



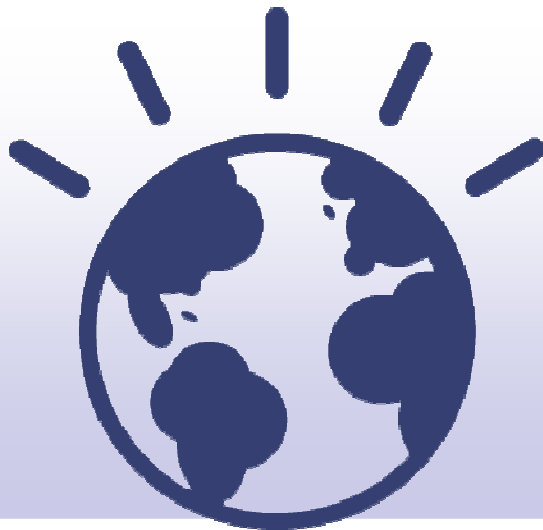
## Energy Aware Scheduling

### Journees CFD Equip@Meso

May 16, 2013

Luigi Brochard, IBM

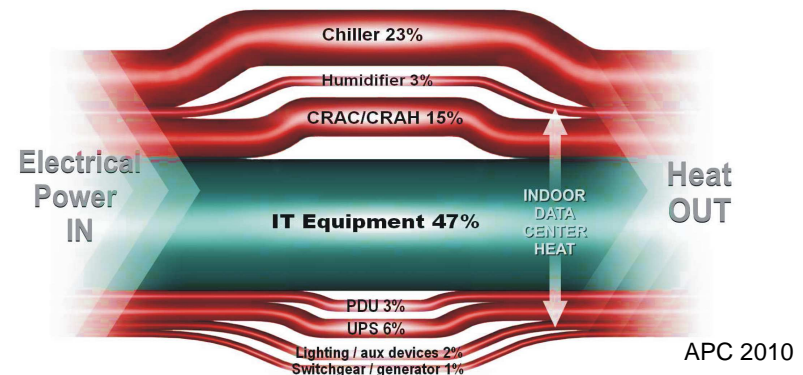
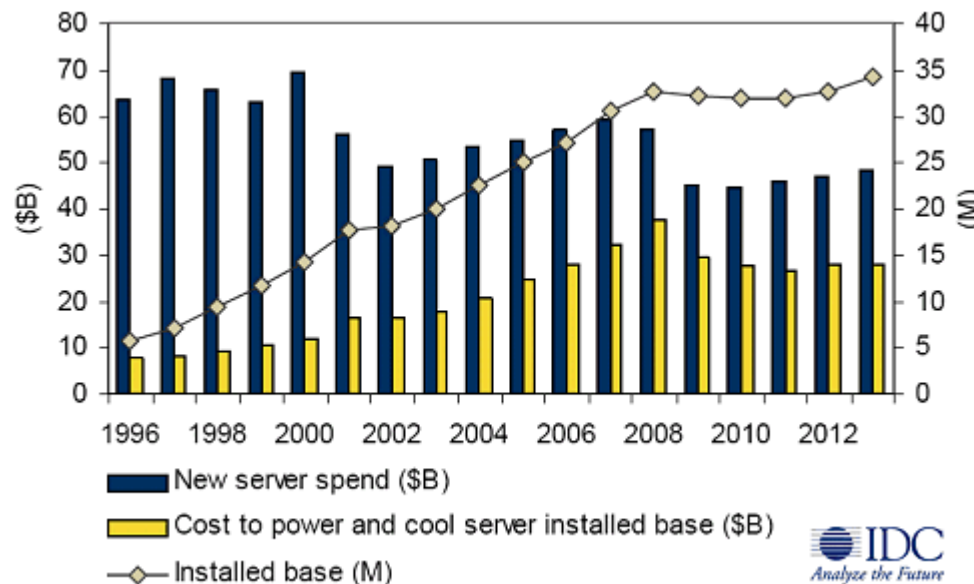
Francois Thomas, IBM



High Performance Computing  
For a Smarter Planet

## Green Datacenter Market Drivers and Trends

- **Increased green consciousness, and rising cost of power**
- **IT demand outpaces technology improvements**
  - ▶ Server energy use quadrupled 1996-2008; It has decreased 2008-2012, as # servers installed
  - ▶ Power costs are more than 50% of new server spending
- **ICT industries consume 2% ww energy**
  - ▶ Carbon dioxide emission like global aviation



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

## The Power Problem

**A 1000 node cluster with 2 x86 sockets, 8 core 2.7 GHz, consumes about 340 KW (Linpack), without cooling**

In Europe (0.15€ per KWh), this will cost about 441K€ per year

In US (0.10\$ per KWh), this will cost about US\$ 295K per year

In Asia (0.20\$ per KWh), this will cost about US\$ 590K per year

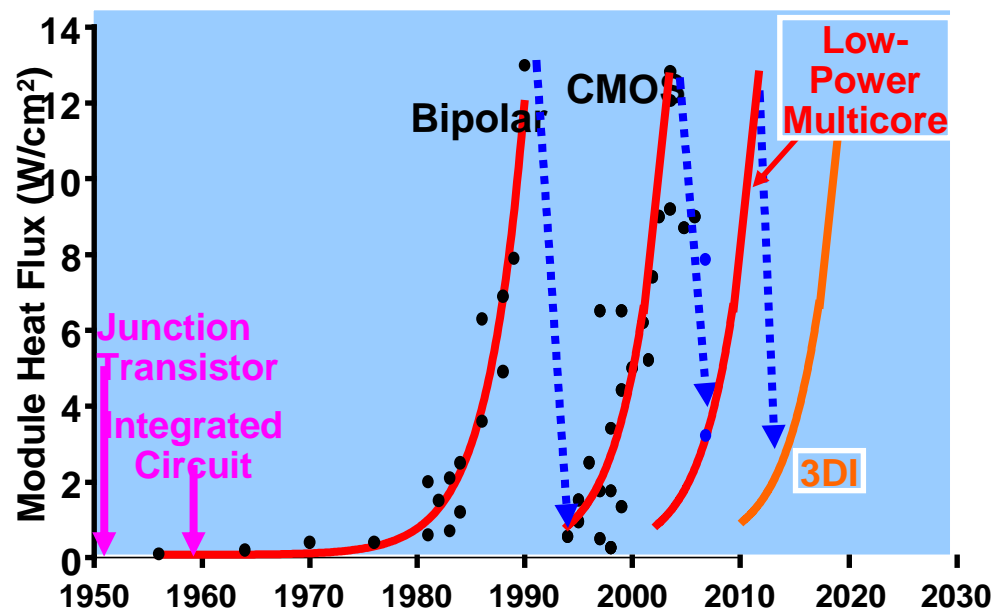
**What about saving 10% to 15% without any infrastructure change?**

# The Power Equation

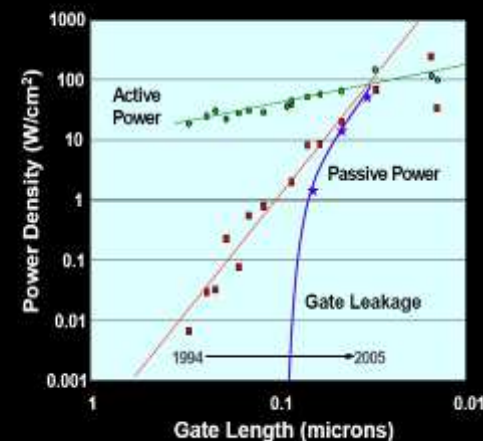
■  $\text{Power} = \text{Capacitance} * \text{Voltage}^2 * \text{Frequency}$

$\text{Power} \sim \text{Capacitance} * \text{Frequency}^3$

- ▶ We have an active power problem
  - Frequency minimisation for active nodes
- ▶ We have a passive power problem
  - Power minimisation for idle nodes



- Power components:
  - ◆ Active power
  - ◆ Passive power
    - Sub-threshold leakage (source-drain leakage)
  - ◆ Gate leakage



---

## Several ways to reduce power

### ■ Use specific processors

- ▶ Low voltage/frequency
- ▶ Many cores at lower frequencies
  - GPUs
  - MICs
  - FPGAs

