Operating Liquid-Cooled Large-Scale Systems: Long-Term Monitoring, Reliability Analysis, and Efficiency Measures[*]

Rohan Basu Roy, Tirthak Patel
Northeastern University

Raj Kettimuthu,William Allcock
Argonne National Laboratory

Paul Rich, Adam Scovel
Argonne National Laboratory

Devesh Tiwari
Northeastern University

***Abstract*—The past decade has seen a rise in the use of liquid cooling due to its energy efficiency. While many previous works have helped make progress toward improving data center cooling, a vast majority of them perform studies on a small system over a short span. The computer systems and HPC community lacks a long-term study highlighting the challenges and solutions in operating a liquid-cooled large-scale data center. We conduct the first detailed characterization of a petascale supercomputer, Mira, over a span of six years. The study is enabled by systematic monitoring of the environmental metrics, and discusses new research avenues, including coolant monitor failures.**

I. Introduction

A high-performing computing data center incurs a considerable energy cost to power up the cooling infrastructure. A significant amount of the operational and maintenance effort is also geared toward keeping the cooling systems running smoothly [6, 13, 62]. The last few years have seen an increase in the use of liquid cooling as the method of choice for cooling supercomputing data centers, especially with the increased requirement for energy-efficient cooling for upcoming exascale computing systems [56]. The high effectiveness of liquid cooling stems from the fact that the liquid water, with a higher density than air molecules, flows adjacent to the hardware, absorbing heat more efficiently than air cooling [6, 13, 62]. This enables liquid cooling to be more energy-efficient, especially when combined with the characteristics of surrounding environmental and climate factors to deliver free data center cooling [6, 13, 62].

Previous research in the area of data center cooling has focused on characterizing the energy efficiency and cooling effectiveness of air-cooled systems [50], proposing different infrastructural and alternative-technology mitigation strategies to make cooling more functional and effectual [10, 31], studying the impact of air quality on the failure rate and reliability of hardware components in an air-cooled (free cooled or otherwise) data center [15, 31, 43, 49]. However, environmental and other factors affecting the efficiency of a liquid-cooled system and their impact on the overall system performance can be significantly different.

To this end, we present the first study that monitors and characterizes the operations of a liquid-cooled production, petascale supercomputer over a span of six years, observing trends in power consumption, utilization, component temperatures, humidity, and failures. We study the 10 PFlops leadership-class IBM Blue Gene/Q Mira supercomputer, located at the Argonne Leadership Computing Facility (ALCF) in Chicago, Illinois, which is cooled using two 1,500-ton water chiller towers with the ability to perform outdoors-temperature-leveraged free cooling when the weather of Chicago permits. We identify several new trends and challenges that need further research effort in academic experimental lab setting to design effective solutions. For example,

1) Even highly utilized production data centers can observe transient fluctuation in utilization due to various scheduling and resource allocation policies. This, in turn, can cause power consumption fluctuations. Opportunistically utilizing the transient resources that become available on capability machines for on-demand and elastic HPC jobs will be the key to avoiding power utilization swings.
2) When new systems are added to an existing data center, it can pose new challenges due to the sharing of the cooling loop. Research prototypes should model and simulate effects from other systems in the data center to demonstrate its potential effectiveness in production environment. They should demonstrate the ability to adapt to changes in the chillers/cooling loop and avoid accidental system outages.
3) At ALCF, data center operators put considerable effort toward achieving homogeneous and adaptive coolant flow rate across the data center – which significantly improves the operational efficiency. Further efforts are needed to monitor and manage the coolant flow rate effectively in real time to identify challenges posed by factors outside the compute cluster. More experimental research effort is

[*] A part of this work was performed during an unprecedented time – around the peak of the COVID-19 pandemic. We were able to conduct this scientific study only because first-line responders and essential workers worked tirelessly to keep our community safe and functioning – we are eternally grateful to them. This paper is dedicated to the memories of all the first-line responders and essential workers who sacrificed their lives trying to keep ours safe.

needed toward quantifying the effects of extended humidity variability on component reliability and application performance via in-house controlled testing in academic labs.

4) We present a detailed analysis of **coolant monitor failures (CMF)** - failures that affect multiple racks and kill hundreds of jobs within a short duration. CMF have not been well characterized in the past in academic literature and hence, mitigation strategies are not mature despite the severe side-effects – it can induce a chain of non-CMF failures too. We provide detailed pointers about what aspects of CMF need better understanding and provide a learning-based solution for predicting CMFs.

II. Background

In this section, we provide a brief overview of the Mira system's computing and cooling infrastructures.

Mira Overview. Mira is a 10 PFlops Blue Gene/Q (BGQ) production system that was stationed at the Argonne Leadership Computing Facility (ALCF). Mira was operational from 01/01/2014 to 31/12/2019, primarily supporting Impact on Theory and Experiment (INCITE) projects, the Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC) projects, and other discretionary projects.

Mira consists of 48 racks, each with two midplanes, each midplane containing 16 node boards. Each node board has 32 compute cards, totaling to 1,024 nodes per rack and 49,152 nodes in the entire system. Each BGQ computation card has a 1,600 MHz PowerPC A2 processor with 18 cores each, where each of the cores can run four hardware threads at 1.6 GHz, and 16 GB of DDR3 memory. However, only 16 cores are available for computation, and the others either remain inactive or are used for communication libraries. Overall, Mira is equipped with 786,432 active cores, 768 TB of memory, with a peak performance of 10 Pflops. Mira is connected throughout by IBM 5D torus interconnect with two GB/s chip-to-chip linkage, which reduces communication latency by minimizing the average number of hops between nodes. Mira has access to a 24 PB file system with a 240 GB/s bandwidth [73].

The 48 racks of Mira are divided into 3 rows, each containing 16 compute racks. In addition to this, each row has two racks of I/O forwarding nodes (IONs), located at the end of each row (for a total of six racks of IONs). This whole infrastructure encompasses 1632 square feet of area and is located in the Theory and Computational Sciences (TCS) Building at Argonne National Laboratory, Chicago, IL. Along with Mira, ALCF also hosts other computing systems such as Theta, a 12 PFlops large-scale production system, and Cooley, a small-scale data analysis and visualization cluster.

Power system. Mira has built-in support for six MW of power draw, but has an average load of four MW. Each rack of Mira is associated with a Bulk Power Module (BPM). It converts AC to DC power and distributes it among the two midplanes of a rack. The BGQ system gets the power from 13.2 kV substations, which supply power to the BPMs via Eaton Digitrip Optim 1050 distribution channels. Each BPM has four 480V three-phase, 60A line cords which are located on the left side of each rack. These cords separate the power distribution BPMs for each of the racks and reduce electrical distribution by about 75% [6, 66].

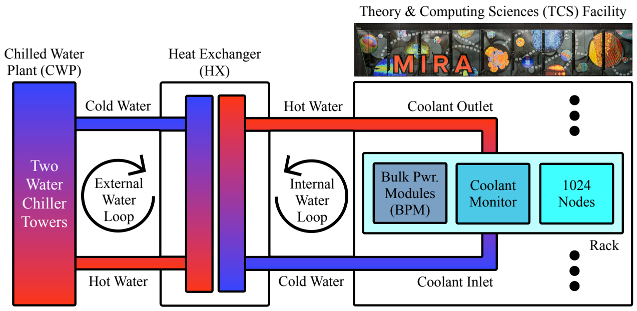
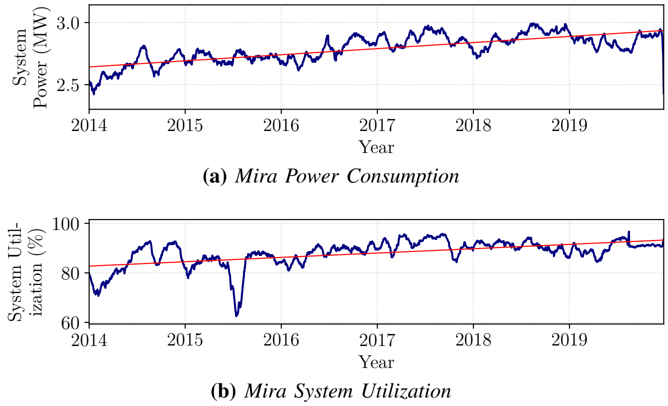


Fig. 1. *Design of Mira's liquid cooling system including the internal and external coolant water loops.*

Coolant system. A significant amount of power consumed by Mira is utilized for the cooling of the system. The BGQ compute racks of Mira undergo liquid cooling. However, other associated infrastructures, including the IONs, are air-cooled. Two 1,500 ton water chiller towers were built at the Argonne Chilled Water Plant (CWP) to support Mira's cooling, as shown in Fig. 1. The CWP is located in a building adjacent to Mira. A closed liquid cooling process loop extends from CWP to under the floor of the data center where Mira is housed. Chilled water from CWP runs through the loop, forming an external water loop. Each rack maintains a separate internal water loop that runs across the walls of the rack to cool the system. Under the floor of each rack, the internal and external water loop forms a junction known as a heat exchanger (HX). The chilled water from the external loop cools up the hot water in the internal loop. The heat generated in the internal loop due to system usage is dissipated to the external loop. The external loop again cools down the water using the chillers.

CWP has a waterside economizer design to induce free cooling capabilities when the weather is favorable. The chiller towers installed in CWP are over-sized to make room for the water generated through free cooling. 17,820 kW-hr energy can be saved per day if 100% of CWP capacity can be produced by the free cooling modules. This potentially saves 2,174,040 kW-hr of energy by not operating the chillers during the colder months (December - March) when the surrounding temperature allows for free cooling [6, 14, 77].

Coolant monitor. BGQ systems have several associated modules to periodically sample environmental sensor data and store them in IBM DB2 environmental database [40]. In this paper, we particularly focus on the *Coolant Monitor* module, present in each of the racks of Mira. It is a module of sensors present beside the coolant outlet and coolant inlet lines of the internal water loop of each rack. The coolant monitor collects rack-level data at a granularity of 300 sec. The monitor also stores the calibration data used to calibrate the sensors. It collects the following sensor data: (1) *data center temperature*, (2) *data center humidity*, (3) *coolant flow rate*,



(a) *Mira Power Consumption*

(b) *Mira System Utilization*

Fig. 2. *The power consumption and system utilization of Mira have increased over the years. The red lines show a linear fit.*

(4) *coolant temperature* (inlet and outlet), and (5) *power*. The temperature and humidity values denote data center conditions near the rack (not node level). The flow rates measure the rate at which water flows through the outlet and inlet lines, and the coolant temperature measures the respective coolant temperatures of the outlet and inlet ports. Power data denotes the aggregate power drawn from all four power enclosures in a rack. The enclosures supply power to the computation nodes and the fans in the power module. The coolant monitor sensors were regularly tested and validated to ensure accurate measurements. Only one sensor (on one rack) was replaced during the six years as it malfunctioned.

The coolant monitor sets alarm thresholds for measured sensor values. If a sensor reading crosses the relevant set threshold, a *Coolant Monitor Failure (CMF)* event is recorded in the *RAS* dataset of Mira. The RAS logs record events which affect the reliability, availability, and serviceability (RAS) of Mira. The severity of a failure event can be *warn* (designating low-risk situations) or *fatal* (identifying a severe error event that leads to a rack-level failure). Apart from coolant monitor failures, the RAS log also captures failures due to BPMs, ethernet adapter cards, BGQ computation cards, and link modules, among others.

III. Identifying and Investigating Temporal System-Level Trends

In this section, we identify and investigate the temporal characteristics of cooling-related parameters, the factors that affect them, and the lessons learned from them.

A. Year-over-Year Trends

Fig. 2(a) and (b) show the increase in the power consumption of Mira and the increase in its utilization, respectively. The system-level power consumption of Mira has increased from ≈2.5 MW at the beginning of 2014 to ≈2.9 MW near the end of 2019 with many fluctuations in between which depend on the system utilization, system failures and crashes, and power outages and blackouts. However, the general trend has been toward increase in power consumption as is shown by the linear fit (red line).

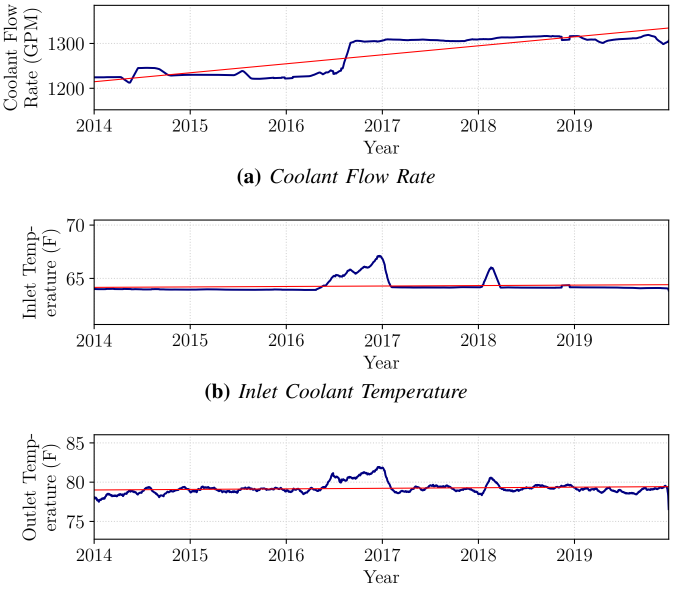
Correspondingly, the system-level utilization of Mira has also increased from ≈80% to ≈93%. At any instant, system utilization is defined as the percentage of nodes on which jobs are running, out of the total number of nodes present (49,152 in case of Mira). Utilization has varied considerably over the years. Under normal operating conditions, Mira maintains a high utilization. However, drop in utilization occurs frequently at both longer and smaller time period for various reasons and has implications for ensuring high operational efficiency.

Although the queue of the high-utilized Mira supercomputer always has a large number of jobs waiting to be scheduled, a transient drop in the utilization can occur (1) when racks are reserved for projects which end up not using them fully, (2) when rack-level or system-level failures occur, (3) when the system has to wait to finish all the smaller jobs to accommodate the occasional large job which runs at full or near-full capacity. Unfortunately, these drops in utilization can also cause undesired power utilization fluctuations.

Opportunity: There is a need to develop more robust methods to "fill in" the idle nodes waiting for a large-job to start. State-of-the-art back-filling job scheduling strategies [39, 45] may not be able to fill all such holes – leading to resource wastage. Although challenging, an opportunity for making traditional HPC jobs more elastic to fill such holes exists [22], and more efforts are needed to utilize these transient resources on a capability machines for on-demand HPC jobs or serverless computation [47, 69]. Similarly, smooth rebooting of large-scale system, often under-investigated, remains a challenging avenue [44].

Next, Fig. 3(a), (b), and (c) show the changes in coolant flow rate, inlet coolant temperature, and outlet coolant temperature from 2014 to 2019, respectively. Mira's coolant flow rate was maintained at around 1250 GPM (≈26 GPM per-rack) until July 2016 at which point it was to increased to about 1300 GPM. The reason for this increase was the addition of Theta supercomputer which shared its internal water loop with Mira. Mira's water loop was installed with the necessary valves and stubs to allow additions in a manner which would not affect its exiting water loop until Theta's piping installation was complete and the final connection needed to be made. Mira and Theta's individual coolant flows were controlled using a flow regulating valve. Therefore, to prevent accidental shutdowns of Mira, the impellers on the coolant loop were upgraded when Theta was added to the loop and the flow rate of coolant to Mira was increased.

Similar to the coolant flow rate, the inlet coolant temperature and the outlet coolant temperatures have remained consistent throughout the years at ≈64 F and ≈79 F, respectively, with only a few exceptions. For example, both the inlet and the outlet coolant temperatures rise in June 2016 and remain high until early 2017. Again, the reason for this increase is the addition of Theta to Mira's water loop which generated higher heat load as Theta was in early testing until the end of 2016.



(a) *Coolant Flow Rate*

(b) *Inlet Coolant Temperature*

(c) *Outlet Coolant Temperature*

Fig. 3. *While the coolant flow rate was increased when Theta cluster was added in 2016, barring a few times during the six years, the inlet and outlet temperatures have remained generally consistent. Coolant flow rate, inlet coolant temperature, and outlet coolant temperature have an overall standard deviation of 41 GPM, 0.61 F, and 0.71 F respectively.*

> **Summary:** When new systems are added to an existing data center, it can pose new challenges due to potential sharing of the chillers/cooling loop. Sufficient attention must be given to regulating the coolant flow rate to prevent accidental system outages and the thermal profile of the data center needs to be re-done.

B. Monthly and Daily Trends

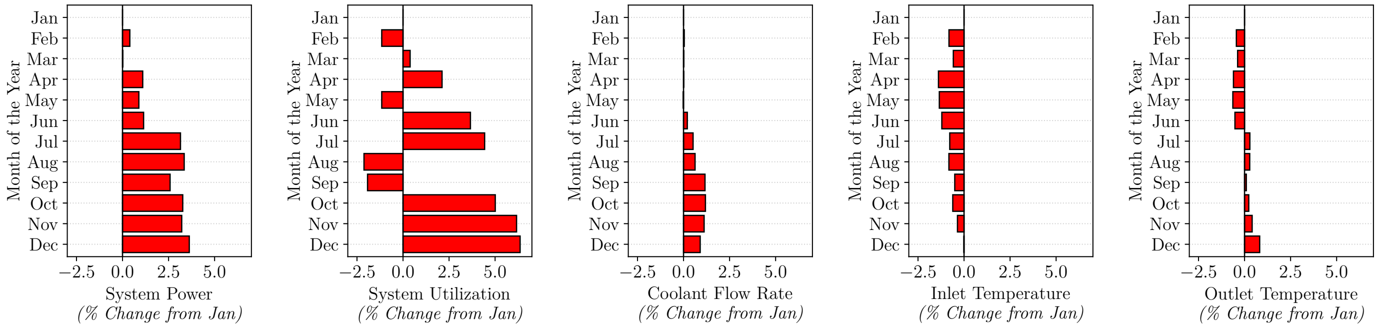
While the year-over-year analysis revealed some interesting results, some parameters show patterns on a finer scale. Therefore, next we present results pertaining to how the parameters vary on monthly and daily basis.

Starting with the per-month granularity, Fig. 4(a) shows that while the power consumption remains low in the months of January through June, it increases significantly in the later half of the year, achieving its peak in December. The reason for this trend is primarily due to the the monthly variation of utilization as shown in Fig. 4(b). As described in Sec. II, Mira primarily supports projects related to the ALCC and INCITE programs. In order to avoid severe resource contention, ALCC projects have their project allocation year from July 1 to June 30 of next year and the INCITE projects have their allocation year from January 1 to December 31. An allocation year is the period during which the units (core hours) allocated to a project at the beginning of the allocation year must be used up by the end of it. Because users tend to run a majority of their jobs near the end of their allocation year deadline to consume all of the allocated core hours, setting the allocation years in the above manner allows the two programs to run their jobs with less contention. However, the INCITE projects tend to be of higher priority and have higher resource requirements, therefore, the system utilization increases during the second half of the year when INCITE projects are run.

> **Opportunity:** Designing the allocation years in this manner has been an effective strategy to avoid resource contention, but it does not always help in terms of the variability of utilization as the resource usage behavior of the projects belonging to the two programs is very different. This can cause the variability in utilization, which has an impact on the variability of the system's power consumption. New innovative mechanisms are needed to avoid variability in system power consumption caused because of resource allocation policies in HPC data centers (e.g., incentive-based trading of compute core hours among users [58] and potentially different project types to smoothen the load and provide fairness [26, 57]).

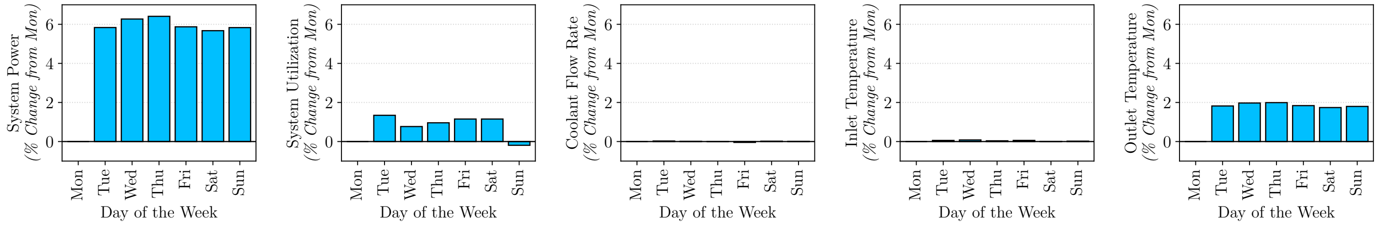
Since higher utilization demands more cooling efforts to regulate the CPU temperatures, the coolant flow rate increases slightly from June to the end of the year (Fig. 4(c)). Due to the water side economizer design of the Chilled Water Plant (CWP) to save power, the chillers remain partially or fully non-operational during the colder months of Chicago (December - March). At this time, the environmental temperature aids with the cooling. Since, environmental cooling is not as effective as cooling by the chillers, the inlet coolant temperature is slightly higher in the colder months than relatively hot months (when the chillers remain operational) as shown in Fig. 4(d). The outlet coolant temperature shows trends which can be directly explained from the system utilization characteristics (Fig. 4(e)). It rises in July and near the end of the year due to the high utilization during those periods. It is minimum in April and May due to relatively low utilization and full operation of the chillers during those months.

The next granularity of behavior we analyze is the daily variation of system parameters during a week, as shown in Fig. 5. First, we observe that the system power is minimum on Mondays (Fig. 5(a)). This is because utilization is also low on Mondays (Fig. 5(b)) due to maintenance activities. They typically last for 6 to 10 hours starting from 9 AM. Though the maintenance does not need to be scheduled every week, but given the high utilization of the machine at all other times, even a small unavailability on Mondays shows up in the temporal trends. During this time, the system does not run any jobs from the users. Instead, burner jobs which perform no useful computation, are run across the system for health monitoring. *This is also done to prevent hardware damage due to excessive cooling when a rack is idle. We found that the cold inlet coolant temperature can damage inactive CPUs and increase the failure rate of nodes when they become busy again, causing them to crash. Therefore, as a workaround, burner jobs were run on the racks to burn cycles.* The damage was caused in many cases due to



(a) *Mira Power Consumption* **(b)** *Mira System Utilization* **(c)** *Coolant Flow Rate* **(d)** *Inlet Coolant Temp.* **(e)** *Outlet Coolant Temp.*

Fig. 4. *The median power consumption and system utilization of Mira are higher during the second half of a year as compared to the first half due to the way that the allocation units are allocated to the projects. Other metrics including the coolant flow rate, inlet coolant temperature, and outlet coolant temperature show less than 1.5% change from January across months.*

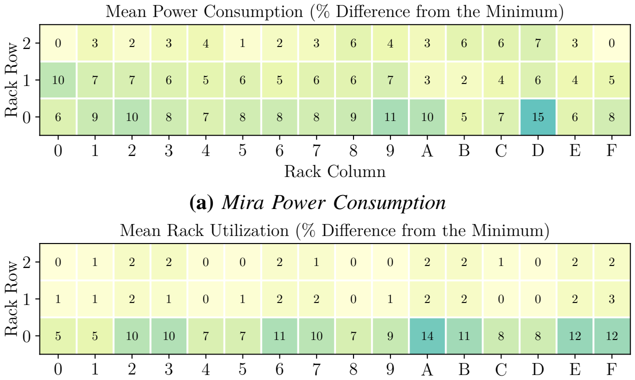


(a) *Mira Power Consumption* **(b)** *Mira System Utilization* **(c)** *Coolant Flow Rate* **(d)** *Inlet Coolant Temp.* **(e)** *Outlet Coolant Temp.*

Fig. 5. *The power consumption of Mira increases by ≈6% on days other than Mondays even though the utilization only increases by ≈1.5% on days other than Mondays. The outlet coolant temperature also shows a 2% increase on non-Monday days, while coolant flow rate and inlet coolant temperature observe no difference. Mira has scheduled maintenance on Mondays which causes a drop in utilization.*

optimal interconnect misalignment of the nodes on reboot, but in general, it was difficult to reproduce, but required a heavy investment of efforts to mitigate the post-effects. Note that the system still consumes considerable amount of power during maintenance tasks. Interestingly, while the utilization increases by ≈1.5% on days other than Mondays, system power consumption increases by ≈6%. As expected, the outlet coolant temperature increases ≈2% on non-Mondays due to the increased utilization (Fig. 5(e)). However, the near-constant flow rate is maintained on Mira on all days with inlet coolant temperatures also being highly consistent regardless of the day of week (Fig. 5(c)-(d)).

> **Summary:** In production systems, keeping compute nodes "warm" is a critical goal, even during maintenance. Further, a slight change in utilization can result in undesirable swing in the power consumption.



(a) *Mira Power Consumption*

(b) *Mira System Utilization*

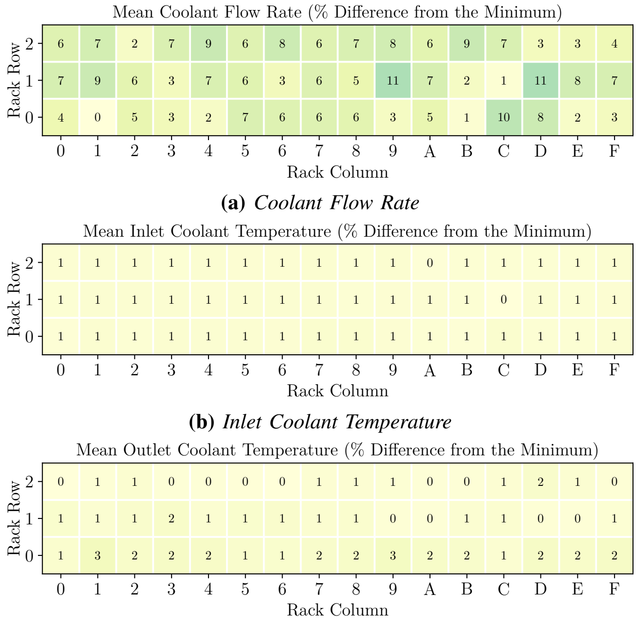
Fig. 6. *The power consumption and utilization are highest in row 0 of racks. However, the power consumption is the highest on rack (0, D), while utilization is the highest on rack (0, A).*

IV. Identifying and Investigating Spatial Trends

Next, we describe how the cooling related metrics vary across the 48 racks of Mira averaged over the full duration.

A. Rack-Level Power Usage & Utilization

First, we study the variation of average power consumption and utilization across the racks of Mira, respectively (Fig. 6(a) and (b)). During the whole production period, power varies significantly among the racks (up to 15 %) as seen in Fig. 6(a). Row 0 of racks always has the highest utilization (Fig. 6(b)) as longer jobs are allocated racks from row 0 (jobs submitted to the *prod-long* queue of Mira). The queue allocation of Mira causes row 0 of racks to have a higher power consumption than the other two racks.



(a) *Coolant Flow Rate*

(b) *Inlet Coolant Temperature*

(c) *Outlet Coolant Temperature*

Fig. 7. *The coolant flow rate varies considerably with a difference of up to 11% among the racks. On the other hand, the inlet and outlet coolant temperatures remain consistent across the racks with maximum difference of 1%-3%.*

While the racks with higher utilization generally have higher power consumption, this correlation is not always direct and one-to-one. For example, rack (0, D) has the highest power consumption, while rack (0, A) has the highest utilization. Another example is rack (2, D), which has a power consumption of 7% more than the minimum, while the same rack has the lowest utilization among all. In fact, the correlation coefficient [52] between average power consumption and utilization of each rack is only 0.45 (on a scale of -1 to 1, a correlation of 1 indicates perfect positive correlation, 0 indicates no correlation, -1 indicates perfect negative correlation).

The reason for this mismatch is that the CPU load of a rack depends on the characteristics of the jobs running on it. A rack gets hotter when it is running CPU intensive jobs. However, such application-level information is not monitored due to potential operational interference with production jobs and hence, is not shown in this paper. Therefore, the correlation between power consumption and rack utilization is lower than expected. Nonetheless, the significant variation in power consumption implies that some racks are running jobs which utilize the allocated cores much more efficiently than others causing an imbalance in the power consumption of racks. Since long running jobs (submitted to *prod-long* queue and allocated in row 0) usually do not underutilize the allocated nodes, both utilization and power consumption are highest in row 0. Other hotspots of utilization can be attributed to certain users submitting high number of compute jobs to certain specific regions, such as nodes in columns 2, 6, A, and B. Users who do this on a prolonged basis were contacted and asked to refrain from overburdening specific racks. Note that apart from utilization, power consumption is also affected by the ability of the cooling system to cool a rack, which again can be determined by the flow rate and coolant temperatures.

B. Rack-Level Coolant Monitor Telemetry

Next, we discuss other cooling-related metrics and how they affect the power consumption of the racks. First, we look at the average coolant flow rate across the racks in Fig. 7(a). The coolant flow rate varies considerably across the racks with up to 11% difference between the minimum and the maximum. *Underfloor cables from the chillers in CWP to the racks can undergo partial blockage in the pipes and the filters due to the complex cable layout, space constraints, and various maintenance tasks. This results in variability in coolant flow rate as measured by the coolant monitors attached with each rack. This variability can have an impact on the power consumption of the racks [57, 60]. To be on the safe side, data center raises the coolant flow rate.*

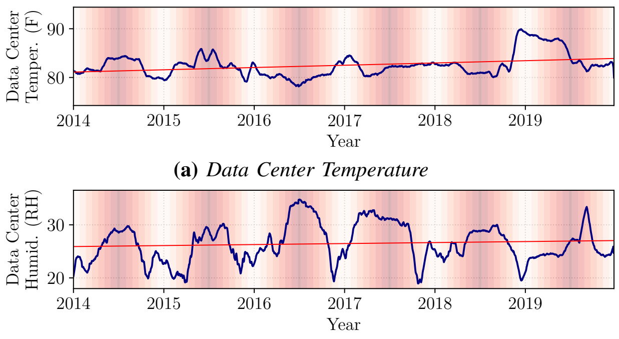
Though the flow rate varies from rack to rack, the chillers maintain a highly consistent coolant temperature and hence the inlet coolant temperature remains almost the same across the racks (Fig. 7(b)). Moreover, while the outlet coolant temperature does not vary as much as the power consumption, it does vary more than the inlet coolant temperature with a difference of up to 3% between the minimum and the maximum (Fig. 7(c)).

Opportunity: Data center operators often conservatively increase the coolant flow rate. Further efforts are required to monitor and manage the coolant flow rate effectively in real-time to identify the challenges posed by factors outside the compute cluster.

V. Ambient Data Center Temperature and Humidity Analysis

Next, we explore the trends in data center temperature and humidity, which, while not directly related to the liquid-cooling system, are important for studying temperature and humidity hotspots. First, we look at the ambient data center temperature and humidity trends from 2014 to 2019 in Fig. 8 (a) and (b), respectively. The figure shows a relatively high variability with temperature varying from 76 F to 90 F and relative humidity varying from 28 RH to 37 RH. The variation in humidity is highly seasonal with increased levels of humidity during the summer months (as shown by the red area in the plot). The reason for this is that ambient data center humidity of the TCS building where Mira is located is affected by the outdoor humidity of Chicago, which is lower in winter months due to the dryer air. However, the data center temperature is regulated with air cooling and therefore, experiences less variability except in the event of power outages, air-cooling system failure, and extreme weather incidents.

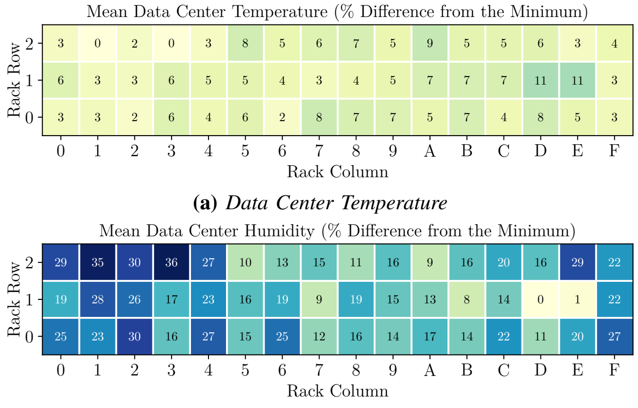
Keeping these temporal trends in mind, next, we study the variability of data center temperature and data center humidity from rack to rack in Fig. 9. Both metrics, but especially



(a) *Data Center Temperature*

(b) *Data Center Humidity*

Fig. 8. *The data center temperature and humidity have varied considerably on a seasonal and year-to-year basis. Humidity especially increases during the summer months (summer months are indicated by the redder area in the figure). Temperature and humidity have an overall standard deviation of 2.48 F and 3.66 RH respectively.*



(a) *Data Center Temperature*

(b) *Data Center Humidity*

Fig. 9. *Data center temperature and especially, data center humidity vary from rack to rack. In fact, data center humidity can have a difference of up to 36% across different racks.*

humidity, vary considerably from rack to rack. In fact, data center humidity can have a difference of up to 36% across different racks while data center temperature can have a difference of up to 11%.

The root cause of this spatial variation is the underfloor airflow characteristics underneath the racks. We found that the air flow is significantly lower near the ends of each row as opposed to the center due to the presence of obstructive surfaces. This causes the humidity to be lower and the temperature to be higher near the last three or four racks on either sides of all three rows. However, localized humidity hotspots, such as rack (1, 8), which is in the center of row 1, also exist. This is due to the presence of airflow blocking objects such as plumbing pipes, air cooling vents, and torus cables, underneath those racks. Thus, the lack of rack of air flow regulation cannot cause the racks to be affected by outdoor weather patterns but also by inadvertently created hotspots.

An important implication of this variability is that different racks can potentially observe different levels of failure rates due to this difference in humidity levels in their surrounding environment. High humidity, vibration and temperature have been shown to increase the error failure rate of data center hardware [5, 15, 17, 46, 49, 68], although in Mira's case, such a correlation was not found.

Summary: Despite best efforts and significant staff time investment, uneven humidity and temperature distribution can continue to pose a challenge for production systems due to multiple external factors which can not be always controlled [5, 46].

Opportunity: More research effort is needed toward (1) quantifying the effect of extended humidity variability, vibration, and temperature on component reliability and application performance via in-house controlled testing in academic labs, and (2) low-cost mitigation strategies for such effects and its cost analysis. Data center staff time and effort invested toward resolving transient humidity and temperature variability could be saved if its costs outweigh the savings.

VI. Coolant Monitor Failures and Its Impact

In the previous sections, we characterized the time-dependent and rack-dependent characteristics of different metrics. A lot of these characteristics are impacted by the failures related to the cooling system – an important type of failure but often under-investigated in computer systems academic literature [3, 11]. Next, we study the characteristics of failures related to the cooling infrastructure and how they influence other system characteristics. Before we begin our analysis we first describe our methodology.

Methodology: The type of coolant monitor failures (CMF) that we analyze are "fatal" failures which cause at least one of the racks to shut down. Note that other types of fatal failures are possible; for example, uncorrectable memory errors causing a job to crash. Here on, a "failure" refers to a CMF unless otherwise stated.

The Blue Gene/Q architecture performs two main control actions for a rack in the event of a coolant monitor fatal failure: (1) close the solenoid valve to cut off coolant flow, and (2) shut off the power supply. This failure is triggered when the dewpoint temperature, which is a composite metric consisting of data center temperature and data center humidity, falls below or becomes almost equal to the data center temperature, resulting in increased possibility of condensation on the electronic hardware. This can damage the electronic components and in extreme cases, cause data center outages.

When a coolant monitor failure (CMF) takes place on a rack, it is followed by a cascade of other failures on the rack that is the epicenter as well as some or all of the 47 other racks on Mira (referred to as "RAS storms"). The actual number of failure messages logged by the coolant monitors during such storms can be upwards of 10,000 failures. However, because the rack shuts down after the first failure, we do not consider

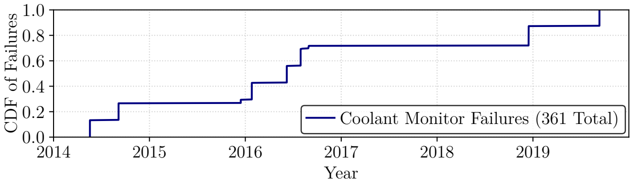


Fig. 10. *Mira has experienced 361 total failures over the six years, with many of them happening during the time when Theta was being added to its water loop (2016).*

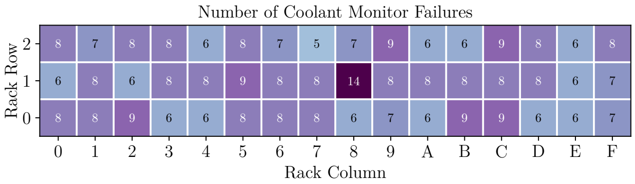


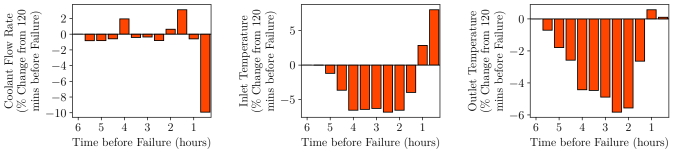
Fig. 11. *The number of CMFs experienced by a rack varies from five failure events (rack (2, 7)) to 14 failure events (rack (1, 8)). No other rack has more than nine failures.*

the other failures as new failures. It can take up to six hours for a rack to come back up, therefore, we ignore all other CMFs on the same rack within a six hour window from the first failure. We do this for individual racks and not for the whole system in order to capture information about how many racks crash once a CMF occurs. For example, if 1000 CMFs take place on eight racks within six hours of each other, we consider these as eight failures and not one failure in order to capture that fact that eight racks were affected as opposed to just one rack or all 48 racks.

Opportunity: Academic researchers have devised strategies to mitigate traditional software and hardware related failures (e.g., memory errors, voltage fault) that impact a few jobs at once and often have limited cascading effecting [18, 24]. In contrast, CMF kills jobs running on at least one rack and multiple racks in most cases. Unfortunately, there is limited understanding of CMF characteristics and mitigation strategies despite its much more severe effects (e.g., killing hundred of jobs in a small span of time).

A. Frequency and Locations of Coolant Monitor Failures

We begin by looking at all coolant monitor failures which took place from 2014 to 2019 in Fig. 10. Mira has experienced 361 total coolant monitor failures during its six years across all the racks. These failure events are not isolated rack-level incidents and in most cases, affect the entire system because the racks are inter-connected and mediate links connecting to each other. For example, if rack (0, A) shuts down due to a failure then rack (0, 9) also fails because it does not have its own clock card and it gets its clock signal through rack (0, A). These links do not have to be based on physical proximity either. For example, if rack (1, 4) fails, then the entire system fails because all racks get their clock signals through rack (1, 4) in Mira's Blue Gene/Q design.



(a) *Coolant Flow Rate* **(b)** *Inlet Temperature* **(c)** *Outlet Temperature*

Fig. 12. *The coolant monitor telemetry shows signs before a failure is about to occur. The inlet and the outlet coolant temperatures drop by 5% and 7%, respectively, up to three hours before a CMF.*

One can expect that the rate of failure should be higher near the beginning of Mira's operation period when operators are still learning about the optimal way of managing the system due to the fact that Mira was the first liquid cooling system setup at ALCF's TCS facility. One would also expect the failure rate to be higher near the end of Mira's operation period due to general wear and tear. Surprisingly, Fig. 10 shows that this is not necessarily the case. Failures happen inconsistently over the six years with 40% of all failures taking place back-to-back in 2016 when Theta was brought on but no failures happening for over a two year period after that until the end of 2018.

Summary: CMF failures do not exhibit traditional bathtub-like behavior. Hence, traditional proactive strategies such as burn-in during early phase may not be sufficiently effective. Academic research effort is needed for early identification and prediction of such failures.

Next, we look at how the 361 CMFs are distributed across the individual racks of Mira in Fig. 11. *Surprisingly, the failure distribution among racks is not necessarily correlated with the utilization (which would cause more hardware wear and tear: Fig. 6) or coolant outlet temperature (which means the racks are hotter for prolonged period of time (Fig. 7)) or data center humidity (which has been shown to cause failures: Fig. 9). For example, rack (1, 8), which has the highest number of CMFs, falls in the bottom half of all racks in terms of all three of these metrics: it has low utilization, low outlet temperature, and low humidity.* More formally, the correlation coefficient between the number of CMFs and rack utilization is -0.21, outlet coolant temperature is -0.06, and data center humidity is 0.06. Thus, none of these markers can be used to predict when and where the CMFs are going to occur, which makes it especially difficult to tackle them.

Summary: Rack-level distribution of CMF failures exhibit limited correlation with rack-level distribution of utilization, temperature, or humidity.

B. The Lead up to a Coolant Monitor Failure

While the macro-level temporal and spatial trends of coolant monitor failures indicate a lack of patterns which can help

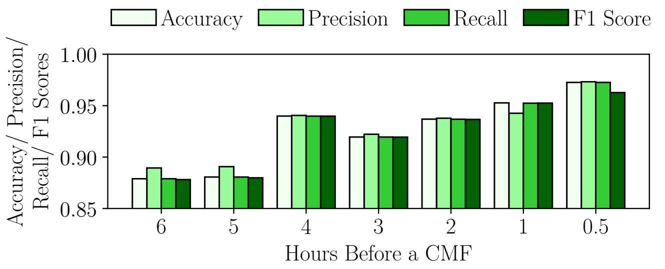


Fig. 13. *The change pattern in the coolant monitor metrics can help predict an impeding CMF with high accuracy up to six hours before its occurrence using a neural network.*

predict when CMFs will occur, given their severity, it is important to be able to predict them even if it is just three to six hours before a failure occurs in order to provide enough time to be able to take corrective and backup measures, and preempt restorative actions. Therefore, we look at the coolant monitor telemetry data up to six hours before a coolant monitor failure occurs and the results are shown in Fig. 12. Fig. 12(a) shows that the coolant flow rate continues to remain relatively stable until just a half hour before a CMF. In fact, in many cases, its rapid and significant decline becomes the cause of the failure. Therefore, it cannot be used as a good prediction feature. On the other hand, when looking at Fig. 12(b) and Fig. 12(c), the results are encouraging. Earlier, in Fig 3 and Fig 7, we saw how stable the metrics tracked by the coolant monitor are temporally and across racks. This is especially true for inlet coolant temperature. However, Fig. 12 shows that the inlet coolant temperature drop by as much as 7%, over four hours before a CMF is about to take place, and then rises by up to 8%, half an hour before a CMF. The outlet coolant temperature also decreases by 5% three hours before a CMF. These two metrics are good indicators of an imminent failure.

Toward that end, we build a simple and effective CMF predictor. This is an example to demonstrate how coolant telemetry can be used toward low-overhead operationally useful tasks. We build a neural network based binary classification model which predicts whether a CMF will occur within the next six hours of time. It uses the change in value of coolant monitor metrics (coolant flow rate, outlet temperature, inlet temperature, system power, data center temperature and humidity) over the past six hours as input features.

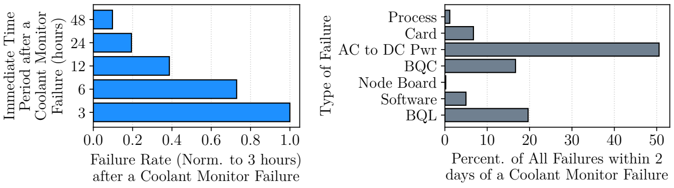
To prepare the training dataset for our model, we collect coolant monitor metrics from six hours before the occurrence of all the CMFs in their respective racks. These data are labeled as class *one* (positive class, a CMF will occur). Next, we collect an equal number of data points for the change in the coolant monitor metrics when no CMF occurred within the next six hours. We evenly collect data throughout the whole period of Mira's production time. These are labelled as class *zero* (negative class, a CMF will not occur). The combination of both the classes of data form our training dataset. We train the neural-network-based prediction model for 50 epochs by dividing this data in a ratio of 3 : 1 : 1 for training, testing and validation, respectively. The neural network consists of three layers between the input and output layers having twelve, twelve, and six neurons in each layer, respectively. Bayesian Optimization, a technique frequently used for hyper-parameter tuning, is used to optimize the architecture of this neural network (number of neurons per layer). Rectified linear unit (ReLU) is used as the activation function for all the layers, except the output layer which uses a sigmoid activation (a commonly used activation function for binary classification which restricts the output between zero and one).

We measure the performance of our prediction model from 30 minutes before an impeding CMF all the way up to six hours before a CMF. We measure accuracy (ratio of correct predictions to total number of predictions), precision (ratio of correct positive class predictions to the sum of correct positive class predictions and incorrect positive class predictions), recall (ratio of correct positive class predictions to the sum of correct positive class predictions and incorrect negative class predictions) and F1 score (harmonic mean of precision and recall) for evaluating the prediction performance. All the prediction performance results are obtained using 5-fold cross validation for robustness against sample selection.

Fig. 13 shows how the predictor performs for all four metrics up to six hours before a CMF. All metrics of performance provide nearly similar values at a given time before a CMF. The prediction model can predict a CMF with 87% accuracy six hours before the event. In fact, it can predict an impeding CMF with as high as 97% accuracy when an impeding CMF is within 30 minutes. Note that the testing set also contains equal number of samples from both positive and negative classes, and the prediction model can predict both classes with high accuracy. Evidently, as the CMF approaches, the prediction performance improves because the coolant monitor metrics change more vigorously. This results in stronger trends which indicate that a CMF might be approaching. Overall, the predictor performs well up to six hours before a CMF, providing ample time to deploy precautionary measures.

Our simple model achieves high prediction quality, but it has certain limitation and areas for improvement. While our predictor requires each rack to be monitored individually, operationally it will be even more useful to have a predictor which even predicts the location of an impeding CMF from the overall coolant telemetry of the datacenter. Second, any proactive measure if a CMF is predicted is likely to incur high overhead since a CMF impacts the whole rack, at minimum. Therefore, the false positives need to minimized as much as possible; it is another area where further scope of improvement is present (our false positive rate is 6%, six hours before a CMF and 1.2%, 30 minutes before a CMF).

Opportunity: The otherwise stable data such as inlet coolant temperature and outlet coolant temperature may observe significant changes up to three to four hours before a coolant monitor failure. This time can be used to checkpoint active jobs, alert data center users, and kick



(a) *Failure Rate* **(b)** *Types of Failures*

Fig. 14. *(a) The rate of failure of non-CMF variety is especially high within three to six hours after a failure. The failure rate drops to 10% of the failure rate at three hours, 48 hours after a CMF. (b) The most common failures to occur after a CMF are "AC to DC Power" and "BQL" related.*

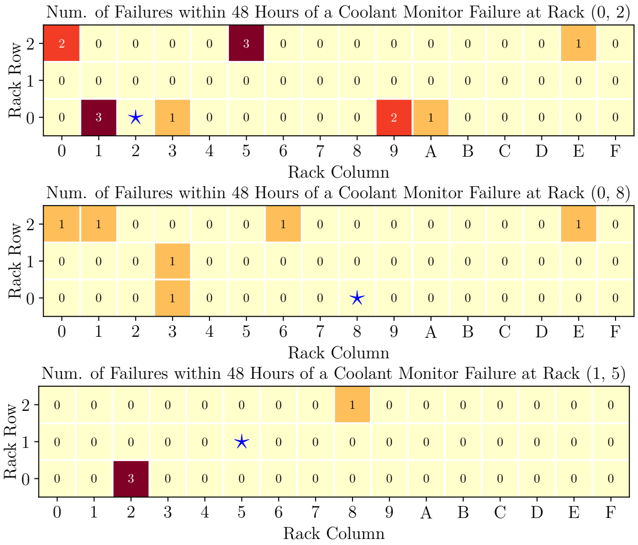


Fig. 15. *The three examples shown here demonstrate that failures after a CMF do not necessarily happen on the same rack as the CMF or even in the vicinity of the epicenter rack (indicated by a blue star). The failures can take place anywhere on the system.*

off backup and restorative actions – more research effort is needed toward emulating and testing such scenarios in research lab setting and devising deployable tools. While it is possible to predict the occurrence of a CMF at the system-level, further improvements are needed to improve the location accuracy and reduce false positives for more effective proactive mitigation actions. As discussed next, predicting CMF occurrence is important, but post-CMF world is even more chaotic and requires more predictive tools.

C. What Follows a Coolant Monitor Failure?

Now that we have analyzed what happens before a coolant monitor failure occurs, next we look at what happens after a coolant monitor failure takes place. Namely, we look at the rate and the type of non-CMF failures that take place. Fig. 14(a) shows the rate of failure within three hours after a CMF to within 48 hours or two days of a CMF. Note that as per our methodology, we do not take cascaded failures into account. It takes on average one hour for a rack to come back up after a non-CMF failure. Therefore, we consider all failures within an hour of each other as a single failure.

According to Fig. 14(a), the failure rate within six hours after a CMF is less than 75% of the failure rate within three hours of a CMF. In fact, the failure rate drops to 10% of the failure rate at three hours, 48 hours after a CMF. This demonstrates that there is heightened risk of non-CMF failures immediately after a CMF. But what types of non-CMF failures take place after a CMF failure?

Fig. 14(b) shows a distribution of the type of failures which the system is most at risk of after a CMF failure. The most common type, in fact, 50% of all non-CMF failures after a CMF failure is a "AC to DC power" failure which denotes the bulk power module failing to convert power at appropriate level. Other significant types of failures include BQC and BQL which are failures caused by Blue Gene/Q Compute and Link modules, respectively. The compute module mainly includes the cores of each node while the link module includes links between higher network topologies, load balancers, and primary and backup devices. While some of these failures can be masked by redundancy, failures in links connecting commodity switches can result in the highest downtime after a failure. Link failures primarily occur due to increased network traffic. The card failure refers to a failure in the clock card which is used to synchronize the nodes. Software failures can range from buggy updates to certain network decisions causing software systems to malfunction. Lastly, process failures, which are rare (< 2%), are mainly caused by various software daemons running in the background.

Next, we look at whether these non-CMF failures occur on the same racks as the original CMF failure which they follow. *Unexpectedly, Fig. 15 shows three examples of CMFs for which the non-CMF failures after a CMF do not necessarily happen on the same rack as the CMF or even in the vicinity of the epicenter rack. The failures can take place on other racks the system because as mentioned above, the racks are inter-linked in a manner which is not necessarily spatially correlated. This makes it especially difficult to predict where non-CMF failures following a CMF are going to occur.*

Summary: The failure rate of non-coolant monitor failures goes up after a coolant monitor failure and these failures can occur on any rack in the system, demonstrating the severity and across-system impact of a coolant monitor failure. We observed that certain types of failures such as "AC to DC power" failure (caused by failures in BPMs) are more likely to occur than others. Moreover, it is also difficult to predict the location of these post-CMF failures as they can happen anywhere on the system.

Opportunity: More research efforts are needed to understand how CMF can manifest itself, induce non-CMF

> failures, and propagate in the system – currently, there is very limited understanding of these critical issues.

D. Discussion

Coolant telemetry: threshold-based monitoring not always sufficient. Typical data center monitoring infrastructure monitors temperature, pressure and humidity levels to detect abnormalities in the system. Usually, there are set threshold levels of these parameters and the system throws off warnings when the corresponding threshold levels are crossed [64, 75]. However, we observed in Sec. VI(B) that not only the level of cooling metrics, but more importantly the change in their values are key features for detecting abnormalities and predicting CMFs. A threshold-based approach is not sufficient for abnormality detection as certain metrics might continue to remain high during periods of high utilization, but that does not signify an impeding failure.

Preventive actions on non-neighboring racks on coolant monitor failures. When a rack undergoes any kind of failure, datacenter operators apply preventive measures to that rack and sometimes to its surrounding ones [9, 25]. But we have observed in Sec. VI that after a CMF, a system-wide *RAS storm* may occur without following a known pattern. It is not necessary that just the epicenter rack and its surrounding racks would get affected. Hence, keeping in mind the severity of the CMFs, datacenter operators should take preventive and precautionary measures for all the racks of the system.

Opportunities for computer architects and system researchers. Software-based checkpointing imposes high overhead and is not practical for production. It leads to resource wastage by idle nodes in many situations (Sec. III(A)). Hardware support for fast checkpoint/ restoration is critical. Also, hardware support for a fast reboot is important toward solving the challenges discussed in this paper (Sec. III(A)). For improving resource utilization and reducing variability in utilization (Sec. III(A-B)), interference- and jitter-free co-location of parallel jobs on the same node is required (e.g., Intel CAT, but not supported by all vendors). Architecture and systems research on mitigating the impact of occasional variation in humidity on CPUs will reduce the overall operational cost of the data centers (Sec. VI). This work can motivate researchers to develop CMF-aware job schedulers and resource management strategies.

VII. Related Work

Previous works have discovered that inefficiencies in the cooling system can be a cause of many data center problems from unnecessarily high energy consumption and high operational and maintenance costs to compromised resiliency characteristics of the center which lead to a high failure rate [7, 12, 23, 27–30, 37, 38, 41, 42, 76]. Consequently, the HPC community has focused research effort toward (1) cooling system development, (2) power management, and (3) reliability of associated infrastructures [32, 33, 48, 53, 59, 61, 70–72].

For increasing energy-efficiency, efforts have primarily been directed toward designing energy-efficient hardware, power capping, dynamic voltage and frequency scaling (DVFS), and energy-efficient job scheduling on HPC and cloud infrastructures [20, 21, 32, 54, 63, 65, 67, 70, 72].

Since power utilization is largely affected by the efficiency of cooling systems, active research continues to be conducted toward developing several kinds of cooling infrastructures. Prior works have developed techniques for cooling-aware and temperature-variation-aware job scheduling and management of HPC systems [34, 51, 70, 74]. Cooling systems also frequently rely on free cooling, which makes the systems prone to severe system outages, a problem which many works have attempted to tackle [10, 17, 19, 31, 43, 55, 68]. However, these works do not provide in-depth insights about how cooling and other environmental factors can affect the performance of a large-scale system over a long period.

Lastly, studies have been performed on how fault rates in specific components of a system are affected by environmental factors such as temperature and humidity [5, 8, 15, 16, 46, 49, 73]. How the cooling system can effectively work on a large-scale computing system to reduce power consumption, operational cost, as well failure rates is still a challenging question, which can be system dependent [1, 2, 4, 13, 35, 36].

VIII. Conclusion

In this paper, we performed the first in-depth study characterizing the operations of Mira supercomputer, and presented several interesting results and insights, especially in the context of temporal variability and rack-wise variability of cooling parameters and the failure rate of the cooling infrastructure. We hope new problems found in operating these production systems, often less-known and under-investigated in academic settings, drive research to mitigate the identified challenges.

Acknowledgment: We are thankful to anonymous reviewers for providing thoughtful feedback despite the pandemic challenges. We are thankful for the support from NSF Award 1910601, Northeastern University, and the Massachusetts Green High Performance Computing Center (MGHPCC) facility, and initial discussions with Zhengchun Liu. The work is also supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences, under Contract DE-AC02-06CH11357. We gratefully acknowledge the computing resources provided by the Argonne Leadership Computing Facility, which is a U.S. Department of Energy, Office of Science User Facility.

References

[1] V. Ahlgren, S. Andersson, J. Brandt, N. Cardo, S. Chunduri, J. Enos, P. Fields, A. Gentile, R. Gerber, M. Gienger, et al. Large-scale system monitoring experiences and recommendations. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 532–542. IEEE, 2018.

[2] V. Ahlgren, S. Andersson, J. M. Brandt, N. Cardo, S. Chunduri, P. Fields, A. C. Gentile, R. Gerber, J. Greenseid, A. Greiner, et al. Cray system monitoring: Successes requirements and priorities. Technical report,

Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia …, 2018.

[3] S. Alkharabsheh, U. L. N. Puvvadi, B. Ramakrishnan, K. Ghose, and B. Sammakia. Failure Analysis of Direct Liquid Cooling System in Data Centers. *Journal of Electronic Packaging*, 140(2), 05 2018. ISSN 1043-7398. doi: 10.1115/1.4039137.

[4] J. Becklehimer, C. Willis, J. Lothian, D. Maxwell, and D. Vasil. Real time health monitoring of the cray xt3/xt4 using the simple event correlator (sec). *Cray Users Group*, 2007.

[5] J. Bhimani, T. Patel, N. Mi, and D. Tiwari. What does vibration do to your ssd? In *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2019.

[6] A. S. Bland, W. Joubert, D. E. Maxwell, N. Podhorszki, J. H. Rogers, and A. N. Tharrington. Contemporary High Performance Computing From Petascale toward Exascale. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States); Center for …, 2013.

[7] N. Bourassa, W. Johnson, J. Broughton, D. M. Carter, S. Joy, R. Vitti, and P. Seto. Operational data analytics: Optimizing the national energy research scientific computing center cooling systems. In *Proceedings of the 48th International Conference on Parallel Processing: Workshops*, pages 1–7, 2019.

[8] G. Callou, P. Maciel, D. Tutsch, and J. Araujo. Models for dependability and sustainability analysis of data center cooling architectures. In *IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN 2012)*, pages 1–6. IEEE, 2012.

[9] J. Cao, M. Lai, Z. Luo, Z. Pang, et al. Efficient management and intelligent fault tolerance for hpc interconnect networks. In *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 343–351. IEEE, 2019.

[10] T. Cao, W. Huang, Y. He, and M. Kondo. Cooling-Aware Job Scheduling and Node Allocation for Overprovisioned HPC Systems. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 728–737. IEEE, 2017.

[11] C. Chen, G. Wang, J. Sun, and W. Xu. Detecting data center cooling problems using a data-driven approach. In *Proceedings of the 9th Asia-Pacific Workshop on Systems*, pages 1–8, 2018.

[12] B.-G. Chun, G. Iannaccone, G. Iannaccone, R. Katz, G. Lee, and L. Niccolini. An energy case for hybrid datacenters. *ACM SIGOPS Operating Systems Review*, 44(1):76–80, 2010.

[13] H. Coles, M. Ellsworth, and D. J. Martinez. "hot" for warm water cooling. In *State of the Practice Reports*, pages 1–10. 2011.

[14] J. Collins, M. E. Papka, B. A. Cerny, and R. M. Coffey. 2016 annual report-argonne leadership computing facility. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States), 2016.

[15] J. Dai, D. Das, M. Ohadi, and M. Pecht. Reliability Risk Mitigation of Free Air Cooling Through Prognostics and Health Management. *Applied energy*, 111:104–112, 2013.

[16] W. Deng, F. Liu, H. Jin, B. Li, and D. Li. Harnessing renewable energy in cloud datacenters: opportunities and challenges. *iEEE Network*, 28 (1):48–55, 2014.

[17] N. El-Sayed, I. A. Stefanovici, G. Amvrosiadis, A. A. Hwang, and B. Schroeder. Temperature Management in Data Centers: Why Some (might) Like it Hot. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, pages 163–174, 2012.

[18] B. Fekade, T. Maksymyuk, and M. Jo. Clustering hypervisors to minimize failures in mobile cloud computing. *Wireless Communications and Mobile Computing*, 16(18):3455–3465, 2016.

[19] W.-c. Feng. The importance of being low power in high performance computing. *Cyberinfrastructure Technology Watch Quarterly (CTWatch Quarterly)*, 1(3):11–20, 2005.

[20] W.-c. Feng and K. Cameron. The green500 list: Encouraging sustainable supercomputing. *Computer*, 40(12):50–55, 2007.

[21] G. Fourestey, B. Cumming, L. Gilly, and T. C. Schulthess. First experiences with validating and using the cray power management database tool. *arXiv preprint arXiv:1408.2657*, 2014.

[22] W. Fox, D. Ghoshal, A. Souza, G. P. Rodrigo, and L. Ramakrishnan. E-hpc: a library for elastic resource management in hpc environments. In *Proceedings of the 12th Workshop on Workflows in Support of Large-Scale Science*, pages 1–11, 2017.

[23] V. W. Freeh, D. K. Lowenthal, F. Pan, N. Kappiah, R. Springer, B. L. Rountree, and M. E. Femal. Analyzing the energy-time trade-off in high-performance computing applications. *IEEE Transactions on Parallel and Distributed Systems*, 18(6):835–848, 2007.

[24] S. Fu and C.-Z. Xu. Exploring event correlation for failure prediction in coalitions of clusters. In *Proceedings of the 2007 ACM/IEEE conference on Supercomputing*, pages 1–12, 2007.

[25] A. Gainaru, F. Cappello, M. Snir, and W. Kramer. Fault prediction under the microscope: A closer look into hpc systems. In *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pages 1–11. IEEE, 2012.

[26] R. Garg, T. Patel, G. Cooperman, and D. Tiwari. Shiraz: Exploiting system reliability and application resilience characteristics to improve large scale system throughput. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2018.

[27] R. Ge, X. Feng, and K. W. Cameron. Performance-constrained distributed dvs scheduling for scientific applications on power-aware clusters. In *SC'05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*, pages 34–34. IEEE, 2005.

[28] R. Ge, X. Feng, W.-c. Feng, and K. W. Cameron. Cpu miser: A performance-directed, run-time system for power-aware clusters. In *2007 International Conference on Parallel Processing (ICPP 2007)*, pages 18–18. IEEE, 2007.

[29] R. Ge, X. Feng, S. Song, H.-C. Chang, D. Li, and K. W. Cameron. Powerpack: Energy profiling and analysis of high-performance systems and applications. *IEEE Transactions on Parallel and Distributed Systems*, 21(5):658–671, 2009.

[30] Í. Goiri, K. Le, M. E. Haque, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini. Greenslot: scheduling energy consumption in green datacenters. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 2011.

[31] Í. Goiri, T. D. Nguyen, and R. Bianchini. Coolair: Temperature- and Variation-Aware Management for Free-Cooled Datacenters. *ACM SIGPLAN Notices*, 50(4):253–265, 2015.

[32] S. Gupta, D. Tiwari, C. Jantzi, J. Rogers, and D. Maxwell. Understanding and exploiting spatial properties of system failures on extreme-scale hpc systems. In *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 37–44. IEEE, 2015.

[33] S. Gupta, T. Patel, C. Engelmann, and D. Tiwari. Failures in large scale systems: long-term measurement, analysis, and implications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–12, 2017.

[34] M. Hennecke, W. Frings, W. Homberg, A. Zitz, M. Knobloch, and H. Böttiger. Measuring power consumption on ibm blue gene/p. *Computer Science-Research and Development*, 27(4):329–336, 2012.

[35] C. Hsu and W. Feng. A Feasibility Analysis of Power Awareness in Commodity-Based High-Performance Clusters. In *2005 IEEE International Conference on Cluster Computing*, pages 1–10, 2005.

[36] S. Huang, S. Fu, S. Pakin, and M. Lang. Characterizing power and energy efficiency of legion data-centric runtime and applications on heterogeneous high-performance computing systems. In *High Performance Parallel Computing*. IntechOpen, 2018.

[37] S. Jana, G. A. Koenig, M. Maiterth, K. T. Pedretti, A. Borghesi, A. Bartolini, B. Hadri, and N. J. Bates. Global survey of energy and power-aware job scheduling and resource management in supercomputing centers. 2017.

[38] K. Kant, M. Murugan, and D. H. Du. Willow: A control system for energy and thermal adaptive computing. In *2011 IEEE International Parallel & Distributed Processing Symposium*. IEEE, 2011.

[39] N. Kumbhare, A. Marathe, A. Akoglu, H. J. Siegel, G. Abdulla, and S. Hariri. A value-oriented job scheduling approach for power-constrained and oversubscribed hpc systems. *IEEE Transactions on Parallel and Distributed Systems*, 31(6):1419–1433, 2020.

[40] G. Lakner, B. Knudson, et al. *IBM System Blue Gene Solution: Blue Gene/Q System Administration*. IBM Redbooks, 2013.

[41] J. H. Laros, K. T. Pedretti, S. M. Kelly, J. P. Vandyke, K. B. Ferreira, C. T. Vaughan, and M. Swan. Topics on measuring real power usage on high performance computing platforms. In *2009 IEEE International Conference on Cluster Computing and Workshops*. IEEE, 2009.

[42] D. Li, B. R. de Supinski, M. Schulz, K. Cameron, and D. S. Nikolopoulos. Hybrid mpi/openmp power-aware computing. In *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)*, pages 1–12. IEEE, 2010.

[43] L. Li, W. Zheng, X. Wang, and X. Wang. Data Center Power Minimization with Placement Optimization of Liquid-Cooled Servers and Free Air Cooling. *Sustainable Computing: Informatics and Systems*, 11:3–15, 2016.

[44] S. Liang, V. Holmes, and I. Kureshi. Hybrid computer cluster with high flexibility. In *2012 IEEE International Conference on Cluster Computing Workshops*, pages 128–135. IEEE, 2012.

[45] D. Lipari. The slurm scheduler design. *SLURM User Group. http://slurm. schedmd. com/slurm_ug_2012/SUG-2012-Scheduling. pdf*, 2012.

[46] S. Lu, B. Luo, T. Patel, Y. Yao, D. Tiwari, and W. Shi. Making disk failure predictions smarter! In *18th {USENIX} Conference on File and Storage Technologies ({FAST} 20)*, pages 151–167, 2020.

[47] A. Luckow and S. Jha. Performance characterization and modeling of serverless and hpc streaming applications. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5688–5696. IEEE, 2019.

[48] O. Mämmelä, M. Majanen, R. Basmadjian, H. De Meer, A. Giesler, and W. Homberg. Energy-aware job scheduler for high-performance computing. *Computer Science-Research and Development*, 27(4):265–275, 2012.

[49] I. Manousakis, S. Sankar, G. McKnight, T. D. Nguyen, and R. Bianchini. Environmental Conditions and Disk Reliability in Free-Cooled Datacenters. In *14th {USENIX} Conference on File and Storage Technologies ({FAST} 16)*, pages 53–65, 2016.

[50] A. Marathe, G. Abdulla, B. L. Rountree, and K. Shoga. Towards a Unified Monitoring Framework for Power, Performance and Thermal Metrics: A Case Study on the Evaluation of HPC Cooling Systems. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 974–983. IEEE, 2017.

[51] V. Morozov, J. Meng, V. Vishwanath, J. R. Hammond, K. Kumaran, and M. E. Papka. Alcf mpi benchmarks: Understanding machine-specific communication behavior. In *2012 41st International Conference on Parallel Processing Workshops*, pages 19–28. IEEE, 2012.

[52] L. Myers and M. J. Sirois. Spearman Correlation Coefficients, Differences between. *Encyclopedia of statistical sciences*, 12, 2004.

[53] B. Nie, J. Xue, S. Gupta, C. Engelmann, E. Smirni, and D. Tiwari. Characterizing temperature, power, and soft-error behaviors in data center systems: Insights, challenges, and opportunities. In *2017 IEEE 25th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 22–31. IEEE, 2017.

[54] M. Ot, T. Wilde, and H. Ruber. Roi and tco analysis of the first production level installation of adsorption chillers in a data center. In *2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 981–986. IEEE, 2017.

[55] E. Pakbaznia, M. Ghasemazar, and M. Pedram. Temperature-aware dynamic resource provisioning in a power-optimized datacenter. In *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, pages 124–129. IEEE, 2010.

[56] L. A. Parnell, D. W. Demetriou, V. Kamath, and E. Y. Zhang. Trends in high performance computing: Exascale systems and facilities beyond the first wave. In *2019 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 167–176. IEEE, 2019.

[57] T. Patel and D. Tiwari. Perq: Fair and efficient power management of power-constrained large-scale computing systems. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, pages 171–182, 2019.

[58] T. Patel, R. Garg, and D. Tiwari. {GIFT}: A coupon based throttle-and-reward mechanism for fair and efficient i/o bandwidth management on parallel storage systems. In *18th {USENIX} Conference on File and Storage Technologies ({FAST} 20)*, pages 103–119, 2020.

[59] T. Patel, Z. Liu, R. Kettimuthu, P. Rich, W. Allcock, and D. Tiwari. Job characteristics on large-scale systems: Long-term analysis, quantification and implications. In *2020 SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1186–1202. IEEE Computer Society, 2020.

[60] T. Patel, A. Wagenhäuser, C. Eibel, T. Hönig, T. Zeiser, and D. Tiwari. What does power consumption behavior of hpc jobs reveal?: Demystifying, quantifying, and predicting power consumption characteristics. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 799–809. IEEE, 2020.

[61] V. A. Patil and V. Chaudhary. Rack aware scheduling in hpc data centers: An energy conservation strategy. *Cluster Computing*, 16(3):559–573, 2013.

[62] M. K. Patterson, S. Krishnan, and J. M. Walters. On energy efficiency of liquid cooled hpc datacenters. In *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 685–693. IEEE, 2016.

[63] N. Rasmussen. Calculating total cooling requirements for data centers. *White paper*, 25:1–8, 2007.

[64] S. Sankar, M. Shaw, K. Vaid, and S. Gurumurthi. Datacenter scale evaluation of the impact of temperature on hard disk drive failures. *ACM Transactions on Storage (TOS)*, 9(2):1–24, 2013.

[65] T. R. Scogland, C. P. Steffen, T. Wilde, F. Parent, S. Coghlan, N. Bates, W.-c. Feng, and E. Strohmaier. A power-measurement methodology for large-scale, high-performance computing. In *Proceedings of the 5th ACM/SPEC international conference on Performance engineering*, pages 149–159, 2014.

[66] W. Scullin and A. Scovel. Lessons from the ibm blue gene series of supercomputers. In *Proceedings of the HPC Systems Professionals Workshop*, pages 1–7, 2017.

[67] H. Shoukourian, T. Wilde, H. Huber, and A. Bode. Analysis of the efficiency characteristics of the first high-temperature direct liquid cooled petascale supercomputer and its cooling infrastructure. *Journal of Parallel and Distributed Computing*, 107:87–100, 2017.

[68] P. Singh, L. Klein, D. Agonafer, J. M. Shah, and K. D. Pujara. Effect of Relative Humidity, Temperature and Gaseous and Particulate Contaminations on Information Technology Equipment Reliability. In *ASME 2015 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with the ASME 2015 13th International Conference on Nanochannels, Microchannels, and Minichannels*, 2015.

[69] J. Spillner, C. Mateos, and D. A. Monge. Faaster, better, cheaper: The prospect of serverless scientific computing and hpc. In *Latin American High Performance Computing Conference*, pages 154–168. Springer, 2017.

[70] K. Tang, D. Tiwari, S. Gupta, P. Huang, Q. Lu, C. Engelmann, and X. He. Power-capping aware checkpointing: On the interplay among power-capping, temperature, reliability, performance, and energy. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 311–322. IEEE, 2016.

[71] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach. *IEEE Transactions on Parallel and Distributed Systems*, 19(11):1458–1472, 2008.

[72] D. Tiwari, S. Gupta, and S. S. Vazhkudai. Lazy checkpointing: Exploiting temporal locality in failures to mitigate checkpointing overheads on extreme-scale systems. In *2014 44th IEEE International Conference on Dependable Systems and Networks*, pages 25–36. IEEE, 2014.

[73] S. Wallace, V. Vishwanath, S. Coghlan, Z. Lan, and M. E. Papka. Measuring Power Consumption on IBM Blue Gene/Q. In *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*, pages 853–859. IEEE, 2013.

[74] S. Wallace, Z. Zhou, V. Vishwanath, S. Coghlan, J. Tramm, Z. Lan, and M. E. Papka. Application power profiling on ibm blue gene/q. *Parallel Computing*, 57:73–86, 2016.

[75] G. Wang, L. Zhang, and W. Xu. What can we learn from four years of data center hardware failures? In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 25–36. IEEE, 2017.

[76] T. Wilde, M. Ott, A. Auweter, I. Meijer, P. Ruch, M. Hilger, S. Kühnert, and H. Huber. Coolmuc-2: A supercomputing cluster with heat recovery for adsorption cooling. In *2017 33rd Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*, pages 115–121. IEEE, 2017.

[77] K. Yoshii, K. Iskra, R. Gupta, P. Beckman, V. Vishwanath, C. Yu, and S. Coghlan. Evaluating power-monitoring capabilities on ibm blue gene/p and blue gene/q. In *2012 IEEE International Conference on Cluster Computing*, pages 36–44. IEEE, 2012.