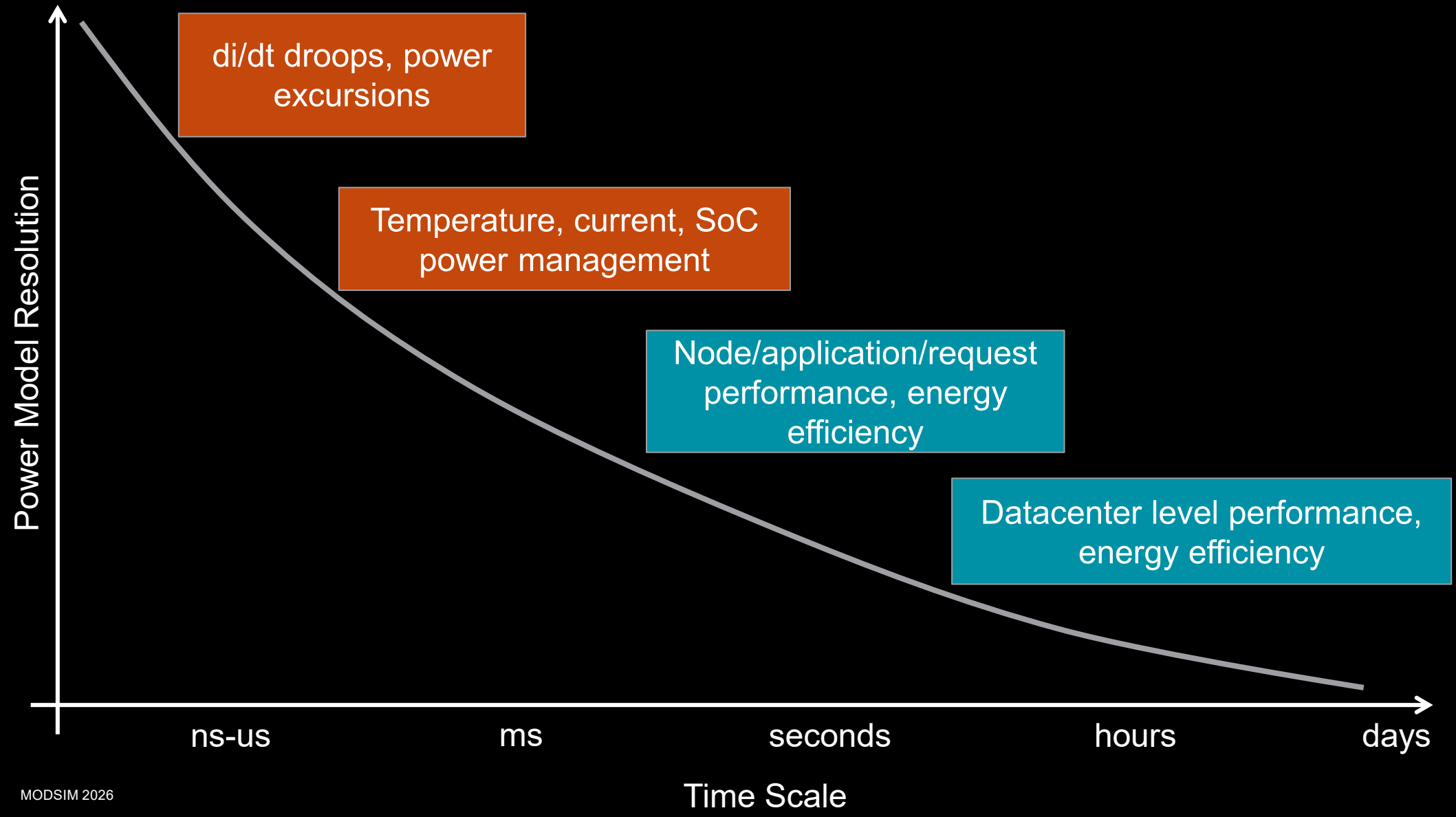
Thinking Bigger With AI

Srilatha (Bobbie) Manne
Senior Fellow
Research and Advanced Development
Advanced Micro Devices, Inc.

AMD
together we advance_

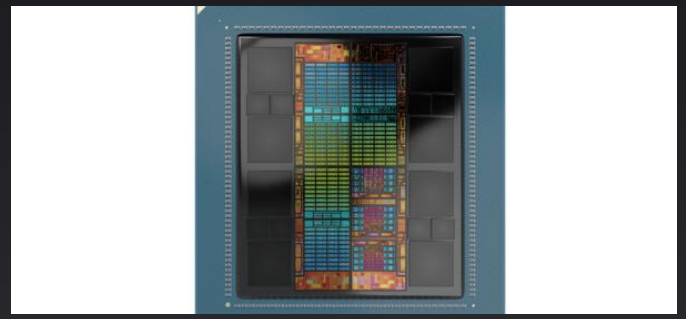
What's Under the Hood of Power Management



Power Bottlenecks at Every Level

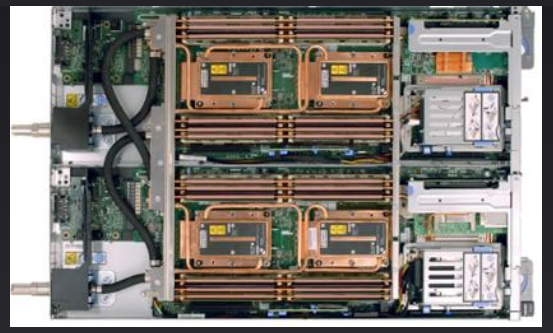
It's not just about energy efficiency

Socket



- **Physical Limits:**
  - Thermal design power (TDP) limits
  - Power delivery infrastructure constraints
  - Temperature constraints

Node



- **Platform Limits:**
  - Cooling solutions
  - Power delivery and reliability
- **Quality of Service**
  - Power = shared resource

