Collecting, Monitoring, and Analyzing Facility and Systems Data at the National Energy Research Scientific Computing Center

Elizabeth Bautista
ejbautista@lbl.gov
Lawrence Berkeley National Laboratory

Melissa Romanus
mromanus@lbl.gov
Rutgers University
Lawrence Berkeley National Laboratory

Thomas Davis
tadavis@lbl.gov
Lawrence Berkeley National Laboratory

Cary Whitney
clwhitney@lbl.gov
Lawrence Berkeley National Laboratory

Theodore Kubaska
tedkubaska@comcast.net
Energy Efficiency HPC Working Group (EE HPC WG)

ABSTRACT

As high-performance computing (HPC) resources continue to grow in size and complexity, so too does the volume and velocity of the operational data that is associated with them. At such scales, new mechanisms and technologies are required to continuously gather, store, and analyze this data in near-real time from heterogeneous and distributed sources without impacting the underlying data center operations or HPC resource utilization. In this paper, we describe our experiences in designing and implementing an infrastructure for extreme-scale operational data collection, known as the Operations Monitoring and Notification Infrastructure (OMNI) at the National Energy Research Scientific Computing (NERSC) center at Lawrence Berkeley National Laboratory. OMNI currently holds over 522 billion records of online operational data (totaling over 125TB) and can ingest new data points at an average rate of 25,000 data points per second. Using OMNI as a central repository, facilities and environmental data can be seamlessly integrated and correlated with machine metrics, job scheduler information, network errors, and more, providing a holistic view of data center operations. To demonstrate the value of real-time operational data collection, we present a number of real-world case studies for which having OMNI data readily available led to key operational insights at NERSC. The case results include a reduction in the downtime of an HPC system during a facility transition, as well as a $2.5 million electrical substation savings for the next-generation Perlmutter HPC system.

CCS CONCEPTS

• **Applied computing → Enterprise data management**; **Data centers**; • **Mathematics of computing → Time series analysis**; *Exploratory data analysis*; • **Hardware → Power and energy**; *Enterprise level and data centers power issues*; • **Information systems →** *Business intelligence*; *Data analytics*.

KEYWORDS

data centers, operations, monitoring, high-performance computing, data collection, operational data analytics, time series data, Green HPC

ACM Reference Format:
Elizabeth Bautista, Melissa Romanus, Thomas Davis, Cary Whitney, and Theodore Kubaska. 2010. Collecting, Monitoring, and Analyzing Facility and Systems Data at the National Energy Research Scientific Computing Center. In *EE HPC SOP 2019: Energy Efficient HPC State of the Practice Workshop, August 05–08, 2019, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

EE HPC SOP 2019, August 05–08, 2019, Kyoto, Japan

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

High-performance computing (HPC) systems that support a range of large-scale scientific applications are continuing to grow in size and complexity on the path to exascale. Operating these machines requires a data center capable of meeting the power, space, infrastructure, and cooling requirements that they demand. Given the complexity and scale of these systems, a number of unique challenges exist in managing HPC data centers, such as high power usage with large fluctuations, providing high-availability and high utilization for users over long-running jobs despite failures, and extensive cooling requirements involving both air and water.

Achieving operational efficiency in this type of environment requires gathering information from all the systems and sources that support the HPC data center, analyzing it, and responding to near-real time events when necessary. However, the nature of this data is heterogeneous, coming from diverse sources, and distributed, as these sources can be in different formats, located across the machine, data center, or external to the facility. The scale of resources in an HPC data center also means that the amount of data that must be collected is proportionally large. The size of the datasets depend on the time-variant, as operational changes can occur at micro- or even nano-second scales. They are also dependent on their rates of collection, resolution, indexing, and availability.

However, the nature of this data is heterogeneous, coming from diverse sources located across the machine, data center, or external to the facility and in different formats. The scale of resources in

an HPC data center also means that the amount of data that must be collected is proportionally large. Further, the datasets are time-variant due to their rates of collection, resolution, indexing, and availability. Some may occur at micro or nano-second intervals while others can be in seconds, minutes, or more. Some of the streaming data needs to be captured and exposed to operations staff in near-real time for correlation. In addition, archived operational data can continue to be useful for data scientists and researchers in identifying historical trends that may help inform decisions about energy efficiency, future procurements, proactive maintenance, or building models for predictive machine learning applications.

Providing the means to ingest and expose this data requires an integrated operational data collection and analytics infrastructure capable of overcoming these challenges while minimizing the impact on the systems themselves and meeting the operational goals of that facility or organization. This paper provides the experiences and lessons learned in creating the Operations Monitoring and Notification Infrastructure (OMNI) for this purpose at the National Energy Research Scientific Computing Center (NERSC), located at Lawrence Berkeley National Laboratory (LBNL, hereafter referred to as Berkeley Lab) in Berkeley, California. OMNI ingests streaming time series data from a variety of sources including the HPC systems at NERSC, other supporting computational infrastructure, environmental sensors, mechanical systems, and more in near-real time. OMNI is built using open-source technologies, such as the Elastic Stack, and currently contains over two years of online operational data, totaling 550 billion records (125 TB of data). The rest of this paper is structured as follows. Section 2 discusses the operational details of the NERSC data center and the building, as well as the potential sources of data in a data center. The design rationale for an HPC facility integrated operational data collection and analytics infrastructure and implementation details of OMNI at NERSC are provided in Section 3. Section 4 provides scenarios of insights gained thus far from OMNI data analysis and Section 5 discusses the lessons learned in creating and implementing OMNI. Section 6 provides future directions for OMNI and a brief conclusion.

2 BACKGROUND

NERSC is the mission scientific computational facility for the Office of Science in the U.S. Department of Energy (DOE) and has operated many high-performance computing systems since its inception at Lawrence Livermore National Laboratory in 1974. Sixteen NERSC systems have appeared on the Top500 [1] list of fastest computing systems in the world. NERSC's mission is to provide HPC and compute resources to science users at high-availability with high-utilization of the machines in order to further the scientific research supported by the DOE office of Science.

The current NERSC HPC data center is located at Shyh Wang Hall. The building is a 140,000 gross-square-foot (GSF) facility that houses both the data center as well as office spaces for Berkeley Lab Computing Sciences division employees spanning NERSC, the Energy Sciences network (ESnet), and the Computational Research Division (CRD). It is comprised of 4 floors – 2 office floors (28,000 square feet each), 1 machine room floor (20,000 square feet with room to expand up to 28,000 square feet), and 1 mechanical level. It is outfitted with a seismic sub-floor and is a LEED®-certified Gold facility, averaging a monthly Level 2 Power Usage Effectiveness (PUE) [5] ratio of 1.07 over the past year.

NERSC currently supports two high-performance systems in the Shyh Wang Hall data center. The first, Edison, is a Cray XC30 machine and has a peak performance of 2.57 petaflops per second, with 134,064 compute cores, 357 terabytes of memory, and 7.56 petabytes of disk[1]. The second is Cori, a Cray XC40 machine, equipped with 2,388 Intel Xeon "Haswell" processor nodes (32 cores each), 9,688 Intel Xeon Phi "Knight's Landing" (KNL) nodes (68 cores each), and a large all-flash burst buffer.

To support these systems, Shyh Wang Hall has an available power capacity of 12.5 megawatts.[2] The maximum possible power capacity for the building, with upgrades, is 42 megawatts. Together, the systems draw an average monthly energy usage of 4.8×10^6 kilowatt-hours.

To maintain energy efficiency and reduce environmental impact, the facility leverages the temperate Berkeley, California climate to cool the data center. The facility does not have traditional chillers (i.e., "air conditioning units") and instead uses a combination of cooled water and evaporative cooling to maintain the data center environment. Air that has been circulated through the data center is mixed with the outside air in order to temper and dehumidify it, as well as to warm the office floors. When additional cooling and dehumidifying is required, cooled water is used to bring down the air temperature prior to it entering the evaporative unit.

In addition to its use in the air handling units, cooled water is also used by the high-performance computing systems to moderate temperature. Water circulates through cooling towers, where it is cooled by evaporation. This outside, open water loop is connected by a heat exchanger to a closed, inside water loop that provides cool water directly to the systems. The water loop additionally provides cooling for air on hot days.

3 OMNI INTEGRATED OPERATIONAL DATA COLLECTION AND ANALYTICS

This section describes the design and implementation of OMNI, an integrated operational data collection and analytics infrastructure that gathers data from a variety of operational and facilities sources across a data center.

3.1 Design

Operational data, especially at the scale of HPC data centers, is large, heterogeneous, and distributed. Time is also an important characteristic of operational data, as changes in the compute environment can occur at nano- and micro-second scales. Examples of operational data include time series data from the environment (e.g., temperature, power, humidity levels, and particle levels), monitoring data (e.g., network speeds, latency, packet loss, utilization or those that monitor the filesystem for disk write speeds, I/O, CRC errors), and event data (e.g., system logs, console logs, hardware failure events, power events essentially anything that has a start and end time). The reporting rate of this data often depends on

[1]Edison will be retired on May 13, 2019
[2]The Center is in the process of being upgraded to 25MW for the upcoming Perlmutter system

several factors including individual properties of the sensor or machine, the size of the data, whether or not continuous monitoring is necessary, and how quickly it is needed for analysis. Some systems do not report data by default and must be instrumented by system administrators.

Given the complex nature of the data, creating a system for collecting it in a production environment is challenging. Based on the data properties and the sources from which it can be collected, the OMNI team identified the core system requirements, as follows:

Scalability. One of the primary concerns with building a data infrastructure for operations and monitoring is the volume of data that is being collected and the need to provide near-real time insights into the systems and sensor data. This is especially true for HPC data centers, where there are thousands of nodes and millions of metrics to be reported to determine the health of systems and facility. In addition, it is critical to minimize the overhead created by sending operational data from multiple sources across the network on any of the data center operations. Thus, designing a data collection infrastructure capable of ingesting new data sources and dynamically scaling to meet the needs of emerging exascale infrastructures is of utmost importance.

High-Availability. Most data centers operate in a 24x7 environment and consequently, their operational data is a continuous monitoring process. The data collection infrastructure is a critical part of operations & monitoring that must be available at all times, even in the presence of other faults or issues at the data center.

Maintainability. Software and hardware evolves and changes over time. The system maintainers must be able to apply rolling patches, upgrades, warm hardware swaps, etc. to parts of the system without affecting the flow of data from the various sources.

Usability. Exposing the data to be utilized by a variety of stakeholders (e.g., site-reliability engineers, consultants, system administrators, researchers) is equally as important as collecting it. Policies for anonymizing data and controlling access to it are important. Providing tools to access the data furthers the possibilities of what can be gained from the data that is collected. The system must provide fast and easy access to the data that it collects for analytics, visualization, and monitoring purposes.

Lifetime Data Retention Policy. Traditional research has focused on minimizing data collection by only collecting data when there is a problem or open research question. However, there are times when a bug or an issue is identified later in time, where historical operational data would be a valuable tool for providing insights. In addition, machine learning models can be trained with historical data. The nature of operations means that virtually anything can happen. The system must collect the data with the goal of storing and saving it forever, thereby providing an asset that can be consulted when new questions arise and a source for statistical modeling and failure prediction.

3.2 Implementation Details

The OMNI cluster is independent of any system in the facility; it is the first system to become available after the power is turned on and the last system to be taken down before the power is turned off.

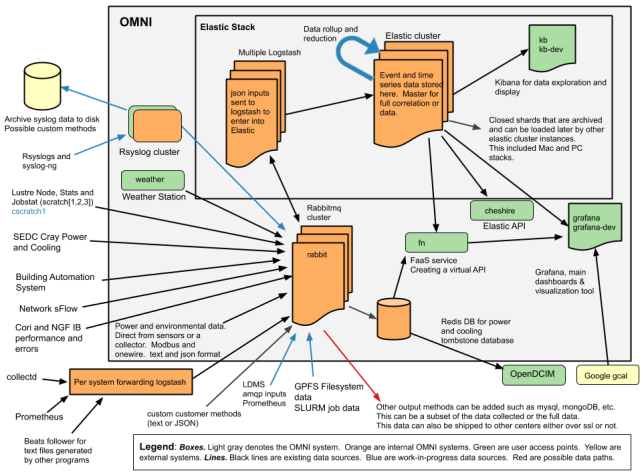

Figure 1: OMNI Integrated Operational Data Collection and Analytics Architecture.

As long as there is power to the facility, OMNI stays on to collect data.

OMNI is implemented using open source software, on-premise hardware, and virtualization technologies. The decision to use open source software avoids vendor lock-in and reduces the cost of the data collection infrastructure for the center. The use of virtual machines and containers in OMNI enables more efficient use of the underlying hardware, facilitates on-demand application provisioning, lowers the cost of hardware maintenance, and allows for high-availability configurations. Accordingly, the use of virtualization for operational data collection leads to lower overall power consumption and cooling requirements, compared to using a bare metal solution alone.

The specific software and technologies used by OMNI were selected based on their abilities to meet the design requirements set forth in Section 3.1. Virtualization is configured and managed using oVirt [8] and Rancher [10]. Data ingestion and storage in near-realtime is enabled via the Elastic Stack [2], an open source distributed search and analytics software stack comprised of different components for ingesting, querying, and visualizing data. The central component of the stack is Elasticsearch [3], a distributed JSON-based RESTful search engine that enables the ingestion and searching of massive amounts of data in near real-time. The Logstash [7] component handles server-side data processing pipelines, i.e., parsing streaming machine log information for pertinent values and forwarding them to Elasticsearch for ingestion. The Kibana [6] component provides a web interface for data discovery, analysis, and visualization of Elasticsearch data, as well as monitoring information and management controls for the Elastic Stack. Additional stack components are available but are not used in the OMNI system. The Elastic Stack is free to use but certain features of it, such as X-Pack security and cross-cluster search, are enabled via license. Since cost and impact to normal NERSC operations is an important consideration in this implementation, OMNI utilizes only the free version.

Figure 7 shows the OMNI data collection architecture and its diverse data sources. The data sources from the NERSC data center include external systems and sensors, such as meters at the electrical substations, information from the water tower that supplies the building's water, and weather and atmospheric data about the air and surrounding Berkeley climate. From the facilities perspective, sensors and metrics inside the data center include building management systems (BACnet, Modbus), i.e., cooled water, air handling units, particle counters, temperatures from rack doors, and earthquake sensors, as well as power readings at the breaker panels, power distribution units (PDUs), and Uninterruptible Power Supplies (UPSs). Metrics from the high-performance computing systems include Cray Power Management Database (PMDB) and System Environment Data Collections (SEDC) data, job information from the Slurm job scheduler, Lustre parallel file system data, and information from the Aries high-speed network. For Cori, there is additional information available for the burst buffer. Other network information from the data center and the Energy Sciences Network (ESnet), DOE's dedicated science network is captured via sFlow, SNMP, and InfiniBand data. In addition, OMNI also captures `syslog` information.

Getting data from the various systems and sensors into Elasticsearch occurs via RabbitMQ [9]. RabbitMQ is a messaging broker that supports multiple messaging protocols and queuing. It can be set up as a high-availability message queue and is capable of writing a data stream's subset of data and directing it elsewhere or to be reindexed (this is especially helpful in smoothing out the burstiness of the data collection process). Data from external sources may be queued directly into RabbitMQ if the format is appropriate or may be first collected from a system via `collectd`, parsed by Logstash, and then queued into RabbitMQ. `collectd` is used to minimize the impact of the data collection process on the underlying system. In addition, using `collectd`, collection rates and plugins can be configured independently for different sources.

To ingest data into Elasticsearch, JSON data is sent from the RabbitMQ cluster to Logstash. As the local aggregation point, Logstash reduces the number of network connections that the central logging clusters need to manage. The system's local server takes the local connections and forwards a single connection to the center logger. It also provides the encryption point so non-encrypted data is only present on the local system or within the local system where the data may be protected by other methods. Logstash also offers the ability to bridge between networks. For example a private network and a logging network. Lastly, Logstash can be used to convert `collectd` UDP packets to TCP, as a means of preventing dropped packets and lost data. Using RabbitMQ, Logstash, and Elasticsearch, OMNI is able to ingest over 25,000 messages per second from heterogeneous and distributed sources in and out of the data center.

Once ingested, Elasticsearch indexes the data for near real-time retrieval and querying. Data may be directly queried from Elasticsearch using the native RESTful APIs or using visualization and data discovery tools, such as Kibana or Grafana [4]. Both tools are web-based and allow for intelligent data delivery to the browser via JSON data structures. They are also specifically designed for visualizing and analyzing extremely large time series datasets. Visualizations may be gathered onto dashboards for providing monitoring information across the data center.

4 RESULTS

The following section describes the analysis and outcomes achieved using operational data from the OMNI data collection infrastructure at NERSC. The initial motivation for collecting operational data was due to a requirement to submit monthly metrics to DOE that help quantify how well the center aligns with its mission of delivering high-availability and reliability to science users. On an operational basis, these datasets allow staff to identify potential problems faster, more effectively diagnose the root causes of issues, and resolve incidents more quickly. Using visualization dashboards for near-real time monitoring, the data is used daily by NERSC staff to determine if the system is being utilized efficiently, if the types of jobs requested are meeting the DOE's mission, if there is a bottleneck in storage requests that is impacting job scheduling, and whether or not there are network latencies impacting data transfers, among other use cases.

As more diverse data is accumulated over long periods of time, the ability to ask more complex and specialized questions becomes possible. Some examples of these types of questions include:

- How efficient are the cooling towers during certain types of atmospheric conditions (i.e., high temperatures with low humidity, high temperatures with high humidity, etc)?
- How does the outside air quality impact the data center ecosystem?
- Are there software bug patterns that can be discerned over an eight-month period and not a three-month period?

Data from OMNI helps provide the team with a holistic view of the HPC data center and the environmental information that contributes to the center's overall status. When problems occur, NERSC staff is in a position to treat the symptoms but also to be able to determine the root cause or see the "big picture." Before a problem occurs, staff is able to see when something is not behaving as expected and can take proactive steps to mitigate risks.

The following sections discuss a number of results that were achieved using operational data analytics. These analyses provided NERSC with results that supported the business decision to collect and analyze operational data in the scenarios described.

4.1 Incorrect Voltage Issues after Data Center Relocation

In December 2015, the NERSC HPC data center geographically relocated from its Oakland, California location to a new facility, Shyh Wang Hall, in Berkeley, California. As part of the business strategy to continuously provide computational resources to scientific users during the transition, one HPC system, Hopper, continued to be available in Oakland, while a new system, Cori Phase I, was delivered to Berkeley and made immediately available to users. The third system, Edison, was migrated from Oakland to Berkeley.

Following the migration, in January 2016, a large job that used a significant portion of Edison's HPC resources finished in the job scheduling queue. Because the job was so large and cleaning up its resources takes some time, the system began to idle and multiple cabinets unexpectedly powered off without warning. Due to the high-availability of the OMNI cluster, NERSC engineers were able to turn to the OMNI datasets for insights about the facility power,

substation power, job scheduling data from slurm, and system metrics despite the fact that these cabinets lost power. A cursory data analysis showed that the cause of this incident was that the cabinets had received an over-voltage event.

By back-tracing the power source from the cabinets back through to the main breakers where it enters the building, it was discovered that the facility's substations were delivering an incoming voltage of 12 kilo-Volts, which was approximately 10% above the expected value. Further investigation uncovered that the building designers had used normal specifications rather than that of a data center in their designs and failed to account for the large-scale power fluctuations that can occur in HPC systems.

During the period after a large-scale job has finished, when the Edison system is idling, the power requirements of the system would change downward. This change caused the substation powering Edison to provide voltages that to the cabinets with idling nodes that were too high. Therefore, as part of the system's power supply self-protect mode, the system's cabinet power supplies shut down. 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.

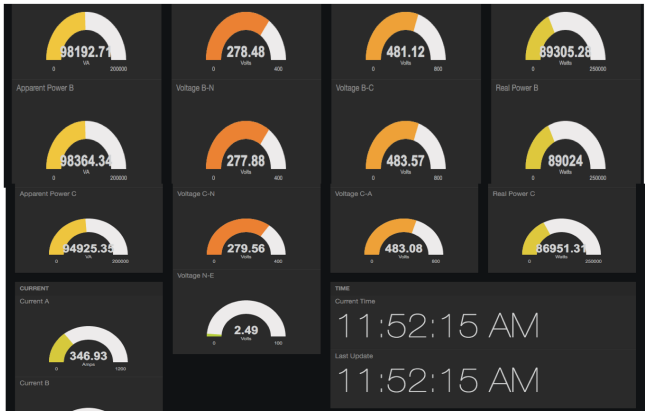

Figure 2: Edison Dashboard Panel Substation 628a3a - Pqube Statistics

Figure 2 is a dashboard of real-time power quality monitoring measured in four different areas of the facility. Column 1 is power coming from the substation, Column 2 is power from a specific panel to a common point in the facility, Column 3 is power from the panel to a specific point, like Edison, and Column 4 is the corrected power. Because this is in real time, it allows us to watch for anomalies like high voltage spikes.

In Figure 3, the graph is the house power to Edison, the blue line corresponding to C1-18. The line drops significantly during the point where nodes idled. As Edison nodes idled, the system's power requirement should have also dropped; it did, but the voltage provided by house power was not low enough.

Without this data, NERSC would not have realized that every panel in the HPC portion of the data center's substation transformer's output voltage needed to be shunted. This determination was made before the delivery of the rest of Cori.

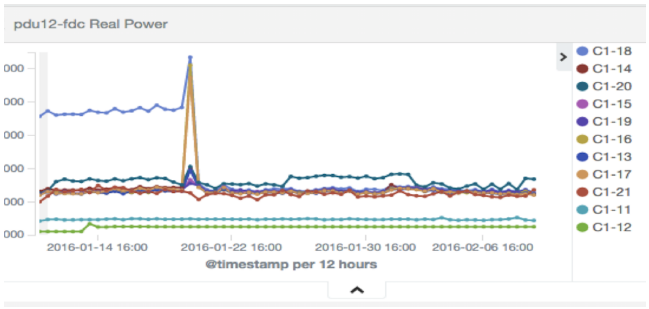

Figure 3: Edison C1-18 PDU Strips

4.2 Facility Power Planning for a Next-Generation HPC System

Each new NERSC high-performance computing system is orders of magnitude larger than its predecessors. The upcoming Perlmutter system, with an expected delivery date in late 2020, is no exception - it will be the largest HPC system to date. In order to support this pre-exascale machine, a number of facility upgrades are required, including upgrading electrical substations for the additional compute power requirements, as well as additional mechanical components, such as air handlers, evaporative coolers, water pumps, and cooling tower cells for controlling the environment in the data center. These cooling and environmental units are typically fed power from a substation that is separate from the compute substations, helping to separate the power used by the compute infrastructure from the power needed to maintain the environment of the data center, i.e., a controlled temperature and humidity that maximizes energy efficiency.

In planning for these additional units to control the data center environment, the power requirements for each of these units must be considered. Prior to the Perlmutter facility upgrades, Shyh Wang Hall utilized one dedicated mechanical substation for powering these non-compute components. To estimate the environmental and cooling power needs based on the units that are being added for Perlmutter, a routine capacity study was performed, using the electrical specifications of each component to calculate the theoretical peak load when all components are working at 100% power. The result of this study recommended the addition of an additional substation to support the increased mechanical power for Perlmutter.

However, the Berkeley Lab Facilities Master Specification permits a different calculation to be used for mechanical load planning in data center upgrades. This calculation requires at least one year's worth of operational data at the facility in which the upgrades will take place. OMNI had that data, and the analysis of it determined that operational power usage was 60 percent of the total power usage from the general specifications. Therefore, they determined that an additional substation was not required to support the new system. Plots from these analyses are shown in Figure 4 and Figure 5.

Figure 4 shows the total power load of all of the compute substations (yellow line, Substations 590 [Non-HPC Compute], 628

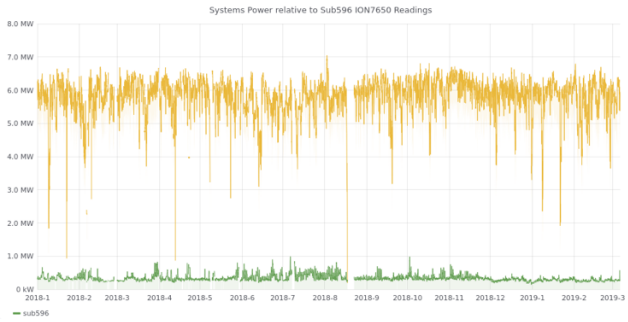

Figure 4: Total Power Load from the Compute Substations vs. Mechanical Power (kW)

[Edison], 612 [Cori], and 613 [Cori]) relative to the mechanical substation (green line, Substation 596) from January 1, 2018 at 12:00AM to present. This figure illustrates the consistency of the total power load of the compute resources at the center and, correspondingly, the largely consistent load of the mechanical substation. In the warmer months (e.g., 2018-6 through 2018-8), there is a higher than average overall demand on the mechanical substation but this demand does not exceed 1MW.

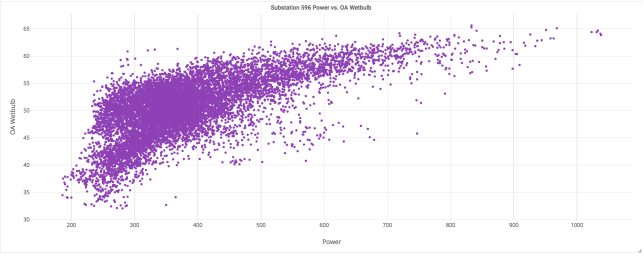

Figure 5: Power (kW) vs. Outside Air Wetbulb Temperature.

Figure 5 shows the load on the mechanical substation but this time relative to the outside air wet bulb. The wet bulb temperature measures the evaporative cooling and accounts for both the temperature and the moisture in the air. Previous analysis of data in OMNI has shown that the outside air wet bulb temperature has the largest influence on the overall power usage effectiveness of the facility. Using the scatterplot in Figure 5, NERSC staff verified that the higher readings above an 800 kW power draw are occurring on days when the wet bulb temperature is around 58F and above (i.e., days when it is both hot and humid outside). Analysis of the plot also shows that the majority of hours over the 15-month timespan are concentrated in that 300-500 kW power consumption range where the wet bulb reading fluctuates between 45-55F. This demonstrates that the Berkeley climate and evaporative cooling has a significant impact on maximizing energy efficiency and reducing unnecessary costs.

Ultimately, having the data available in OMNI enabled the Facilities team to analyze the data and come to the conclusion that Perlmutter's load would not require a new substation, saving NERSC about $2M.

An indirect benefit of analyzing these operational data produced some insights regarding the Slurm job scheduler and demonstrated why the power had a stable draw over time. Going back to Figure 4, there are some highs and lows over time, however, compute power (the yellow line) generally stays within the 6.0 MW and correlates with mechanical power (green line) also staying relatively even, mostly below 1.0 MW. The engineers expected to have more variations in the fluctuations but discovered when correlating this to Slurm data, that the rules to backfill a large job make the queues in general consistently full.

For example, when the system prepares for a large job requiring 1000 nodes for 6 hours, it takes time for these nodes to be prepared and ready. While waiting, should a job that requires a smaller amount of nodes and a completion time before the large job would start, the scheduler will have that job run. As a result, the large system queues are mostly always filled which is why the power usage for the compute systems stays stable. If the large system is consistently at a known load across a time period, then it follows that power usage will be the same across time.

Examining these two data sets in parallel provided confirmation to the HPC systems staff that they have efficiently managed the queue requirements and the job scheduler. Further, this data confirmed to Facilities and the Energy Efficiency groups how NERSC is able to achieve a stable draw on power relating to the compute systems. Having diverse data in OMNI allowed multiple groups to see insights into not only power usage but also job scheduler efficiency.

4.3 Collaborations with vendors

OMNI's scale of the data over time, allowed the facility to collaborate with vendors in helping solve issues not previously possible as highlighted in the scenarios described below. For example, the OMNI team specifically wanted to collect Cray system data about power, temperature, fan speeds, water temperature and jobs running for all the nodes, slots and cabinets. Much of this information is internal to the system and previously not exposed to their clients.

In collaboration with NERSC systems engineers, Cray created an API plugin architecture called xtpmd whose daemon runs on the Cray SMW and handles SEDC and high-speed power telemetry data as a stream before it is injected into the PMDB.

Using the API, NERSC engineers created a process where sites can independently write their own plugins to gather this data. The plugins can export data off Cray XC systems using Kafka, Redis Pub/Sub, and RabbitMQ. NERSC implemented this and streamed the data into RabbitMQ. As a result of this collaboration to gather system data into OMNI, this plugin is now available to the community. Figure 6 shows a partial display of a Cori power dashboard from OMNI.

While this is not a direct result of data analysis in OMNI, this was an important step in getting a previously unavailable external source of data into the collection. Further, this method not only served NERSC but also served the community at large.

As a result of collecting SEDC and telemetry data into OMNI, NERSC and Cray investigated a random anomaly occurring in systems. Using almost one year of data, engineers analyzed the data and correlated the anomaly to a pattern of system events. This

Figure 6: Cori power dashboard - partial display.

analysis resulted in Cray adjusting a thermal alarm setting in the HPC system. This solution was made possible by analyzing OMNI historical data.

There is also future work relating to implementing variable fan speeds with the potential to lower power usage on the HPC system that can now be done with SEDC data in OMNI.

4.4 Arc Flash

On December 31, 2018, the facility had an arc flash and experienced a level one fire alarm, the lowest alarm that detected smoke but not fire, as a result of a damaged bus bar. Once safe to return to the facility, during the inspection with the fire department, Operations staff notice the temperature was warmer than usual on the HPC floor. After examining several sets of environmental data from OMNI in correlation to the Building Management System (BMS) software, the team discovered that the air handlers had been automatically turned off, which is the correct action should a fire occur. However, restarting the air handlers required the Berkeley Lab fire alarm system to be reset and can only be completed by an authorized fire technician. Restarting the air handlers also required an authorized facility engineer who had access to the Lab's BMS software.

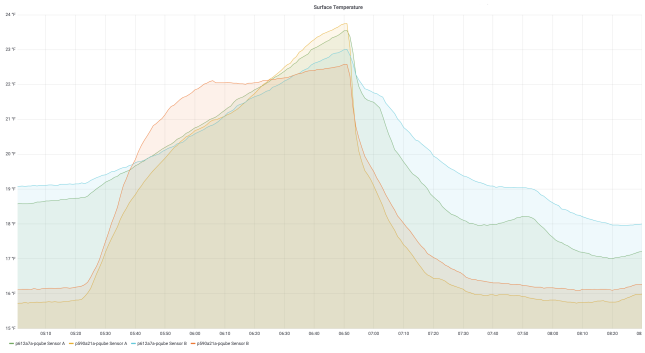

Figure 7: PQube Sensor Surface Temperature immediately after Arc Flash Event.

Figure 7 is a graph of the surface temperature as recorded from 4 PQube sensors in the data center. Immediately after the arc flash event, the graph shows a significant increase in temperature for all sensors from 16 degrees Celsius all the way to 22 degrees Celsius. OMNI data show trends that the room temperature was rising 1 degree per minute on average with the possibility that some equipment's maximum temperature rating would be reached before others. While this is not a direct analysis of data, visualizing BMS data in OMNI dashboards that show temperature trends allowed Operations staff to prioritize which equipment needed immediate attention.

As a result, these processes have been examined and the facility has determined new actions to take in cooperation with the appropriate Berkeley Lab departments to ensure the safety of the assets on the HPC floor.

5 LESSONS LEARNED

Fourteen years ago, when NERSC moved to the Oakland Scientific Facility, the Operations team instrumented what they needed as they needed them to collect environmental data. Moving back to Berkeley, they wanted to ensure that the brand new facility was an opportunity from the ground up to install the instrumentation and build the infrastructure around it. When the management team saw the potential of centralized data for analysis, correlation, business decisions, capacity planning, facility planning, etc, the collection became a necessity.

5.1 Philosophy differences

Operational data, not just environmental data. OMNI initially collected data from the sensors and devices on the HPC floor to monitor power, humidity, air flow, temperature, water pressure, etc. These measurements are essential to maintaining the facility and to creating the required environment for the large systems. Soon after, other NERSC groups saw the potential of a centralized database of collected data, thus, system logs, network data, disk I/O, border traffic, etc. are being requested to also be included. The initial system configured is a 4u/4 node system that grew by a factor of 4 in the second year and another factor of 4 by the third year.

In hindsight, a discussion of the potential of this type of data collection should have happened earlier however, we do not believe this would have been possible because no one saw its true potential until a significant amount of heterogeneous data was collected. The collection started as environmental data for the facility and eventually became operational data for NERSC.

How much and which data? OMNI collects, centralizes and archives data already being collected in a standard format. This can potentially be a lot of data. This idea is counter intuitive to the philosophy of "collect data to answer a specific question." Many organizations tend to collect data they can immediately analyze and gain insight. People did not understand why to "collect all the data" and we are constantly asked, "how do you know which data to collect?".

For the OMNI team, it is better to collect 100% of the data and be able to only analyze 80% than to collect only 80% of the data and to have something missing. Our worst case scenario is that someone wanted to analyze a year of data set A and B in comparison to data set C only to find out that data set C is not being collected. How do we know which data to collect? If it is being collected, it should be stored in OMNI.

What other groups did not appreciate is, once the data is stored in OMNI, the data is no longer a group or individually owned

but belongs to NERSC. Eliminating ownership issues breaks down barriers to information and insight. The OMNI team will do the work to ensure the data is segmented; however, the purpose of placing the data in a centralized location is to ensure all groups of data can be analyzed together, correlated together, used together to answer complex questions.

Who does the analysis and correlation? Once the data resided in OMNI, there was confusion on who did the analysis and correlation. Granted, Operations staff did their own analyses and correlations to insights that are operational and used to inform daily decisions, diagnoses, early detection of issues and some planning.

Because we collected the data, it is assumed we are also data scientists and were asked: "what new insights did you get from your collection?" As system administrators, our expertise is in acquiring the collection, its storage and the availability of the data. It is a challenge to take the leap from this function into one of a data scientist.

We should have communicated our intentions from the beginning much better, that we wanted to centralize the storage of data collection, not necessarily to be the group to analyze the data for other groups.

5.2 Growth and scalability

The initial configuration of OMNI is less of a challenge to implement for NERSC because of the facility's non-classified status. NERSC has its own subnet, managed by its own networking team. Data from its sources are not moving outside of the NERSC subnet, possibly unlike that at other Labs or organizations.

The OMNI team's decision to use open source software is a business decision. Over the years, it had been their experience with a paid product that they are constantly asking for and waiting for updates, features, and other improvements. The cost of the license can be an issue and the group faced the dilemma of either paying for a license and having to purchase reduced hardware or use open-source in order to be able to purchase more hardware. Besides, when there is an open-source alternative, management would require to, at minimum, try the open-source version first.

Initially purchasing as much hardware as possible is important to grow to scale and this is where we wanted to spend our budget. Once there was enough data collected and we began to scale, the team was quickly able to troubleshoot the problems and shared solutions with the community. Solving problems was a much better fit for the team than reporting a problem and waiting. However, supporting the infrastructure and making the data highly available became a challenge especially when the issues they encountered are not those of the community. The OMNI data collection scale is so large and the data so diverse that there is no one to ask for advice and the team is on their own to solve it.

Over the years, the team saved close to $350K in licensing costs yearly and continue to save that as long as the open source product is used. Today there are features in the paid product that would be nice to have, but we hope they will eventually be addressed as another open source product. For now, the team would not exchange their decision to remain with the open source product and maintain their flexibility to solve problems they encounter in the future.

6 CONCLUSION

The OMNI team plans the following expansion in the near future:

- upgrading the storage drives to faster medium from SATA to non-volatile memory express (NVME)
- replacing collectd where Prometheus can be used in preparation for Perlmutter
- using Kubernetes as OMNI's container management system
- increasing the ingest rates to 100k/second or more, potentially bypassing RabbitMQ and directly sending data streams to Victoria Metrics. This extremely fast time-series database is much more efficient for long-term storage and can directly ingest data from Prometheus.

OMNI continues to grow and new data sources are constantly being identified to be incorporated into the collection, for example, Lightweight Distributed Metric Service (LDMS) data. Using OMNI, NERSC management is able to provide the necessary metrics to DOE on a monthly basis. Operational teams are able to use real-time data to make the best daily decisions to keep the HPC systems highly available and highly utilized by the scientific community. Many groups use the OMNI data to gain new observations on how to improve performance, utilization and to gain insights from the correlated information. The data has been used to lower costs, save hardware, assist with business decisions and influence collaborations. NERSC continues to achieve operational efficiency in its new facility and the use of OMNI datasets helps makes this possible.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Energy Efficient HPC Working Group Operational Data Analytics (ODA) Team with special thanks to Torsten Wilde, Ghaleb Abdulla, and Natalie Bates for their technical leadership of the ODA Team and for spearheading the idea of this case study.

We also acknowledge Norman Bourassa at Berkeley Lab for encouraging us to document this case study.

This research 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.

REFERENCES

[1] Jack J. Dongarra, Hans W. Meuer, and Erich Strohmaier. [n. d.]. Top500. ([n. d.]). Retrieved May 7, 2019 from https://www.top500.org/
[2] Elastic 2019. Elastic Stack. (2019). Retrieved May 9, 2019 from https://www.elastic.co/products/
[3] Elasticsearch 2019. Elasticsearch. (2019). Retrieved May 9, 2019 from https://www.elastic.co/products/elasticsearch
[4] Grafana 2019. Grafana. (2019). Retrieved May 9, 2019 from https://grafana.com/
[5] The Green Grid. 2007. The Green Grid power efficiency metrics: PUE and DCiE. (2007).
[6] Kibana 2019. Kibana. (2019). Retrieved May 9, 2019 from https://www.elastic.co/products/kibana
[7] Logstash 2019. Logstash. (2019). Retrieved May 9, 2019 from https://www.elastic.co/products/logstash
[8] oVirt 2019. oVirt. (2019). Retrieved May 9, 2019 from https://ovirt.org/
[9] RabbitMQ 2019. RabbitMQ. (2019). Retrieved May 11, 2019 from https://www.rabbitmq.com/
[10] rancher 2019. Rancher. (2019). Retrieved May 10, 2019 from https://rancher.com/