CAS2K13
Energy Aware Application Deployment

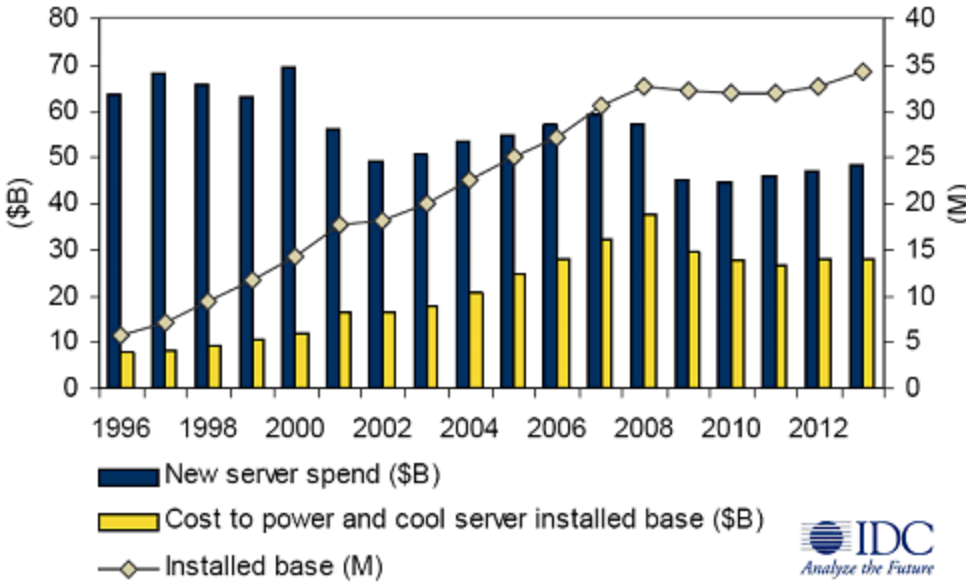
September 11, 2013

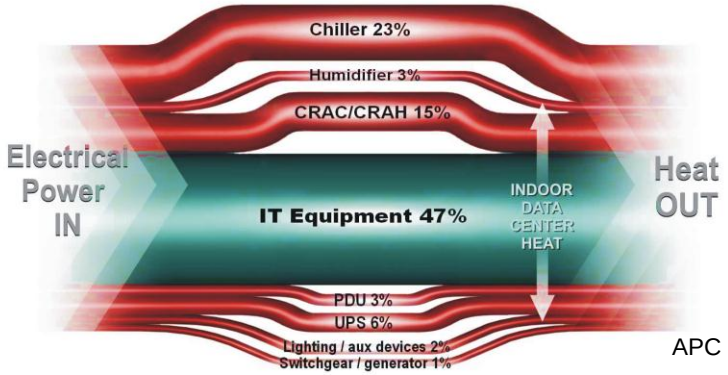
Don Grice

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





APC 2010

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

IBM Energy Aware Application Deployment

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?

Several ways to reduce power

- **Use specific processors**
  - Low voltage
  - Many cores at lower frequencies – SW implications
    - GPUs
    - MICs
    - FPGAs

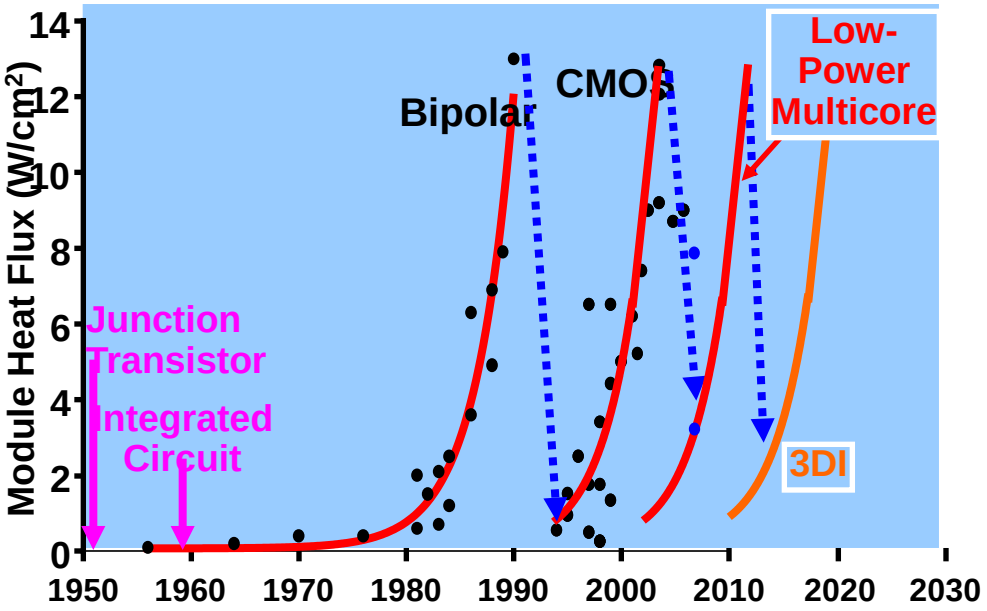
- **Use any processors with:**
  - New cooling – Warm water cooling and heat reuse –> PUE<1?
  - **New software**

