Power-Aware Performance: Modeling for the Physical Limits of Extreme Computing*

Marko Scrbak
Senior Member of Technical Staff
AMD

Srilatha Manne
Senior Fellow
AMD

AMD
together we advance_

* Presented at the Workshop on Modeling & Simulation of Systems and Applications, August 2025.

“Power is the currency of performance!”

Samuel Naffziger

SVP and Corporate Fellow, AMD

A New Era in Computing

- **The Rise of AI:**
  - Explosive growth of Large Language Models (LLMs), deep learning, and complex AI workloads
  - Demand for unprecedented computational power

- **Extreme-Scale Computing:**
  - Massive data centers, supercomputers, and specialized AI accelerators
  - Pushing hardware to its limits

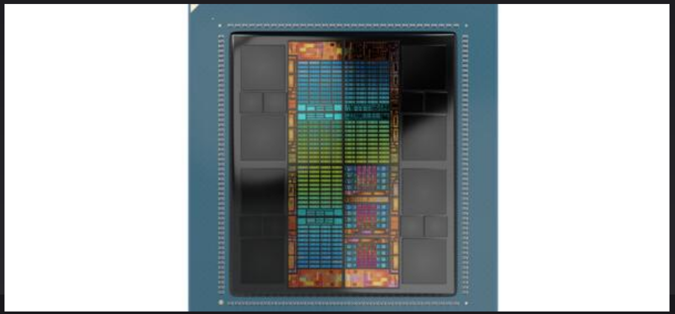
As we push the boundaries of extreme-scale computing in the AI era, power and energy have become first-order constraints

at every level of the solution stack!

Power Bottlenecks at Every Level

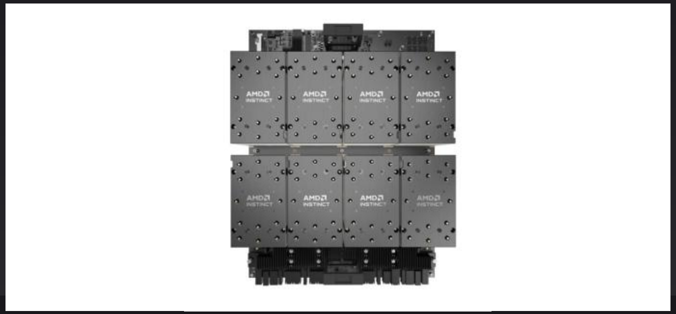
It's not just about energy efficiency

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

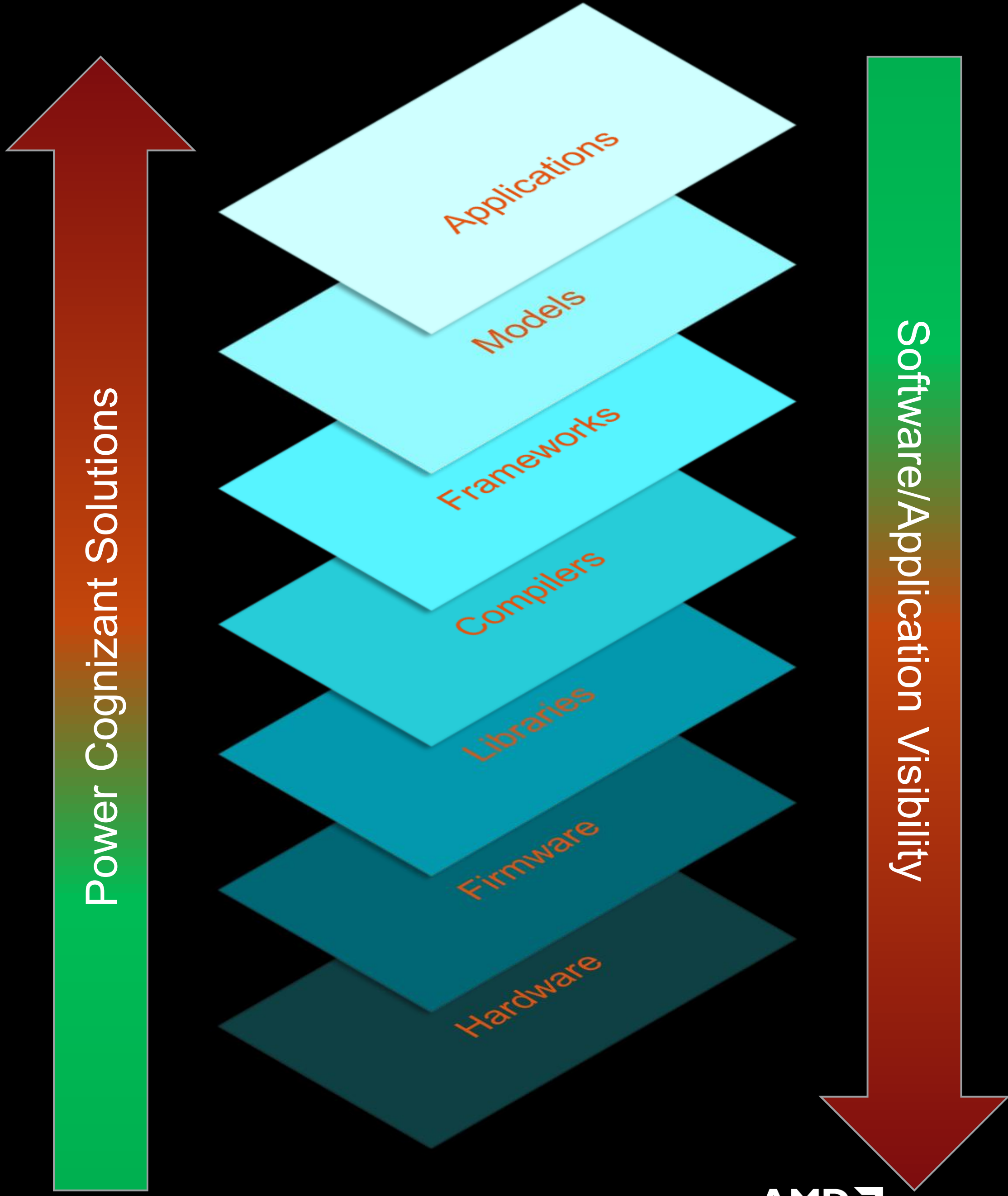
$$perf = f(power)$$

How do we model f ?

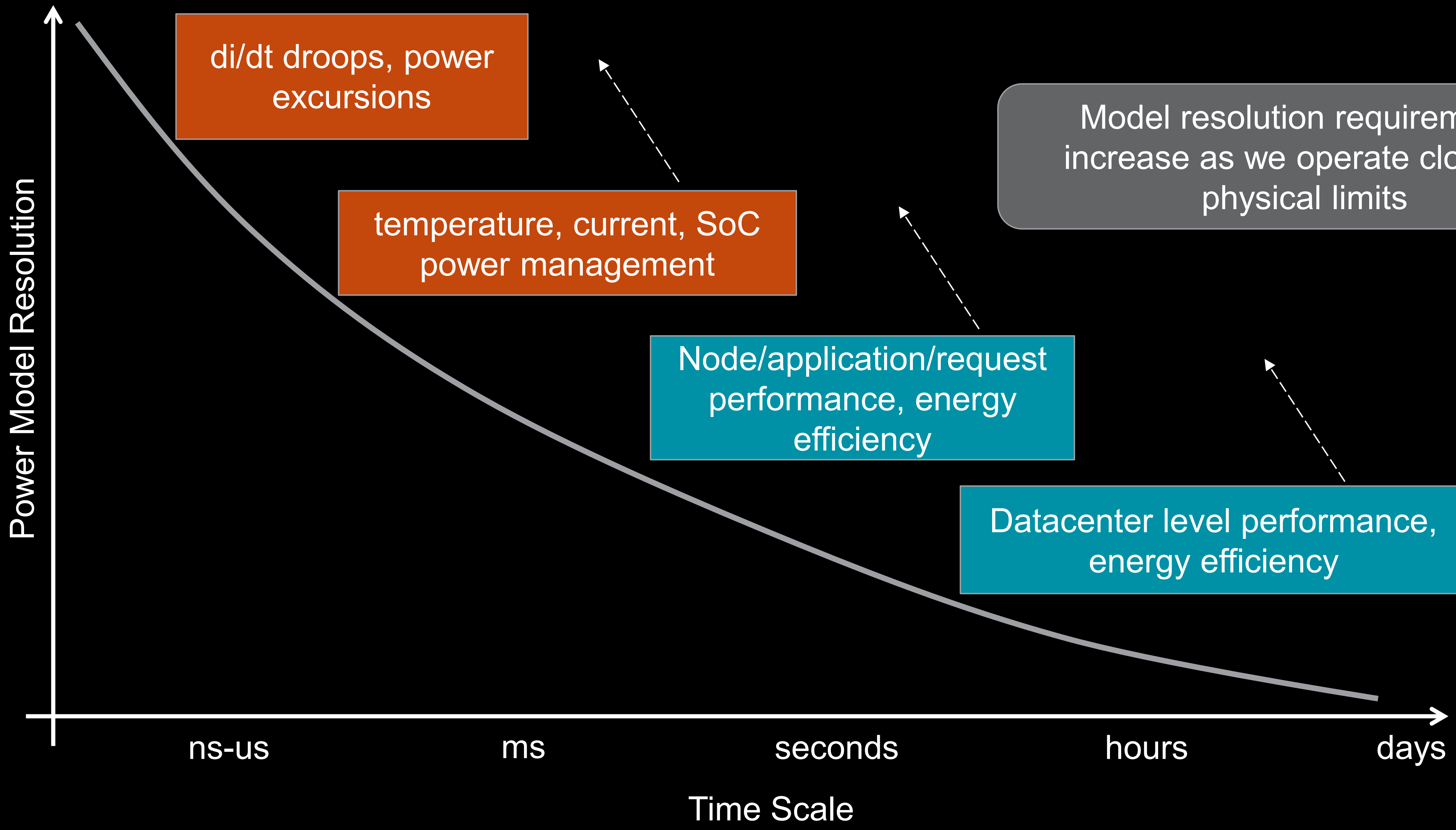
We Want Cross-Stack Transparency

- 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 models with higher returns on performance



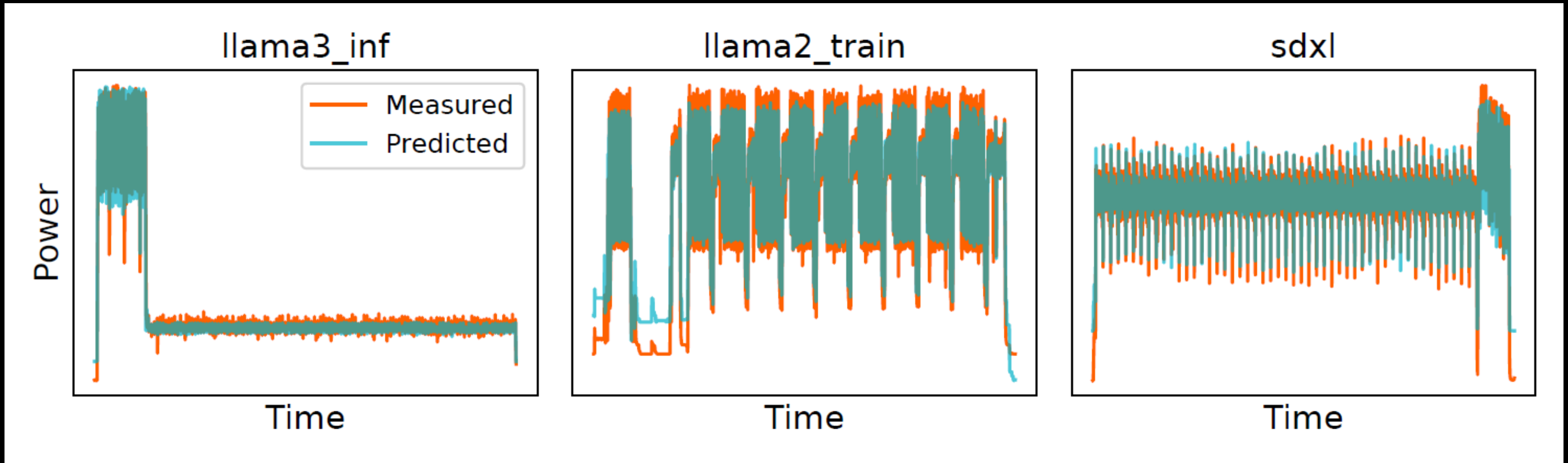
How Much Detail Do We Need?


Power Model Resolution
di/dt droops, power excursions
temperature, current, SoC power management
Node/application/request performance, energy efficiency
Datacenter level performance, energy efficiency
ns-us
ms
seconds
hours
days
Time Scale

Model resolution requirements increase as we operate closer to physical limits

Power Modeling (Estimation)

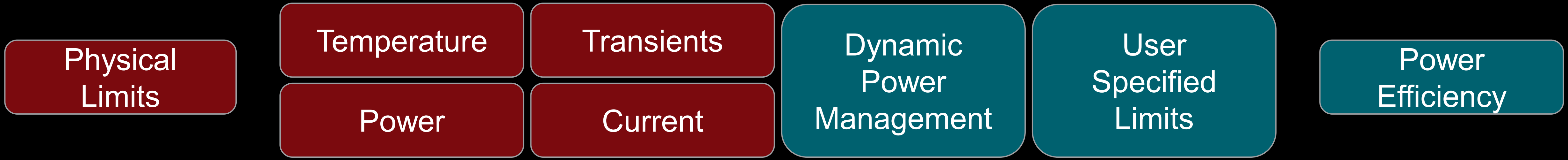
- **Power meter** => Correlate hardware events to power/energy values to build a digital power meter
  - Can be done at fine-grain resolution
  - E.g. Running Average Power Limit (RAPL)



- **Modelling performance as a function of power is complex (no simple correlation)**

What You Ask for IS NOT Always What You Get

- Performance models, compilers, SW optimizations, etc. assume a fixed frequency => **NOT THE CASE**
- Power management firmware (PMFW) and hardware limit operating frequency



Physical Limits

Temperature

Transients

Power

Current

Dynamic Power Management

User Specified Limits

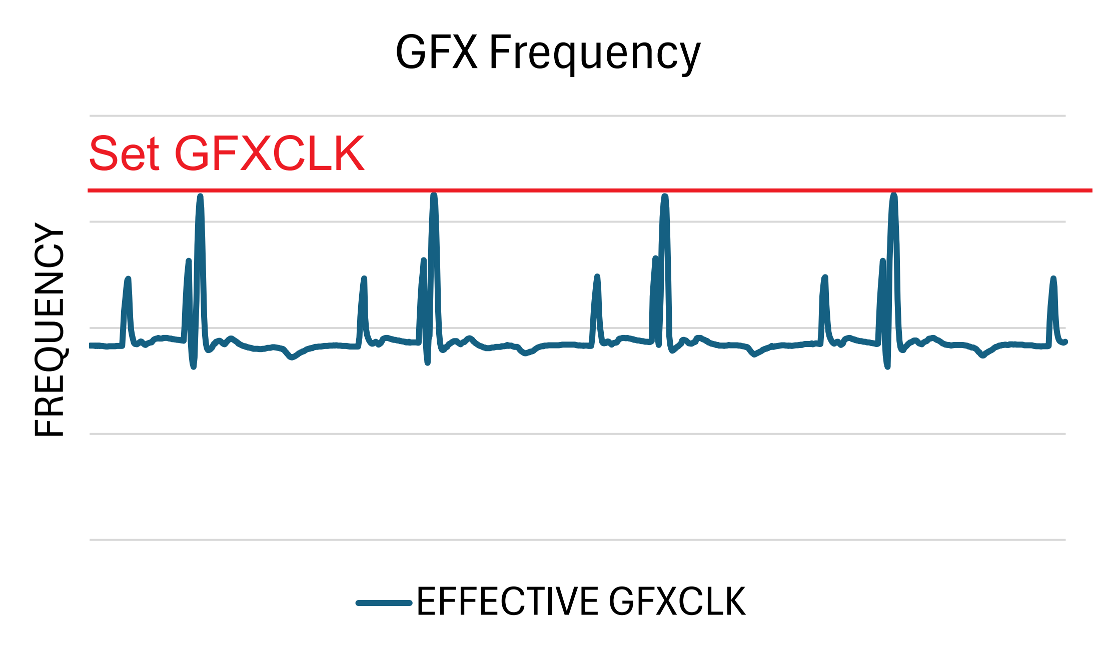
Power Efficiency

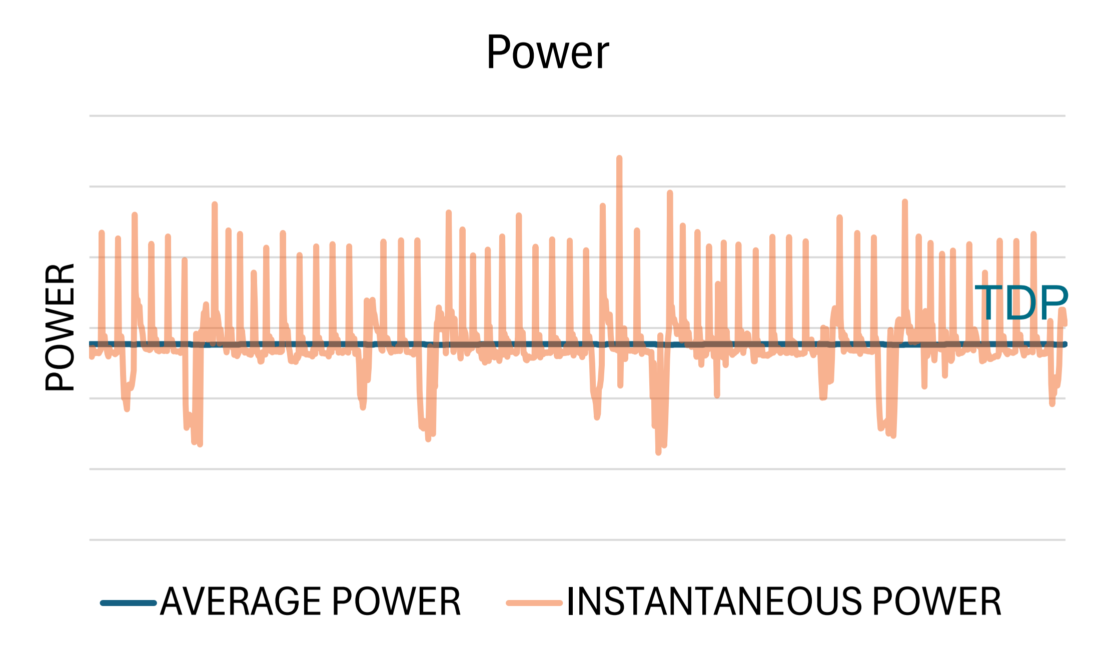
- Hardware clock modulation, firmware managed DVFS
- PMFW implements a set of algorithms and rules to manage/slosh power efficiently

Traditional Computing	Extreme Computing
$f_{effective} \approx f_{target}$	$\boldsymbol{f_{effective} < f_{target}}$

- Leads to performance non-determinism and variability and complicates modeling

Averages Hide the Details

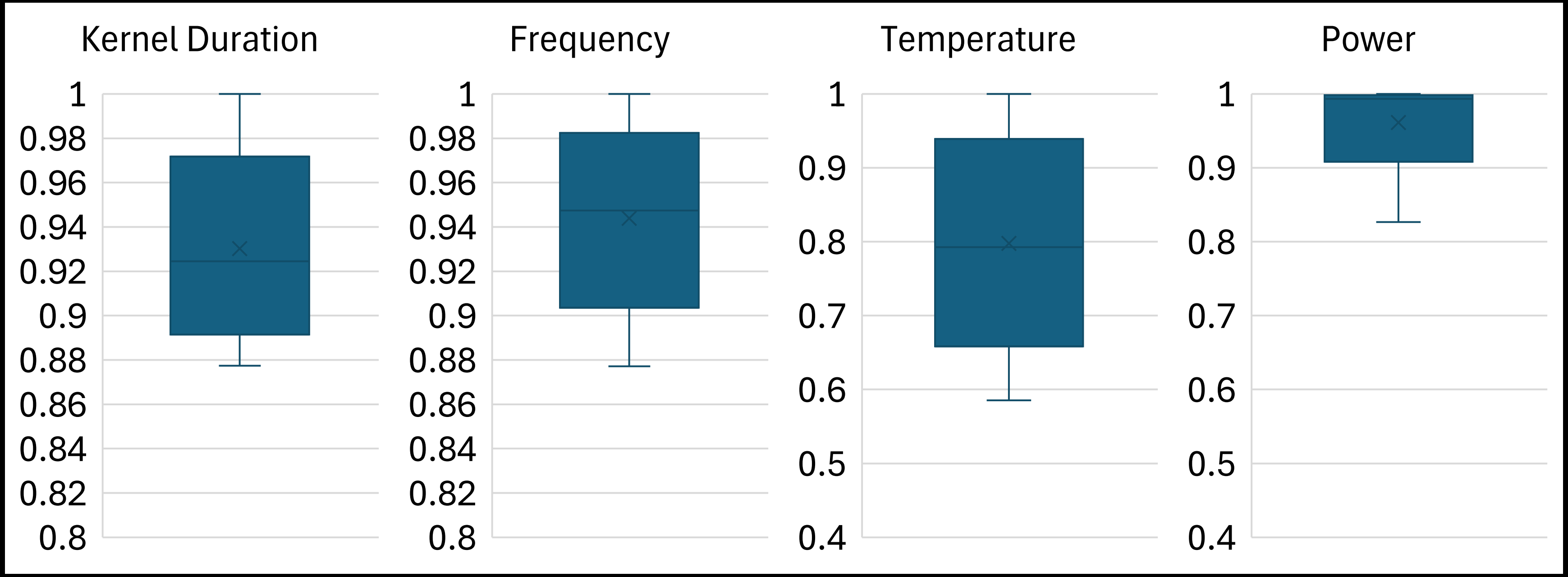




- AMD Instinct™ MI250 Accelerator running rocHPL with a set frequency of 1700MHz
- “Instantaneous Power” = high resolution average power (1ms)
- Zoomed in to 1s of total execution time
- High frequency power transients result in frequency modulation
  - Not visible when observing average power

At-scale Details Matter

- High temperature
  => higher leakage and less power available for performance
  => thermal throttling
- Slowest node impacts overall performance in synchronized run [1]



- Results for SGEMM on Longhorn cluster [1]

- LLAMA 3 diurnal 1-2% throughput variation based on time-of-day because of higher mid-day temperatures [2]

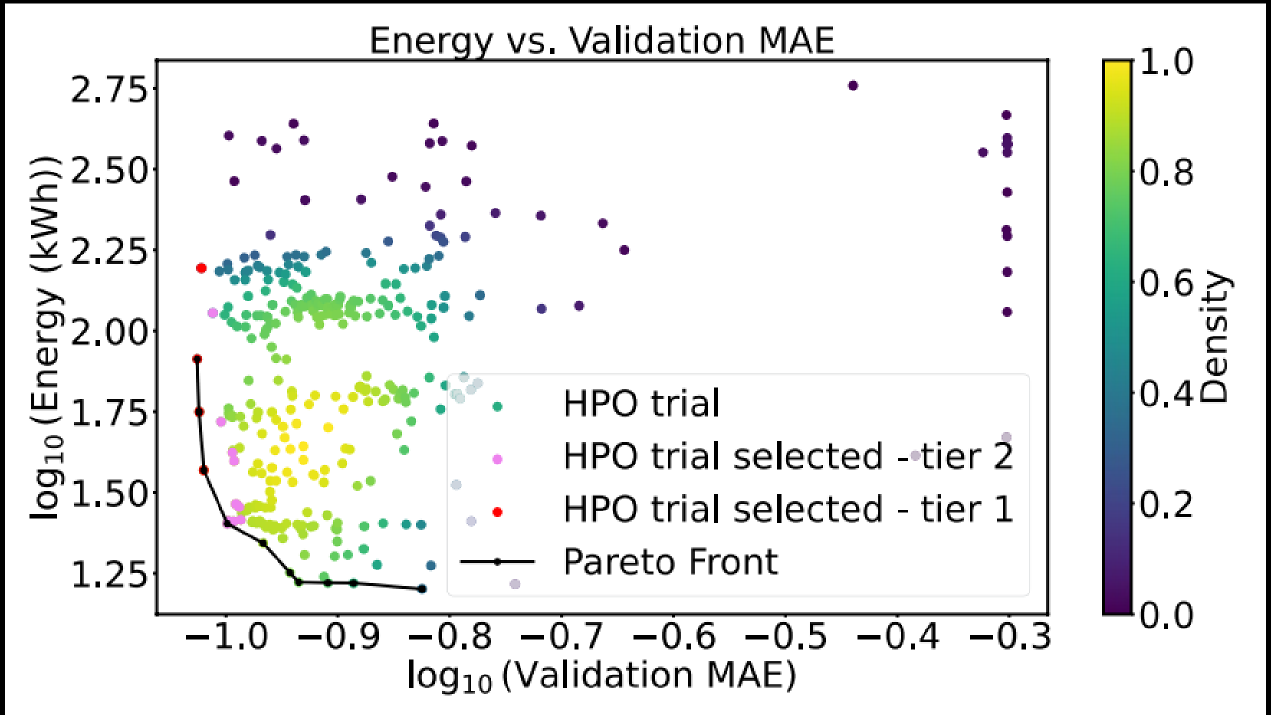
[1] Sinha, Prasoon, et al. "Not all GPUs are created equal: characterizing variability in large-scale, accelerator-rich systems." *SC22.* IEEE, 2022.
[2] Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv e-prints* (2024): arXiv-2407.

Omnistat – Details at Scale

Omnistat is a set of Python utilities and data collectors to support scale-out cluster telemetry targeting AMD Instinct™ GPUs/APUs

- Tracks metrics available via AMD SMI interface(s)
  - GPU and HBM utilization
  - GPU power, clock frequencies, thermals , power-caps
  - RAS error counts, throttling events, HW counters
- Low overhead (target 1% or less)
- Resource manager job tracking (SLURM, Flux, PBSPro)

Open-source: https://github.com/ROCm/omnistat

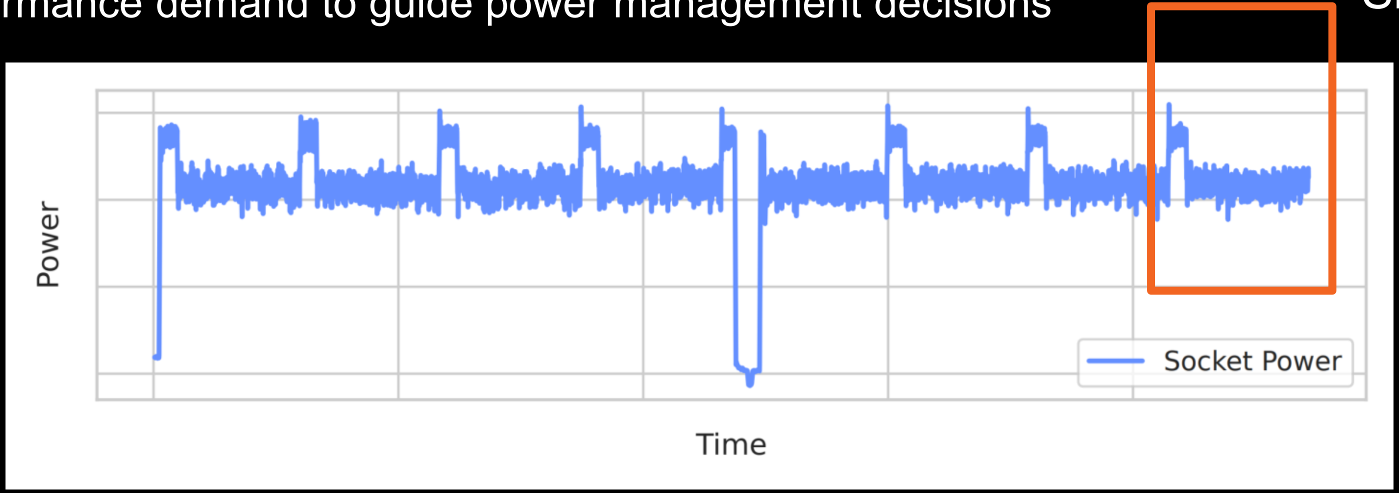


Energy efficiency hyperparameter optimization of graph foundational models trained on Frontier [3]

[3] Lupo Pasini, Massimiliano, et al. "Scalable training of trustworthy and energy-efficient predictive graph foundation models for atomistic materials modeling: a case study with HydraGNN." *The Journal of Supercomputing* 81.4 (2025): 618.

Reaction Vs. Prediction

- Current models are reactive
  - Models base their behavior on past actions
- We want predictive models, perhaps using AI or hints
  - Predict future performance demand to guide power management decisions



Llama 3.1 8B [2] – 4K/output input [reference]

- Examples
  - Predict which VMs will be active when, based on learning over large amounts of past behavior
  - LLMs repeat the same layer N times (32 for 8B, 126 for 405B)
  - Predict number of output tokens

Modeling $perf = f(power)$ is Difficult … but Necessary

Takeaways

Frequency is not constant! (What You Ask for IS NOT Always What You Get!)

Performance non-determinism and variability complicate modeling

Predictive models are more attractive

Averages hide the details and at scale details matter

Expose fine-grain telemetry to the SW stack

Create interfaces between HW and SW for power

References

- [1] Sinha, Prasoon, et al. "Not all GPUs are created equal: characterizing variability in large-scale, accelerator-rich systems." *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2022.

- [2] Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv e-prints* (2024): arXiv-2407.

- [3] Lupo Pasini, Massimiliano, et al. "Scalable training of trustworthy and energy-efficient predictive graph foundation models for atomistic materials modeling: a case study with HydraGNN." *The Journal of Supercomputing* 81.4 (2025): 618.

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 for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, 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 CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Third-party content is licensed to you directly by the third party that owns the content and is not licensed to you by AMD. ALL LINKED THIRD-PARTY CONTENT IS PROVIDED "AS IS" WITHOUT A WARRANTY OF ANY KIND. USE OF SUCH THIRD-PARTY CONTENT IS DONE AT YOUR SOLE DISCRETION AND UNDER NO CIRCUMSTANCES WILL AMD BE LIABLE TO YOU FOR ANY THIRD-PARTY CONTENT. YOU ASSUME ALL RISK AND ARE SOLELY RESPONSIBLE FOR ANY DAMAGES THAT MAY ARISE FROM YOUR USE OF THIRD-PARTY CONTENT.

AMD