Benchmarking, Power and Thermals in the Data Center

8 May 2007

Andreas Hirstius

Background

- Performance and power consumption of a data center are the most important properties
- We have power distribution limit of (only) 2.5MW
- The metrics we use to describe our machines are
  - Performance per Watt
  - Performance per CHF

- Efficiency is (almost) everything...

Benchmarking

- Our software scales with SPECInt2000!
- The published results (spec.org) are highly optimized
- The physicists are very conservative when it comes to optimizations by the compiler
  - ➔ accuracy is everything!
  - *gcc* is used with the "*-O2 -fPIC -pthread*" options

- When compiling SPECInt the resulting code is up to 50% slower than the best possible result
- SPECInt2000 with a specially prepared configuration file is used for our tenders

Benchmarking II

- The frameworks used at CERN have their own set of benchmarks
  - ROOT for data analysis
  - GEANT4 for simulation
  - CLHEP as collection of frequently used routines
- Those frameworks are huge (millions of lines of code)
- Difficult to handle for optimization and debugging purposes
  - openlab extracted “snippets” - small, self contained pieces
    - very effective for testing new compiler versions
    - they make communication with the compiler writers easier

Benchmarking III

- Modern processors have powerful performance monitoring capabilities
  - Itanium2 was the first CPU with a usable performance monitoring unit, now all CPUs have one
  - we work closely with the main developer of the *perfmon* tool
  - perfect tool set to get a detailed look at the applications (almost) without disturbing them

Our machines

Hardware:

- Dual Nocona 2.8GHz (32-bit) ➞ 2 cores
- Dual Dempsey 3.2GHz (64-bit) ➞ 4 cores (beta system)
- Dual Woodcrest 2.66GHz (32/64-bit) ➞ 4 cores (beta system)
- Dual Clovertown 2.4GHz (64 bit) ➞ 8 cores (beta system)
- Itanium2 1.3GHz, 1.5GHz and 1.6GHz
- Montecito 1.6GHz (beta system)

Benchmark:

- SPEC2000 version 1.3
  - SPECInt with icc and gcc
  - SPECbase measurements
    - 1, 2, 4, n/2, n, 1.5*n parallel jobs, “manually” started
    - Benchmarks run independently – out-of-sync

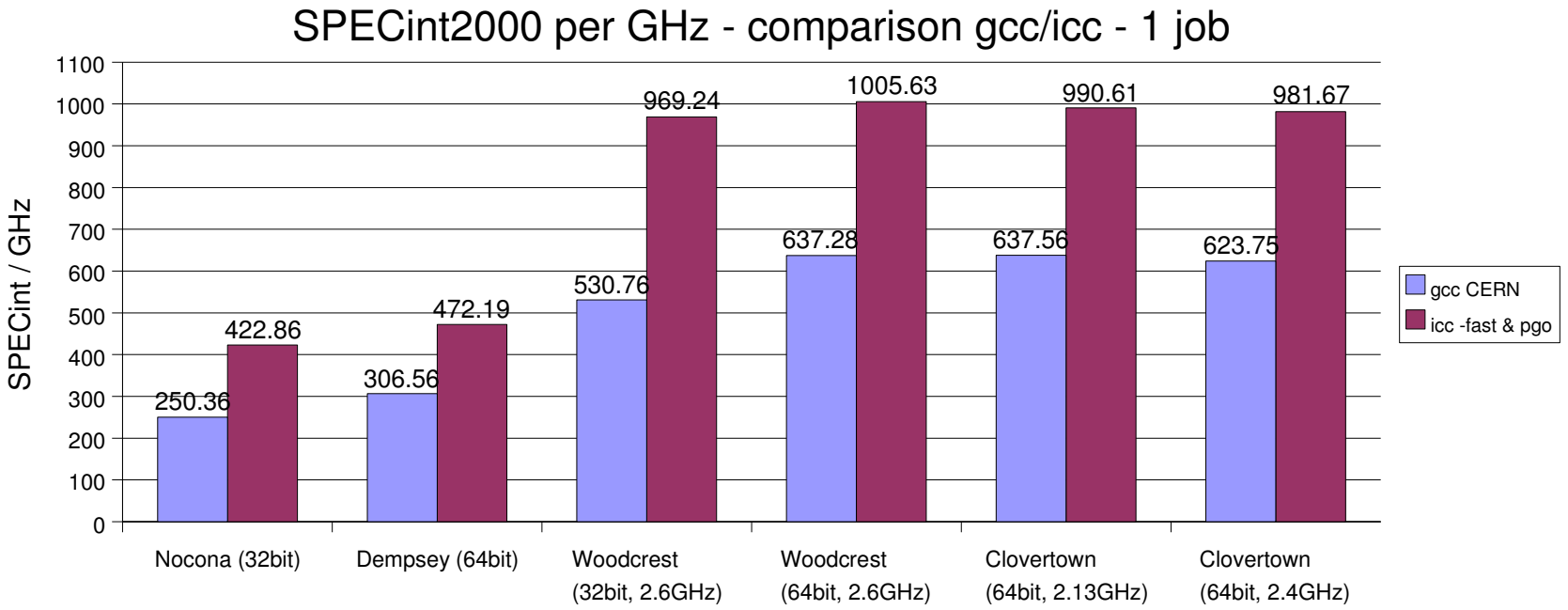
(n=# of cores)

