VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

ENERGY EFFICIENCY FEATURES OF THE MODERN HPC HARDWARE AND ENERGY CONSUMPTION MEASUREMENT

Ondřej Vysocký
IT4Innovations

June 6, 2023

EURO CZECHIA

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

MINISTRY OF EDUCATION, YOUTH AND SPORTS

ENERGY EFFICIENCY

- How to improve energy efficiency?

ENERGY EFFICIENCY

- How to improve energy efficiency?
- Improve (parallel) efficiency of your code!

ENERGY EFFICIENCY

- How to improve energy efficiency?
- Improve (parallel) efficiency of your code!
- So, we can go home today...

ENERGY EFFICIENCY

- How to improve energy efficiency?
- Improve (parallel) efficiency of your code!
- So, we can go home today... or not?

ENERGY EFFICIENCY

$$EnergyEfficiency = \frac{Rmax}{power}[Flops/W]$$

ENERGY EFFICIENCY

$$EnergyEfficiency = \frac{Rmax}{power} [Flops/W]$$

ENERGY EFFICIENCY

- Why should I improve energy efficiency (and potentially reduce performance) of my application?

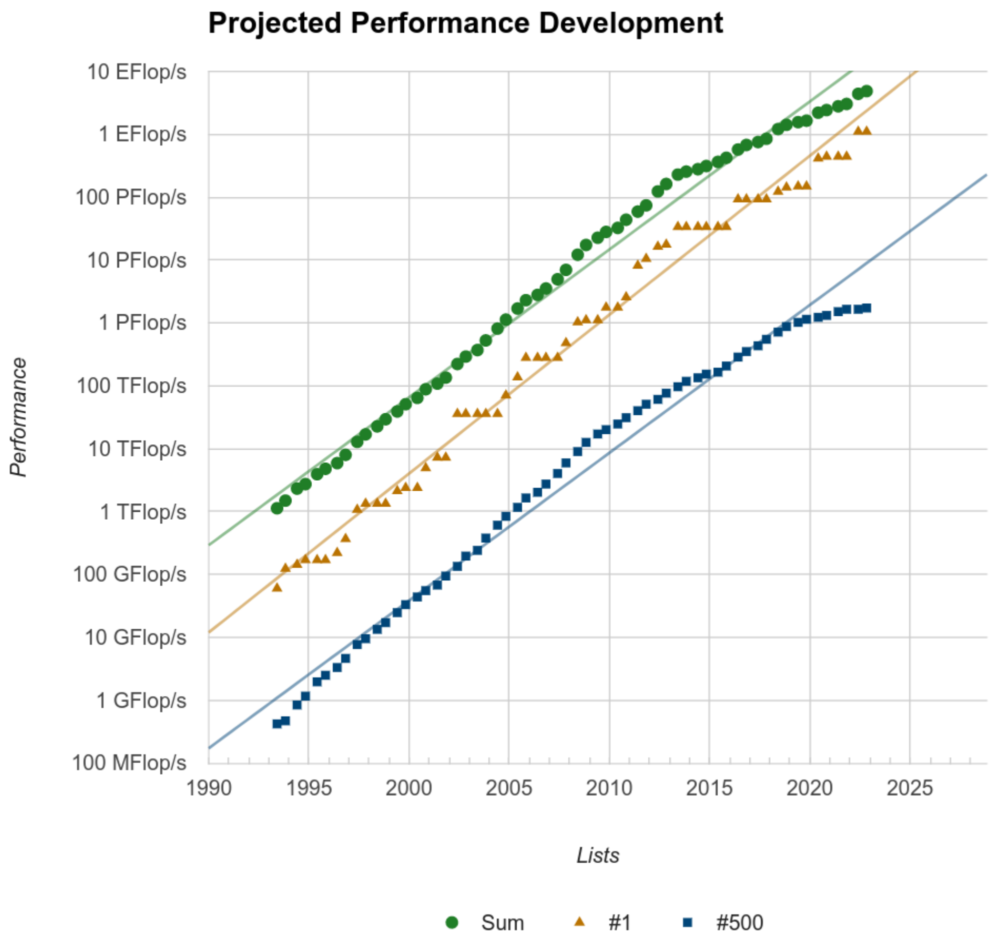
ENERGY EFFICIENCY

- Why should I improve energy efficiency (and potentially reduce performance) of my application?

- **Better energy efficiency does not imply worse performance**

END OF MOOR'S LAW

- Scaling
  - Power wall
  - Target 20 MW power limit for exascale
    - = 50 GFlop/W
    - Soft limit
- General hardware optimised for all possible workloads => silicon area wasted to maximize single thread performance
  - New heterogenous hardware – GPU, FPGA, …
  - Specialized computing units



Top500 11/2022

TOP500

Rmax = Linpack Performance
Rpeak = Theoretical Peak

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**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
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
8	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589
9	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646
10	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61.44	100.68	18,482

TOP 500 The List.

11/2022

TOP500

TOP 500 The List.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)		
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100	**2021**	**x1.9 = 57 MW**
2	**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	**2020**	**x3.3 = 20 MW**
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016	**2022**	**x6 = 34 MW**
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610	**2022**	**x5 = 44 MW**
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096	**2018**	**x8 = 60 MW**
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438	**2018**	**x8 = 123 MW**
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371	**2016**	**x11 = 28 MW**
8	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589	**2021**	**x13 = 34 MW**
9	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646	**2020**	**x10 = 185 MW**
10	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61.44	100.68	18,482	**2018**	

Rmax = Linpack Performance
Rpeak = Theoretical Peak

Exascale goal is

20 MW limit

=

50 GFlops/watts

11/2022

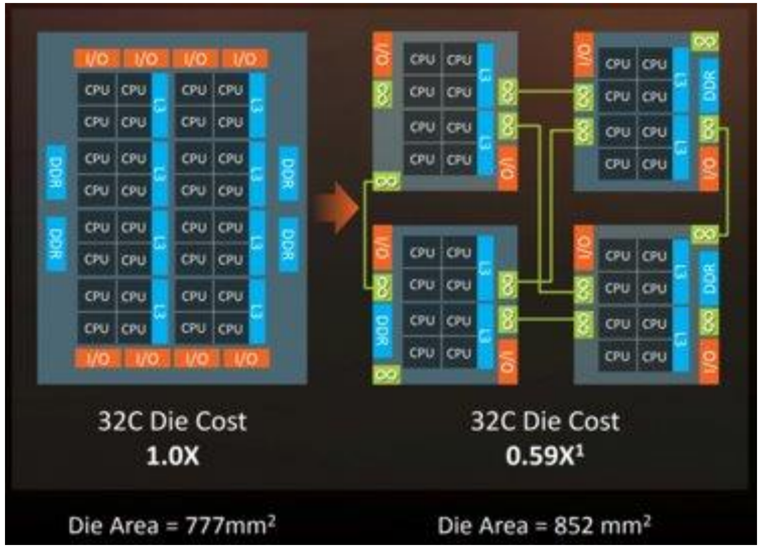
Exascale computing

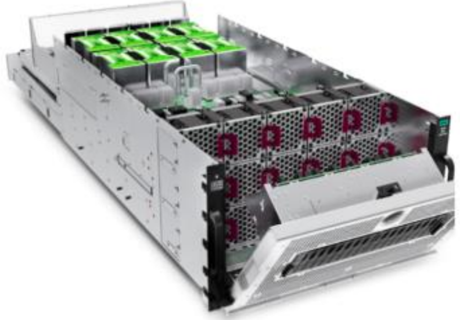
11/2016

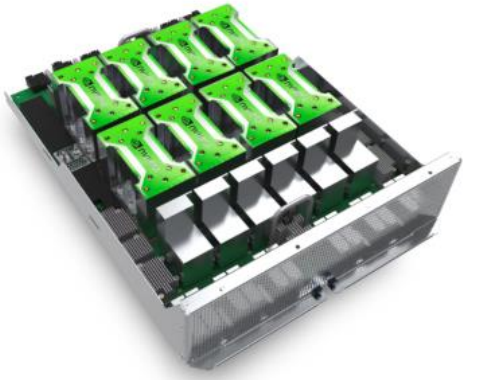
Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371 x8 = 123 MW
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808 x19 = 340 MW
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209 x38 = 310 MW
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890 x50 = 395 MW
5	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939 x37 = 145 MW

HARDWARE TRENDS

- CPUs
  - Rising number of cores
  - Chiplets (tiles)
  - Purpose specific units
    - AI, crypto, matrix calculation
- GPUs
  - Tensor cores
- FPGAs







GREEN500

The GREEN 500

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	255	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States **Nvidia H100**	8,288	2.88	44	65.396
2	34	**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 **AMD MI250X**	120,832	19.20	309	62.684
3	12	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France **AMD MI250X**	319,072	46.10	921	58.021
4	17	**Setonix – GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Pawsey Supercomputing Centre, Kensington, Western Australia Australia **AMD MI250X**	181,248	27.16	477	56.983
5	77	**Dardel GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE KTH - Royal Institute of Technology Sweden **AMD MI250X**	52,864	8.26	146	56.491
6	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States **AMD MI250X**	8,699,904	1,194.00	22,703	52.592
7	3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland **AMD MI250X**	2,220,288	309.10	6,016	51.382
8	483	**amplitUDE (GPU Partition)** - MEGWARE D50DNP, Xeon Platinum 8480+ 56C 2GHz, NVIDIA H100 80GB PCIe, Infiniband NDR, MEGWARE University of Duisburg-Essen Germany **Nvidia H100**	6,768	1.95	38	51.343
9	70	**Goethe-NHR** - Supermicro AS-4124GS-TNR, AMD EPYC 7452 32C 2.35GHz, AMD Instinct MI210 64 GB, Mellanox InfiniBand EDR, MEGWARE / Supermicro Universitaet Frankfurt Germany **MI210**	96,768	9.09	195	46.543
10	187	**ATOS THX.A.B** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos Atos France **Nvidia A100**	25,056	3.50	86	41.411

6/2023

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

ENERGY EFFICIENCY

$$EnergyEfficiency = \frac{Rmax}{power} [Flops/W]$$

Power management and monitoring

ENERGY

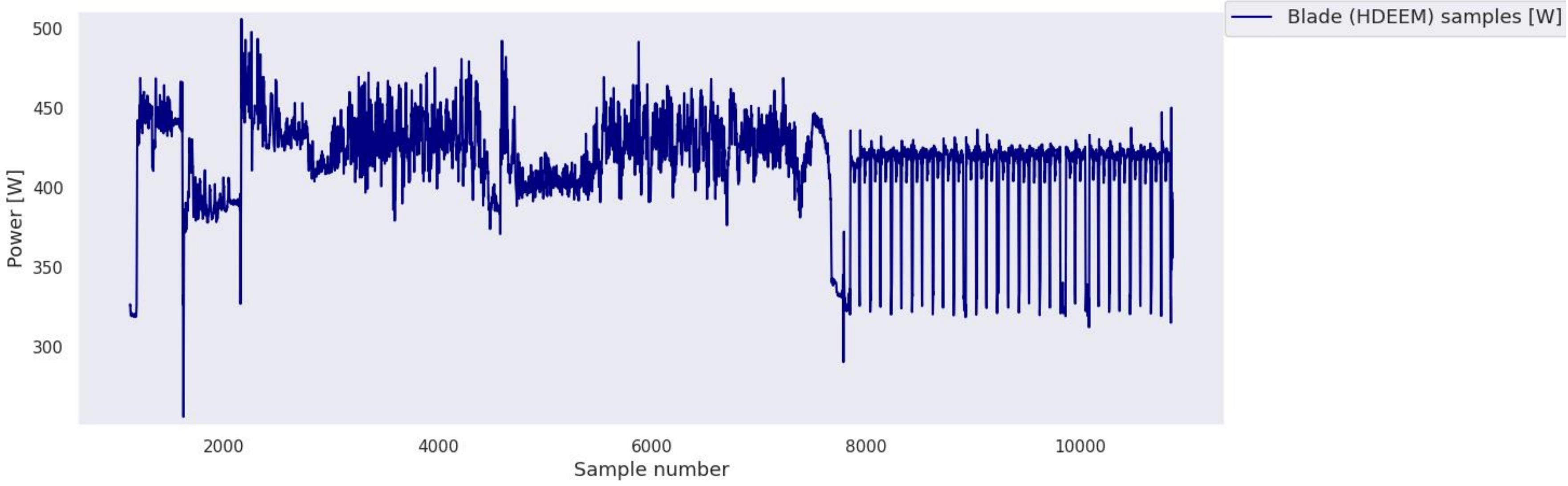
$$Energy = Power \times Time$$

- 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

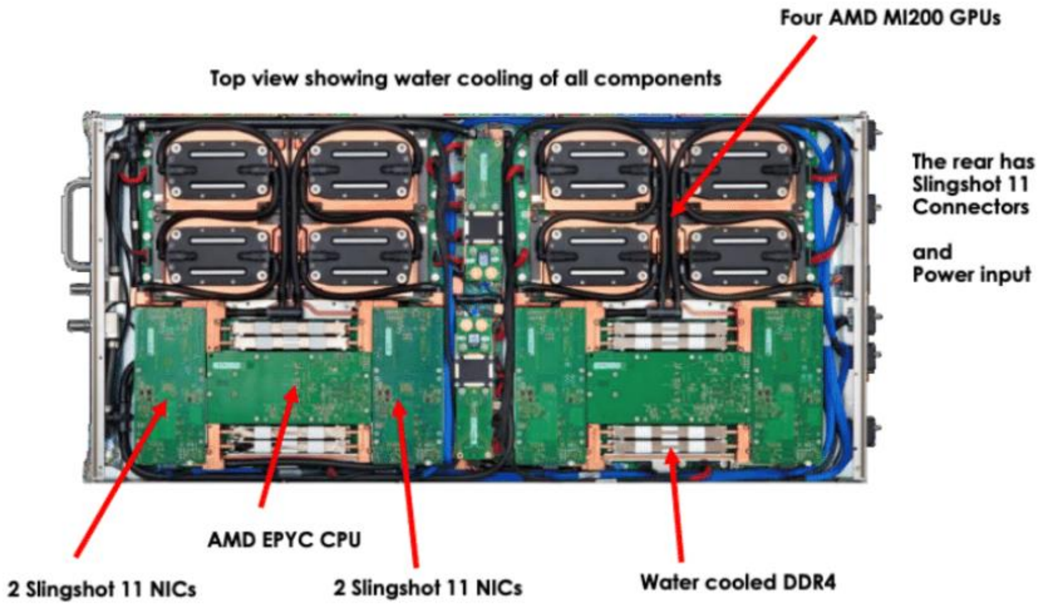
$$Energy = Power \times Time$$

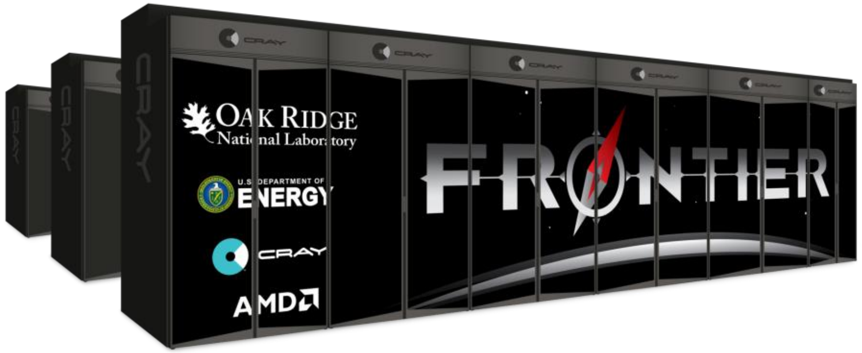
- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J

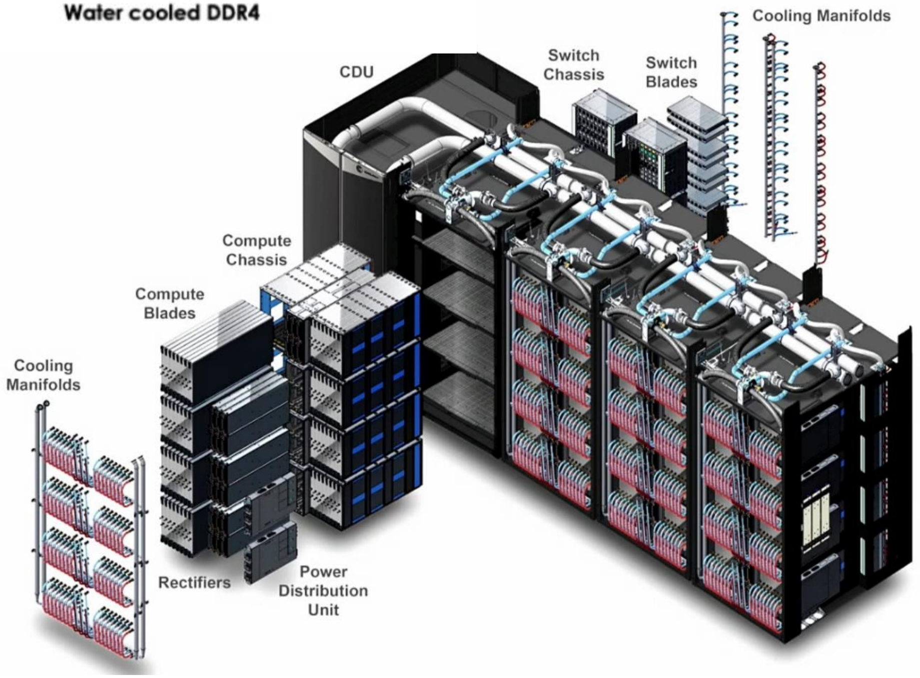
$$Energy(t) = \int_0^t Power(x)\,\mathrm{d}x \approx \frac{\sum_{i=0}^{n} PowerSample_i}{SamplingFrequency}$$

POWER MONITORING

- On node components
  - CPU, GPU, memories, NIC
- Node
- Rack
- Sytem
- Data hall
- Building





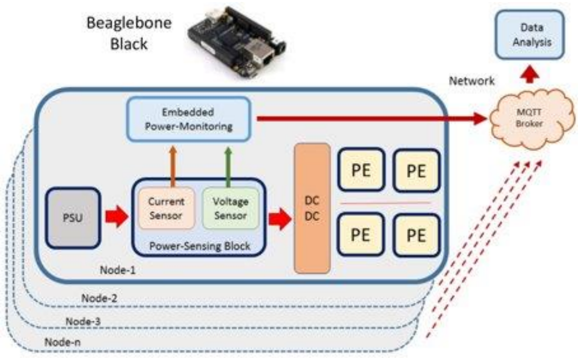


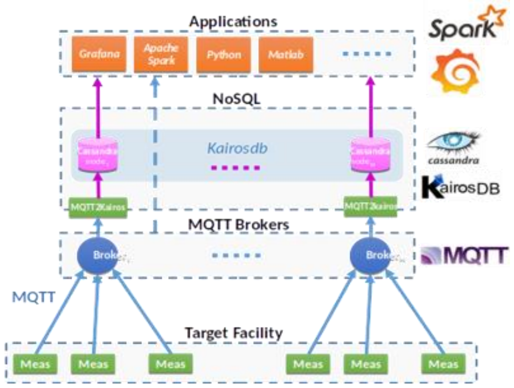
Images of HPE Cray, OLCF Frontier

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*

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

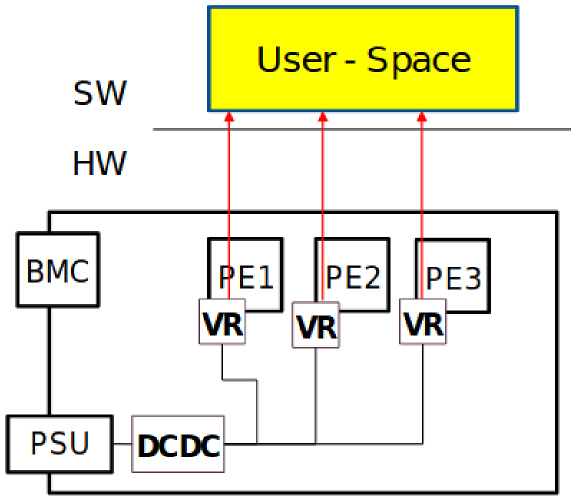




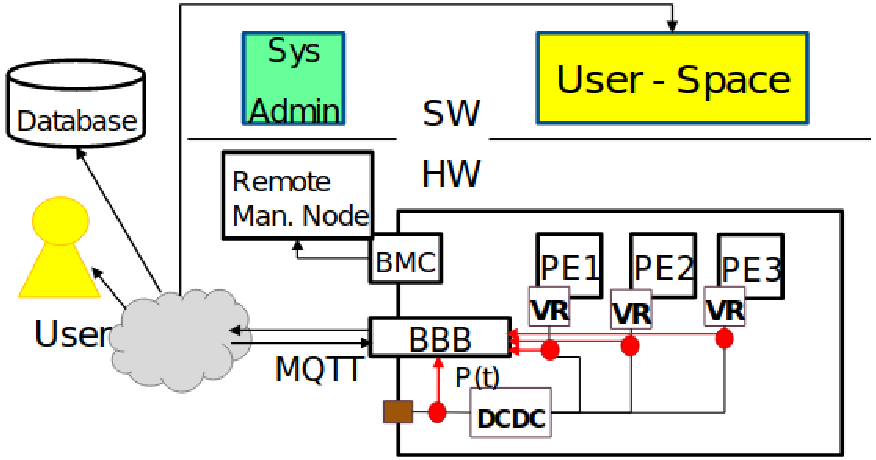
IN- AND OUT-OF-BAND POWER MONITORING

- In-band
  - Vendor dependent
  - HW performance counters
- Out-of-band
  - High overhead of communication -> can be avoided by reading post mortem
  - (usually) fine-grain power measurement
  - Custom sensors

RAPL



DiG



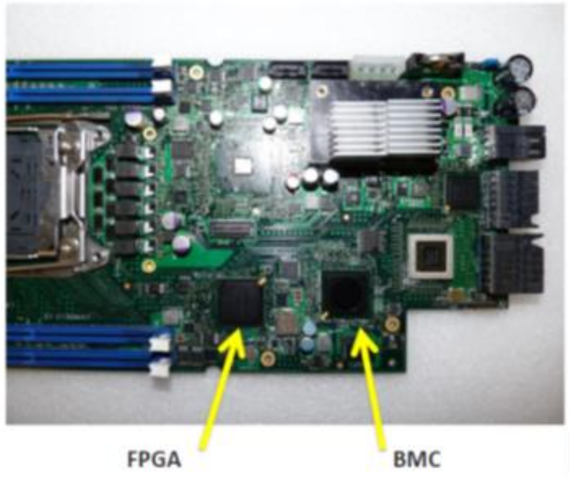
Img source, Antoniu Libri (UNIBO)

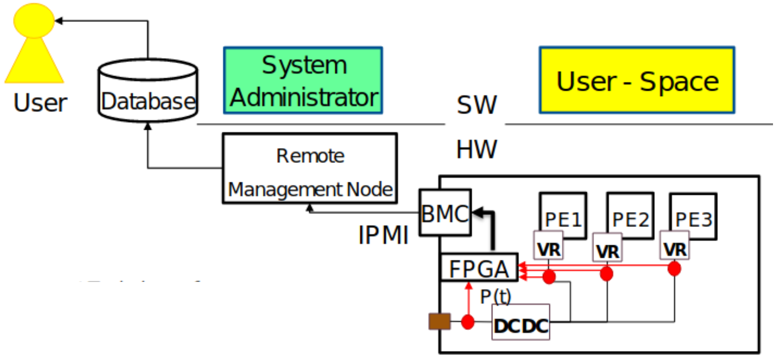
HIGH DEFINITION ENERGY EFFICIENCY MONITORING (HDEEM)

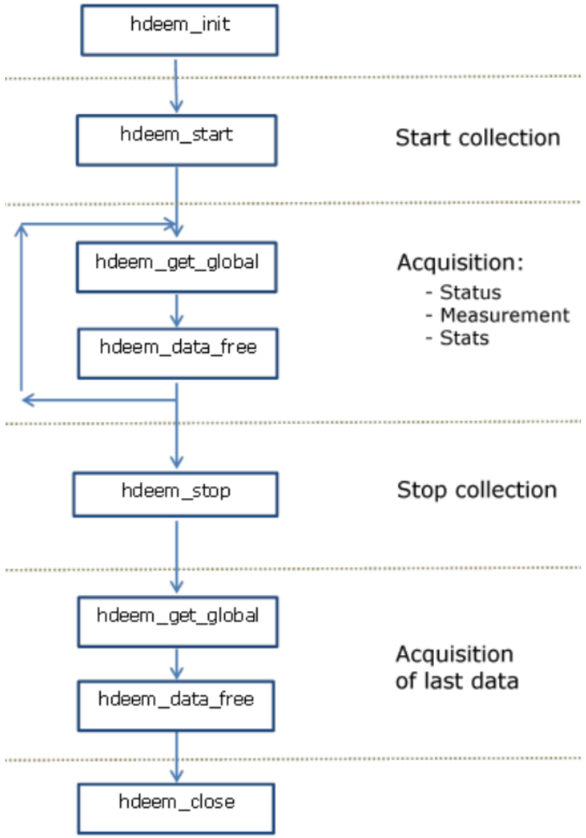
- Bull|Atos technology available for production systems (Bullx B7xx and Bull Sequana)
- On board out-of-band technology for power monitoring

- 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







POWER KNOBS

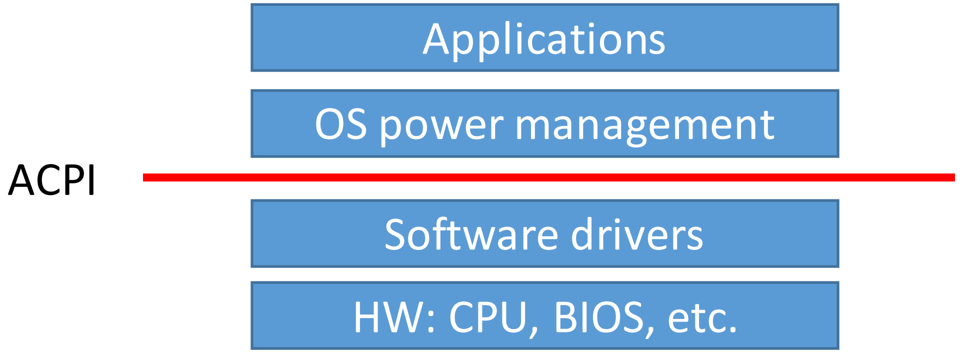
- Intel
  - CPU - core frequency, uncore frequency, power capping
  - ACC (PVC) - GPU frequency, memory frequency, power capping
  - ACC (KNL) - core frequency, power capping
- AMD
  - CPU - core frequency, power capping, Data Fabric frequency
  - ACC - power capping, frequency – system, Data Fabric, display controller, SOC, memory, PCIe
- Nvidia
  - GPU - SM frequency, memory frequency, power capping
- IBM
  - CPU - core frequency, power capping
    + GPU and node power capping
- ARM
  - A64FX - core frequency, FLA (floating-point ops) and EXA (integer ops) pipelines elimination, memory frequency
  - EPI - core frequency, power capping, ???

This list is incomplete

OS POWER MANAGEMENT

- ACPI (Advanced Configuration and Power Interface) is an open industry specification establishes industry-standard interfaces enabling OS-directed configuration, power management, and thermal management of mobile, desktop, and server platforms.
- ACPI defines performance states (P-States)
- P-States correspond to different performance levels that are applied while the processor is actively executing instructions
- Intel CPUs from Haswell architecture provide Voltage regulators per core, so each core has its own P-State

- Scaling driver
  - Acpi-cpufreq, intel_psate, ...
- Scaling governor
  - Performance, powersave, userspace, ondemand, conservative
- Intel hardware P-state

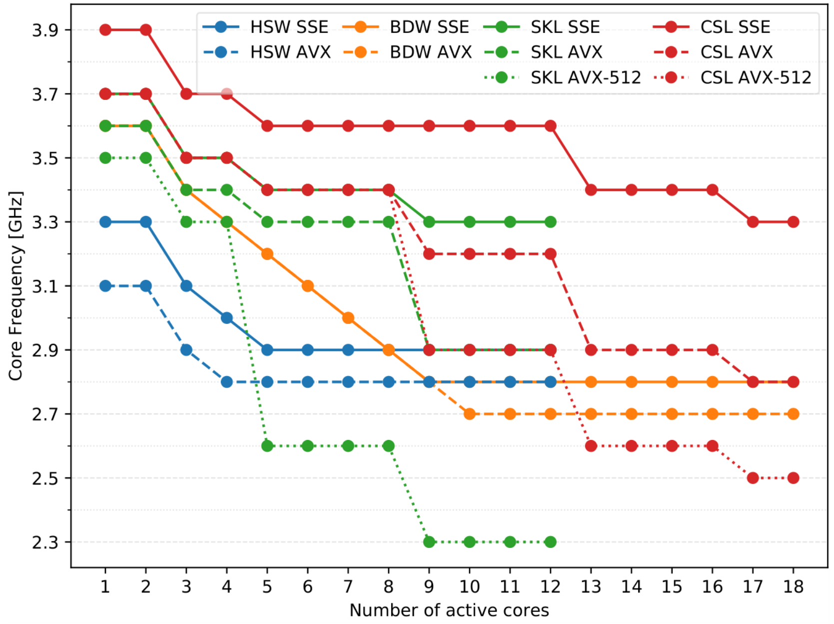


INTEL CPU CORE TURBO FREQUENCY

- Turbo Boost is a technology that opportunistically allows the processor to run faster than the nominal frequency if the CPU is operating below power, temperature and current limits

- There are three different levels of the turbo core frequency based on instruction set – SSE, AVX/AVX2, AVX-512
- The turbo frequency limit also relies on the number of active cores
- Turbo Boost frequency is selected by the firmware of the CPU – no OS control

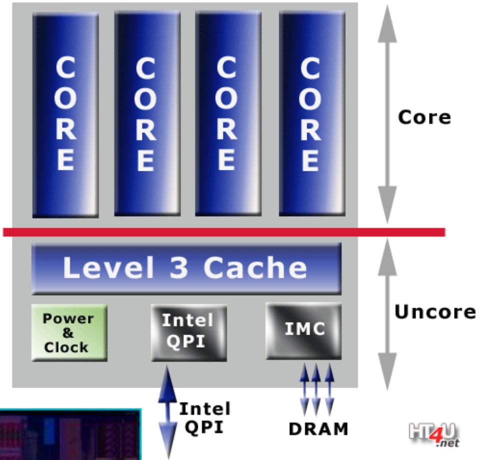
- Be careful when using an islands of AVX instructions, there is always a transition latency



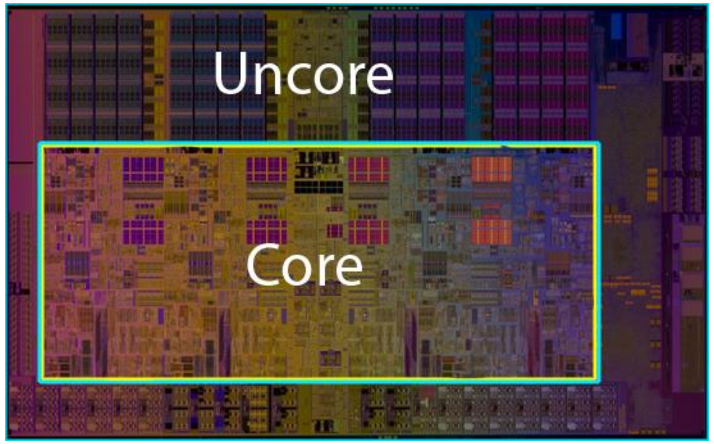
INTEL CPU UNCORE FREQUENCY

- MSR MSR_UNCORE_RATIO_LIMIT (0x620)

- frequency of subsystems in the physical processor package that are shared by multiple processor cores
- last level cache, on-chip ring interconnect or the integrated memory controllers, etc.
- occupies approximately 30 % of a chip area
- specification of the maximum and minimum limit



620H		MSR UNCORE_RATIO_LIMIT	Package	Uncore Ratio Limit (R/W) Out of reset, the min_ratio and max_ratio fields represent the widest possible range of uncore frequencies. Writing to these fields allows software to control the minimum and the maximum frequency that hardware will select.
		63:15		Reserved
		14:8		MIN_RATIO Writing to this field controls the minimum possible ratio of the LLC/Ring.
		7		Reserved
		6:0		MAX_RATIO This field is used to limit the max ratio of the LLC/Ring.



INTEL RUNNING AVERAGE POWER LIMIT (RAPL)

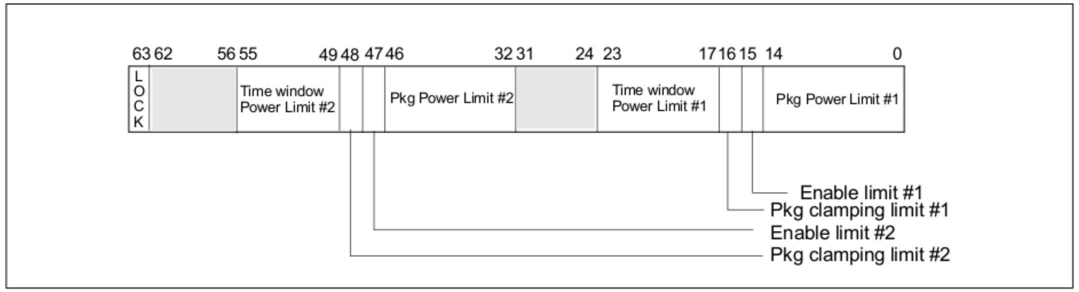
- Sysfs: /sys/devices/virtual/powercap/intel-rapl/intel-rapl:X/intel-rapl:0:Y

- 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

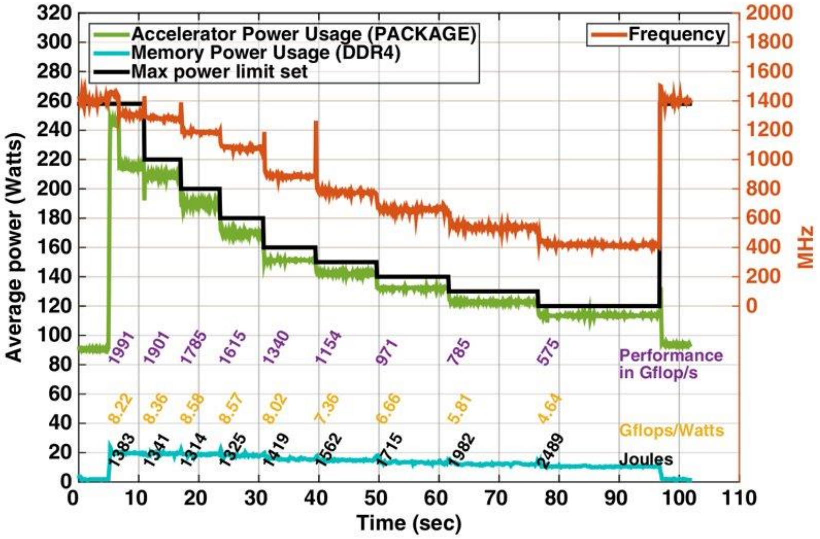
INTEL RUNNING AVERAGE POWER LIMIT (RAPL)

- MSR MSR_PKG_POWER_LIMIT (0x610)

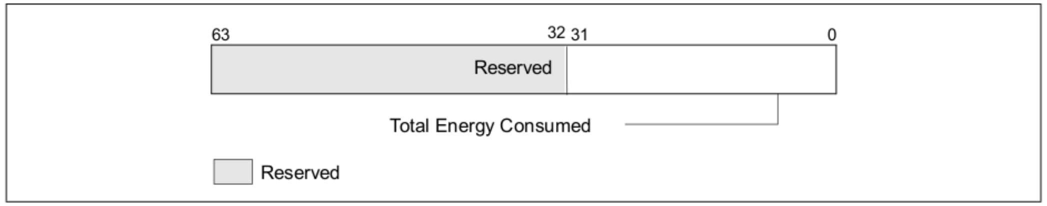


- MSR MSR_RAPL_POWER_UNIT (0x606)
  - Power units
  - Energy status units
  - Time units

- **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)

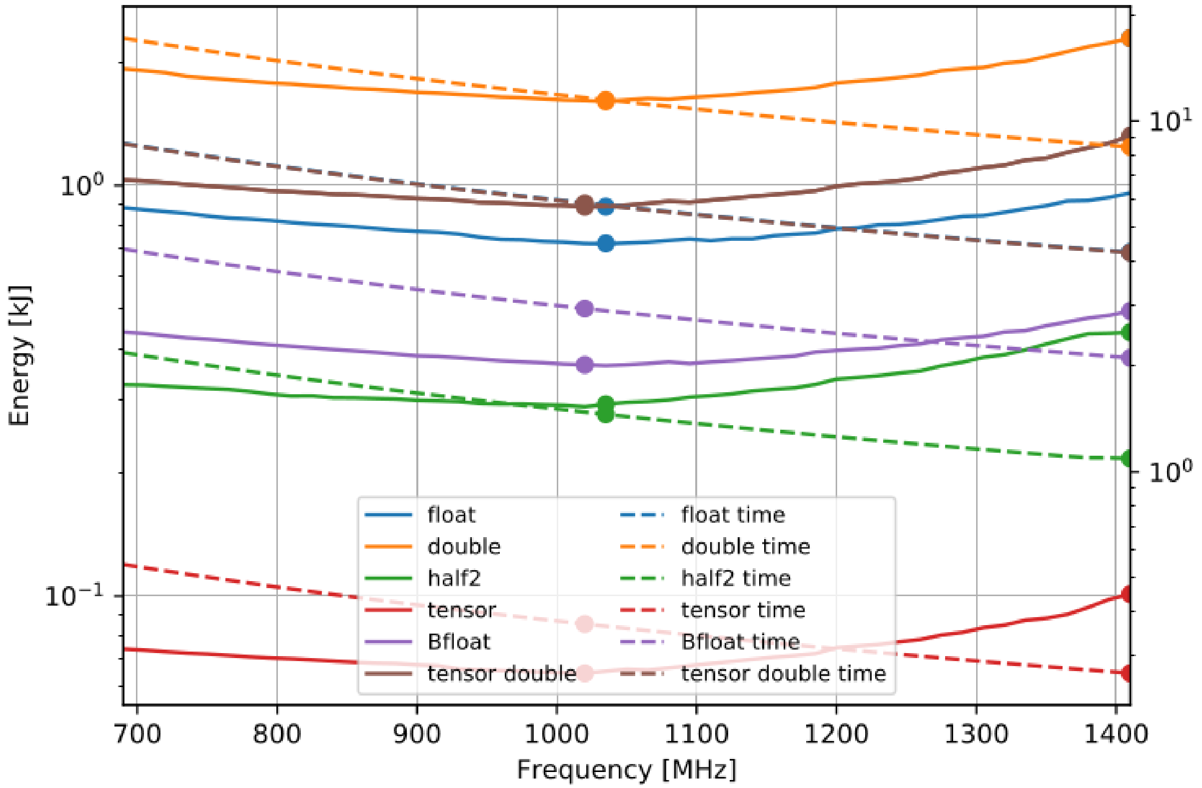


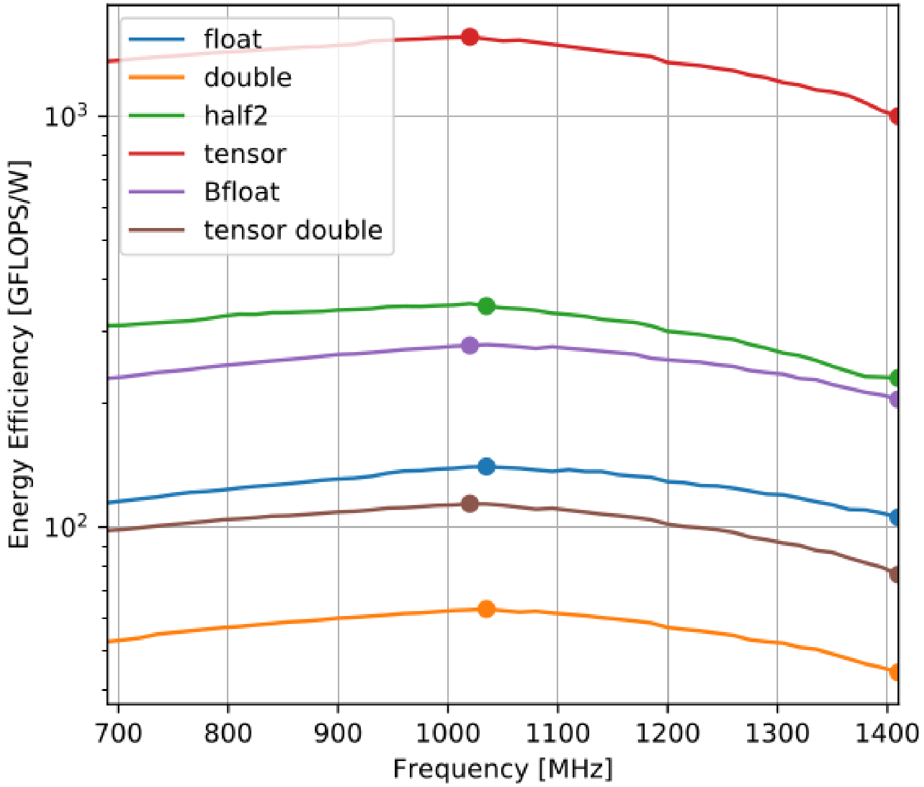
Haidar et al: Investigating power capping toward energy-efficient scientific applications



NVIDIA GPU SM FREQUENCY TUNING

- A100-SXM4
  - 1410 MHz SM max turbo frequency
  - 1095 MHz SM nominal frequency
- Mandelbrot benchmark





Spetko et al: DGX-A100 Face to Face DGX-2—Performance, Power and Thermal Behavior Evaluation.

NVIDIA GPU SM FREQUENCY TUNING

- A100-SXM4
  - 1410 MHz SM max turbo frequency
  - 1095 MHz SM nominal frequency
- Mandelbrot benchmark

	Frequency [MHz]	Time [s]	Time Difference	Energy [J]	Energy Savings	Performance [TFLOPS]	Energy Efficiency [GFLOPS/W]
double	1410	8.43		2285		9.71	35.86
	1035	11.49	136.19%	1601	29.91%	7.13	51.16
float	1410	4.23		958		19.37	85.53
	1035	5.76	136.13%	721	24.76%	14.23	113.68
Bfloat	1410	2.11		494		38.75	165.74
	1035	2.88	136.19%	364	26.40%	28.46	225.17
half2	1380 *	1.09		439		75.04	186.69
	1020	1.48	135.65%	289	34.15%	55.32	283.53
tensor half	1410	0.27		101		307.02	810.80
	1020	0.37	138.18%	65	35.86%	222.18	1264.15
tensor double	1410	4.21		1321		19.44	62.02
	1020	5.82	138.19%	887	32.82%	14.07	92.32

Spetko et al: DGX-A100 Face to Face DGX-2—Performance, Power and Thermal Behavior Evaluation.

A100 VS V100

A100

	Frequency [MHz]	Time [s]	Time Difference	Energy [J]	Energy Savings	Performance [TFLOPS]	Energy Efficiency [GFLOPS/W]
double	1410 1035	8.43 11.49	 136.19%	2285 1601	 29.91%	9.71 7.13	35.86 51.16
float	1410 1035	4.23 5.76	 136.13%	958 721	 24.76%	19.37 14.23	85.53 113.68
Bfloat	1410 1035	2.11 2.88	 136.19%	494 364	 26.40%	38.75 28.46	165.74 225.17
half2	1380 * 1020	1.09 1.48	 135.65%	439 289	 34.15%	75.04 55.32	186.69 283.53
tensor half	1410 1020	0.27 0.37	 138.18%	101 65	 35.86%	307.02 222.18	810.80 1264.15
tensor double	1410 1020	4.21 5.82	 138.19%	1321 887	 32.82%	19.44 14.07	62.02 92.32

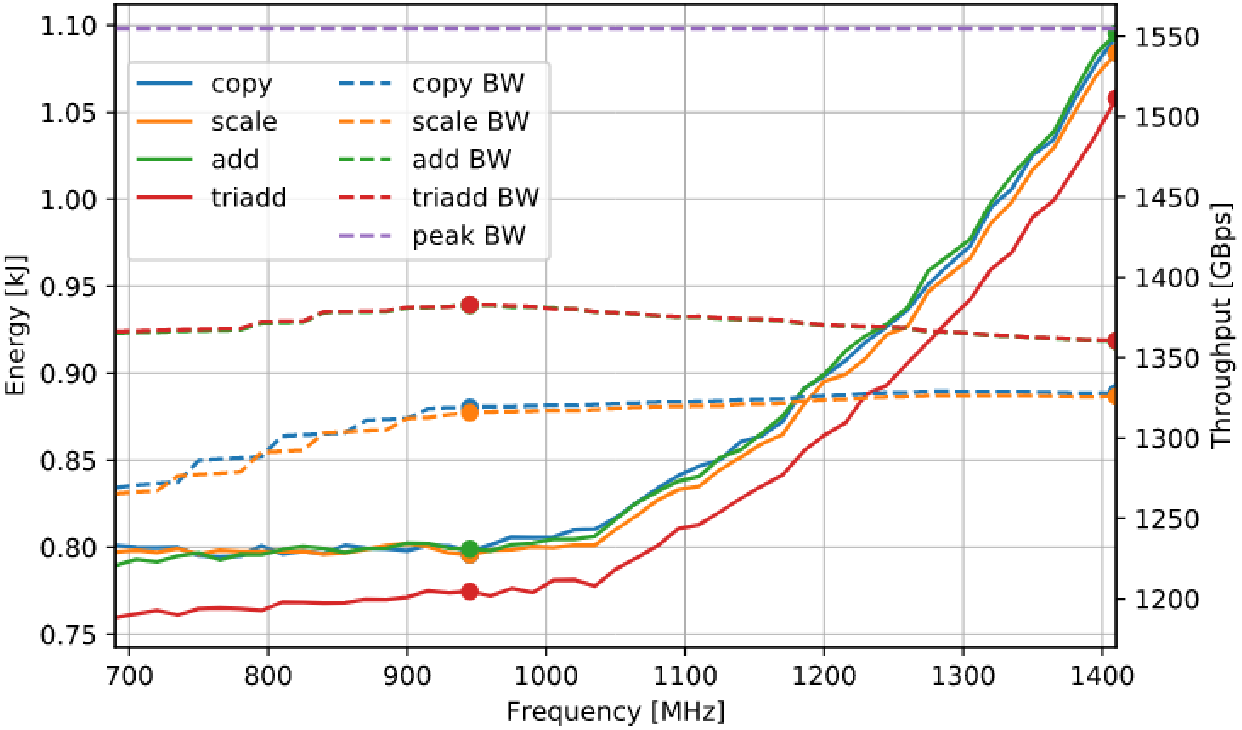
V100

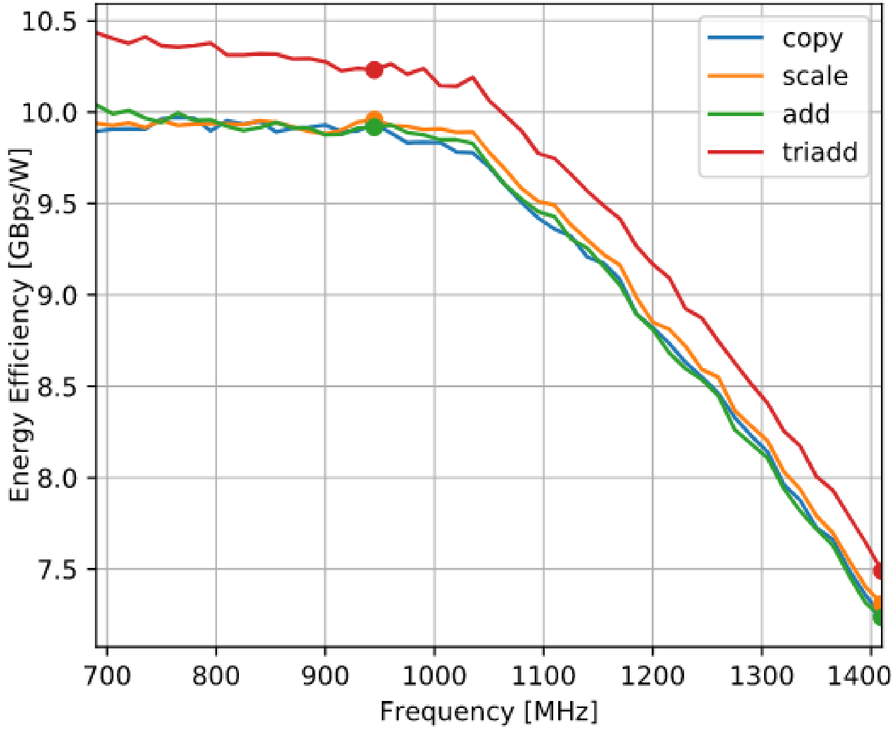
	Frequency [MHz]	Time [s]	Time Difference	Energy [J]	Energy Savings	Performance [TFLOPS]	Energy Efficiency [GFLOPS/W]
double	1597 1050	10.02 15.25	 152.16%	3303 2015	 39.01%	8.17 5.37	24.80 40.67
float	1597 1057	5.01 7.57	 150.99%	1596 982	 38.50%	16.34 10.82	51.33 83.46
half2	1597 1057	2.51 3.78	 151.05%	870 531	 38.97%	32.69 21.64	94.18 154.30
tensor half	1597 1057	0.63 0.95	 151.04%	219 132	 39.58%	130.65 86.50	374.90 620.51

GPU TUNING

STREAM benchmark

A100 + core frequency tuning





GPU TUNING

STREAM benchmark

A100 + core frequency tuning

	Frequency [MHz]	Time [s]	Time Difference	Energy [J]	Energy Savings	Throughput [GBps]	Energy Efficiency [GBps/W]
copy	1410 945	5.97 6.01	 100.69%	1094 798	 27.07%	1328.16 1319.02	7.25 9.94
scale	1410 945	5.98 6.02	 100.77%	1084 796	 26.59%	1325.90 1315.73	7.31 9.96
add	1410 945	5.83 5.73	 98.39%	1095 799	 27.05%	1360.35 1382.62	7.23 9.92
triadd	1410 945	5.82 5.73	 98.39%	1058 775	 26.79%	1360.62 1382.92	7.49 10.23

EE HPC centers

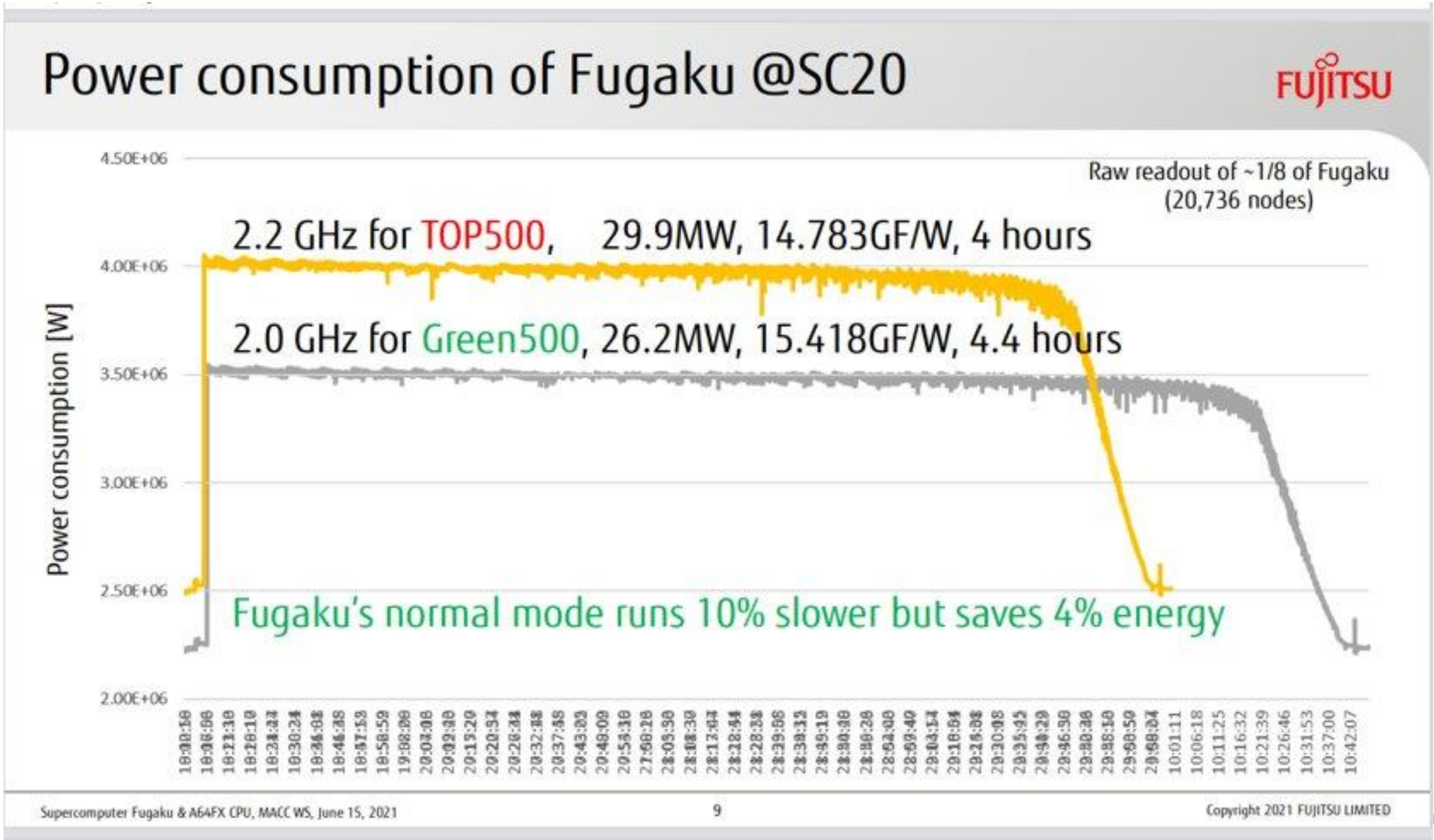
ENERGY AND POWER AWARE HPC CENTERS

RIKEN Fugaku:

- #1 in Top500 since 6/2020
- Using Fujitsu A64FX (48 compute cores + 4 assistant cores for OS daemon and MPI offload)
  - No TDP, no nominal frequency => no turbo frequency
  - Available frequencies 1.6, 2.0, or 2.2 GHz
- User-controlled options
  - Power mode (scheduler option)
    - **Normal** - 2.0 GHz frequency
    - **Boost** - 2.2 GHz frequency
    - **ECO** – 2.0 GHz frequency + use one of two FP units only + reduces its standby power
    - **Boost ECO** - 2.2 GHz frequency + FPU elimination
  - Core retention ON/OFF
    - Eliminates standby power idle CPU cores
- See: https://sites.google.com/view/rikenfugakushowcase/home

CPU core frequency	1.8	2.0	2.2	GHz
Peak DP perf (FP64)	2.7	3.0	3.3	TFLOPS
Peak SP perf (FP32)	5.5	6.1	6.7	TFLOPS
Peak HP perf (FP16)	11	12	13	TFLOPS
Memory peak bandwidth	1024			GB/s

ENERGY AND POWER AWARE HPC CENTERS



ENERGY AND POWER AWARE HPC CENTERS

LRZ SuperMUC-NG:

- #8 in Top500 in 11/2018, Rmax 20 PFlops
- Using Intel Xeon Platinum 8174 (24 cores)
    - Intel default
        - 240 W TDP
        - 2.4 GHz CPU uncore frequency
        - Turbo CPU core frequencies
            - 3.9 GHz SSE, 3.8 GHz AVX-2, 3.8 GHz AVX-512
    - LRZ Default
        - 205 W power limit (-14.6%)
        - 1.8 GHz CPU uncore frequency
        - Turbo CPU core frequencies
            - 3.7 GHz SSE, 3.6 GHz AVX-2, 3.5 GHz AVX-512
- All jobs executed under Energy Aware Runtime (EAR)

See:
https://doku.lrz.de/display/PUBLIC/Details+of+Compute+Nodes

https://doku.lrz.de/display/PUBLIC/Energy+Aware+Runtime

ENERGY AND POWER AWARE HPC CENTERS

CINECA's systems:

- It is possible to access and change all the power knobs without special permission on all CINECA's systems, the SLURM scheduler takes care to restore a default configuration after the termination of power-aware jobs.

