EESP 2026 Workshop

November 15, 2026

DOMAIN-AWARE POWER REBALANCING UNDER NODE-LEVEL POWER CAPS ON AURORA

Solomon Bekele[1], Nicholas John Eliopoulos[2], Swann Perarnau[1], Servesh Muralidharan[1], and Brice Videau[1]

[1] ARGONNE NATIONAL LABORATORY, LEMONT, IL, USA
[2] PURDUE UNIVERSITY, WEST LAFAYETTE, IN, USA

U.S. DEPARTMENT of ENERGY

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

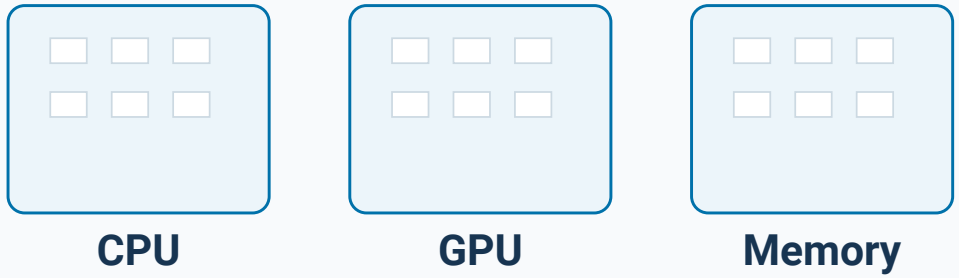
Argonne NATIONAL LABORATORY

From chip-level controls to scalable node power caps

Power and energy constrain HPC operating cost, cooling, and system capacity.
Fine-grained controls work inside chips. Facilities need power limits that scale across the machine.

INSIDE A NODE

Built-in device controls

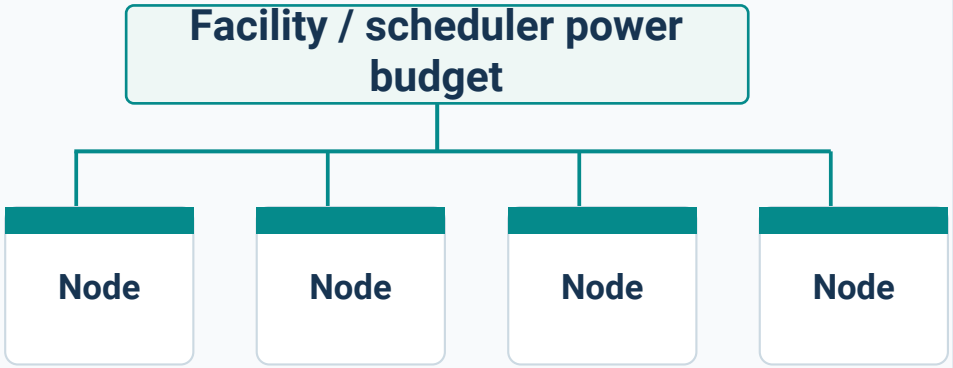


DVFS Clock gating Power gating

→ ×N

AT FACILITY SCALE

One enforceable cap per node



Simple to budget, schedule, and enforce

The challenge: coordinating every CPU, GPU, and memory-domain policy centrally does not scale.

The remaining question: how is the capped power shared inside a node?

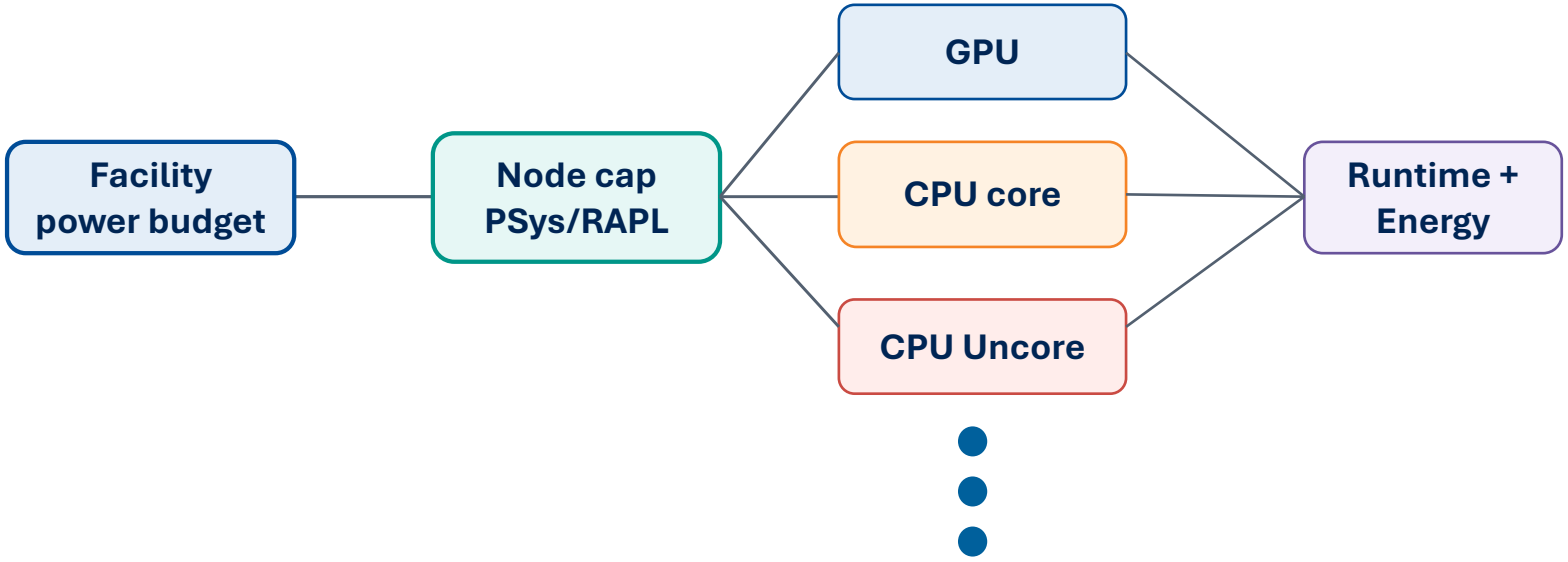
U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

