Performance Variability Due to Job Placement on Edison

Dylan Wang[1], Abhinav Bhatele[2] (advisor), Dipak Ghosal[1] (advisor)
[1]University of California, Davis, [2]Lawrence Livermore National Laboratory

Abstract

Some scientific applications running on machines like the Edison supercomputer can suffer from high variability in run-time. This leads to difficulties in debugging and optimization for end users and less efficiency for the supercomputing facility. The objective of this research is to characterize the application run-time performance and identify the root cause of the variability on machines with the Aries network.

Edison

- 133,000+ cores
- Global Network Bandwidth of 23.7TB/s
- 5500+ nodes (two 12-core Ivy Bridge processors per node)

Cray Aries Interconnect

- Aries router = 4 nodes, 40 network tiles with 1 incoming and 1 outgoing port, creating the 3-tier "Dragonfly" topology

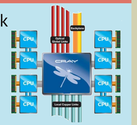

Figure 1: The Aries router

1) Chassis: 16 Aries routers
2) Group: 6 Chassis
3) System: 15 Groups

- Aries routers are capable of adaptive, non-minimal routing

Figure 2: A chassis

Figure 3: A cabinet making half a group

Figure 4: The optical cables connecting groups

Experiment Setup

- The application MILC was run on Edison multiple times
  - Three allocation sizes: 256, 384, and 512 nodes
  - To avoid OS noise, we skip first/last core of each socket
- MILC is computation-constant application that communicates in a weak scaling 4D stencil

- Our first goal is to find metrics that strongly correlate with the application's performance.

We are collecting the following pieces of information for each run:
1) Program run time
2) The state of the machine
3) MPI Profiler (mpiP) report
4) Hardware counter values

Data Analysis

- Determining metrics that simultaneously "describe" an allocation and the interference due to congestion on the routers and also correlates with application performance is challenging.

- We identified the following metrics in three catagories for correlation tests:
  - Allocation-related
    - The coordinates of the allocation given
  - MPI-related
    - MPI behavior reported by mpiP
  - Counter-related
    - The load, bandwidth, and congestion of the routers and links

Visualization

The figures below are 2D visualizations of the entire system at two different machine states.

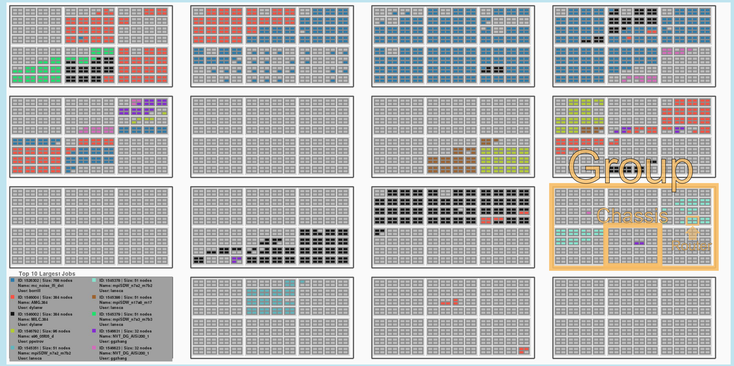

Figure 5: Good run (322 seconds)

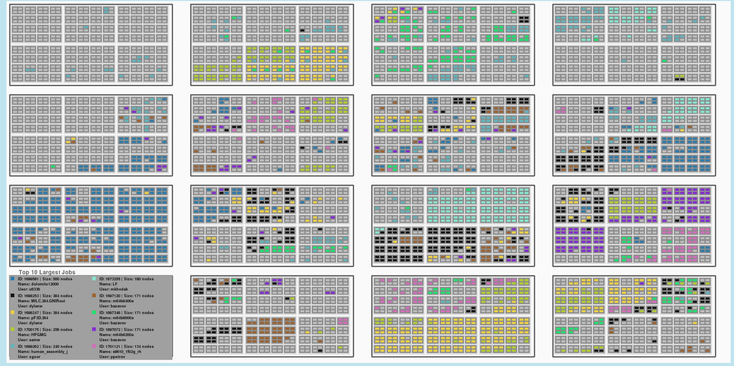

Figure 6: Bad run (857 seconds)

Figures 5 and 6 above show the visualization of the system at two different times corresponding to two different runs of MILC in black. What are the reasons for the large variability between the two runs?

Results and Analysis

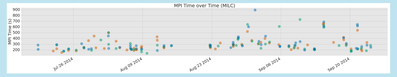

Figure 7

Mean	SD	IQR	CV	0%	25%	50%	75%	100%	n
304s	129s	110s	43%	143s	221s	272s	332s	896s	165

Table 1

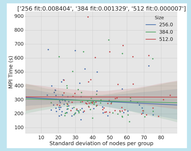

Figure 8: StdDev of nodes/group

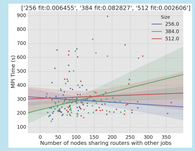

Figure 9: Number of nodes sharing routers with other jobs

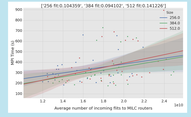

Figure 10: Average router load

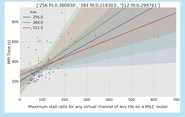

Figure 11: Max stall per incoming flit for any port of a router

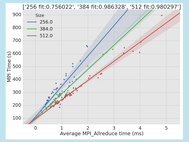

Figure 12: Average MPI_Allreduce

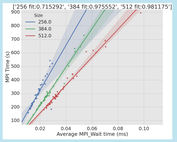

Figure 13: Average MPI_Wait

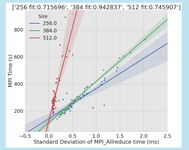

Figure 14: Stddev MPI_Allreduce

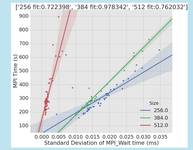

Figure 15: Stddev MPI_Wait

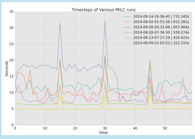

Figure 16: Breakdown of MILC timestep duration

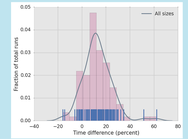

Figure 17: Time difference for minimal versus dynamic routing

Conclusions

- Table 2's CV suggests an approximate 43% in variation for job MPI time
- Our node allocation metrics individually show almost no correlation, meaning placement is not significant or is covered by general interference
- Hardware counters may correlate with MPI performance, though the results are not conclusive

Future Work

- Run other applications such as pF3D which have different communication patterns
- Controlled expirements such as finding a base case time by reserving a whole group

Thanks to Timo Bremer (LLNL), Todd Gamblin (LLNL), Martin Schulz (LLNL), Yarden Livnat (Utah), Nick Wright (LBNL), Katherine Isaacs (UC Davis), and Andrew Titus (MIT)

Reference: Abhinav Bhatele, Kathryn Mohror, Steven H. Langer, and Katherine E. Isaacs. 2013. There goes the neighborhood: performance degradation due to nearby jobs. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '13). ACM, New York, NY, USA, Article 41 , 12 pages. DOI=10.1145/2503210.2503247 http://doi.acm.org/10.1145/2503210.2503247

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.
This work was funded by the Laboratory Directed Research and Development Program at LLNL under project tracking code 13-ERD-055 (LLNL-POST-662284).