Long term failure analysis of 10 petascale supercomputer

Fumiyoshi Shoji[1], Shuji Matsui[2], Mitsuo Okamoto[3], Fumichika Sueyasu[2],
Toshiyuki Tsukamoto[1], Atsuya Uno[1] and Keiji 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

Computer simulations create the future

Abstract

This report presents our analysis of the failures of the K computer at RIKEN AICS. We analyzed the failure rates of the major parts and compared some of the results with the case of Blue Waters. Our analysis showed that CPU and DIMM failure rates of the K computer are much lower than those of Blue Waters.

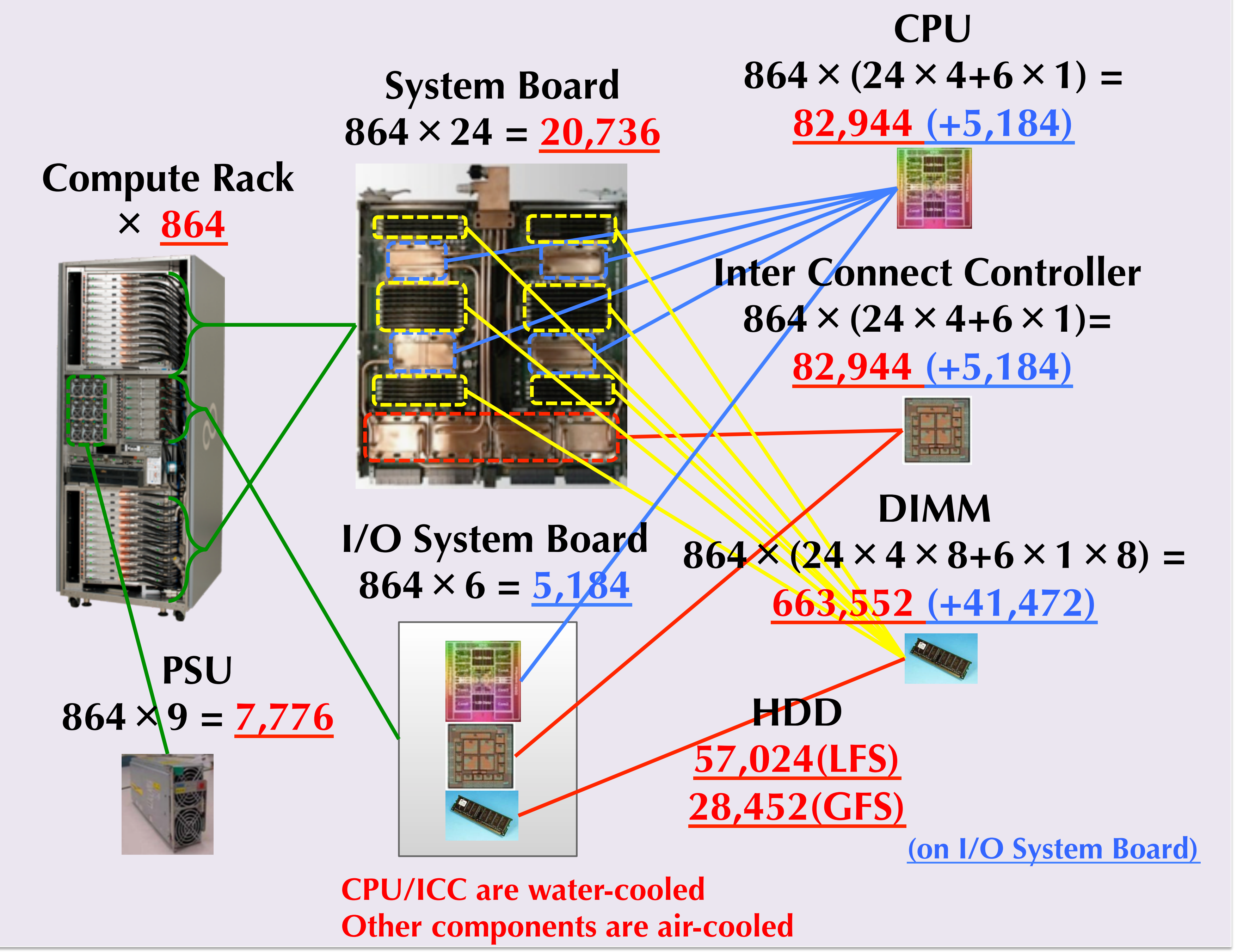
We also evaluated system availability and analyzed the causes of system failures. The K computer achieved more than 93% availability with approximately 4% of its downtime because of scheduled maintenance. System failures only accounted for 2.23% of the downtime, and most of these were caused not by node downs but by file system failures.

