Liu RT, Chen ZN. A large-scale study of failures on petascale supercomputers. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(1): 24–41 Jan. 2018. DOI 10.1007/s11390-018-1806-7

A Large-Scale Study of Failures on Petascale Supercomputers

Rui-Tao Liu[1] and Zuo-Ning Chen[2], *Fellow, CCF*

[1] *State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214215, China*
[2] *National Research Center of Parallel Computer Engineering and Technology, Beijing 100190, China*

E-mail: liuruitao@wo.cn; chenzuoning@vip.163.com

Received July 29, 2017; revised December 7, 2017.

Abstract With the rapid development of supercomputers, the scale and complexity are ever increasing, and the reliability and resilience are faced with larger challenges. There are many important technologies in fault tolerance, such as proactive failure avoidance technologies based on fault prediction, reactive fault tolerance based on checkpoint, and scheduling technologies to improve reliability. Both qualitative and quantitative descriptions on characteristics of system faults are very critical for these technologies. This study analyzes the source of failures on two typical petascale supercomputers called Sunway BlueLight (based on multi-core CPUs) and Sunway TaihuLight (based on heterogeneous manycore CPUs). It uncovers some interesting fault characteristics and finds unknown correlation relationship among main components' faults. Finally the paper analyzes the failure time of the two supercomputers in various grains of resource and different time spans, and builds a uniform multi-dimensional failure time model for petascale supercomputers.

Keywords petascale supercomputer, fault characteristic, correlation relationship, multi-dimension, failure time model

1 Introduction

For the requirements of scientific research and engineering applications, the supercomputers are becoming more and more powerful, and their scales are more and more tremendous. The supercomputers ranging from 10 petaflops to 100 petaflops have hundreds of thousands of CPUs and more. With the increasing scale and complexity, the supercomputer is facing an unprecedented challenge in the reliability and availability. Although there have been various types of methods to promote components' reliability and availability, the system's scale and complexity still prevail over their improvement. Some researches show that MTBF (mean time between failures) of the future exascale supercomputers is $O(1 \text{ day})$[1-2], or even only half an hour[3]. The new generation resilience technology is urgently required to improve the system reliability and availability, and to decrease the cost of fault tolerance for the next generation supercomputers. Proactive failure avoidance technologies based on fault prediction and live migration, reactive fault tolerance based on adaptive checkpoint/restart, and scheduling technologies to improve the reliability of job running environment are promising resilience technologies for exascale systems. They all need comprehensively qualitative and quantitative description on characteristics of system faults.

Many researchers have studied fault characteristics on large-scale parallel computing systems. The researchers have analyzed the distribution of coarse types of faults (such as human, environment, network, software and hardware faults) and investigated the time between those types of faults on some high-performance computers (less than 1 petaflop) from LANL (Los Alamos National Laboratory)[4]. Some researchers gave temporal and spatial characteristics of faults on network, application and I/O, and analyzed correlation relationship between non-fatal and fatal faults[5-6]. The classification of faults is coarse and lack of precise and exact description on temporal distribution and cause analysis of failures.

Other researchers have made use of RAS (reliability, availability, and serviceability) and job logs to co-analyze failure characteristics on Bluegene/P[7-8]. They found that the time between failures of the whole ma-

Regular Paper

The work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB0200502.

chine meets the Weibull distribution. However their work has no further fine-grained analysis on the time between failures of each component and thus has limited effects on system fault tolerance.

INRIA found that different types of faults have different distribution models and gave a correlation analysis among faults each of components on Mercury high performance cluster[9]. The classification of faults is not so precise because of coarse data derived from the system's logs (six types of errors in total: SCSI, NFS, PBS, Not Connected, memory and CPU cache). There is a lack of fine-grained analysis of faults on each component in the study.

Recently, Oak Ridge National Laboratory has characterized and quantified different kinds of soft-errors on the Titan supercomputer's GPU nodes and uncovered characteristics and cause of GPU errors and the correlation between GPU errors[10]. However for the lack of logging time when faults occur, fault data was collected with the granularity of one batch job and normalized to 24 hours as the maximum precision.

In addition, DRAM errors[11] and disk drive errors[12] in a large cluster of commodity servers and failure characteristics in the cloud[13-15] were discussed. Fault prediction on HPC and cluster systems has also been studied based on event logs[16-20].

So far there is a lack of comprehensive and fine-grained analysis of fault characteristics on supercomputers whose scales are ranging from 1 petaflop to 100 petaflops. This is because it is hard to get comprehensive and complete data about detailed faults.

The development and construction of Sunway BlueLight and TaihuLight supercomputers is an important timeline for building the next generation exascale systems. Sunway BlueLight consists of 8 575 CPUs or 137 200 computing cores. Sunway TaihuLight consists of 40 960 CPUs or 10 649 600 heterogeneous computing cores[①]. Sunway BlueLight based on multi-core CPUs and Sunway TaihuLight based on heterogeneous many-core CPUs have different types of running modes because of their different architectures. Although many types of reactive and proactive fault tolerance technologies have been adopted for the two machines, there are still many difficulties to obtain high availability. The main challenges are as follows.

1) The checkpoint/restart on the system level is hard to be conducted because of the high cost of checkpoint on large-scale jobs.

2) There are so many various types of faults on the two machines. It is hard to identify the characteristics of those faults and give a precise description of failure time.

3) Although some proactive resilience method has been adopted such as living migration, the just time to do it is still hard to be determined.

The comparative analysis of the faults on the two machines helps the understanding of fault characteristics on large parallel computing systems and gives an important reference for further improving system design, realizing fault prediction, and promoting availability. For example, the optimum checkpoint interval can be calculated and used by the failure time model. The fault prediction is easier to conduct according to the model and the correlation relationship between faults.

In this work, we bring out a quantitative analysis method of describing various faults on multi-dimension, and lay a theoretical and engineering basis on technologies of reliability evaluation and fault prediction. The main contributions of this paper are as follows.

1) We present the way of fault collection and classification, and obtain the root cause of failures on Sunway series supercomputers.

2) We describe the characteristics of DRAM (dynamic random access memory) single bit errors (SBEs) and multiple bit errors (MBEs), and find the correlation relationship among the two types of DRAM errors and CPU nodes.

3) We analyze the time between failures of Sunway series supercomputers and build a uniform multi-dimensional failure time model.

We introduce the system organizations and the way of fault collection on Sunway series high performance computers, and comparatively analyze the root cause of failures on Sunway BlueLight and TaihuLight. We study the characteristics of DRAM SBEs and MBEs interested by fault tolerance technologies, and especially analyze warning correlation between the two types of DRAM errors. Our study gives a detailed analysis on the conditions DRAM faults occurring and a deep understanding of DRAM SBEs and MBEs.

At last, we give a large-scale study of time between failures with various grains of resources, such as CPU nodes, computing cards, and the whole mainframe in different time spans on Sunway series supercomputers. Then we build a uniform failure time model on multi-dimension for petascale supercomputers.

[①] http://www.top500.org, Dec. 2017.

The rest of the paper is organized as follows. Section 2 describes organizations, architectures, and fault collection and classification of Sunway series supercomputers. Section 3 analyzes the root cause of failures and identifies the main components impacting on system reliability. Section 4 describes the method of analyzing correlation relationship among various types of DRAM errors and CPU nodes. Section 5 analyzes the time between failures on Sunway series supercomputers and presents a uniform multi-dimensional failure time model for reliability improvement and failure prediction. Finally, we conclude the paper with our contribution and future work plan in Section 6.

2 Background and Methodology

In this section, we give an overview of organizations and architectures of Sunway BlueLight and TaihuLight. Further we introduce the collection of fault data online and the classification of faults.

2.1 Organizations and Architectures of Sunway BuleLight and TaihuLight Supercomputers

The Sunway BlueLight supercomputer is a large-scale parallel computing system with peak performance of 1.07 petaflops[2]. The machine is composed of computing and interconnecting, storage, maintaining and monitoring, power and cooling systems. The computing and interconnecting system is composed of CPU nodes and their network. Each cabinet has four super nodes, each of which has 32 computing cards and one network backplane. Each computing card is composed of eight CPU nodes and one integrated network card. One CPU node has DDR3-1066 memory chips and one high performance multi-core CPU called Sunway1600. The CPU integrates 16 general computing cores, multi-path memory controllers and one system interconnecting interface.

The Sunway TaihuLight supercomputer is a large-scale parallel computing system with peak performance of 125.43 petaflops. Each cabinet has four super nodes, each of which has 32 computing cards and one network backplane. There are eight CPU nodes and one integrated network card on each computing card. Each super node has 256 CPU nodes and their interconnecting. The super node supports computing-intensive, communication-intensive and I/O-intensive applications efficiently.

The main difference between Sunway TaihuLight and Sunway BlueLight is that the CPU node on TaihuLight has one heterogeneous many-core CPU called Sunway 26010, in which there are 260 general computing cores. Sunway 26010 uses a computing array and distributed shared memory combined architecture. The Sunway 26010 CPU has four groups of computing cores, each of which supports cache coherence. One computing and controlling core and one computing array are contained in each group of the computing cores. The computing array is composed of 64 computing cores, one array controller and one level-2 instruction cache. There are multipath DDR3 memory controllers and PCIE (peripheral component interconnect express) on the Sunway 26010 chip. The CPU node has a high-capacity DDR3-2133 memory chip and a PCI-E 3.0 interface, providing the capability of high bandwidth of data exchange.

The organization of the Sunway series supercomputers is shown in Fig.1.