- SPEC2006 version 1.0
  - runtime is ~6 - 8 times longer

Some Results

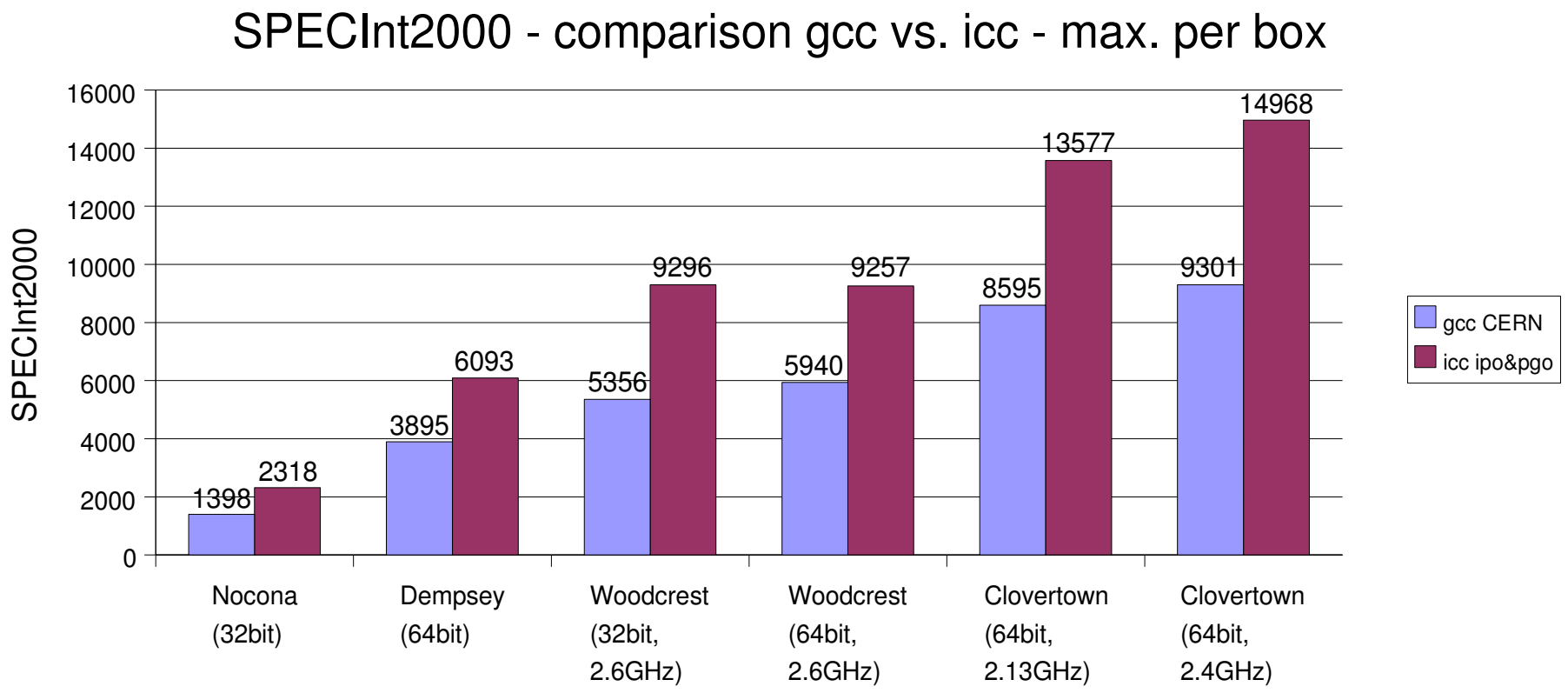
The measured SPECInt2000 for different CPUs and different compilers (and compiler options).

