E4 COMPUTER ENGINEERING

Sustainable and Autonomous HPC: Reconfigurable Architectures for Data-Intensive Science

Giordano Mancini, E4 Computer Engineering, CTO

HPC-AI Advisory Council

Locarno, April 21, 2026

E4 INNOVATION HISTORY

E4 COMPUTER ENGINEERING

2002
E4
FOUNDATION

2009
FIRST
HYBRID CPU-GPU
NVIDIA TESLA
CLUSTER

2012
SGI ITALIA
AQUISITION

2016 2017
D.A.V.I.D.E.
FIRST
OPENPOWER SERVER
& PETAFLOPS CLUSTER

299° IN TOP500 (JUNE 2017)
14° IN GREEN500 (JUNE 2017)

2019
SW INTEGRATION
& CI/CD FACTORY

EUROHPC
JU CALLS

2022
BOLOGNA TECNOPOLO
(**ECMWF** / **CINECA**)
+
FIRST
RISC-V BASED *CLUSTER*
(Monte Cimone)

CLOUD-NATIVE
AI PLATFORM

2023
DATACENTER
CHIRON LAB
NVIDIA GRACE &
NVIDIA GRACE
HOPPER

2024
GenAI
ON
PREMISE

GAIA
@ BOLOGNA
TECNOPOLO
(CINECA)

2025
E4
SPAIN

NEW HQ
RUBIERA

2026

2005
FIRST
INFINIBAND
CLUSTER

2013
FIRST
ARM+GPU SERVER
FIRST
ARM CLUSTER

PEDRAFORCA
CLUSTER

2018
E4 ANALYTICS
E4 ANALYTICS
FOUNDATION

EPI SGA1
PARTNER OF

2021
E4
DEUTSCHLAND

EPI SGA2

2023
EDOCC

2024
GAIA

2024
2025

CINECA | MARCO POLO

E4 COMPUTER ENGINEERING

CINECA

Year	*2024~2025*
Customer	*CINECA @Bologna Tecnopolo*
Solution	*Installation*
Partners	*Lenovo, VAST Data*
	*E4 deploys, integrates and supports multi-tier storage infrastructure based on **VAST Data** and additional storage architecture based on HDDs and TAPE*
	• *228 **Lenovo** SD665 V3 CPU nodes equipped with AMD EPYC 9745 128C 400W 2.3GHz, 320W-400W CPU.* *The CPU partition will provide 4,043 PF of HPL performance* • *90 **Lenovo** SD650-N V3 GPU nodes each with 4 x Nvidia H100 80GB HBM3, 2 x Intel Emerald Rapids 8592+ 64C 1.9Ghz.* *The GPU partition will provide 15.24 PF HPL*

CINECA | GAIA

CINECA

E4 COMPUTER ENGINEERING

Year	*2024~2025*
Customer	*CINECA @Bologna Tecnopolo*
Solution	*Design, Installation, On-Premise maintenance*
Partners	*DELL, VAST Data*
	*E4, in partnership with **Dell** and **VAST Data**, is building a solution that can increase the number of available vCPUs by 10 times, improve storage capacity and AI-based data management of **CINECA**'s **Galileo 100+ (GAIA)**, Italy's largest cloud for public research*
	E4 has supplied the system with: • *Examon HPC Monitoring (Exascale Monitoring) for Data Collection and Analysis* • *E4 Medooza Cluster Management Suite* • *On-site permanent support with 60 minutes intervention for critical incident*

Currently under evaluation two EuroJU AI factories … fingers crossed!

EuroHPC Joint Undertaking

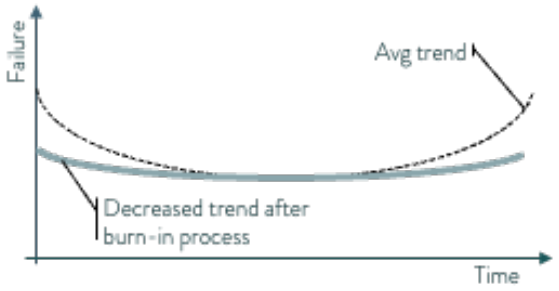
OPERATIONS & SUPPORT

E4 has the facilities, logistics and expertise to ***pre-assemble systems****, up to entire racks, in its own laboratory, where every hardware and software component is* ***checked****,* ***updated*** *and* ***tested*** *before delivery to the customer.*

E4 has the facilities, logistics and expertise to ***pre-assemble systems****, up to entire racks, in its own laboratory, where every hardware and software component is* ***checked****,* ***updated*** *and* ***tested*** *before delivery to the customer.*

The attention paid to every stage of the process is well illustrated by the ***burn-in process****, which consists of a minimum 72-hour test phase during which the system is run at maximum capacity.*

This activity gives back the guarantee to the customer to minimize the failure of components during their lifecycle.



When in production, E4 is able to support the customer in any kind of issue, ***remotely*** *or* ***with on-premise specialistic personnel****. From direct activities on hardware to RMA management and logistics, the team could guarantee the maximum reliability.*

Timeline view

Loads data on

trigger for

Interacts with

Operators

RK	NAME	OPENSTACK_NAME	BMC_IP	BMC_USER	BMC_PWD	ROLE
RK020289	idnode21	idnode13	172.16.4.111	ADMIN	ADMIN	SEED
RK020619	idnode22	controller01	172.16.4.112	ADMIN	ADMIN	CONTROLLER
RK020620	idnode23	controller02	172.16.4.113	ADMIN	ADMIN	CONTROLLER
RK020621	idnode24	controller03	172.16.4.114	ADMIN	Adminadmin2	CONTROLLER

