Instituto de Física de Cantabria

CSIC
Consejo Superior de Investigaciones Científicas

UC | Universidad de Cantabria

GreenDIGIT | All Hands Meeting
EGI Conference 2024
October 1st 2024

IFCA + AI4EOSC Power Consumption & Environmental Impact Evaluation

Design, implementation, challenges

Álvaro López García & Jaime Iglesias
(on behalf of IFCA team)
{aloga,iglesias}@ifca.unican.es

Advanced Computing and e-Science Group
https://advancedcomputing.ifca.es

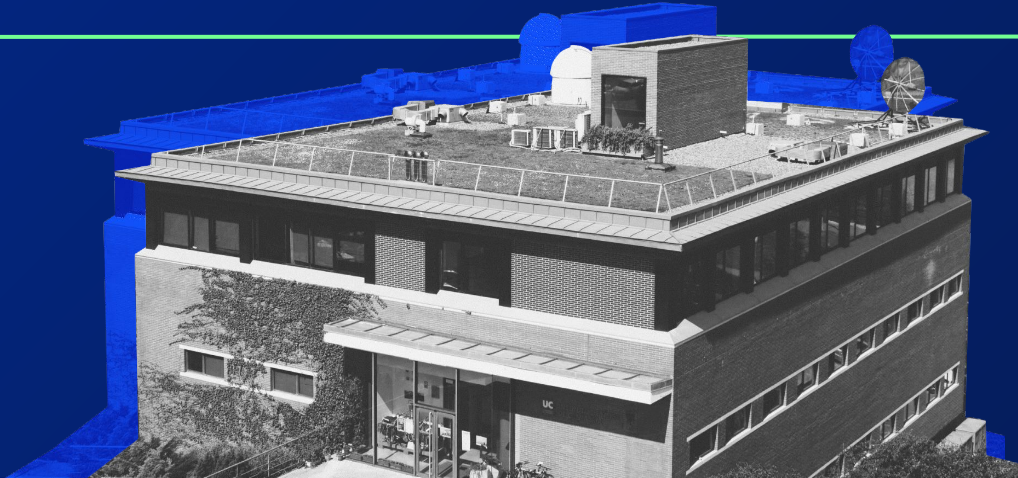


Table of contents:

Introduction

Framing the background and problem

Introduction & motivation

AI4 | eosc

AI4EOSC Platform is delivering an easy to use toolbox to develop, share and deploy Artificial Intelligence (AI) and Machine Learning (ML) models within the European Open Science Cloud (EOSC) context.

- Easy to use, developer (data scientist) focus.
- Transparent access to underlying e-Infrastructures
  - 5 federated datacenters, > 1700 CPU, > 70 GPUs

- SotA AI/ML models can leverage significant computing power during the training phase (but also for inference).
- AI4EOSC aims to make users aware of the impact produced by training and deploying AI models.

Power and impact measurement

AI4 eosc

- Produce metrics at datacenter level.
  - PUE, energy quality, efficiency, hardware characteristics.
  - Currently static, working to make data collection dynamic.
  - **Objective**: improve job allocation within the platform.
- Produce metrics at model level.
  - Impact on training phase. (i.e. how much does it *cost* to build a model)
  - Impact on inference/prediction. (i.e. how much does it *cost* to use the model)
  - **Objective**: make developers, scientists and final users aware of the impact of a given model.
- Work in progress:
  - Metadata (i.e. how to publish), automatic data collection, integration with external data sources.

GHG Protocol's ICT Sector Guidance

- (One of the) Most adopted protocols to report GHG emissions in IT.
- It accounts for all GHG emissions by allocating all of the emissions/power consumption of the data centre/server to the VMs.
- Drawback. Some factors are out of user control: single VM in a server, will account higher than one with over VMs running on the same host.
- It focuses on completeness and not on accuracy.

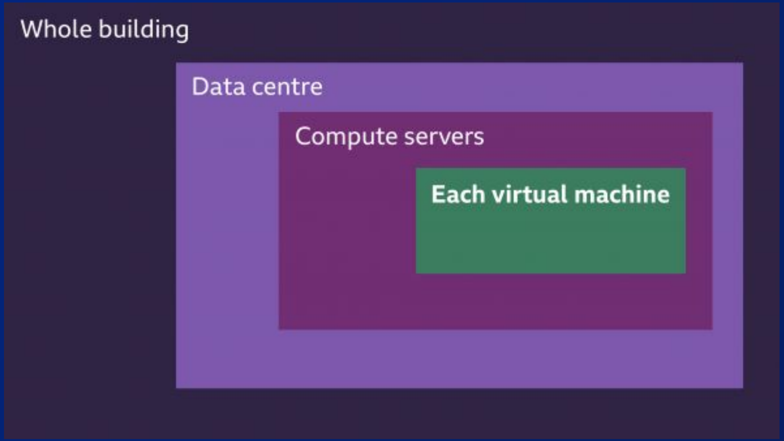


Figure: The GHG Protocol's approach to allocation of GHG emissions to a virtual machine. (https://www.bbc.co.uk/rd)

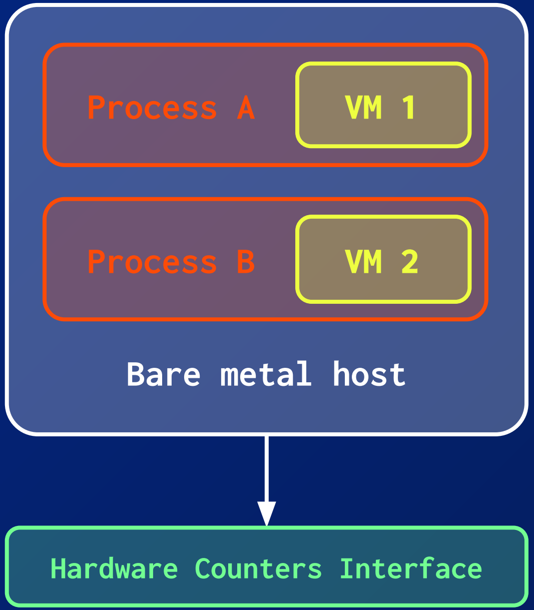
Approach

How AI4EOSC aims to gather (bottom-up) energy consumption

Approach Followed

Step 1: Get the power consumption of cloud virtual machine

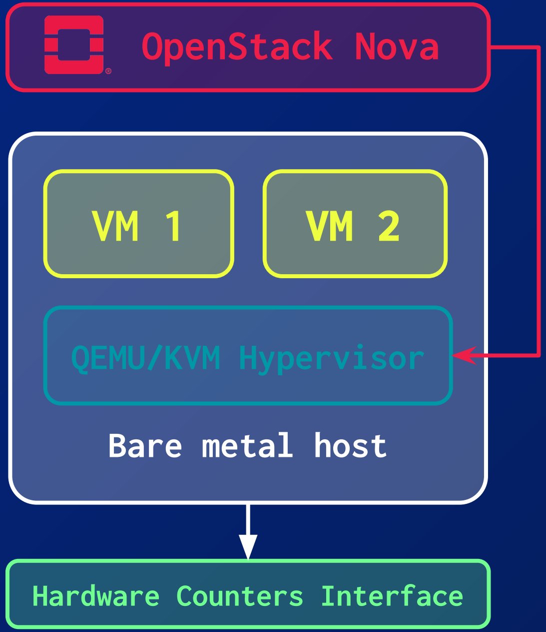
- **Restriction:** The only way to measure VM power consumption is at infrastructure level (i.e. bare metal host, where it is allocated).
  - Because we depend on kernel access to the hardware counters.
- **Assumption:** We can measure power consumption of each process running on bare metal host.
- **Statement:** Each running VM appears as a process on the server.
- **Result:** The watts consumed by a VM are the watts consumed by the process that executes it.



