Energy-Efficiency Research at ZIH

Or: What the HAEC is HDEEM?

Daniel Hackenberg (daniel.hackenberg@tu-dresden.de)
Thomas Ilsche (thomas.ilsche@tu-dresden.de)
Joseph Schuchart (joseph.schuchart@tu-dresden.de)
Robert Schöne (robert.schoene@tu-dresden.de)
Agenda

- What is ZIH?
- What is HDEEM?
- What is HAEC?
- What is the rest?
- What is the future?
Center for Information Services and HPC (ZIH)

- Central scientific unit at TU Dresden
- Providing infrastructure and qualified service for TU Dresden and Saxony
- Competence Center for „Parallel Computing and Software Tools“
- Nvidia CUDA Center of Excellence
- Intel Parallel Computing Center
- Energy-Efficiency Cooperation with Bull SAS
Taurus: Petaflop Machine in Dresden

- 1456 nodes 2x Intel Xeon E5-2680 v3 (12 cores) @ 2.50GHz
- 64 nodes 4x NVIDIA Tesla K80
- HDEEM infrastructure

Fotos: Robert Gommlich/ZIH
Challenges of EE Measurements

- **High Resolution**
  - **Temporal:** usually ~1s – 1ms
  - **Spatial:** data center → rack → chassis → node → VR

- **Correctness:**
  - Internal vs external sampling
  - Instantaneous vs accumulated sampling
Average Power Sampling

Challenges of EE Measurements

- **Aliasing** due to low sampling rate

![Graph showing 2 Hz (0.25s high/0.25s low) and 1/8 Hz (4s high/4s low) signals](image-url)
Challenges of EE Measurements

- **High Resolution**
  - **Temporal:** usually ~1s – 1ms
  - **Spatial:** data center → rack → chassis → node → VR

- **Correctness:**
  - Internal vs external sampling
  - Instantaneous vs accumulated sampling

- **Accuracy:**
  - Time: rather easy
  - Voltage and Amps?
    → Calibration required!
Energy-Efficiency Research @ZIH

Phase 2 High Res
Energy Accounting

1 node 100 nodes 1k nodes full HPC system

1 Sa/s 1k Sa/s >10k Sa/s
High Definition Energy Efficiency Monitoring

Live viewers, e.g. Website

Post-Mortem Analysis, e.g.
- Get energy of job
- Get energy for node
- Analyze long term energy savings since optimization
- Websites

Compute Node

Admin Node

Dataheap Server

Database
The Classics: Live System View
The Classics: History View

- Power at $\geq 1$ sample/s, temperature at $> 1$ sample/min
- All components: nodes, switches, storage etc.
- Full history, no data reduction/dumping
COSMO on 4 nodes of Taurus phase 1 with power trace
(64 application processes above, node power profile below)
HDEEM Phase 2 Power Acquisition Scheme

1. High-freq. power sampling
2. Analogue low-pass filtering
3. A/D conversion
4. Digital filtering
5. Flush power data to BMC
6. Query data from BMC

- 2xCPU, 4xRAM measurement at voltage regulators
  - 1000 samples/s
- Blade sensor analog filter
  - 500 Hz
- A/D converter
  - 8000 samples/s
- FPGA short-term buffer
  - digital filter 100 samples/s
- Dataheap collection and history charts
  - 1 sample/s
- IPMI accounting and profiling
  - 1 sample/s
HDEEM: Hardware
HDEEM: Some Lessons to Share

- IPMI standard protocol limited to 1 Byte
  - Use proprietary protocol extensions
- Timestamp/Energy tuple:
  - Everything else is hard to get right (see RAPL)
  - Used in start-stop-interface
- Easy access: C-API & CLI tools
  - Use technologies that work for unprivileged users
- Accuracy is hard to maintain at 2%

http://tu-dresden.de/zih/hdeem
Energy-Efficiency Research @ZIH

1 node 100 nodes 1k nodes full HPC system

HAEC

1 Sa/s > 10k Sa/s

HDEEM
Phase 2 High Res

Energy Accounting

COOL silicon

1 Sa/s

HDEEM

TECHNISCHE UNIVERSITÄT DRESDEN

Joseph Schuchart
HAEC - Fine grained measurement of single nodes

**Apollo**: AMD Bulldozer system, 3x16 cores
- ATX (12V, 5V, 3.3V)
- Augmentation on mainboard (using Voltage Regulators)
  - Per socket: Core, NB, MEM, Temp
  - Extra regulator for all sockets
- ~10 kHz sampling rate per channel

https://tu-dresden.de/zih/haec
HAEC - Fine grained measurement of single nodes

**Artemis**: Intel Sandy Bridge system, 2x8 cores

- Instrumented using Shunts:
  - ATX (12V, 5V, 3.3V)
  - 12V per socket
  - SATA, FAN

- Riser card instrumentation:
  - PCIe (network or GPU with 2x8pin GPU plug)
  - DDR3 (initially one module)

https://tu-dresden.de/zih/haec
HAEC - Fine grained measurement of single nodes

National Instrument Measurement Hardware

- NI 6255:
  - 80 analog input channels
  - 7,5kSa/s per channel

- NI 6321:
  - 16 analog input channels
  - 500 kSa/s per channel

Custom data management server
- Input to Dataheap
- Live-Data subscription

HAEC – Fine grained measurement of single nodes

application activity

time

Power consumption measurements

HAEC – Fine grained measurement of single nodes

application activity

Power consumption measurements

Joseph Schuchart

HAEC – Fine grained measurement of single nodes

<table>
<thead>
<tr>
<th>Time</th>
<th>0.24225 s</th>
<th>+75 µs</th>
<th>+150 µs</th>
<th>+225 µs</th>
</tr>
</thead>
</table>

"artemis/s0-fast/watts" over Time

<table>
<thead>
<tr>
<th>Time</th>
<th>10.83 ms</th>
<th>+10 µs</th>
<th>+20 µs</th>
<th>+30 µs</th>
<th>+40 µs</th>
<th>+50 µs</th>
<th>+60 µs</th>
</tr>
</thead>
</table>

"apollo/s0/core/watts" over Time

Synthetic workload and the power consumption measured with 500 kSa/s. Low load (sqrt): orange, thread synchronization: light blue, High load (compute): dark blue.

What is the rest? FIRESTARTER!

- Goal: Maximize power consumption to stress system components
- Prime95
  - Default configuration performs computation with different parameters
- Linpack
  - No constant, long-running computation
- FIRESTARTER:
  - Download, run, burn 😊
  - http://tu-dresden.de/zh/firestarter/


What is the future?

- Final calibration of HDEEM board measurement
- First projects using HDEEM (READEX, Prof. Chapman’s group)
- Implementation of plugins for the Sandia PowerAPI

We are always looking for interesting projects to use the HDEEM infrastructure!

Questions? Please feel free to contact energy-group@fusionforge.zih.tu-dresden.de
Thank you!
References

- Thomas Ilsche, Daniel Hackenberg, Stefan Graul, Robert Schöne, Joseph Schuchart: *Power Measurements of Compute Nodes: Improving Sampling Rates, Granularity, and Accuracy*, 2015, accepted for publication.


