EVALUATING THE IMPACT OF SPIKING NEURAL NETWORK TRAFFIC ON EXTREME-SCALE HYBRID SYSTEMS

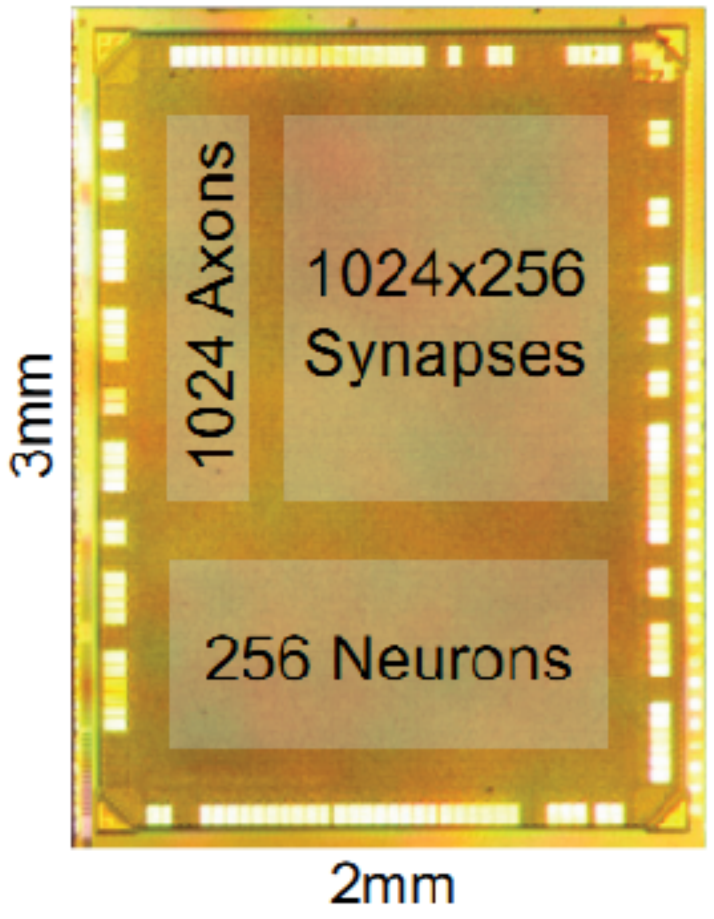
Noah Wolfe[1], Mark Plagge[1], Misbah Mubarak[2], Robert Ross[2], Christopher Carothers[1]

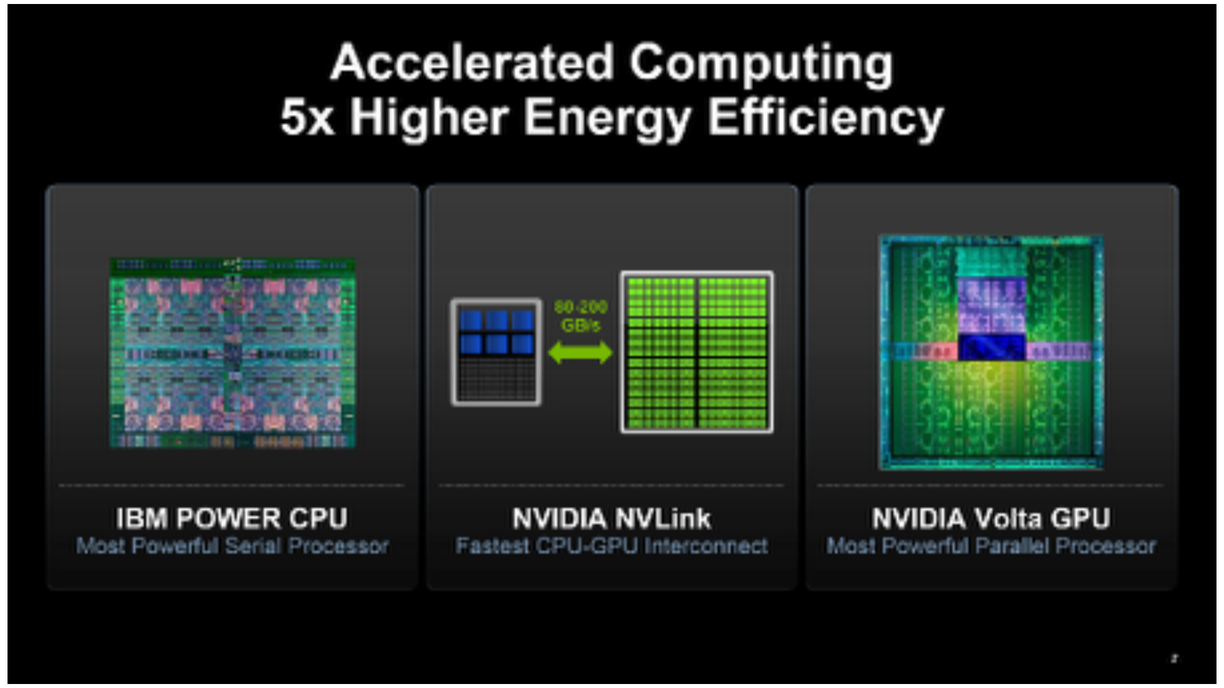
[1] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180
[2] Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439

Animations: Caitlin Ross[1]

November 12th, 2018

Hybrid Neuromorphic Supercomputer?





The question: how might a neuromorphic "accelerator" type processor be used to improve the HPC application performance, power consumption and overall system reliability of future exascale systems?

Driven by the recent DOE SEAB report on high-performance computing which highlights the neuromorphic architecture as one that *"is an emergent area for exploitation".*

Address using parallel systems simulation

Simulation Workflow

TrueNorth (Neuromorphic Design)

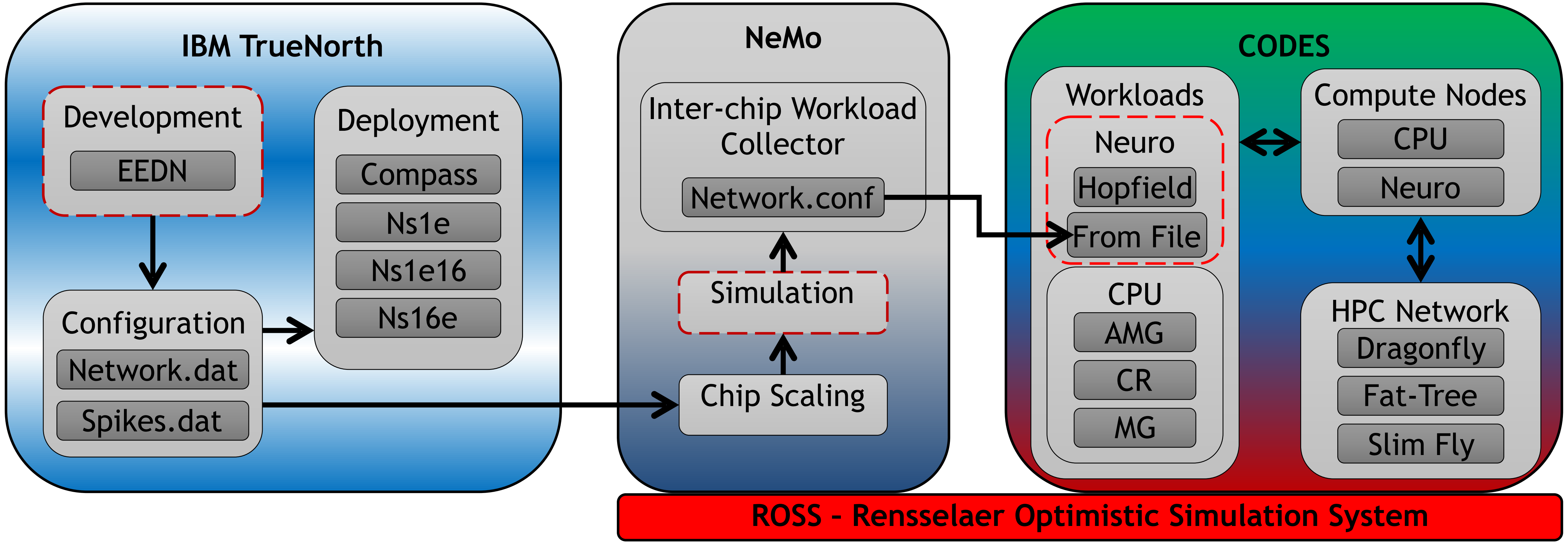
- Ecosystem for building real neural network applications for the IBM neuromorphic processor
- Spiking neuron based architecture with 4,096 neurosynaptic cores resulting in 1M neurons

NeMo (Neuromorphic Scaling)

- Framework for simulating general purpose neuromorphic processors
- Capable of generating partially synthetic multi-chip workloads from TrueNorth applications
- Built on ROSS

CODES (HPC)

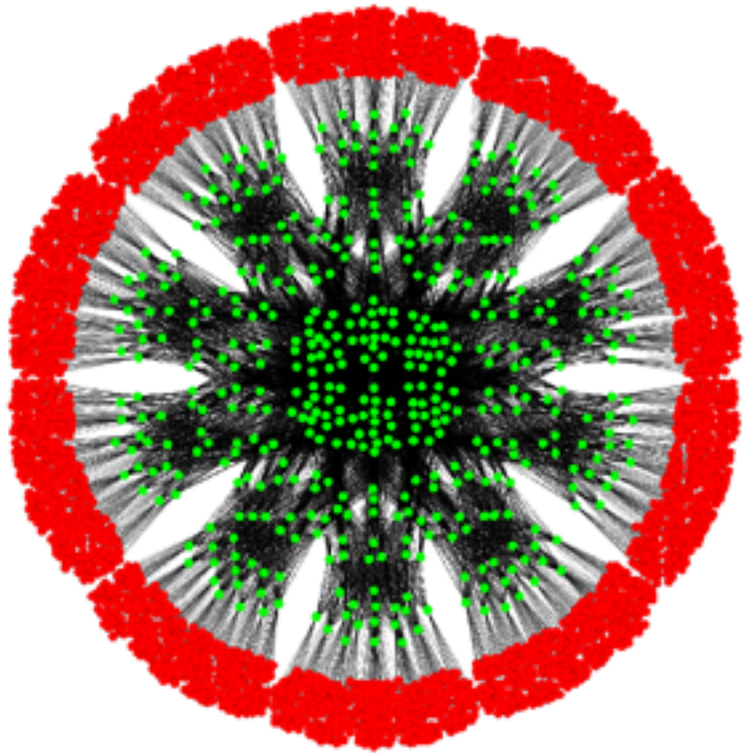
- High-fidelity packet-level framework for exploring design of HPC interconnects and workloads
- Synthetic or application trace network workloads
- Built on ROSS



HPC Interconnection Networks

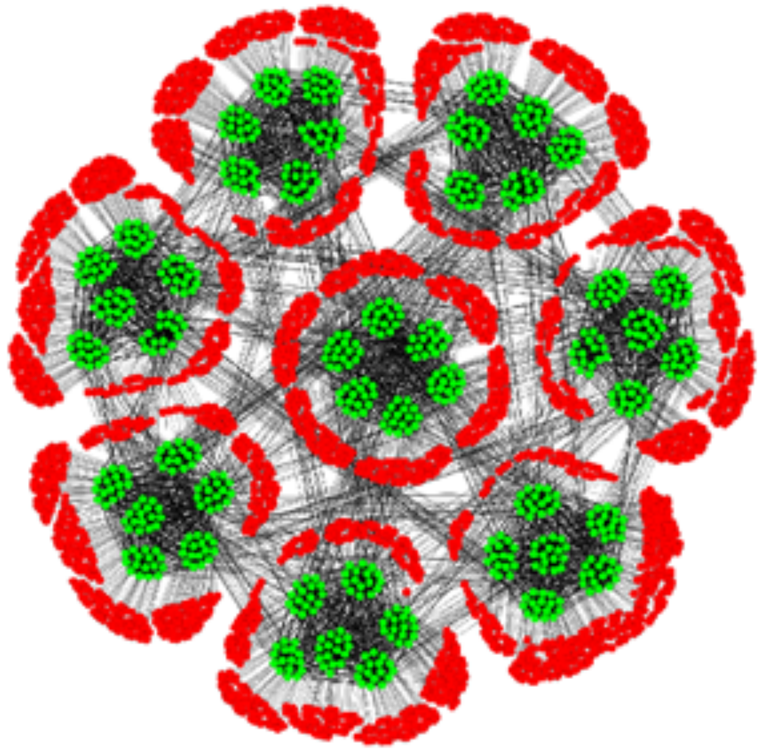
Fat-Tree

- Nodes: 3,240
- Routers: 468
- Links: 9,720
- Levels: 3
- *Net Diameter: 4
- Routing: Static
- WC Hop Count: 6
- Inj. BW: 316 Tbps
- Net BW: 648 Tbps



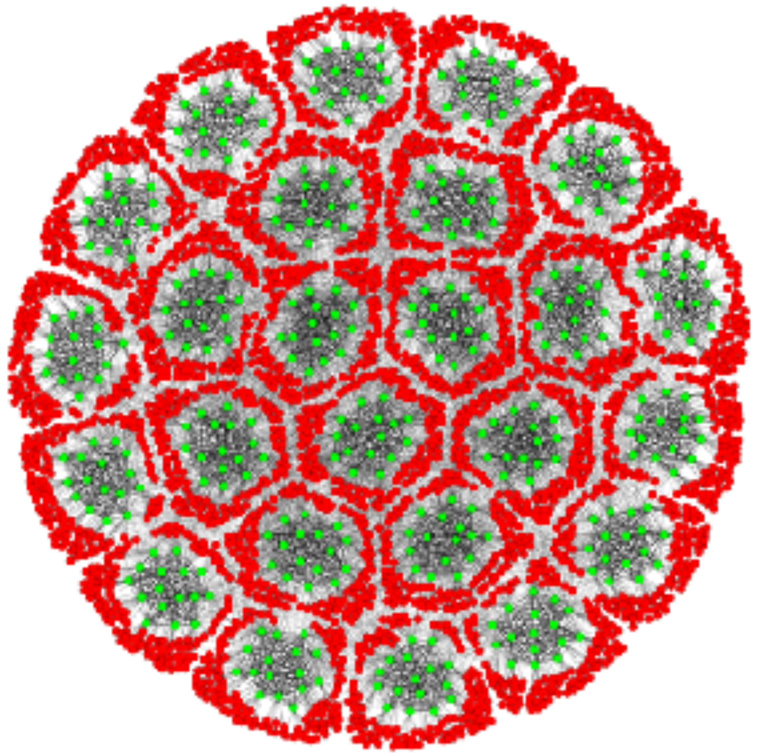
Dragonfly-2D

- Nodes: 3,072
- Routers: 768
- Links: 11,424
- Groups: 8
- *Net Diameter: 5
- Routing: Adaptive
- WC Hop Count: 12
- Inj BW: 300 Tbps
- Net BW: 835 Tbps



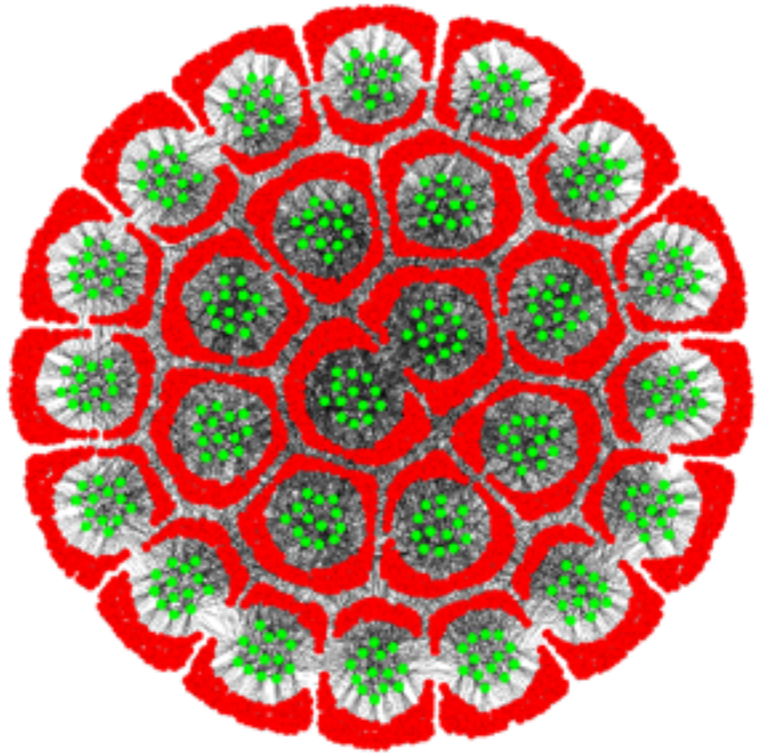
Dragonfly-1D

- Nodes: 3,200
- Routers: 400
- Links: 8,600
- Groups: 25
- *Net Diameter: 3
- Routing: Adaptive
- WC Hop Count: 7
- Inj BW: 313 Tbps
- Net BW: 540 Tbps



Slim Fly

- Nodes: 3,042
- Routers: 338
- Links: 6,253
- Groups: 26
- *Net Diameter: 2
- Routing: Adaptive
- WC Hop Count: 6
- Inj BW: 297 Tbps
- Net BW: 321 Tbps



*Net Diameter doesn't include NIC to router and router to NIC hops

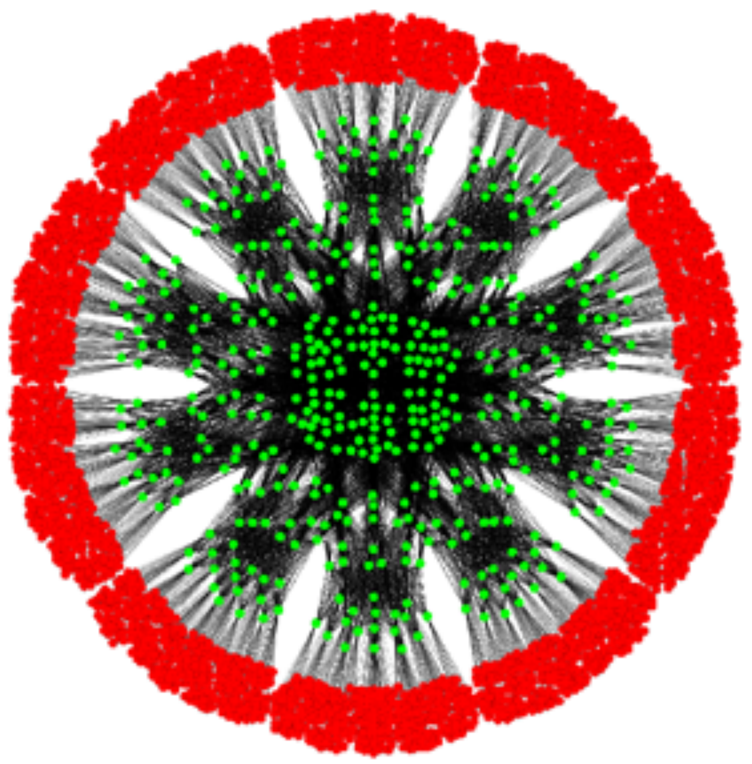
WC: worst-case, Inj: injection, BW: bandwidth, Net: network

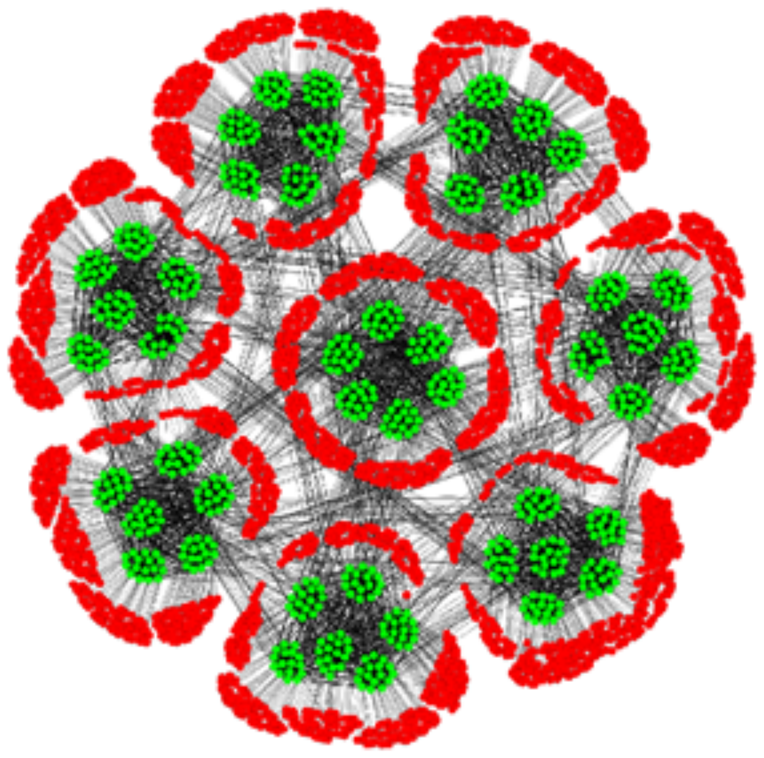
Simulation Parameters

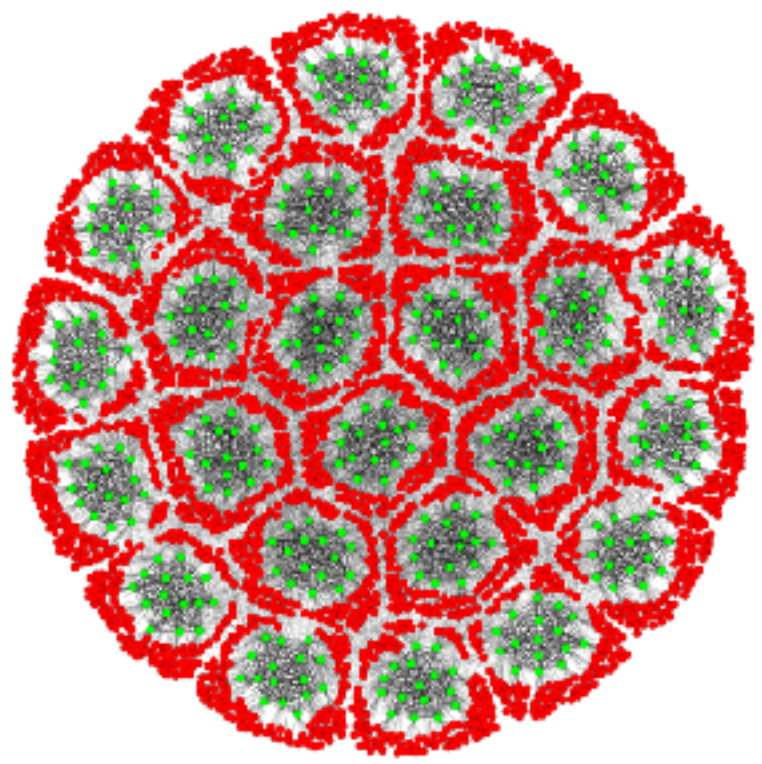
Fat-Tree	Dragonfly-2D	Dragonfly-1D	Slim Fly

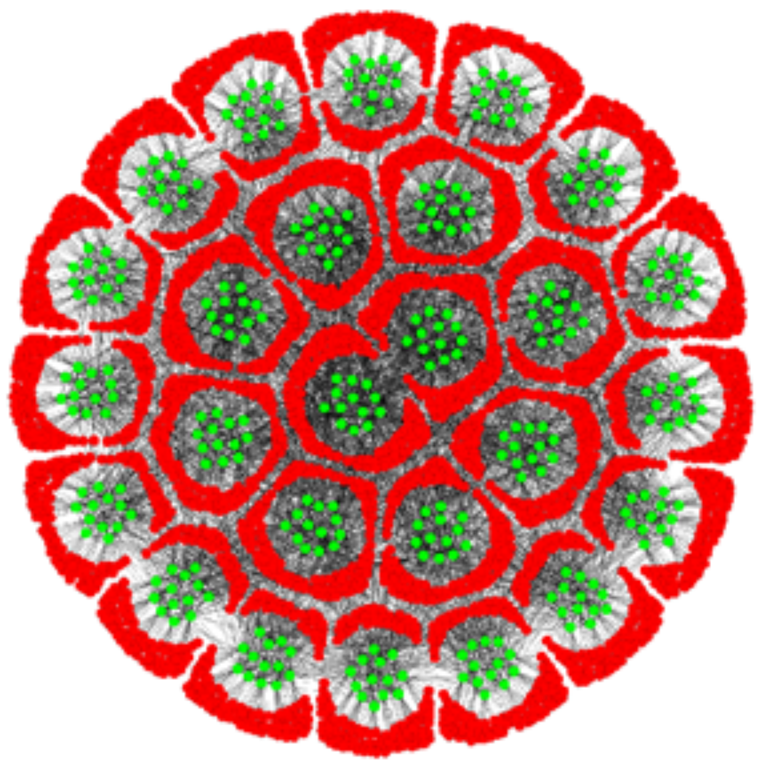
- Router Latency: 90ns
- NIC Latency: 1.5us
- MPI Latency: 2.5us
- Buffer Space per VC: 64KB
- Link Speed: 100Gbps

- Job Mapping: Contiguous
- System Utilization:
  - Neuromorphic: 33%
  - CPU: 33-55%









WC: worst-case, Inj: injection, BW: bandwidth, Net: network

Application Workloads

Neuromorphic Workloads

Workload	*Chips*	*Connectivity*	*Spikes/Tick*	*MB/Tick*	*Waits*
MNIST	1234	15 layers	577K	4.4MB	0
CIFAR	1024	15 layers	647K	4.9MB	0
Hopfield	1024	all-to-all	10.2M	80MB	0

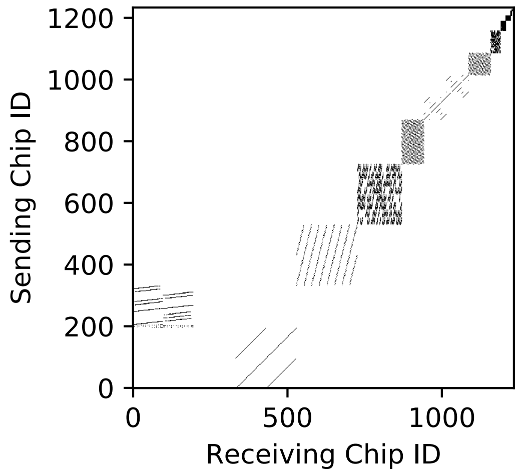
A tick is 1ms of simulated time.

CPU Workloads

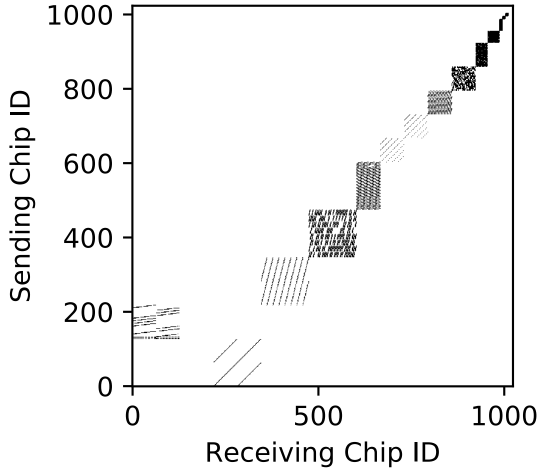
Workload	*Processes*	*End Time*	*Msgs*	*Msg Size*	*Waits*
AMG	1,728	0.50ms	2.2M	0.79KB	101.1K
CR	1,000	258.48ms	39.9M	7.95KB	39.9M
MG	1,000	5.51ms	0.5M	9.30KB	248K

End time is the virtual time to replay the workload through the Fat-Tree configuration. Msg size is the average size of all messages transfered across all processes.

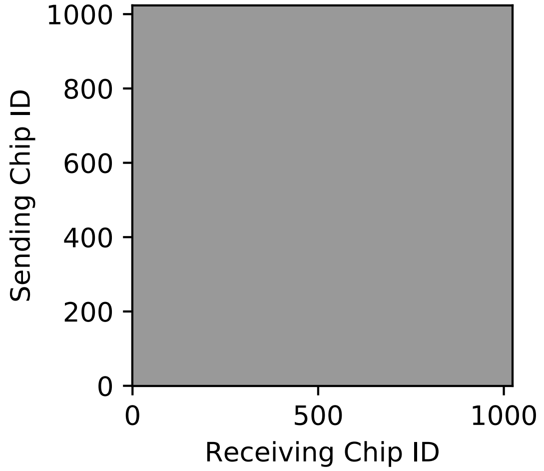
MNIST



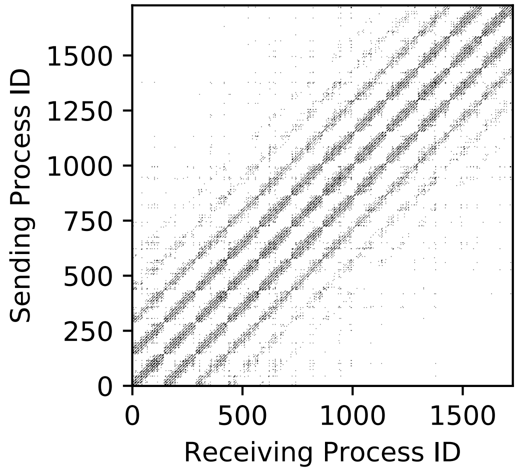
CIFAR



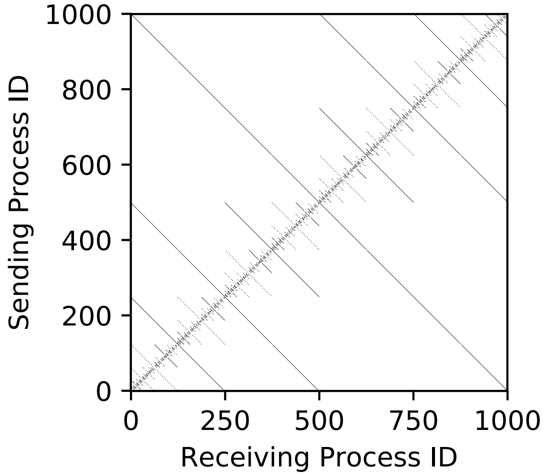
Hopfield



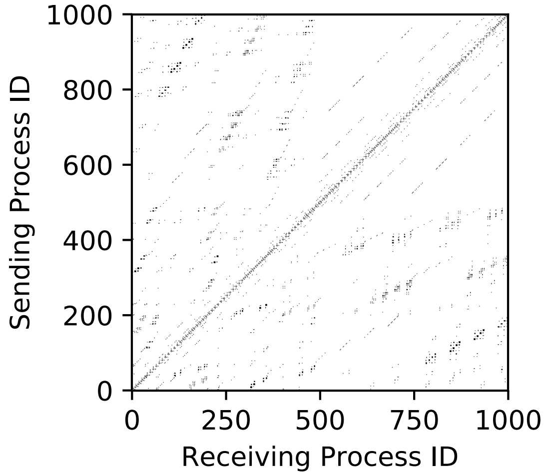
AMG



CR



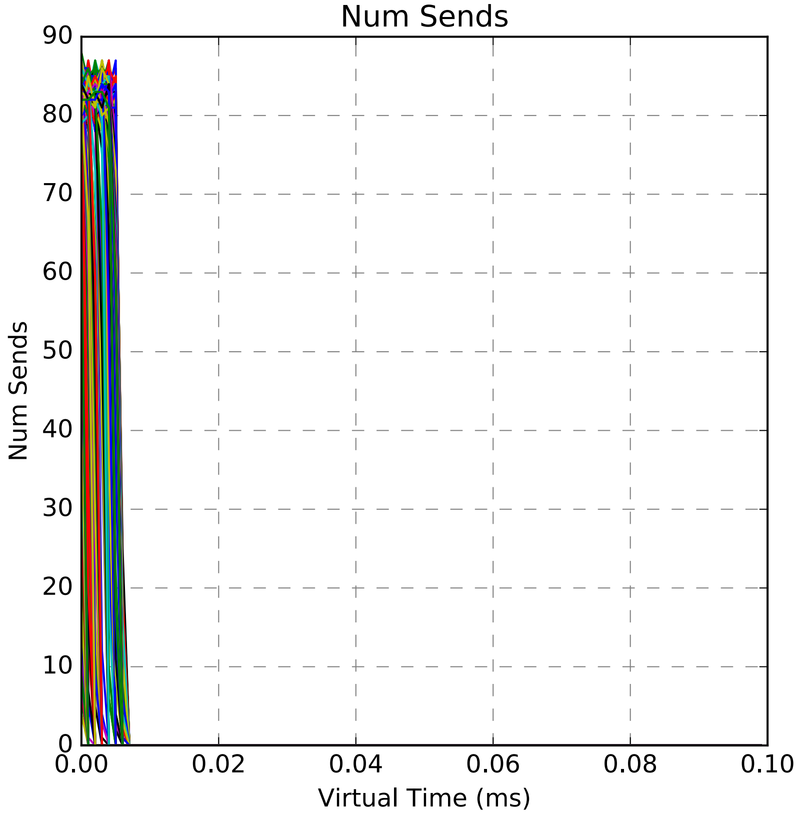
MG

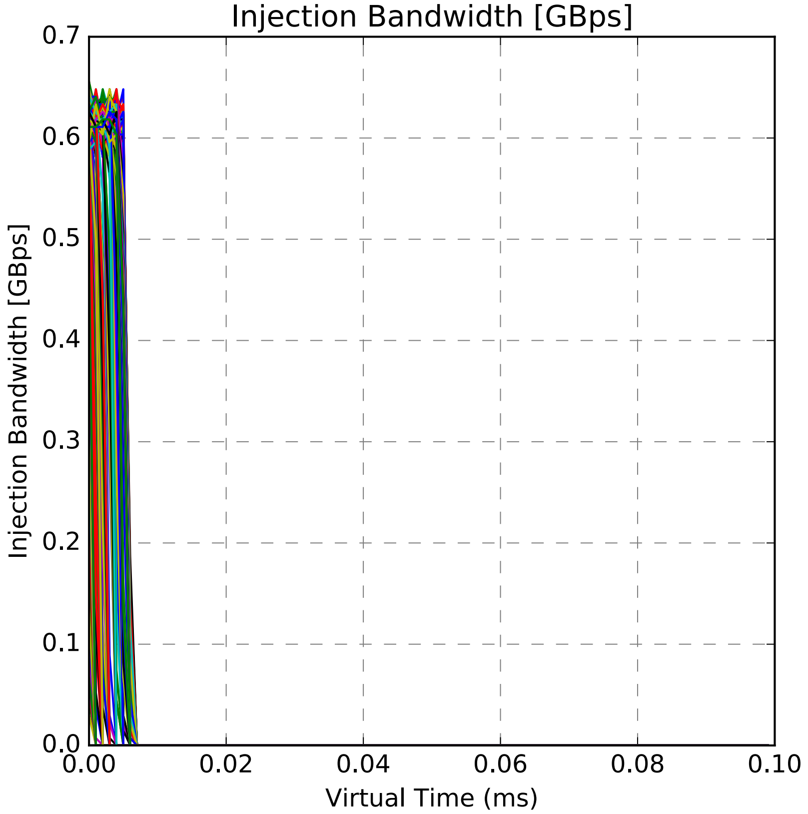


Application Workloads

MNIST
(Neuromorphic)

- **Description**: Convolutional neural network for handwritten digit classification
- **Communication Pattern**: Periodic injection of 8B messages between 15 interconnected layers of neurons
- **Trace Size**: 1,234 neuromorphic chips





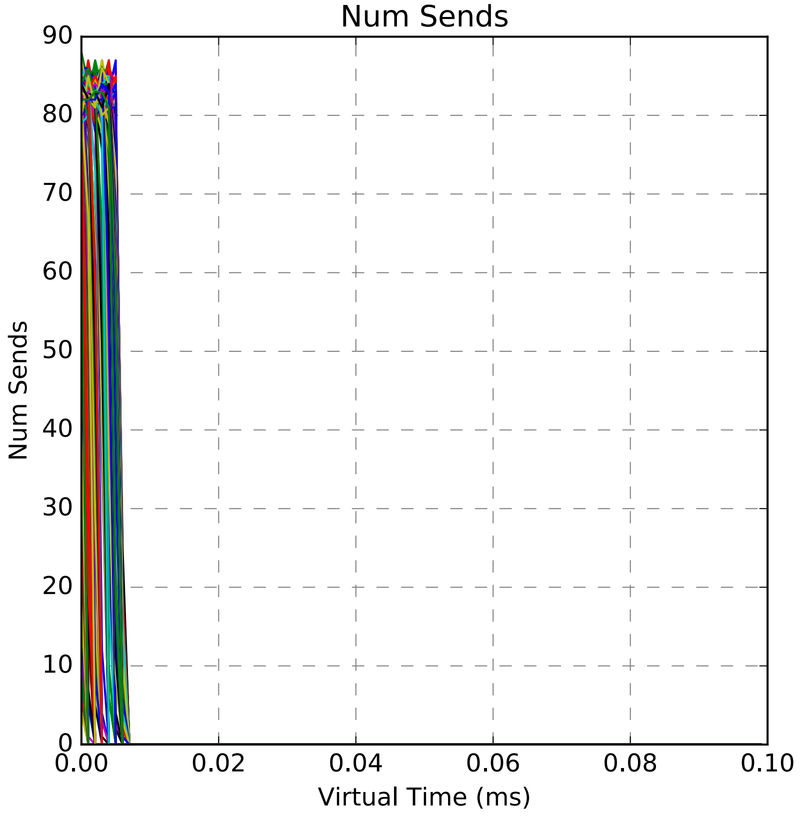
MNIST 1234 NEUROMORPHIC CHIP WORKLOAD

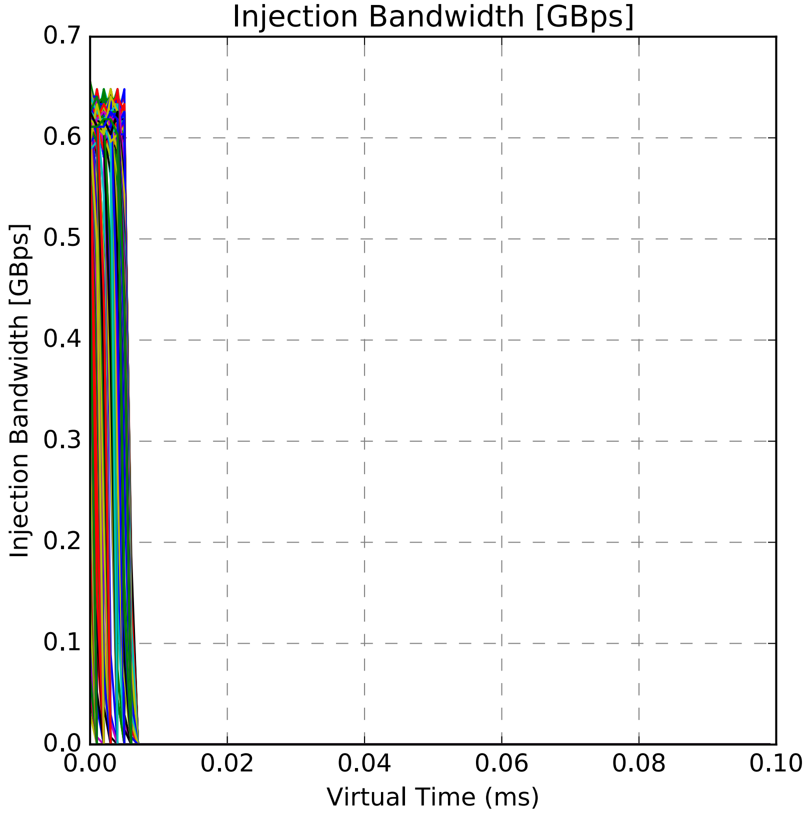
SLIMFLY NETWORK

Application Workloads

MNIST
(Neuromorphic)

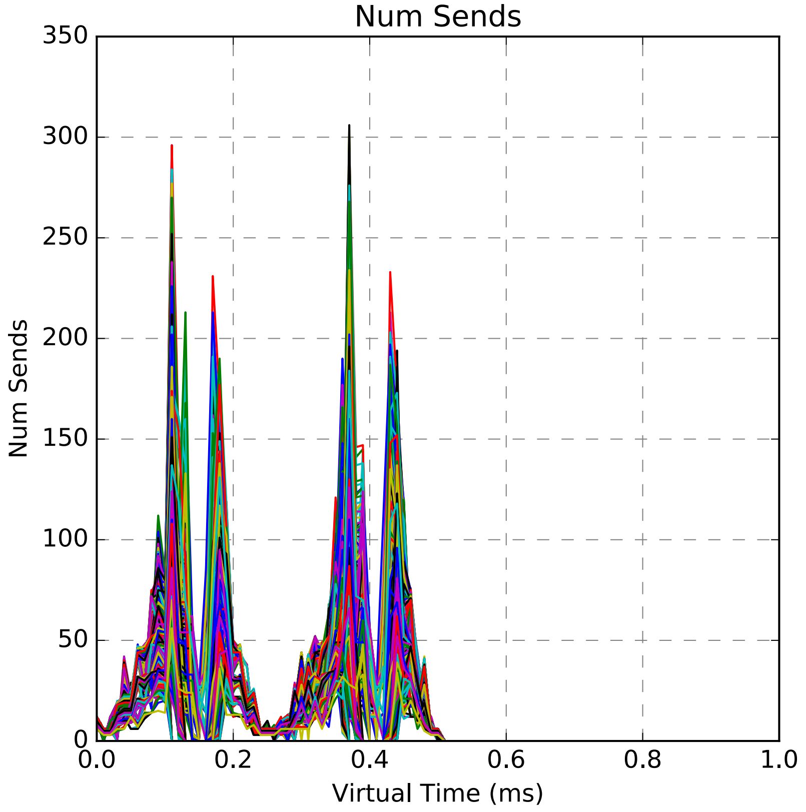
- **Description:** Convolutional neural network for handwritten digit classification
- **Communication Pattern:** Periodic injection of 8B messages between 15 interconnected layers of neurons
- **Trace Size:** 1,234 neuromorphic chips

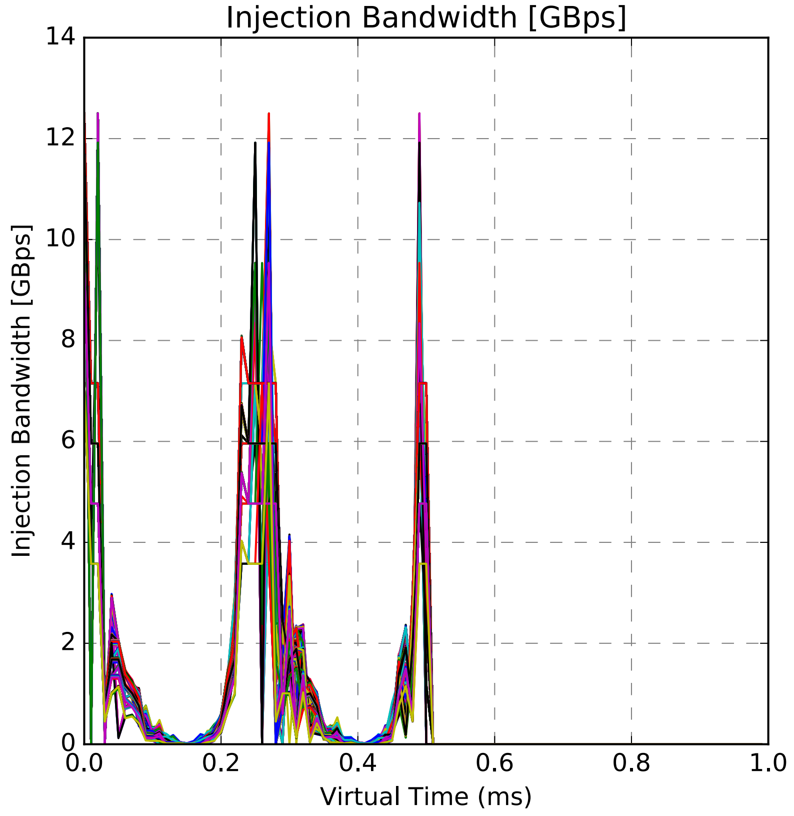




Algebraic Multigrid Solver
(HPC)

- **Description:** Parallel Algebraic Multigrid Solver (AMG) used for unstructured grids developed at LLNL
- **Communication Pattern:** Bursty periods of ~1KB messages following 3D nearest neighbor
- **Communication Time:** 52.9% of runtime
- **Trace Size:** 1,728 CPU processes



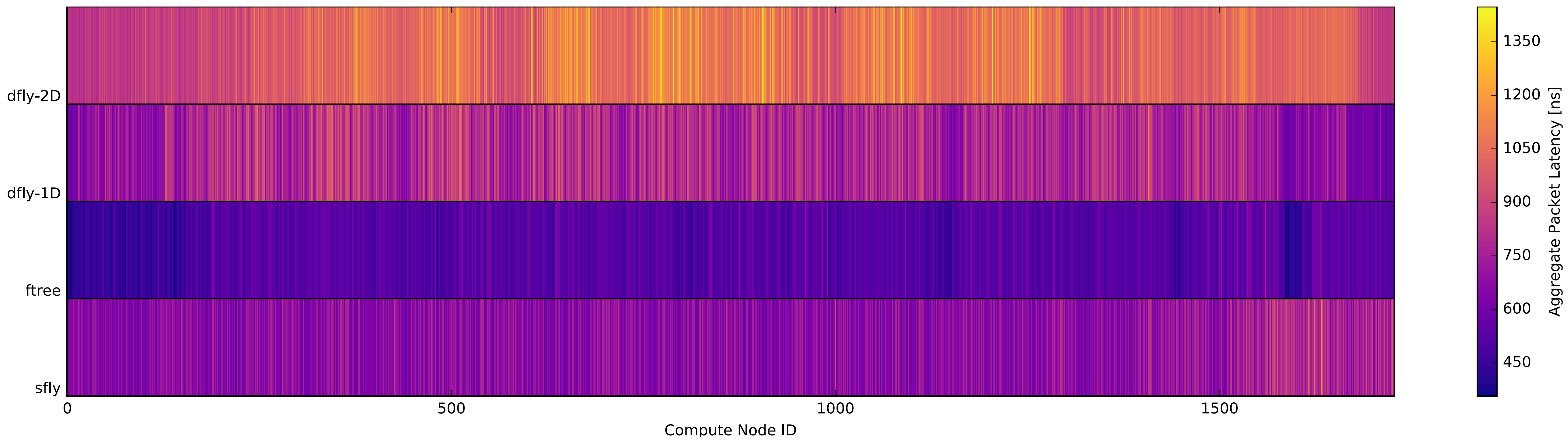


AMG 1728 MPI RANK WORKLOAD

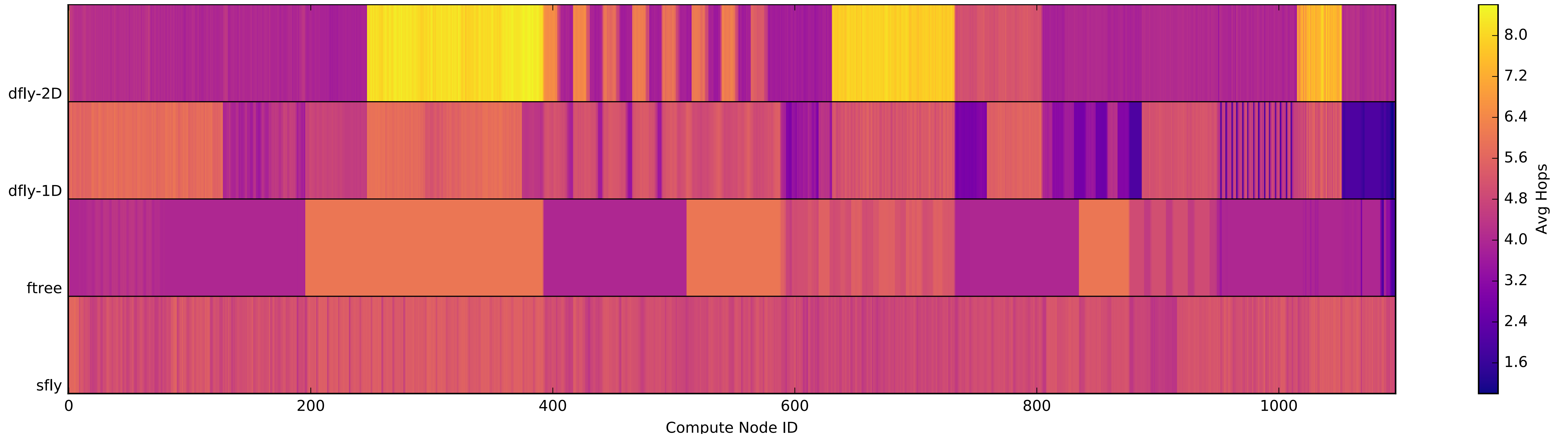
SLIMFLY NETWORK

Homogeneous Results

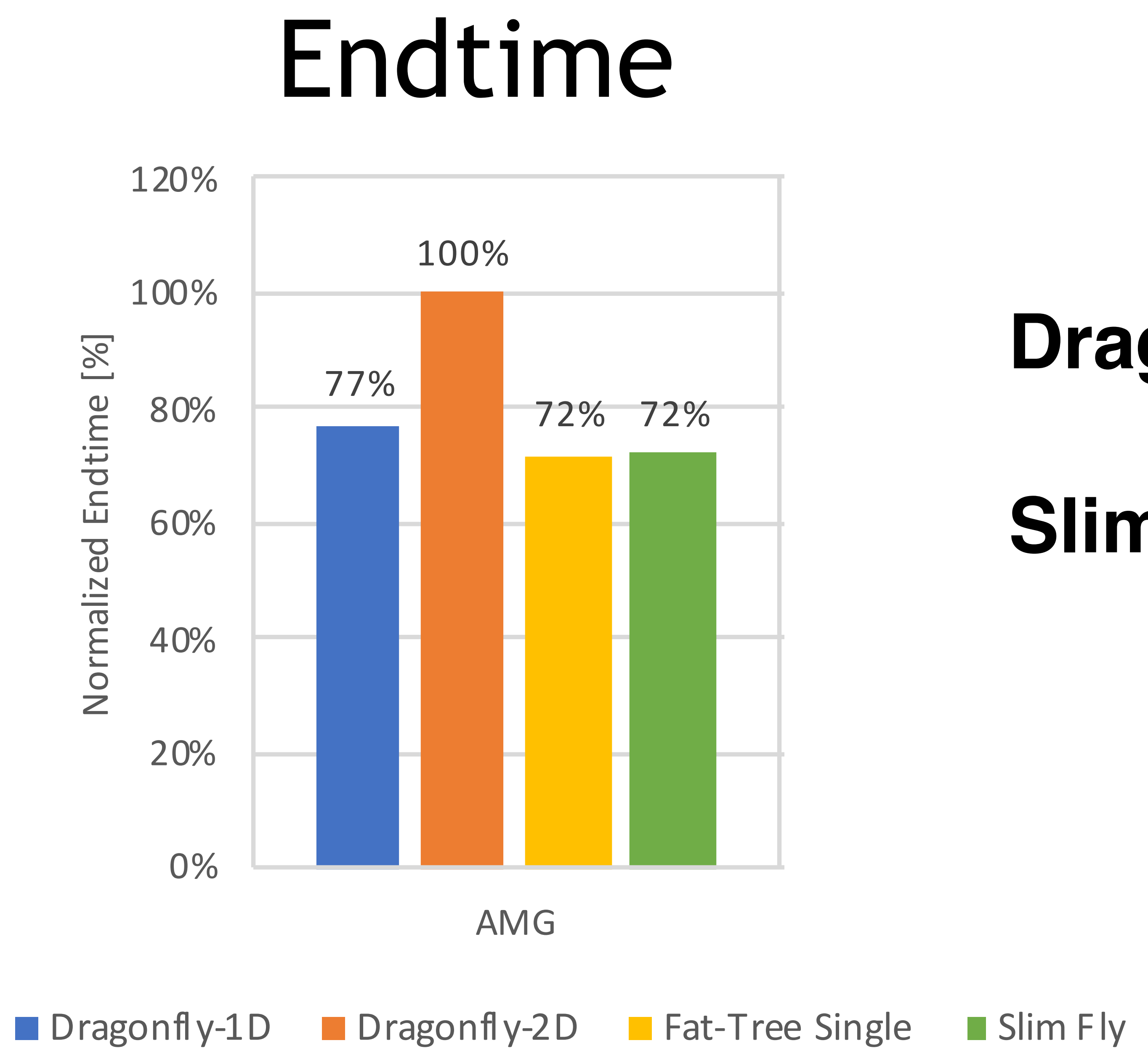
AMG



MNIST



Homogeneous Results
(AMG)

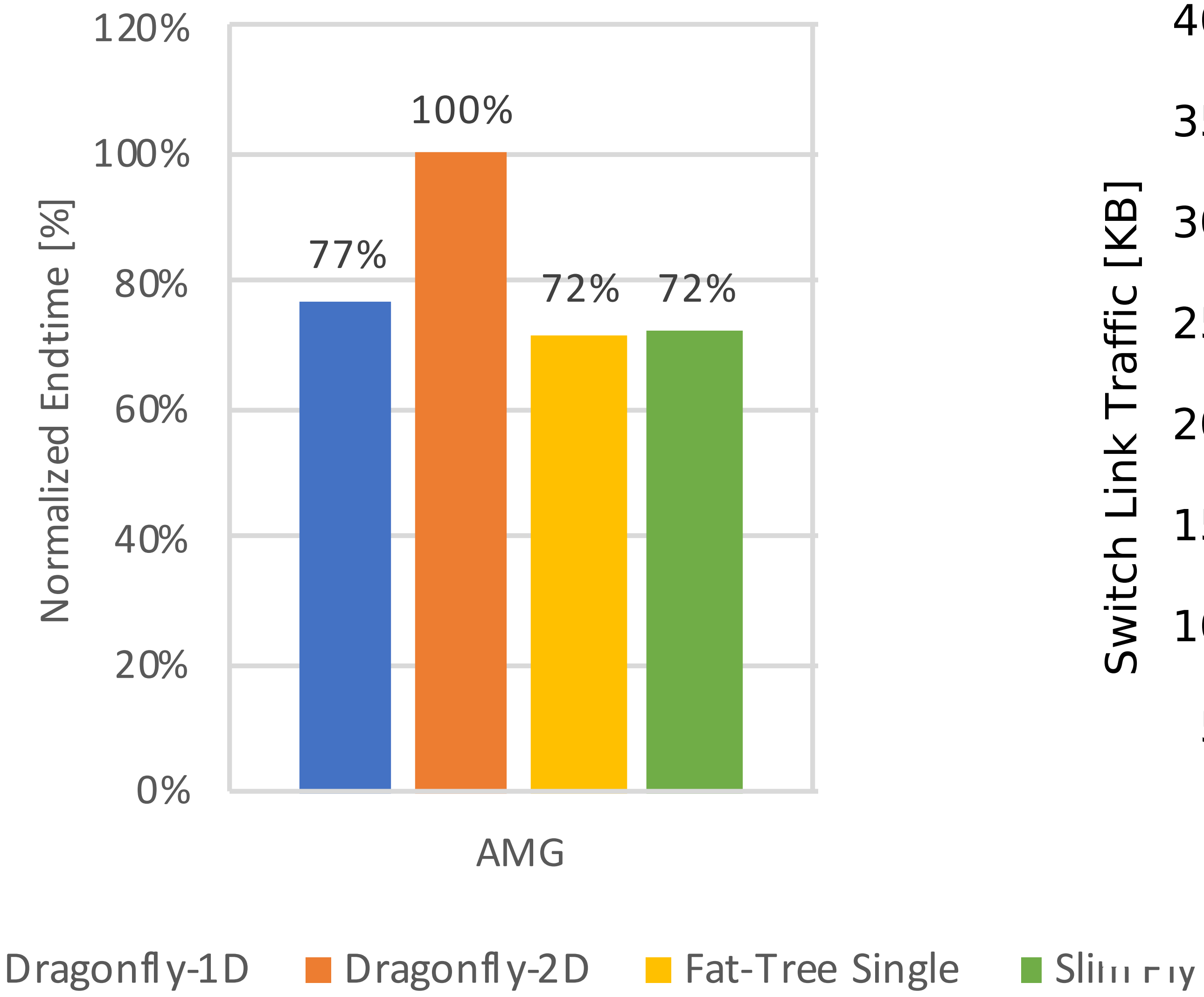


Dragonfly-1D 23% faster than Dragonfly-2D

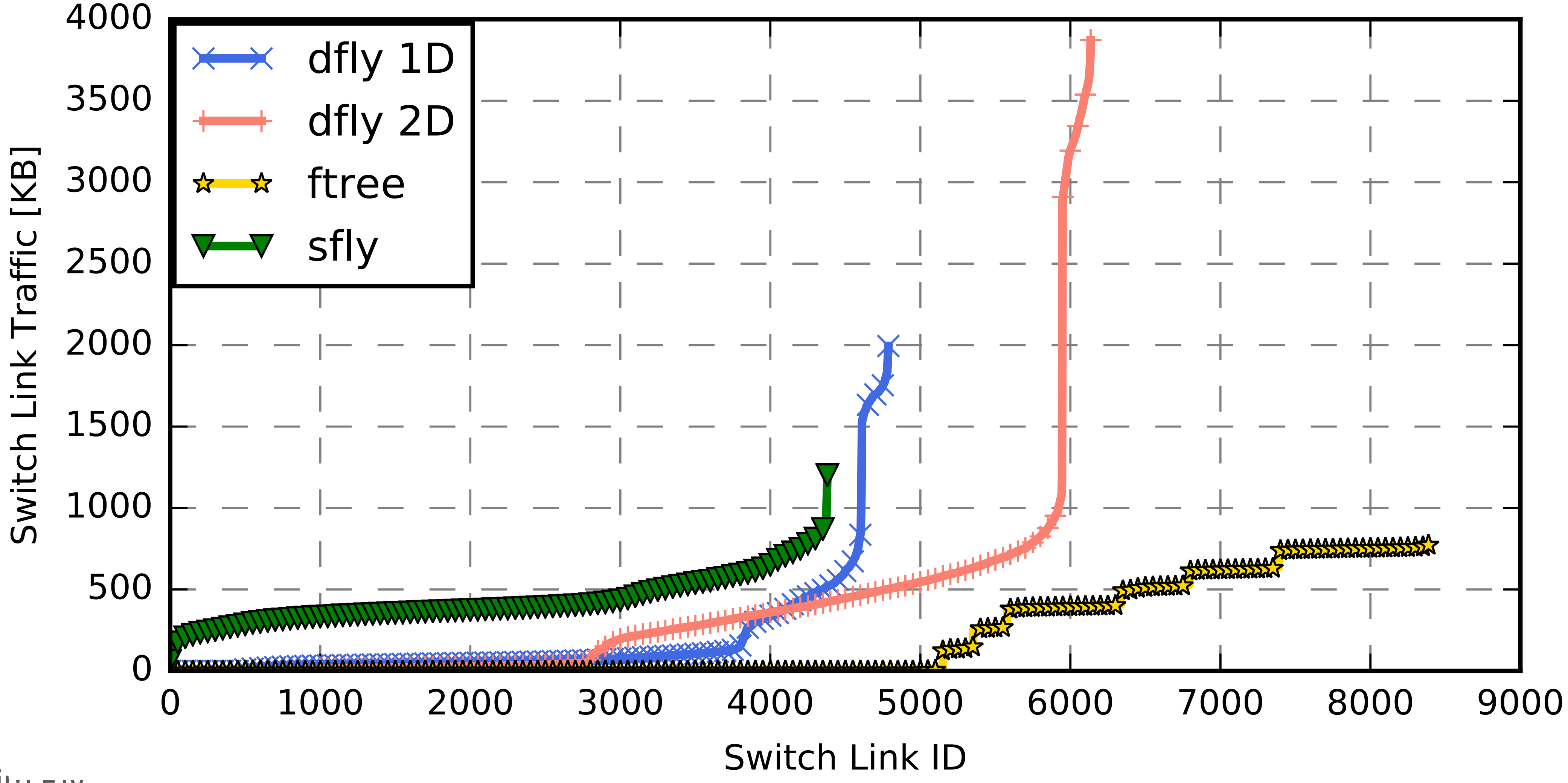
Slim Fly & Fat-Tree 28% faster than Dragonfly-2D

Homogeneous Results
(AMG)

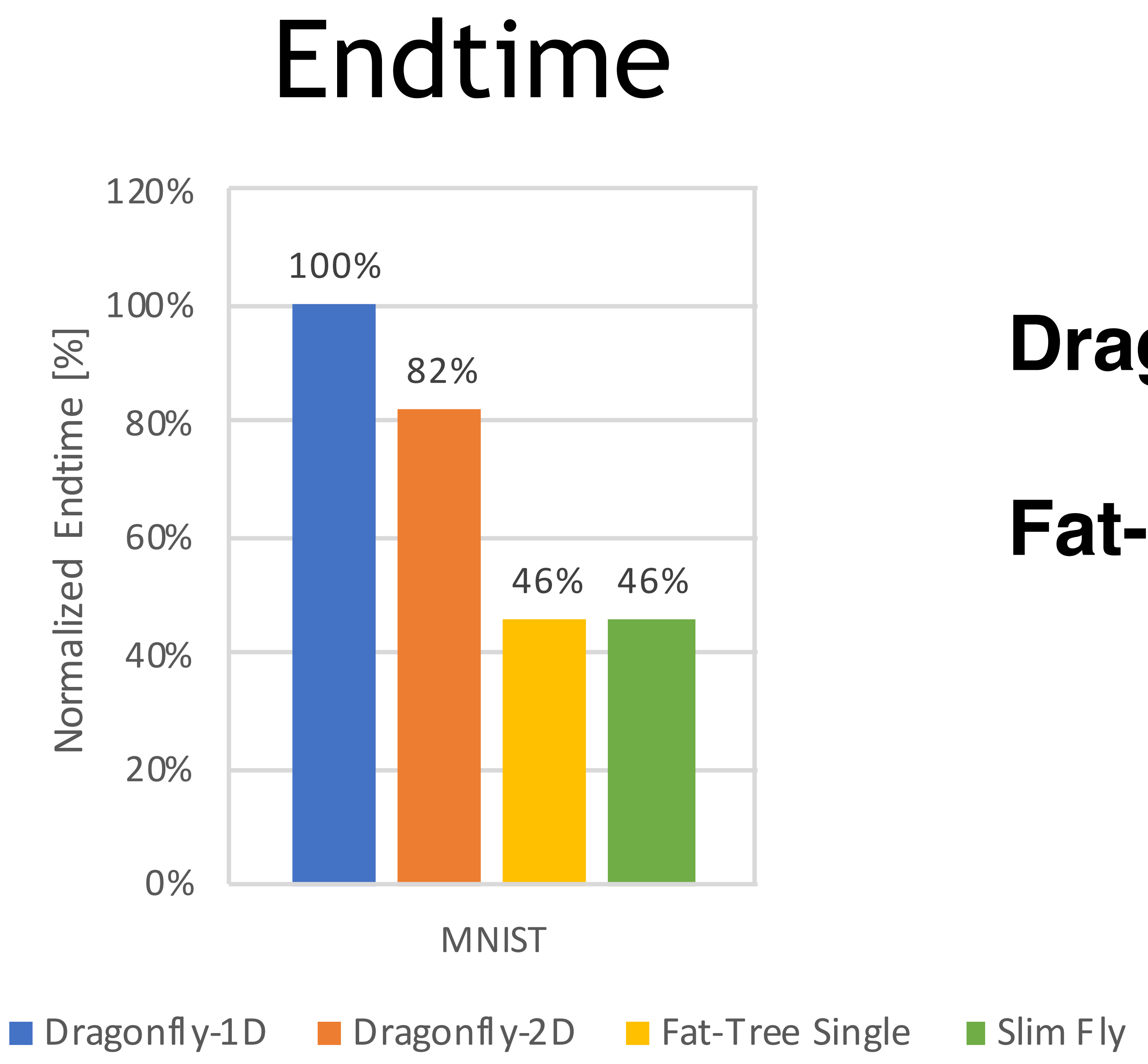
Endtime



Link Traffic



Homogeneous Results
(MNIST)

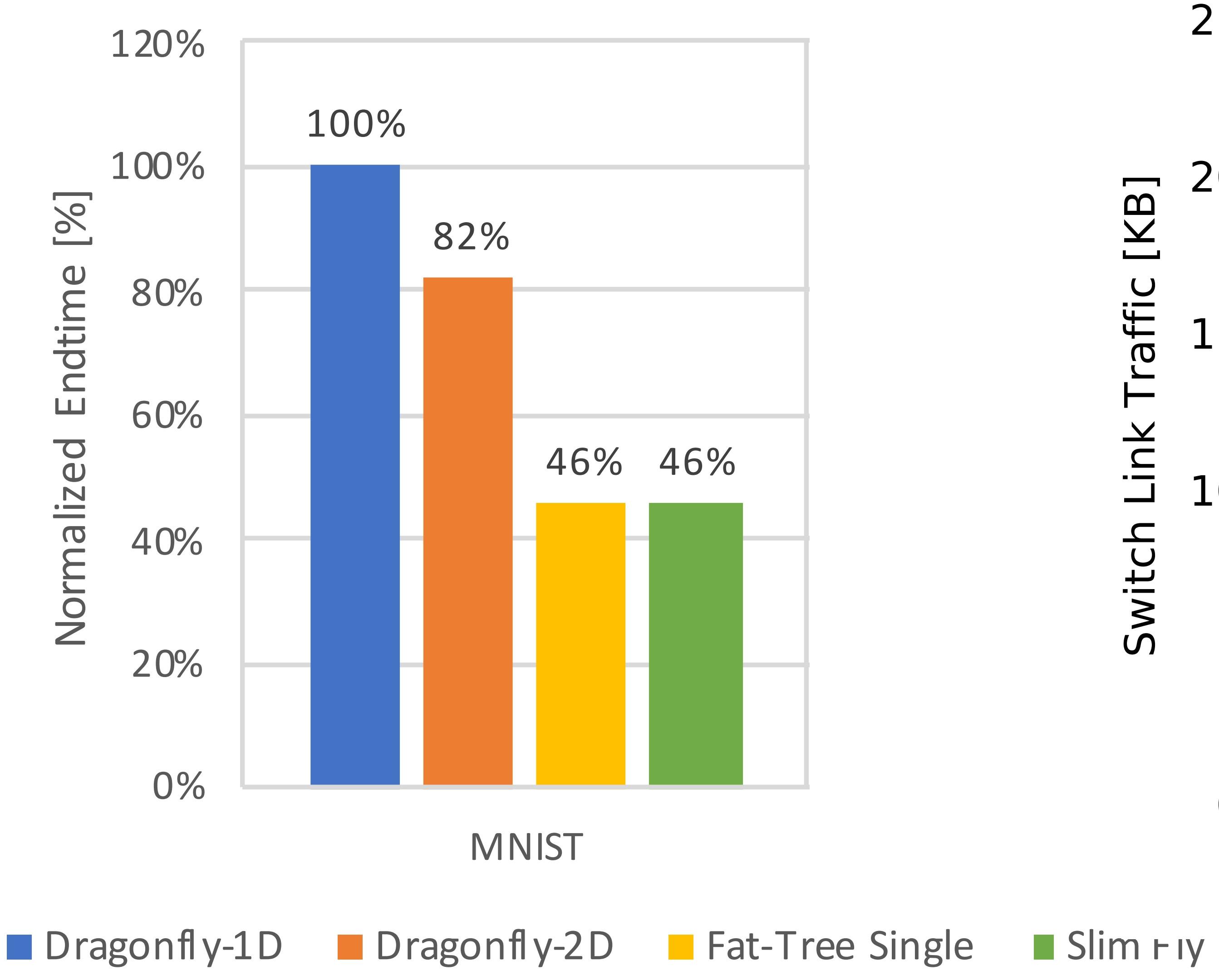


Dragonfly-2D 18% faster than Dragonfly-1D

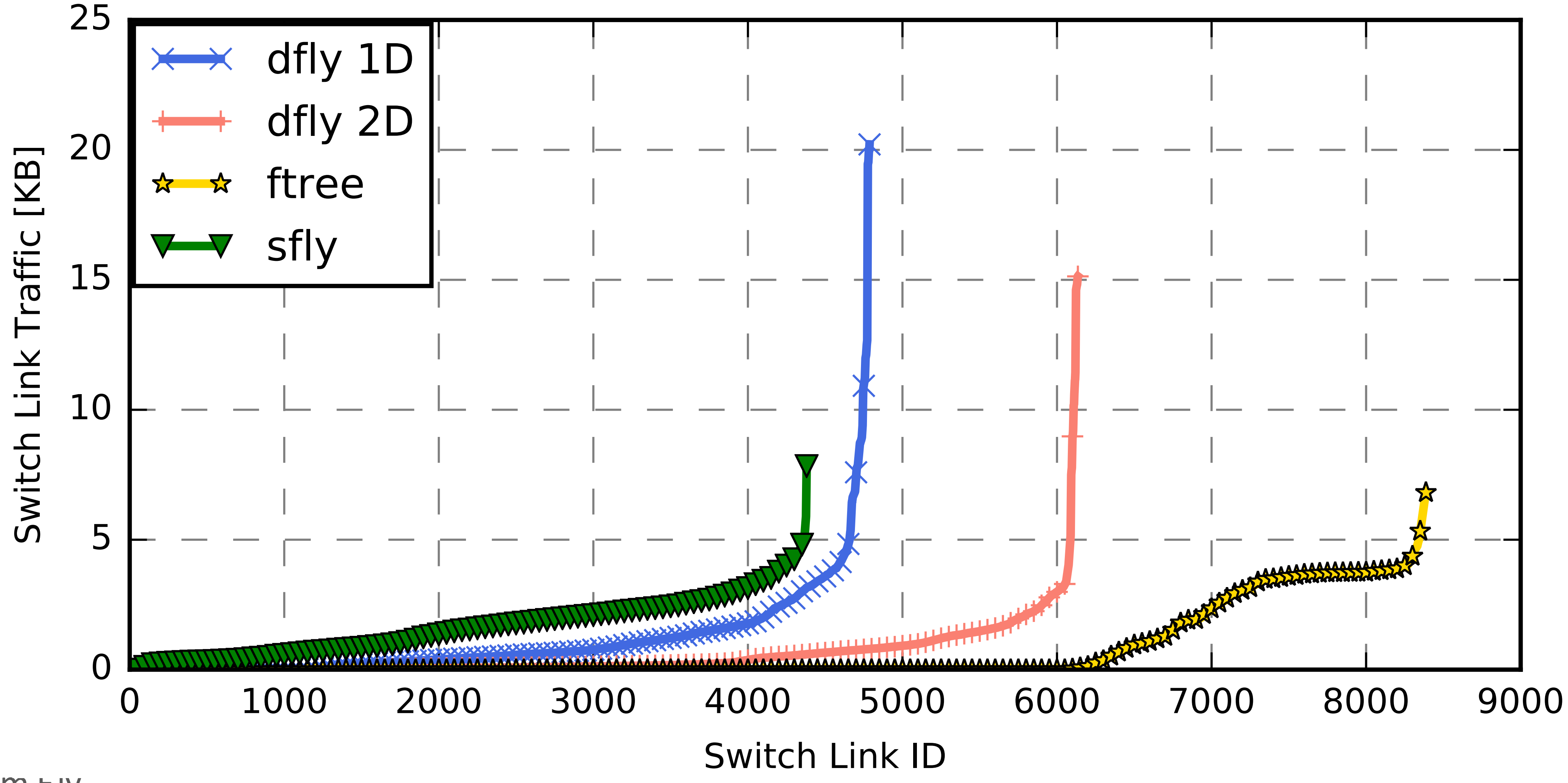
Fat-Tree & Slim Fly 54% faster than Dragonfly-1D

Homogeneous Results
(MNIST)

Endtime

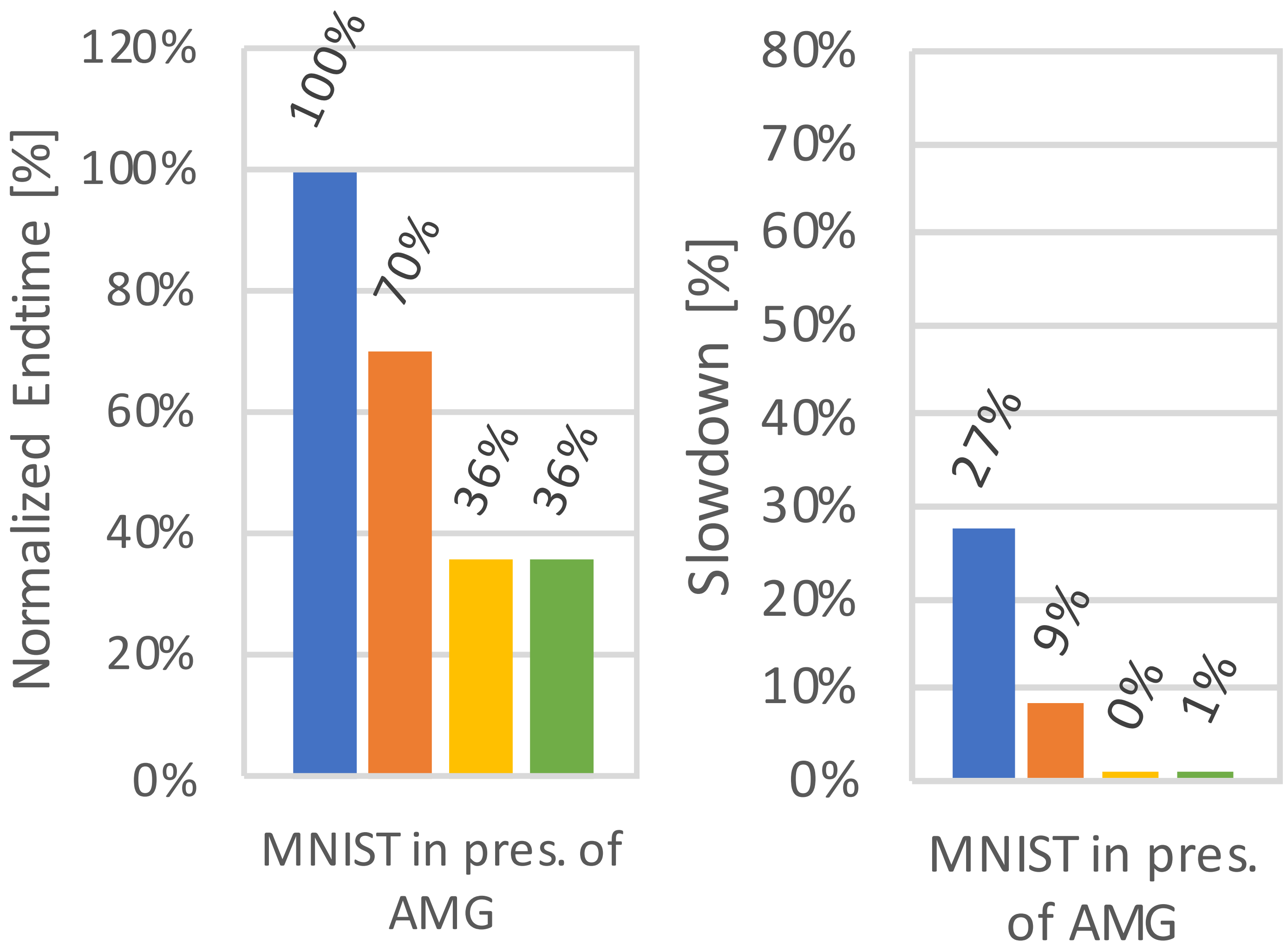


Link Traffic

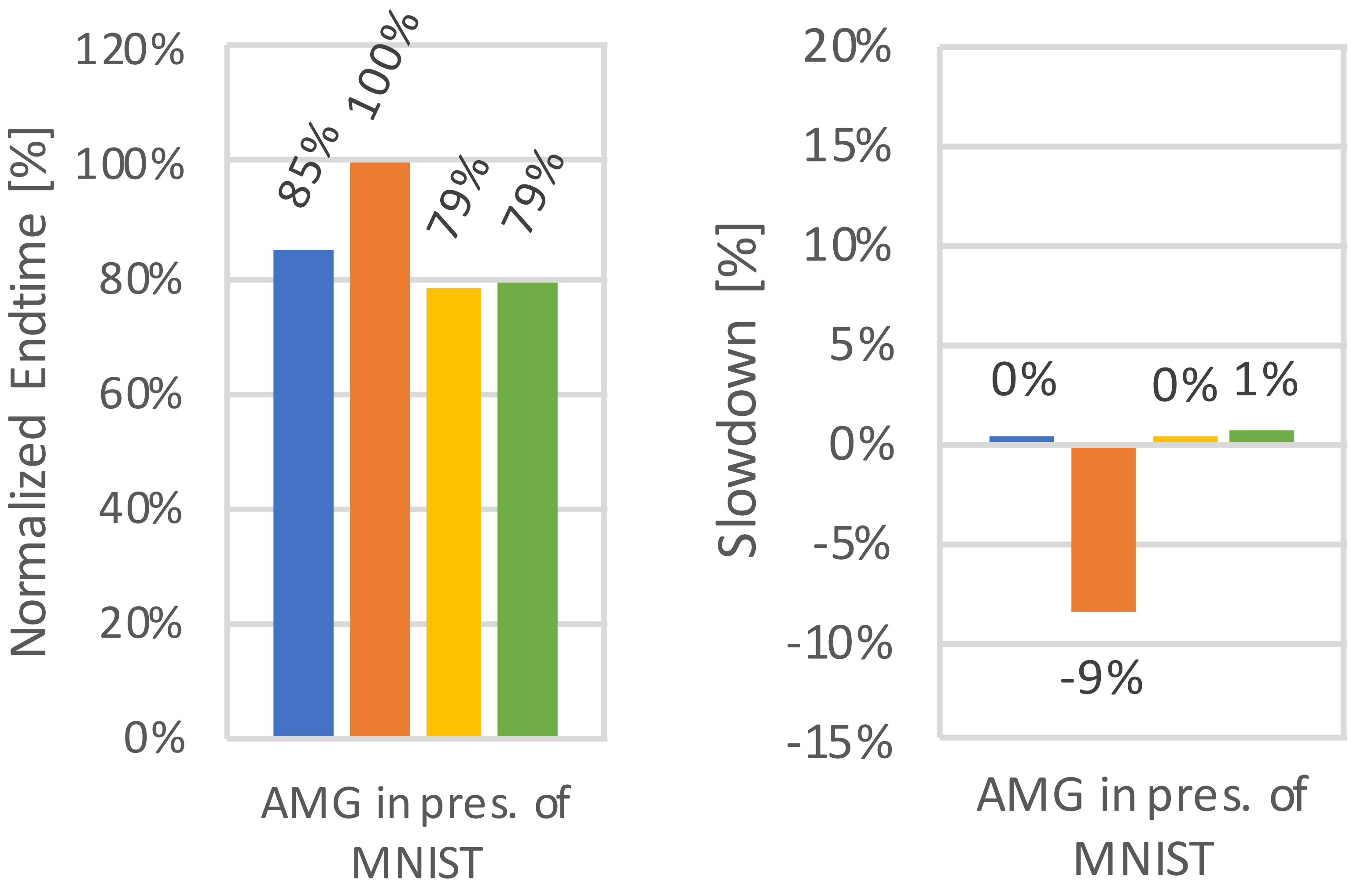


Hybrid Results
(Endtime & Slowdown)

MNIST



AMG



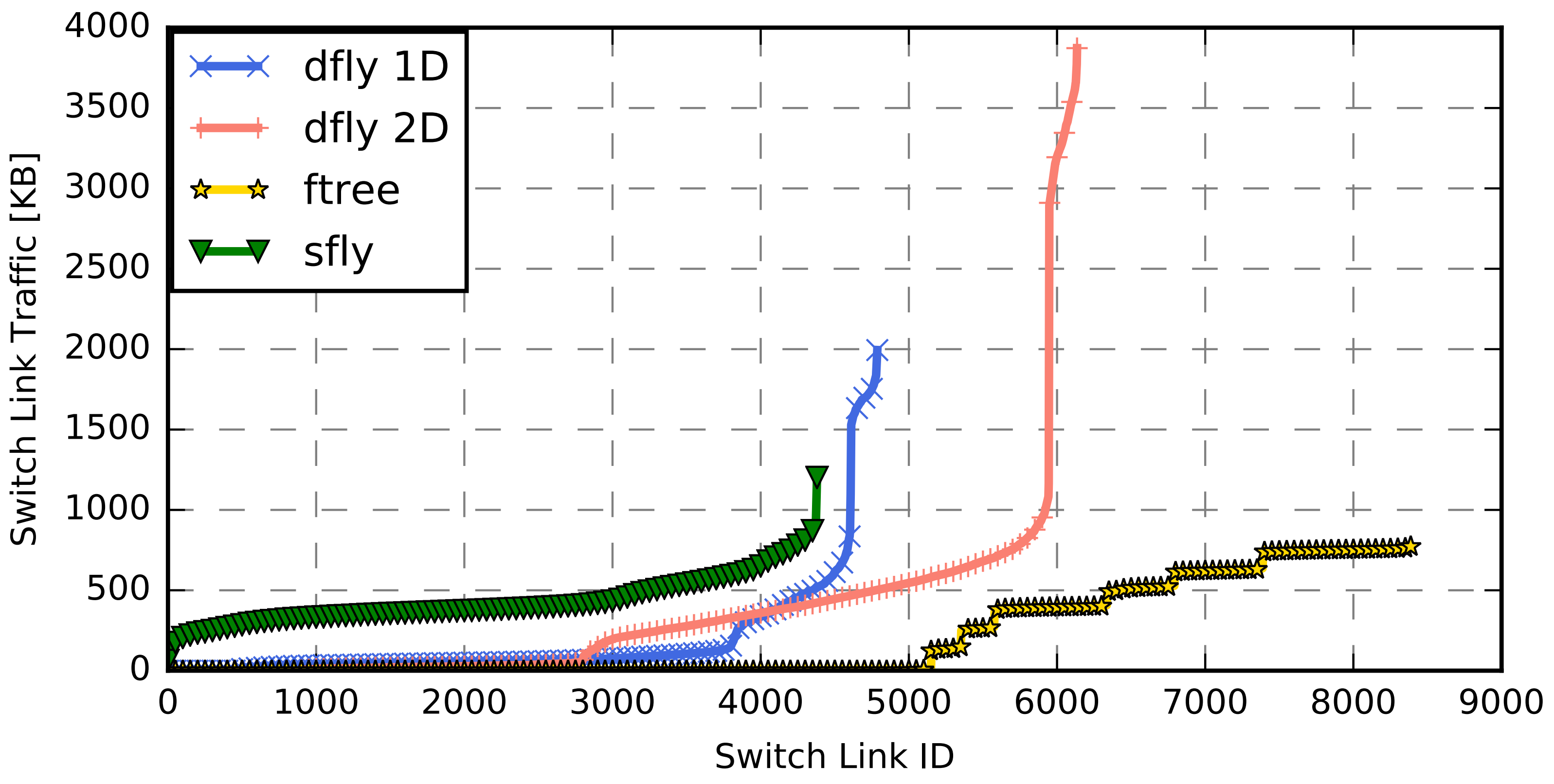
27% slowdown on Dragonfly-1D
9% slowdown on Dragonfly-2D

9% improvement on Dragonfly-2D

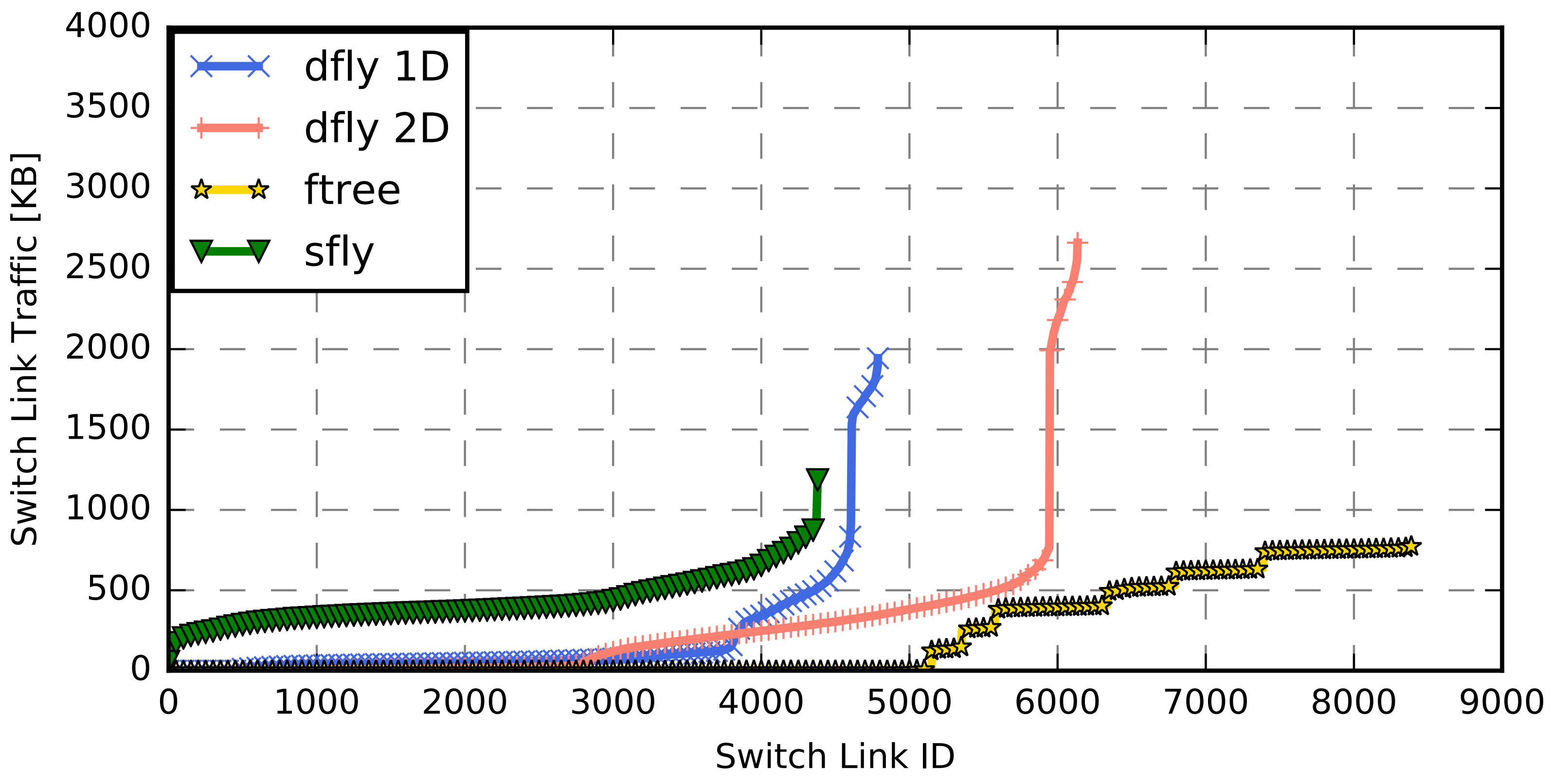
Dragonfly-1D Dragonfly-2D Fat-Tree Single Slim Fly

Hybrid Results
(Link Traffic)

AMG



Hybrid



30% decrease in maximum link traffic for Dragonfly-2D

Summary

- Neuromorphic architectures such as IBM TrueNorth pose **interesting design questions** for future **Extreme-Scale HPC** systems.
- Using IBM TrueNorth, NeMo, CODES we can study network performance of real deep learning **neuromorphic applications at scale** in an HPC environment.
- Preliminary analysis shows **Fat-Tree and Slim Fly** HPC network topologies are better able to **minimize interference** between neuromorphic and traditional HPC applications than Dragonfly.
- **Neuromorphic** workloads representing convolutional neural network and Hopfield network applications **pose little effect on** traditional **CPU applications** when running in parallel in a multi-job hybrid HPC environment.
- Traditional **CPU** network workloads can **significantly effect** performance of **Neuromorphic** application workloads.

Future Work

- Investigating additional **neuro workloads** and approaches for **workload scaling** and **chip mapping**.
- **Improving Dragonfly** configuration performance by investigating **minimal path bias**, **adaptive thresholds**, and **job allocations** to compute nodes.
- Studying **coexistence** of **multiple neuromorphic** applications.