CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

EMOI: CSCS Extensible Monitoring and Observability Infrastructure

PASC24

Jonathan Coles, CSCS

June 3, 2024

DISCLAIMER: I’m just the messenger!

This work is a collaboration of many people at CSCS.

Massimo Benini
Michele Brambilla
Dino Conciatore
Monica Frisoni
Mathilde Gianolli
Gianna Marano
Gianni Ricciardi

Benjamin Cumming
Jean-Guillaume Piccinali
Jonathan Coles

All results and conclusions presented today are on early access and pre-acceptance systems.

Agenda

The ALPS Supercomputer at CSCS

Power Concerns

Measuring and Validating
Power with EMOI

Understanding
Power Usage

Applications

Outlook

Alps: a multi-tenant HPE Cray EX system with heterogeneous resources

The **User Lab** scale-out on the Grace-Hopper platform will be the largest tenant on Alps.

It will replace Daint-GPU with the same number of Grace-Hopper modules (>5000) as there are on Daint-GPU today.

2020

Phase 0

- 2x AMD Rome 64-core CPUs
- Currently used in Eiger

2022

Phase 1

- AMD Milan CPUs
- 4x NVIDIA A100 *or* 4x AMD Mi250x GPUs

2024

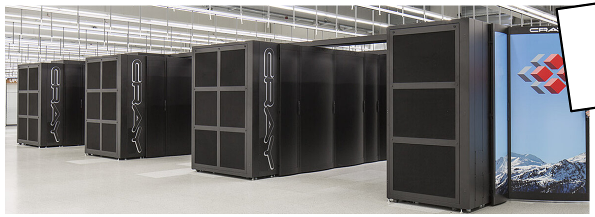
Phase 2, Q1-Q2

- 4x GraceHopper SuperChip Modules
  - 4x Grace 72-core ARM CPUs
  - 4x Hopper H100 GPUs

Phase 3, Q3

- AMD EPYC CPUs
- AMD INSTINCT MI300A GPUs

CSCS HPC Power Consumption



2x GPUs but ~4x power!

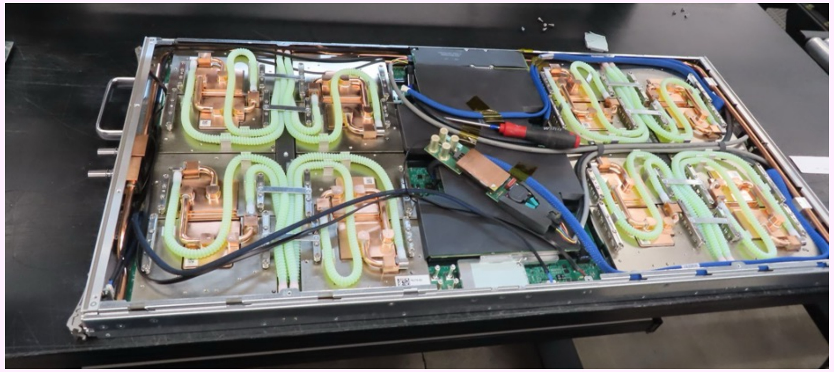
Piz Daint
- 5,700 P100 nodes
- 1 node → 300 W each under full load

Alps
- > 2,500 GH200 nodes
- 1 node → 2600 W under full load

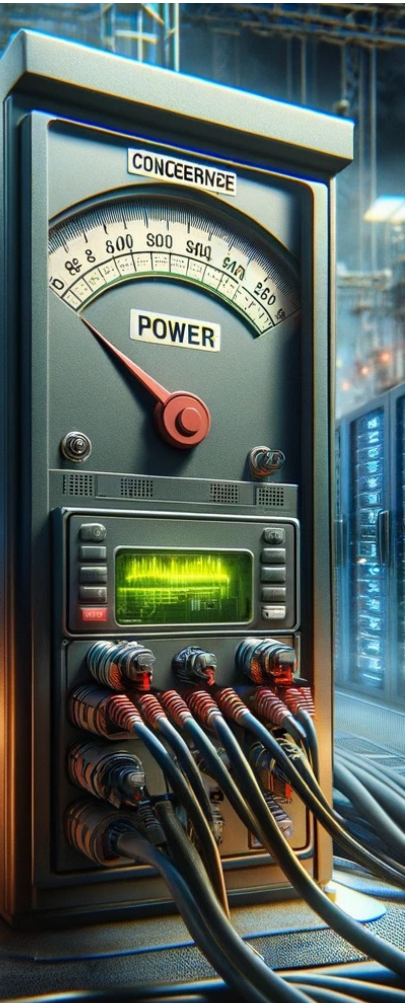
Energy is currently ~10% of User Lab expenditures

ALPS Blade

- 2 Blanca Peak (EX254n) nodes each with 4x GraceHopper modules
- Each GH200 module:
  - 72 core ARM Neoverse v2 CPU with 128 GB LPDDR
  - H100 with 96 GB HBM2e
  - Thermal design power (TDP) = 800W
  - Will be power capped



Power Concerns at HPC Centers



Understanding how we use energy is vital.

