Simulating and Visualizing Traffic on the Dragonfly Network

Abhinav Bhatele, Nikhil Jain, Yarden Livnat, Valerio Pascucci and Peer-Timo Bremer

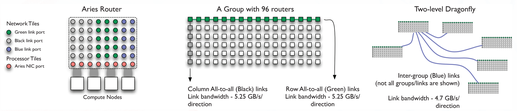
Abstract — The dragonfly topology is a popular choice for building high-radix, low-diameter networks with high-bandwidth links. Even with a powerful network, preliminary experiments on Edison at NERSC have shown that for communication-heavy applications, job interference and thus presumably job placement remains an important factor. In this poster, we explore the effects of job placement, parallel workloads and network configurations on network throughput to better understand inter-job congestion and interference. We use a simulation tool called Damselfly to model the network behavior of Edison and study the impact of various system parameters and configurations on network throughput.

Methodology and Tools

Damselfly: Analytical network simulation tool to model congestion on dragonfly networks
- Outputs the steady-state traffic on all ports of each router on the network

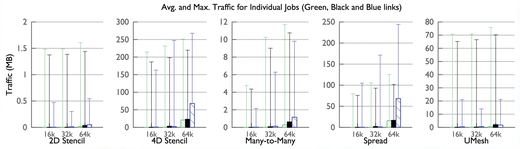
Modifications to Damselfly:
- Support to simulate arbitrary connections between routers in a dragonfly topology
- Enable simulation of multi-job workloads with user-defined placement
- Ability to attribute link traffic to individual jobs within a workload

DragonView: Visual analytics tool to understand network traffic and throughput
- Radial view shows the job placement and inter-group (blue) links
- Matrix views show the intra-group (green and black) links



Simulation Setup

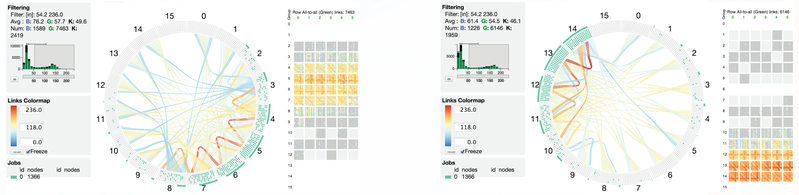
- Use the network topology of Edison provided by NERSC system administrators
- Simulate an allocation policy that resembles the batch queue on Edison
- Five communication patterns representative of codes run at NERSC: 2D Stencil, 4D Stencil, Many-to-many, Spread, Unstructured Mesh (UMesh)



Average and maximum traffic on green, black and blue links for individual jobs with different communication patterns and core counts (16k, 32k, 64k cores).

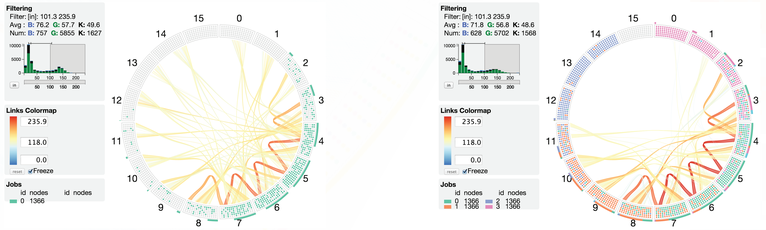
Individual Jobs

Traffic on blue (radial view) and green (matrix view) links for individual jobs running alone on 64k cores of Edison.



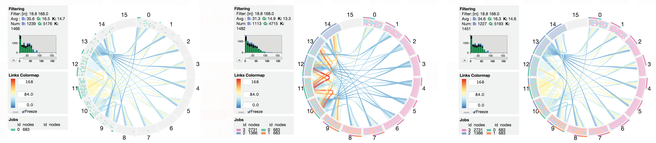
Job placement affects the traffic on blue and green links for a simulation of 4D Stencil running on 32k cores. On the left, the job is scattered which leads to lower maximum traffic (191 MB) but higher average traffic per link (51.4 MB) in comparison to the job on the right which is more compact (maximum traffic: 236 MB, average traffic: 51.4 MB).

Inter-job Interference



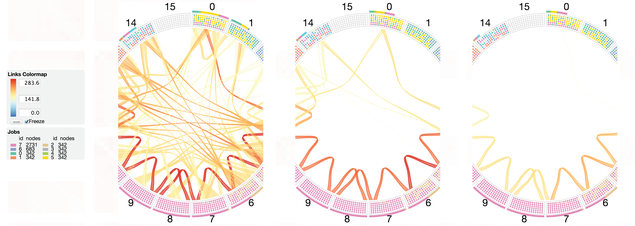
When Job 0 (4D Stencil) is run in a workload along side other jobs (right), the number of links with traffic above a certain threshold decreases and overall maximum traffic on blue links increases (231 MB) as opposed to when it is run individually (191 MB, left). In the parallel workload run, Job 0's traffic is confined to fewer blue links in order to share bandwidth with other jobs.

Inter-job Interference (contd.)



Maximum traffic on blue links increases when Job 0 (Spread) is run in a workload along side other jobs (middle) versus when run individually (left). However, if other jobs are not communication-intensive, the effect on a particular job (Job 0) might be minimal (right).

Network Wiring



Network health of a dragonfly-based system when simulating an eight-job parallel workload. Left to right: Different network configurations using two, three and five inter-group links per router pair. The additional links remove hot-spots and reduce the overall average network load.

Summary

- Black links are usually not congested on Edison because of 3 links per router pair
- In a parallel workload, communication of each job gets restricted to fewer inter-group links
- Using additional blue links reduces the overall congestion but at a significant dollar cost

Project page: https://computation.llnl.gov/project/extreme-computing/interconnection-networks.php
Software download: https://github.com/scalability-llnl

COMPUTATION

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-676008).

ASC