1.Objective

Analyzing failures on extremely large scale supercomputers, such as the K computer, are useful 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 improvement for the development of the K computer's successors
- To share operational experience with other supercomputer centers and assist in developing best practices

2.Number of major parts



Compute Rack × 864

System Board 864 × 24 = 20,736

CPU 864 × (24 × 4+6 × 1) = 82,944 (+5,184)

Inter Connect Controller 864 × (24 × 4+6 × 1)= 82,944 (+5,184)

DIMM 864 × (24 × 4 × 8+6 × 1 × 8) = 663,552 (+41,472)

I/O System Board 864 × 6 = 5,184

PSU 864 × 9 = 7,776

HDD 57,024(LFS) 28,452(GFS)

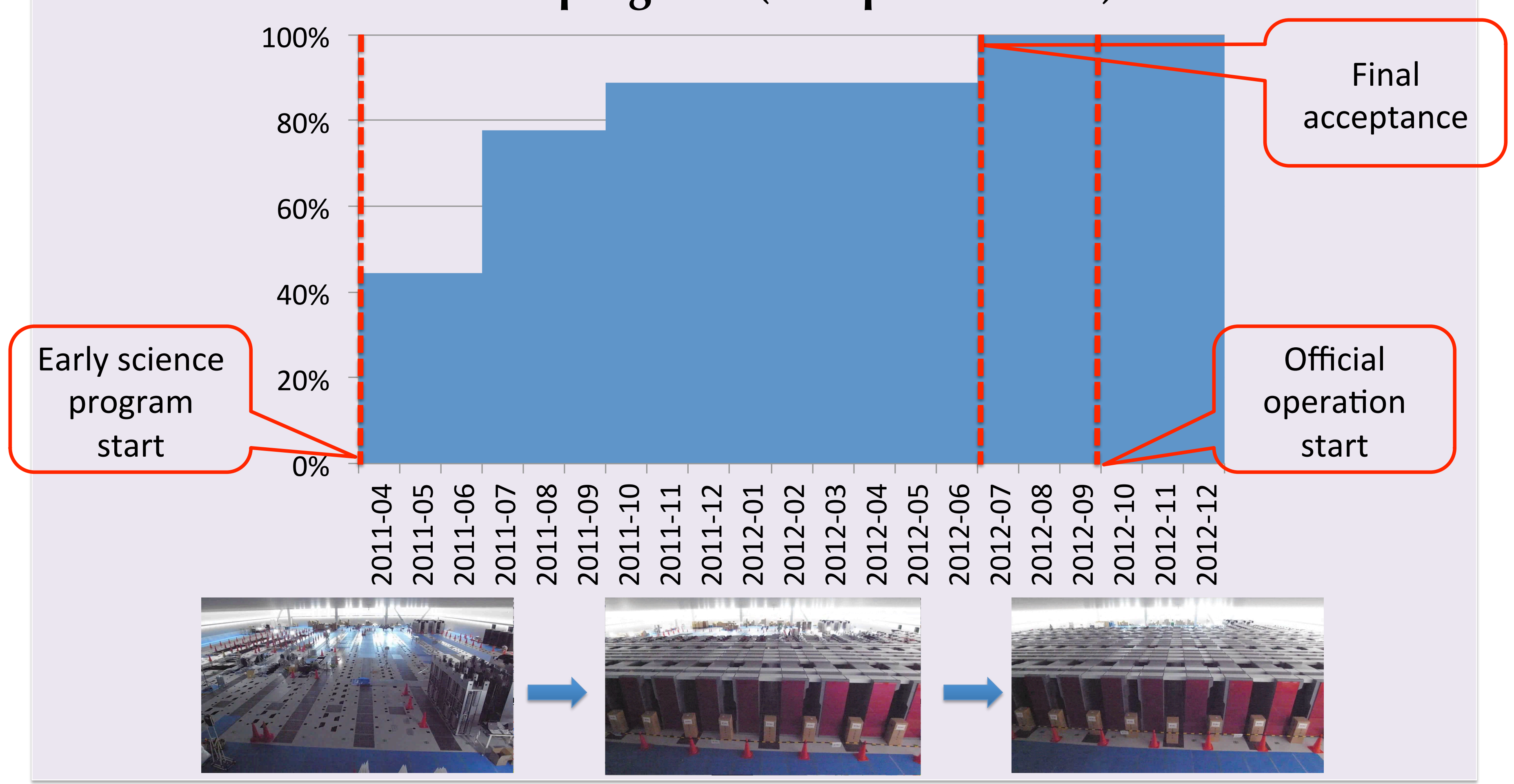
(on I/O System Board)

