VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

10 YEAR ANNIVERSARY

MERIC

MERIC: ENERGY EFFICIENCY DATA CENTER SOFTWARE SUITE

Ondřej Vysocký
IT4Innovations

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

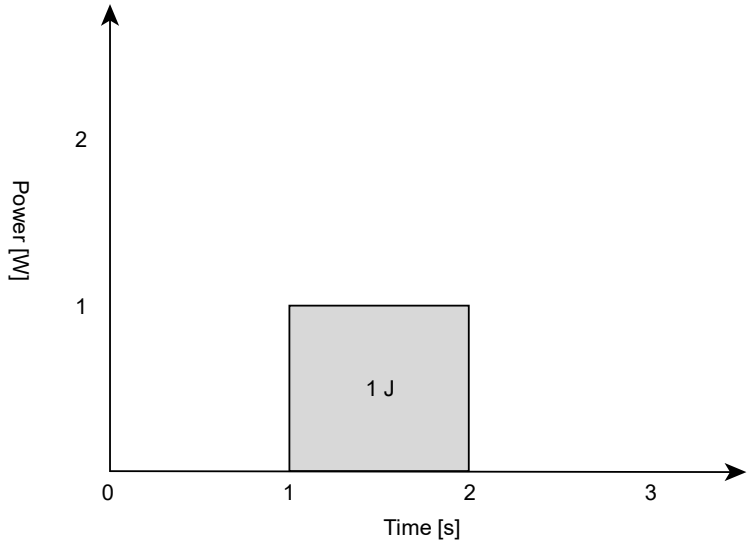
MINISTRY OF EDUCATION, YOUTH AND SPORTS

EuroHPC (2025)

ENERGY

$$Energy = Power \times Time$$

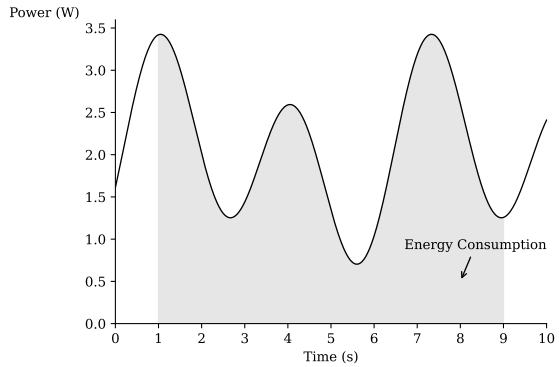
- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J



ENERGY

$$Energy = Power \times Time$$

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

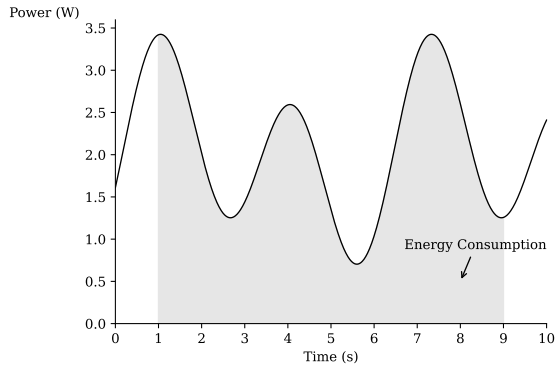


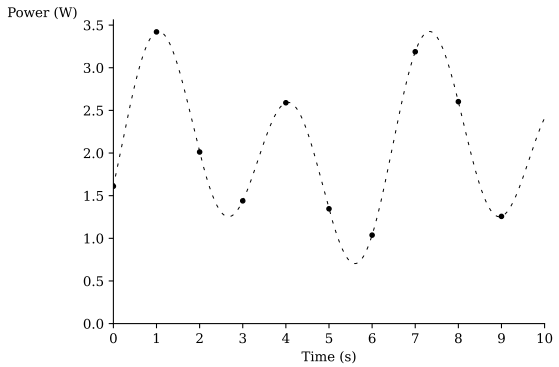
Img source, Luís Cruz (TU Delft)

ENERGY

$$Energy = Power \times Time$$

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



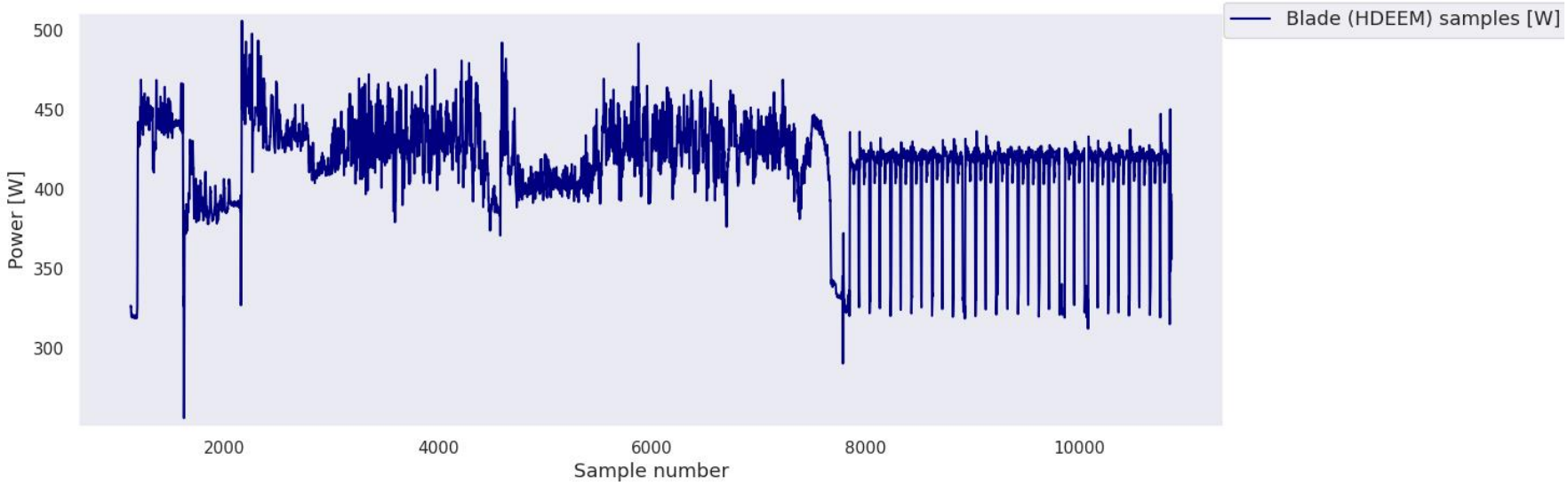


Img source, Luís Cruz (TU Delft)

ENERGY

$$Energy = Power \times Time$$

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

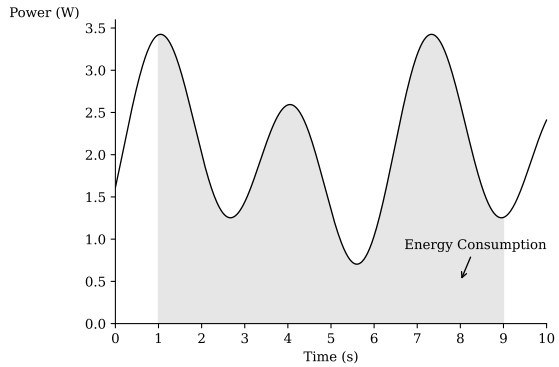


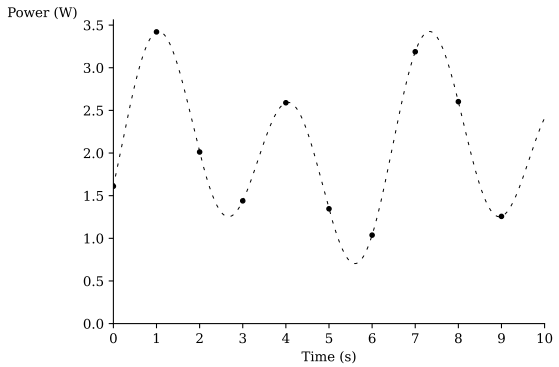
ENERGY

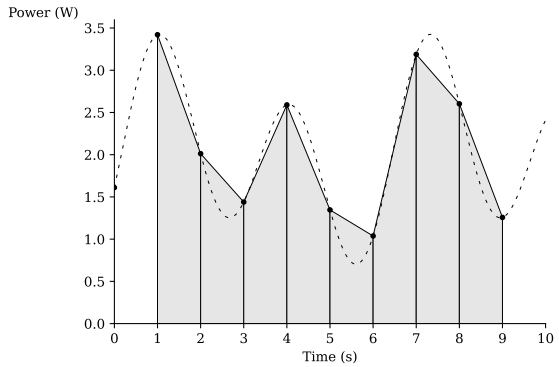
$$Energy = Power \times Time$$

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

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

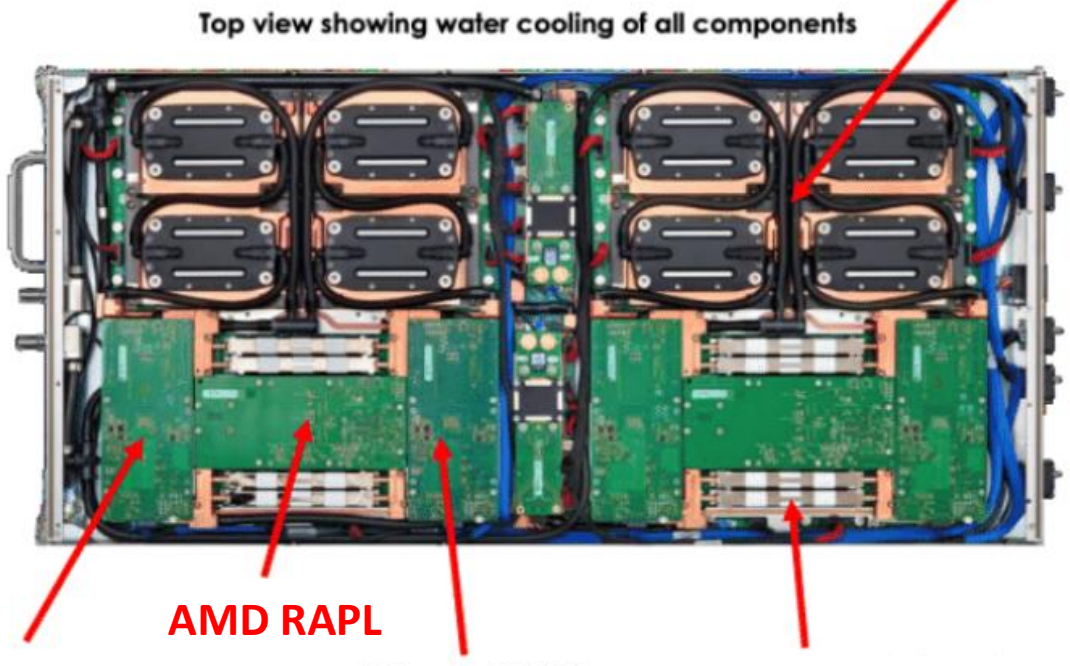






Img source, Luís Cruz (TU Delft)

POWER MONITORING



Images of HPE Cray, OLCF Frontier

NODE POWER BASELINE

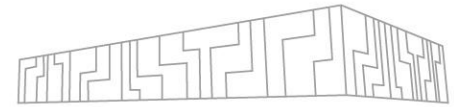
- High frequency *energy* measurement of *some components*
  - Missing energy consumption of the remaining parts of the cluster
- Low frequency *power* monitoring of the whole *node*
  - Unreliable energy measurement for short and medium length regions/applications

- 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

$$NodeEnergy = CPU_{IB} + GPU_{IB} + NodePowerBaseline \times NodeUtilization \times time$$

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

KAROLINA ACN POWER BASELINE



GPU	0/8	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]

- Node power baseline impacted by HW manufacturing variability and node location
- Not using node-specific value, not to penalize user for allocations "worse" nodes

ENERGY EFFICIENCY

$$EnergyEfficiency = \frac{Performance}{Power}$$

ENERGY EFFICIENCY - GREEN500

$$EnergyEfficiency = \frac{R_{max}}{\overline{P}(R_{max})}$$

- LINPACK Rmax/W
- FLOPs/W
- LUPs/W
- …

Green500 Data

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	293	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	44	**Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684

11/2023

https://www.top500.org/files/green500/tutorial.pdf

TOP500 LIST HPL

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)	
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581	**58.9 GF/W**
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607	**52.5 GF/W**
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698	**26.2 GF/W**
4	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84		
5	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461	**56.6 GF/W**
6	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899	**14.8 GF/W**
7	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	574.84	7,124	**61.1 GF/W**
8	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107	**51.6 GF/W**

Exascale goal is
50 GFlops/Watt = 20 MW system

11/2024

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

GREEN500

The GREEN 500

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)	
1	420	**KAIROS** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN CALMIP / University of Toulouse - CNRS France	13,056	3.05	46	73.282	Nvidia GH200
2	171	**ROMEO-2025** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, Red Hat Enterprise Linux, EVIDEN ROMEO HPC Center - Champagne-Ardenne France	47,328	9.86	160	70.912	Nvidia GH200
3	225	**Levante GPU extension** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN DKRZ - Deutsches Klimarechenzentrum Germany	35,904	6.75	110	69.426	Nvidia GH200
4	213	**Isambard-AI phase 1** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE University of Bristol United Kingdom	34,272	7.42	117	68.835	Nvidia GH200
5	286	**Otus (GPU only)** - ThinkSystem SD665-N V3, AMD EPYC 9655 96C 2.6GHz, NVIDIA H100 SXM5 80GB, Infiniband NDR, Rocky Linux 9.4, Lenovo Universitaet Paderborn - PC2 Germany	19,440	4.66		68.177	Nvidia H100
6	73	**Capella** - Lenovo ThinkSystem SD665-N V3, AMD EPYC 9334 32C 2.7GHz, Nvidia H100 SXM5 94Gb, Infiniband NDR200, AlmaLinux 9.4, MEGWARE TU Dresden, ZIH Germany	85,248	24.06	445	68.053	Nvidia H100
7	334	**SSC-24 Energy Module** - HPE Cray XD670, Xeon Gold 6430 32C 2.1GHz, NVIDIA H100 SXM5 80GB, Infiniband NDR400, RHEL 9.2, HPE Samsung Electronics South Korea	11,200	3.82	69	67.251	Nvidia H100
8	96	**Helios GPU** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cyfronet Poland	89,760	19.14	317	66.948	Nvidia GH200
9	426	**AMD Ouranos** - BullSequana XH3000, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Infiniband NDR200, RedHat Enterprise Linux, EVIDEN Atos France	16,632	2.99	48	66.464	AMD MI300A
10	70	**Portage** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, RHEL 8.9, HPE Hewlett Packard Enterprise United States	129,024	24.50	370	66.277	AMD MI300A

11/2025

IRSCHING POWER STATIONS

20 MW ~ 1%

POWER DEMANDS

CNBC

Search quotes, news & videos

MARKETS BUSINESS INVESTING TECH POLITICS VIDEO INVESTING CLUB JOIN

ENERGY

Google signs deal with nuclear company as data center power demand surges

PUBLISHED MON, OCT 14 2024•3:00 PM EDT | UPDATED TUE, OCT 15 2024•11:03 AM EDT

Pippa Stevens
@PIPPASTEVENS13

SHARE

KEY POINTS

- Google said
  Power.
- Tech compa
  data centers
- Google said

The Washington Post

Sign in

Business Economy Economic Policy Personal Finance Work Technology Business

Microsoft deal would reopen Three Mile Island nuclear plant to power AI

The owner of the shuttered Pennsylvania plant plans to bring it online by 2028, with the tech giant buying all the power it produces.