Redfish

Redis

Pre Kayobe

netboxlabs

From here, Netbox becames SSOT for a Given project

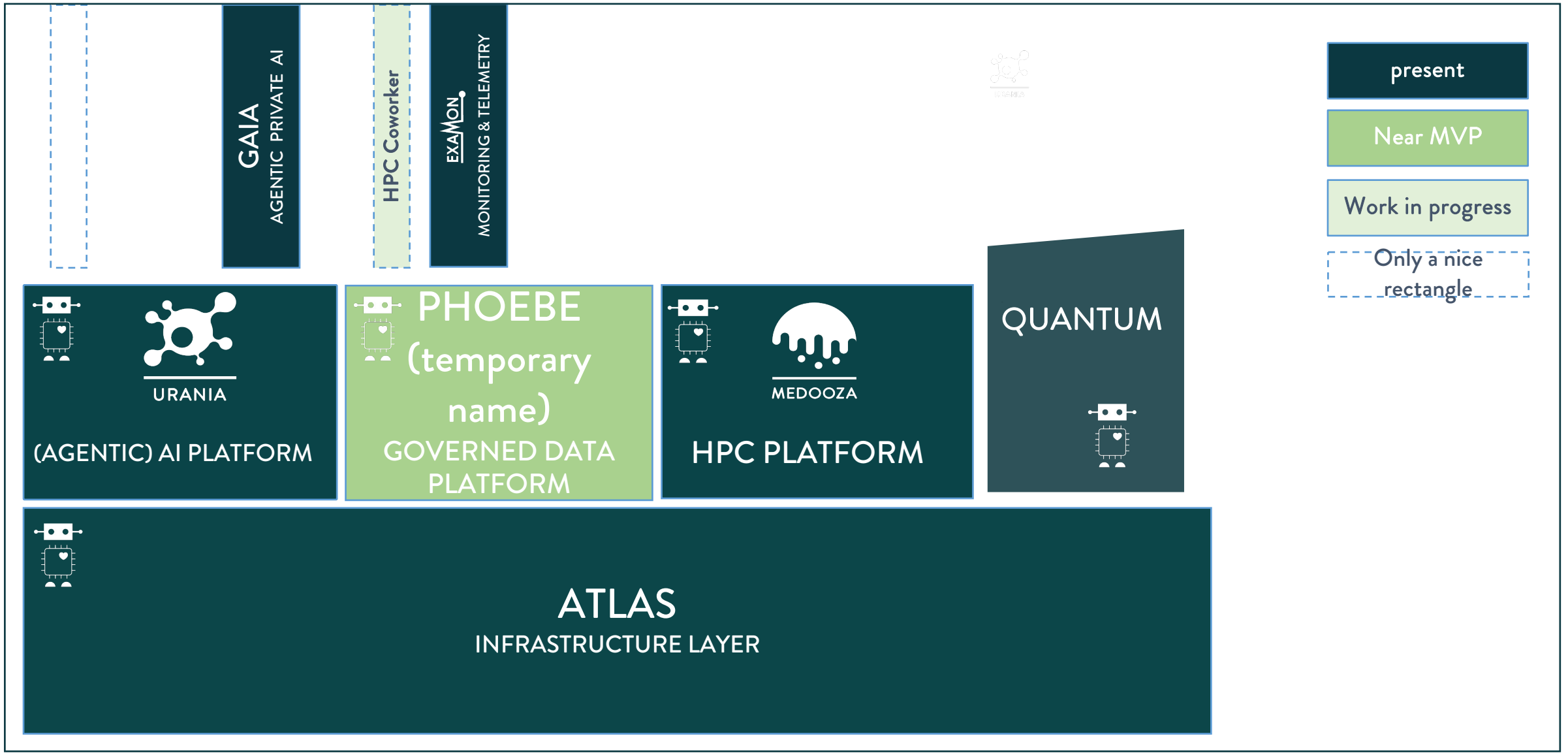
Medooza

get available secrets + set medooza secrets

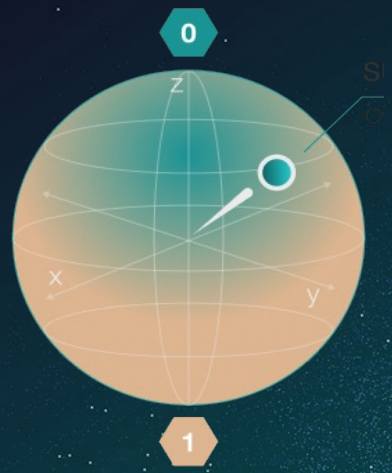
integrate network-allocation.yml info

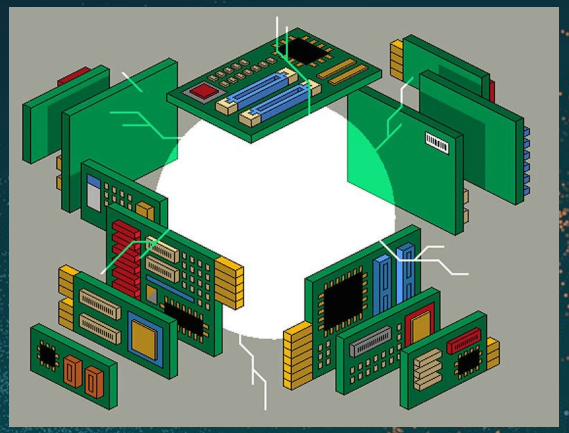
E4 SW SUITE

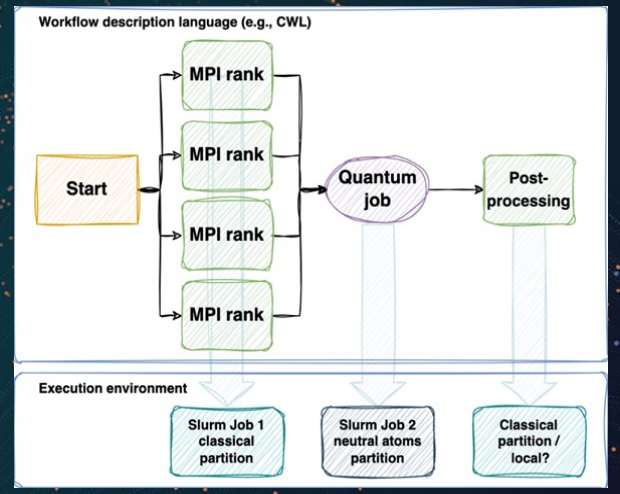
E4 COMPUTER ENGINEERING



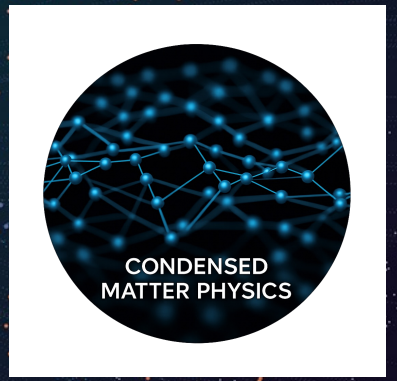
E4 COMPUTER ENGINEERING

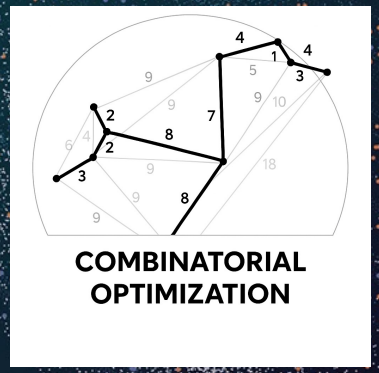


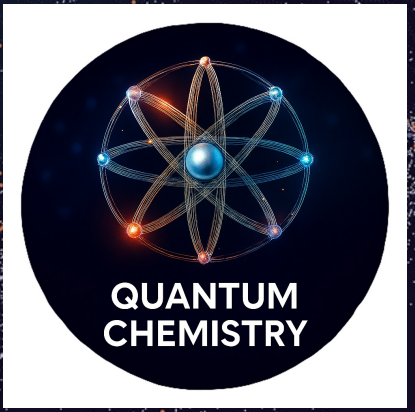




How not to Decelerate your HPC Workloads with Quantum Computing



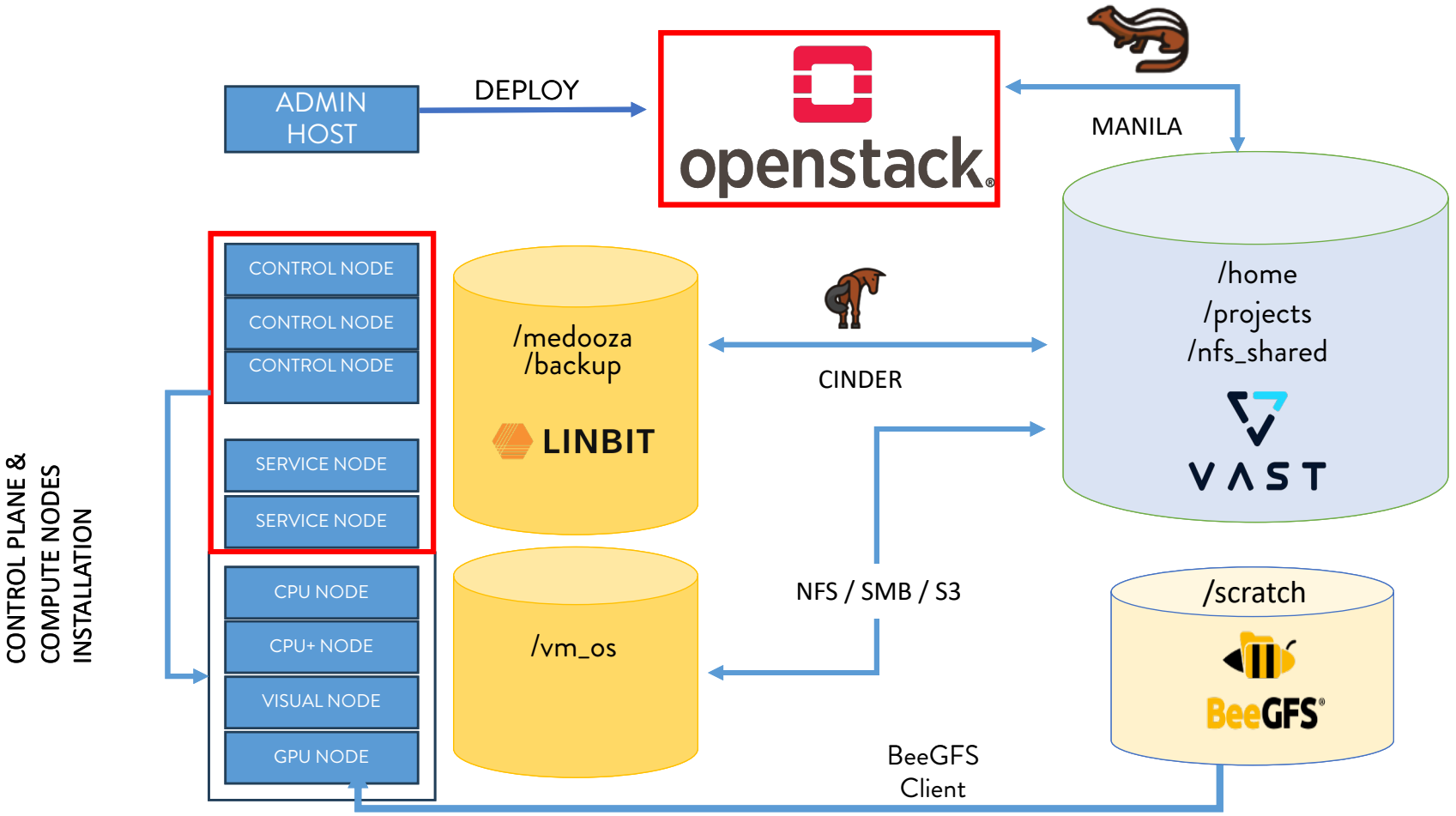




Gabriella Bettonte & Roberto Rocco – E4 Computer Engineering

23 rd April 2026

IIT | RAISE – EXAMPLE SYSTEM



- *HPC oriented cloud*
- *Simulation & inference*
- *Limited number of tenants, trusted*
- *High tenant privileges*

- *Basic infrastructure can be scaled out*
- *Workloads by VM or orchestrated containers*
- **Different kind of storage on demand**

WHY OPEN STACK?

Performance

- Bare Metal Provisioning (Ironic)
  Deploy physical servers with cloud-like agility. Zero hypervisor overhead for maximum FLOPs/latency-sensitive workloads
- AI Hardware Acceleration
  Native support for GPU Passthrough (PCIe) and vGPU (NVIDIA vGPU/MIG) via Nova
- Topology Awareness
  NUMA-aware scheduling and CPU pinning ensure training jobs land on the most efficient cores/memory