### ■ Use any processors with:

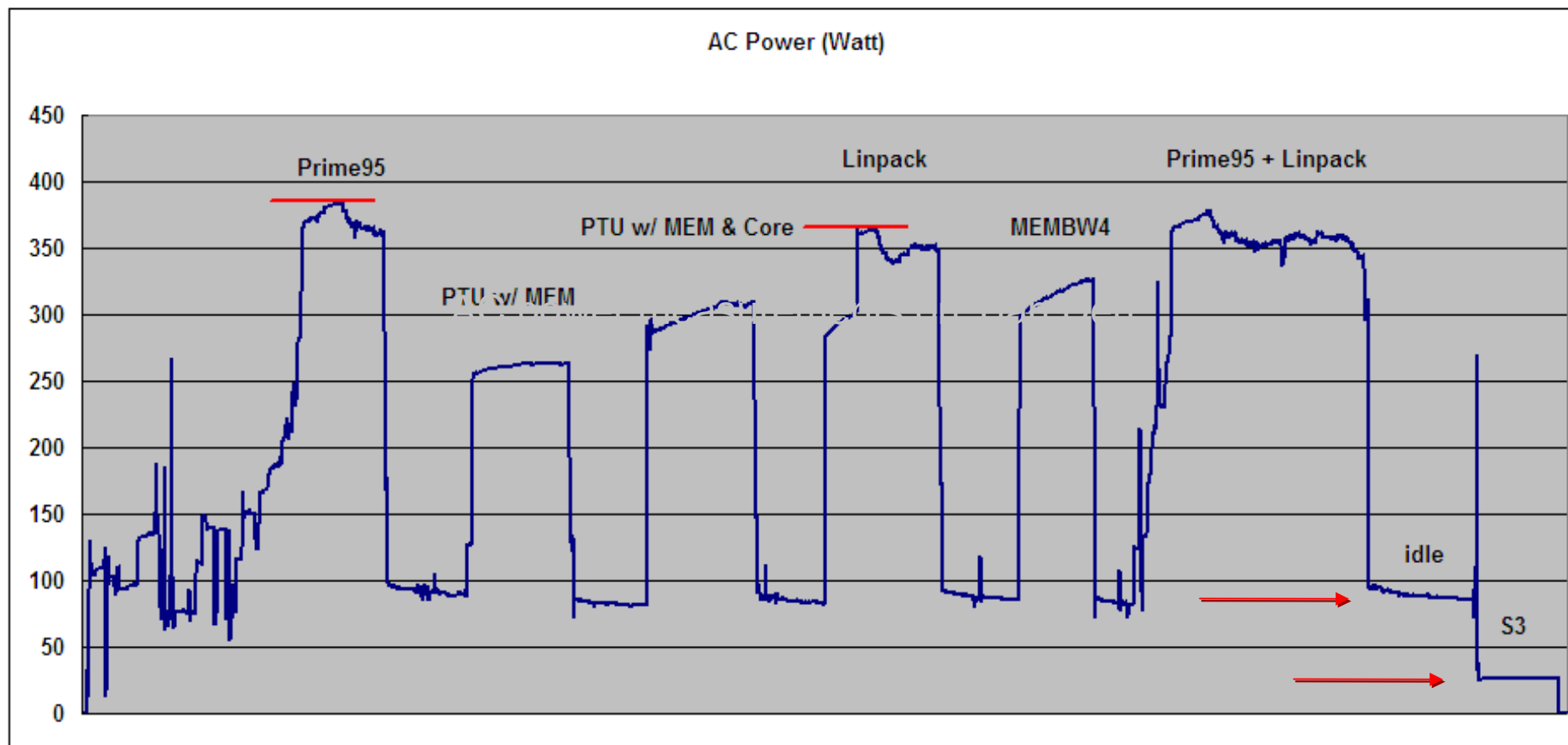
- ▶ New cooling
- ▶ New software

# Green500

## Top 20 du classement Green500 de novembre 2012

Site	Manufacturer	Computer	Mflops/Watt
1 Sciences/University of Tennessee	Appro	Appro Green Blade Xeon E5-2670 8C 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P	2499
2 King Abdulaziz City for Science and Technology	Adtech	Adtech , Xeon E5-2650 8C 2.000GHz, Infiniband FDR, AMD FirePro S10000	2351
3 DOE/SC/Oak Ridge National Laboratory	Cray Inc.	Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x	2143
4 Swiss Scientific Computing Center (CSCS)	Cray Inc.	Cray XK7 , Opteron 6272 16C 2.100GHz, Cray Gemini interconnect, NVIDIA Tesla K20 Kepler	2243
5 Forschungszentrum Juelich (FZJ)	IBM	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	2102
6 Consortium/University of Toronto	IBM	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	2101
7 DOE/NNSA/LLNL	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2101
8 IBM Thomas J. Watson Research Center	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2101
9 IBM Thomas J. Watson Research Center	IBM	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	2101
10 Ecole Polytechnique Federale de Lausanne	IBM	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	2101
11 Computational Modelling, University of Warsaw	IBM	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	2101
12 DOE/SC/Argonne National Laboratory	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2101
13 DOE/SC/Argonne National Laboratory	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2101
14 Rensselaer Polytechnic Institute	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2101
15 University of Rochester	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2101
16 /KEK	IBM	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	2099
17 /KEK	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2099
18 University of Edinburgh	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2099
19 Daresbury Laboratory	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2099
20 CNRS/IDRIS-GENCI	IBM	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	2099

## AC power measurements on dx360m4



---

## What can do from a software perspective ?

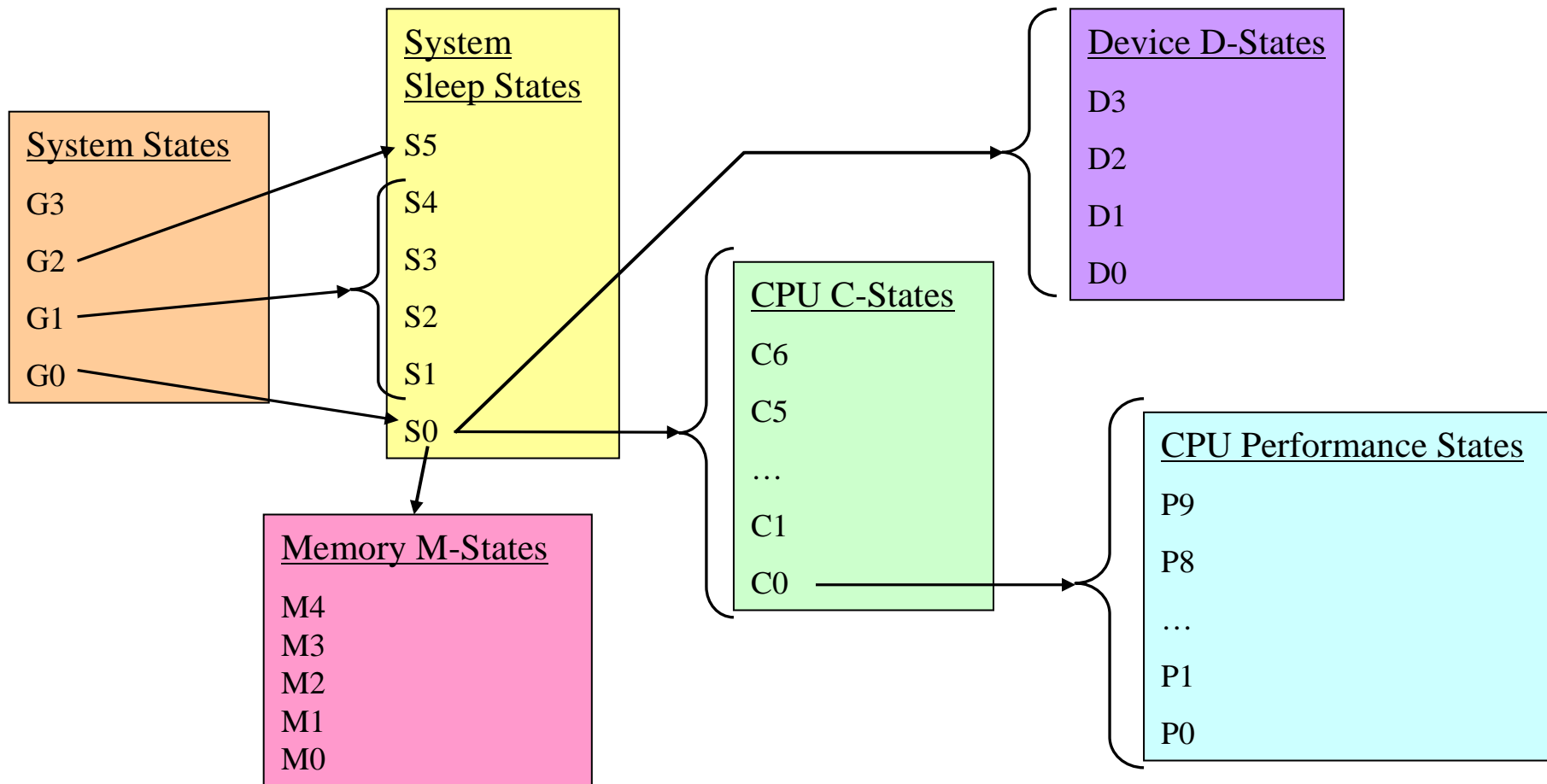
- **Reduce power of inactive nodes**
  - by C- or S-states
- **Reduce power of active nodes**
  - by P-state / CPUfreq



# ACPI State Hierarchy



- ACPI =Advanced Configuration and Power Interface (<http://www.acpi.info/>)
- The ACPI specification defines several system and component states designed to save power.



