Workload Characterization and Evolutionary Analyses of Tianhe-1A Supercomputer

Jinghua Feng[1,2] , Guangming Liu[1,2]
Jian Zhang[2], Zhiwei Zhang[2] , Jie Yu[1] and Zhaoning Zhang[1]

[1] College of Computer, National University of Defense Technology, Changsha, China
[2] National Supercomputer Center in Tianjin,Tianjin, China
{fengjh,liugm,zhangjian,zhangzw,yujie}@nscc-tj.gov.cn

Abstract. Currently, supercomputer systems face a variety of application challenges, including high-throughput, data-intensive, and stream-processing applications. At the same time, there is more challenge to improve user satisfaction at the supercomputers such as Tianhe-1A, Tianhe-2 and TaihuLight, because of the commercial service model. It is important to understand HPC workloads and their evolution to facilitate informed future research and improve user satisfaction.

In this paper, we present a methodology to characterize workloads on the commercial supercomputer (users need to pay), at a particular period and its evolution over time. We apply this method to the workloads of Tianhe-1A at the National Supercomputer Center in Tianjin. This paper presents the concept of quota-constrained waiting time for the first time, which has significance for optimizing scheduling and enhancing user satisfaction on the commercial supercomputer.

Keywords: HPC, workload, quota-constrained, scheduling.

1 Introduction

High Performance Computing (HPC) is a mainstream for performing large-scale scientific computing [1, 2]. Currently, large scientific computations that include high-throughput, data-intensive jobs, and stream-processing are increasingly becoming more common in HPC centers. It is important to understand HPC workloads and the first step in understanding workload is to understand the evolution of workload on the current systems. Previous works have been on workloads on various grids [3] and cloud [4] systems. However, these studies were earlier and not the same as the current workloads. The research on Carvers and Hopper at the National Energy Research Institute (NERSC) [5], and Mira at the Argonne Leadership Computing Facility (ALCF) [6], none of these supercomputers are representative of commercial supercomputers.

In this paper, we first give the details about the methodology for characterizing workloads, including the process for submitting jobs on the commercial supercomput-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44

er, system description, data source, definition and calculation of various variables, especially the quota-constrained waiting time.

Because there are more than 70% jobs are single node jobs, and they only occupy about 10% of the CPU Hours. We divided the jobs into two kinds, single and multi.

In this paper , we provide an evolutionary analysis of the Tianhe-1A supercomputer. We study the trend of runtime, waiting time, core time from 2011 to 2017 about the two kinds of jobs. Especially for the waiting time, this paper analyzed the relationship between the quota-constrained waiting time and the waiting time, runtime, and job size, which is instructive for the future optimizing scheduling and enhancing user satisfaction on the commercial supercomputer.

2 Background and Related Work

HPC schedulers use the FCFS (First-Come, First-Served) [9] and backfilling [10] techniques to achieve the highest system utilization possible with a reasonable turnaround. On commercial supercomputer, FCFS is more often chosen for business fairness reasons.

Currently, there are some researches focus on workload characterizations, [14] presented the history of HPC system development and applications in China, HPC centers and facilities, major research institutions, but it's before 2010. [5] discover investigate .. the evolution trend of Hopper and Carver, [15] analyzed the characterization of the workload on google compute clusters using the k-means algorithm. [7, 8] analyzed the system features of three supercomputers (Hopper, Edison, and Carver). [11] analyzed the I/O features of 6 years of applications on three supercomputers, Intrepid, Mira, and Edison.

In fact, the characteristics of job scheduling and system workload have changed a lot on commercial supercomputer, which are the focuses of this paper.

3 Methodology

In this section, we present the system and workloads in focus for our investigation and elaborate on the key parameters studied.

3.1 Data Source

All workload analysis is performed on the job summary entries from the SLURM [13] workload manager logs. The data includes seven years and 10735864 jobs. The data after filtering and parsing is reduced to 3 GB. Because there are more than 70% jobs are single node jobs, and they only occupy about 10% of the CPU Hours. In the paper, we divided the jobs into two datasets, single-node jobs and multi-nodes jobs.

The data fields consist of Jobid, Submit time, Start time, End time, Alloccpus, State and so on. And we complement existing scientific workload characterizations work [12] by adding the quota-constrained waiting time. Especially, we used 2016-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44

2017 data to fully analyze the relationship between the quota-constrained waiting time and the waiting time, runtime, job size.

Table 1. Workload of Tianhe-1A from 2011 to 2017

No. of Jobs	2011	2012	2013	2014	2015	2016	2017	total
Single node	21205	561954	640614	899509	1610001	2488108	1769195	7990586
Multi nodes	86784	231600	326046	397861	472651	597243	633093	2745278
Sum	107989	793554	966660	1297370	2082652	3085351	2402288	10735864

3.2 Systems Description

Tianhe-1A is the world's top supercomputer in 2010 at the National Supercomputer Center in Tianjin. It has been in service since 2011 and has been in operation for more than seven years. It is a typical representative of commercial supercomputer.

Tianhe-1A supercomputer consists of 7168 computing nodes (12 cores, 24GB of memory per node) with a peak performance 4.7PFlops. The workload is composed of applications that belong to a wide range of scientific fields including Chemistry, Material Science, Climate Research, Astrophysics, Life Sciences, and Basic Science.

3.3 Analysis variables

When using the commercial supercomputer, users need to pay for the computing, and they usually use resources in accordance with the contract, the contract mainly limits the size of the total resources that users submit jobs, which is the quota constraint.

For example, if the user can submit k jobs, each job occupies the resource N_j, and the user's quota constraint is M, thus $\sum_{j=1}^{k} N_j \leq M$.

Fig.1 describes the process of submitting a job under the quota restraint environment. After the user submits the job, the workload manager first performs quota check, if including the job, the sum of user's resource has not exceeded the quota. Then, proceeds to the next step for resource check, On the contrary, the job needs to wait.

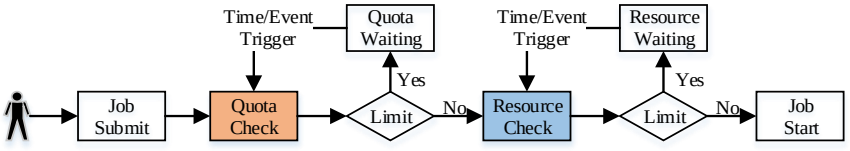


Fig. 1. The step from job submit to job start on a quota-constrained supercomputer

The job has Submit time (t_{sub}), Start time (t_{str}) and End time (t_{end}). The runtime of job j is the timespan between End time and Start time (t_{end}- t_{str}); the waiting time of job j (W_j) is the timespan between Start time and Submit time (t_{str}- t_{sub}); the response time

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44

of job j (R_j) is the timespan between End time and Submit time (t_{end}- t_{sub}). And the core time is defined as the total CPU time of the job.

If a user has a quota constraint (size of cpus), the waiting time consists of two parts: waiting time caused by quota-constrained and resource-constrained

$$W_j = Wq_j + Wr_j. \quad (1)$$

4 Trend Analysis

Fig.2 shows the percentages of job count and core time for each year, taking the total number of jobs and the core time respectively from 2011 to 2017. According to Fig. 2 we can see that from 2012, the number of single node jobs is much bigger than multi-node jobs (basically keeping the ratio of 7:3), but the actual core time multi-nodes jobs is much larger than single node jobs (basically keeping the ratio of 9: 1). We can also see that from 2011 to 2016, the number of jobs and the core time used are all increasing (the average utilization rate of resources in 2016 is over 85%), a slight decrease from 2017 in 2016.

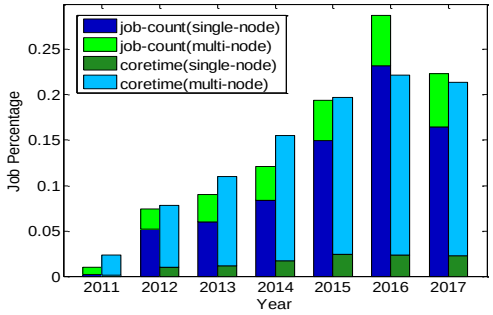


Fig. 2. The percentage of job count and core time from 2011 to 2017

4.1 Trend Analysis

Fig. 3 shows the box plot (the vertical axis is logarithmic) for job runtime from 2011 to 2017. The runtime time of single node jobs is significantly lower than that of multi-nodes jobs from 2013 to 2017, and median of single node jobs is 88,118,26,19,52 seconds. Because 2016 submitted the largest number of single node jobs (2488108), and the operation of the lower runtime, so the median in 2016 is the lowest.

Fig. 4 The core time shows a similar trend with runtime in both job types.

Fig. 5 shows the box plot (the vertical axis is logarithmic) for job waiting time from 2011 to 2017. There are several phenomena that deserve our attention. The first point is the higher waiting time value of single node jobs in 2012, median reached 2490 seconds, because some users submitted a large number of consecutive single node jobs, making these jobs increase the quota waiting time, so the overall waiting time increases; Second, from 2013 to 2016, the number of jobs and the core time are

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44

increasing. However, the waiting time value changes are not very obvious. Even the waiting time of the multi nodes jobs, the median value is decreasing. The median value gradually drops from 1156 seconds to 788 seconds. This point is different from [5] , the waiting time of Hopper supercomputer gradually increased, because of the number of jobs increasing. Third, in 2017, the waiting time for multi-node jobs has dropped because users can use more clustered systems, resulting in an increase of the inter-arrival of jobs.