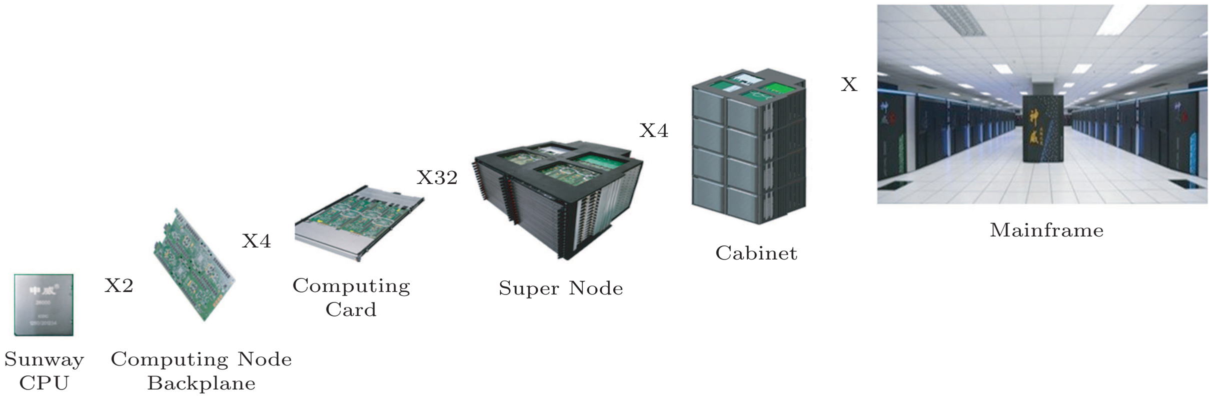


Fig.1. Organization of the Sunway series supercomputers.

[2] http://www.top500.org, Dec. 2017.

2.2 Framework of Faults Collection and Methodology of Classification

An infrastructure of monitoring and maintaining system is designed and deployed on the Sunway series supercomputers, which uses a distributed scalable architecture to monitor, diagnose, and maintain major objects of the whole machines including the mainframe.

The framework consists mostly of the baseboard management controller (BMC) and the processor management (PM), as shown in Fig.2. BMC monitors and maintains eight CPUs and their memory in each computing card. BMC is the fundamental component of the whole maintaining system, down to provide multiple parallel channels to connect with maintained components and up to connect with management network to constitute a scalable system. PM monitors and manages all the computing network cards, BMCs and Ethernet cards on each super node.

The diagnosing system includes three parts: backplane-level fault positioning programs, system-level fault positioning programs, and programs to save and store fault data. The backplane-level fault positioning programs collect and locate various types of faults on super nodes in real time. The system-level fault positioning programs reside on the central console to analyze and locate faults on the mainframe. The programs for saving faults identify fault levels and store the faults of the whole machine into database in a rapid and centralized way.

Various types of sensors covering the whole system are set to collect and analyze faults and their related running information on the machine. For example, there are fault sensors for CPU cores, controlling units, memory and HCA chips on BMCs and related temperature sensors. The data warehouse technologies are used to save and store the fault scenes and related information. The big data of fault analysis is built for the supercomputers.

To ensure catching the authentic faults and avoiding fault propagation, the software (SW) and hardware (HW) co-design method has been adopted. On the hardware side, passive and active isolation are taken in CPU nodes and other main components. For example, when the computing core detects its uncorrected errors or faults, it isolates itself actively to block all requests forwarded to the others. The response to the isolated computing core is discarded. Fault components can also be set to be isolated by the software. On the software side, when uncorrected errors or faults of components are detected, the software infrastructure makes related applications fault tolerant and isolates fault components immediately.

In general, each record of fault logs has a number of attributes. They are the objects where the fault occurs, the fault identifier, the fault name, the time when the fault occurs, the real-time temperature, the fault scene, etc. The objects where faults occur include CPU cores, CPU controlling units, memory, HCAs, and components for power, maintaining and cooling. The time when faults occur is precise to seconds. The real-time temperature logs the CPU's real-time temperature when faults occur. The fault scene saves the detailed information about faults. There is also correlation information among faults, computing resources, and running jobs in the system in order to analyze their correlation.

Filtering and compression of fault records are needed for fault analysis. Filtering removes the unqualified records online, for example, the records with null items. Compression eliminates redundant records of faults, for example, the duplicated fault records with the same fault type, time and location. The filtering and the compression of fault data are efficiently implemented to optimize data source for analysis by SQL retrieve technology of database.

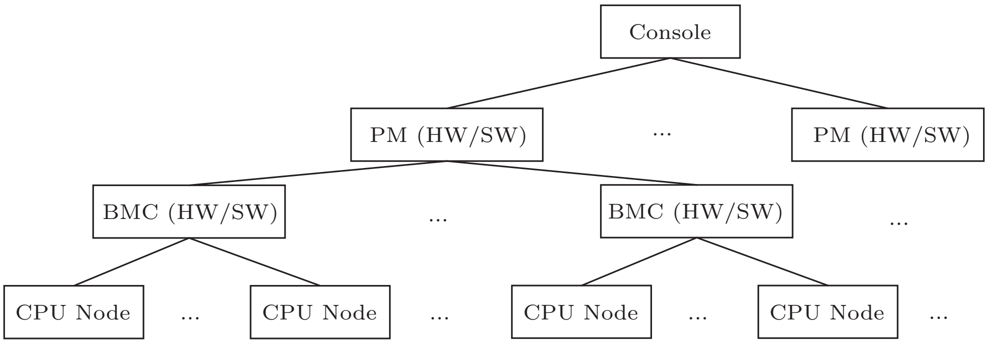


Fig.2. Scalable framework of fault collection.

In order to analyze conveniently, we classify various types of detailed faults in the machine into some "big" types, such as CPU (excluding DRAM) fault, DRAM fault, CPU node fault (including CPU and DRAM), computing card fault, interconnecting fault, maintaining fault, and power and cooling fault. Each "big" type is further classified into precise types. For example, DRAM fault is classified into DRAM single bit error (SBE) and DRAM multiple bit error (MBE), and CPU faults are classified into various computing components' errors and related controlling components' errors.

According to the fault level and the way to deal with, each type of faults is further classified into non-fatal or fatal faults. The non-fatal faults are the incorrect states which do not lead to system failure or can be corrected by hardware itself automatically. The fatal faults are the incorrect states that will lead to system failure and must be corrected by software.

3 Analyzing the Cause of Failures

In this section, the comparative analysis on faults of Sunway BlueLight and TaihuLight gives the cause of system failures and the proportion of various components' faults. It finds the key components mostly impacting on system reliability and gives an important guide for the design of system reliability and fault tolerance

3.1 Faults of Sunway BlueLight

The main components of faults on Sunway BlueLight are CPU, DRAM, interconnecting network, maintaining and diagnostic system, cooling system, and power system. After three years (2011∼2014) of study of fault data on Sunway BlueLight, the fault distribution is shown as follows.

1) Proportion of faults (including non-fatal and fatal faults) on main components, as shown in Fig.3(a). It can be seen that the memory accounts for the most proportion of faults over 90%. The second is the interconnecting network, the third is CPU.

2) Proportion of fatal faults on main components, as shown in Fig.3(b). It can be seen that the faults on CPU, DRAM and interconnecting system dominate almost 90% of the faults on the whole machine after filtering out non-fatal faults. The three types of components are the major components in the mainframe.

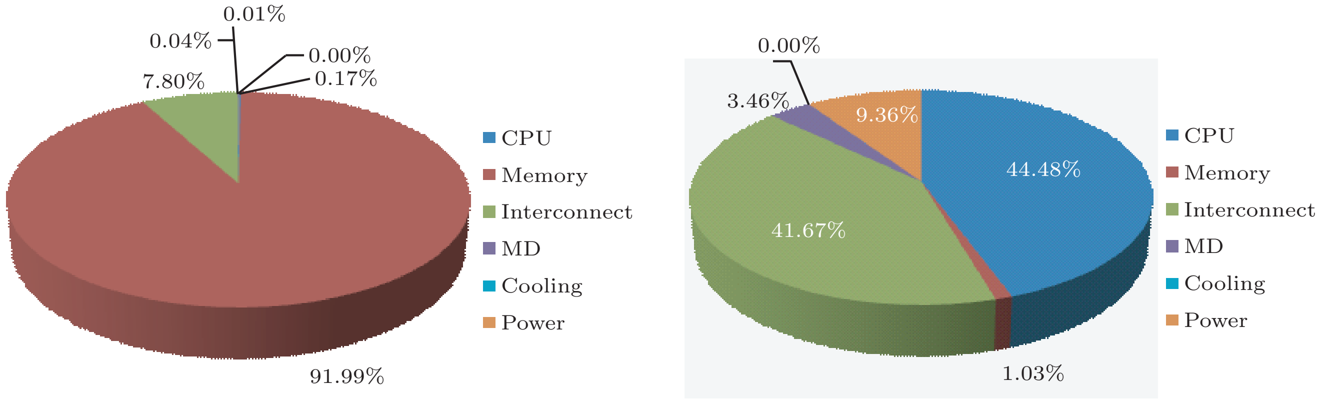
3.2 Faults of Sunway TaihuLight

The main components where faults occur on Sunway TaihuLight are similar to those of Sunway BlueLight. After about two years (2016∼2017) of study of fault data on Sunway TaihuLight, the fault distribution is shown as follows.

1) Proportion of faults (including non-fatal and fatal faults) on main components as shown in Fig.4(a) It can be seen that the memory accounts for almost half of all faults. The second is CPU (40%) and the third is the power system (9%). The faults of the interconnecting system are about 1%.

2) Proportion of fatal faults on main components as shown in Fig.4(b). It can be seen that the faults on CPU, DRAM and the interconnecting system dominate 90% of the faults on the whole machine after filtering out non-fatal faults.

The comparison of Fig.3 and Fig.4 shows that non-fatal faults of DRAM dominate a large proportion of



(a) (b)

Fig.3. Fault distribution of Sunway BlueLight. (a) Fault distribution (including non-fatal and fatal faults). (b) Fatal fault distribution. MD: maintaining and diagnostic system.

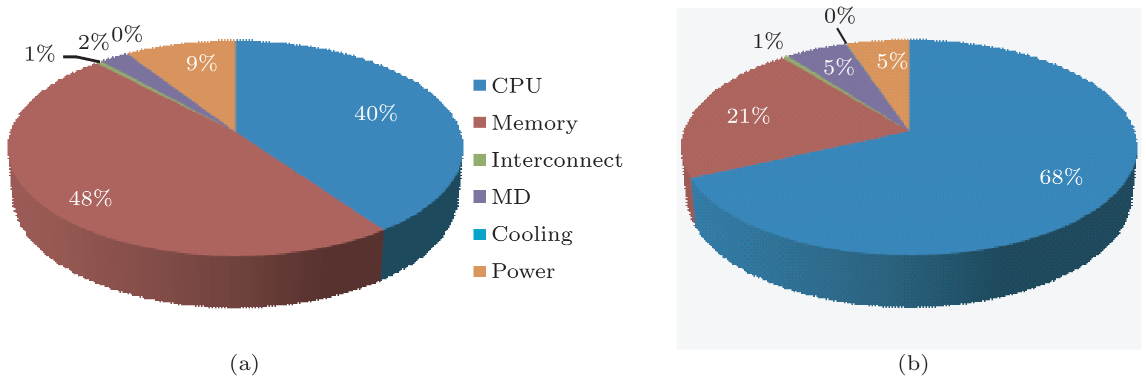


Fig.4. Fault distribution of Sunway TaihuLight. (a) Fault distribution (including non-fatal and fatal faults). (b) Fatal fault distribution. MD: maintaining and diagnostic system.

both the DRAM faults and the whole machine faults. The non-fatal faults here are DRAM SBEs which are correctable. The proportion of CPU faults in Fig.3(a) is less than that in Fig.4(a) because of more DRAM SBEs on Sunway BlueLight than on Sunway TaihuLight. The mainframe is composed of CPUs, DRAM chips and interconnecting devices. It can be seen that, as the major computing system of the two machines, the mainframe accounts for the most system faults on Sunway BlueLight with 87.18% and on Sunway TaihuLight with 90%. The cooling system and the power system have higher reliability for their hot spare configurations.

Conclusion 1. *On Sunway BlueLight and Sunway TaihuLight, DRAM SBEs dominate a large proportion of DRAM faults and the faults of the whole machine. The mainframe composed of CPUs, memory chips and the interconnecting system is the major source of failures on petascale supercomputers.*

With support of DRAM ECC on CPUs and memory controllers, DRAM SBEs can be corrected online automatically. This improves system reliability greatly. The mainframe has the largest effect on reliability, and needs to be specially analyzed and improved for fault tolerance.

4 Analyzing DRAM SBEs and DRAM MBEs

DRAM faults (including non-fatal and fatal) dominate a large proportion of faults on the system (91% on Sunway BlueLight and 48% on Sunway TaihuLight). Therefore analyzing DRAM faults in depth is necessary for the design of RAS (reliability, availability and serviceability). In this section, we first count resource utilization and jobs' running, and then analyze DRAM SBEs, DRAM MBEs, and the correlation relationship between them.

4.1 Workloads on the Machines

We have monitored Sunway BlueLight and TaihuLight about their daily running and computing resource utilization for a long time: 2011∼2016 on Sunway BlueLight and 2016∼2017 on Sunway TaihuLight. It shows that Sunway BlueLight and TaihuLight have high utilizations of computing resources up to 90% and 95% respectively.

The proportions of various types of jobs on Sunway BlueLight are shown in Fig.5(a). The statistic shows that the majority of jobs (above 99%) are computing-intensive, and the others are memory-intensive (46%), communication-intensive (4%) and I/O-intensive (21%).

The proportions of various types of jobs on Sunway TaihuLight are shown in Fig.5(b). It also shows that the majority of jobs (above 99%) are computing-intensive. The proportions of memory-intensive, communication-intensive and I/O-intensive jobs are also high enough up to about 53% respectively.

4.2 Analyzing DRAM SBEs

In this subsection, we first count DRAM SBEs according to daily running of Sunway BlueLight and Sunway TaihuLight. Then we conduct an experiment to find the correlation between DRAM SBEs and the reliability of components. Finally we obtain the characteristics of DRAM SBEs.

4.2.1 *Methodology and Principle*

In order to improve the reliability during the system's running, administrators may hot plug those CPU nodes where faults occur or jobs suspend to stop frequently into a special backup cabinet. The backup cabinet is used to diagnose and recover those CPU nodes.

The CPU nodes in the backup cabinet are initialized to run the operating system normally without any jobs on them. After a period of time the backup cabinet will be filled with CPU nodes subject to faults.

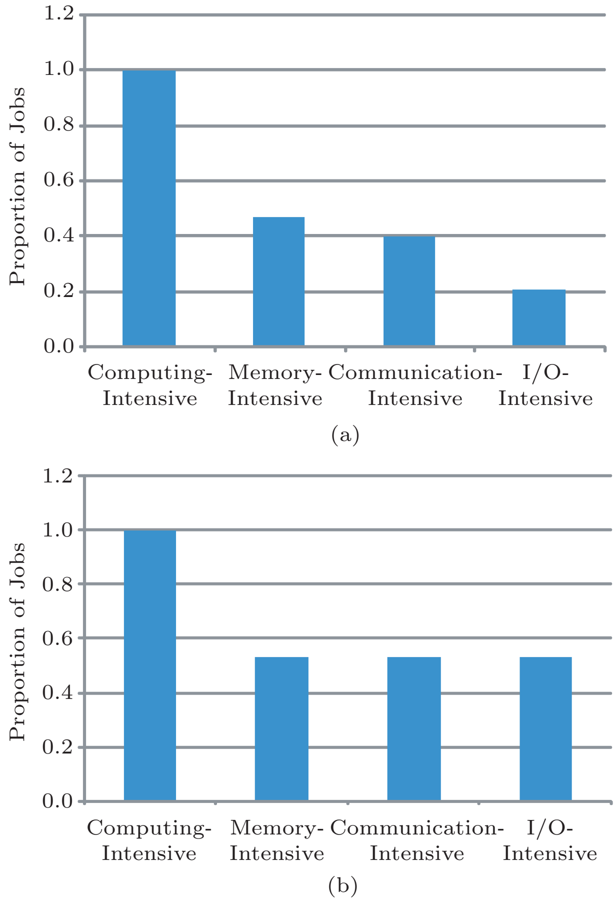


Fig.5. Proportion of jobs. (a) BlueLight. (b) TaihuLight.

We count the number of DRAM SBEs in each cabinet on Sunway BlueLight. The x-axis represents the CPU number and the y-axis presents the number of SBEs. The red dot represents that there are no DRAM SBEs on the corresponding CPU. The green dot represents the number of DRAM SBEs. Typical distribution of DRAM SBEs on each working cabinet is shown in Fig.6(a). Intuitively, although some CPUs have a big number of DRAM SBEs, the proportion of CPUs with DRAM SBEs in the cabinet is not so big. Distribution of DRAM SBEs in the backup cabinet is shown in Fig.6(b). It can be seen that the number of CPUs with DRAM SBEs in the backup cabinet turns to be very large obviously. We note that there are no jobs running in the backup cabinet; however the proportion of CPUs with DRAM SBEs is much higher than that of other cabinets.

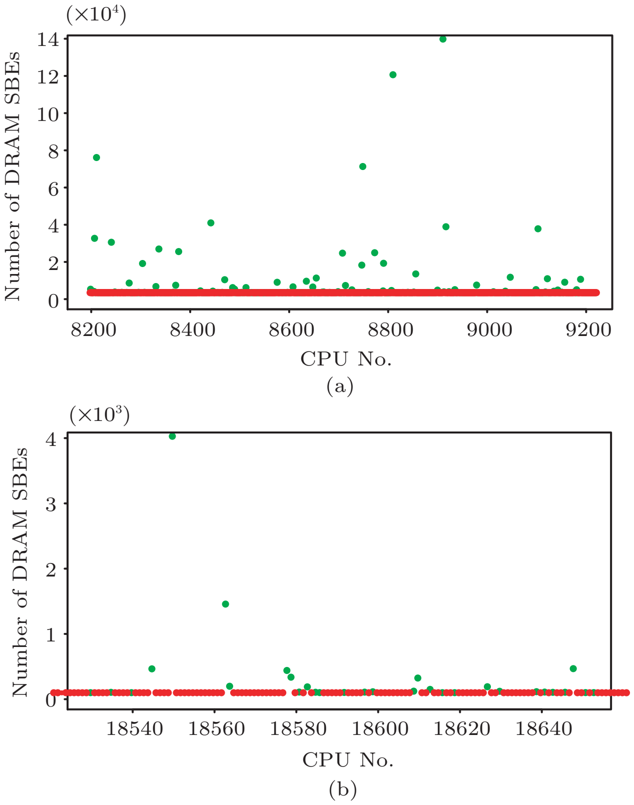


Fig.6. Distributions of DRAM SBEs in cabinets on Sunway BlueLight. (a) Typical cabinet 8. (b) Backup cabinet 18.

The above-mentioned situations are also the same on Sunway TaihuLight, as shown in Fig.7.

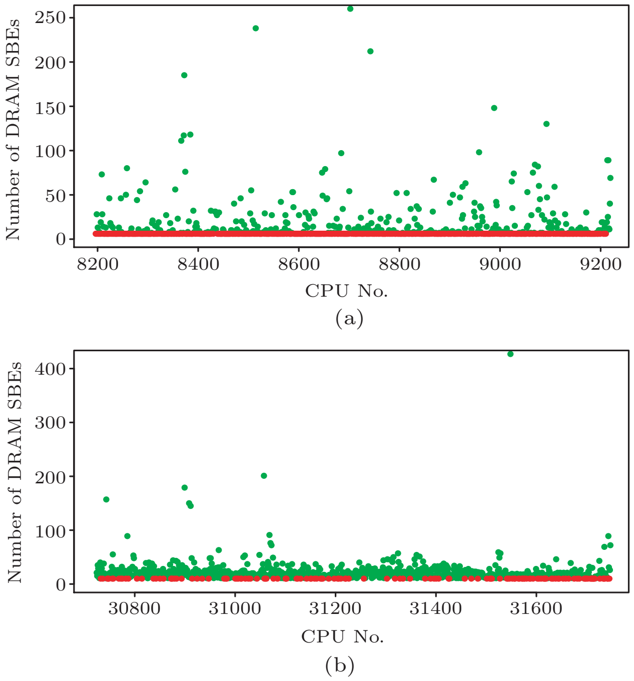


Fig.7. Distributions of DRAM SBEs in cabinets on Sunway TaihuLight. (a) Typical cabinet 8. (b) Backup cabinet 30.

