JCAHPC

Activity Report from JCAHPC and ITC, Univ. of Tokyo

Toshihiro Hanawa

Joint Center for Advanced High Performance Computing /

Information Technology Center

The University of Tokyo

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

筑波大学
計算科学研究センター
Center for Computational Sciences

Oakforest-PACS

- Full Operation started on December 1, 2016
- 8,208 Intel Xeon/Phi (KNL), 25 PF Peak Performance
  - Fujitsu
- Ranking History since Nov. 2016 (rank in Japan)

	Nov. 16	Jun. 17	Nov. 17	Jun. 18	Nov. 18	Jun. 19
TOP500	6 (1)	7 (1)	9 (2)	12 (2)	14 (2)	16 (2)
HPCG	3 (2)	5 (2)	6 (2)	9 (3)	9 (3)	9 (3)
Green500	6 (2)	21	22	24	29	29

- **JCAHPC: Joint Center for Advanced High Performance Computing)**
  - University of Tsukuba
  - University of Tokyo
    - The system was installed at Kashiwa-no-Ha (Leaf of Oak) Campus/U.Tokyo, which is located between Tokyo and Tsukuba
  - http://jcahpc.jp

Computation node (Fujitsu PRIMERGY) with single chip Intel Xeon Phi 7250 (Knights Landing, 68 cores, 3+ TFLOPS) and Intel Omni-Path Architecture card (100Gbps)

Chassis with
8 nodes, 2U size

Spec:
4.24 MW (incl. Cooling)
3.44 MW (w/o Cooling)

Green500:
2.72 MW
13.55 PF => 4.98 GF/W
#6 (Nov.'16) => #21 => #22 => #24 => #29 => #29

15 Chassis with 120 nodes per Rack

Reedbush systems in ITC, U-Tokyo (not JCAHPC)

- Supercomputer System for Data Analyses & Scientific Simulations
  - HPE (ex-SGI)

	Reedbush-U	Reedbush-H	Reedbush-L
CPU	Intel Xeon E5-2695v4 (Broadwell-EP, 2.1GHz 18core) x 2		
Memory	256 GiB (153.6GB/sec)		
GPU	NA	NVIDIA Tesla P100 : (5.3TF, 720GB/sec, 16GiB, NVLink) x 2	NVIDIA Tesla P100 : (5.3TF, 720GB/sec, 16GiB, NVLink) x 4
Interconnect	IB EDR x 1ch	IB FDR x 2ch	IB EDR x 2ch
Topology	Full-bisection Fat-tree		
# of Nodes	420	120	64
FLOPS	508.0 TF	145.2 TF(CPU)+1.27 PF(GPU)= 1.42 PF	77.4 TF(CPU)+1.35 PF(GPU)= 1.43 PF
Operation	Jul. 2016	Mar. 2017	Oct. 2017

Green500: 199 kW, 459.8 TF
2.31 GF/W
#69 (Nov.'16) => #78 => n/a

Green500: 93.6 kW, 802.4 TF
8.57 GF/W
#11 (Jun. '17) => #16

Green500: 79.0 kW, 805.6 TF
10.167 GF/W
#11 (Nov. '17)

Oakbridge-CX (OBCX)
@ ITC, U-Tokyo (not JCAHPC, but same building)

- Intel Xeon Platinum 8280 (Cascade Lake, CLX), Fujitsu
  - 28core, 2.7GHz x 2 socket, 192 GB
  - 1,368 nodes, 6.61 PF peak, 385.1 TB/sec
  - 4.2+ PF for HPL, #45 in 53rd Top500 (June 2019)
  - Fast Cache: SSD's for 128 nodes: Intel SSD, BeeGFS
    - 1.6 TB/node, 3.20/1.32 GB/s/node for R/W
    - Staging, Check-Pointing, Data Intensive Application
    - 16 of these nodes can directly access external resources (server, storage, sensor network etc.)
  - Network: Intel Omni-Path, 100 Gbps, Full Bi-Section
  - Storage: DDN EXAScaler (Lustre) 12.4 PB, 193.9 GB/sec
- Power Consumption: 950.5 kVA
  - 846 kW @HPL, 5.0+ GF/W, #28 in Green500 (June 2019)
- Operation Starts: July 1st, 2019

JPY (=Watt)/GFLOPS Rate
Smaller is better (efficient)

System	JPY/GFLOPS
Reedbush-U (HPE) (Intel BDW)	61.9
Reedbush-H (HPE) (Intel BDW+NVIDIA P100x2/node)	15.9
Reedbush-L (HPE) (Intel BDW+NVIDIA P100x4/node)	13.4
Oakforest-PACS (Fujitsu) (Intel Xeon Phi/Knights Landing)	16.5
Oakbridge-CX (Fujitsu) (Intel Cascade Lake (CLX))	20.7

Specification

What we requested in the specification for procurement about energy efficient concerns

- Total number of pages of specification (in Japanese):
  - Oakforest-PACS(OFP) on JCAHPC: 81 pages
  - (Reedbush on ITC, U-Tokyo: 87 pages, Oakbridge-CX: 82 pages)
- Related to Power and Energy efficiency: 4~5 pages
  - Requisite related to installation
  - Technical requisite: Job management system
  - Technical requisite: Automatic operation & Operation Management Support Functions
- OFP procurement included cooling system like chiller, cooling tower, and air-conditioner since requirements and proposals by vendors might be varied.

