EAS

Energy Aware Runtime for Sustainable Data Centers

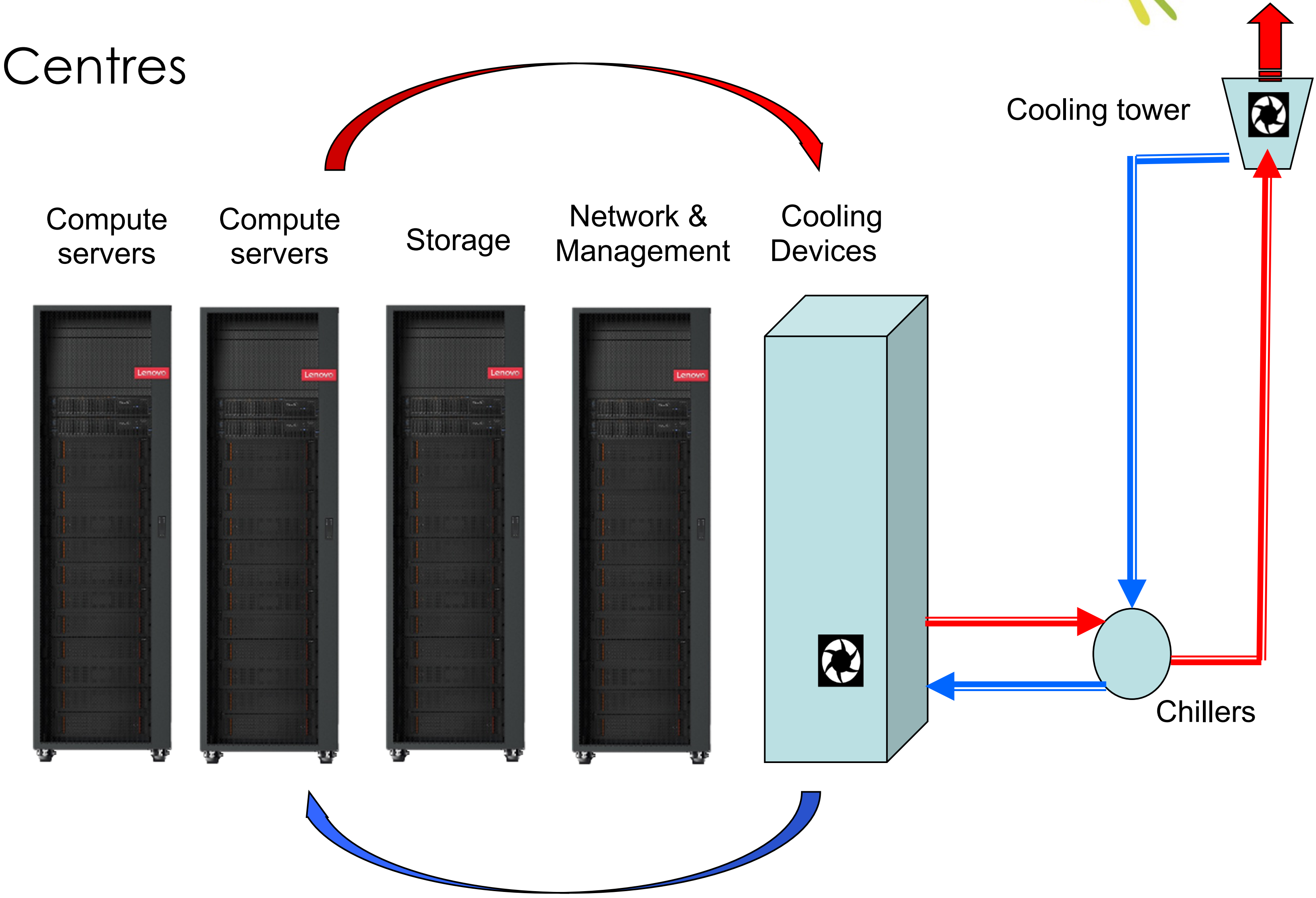
Energy Efficient Data Centers

Energy Consumption in Data Centres

- High Energy & CO2 footprint
- IT Equipments & Cooling
- Improving Energy Efficiency

Total Energy = IT Equipments Energy x PUE

  - Reducing the PUE
    - Optimizing airflow
    - Advanced cooling
  - Optimizing IT Energy
    - Hardware consolidation
    - **EAR**

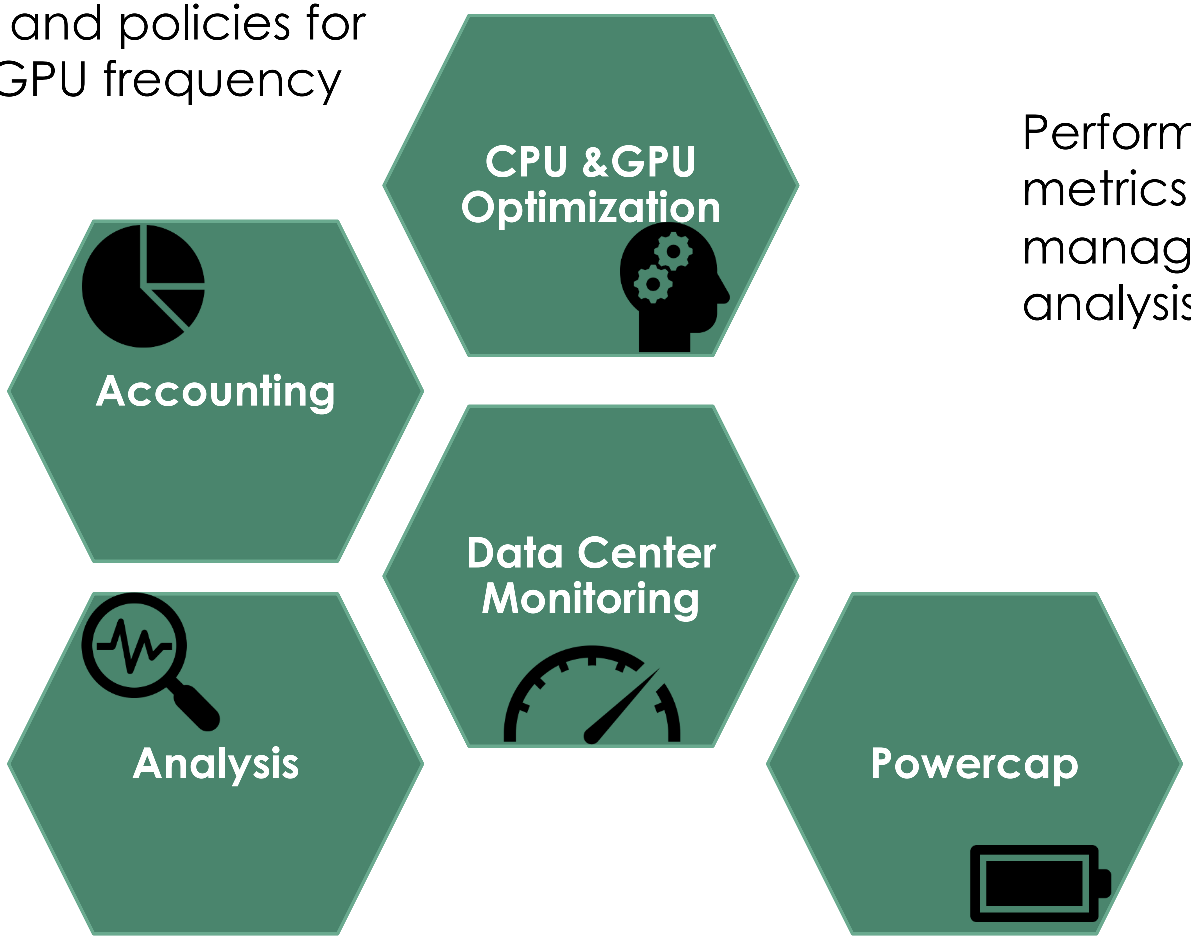


Data Center Energy Components

EAR provides….

Energy models and policies for CPU/Memory/GPU frequency selection

Data reporting for accounting, billing, and system analytics



CPU &GPU Optimization

Accounting

Data Center Monitoring

Analysis

Powercap

Performance and Power metrics for system management and job analysis

Analysis and classification of job metrics for energy and power optimization

Power control to guarantee data center operational limits

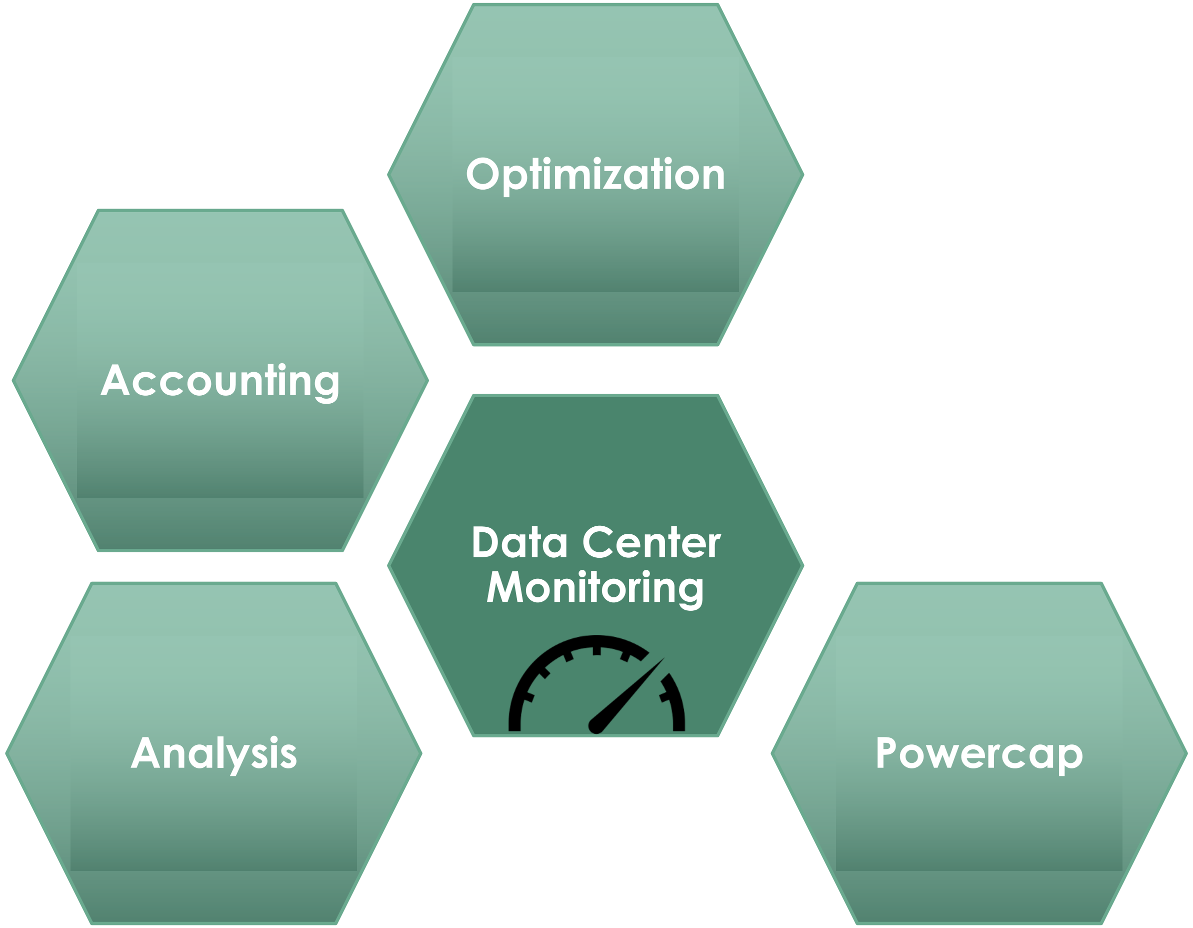
EAR Version 5

- Introduce **Energy Optimization** for **GPUs** running **AI and HPC workloads**
- **Full** Data Center **monitoring: from Compute servers to Network and Storage**
- Support **Workflows** on top of Jobs
- First Implementation of **European Power stack API** (Regale)

EAR successful installations

- EAS/EAR major installations
  - **LRZ Germany**, (**SuperMUC-NG** since 2019 and **SuperMUC-NG2**)
    - Phase 1: Lenovo 6700 2 x Intel Xeon Platinum 8174 24C 3.1GHz
    - Phase 2: Lenovo 240 nodes with 2 x Intel Sapphire Rapid + 4 Intel Ponte Vecchio
  - **SURF Netherland, Snellius**
    - CPU partition: Lenovo 500 nodes AMD Rome and 786 nodes AMD GENOA
    - GPU partition: Lenovo 72 nodes 2x intel Icelake + 4 NVIDIA GPU A100
  - **BSC Spain, (MN5)** on both GPP and ACC partitions (2023/2024)
    - GPP partition: Lenovo 7200 nodes with Intel Sapphire Rapid
    - ACC partition: BullSequana XH3000, 1110 nodes with 2x Intel Sapphire Rapid+ 4 NVIDIA H100
  - POC underway at **EDF France (Cronos)**
    - CRONOS - BullSequana X, 1995 nodes with 2xXeon Platinum 8260 24C 2.4GHz

EAR provides….Monitoring



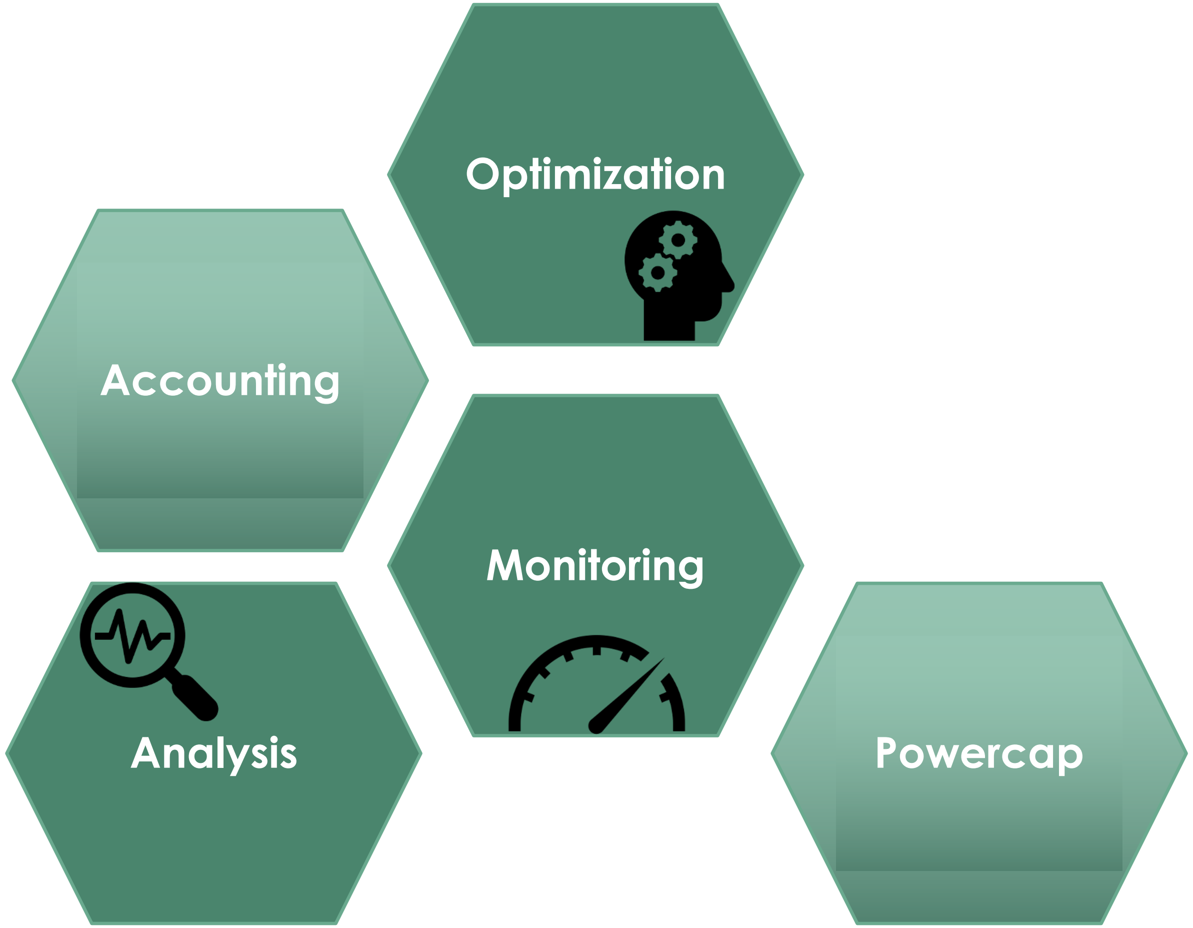
Monitoring

- **Job Monitoring**
  - Powerful non-intrusive application **monitoring**
  - **100% dynamic, no code modifications**
  - **Runtime signatures**:
    - Performance: Time, CPI, Memory Bandwidth (GB/sec), Gflops, IO MB/sec, MPI activity, GPU utilization, GPU Memory utilization, …
    - Power metrics : Node, CPU, DRAM, GPU

- **Computational nodes Monitoring**
  - Extensible **monitoring** : Power, CPU frequency, temperature, etc
  - Multiple sources of data: inband IPMI, GPU, RAPL…
  - Intel, AMD, NVIDIA
  - Extensible **report** : MariaDB, Postgres, Sysfs, Prometheus (wip),…
  - Basic **alerts** for power and temperature

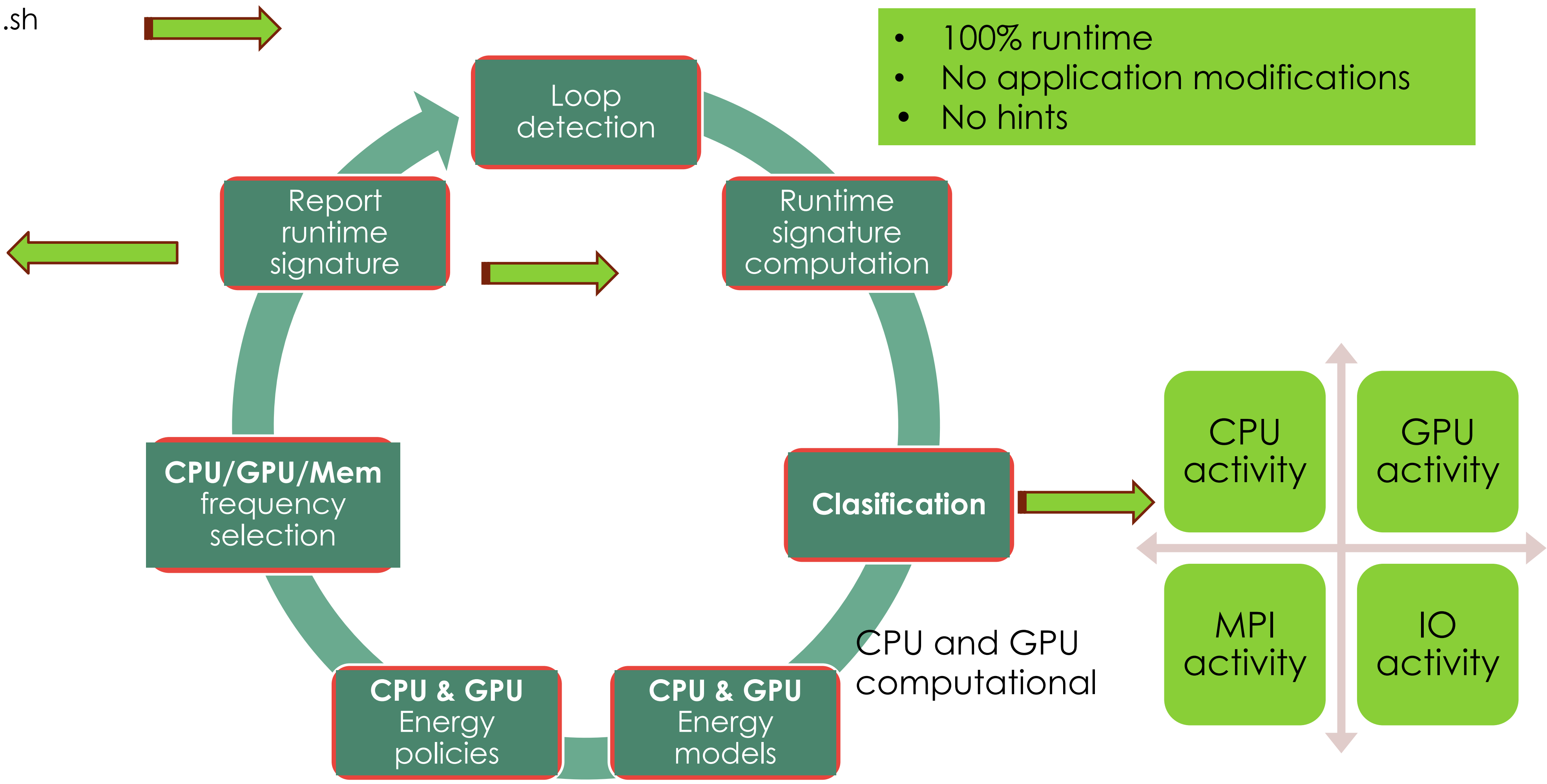
- **Data Center monitoring**
  - AC power for compute, storage and network
  - Report to DB
  - Possible integration with EAR powercap service

EAR provides(among others)….Runtime analysis and optimization



The optimization loop

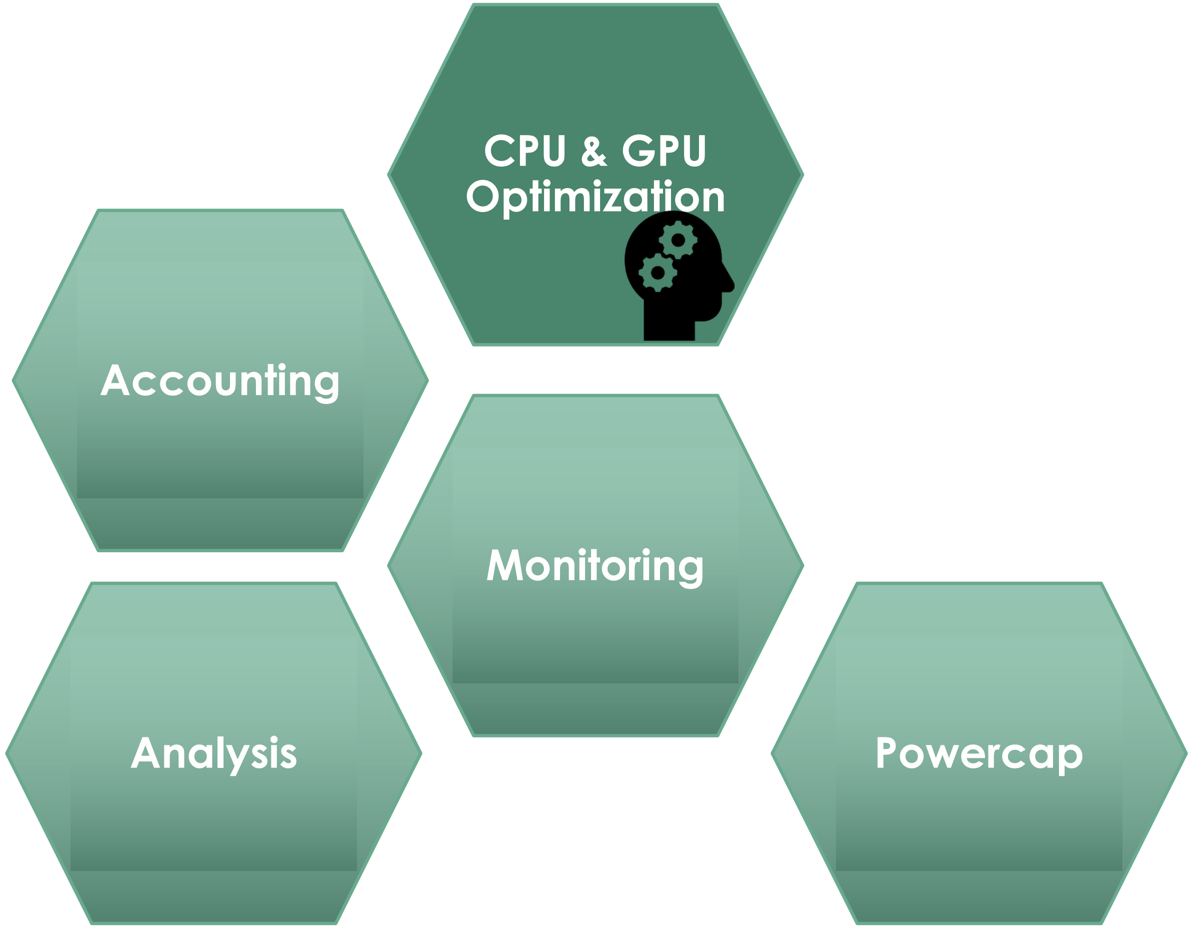
sbatch myapp.sh



- 100% runtime
- No application modifications
- No hints

CPU and GPU computational

EAR provides….Optimization



CPU & GPU Optimization

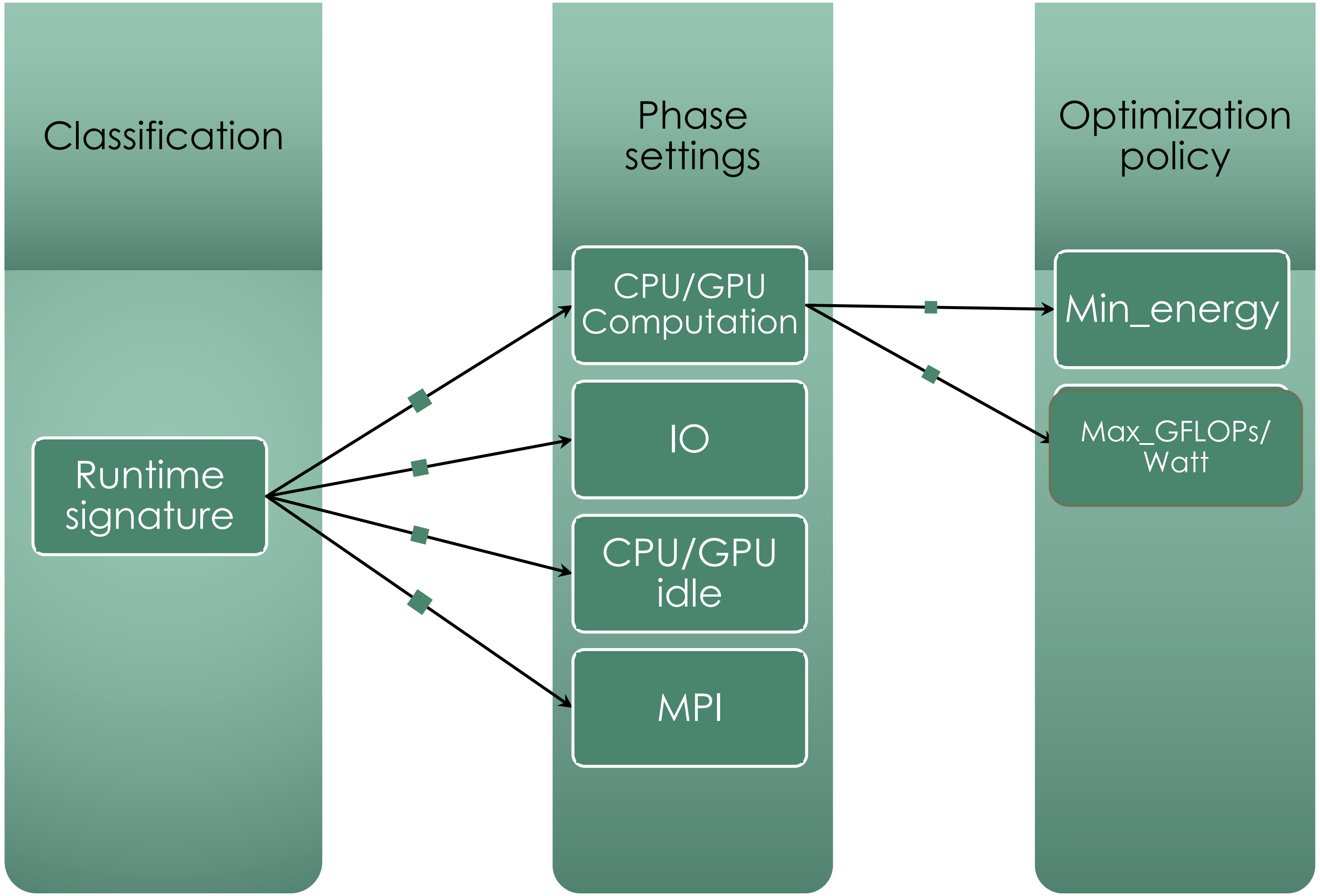
Accounting

Monitoring

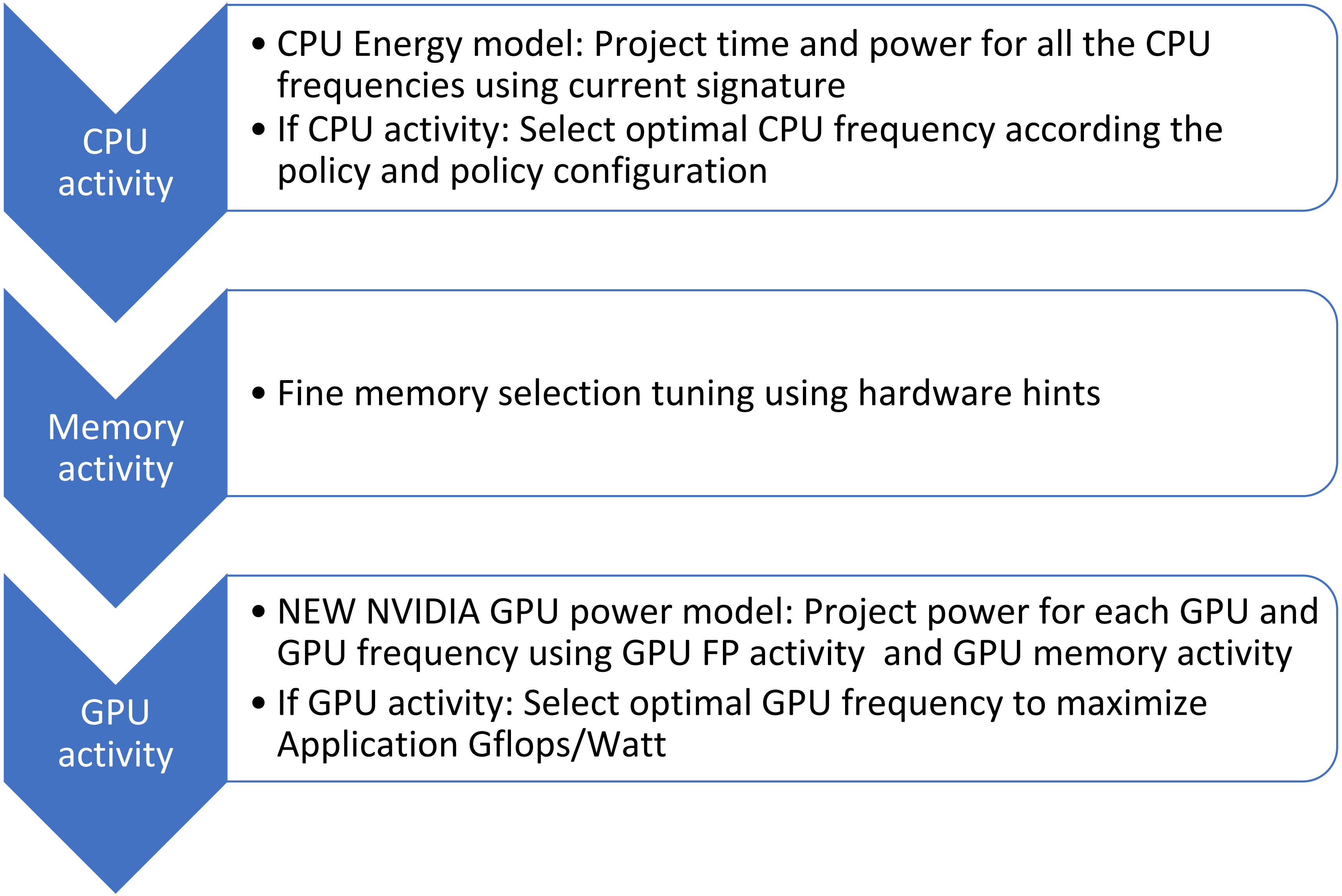
Analysis

Powercap

Optimization



Energy optimization for computational phases



CPU activity

- CPU Energy model: Project time and power for all the CPU frequencies using current signature
