Long term failure analysis of 10 petascale supercomputer

F. Shoji[1], S. Matsui[2], M. Okamoto[3], F. Sueyasu[2],
T. Tsukamoto[1], A. Uno[1] and K. Yamamoto[1]

[1] Operations and Computer Technologies Division, RIKEN AICS
[2] Technical Computing Solutions Unit, Fujitsu Limited
[3] IT Infrastructure Business Group, Fujitsu Systems West Limited

AICS

FUJITSU

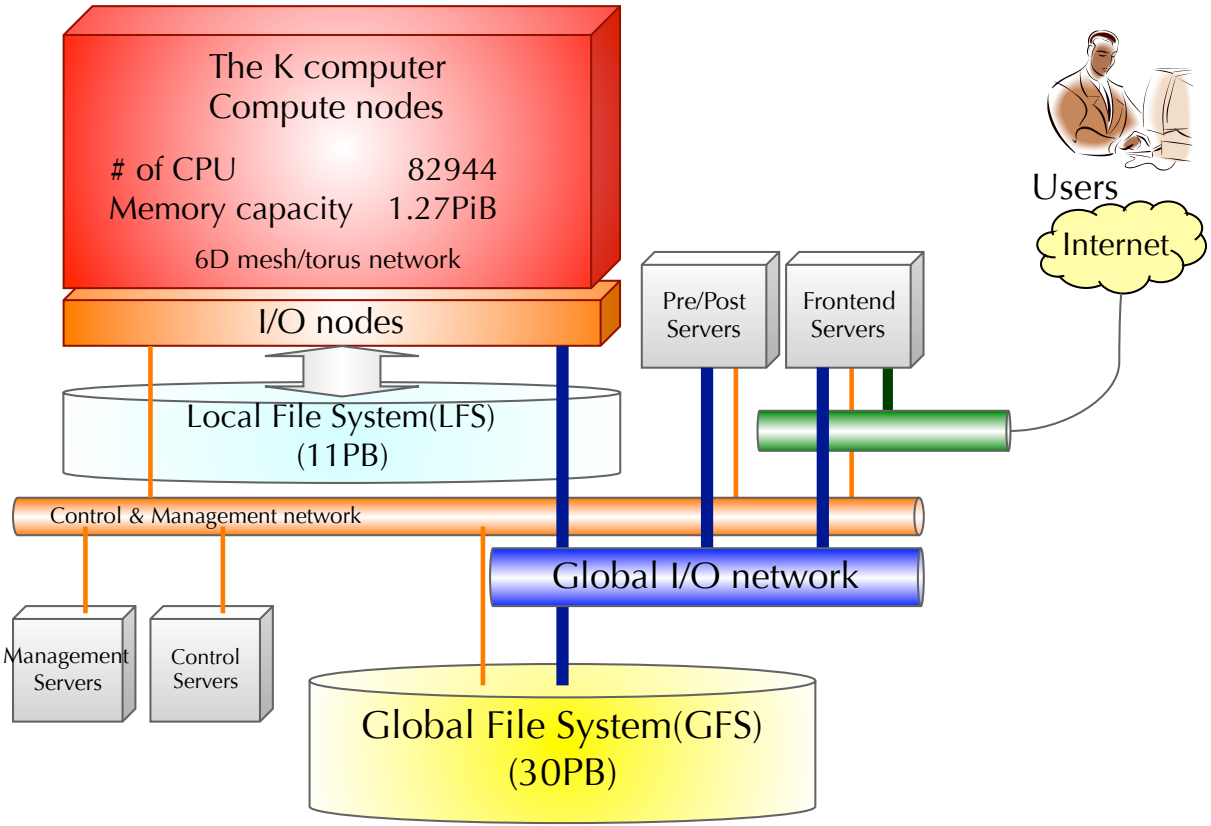
京
K computer

Objectives

- Analyzing failures on extremely large scale supercomputers, such as the K computer, is very important for the following reasons:
  - To optimize operation against failures and reduce downtime
  - To reveal and repair weaknesses in hardware and software
  - To clarify factors that require improvements for the development of the K computer's successor
  - To share operational experiences with other supercomputer centers and assist in developing best practices

