Embrace the Chaos

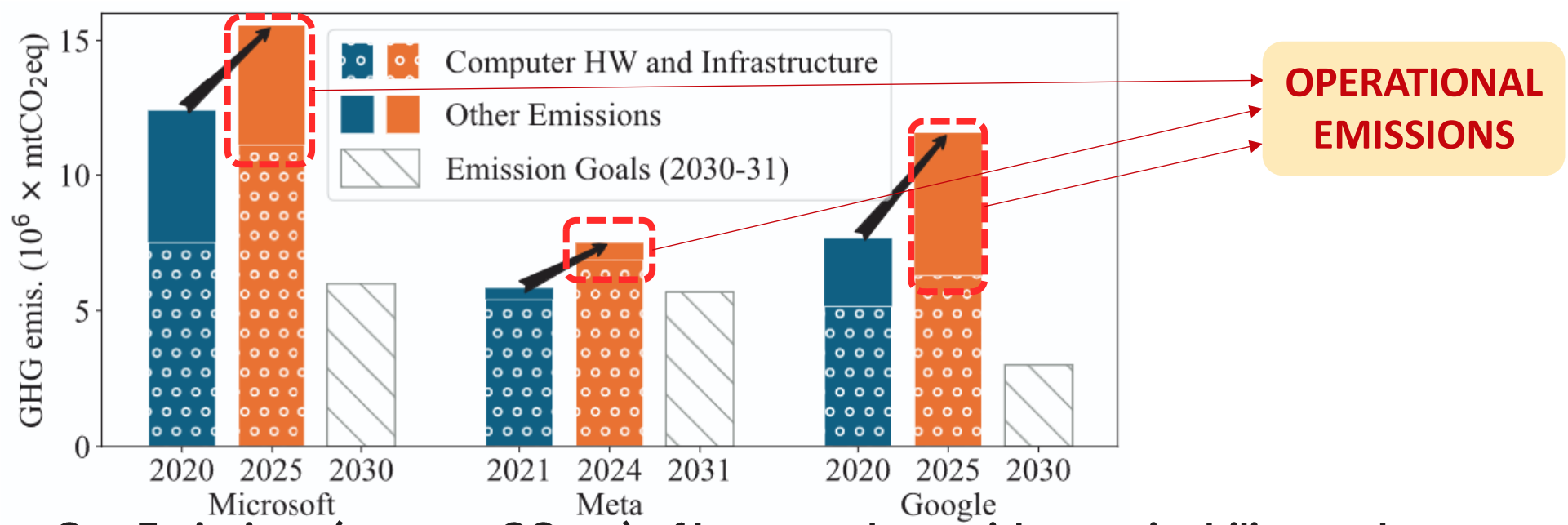
Harnessing Variability for Large-Scale System Efficiency

Sustainable Computing Architecture: From Silicon to Datacenter
Co-Designing AI Systems Across the Stack
MICRO 2026

Rutwik Jain
Advisors: Matthew Sinclair, Shivaram Venkataraman
Ph.D. Candidate, University of Wisconsin-Madison

AI is driving exponential compute growth

… and with it, emissions



Greenhouse Gas Emissions ($_{10^6}$ ×mtCO$_2$eq) of hyperscalars, with sustainability goals

[Liu et al. CAPA: Manufacturing Carbon Estimation for Advanced-Packaged Architectures, **ISCA 2026**]

The gap between current emissions and sustainability goals is widening

Systems are massively power constrained

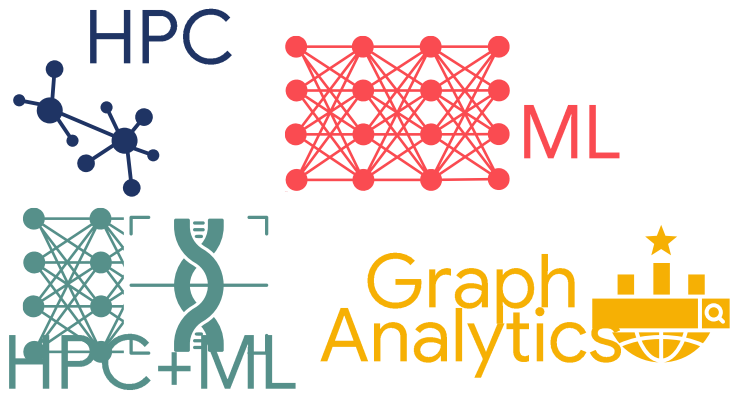
Application Software

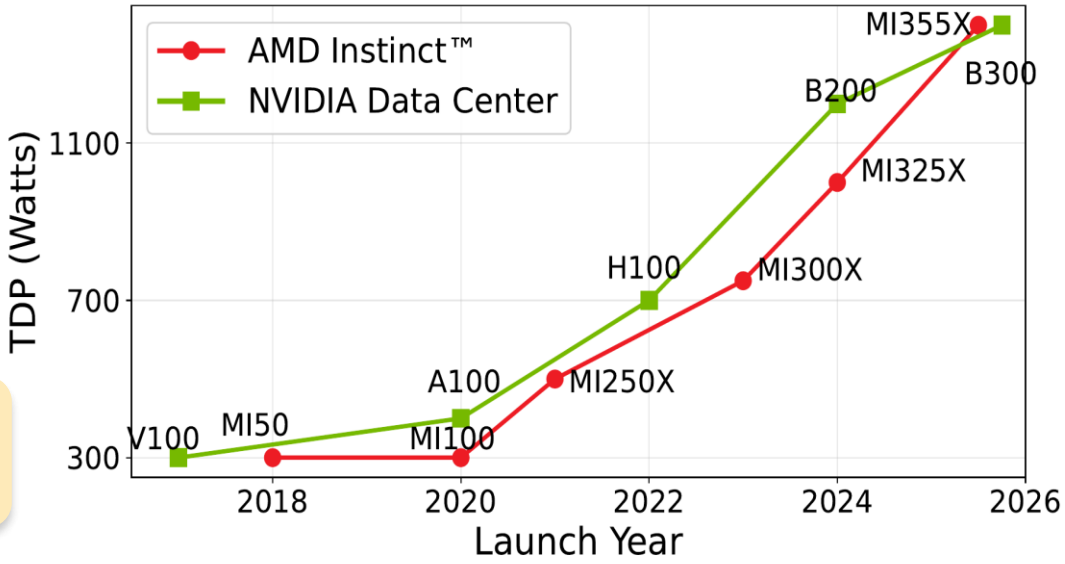
voracious scaling demands

Systems

post-Moore scaling – extreme TDP limits

GPU Hardware





Systems are massively power constrained



Application Software

voracious scaling demands

Systems

OAK RIDGE National Laboratory
Frontier
64,000 MI250X GPUs
(24.6 MW)

Lawrence Livermore National Laboratory
El Capitan
44,544 MI300A GPUs
(35 MW)



