International Conference on
Energy-Aware High Performance Computing
September 16 – 17, 2010

Energy Management for HPC with IBM

Klaus Gottschalk
IBM Germany
gpttschalk @de.ibm.com

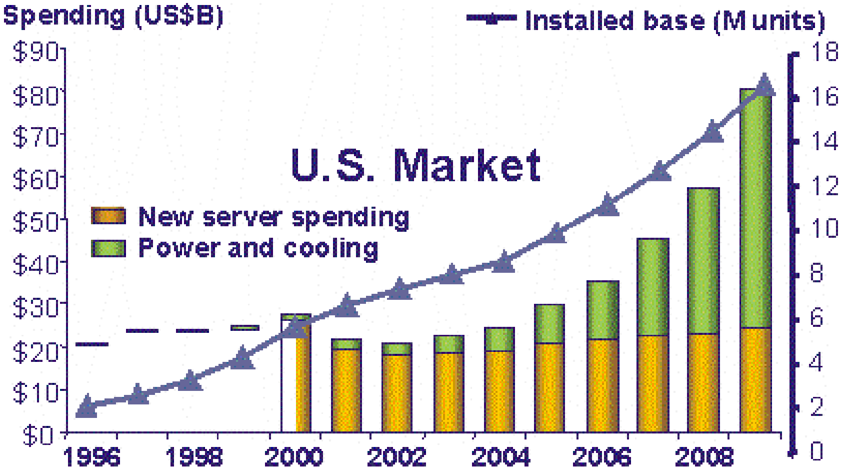
Green Datacenter Market Drivers and Trends

- Increased green consciousness, and rising cost of power

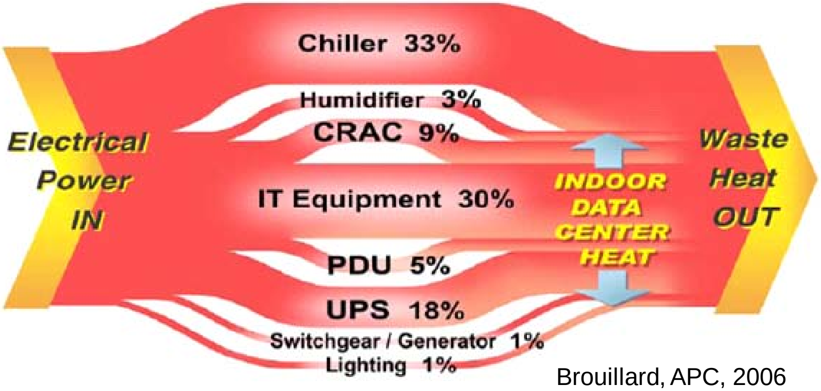
- IT demand outpaces technology improvements
  - Server energy use doubled 2000-2005; expected to increase15%/year
  - 15 % power growth per year is not sustainable
  - Koomey Study: Server use 1.2% of U.S. energy

- ICT industries consume 2% ww energy
  - Carbon dioxide emission like global aviation

al Actions Needed



Source IDC 2006, Document# 201722, "The impact of Power and Cooling on Datacenter Infrastructure, John Humphreys, Jed Scaramella"



Brouillard, APC, 2006

Future datacenters dominated by energy cost; half energy spent on cooling

How much does it costs ?

HPC System with 10 Racks, each 36kW

- CRAC Cost for 100 kW Unit: $24k H/W & $15.6k Install
- CDU Cost for 150 kW Unit: $26k H/W & $7.6k Install
- Chiller Cost for 100 kW Capacity: $50k H/W & $10k Install
- Cost of Electricity: $0.20 / kWh (0.13 Euro /kWh)