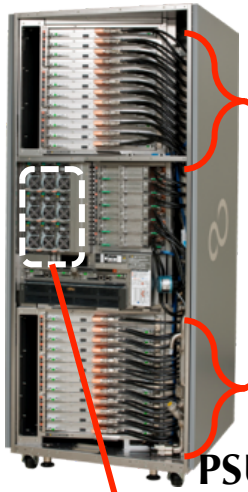
- However, this kind of researches and activities are not so much.

The K computer overview



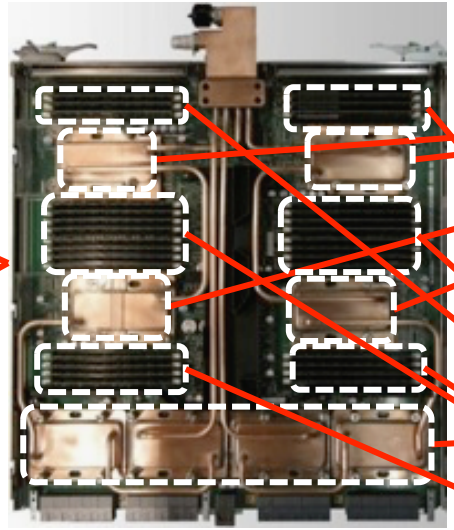
Number of major parts

Compute Rack
× 864

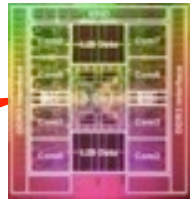


PSU
864 × 9 = 7,776

System Board
864 × 24 = 20,736



CPU
864 × (24 × 4) = 82,944



Inter Connect Controller
864 × (24 × 4)=82,944

DIMM
864 × (24 × 4 × 8) = 663,552

CPU/ICC are water-cooled(inlet:15℃ outlet:17℃)
Other components are air-cooled

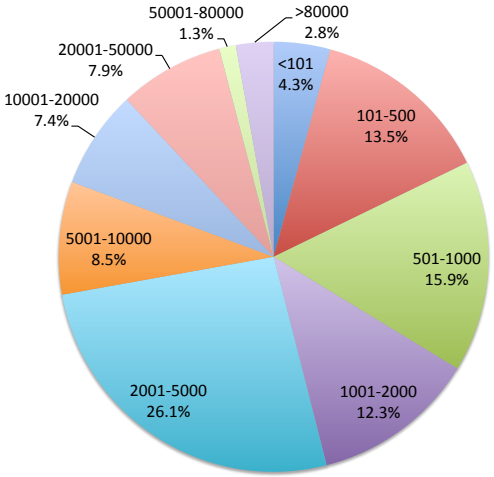
When a failure of CPU/ICC/System Board occurred then the system board will be replaced.
(For DIMM failure, the DIMM will be replaced.)

Usage status

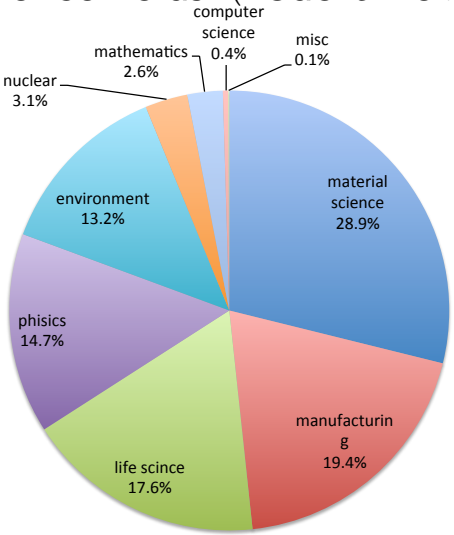
2012/09/28 - 2015/06/30

- Registered subjects/users : **~150/1200**
- Average number of executed jobs : **1193.2/day**
- Average number of active users : **109.5/day**

Job size (node*time based)



Science fields (node*time based)



Failure analysis on K computer

- K computer consists of extremely many parts and components.
- K computer always works with high load and is used by various types of jobs and users.
- Failure events are expected to occur more frequently than the others.