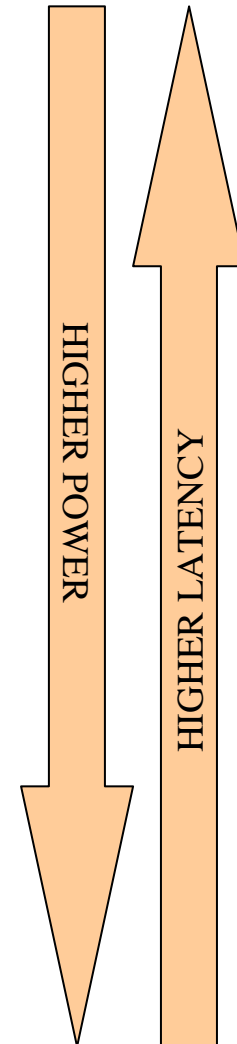
Implementation of power saving states is necessary to recapture lost power when a server or components in a server are idle.

# S-States



S-states are system sleep states that define what **sleep** state the entire server is in.

S-State	G-State	BIOS Reboot	OS Reboot	Comments
S5	G2	Yes	Yes	Server is in a soft off state. When turned back on, the server must completely reinitialize with POST and the operating system.
S4	G1	Yes	No	A.K.A “Hibernate”, “Suspend-to-disk”. The state of the operating system (all memory contents and chip registers) is saved to a file on the HDD and the server is placed in a soft-off state.
S3	G1	No	No	A.K.A “Standby”, “Suspend-to-RAM”. The state of the chipset registers is saved to system memory and memory is placed in a low-power self-refresh state. To preserve the memory contents, power is supplied to the DRAMs in S3 state.
S2	G1	No	No	CPU caches are powered down.
S1	G1	No	No	A.K.A “Idle”, “Standby” –if S3 not supported. Typically, when the OS is idle, it will halt the CPU and blank the monitor to save power. No power rails are switched off. This state may go away on future servers.
S0	G0	No	No	System is fully on but some components could be in a power savings state.



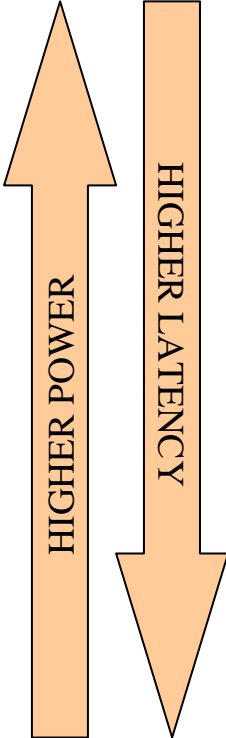
# C-States



C-states are **CPU power saving** states. The CPU transitions to C-states higher than C0 when it is **idle** after a period of time. All C-states occur when the server is in S0 state and G0 state.

Example implementation for Intel Nehalem architecture:

C-State	CPU Core Description	CPU Package Description
C0	CPU core is fully on	CPU package is fully on
C1	CPU core clock is stopped	NA –C1E is the package C1 state. This naming is unique to C1E and is leftover from older CPUs.
C1E	NA –package only state	At least one CPU core is in C0/C1 state and all others are in higher numbered C-states. VRD switches to minimal voltage state.
C3	C1 + CPU core caches are flushed	At least one CPU core is in C3 state and all others are in higher numbered C-states. C1E + Some uncore CLKs stopped and <b>memory placed in fast self-refresh</b>
C6	C3 + CPU cores are powered down. CPU core state stored in last level cache	At least one CPU cores is in C6 state and all others are in higher numbered C-states. C3 package + CLKs stopped on most of the uncore and <b>memory placed in slow self-refresh</b>
C7	C6 + last thread flushes remaining caches ways	All CPU cores are in C7. CPU package doesn't need to wake up for snoops.



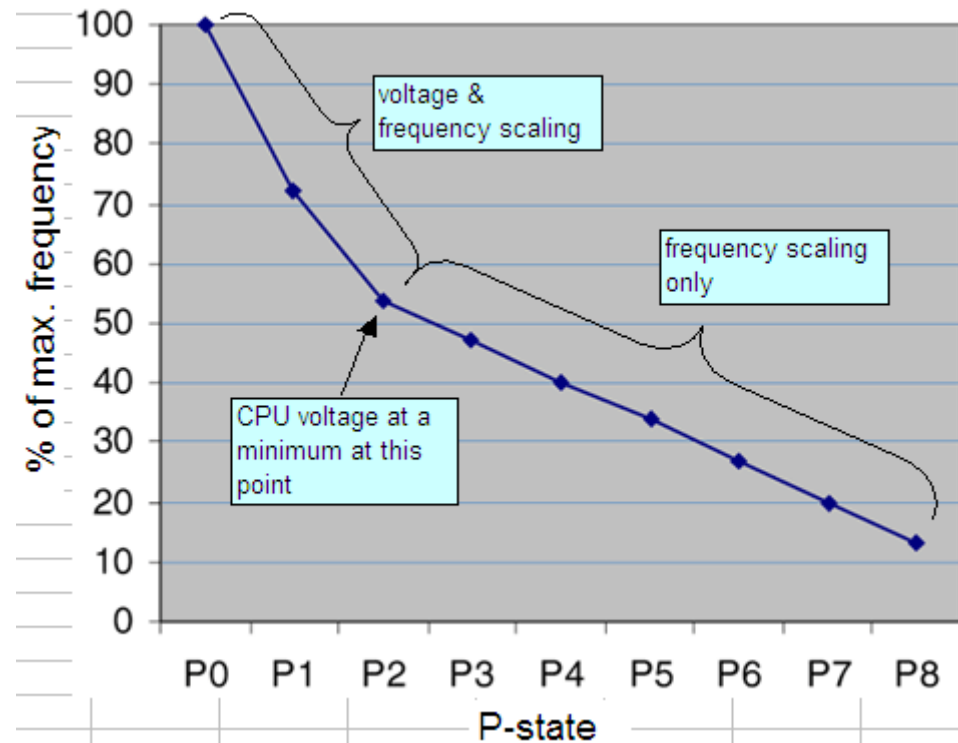
- CPU package refers to all the hardware contained in a CPU chip. “Uncore” refers to all the hardware except for the CPU cores.
- C-states can operate on each core separately or the entire CPU package. Core C-state transitions are controlled by the OS. Package C-state transitions are controlled by the hardware.
- The number of C-states and the savings associated with each is dependent on the specific type and SKU of CPU used.

# P-States



P-states are **CPU performance** states. The OS places the CPU in different P-states depending on the amount of power needed to complete the current task. For example, if a 2GHz CPU core only needs to run at 1 GHz to complete a task, the OS will place the CPU core into a higher number P-state.

P-State	Comments
P0	<b>CPU core running at maximum frequency and voltage</b>
P1	
P2	
P3	
P4	
P5	
P6	
P7	
P8	
P9	
P10	<b>CPU core running at minimum frequency and voltage</b>



- P-states operate on each core separately. The OS controls the transitioning among the P-states.
- The number of P-states and the frequency and voltage associated with each is dependent on the specific type and SKU of CPU used.
- BIOS can restrict the total number of P-states revealed to the OS and can change this on-the-fly. This feature is useful for power capping a system.



## Intel Xeon P-States

Processor	<b>P0</b> (GHz)	<b>P1</b> (GHz)	<b>P2</b> (GHz)	<b>P3</b> (GHz)	<b>P4</b> (GHz)	<b>P5</b> (GHz)	<b>P6</b> (GHz)	<b>P7</b> (GHz)	<b>P8</b> (GHz)	<b>P9</b> (GHz)	<b>P10</b> (GHz)	<b>P11</b> (GHz)
Intel® Xeon® X5670 <b>Turbo Enabled</b>	3.20 3.06	2.93	2.80	2.67	2.54	2.40	2.27	2.14	2.00	1.87	1.74	1.60
Intel® Xeon® X5670 <b>Turbo Disabled</b>	2.93	2.80	2.67	2.54	2.40	2.27	2.14	2.00	1.87	1.74	1.60	

## Xeon 2 S C-States

Processor	Intel® Xeon® 5600 Series (Westmere-EP)	Intel® Xeon® 5500 Series (Nehalem-EP)	Intel® Xeon® 5400 Series (Harpertown)
Core C1	✓	✓	✓
Package C1/C1E	✓	✓	✓
Core C3	✓	✓	✓
Package C3	✓	✓	
Core C6	✓	✓	
Package C6	✓	✓	

		Processor	Intel® Xeon® 5600 Series (Westmere)	Intel® Xeon® 5500 Series (Nehalem)	Intel® Xeon® 5400 Series (Harpertown)
Interrupt Service Latency	Core C1		<1us <sup>2,3</sup>	<1us <sup>2,3</sup>	<0.1us <sup>3</sup>
	Package C1E		<1us <sup>2,3</sup>	<1us <sup>2,3</sup>	~5us
	Core C3		TBD, being measured	<40us <sup>2,3</sup>	TBD
	Package C3		TBD, being measured	<100us <sup>2,3</sup>	Not supported
	Core C6		TBD, being measured	<60us <sup>2,3</sup>	Not supported
	Package C6		TBD, being measured	<100us <sup>2,3</sup>	Not supported
Processor Power	Core C1		Not specified	Not specified	Not specified
	Package C1E		22-40W <sup>1</sup>	~30W <sup>1</sup>	~16W <sup>1</sup>
	Core C3		Not specified	Not specified	Not specified
	Package C3		18W-33W <sup>1</sup>	~26W <sup>1</sup>	Not supported
	Core C6		~0W	~0W	Not supported
	Package C6		8W-14W <sup>1</sup>	~10W <sup>1</sup>	Not supported

Source: Intel latency and processor power from data sheets and/or BIOS writers guide.

<sup>1</sup> Varies by SKU

<sup>2</sup> Intel general guidance for average entrance/exit latency times from Intel BIOS writers guide March 2010

<sup>3</sup> Collision of P-state transition with C-state exit can add ~2us latency



- **Between Vmax and Vmin, frequency is changed with voltage**
- **Frequency reduction implies power reduction**
  - But not like  $f^3$  which is true for the processor but not for the other node component
- **Frequency reduction implies performance reduction**
  - Best as  $f$ , but could be less depending on the application/use case profile



# Power and Performance of JS22 and HS21

## JS22 4.0 GHz

Application	Average Power (watts)					
	Total	CPU	DIMM	Other	CPI	GBS
416.gamess	289	87	14	102	1,3	0,0
433.milc	306	76	51	103	6,8	16,3
435.gromacs	292	87	15	102	1,5	0,7
437.leslie3d	326	85	50	105	2,6	16,5
444.namd	296	89	14	104	1,4	0,3
454.calculix	301	91	18	103	1,0	1,9
459.GemsFDTD	315	80	49	106	5,1	15,8
481.wrf	311	84	39	103	1,5	12,7
Idle	212	48	14	102		

## HS21 2.8 GHz

Application	Average Power (watts)					
	Total	CPU	DIMM	Other	CPI	GBS
416.gamess	366	106	15	62	0,6	0,0
433.milc	321	64	30	66	9,8	6,2
435.gromacs	363	102	17	63	0,6	1,2
437.leslie3d	328	68	30	67	8,6	6,3
444.namd	356	100	15	64	0,7	0,2
454.calculix	379	106	20	64	0,6	2,2
459.GemsFDTD	323	66	29	66	9,5	6,1
481.wrf	329	69	29	66	5,2	6,1
idle	210	24	15	66		

Systems	Processors	Nominal Frequency	Memory	Tools
JS22 2 Sockets 2 cores	IBM Power6 2 core processor	4 GHz	4 x 4GB, 667 MHz DDR2	AEM, Amester
HS21 2 Sockets 2 cores	Intel Harpertown 4 core processor	2.86 GHz	8 x 2GB, 667 MHz DDR2	AEM, Amester

“CPU” includes N processor cores, L1 cache + NEST (memory, fabric, L2 and L3 controllers,..)

“Other” includes, L2 cache, Nova chip, IOChip VRM losses, etc.

# Power and Performance of p575 and

## p755

p575 4.7 GHz

Application	Average Power (watts)						
	Total	CPU	DIMM	L3	CPI	GBS	
416.gamess	4136	2754	263	741	205	1,4	0,2
433.milc	4103	2318	612	807	203	8,9	137,6
435.gromacs	4002	2624	263	741	204		
437.leslie3d	4466	2665	611	816	204	2,1	133,4
444.namd	4103	2722	264	741	205	1,4	0,8
454.calculix	4293	2883	285	747	205	1,1	7,7
459.GemsFDTD	4154	2379	599	805	205	5,0	126,8
481.wrf	4237	2658	426	778	205	1,7	51,3
idle	2548	1350	262	728	176		

p750 3.55 GHz

Application	Average Power (watts)						
	Total	CPU	DIMM	Other	CPI	GBS	
416.gamess	984	724	110	151	0,6	0,1	
433.milc	1084	646	268	171	2,7	40,0	
435.gromacs	1002	705	144	153	0,7	1,0	
437.leslie3d	1141	699	266	176	0,9	39,4	
444.namd	1001	717	135	149	0,7	0,4	
454.calculix	1041	747	142	151	0,5	3,1	
459.GemsFDTD	1070	645	264	161	2,0	38,5	
481.wrf	1147	750	234	163	0,6	29,2	
Idle	768	517	98	153			

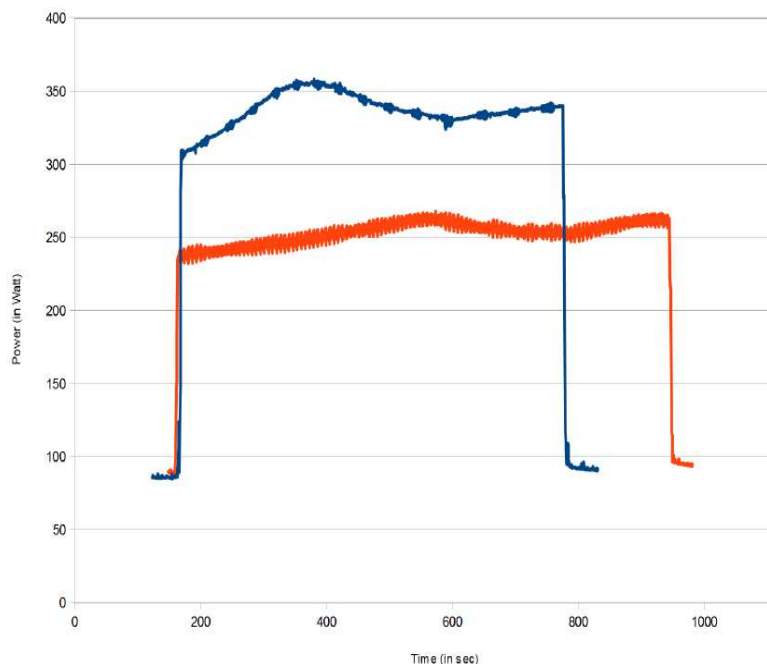
	Memory	Tools
4.7 GHz	64 x 2GB, 667 MHz DDR2	Another lab tool
3.55 GHz	32x 4GB, 1066 MHz DDR3	AEM, Amester

“Core” includes N processor cores, L1 cache + NEST (memory, fabric, L2 and L3 controllers,..)

“Other” includes, L2 cache, Nova chip, IOChip VRM losses, etc.

## Example: what happens when you just change frequency

Quantum ChromoDynamics Application



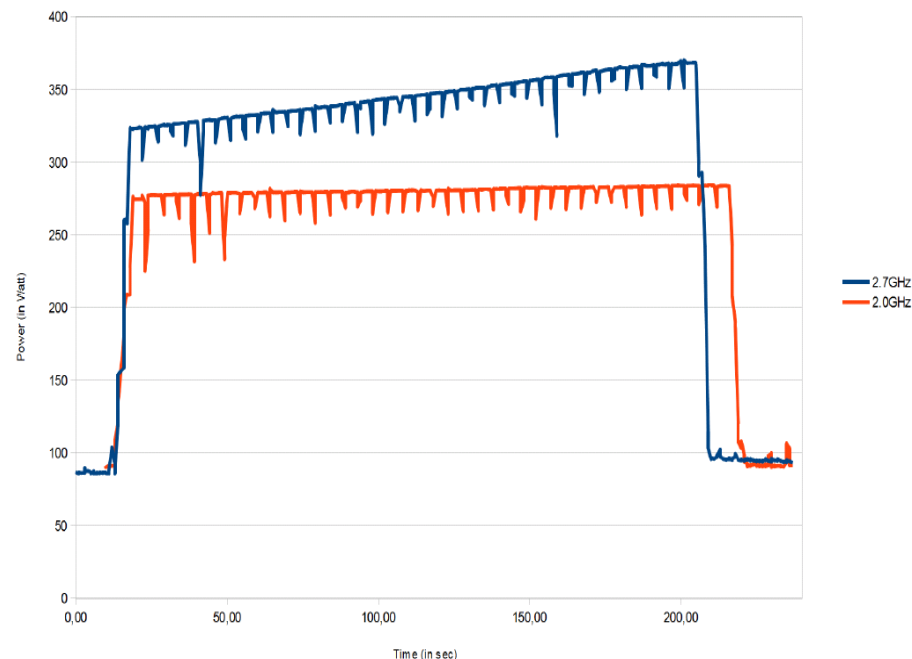
$\Delta f = -26\%$

$\Delta \text{Power} = -26\%$

$\Delta \text{Time} = +26\%$

$\Delta \text{Energy} = \sim 0\%$

Astrophysics Application



$\Delta f = -26\%$

$\Delta \text{Power} = -17\%$

$\Delta \text{Time} = +5\%$

$\Delta \text{Energy} = -12\%$

# How to find the performance/power trade-off ?

---



- **Build a performance model**
  - Which depends on the processor/node and the application
- **Build a power consumption model**
  - Which depends on the processor/node and the application
- **Express the trade-off by simple policy which the user/sysadmin will select**

