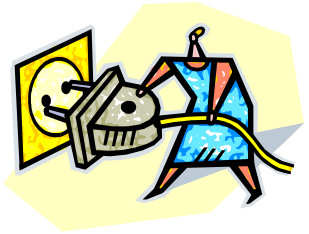
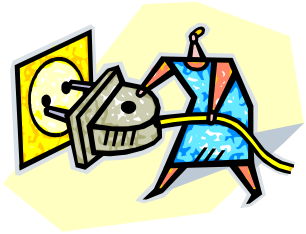


DOE Best Practices Workshop
Power Management
San Francisco, Sept. 28-29, 2010

Breakout 1c:
Power-aware OS features and scheduling

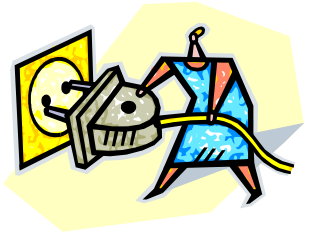
James H. Laros III
jhlaros@sandia.gov
Sandia National Labs
Marcus Epperson (SNL), Natalie Bates (EEHPCWG)



Breakout participants

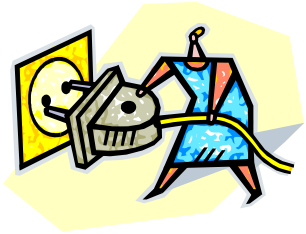
lucky ~~13~~ 12

- Jim Laros (Lead)
- 1st - Marcus R. Epperson(Co-Lead, Notetaker)
- 1st - Natalie Bates (Notetaker)
- 1st - Jacques Noe
- 1st - Jim Garlick
- 1st - Mark A. Grondona
- 1st - Tisha Stacey
- 1st - Mike Lang
- 1st - Myra Branch
- 1st - Mary Zosel
- 1st - Kimberly C. Cupps
- 1st - Michael Knobloch



Goals

- Foster a shared understanding of power management issues in the context of HPC centers.
- Identify top challenges and open issues.
- Share best practices and lessons learned.
- Establish communication paths for managerial and technical staff at multiple sites to continue discussion on these topics.
- Discuss roles and benefits of HPCC stakeholders.
- Present findings to DOE and other stakeholders.
- CAPTURE YOUR THOUGHTS!!

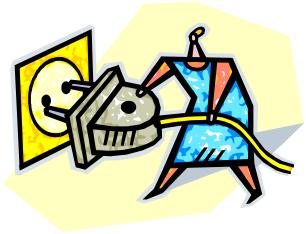


Topics for Discussion

This breakout session will focus on both hardware and software issues related to achieving power efficiency. Example issues include but are not limited to:

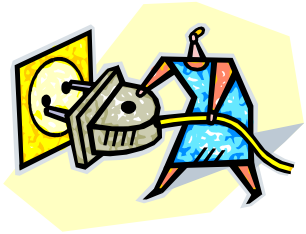
- Advanced Power Management (APM) features available on current and future architectures (frequency scaling, sleep/low power states, dynamic voltage transitions);
- Available OS interfaces to APM features;
- OS techniques to leverage APM features (independent of applications);
- OS interfaces exposed to enable higher level exploitation of APM features;
- OS abstraction of underlying APM features;
- What, if any, features to expose directly to the application;
- Power/performance trade-offs;
- Power aware scheduling;
- Scheduling benefits and impacts of power aware scheduling.

These issues are largely interdependent and must be considered from the system perspective. In addition, power efficiency issues and techniques necessary for HPC-class platforms likely differ greatly from commodity approaches developed for PC and enterprise class platforms. Our goal will be to identify obstacles and opportunities specific to HPC in this emerging area.



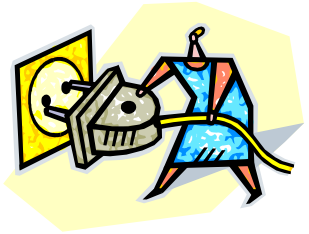
Outline of Breakout Discussion

- Introductions
- Define our scope
  - Capability and Capacity?
  - Can they be approached in the same way? Overlap?
- Project Descriptions
  - Energy-Efficient Cluster Computing (eeClust) http://www.eeclust.de - Michael Knobloch
  - Fit4Green http://www.fit4green.eu - *Michael Knobloch*
  - Less Watts – http://www.lesswatts.org (Linux specific?) - *All*
  - Ongoing work at Sandia Labs – *James Laros*
  - Others? - *All*
- Breakout Slides – Cross Cut Questions
  - Keep in mind….
    - Capability vs. Capacity – does it make a difference
    - Where can our community have the largest impact
    - What is most important to us? Trade-offs possible?



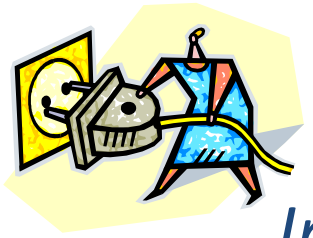
Poll

- Should OS play a role in power management, savings etc?
  - A: yes
- Does Linux provide what we need?
  - A: no, might be able to modify
- What areas should future OS's target?
  - A: power during idle
  - A: deterministic power management (don't introduce jitter)
- Is there an acceptable power-performance trade-off?
  - A: yes, what the trade-off is might be a harder to define question
- Others?