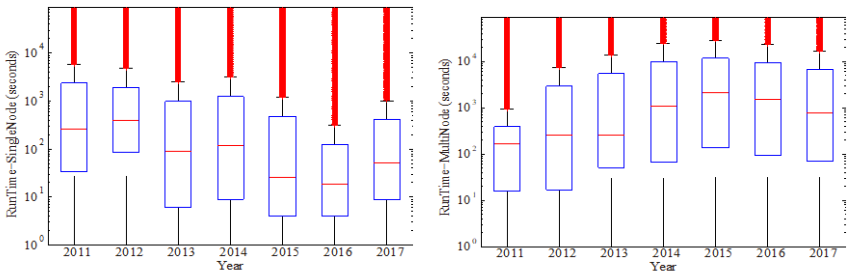


Fig. 3. The runtime of the two kinds of jobs from 2011 to 2017. Left (a), Right (b)

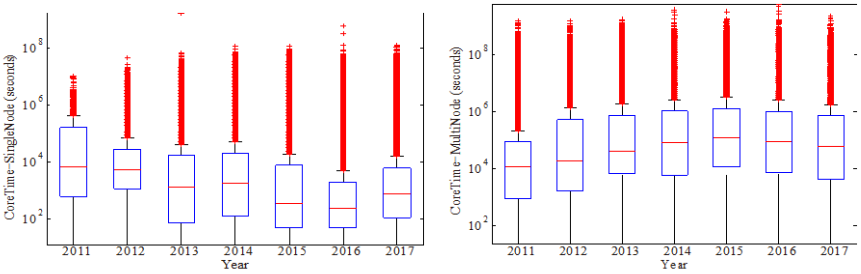
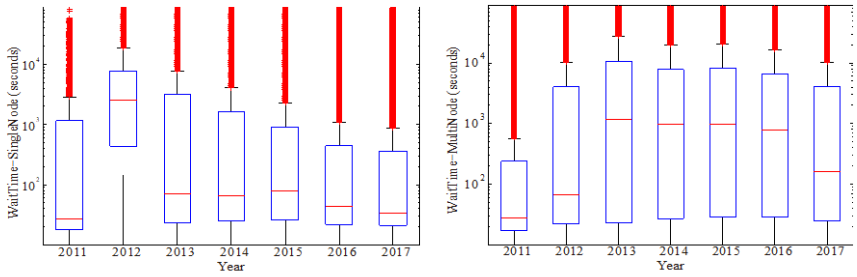


Fig. 4. The core time of the two kinds of jobs from 2011 to 2017. Left (a), Right (b)



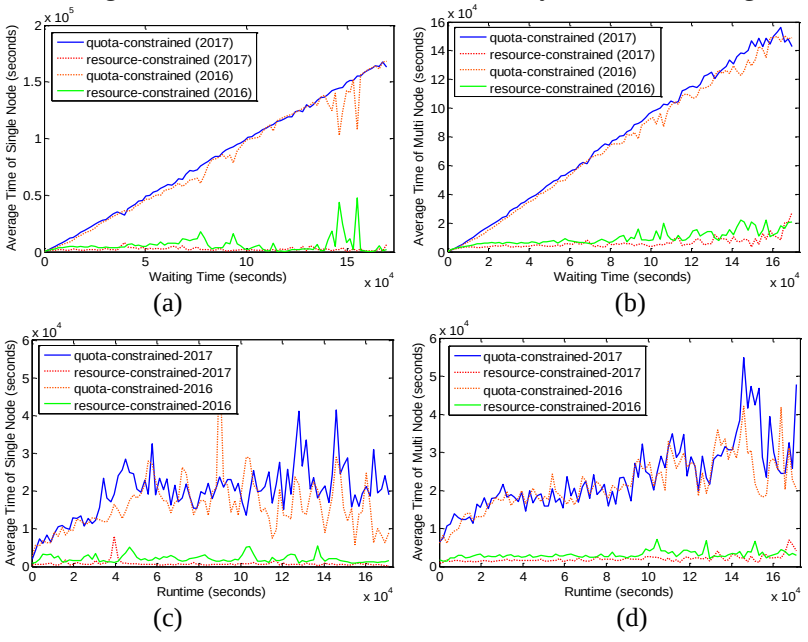
ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44

Fig. 5. The waiting time of two kinds of jobs from 2011 to 2017. Left (a), Right (b)

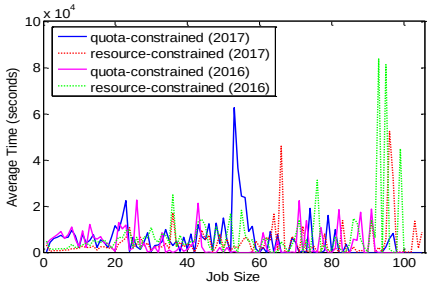
4.2 Analyze the Characteristics of Waiting Time Caused by Quota-constrained

Fig.5 (b) shows that the waiting time tends to decrease as the number of jobs increases. In order to figure out the reason, we used 2016-2017 data to fully analyze the relationship between the quota-constrained waiting time and the runtime, waiting time, job size. Fig.6 shows the changing trends of Average waiting times influenced waiting time, runtime, and job size, respectively.

Fig.6 (a), (b) the trend is similar. With the increase of waiting time, quota-constrained waiting time increases rapidly and near linearly, and takes up the main proportion of waiting time. However, the increase of resource-constrained waiting time is not obvious, and the proportion is smaller. Note that in commercial supercomputer, the waiting time is mainly due to the quota constraint, it is difficult to effectively reduce the waiting time of the job without changing the business restriction of the quota constraint. The quota-constrained waiting time is mainly affected by the job submission behavior of users. If users submit jobs frequently, many jobs wait for quota constraints, which increases the quota-constrained waiting time. The phenomenon in Fig. 5 (b) is also caused by a change in the user's submission behavior and a decrease in the quota-constrained waiting time. In the future, we will do the research to understanding the user behavior, in order to effectively reduce the waiting time.



ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44



(e) 110 kinds of job size from small to large

Fig. 6. Average of waiting time caused by quota-constrained and resource-constrained as a function of (a)(b) Waiting time, (c)(d) Runtime, (e) Job Size

Fig.6 (c) and (d) show a similar trend. As the runtime increases, the quota-constrained waiting time increases obviously and dominates.

Fig. 6 (e) shows that the two-part waiting time is not very affected by the job size when the job size is small, but when the job size is larger, the resource-constrained waiting time rapidly increases to the dominant factor.

5 Conclusions

In this paper, we present a methodology to characterize workloads on the commercial supercomputer, at a particular period and its evolution over time. We apply this methodology to the workloads of Tianhe-1A at the National Supercomputer Center in Tianjin. This paper presents the concept of quota-constrained waiting time for the first time, which has significance for optimizing scheduling and enhancing user satisfaction on the commercial supercomputer. In the future, we will do the research to understanding the user behavior, in order to effectively reduce the waiting time.

References

1. A. Geist et al. A survey of high-performance computing scaling challenges. The International Journal of High Performance Computing Applications 33(1), 104-113，(2017).
2. D. A. Reed and J. Dongarra Exascale computing and big data. Communications of the ACM, 58(7), (2015).
3. Iosup, Alexandru, et al. The Grid Workloads Archive. Future Generation Computer Systems 24(7),672-686 (2008)
4. Di, S., et al. Characterization and Comparison of Cloud versus Grid Workloads. IEEE International Conference on CLUSTER Computing (CLUSTER), (2012)
5. Gonzalo P. Rodrigo, et al. HPC System Lifetime Story: Workload Characterization and Evolutionary Analyses on NERSC Systems. International Symposium on High-Performance Parallel and Distributed Computing (HPDC), (2015).
6. S. Schlagkamp et al. Consecutive Job Submission Behavior at Mira Supercomputer. International Symposium on High-Performance Parallel and Distributed Computing (HPDC), (2016).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44

7. Gonzalo P. Rodrigo, et al. Towards Understanding Job Heterogeneity in HPC: A NERSC Case Study. IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, (2016).
8. Gonzalo P. Rodrigo, et al. Towards understanding HPC users and systems: A NERSC case study. Journal of Parallel & Distributed Computing 111,206-221 (2017).
9. D. G. Feitelson, et al. Parallel job scheduling, a status report. International conference on Job Scheduling Strategies for Parallel Processing, pp.1–16. Springer, (2005).
10. D. A. Lifka. The ANL/IBM SP Scheduling System. The Workshop on Job Scheduling Strategies for Parallel Processing,pp.295-303. Springer, (1995).
11. Huong Luu, et al. A Multiplatform Study of I/O Behavior on Petascale Supercomputers. International Symposium on High-Performance Parallel and Distributed Computing (HPDC), (2015)
12. D. Feitelson. Parallel workloads archive. http://www.cs.huji.ac.il/labs/parallel/workload, 2018/02/11.
13. AB Yoo, et al. SLURM: Simple Linux Utility for Resource Management. Lecture Notes in Computer Science 2862(2862), 44-60 (2002).
14. Ninghui Sun, et al. High-performance Computing in China: Research and Applications. International Journal of High Performance Computing Applications, pp.363-409, (2010).
15. A.K. Mishra, et al. Towards characterizing cloud backend workloads: insights from google compute clusters, Acm Sigmetrics Performance Evaluation Review, 37(4), 34-41 (2010).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_44