The hidden problem: power cap ≠ power allocation

Facility view
“Keep each node within a power budget.”

Runtime reality
“Which domain gets the limited budget?”

Energy outcome
Runtime × power determines total board energy.



On heterogeneous exascale nodes, a board-level cap can reduce instantaneous power while still wasting energy if firmware under-provisions the progress-critical domain.

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Aurora platform and experimental setup

GEOPM-based telemetry and control across a heterogeneous compute node.

Setup

Aurora compute node

Intel Xeon CPU Max 52 cores	GPU 1	GPU 2	GPU 3
Intel Xeon CPU Max 52 cores	GPU 4	GPU 5	GPU 6

6 × Intel Data Center GPU Max 1550

Power Controls

Node power cap
1,800–3,500 W

GPU DVFS
800–1,600 MHz

CPU uncore DVFS
800–2,300 MHz

SPEChpc2021

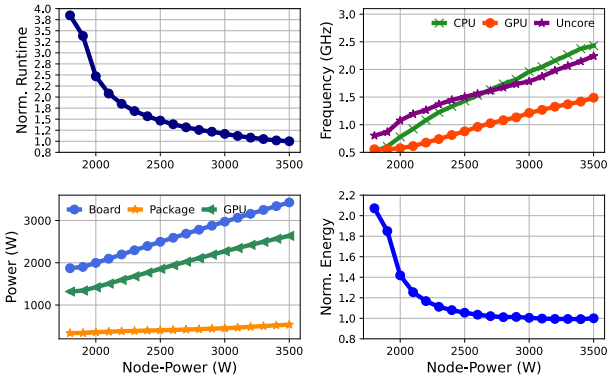
Aurora exposes multiple domain controls, and GEOPM makes them measurable and controllable for systematic power/energy experiments.

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

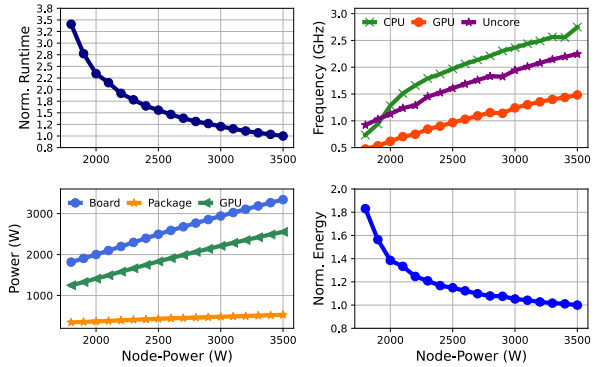
Node caps are scalable, but allocation is opaque

