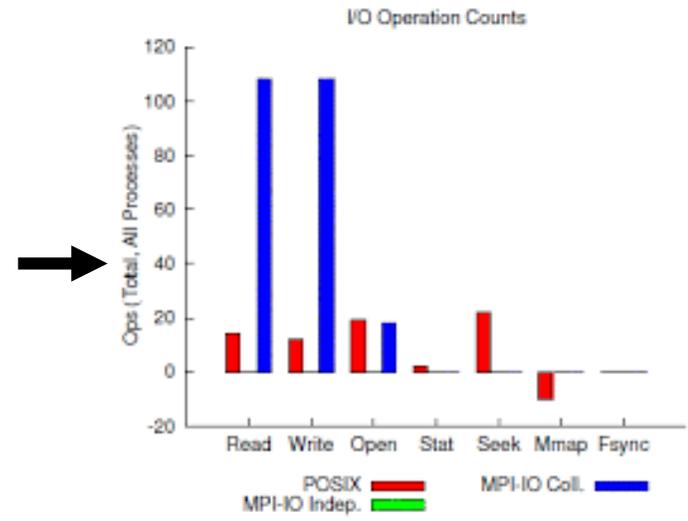
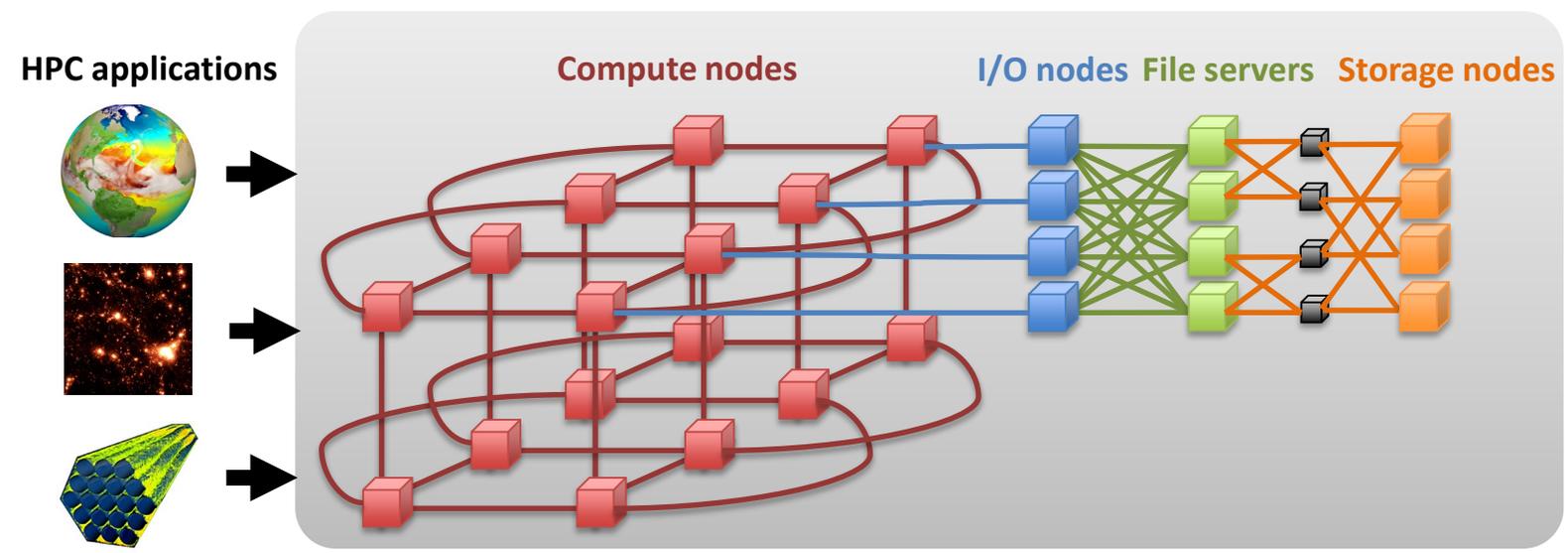


