SAND2015-10083C

Exceptional service in the national interest

Sandia National Laboratories

Overtime: A Tool for Analyzing Performance Variation due to Network Interference

Ryan E. Grant, Kevin T. Pedretti, Ann Gentile

U.S. DEPARTMENT OF ENERGY

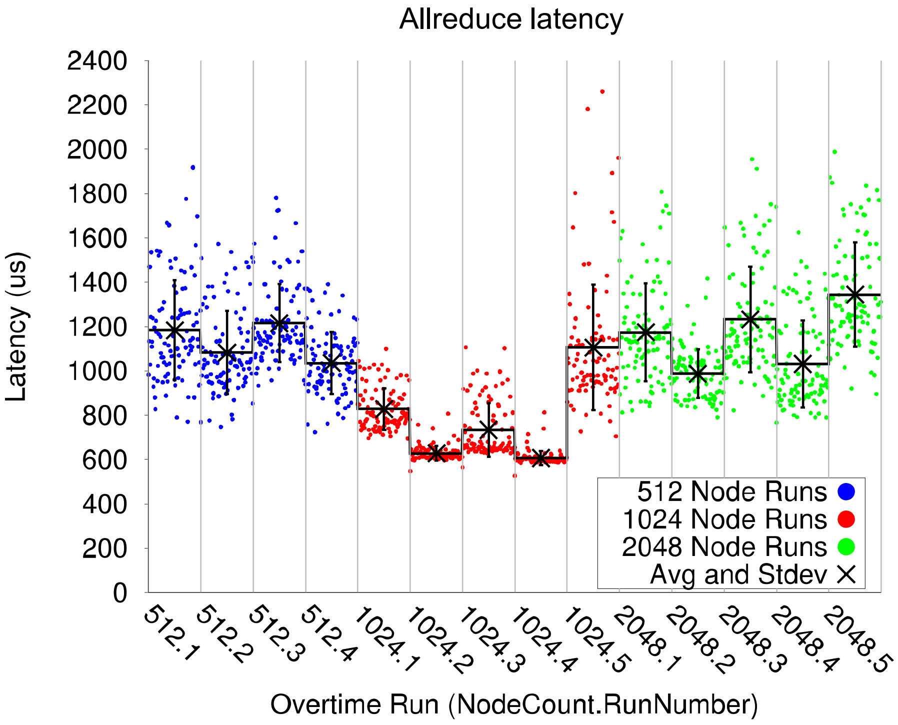
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

CCR Center for Computing Research

Performance Variation in MPI

- Performance variation in MPI can have significant impact on code performance
- Latencies can range almost 4X for a single allreduce operation over different runs
- Understanding this variation is key to leveraging system performance for jobs
- Network conditions need to be understood when jobs are placed on a system
- Understanding the intersection of different jobs communications is difficult



Performance Variation in MPI

- Not all performance variation is due to network congestion/interference
- OS noise can cause this issue as well
- What variation is caused by the network and what variation is caused by other factors?
- Goal:
  - Determine impact on MPI point to point and collectives over time on a production system
  - Mitigate OS noise impact by not fully loading all CPUs on nodes under test
  - Correlate network performance counter data with observed performance
  - Characterize network interference over time with observed causes

Network Performance Variation

- Difficult to attribute to a single factor
- Normally caused by an intersection of multiple jobs behaviors
- Jobs utilize the network at different times during execution
  - Makes determining network conditions from job list alone difficult
  - Communication frequency dependent on application and workload
- Shared network resources can be hard to reason about
- Some network topologies make reasoning easier
  - Using a 3D torus
    - Easier to reason about job placement and traffic patterns

