SC17

Denver, CO | hpc connects.

E2 HPC Working Group

Total Cost of Ownership and HPC System Procurement

Total Cost of Ownership and HPC System Procurement

SC17: Birds of Feather

Tuesday Nov 14, 12:15-1:15 Room 605

Steven Martin, Cray Inc. (Moderator)

Opening Statements

- Over the past 10 years, we have started to include more operational and facility elements in our procurement documents.
  - Should we be more systematic and collective in our approach?
  - How do we refine this process moving forward?

- What are the site constraints and trade-offs in money that goes into preparing the site verses the machine performance?

- How are they articulated in the procurement process documents and discussions?

Opening Statements

- Each site has unique constraints and trade-offs.

- How do you decide which of the site specifics to include in the procurement process and how is that communicated to the vendor?

- Chillerless operation, power distribution and power/energy limits :
  - What are the absolute constraints?
    - What are the trade-offs and how are decisions made?
  - How do you communicate this to the vendors during the procurement process?
  - How do you write it into your procurement documents?

Introduction: Panelists

- Gert Svensson, KTH:
  - PRACE TCO and Procurement
- Eric Boyer, GENCI:
  - TCO model
- Jim Rogers, ORNL:
  - Chiller less cooling
- Jason Hick, LANL:
  - ACES and power distribution
- Toshio Endo, Tokyo Institute of Technology:
  - Power/energy limits
- Steven Roberts, IBM:
  - Vendor neutral presentation

PRACE TCO and Procurements

Gert Svensson, KTH

What is TCO?

- Total Cost of Ownership (TCO) is an analysis that places a **single value** on the complete lifecycle of a purchase

- The concept of TCO is to determine a product's **true cost** not just based on the price but on all the costs it takes to purchase and maintain the product during the **entire time** it is owned.

Prace White Paper

Willi Homberg et al., Ways to Decrease TCO and Infrastructure Related Costs, PRACE-4IP White Paper

Ways to Decrease TCO and Infrastructure Related Costs

Willi Homberg
Jülich Supercomputing Centre, JSC, Germany

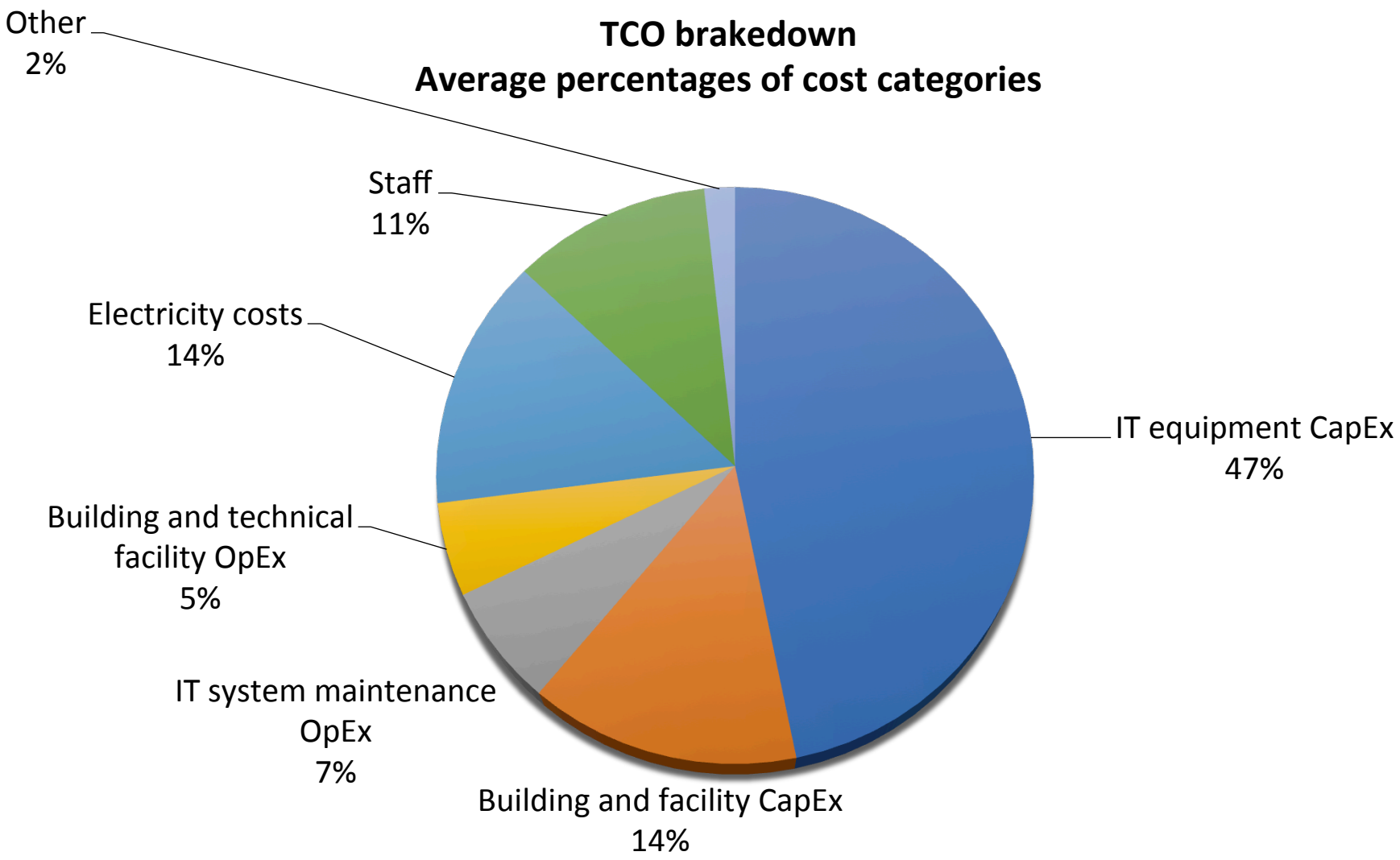
Contributors: Eric Boyer (GENCI), Radoslaw Januszewski (PSNC), Norbert Meyer (PSNC), Dirk Pleiter (JSC), Francois Robin (CEA), Philippe Segers (GENCI), Gert Svensson (KTH), Mateusz, Tykierko (WCSS), Torsten Wilde (LRZ) and all PRACE partners contributing to the TCO survey

Soon available on http://www.prace-ri.eu/white-papers/

Sections of the report

1. TCO cost categories
2. Strategy for reducing costs
3. Survey on state of the art costs and best practices in cost reduction
4. TCO based procurement of supercomputers

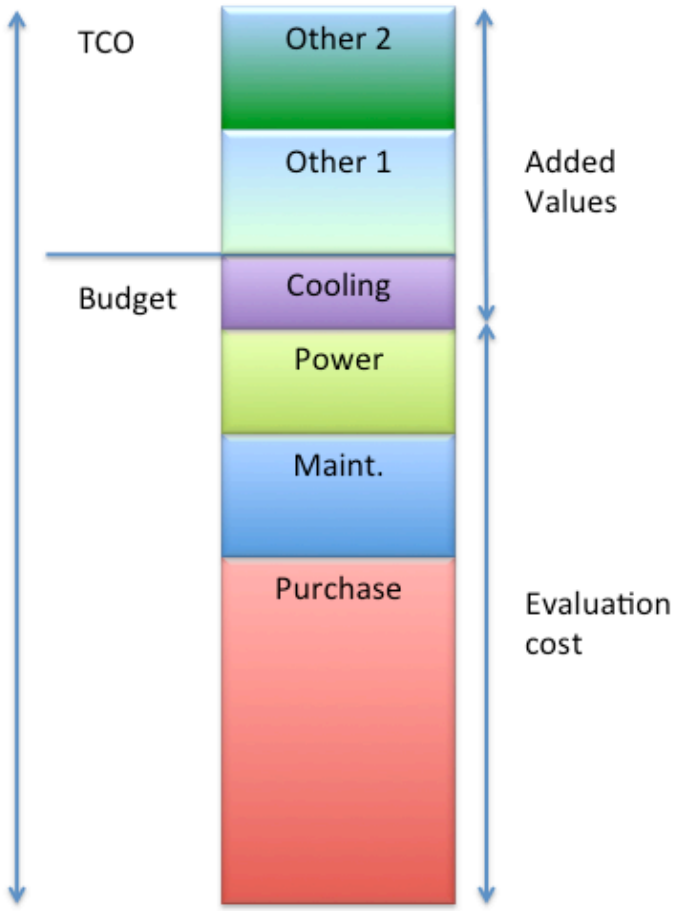
Example of survey results



Experience of TCO based procurements

- KTH have used TCO based procurements in two generations of main systems with good results
- Many advantages to use TCO model
  - “True cost” used for evaluation of purchase
  - Energy growing part of the cost
  - Taking energy efficiency and heat re-use into account
- Many “excuses” not to use TCO model
  - Hard to estimate all cost but better to estimate (guess) than skip
  - Too hard work
  - Budget not in one lump

A TCO method for split budgets



- In many organizations budgets are split in different areas and funds can't be transferred between the areas
- In KTH case: one budget for purchase, maintenance, power and cooling
- Solved by introducing additional minimum requirements
- Example: purchase + maint. + power + cooling < budget
- The vendor can choose the largest system that fulfills this
- Select the bid that fulfils all minimum requirements with the lowest evaluation cost

TCO Model

Eric Boyer, GENCI

TCO based evaluation : Why

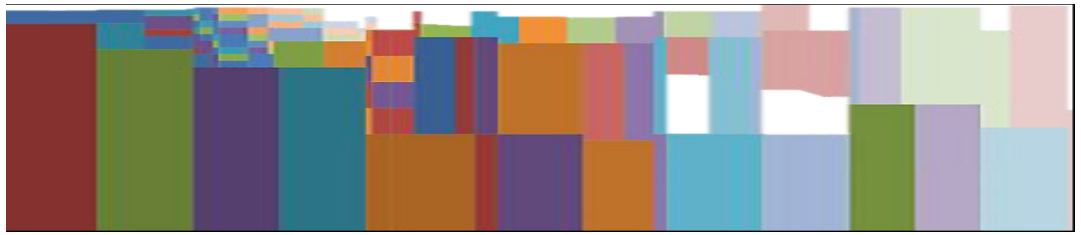
HPC solution life time TCO trends

- Purchase cost proportion (CAPEX) is decreasing
- Energy cost is increasing
- Additional infrastructure solution related are sometimes required
- Workload profile has a significant impact on energy footprint

Objectives

- Best fit for the targeted workloads
- Incentive for :
  - Bids with the best balance between components
  - Setting up energy optimization environments
- Best value for money !

TCO based evaluation : How



Methodology

- Based upon a custom workload
  - ❑ As much as close to the targeted activity during machine life time
  - ❑ Composed by a set of n “production” codes run simultaneously on a given number of nodes = pattern
- Costs already included in evaluation of 5 years
  - ❑ Machine acquisition cost
  - ❑ Maintenance
  - ❑ Specific staff
  - ❑ Energy cost
    - Computation nodes and interconnect
    - Storage and service nodes
    - Power and cooling components (internal)
  - ❑ Infrastructure specific add-on

➔ Actual energy consumption is measured during workload execution

TCO based evaluation : How (cont'd)

Variable definition :

t : wallclock time of the workload
ECt : Energy consumption by computational nodes
EDt : Energy consumption from disks and service nodes
EUt : Energy consumption from internal cooling devices
c : total core count of the proposed configuration
cwl : core count used for the selected workload
T : Number of second in 5 years
N : Number of runs of the workload during 5 years
ET : Total energy consumption during 5 years
EC : Energy of computational part during 5 years
ED : Energy of storage and service nodes during 5 years
EU : Energy of internal cooling device during 5 years
TCOE : TCO due to energy consumption during 5 ans
PUEair : Power Usage Effectiveness for air cooled devices
PUEcoldw : Power Usage Effectiveness for "Cold water" cooled devices
PUEhotw : Power Usage Effectiveness for "Hot water" cooled devices

TCO based evaluation : How (cont'd)

Energy calculations :

$$E_C = E_{Ct} * \frac{c}{c_{wl}} * PUE_{compute} * \frac{T}{t}$$

$$E_D = E_{Dt} * PUE_{disk} * \frac{T}{t}$$

$$E_U = E_{Ut} * PUE_{cdu} * \frac{T}{t}$$

With

$$\boldsymbol{PUE_{compute}} = 1 + \left[(PUE_{air} - 1) * \%_{air-for-compute}\right] + \left[(PUE_{coldw} - 1) * \%_{coldw-for-compute}\right] + \left[(PUE_{hotw} - 1) * \%_{hotw-for-compute}\right]$$

$$\boldsymbol{PUE_{disk}} = 1 + \left[(PUE_{air} - 1) * \%_{air-for-disk}\right] + \left[(PUE_{coldw} - 1) * \%_{coldw-for-disk}\right] + \left[(PUE_{hotw} - 1) * \%_{hotw-for-disk}\right]$$

$$\boldsymbol{PUE_{cdu}} = 1 + \left[(PUE_{air} - 1) * \%_{aircdu}\right] + \left[(PUE_coldw - 1) * \%_{holdw-for-cdu}\right] + \left[(PUE_{hotw} - 1) * \%_{hotw-for-cdu}\right]$$

TCO based evaluation : How (cont'd)

Then :

$$E_T = E_C + E_D + E_U$$

$$E_T = \left[\left(E_{Ct} * \frac{c}{c_{wl}} * PUE_{compute}\right) + (E_{Dt} * PUE_{disk}) + (E_{Ut} * PUE_{cdu})\right] * \frac{T}{t}$$

$$TCO_E = E_T * EnergyCost$$

$$TCO = Purchase + Maintenance + Infrastructure + Staff + TCO_E$$

Where:

- *Purchase* is the acquisition cost of the machine
- *Maintenance* is the maintenance cost over 5 years
- *Infrastructure* is the specific cost of infrastructure for the machine
- *Staff* is the cost for complementary staff for machine exploitation over 5 years

Cost of the workload (TCO_b) takes into account the number of time (N) it twill be possible to run it over 5 years

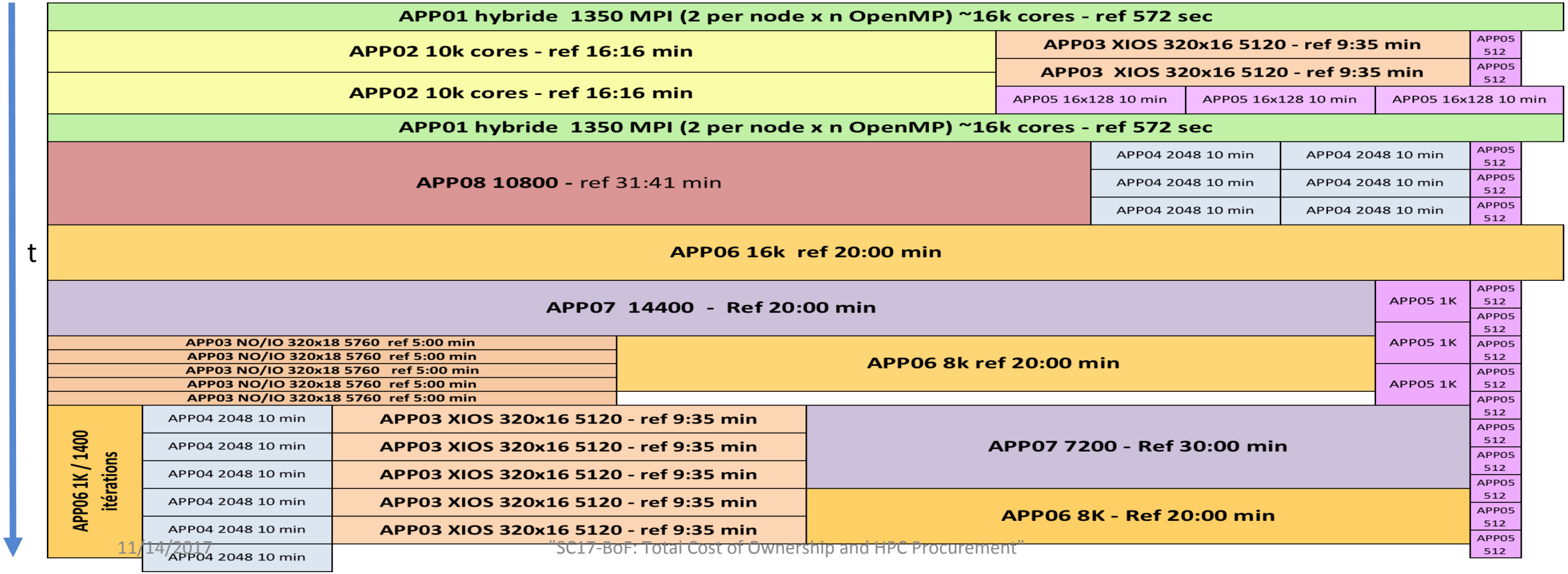
$$N = \frac{c}{c_{wl}} * \frac{T}{t}$$

$$TCO_b = {}^{TCO}/_{N}$$

TCO based evaluation : Implementation (cont'd)

Workload building for TGCC procedure :

- Build on the basis of reference time
- Any other arrangement is allowed
- Final view of the optimized pattern running on ~16k cores during ~3 hours



TCO based evaluation : Implementation (cont'd)

General conditions :

- The proposal has a fixed budget for machine acquisition and maintenance
- Some variables may be added or removed depending on the context (related to water temperature for example)
- The candidate can tune the workload :
  - Ordering for optimal "TETRIS"
  - Frequency adjustment allowed (to lower the energy envelope) on a per job basis
  - Specific code optimization is allowed

Requisites :

- Energy measurement must be available at appropriate grain

Issues :

- Power of 2 era has ended for core count
- Hybrid architectures are introducing a higher level of complexity for comparison
- Inclusion of potential energy reuse in the model

→ Workload design must adapt

TCO based evaluation : Implementation

Used by GENCI for the two last procurements :

- OCCIGEN facility @CINES in production at CINES 3,5 Pflops since end of march.
  - First lot 2,1 Pflops (haswell based) TCO cost : 274,8€ / pattern
  - Second lot + 1,4 Pflops (broadwell based) TCO cost : 247,1€ / pattern

- CURIE replacement @TGCC IRENE to be installed early 2018

  - ❑ Atos/BULL SEQUANA system
  - ❑ 9 Pflop/s peak for the first phase
  - ❑ Mixed Intel processor technology :
    - ➢ KNL 7250
    - ➢ Skylake 8168 24 core 2.7 GHz
  - ➢ TCO cost : 666,21€ / pattern

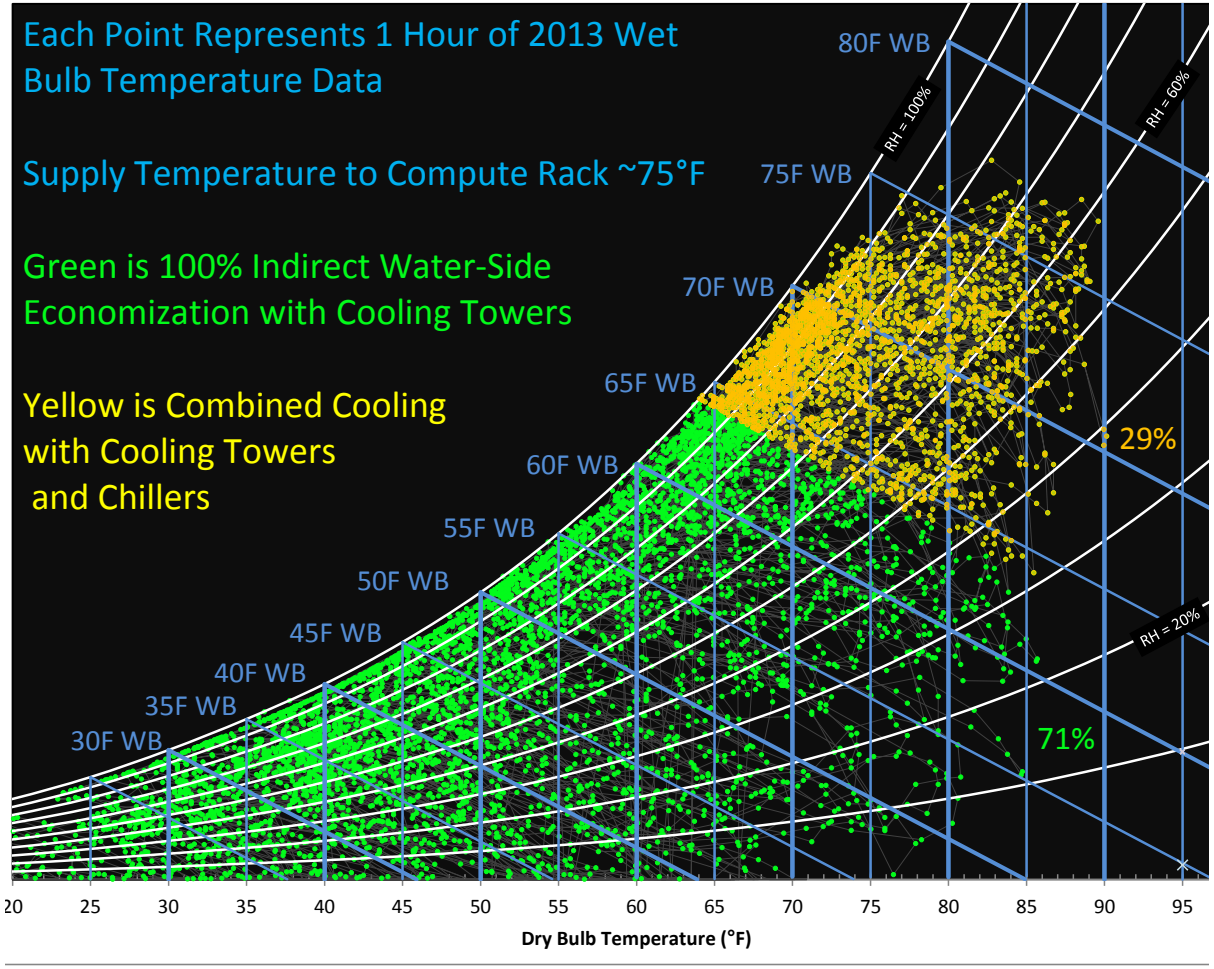
➔ And for the next procurement for HPC platforms at IDRIS (CNRS computing center) to be published in the few weeks

SC'17 bof224 / Total Cost of Ownership and HPC System Procurement

Jim Rogers

Director, Computing and Facilities
National Center for Computational Sciences
Oak Ridge National Laboratory

Water Side Economization for Summit's MWT Loop



"Yellow" hours require a contribution from the chilled water facility

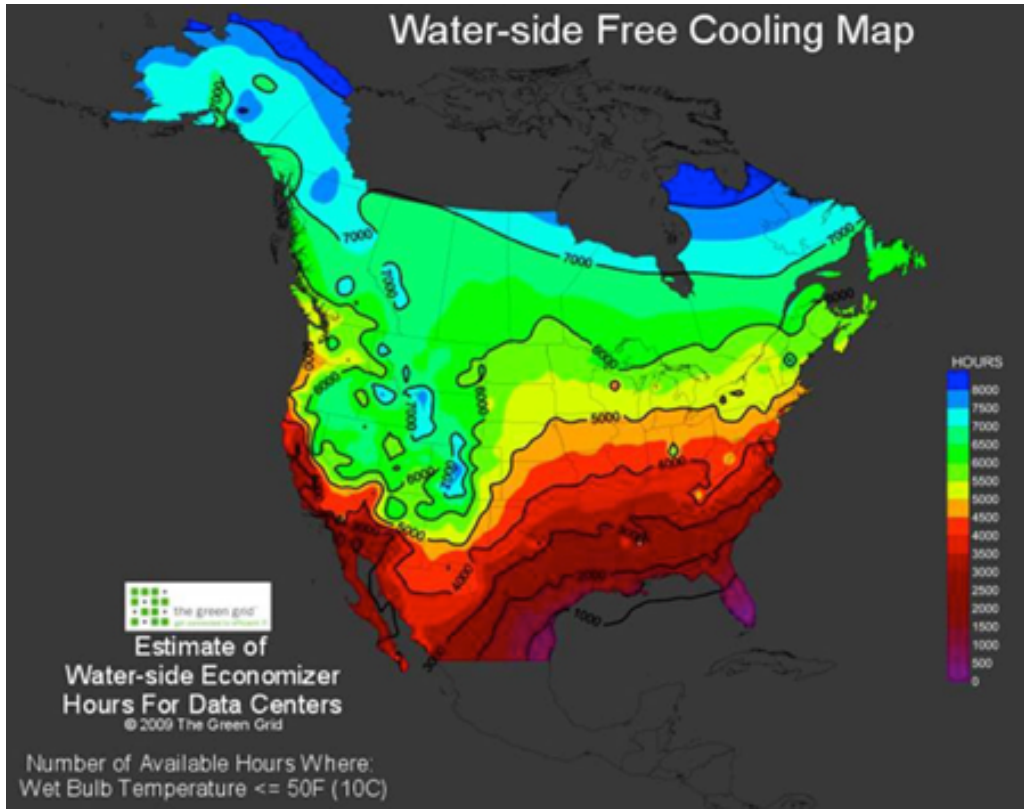
- CAPEX for chillers, trim HX, facilities, PLC
- OPEX with a PUE that can reach 1.4 at ORNL

One opportunity to reduce these CAPEX and OPEX costs involves eliminating the need for chilled water for HPC (but not entirely)

100% Economization -> ~1.1 PUE

Projecting water-side economizer effectiveness

- At ASHRAE W1 liquid cooling class; supply temp to 62.6F, Oak Ridge, TN can take advantage of about 4,000 hours per year where conditions allow the use of water-side economizers
- Increase this supply temperature to 85F, and we can run computers without chillers, year-round.
- Leverage this in our next HPC procurement…



Hours of ideal water-side economizer conditions in the United States. (Courtesy of DOE and The Green Grid)

CORAL-2 Facility Requirement for ORNL

For an HPC system targeted for delivery in 2021

From the DRAFT Statement of Work

"It is assumed that the Offeror will use a facility water supply temperature described by ASHRAE Technical Committee 9.9 Liquid Cooling Guidelines, as Liquid Cooling Class W3, or higher for the compute racks. The CORAL system's compute racks must operate without the direct use of chilled water for all hours of the year given the sites' climatic zones."

ORNL will provide chilled water as part of fresh air pressurization, air conditioning, and relative humidity control of the room. No more than 2% of the system load may escape as parasitic heat to the data center ambient.

ACES and Power Distribution – Why Wye and What Watt?

Jason Hick, LANL

ACES

Alliance for Computing at Extreme Scale

- Partnership since 2008 between SNL and LANL to design, procure, deploy and support capability and advanced technology systems at the Strategic Computing Complex at LANL.

Cielo in 2010, ~1.3 PF

Trinity in 2016, ~40 PF

Coming soon …

Crossroads in 2020

HPC Facility considerations for power distribution

Efficiency (PUE)

Wiring capabilities (substations, switchboards, breakers, 4-wire, 5-wire)

- Cost of wiring
- Labor availability/cost to run wiring
- Floor loading

Voltage available

Location of power (length of cables)

HPC Vendor considerations for power distribution

Efficiency (PUE)

Availability of parts (commodity vs. single supplier)

Cost of parts

Packaging (will it fit in racks/design, is the solution taking up too much space?)

Flexibility (can we support multiple power configurations?)

Safety (ensuring enough clearance to ground)

Pros and Cons of Delta vs Wye

Delta

- Easier to balance a three phase load
- More efficient for stepping down voltages
- Can be electrically ungrounded
- Tough to find faults on the system
- Single phase systems cannot be direct fed
- No neutral

Wye

- Having a neutral makes downstream feeds easier to balance
- No wild legs; all single phase loads can be connected to any point in the wye
- Can deliver any power downstream (e.g. 277/120 single phase, 208 3-phase, 480 single or 3-phase).

Where is industry headed with required power distribution?

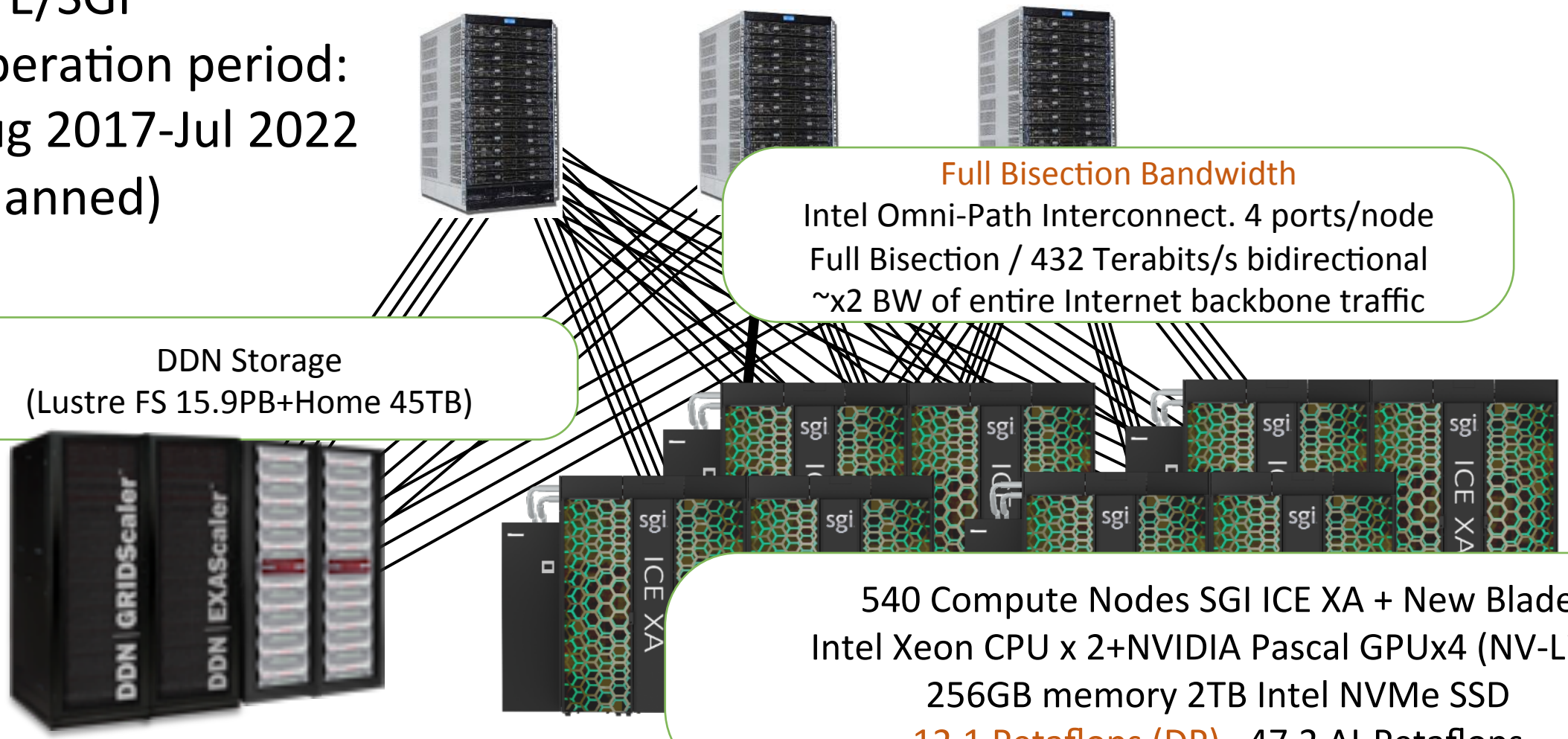
Looking to improve how we express these trade-offs in our procurements.

TCO of Tokyo Tech TSUBAME3.0 Supercomputer

Toshio Endo, Tokyo Institute of Technology

Overview of TSUBAME3.0
BYTES-centric Architecture with 2160 GPUs, 540 nodes

Contractor:
HPE/SGI
Operation period:
Aug 2017-Jul 2022
(planned)

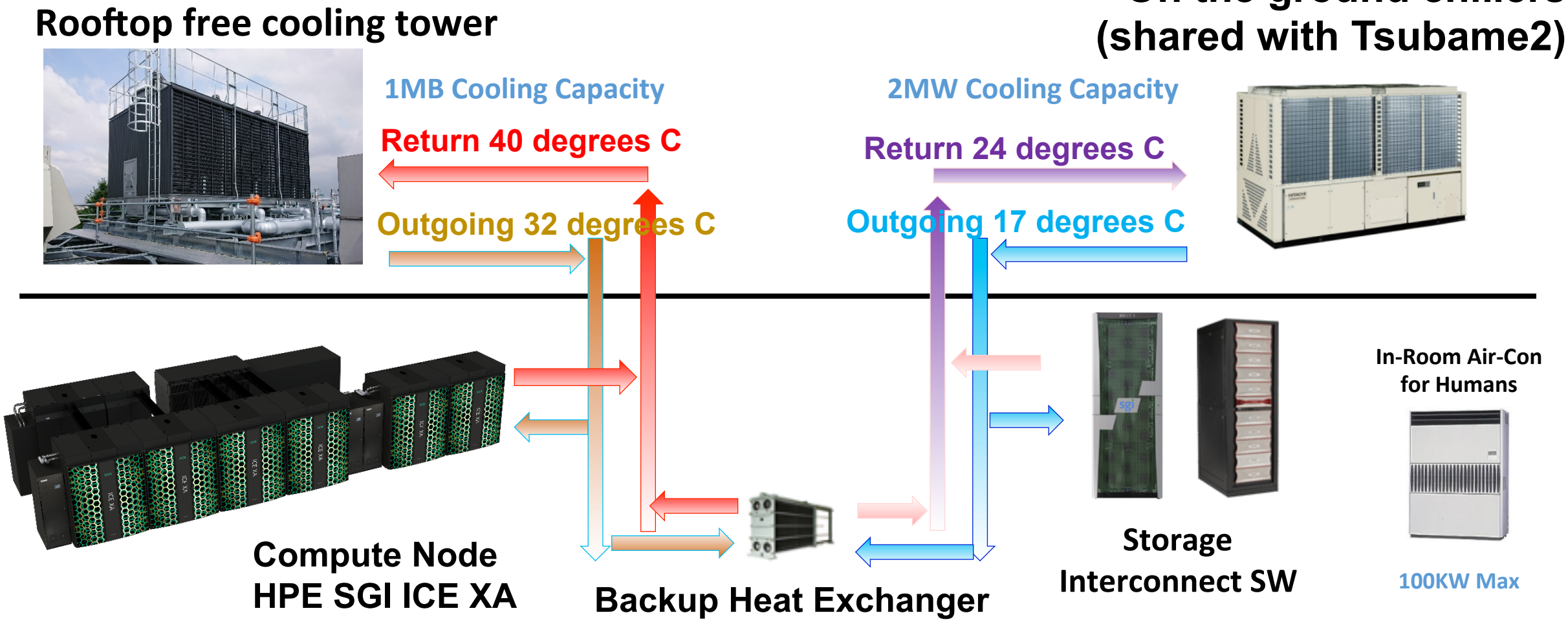


Full Bisection Bandwidth
Intel Omni-Path Interconnect. 4 ports/node
Full Bisection / 432 Terabits/s bidirectional
~x2 BW of entire Internet backbone traffic