Gauge: An Interactive Data-Driven Visualization Tool for HPC Application I/O Performance Analysis

Eliakin del Rosario, Mikaela Currier,
Mihailo Isakov, Michel A. Kinsy
Adaptive and Secure Computing
Systems (ASCS) Laboratory
Texas A&M University

Sandeep Madireddy, Prasanna Balaprakash,
Philip Carns, Robert B. Ross
Mathematics and Computer
Science Division
Argonne National Laboratory

Manually Analyzing HPC Jobs Is Inefficient

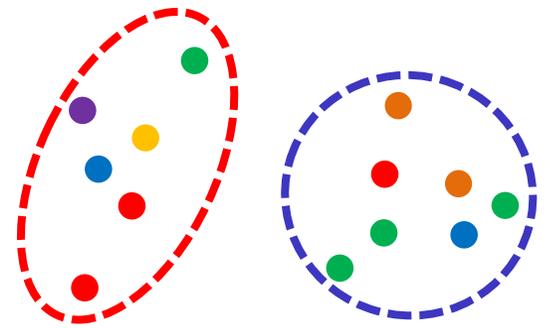


Darshan
 LMT, LDMS
 LDMS, UFM
 LMT, mmpfmon, DDNtool

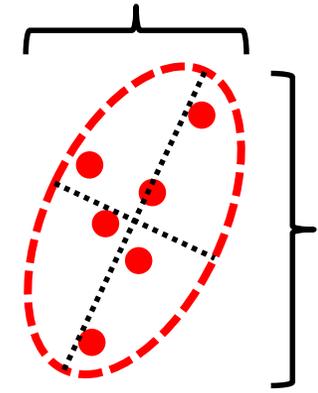
- HPC jobs are not organized by similarity, it's hard to navigate them
- Hard to derive insight from a bulk of job logs
- Effort spent on analyzing a specific job does not speed up future analysis

Scaling Analysis Through Grouping HPC Jobs

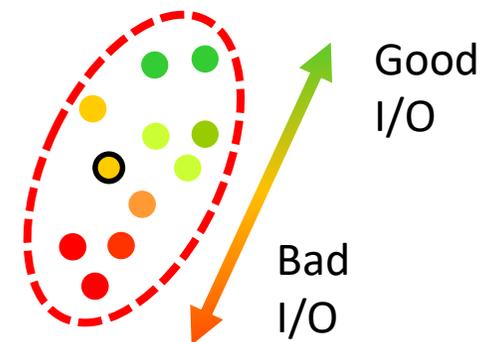
- HPC I/O experts can provide deep insight on a specific job, but it's hard to reuse their effort between jobs or users
- Researchers may benefit from comparing their jobs against similar runs, but how to find them?
- There is insight about workloads or the system that can only be gained by observing jobs in bulk



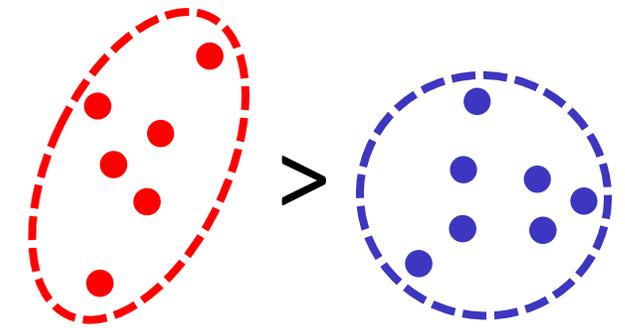
How can we group jobs together?



What are the key characteristics of the group itself?



How does this job's performance rank with the rest of the group?

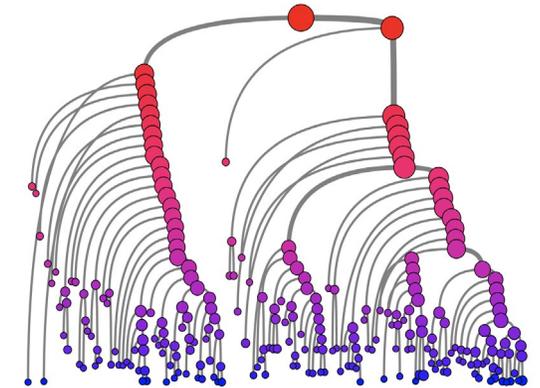


How does the group compare with other groups?