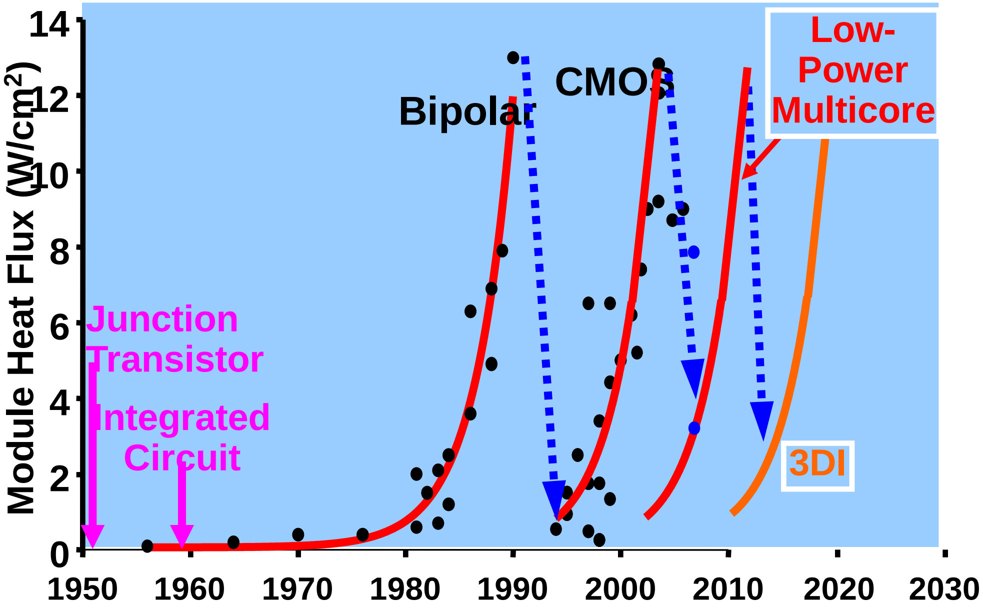
The Power Problem

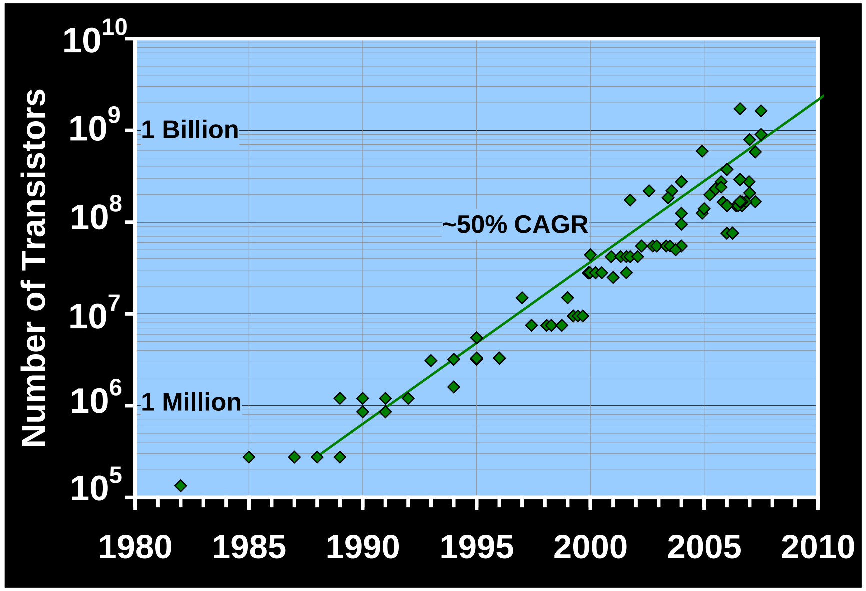
Power = Capacitance * Voltage^2 * Frequency

=> Power ~ Frequency^3

- Power Consumption of 2 cores @80% frequency consumes as much as 1 core at 100% frequency
- We have a frequency problem:
  - Power per chip is constant due to cooling

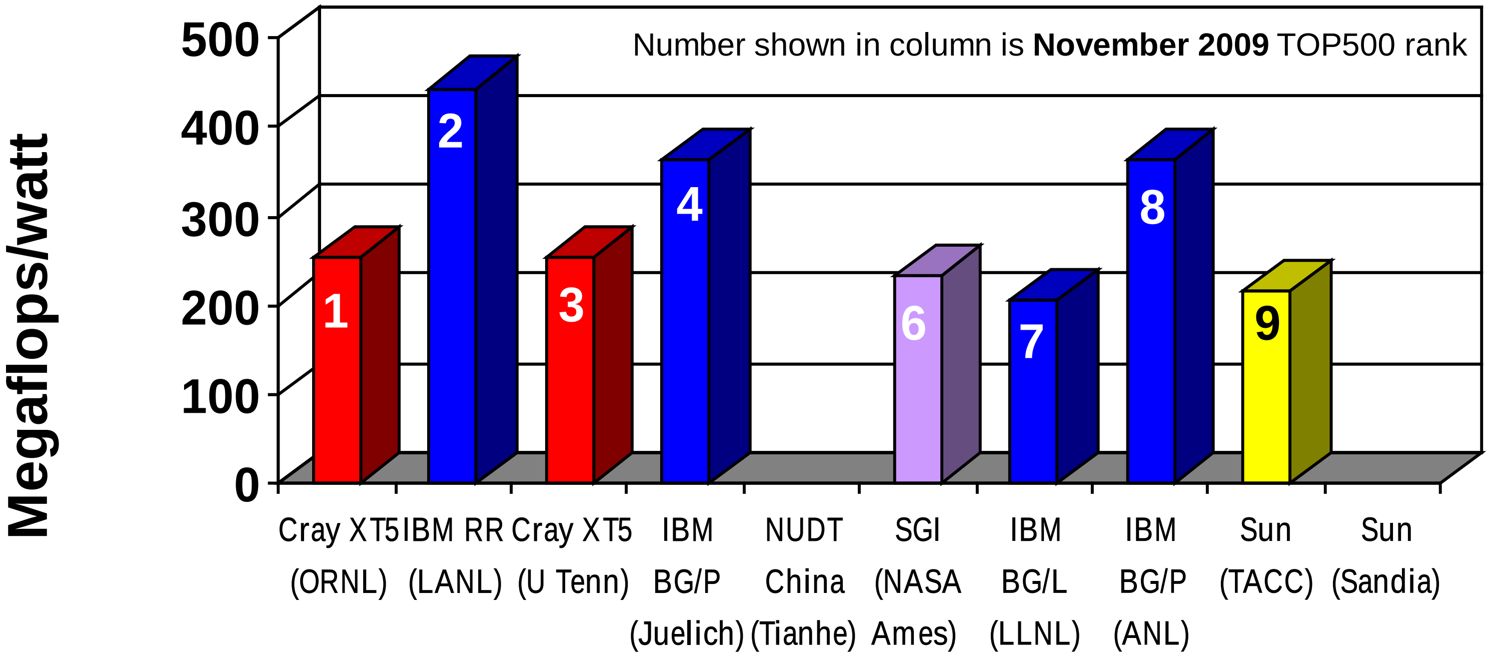
    => multicores at constant frequency
- And there is a passive power problem
  - Smaller lithography
    - ➢ more leakage current
    - ➢ more idle power





Power your planet.

Power Efficency (MF/w) of Top 10 TOP500 Systems



Rank	Site	Mfgr	System	Rmax	MF/w	Relative	
1	ORNL	Cray	Jaguar XT5 HE 2.6 GHz 6C Opteron	1759	253	1.76	6.9 Mw
2	LANL	IBM	Roadrunner QS22/LS21	1042	444	1	2.3 Mw
3	U of Tenn	Cray	Kraken XT5 HE 2.6 GHz 6C Opteron *	831.7	253	1.76	
4	Juelich	IBM	Blue Gene/P	825.5	364	1.22	2.2 Mw
5	NUDT	Self	Intel Nehalem/AMD Radeon GPU *	563.1			
6	NASA Ames	SGI	QC 3.0 Xeon	544.3	232	1.92	
7	LLNL	IBM	Blue Gene/L	478.2	205	2.16	
8	ANL	IBM	Blue Gene/P	458.6	364	1.22	
9	TACC	Sun	2.3 GHz QC Opteron	433.2	217	2.05	
10	Sandia	Sun	Red Sky 2.93 Nehalem *	423.9			

Source: www.top500.org
*Notes: * Kraken power is scaled down from Jaguar, NUDT, Sandia/Sun did not provide power numbers*

Rack to Rack: Power 755 Compared to Power 575 (POWER6)

	Power 755	Power 575
Cores/chip	8	4
Total cores	32	32
Frequency	3.3 GHz	4.7 GHz
Memory (max)	256 GB	256 GB
Cooling	Air	Water
Cores/rack Rack type	320 19”	448 24”
Power (Watts) (Linpack)	1650	5400

Each Power 755 node offers the same core count as Power 575 with:

