Analysis of the Slingshot high-speed network

Thomas Sorkin[1, 2] William Shanks [1] Storm Knight [2]

[1]NSF National Center for Atmospheric Research [2]University of San Francisco

Background

The flagship supercomputer at the National Science Foundation (NSF) National Center for Atmospheric Research (NCAR), Derecho, uses the novel **Slingshot** interconnect to link its 2570 nodes. Developed by Hewlett Packard Enterprise (HPE), Slingshot is an Ethernet-based high-speed network using a **Dragonfly** topology.

The purpose of this project is to create visualizations of Derecho's underlying network. These visualizations inform the High-Performance Computing Research and Development Group (HPCRD) at NSF NCAR about the network's performance behavior.

Two systems were analyzed for this project: Derecho and Gust

- **Derecho** has 2488 CPU nodes, 82 GPU nodes, and 200 switches.
- **Gust** is the test system for Derecho. Gust has 16 CPU nodes, 2 GPU nodes, 10 switches, and mimics Derecho's design at a smaller scale.

Counter Collection

In order to gather data, we used **dump_counters**, a built-in utility on Slingshot switches. This utility provides counters for each of the 64 ports on an individual switch. Counters range from bandwidth usage to congestion-tracking counters. In order to consolidate and aggregate the influx of data, we used **Telegraf**, a metric collection tool, to gather counter output onto a fabric manager node. The fabric manager can then send the cleaned data to an admin node, which can in turn send the data to **TimescaleDB** for storage.