- **Marconi**
  - Intel Xeon 8160 Skylake, 24 cores, 150 W TDP
  - User-controlled knobs - Power capping, frequency scaling, power driver

```
$ srun -A $PROJECT
--partition=skl_usr_prod
--gres=msrsafe,sysfs --exclusive
```

- **Marconi100**
  - #9 Top500 6/2020
  - IBM POWER9 AC922, 16 cores
  - User-controlled knobs - Power capping, frequency scaling, power driver

```
$ srun -A $PROJECT
--partition=m100_usr_prod
--gres=sysfs --exclusive
```

- **Galileo100**
  - Intel Xeon 8260 Cascade lake, 24 cores, 165 W TDP
  - Support under development
  - User-controlled knobs - Power capping, frequency scaling, power driver

```
$ srun -A $PROJECT
--partition=g100_usr_prod
--gres=msrsafe,sysfs --exclusive
```

ENERGY AND POWER AWARE HPC CENTERS

IT4I's systems:

- It is possible to access and change power knobs and monitor energy consumption
- **Barbora**
  - Intel Xeon Cascade Lake 6240, 18 cores, 150 W TDP / Intel Skylake Gold 6126, 12 cores, 120 W TDP + Nvidia V100, 300 W TDP
  - User-controlled knobs
    - CPU: Power capping, core + uncore frequencies scaling, power driver
    - GPU: Power capping, Mem + SM frequencies scaling
  - Power monitoring – Intel RAPL, Atos|Bull HDEEM / Intel RAPL, Nvidia NVML
- **Karolina**
  - AMD EPYC 7h12, 64 cores, 280 W TDP / AMD EPYC 7763, 64 cores, 280 W TDP + Nvidia A100, 400 W TDP
  - IT4I settings: **7h12** 3.3GHz -> 2.1GHz, **7763** 3.5GHz -> 2.6GHz, **A100** 1.41GHz -> 1.29GHz
  - User-controlled knobs
    - CPU: Power capping, core frequency scaling
    - GPU: Power capping, Mem + SM frequencies scaling
  - Power monitoring – AMD RAPL / AMD RAPL, Nvidia NVML

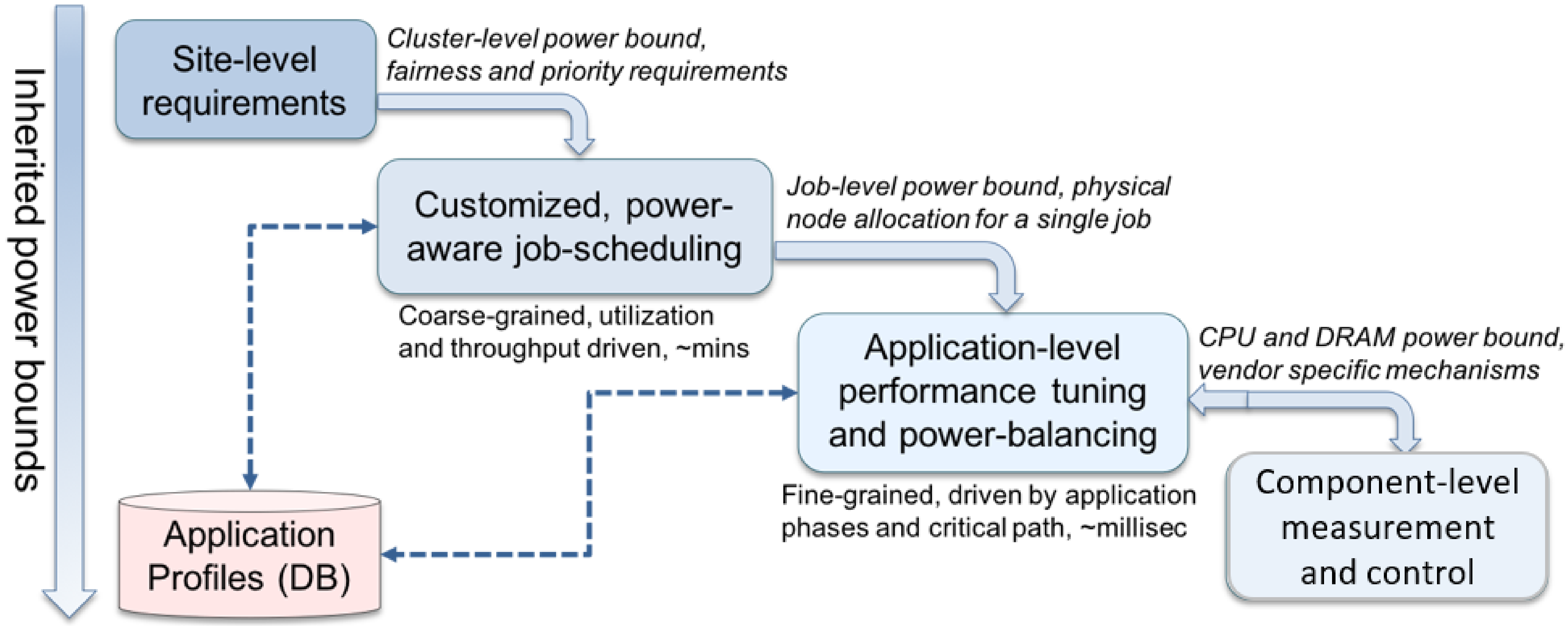
ENERGY AND POWER AWARE HPC CENTERS

NSCC's system:

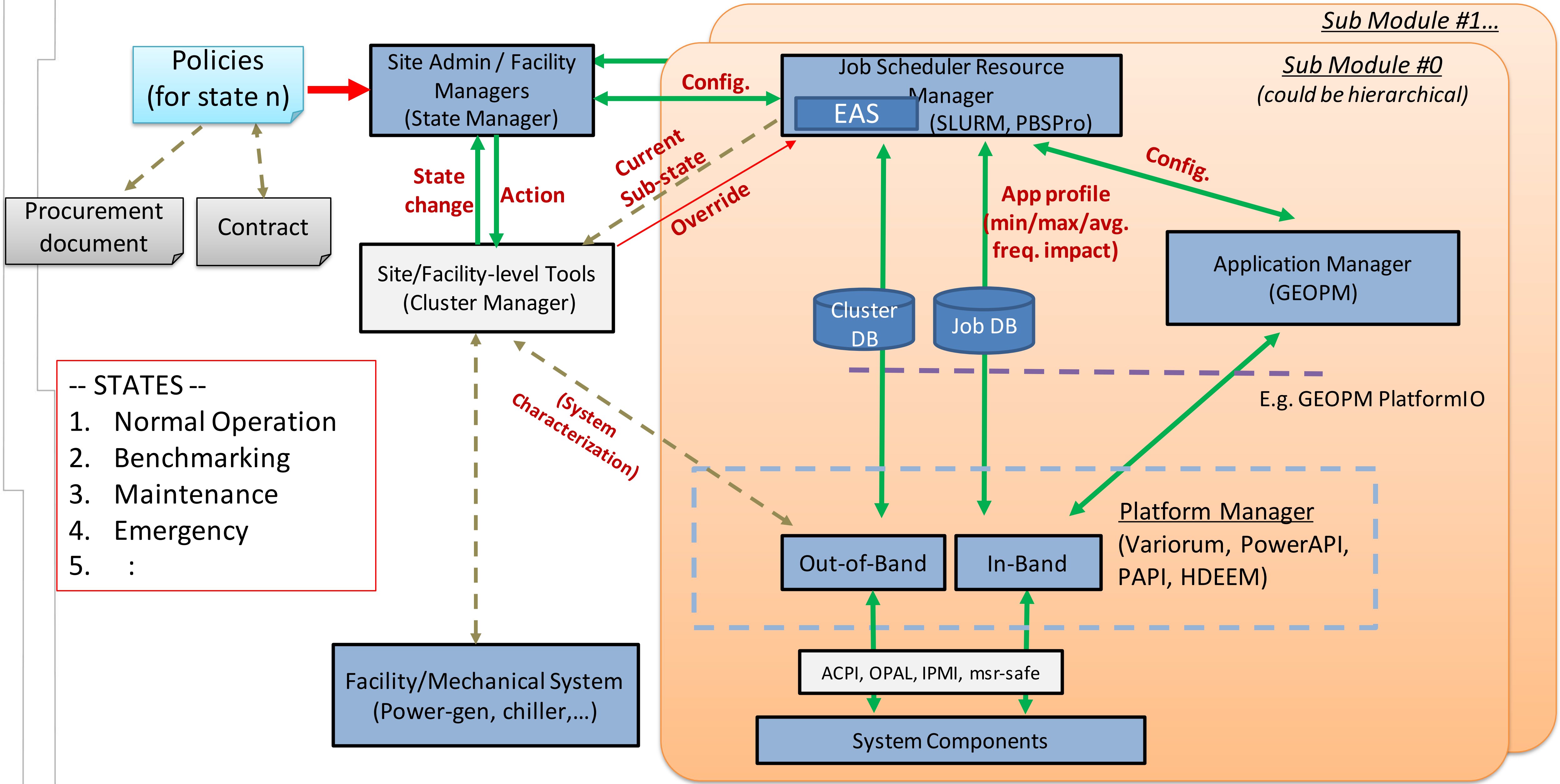
- It is possible to access and change power knobs and monitor energy consumption
- **Devana**
  - Intel Xeon Gold 6338, 32 cores, 205 W TDP / Intel Xeon Gold 6338, 32 cores, 205 W TDP + Nvidia A100, 400 W TDP
  - User-controlled knobs
    - CPU: core + uncore frequencies scaling, scaling governor
    - GPU: none
  - Power monitoring – Intel RAPL / --, Nvidia NVML

```
$ srun -A $PROJECT
--partition=ncpu
--gres=msrsafe,sysfs --exclusive
```

HPC POWERSTACK ARCHITECTURE



HPC POWERSTACK ARCHITECTURE



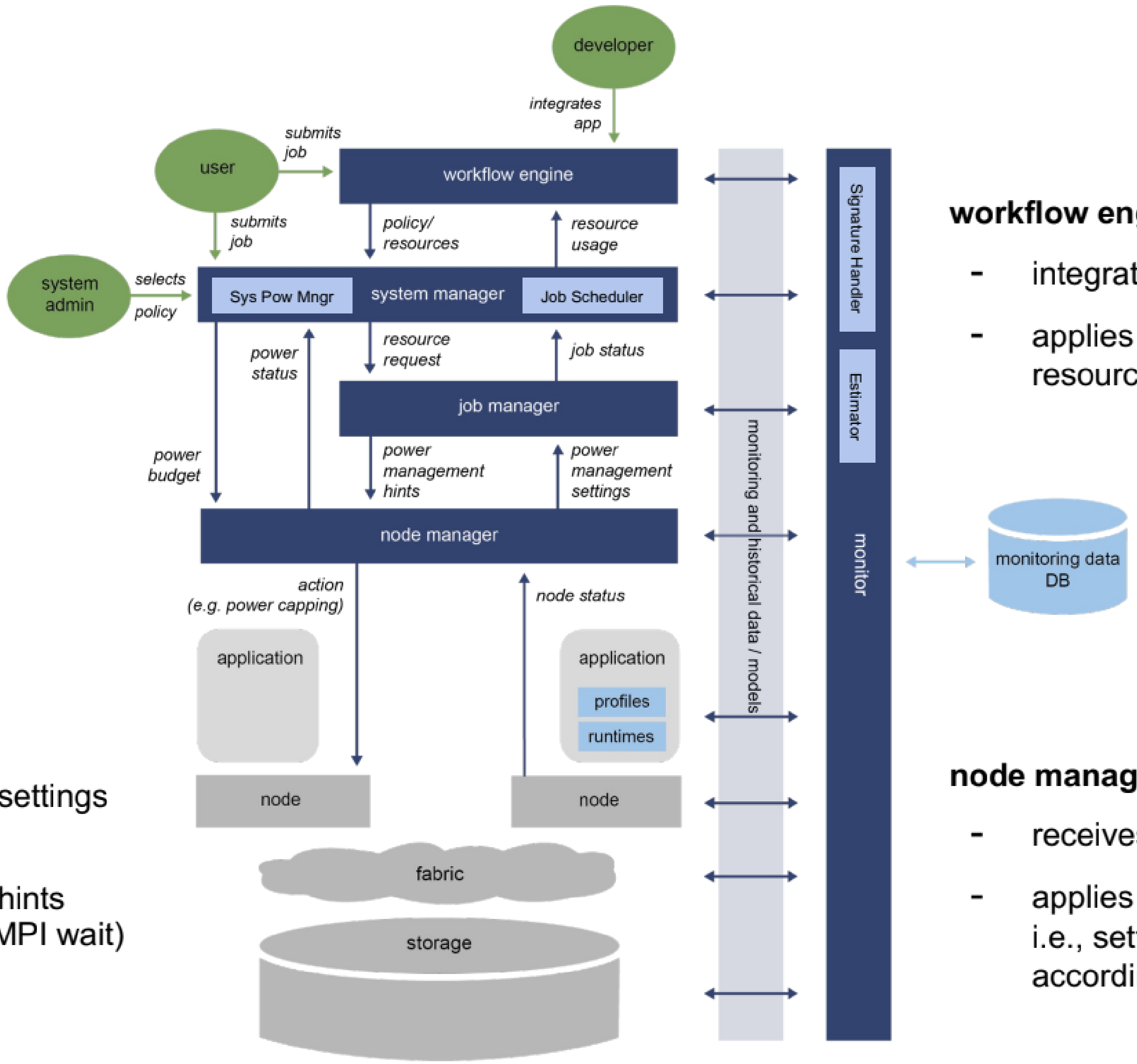
REGALE ARCHITECTURE

system manager

- receives:
  - admin policy
  - job requests
  - application profiles
  - job status
- decides on:
  - job launch
  - resource allocation
  - power budgets

job manager

- receives:
  - policy and resources
  - power management settings
- provides
  - power management hints
    (e.g., idle event like MPI wait)



workflow engine

- integrates with user applications
- applies optimized and elastic resource management

node manager

- receives power budget and hints
- applies actions at the node level, i.e., setting up HW knobs accordingly

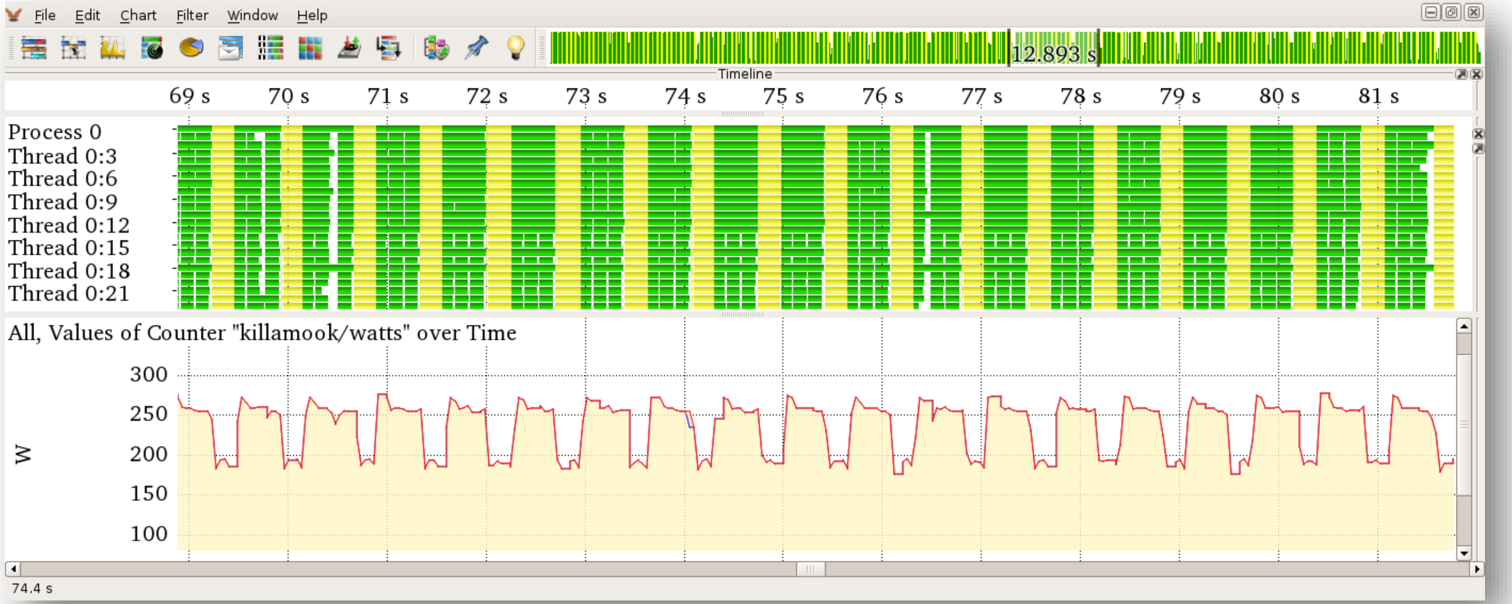
Energy-aware dynamic tuning

EE RUNTIME SYSTEMS

- **BSC EAR**
  - CPU core frequency tuning based on executed instructions
- **CINECA/UniBo COUNTDOWN**
  - CPU core frequency tuning during MPI communication phases
- **LLNL Conductor**
  - Power overprovisioning per loop iteration
- **Atos BDPO**
  - CPU core frequency tuning based on HW metrics sampling to identify HW execution phase
  - CPU core frequency tuning during MPI communication phases
- **Intel GEOPM**
  - Power overprovisioning or CPU core frequency tuning for instrumented regions (manual, MPI, OMP)

READEX PROJECT

- Applications exhibit dynamic behavior
  - Changing resource requirements
  - Computational characteristics
  - Changing load on processors over time



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

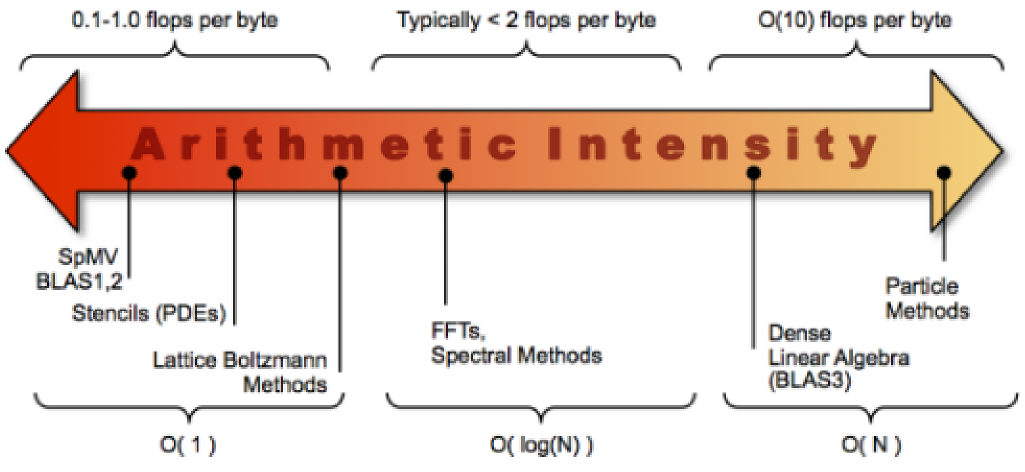
TECHNISCHE UNIVERSITÄT DRESDEN

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

NUI Galway OÉ Gaillimh

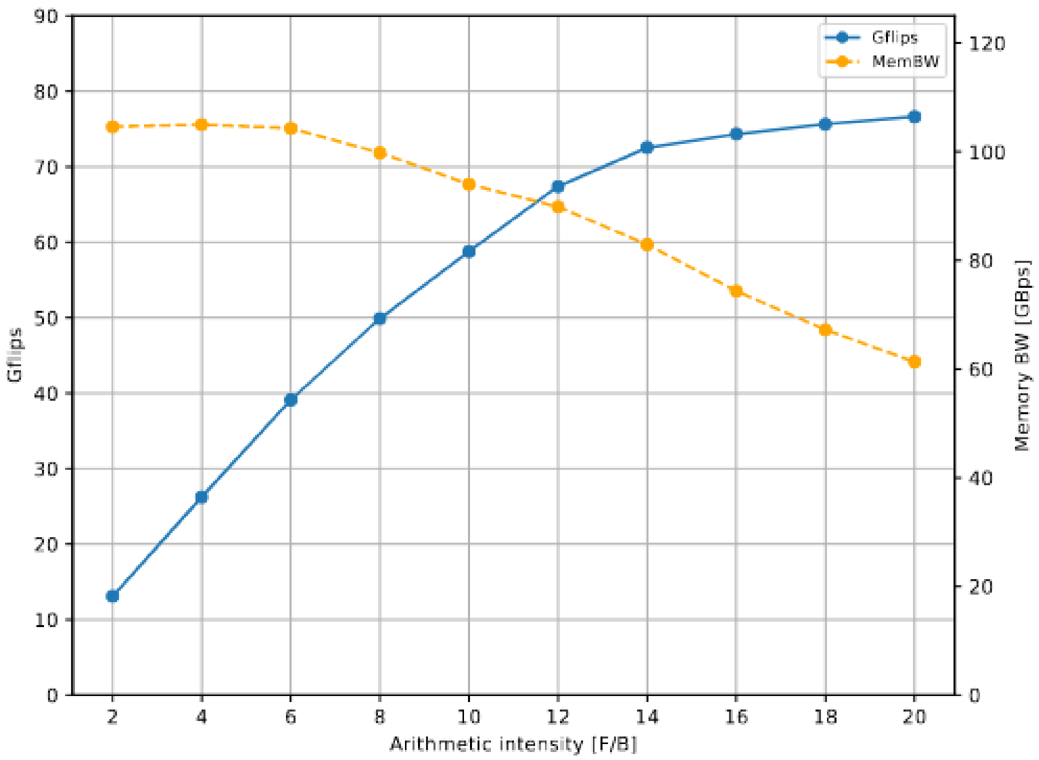
- Goal was to create a tools-aided methodology for automatic tuning of parallel applications
- Dynamically adjust system parameters to actual resource requirements

DYNAMICITY



(a) Arrow presenting a range of applications of various arithmetic intensities

https://crd.lbl.gov/divisions/amcr/computer-science-amcr/par/research/roofline/introduction/

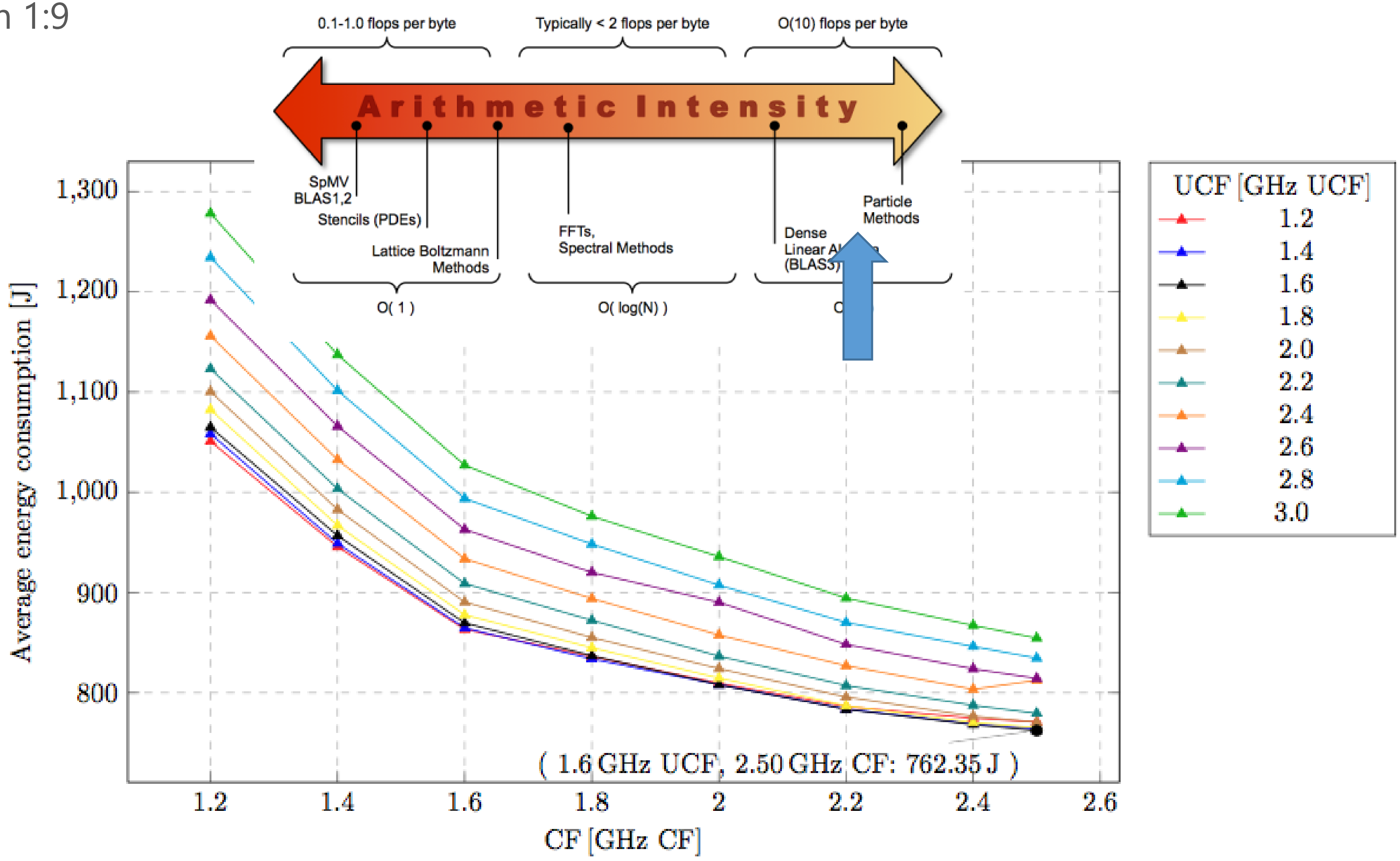


(b) Roofline model of the Intel Xeon Gold 6240 processor when executing a workload of AVX-512 instructions.

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

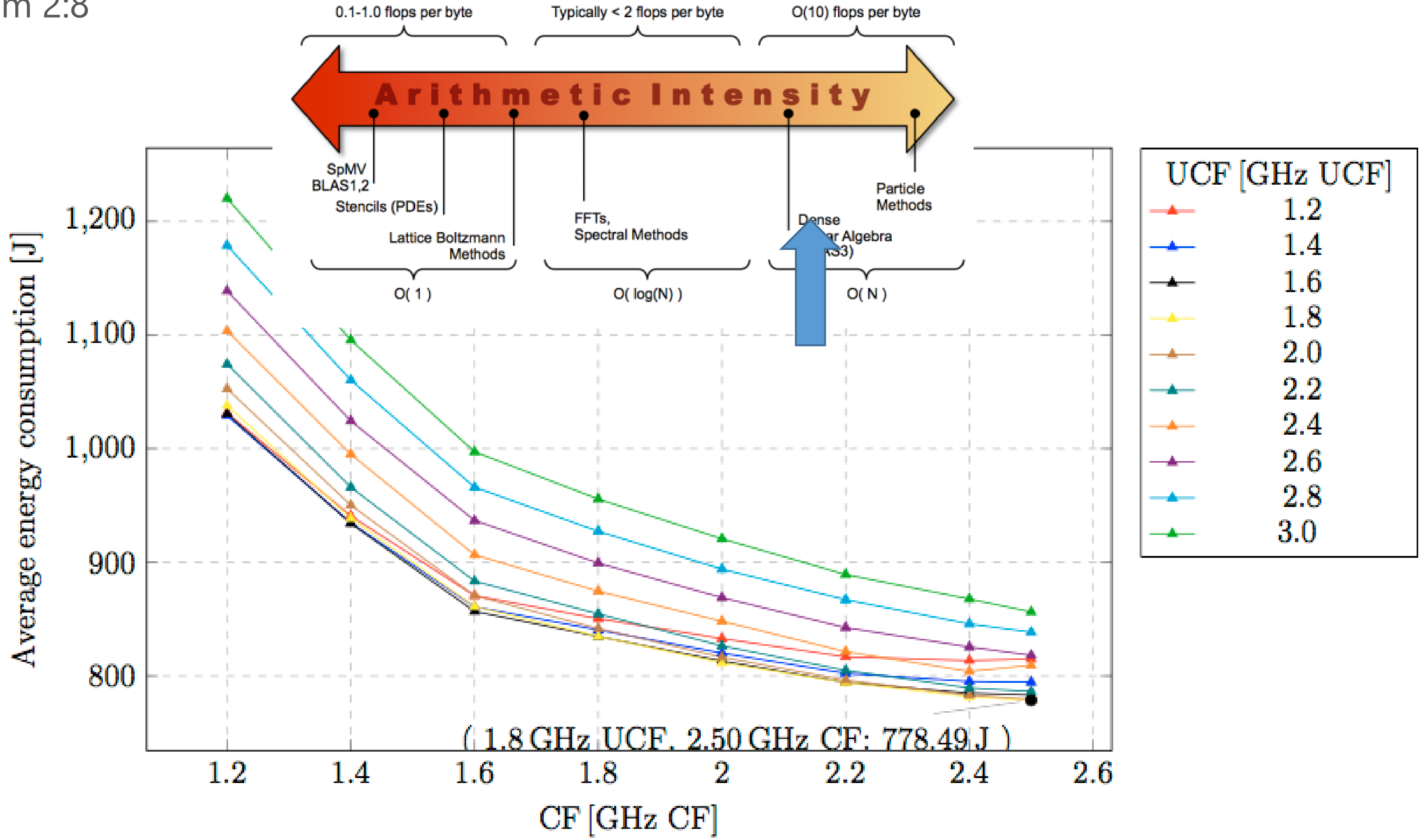
STATIC TUNING FOR VARIOUS AI

Ratio from 1:9



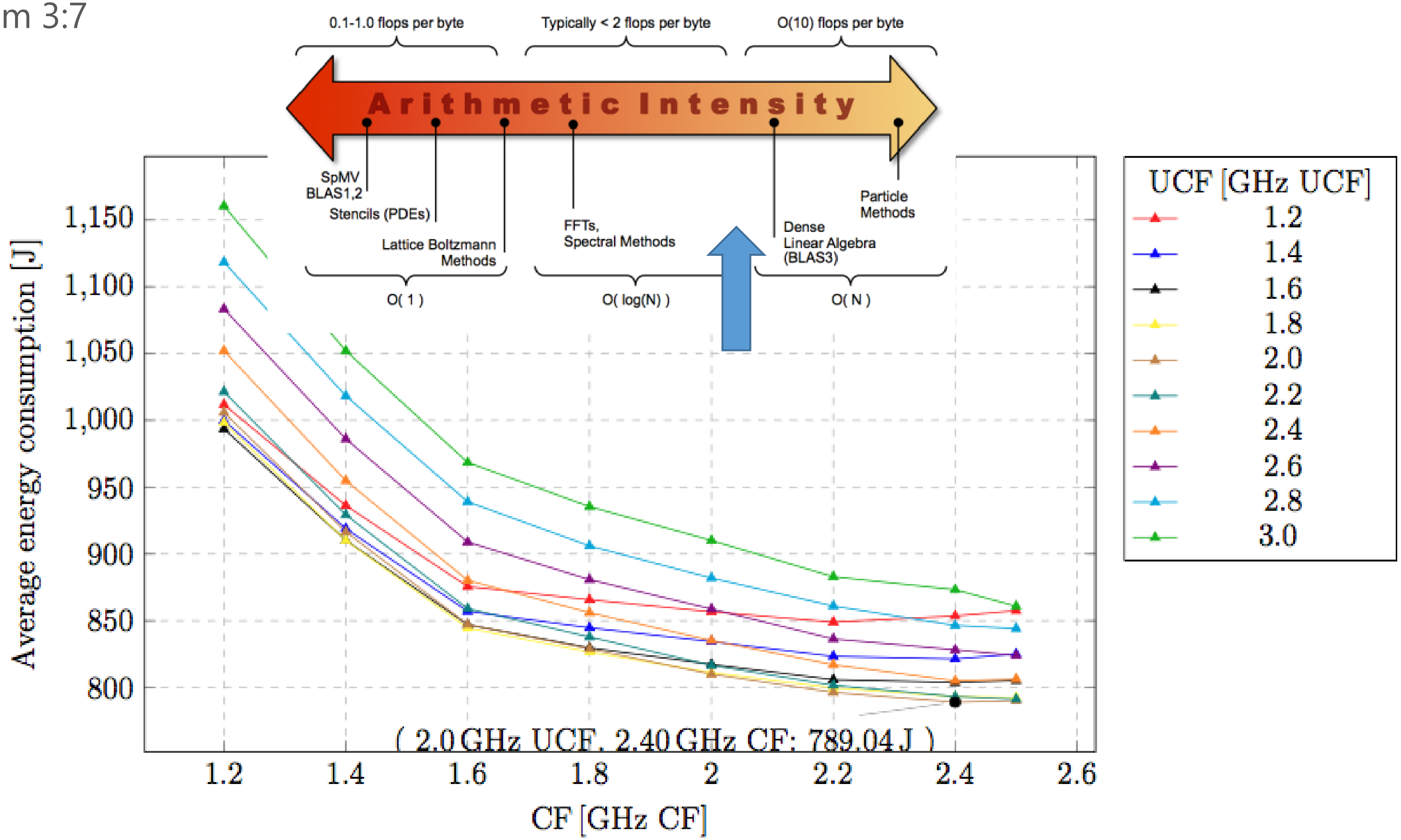
STATIC TUNING FOR VARIOUS AI

Ratio from 2:8



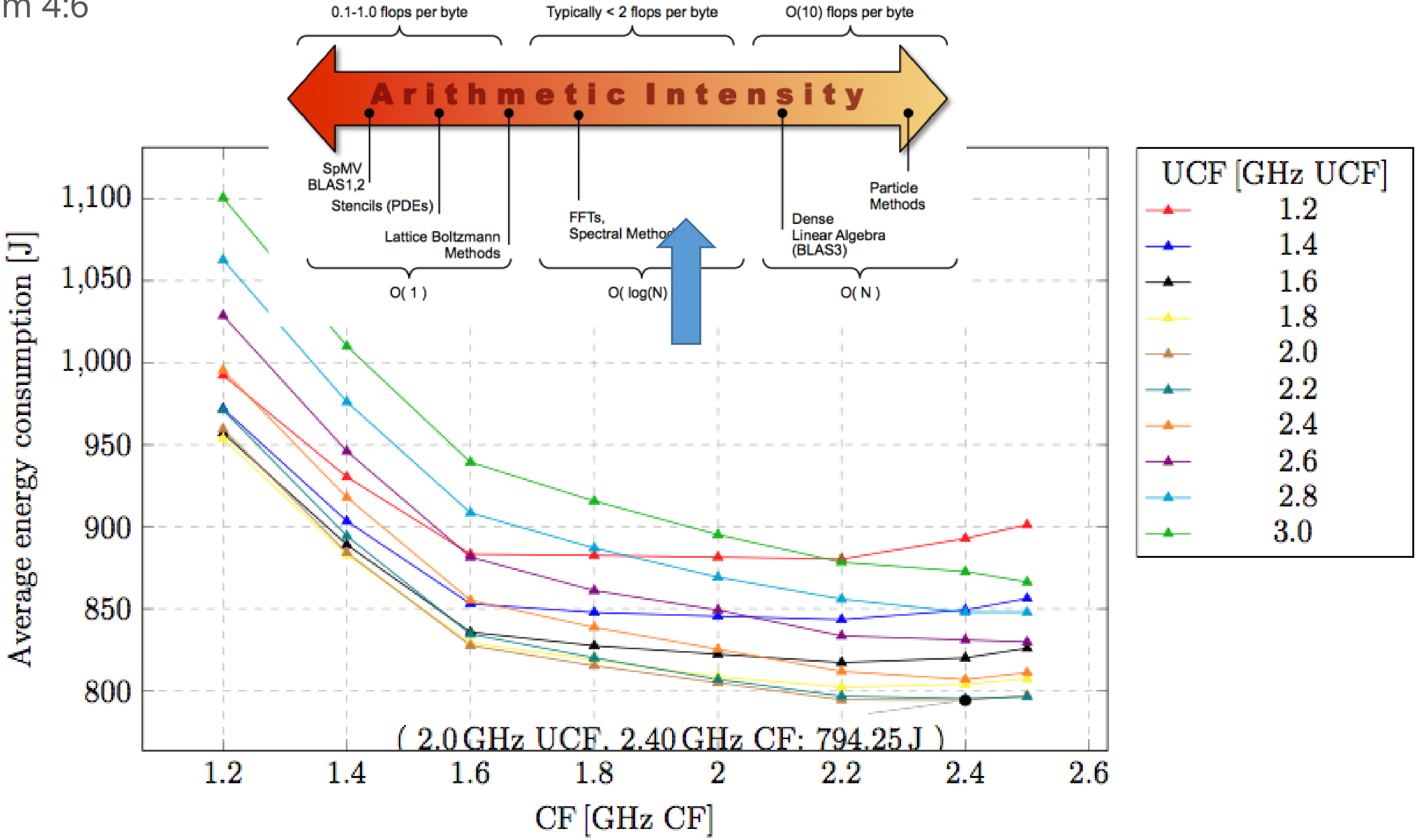
STATIC TUNING FOR VARIOUS AI

Ratio from 3:7



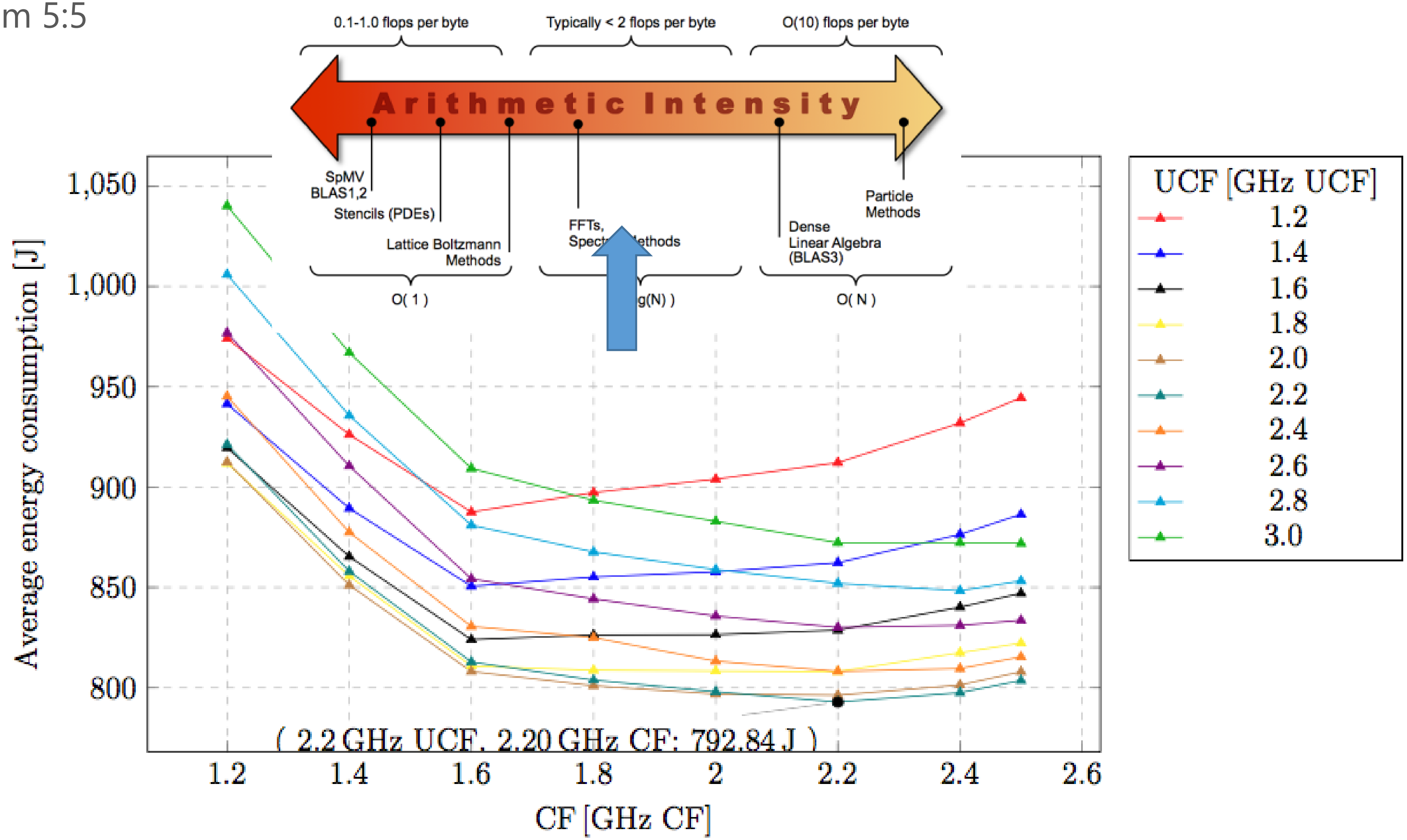
STATIC TUNING FOR VARIOUS AI

Ratio from 4:6



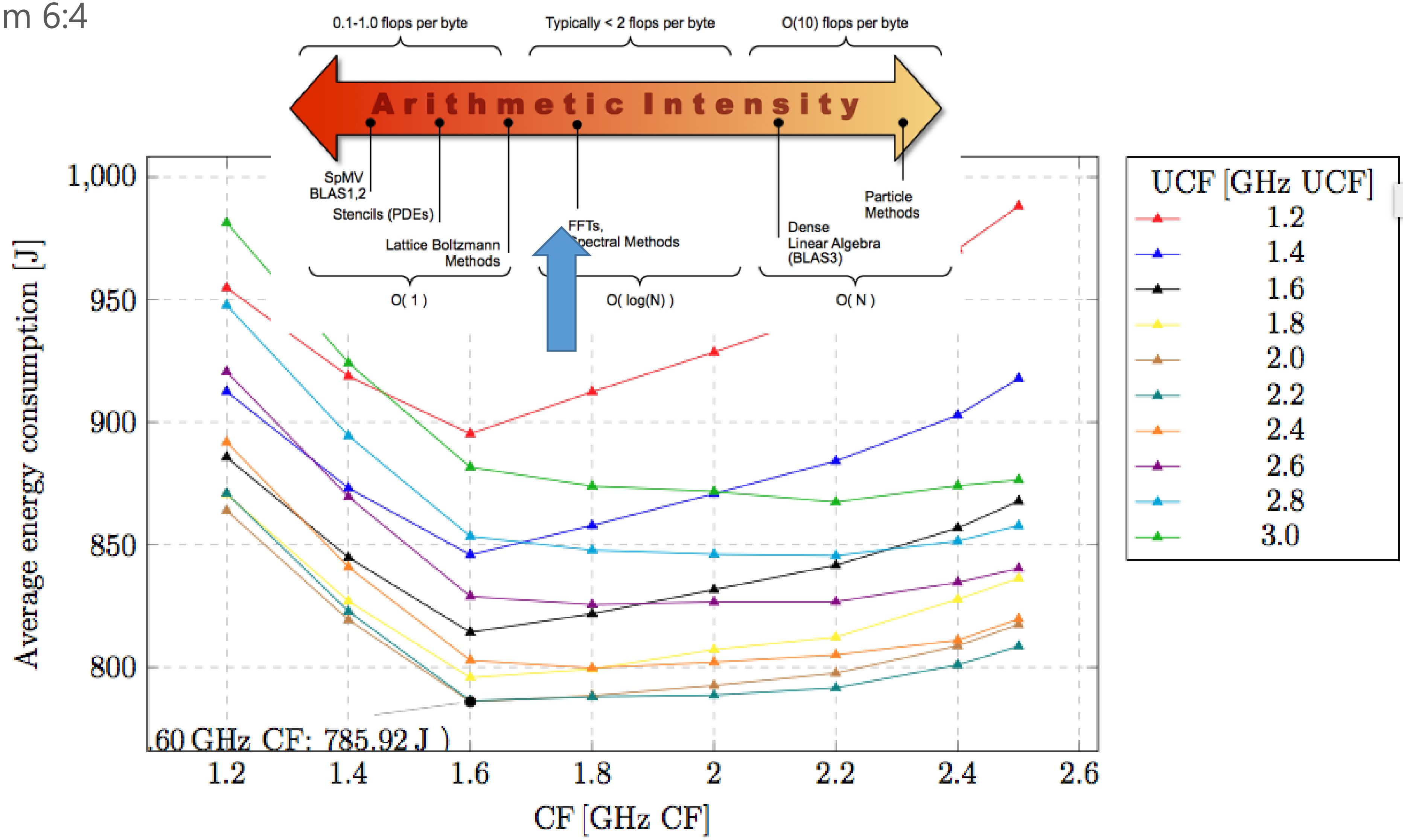
STATIC TUNING FOR VARIOUS AI

Ratio from 5:5



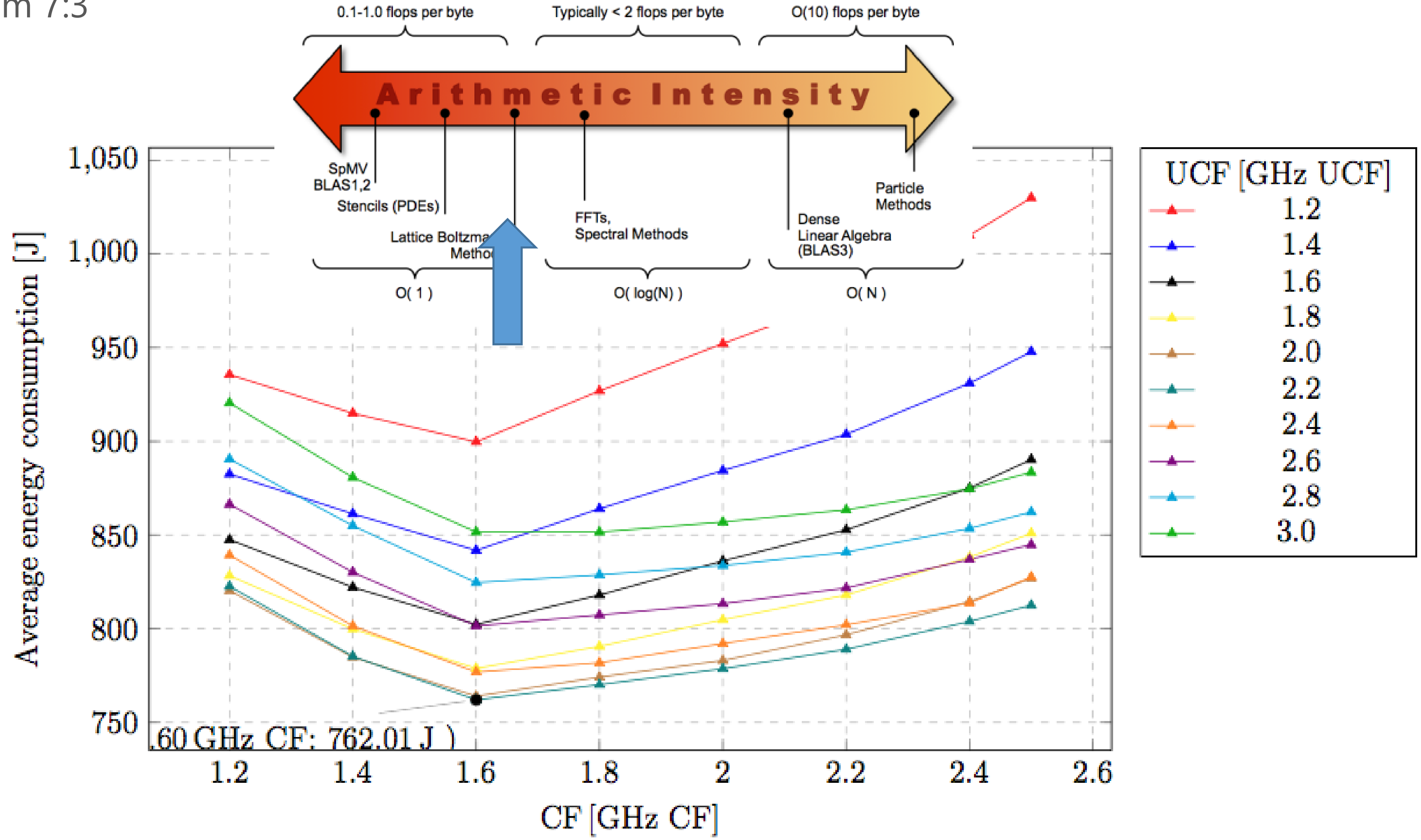
STATIC TUNING FOR VARIOUS AI

Ratio from 6:4



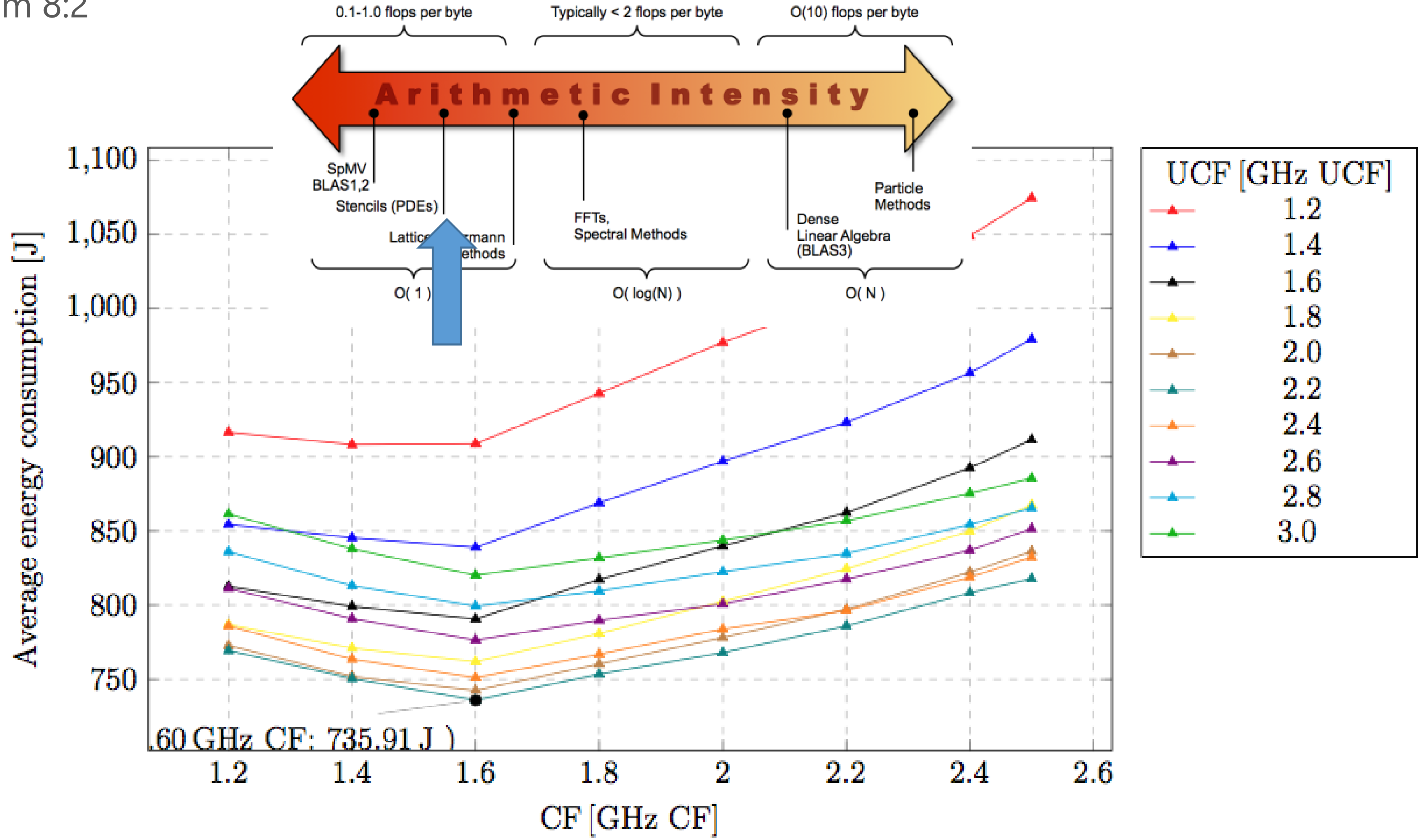
STATIC TUNING FOR VARIOUS AI

Ratio from 7:3



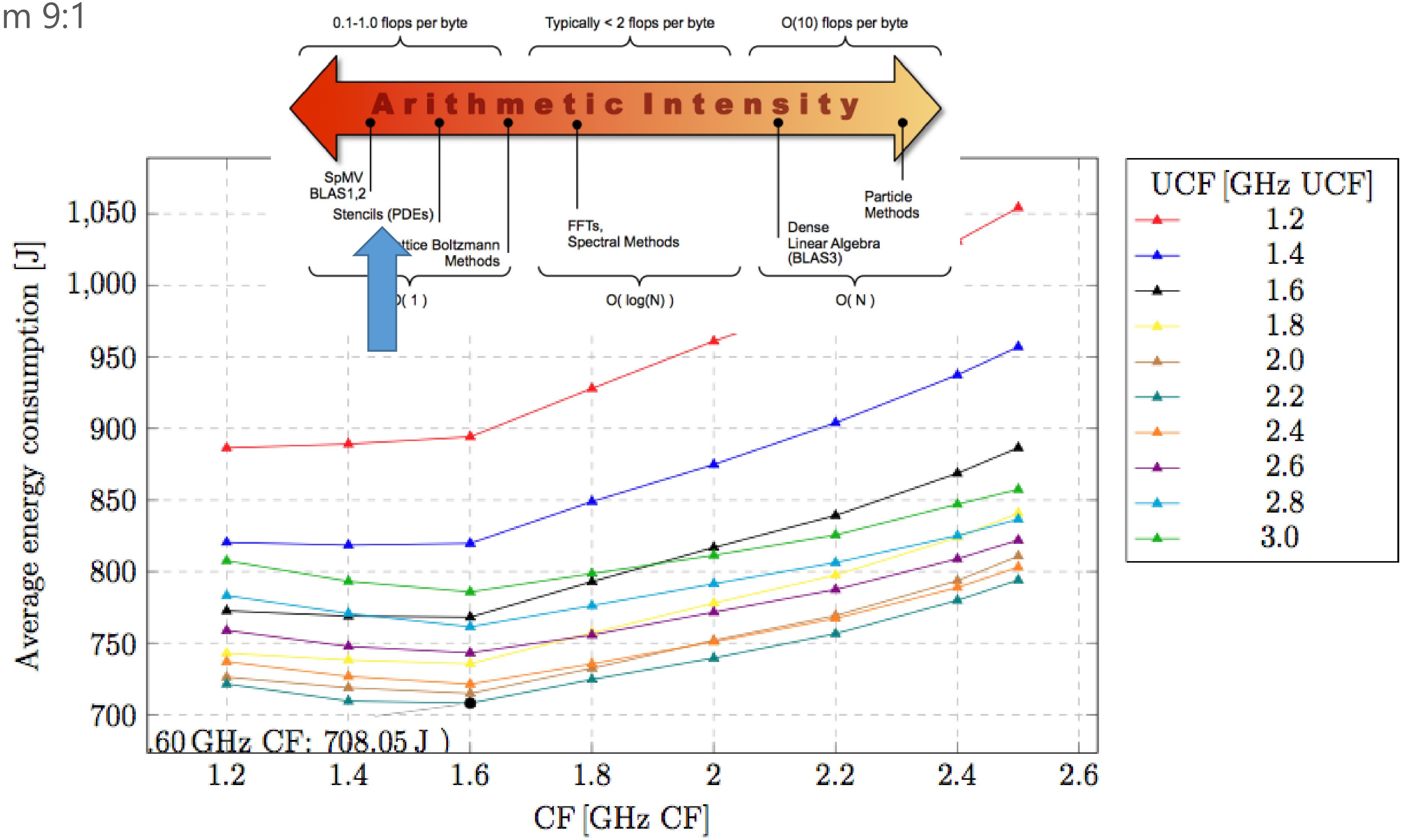
STATIC TUNING FOR VARIOUS AI

Ratio from 8:2



STATIC TUNING FOR VARIOUS AI

Ratio from 9:1



DYNAMIC TUNING

```
int main(void) {

    // Initialize application
    // Initialize experiment variables

    int num_iterations = 2;
    for (int iter = 1; iter <= num_iterations; iter++) {
       // Start phase region

       laplace3D();              // significant region
       residue = reduction(); // insignificant region
       fftw_execute();           // significant region

       // End phase region
    }

    // Post-processing:
    // Write noise matrices to disk for visualization
    // Terminate application

    MPI_Finalize();
    return 0;
}
```

Phase region

Significant region

Significant region

FREQ=2 GHz

FREQ=1.5 GHz

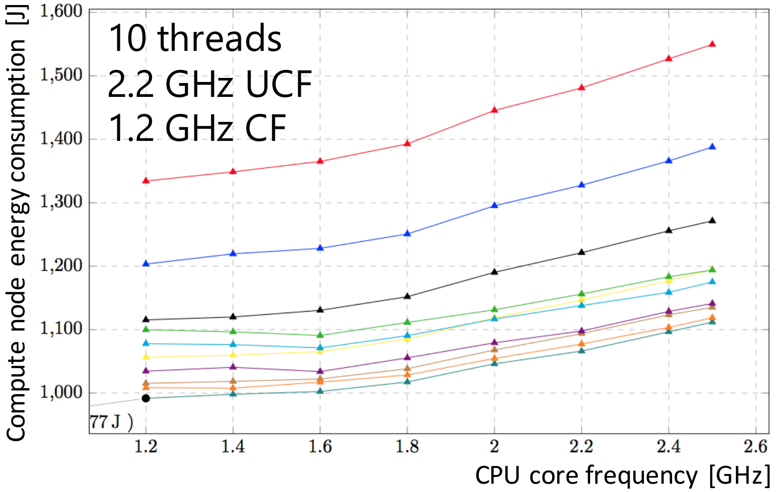
HW PARAMETERS TUNING

Behavior of the simple application with two kernels

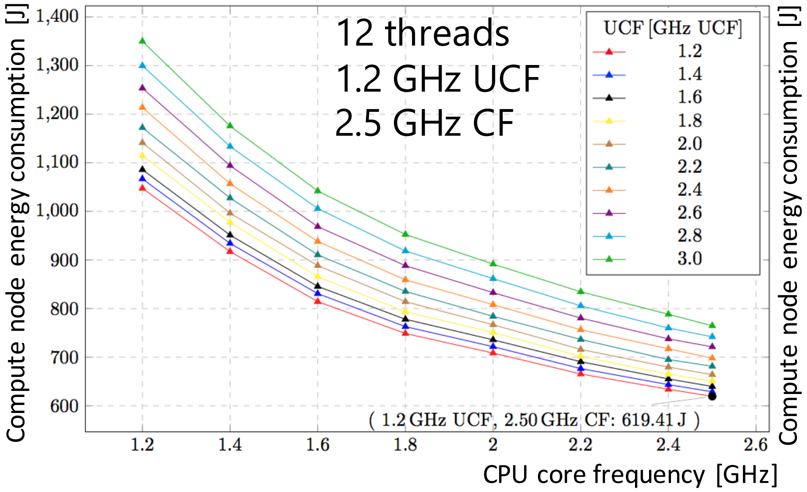
- Low computational intensity – DGEMV
- High computational intensity – DGEMM
- Tuning of three parameters
  - CPU core and uncore frequency, number of OpenMP threads

Two kernels with 1:1 workload ratio	Energy consumption	Energy savings	
Default settings	2017 J	-	-
Static tuning	1833 J	184 J	9%
Dynamic tuning	1617 J	400 J	20%

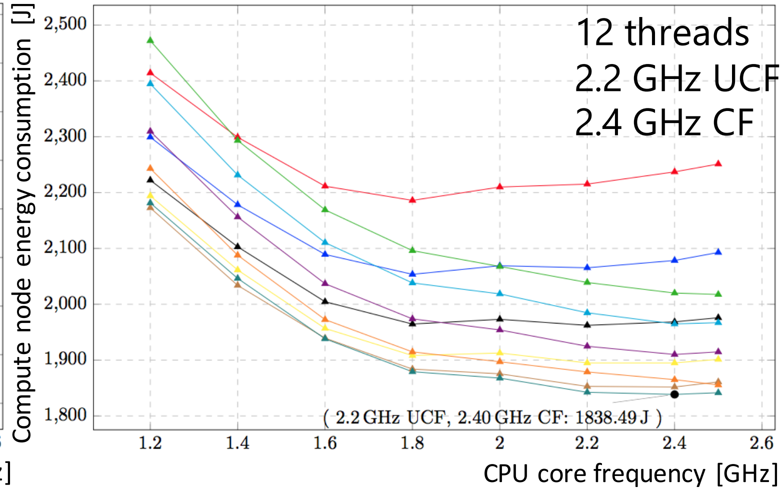
Low CI (DGEMV)



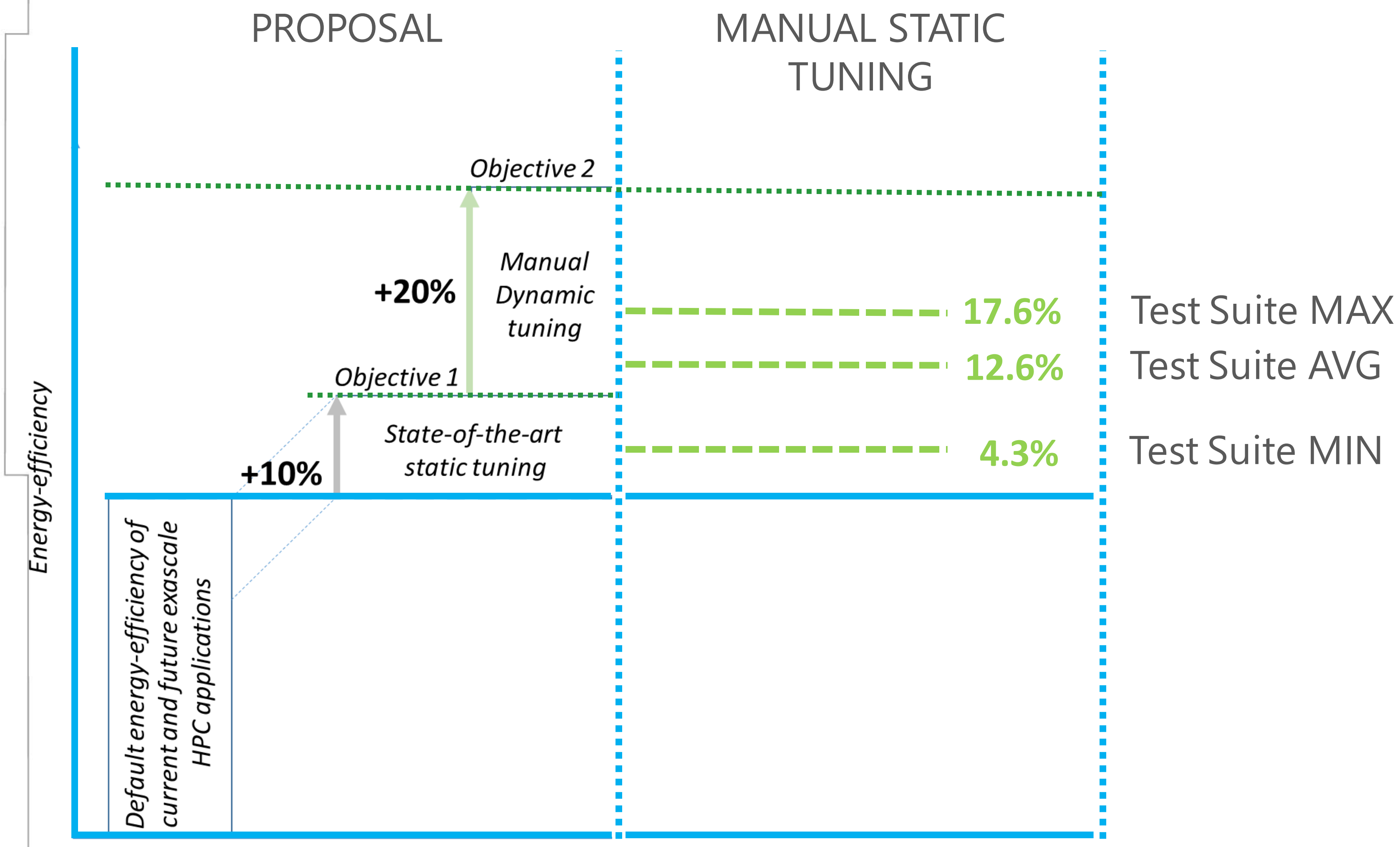
High CI (DGEMM)



Static tuning for both kernels

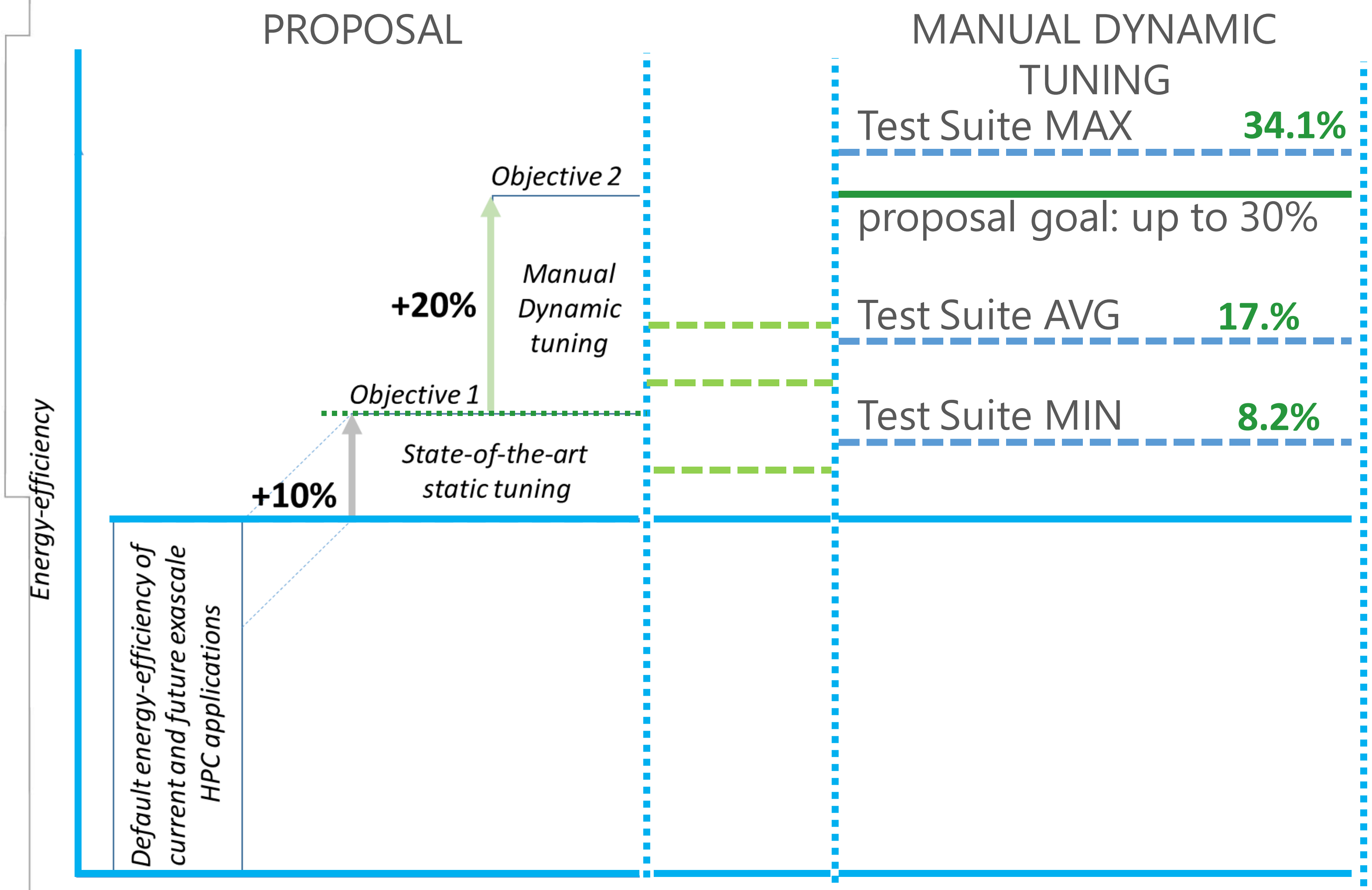


STATIC TUNING



Software	Static tuning savings
AMG2013	12.5 %
Blasbench	7.4 %
Kripke	11.5 %
Lulesh	17.6 %
NPB3.3	11.0 %
BEM4I	15.7 %
INDEED	17.6 %
ESPRESO	4.3 %
OpenFOAM	15.9 %
Average	**12.6 %**

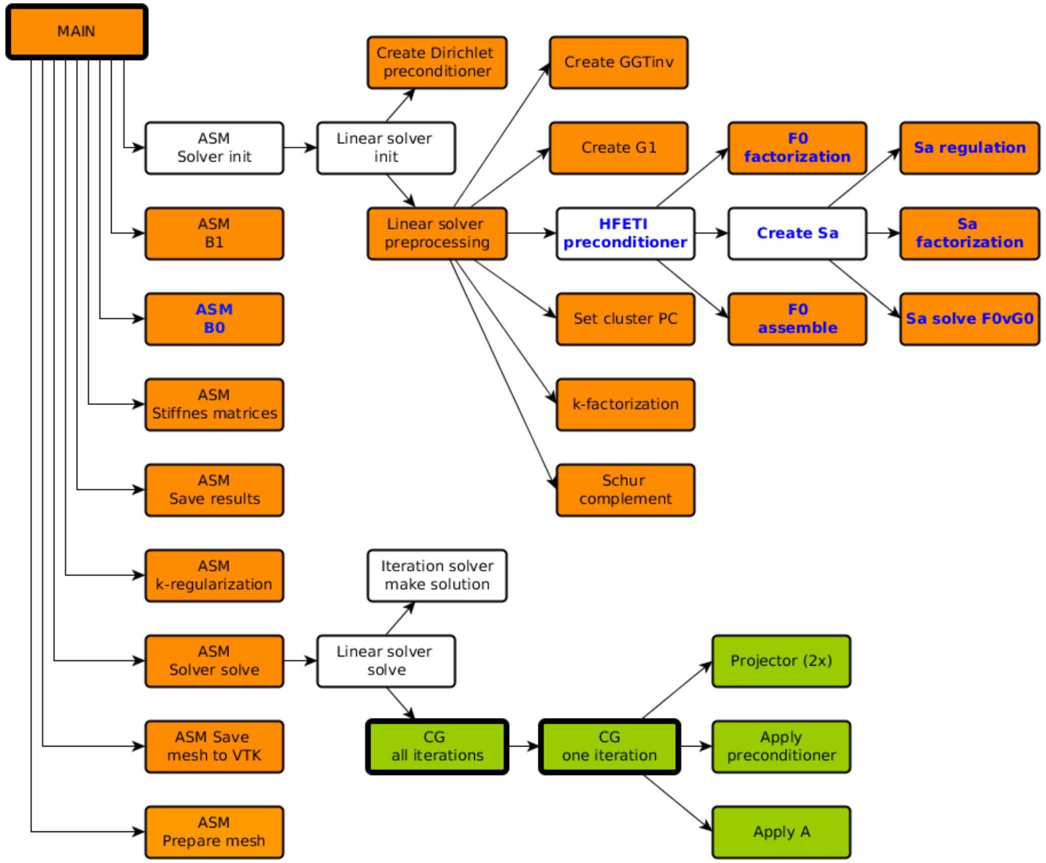
DYNAMIC TUNING



Software	Dynamic tuning savings
AMG2013	12.5 %
Blasbench	15.3 %
Kripke	18.5 %
Lulesh	18.7 %
NPB3.3	11.0%
BEM4I	34.1 %
INDEED	19.5 %
ESPRESO	8.2 %
OpenFOAM	20.1%
Average	**17.5 %**

IMPROVING PERFORMANCE AT SCALE

- Strong scaling of ESPRESO FEM code
  - Improved performance and energy consumption



#nodes	Default time [s]	Default energy [kJ]	Tuned time [s]	Tuned energy [kJ]	Time savings [s]	Energy savings [%]
1	129.3	37.2	143.7	34.3	**-11.1**	**8.0**
2	68.6	39.8	75.5	36.5	**-10.1**	**8.2**
4	33.2	38.0	35.6	34.3	**-7.2**	**9.8**
8	21.5	49.6	22.9	44.7	**-6.8**	**9.9**
16	13.4	60.8	14.3	53.5	**-6.3**	**12.1**
32	7.7	62.2	7.2	50.6	**6.1**	**18.7**
64	4.0	69.9	3.6	52.4	**9.3**	**25.0**
128	3.6	119.6	2.8	80.1	**22.2**	**33.0**

BEM4I

Application runtime	assemble_k [s]	assemble_v [s]	gmres_solve [s]	print_vtu [s]	main [s]
default runtime	5.4	5.9	10.2	5.6	27.3
static tuning runtime	9.8	10.6	6.1	2.4	29.0
dynamic tuning runtime	7.0	7.2	7.9	2.1	24.3

static savings [%]	-82.3%	-79.1%	40.5%	56.8%	-6.2%
dynamic savings [%]	-30.6%	-20.9%	23.2%	62.9%	10.9%

```
"static": {
    "FREQUENCY":  "25",                  <---------  2.5 GHz
    "NUM_THREADS":  "12",                <---------  12 OpenMP threads
    "UNCORE_FREQUENCY":  "22" }          <---------  2.2 GHz
```

Hardware: dual socket system with 2x12 CPU cores – *"standard HW"* in HPC centres

Region description:

- **assemble_k** and **assemble_v** – high utilization of vector units, extreme level of optimization – fully compute bound great utilization of both sockets and all cores
- **gmres_solve** – uses DGEMV from MKL – memory bound, suffers on NUMA effect; this routine is more efficient on single socket
- **print_vtu** – single threaded I/O and network bound region why stores data to a file on LUSTRE system

```
"assemble_k": {
   "FREQUENCY": "23",
   "NUM_THREADS": "24",
   "UNCORE_FREQUENCY": "16"
},

"assemble_v": {
   "FREQUENCY": "25",
   "NUM_THREADS": "24",
   "UNCORE_FREQUENCY": "14"
},

"gmres_solve": {
   "FREQUENCY": "17",
   "NUM_THREADS": "8",
   "UNCORE_FREQUENCY": "22"
},

"print_vtu": {
   "FREQUENCY": "25",
   "NUM_THREADS": "6",
   "UNCORE_FREQUENCY": "24"
}
```

BEM4I

Compute node energy	assemble_k [J]	assemble_v [J]	gmres_solve [J]	print_vtu [J]	main [J]
default energy	1476	1484	2733	1142	6872
static tuning energy	1962	2015	1366	420	5792
dynamic tuning energy	1467	1462	1259	293	4531
static savings [%]	-33.8%	-35.8%	50.0%	63.2%	15.7%
dynamic savings [%]	0.6%	1.5%	53.9%	74.3%	34.1%

```
"static": {
      "FREQUENCY":  ”25",                 <---------  2.5 GHz
      "NUM_THREADS":  ”12",               <---------  12 OpenMP threads
      "UNCORE_FREQUENCY":  ”22” }         <---------  2.2 GHz
```

Large energy savings is combination of optimal HW settings and runtime savings due to mitigation of NUMA effect by optimal settings of OpenMP threading

- Without savings in runtime caused by similar application will
  - Energy savings approx. 15 – 20%
  - Runtime savings approx. -15%

```
"assemble_k": {
   "FREQUENCY": "23",
   "NUM_THREADS": "24",
   "UNCORE_FREQUENCY": ”16”
},

"assemble_v": {
   "FREQUENCY": ”25",
   "NUM_THREADS": "24",
   "UNCORE_FREQUENCY": ”14”
},

"gmres_solve": {
   "FREQUENCY": ”17",
   "NUM_THREADS": ”8",
   "UNCORE_FREQUENCY": ”22”
},

"print_vtu": {
   "FREQUENCY": "25",
   "NUM_THREADS": ”6",
   "UNCORE_FREQUENCY": ”24”
}
```

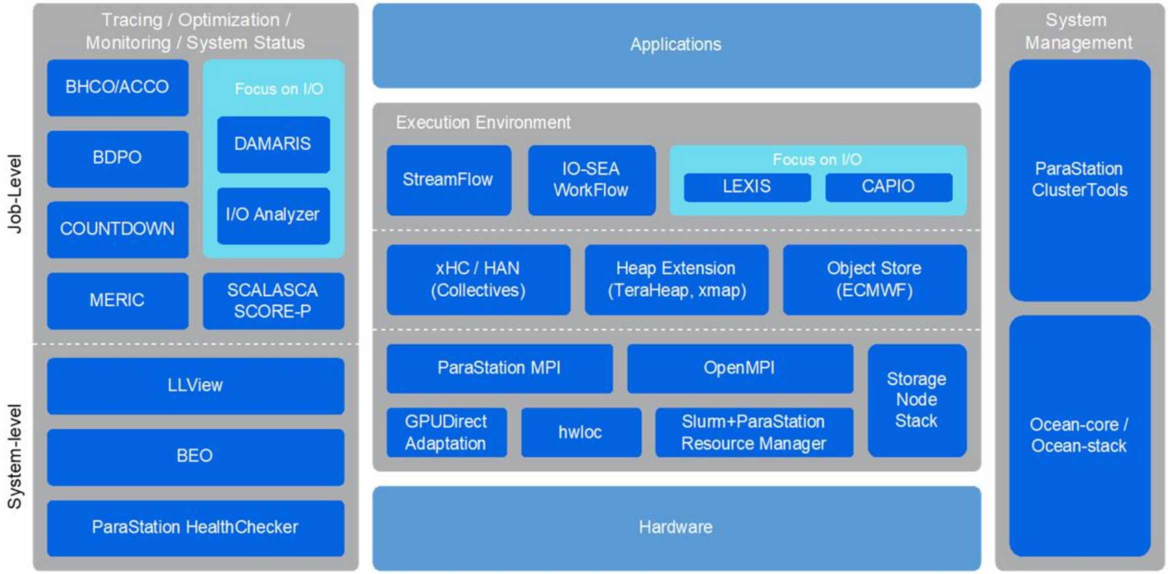
MERIC LIBRARY

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

MERIC

READEX

Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing



EUPEX European Pilot for Exascale

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

POWER KNOBS

- Intel
  - CPU - core frequency, uncore frequency, power capping
  - ACC (PVC) - GPU frequency, memory frequency, power capping
  - ACC (KNL) - core frequency, power capping
- AMD
  - CPU - core frequency, power capping, Data Fabric frequency
  - ACC - power capping, frequency – system, Data Fabric, display controller, SOC, memory, PCIe
- Nvidia
  - GPU - SM frequency, memory frequency, power capping
- IBM
  - CPU - core frequency, power capping
    + GPU and node power capping
- ARM
  - A64FX - core frequency, FLA (floating-point ops) and EXA (integer ops) pipelines elimination, memory frequency
  - EPI - core frequency, power capping, ???
  - Jetson - core frequency, memory frequency

MERIC PARAMETERS

export **MERIC_FREQUENCY**=2400MHz

export **MERIC_UNCORE_FREQUENCY**=2GHz

export **MERIC_NUM_THREADS**=24

export **MERIC_MEASURE**=RAPL,HDEEM-S

export **MERIC_COUNTERS**=papi

| And many more, see MERIC README

MERIC API

- void **MERIC_Init**()
  - At the beginning of the main() or in case of MPI applications follows after MPI_Init()
- void **MERIC_Close**()
  - At the end of application run, but before MPI_Finalize()

- void **MERIC_MeasureStart**(const char * regionName)
- double **MERIC_MeasureStop**(const char * regionName)
- double **MERIC_MeasureStopStart**(const char * regionName)
  - Optimized transition, removes switching to configuration of the parent region
- void **MERIC_CaptureScope**(const char * regionName)
  - Resource Acquisition Is Initialization (RAII)

- void **MERIC_IgnoreStart**()
- void **MERIC_IgnoreStop**()

STATIC TUNING WITHOUT INSTRUMENTATION

- tools/energyMeasuereStart + tools/energyMeasureStop
  - Commandline energy measurement
  - The tuneable parameters also possible to specify

```
$ ./energyMeasureStart -e RAPL
$ sleep 5
$ ./energyMeasureStop -e RAPL
Runtime [s] = 5.03672
RAPL_RAM_0 [J] = 38.2296
RAPL_RAM_1 [J] = 27.3747
RAPL_PCKG_0 [J] = 249.266
RAPL_PCKG_1 [J] = 256.062

RAPL Energy consumption [J] = 570.932
```

energyMeasureStart parameters:

- -e = energy measurement system "RAPL" or "NVML"
- -c = CPU core frequency [Hz]
- -u = CPU uncore frequency [Hz]
- -t = #OpenMP threads
- -p = power capping power limit [mW]
- -w = power capping time window [ms]
- -s = GPU SM frequency [Hz]
- -r = GPU memory frequency [Hz]
- -g = GPU power capping power limit [mW]

energyMeasureStop parameters:

- -e = energy measurement system "RAPL" or "NVML"
- -b = node baseline (static) power [W]
- -q = print the overall consumed energy only [J]

STATIC BINARY INSTRUMENTATION

Tool using Dyninst library (or MAQAO library) to produce a new binary that contains MERIC instrumentation

- Inserts all the necessary shared libraries dependencies
- Inserts MERIC_Init() and MERIC_Close()
  - In case of MPI applications generates also a new binary of MPI library that contains these functions
  - LD_PRELOAD=$(pwd)/libmpi.so mpirun -n $NUMPROC ./application [APP_PARAMS]
- Instruments all the selected application's functions
  - Detects selected functions in the binary and changes the instructions of the function to add MERIC_MeasureStart("funcName") call at the function beginning and MERIC_MeasureStop() call as the last function instruction
- How to select functions to instrument?
  - any profiler
  - or TIMEPROF (part of MERIC repository) provides runtime of the instrumented functions (application binary can be also instrumented with TIMEPROF using dinst_instrument.cpp tool)

DYNAMISM INVESTIGATION

- Dynamism investigation = running the application in different configurations
- MERIC stores measurements for each configuration for each instrumented application region
- **systemInfo** tool provides an overview what is the current status of the CPU and what are the available configurations

$ **meric/tools/systemInfo**

 # SYSTEM INFORMATION
Sockets per Node: 2
Cores per Socket: 8
Threads per Core: 2

 # CPU FREQUENCIES
Current scaling driver: intel_pstate
Current scaling governor: powersave
Available governors: performance powersave
Hardware controlled P-State: not available
Turbo CPU core frequencies: 3400000(1) 3400000(2) 3200000(3) 3100000(4) 3000000(5) 2900000(6) 2800000(7) 2800000(8) kHz(#cores)
Nominal CPU core frequency: 2600000 kHz
Min CPU core frequency: 1200000 kHz
Max CPU uncore frequency: 3000000 kHz
Min CPU uncore frequency: 1200000 kHz

 # RAPL POWER LIMITS
RAPL time window unit: 976.562 us
PKG max power limit: 180 W
PKG min power limit: 34 W
DRAM max power limit: 36 W
DRAM min power limit: 16.5 W

 # DEFAULT RAPL POWER LIMITS
PKG power limit #1: enabled + clamping enabled
PKG power limit #1: 90 W
PKG time window #1: 1 s
PKG power limit #2: enabled + clamping enabled
PKG power limit #2: 108 W
PKG time window #2: 0.0078125 s

 # AVAILABLE ENERGY MEASUREMENT SYSTEMS
RAPL

DYNAMISM INVESTIGATION

MERICwrapper

- Provides algorithms for state space search – the tool will execute the application in various configurations to find the optimal one for each region
- A json configuration file:

```
{
    "MPI" : "true",
    "PARAMETERS" : {
        "FREQUENCY" : {
            "MAX" : 3600000000,
            "MIN" : 1200000000,
            "STEP":  200000000
        },
        "UNCORE_FREQUENCY" : {
            "MAX" : 2800000000,
            "MIN" : 1200000000,
            "LIST": [2800000000,2100000000,
              1600000000,1200000000]
        },
        "NUM_THREADS" : {
            "MAX" : 36,
            "MIN" : 1,
            "STEP":  4
        }
    },
    "MERIC" : {
        "MEASURE" : "RAPL",
        "PWRCAP_POWER" : 0,
        "PWRCAP_TIME"  : 0,
        "COUNTERS" : "msr",
        "AGGREGATE"  : 1,
        "CONTINUAL"  : 1,
        "DETAILED"  : 1,
        "SAMPLES"   : 0,
        "BARRIERS"  : "all",
        "OUTPUT_DIR" : "mericMeasurement"
    },
    "ALGORITHM" : {
        …
    }
}
```

```
"ALGORITHM" : {
    "NAME" : "EXHAUSTIVE"
}
```

```
"ALGORITHM" : {
    "NAME" : "EVO",
    "PARAMETERS" : {
        "END_CONDITION" : 2,
        "POPULATION" : 10
    }
}
```

```
"ALGORITHM" : {
    "NAME" : "PSO",
    "PARAMETERS" : {
        "END_CONDITION" : 2,
        "PARTICLES" : 10,
        "CONST_CP"  : 2.05,
        "CONST_CG"  : 2.05,
        "CONST_W"   : 0.9
    }
}
```

ANALYSIS WORKFLOW

1. **Application profiling**
2. **Identification of significant regions**
3. **Application instrumentation**
4. **Application analysis**
5. **Optimal configuration identification**
6. **App’s behavior visualization**
7. **Production runs and dynamic tuning**

Binary instrumentation tool

Dyninst / MAQAO

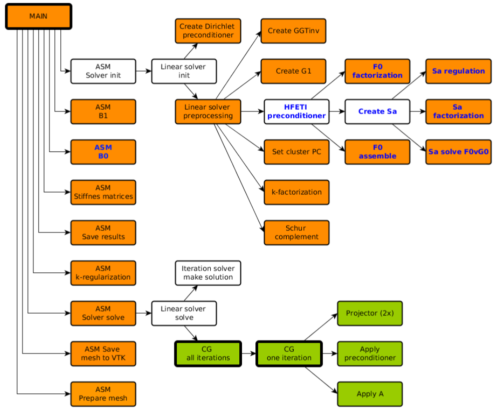
MERIC wrapper

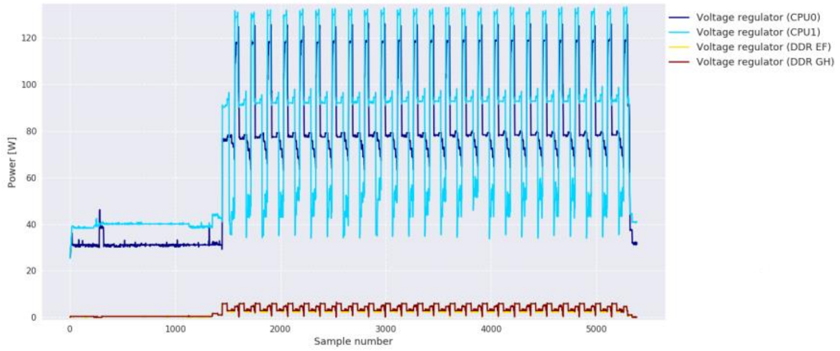
RADAR visualizer

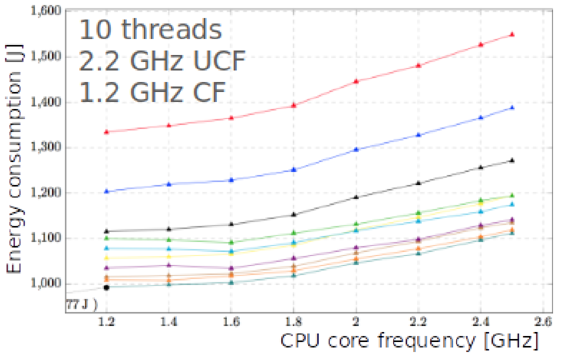
PyQt5 tool

RADAR VISUALIZER

- *Visualisation of applicatin behavior in various configuration*
- Tables
  - Overall application evaluation
  - Summary of nested regions' behavior
  - Each region behavior description
- Heatmaps
- Plots
- Power timeline
  - Cluster analysis
- Call-path graph





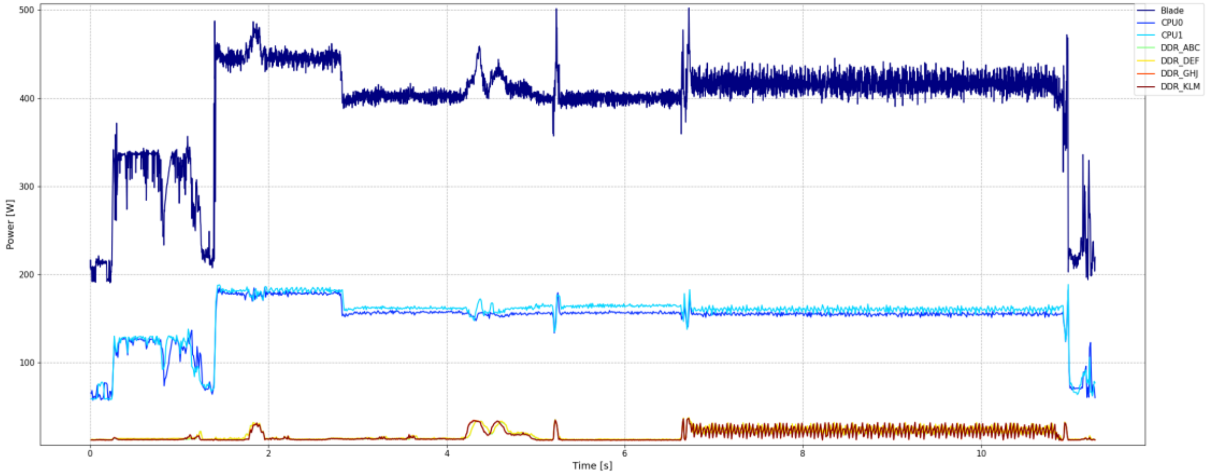


	Default settings	Default values	Best static configuration	Static savings	Dynamic savings
Runtime of function [s], Job info - rapl	3.0GHz, 2.5GHz	1.97s	3.0GHz, 2.5GHz	0.00s (0.00%)	0.015s of 1.97s (0.76%)
Energy summary, COUNTERS - rapl:	3.0GHz, 2.5GHz	800.37	2.4GHz, 2.5GHz	19.70 (2.46%)	46.52 of 780.67 (5.96%)
Run-time change with the energy optimal settings	+0.14s (107.04 % of default time)				

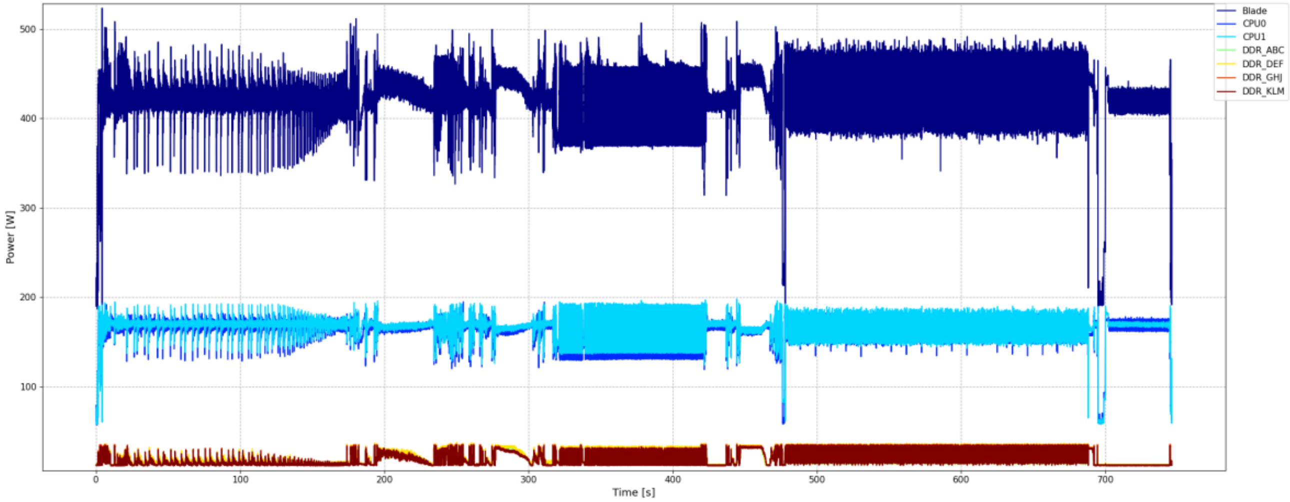
Uncore freq [GHz] / Core freq [GHz]	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
1.2	13,200.02	12,717.1	12,621.78	12,410.62	12,380.68	12,507.38	12,774.16	13,108.6	13,604.2	14,040.8
1.4	13,161.9	12,597.78	12,125.18	12,065.52	12,074.54	12,173.36	12,312.24	12,802.26	13,095.84	13,450.8
1.6	13,320.66	12,640.76	12,256.22	12,033.62	11,966.36	11,992.7	12,372.04	12,579.22	13,126.44	13,370.24
1.8	13,878.04	13,082.66	12,700.92	12,457.08	12,373.86	12,445.98	12,574.6	12,831.82	13,081.62	13,296.04
2	14,218.58	13,327.12	12,902.62	12,544.82	12,456.82	12,494.8	12,680.32	13,038.86	13,207.38	13,474.8
2.2	14,625.62	13,849.58	13,240.14	12,851	12,760.98	12,802.24	12,993.44	13,260.38	13,497.6	13,767.62
2.4	15,083.2	14,412.62	13,568.68	13,447.18	12,973.38	13,238.6	13,332.7	13,388.7	13,777.68	14,030.66
2.5	15,554.96	14,465.2	13,991	13,553.84	13,300.24	13,354.46	13,472.36	14,179.16	14,083.06	14,231.3

SERVICES

- **waLBerla**



- **LaBS**



DRIVING THE EXASCALE TRANSITION

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

SERVICES

- **waLBerla**

	Default	Static tuning	Dynamic tuning constant runtime	Dynamic tuning
Runtime [s]	66.86	70.75	66.64	77.69
Energy consumption [kJ]	99.04	90.88	91.24	80.17
Solver energy-efficiency [MLups/W]	0.548	0.667	0.595	0.734
Runtime extension [%]	-	5.82	−0.4	16.2
Energy savings [%]	-	9.2	7.9	19.1

- **LaBS**

	Default	Static tuning	Dynamic tuning constant runtime	Dynamic tuning
Runtime [s]	1797.9	1942.73	1807.13	1871.14
Energy consumption [kJ]	3102.3	1942.73	2726.7	2496.71
Solver energy-efficiency [MLups/W]	0.054	0.059	0.056	0.056
Runtime extension [%]	-	8.1	0.5	4.1
Energy savings [%]	-	15.1	12.1	19.5

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.

EURO
CZECHIA

ACKNOWLEDGEMENT

- This work was supported by the **READEX** project - the European Union's Horizon 2020 research and innovation programme under grant agreement No. 671657.
- This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (**NPS II**) project "IT4Innovations excellence in science - LQ1602" and by the IT4Innovations infrastructure which is supported from the Ministry of Education, Youth and Sports of the Czech Republic through the **e-INFRA** CZ (ID:90140).
- This work was supported by the **POP2** project - the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824080.
- This work was supported by the **SCALABLE** project. This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956000. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, the CzechRepublic.
- This work was supported by the **EUPEX** project - the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101033975.

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

e-INFRA CZ

SCALABLE

EUPEX European Pilot for Exascale

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | IT4INNOVATIONS NÁRODNÍ SUPERPOČÍTAČOVÉ CENTRUM