

Full-System Modeling and Simulation: Contributions Towards Coupling, Contention, and I/O

WARWICK
THE UNIVERSITY OF WARWICK



D. G. Chester (Warwick), S. A. Wright (York), S. D. Hammond (Sandia), T. Law (AWE), R. Smedley-Stevenson (AWE), S. Maheswaran (AWE), S. A. Jarvis (Warwick)

Problem Statement

- Predict time to solution for operational systems
 - Performance predictions are inaccurate when machine utilization is high
- Support co-design of network interconnects, topology and applications at scale
 - Optimize performance-per-price-point

Tools and Techniques

- Congestion patterns from GPCNeT (NERSC) [1]
- Caliper (LLNL) to capture application behavior
- Memory subsystem and communication benchmarks
 - Stream
 - LMbench
 - Intel MPI Benchmarks
- SST Core and SST Elements 9.0 for simulating the network stack [2]
- Rebuild combine communication patterns in Ember Motifs

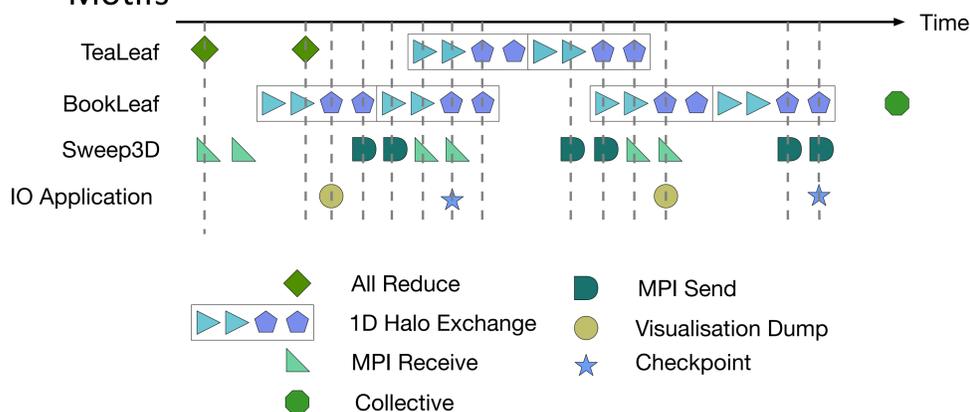


Fig 1: Contention induced by multiple applications communications patterns.

System Modeling

- Astra is a 1.5 PFLOP/s ARM Machine at SNL featuring Mellanox EDR in a tapered (2:1) fat tree
- From a validated model congestion patterns and PingPong can be combined
- All-to-all Congestion pattern; using all other nodes on the system

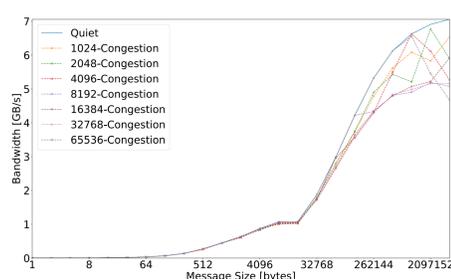


Fig 2: Simulated Bandwidth with varying congestion size

Results

- Congestion has detrimental effects on collective operations.

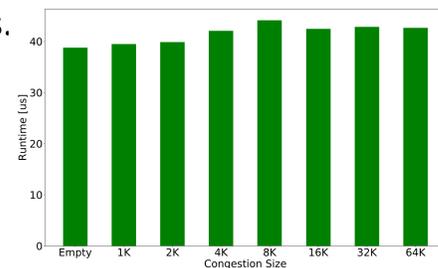


Fig 3: Simulated All Reduce (128 Nodes) times with varying congestion message size.

- Ternary plots provide a information rich graphic showing where switch ports spent their time [3]
- Below we have 4 motifs contending for the network with different congestion message sizes
 - Sweep3D
 - 2D Halo Exchange
 - All Reduce
 - Congestion Pattern – All-to-all

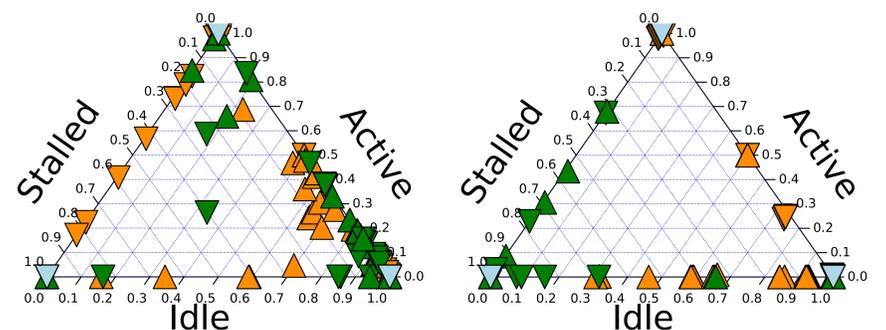


Fig 4: Ternary plots showing effect on network ports for 1K (left) and 64K (right) congestion message sizes; level 0 switch ports orange, level 1 switch ports green and level 2 switch ports blue

Tab 1: Average Time switch ports spent in the different states.

Switch	Active		Idle		Stalled	
	1K	64K	1K	64K	1K	64K
Level 2	86.11%	86.11%	11.11%	11.11%	2.78%	2.78%
Level 1	3.13%	0.56%	93.74%	95.24%	3.36%	4.41%
Level 0	37.68%	29.8%	57.67%	57.37%	5.73%	12.83%

Future Work and Extensions

We look to:

- Validate congestion modeling against representative systems
- Simulate congestion management techniques
- Model and simulate more systems and software

Acknowledgements

This work was supported by the UK Atomic Weapons Establishment under grant CDK0724 (AWE Technical Outreach Programme). Professor Stephen Jarvis is an AWE William Penney Fellow. Benchmarking time during the early access phases of Astra was made possible by the Vanguard system prototype program which is sponsored by the NNSA/ASC. We are grateful to all of the administrator staff and the technical support team from HPE for enabling productive access to the Astra system.



[1] T. Groves, S. Chunduri, and P. Mendygral, "Global Performance and Congestion Network Test - GPCNeT", <https://xgmlab.cels.anl.gov/networkbench/GPCNET> (accessed July 20, 2019), 2019.

[2] A. F. Rodrigues, K. S. Hemmert, B. W. Barrett, C. Kersey, R. Oldfield, M. Weston, R. Risen, J. Cook, P. Rosenfeld, E. Cooper-Balls et al., "The Structural Simulation Toolkit", SIGMET- RICS Performance Evaluation Review, vol. 38, no. 4, pp. 37–42, 2011.

[3] T. Groves, R. E. Grant, S. Hemmer, S. Hammond, M. Levenhagen, and D. C. Arnold, "(SAI) Stalled, Active and Idle: Characterizing Power and Performance of Large-Scale Dragonfly Networks", in 2016 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2016, pp. 50–59.

© British Crown Owned Copyright 2019/AWE. Published with permission of the Controller of Her Britannic Majesty's Stationery Office. This document is of United Kingdom origin and contains proprietary information which is the property of the Secretary of State for Defence. It is furnished in confidence and may not be copied, used or disclosed in whole or in part without prior written consent of Defence Intellectual Property Rights DGDCDIPR-PL—Ministry of Defence, Abbey Wood, Bristol, BS34 8JH, England.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525