Approach Followed

Step 2: Share with the cloud VM each own measurements

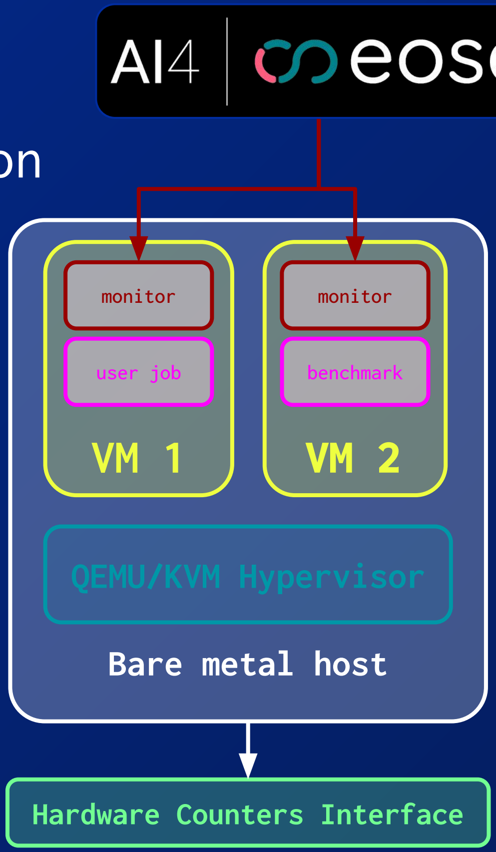
- **Restriction:** The measurement has to be available within the platform, i.e. available on each virtual machine.
- **Assumption:** The hypervisor (QEMU/KVM) can share information between both systems.
  - It has to be able to share information in a variable multi-tenant computing environment (i.e. cloud infrastructure operated by OpenStack)
- **Result:** The VM has access to its own power consumption metrics.



Approach Followed

Step 3: Measure workload energy consumption

- **Restriction:** The platform needs to track the energy consumption for workloads running inside it
- **Assumption:** Once measurements are available, we can instrument user workloads with different software (CodeCarbon, perun, etc.) to measure energy consumption
    - Exploit this to run platform level benchmarks to assess energy efficiency, independently of user workloads
- **Result:** The platform software (AI4OS) has access to energy consumption metrics, both platform benchmarks or real user workloads



Approach Followed

Step 4: Publish metrics and make users aware

- **Restriction:** Publish metrics following an open, common and agreed schema, with crosswalks between different formats
- **Assumptions:**
  - Raw data, once generated, is easy to collect from the platform.
  - AI4EOSC model metadata can be easily extended to include CO_2/energy/impact metrics.
  - Some ontologies and data formats exist to publish workload metrics, and IT systems impact.
- **Result:** AI4EOSC can make decisions on workload placement. Users can be more aware of the impact of their activities.

Step 1.1:

Measuring VM Power Consumption

Scaphandre Agent

Scaphandre

Is an **open-source metrology agent** dedicated to electric power and energy consumption metrics created by Hubblo.

The goal is to permit to anyone to measure the power consumption of its tech services and get this data in a convenient form, sending it through any monitoring or data analysis toolchain.

Relevant Features:

- Measuring power/energy consumed on bare metal hosts.
- Measuring power/energy consumed of QEMU/KVM VM from the host.
- Exposing power metrics of a VM, to allow manipulating those metrics in the VM as if it was a bare metal machine.
- Exposing metrics as a Prometheus (HTTP) exporter.

Scaphandre Internals

Sensors:

- For GNU/Linux: **Powercap/RAPL**
  - Requires Intel / AMD x86 CPUs, produced after 2012.

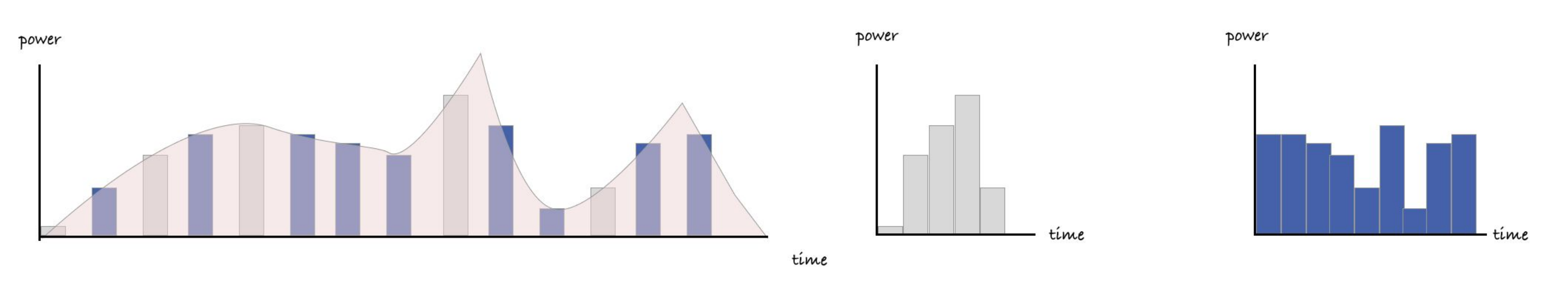
Exporters:

- **QEMU:** Computes energy consumption metrics for each QEMU/KVM virtual machine found on the host. Exposes those metrics as filetrees compatible with the powercap_rapl sensor.

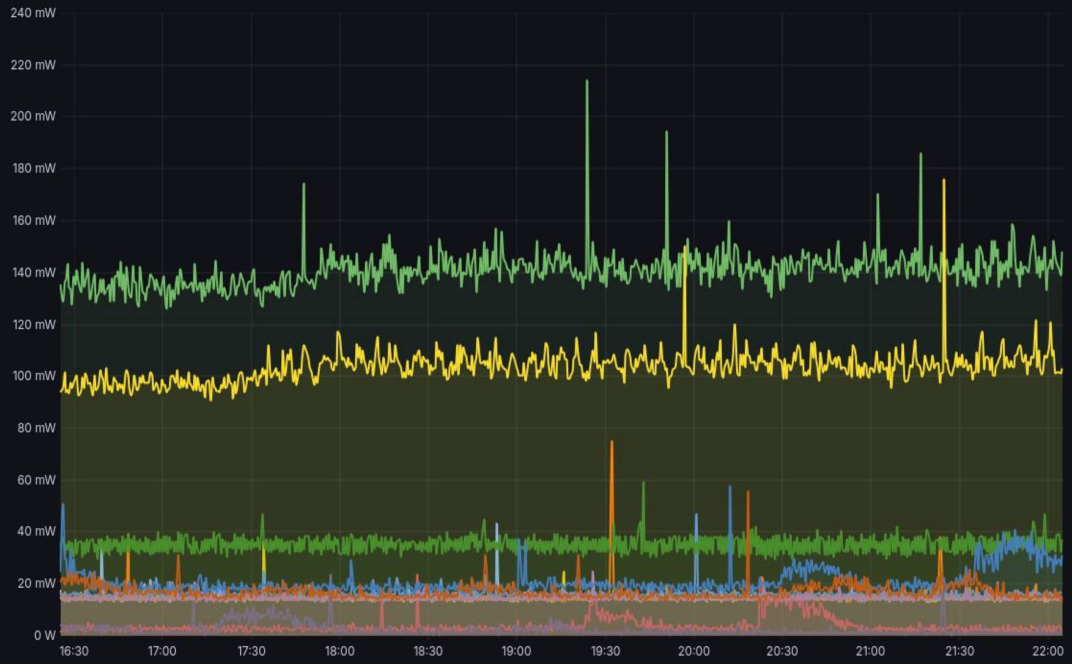
- **Prometheus:** Exposes power consumption metrics on an HTTP endpoint.