Spec.: Power Consumption

Requisite Related to Installation

- The proposed system must fulfill the following power limitation:
  - To propose the components which can be connected to 6.00 MVA of the total power supply.
  - However, available power is 4.60 MVA including cooling facility in maximum simultaneously.

To allow exceeding facility power limitation, we need power capping feature.

Spec.: Power Measurement

Requisite Related to Installation

- To provide the functions which can monitor and record power consumption for entire system in real time.
  - To provide the display function by GUI for monitoring power consumption in real time.
  - To enable Level 2 measurement based on the Energy Efficient High Performance Computing Power Measurement Methodology 2.0
  - To provide the functions which can automatically create reports and graphs based on recorded power consumption by day, month.
  - To provide the alert function to system administrators when future forecast of power consumption based on current power consumption might exceed predefined threshold.

Spec.: Power capping (1/2)

Job Management System

- If the Job Management System has a power capping function, additional points are given.
  - As a condition for additional points, to provide the function of system operation under the restriction of power consumption by forced job termination according to the priority or power-aware function with CPU frequency control etc. when power consumption of each job is based on the actual measurement during execution

Power capping enables larger system than facility's electricity limitation. (OFP does not reach limitation…)

Spec.: Power capping (2/2)

Job Management System (cont'd.)

- If the system has the job scheduling function based on the remaining power resource amount and the assumed power consumption of each job under the power capping, additional points are given. The function must include following elements. (omitted)

Power-aware job scheduling by considering electricity as resource for computing

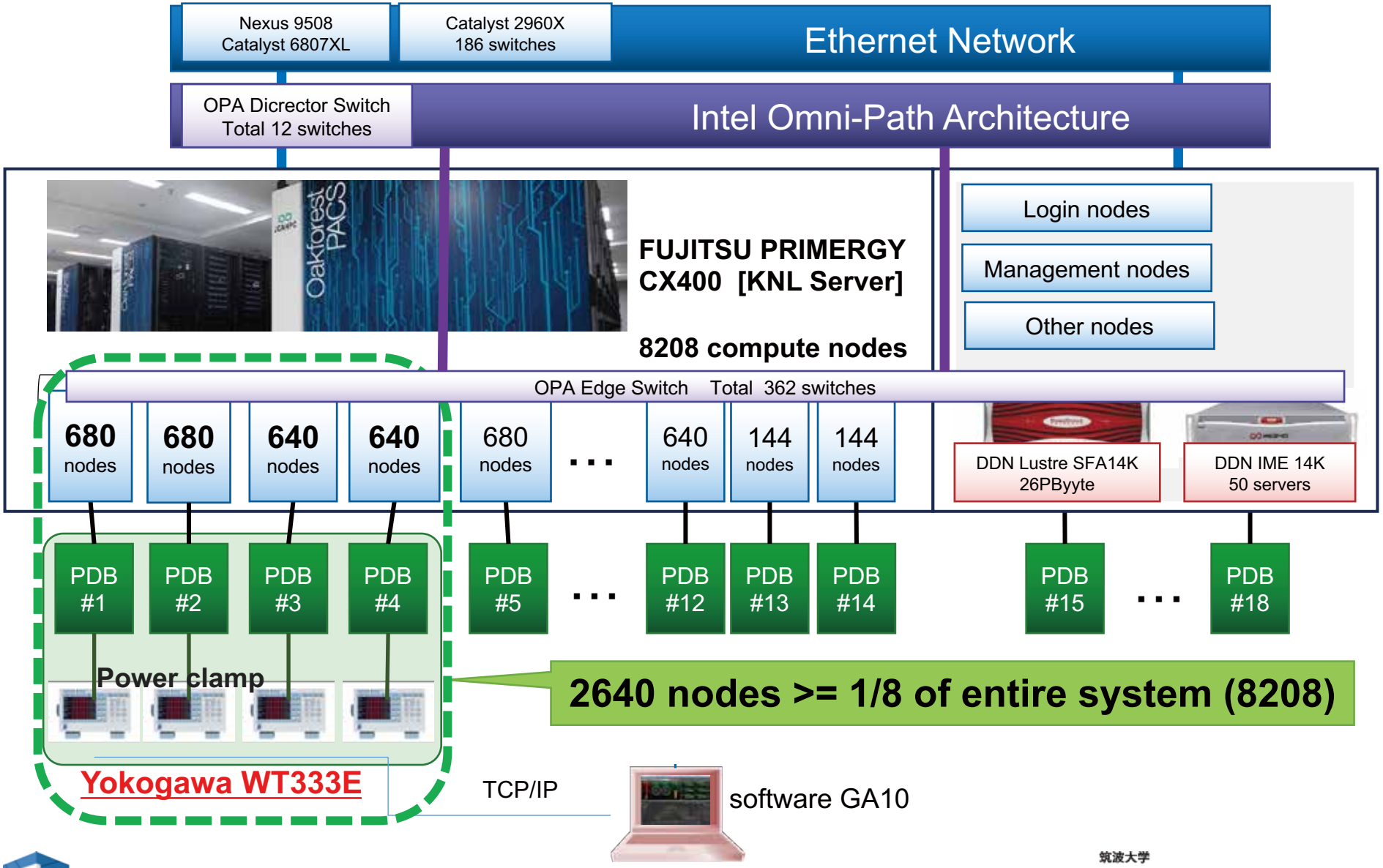
Spec: Power Saving Operation

Automatic Operation & Operation Management Support Functions

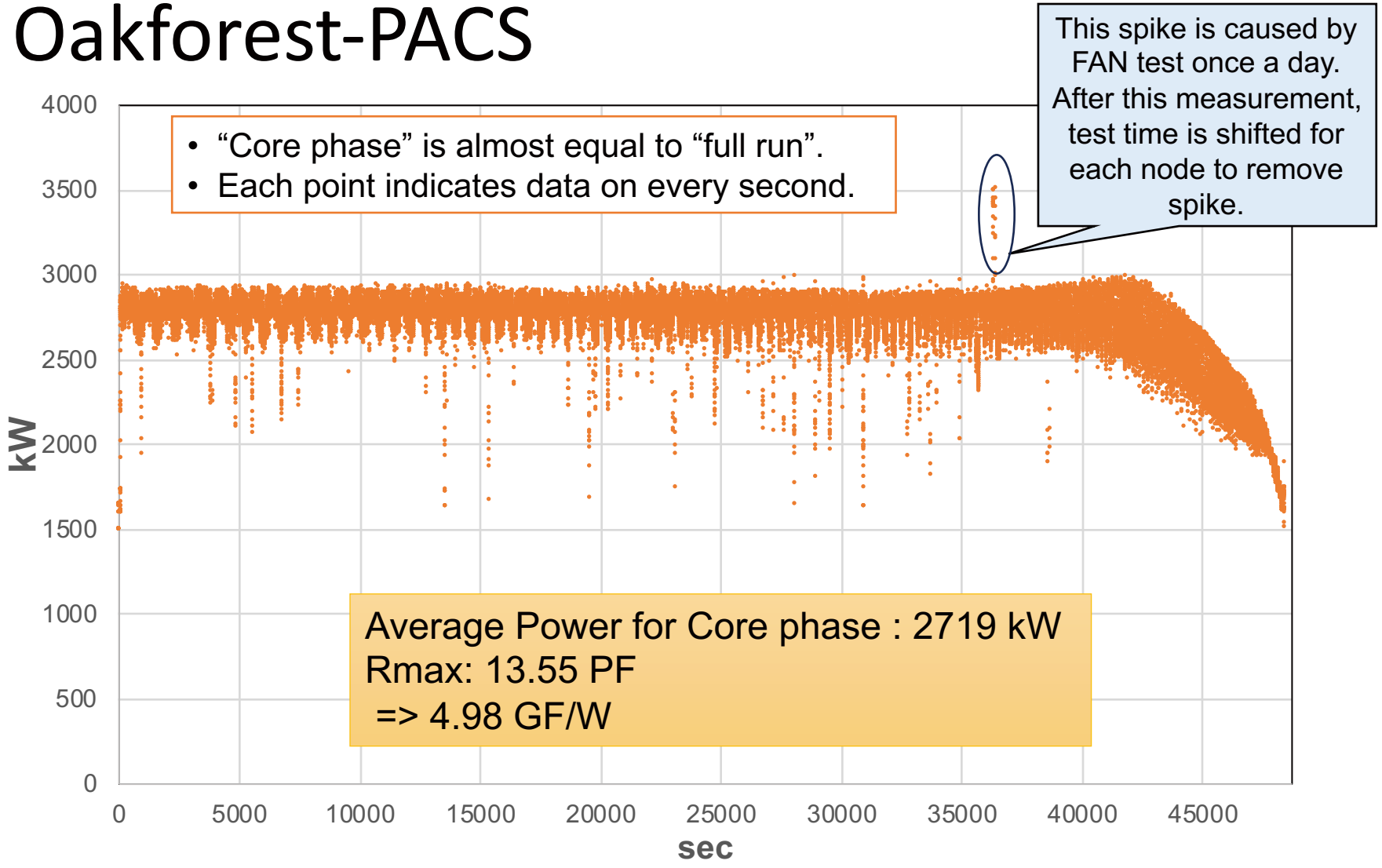
- To enable power saving operation according to changeable power supply capability at daytime, nighttime, weekend and so on. Power saving operation is defined as shrinking operation with several computing node groups, related switches, etc. specified by the system administrator shut down.

Remedy for request of saving electricity in summer or winter time frame, or for electricity shortage due to disaster

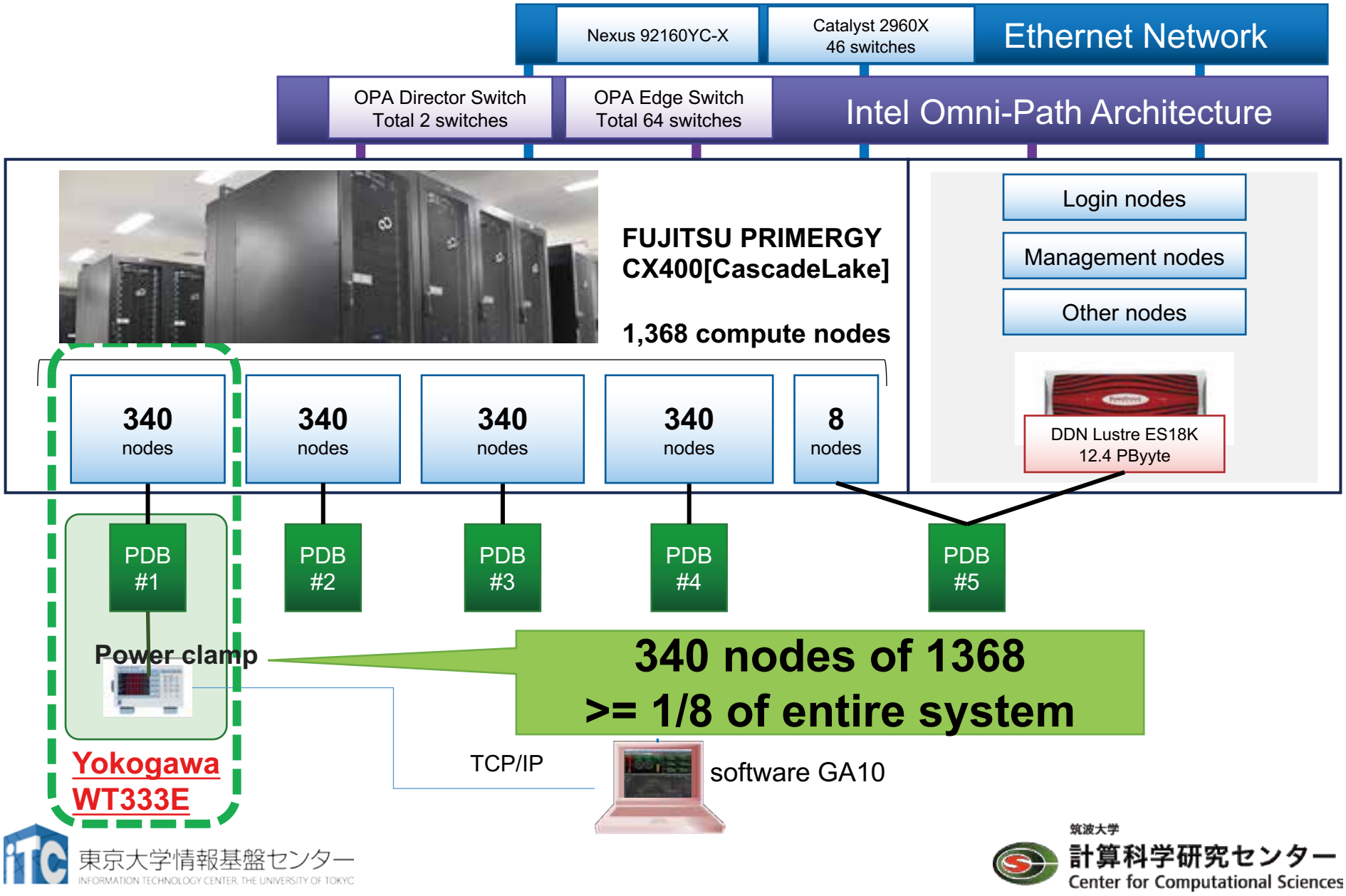
Power Monitoring for Oakforest-PACS



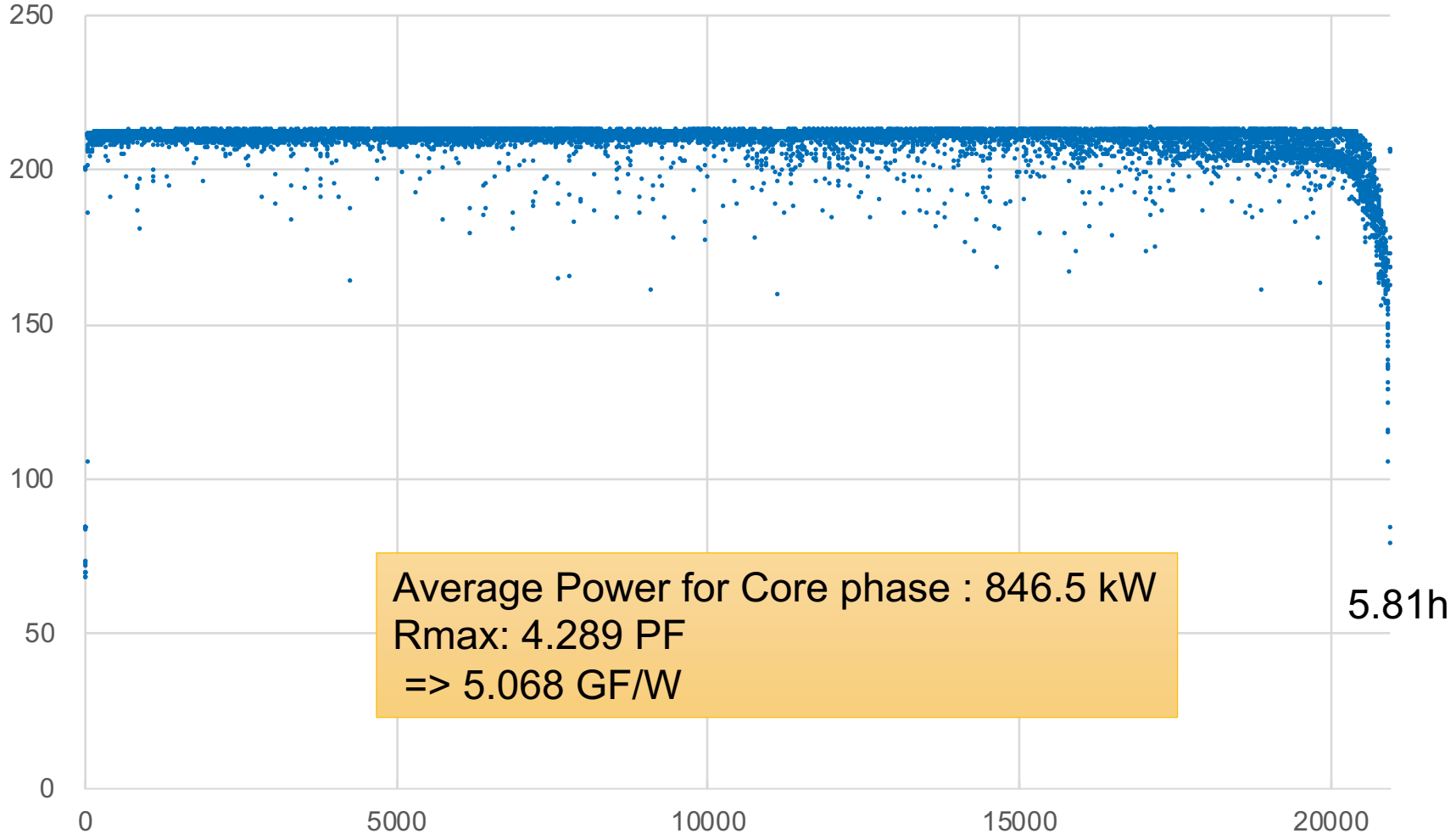
Transition of Power Consumption on Oakforest-PACS



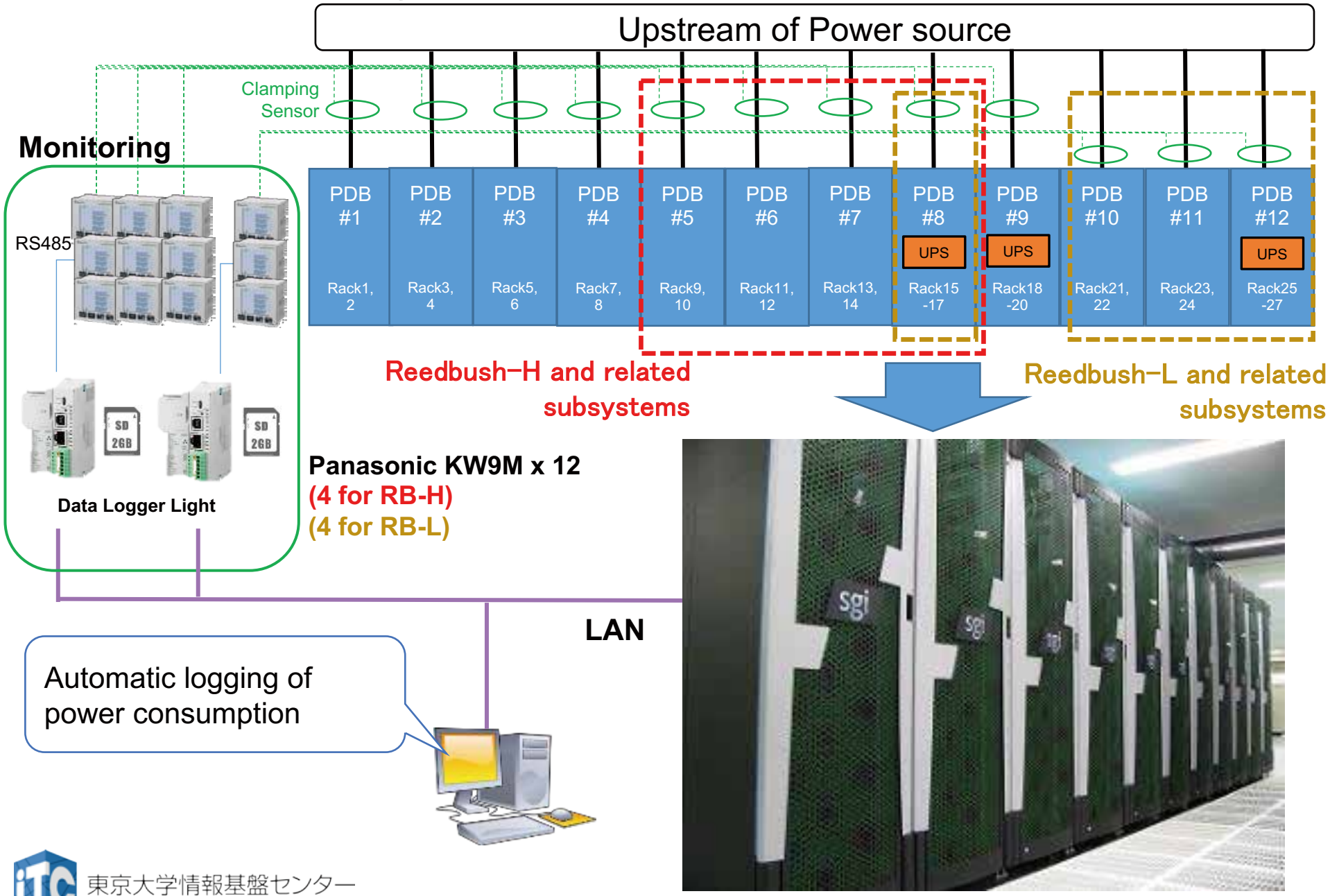
Power Monitoring for Oakbridge-CX



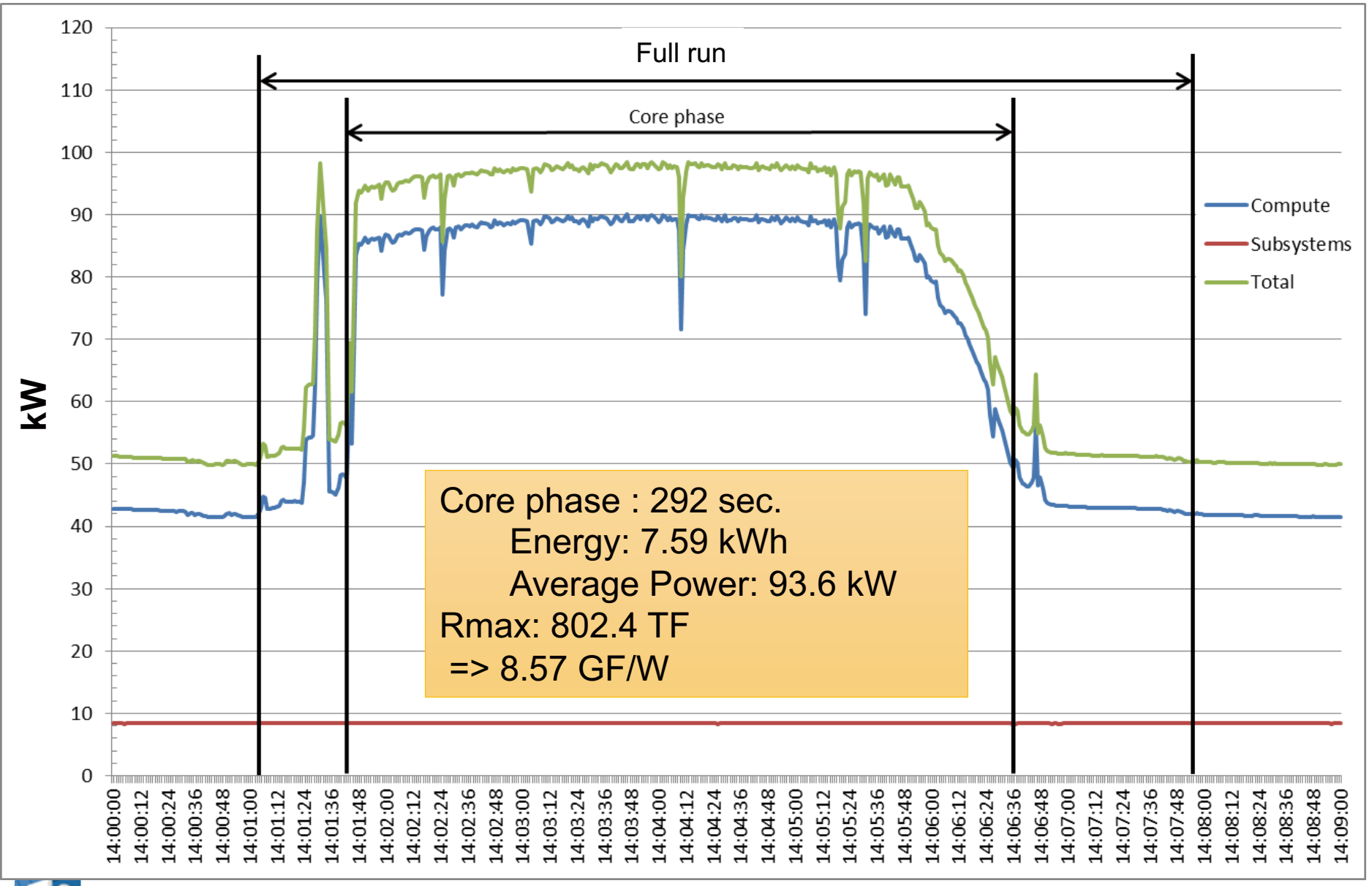
Transition of Power Consumption on Oakbridge-CX (HPL Start to End)



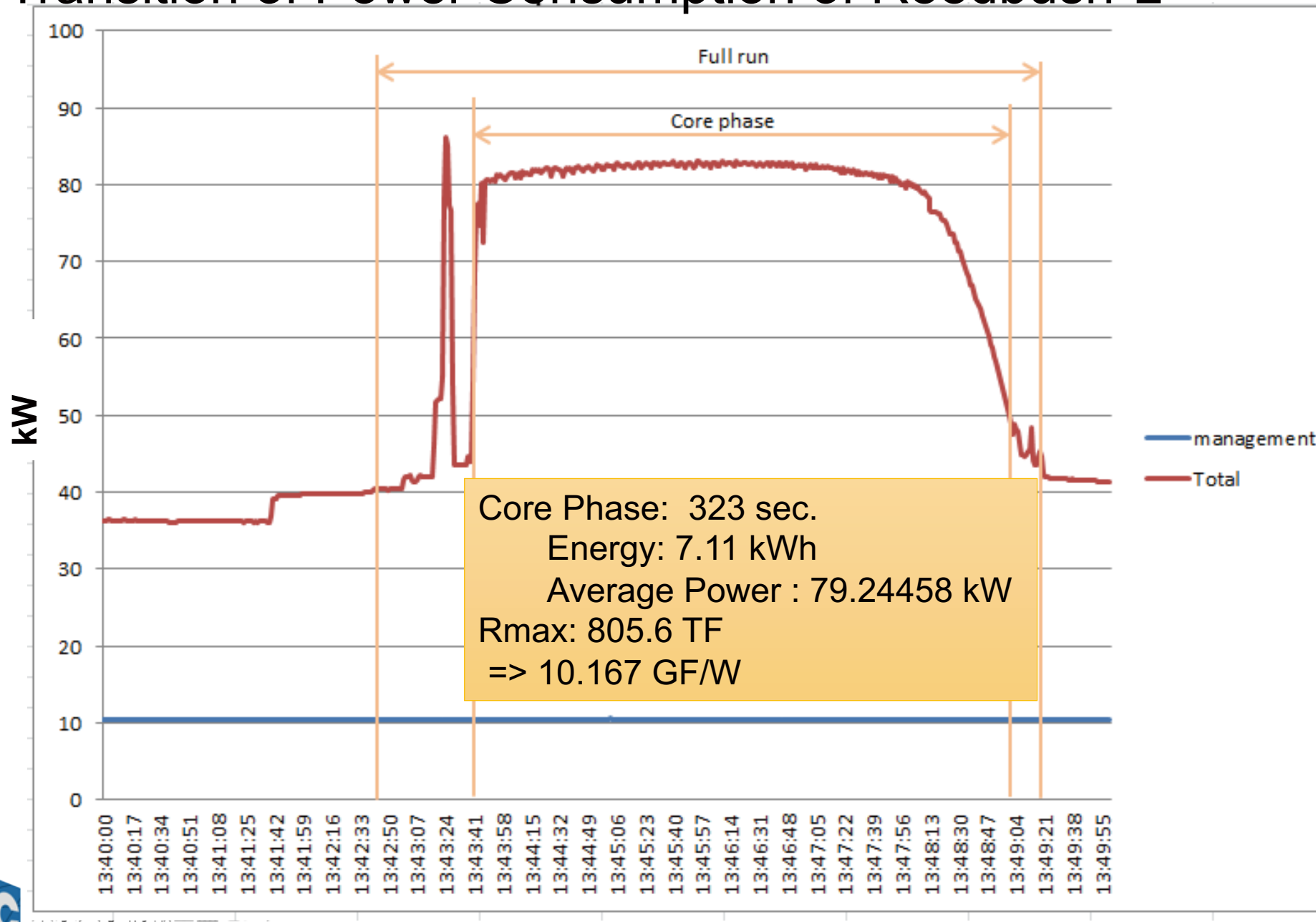
Power Monitoring for Reedbush-U, -H, and -L



Transition of Power Consumption on Reedbush-H

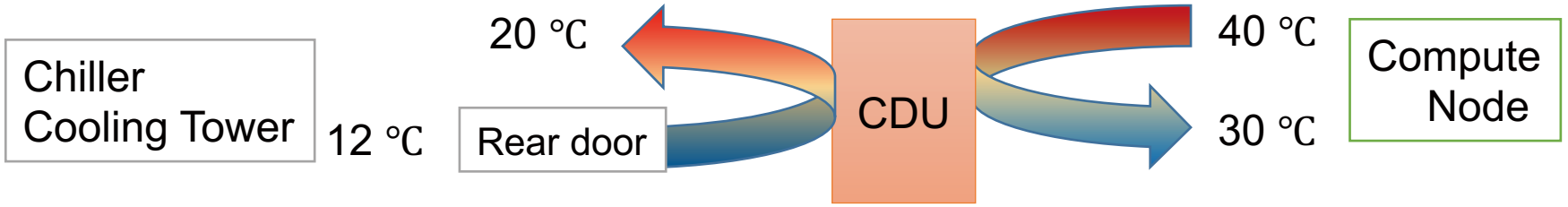


Transition of Power Consumption of Reedbush-L

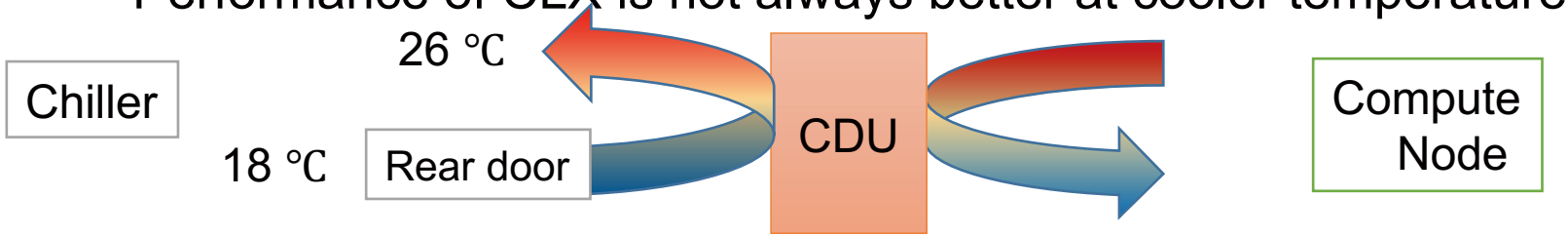


Cooling

- Temperature affects CPU freq by “Turbo boost”…
- Oakforest-PACS
  - Performance of KNL is better as cool as possible…



- Oakbridge-CX
  - Performance of CLX is not always better at cooler temperature



- Reedbush: Air cooled
  - Water cooling facility could not be installed

JCAHPC

Thank you !!

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

筑波大学
計算科学研究センター
Center for Computational Sciences