The result is normalized to SPECInt per GHz for a better comparison



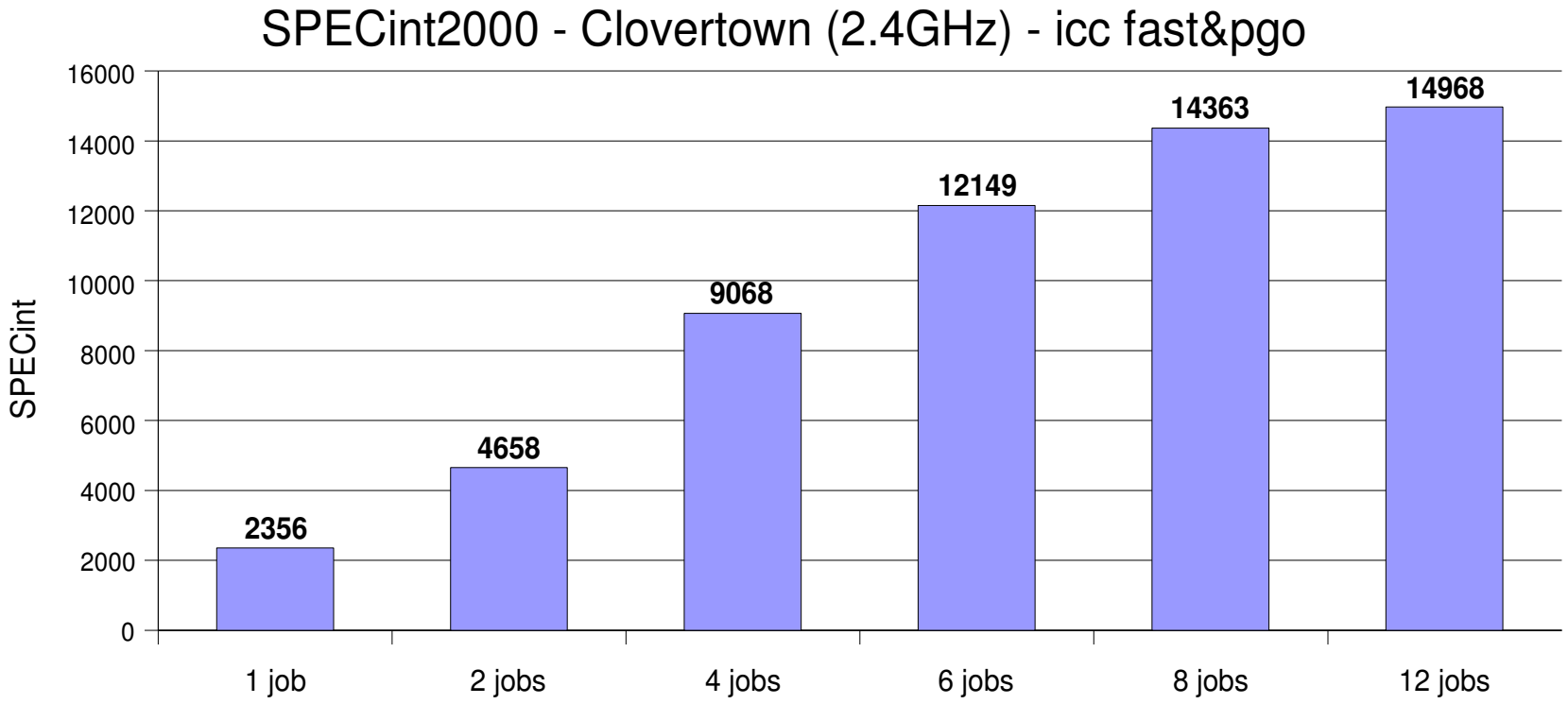
More results

The measured SPECInt2000 for different CPUs and different compilers (and compiler options).

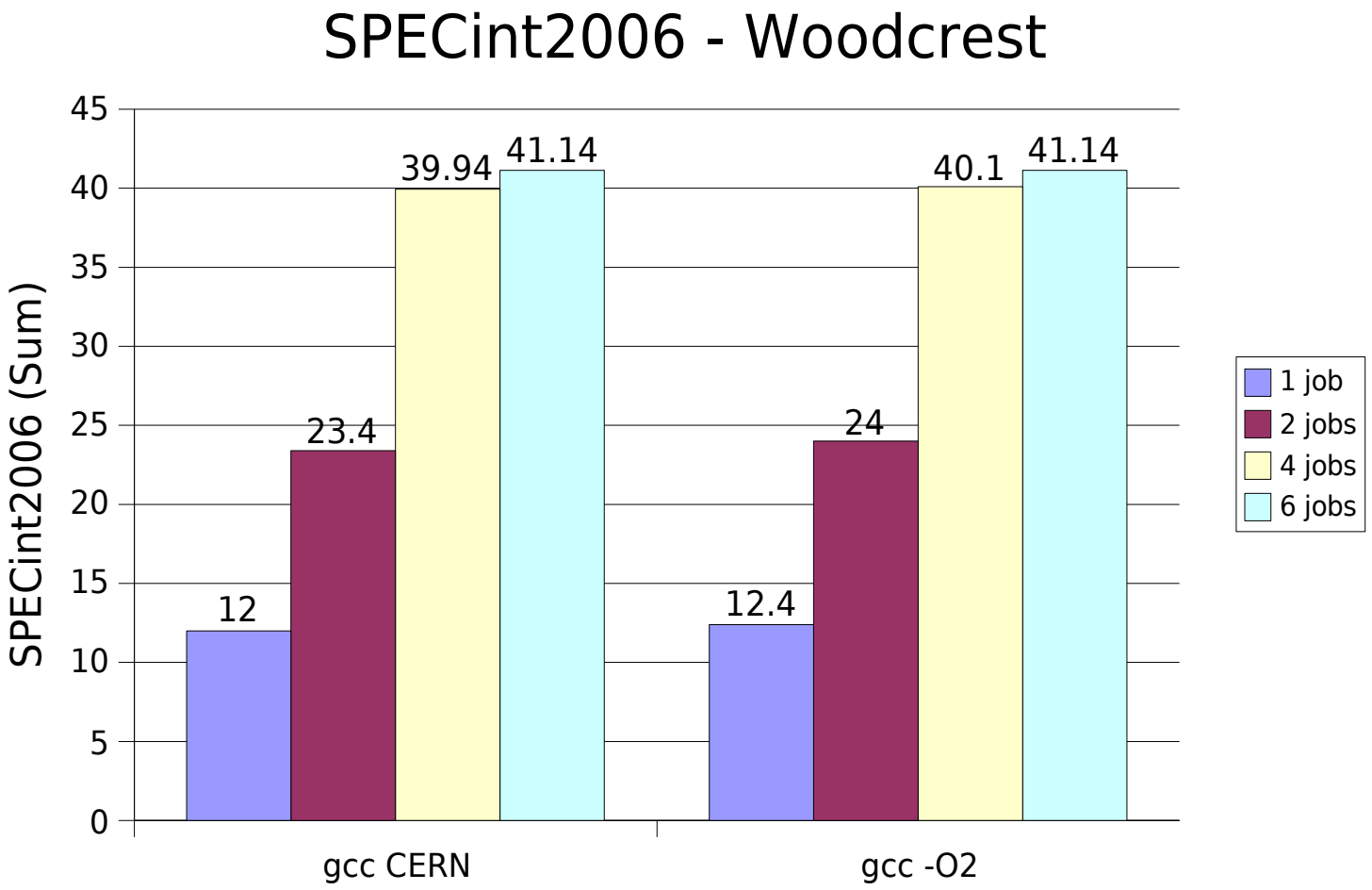


Clovertown - SPECInt2000

- performance comparable, but slightly worse than Woodcrest
- on a 2.4GHz machine the max. aggregate is about 15000 SPECint



Woodcrest – SPECInt2006



Background on Power consumption

- Only the power consumption of the entire box is important
- The main consumers are
  - Server CPUs consume 65-130W (lower power versions do exist as well)
    - 130-260W for a dual-socket system
  - Memory (under load)
    - DDR2 consumes ~4-5W per GB
    - FB-DIMM consume ~10W per 1GB module
      - ~5W for the memory (as DDR2 memory)
      - ~5W for the controller chip (AMB)
    - Clovertown system with 2GB per core: 160W (!!)
- The power consumption varies significantly under different load conditions!!