How to get the consumption of one process?

- Each VM of the platform is a QEMU/KVM process on the bare metal host.
- The idea is to only measure the power consumption when the process is running.
- Data about process statistics are stored in */proc/stat*.



Source: https://hubblo-org.github.io/scaphandre-documentation/explanations/how-scaph-computes-per-process-power-consumption.html scloud01 VMs / Source: Scaphandre



Name	Last *	Min	Max	Mean
scloud01 - instance-00100af2 - 3d846f22-c008-4a4e-adc8-738df9c949a4	148 mW	126 mW	214 mW	141 mW
scloud01 - instance-00101023 - 3e4e43bc-9f7f-4cf7-b5b3-db8a6ec151a7	13.0 mW	13.0 mW	36.3 mW	14.3 mW
scloud01 - instance-0010109b - 70c22e2f-d9fd-460e-ade4-02656f1d9878	15.1 mW	13.0 mW	43.2 mW	14.5 mW
scloud01 - instance-001010a4 - 7583358a-d42f-468a-b7ee-ed58507e7b1c	13.7 mW	13.0 mW	74.7 mW	14.8 mW
scloud01 - instance-000fdafe - 81b896e3-0c15-4570-a69c-e22410238a07	1.37 mW	684 µW	23.2 mW	3.61 mW
scloud01 - instance-001008c4 - b83cd8e9-a0b2-4c3b-af71-3f7cf31fa020	16.4 mW	14.4 mW	46.6 mW	16.1 mW
scloud01 - instance-00100f2a - b8e0a6ff-9dca-49fb-b793-2ca790e63d56	14.4 mW	13.0 mW	24.7 mW	14.5 mW
scloud01 - instance-00100a98 - c4a922a7-5994-4b11-b89a-8bcdd74ef5de	685 µW	0 W	23.3 mW	2.08 mW
scloud01 - instance-000fdb01 - c94950cc-9d92-4533-ab0f-e7671c04e175	37.0 mW	29.4 mW	58.9 mW	35.0 mW
scloud01 - instance-00100aef - cbaad9db-689b-40e5-baf0-4662b18268c6	103 mW	90.4 mW	176 mW	103 mW
scloud01 - instance-000fbbea - edcff64c-ce57-4cb8-8dd0-2cdd95d644f6	30.8 mW	15.1 mW	57.6 mW	20.7 mW
scloud01 - instance-00100af5 - fbb97ae0-7891-4421-be67-53096d1b40d1	14.4 mW	13.0 mW	55.5 mW	16.7 mW

Developed by Advanced Computing and e-Science Group
With Scaphandre, Collectd, Telegraf, InfluxDB and Grafana

IFCA Instituto de Física de Cantabria

Step 1.2:

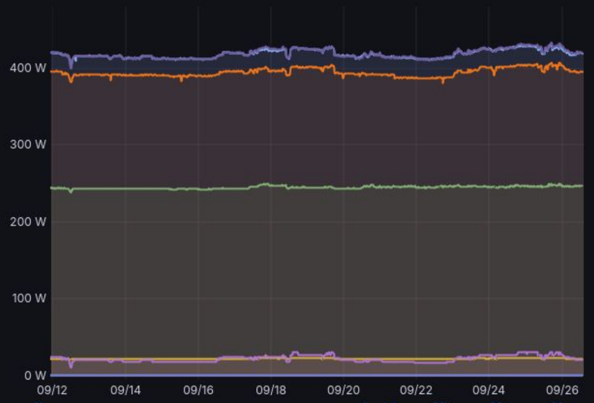
Measuring GPU Power Consumption

Because not everything are CPUs

Comparison of different methods of measurement on a host without GPU

scloud01

scloud01 / Source: IPMI (instantaneous)



Name	Last *	Min	Max	Mean
scloud01 - Sys Power power_unit (19.1)	418 W	399 W	432 W	418 W
scloud01 - Sys Fan Pwr power_unit (19.4)	20 W	10.5 W	30 W	21.7 W
scloud01 - PSU 2 DC Out Pwr power_unit (19.8)	0 W	0 W	0 W	0 W
scloud01 - PSU 2 AC In Pwr power_unit (19.6)	0 W	0 W	0 W	0 W
scloud01 - PSU 1 DC Out Pwr power_unit (19.7)	393 W	380 W	406 W	393 W
scloud01 - PSU 1 AC In Pwr power_unit (19.5)	418 W	404 W	430 W	417 W
scloud01 - Mem Power power_unit (19.3)	21.9 W	20.3 W	22.7 W	21.5 W
scloud01 - CPU Power power_unit (19.2)	246 W	238 W	249 W	244 W

scloud01 / Source: RAPL (average)

Name	Last *	Min	Max	Mean
RAPL total	265.570 W	260.738 W	268.735 W	264.363 W
RAPL package-1	122.288 W	121.168 W	124.182 W	121.938 W
RAPL package-0	121.441 W	118.278 W	122.790 W	120.793 W
RAPL dram-1	11.115 W	10.512 W	11.480 W	10.981 W
RAPL dram-0	10.728 W	10.242 W	11.109 W	10.654 W

scloud01 / Source: Scaphandre (average)

Name	Last *	Min	Max	Mean
scloud01 - scaph_host_power	265.576 W	257.510 W	268.860 W	264.362 W

scloud01: IPMI (CPU+Mem) / IPMI (system) last difference ratio — 64.578%

scloud01: RAPL / IPMI (system) last difference ratio — 63.993%

scloud01: Scaphandre / IPMI (system) last difference ratio — 64.001%

Comparison of different methods of measurement on a host with GPU

gpumad20

gpumad20 / Source: IPMI (instantaneous)

Name	Last *	Min	Max	Mean
gpumad20 - Sys Power power_unit (19.1)	230 W	220 W	285 W	226 W
gpumad20 - Sys Fan Pwr power_unit (19.4)	15 W	15 W	15 W	15 W
gpumad20 - PSU2 DC Out Pwr power_unit (19.6)	95 W	85 W	120 W	90.6 W
gpumad20 - PSU2 AC In Pwr power_unit (19.6)	100 W	90 W	125 W	96.2 W
gpumad20 - PSU1 DC Out Pwr power_unit (19.5)	115 W	110 W	145 W	117 W
gpumad20 - PSU1 AC In Pwr power_unit (19.5)	120 W	120 W	160 W	129 W
gpumad20 - Mem Power power_unit (19.3)	11 W	10 W	15 W	10.6 W
gpumad20 - CPU Power power_unit (19.2)	62 W	52 W	108 W	57.3 W

gpumad20 / Source: RAPL (average)

