Conference or Workshop Item:

Chester, Dean, Wright, Steven A. orcid.org/0000-0001-7133-8533, Hammond, Simon D. et al. (4 more authors) (2019) Full-System Modeling and Simulation: Contributions Towards Coupling Contention and I/O. In: Workshop on Modeling & Simulation of Systems and Applications, 14-16 Aug 2019.

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

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

Problem Statement

Production machine performance has large variability. On the UK National Supercomputing Service, the time a job takes to complete can vary by as much as 53%. Load imbalance and shared resource contention are largely responsible, but we find that previous efforts to model application/architecture performance do not typically take these into account.

In this research we model and simulate network contention, which allows us to explore the impact of multiple interacting jobs and approaches to alleviate these effects, including network re-design and communication-staging within applications. We show the utility of this work on a variety of systems and interacting applications.

Tools and Techniques

We make use of the Structural Simulation Toolkit (SST V9.0.0) [1]. Models are developed by benchmarking key system components of our target architectures, including the memory subsystem and the network interconnect, using Stream, LMBench and the Intel MPI Benchmarks.

Applications are profiled with Caliper and we combine computation and communication patterns in SST. All application/architecture simulations are validated on quiet (unloaded) systems to ensure that they are accurate.

In order to model and simulate the impact of multiple interacting jobs, we build on techniques first developed in GPCNeT [2]. An example of contending communication patterns from four simultaneously executing applications can be found in Figure 1.

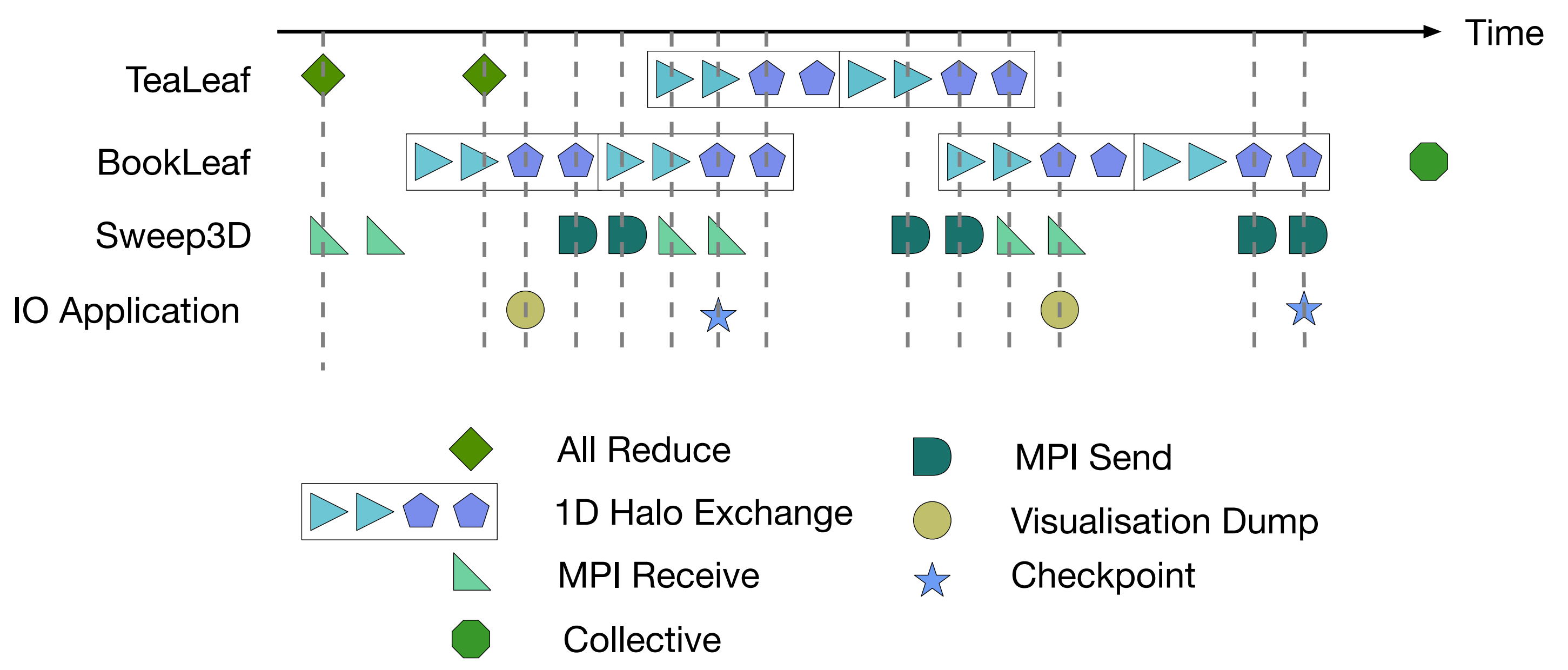

Figure 1: Contention induced by multiple applications communications patterns.

System Modeling

Much of our work has been conducted on Astra, a 1.5 PFLOP/s supercomputer at Sandia National Laboratories. Astra is comprised of 5,184 Cavium ThunderX2 central processing units, each with 28 processing cores based on the Arm V8 64-bit core architecture, with a Mellanox EDR tapered (2:1) fat tree interconnect. Figure 2 demonstrates the simulated bandwidth as congestion effects are explored through the simulator.

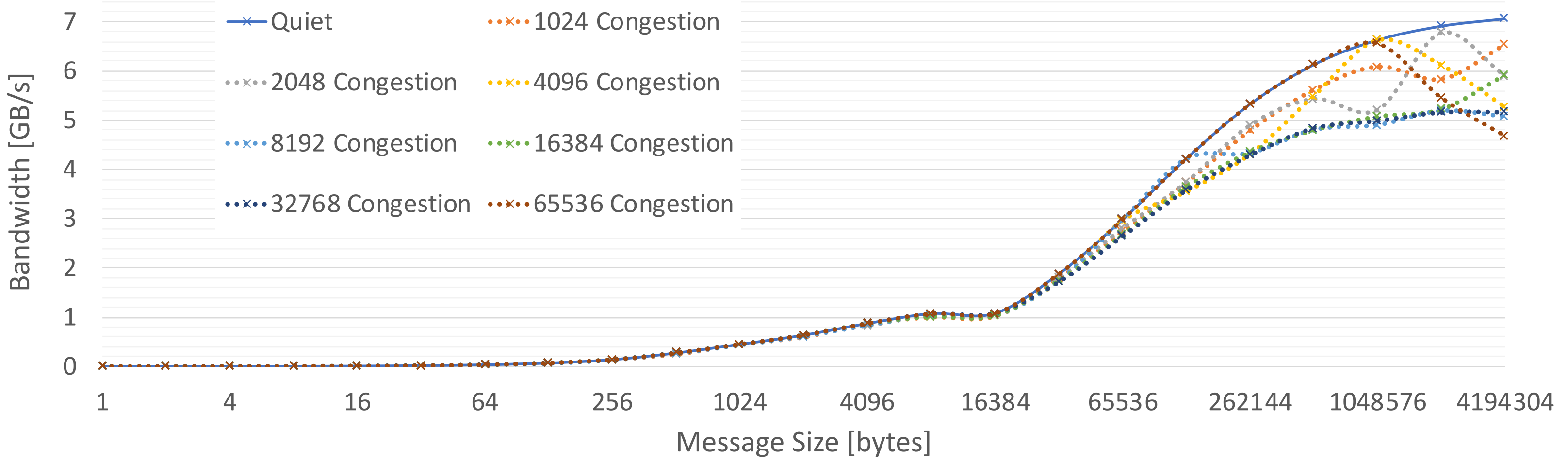

Figure 2: Simulated Bandwidth with varying congestion size

Results

Figure 3 shows ternary plots which provide information on network switch ports [3]. In these experiments four communication motifs (for Sweep3D, a 2-D Halo Exchange, an All Reduce benchmark and a congestion pattern generator) contend for the network with differing congestion message sizes. The average time for the All Reduce increases by 1.3 µs for the 64 K congestion messages compared to a quiet system.

Table 1 shows the impact on the Active, Idle and Stalled states as the contention messages are increased from 1 K to 64 K in size. All results are generated by the SST simulator.

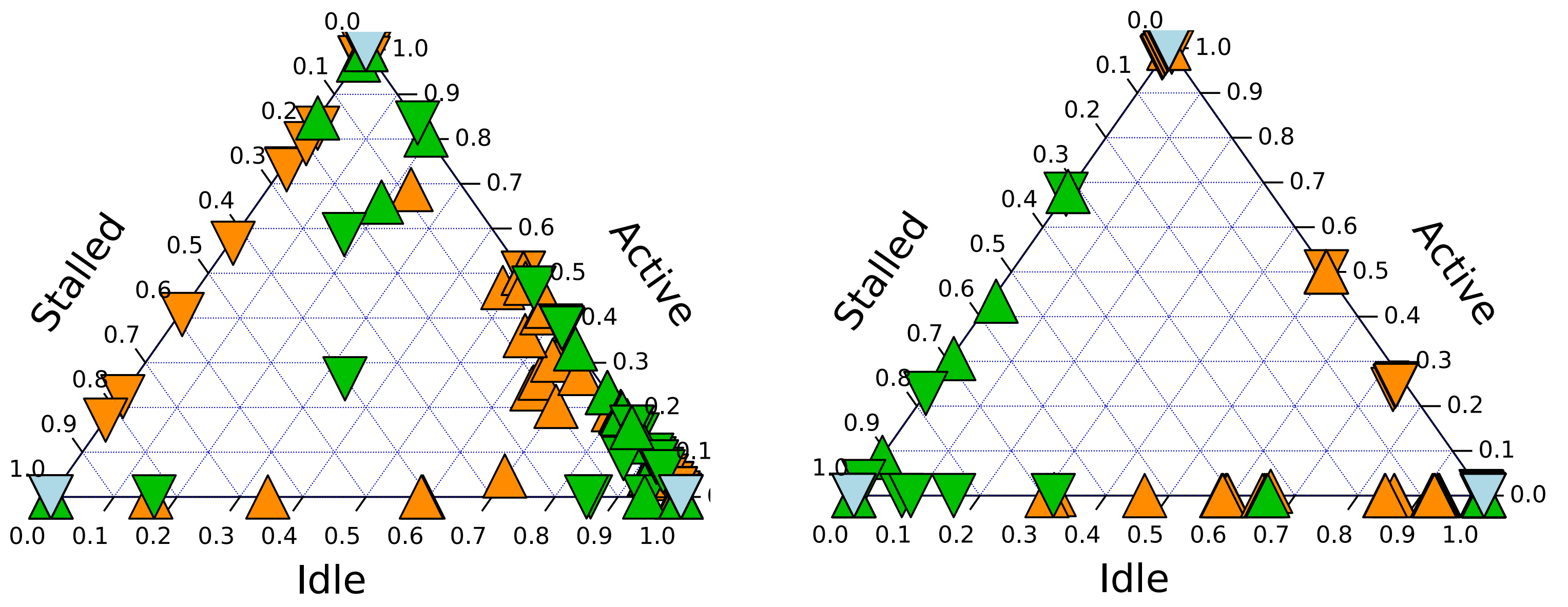

Figure 3: Ternary plots showing effect on network ports for 1 K (left) and 64 K (right) congestion message sizes; level 0 switch ports orange, level 1 switch ports green and level 2 switch ports blue

	Active		Idle		Stalled	
Switch	1 K	64 K	1 K	64 K	1 K	64 K
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%

Table 1: Percentage Time switch ports spent in the different states.

Future Work and Extensions

We are seeking to validate the congestion modeling against representative systems and workloads. In addition we are exploring a variety of congestion management techniques (at the application level) and system changes (at the architectural level) to alleviate these issues for future production systems.

References

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

[2] S. Chunduri, T. Groves, P. Mendygral, B. Austin, J. Balma, K. Kandalla, K. Kumaran, G. Lockwood, S. Parker, S. Warren, N. Wichmann, N. J. Wright, "GPCNeT: Designing a Benchmark Suite for Inducing and Measuring Contention in HPC Networks", International Conference on High Performance Computing, Networking, Storage and Analysis (SC'19), November 16, 2019

[3] T. Groves, R. E. Grant, K. S. Hemmert, S. D. 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.

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.

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