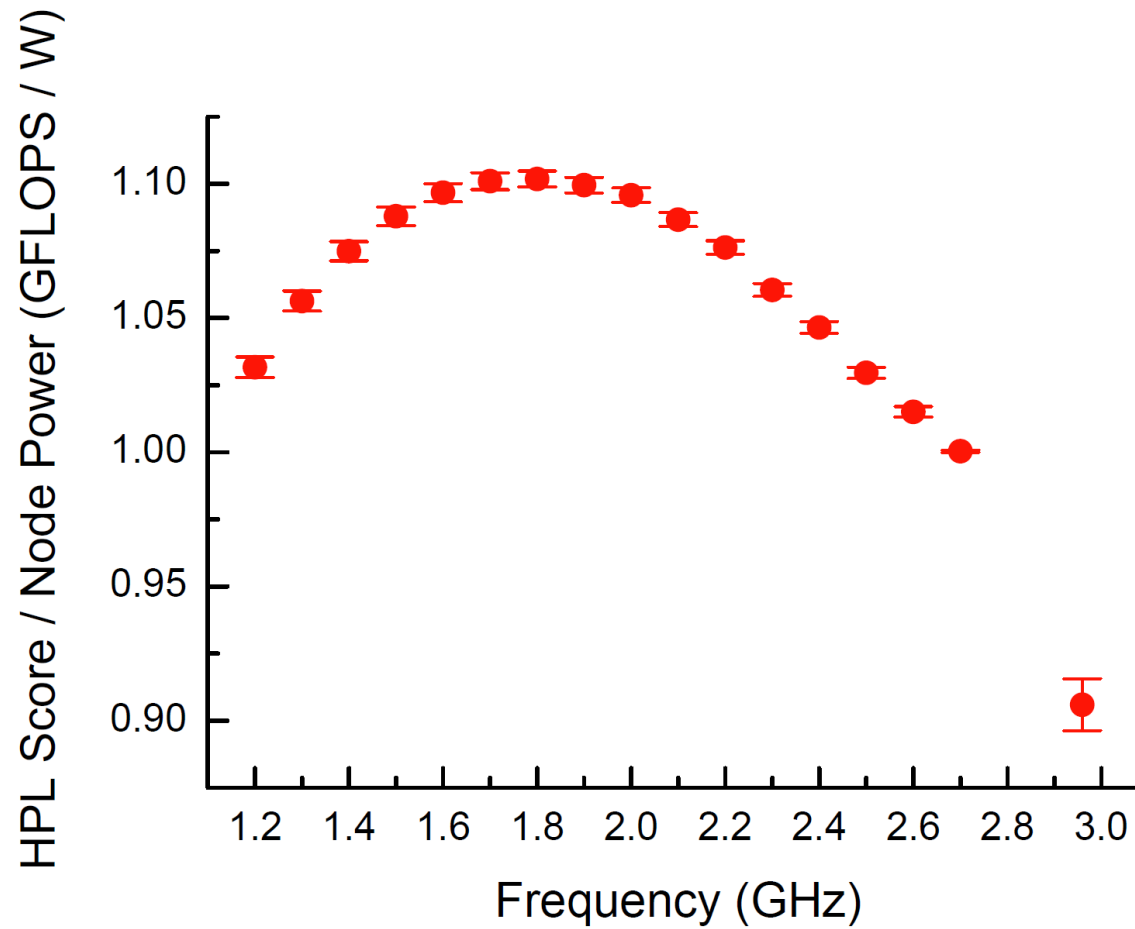
# Is it worth using Turbo ?

---



- **Not really**
- **Not yet**

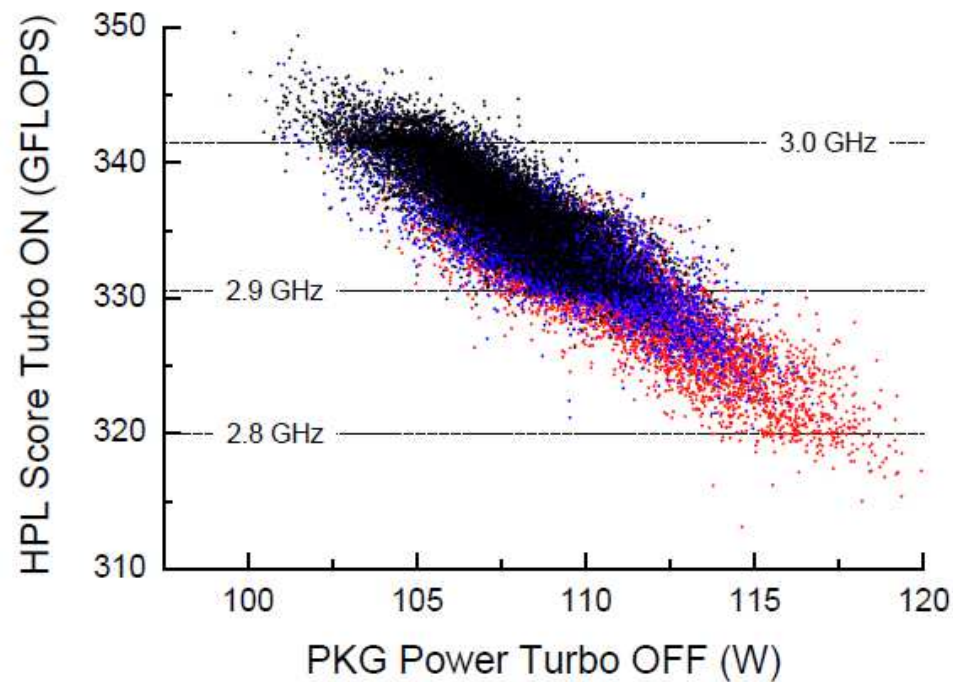
# Energy Efficiency IBM iDataPlex DWC dx360EM4



IBM Energy Aware Scheduling

# IBM System x iDataPlex Direct Water Cooled dx360 M4

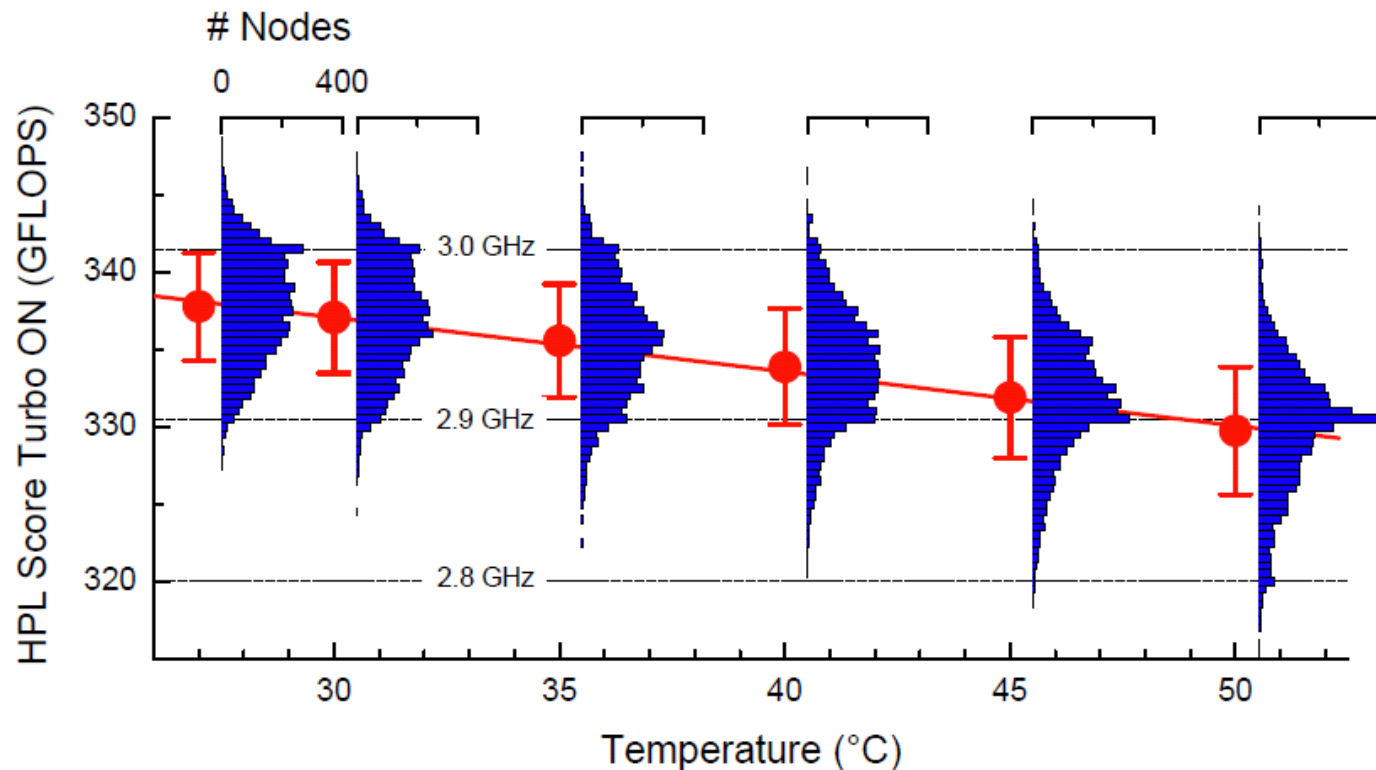
**2x Intel SB-EP 2.7 GHz 130 W. 8x Samsung 4**



Ingmar Meijer, 2012

# IBM System x iDataPlex Direct Water Cooled dx360 M4

**2x Intel SB-EP 2.7 GHz 130 W. 8x Samsung 4**



$$\partial \text{HPL} / \partial T = -0.350 \pm 0.013 \text{ GFLOPS} / \text{K}$$

Ingmar Meijer, 2012

IBM Energy Aware Scheduling

© 2013 IBM Corporation



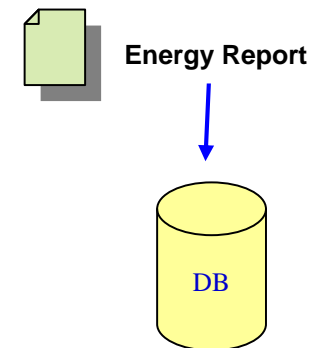
# IBM Energy Aware Scheduling

## ■ Report

- ▶ temperature and power consumption per node
- ▶ temperature, power consumption and energy per per job
- ▶ total power consumption of the cluster

## ■ Optimize

- ▶ Reduce power of inactive nodes
- ▶ Optimize energy of active nodes



## Features available to reduce and control power

### ■ xCAT

#### ► 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 consumed info, CPU usage, fan speed, environment temperature
- Set: Power saving mode , Power capping value, Deep Sleep (S3 state)

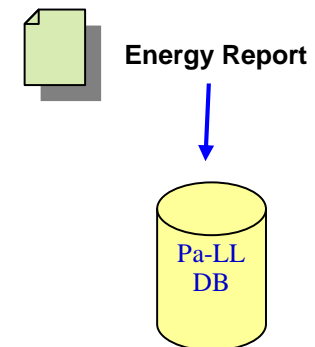
### ■ LL (and soon LSF)

#### ► Report power and energy consumption per job

- Energy report is created and stored in the DB

#### ► Optimize power and energy consumption per job

- Optimize power of idle nodes:
  - set nodes at lowest power consumption when no workload is scheduled on this set of nodes
- Optimize power of active nodes:
  - set nodes at optimal processor frequency according to an energy policy for a given parallel workload (i.e minimize energy with maximum performance degradation)



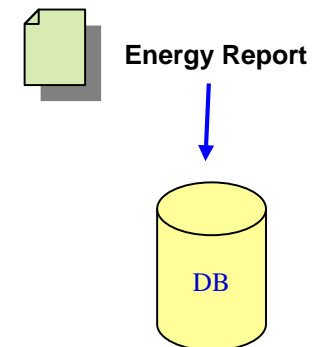
# IBM software to monitor and reduce power

## ■ Report

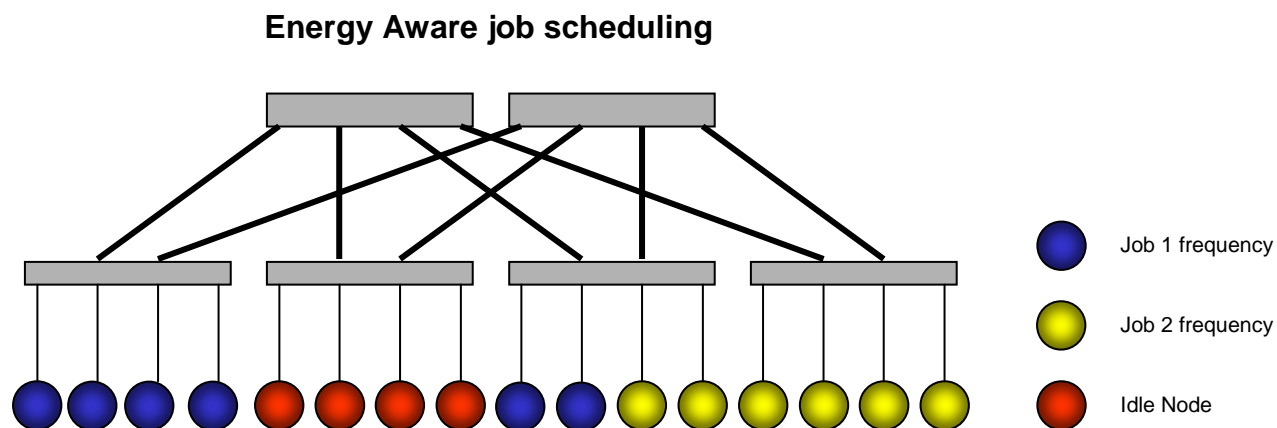
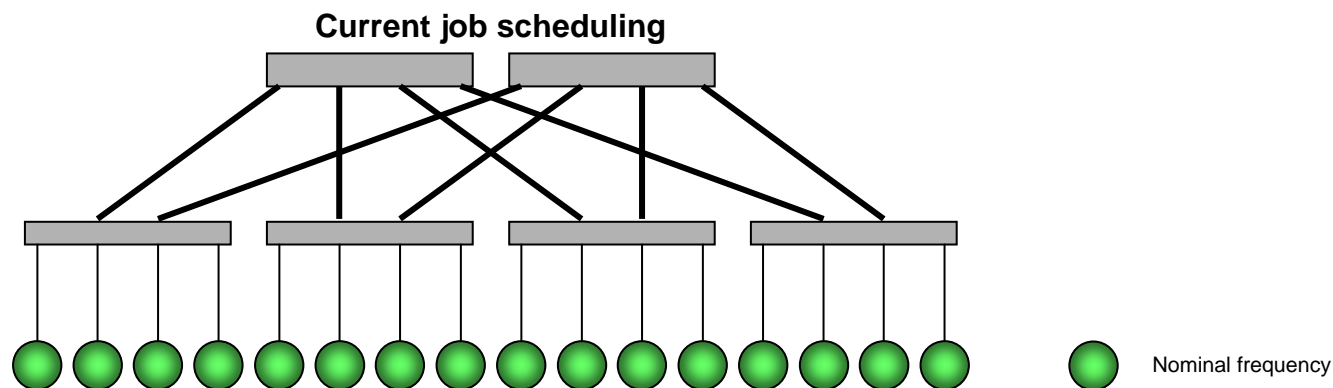
- ▶ Temperature, fan speed and power consumption per node
- ▶ power consumption, energy and performance per job

## ■ Optimize

- ▶ Reduce power of inactive nodes
- ▶ Reduce power of active nodes



# Energy Aware Scheduling



Before each job is submitted, change the state/frequency of the corresponding set of nodes to match a given energy policy defined by the Sys Admin

## How LL-EAS manages idle nodes

- When a job has completed on a set of nodes, LL set those nodes in a state which does let the OS to turn them into C6 state
- When nodes are idle and no jobs are in queue, LL will ask xCAT to put them into S3 state according to the idle power policy
  - Idle power policy is determined by the system admin
- When new jobs are submitted which, according to the idle power policy, require nodes to be awaked , LL asks xCAT to resume the desired nodes from S3 before it submits the job

## LL-EAS phases to set optimal frequency for jobs

- **Learning phase**
  - LL evaluates the power profile of all nodes and store it in the xCAT/LL DB
- **System admin defines a default frequency for the cluster**
  - Can be nominal frequency or a lower frequency
- **User submit a job**
  - User submit his job with a tag
  - Job is run at default frequency
  - In the background:
    - LL measures power, energy, time and hpm counters for the job
    - LL predicts power(i), energy(i), time (i) if job was run a different frequency i
  - LL writes Energy report for the job in the xCAT/LL DB
- **User resubmit a job with same tag**
  - Given the energy policy and the tag, LL determines optimal frequency j
  - LL set nodes for the job at frequency j
  - In the background:
    - LL measures power, energy, time and hpm counters for the job
    - LL compares measurement and prediction, and provide correction actions if needed
  - LL add new record with new energy report for the job in the xCAT/LL DB