Updated September 20, 2024

NuclearNewswire

INDUSTRY

Amazon buys nuclear-powered data center from Talen

Thu, Mar 7, 2024, 2:01PM | Nuclear News

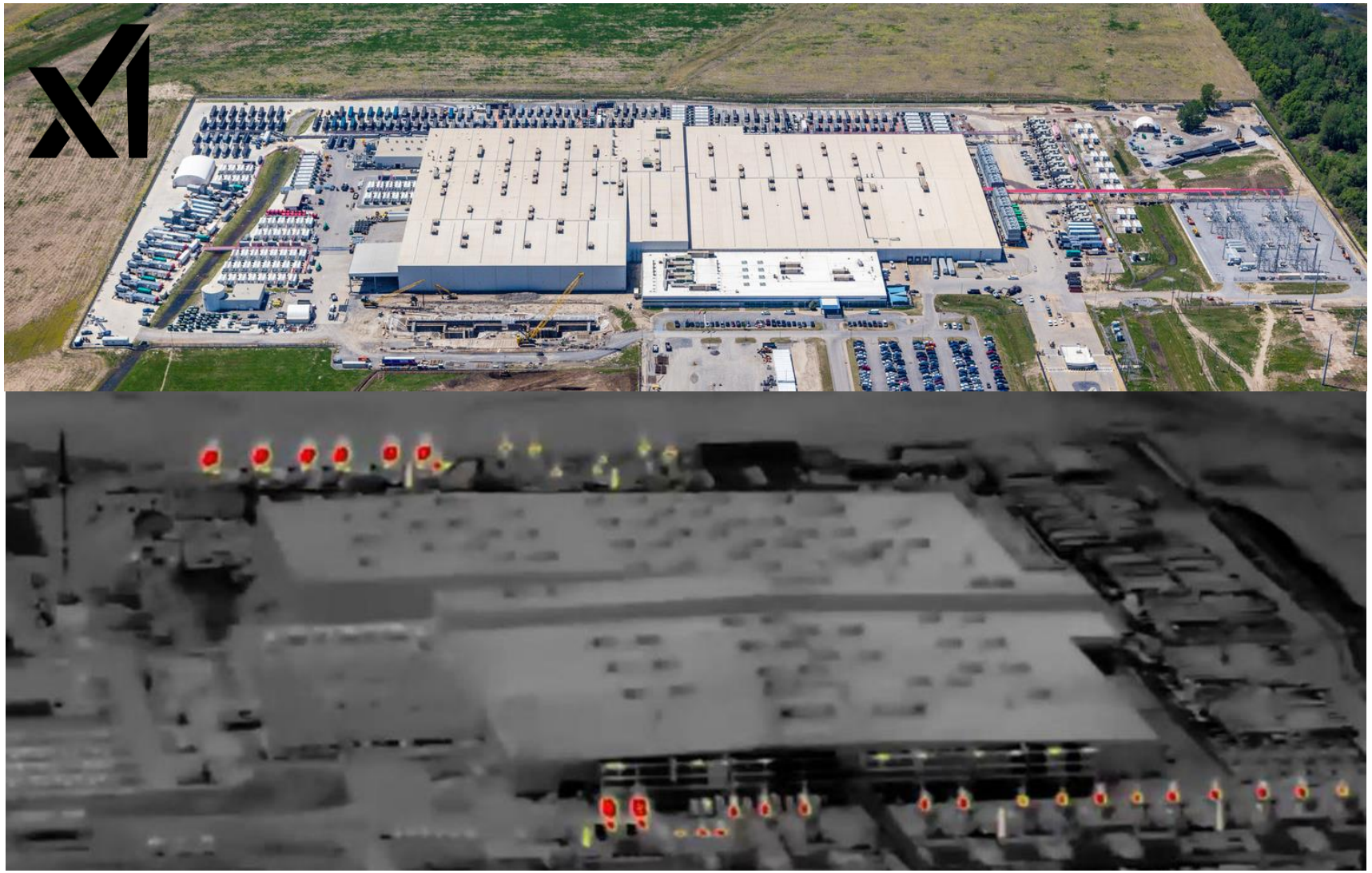
quehanna nuclear plant in Salem Township, Penn., along with the data center in ground. (Photo: Talen Energy)

n Energy announced its sale of a 960-megawatt data center campus oud service provider Amazon Web Services (AWS), a subsidiary of zon, for $650 million.

data center, Cumulus Data Assets, sits on a 1,200-acre campus in sylvania and is directly powered by the adjacent Susquehanna m Electric Station, which generates 2.5 gigawatts of power.

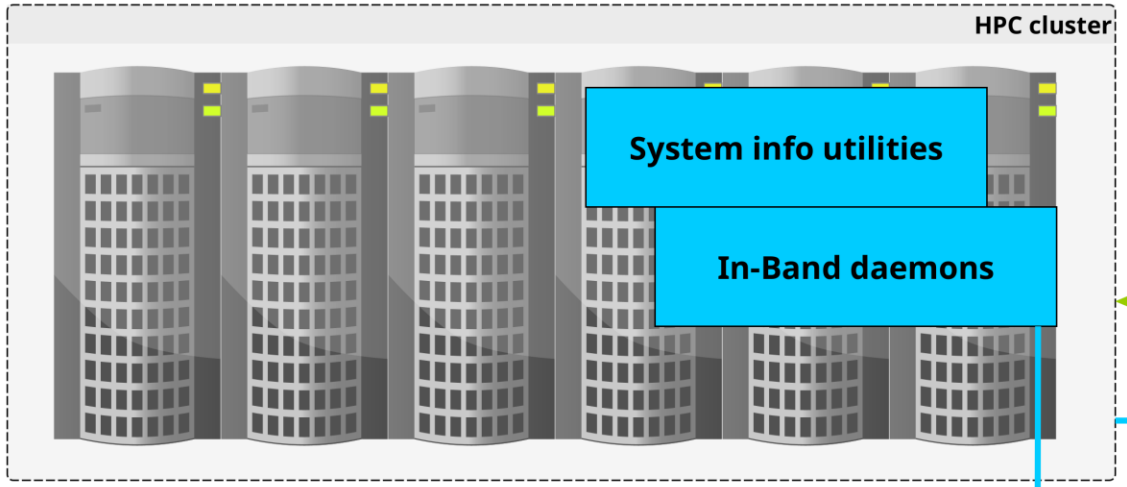
POWER DEMANDS

POWER DEMANDS



5/2025 https://edition.cnn.com/2025/05/19/climate/xai-musk-memphis-turbines-pollution
4/2025 https://www.capacitymedia.com/article/musks-xai-data-centre-allegedly-running-illegal-gas-turbines

MONITORING & BUDGETING



MERIC

legend
system power monitoring
system power management

Out-Of-Band daemons (VM)

System monitoring DB (VM)

VR visualization HPCMonitor4B

Data analytics (VM)

Grafana dashboards & alerting (VM)

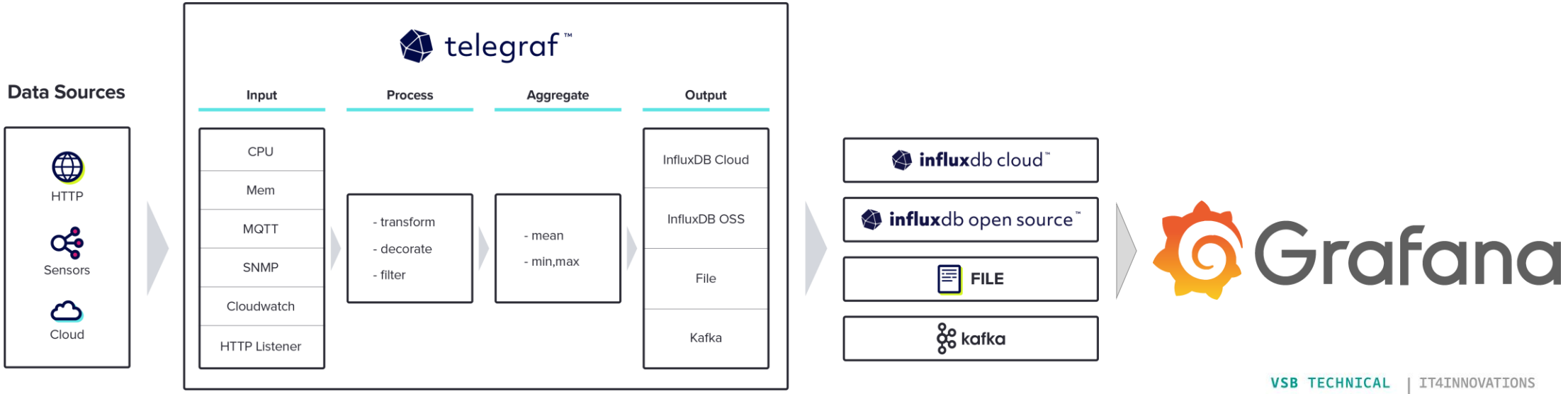
Job energy/CO2e budgeting DB & service (VM)

Job energy/CO2e budgeting API (login nodes)

slurm workload manager

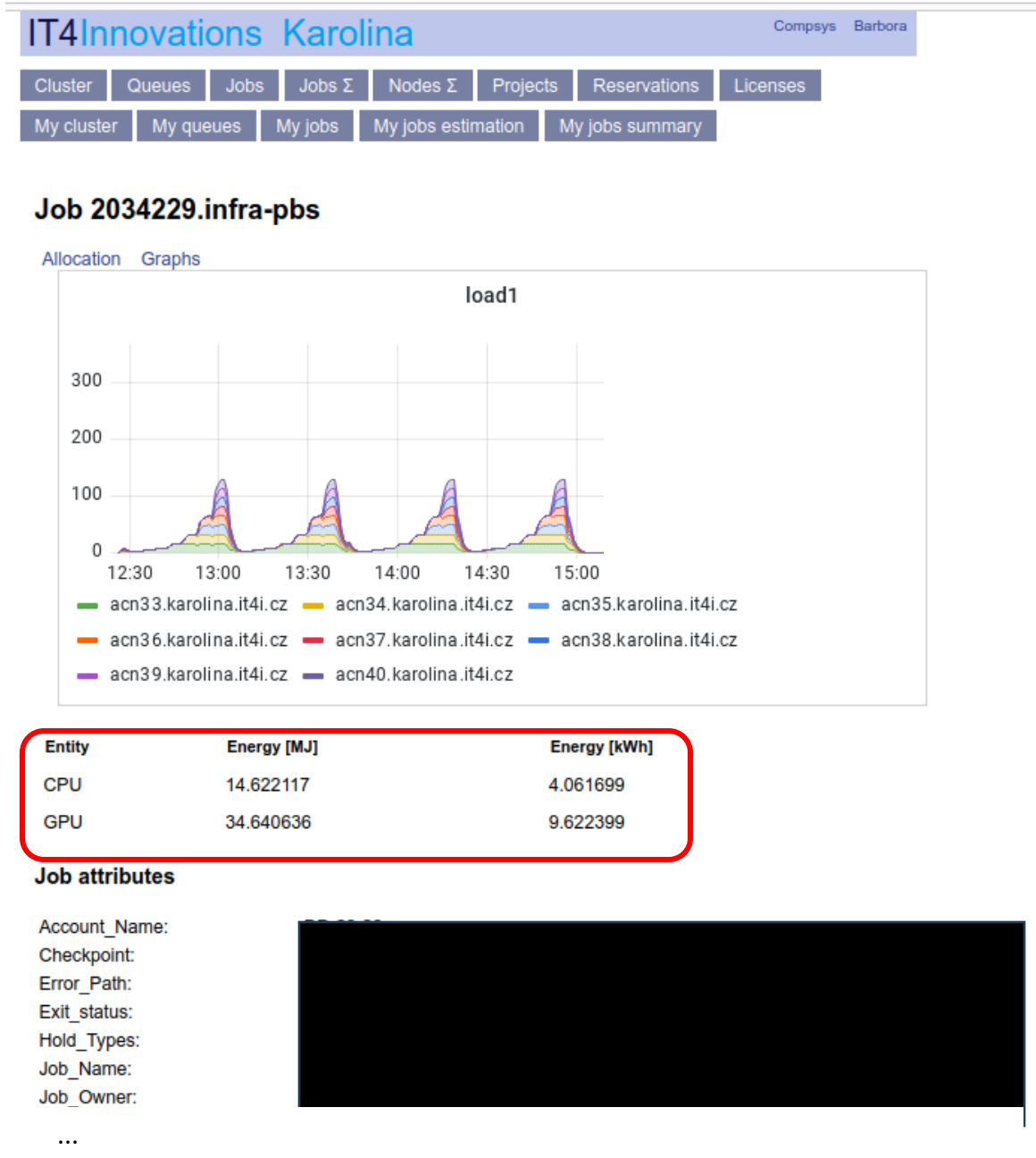
POWER MONITORING

- Power management related daemons
  - Power, energy, temperature, utilization, etc.
  - Both in-band & out-of-band (node, chassis, PDU)
  - Alerting
- Libraries executed
- Scheduler-provided data
  - Job name, owner, project, set of nodes, start & stop timestamp



JOB BUDGETING

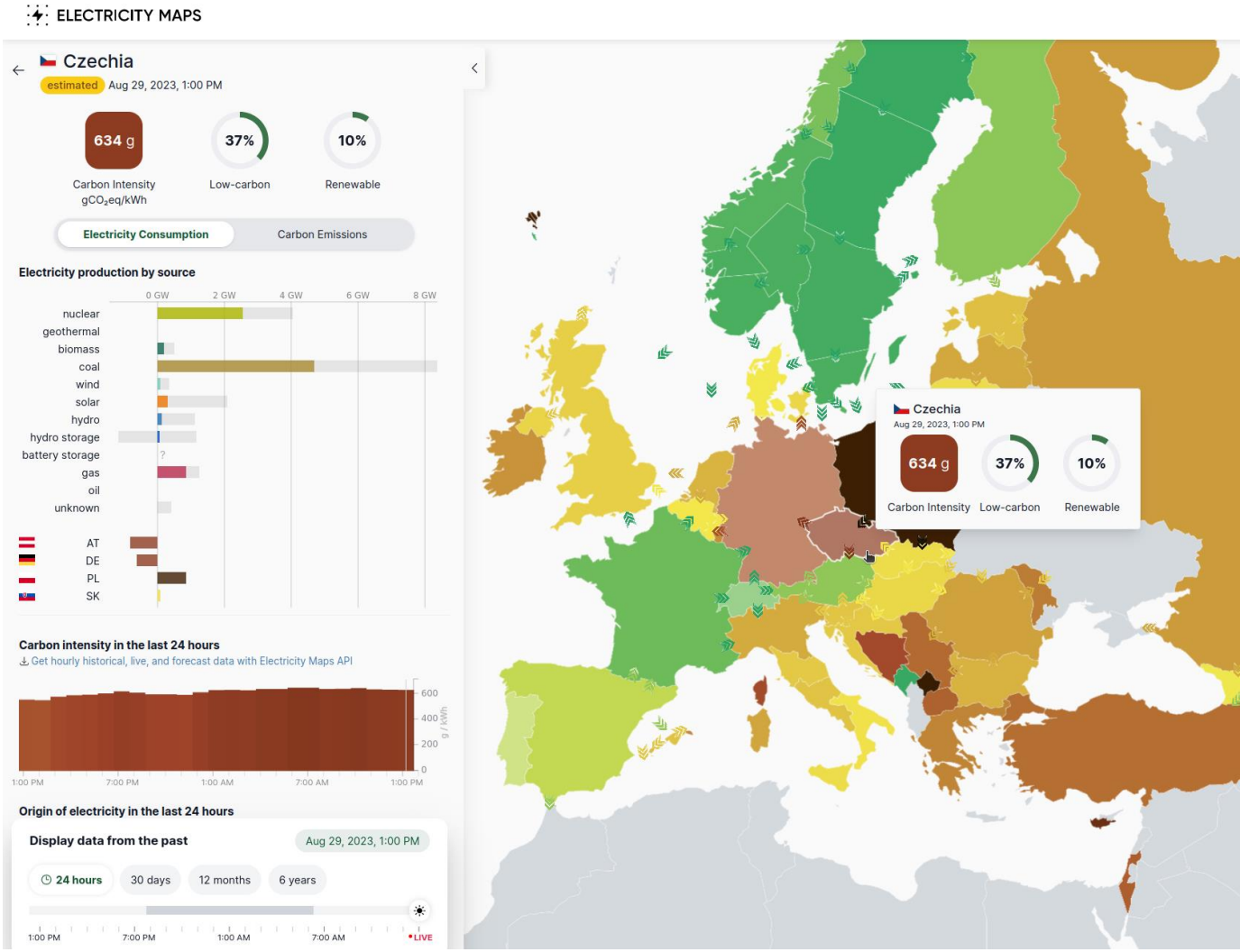
- Support for Slurm and PBS
- Energy consumption reported for each component type
  - CPU
  - GPU
  - Node = IB + OOB
  - Overall = Node * PUE
  - CO2eq



CO_2e

- CO_2 emissions calculation:
  - **CO_2Emissions[g]** = NodeEnergy[kWh] ∗ CarbonIntensity[gCO2/kWh]
- https://app.electricitymaps.com/
- Hourly CO2e/kWh per zone (country)
- API to get the data