DDN Storage
(Lustre FS 15.9PB+Home 45TB)

540 Compute Nodes SGI ICE XA + New Blade
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
256GB memory 2TB Intel NVMe SSD
12.1 Petaflops (DP) , 47.2 AI-Petaflops

Energy Efficient Cooling based on Warm Water



Partial PUE is expected to ~1.03 (estimation by HPE/SGI)

Estimated TCO of TSUBAME3.0

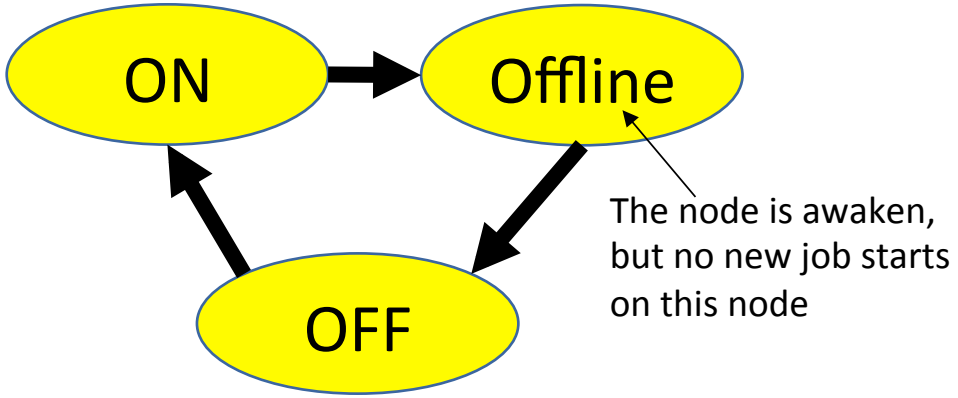
Facility preparation • Floor (1000kg/m2) • Power supply (new 415V) • Waterworks	$~1M
System rental fee • Hardware, incl. guarantee • Cooling facility • License fee for software • Installation	$~7M/yr × 5yr
Electric fee	$0.7~1M/yr × 5yr
Employing system engineers	$~0.3M/yr ×5yr

Energy Saving Efforts in TSUBAME2

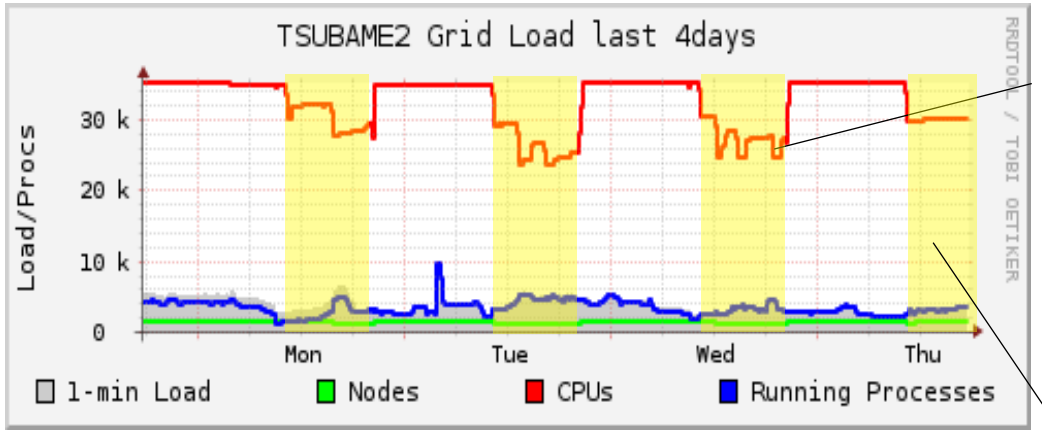
We have developed <u>system power capping method</u>

- The limit is changed in day/night
- Works cooperatively with job scheduler

Transition of modes of each node (simplified)



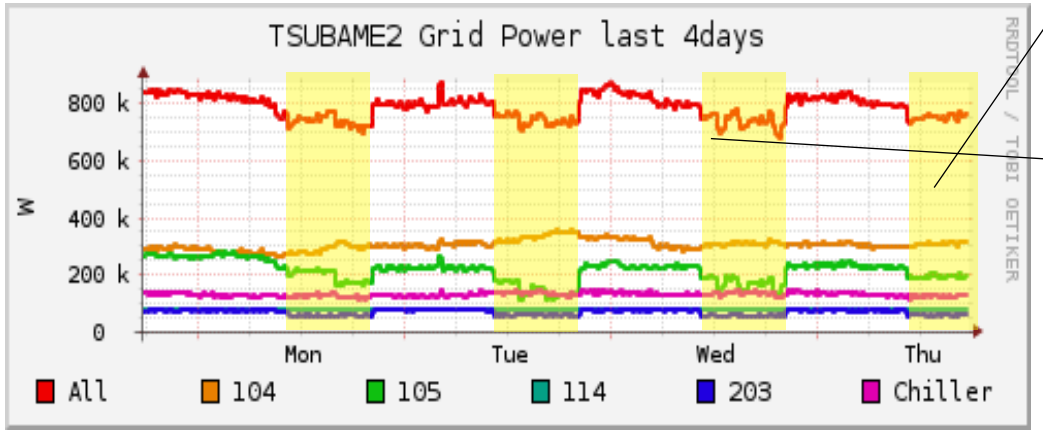
The number of working CPU cores



The number of nodes are controlled dynamically

Daytime

Power Consumption



The power is capped by 800 kW successfully

We are planning to introduce a similar method to TSUBAME3.0 to reduce TCO, but we need to prepare development fee (that will be lower than the first time)

Energy Efficient HPC WG Vendor Neutral Insight on Trends Driving TCO

SC17: Birds of Feather

Tuesday Nov 14, 12:15-1:15 Room 605

Steven Roberts, IBM (presenter)
Steve Martin, Cray
Tim McCann, HPE
Austin Shelnutt, DELL
Toshiyuki Shimizu, Fujitsu
Vinod Kamath , Lenovo
Ludovic Saugé, Bull/Atos

Electrical Trends

- Higher Voltages and High Amperes
  - 2017 showed several 480+v requirements
    - This reduces TCA as it eliminates space and cost of 208/120v step-down transformers
    - May increases TCA requirements due to facility current limiting fuse requirements based on fault current calculations
    - TCO is increased as technicians may be required to have NFPA-70E certification and PPE (Personal protective equipment)

- PUE / iTUE / SERT
  - PUE requests lack uniform definition – in fact some data centers are reporting slightly over 1.0??
  - For discussion – SERT verse iTUE?
  - Recommendation: use The Green Grid method (https://www.thegreengrid.org/en/resources/library-and-tools/20-PUE%3A-A-Comprehensive-Examination-of-the-Metric)

- Hardwired PDUs
  - Saves TCA plug cost ($600 to $1k per rack for PDU pin-and-sleeve plug and receptacle)
  - Moves certification from UL listed to UL Recognized
  - Trade-off is electrical authority required to approve installation and electricians are needed to install/ remove the PDU from service

Cooling Trends

- Medium/Large installs are trending to water cooled
  - Vendors forecasting a 75%/25% split of water to air cooled
  - TCA is 20% more for water-cooled systems
  - Procurements do not typically provide a holistic approach between compute and facility requirements
    - Europe and now some US bids are even moving to warm water (40C)

- Beginning trend to specify # watts of idle power
  - Drives up component costs to turn off function from rack to silicon
  - Why is this important in data centers that boast > 80% utilization?

- Measurement frequency specification for node components over constrains design
  - Proposal: leverage what is already available for system reliability and telemetry requirements and what is truly actionable for rack level energy considerations

Invited Vendor Participants

- Vinod Kamath, Lenovo
- Ludovic Saugé, Bull/Atos
- Toshiyuki Shimizu, Fujitsu
- Tim McCann, HPE
- Steven Roberts, IBM
- Austin Shelnutt, Dell
- Steven Martin, Cray

Q&A:

Has your organization participated in the EEHPC WG on Procurement Considerations?

- Contact: Natalie Bates natalie.jean.bates@gmail.com

Have you seen the 2014 Energy Efficiency Considerations for HPC Procurement Document?

https://eehpcwg.llnl.gov/pages/compsys_pro.htm

SC17 EE HPC WG Primer

https://eehpcwg.llnl.gov/pages/conf_sc17a.htm

Energy Efficient HPC Working Group
Super Computing Conference 2017

Home | Webinars | Links & Events | Meetings | Groups | Conferences | Contacts | Join Us | Publications

- ISC10 Thru 13 Presentations
- ISC 14 Presentations
- SC10 Presentations
- SC11 Presentations
- SC12 Presentations
- SC13 Presentations
- SC14 Presentations
- SC15 Presentations
- SC17 Conference

Eighth Annual Workshop for the Energy Efficient [illegible] E HPC WG)

Abstract:

This annual workshop is organized by the Energy Efficient HP[illegible] Work[illegible] [illegible]es th[illegible] gap between facility and IT system with regards to energy efficiency analysis and improvements [illegible]or sustainable exascale computing, powe[illegible] and energy are a main concern, which can only be addressed by taking a holistic view combi[illegible]ng the HPC facility, HPC system, HPC sy[illegible]em software, and the HPC application

Other EE HPC WG Technical Sessions

- **Research Poster: Tuesday- Thursday 8:30-5:00**
  - "Global Survey of Energy and Power-aware Job Scheduling and Reso[illegible]
- **Birds of Feather: Tuesday 12:15-1:15 Room 605**
  - "Total Cost of Ownership and HPC System Procurement"
    - Energy Efficiency Considerations for HPC Procurement Documents: 2017
    - EE HPC WG Liquid Cooling Controls Team Whitepaper. June 11, 2017
    - Systematic Approach For Universal Commissioning [illegible] For Liquid-Cooled Systems
- **Birds of Feather: Tuesday 5:15-7:00 Room 712**
  - "Power API, GEOPM and Redfish: Open Interfaces for Power/Energy Measure[illegible] Control"
    - EE HPC WG Liquid Cooling Controls Team Whitepaper. June 11, 2017
- **Birds of Feather: Wednesday 12:15-1:15 Room 501/502**
  - "State of the Practice: Energy and Power Aware Job Scheduling and Resource Management"
    - Whitepaper: A Survey of Energy and Power Aware Job Scheduling and Resource Man[illegible] Supercomputing Centers
- **Birds of Feather: Wednesday 5:15-7:00 Room 402/403/404**
  - "Green500: Trends in Energy Efficient Supercomputing"
    - Energy Efficient High Performance Computing Power Measurement Methodology (vers[illegible]
- **Panel: Friday 10:30-noon Room 201/203**
  - "Energy Efficiency Gains from Software: Retrospectives and Perspectives (Version 2.0 RC 1.0)"
    - This panel will explore what HPC software capabilities were most helpful over the past years in improving HPC system energy efficiency? It will then look forward; asking in what layers of the software stack should a priority be put on introducing energy-awareness; e.g., runtime, scheduling, applications? What is needed moving forward? Who is responsible for that forward momentum?
    - Moderator/Panelists: Dan Reed, Satoshi Matsuoka, Sadaf Alam, Bill Gropp and John Shalf

Links to relevant BoF Documents

SC17

Denver, CO | hpc connects.

E2 HPC Working Group

Sustainably Supporting Science

through committed community action