## LL-EAS energy policies available

### ■ **Predefined policy**

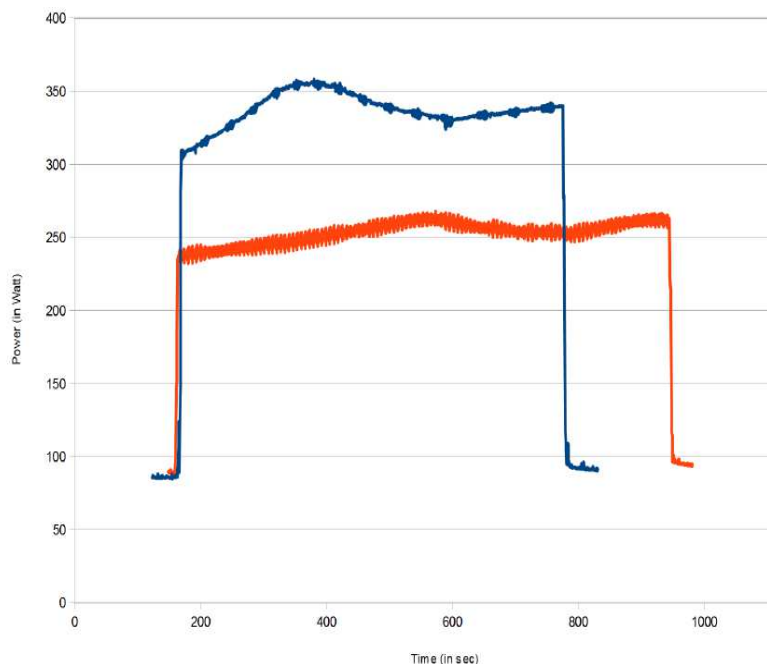
- Minimize Energy within max performance degradation bound of X%
  - LL will determine the frequency (lower than default) to match the X% performance degradation while energy savings is still positive
- MinimizeTime to Solution
  - LL will determine a frequency (higher than default) to match a table of expected performance improvement provided by sysadmin
  - This policy is only available when default frequency < nominal frequency
- Set Frequency
  - User provides the frequency he wants his jobs to run
  - This policy is available for authorized user only
- Policies thresholds are dynamic, i.e values can be changed any time and will be taken into account dynamically

### ■ **Site provided policy**

- Sysadmin provides an executable which set the frequency based on the information stored in the DB

## Example: what happens when you just change frequency

Quantum ChromoDynamics Application



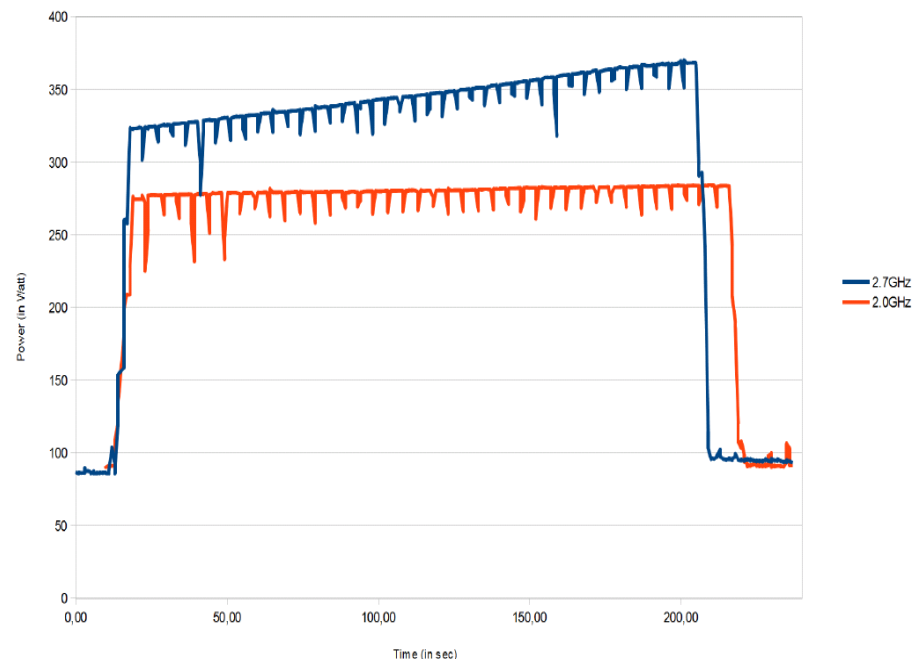
$\Delta f = -26\%$

$\Delta \text{Power} = -26\%$

$\Delta \text{Time} = +26\%$

$\Delta \text{Energy} = \sim 0\%$

Astrophysics Application



$\Delta f = -26\%$

$\Delta \text{Power} = -17\%$

$\Delta \text{Time} = +5\%$

$\Delta \text{Energy} = -12\%$



## Example: how to submit a job first time

```
#!/bin/bash
# @ job_name = test
# @ account_no = 99999
# @ class = parallel
# @ job_type = MPICH
# @ network.MPI = sn_all,,US
# @ total_tasks = 128
# @ node = 8
# @ output = $(jobid)_output
# @ error = $(jobid)_error
# @ initialdir = /bench/gpfs/fs1/users/fthomas/lleas/Astrophysics
# @ node_usage = not_shared
# @ energy_policy_tag = Astro
# @ energy_output = energy.dat
# @ queue

. ~/.bashrc
```

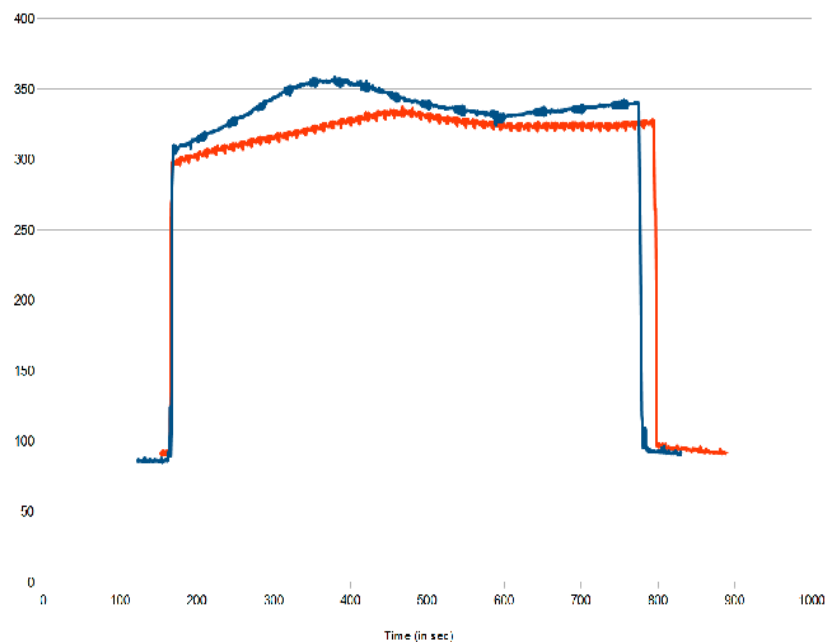
## Example: how to submit a job with a policy

```
#!/bin/bash
# @ job_name = test
# @ account_no = 99999
# @ class = parallel
# @ job_type = MPICH
# @ network.MPI = sn_all,,US
# @ total_tasks = 128
# @ node = 8
# @ output = $(jobid)_output
# @ error = $(jobid)_error
# @ initialdir = /bench/gpfs/fs1/users/fthomas/lleas/Astrophysics
# @ node_usage = not_shared
# @ energy_policy_tag = Astro
# @ energy_output = energy.dat
# @ max_perf_decrease_allowed = 5
# @ queue

. ~/.bashrc
```

## Example: what happens with max perf degrad policy=5%

Quantum ChromoDynamics Application



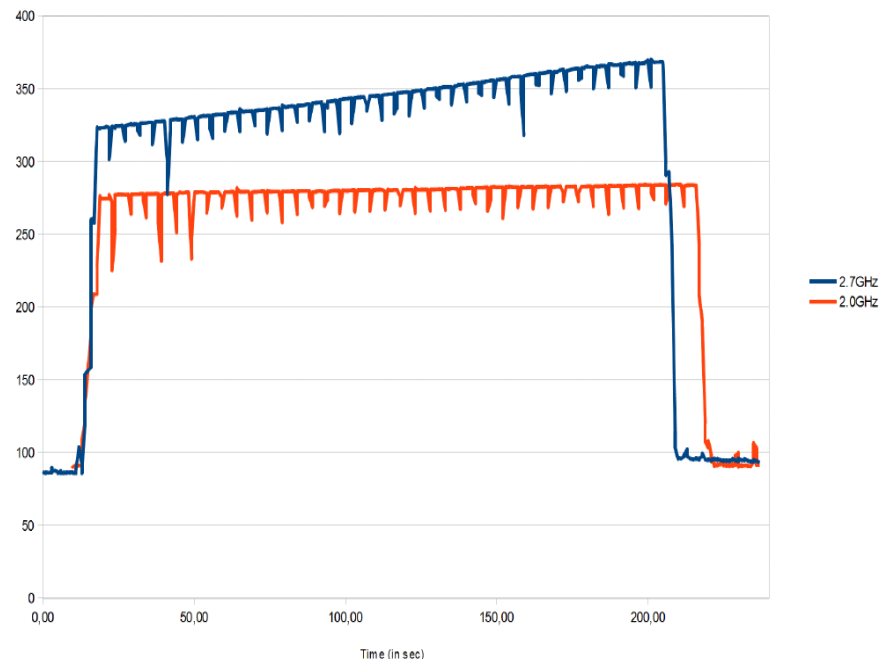
**f= 2.6 GHz**

**$\Delta$ Power=-5%**

**$\Delta$ Time=+2%**

**$\Delta$ Energy=-3%**

Astrophysics Application



**f=2.0 GHz**

**$\Delta$ Power=-17%**

**$\Delta$ Time=+5%**

**$\Delta$ Energy=-12%**

## Examples of savings

- **1000 node cluster, 0.15€ per KWh**
  - Linpack power consumption per year = 442K€
- **Inactive nodes**
  - ▶ With 80% workload activity and nodes in S3 half of the idle time (10% of overall time)
  - ▶ Savings per year = 24.5 K€
- **Active nodes**
  - ▶ With a 3% performance degradation threshold, , about 8% power ca be saved (see examples)
  - ▶ Savings per year = 20.4 K€
- ▶ **Total savings: 45K€, ~10%**



