VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

HPC HW POWER MONITORING

CASTIEL 2

Ondřej Vysocký
IT4Innovations

19.1. 2024

EURO CZECHIA

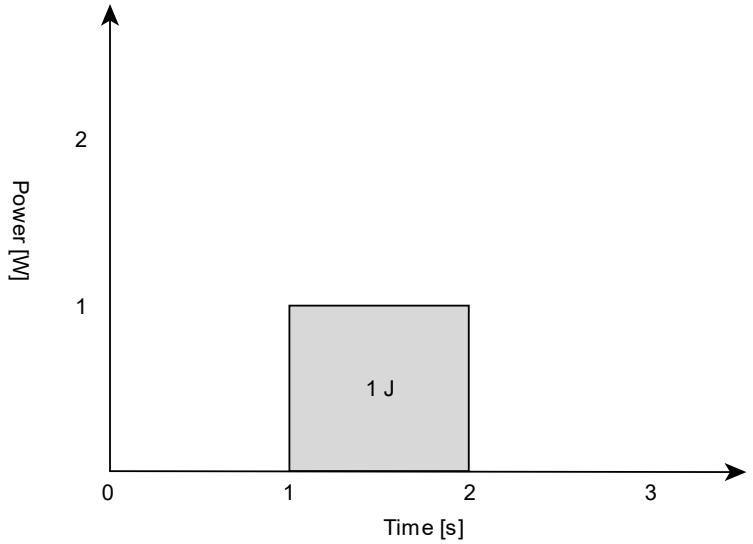
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

ENERGY

$$Energy = Power \times Time$$

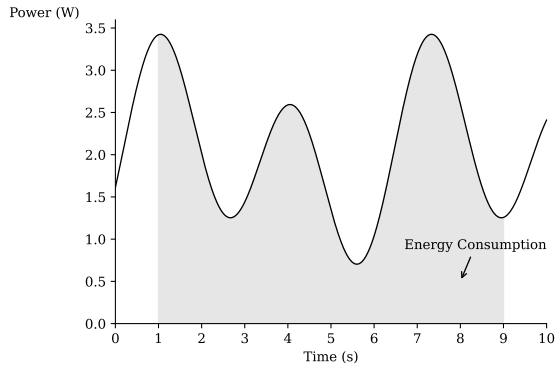
- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J



ENERGY

$$Energy = Power \times Time$$

- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J

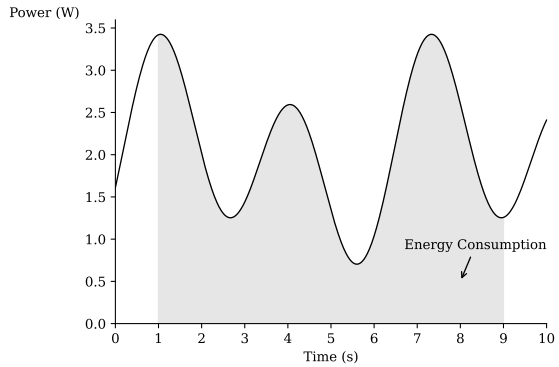


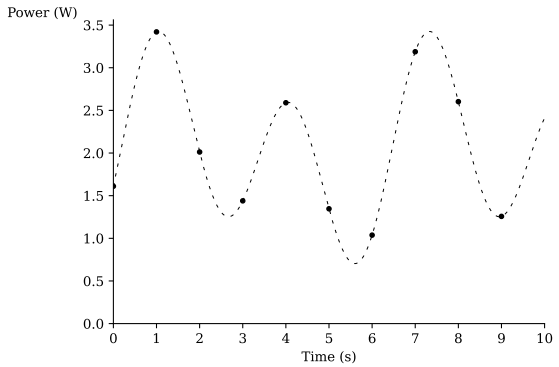
Img source, Luís Cruz (TU Delft)

ENERGY

$$Energy = Power \times Time$$

- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J



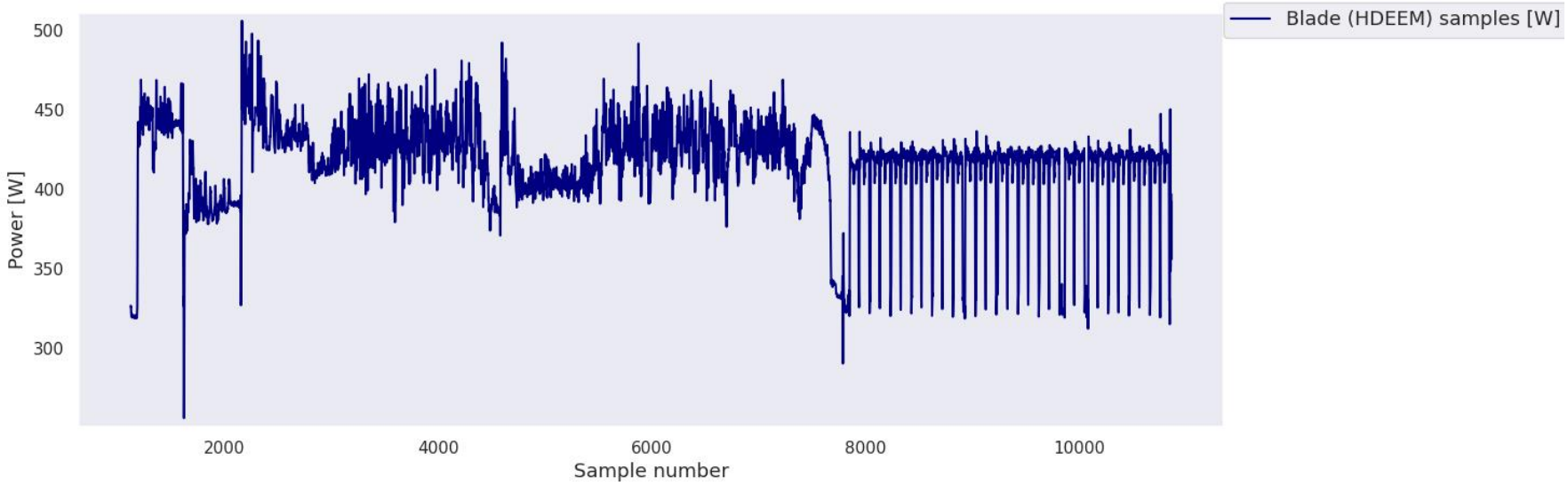


Img source, Luís Cruz (TU Delft)

ENERGY

$$Energy = Power \times Time$$

- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J

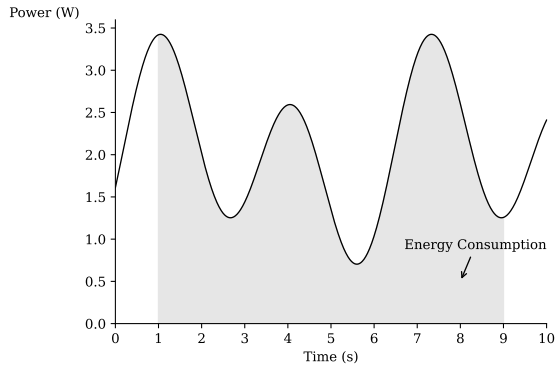


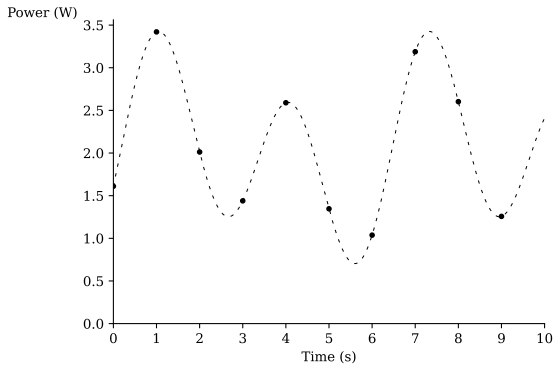
ENERGY

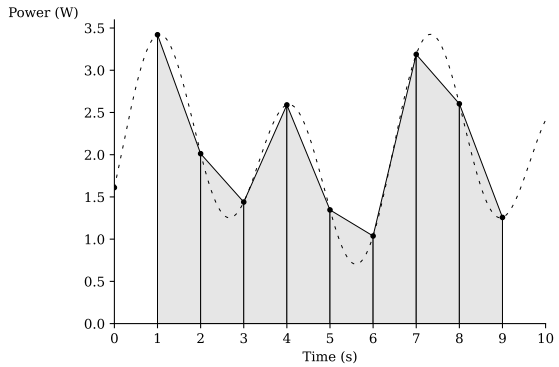
$$Energy = Power \times Time$$

- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J

$$Energy(t) = \int_0^t Power(x)\,\mathrm{d}x \approx \frac{\sum_{i=0}^{n} PowerSample_i}{SamplingFrequency}$$



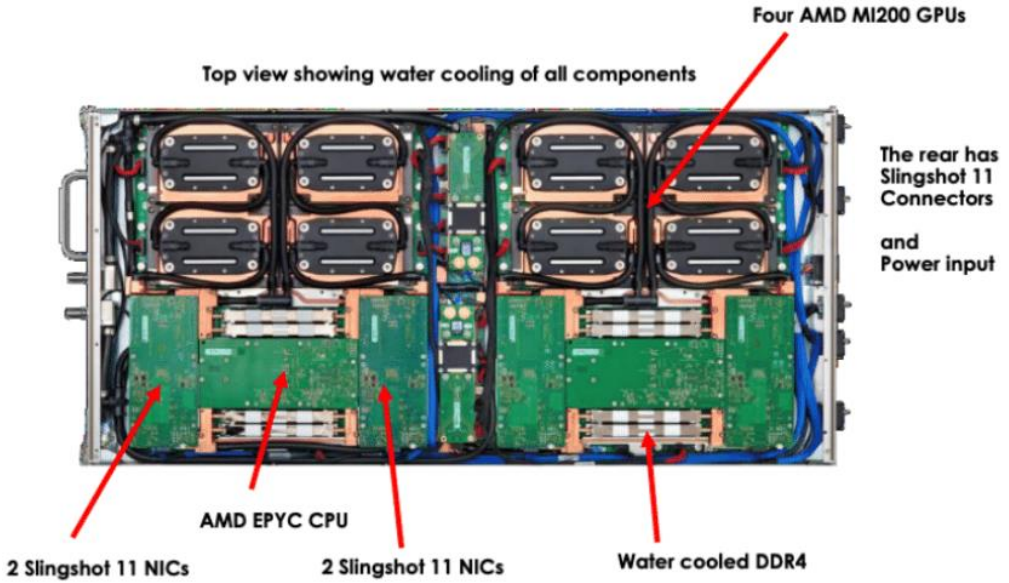


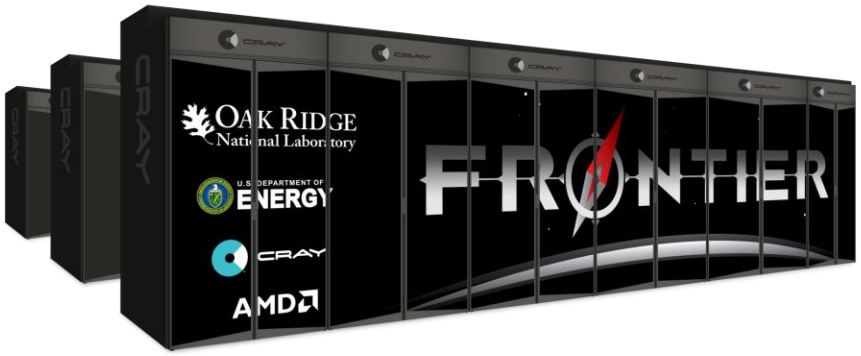


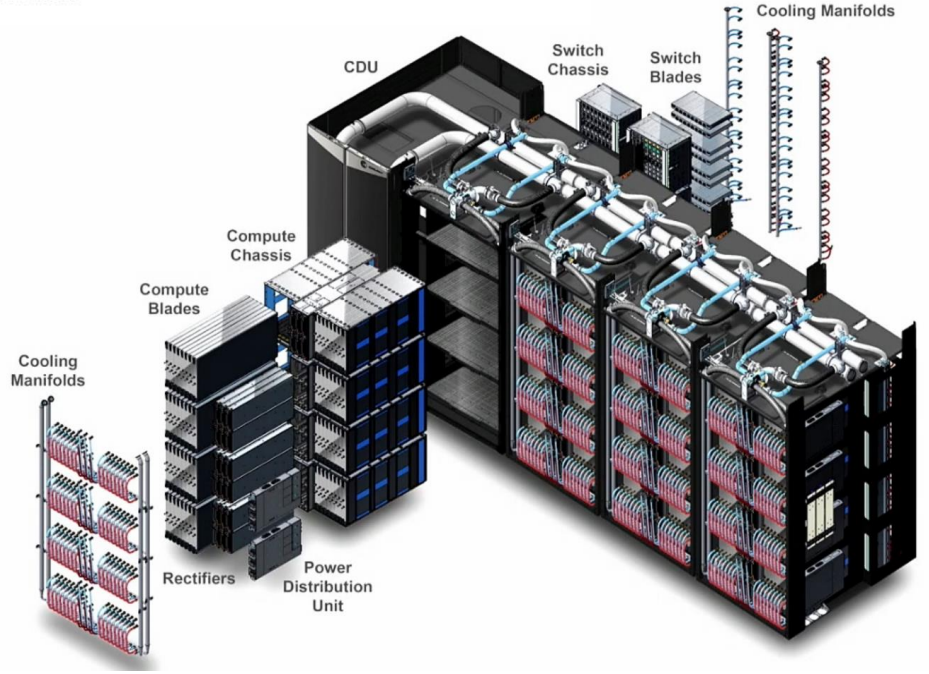
Img source, Luís Cruz (TU Delft)

POWER MONITORING

- On node components
  - CPU, ACC, memories, NIC
- Node
- Rack
- System
- Data hall
- Building



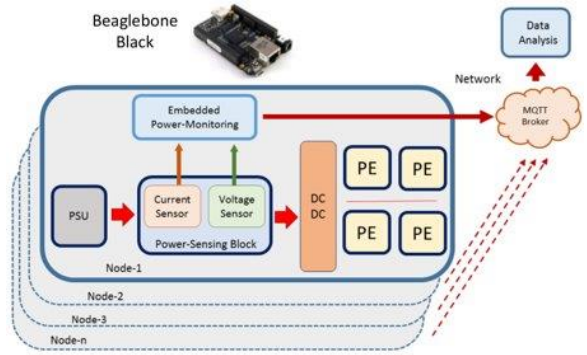


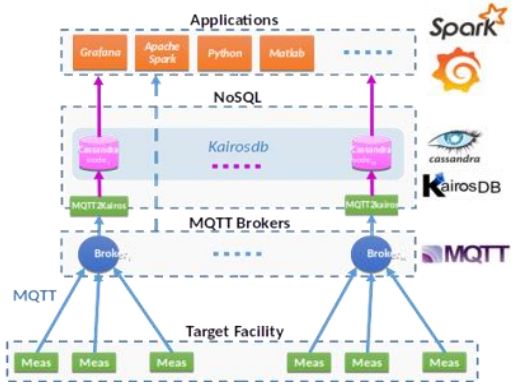


Images of HPE Cray, OLCF Frontier

POWER MONITORING SYSTEMS FOR HPC

name	original purpose	out/in band	sampling rate	sensors
ADEPT	energy measurement	out	1 MHz	blade, CPUs, DRAMs, ACC, HDD, NIC
DiG	anomaly monitoring	out	50 kHz	blade
HDEEM	energy measurement	out	1 kHz / 100 Hz	blade, CPUs, DRAMs, NIC*, VAUX*
NVML	power management	in	<66.7 Hz [82]	GPU
OCC	power management	in[2]	4 kHz	blade
PI	energy measurement	both	1 kHz	CPUs, DRAMs, ACC
PM2	energy measurement	out	1 kHz / 3 kHz	8 sensors
RAPL	power management	in	1 kHz	Package, DRAM*, PP0*, PP1*, Platform*



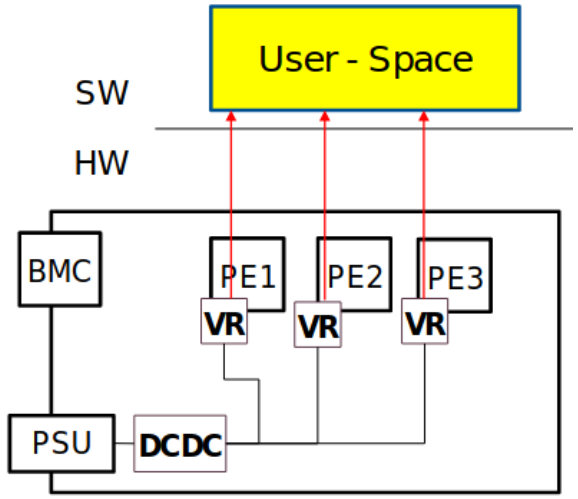


$$Energy(t) = \int_0^t Power(x)\,\mathrm{d}x \approx \frac{\sum_{i=0}^{n} PowerSample_i}{SamplingFrequency}$$

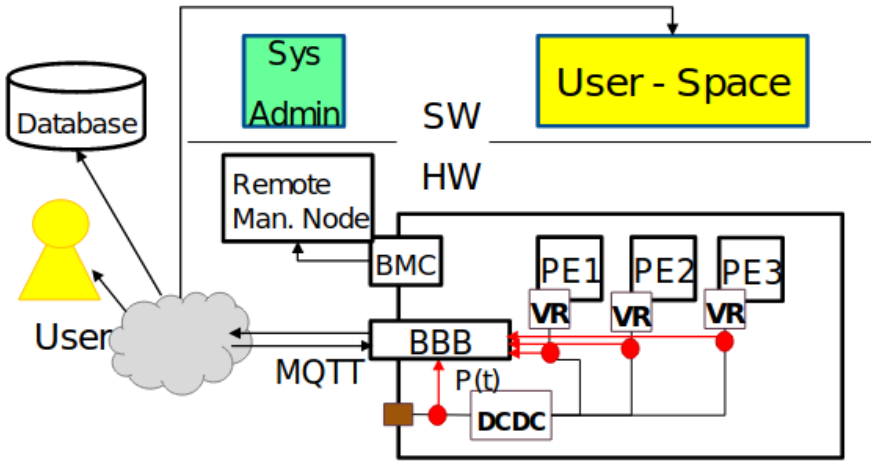
IN- AND OUT-OF-BAND POWER MONITORING

- In-band
  - Vendor specific
  - HW performance counters
- Out-of-band
  - No overhead of monitoring, high overhead of exposing to user-space
  - Custom sensors

RAPL



DiG



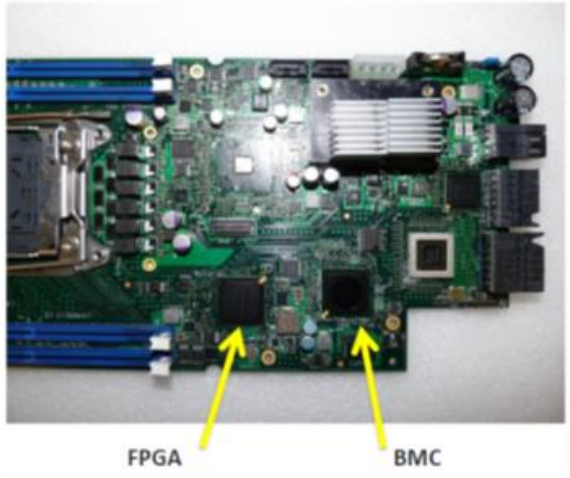
Img source, Antonio Libri (UNIBO)

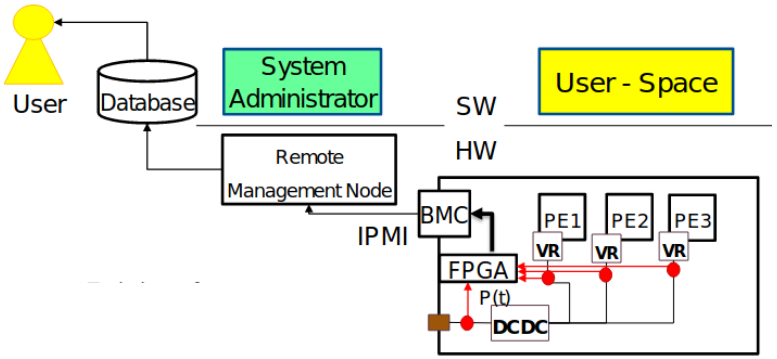
HIGH DEFINITION ENERGY EFFICIENCY MONITORING (HDEEM)