Total Power

- Increased electricity costs in Europe since 2023
- Top machines are hungry
  - 22 MW for Frontier
  - 38(!) MW for Aurora
  - ~7 MW for ALPS in production

Top 500	Green 500	Top 6 Systems June 2024	Rmax [PFlops/s]	T500 Power [kW]	G500 Power Eff. [GF/W]
1	7	Frontier / ORNL	1,206.00	22,786	62.58
2	42	Aurora / ANL	1,012.00	38,698	26.15
3	-	Eagle / MS Azure	561.20	-	-
4	68	Fugal / RIKEN	442.01	29,899	15.41
5	12	LUMI / EuroHPC	379.70	7,107	53.42
6	14	Alps / CSCS	270.00	5,194	51.98

Efficiency

- Top systems are getting more energy efficient.
- More science, but not less energy.

Green 500	#1 System Efficiency [GF/W]	Approx. Relative Change [%]
2019 - June	15.1	-
2019 - Nov.	16.9	11
2020 - June	21.1	20
2020 - Nov.	26.2	19
2021 - June	29.7	12
2021 - Nov.	39.4	25
2022 - June	62.7	37
2022 - Nov.	65.1	4
2023 - June	65.4	0
2023 - Nov.	65.4	0
2024 - June	72.7	10

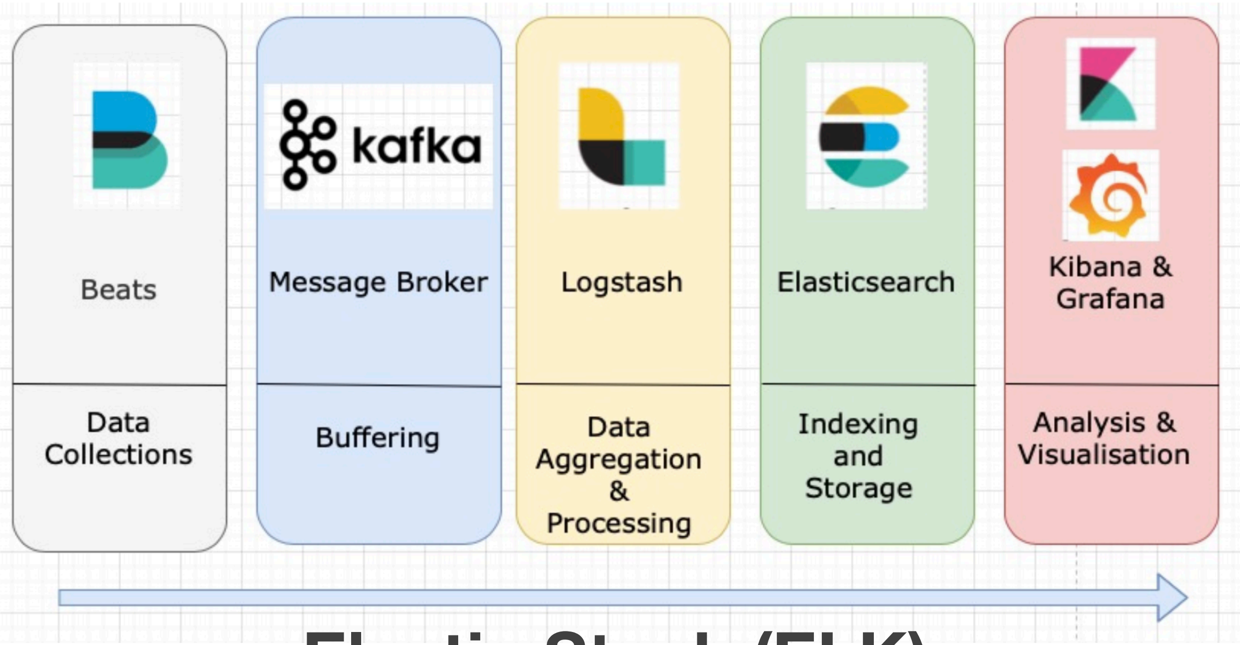
Extensible Monitoring and Observability Infrastructure

DWDI

DATA WAREHOUSE DATA INTELLIGENCE

EMOI aims to minimize the burden on human operators who manage the **deployment**, **monitoring**, **maintenance**, and **upgrading** of the observability infrastructure.

