Operation of the K computer and the facilities

F.Shoji, A.Kuroda, K.Minami, T.Tsukamoto, A.Uno and K.Yamamoto

Operations and Computer Technologies Division, RIKEN AICS

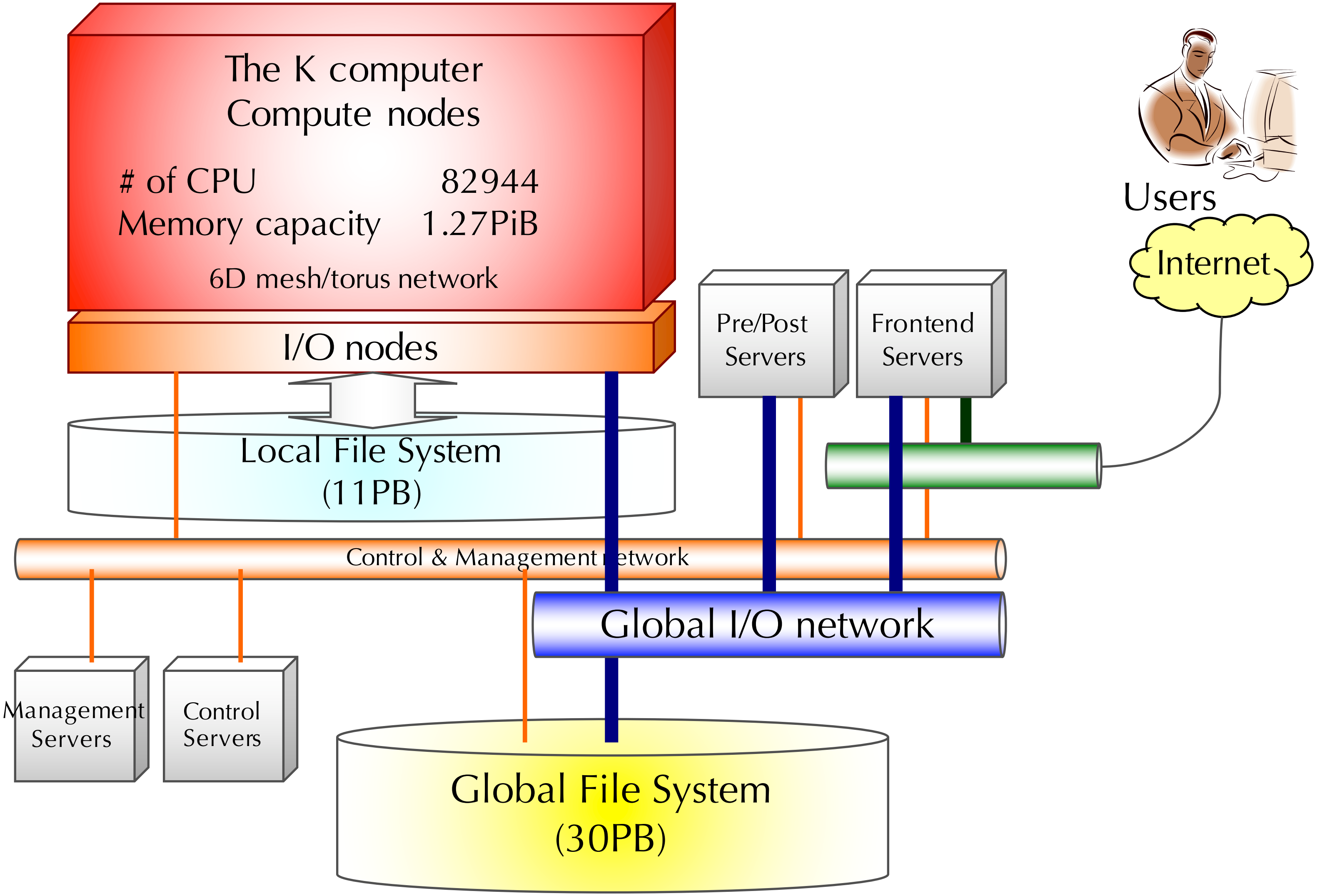
Outline

- The K computer and operation status
- Failure analysis
- The facilities and energy efficiency
- Summary

Outline

- The K computer and operation status
- Failure analysis
- The facilities and energy efficiency
- Summary

The K computer overview

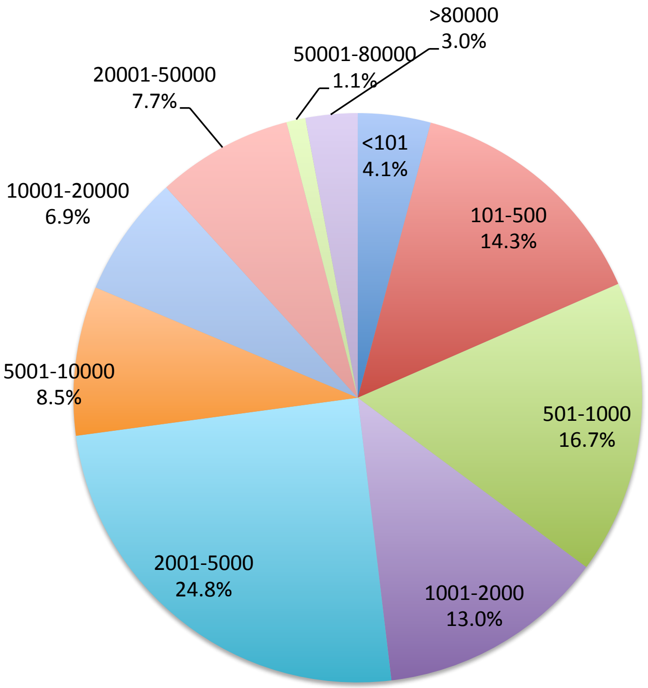


Users/Jobs on K computer

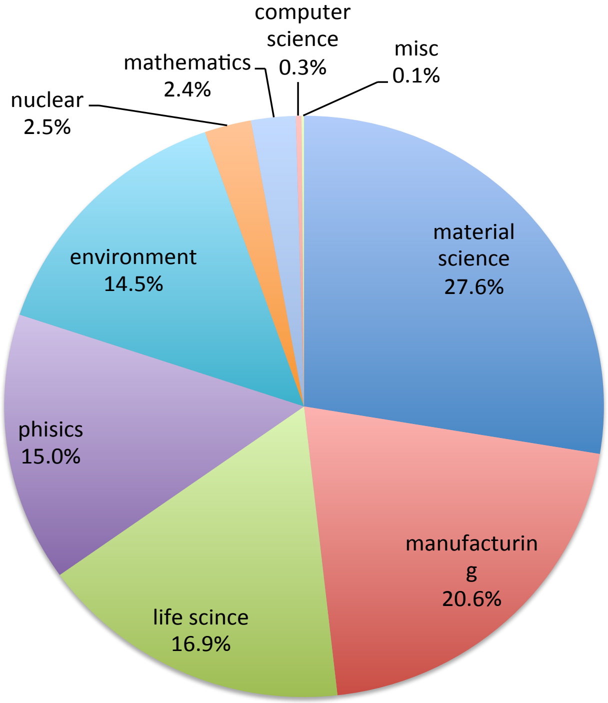
2012/09/28 - 2016/01/31

- Registered projects/users : **~150/1200 per FY**
- Average number of executed jobs : **1275.0/day**
- Average number of active users : **113.4/day**

