Energy Aware Computing

Power Consumption of Clusters Control and Optimization

PDP2014, Feb 12-14, Turin

Luigi Brochard (luigi.brochard@fr.ibm.com)
Raj Panda (panda@us.ibm.com)
Francois Thomas (ft@fr.ibm.com)

ethink High Performance Computing.
ata-intensive. Energy-efficient. Intuitive.

The Power Problem

A 1000 node cluster with
2 x86 sockets, 8 cores, 2.7 Ghz
consumes **340 kW** (Linpack)
not including cooling

In Europe (0.15€ per Kwh)
441K€ per year
In US (0.10$ per Kwh)
US$ 295K per year
In Asia (0.20$ per Kwh)
US$ 590K per year

Several ways to reduce power

Use better cooling (Direct Water Cooling)
Reduce power distribution losses
Choose processors with high Flops/Watt
Use power and energy aware software tools
Tune the applications

Several ways to reduce power

Data center (PUE reduction)

- Use better cooling (Direct Water Cooling)
- Reduce power distribution losses

Hardware, microprocessor technologies

- Choose processors with high Flops/Watt

Software

- Use power and energy aware software tools
- Tune the applications

Several ways to reduce power

Before your RFP starts

- Use better cooling (Direct Water Cooling)
- Reduce power distribution losses

Outcome of your RFP

- Choose processors with high Flops/Watt

During the lifetime of you supercomputer

- Use power and energy aware software tools
- Tune the applications

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
JS22 2 Sockets 2 cores	IBM Power6	4 GHz	4 x 4GB, 667 MHz DDR2
HS21 2 Sockets 4 cores	Intel Harpertown	2.86 GHz	8 x 2GB, 667 MHz DDR2

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

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

Power and Performance of iDataplex dx360 M4

Idataplex dx360 M4 – dual Sandy Bridge 2.7 Ghz (SSE42 binaries)

Application	Average Power (watts)				Perf metrics	
	Total	Core	DIMM	Other	CPI	GBS
416.gamess	275	100	5	71	0.9	0.3
433.milc	330	99	55	77	2.3	68.6
435.gromacs	260	95	5	65	1.2	5.0
437.leslie3d	332	99	57	78	3.1	65.0
444.namd	252	92	5	64	0.9	1.0
454.calculix	274	96	8	74	0.8	11.6
459.GemsFDTD	320	95	57	73	2.4	63.1
481.wrf	330	98	53	82	1.8	65.1
idle	85	6	5	68		

Idataplex dx360 M4 – dual Sandy Bridge 2.7 Ghz (AVX binaries)

Application	Average Power (watts)				Perf metrics	
	Total	Core	DIMM	Other	CPI	GBS
416.gamess	275	100	5	71	0.9	0.3
433.milc	327	97	55	78	2.4	68.5
435.gromacs	264	97	5	65	1.3	4.9
437.leslie3d	335	101	56	77	4.5	65.0
444.namd	253	90	5	68	1.0	1.0
454.calculix	281	100	8	73	0.9	12.5
459.GemsFDTD	320	95	57	73	2.4	62.5
481.wrf	332	101	53	77	2.2	65.2
idle	85	6	5	68		

Systems	Processors	Nominal Frequency	Memory
iDataplex dx360M4 2 Sockets 8 cores	Intel Sandy Bridge	2.7 GHz	8 x 16GB, 1600 MHz DDR3

Power and Performance comparison of Nehalem and Sandy Bridge systems (3-4 years apart)

Application	Instances/hour		Energy/instance	
	NHM	SNB	NHM	SNB
416.gamess	35	83	24	12
433.milc	69	145	12	8
435.gromacs	91	242	9	4
437.leslie3d	51	100	17	12
444.namd	75	159	11	6
454.calculix	94	223	9	4
459.GemsFDTD	40	84	21	14
481.wrf	72	145	12	8

Throughput per core is **conserved**

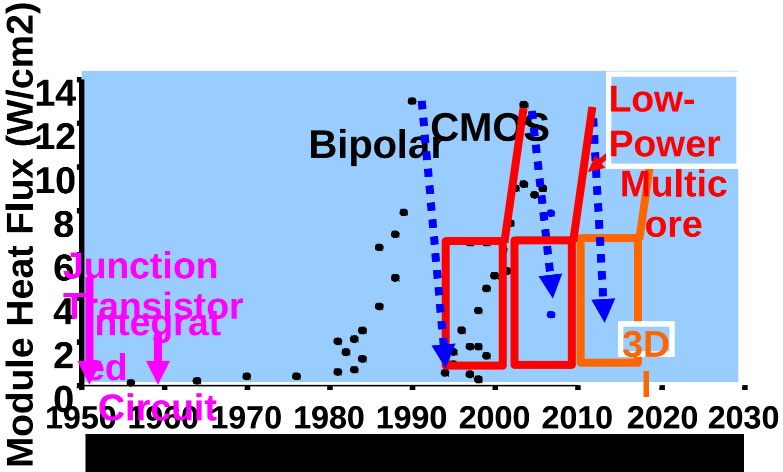
Energy per job is **halved** (not exactly true for memory intensive jobs)

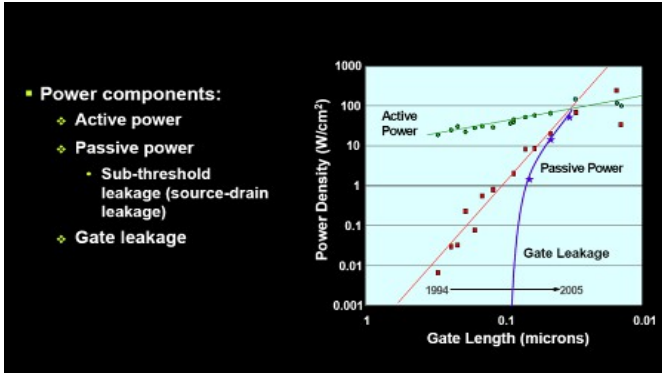
The Power Equation

Power=capacitance*voltage^2*frequency
Power~capacitance*frequency^3

Active power problem
- Control frequency of active nodes

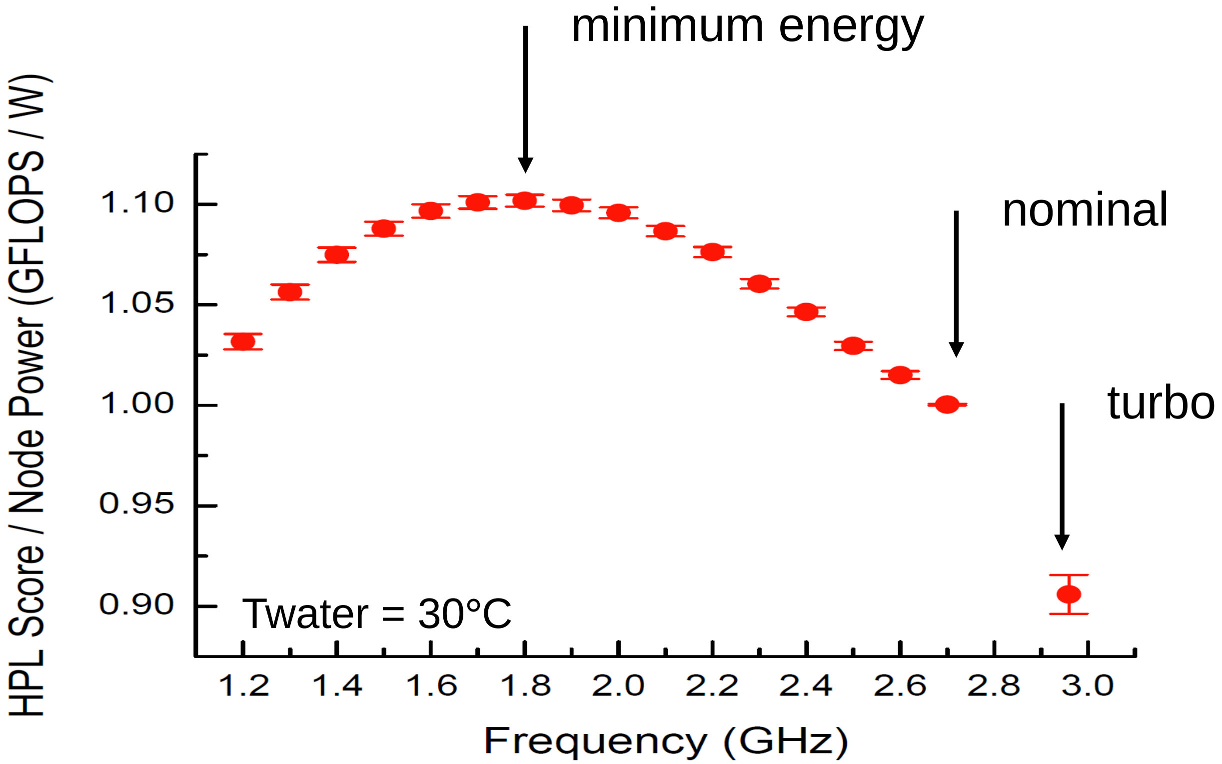
Passive power problem
- Minimize idle nodes power





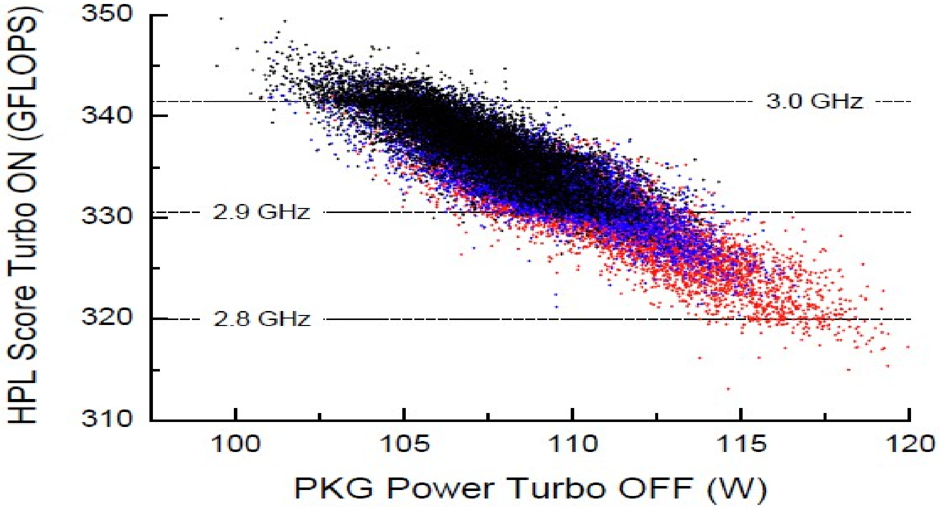
Is it worth using Turbo ?

Energy Efficiency IBM iDataPlex DWC dx360 M4



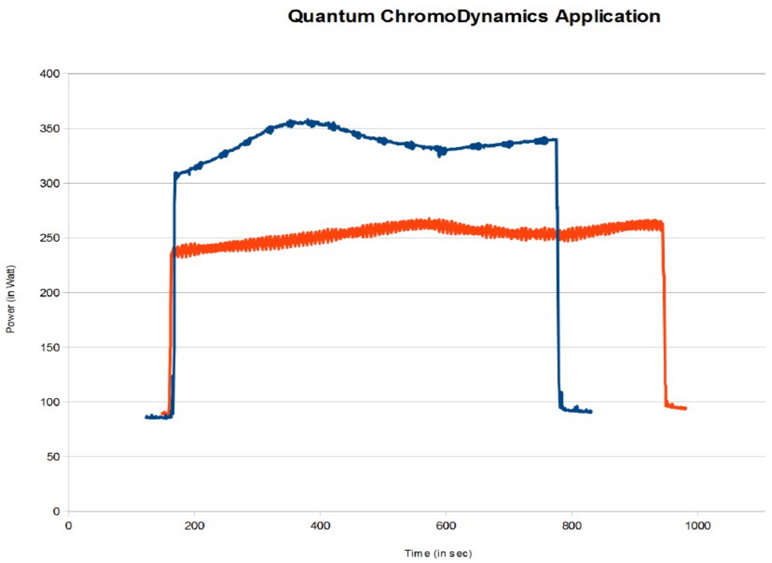
IBM System x iDataPlex Direct Water Cooled dx360 M4

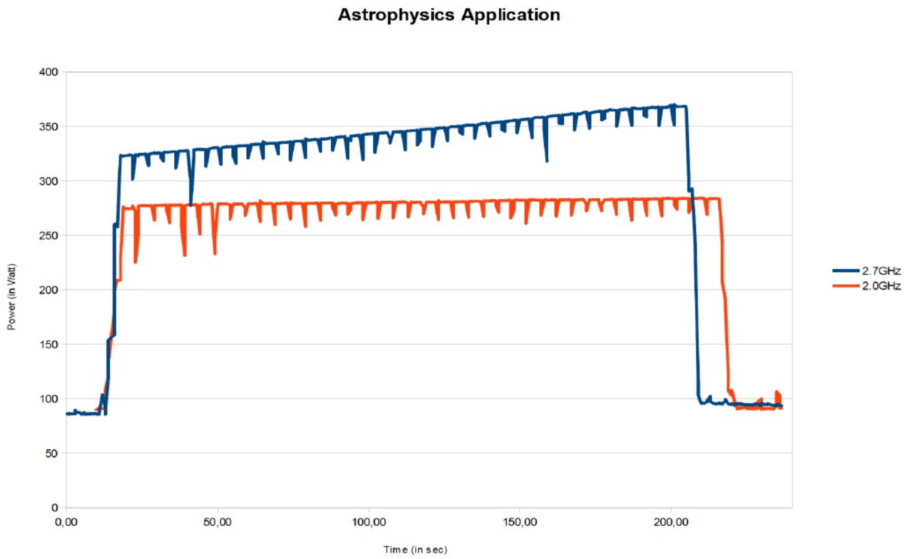
2x Intel SB-EP 2.7 GHz 130 W. 8x 4 GB.



Ingmar Meijer, 2012

What happens when you just lower the frequency ?





Δf=-26%
ΔPower=-26%
ΔTime=+26%
ΔEnergy=~0%

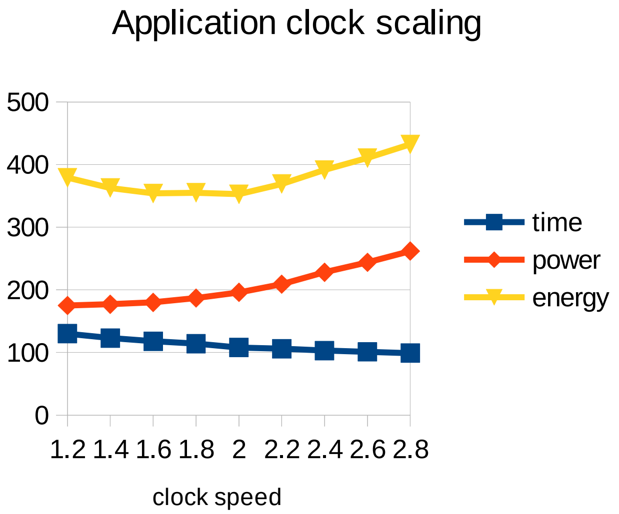
Δf=-26%
ΔPower=-17%
ΔTime=+5%
ΔEnergy=-12%

How do we find the performance/power trade-off ?

Monitor the application (hpm counters, power)
- Done transparently by the job scheduler

Build a performance and a power model
- Taking into account the processors/nodes
- And the application's characteristics

Introduce energy policies
- METS : Minimize Energy To Solution
- MTTS : Minimize Time To Solution
- MxTS : Minimize x to Solution



What can we do from a software perspective ?

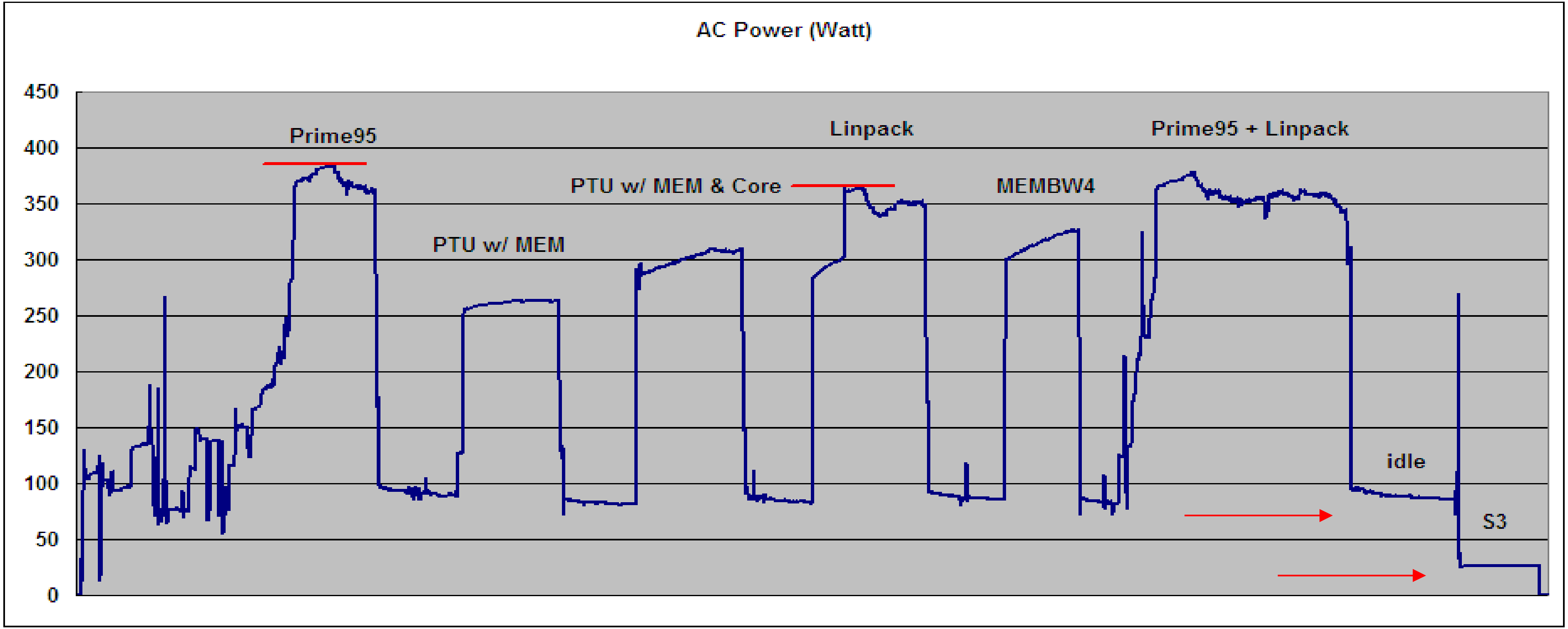
Reduce power of inactive nodes

- by C- or S-states

Reduce power of active nodes

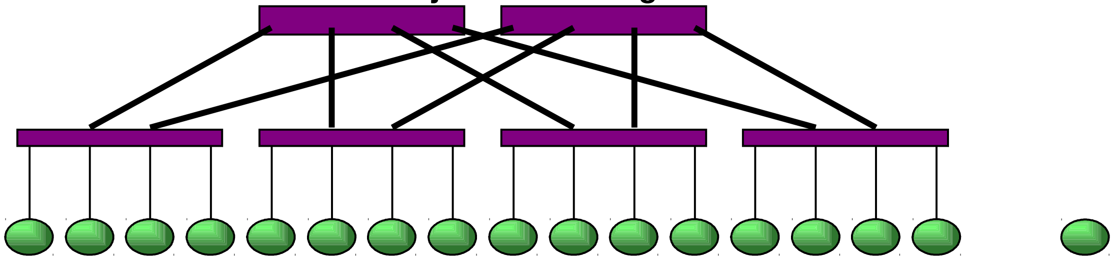
- by P-state / CPUfreq
- by memory throttling

Active and Idle power measurements on dx360m4

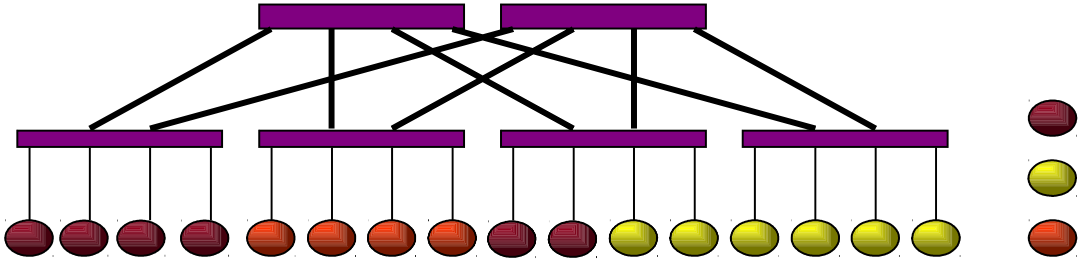


Energy Aware Scheduling (EAS)

Current job scheduling



Energy Aware job 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

LSF-EAS energy policies available

Minimize Energy To Solution

- subject to a maximum performance degradation of X%

Minimize Time To Solution

- frequency higher than default
- if default is not nominal
- subject to minimum performance improvement with clock speed

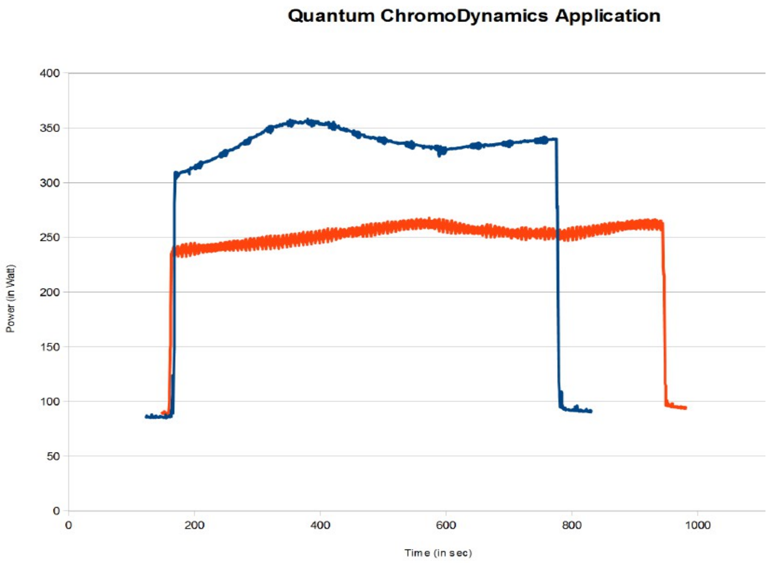
Set Frequency

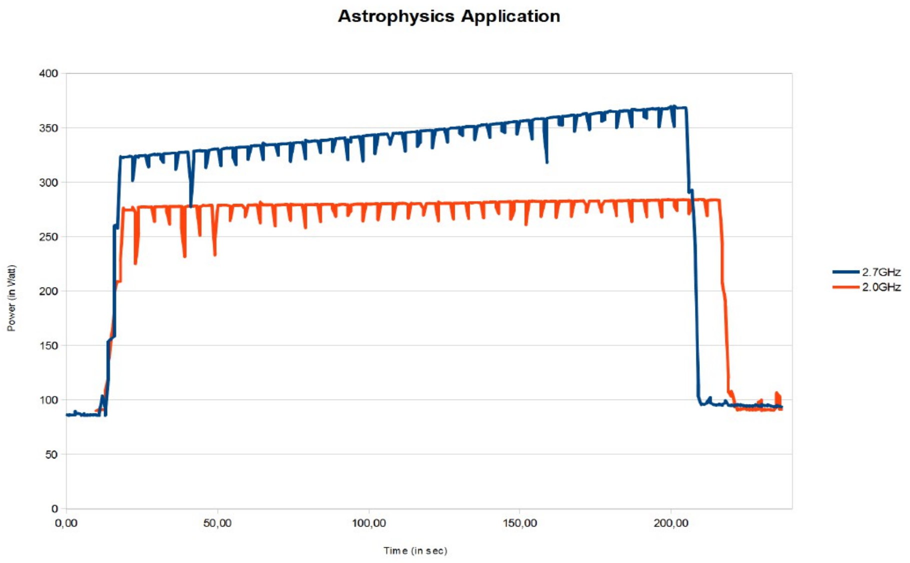
- (privileged)user specified

Site provided policy

- Sysadmin provides an executable to set frequency based on site local criteria

Example: what happens when you just change frequency





Δf=-26%
ΔPower=-26%
ΔTime=+26%
ΔEnergy=~0%

Δf=-26%
ΔPower=-17%
ΔTime=+5%
ΔEnergy=-12%

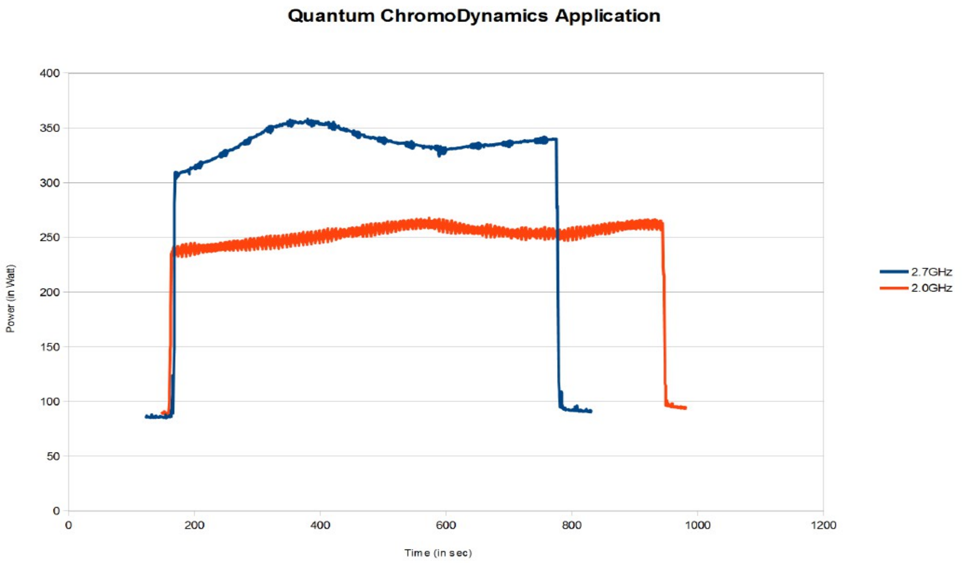
Example: how to submit a job first time

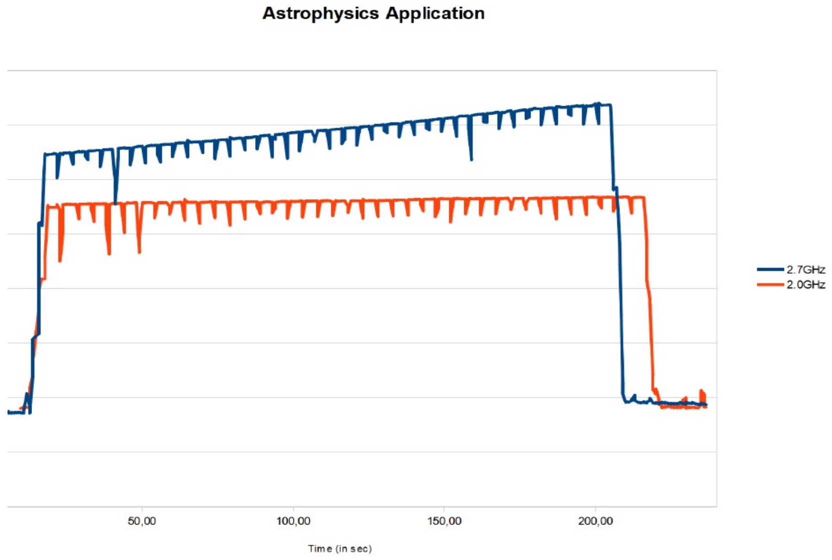
```
#!/bin/bash
# @ job_name = test
# @ account_no = 99999
# @ class = parallel
# @ job_type = MPICH
# @ network.MPI = sn_all,,US
# @ total_tasks = 128
# @ node = 8
# @ output = $(jobid)_output
# @ error  = $(jobid)_error
# @ initialdir = /bench/gpfs/fs1/users/fthomas/lleas/Astrophysics
# @ node_usage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





f= 2.6 GHz
ΔPower=-5%
ΔTime=+2%
ΔEnergy=-3%

f=2.0 GHz
ΔPower=-17%
ΔTime=+5%
ΔEnergy=-12%

Savings example

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 saved (cf examples)
Savings per year = 20.4 K€

Total savings: 45K€, ~10%

3 PFlops SuperMUC system at LRZ

Fastest Computer in Europe (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

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€
- **ISC'14 : “A Case Study of Energy Aware Scheduling on SuperMUC”, Axel Auweter.**

Thank you !

High Performance Computing
For a Smarter Planet