Assessment of Variation

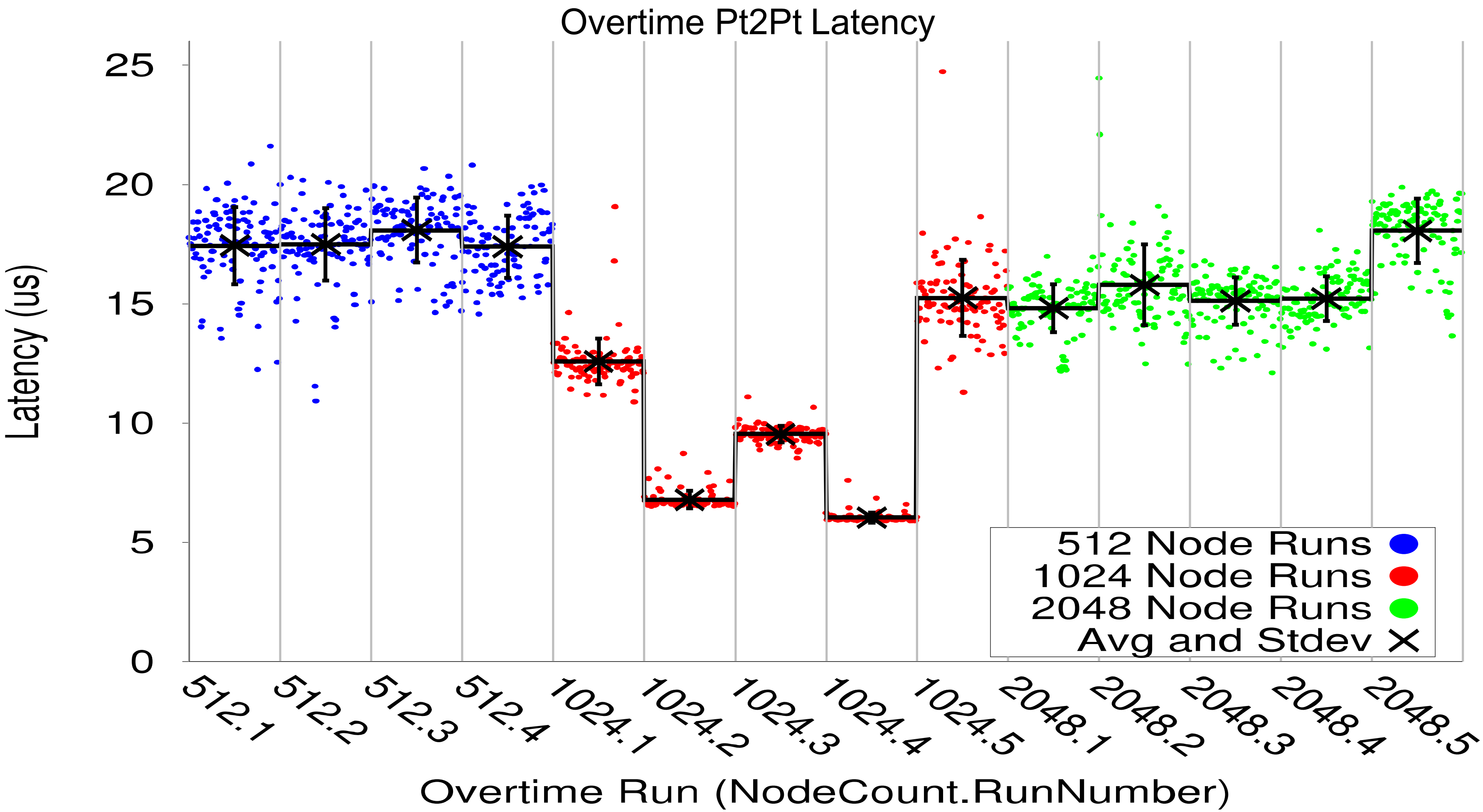
- In order to assess network performance variation from an MPI perspective – Overtime

- Overtime is a tool that measures performance and record network performance counters
  - MPI pt2pt latency, bandwidth, and all-reduce performance
  - Alternates between MPI performance and observation of system with no communication
  - Sleep periods are adjustable – default to exact time period of previous MPI tests so network counters are comparable
  - Leverages rich set of network performance counters for Cray Gemini networks

Experimental Setup

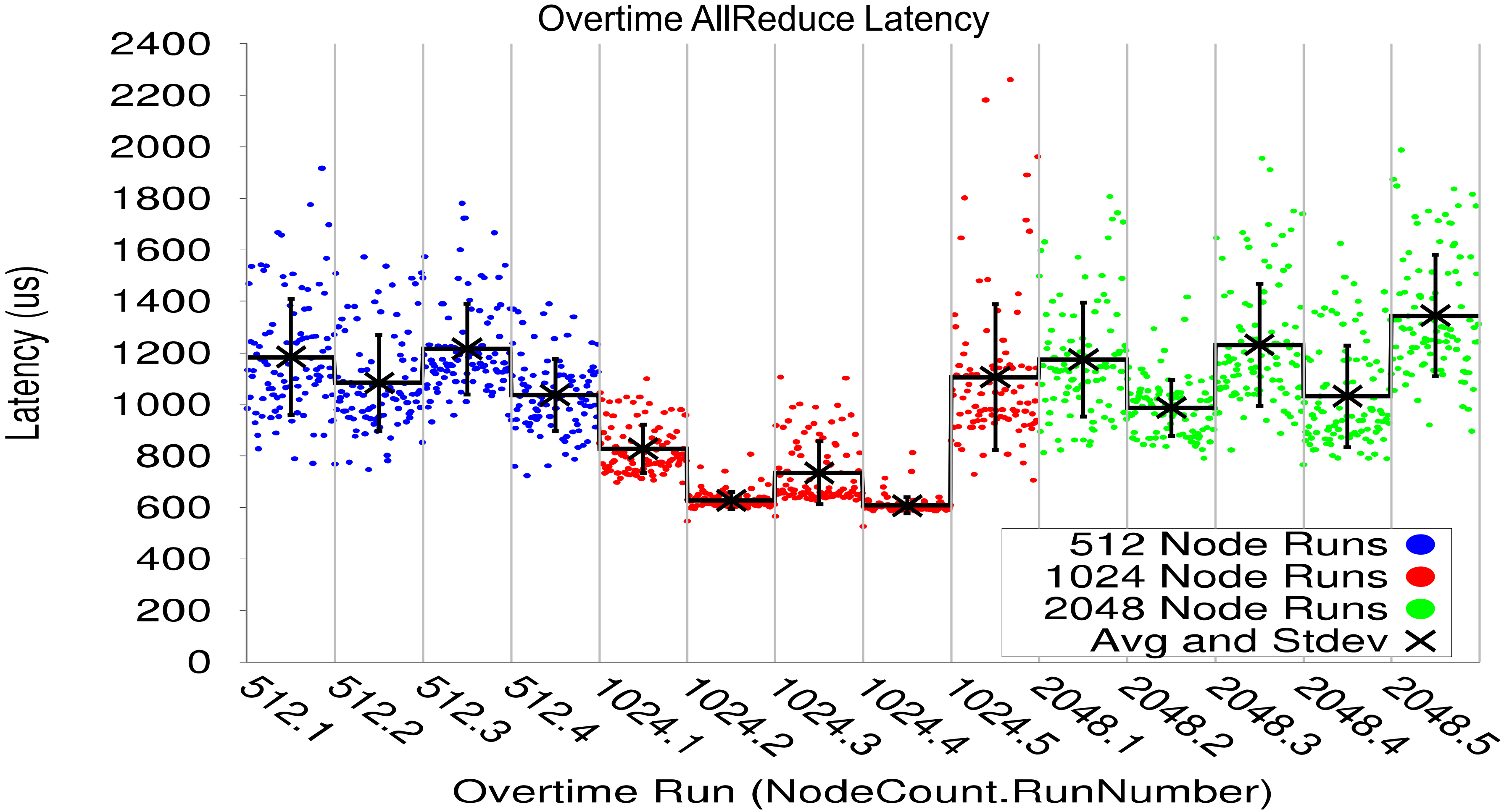
- Testing performed on the Blue Waters system as the National Center for Supercomputing Applications
- 22,640 Cray XE nodes and 4,228 Cray XK7 nodes
  - Only used XE nodes
- 237 XE cabinets, 44XK cabinets, 13.34 PF peak
- All tests performed during regular production time
- Cray Gemini 3D torus network
  - 24x24x24
  - 13,824 Gemini chips per system
    - Each Gemini connected to 6 neighbors
  - 2 compute nodes share a Gemini
  - Peak injection bandwidth 9.6GB

Results and Observations



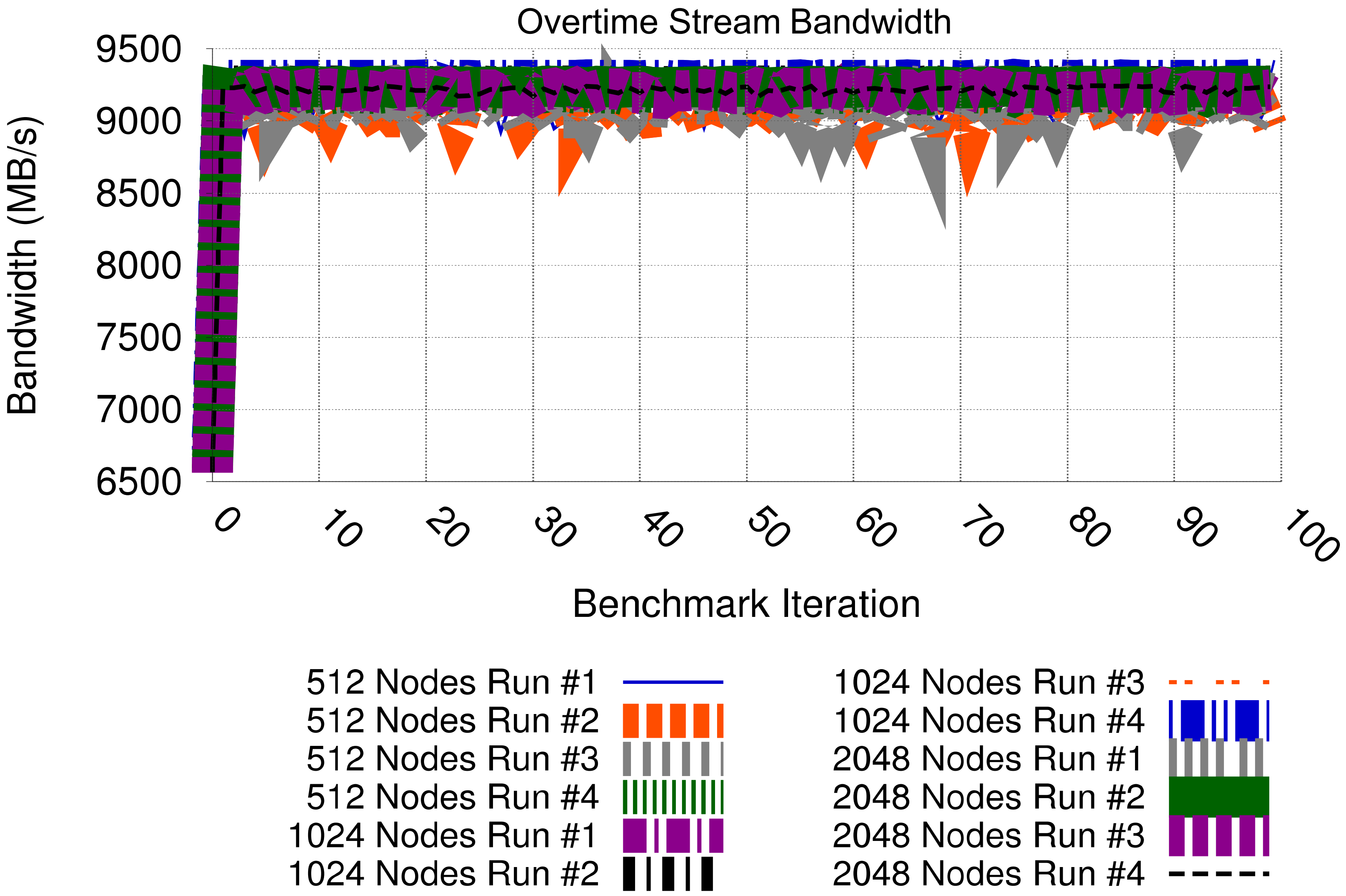
- Latency swings over time can be significant
- Although relatively stable in a sampling period (1 hour)

Results and Observations



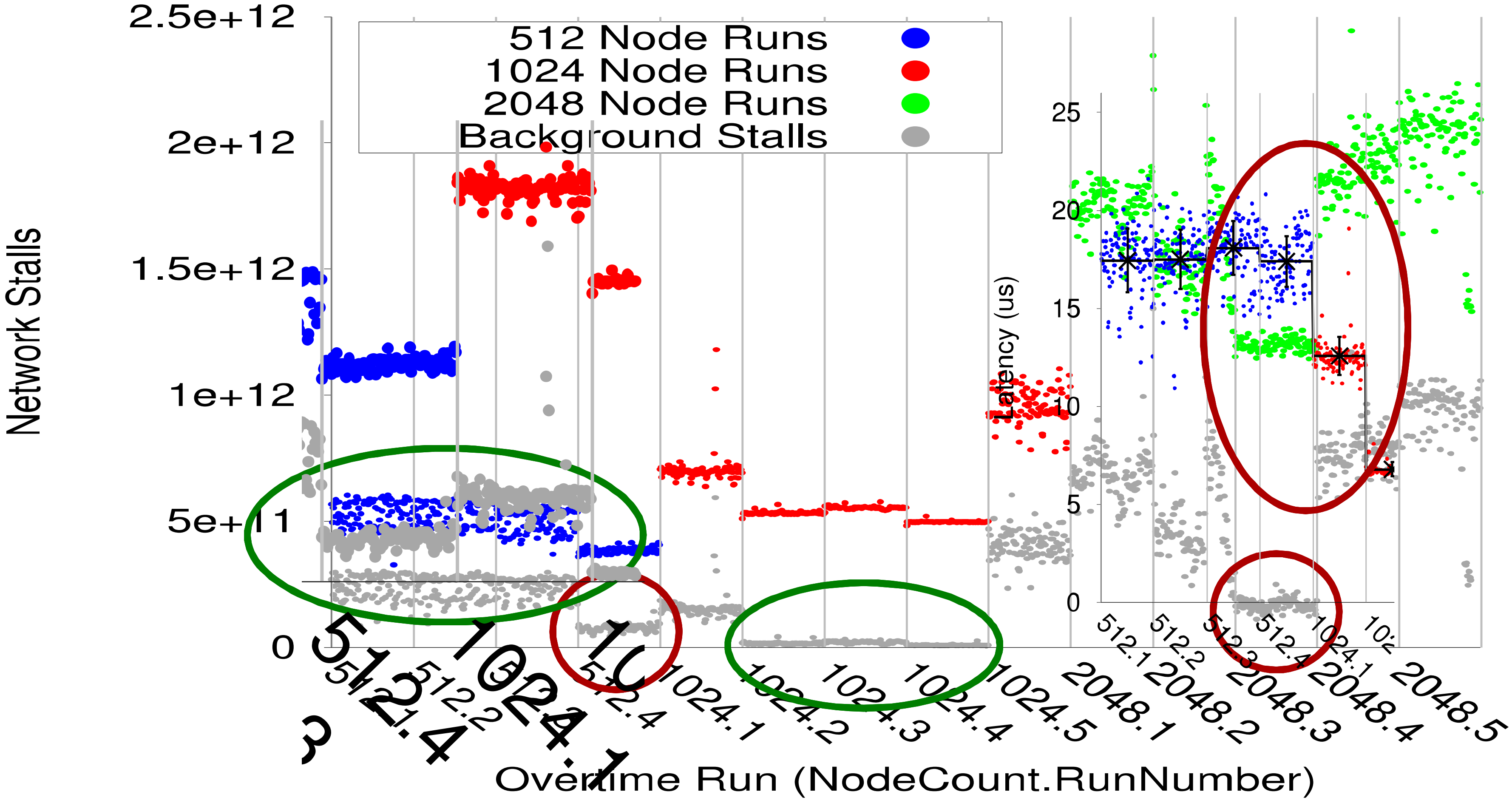
- All-reduce sees variation, but with higher std. dev.
- No major trends with obvious changes within time periods
- Bandwidth is much less impacted over time

Results and Observations



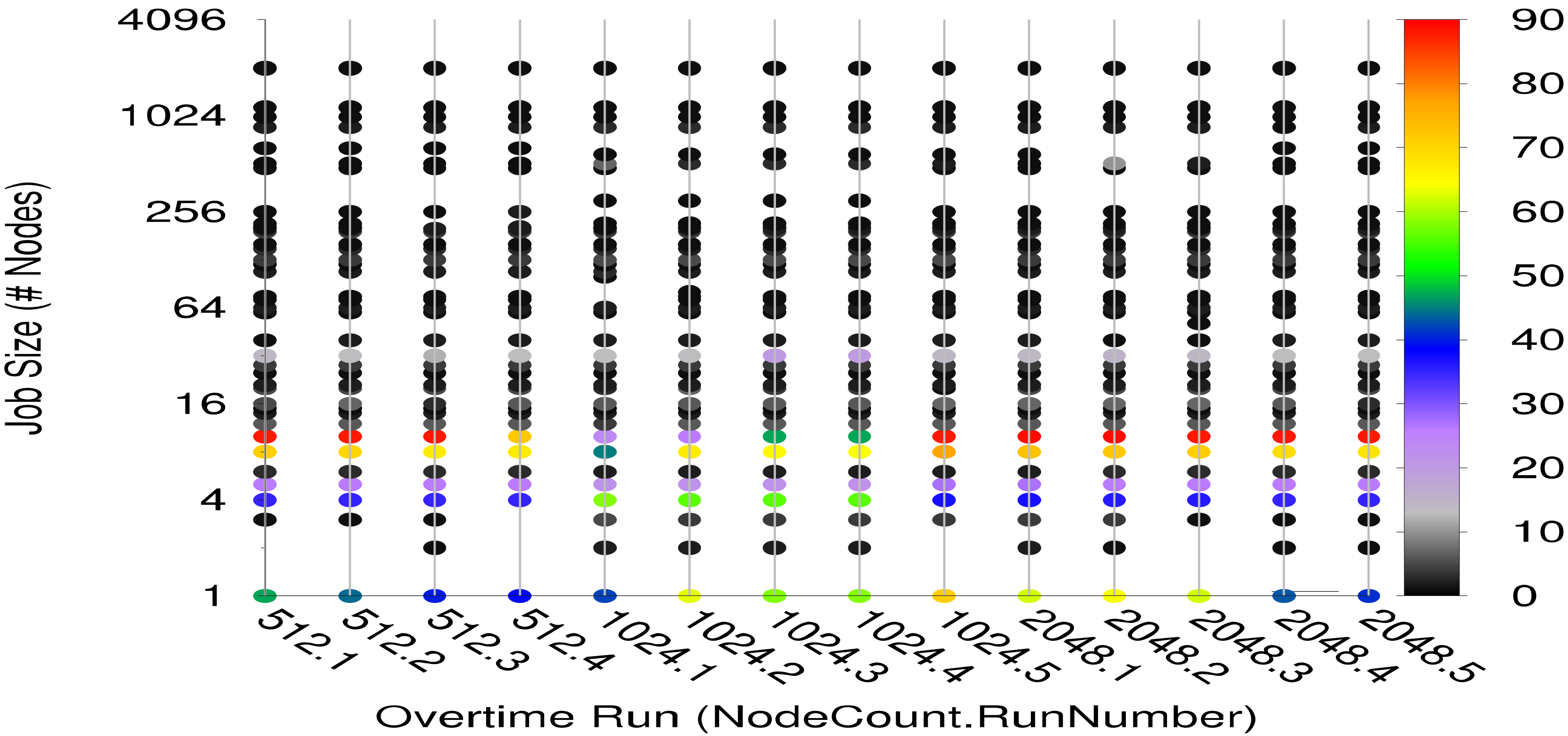
- Bandwidth is much less impacted over time than latency or reductions

Network Performance Counters



- To understand, take a look at network performance counters
- Some correlation between idle stalls and observed perf
- Important exceptions to this observation – multiple factors

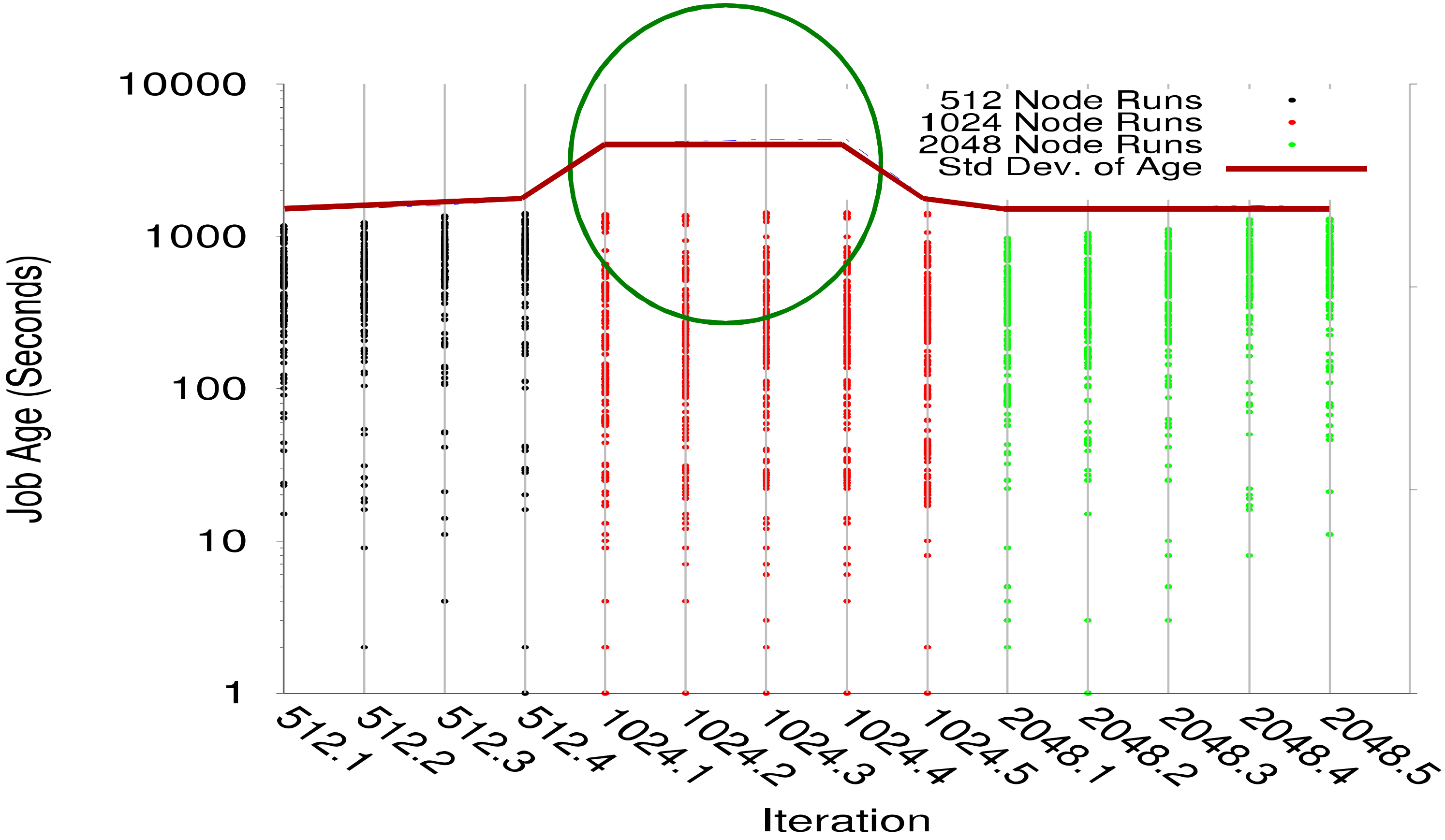
Job Mix



Job size with color coding of number of jobs of that size executing

- Is job mix playing a role?
- Job sizes are relatively regular

Job Age



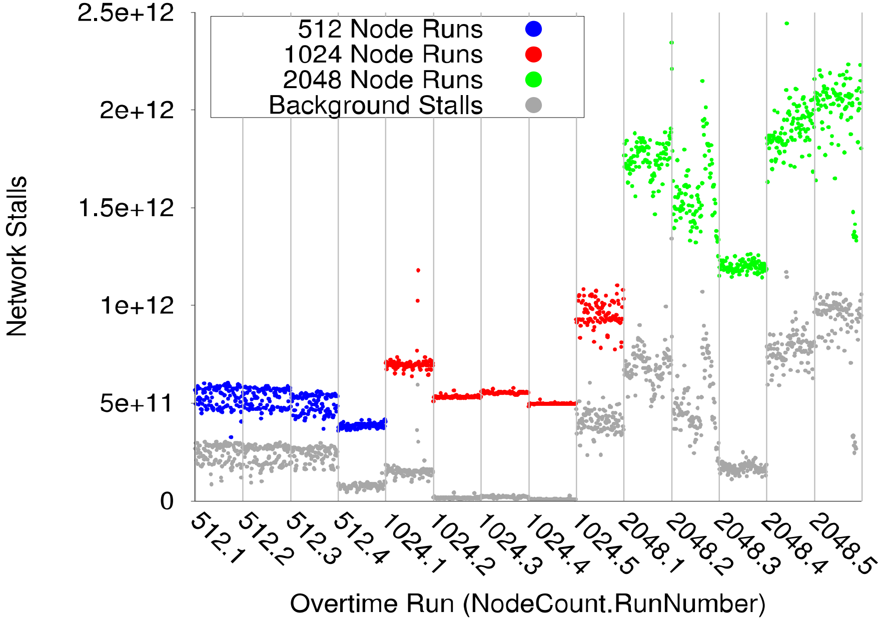
- Could age of the jobs running be a factor
  - Jobs starting up could cause network variation

Observed Stalls vs. Job Age

- If job age correlates well with best runs why don't we see a change in resting stalls over other runs?
  - If job startup/completion is causing network interference, it should show up in the observed stalls over time…

- Conclusion:
  Network stalls are not sufficient to understand or predict network performance

- Separating stalls out by links (x,y,z) doesn't provide further insight either



The Good News

- Although multiple factors are at play when trying to predict performance based on job age/network counters…
- Network performance is relatively steady for 10-60 minute periods
- Actively measuring network performance provides reliable feedback
- Network measurements during idle periods provide reasonable feedback
  - With some false positives (optimistic network prediction)
- Overtime could be used to assess a potential job allocation
  - Determine if the predicted network performance matches requirements

Conclusions

- Multiple factors make network performance prediction difficult but possible to achieve (not 100% of the time)

- Overtime tool will be available for others to use
  - Part of the Sandia Microbenchmarks
  - http://www.cs.sandia.gov/smb/

- Use for:
  - Assessing job placement for fulfilling networking requirements
  - Composing with application variation studies to understand networking variation independently
  - Studying network interference on other networks (e.g. Aires)
  - Evaluating periods of network variability on other systems

Thank you

Questions?

Acknowledgments:

This work was funded through the Computational Systems and Software Environment sub-program of the Advanced Simulation and Computing Program funded by the National Nuclear Security Administration