Sandia Power Project

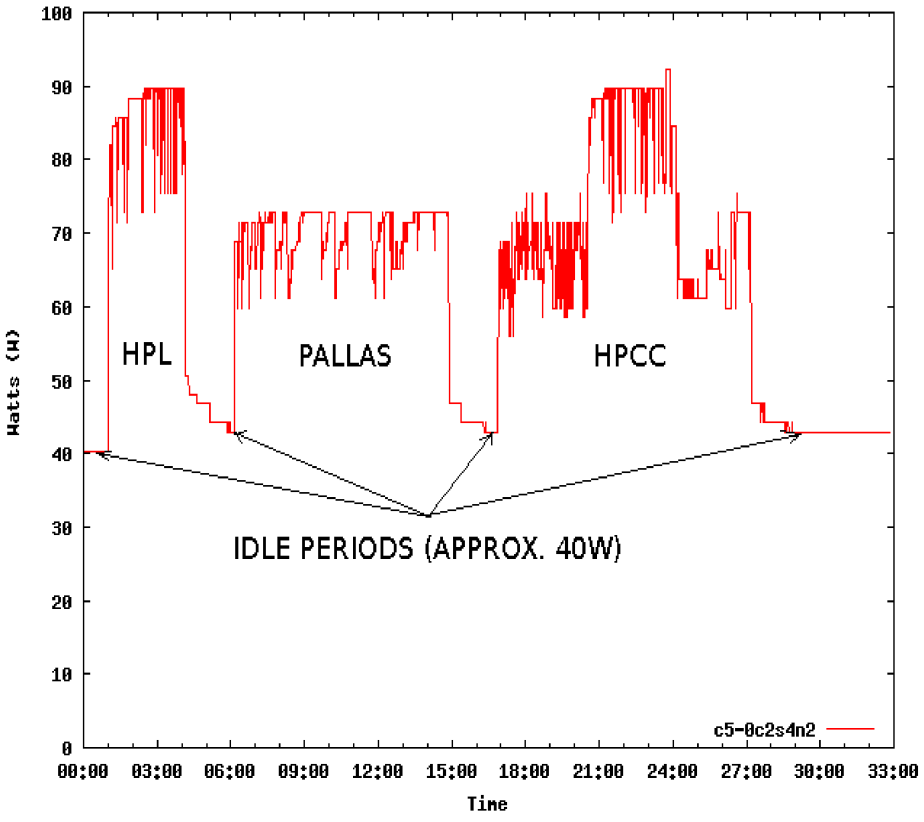
- Lots of thought put into this name…

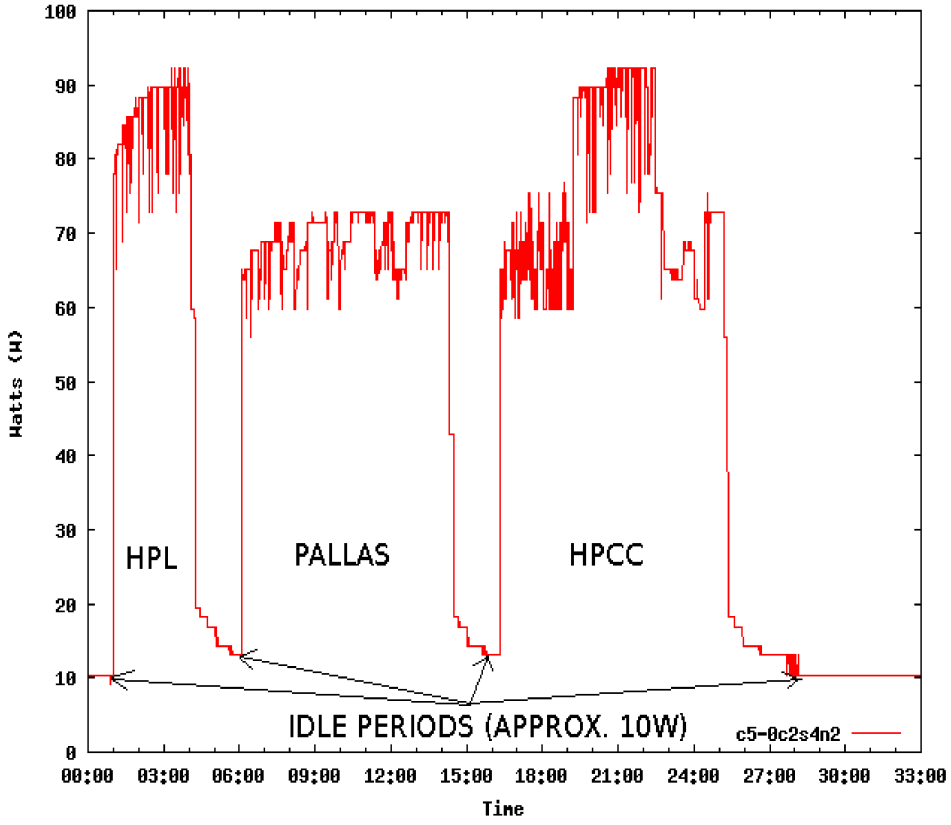


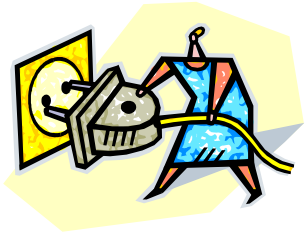
Sandia power Project

Initial Target: CNL vs. Catamount IDLE Draw

low hanging fruit

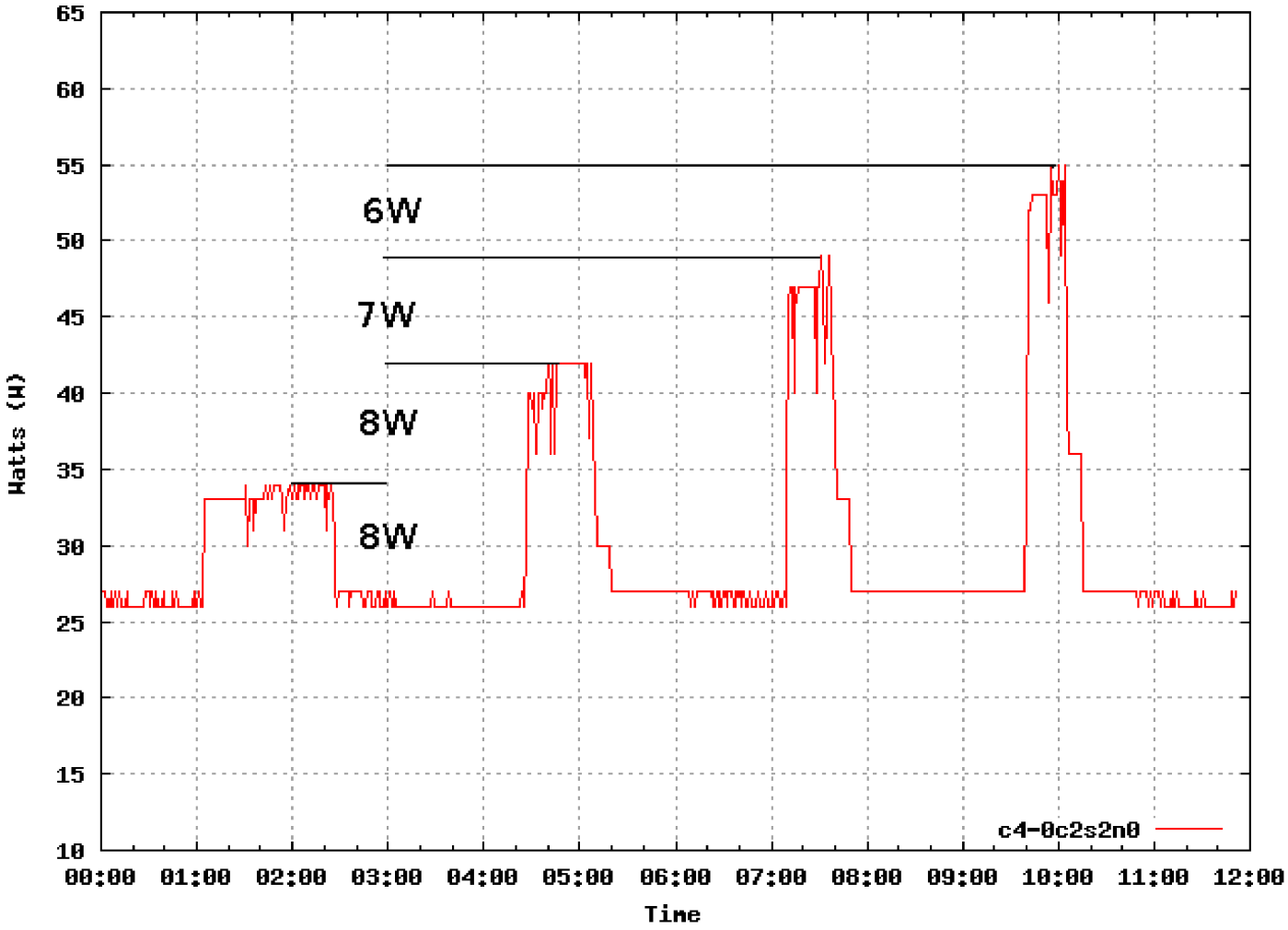


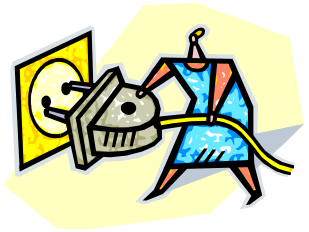




Sandia Power Project

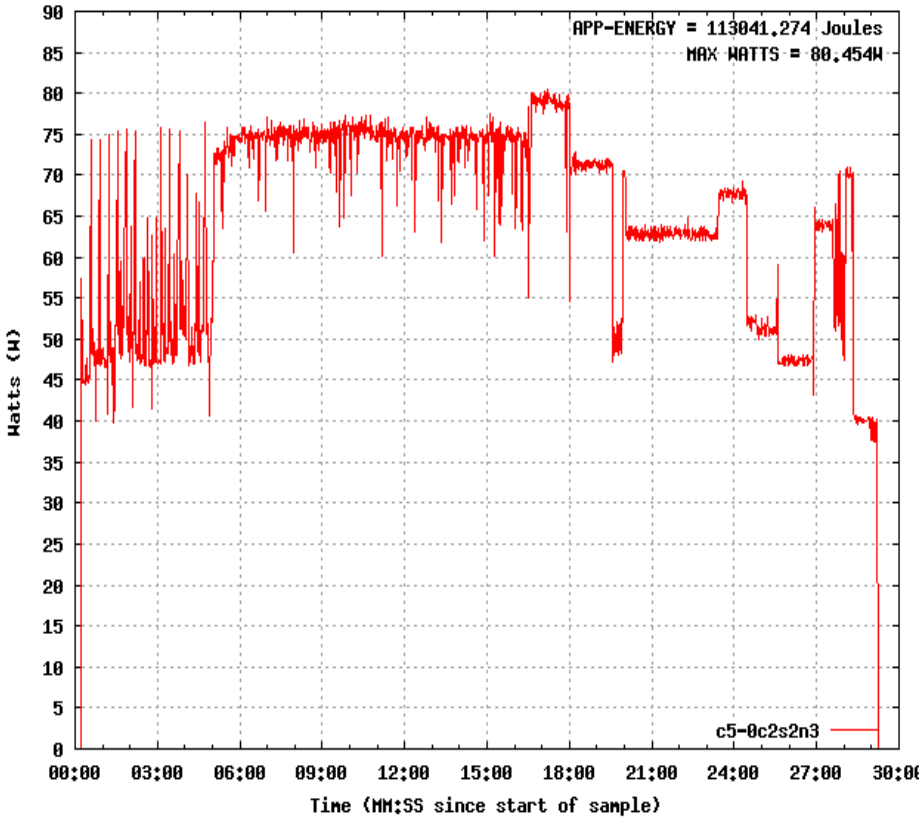
Halt individual cores when not in use…

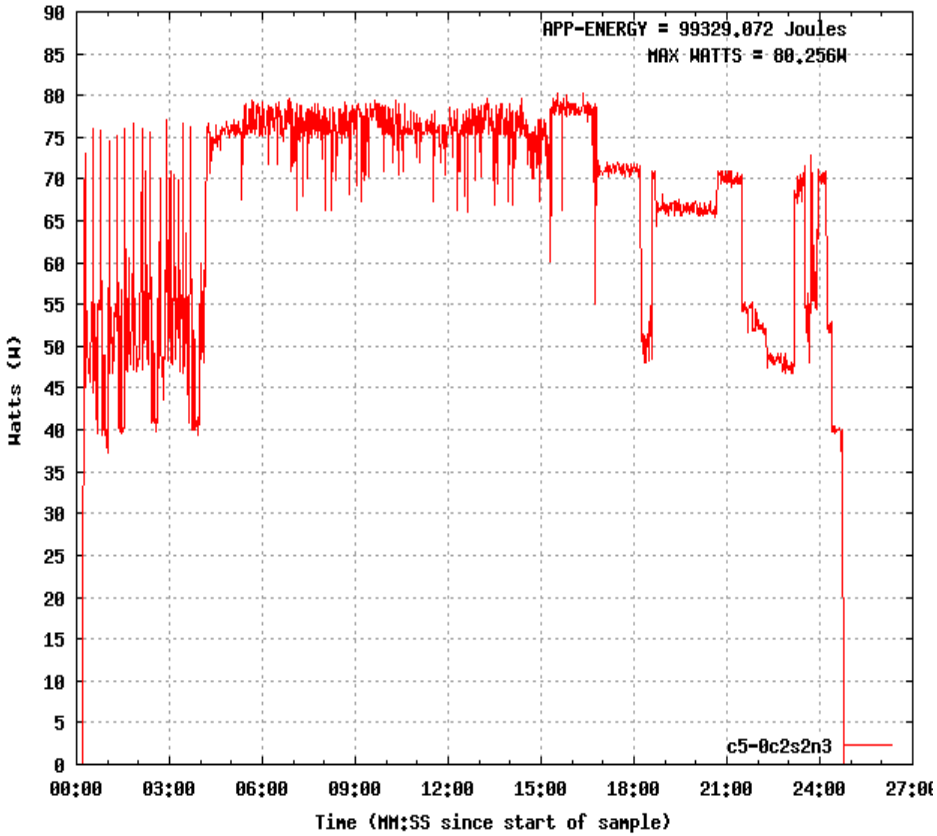




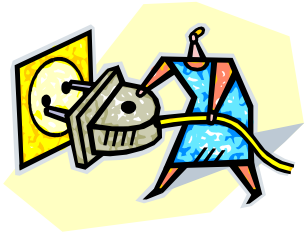
Sandia Power Project

Application Energy Signatures





- HPCC
  - 16% faster on Catamount, 13% more energy used on CNL
- Obvious but important, longer runtime = more power
- How do other things that affect performance affect power use?
  - Noise, for example

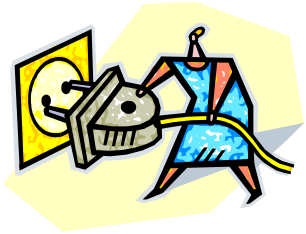


Sandia Power Project

Impact of Noise on Power

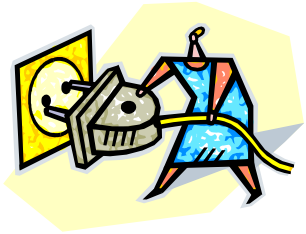
