Mitglied der Helmholtz-Gemeinschaft

JÜLICH FORSCHUNGSZENTRUM

Mapping Fine-Grained Power Measurements to HPC Application Runtime Characteristics on IBM POWER7

September 2nd, 2013 | Michael Knobloch

The Exascale Innovation Center

Exa Scale
Innovation Center

EIC - Exascale Innovation Center

- Project partners: IBM Germany R&D and JSC
- Goal: Co-Design for next-gen of Supercomputers
- One work-package on power and energy-efficiency
  - Investigation of power consumption on Blue Gene (EnA-HPC’11)
  - Fine-grained power measurements on POWER7 (this work)
  - Energy modelling on POWER7 (to be published)

Test system – IBM Power 720

- 4-Core 3.0 GHz processor (Pseries, 8202-E4B)
  - 96 GFLOPS peak
  - 4 SMT threads per core
  - 64 kB L1 cache per core
  - 256 kB L2 cache per core
  - 16 MB L3 cache (shared)

- 16 GB memory, 2x 300 GB 10K RPM SAS disk
- **TPMD (Thermal Power Management Device)**
- **External power distribution unit (Raritan DPXS12A-16)**
  - 3 s measurement interval, 1 W resolution

Amester

Amester

IBM Automated Measurement of Systems for Temperature and Energy Reporting software.

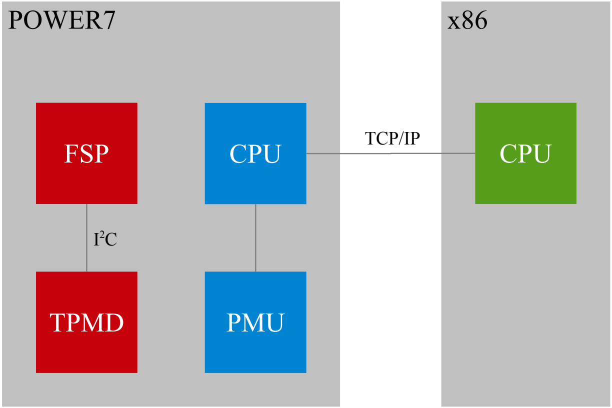
- Tool for monitoring and controlling power consumption of (IBM) servers – x86 and POWER
- Developed by Charles Lefurgy, IBM Research, Autisn, TX
- Histograms, traces for any sensor
- Scripting
  - Tcl command line
  - Send any IPMI command to measured system (ipmicmd)
  - On-line (50 ms interval) and off-line (buffered, 16 MB, 1 ms sampling) modes

Amester (condt.)

- Sensor data collection
  - Whole system power data collection (CPU, Memory, Fans, IO, Storage)
  - CPU temperature, processor speed, system utilization, instructions per second, memory bandwidth

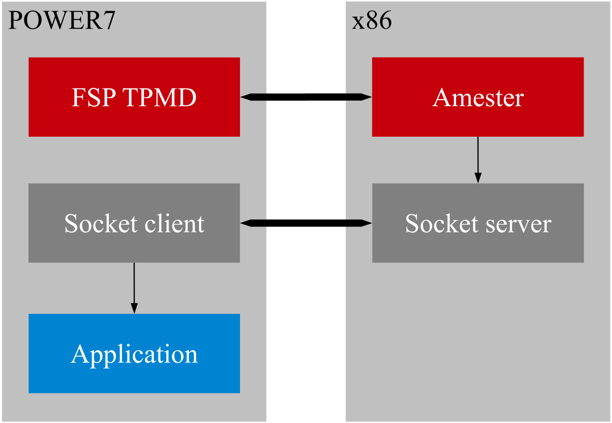
Sensor name	Units	Time scale	Description
PWR1MS	W	Instantaneous	Node power consumption
PWR1MSP0	W	Instantaneous	Processor power consumption
PWR1MSMEM0	W	Instantaneous	Memory power consumption
PWR32MS	W	avg. over last 32 ms	Node power consumption
PWR32MSP0	W	avg. over last 32 ms	Processor power consumption
PWR32MSMEM0	W	avg. over last 32 ms	Memory power consumption
IPS32MS	Mips	Every 32 ms	Instructions per second rate

