VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

ENERGY EFFICIENCY ANALYSIS OF PARALLEL APPLICATIONS

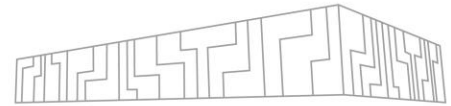
Ondřej Vysocký
IT4Innovations

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

MINISTRY OF EDUCATION, YOUTH AND SPORTS

ENERGY EFFICIENCY

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



1] Measure performance

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

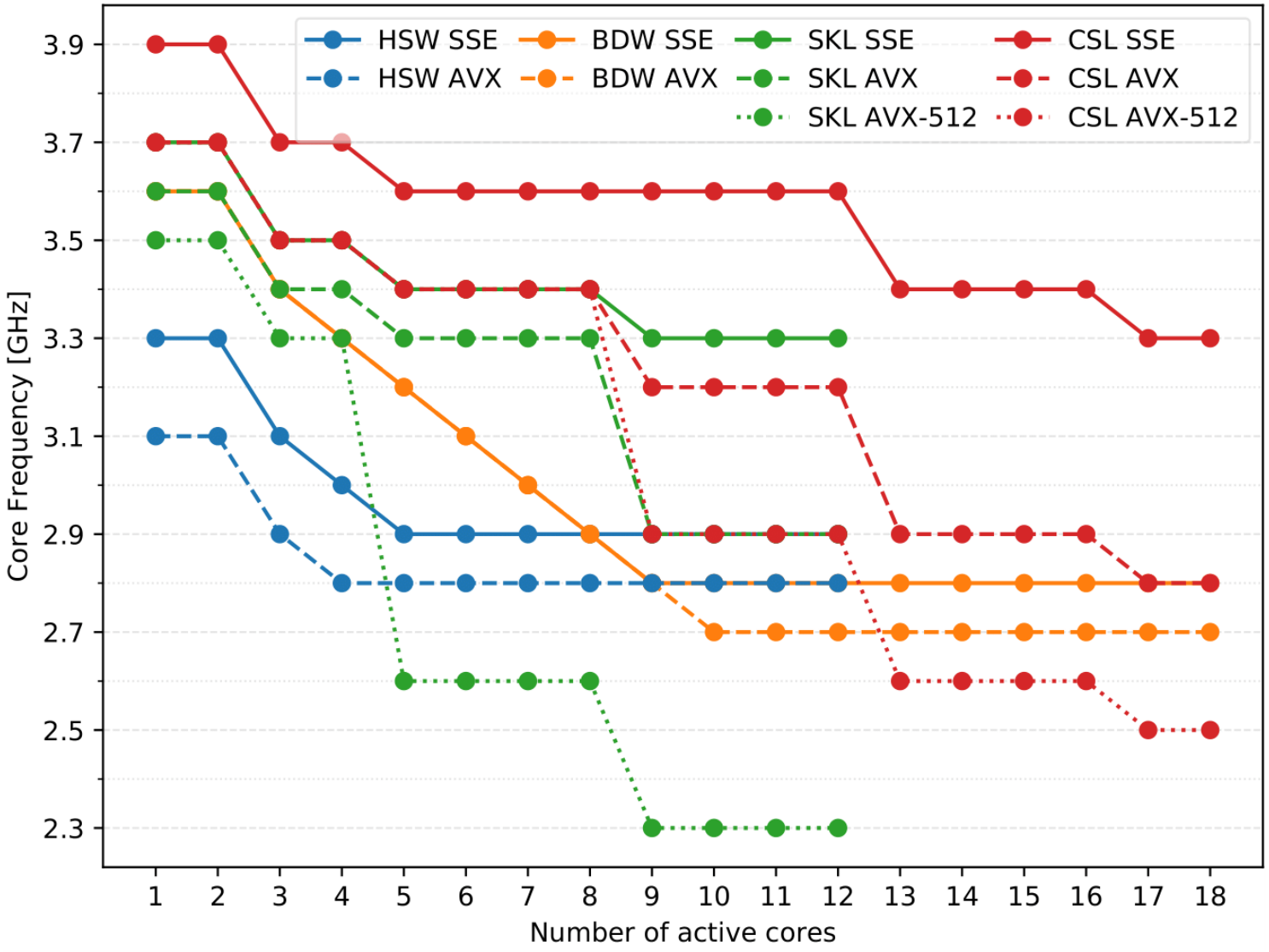
https://www.top500.org/files/green500/tutorial.pdf

PERFORMANCE

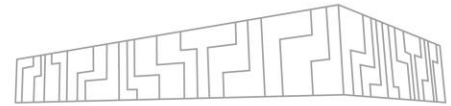
- Reported by the application (solver)
  - FLOPs, MLUPs, ...
- Evaluate FLOPs

event name	FLOPs/inst	#inst.	FLOPs
SCALAR_DOUBLE	1	7240374183451	9164881641636
SCALAR_SINGLE	1	154346452896	154346452896
128B_PACKED_DOUBLE	2	557726129284	1115452258568
128B_PACKED_SINGLE	4	254594513	1018378052
256B_PACKED_DOUBLE	4	2725451737	10901806948
256B_PACKED_SINGLE	8	0	0
512B_PACKED_DOUBLE	8	0	0
512B_PACKED_SINGLE	16	0	0
Runtime [s]	105.38	**Sum [FLOPs]**	8522093079915
		GFLOPs/s	80.87

INSTRUCTION SET vs FREQUENCY



HSW - Xeon E5-2680v3
BDW - Xeon E5-2697v4
SKL - Xeon Gold 6126
CSL - Xeon Gold 6240

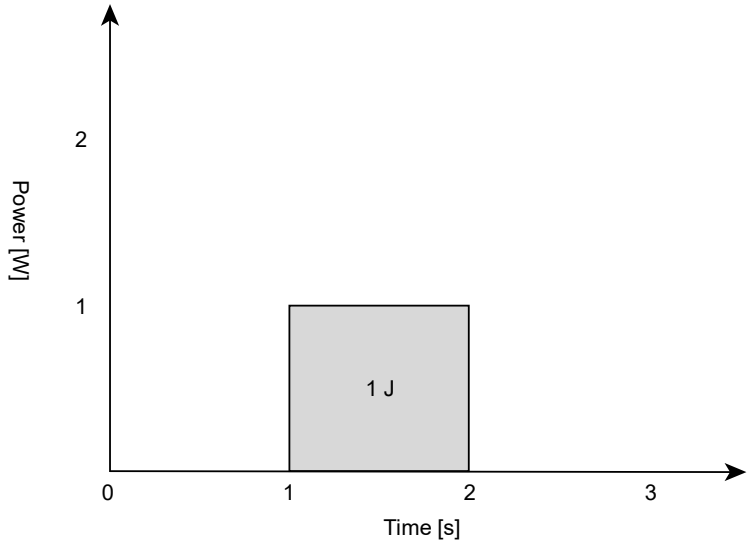


2] Measure energy

ENERGY

$$Energy = Power \times Time$$

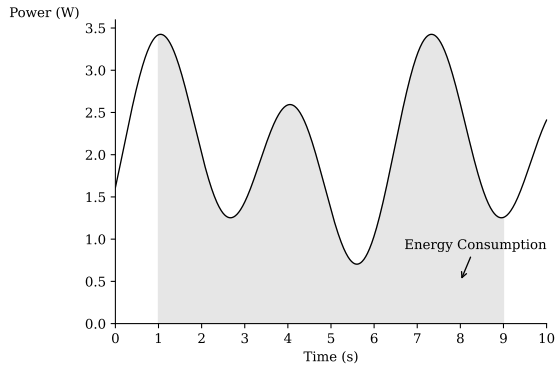
- Power [W]
- 1 W * 1 s = 1 J
- 1 W * 1 h = 1 Wh = 3 600 J



ENERGY

$$Energy = Power \times Time$$

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

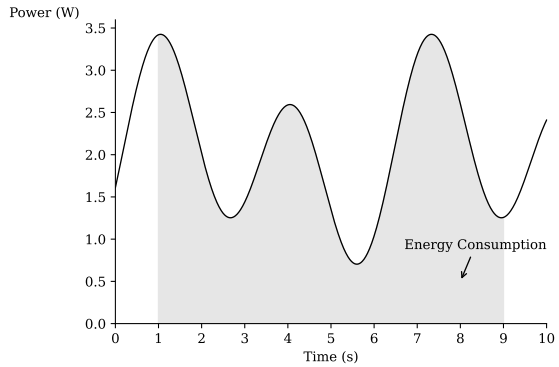


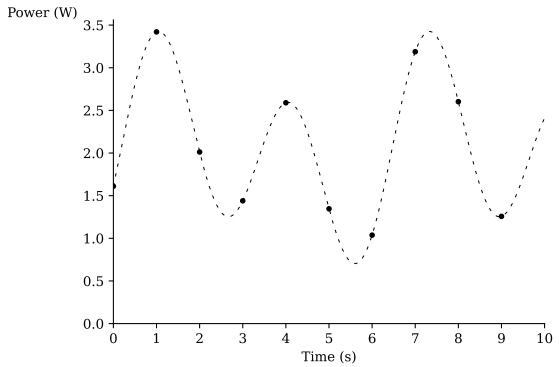
Img source, Luís Cruz (TU Delft)

ENERGY

$$Energy = Power \times Time$$

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



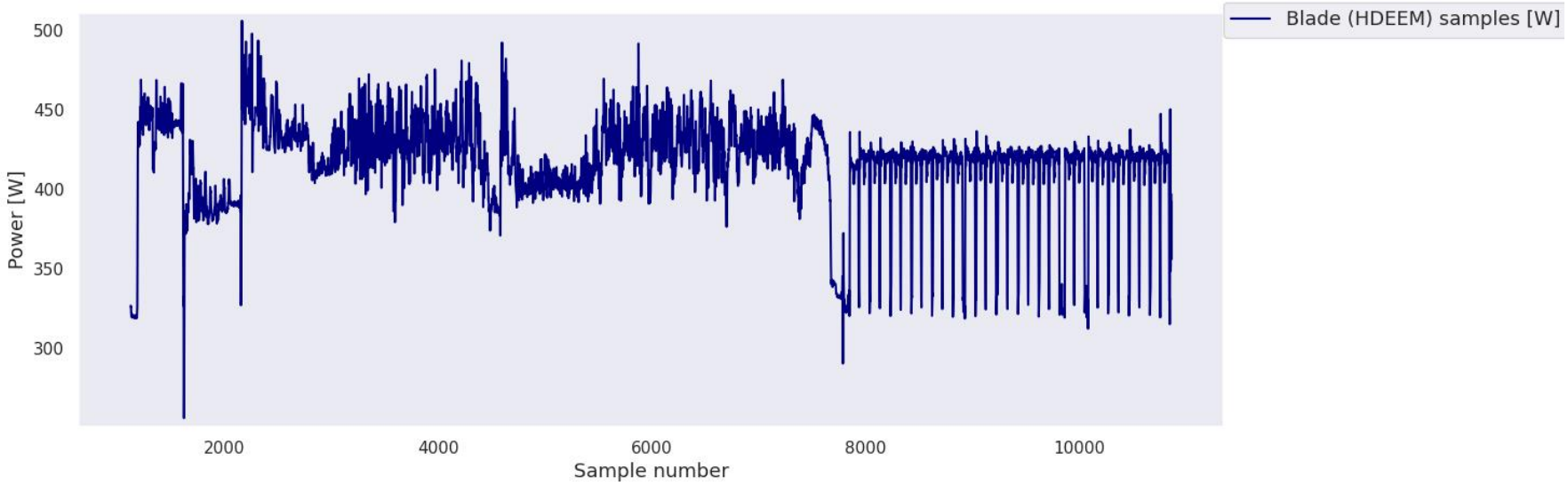


Img source, Luís Cruz (TU Delft)

ENERGY

$$Energy = Power \times Time$$

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

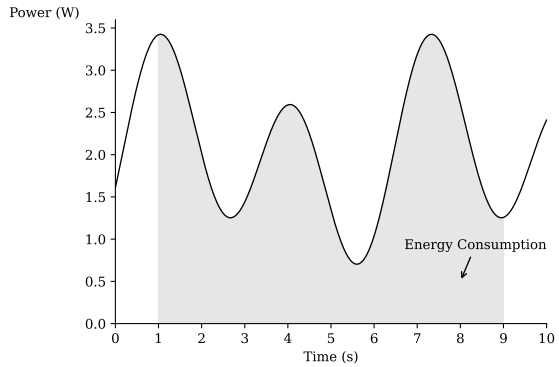


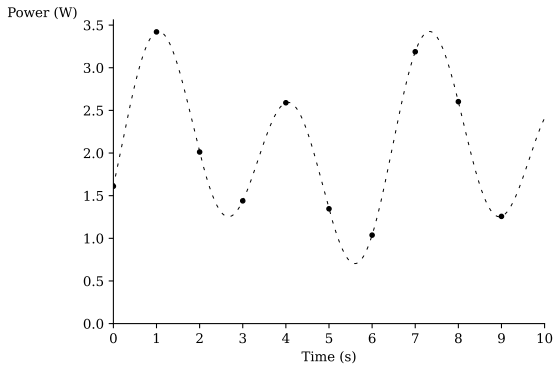
ENERGY

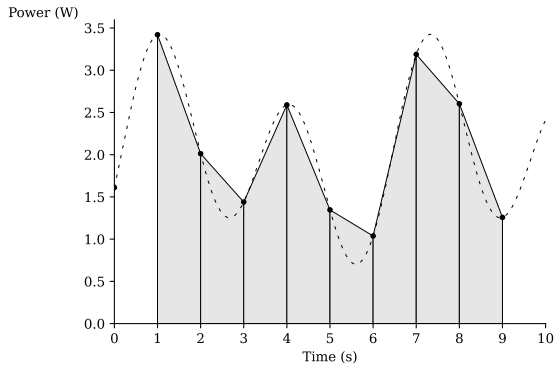
$$Energy = Power \times Time$$

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

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

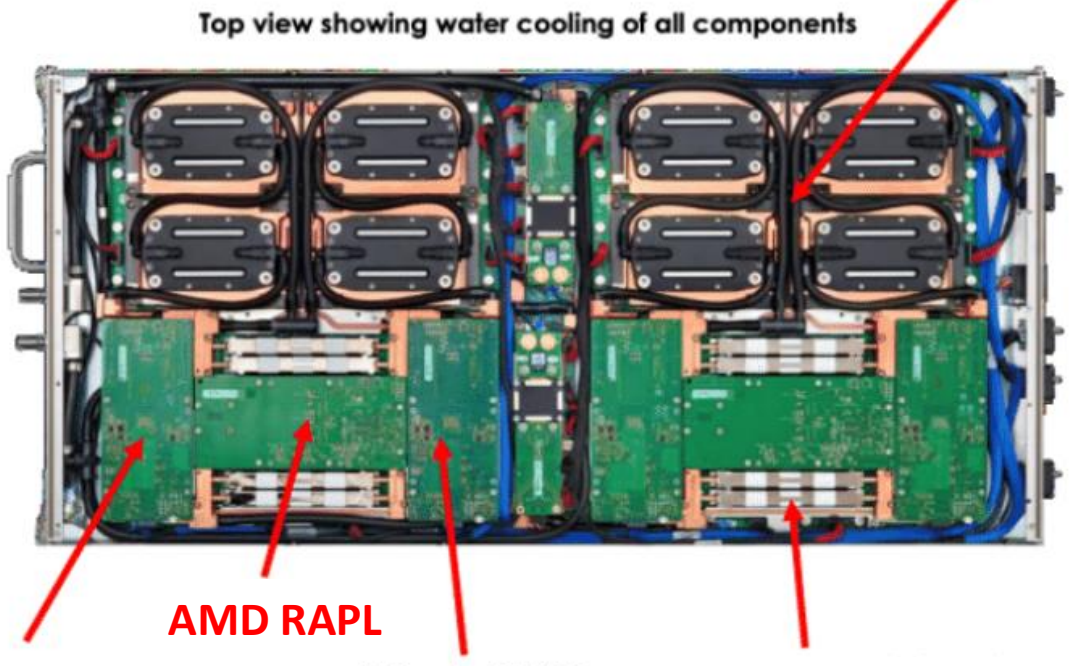




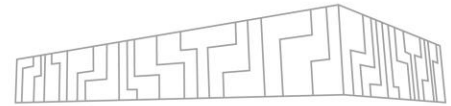


Img source, Luís Cruz (TU Delft)