Failure statistics of K computer should be useful for failure analysis of supercomputer.

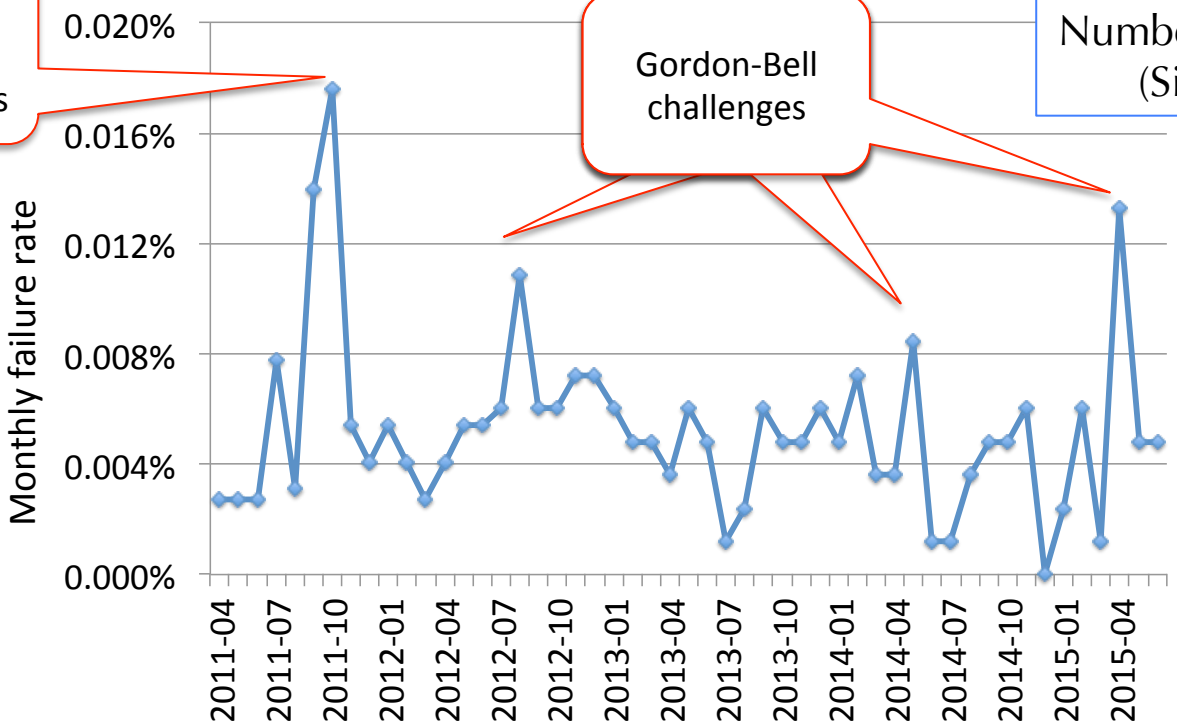
Failure trends

Monthly Failure Rate of CPUs

Full node LINPACK measurements

Gordon-Bell challenges

Number of CPUs=**82,944** (Since July 2012)