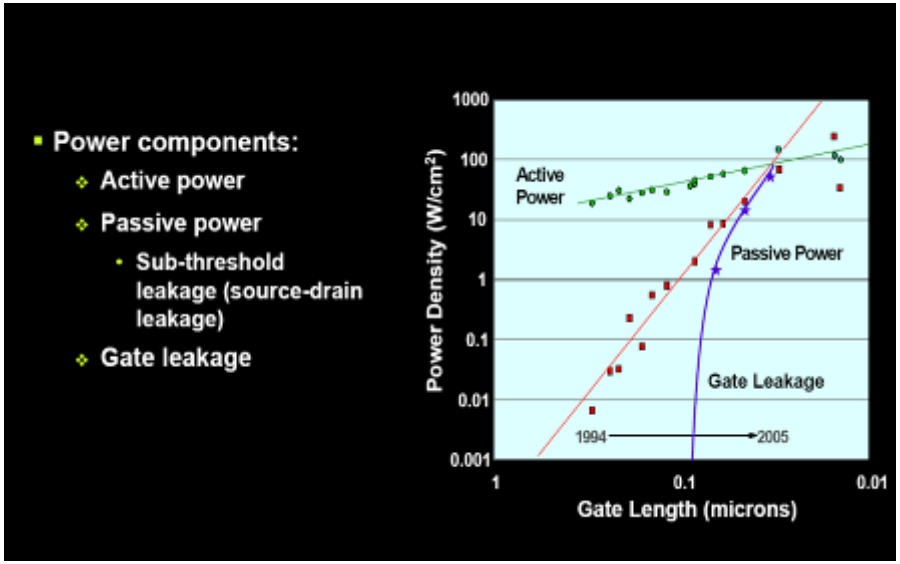
The Power Equation

- **Power = Capacitance * Voltage^2 * Frequency + DC Leakage Power**

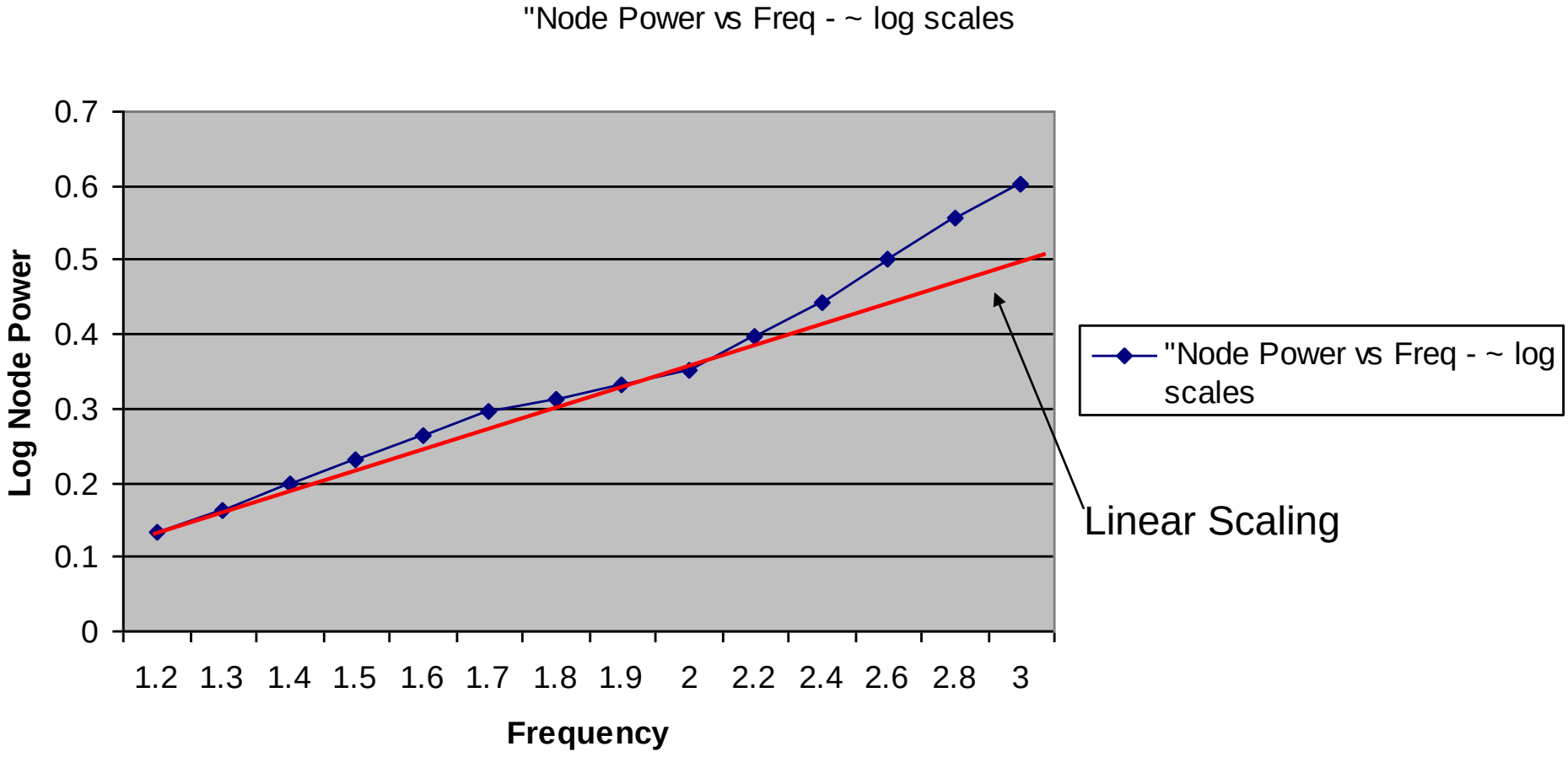
  - We have an active power problem
    - Frequency minimisation for active nodes

  - We have a passive power problem
    - Power minimisation for idle nodes

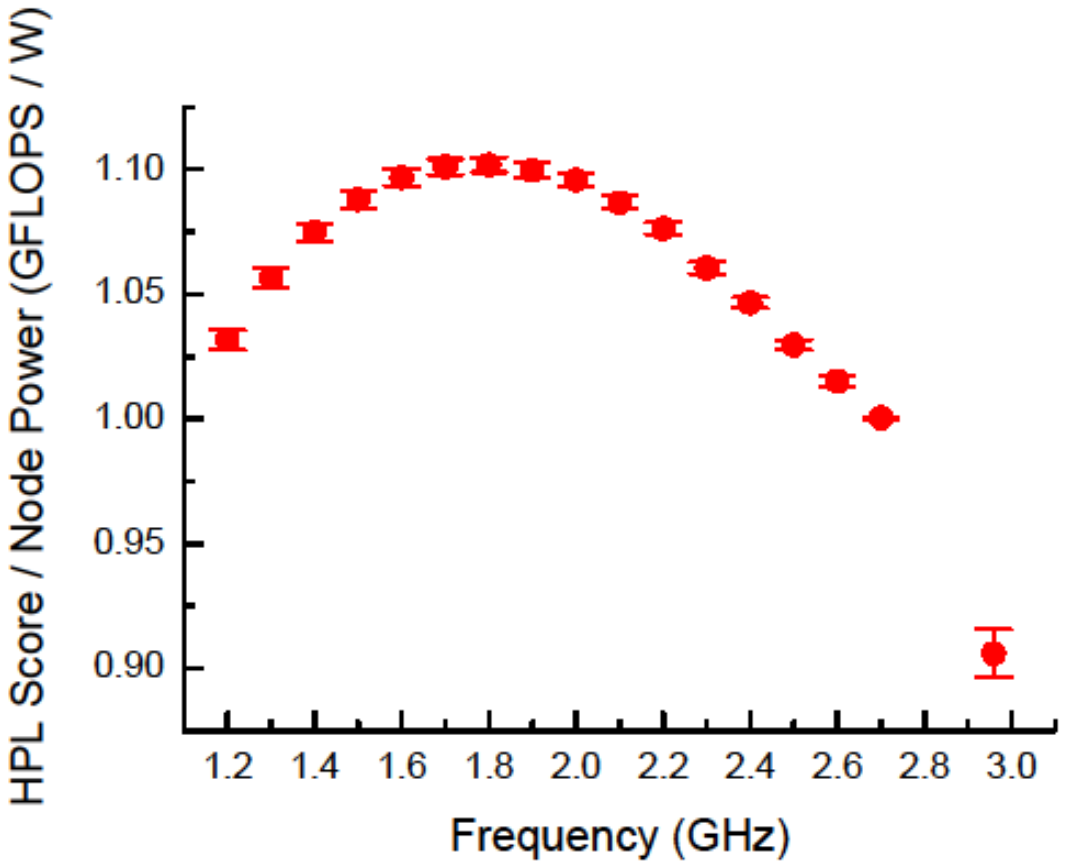




Non Linearity of Server Power at Higher Frequencies

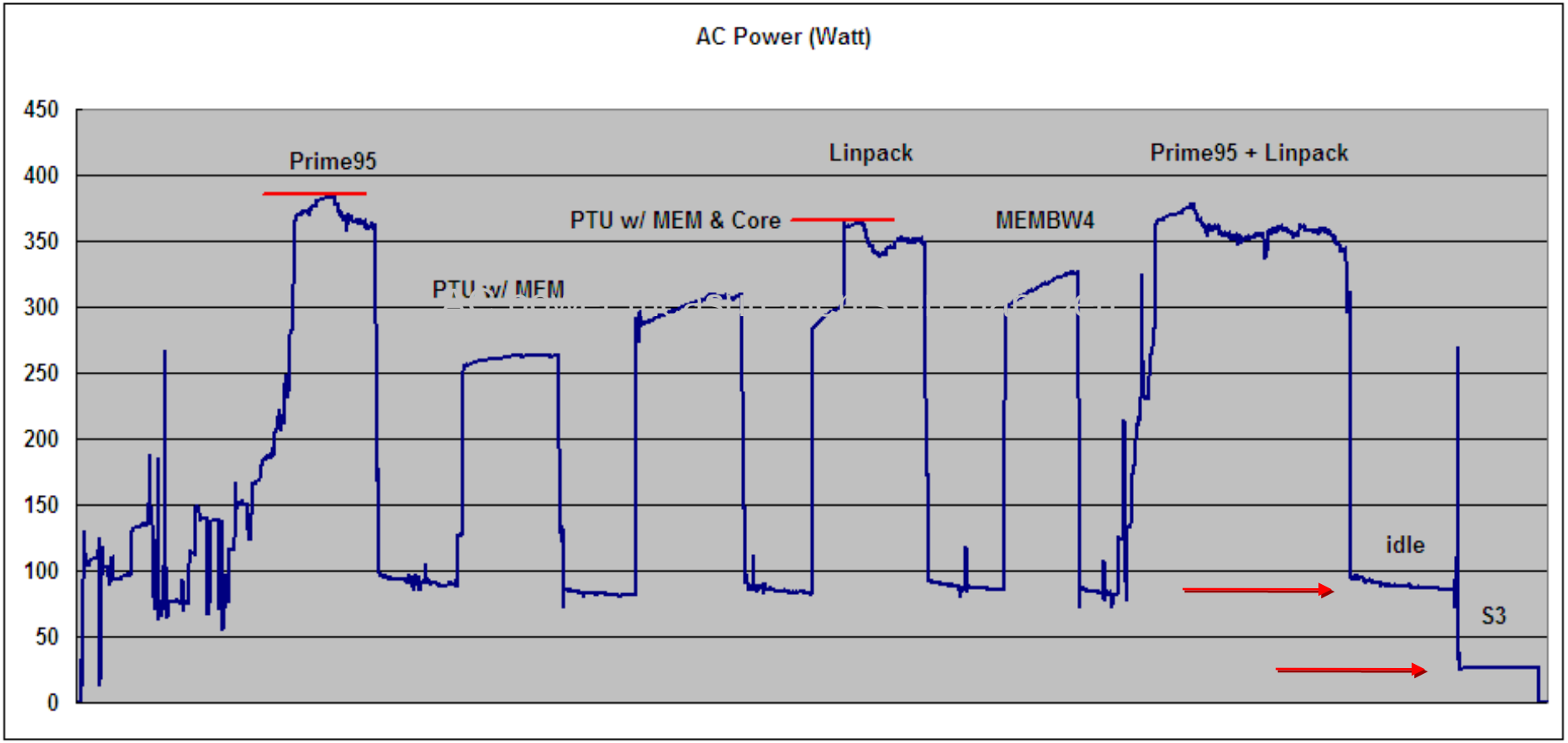


Power Efficiency of Linpack vs Frequency
Need to look at Total Energy as Well

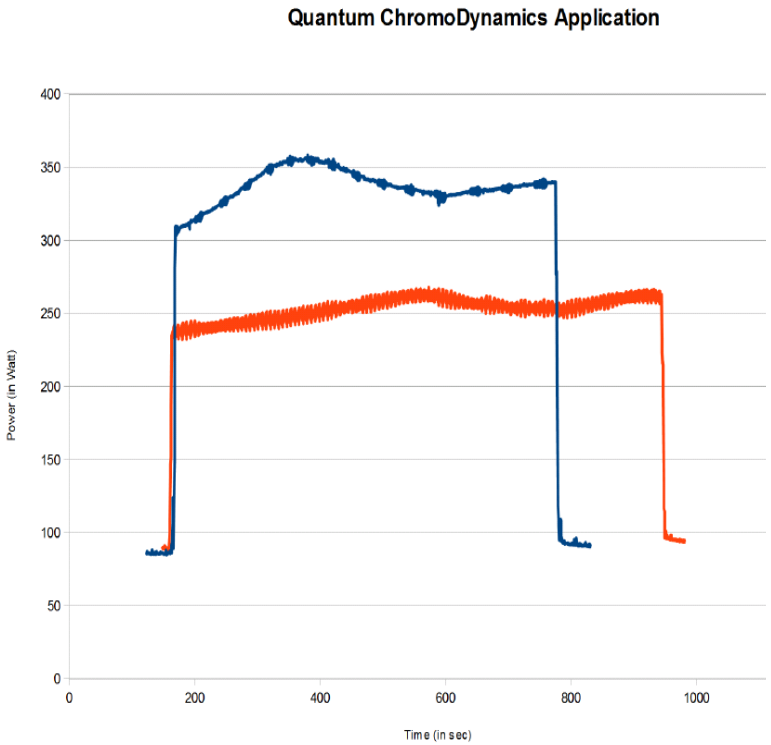


Ingmar Meijer, 2012

AC power measurements on dx360m4



Example: what happens when you just change frequency

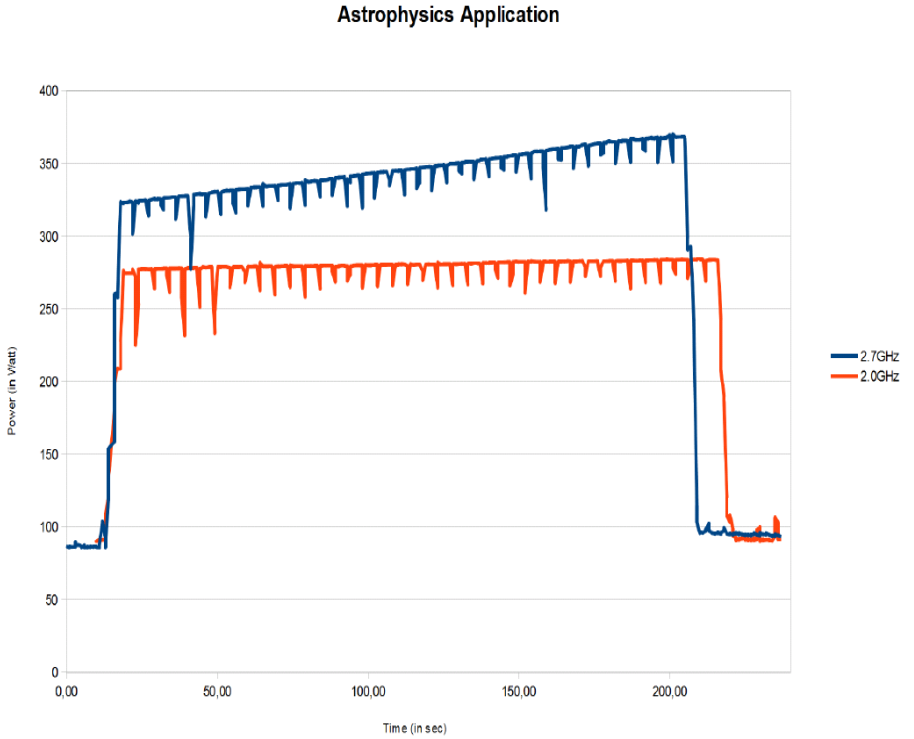


∆f=-26%

∆Power=-26%

∆Time=+26%

∆Energy=~0%



∆f=-26%

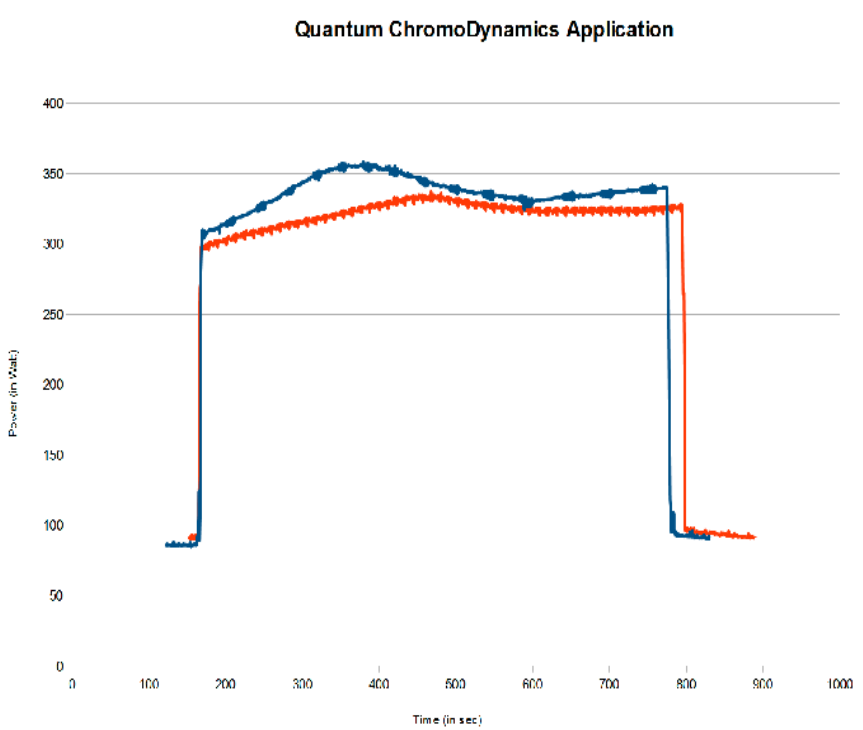
∆Power=-17%

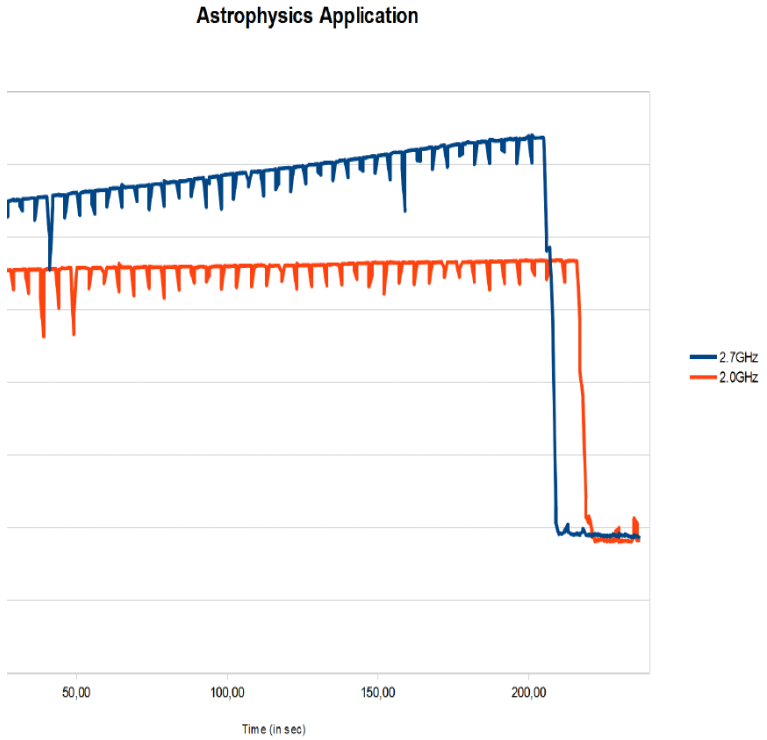
∆Time=+5%

∆Energy=-12%

IBM Energy Aware Application Deployment

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%

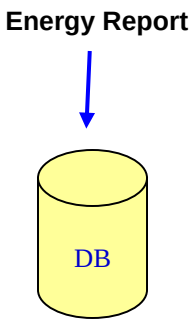
IBM Energy Aware Scheduling

Report

- Temperature, fan speed and power consumption per node
- power consumption, energy and performance per job
- total power consumption of the cluster

Optimize

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



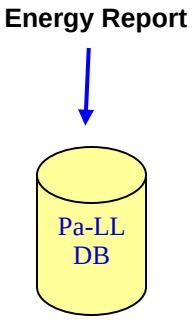
Features available to reduce and control power

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)

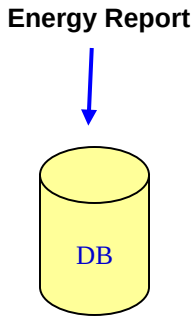
LL

- 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)



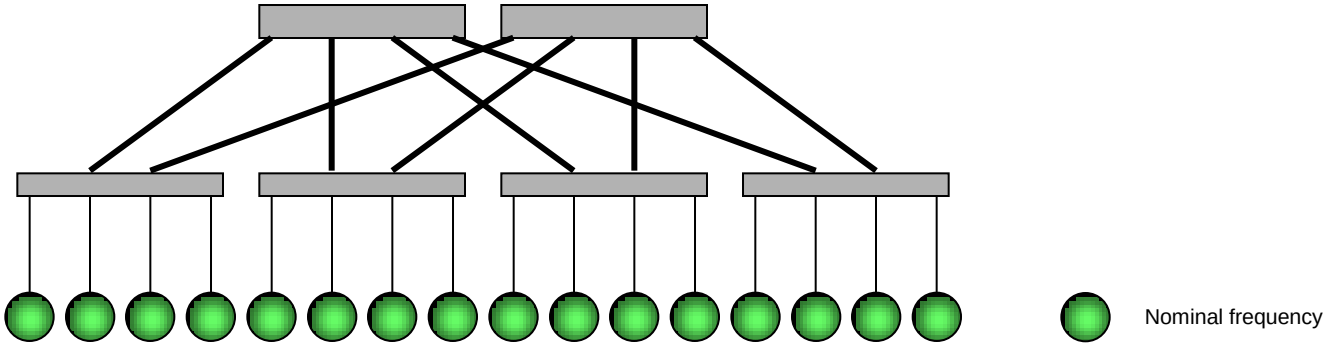
IBM software to monitor and reduce power

- **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

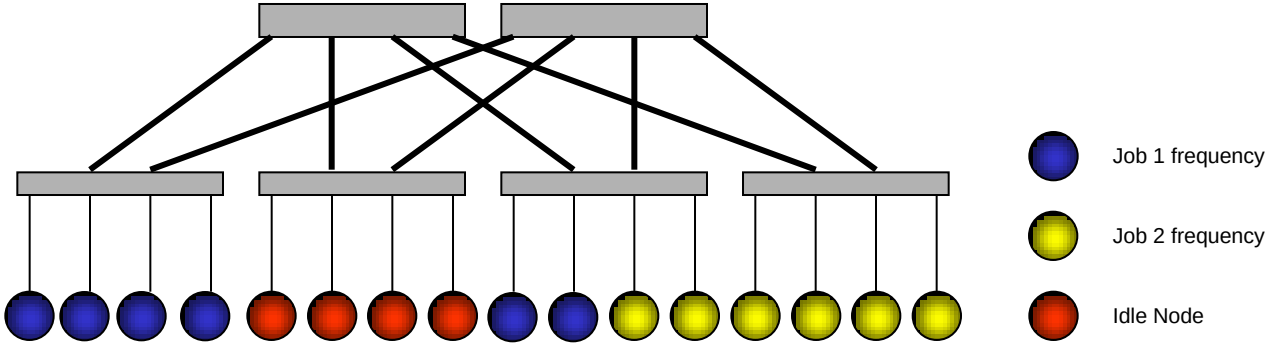


Energy Aware Scheduling

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

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 with max performance degradation = 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 hos 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

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

perf., power

Clock	CPI	Time	Power	Energy	PerfVa	PwrVa	EnyVar
2700	1,075	509	308	0,0435	0	0	0
2600	1,062	522	290	0,0420	-2,6%	5,8%	3,3%
2500	1,038	531	280	0,0413	-4,3%	8,8%	4,9%
2400	1,015	540	275	0,0413	-6,2%	10,6%	5,0%
2300	0,994	552	261	0,0400	-8,5%	15,3%	8,0%
2200	0,972	565	255	0,0399	-10,9%	17,2%	8,1%
2000	0,932	596	237	0,0393	-17,1%	22,8%	9,6%
1900	0,908	611	228	0,0386	-20,0%	25,9%	11,1%
1800	0,894	635	220	0,0388	-24,7%	28,4%	10,8%
1700	0,877	659	212	0,0388	-29,6%	31,1%	10,7%
1600	0,848	677	207	0,0390	-33,0%	32,6%	10,4%
1500	0,831	708	199	0,0392	-39,2%	35,2%	9,8%
1400	0,821	750	188	0,0391	-47,3%	38,9%	10,0%
1300	0,807	794	179	0,0394	-55,9%	41,9%	9,4%
1200	0,797	849	170	0,0400	-66,7%	44,8%	7,9%

Clock	CPI	Time	Power	Energy	PerfVa	PwrVar	EnyVar
2700	0,661	304	290	0,0244	0	0	0
2600	0,651	311	273	0,0236	-2,2%	5,7%	3,6%
2500	0,645	320	263	0,0234	-5,3%	9,2%	4,4%
2400	0,634	328	257	0,0235	-7,9%	11,1%	4,1%
2300	0,626	338	244	0,0229	-11,1%	15,6%	6,2%
2200	0,620	350	237	0,0231	-15,2%	18,1%	5,6%
2000	0,598	372	222	0,0229	-22,2%	23,3%	6,3%
1900	0,593	387	213	0,0229	-27,4%	26,4%	6,2%
1800	0,584	403	206	0,0230	-32,5%	29,0%	5,9%
1700	0,581	424	199	0,0234	-39,6%	31,4%	4,2%
1600	0,575	446	194	0,0240	-46,7%	33,2%	1,9%
1500	0,571	473	186	0,0244	-55,5%	35,8%	0,1%
1400	0,566	502	175	0,0244	-65,1%	39,5%	0,1%
1300	0,563	538	167	0,0249	-76,9%	42,3%	-2,0%
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

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**

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%**

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€

Next Steps

- Data Aware Scheduling
  - Don’t Power Up Compute Nodes until the Data is Available
  - Include Data Flow and pre-staging in the Job Definition
- Compute Aware Powering
  - Only Power Circuitry needed for Computation
  - Not all cores, not all accelerators, etc
  - Latency from idle to operational will be a HW design characteristic
- Power Fluctuation Control
  - Analogous Issue to Wind and Solar production leveling
  - As machine power becomes more significant to Utility Suppliers fluctuations in Machine power will be an issue
  - Job Start/Stop, Job Steps?
  - Spending Energy to keep load level is not the right answer

THANK YOU

Any Questions?

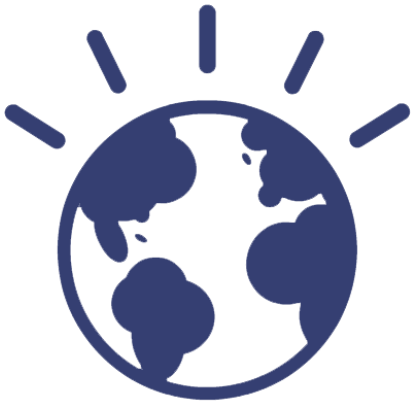
High Performance Computing

For a Smarter Planet

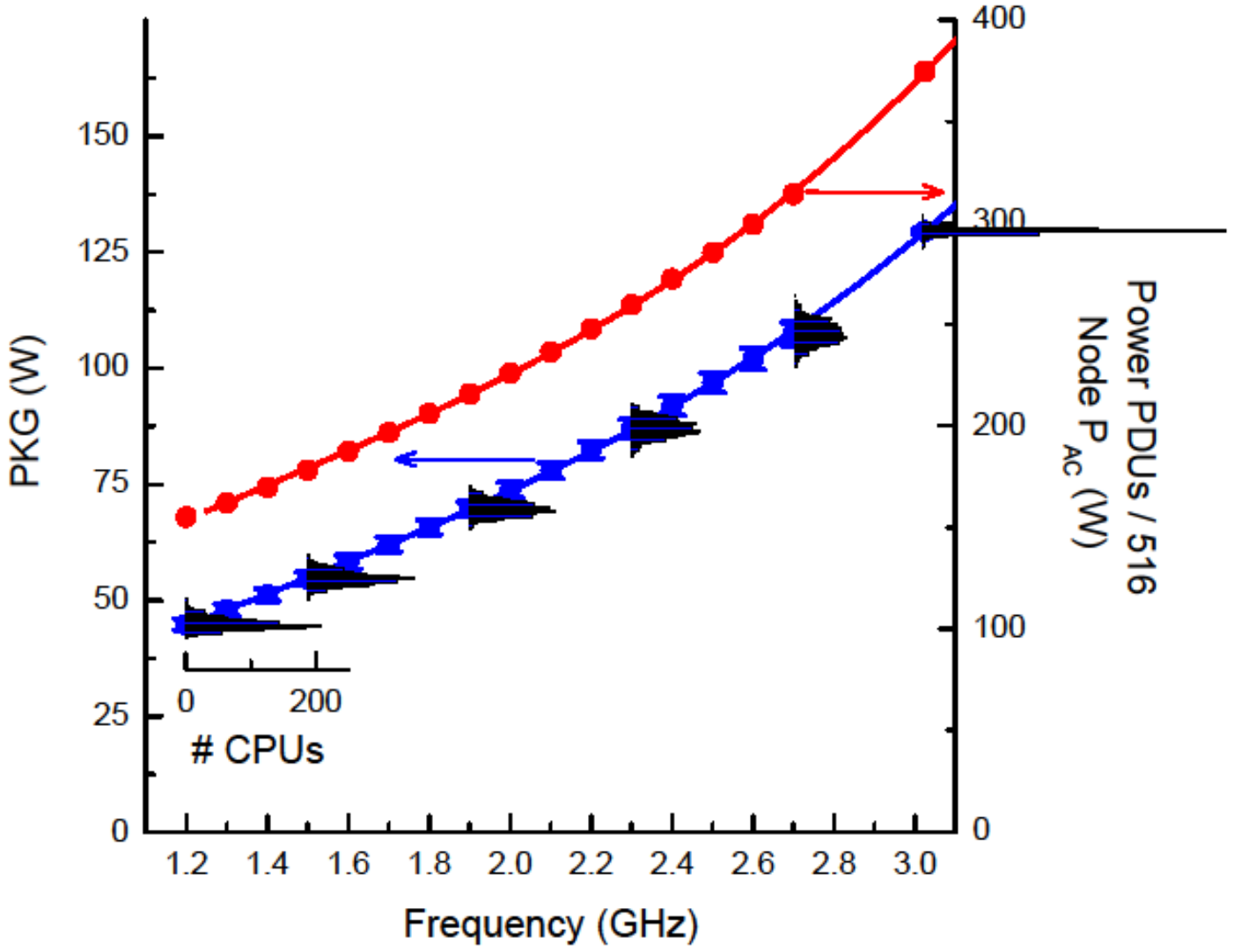
Back UP

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**



Power Consumed by the Server vs Frequency



Ingmar Meijer, 2012

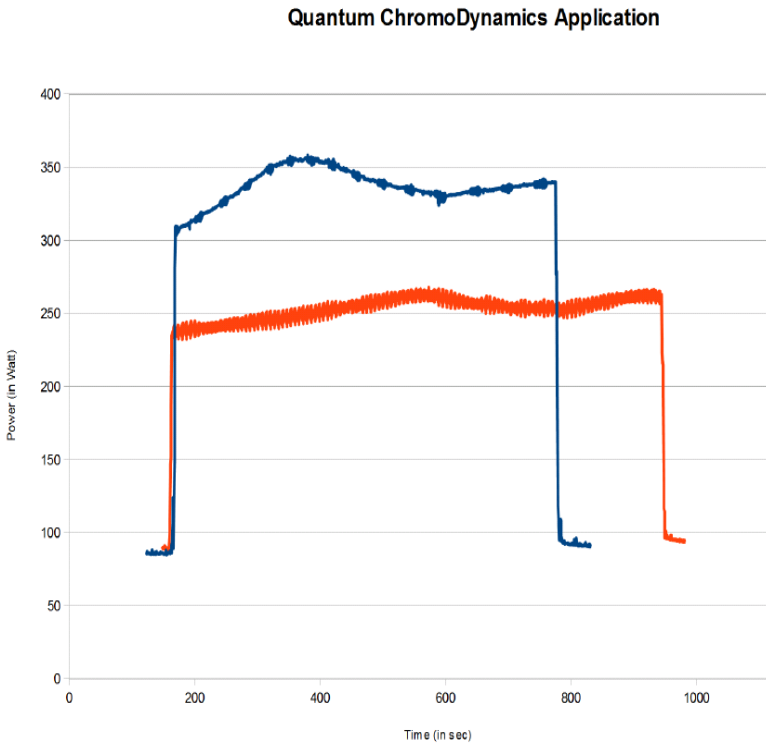
Example: how to submit a job first time

```
#!/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 when you just change frequency

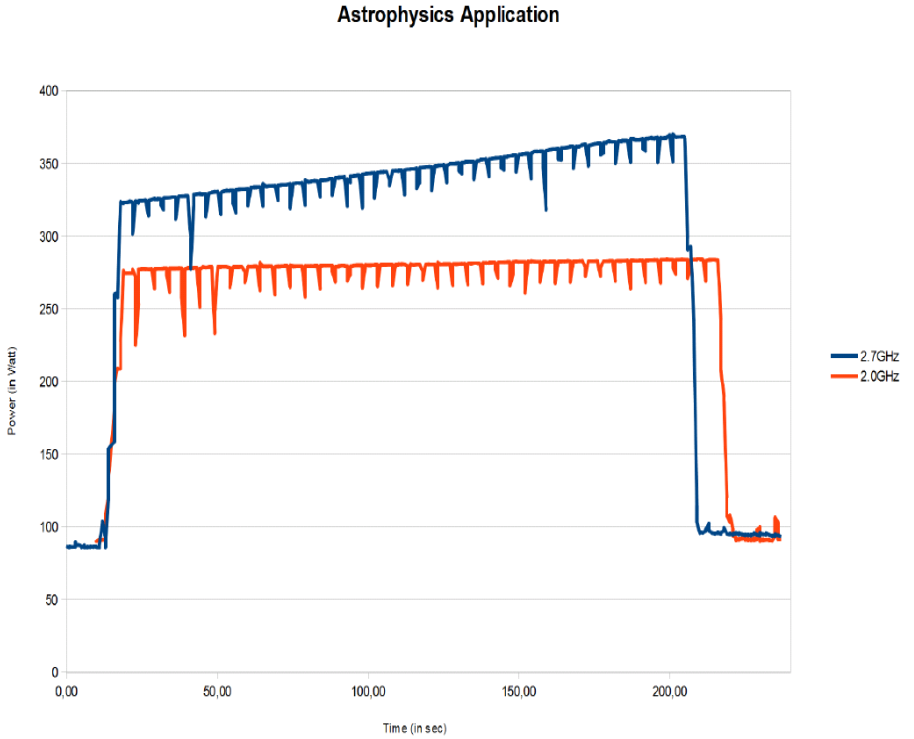


∆f=-26%

∆Power=-26%

∆Time=+26%

∆Energy=~0%



∆f=-26%

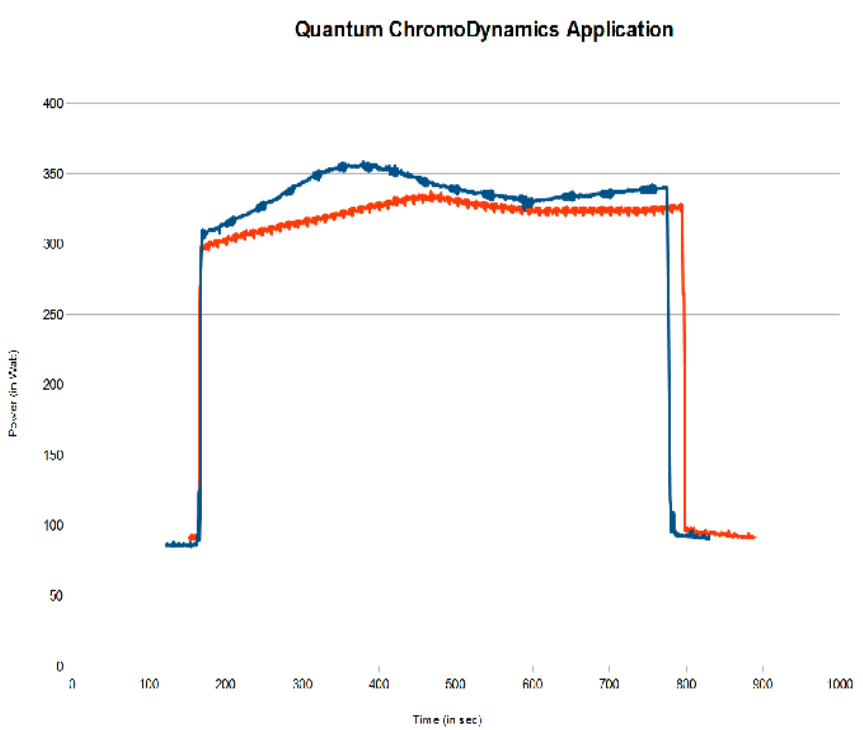
∆Power=-17%

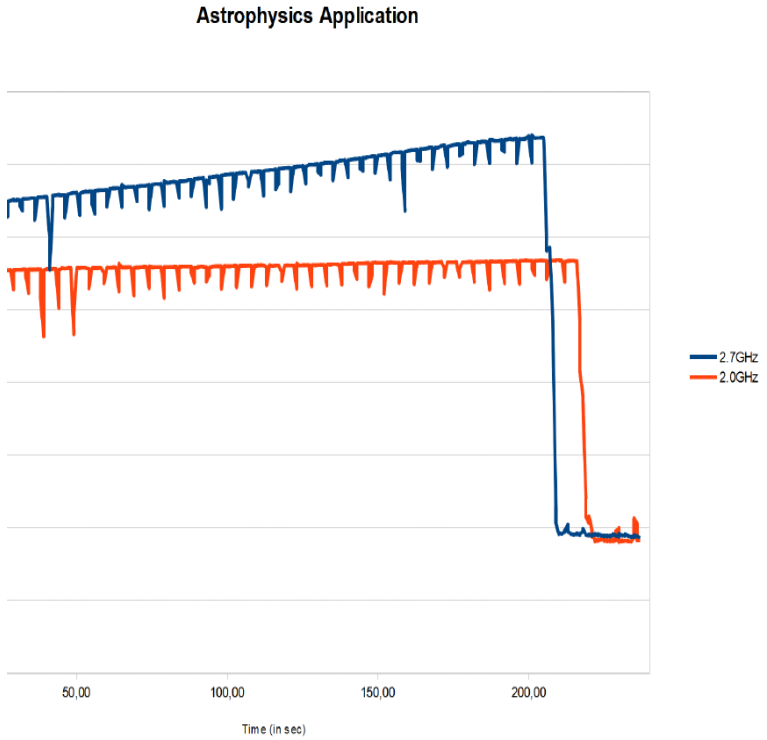
∆Time=+5%

∆Energy=-12%

IBM Energy Aware Application Deployment

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%

AC power measurements on dx360m4