Bull

Architect of an Open World™

High Definition Energy Efficiency Monitoring

TECHNISCHE UNIVERSITÄT DRESDEN

September 2nd, 2014

Marc Simon, Daniel Hackenberg, Thomas Ilsche, Joseph Suchard, Robert Schöne, Wolfgang E. Nagel, Yiannis Georgiou

Abstract

While the cost of energy becomes the limit factor in growing computer center capacity, it becomes more and more important to have a precise measurement of the power consumption of each component of the cluster.

The need for more precision implies a higher sampling rate on the sensor values. In addition, a better time resolution allows the user to match power consumption to the internals of his code. The volume of data generated make it necessary to move a part of the measuring work closer to the components on the managed board. A first step is to move the work of calibrating the sensors and aggregating data at the BMC level. In a second step, a dedicated FPGA keep a track of power consumption at high speed and make it available to the BMC. While agregated data may still be read through the administration network for accounting purpose, a high frequency recording of current variation is made available to the host through a PCIe link.

After a prototype based on water cooled B710 blades for HPC and IvyBridge processors, this solution for high precision energy measurement will be available on the next generation of B720 blades with Haswell processors.

The power measurement challenge

Architect of an Open World™

TECHNISCHE UNIVERSITÄT DRESDEN

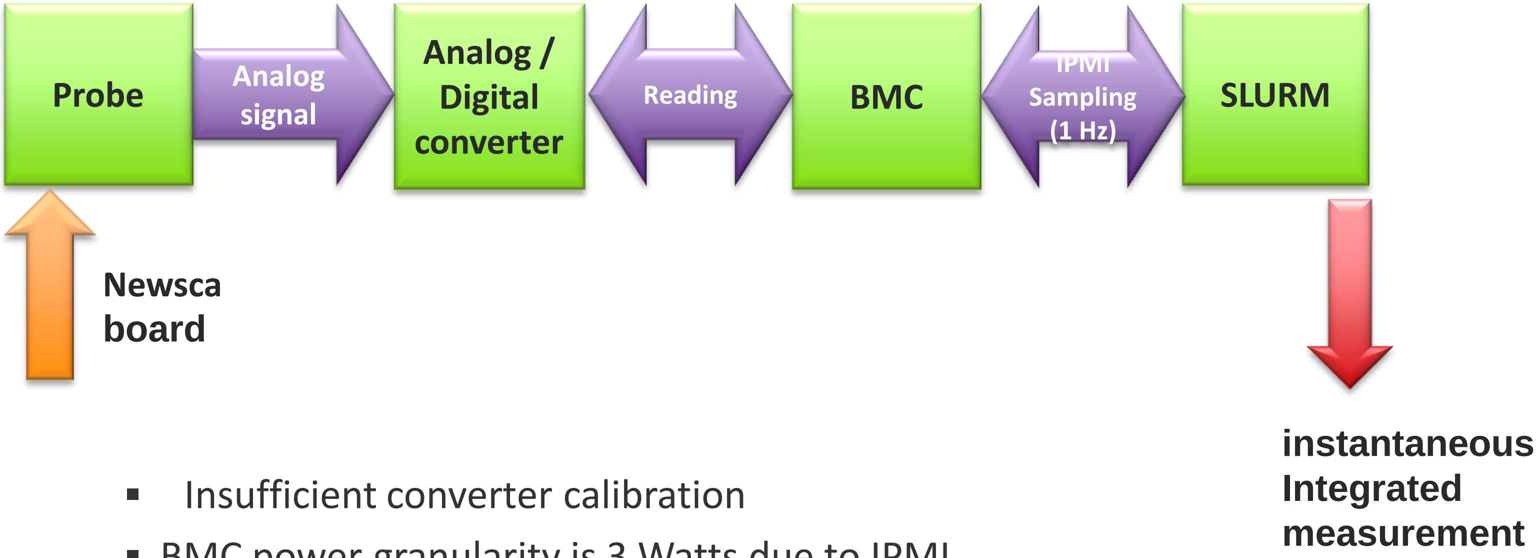
Need for more accurate measurements

- HPC users need use power measurement for several purposes:
- Energy accounting
  - Need to have per node data
  - Energy aggregation through time linked to resource allocation
- Long time efficiency evaluation
  - The data center efficiency must be evaluated by long time (typically 1 year) data aggregation
- Fine grain optimization
  - Need power measurement with granularity below 1 second
  - Need to separate CPU, DIMM, Disks, Network

Energy efficiency measurement

- PUE / TUE
  - With water cooling, the frontier between host and computer room is blurred
  - Need to compare systems at component level (CPU, DIMMs). Server or blade level is no longer sufficient
  - Some data available at Voltage Regulator (VR) level can be used.

Newsca 1 perimeter (no fpga)



- Insufficient converter calibration
- BMC power granularity is 3 Watts due to IPMI constraints
- Internal BMC sampling rate at 4Hz for energy measurement
- SLURM polling to BMC (1Hz) through IPMI

Power measurement problematics

□ Calibrate the sensors

- Embedded sensors have a poor precision (5% - 10%)
- Need a calibration depending on each sensor

□ Time resolution

- Communications between the BMC and administration server are slow
- Need to collect and buffer data closer to the sensors

□ Aliasing

- Simply sampling values show aliasing effects. High frequencies must be filtered and oversampled.

□ Data collection

- Statistics must be made available to a remote administration server

□ Overhead

- All this must be done with little or no overhead on compute nodes!

On board embedded FPGA architecture

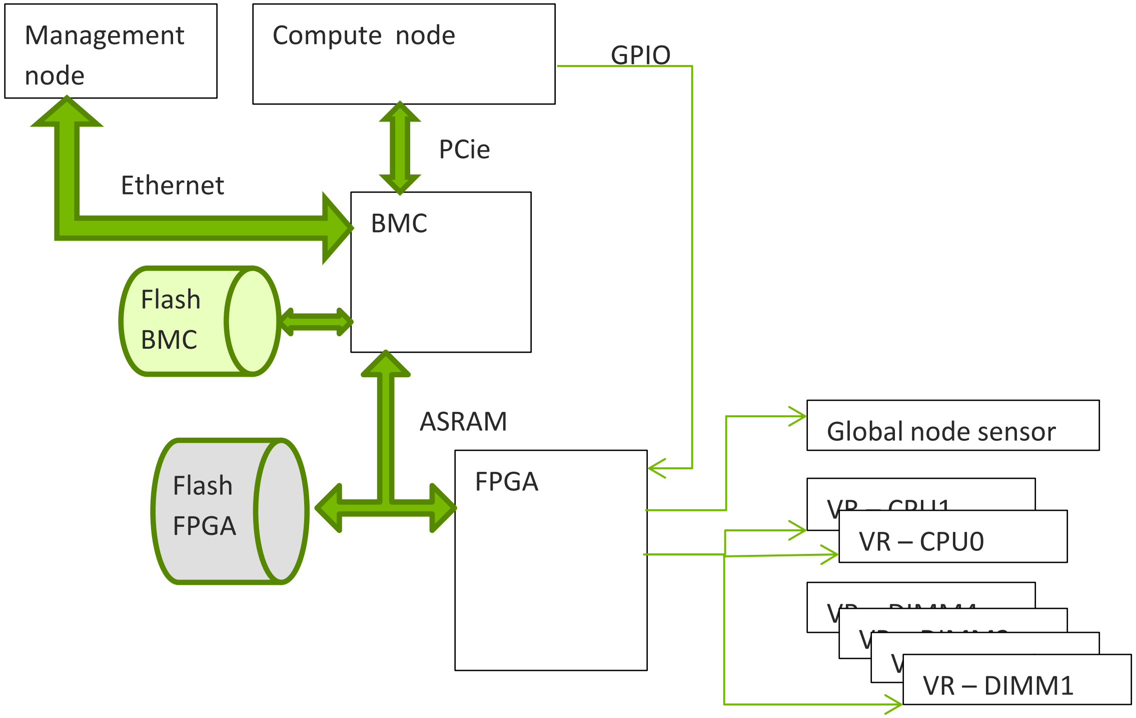
Architect of an Open World™

TECHNISCHE UNIVERSITÄT DRESDEN

Embedded architecture

□ FPGA implementation

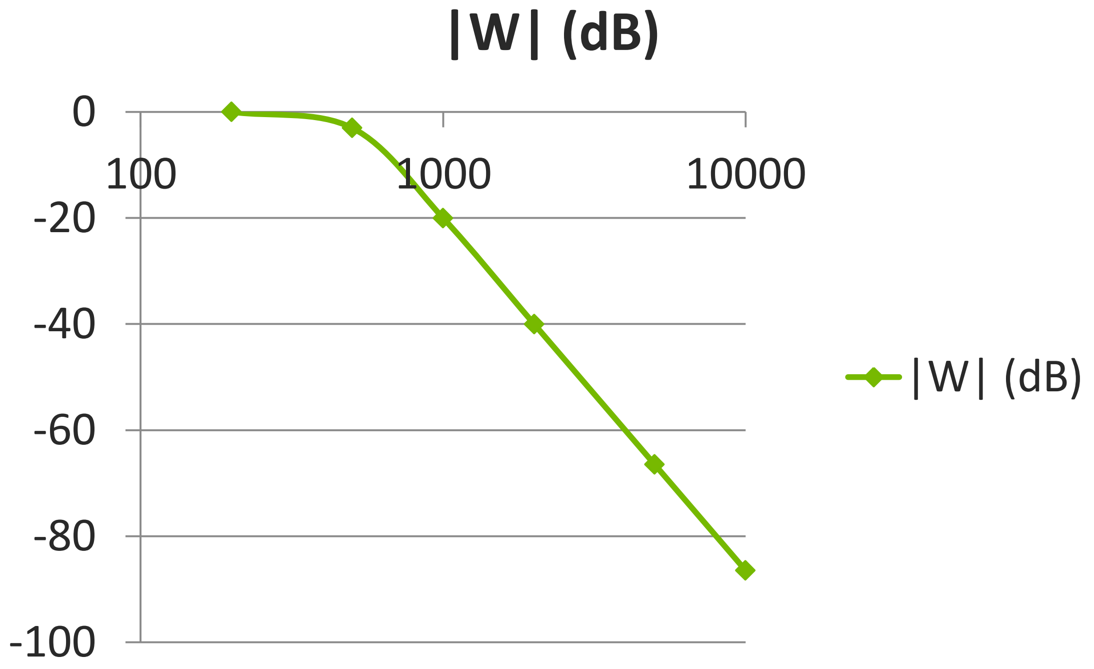
- Most work is offloaded to a FPGA
- Data buffering is done into the BMC



Anti-aliasing

□ Analogic filtering

- The analogic data are filtered by a 2nd level filter
- Sampling frequency so that remaining harmonics are below 1% (-66dB)



- Cut off @ 500 Hz => sampling at 8000 Hz

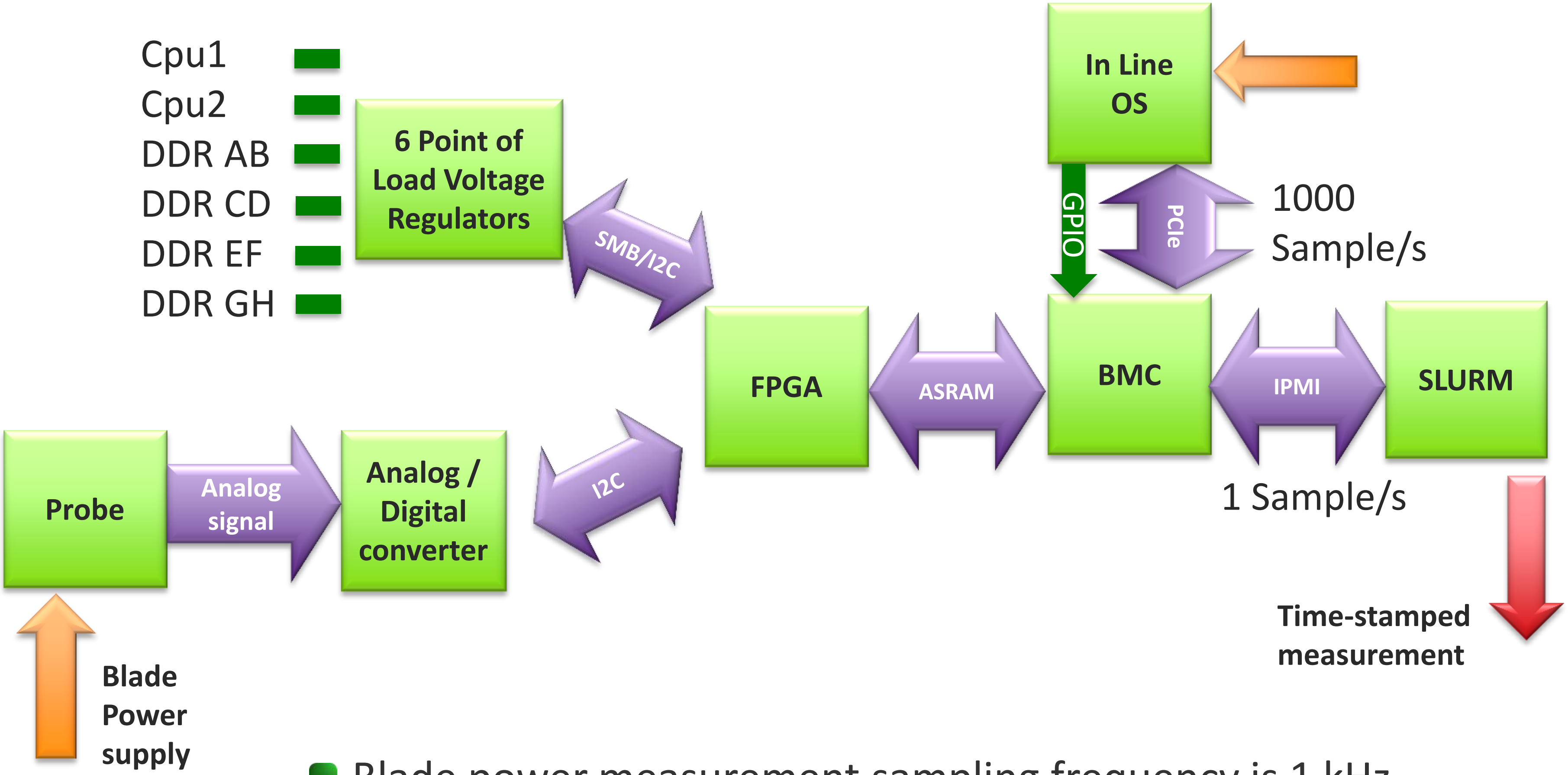
1) High precision power sampling

- New technologies make possible more precision
  - GPIO allows to drive the FPGA from the operating system with very short latency
  - More memory available on the BMC (pilot3)
  - PCIe access to the BMC from host
- High speed sensor reading
  - Values are stored at 1000 samples/second for the blade sensor
  - VR data are stored at 100 samples/second
- No host overhead
  - Buffering in the BMC allow zero overhead measurement for several hours.

2) Long term energy reporting

- Cumulative energy counter
  - FPGA’s internal clock allow regular sampling (independant of BMC and host activity)
  - Instantaneous power values are cumulated to give an energy counter.
- SLURM
  - SLURM is a Job Manager that schedules tasks on the HPC nodes and collect the corresponding energy consumption
  - Starts and reads the energy counter through each BMC through IPMI

Power acquisition block diagram



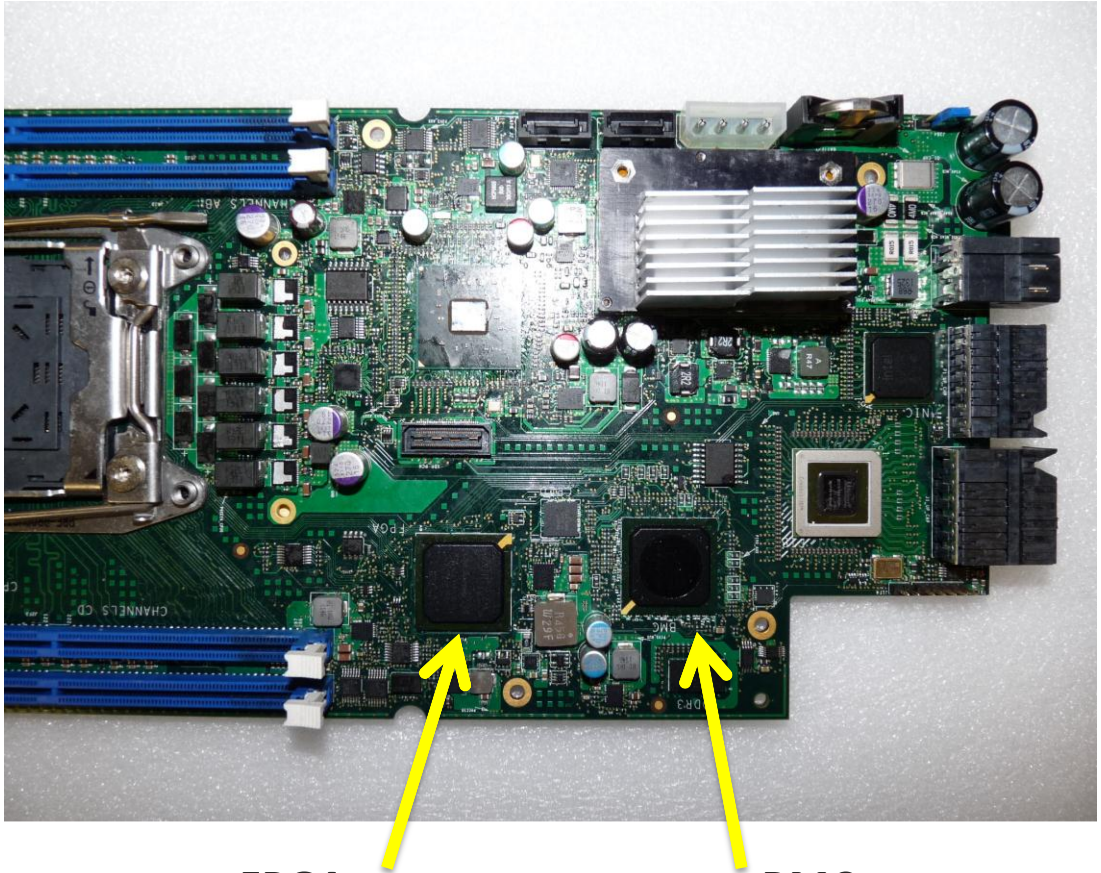
- Blade power measurement sampling frequency is 1 kHz. Accuracy < 2% above 100W (including aliasing).
- Cpu and DDR power measurement is sampled at 100 Hz

Implementation

Architect of an Open World™

TECHNISCHE
UNIVERSITÄT
DRESDEN

B720 layout

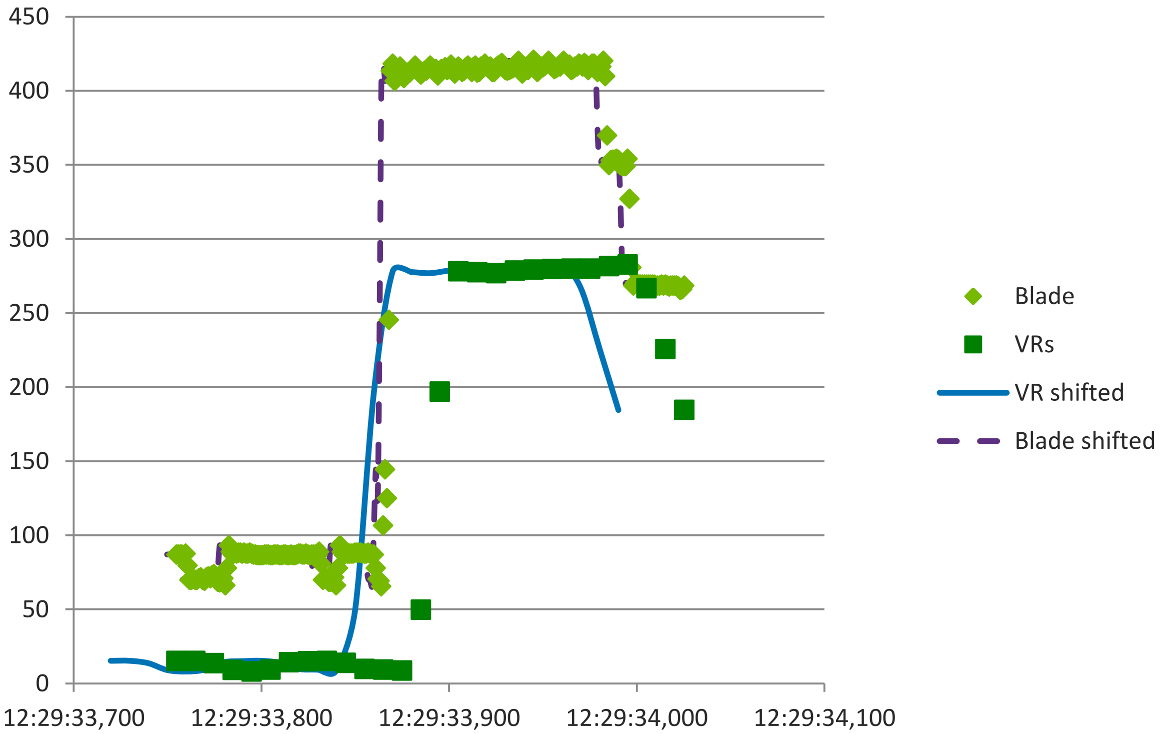


FPGA

BMC

Real time measurements

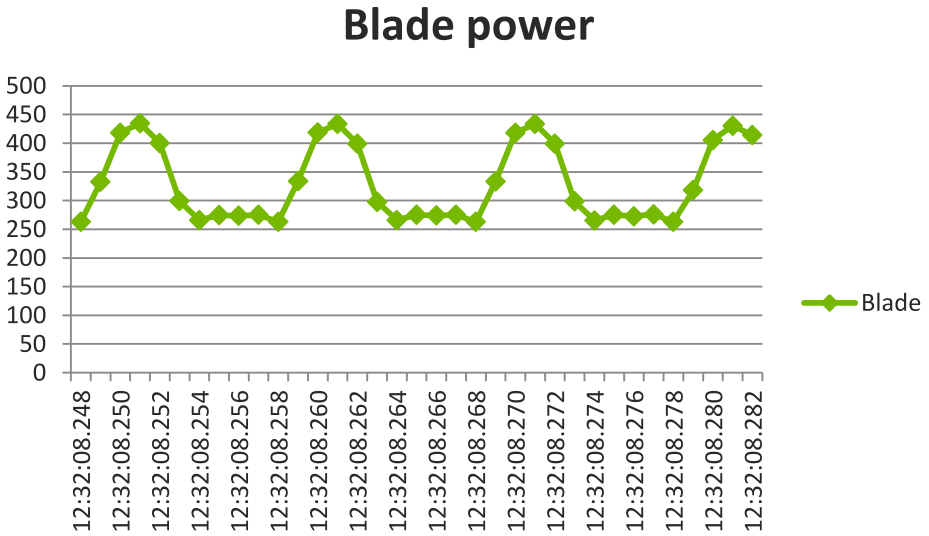
▢ Response to a short impulse



- Filters applied on raw collected values produce a shift that has to be compensated

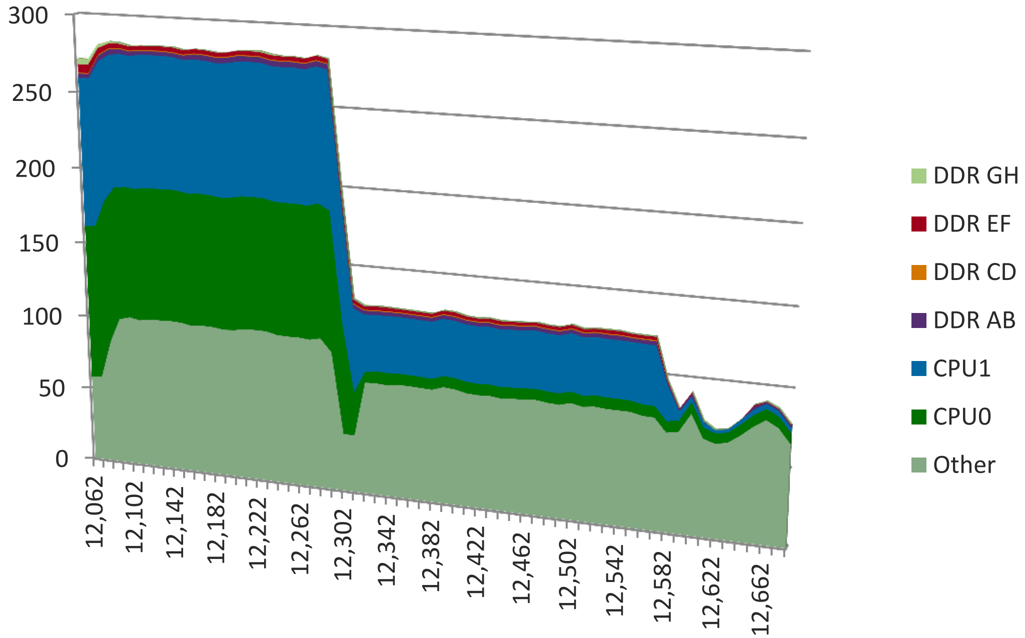
Real time measurement

- □ Compute intensive every 10 ms



Real time measurement

□ Profile of task ending



- Relative importance of CPU / DIMMs / Other can be seen depending on the instantaneous load
- « other » include BMC, network, HDD,…

Bull

Architect of an Open World™