

Blue Waters System and Component Reliability

Brett Bode, David King, Celso L. Mendes,
William T. Kramer, Saurabh Jha
National Center for Supercomputing Applications
University of Illinois
Urbana, Illinois, USA
{brett,kingda,cmendes,wtkramer,sjha8}@illinois.edu

Roger Ford, Justin Davis and Steven Dramstad
Cray, an HPE Company
Champaign, Illinois, USA
{roger.ford,justin.davis,steven.dramstad}@hpe.com

Abstract—The Blue Waters system, installed in 2012 at NCSA, has the largest component count of any system Cray has built. Blue Waters includes a mix of dual-socket CPU (XE) and single-socket CPU, single GPU (XK) nodes. The primary storage is provided by Cray’s Sonexion/ClusterStor Lustre storage system delivering 35PB (raw) storage at 1TB/sec. The statistical failure rates over time for each component including CPU, DIMM, GPU, disk drive, power supply, blower, etc and their impact on higher level failure rates for individual nodes and the systems as a whole are presented in detail, with a particular emphasis on identifying any increases in rate that might indicate the right-side of the expected bathtub curve has been reached. Strategies employed by NCSA and Cray for minimizing the impact of component failure, such as the preemptive removal of suspect disk drives, are also presented.

Keywords—Failure analysis, system management.

I. INTRODUCTION

Blue Waters is a 27,648 node Cray XE6/XK7 system constructed in 2012 that went into production in 2013 and will remain in production through 2021. Blue Waters is the largest system Cray has built in terms of cabinet count (288) and compute node count (26,862). Cray also provided the high-speed storage for Blue Waters in the form of over 26PB of usable disk space in Cray Sonexion (aka ClusterStor) Lustre appliances. Until the end of 2019 Blue Waters served as the leadership supercomputer for the National Science Foundation (NSF), providing large-scale computing to NSF researchers across the nation covering a wide gamut of science domains, from elementary particle physics and computational biology to weather and climate modeling to star formation and cosmology [1]. One unexpected project was the use of Blue Waters to produce highly accurate Digital Elevation Maps (DEM) from optical satellite image data. Blue Waters has proven so successful at this job that the system will continue to operate through 2021 to focus on DEM production.

The longevity of Blue Waters is enabled by an excellent track record for reliability that is the subject of this paper. Blue Waters’ extreme scale can be a challenge for reliability but offers an excellent opportunity for studying long-term component failure rates. In this paper we examine the full spectrum of significant failure modes, from the failure rates of individual components (CPU, DIMMs, etc) to the impact

of external forces such as power and cooling issues. Any differences from standard Cray practice will be discussed.

The remainder of this paper is organized as follows. Section II briefly reviews the history leading to Blue Waters design and deployment. Section III presents reliability aspects for individual types of system components, like processors, memory chips, disk drives, and others. Section IV offers a wider view of reliability, covering failure rates for larger portions of the system. Section V describes the accumulated Blue Waters raw monitoring data shared in a public website, which would hopefully be of interest to the HPC community. Finally, section VI addresses the current scenario of spare parts, and section VII concludes our presentation.

II. BRIEF HISTORY OF BLUE WATERS

The Blue Waters project started in 2007, when NCSA was awarded a grant from NSF to deploy a sustained-petaflop system that could provide advanced computing capabilities to serve the science and engineering communities [2]. At the beginning of the project, a selection of key science and engineering applications was composed, and NCSA staff worked extensively with developers of those applications to ensure that their codes would be ready to effectively scale up to a petaflop level of sustained performance [3].

The deployment of Blue Waters hardware was started in early 2012, when Cray installed at NCSA a 32-cabinet partition that could be used by selected Early-Science applications. This partition contained only AMD Interlagos processors. During the Summer and Fall of 2012, major remaining parts of the system were installed, including 244 additional cabinets with a mix of CPUs and GPUs, full interconnection network, and the final storage sub-system.

After intensive on-site testing by NCSA with assistance from Cray [4], Blue Waters was formally accepted by NSF near the end of 2012 [5]. For acceptance, beyond traditional functionality tests, sustained-petascale performance was measured on a set of fully functional scientific applications [6], using a metric based on a method that considers time-to-solution as the key factor in evaluation of performance [7].

Early in 2013, a nearline storage component was added, containing a high-capacity tape sub-system. The operation of Blue Waters in a production manner started in April 2013. That operation was briefly interrupted in the Summer of 2013 to integrate 12 additional cabinets comprising exclusively GPU nodes, aiming to further promote the adoption of GPU-

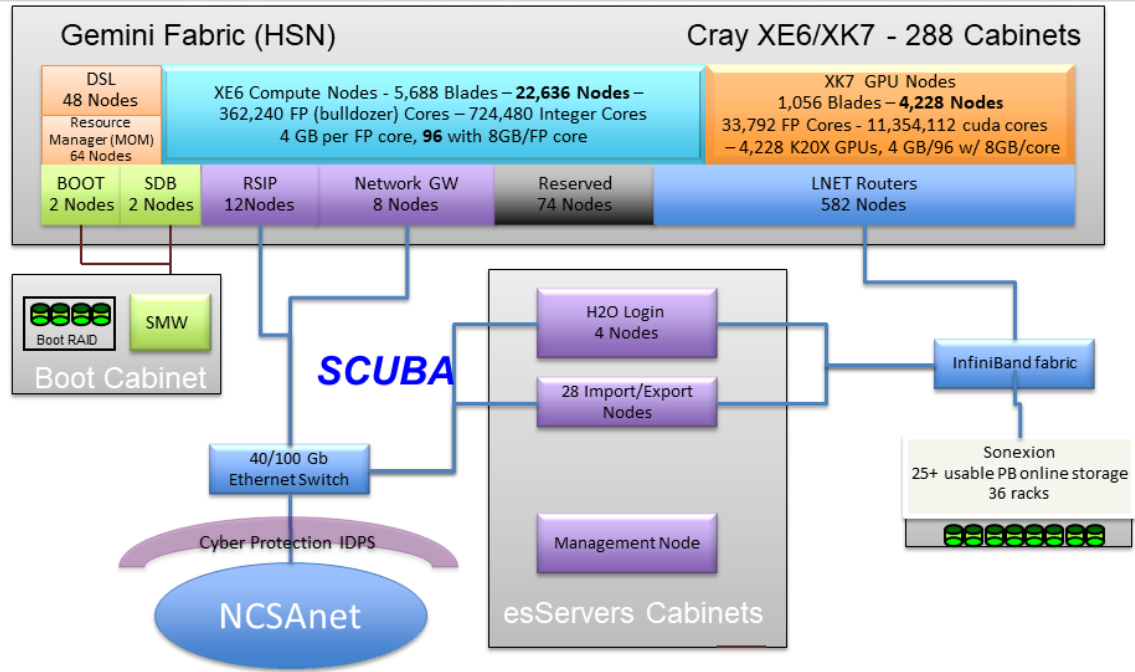


Figure 1 Overview of Blue Waters System

computing by the scientific community. This integration expanded the Blue Waters physical topology to its final configuration, comprising 288 cabinets of compute nodes [8].

In summary, Blue Waters has the organization depicted in Figure 1. The compute partition contains 27,648 nodes, including XE nodes (dual-socket AMD processors) and XK nodes (an AMD processor and an NVIDIA K20X GPU). The system has a combined total of 57,930 CPU processors and 4,228 GPUs. There are 201,568 DIMMs in the memory system, and 17,712 disk drives, with 2 TB each, for storage.

During the period when Blue Waters was serving NSF-based allocations, the job scheduling policy favored large jobs. To improve the spatial geometric shapes of the set of nodes allocated to such jobs, a topology-aware job scheduler

was developed and successfully deployed [9]. This special job scheduler was deactivated in 2020, when the new workload consisted predominantly of jobs with fewer nodes.

III. COMPONENT RELIABILITY

We start our discussion on reliability by presenting, in this section, observed failure data for each individual type of system component. This analysis is useful to show the actual behavior from each particular class of components across more than eight years of Blue Waters operation in production.

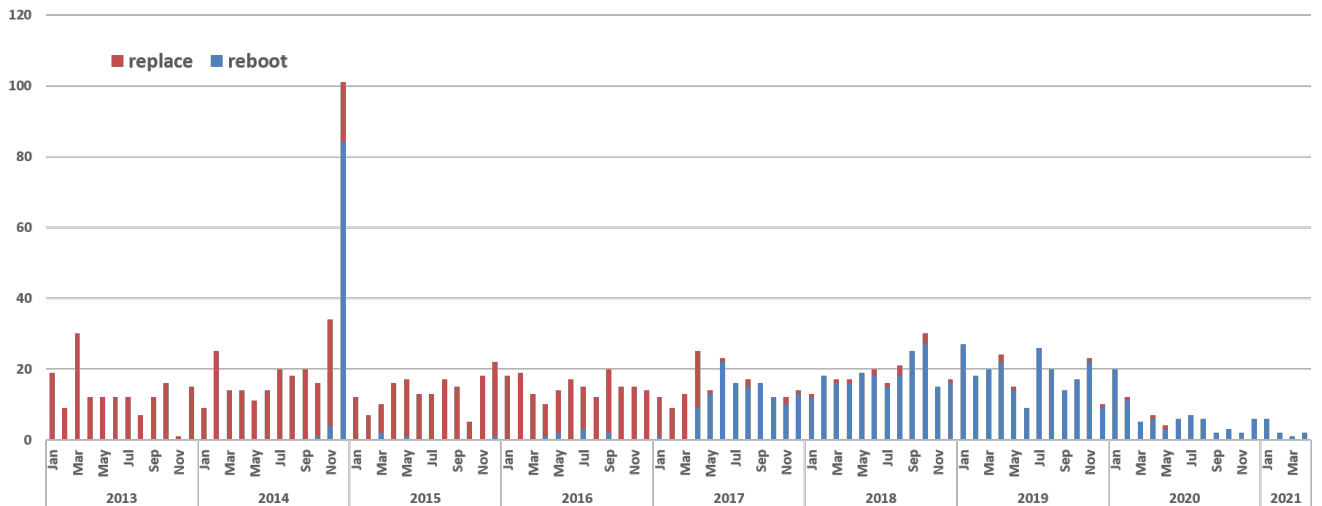


Figure 2 AMD processor faults

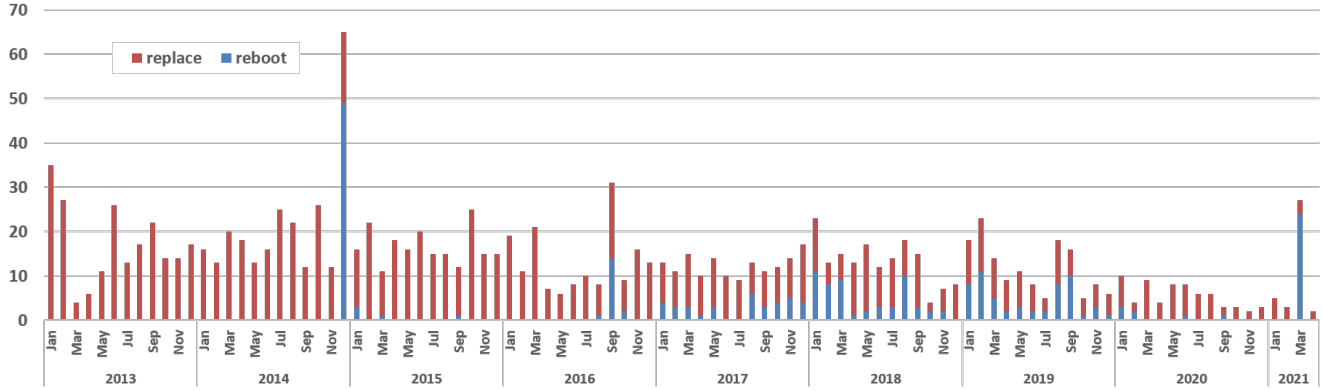


Figure 3 Monthly DIMM fault events

A. AMD Processor Faults

The monthly failure rate for the AMD Opteron/Interlagos processors, the CPUs in Blue Waters nodes, are given by Figure 2. This figure shows the total number of failures observed in each month since 2013. In the first years of operation, Cray would immediately replace any processors from nodes detected as failing. As time went on, the removed processors were tested offline, and most of those failures proved to be transient failures. Thus, a simple node-reboot would be sufficient to recover the failing nodes. This rebooting practice was adopted in 2017, as the first attempt to bring a failing node back to normal operation. If the node still failed after the reboot, then the processor was actually replaced. From a total of 1,525 processor fault events, only 140 were confirmed to correspond to failed processors.

The observed spike in failures for December 2014, initially attributed to the processors, was later identified to be in fact due to a memory issue, as we discuss in the next subsection. Overall, the AMD processors have shown very stable behavior, and the low number of failures observed after January 2020 reflects the transition in the system workload, as the NSF allocations expired at the end of 2019. After that

point, the new workload has been less CPU-intensive and more demanding in terms of both GPU computing and I/O.

B. DIMM Faults

The monthly number of DIMM fault events is presented in Figure 3. Just like in processors, DIMM failure events initially led to replacement of the underlying DIMMs, but later on Cray started to work around those node failures by simply rebooting the nodes. However, for DIMMs, rebooting a node was not as effective in fixing the failure as it was for processors, and actual replacements were required.

Blue Waters employs DIMMs made by four different manufacturers, as shown in Table 1. The majority of DIMMs currently installed comes from Micron, and that is also the type of DIMM that has been in the system for the longest time. Hence, it accumulates the highest percentage of failures.

The three spikes observed in Figure 3, for Dec/2014, Sep/2016 and Mar/2021, correspond to “row-hammer” events in the DIMMs [10]: due to a design problem, certain memory access patterns on writes cause changes in neighboring memory cells. When those changes are detected by the processor it triggers a node interrupt. On Blue Waters, this problem was observed only on the Samsung DIMMs, and

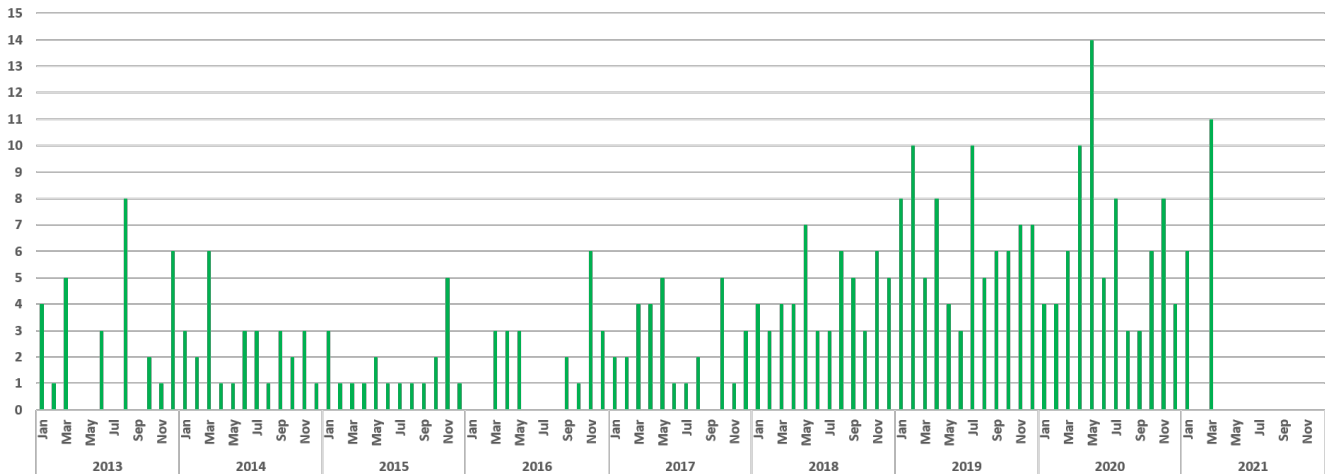


Figure 4 Monthly GPU faults

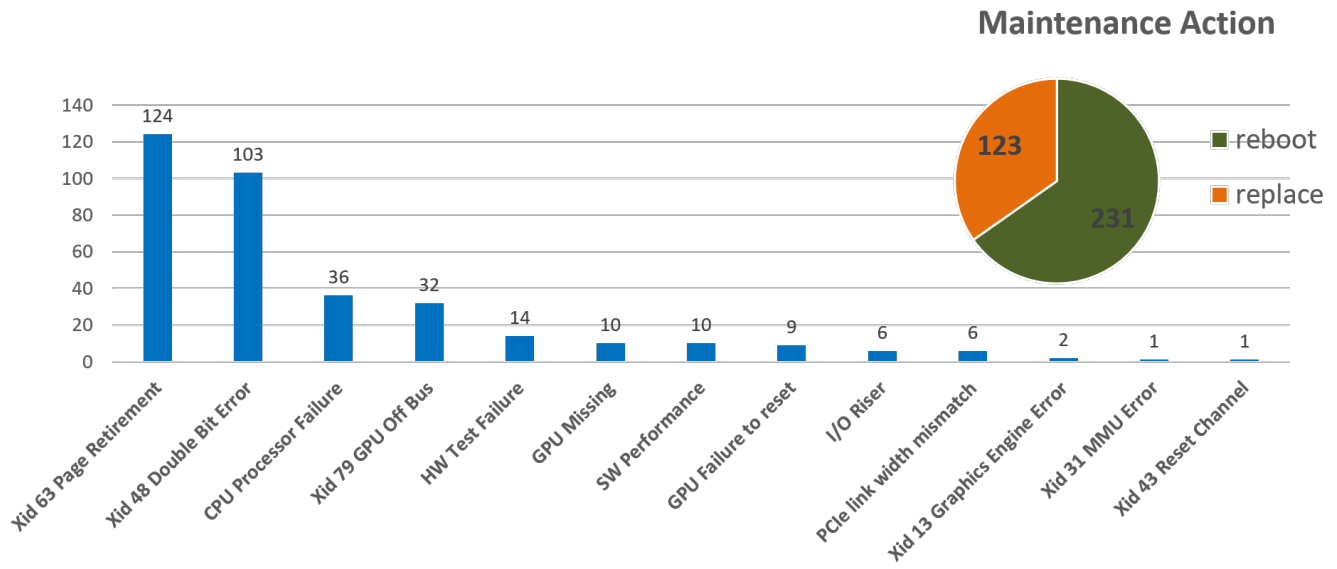


Figure 5 Reason for the observed GPU faults and corresponding maintenance action

because it is associated to particular kinds of memory access, it is triggered only by a few applications. The problem was corrected in each case by working with the application team to slightly modify their code or often just change the compiler optimization level.

Table 1- Type and number of DIMMs in Blue Waters

Manufacturer	Installed	Failures	Per installed	Model
Micron Technology	92160	574	0.62%	DIMM 8GB MICRON PC3-12800
Samsung	51928	298	0.57%	DIMM 8GB SAMSUNG PC3-12800
Hynix Semiconductor	57288	170	0.30%	DIMM 8GB HYNIX PC3-12800
Qimondo	192	0	0.00%	DIMM 4GB QIMONDA PC2-6400
Totals	201568	1042	0.52%	

C. GPU Faults

Although other large Cray systems employing GPUs, like ORNL's Titan, have presented a high number of failures in the past [11], the number of observed GPU failures on Blue Waters was moderate. Figure 4 shows the number of GPU failures by month, indicating that there was a noticeable increase in the number of monthly failures in recent years. The average GPU utilization was high and steady through the end of 2019 but dropped substantially after the start of 2020 with a significant portion of XK node usage by CPU only applications. Since the GPU failure rate is close to flat from 2018 through 2021, we conclude that the failures are simply age related rather than load related. Despite the modest increase in the failure rate the number of failures is still well

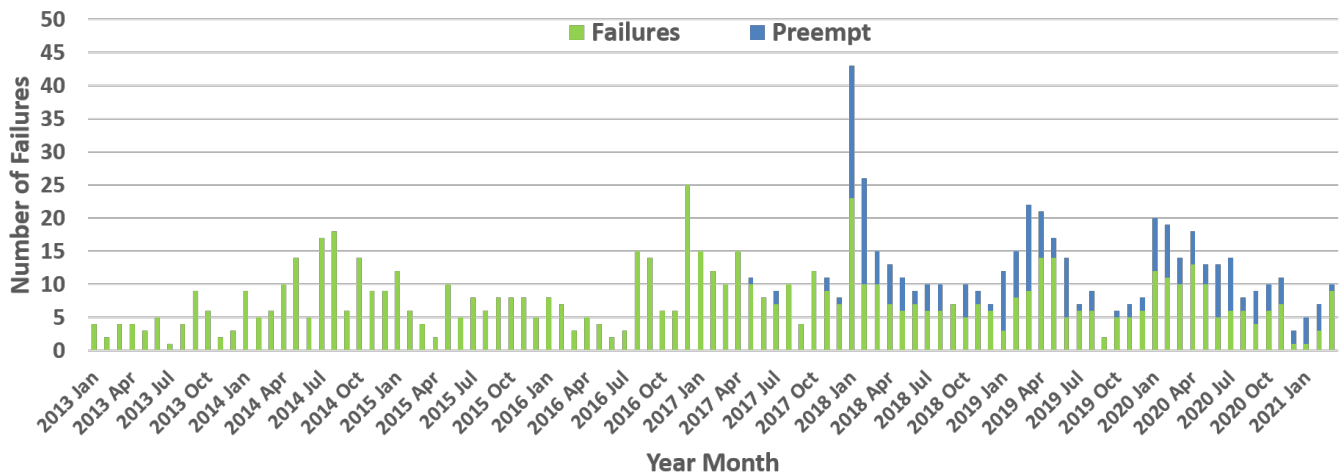


Figure 6 Number of 2TB disk drive replacements per month

within manageable levels, and we have sufficient spare parts to last several years.

The reasons for the GPU failures are detailed in **Error! Reference source not found.**, showing that most of the faults are due to either page retirement or double-bit errors in the GPU memory structure. Such errors are well documented in the literature [12], and their occurrence is not surprising for a large system with the dimensions of Blue Waters. Nearly one third of the failures required replacement of the GPUs, whereas the remaining failures could be properly managed with a reboot of the corresponding XK node.

The low number of observed GPU replacements (i.e. only 123 in more than 4,200 parts, or less than 3%) is even more remarkable if we note that most of these GPUs were shipped by NVIDIA directly to NCSA and were in the same manufacturing batch as those shipped to ORNL for the Titan system. In the Fall of 2012, when most of Blue Waters was already in place undergoing testing, the K20X GPUs, which had just entered production at NVIDIA, were installed by Cray personnel into empty sockets of the XK nodes. Thus, these nodes did not even have the extensive factory-testing that is typically done by Cray before deploying a system at the customer's site. Nevertheless, during system acceptance, specific tests were conducted to verify the proper behavior of all XK nodes.

D. Hard Drive Failures

For a system with more than 17,000 traditional (i.e. mechanical) hard disk drives, one would expect the storage sub-system to be a critical component for reliability. However, Blue Waters disks continue to present very good behavior, despite the age of the system. Figure 6 depicts the number of disk replacements by month, including the replacements required due to actual failures and the replacements recommended based on some degraded metrics observed for the drives.

Starting in January 2018, Cray implemented a more rigorous policy for preemptively replacing disk drives: they replaced any drive that would either (a) contain more than 1,000 replaced sectors, or (b) achieve more than 100 uncorrected reads/writes, or (c) present consistently slow response. Under this new policy, disks were replaced more often, with a monthly average of 11.1 replaced drives for the last three years, whereas the lifetime rate for the system is 9.7 replaced drives per month. Nevertheless, for the past 12 months, the replacement rate was 10.1 drives per month, showing that the storage sub-system of Blue Waters has not yet reached the extreme of the bathtub curve normally expected for aging components [13].

This new disk replacement policy was motivated by the spike in failures observed in January 2018: two drives in the same RAID-set failed, and during the rebuild of that RAID-set, a third drive in the same set was preemptively failed due to a high rate of its observed errors. This required manual reconstruction of that RAID-set, and luckily no data loss occurred. To minimize the likelihood of reoccurrence of such failures in a given RAID-set, the new replacement scheme was adopted. In addition, the pre-emptive failure logic was

modified such that it would not fail a drive while that drive was part of a rebuild process.

E. Liquid Cooling System Failures

The Cray XE/XK rack design employs a mix of liquid and air cooling. External to the cabinets, Liebert XDP heat exchanger units cool RF134A refrigerant using facility chilled water. The RF134A provides liquid cooling that serves as the basis for the cooling of a given cabinet [14]. At the bottom of each cabinet, there is a blower that blows air from the lower to the upper part of the cabinet, providing cooling air through the compute blades that are positioned vertically. The air that circulates internally in the cabinet is cooled by the liquid that is provided externally by the XDPs. The external liquid cooling mechanism is supported by an advanced structure [15] in the building where Blue Waters is installed.

For Blue Waters, each XDP unit feeds four compute cabinets and in normal operation the design is temperature room neutral but is not a closed loop. When an XDP unit fails, the exhaust air from the affected four racks quickly heats up and can easily exceed the maximum inlet temperature for the racks, causing them to power off. Since the exhaust air mixes with neighboring cabinets, all four racks do not always fail. In addition, if the XDP issue can be anticipated, the rear doors of the affected cabinets can be propped open and vent tiles arranged to provide facility air from the raised floor to keep the racks running while the XDP is serviced. In the Blue Waters facility, one challenge for this system is that the cooling water supplied to the XDPs has two sources. In the cooler months, onsite evaporative cooling towers provide water at up to 60 degrees F, while the rest of the year mechanically chilled water is provided at 43 degrees F. This temperature range stresses the water control valves, which has resulted in the valve controls being a significant maintenance issue. In addition, the original pump gaskets degraded over time, resulting in the refrigerant leaking out and potentially shutting down four compute cabinets. Ten single or multiple compute rack interrupts were attributed to XDP issues. However, after many more issues were proactively detected and corrected without impacting the compute system, all pump gaskets and valve control arms were proactively replaced.

F. Blower Failures

As mentioned in the previous subsection, each Cray XE/XK rack utilizes a single, 7.5HP blower to keep the rack cool. When that blower fails the rack almost immediately powers off due to exceeding the thermal limits. Fortunately, blower failures are rare, with ten total failures in eight years and a maximum of two in one year.

G. Power Supplies

Each Cray XE/XK cabinet utilizes seven power supplies to convert the 480V AC input to DC for distribution inside the rack. The power supplies are designed with redundancy such that one can fail without impacting the rack. However, in certain failure modes an arc is generated that creates an inrush current high enough to trip the facility supply breaker,

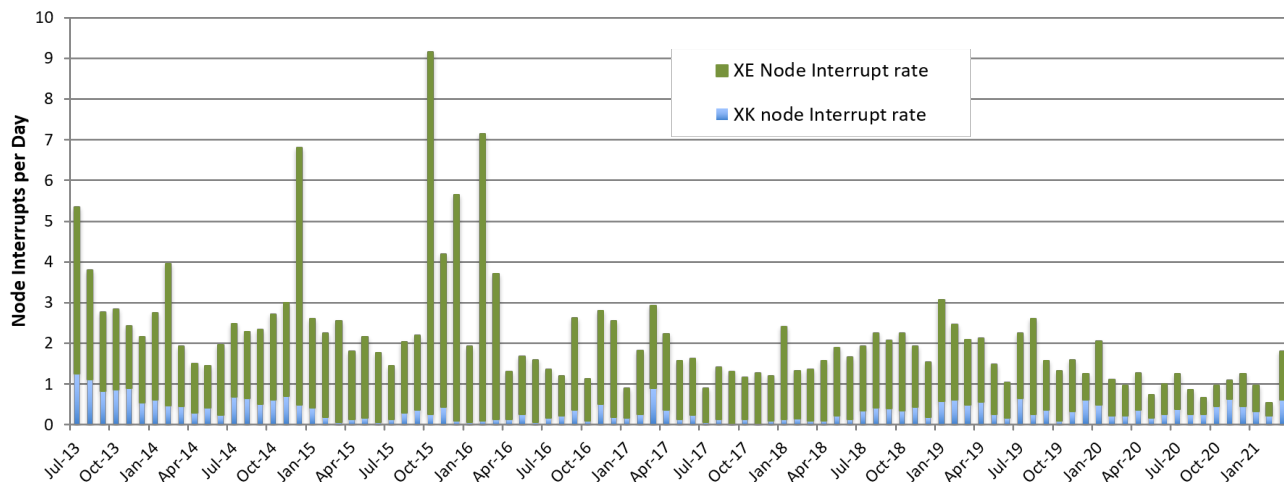


Figure 7 Daily node interrupt rates by month

taking down the cabinet. This unusual failure mode has caused six single rack failures in the eight years of operations.

IV. SYSTEM-WIDE RELIABILITY

We now discuss failures of a wider scope, like those causing interruptions in an entire node or even in the full system. Thanks to good engineering and careful maintenance, these interruptions have not caused severe downtimes along the lifespan of Blue Waters so far. Nevertheless, it is still instructive to analyze their frequency and main causes.

A. Daily Node Interrupt Rate

The daily node failure rate for Blue Waters is in **Error! Reference source not found.**, for both XE (dual CPUs) and XK (CPU+GPU) nodes. As the figure shows, the failure rate has been quite stable. In the first years of operation, many node failures were caused by software, in particular due to installation of new software releases. In the latest years, a higher proportion of failures have been caused by the

hardware, since software updates have been much less frequent in this period.

Over the lifetime of Blue Waters, we have observed an average of 2 node failures per day. Since 2016, as the system software became more mature, the rate dropped to 1.6 node failures per day. Furthermore, over the last 12 months, we have observed an average of only one node failure per day, which is quite impressive for a system with more than 27,000 nodes.

Because there are no periodic interruptions for preventative maintenance on Blue Waters, whenever a node fails, and cannot be rebooted by software, that node is left down until there is an opportunity for its replacement. When the number of nodes down goes beyond a certain value agreed upon by Cray and NCSA, Cray conducts (at most once a week) a procedure of “warm-swapping” blades containing failed nodes. This is preceded by a rerouting of the high-speed interconnect, such that the affected blades can be safely powered down and physically replaced while the rest of the system continues to be in regular operation. After the blades

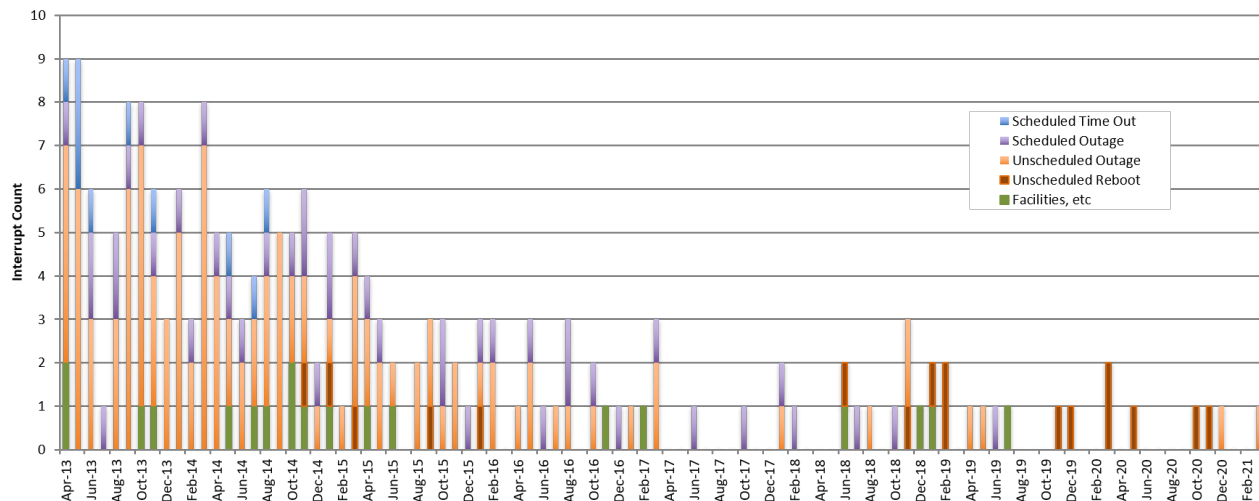


Figure 8 System-wide interrupts

are replaced and powered up, the interconnect is reconfigured back to its original routing. The removed nodes are then tested offline by Cray, in a test cabinet, to diagnose their status.

B. System-Wide Interrupts

Various different factors can lead to a system-wide outage in a large system. **Error! Reference source not found.** shows the number of monthly system-wide interrupts since Blue Waters entered into production. As in the case of node interrupts, most outages in the first three years were caused by software issues. Since 2016, however, the number of interrupts in a month has been stable and quite low. Because the system software is no longer updated frequently, other reasons led to the recent outages, such as occasional hardware failures in the interconnect or problems in the storage sub-system caused by particular access patterns in the workload.

The second notable aspect is related to unscheduled reboots. Rebooting a large system like Blue Waters is a very costly operation, which can take several hours depending on how the system was last shut down. Fortunately, as noted in **Error! Reference source not found.**, the number of such operations was typically low on any given year. Furthermore, given the policy of avoiding regular maintenance operations as already discussed in the previous subsection, the time wasted by powering up or down the system was kept to a minimum.

C. Power Events

Power is provided to the Blue Waters facility via four 8MW 13.8KV feeds from the University of Illinois' 100KV substation providing 24 MW of usable power, with the option of using multiple feeds for power diversity. Since this power comes from the local utility it is susceptible to local severe weather events, though those interrupts are usually less than a second. Due to the capital and operating costs, no Blue Waters equipment utilizes a UPS or other similar backup power. The Blue Waters storage sub-system utilizes two feeds to each rack, which has proven to be very effective at preventing power-related issues. However, each Cray XE/XK rack allows only a single power feed. Thus, 288 XE/XK racks are distributed across the four feeds and an interrupt on any of the feeds has the potential to power off $\frac{1}{4}$ of the system, requiring a reboot to recover system operation. In eight years of operations there have been 15 full system outages due to facility power issues. Ten of those issues were caused by external weather-related events while the others were a mixture of component failure and human error.

V. BLUE WATERS DATA AVAILABILITY

Blue Waters is possibly one of the most intensively monitored supercomputer in the world. In addition to its large size, which naturally leads to a huge volume of data from various metrics, the management of the system involves collection and analysis of several distinct kinds of operating system and application data. The monitoring of the physical machine uses a holistic approach that collects information from all system components [16], using a scalable, light-

weight mechanism that captures that information from each source with a frequency of one sample per minute [17].

In addition to hardware-related data, extensive logs from activity in the system are maintained, aiming to support detailed analysis of system behavior, application performance, or any other investigation of interest. To offer the HPC community a rich source of real-life system data, many of these logs, properly anonymized, are publicly available via the Globus transfer tool at the following site:

<https://bluewaters.ncsa.illinois.edu/data-sets>

The kind of available data consists of the following:

- Collected metrics from hardware sub-components
- Various system-related logs
- Logs from the job scheduler and job execution
- Darshan logs from I/O activity in applications
- Logs of user's view for Quality-of-Service of storage

VI. SPARE PARTS

Even though the component failure rates on Blue Waters are low, there are still parts failing that need to be replaced. Since Cray XE/XK systems have not been produced in over five years, HPE/Cray no longer stocks replacement parts. However, NCSA and HPE/Cray anticipated this situation and HPE/Cray placed multiple complete racks of equipment retired from other customers at NCSA to use as spare parts. The exception is hard disk drives, which are replaced with new drives, but since 2TB drives are no longer available replacements are now 8TB in size, though the extra space is not used. With these arrangements, we believe there are sufficient spare parts to operate the full Blue Waters system for multiple additional years.

VII. CONCLUSION

The data presented in this paper clearly shows that it is possible to run HPC systems, even extreme scale systems such as Blue Waters, for much longer time periods than is normal. Even after eight full years of production operation, failure rates are stable to only slightly increasing. This is perhaps most surprising for the hard drives, which have now been in continuous operation for nine years. Overall, we have sufficient parts to continue to operate Blue Waters for several more years. It is much more likely that Blue Waters will be shut down due to its power efficiency or its aging software stack than its reliability.

ACKNOWLEDGMENTS

The Blue Waters sustained-petascale computing project is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993), the National Geospatial-Intelligence Agency and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. We thank the rest of the Blue Waters staff and the HPE/Cray site team, in particular Mark Dalton, for their efforts in keeping Blue Waters in excellent operational shape for the last eight years.

REFERENCES

- [1] G. H. Bauer, B. Bode, J. Enos, W. T. Kramer, S. Lathrop, C. L. Mendes and R. R. Sisneros, "Best Practices and Lessons from Deploying and Operating a Sustained-Petascale System: The Blue Waters Experience," in *Proceedings of Supercomputing'18*, Dallas, TX, 2018.
- [2] W. Kramer, "Blue Waters - A Super System for Super Challenges," in *Proceedings of the Cray User Group Annual Meeting - CUG'2012*, Stuttgart, 2012.
- [3] G. H. Bauer, T. Hoefler, W. T. Kramer and R. A. Fiedler, "Analyses and modeling of applications used to demonstrate sustained petascale performance on Blue Waters," in *Proceedings of Cray User Group Meeting (CUG-2012)*, Stuttgart, 2012.
- [4] C. L. Mendes, B. Bode, G. H. Bauer, J. Enos, C. Beldica and W. T. Kramer, "Deployment and Testing of the Sustained Petascale Blue Waters System," *Journal of Computational Science*, vol. 10, pp. 327 -- 337, 2015.
- [5] C. L. Mendes, B. Bode, G. H. Bauer, J. R. Moggli, C. Beldica and W. T. Kramer, "Blue Waters Acceptance: Challenges and Accomplishments," in *Proceedings of CUG-2013*, Napa, CA, 2013.
- [6] G. Bauer, V. Anisimov, G. Arnold, B. Bode, R. Brunner, T. Cortese, R. Haas, A. Kot, W. Kramer, J. Kwack, J. Li, C. Mendes, R. Mokos and C. Steffen, "Updating the SPP benchmark suite for extreme-scale systems," in *Proceedings of Cray User Group Meeting (CUG-2017)*, Redmond, WA, 2017.
- [7] W. T. Kramer, "PERCU: A Holistic Method for Evaluating High Performance Computing Systems," University of California at Berkeley, Berkeley, 2008.
- [8] C. L. Mendes, G. H. Bauer, W. T. Kramer and R. A. Fiedler, "Expanding Blue Waters with Improved Acceleration Capability," in *Proceedings of CUG-2014*, Lugano, Switzerland, 2014.
- [9] J. Enos, G. Bauer, R. Brunner, S. Islam, R. A. Fiedler, M. Steed and D. Jackson, "Topology-aware job scheduling strategies for torus networks," in *Proceedings of Cray User Group Meeting (CUG-2014)*, Lugano, Switzerland, 2014.
- [10] M. Jung, C. C. Rheiländer, C. Weis and N. Wehn, "Reverse Engineering of DRAMs: Row Hammer with Crosshair," in *Proceedings of the Second International Symposium on Memory Systems (MEMSYS)*, Alexandria/VA, 2016.
- [11] J. Rogers, "Statistical Analysis of Titan Reliability as it reaches End of Life," in *Proceedings of the Annual Cray User Group Meeting (CUG)*, Montreal, 2019.
- [12] Tiwari, D. et al, "Understanding GPU errors on large-scale HPC systems and the implications for system design and operation," in *Proceedings of the IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, Seoul, South Korea, 2015.
- [13] G. A. Klutke, P. C. Kiessler and M. A. Wortman, "A critical look at the bathtub curve," *IEEE Transactions on Reliability*, vol. 52, no. 1, pp. 125-129, 2003.
- [14] D. Maxwell, M. Ezell, M. Donovan, C. Layton and J. Becklehimer, "Monitoring Cray Cooling Systems," in *Proceedings of the Annual Cray User Group Meeting (CUG)*, Lugano, 2014.
- [15] K. Chung, V. Formicola, Z. T. Kalbarczyk, R. K. Iyer, A. Withers and A. J. Slagell, "Attacking Supercomputers Through Targeted Alteration of Environmental Control: A Data Driven Case Study," in *Proceedings of the International Workshop on Cyber-Physical Systems Security (CPS-Sec)*, Philadelphia, 2016.
- [16] Showerman, M. et al, "Large Scale System Monitoring and Analysis on Blue Waters using OVIS," in *Proceedings of the Cray User Group Annual Meeting (CUG)*, Lugano, 2014.
- [17] "Lightweight Distributed Metric Service (LDMS)," Open Grid Computing and Sandia National Laboratories/NTESS. Available Open Source, 2018. [Online]. Available: <https://github.com/ovis-hpc/ovis>. [Accessed 8 April 2021].