Towards a systematic analysis of cluster computing log data: the case of IBM BlueGene/Q

Alina Sirbu, Ozalp Babaoglu
Department of Computer Science and Engineering, University of Bologna
Mura Anteo Zamboni 7, 40126 Bologna, Italy
Email: alina.sirbu@unibo.it, ozalp.babaoglu@unibo.it

Abstract—Management of large computing infrastructures is increasingly challenging and costly. While log data is often the basis for models and predictions to aid both automatic and human interventions, current use is typically limited to restricted datasets, and as such, is not applicable to entire systems. Data integration is necessary to build models that are accurate enough to be useful in real systems. In this paper we analyze four datasets reporting on power consumption, temperature, workload and hardware/software events for an IBM BlueGene/Q system. We show that the system handles a very rich parallel workload, with low correlation among components in terms of temperature and power, but higher correlation in terms of events. Power and temperature correlate positively very well, while events display negative relations with load and power consumption. The aim of the study is a systematic characterization of the system and identification of correlation patterns that can aid further modeling and prediction efforts.

I. INTRODUCTION

Recent technological developments have resulted in large computing infrastructures becoming critical for many applications, not only for specialized tasks, but also for governing many aspects of our daily lives. This in turn renders their continued and correct functioning of utmost importance. At the same time, increases in the size and complexity of such systems has significantly hampered their manageability. Although progress has been made in modeling and predicting the behavior of large computing infrastructures, results have not reached a level of accuracy necessary for wide-spread industry adoption. This leaves the management of large computing systems a difficult activity that continues to rely heavily on human interventions [1]. One possible cause for this disconnect is the narrow scope of most analyses limiting their focus to particular subsystems rather than considering the system as a whole, which is a necessary condition for achieving high performance [2].

Large computing systems produce large amounts of data in the form of logs tracing resource consumption, errors and various events. These data can be put to use for building predictive models of system behavior in an effort to tackle the management challenges. With recent progress in Data Science and Big Data tools, this approach to system management is becoming increasingly promising [3]. It is also becoming increasingly feasible to integrate data from multiple sources describing different components of the whole system to obtain more realistic models towards more reliable predictions.

In this paper we present the results of study aimed at integrating data from a multitude of sources in an effort to understand the behavior and relations among the components for a large computing system. We have analyzed four different datasets obtained from different subsystems of a 10-rack IBM BlueGene/Q supercomputer [4]. These include power and temperature logs, workload traces and hardware/software events during an overlapping period of about 3 months. We fist present studies considering each of the datasets independently seeking significant features that describe system behavior and correlations among system components. Next we consider multiple datasets taken together looking for correlations among various metrics so as to identify further possible relations to be used in future modeling and prediction studies.

The contributions of this paper are twofold. First, we provide a characterization of a BlueGene/Q machine from thermal, power consumption, workload and event log perspectives, stressing the relations between different components of the system. We show that power consumption and thermal behavior are independent among components, yet events exhibit significant correlations, indicating possible propagation of errors. Secondly, an integrated analysis looking at the four datasets together reveals significant positive correlation between power consumption and temperature, yet a weaker negative correlation between hardware/software events and power consumption or workload.

The rest of the paper is organized as follows. In the next section we describe the data and the system which it comes from. We then proceed to the data analysis, both for individual datasets and correlated features, in Section III. Section IV discusses related work and Section V concludes the paper with some final remarks and ideas for future work.

II. DATA DESCRIPTION

The source of our data is Fermi [5], an IBM BlueGene/Q system run by CINECA, the non-profit consortium of 69 Italian universities and 3 Public Administration Institutions, operating the largest data center in Italy. The system has a total of 163,840 computing cores and handles HPC workload from the consortium’s academic partners. Fermi consists of 10 racks, each containing 2 mid-planes. Each mid-plane has 16 node-boards with 32 16-core nodes per node-board. Each mid-plane is powered by 18 so-called bulk power modules. System logging is based on standard BlueGene/Q tools [4]. One such tool is the Mid-plane Manager Control System which is in charge of environmental monitoring. This is where we extract power and temperature information for the system. A different tool is the Machine Controller, which controls
Table I: Four datasets that are analyzed

