

Power Management on Oakforest-PACS (JCAHPC)



Toshihiro Hanawa

Joint Center for Advanced High
Performance Computing /
Information Technology Center
The University of Tokyo

Oakforest-PACS

- Full Operation started on December 1, 2016
- 8,208 Intel Xeon/Phi (KNL), 25 PF Peak Performance
 - Fujitsu
- Nov. 2016: TOP500 #6 (#1 in JP), HPCG #3 (#2), Green500 #6 (#2)
- Jun. 2017: TOP500 #7 (#1 in JP), HPCG #5 (#2), Green500 #21
- Nov. 2017: TOP500 #9 (#2 in JP), HPCG #6 (#2), Green500 #22, IO500 #1
- Jun. 2018: TOP500 #12 (#2 in JP), HPCG #7 (#2), Green500 #25, IO500 #1
- Nov. 2018: TOP500 #14 (#2 in JP), HPCG #9 (#2), Green500 #29, IO500 #?

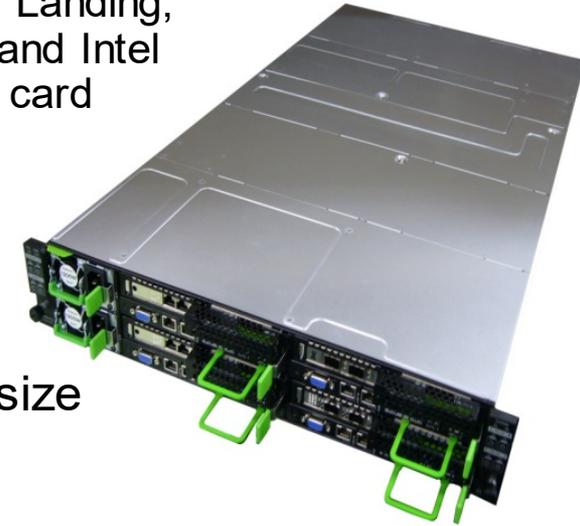
JCAHPC: Joint Center for Advanced High Performance Computing

- 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 (max. rating):
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)

15 Chassis with 120 nodes per Rack

BoF of Energy Efficiency
Considerations for HPC
procurements

Supercomputer Procurement Rule in Japan

- Basically, we can only purchase the available products at the time when the system is installed.
- Each vendor must propose an appropriate system which consists of components which are already available, or brand-new products under development with estimation at bidding.
- The actual installed system must fulfill proposed performance numbers before at acceptance validation.
- The actual product name or vendor preference cannot be described in specification.
- Disclosure of budget amount is prohibited.

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)
- 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.
 - (omitted)
 - 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

Thank you !!

Why we (JCAHPC and U-Tokyo) used Level 2/3?

- They reflect more realistic system potentials
 - Typically Level 1 number becomes better than L2/L3
 - On level 1, power of interconnect can be estimated, and power of subsystem can be omitted
- Our philosophy:
 - Measure the system performance under the same condition, configuration, and assumptions as the product run phase as much as we can
- Comments to measurement methodology of Green500:
 - About the ranking of Reedbush-L
 - Perf/W in Top500: #10, but #11 in Green500
 - “Power optimized number” is allowed to submit for Green500 but no one knows this fact without seeing Excel sheet
 - ex.) Power optimized run on Reedbush-L
 - => 770.0TFLOps, 10615.92 Mflops/W, can be 8th in current Green500
 - Only 2 systems are measured by L2/L3 in Top 10 of Green500 but only few knows...

Table 3.1: Summary of aspects and quality levels

Aspect	Level 1	Level 2	Level 3
1a: Granularity	One power sample per second	One power sample per second	Continuously integrated energy, voltage and current sampled at 5 kHz for AC / 120 Hz for DC
1b: Timing	At equal intervals across the entire core phase of the run, which must be at least one minute	At equal intervals across the full run	At equal intervals across the full run
1c: Measurements	Core phase average power	<ul style="list-style-type: none"> • Core phase average power • Full run average power • 10 average power measurements in the core phase • Idle power 	<ul style="list-style-type: none"> • Core phase total energy • Full run total energy • 10 total energy measurements in the core phase • Idle power
2: Machine fraction	The entire system; at least 40 kW; or the largest of $\frac{1}{10}$ of the compute subsystem, 2 kW, or 15 compute nodes	The greater of $\frac{1}{8}$ of the compute-node subsystem or 10 kW or 15 compute nodes - alternatively the entire system	The whole of all included subsystems
3: Subsystems	Compute-nodes	Compute-nodes	All

BoF of Energy Efficiency Considerations for HPC

		<ul style="list-style-type: none"> • 10 average power measurements in the core phase • Idle power 	<ul style="list-style-type: none"> • 10 total energy measurements in the core phase • Idle power
2: Machine fraction	The entire system; at least 40 kW; or the largest of $\frac{1}{10}$ of the compute subsystem, 2 kW, or 15 compute nodes	The greater of $\frac{1}{8}$ of the compute-node subsystem or 10 kW or 15 compute nodes - alternatively the entire system	The whole of all included subsystems
3: Subsystems	Compute-nodes measured, interconnect measured or estimated	Compute-nodes measured, all participating subsystems measured or estimated	All participating subsystem must be measured
4: Location of measurement	Upstream of power conversion OR Conversion loss modeled with manufacturer data	Upstream of power conversion OR Conversion loss modeled with off-line measurements of a single power supply	Upstream of power conversion OR Conversion loss measured simultaneously
4b: Meter accuracy	5%	2% (see Section 3.1)	1% (see Section 3.1)