A Study of Network Congestion in Two Supercomputing High-Speed Interconnects

Saurabh Jha*, Archit Patke*, Jim Brandt^, Ann Gentile^, Mike Showerman^, Eric Roman^^, Zbigniew Kalbarczyk*, Bill Kramer*, Ravi Iyer*

** UIUC/NCSA, ^ Sandia National Lab, ^^ NERSC*

Email: sjha8@Illinois.edu

Presented on August 16, 2019 at HOTI 26

DOE HPC Facilities

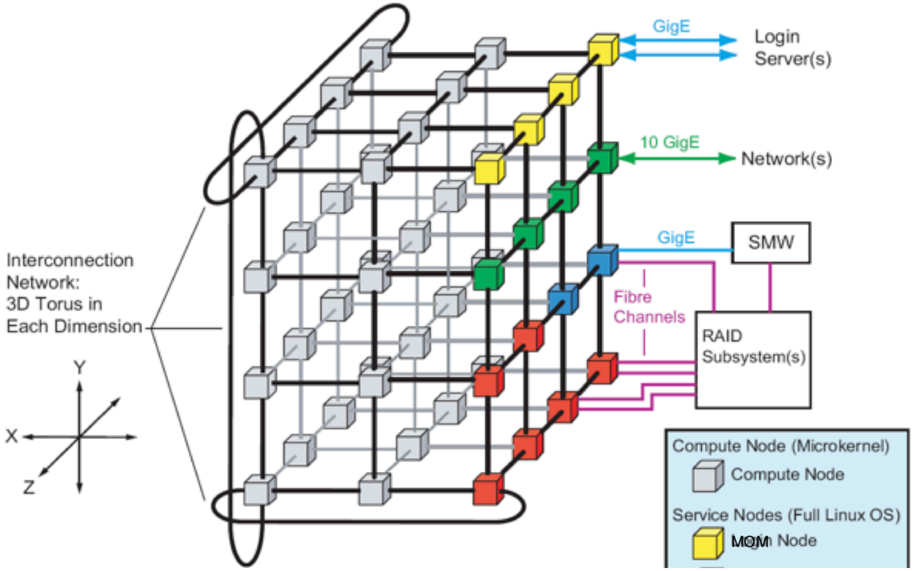
Source: hpcwire

HPC interconnect technologies:

- Cray Gemini (3D Torus), Aries (DragonFly), Slingshot (DragonFly)
- Mellanox InfiniBand (Non-blocking Fat Tree)
- IBM BG/Q proprietary technology (5D Torus)

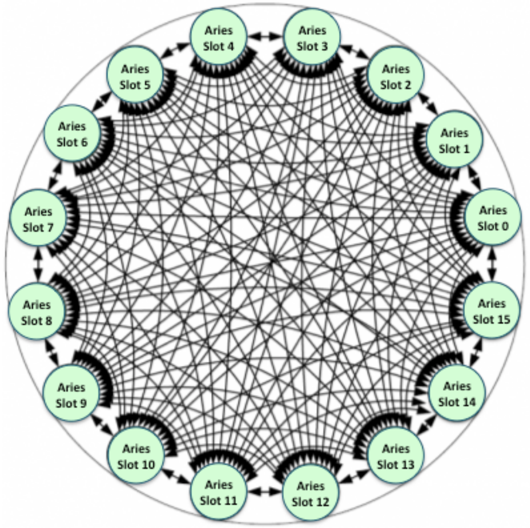
Studied HPC Systems

NCSA Blue Waters



Interconnect: Cray Gemini (3-D Torus)

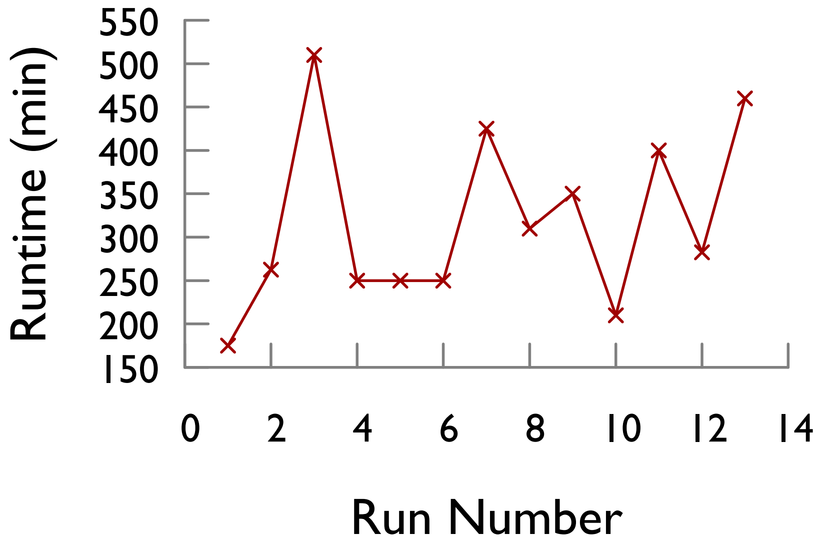
NERSC Edison



Interconnect: Cray Aries (DragonFly)

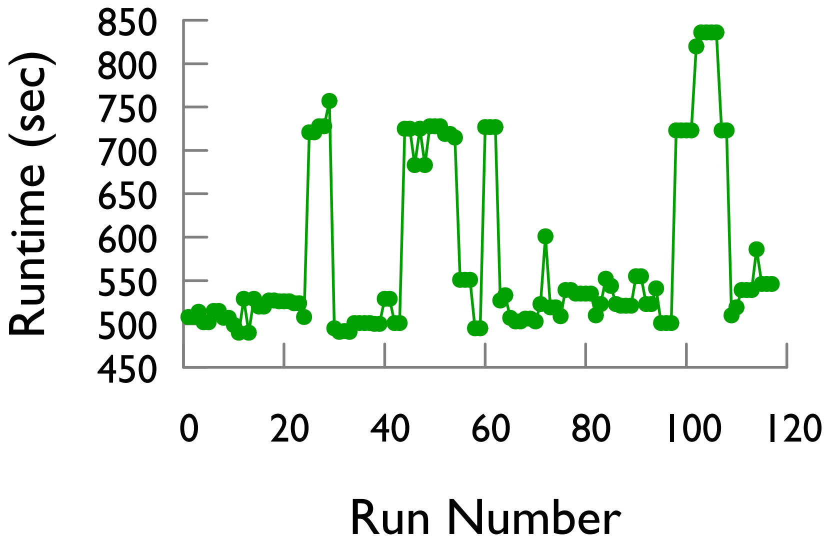
Mystery of Application Performance Variation

Mean Runtime : 318 mins
Standard Deviation in Runtime: 103 mins



NAMD Completion Time (*NCSA Blue Waters*)

Mean Runtime : 576 secs
Standard Deviation in Runtime: 100 secs

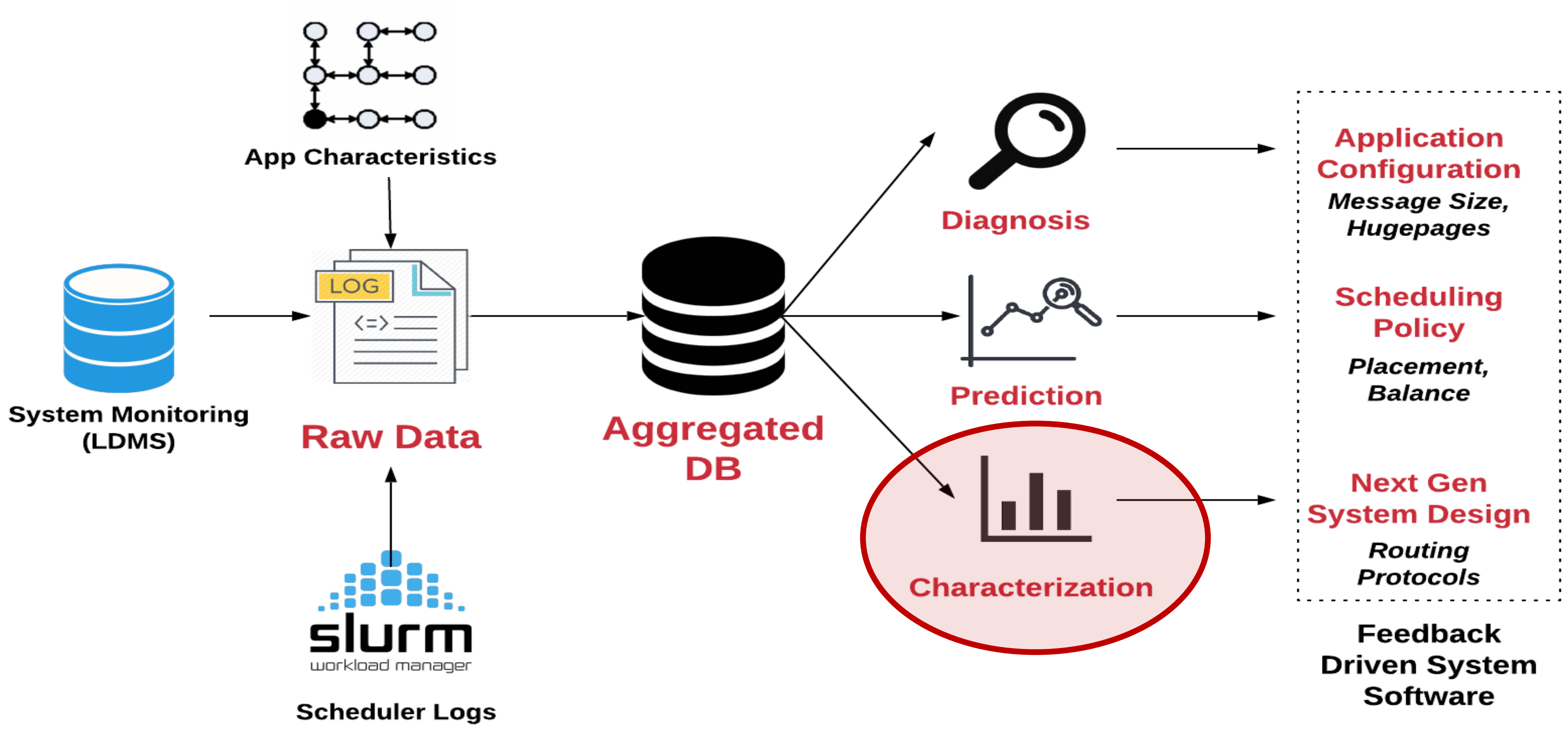


MILC Completion Time (*NERSC Edison*)

Variation caused by

- Interference from other applications
- Non-optimal configuration settings (too many knobs!)
  e.g., huge pages, placement, message size, node sharing

Monet: End-to-End Interconnect Monitoring Workflow



Data Measurement and Metrics

- Congestion measured in terms of Percent Time Stalled (P_{TS})

$$P_{TS} = 100 * \frac{T_{is}}{T_i}$$

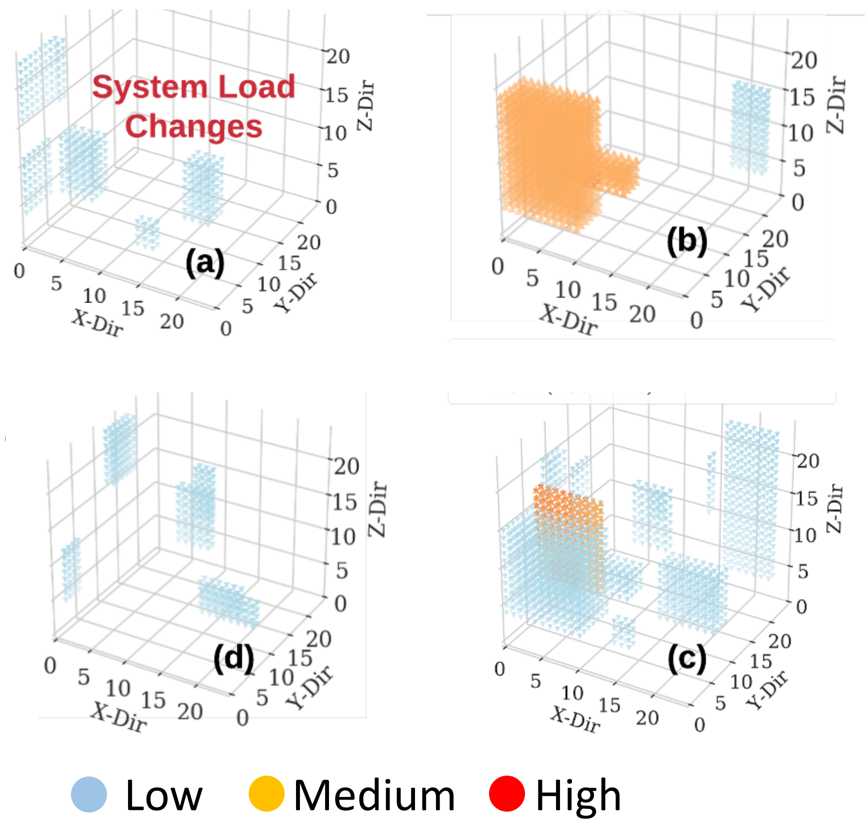
T_{is} is the time spent in stalled state

T_i is the measurement interval

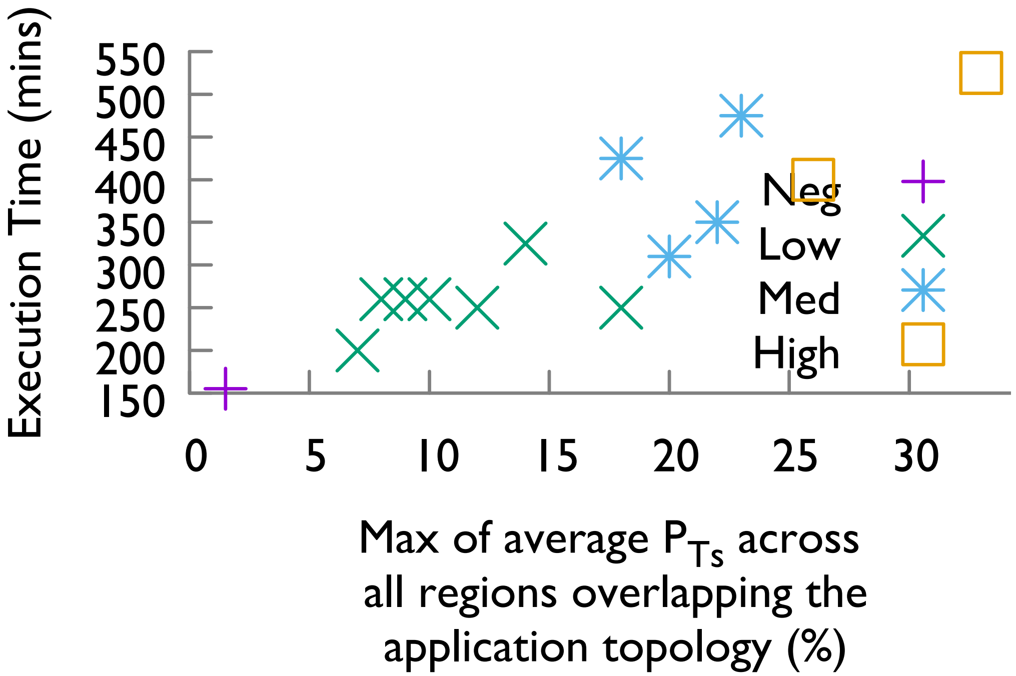
- On Blue Waters: T_i is 60s, data gathered: ~700 MB/minute
- On Edison: T_i is 1s, data gathered: ~7.5 GB/minute

Characterizing Congestion On Toroidal Networks

Neg: 0-5%, Low: 5-15%, Medium: 15-25%, High > 25%



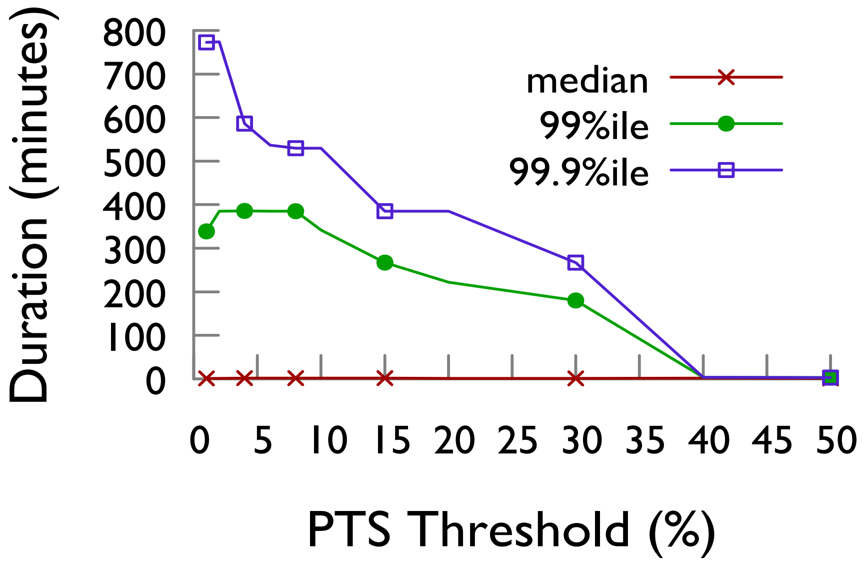
Congestion Cloud (NCSA Blue Waters)



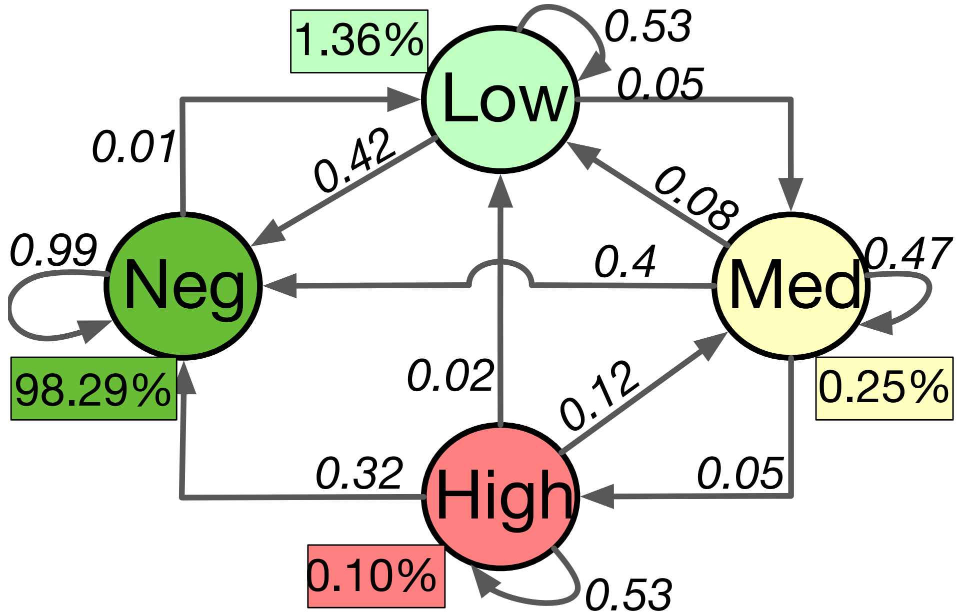
Correlation Between NAMD Completion Time and Network Congestion (NCSA Blue Waters)

Performance Variation Depends on Duration, Size and Intensity of Congestion Clouds

Characterizing Congestion On Toroidal Networks



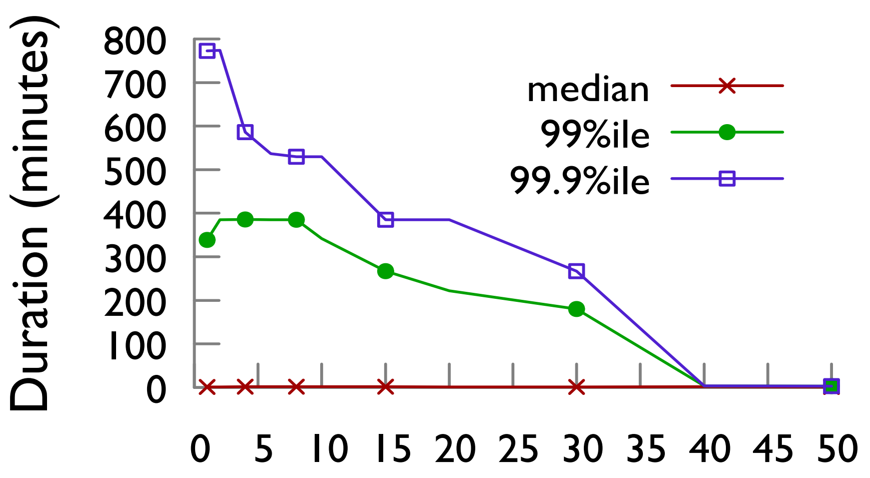
Hotspot Link Duration (NCSA Blue Waters)



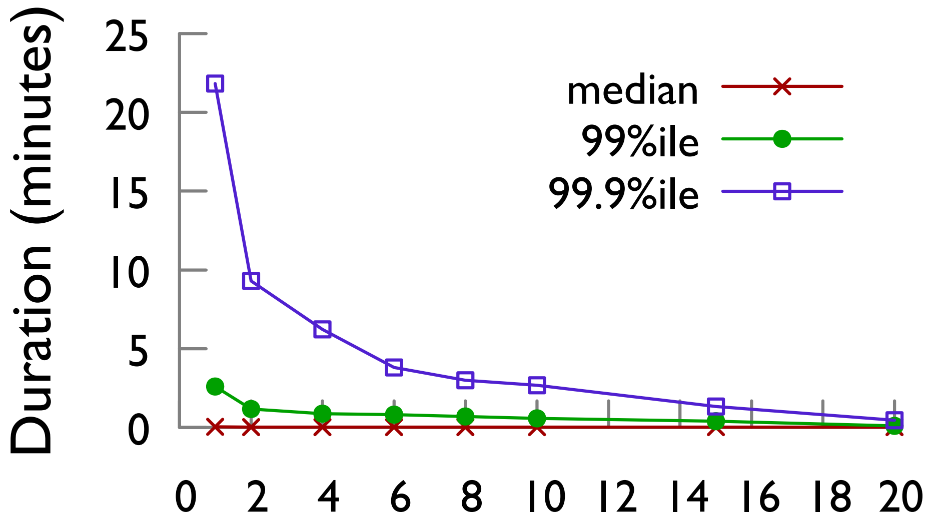
Congestion cloud: % of total system-time

- Continuous presence of highly congested links
- 95% of the operation time contained a region with a min size of 20 links.
- Max size of 6904 (17%) links
- Average congestion duration: 16 minutes, 95th percentile: 16 hours

Comparing with DragonFly Network – Link Hotspot Characterization



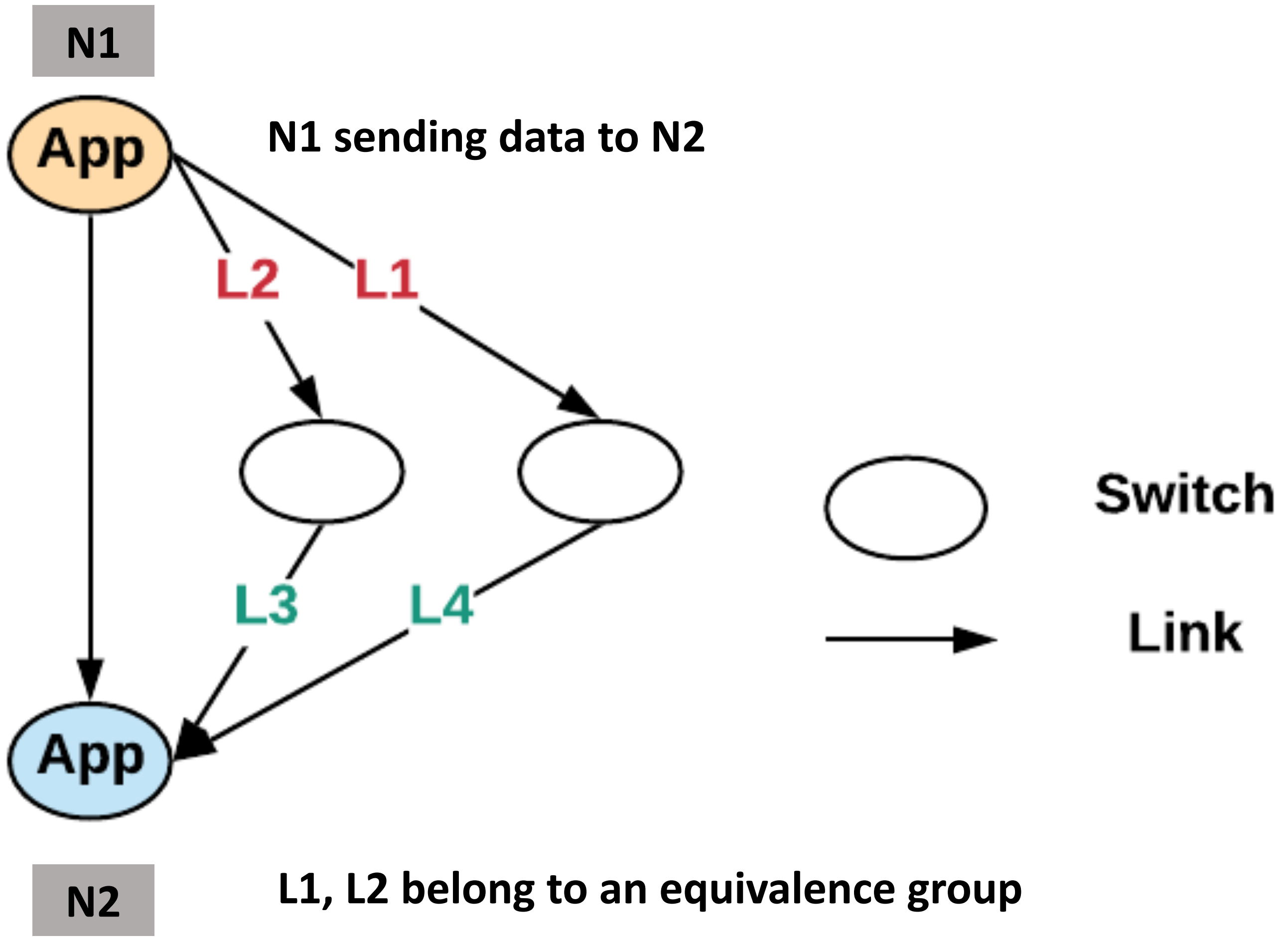
Hotspot Link Duration (NCSA Blue Waters)



Hotspot Link Duration Aries (NERSC Edison)

- Duration of continuous congestion on links significantly reduced in DragonFly
- Hotspots continue to exist

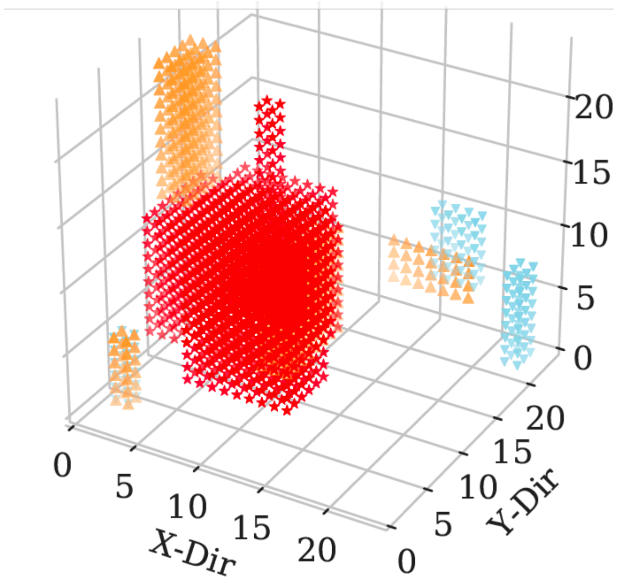
Measuring Load Imbalance: Equivalence Groups



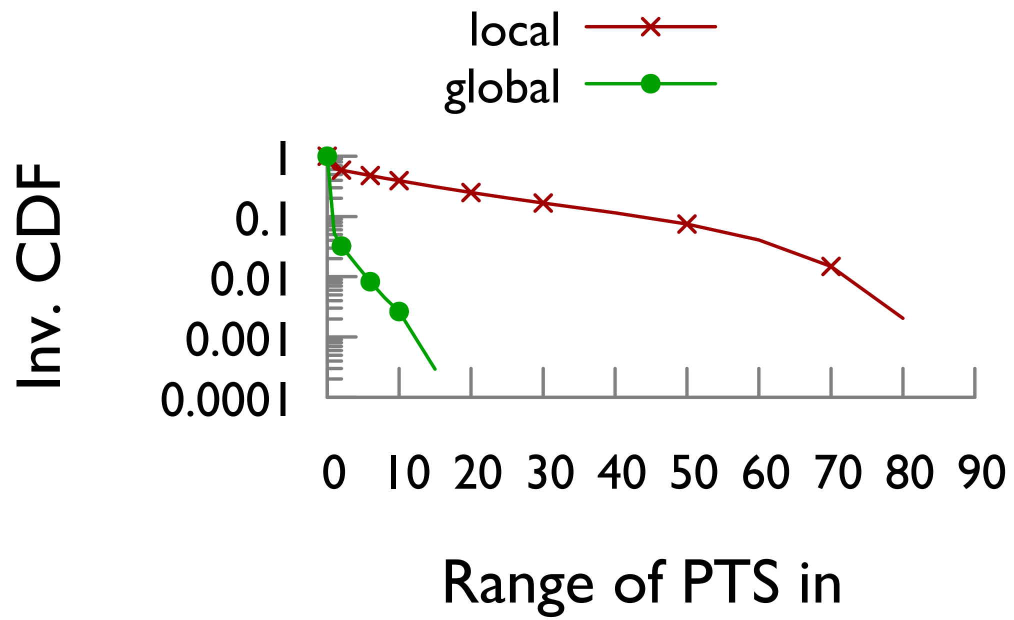
Congestion Imbalanced Scenario	
Link	Congestion (PTS)
L1	20%
L2	10%

Congestion Balanced Scenario (Ideal)	
Link	Congestion (PTS)
L1	15%
L2	15%

Comparing with DragonFly Network – Load Imbalance Characterization



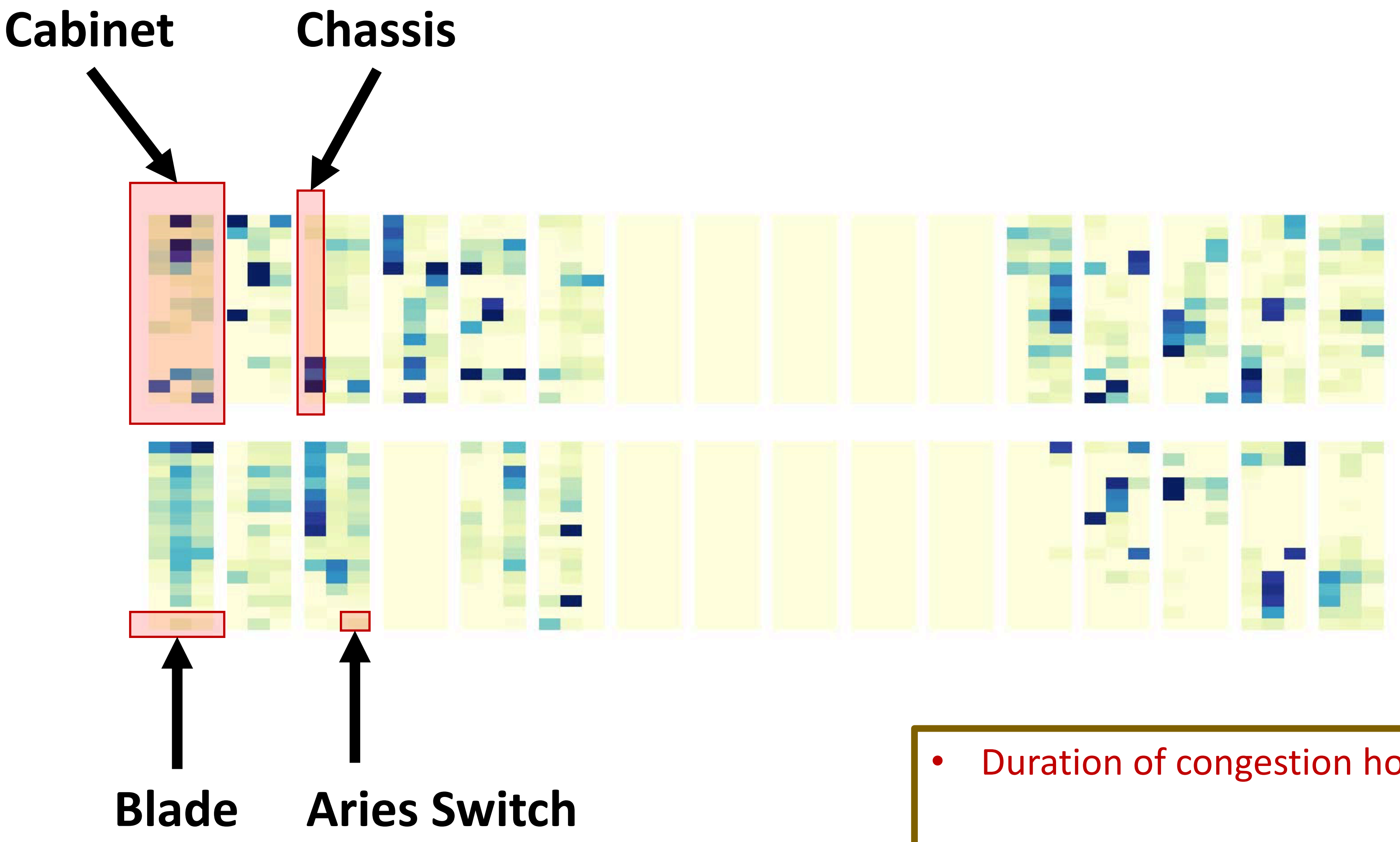
Existence of congestion cloud for long duration (NCSA Blue Waters)



(Cray Testbed System)

- Existence of load imbalance on local links
- Load balance significantly improved on Global links

Demo – 5 minutes of congestion viz on Edison



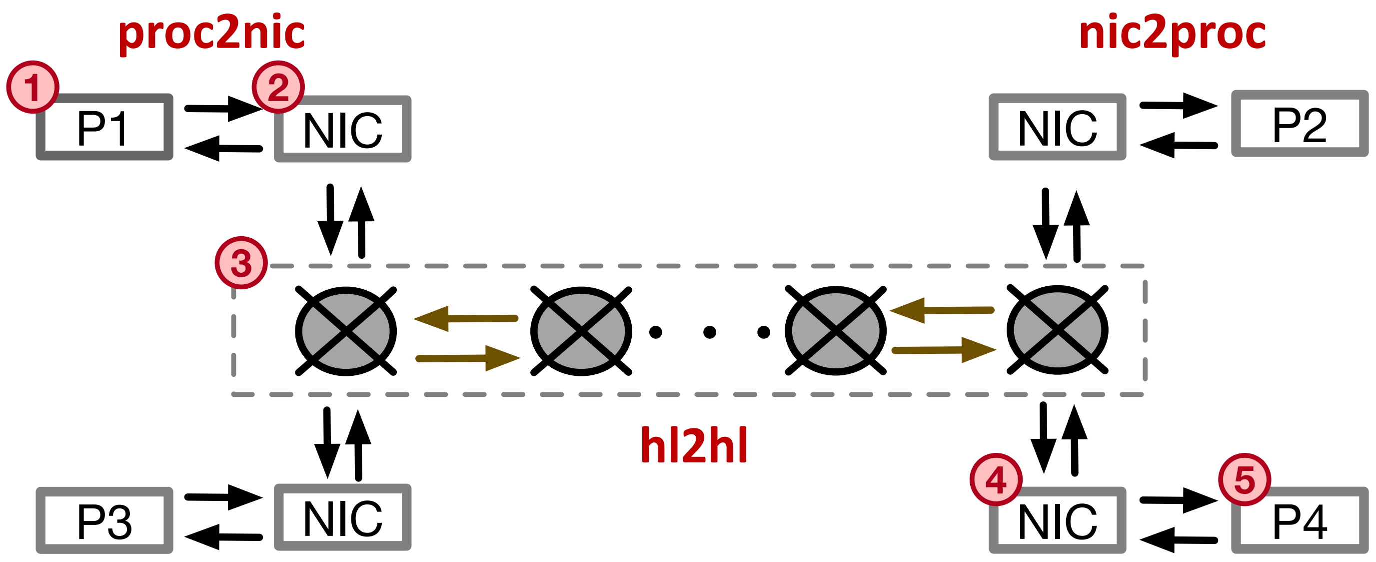
>15% PTS is dark blue
0% - 15% green gradient

- Duration of congestion hotspots is short
- Hotspots move rapidly and application continues experience congestion

Going Forward: Taking Systems Approach for Minimizing Performance Variation

Sources of Contention

Traffic flowing from P1 to P4

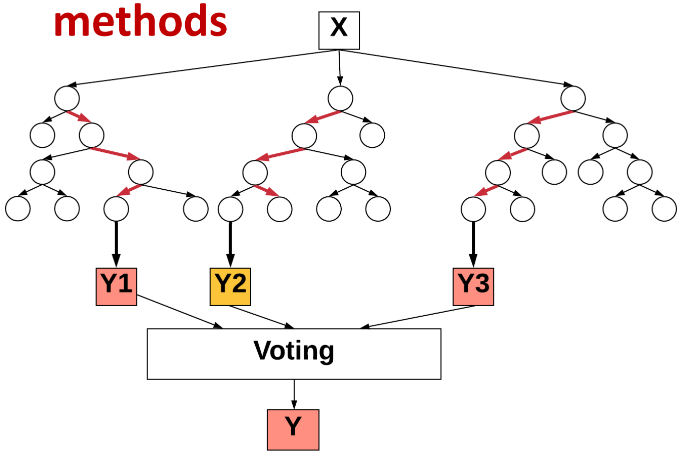


Map	Contention	Reasons
1	<src> processor	Cache conflict, TLB, OS resources
2 4	<src>/<dst> NIC	NIC buffers busy
3	Network Switches	Head of line blocking due to busy buffers
5	<dst> processor	<dst> processor not ready to receive data

Multiple sources for contention/congestion

ML-driven Performance Prediction in Aries

1. **Increase observability into system and applications**
2. **Backpressure model for contention using probabilistic methods**
3. **Use explainable machine-learning methods**



Ensemble Models

k-Fold

Cross Validation

- Diagnosis for understanding performance variation
- Understanding sensitivity to each configuration parameter
- Scheduling policy

Feature Type	**Feature List**	**Avg R2**
Application and Scheduler Specific	*app,hugepages,placement, balance, avg hops*	0.64
Network Specific	*nic2proc,proc2nic,hl2hl* PTS & SPF	0.86
All features	All the above	0.91

Conclusion and Future Work

Conclusion

- Congestion studies across generations of production systems help improve understanding of
  - App. performance variation – diagnostic models
  - Parameter tuning – selecting optimal parameters
  - Network design - load imbalance continues to be a problem

Future work

- Characterize and understand upcoming QoS features in Slingshot
- ML-driven scheduling for HPC kernels

References

S. Jha, A. Patke, B. Lim, J. Brandt, A. Gentile, G. Bauer, M. Showerman, L. Kaplan, Z. Kalbarczyk, W. T. Kramer, R. Iyer (2020). Measuring Congestion in High-Performance Datacenter Interconnects. *17th USENIX Symposium on Networked Systems Design and Implementation* ***(NSDI)***.

S. Jha, A. Patke, J. Brandt, A. Gentile, M. Showerman, E. Roman, Z. Kalbarczyk», W. T. Kramer, R. Iyer (2019). A Study of Network Congestion in Two Supercomputing High-Speed Interconnects. *2019 IEEE 26th Annual Symposium on High-Performance Interconnects* ***(HOTI)***.

S. Jha, J. Brandt, A. Gentile, Z. Kalbarczyk, R. Iyer (2018). Characterizing Supercomputer Traffic Networks Through Link-Level Analysis. *2018 IEEE International Conference on Cluster Computing*

Saurabh Jha, Shengkun Cui, Tianyin Xu, Jeremy Enos, Mike Showerman, Mark Dalton, Zbigniew T. Kalbarczyk, William T. Kramer, Ravishankar K. Iyer (2019). Live Forensics for Distributed Storage Systems. *arXiv e-prints*.

PDF Cite