| Dataset | Time span (2014) | Time resolution | Component | Total records |
|---------|------------------|-----------------|-----------|---------------|
| Power   | 28 March - 25 July | 5 min           | Bulk Power Module | 9,655,298 |
| Temperature | 23 April - 25 July | 15 min         | Node-board | 2,648,331 |
| Workload | 1 May - 27 July | NA              | System    | 78,128 |
| RAS     | 23 April - 25 July | NA              | All       | 774,555 |

Table I summarizes the four datasets. These span similar time periods, with an overlap of about three months, and include information at different levels of the system obtained from sensors on different components.

Power logs were extracted from the bulk power monitor of Fermi, which reports input and output voltages and current levels for each bulk power module, with a 5-minute resolution. We used input values to compute power levels. By summing the power levels over the different components, we obtained time series of power consumption for individual mid-planes, racks and for the entire system. Power at the node-board level cannot be reliably computed from these data since 18 bulk power modules are used to power 16 node-boards (redundant system).

Temperature logs were extracted from the node-board monitor, which reports two temperatures for each node-board, with a 15-minute resolution. Using these values we were able to compute time series of average temperatures at node-board, mid-plane, rack and system levels.

Workload data was reported by CINECA in a more aggregated form: a list of jobs with the date of completion, running time, number of cores and queue time. This allowed us to compute the CPU time used by every job, as well as time series of total daily CPU times, number of cores and queue times. Since only the date of job completions (and not the exact times) are available in the data, these totals are approximate, yet they give a very good indication of the daily load at system level. No load information at other levels (node-board, mid-plane, rack) was available at this time.

RAS logs were the fourth dataset available. They consist of hardware and software events from all system components. Events are labeled based on severity: our dataset includes three event categories with 163,134 FATAL, 473,982 WARN and 137,438 INFO events. Also, the exact time and exact location of each event is included. Using these data, we computed the distribution of inter-event times at system level, for each severity category. We also obtained time series monitoring the number of events in each category at 24-hour and 5-minute resolutions. These were obtained at node-board, mid-plane, rack and system levels.

III. DATA ANALYSIS

While each of the above datasets may provide very useful insight into the functioning of the Fermi system, combining them for an integrated analysis has even greater potential. Consequently, in this section we first perform an analysis of each dataset individually, identifying their features and comparing them. This is followed by an integrated analysis to study how different metrics from different subsystems correlate.
A. Power logs

The system specifications for an IBM Blue/GeneQ machine declare the typical power consumption to be around 65kW, with a maximum of 100kW per rack \cite{6}. However, real consumption varies depending on system load and state of components, e.g., how many nodes are up. Figure 1 displays the distribution of power consumption sampled at 5-minute intervals. Three different distributions are shown: total power at system level followed by power consumption by each rack and mid-plane. All distributions are centered around the official declared average values: 650kW at the system level consisting of 10 racks, 65kW at the rack level and 32.5kW at the mid-plane level, confirming the specifications. As we move from higher levels to lower ones, the distribution becomes broader. While total consumption is mostly between 50kW and 70kW per rack, with a bell-shaped distribution, when looking at individual racks additional peaks emerge, with some racks showing power consumptions up to 90kW for some time windows, but also frequent values under 50kW. The additional peaks become even more evident at the mid-plane level. This difference in distributions shows that power consumption is very heterogeneous. This needs to be taken into account when modeling power consumption, and indications are that while predicting overall system power consumption might be easier due to a large stability in time, finer grained predictions at rack or mid-plane levels might produce more accurate results.

Since power consumption at rack and mid-plane levels appears to be more heterogeneous, it is interesting to see if power values are similar in time when comparing different components (racks or mid-planes). Considering that power consumption depends on system load, load balancing algorithms should make it so that the different parts of the system work equally. So power consumption should be correlated among components of the system (racks or mid-planes), unless load is too low. Figure 2 shows Pearson correlations of power consumption between rack pairs and between mid-plane pairs. It is clear that at both levels, correlations are in general very low. Only a few mid-plane pairs seem to have correlations above 0.5 which are only moderate. Given that in section III-C we will observe a high general load of the system, this may indicate that load is not very well balanced.

B. Temperature logs

Distribution of average temperatures has also been analyzed. Figures 3 and 4 show histograms of average tem-
temperatures sampled every 15 minutes. At the overall system level, again the distribution is bell-shaped and narrow with few exceptions, with one mode around 50°C. As we zoom in at different component levels, the distribution becomes again wider with additional peaks showing up again at very high and very low temperatures. Individual node-boards can reach up to 75°C, significantly larger than the system average. This shows again how at different levels the system appears to behave differently, with greater heterogeneity in time at the finer grain logs.

Looking again at correlations among different components of the same type (Figure 5), a pattern similar to power consumption is observed. With very few exceptions, temperatures among components are not correlated. In terms of thermal isolation, this is good news, since having one hot node-board does not mean all surrounding node-boards are hot as well. However since power consumption showed the same pattern, this can be considered an additional evidence that workload is not very well balanced in the system.

C. Workload logs

A first important question in terms of workload is what type of jobs are submitted to the system. For the 78,128 jobs that ran on the system during the logged period, Figure 6 displays the distribution of different job attributes: total CPU time, total running time and number of cores used. In terms of time requirements, jobs are very heterogeneous as evidenced by a long tailed CPU time distribution, with a few very heavy jobs and many short jobs present. Effective running times are
bimodal, with many short jobs and many long jobs (all running times under 24 hours), and slightly fewer medium-length jobs. The number of cores per job, on the other hand, is not as heterogeneous. Jobs require only eight different values for the core count, most using over 100 cores and up to 32768. So, in general, the jobs that are run are highly parallel.

The structure of the workload dataset allows for analysis of patterns in time only at system level and at 24-hour resolution. Figure 7 shows the total CPU time for jobs completed each day. Note that this does not represent the exact system load for that day. The data only contains the date of completion for each job, not the exact time, so it is not possible to check exactly how many hours each job ran in a day. Jobs completed on one day could have been started the previous day. However the total time used by jobs completed in the same day still gives an indication of system load. A roughly bell-shaped distribution is obtained again, with a mean around 13 billion seconds (around 400 years), indicating very high levels of system load for Fermi.

D. RAS logs

For RAS events, we first look at inter-event times at the system level. Figure 8 shows the distribution of inter-event times, in seconds for the three event categories. The plot is a rank plot: the time intervals between events were sorted in
descending order, and then their values plotted on the vertical axis. The inter-event times do not appear to follow a known distribution. FATAL events show a few very large intervals and then many very small intervals. This indicates a pattern with spikes of many events in short periods of time but with large breaks between them. INFO and WARN events do not show very large inter-event times, and a smaller fraction of very short intervals, indicating they are more evenly spread across time.

Further analysis of the time distribution of events is performed by counting the number of events in each category at different time windows. We use two window sizes, to explore different time resolutions: 24-hour and 5-minute windows. Figure 9 displays the 24-hour resolution time series and their relative correlations. It is clear how WARN and INFO events are much more common, with most days showing some instances. While FATAL events come in spikes in a few of the monitored days. At the same time, it appears that daily INFO and WARN events are highly correlated, and so are WARN and FATAL events. However INFO and FATAL events seem to appear together less frequently. This could mean that INFO and WARN events could be used for predicting FATAL events at this time resolution. Similar data is shown for the finer grained time resolution (5-minute windows) in Figure 10. Here we display only histograms for clarity, since there are significantly more values in the time series. The histograms again show that for FATAL events, there are fewer time windows with low number of events compared to the other two categories, and a larger number of windows with many FATAL events. Correlations between event categories decrease at this resolution level, indicating that there might be a time shift between events that...
Fig. 12: Correlation between datasets at two resolutions (24 hours or 5 minutes) for overall system measures.

does not affect 24-hour correlations but decreases them at 5-minute resolution.

A different question is whether events correlate not only across categories, but also within the same category and across different components. Figure 11 shows the distribution of correlations between racks, mid-planes and node-boards for FATAL events. Similar results were obtained for the other two categories. Unlike power and temperature, FATAL events have higher correlation across components, with a significant number of component pair correlations over 0.5. This indicates that failures may propagate across components, while temperature and power consumption are more independent.

E. The big picture

The analysis of individual datasets has shed some light into the functioning of the Fermi system, and where relations between components might exist. Here we look at the four datasets together to uncover further relations between the different components and logs.

A first analysis is at the system level, looking at the different measures for the overall system rather than at individual

racks, mid-planes or node-boards. Figure 12 shows correlations between the different datasets for two time windows: 24 hours and 5 minutes. Workload data can only be analyzed on a daily basis, so it is not included at the 5-minute resolution.

As we have seen earlier, RAS events correlate very well
at 24-hour resolution and very little at 5-minute resolution. There also seems to be a strong relation between temperature and power, at both resolutions. This is a well known correlation for computer systems, and has been observed in other systems as well [2].

Workload measures do not appear to correlate among themselves, nor with power consumption, at 24-hour resolution. Better correlations could be obtained by zooming in at rack, mid-plane or node-board levels, however the structure of our workload dataset does not allow for this analysis. This prompts for changes in workload log structure in Fermi. Low correlation is present between system temperature and CPU time. Again, this should increase at other component levels. Some negative relation appears to exist between RAS events and CPU time, which is somewhat counterintuitive: one would expect more events to appear when the system works harder. However, this could be due to the coarse-grained time resolution. When looking at daily patterns, it is quite possible for large numbers of RAS events to have resulted in system failure, which in turn resulted in fewer completed jobs. This could explain the negative relation.

A negative correlation also appears between power/temperature and RAS events, again rather counterintuitively. This means that when there are more events, the system is working and consuming less, which can be due to complete shutdown of components after frequent RAS events. So, when trying to predict power consumption or FATAL events, one needs to take into consideration the negative dependence. Similar to RAS events, when increasing resolution (5-minute windows) the correlations dampen out. This indicates possible cross-correlations, i.e. correlations obtained between the two time series after shifting one of them by a specific time, due to a time shift in the events themselves.

The system level analysis might enhance some correlations, when behavior of components propagate or dampen them out by averaging over components. For this reason, we have also analyzed correlations at rack, mid-plane and node-board levels. Figure 13 shows correlations between available datasets at each level. Again, a strong relation between temperature and power consumption is visible. This suggests that in making predictions, using only one of the two features might suffice. In the case of BlueGene/Q, this is good news since power logs at node-board level are not available.

At rack level, RAS events are well correlated, but then correlations decrease as the level of detail increases (moving from rack to node-board level). This could be an artifact of rack-level events including events related to components other than node-boards. So, as components become more specific, the number of total events decreases, resulting in a very sparse dataset for which the Pearson correlation might be unsuitable.

IV. RELATED WORK

Log analysis for large computing infrastructures has been the focus of numerous studies. Recently, Google have released two workload traces for one of their clusters, triggering a flurry of analyses trying to characterize their system. General statistics have been provided [8], [9], [10], and more heterogeneity compared to grids has been observed [11]. Some modeling work has also appeared based on these data. In terms of resource usage, classes of jobs and their prevalence were used to characterize workloads and generate new ones [12], [13], or real usage patterns were replaced by the average utilization [14]. A simulator has also been built to reproduce the Google cluster load [15]. Important insight has been obtained from these data, however they only look at one part of the system, i.e. the workload. In modeling and prediction studies, it is important to integrate data from different components and sources. Other traces have also been analyzed in the past [16], [17], [18], and tools for analysis developed [19], but again concentrating on one type of data only, such as workload or failure traces. Here we integrate various datasets for a systematic analysis.

For IBM BlueGene, some previous work on failure prediction using RAS logs exists. In [20] failure prediction in a BlueGene/Q cluster is attempted. Their aim is to combine low-level hardware failure with high level kernel logs, however they stop at classification using RAS logs only. They analyze several classification tools (SVM, customized KNN, feature selection, rule based models) and obtain best performance with the KNN classifier. Another comparison of different classification tools for failure prediction in an IBM BlueGene/L cluster is [21]. They analyze SVMs, neural networks and rule based classifiers and also compute feature significance. They try to predict whether different event categories will appear in a time window based on previous windows. These predictive studies look only at RAS events, while integrating further data related to power consumption, temperature and workload could improve prediction accuracy significantly. Here we provide the first step towards such an analysis.

Most previous methods concentrate on workload or failure events, however recent work also includes thermal analysis of systems. In [22] multi-component chiller systems used in data centers are modeled. They build a Dynamic Bayesian Network model of the system using time series data from different sensors. Their aim it to control utilization levels of such equipment in order to optimize power consumption and lifetime. Again, this study does not integrate other types of data in the analysis. Some integration is performed by a very recent study from Google [3], who model Power Usage Effectiveness of their system using thermal information (temperatures, humidity, etc) together with overall system load, using an Artificial Neural Network model. Recently, a novel monitoring system for several criteria has been built [4]. This monitors Eurora, another HPC machine from CINECA, the same organization that provided our data. Their aim is thermal characterization of the system, however they do record several types of data besides power and chiller characteristics, including workload and machine status. These data could be, in principle, used for predictive and modeling studies in the future.

V. CONCLUSIONS

Given the need for a systematic integrative analysis of large computing infrastructures, this paper has presented an analysis of four data sets describing different subsystems of an IBM BlueGene/Q machine. Temperature, power consumption, workload and RAS logs were studied independently to characterize the system systematically and then together to
identify features of the data that might help in future integrated predictive or modeling studies.

The power consumption data showed that subcomponents of the system do not have correlated power consumptions, suggesting that load balancing might need attention for this system. Workload analysis showed high levels of load of the system, on average, with highly parallel jobs. RAS events of different severity showed correlation when looking at hourly patterns, but not when time resolution was increased to 5-minute windows. This suggests that there might be cross-correlations in the data that should be investigated further for modeling and prediction purposes. Across components, correlations were higher than in the case of power consumption, indicating that failures tend to propagate between components.

Not surprisingly, integrated data analysis showed that power consumption and system temperatures correlate very well, regardless of the space or time resolution. This indicates that, for the future, it might be useful to select only one of the two features for data analysis, in order to decrease dimensionality, which is good news also because power logs do not reach the same spatial resolution as temperature logs. Unexpectedly, when looking at daily levels of workload and power consumption versus RAS events, negative correlations were found, which can be explained by components turning off when errors appear and thus resulting in lower power consumption and less work completed. This has to be taken into account in predictive studies, where one would expect more events to be logged when the system is busier.

This study will continue with data external to the computing infrastructure, such as the water cooler and indoor air conditioning, together with data external to the data center, such as weather and seismic activity. RAS events will be further investigated to identify which of them indicate full shutdown of components, in order to verify our hypothesis for the negative correlations. Also, a predictive study will be undertaken, to use the correlations we already obtained to predict component failure. For this objective, cross-correlations will be also investigated.

ACKNOWLEDGMENTS

We are grateful to the HPC team at CINECA for sharing with us the log data related to the Fermi system and for helpful discussions.

REFERENCES

[1] L. Abraham, J. Allen, and O. Barykin, “Scuba: diving into data at facebook,” Proceedings of the VLDB Endowment, vol. 6, no. 11, pp. 1057–1067, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2536233

[2] F. Salfner, M. Lenk, and M. Malek, “A survey of online failure prediction methods,” ACM Computing Surveys (CSUR), vol. 42, no. 3, pp. 1–68, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1670680

[3] J. Gao, “Machine learning applications for data center optimisation,” Google White Paper, 2014.

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

[5] F. Falciano and E. Rossi, “Fermi: the most powerful computational resource for italian scientists,” EMBnet journal, vol. 18, no. A, pp. p–62, 2012.

[6] J. Milano, P. Lembke et al., IBM System Blue Gene Solution: Blue Gene/Q Hardware Overview and Installation Planning. IBM Redbooks, 2013.

[7] A. Bartolini, M. Cacciari, C. Cavazzoni, G. Tecchiolli, and L. Benini, “Unveiling eurora-thermal and power characterization of the most energy-efficient supercomputer in the world,” in Proceedings of the conference on Design, Automation & Test in Europe. European Design and Automation Association, 2014, p. 277.

[8] Z. Liu and S. Cho, “Characterizing Machines and Workloads on a Google Cluster,” in 8th International Workshop on Scheduling and Resource Management for Parallel and Distributed Systems (SRMPDS), 2012.

[9] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, “Heterogeneity and Dynamicity of Clouds at Scale: Google Trace Analysis,” in ACM Symposium on Cloud Computing (SoCC), 2012.

[10] ——, “Towards understanding heterogeneous clouds at scale: Google trace analysis,” Carnegie Mellon University Technical Reports, vol. ISTC-CC-TR, no. 12-101, 2012.

[11] S. Di, D. Kondo, and W. Cirne, “Characterization and Comparison of Google Cloud Load versus Grids,” in International Conference on Cluster Computing (IEEE CLUSTER), 2012, pp. 230–238.

[12] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, “Towards Characterizing Cloud Backend Workloads : Insights from Google Compute Clusters,” Sigmetrics performance evaluation review, vol. 37, no. 4, pp. 34–41, 2010.

[13] G. Wang, A. R. Butt, H. Monti, and K. Gupta, “Towards Synthesizing Realistic Workload Traces for Studying the Hadoop Ecosystem,” in 19th IEEE Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), 2011, pp. 400–408.

[14] Q. Zhang, J. L. Hellerstein, and R. Boutaba, “Characterizing Task Usage Shapes in Google’s Compute Clusters,” in Proceedings of the 5th International Workshop on Large Scale Distributed Systems and Middleware, 2011.

[15] A. Balliu, D. Olivetti, O. Babaoglu, M. Marzolla, and A. Sirbu, “Bidal: Big data analyzer for cluster traces,” in Informatika (BigSys workshop), vol. 232. Gi-Edition Lecture Notes in Informatics, 2014, pp. 1781–1795.

[16] S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, “An Analysis of Traces from a Production MapReduce Cluster,” in 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, no. December, 2010.

[17] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz, “The Case for Evaluating MapReduce Performance Using Workload Suites,” 2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems, pp. 390–399, Jul. 2011. [Online]. Available: http://ieeexplore.ieee.org/tpdocs/epic03/wrapper.htm?arnumber=6005383

[18] Y. Chen, S. Alsphaugh, and R. H. Katz, “Design Insights for MapReduce from Diverse Production Workloads,” Technical Report, University of California Berkeley, vol. UCB/EEECS-2, 2012.

[19] B. Javadi, D. Kondo, A. Isoup, and D. Epema, “The Failure Trace Archive: Enabling the comparison of failure measurements and models of distributed systems,” Journal of Parallel and Distributed Computing, vol. 73, no. 8, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0743731513000634

[20] R. Dudko, A. Sharma, and J. Tedesco, “Effective Failure Prediction in Hadoop Clusters,” University of Idaho White Paper, pp. 1–8, 2012.

[21] Y. Liang, Y. Zhang, H. Xiong, and R. Sahoo, “Failure Prediction in IBM BlueGene/L Event Logs,” Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 583–588, Oct. 2007. [Online]. Available: http://ieeexplore.ieee.org/tpdocs/epic03/wrapper.htm?arnumber=4470294

[22] D. Patnaik, M. Marwah, R. K. Sharma, and N. Ramakrishnan, “Data mining for modeling chiller systems in data centers,” Advances in Intelligent Data Analysis IX, pp. 125–136, 2010. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-13062-5_13