From Fig.6 and Fig.7, we intuitively guess that there is some correlation between DRAM SBEs and the reliability of components. To figure it out, we conduct an experiment as follows.

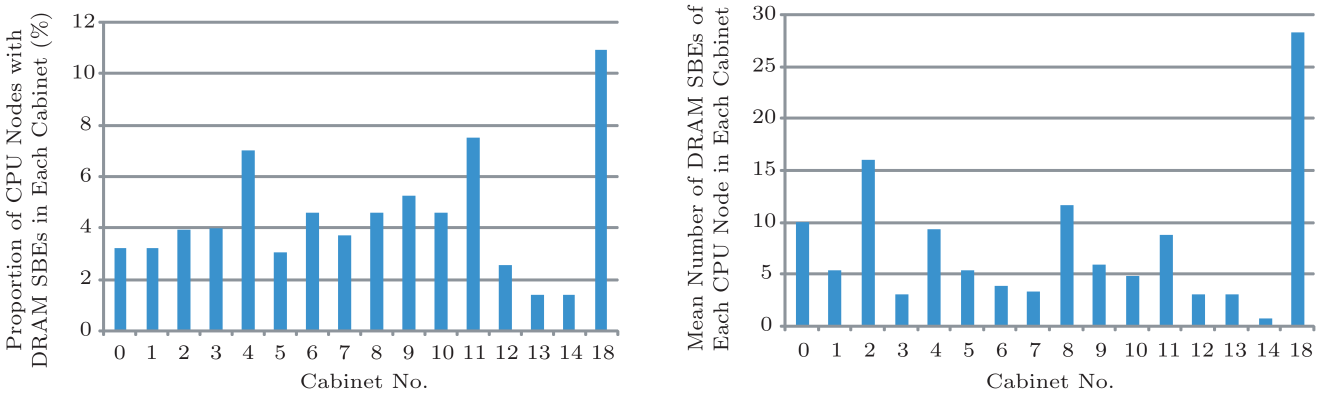
For the CPU nodes (including CPU and DRAM) on which transient failures (fatal faults of CPU or DRAM) occur or jobs suspend to stop for many times, we plug out them into the backup cabinet and diagnose them online (with OS running and no jobs). After a long period of time, the backup cabinet has a lot of CPU nodes being diagnosed online. At this time, we start monitoring and logging DRAM SBEs on all CPU nodes of Sunway BlueLight for 2 weeks (2012.2.22∼2012.3.7) and those of Sunway TaihuLight for 1 month (2017.01.01∼2017.01.31). All the cabinets excluding the backup one have jobs running during the period. The resource utilization is shown in Subsection 4.1.

4.2.2 *Analysis of DRAM SBE Characteristics*

According to the experiment, we perform a statistical analysis of the CPU nodes with DRAM SBEs in each cabinet. We get the proportion of CPU nodes with DRAM SBEs and the mean number of DRAM SBEs per CPU node in each cabinet on Sunway BlueLight and Sunway TaihuLight as shown in Fig.8 and Fig.9 respectively.

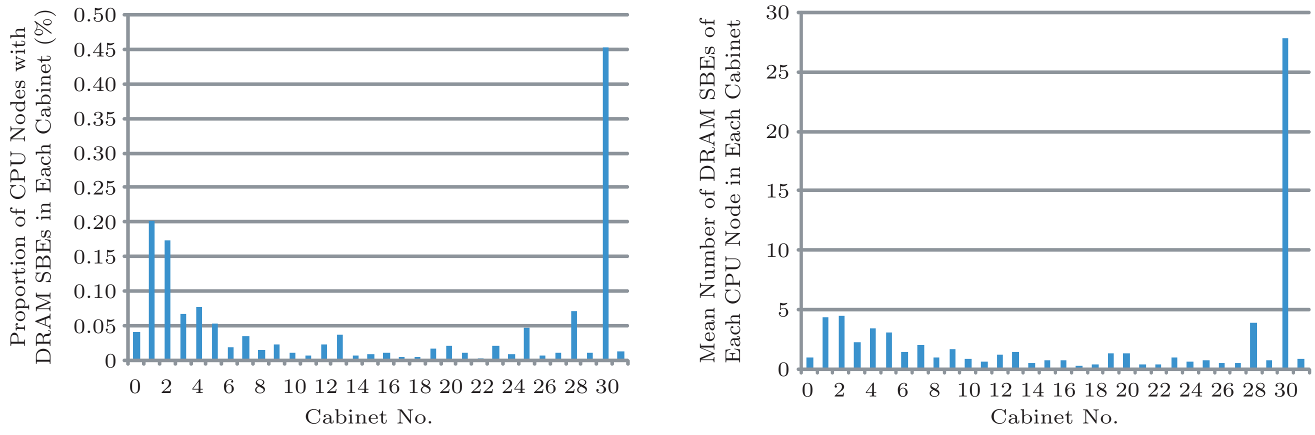
From Fig.8(a), we can see that the proportion (near to 11%) of CPU nodes with DRAM SBEs in the backup cabinet (cabinet 18) is much higher than that of other cabinets with jobs, even twice or more times. From Fig.8(b), we can also see that the mean number of DRAM SBEs of each CPU node in the backup cabinet is much larger than that of the other cabinets, even twice or more times.

The above-mentioned situations also exist on Sun-



(a) (b)

Fig.8. Characteristics of DRAM SBEs on Sunway BlueLight. (a) Proportion of CPU nodes with DRAM SBEs in each cabinet. (b) Mean number of DRAM SBEs of each CPU node in each cabinet.



(a) (b)

Fig.9. Characteristics of DRAM SBEs on Sunway TaihuLight. (a) Proportion of CPU nodes with DRAM SBEs in each cabinet. (b) Mean number of DRAM SBEs of each CPU node in each cabinet.

way TaihuLight, as shown in Fig.9. The proportion of CPU nodes with DRAM SBEs in the backup cabinet (cabinet 30) is much higher than that of the other cabinets with jobs, even 10 or more times. The mean number of DRAM SBEs of each CPU node in the backup cabinet is twice larger than that of the other cabinets, even up to 10 times.

A common situation is found in the two machines. The proportion of CPU nodes with DRAM SBEs and the mean number of DRAM SBEs of each CPU node in the backup cabinet are both significantly larger than those of the other cabinets with jobs running daily respectively. From the experimental methodology it can be seen that the reliability and the stability of CPU nodes in the backup cabinet are lower than those in the other cabinets.

Conclusion 2. *On peta-scale supercomputers, DRAM SBEs may have no correlation with jobs running and have direct correlation with the reliability of CPU nodes or DRAMs. The lower the reliability of CPU nodes or DRAMs is, the higher the probability of DRAM SBEs is.*

Whenever the accumulative count of DRAM SBEs on certain CPU node increases significantly, the reliability of the CPU node decreases significantly. For daily management, the CPU node can be replaced in time to improve system reliability. For resource scheduling, the CPU node can be allocated with a lower probability to improve the reliability of jobs. For fault tolerance, the CPU node can be proactively migrated and replaced online to reduce the cost of fault tolerance.

4.3 Analyzing the Correlation Between DRAM SBEs and MBEs

The relationship between DRAM SBEs and MBEs is generally considered as cause and effect, that is, DRAM SBEs occurring many times on certain CPU node will cause DRAM MBEs to occur successively. There is an infrastructure on warning of DRAM MBEs with DRAM SBEs on Sunway series supercomputers. The infrastructure brings out more complexity on CPU design, more chip area and more cost of power and fault tolerance. In addition, the effect of DRAM MBEs on DRAM MBEs also needs to be analyzed.

4.3.1 Methodology and Principle

Sequential rule mining is used to find the correlation among time-ordered data item sets[21-25]. The mined sequential rules are used to demonstrate the correlation or the cause and effect among some objects. We can use sequential rules to describe the correlation between DRAM SBEs and MBEs without time window constraints. Supposing that there is causal relationship between DRAM DBEs and MBEs, there must be sequential rules as follows,

$$<DRAM\ SBE, DRAM\ SBE, ..., DRAM\ SBE> \\ \rightarrow <DRAM\ MBE> .$$

The sequential rules mean that some DRAM SBEs which have occurred are followed by a DRAM MBE. The rule can be used to verify the causal relationship between the DRAM SBE and the DRAM MBE.

The formalization of the correlation between the DRAM SBE and the DRAM MBE is shown as follows:

$$<X_1, X_2, ..., X_i>\rightarrow<Y>,$$

where X_i and Y are items, $< X_1, X_2, ..., X_i >$ is an i-sequence ordered by time ascending, $< Y >$ is a 1-sequence, X_i indicates a DRAM SBE, Y indicates a DRAM MBE or DRAM fatal fault, $i \geqslant 1$, $i \in \mathbb{N}$ and $\mathbb{N}$ is the set of natural numbers.

In hardware design, DRAM SBEs and DRAM MBEs are two different types of errors; therefore there is an equation: $\{X_1, X_2, ..., X_i\} \cap \{Y\} = \emptyset$.

Only if $\exists i \in N$ to make $<X_1, X_2, ..., X_i>\rightarrow<Y>$ true, it can be said that there is a causal relationship between DRAM SBEs and DRAM MBEs. The count of such i may be one or more than one.

Whatever the count of such i is, there is always a requirement for support and confidence:

$$\begin{cases} s(<X_1, X_2, ..., X_i> \rightarrow <Y>) \\ = \dfrac{\sigma(<X_1, X_2, ..., X_i, Y>)}{cpu\_count} \geqslant min\_sup, \\ c(<X_1, X_2, ..., X_i>\rightarrow<Y>) \\ = \dfrac{\sigma(<X_1, X_2, ..., X_i, Y>)}{\sigma(<X_1, X_2, ..., X_i>)} \geqslant min\_conf, \end{cases}$$

where $\sigma(X)$ is the count of frequency, namely, the number of CPU nodes on which the fault sequence X occurs, cpu_count is the count of all CPU nodes, min_sup is the minimum support, and min_conf is the minimum confidence.