## BQCD : Energy report for 1K and 8K tasks ,

perf., power

Clock	CPI	Time	Power	Energy	PerfVa	PwrVa	EnyVar	Clock	CPI	Time	Power	Energy	PerfVa	PwrVar	EnyVar
2700	1,075	509	308	0,0435	0	0	0	2700	0,661	304	290	0,0244	0	0	0
→ 2600	1,062	522	290	0,0420	-2,6%	5,8%	3,3%	→ 2600	0,651	311	273	0,0236	-2,2%	5,7%	3,6%
2500	1,038	531	280	0,0413	-4,3%	8,8%	4,9%	2500	0,645	320	263	0,0234	-5,3%	9,2%	4,4%
2400	1,015	540	275	0,0413	-6,2%	10,6%	5,0%	2400	0,634	328	257	0,0235	-7,9%	11,1%	4,1%
2300	0,994	552	261	0,0400	-8,5%	15,3%	8,0%	2300	0,626	338	244	0,0229	-11,1%	15,6%	6,2%
2200	0,972	565	255	0,0399	-10,9%	17,2%	8,1%	2200	0,620	350	237	0,0231	-15,2%	18,1%	5,6%
2000	0,932	596	237	0,0393	-17,1%	22,8%	9,6%	2000	0,598	372	222	0,0229	-22,2%	23,3%	6,3%
1900	0,908	611	228	0,0386	-20,0%	25,9%	11,1%	1900	0,593	387	213	0,0229	-27,4%	26,4%	6,2%
1800	0,894	635	220	0,0388	-24,7%	28,4%	10,8%	1800	0,584	403	206	0,0230	-32,5%	29,0%	5,9%
1700	0,877	659	212	0,0388	-29,6%	31,1%	10,7%	1700	0,581	424	199	0,0234	-39,6%	31,4%	4,2%
1600	0,848	677	207	0,0390	-33,0%	32,6%	10,4%	1600	0,575	446	194	0,0240	-46,7%	33,2%	1,9%
1500	0,831	708	199	0,0392	-39,2%	35,2%	9,8%	1500	0,571	473	186	0,0244	-55,5%	35,8%	0,1%
1400	0,821	750	188	0,0391	-47,3%	38,9%	10,0%	1400	0,566	502	175	0,0244	-65,1%	39,5%	0,1%
1300	0,807	794	179	0,0394	-55,9%	41,9%	9,4%	1300	0,563	538	167	0,0249	-76,9%	42,3%	-2,0%
1200	0,797	849	170	0,0400	-66,7%	44,8%	7,9%	1200	0,556	575	158	0,0252	-89,2%	45,4%	-3,2%

# UM: Energy Report

perf., power



Clock (MHz)	CPI	Time (s)	Power (Watt)	Energy (KW/h)	PerfVar (%)	PowerVar(%)	EnergyVar (%)
2700	0,986	158	274	0,0120	0	0	0
→ 2600	0,977	163	259	0,0117	-2,9%	5,3%	2,6%
2500	0,970	168	249	0,0116	-6,2%	9,1%	3,4%
2400	0,956	172	243	0,0116	-9,1%	11,3%	3,2%
2300	0,946	178	232	0,0114	-12,6%	15,4%	4,7%
2200	0,938	184	224	0,0115	-16,8%	18,2%	4,4%
2000	0,915	198	210	0,0115	-25,2%	23,4%	4,0%
1900	0,905	206	202	0,0116	-30,5%	26,3%	3,8%
1800	0,897	216	195	0,0116	-36,5%	28,9%	3,0%
1700	0,891	227	188	0,0119	-43,6%	31,3%	1,3%
1600	0,880	238	183	0,0121	-50,6%	33,2%	-0,6%
1500	0,873	252	175	0,0123	-59,4%	36,0%	-2,1%
1400	0,867	268	166	0,0123	-69,6%	39,5%	-2,6%
1300	0,861	287	158	0,0126	-81,4%	42,4%	-4,5%
1200	0,854	308	149	0,0127	-94,9%	45,6%	-6,0%

# Saturne: Energy Report

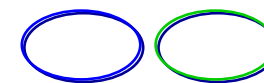
perf., power



Clock (MHz)	CPI	Time (s)	Power (Watt)	Energy (KW/h)	PerfVar (%)	PowerVar(%)	EnergyVar (%)
2700	0,618	109	248	0,0075	0	0	0
→ 2600	0,609	111	236	0,0073	-2,3%	4,9%	2,7%
2500	0,607	116	226	0,0072	-6,1%	9,1%	3,6%
2400	0,599	119	219	0,0072	-9,1%	11,9%	3,8%
2300	0,594	123	210	0,0072	-12,7%	15,3%	4,5%
2200	0,592	128	201	0,0072	-17,6%	18,9%	4,5%
2000	0,575	137	189	0,0072	-25,5%	23,8%	4,3%
1900	0,573	144	183	0,0073	-31,8%	26,4%	3,1%
1800	0,566	150	176	0,0073	-37,4%	29,2%	2,7%
1700	0,566	158	171	0,0075	-45,4%	31,3%	0,1%
1600	0,565	168	165	0,0077	-54,2%	33,7%	-2,3%
1500	0,564	179	157	0,0078	-64,1%	36,6%	-4,1%
1400	0,560	190	149	0,0079	-74,7%	39,9%	-5,0%
1300	0,559	205	142	0,0081	-87,9%	42,7%	-7,7%
1200	0,553	219	133	0,0081	-101,2%	46,3%	-8,1%

# Ramses: Energy Report:

perf., power

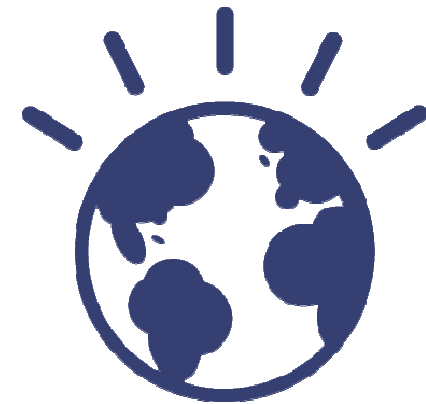


Clock (MHz)	CPI	Time (s)	Power (Watt)	Energy (KW/h)	PerfVar (%)	PowerVar(%)	EnergyVar (%)	Clock (MHz)
2700	3,639	189	288	0,0151	0	0	0	2700
2600	3,619	189	275	0,0144	0,0%	4,7%	4,7%	2600
2500	3,525	190	269	0,0142	-0,5%	6,7%	6,2%	2500
2400	3,442	191	263	0,0140	-1,1%	8,7%	7,7%	2400
→ 2300	3,370	193	256	0,0137	-2,1%	11,4%	9,5%	2300
2200	3,274	195	248	0,0134	-3,2%	14,0%	11,3%	2200
2000	3,164	200	239	0,0133	-5,8%	17,0%	12,2%	2000
1900	3,058	203	232	0,0131	-7,4%	19,7%	13,8%	1900
1800	3,023	206	224	0,0128	-9,0%	22,5%	15,5%	1800
1700	2,948	211	217	0,0127	-11,4%	24,8%	16,3%	1700
1600	2,815	215	210	0,0125	-13,8%	27,2%	17,2%	1600



## Functions planed in LSF

- Energy Aware Scheduling is being ported into LSF
  - ▶ First features to be available 2Q13
    - Energy report (with no prediction)
    - Idle node power management
    - Set frequency policy
  - ▶ Full features available 3Q13
    - Full energy report
    - All Energy Policies



## 3 PFlops SuperMUC system at LRZ

### ■ Fastest Computer in Europe on Top 500 June 2012

- ▶ 9324 Nodes with 2 Intel Sandy Bridge EP CPUs
- ▶ 3 PetaFLOP/s Peak Performance
- ▶ Infiniband FDR10 Interconnect
- ▶ Large File Space for multiple purpose
  - 10 PetaByte File Space based on IBM GPFS with 200GigaByte/s aggregated I/O Bandwidth
  - 2 PetaByte NAS Storage with 10GigaByte/s aggregated I/O Bandwidth



### ■ Innovative Technology for Energy Effective Computing

- ▶ Hot Water Cooling
- ▶ Energy Aware Scheduling

### ■ Most Energy Efficient high End HPC System

- ▶ PUE 1.1
- ▶ Total Power consumption over 5 years to be reduced by ~ 37% from 27.6 M€ to 17.4 M€