- 40-50% Improvement in Performance
- Air Cooling vs. Water Cooling
- 1/3 of the Energy Consumption
- 37% Improvement in floor space for a 64 node configuration
- Green500 ~ 495 MFlops/Watt

IBM EnergyScale functions

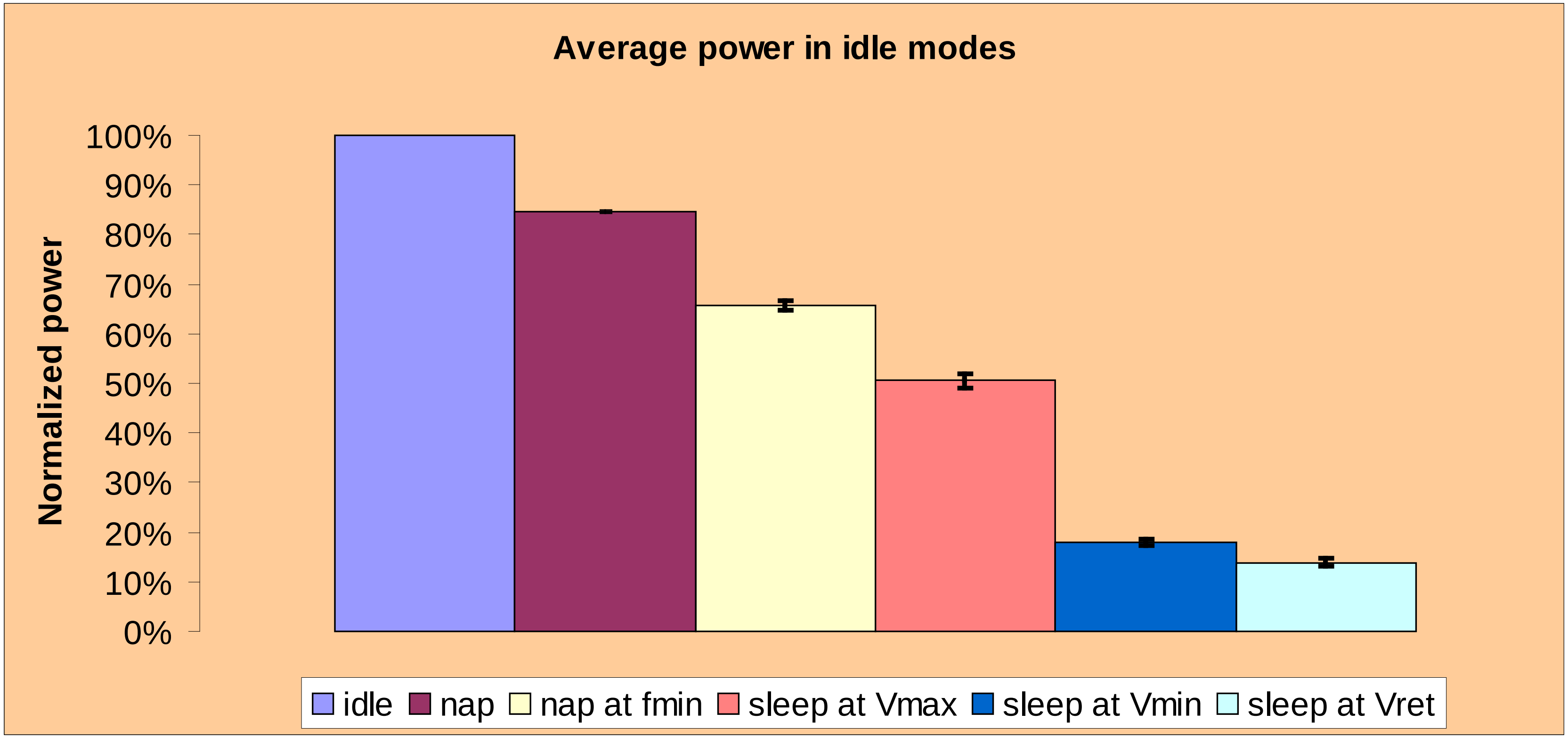
- **Power / Thermal Trending**
  - Collect and report power consumption, inlet and exhaust temp
- **Power Capping**
  - **Guaranteed (Hard Cap)**
    - Enforces a power cap via Dynamic Frequency and Voltage Slewing
  - **Soft Power Cap**
    - Attempted lower cap, but not guaranteed.
- **Energy Management Modes – Enhanced for POWER7**
  - **Static Power Save (SPS)**
    - Save power via a fixed voltage and frequency drop – as much as 30% down for P7
  - **Dynamic Power Save (DPS)**
    - Optimize power vs performance using Dynamic Voltage and Frequency Slewing
    - Will provide performance boost at very high utilization
    - Will save power at most utilizations
  - **Dynamic Power Save - Favor Performance (DPS-MP)**
    - Will provide performance boost at most utilizations
    - Will save power only at very low utilization

