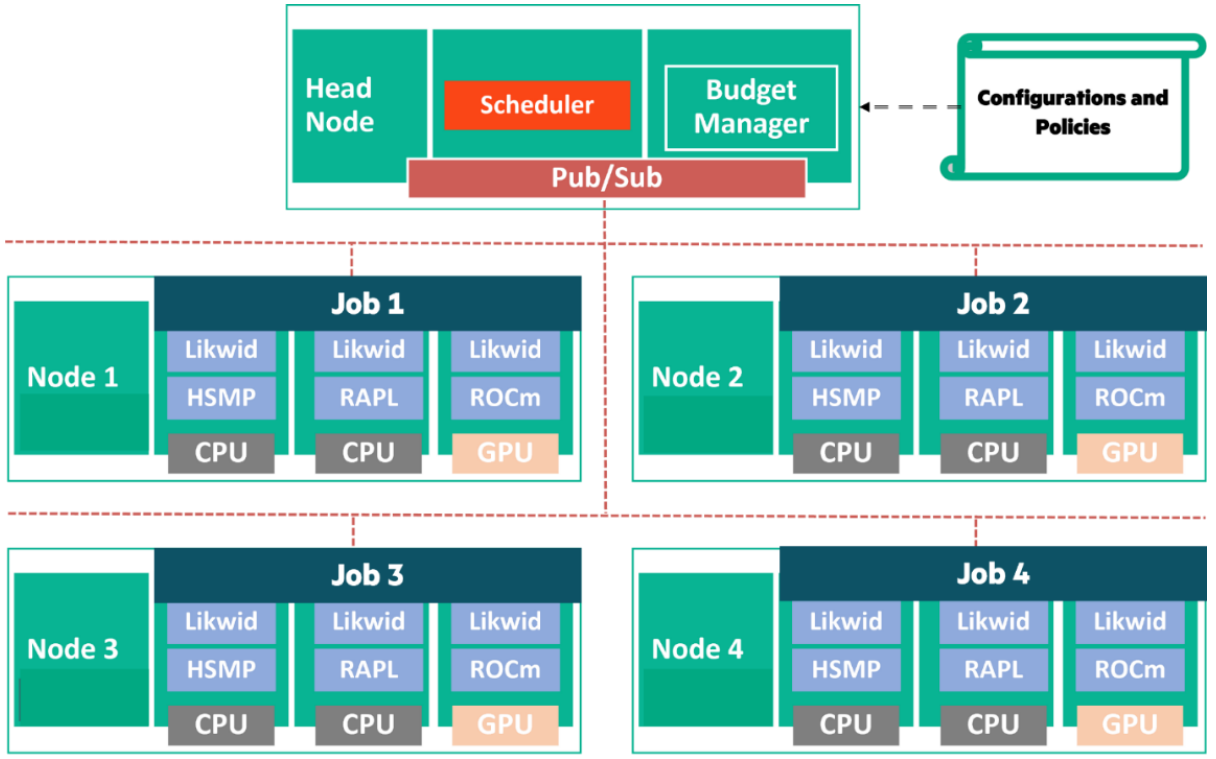
Energy and Cost-Aware Control at Scale: Multi-Dimensional Frequency Steering on AMD MI300A

Christian Simmendinger (HPE), Marcel Marquardt (HPE), Elias Klett (HPE), Ralf Schneider (HLRS)

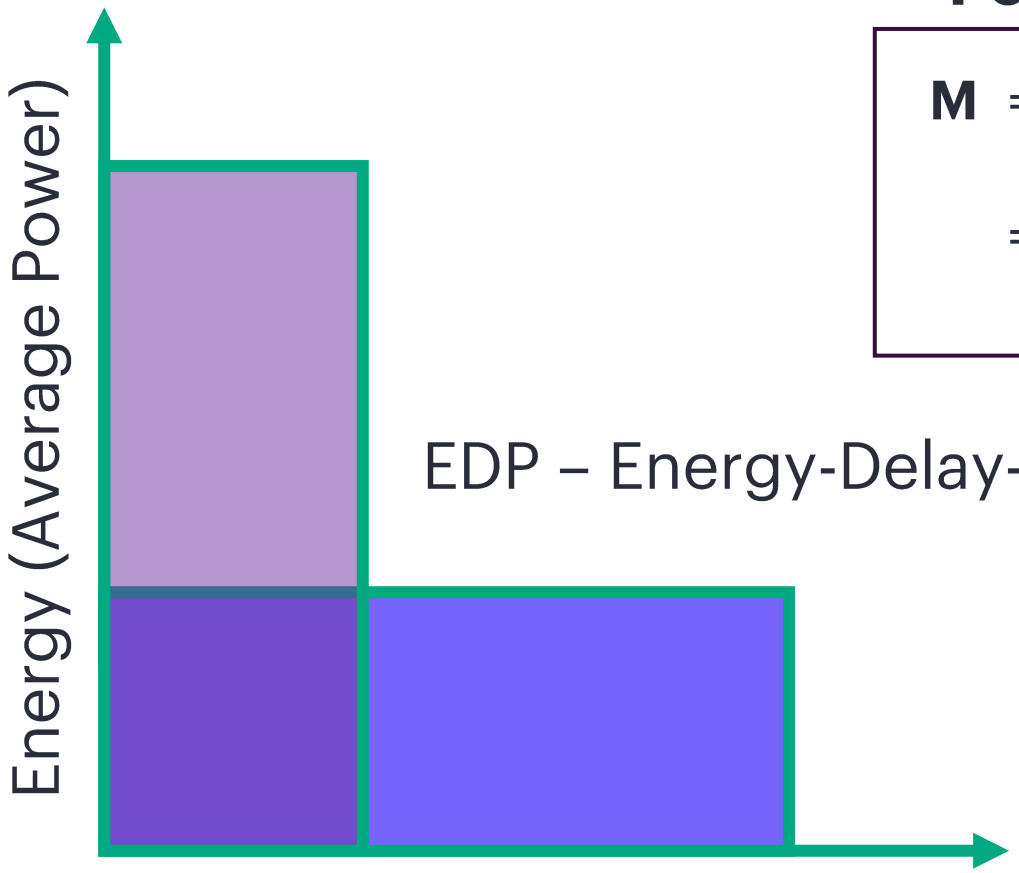
November 2026, EESP Workshop SC26

PowerSched Architecture

- Budget Manager
- Application Power Aware Runtime (APAR) daemon per node.
- Pub/Sub architecture
- Dynamic Global Policies



Dynamic Global Policies

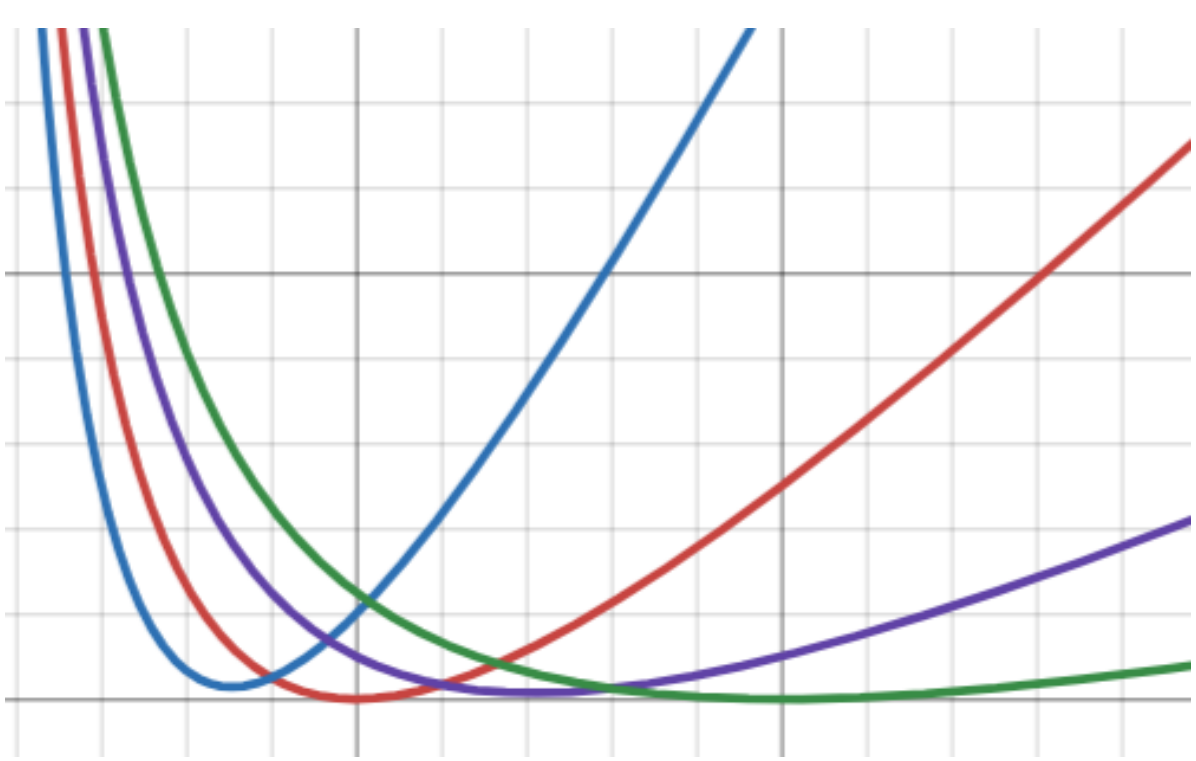


EDP – Energy-Delay-Product (Energy per Job)

Powersched Metric

$$\mathbf{M} = \text{ENERGY} \cdot \text{DELAY} = \frac{(\mathbf{C} + \text{ENERGY_PER_SOCKET})}{\text{MEMORY_ACTIVITY}}$$

Dynamic Global Policies



$$M = \frac{(C + ENERGY_PER_SOCKET)}{MEMORY_ACTIVITY}$$

Const factor C

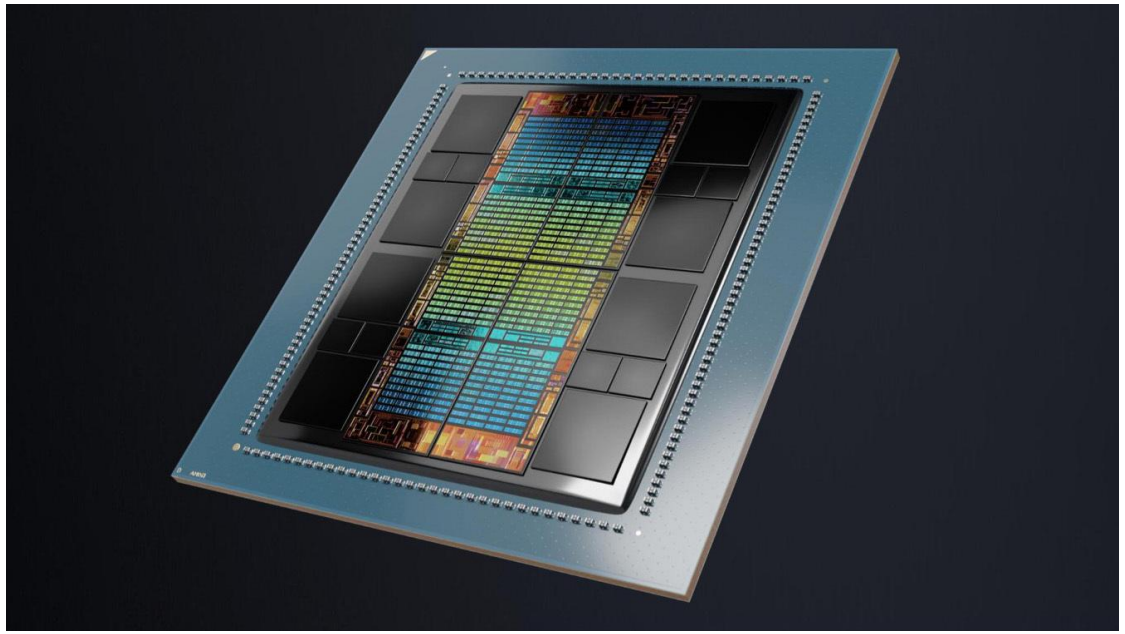
- accounts for static contributions from compute nodes and infrastructure.
- sets the tradeoff between improved energy efficiency and performance degradation of applications.

Multi-Dimensional Frequency Steering Search Algorithms

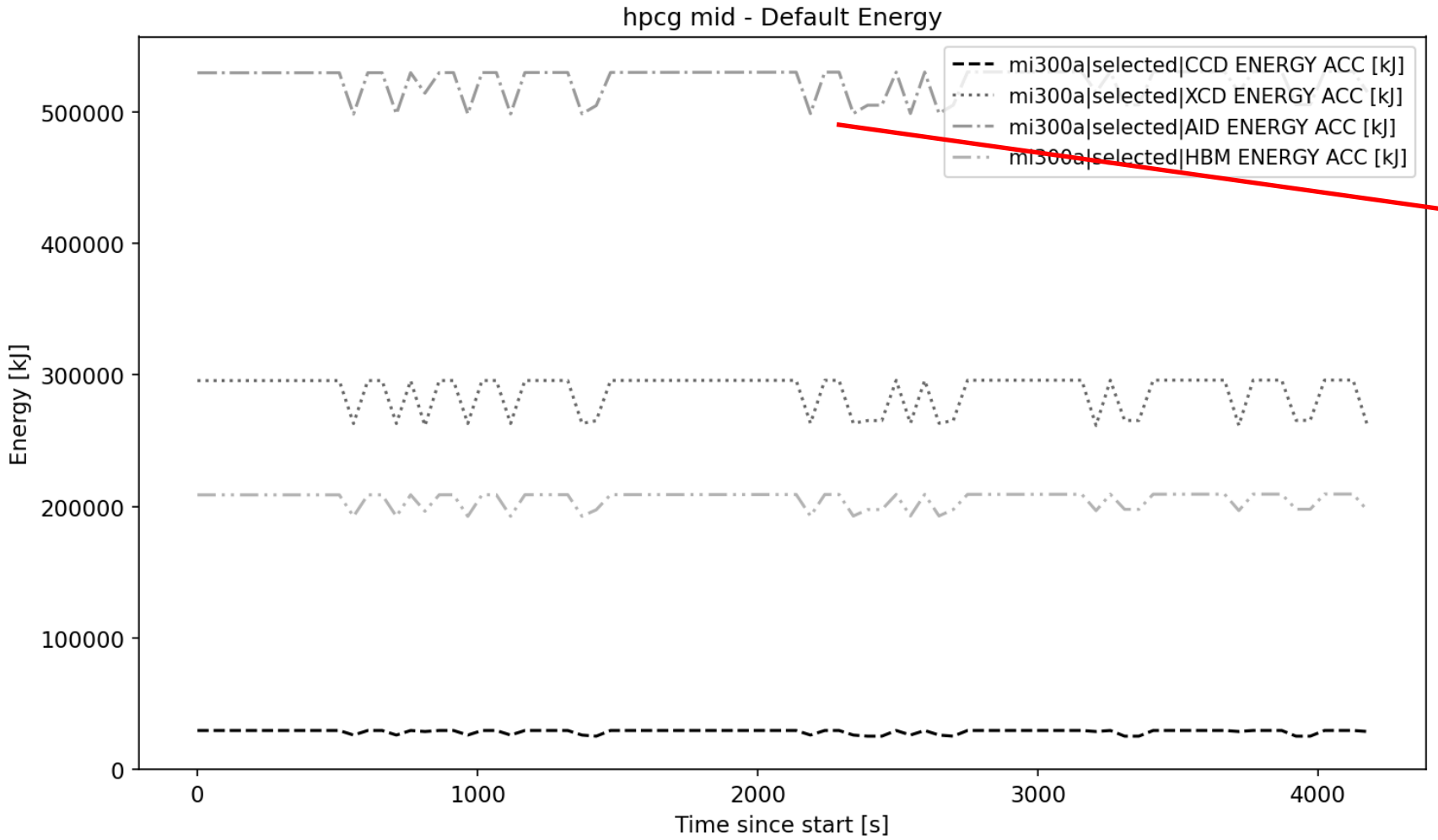
HLRS Hunter – HPE Cray EX4000 (AMD MI300A)

Multi Dimensional frequency and power steering of the AMD MI300A chip.

- **CPU boost frequency**
- **GPU shader frequency**
- GPU memory frequency
- Socket power
- **GPU interposer fabric clock**

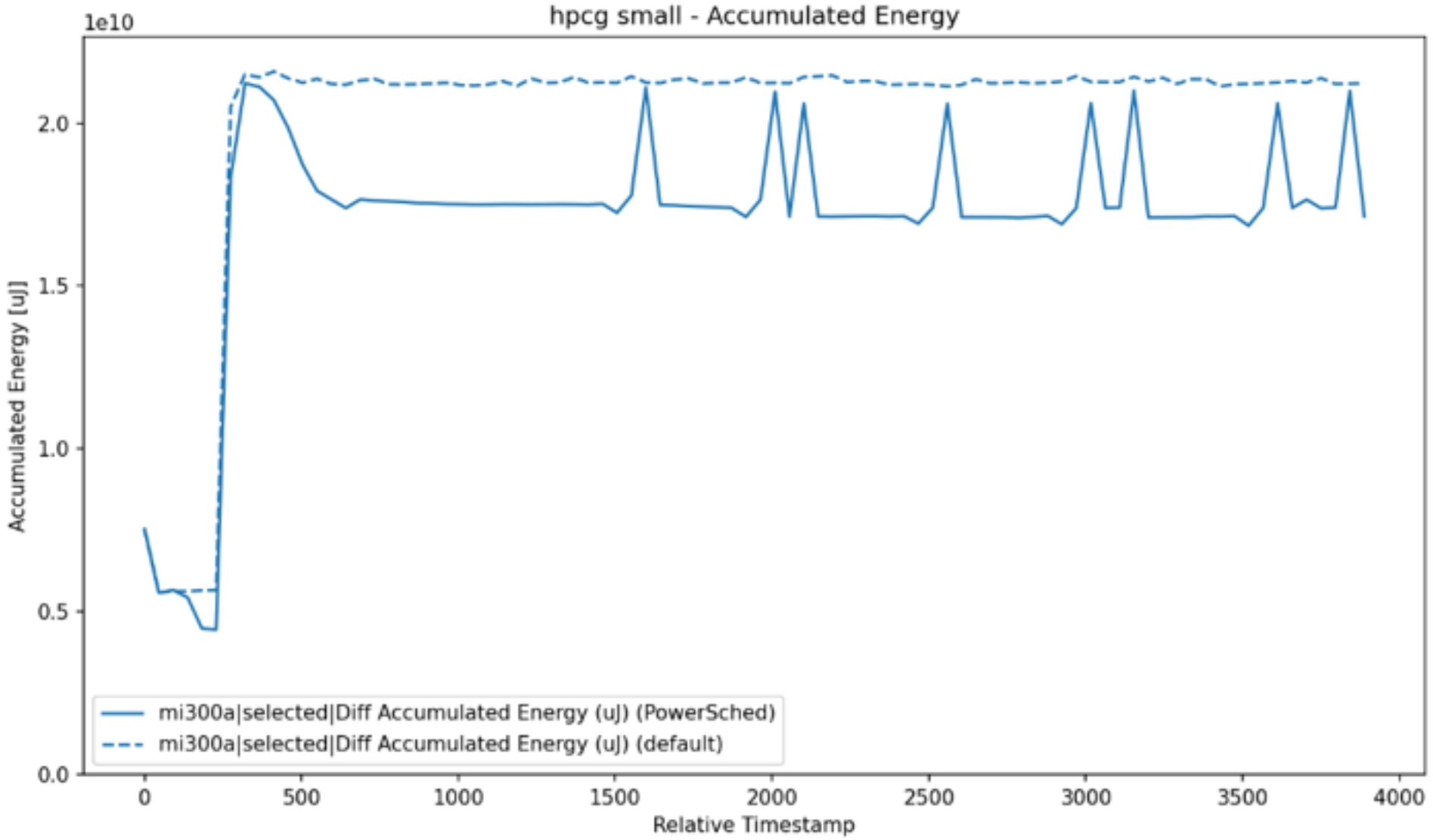


MI300A - Energy (rocHPCG)



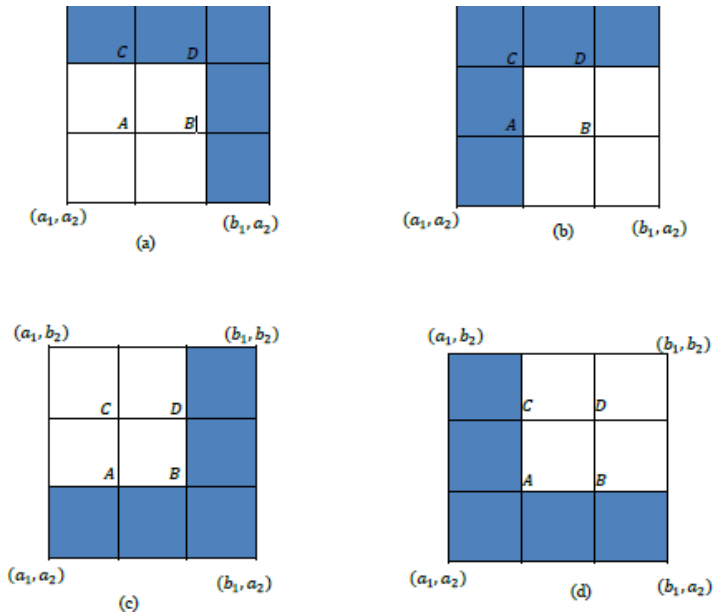
Interposer die (AID)
Measurements from e-smi

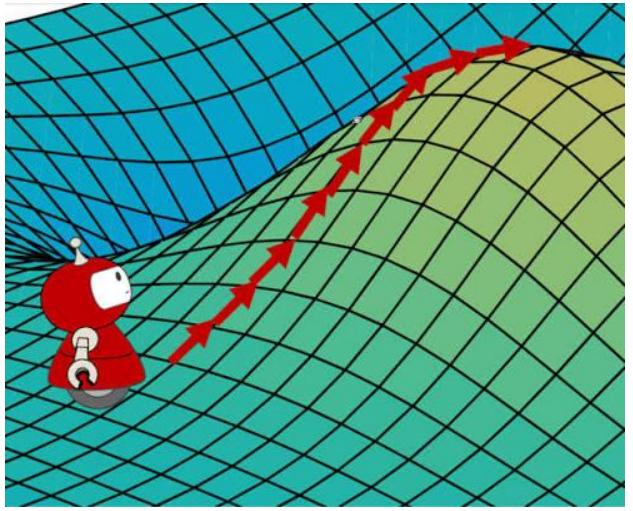
MI300A - Energy (rocHPCG)



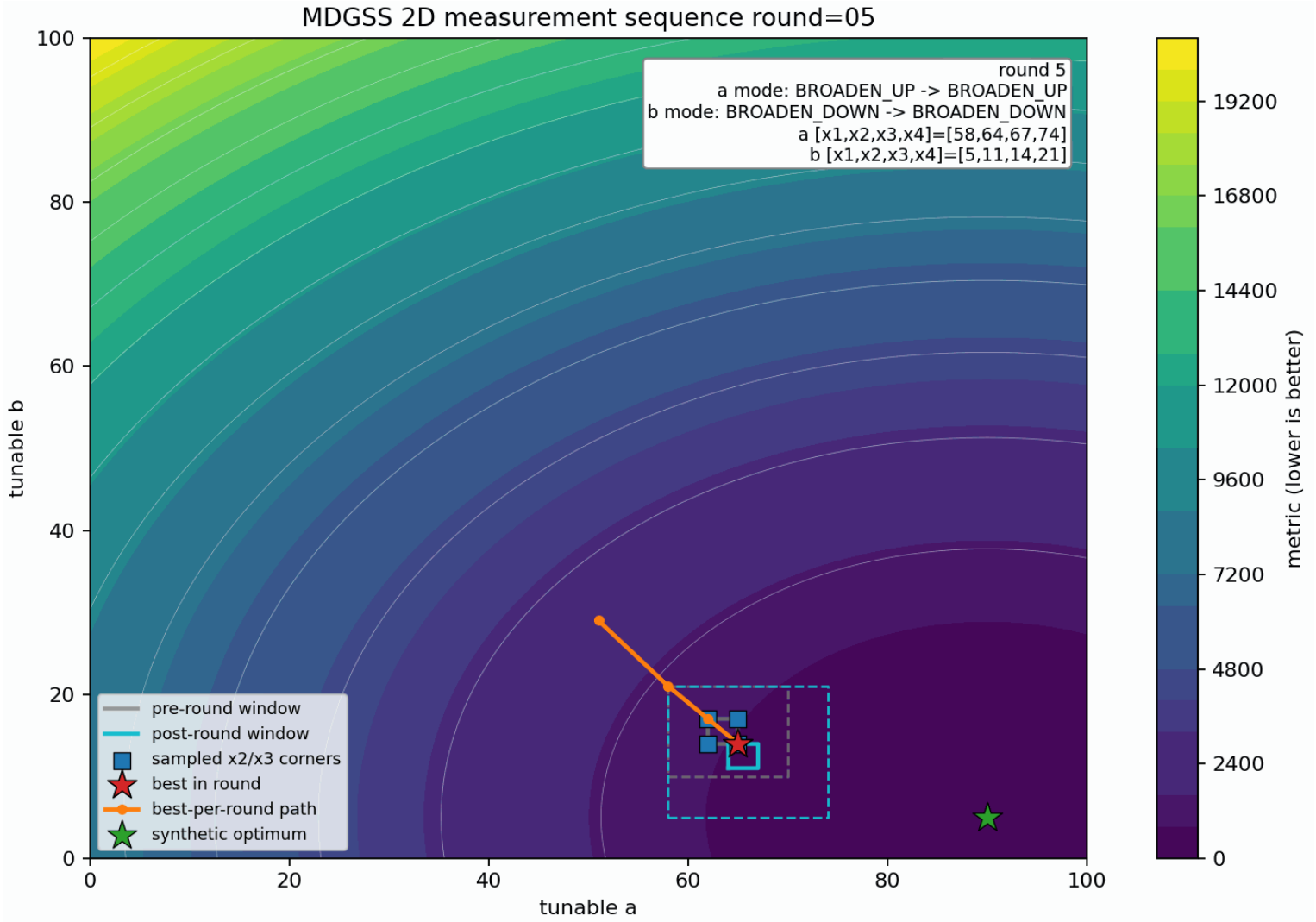
Multidimensional Frequency Steering

- Multidimensional Golden Section Search
- Hillclimb
- Ternary Discrete Gradient Descent





GSS -> Multi-Dimensional GSS

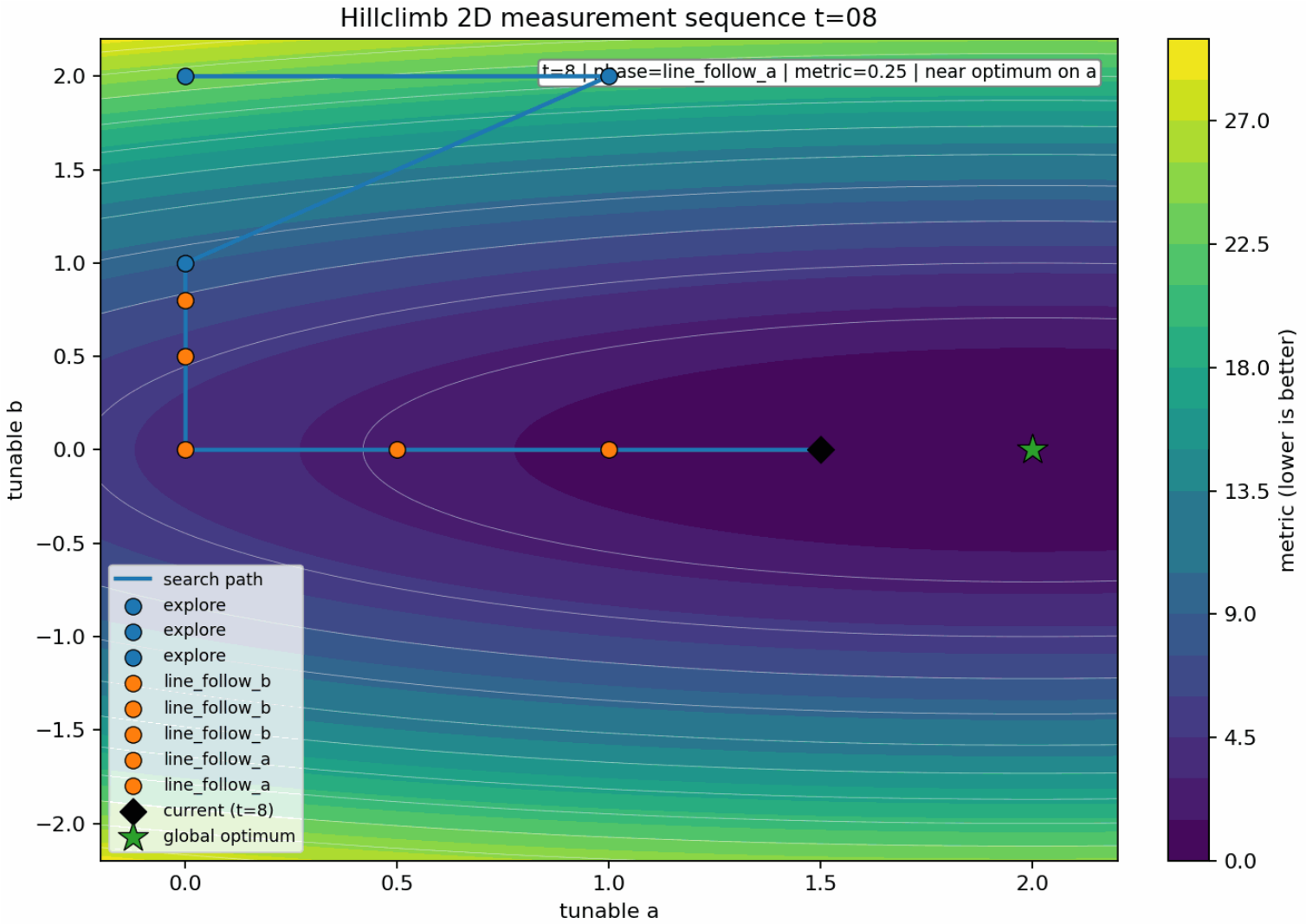


Multi-Dimensional GSS - Issues

- Not suitable for discrete tunables, e.g fclk, mclk have just 4 values.



Hillclimb

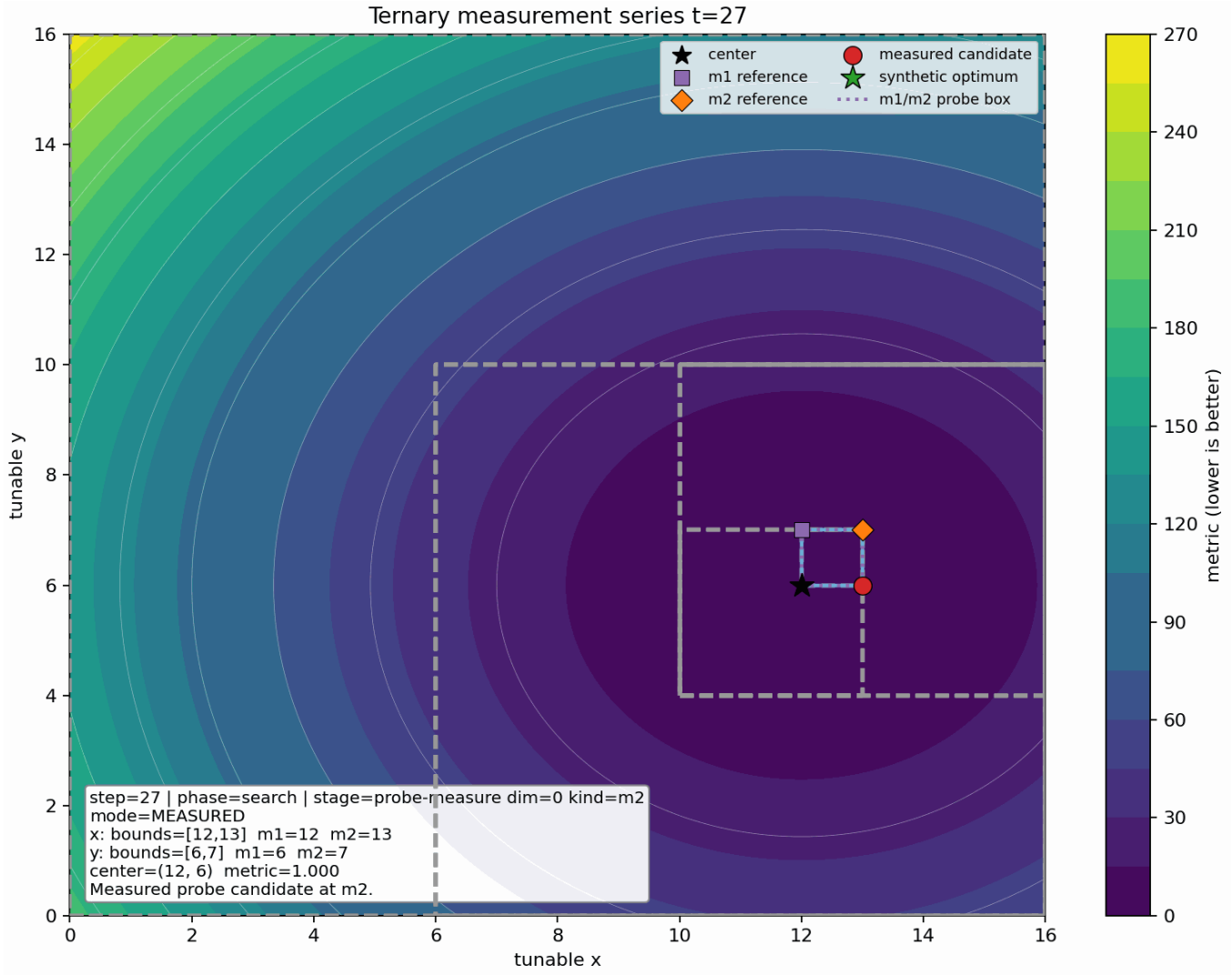

Hillclimb 2D measurement sequence t=08
t=8 | phase=line_follow_a | metric=0.25 | near optimum on a
tunable b
tunable a
metric (lower is better)
search path
explore
explore
explore
line_follow_b
line_follow_b
line_follow_b
line_follow_a
line_follow_a
current (t=8)
global optimum

Hillclimb - Issues

- **Slow** convergence
- On AMD mi300a all **clock dimensions are coupled** via socket power (e.g. boost and sclk).
- Some dimensions are coupled through mutual bottlenecks (fclk, mclk)
- Coupled dimensions can require multiple tunable changes to restore throughput.
  (e.g. raising sclk without raising fclk and vice-versa may raise energy without improving performance.
- Search can remain trapped in local minima during state transitions.



Ternary Discrete Gradient Descent



Ternary Discrete Gradient Descent

Monitor phase

- Monitoring instead of continuous probing to mitigate ‘bad’ neighbor measurement

Restarts

- Local/global restarts + cross-neighbor probes mitigate lock-in.

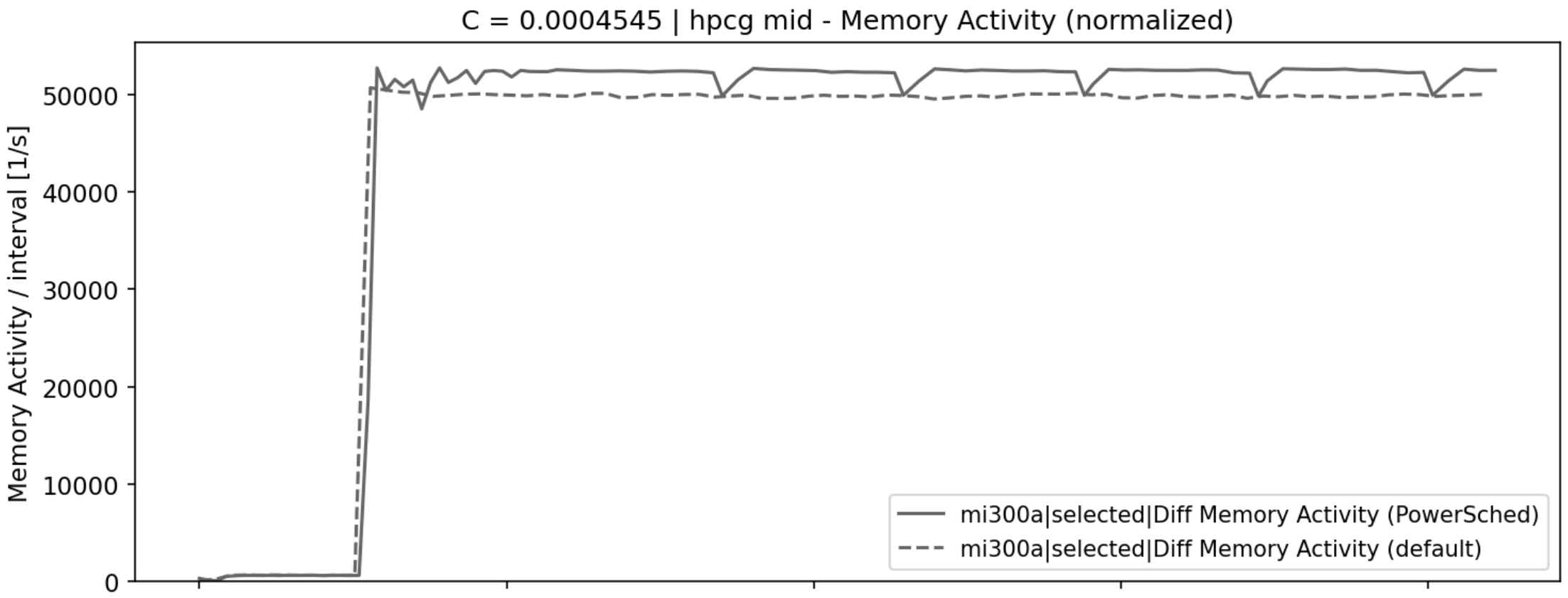
Wait/probe/measure cycle.

- Wait phase suppresses transients (application, cache, thermal ramps, sensor noise, mpi).
- Probe phase tests candidates without immediate commitment
- Measure phase accepts only for candidates near current best

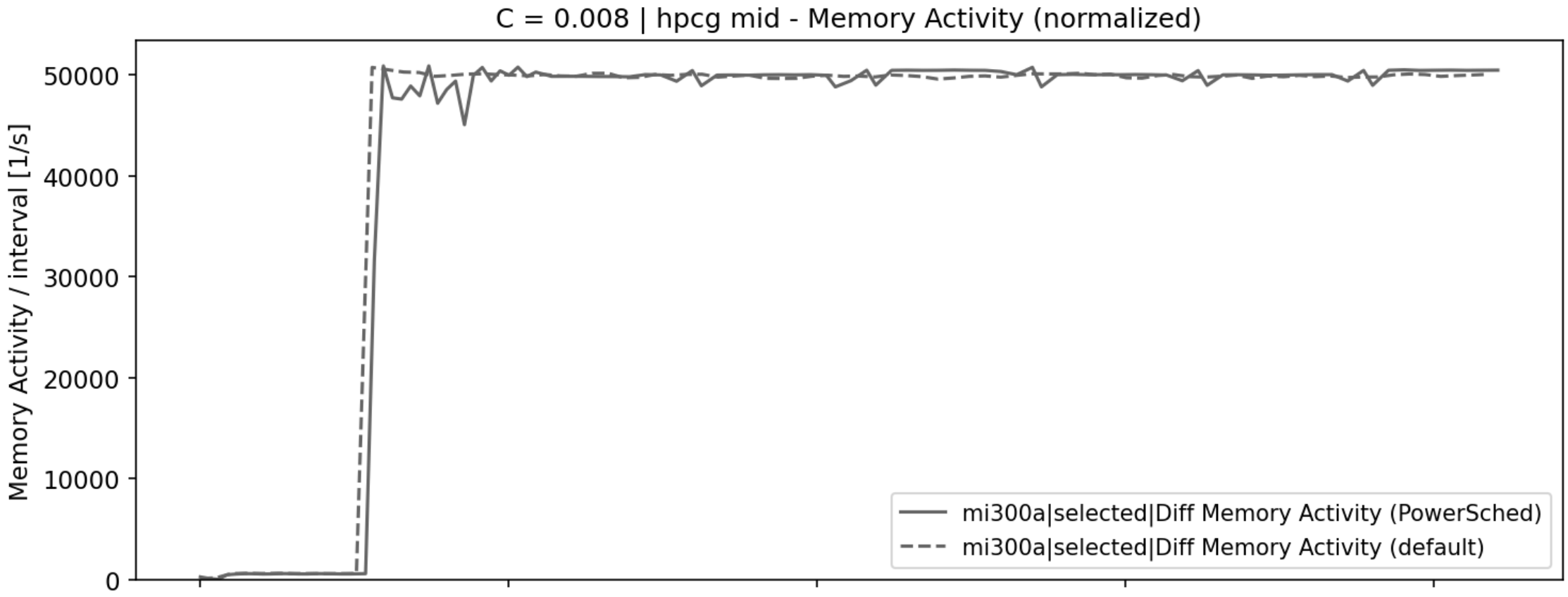


Ternary Discrete Gradient Descent Results on MI300A

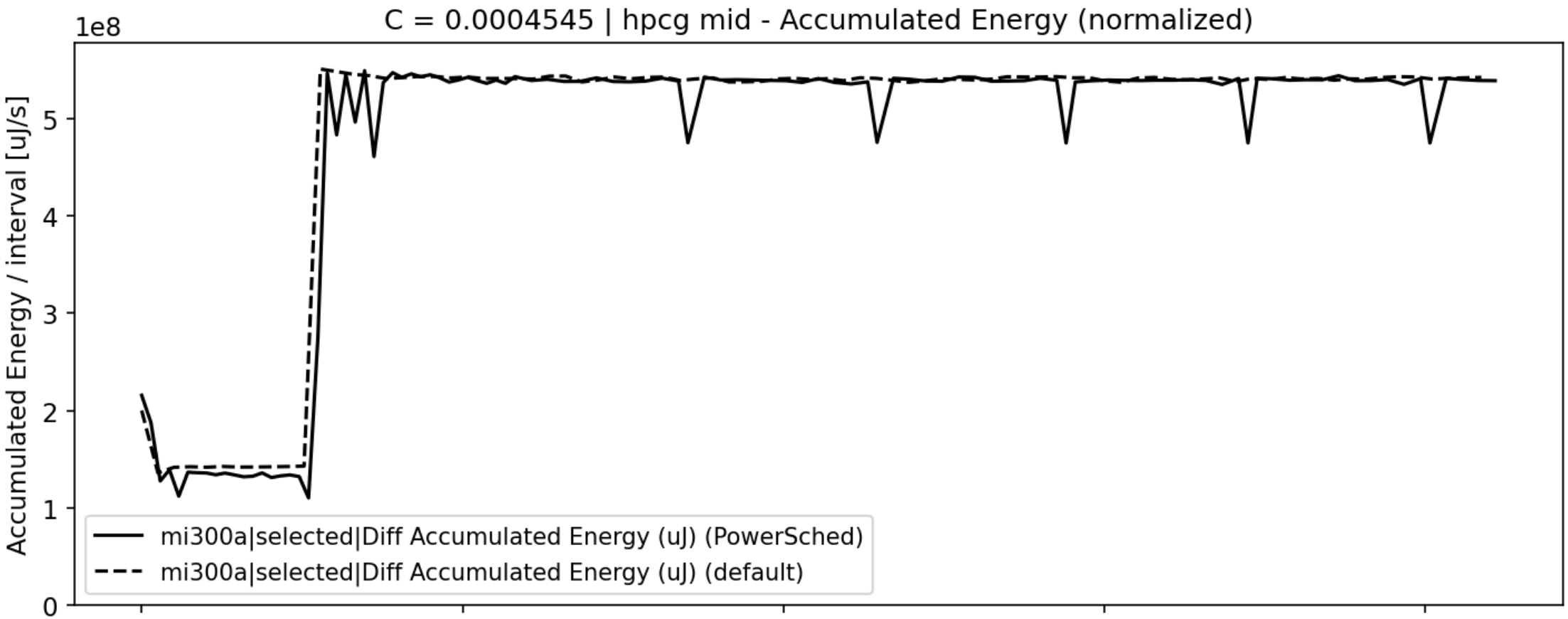
rocHPCG – Performance Proxy (C=0.0004545)



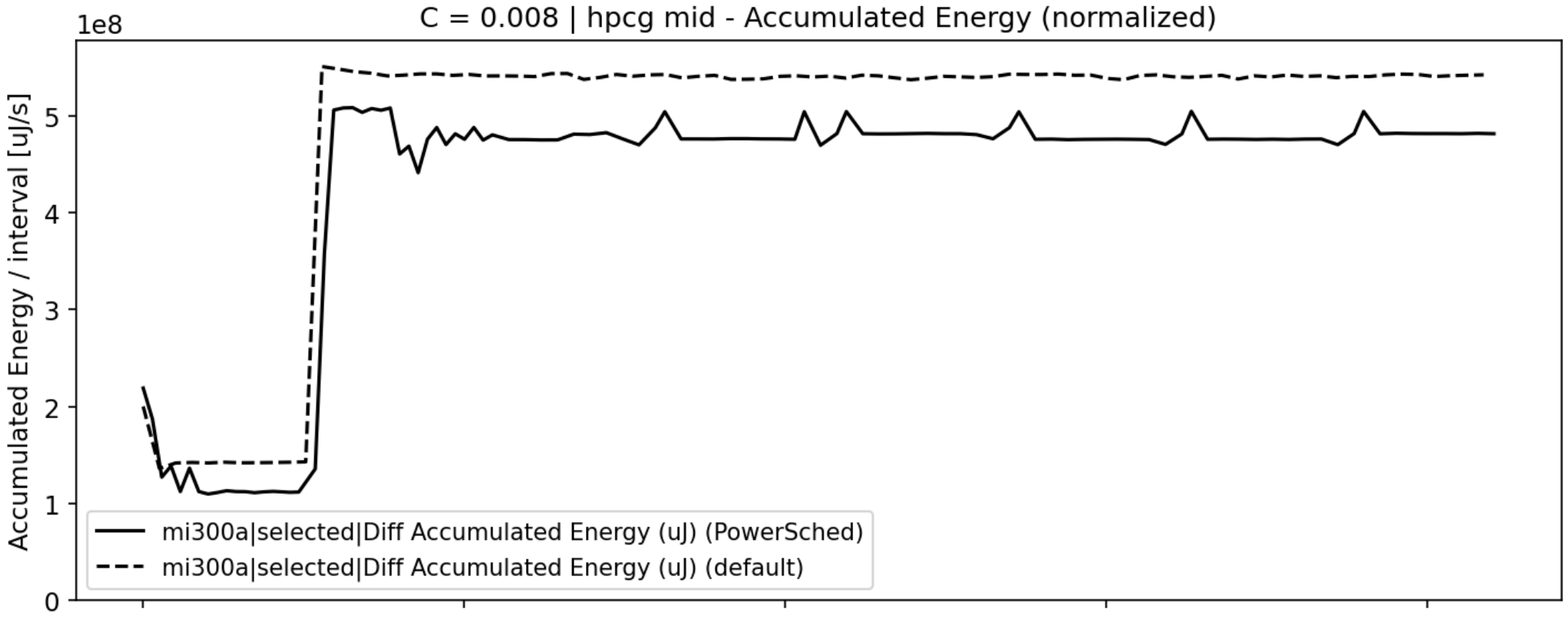
rocHPCG – Performance Proxy (C=0.008)



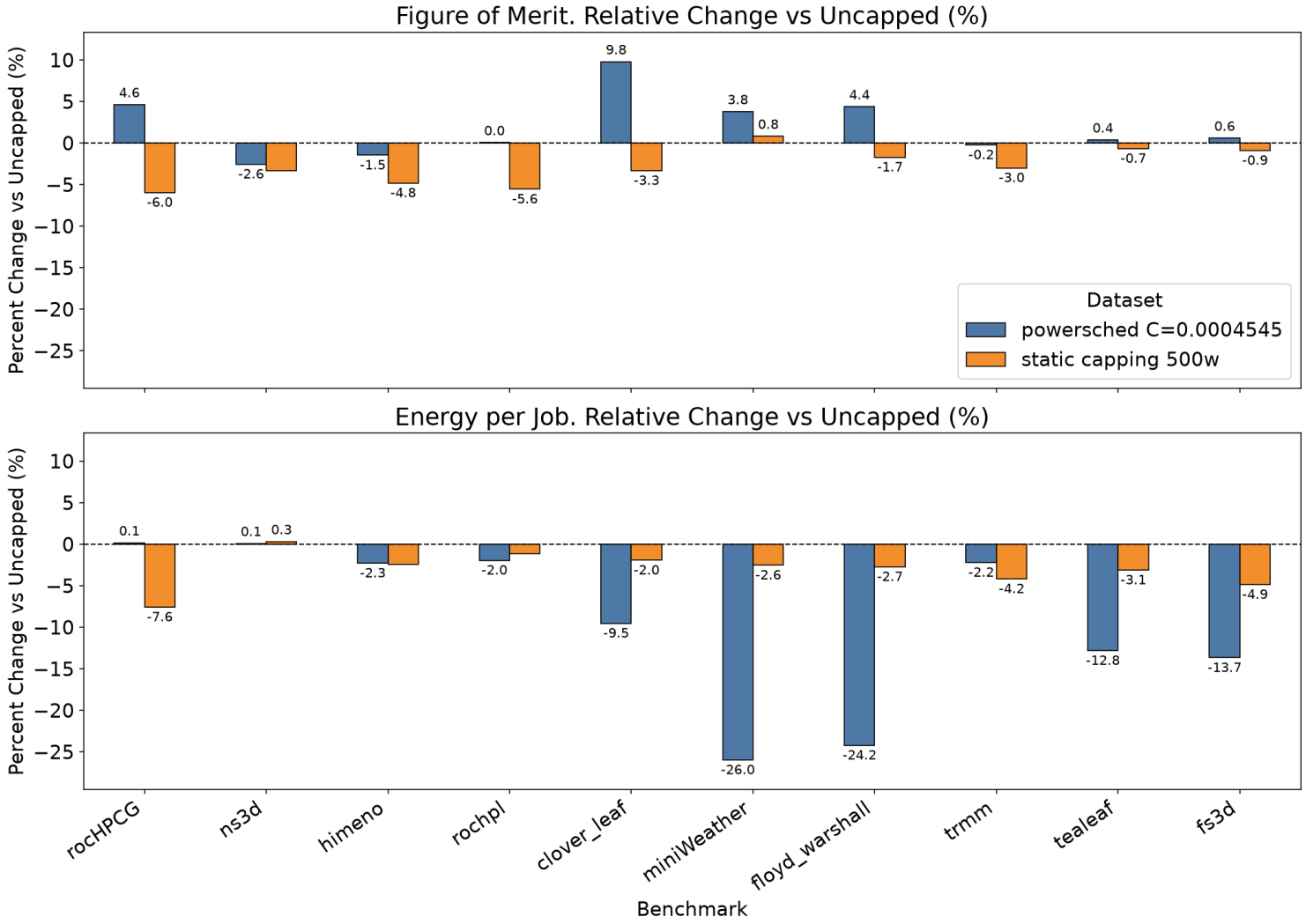
rocHPCG - Energy (C=0.0004545)



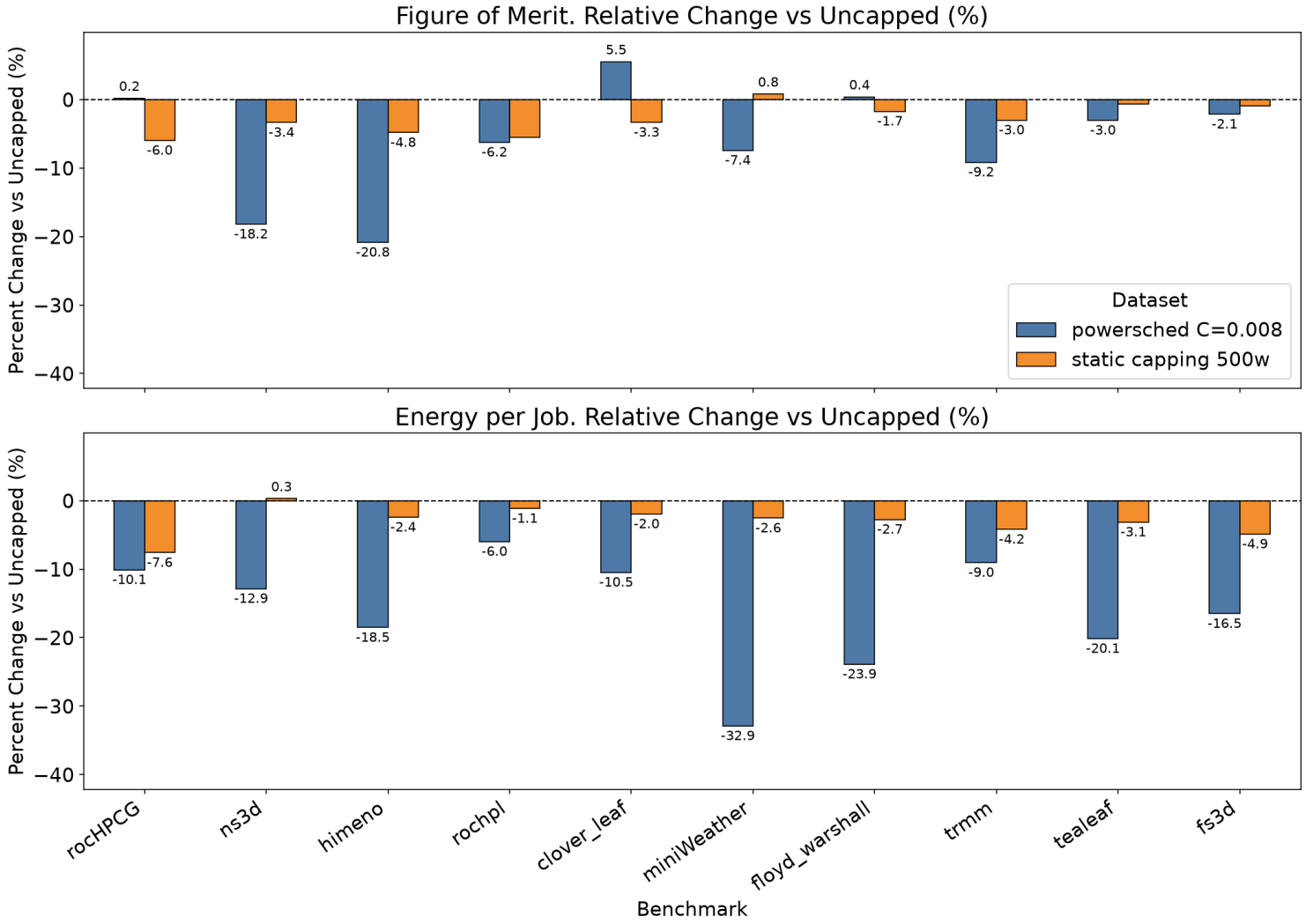
rocHPCG – Energy (C=0.008)



C=0.0004545


Figure of Merit. Relative Change vs Uncapped (%)
Percent Change vs Uncapped (%)
10
5
0
−5
−10
−15
−20
−25
4.6
-6.0
-2.6
-1.5
-4.8
0.0
-5.6
9.8
-3.3
3.8
0.8
4.4
-1.7
-0.2
-3.0
0.4
-0.7
0.6
-0.9
Dataset
powersched C=0.0004545
static capping 500w
Energy per Job. Relative Change vs Uncapped (%)
Percent Change vs Uncapped (%)
0.1
-7.6
0.1
0.3
-2.3
-2.0
-2.0
-9.5
-2.0
-26.0
-2.6
-24.2
-2.7
-2.2
-4.2
-12.8
-3.1
-13.7
-4.9
rocHPCG
ns3d
himeno
rochpl
clover_leaf
miniWeather
floyd_warshall
trmm
tealeaf
fs3d
Benchmark

C=0.008


Figure of Merit. Relative Change vs Uncapped (%)
Percent Change vs Uncapped (%)
0
−10
−20
−30
−40
0.2
-6.0
-18.2
-3.4
-20.8
-4.8
-6.2
5.5
-3.3
-7.4
0.8
0.4
-1.7
-9.2
-3.0
-3.0
-2.1
Dataset
powersched C=0.008
static capping 500w
Energy per Job. Relative Change vs Uncapped (%)
Percent Change vs Uncapped (%)
0
−10
−20
−30
−40
-10.1
-7.6
-12.9
0.3
-18.5
-2.4
-6.0
-1.1
-10.5
-2.0
-32.9
-2.6
-23.9
-2.7
-9.0
-4.2
-20.1
-3.1
-16.5
-4.9
rocHPCG
ns3d
himeno
rochpl
clover_leaf
miniWeather
floyd_warshall
trmm
tealeaf
fs3d
Benchmark

Results AMD MI300A

Notes

- FoM includes startup, transitions.
- Energy measured by PM_counters, entire run.

Observations

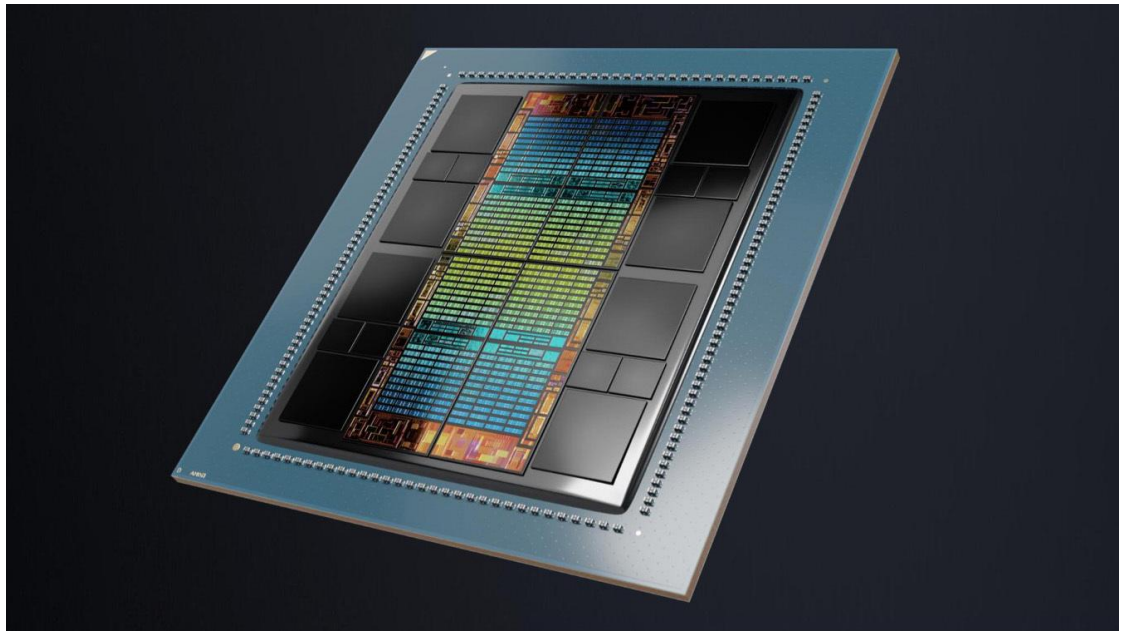
- Static 500W cap: -2.8% performance, -3.1% energy, -5.9% power (avg).
- Performance (C=0.0004545): +1.7% performance and -8.5% energy.
- Energy efficiency (C=0.008): -5.8% performance, -15.0% energy, -19.9% power.



HLRS Hunter – HPE Cray EX4000 (AMD MI300A)

Preliminary results on AMD MI300A at **550W**

- 0-10% improved performance
- 0-35% improved energy efficiency – (35% only with substantial performance impact)



Next steps

- **AI workloads** on MI300A
- Support for **Nvidia**
- **Training for Offline-RL**

Thank You

Christian Simmendinger (HPE)
christian.simmendinger@hpe.com



© 2026 Hewlett Packard Enterprise Development LP