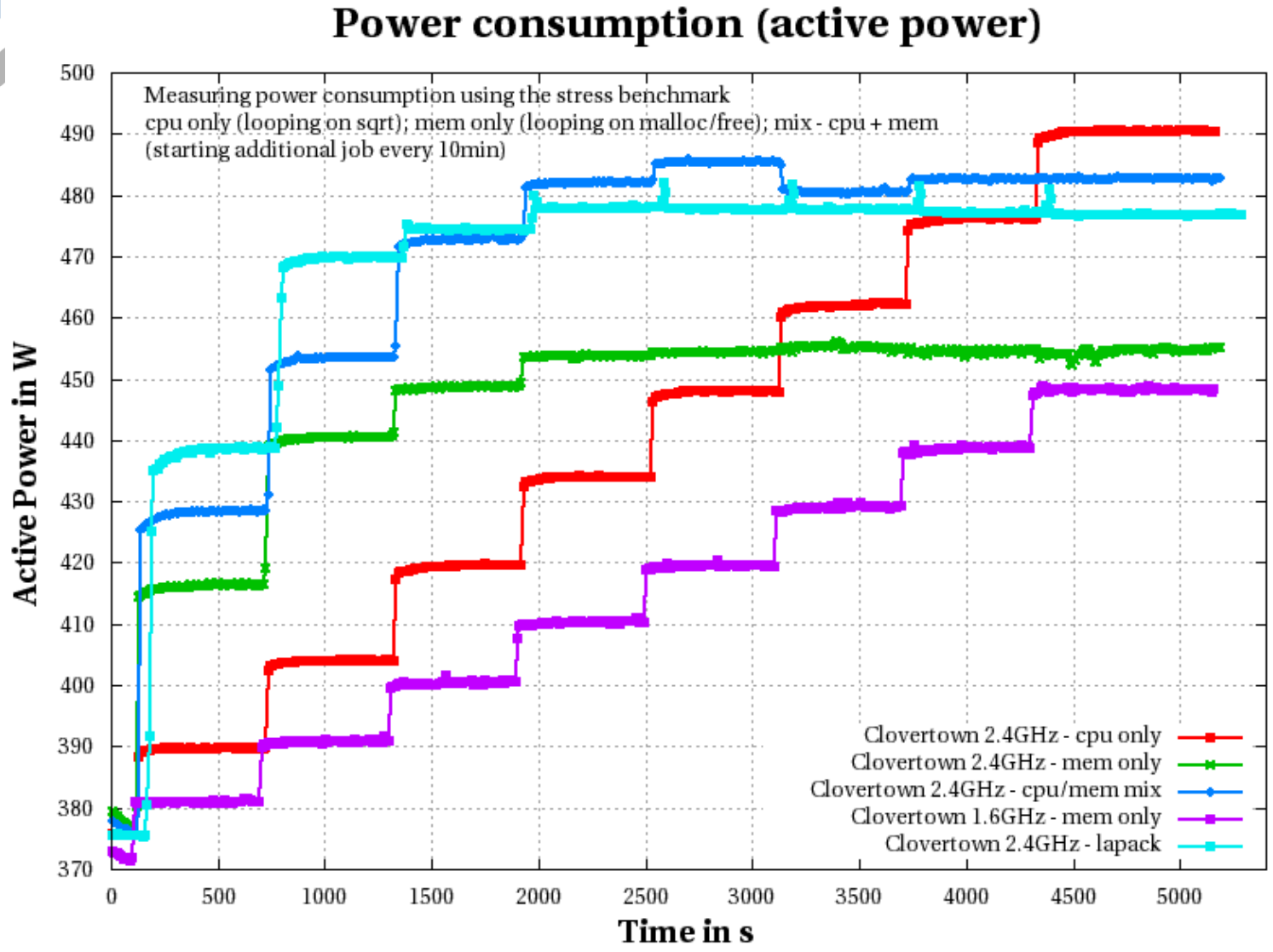
Power Consumption - Measurements

Measurement of the power consumption of

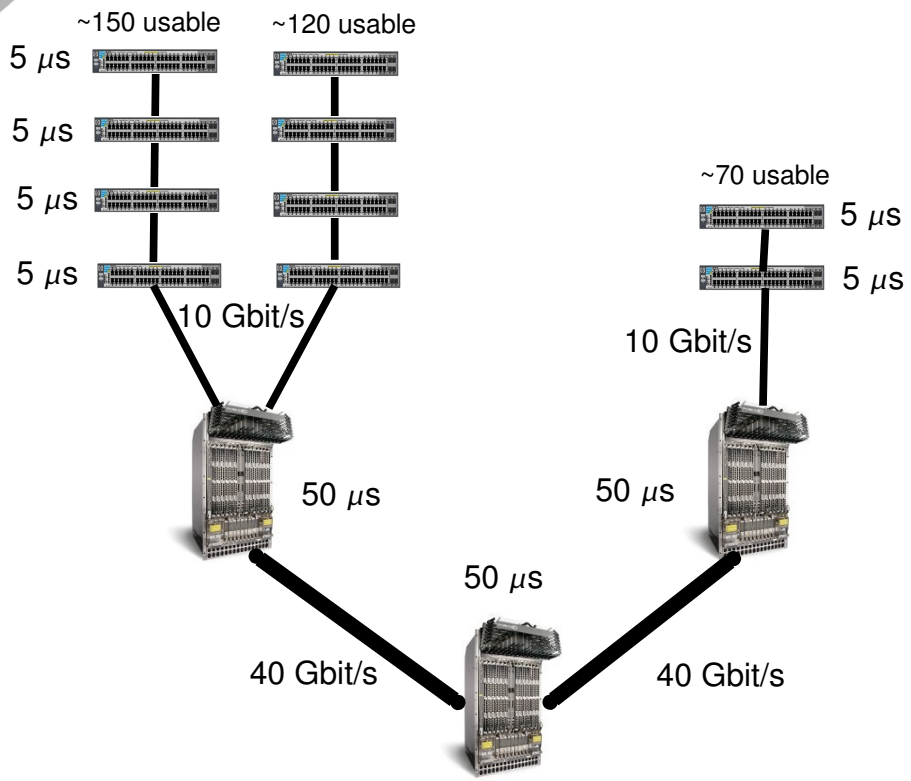
- **Dempsey**
  - 4 GB FB-DIMMs (DDR2 based) (~10W per DIMM)
  - 3 disks (~11W idle, ~16W active)
- **Woodcrest**
  - 4 GB FB-DIMMs (DDR2 based) (~10W per DIMM)
  - 3 disks (~11W idle, ~16W active)
- **Clovertown**
  - 12 GB FB-DIMMs (DDR2 based) (~10W per DIMM)
  - 4 disks (~11W idle, ~16W active)

	Clovertown	**Woodcrest**	**Dempsey**
Max. utilisation	~460W	~290W	~410W
SPECint/Watt icc & pgo	32.6/31.9	31.9/29.8	14.9/14.2
SPECint/Watt gcc CERN	20.2/20.0	20.5/19.2	9.5/9.0
(max. power consumption while running SPECInt)	actual/normalised to 2GB/core	actual/normalised to 2GB/core	actual/normalised to 2GB/core

Power Consumption – Measurements II



The TOP500 setup and result



- 340 machines
  - Dual Woodcrest 3GHz
  - 8GB RAM

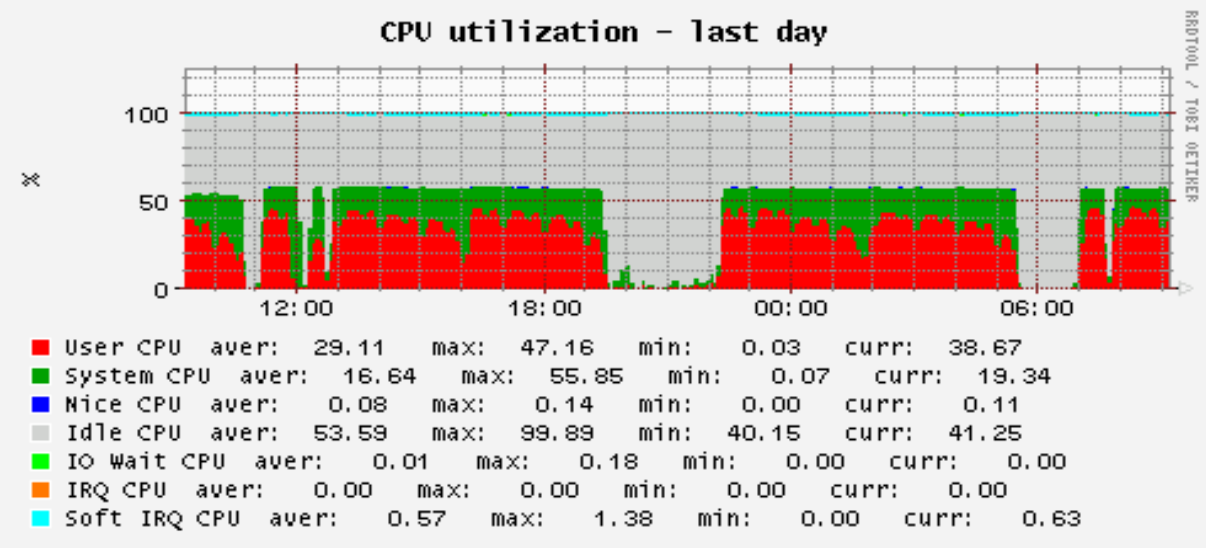
➔ 1360 cores

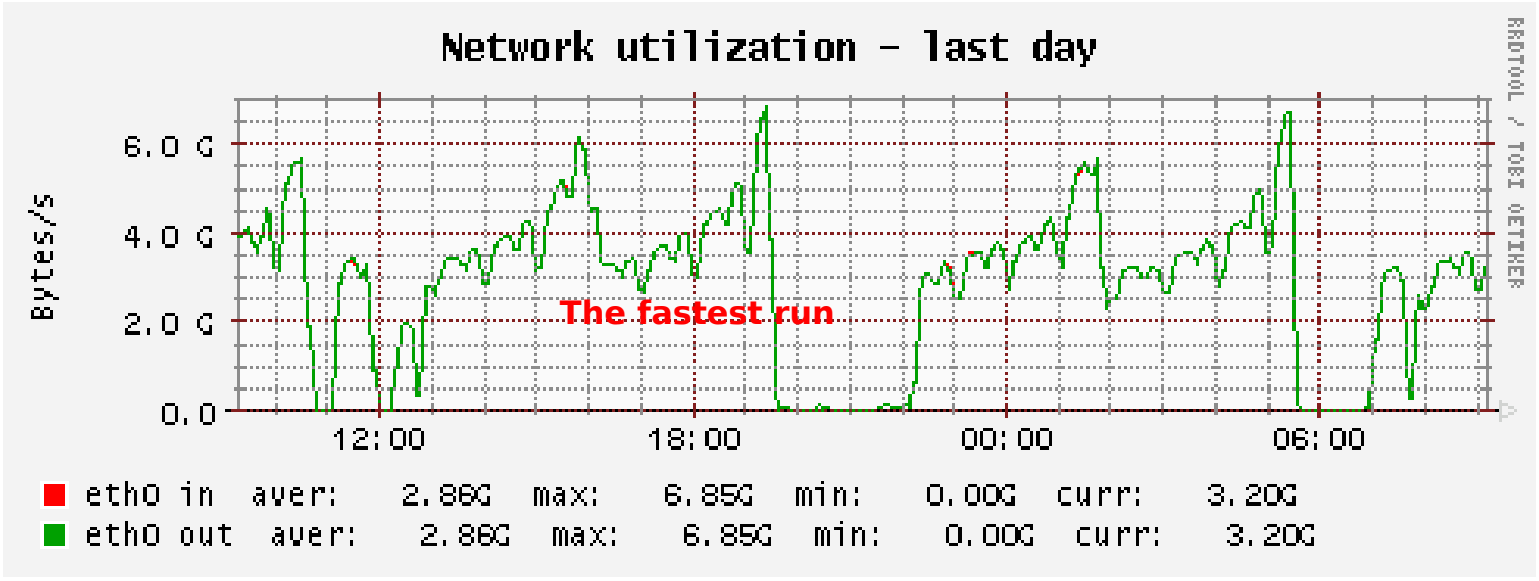
... delivering:

8329 GFlops

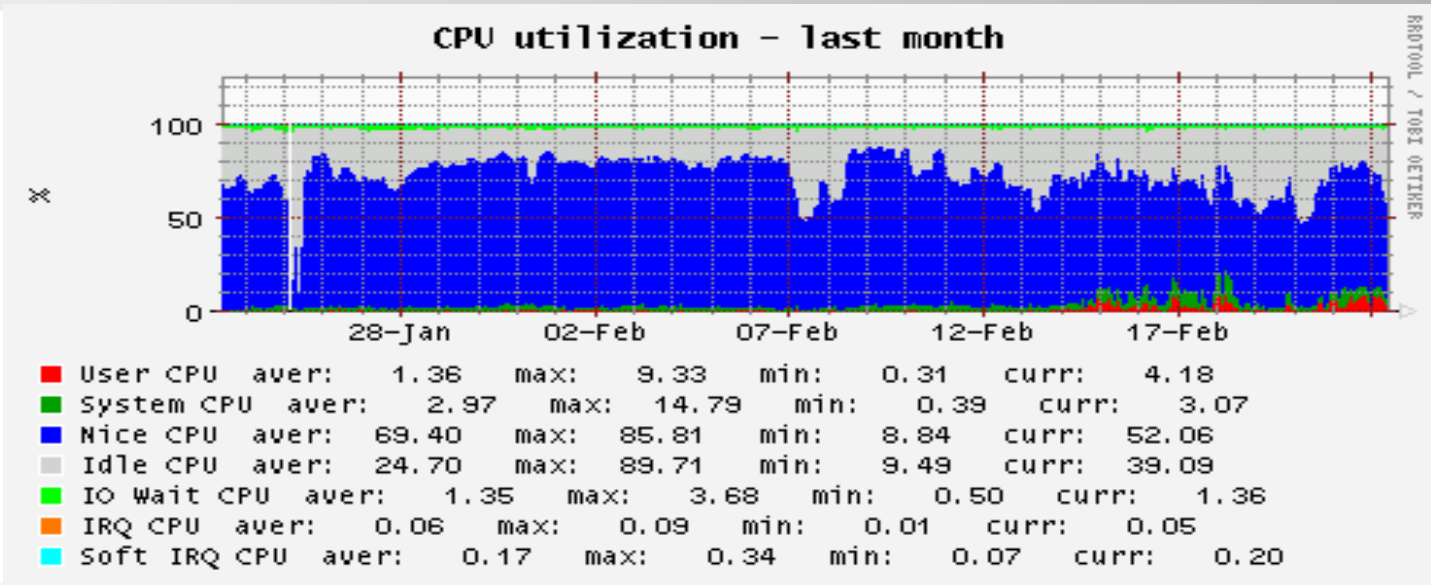
- 6.12 GFlops per core
- 51% efficiency !

