Characterizing and Improving Power and Performance in HPC Networks

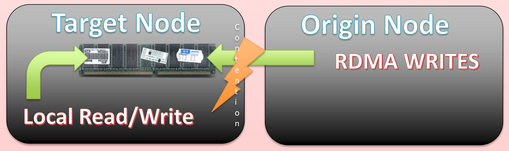
Taylor Groves advised by Dorian Arnold (UNM) mentored by Ryan Grant (SNL)

Networks are the backbone of modern HPC systems. They serve as a critical piece of infrastructure, tying together applications, analytics, storage and visualization. Despite this importance, we have not fully explored how evolving communication paradigms and network design will impact scientific workloads. As networks expand in the race towards Exascale, we must reexamine this relationship, so that the community better understands (1) characteristics and trends in HPC communication, (2) how to best design HPC networks and (3) opportunities in the future to save power or enhance the performance. **My thesis is that I can improve application performance and system power usage by gaining a detailed understanding of HPC communication on both the network endpoints and fabric; specifically, I address the problem of network-induced memory contention, quantify the power/performance tradeoffs for dragonfly topologies in HPC networks, and increase the scalability/responsiveness of large-scale network monitoring.** This dissertation highlights opportunities for improving network performance and power efficiency, while uncovering pitfalls (and mitigation strategies) brought about by shifting trends in HPC communication.

Network-induced Memory Contention (NiMC)

IPDPS 16

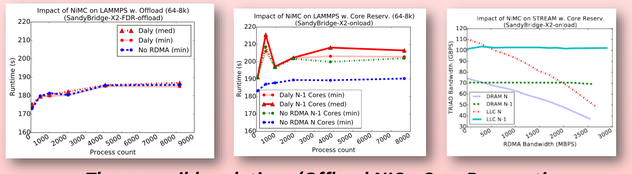
Taylor Groves, Ryan Grant, Dorian Arnold



Remote operations have potential to create contention

Machine	Triad no RDMA (GB/s)	Triad w. RDMA (GB/s)	Diff. (GB/s)	Diff. %
Westmere @ 800MHz, 1066MHz (offload)	12.9, 16.8	9.7, 12.8	-3.2, -4.0	-25%, -24%
Lisbon @ 800MHz, 1066MHz, 1333MHz (offload)	14, 17.9, 19.7	10.8, 14.3, 16.5	-3.2, -3.6, -3.2	-23%, -20%, -16%
Piledriver @ 1600MHz (onload)	12.4	7.4	-5	-40%
Piledriver @ 1866MHz (onload)	12.7	5.6	-7.1	-56%
SandyBridge-X2 (offload)	77.8	77.6	-0.2	0%
SandyBridge-X2 (onload)	73.4	36.1	-37.3	-51%
Xeon-Phi (on-chip, offload)	126.4	121.7	-4.7	-4%
Haswell-X2 (offload)	116.6	116.9	+0.3	0%

STREAM performance with(out) NiMC for varying Architectures



Three possible solutions (Offload NICs, Core Reservation, and RDMA Bandwidth Throttling)

Random Forests to detect presence of NiMC, predict impact and rank feature importance.

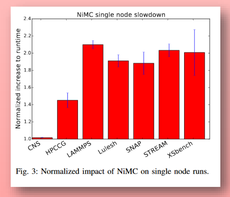
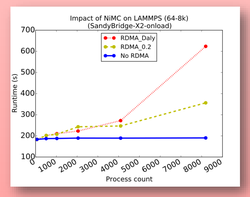
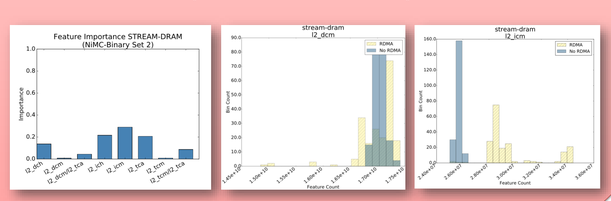


Fig. 3: Normalized impact of NiMC on single node runs.



NiMC impacts many workloads, increasing at scale



Feature importance and histograms of least/most important feature

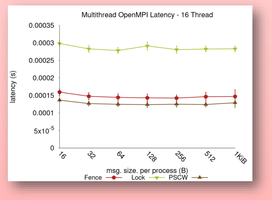
Multithreaded RMA Benchmarks for MPI

CCgrid 16

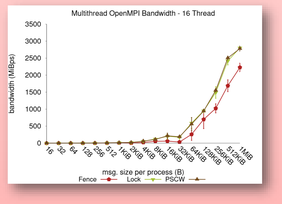
Matthew Dosanjh, Taylor Groves, Ryan Grant, Patrick Bridges, Ron Brightwell

- RMA and Multithreaded becoming increasingly prevalent
  - Core count increasing and more applications using one-sided
- Evaluate correctness and performance of RMA+Multithreaded

Introducing the RMA-MT benchmark suite

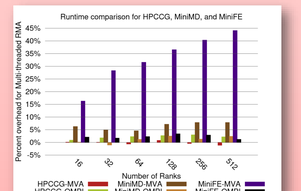


Latency



Bandwidth

Message Rate



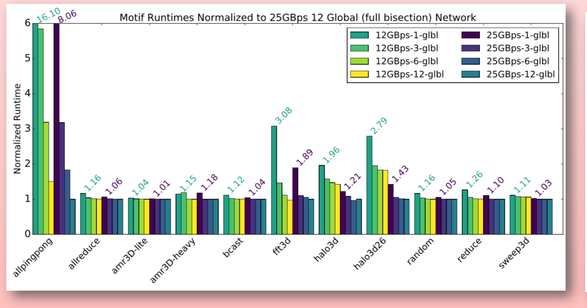
Converted HPCCG MiniMD & MiniFE

- **Supports Four Synchronization Methods**
- **RMA Puts and Gets**

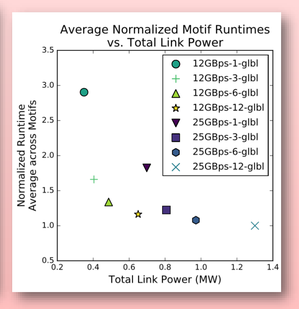
Stalled, Active and Idle: Power and Perf. of Large Dragonfly Networks

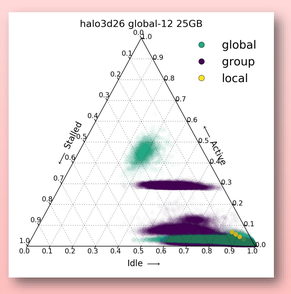
Cluster 16

Taylor Groves, Ryan Grant, Scott Hemmert, Simon Hammond, Michael Levenhagen, Dorian Arnold



Performance and power costs for different degrees of tapering number of global links and reducing link width





- **Detailed simulation of 100,000 node Dragonfly Networks**
- **Explored topology design decisions for 11 HPC workloads**
- **Introduced metric and visualization for characterizing port level network performance at largescale**

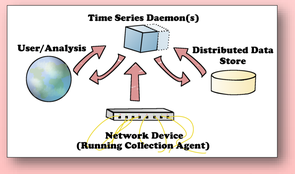
Monitoring Largescale Networks and Modeling Data Aggregation

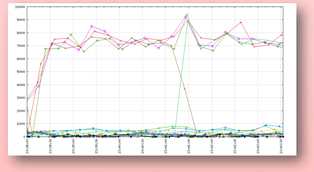
HPCMASPA 15
NDM 13

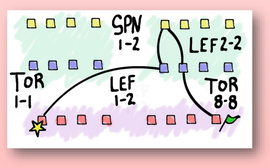
Taylor Groves, Samuel Gutierrez, Yihua He, Dorian Arnold

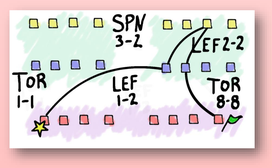
- Current Network Monitoring not scalable e.g. SNMP, or OpenSM

Implement a on-switch push-based monitoring agent



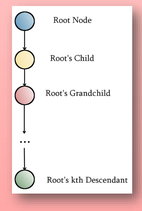


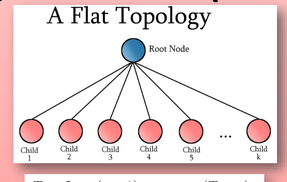




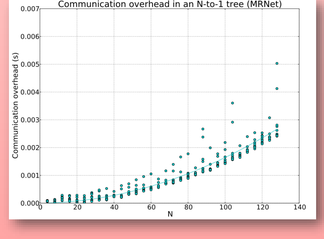
Improved monitoring responsiveness detects route changes otherwise missed by previous SNMP implementation

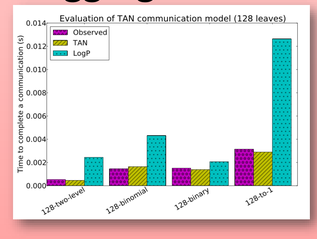
LogP extension (TAN): accurately model data aggregation





$$T_i = L + o(x+1) + g + \max(T_{child}),$$
$$(i \in N, \quad i \neq leaf)$$
$$T_{leaf} = C$$





Least Squares regression of tree fanout provides greater accuracy than simple additive function of original LogP model