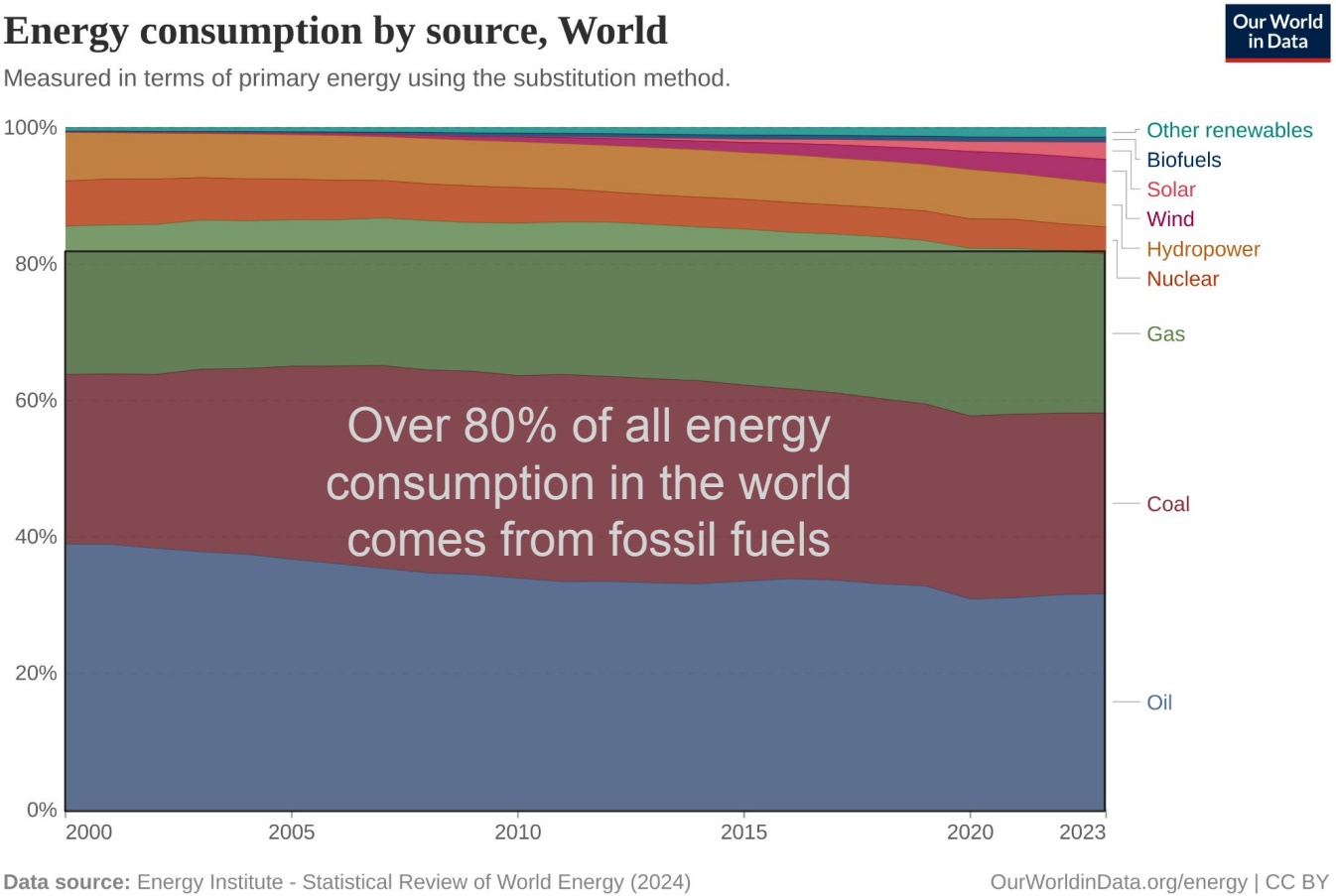
Energy measurements in HPC Architectures

[CMP223] Computer Systems Performance Analysis

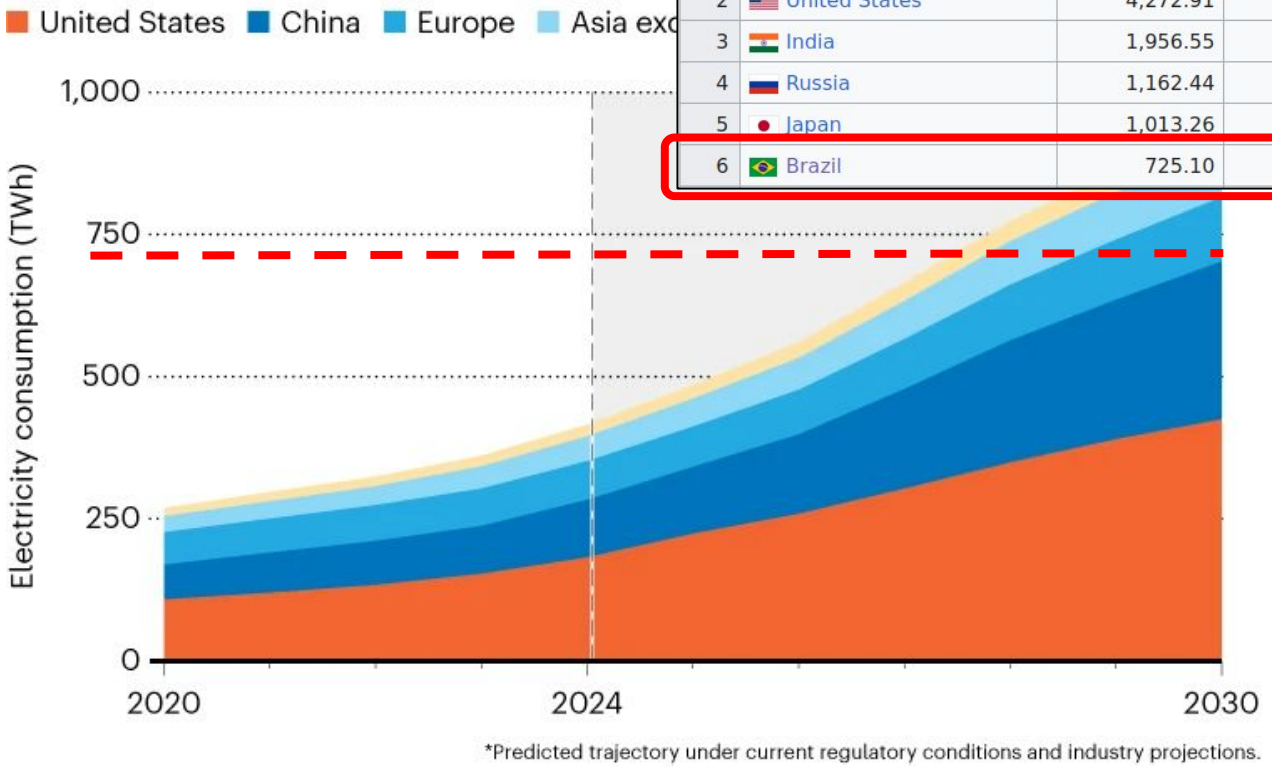
Laura Soares
Otho Marcondes
December/25


Energy consumption by source, World
Measured in terms of primary energy using the substitution method.
Our World in Data
100%
80%
60%
40%
20%
0%
2000
2005
2010
2015
2020
2023
Other renewables
Biofuels
Solar
Wind
Hydropower
Nuclear
Gas
Coal
Oil
Over 80% of all energy consumption in the world comes from fossil fuels
Data source: Energy Institute - Statistical Review of World Energy (2024)
Note: "Other renewables" include geothermal, biomass, and waste energy.
OurWorldinData.org/energy | CC BY

[energy mix]

DATA-CENTRE ENERGY GROWTH

China and the United States are predicted to acc
the global growth in electricity consumption by



	Location	Consumption (TWh)	Per capita (MWh)	Year
	World	29,664.60	3.67	2023
1	China	9,443.07	6.64	2023
2	United States	4,272.91	12.44	2023
3	India	1,956.55	1.36	2023
4	Russia	1,162.44	7.99	2023
5	Japan	1,013.26	8.15	2023
6	Brazil	725.10	3.43	2023

[nature | wikipedia]

