An Operations Monitoring and Notification Infrastructure (OMNI) for Exascale Data Centers

Melissa Romanus[1,2], Elizabeth Bautista[1], Thomas Davis[1], Cary Whitney[1]
[1] National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[2] Rutgers University, Piscataway, NJ 08854

Introduction

- Operational data (OD) from data centers is useful in monitoring and alerting, tracking down issues, building machine learning models, engaging users, mitigating risk, automating tasks, and more.
- Extreme-scale HPC machines generate correspondingly large amounts of OD.
- As systems continue to grow in scale and complexity, new innovations are required for collecting OD and transforming it into actionable insights.
- This work details the creation of the **Operations Monitoring and Notification Infrastructure (OMNI)** for collecting and managing Big OD across the National Energy Research Scientific Computing Center (NERSC) at the Lawrence Berkeley National Laboratory in Berkeley, California, USA.
  - OMNI is capable of meeting the unique Big Data needs of HPC data centers.
  - It provides a centralized repository for data collection and a common access point for users to analyze OD that initiated anywhere in the center.

NERSC HPC Data Center Operations

- NERSC is the primary scientific computing facility of the US Department of Energy Office of Science and is home to two Cray HPC machines, Edison (#124 Top500 November 2018) and Cori (#12 Top500 November 2018).
  - Upcoming system Perlmutter, expected 2020, will have 3x compute power of Cori
- It is also an extremely energy-efficient data center, requiring no mechanical chillers. Instead, it leverages the temperate climate of Berkeley, together with evaporative cooling, to cool the facility, office spaces, and compute infrastructure.
  - Average Monthly Power Usage Efficiency: 1.07

Table 1, Examples of Operational Data at NERSC

Machine Metrics	Network Counters and Flow	Particle Sensors
Job Scheduling	Application Characteristics	Power Usage & Distribution
Node Status	Filesystem I/O	CPU Usage & Temperature
Air Circulation	Outdoor Weather Station	Syslogs
Water Tower Fans & Pump Status	Cooled Water Temperature & Pressure	Room, Rack, and Cabinet Temperature & Humidity

- **Characteristics and Challenges of OD:** *real-time, continuous, streaming, heterogeneous, distributed, time series—based, 24x7, high volume, high cardinality*

OMNI Data Collection Architecture

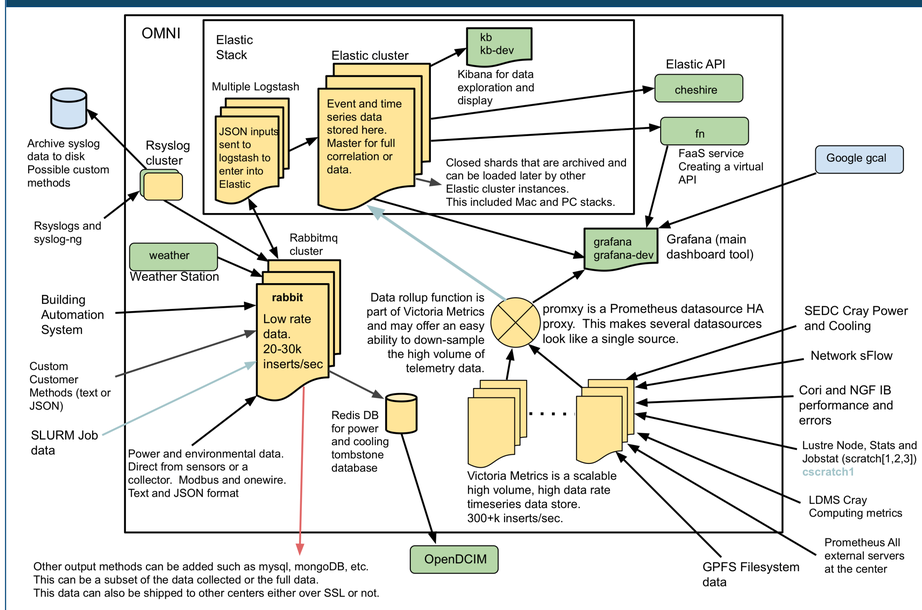

Figure 1, Data Collection Architecture. Sources include weather station, water towers, substation power, ESnet, building management systems, internal power at breakers/ PDUs/ UPS, Cray system environmental & power data, job characteristics, Lustre filesystem, burst buffer, InfiniBand high-speed network, particle counters, earthquake sensors, and more.