Gust Node Graph

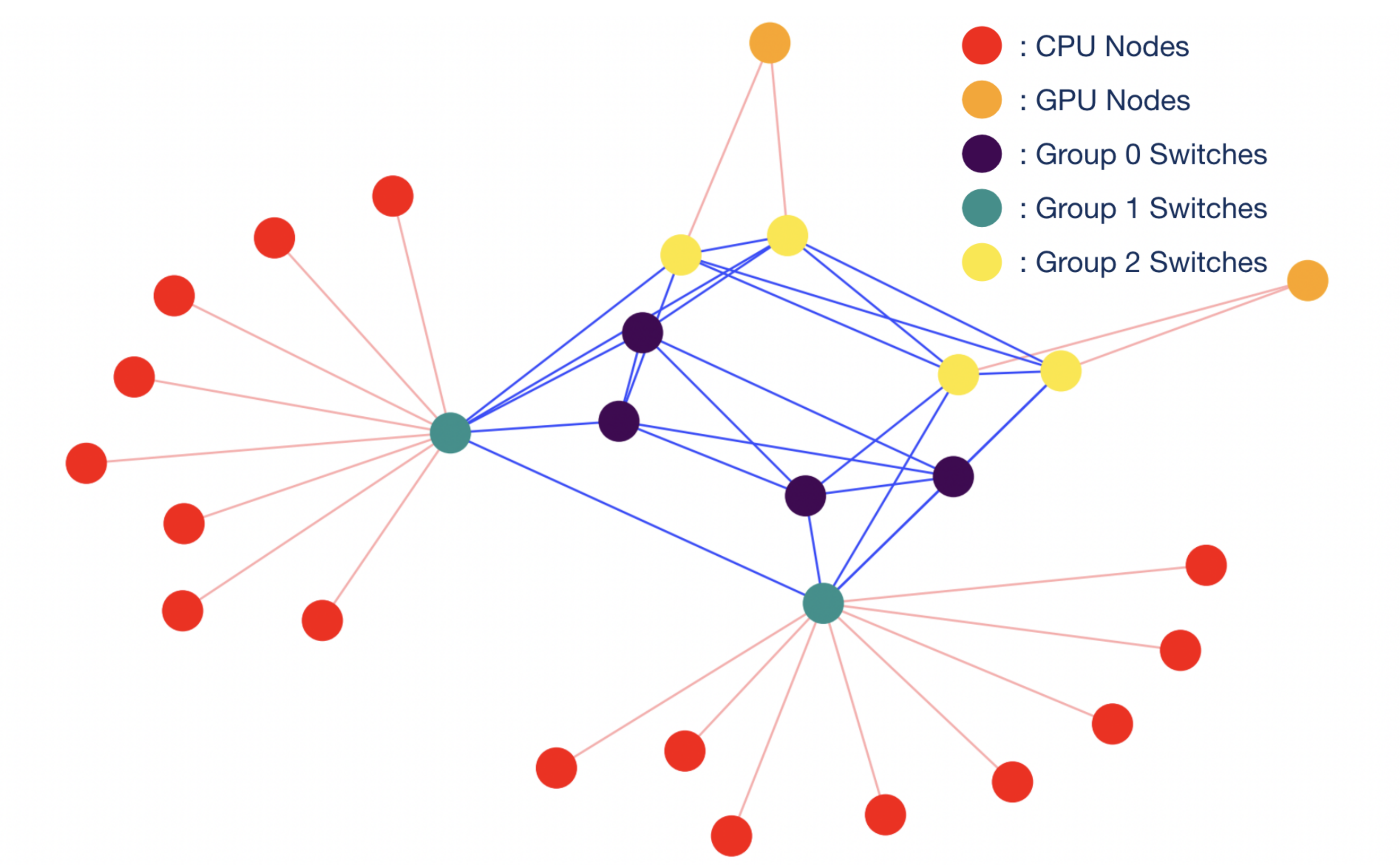

Figure 1. Visualization of Gust using the networkx and matplotlib Python packages

