Reading between the lines of failure logs: Understanding how HPC systems fail

Nosayba El-Sayed Bianca Schroeder
Department of Computer Science
University of Toronto
Toronto, Canada
{nosayba, bianca}@cs.toronto.edu

***Abstract*—As the component count in supercomputing installations continues to increase, system reliability is becoming one of the major issues in designing HPC systems. These issues will become more challenging in future Exascale systems, which are predicted to include millions of CPU cores. Even with relatively reliable individual components, the sheer number of components will increase failure rates to unprecedented levels. Efficiently running those systems will require a good understanding of how different factors impact system reliability.**

In this paper we use a decade worth of field data made available by Los Alamos National Lab to study the impact of a diverse set of factors on the reliability of HPC systems. We provide insights into the nature of correlations between failures, and investigate the impact of factors, such as the power quality, temperature, fan and chiller reliability, system usage and utilization, and external factors, such as cosmic radiation, on system reliability.

I. Introduction

System reliability is one of the major challenges in running and designing high-performance computing (HPC) systems. As architectural constraints limit the speed of individual devices, the component count in HPC systems is continuously growing. For example, future exascale systems are expected to combine the compute power of millions of CPU cores. Efficiently running systems at such scale will require a good understanding of their failure behavior.

In this paper we conduct an analysis of a decade of field data made available by Los Alamos National Lab. While previous work [12] has provided a high-level, general statistical summary of this data set, in this work we are particularly interested in identifying factors or circumstances that are predictive of future failures. Understanding what those factors are can help operators mitigate them, or take proactive measures against impending failures in cases where they cannot be avoided.

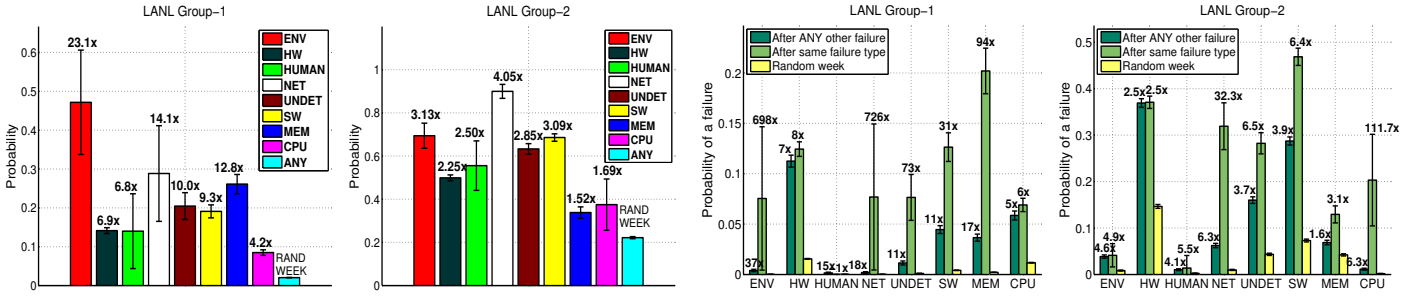
While there have been a number of papers analyzing failures in HPC systems, see for example [4]–[6], [10], [13], this prior work tends to be concerned with deriving statistical models that capture the observed failure process. For example, work that studies correlations between failures (which are relevant for predicting future failures and hence fall into the category of events we are interested in, in this work) usually does so by statistically modeling the empirical distribution of the inter-arrival time between failures or analyzing the auto-correlation function of the observed sequence of failures. While statistical models are very useful, for example in driving simulations or analyses of HPC systems, they are not all that helpful for operators in developing a good intuition for how and why their systems fail.

The goal of our work is to answer a set of specific questions to improve our understanding of failures in HPC systems, rather than providing a statistical model of failures. After providing a summary of the data set we use in our work in Section II, Section III looks into correlations between failures, including questions such as which failure types are most likely to generate follow-up failures. In Section IV we study whether some nodes are more likely to fail than others and why. Section V and Section VI address the question of how usage affects the reliability of a node. Sections VII, VIII, IX investigate the impact of environmental factors on node reliability, including the effect of the quality of power, the effect of temperature, and external factors, such as cosmic radiation. Finally, in Section X we put different pieces of our work together by performing a joint regression analysis including a diverse set of factors.

II. The data

Our study is based on failure data collected at 10 different high-performance computing (HPC) clusters at Los Alamos National Lab over a period of 9 years and is publicly available at [1]. We divided the 10 clusters into two different groups, based on their hardware architecture. Group-1 includes seven systems that are based on 4-way SMP (Symmetric Multi-Processing) nodes with one or two network interfaces (NICs) and a varying amount of main memory per node. In total these systems have 2848 nodes and 11392 processors. On the LANL web page, where the data is available, these systems correspond to the systems with IDs 3, 4, 5, 6, 18, 19 and 20. Group-2 includes 3 systems that are based on NUMA (Non-Uniform Memory Access) technology and contain a smaller number of nodes, but a larger number (typically 128) of processors per node. In total the systems in group-2 contain 70 nodes and 8744 processors, and correspond to the systems with IDs 2, 16, and 23 on the LANL web page.

For each of the systems the data contains records of all node outages that occurred during the measurement period, including information on the root cause of the node outage, the time when the outage happened and the ID of the node that was affected. The root cause of each failure falls into one of six high-level categories: *environment* failures, including power-outages for instance; *hardware* failures; failures resulting from



(a) *The probability that any node-failure follows a failure of type X.* (b) *The probability that a failure of type X follows other node-failures.*

Fig. 1. Correlations between failures in the same node

human-errors; *software* failures; *network* failures; and *undetermined*, whenever the root cause of the failure is unknown. The process of assigning failures to categories in LANL over the 9 years that the data spans was done by system administrators according to classification rules developed jointly by hardware engineers, administrators and operations staff [12]. Besides the high-level categorization of root causes, for many failures more detailed information is available, such as the hardware component responsible for a hardware failure.

In addition to logs of node outages, for some of the systems there is data on usage and the physical layout of nodes in the machine room available. In particular, group-1 systems have "machine layout" files that describe the position of each node inside a rack, and the location of a rack inside the server room. Additionally, detailed data on usage is available for two LANL systems: Systems 8 and 20. The usage records contain for each job information on the job submission time, job dispatch time (the time the job got dispatched from the queue to start running), job end-time, the number of requested processors and the ID(s) of the node(s) that this job was assigned to.

III. How are failures in HPC systems correlated?

The first question we are addressing in our work is how failures in HPC systems are correlated with each other. Discovering correlations between failures in HPC systems serves two purposes. First, it helps create a deeper understanding of their underlying root causes. Second, it helps in the prediction of failures, which is useful, for example, for scheduling application checkpoints or for designing job migration strategies.

Rather than building formal statistical models of correlations, we are interested in providing intuitive insights into correlations by answering questions, such as what types of failures increase the probability of future failures and by how much is the failure probability increased after a prior failure.

In order to quantify these dependencies, we use the data to determine the probability of a node failure in the time window following a previous failure and compare this probability to the probability of a node failure in a random window. We look at time windows of different lengths, including one day, one week and one month, and perform the calculations at three different spatial granularities: node level, rack level and system level. To test the statistical significance of our results all graphs include 95% confidence intervals. We also perform two-sample hypothesis tests to measure the significance of the difference between probabilities.