Similarly, the correlation between DRAM MBEs and DRAM MBEs can also be described with sequential rules.

Generally, the support may be low, but the confidence must be high enough so as to make the sequential rule true.

We use GSP[21] algorithm to mine and analyze the correlation between DRAM SBEs, DRAM MBEs or fatal faults taken from the system fault database. The data preprocessing includes three steps. The first step is to filter data and remove redundancy. It filters out non-DRAM faults from all types of faults and removes redundant fault data with the same location, time and fault type. The second step is to sort data. It makes data ordered by CPU number and time of occurrence and obtains a fault sequence database. The third step is to specify and map fault types. It maps multiple detailed types of DRAM single bit errors to DRAM_SBE and multiple detailed types of DRAM multiple bit errors to DRAM_MBE. At last the mapped fault sequence database is obtained, as shown in Table 1.

Table 1. Sequence Database of DRAM Faults

CPU No.	Occurrence Time	Fault Event
1	2011.12.15 11:20:35	DRAM_SBE
1	2011.12.16 14:08:11	DRAM_MBE
⋮	⋮	⋮

4.3.2 Analysis of Correlation

The DRAM faults occurring in the period of two years (2011.12∼2013.12) are taken from the fault database on Sunway BlueLight. The big dataset has fault records of about 5 190 000. We use GSP[21] algorithm to mine sequential rules. When the support of sequential rules is low, the calculation and the required memory increase significantly. In order to improve efficiency, the dataset of DRAM faults is divided into multiple partitions (by a time span of every three months).

In order to analyze the correlation between DRAM SBEs and DRAM MBEs or the correlation between DRAM MBEs and DRAM MBEs, the frequent sequences containing DRAM MBEs need to be extracted. The support needs decreasing constantly until the frequent sequences containing DRAM MBEs occur. The frequent sequences involving the correlation between DRAM SBEs and DRAM MBEs or DRAM MBEs and DRAM MBEs are selected from the generated frequent sequences. The sequential rules under various supports in specified time spans are shown in Table 2. As the support decreases, the antecedents of mined

Table 2. Examples for Sequential Rules with the Correlation Between DRAM SBEs and DRAM MBEs

Time Span	Minimum Support (%)	Frequent Sequences with DRAM_MBE as the Last Item and the Frequency Counts	Sequential Rule with DRAM_MBE on Right Side
2012.9.1 ∼ 2012.12.1	3	<{DRAM_SBE}{DRAM_MBE}> (21) <{DRAM_MBE}{DRAM_MBE}> (22)	<{DRAM_SBE}>→<{DRAM_MBE}:conf=21/525 =4% <{DRAM_MBE}>→<{DRAM_MBE}>:conf=22/22 =100%
	1	<{DRAM_SBE}{DRAM_MBE}> (21) <{DRAM_MBE}{DRAM_MBE}> (22) <{DRAM_SBE}{DRAM_SBE}{DRAM_MBE} > (13) <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE} > (15) <{DRAM_SBE}{DRAM_SBE}{DRAM_SBE} {DRAM_MBE}> (9) <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE} {DRAM_MBE}> (11) <{DRAM_SBE}{DRAM_SBE}{DRAM_SBE} {DRAM_SBE}{DRAM_MBE}> (6) <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE} {DRAM_MBE}{DRAM_MBE}> (10) <{DRAM_SBE}{DRAM_SBE}{DRAM_SBE} {DRAM_SBE}{DRAM_SBE}{DRAM_MBE} > (5) <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE} {DRAM_MBE}{DRAM_MBE}{DRAM_MBE} > (9)	<{DRAM_SBE}>→<{DRAM_MBE}:conf =21/525=4% <{DRAM_SBE}{DRAM_SBE}> →<{DRAM_MBE}>:conf=13/410=3.1% <{DRAM_SBE}{DRAM_SBE}{DRAM_SBE}> →<{DRAM_MBE}> :conf =9/360=2.5% <{DRAM_SBE}{DRAM_SBE}{DRAM_SBE} {DRAM_SBE}>→<{DRAM_MBE}>:conf =6/333=1.8% <{DRAM_SBE}{DRAM_SBE}{DRAM_SBE} {DRAM_SBE}>→<{DRAM_SBE} {DRAM_MBE}>:conf=5/313=1.5% <{DRAM_MBE}>→<{DRAM_MBE}>:conf =22/22=100% <{DRAM_MBE}{DRAM_MBE}> →<{DRAM_MBE}>:conf=15/22=68% <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE}> →<{DRAM_MBE}>:conf =11/15=73% <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE} {DRAM_MBE}>→<{DRAM_MBE}>:conf =10/11=90% <{DRAM_MBE}{DRAM_MBE}{DRAM_MBE} {DRAM_MBE}{DRAM_MBE} > →<{DRAM_MBE}>:conf=9/10=90%

sequential rules with a formalization of $<SBE_1$, ..., $SBE_i>\to<MBE>$ are longer and longer and at the same time the confidence of the rules decreases increasingly. The reason is that the count of $<SBE_1$, ..., $SBE_i, MBE>$ is decreasing rapidly in comparison with that of $<SBE_1, ..., SBE_{i-1}, MBE>$ while the count of $<SBE_1, ..., SBE_i>$ is just little less than $<SBE_1, ..., SBE_{i-1}>$ in proportion.

The maximum confidence of the rules that describe the casual relationship between some DRAM SBEs and a DRAM MBE is shown in Fig.10. It can be seen that the maximum confidence of the rules that describe the correlation between some DRAM SBEs and a DRAM MBE is only 4%. It proves that DRAM SBEs do not necessarily cause a DRAM MBE.

The confidence of the sequential rules that describe the correlation between DRAM MBEs and DRAM MBEs is also shown in Fig.10. The minimum confidence is 60%, the maximum is 100%, and the average is 88%. Therefore some DRAM MBEs often cause a following DRAM MBE.

The further analysis finds that there are similar conclusions on Sunway TaihuLight.

Conclusion 3. *In petascale supercomputers, there is no obvious casual relationship between DRAM SBEs and DRAM MBEs. DRAM SBEs do not necessarily cause a DRAM MBE. A DRAM MBE is often followed by a DRAM MBE.*

The circuit of DRAM SBE warning of DRAM MBE on the chip can be removed. The complexity, area and power of the CPU chip could be reduced; thus the reliability could be improved. The DRAM MBEs may be used to predict a following DRAM MBE for proactive fault tolerance.

5 Analyzing Failure Time on Multi-Dimension

Zheng *et al.*[7] showed that the failure rate may have no correlation with the workload of small jobs, and the workload of wider jobs may have a significant impact on the failure rate. In Sunway BlueLight and TaihuLight, the majority of jobs are wider jobs. Especially in Sunway TaihuLight, the jobs using the whole machine are common. We take cabinet 8 as an example to see how applications influence the faults. The statistics show that memory intensity may bring out a large number of memory faults, the computing intensity may bring out a large number of CPU faults, and the interconnecting has high reliability, as shown in Table 3.

The quantitative analysis on failure time can be used to get characteristics of fault distribution and failure rate. This constitutes a theoretical and engineering basis for failure prediction.

The failures mean that some fatal faults in the system could not be repaired by hardware itself automatically and the system could not continue running. For non-fatal faults, they can be repaired automatically by the hardware's error correction (e.g., single error correction by ECC). For the fatal faults, they could not be repaired automatically. They lead to the interruption of the system's running or system failure. The fatal faults need to be corrected by out-of-band system and software. Analysis of failure time aims to find a model of the time between failures for petascale supercomput-

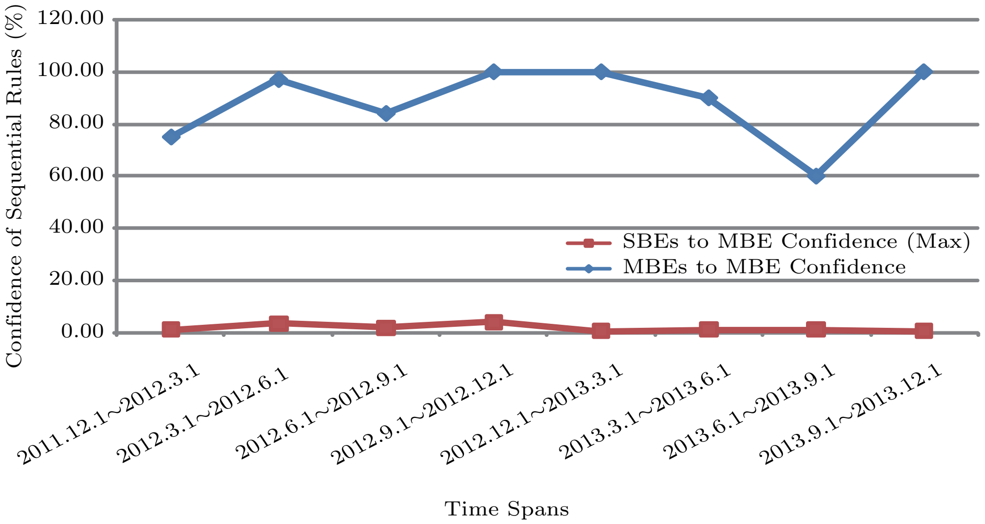


Fig.10. Confidence of sequential rules in different time spans.

Table 3. Effect of Application on Faults in the Typical Cabinet of Sunway TaihuLight

Application	Time Span (Month)	Cabinet	Number of Memory Faults	Number of CPU Faults	Number of Interconnect Faults
Computing-intensive	2014.10	8	93	33	0
Computing-intensive	2014.12	8	7	5	0
Computing intensive	2015.01	8	51	8	0
Computing- and memory-intensive	2015.09	8	995	136	0
Computing- and memory-intensive	2015.10	8	527	867	0
Communication- and memory-intensive	2015.11	8	195	58	0
Communication- and memory-intensive	2015.12	8	115	53	0
Communication- and memory-intensive	2016.08	8	284	20	0
Communication- and memory-intensive	2016.09	8	277	85	0

ers and to give a quantitative description for the failure time.