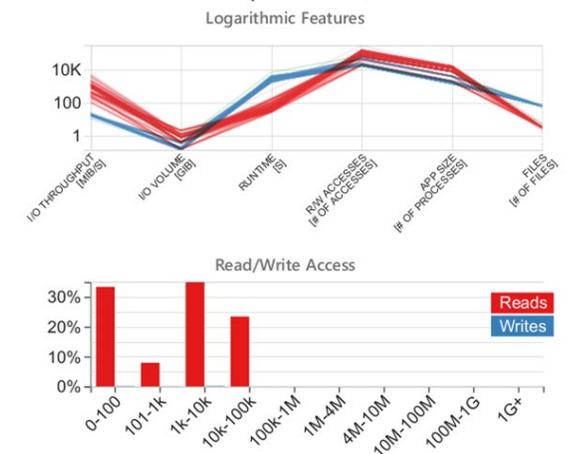
Gauge: HPC I/O Visualization Tool

- Gauge is a web-based, data-driven, highly interactive exploration and visualization tool for diagnosing HPC I/O behaviors
- Gauge analyzes HPC I/O logs, groups / clusters similar jobs together, and creates a cluster hierarchy of jobs running on the system
 - Gauge allows I/O experts and facility operators to better scale their efforts when analyzing HPC jobs
 - I/O experts can analyze groups of similar jobs to find patterns not detectable when analyzing single runs
 - The hierarchy helps facility staff better understand the workloads running on their systems
 - Gauge allows researchers to find jobs that look similar to theirs, which might help to optimize their jobs or better understand I/O bottlenecks
- Gauge provides cluster-level visualizations, and lets users visualize clusters at the 'right granularity'

Sample Gauge Hierarchy

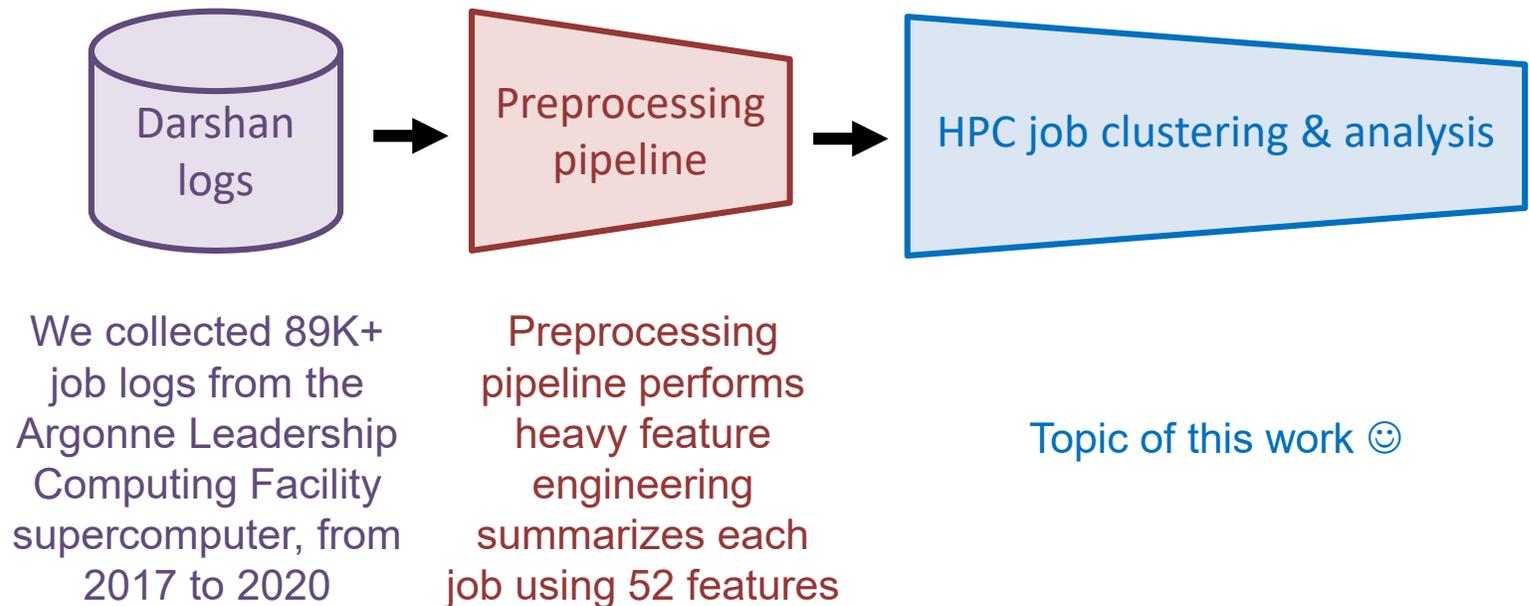


Sample Gauge Cluster Visualization



Preliminaries: Logging and Data Pipeline

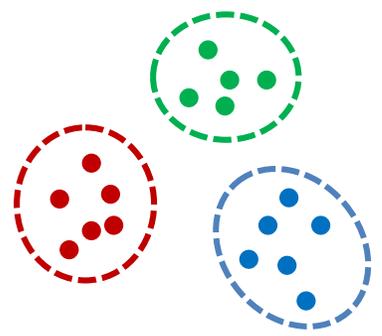
- Gauge is analyzes HPC I/O logs, with one log per job
- Dataset used in this work consists of 89K jobs, with each job described by 52 features
- Techniques described here are generally applicable, and are not necessarily tied to the logging tools used



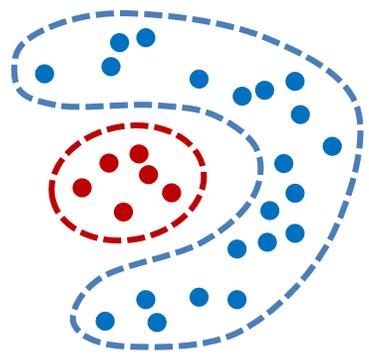
For more details on the data preprocessing and clustering, see our **SC20** paper “HPC I/O Throughput Bottleneck Analysis with Explainable Local Models”

Difficulties of Clustering HPC Jobs

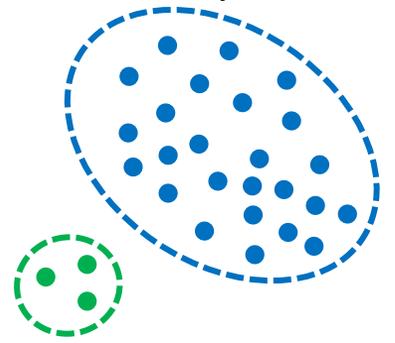
We don't know what to expect from clustering:



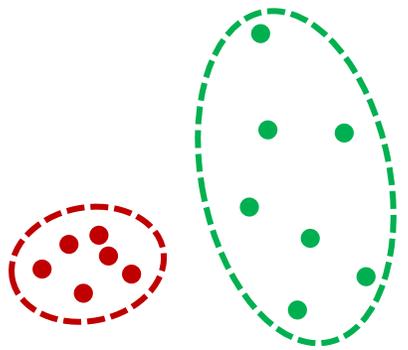
How many clusters?



What shape?

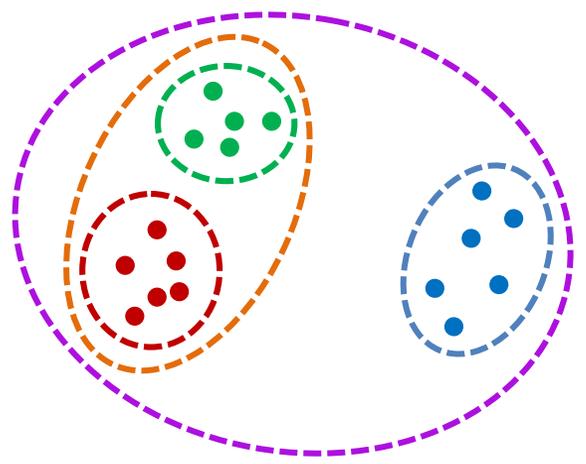


What size?

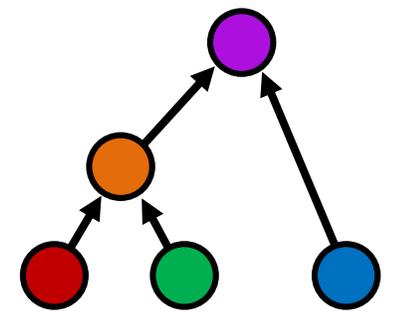


What density?

HPC jobs:



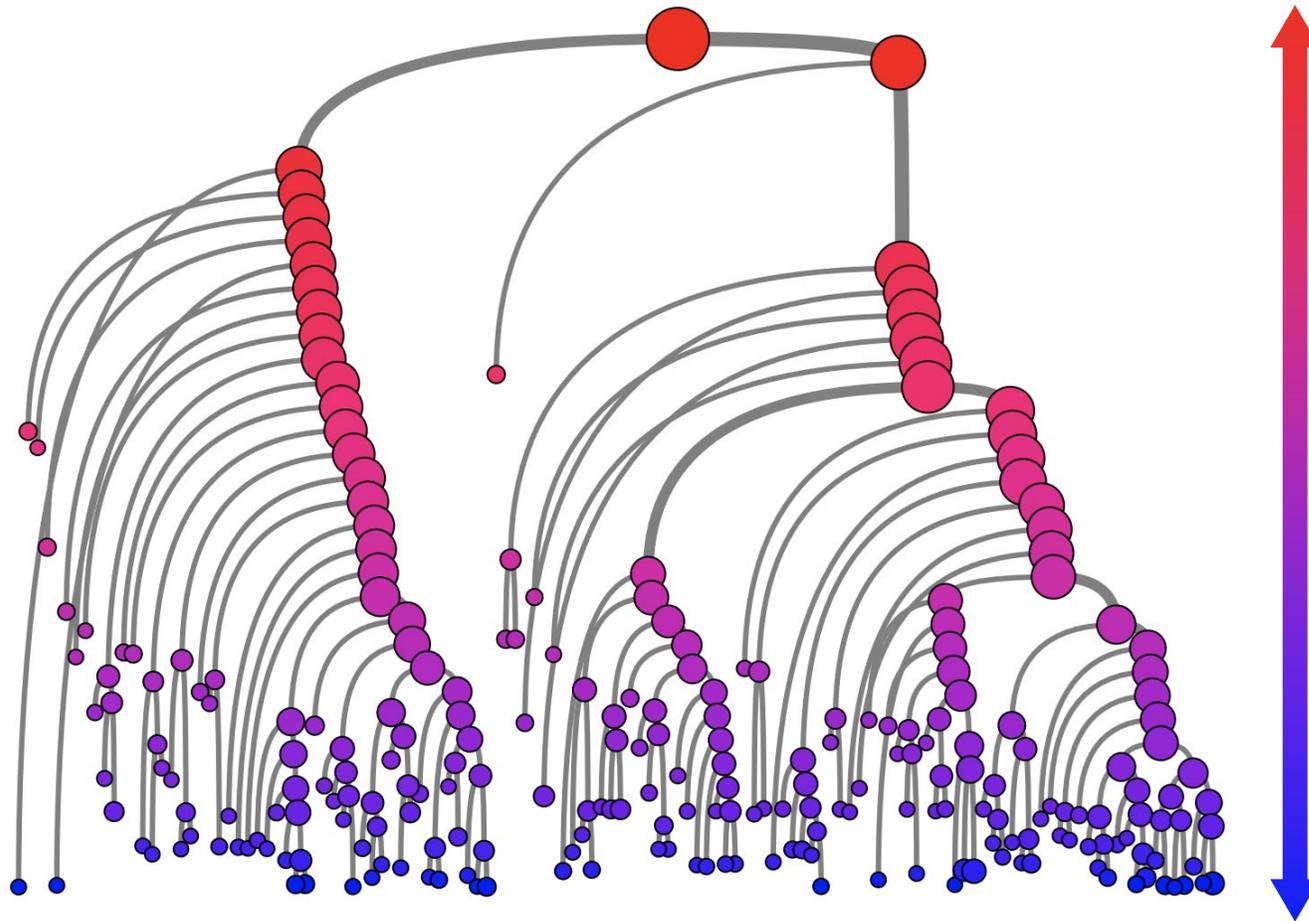
Clusters of HPC jobs:



- To build a hierarchy of clusters we use HDBSCAN - a hierarchical agglomerative clustering algorithm
- HDBSCAN is robust to varying numbers of clusters, cluster sizes, densities, and shapes
- Hierarchy helps to explore & select right granularity

Gauge Hierarchy

- Each node is a group of jobs
- Branches show which clusters merge together
- Node size shows # of jobs in cluster
- User can select a node to bring up cluster information



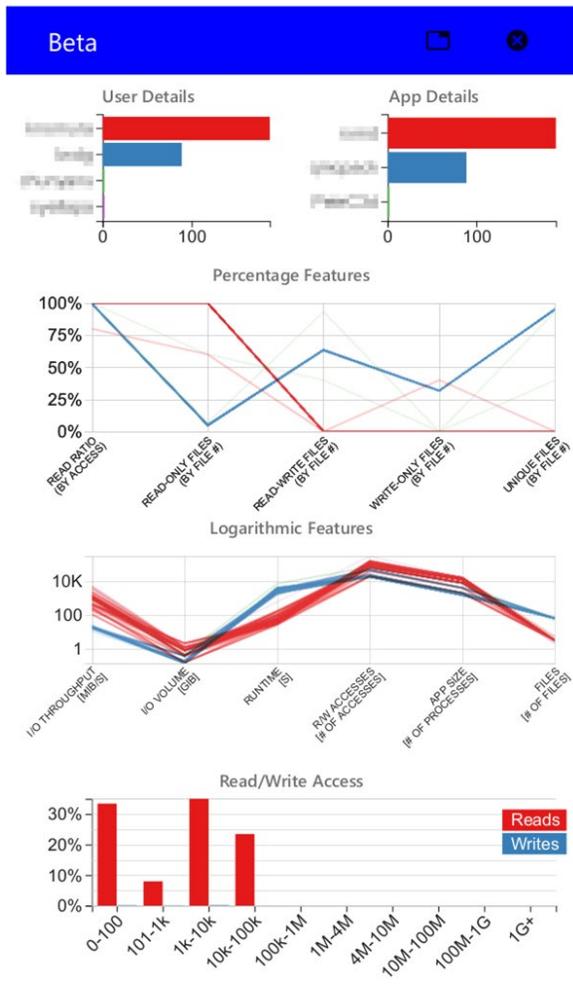
Larger, sparser
clusters

Node height
shows cluster
density

Smaller, denser
clusters

Gauge Per-Cluster Visualization

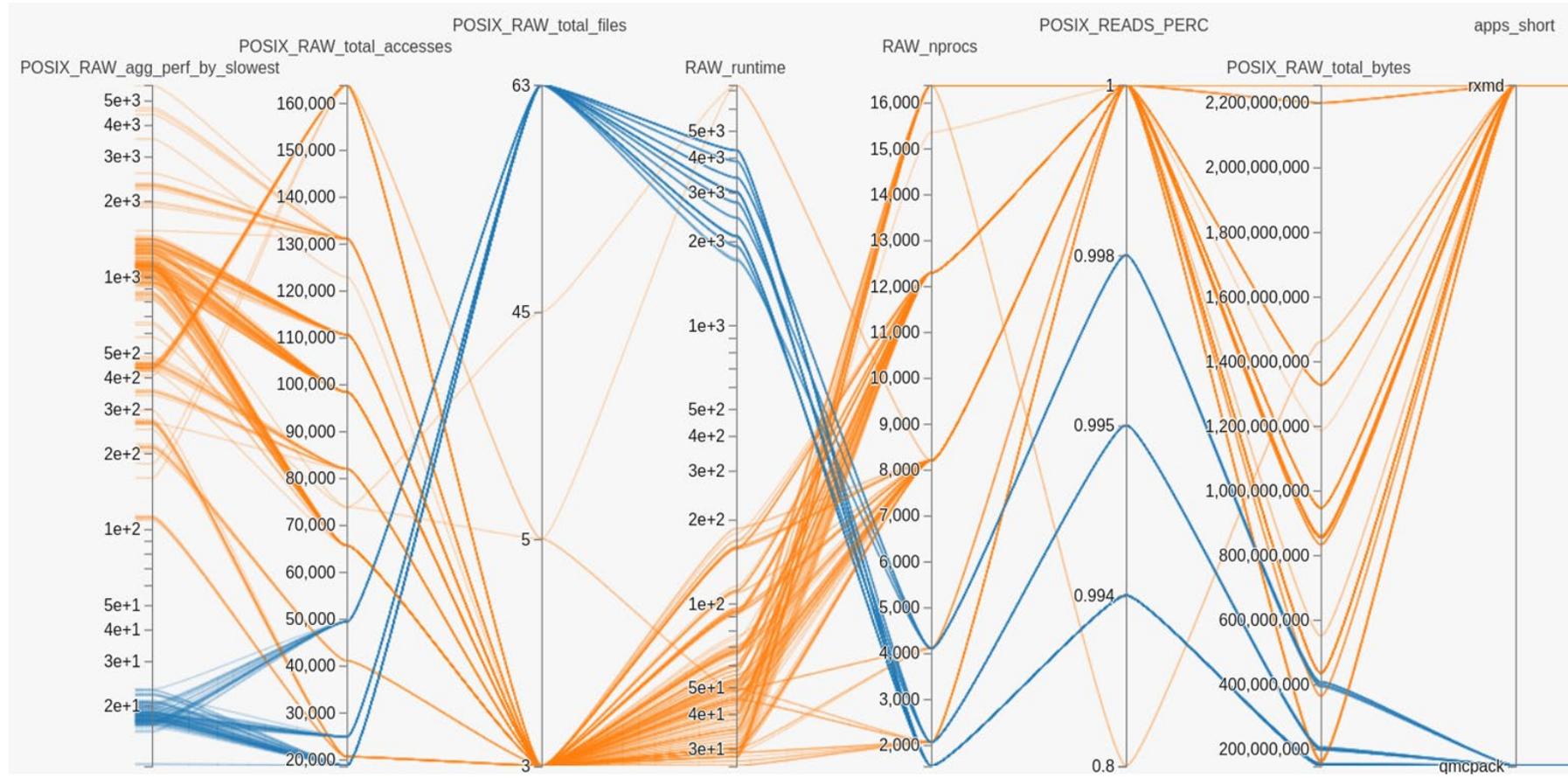
- When the user clicks on a node, a new cluster column opens up (right)
- 5 different graphs show most important info about a cluster
- Graphs are interactive, user can set color-by-user or color-by-application
- User can open up a full-page parallel coordinates plot



- Cluster name & options
- User and application details: breakdown of users and applications
- Percentage features: job features represented as a ratio in 0-100 %
- Absolute features: job features that don't have a known range
- Breakdown of accesses by access size and read / write properties

Gauge Cluster Parallel Coordinates Plot

- Gauge offers a full page parallel coordinates plot
- Each broken line is a specific job
- Each column is a feature. 50+ features to select from!
- Keep or exclude any range of jobs