Data Center

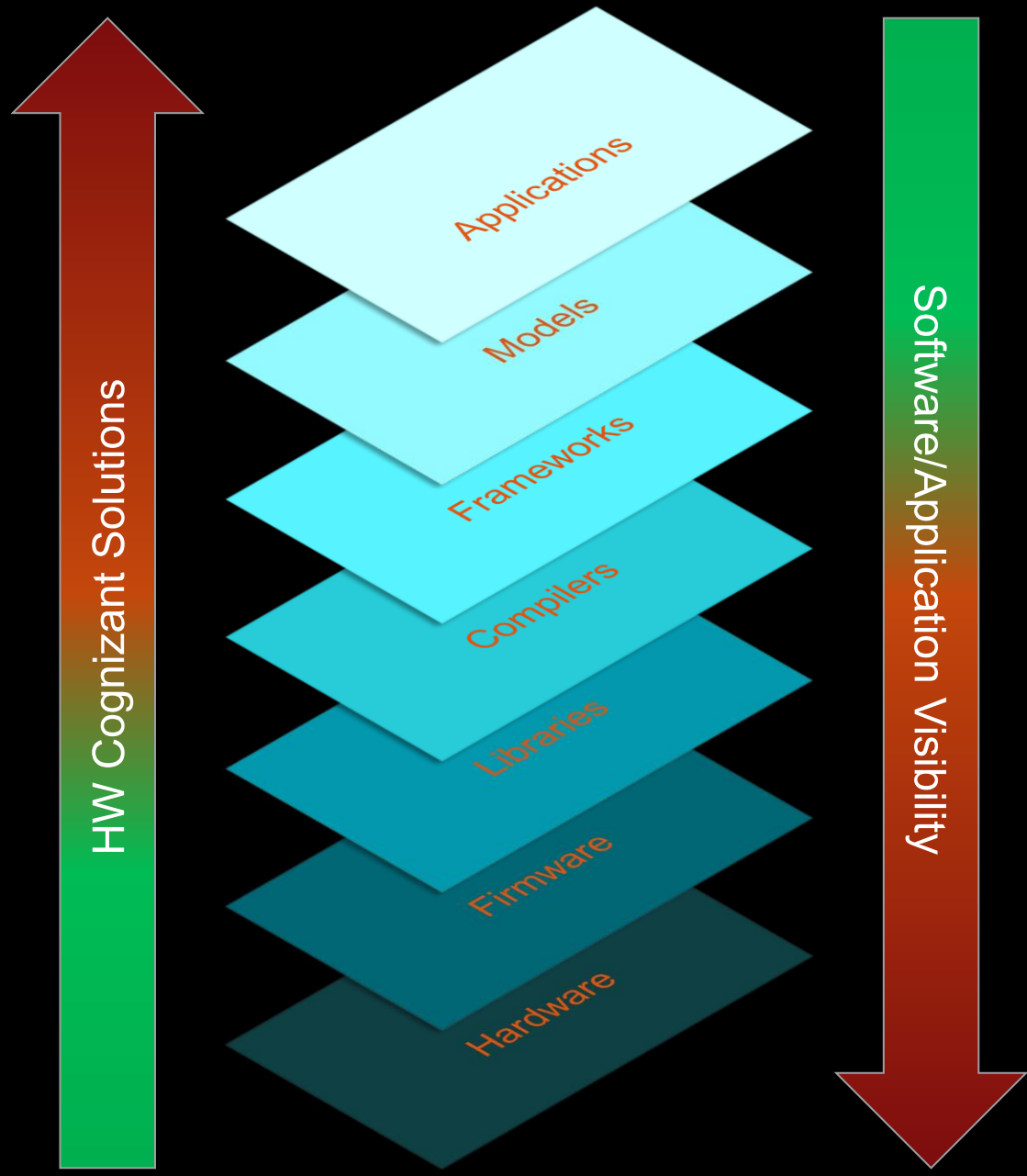
- **Infrastructure Impact:**
  - Data Center level voltage swings
- **Operational & Procurement Costs → TCO**
- **Environmental Impact:**
  - Growing carbon footprint

$$perf = f(power)$$

We Want Cross-Stack Optimization

- Power is not incorporated into decisions at the higher levels of software stack
- Lower levels oblivious of the software stack
- **It is critical to integrate all the "power concepts" into the SW stack**
- **And make SW more "visible" at the hardware levels**

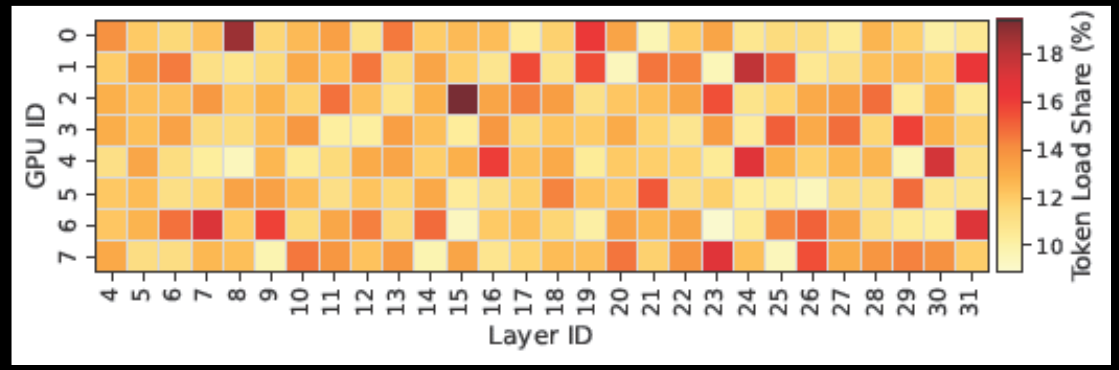
Goal:
Power management with higher returns on performance



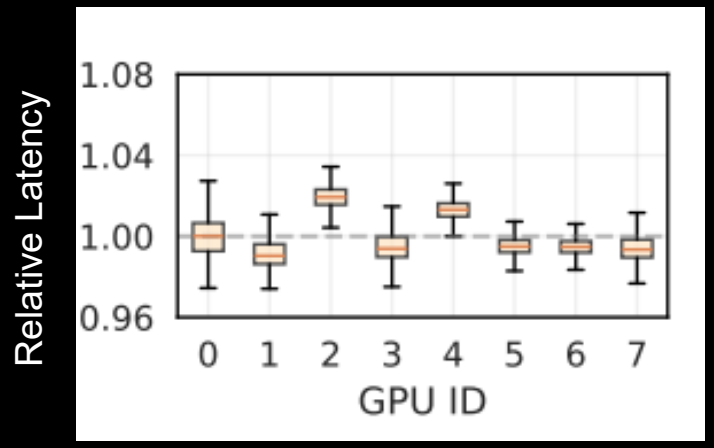
Power & Performance → Two Sides of the Same Coin

Model Sparsity

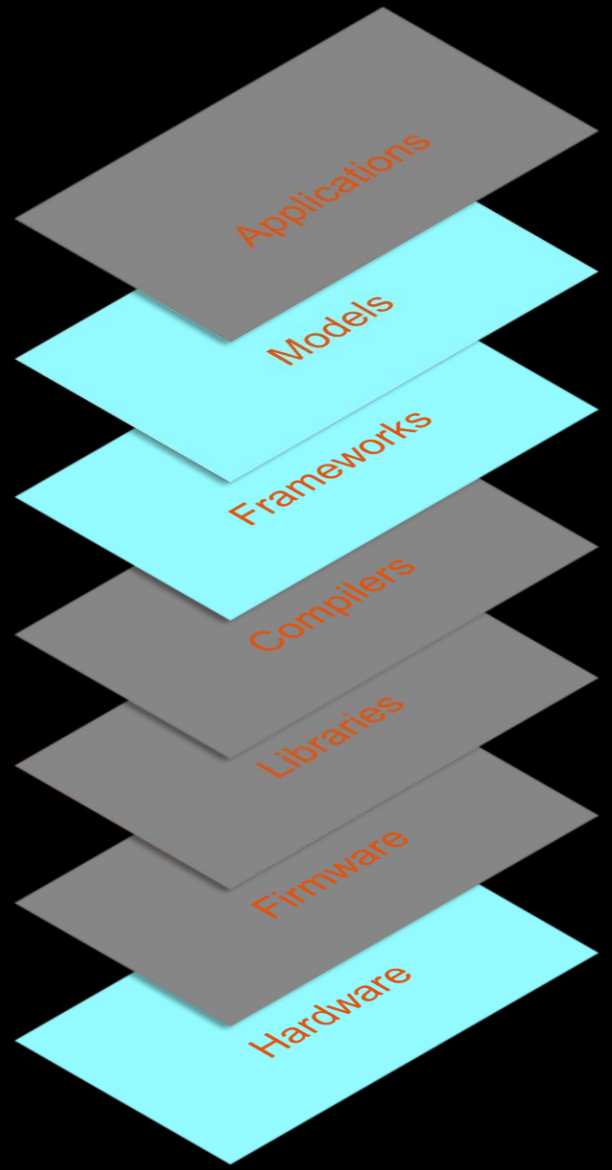
Sparse Expert Routing → Uneven GPU Usage } EPLB for load balancing