Noise	Freq	Duration	Diff Run-time	Diff App Energy (AVG)
2.5%	10Hz	2500µs	4.0%	4.0%
1%	10Hz	1000µs	1.7%	1.9%
2.5%	100Hz	250µs	2.6%	2.5%
2.5%	1000Hz	25µs	2.6%	2.5%
1%	1000Hz	10µs	0.1%	0.1%
10%	**10Hz**	**10000µs**	**21.6%**	**21.0%**

*% Noise = ((Frequency(Hz) * Duration(µs))÷(1*10^6)) * 100*



Sandia Power Project
Some recent results

- Quantify impact in dollars, 300-500k in idle power alone
- Can we save power when running applications?
  - Go into lower power state/frequency while waiting...
- Reduce frequency runs without affecting performance?
  - Little to no impact on run-time, large power savings?
  - Early Results:
    - AMG 32% Power Savings costs 3% in performance
- Does network imbalance impact Power?
  - Less bandwidth?
  - Higher latency?
  - Many more questions need to be answered (quantified)



Work at Julich

- Energy-Efficient Cluster Computing (eeClust) http://www.eeclust.de - Michael Knobloch
- Fit4Green http://www.fit4green.eu - Michael Knobloch

JÜLICH
FORSCHUNGSZENTRUM

Green IT @ JSC

THE GREEN 500 #1

Michael Knobloch
m.knobloch@fz-juelich.de

HPC Best Practices: Power Management, September 28/29, 2010

Outline

- Jülich's dual supercomputer concept
- Partnerships and projects
  - Exascale projects
  - PRACE
  - Fit4Green
  - eeClust

Jülich Supercomputing Centre

Jugene:

IBM BlueGene/P

294912 processors

1 Petaflops

2.3 MW

Juropa:

Bull, Sun

3288 procs / 26394 cores

308 Teraflops

1.5 MW

Exascale Projects @ JSC

- 

Exa Scale Innovation Center (with IBM)
  - Energy efficiency
  - Chip/processor technology
  - Application development
  - 2015 exascale prototype

- ExaCluster Laboratory (together with Intel and ParTec)
  - Cluster technology
  - Commodity hardware
  - Accelerators
  - Cluster management software

The PRACE Project

EU approved the PRACE Preparatory Phase Project
(Grant: INFSO-RI-211528)

- Project start (preparation phase): Jan 2008
  - Project budget: 20 M€
  - Evaluation of future technologies
  - Installation of 6 production system prototypes
  - PRACE benchmark suite
- 20 members from European countries (2010)
- Implementation phase (PRACE-1IP) started on July 1, 2010

The Fit4Green Project

http://www.fz-juelich.de/jsc/grid/FIT4Green
http://www.fit4green.eu

- Project start: Jan 2010
  - Funded by the European Union (EU)
  - 10 partners from European countries

- Targets at ICT energy reducing
  - Creating an energy-aware layer of plug-ins for data center automation
  - Without giving up on compliance to Service Level Agreements (SLA) and Quality of Service (QoS) metrics
  - Run pilots using three representative data center typologies
    - Service/Enterprise Portal
    - HPC Grid
    - Cloud

Fit4Green: Envisioned Optimization Areas

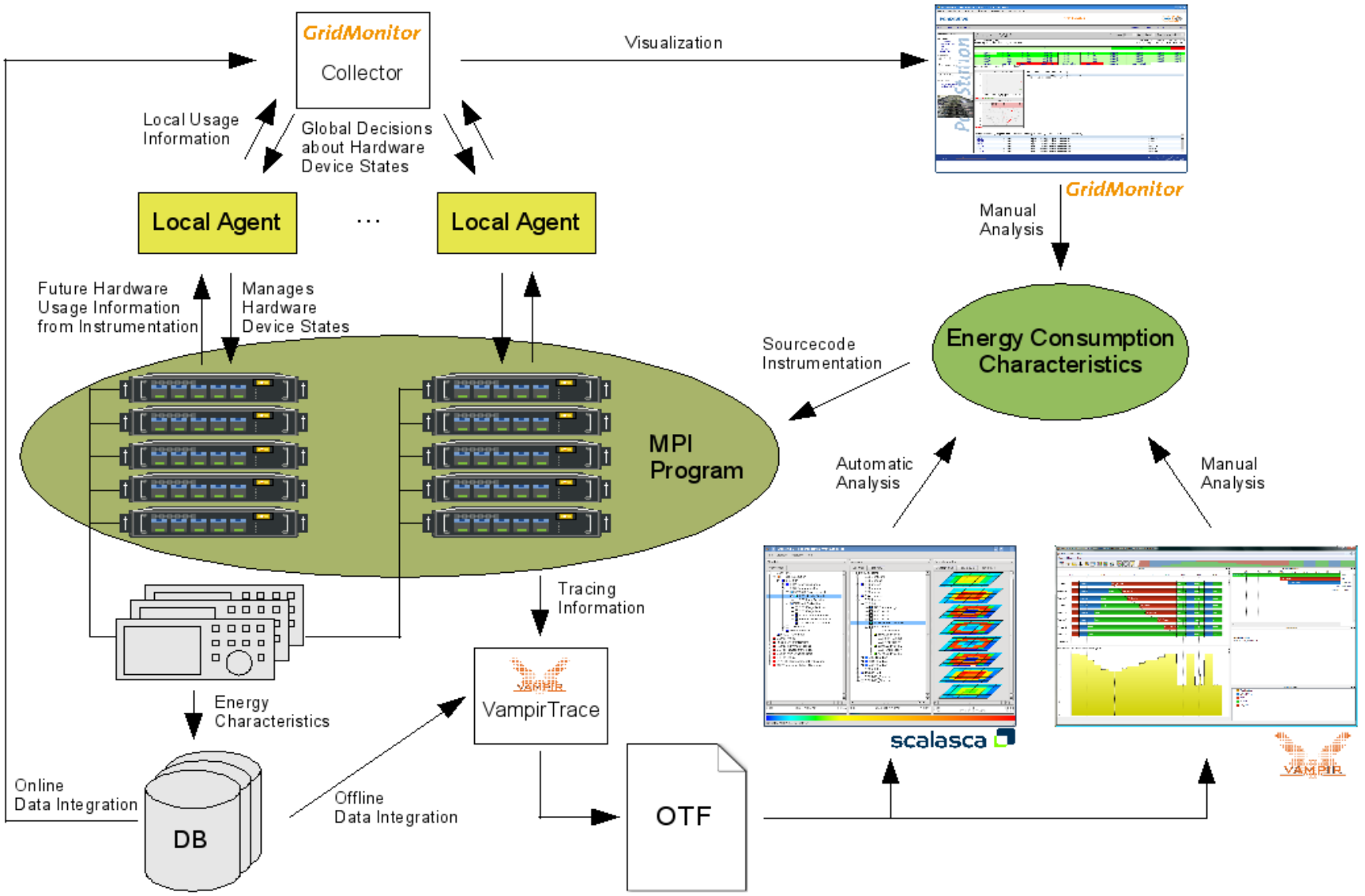
- Energy-aware scheduling
  - Assign most energy-effcient machine for the job
  - Consider night/day cycles and power costs
  - Automatically turning off unused cluster equipments
    - Nodes
      - draining of job queues preluding a maintenance time
      - reservation phase before starting a multi-node job
    - Cores which are not required by application
  - Assign jobs so that nodes can sleep as long as possible
- Introducing power-aware brokering strategies in the Grid federation
- Application focus: PEPC, WRF, ...