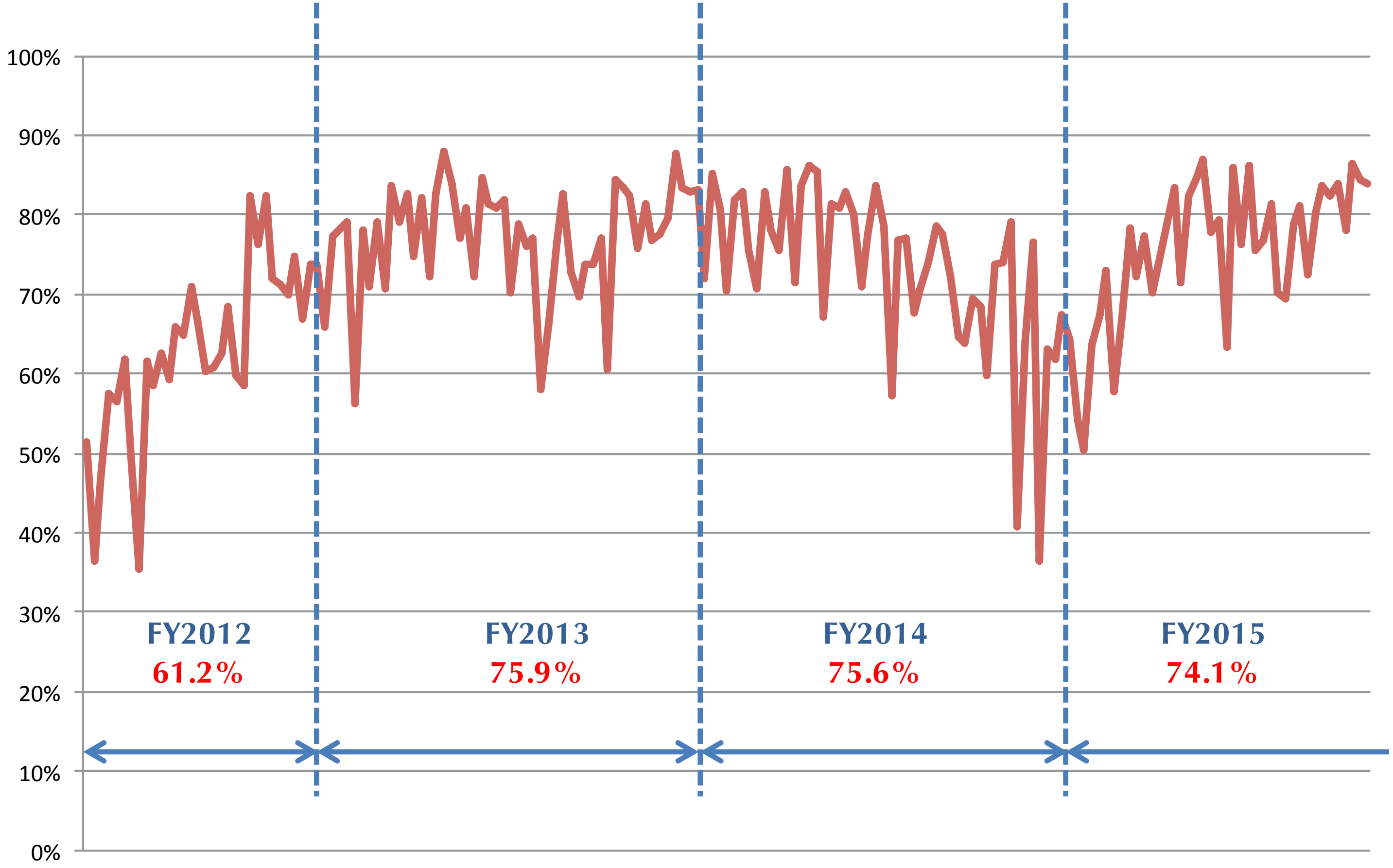
Job scale (node*time based)



Science fields (node*time based)

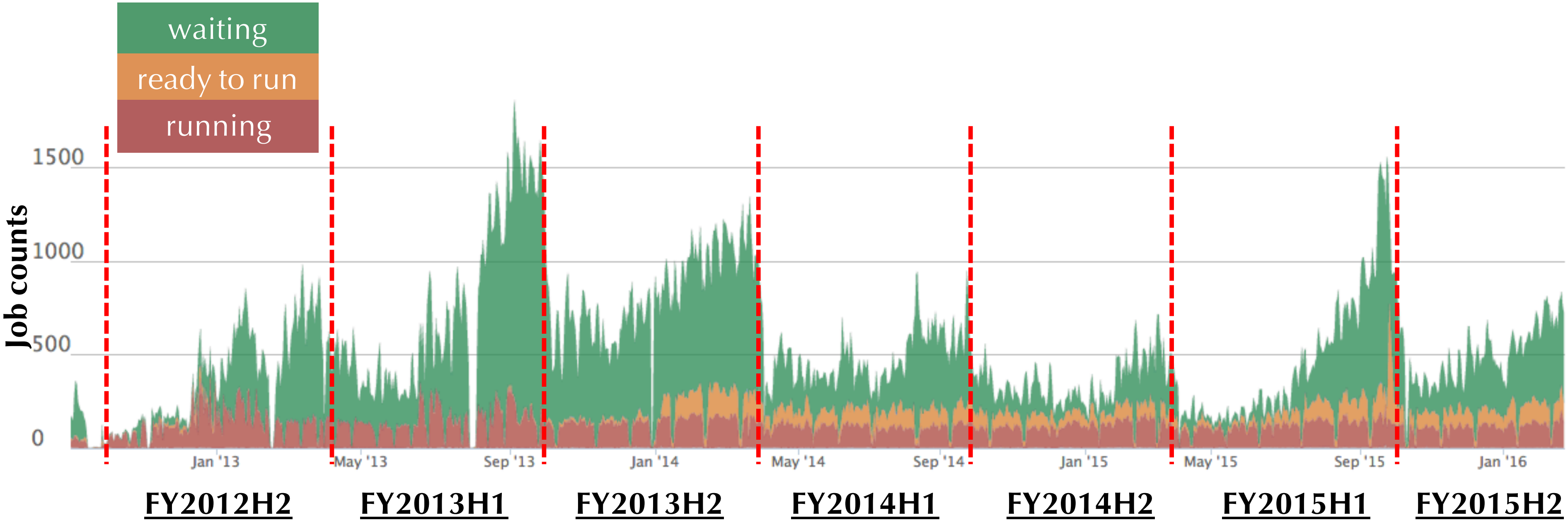


Usage rate



- Average usage rates keep about 75% without FY2012.
- From FY2014Q4 to FY2015Q1, usage rate decreased.
    - At FY2014Q4, some projects spent their node hours completely.
    - At the beginning of FY2015, startup of usage was slower.

Waiting jobs



- Job congestions at the end of half are always happened.
- Job congestions at FY2013 was caused by too much overbooking of computing time.
- We changed resource allocation rate from 100% to 85% at FY2014 and 88% at FY2015.
  - Overbooking rate (resource allocation rate / usage rate):
    - 131.8%(FY2013) -> 112.4%(FY2014) -> 118.8%(FY2015)
- At the beginning of FY2015H1, startup of usage was slower than the other half. It cause the severe congestion at the end.
  - Some new policies to encourage quick startup are introduced.

Outline

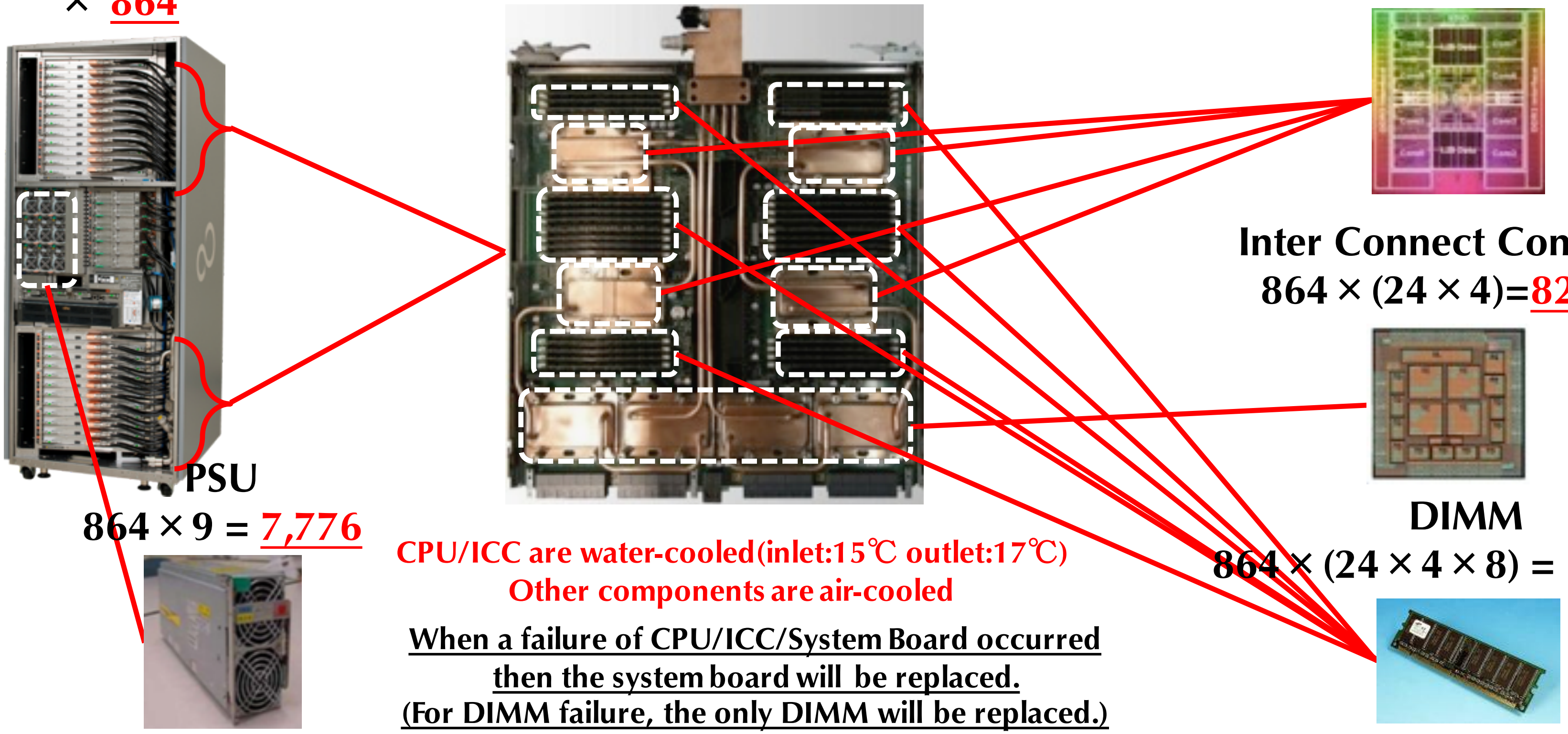
- The K computer and operation status
- Failure analysis
- The facilities and energy efficiency
- Summary

Failure analysis of 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 include significant information and is expected to be useful for general failure analysis of supercomputer.

Number of major parts



Compute Rack
× 864

System Board
864 × 24 = 20,736

CPU
864 × (24 × 4) = 82,944

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

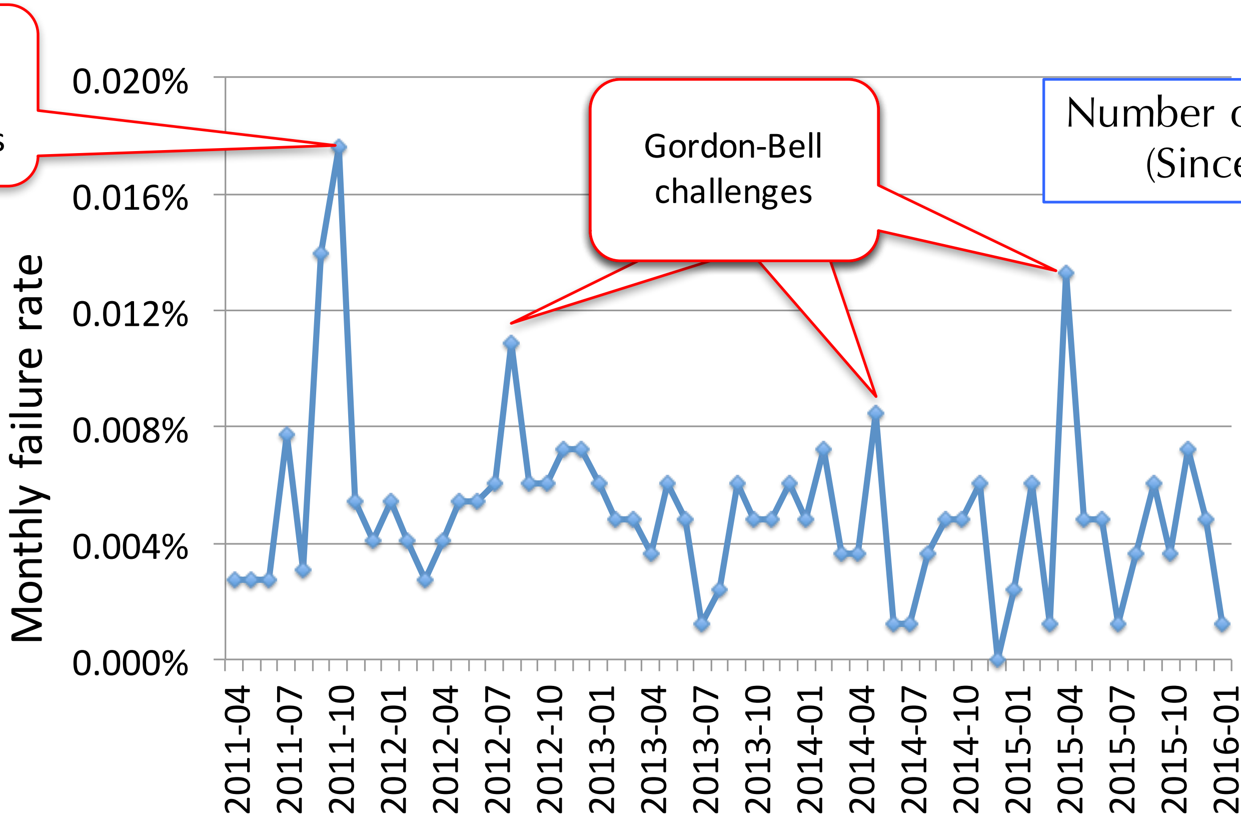
DIMM
864 × (24 × 4 × 8) = 663,552

PSU
864 × 9 = 7,776

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 only DIMM will be replaced.)

Monthly Failure Rate of CPUs



Full node LINPACK measurements

Gordon-Bell challenges

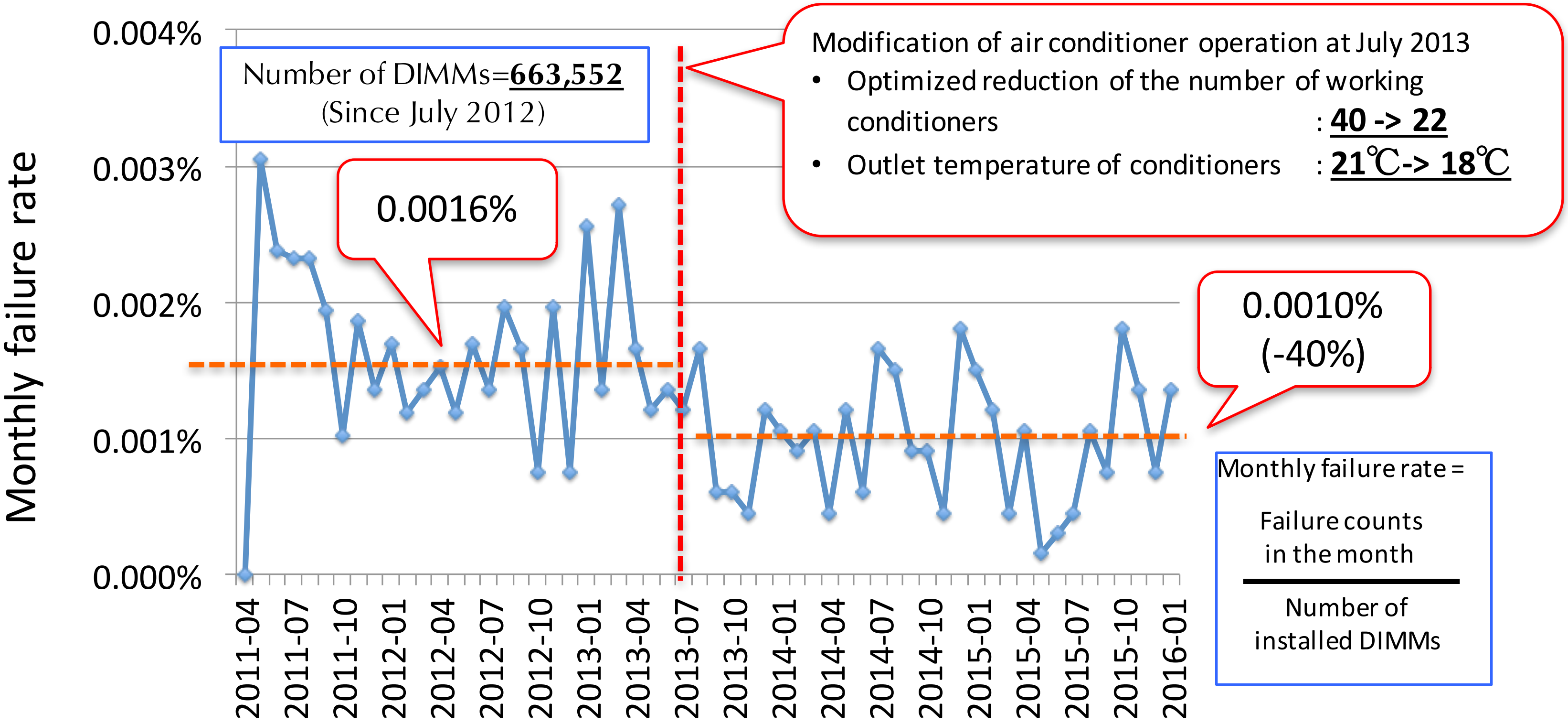
Number of CPUs=**82,944**
(Since July 2012)

Monthly failure rate =

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

Failure trend of CPUs is almost stable except after high load events

Monthly Failure Rate of DIMMs



Number of DIMMs=**663,552**
(Since July 2012)

Modification of air conditioner operation at July 2013

- Optimized reduction of the number of working conditioners : **40 -> 22**
- Outlet temperature of conditioners : **21℃-> 18℃**

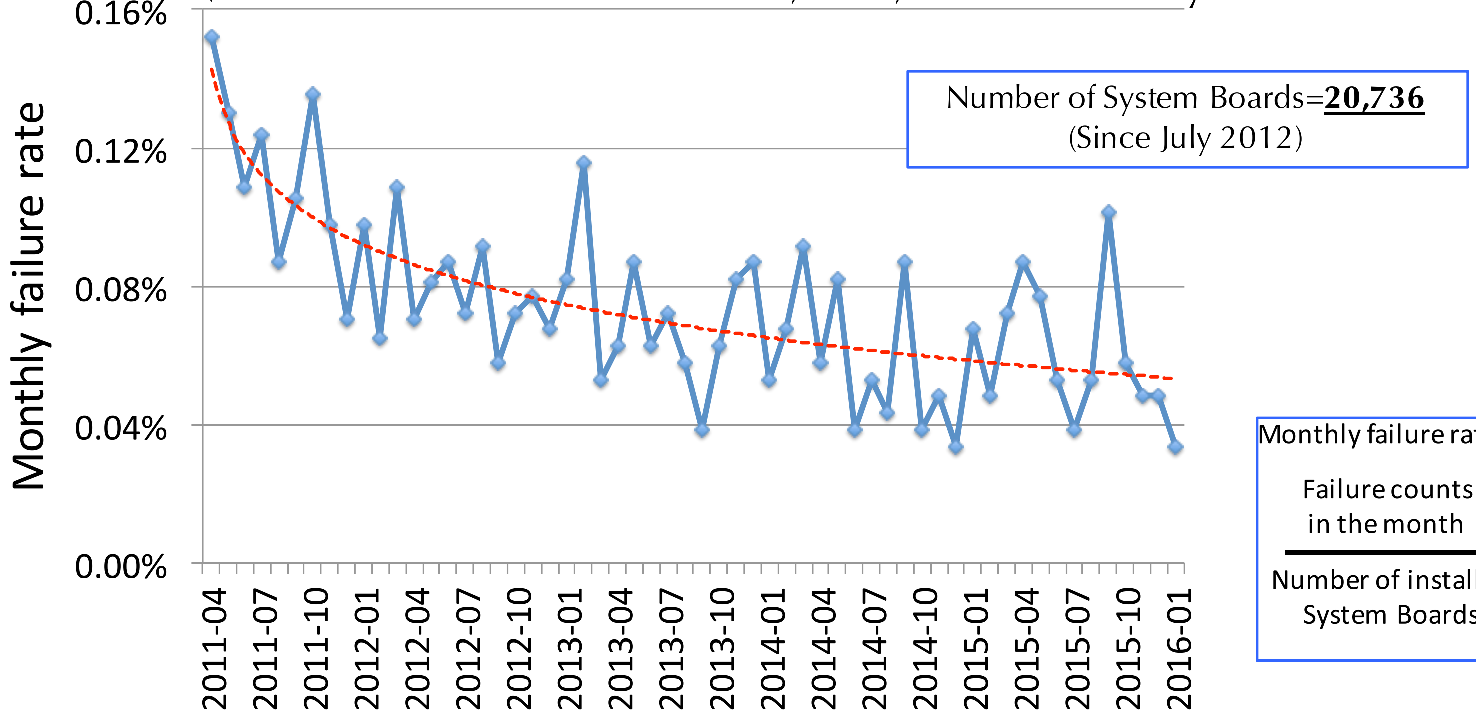
Monthly failure rate =

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

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 = $\dfrac{\text{Failure counts in the month}}{\text{Number of installed System Boards}}$

Failure rate of system boards seems to reach to the plateau
Average failure counts (= maintenance operation) ~ 14 / month

Comparison with Blue Waters

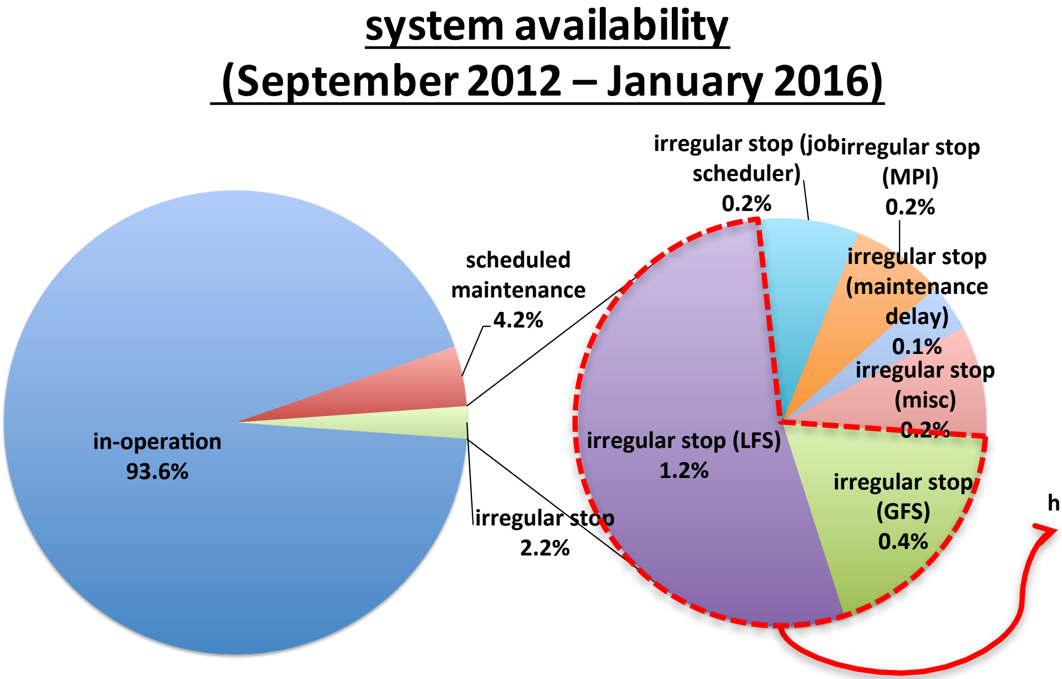
FIT : Failure In Time (1FIT = 1 failure per 10^9 hours)

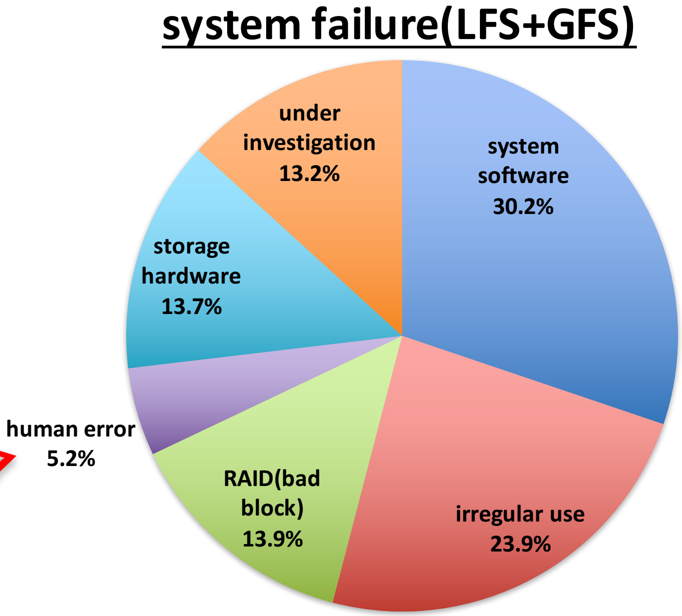
	K computer (April 2011 – January 2016)			Blue Waters(*)		
	Number of parts	FIT	FIT/GB	Number of parts	FIT	FIT/GB
CPU	82,944	69.86	N/A	49,258	265.15	N/A
DIMM	663,552	17.63	8.82	197,032	127.84	15.98

(*) 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 1/4 compared to Blue Waters.
- For DIMM, FIT/GB is about 1/2.

System availability





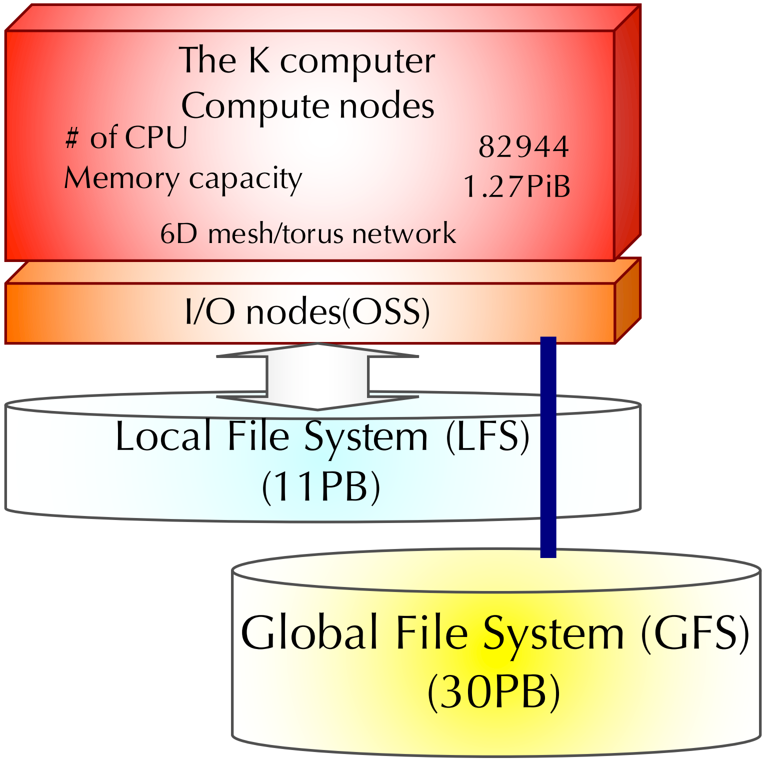
- System availability : **93.6%**
- Scheduled availability : **97.6%**
  - 91.0% (Blue Waters 2015[*])
- More than 60% of system failure time was due to file system(LFS and GFS) failures.

- System software bugs/invalid settings (30.2%)
- MDS/OSS down due to irregular use (23.9%)
- Partial RAID system failures (13.9%)

…

[*] The 2015 Blue Waters Annual Report book: https://bluewaters.ncsa.illinois.edu/documents/10157/27cb9800-01c1-49be-a7aa-a210ad14d21b

Consideration of LFS failures



Design concept for 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 huge 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.

MTBF/MTTR (Sep.2012-Jan.2016)

MTBF
(Mean Time Between (system wide) Failures)

$$= \frac{\text{(Real time)} - \text{(Maintenance time)} - \text{(Irregular stop time)}}{\text{(System wide irregular stop counts)}}$$

$$= \frac{27402.8 \text{ hours}}{62 \text{ times}} = 442.0 \text{ hours} = \underline{\textbf{18.4 days}}$$

11.2 days (Blue Waters 2015[*])

[*]The 2015 Blue Waters Annual Report book:
https://bluewaters.ncsa.illinois.edu/documents/10157/27cb9800-01c1-49be-a7aa-a210ad14d21b

MTTR
(Mean Time To Recovery)

= Average (System wide irregular stop time)

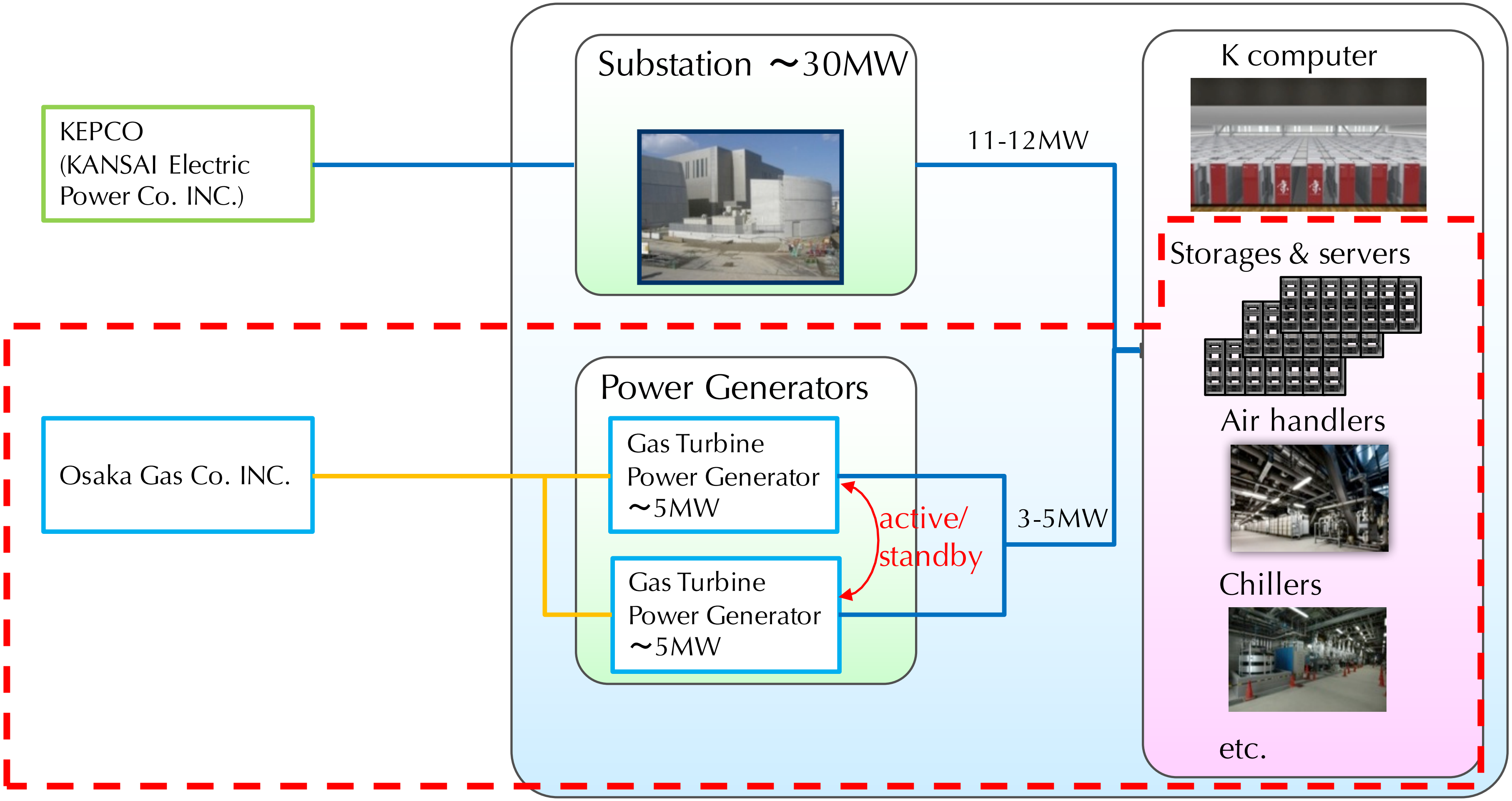
= **10.6 hours** (Max. 49.3 hours (October 2012))

Outline

- The K computer and operation status
- Failure analysis
- The facilities and energy efficiency
- Summary

Power supply

Total power consumption:14-16MW



KEPCO (KANSAI Electric Power Co. INC.)

Substation ～30MW

11-12MW

K computer

Storages & servers

Air handlers

Chillers

etc.

Osaka Gas Co. INC.

Power Generators

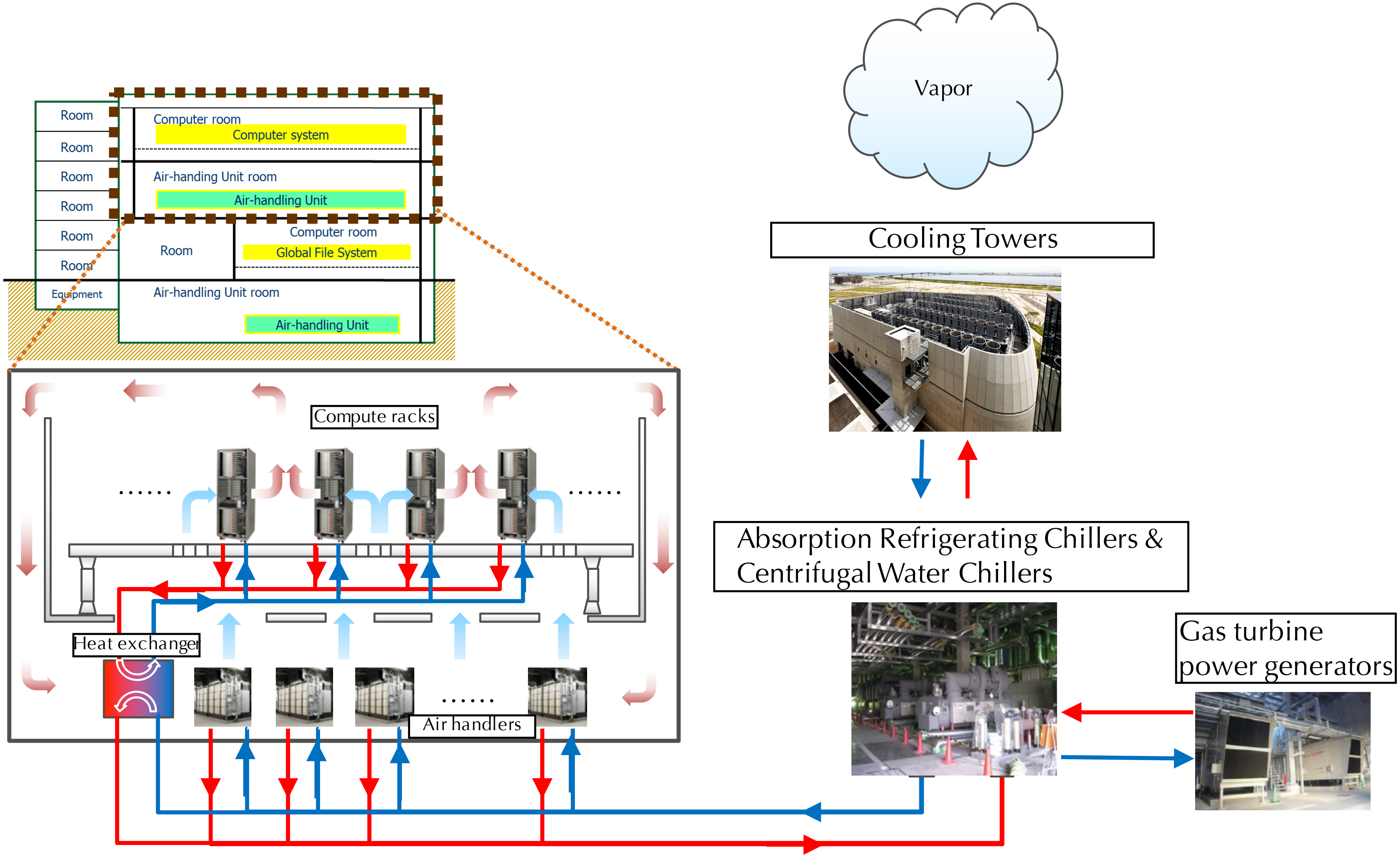
Gas Turbine Power Generator ～5MW

Gas Turbine Power Generator ～5MW

active/ standby

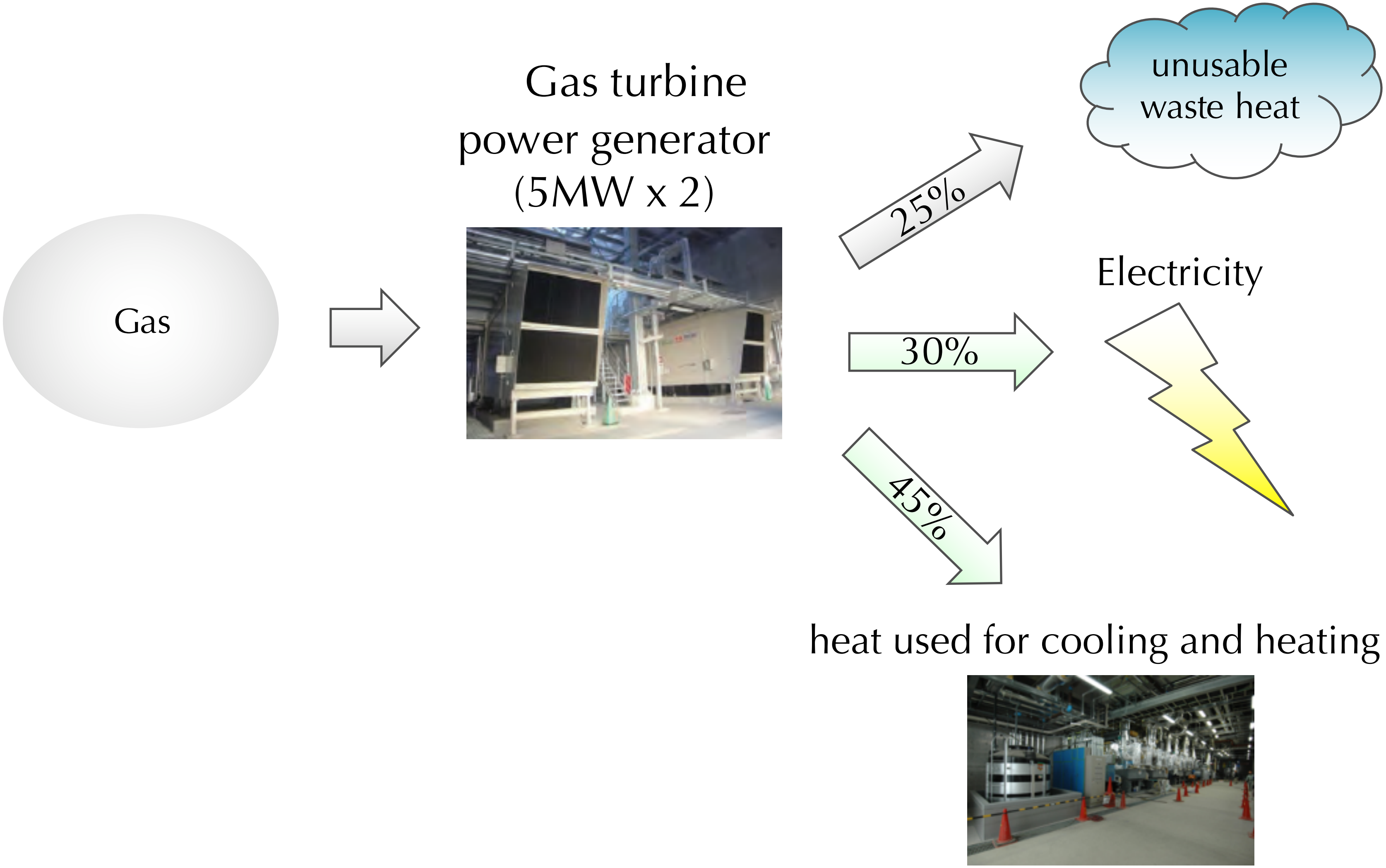
3-5MW

Cooling



Room
Room
Room
Room
Room
Room
Equipment
Computer room
Computer system
Air-handing Unit room
Air-handling Unit
Room
Computer room
Global File System
Air-handling Unit room
Air-handling Unit
Compute racks
……
……
Heat exchanger
……
Air handlers
Vapor
Cooling Towers
Absorption Refrigerating Chillers & Centrifugal Water Chillers
Gas turbine power generators

Gas co-generation system



Gas

Gas turbine
power generator
(5MW x 2)

25%

unusable
waste heat

30%

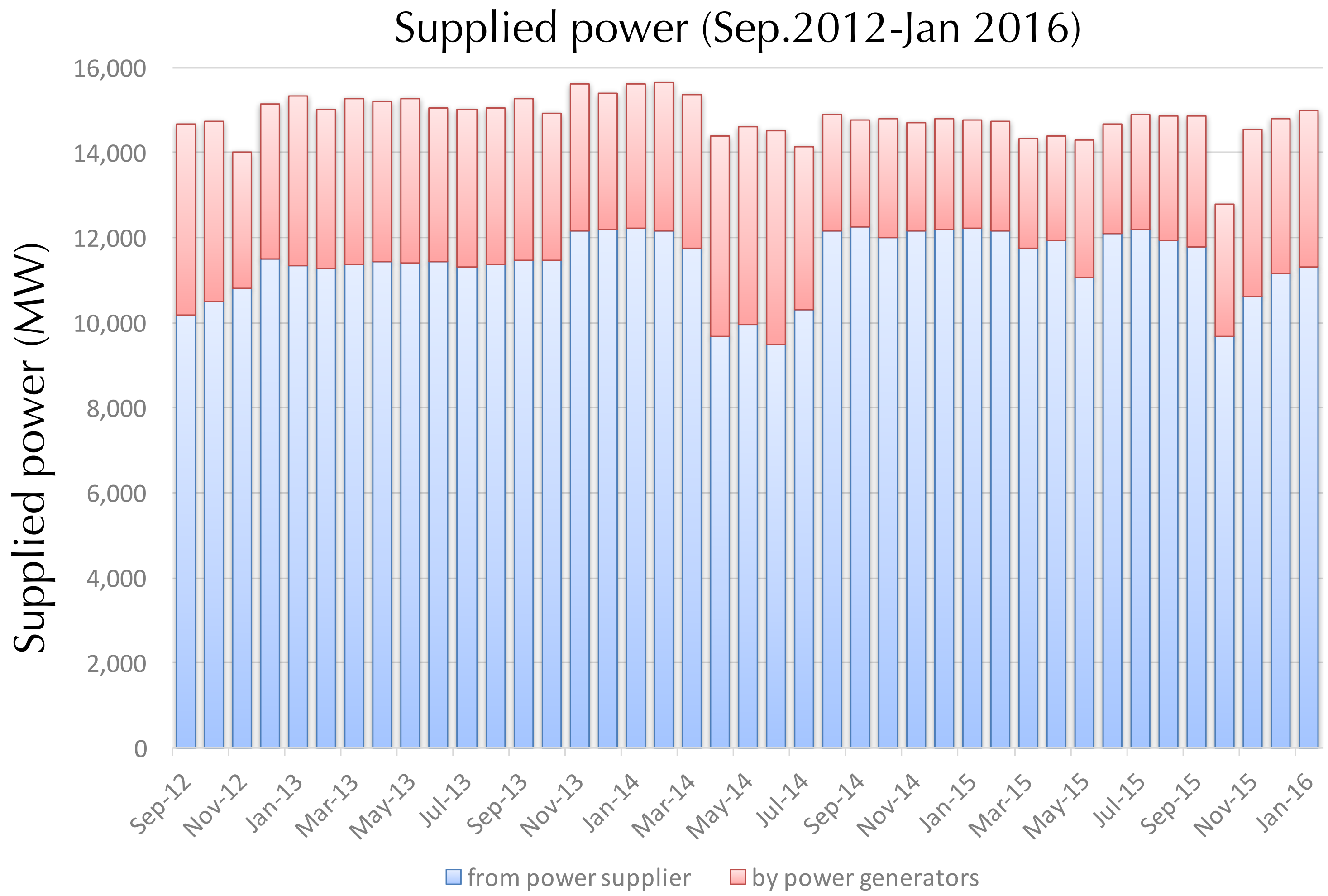
Electricity

45%

heat used for cooling and heating

Co-generation system enable to achieve higher energy efficiency
by re-using waste heat for cooling/heating

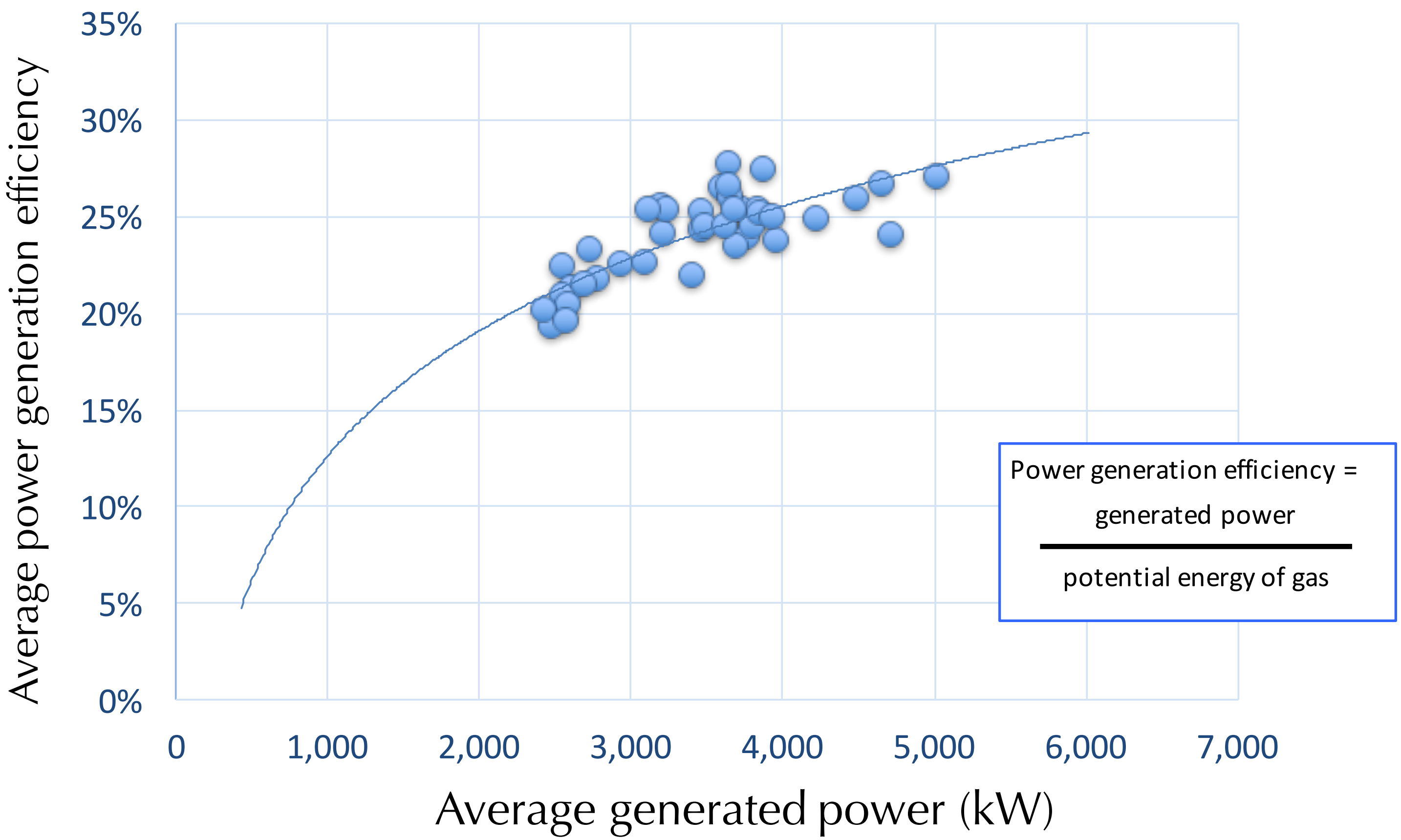
Power supply



- Average total power supply : 14.83MW
- Generated power contributes 17~35% of total power supply.

Power generation efficiency

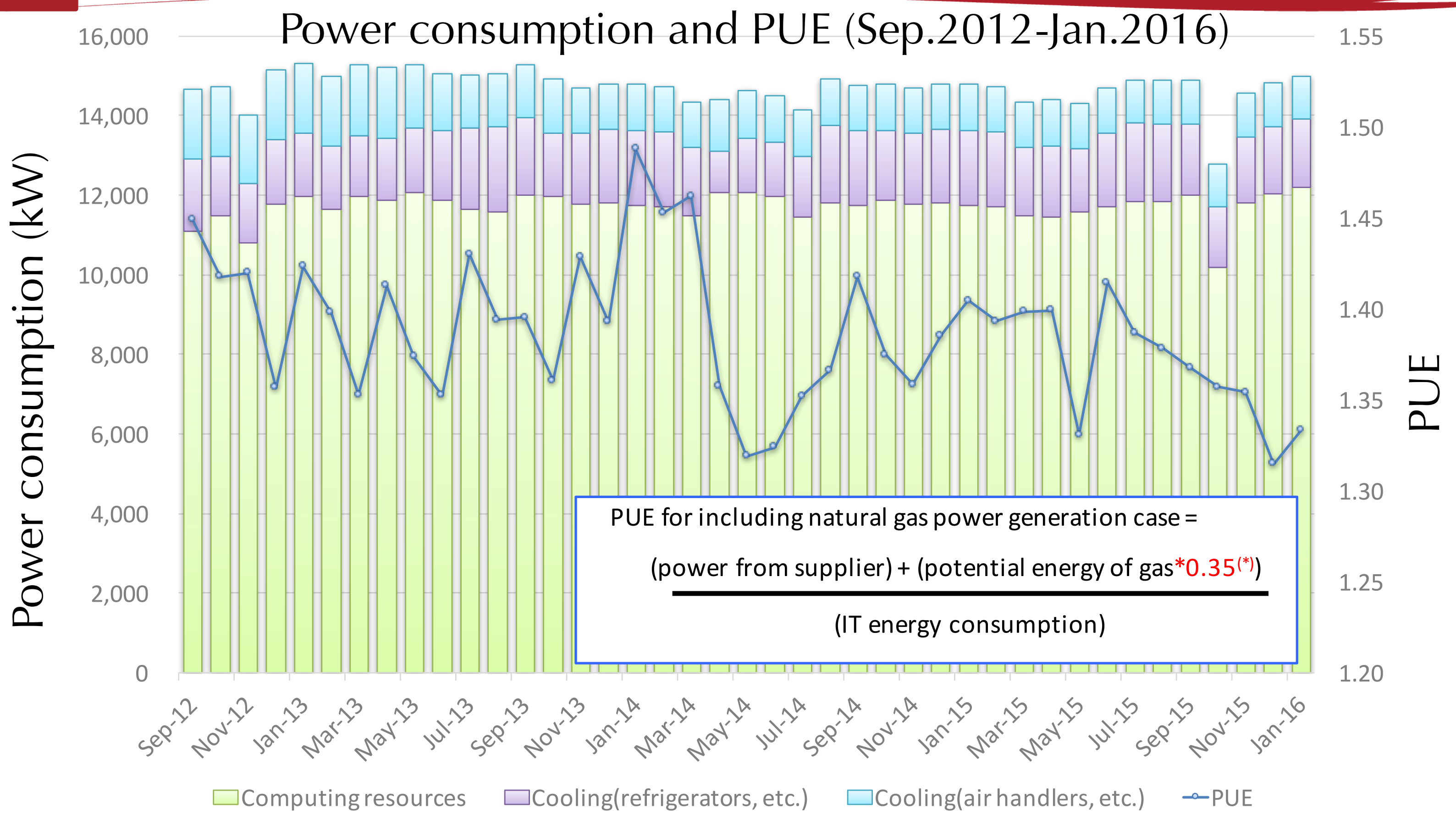
power generation efficiency (Sep.2012-Jan.2016)



Power generation efficiency = $\dfrac{\text{generated power}}{\text{potential energy of gas}}$

- As generated power increase, power generation efficiency also increase.
- Power generation efficiency flactuates between 20% to 30% (23.9% in average).

Power consumption and PUE



PUE for including natural gas power generation case =

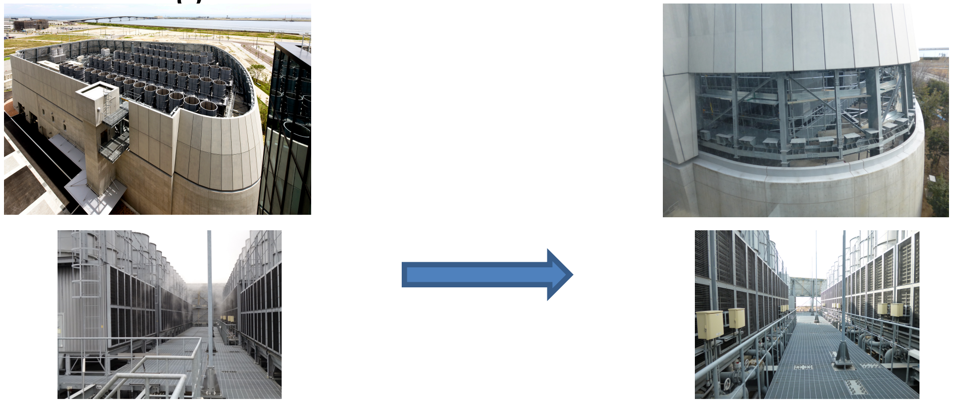
$$\frac{\text{(power from supplier)} + \text{(potential energy of gas}*0.35^{(*)}\text{)}}{\text{(IT energy consumption)}}$$

(*) Harmonizing Global Metrics for Data Center Energy Efficiency : Global Taskforce Reaches Agreement on Measurement Protocols for GEC, ERF, and CUE – Continues Discussion of Additional Energy Efficiency Metrics , October 2, 2012

- Due to power generation loss (35% - 23.9%) PUE tends to be higher.

How can we improve energy efficiency?

- **<u>To improve power generation efficiency</u>**
  - Drive power generator at peak.
    - It was not cost effective because gas rate was so higher for past few years.
- **<u>To improve cooling efficiency</u>**
  - Optimizing air handler operation
    - Working air handlers:40 -> 30
    - Working fan in handlers: 2 -> 1
      - Power consumption could be reduced to be ½ but 70% of the air flow could be kept.
      - Totally 703kW(=40%) is saved.
  - Cooling tower modified (Feb.2016)
    - Ventilation of cooling tower was not effective due to bad design.



Summary

- **Availability, reliability and failure rates**
  - The K computer achieves high availability (93.6%), reliability (MTBF:18.4days) and low failure rates (FIT_{CPU} and FIT_{DIMM} are 1/4 and 1/2 compared to BW).
  - More than 60% of system failure time was due to file system failures.
    - 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.

- **Energy efficiency**
  - We have already done some improvements about air handlers and reduced 40% of power for air handlers.
  - To drive generator at almost peak, power generation loss will be reduced and PUE will also be improved.

Thank you for your attention