Artificial Intelligence for Data Center Operations (AI Ops)

Austin Todd,[1] Avi Purkayastha,[1] Hilary Egan,[1]
David Sickinger,[1] Matthew Eash,[1] Sergey Serebryakov,[2]
Jeff Hanson,[2] Matt Slaby,[2] Nick Wunder,[1] Nalinrat Guba,[1]
Kristin Munch,[1] Tahir Cader,[2] and Caleb Phillips[1]

1 National Renewable Energy Laboratory
2 Hewlett-Packard Enterprise

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-2C00-79712
May 2021

Artificial Intelligence for Data Center Operations (AI Ops)

Austin Todd,[1] Avi Purkayastha,[1] Hilary Egan,[1]
David Sickinger,[1] Matthew Eash,[1] Sergey Serebryakov,[2]
Jeff Hanson,[2] Matt Slaby,[2] Nick Wunder,[1] Nalinrat Guba,[1]
Kristin Munch,[1] Tahir Cader,[2] and Caleb Phillips[1]

1 National Renewable Energy Laboratory
2 Hewlett-Packard Enterprise

Suggested Citation
Todd, Austin, Avi Purkayastha, Hilary Egan, David Sickinger, Matthew Eash, Sergey Serebryakov, Jeff Hanson, Matt Slaby, Nick Wunder, Nalinrat Guba, Kristin Munch, Tahir Cader, and Caleb Phillips. 2021. *Artificial Intelligence for Data Center Operations (AI Ops)*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-2C00-79712. https://www.nrel.gov/docs/fy21osti/79712.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-2C00-79712
May 2021

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy and Hewlett-Packard Enterprise. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Acknowledgments

This research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory. This work was partially supported by a research agreement with Hewlett Packard Enterprise (HPE). The authors would also like to thank Vijay J. Arokia-Raj, Deepak Nanjundaiah, Subrahmanya V. Joshi, Amarnath Chilumukuru, and Cornel Boac at HPE for their ongoing support and collaboration.

1 Introduction

The U.S. Department of Energy's National Renewable Energy Laboratory (NREL) hosts the world's most energy-efficient High-Performance Computing (HPC) data center (NREL, 2018). The data center is located inside the Energy Systems Integration Facility (ESIF), a 17,000 m^2 research facility on NREL's campus in Golden, Colorado that was rated Platinum by the Leadership in Energy and Environmental Design (LEED) certification program for its energy efficient design and construction. The ESIF contains integrated energy laboratory space, a hydrogen storage facility, office space, state-of-the-art 3D visualization rooms, and the HPC data center. A holistic approach was taken to design the 930 m^2 data center, with a goal to capture waste heat and to facilitate the efficient use of energy resources.

Data centers are energy-intensive facilities that typically rely on air circulation to remove heat generated by the IT equipment. While this is effective in cooling the IT equipment, it requires a significant amount of facility energy to both cool the ambient air and to run fans that distribute this cool air to the individual components. With data center energy consumption nationally over 70 billion kWh per year (representing almost 2% of energy consumption in the United States) and increasing by roughly 4% annually (17, Masanet et al., 2020), there is a clear need to improve data center energy efficiency in order to drive down energy consumption in these data centers. In fact, data center efficiency has steadily increased since 2007. The energy efficiency is often measured by the power use effectiveness (PUE), which is the ratio of all energy required to run the data center to the energy required to run just the IT equipment, with values close to 1 that indicate better energy efficiency. Improvements in data center designs have led industry average PUE values across both water- and air-cooled systems together to fall from 2.5 in 2007 to 1.67 in 2019 (17).

NREL's ESIF data center accomplishes significantly greater facility energy consumption reduction compared to other data centers nationally by moving away from a traditional design with rows of air-cooled components and moving to component-level warm-water liquid cooling that efficiently removes heat from the data center. The waste heat captured from IT equipment can be reused within ESIF or rejected to the atmosphere without any mechanical cooling (which eliminates expensive and inefficient chillers). This design has allowed the ESIF data center to maintain a trailing 12-month average PUE of 1.06 or better since opening in 2013, indicating that on average only 6% of the energy entering the data center is consumed by the facility to deliver power and cooling to IT equipment. This is a significant improvement over the PUE average of data centers nationwide. In fact, NREL's ESIF data center is also distinguished from other data centers in the United States by its reuse of energy (e.g., heat) generated by the data center for heating of the ESIF which accounted for 10.5% of annual IT equipment heat rejection during the first year of operation (1717). The ESIF data center was also designed to reduce water use through utilization of a thermosyphon hybrid cooling system which reduced water consumption by 7.9 million liters over a 2-year period from August 2016 to August 2018.

In addition to energy efficiency, as we move to Exascale systems, a related concern is data center resiliency. HPC data centers such as the one at NREL's ESIF will increasingly need to rely on automation to keep pace with exascale growth in compute capability and to manage and optimize the data center environment and facility resources. Artificial intelligence (AI) and machine learning (ML) approaches provide the means to improve HPC data center efficiency (energy, operational, and managerial efficiency) and resiliency by learning historical trends and training models to operate on real-time data collected from both IT and facilities sources. The goal of coupled improvement of data center resiliency and energy efficiency through automated data collection and AI has led to a multi-year, multi-staged collaboration between NREL and Hewlett-Packard Enterprise's Advanced Technology Group, referred to as *Artificial Intelligence for Data Center Operations* (AIOps). The extended efforts within the AIOps project include a common goal of building capabilities for an advanced smart facility and demonstration of data collection and AI modeling techniques in the ESIF data center.

Given the complexity of a system such as the ESIF data center, building data-driven tools for monitoring and operating the system requires a large variety and quantity of sensors to adequately capture the computing facility and related IT energy use data. NREL's set of sensors measure not only power consumption from IT equipment, but also metrics about network use, storage, various system components (e.g., temperature, pressure, flow rate, valve states, fan speeds) internal to the data center, and external environmental conditions. Through this ongoing effort to build an advanced smart facility, over one million metrics are recorded per minute using state-of-the-art streaming data architecture and software to capture and understand the state of the system in real time.

This streaming data platform allows for innovation through the development of data-driven analytics applications that support maintenance and operation of the system. Of particular importance to data center operations is the implementation of cooling system control algorithms to help keep air and water temperatures within appropriate operational ranges, anomaly detection algorithms to identify potential thermal leaks or device failures, and optimization algorithms to maximize the overall resiliency and efficiency of the data center. Nearly every device (motors, fans, valves) can be set to control the energy efficiency of the overall system, and the AIOps project seeks to build capabilities that utilize data and device controls that can eventually lead to automated data center operations.

This technical report describes progress and findings on the first two years of the AIOps collaboration, in particular:

- Details of the overall streaming data and software architecture that have been developed to support the capture and accessibility of the complete set of facility and IT energy data from hundreds of sensors, meters, and controls.

- Dashboards, visualization tools, and algorithms that have been developed to better monitor the state of the various components of the data center.

- Preliminary results from ongoing research into understanding the power footprint of certain application job types and their impact on the data center's cooling resources.

In the next section, we provide background on the data center design and motivation for this effort. This section is followed by detailed descriptions of methods and results. Finally, we conclude with a summary of impacts and discussion of ongoing work and next steps.

2 Background

In building its data center, NREL's vision was to create a showcase facility that demonstrates best practices in data center sustainability and serves as an exemplar for the scientific computing community. The innovation focused on three critical aspects of data center sustainability:

- Efficiently cool the IT equipment using direct, component-level liquid cooling with a PUE design target of 1.06 or better.
- Capture and reuse the waste heat produced.
- Minimize the water used as part of the cooling process.

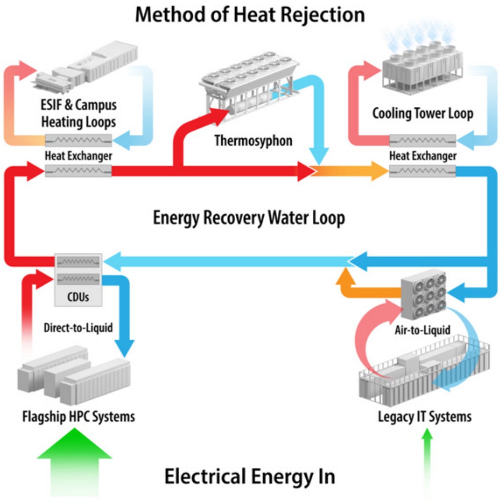


Figure 1: Cooling system representation for the HPC data center in the ESIF

The ESIF data center accomplishes these goals through high efficiency architectures, instrumentation, and monitoring. Electrical energy supplied to the data center IT systems (shown along the bottom of Figure 1) is converted to thermal energy, with the majority of IT energy utilized by the flagship HPC systems that are direct liquid-cooled. The liquid cooling approach involves a cooling distribution unit (CDU), which interfaces with the facility cooling loop and provides cooling liquid at the appropriate temperature, pressure, and chemistry for the IT equipment. There are also legacy IT systems that are traditionally air-cooled, and through the use of fan walls, the heat from these systems is also transferred to a closed facility loop called the Energy Recovery Water (ERW) loop. To maximize energy efficiency, there are no compressor-based cooling systems. The three heat-rejection options for this IT load operate in the following hierarchy and are shown along the top of the diagram in Figure 1 (priority is indicated from left to right):

- When possible, heat energy from the energy recovery loop is transferred through the energy recovery heat exchanger to the ESIF building process hot water loop to help heat the building or campus.
- After reuse potential is exhausted and when temperatures permit, heat is dissipated through the thermosyphon cooler (an advanced dry cooler that uses refrigerant in a passive cycle to dissipate heat) to economize water use. More details on how the Thermosyphon reduces water usage can be found in 17.
- Finally, the remaining heat is transferred from the ERW loop to a tower water open loop via the cooling tower heat exchanger. The resulting ERW supply temperature is 24°C or lower.

The data center was designed to support up to 10 MW of total IT load at full capacity. However, the current system capacity is 5 MW with typical operation about 2 MW. There are two different mechanical rooms in the data center - one of which is located directly below the data center floor where the hydronic distribution system for racks and the heat exchanger for energy recovery are located. There is also an outdoor platform where the thermosyphon and multiple cooling towers are located.

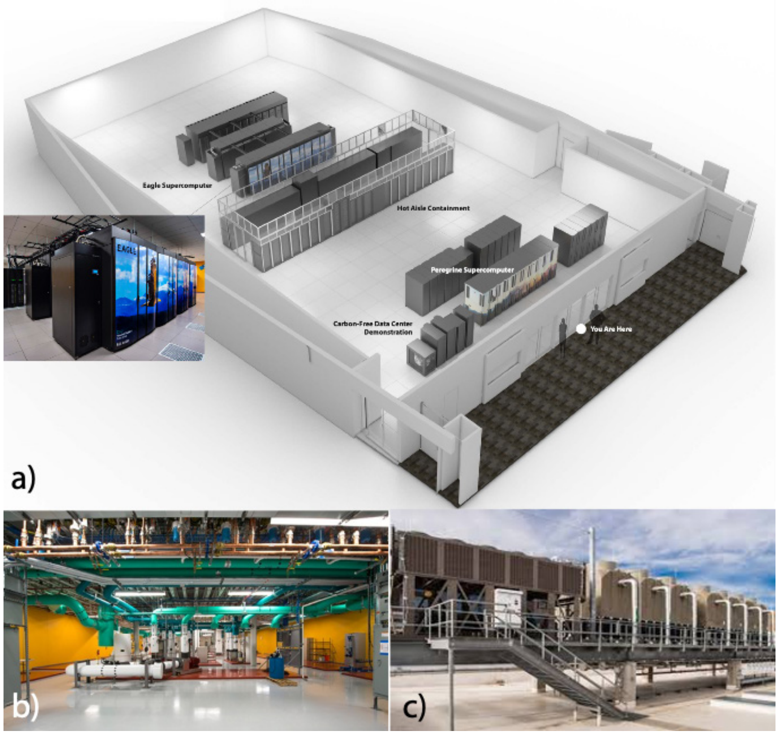


Figure 2: NREL ESIF data center in Golden, Colorado, USA.

The largest IT energy component inside the ESIF data center is the flagship HPC system. NREL's first flagship HPC system, named Peregrine, was located on the data center floor near a visitor window (Figure 2) and was deployed in two separate phases. Peregrine was active from the opening of the ESIF building in 2013 until it was retired in August 2019. Peregrine was the first installation of the HP Apollo Liquid-Cooled Supercomputing Platform and consisted of 2,592 compute nodes and a 2.25 petabyte data storage system.

The current flagship system, named Eagle, entered production in January 2019. The Eagle system resides in the back half of the data center that now consists of ten HPE 8600 E-Cells with Intel Xeon Gold Skylake processors, containing a total of 2,618 nodes. Each E-Cell consists of two 42U high racks in a sealed unit that uses closed-loop cooling technology. The ten E-Cells, along with five corresponding CDUs take up two rows (the floor graphic in Figure 2 shows the initial deployment of seven E-Cells and four CDUs). The Eagle system also includes a 3rd row of HPE Adaptive Rack Cooling System (ARCS), which houses “Big Memory” nodes and special-use nodes that also include 100 GPUs for graphics processing. Eagle also has a 17 petabyte data storage system with a parallel, high-performance Lustre filesystem.

3 Methods

3.1 Streaming Data Architecture

Over 1 million metrics are collected per minute related to the Eagle system, and more than 4,000 metrics are collected per minute related to the ESIF data center and facilities. The facility metrics consist of power, temperature, flow rate, pressure, and other states (e.g., alarm, position, speed) for facility and data center components including data center cooling towers and thermosyphons, pumps, fan walls, heat exchangers, hydronic loops, and environmental conditions (e.g., outdoor air temperatures and humidity). The Eagle metrics consist of integrated Eagle job logs; node metrics such as memory, disk, network, processor, and GPU utilization, plus hardware power and temperatures; and InfiniBand, Lustre, and application metrics. In addition, rack-level data such as air temperature, fan speeds, rack hardware, water temperatures, and inverter data are collected, as well as data from the CDUs and ARCS that include temperatures, flow rates, and pressures. Many of these metrics are collected every few seconds, while some are collected at 1-minute intervals.

Such a large volume and velocity of data requires a system that can effectively handle millions of simultaneous data streams but is also resilient to downtime and lags in reporting. The data architecture design for the collection of data in the ESIF data center therefore considers the data sources, data frequencies, the movement of data, and the eventual storage and use of the data. The goal of the data collection architecture is to provide a scalable infrastructure suitable for collecting, managing and processing streaming data from multiple heterogeneous data sources. The integration of the HPC node-level metrics, jobs data, and facility data in a single platform is extremely valuable for improving data center resiliency and for future applications of energy optimization, such as job scheduling based on power profile or optimizing water temperature setpoints and cooling power configurations.

The data architecture was implemented with a focus on open-source platforms and highly scalable systems. The data sources either push data to a device historian (a single-node Influx database running at the network interface to enable meter collection) or push data directly to a five-node Apache Kafka streaming data cluster (Figure 3).

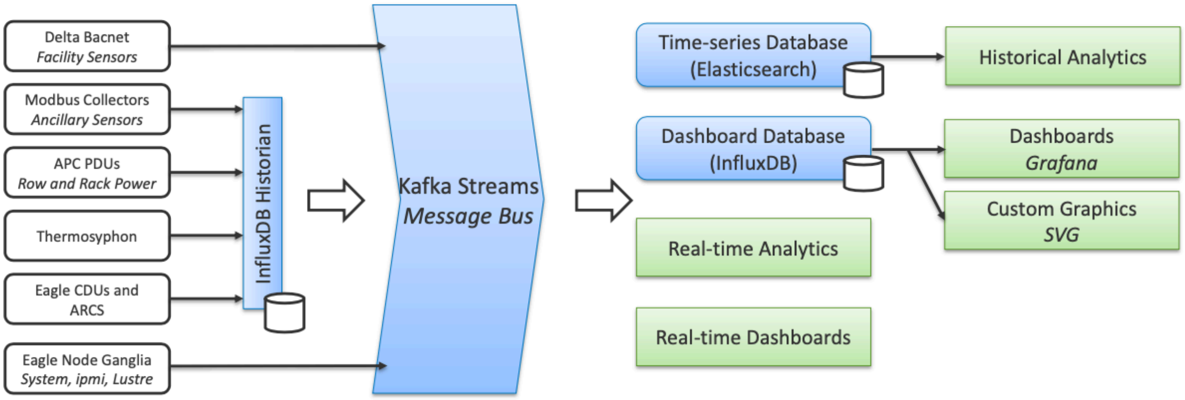


Figure 3: NREL HPC data architecture design

Data collected into the device historian is periodically queried and pushed to the streaming data cluster. The data streams are then accessible from a number of clients for either real-time visualizations and analytics or are collected into a time-series cluster for archival storage. The time-series cluster is an Apache Druid installation: an open-source, distributed data store that is designed to quickly ingest massive quantities of event data and allows for real-time analytics on top of the data. The data is persistent in the time-series cluster for historical analysis and interactive dashboards over the entire dataset. The persistent storage of the data is particularly useful when testing new predictive analytics methods, as the entire historical dataset is available for training and validation.

3.2 Data Center Metrics

One important use of the data collection capabilities is to obtain a clearer understanding of the data center performance relative to power usage effectiveness (PUE) and energy reuse effectiveness (ERE). The PUE and ERE are critical to understanding the real-time and long-term performance of the ESIF data center. Numerous data points are involved in the PUE and ERE calculations for both the facility and equipment. These readings are then used to calculate the PUE, which is defined as

$$PUE = {}^{(Facility\ Energy + IT\ Energy)}\!\big/_{IT\ Energy}$$

and ERE is defined as:

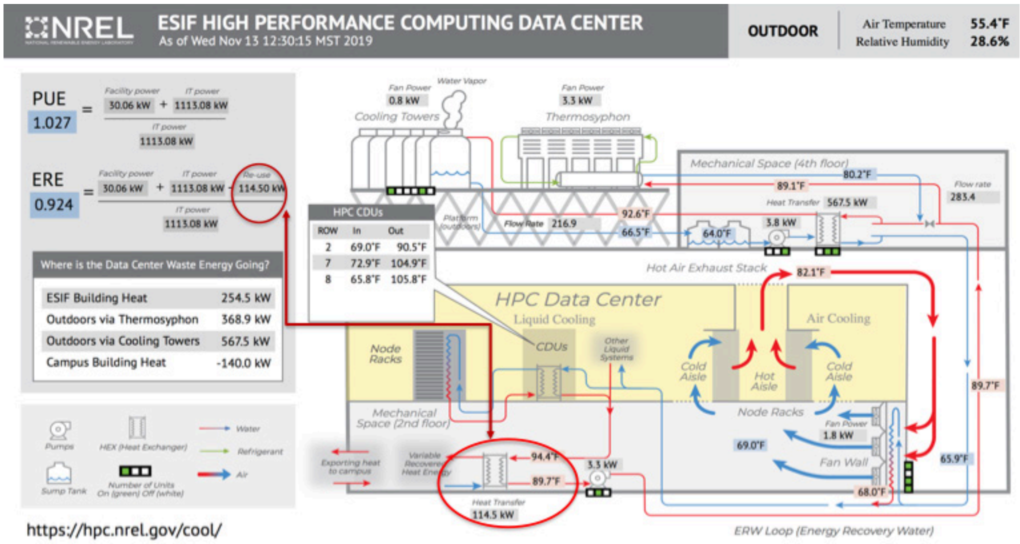
$$ERE = {}^{(Facility\ Energy + IT\ Energy - Reuse\ Energy)}\!\big/_{IT\ Energy}.$$

3.3 Anomaly Detection

Exascale data centers can be expected to be more highly instrumented and complex than today's already complex HPC data centers. With significantly more data being collected at much faster rates, the management of future data centers will be more difficult and more prone to failures. Through the AIOps Collaboration, NREL and HPE aim to introduce automatic, rapid, real-time, and highly scalable anomaly detection to the data center in order to:

- Reduce equipment failure and downtime to increase data center resiliency.
- Introduce advanced monitoring techniques to reduce false alarms, simplify data center management, and enhance root-cause analysis of anomalous data.
- Provide a means of scalability for monitoring data feeds more broadly by focusing only on anomalous results.

As part of this collaboration, HPE has developed AI and ML tools to develop a robust anomaly detection capability that performs in real-time, automatically, and at massive scale. An end-to-end anomaly detection pipeline was deployed in the ESIF Data Center in June 2020, with the pipeline operating on



ESIF HIGH PERFORMANCE COMPUTING DATA CENTER
As of Wed Nov 13 12:30:15 MST 2019

OUTDOOR — Air Temperature 55.4°F, Relative Humidity 28.6%

PUE 1.027 = (Facility power 30.06 kW + IT power 1113.08 kW) / IT power 1113.08 kW

ERE 0.924 = (Facility power 30.06 kW + IT power 1113.08 kW - Re-use 114.50 kW) / IT power 1113.08 kW

Where is the Data Center Waste Energy Going?

ESIF Building Heat	254.5 kW
Outdoors via Thermosyphon	368.9 kW
Outdoors via Cooling Towers	567.5 kW
Campus Building Heat	-140.0 kW

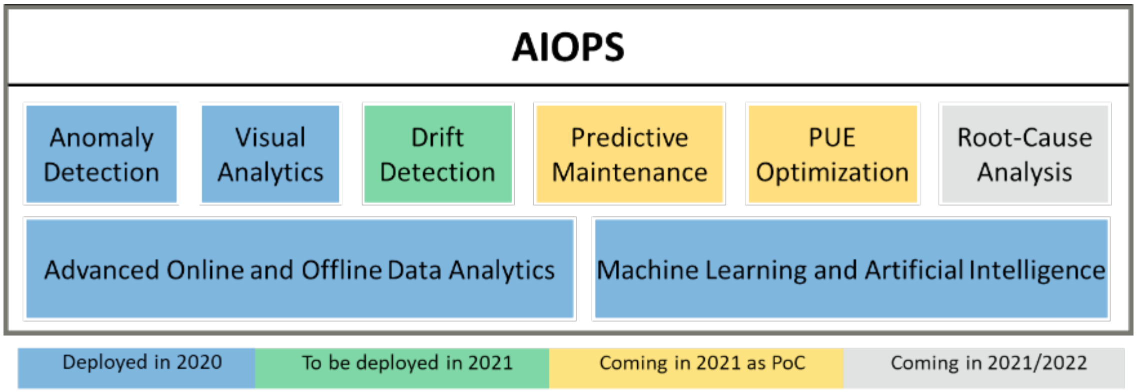
HPC CDUs

ROW	In	Out
2	69.0°F	90.5°F
7	72.9°F	104.9°F
8	65.8°F	105.8°F

https://hpc.nrel.gov/cool/

Figure 4: Grafana Cooling Tower dashboard for the ESIF data center

real-time IT (Eagle cluster) and facilities data. The current components and status of the AIOps stack are shown in Figure 5. The anomaly detection pipeline uses both uni-variate and multi-variate models. Multi-variate models are applied to single devices, such as a single CDU, and data center-wide, with models configured to ingest data from multiple devices (e.g., a computational node, a network switch, a CDU, a cooling tower, etc.) in order to detect anomalies in metrics correlated across multiple disparate devices in the data center. Multi-variate anomaly detection can lead to a reduction in false positives and false negatives since anomalous behavior is considered in the context of multiple correlated metrics whose joined behavior is known to a model. As part of the AIOps stack, AIOps ML includes Statistical methods (Z-Score, (double) MAD, Tukey, Entropy-based) for univariate anomaly detection models, machine learning and deep learning methods such as forecasting-based (ARIMA, LSTM), and reconstruction-based (PCA, autoencoders) anomaly detection for both uni-variate and multi-variate models. The AIOps ML stack provides components to support end-to-end ML workflows, from loading, cleaning, and visualizing data to training various anomaly detection models and evaluating and deploying models. AIOps ML supports supervised and unsupervised model evaluation. The basis of unsupervised evaluation is the store of realistic uni-variate and multi-variate models of anomalous behavior. AIOps ML takes nominal metrics, injects artificially generated anomalous data, and runs



AIOPS

Anomaly Detection | Visual Analytics | Drift Detection | Predictive Maintenance | PUE Optimization | Root-Cause Analysis

Advanced Online and Offline Data Analytics | Machine Learning and Artificial Intelligence

Deployed in 2020 | To be deployed in 2021 | Coming in 2021 as PoC | Coming in 2021/2022

Figure 5: AIOps technology stack

models against the modified data. This feature enables more comprehensive analysis of the models' performance, and better model tuning.

In support of operational resiliency, the streaming data and analytics platform was initially deployed with a pipeline for detecting anomalies in the cooling infrastructure using historical and real-time data from the Eagle supercomputer and the ESIF data center (Figure 6). For this application, CDU and Cooling Rack Controller (CRC) data streams are pushed to the Kafka cluster. These data streams are then used as input into an anomaly detection framework that implements a variety of anomaly detection models (discussed above) on real-time data to detect abnormal activity for the CDUs and CRCs. For this case, both univariate and multivariate models are used in the anomaly detection, where the anomaly score is determined from the reconstruction or forecasting error discovered by the model. The anomaly detection model then sends the anomaly scores, current anomaly thresholds, and anomaly labels back through the Kafka cluster for use in alerts, dashboards, and analysis programs. Historical datasets were used to train the models offline, while the near real-time model is run on streaming data from the Kafka feeds.

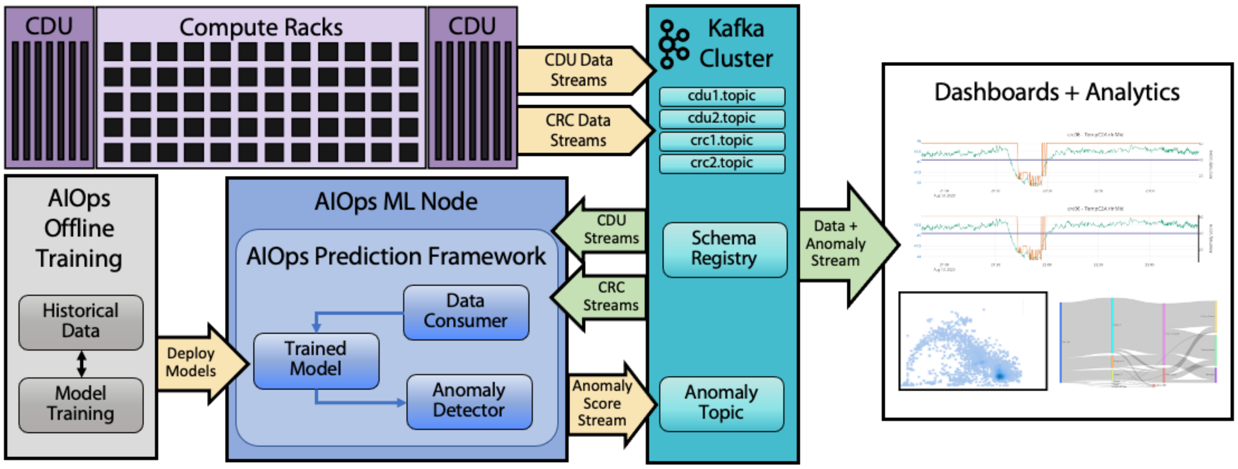


Figure 6: Streaming data architecture for the AIOps project

4 Results

4.1 Dashboards

The NREL Advanced Computing Operations (ACO) team has implemented a number of dashboards to better understand the state of the data center systems, to monitor energy efficiency (PUE and ERE), and to display facility metrics related to thermal loops. While these dashboards help describe the numerous facility metrics with various layers of granularity, the complexity of manually monitoring such a system is a cumbersome task. This stems from the large number of simultaneous data streams that require monitoring as well as the compounding impacts of a large number of potential adjustments that can be made to nearly every device in the facility cooling system to achieve optimal system performance.

The ACO team has also found that set-points, alarms, and dashboards are not always sufficient to identify anomalies in the system. For example, cooling tower configurations and water temperature set-points have a tendency to be set once and left alone if the overall system is working well. This creates opportunities for improved efficiency through automated set-point control. In addition, facility events from pump and cooling tower failures are not always easy to identify through dashboards, highlighting the need for a new approach to monitoring and system optimization.

There are additional challenges to dashboard building with such a large amount of data. For example, high frequency anomalies may not be seen on dashboards that aggregate sensor values by averaging over longer frequencies. In addition, some activity requires investigation at a temporal resolution that is typically not used for data center operations management (e.g., the need to look at graphs of multiple data points within a single minute). Finally, some activities require a look at data at different scales, such as the loss of a 2 kW fan that does not manifest itself when viewing at 1 MW scale of system.

The following describes an example case of a temporal resolution issue with a Grafana dashboard NREL created for the Eagle CDUs (responsible for transferring thermal energy from racks to the facility) for a period from September 19, 2019 to September 21, 2019 (Figure 7).

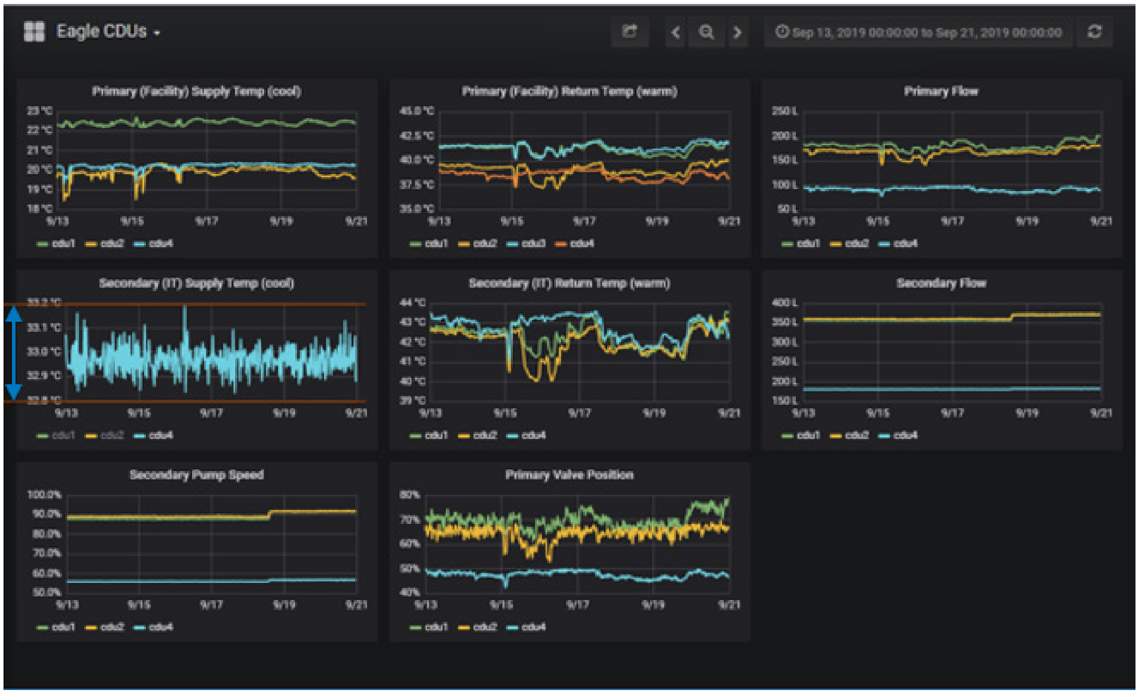


Figure 7: Eagle CDU Dashboard (top row: primary supply temperature, primary return temperature, primary flow; middle row: secondary supply temperature, secondary return temperature, secondary flow; bottom row: secondary pump speed, primary valve position)

The different graphs show the supply and return temperatures on both the facility and secondary side water loops, along with flow rates, secondary pump speed, and primary valve position. The middle graph on the left side shows the secondary water supply temperature (that is used to cool the nodes) is around 33°C. On occasion CDU 4 spikes by about 0.4°C but remains relatively stable when viewing a week's worth of data.

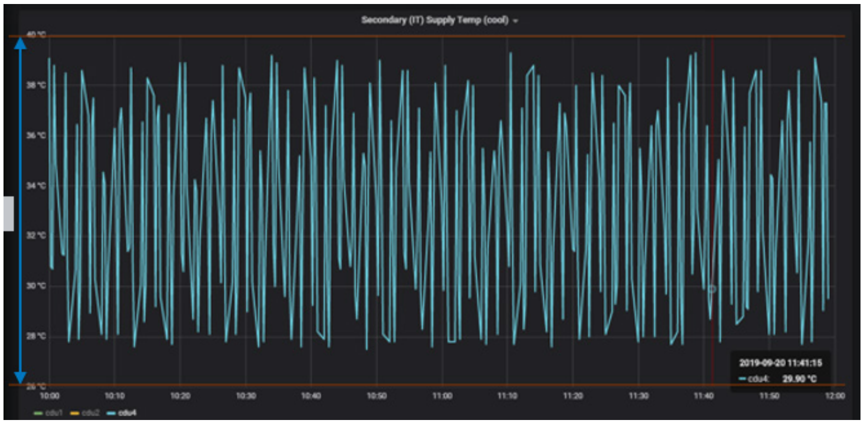


Figure 8: Eagle CDU-4 secondary supply temperature for a 2-hour period from 10:00 to 12:00 on September 20, 2019.

A more granular view of one single day (September 19) indicates that the range in temperature swings is still < 1°C. However, when the graph is resolved to display only a 2-hour period, it becomes clear that the temperature swings are actually > 10°C and up to 14°C (Figure 8), which is abnormal behavior. Further investigation by the ACO team indicated that the temperature swings were the result of a facility-side control valve within the CDU that was cycling widely every 30 seconds to 1 minute, causing the oscillating temperature swings. The behavior would last a few days and then return to normal oscillations of less than 1°C. This behavior also repeated itself over many months. The AIOps system was able to detect this anomalous behavior on the very first cycle. By contrast, if the valve had vailed, it would have caused substantial downtime for 4 racks.

4.2 Anomaly detection

The fact that this behavior in the CDUs went unnoticed for many months provided a great test case to evaluate the AIOps anomaly detection model on data from this time period. The existing aggregated dashboards missed the abnormal valve cycling behavior, but alerts from the anomaly detection model during this time period identified an issue (Figure 9). While the ACO team was aware that CDU 4 was specified to handle a full load, the initial configuration of this CDU had it loaded up to 50% of its maximum load. The valve cycled widely at times as a result of the light loading, creating a response in the resulting CDU temperature and prompting the ACO team to make the necessary adjustments so that the valve did not cycle as widely as before. While these types of adjustments are necessary to maintain optimal performance of the data center, it requires a great deal of manual effort and some luck to identify these types of system events. The anomaly detection pipeline developed by NREL and HPE facilitates the identification of these types of events through an automated, data-driven approach. Given the vast number of sensors to monitor at NREL, prior dashboards have required down-selection and manual curation of which sensors appear on the dashboards. In 2015, a 3-way valve failure that led to system shut down did not appear to be a high priority item to monitor but caused NREL to lose 20,000 node-hours in the shutdown. Motivated by this work, a key ongoing priority is automation around the monitoring and selection of sensors. This is a fundamental paradigm shift in how dashboards are built and used, allowing data center operators to monitor everything and focus on key anomalous events.

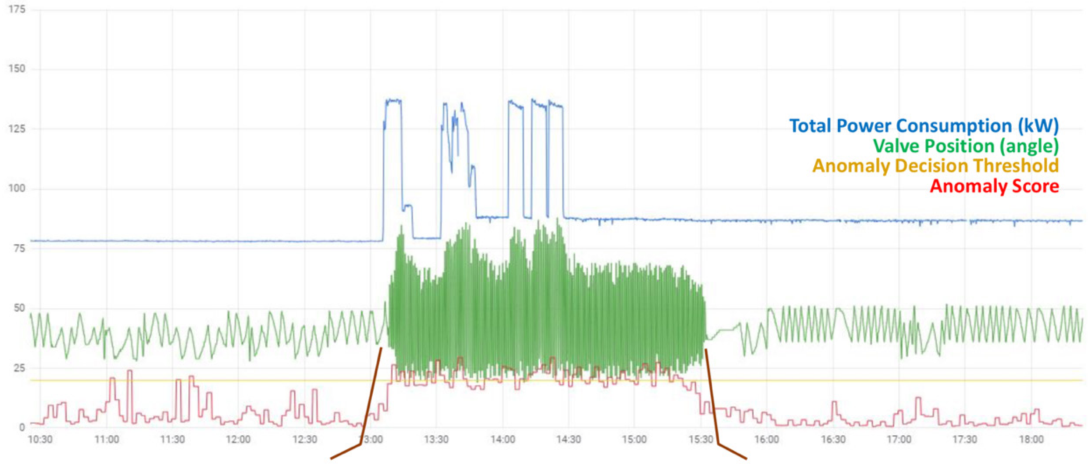


Figure 9: Example of CDU Valve cycling observed by the AIOps anomaly detection model. Blue lines indicate the total power consumption of the CDU, the green line indicates the angle of valve position, the red line indicates the anomaly score, and the yellow line indicates the threshold used to determine the presences of anomalous behavior. The thick dark red lines indicate the start and end of anomalous behavior.

4.3 Energy Balance

The consistent collection, storage, and persistence of the streaming data architecture has also allowed NREL researchers to obtain real-time estimates of the data center's energy balance. The calculation of the energy balance has been implemented in a python-based software tool, which measures the energy balance of the entire data center by grouping energy consumption metrics by resource component (e.g., Eagle, Lights & Plugs, Air-to-liquid cooling) and calculating the amount of energy that is lost to the ESIF (Figure 10). This version of the software builds upon the energy balance shown above by ingesting real-time data from the Druid database, includes all of the data from NREL's Eagle system and associated electrical subsystems, and incorporates sensors from new additions to the system.

The goal of producing a real-time energy balance is to account for all the electrical energy that is input into the data center from the grid and to account for its breakdown in usage amongst the different facility equipment, devices, and IT equipment, as well as the subsequent method of capture and rejection of that energy. This model therefore allows us to capture changes at any of the phases of energy usage, capture, and rejection and can serve as a first-level monitoring tool for the data center for any undetected anomalies in energy consumption or reuse.

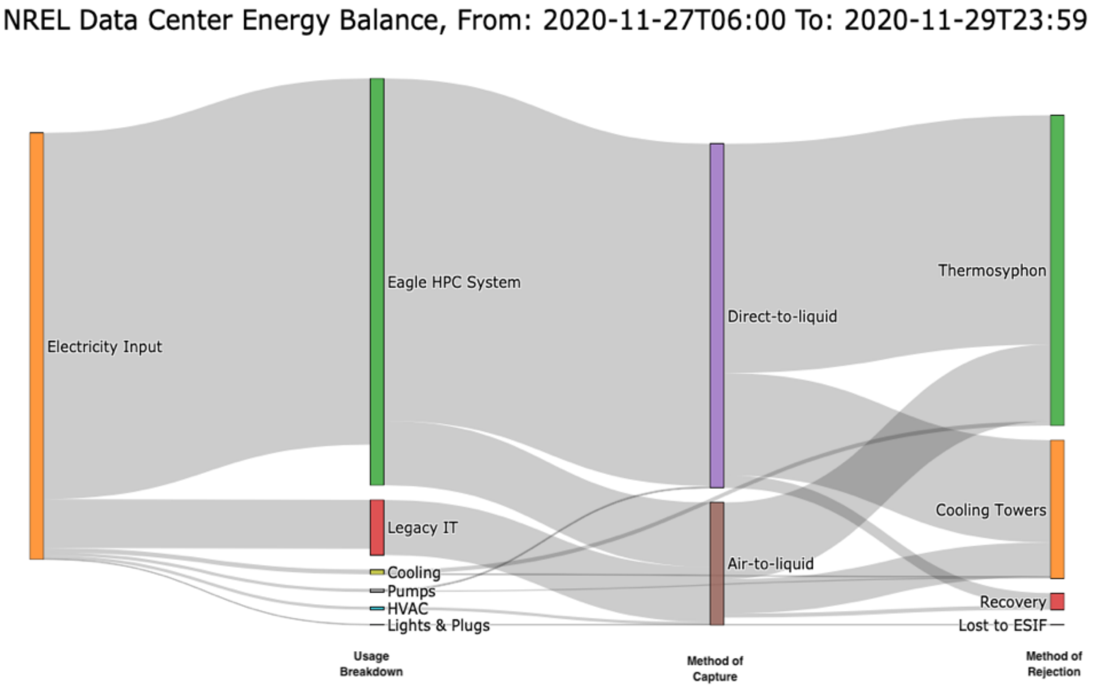


Figure 10: Sankey representation of Energy Model for the ESIF data center, demonstrating how electricity input (far left) into the data center is used (middle left), captured (middle right), and rejected (far right) as part of the data center design. Vertical bars represent the representative amount of energy used/captured/rejected by the individual components.

4.4 Application Power Usage

In addition to enabling real-time anomaly detection pipelines and improved real-time energy balance calculations, the streaming data architecture in the ESIF data center has also allowed NREL researchers to investigate the power footprint of individual jobs on Eagle and their associated cooling resource requirements. The power footprints of running jobs on Eagle gives insight into the collective energy that is consumed by these jobs and that contribute to the total IT energy consumption. Prior work at NREL (e.g., 17 and 17) has indicated that further reductions in PUE/ERE can be achieved through optimization of the timing of jobs. In that work, application-level power usage of a production system was characterized for the Peregrine supercomputer and potential methods for predicting power usage were explored based on a priori and in situ characteristics about application job type. Other related work has shown the effects of different approaches to power management, such as power throttling strategies like core parking, which for certain applications are optimal in minimizing performance losses while maximizing power savings leading to further reductions in PUE/ERE (Purkayastha et al., 2018). Other strategies like frequency reduction have also shown benefits in reducing power consumption without affecting performance for a different class of applications. These initial studies also demonstrated potential use cases of these methods through a simulated power-aware scheduler for the different approaches.

Current research being undertaken at NREL in collaboration with HPE as part of the AIOps project seeks to extend the use case of power usage prediction and to build a prototype implementation. To enable similar research in this area, a publicly available dataset[1] has been constructed consisting of three months of job data including anonymized SLURM metadata and derived node level power metrics per job. The dataset was constructed from three months of data over different seasons (December 2019, April 2020, and August 2020), selected for periods where the system was operating under nominal conditions. Jobs lasting less than five minutes, those with missing metadata, those run on GPU or special-use nodes, and those run by end use rather than users were all removed from the dataset. This dataset was then combined with executable information from XALT metadata (Agrawal 2014) to create a set of data labeled by application. An anonymized release including application labels is planned for a future release to compliment the set of publicly released data.

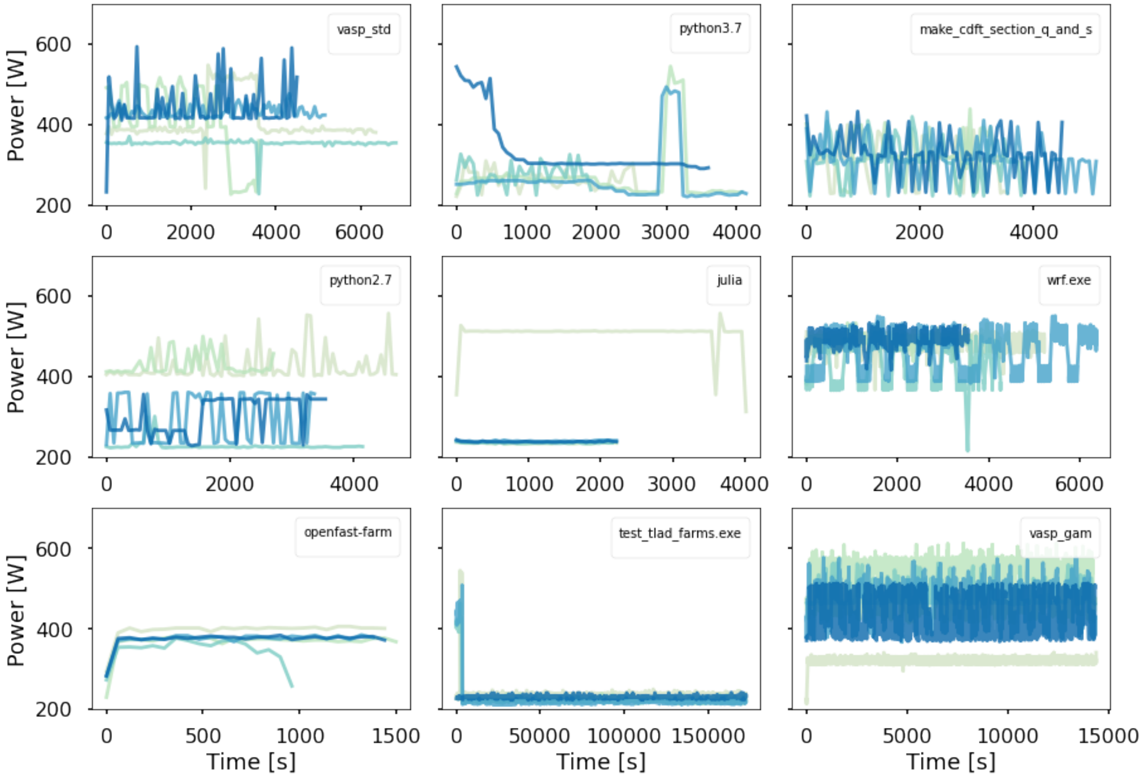


Figure 11: Three randomly sampled job power profiles per application.

Analysis of the power footprint of these jobs has indicated that there are distinct job power profiles; however, this footprint is not entirely describable by the associated job executable. Figure 11 shows three randomly sampled job power time-series for the top nine application types run on Eagle. Note that there are substantial differences within a given application (e.g., python3.7) as well as between applications (e.g., wind plant simulation (openfast-farm) and atomic structure analysis (vasp)). Many of these applications with widely varying power profiles are indicative of a very broad job type classification with many different underlying use cases. In most of these cases the mean and standard

[1] https://data.nrel.gov/submissions/152

deviation of these power time-series are fairly well defined, although others do show interesting time varying behavior (e.g., python3.7, wrf.exe).

Figure 12 shows the entire distribution of average and standard deviation of power profiles for jobs by each category.

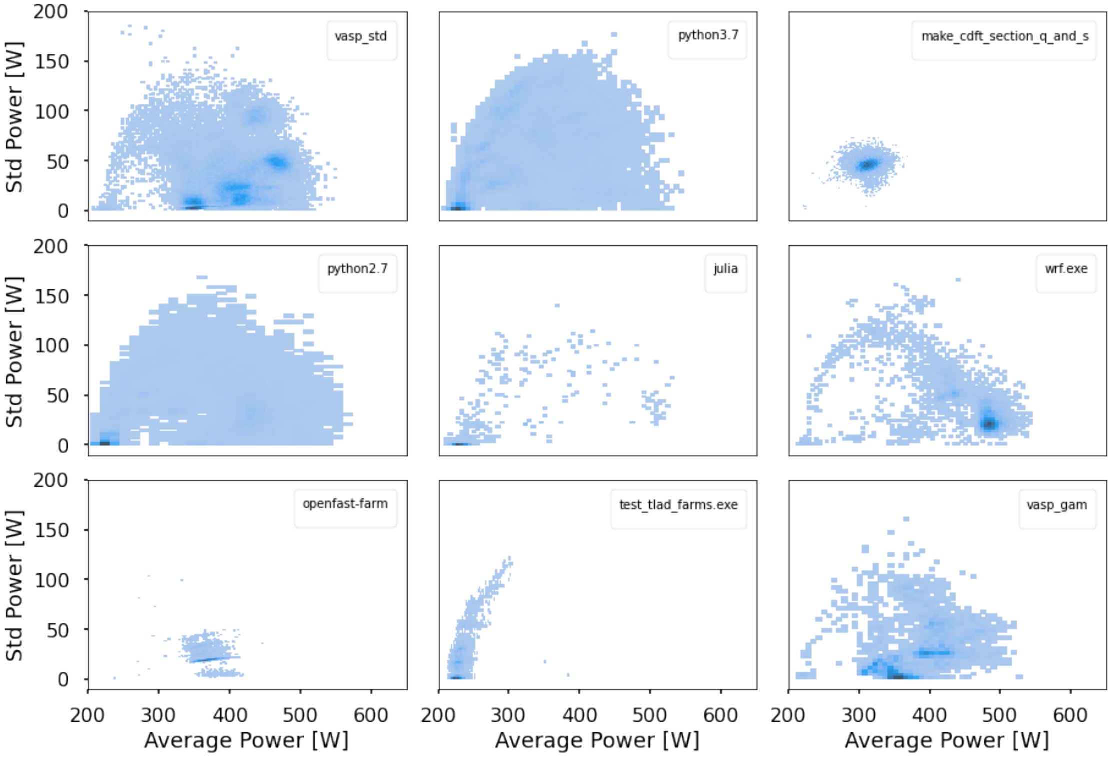


Figure 12: Histogram of all labeled application jobs within the 3 month sample period, with average job power on the x-axis and standard deviation of job power on the y-axis.

While the overall spread of the various distributions is quite large in many cases (once again indicating a broad job type classification with many underlying use cases), there are indications of centroids that are in some cases quite well defined across all jobs for a given application (e.g., make_cdft_section_q_and_s, openfast-farm). While the job executable can be a strong indicator of power footprint in many cases, a more complicated and multi-metric model would be necessary to be fully predictive.

5 Conclusions

NREL's HPC data center in the ESIF has been built with energy efficiency as a key component of its design. While the data center has maintained an impressive 12-month running average PUE of 1.06 or less, there are still opportunities to improve the energy efficiency of the center through automated, data-driven operations. NREL's collaboration with HPE through the AIOps project has thus far demonstrated some of the potential optimizations that are possible with scalable data architecture for real-time monitoring and analytics.

The ESIF data center is now equipped with state-of-the-art scalable streaming data architecture for data collection, which requires accessibility of data from facilities, compute infrastructure, and HPC jobs data streams. This design will enable NREL to continue to optimize and manage the energy efficiency of the ESIF data center and to build new capabilities for ML approaches that improve the HPC operational efficiency. Furthermore, the streaming data architecture capabilities will enable a variety of other projects across the NREL research community, which will be able to utilize modern data architectures for their dashboards and analytics.

From a monitoring perspective, the AIOps project has contributed to the development of additional dashboarding capabilities that allow the user to view real-time alerts across all of the CDU and CRC sensors via an anomaly detection pipeline. These additional capabilities will facilitate quicker response times to future failure events in the data center and detect event onsets before they become critical, minimizing downtime and potentially saving hardware systems from catastrophic failure.

Taking this collaboration forward, there are several key threads of ongoing work: (1) a power-aware scheduler for the Eagle supercomputer is being developed, which would allow power-intense jobs to be run at times of the day when renewable energy is being generated on campus or when energy reuse can be maximized. A power-aware scheduling system could be further expected to ramp up cooling resources ahead of an expectedly intensive job to further maximize energy efficiency in the data center. (2) Research is currently underway to understand the relationship between not only a job's application type and its power profile, but also that job's impact on required cooling resources. This understanding will help to better inform a predictive scheduler. (3) The team is using existing data to produce a model for prediction or forecasting PUE for the weeks or months ahead, which would allow for more optimal utility planning, among other benefits. (4) In 2021/2022 HPE will introduce AIOps Analytics, along with improvements to the AIOps Runtime stack. AIOps Analytics will include elements of root-cause analysis and trustworthy AI for more comprehensive and reliable interpretation of models' decisions. AIOps Runtime will be capable of monitoring model performance, including identification of data drift and concept drift, and capable of triggering automatic model re-training. AIOps anomaly detection operates in real-time and at scale (measured in two dimensions: metrics and ML models); it is one of the first such solutions to be introduced in data centers.

Taken together, these efforts will inform future supercomputer procurement efforts as to the type of resources that are used, how efficiently they are used, and ways that the NREL HPC community can improve its practices and help steer advancement in the design and widespread adoption of energy efficient data center practices, significantly decreasing the carbon cost of leadership class computing while also reducing maintenance costs and improving system reliability.

References

NREL, 2018: NREL Garners Top Sustainability Honor at Data Center Dynamics Awards. https://www.nrel.gov/news/program/2018/nrel-garners-top-sustainability-honor-at-data-center-dynamics-awards.html

Masanet, E., and A. Shehabi, N. Lei, S. Smith, J. Koomey, 2020: Recalibrating global data center energy-use estimates. *Science*, **367** (6481), 984-986. dx.doi.org/10.1126/science.aba3758

Ayanoglu, E., 2019: Energy Efficiency in Data Centers. *IEEE ComSoc*. https://www.comsoc.org/publications/tcn/2019-nov/energy-efficiency-data-centers

Shehabi et al. 2016: https://eta.lbl.gov/publications/united-states-data-center-energy

Lawrence, A. 2019: Is PUE actually going UP? Uptime Institute. https://journal.uptimeinstitute.com/is-pue-actually-going-up/

Sickinger, D., O. Van Geet, S. Belmont, T. Carter, and D. Martinez, 2018: Thermosyphon Cooler Hybrid System for Water Savings in an Energy-Efficient HPC Data Center: Results from 24 Months and Impact on Water Usage Effectiveness. Tech. Rep. NREL/TP-2C00-72196. National Renewable Energy Laboratory.

Bugbee, Bruce., C. Phillips, H. Egan, R. Elmore, K. Gruchalla, A. Purkayastha, 2017: “Prediction and characterization of application power use in a high-performance computing environment”. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10 (3): 155–165. doi:10.1002/sam.11339. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11339.https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11339.

Elmore, R., K. Gruchalla, C. Phillips, A. Purkayastha, N. Wunder, 2016: An Analysis of Application Power and Schedule Composition in a High Performance Computing Environment. Tech. Rep. NREL/TP-2C00-65392. National Renewable Energy Laboratory.

Carter, T., Z. Liu, D. Sickinger, K. Regimbal, and D. Martinez, 2017: Thermosyphon Cooler Hybrid System for Water Savings in an Energy-Efficient HPC Data Center: Modeling and Installation. Tech. Rep. NREL/TP-2C00-66690. National Renewable Energy Laboratory.

Purkayastha, A., S. Hammond, R. Nagappan, and M. Alt, 2018: “Holistic Approaches to HPC Power and Workflow Management”. In Proceedings for *The 9th International Green and Sustainable Computing Conference*.

Agrawal, K., M. R. Fahey, R. McLay, and D. James, 2014: User Environment Tracking and Problem Detection with XALT, *Proceedings of the First International Workshop on HPC User Support Tools, HUST '14*. dx.doi.org/10.1109/HUST.2014.6.