Benchmarking topics at CERN

Helge Meinhard / CERN-IT

HEPiX, GSC St Louis MO USA

06 November 2007

Outline

- SPEC 2006 at CERN
- Recent calls for tender
  - SPEC 2000
  - Adjudication
  - Power consumption
  - Results
- LINPACK / Top 500
- SPEC Power

CERN and SPEC 2006

- By far not as advanced as INFN and GridKA
  - Initial tests, some comparisons started
- Procurements so far using SPEC 2000
  - Introduced SPEC 2000-based adjudication 1.5 years ago
  - Some learning curve on vendor side
  - Series of tenders ran since
  - Some gap until next tenders, will consider migrating

CERN tenders and SPEC 2000

- SPEC defines an application suite, but not an environment
  - Vendors submitting SPEC results optimise OS, compiler, compiler flags, other conditions
  - For our tenders, we want that SPEC rating reflects as closely as possible the value of a machine in our environment and for our use case – farm processing of user jobs
    - Fix OS (RedHat Enterprise 4 x86_64)
    - Fix compiler (RHES 4 gcc system compiler)
    - Fix compilation options (`-O2 –fPIC –pthread`)
    - As many SPEC runs in parallel as there are CPU cores in the machine

CERN tenders: Adjudication

- Example of our past two tenders for worker nodes:
  - Purchase price of as many nodes as are required to achieve adjudication quantity (2 MSPECint2000)
  - 300 CHF per system unit (aka mainboard) for CERN infrastructure cost
  - 50 CHF per system unit if dedicated line required for IPMI
  - 6 CHF/VA of power consumed

CERN tenders – power: why 6 CHF/VA?

- Elements taken into account for farm nodes:
  - Power consumption of machine over 4 years
  - Cooling power for machine over 4 years
  - Depreciation of infrastructure cost
    - Following industry practice, assuming 10 years' lifetime of infrastructure
    - Add 40% of infrastructure per VA
- For equipment in critical area (dual UPS, Diesel generator) we use 10 CHF/VA

CERN tenders: power consumption

- No widespread standard benchmark available
- Procedure defined to be run by bidders
  - Fully configured enclosure (e.g. blade chassis filled up with blades)
  - SLC4 x86_64 installed
  - Run idly, and fully loaded
    - Fully loaded: 50% cores run CPUburn, 50% run LAPACK
  - For worker nodes, use average of 80% loaded + 20% idle
- High-precision power meter recommended
- Only interested in apparent power (VA) in primary AC circuit (and in power factor > 0.9)

CERN tenders: penalties

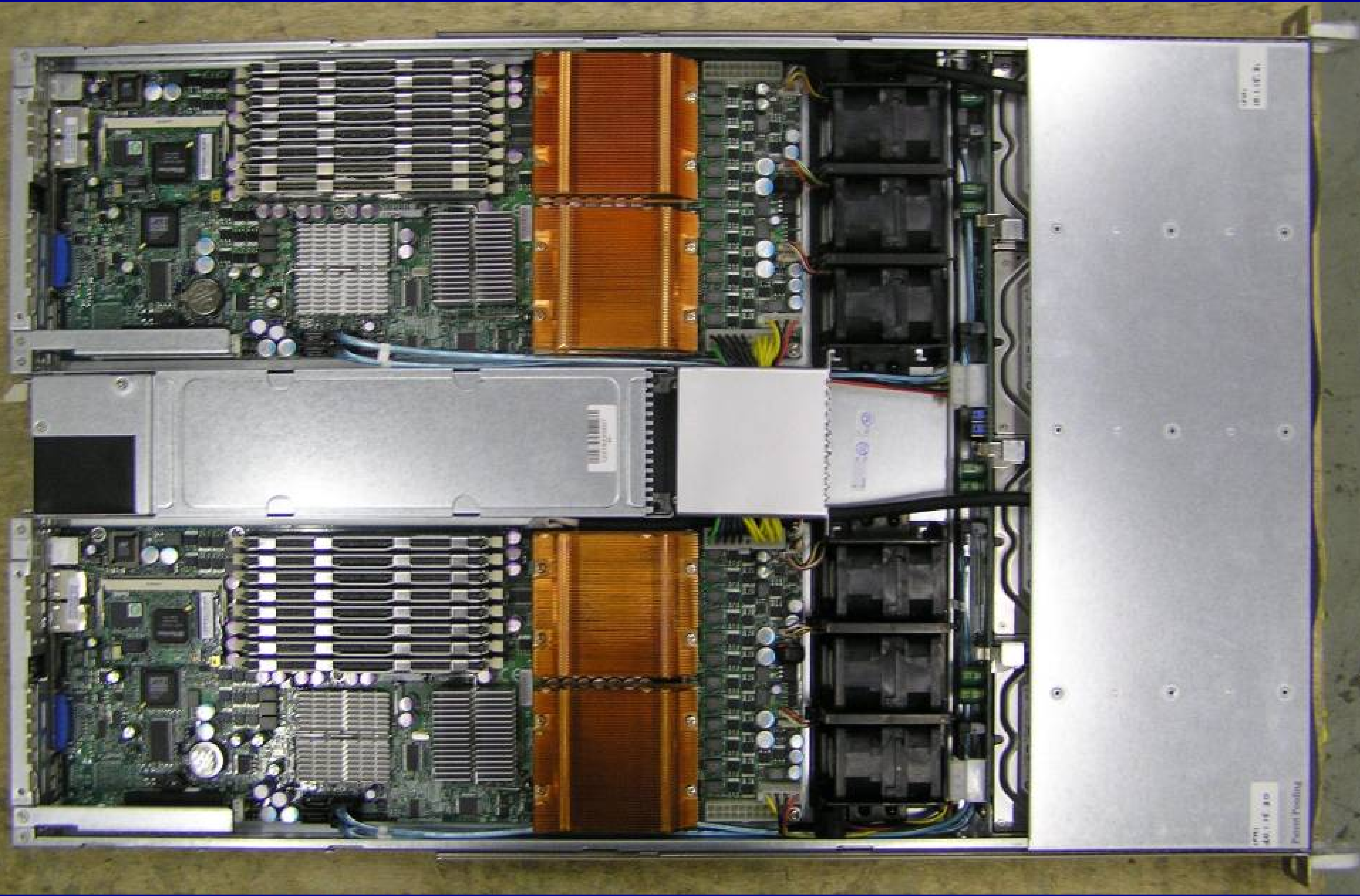
- If box performance is >1.5% lower than indicated: At CERN’s discretion
  - Request corresponding number of nodes for free
  - Pay only pro-rata amount of bill
  - Send the batch back
- If power consumption is >5% higher than indicated: At CERN’s discretion
  - Subtract corresponding amount from bill (6 CHF/VA)
  - Send the batch back

CERN tenders: experience

- Bit of a learning curve for vendors
  - A little less so for SPEC, a little more so for power
- Some vendors don’t seem to measure power, but use some internal spreadsheet tools to estimate
  - Usually found too high, sometimes even by a long way
- No big problems anyway
  - Vendors understand why we are proceeding this way

CERN tenders: results

- CPU tender for 3 x 2 MSI2k open for different form factors
  - Had classical 1U pizza boxes and blade systems in mind
  - Got something else – Supermicro Atoca (2 slim mainboards in a 1U chassis) as number 1, 2 and 3
- CPU performance (rather) independent of form factor
- Power: a little surprise…
  - Twins: 35 mVA / SI2k
  - Blades: 35…42 mVA / SI2k
  - Classical 1U pizza boxes: 37…66 mVA / SI2k



CERN tenders for disk servers

- In first round, used power consumption only for worker nodes
- Encouraged by good experience, did the same for disk servers in second round
- Allowed us to open up from storage-in-a-box only to solutions with a 1U front-end server and an external disk extension
  - Two-box solutions competitive on purchase price, but not including power element

December 2006 CPUs: LINPACK (1)

- Proposed and supported by Intel
- Theoretical max: 30 TFlops (48 GFlops per machine)
- Very little experience with parallel computing at CERN, in particular MPI
- Other systems in Top500 are either huge multiprocessor machines or clusters with low-latency interconnects; our setup: factor 60 higher latencies
- Standard machine setup with all daemons, no special tuning
- Intel MKL, Intel MPI

From my presentation in Hamburg

December 2006 CPUs: LINPACK (2)

- Started with 530 machines, first tests run successfully with 256 machines
- One batch of three had to be taken out - networking problems
- Linpack tuning required to avoid bottlenecks in 10 Gbit/s uplinks from switches to routers
- In the end: 340 machines (1360 cores) achieving 8’329 GFlops
  - N=530’000; NB=104; P=16; Q=85
  - 25 GFlops per machine = 51% of theoretical max
  - Would have been position 79 if submitted for SC fall 2006

From my presentation in Hamburg

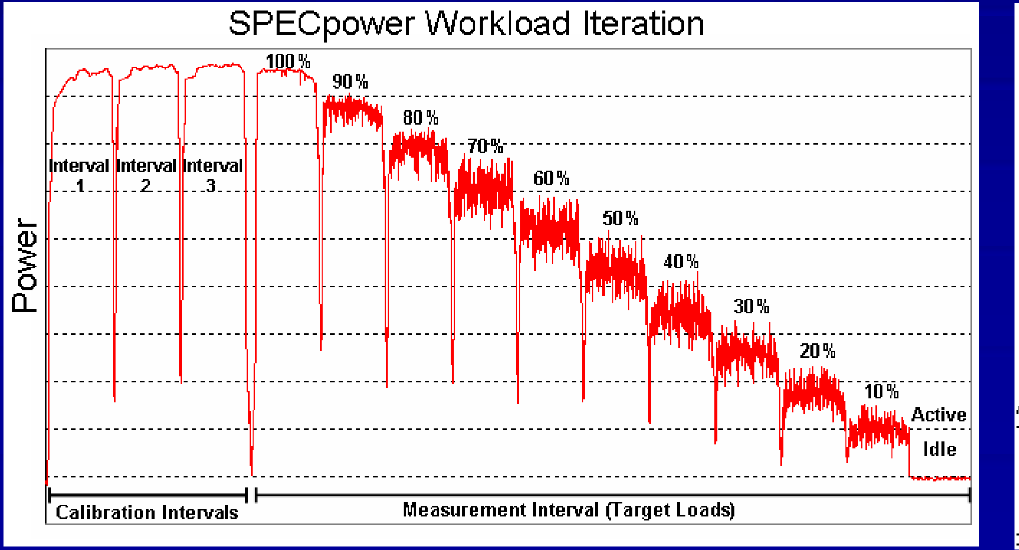
LINPACK for Top 500

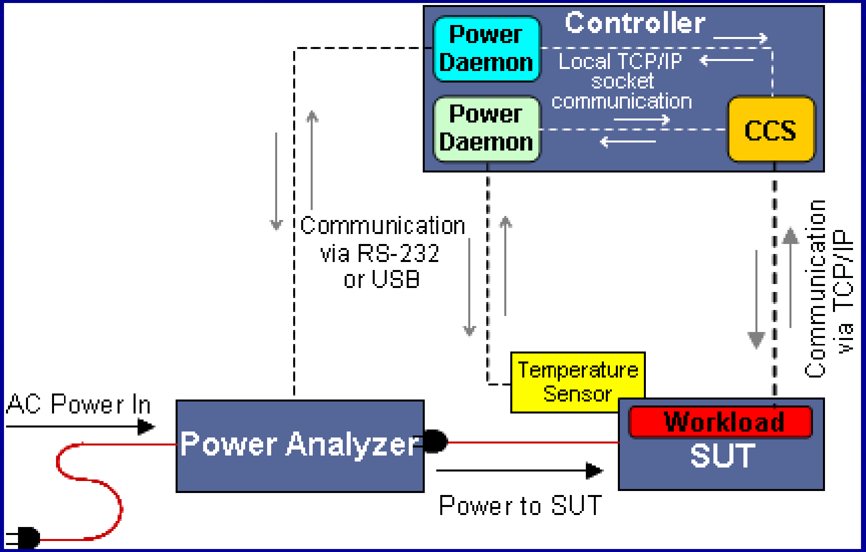
- Result of 8’329 GFlops submitted to SC June 2007 in Dresden
- Obtained position 115
- Will try and redo with massive delivery of 620 twin-based dual Clovertown systems

Future: SPEC Power

Slides courtesy of Alex Iribarren

- Latest SPEC benchmark, currently beta
- Purpose: reliably measure power consumption at different usage levels
- Methodology + Software framework + Workload (currently only SPECjbb2005)





SPEC Power: why we're interested

- Well-defined methodology
  - Minimum requirements for power meters
  - Defined environmental conditions
  - Strict run and reporting rules
- Extensible software framework
  - Use our own workload
- “Run SPEC Power with this workload”
  - We get repeatable and comparable results

CERN and SPEC Power: Current status

- Early contacts with members of the SPEC Power working group – SPEC very interested in feedback
- CERN gave feedback based on discussions and documents
- We have received the beta kit of SPEC Power (today!)
- Tests will start next week, and run until end November
- Will try to report at next HEPiX

Conclusion

- Significant steps made, and still being made, towards HEP-wide solutions compatible with industry standards
- Still a lot of work ahead of us…