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

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

Elizabeth Bautista
ejbautista@lbl.gov
Lawrence Berkeley National Laboratory

Thomas Davis
tadavis@lbl.gov
Lawrence Berkeley National Laboratory

Cary Whitney
clwhitney@lbl.gov
Lawrence Berkeley National Laboratory

ABSTRACT

Real-time operational data from high-performance computing (HPC) data centers is growing in volume and velocity on the path to exascale. Efficient methods to monitor these resources and identify problems that arise in near-realtime throughout the data center is of utmost importance in delivering high-availability to the computational scientists that make use of them. This poster describes the design of an Operations Monitoring and Notification Infrastructure (OMNI) at the National Energy Research Scientific Computing (NERSC) center at Lawrence Berkeley National Laboratory, capable of continuously gathering, storing, and analyzing data from systems and sensors throughout the data center in near-realtime. 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.

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 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).

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. 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 (defined to be measured from the PDU outputs in terms of equipment, utility inputs in terms of facility, at an interval of hourly and daily) Power Usage Effectiveness (PUE) [2, 3] ratio of 1.07 over the past year.

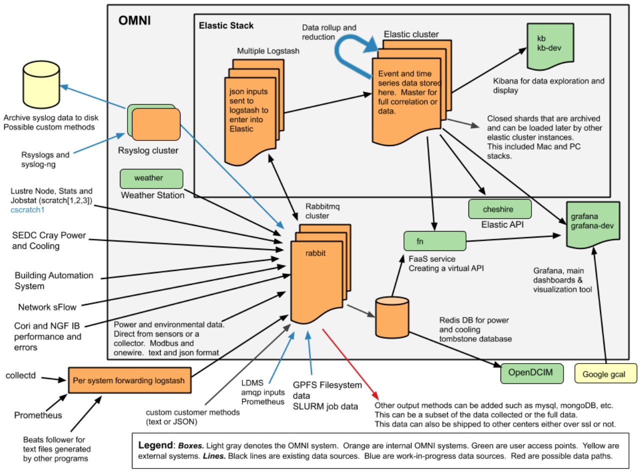
OMNI
Elastic Stack
Data rollup and reduction
Elastic cluster
Multiple Logstash
json inputs sent to logstash to enter into Elastic
Event and time series data stored here. Master for full correlation or data.
kb
kb-dev
Kibana for data exploration and display
Closed shards that are archived and can be loaded later by other elastic cluster instances. This included Mac and PC stacks.
Archive syslog data to disk
Possible custom methods
Rsyslog cluster
Rsyslogs and syslog-ng
weather
Weather Station
Lustre Node, Stats and Jobstat (scratch[1,2,3])
cscratch1
Rabbitmq cluster
cheshire
Elastic API
grafana
grafana-dev
SEDC Cray Power and Cooling
rabbit
fn
FaaS service
Creating a virtual API
Building Automation System
Grafana, main dashboards & visualization tool
Network sFlow
Cori and NGF IB performance and errors
Power and environmental data. Direct from sensors or a collector. Modbus and onewire. text and json format
Redis DB for power and cooling tombstone database
collectd
Per system forwarding logstash
OpenDCIM
Google gcal
Prometheus
LDMS
amqp inputs
Prometheus
GPFS Filesystem data
SLURM job data
custom customer methods (text or JSON)
Other output methods can be added such as mysql, mongoDB, etc. This can be a subset of the data collected or the full data. This data can also be shipped to other centers either over ssl or not.
Beats follower for text files generated by other programs
Legend: Boxes. Light gray denotes the OMNI system. Orange are internal OMNI systems. Green are user access points. Yellow are external systems. Lines. Black lines are existing data sources. Blue are work-in-progress data sources. Red are possible data paths.

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

3 OMNI INTEGRATED OPERATIONAL DATA COLLECTION AND ANALYTICS

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 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:

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. 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.

Figure 1 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. All data collected is time synced to a local stratum, an network time protocol (NTP) server. The data streams themselves are not synchronized to each other. Data is ingested into an ElasticSearch [1] cluster via RabbitMQ [4]

4 RESULTS

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 results show a deep historical power analysis of the Cori supercomputer's power consumption, an evaluation of whether or not a new mechanical substation was needed for an upcoming pre-exascale system, how data led to cooled water tower pump upgrade, as well as how OMNI data was utilized for realtime network monitoring at the center.

REFERENCES

[1] Elasticsearch 2019. Elasticsearch. (2019). Retrieved May 9, 2019 from https://www.elastic.co/products/elasticsearch

[2] The Green Grid. 2007. The Green Grid power efficiency metrics: PUE and DCiE. (2007).

[3] The Green Grid. 2012. *PUE: A Comprehensive Examination of the Metric.* Technical Report.

[4] RabbitMQ 2019. RabbitMQ. (2019). Retrieved May 11, 2019 from https://www.rabbitmq.com/