Legend: Big box is OMNI services. Yellow internal OMNI services. Green are User access points. Blue are external systems. Black lines are existing data sources. Blue lines are work in progress data sources. Red lines are possible data paths.

Built with Open-Source Technology, Data Exchanged in JSON
Ingestion, Querying, Indexing: RabbitMQ, Logstash & Elasticsearch
Exposed via Frontend Tools: [Web]: Kibana, Grafana [Functions as a Service (FaaS)]: fn
Extensible to More Advanced Analytics Tools: APIs for programming languages like python, PHP, Perl, Ruby, Java, JavaScript, and more

Current Ingest Rate: 25K New Records/Sec.
Online, Searchable Data: 125TB - 500 Bil. Records 2+ years of historical data Query entire dataset in milliseconds to seconds
Current Max Capacity: 172TB (upgradable)

Case 1: Detecting Voltage Fluctuation at New Data Center

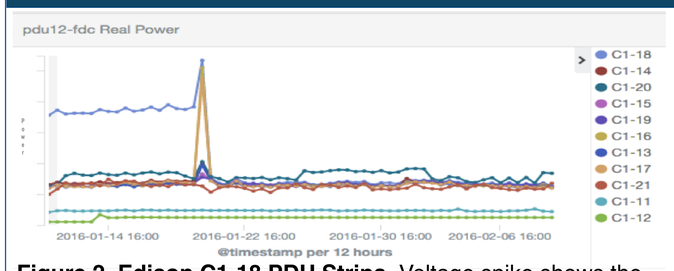

Figure 2, Edison C1-18 PDU Strips. Voltage spike shows the system response to large job idling at the PDU-level.

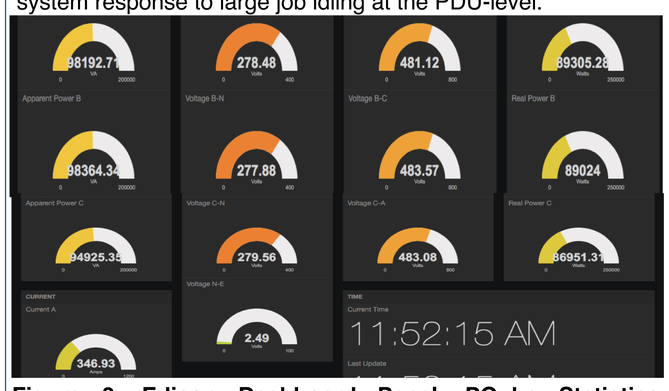

Figure 3, Edison Dashboard Panel, PQube Statistics. Thresholds give quick visual indicators of voltage issues.

- In December 2015, the NERSC data center relocated from Oakland, CA to Berkeley, CA.
- Shortly after the move, in January 2016, multiple cabinets of the Edison HPC system began unexpectedly powering off without warning during the cleanup phase after large jobs would finish.
- Metrics were unavailable from the nodes and CPUs directly but OMNI's high-availability data collection allowed NERSC engineers to examine the events surrounding the failure.
- Incoming power supplied to the new data center was found to be 12 kilovolts, 10% above expected.
- Building designers had used normal office building specifications rather than those of a data center in their designs and failed to account for the large-scale power fluctuations that can occur in HPC systems.
- Until the substation's output was shunted, large jobs could not be run on Edison. A shunt is used to control the high-voltage that can occur when there is a sudden loss of power demand.

Case 2: Facility Planning for Next-Generation System

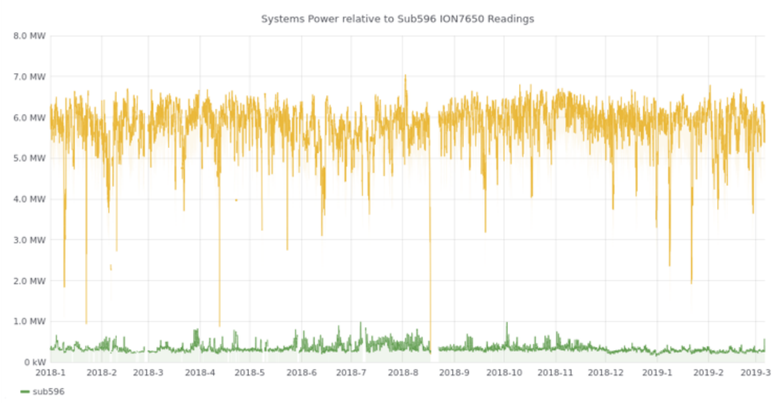

Figure 4, Total Power Load from the Compute Substations vs. Mechanical Power (MW). At peak, the mechanical substation uses only ~1MW of power.

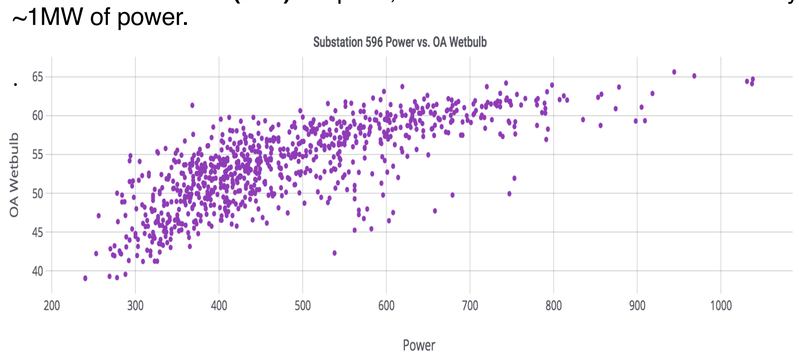

Figure 5, Substation 596 Power v. Outside Air Wetbulb. In warmer temperatures with more humidity, the evaporative cooling functionality activates and draws more power.

- When planning facility upgrades for the upcoming NERSC Perlmutter system, the LBNL Project Planning code dictated that the need for a new substation to power the mechanical upgrades (evaporative cooling air handling units, chilled water, etc.) should be computed by either (1) summing the peak power usage of each device (as specified by the manufacturer) that will be powered by the substation **or** (2) using at least one year's worth of OD.
- The calculation from Method (1) showed that another mechanical substation would be needed at a cost of $2.5 million.
- However, OMNI already contained over 2 years worth of online OD of the facility when two HPC systems, Cori & Edison, were both fully operating at the NERSC data center.
- Using Method (2), engineers showed via ODA that the actual demand on the mechanical substation was much lower than the peak usage rating of each device and proved that NERSC's existing mechanical substation could handle the new load.
- ***Resultant Savings: $2.5 million***

Case 3: Arc Flash Event

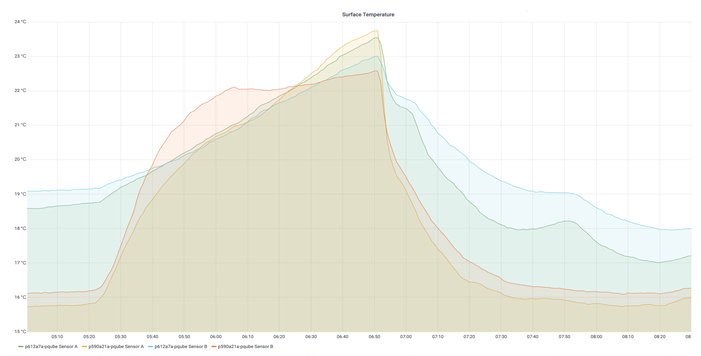

Figure 6, PQube Sensor Surface Temperature Immediately after Arc Flash Event. A steep rise in temperature at ~5:23AM demonstrates its severity and impact on surrounding infrastructure.

- On December 31, 2018, the Cori supercomputer experienced an arc flash due to a damaged bus bar, triggering a fire alarm.
- During the walkthrough with the fire department to clear the facility, site-reliability engineers (SREs) felt that the machine room floor was warmer than normal.
- Using OMNI data to monitor all systems, an SRE quickly identified that the air handling vents had been closed in response to the fire alarm and took action to reopen them.
- Being able to quickly identify the root cause of an observation minimized the impact of the event to other NERSC infrastructure.