Finding

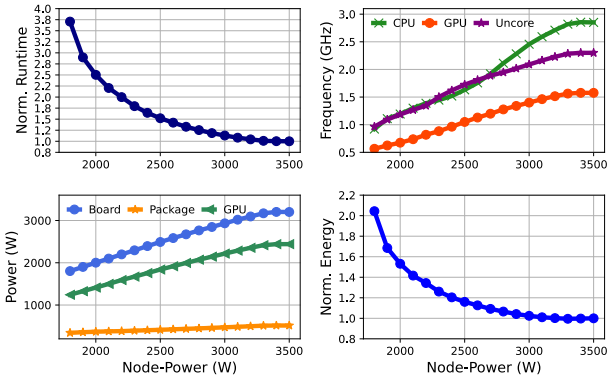
519.clvleaf_t



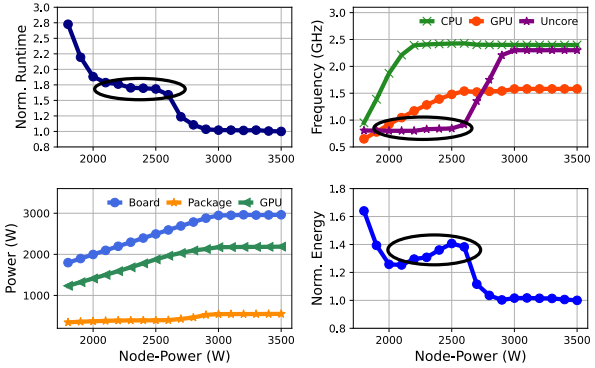
513.soma_t



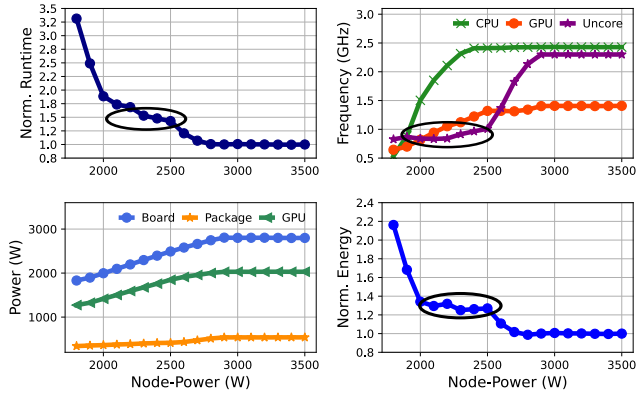
505.lbm_t



535.weather_t



518.tealeaf_t



General trend

Increasing the node cap usually improves runtime.

Some exceptions

At low caps, the CPU Uncore can remain pinned near its minimum.

Energy consequence

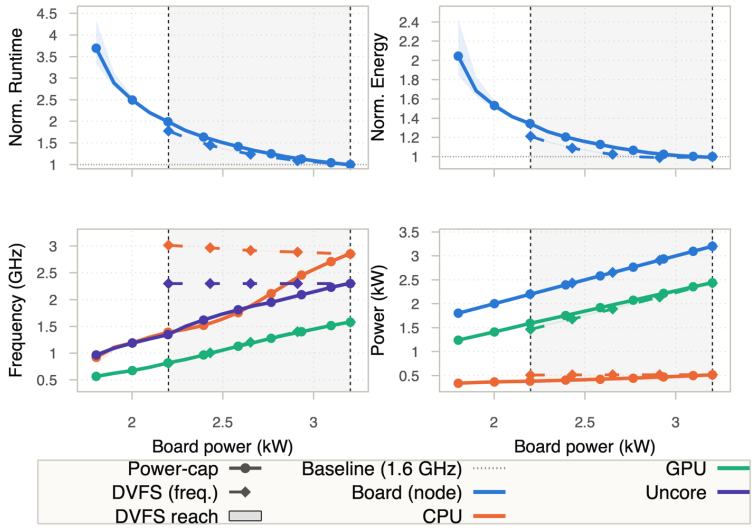
The runtime stretch can outweigh the lower instantaneous power.

Reducing power is not sufficient; the limited budget must reach the domain that controls application progress.

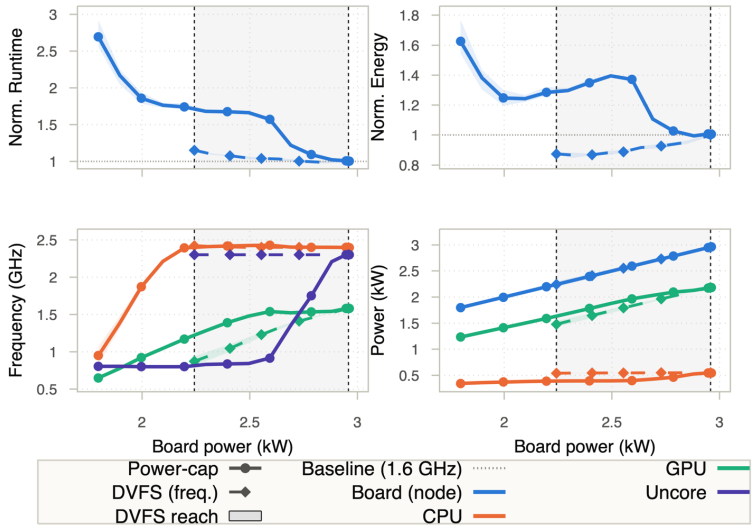
U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Diagnosis: equal board power can still have different energy outcomes

505.lbm_t



535.weather_t



Observation

GPU DVFS and node caps can reach similar board power.

Difference

The uncore can be retained under GPU DVFS but reduced under a node cap.

Control opportunity: keep the node cap active, but nudge the domain allocation when data movement suggests higher uncore frequency is valuable.