- If CPU activity: Select optimal CPU frequency according the policy and policy configuration

Memory activity

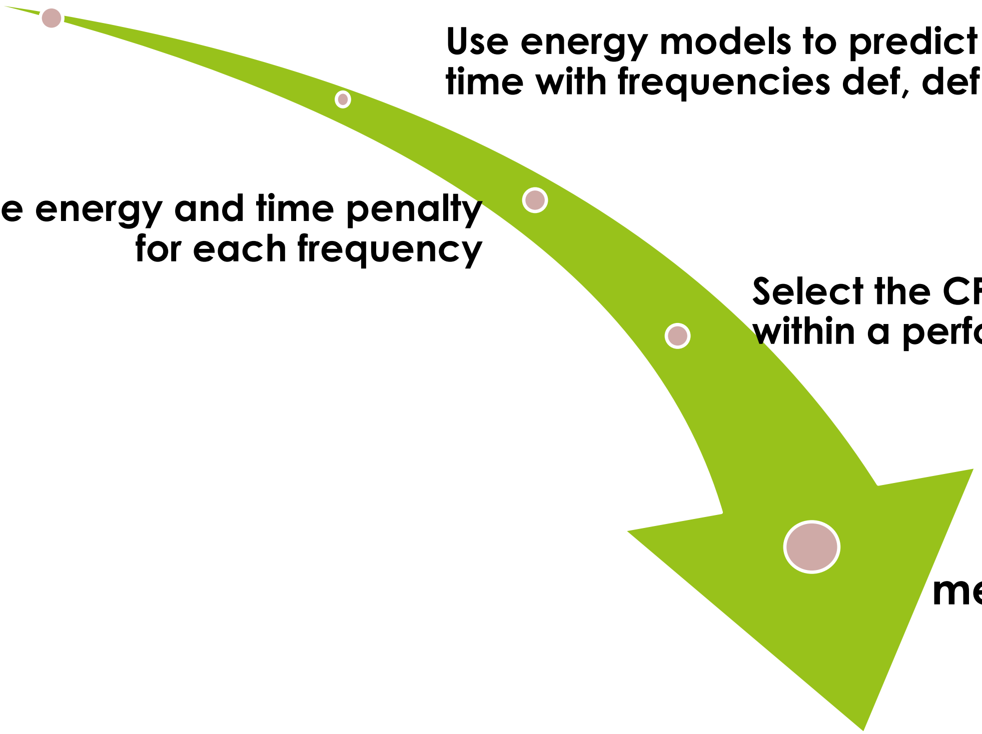
- Fine memory selection tuning using hardware hints

GPU activity

- NEW NVIDIA GPU power model: Project power for each GPU and GPU frequency using GPU FP activity and GPU memory activity
- If GPU activity: Select optimal GPU frequency to maximize Application Gflops/Watt

CPU Optimization: Minimize energy to solution

Save energy by reducing CPU frequency (DVFS)



Execute one "iteration" at nominal frequency and compute runtime signature

Use energy models to predict power and time with frequencies def, def-1, def-2…

Compute energy and time penalty for each frequency

Select the CPU frequency minimizing energy within a performance penalty limit

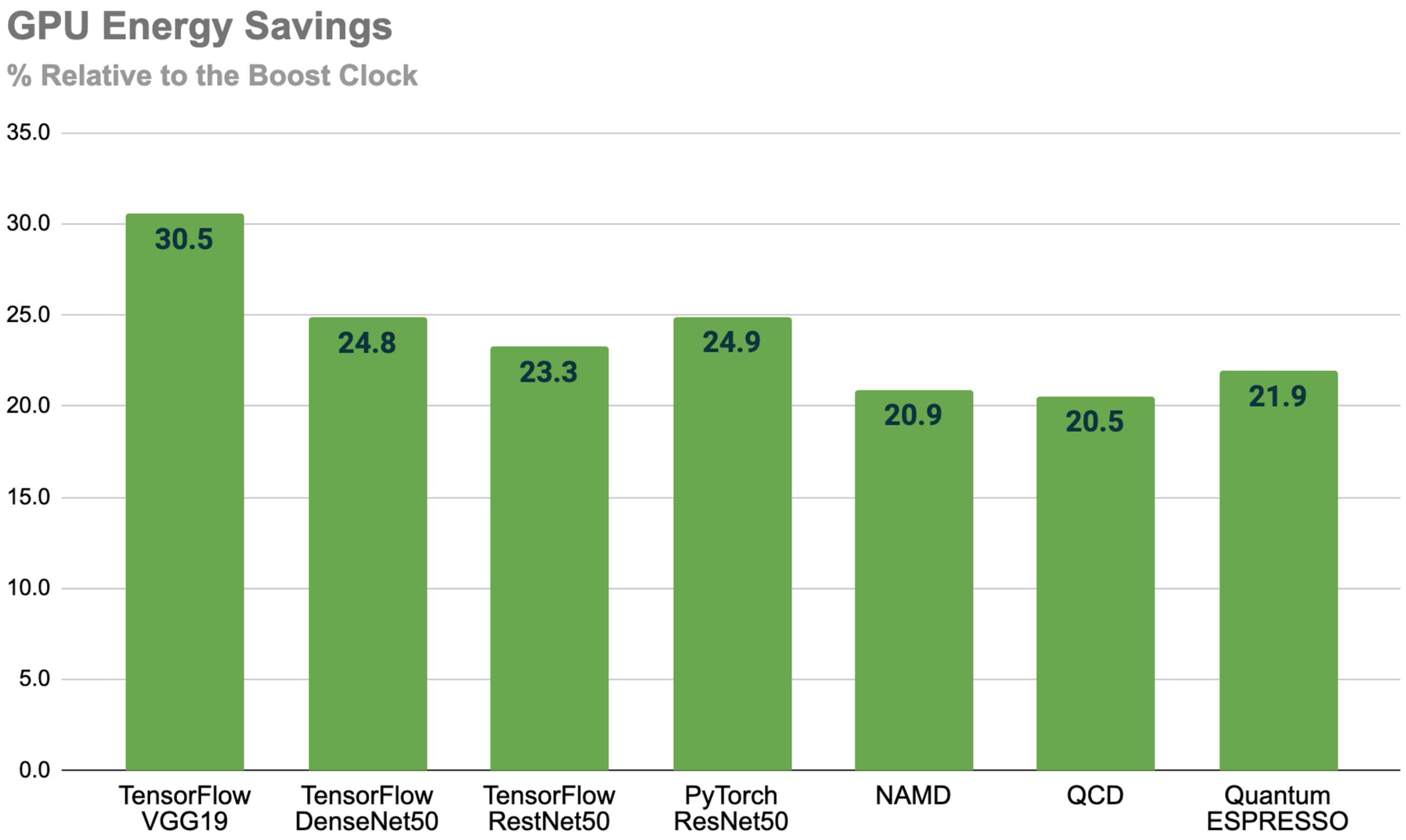
Select memory/GPU frequency

NVIDIA GPU optimization with EAR

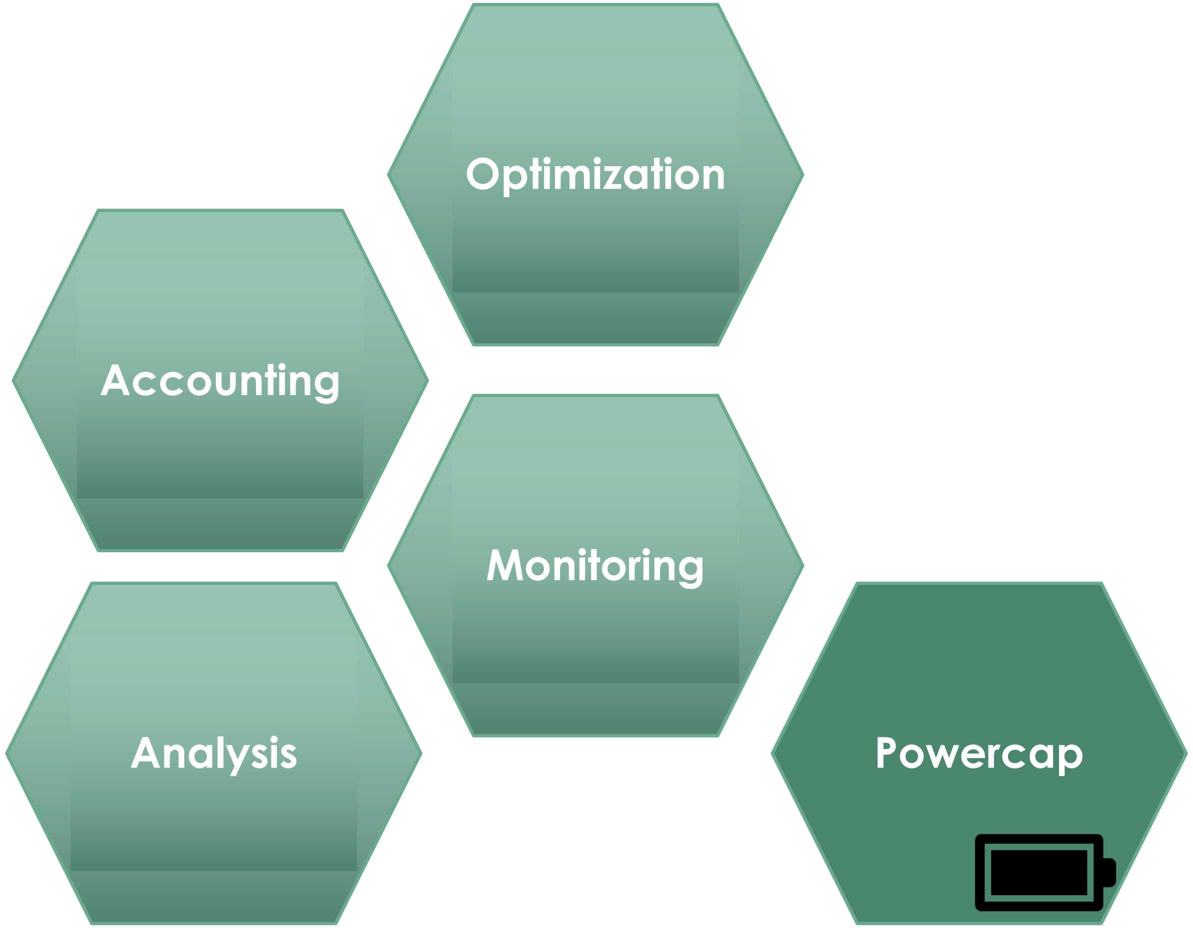
- Extended GPU metrics + GPU power model + GPU optimization policy
- GPU metrics
  - Based on DCGMI/NVML
  - **Performance counters + activity ratios**
  - **More semantics than just utiization**
- GPU power model
  - Floating Point activity characterize the utilization of FP and tensor instructions
  - DRAM activity characterize GPU memory utilization
- GPU optimization policy
  - GPU signature computed at runtime
  - Power projections for all the GPU frequencies (per-GPU)
  - **Optimization metric** computed:
    - **CPU+GPU GFlops/Node power (W)**
  - **Optimization function**: Max

GPU energy savings on AI & HPC workloads

- Energy = Power x Time
- Evaluation computed in 2 x Icelake + 4 x NVIDIA A100 (Snellius cluster)



EAR provides….Powercap



Optimization

Accounting

Monitoring

Analysis

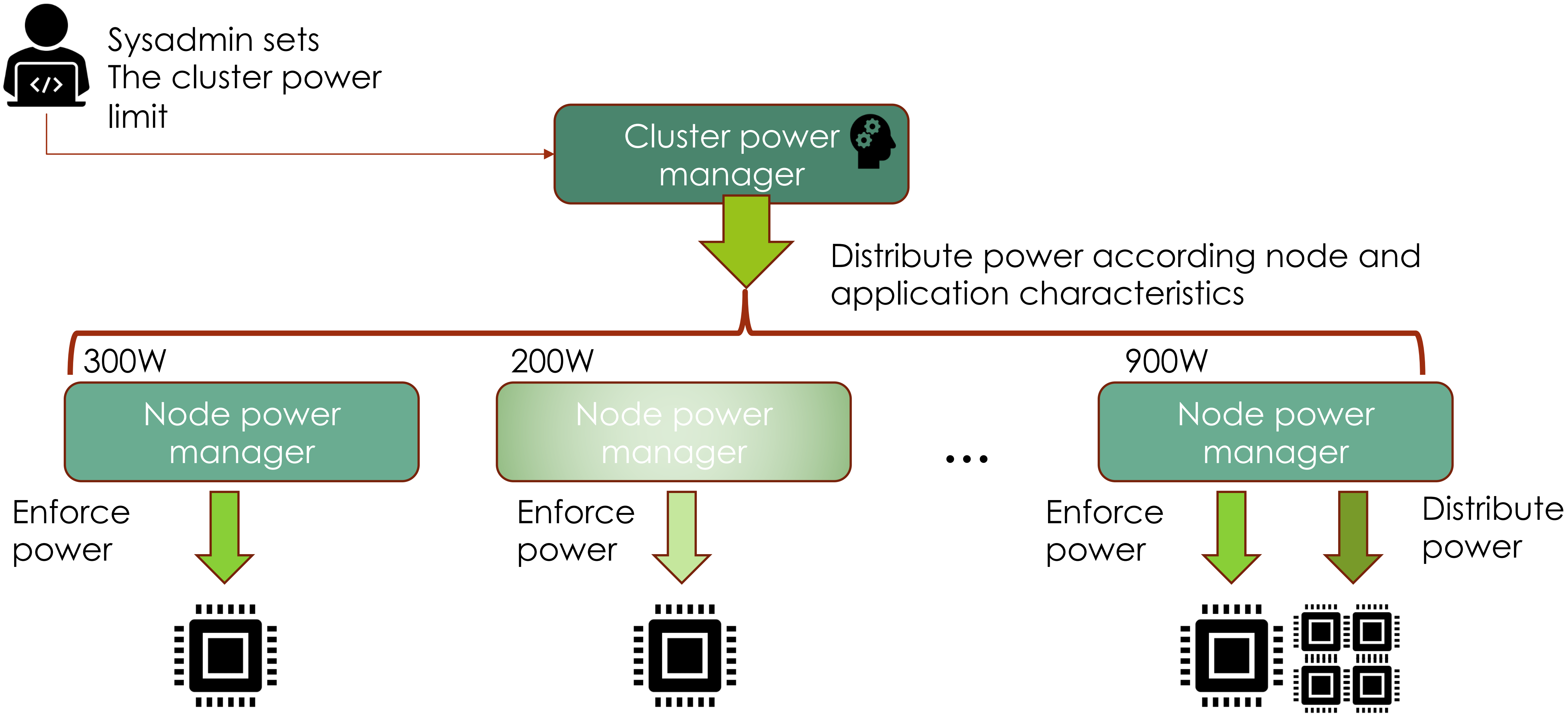
Powercap

EAR powercap summary

- **EAR Node powercap manager enforces node power limit**
    - Extensible through plugins: CPU, GPU
    - Dynamic intra-node power re-allocation based on application activity
- **Cluster power manager distributes power to computational nodes**
    - Hierarchical architecture for large scale clusters
    - Two algorithms offered: soft and hard powercap
- **EAR runtime library informs the EAR node powercap manager of application activity and power requirements**

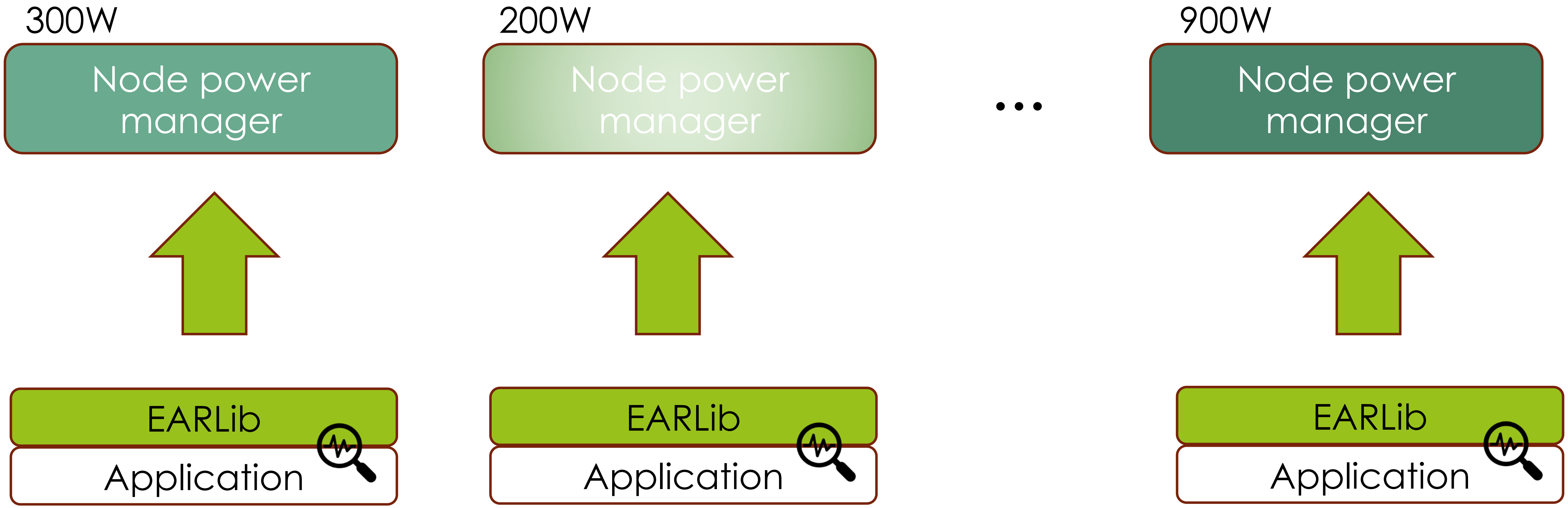
Powercap (I): Initial distribution

Cluster power manager distributes power and node power manager enforces power



Powercap(II): Application feedback

Application (through EARlib) informs each node about its power needs



300W

Node power manager

EARLib

Application

200W

Node power manager

EARLib

Application

…

900W

Node power manager

EARLib

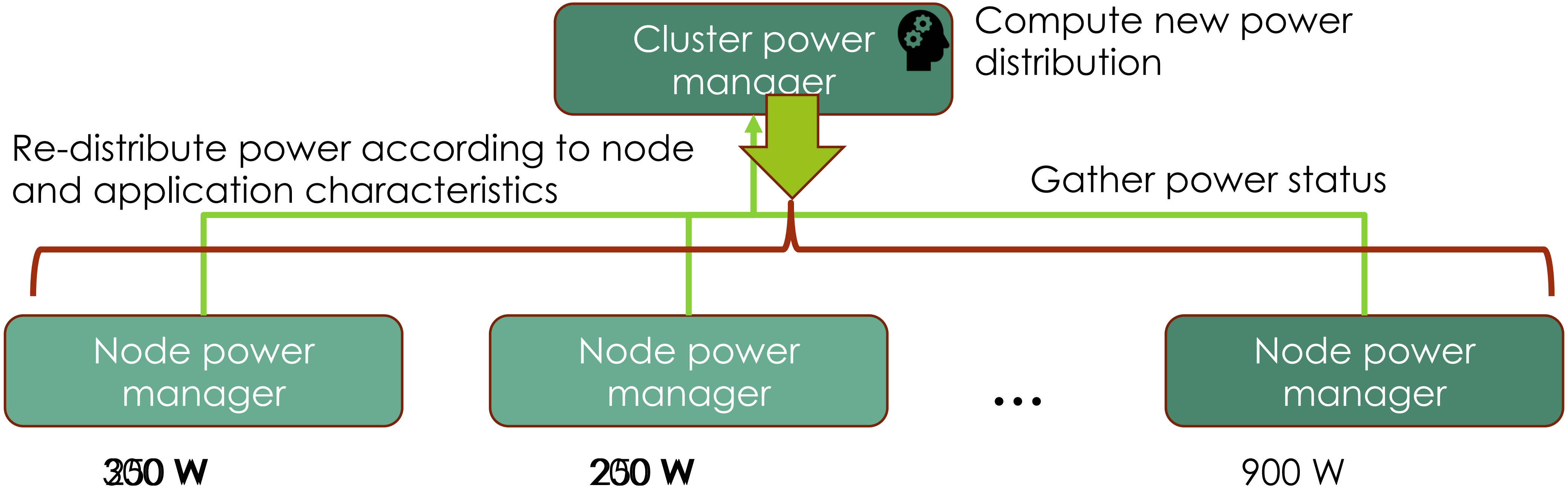
Application

I can release power because I don’t need so much power

I need more power because I will be more efficient

I need more power because I will be more efficient

Powercap(III): Dynamic power reallocation



Cluster power manager

Compute new power distribution

Re-distribute power according to node and application characteristics

Gather power status

Node power manager

Node power manager

...

Node power manager

300 W

250 W

200 W

250 W

900 W

EAR provides….Accounting

Optimization

Accounting

Monitoring

Analysis

Powercap

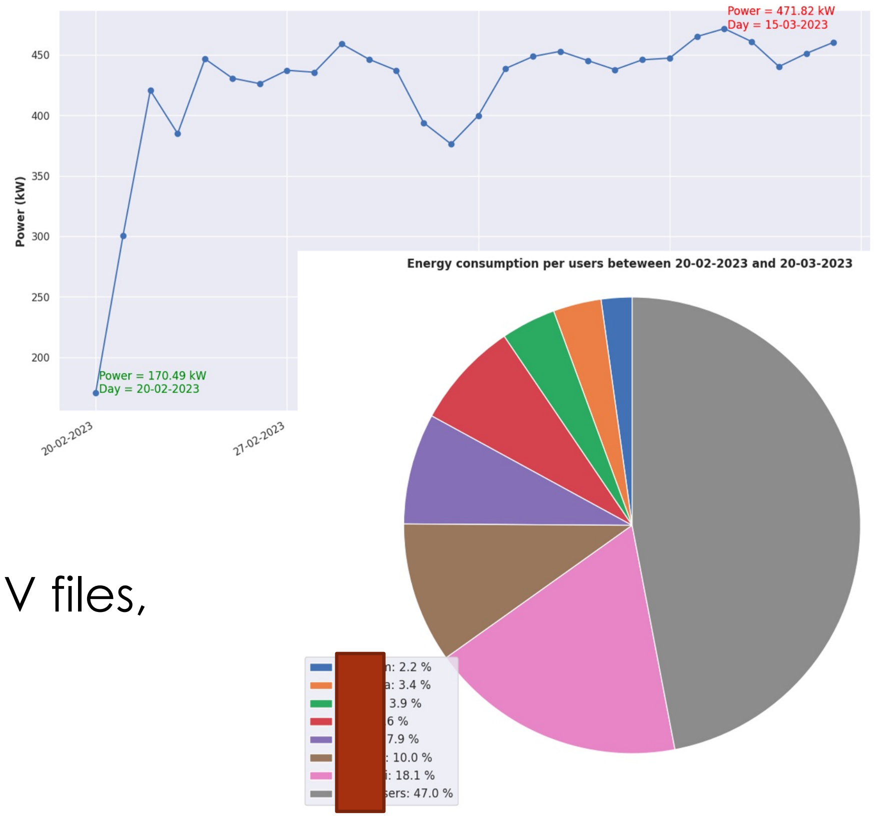
Data analysis with ear-system-analytics and ear-job-analytics

EAR reports job metrics and system Telemetry through plugins

Multiple plugins can be loaded at the same time

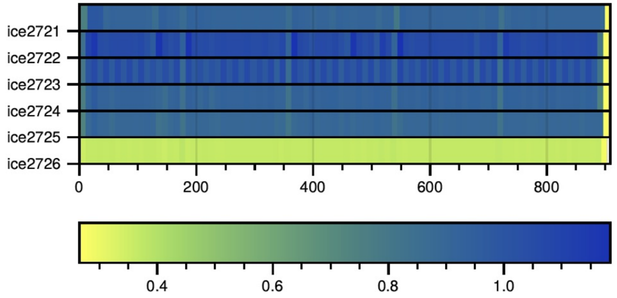
Plugins included by default: DB, CSV files, Paraver traces, Prometheus (WIP), etc

Average daily power



Energy consumption per user

Cycles per Instruction



Memory bandwith(GB/s)

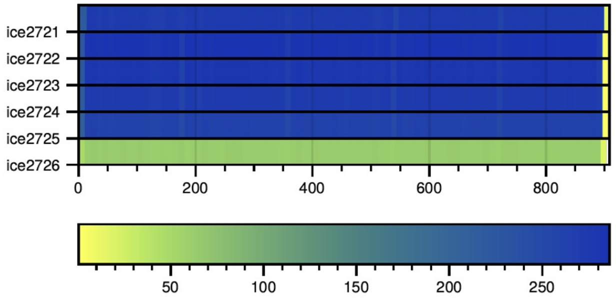


Figure 7: Memory bandwidth (GB/s) over the time.

ear-system-analytics

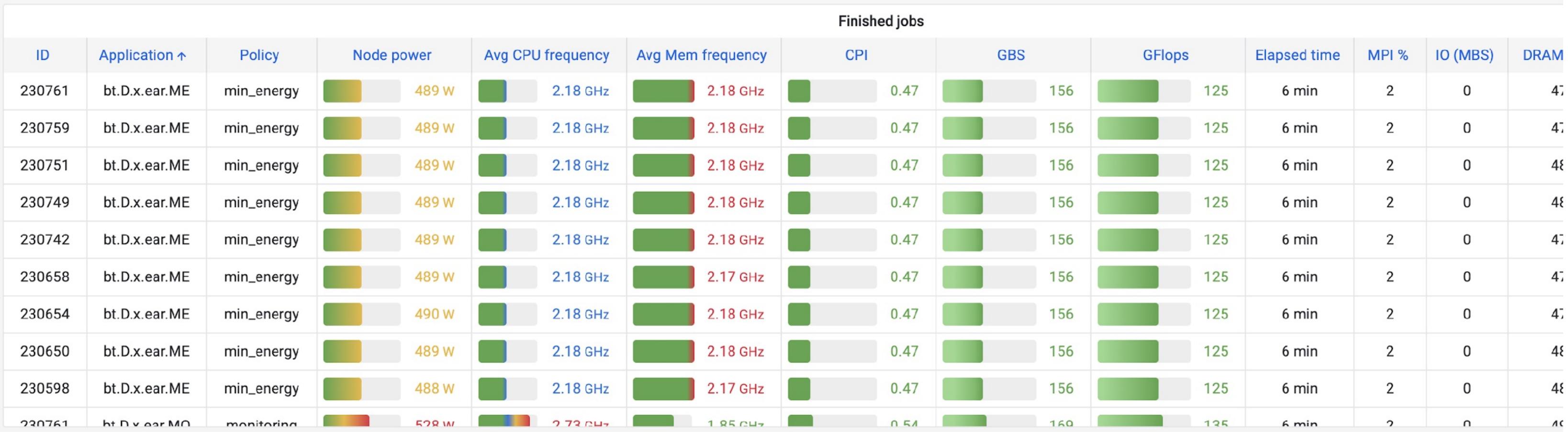
ear-job-analytics

Valuable system and Workload statistics can be computed using EAR data: power over the time, energy per user, job performance & power characteristics, etc

Data visualization

Jobs data can be seen with tools such as Grafana

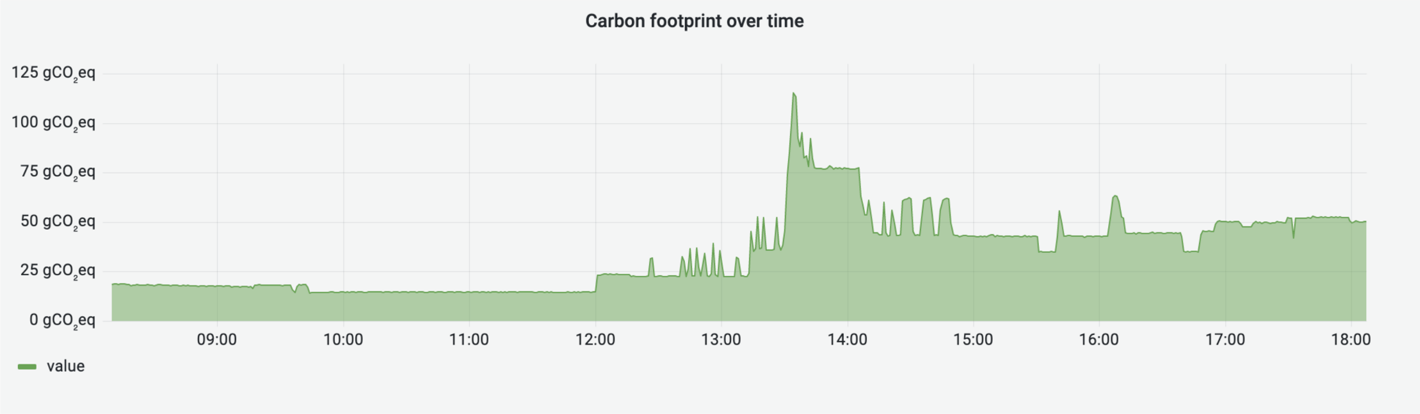
Job metrics: CPU frequency, CPI, Memory bandwith, Gflops, etc



Finished jobs

ID	Application ↑	Policy	Node power	Avg CPU frequency	Avg Mem frequency	CPI	GBS	GFlops	Elapsed time	MPI %	IO (MBS)	DRAM
230761	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	47
230759	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	47
230751	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	48
230749	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	48
230742	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	47
230658	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.17 GHz	0.47	156	125	6 min	2	0	47
230654	bt.D.x.ear.ME	min_energy	490 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	47
230650	bt.D.x.ear.ME	min_energy	489 W	2.18 GHz	2.18 GHz	0.47	156	125	6 min	2	0	48
230598	bt.D.x.ear.ME	min_energy	488 W	2.18 GHz	2.17 GHz	0.47	156	125	6 min	2	0	48
230761	bt.D.x.ear.MO	monitoring	528 W	2.73 GHz	1.85 GHz	0.54	169	135	6 min	2	0	49

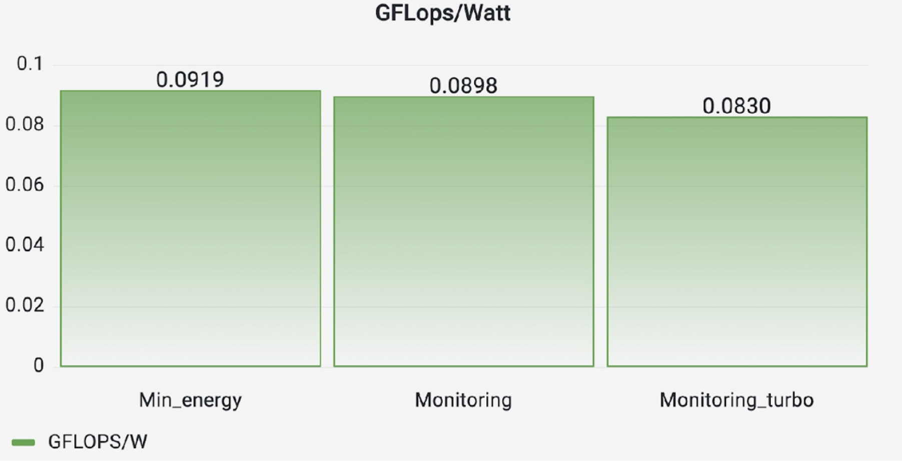
System metrics: Carbon footprint

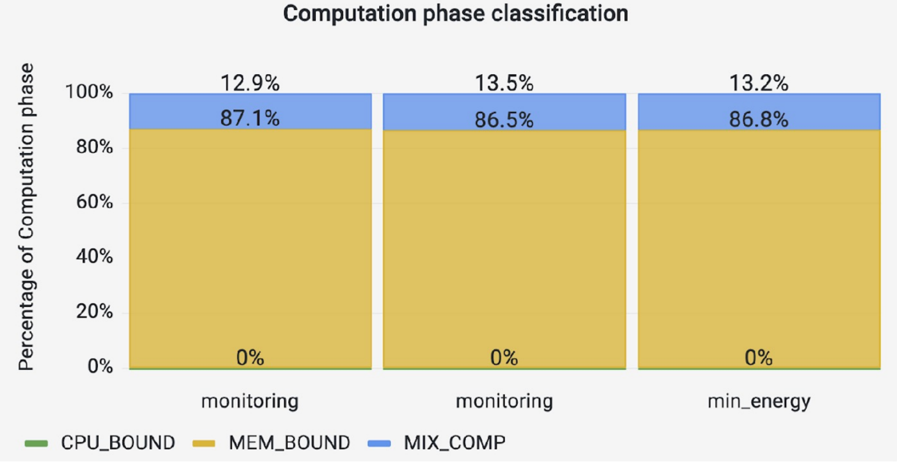


Data visualization

Job metrics visualization based on CSV files. Graphs compare results with different policies

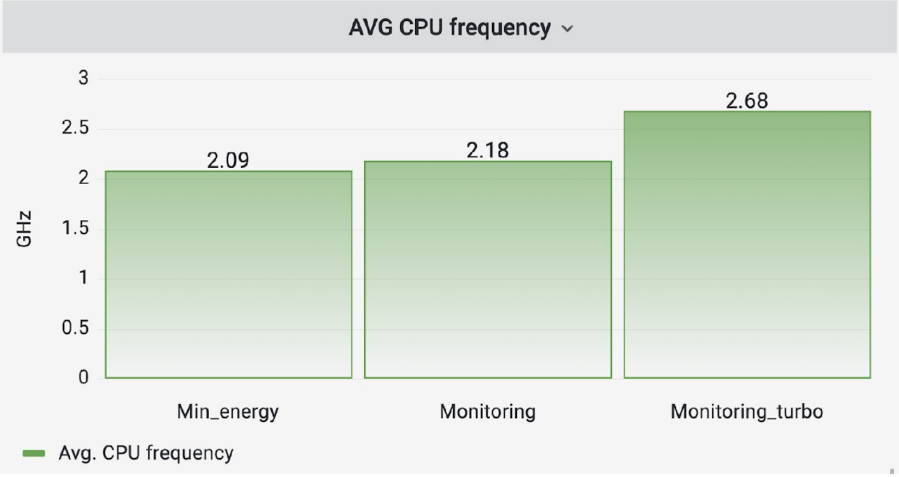
Gflops/W

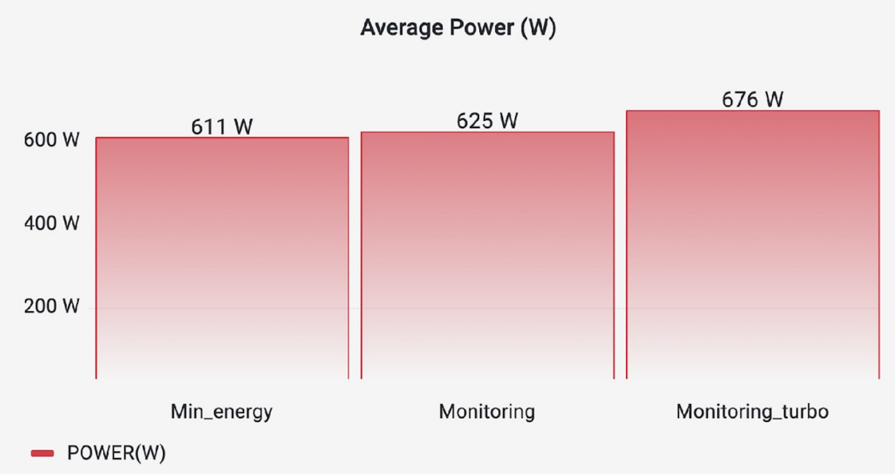

GFLops/Watt


Computation phase classification

Percentage of time
in CPU/MEM/MIX computation

CPU Freq. (GHz)

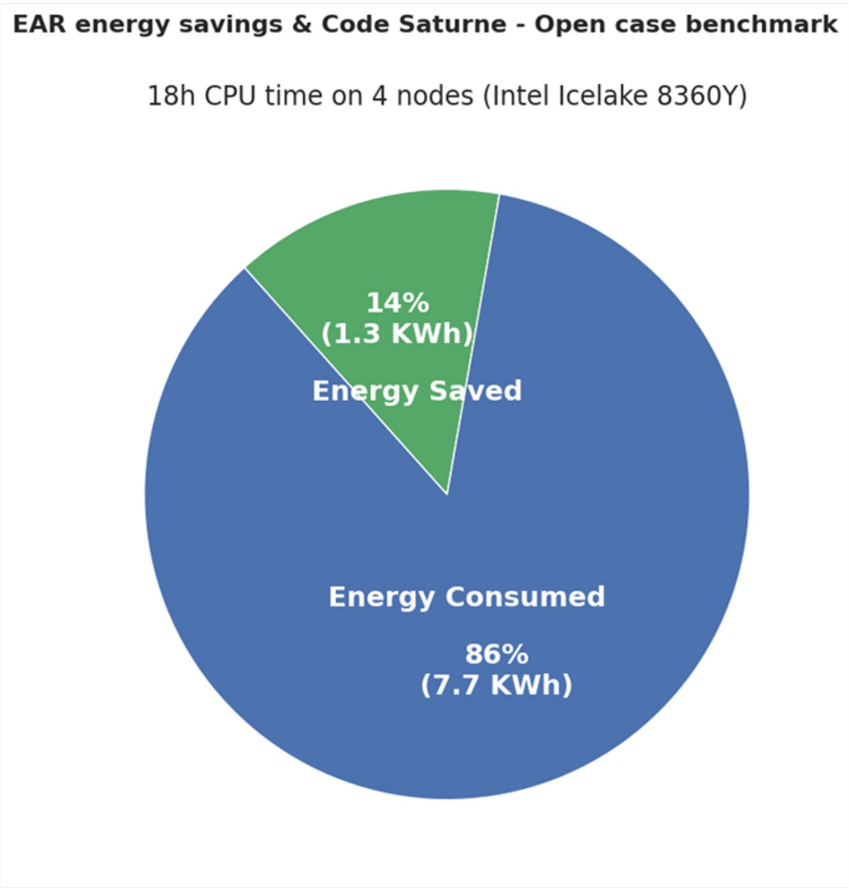

AVG CPU frequency


Average Power (W)

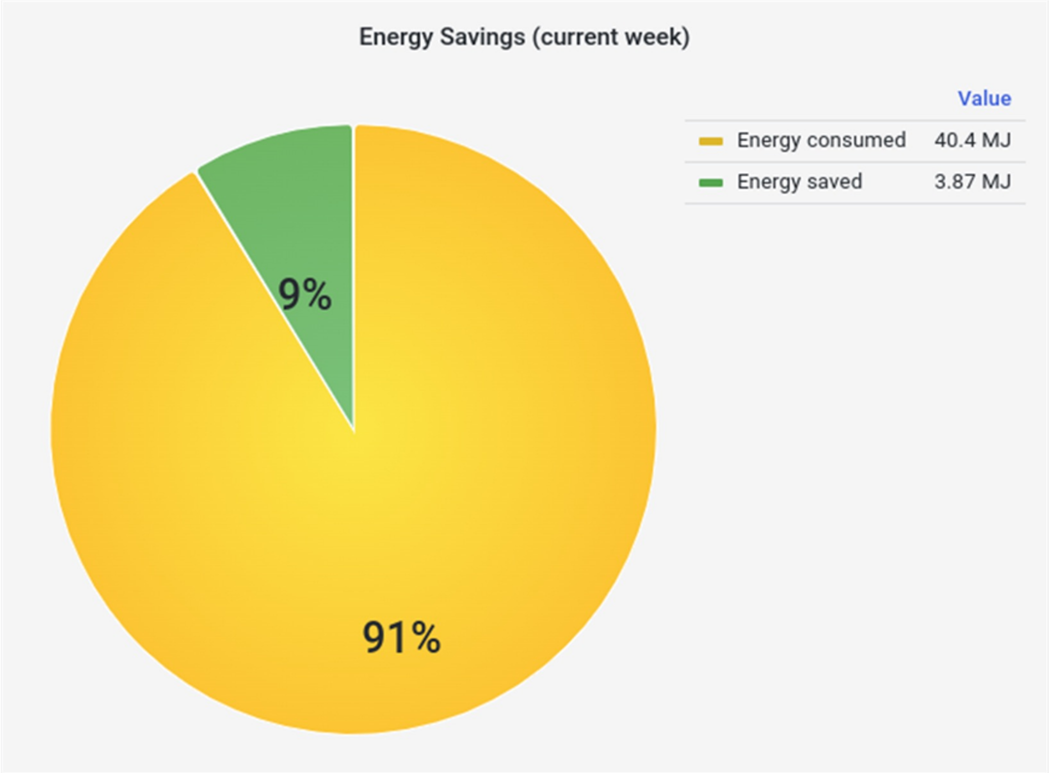
Average power (W)

Energy savings estimation

Job metrics visualization using eas-system-analytics and eas-job-savings tool



EAR energy savings* per job



EAR Cluster energy savings* per - period

(*): energy savings % = node power savings % – time penalty %

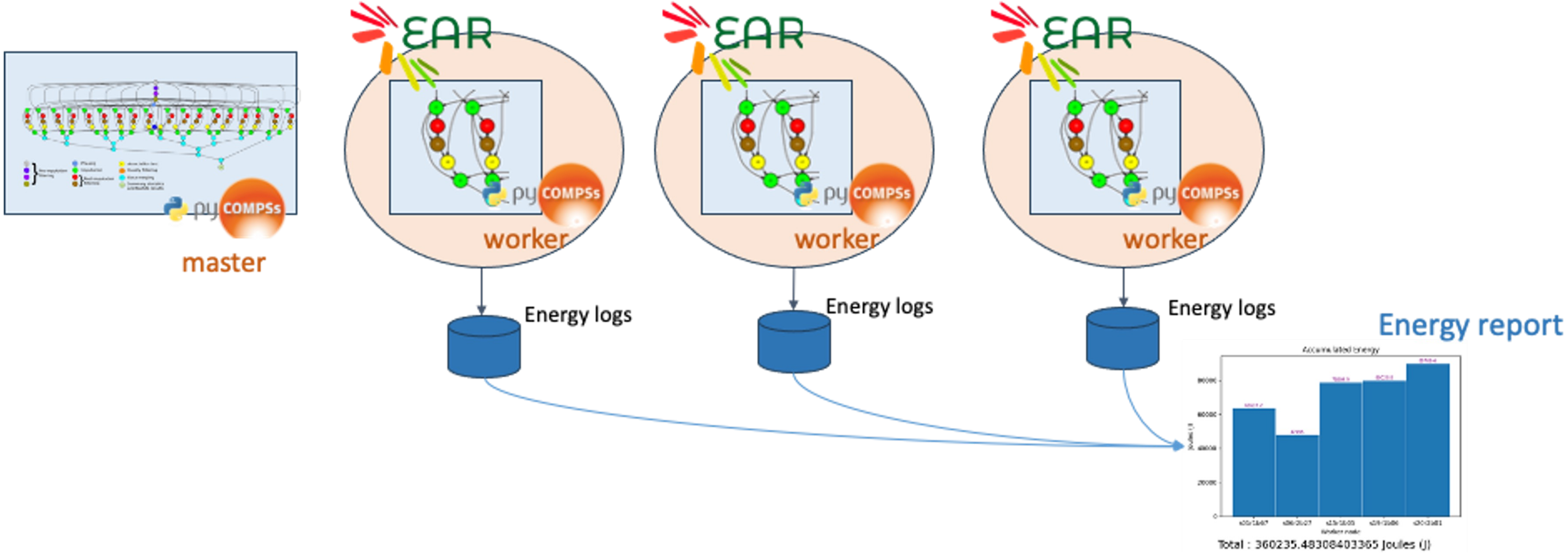
EAR Data center monitoring

EDCMON

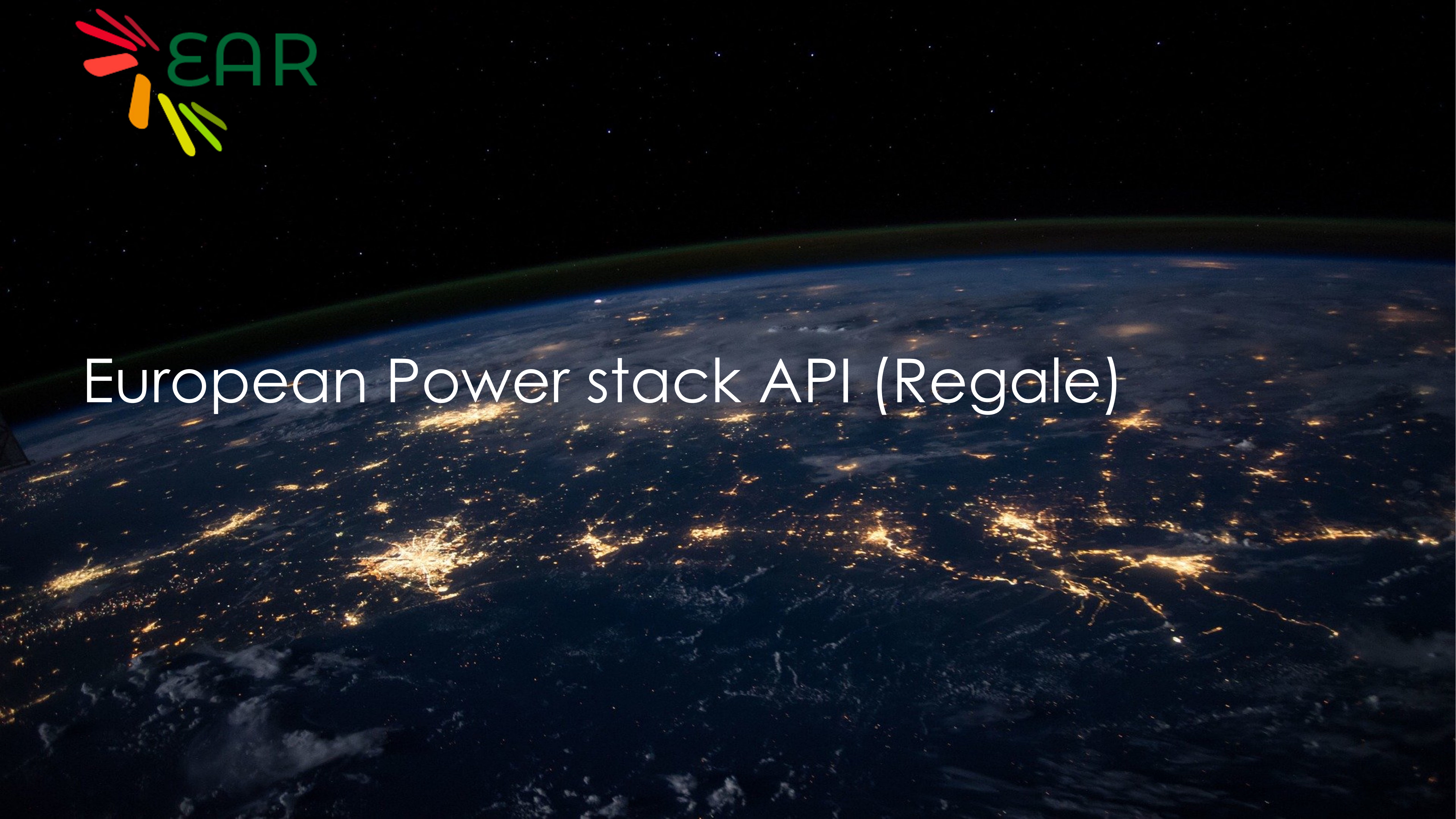
- New EAR service to monitor additional elements in the Data center apart from Compute Nodes
  - Storage
  - Network
  - Management
- EDCMON is extensible based on plugins
  - Period for monitoring
  - List of PDUs
  - AC power
  - Report strategy: Log, Prometheus, etc

Workflows support: PyCOMPSs

EAR + BSC PyCOMPSs integration



- EAR and PyCOMPSs are integrated to monitor and optimize individual tasks in workflows
- Extensions to support multiprocess python
- Extensions to support multiple applications (workers) runnin in same jobid/stepid context



European Power stack API (Regale)

European power Stack

- EAR team is an active partner in the design and development of the European initiative to create a power stack architecture and API
- Standardization effort done in the REGALE project
- EAR5.0 architecture is compliant with the proposed REGALE architecture
  - Implements the Node Manager API

- https://regale-project.eu/

EAR

https://www.eas4dc.com