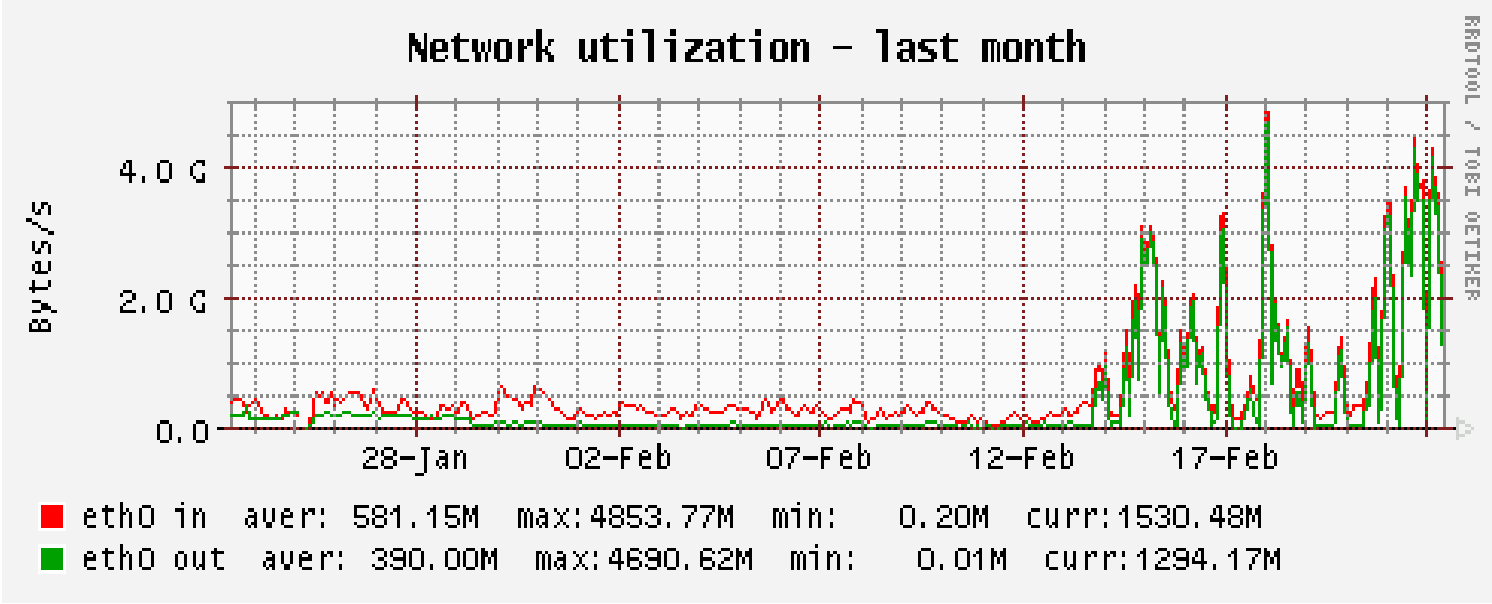
... how it looks in the monitoring





… compared to the entire lxbatch farm





Conclusions

CERN IT achieved a remarkable performance with High Performance Linpack and Intel MPI

8329 GFlops with 1360 cores

(6.12 GFlops per core ≙ 51% efficiency)

- setup not optimized (h/w or s/w wise)
- for our type of (network) setup extremely good result
  - other GigE based clusters: 19 - 67 % efficiency
- would be rank #79 in current list
- being submitted to the TOP500 committee
  (as soon as the submission webpage is online again)
- HPL is extremely sensitive to it's parameters ...

Acknowledgments

- Intel:
  - Sergey Shalnov
  - Intel MPI people
- CERN
  - Ulrich Schwickerath
    - LSF and general help (debugging, etc.)
  - Veronique Lefebure
    - installation of machines
  - Nick Garfield (and others in CS group)
    - looking into the network setup