POWER7 Power Save States

Power Save States	Freq (Max)
Pstate0	**1.1**
Pstate1	**1.0**
Pstate2	**0.9**
Pstate3	**0.8**
Pstate4	**0.7**
Pstate5	**0.6**
Pstate6	**Fmax**
Pstate7	**0.50**

Architected Idle Modes (OS+hypervisor managed)

- *Nap* - clocks off for execution units and L1 caches within core
  - Optional *auto* frequency drop support for additional power reduction
- *Sleep* – clocks off for entire chiplet, caches flushed prior to entry.
  - Optional *auto* voltage drop to *latch-state-retention level* for additional power reduction when all cores sleep.



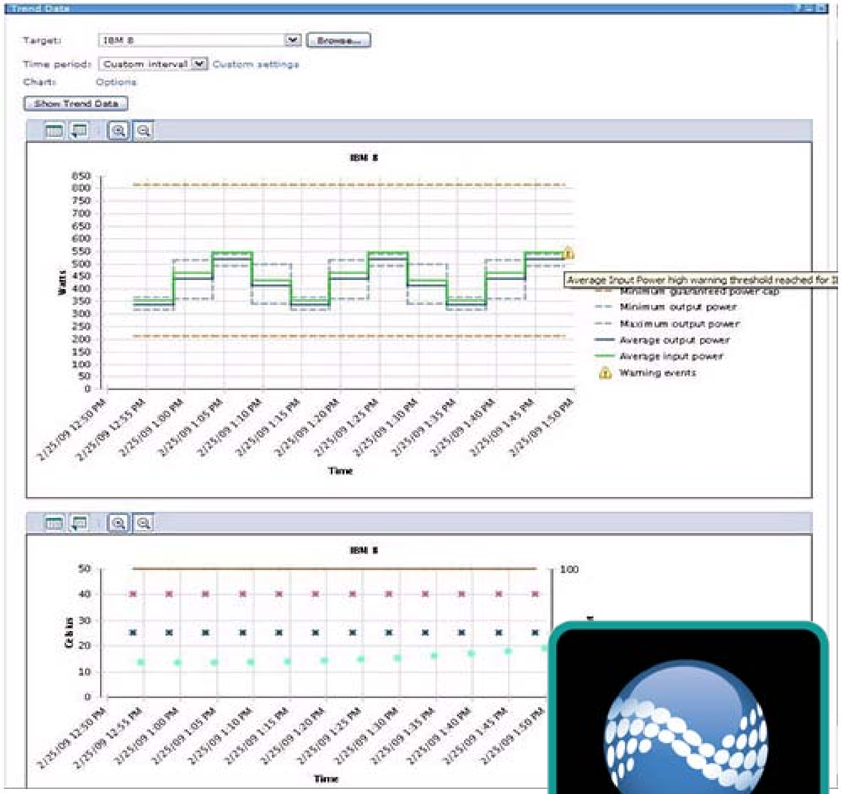
Power your planet.

IBM Systems Director Active Energy Manager (AEM)

Monitoring energy in a data center lets you begin to manage it

- AEM is a cornerstone of the IBM energy management framework
- Measure , Monitor, and control energy usage
- Power and Thermal Measurement
- Supports System x, POWER, and z System natively
- Supports other equipment via external sensors
- Integrates with Infrastructure Management
- Integrates with Enterprise Management

Active Energy Manager



Trend Data

Target: IBM 8

Time period: Custom interval

Charts: Options

IBM® Systems Director

Active Energy Management: Energy Scale™

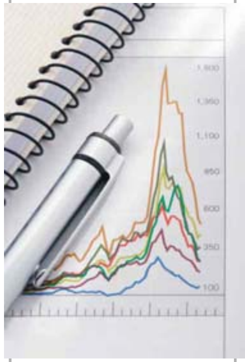
- Active Energy Manager is configurable using IBM Systems Director
- Offers 4 modes of energy management (frequencies are for p755 @ 3.3 GHz
- Normal
  - Static: Active processor frequency set at 100% of nominal (3.3 GHz)
- Static Power Saver (SPS)
  - Static: Active processor frequency set at 2.3 GHz, 30% below nominal.
  - Folding will set idle cores to Nap (1.65 GHz) or to Sleep (0 GHz)
    - available with AIX 6.1J and above
  - Maximum energy savings – used for long periods of low utilization
- Dynamic Power Saver (DPS)
  - Processor frequency is set based on processor core utilization
  - Un-utilized cores set to 1.65 GHz and ramped up as utilization increases to maximum 90% of nominal frequency (2.97 GHz)
  - This feature prefers power savings over performance
- Dynamic Power Saver – Maximum Performance (DPS-MP)
  - Processor frequency is set based on processor core utilization
  - Un-utilized cores set to 1.65 GHz and ramped up as utilization increases to maximum 110% of nominal frequency (3.53 GHz)
  - This feature prefers maximum performance over power savings

IBM Systems Director Active Energy Manager V4.2 (con't)

- *AEM application supported on:*
  - *Windows, AIX, and Linux (x86, POWER, and System z)*
- *Web-based user interface requiring only a browser*
- *Energy thresholding*
  - Enables a user to set an energy or temperature threshold and be notified
  - when it is reached (or allow an action to automatically be taken)
- *Soft power capping (an option within power capping)*
  - Ability to set a lower energy cap value to enable clients to save energy
- *Easily set power caps on multiple systems*
- *Group capping (an option within power capping):*
  - Enables a user to set an energy cap for a group of servers (such as all the
  - servers in a rack)
- Data to aid in server power on/off scenarios
  - Understand time to IPL and standby power
  - Number of lifetime IPLs and reliability threshold (P7 only)

IBM Systems Director Active Energy Manager V4.2 (con't)

- ***Support for Facility Providers***
  - – Receive data and alerts from infrafrastructure management applications:
    - Emerson-Liebert' SiteScan
    - Eaton's Power Xpert and Foreseer
    - APC's InfraStruXure Central
  - – Ability to correlate Power Dist and Cooling with IT Resources
  - – Retrieve temperature & power data using sensors and meters from:
    - Synapsense
    - iButton
    - Sensitronics
    - SmartWorks
    - Arch Rock



- ***Command Line Interface (CLI) support***
  - – Provides the ability to script commands and actions
- ***REST API Northbound Interface Support***
  - – Provides Integration ability with other Management SW (eg. Tivoli
- ***Automation***
  - – Easy-to-use wizards provide a way to set energy policies and automate events based on energy alerts and data

IBM HPC Software Stack
Energy Awareness Goals for xCAT and LoadLeveler Power

- **IBM xCAT Cluster Management**
  - Manage power consumption on an ad hoc basis
    - For example, while cluster is being installed, or when there is high power consumption in other parts of the lab for a period of time
    - Query: Power saving mode, Power capping value, power consumed info, CPU usage, fan speed, environment temperature
    - Set: Power saving mode and Power capping value

- **IBM Tivoli Workload Scheduler LoadLeveler**
  - Optimize energy consumption:
    - Set Processor Frequency to Minimum (Sleep) for idle nodes with no scheduled workload
    - Set Optimal Processor Frequency to Minimize energy consumption with minimum performance degradation for nodes with a parallel workload

Power and Energy Aware LoadLeveler Features

- **Goals**
  - Identify idle nodes in the cluster and put them in the lowest power mode
  - Provide to system administrators query capability on historical usage of power and energy by workload, user, etc.
  - Reduction of energy consumption on workloads with minimal impact to performance
- **Choices for system admin**
  - Decide to use Energy Optimize policy or not on his system
  - Decide the max performance degradation one application will be impacted by, if the Energy policy is applied

- If Energy Policy is on, policy is applied only to jobs that match the performance degradation criteria
- System admin can query LL DB to evaluate the impact of the potential policy on performance degradation and energy saving

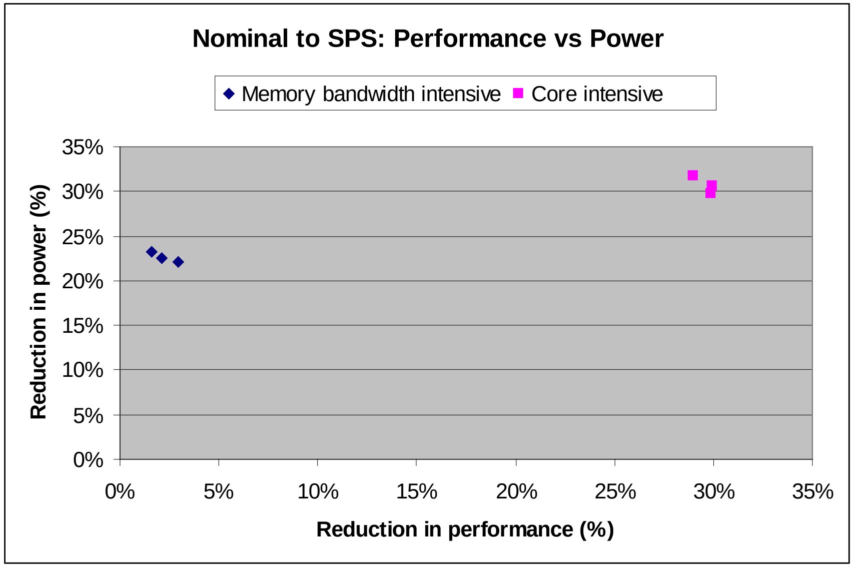
Power and Energy Aware Scheduling in Loadleveler

- User submits his job as usual
- At the end of the job, a energy report and a tag is produces. The report provides:
  - Actual power and energy consumed, and traditional elapsed time, at nominal frequency
  - Power/Energy savings if other frequency was used
- Report and tag are stored in LoeadLeveler Energy DB
- System Admin browse the DB to decide if he applies the minimum energy policy and with a limit on the allowable performance degradation
  - Frequency will be within total cluster power consumption and max power limit (TDP) per node
- When user resubmits his job (preferably with the tag), his job will run at optimal frequency if projected performance variation is within the policy criteria. If not, it will run at nominal frequency.
- Energy report is sent to user and stored in LoadLeveler DB

Active Energy Management: Energy Scale on POWER7

- Active Energy Manager is configurable using IBM Systems Director
- Offers 3 modes of energy management
- Static Power Saver (SPS)
  - Static: Active processor frequency set at 30% below nominal (2.31GHz)
  - Folding will set idle cores to Nap (1.65 GHz) or to Sleep (0 GHz)
  - Maximum energy savings – used for long periods of low utilization
- Dynamic Power Saver (DPS)
  - Processor frequency is set based on processor core utilization
  - Un-utilized cores set to 1.65 GHz and ramped up as utilization increases to maximum 90% of nominal frequency (2.97 GHz)
  - This feature prefers power savings over performance
- Dynamic Power Saver – Maximum Performance (DPS-MP)
  - Processor frequency is set based on processor core utilization
  - Un-utilized cores set to 1.65 GHz and ramped up as utilization increases to maximum 107% of nominal frequency (3.53 GHz)
  - This feature prefers maximum performance over power savings

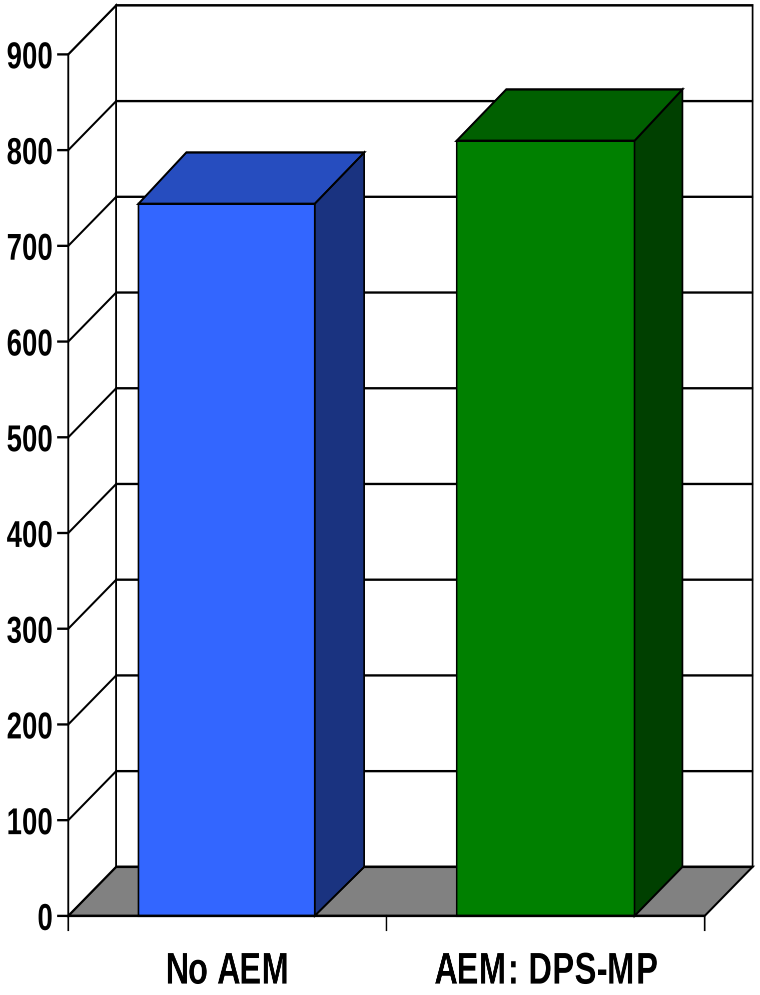
SPEC Benchmark	Performance Characteristic
416.gamess	Core intensive
433.milc	Mem. bandwidth intensive
435.gromacs	Core intensive
437.leslie3d	Mem. bandwidth intensive
444.namd	Core intensive
459.GemsFDTD	Mem. bandwidth intensive



Correlation of performance and power consumption

DPS-MP on p755

- Linpack Benchmark with Active Energy Manager
- AEM “Over–Clocking” support with DPS-MP
  - Dynamic Power Save Favor Performance
- Test environment
  - Power 755 32core @ 3.3GHz
  - XLC V11.0 beta
  - ESSL V5.1 beta
  - PE V5.2
- In case of AEM=OFF
  - 692.7 Gflops* (82.0% efficiency)
- In case of AEM=ON
  - DPS-MP
  - 753.6 Gflops* (89.2% efficiency)
- Using AEM: 8.8% Performance Gain
  - 753.6 Gflops / 692.7 Gflops



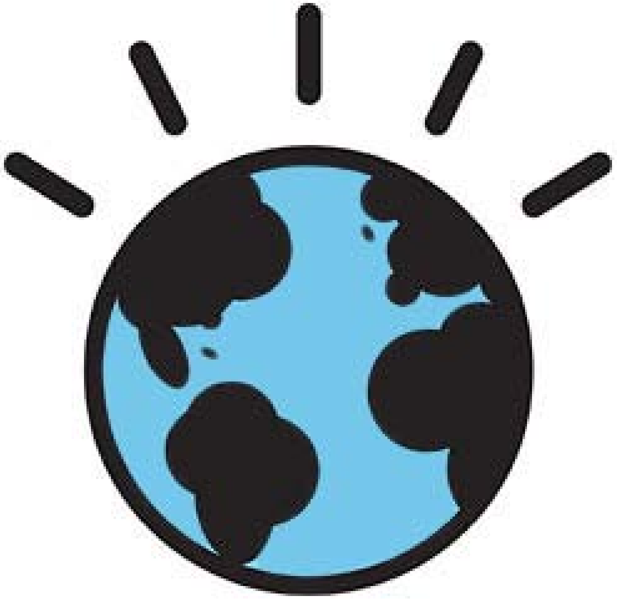
* Power 755 performance projected from actual Power 550 results

Summary

- POWER7 is bringing significant improvement in performance along with energy-efficiency and dynamic power management support.
  - 4X number of cores, significantly wider DVFS range, new idle modes.

- We have software offer to manage power and energy :
  - AEM, xCAT (for HPC) and in the future LL

- Using those tools, our customer can save quite a lot on their energy bill:
  - Between 40 and 80% when nodes are idle
  - Up to 20% with no performance degradation for memory bound workloads.



Power your planet.

Smarter systems for a Smarter Planet.