Case 4: Vendor Collaborations

- Collecting data from the NERSC supercomputers into OMNI has led to a mutually beneficial relationship with the Cray HPC vendor. NERSC's Cori, Edison, and upcoming Perlmutter machines are all Cray systems.
- Historically, internal temperature sensors and power readings of the HPC systems were collected by Cray and not exposed to customers. To ingest the data into OMNI, NERSC engineers worked with Cray to write an API that allows customers to access and export the power and systems environmental data. This API was then released to the community expanding the benefit of it beyond NERSC to all Cray customers.
- Cray keeps machine data on the order of months. When Cray identified a bug in the system, they asked NERSC for 1-year of OMNI data to examine the behavior and identify the root cause and timeframe.

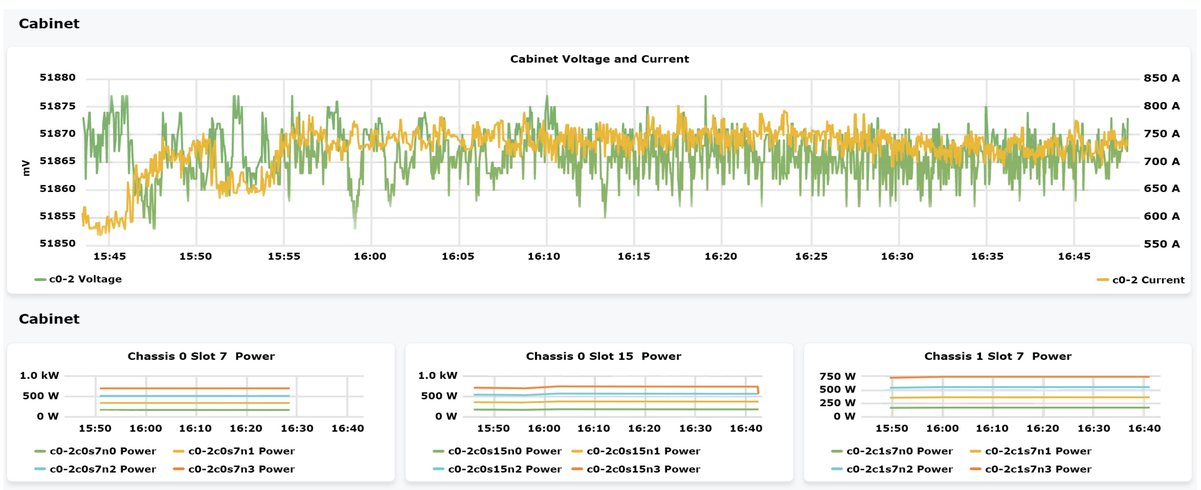

Figure 7, Partial Cori power dashboard. OMNI data obtained through the new Cray PMDB API (from collaboration).

Lessons Learned

- ***Operational Data >> Environmental Data Alone.*** Collecting machine, network, and filesystem logs and metrics together with sensor and environmental data provides a holistic view of the data center, thereby enabling research opportunities that span environmental & machine—level questions and allowing us to identify root causes faster in the event of a bug or system issue.
- ***Collect Everything & Save It Forever.*** From years of operational experience, NERSC engineers know firsthand that questions or bugs are often discovered over time – having historical OD to analyze or refer to when this occurs would aid in answering those questions or discovering when a problem was introduced. Thus, OMNI's goal is to have the data that can provide answers for questions that are not yet known.
- ***Data Stewardship of Centralized Data.*** Collecting data from different teams at NERSC into a centralized location created greater sharing and collaboration across the organization. However, it also led to an assumption that the OMNI team would be responsible for data analysis across datasets or for tracking down issues with data belonging to other teams. Creating clear data stewardship policies for new datasets would have helped to outline the responsibilities of a team or individual in regards to the data they are contributing (including data analysis efforts) and provided a point-of-contact when necessary, while still allowing for open collaboration and sharing across the organization.
- ***Growth & Scalability.*** When OMNI was architected, no other HPC data centers were collecting data at the size, scale, and rate that NERSC was targeting. Choosing open source software allowed the OMNI team to design a system that fit their unique needs without having to wait for a licensed company to design one for them or solve issues on their behalf. In addition, it enabled upfront costs to be primarily allocated to hardware, which was of key importance at such scales.

This work used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

Scan Here to keep up with the Latest OMNI News at NERSC!