Derecho Node Graph

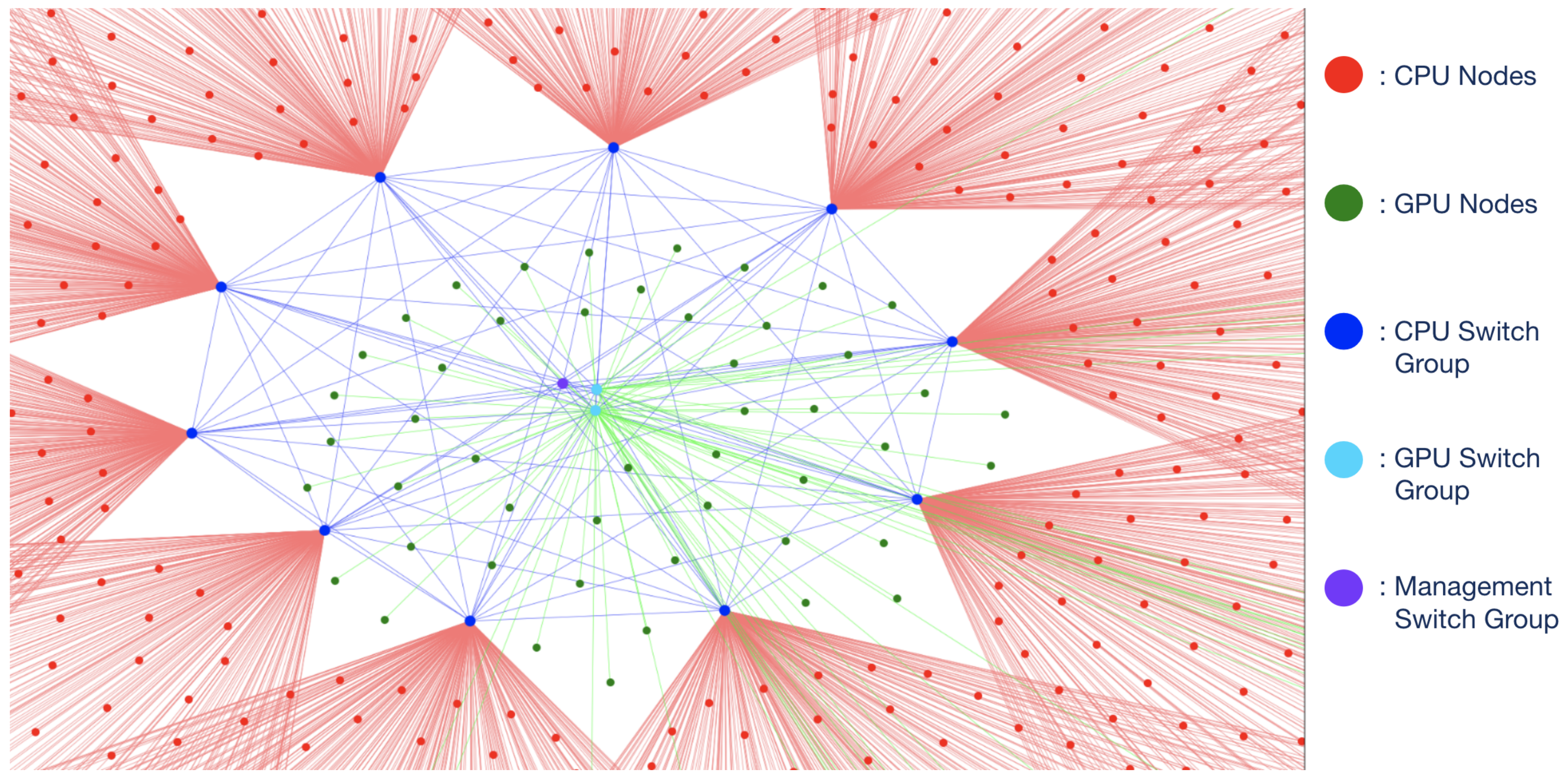

Figure 2. Visualization of the entire Derecho network. Switches are grouped together based on Dragonfly groups

Derecho Switch Group 1 Node Graph with Counters

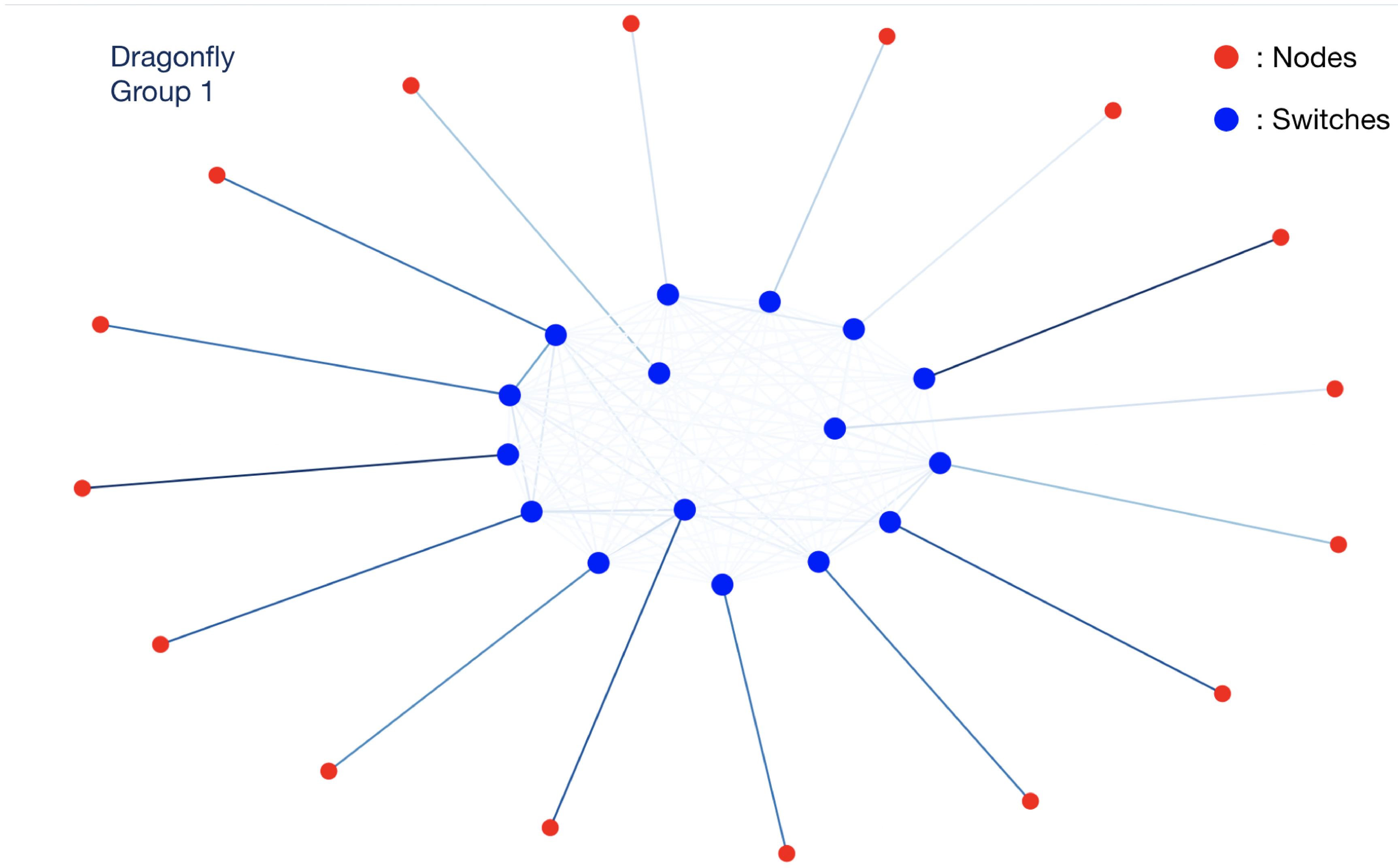

Figure 3. Visualization of an individual switch group on Derecho, with counter data added. Darker links correspond to increased bandwidth usage (bits/second)

After testing several different visualization tools, node graph visualizations were eventually built with the **networkx** python package, and displayed with the **matplotlib** python package.

Other visualization tools we tested include: **Graphviz**, **pyvis**, and **Gephi**.

Grafana Heatmaps

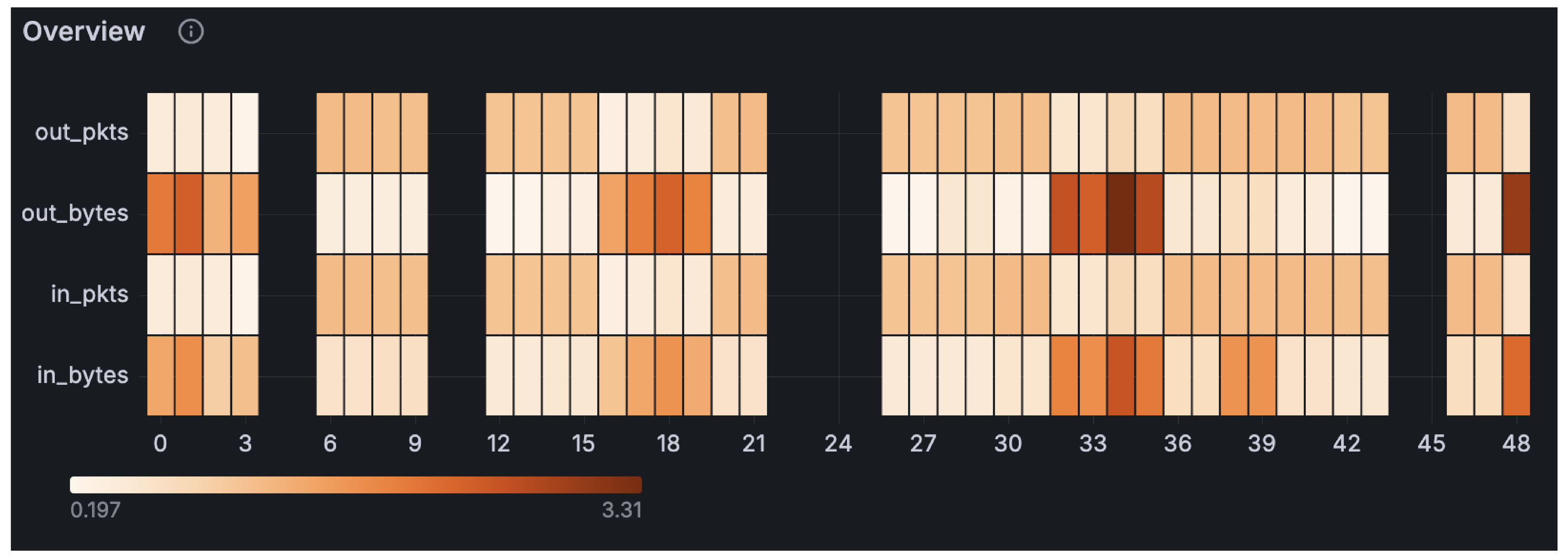

Figure 4. Each Grafana dashboard shows the average normalized value of different counters. A darker color signifies a higher value. The numbers below the heatmap correspond to the ports on a switch

