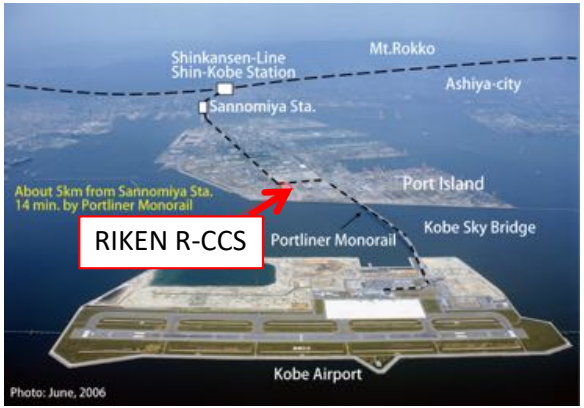
Facility enhancement and operation for “Fugaku”

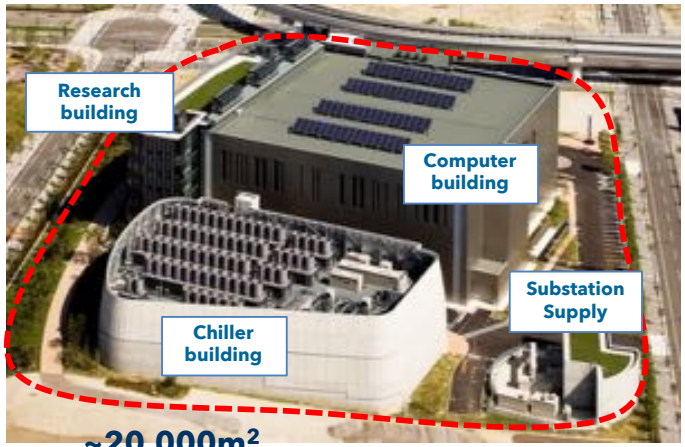
Fumiyoshi Shoji
Operations and Computer Technologies Div., R-CCS, RIKEN
@ EEHPCWG annual meeting 2019
November 18, 2019

RIKEN R-CCS

- **Established on April 1, 2018 (former name: RIKEN AICS(2010-2017)).**
- **Missions**
  - Manage the operations and enhancement of the K computer/Fugaku.
  - Promote collaborative projects with a focus on the disciplines of computational and computer sciences.
  - Plot and develop Japan's strategy for computational science, including defining the path to exa-scale computing. -> Flagship2020 project

423km (263miles) west of Tokyo



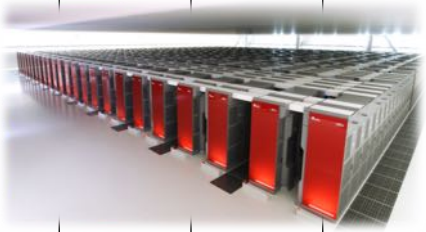


~20,000m²

K computer retired

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019

I moved to RIKEN(November 2005)