EMOI Observability Cluster Components

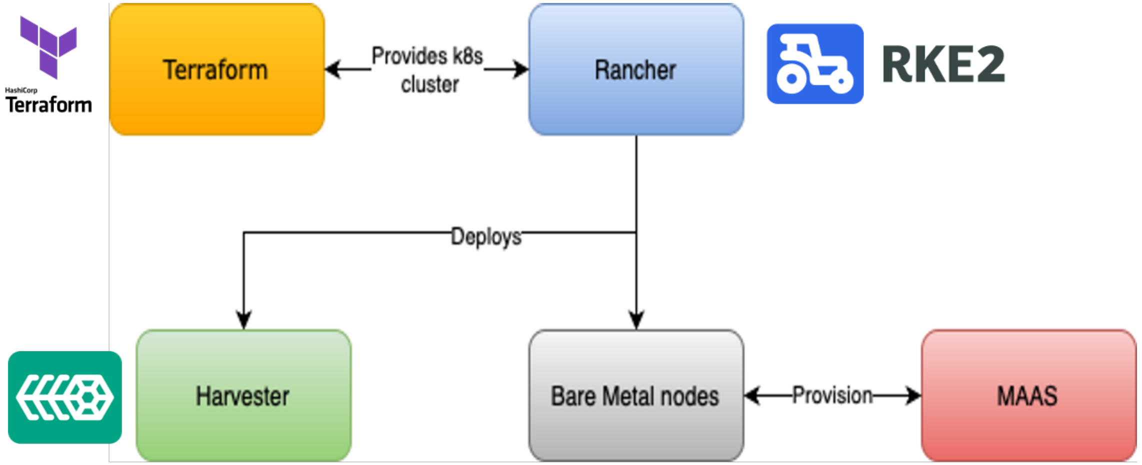


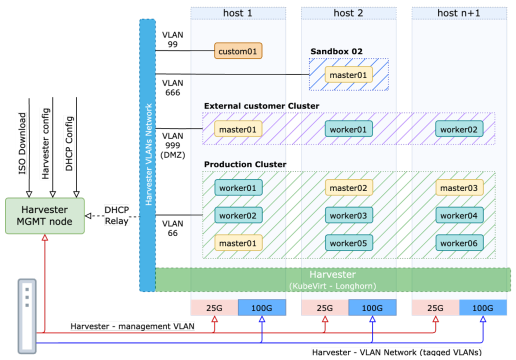
Elastic Stack (ELK)

- **Beats**: lightweight data shippers
- **Kafka**: message broker, push model, implements streaming telemetry and acts as a buffer
- **Logstash**: data processing pipeline
- **Elasticsearch**: distributed search and analytics engine designed for storing large volumes of data
- **Kibana & Grafana**: visualization tools, build dashboards, view and analyze data

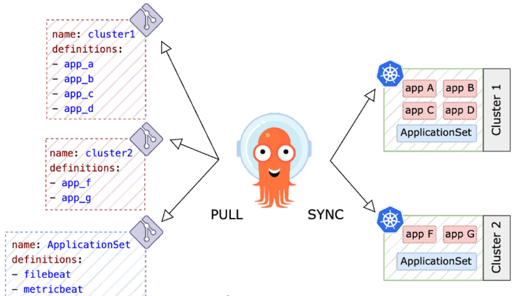
Dynamic deployments of an Observability Cluster

- **Flexible**: Multiple physical or virtual Kubernetes cluster dynamically deployed to accommodate custom workflows or external customers
- **Scalable**: provide horizontal scalability to meet changing demands
- **Automated**: apply Infrastructure-as-Code (IaC) principles and git-ops approach

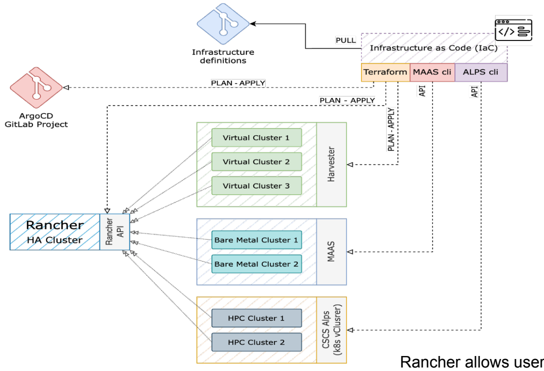


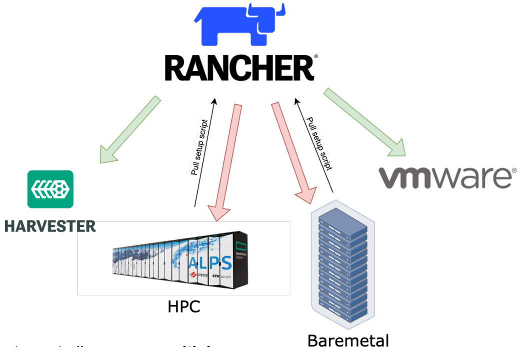


Harvester uses **Kubernetes** as its orchestration engine, allowing for effective management of resources and workloads



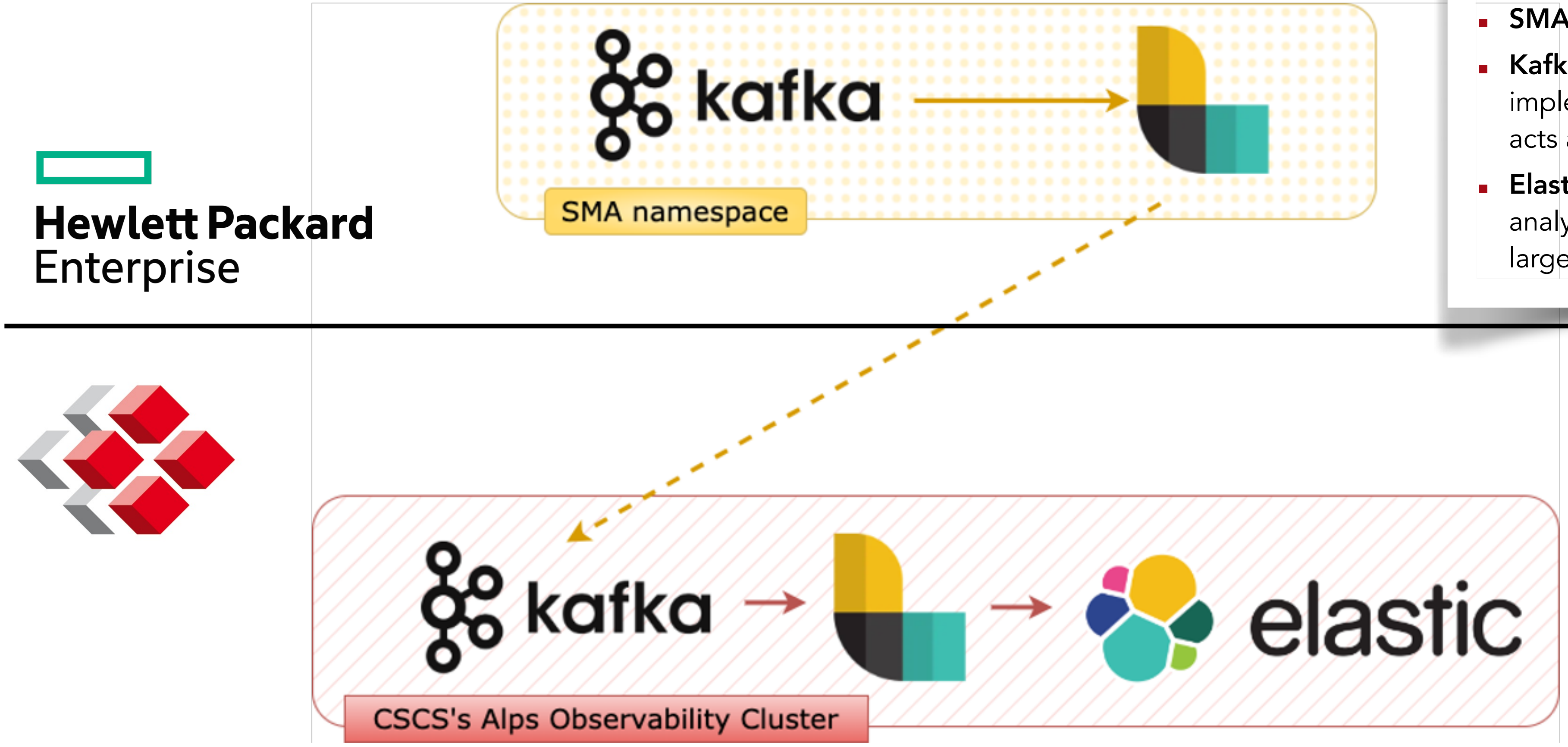
Argo CD is an open-source continuous delivery tool for Kubernetes following a GitOps methodology.





Rancher allows users to centrally manage **multiple Kubernetes clusters**, regardless of their location or provider, from a single platform.

Integration with HPE's CSM-SMA Kafka Bus



- **CSM**: Cray Service Management
- **SMA**: System Monitoring Application
- **Kafka**: message broker, push model, implements streaming telemetry and acts as a buffer
- **Elasticsearch**: distributed search and analytics engine designed for storing large volumes of data

Collecting Energy Data

The energy usage at Node level can also be accessed with the Slurm **`sacct`** command.

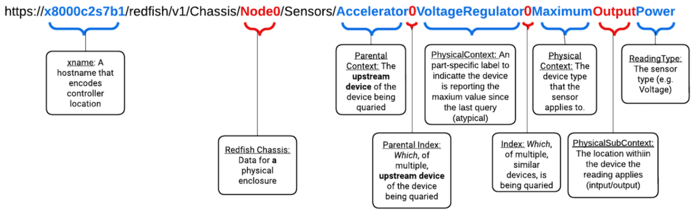
Cray Power Measurement data: Consumed Energy at Node, CPU and GPU levels can be read from **`/sys/cray/pm_counters/`** sysfs files. Default collection rate is 10 Hz.

Telemetry data: HPE/Cray sensors are published via the Redfish restful API, using the Sensor schema. Default collection rate is 1 Hz.

```
┌ santis:santis-ln001 /users/bcumming
└── cat /sys/cray/pm_counters/accel0_energy
73715541 J 1714148567299497 us
```

	node	memory	cpu	cpu0	cpu1	cpu2	cpu3	accel0	accel1	accel2	accel3
zen2/zen3	+	+	+								
a100	+	+	+					+	+	+	+
gh200	+		+	+	+	+	+	+	+	+	+
mi250x	+	+	+					+	+	+	+

- Specific counters are available on different node types.
- Provide accumulated and instantaneous power readings for components on a node.



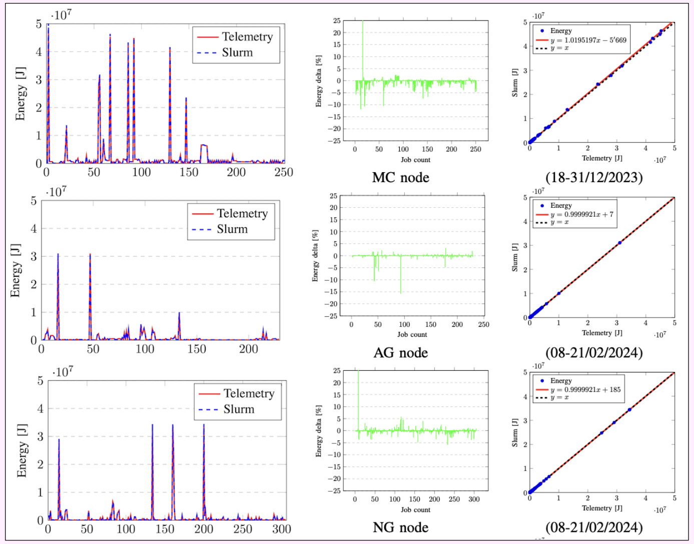
x1002c3s7b1n0

- **x** [cabinet]
- **c** [chassis]
- **s** [blade]
- **b** [bmc controller]
- **n** [node]

Validating Energy Data

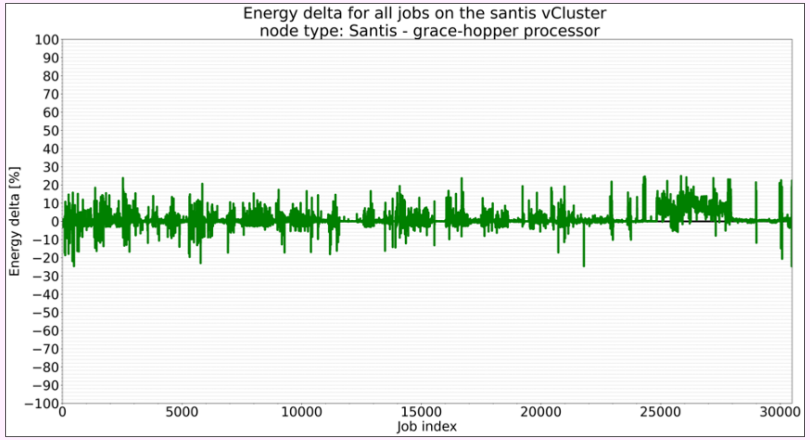
We validate data by comparing the energy data collected from slurm/pm_counters (sysfs) with the data collected from telemetry (redfish).

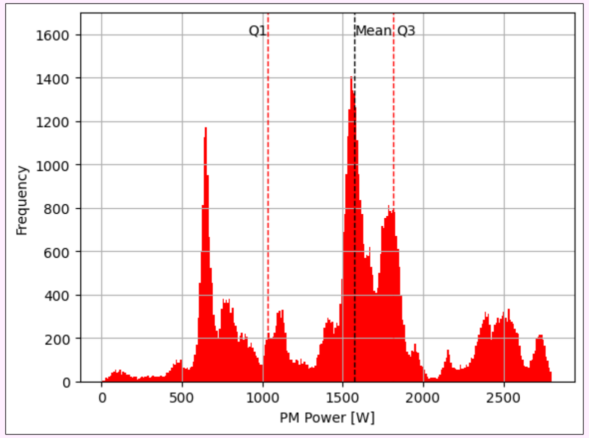
Non-GraceHopper Nodes

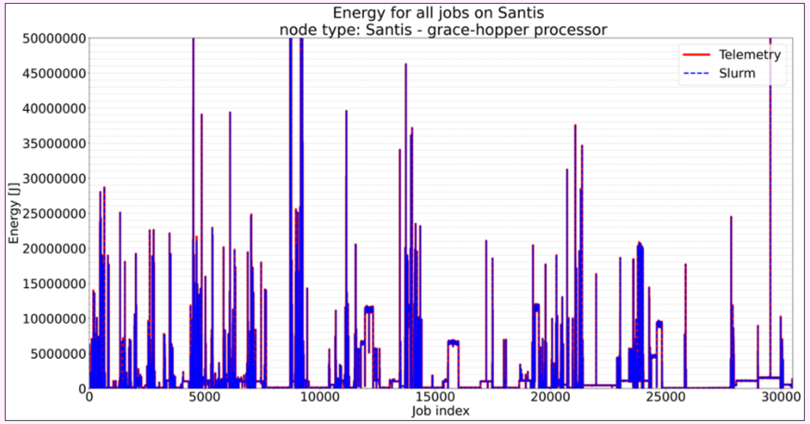


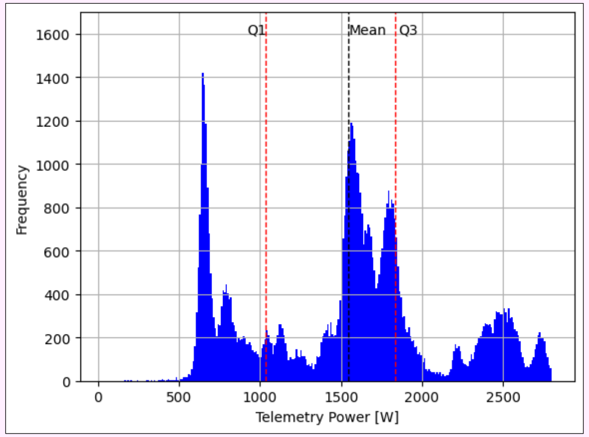
Node Architecture	CPUs / Node	GPUs / Node
EX425 Windom (MC)	2 AMD 64-core	0
EX325A Bard Peak (AG)	1 AMD 64-core	4 AMD MI200
EX325N GrizzlyPeak (NG)	1 AMD 64-core	4 NVIDIA A100

GraceHopper Nodes









Slurm vs Telemetry

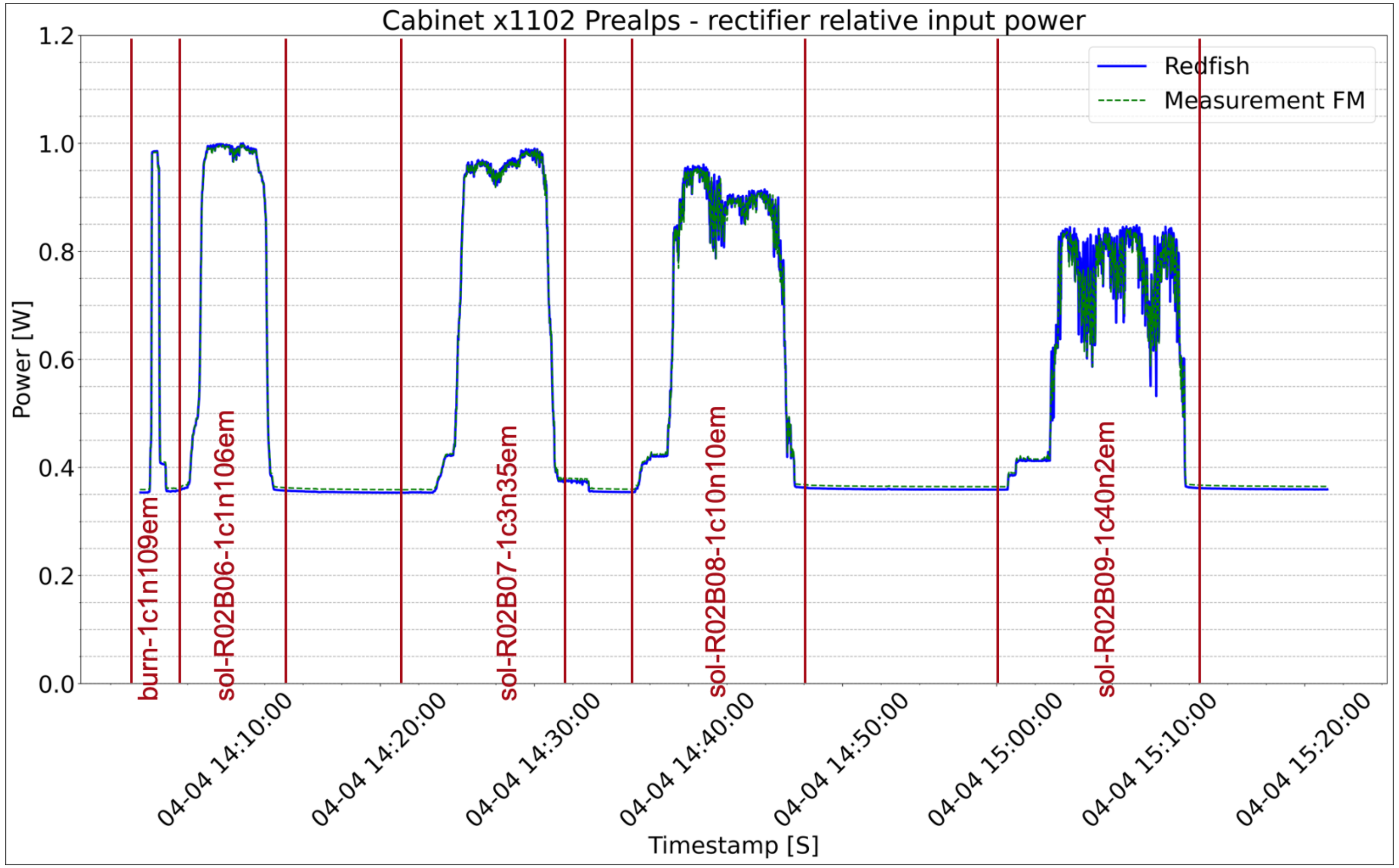
83,845 jobs: a mix of small, medium, and large power-intensive jobs

CSCS

ETH zürich

Validating Energy Data

We validate data by comparing the energy data collected from telemetry (redfish) to a physically attached power meter.



Understanding Power Consumption

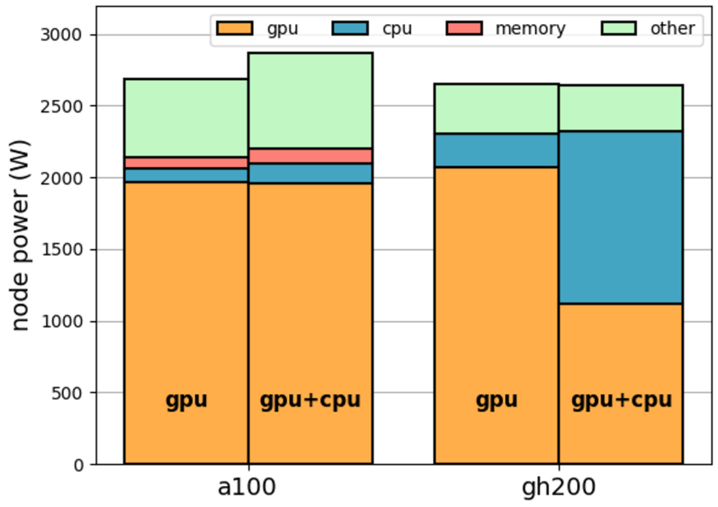
CSCS developed nodeburn

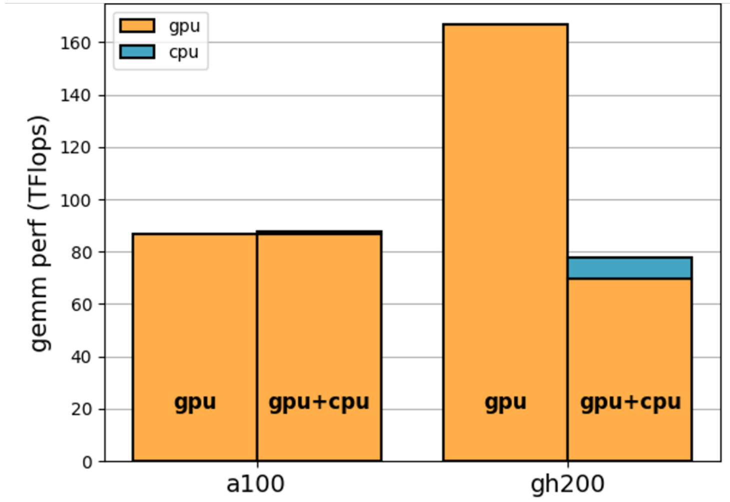
- Runs DGEMM back-to-back on the CPU and GPU
- Measures CPU and GPU performance
- Measures CPU and GPU power consumption
- github.com/eth-cscs/node-burn

```
santis:santis-ln001 /bret/scratch/cscs/bcumming/software/node-burn/build
└── OMP_NUM_THREADS=64 srun -n4 -N1 ./burn -ggemm,16000 -cgemm,6000 -d180 --batch
nid005898:gpu     347 iterations, 15785.09 GFlops,   180.1 seconds,     6.144 Gbytes
nid005898:cpu     860 iterations,  2063.85 GFlops,   180.0 seconds,     0.864 Gbytes
nid005898:gpu     348 iterations, 15828.03 GFlops,   180.1 seconds,     6.144 Gbytes
nid005898:cpu     858 iterations,  2059.10 GFlops,   180.0 seconds,     0.864 Gbytes
nid005898:gpu     413 iterations, 18775.46 GFlops,   180.2 seconds,     6.144 Gbytes
nid005898:cpu     846 iterations,  2030.11 GFlops,   180.0 seconds,     0.864 Gbytes
nid005898:gpu     426 iterations, 19368.84 GFlops,   180.2 seconds,     6.144 Gbytes
nid005898:cpu     838 iterations,  2008.96 GFlops,   180.2 seconds,     0.864 Gbytes
nid005898:power [total 2639, cpu 1203, gpu0 290, gpu1 292, gpu2 266, gpu3 269, cpu0 304, cpu1 304, cpu2 304, cpu3 292]
```

Experiment on A100 and GH200 nodes:

- Run GPU and CPU+GPU DGEMM
- GH200 had power cap of 620W





On A100 (Grizzly Peak)

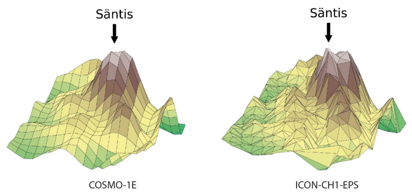
- GPU power consumption is unaffected by CPU workload
- CPU TDP is separate from GPU TDP
- The A100 GPU throughput is unaffected by CPU workload.

On GH200 (Blanca Peak)

- CPU+GPU power consumption is constant (~2,600W total)
- The CPU's power requirements are *prioritised* over the GPU
- This is "power sloshing" or "**power steering**"

Using 72 Grace cores halves H100 performance!

Measuring real-world applications



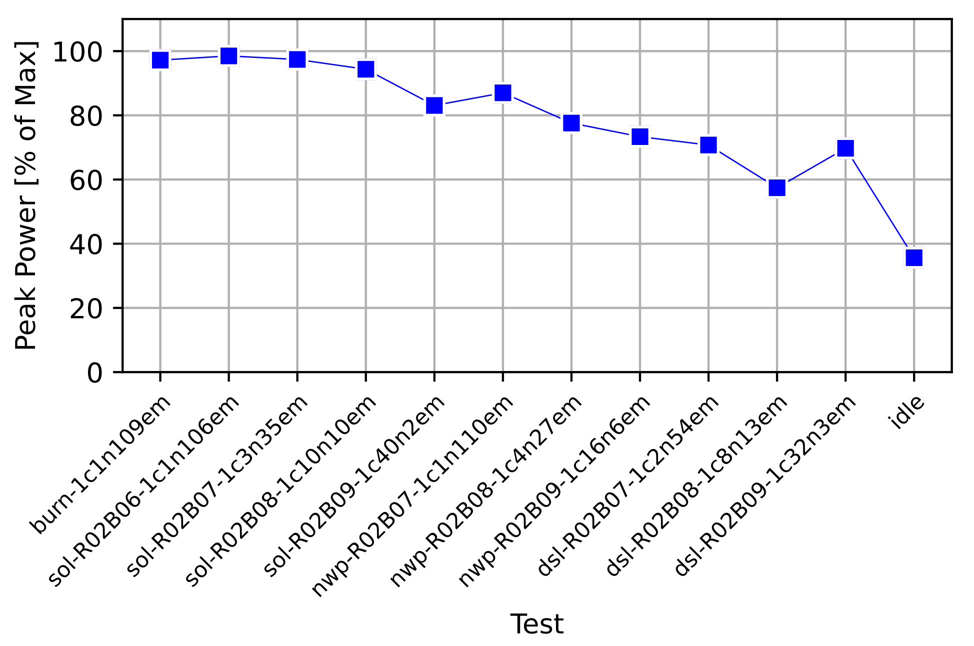
The Icosahedral Nonhydrostatic Weather and Climate Model

- ICON is an atmospheric model used for numerical weather prediction and climate modeling

Implementations		
NWP Standard Version Numerical Weather Prediction	**SOL** NVIDIA "speed of light" Implemented with CUDA	**DSL** Python "Domain Specific Language" generated stencils

Experiment: Fill a single cabinet with ICON

- On Daint, ICON uses ~180W/node → Relatively low ~50% of DGEMM
- On GH nodes, ***ICON consumes >80% of peak*** (peak ~ 2600W / node)

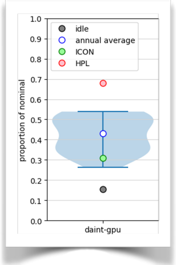


Model	Resolution
R2B6	40 km
R2B7	20 km
R2B8	10 km
R2B9	5 km

Concluding Remarks & Outlook

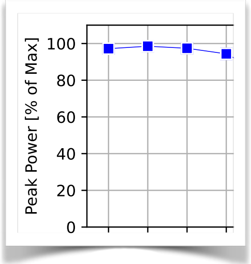
The Old:

- “Normal workloads” consumed on average 40% of nominal power
- ICON consumption was 30%
- Even nodeburn or HPL (High-Performance Linpack) could only achieve 70%



The New:

- We expect large well-optimised codes to use > 90 %
- We have to be prepared for normal workloads to consume peak power.
- CPUs and GPUs are no-longer independent entities.



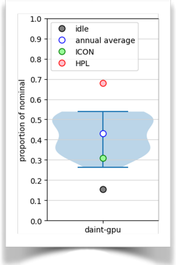
Tools like EMOI provide critical insight into how applications behave and consume power.

This insight will be needed to develop efficient and power conscious applications.

Concluding Remarks & Outlook

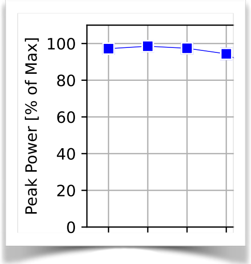
The Old:

- "Normal workloads" consumed on average 40% of nominal power
- ICON consumption was 30%
- Even nodeburn or HPL (High-Performance Linpack) could only achieve 70%



The New:

- We expect large well-optimised codes to use > 90 %
- We have to be prepared for normal workloads to consume peak power.
- CPUs and GPUs are no-longer independent entities.



Tools like EMOI provide critical insight into how applications behave and consume power.

This insight will be needed to develop efficient and power conscious applications.