Performance Metrics for the Chimera Cluster

Overview:

The Chimera Cluster at UMB is in need of a way to collect and visualize performance metrics.

Background:

The project will initially be hosted on a virtual machine provided by the CS department. Initial metrics may be collected for the CS GPU server, babbage.cs.umb.edu, which includes a GeForce RTX 4080 GPU (part of the nfs/ldap environment).

The project will use Grafana to visualize GPU and other metrics exported from the Chimera cluster. The exporter will be determined by the team during the project, most likely either Prometheus or Telegraph. The project will first look at standard metrics such as: CPU usage, memory usage, system load, GPU memory usage, GPU utilization (for non-MIG GPUs), GPU power consumption, and GPU temperatures.

Then the team will run GPUprograms on the test system and collect individual run time metrics on the test system and/or collect and display real time metrics from the production system (Chimera).

As part of annual updates to the CS website, the team will create an article to post regarding Chimera availability and usage. The team will also document usage of the new metrics web application.

First steps:

Get the VM set up on the CS server and access it
Get Grafana metrics working for babbage or an older GPU
Create the DB to store the metrics
Import the metrics exported from Chimera
Test the data collection from Chimera
Test the DB storage of the metrics

Additional Features:

Next steps for a cluster is to implement the ability to drill down – see a full cluster view, select individual nodes, select individual GPUs
Also, to drill down into different time windows.