CLOUD ENERGY CONSUMPTION

How to estimate energy in restricted environments

GREEN CODING;

Green Coding Berlin

We do open source tools for energy measurement

- Work with many NGOs and open source communities
- We do educational talks at organizations

axel springer_ | WBS CODING SCHOOL | SDIA | Bits & Bäume | IRON HACK

Agenda

1. Techniques of energy measurement
2. Tools
3. Cloud restrictions
4. SPECPower Database / Teads data
5. Modelling approaches
6. Our open source model

First you need to measure the energy

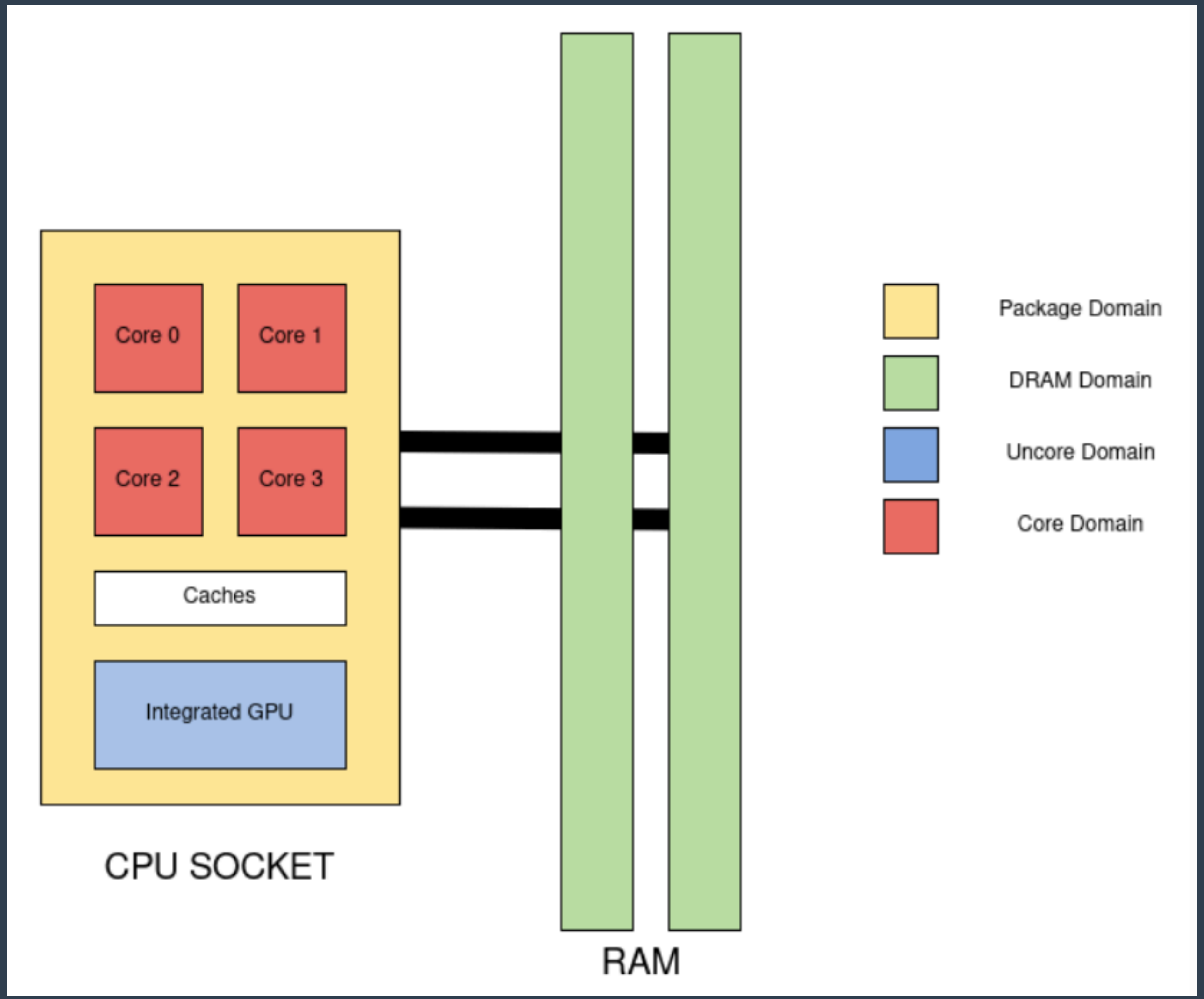
General ways how to measure power. no accepted approach

- Electrical
  - AC energy meter in server rack / wall socket (Buffer effects in PSU)
  - DC inline current measurement of PSU (hard to integrate, power lines not exclusive, current mirror or sampling)
- Software
  - CPU % and / or memory usage (robust transfer model?, idle power?)
- Sensors / On-Chip
  - RAPL (opaque validity / security concerns)
- Instruction based
  - Decompile and count Instructions (reference values, decompile?)
  - perf (statistical approach, what type of insctruction occured?, strongly architecture dependent)

Details on RAPL

The most used technology atm

- Energy measurement capabilities on most modern Intel/AMD processors
- Measure:
  - CPU Energy per Core / Package
  - RAM
  - Integrated GPU
- Software model of capacitor readings on mainboard
  - Resolution 1ms / 15.3 microJoules
- Exposed in Linux kernel through device



Source: https://pyjoules.readthedocs.io/en/stable/devices/intel_cpu.html

GREEN CODING;

Agenda

1. Techniques of energy measurement
2. Tools
3. Cloud restrictions
4. SPECPower Database / Teads data
5. Modelling approaches
6. Our open source model

</> GREEN CODING;

Ready to use tools

codecarbon.io

- Python
- RAPL-based
- NVIDIA GPU support

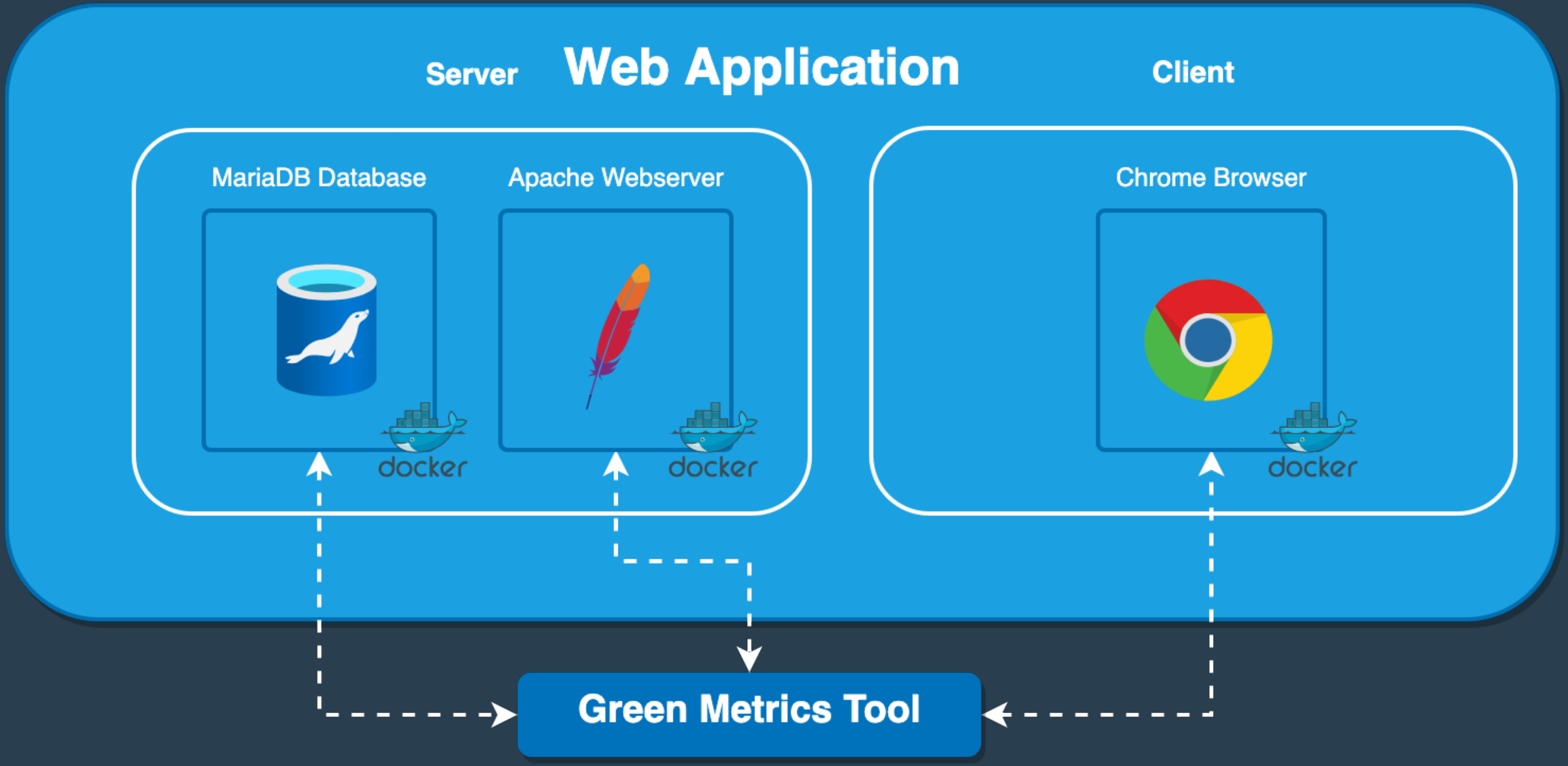
```
import tensorflow as tf

from codecarbon import Emission
                          EmissionsTracker            codecarbon
mnist = tf.keras.da       OfflineEmissionsTracker     codecarbon
                          Press ⏎ to insert, ⇥ to replace  Next Tip
(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0


model = tf.keras.models.Sequential(
    [
        tf.keras.layers.Flatten(input_shape=(28, 28)),
```

Ready to use tools

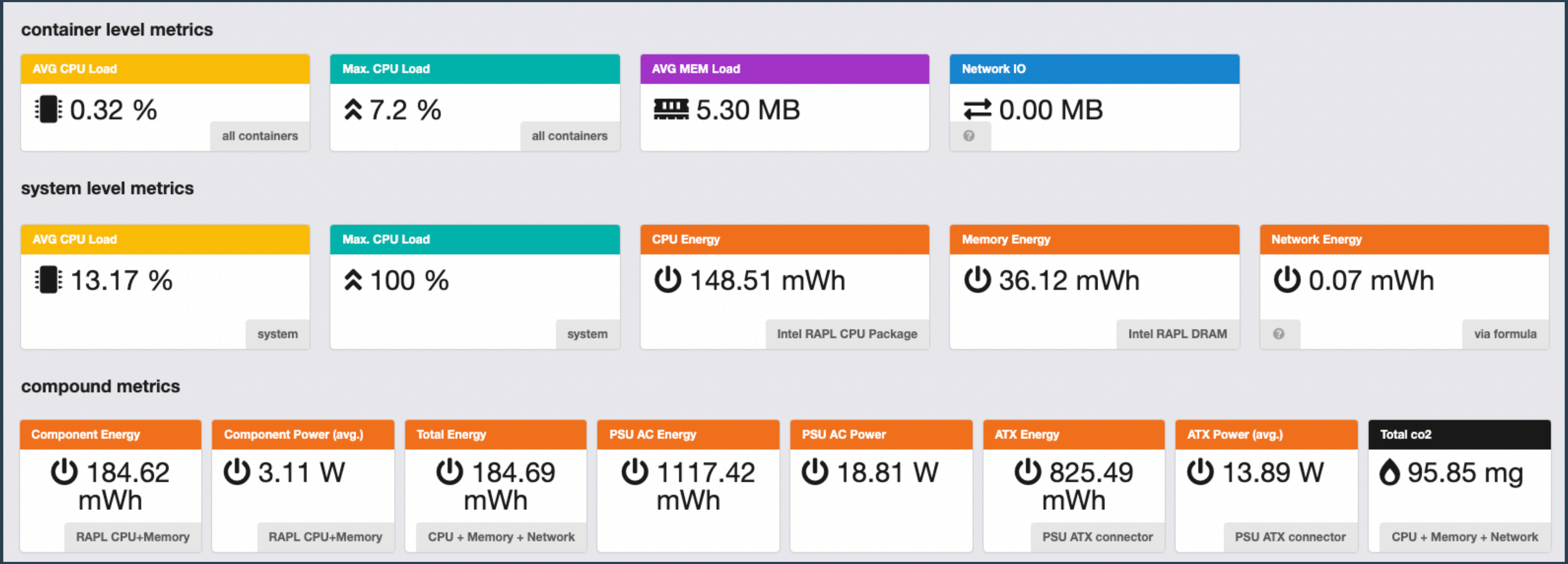
Green Metrics Tool - container focused tool