POWER MONITORING



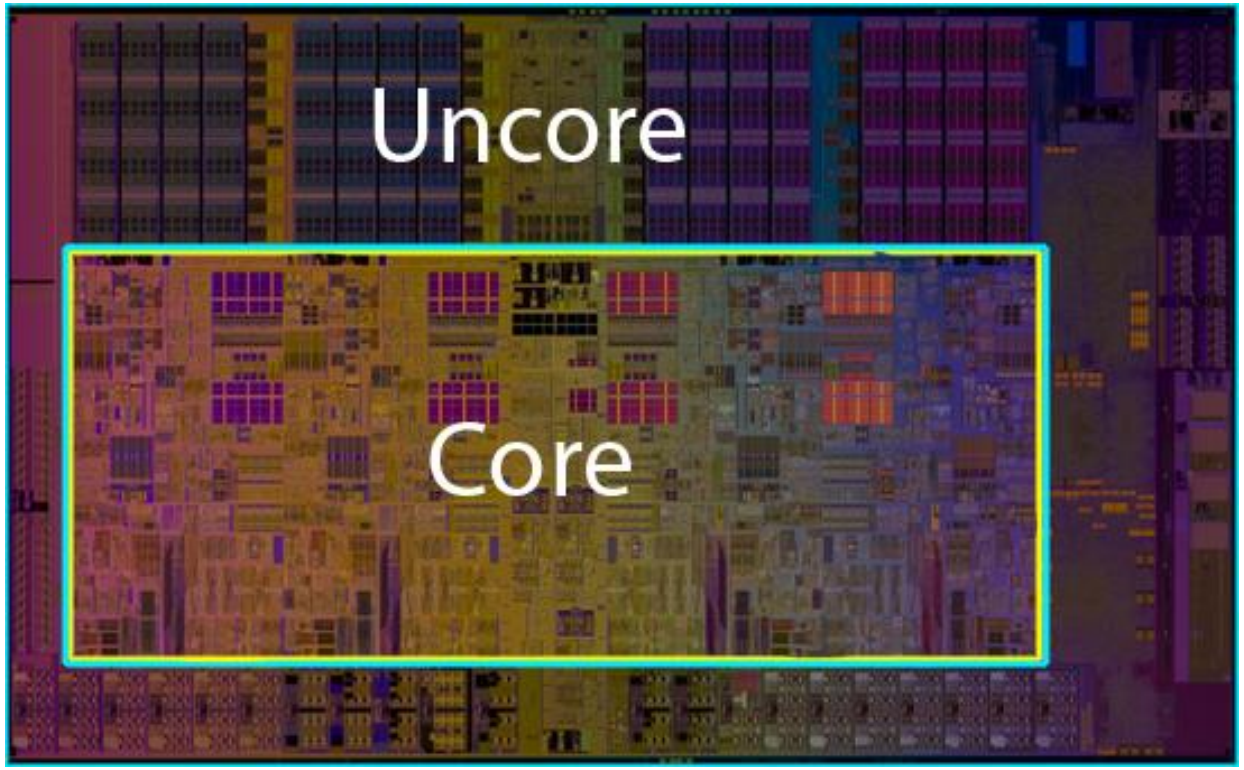
Images of HPE Cray, OLCF Frontier

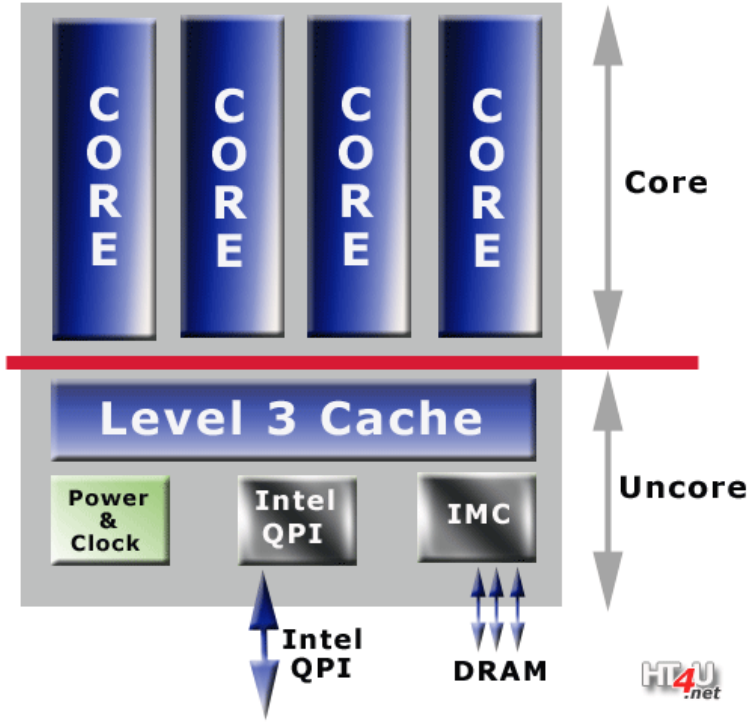


3] Know your hardware

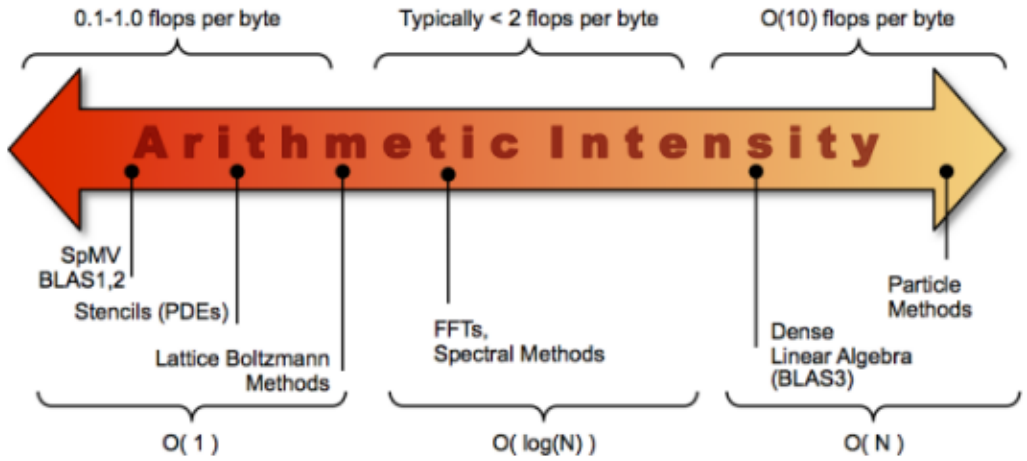
INTEL CPU UNCORE FREQUENCY

- 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

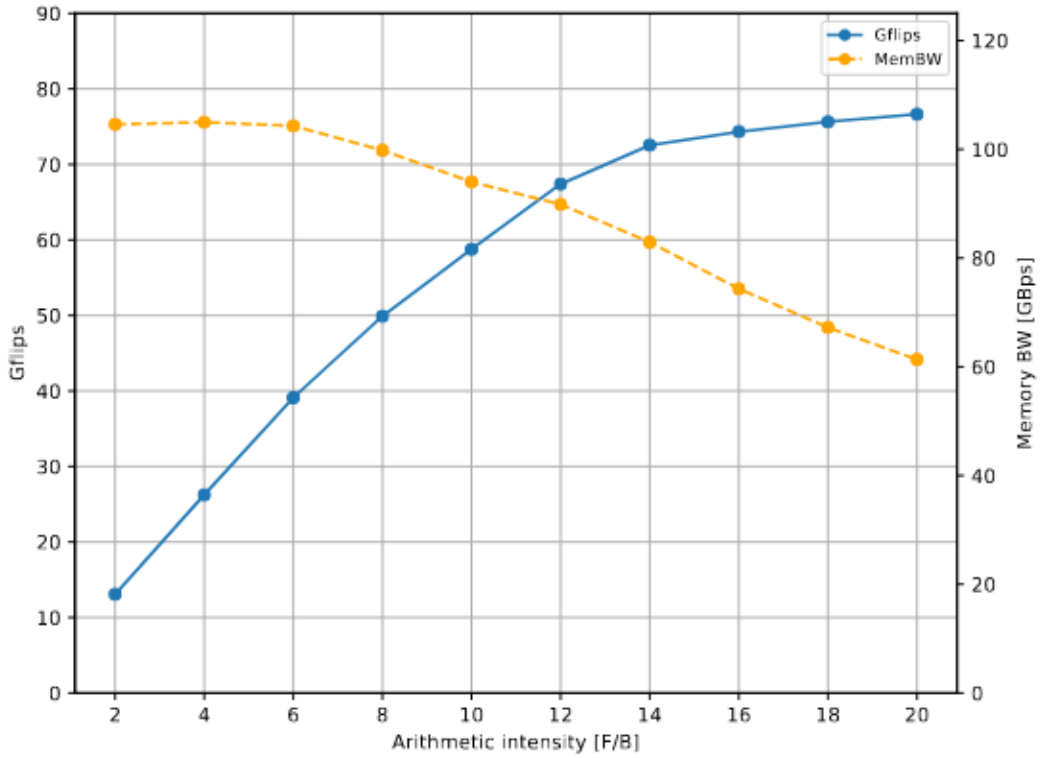




WORKLOAD BOUNDNESS



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

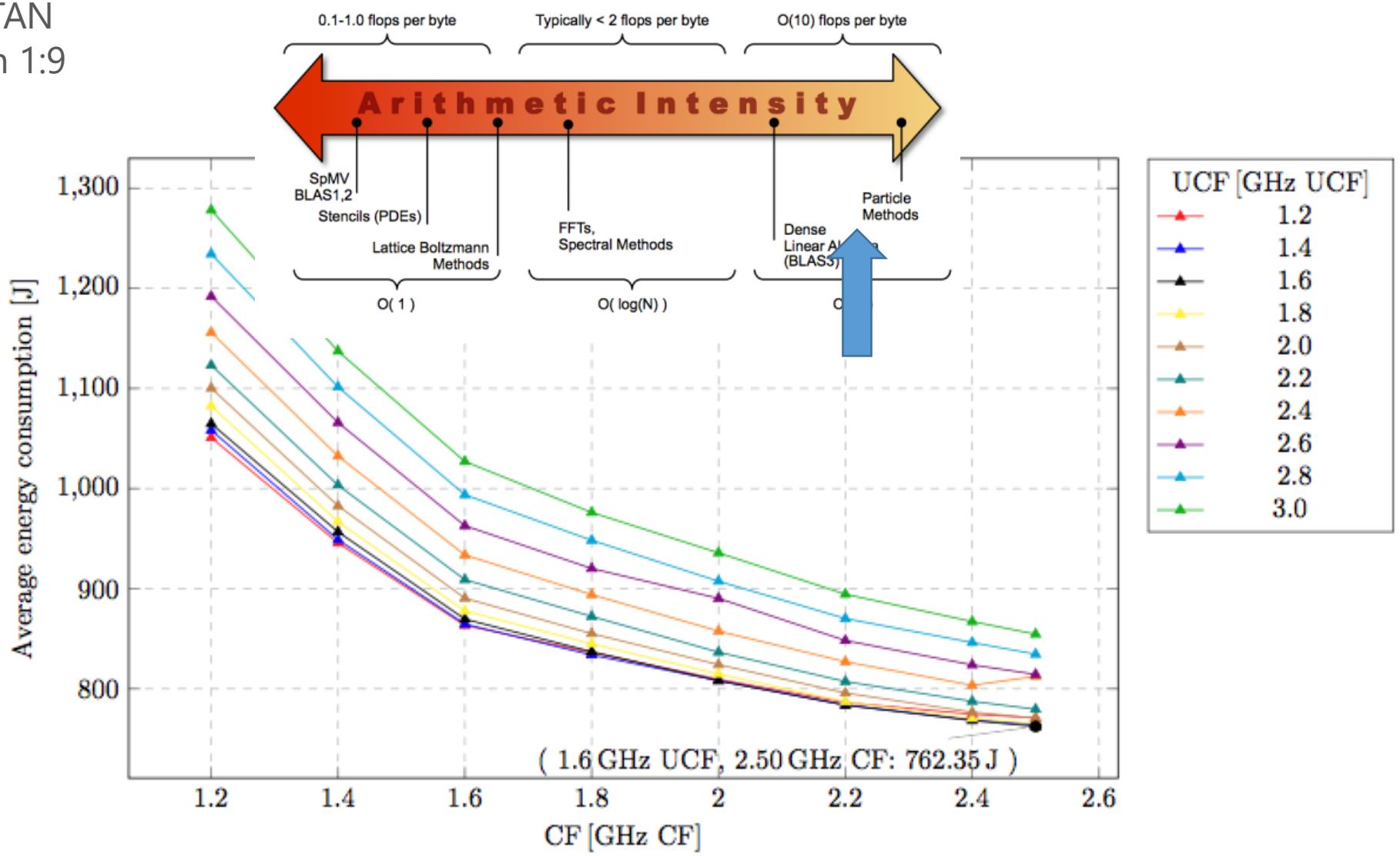


(b) Roofline model of the Intel Xeon Gold 6240 processor when executing a workload of AVX-512 instructions.

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

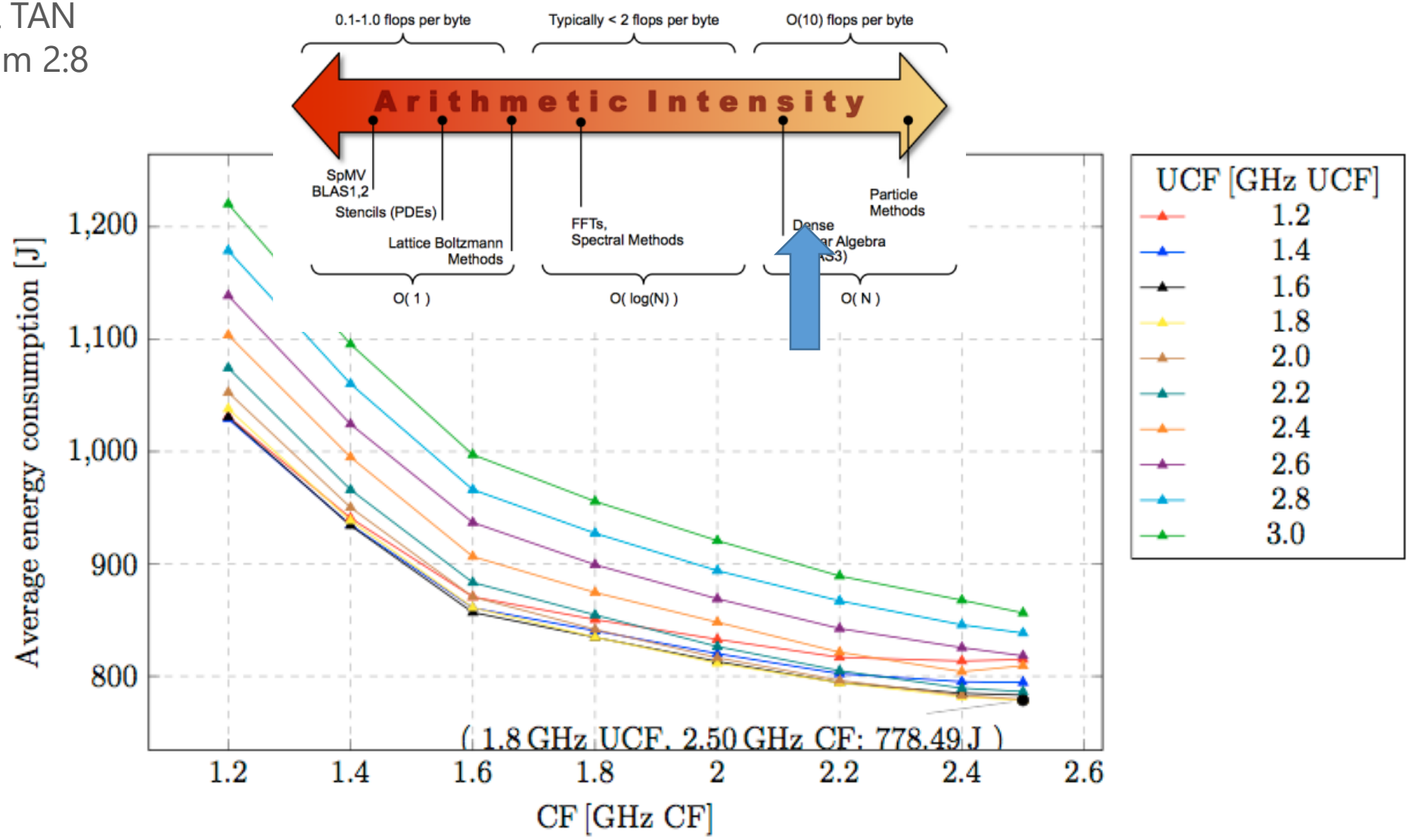
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 1:9



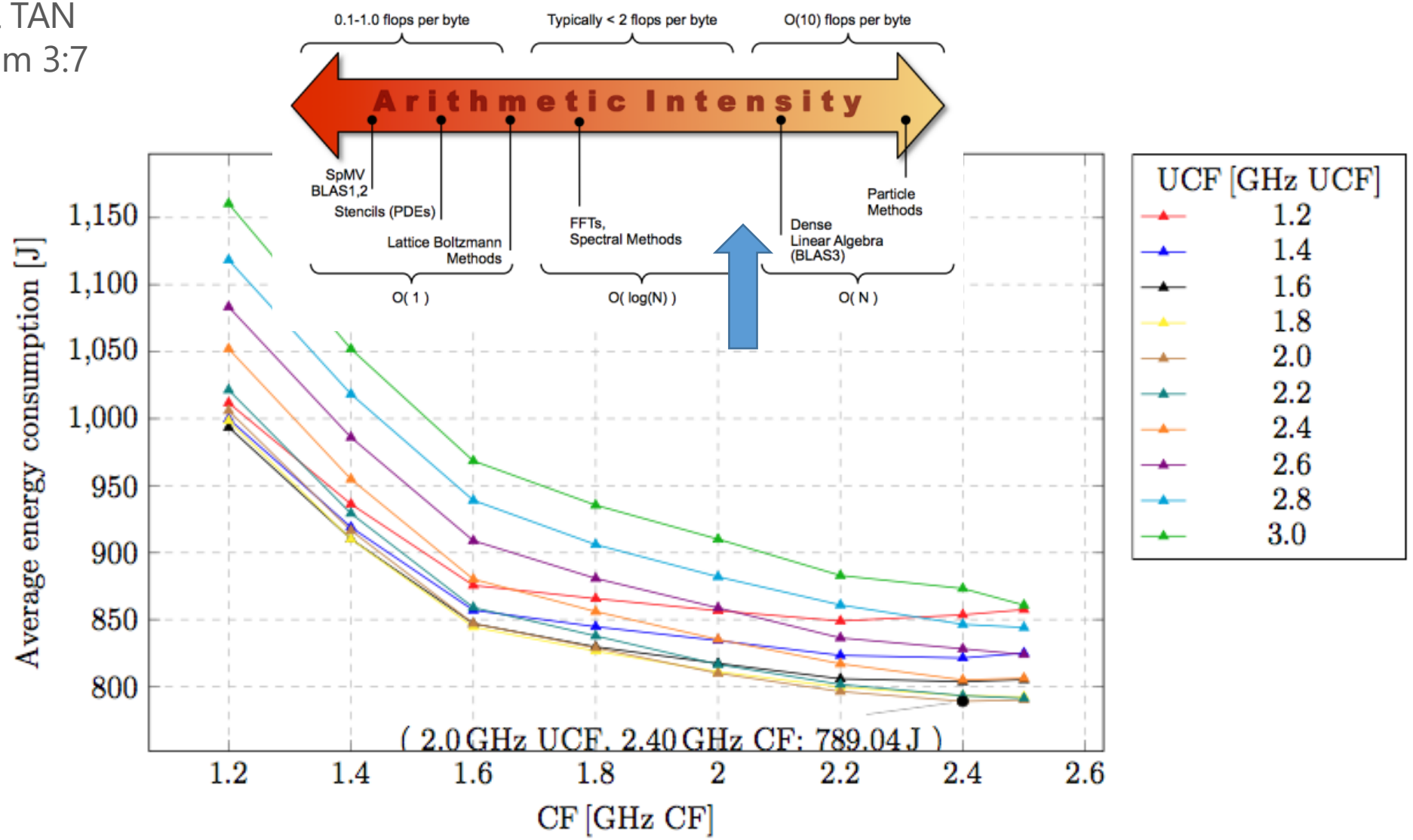
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 2:8



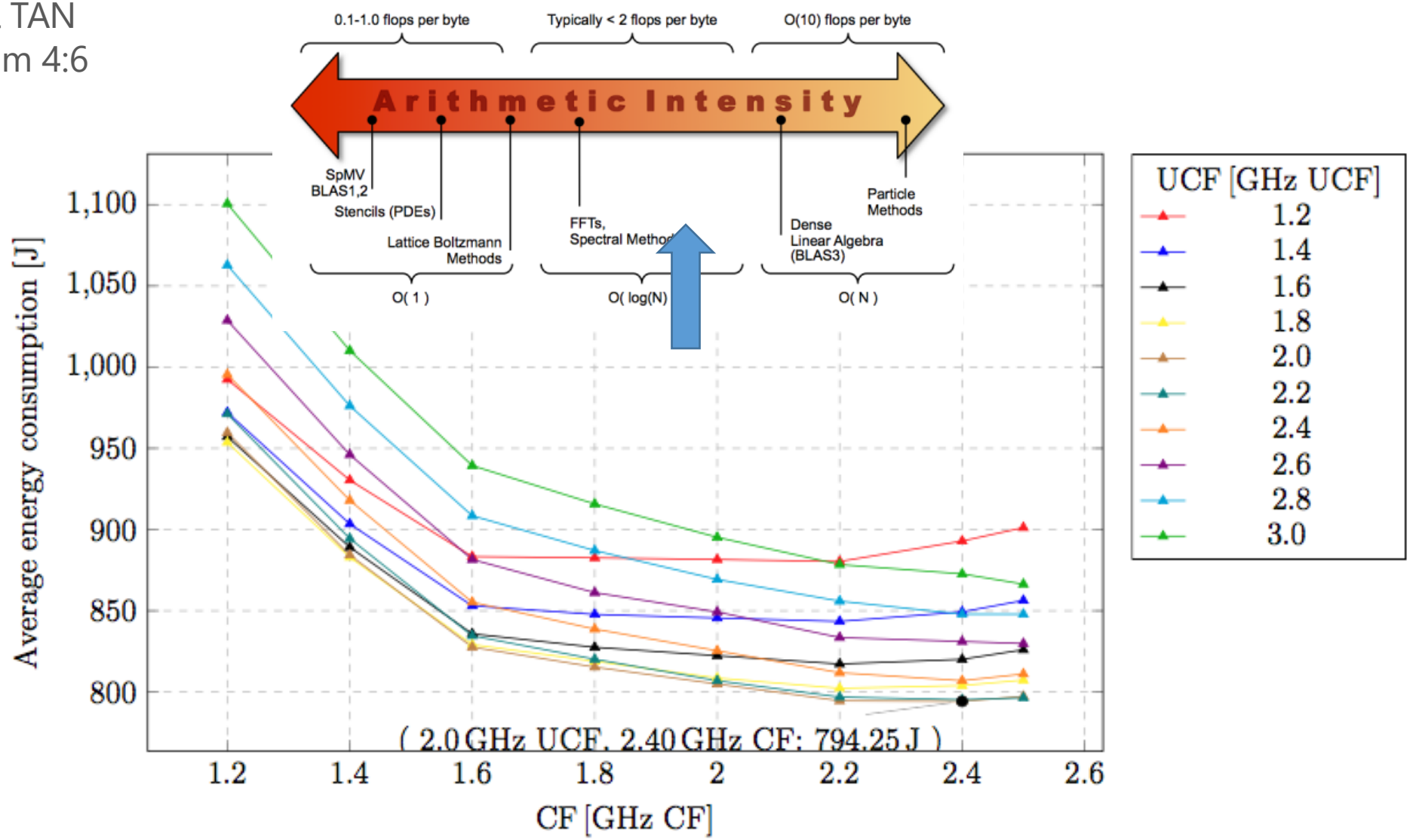
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 3:7



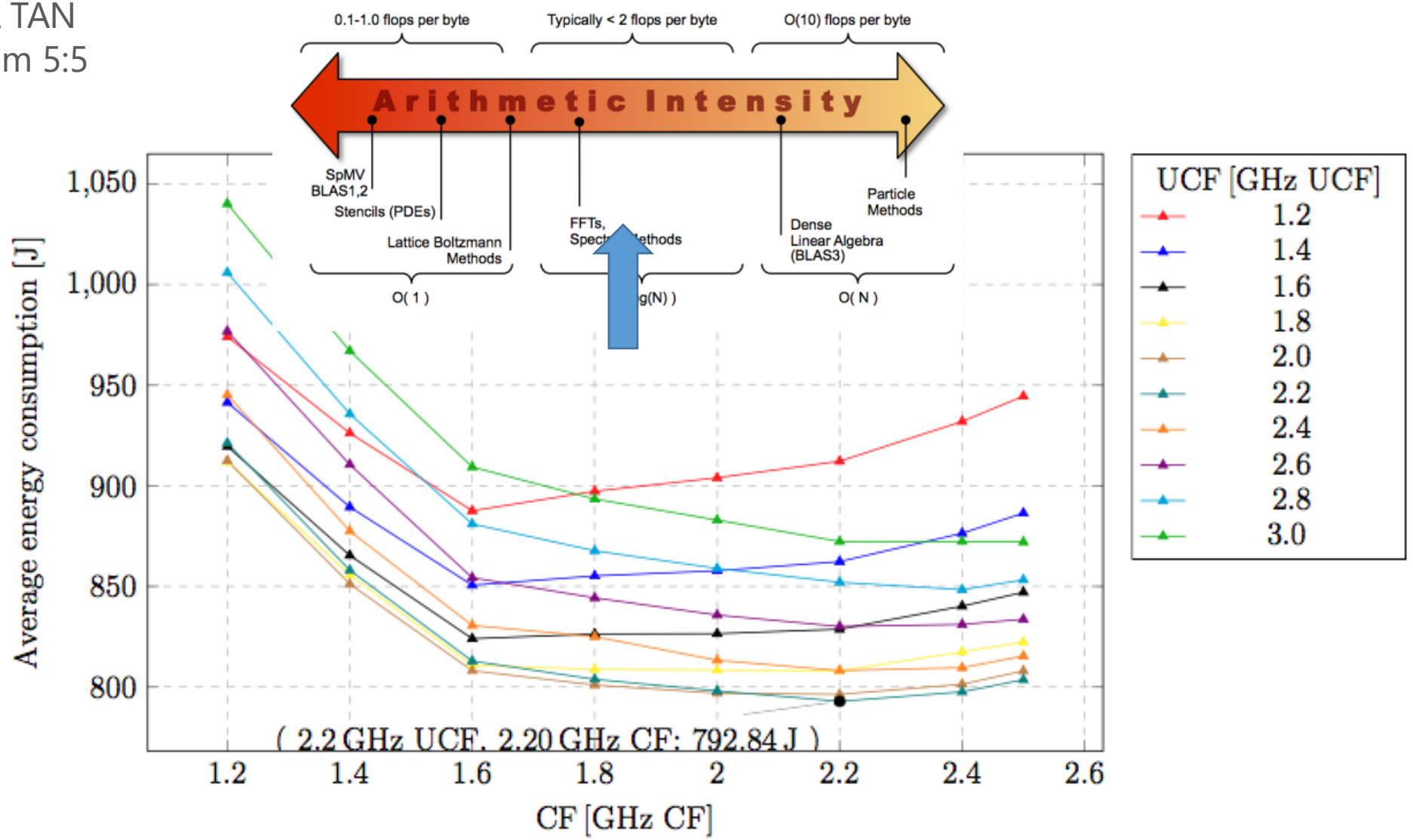
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 4:6



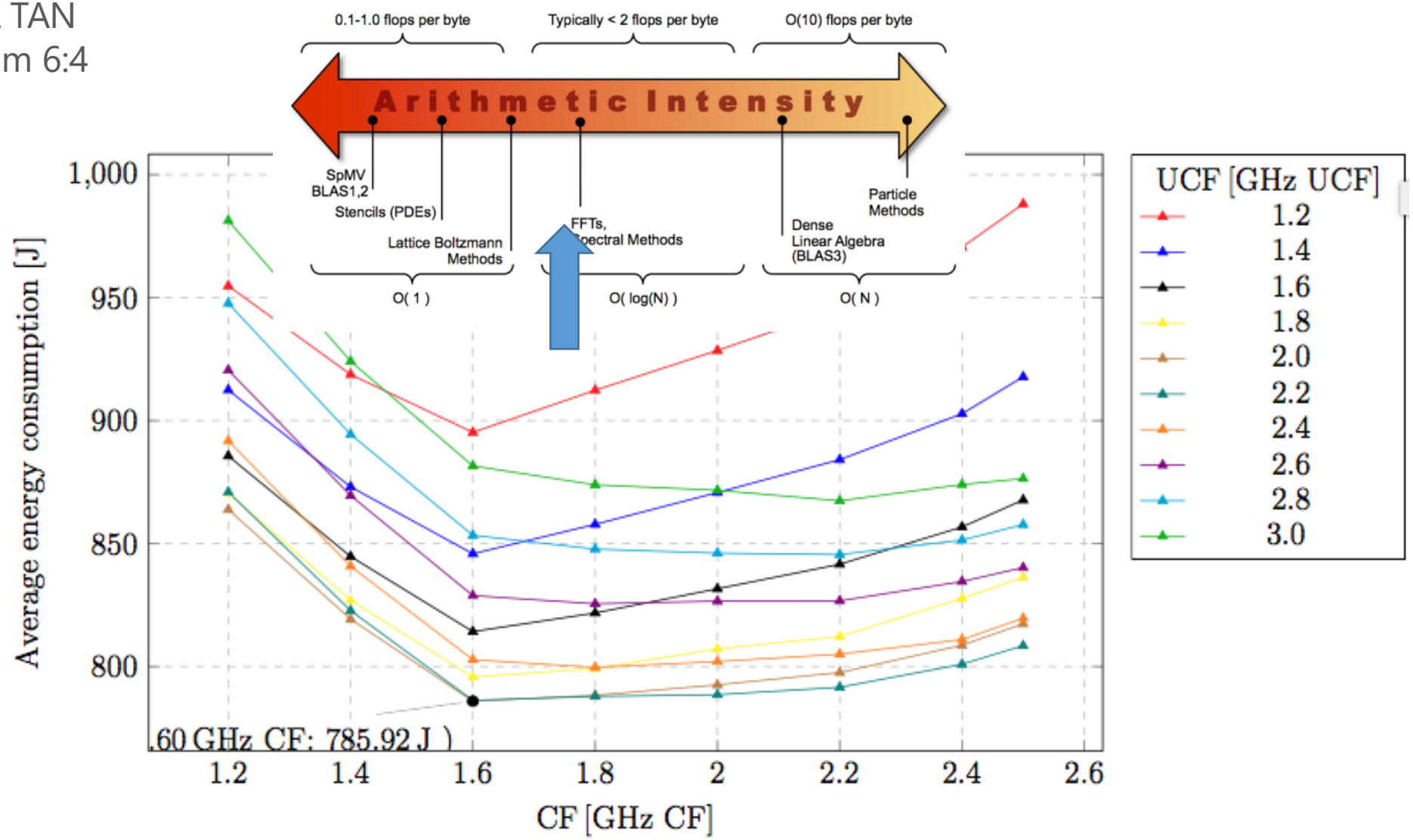
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 5:5



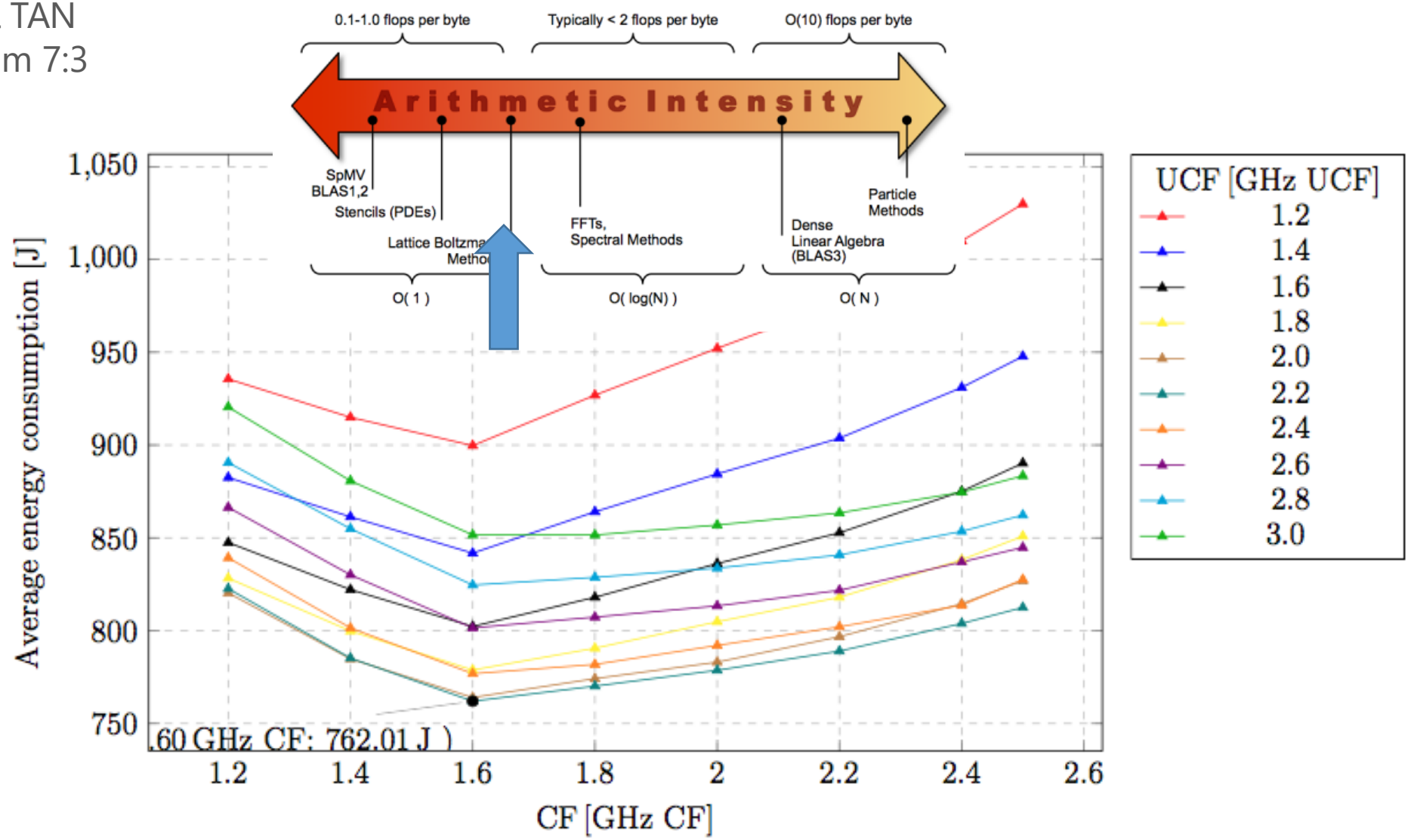
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 6:4



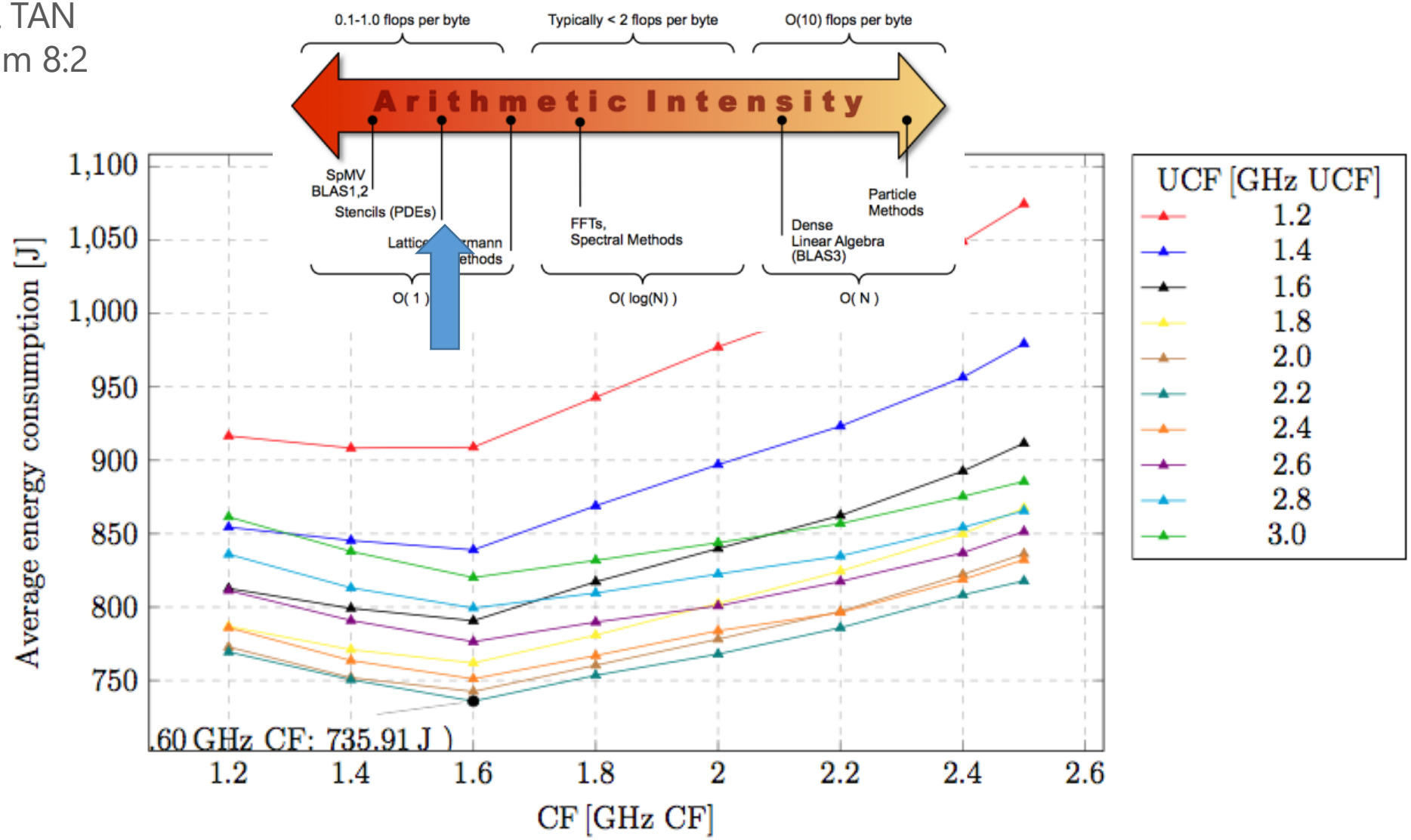
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 7:3



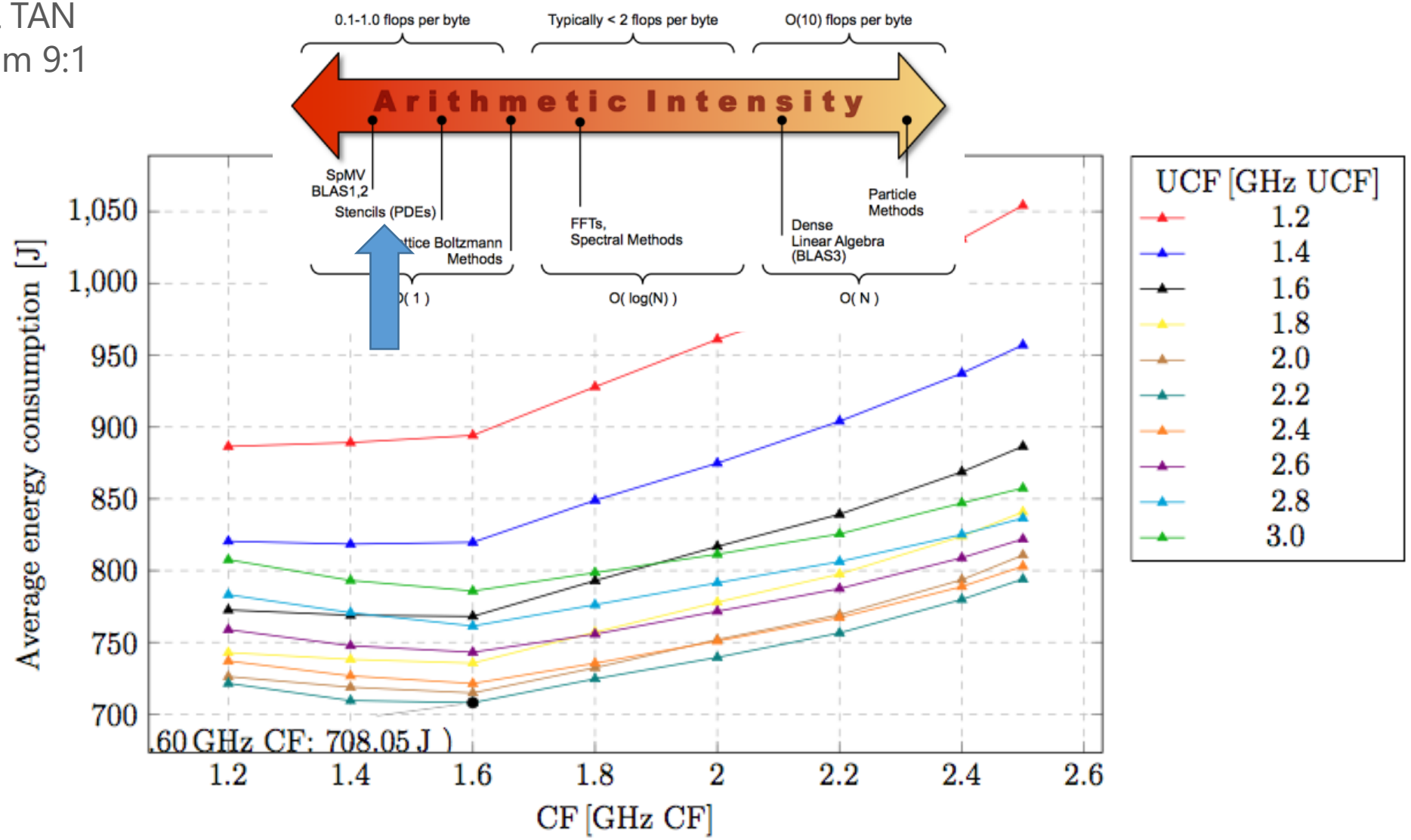
ARITHMETIC INTENSITY

GEMV & TAN
Ratio from 8:2

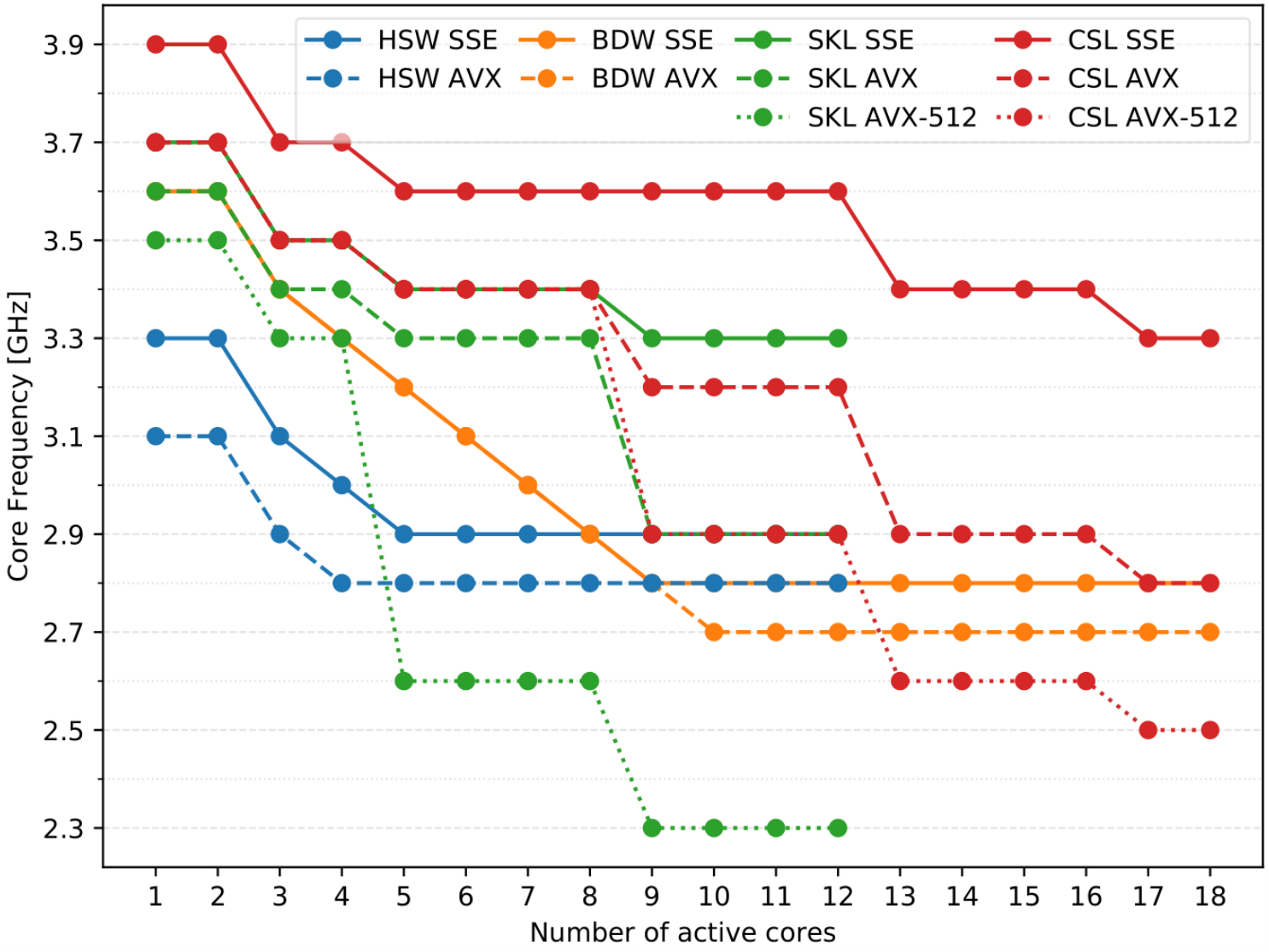


ARITHMETIC INTENSITY

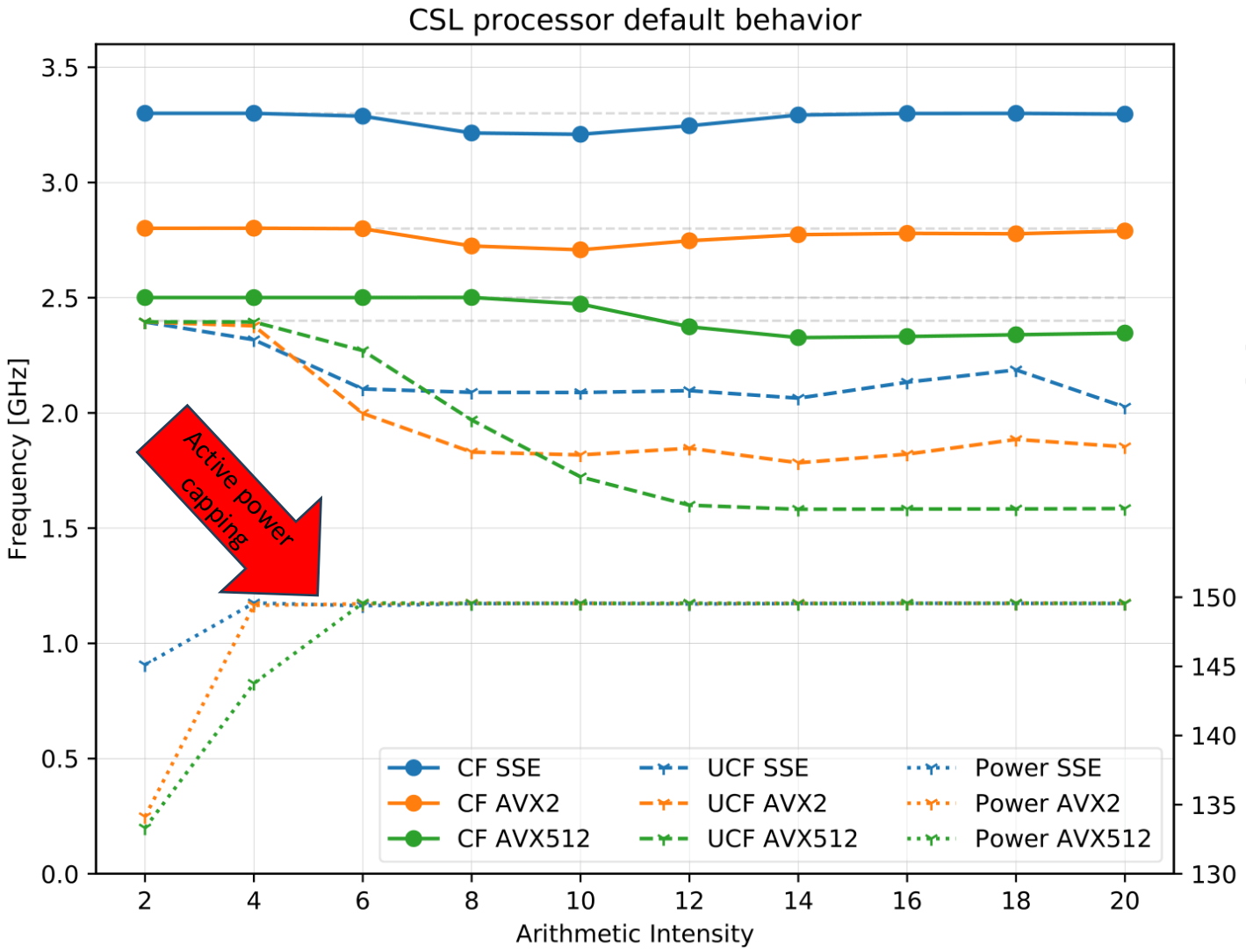
GEMV & TAN
Ratio from 9:1



MAXIMUM CORE FREQUENCY



CASCADELAKE CPU BEHAVIOR

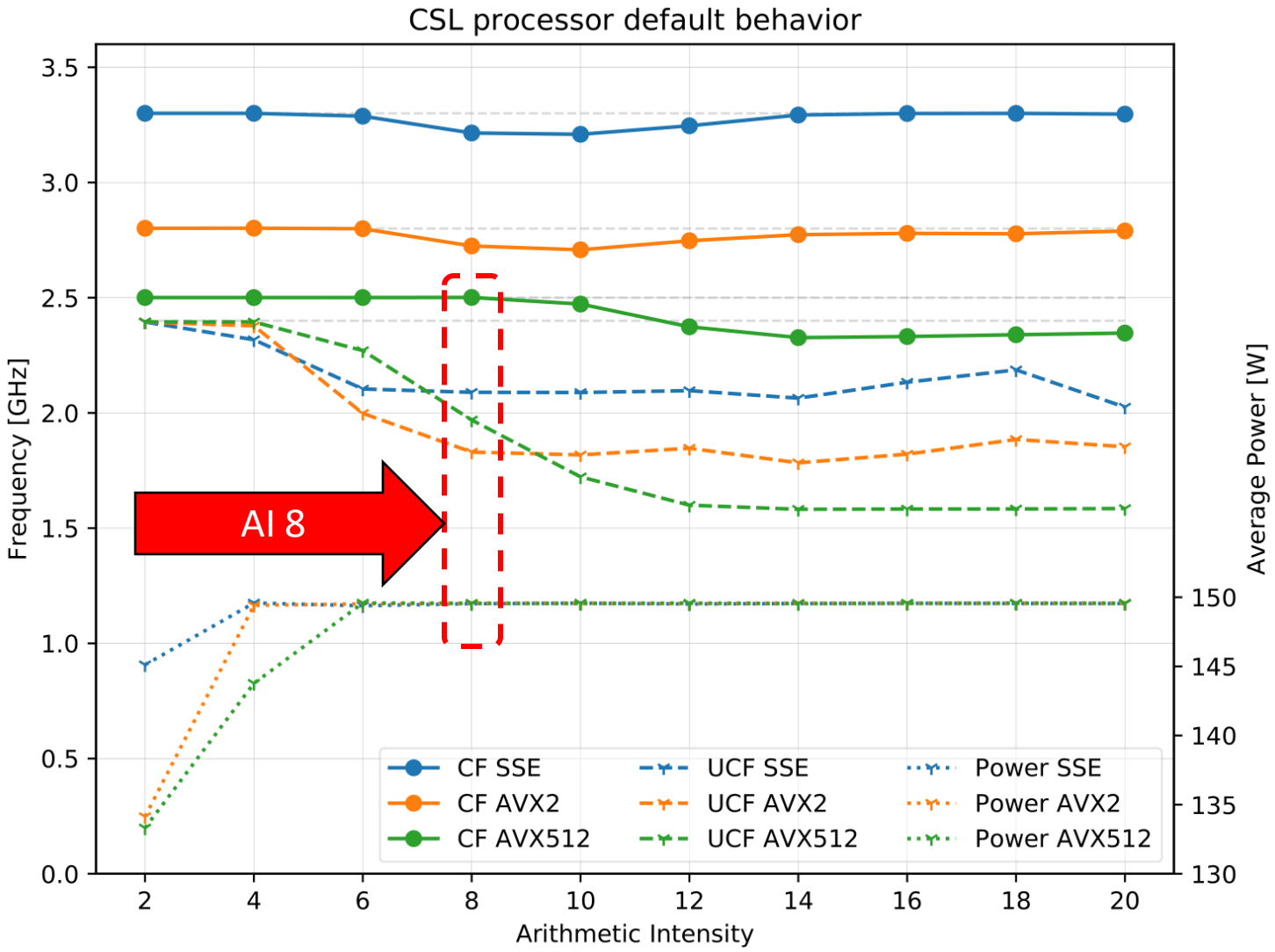


- Power capping system downscales CPU core and uncore frequencies to keep power consumption at the power limit
- Intel RAPL does not reflect the arithmetic intensity of the workload

Thermal Design Power

- Different e-e analysis results for a different power limit (default power limit = TDP)

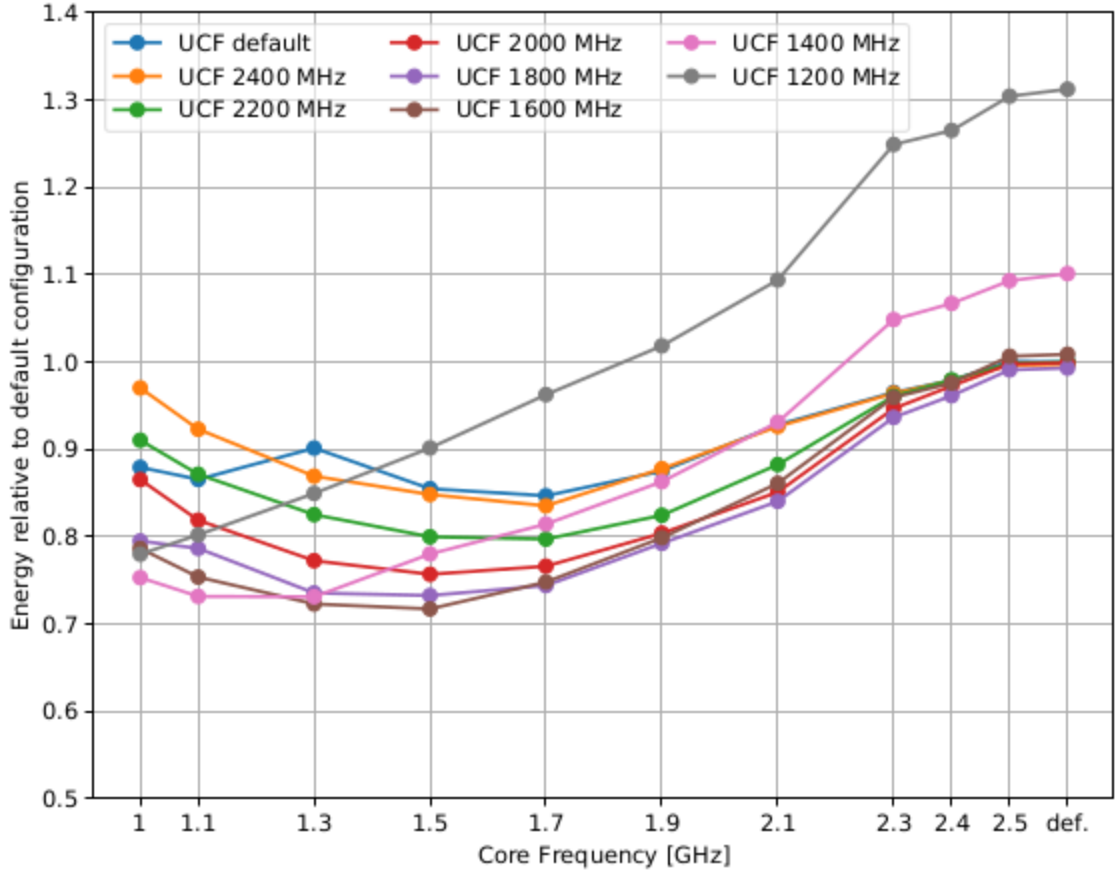
CASCADELAKE CPU BEHAVIOR



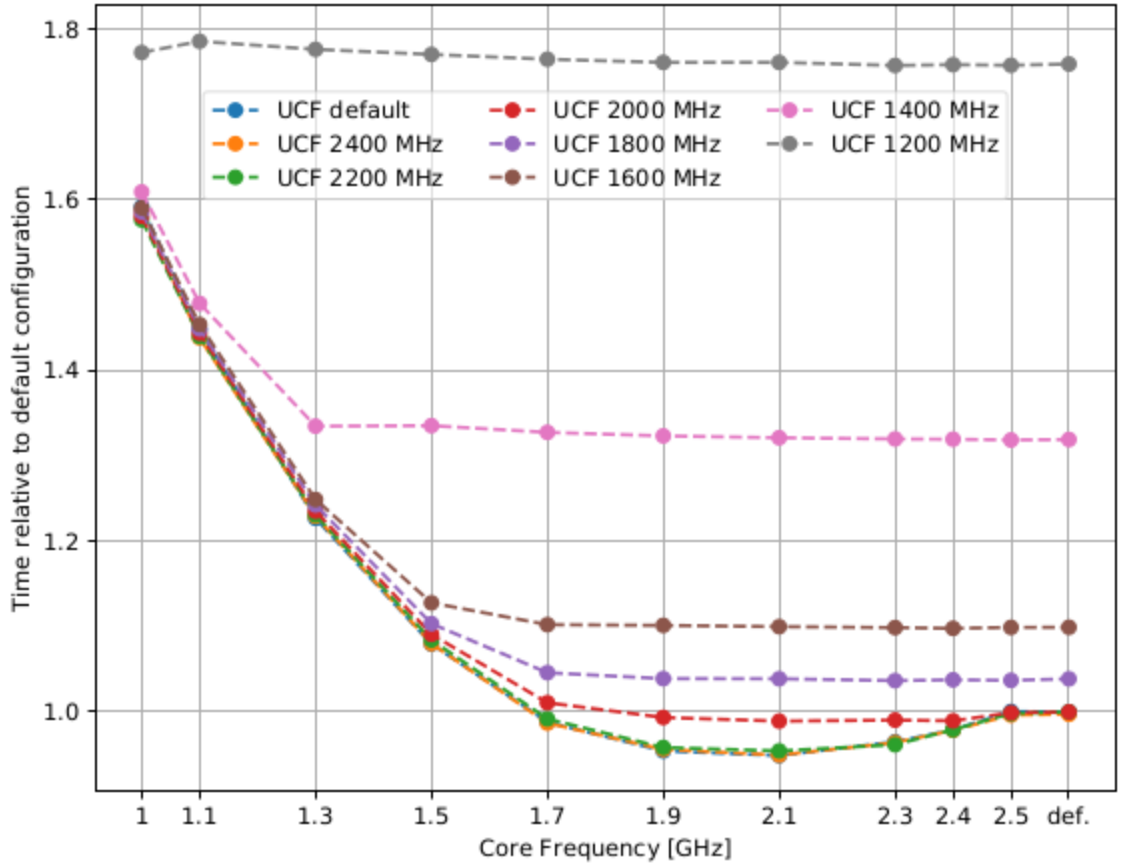
- AVX-512
- Arithmetic intensity 8
- RAPL keeps core frequency at its maximum while downscales uncore frequency

CPU FREQUENCIES SCALING

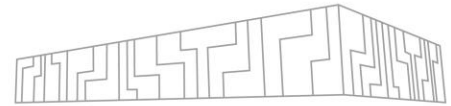
CPU energy



Time



- AVX-512, arithmetic intensity 8
- 1.9 GHz core frequency, 2.2 GHz uncore frequency
- => **18% CPU energy savings, 3.5% runtime improvement**
- => **15% node energy savings**



4] efficiency improvement

STATIC TUNING

Time

uncore [GHz] / core [GHz]	1.2	1.4	1.6	1.8	2	2.2	2.4
1.3	103.84	96.29	92.74	90.54	89.34	89.58	87.7
1.5	84.87	76.01	70.79	69.15	68.21	66.83	66.53
1.7	69.34	60.64	55.47	53.77	52.19	51.79	50.31
1.9	58.78	49.49	43.81	41.47	44.04	61.13	60.95
2.1	65.4	56.55	38.51	31.77	30.24	30.2	28.5
2.3	41.53	31.78	25.89	23.82	22.03	21.07	20.1
2.5	35.68	25.47	19.09	17.67	15.42	14.17	13.73
2.6	32.9	22.5	16.47	14.07	12.38	11.09	10.46
2.7	30.15	20.25	14.06	11.35	10.38	8.49	7.92
2.8	27.65	17.58	10.74	8.84	7	6.17	5.27
2.9	25.52	14.94	8.75	6.35	5.26	3.52	2.99
3	23.5	13.06	6.28	4.34	2.76	1.73	1.26
3.1	21.75	10.84	4.75	2.18	0.57	−0.55	−0.83
3.2	19.9	9.26	2.47	0.91	−1.2	−1.45	−1.6
3.3	18.49	7.83	1.02	−0.86	−1.64	−1.65	−1.56

Node energy (HDEEM)

uncore [GHz] / core [GHz]	1.2	1.4	1.6	1.8	2	2.2	2.4
1.3	9.27	7.86	8.48	9.52	12.78	17.55	21.76
1.5	3.25	0.77	0.08	1.32	4.45	7.73	12.26
1.7	−0.69	−3.28	−4.06	−3.09	−0.83	2.75	6.21
1.9	−3.29	−6.45	−7.76	−7.36	−2.46	12.82	17.32
2.1	−4.58	−8.1	−7.28	−9.93	−8	−5.2	−2.06
2.3	−5.86	−9.74	−11.49	−11.16	−9.55	−7.01	−4.11
2.5	−4.85	−9.3	−11.56	−10.84	−9.74	−7.62	−4.51
2.6	−4.69	−9.38	−11.42	−11.49	−10.09	−8.04	−5.16
2.7	−3.97	−8.39	−10.69	−11.09	−9.12	−7.62	−4.75
2.8	−3.14	−7.85	−10.74	−10.52	−9.34	−7.01	−4.43
2.9	−1.14	−6.46	−8.89	−9.17	−7.35	−5.87	−2.97
3	1.02	−4.35	−7.43	−7.3	−5.94	−3.81	−1.08
3.1	3.53	−2.43	−4.93	−5.44	−4.16	−2.43	−0.27
3.2	6.15	0.24	−3.04	−2.64	−2.24	−0.53	0.11
3.3	9.63	3.38	−0.46	−0.78	−0.3	0.16	0.46

STATIC TUNING

various performance degradation limits

various power monitoring systems

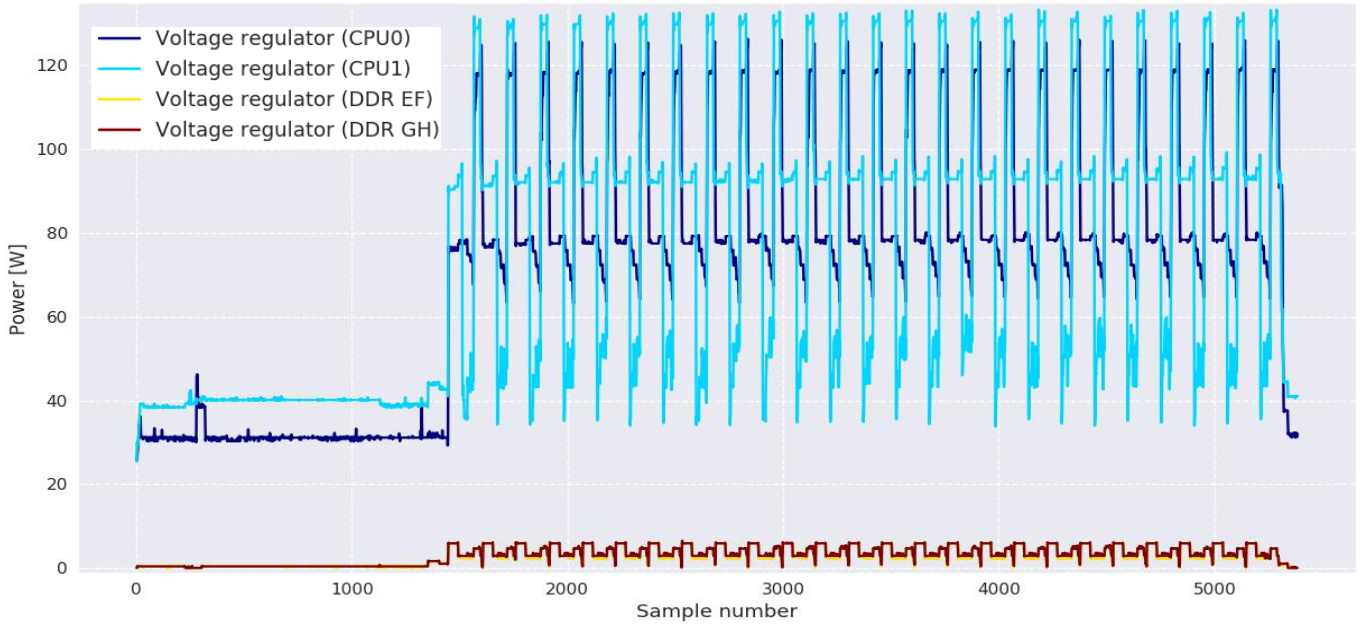
	default		-2 % limit	-5 % limit	-10 % limit	unlimited
Runtime [s]	257.07	**Performance penalty [%]**	2.18	4.34	8.84	25.9
Energy [kJ]		**Energy savings [%]**				
HDEEM	881.28	**HDEEM**	5.44	7.30	10.52	11.49
RAPL	668.07	**RAPL**	6.85	9.44	13.92	18.42
Eff. [MFLOPs/W]		**Eff. [MFLOPs/W]**				
HDEEM	10.40	**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

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

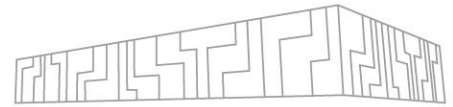
NUI Galway OÉ Gaillimh

intel

gns

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

DYNAMICITY 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
}
```

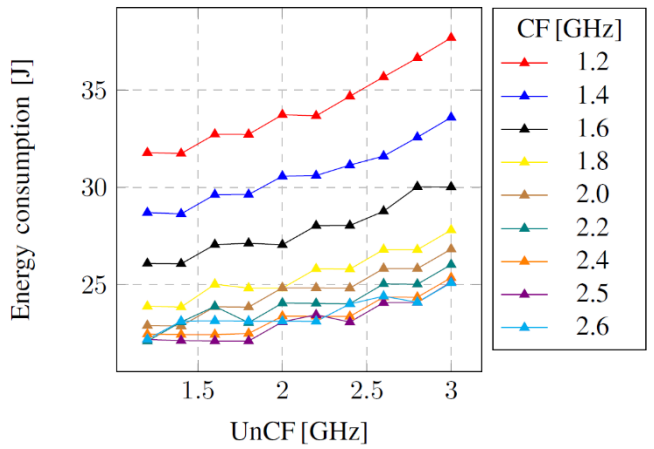
memory bound region
CPU core freq = 1.6 GHz

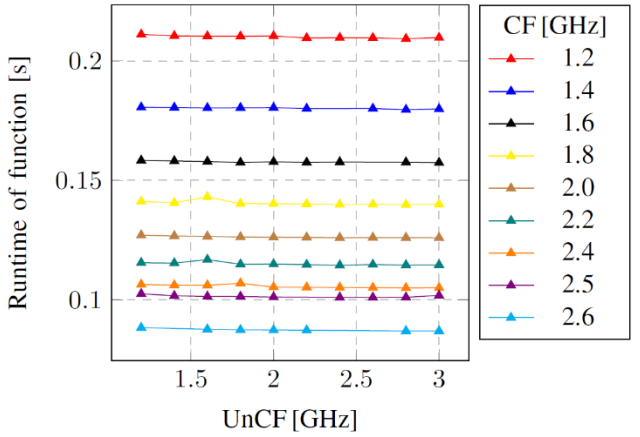
compute bound region
CPU core freq = 2.5 GHz

LATTICE BOLTZMANN BENCHMARK

Collide

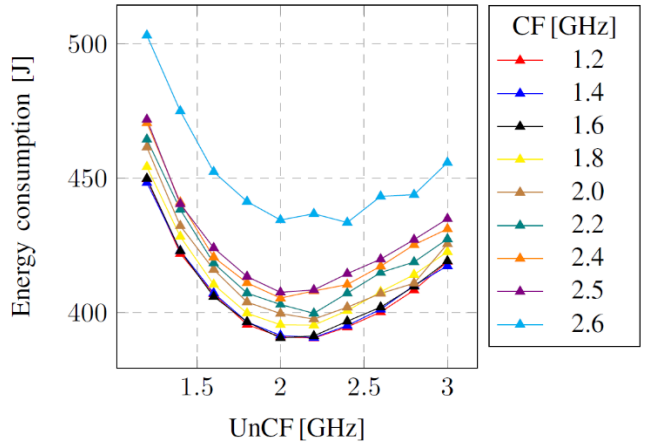
compute bound

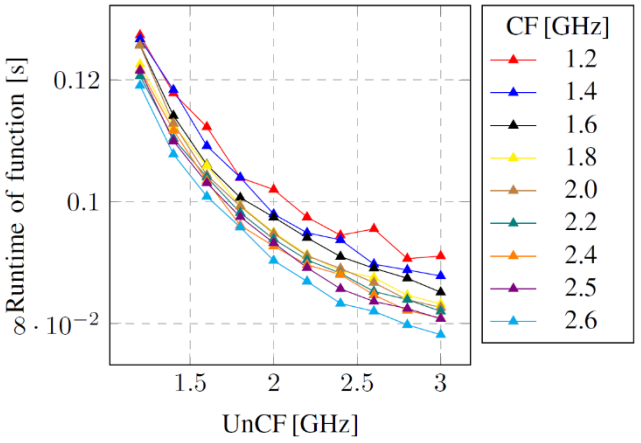




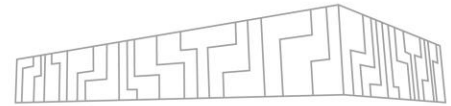
Propagate

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 %**



5] MERIC runtime system

MERIC RUNTIME SYSTEM

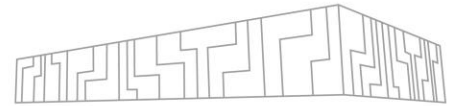
- MERIC runtime system provides dynamic application tuning
  - Lightweight & easy to install & easy to use
  - C/C++ API, Fortran module, Python module
  - MPI, OpenMP, CUDA parallelization
- Performance and power aware
- Support for a wide range of architectures
  - x86
  - IBM OpenPOWER
  - ARM
  - 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

SIMPLE ENERGY MEASUREMENT



mericStatic

```
$ mericStatic -e RAPL -- <application> [application parameters]
$ mericStatic –eval
```

OR

```
$ srun --overlap ./mericStatic -e RAPL,NVML -- ./application [application parameters] &
$ srun --overlap ./mericStatic -- stop
$ mericStatic –eval
```

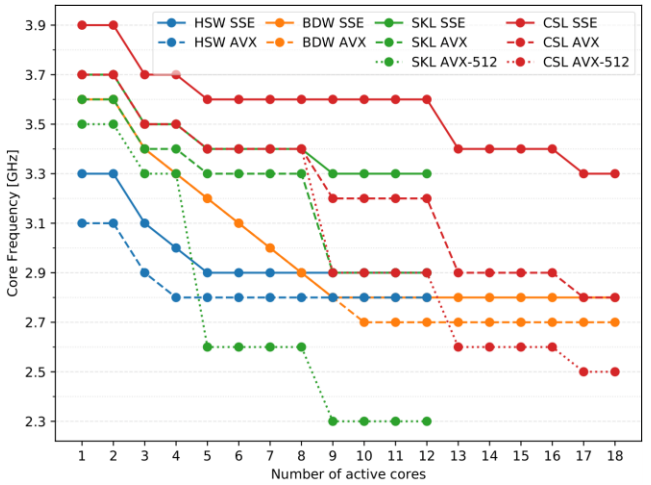
```
[nct00018@gs23r3b72 meric]$ mericStatic -e RAPL -- sleep 5
measurement_id : 36276003
[nct00018@gs23r3b72 meric]$ mericStatic -- eval
measurement_id : 36276003
Max Runtime [s] = 5.004
RAPL Energy consumption [J] = 1445.562


Total Energy consumption [J] = 1445.562
Total Energy consumption [Wh] = 0.402
```

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									

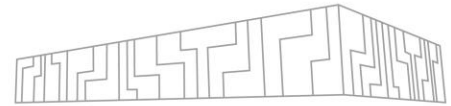
SYSTEM INFO 1/2

Detailed information about available hardware



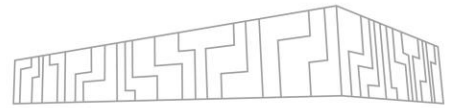

```
$ systemInfo
# SYSTEM INFORMATION
    CPU name:           Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz
    Sockets per node: 2
    Cores per socket: 18
    Threads per core: 1

# CPU FREQUENCIES
    Current scaling driver:      acpi-cpufreq
    Current scaling governor:    performance
    Available governors:         conservative ondemand userspace powersave performance
    Hardware controlled P-State: disabled
    Turbo CPU core frequencies:  3900000(2) 3700000(4) 3600000(8) 3600000(12) 3400000(16) 3300000(18)
kHz(#cores)
    Nominal CPU core frequency:  2600000 kHz
    Min CPU core frequency:      1000000 kHz
    Available CPU core frequencies:  2601000 2600000 2500000 2400000 2300000 2100000 2000000 1900000
1800000 1700000 1600000 1500000 1300000 1200000 1100000 1000000 kHz
    Max CPU uncore frequency:    2400000 kHz
    Min CPU uncore frequency:    1200000 kHz
...
```

SYSTEM INFO 2/2



```
...
# CPU POWER LIMITS
    RAPL time window unit: 976.562 us
    PKG max power limit:   376 W
    PKG min power limit:   69 W
    DRAM max power limit:  26.75 W
    DRAM min power limit:  4.5 W

# DEFAULT CPU POWER LIMITS
    PKG power limit #1: enabled + clamping disabled
    PKG power limit #1: 150 W
    PKG time window #1: 1 s
    PKG power limit #2: enabled + clamping enabled
    PKG power limit #2: 180 W
    PKG time window #2: 0.000976562 s
    DRAM power limit:   disabled + clamping disabled
    DRAM power limit:   0 W
    DRAM time window:   0.000976562 s

# AVAILABLE POWER MONITORING SYSTEMS
    RAPL
    HDEEM
```

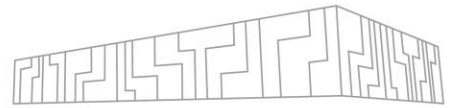
ANALYSIS WITHOUT HW TUNING



- Performance (FLOPs) evaluation
- Energy consumption
  - node energy consumption
  - CPU+GPU energy performance counters
- Energy efficiency
- Carbon footprint
- Vectorization ratio
- CPU (core/uncore) frequency
- Memory bandwidth
- Computation intensity
- Power consumption timeline

- Various hardware platforms, various power monitoring systems

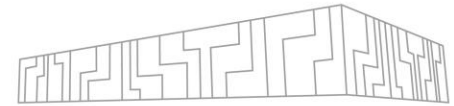
ADDITIONAL ANALYSIS STEPS



- Application instrumentation
- Default HW configuration execution
- Multiple executions of the application in a variety of HW configurations
- Evaluation of the trace files

- Various power limits

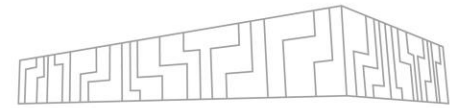
LIBRARY API



Instrumentation

```
int main(int argc, char **argv)
{
MERIC_Init();
MERIC_MeasureStart("A");
do_work_A(2);
MERIC_MeasureStop("A");

MERIC_MeasureStart("B");
do_work_B(5);
MERIC_MeasureStop("B");

MERIC_Close();
return 0;
}
```

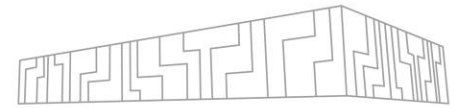
APPLICATION ANALYSIS



- Manual exhaustive state-space-search (old-fashioned way)
- MERICwrapper will identify the optimal configuration from the data

```
export MERIC_OUTPUT_DIR="MERICdir"
export MERIC_MEASUREMENT=RAPL
for CF in 30 25 20
do
   for UCF in 24 22 20
   do
      export MERIC_FREQUENCY=$CF"00MHz"
      export MERIC_UNCORE_FREQUENCY=$UCF"00MHz"
      LD_PRELOAD=libmpi_sbi.so mpirun ./a_sbi.out --app_param
   done
done
$MERIC_PATH/bin/wrapper -a $MERIC_OUTPUT_DIR -m RAPL
```

APPLICATION TUNING



- Automatized state-space-search using MERICwrapper

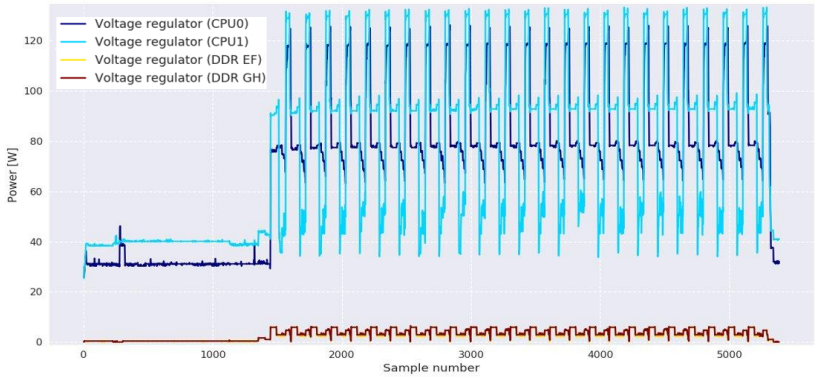
```
$ $MERIC_PATH/bin/wrapper -s config.json -- ./a_meric.out
```

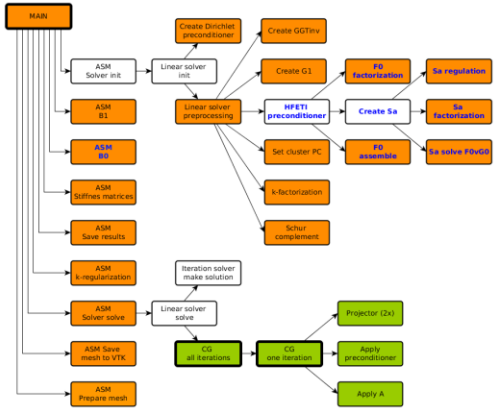
```
$ $MERIC_PATH/bin/wrapper -s config.json -- \
LD_PRELOAD=libmpi_meric.so mpirun ./a_meric.out --app_param
```

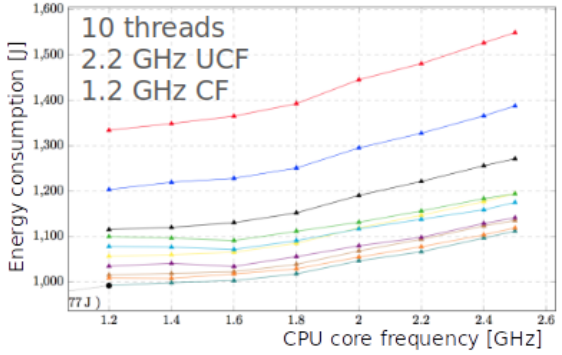
- Configuration file specifies what
  - HW configurations to test
  - State-space-search algorithm
  - Used energy measurement mode
  - MERIC output configuration

RADAR VISUALIZER

- *Visualisation of applicatin behavior in various configuration*
- Tables
  - Overall application evaluation
  - Summary of nested regions' behavior
  - Each region behavior description
- Heatmaps
- Plots
- Samples timeline
- Trace timeline
- 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

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