A. Correlations between failures within a node

In the first part of our correlation study we only focus on correlations between failures in the same node, i.e. we are asking the question whether current failure behavior of a node is predictive of its future failure behavior.

1) How does a failure affect the likelihood of later failures?: As a starting point, we calculate the daily and weekly probability of a node failure for group-1 and group-2 systems, i.e. the probability that a random node will fail in a random day/week. We then compare those probabilities against the probability of a node failing during a day or week following another failure (of any kind).

We find that the unconditional probability of a node failure on a random day is 0.31% and 4.6% for group-1 and group-2 systems, respectively. We observe that the daily failure probability is markedly higher during the 24 hours following a prior failure: 7.2% and 21.45% for group-1 and group-2 systems, respectively, which corresponds to roughly a 20X increase and 5X increase for groups 1 and 2, respectively. We observe similar, albeit somewhat weaker trends, for the entire week following a failure: the failure probability of a node in a given week increases from 2.04% to 15.64% in group-1 and from 22.5% to 60.4% in group-2.

2) Does the type of a failure affect the chance of follow-up failures?: Since we have information on the root cause of failures an interesting question is whether some types of failures increase the probability of follow-up failures more than others. To answer this question Figure 1-(a) shows the probability that a given node will fail within the one-week period following a failure of a particular type. The failure type is any of the six different categories of root causes that are distinguished in LANL: Environment, hardware, human error, network, software or undetermined failures. Each bar in the figure corresponds to one of those failure types. To provide a baseline, the right-most bar shows the probability for a node failing on a random week (not necessarily preceded by a failure).

Based on Figure 1-(a), we make several interesting observations. First, all types of failures increase the probability of failure in the following week, most commonly by factors of 7-10X in group-1 systems and factors of 2-3X in group-2 systems. For some cases, such as network or environmental failures in group-1 systems, the increase in failure probability

is more than 10X compared to a random week. We also note that prior failures increase the likelihood of later failures to significant levels. For example, while the probability of failure in a random week is only 2.04% in group-1 systems, chances are 30-50% that a node will experience a failure in the week following a network or environmental failure.

The second interesting observation is that the overall trends are very similar for group-1 and group-2 systems. In both cases the increase in failure probabilities is highest following a network or environmental failure. For group-1 systems a network or environmental failure increases the probability that a node will fail in a given week by a factor of 14-23X, and for group-2 systems it increases the failure probability by a factor of 3-4X.

We note that the factor increases are in general smaller for group-2 systems, since their baseline probability is higher. The probability for a node to experience a failure in a given week is 22.5% for a group-2 node (compared to only 2.04% for a group-1 node), which means the failure probability can not increase by more than a factor of 5X. The reason for the higher failure rates in group-2 systems is that the nodes in those systems are of a different type: they are NUMA nodes with 128 processors per node, compared to SMPs with 4 processors per node for group-1 systems, and the larger component count leads to higher failure rates.

3) Does the type of a failure predict the type of a follow-up failure?: Often it might be useful to know what type of failure to expect in the future. For example, are failures of type X usually followed by failures of type Y? To answer this question we computed all pairwise probabilities $p(x, y)$, where $p(x, y)$ is the probability of a failure of type Y in a week following a failure of type X, and compare this to the probability of a type Y failure in a random week.

Our first observation is that a failure always significantly increases the probability of a follow-up failure of the same type, and more so than a random failure. Figure 1-(b) shows the probability of a failure of type X in the week following a failure of type X, compared to the week following any type of failure, and compared to a random week. We observe that the increase in the failure likelihood can be dramatic. For example for group-1 systems, the probability of an environmental or a network failure in a given week increases by a factor of around 700X (to absolute values above 7%) if a failure of the same type was observed previously.

Besides correlations between failures of the same type, we notice significant correlations between network, environmental and software problems, i.e. each of these three types increases the follow-up probability of a failure of one of the other two types. We have been in discussions with operators at LANL and have not been able to come up with a clear explanation for these correlations. A closer analysis of the correlations between these three error types revealed that there are a few nodes who happen to have a relatively large number of network, environmental and software problems. It is possible that the correlation is biased by a few nodes that coincidentally had a large number of these three types of failures and does not imply a causal relationship.

4) How are hardware failures correlated?: We pay special attention to hardware failures since these are the single most common failure category: 60% of all failures are attributed to hardware problems. Our data set contains more detailed information on the root cause of hardware failures. The data shows that by far the most common types of hardware failures are due to problems with memory or CPU. 20% of hardware failures are attributed to memory and 40% are attributed to CPU.

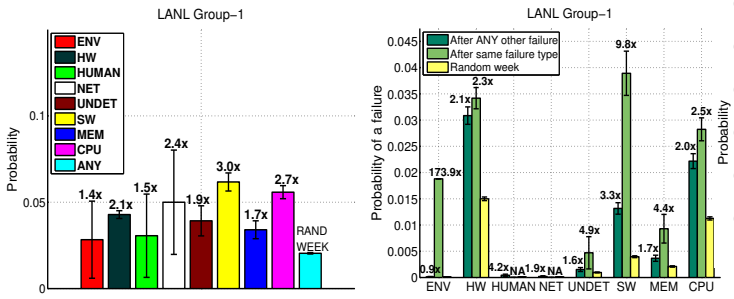
When repeating a correlation analysis similar to the one performed for the high-level failure categories, we find that past failures significantly increase the future probability of memory and CPU failures. In the week following a memory failure the probability of experiencing an additional memory failure is 20.23% for group-1 systems, a factor of nearly 100X increase over the probability of 0.21% in a random week. For group-2 systems, the weekly probability of a memory failure increases from 4.2% to 12.6%. All increases are statistically significant based on the two-sample hypothesis test.

The strong correlations between hardware-related failures allow us to draw some conclusions about the nature of these failures. Based on discussions with people at LANL, node failures that are attributed to memory or CPU problems are usually due to bit corruption events that go beyond what the built-in ECC can correct. This type of data corruption could either be due to soft errors, which are caused by random events, such as cosmic rays or random noise, or it could be due to hard errors, i.e. problems with the underlying hardware. The strong correlation between those errors points to hard errors as the more likely source of the problem, as one would not expect correlation between random events, such as cosmic rays. We study the impact of cosmic rays on hardware failures in Section IX.

B. Correlations between failures within a rack

The data for group-1 systems also includes information on the machine room layout, including the rack layout, which allows us to study how failures in different nodes in the same rack are correlated. We begin with the probability of a node failing (with a failure of any type) within a week following a failure (of any type) of another node in the same rack. We find that this probability is 4.6%, which is more than double the probability of a node failing in a random week (which is 2.04%). The increase in the daily probability is higher: the failure probability on a day following the failure of another node in the rack is 1.2%, which is nearly a factor of 3X higher than the baseline probability of 0.31%.

As we did in the case of correlations within the same node, we also looked at which failure types have the biggest effect on the probability of another node failing later on in the same rack. The results are shown in Figure 2 (left). We observe some increase in the failure probability for all types of failures, although with factors of 1.4–3X these are markedly lower than the increase of failures in the same node. Statistical testing with the two-sample hypothesis test allows us to conclude only for software failures that the probability of follow-up failures is significantly increased.



(a) *The probability that a node-failure of type X is followed by any failure in another node on the same rack.*

(b) *The probability that a node-failure of type X follows any failure in another node on the same rack.*

Fig. 2. Correlations between failures in the same rack

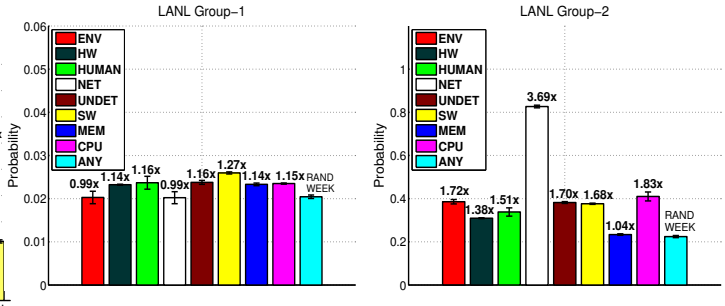


Fig. 3. Correlations between failures in the same system. Each bar corresponds to the probability that a node-failure of type X is followed by any failure in another node in the same system.

When looking at pairwise correlations, i.e. the probability of a failure of type y within a week of a failure of type x, we find that a failure of a particular type always increases the probability of the same type of failure within the following week. Moreover, this increase is much larger than the increase for the same type of failure following a random failure (i.e. not necessarily the same type). Figure 2 (right) summarizes our results. We observe an increase in failure probability as high as 170X for environmental failures and nearly 10X for software failures. All increases are statistically significant based on the two-sample hypothesis test.

Finally, we take a look specifically at hardware failures as these are the most common type of failure. We find that both memory and CPU failures experience a significant increase in probability in the day or week following another failure of the same type. This observation provides some room for hypotheses explaining the cause of such errors. One possible explanation might be that nodes in the same rack share similar environmental factors, such as the quality of the supplied power. This observation, combined with the strong effect of environmental failures on the frequency of follow-up failures motivates us to study environmental failures in more detail in Section VII.

C. Correlations between failures in the same system

In this section we ask the question of whether and how failures between different nodes in the same system (not necessarily in the same rack) are correlated. We find that the weekly probability of a node experiencing a failure does increase after another node in the same system had a failure, however the increase is significantly smaller than for nodes in the same rack: in group-1 systems the weekly probability of a node experiencing a failure increases from 2.04% to 2.68% and for group-2 systems it increases from 22.5% to 35.3%. Both increases are not significant enough to allow the rejection of the hypothesis that a node failure does not increase the likelihood of follow-up failures in nodes within same system, based on the two-sample hypothesis test.

The results are more interesting when breaking them down as a function of the failure type. Figure 3 shows the probability that a node in a system will fail within a week following a failure of type X (where X can be: *environment*, *hardware*, *human-errors*, *network*, *software*, *memory*, *CPU failures*, or *undeterminted*). We observe that software, hardware and human failures in a node in group-1 systems increase the probability that also other nodes in the system will see failures. The increase following software failures (a factor of 1.27X) is statistically significant based on the two-sample hypothesis test. For group-2 systems, all types of failures show an increase in Figure 3, but by far the biggest increase, with a factor of 3.69X, is observed following a network failure. The two-sample hypothesis test allows us to show that all failure types, except hardware and human, increase the chance of follow-up failures in other nodes significantly.

IV. Do some nodes in a system fail differently from others?

A. Do some nodes fail more frequently than others?

Figure 4 shows the total number of failures for each node in systems 18, 19 and 20 (the three largest systems of all LANL systems in terms of number of nodes: 1024, 1024 and 512 nodes, respectively). The graphs show that in all systems a single node (the node with ID 0) had significantly more failures than rest of nodes. For example, for system 20 node 0 reported 19 times more failures than the average node and for system 19 node 0 reported more than 30 times higher failure rates than the average node. To test the significance of differences between failure rates in nodes, we performed chi-square tests for differences between proportions: with 99% confidence level we are able to reject the null hypothesis that all nodes in each system had equal failure rates (p-value $<$ 2.2e-16). Interestingly, even when repeating the same analysis after removing node 0 we can still reject the hypothesis that all nodes in each system had equal failure rates.

B. Are the failure characteristics of failure prone nodes different from other nodes?

We are interested to find out whether the increased number of failures in some nodes is due to an increased number of failures of a particular type or due to generally increased failure rates. To answer this question we compare in Figure 5 the relative breakdown of the different failure types for failure prone nodes against the remainder of the system, and we compare in Figure 6 for each failure type the probabilities of a node failure of this type in failure prone nodes vs the rest of the nodes in the systems. In Figure 6 each plot contains three pairs of bars for each of the three systems, where each pair corresponds to a timespan: day, week or month. The numbers on top of the bars indicate the factor increase in failure probability in a failure prone node compared to an average node.

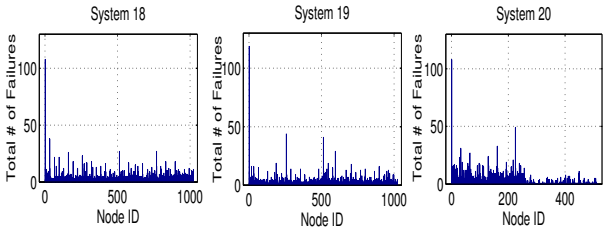


Fig. 4. Total number of failures as a function of Node-ID

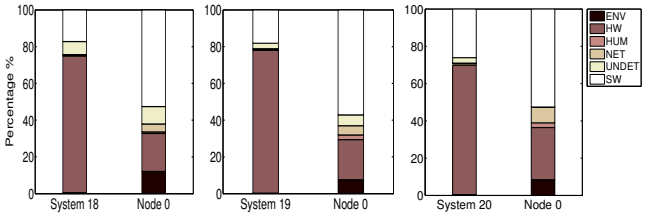


Fig. 5. Root cause breakdown in failure prone nodes vs other nodes

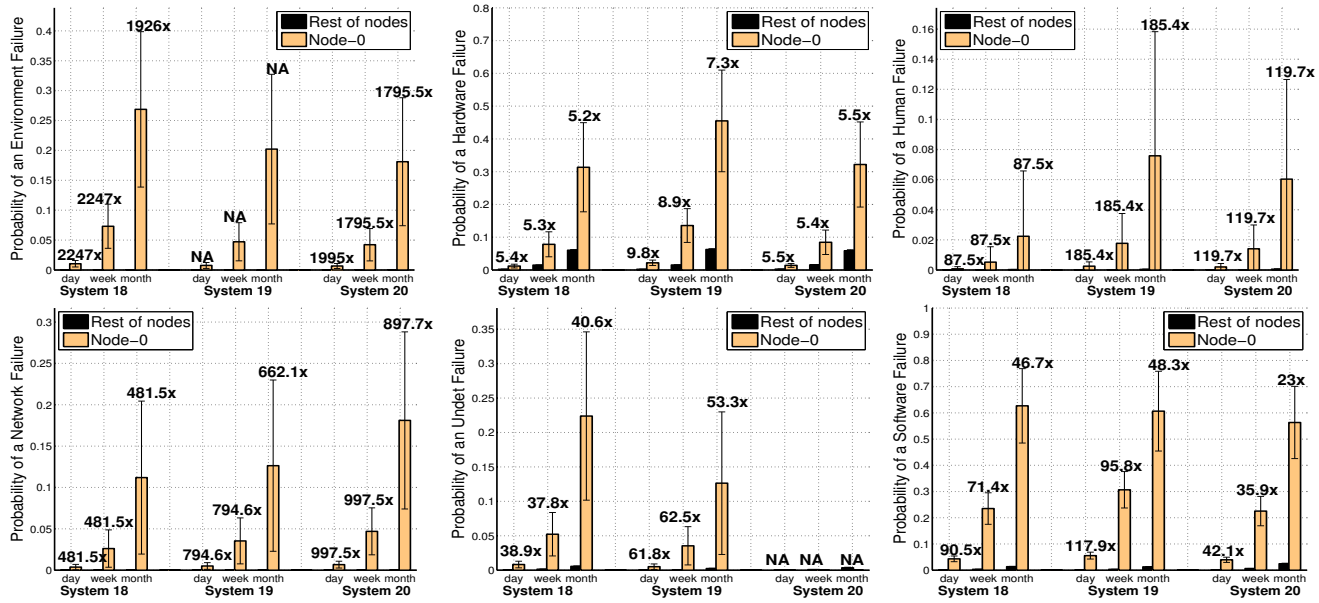


Fig. 6. The probability of different failure types in failure prone nodes compared to the rest of the nodes in a system.

The first observation we make based on Figure 6 is that node 0 exhibits increased failure probabilities for all types of failures, so the higher failure rate in those nodes cannot be attributed to a particular type of failure. However, we observe that the increase in failure probabilities is particularly high for environmental and network failures, with factors of increase in the 2000x and 500x-1000x range, respectively. Software failure rates are also significantly higher in node 0 than the remainder of the system (factors of 36X up to 118X). The increase in the probability of hardware failures is modest in comparison, but still significant with factors in the 5–10X range. To formalize our results we repeat the chi-square test for differences between proportions separately for each failure type. The only failure type where we fail to reject the null hypothesis that nodes fail with equal rates is for failures due to human errors; for all other failure types the test rejects the null hypothesis with 99% confidence.

Turning to Figure 5, which shows the relative breakdown of failures by root cause for the failure prone nodes compared to the whole system, we observe a higher percentage of software, environment and network failures in the failure prone nodes. This observation is in agreement with our findings in Figure 6, which indicate that those three failure types have a higher factor increase in the failure prone nodes than other failure types. It is interesting to note that in the failure prone nodes the dominant failure mode shifts from hardware failures to software failures.

C. *Why do some nodes fail more frequently than others?*

One might wonder what the reason for the high variability in failure rates between nodes in the same system is, in particular since all nodes within a system typically use the same type of hardware. One possible explanation are statistical effects due to the strong correlations between failures in the same node (recall Section III). Once a node is "unlucky" and starts to develop failures, a large number of correlated follow-up failures might bring the total failure rate of a node way above the average.

Another hypothesis we investigated is the effect of a node's position in the machine room or inside the physical rack. For a few systems where we had information on the layout of nodes in the machine room we checked whether the location in the machine room or the location of a node within a rack played any role, but we could not find any clear patterns that certain areas in the machine room were more likely to be correlated with higher error rates.

One more hypothesis that we tested is whether usage has an effect on the failure rate of a node and whether the failure prone nodes were used differently from other nodes. We will look at our analysis of usage in the following two sections.

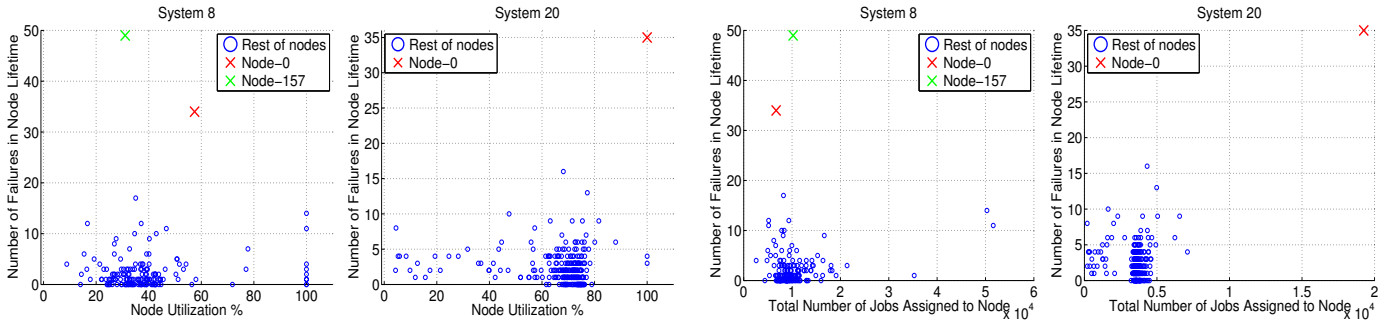


Fig. 7. The impact of usage on node reliability. (a) Node failures vs node utilization. (b) Node failures vs node jobs

V. What is the effect of usage on a node's reliability?

The effect of system workload on system reliability was studied in a series of papers by Iyer et al. [8], [9] and Castillo et al. [2]. However, these papers date back to the early 1980's and don't necessarily translate to modern HPC systems.

We therefore used job logs that are available for two of LANL's systems, system 8 (where we have a total of 763,293 job records), and system 20 (with a total of 477,206 job records), to study whether the way a node is used affects its failure rate. These two systems are representative of two larger groups of LANL systems, where all systems within the same group shared a similar hardware architecture and ran very comparable workloads.

We consider the effect of two simple usage metrics, one is the average node utilization (where we define a node as being utilized if at least one job is currently assigned to it, and idle otherwise) and the other one is the number of jobs that were scheduled on a node throughout its lifetime. We begin by plotting the number of failures a node experiences against the node's average utilization (see Figure 7-(a)) and against the number of jobs served by the node (see Figure 7-(b)). We have marked nodes with particularly high failure rates with special markers. This includes node 0, which we discussed in the previous section.

We observe that in both systems where we have usage information available the failure prone node 0 tends to be among the nodes with the highest utilization and the largest number of jobs assigned to it. We formalized our observation by looking at the Pearson correlation coefficient between the number of jobs assigned to a node and the number of failures experienced by the node. For both systems we observe clearly positive correlation coefficients of 0.465 and 0.12, respectively. However, repeating our analysis after removing node 0 reduces the correlation to insignificant levels, which lets us conclude that the strong linear correlation between usage and failures, as captured by Pearson's coefficient, is mostly due to node 0. In discussions with operators at LANL we have been told that node 0 in most systems has a special role where it is used as a login node for users and/or is used to schedule and launch jobs.

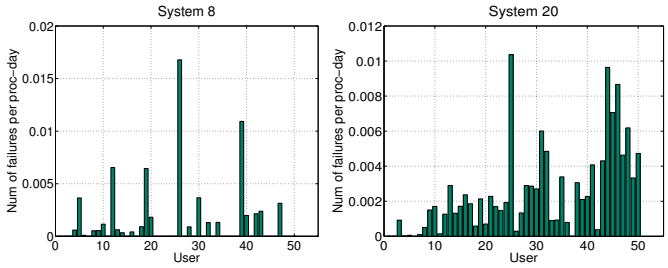


Fig. 8. Distribution of failed jobs over different users.

VI. Are some users more prone to node failures than others?

As a follow-up question on the relationship between usage and failures we used the job logs to test whether certain users are more likely to experience job failures than others. Here we only include job failures that are caused by failures in one of the underlying nodes, rather than a failure of a user's application software. The two systems that have job logs available (systems 8 and 20) both have more than 400 different users. For each system, we focus on the 50 heaviest users in terms of the number of processor-days that they used on those systems.

The two graphs in Figure 8 show for each of the 50 heaviest users the average number of failures this user experienced per processor-day that this user utilized the system. Visual inspection shows a large discrepancy between the failure rates experienced by different users. We also formally verified that the difference in failure rates between users is statistically significant by using Poisson regression to fit a full (saturated) model (with users' actual failure counts and usage periods), and a common failure rate model (where all users have the same failure rate). We then applied Analysis of Variance (ANOVA) test and found with 99% confidence level that the saturated model is significantly better than the common rate model.

In conclusion, we find that the way a node is exercised affects its failure rates. This might for example be because some users run applications that are more likely to exercise a buggy code path in some system software or because their application is more likely to exercise a hardware component in an access pattern that makes intermittent or hard errors more likely to manifest themselves.

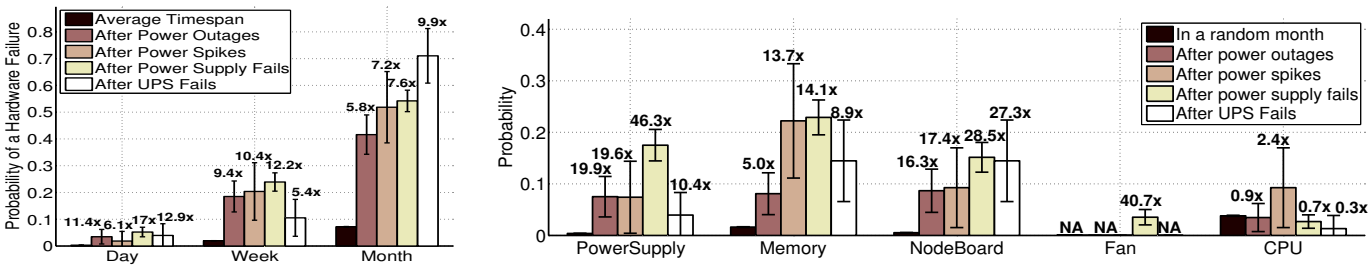


Fig. 10. Impact of power problems on hardware failures

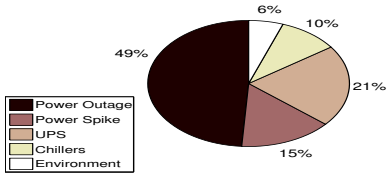


Fig. 9. Breakdown of environmental failures in LANL systems

VII. What is the impact of environmental factors, in particular problems related to power?

We have observed in Section III that environmental failures cause a steep increase in the probability of follow-up failures. A node with an environmental failure has a chance of 47.2% and 69.4% for group-1 and group-2 systems, respectively, of experiencing another failure within a week. This observations warrants a closer look at what environmental failures are and how they affect other failures.

The LANL data provides a breakdown of the high-level root cause category of environmental failures into lower-level sub-categories. Figure 9 presents a breakdown of the observed environmental failures. We observe that the majority of those failures are related to problems with power in the datacenter, in particular either power outages, power spikes or UPS failures. In the remainder of this section we study how power issues affect the two most common types of failures, hardware and software failures. In addition to power outages, spikes and UPS failures recorded as part of environmental failures, we also take into account the effect of problems with the power supply unit of individual servers, which are recorded as hardware problems.

A. How do power problems affect hardware failures?

Figure 10 (left) shows the probability that a node will experience a hardware failure within a day (left-most set of bars), a week (middle set of bars) and a month (right-most set of bars) after experiencing a power outage, a power spike, a power supply failure or a UPS failure, compared to the probability of a hardware failure in a random day, week, month (i.e. not necessarily preceded by a power issue).

We observe that generally after power issues the probability of seeing hardware failures in LANL nodes is significantly increased. Interestingly, while power outages and power supply failures caused a significant increase in hardware failures both in the short-term (within a day following the power problem) and in the long-term (within a month following the power problem), the effect of power spikes is more apparent at longer timespans. In the long-term, all four types of power issues lead to an increase in the hardware failure probability by factors of 5-10X.

1) What types of hardware failures are most affected by power problems?: Figure 10 (right) shows the probabilities for different types of hardware failures to occur within a month of a power outage, power spike, power supply failure or a UPS failure, compared to the probabilities of those failures in a random month (not preceded by power issues).

We observe that a large range of hardware components, including memory DIMMs, node boards, and power supplies, show markedly increased failure rates following power problems. The only component that showed no clear signs of increased failure rates after any of the power problems are CPUs. For the other components the degree at which failure rates increase depends on the type of power problem that preceded. After power outages the node board and power supply show the biggest increase in their failure rates (factors of 16-20X). These components also show similar failure rates following power spikes. Memory DIMMs show a higher failure rate following power spikes, compared to power outages, with an increase of 13.7X compared to 5X. For all components the increase in failure rates is strongest following a power supply failure, and ranges from more than 40X for fans and power supplies, to 14X and 28X for memory DIMMs and node boards. Two components show high failure rates following UPS failures: node boards (27.3X increase) and memory DIMMs (8.9 increase).

2) Do power problems cause issues in addition to node failures?: When analyzing the LANL data to investigate the consequences of power problems, we also made another interesting observation. In addition to the clearly increased number of node outages due to failures following a power problem, we observe a large increase in the number of non-scheduled maintenance events related to hardware problems. Within a month after a power outage or power spike, around 25% of affected nodes need to undergo unscheduled downtime due maintenance. This is an increase of nearly 90X in the frequency of unscheduled maintenance compared to a random month in a node's lifetime. In the month after a power supply failure maintenance activity is also markedly increased: a node has an 8% chance of requiring hardware-related maintenance work within a month after a power supply failure, which is lower than after a power outage or spike, but still nearly 30X higher than in a random month. Failures in the UPS system had the strongest effect, increasing a node's chance of undergoing unscheduled maintenance by a factor of 100X (28% chance).

These results indicate that problems with power not only lead to hardware problems that cause a node to fail, but also a significant amount of downtime due to unscheduled maintenance.

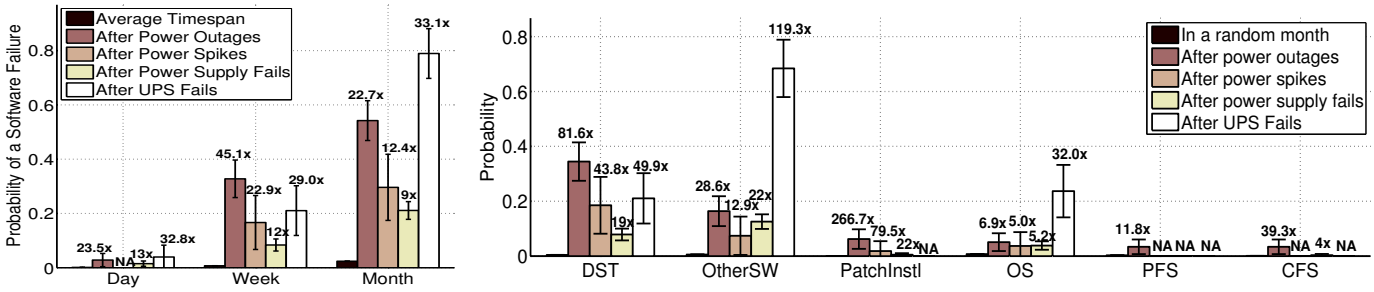


Fig. 11. Impact of power problems on software failures

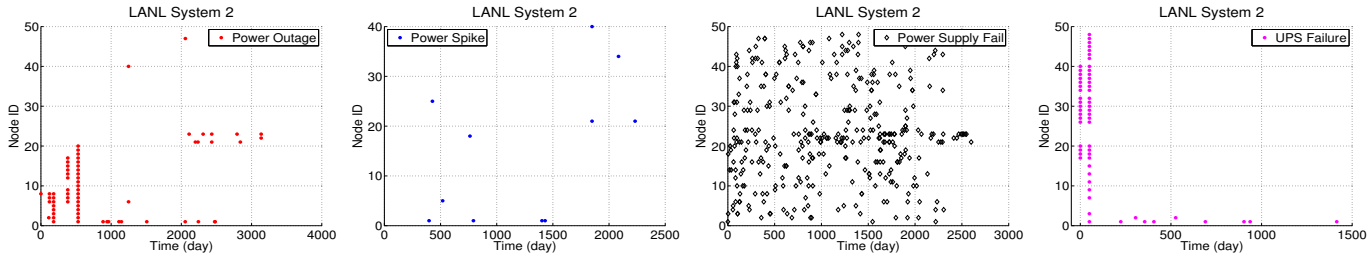


Fig. 12. Distribution of power-related failures across nodes over time (LANL System 2)

B. How do power problems affect software failures?

Figure 11 (left) shows the probability that a node will experience a software failure within a day (left-most set of bars), a week (middle set of bars) and a month (right-most set of bars) after experiencing a power outage, a power spike, a power supply failure or a UPS failure, compared to the probability of a software failure in a random day, week, month (i.e. not necessarily preceded by a power issue).

As was the case for hardware failures, we observe that after power issues the probability of seeing software failures in LANL nodes is significantly increased. We observe the strongest effect for power outages and UPS failures, which increase the probability of a software failure within a week by factors of 45X and 29X, respectively. Power spikes and power supply failures had a somewhat weaker effect, with factors of 10-20X, but still very strong. All four types of power problems show longer-term effects, as evidenced when looking at the software failure rates following the month of power problem, although the effects are weaker than the weekly ones (except for UPS failures).

1) What types of software failures are most affected by power problems?: Figure 11 (right) shows a breakdown of software failures into their more detailed underlying root causes and for each of these underlying root causes the associated probability within a month after a power outage, power spike, power supply failure and UPS failure. We observe that the majority of the software-related outages following power issues are related to the system's distributed storage system (DST). Some additional issues are related to Parallel File System (PFS) and the Cluster File System (CFS).

In summary, we observe that a large fraction of software issues created by power problems are related to data storage (either the distributed storage system or the file system), rather than general operating system issues or other software issues. While the data does not provide details on the nature of those storage and file system failures, the loss of power likely led to some inconsistency in the storage or file system state. All file and storage systems for HPC installations provide mechanisms to protect against loss of consistency or persistence in the case of crashes or power outages, so it's interesting to observe that despite those efforts power problems still remain a high risk factor for those systems.

C. How are power problems laid out in time and space?

Figure 12 illustrates how the four different types of power problems (outages, spikes, UPS and power supply failures) are layed out in time and space using the data for all System 2 nodes. We chose System 2 as it provides the largest data set on power issues. We observe that the different power problems vary in how they are correlated in time and space. While power outages and UPS failures show clear correlations between nodes and also over time within the same node, power spikes tend to happen in more random unpredictable ways. Power supply failures are the most common type of power-related failure and show only correlations within the same node.

VIII. How does temperature affect failures?

Understanding the effect of temperature on system reliability is important as a large fraction of a datacenter's energy bill goes into cooling. Recent work [3] indicates that the impact of temperature on hardware components might be much weaker than often assumed and reports for some types of errors, such as DRAM errors no correlation at all.

A. How does average temperature affect failures?

LANL has provided event logs for some of their systems, in addition to the logs of failure outages. For one of LANL's systems (system 20) periodic temperature measurements from a motherboard sensor are available. It is worth noting that the cooling mechanism used in the facilities hosting the rest of LANL's systems was similar to system 20's: hot-aisle cold-aisle air cooling through perforated floors.

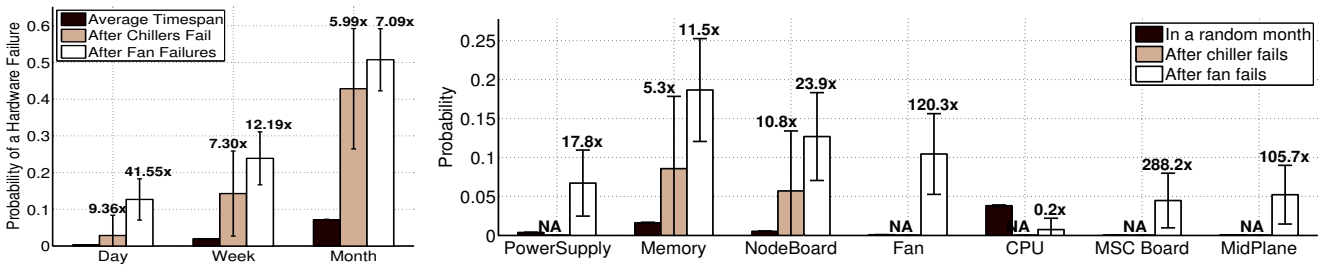


Fig. 13. Impact of temperature related problems on hardware failures

The work in [3] uses this data for system 20 to study how node outages, either due to hardware failures in general or DRAM failures in particular, change with temperature. Their results show no correlation between a node's average temperature and node outages, within the temperature range that the data comprises.

We have formalized the work in [3] by using regression analysis to model the occurrence of node outages due to hardware failures as a function of a node's average temperature. We used two commonly used regression models, Poisson regression and negative binomial regression. In agreement with [3], we find that average temperature is not correlated significantly with the occurrence of hardware failures. When repeating our analysis for CPU and DRAM failures separately, we also find that average temperature is not significant to either type of failure.

B. How do temperature excursions affect failures?

Previous work [3] has only considered the effect of average temperature, but not looked at the effect of temporary excursions to very high temperatures. Therefore, rather than looking at average temperature, we repeat the same regression analysis as above for the maximum temperature observed for a node and the variance in temperature across all temperature recordings for a node. We find that the maximum temperature and temperature variance are insignificant to the occurrences of hardware failures in general and CPU and memory failures in particular.

Temperature recordings are only available for one system and since they consist of periodic samples, they might miss brief periods of very high temperatures. For a broader study of the effects of brief periods of high temperature we look at the impact that a fan failure or a failure in the chiller system has on the nodes. Fan and chiller failures will lead to temporarily increased temperatures at a node, and depending on whether it's a complete or partial failure can lead to to extreme temperatures inside a node, making a node shutdown necessary.

Figure 13 (left) shows the impact of fan failures and chiller failures on hardware failures. The graph shows the probability that a node will experience a hardware failure within a day, week and month following a fan or a chiller failure, compared to the probability of a hardware failure in an average day, week and month. We observe clearly increased hardware failure rates following fan and chiller failures for all timespans. Fan failures have a stronger effect for all timespans, with a factor of 40X increase in hardware failure rates on the day following a fan failure (compared to a random day). Chiller failures had a weaker effect across the different timespans, with factors of 6-9X increase in hardware failure rates.

We also ask what type of hardware failures are likely to follow fan and chiller failures. Figure 13 (right) shows for each of the hardware components with corresponding records in the data the probability of failure within a month after a fan or a chiller failure, compared to a failure of that component in a random month. We find that all hardware components, except for CPUs, show an increase in the failure rate following a fan failure. We find that for memory, node boards, and power supplies the order of magnitude of the increase is similar to the one observed after power problems, with factors of 10-20X. In addition, we observe significant increases in failure rates for two types of boards, MSC boards and midplanes, which we did not observe in the case of power problems. One of the largest increases in failure rates, a factor of 120X, occurs for fans, which is maybe not surprising given that we have observed previously that most failure types have the strongest correlation with events of the same type. Chillers failures seem to only affect two components: memory DIMMs and node boards, with 5.3X and 10.8X increases in their probabilities, respectively.

Overall, our analysis shows that hardware components are well able to tolerate higher average temperatures within the ranges that are typically observed in a datacenter. The harmful effects of temperature mostly stem from short periods of extremely high temperatures, for example due to the failure of a fan in the system.

IX. External factors: cosmic radiation

It is known that high rates of cosmic radiation can lead to soft errors due to bit flips in DRAM or on system buses. If the built-in error correcting codes (ECC) are not strong enough to correct the corrupted bits, those errors will lead to a machine crash or shutdown. Cosmic rays and their effect on system reliability are a major concern, and, for example in the case of DRAM errors, most of the existing work on DRAM reliability focuses on the effects of cosmic radiation.

Since the LANL data spans a very long time period (nearly a decade), it covers almost an entire solar cycle (typically 11 years long), including several solar flares. Records of high-energy neutron counts that are produced by cosmic rays in the atmosphere are collected at many neutron monitor (NM) stations around the world. We use data of 1-minute resolution neutron counts collected at a NM station in Climax, Colorado (geographically close to Los Alamos National Lab) [11].

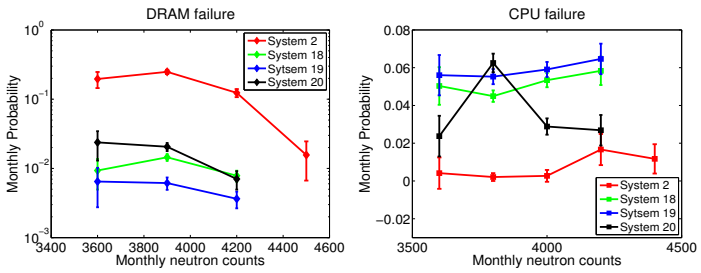


Fig. 14. Probability of CPU and memory failures as a function of average monthly neutron counts

TABLE I
SUMMARY OF REGRESSION VARIABLES

Variable	Category	Description
fails_count *(response variable)*	Failures	This is the response variable; the total occurrences of node outages in a node's lifetime.
Input Variables		
avg_temp	Temperature	The average ambient temperature of a node
max_temp	Temperature	The maximum temperature reported by a node
temp_var	Temperature	The variance of all temperatures reported by a node
num_hightemp	Temperature	The number of severe temperature warning messages reported by a node (i.e. when its ambient temperature exceeds 40C)
num_jobs	Usage	The number of jobs that were assigned to the node in the observation period
util	Usage	The utilization of a node during the observation period
PIR	Layout	Position in Rack: the position of a node inside the physical rack (1=most bottom, 5=top)

We use this data to analyze whether periods of increased cosmic rays are correlated with a higher rate of hardware errors, in particular failures related to DRAM and the CPU.

We begin by studying whether the likelihood of a node outage due to DRAM failure changes with neutron flux levels. Figure 14 (left) shows the monthly probability of a DRAM failure as a function of the monthly average neutron counts-per-minute, for LANL Systems 2, 18, 19 and 20. We focus our analysis on the LANL systems that span the longest lifetimes, or consist of the largest numbers of nodes, across all systems. We find that months with higher neutron rates are not associated with higher rates of DRAM failures. These results might be unexpected, since cosmic rays are known to increase soft error rates in DRAM. One possible explanation is that while increased rates of cosmic rays might lead to a higher number of corrupted bits, the types of corruption caused by those events might usually be correctable with the built-in ECC. This explanation agrees with recent findings in [7], which provide evidence that most node outages that are due to errors in DRAM are likely caused by hard errors, i.e. problems with the underlying hardware, rather than random events, such as cosmic rays.

Cosmic ray-induced neutrons can also cause CPU faults, possibly leading to a machine crash or shutdown. We repeat our correlation analysis using data on node outages that were attributed to CPU failures, rather than outages due to DRAM problems (see Figure 14 (right)). We observe that in three systems (2, 18 and 19), CPU failures were slightly more likely to occur in months with relatively higher neutron rates.

TABLE II
POISSON REGRESSION COEFFICIENTS

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	2.0232	0.8288	2.44	0.0146
avg_temp	0.0546	0.0337	1.62	0.1046
max_temp	-0.0705	0.0339	-2.08	0.0373
temp_var	0.0253	0.0333	0.76	0.4479
num_hightemp	0.0210	0.0698	0.30	0.7639
num_jobs	0.0004	0.0001	7.17	0.0000
util	-0.0268	0.0050	-5.34	0.0000
PIR	-0.0262	0.0358	-0.73	0.4654

TABLE III
NB REGRESSION COEFFICIENTS

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.5478	1.1930	1.30	0.1945
avg_temp	0.0499	0.0462	1.08	0.2802
max_temp	-0.0510	0.0475	-1.07	0.2828
temp_var	0.0252	0.0449	0.56	0.5744
num_hightemp	0.0021	0.0937	0.02	0.9820
num_jobs	0.0004	0.0001	3.86	0.0001
util	-0.0248	0.0073	-3.42	0.0006
PIR	-0.0345	0.0488	-0.71	0.4794

X. PUTTING IT ALL TOGETHER

In the previous sections we have looked at a number of factors and their impact on node outages in HPC systems in isolation. Our goal in this section is to put all the different pieces together. Rather than studying the individual effect of these factors separately, we now ask the question of what the collective effect of these different factors, combined, is on long-term HPC node reliability. The only LANL system that allows us to explore this question is system 20 where we have data on node outages, node usage, physical layout and ambient temperature.

We use regression analysis to model occurrences of node outages in system 20 as a function of node usage, physical location and temperature. More precisely, we use the total number of outages a node experiences during the data collection period (due to any type of failure) as our response variable, which we try to express by the set of predictor (explanatory) variables summarized in Table I. We use two commonly used regression models, Poisson regression and negative binomial regression.

The results of applying Poisson regression and negative binomial regression are shown in Tables II and III, respectively. The two right most columns show the test statistic and the p-value, respectively, that the null hypothesis that each predictor's coefficient is zero given that the rest of the predictors are in the model. Interestingly, we observe similar results for both models: The predictors *num_jobs* (i.e. the number of jobs assigned to a node during the observation period) and *util* (i.e. the node's average utilization) are each statistically significant in both models; with 99% confidence, we can reject our null hypothesis and conclude that each one of them is statistically different from zero given that the rest of the coefficients are in the model. Since we know from Section V that node 0 in this system exhibited a strong linear correlation between usage variables and number of failures, we reran our regression models after removing node 0 from the data and found that utilization remains significant to the model, although the significance level drops slightly.

In addition to usage variables, we observe that for the Poisson model *max_temp* is statistically significant to the frequency of node outages (recall that maximum temperature was found insignificant to *hardware* failures in particular, in Section VIII). However, when rerunning the model with only the significant predictors, the significance level of max_temp in the Poisson model drops.

We find these results to be a strong indicator that a node's usage and utilization levels have a stronger impact on a node's failure rates than other factors, such as its ambient temperature or physical location inside a rack.

XI. Lessons learned

Below we highlight some of the key findings of our work and derive some lessons we learned from our analysis.

• In agreement with prior work, we observe strong correlations between failures in HPC systems. During the day following a failure, a node is 5–20X more likely to experience an additional failure, when compared to a random day. Similar, albeit weaker trends exist across the nodes in the same rack: a node's failure probability is increased by a factor of 3X during the day following another node failure in the same rack. The trends were significantly weaker when looking at other nodes in the same system (not in the same rack). We found that software and network failures in one node increase the probability of subsequent failures of other nodes in the same system by factors of 1–3X.

• Interestingly, we observe that some types of failures increase the likelihood of follow-up failures more than others. In particular, environmental failures (such as power outages) and network failures have a very strong effect on subsequent failures: 30–50% of nodes experience at least one failure in the week following a network or environmental failure, compared to only 2% in an average week. These observations are critical for creating effective failure prediction models, as they imply that such models should not only account for correlations between failures in time and space, but also consider the *root-causes* of failures.

• The strong chance of follow-up failures after environmental failures, which in our data are mostly due to power outages, motivated us to study the effects of power problems more broadly. We considered four different events: power outages, power spikes, UPS failures, and failures of a node's power supply units, and found that they all lead to significantly increased hardware failure rates, as well as unscheduled maintenance events.

• Our observations on increased failure rates in memory DIMMs and node boards following power spikes, UPS failures and power supply problems suggest that after such events one might want to thoroughly inspect these hardware components for problems. Suspected fans should also be properly inspected in the case of a power supply failure since they were 40X more likely to fail in the following month, than in an average month. In general, we find that a bad or failing power supply can lead to many auto-correlated node outages and therefore should be quickly fixed or replaced.

• Power outages have another interesting effect: significantly increased rates of software issues. A large fraction of the software failures following within a month of a power outage were either related to the distributed storage system or the file system. This observation might hold evidence that stronger mechanisms are required to protect storage and file system consistency in the face of power outages.

• The large cost of datacenter cooling motivated us to study the effect of temperature on node reliability. We do not observe a significant correlation between the average temperature at a node and its likelihood of failure. However, when studying the effect of node outages due to fan or chiller failures, which likely cause a brief period of very high temperatures inside a node, we do observe a strong increase in the subsequent rate of failures. Hardware components most strongly affected were MSC boards and midplanes (>100X increase in failure probabilities), but memory DIMMs, power supplies and node boards also experienced increased failure probabilities (>10X increase). This shows that hardware components are well able to tolerate higher average temperatures within the ranges that are typically observed in a datacenter. The harmful effects of temperature mostly stem from short periods of extremely high temperatures, for example due to a fan failure.

• Another environmental factor we studied is cosmic ray-induced neutron flux, which can lead to increased soft error rates. Interestingly, we observe no effect on failures due to DRAM errors, which might indicate that built-in error correcting codes are generally sufficient to mask bit flips in DRAM due to soft errors (and that those DRAM errors that lead to node outages are more likely due to hard errors). On the other hand, CPU failure rates, which did not show a strong correlation with other types of failures or environmental factors, such as power or temperature, are positively correlated with cosmic rays-induced neutron flux.

• When studying the effect of a node's usage on its failure rate, we find that nodes with higher utilization and a higher number of jobs assigned to them experience higher failure rates. Moreover, when studying the number of failures experienced by different users of the system, we find that some users experience a significantly higher failure rate per processor-day of usage of the system. Since we exclude problems caused by the users' application software, this skew is not due to users' varying abilities to write stable code. Instead, we conclude that the way a node is exercised affects its failure behaviour.

• We observe that some nodes fail significantly more frequently than others, even in systems where all nodes are identical in terms of their hardware. When we looked more closely at the most failure prone nodes in LANL's systems, we found that they encountered higher-than-average rates of all types of failures, but the increase was strongest for software, network and environment failures. One of the possible reasons that we investigated is the location of a node within the machine room, but we find no indication that certain areas in the machine room are more failure prone than others. Instead, we find that the failure prone nodes were typically used differently from the rest of nodes.

• When performing a joint regression analysis, where we model node reliability as a function of different aspects of physical location, temperature and usage, we found that usage related variables were the most significant.

ACKNOWLEDGMENTS

We would like to thank LANL for making their data publicly available and for answering our questions about the data. We would also like to thank the US National Oceanic and Atmospheric Administration for making the neutron count data from Colarado NM station publicly available. Special thanks goes to Sotirios Damouras for his insightful answers to our statistics questions. Finally, we thank our anonymous reviewers for their feedback and comments on the paper. This work has been funded by an NSERC discovery grant.

REFERENCES

[1] Operational Data to Support and Enable Computer Science Research, Los Alamos National Laboratory. http://institute.lanl.gov/data/fdata/.

[2] X. Castillo and D. Siewiorek. Workload, performance, and reliability of digital computing systems. In *Proc. of International Symposium on Fault-Tolerant Computing*, 1981.

[3] 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 *Proc. of SIGMETRICS 2012*.

[4] D. Ford, F. Labelle, F. I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan. Availability in globally distributed storage systems. In *Proc. of OSDI'10*, 2010.

[5] S. Fu and C.-Z. Xu. Exploring event correlation for failure prediction in coalitions of clusters. In *Proc, of SC'07*, 2007.

[6] E. Heien, D. Kondo, A. Gainaru, D. LaPine, B. Kramer, and F. Cappello. Modeling and tolerating heterogeneous failures in large parallel systems. In *Proc, of SC'11*, 2011.

[7] A. A. Hwang, I. A. Stefanovici, and B. Schroeder. Cosmic rays don't strike twice: understanding the nature of DRAM errors and the implications for system design. In *Proc. of ASPLOS'12*, 2012.

[8] R. K. Iyer and D. J. Rossetti. Effect of system workload on operating system reliability: A study on ibm 3081. *IEEE Trans. Softw. Eng.*, 11(12), Dec. 1985.

[9] R. K. Iyer, D. J. Rossetti, and M. C. Hsueh. Measurement and modeling of computer reliability as affected by system activity. *ACM Trans. Comput. Syst.*, 4(3), Aug. 1986.

[10] Y. Liang, Y. Zhang, M. Jette, A. Sivasubramaniam, and R. Sahoo. BlueGene/L failure analysis and prediction models. In *Proc. of DSN'06*, 2006.

[11] National Oceanic and Atmospheric Administration. Cosmic ray neutron monitor data. *http://www.ngdc.noaa.gov/stp/solar/cosmic.html*.

[12] B. Schroeder and G. Gibson. A large-scale study of failures in high-performance computing systems. In *Proc. of DSN'06*.

[13] T. Thanakornworakij, R. Nassar, C. Leangsuksun, and M. Paun. The effect of correlated failure on the reliability of HPC systems. In *Proc. of Parallel and Distributed Processing with Applications Workshops (ISPAW)*, 2011.