Soft Failures in Large Datacenters

Sriram Sankar and Sudhanva Gurumurthi

Abstract— A major problem in managing large-scale datacenters is diagnosing and fixing machine failures. Most large datacenter deployments have a management infrastructure that can help diagnose failure causes, and manage assets that were fixed as part of the repair process. Previous studies identify only actual hardware replacements to calculate Annualized Failure Rate (AFR) and component reliability. In this paper, we show that service availability is significantly affected by soft failures and that this class of failures is becoming an important issue at large datacenters with minimum human intervention. Soft failures in the datacenter do not require actual hardware replacements, but still result in service downtime, and are equally important because they disrupt normal service operation. We show failure trends observed in a large datacenter deployment of commodity servers and motivate the need to modify conventional datacenter designs to help reduce soft failures and increase service availability.

Index Terms—Datacenter, Reliability, Characterization, Management.

1 Introduction

LARGE enterprises house hundreds of thousands of servers in globally distributed datacenters to support the growth in online services and cloud computing. The cost of providing such services is a significant expenditure for these enterprises, and hence infrastructure efficiency is a key value consideration. As with any large system, a datacenter experiences failures at different levels of its architecture, ranging from critical power infrastructure to server component. The datacenter itself becomes a warehouse-scale computer [2] and failures can cause service unavailability for hosted services. Among component failures, some result in actual replacement of hardware components; however, another class of datacenter failure does not result in actual replacement while still causing service downtime. This is an important class of failures to look at, because they can cause service unavailability, and incur the cost for a technician to identify and address the reason for failure. We classify these failures as ***soft failures***. This term has already been used in literature for transient failures in CPUs and memory (for example, particle strikes) [5], but it is uncommon in the context of large datacenters. In this paper, we analyze failure data collected from large-scale datacenter deployments housing tens of thousands of commodity servers. We then identify trends in fixes made for the failures and highlight the soft failures issue. We then propose methodologies that can help increase the overall availability for the datacenter while masking soft failures.

Previous studies have looked at hardware errors in large datacenters [6, 8, 9, 10]. However, there has been minimal work that looked at soft failures. This study characterizes soft failures in large clusters of machines using data collected over a year of observation. We attempt to answer some non-trivial questions including, the downtime experienced by a machine when it experiences a soft failure, probability of a machine having a soft failure once it has already experienced soft failures, the categories of most likely fixes following a soft failure, average number of days to the next failure event following a soft failure, and time histograms describing other interesting trends.

Following our characterization of soft failures, we present a discussion on datacenter architecture and services that these failures affect, and propose possible ways of addressing these failures. Since the underlying root causes of such failures are not fully understood, our objective in this early work is to highlight the issue of soft failures, and to identify possible research avenues for potentially significant solutions to this problem.

2 Soft Failure Characterization

2.1 Percentage of Soft Failures

Figure 1 shows the percentage of soft failures in a population of machines belonging to the same cluster, which contains tens of thousands of machines. These clusters include servers hosting representative large scale online services like Websearch, Cosmos and Email [4]. The servers are designed for scale-out scenarios with cost optimization around dual processors and enterprise SATA disk drives [4]. The automated management layer [3] classifies failures and pushes a machine from healthy operational state to a repair state, where it is triaged before actual repair is done.

Sriram Sankar is with Microsoft Corporation and the University of Virginia. E-mail: sriram.sankar@microsoft.com
Sudhanva Gurumurthi is with AMD Research, Advanced Micro Devices Inc., and the University of Virginia.
E-mail: Sudhanva.Gurumurthi@amd.com

Manuscript submitted: 11-May-2013. Manuscript accepted: 15-Jun-2013. Final manuscript received: 5-Jul-2013

.

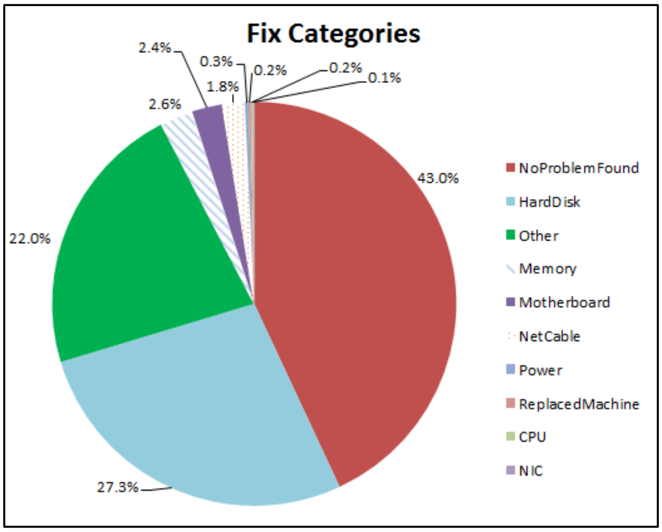

Figure 1 Percentage of Fixes for Failures in Datacenters

Figure 1 shows that 43% of the total fixes performed in this cluster belong to the category *No Problem Found*, which assigns these failures to a category describing machines that reported a problem even though no actual failures occured. In most of these cases, hard-power recycling (as contrasted to an autonomous soft power recycle) fixes some portion of the problem. In other cases, physically reseating the hard disk drives or the memory DIMMs resolves the issue due to transient configuration issues with these components. Figure 1 also shows other categories of actual failures in a datacenter; the next biggest category is due to hard disk drives. Hard disk drive failures as also documented in other previous works is one of the most common cases of datacenter server failures [6]. A significant portion of the fixes include *Other*, which includes firmware fixes, fan cages, SATA cables, and other components in servers that are not listed. The other component failures are minimal compared to these big contributors, and the most common error cause involves no actual hardware issue.

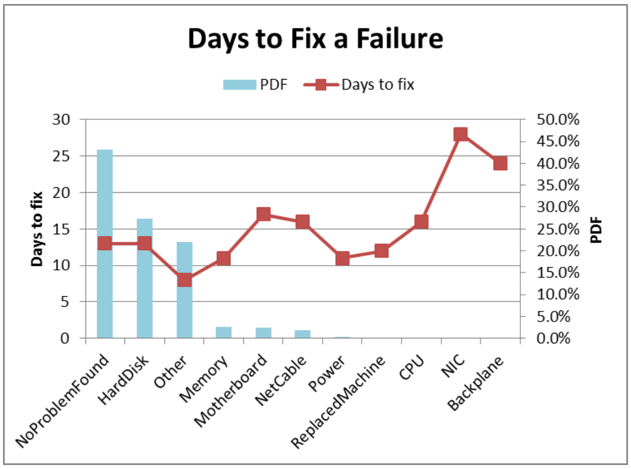

Figure 2: Average days for fixing a failure

2.2 Downtime due to Soft Failures

To understand downtime statistics, it is crucial to understand the service model in large datacenters. Most datacenters that host online and cloud computing services are highly redundant. It is rare that an application has a mission-critical dependency on one server or has just a single instance application model (in which one instance failure means that the single instance and data associated with it has to be recovered). Hence, fixing failures in a datacenter occurs by a batch processing model, where several failures are fixed at a time. This saves the amount of time a technician has to visit a datacenter, and decreases the load when scaling up to several thousands of servers.

Figure 2 shows the average number of days it takes to fix failures belonging to different types. This measure could be taken as a relative metric, to see how *No Problem Found (NPF)* fixes compare to failures that are fixed by hard disk replacement. As can be seen from the figure, *No Problem Found* failures take the longest time to fix: there was a failure, however the technician is not able to identify the failure correctly, and thus returning it to healthy state. In most other cases, the average time to fix the failure is significantly shorter than *No Problem Found* case, because it is easy to identify the component where the failure occurs. NIC and Backplane are exceptions, because those repairs involve taking out the entire enclosure.

2.3 Recurrent Soft Failure Probability

The probability of a machine having another soft failure, after encountering a specified number of soft failures is plotted in Figure 3. As can be observed from the figure, there is no significant increase in the probability of another failure, until the third occurrence following the first occurence of a soft failure, however, after that, there is a sharp increase in probability of soft failures. In fact after the fifth consecutive soft failure, one out of every two machines encounters another soft failure. This is a very useful statistic to have, since we can formulate machine repair strategies using this probability.

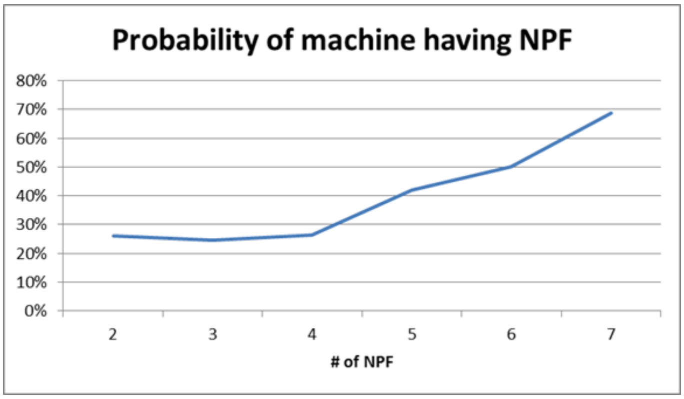

Figure 3: Probability of machines to have recurrent No Problem Found failures

2.4 Next Fix After a Soft Failure

In this section, we identify the likelihood of the next fix after a soft failure fix categorized as *No Problem Found.* Essentially, we want to identify whether a subsequent fix on the same machine yielded a different fix, or we remained unable to diagnose the issue, and classified it as a soft failure. Figure 4 shows the immediate next fix following a *No Problem Found* fix. The most immediate fix after an initial *No Problem Found* ends up being another *No Problem Found*, which is the most likely cause of the recurrent behavior we saw in Figure 3.

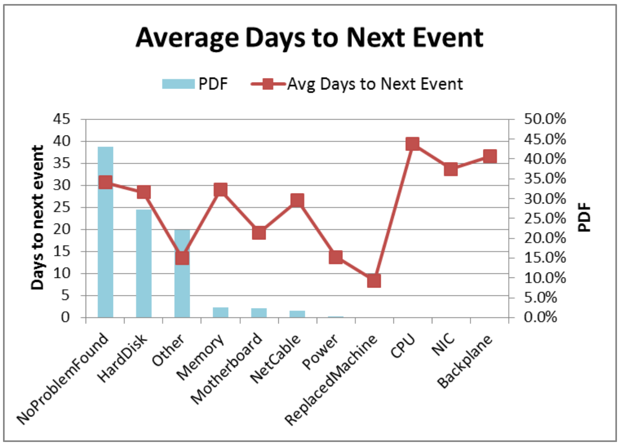

Figure 4: Actual Fixes Compared to NPF Fixes

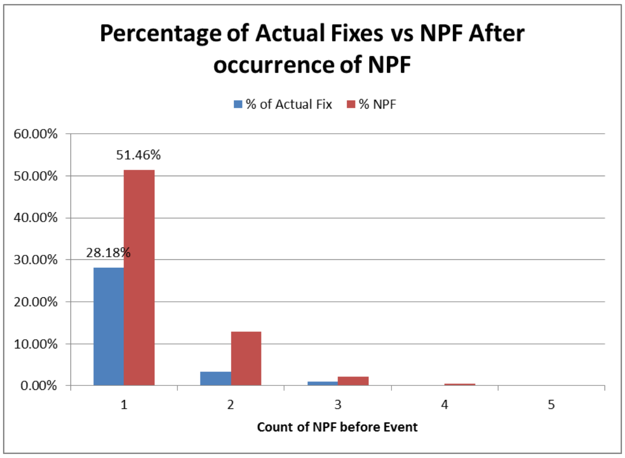

Figure 5: Subsequent fix type after a Soft Failure

In addition to plotting the next fix, after a *No Problem Found* event on the left vertical axis, we plot the average days to the next event on the right vertical axis. This is a measure of the efficiency of the fix: if the fix were made correctly, then it would take a lot longer for a subsequent failure to happen. Soft failures seem to happen at around the same frequency as hard disk drives, and do not show any particularly interesting inter-arrival behavior.

2.5 Probability of Soft Failures Leading to an Actual Failure

It is natural to investigate the correlation between the number of soft failures in a machine and actual failures experienced by the machines. In order to evaluate this question, we identify the count of *No Problem Found* failures beore an actual hardware fault in a machine. We see in Figure 5 that the probability of the first event being a soft failure is much higher than the probability of having an actual failure. In all machines that had only two failures, the first failure was a soft failure 51% of the time, whereas the first failure was an actual failure only 28% of the time.

3 Possible Architectural Approaches to Counter Soft Failures

As we observed from the data presented in Section 2, Soft failures in a datacenter are very hard to root-cause and even harder to fix. The typical fix-only approach to this problem is an iterative approach, where technicians fix a potential failing part, and keep fixing other parts till the issue “goes away”. We argue that a system-wide approach to solving this issue is needed. In this paper, we present three primary approaches at solving this issue in a large datacenter, 1) Process modifications based on our characterization 2) Architectural modifications to hardware components in the datacenter and 3) Software approaches that can identify and mask the fault. Because this is an open problem, we also discuss several possible avenues of research at the end of this section.

3.1 Process Modifications

From our earlier section, we observed some important characteristics of ‘Soft failures’. To summarize, soft failures were likely to occur in one out of every two machines that had five consecutive failures (Figure 3), and these failures occur at inter-arrival rates similar to those of hard disk drives (Figure 4). An easy process modification is to remove the machine with five consecutive soft failures from the datacenter. Note that this replacement would be done specific to each environment based on cost models. Since we are dealing with a very high rate of consecutive failures, this process modification can be more efficient than leaving the machine in production. Another process modification is to increase technician training that can prioritize soft failure fixes, since these constitute the most frequent errors seen in the datacenter.

3.2 Hardware Modifications to Handle Soft Failures

We often find that soft failures are caused by incorrect racking and incorrect cabling of components. It is hard to identify the component that could have failed; if no components actually failed. The failures that get tagged as *No Problem Found* can be partly addressed by removing cabling from inside the server chassis. A chassis could be designed as a collection of servers inside an enclosure with a shared backplane, and plug in place hard disk drives, NIC and PCIe cards. This would help eliminate the need for a lot of cables that are present in a server enclosure, including SATA cables that connect hard disk drives, network cables that connect from a switch to the server, etc. Even in this case, there could be connector end-point issues that might cause *No Problem Found* errors. The growth of System-on-chip designs and introduction of non-volatile memory designs [7] in this paradigm might remove the need for external connectors in the future.

Another hardware modification that could be employed to handle soft failures at scale is to identify opportunities for provisioning components in alternate topologies. For instance, datacenters currently deploy centralized UPS systems as transitory power source, while switching over to a generator backup. New topologies could deploy distributed smaller sized UPS systems [1], in order to help minimize the impact of transient failures of large components. This architectural change might increase maintenance costs, and also service costs. However it might increase the possibility of fault tolerance to both regular hardware failures and soft failures.

3.3 Software approaches for Soft failures

Soft failures are like any other failure in a datacenter, and software and system redundancy approaches that are employed in cloud services can mask the fault from being visible at the service level. There are sub-system level error detection capabilities in processors, memory controllers and SMART counters in hard disk drives that can detect potential failure modes. However, there are very few diagnostic tests targeted to detect potential soft failures at the system level. Hence it is imperative to build a decision tree, and store the history of server behavior and repair fixes, so we can identify possible server behavior that can denote a soft fault. For instance, if a server does not fail any of the diagnostic tests, but fails to boot or image, then there might be a possible soft failure. Also, if the same server is encountering repeated transient errors, but passes all diagnostic tests, then it could be a potential candidate that is encountering soft failures. Such behavioral identifications could be made by a management software layer similar to Microsoft's Autopilot [3] or Google's System Health Infrastructure [6].

3.4 Discussion

In a large scale datacenter, several other factors might contribute to soft failures, including transient faults in server, memory, PCIe and other silicon [5], system level constructs like position of servers in rack, rack design, temperature relationships, humidity, and rack vibrations. Extensive large scale data collection and analysis needs to be conducted to identify possible relationships, which is part of our future work on this subject.

4 Related Work

Most of the previous studies in this space only look at actual hardware replacements to compute failure rates. For instance, a study on disk failures in large scale infrastructure at Google looks specifically only at failed disk drives [6], and a recent study at Microsoft also looks only at hardware replacements [8]. Other large scale studies that look at hardware replacements include work on memory failures at Google [9], and failures at HPC clusters [10]. Hardware reliability for datacenters was studied in [11] with emphasis on failure trends and the authors note successive failure probabilities, but do not investigate transient failures in depth. Transient failures are not an unknown issue at microprocessor scale [5]. However, at a large scale datacenter, such a phenomenon is not typically being looked at as a systemic issue in large scale datacenters.

5 Conclusion

This work highlights a growing problem that is not categorized under the traditional challenges that datacenters face. The research community has yet to focus on *soft failures* as a challenge that could be solved using both hardware and software architectural approaches. In this paper, we present data that shows that soft failures are indeed something that ought to be addressed immediately, and we also identify early approaches at solving this growing issue in large scale datacenters.

References

[1] S. Govindan, D. Wang, L. Chen, A. Sivasubramaniam, and Bhuvan Urgaonkar. "Towards realizing a low cost and highly available datacenter power infrastructure." In Proceedings of the 4th Workshop on Power-Aware Computing and Systems, p. 7. ACM, 2011.

[2] U. Hoelzle , L. A. Barroso, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Morgan and Claypool Publishers, 2009

[3] M. Isard, Autopilot: Automatic Data Center Management, in Operating Systems Review, vol. 41, no. 2, pp. 60-67, April 2007

[4] C. Kozyrakis, A. Kansal, S. Sankar, and K. Vaid. "Server engineering insights for large-scale online services." Micro, IEEE 30, no. 4 (2010): 8-19.

[5] S. Mukherjee. Architecture design for soft errors. Morgan Kaufmann, 2008.

[6] E. Pinheiro, W. D. Weber, and L. A. Barroso. Failure trends in a large disk drive population. In Proc. of the FAST '07 Conference on File and Storage Technologies, 2007

[7] P. Ranganathan and J. Chang. "(Re) Designing Data-Centric Data Centers." Micro, IEEE 32.1 (2012): 66-70.

[8] S. Sankar, M. Shaw and K. Vaid. “"Impact of Temperature on Hard Disk Drive Reliability in Large Datacenters", IEEE/ IFIP International Conference on Dependable Systems and Networks, Hong Kong, June 2011.

[9] B. Schroeder , E. Pinheiro , W. Weber, “DRAM errors in the wild: a large-scale field study”, Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems, June 15-19, 2009, Seattle, WA, USA

[10] B. Schroeder, G. Gibson. "A large scale study of failures in high-performance-computing systems." International Symposium on Dependable Systems and Networks (DSN 2006).

[11] K. V. Vishwanath and N. Nagappan. "Characterizing cloud computing hardware reliability." In Proceedings of the 1st ACM symposium on Cloud computing, pp. 193-204. ACM, 2010.