- Scalable: new resoures may be added with ease supporting different models of life cycle. Policies and dynamic allocation

AI Engine

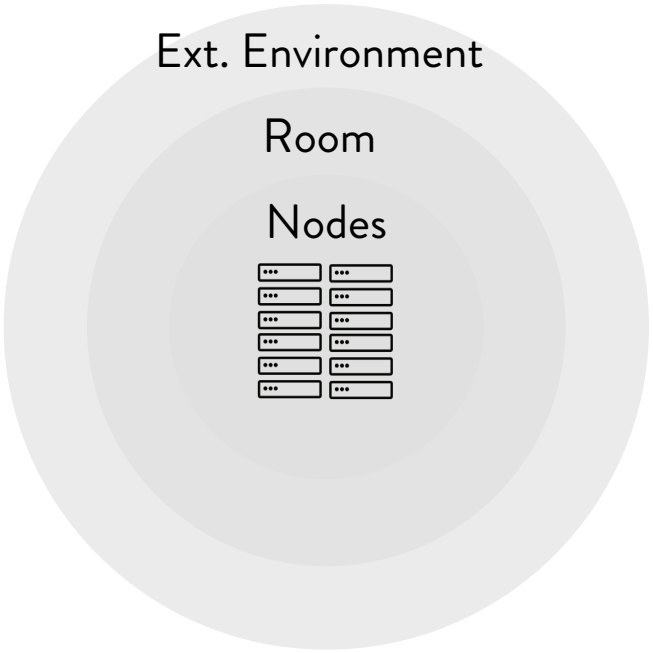
- Kubernetes on Demand (Magnum)
  Automated deployment of certified K8s clusters. Provides the substrate for Kubeflow, Ray, or PyTorch distributed training
- Multi-Tenancy
  Hard isolation between research teams or projects (namespaces, quotas, network segmentation) within the same physical cluster

Open and sovereign ecosystem

- Agnostic Vanilla, community driven solutions ensure maximum control of own stack, no explicit or hidden lock-in
- Flexible: the on prem cloud ensures adaptability to every type workload and ample reconfiguration possibilities

THE HPC MONITORING CHALLENGE

Modern HPC clusters generate millions of measurements per hour across incompatible interfaces.
No single tool sees the full picture.



Fragmented data sources

IPMI / BMC sensors
Out-of-band hardware telemetry per node

GPU Telemetry (NVML / DCGM)
Power, temperature, utilization, errors

Job Scheduler (SLURM/PBS)
Job accounting, energy, resource usage

Prometheus, OTeL, ...
Application and system-level metrics

SNMP, Modbus, BACnet, OPC UA, ...
Site-level equipment and BMS

What this costs you

Hardware faults go undetected
No Cross-domain correlation means problem hides in plain sight

Energy blindspot
Per-job energy is unknown without combining Power measurements + scheduler data

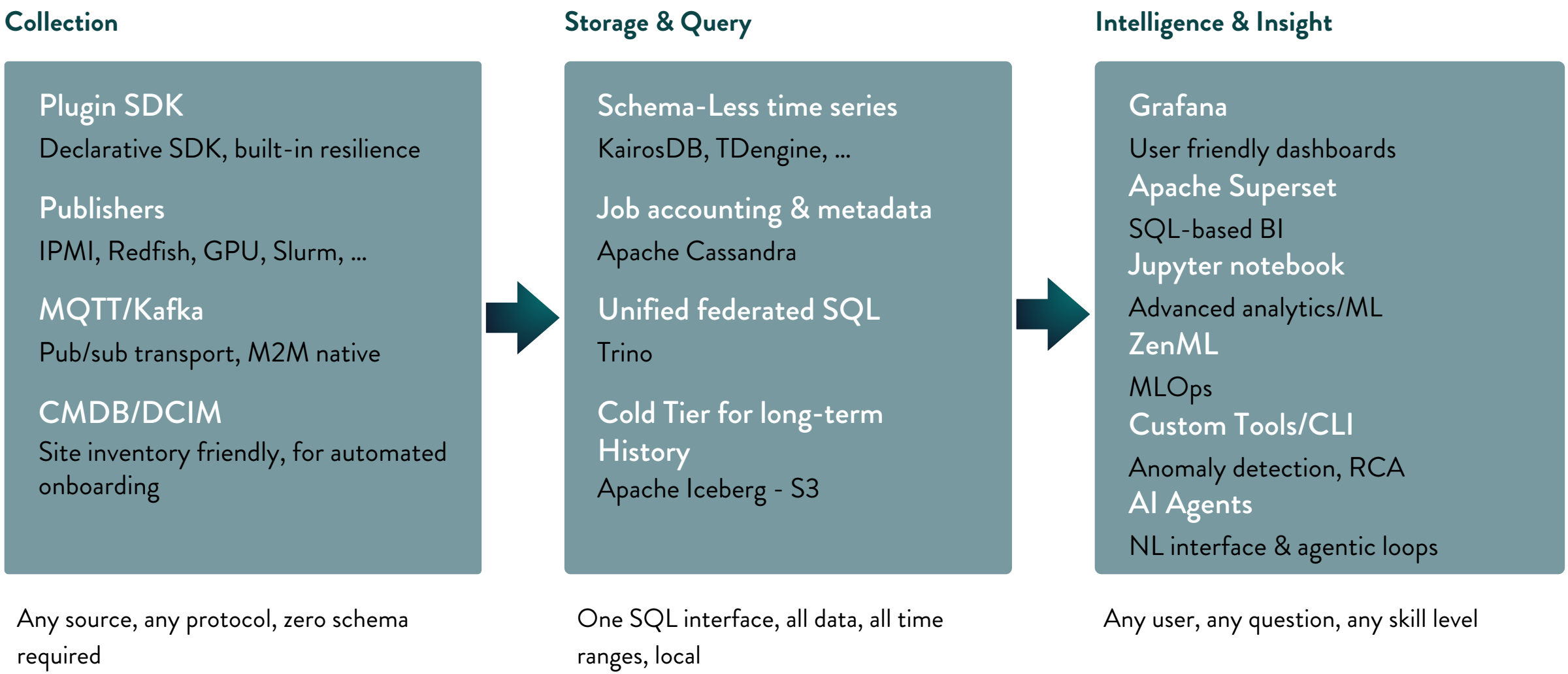
Expert-only investigation
Incident response requires querying multiple systems manually

ML models train on partial data
Research built on siloed data misses key predictive signals

EXAMON - A DATA LAKEHOUSE FOR HPC

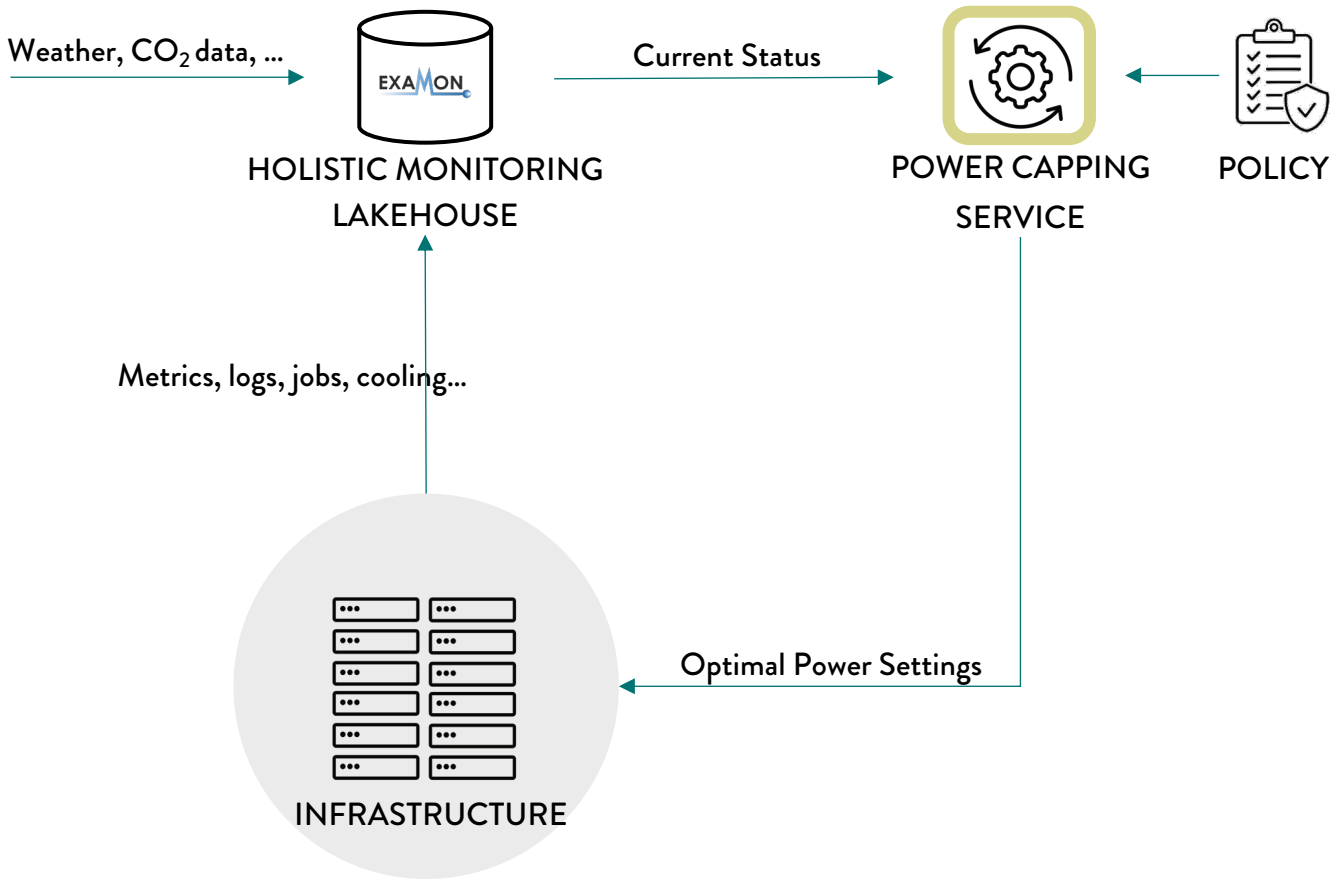
E4 COMPUTER ENGINEERING

Every signal from the infrastructure flows into a single, schema-less, scalable store and immediately available to operators, researchers, and AI agents.



POWER CAPPING SOLUTION

GPU-dense AI Factory clusters demand continuous power visibility, per-job energy accountability, and active power control. ExaMon provides all three

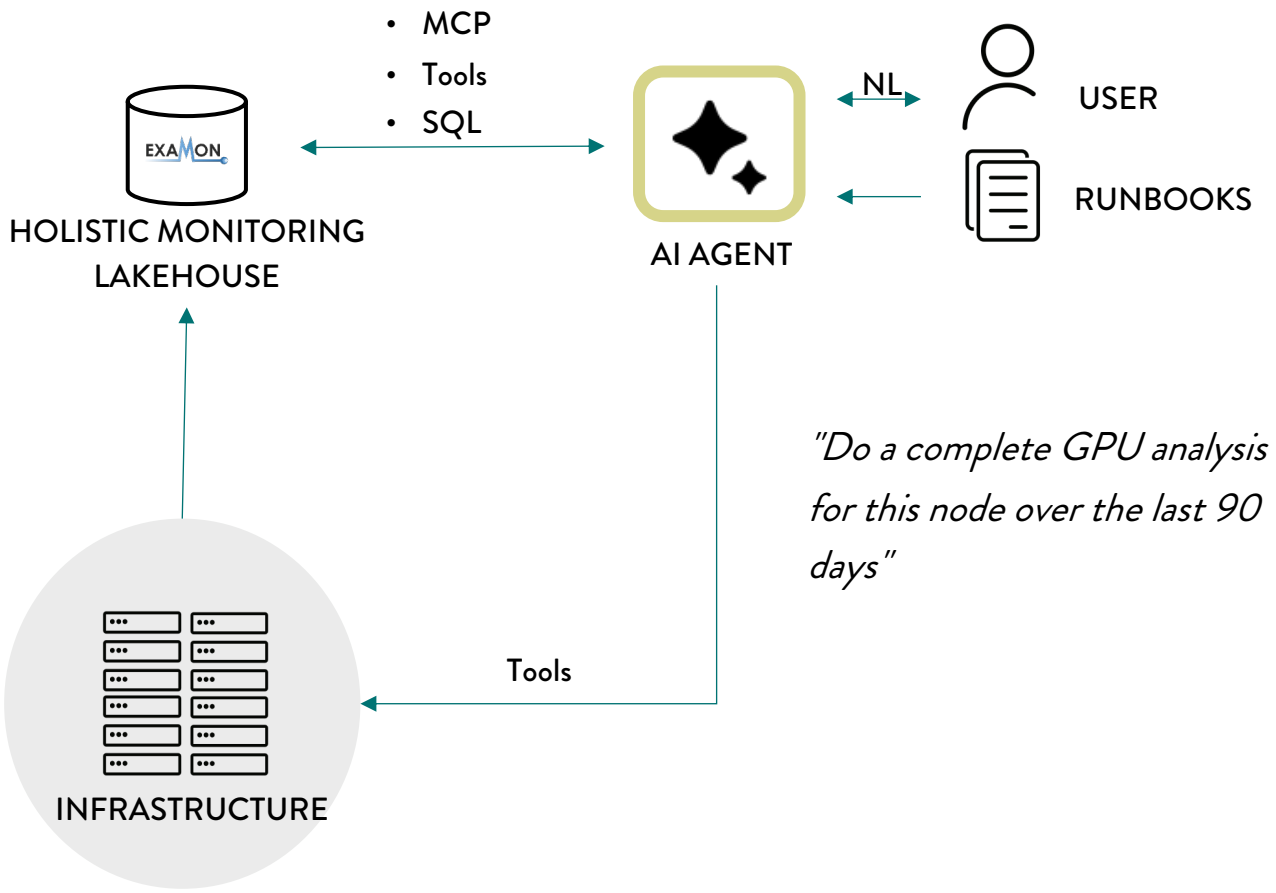


Features

- Out-of-band IPMI power cap commands
- Support for vendor specific interfaces: (Intel, AMD, NVIDIA, DELL, ...)
- Open-Source tools: Variorum, EAR, ...
- Node-level and cluster-level policies
- Closed-loop integration with job scheduler
- Operate within facility power envelope
- Foundation for carbon-aware scheduling
- Historical data enables ML/AI based approaches

AI AGENT

AI agents help users to analyze data taking advantage of the full SQL compatibility enabled by ExaMon.



Features

- **Runbooks**

  Teach the LLM investigation workflows: metric discovery, time-series analysis, job failure diagnosis, RCA, ...

- **Tools**

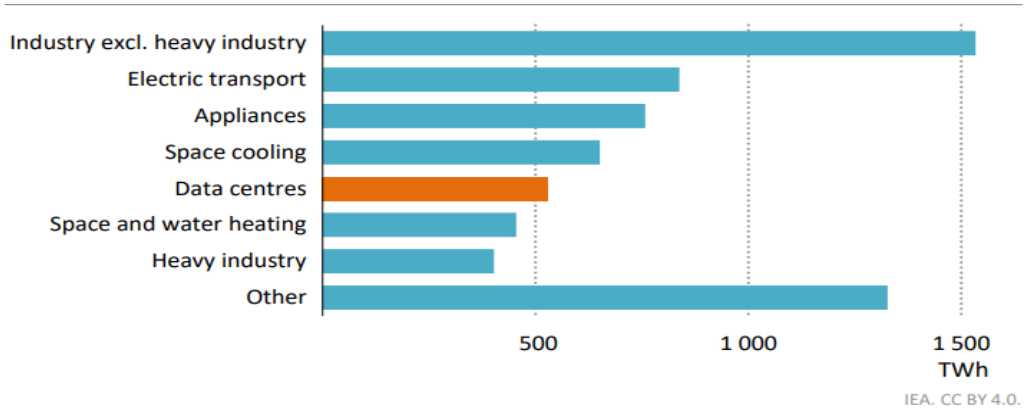
  CLI tools used by the LLM to interact correctly with the DB and the infrastructure (Actions whitelist).

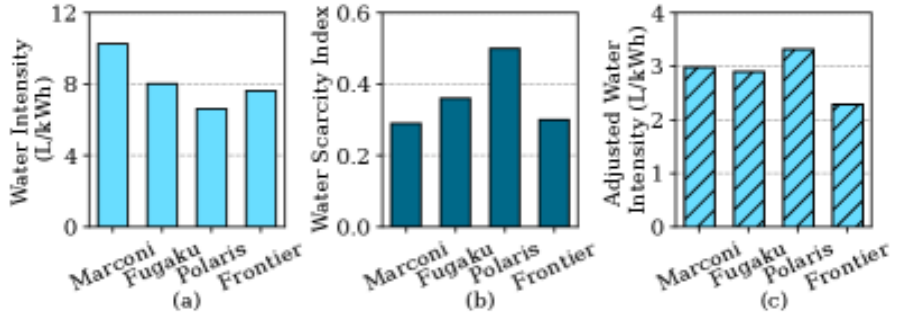
- **Local LLM**

  Runs entirely on-premises. Data never leaves the facility. Model-agnostic: tested with multiple open-weight models.

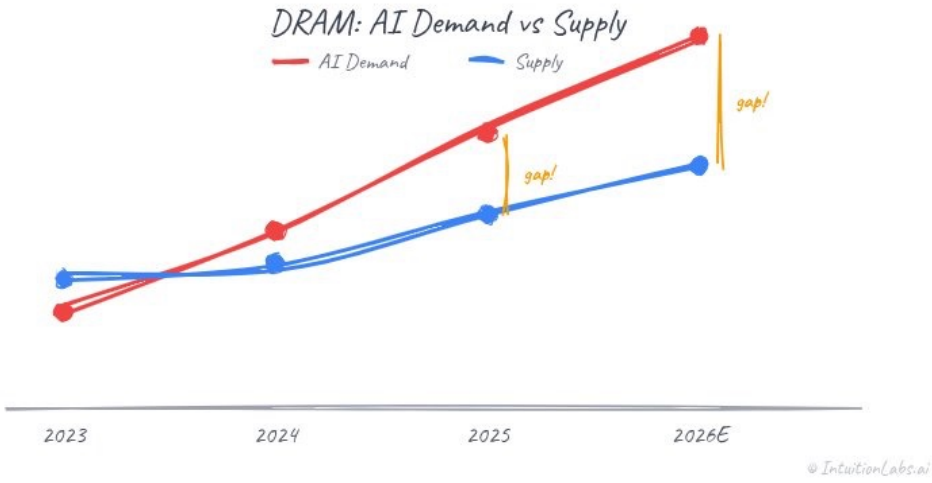
Energy and environmental footprint: TTS is not the right metric

"The International Energy Agency (IEA) projects that global data center electricity consumption, which stood at approximately 460 terawatt-hours (TWh) in 2022, could reach (Base Case) ~ 1,000 TWh by 2026, effectively matching the total electricity demand of a country the size of Japan" Energy and AI report 2025, IEA.

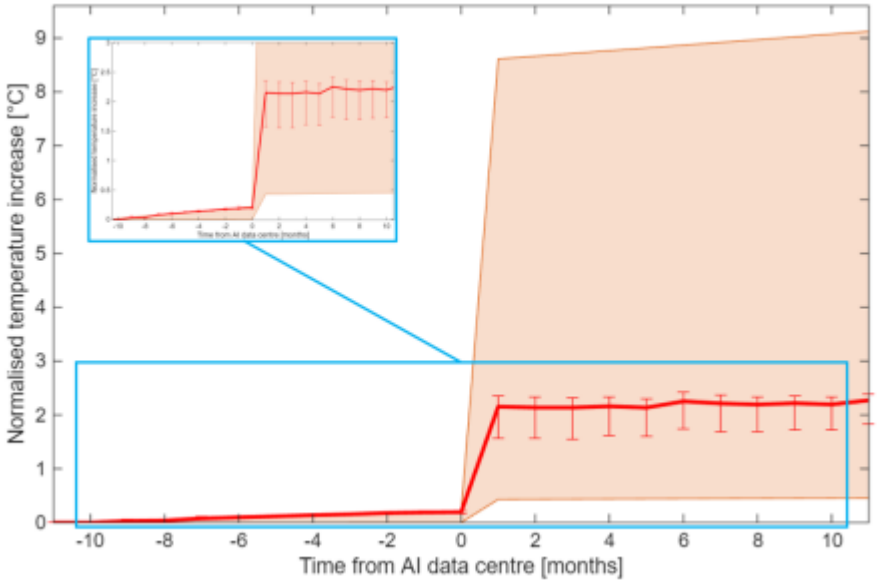




Annual average water intensities (water footprint) across different regions, (b) water scarcity index of different regions using AWARE-global data, and (c) the adjusted water intensity after combining the water intensity with the water scarcity index. arXiv:2510.00471v1 [cs.DC] 01 Oct 2025



Temperature increase through time over the AI hyperscalers locations centred around the time of start of operations (i = 0). https://arxiv.org/abs/2603.20897v1



URANIA

E4 COMPUTER ENGINEERING

DATA PREPARATION → MODEL DEVELOPMENT → MODEL TRAINING → MODEL VALIDATION → MODEL INFERENCE → MODEL MONITORING

URANIA CORE

URANIA BASIC SERVICES

- WEB UI FOR ADMINISTRATORS
- OBSERVABILITY STACK
- CONTAINER IMAGE REGISTRY
- GIT REPOSITORY
- OIDC IDENTITY PROVIDER
- HIGH PERFORMANCE S3 STORAGE

AI-OPTIMIZED KUBERNETES INFRASTRUCTURE

AI-OPTIMIZED KUBERNETES INFRASTRUCTURE

- CNCF OSS CONTAINER ORCHESTRATOR
- GPU & NETWORK KUBERNETES OPERATOR
- HIGH-THROUGHPUT SHARED VOLUMES
- HIGH-IOPS LOCAL VOLUMES

PHOEBE

E4 COMPUTER ENGINEERING

A unified on-premises data platform that enables developers to transform raw data into governed assets and perform 360° analysis

URANIA Data Platform

- ICEBERG LAKEHOUSE (ACID) WITH VERSIONING AND TIME-TRAVEL
- UNIFIED DATA CATALOG (BRANCHING, METADATA VERSIONING, ...)
- FAST SQL ENGINE FOR MEDIUM-SIZE DATA-SETS
- DISTRIBUTED SQL ENGINE FOR LARGE DATA-SETS
- DISTRIBUTED DATA PROCESSING WITH SPARK AND RAY OPERATORS
- NOTEBOOK-BASED DATA DEVELOPMENT ENVIRONMENTS
- INTERACTIVE BI FOR DATA EXPLORATION & VISUALIZATION

Backed by S3 services provided by Urania Core

URANIA & EXAMON to run benchmarks

E4 COMPUTER ENGINEERING

Context & Experimental Setup

Hardware: NVIDIA GB200 NVL72

- GPU Blackwell SM100: 192 GB HBM3e, ~8 TB/s bandwidth
- TDP: 1,200 W/GPU · Liquid cooling
- CPU: NVIDIA Grace ARM64 (NVLink-C2C)
- vLLM v0.18.0 · CUDA 13.0 · Rockåy Linux 10.1

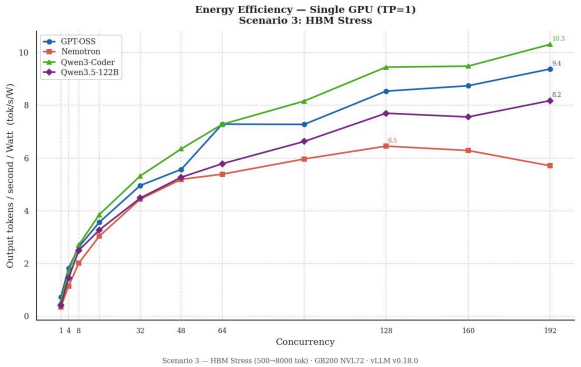
3 Benchmark Scenarios

Scen.	Input	Output	Goal
S1 Agentic	60k tok	1k tok	KV cache hit (long ctx)
S2 MaxPar	1k tok	1k tok	Peak throughput
S3 HBM	500 tok	8k tok	HBM bandwidth, pure decode

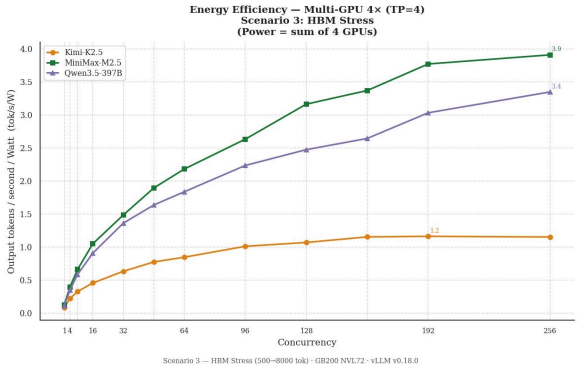
7 Models Tested

Model	TP	Architecture	Quantization
GPT-OSS 120B	1	Sparse MoE	MXFP4 native
Nemotron-3-Super	1	Mamba-2 + MoE	NVFP4 + DeepGEMM
Qwen3-Coder-Next	1	GDN + MoE	AWQ 4-bit (Marlin)
Qwen3.5 122B	1	GDN + MoE	NVFP4 community
— TP=4 —			
Kimi K2.5 (~1T)	4	MoE + MLA	NVFP4 + DeepGEMM
MiniMax M2.5	4	Lightning Attn + MoE	NVFP4 + DeepGEMM
Qwen3.5 397B	4	GDN + MoE	NVFP4 + DeepGEMM

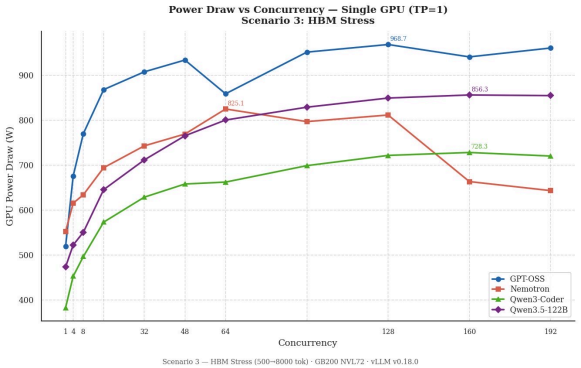
Energy Efficiency & Power Draw — Decode Phase



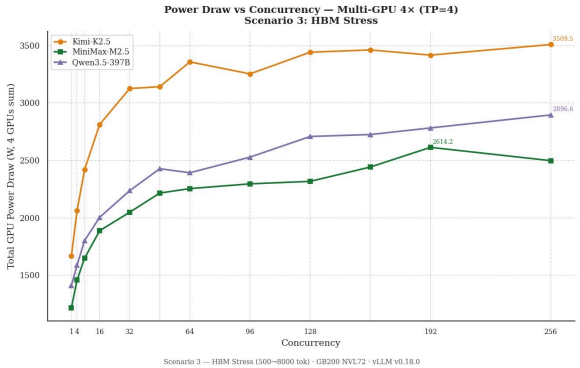
tok/s/W — Single GPU (TP=1)



tok/s/W — Multi-GPU TP=4



GPU Power Draw (W) — Single GPU



Total Power Draw (W, 4 GPUs) — TP=4

Energy efficiency — what drives tok/s/W

tok/s/W = throughput / power: efficiency wins who maximizes tokens produced per Watt.
Qwen3-Coder (10.2): smaller weight footprint (45 GiB) + low DRAM_ACTIVE (34%) ⟶ less HBM read per token.
Nemotron (6.1): same weight as GPT-OSS (~69 GiB) but SSM overhead cuts throughput; Watts remain similar.
Multi-GPU: MiniMax (31 GiB/GPU) 3.3× more efficient than Kimi K2.5 (139 GiB/GPU) — weight footprint is decisive.

Efficiency by category (S3)

Model	tok/s/W	Config
Qwen3-Coder	10.2	1 GPU
GPT-OSS 120B	9.3	1 GPU
Qwen3.5 122B	8.2	1 GPU
Nemotron-3-Super	6.1	1 GPU
MiniMax M2.5	3.77	4 GPU
Qwen3.5 397B	2.40	4 GPU
Kimi K2.5	1.15	4 GPU

Architecture–efficiency correlation is direct: tok/s/W correlates with DRAM_ACTIVE.

Qwen3-Coder leads (10.2): its 45 GiB of weights require fewer HBM reads than GPT-OSS at 69 GiB (9.3). The advantage holds despite Marlin WNA16 not using Blackwell FP4 Tensor Cores.

MiniMax (33% DRAM) beats Kimi K2.5 (53% DRAM) in multi-GPU by 3.3×: decisive factor is weight footprint per GPU (31 vs 139 GiB), not MoE sparsity.

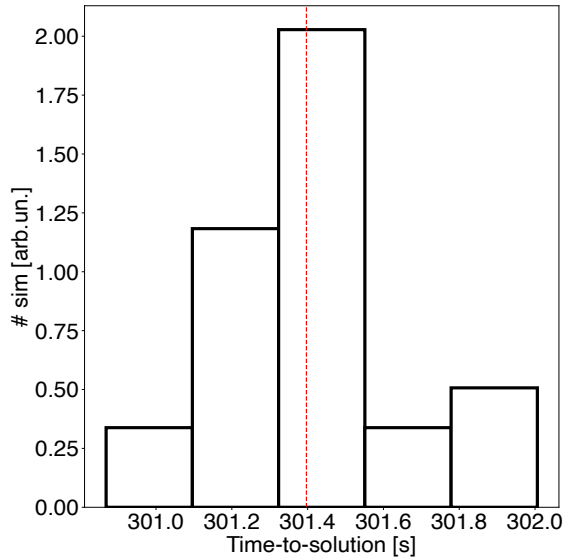
GPT-OSS: peak power 969 W (c=128), stable in 900–960 W band from c=32+. Kimi K2.5: power saturates at ~3,510 W total (4 GPUs) already at c=32.

RISC-V ACCELERATORS: LESS ENERGY CONSUMPTION

E4 COMPUTER ENGINEERING

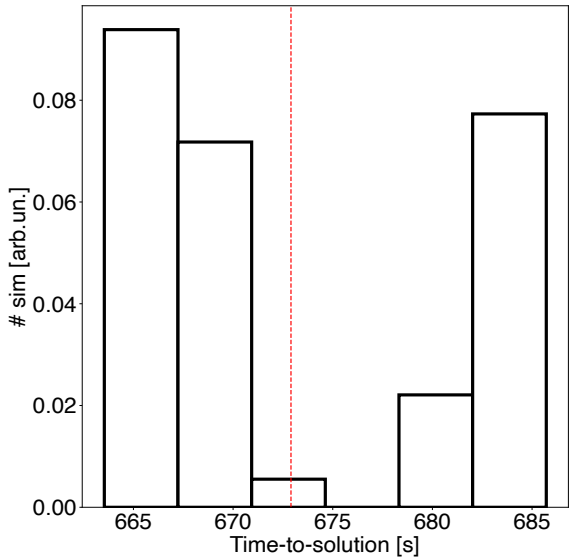
J.L. Almerol et al. (2025), arXiv pre-print, arXiv:2509.19294

Time-to-solution (TTS) over ~50 simulations



1 MPI task +
1 OMP thread +
1 Tenstorrent n300

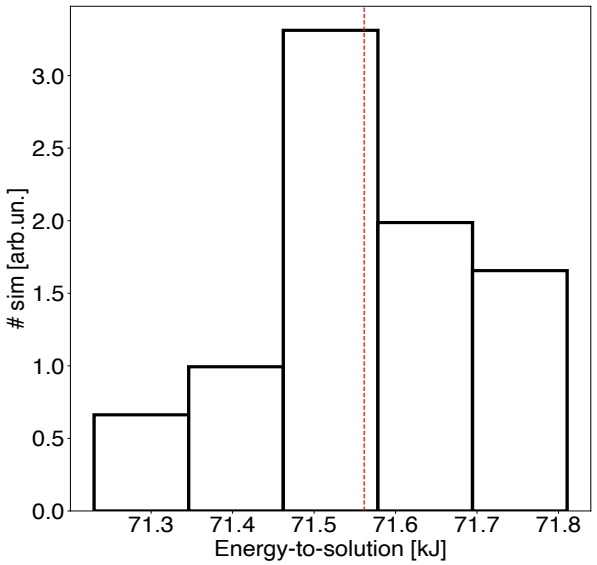
TTS_{avg} = 301.40 ± 0.24 s



1 MPI task +
32 OMP threads +
AVX-512 intrinsics

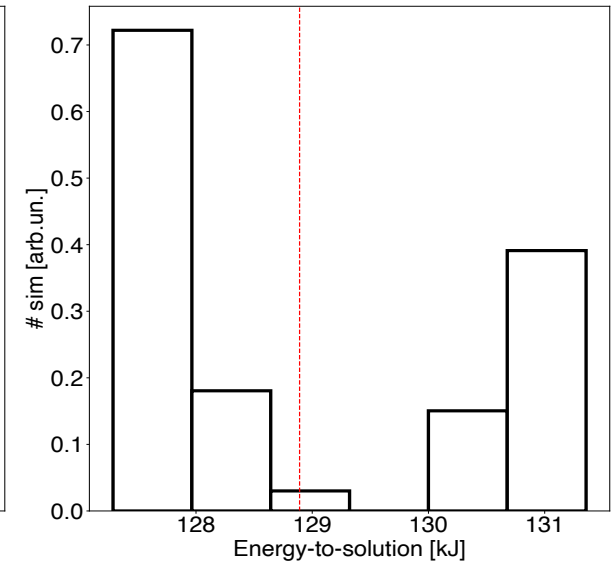
TTS_{avg} = 672.90 ± 7.83 s

Energy-to-solution (ETS) over ~50 simulations



1 MPI task +
1 OMP thread +
1 Tenstorrent n300

ETS_{avg} = 71.56 ± 0.13 kJ



1 MPI task +
32 OMP threads +
AVX-512 intrinsics

ETS_{avg} = 128.89 ± 1.52 kJ

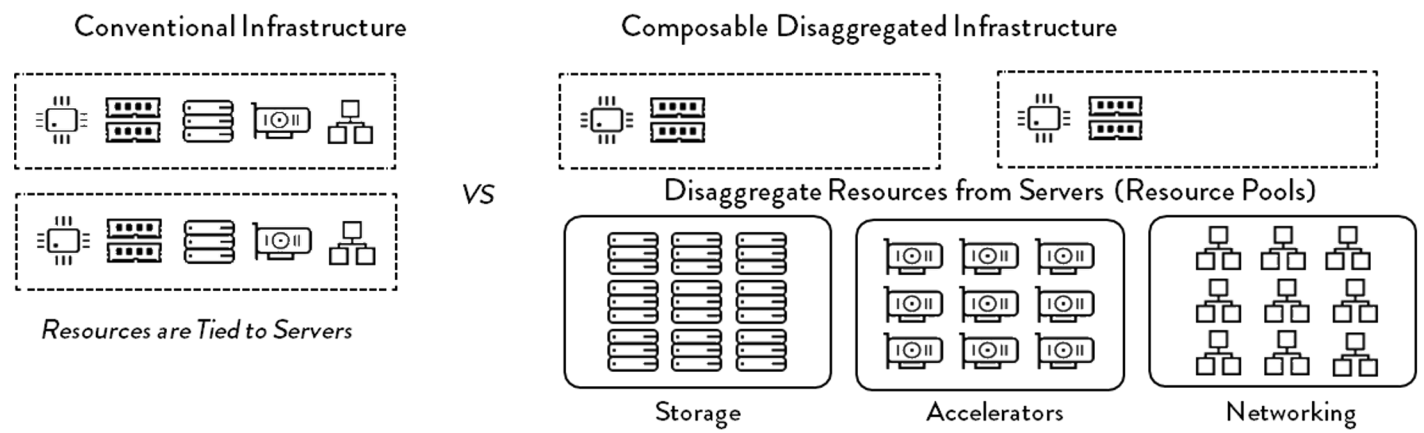
RISC-V4ASTRO

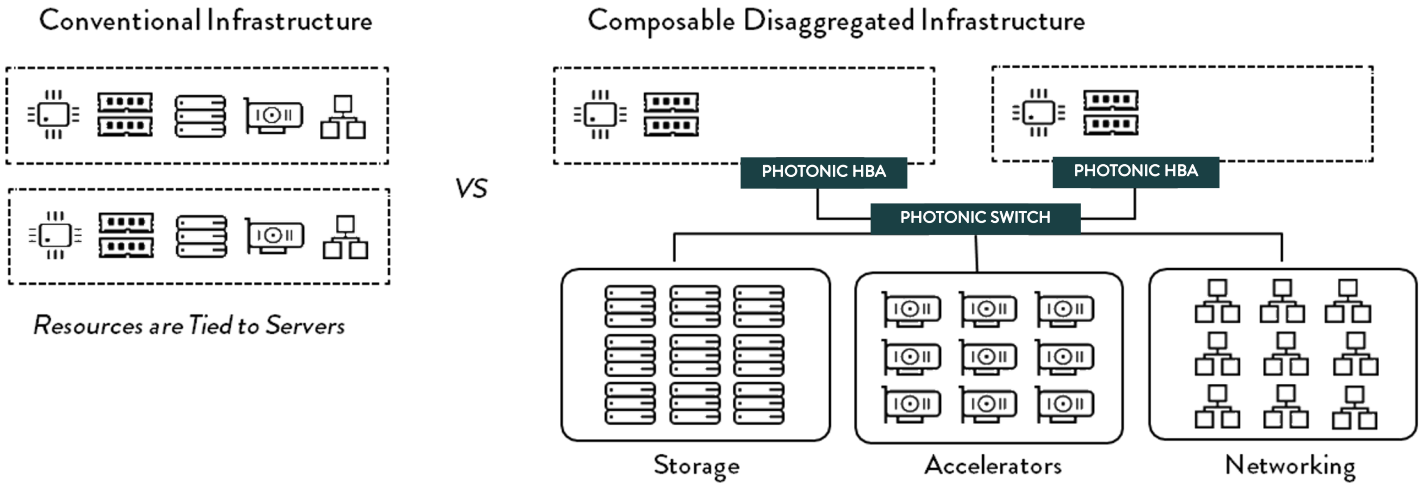
SPACE

EFFICIENCY AND CONSUMPTION:

COMPOSABLE DISAGGREGATED ARCHITECTURE

E4 COMPUTER ENGINEERING





INSTEAD OF RESIDING IN SERVERS, RESOURCES ARE
DISAGGREGATED INTO POOLS

RESOURCES CAN ALSO BE INTERCONNECTED THROUGH
PHOTONIC TECHNOLOGY

USING EXAMON AND SLURM ON THE FLY WE CAN MEASURE EFFICIENCY TRADEOFF AND ACCOMMODATE USER REQUESTS

R&D ACTIVITIES RUN IN E4 LAB WITH PHOTONIC SHOW NO PERFORMANCE LOSS ON MOLECULAR DYNAMICS TEST

CONCLUSIONS AND PERSPECTIVES

E4 COMPUTER ENGINEERING

Composable Disaggregation for Zero-Waste Infrastructure: Moving away from static, monolithic server nodes toward true Composable Disaggregated Infrastructure (CDI) enabled by photonic fabrics. Photnic Fabrics is a technology in which EU should bet? How to balance flexibility in architecture with further potential lock in (with respect to HW abstraction)?

RISC-V as the Engine for Sovereign, Low-Power Acceleration: beyond strategic autonomy, RISC-V can provide a vital pathway to sustainable exascale computing by creating specialized accelerators for either traditional FP64 workloads or low precision inference.

Energy-Aware Computational Experiments: Shifting the user paradigm: providing energy budgets instead of compute hours would shift responsibility in the wrong plance? Ask the system to "Run this potential energy surface scan within 2 weeks optimized for the lowest possible carbon footprint," or "Complete this LLM training within a strict 500 kWh energy budget").

E4 COMPUTER ENGINEERING

THANK YOU

CONTACTS

info@e4company.com
support@e4company.com
sales@e4company.com

E4 Computer Engineering SpA
Via Martiri della Libertà, 66 . 42019 Scandiano (RE) - Italy
Tel. +39 0522 991811