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

U-PullUp: cap-preserving uncore biasing plug-in agent

Use GEOPM telemetry to decide when to request higher CPU uncore frequency while the node power cap remains active

U-PullUp

Problem

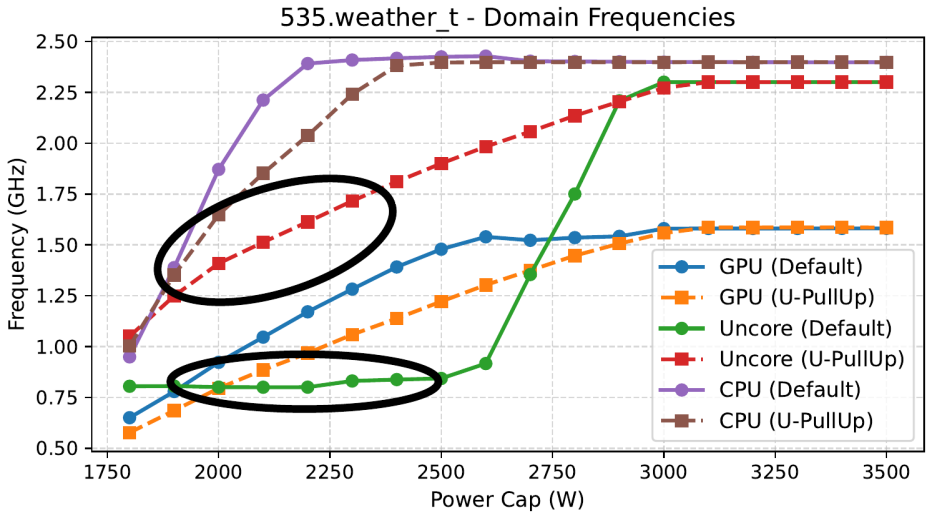
Under a low node cap, firmware can under-provision CPU uncore for data-movement-sensitive phases.

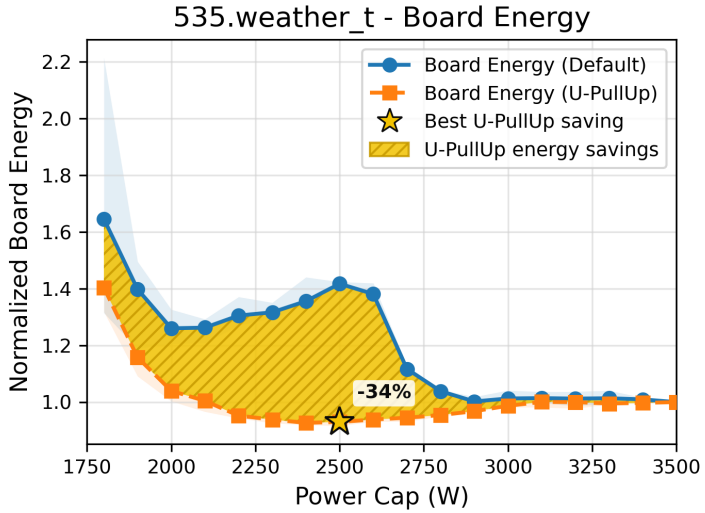
Runtime signal
copy-engine activity ≈ data movement

Power condition
cap below low-power threshold

Action
request max CPU uncore frequency

GEOPM keeps privileged controls mediated; PSys/RAPL cap still bounds total board power.





The mechanism is complementary, not a replacement for platform power management.

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

All nine SPEChpc workloads: measured best and worst outcomes

Workload	Best saving	Cap	Worst	Cap
535.weather_t	34.2%	2.5 kW	+0.3%	3.5 kW
534.hpgmgfv_t	30.6%	1.8 kW	-1.0%	3.4 kW
528.pot3d_t	28.1%	1.8 kW	-1.9%	3.1 kW
521.miniswp_t	25.5%	1.8 kW	-1.5%	2.7 kW
518.tealeaf_t	21.5%	2.5 kW	-1.9%	3.3 kW
519.clvleaf_t	17.6%	1.9 kW	-2.0%	2.4 kW
532.sph_exa_t	13.3%	1.8 kW	-1.5%	2.7 kW
505.lbm_t	4.6%	2.0 kW	-3.5%	2.4 kW
513.soma_t	1.1%	1.8 kW	-3.4%	2.5 kW

Mean of per-workload BEST savings: 19.6% | Mean of per-workload WORST savings: −1.8%

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Conclusions

Node caps alone do not determine energy efficiency

At the same board power, the CPU, GPU, and uncore can receive different shares of the budget.

U-PullUp improves the allocation when data movement matters

A GEOPM-guided uncore frequency request preserves the configured node cap.

The benefit is substantial—but workload dependent

Up to 34.2% energy savings for weather, 30.6% for hpgmgfv, and 28.1% for pot3d; small regressions also occur.

Next step: test broader workloads and phase-aware domain control.

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

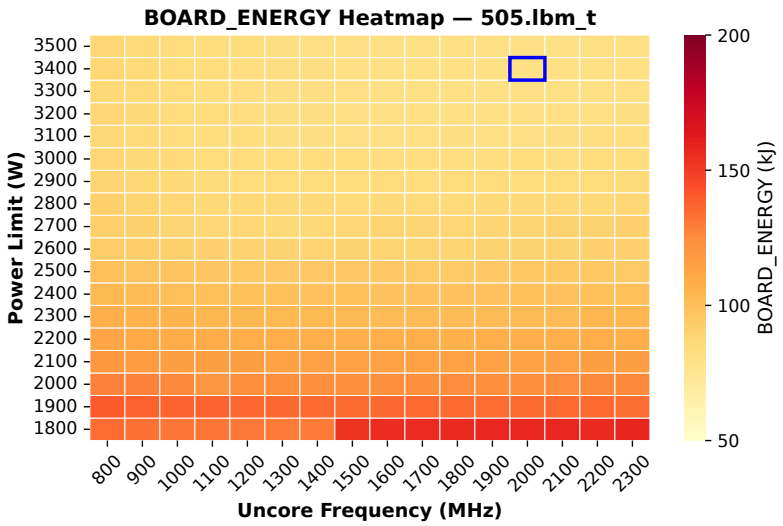
THANK YOU!

Argonne NATIONAL LABORATORY

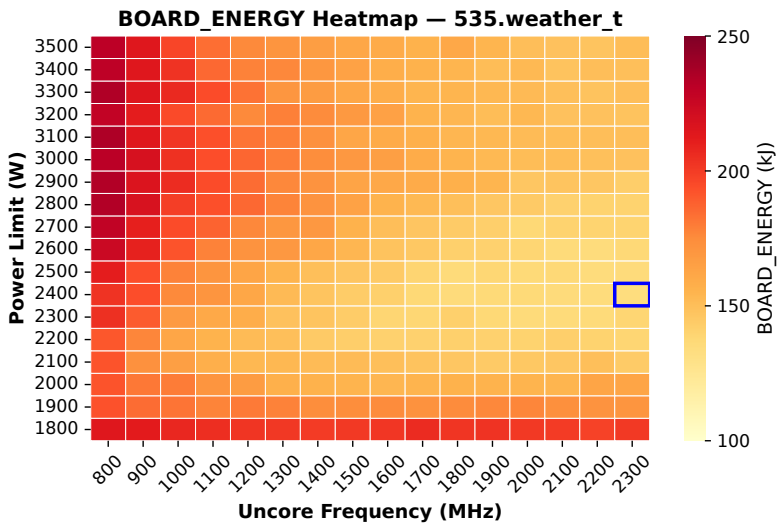
U.S. DEPARTMENT of ENERGY

This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory and is based on research supported by the U.S. DOE Office of Science-Advanced Scientific Computing Research Program, under Contract No. DE-AC02-06CH11357.

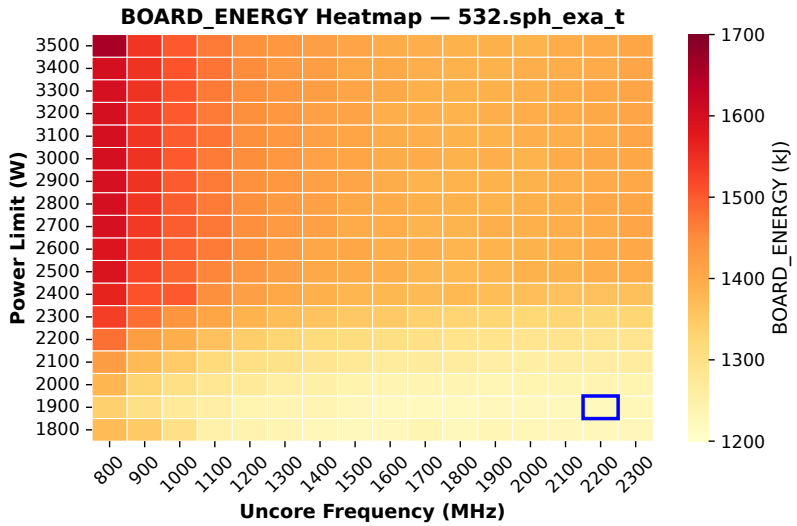
Joint cap–uncore measurements identify the missing resource



Mostly cap-sensitive



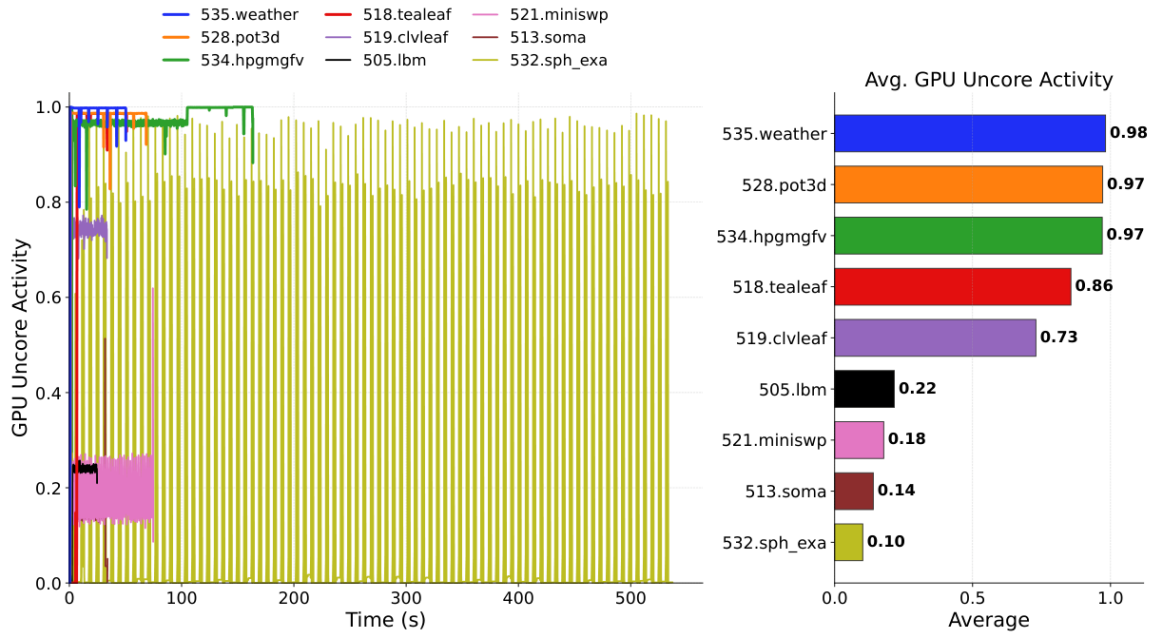
Strong uncore sensitivity



High uncore helps under low cap

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Measured GPU copy-engine activity provides a runtime signal



High mean activity

weather 0.98

pot3d 0.97

hpgmgfv 0.97

Low mean activity

lbm 0.22

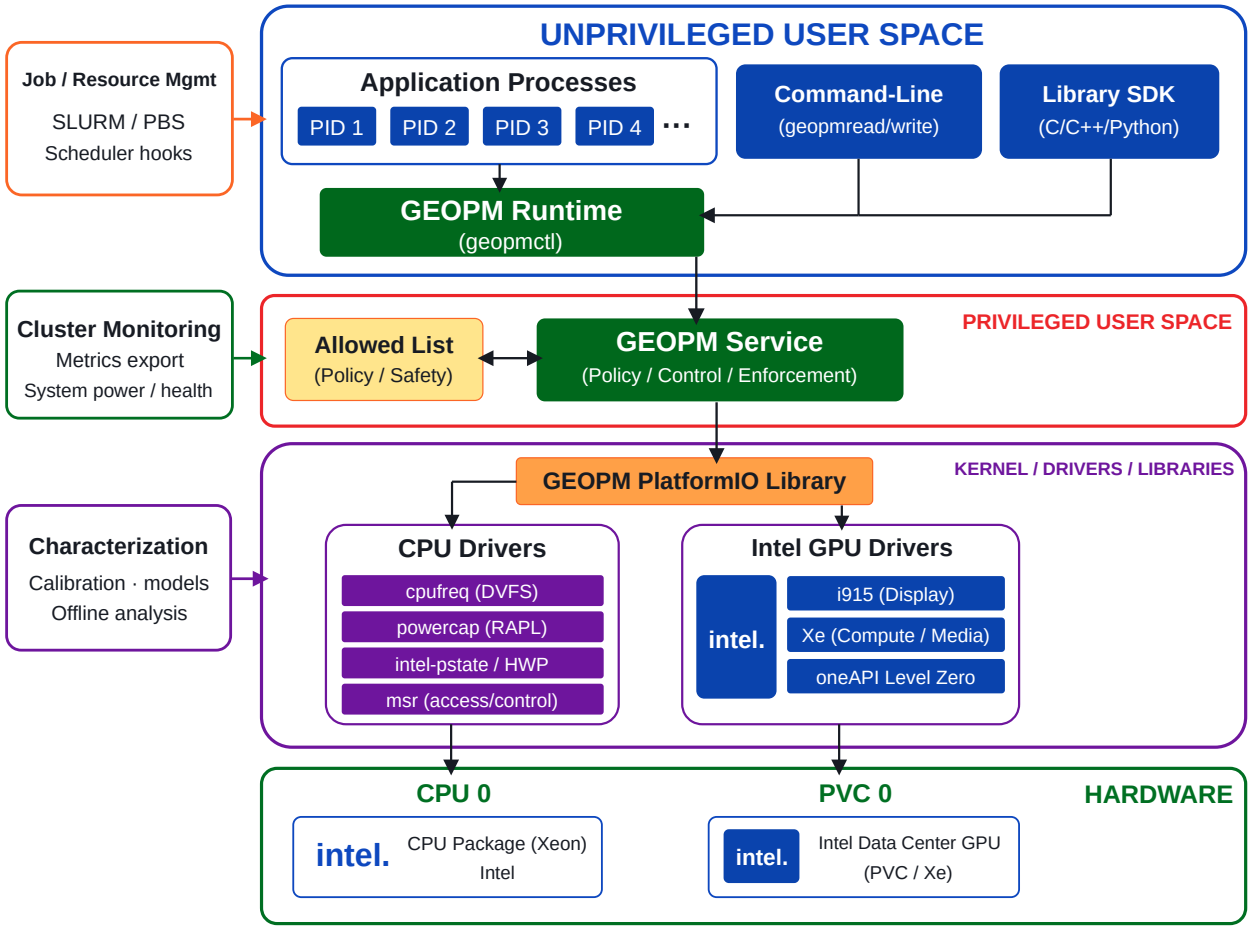
soma 0.14

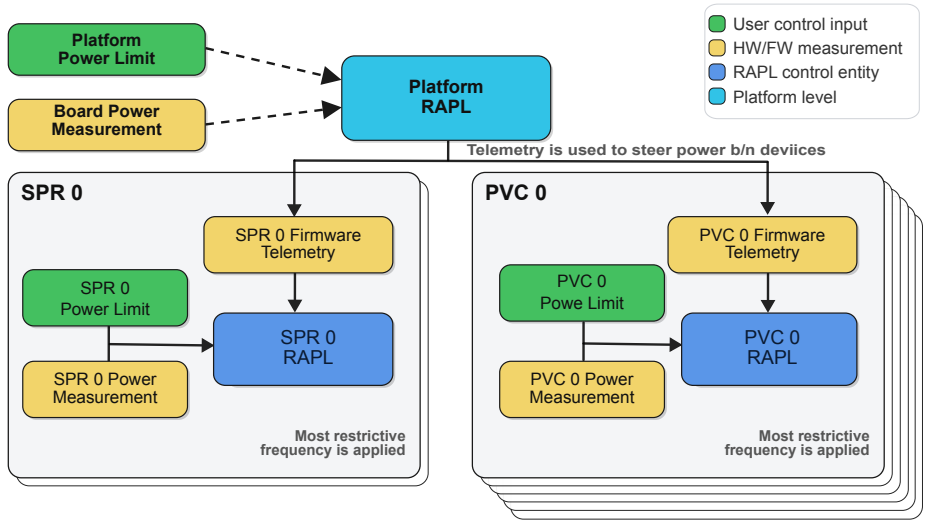
sph_exa 0.10

Source: Domain-Aware Power Rebalancing on Aurora, Fig. 7 (original measured data)

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Node power cap implementation on Aurora via Psys RAPL





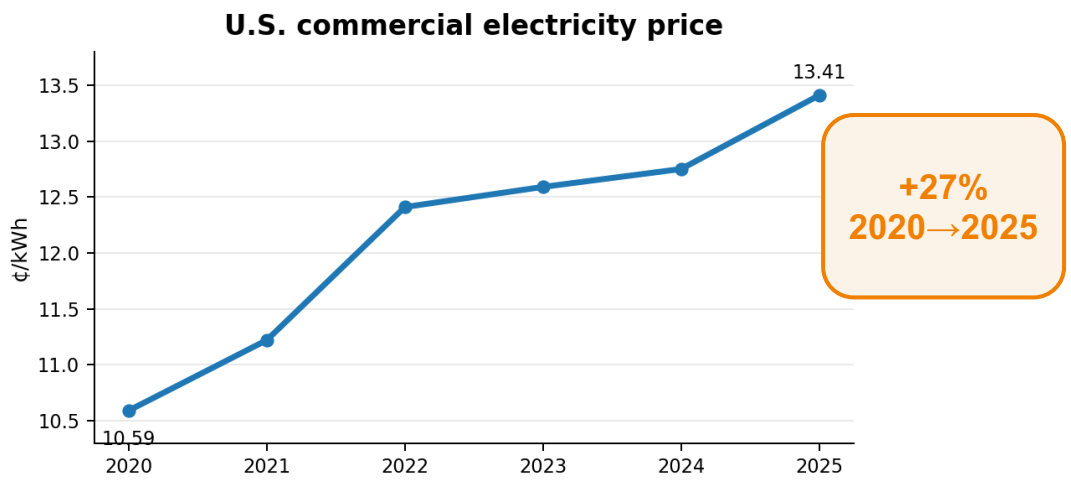
U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Power and energy are becoming first-order constraints

Before choosing a control mechanism, the pressure comes from cost, demand growth, and infrastructure limits

Motivation

Cost pressure

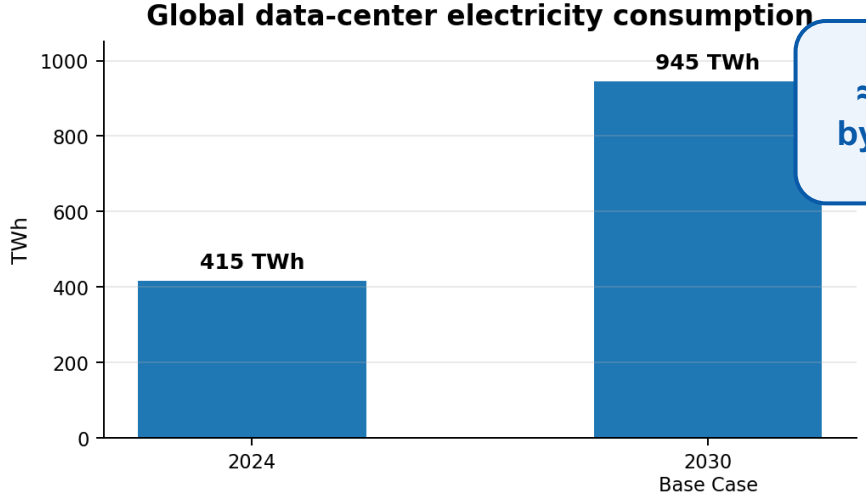


+27% 2020→2025

Commercial electricity prices rose from 10.59 to 13.41 ¢/kWh.

Source: EIA Electric Power Monthly, Table 5.3 (annual commercial average).

Demand pressure



≈2.3× by 2030

IEA projects data-center electricity use to more than double by 2030.

Source: IEA, Energy and AI (Base Case: ~415 TWh in 2024, ~945 TWh in 2030).

Motivation: energy is no longer just a local tuning metric — it is a cost, infrastructure, and sustainability constraint.

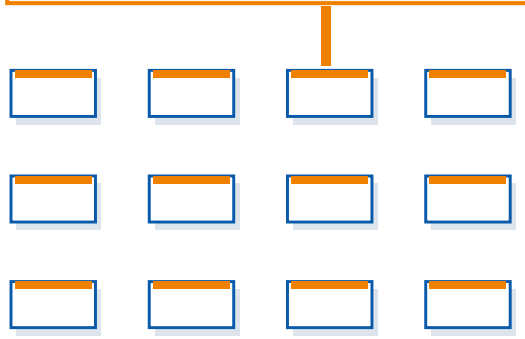
U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Why facilities need node-level power caps

A cap is not only an energy knob; it is a facility operating mechanism for power, grid, cooling, and scheduling constraints

Facility control view

Site power budget / capacity limit



One scalable policy: assign a cap per node

The facility can reason about the machine as a collection of bounded-power nodes, instead of micromanaging every device.

Representative facility use cases

SYS — System-level control
Keep total machine power under a site or system-level envelope.

GRID — Grid capacity
Respect utility interconnect, transformer, or contracted demand limits.

THERM — Thermal envelope
Prevent cooling overloads and reduce heat-density hot spots.

SCHED — Scheduler admission
Treat power as a schedulable resource, just like node count, for jobs and reservations.

COST — Peak-cost management
Reduce exposure to demand charges or expensive operating periods.

SAFE — Reliability margin
Avoid breaker trips, throttling cascades, and emergency shutdowns.

Node caps solve the facility-scale control problem, but they still leave a hidden in-node allocation problem.

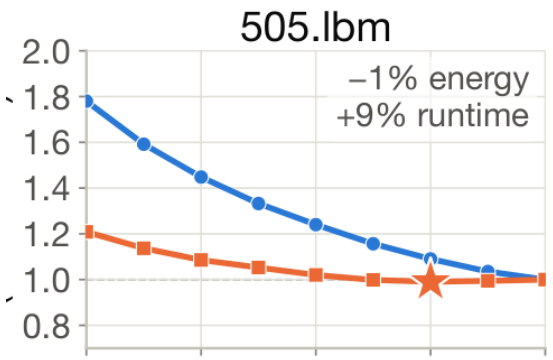
U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

Backup — Static GPU DVFS also depends on workload

Energy–runtime trade-offs from the original measurement sweep.

BACKUP

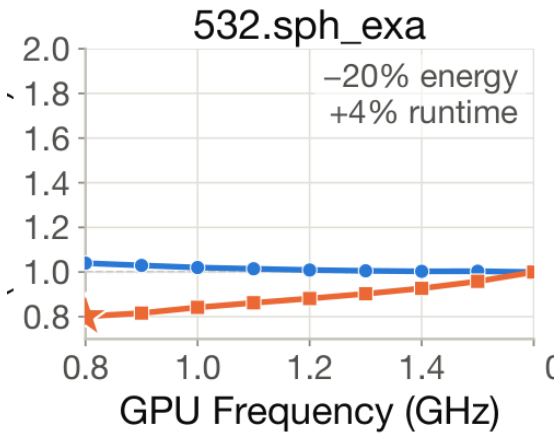
505.lbm_t



Best energy saving: 1.0%

GPU-compute intensive

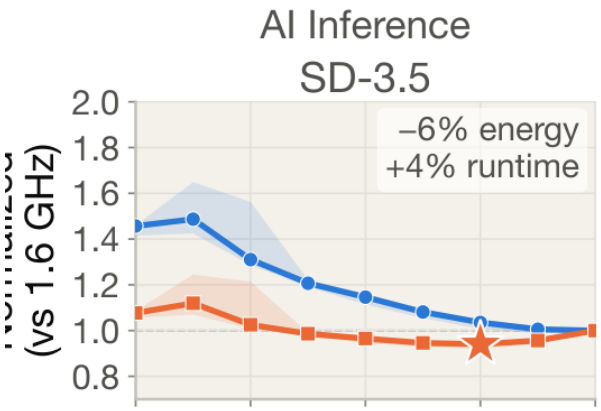
532.sph_exa_t



Best energy saving: 19.8%

Bursty / idle-rich

SD-3.5



Best energy saving: 6.0%

AI image generation

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY