Understanding the Impact of Fat-Tree Network Locality on Application Performance

RICE

Philip Taffet

Advisors: Ian Karlin, Edgar A. Leon, John Mellor-Crummey

Lawrence Livermore National Laboratory

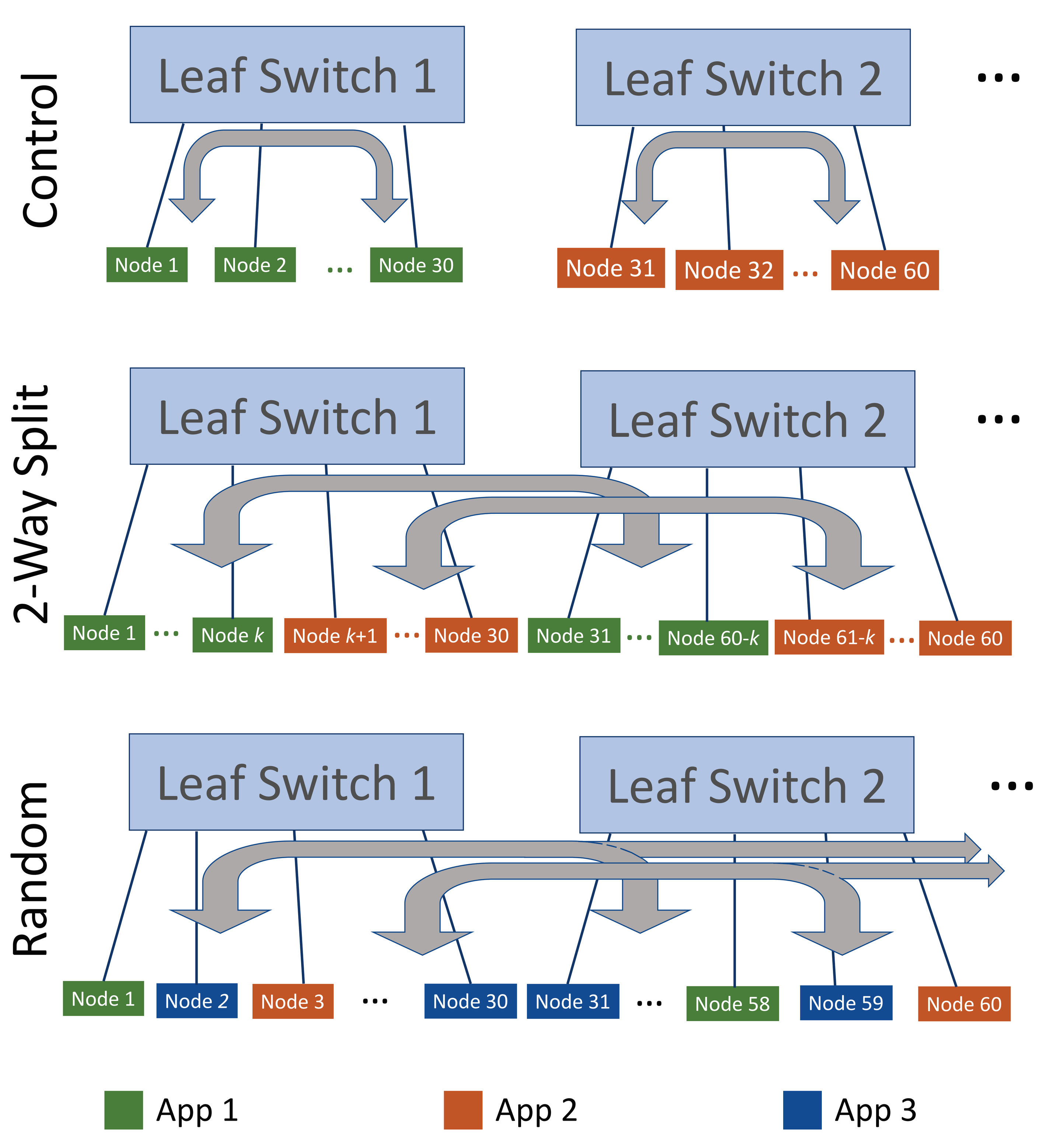
Abstract

Network congestion can be a significant cause of performance loss and variability for many message passing programs. However, few studies have used a controlled environment with virtually no other extraneous sources of network traffic to observe the impact of application placement and multi-job interactions on overall performance.

We study different placements and pairings for three DOE proxy applications. We observe that for jobs sizes typical for an LLNL commodity cluster, the impact of congestion is less significant than previously believed. In addition, in most cases, the cyclic MPI task mapping strategy increases performance despite also increasing total network traffic.

Experimental Setup

All experiments ran on 186 nodes of Quartz, an LLNL commodity cluster, during a dedicated access time. Quartz uses dual Intel Xeon E5-2695 (Broadwell) with 18 cores per socket. Quartz's interconnect is a 100Gbps Omni-Path fat-tree network with a 2:1 taper after the first level.



Additional experiments:

- **3-Way Split:** each application spans three leaf switches, and each leaf switch has some nodes running each application
- **With Bully**: similar to the 2-Way Split with k=15, but each application is paired with MPIBW or MPIPPS (see right)
- **In The Wild**: each application runs on a normally scheduled allocation

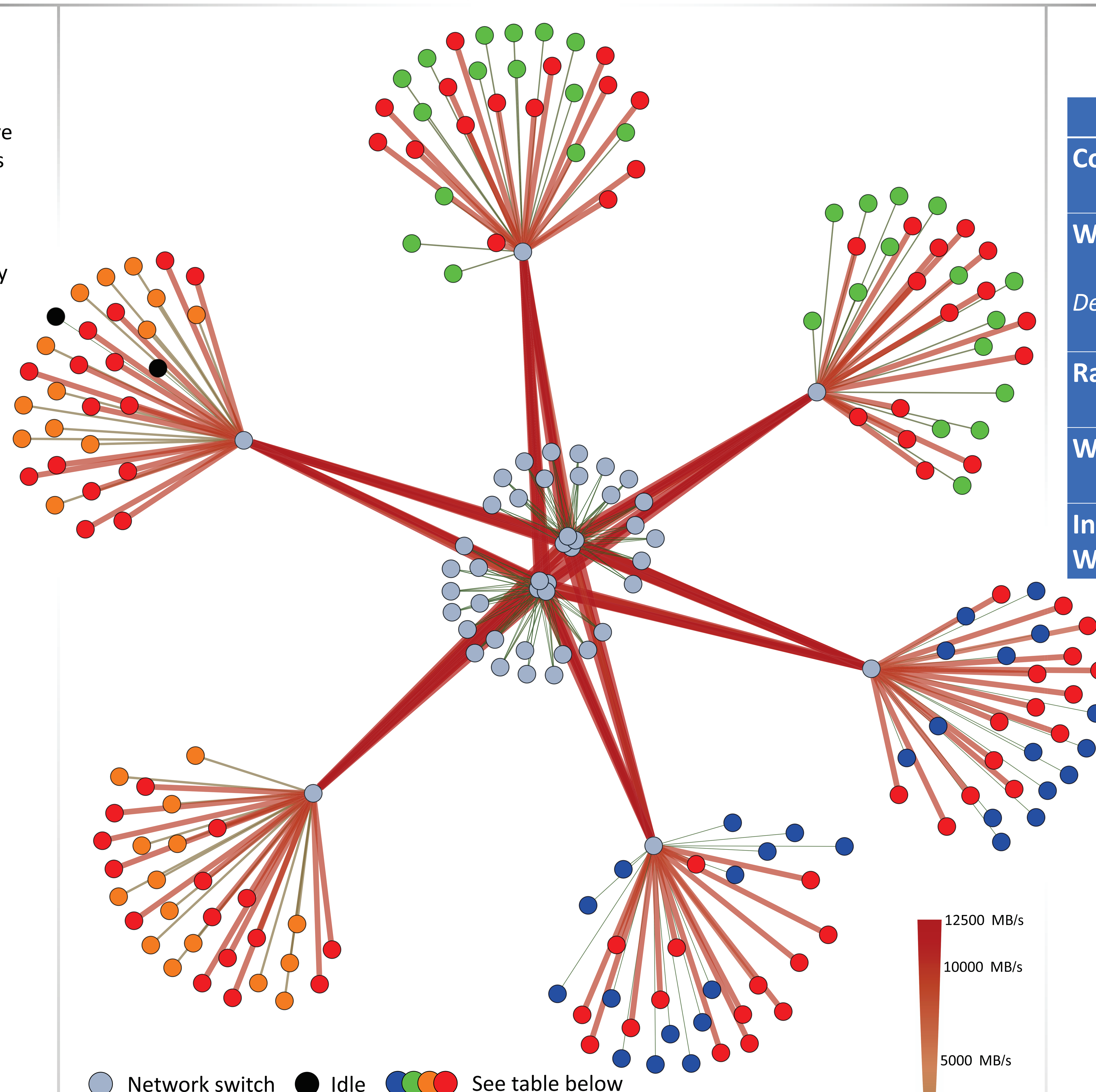


Figure 1: The state of the portion of Quartz used during one of the **With Bully** experiments. Link utilization is measured over 30 seconds. Note how dramatically link utilization varies between applications. This experiment did not prescribe running MPIPPS on any nodes.

Applications

App Name	Description	Point-to-point messages	Collective Operations
AMG	Algebraic multigrid solver for linear systems on unstructured meshes	Many, small to medium	Many, small
UMT	Radiation transport proxy app	Many, medium	Some, small
pF3D	Laser-plasma interaction simulator	Many, small to large	Many, small to medium
MPIBW	Bidirectional point-to-point MPI bandwidth benchmark	Many, large	None
MPIPPS	MPI message-rate benchmark	Many, small	None

Network Placement Sensitivity

	AMG	UMT	pF3D
Control	1.75×10^{9}	1.52×10^{10}	3.88×10^{1}
Worst	1.74×10^{9} (-0.6%)	1.51×10^{10} (-0.4%)	3.82×10^{1} (-1.5%)
Description	*Split with pF3D, k=5*	*Split over 3 switches*	*Split over 3 switches*
Random	1.74×10^{9} (-0.2%)	1.51×10^{10} (-0.4%)	3.81×10^{1} (-1.7%)
With Bully	1.74×10^{9} (-0.3%)	1.52×10^{10} (0.0%)	3.89×10^{1} $(+0.2\%)$
In The Wild	1.72×10^{9} (-1.6%)	1.51×10^{10} (-0.2%)	3.83×10^{1} (-1.2%)

App-specific figures-of-merit reported. Higher is better. Green values are significantly different from Control at $\alpha = .01$ and orange values are significantly different from Control at $\alpha = .05$.

Cyclic vs. Block Mapping

	AMG	UMT	pF3D
Control	+3.5%	+1.5%	+9.8%
Random	+4.9%	+1.2%	+2.5%
With Bully	+3.1%	+1.8%	+9.6%
In The Wild	+4.4%	+4.2%	+8.9%

Positive values indicate cyclic distribution was faster than block. Green values are significantly different from Control at $\alpha = .01$.

Conclusions

- Cyclic task mapping performs better than block except when the network is congested and task placement is poor
- AMG and UMT are minimally affected by congestion and poor placement
- pF3D is somewhat sensitive to network placement, congestion, and task mapping
- In The Wild: UMT is unaffected, pF3D is between perfect and poor placement, and AMG is sometimes significantly worse

Future Work

- We plan to evaluate larger jobs and a 4:1 tapered network topology
- We will compare measured results to network simulator predictions
- Network simulators, if accurate, will help explain the mechanisms that result in congestion and cause these performance impacts

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.