Hardware Setup



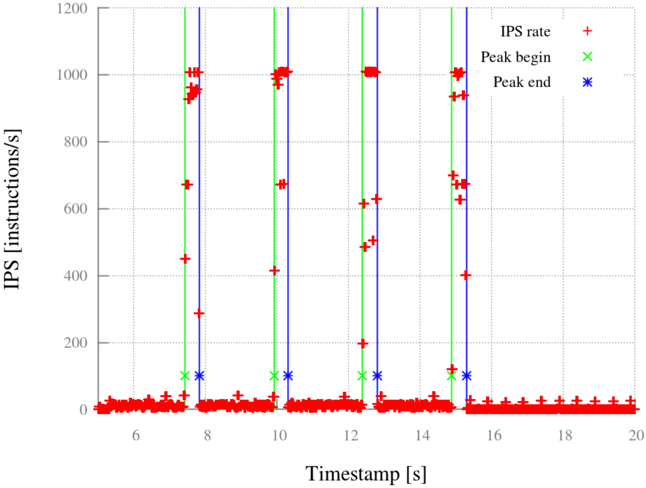
🟦 POWER7 CPU time domain 🟥 POWER7 FSP time domain

Software Setup

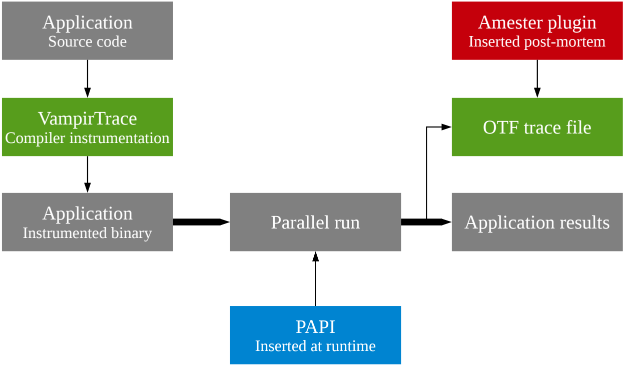


🟦 POWER7 CPU time domain 🟥 POWER7 FSP time domain

Time-stamp synchronization



VampirTrace Workflow



Application
Source code

VampirTrace
Compiler instrumentation

Application
Instrumented binary

Parallel run

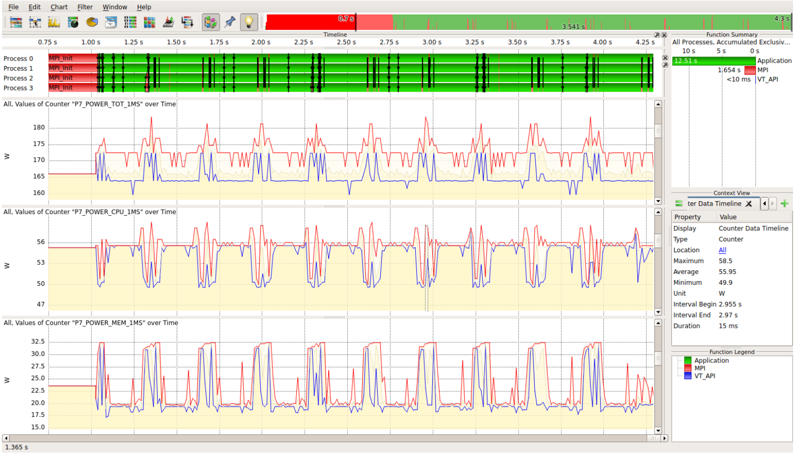
PAPI
Inserted at runtime

Amester plugin
Inserted post-mortem

OTF trace file

Application results

PEPC: Full run (without initialization)

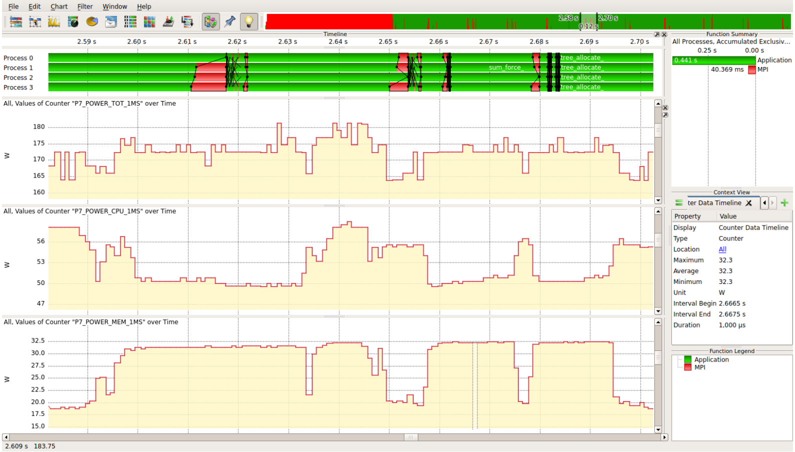


Property	Value
Display	Counter Data Timeline
Type	Counter
Location	All
Maximum	58.5
Average	55.95
Minimum	49.9
Unit	W
Interval Begin	2.955 s
Interval End	2.97 s
Duration	15 ms

Function Legend

- 🟩 Application
- 🟥 MPI
- 🟦 VT_API

PEPC: 1 Iteration

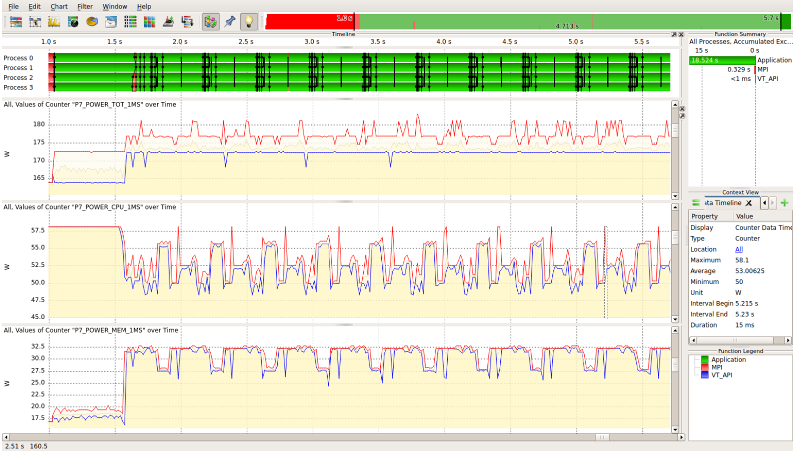


Property	Value
Display	Counter Data Timeline
Type	Counter
Location	All
Maximum	32.3
Average	32.3
Minimum	32.3
Unit	W
Interval Begin	2.6665 s
Interval End	2.6675 s
Duration	1,000 µs

Function Legend

- 🟩 Application
- 🟥 MPI

MP2C: Full run (without initialization)



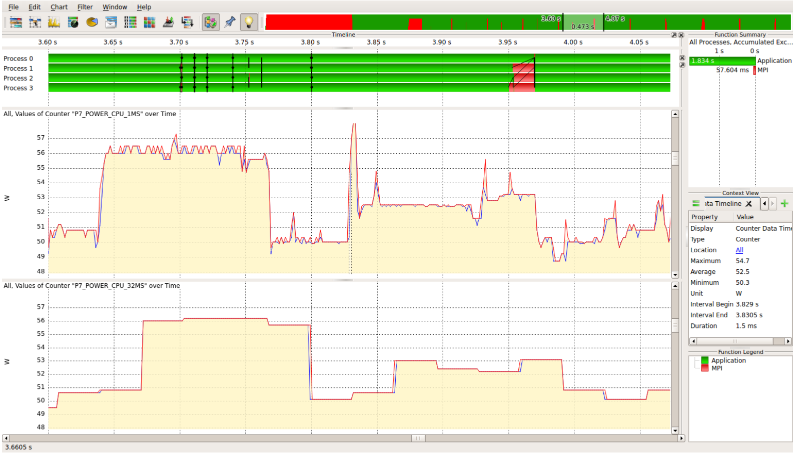
All Processes, Accumulated Exc...

15 s	0 s
18.524 s	Application
0.329 s	MPI
<1 ms	VT_API

Property	Value
Display	Counter Data Time
Type	Counter
Location	All
Maximum	58.1
Average	53.00625
Minimum	50
Unit	W
Interval Begin	5.215 s
Interval End	5.23 s
Duration	15 ms

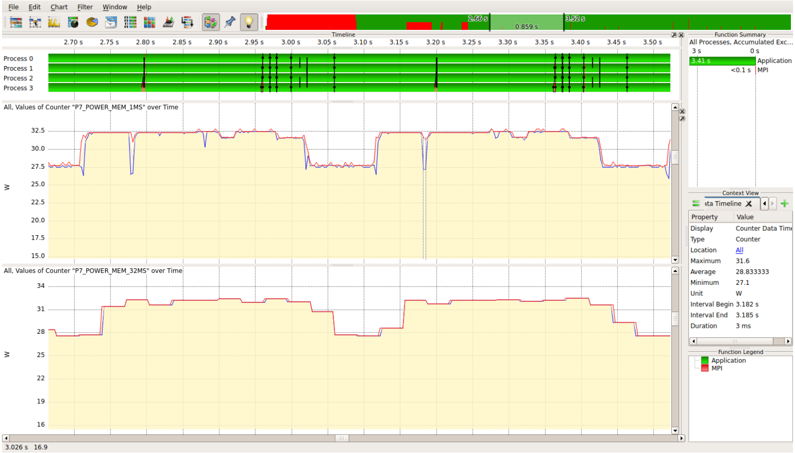
- 🟩 Application
- 🟥 MPI
- 🟦 VT_API

MP2C: CPU Counter Resolution Comparison


File Edit Chart Filter Window Help
3.60 s
0.473 s
4.07 s
Timeline
3.60 s 3.65 s 3.70 s 3.75 s 3.80 s 3.85 s 3.90 s 3.95 s 4.00 s 4.05 s
Process 0
Process 1
Process 2
Process 3
All, Values of Counter "P7_POWER_CPU_1MS" over Time
W
57 56 55 54 53 52 51 50 49 48
All, Values of Counter "P7_POWER_CPU_32MS" over Time
W
57 56 55 54 53 52 51 50 49 48
3.6605 s
Function Summary
All Processes, Accumulated Exc...
1 s 0 s
1.834 s Application
57.604 ms MPI
Context View
ata Timeline
Function Legend
Application
MPI

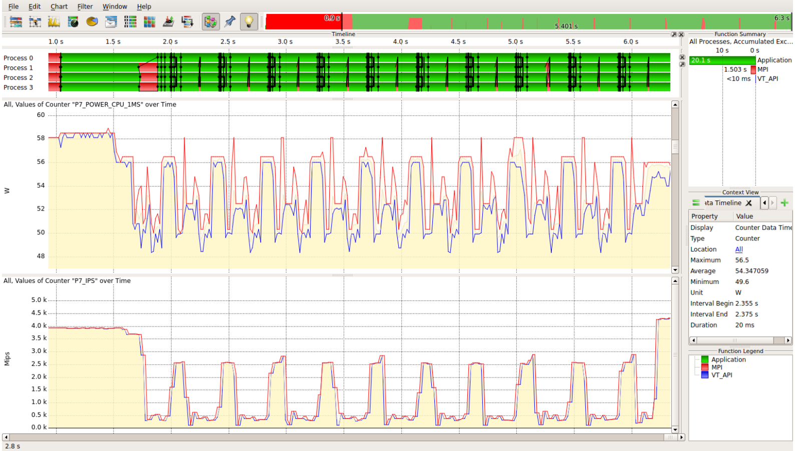
Property	Value
Display	Counter Data Time
Type	Counter
Location	All
Maximum	54.7
Average	52.5
Minimum	50.3
Unit	W
Interval Begin	3.829 s
Interval End	3.8305 s
Duration	1.5 ms

MP2C: MEM Counter Resolution Comparison



Property	Value
Display	Counter Data Tim
Type	Counter
Location	All
Maximum	31.6
Average	28.833333
Minimum	27.1
Unit	W
Interval Begin	3.182 s
Interval End	3.185 s
Duration	3 ms

MP2C: CPU Power and IPS

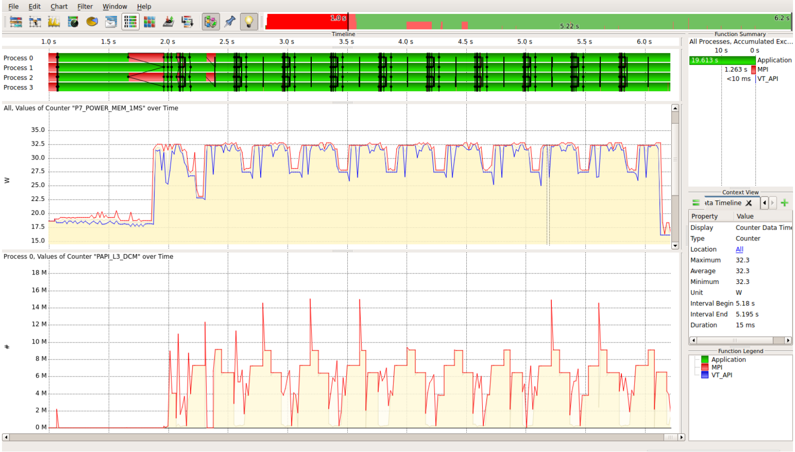


Property	Value
Display	Counter Data Time
Type	Counter
Location	All
Maximum	56.5
Average	54.347059
Minimum	49.6
Unit	W
Interval Begin	2.355 s
Interval End	2.375 s
Duration	20 ms

Function Legend

- 🟩 Application
- 🟥 MPI
- 🟦 VT_API

MP2C: MEM Power and L3 Data Cache Misses



Property	Value
Display	Counter Data Time
Type	Counter
Location	All
Maximum	32.3
Average	32.3
Minimum	32.3
Unit	W
Interval Begin	5.18 s
Interval End	5.195 s
Duration	15 ms

Outlook

Conclusions

- Fine-grained power measurements help to better understand application power consumption
- Amester requires complicated setup
- Mapping to other metrics can be difficult due to timing issues
- **Correlation does not imply causation**

Future Work

- Energy modelling (to be published)
- Integration in Score-P and Scalasca