GREEN CODING;

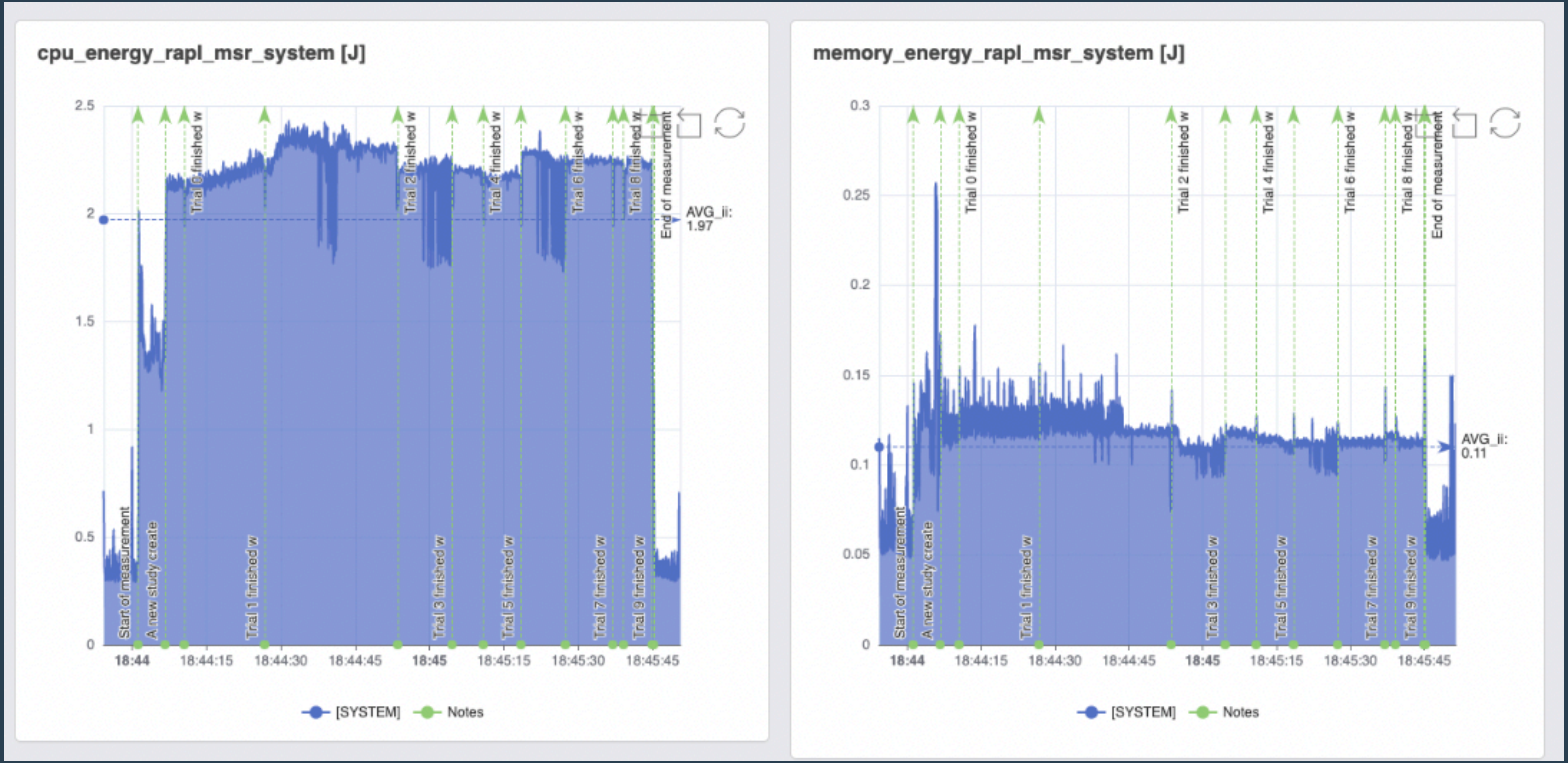
What the ouput looks like

Cumulative metrics on system / container level



What the output looks like

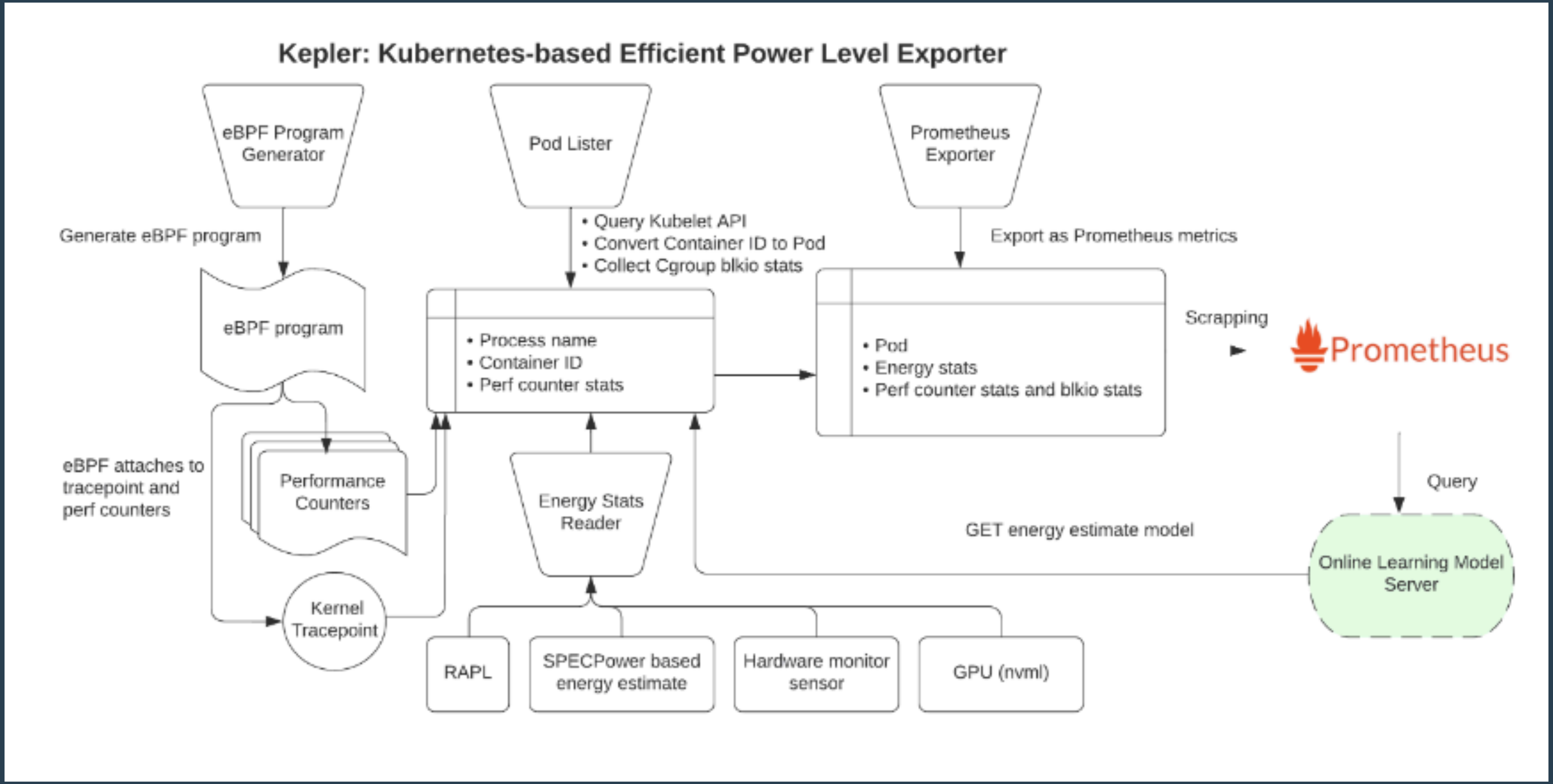
Charts ... obviously ;)



GREEN CODING;

Looking into distributed environments

Introducing Kepler: http://sustainable-computing.io



Very interesting on bare metal machines. Unclear how it works in cloud environments

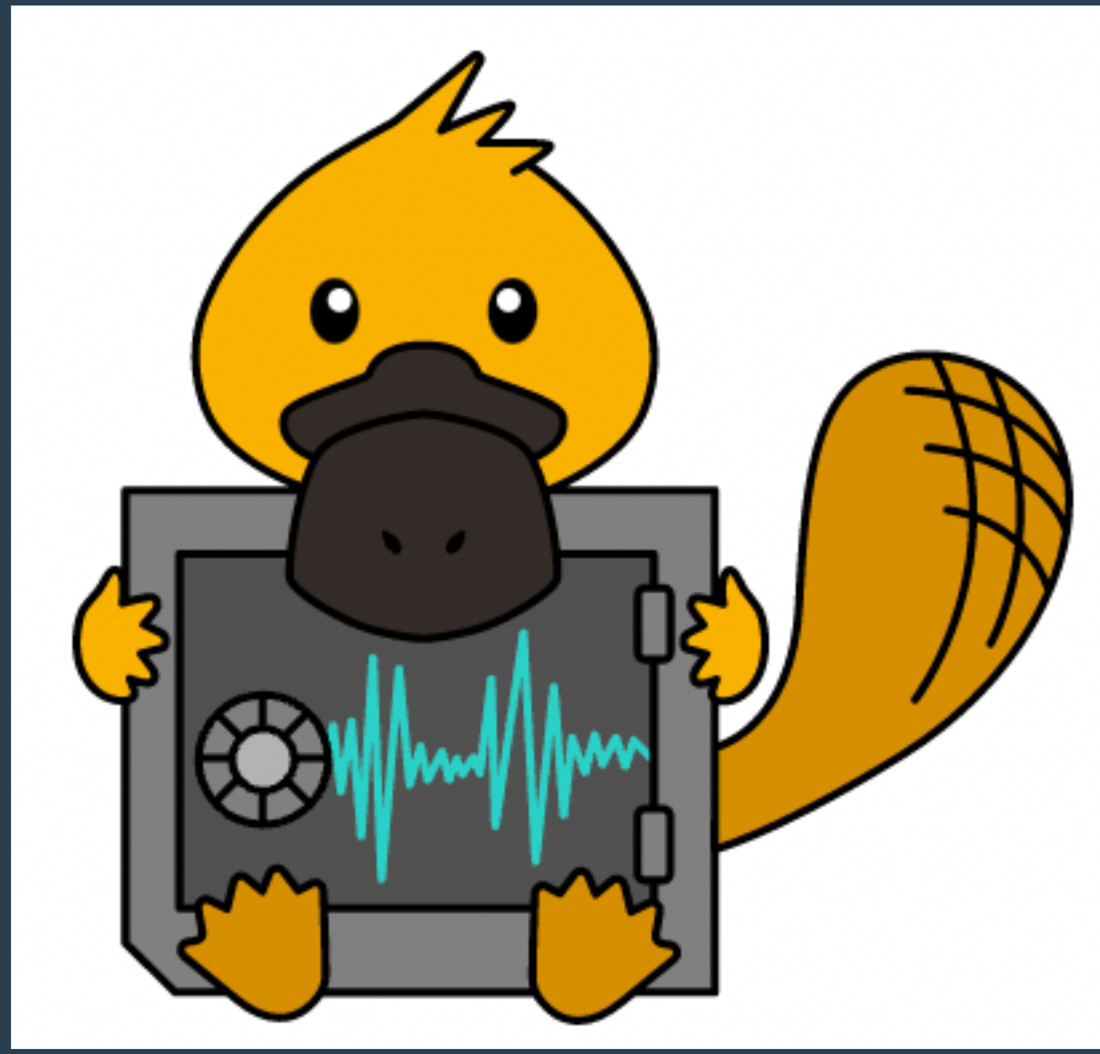
GREEN CODING;

Agenda

1. Techniques of energy measurement
2. Tools
3. Cloud restrictions
4. SPECPower Database / Teads data
5. Modelling approaches
6. Our open source model

Cloud restrictions

RAPL was nerfed and deactivated



- Platypus is a timing attack using RAPL to extract secret keys vom SGX
- Lead to most hypervisors not forwarding PMUs anymore
- Other restrictions include obfuscating RAPL when SGX is enabled

PLATYPUS

WITH GREAT POWER COMES GREAT LEAKAGE

GREEN CODING;

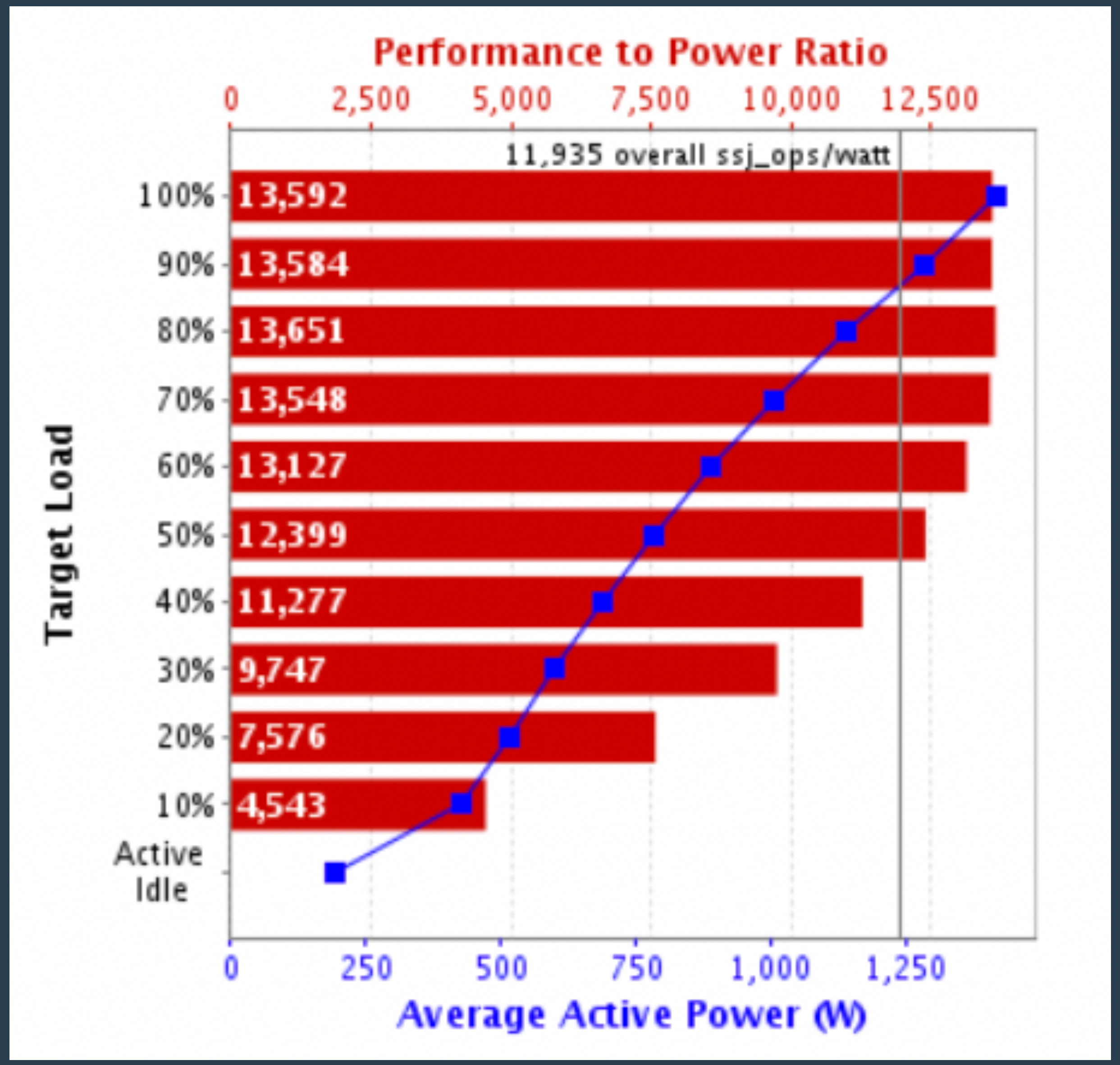
Agenda

1. Techniques of energy measurement
2. Tools
3. Cloud restrictions
4. SPECPower Database / Teads data
5. Modelling approaches
6. Our open source model

SPECPower 2008

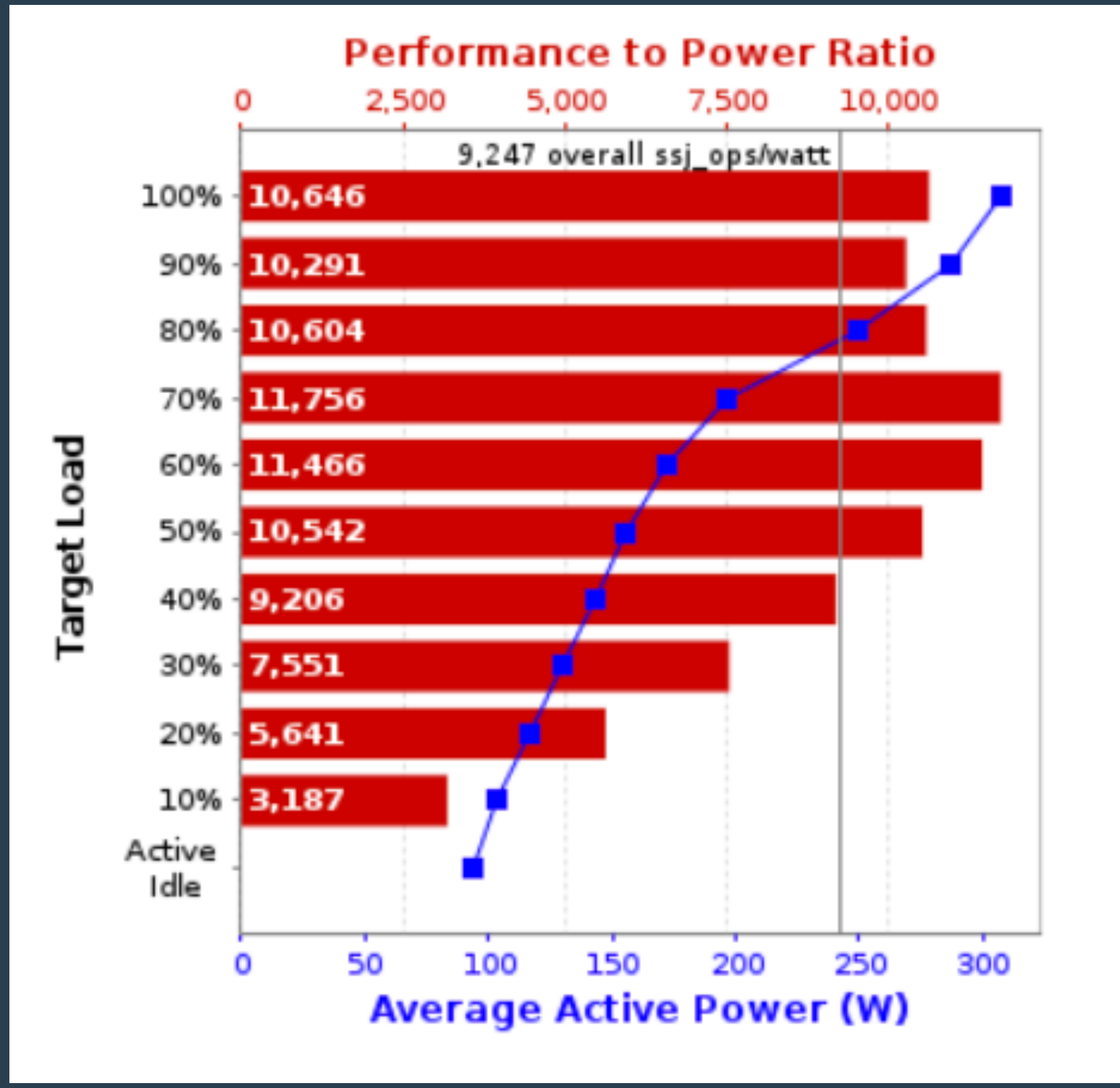
Proprietary Java based energy benchmark - 836 machines

High Idle, but almost linear



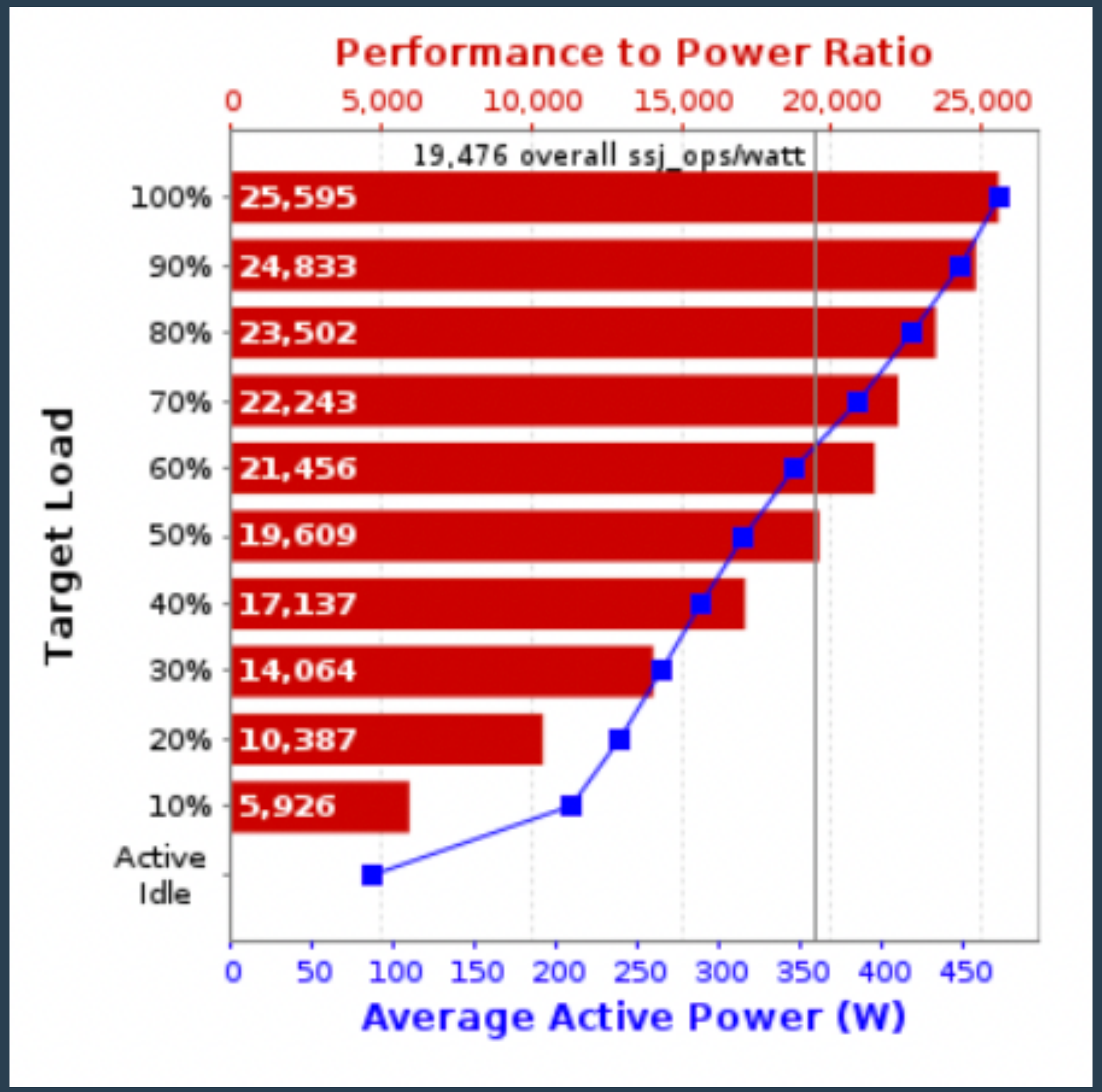
ASUSTeK Computer Inc. RS720Q-E9-RS8 (2019)

50% Power increase at 70% utilization



Hewlett Packard Enterprise ProLiant DL110 Gen10 Plus

Idle optimized



QuantaGrid D43K-1U (2022)

GREEN CODING;

SPECPower 2008

No uniform setup of machines - Just configuration has to be documented

System Under Test Notes

- kernal parameter:pcie_aspm=force pcie_aspm.policy=powersupersave
- Benchmark started via ssh
- cpupower frequency-set -g ondemand
- echo -n 98 > /sys/devices/system/cpu/cpufreq/ondemand/up_threshold
- echo -n 1000000 > /sys/devices/system/cpu/cpufreq/ondemand/sampling_rate
- echo 0 > /proc/sys/kernel/nmi_watchdog
- sysctl -w kernel.sched_migration_cost_ns=6000
- sysctl -w kernel.sched_min_granularity_ns=10000000
- sysctl -w vm.swappiness=50
- sysctl -w vm.laptop_mode=5
- N/A:The test sponsor attests, as of date of publication, the CVE-2017-5754(Meltdown) is mitigated in the system as tested and documented.
- Yes:The test sponsor attests, as of date of publication, the CVE-2017-5753(Spectre variant 1) is mitigated in the system as tested and documented.
- Yes:The test sponsor attests, as of date of publication, the CVE-2017-5715(Spectre variant 2) is mitigated in the system as tested and documented.

Boot Firmware Settings

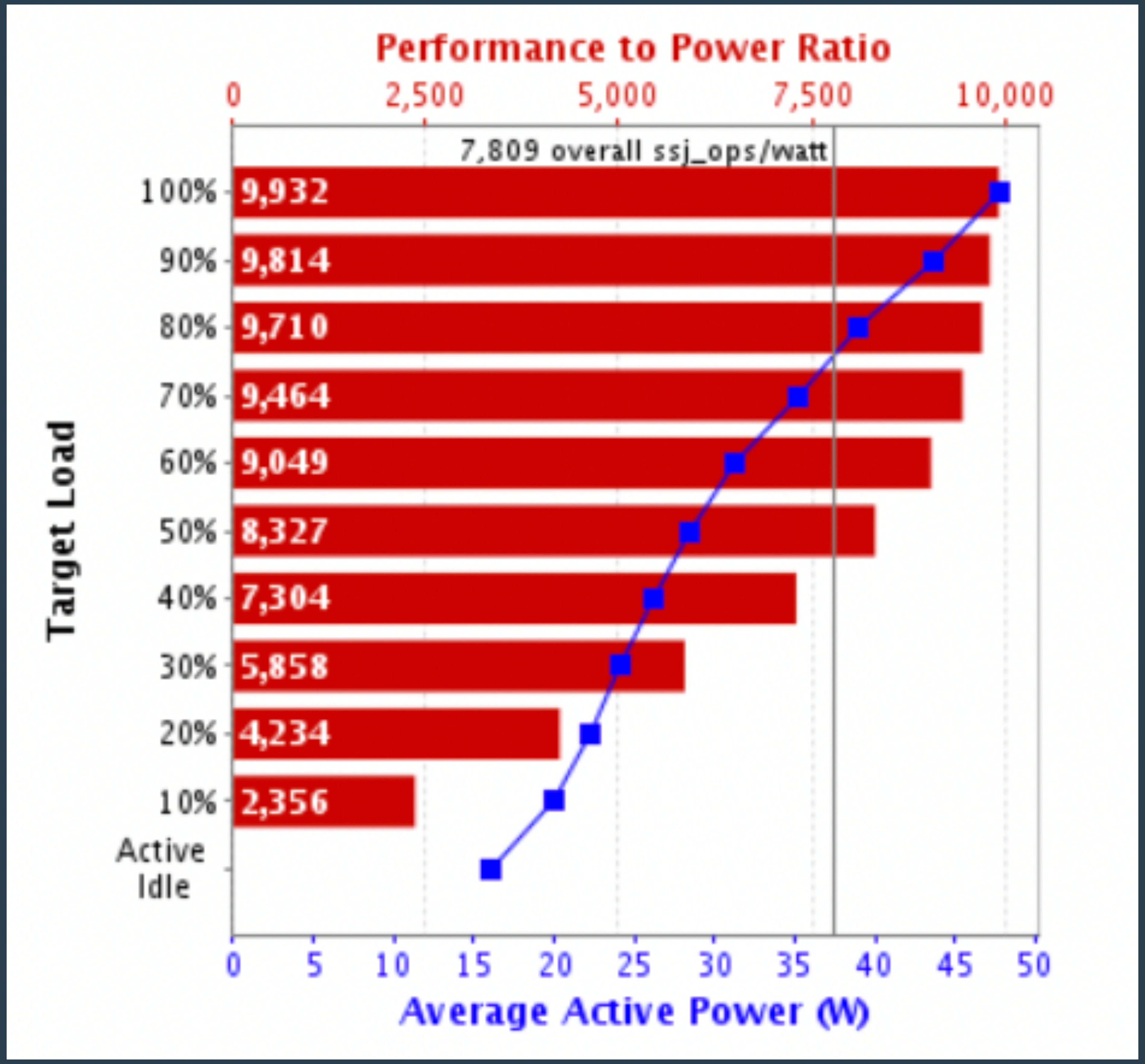
- SVM Mode = Disabled
- SMEE = Disabled
- Core Performance Boost = Disabled
- SR-IOV Support = Disabled
- L1 Stream HW Prefetcher = Disable
- L2 Stream HW Prefetcher = Disable
- Power Balancer = Auto
- DRAM scrub time = Disabled
- NUMA nodes per socket = NPS4
- DRAM Power Down Enable = Enabled
- APBDIS = 1
- Fixed SOC Pstate = P3
- Memory clock speed = 1333 MHz
- EfficiencyModeEn = Enabled
- ACPI SRAT L3 Cache As NUMA Domain = Enabled
- Memory Interleaving = Disabled
- xGMI max speed = 9.6Gbps
- EDC Control = Manual
- EDC = 240
- EDC Platform Limit = 240
- XHCI Controller1 enable = Disabled
- SATA Enable = Disabled

- This means variables rise from 100 initially to 200+ with also many categorical variables

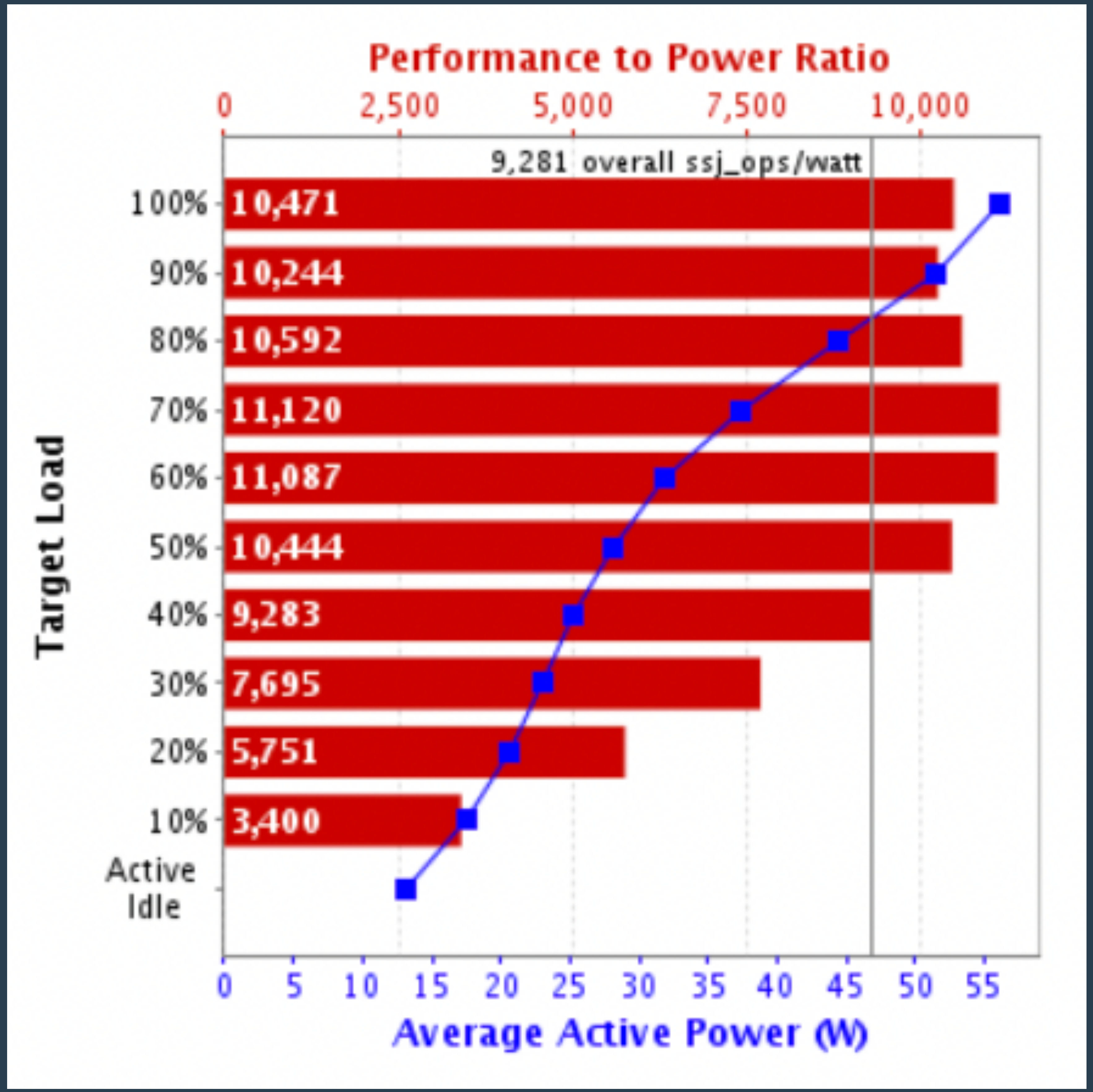
GREEN CODING;

SPECPower Settings example

TurboBoost on/off - Some settings have strong effect, some don't



Turbo Boost off



Turbo Boost on - Performance-to-Power drops

GREEN CODING;

Teads dataset

AWS only - RAPL measurements extrapolated and enriched

1	Product						ige Consumption (in Watts)				
2	Vendor	Product Name	Region	Test Date	Total Number of vCPU	Memory Quantity (in GB)	PkgWatt CPUStress 10%	PkgWatt CPUStress 20%	PkgWatt CPUStress 30%	PkgWatt CPUStress 40%	PkgWatt CPUStress 50%
3	AWS	c5n.metal	Oregon	avr.-21	72	192	135	174	212	249	293
4	AWS	c5.metal (underclocked)	North Virginia	févr.-21	96	192	146	194	225	244	263
5	AWS	c5.metal	Paris	juin-21	96	192	176	241	299	375	448
6	AWS	c5.metal*	Oregon	juin-21	96	192	174	243	299	372	441
7	AWS	r5.metal (underclocked)	North Virginia	févr.-21	96	768	148	188	205	222	258
8	AWS	r5.metal	North Virginia	juin-21	96	768	138	178	224	272	307
9	AWS	m5.metal (underclocked)	North Virginia	févr.-21	96	384	127	150	188	214	224
10	AWS	m5.metal	Oregon	juin-21	96	384	147	193	246	298	344
11	AWS	z1d.metal	North Virginia	mars-21	48	384	148	198	245	305	361
12	AWS	m5zn.metal	North Virginia	mars-21	48	192	147	212	270	331	381
13	AWS	i3.metal	Oregon	avr.-21	72	512	100	126	152	178	205
14	Scaleway	GP-BM1-S			8	32	10	16	17	15	15

https://medium.com/teads-engineering/building-an-aws-ec2-carbon-emissions-dataset-3f0fd76c98ac

Agenda

1. Techniques of energy measurement
2. Tools
3. Cloud restrictions
4. SPECPower Database / Teads data
5. Modelling approaches
6. Our open source model

SDIA modelling approach

Linear model with CPU focus

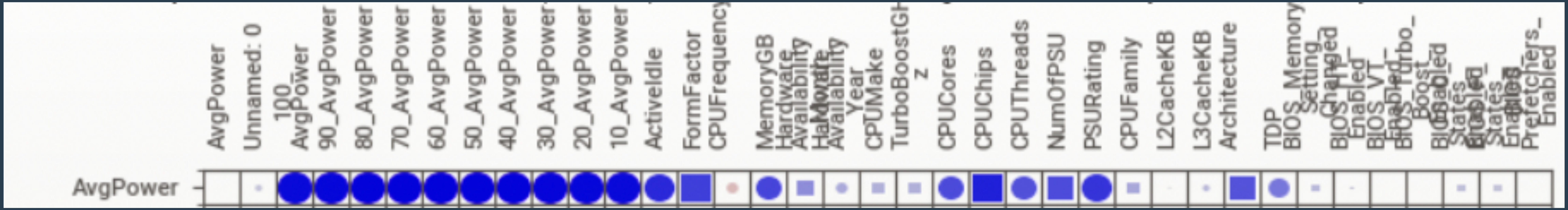
- P_machine = (Chips * TDP) / alpha
- => alpha typically assumed to be 0.65 in align with many averages from literature
- Model cannot handle GPU / Memory intensive machines
- Model cannot predict idle power (linear extrapolation to 0)

- Model puts most focus on embodied carbon and data center coefficients like PUE, mineral use etc. !

Linear modeling (OLS)

Curves are almost linear - most introspective

- First look at Associations
- Interesting: TDP, MemoryGB, CPUCores / CPUChips
- Interesting but not known: FormFactor, Architecture, PSURating / NumOfPSU

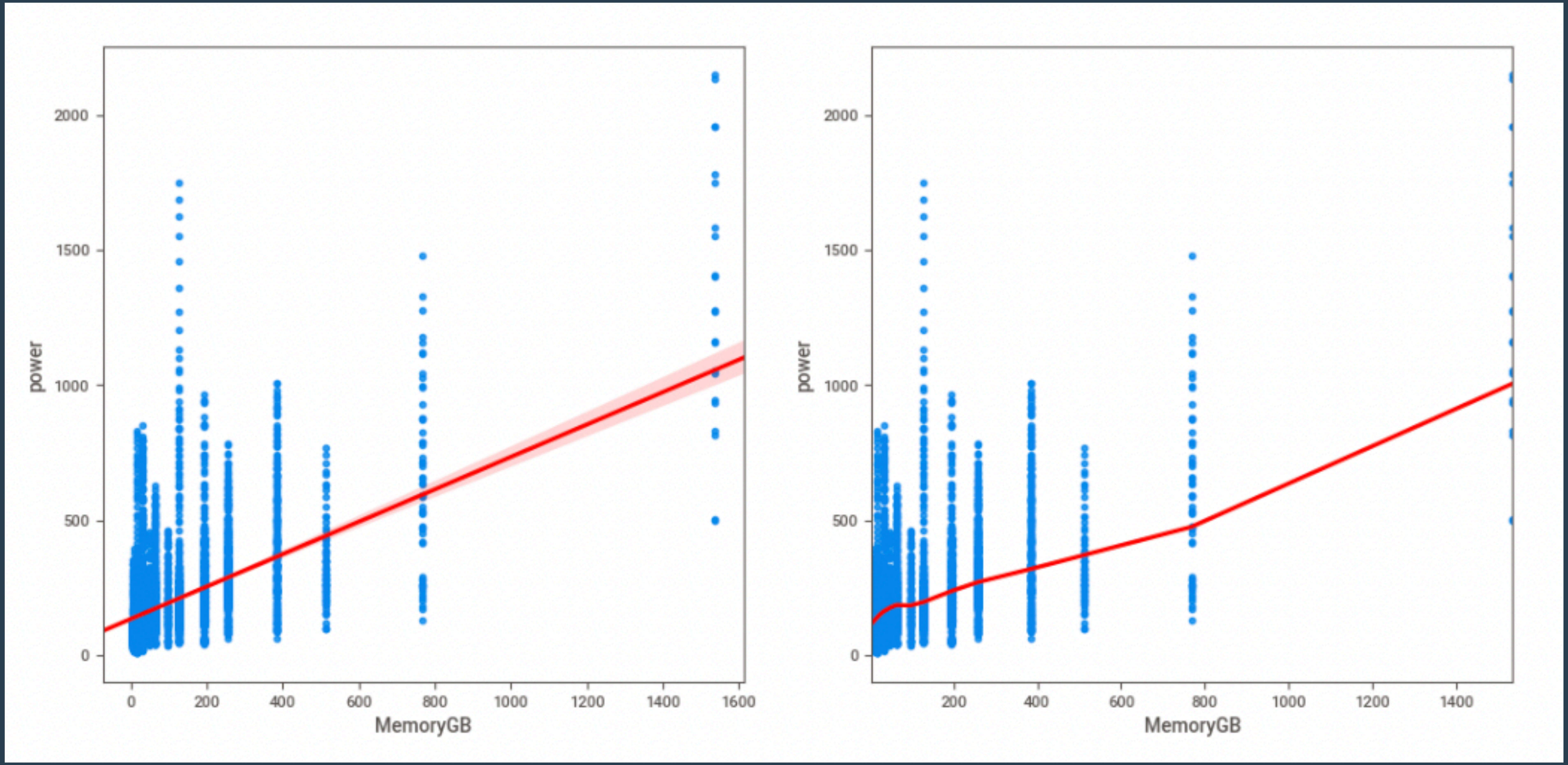


https://www.kaggle.com/code/arne3000/spec-power-eda-pass-2

GREEN CODING;

Linear modeling (OLS)

Real variable distributions shows to be highly scattered



GREEN CODING;

Similar work from academia

Interact DC paper in collab with university of East London

Interact: IT infrastructure energy and cost analyzer tool for data centers

Nour Rteil[a,c,*], Rabih Bashroush[a,b], Rich Kenny[c], Astrid Wynne[c]

[a] *University of East London, London, E16 2RD, United Kingdom*
[b] *Uptime Institute, London, E1 7LS, United Kingdom*
[c] *Techbuyer, Harrogate, HG2 8PB, United Kingdom*

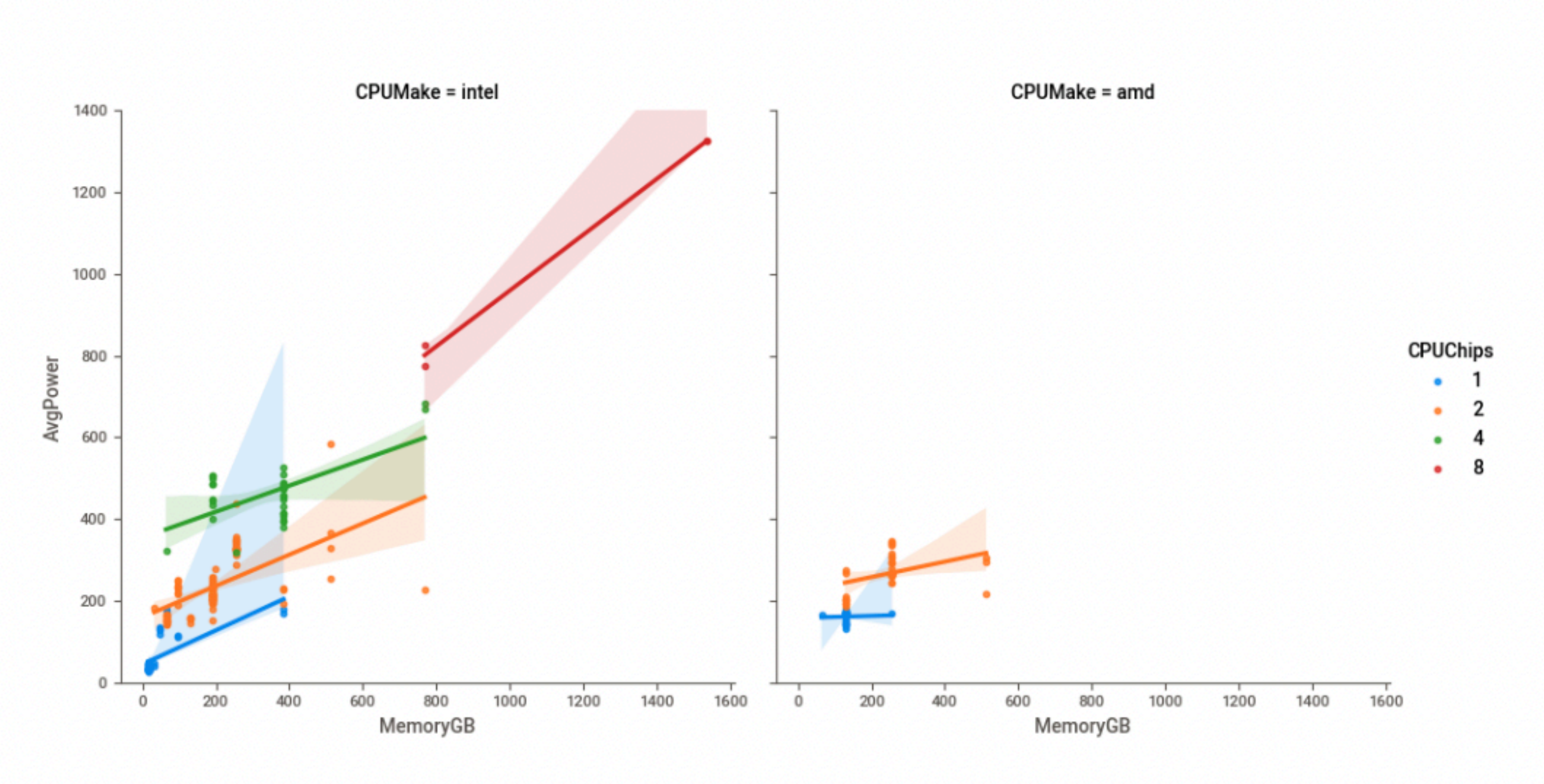
Table 3
Summary of selected features for the model.

User Input	Extracted features
Server model	Hardware vendor Release year Form factor
CPU model	Cores per chip Threads per core Frequency
Number of chips populated	Number of chips
Memory capacity	Memory capacity

GREEN CODING;

Our model and its parameters

Incorporated some splittings from InteractDC and our own



Splitting gets the scattering quite a bit down

GREEN CODING;

Our model and its parameters

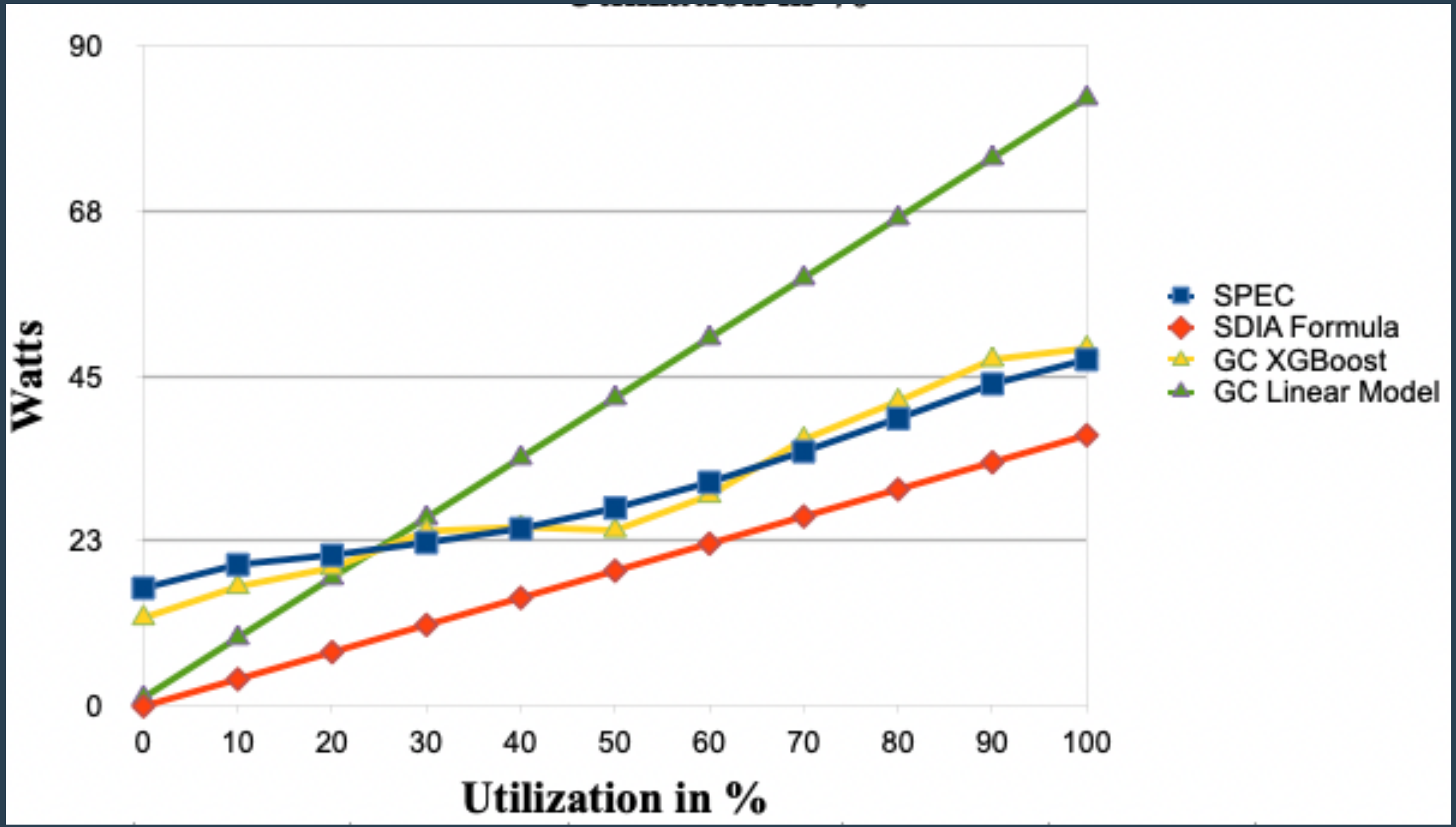
Incorporated some splittings from InteractDC and our own

```
parser.add_argument("--cpu-chips", type=float, help="Number of CPU chips", default=1)
parser.add_argument("--cpu-freq", type=float, help="CPU frequency")
parser.add_argument("--cpu-cores", type=float, help="Number of CPU cores")
parser.add_argument("--tdp", type=float, help="TDP of the CPU")
parser.add_argument("--ram", type=float, help="Amount of DRAM for the bare metal system")
parser.add_argument("--vhost-ratio", type=float, help="Virtualization ratio of the system.
```

Our final set of variables

Our model against baseline and against SDIA

In-sample performance of SPECPower

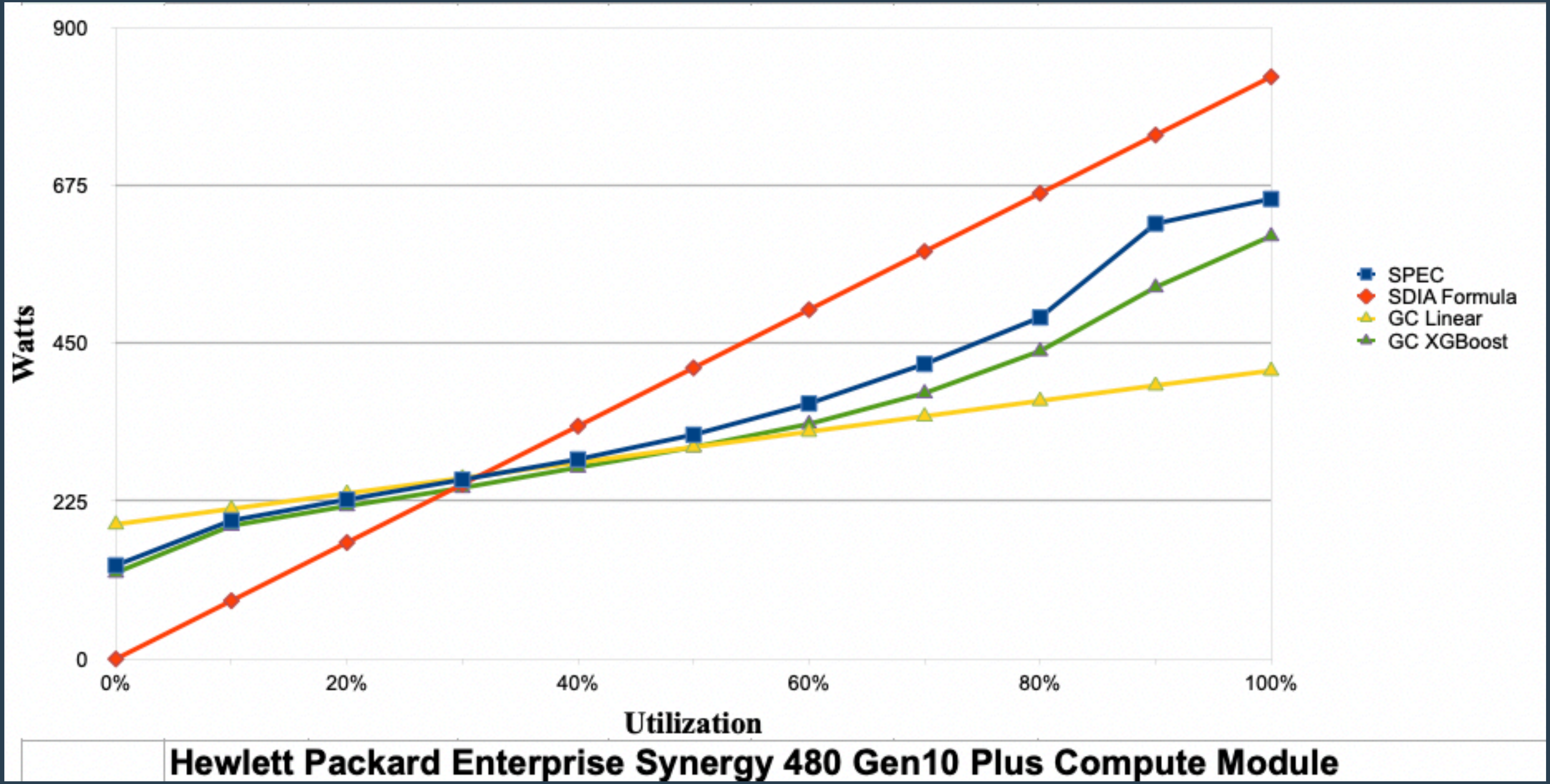


In sample performance

GREEN CODING;

Our model against baseline and against SDIA

Out of sample performance of SPECPower



Out of sample performance

GREEN CODING;

Our model - In-sample performance

Compared to Interact DC

- In-sample neg_mean_absolute_error in Repeated K-Folds (5,10)
  - K-fold CV score range: -16.00 < -14.06 < -12.92
- When using their variable set:
  - K-fold CV score range: -14.99 < -12.96 < -10.81

Table 4
Error evaluation for each model.

Label	Model	Average Mean absolute error (MAE)	Standard deviation for MAE scores
Power at idle (W)	RF	−14.13	4.35
	GB	−15.73	4.02
	KNN	−26.89	5.33
	ANN	−18.43	4.28
Power at full load (W)	RF	−26.45	6.17
	GB	−30.77	5.65
	KNN	−78.65	11.19
	ANN	−49.05	9.52

Full comparison against InteractDC tricky, as they only show 0% load and 100% load error values

Our model: Out of sample performance

Compared to Interact DC

- Better at 0% / 100%
- Worse at 25%

Model	Memory details	Actual idle power	Predicted idle power
DL380 G9	x2 16GB	76.2	66.9

Model	Memory details	Actual full power	Predicted full power
DL380 G9	x2 16GB	333.8	261.6

Summary of actual vs calculated power at 25 % load.

Model	Memory details	Actual avg. power	Calculated avg. power
DL380 G9	x2 16GB	118.1	115.6

```
arne@mintbook:~/Sites/green-coding/spec-power-model (main*) $ \
py3 xgboost_model.py --cpu-chips 2 --tdp 135 --ram 32 --cpu-freq 2900 --cpu-cores 24
Model will be restricted to the following amount of chips: 2.0
Model will be trained on: Index(['HW_CPUFreq', 'CPUCores', 'TDP', 'HW_MemAmountGB', 'utilization'], dtype='object')
Sending following dataframe to model:
    HW_CPUFreq  CPUCores    TDP  HW_MemAmountGB  utilization
0       2900.0      24.0  135.0            32.0            0
vHost ratio is set to  1.0
0
76.21349334716797
25
149.61862182617188
100
367.47601318359375
```

GREEN CODING;

Teads modelling approach

Through measurement and extrapolation

- AWS EC2 bare metal machines have been benched with `turbostress`
  - effectively stress-ng with 10% load steps and RAPL CPU/DRAM readings

- Non measured machines are derived by TDP per Watt (linear interpolation)
- Total machine is CPU+DRAM+GPU+(Delta for full machine)

- Details: https://medium.com/teads-engineering/building-an-aws-ec2-carbon-emissions-dataset-3f0fd76c98ac

Our model compared to the Teads data

Comparing against their datasheet

Instance type	Delta Full Machine	Instance @ Idle	Instance @ 10%	Instance @ 50%	Instance @ 100%
m5n.metal	84,0	198,4	320,8	567,8	784,3
not actually measured :		140.99	199.78	299.98	570.29

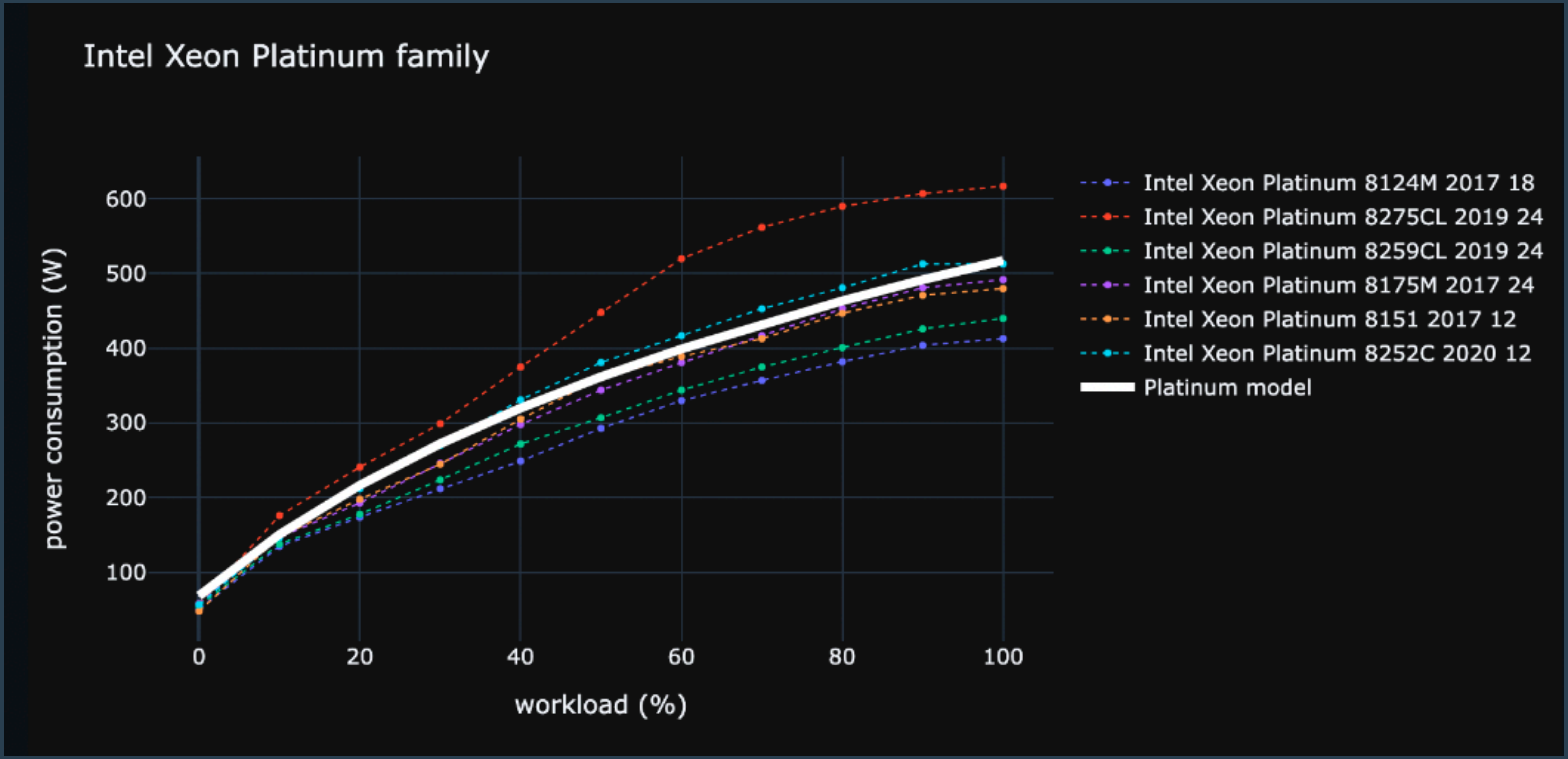
1	Instance type	Delta Full Machine	Instance @ Idle	Instance @ 10%	Instance @ 50%	Instance @ 100%
62	c5n.metal	96,0	184,0	298,3	507,7	689,1
	RAPL extrapolated:		147	192	258	525

- We are underestimating! ... but there is also not falsification!

GREEN CODING;

Our model compared to the Teads data

Teads data is is also logarithmic in shape - Different from what we see in SPECPower



https://github.com/Boavizta/Energizta/tree/dev/api

GREEN CODING;

Are AI models maybe interesting?

Some relevant studies

- Interact Paper already tried a "vanilla" RNN and KNN network
- Accuracy was lower than GB models
- Assumption was that there are just not enough datapoints (~800) to work with
- Maybe you have a different approach / better AI model?
- => Shout out to Jens from the AI idea workshop! (KI-Ideenwerkstatt)
  - BMUV sponsored project to implement AI solutions for climate and environmental protection

Community questions

Related to energy settings in linux

- What other settings do you know?
- Do you know what Cloud Vendors / Hosters typically tune / configure?
- Do you know the defaults of the cloud or where to get them?

Thank you!

Time for Q&A

- Check out our website and blog & newsletter: https://www.green-coding.org

- https://www.linkedin.com/in/arne-tarara / arne@green-coding.org

Want deeper dives and more details?

Follow Green-Coding.org

- Check out our website and blog & newsletter: https://www.green-coding.org
- Demo Open Data Repository: https://metrics.green-coding.org

- Meetup group: https://www.meetup.com/green-coding
- If you wanna present your green software case, please hit us up!

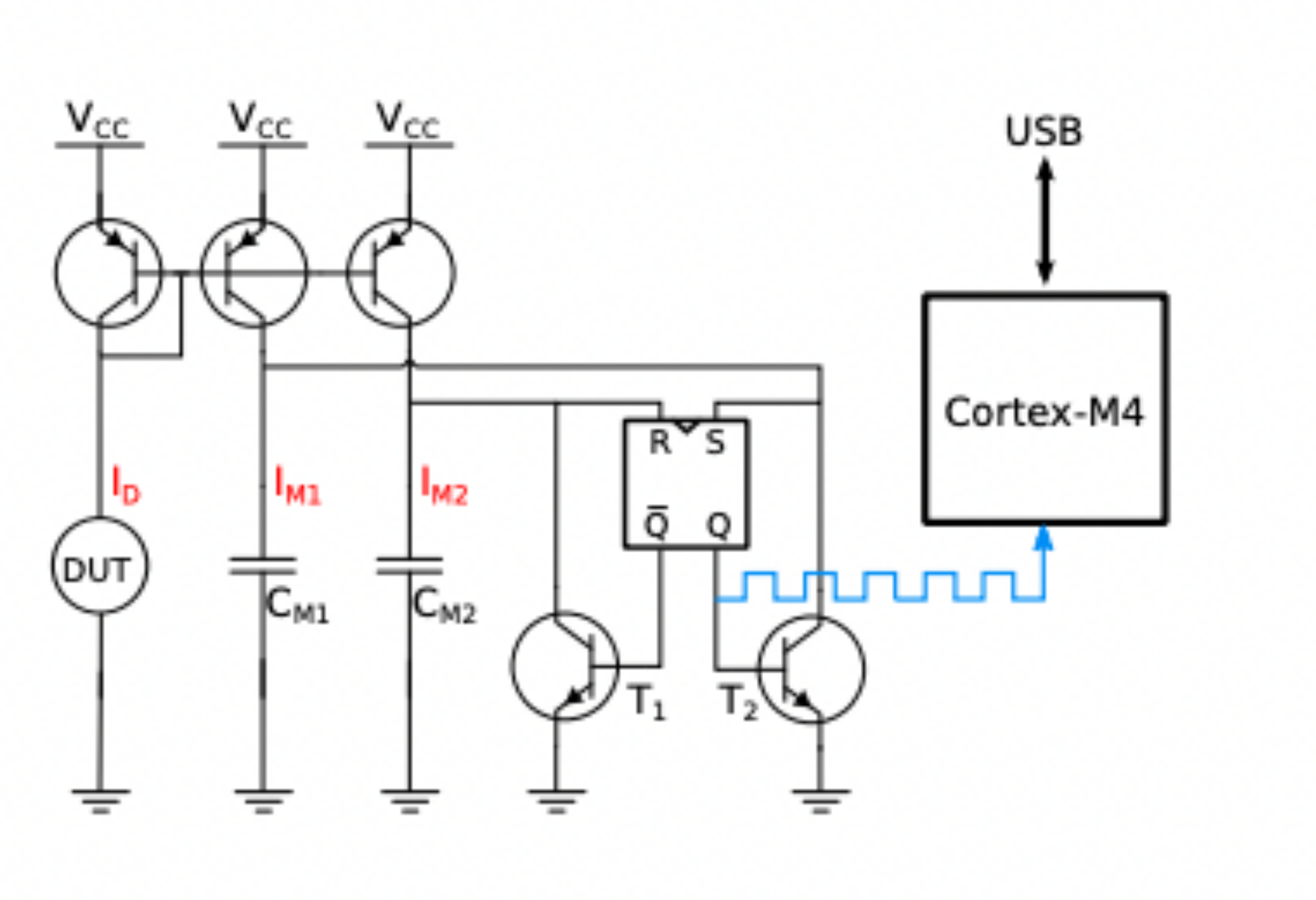
- We are looking for contributers: https://github.com/green-coding-berlin/green-metrics-tool

- https://www.linkedin.com/in/arne-tarara / arne@green-coding.org

</> GREEN CODING;

DC energy measurements

What is the best method possible?



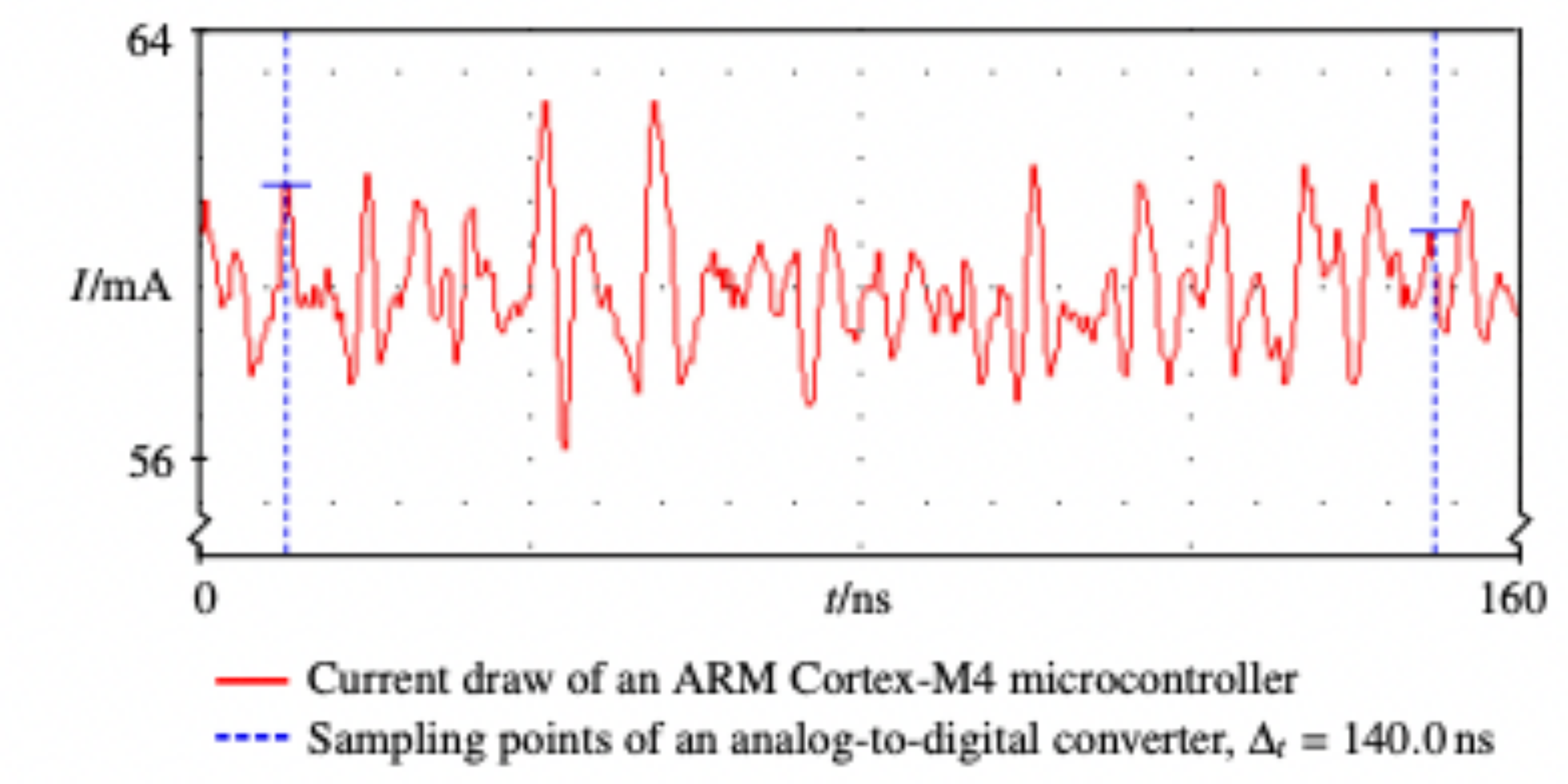


Figure 3: The undersampling of analog-to-digital converters leads to inaccurate energy measurement results.

GREEN CODING;

SPECPower 2008

No uniform setup - Configuration just has to be documented

- Hyperthreading
- SVM / VT-X / VT-D / AMD-V
- Hardware prefetchers (Adjacent Line, DCU, Spatial ...)
- P-States
- C-States
- Memory Speed and Timings
- TurboBoost

Settings in SPECPower

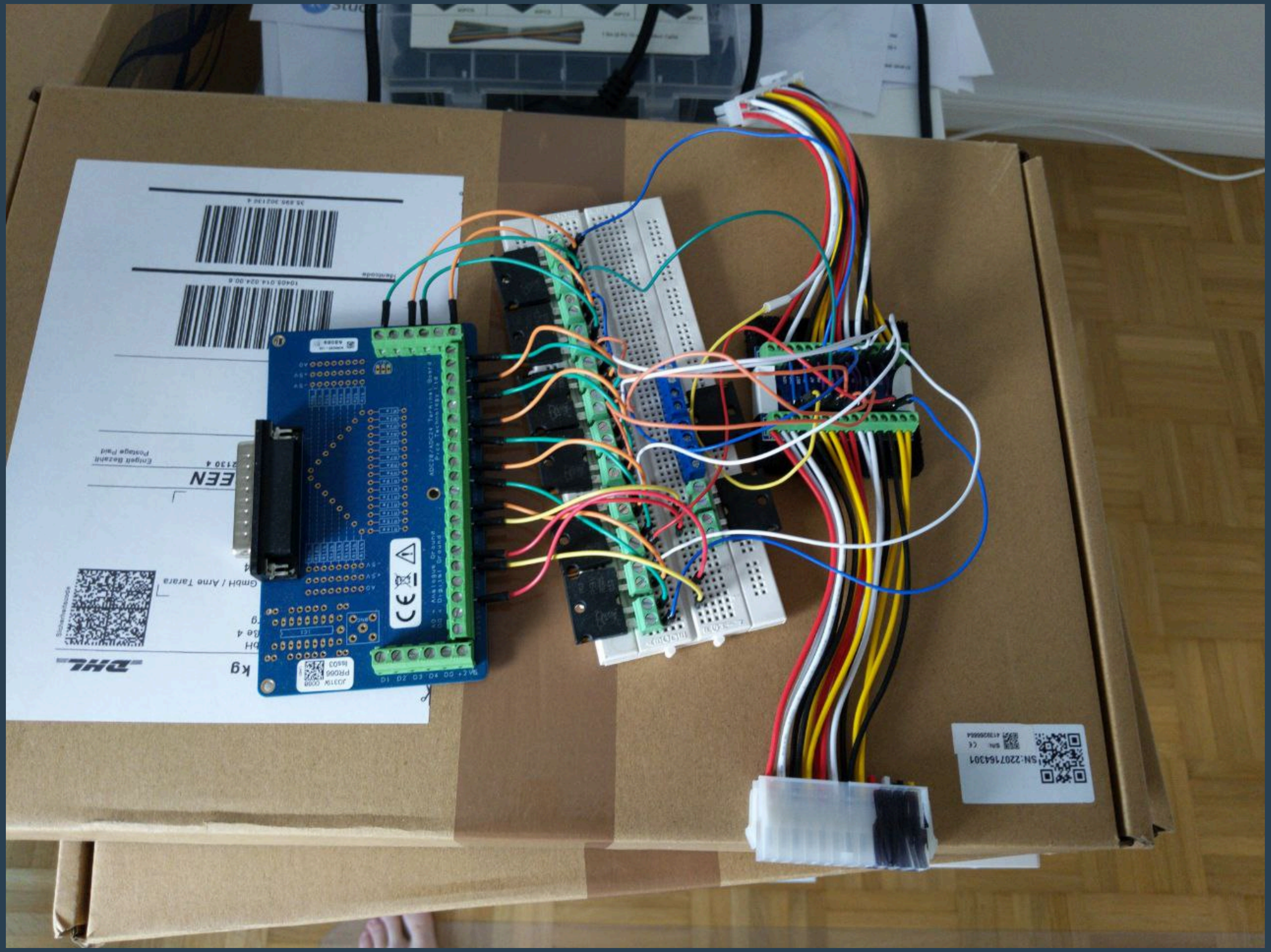
Settings we did not look at / did not find

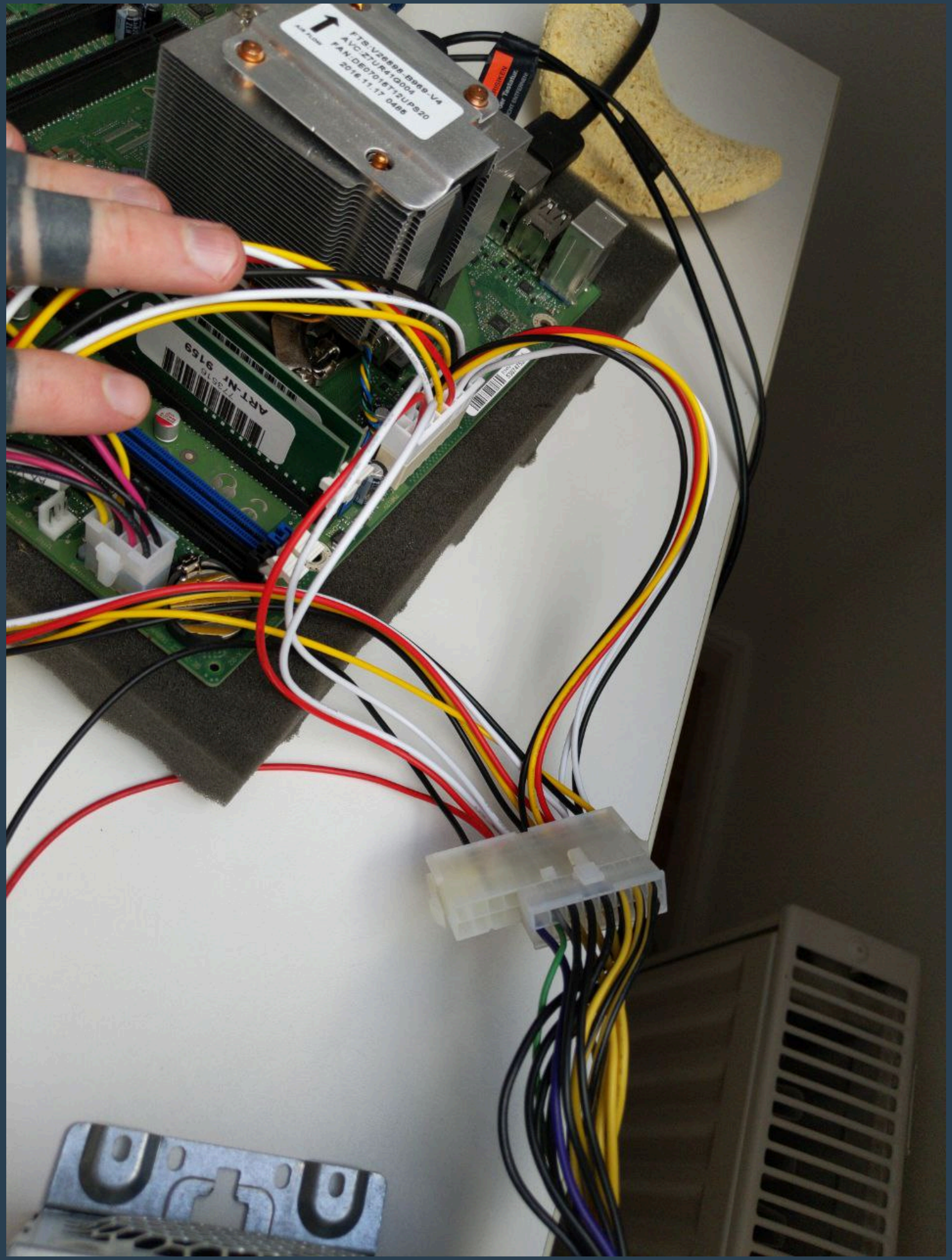
- Fixing workloads to cores
- "Maximum Processor State: 100%" ?
- SATA Controller = Disabled
- Uncore Frequency Override = Power balanced
- DEMT -enabled / EfficiencyModeEn = Enabled
- Memory Data Scrambling: Disable / Set "Memory Patrol Scrub = Disabled"
- ASPM Support - Power saving for PCIe
- SGX enabled / disabled
- .....

</> GREEN CODING;

DC energy measurements

How good is good enough?





GREEN CODING;