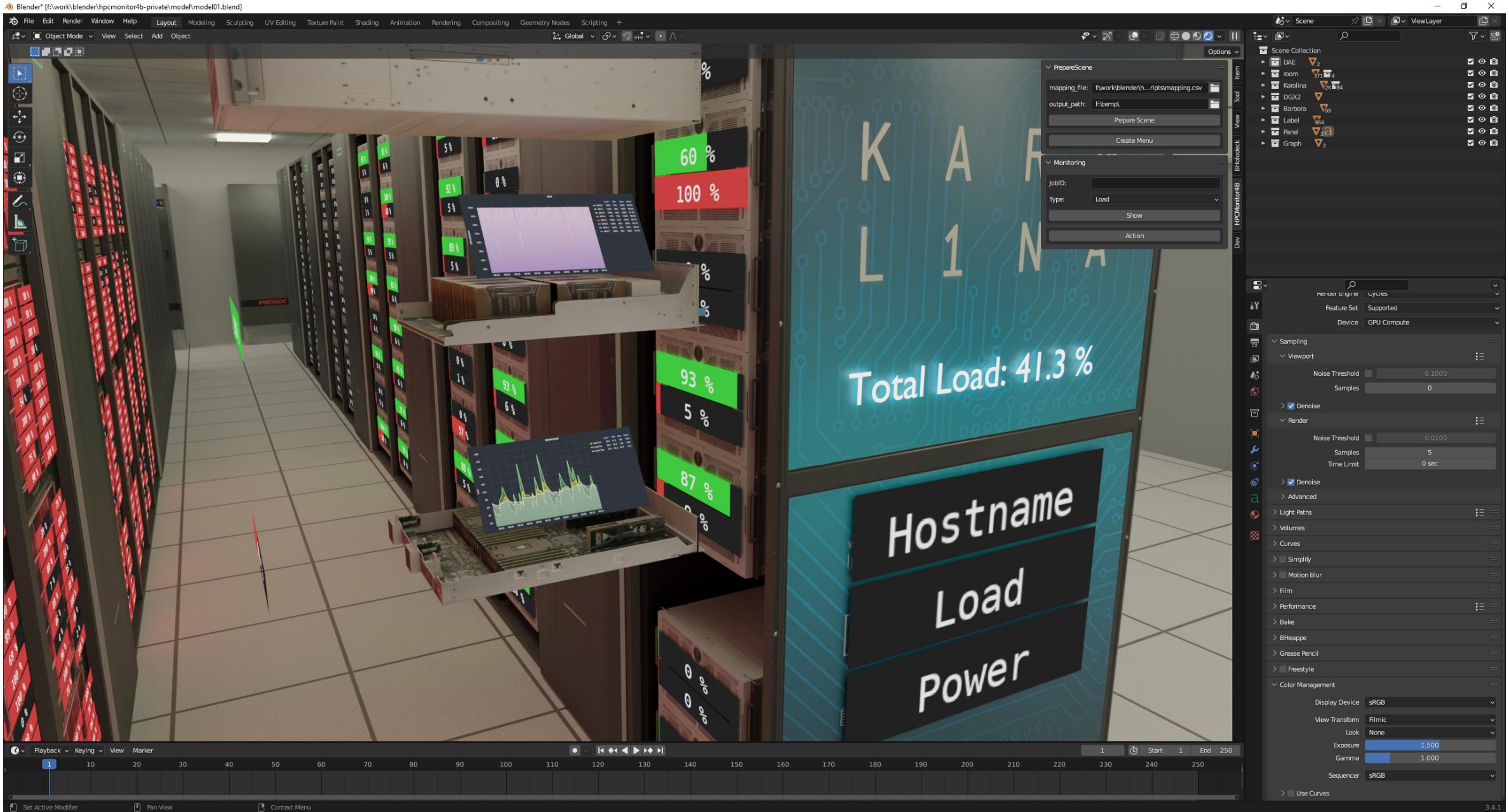
- User must understand the values
  - Kilometers driving a car (Toyota Corolla ST 1.8 Hybrid, 103 gCO2e/Km)
  - Hours of flying (Boeing 747-400, 92 kg CO2e per passenger per hour)



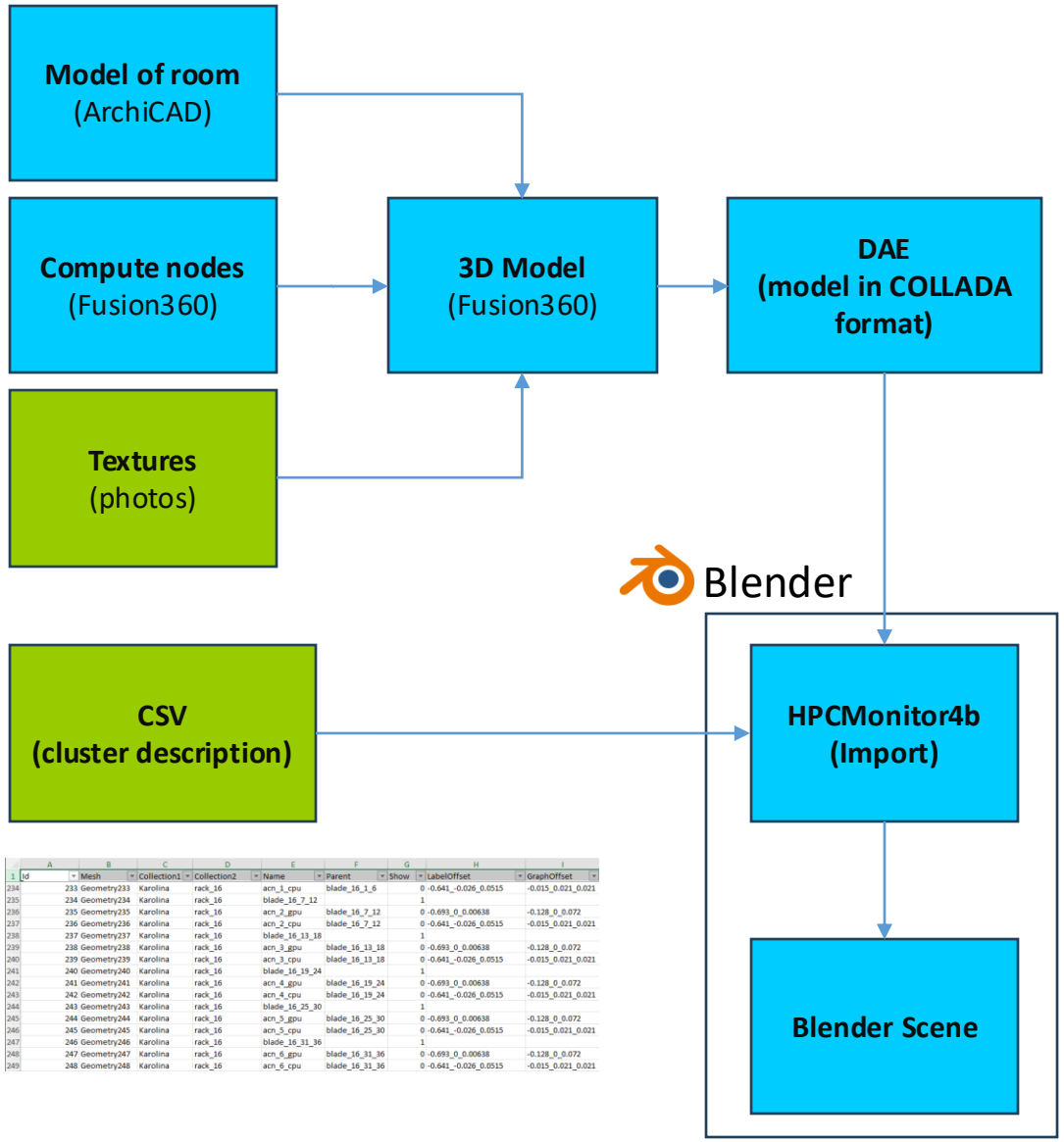
VR VISUALIZATION

EXADIGIT

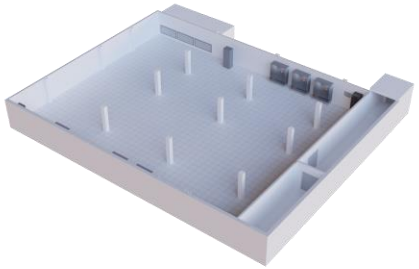
blender®

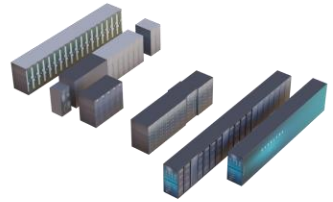


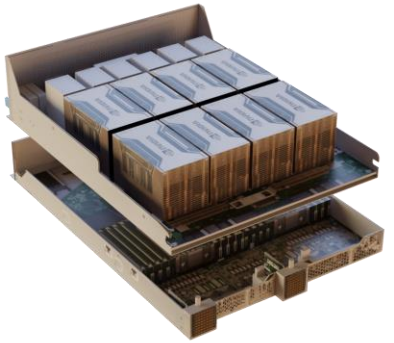
CREATION OF A 3D MODEL



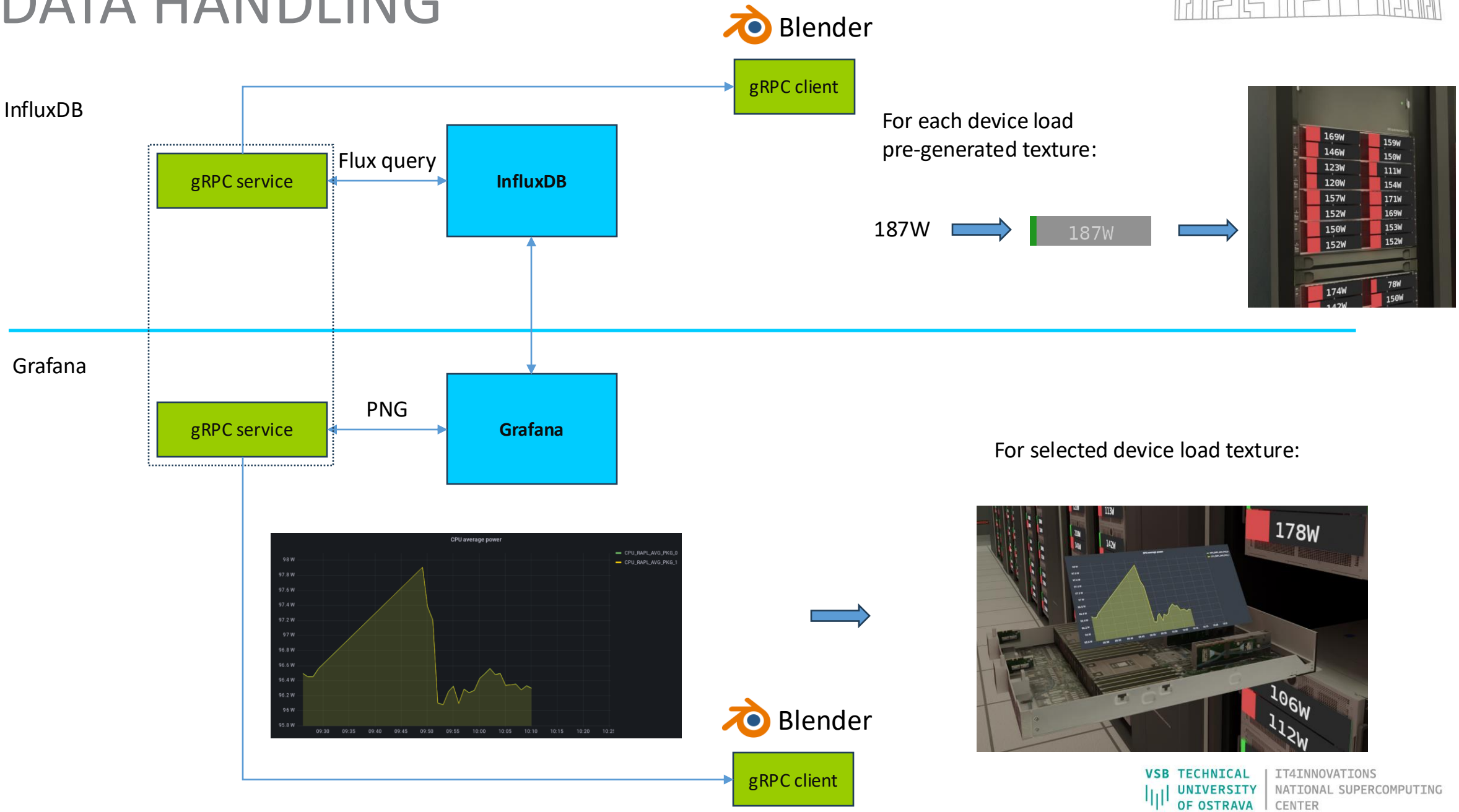
	A	B	C	D	E	F	G	H	I
1	Id	Mesh	Collection1	Collection2	Name	Parent	Show	LabelOffset	GraphOffset
234	233	Geometry233	Karolina	rack_16	acn_1_cpu	blade_16_1_6	0	-0.641_-0.026_0.0515	-0.015_0.021_0.021
235	234	Geometry234	Karolina	rack_16	blade_16_7_12		1		
236	235	Geometry235	Karolina	rack_16	acn_2_gpu	blade_16_7_12	0	-0.693_0_0.00638	-0.128_0_0.072
237	236	Geometry236	Karolina	rack_16	acn_2_cpu	blade_16_7_12	0	-0.641_-0.026_0.0515	-0.015_0.021_0.021
238	237	Geometry237	Karolina	rack_16	blade_16_13_18		1		
239	238	Geometry238	Karolina	rack_16	acn_3_gpu	blade_16_13_18	0	-0.693_0_0.00638	-0.128_0_0.072
240	239	Geometry239	Karolina	rack_16	acn_3_cpu	blade_16_13_18	0	-0.641_-0.026_0.0515	-0.015_0.021_0.021
241	240	Geometry240	Karolina	rack_16	blade_16_19_24		1		
242	241	Geometry241	Karolina	rack_16	acn_4_gpu	blade_16_19_24	0	-0.693_0_0.00638	-0.128_0_0.072
243	242	Geometry242	Karolina	rack_16	acn_4_cpu	blade_16_19_24	0	-0.641_-0.026_0.0515	-0.015_0.021_0.021
244	243	Geometry243	Karolina	rack_16	blade_16_25_30		1		
245	244	Geometry244	Karolina	rack_16	acn_5_gpu	blade_16_25_30	0	-0.693_0_0.00638	-0.128_0_0.072
246	245	Geometry245	Karolina	rack_16	acn_5_cpu	blade_16_25_30	0	-0.641_-0.026_0.0515	-0.015_0.021_0.021
247	246	Geometry246	Karolina	rack_16	blade_16_31_36		1		
248	247	Geometry247	Karolina	rack_16	acn_6_gpu	blade_16_31_36	0	-0.693_0_0.00638	-0.128_0_0.072
249	248	Geometry248	Karolina	rack_16	acn_6_cpu	blade_16_31_36	0	-0.641_-0.026_0.0515	-0.015_0.021_0.021