Name	Last *	Min	Max	Mean
RAPL total	66.935 W	63.892 W	90.355 W	67.471 W
RAPL package-0	29.978 W	29.568 W	42.037 W	30.468 W
RAPL package-1	26.419 W	23.502 W	37.365 W	26.323 W
RAPL dram-0	5.834 W	5.732 W	7.080 W	5.943 W
RAPL dram-1	4.704 W	4.630 W	6.173 W	4.745 W

gpumad20 / Source: Scaphandre (average)

Name	Last *	Min	Max	Mean
gpumad20 - scaph_host_power	67.451 W	63.874 W	90.344 W	67.467 W

gpumad20: IPMI (CPU+Mem) / IPMI (system) last difference ratio — 31.739%

gpumad20: RAPL / IPMI (system) last difference ratio — 29.322%

gpumad20: Scaphandre / IPMI (system) last difference ratio — 29.327%

Step 2:

Sharing with the cloud VM its own measurements

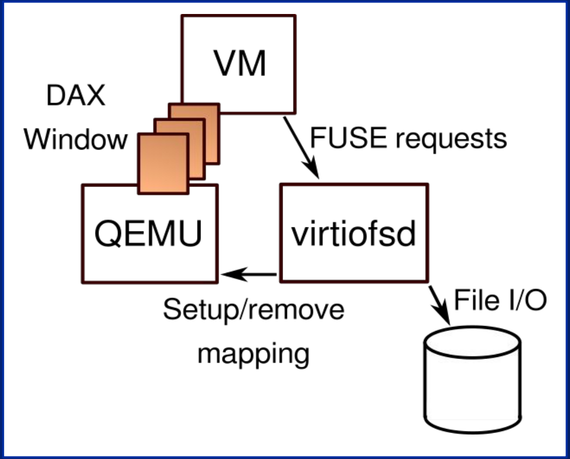
Not an easy thing

Original Scaphandre idea to share metrics

Use virtiofs to share the VM metrics through the hypervisor

Shared file system purposed for virtualization, it is specifically designed to take advantage of the locality of virtual machines and the hypervisor, using memory shared pages.

virtiofs



Version incompatibility with (our) current OpenStack Cloud

Its supported since v6.2.0 of libvirt.

- ➔ Now currently using v6.0.0 on OpenStack Nova 23.0.0
  - ◆ Minimum Nova version is 22.0.0

BUT!!

- ➔ Now currently using v6.0.0 on Ubuntu 20.04 on bare metal host.
  - ◆ Minimum Ubuntu version is 21.04

Lack of support for automation in OpenStack

A new feature in OpenStack Nova is required to provide virtiofs mounts.

Nova needs to manage the xml definition of the VM. Nova fully manages this file, and only nova can change it.

There are some nova-specs proposing changes to support this feature:

https://specs.openstack.org/openstack/nova-specs/specs/2023.1/approved/virtiofs-scaphandre.html

Step 3:
Measure workloads energy consumption

We are not still there

Measuring user and platform consumption

- Instrument user-level workloads to obtain direct consumption of user tasks.
  - Scaphandre (at VM level) + sidecar tasks (CodeCarbon, perun, BOAgent).
  - This requires resource provider cooperation to expose measures from the host.
- Platform-level metrics gathered automatically (WIP)
  - CPU/GPU efficiency + energy consumption
  - Hardware characteristics for multi-criteria impact assessment
  - Consider to also read data from resource provider side (IMPI)

Step 4:

Publish metrics and make users aware

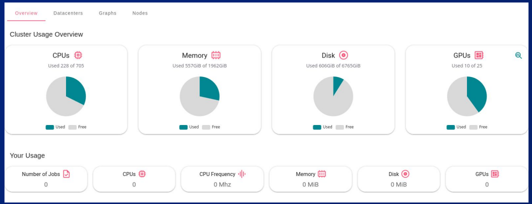
But we are starting to be here

Integration with external data sources

- (WIP) Integration with Spanish Power Grid and ENTSO-E to obtain real time data.
  - Sampling energy source, CO2 emissions (and other metrics).
- Integration of platform-level metrics (energy consumption) with power grid data to assess datacenter energy quality.
  - Automate energy quality metric
- Inform users about the impact of their deployments and prioritise *greener* datacenters at scheduler level

Metrics

- Integrate user-workload metrics into user deployment information and user statistics (real time data)
- AI4EOSC model (i.e. user facing) metadata is based on an open JSON Schema
  - https://github.com/ai4os/ai4-metadata
  - V2.X extension to include AI module impact
- AI4EOSC as an IT system metrics
  - No schema defined so far



Next steps:

Environmental Impact Evaluation

Moving ahead from simple CO_2 or energy consumption

Assessing environmental impact

Boavizta

- Energy consumption and carbon footprint (scope 1 and scope 2) are not enough and other more complex multi-criteria metrics are needed.
- Boavizta is an inter-organizational working group dedicated to evaluate the environmental impact of digital technologies across organizations.
    - Cover manufacturing and use (scope 2 and scope 3)
    - **GHG protocol** and **Life Cycle Assessment** methodology (ISO 14040/14044) are used as a reference to obtain the evaluation metrics.
- Limitations: Fine-grained hardware information (manufacturer, model), specialized hardware (GPU, Infibiband, etc) and other IT components (SAN, NAS, etc) are not yet covered.

Boavizta Tool Stack

Boaviztapi: An API to evaluate the environmental impacts of digital products and services based on their configuration and usage.

Boagent: A modular agent to evaluate the environmental impacts of a server or an application.

- Needs Scaphandre to get power consumption metrics.

Energizta: A collaborative project to collect and report open data on server energy consumption.

Environmental Impact Metrics

1. Greenhouse Gas emissions / Global Warming Potential
    a. Use (LCA) / scope 2 (GHG protocol)
    b. Manufacturing (LCA) / scope 3 (GHG protocol)

2. Primary energy usage: PE
    a. Use (LCA)
    b. Manufacturing (LCA)

3. Abiotic Resources Depletion / Abiotic Depletion Potential
    a. Use (LCA)
    b. Manufacturing (LCA)

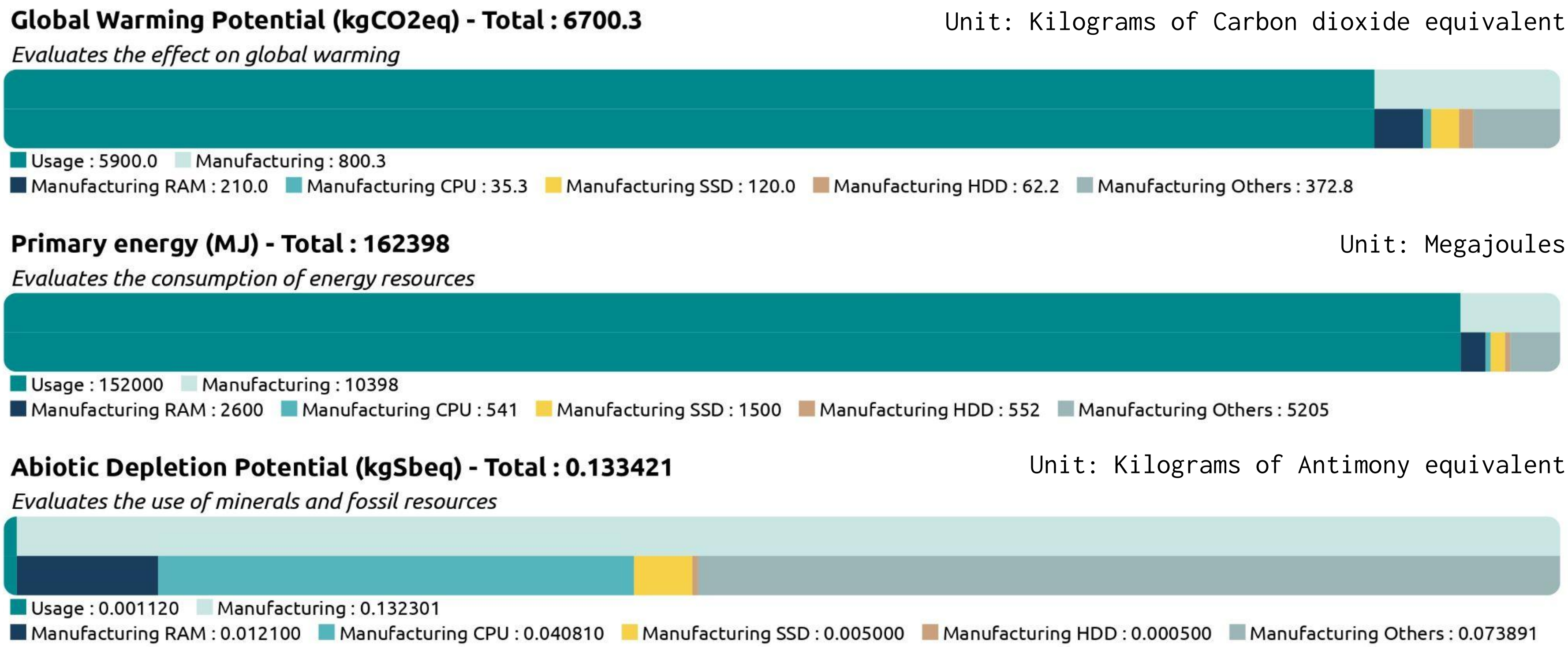


Figure: DataVizta Dashboard for Multicritera server impacts during a lifespan of 5 years and an average consumption of 150 Watts. Server is configured as 2 Skylake CPU, with 16 cores each, 150 Watts TDP, 4 Samsung DRAM DIMMs of 32GB each, 4 Micron SSD of 1TB each, 2 HHD, Rack Type and 2 PSU. Example at https://dataviz.boavizta.org/serversimpact

Wrap up

Some thoughts, as far as we have reach

Some thoughts

- Not possible to instrument the whole OpenStack cloud stack without cooperation from resource providers.
    - And they may not have the incentive to do so.
- CPU + Memory power consumption does not fully represent a VM's energy usage.
    - GPUs and other specialized hardware must be taken into account.
- CPU + Memory + GPU is not the complete server power consumption.
    - Even more cooperation from resource providers (i.e. IMPI) is needed to get the global picture.

Conclusions

- Energy consumption and PUE are not the only metrics to consider.
    - Consider other metrics, harder to measure (e.g. Water Usage Effectiveness), integrate with existing data sources for energy sources, etc.
- Energy consumption is the first step…
    - (and we are working on automating it).
- … but we need to move towards other complex multicriteria environmental impact metrics, integrating on-line data with existing databases in order to provide accurate metrics.
    - Global Warming Potential, Primary Energy, Abiotic Depletion potential.

References

- AI4EOSC Platform - Build AI models in the EOSC. https://ai4eosc.eu/platform/
- Tandon, Sonal. “Environmental Reporting Dashboards for OpenStack from BBC R&D.” *Superuser* (blog), February 9, 2022. https://superuser.openinfra.dev/articles/environmental-reporting-dashboards-for-openstack-from-bbc-rd/
- Scaphandre Docs. https://hubblo-org.github.io/scaphandre-documentation/
- Libvirt virt driver OS distribution support matrix. https://docs.openstack.org/nova/latest/reference/libvirt-distro-support-matrix.html
- NVIDIA Developer - DCGM Documentation. https://developer.nvidia.com/dcgm
- Boavizta. “Tools | Boavizta.” https://www.boavizta.org/en/tools

AI4EOSC - more information

AI4 | eosc

Demos & posters:

- Secure personalized federated learning within the AI4EOSC platform (Oct 2nd - 14:30)
- AI Inference Pipeline Composition with AI4Compose and OSCAR (Oct 2nd - 14:00)
- Poster: Analysis of Transitioning from Centralized Federated Learning to Decentralized Federated Learning: A Case Study on Thermal Anomalies Detection using UAV-Based Imaging
- Poster: Using AI4EOSC platform for integrated plant protection use case

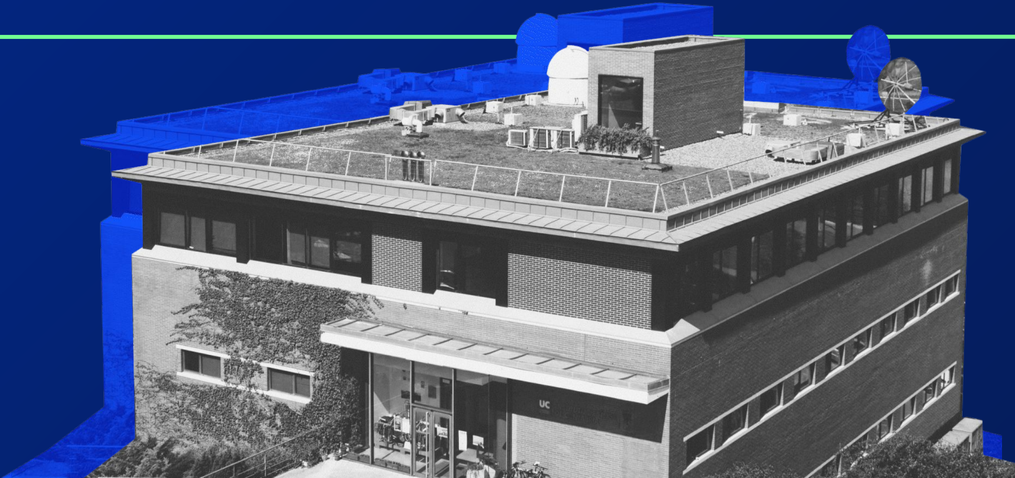
Session: Processing Research Data with AI and ML (Oct 3rd - 09:00 to 12:30)

- iMagine: an AI platform supporting aquatic science use cases (Oct 3rd - 09:10)
- AI4EOSC as a toolbox to develop and serve AI models in the EOSC (Oct 3rd - 11:00)
- MLOps: from global landscape to practice in AI4EOSC (Oct 3rd - 11:15)
- Leveraging MLflow for Efficient Evaluation and Deployment of LLMs (Oct 3rd - 11:30)
- Comparative Study of Federated Learning Frameworks NVFlare and Flower for Detecting Thermal Anomalies in Urban Environments (Oct 3rd - 11:45)
- Distributed computing platform on EGI Federated Cloud (Oct 3rd - 12:00)

Instituto de Física de Cantabria | CSIC Consejo Superior de Investigaciones Científicas | UC Universidad de Cantabria

GreenDIGIT | All Hands Meeting
EGI Conference 2024
October 1st 2024

Thanks for your attention



Advanced Computing and e-Science Group
https://advancedcomputing.ifca.es

Backup slides

Step 1.1:

Measuring VM Power Consumption

Scaphandre Agent

RAPL Domains

RAPL stands for *"Running Average Power Limit"*, it is a feature on Intel/AMD x86 CPU's that allows to set limits on power used by the CPU and other components.

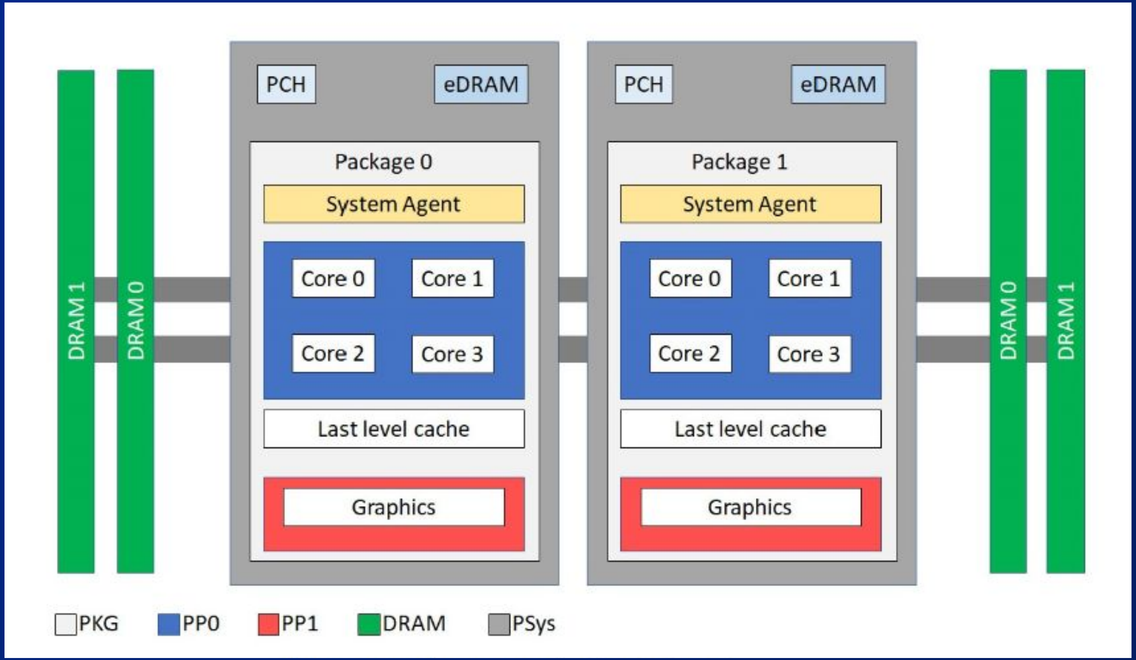


Figure: Power domains considered in RAPL interface.
Source: https://github.com/bpetit/awesome-energy#rapl

Interesting metrics exposed

- **scaph_host_power_microwatts**: Aggregation of several measurements to give a try on the power usage of the the whole host, in microwatts. It might be the same as **RAPL PSYS** or a combination of **RAPL PKG** and **DRAM domains**.
- **scaph_process_power_consumption_microwatts**: Power consumption per process, in microwatts.
  - Metadata:
    - ***exe***: is the name of the **executable** that is the origin of that process.
      - In our case the exe to track is: */usr/bin/qemu-system-x86_64*
    - ***cmdline***: whole command line with the executable path and its parameters.
      - Some relevant parameters: **vmname** and **vm-uuid**
    - ***instance***: label to filter the metrics by the host.
    - ***pid***: is the process **id**

Design & Implementation

AI4EOSC bottom up solution

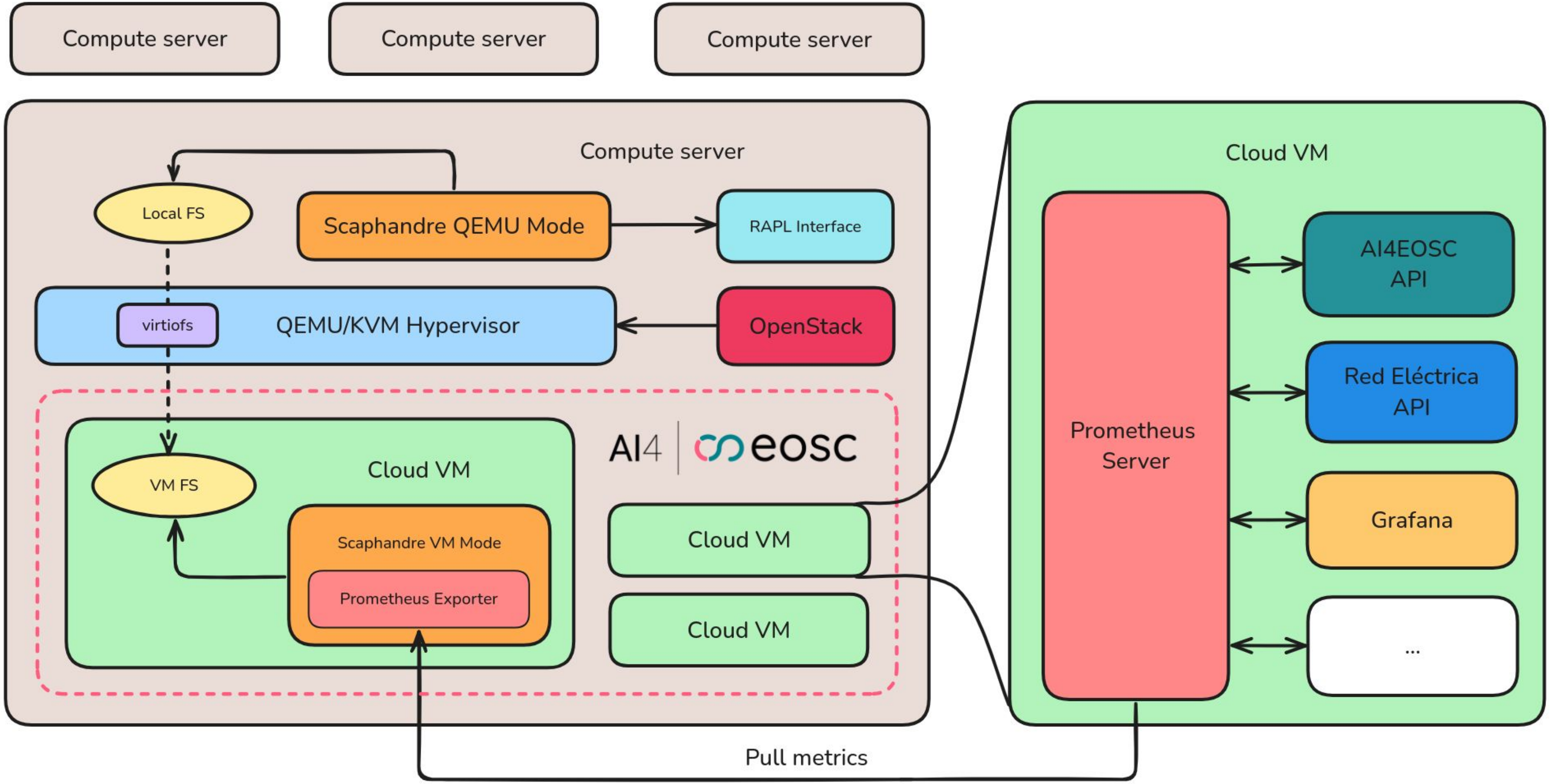
Open Source Tool Stack

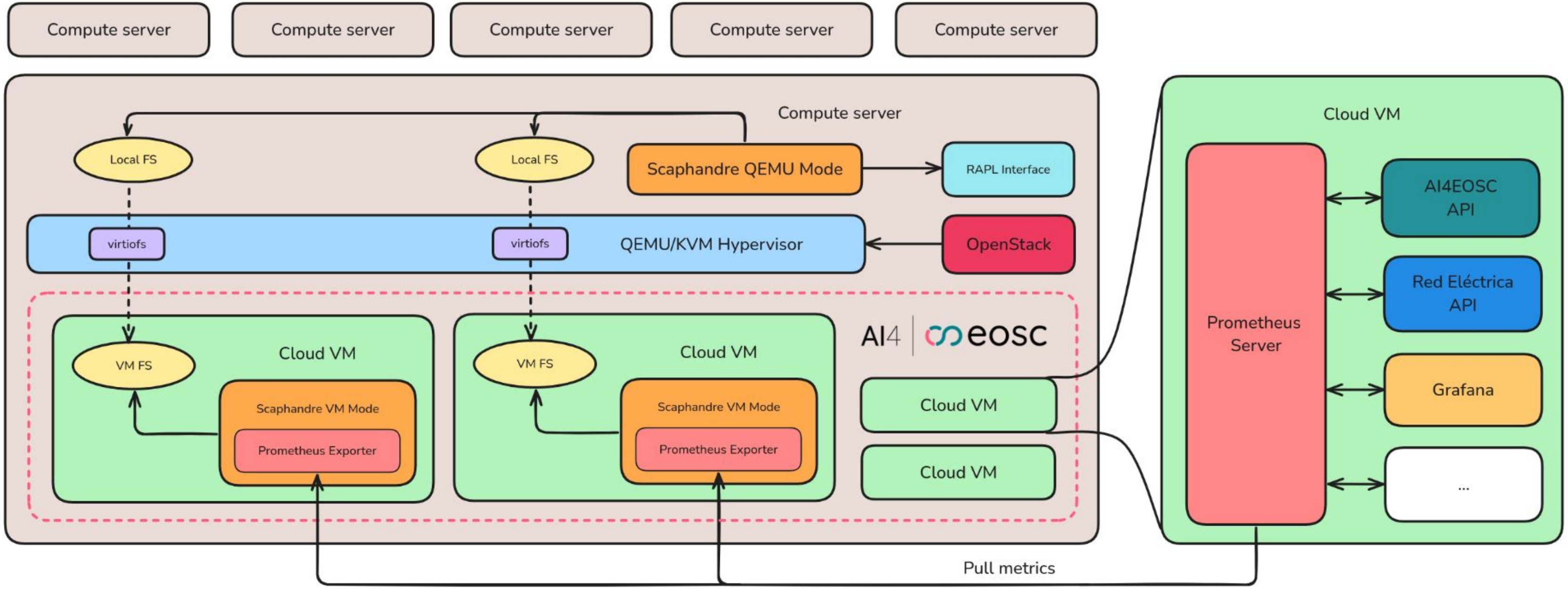
Scaphandre Agent	Prometheus Server	Grafana Server	AI4EOSC API & Dashboard	Power Grid API
Scrape power consumption metrics	Store time series data	Visualize data	Visualize data	Get quality of energy

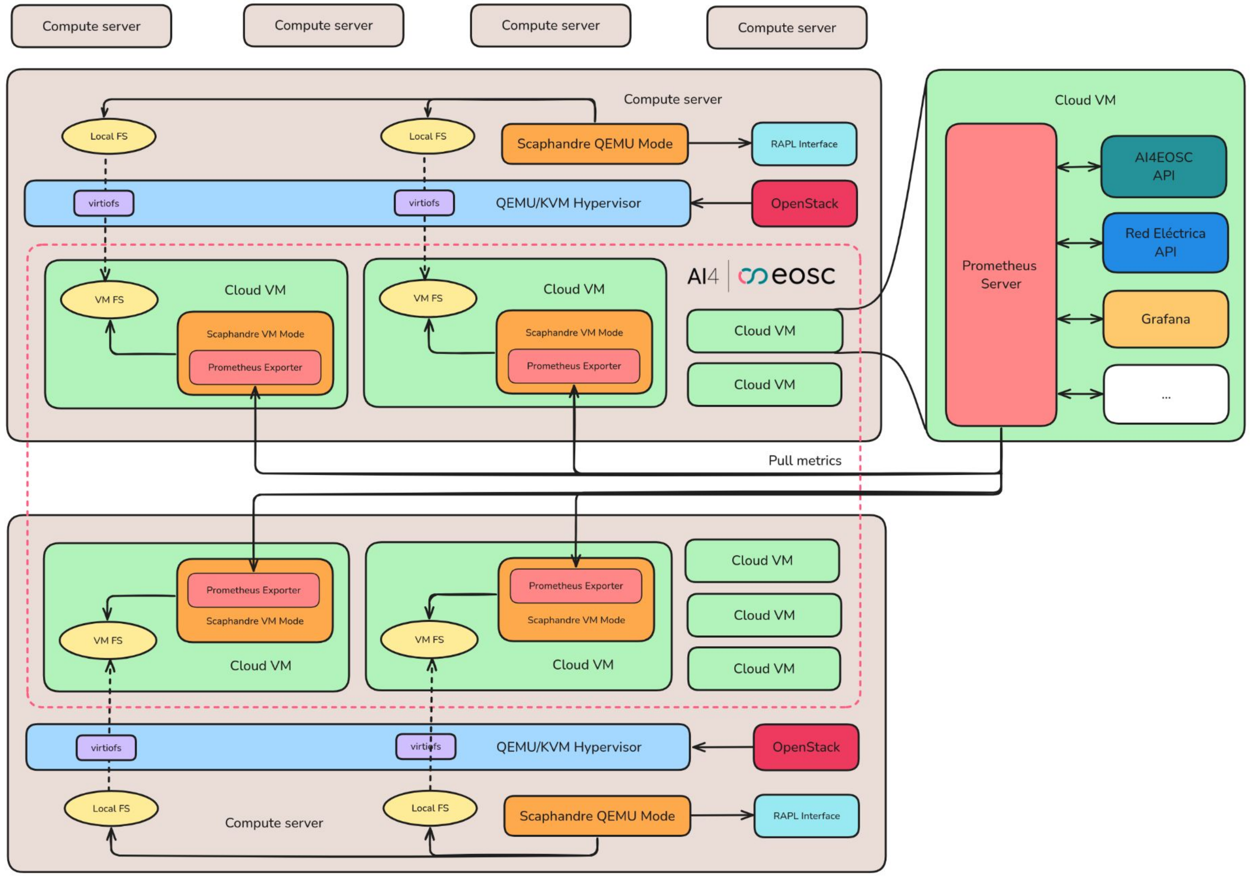
AI4 eosc

red eléctrica
Una empresa de Redeia

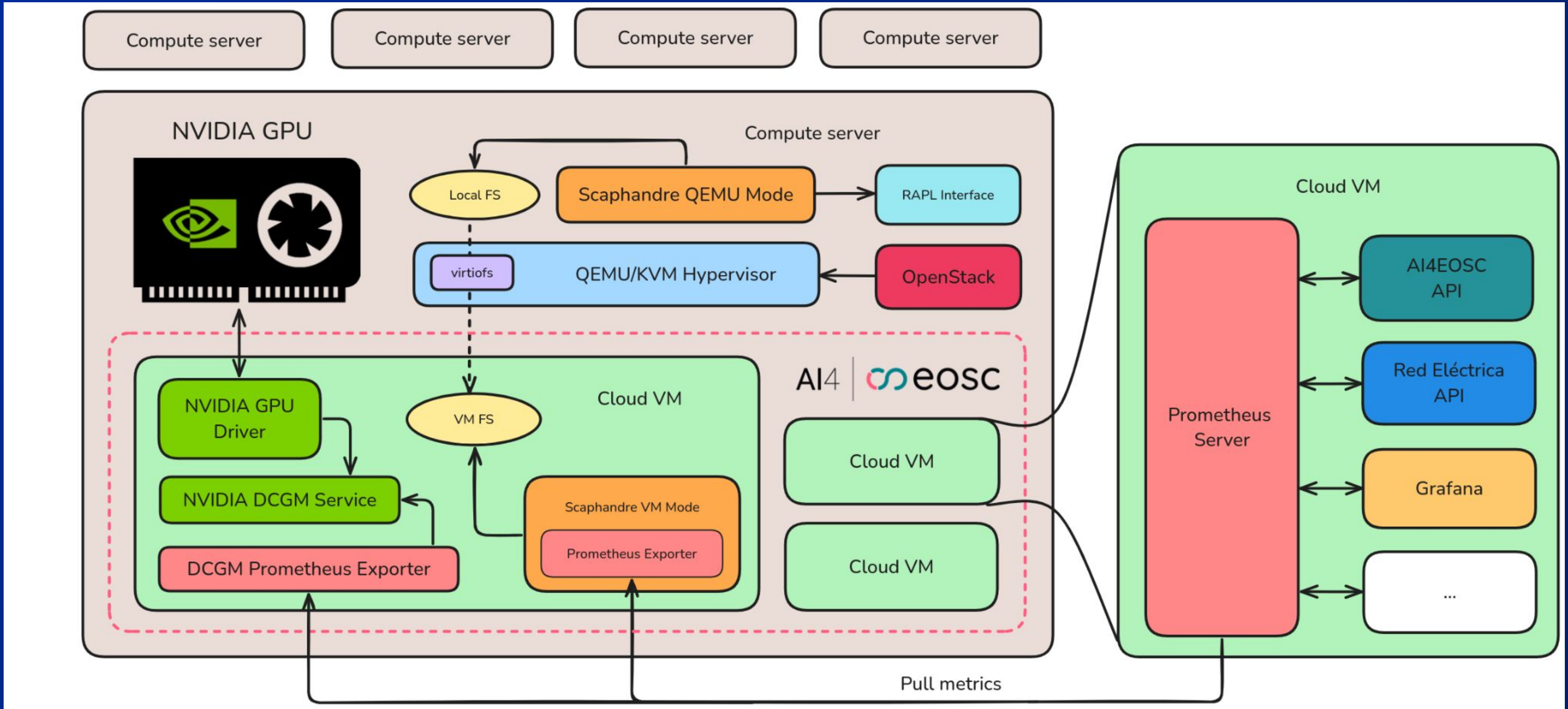
entsoe


Compute server
Compute server
Compute server
Compute server
Local FS
Scaphandre QEMU Mode
RAPL Interface
virtiofs
QEMU/KVM Hypervisor
OpenStack
AI4 eosc
Cloud VM
VM FS
Scaphandre VM Mode
Prometheus Exporter
Cloud VM
Cloud VM
Cloud VM
Prometheus Server
AI4EOSC API
Red Eléctrica API
Grafana
...
Pull metrics


Compute server
Compute server
Compute server
Compute server
Compute server
Compute server
Local FS
Local FS
Scaphandre QEMU Mode
RAPL Interface
virtiofs
virtiofs
QEMU/KVM Hypervisor
OpenStack
Cloud VM
VM FS
Scaphandre VM Mode
Prometheus Exporter
Cloud VM
VM FS
Scaphandre VM Mode
Prometheus Exporter
AI4 eosc
Cloud VM
Cloud VM
Cloud VM
Prometheus Server
AI4EOSC API
Red Eléctrica API
Grafana
...
Pull metrics


Compute server
Compute server
Compute server
Compute server
Compute server
Local FS
Local FS
Scaphandre QEMU Mode
RAPL Interface
virtiofs
virtiofs
QEMU/KVM Hypervisor
OpenStack
AI4 eosc
VM FS
Cloud VM
Scaphandre VM Mode
Prometheus Exporter
VM FS
Cloud VM
Scaphandre VM Mode
Prometheus Exporter
Cloud VM
Cloud VM
Cloud VM
Prometheus Server
AI4EOSC API
Red Eléctrica API
Grafana
...
Pull metrics
Prometheus Exporter
Scaphandre VM Mode
VM FS
Cloud VM
Prometheus Exporter
Scaphandre VM Mode
VM FS
Cloud VM
Cloud VM
Cloud VM
Cloud VM
virtiofs
virtiofs
QEMU/KVM Hypervisor
OpenStack
Local FS
Compute server
Local FS
Scaphandre QEMU Mode
RAPL Interface

NVIDIA Data Center GPU Manager (DCGM)



Grafana GPU real consumption

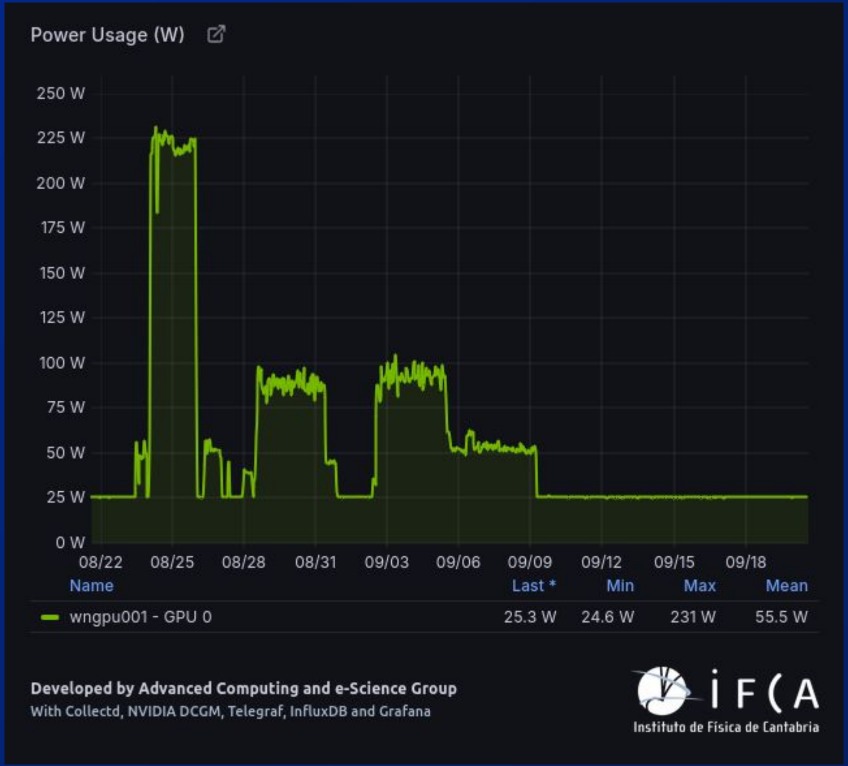


Figure: Chart of the IFCA's Grafana Dashboard to represent the power consumption in watts of each of the GPU as a time series.