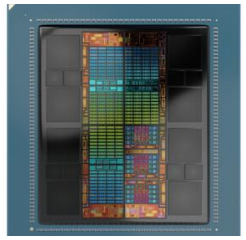
Hyperscalers
~millions
of accelerators
~GW power budgets

post-Moore scaling – extreme TDP limits

GPU Hardware

Power bottlenecks at every level

socket

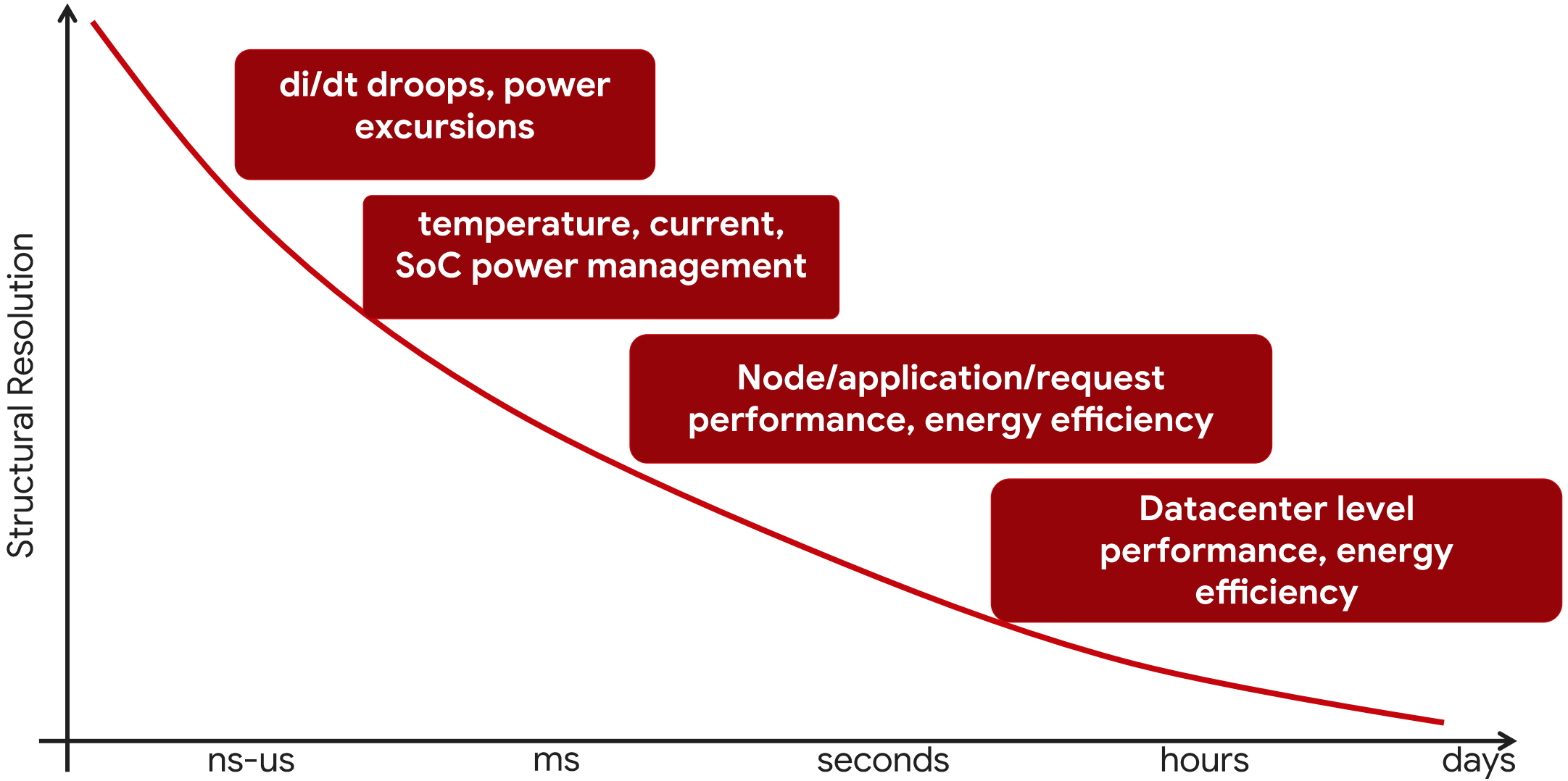


node



datacenter





Systems are massively power constrained

Application Software

voracious scaling demands



Challenge

How to **manage system resources** like **power** for the overall system to run **efficiently**

My Research

Co-optimizing power and performance at scale, through cross-stack approaches spanning hardware, power management (PM) firmware, systems, and applications.

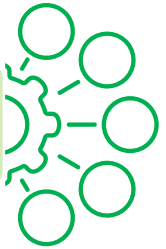
GPU Hardware

My Research



Characterizing Variability

Not All GPUs are Created Equal [SC'22]
AI Benchmark Carpentry [arXiv'25]



Current Systems

PAL [SC'24]
Minos [SIGMETRICS'26]
MLCommons Ontology [arXiv'25]

Application Software

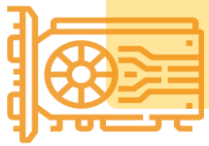
Systems

Scheduler

Hardware

PM

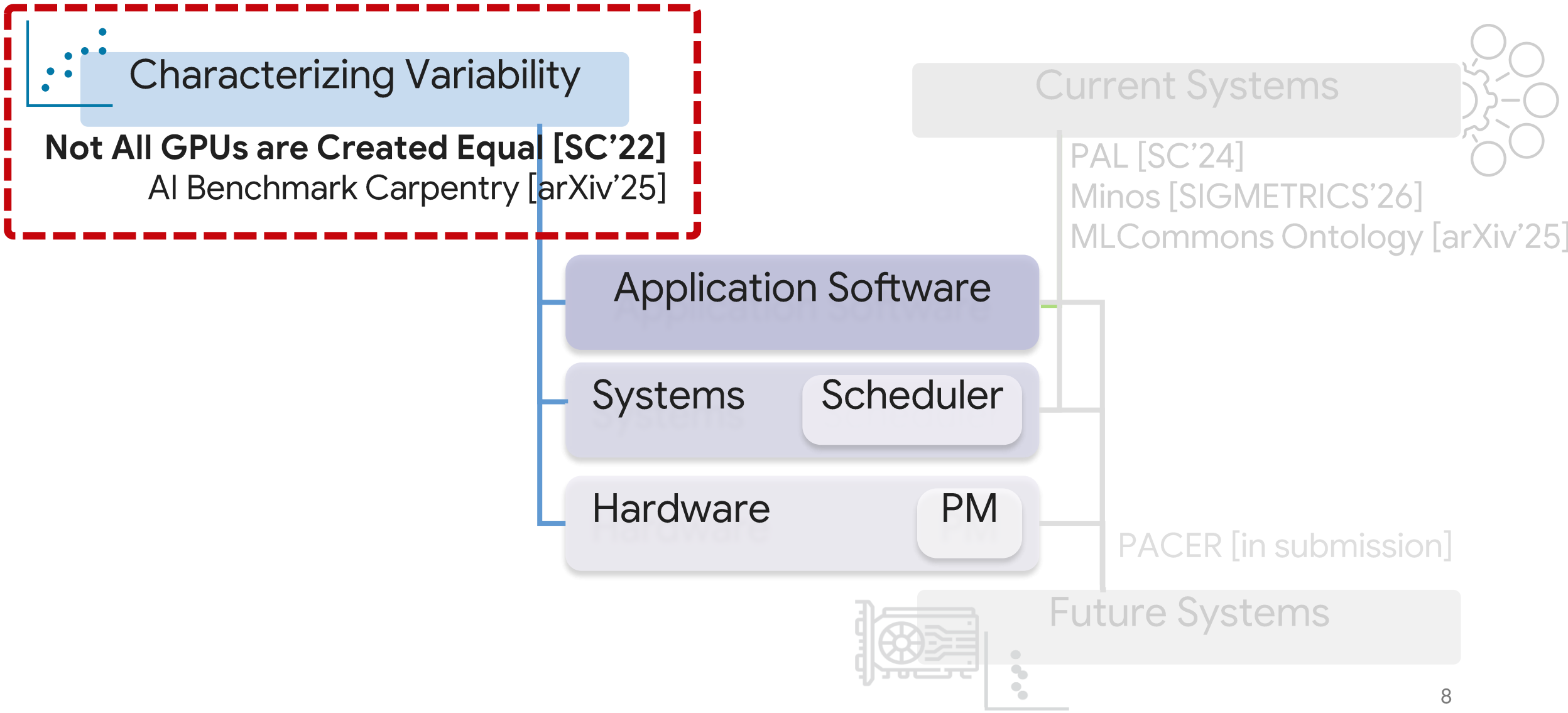
PACER [in submission]



