Advanced School on PARALLEL COMPUTING

Energy Benchmarks

Andrew Emerson,
a.emerson@cineca.it

CINECA

What are application benchmarks (in HPC) and why do them ?

- *Benchmarks compare the performance of an application code according to different parameters such as no. of cores, type of architecture, program version and input, ...*
- *Hardware vendors, computer centres and other organisations (e.g. PRACE) often provide "benchmark suites" which may be used to provide:*
    1. *a resource of application codes and datasets for hardware procurement.*
    2. *data to help users decide during project preparation which system to choose and how much time to ask for.*
    3. *data for "currency conversion" of CPU hours between different systems (e.g. PRACE Tier-1).*
- *For a user starting an HPC project, should be standard practice to benchmark application code with the required input on the target system* **before** *starting production runs.*

PRACE Unified European Application Benchmark Suite (UEABS)

Scientific Area	Application code
Particle Physics	*QCD*
Classical MD	*NAMD, GROMACS*
Quantum MD	*Quantum Espresso, CP2K, GPAW*
CFD	*Code_Saturne, ALYA*
Earth Science	*NEMO, SPECFEM3D*
Plasma Physics	*GENE*
Astrophysics	*GADGET*

[1] *Selection of a Unified European Application Benchmark Suite, J. Mark Bull and Andrew Emerson,* *http://www.prace-ri.eu/IMG/pdf/Selection_of_a_Unified_European_Application_Benchmark_Suite.pdf*
[2] *Unified European Applications Benchmark Suite, J. Mark Bull et al, http://www.prace-ri.eu/ueabs*

Advanced School on PARALLEL COMPUTING

PRACE UEABS

- *Each code was benchmarked for 3 different datasets ("small", "medium" and "large") on PRACE Tier-0 and Tier-1 systems;*
- *First version of PRACE UEABS concentrated only on "standard" CPU cores (i.e. no GPUs or accelerators).*

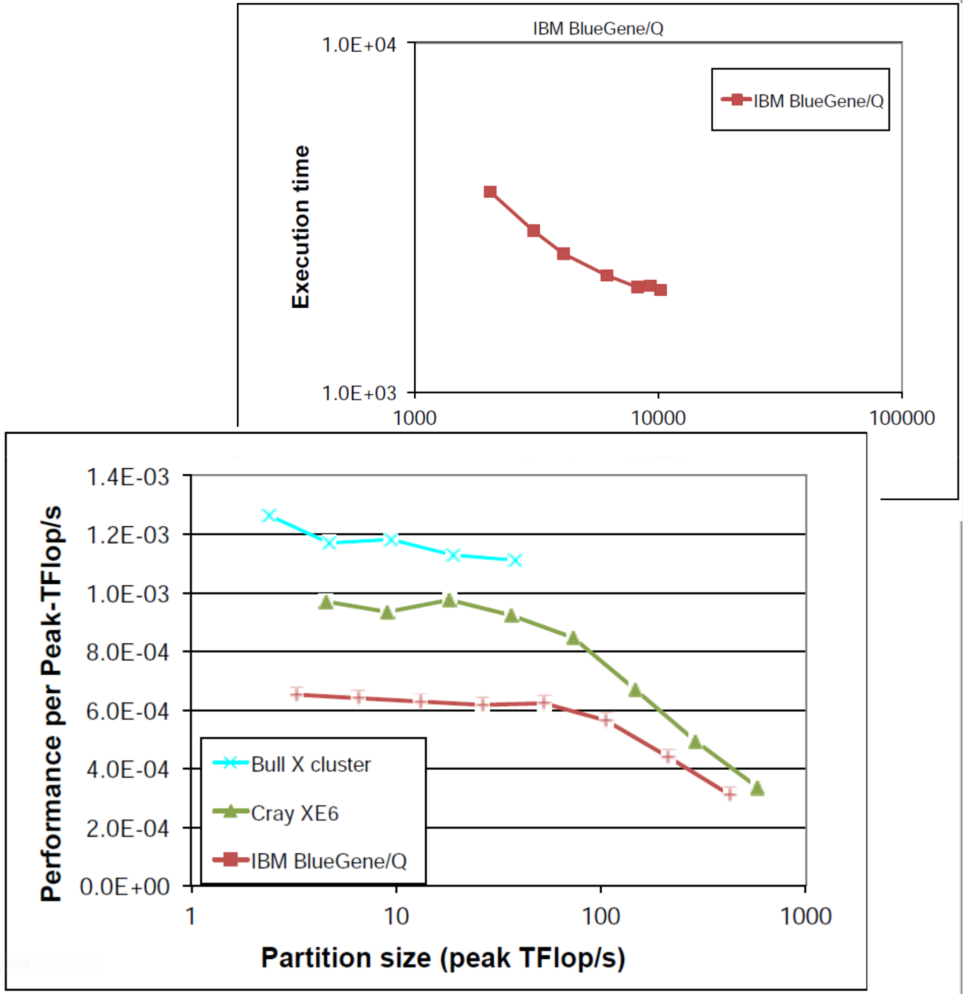


Figure 38 Performance per Peak-TFlop/s of QCD for Kernel E

CINECA

PRACE UEABS – how to measure "performance"

- *"Performance" can be domain-specific so for the UEABS two domain-independent metrics were used;*
  1. *execution time (i.e. time in seconds or 1/time to complete the run).*
  2. *performance (1/time) per Peak-TFlop/s as function of the partition size in Peak-Tflops.*
- *This second metric allows codes to be compared between different platforms.*

PRACE UEABS – QCD

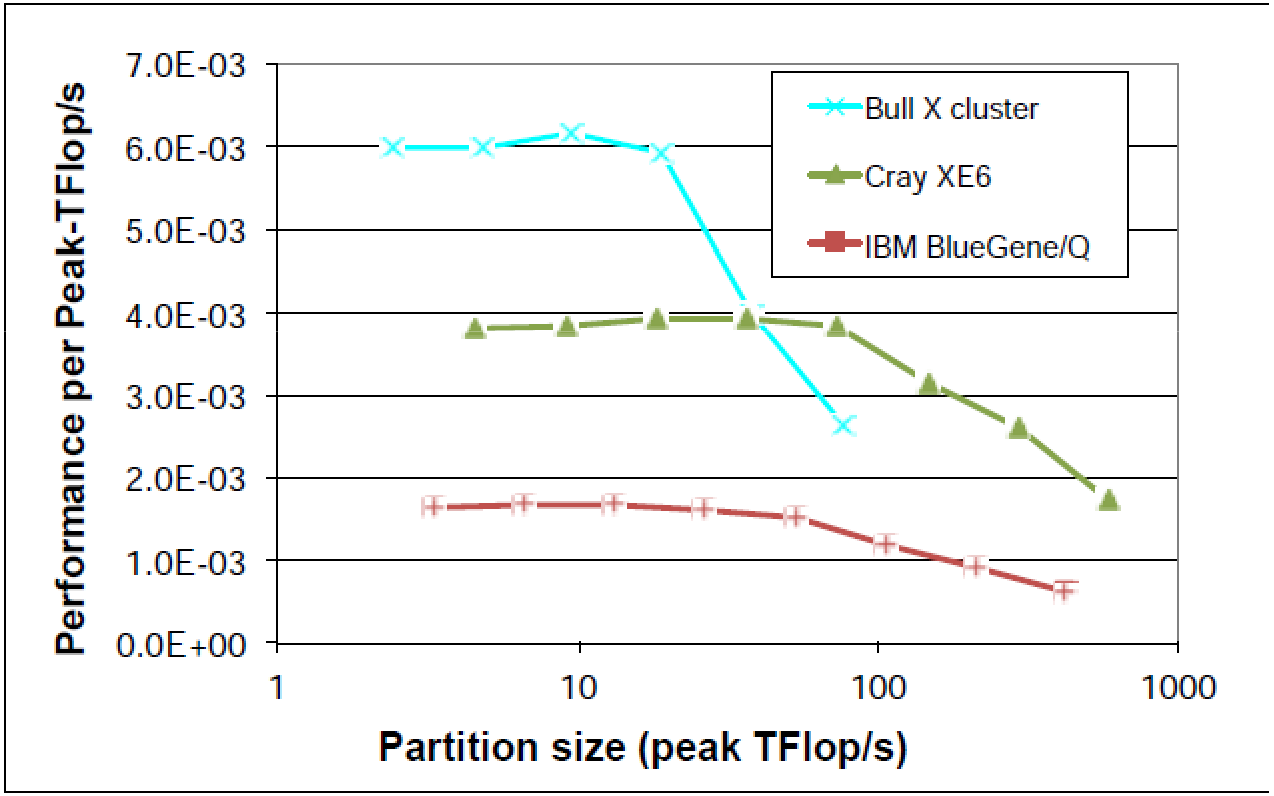


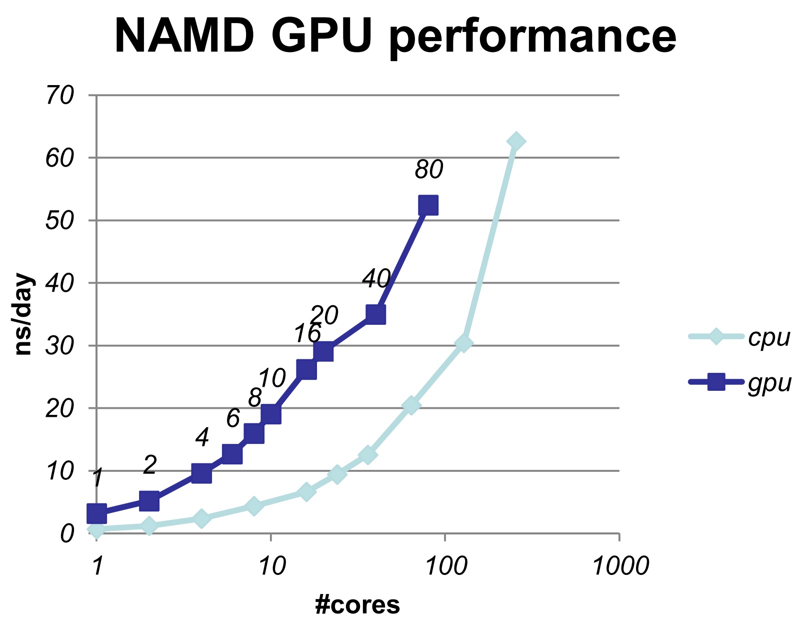
Figure 32 Performance per Peak-TFlop/s of QCD for Kernel B

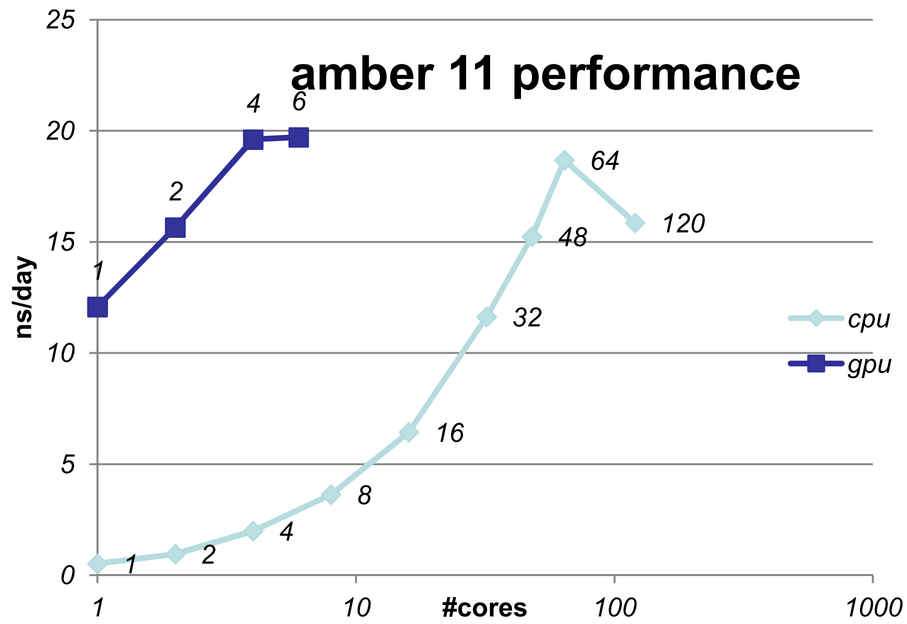
horizontal line indicates ideal scaling

And accelerators ?

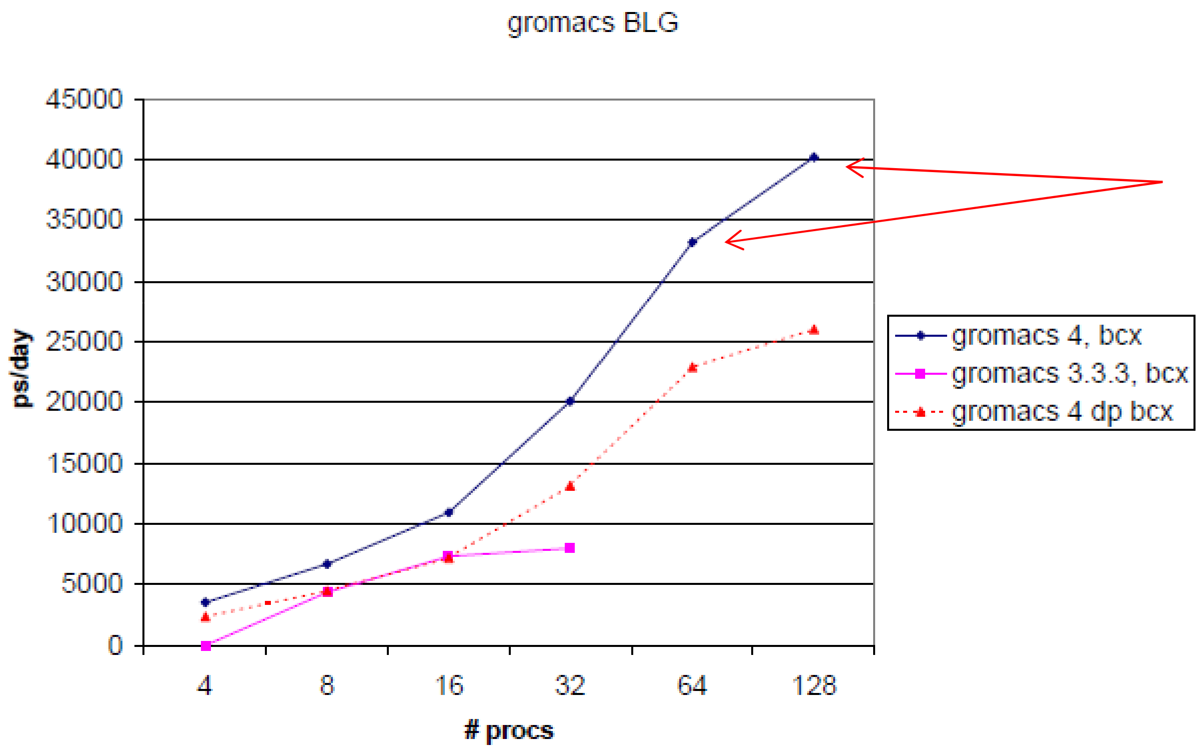
- *GPU-enabled versions of some codes can bring enormous speedups when compared to "traditional" cores.*
- *Thus, even if in cases where the overall maximum performance is not exceeded, by using few cores GPU-enabled codes can be more "cost effective".*
- *Same argument used for other accelerators such as Intel's Xeon PHI (MIC) technology.*

Examples: Amber and NAMD (Molecular Dynamics)





What about "cost" ?



Most high performing option not always the most cost efficient option!

So what about...

- *An "energy metric" for energy cost ?*
- *Since we are told that energy efficiency and power consumption are important it makes sense to measure this as well.*

Crowned the greenest supercomputer, the Tsubame-KFC system at the Tokyo Institute of Technology, hit a record 4.5 gigaflops per watt. That's about 25 percent more efficient than the list's number-two, Cambridge University's Wilkes, at 3.6 gigaflops per watt. In third place was the system at Japan's Center for Computational Sciences, at the University of Tsukuba, at 3.5 gigaflops per watt. .
Green TOP500

Green Top 500

The Green500 List

Listed below are the November 2013 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

* Performance data obtained from publicly available sources including TOP500

Estimating Energy consumption

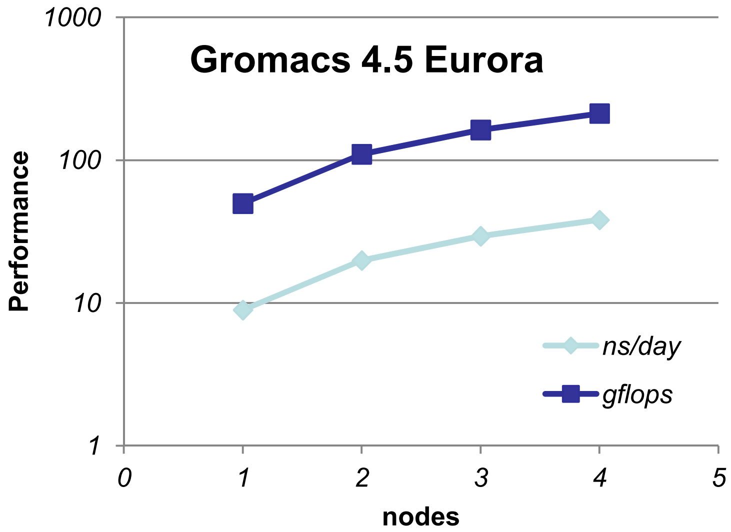
- *If your application can output Gflops can use that estimate energy needed to run your program*
- *Case study Gromacs (Molecular Dynamics). Run identical runs as a function of #nodes*

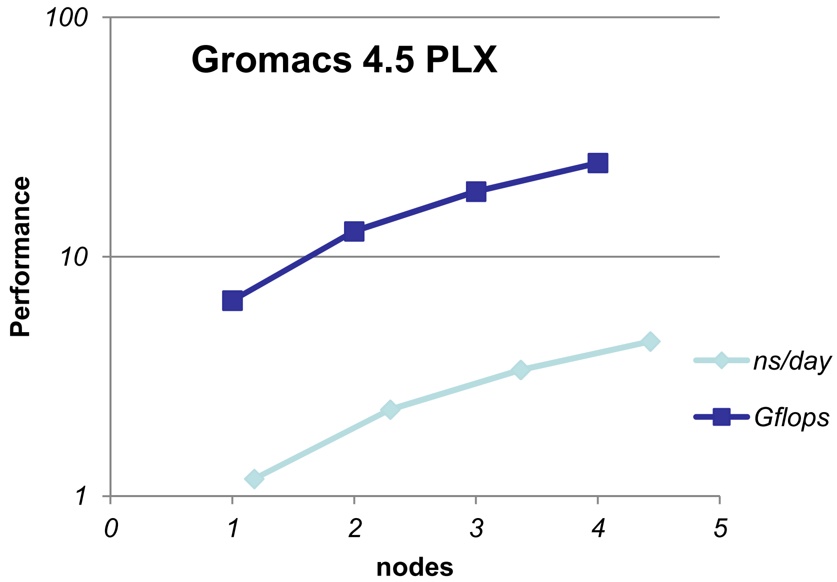
```
Parallel run - timing based on wallclock.

               NODE (s)   Real (s)      (%)
       Time:     45.318     45.318    100.0
               (Mnbf/s)   (GFlops)   (ns/day)  (hour/ns)
Performance:   2751.193    212.351     38.135      0.629
Finished mdrun on node 0 Wed Feb 12 22:11:36 2014
```

Computational chemists use ns/day as performance – directly indicates how much "scientific work" can be done.

Estimating Energy consumption

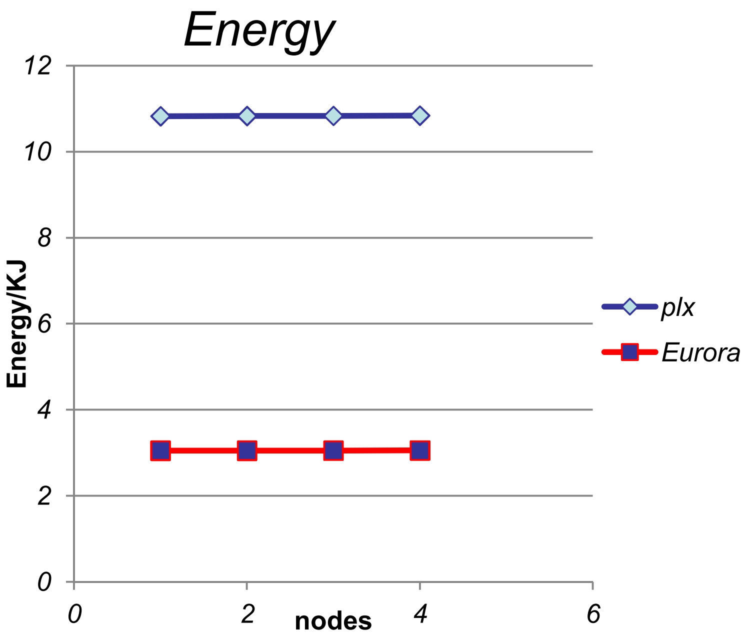
- *Hopefully trend in Gflops should mirror trend of the usual performance metric for your application.*