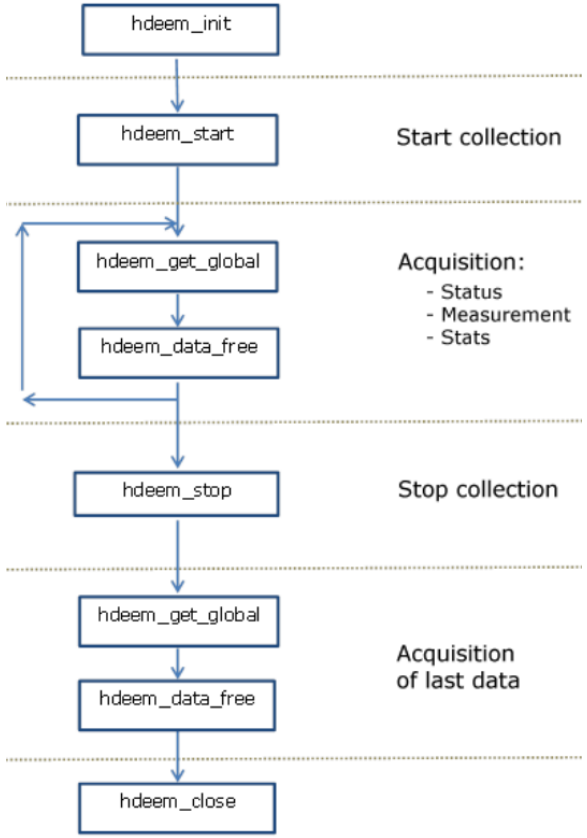
- Bull|Atos technology available for production systems (Bullx B7xx and Bull Sequana)
- On board out-of-band technology for power monitoring

- Power domains:
  - Blade (1kHz)
  - VRs (100 Hz) CPUs, DRAMs, NIC*, VAUX*

- 2% of accuracy uncertainty
- C library as well as command line utility:
  - startHdeem
  - stopHdeem
  - checkHdeem
  - printHdeem
  - clearHdeem



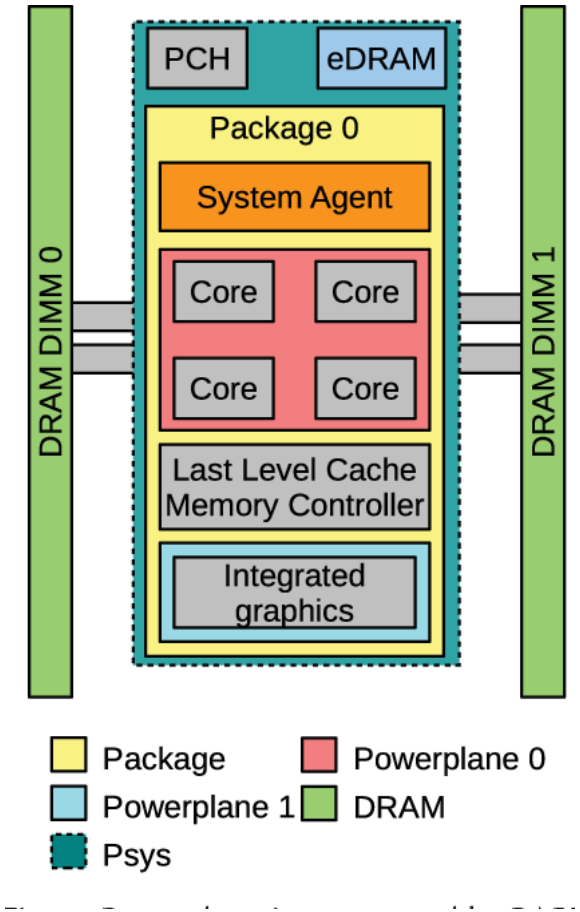




Img source, Antonio Libri (UNIBO)

INTEL RUNNING AVERAGE POWER LIMIT (RAPL)

- Sysfs: /sys/devices/virtual/powercap/intel-rapl/intel-rapl:X/intel-rapl:0:Y
- Power domains:
  - **Package**: limits the power consumption for the entire package of the CPU, this includes cores and uncore components
    - short (1.2 * TDP, ~ milliseconds) and long window (TDP, ~ second)
  - **DRAM**: is used to power cap the DRAM memory = memory monitoring, P-State scaling
    - only for server architectures, no client
    - single time window
    - in default is turned off
  - **PP0/Core**: is used to restrict the power limit only to the cores of the CPU
    - no new server
    - single time window
  - **PP1/Graphic**: is used to power limit only the graphic component of the CPU
    - no server
    - single time window
  - **PSys/Platform**: controls entire System on Chip
    - short and long window
    - available from Skylake architecture
    - requires support from vendor

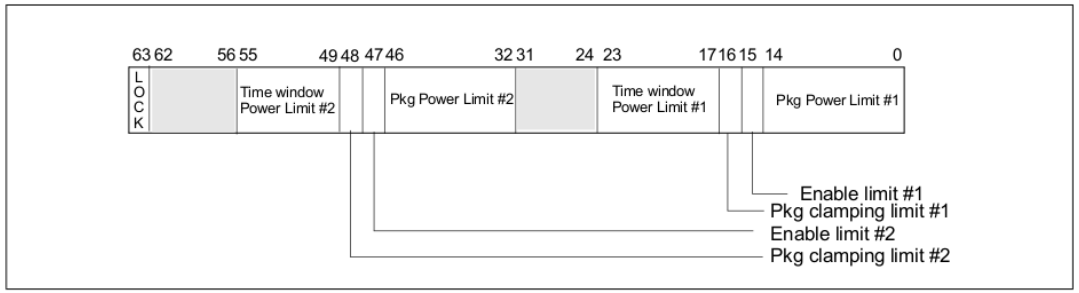


Domain	Machine Specific Register	Address
Package	MSR_PKG_POWER_LIMIT	0x610
DRAM	MSR_DRAM_POWER_LIMIT	0x618
PP0	MSR_PP0_POWER_LIMIT	0x638
PP1	MSR_PP1_POWER_LIMIT	0x640
Platform	MSR_PLATFORM_POWER_LIMIT	0x65C

Img src Khan, K. et al. "RAPL in Action."

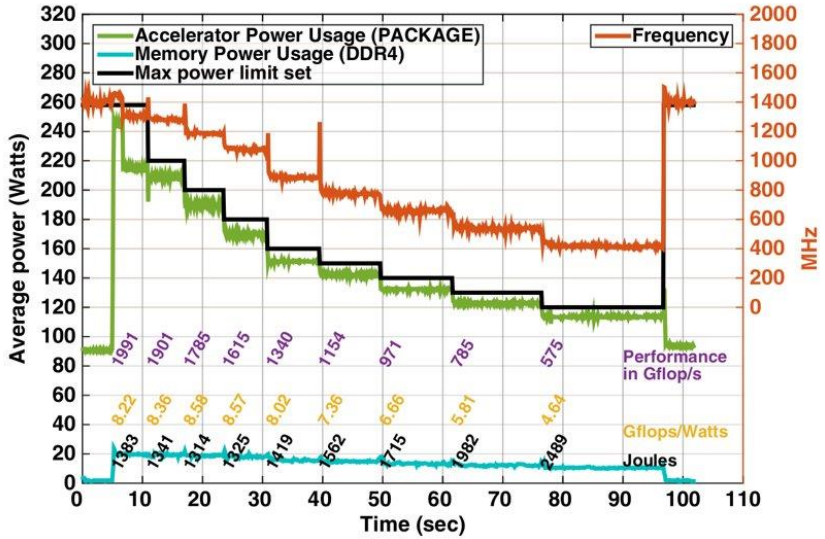
INTEL RUNNING AVERAGE POWER LIMIT (RAPL)

- MSR MSR_PKG_POWER_LIMIT (0x610)

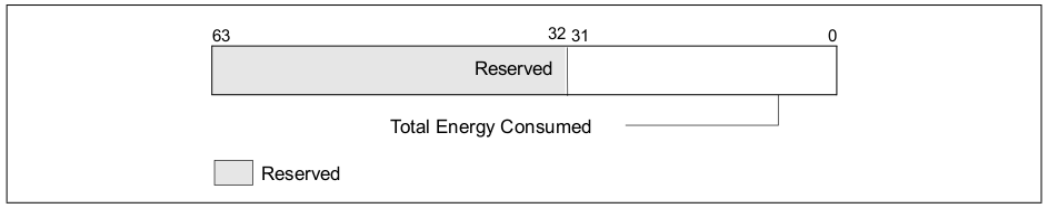


- MSR MSR_RAPL_POWER_UNIT (0x606)
  - Power units
  - Energy status units
  - Time units

- **Energy consumption measurement**
  - MSR MSR_PKG_ENERGY_STATUS (0x611)
  - MSR MSR_DRAM_ENERGY_STATUS (0x619)
  - MSR MSR_PP0_ENERGY_STATUS (0x639)
  - MSR MSR_PP1_ENERGY_STATUS (0x641)
  - MSR MSR_PLATFORM_ENERGY_COUNTER (0x64D)



Haidar et al: Investigating power capping toward energy-efficient scientific applications



AMD RAPL

- The same interface as Intel RAPL
- For energy consumption report, not used for power capping
- Power domains:
  - **Package**:
    - PKG domain includes of cores and also IO die and other in socket components [1]
  - **Core**:
    - Each CPU core separately

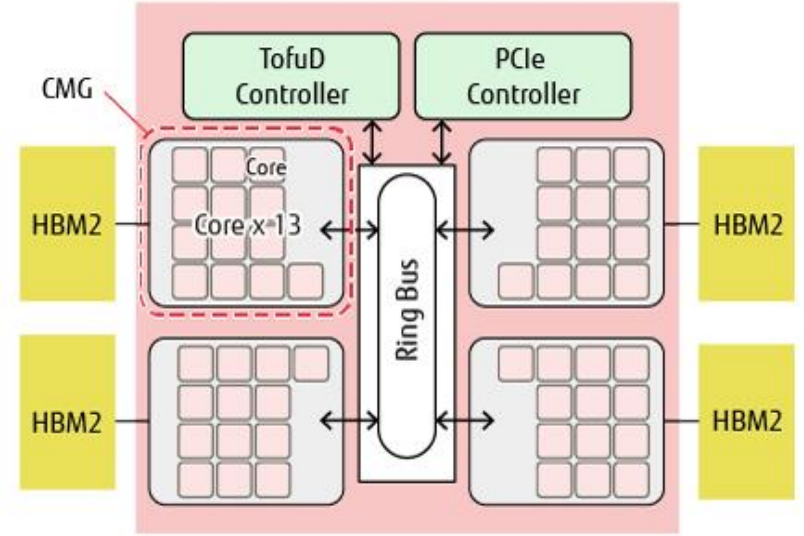
Power domain	Model specific register	Address
Core	Core::X86::Msr::CORE_ENERGY_STAT	0xC001029A
Package	Core::X86::Msr::PKG_ENERGY_STAT	0xC001029B
Energy unit		
RAPL power unit	Core::X86::Msr::RAPL_PWR_UNIT	0xC0010299

[1] https://github.com/amd/amd_energy/issues/1

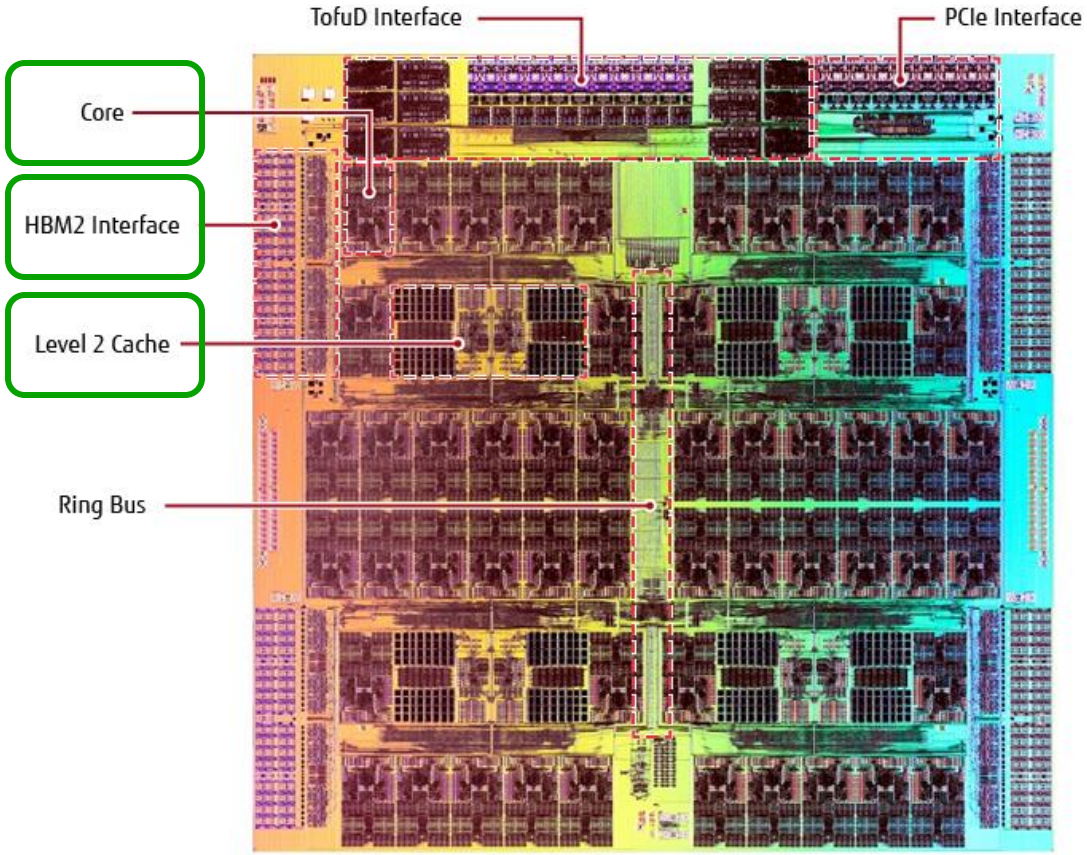
FUJITSU A64FX

- Power domains:
  - **Core**
  - **Memory – Core Memory Group (CMG) local memory**
  - **L2 cache (LLC)**
- Sampling frequency:
  - every cycle of the domain

No energy consumption reported if the HW is idle

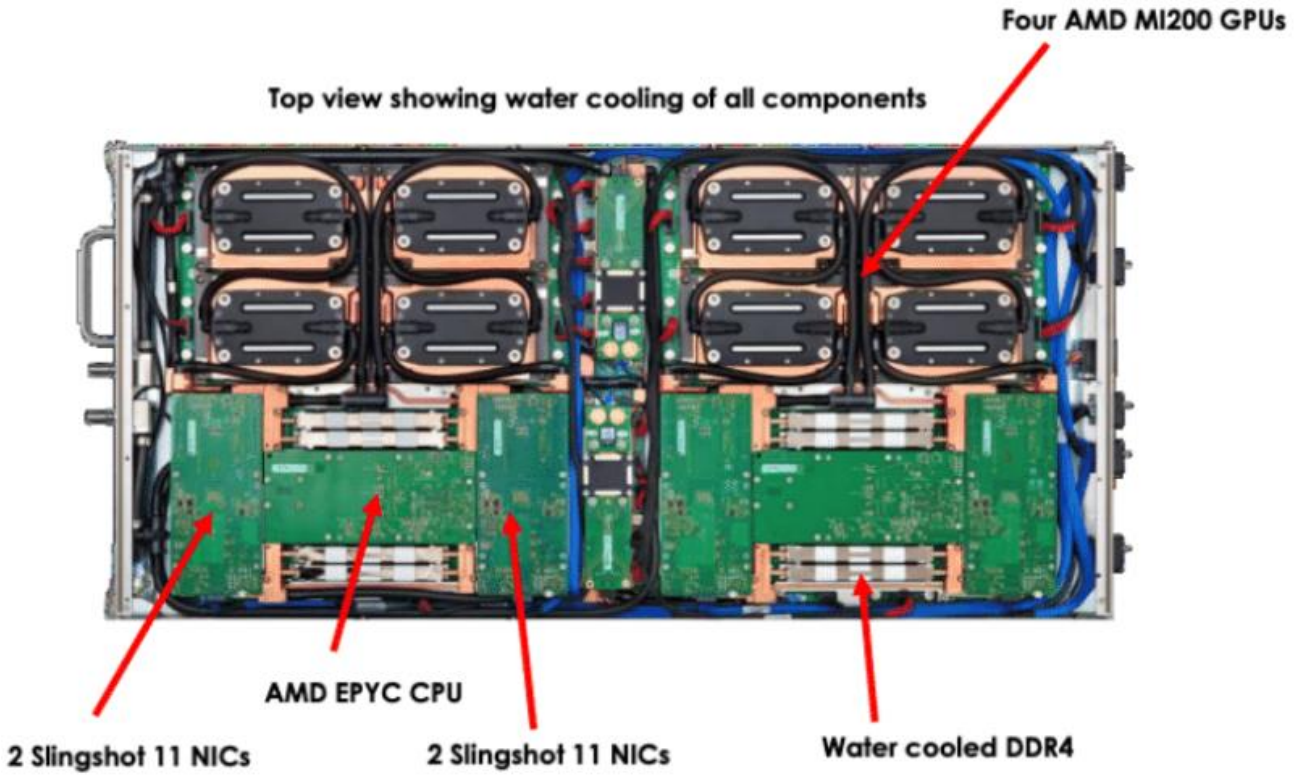


HBM2: High Bandwidth Memory 2



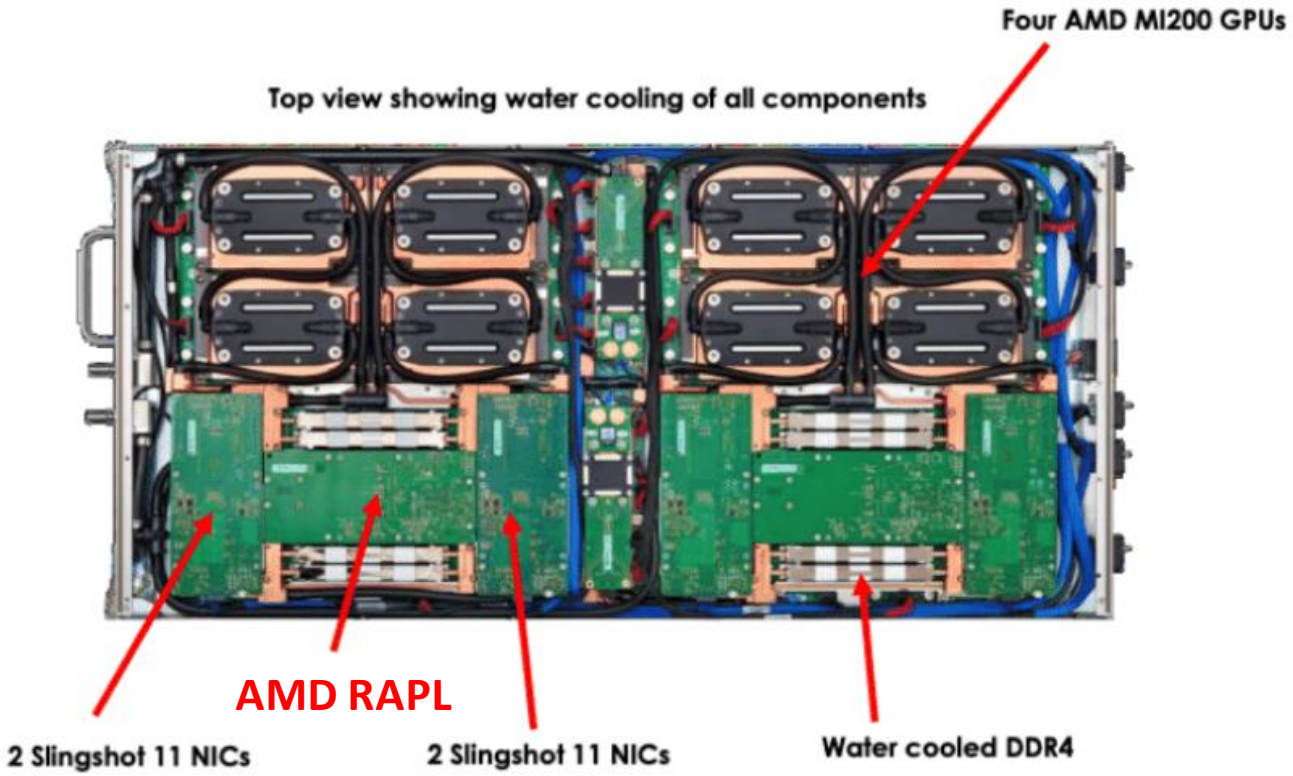
Img source, Fujitsu

POWER MONITORING



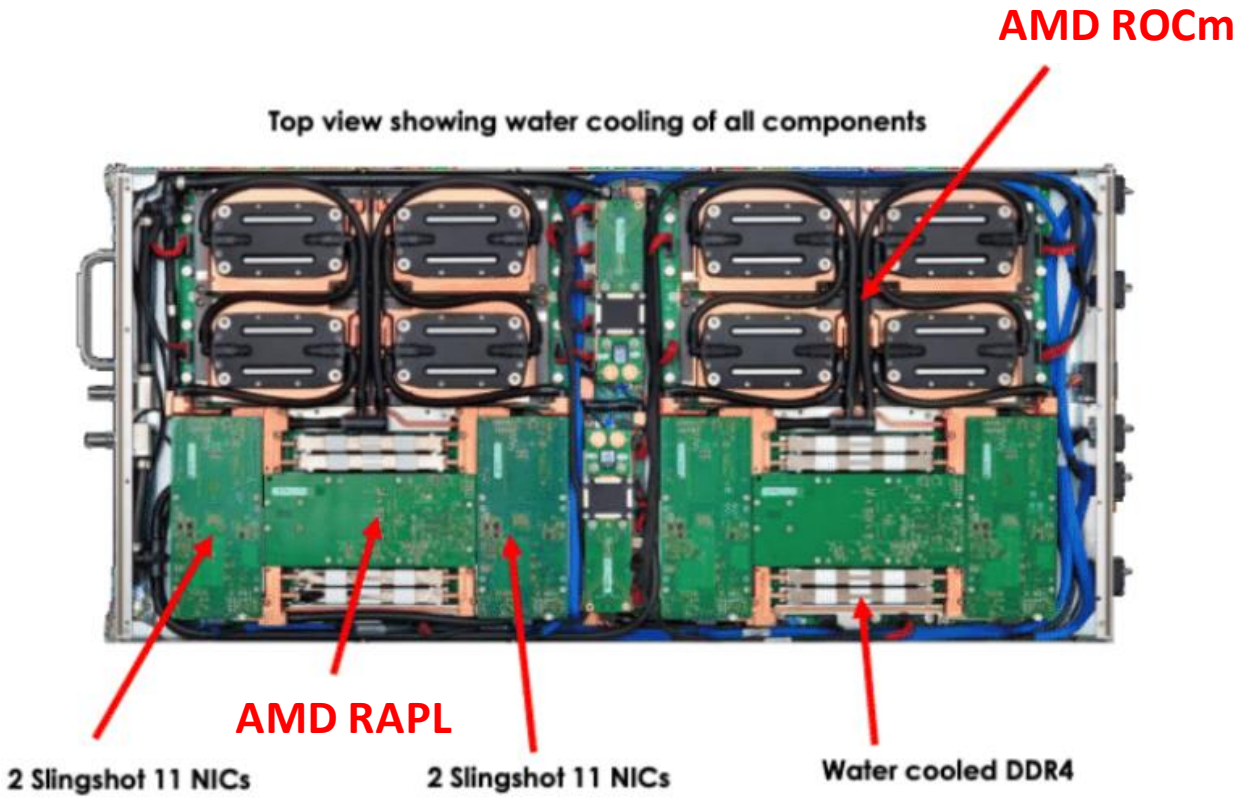
Images of HPE Cray, OLCF Frontier

POWER MONITORING



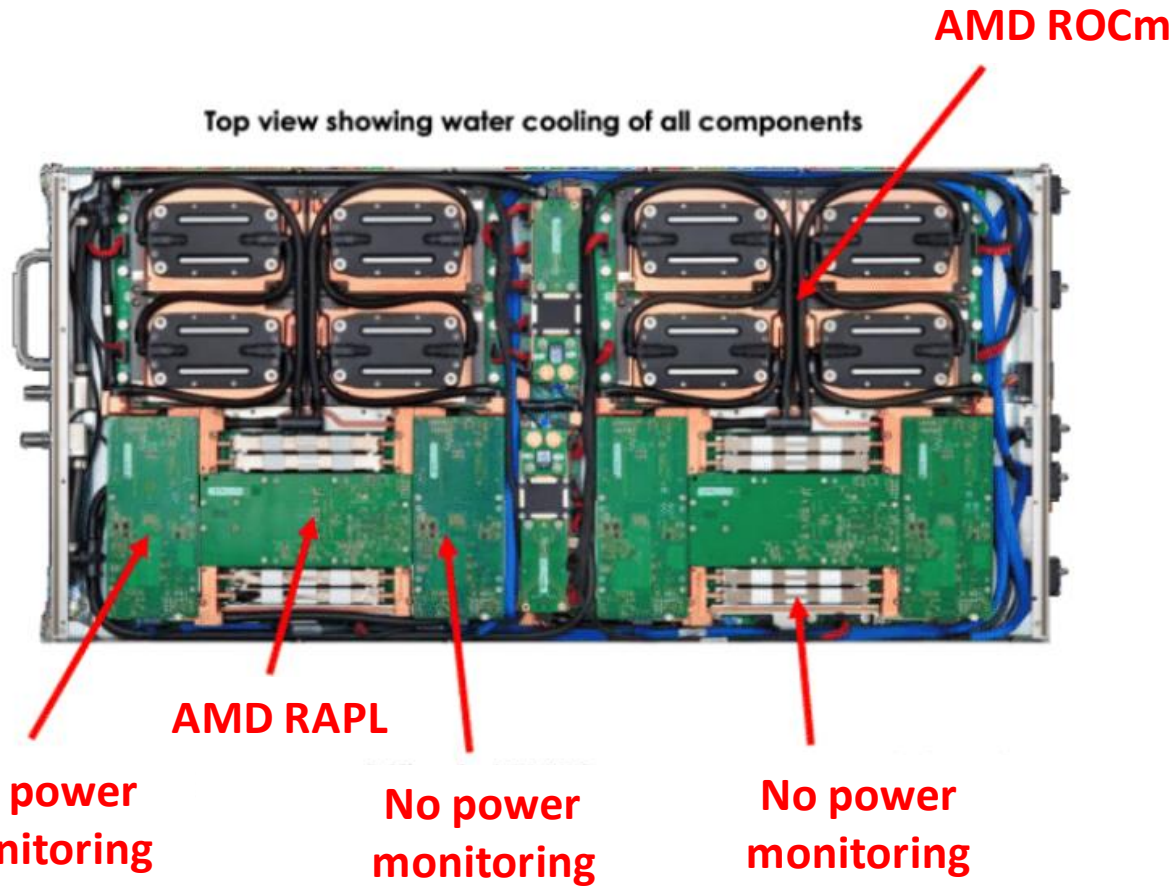
Images of HPE Cray, OLCF Frontier

POWER MONITORING



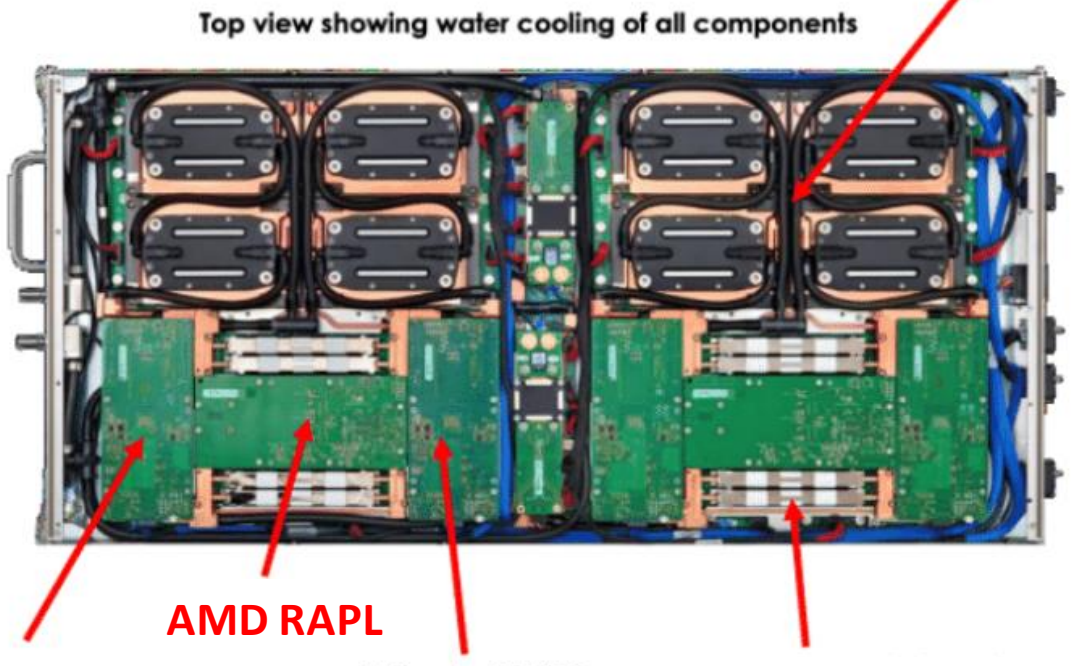
Images of HPE Cray, OLCF Frontier

POWER MONITORING



Images of HPE Cray, OLCF Frontier

POWER MONITORING



Images of HPE Cray, OLCF Frontier

POWER BASELINE

- High frequency *energy* measurement of *some components*
  - Missing energy consumption of the remaining parts
- Low frequency *power* monitoring of the whole *node*
  - Unreliable energy measurement for short and medium length regions

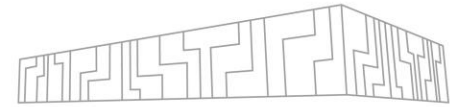
- To estimate power consumption of non-monitored on-node components we evaluate their power consumption from node power monitoring by loading the node by a uniform workload

- Example:

$$Energy = AMD\ RAPL + AMD\ ROCm + power\ baseline \times time$$

- Power baseline is system-specific, should be evaluated for each system individually

KAROLINA ACN POWER BASELINE



GPU	0/0	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
GPU7	55	54	54	54	54	56	398	397	398
GPU6	52	52	51	51	51	397	398	396	397
GPU5	51	51	51	50	50	51	51	55	398
GPU4	52	52	51	51	51	52	52	398	399
GPU3	53	60	398	398	398	398	398	398	398
GPU2	52	398	397	398	398	397	398	397	398
GPU1	55	54	54	57	396	398	398	398	398
GPU0	52	52	51	398	398	394	394	393	398
CPU0	92	94	92	94	96	97	96	99	99
CPU1	95	95	99	99	98	98	102	100	102
SUM CPU	187	189	191	193	194	195	198	199	201
SUM GPU	422	773	1107	1457	1796	2143	2487	2832	3184
ILO avg	1129	1490	1842	2210	2570	2930	3290	3650	4010
CPU+GPU	609	962	1298	1650	1990	2338	2685	3031	3385
BASELINE	520	528	544	560	580	592	605	619	625

GPU load

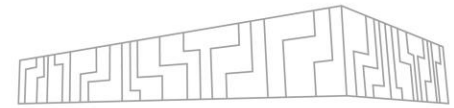
power [W]

CPU	0/0	16/0	32/0	48/0	64/0	64/16	64/32	64/48	64/64
CPU0	92	153	212	268	286	286	286	286	286
CPU1	95	95	95	95	95	154	211	265	285
SUM CPU	187	248	307	363	381	440	497	551	571
SUM GPU	423	425	425	423	422	421	422	421	422
ILO avg	1131	1196	1266	1329	1360	1423	1488	1555	1591
CPU+GPU	610	673	732	786	803	861	919	972	993
BASELINE	521	523	534	543	557	562	569	583	598

CPU load (cores)

power [W]

MERIC LIBRARY



- Performs tracing, (currently) no sampling
- Provides energy consumption of an application or its instrumented regions
- Supported power monitoring systems
  - Intel/AMD RAPL
  - OCC
  - ATOS HDEEM
  - NVML
  - DiG
  - A64FX

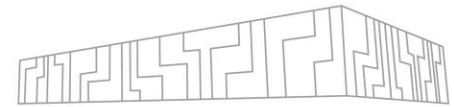
MERIC

```
[vys0053@p10-intel01.cs p10-intel]$ mericStatic -e RAPL -- sleep 2
Runtime [s] = 2.00712
RAM_0 [J] = 6.27963
RAM_1 [J] = 4.47025
PCKG_0 [J] = 332.066
PCKG_1 [J] = 321.863

Energy consumption [J] = 664.679
```

- https://code.it4i.cz/vys0053/meric

SUMMARY



- Power monitoring systems may have various power domains, which may overlap
- Performance counters usually not in joules, but in a specific joule fraction
- Tools such as MERIC unifies access to energy measurement from various power monitoring systems
- **Always** read description of the power monitoring system to understand reported energy consumption correctly
- Add power baseline to on-node components energy consumption for comparable energy measurements from various HW platforms

ACKNOWLEDGEMENT

Funded by the European Union. This project has received funding from the European High Performance Computing Joint Undertaking (JU) and Germany, Bulgaria, Austria, Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, the Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia under grant agreement No 101101903.
This project has received funding from the Ministry of Education, Youth and Sports of the Czech Republic (ID: MC2301).

EURO
CZECHIA

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | IT4INNOVATIONS NÁRODNÍ SUPERPOČÍTAČOVÉ CENTRUM