GPU Variation When Power Limited



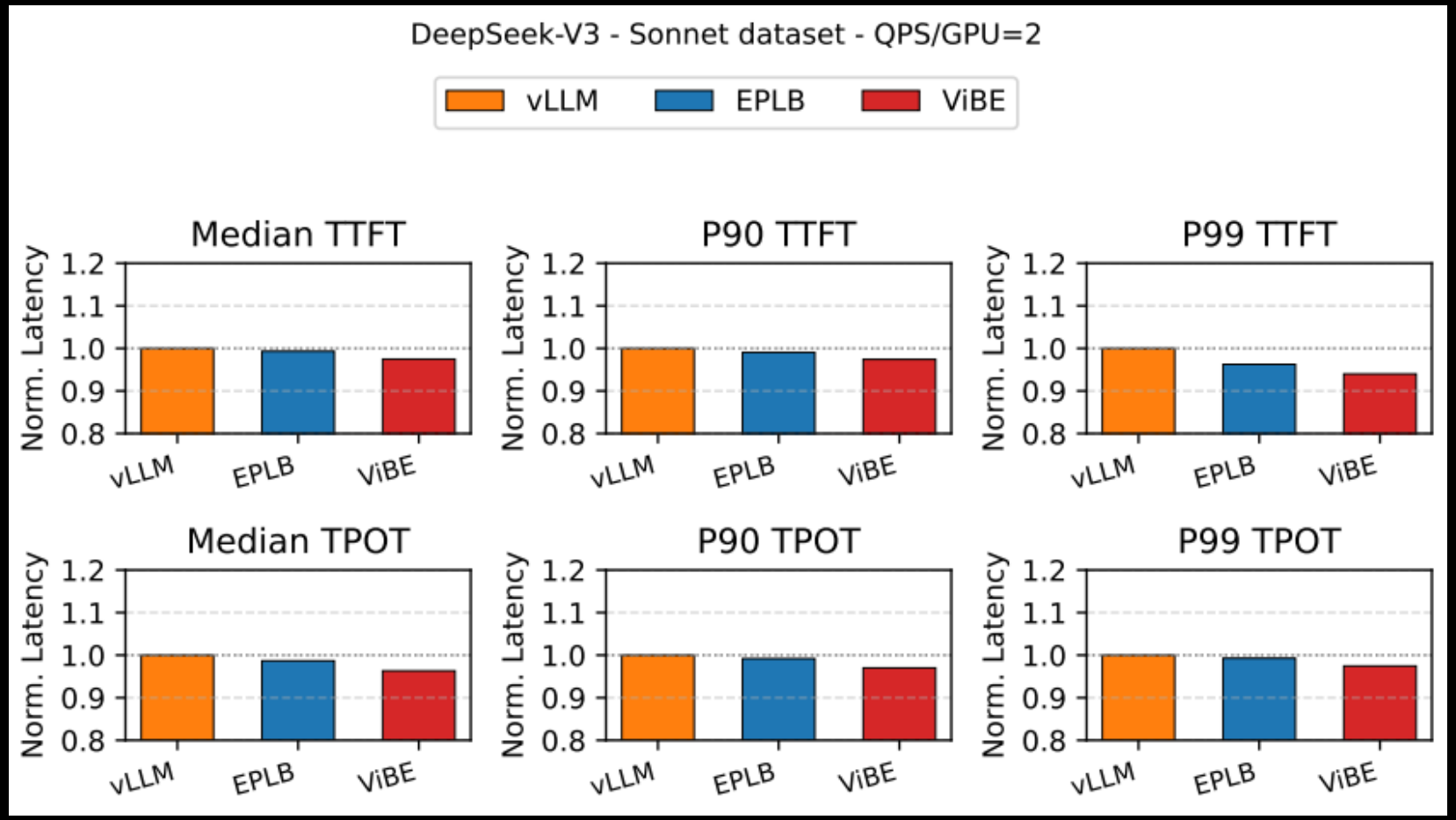
Solution: Lean into HW variability to improve application performance