Future Systems



Snapshot 1
Not All GPUs Are Created Equal

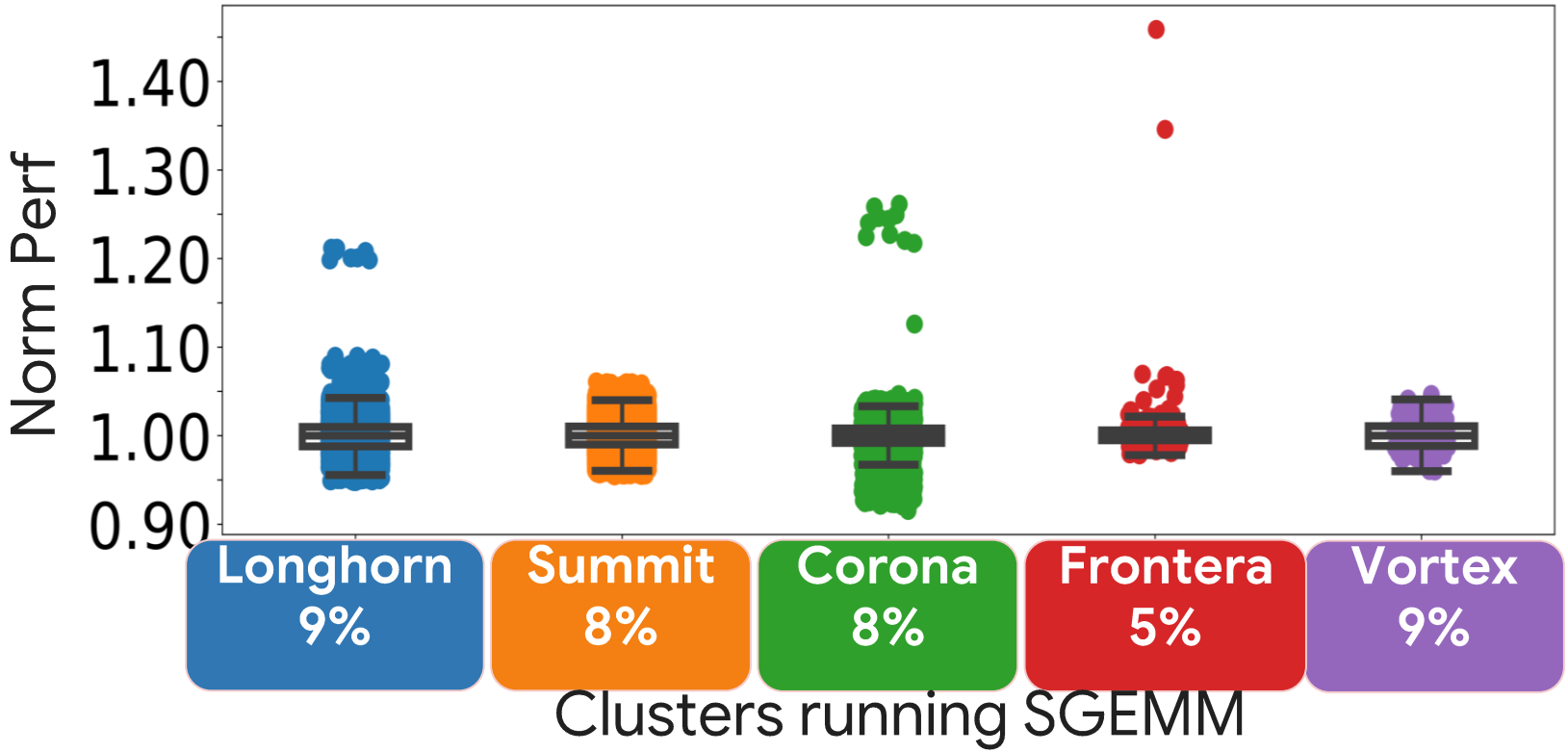


Not All GPUs Are Created Equal

- GPUs exhibit variability because of:
  - Local Power Management (PM)
  - Thermal/non-uniform cooling effects
  - Process variation
- **Impact**
  - Users: Hard to get **repeatable, consistent** performance
  - SysAdmins: **Resource underutilization** for multi-GPU jobs – faster GPUs wait for stragglers

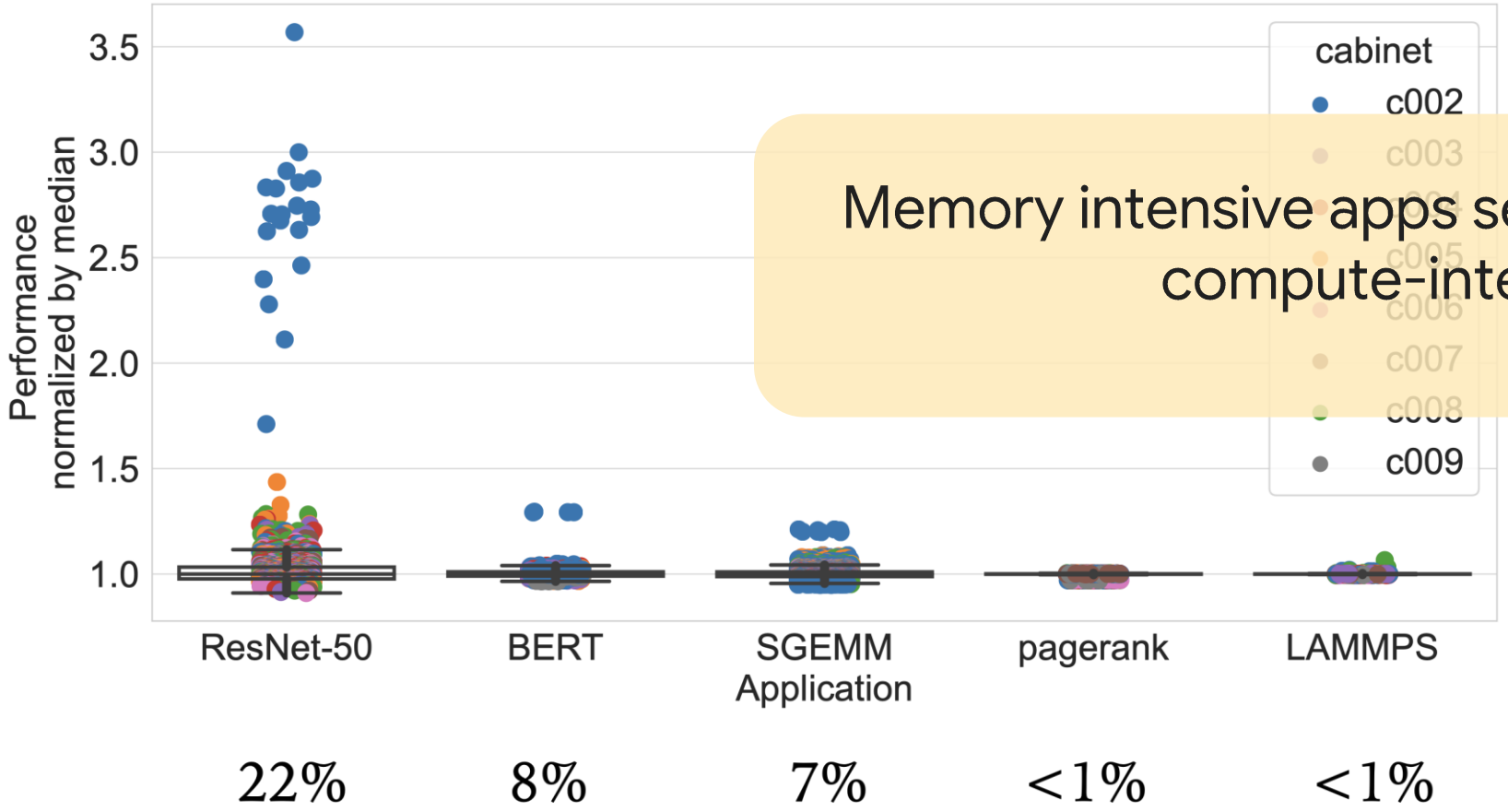
    $\Rightarrow$ higher energy usage for same work

Takeaway 1
Consistent variability across clusters



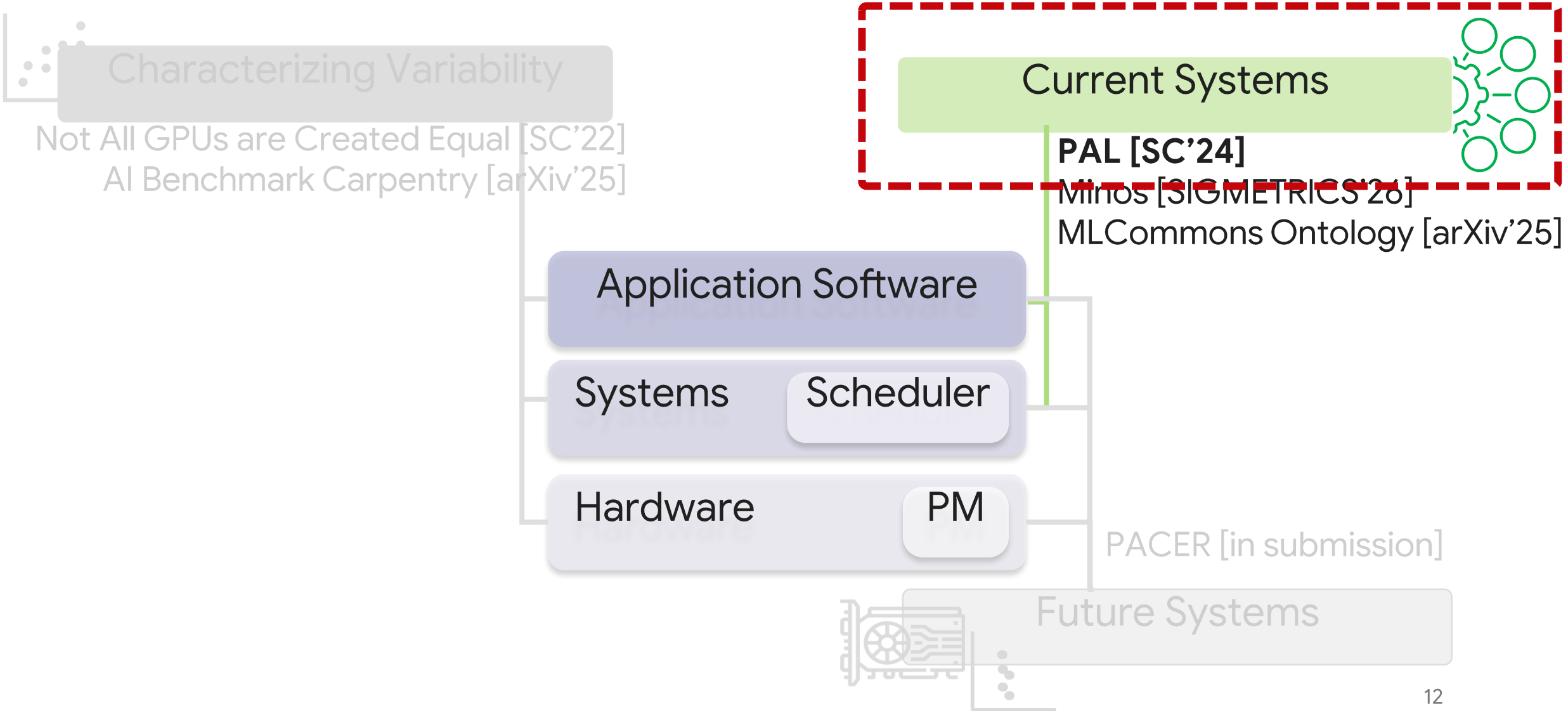
Consistent GPU perf variation **(avg 8%, up to 22% per app**, up to **1.5x** slower outliers)

Takeaway 2
Variability is application-dependent

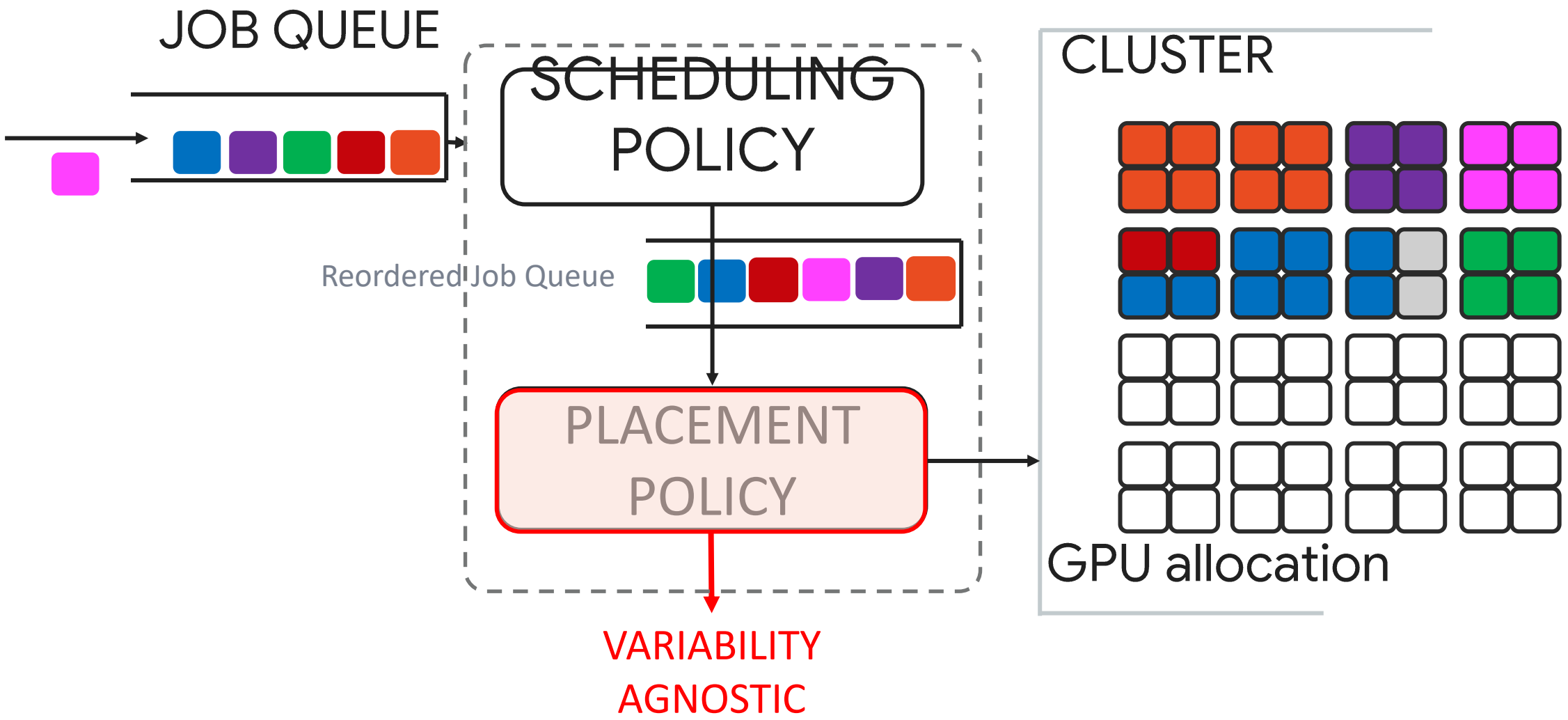


Memory intensive apps see lower variability than compute-intensive ones

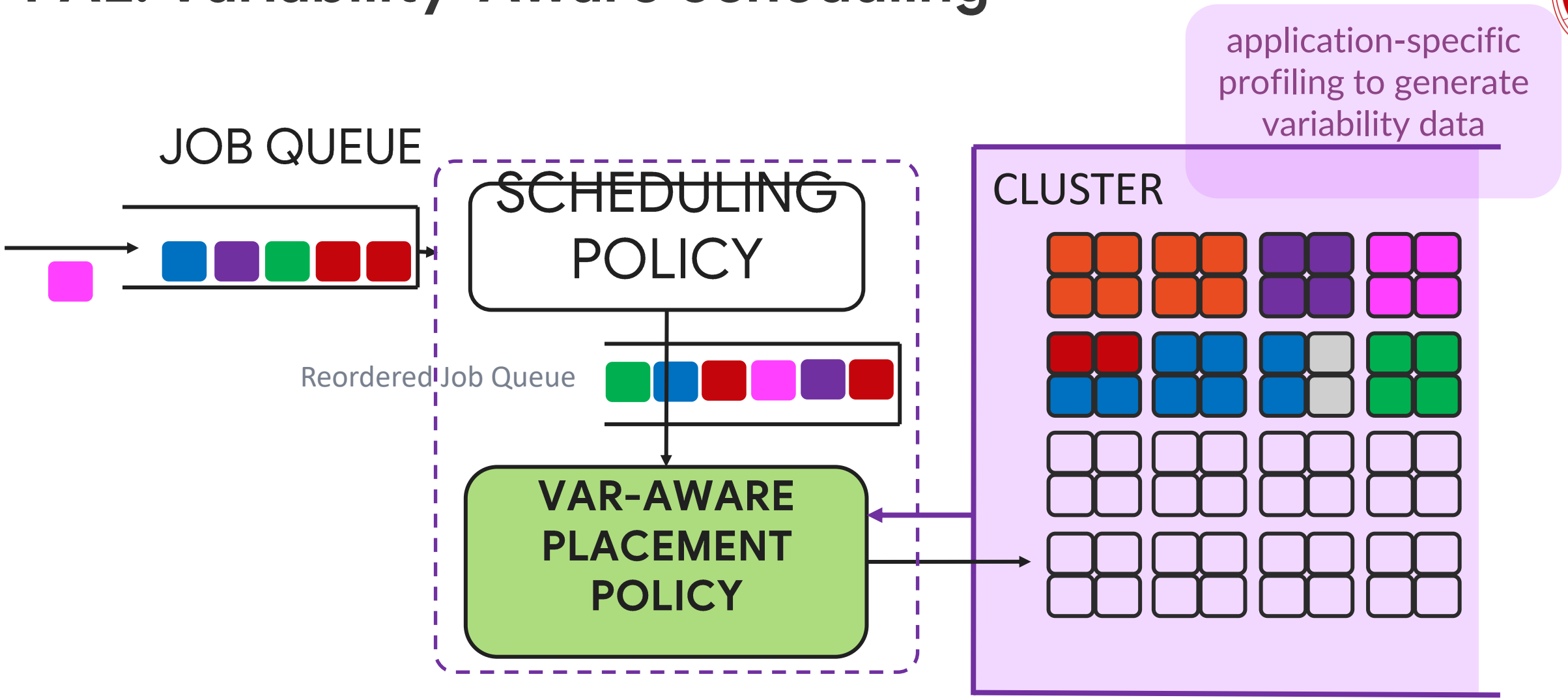
Snapshot 2
PAL: Variability-Aware Scheduling



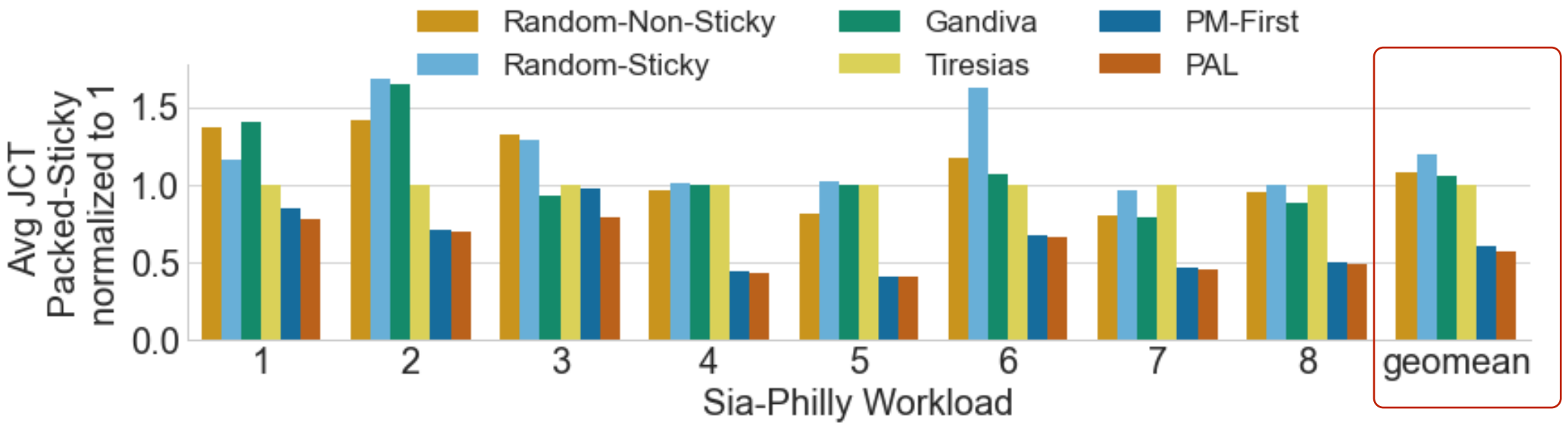
PAL: Variability-Aware Scheduling [SC'24]



PAL: Variability-Aware Scheduling

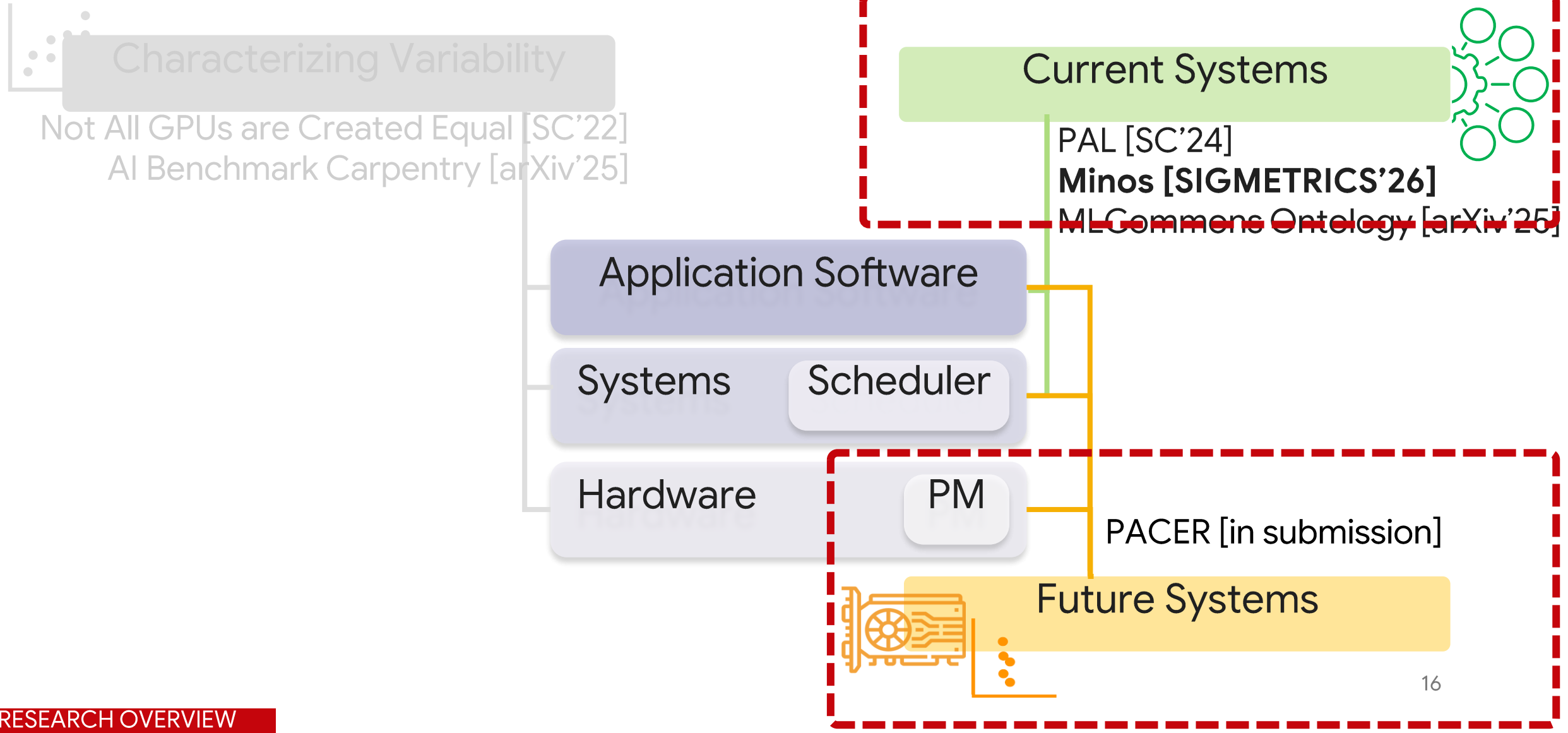


Reduced JCT, Makespan; Improved cluster utilization

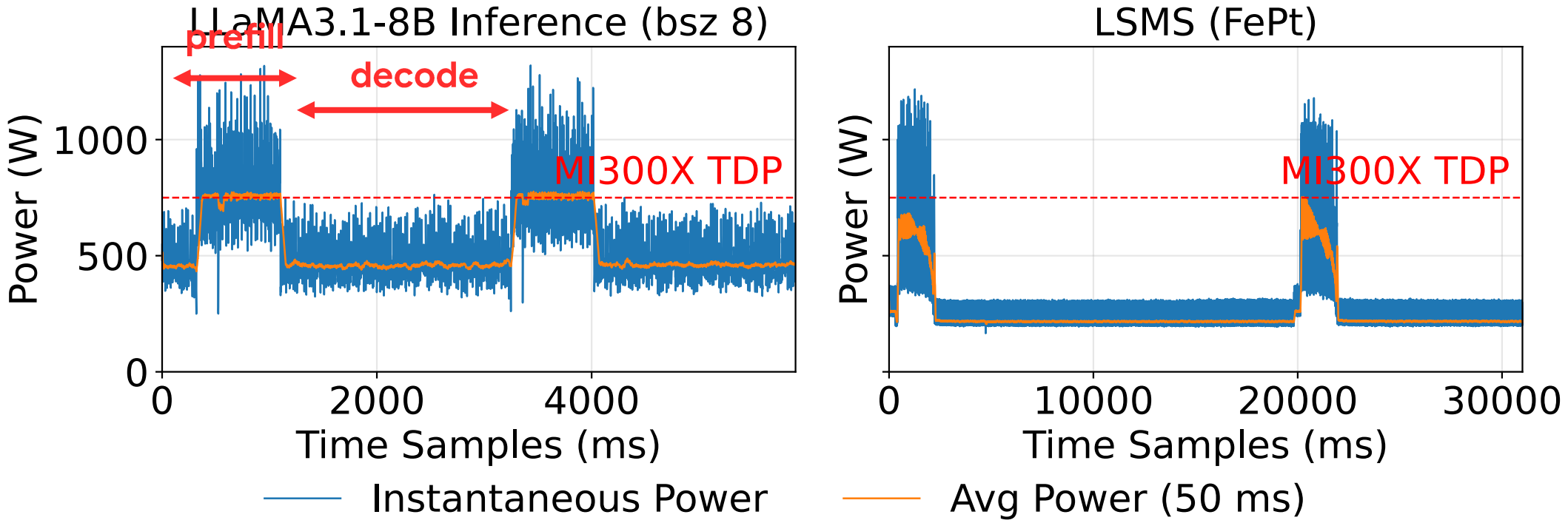


PAL improves
makespan by **47%**
cluster utilization by **28%**
Job Completion Time (JCT) by **43%** geomean
over SOTA

Snapshot 3
Power Excursions



Power Excursions



- **Power excursions** are **spikes above the thermal design power (TDP)** of the GPU
- Simultaneous **power excursions** across many GPUs can breach power provisioning limits.
- Trigger GPU-level protection mechanisms that throttle frequency and degrade performance.

Excursion Specification

OCP Power-Spike Specification

Excursion Design Power (×TDP)	Duration
2×TDP	≤ 20 µs
1.75×TDP	≤ 2 ms
1.5×TDP	≤ 5 ms
1.3×TDP	≤ 10 ms
1.1×TDP	≤ 20 ms
1xTDP	indefinitely

[Open Accelerator Infrastructure (OAI) Specification, Open Compute Project (OCP)]

Q. Are all excursions bad?

A. Not all of them.
We studied excursions and mitigation mechanisms in HW and FW [**PACER**] to find

- Low-magnitude excursions (up to 1.3xTDP) allow kernels to **computationally sprint** for short bursts of time, improving performance
- Higher-magnitude excursions can cause severe throttling.
- **Manage, don't mitigate**

Developed a dynamic frequency management algorithm that improves **median TTFT** of online LLM serving by **12%** on production systems, within 3% of an oracle solution

Q. Can I predict power behavior for unseen workloads?

Power efficiency optimizations, including PAL and PACER, need workload-specific information & incur profiling overheads

A. Minos [SIGMETRICS'26]

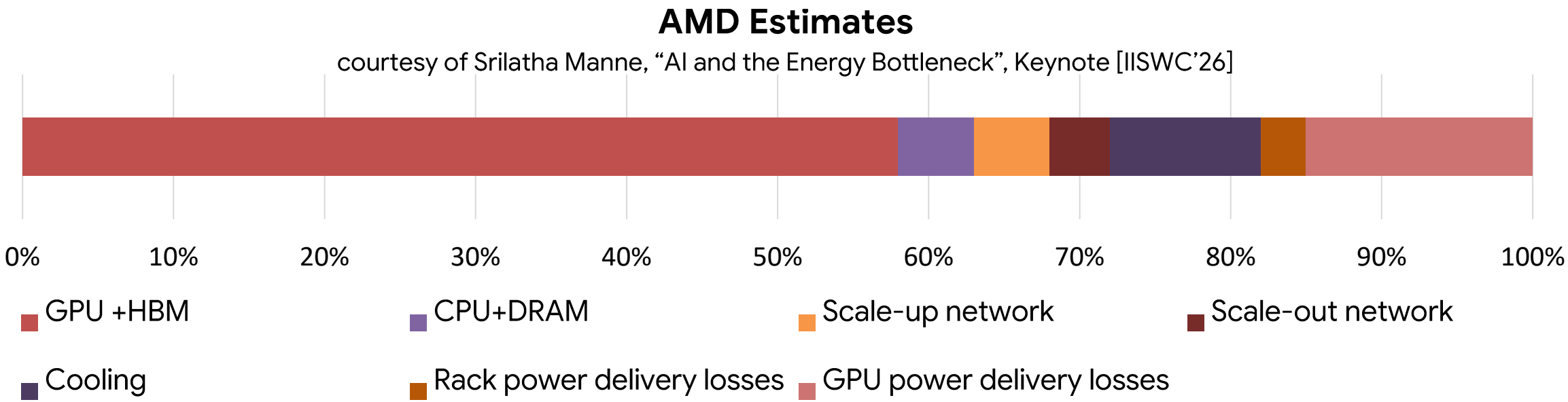
- Combines novel **power** classification with **performance**-based classification to group similar workloads
- Low overhead approach that reduces profiling time by **90%**
- Accurately predicts **peak power and performance scaling** under different frequencies.
- Minos achieves a mean error of **4%** for peak power predictions for unseen workloads



Looking to the future

Power Management & Energy Efficiency

Where is the power going?



GPU: ~50-60% of total datacenter power
Cooling, power conversion losses, scale-up network switches: 33%

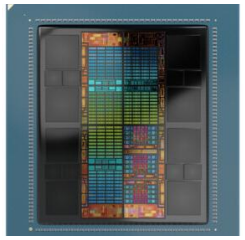
Power consumption and optimization is now the primary limiter to future perf and efficiency gains

- Lower energy per flop
- Lower energy per memory byte
- Lower energy per network byte

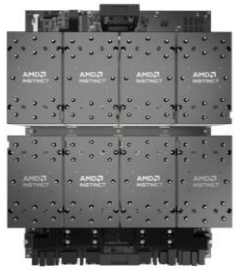
Power bottlenecks at every level

chiplet

socket

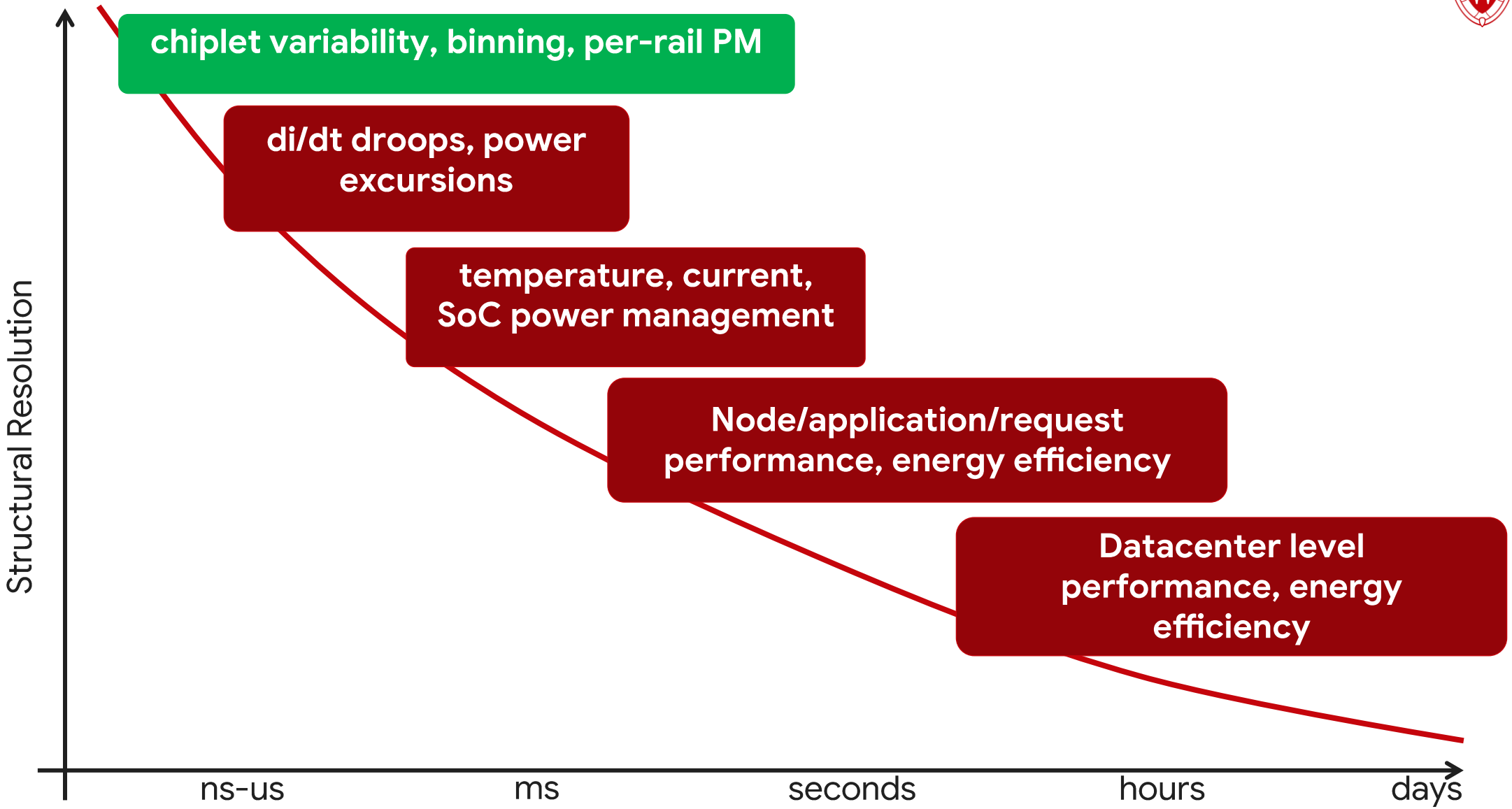


node

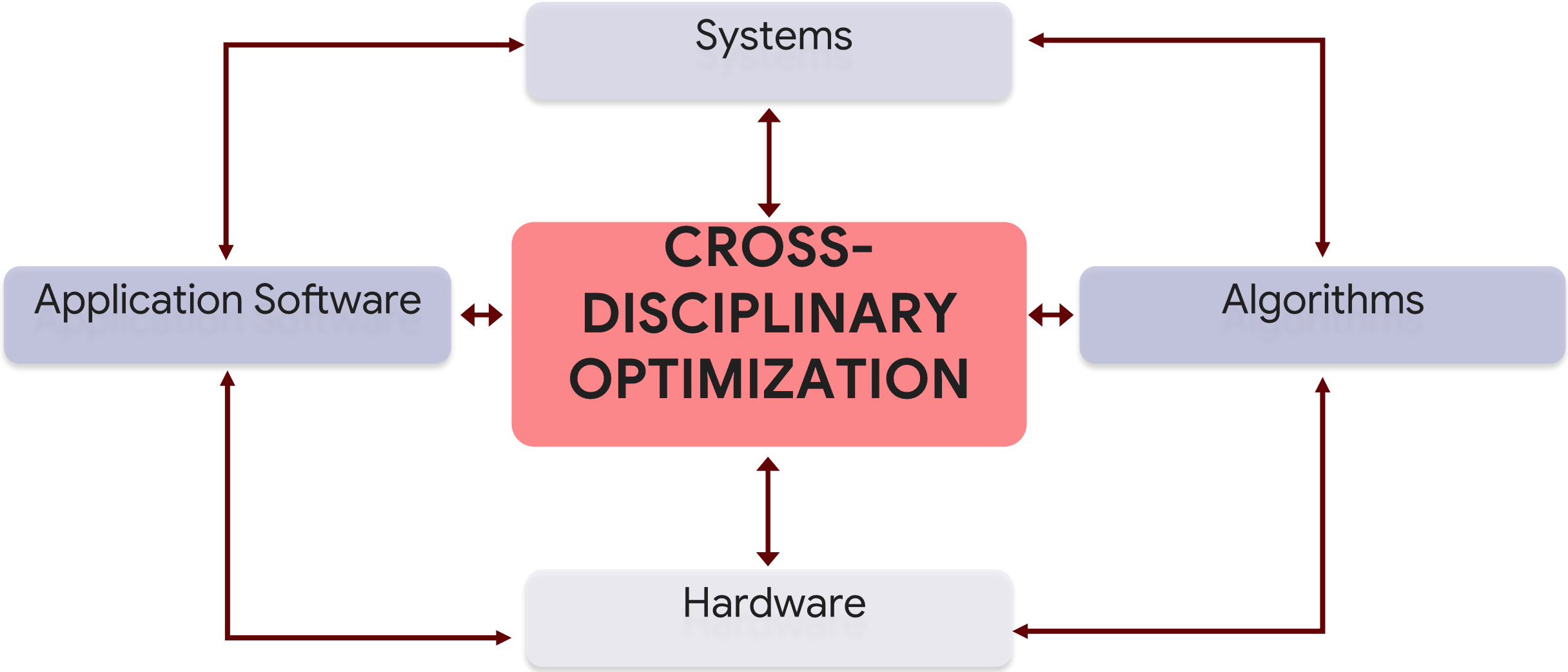


datacenter





Holistic co-design across the stack

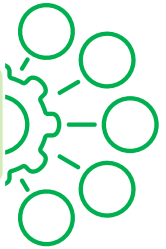


Thank you! Questions?



Characterizing Variability

Not All GPUs are Created Equal [SC'22]
AI Benchmark Carpentry [arXiv'25]



Current Systems

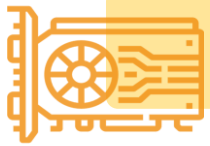
PAL [SC'24]
Minos [SIGMETRICS'26]
MLCommons Ontology [arXiv'25]

Application Software

Systems Scheduler

Hardware PM

PACER [in submission]



Future Systems

Collaborations across the stack:

- OakRidge National Labs (**ORNL**)
- **MLCommons** Science & HPC Working Group
- **AMD** and **NVIDIA**


WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON