LEIBNIZ RECHEN ZENTRUM

Autotuning the Energy Consumption

{carmen.navarrete, carla.guillen, wolfram.hesse,
matthias.brehm, david.brayford}@lrz.de

Overview

- Processors operating at lower clock speed consume proportionately less power and generate less heat.
- Dynamic scaling of the clock speed gives some control in power consumption, when not operating at full capacity.
- Lower processor frequency does not necessarily reduce energy consumption (application will take longer).

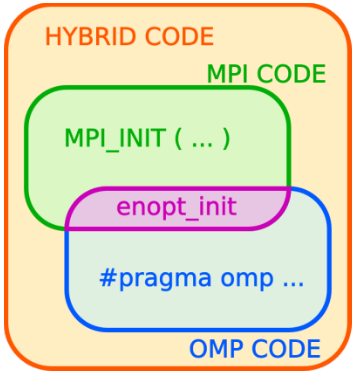
Overview

5 Governors :

- Static (no thresholds)
  - Peformance : Max(frequency).
  - Powersave : Min(frequency).
  - Userspace : User defined frequency.
- Dynamic:
  - Ondemand : Single threshold increases and decreases the frequency step size.
  - Conservative : Dual threshold (up & down frequency) reduces the possibility of oscillation between frequency steps.

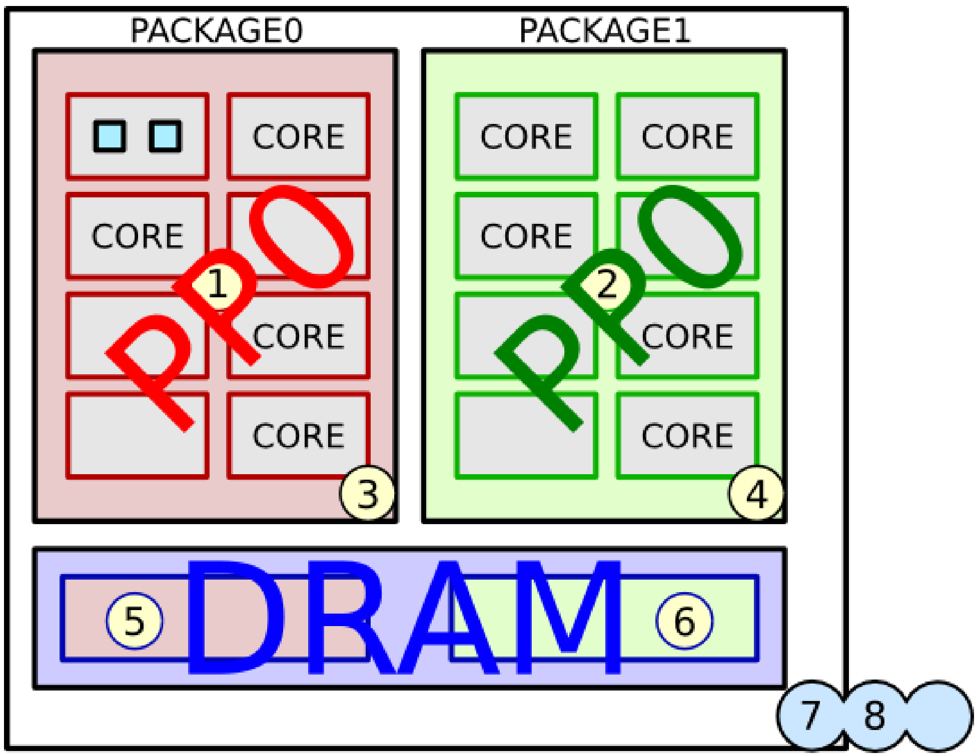
Features

- Written in C++
- Bindings for C and Fortran codes
- Support for:
  - Parallel codes: MPI, OpenMP and Hybrid.
  - Sequential codes.
- Socket and node level counter measurements.
- Compatible with PAPI v4 and PAPI v5 headers
- Provides accesses to kernel mode operations:
  - Changing CPUFreq infrastructure parameters.
  - Accesses to the MSR devices.



SandyBridge microarchitecture

- SandyBridge sensors



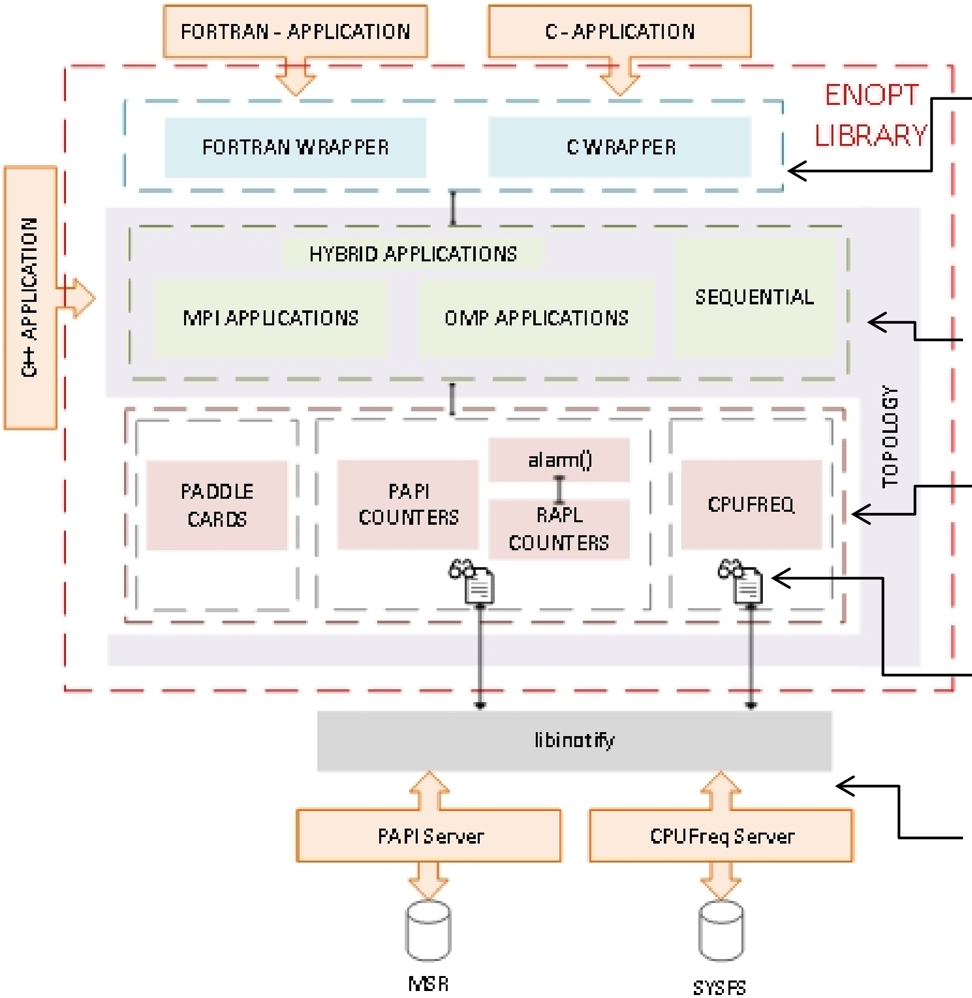
1-6: RAPL (Running Average Power Limit) Counters.

7-8: IBM AEM (Advanced Energy Management) Kernel Module.

1 & 2: Energy of the 8 cores.
3 & 4: Energy of the complete Package (core + uncore).
5 & 6: DRAM Energy.

7: DC Counter.
8: AC Counter.

Components



Allow access the library kernel from different languages.

Discover of processes topology: register and handshake.Election of the master process per node: node level counters.

Counter layer: Interface for counter commands

Communication between server and library done through a special file.

Communication with the Linux kernel subsystem

Metrics

- PAPI - RAPL
  - PAPI_TOT_CYC
  - PAPI_TOT_INS
  - PAPI_L3_TCM
  - PACKAGE_ENERGY:PACKAGEx
  - PP0_ENERGY:PACKAGEx
  - DRAM_ENERGY:PACKAGEx
- Paddle Card (HWMON kernel driver)
  - AC Counter
  - DC Counter
- Time metrics
- Future: other counters: temperature, fan, network counters…

Validation

- Comparison of measurements with three external tools.

Tool	DRAM	SOCKET	NODE	RACK
LIKWID	X	X		
PAPI-RAPL	X	X		
PaddleCard			X	
PDU				X

Tool	Technology	Resolution
LIKWID	MSR	1ms
PAPI-RAPL	MSR	1ms
PaddleCard	Ibmaem-HWMON	300 ms
PDU	Power meter	1 min

Validation

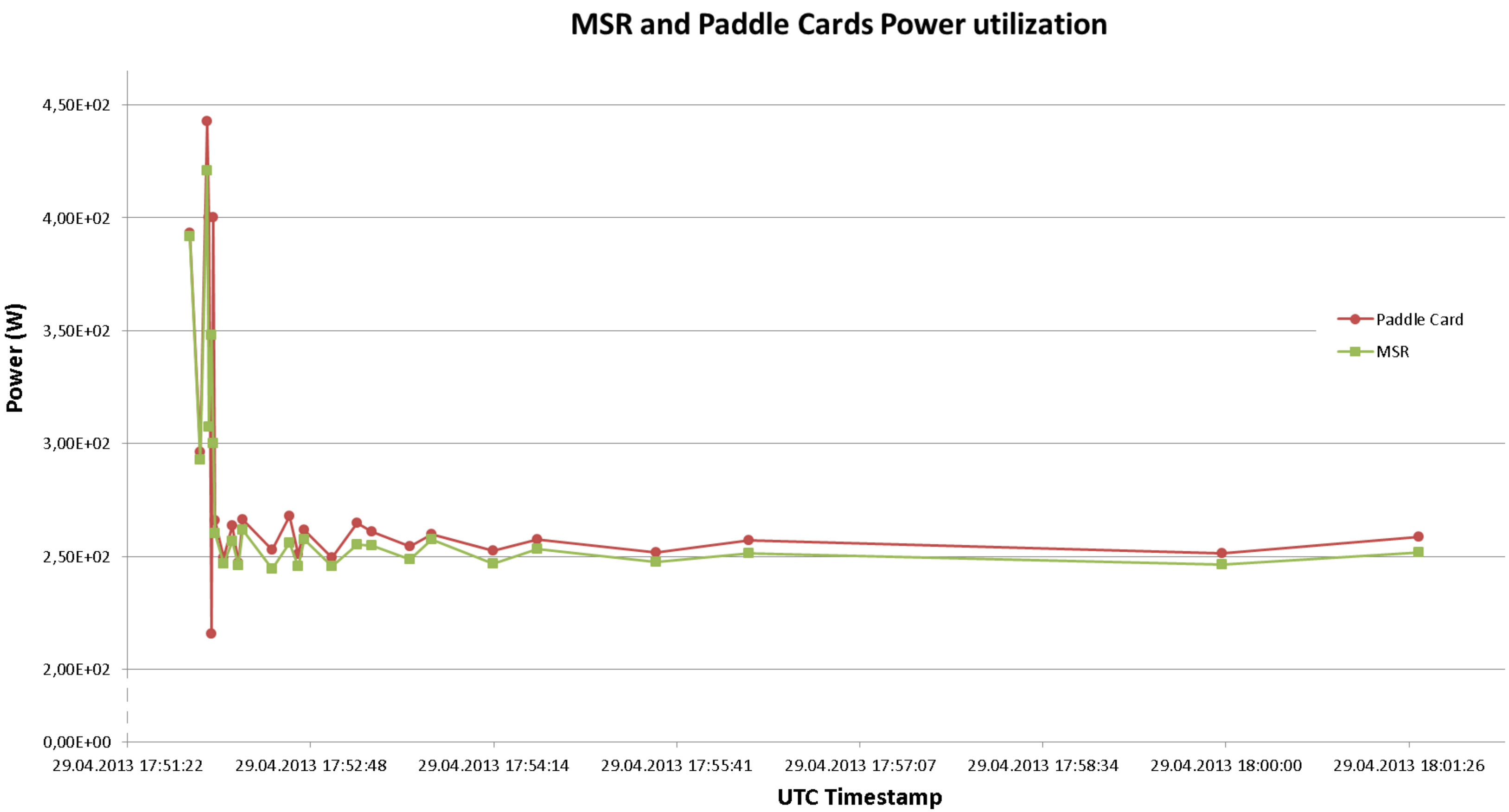
- sleep(10) command

	LIKWID	**RAPL**	**IBMAEM**	IBMAEM
PKG0	105 J	**103 J**	**-**	-
PKG1	-	**104 J**	**-**	-
DRAM0	25 J	**25.5 J**	**-**	-
DRAM1	-	**25.5 J**	**-**	-
DC	-	**-**	**491 J**	449 J
E/Node	260 J	**257 J**	**245.5 J**	224.5 J

Bayerische Akademie der Wissenschaften

Validation

- MSR and Paddle Card comparison



Bayerische Akademie der Wissenschaften

Tests and Results

- APEX-MAP benchmark
  - Generates artificial calculations and memory accesses for measurement purposes.
  - Assumes that performance behavior of scientific apps can be modeled by a set of specific performance factors.
  - Simulate compute and memory bound applications.

  - Developed by the Laurence Berkeley National Laboratory
  - Specific performance factors: memory bandwidth and FLOPS

Plugin for the Energy Consumption via CPUFreq

- Aim
  - Optimize the energy consumption of an arbitrary application, by choosing the best combination of frequencies for each code region.
- Integration with periscope
  - The start of each code region calls (per callback) the corresponding library function to change:
    - The CPU governor
    - The CPU frequency
  - The code is executed for each combination of frequencies and governors, looking for the minimum energy consumption.

Energy model

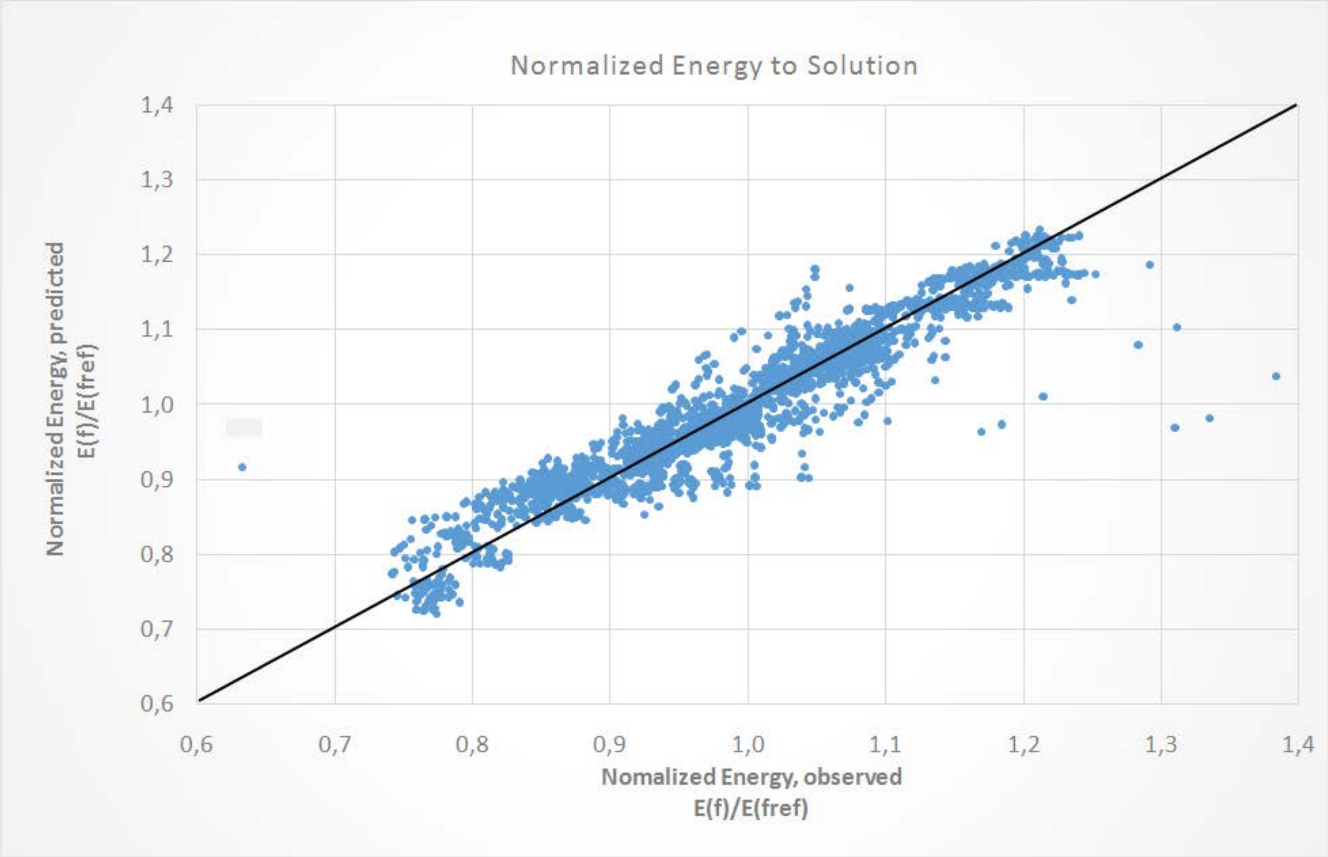
- Used by loadleveler to minimize the energy-to-solution

$$PWR_{Fn} = PWR_{F0} * func_0(CPI_{F0}, L2_{F0}, L3_{F0}, GIPS_{F0}, GBS_{F0}, \ldots)$$
$$T_{Fn} = T_{F0} * func_1(CPI_{F0}, L2_{F0}, L3_{F0}, GIPS_{F0}, GBS_{F0}, \ldots)$$

  – CPI, L2, L3, GIPS, GBS… are measured at nominal frequency F0.
  – Coefficients of functions are measured for the given platform at all possible frequencies.
  – Hides the dependency of GIPS and GBS of a given clock frequency

Energy model

- Predicted energy use (model) vs Observed energy used (measured)



Bayerische Akademie der Wissenschaften

LEIBNIZ RECHEN ZENTRUM

Your turn!

Questions?