Conceptual/Detailed design for facility

Conceptual/Detailed design for K computer

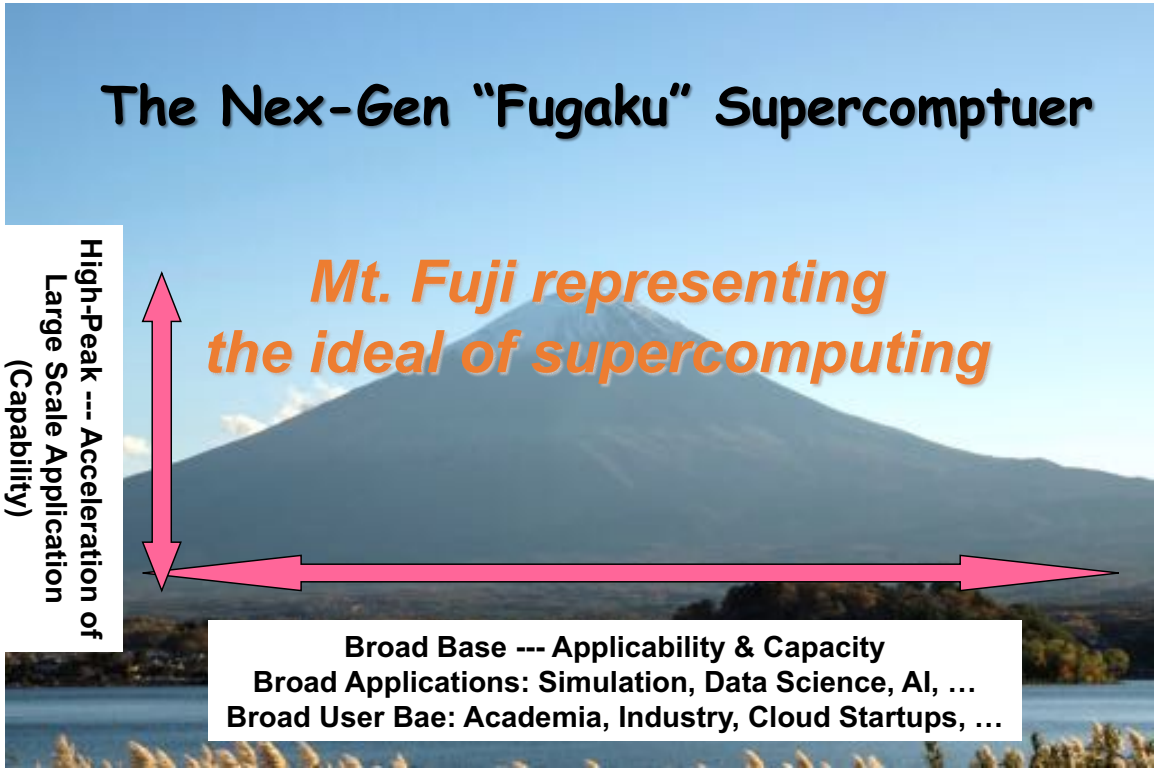
8 years and 5 months
(Apr. 2011-Aug.2019)

Early access

Official Operation

- **Achievements:**
  - **TOP500 #1** **x 2**
  - **Graph500 #1** **x 10**
  - **HPCG #1** **x 3**
  - **Gordon Bell prize winner** **x 2**

Fugaku – successor of K computer -



ふ が く
富 岳

Presentation by Satoshi Matsuoka @EEHPC SOP Workshop 2019

https://sites.google.com/view/eehpcsop2019/

Prototype of Fugaku in the latest Green500 list

Green500 List for November 2019

Listed below are the November 2019 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Note: Shaded entries in the table below mean the power data is derived and not meassured.

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	159	**A64FX prototype** - Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D , **Fujitsu** Fujitsu Numazu Plant Japan	36,864	1,999.5	118	16.876
2	420	**NA-1** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , **PEZY Computing / Exascaler Inc.** PEZY Computing K.K. Japan	1,271,040	1,303.2	80	16.256
3	24	**AiMOS** - IBM Power System AC922, IBM POWER9 20C 3.45GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100 , **IBM** Rensselaer Polytechnic Institute Center for Computational Innovations (CCI) United States	130,000	8,045.0	510	15.771

https://www.top500.org/green500/lists/2019/11/

K to Fugaku

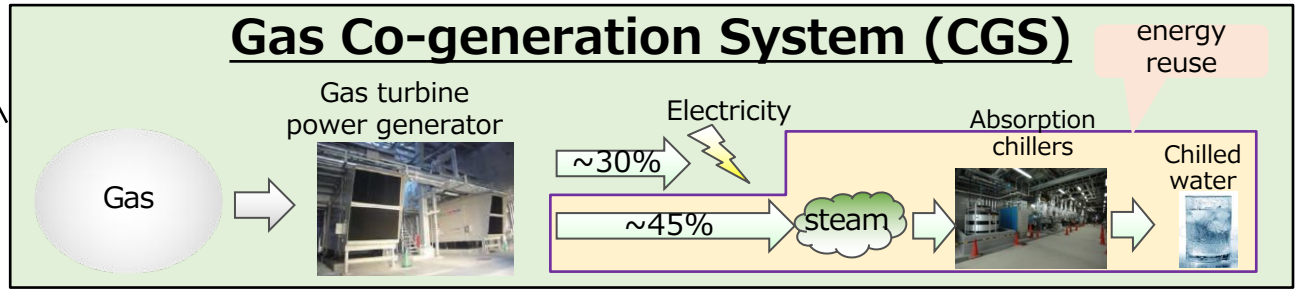
	K computer 	Fugaku 	
Official operation start	2012	2021 (planned)	10 years
CPU Architecture	SPARC64VIIIfx	A64FX (Armv8.2-A SVE)	
Peak performance	11.28 PF	500+ PF	50x
# of node	82,944	150k+	2x
Voltage	3-phase AC 200V	->	
Peak Power	15MW	30-40MW (design target)	2x
Cooling ratio (water vs air)	65:35	90:10	

Facility for K computer

- **Gas Turbine Power Generator 5MW x 1**
- **Chiller(absorption) 1700USRt x 2**

x 2 (active/standby)

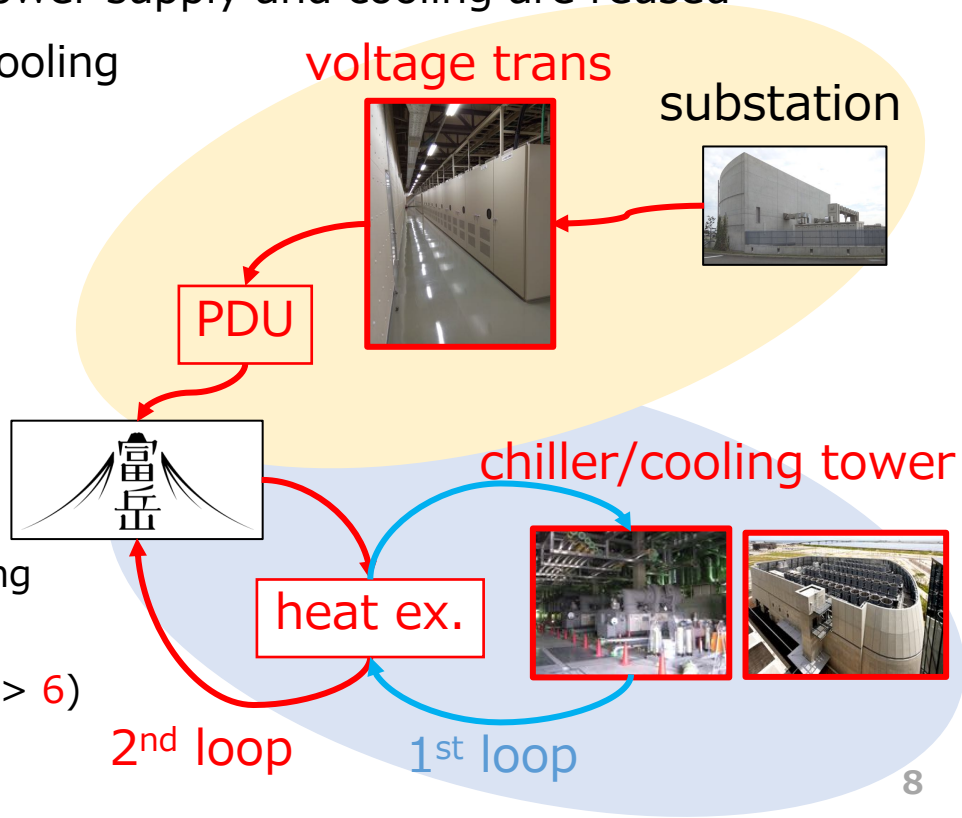
- **Chiller(Centrifugal) 1400USRt x 2, 700USRt x 1**
- **Air handlers, Cooling towers, etc.**
- **cold water cooling based design (inlet:15C, outlet:17C)**



- **2-stage voltage transformation**
  - Supplier-(77,000V)-> Substation-(6,600V)-> Transformer-(200V)-> K computer
- **PUE: 1.35 (Best case)**

Facility enhancement

- **Policy : Current facility resources should be reused as possible**
  - building, floor and all equipment for power supply and cooling are reused
  - no options without cold water based cooling
- **Enhancement**
  - Electric
    - add power line between substation and Fugaku
    - add transformers and PDUs
  - Cooling
    - add centrifugal (electric) chillers w/ cooling towers (700USRt x 1->3)
    - add zone in 2nd water loop (# of zone 5 -> 6)
    - increase heat exchanger capability

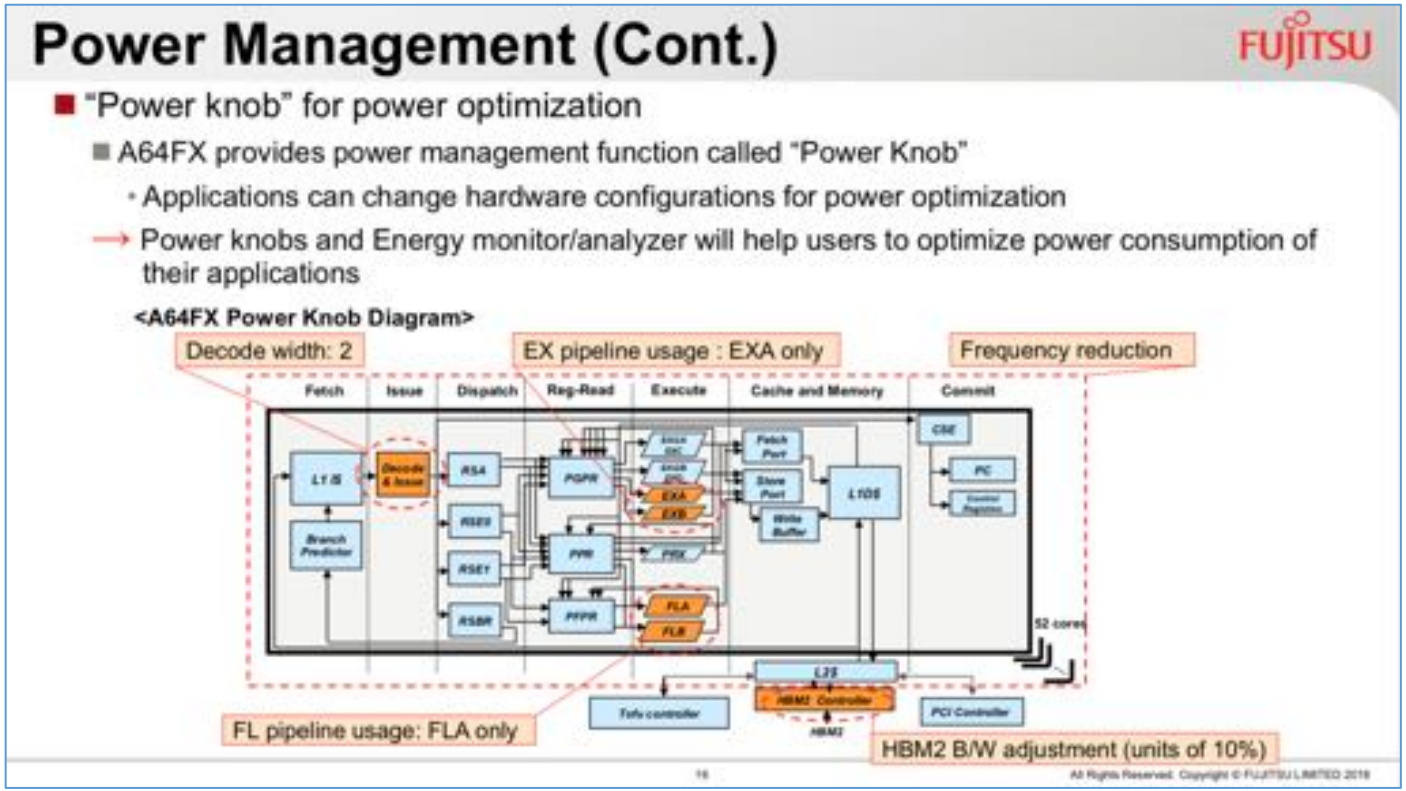


Challenges

- **Energy saving**
  - to reduce running cost
- **Power capping and fluctuation**
  - to prevent damages for system and cooling equipment by under/over cooling

New functions for energy saving

Fujitsu's presentation @ Hot Chips30
https://www.fujitsu.com/jp/Images/20180821hotchips30.pdf



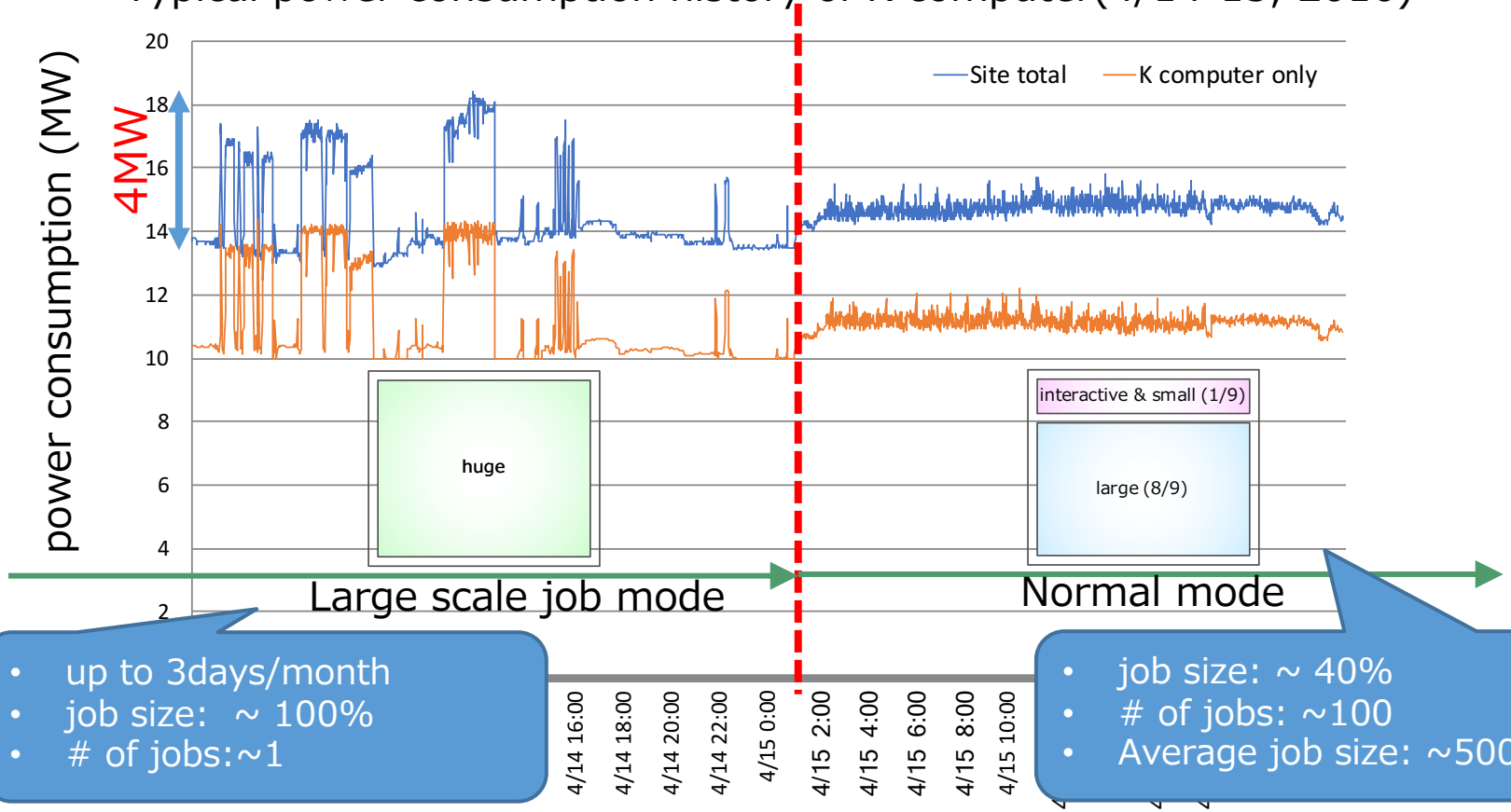
Discussion

How can we motivate users for energy saving?

- **Control power knob setting**
  1. all power knobs are turn off at default (start from minimum saving)
     - admin searches for jobs that are wasting power from profiling data
     - admin requests user to turn on the knob
  2. all power knobs are turn on at default (start from maximum saving)
     - user shows to admin that using the knob reduces (keeps) energy-to-solution for his/her job
     - admin allow the user to turn off the knob
- **Grant incentives depending on the contribution to the power saving**
  - additional node hours, higher priority, etc.
  - How can we evaluate the contributions (as-is <-> tuned)?
- **Change resource allocation unit**
  - node x hours -> energy
  - How can we keep fairness between applications which have different power profile?

Impact of full scale job

Typical power consumption history of K computer(4/14-15, 2016)



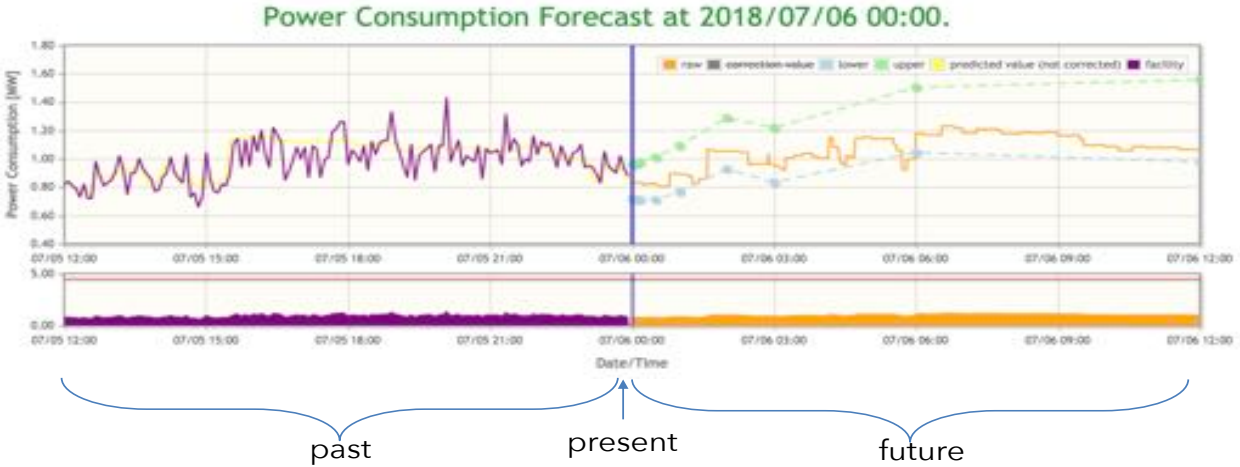
- up to 3days/month
- job size: ~ 100%
- # of jobs:~1

- job size: ~ 40%
- # of jobs: ~100
- Average job size: ~500

Power capping

Very important to prevent damages for system and cooling equipment

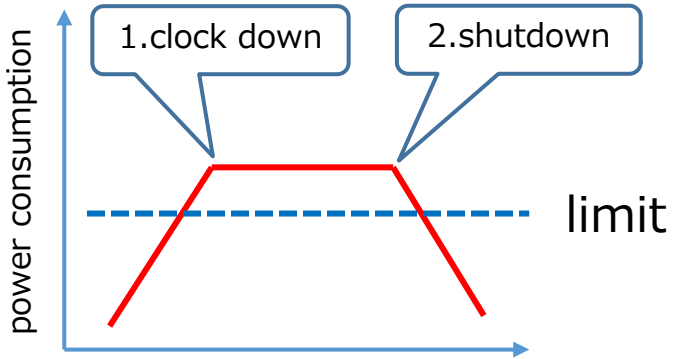
1. In normal mode, forcasting the power of the job from job history data and preparing additional chillers if needed



2. In large scale mode, preview process help us to identify power profile of the job
   - User must execute a small version (10% of full system) of the large scale job before large scale mode period.
   - We evaluate the power profile of the job and decide to admit to execute the job or not.

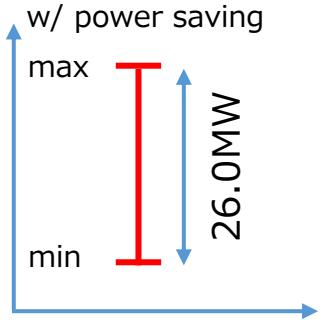
Power capping (Cont'd)

3. If unexpected power exceed is occur unfortunately, a monitoring system terminates the job(s) automatically.
   - The job(s) are selected to minimize node hours lost
4. FPGA based power capping unit (/16 node)
   1. When power consumption of CPU exceeds a limit, the unit can decrease the CPU clock on16 node without OS control
   2. If power consumption still exceeds the limit, the unit can shut down the 16 node



How can we prevent influence by huge power fluctuation

- Fugaku has function that save power at the idle
- When the function is enable, power gap between idle and peak is 26MW !
- It is impossible for chillers to follow the change of heat load



Chiller type	Quantity	Capability(MW)	Ratio	min to max	Minimum output
Absorption	4	5.98	58.1%	45min	10%
Centrifugal	2	4.93	24.0%	4min	20%
Centrifugal	3	2.46	17.9%	4min	20%

Cooling capability for quick change of heat load

(4.92x2+2.46x3)*(1.0-0.2) = 13.79MW(in 4min)

- During the large scale job mode, the function is disable, the gap decreases up to 8.3MW (the lower band increases)
- This is controllable range by centrifugal chillers throttling only