Monitoring and Controlling Power Usage on Cray XC30

Wilfried Oed
Principal Engineer

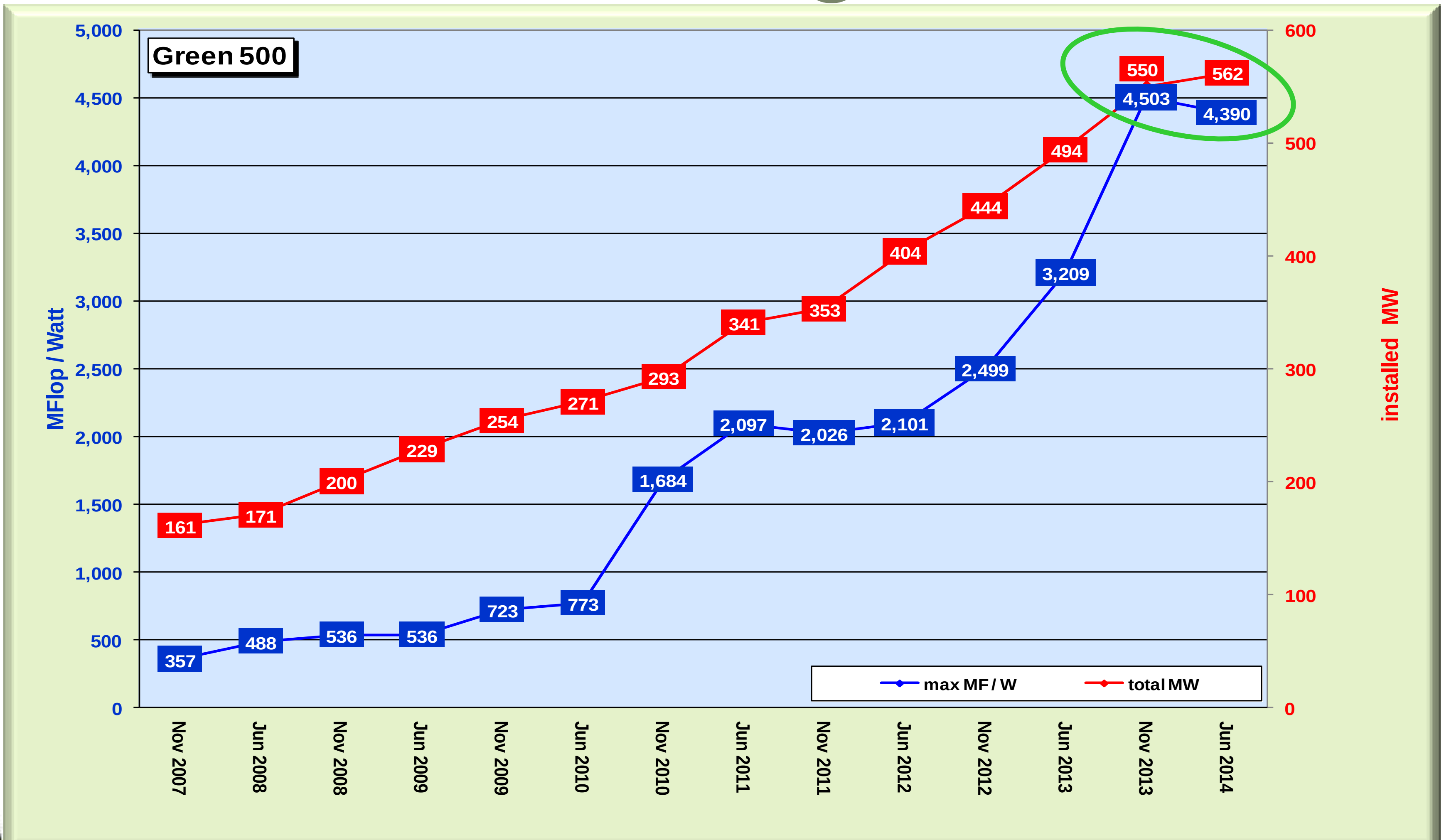
September 2, 2014

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Power Consumption – Quo Vadis ?

- MFLOPS per Watt continues to improve **???**
  - Emphasizes pure floating-point (HPL)
  - The goal for EXAflop is 20 MW => 50 GF / W
- Slowed down rise on installed power consumption



A Look at the current Green500

Green list	Green500 rank	Top500 rank	R_{max}	total power	Mflop / Watt	total cores	R_{Peak}
Nov 13	1	311	125,100	28	4,503	2,720	217,664
Jun 14	1	437	151,800	35	4,390	2,720	217,824

Performance per Watt drops

Apparently same system

R_{max} increases by 21.3% and power by 24.5%

Would have fallen out of the Top500

Green list	Green500 rank	Top500 rank	R_{max}	total power	Mflop / Watt	total cores	R_{Peak}
Jun 14	1	437	151,800	35	4,390	2,720	217,824
Jun 14	2	201	191,100	53	3,632	5,120	367,565
Jun 14	3	165	277,100	79	3,518	4,864	364,288
Jun 14	4	421	153,600	44	3,459	3,036	209,880
Jun 14	5	6	5,587,000	1,754	3,186	115,984	7,788,853
Jun 14	6	185	254,900	81	3,131	5,720	384,124
Jun 14	7	181	260,300	86	3,020	5,376	333,481
Jun 14	8	13	2,739,000	928	2,952	76,032	5,735,685
Jun 14	9	11	3,003,000	1,067	2,813	62,640	4,006,350
Jun 14	10	455	146,241	55	2,678	3,264	208,760

The leading systems in the Green500 are rather small

2 systems are closely related to their Top500 position

Only one system is under the first 10 in both Green500 **and** Top500

Energy Efficient Supercomputing

- The energy required to produce scientific results can be characterized by the following simple equation

$$\text{Energy to result} = \underbrace{\text{Power Draw Per Node} \times \text{Node Hours Used}}_{1} \times \underbrace{PUE}_{\text{1-Point something}}$$

In practice, this is probably more important…

Most of the focus on “green computing” has been here

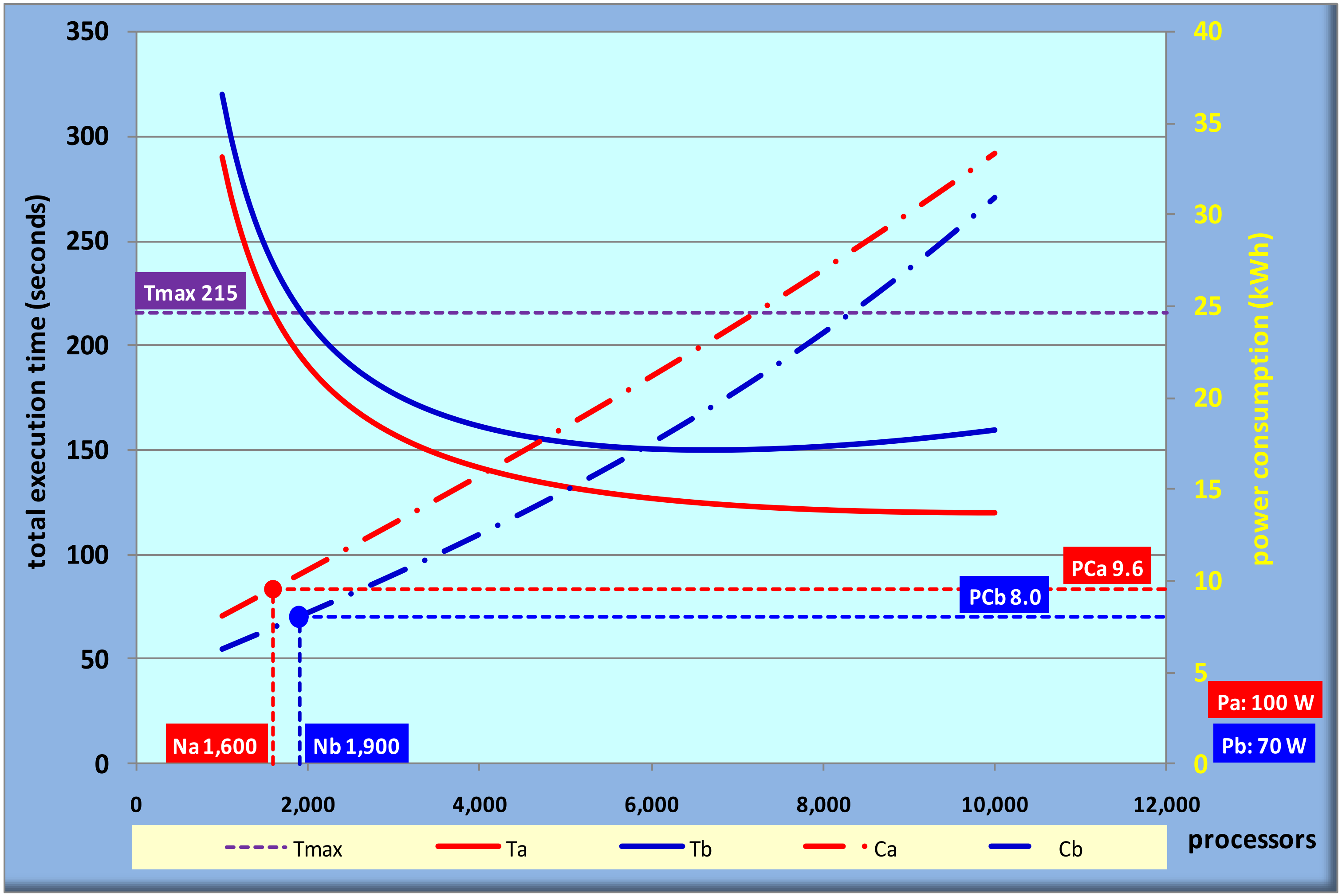
Solving a given Problem in T_{max}

Recap: http://www.ena-hpc.org/2011/talks/oed-slides.pdf

- The lower power processor always requires less power on a per core basis
- At low core counts (higher time to solution) the lower powered processor is more energy efficient , as only a few additional cores are required

Note: this is an arbitrary example for demonstrating certain effects

neither based on actual systems nor applications



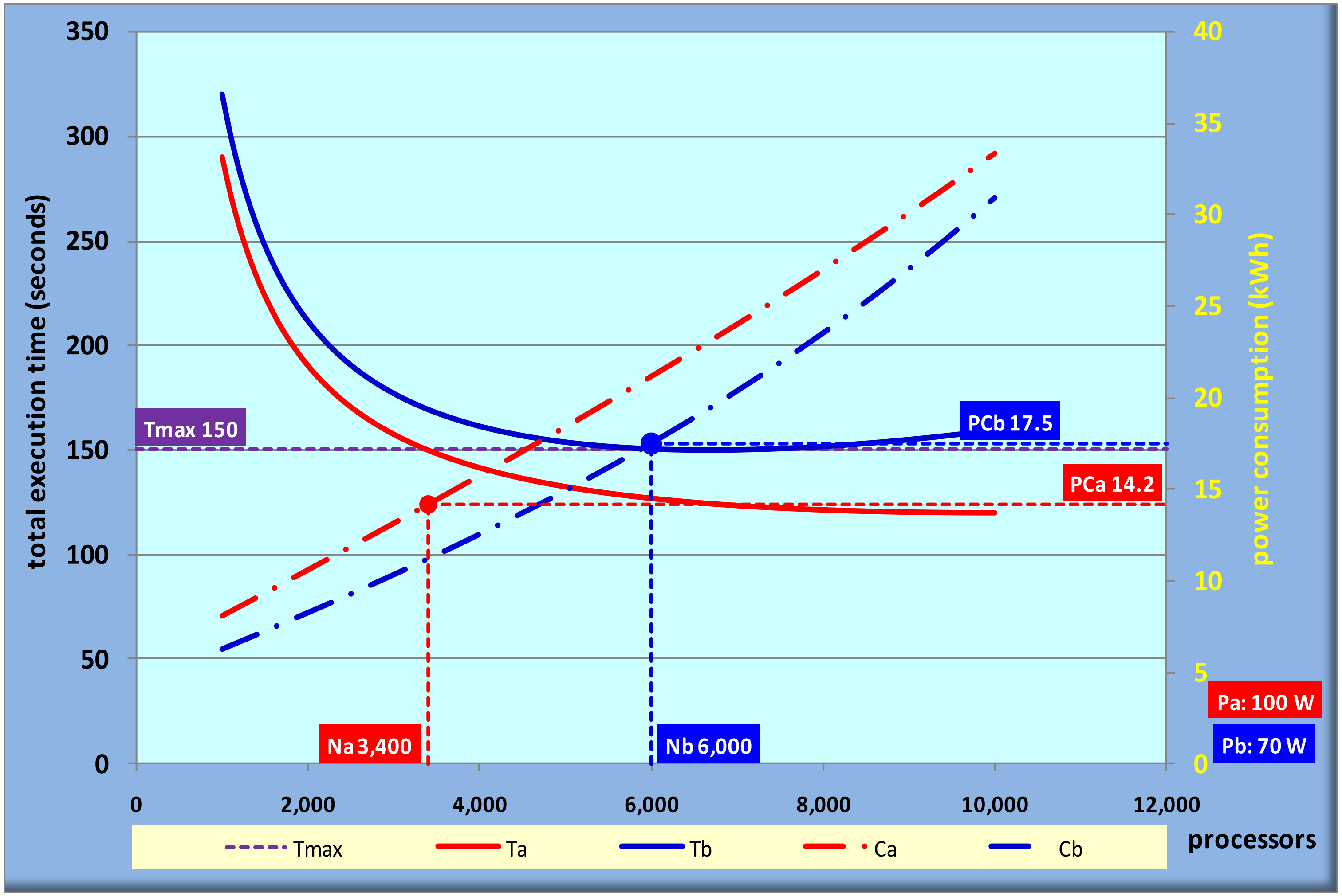
Solving a given Problem in (a lower) T_{max}

Recap: http://www.ena-hpc.org/2011/talks/oed-slides.pdf

- The lower power processor always requires less power on a per core basis
- At higher core counts (lower time to solution) the lower powered processor is less energy efficient, as far more cores are required

Note: this is an arbitrary example for demonstrating certain effects

neither based on actual systems nor applications



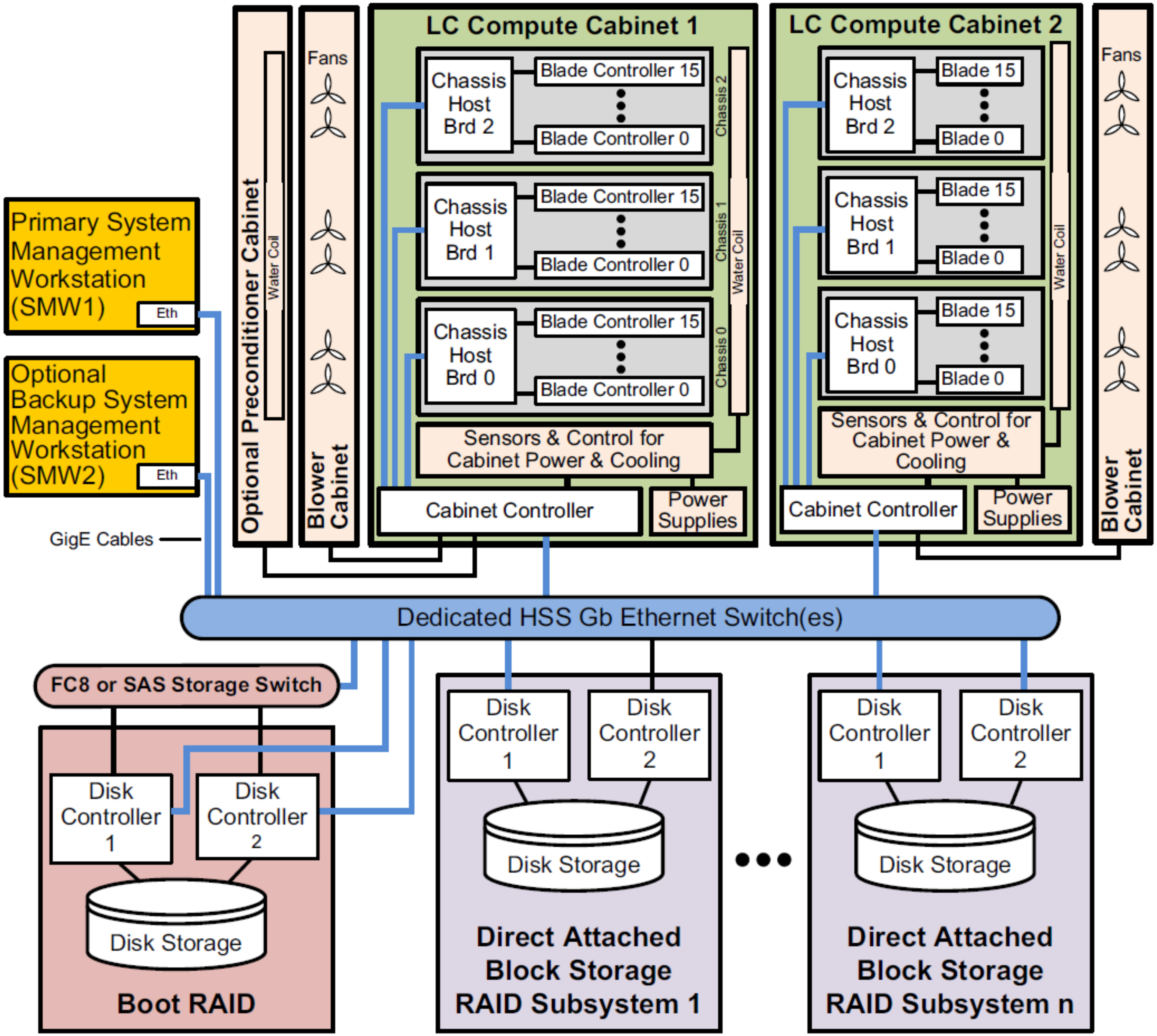
Cray XC30 In-Band Monitoring

- System /sys/cray/pm_counters

```
/sys/cray/pm_counters/accel_energy:24675886 J
/sys/cray/pm_counters/accel_power:22 W
/sys/cray/pm_counters/accel_power_cap:0 W
/sys/cray/pm_counters/energy:71224823 J
/sys/cray/pm_counters/freshness:4516770
/sys/cray/pm_counters/generation:9
/sys/cray/pm_counters/power:62 W
/sys/cray/pm_counters/power_cap:425 W
/sys/cray/pm_counters/startup:1396011015159068
/sys/cray/pm_counters/version:1
```

- Intel RAPL counters
  - PAPI
  - CrayPat
- In-Band monitoring is intrusive and non-scalable
  - Applicable for code tuning

Cray XC30 Hardware Supervisory System (HSS)



- Integrated, independent (Out-Of-Band) hardware und software system
  - SMW
  - Cabinet controllers
  - Blade controller (one per blade)
  - Several thermal and voltage sensors
  - Hardware and Software Status
  - System Startup / Shutdown
- Power monitoring
  - 1 Hz resolution over **all** blades (nodes) of the **entire** system
  - Non-intrusive
  - Data archived on the SMW in the PMDB

Cray XC30 **Out-Of-Band** Monitoring

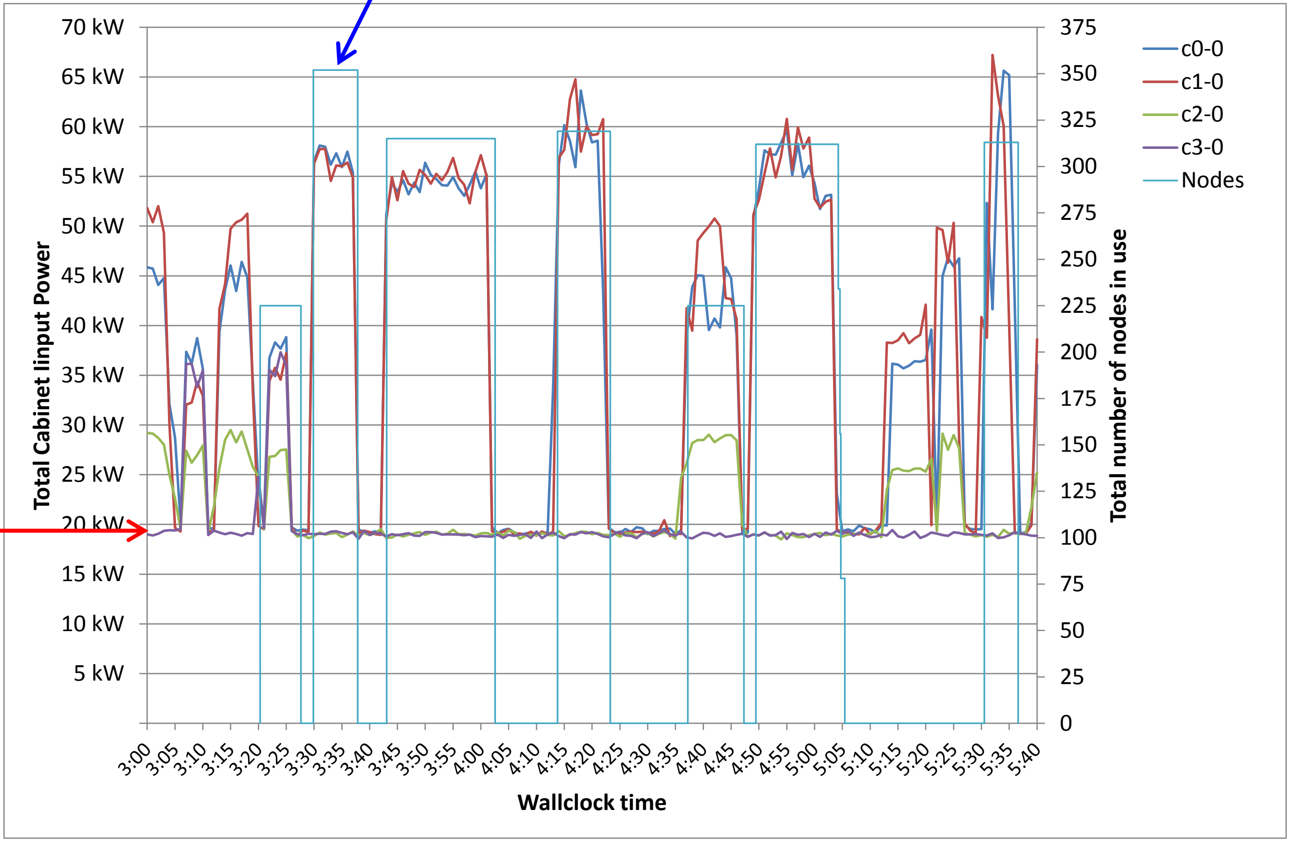
- Out-of-Band monitoring is non-intrusive and scalable
  - Can run it in production
  - Enables energy accounting
- System Environmental Data Collection (SEDC)
  - Voltage, current, temperature, pressure, fan-speed, …
  - Readings updated once per minute
  - Data written to flat-files on SMW
- High-speed power/energy data collection
  - Data Cabinet, Blade, Node, and [Accelerator] data
  - Blade level data collection at 10 Hz
- Power Management Database (PMDB)
  - Cabinet-level Power (+blowers)
  - Blade- and Node-level data at 1 Hz

Power Monitoring

- Out of band monitoring
  - Sensors on every blade
  - Data aggregation in cabinet
  - Data collection in SMW
  - Logging to database
  - Query interface
- Detailed data available on Power Management Database (PMDB)

number of nodes used

idle power



P-state Control Example

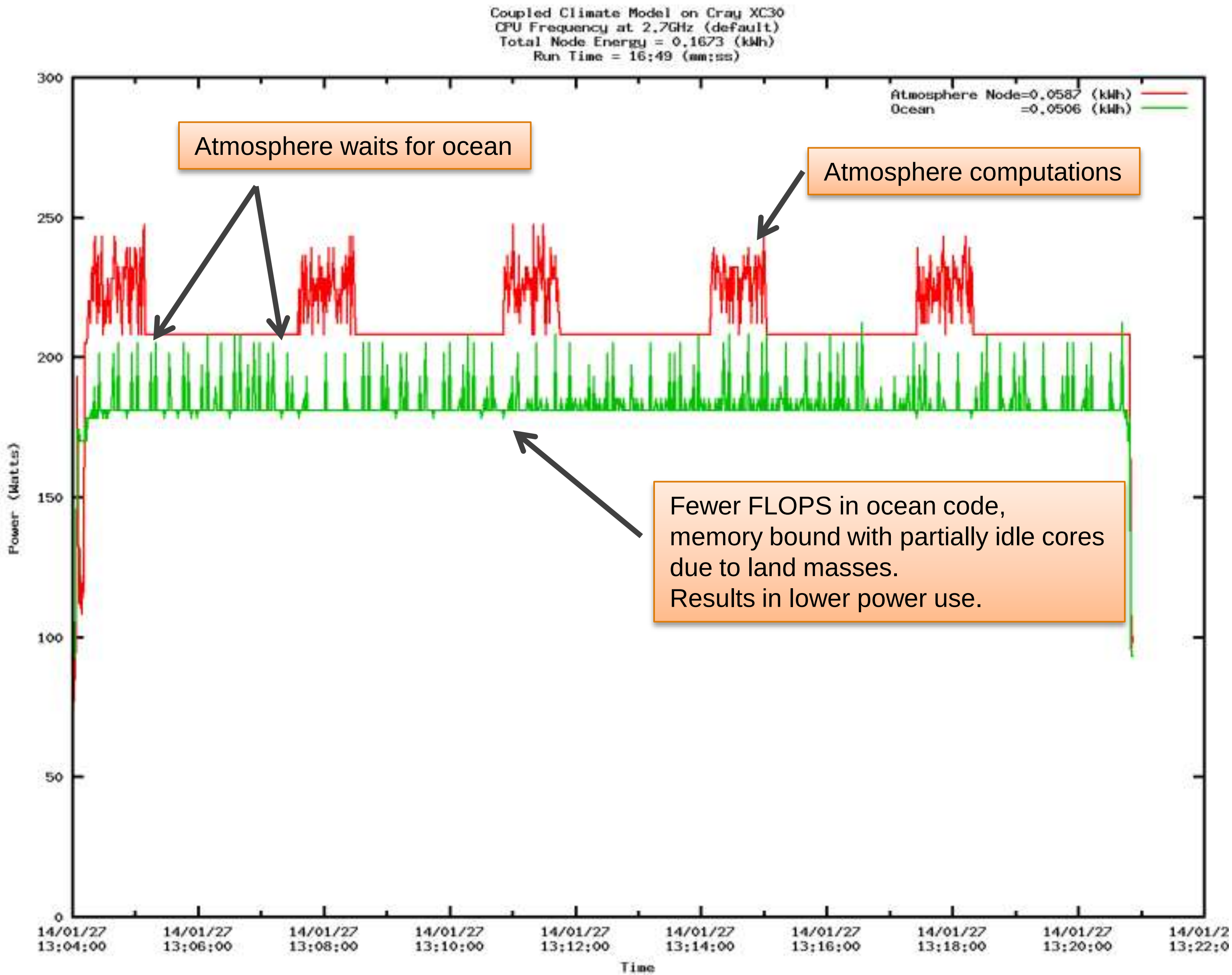
- Coupled climate code with imbalanced atmosphere and ocean components was run on **2.7 GHz** Ivy Bridge nodes and, since the atmosphere portion was spending a lot of time waiting on the ocean, atmosphere nodes were **capped at 2.3 Ghz**
- aprun -n 24 ./ATM.exe : -n1 -N1 env OMP_NUM_THREADS=12 ./OCN.exe

```
  APID    | Joules |            KWh           |      Runtime
---------+--------+-------------------------+-----------------
 5615481 | 607712 | 0.168808888888888889 | 00:16:49.698598
 Component   | NID | Joules
------------+-----+--------
 c3-0c2s3n0 | 716 | 210805
 c3-0c2s3n1 | 717 | 213097
 c3-0c2s5n2 | 726 | 183810
```

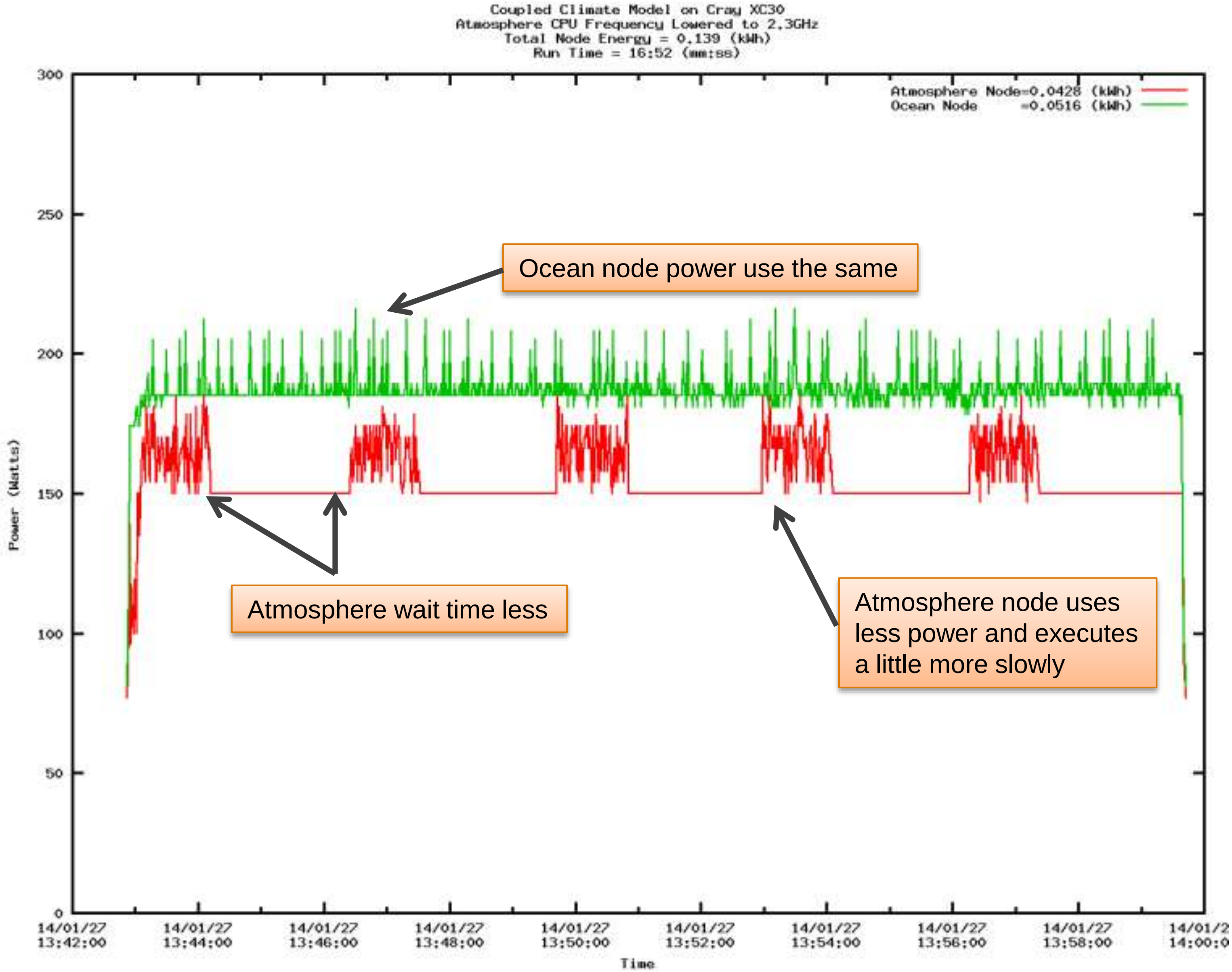
- aprun **--p-state 2300000** -n 24 ./ATM.exe : -n1 -N1 env OMP_NUM_THREADS=12 ./OCN.exe

```
  APID    | Joules |            KWh           |      Runtime
---------+--------+-------------------------+-----------------
 5615548 | 509155 | 0.141431944444444444 | 00:16:51.833913
 Component   | NID | Joules
------------+-----+--------
 c3-0c2s4n3 | 723 | 164821
 c3-0c2s5n0 | 724 | 156068
 c3-0c2s5n1 | 725 | 188266
```

Power Monitoring



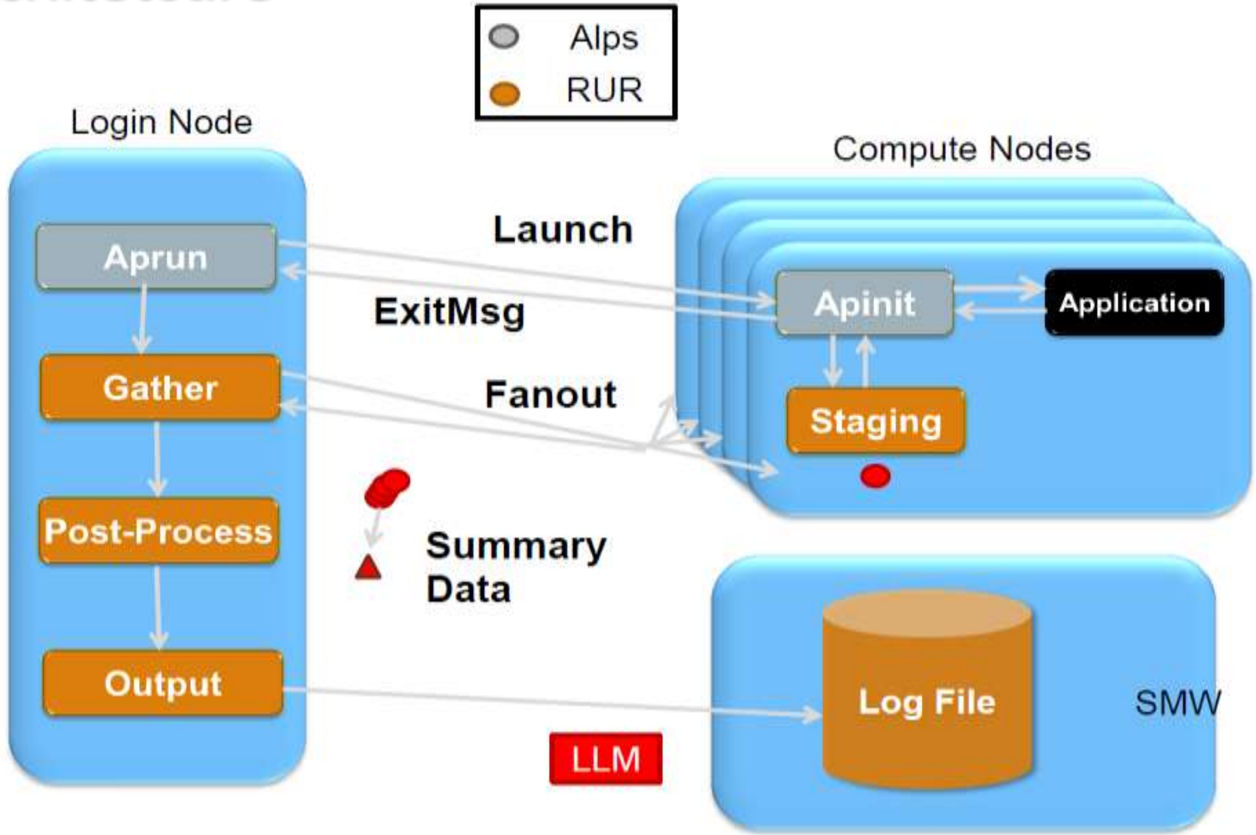
Power Monitoring



RUR – Resource Utilization Reporting

- RUR is a **scalable** accounting tool for gathering diverse usage data, and reporting to systems administrators
- RUR is a **plugin-based** architecture, allowing the collected data to grow
  - Data plugins change what data is collected, output plugins change where the collected data is written
- RUR Phases
  - **Data Staging** on compute nodes in two phases
    - Collect data before the application run
    - Collect data after the application run
    - The staged data is the delta of the two
    - Plugin-specific
  - **Data collection** on the login/mom node from all compute nodes
    - In the future can also be launched in batch epilogue
    - Uses a resilient fanout tree, with a timeout
  - **Post-processing** built-in support for sum, min, max, mean, and histogram operations
  - **Logging/storage** of the output

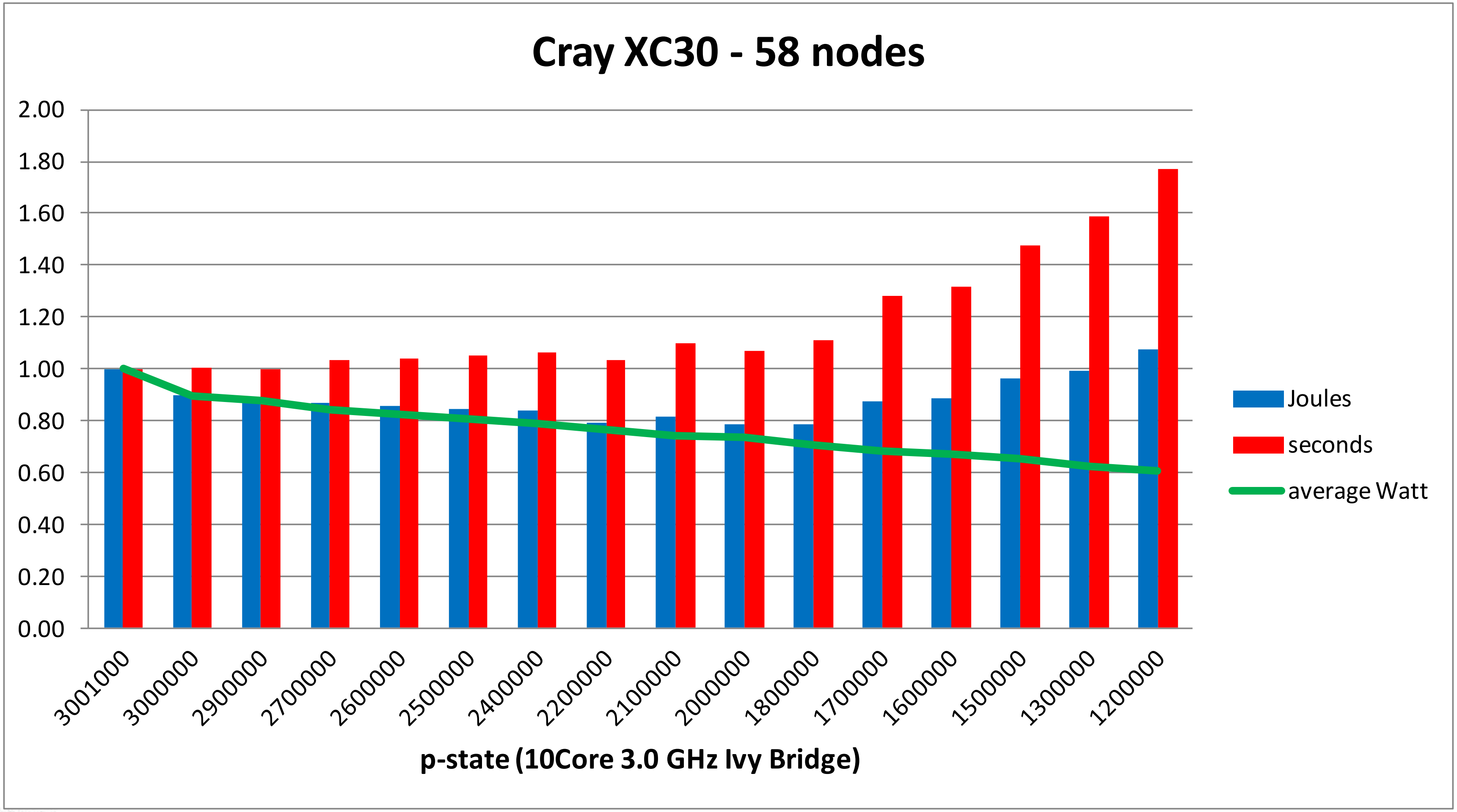
RUR Architetcure



- **LLM**: logs statistics to smw: /var/opt/cray/log/current/messages-date
- **File**: writes statistics to text file writeable by MOM node (path set in config file)
- **User**: writes statistics to text file in user's home directory
  - May redirect to other location with write permissions

Running at various p-states (monitoring by RUR)

- Reducing energy consumption by lowering the clock frequency
  - Runtime expansion initially acceptable
  - This is application dependent !



Recent Work presented at CUG 2014

Cray XC30 Power Monitoring and Management

Steven J. Martin
Cray Inc.
Chippewa Falls, WI USA
stevem@cray.com

Matthew Kappel
Cray Inc.
St. Paul, MN USA
mkappel@cray.com

User-level Power Monitoring and Application Performance on Cray XC30 Supercomputers

Alistair Hart, Harvey Richardson
Cray Exascale Research Initiative Europe
King's Buildings
Edinburgh, UK
{ahart,harveyr}@cray.com

Jens Doleschal, Thomas Ilsche, Mario Bielert
Technische Universität Dresden,
ZIH
Dresden, Germany
{jens.doleschal,thomas.ilsche,mario.bielert}@tu-dresden.de

Matthew Kappel
Cray Inc.
St. Paul MN, USA
mkappel@cray.com

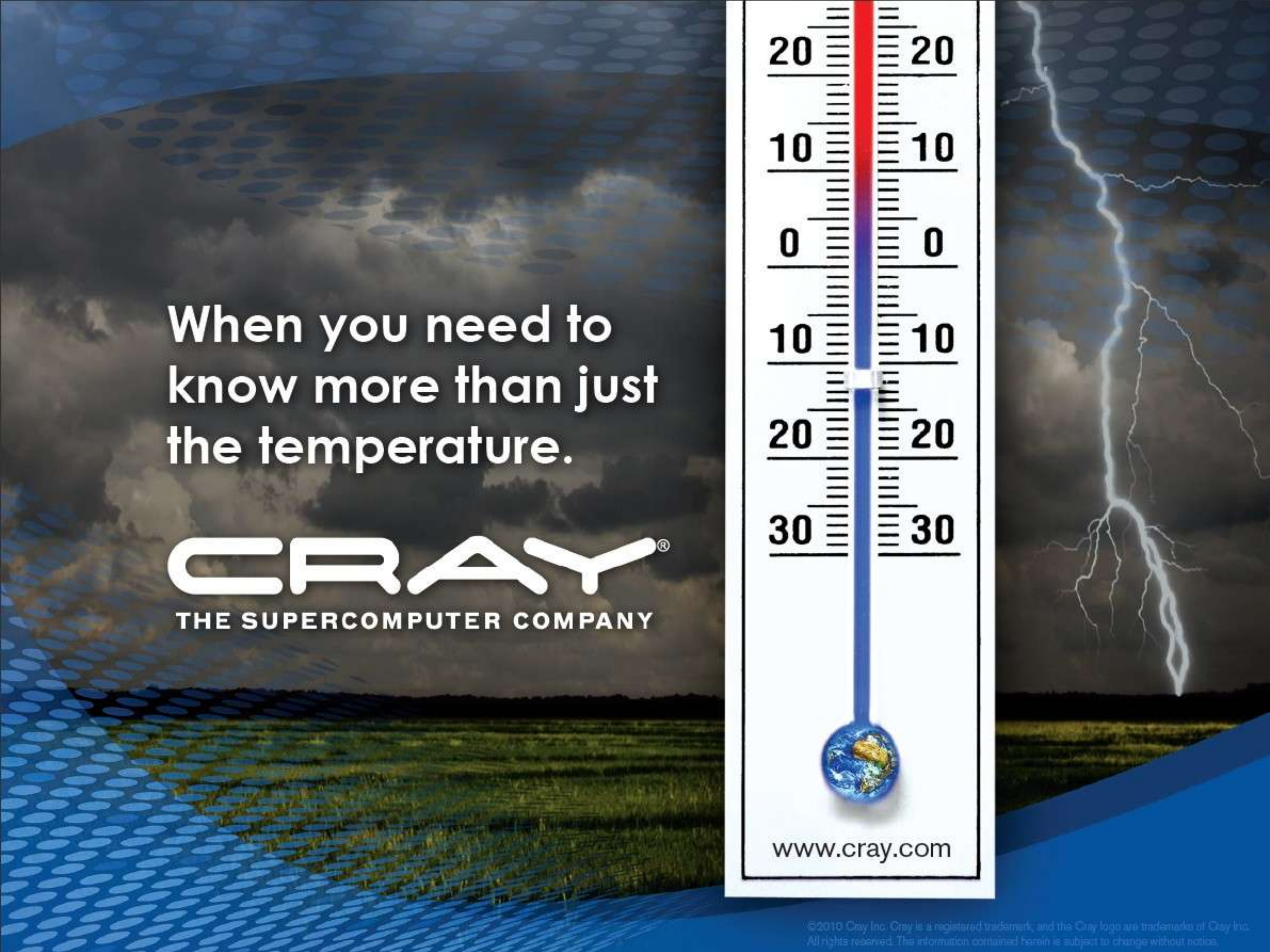
First Experiences With Validating and Using the Cray Power Management Database Tool

Gilles Fourestey*, Ben Cumming*, Ladina Gilly*, and Thomas C. Schulthess*†‡
* Swiss National Supercomputing Center, ETH Zurich, 6900 Lugano, Switzerland
† Institute for Theoretical Physics, ETH Zurich, 8093 Zurich, Switzerland
‡ Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
Email: {gilles.fourestey, ben.cumming, ladina.gilly, thomas.schulthess}@cscs.ch

- CUG Proceedings are publicly available typically 6 months after the meeting at https://cug.org/

Summary

- Improving energy efficiency
  - The biggest innovations will have to come from technology
    - Remember: the goal for EXAflop is 20 MW or 50 GF / W
  - Apply power capping where applicable
    - But beware, overall power consumption may end up to be higher due to more cores or longer execution time
  - Run a proper mix
    - Avoid peak usage by energy aware scheduling
- Monitoring and conclusions
  - Required is the ability to measure performance and energy consumption on an application level
  - **TUNE** your application (a truck has good mileage if fully loaded)
  - Scalability is a decisive factor on time to solution and consequently on power efficiency



When you need to know more than just the temperature.
20 20
10 10
0 0
10 10
20 20
30 30
www.cray.com
THE SUPERCOMPUTER COMPANY
©2010 Cray Inc. Cray is a registered trademark, and the Cray logo are trademarks of Cray Inc.
All rights reserved. The information contained herein is subject to change without notice.