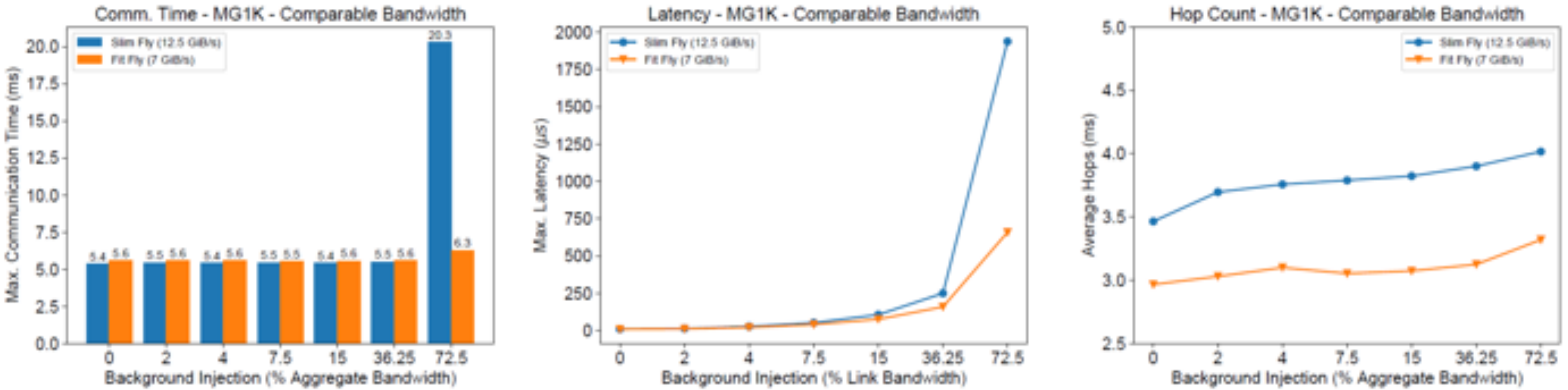


(a) Application Communication Time (b) Latency (c) Hop Count

Figure 6: Synthetic interference experiments on the AMG1728 trace workload with 1,000 synthetic background ranks. Link bandwidth of Slim Fly in this case is 12.5 GiB/s (≈InfiniBand EDR) while Fit Fly is 7 GiB/s (≈InfiniBand FDR). Total aggregate bandwidth is calculated by $B_L \cdot P$, where B_L is the bandwidth of each link in the network.

Equalized Bandwidth (12.5GiB/s) – MG1000

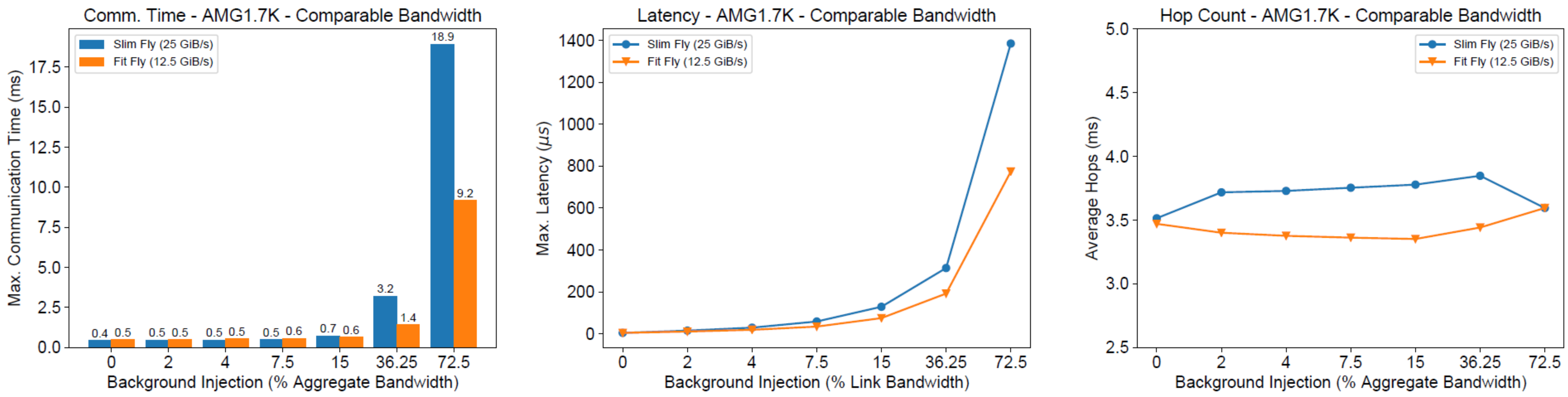
Tested Configuration Pair	
Slim Fly (12.5GiB/s)	**Fit Fly (7GiB/s)**



(a) Application Communication Time (b) Latency (c) Hop Count

Figure 7: Synthetic interference experiments on the MultiGrid1000 trace workload with 1,000 synthetic background ranks. Link bandwidth of Slim Fly in this case is 12.5 GiB/s (≈InfiniBand EDR) while Fit Fly is 7 GiB/s (≈InfiniBand FDR). Total aggregate bandwidth is calculated by $B_L \cdot P$, where B_L is the bandwidth of each link in the network.

Equalized Bandwidth (25GiB/s) – AMG1728

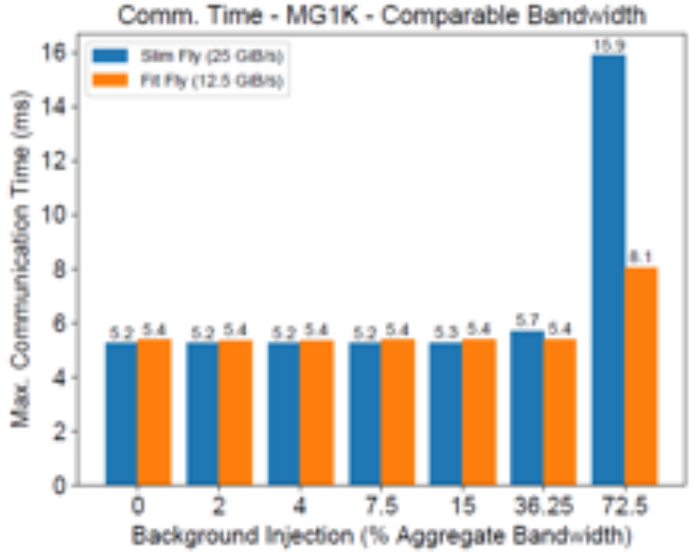


(a) Application Communication Time **(b) Latency** **(c) Hop Count**

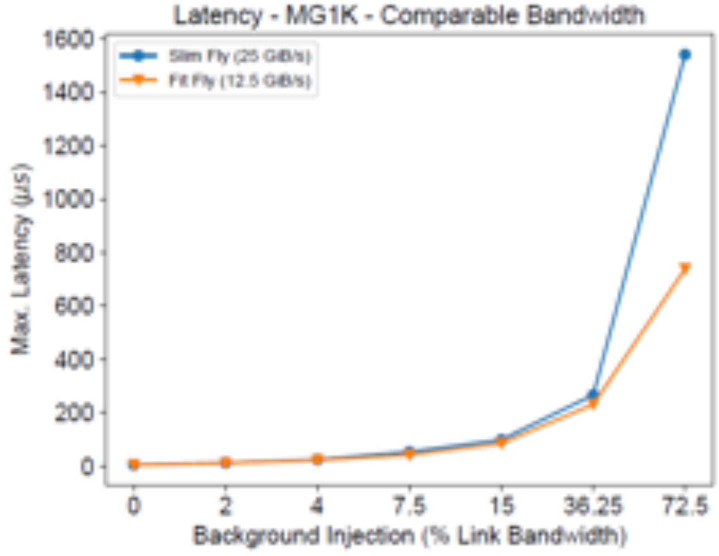
Figure 8: Synthetic interference experiments on the AMG1728 trace workload with 1,000 synthetic background ranks. Link bandwidth of Slim Fly in this case is 25 GiB/s (≈InfiniBand HDR) while Fit Fly is 12.5 GiB/s (≈InfiniBand EDR). Total aggregate bandwidth is calculated by $B_L \cdot P$, where B_L is the bandwidth of each link in the network.

Tested Configuration Pair	
Slim Fly (25GiB/s)	Fit Fly (12.5GiB/s)

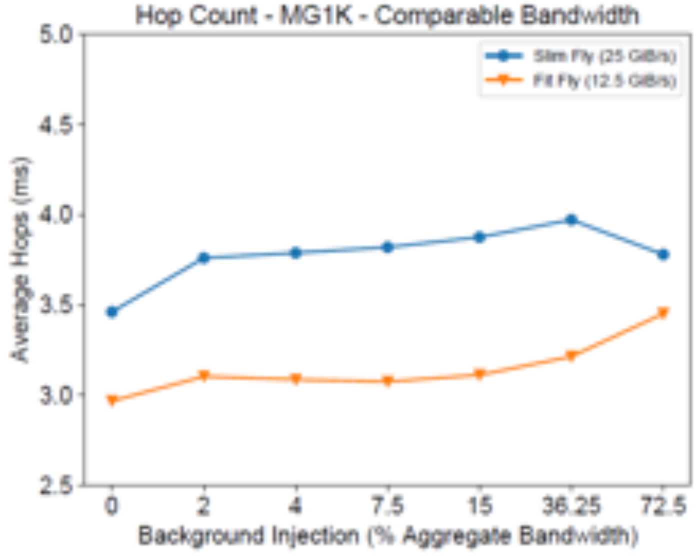
Equalized Bandwidth (25GiB/s) – MG1000



(a) Application Communication Time



(b) Latency



(c) Hop Count

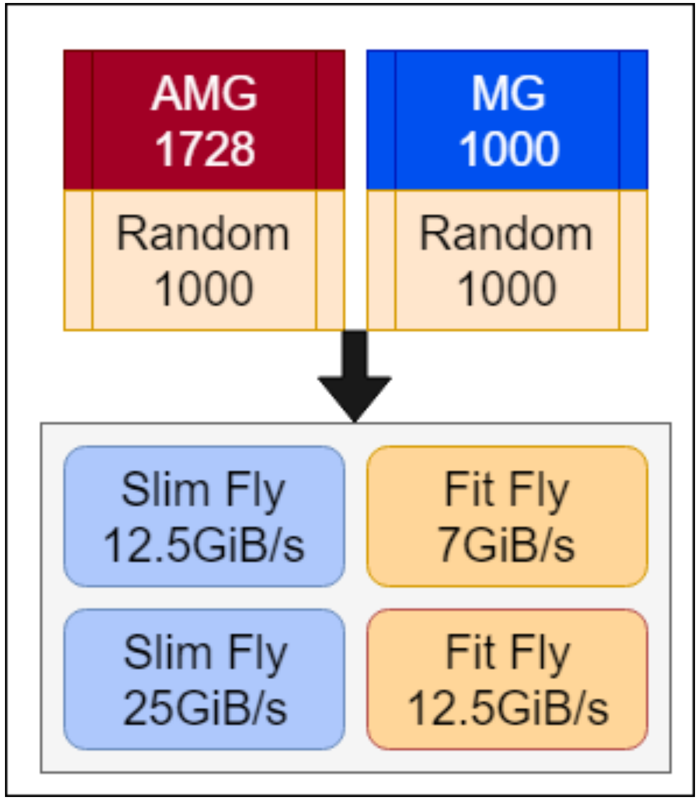
Figure 9: Synthetic interference experiments on the MultiGrid1000 trace workload with 1,000 synthetic background ranks. Link bandwidth of Slim Fly in this case is 25 GiB/s (≈InfiniBand HDR) while Fit Fly is 12.5 GiB/s (≈InfiniBand EDR). Total aggregate bandwidth is calculated by $B_L \cdot P$, where B_L is the bandwidth of each link in the network.

Tested Configuration Pair	
Slim Fly (25GiB/s)	Fit Fly (12.5GiB/s)

Discussion: Equalized Bandwidth

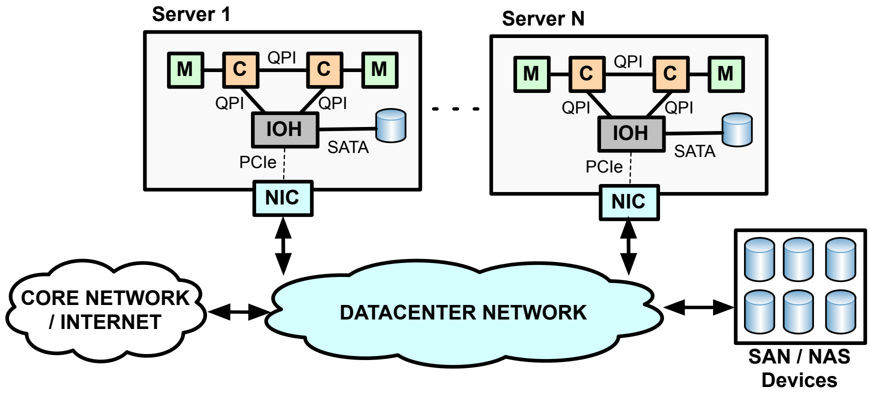
- Equalizing the aggregate bandwidth across networks slightly reduced the advantage that Fit Fly had
  - Fit Fly still pulled ahead
    - Greater interference resilience
- Additional planes of routers give less chance for any two packets to interact
  - Less interference
  - Less buffer wait time
  - Increased Application Performance
- **More planes of cheaper routers may be a better option to single-plane-high-bandwidth networks**

Experiment Set 2
Equalized Bandwidth

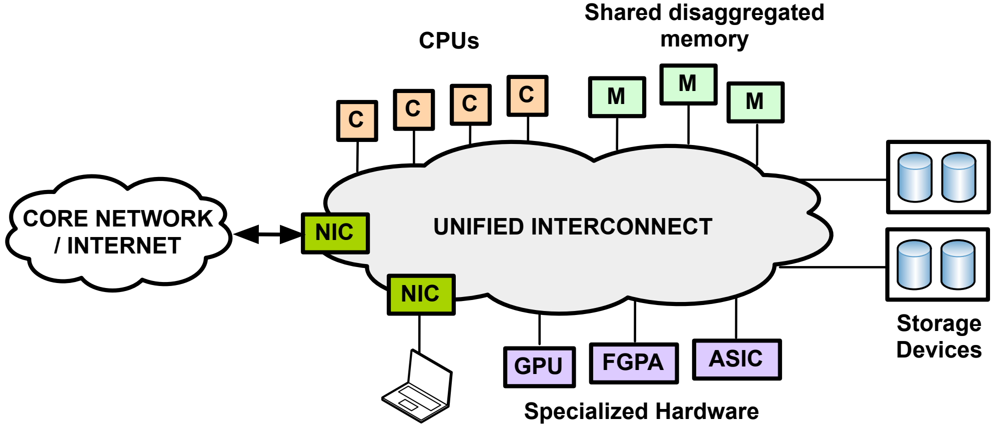


Conclusion & Future Work

- Slim Fly networks show promise in comparison to current networks designs.
  - Fit Fly appears to yield better performance than Slim Fly for tested workloads even at higher cost
  - Additional routers planes lower chance of interference
  - Equalizing overall bandwidth throughput, additional planes give strong advantage
- Multi-rail, multi-plane design lead toward considering disaggregated SC network architectures in the future
- **CODES provides a strong environment for answering "What If…" questions and fostering future innovation in the field of HPC interconnection networks**



(a) Current datacenter



(b) Disaggregated datacenter

From Gao et al OSDI 2016.