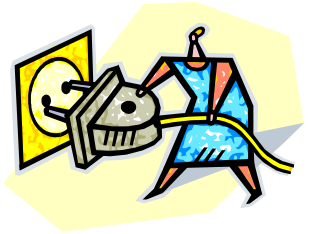
The eeClust Project

http://www.eeclust.de

- Project start: April 2009
  - Funded by the BMBF
  - 4 partners in Germany (JSC, ZIH, Uni Hamburg, ParTec)

- Targets energy-efficient HPC
  - Energy savings without performance degradation
  - Usage of hardware energy saving options whenever possible → need to know in advance which hardware is (not) used
  - Collect traces of program execution → define energy consumption characteristics
  - Instrument program before next execution → OS daemon invokes power management features at right time


GridMonitor
Collector
Visualization
Local Usage Information
Global Decisions about Hardware Device States
GridMonitor
Local Agent
...
Local Agent
Manual Analysis
Future Hardware Usage Information from Instrumentation
Manages Hardware Device States
Energy Consumption Characteristics
Sourcecode Instrumentation
MPI Program
Automatic Analysis
Manual Analysis
Tracing Information
VampirTrace
Energy Characteristics
scalasca
Online Data Integration
DB
Offline Data Integration
OTF



Work at LANL - Michael Lang

Energy/Performance Observations

- Looked at node power for recent systems
  - Istanbul
  - Nehalem
- In both cases
  - Best energy utilization was to max power/frequency and shorten time to solution.
  - Overall issue idle node power.

Number of usefull cores

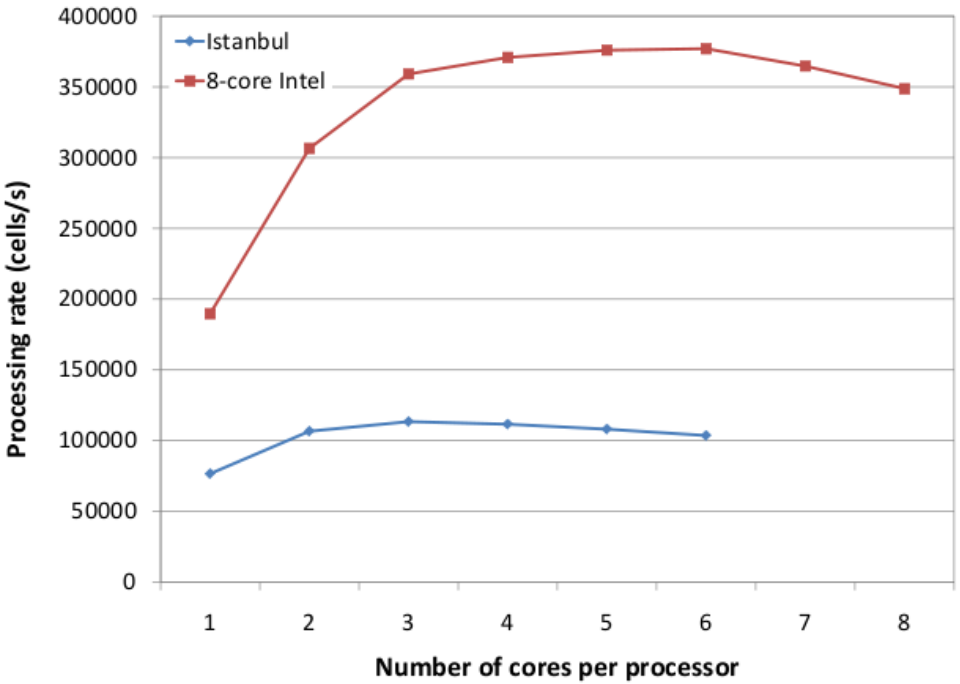


Fig. 3. SAGE performance scaling.

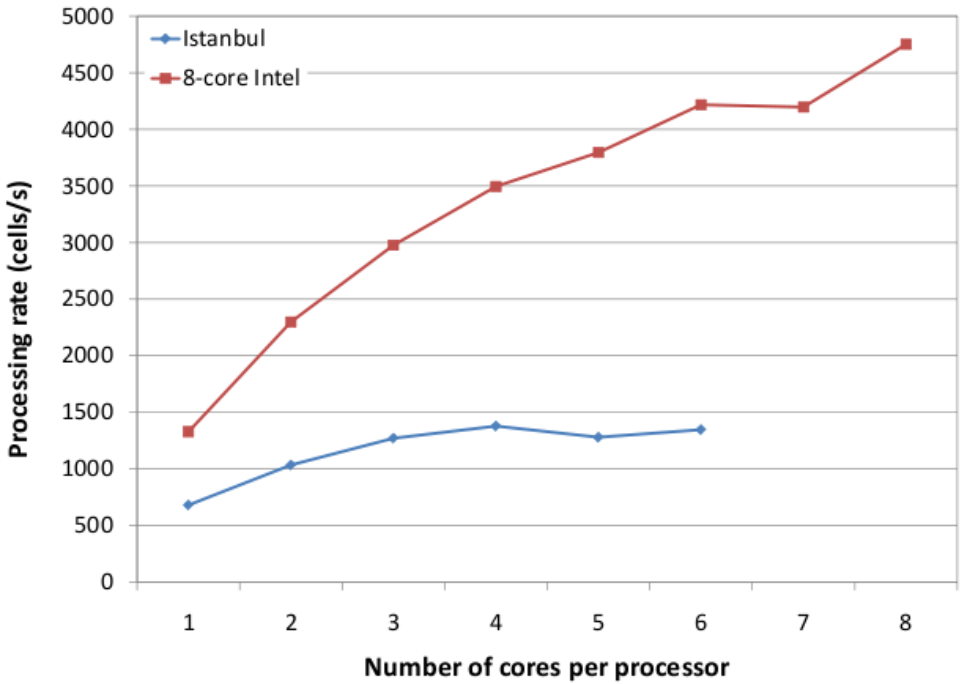


Fig. 4. PARTISN performance scaling.

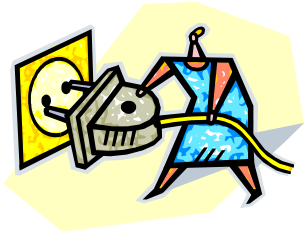
"Characterizing the Impact of Using Spare-Cores on Application Performance" Europar 2010

Other power issues

- Processor managed power problems
  - Unpredictable performance
  - Performance is reduced to slowest node for Bulk Synchronous codes
-

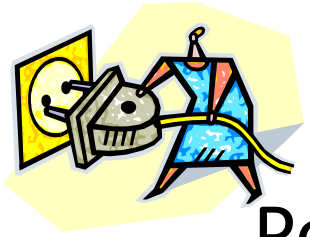
What we need

- Need cluster wide energy management
  - Control freq
  - Ablility to power down individual cores rather than sockets (Match app to memory BW)
    -



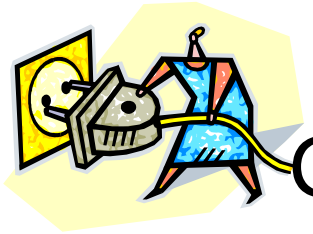
Experience
Novel / Interesting Approaches

- See project presentations
  - Early stages, most efforts are at least somewhat Novel
- Continue/extend monitoring at the HW level
  - E.g. Power7 , CrayXT , BlueGene
    - Information (sensors) at fine granularity and high frequency
- Standardize and make available to OS and/or application and tools
  - See standards slide
- Observation: Hard/Impossible to project large scale effects from small scale runs
  - Testing and verification at large scale
- High percentage Application and architecture dependent



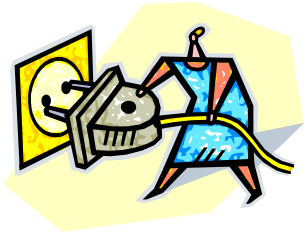
Best Practices in Power-aware OS features and scheduling

- Save power during idle (low hanging fruit – pick it)
  - But: no standard way to do this
    - see standards
  - dependent upon architecture, OS, application
- Application power signatures
  - Standard way of collecting that enables things like directives in applications, tuning, targeting application specific power efficiency
  - Observing effect
- Consider the trade-off between performance and power
  - What is the acceptable trade-off?
  - Who decides? - User, application, site specific?
  - Cost model including power
    - Pay for what you use influences how you use



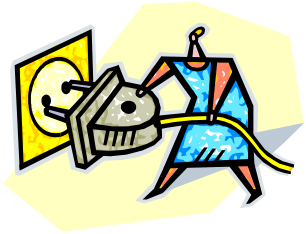
Gaps Looking Forward to New Systems

- More capability to monitor component level power (CPU and beyond)
- More ability to control power and frequency features
  - OS, tools, application
- Need SW ability to monitor
  - OS, application, tools, application
  - E.g. Compiler – decide when to turn or floating point unit
- Ability to turn off power and frequency features that are detrimental
  - Designed for enterprise sometimes harms HPC
- Interface with facilities
  - How does what we do affect facilities and vice versa
  - Power management affect on facilities and platform hardware
- Overhead of power management
  - State changes are not free



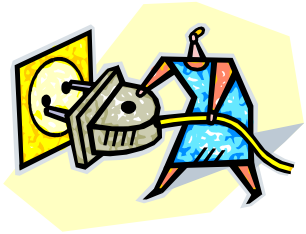
Evolve or start over for future systems?

- Linux or light weight kernels certainly provide a basis, but they might be inadequate or even a poor basis
- Scheduling, do we evolve or have to start over?
  - We do have a basis for evolution
  - Charging for power consumption
  - Dynamic power based scheduling (Macro and/or micro)
  - Backfill based on available power
- We have some basis for evolution in instrumentation and control
  - Power7, CrayXT, BlueGene, (x86 control, little instrumentation)
  - What little we have we use poorly at the moment.
- Maximum efficient utilization of power envelope
  - Power aware scheduling
  - Metering, OS and architectural knobs
  - Schedule based on variable cost of power (time of day, time of year)



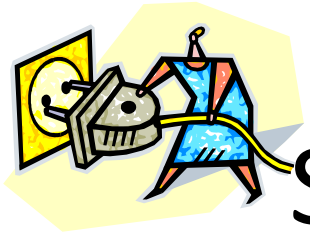
Issues shared with large commercial centers

- There is a difference between capability and capacity
  - There is overlap but mostly at low level (architecture)
  - How it is managed/leveraged might vary drastically
- We do share the need for advanced architecture features for power management and load balancing- largest overlap
- Cost; TCO
- Potential change in overlap as time progresses
  - could be more or less, undefined



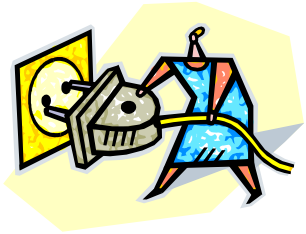
Hardware/facility/system interfaces to influence

- Incorporate appropriate power management/monitoring "knobs" into chip/component architecture board design etc.
- Expose "knobs"!!!!
- Need to integrate facilities and platform power management
- Adding/standardizing board level instrumentation to monitor power at fine granularity/component level
  - More from the SW perspective



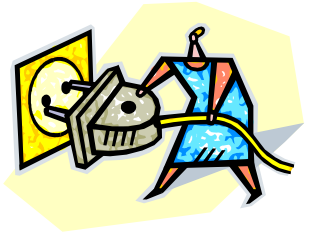
Status of (de facto) standards

- ACPI – Does this “standard” serve our long term purposes?
- Standardize application and tools API for power management
- Standardize what OS exposes vs. expose everything
  - Notice we didn’t say standardize how OS interfaces with hardware (ACPI) we didn’t go that far.
- HPC involvement in standards development
  - e.g. must scale
- Trade-off between fine grained control and HW agnostic standards
  - Might get lowest common denominator
- If power is added to scheduling parameters, do it in a standard way across architectures
  - an apple is an apple on every platform whether supported or not



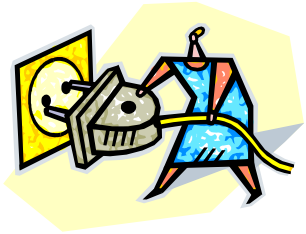
Other key findings

- Discussed overlap with commercial, can we focus on overlap within the community?
- Wider HPC community standards development and implementation
  - Overlap more significant
  - Outcome more useful
  - Community as a whole has more influence



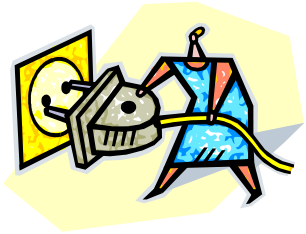
Re-occurring Themes

- INFORMATION Critical
  - Cannot affect without understanding effect
  - Component level instrumentation critical
- IT DEPENDS
  - Variables include Architecture, OS, Application etc….
  - Level of exposure



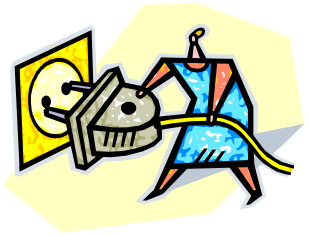
Related Efforts, Publications, etc.

- *"Topics on Measuring Real Power Usage on High Performance Computing Platforms",* James H. Laros III et.al.
- *"Profile-Based Energy Reduction for High-Performance Processors", Michael Huang et.al.*
- *"Analysis of Dynamic Voltage Scaling for System Level Energy Management",* Gaurav Dhiman et.al.
- "*Implications of Historical Trends in the Electrical Efficiency of Computing",* Jonathan G. Koomey et.al.
- *"Memory-aware Scheduling for Energy Efficiency on Multicore Processors",* Andreas Merkel et.al.
- *"Compiler-Directed Dynamic Voltage/Frequency Scheduling for Energy Reduction in Microprocessors",* Chung-Hsing Hsu et.al.
- *"Semantic-less Coordination of Power Management and Application Performance",* Aman Kansal et.al
- *"Energy-Efficient Processor Design Using Multiple Clock Domains with Dynamic Voltage and Frequency Scaling",* Greg Semeraro et.al.
- *"Power and Performance Trade-Offs in Contemporary DRAM System Designs for Multicore Processors",* Hongzhon Zheng et.al.
- *"Empirical Analysis on Energy Efficiency of Flash-based SSDs",* Euiseong Seo



Information and Links

- http://www.acpi.info - Advanced Configuration & Power Interfaces Specifications
- http://www.intel.com/technology/iapc/acpi/ - Intel ACPI related information and specifications
- http://developer.amd.com/cpu/apml/Pages/default.aspx- AMD Advanced Power Management Link and ACPI related information and specifications
- BIOS and Kernel Developers guide (BKDG) pick your family - http://developer.amd.com/documentation/guides/Pages/default.aspx



Additional Material
ACPI

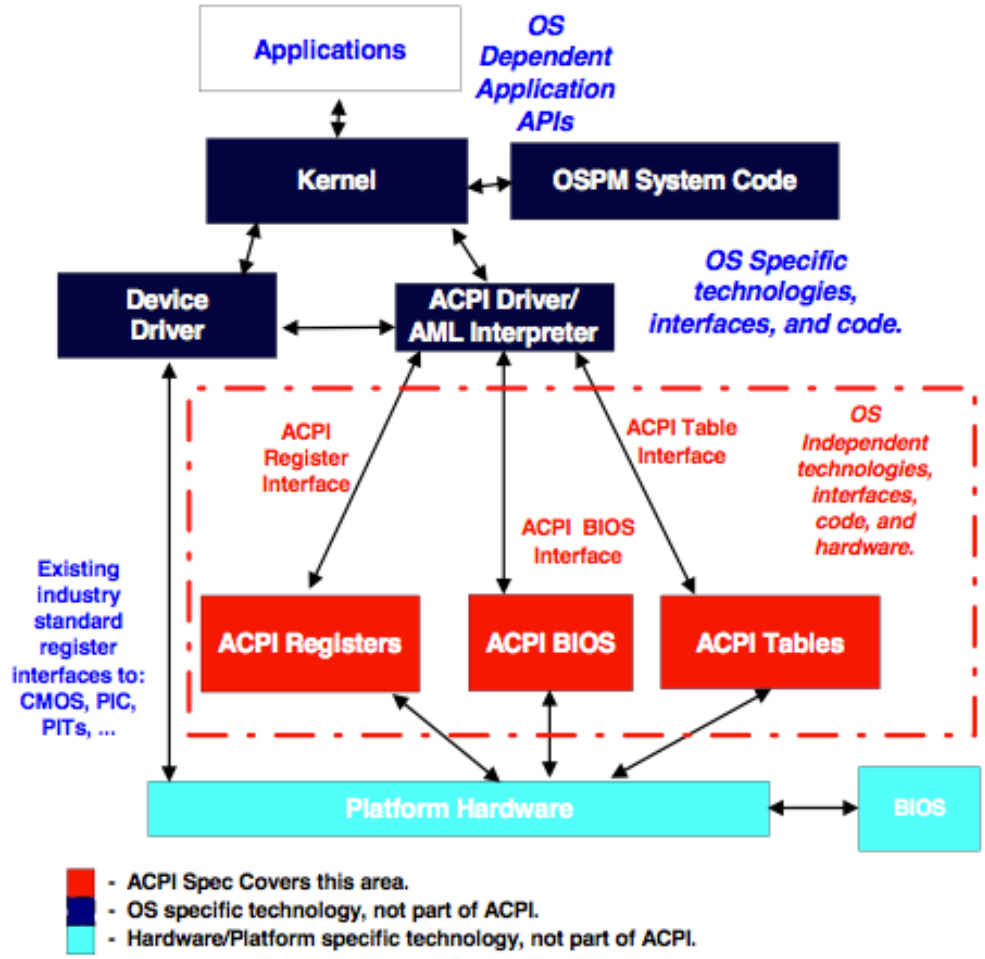


Figure 1-1 OSPM/ACPI Global System

From ACPI 3.0 Specification