From failure root cause on Sunway BlueLight and TaihuLight, we can see that the mainframe (including CPU, memory, and interconnecting) has the most effect on system reliability. In this section, the time between failures of the mainframe on Sunway BlueLight and TaihuLight is analyzed in spatial dimension (CPU nodes, computing cards and the whole mainframe) and time dimension (different time spans). Finally, a uniform model for failure time distribution is built.

5.1 Sensors Deployment and Data Collection

The mainframe of Sunway BlueLight consists of general multi-core CPUs called Sunway1600, DDR3-1066 memory chips and an IBA-compatible network. There are a multi-core CPU and DRAM chips on each CPU node.

The mainframe of Sunway TaihuLight consists of heterogeneous many-core CPUs called Sunway26010, DDR3-2133 memory chips and an IBA-compatible network. There are 260 general cores on the Sunway26010. The CPU has a heterogeneous architecture with computing arrays and distributed shared memory combined.

In the two machines, eight CPU nodes and an integrated IBA card constitute one computing card which is a pluggable unit to be replaced and repaired easily. The failures of computing cards mainly derive from fatal faults of CPU nodes and IBA cards. The fatal faults of IBA cards include those of components and power supply. The fatal faults of CPU nodes include those of CPUs and DRAM chips.

The fatal faults of Sunway1600 are protocol component errors, successive checking errors, OS panic, invalid instruction errors, instruction stream faults, data stream faults, errors of accessing memory component, cache uncorrectable errors, PCI-E faults, etc. The fatal faults of memory are DRAM MBEs, uncorrectable errors of tags on memory controller, memory controller errors, etc.

The fatal faults of Sunway26010 are those from master cores, slave core arrays, protocol transformation components, memory controllers, NOC (network-on-chip) and system interfaces, etc.

In the two machines, the root cause of failures on the mainframe is consistent with that of the computing card. Analysis of reliability on the mainframe can use the serial model of computing cards.

There are 80 various types of faults in total and about 40 types of fatal faults causing the mainframe failures on Sunway BlueLight. There are more than 170 types of faults in total and 108 types of fatal faults causing the mainframe failures on Sunway TaihuLight.

According to the failure root causes of the two supercomputers, the sensors covering all fault units are deployed by the distributed infrastructure of fault collection.

1) *Sensors of CPU Node Faults.* BMC scans CPU's total error signal and fault registers periodically. The faults of the CPU node are found through CPU total error signal and identified through fault registers.

2) *Sensors of IBA Card Faults.* BMC scans the signals of physical and logical links on the HCA chips periodically, and finds their faults through analyzing link status on the HCA chips.

3) *Sensors of Power Supply Faults.* BMC uses the interface of I^2C to scan status of power supply on CPU boards and HCA boards periodically, and find their power controller's faults.

4) *Sensors of Temperature.* BMC uses the interface of I^2C to obtain temperatures of CPU and HCA through their inner temperature sensors.

5) *Sensors of Software*. On the OS of each CPU node, the customized monitoring program on the user level is used to collect various types of software faults and hardware fault scenes periodically.

The sensors cover all the computing units on Sunway BlueLight and Sunway TaihuLight. They ensure the completeness and correctness of original fault data. The data obtained by distributed sensors is stored in real time, and used to build big data for analyzing system faults.

Time spans of the original fault data are three years of Sunway TaihuLight (2014.7.1∼2017.10.1) and five years of Sunway BlueLight (2011.7.1∼2016.7.1). The original data is filtered to remove redundancy and obtain only failure related data by the database SQL technology.

5.2 Analysis of Time Interval Between Failures

In this subsection, we analyze the failure time of CPU nodes, computing cards and the mainframe to illustrate the failure model of petascale supercomputers.

5.2.1 Methodology

The exponential distribution ($T \sim E(\lambda)$), lognormal distribution ($T \sim LN(\mu, \sigma^2)$), Weibull distribution ($T \sim W(m, \eta)$) and Gamma distribution ($T \sim \Gamma(\alpha, \lambda)$) are typical life distributions. The probability density functions of the four distributions are shown in Table 4[26].

These mathematical models are selected and used to analyze the failure time of each fault unit comparatively on the same temporal and spatial dimension. The maximum likelihood estimation is used to parameterize each of the distributions to fit the empirical cumulative distribution as well as possible. We use Kolmogorov-Smirnov (K-S) test to evaluate the fit of each distribution with the empirical data. The produced p-value between 0 and 1 is used to evaluate the goodness of fit. Lower p-value indicates a worse fit, and conversely, the better. In general, a distribution has a good fit with the empirical data when the p-value is larger than a standard threshold of 0.05. p-value below the threshold indicates that the data did not come from the corresponding distribution.

We use the period of one year as the time interval. The running period of Sunway TaihuLight is divided into three time spans. The time spans are shown in Table 5.

5.2.2 Analyzing Failure Time of Sunway TaihuLight

We select two random CPU nodes, two random computing cards, and the mainframe on Sunway TaihuLight as examples, as shown in Table 5. Then we analyze their time intervals between failures in divided time spans. According to the actual failure data, the time between failures on the CPU nodes, the computing cards and the mainframe in different time spans are analyzed, as shown in Fig.11.

The green curve depicts the empirical failure time and the other curves depict the corresponding Weibull, Gamma, exponential and lognormal distribution fitted.

From Fig.11, it can be visually found that during the three different time spans, the exponential distribution worst fits the actual failure data, while Weibull, Gamma and lognormal distributions better fit the data.

With the actual failure data, the MLE (maximum likelihood estimation) is used to estimate the parameters of corresponding exponential, lognormal, Gamma and Weibull distributions. The parameters and the fitness of the corresponding distributions are shown in Table 5. The p-values of the K-S test with the four distributions are also listed.

We take the CPU node (4-3-26-1-1) as an example.

In the first time span (2014.7.1∼2015.9.1), we find the exponential distribution worst fit the actual data with the p-value of 0.006 416 931. This indicates that the empirical failure data does not come from the exponential distribution. With the actual failure data, we find that Weibull, Gamma and lognormal distributions have good fit with p-values of 0.876 947 4, 0.759 551 2 and 0.123 488 6 respectively. The Weibull distribution

Table 4. Typical Life Distributions

Distribution	Failure Probability Density Functions and Parameters
Exponential	$f(t) = \lambda e^{-\lambda t}, \lambda > 0, t > 0$; λ is the failure rate parameter
Lognormal	$f(t) = \frac{1}{\sqrt{2\pi}\sigma t}\exp(-\frac{(\ln t-\mu)^2}{2\sigma^2}), t > 0, \sigma > 0$; μ and σ are the mean and the standard deviation of the logarithmically transformed variables, respectively
Weibull	$f(t) = \frac{m}{\eta}(\frac{t}{\eta})^{m-1}\exp(-(\frac{t}{\eta})^m), t > 0$; m is the shape parameter and η is the scale parameter or characteristic life
Gamma	$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t}, t > 0$; α is the shape parameter, $1/\lambda$ is the scale parameter, and $\Gamma(\cdot)$ is the usual Gamma function

Table 5. Distributions and Their Parameters of TBF on Sunway TaihuLight

Epoch	Time	Object	Failure	Model Parameters and K-S Test			
	Span		Count	Weibull(m, η)	Gamma(α, λ)	Exponential(λ)	Lognormal(μ, σ)
1	2014.7.1 ~ 2015.9.1	CPU node (4-0-2-1-1)	37	m =0.651 336 5 η = 546 927.6 p-value=0.855 171 9	α = 0.618 527 4 λ = 8.716 579e-07 p-value=0.994 287 4	λ = 1.409 247e-06 p-value= 0.185 067 8	μ = 12.245 47 σ = 2.352 053 p-value=0.339 683 5
		CPU node (4-3-26-1-1)	79	m =0.574 581 1 η = 432 871.8 p-value=0.876 947 4	α = 0.538 868 7 λ = 8.387 872e-07 p-value= 0.759 551 2	λ = 1.556 571e-06 p-value= 0.006 416 931	μ = 11.901 26 σ = 2.502 115 p-value=0.123 488 6
		Card (4-0-2)	48	m =0.530 579 1 η = 457 065.1 p-value=0.677 926 4	α = 0.295 431 2 λ = 3.820 866e-07 p-value= 0.392 685 6	λ = 1.293 318e-06 p-value= 0.009 532 383	μ = 11.878 94 σ = 2.660 758 p-value=0.137 873 9
		Card (0-0-15)	53	m =0.446 447 5 η = 359 889.9 p-value=0.958 894 5	α = 0.225 449 5 λ = 2.575 918e-07 p-value= 0.485 646 5	λ = 1.142 57e-06 p-value=0.000 449 257 6	μ = 11.506 81 σ = 2.725 863 p-value=0.379 197 7
		Mainframe	13 173	m =0.596 615 5 η = 893.236 4 p-value=0.472 270 9	α = 0.388 771 5 λ = 0.000 280 522 5 p-value=0.089 615 62	λ = 0.000 721 561 5 p-value=6.077 687e-05	μ = 5.879 552 σ = 1.844 73 p-value=0.262 231 6
2	2015.9.1 ~ 2016.8.1	CPU node (4-0-2-1-1)	75	m =0.808 786 9 η = 29 529.02 p-value=0.388 657 9	α = 0.419 149 8 λ = 1.230 676e-05 p-value=0.113 027 2	λ = 2.936 125e-05 p-value=0.045 869 24	μ = 9.671 269 σ = 1.221 018 p-value=0.687 602 4
		CPU node (4-3-26-1-1)	57	m =0.473 529 6 η = 157 912.6 p-value=0.178 391 5	α = 0.163 277 2 λ = 3.959 475e-07 p-value=2.430 402e-05	λ = 2.425 003e-06 p-value=5.999 534e-11	μ = 10.893 58 σ = 2.089 942 p-value=0.361 124
		Card (4-0-2)	76	m =0.808 786 9 η = 29 529.02 p-value=0.388 657 9	α = 0.419 149 8 λ = 1.230 676e-05 p-value = 0.113 027 2	λ = 2.936 125e-05 p-value=0.045 869 24	μ = 9.671 269 σ = 1.221 018 p-value=0.687 602 4
		Card (0-0-15)	66	m =0.506 483 η = 234 787.4 p-value=0.464 502 2	α = 0.236 669 7 λ = 4.717 649e-07 p-value=0.071 949 23	λ = 1.993 347e-06 p-value=2.708 325e-06	μ = 11.330 82 σ = 2.041 925 p-value=0.792 034 6
		Mainframe	17 236	m =0.602 770 1 η = 490.093 2 p-value=0.242 759 8	α = 0.213 720 3 λ = 0.000 256 960 3 p-value=0.000 162 547 5	λ = 0.001 202 32 p-value=2.689 054e-05	μ = 5.380 194 σ = 1.586 996 p-value=0.812 874 6
3	2016.8.1 ~ 2017.10.1	CPU node (4-0-2-1-1)	0	–	–	–	–
		CPU node (4-3-26-1-1)	29	m =0.558 366 2 η = 545 029.5 p-value=0.994 984 7	α = 0.280 416 7 λ = 3.109 955e-07 p-value = 0.474 211 2	λ = 1.109 048e-06 p-value =0.068 346 42	μ = 12.166 21 σ = 2.296 362 p-value=0.543 369 5
		Card (4-0-2)	23	m =0.308 258 5 η = 90 703.32 p-value=0.648 566 9	α = 0.119 835 3 λ = 1.247 365e-07 p-value = 0.124 993 4	λ = 1.040 899e-06 p-value=2.241 575e-09	μ = 9.718 841 σ = 3.400 313 p-value=0.955 907 9
		Card (0-0-15)	2	–	–	–	–
		Mainframe	14 047	m =0.594 204 4 η = 2 012.979 p-value=0.470 772 1	α = 0.688 756 1 λ = 0.000 240 177 2 p-value=0.041 091 49	λ = 0.000 348 711 5 p-value=0.001 730 453	μ = 6.589 961 σ = 2.178 682 p-value=0.183 466 9

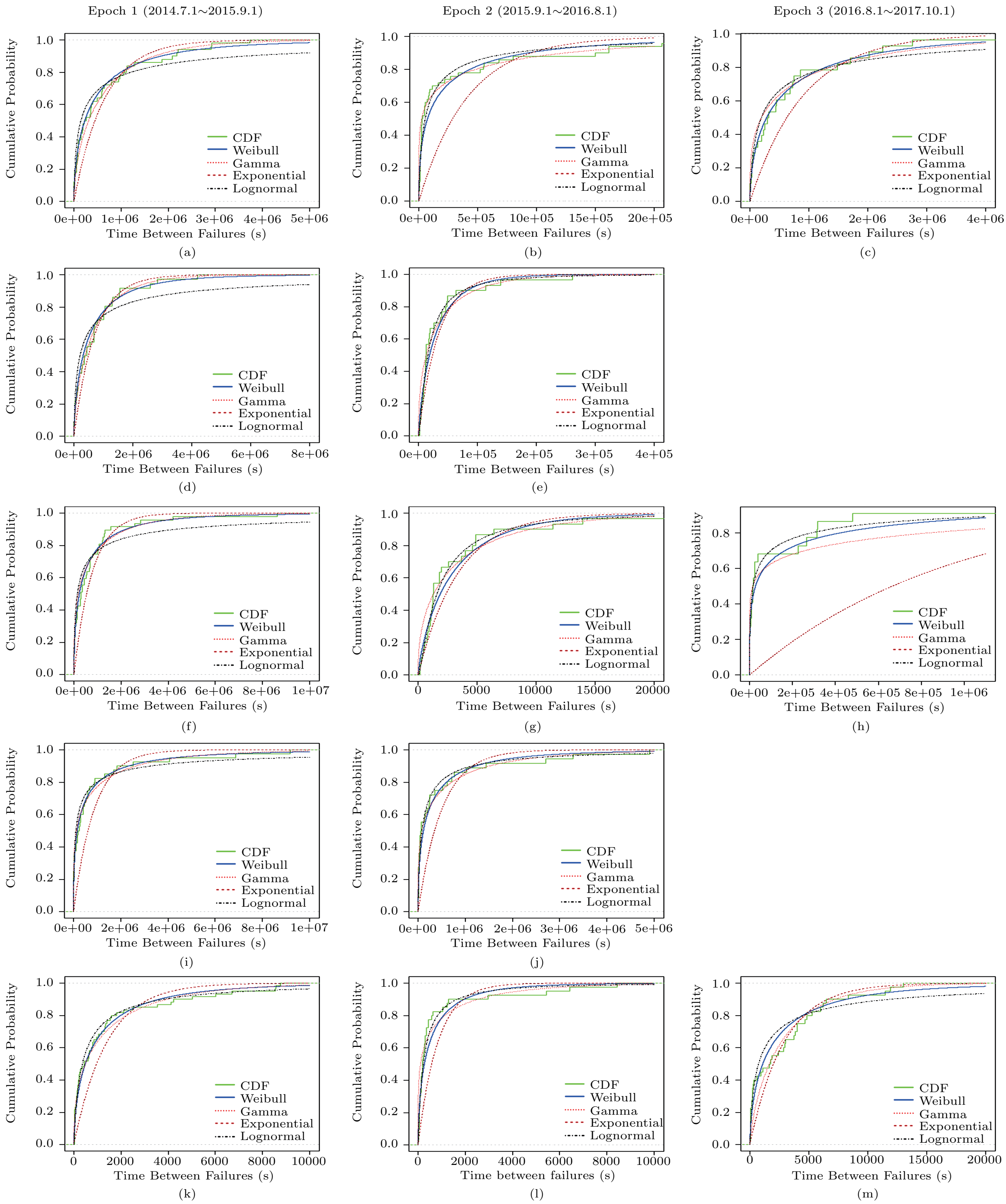
fits best. The second is the Gamma distribution, and the third is the lognormal distribution.

In the second time span (2015.9.1~2016.8.1), the exponential and Gamma distributions do not fit the empirical failure data with p-values of 5.999 534e-11 and 2.430 402e-05 respectively. The lognormal distribution fits the best and the Weibull distribution fits better (with p-values of 0.361 124 and 0.178 391 5 respectively).

In the third time span (2016.8.1~2017.10.1), the four distributions all fit well with the actual failure data. The Weibull distribution fits the best with p-value of 0.994 984 7, the second is the lognormal distribution with p-value 0.543 369 5, the third is the Gamma distribution with p-value of 0.474 211 2, and the last is the exponential distribution with p-value of 0.068 346 42.

In each time span, the shape parameter m of corresponding Weibull distribution is all less than 1. According to the failure rate function of the Weibull distribution $\lambda(t) = \frac{m}{\eta}(\frac{t}{\eta})^{m-1}$, this indicates that for the CPU node (4-3-26-1-1), as time increases, the probability of failures under the conditions that no failures occurred is ever decreasing in a unit of time or the near future.

In addition, it can be found that the Weibull distribution best fits the empirical failure data on the CPU node (4-3-26-1-1). However the parameters of the cor-



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m)

Fig.11. Distributions of TBF with different grains of resource and different time spans on Sunway TaihuLight. (a) CPU(4-3-26-1-1) distribution of TBF in epoch 1. (b) CPU(4-3-26-1-1) distribution of TBF in epoch 2. (c) CPU(4-3-26-1-1) distribution of TBF in epoch 3. (d) CPU(4-0-2-1-1) distribution of TBF in epoch 1. (e) CPU(4-0-2-1-1) distribution of TBF in epoch 2. (f) Card(4-0-2) distribution of TBF in epoch 1. (g) Card(4-0-2) distribution of TBF in epoch 2. (h) Card(4-0-2) distribution of TBF in epoch 3. (i) Card(0-0-15) distribution of TBF in epoch 1. (j) Card(0-0-15) distribution of TBF in epoch 2. (k) Sunway TaihuLight mainframe: distribution of TBF in epoch 1. (l) Sunway TaihuLight mainframe: distribution of TBF in epoch 2. (m) Sunway TaihuLight mainframe: distribution of TBF in epoch 3.

responding Weibull distribution in different time spans have considerable difference from each other (the shape parameter of $m = 0.574\,581\,1$ and the scale parameter of $\eta = 432\ 871.8$ in epoch 1; $m = 0.473\ 529\ 6$ and $\eta = 157\ 912.6$ in epoch 2; $m = 0.558\ 366\ 2$ and $\eta = 545\ 029.5$ in epoch 3 respectively). This indicates that the parameters of the same failure distribution model on the same component are varying as the time changes. And the number of failures on CPU node (4-3-26-1-1) also varies in three different time spans. With the further analysis, it is found that the CPU's workload in the three time spans has different features.

Similarly, for the other CPU nodes, computing cards and the mainframe listed in Table 5, the analysis of failure data shows that the Weibull distribution best fits the empirical TBF (time between failures). It is also found that among the fault units with the same type but different locations, there have different parameters of Weibull distribution corresponding to the empirical TBF (for example, in epoch 1, the parameters of Weibull distribution corresponding to TBF of Card (4-0-2) are $m = 0.530\,579\,1$, $\eta = 457\,065.1$, while the parameters of Weibull distribution corresponding to TBF of Card (0-0-15) are $m = 0.446\,447\,5$, $\eta = 359\ \ 889.9$).

In fact, the analysis of all the fault units with different grains on Sunway TaihuLight shows that the Weibull distribution best fits the empirical failure data. The shape and characteristic life parameters of Weibull distribution corresponding to the same object are also different from one another in different time spans. This indicates that although the Weibull distribution can be used to depict the failure time distribution model of the same component, the parameters of the model vary as the workload changes.

The comprehensive analysis shows that for all the fault units on Sunway TaihuLight, the Weibull distribution can be used to quantitatively describe the TBF and has the following characteristics.

1) For the same component the parameters of Weibull distribution corresponding to the empirical TBF vary as the time and the workload change.

2) For the same type of components with different locations, the parameters of Weibull distribution corresponding to the empirical TBF may have difference.

3) For the same component in different time spans, the failure rates (or the counts of faults) may have difference.

We also conduct a comprehensive analysis of TBF on Sunway BlueLight and find that the failure time model of Sunway BlueLight is similar to that of Sunway TaihuLight with the characteristics described above.

5.3 Uniform Model of Failure Time on Multi-Dimension

From Fig.1, it can be found that organizations of Sunway BlueLight and TaihuLight have some similarity. The main difference of the two machines is that they use homogeneous multi-core CPUs and heterogeneous many-core CPUs respectively. Sunway many-core CPU with 260 cores is much more complex than Sunway multi-core CPU with 16 cores. Sunway TaihuLight has much larger scale and complexity than Sunway BlueLight. The two machines use different parallel computing models. Sunway BlueLight supports message-based and shared memory computing models. In addition to the above two models, Sunway TaihuLight supports accelerating computing model to exploit the computing capability of slave cores.

The analysis of TBF for the two supercomputers shows that although the supercomputers based on multi-core and many-core CPUs have much difference in system complexity, computing model and the work load, they have the same failure time model.

The analysis of the failure time on multi-dimension (with different grains such as one CPU node, one computing card and the mainframe, and in different time spans) for Sunway BlueLight and Sunway TaihuLight shows that the Weibull distribution best fits the empirical time between failures in different time spans and different grains of resource. The TBF of the two typical supercomputers could be quantitatively described by a uniform mathematical model, namely, the Weibull distribution.

The failure probability density function is $f(t) = \frac{m}{\eta}(\frac{t}{\eta})^{m-1}\exp(-(\frac{t}{\eta})^{m}), t \geqslant 0$. The failure time distribution model is represented as $TBF \sim W(m, \eta), m > 0$, where m is shape parameter and η is characteristic life.

Conclusion 4. *On petascale supercomputers, the Weibull distribution best fits the empirical time between failures on multi-dimension. For the same component, the parameters of Weibull distribution corresponding to the empirical TBF vary as the time and workload change. The component's failure rates (or the counts of faults) may have difference in different time spans. For the same type of components with different locations, the parameters of Weibull distribution corresponding to the empirical TBF may have difference.*

The establishment of the failure time model on multi-dimension makes the reliability and availability of petascale supercomputers to be precisely quantified. In order to evaluate the reliability of petascale supercomputers, a mathematical expectation of the TBF model is just computed. By quantification of the time between failures, the time of checkpoint on system or application level can be dynamically adjusted accordingly. This eliminates the unnecessary overhead from the unreasonable checkpoint interval. The resource allocation and job scheduling based on ever varying reliability can also be conducted to improve the reliability of the job running environment. At last, the quantitative model on the time between failures lays a theoretical and engineering basis for technologies of fine grained failure predication.

6 Conclusions

In the paper, we analyzed the root cause of failures during the Sunway series supercomputers' running and obtained the fault characteristics of petascale supercomputers represented by Sunway BlueLight and Sunway TaihuLight in details. The statistic analysis showed that the mainframe is the major source of failures on petascale supercomputers. Our experiment showed that DRAM SBEs may have no correlation with jobs running and have direct correlation with the reliability of CPU nodes or DRAMs. The lower the reliability of CPU nodes or DRAMs is, the higher the probability of DRAM SBEs is. We further gave an analysis about whether there is a correlation between DRAM SBEs and DRAM MBEs. We found that there is no obvious causal relationship between DRAM SBEs and DRAM MBEs. The DRAM SBEs do not necessarily cause a DRAM MBE, and a DRAM MBE is often followed by a DRAM MBE. We also analyzed the failure time on the two supercomputers and found that the Weibull distribution best fits the empirical time between failures on multi-dimension. The model quantifies the time when system failures occur.

Our research lays a basis for the new generation of resilience technologies. As our future plan, we are developing a hybrid model of failure prediction and building an effective resilience technology based on the study for the next generation supercomputers.

References

[1] Cappello F. Resilience: One of the main challenges for exascale computing. Technical Report of the INRIA-Illinois Joint Laboratory, 2011.

[2] Kusnezov D, Binkley s, Harrod B, Meisner B. DOE exascale initiative. Technical Report of US Department of Energy (DOE), 2013. https://energy.gov/downloads/doe-exascale-initiative, Dec. 2017.

[3] Kogge P, Bergman K, Borkar S *et al.* Exascale computing study: Technology challenges in achieving exascale systems. 2008. http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf, Dec. 2017.

[4] Schroeder B, Gibson G A. A large-scale study of failures in high-performance computing systems. *IEEE Transactions on Dependable and Secure Computing*, 2010 7(4): 337-350

[5] Liang Y, Zhang Y, Jette M, Sivasubramaniam A, Sahoo R. BlueGene/L failure analysis and prediction models. In *Proc. the 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks* (*DSN*), June 2006, pp.425-434.

[6] Zheng Z, Lan Z, Park B H *et al.* System log pre-processing to improve failure prediction. In *Proc. IEEE/IFIP International Conference Dependable Systems and Networks*, June 29-July 2, 2009.

[7] Zheng Z, Yu L, Tang W *et al.* Co-analysis of RAS log and job log on Blue Gene/P. In *Proc. the 2011 IEEE International Parallel & Distributed Processing Symposium*, May 2011 pp.840-851.

[8] Zheng Z, Lan Z. Reliability-aware scalability models for high performance computing. In *Proc. IEEE International Conference Cluster Computing and Workshops*, Aug. 31-Sept. 4, 2009.

[9] Heien E, LaPine D, Kondo D *et al.* Modeling and tolerating heterogeneous failures in large parallel systems. In *Proc. the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov. 2011, Article No. 45.

[10] Nie B, Tiwari D, Gupta S *et al.* A large-scale study of soft-errors on GPUs in the field. In *Proc. the 2016 IEEE International Symposium on High Performance Computer Architecture* (*HPCA*), March 2016, pp.519-530.

[11] Schroeder B, Pinheiro E, Weber W. DRAM errors in the wild: A large-scale field study. In *Proc. the 11th International Joint Conference on Measurement and Modeling of Computer Systems*, June 2009, pp.193-204.

[12] Pinheiro E, Weber W, Barroso L A. Failure trends in a large disk drive population. In *Proc. the 5th USENIX Conference on File and Storage Technologies*, February 2007, pp.17-28.

[13] Gunawi H S, Hao M, Suminto R O *et al.* Why does the cloud stop computing?: Lessons from hundreds of service outages. In *Proc. the 7th ACM Symposium on Cloud Computing*, October 2016, pp.1-16.

[14] Gunawi H S, Hao M, Leesatapornwongsa T *et al.* What bugs live in the cloud? A study of 3000+ issues in cloud systems. In *Proc. the ACM Symposium on Cloud Computing*, November 2014, pp.1-14.

[15] Huang P, Guo C, Zhou L *et al.* Gray failure: The Achilles' heel of cloud-scale systems. In *Proc. the 16th Workshop on Hot Topics in Operating Systems*, May 2017, pp.150-155.

[16] Zheng Z, Lan Z, Gupta R *et al.* A practical failure prediction with location and lead time for Blue Gene/P. In *Proc. the 2010 International Conference Dependable Systems and Networks Workshops* (*DSN-W*), June 28-July 1, 2010.

[17] Sahoo R K, Oliner A J, Rish I *et al.* Critical event prediction for proactive management in large-scale computer clusters. In *Proc. the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2003, pp.426-435.

[18] Gu J, Zheng Z, Lan Z *et al.* Dynamic meta-learning for failure prediction in large-scale systems: A case study. In *Proc. the International Conference on Parallel Processing*, Sept. 2008.

[19] Gainaru A, Cappello F, Snir M *et al.* Fault prediction under the microscope: A closer look into HPC systems. In *Proc. the International Conference on High Performance Computing, Networking, Storage and Analysis*, November 2012, Article No. 77.

[20] Lu X, Wang H Q, Zhou R J *et al.* Autonomic failure prediction based on manifold learning for large-scale distributed systems. *The Journal of China Universities of Posts and Telecommunications*, 2010, 17(4): 116-124.

[21] Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In *Lecture Notes in Computer Science 1057*, Apers P, Bouzeghoub M, Gardarin G (eds.), June 2005.

[22] Mannila H, Toivonen H, Verkamo A I. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1997, 1(3): 259-289.

[23] Joshi M, Karypis G, Kumar V. A universal formulation of sequential patterns. Technical Report, No.99-021, University of Minnesota. https://www.cs.umn.edu/research/technical_reports/view/99-021, Dec. 2017.

[24] Fournier-Viger P, Wu C W, Tseng V S *et al.* Mining sequential rules common to several sequences with the window size constraint. In *Proc. the 25th Conference on Advances in Artificial Intelligence*, May 2012, pp.299-304.

[25] Fournier-Viger P, Wu C W, Tseng V S *et al.* Mining partially-ordered sequential rules common to multiple sequences. *IEEE Transactions on Knowledge and Data Engineering*, 27(8): 2203-2216.

[26] Zhang Z. Reliability Theory and Engineering Application. Beijing: Science Press, 2012. (in Chinese)

Rui-Tao Liu received his Bachelor's degree in computer science and technology from National University of Defense Technology (NUDT), Changsha, in 2000. He then received his Master's degree in computer software and theory from Jiangnan Institute of Computing Technology, Wuxi, in 2004. He is currently an engineer and Ph.D. candidate in State Key Laboratory of Mathematical Engineering and Advanced Computing (MEAC), Wuxi. His research interests include high-performance computing, parallel operating system, fault tolerance, big data, etc.

Zuo-Ning Chen received her Master's degree in computer application technology from Zhejiang University, Hangzhou, in 1999. She is the chief engineer of National Research Center of Parallel Computer Engineering and Technology, Beijing. She is an adjunct professor in computer science and technology, Tsinghua University, Beijing, and an academician of the Chinese Academy of Engineering. Her current research interests include big data computing, cloud computing, and high performance computing. She has made important contributions in the field of computer software and high-end computers and received the Special and First Prizes of the National Science and Technology Progress Award of China.