ViBE: Co-Optimizing Workload Skew and Hardware Variability for MoE Serving

Model/HW Co-Optimization

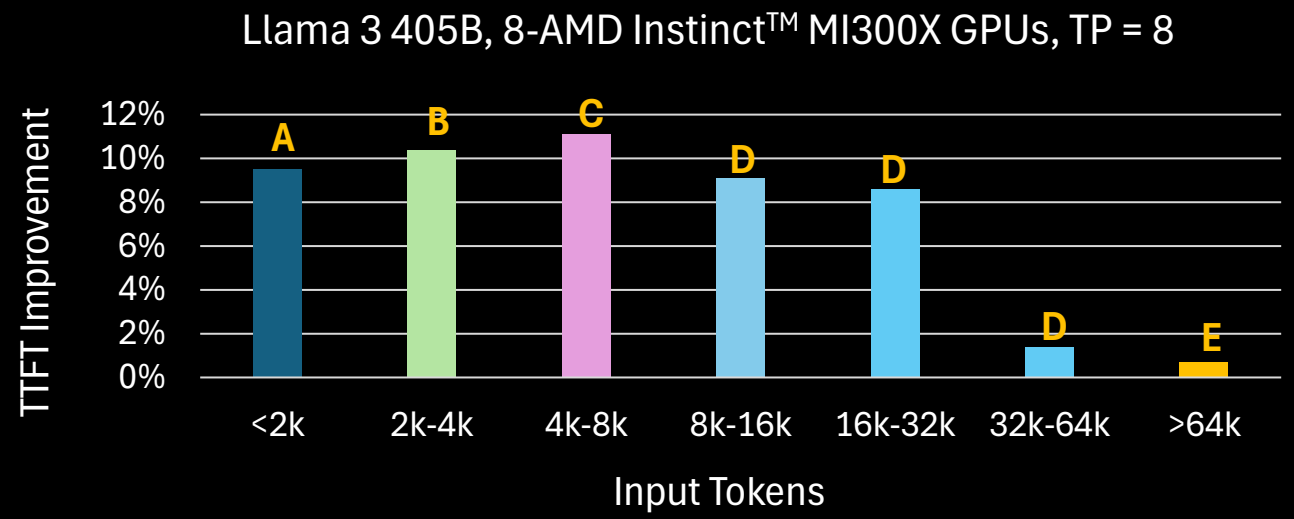
Lower Is Better



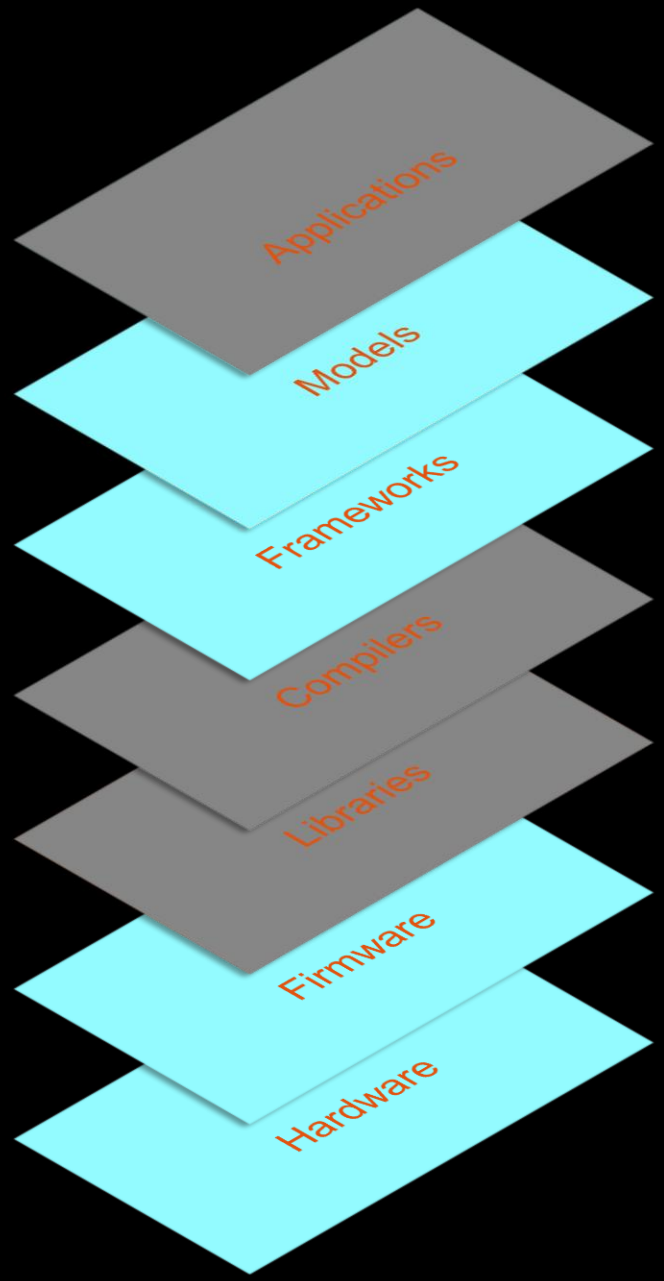
TTFT: Time To First Token
TPOT: Time Per Output Token

- Variations manifest when GPU is power limited
- Prefill (TTFT) is more power limited than decode (TPOT)

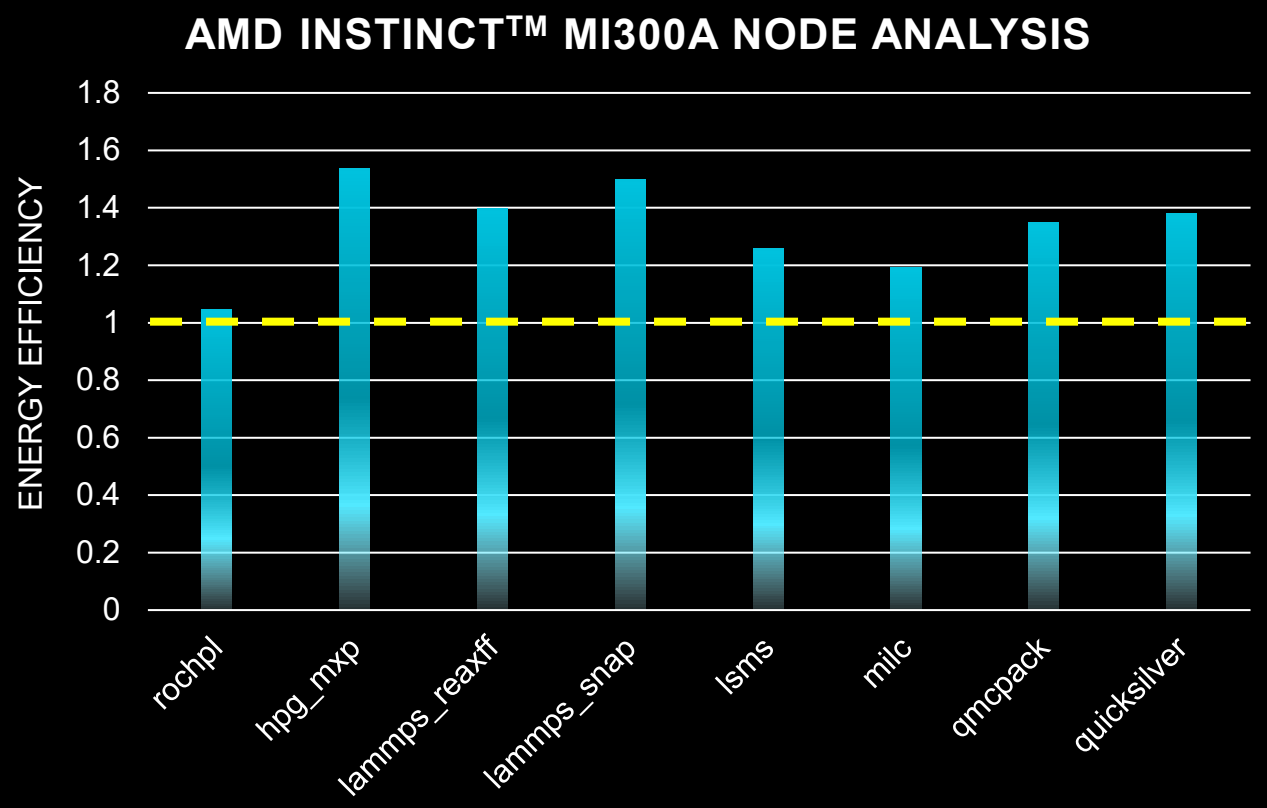
Power Management & Model Configuration

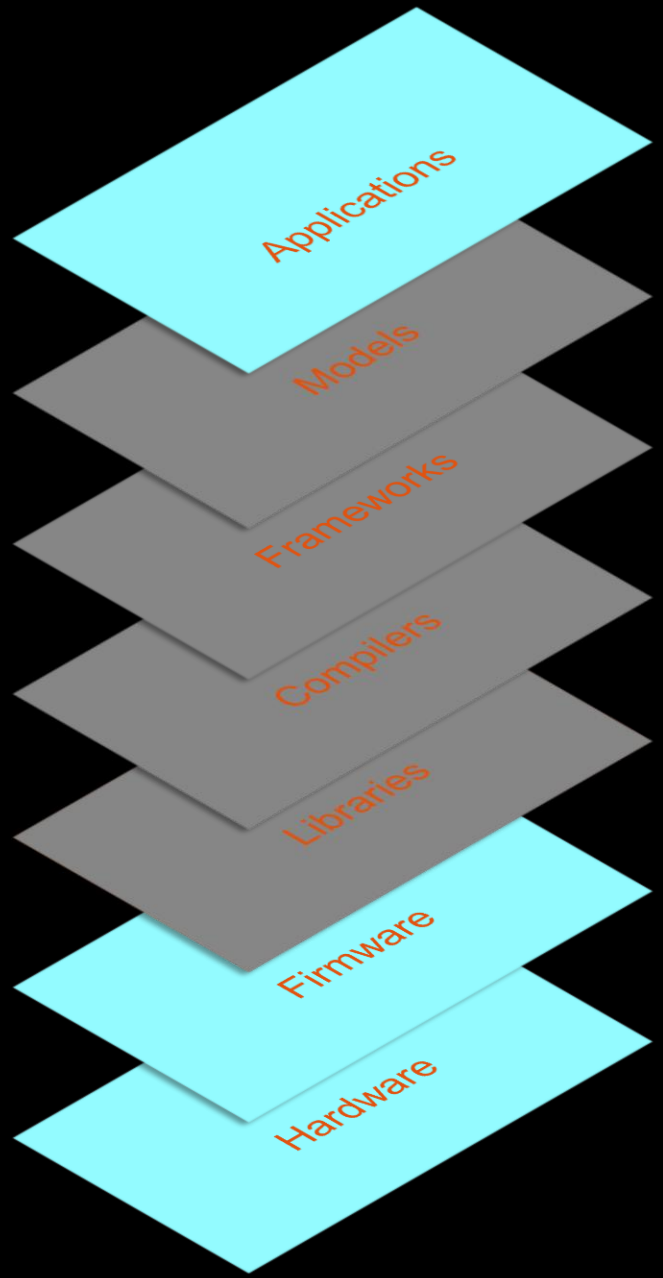


- Dense models such as Llama3 benefit from PMFW modifications based on input tokens
- Sparse models such as DeepSeek V3 prefer default settings



Energy Efficiency and Application Behavior

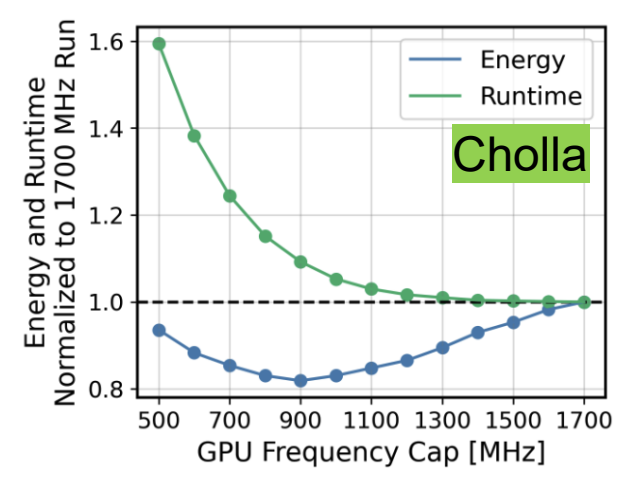


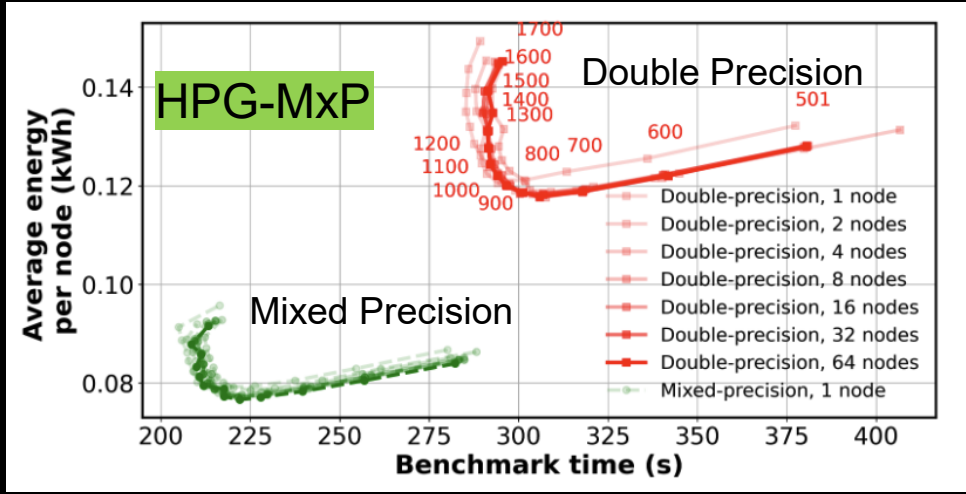


System Level Optimization

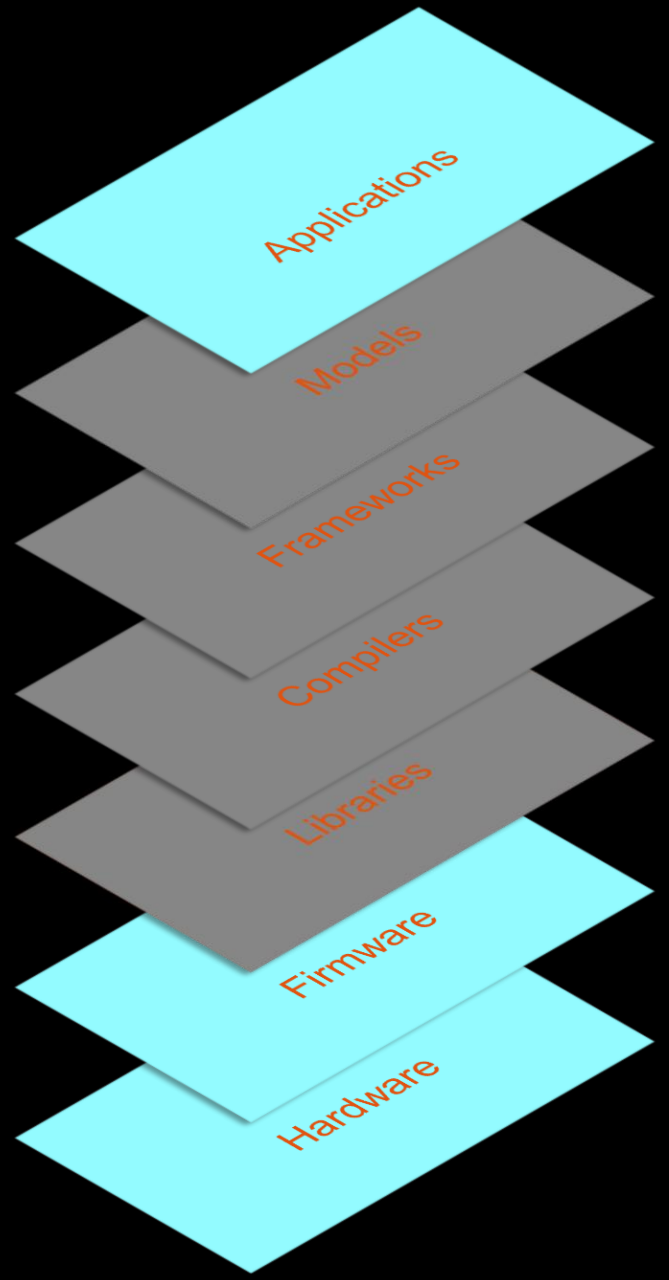
ORNL–Frontier 64 nodes

Workload	Energy Reduction	Perf Impact	Frequency Cap [MHz]
Cholla (CFD)	~15%	~2%	1100
HPG-MxP (linear algebra)	~15%	~3%	1000

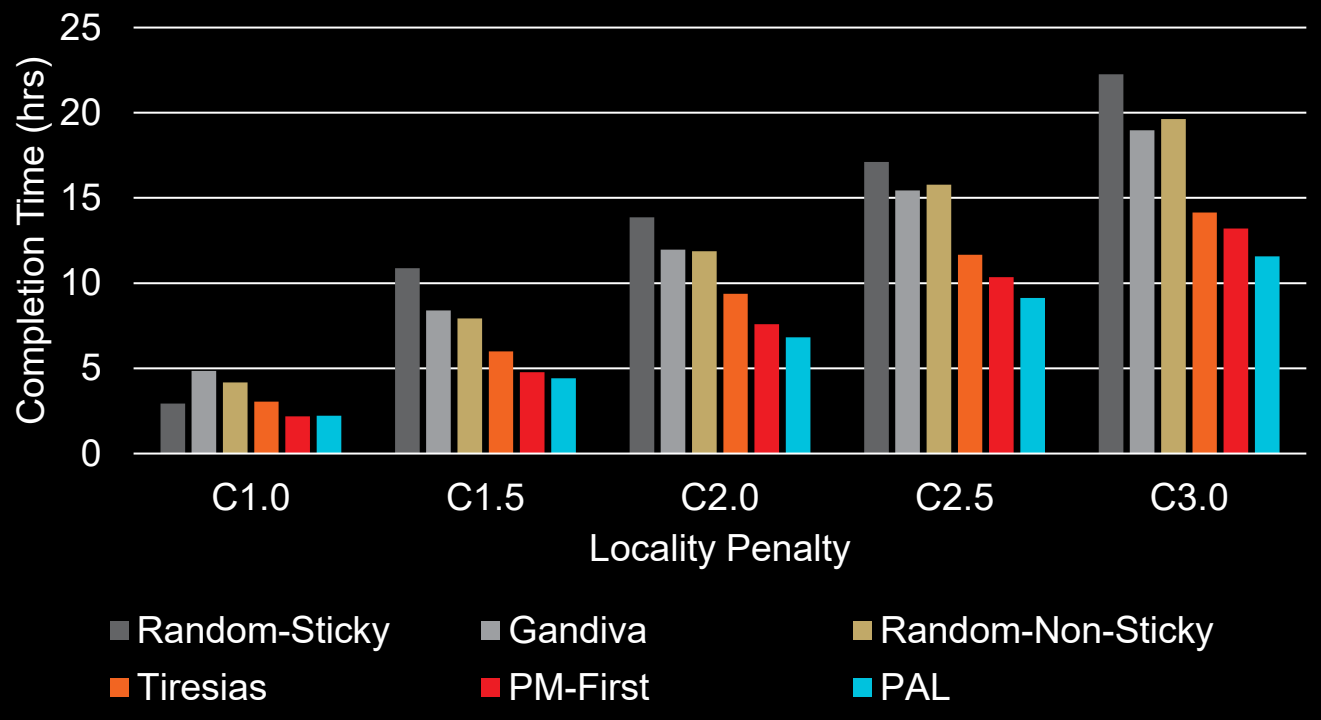




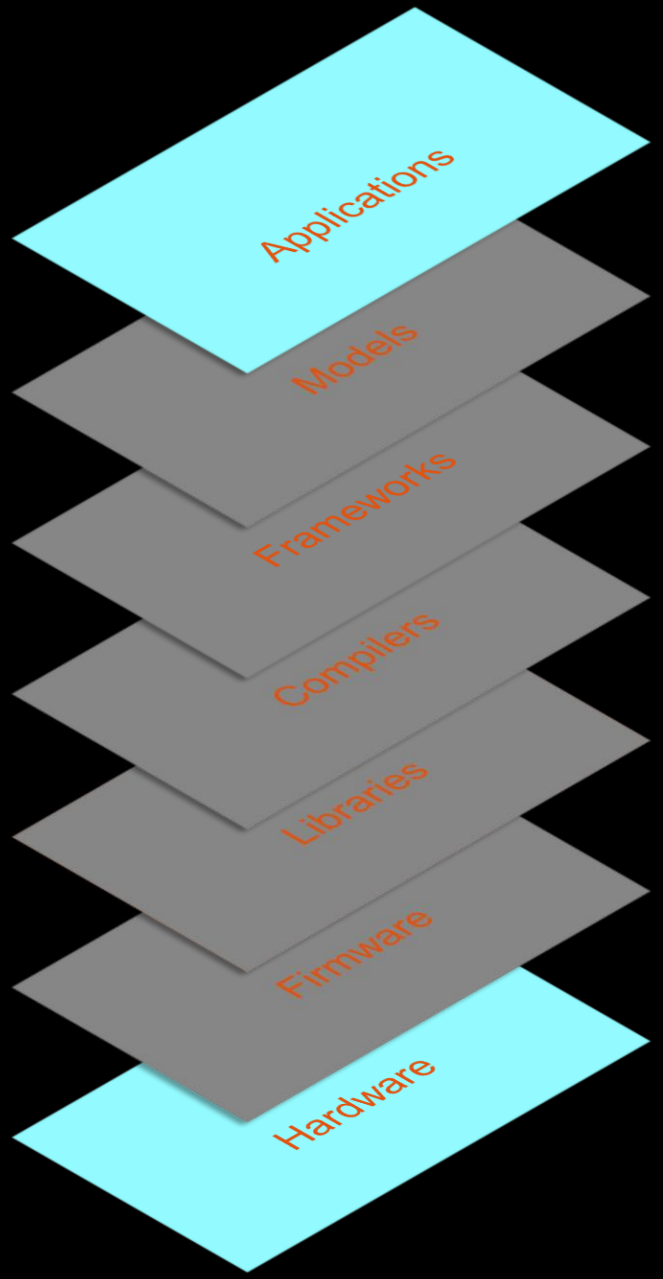
- *Compute-bound workloads (HACC-cosmology, HPL) do not benefit from frequency caps*



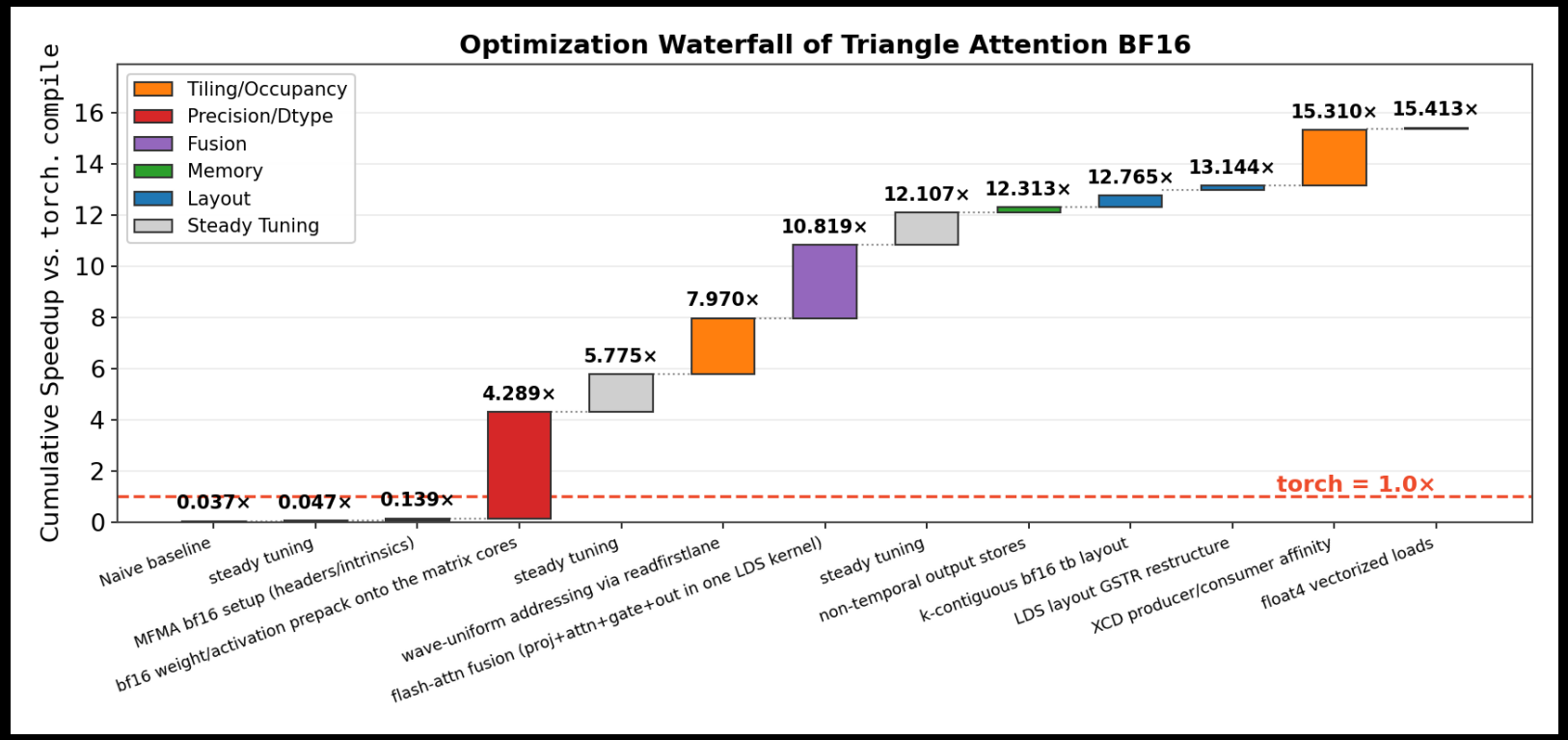
Balancing System Variability/Network Latency

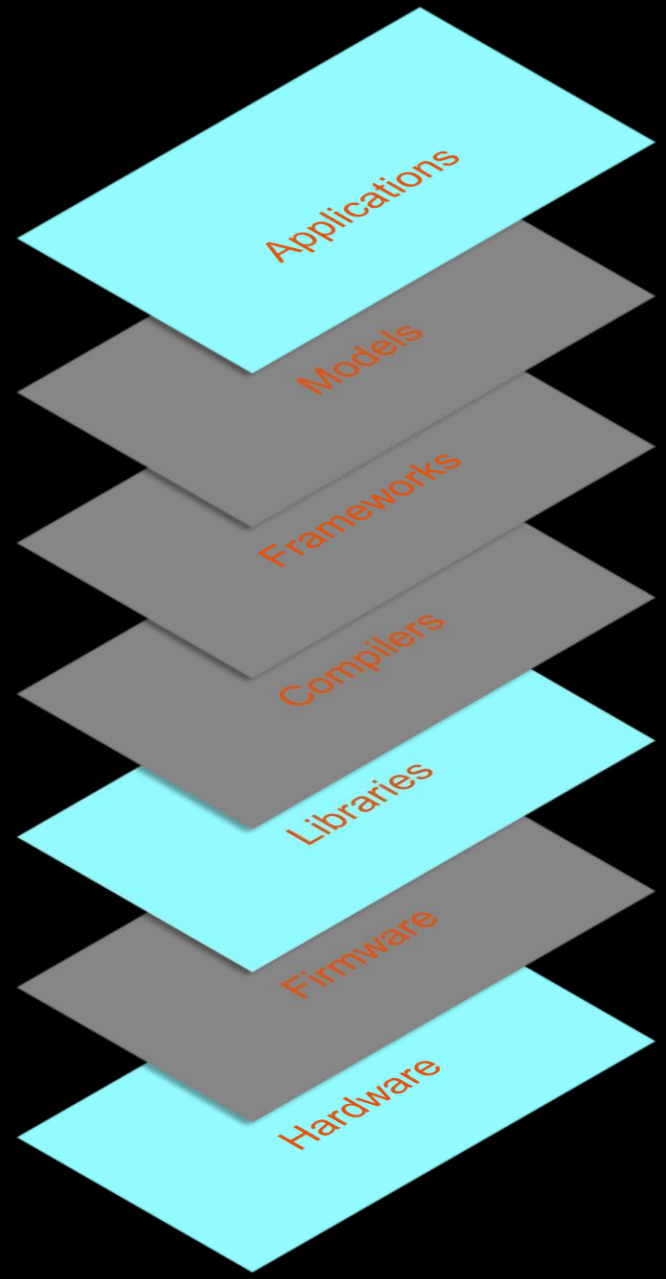


PAL: A Variability-Aware Policy for Scheduling ML Workloads in GPU Clusters



Performance Through Software Optimization





Algorithms presented in Hawkeye: Hardware-Aware GPU Kernel Optimization with Minimal Supervision

ROCm™ Hyperloom

An end-to-end agentic system.

Workload in.
Optimized, production code out.

No human in the loop.

Let me show you how.

ROCm™ Hyperloom

Agentic Workload Optimization

HOW IT WORKS

Input

Workload
Inference

Workload Profiling

TraceLens Agent

Intellikit
Low-level GPU Profiling

Magpie
Trace Collection

TraceLens
Trace Analysis

Gap Analysis and Bridge Planning

Kernel + Self-Evolving Optimization

Arbor – Self-Evolving Optimization Search

Agent loop: Think → Decide → Implement → Benchmark

Dynamic Specialist Agent + Knowledge Base

Kernel Optimization AGENTS

ClaudeCode | GEAK OpenAI Codex | OtherAPIs

Apply Changes

Prepare PR
with optimized code

The Loop Delivered.

14,000+ models. Autonomous optimization. No human in the loop.

At Scale

14,000+
models

Autonomous optimization across:
- Text Gen
- Text-to-Image
- Image-Text-to-Text

Automation

11,000+
positive gain

- Optimization + CI/CD running continuously
- Per-run leaderboard
- Zero manual benchmarking

No Human In The Loop

+38.36%
avg geomean gain

- Self-evolving Optimization search + Knowledge base
- GEAK
- Claude Code · OpenAI Codex

Lead Time

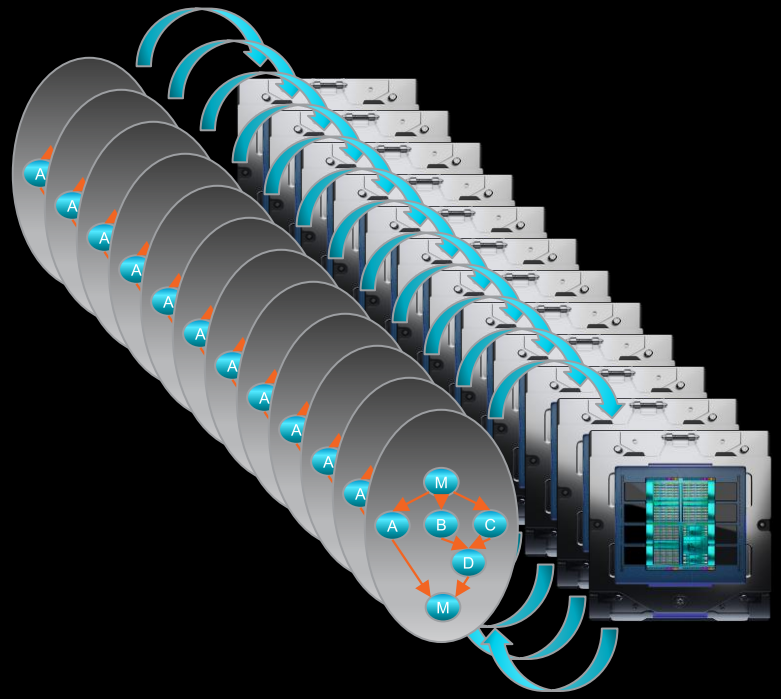
2
months

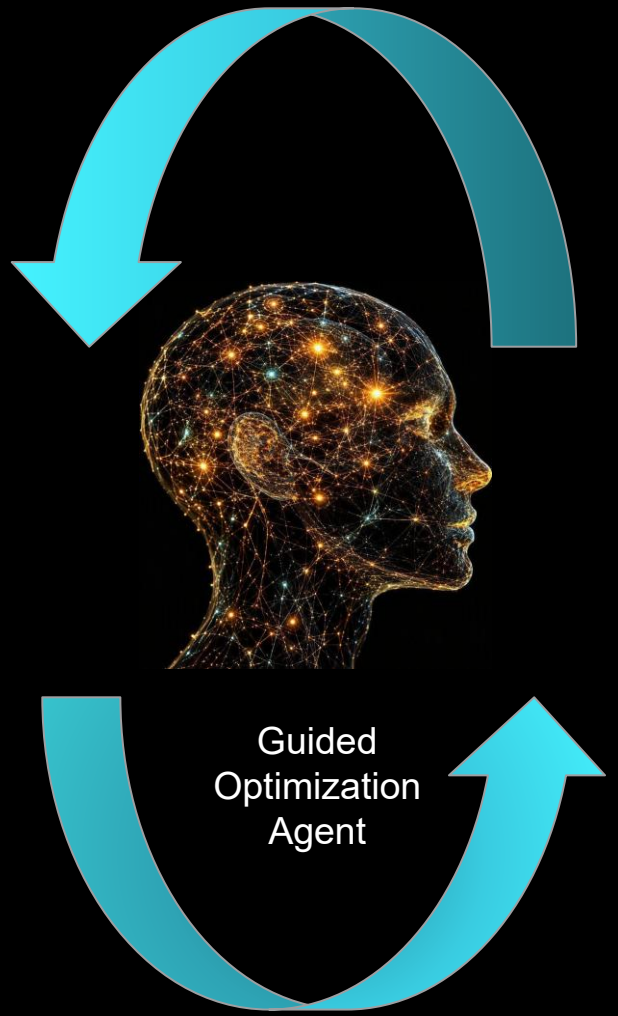
From prototype to production CI
- 437 merged PRs
- 275 fixed issues
- 8 releases shipped

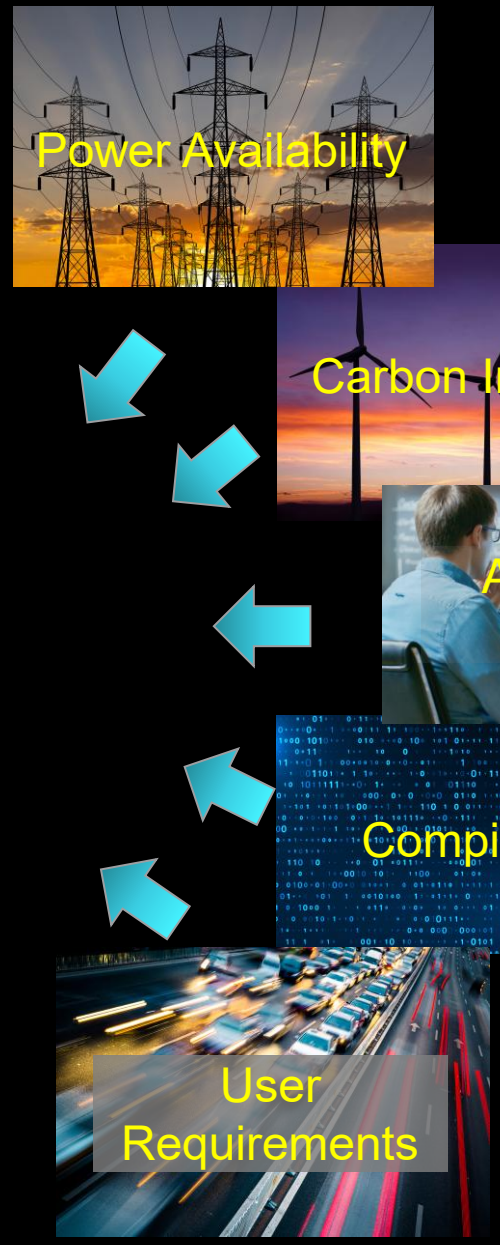
Top Models:

- GPTOSS 120B
- MiniMax M3
- Kimi K2.6
- Qwen 3.5
- DeepSeek v4
- GLM 5.2
- MiniMax2.5
- Qwen3 14B
- DS v4 Pro
- Kimi K2.7

Opportunity



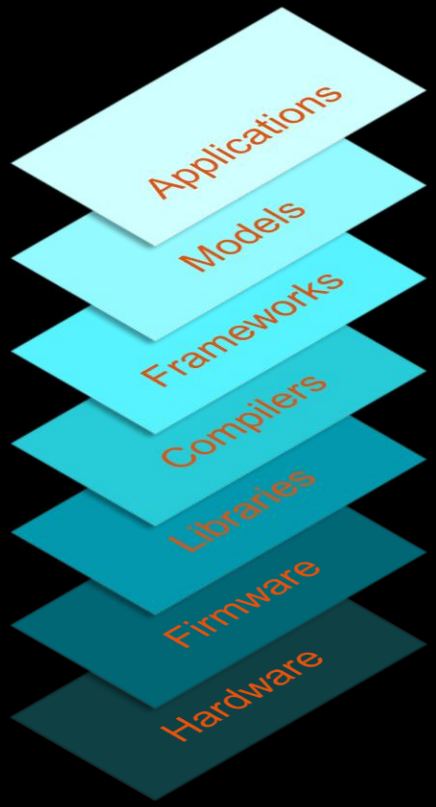




Complexity

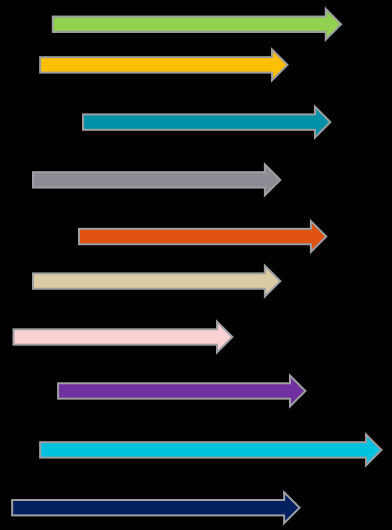
Communication

User/Operator requirements, application type/input, model configurations, framework settings, firmware settings, hardware type



Controls

Telemetry

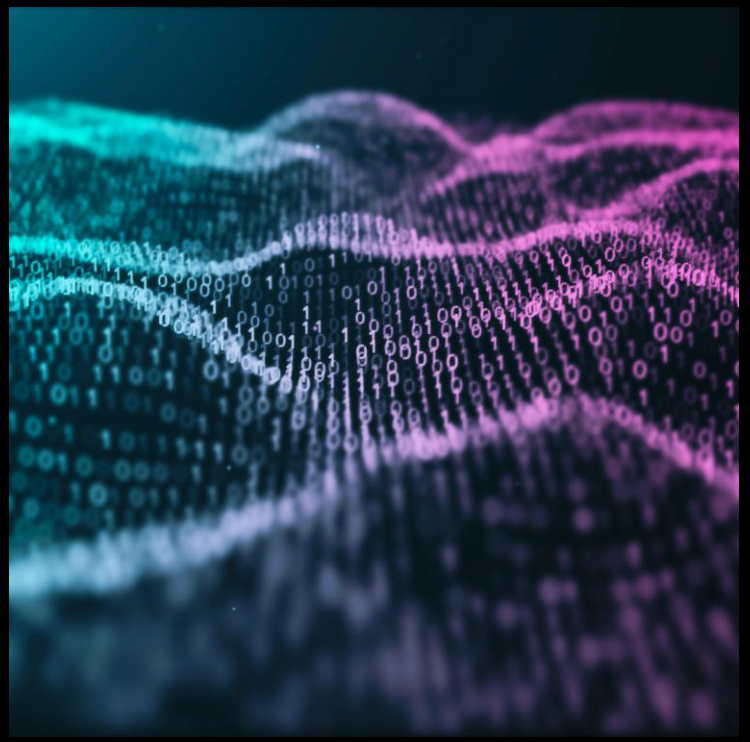


Secure *Agile* *Accurate* *Autonomous* *Fair* *Robust* *Stable* *Auditable* *Timely* ***SAFE***

Centralized and/or Distributed Intelligent Management

Conclusion

Configurable software/hardware plus detailed telemetry



Intelligent optimization across SLOs, efficiency goals, user requirements, etc

Agentic Frameworks For Bespoke Solutions

Copyright and disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate releases, for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED 'AS IS." AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION


AMD