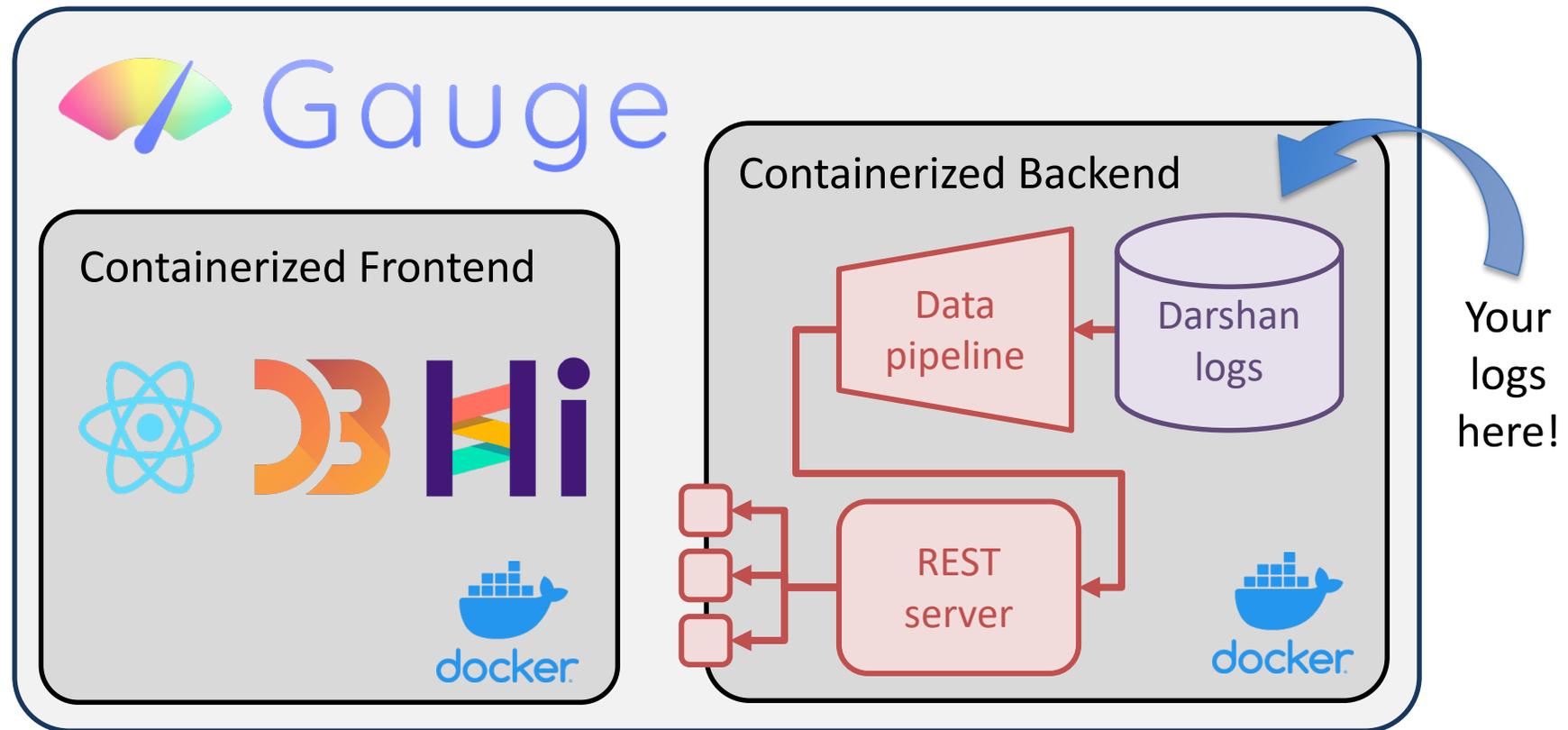


Case Study

- Please watch our video presentation for a viewing of our case study.

Running Your Own Gauge Instance

- A Gauge instance visualizes a single HPC system
- Built with extensibility in mind – easy to add new visualizations
- Simple setup on new systems, just add your logs and spin up docker containers
- **Contact us!**



Conclusions

- Gauge presents a new method for grouping and visualizing HPC data
 - While first developed for the HPC I/O domain, can be used on system data in general
- With Gauge, facility experts can more easily analyze logs in bulk
 - Useful for diagnosing a problematic application or simply exploring workloads running on the system
- Researchers can use Gauge to view their past runs
 - Useful for better understanding an application's I/O behavior, what researchers can do to improve, or how they rank among their peers
- Gauge is open-source and simple to deploy
 - Contact us for help in applying it to new systems and domains!