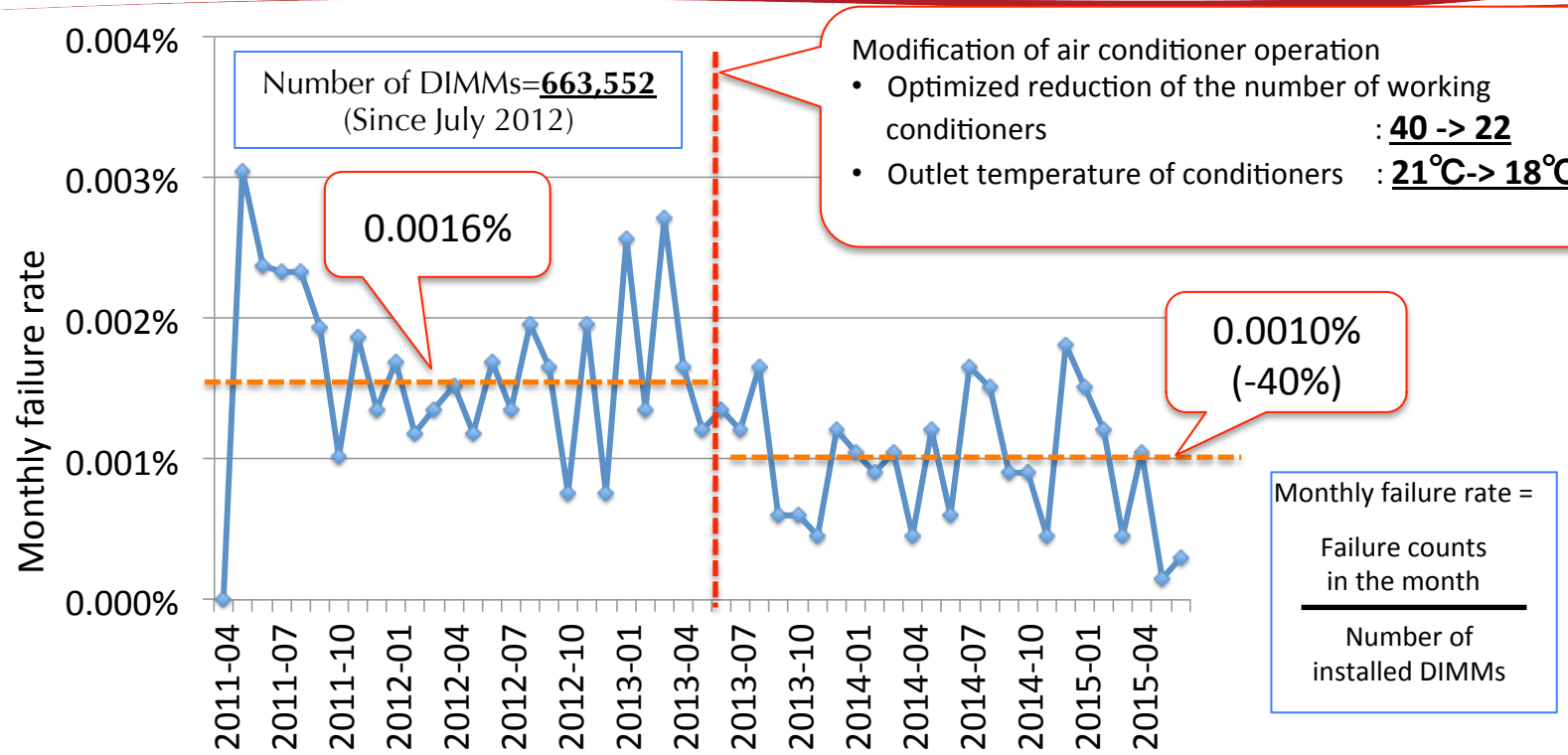


Monthly failure rate =

$$\frac{\text{Failure counts in the month}}{\text{Number of installed CPUs}}$$

Failure trend of CPUs is almost stable except high load terms

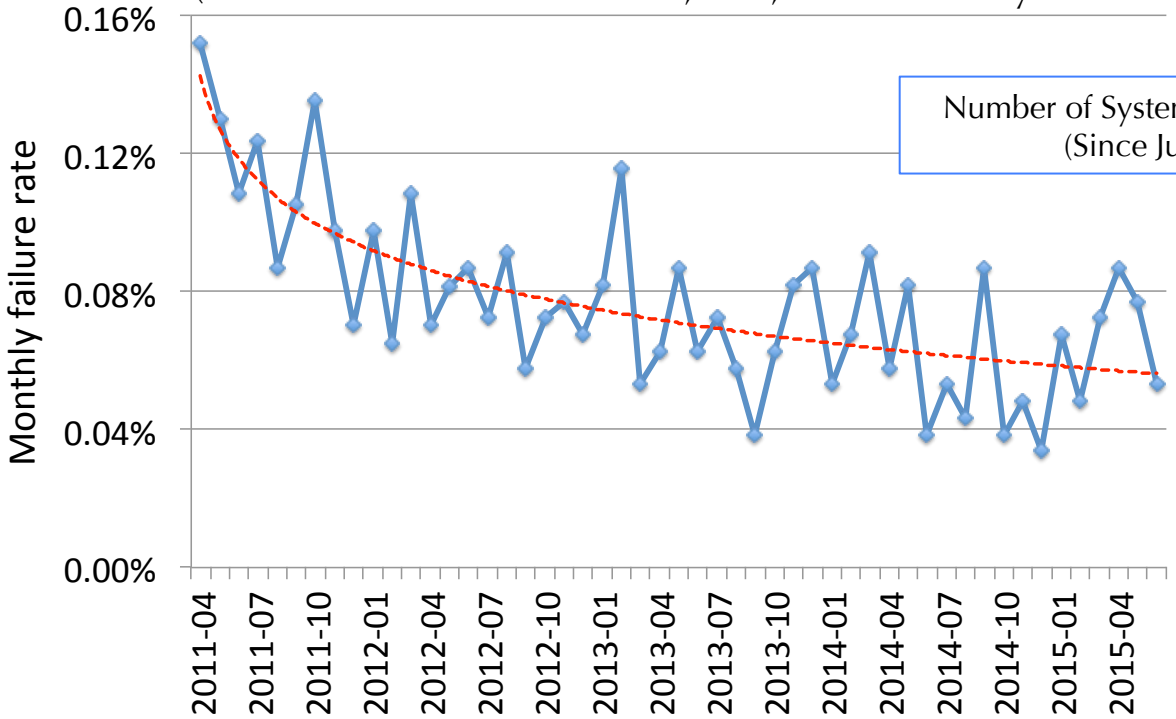
Monthly Failure Rate of DIMMs



Failure trend of DIMMs was changed to be lower at the modification of air conditioner operation in July 2013

Monthly Failure Rate of System Boards

(includes the failures of CPU, ICC, DIMM and System Board itself)



Number of System Boards=**20,736**
(Since July 2012)

Monthly failure rate =

Failure counts in the month

Number of installed System Boards

Failure rate of system boards seems to reach to the plateau

Failure rates

Comparison with Blue Waters

AFR : Annual Failure Rate (Average failure rate per year)
FIT : Failure In Time (1FIT = 1 failure per 10^9 hours)

	K computer (April 2011 – June 2015)				Blue Waters(*)			
	Number of parts	AFR	FIT	FIT/GB	Number of parts	AFR	FIT	FIT/GB
CPU	82,944	0.06%	72.00	N/A	49,258	0.23%	265.15	N/A
DIMM	663,552	0.016%	18.02	9.01	197,032	0.112%	127.84	15.98

(*) C. Di Martino et al., Lessons learned from the analysis of system failures at petascale: the case of blue waters. 44th international conference on Dependable Systems and Networks (DSN 2014), 2014.

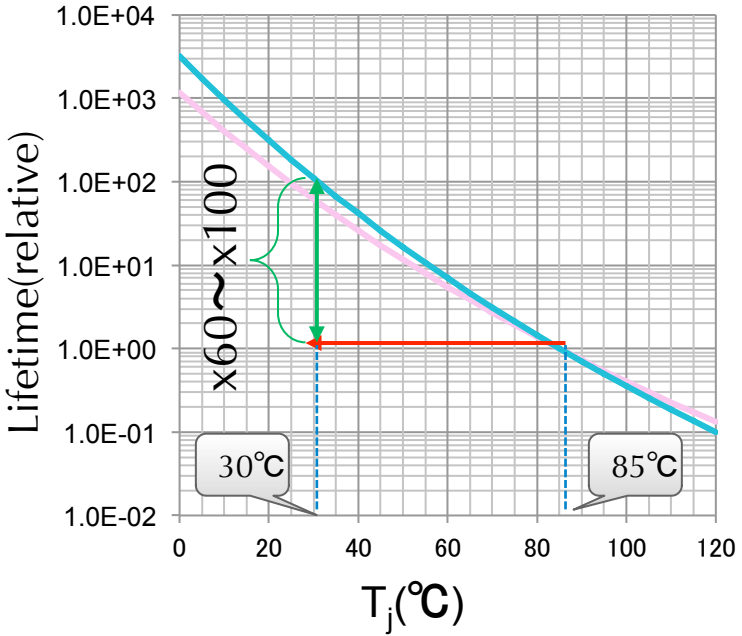
- CPU failure rates of the K computer are about quarter compared to that of Blue Waters.
- For DIMM, FIT/GB is about half.

Consideration of low failure rates

Arrhenius's law

$$k = Ae^{-\frac{E_a}{k_B T}}$$

k : chemical reaction rate constant
A : constat
E_a : activation energy
k_B : Boltzmann constant
T : temperature



According to our early estimation, if junction temperature T_j of CPU could be decreased from 85℃ to 30℃ then

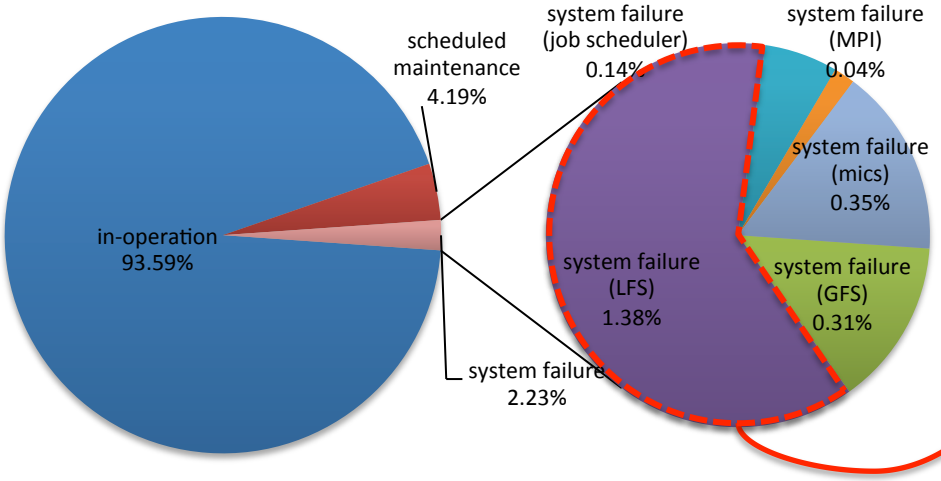
relative life time will be longer x60～x100

Low T_j seems to contribute to lower failure rates

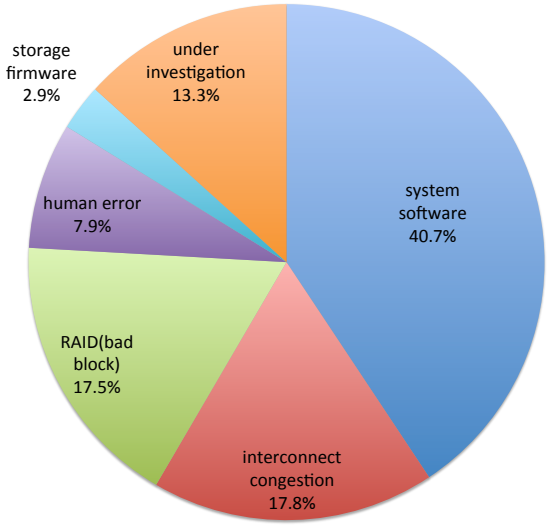
Whole system failures

System availability

system availability (September 2012 – March 2015)



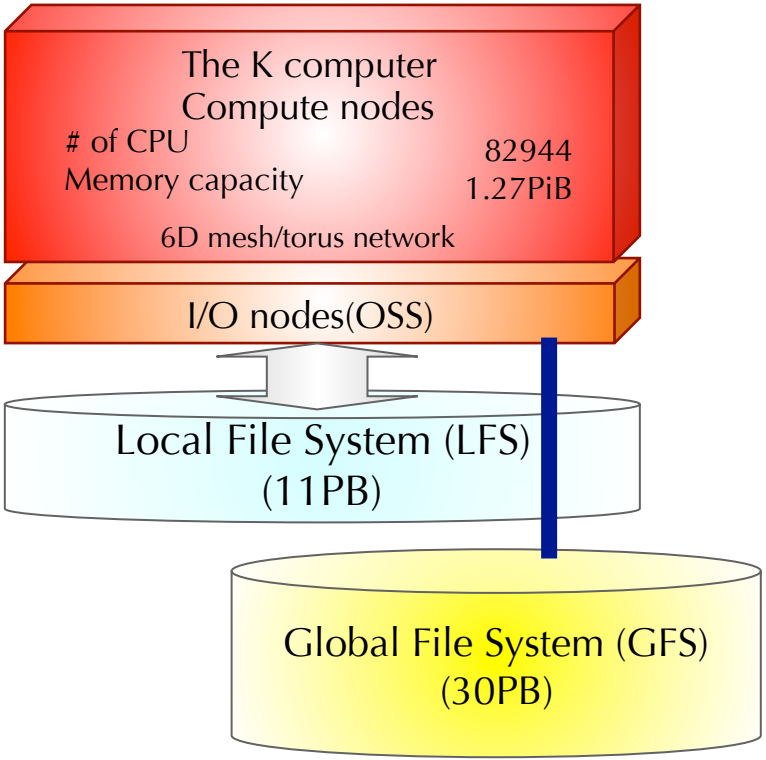
system failure(LFS)



- System availability achieved more than 93%
- More than 60% of system failure time was due to local file system(LFS) failures.

- System software bugs(40.7%)
- MDS/OSS down due to interconnect congestion(17.8%)
- Partial RAID system failure (17.5%)
- …

Consideration of LFS failures



Design concept from user requirements:

- LFS consists of many OSSes and OSTs to realize higher bandwidth.
  - OSS: 2592, OST:5184 (GFS OSS:90, OST:2880)
- LFS is configured as one volume to provide a shared area.

Results:

- Larger number of OSSes and OSTs revealed the many potential bugs in the system software and many severe failures were caused by such bugs.
- LFS down means all service stop, because it is a single failure point.

Lessons learned:

- Do not configure a file system with larger number of OSSes and OSTs to avoid potential bugs.
- Do not make one huge volume to avoid a single point failures.

Summary and Outlook

- On analyzing the failures occurred on the K computer, we found the followings:
  1. Failure trend of CPUs is almost stable except high load terms.
  2. Failure trend of DIMMs was changed to be lower at the modification of air conditioner operation in July 2013
  3. CPU and DIMM failure rates of the K computer are about quarter and half compared to those of Blue Waters, respectively.
     1. Low Tj seems to contribute to lower failure rates.
  4. System availability achieved more than 93%, and more than 60% of system failure time was due to local file system(LFS) failures.
     1. **Do not configure a file system with larger number of OSSes and OSTs to avoid potential bugs.**
     2. **Do not make one huge volume to avoid a single point failures.**

- A detailed analysis of relations between the failures and the factors such as accumulated job processing time, temperature is now in progress.

Acknowledgements

- Fujitsu system engineers and customer engineers who work for the operation support and maintenance of K computer.
- RIKEN AICS Operations and computer technologies division members who work for the operation and enhancement of K computer.
- K computer development project team members of Fujitsu and RIKEN.