CPU/ICC are water-cooled
Other components are air-cooled

3.Installation&Operation

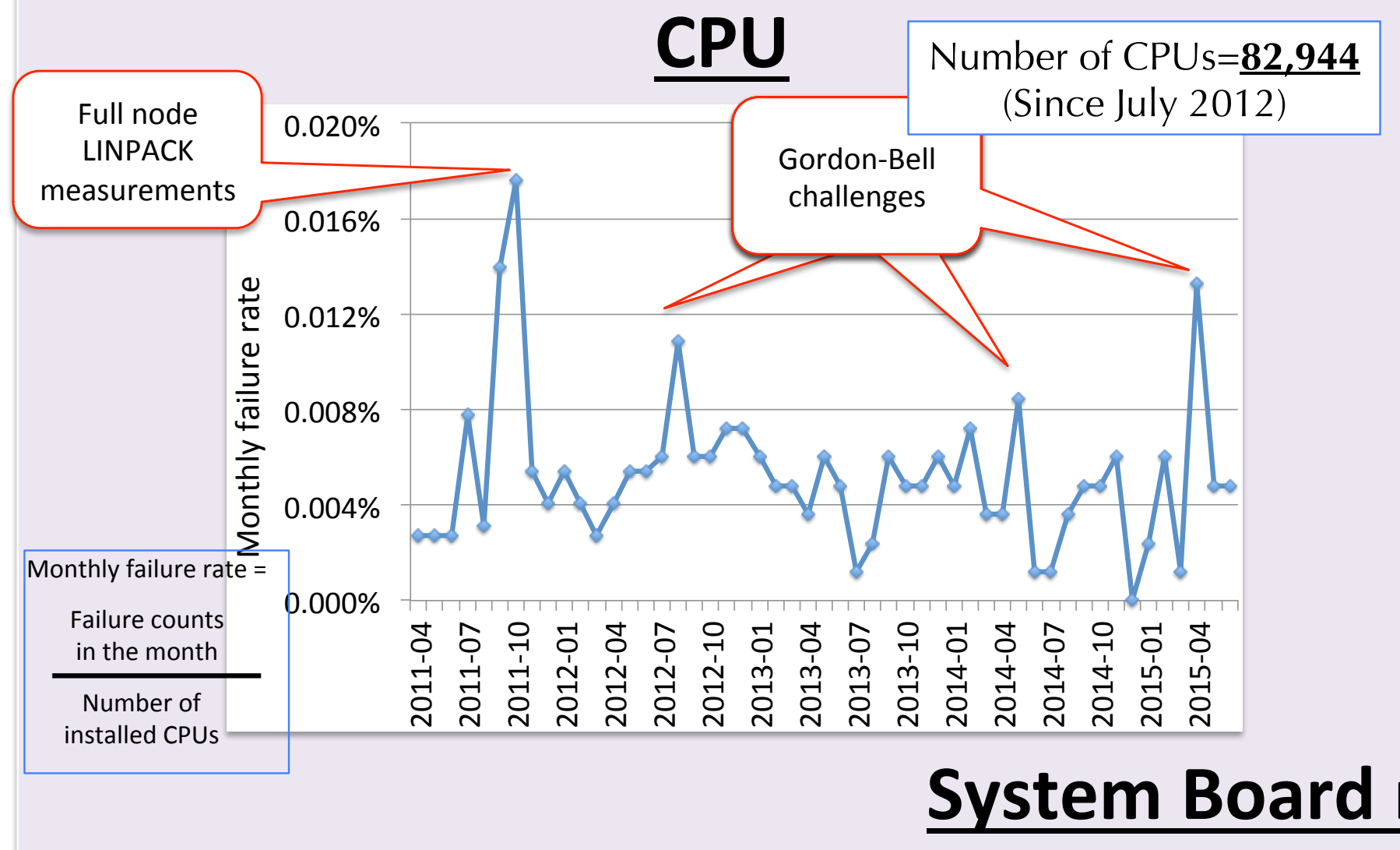
- Installation of the K computer began in August 2010 and has continued in a series of steps.
- The Early Science Program for limited users was started in April 2011.

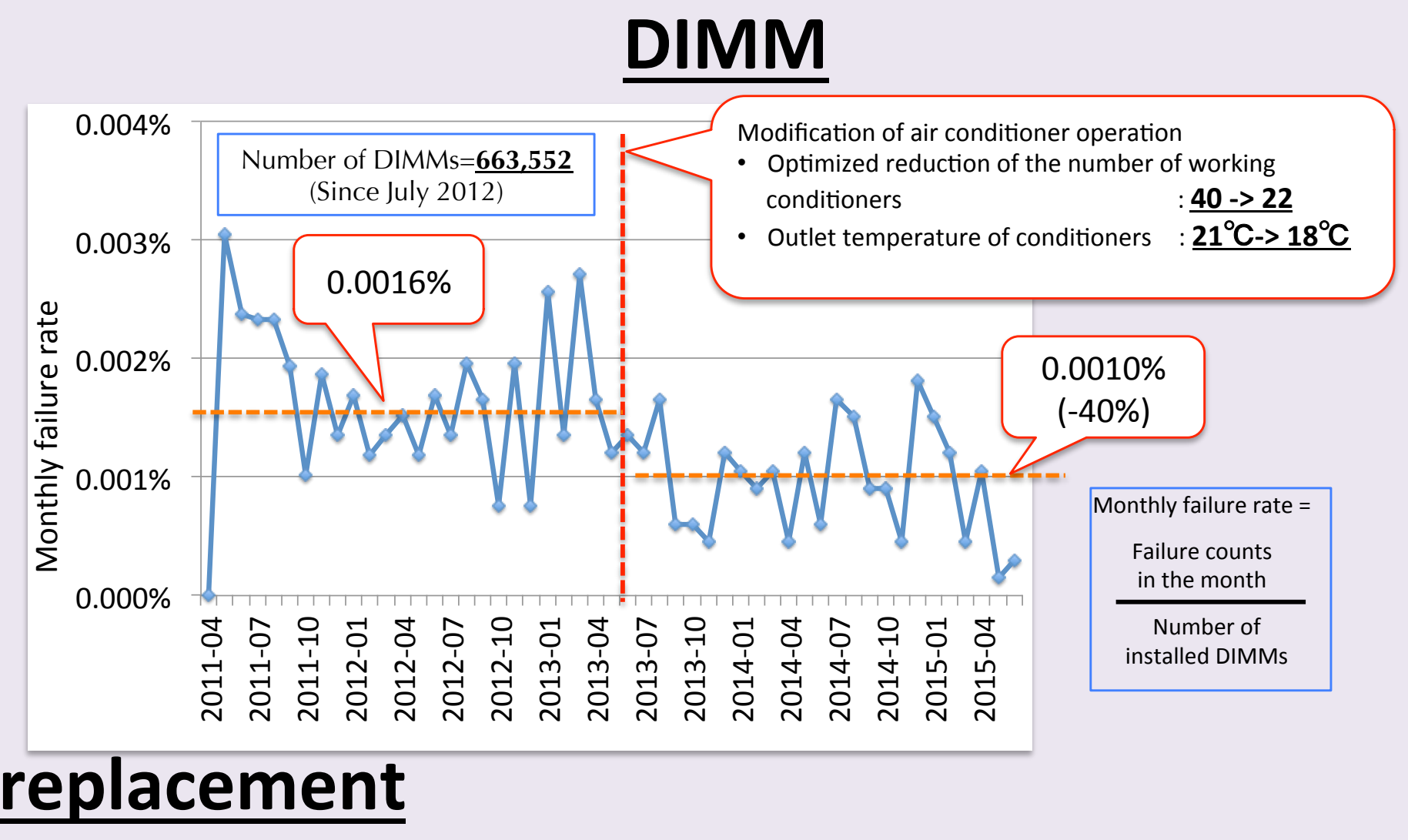
Installation progress (acceptance base)



4.Results

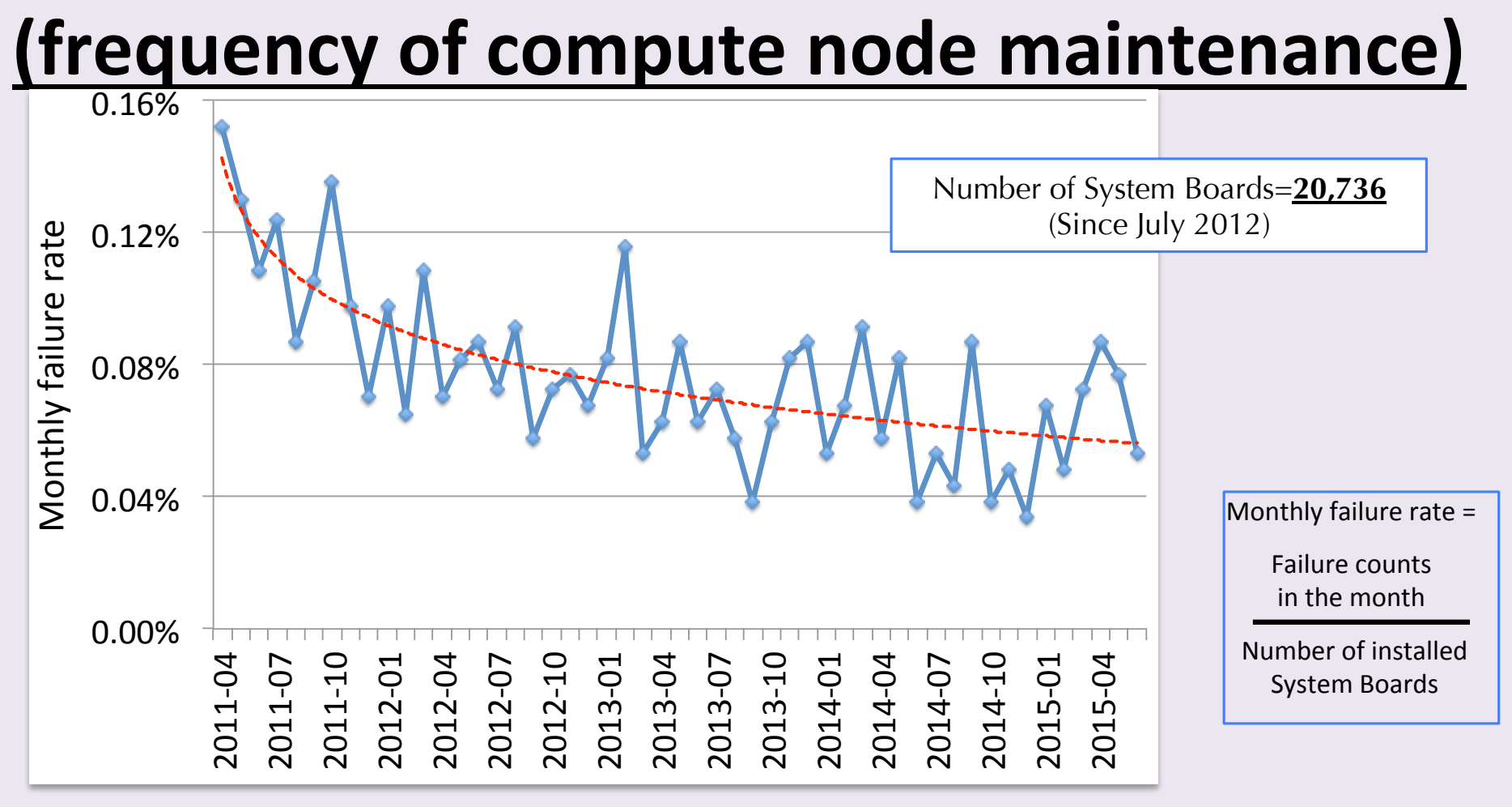
Trends in failures of major parts





System Board replacement

(frequency of compute node maintenance)



- Failure trend of CPUs is almost stable except high load terms.
- Failure trend of DIMMs was changed to be lower at the modification of air conditioner operation in July 2013.
- Failure rate of system boards seems to reach to the plateau.

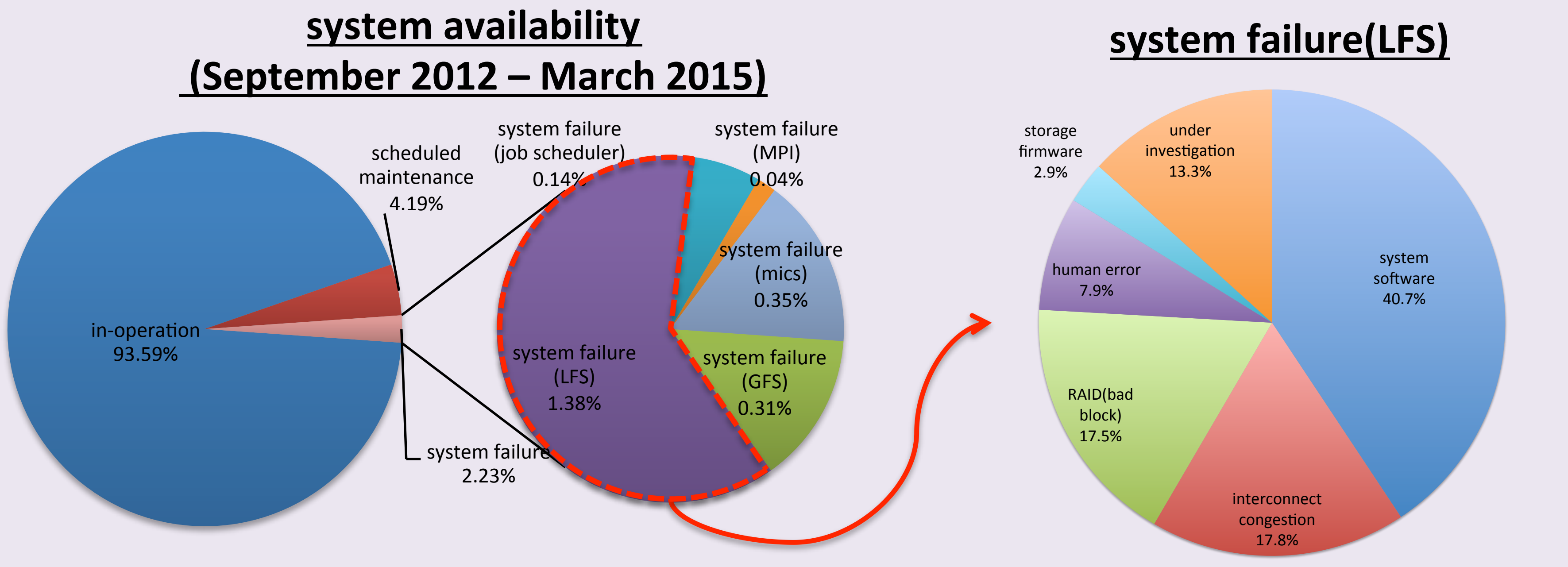
Failure rates

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[2]			
	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

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

System availability



- System availability of over 93% has been achieved since September 2012.
- Approximately 60% of system failure time was due to local file system failures.
  - The failure time consists of system software bugs(40.7%), MDS/OSS down due to interconnect congestion(17.8%), Partial RAID failure for some blocks(17.5%), human errors(7.9%), etc.

5.Summary&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.
  4. System availability achieved more than 93%, and more than 60% of system failure time was due to local file system(LFS) failures.
- A detailed analysis of relations between the failures and the factors such as accumulated job processing time, temperature is now in progress.

References

[1] K. Yamamoto et.al., The K computer operations: Experiences and statistics. International Conference on Computational Science (ICCS2014), 2014.
[2] 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.

RIKEN Advanced Institute for Computational Science (AICS)