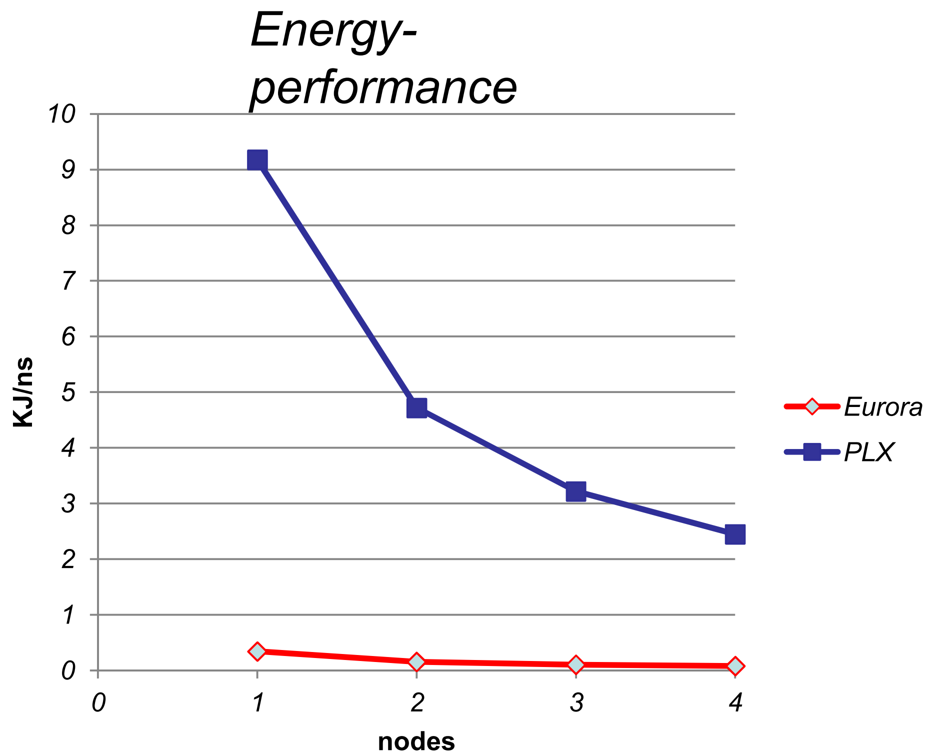




Estimating energy consumption

- *Sustained peak of Eurora ~3.15 Gflops/watt, PLX ~ 0.91 Gflops/watt*
- *Knowing wall time of job + Gflops can calculate total energy of job.*





Estimating energy consumption

- *Very crude method of estimating energy – assumes consuming flops at published power rate (peak, sustained, etc.). Most applications much less efficient than peak (e.g. 20%).*
- *No indications of energy due to network, cooling, disks etc.*
- *Not all applications provide Gflops output.*
- *Need to actually* ***measure*** *the energy consumption by hardware during job run.*

Power Data Aggregation Monitor (PowerDAM)

- *Developed by Leibniz Supercomputing Centre.*
- *PowerDAM monitors both physical sensors as well as “virtual sensors” and provide visualization for factors such as power draw, utilization rate, and average CPU temperatures.*
- *Can be used to measure energy consumed during a single batch job.*

powerDAM commands

Measures directly the energy in kWh (=3600 kJ).
Current implementation still very experimental.

```
ets --system=Eurora --job=429942.node129

EtS is: 0.173056 kWh
Computation:    99 %
Networking:     0 %
Cooling:        0 %
Infrastructure: 0 %
```

Gromacs 4.6 energy consumption via ets

PBS Job id	nodes	Clock freq (GHz)	#gpus	Walltime (s)	Energy (kWh)	Perf (ns/day)	Perf-Energy (ns/kJ)
429942	1	2	0	1113	0.17306	10.9	69.54724
430337	2	2	0	648	0.29583	18.6	62.87395
430370	1	3	0	711	0.50593	17.00	33.60182
431090	1	3	2	389	0.42944	31.10	72.42023

Observations (based on v. limited data):

1. *Previous (Gflop) estimates clearly inaccurate.*
2. *High-clock frequency relatively inefficient.*
3. *In this example use of GPUs really is most efficient, but for 1 node not by that much cf. 2 GHz proc.*

Summary

- *Benchmarks are essential during project preparation and production for estimating resource requirements.*
- *Until recently 1/walltime or field-related metric (e.g. ns/day) used exclusively for assessing "performance". Now focus switching to "energy performance". Need compromise between application performance and cost.*
- *Rough guide can be obtained knowing app performance in Gflops and machine performance, but likely to be very inaccurate.*
- *Need instead hardware monitoring. With accurate energy data/job, users can tune application parameters to balance their energy requirements or write low-energy applications. Future schedulers could prioritise low-energy jobs.*
- *.*