Energy measurements in HPC architectures

- From an observability standing point:
  - It is increasingly critical to have energy monitoring tools in data-centres
  - Optimizing energy performance depends on monitoring
  - Allows power management initiatives
- From the application standing point:
  - Energy measurement can be helpful during an experiment

Proposal & Computational Object

- Perform energy measurements on a cluster (computational object)
- Compare the energy use of different applications
  - Idle
  - Stress package
  - LU factorization (StarPU + Chameleon)

Partition	CPU	RAM	Accelerator	Disk	Motherboard
poti[1,2,3,4,5]	Intel(R) Core(TM) i7-14700KF, 3.40 GHz, 28 threads, 20 cores	96 GB DDR5	NVIDIA GeForce RTX 4070	1.7 TB SSD, 119.2 GB NVME	Gigabyte Technology Co., Ltd. Z790 UD AX

Design of Experiments

- Idle
- Stress
  - Machines in use (1~5)
  - CPU Workload (3 levels)
  - IO Operations (3 levels)
  - Memory Operations (3 levels)
- LU Factorization
  - Factor and levels: machines in use (1~5)

Stress

- CPU, IO and memory levels:

time	cpu	io	mem
5m	8	8	8
5m	12	12	0
5m	24	0	0
5m	0	24	0
5m	12	0	12
5m	0	12	12
5m	0	0	24

- For every # of concurrent nodes
  - From 1 to 5 concurrent nodes, for a total of 35 runs
  - Randomized, 1 minute sleep between runs
- Three replications
  - (Allocating the whole partition is hard)

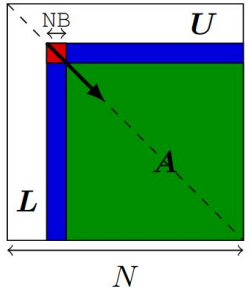
LU Factorization

- Matrix size of 60000, block size of 100, no GPU
- Factor and levels: # of concurrent nodes, from 1 to 5
  - 3 repetitions each (15 experiments total)

$$
\begin{aligned}
Ax &= b \\
(LU)\,x &= b \\
L\,(Ux) &= b
\end{aligned}
$$

$$Ly = b \quad \text{e} \quad Ux = y$$

```
for (k = 0; k < N; k++)
  DGTRF-NOPIV(RW, A[k][k]);
  for (m = k+1; m < N; m++)
    DTRSM(RW, A[m][k], R, A[k][k]);
    DTRSM(RW, A[k][m], R, A[k][k]);
  for (n = k+1; n < N; n++) // Update
    for (m = k+1; m < N; m++)
      DGEMM(RW, A[m][n], R, A[m][k],
                         R, A[k][n]);
```



Instrumentation: Network-manageable Rack Power Distribution Unit (PDU)

- The PDUs (the power outlet) used by the nodes are connected to the internal network of the cluster
  - Access using SSH
  - Answers to SNMP requests
- Provide energy measurements

Metrics: script making SNMP requests

```
while $SECONDS -lt $run_time
    echo `date`
    snmpget etc $IP PowerNet-MIB::ePDUDeviceStatusEnergy.1
    snmpget etc $IP PowerNet-MIB::ePDUDeviceStatusActivePower.1
    sleep $sleep
done
```

time (YYYY-MM-DD HH:MM:SS)
energy (kWh, cumulative)
active power (kW)

Tools we used

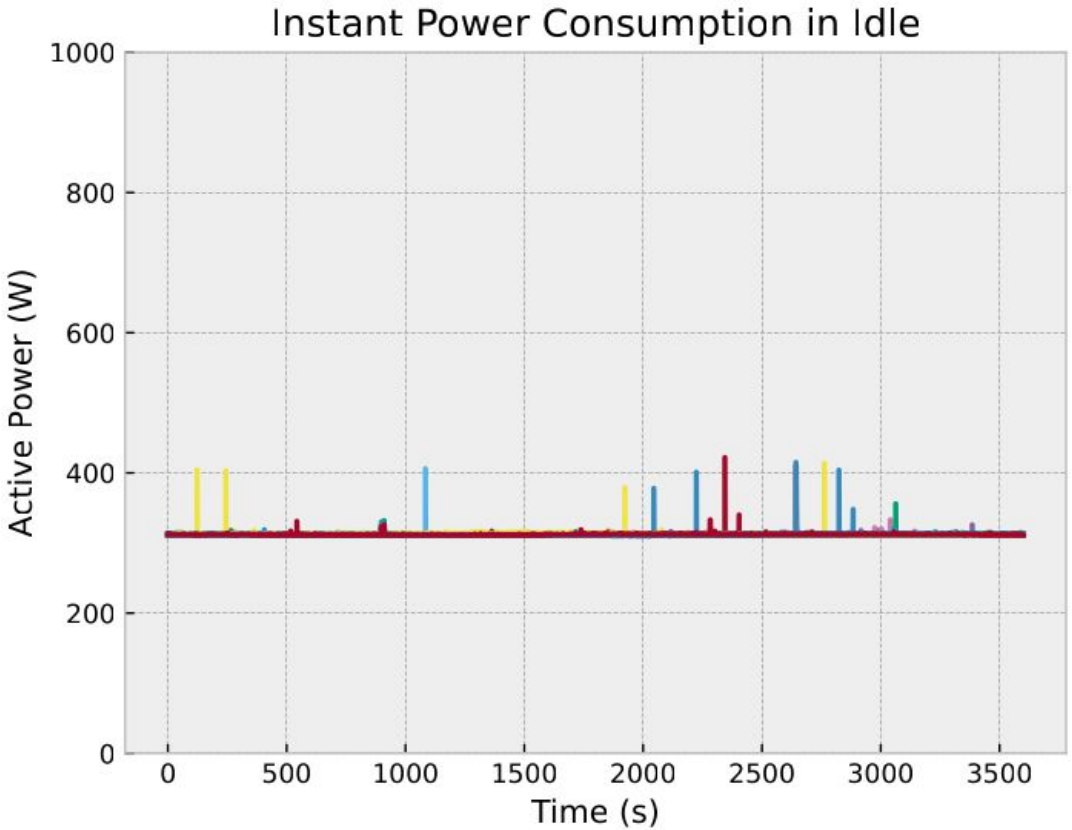
- Slurm scripts for queueing the experiments
- Shell script for the SNMP requests
- Jupiter notebook *and* .org for data analysis (literal programming)
- Git for versioning

❌ Overleaf for the report

❌ Google slides for this presentation 😔

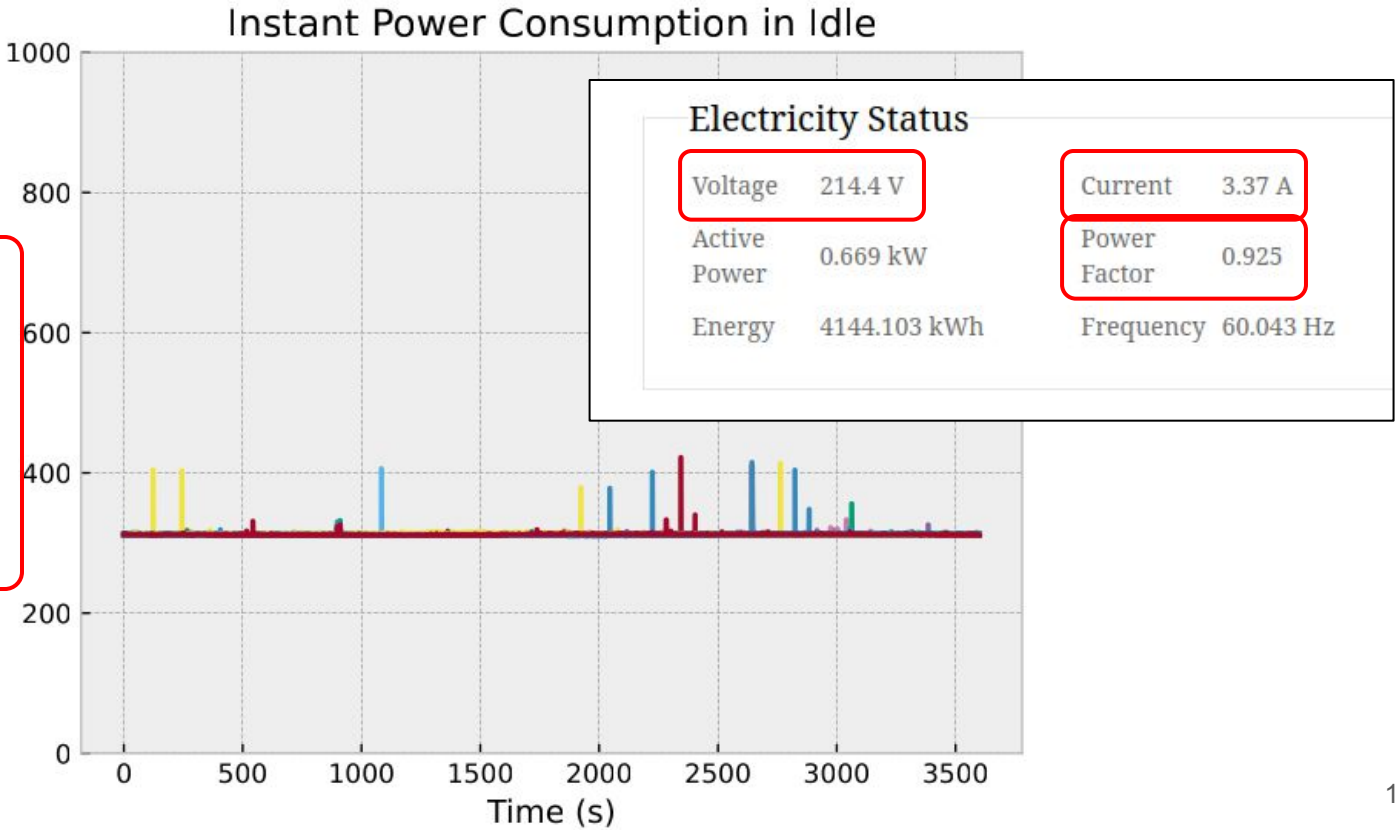
Stage 3: Results 🔋

Poti partition: Idle

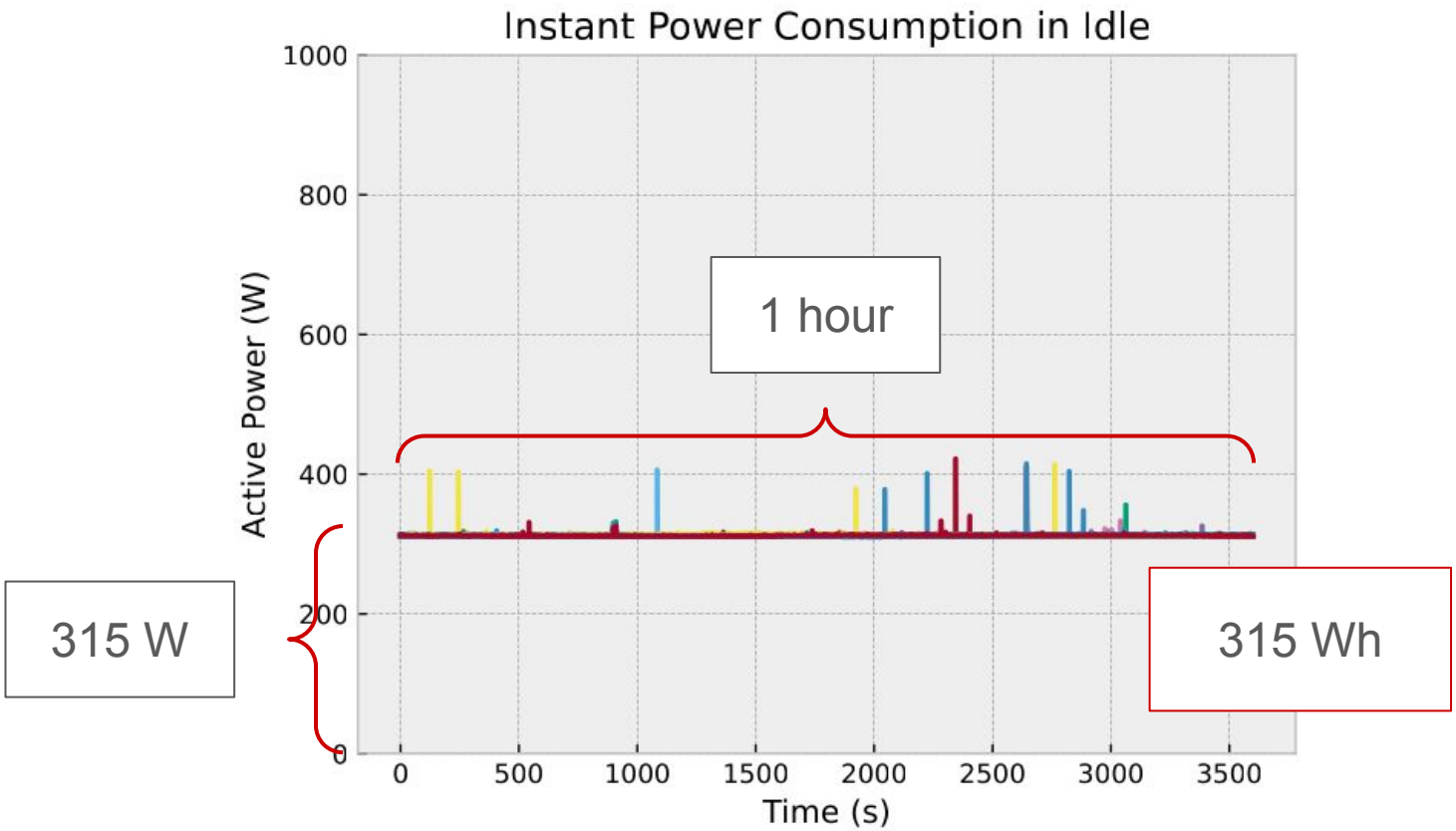


Poti partition: Idle

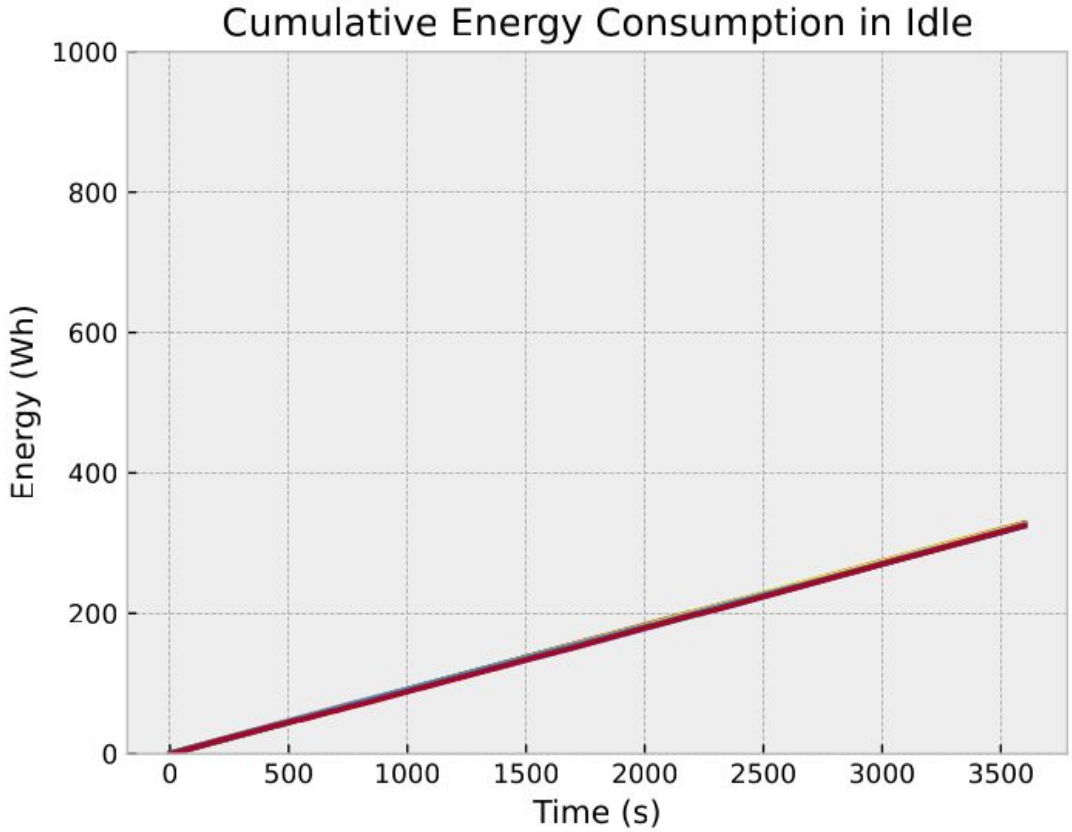
$V \times I \times \cos\phi$



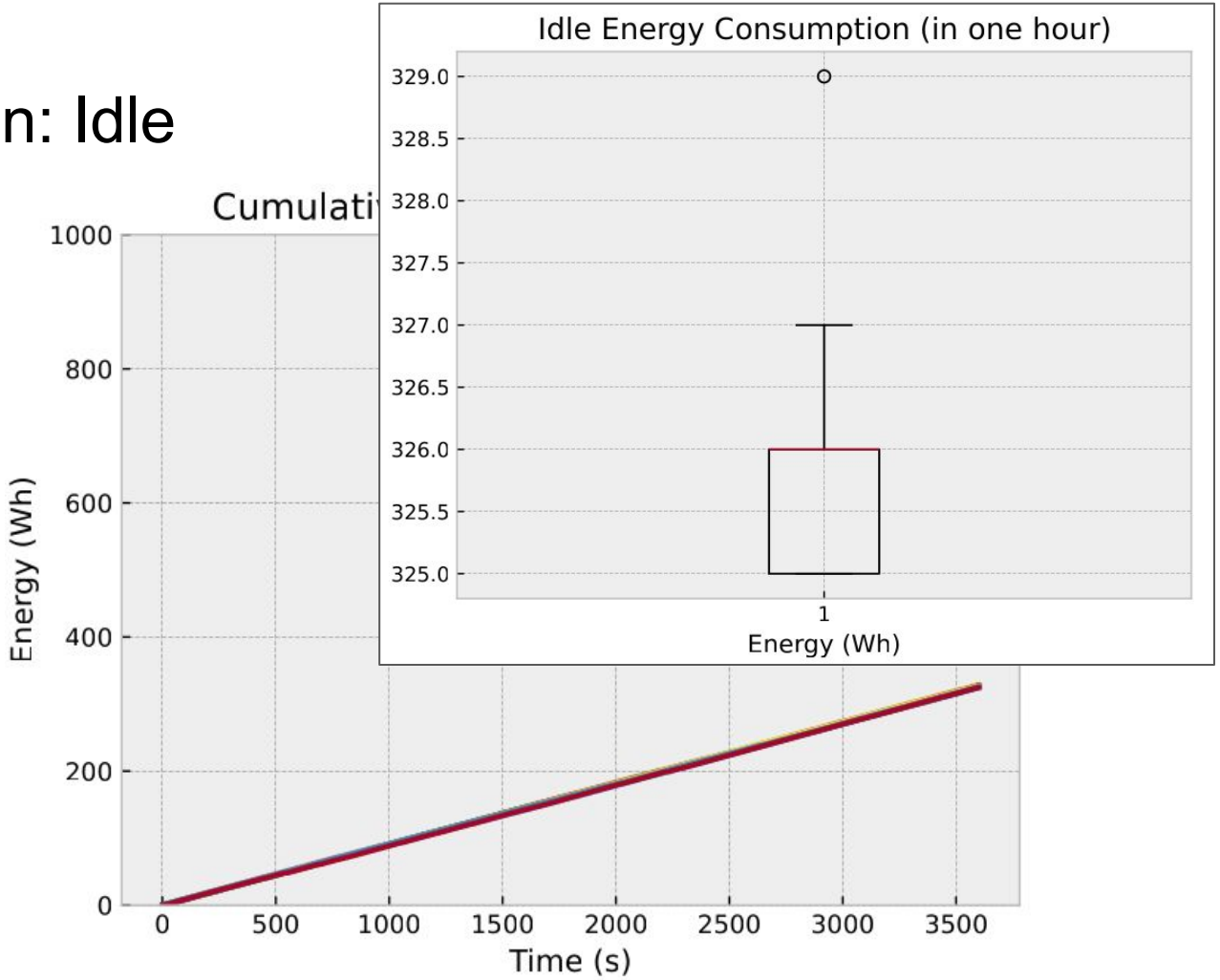
Poti partition: Idle



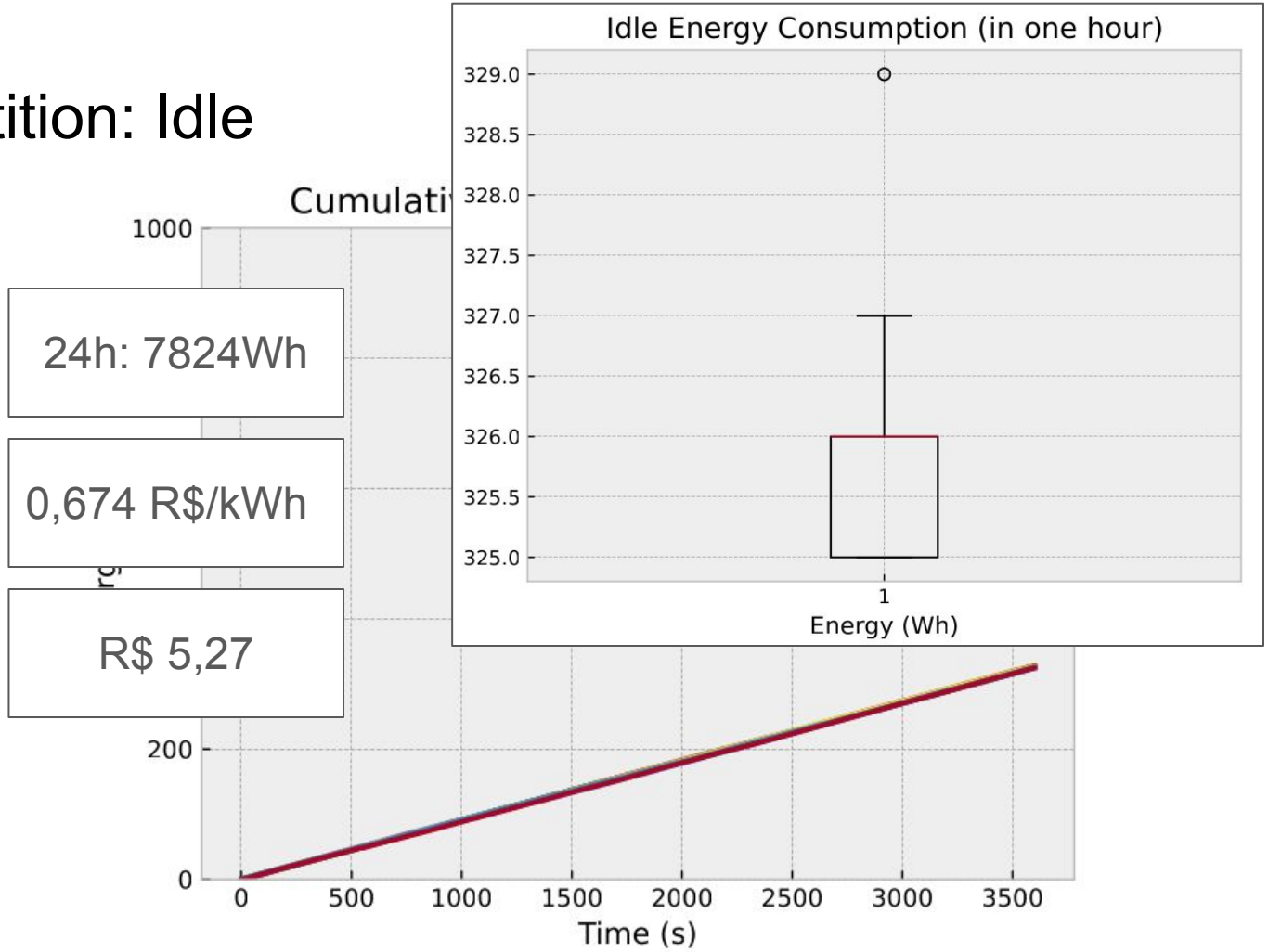
Poti partition: Idle



Poti partition: Idle

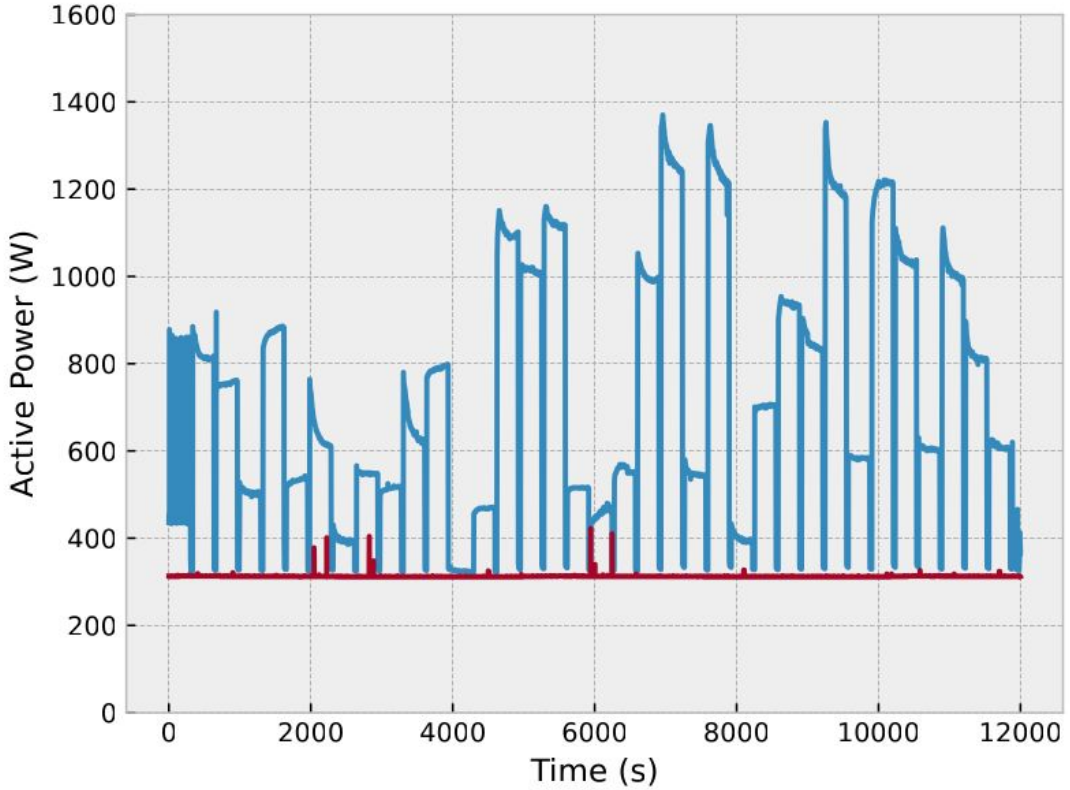


Poti partition: Idle



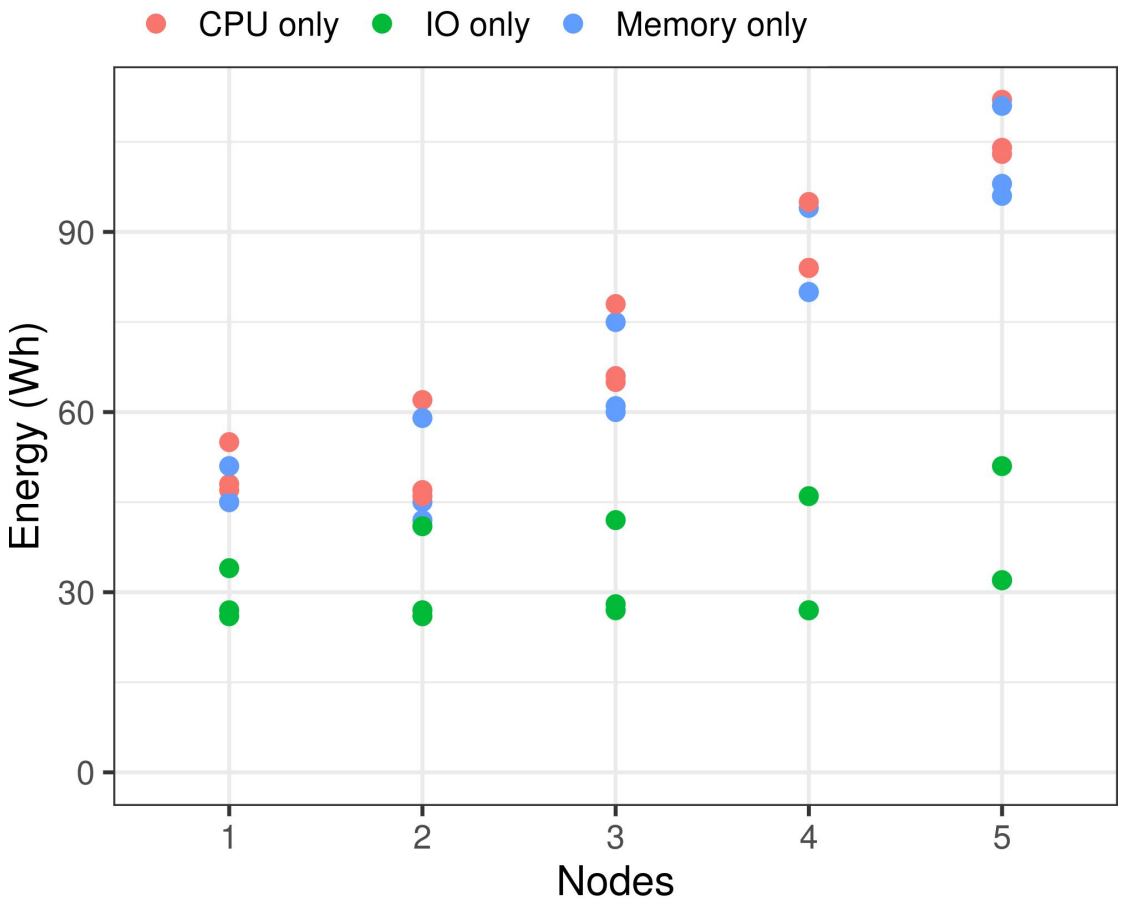
Stress

- 5 minutes each run
- Alternates CPU, io, memory operations and number of concurrent nodes
- Idle for scale



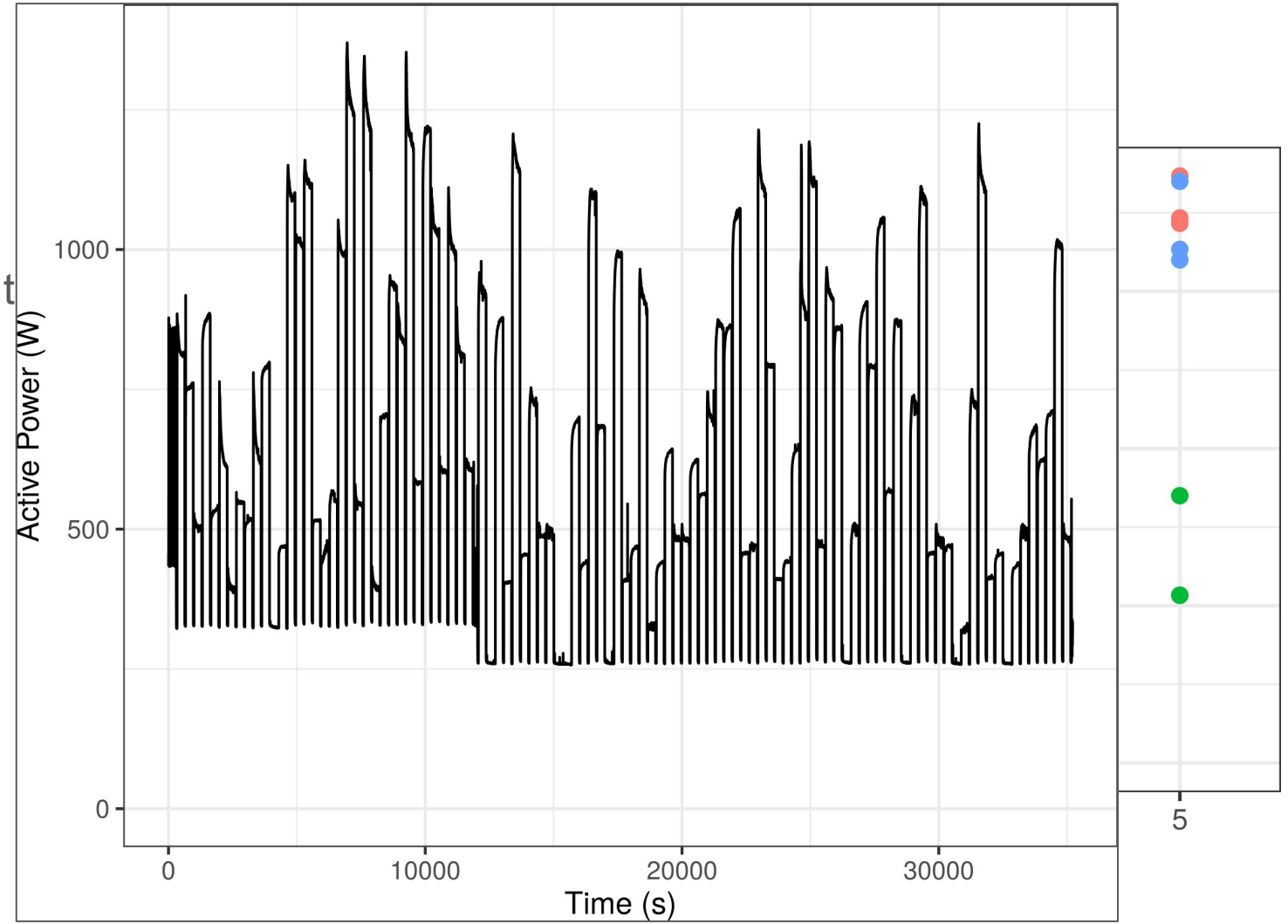
Stress

- Now with the repetitions
- Problem?

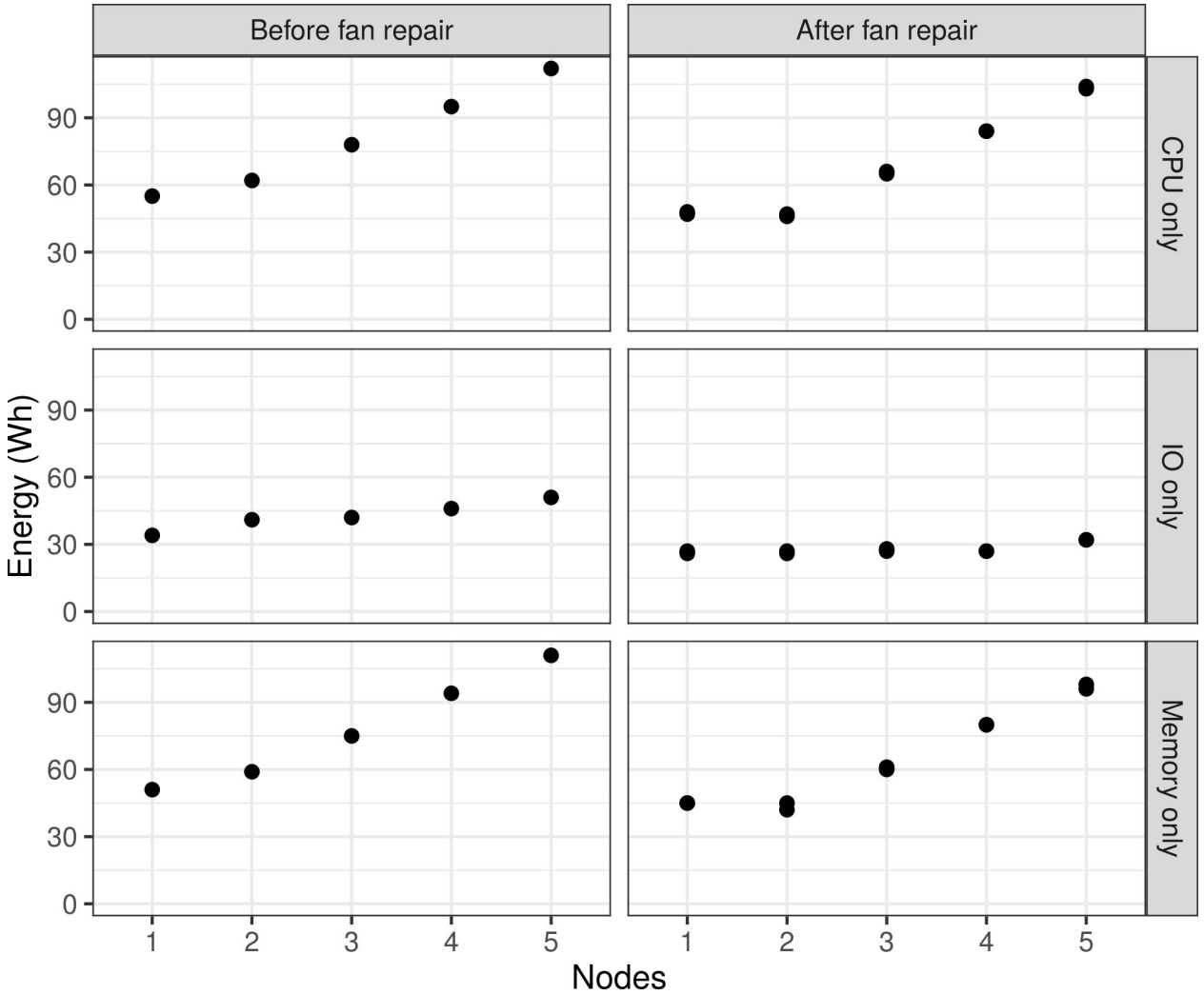


Stress

- Now with t
- Problem?



Stress


Before fan repair
After fan repair
CPU only
IO only
Memory only
Energy (Wh)
0
30
60
90
1
2
3
4
5
Nodes

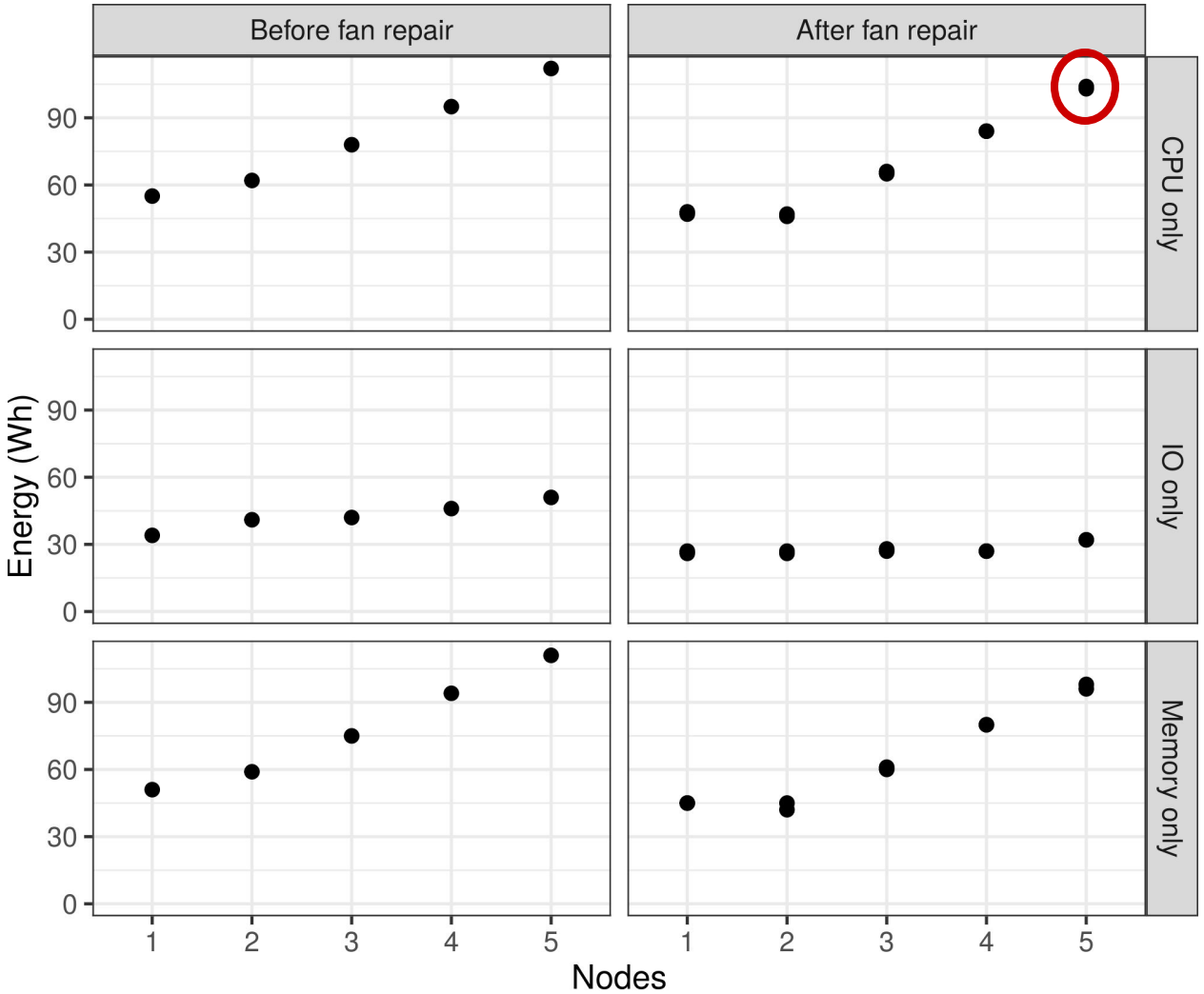
Stress

105Wh in 5 min

1h: 1260Wh

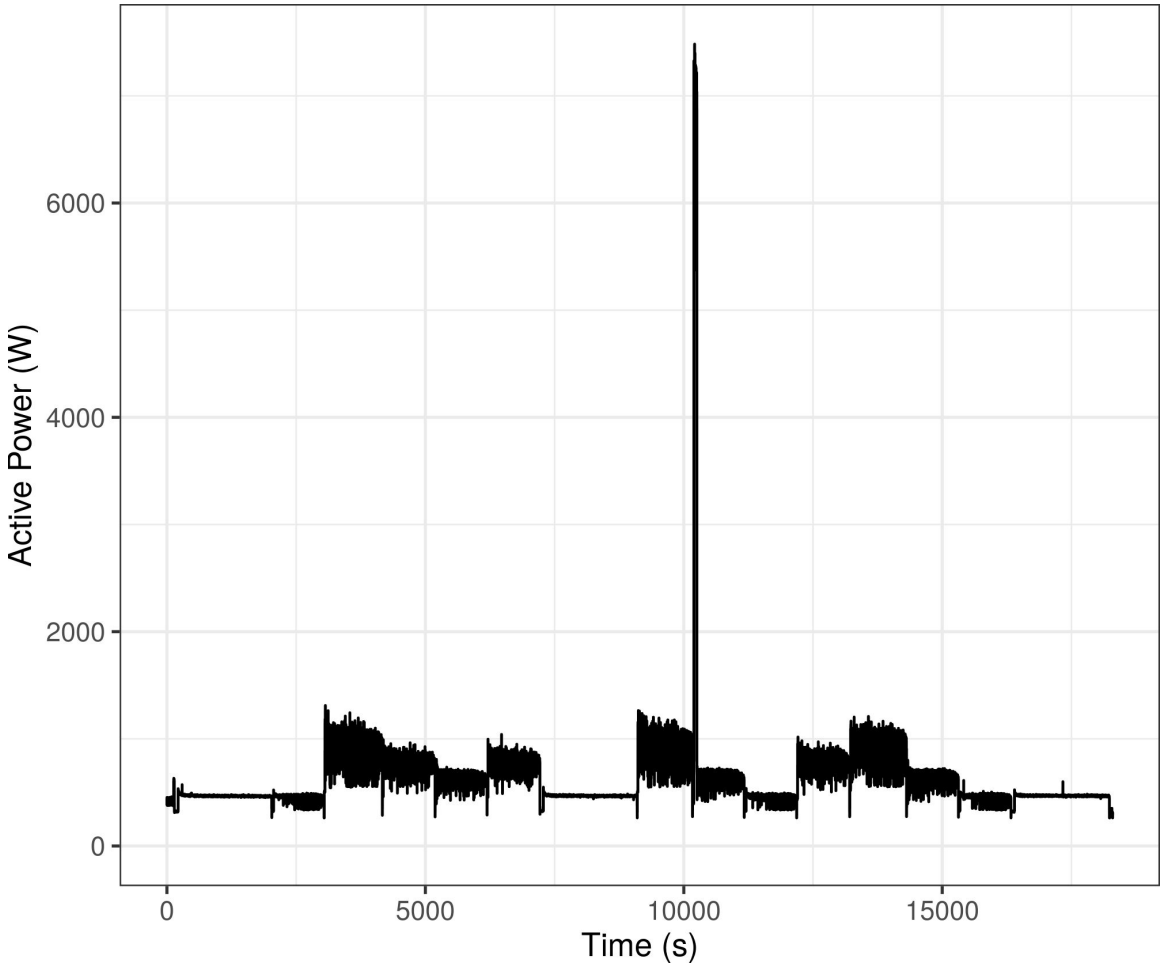
24h: 30,2kWh

R$ 20,38

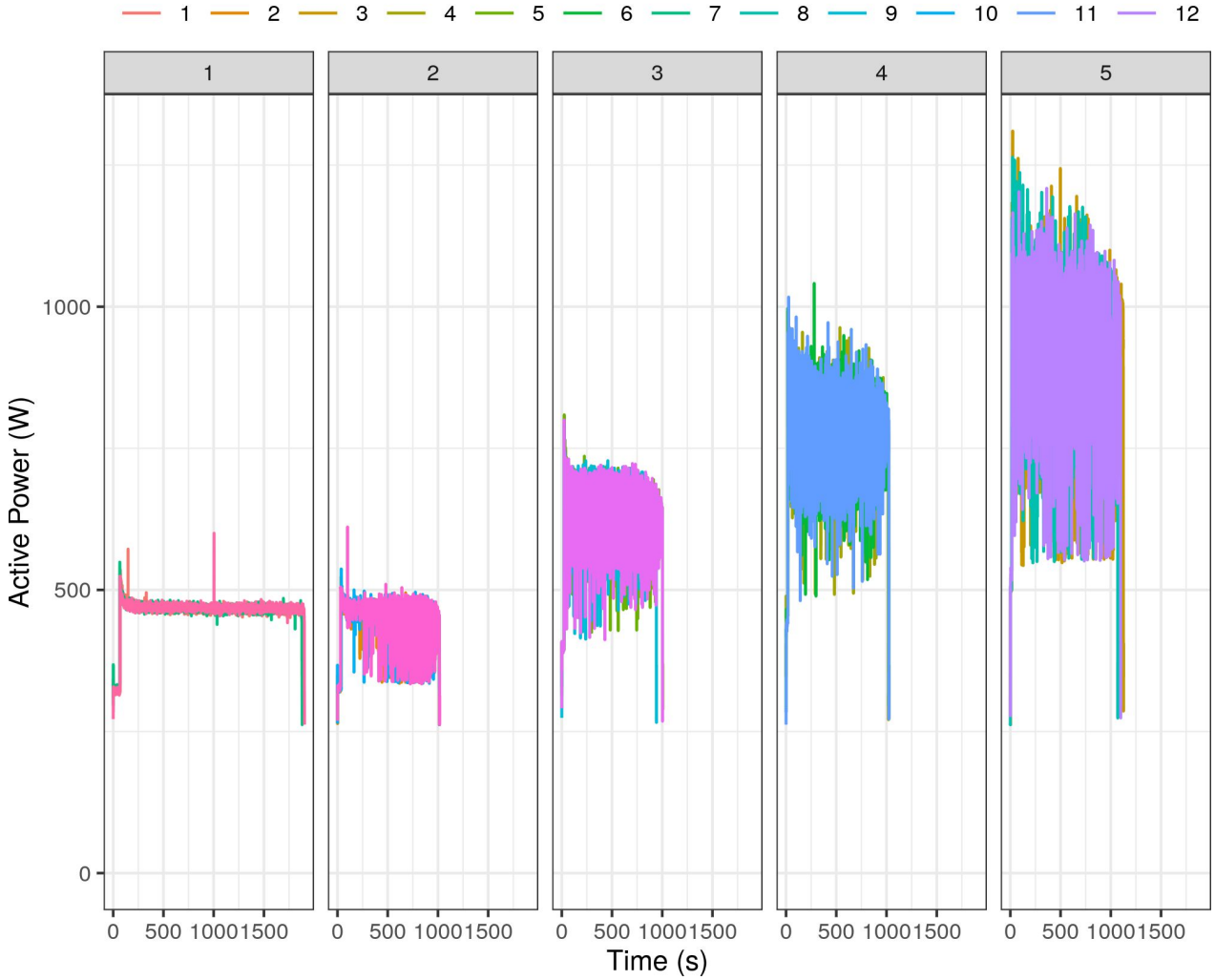


LU Factorization

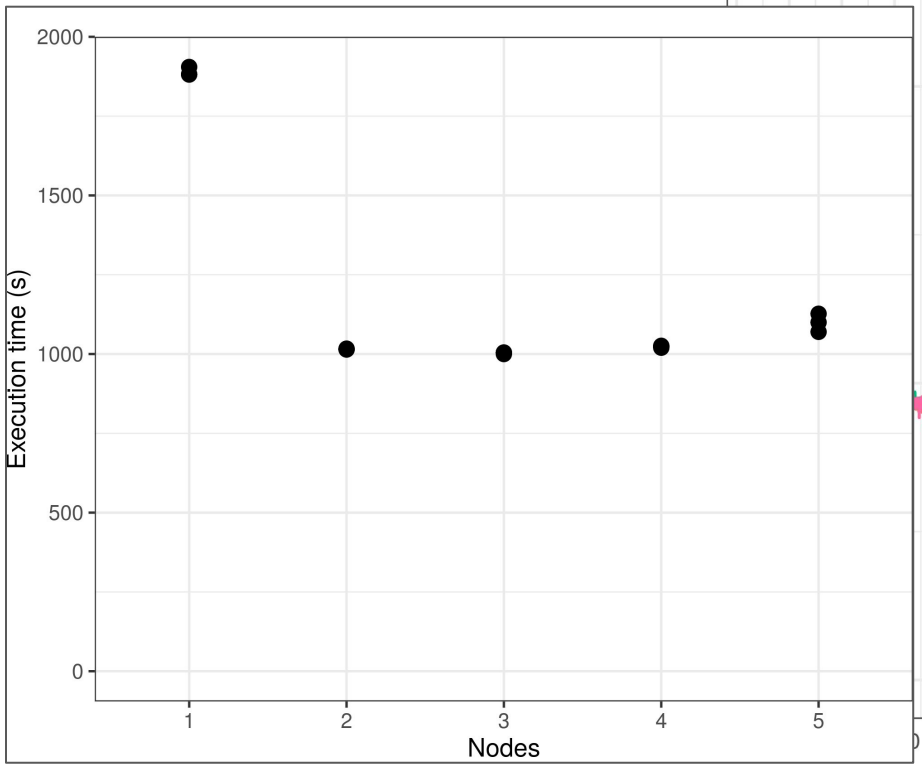
- Outlier
  - Measurement problem?

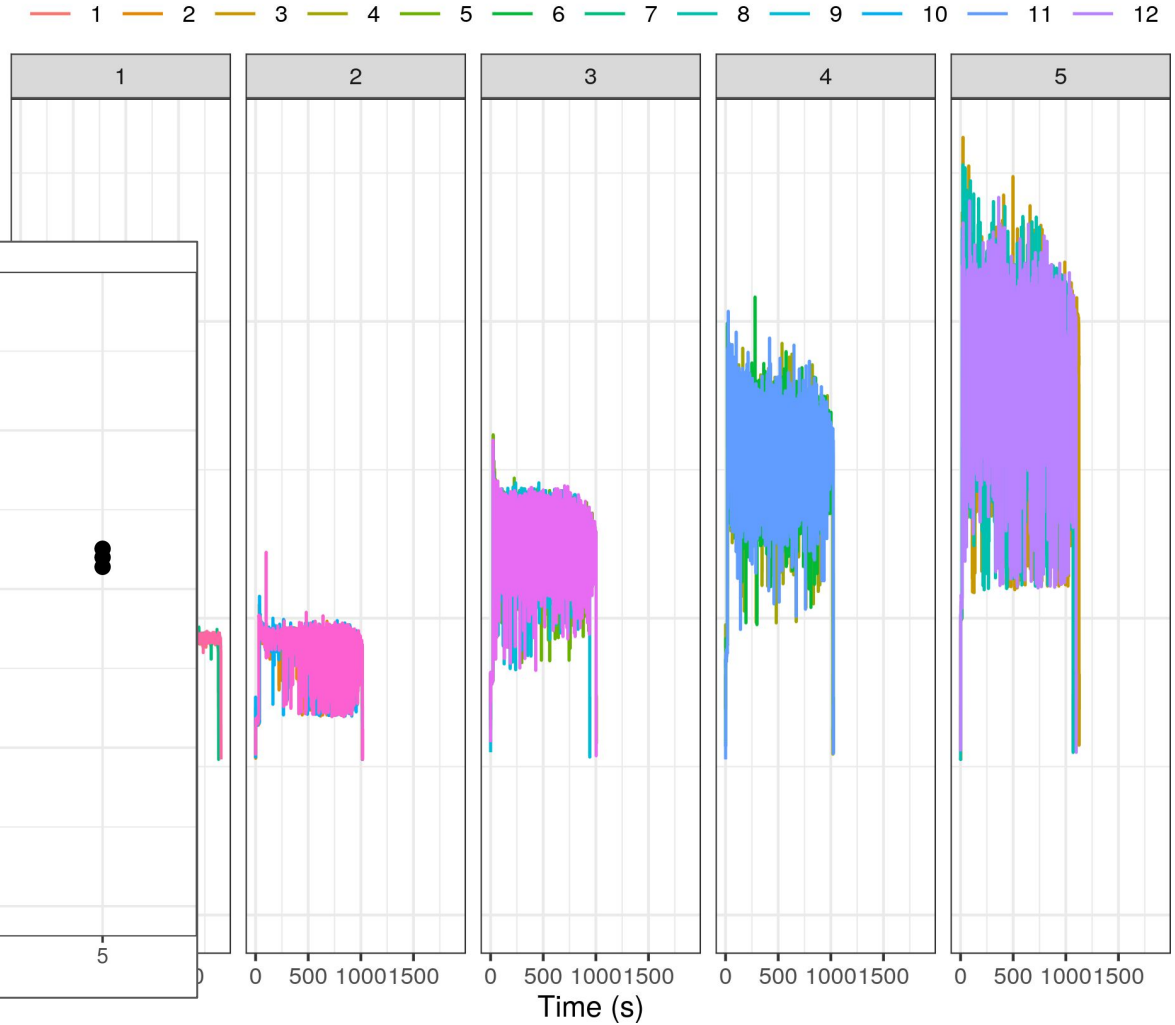


LU Factorization



LU Factorization





Limitations

- Fix the weird peaks in measurement
  - Combine other measurement tools with the PDU, like RAPL counters at the machine hardware (might cause intrusion)
- Needs way more repetitions
  - Hard to set apart the application normal behavior with problems in the measurement
  - More repetitions would allow us to run a statistical model

Conclusion & next steps

- Energy measurements & observability tools are useful in the cluster
  - Besides energy optimization, it can help users to find problems in their applications, hardware problems, and so on
- Work in progress about observability tools at the cluster (Prometheus, Grafana, etc)
- Maybe use comparisons other than R$/kWh, like water per kWh

Thank you!! ✨🧑‍🎓🧠💖🔋

Any questions?

lrsoares@inf.ufrgs.br
otho.marcondes@inf.ufrgs.br

References 1

[energy mix] https://ourworldindata.org/energy-mix

[nature] https://www.nature.com/articles/d41586-025-01113-z

[wikipedia] https://en.wikipedia.org/wiki/List_of_countries_by_electricity_consumption

[Schneider Electric] https://eshop.se.com/in/blog/post/difference-between-active-power-reactive-power-and-apparent-power.html

[wikipedia 2] https://pt.wikipedia.org/wiki/Valor_eficaz

[ICPADS] Lucas Leandro Nesi, Lucas Mello Schnorr, Arnaud Legrand. Communication-Aware Load Balancing of the LU Factorization over Heterogeneous Clusters. IEEE International Conference on Parallel and Distributed Systems (ICPADS), Dec 2020, Hong Kong, France. hal-02633985