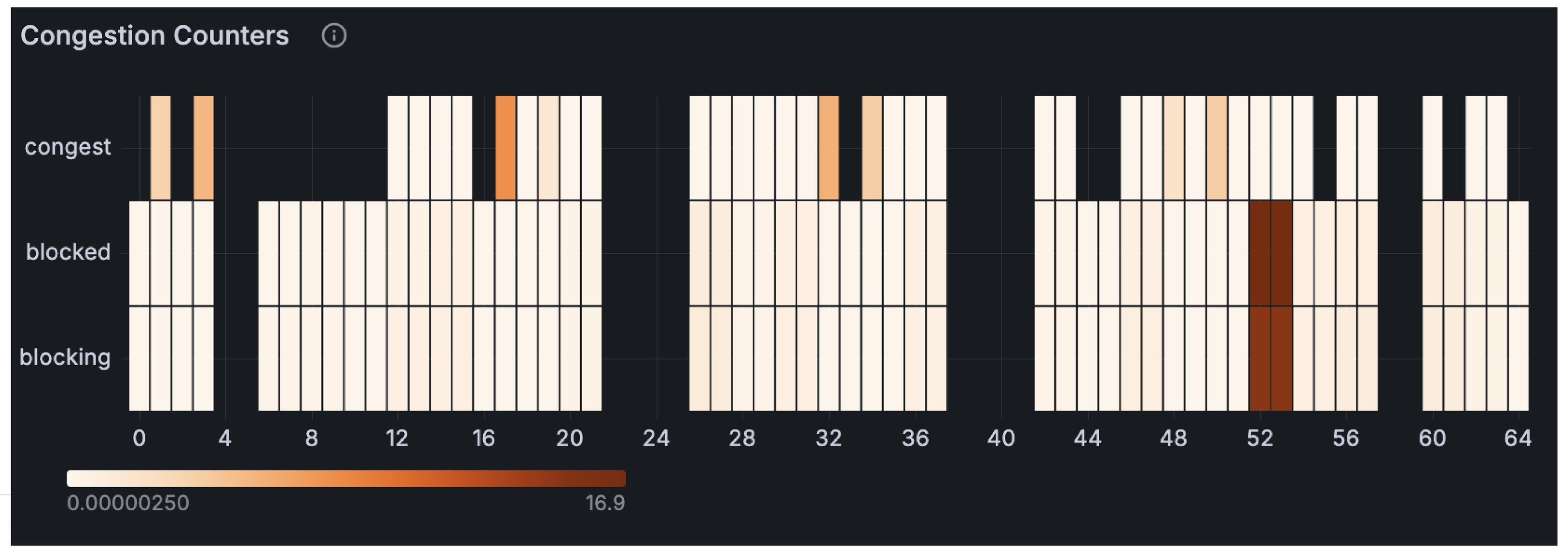

Figure 5. This Grafana dashboard displays congestion-related counters for an individual switch. *congest* is the number of congestion ACK packets received over that port. *blocked* and *blocking* represent the frequency that a port is blocked or blocking because of congestion.

Findings and Applications

Analysis of the network using these visualizations has shown that:

- The network is **most active** between the hours of 2PM and 4PM
- The network is **least active** between the hours of 2AM and 5AM

These visualizations allow the HPCRD Group to identify switches with more traffic or congestion. Furthermore, the node graphs and Grafana heatmaps operate in tandem to get different viewpoints of the network. The node graphs visualize either the entire network or an individual switch group, while the Grafana heatmaps show counters for each port on an individual switch.

Future Work

- Examine network usage over several days to identify more active days of the week
- Combine these visualizations with job scheduler data to identify a correlation between jobs and their impact on the network
- Create snapshots of the network and network usage at specific moments in time to analyze the network in the recent past.

Acknowledgements My mentors William Shanks and Storm Knight for their guidance throughout the program, the HPCRD team for their support, the 2024 SIParCS programming team for a wonderful summer, and all my fellow 2024 SIParCS interns. Additionally, NSF NCAR for funding this research.