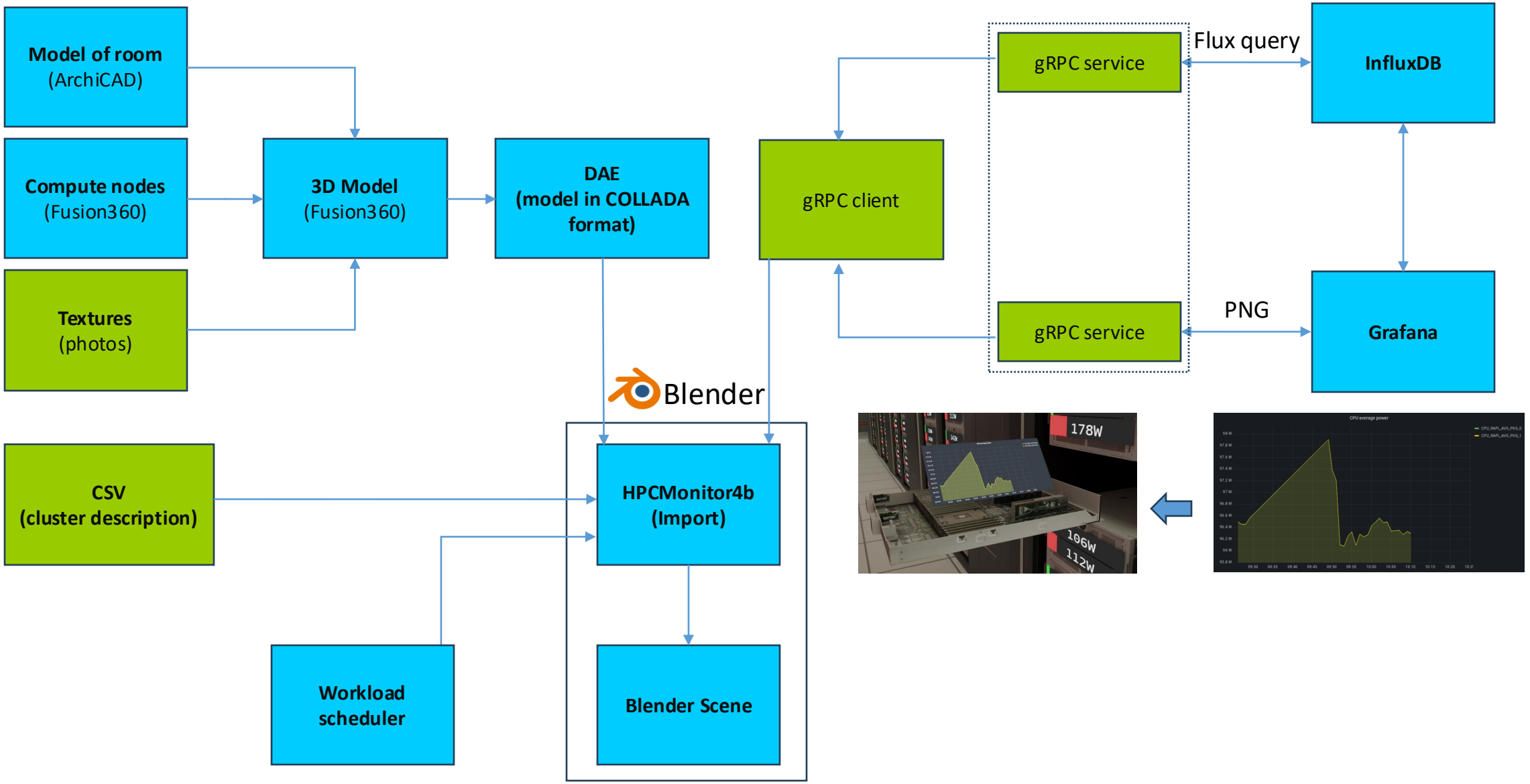




DATA HANDLING

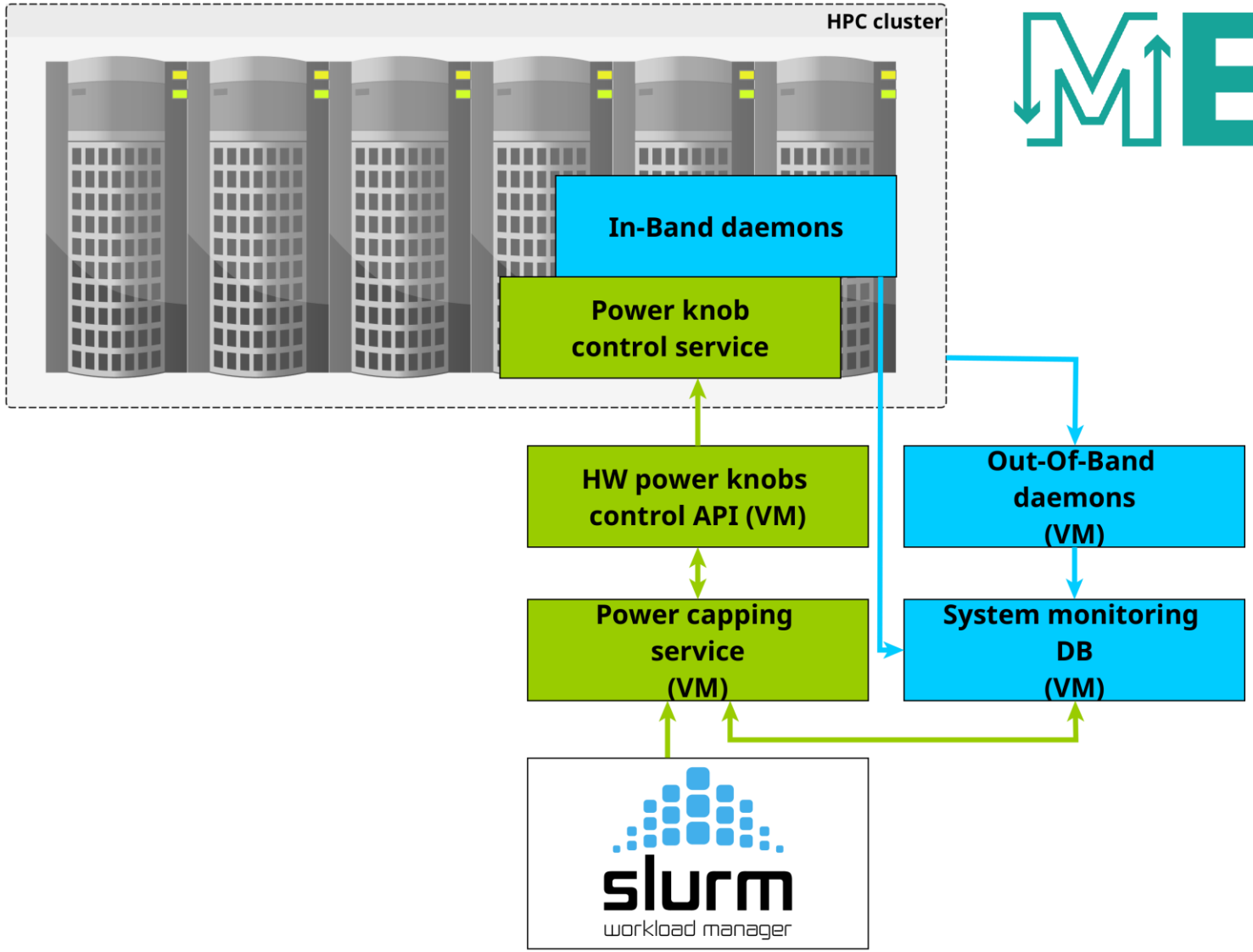


CREATION OF A 3D MODEL



VR VISUALIZATION

POWER CAPPING SERVICE



MERIC

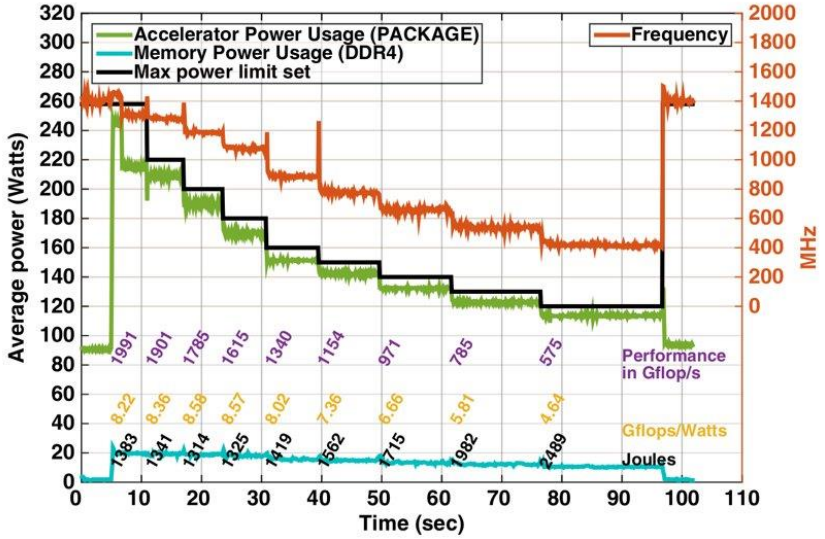
legend

system power monitoring

system power management

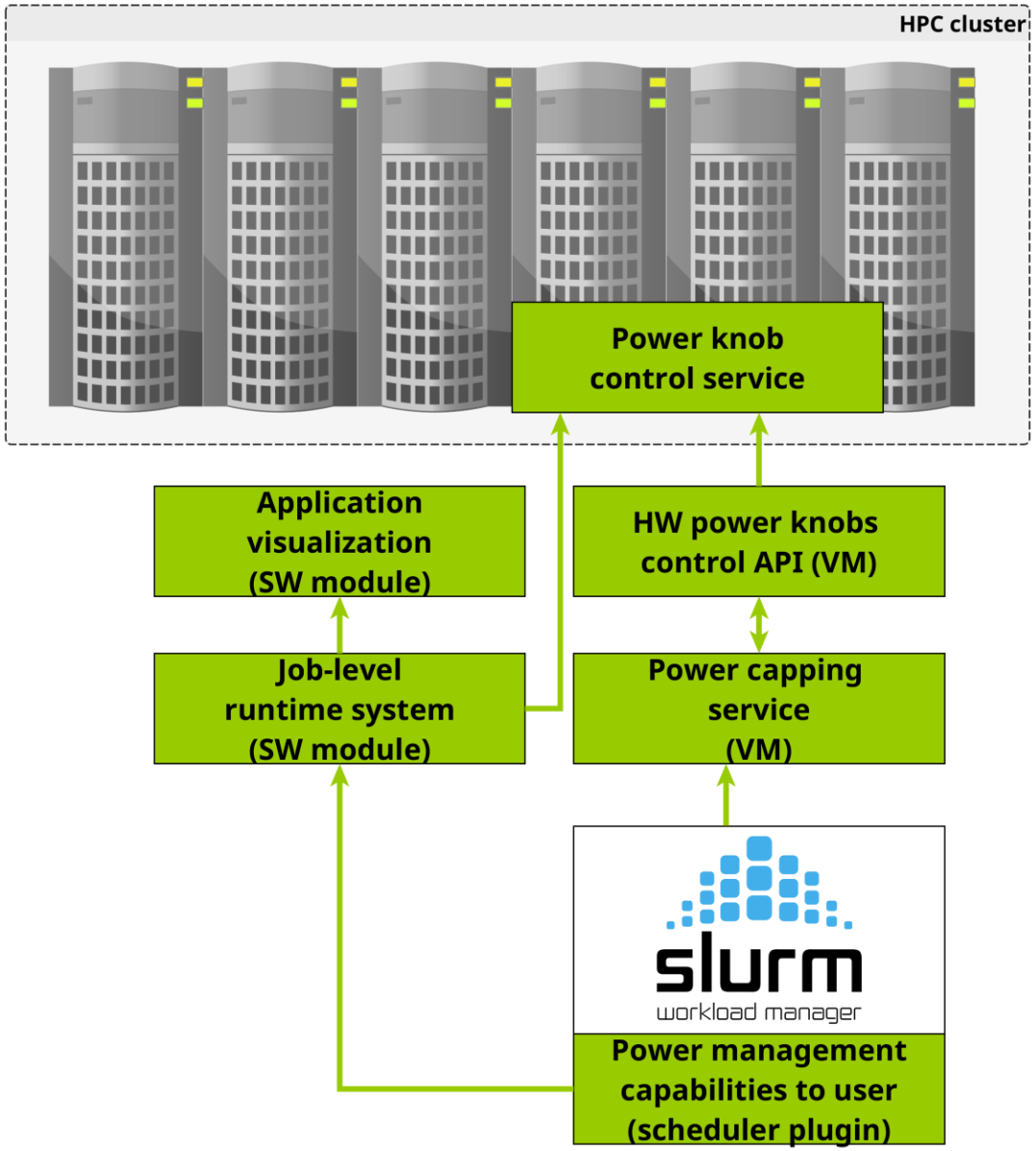
POWER CAPPING SERVICE

- Reacts on new data coming from the system monitoring
- Modifies CPU and GPU power limit
- Inspired by the Intel Running Average Power Limit (RAPL)
  - Using short and long running window
- Admins set a list of rules



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

JOB-LEVEL RUNTIME SYSTEM



MERIC

legend
system power monitoring
system power management

MERIC RUNTIME SYSTEM

- MERIC runtime system provides dynamic application tuning
  - Lightweight & easy to install & easy to use
  - C/C++ API, Fortran module, Python module
  - MPI, OpenMP, CUDA parallelization
- Performance and power aware
- Support for a wide range of architectures
  - x86
  - IBM OpenPOWER
  - ARM
  - Nvidia/AMD GPUs
- Power monitoring systems
  - Intel/AMD RAPL
  - OCC
  - ATOS HDEEM
  - NVML
  - ROCm
  - HWMON *(Nvidia Grace, GraceHopper, AMD RAPL)*
  - A64FX

- **Application energy consumption measurement**
- **Application dynamism & energy efficiency behavior analysis**
- **Dynamic HW power knobs tuning for energy savings**
- **HW & SW power management co-design**

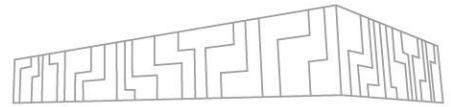
CPU freq, GPU freq,
memory freq, power limit,
#active CPU cores

https://code.it4i.cz/energy-efficiency/meric-suite/meric

ENERGY MEASUREMENT

		JUPITER	LUMI	Leonardo	MN5	Vega	MeluXina	Karolina	Discoverer	Deucalion
AMD CPU	intel_rapl kernel module									
	amd_energy kernel module									
	msr_safe kernel module									
	perf									
Intel CPU	intel_rapl kernel module									
	msr_safe kernel module									
	perf									
A64FX	perf									
Nvidia GPU	NVML									
AMD GPU	ROCm									
GraceHopper	HWMON									

PERFORMANCE METRICS



- Avg CPU core frequency

$$average\ CPU\ core\ frequency = \frac{APERF}{MPERF} * Nominal\ frequency$$

- CPU uncore frequency
- Computational intensity

$$Arithmetic\ Intensity = \frac{Floating\text{-}point\ operations\ executed}{Data\ movement}[FLOPs/B]$$

$$Computational\ Intensity = \frac{Instructions\ executed}{L3\ cache\ misses}$$

- Power capping activity ratio
- Temperature
- Vectorization ratio

STATIC TUNING

various performance degradation limits

various power monitoring systems

	default		-2 % limit	-5 % limit	-10 % limit	unlimited
Runtime [s]	257.07	**Performance penalty [%]**	2.18	4.34	8.84	25.9
Energy [kJ]		**Energy savings [%]**				
HDEEM	881.28	**HDEEM**	5.44	7.30	10.52	11.49
RAPL	668.07	**RAPL**	6.85	9.44	13.92	18.42
Eff. [MFLOPs/W]		**Eff. [MFLOPs/W]**				
HDEEM	10.40	**HDEEM**	11.00	11.22	11.62	11.75
RAPL	13.72	**RAPL**	14.73	15.15	15.94	16.81
		CPU configuration				
		Core freq. [GHz]	3.1	3.0	2.8	2.3
		Uncore freq. [GHz]	1.8	1.8	1.8	1.6

HW configuration tuning

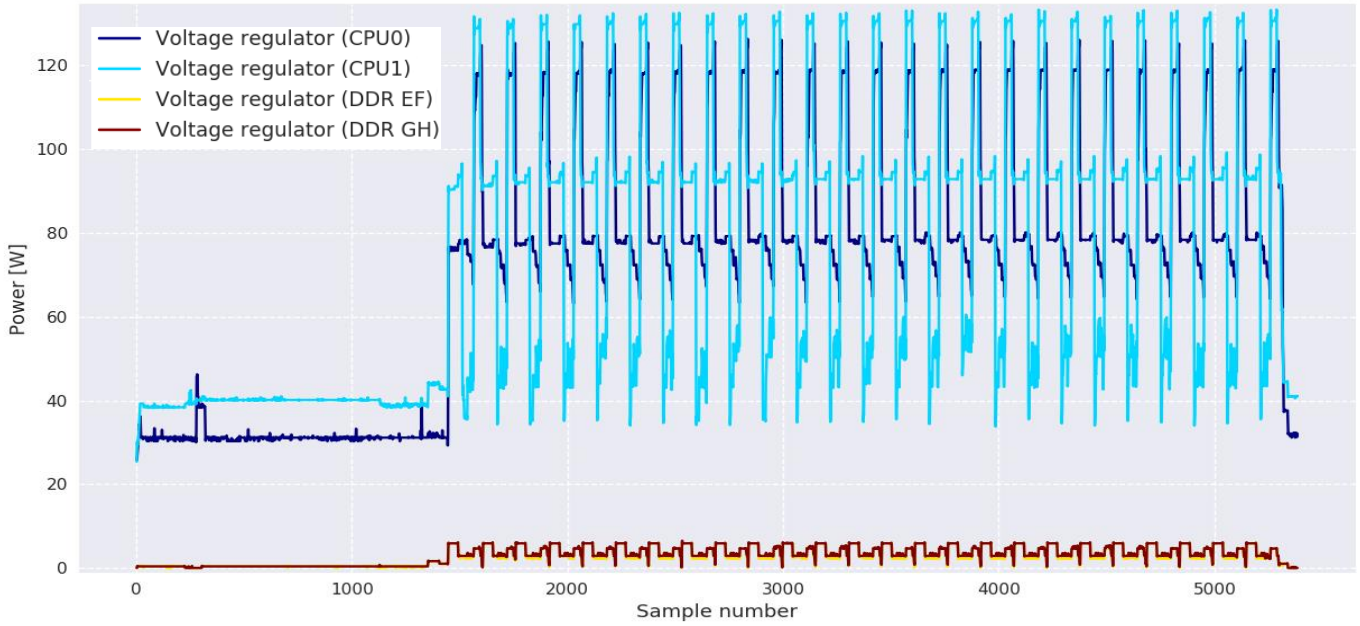
EXAMPLE OF PLATFORM COMPARISON FOR FLEUR

Hardware	Energy efficiency	Node energy consumption	Monitoring system	HW configuration	Runtime
AMD Zen2 (Rome)	1.78 GFLOPs/W	53.36 kJ	AMD RAPL + baseline	default	109 s (100%)
	1.82 GFLOPs/W	52.00 kJ (-3%)		CF 2.9 GHz	101%
	1.94 GFLOPs/W	48.81 kJ (-9%)		CF 2.1 GHz	107%
AMD Zen3 (Milan)	1.67 GFLOPs/W	56.96 kJ	AMD RAPL + baseline	default	93 s (100%)
	1.79 GFLOPs/W	53.05 kJ (-7%)		CF 2.7 GHz	101%
	1.91 GFLOPs/W	49.73 kJ (-13%)		CF 2.0 GHz	112%
Intel Cascade lake	1.00 GFLOPs/W	94.94 kJ	HDEEM	default	217 s (100%)
	1.04 GFLOPs/W	91.26 kJ (-4%)		CF 2.8 GHz, UCF 2.2 GHz	101%
	1.13 GFLOPs/W	84.51 kJ (-11%)		CF 1.9 GHz, UCF 1.8 GHz	123%
Intel Sapphire Rapids w. HBM	1.77 GFLOPs/W	73,31 kJ	Intel RAPL + baseline	default	82 s (100%)
	1.82 GFLOPs/W	71,83 kJ (**-2%**)		CF 3.1 GHz, UCF 1.8 GHz	101%
	1.82 GFLOPs/W	71.83 kJ (-2%)		CF 3.1 GHz, UCF 1.8 GHz	101%
Intel Sapphire Rapids w. DDR memory	1.43 GFLOPs/W	90.22 kJ	Intel RAPL + baseline	default	100 s (100%)
	1.47 GFLOPs/W	88.48 J (-2%)		CF 2.9 GHz, UCF 2.0 GHz	101%
	1.54 GFLOPs/W	86.50 kJ (-4%)		CF 2.3 GHz, UCF 1.8 GHz	110%
Nvidia A100	--	180.6 kJ	AMD RAPL + NVML + baseline	default	111 s (100%)
	--	169.3 kJ (-6%)		1230 MHz	101%
	--	166.3 kJ (-8%)		990 MHz	104%
IBM Power10	0.459 GFLOPs/W	198.6 kJ	PDU	default	199 s
Fujitsu A64FX	0.321 GFLOPs/W	282.5 kJ	perf. counters + baseline	default	812 s

READEX METHODOLOGY

READEX
Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing

- H2020 READEX, 2015-2018
- Complex parallel application has different requirements during execution, so it gives a possibility to be dynamically tuned for energy savings without performance penalty.



- Goal was to create a tools-aided methodology for automatic tuning of parallel applications. Dynamically adjust system parameters to actual resource requirements.

TECHNISCHE UNIVERSITÄT DRESDEN

NTNU

IT4Innovations national supercomputing center

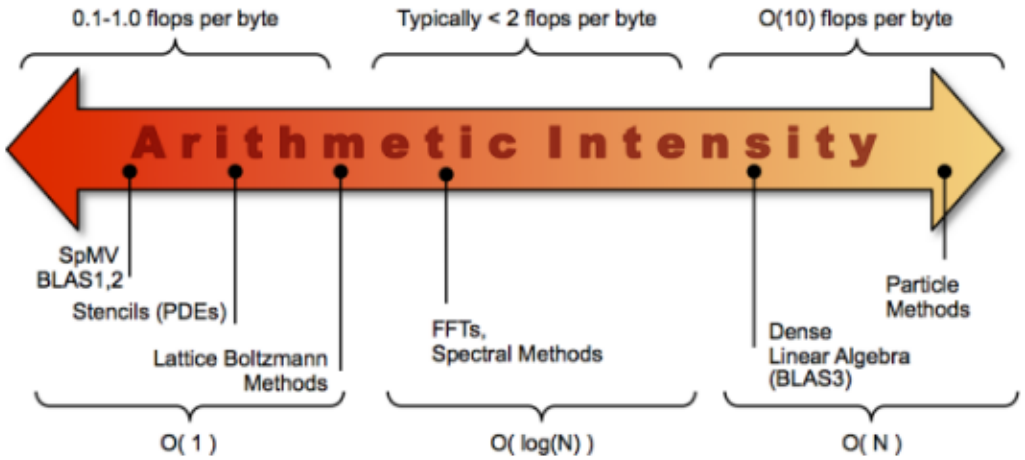
NUI Galway OÉ Gaillimh

intel

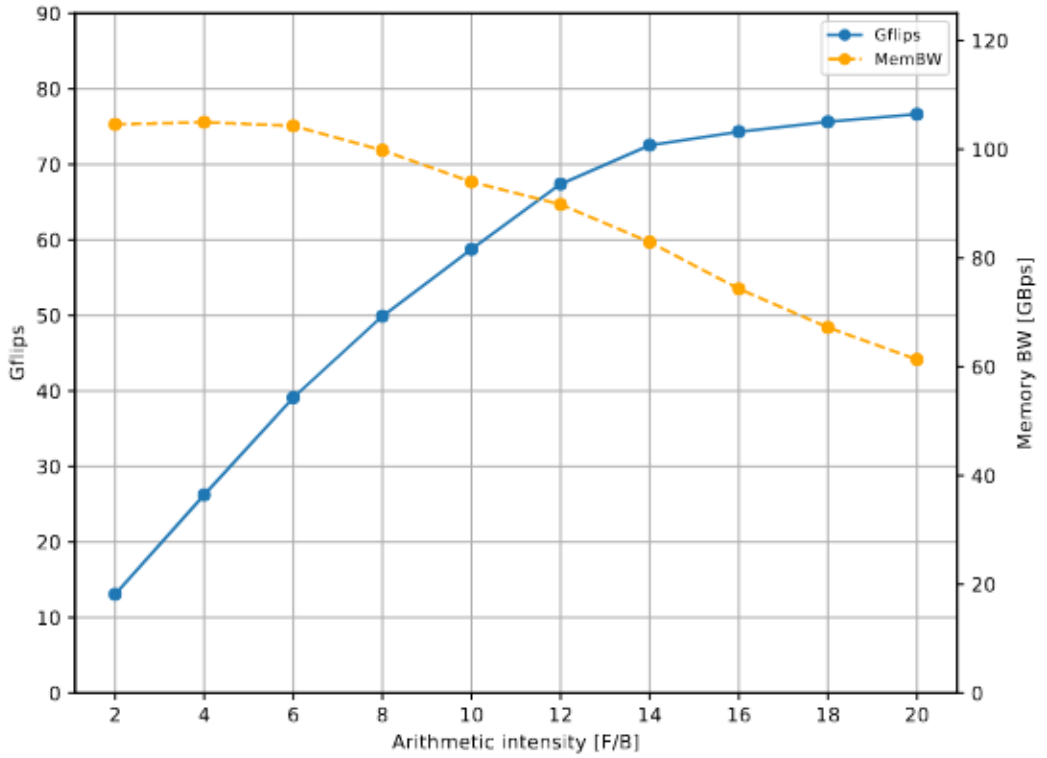
gns

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

DYNAMICITY



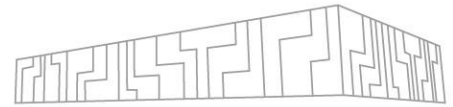
(a) Arrow presenting a range of applications of various arithmetic intensities [54].



(b) Roofline model of the Intel Xeon Gold 6240 processor when executing a workload of AVX-512 instructions.

memory bound, compute bound, communication, I/O, etc.

DYNAMICITY EXPLOITATION



```
void lbm(void) {
  // init application parameters
  // init mass
  for (int iter = 0; iter < NITER; iter++) { // phase region
    update_hallos();  // insignificant region

    propagate();      // significant region

    collide();        // significant region
  }
  // post-processing
  // store output
  // terminate application
}
```

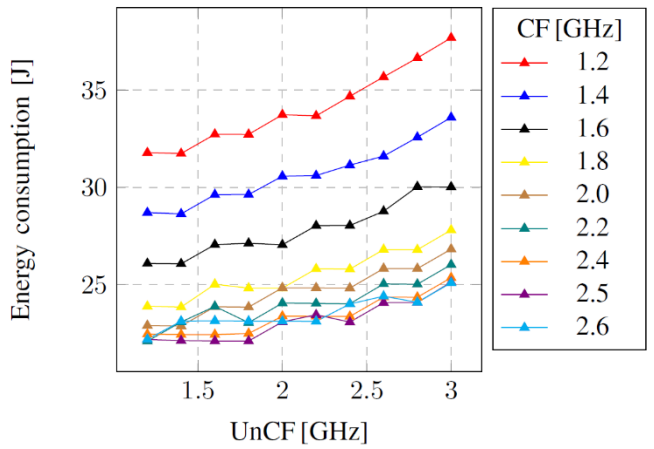
memory bound region
CPU core freq = 1.6 GHz

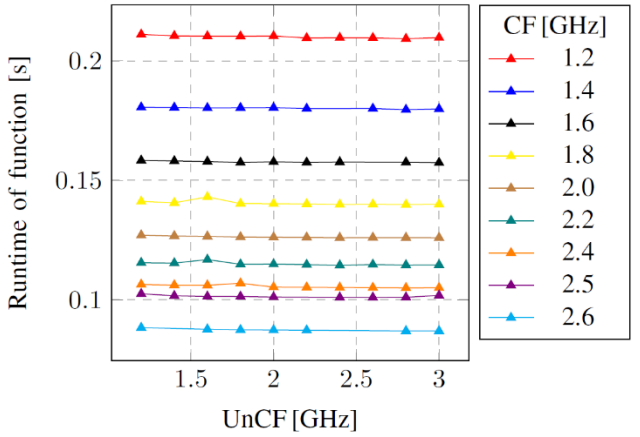
compute bound region
CPU core freq = 2.5 GHz

LATTICE BOLTZMANN BENCHMARK

Collide

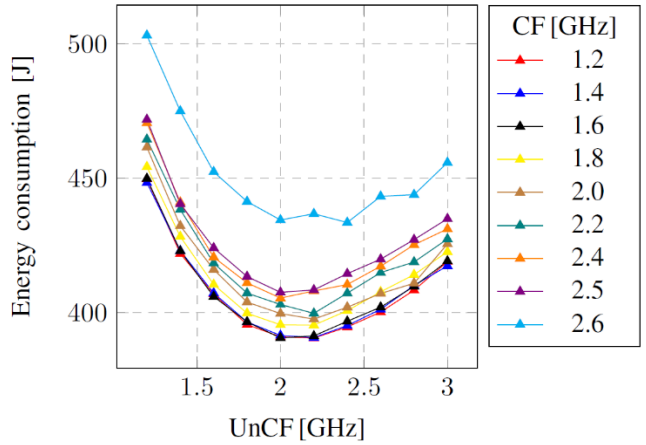
compute bound

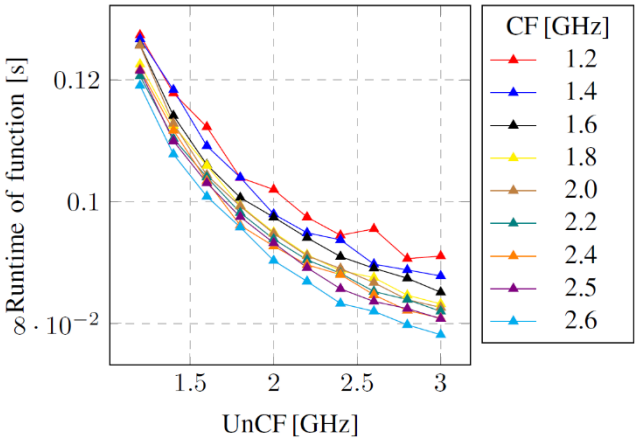




Propagate

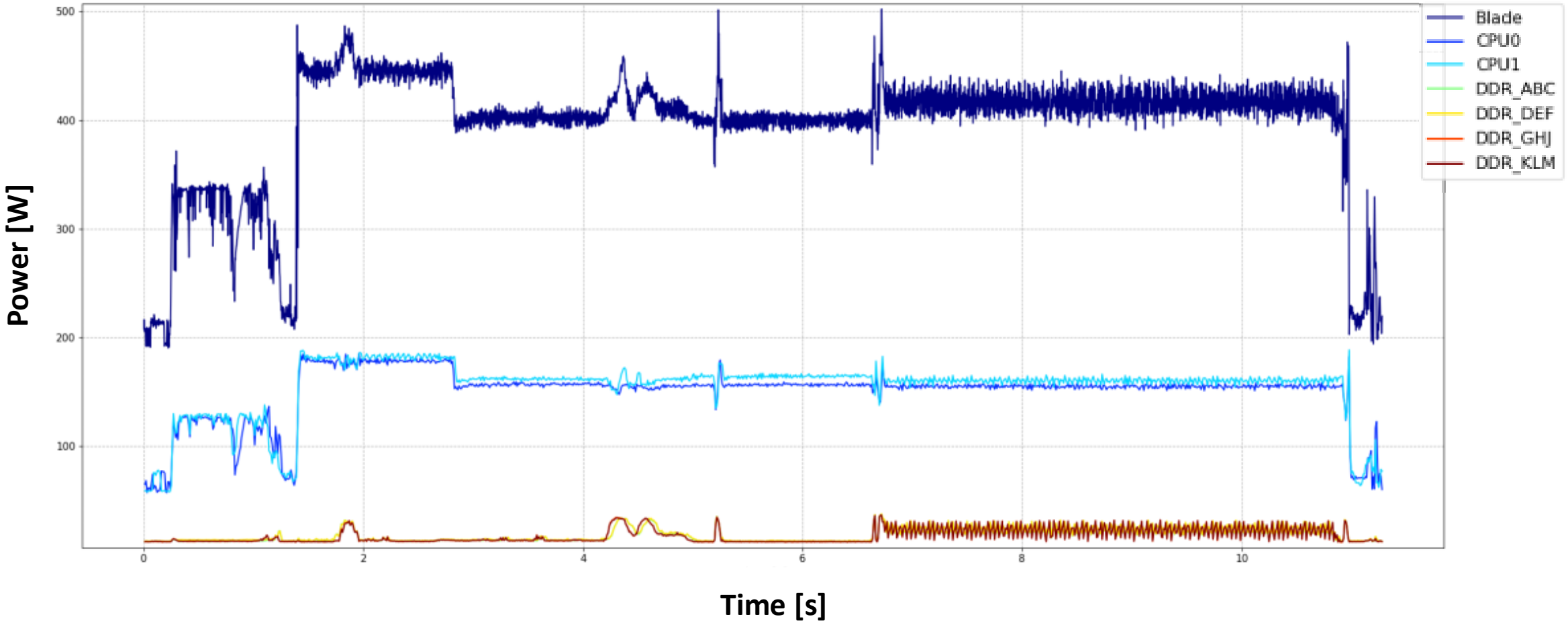
memory bound



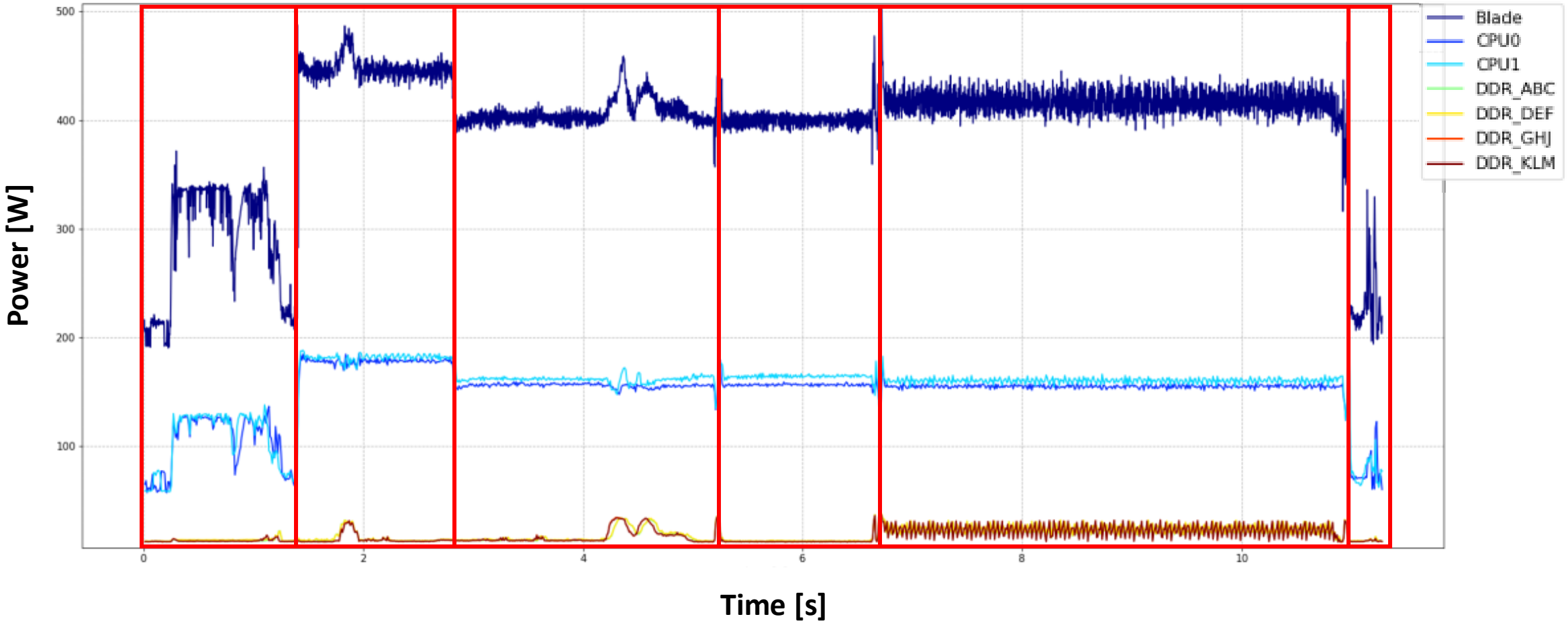


Default		Static savings		Dynamic savings	
time	energy	time	energy	time	energy
24 s	5.7 kJ	-11.6 %	7.8 %	**-1.5 %**	**19.8 %**

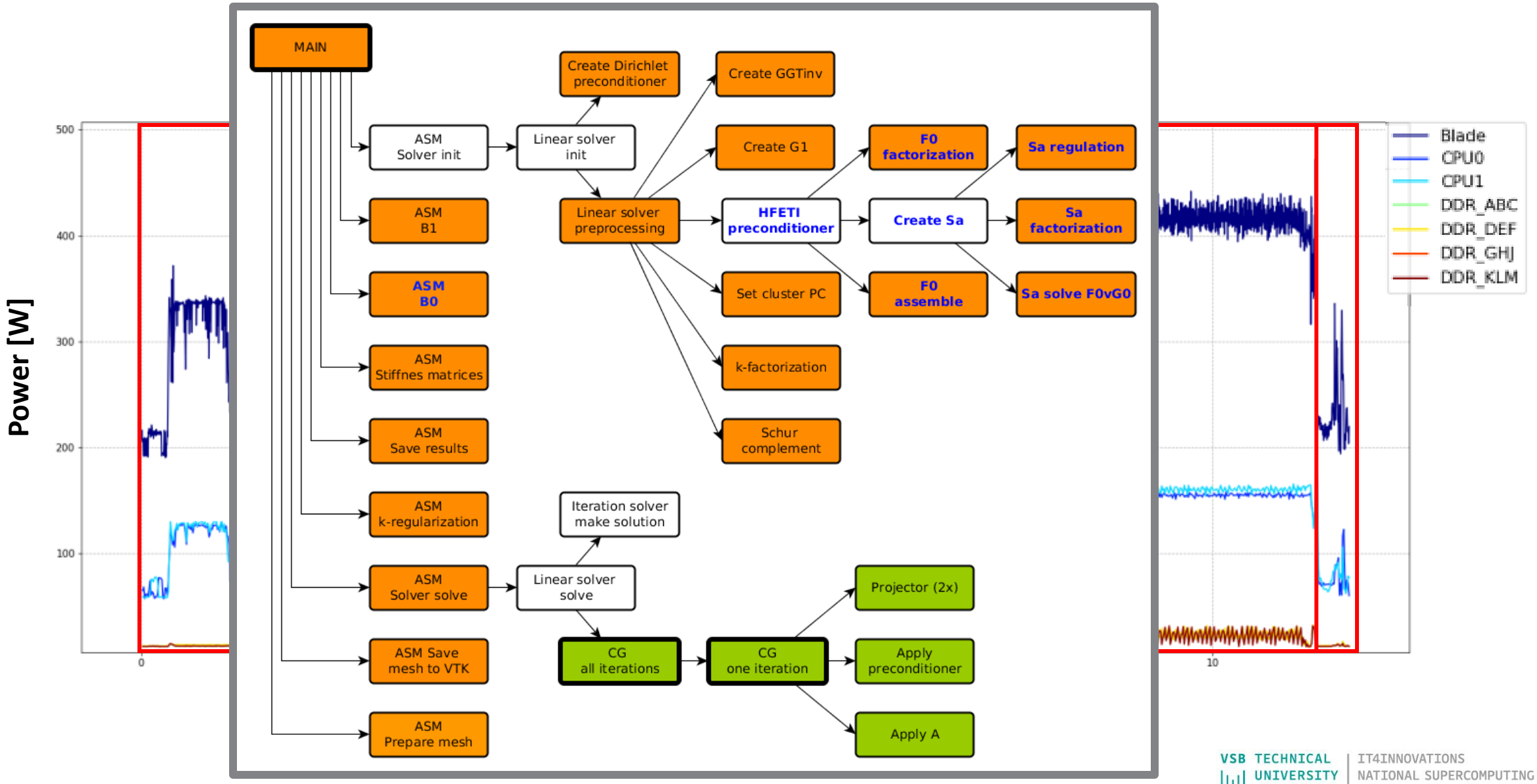
APPLICATION POWER CONSUMPTION



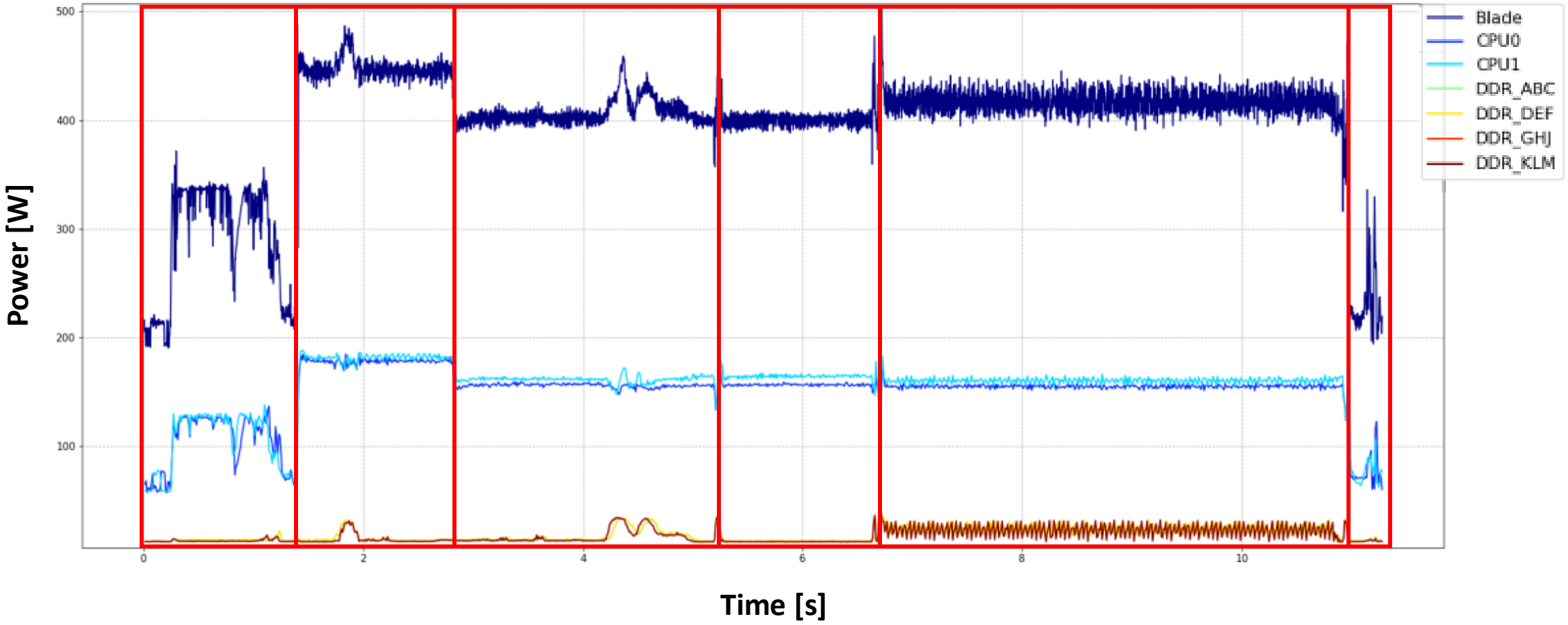
AUTOMATIC INSTRUMENTATION



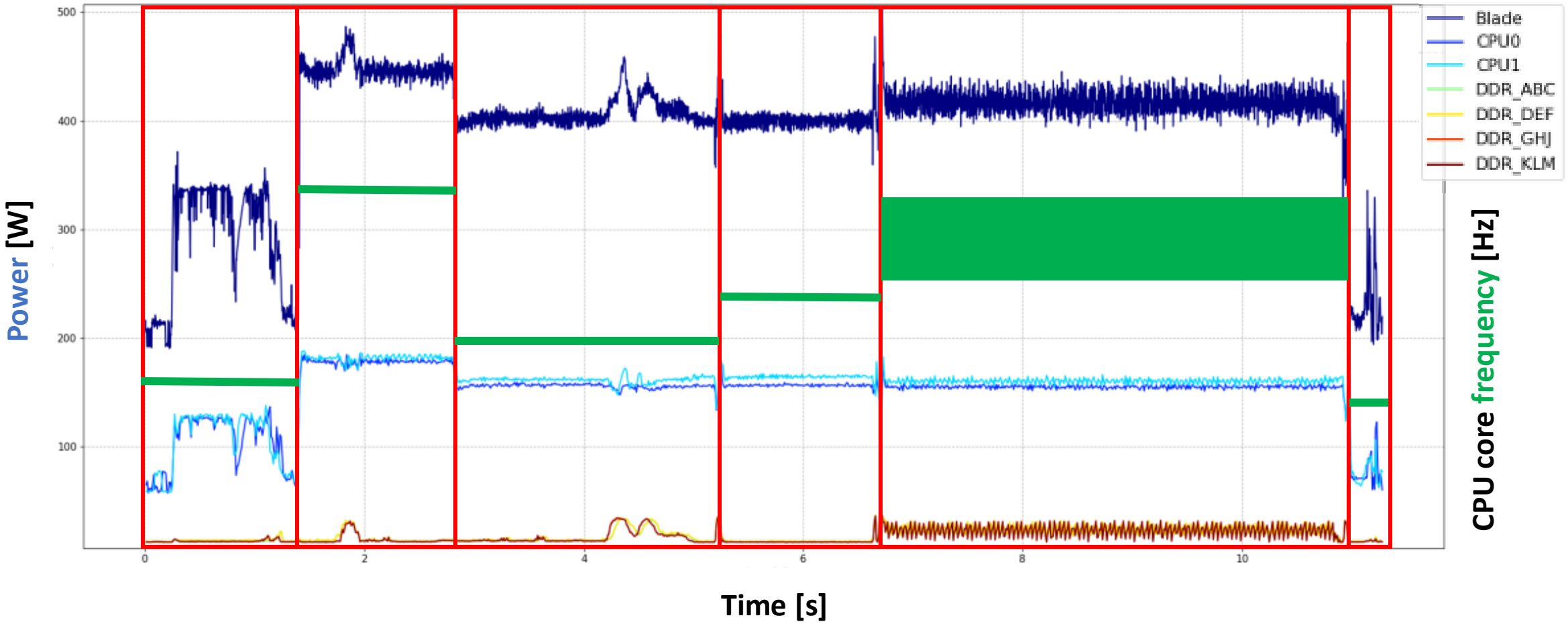
AUTOMATIC INSTRUMENTATION



AUTOMATIC INSTRUMENTATION

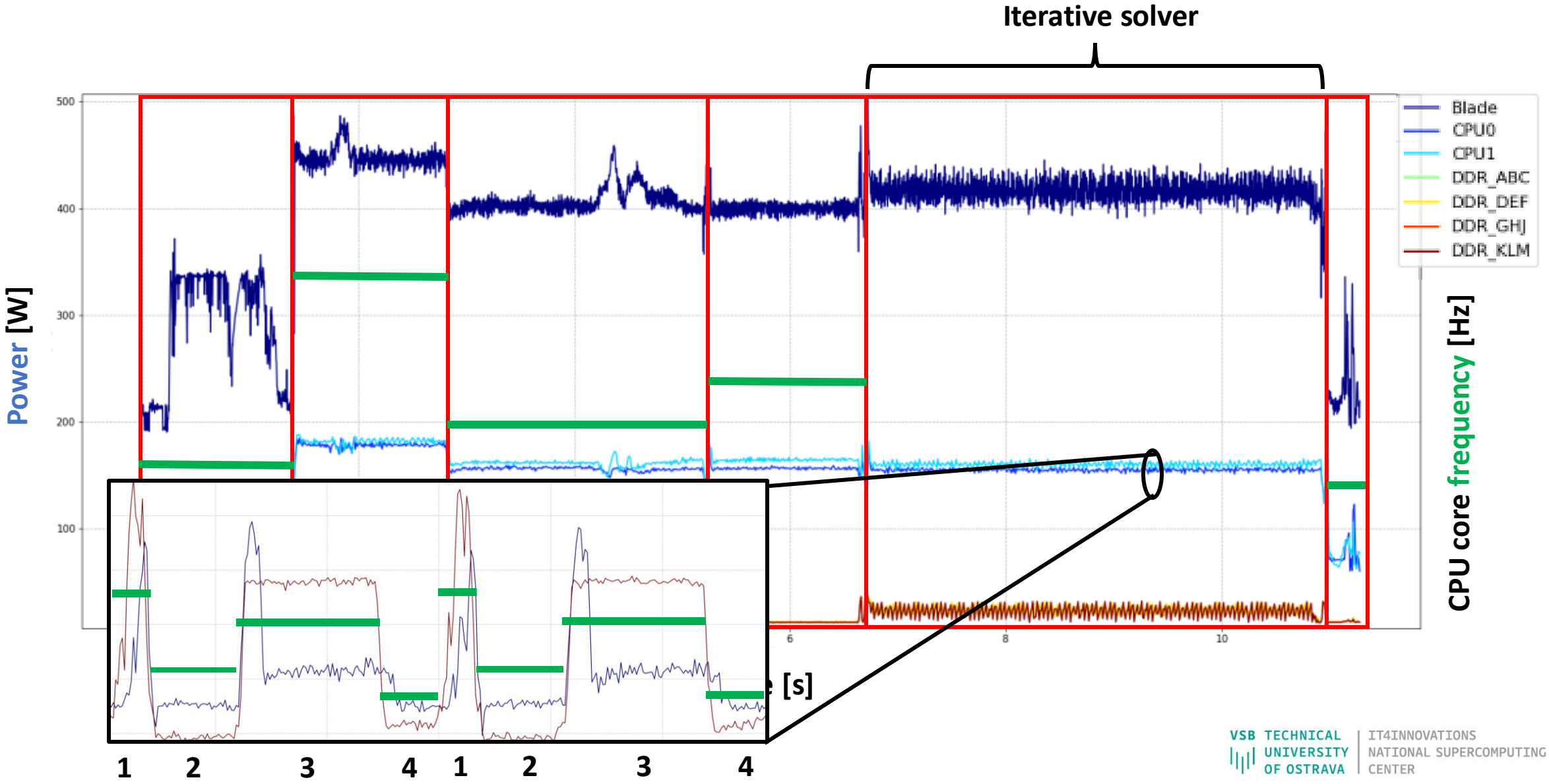


AUTOMATIC CONFIG. IDENTIFICATION



- Currently implemented exhaustive search, particle swarm optimization, genetic algorithm, and sequential frequency downscaling

EXTREME GRANULARITY



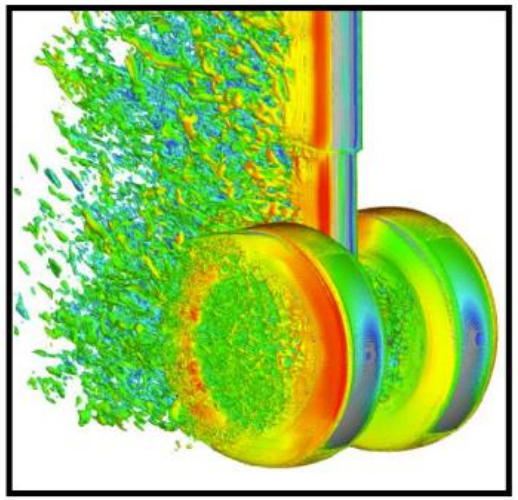
DYNAMIC TUNING

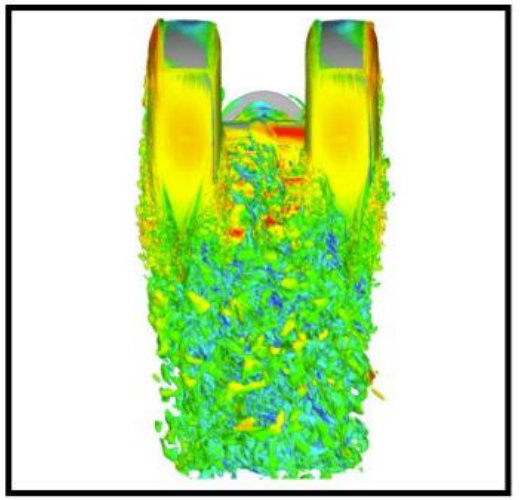
SCALABLE

AIRBUS

ProLB

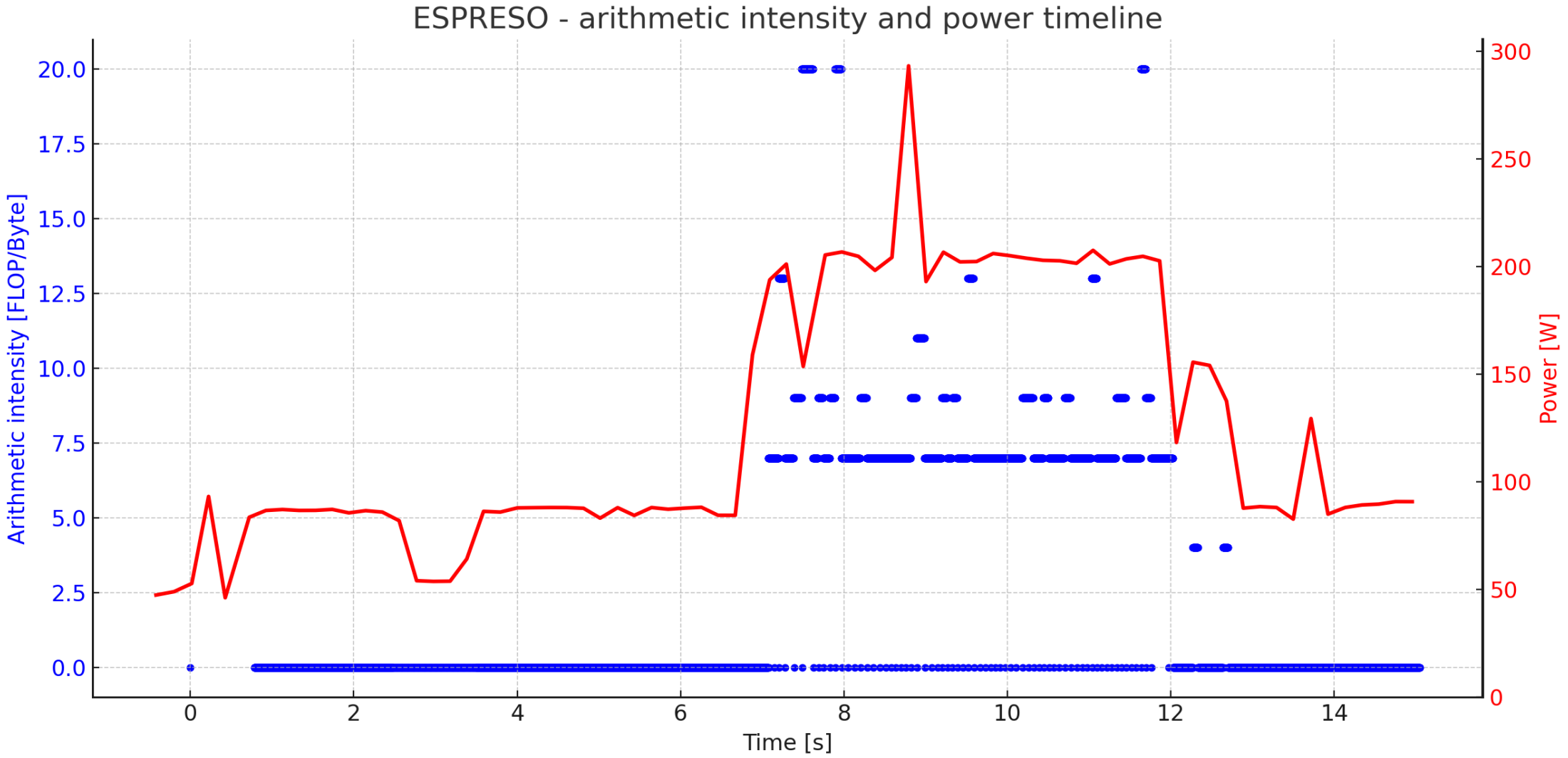
	Default	MERIC tuning no penalty	MERIC tuning limited penalty
Runtime [s]	1797.9	1807.1	1871.1
Energy consumption [kJ]	3102.3	2726.7	2496.71
Solver energy efficiency [MLups/W]	0.054	0.056	0.056
Runtime extension [%]	-	**0.5**	**4.1**
Energy savings [%]	-	**12.1**	**19.5**



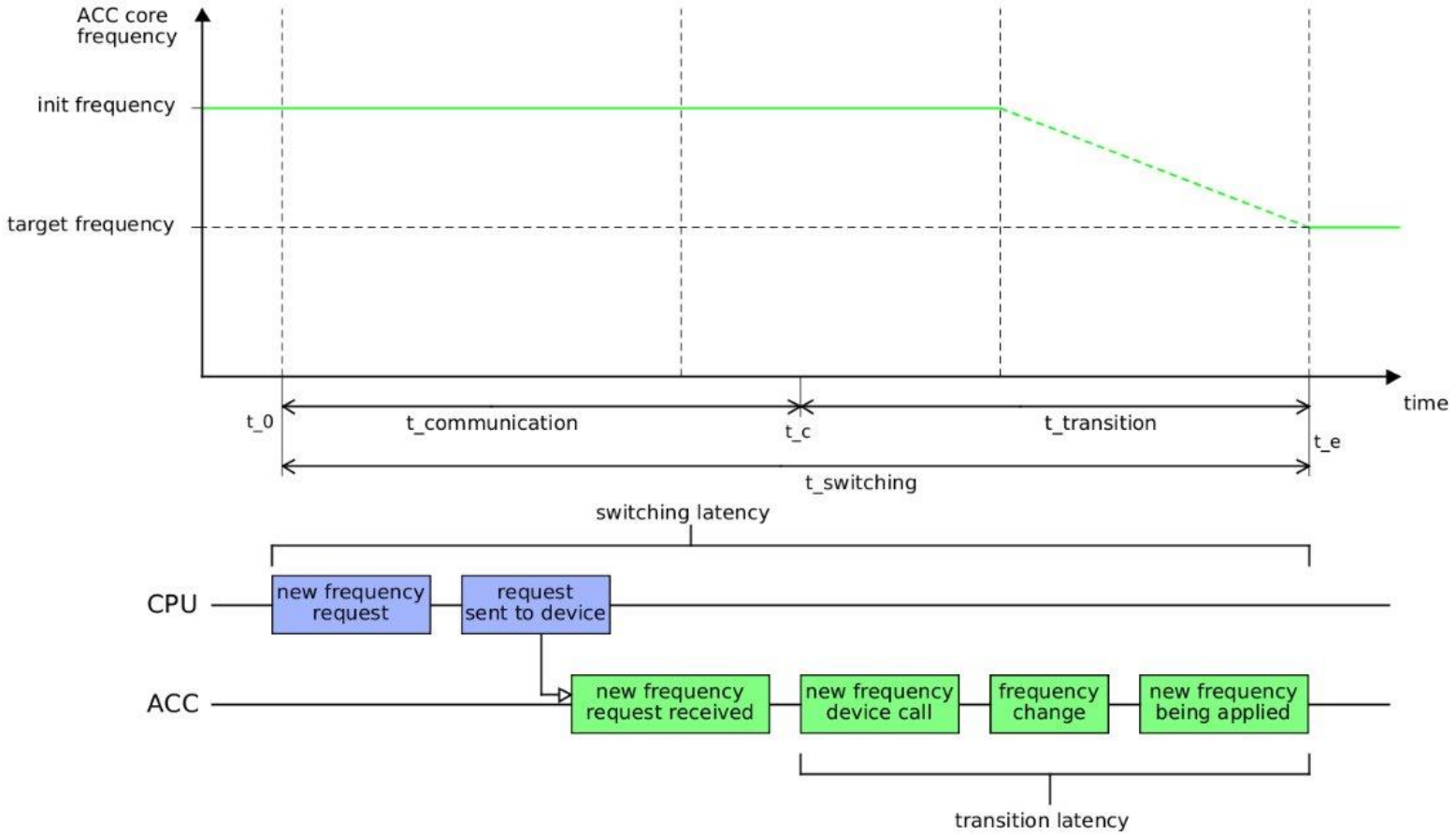


VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

SAMPLING-BASED GPU TUNING



GPU FREQUENCY CHANGE



Visualization of CPU to ACC communication while issuing the ACC frequency change request.
The dashed line shows the frequency change.

GPU SWITCHING LATENCY

A100 SXM-4 Switching latency [ms] (worst case)

Initiabl Frequency [MHz] \ Target Frequency [MHz]	705	750	795	840	885	930	975	1020	1065	1095	1125	1170	1215	1260	1305	1350	1395	1410
705		13.175	15.740	14.435	12.061	15.706	15.880	14.954	16.162	16.279	16.582	14.950	16.275	16.782	15.876	16.305	16.759	16.851
750	19.790		12.689	12.073	13.063	14.043	14.037	14.162	15.547	15.589	14.270	14.729	15.528	15.195	15.410	15.696	16.865	15.383
795	20.698	18.372		14.154	14.144	14.313	14.301	12.572	14.997	10.054	15.277	12.366	15.717	15.260	15.619	15.831	15.569	15.838
840	18.588	19.666	18.922		13.130	12.576	12.923	13.448	14.457	14.757	14.488	14.091	9.693	13.406	13.187	14.686	13.736	14.740
885	19.660	19.453	20.972	21.123		13.403	13.735	11.328	14.856	13.929	14.496	15.637	15.110	13.537	14.652	14.344	13.143	15.410
930	20.130	21.858	20.694	20.033	20.991		13.321	13.966	15.547	14.956	14.340	14.789	15.164	13.607	14.390	13.669	14.928	14.978
975	19.890	19.999	20.139	21.371	21.196	21.185		12.480	14.553	14.272	14.254	14.493	15.644	13.642	13.777	13.924	15.153	15.644
1020	19.825	19.852	11.992	18.923	20.218	21.078	18.704		13.448	12.906	13.938	14.039	11.948	13.483	13.810	13.708	12.957	14.497
1065	20.542	21.079	19.891	22.176	18.314	11.216	9.917	14.226		13.000	13.569	13.678	14.371	14.391	14.385	13.876	13.945	14.274
1095	21.411	21.017	21.501	19.532	21.086	20.565	21.416	14.637	13.064		14.303	13.991	14.962	13.886	15.383	13.953	14.524	14.111
1125	21.134	22.471	22.057	21.164	10.792	12.393	11.179	15.783	14.083	11.132		13.668	14.833	14.449	13.508	14.439	14.162	14.023
1170	20.967	12.648	20.893	20.764	21.025	19.979	21.309	13.454	14.887	14.624	13.591		14.747	14.234	14.215	14.868	14.333	15.207
1215	19.833	19.618	13.293	21.047	19.976	19.494	21.020	15.298	12.809	13.921	13.677	13.047		11.520	11.407	11.988	13.986	11.485
1260	22.285	20.781	21.693	21.404	21.510	20.645	21.647	14.452	13.653	16.554	13.022	14.958	14.675		14.400	12.241	13.589	15.025
1305	22.100	20.459	20.336	20.463	20.332	20.448	20.859	13.478	13.443	14.610	13.065	14.354	13.984	14.197		14.075	12.431	14.800
1350	21.173	20.806	22.507	20.819	9.645	20.300	21.699	13.960	12.734	12.771	17.239	15.821	14.039	9.132	14.373		13.360	14.307
1395	20.659	20.619	11.676	20.246	19.123	20.544	21.471	13.246	14.317	13.682	12.892	13.031	12.439	7.825	8.306	13.565		12.690
1410	21.291	19.838	20.103	19.717	20.079	21.033	20.266	16.520	14.720	13.996	14.013	14.304	14.118	13.544	14.416	13.520	14.931	

Velicka et al "Methodology for GPU Frequency Switching Latency Measurement" (2025)
https://arxiv.org/abs/2502.20075

GPU SWITCHING LATENCY

RTX Quadro 6000

14.1 - 350.4 ms (worst case)

A100 SXM-4

7.8 - 22.5 ms (worst case)

GH200

5.6 - 477.3 ms (worst case)

Target frequency [MHz]

Initial freqquency [MHz]	750	810	930	990	1050	1110	1170	1290	1350	1410	1440	1470	1560	1650
750		14.757	15.520	240.940	140.414	139.973	140.372	142.028	156.513	155.058	140.038	140.314	22.140	39.474
810	20.958		14.581	239.662	135.973	135.871	136.860	136.733	136.313	135.954	137.328	153.091	26.279	34.749
930	21.215	20.533		350.436	136.667	137.396	137.767	137.394	137.618	136.853	137.087	137.220	21.537	27.449
990	23.249	23.276	237.908		150.736	135.646	137.831	137.386	137.730	138.337	137.442	138.543	24.165	18.963
1050	21.991	22.009	20.750	236.168		146.891	136.752	136.846	137.179	136.717	137.396	136.715	24.080	19.163
1110	21.627	21.657	237.367	237.798	136.466		135.531	137.000	135.464	136.386	137.681	136.025	134.215	19.291
1170	19.987	20.790	21.325	237.326	133.325	136.582		136.137	136.194	134.668	136.909	137.415	26.018	19.245
1290	21.082	19.404	236.920	237.140	136.657	137.103	136.332		136.015	135.909	19.576	137.384	31.490	19.289
1350	21.431	20.770	237.051	238.319	136.317	136.280	137.502	135.760		127.195	135.310	136.900	133.306	19.741
1410	21.221	21.066	21.438	238.073	65.695	136.299	136.986	135.595	135.130		135.783	136.311	134.966	19.898
1440	20.706	20.869	237.398	237.945	137.192	137.319	136.778	151.153	136.588	136.035		133.685	126.084	19.185
1470	21.088	23.748	20.072	136.930	137.490	137.056	138.221	135.844	135.410	136.260	187.963		125.804	20.022
1560	21.751	20.837	238.556	136.907	135.078	31.110	136.955	136.212	136.250	133.520	169.922	127.991		20.036
1650	21.356	20.530	20.947	237.801	136.664	135.688	135.312	134.628	135.600	135.463	190.916	127.308	14.135	

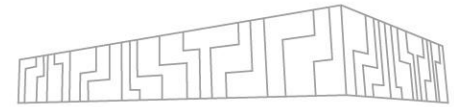
Target Frequency [MHz]

Initiabl Frequency [MHz]	705	750	795	840	885	930	975	1020	1065	1095	1125	1170	1215	1260	1305	1350	1395	1410
705		13.175	15.740	14.435	12.061	15.706	15.880	14.954	16.162	16.279	16.582	14.950	16.275	16.782	15.876	16.305	16.759	16.851
750	19.790		12.689	12.073	13.063	14.043	14.037	14.162	15.547	15.589	14.270	14.729	15.528	15.195	15.410	15.696	16.865	15.383
795	20.698	18.372		14.154	14.144	14.313	14.301	12.572	14.997	10.054	15.277	12.366	15.717	15.260	15.619	15.831	15.569	15.838
840	18.588	19.666	18.922		13.130	12.576	12.923	13.448	14.457	14.757	14.488	14.091	9.693	13.406	13.187	14.686	13.736	14.740
885	19.660	19.453	20.972	21.123		13.403	13.735	11.328	14.856	13.929	14.496	15.637	15.110	13.537	14.652	14.344	13.143	15.410
930	20.130	21.858	20.694	20.033	20.991		13.321	13.966	15.547	14.956	14.340	14.789	15.164	13.607	14.390	13.669	14.928	14.978
975	19.890	19.999	20.139	21.371	21.196	21.185		12.480	14.553	14.272	14.254	14.493	15.644	13.642	13.777	13.924	15.153	15.644
1020	19.825	19.852	11.992	18.923	20.218	21.078	18.704		13.448	12.906	13.938	14.039	11.948	13.483	13.810	13.708	12.957	14.497
1065	20.542	21.079	19.891	22.176	18.314	11.216	9.917	14.226		13.000	13.569	13.678	14.371	14.391	14.385	13.876	13.945	14.274
1095	21.411	21.017	21.501	19.532	21.086	20.565	21.416	14.637	13.064		14.303	13.991	14.962	13.886	15.383	13.953	14.524	14.111
1125	21.134	22.471	22.057	21.164	10.792	12.393	11.179	15.783	14.083	11.132		13.668	14.833	14.449	13.508	14.439	14.162	14.023
1170	20.967	12.648	20.893	20.764	21.025	19.979	21.309	13.454	14.887	14.624	13.591		14.747	14.234	14.215	14.868	14.333	15.207
1215	19.833	19.618	13.293	21.047	19.976	19.494	21.020	15.298	12.809	13.921	13.677	13.047		11.520	11.407	11.988	13.986	11.485
1260	22.285	20.781	21.693	21.404	21.510	20.645	21.647	14.452	13.653	16.554	13.022	14.958	14.675		14.400	12.241	13.589	15.025
1305	22.100	20.459	20.336	20.463	20.332	20.448	20.859	13.478	13.443	14.610	13.065	14.354	13.984	14.197		14.075	12.431	14.800
1350	21.173	20.806	22.507	20.819	9.645	20.300	21.699	13.960	12.734	12.771	17.239	15.821	14.039	9.132	14.373		13.360	14.307
1395	20.659	20.619	11.676	20.246	19.123	20.544	21.471	13.246	14.317	13.682	12.892	13.031	12.439	7.825	8.306	13.565		12.690
1410	21.291	19.838	20.103	19.717	20.079	21.033	20.266	16.520	14.720	13.996	14.013	14.304	14.118	13.544	14.416	13.520	14.931	

Target Frequency [MHz]

Initiabl Frequency [MHz]	705	795	885	975	1095	1170	1260	1275	1290	1350	1410	1500	1665	1770	1830	1875	1920	1980
705		12.370	24.447	10.810	12.011	11.304	245.369	260.804	10.902	12.913	18.105	18.417	17.168	24.686	22.362	10.648	261.650	22.346
795	9.999		18.865	20.843	9.912	9.620	245.403	201.442	9.711	14.533	22.467	22.072	18.362	17.251	15.133	12.962	246.805	22.572
885	12.275	15.174		16.235	11.466	15.884	245.967	256.423	15.500	12.634	14.566	15.200	13.015	16.203	13.015	14.424	294.722	32.591
975	22.044	8.578	23.053		10.730	13.809	246.252	262.165	24.325	21.214	11.174	22.791	13.539	18.346	10.954	16.261	290.120	24.026
1095	13.463	22.892	10.898	15.285		13.977	477.318	263.368	10.369	14.203	11.363	12.715	11.490	15.582	11.562	10.375	302.414	19.479
1170	22.652	14.539	18.161	13.348	12.661		245.255	261.746	20.976	21.091	21.628	12.720	12.777	22.192	17.463	22.192	289.421	20.285
1260	21.902	17.604	18.593	17.761	20.199	18.610		17.852	20.461	19.380	16.872	11.227	18.456	18.743	17.533	19.690	304.845	14.879
1275	23.206	20.660	22.838	20.896	22.629	23.195	24.814		23.143	20.280	15.901	22.285	20.655	19.650	21.766	20.122	303.514	13.227
1290	9.704	20.108	20.516	8.661	19.967	21.387	11.955	9.826		10.780	16.712	18.380	10.740	17.160	146.854	25.762	306.059	174.150
1350	21.346	20.448	21.407	15.290	22.037	16.080	281.823	270.188	21.962		5.572	11.152	20.952	19.520	18.065	19.853	303.097	20.524
1410	7.144	10.168	17.922	11.711	7.708	10.455	289.860	288.611	40.594	89.338		193.691	207.739	32.955	206.769	201.386	290.660	210.565
1500	20.659	21.880	7.269	12.714	22.527	13.647	214.905	214.531	22.739	13.191	9.376		12.751	19.202	11.235	20.709	302.169	12.840
1665	12.943	21.323	21.112	6.836	12.932	7.181	307.031	294.634	16.823	10.449	20.144	24.922		16.539	14.068	14.518	255.089	24.521
1770	10.340	23.587	16.131	18.074	23.233	20.063	113.783	302.975	37.616	43.385	20.048	203.689	203.899		85.533	159.767	301.611	206.412
1830	10.835	8.987	13.653	13.675	10.687	15.608	464.768	23.398	15.829	6.459	14.277	20.801	16.805	10.652		16.332	282.039	14.175
1875	20.748	20.862	21.918	23.699	21.647	23.388	16.524	450.205	24.392	14.387	12.933	12.984	9.518	20.437	11.159		274.134	16.785
1920	15.406	11.156	13.092	13.803	11.025	22.658	23.297	10.878	11.264	10.358	11.811	20.946	10.871	10.320	13.080	20.142		10.916
1980	18.167	22.732	6.381	10.961	23.524	11.712	8.805	411.839	23.071	11.255	10.035	20.425	11.432	20.486	10.398	20.164	298.295	

CPU vs GPU LATENCY

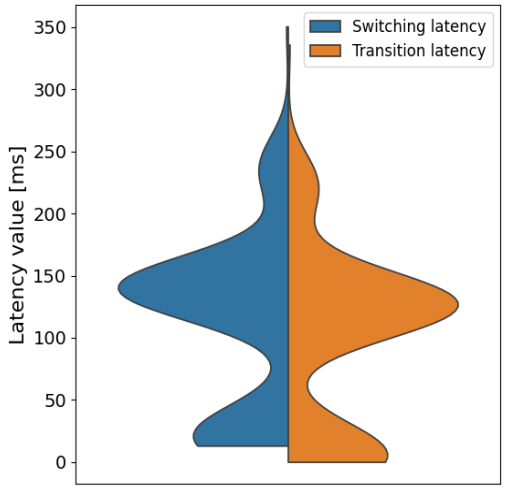


Hardware (Architecture)	Transition latency upper limit
Intel Xeon-SP 6154 (Skylake)	550 us
Intel Xeon X5650 (Westmere)	65 us
Intel Xeon E3-1240 (SandyBridge)	69 us
Intel Core i7-3770 (IvyBridge)	51 us
Intel Core i9-12900 (Alder Lake)	300 us (P-core), 120 us (E-core)
AMD EPYC 7502 (Rome)	1.1 ms

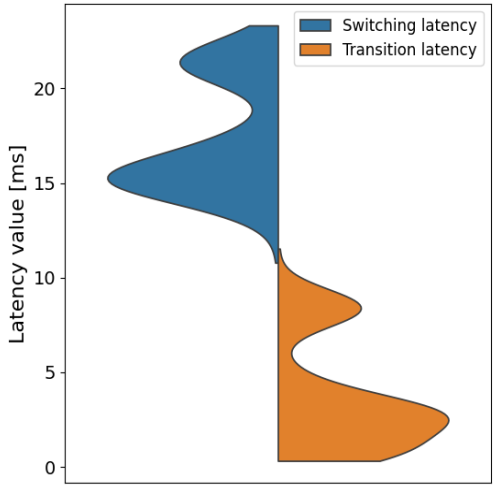
Device (CUDA Architecture)	Switching latency upper limit
RTX Quadro 6000 (Turing)	350 ms
A100 SXM-4 (Ampere)	23 ms
GH200 (Hopper)	477 ms

SWITCHING vs TRANSITION LATENCY

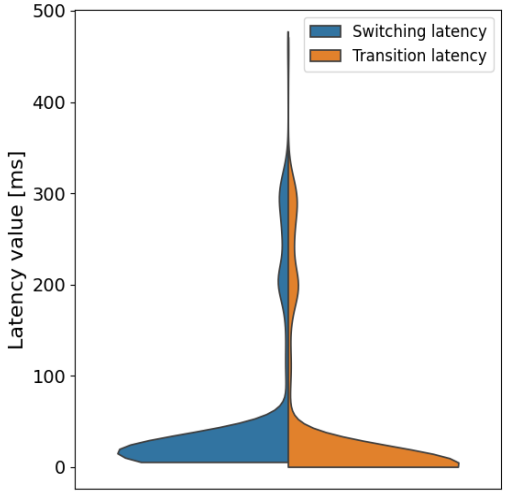
RTX Quadro 6000



A100 SXM-4



GH200



GPU	Switching latency range [ms]	Transition latency range [ms]
RTX Quadro 6000	0.558 - 350.436	0.098 - 335.769
A100 SXM-4 (all four devs)	4.435 - 22.716	0.110 - 11.536
GH200	4.914 - 477.318	0.082 - 471.078

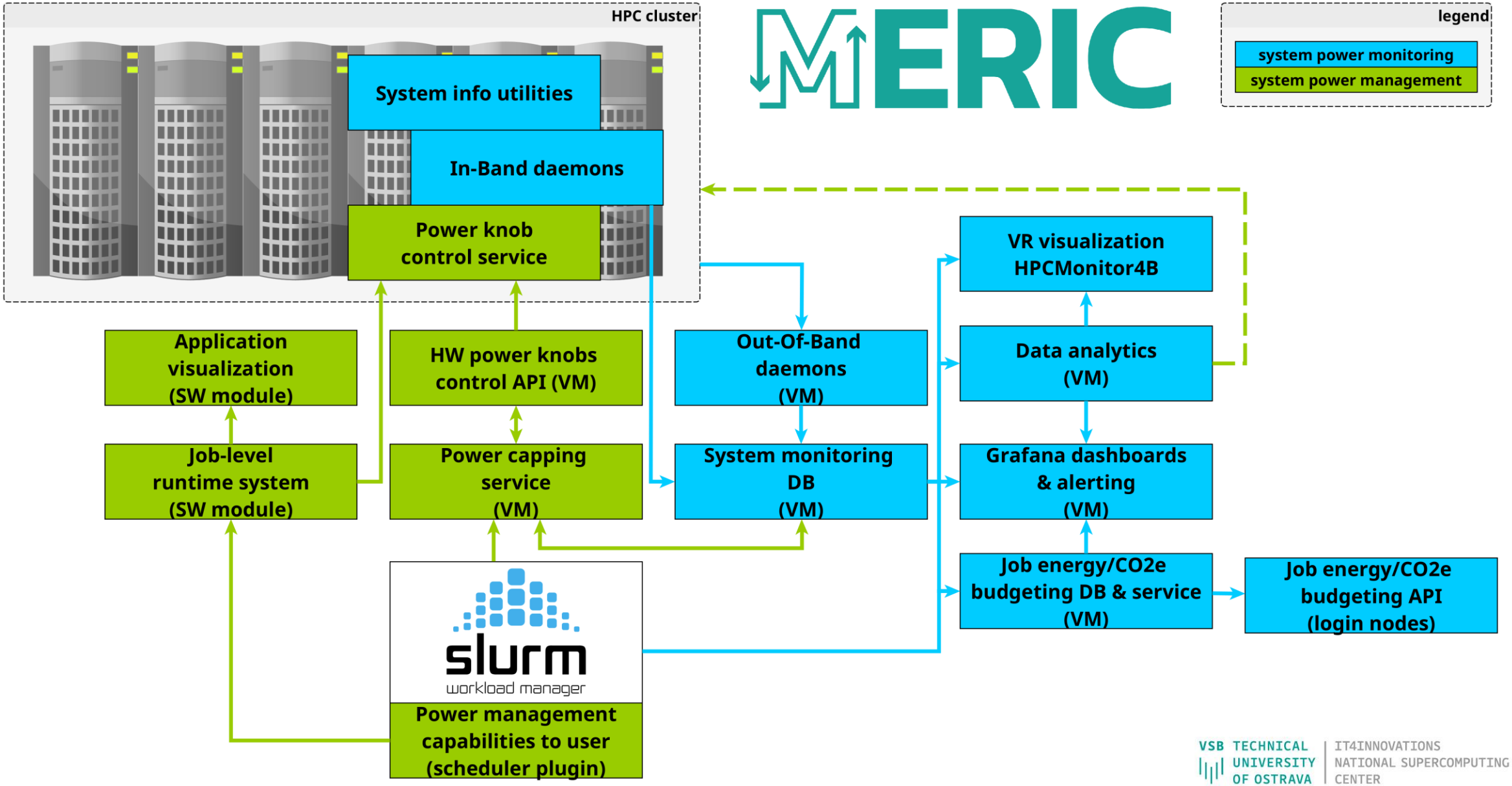
ENERGY EFFICIENCY SERVICES

- How much energy does my application consume? What is its carbon footprint?
- Which parts of the code are power hungry? Does it activate power capping?
- How energy efficient the code is?
- Which hardware platform is the most energy efficient for my code?
- Which parts of the application may give opportunity for energy savings?
- How much energy can be saved by static versus dynamic tuning of power management knobs without impacting application performance? And if the performance penalty is 5%, 10%, … ?

- Does my hardware power/thermal management work as intended?
- When is the capping mechanism a performance-limiting factor?

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

MERIC: ENERGY EFFICIENCY SW SUITE FOR HPC



SEANERGYS
Energy Efficient Exascale

Reduce energy waste & operating costs while maximizing the scientific and industrial benefits for Europe's investments in HPC and AI infrastructures

Design a production-quality SW suite for energy-efficient operation of European HPC systems

- Create a **holistic monitoring infrastructure** and **common data repository for operational HPC/AI data**
- Develop an **advanced AI-based data analytics framework for HPC operational data**
- Implement a **dynamic resource management system** to optimise use of resources and energy efficiency in heterogeneous HPC/AI systems and to enable dynamicity

Validate the SEANERGYS SW suite **in operational environments** and make it available under permissive licenses

Project details

- EuroHPC JU R&D project
- Total grant 32.9 M€
- June 1, 2025 – May 31, 2029

SEANERGYS MODULAR ARCHITECTURE

Interaction of three main modules

Monitoring

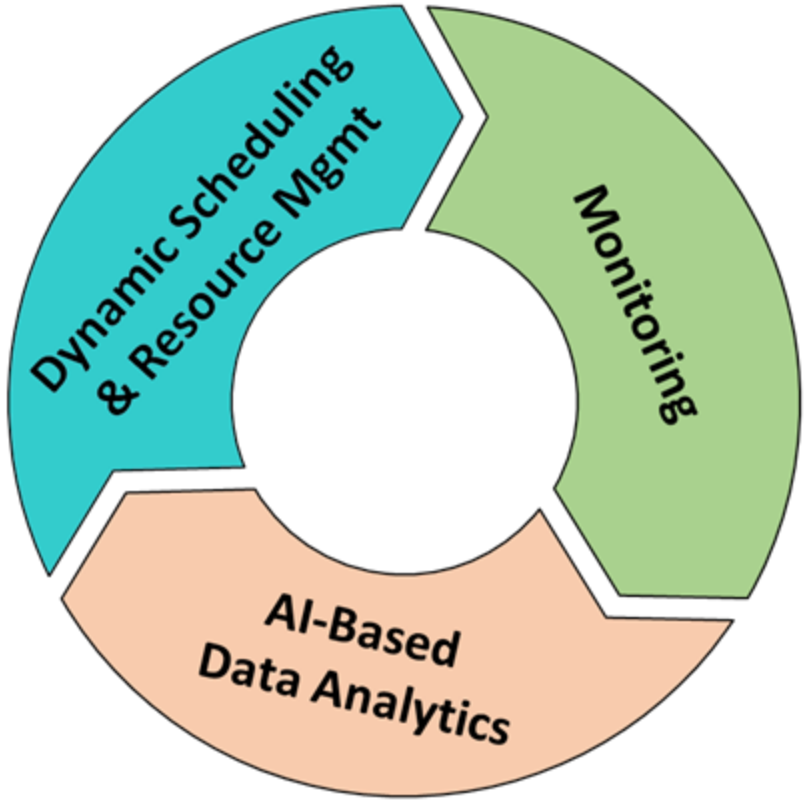
- Scalable system/facility/environment sensor data capture
- Inclusion of non-structured data
- Establishment of a system data plane

AI-based data analytics

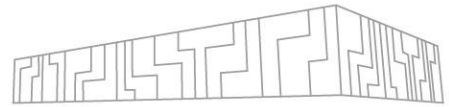
- Automatic workload characterisation and prediction
- Directing scheduling & resource mgmt. decisions
- Actionable feedback to application developers and users

Dynamic scheduling & resource management

- Dynamic control of system & facility operating points
- Exploit dynamicity/malleability
- Right-size moldable workload resource requirements



TEN YEARS OF DEVELOPMENT



Energy-efficiency services

SCALABLE

MAX
DRIVING THE EXASCALE TRANSITION

SPACE

dare

MERIC development

POP

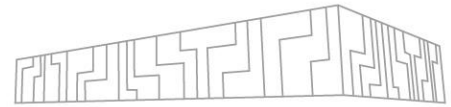
READEX
Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing

SEANERGYS
ENERGY EFFICIENT EXASCALE

e-INFRA CZ

EUPEX
European Pilot for Exascale

INDUSTRIAL PARTNERS



AIRBUS

CS GROUP

RENAULT

SKODA

EXELIZ SOLUTIONS

Teratec

HPC Now!

ParTec MODULAR SUPERCOMPUTING

FUJITSU

EVIDEN

Codasip

SIPEARL

Hewlett Packard Enterprise

AMD

E4 COMPUTER ENGINEERING

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Ondřej Vysocký
meric@it4i.cz

IT4Innovations National Supercomputing Center
VSB – Technical University of Ostrava
Studentská 6231/1B
708 00 Ostrava-Poruba, Czech Republic
www.it4i.cz

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER