

3RD INTERNATIONAL WORKSHOP ON ENERGY EFFICIENCY WITH SUSTAINABLE PERFORMANCE

Sub-Cycle Measurement of AI Workload Energy, Flexibility, and Synchronization on Production GPU Racks

Kevin Menear, Deepthi Vaidhynathan, Kumaraguru Prabakar

National Laboratory of the Rockies

National Laboratory of the Rockies

SC26 hpc unites | Chicago

OPERATIONS ANALYSIS DATA

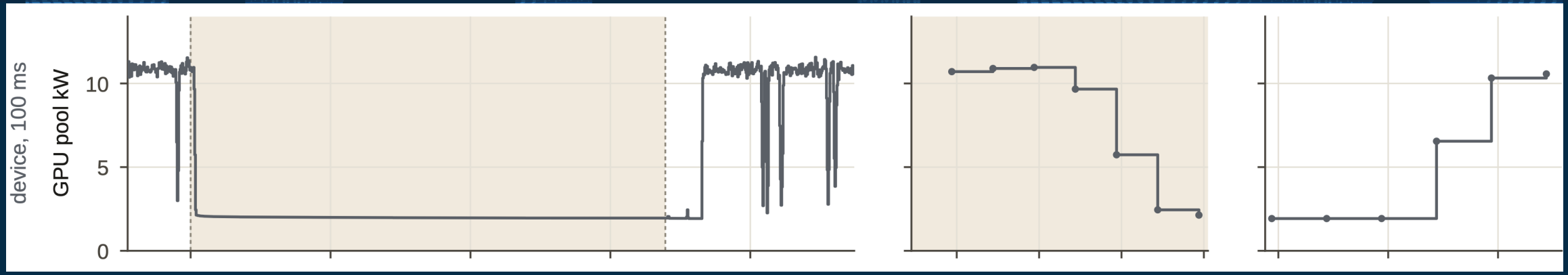
OPERATIONS
ANALYSIS
DATA
TELEMETRY

3-Tier Telemetry Stack

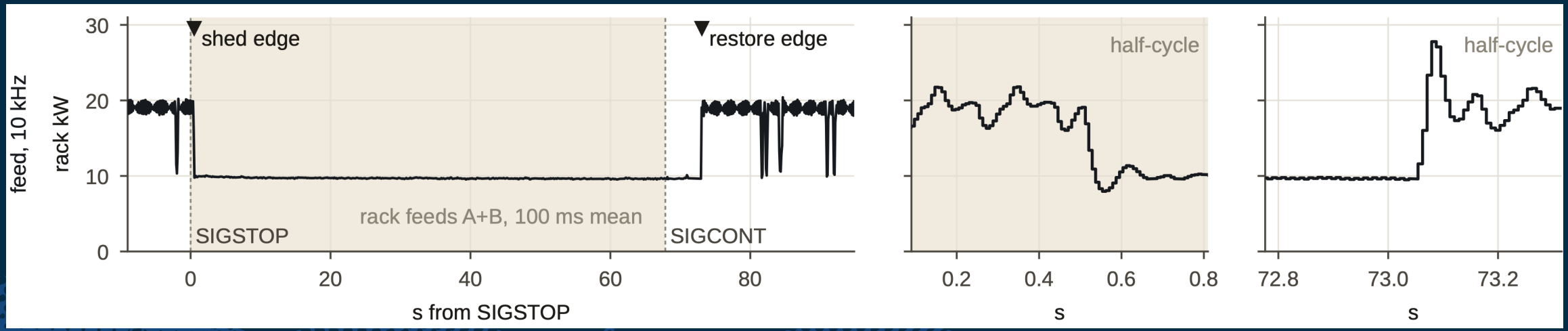
- **Facility meters**
  - Real power and current per circuit (ABB breakers).
  - Read every 5 seconds, archived as 1-minute averages.
- **Onboard meters**
  - Per-GPU power from NVIDIA DCGM, updated every 100 ms.
- **10 kHz meters on the power distribution unit (PDU) feeds:**
  - Current on each PDU feed, with voltage from the panel main feed.
  - Two PDUs per rack, with 9–10 nodes of 4 NVIDIA H100 GPUs each (36–40 GPUs per rack).

What do 10 kHz meters reveal for GPU workloads?

Onboard GPU meters



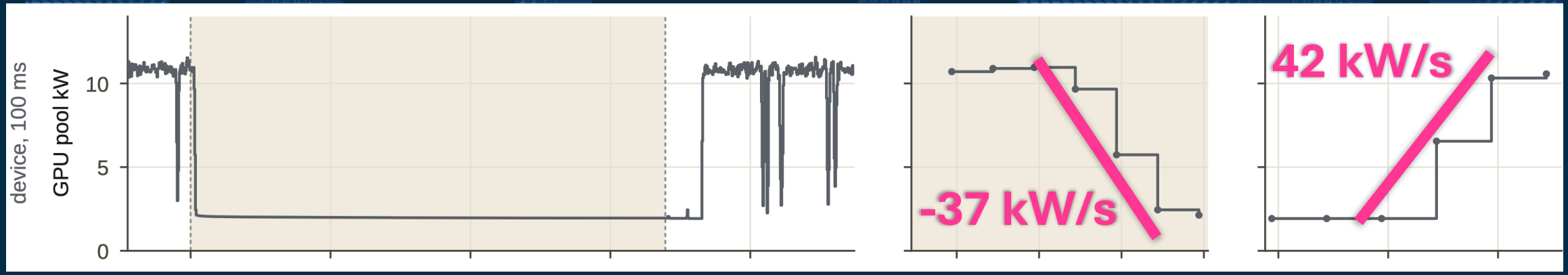
10 kHz meters



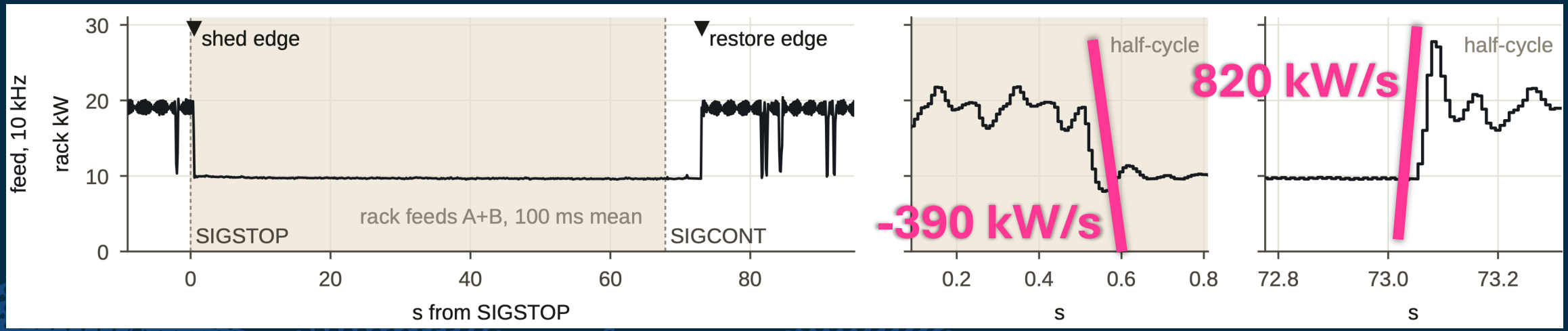
16-GPU training job paused (SIGSTOP), held ~60 s, resumed (SIGCONT)

SC26

Onboard GPU meters



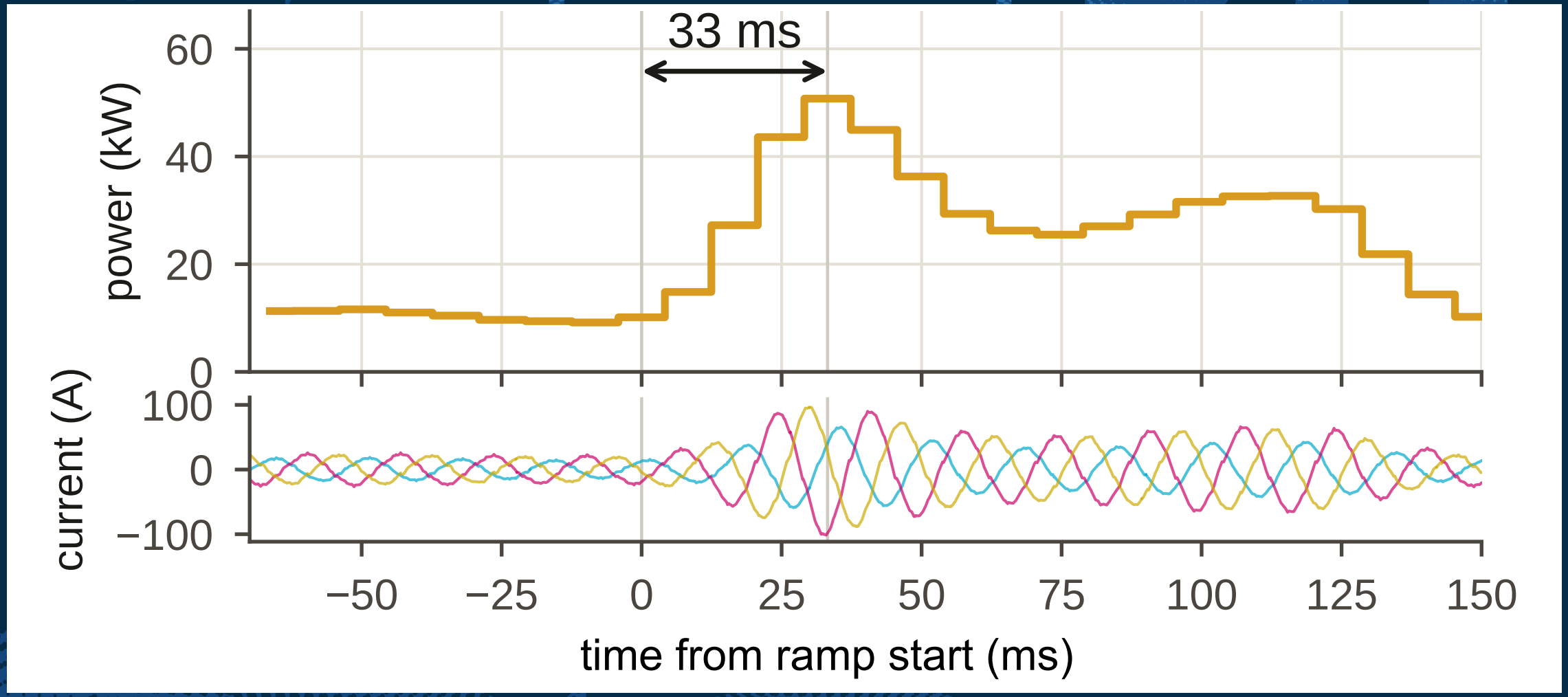
10 kHz meters



Real ramp rates are faster by an order of magnitude

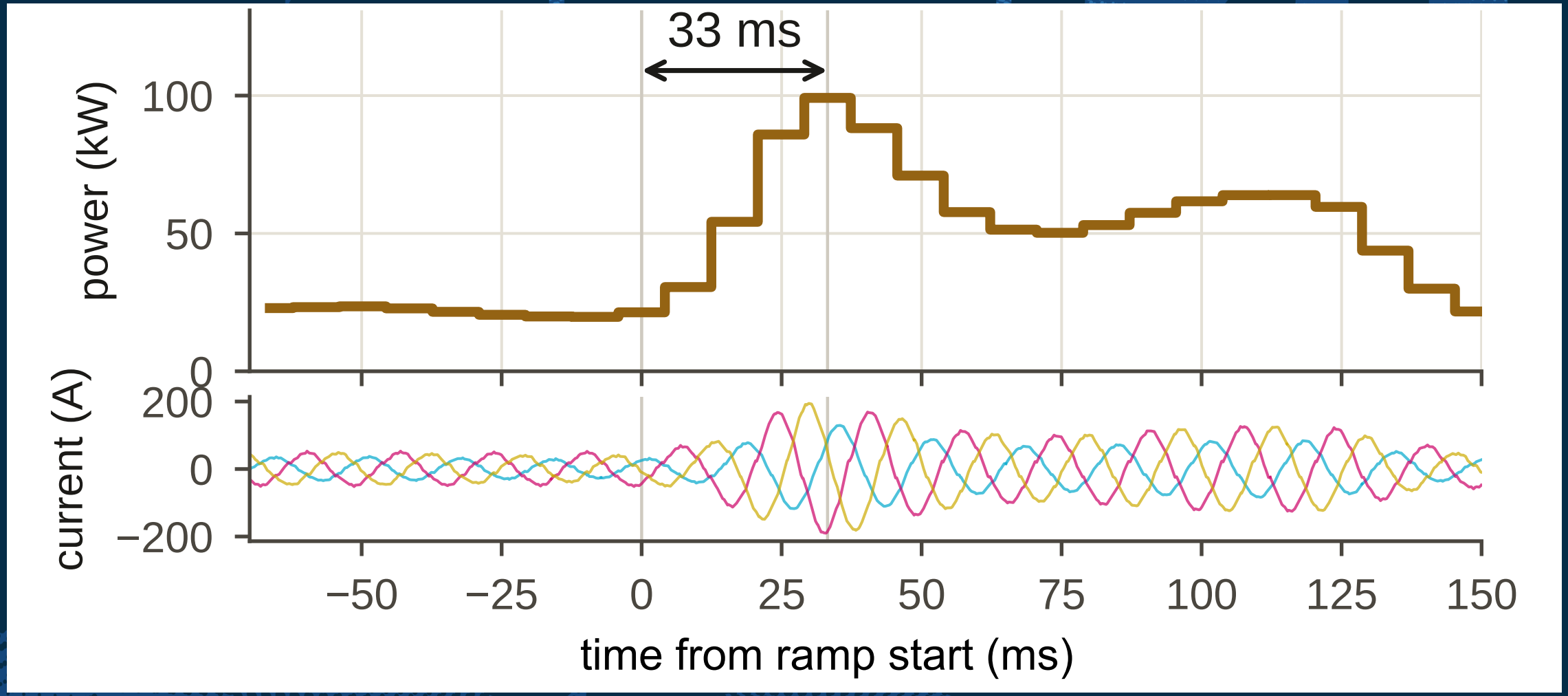
SC26

One-Rack Training



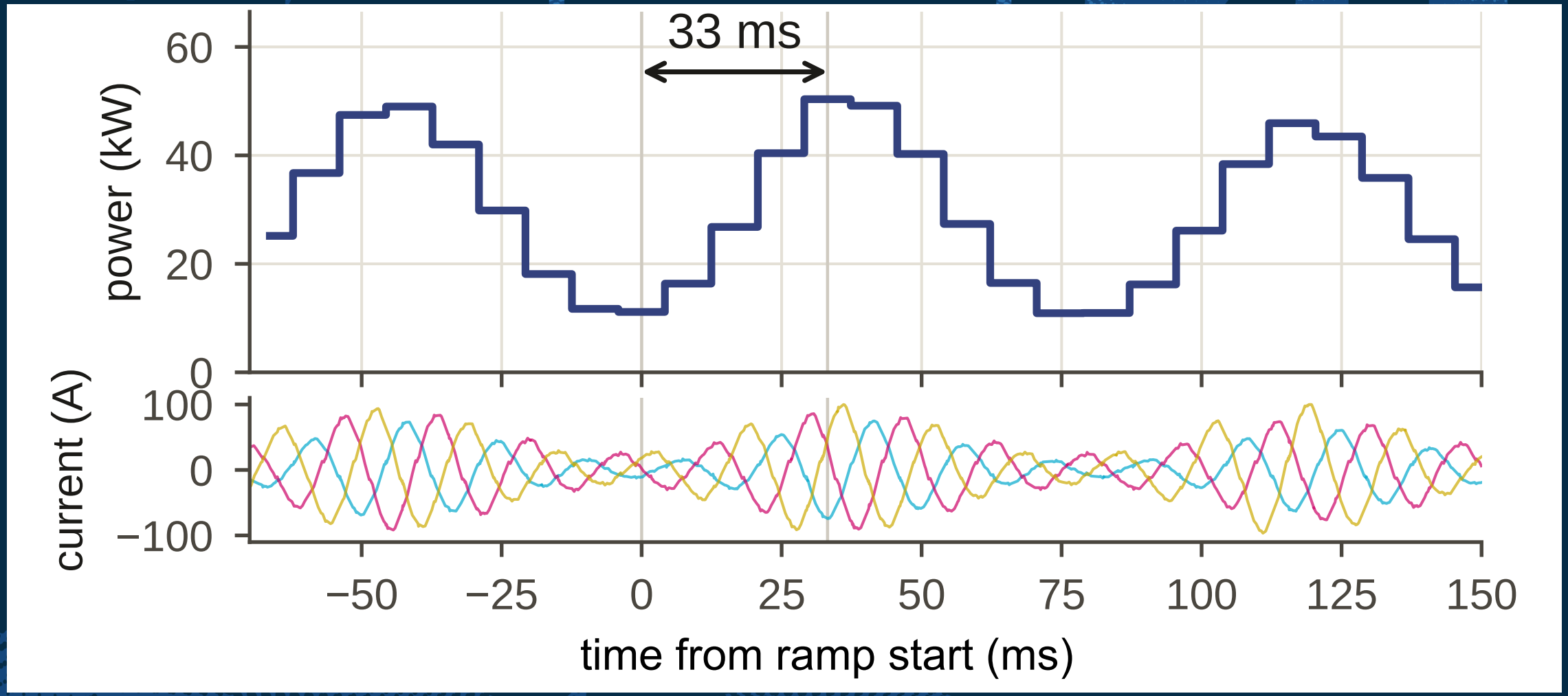
Llama 3.1 8B pretraining, 32 GPUs on one rack

Two-Rack Training



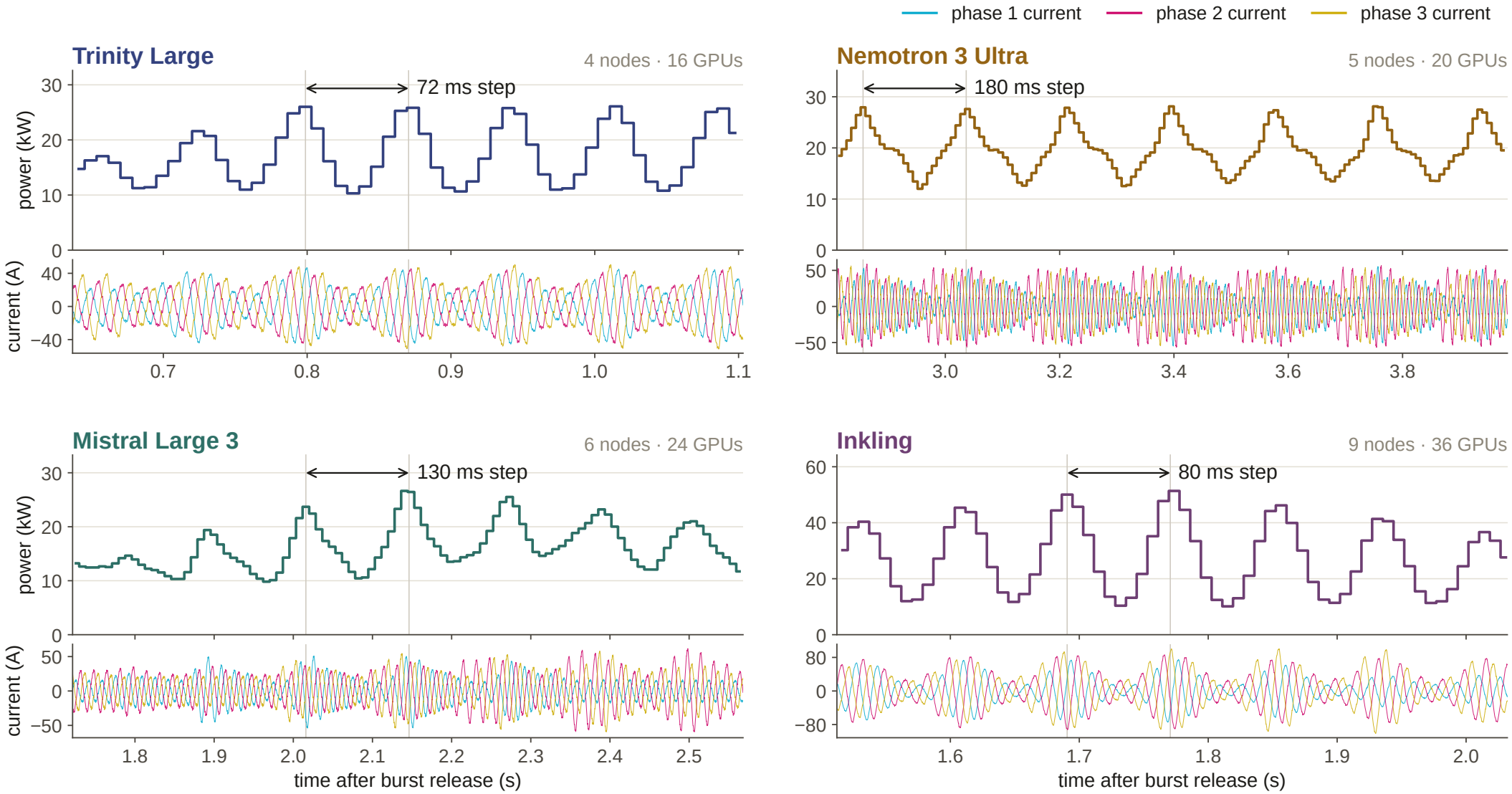
Llama 3.1 8B pretraining, 64 GPUs total, 32 GPUs on each rack

One-Rack Inference: Chunked Prefill



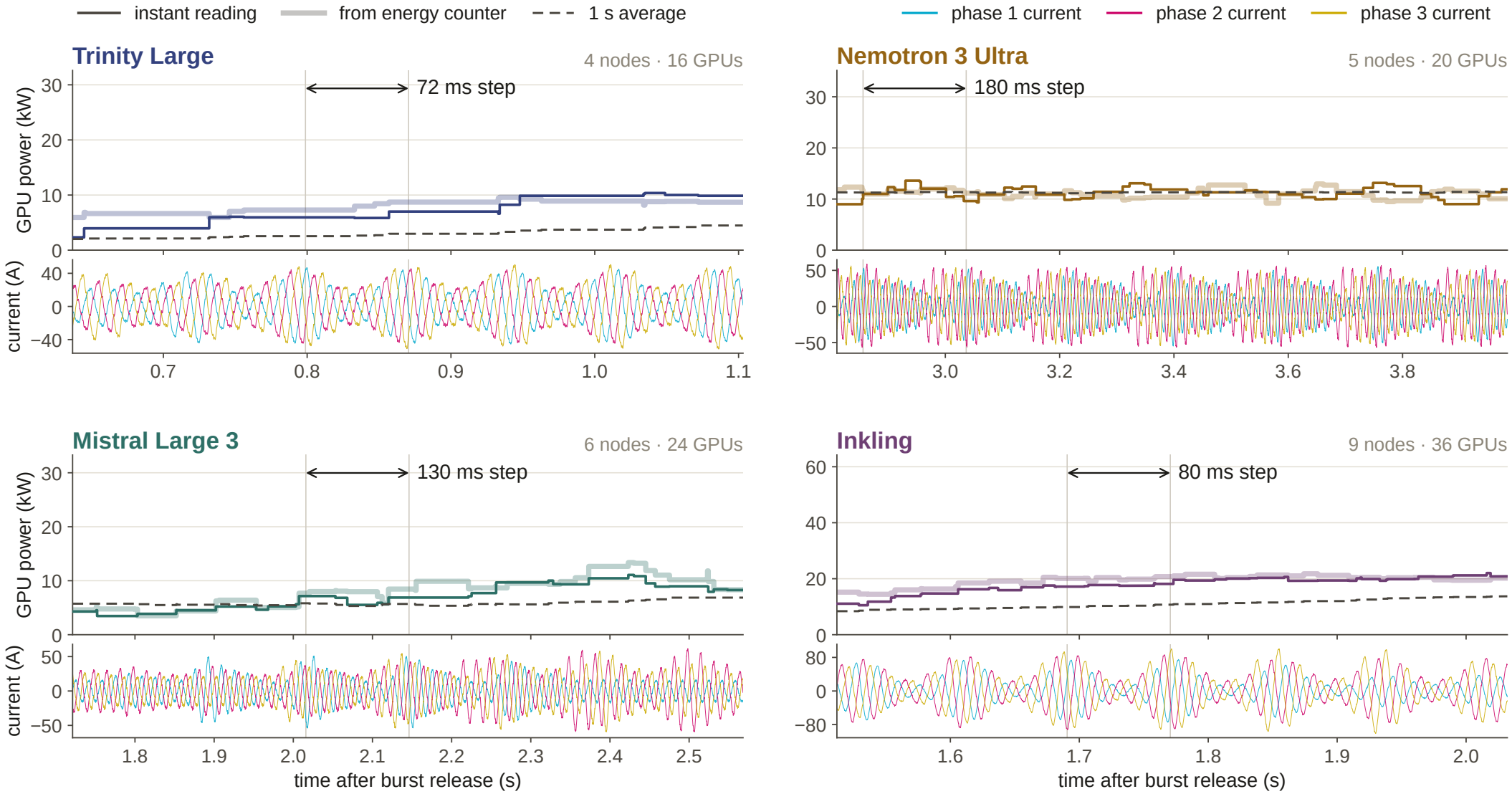
Inkling (975B MoE), 36 GPUs: each tooth is one 8192-token prefill chunk

One-Rack Inference Prefill: Different Models



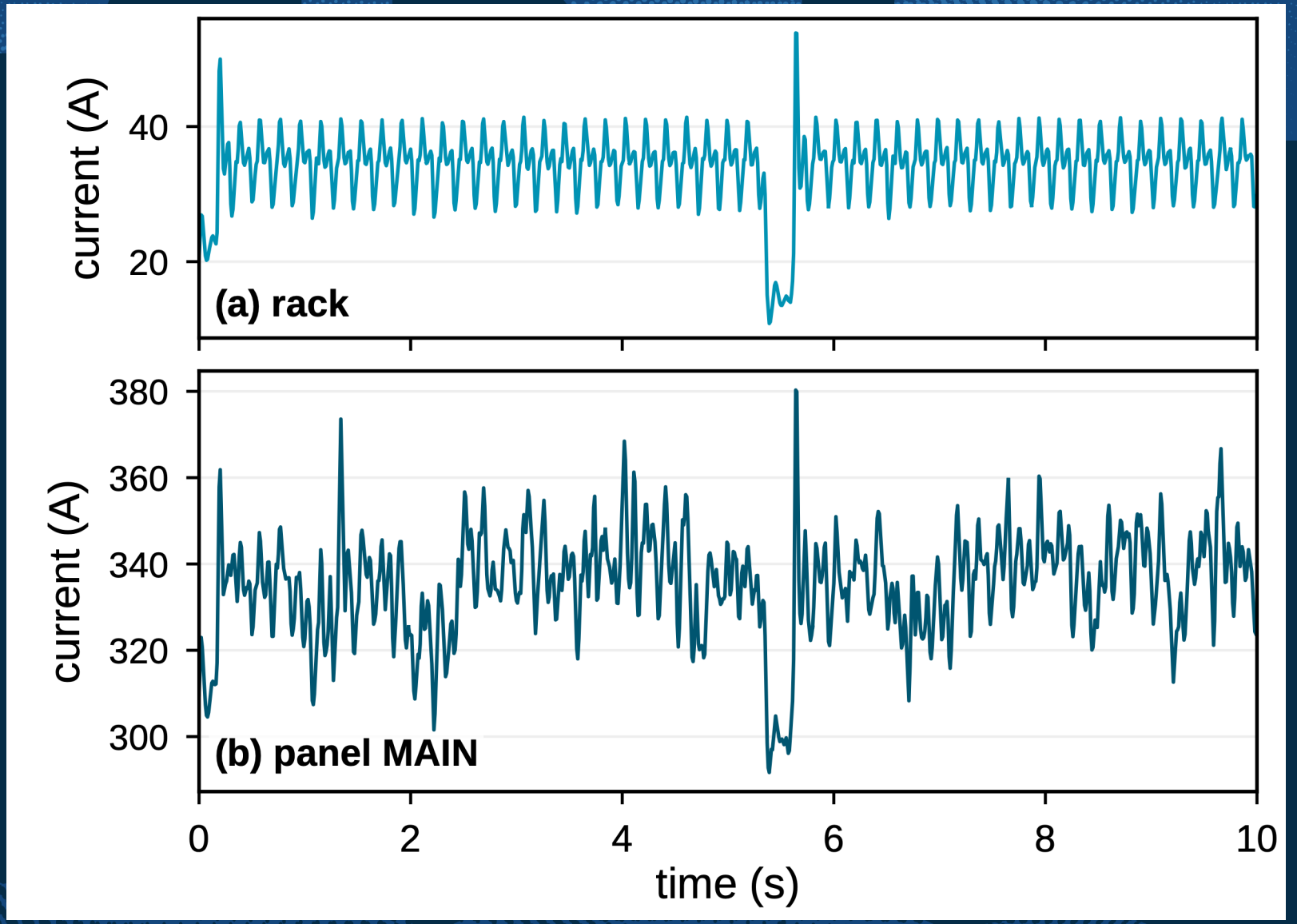
Four different models, each hit with a burst of prompts

One-Rack Inference Prefill: Different Models



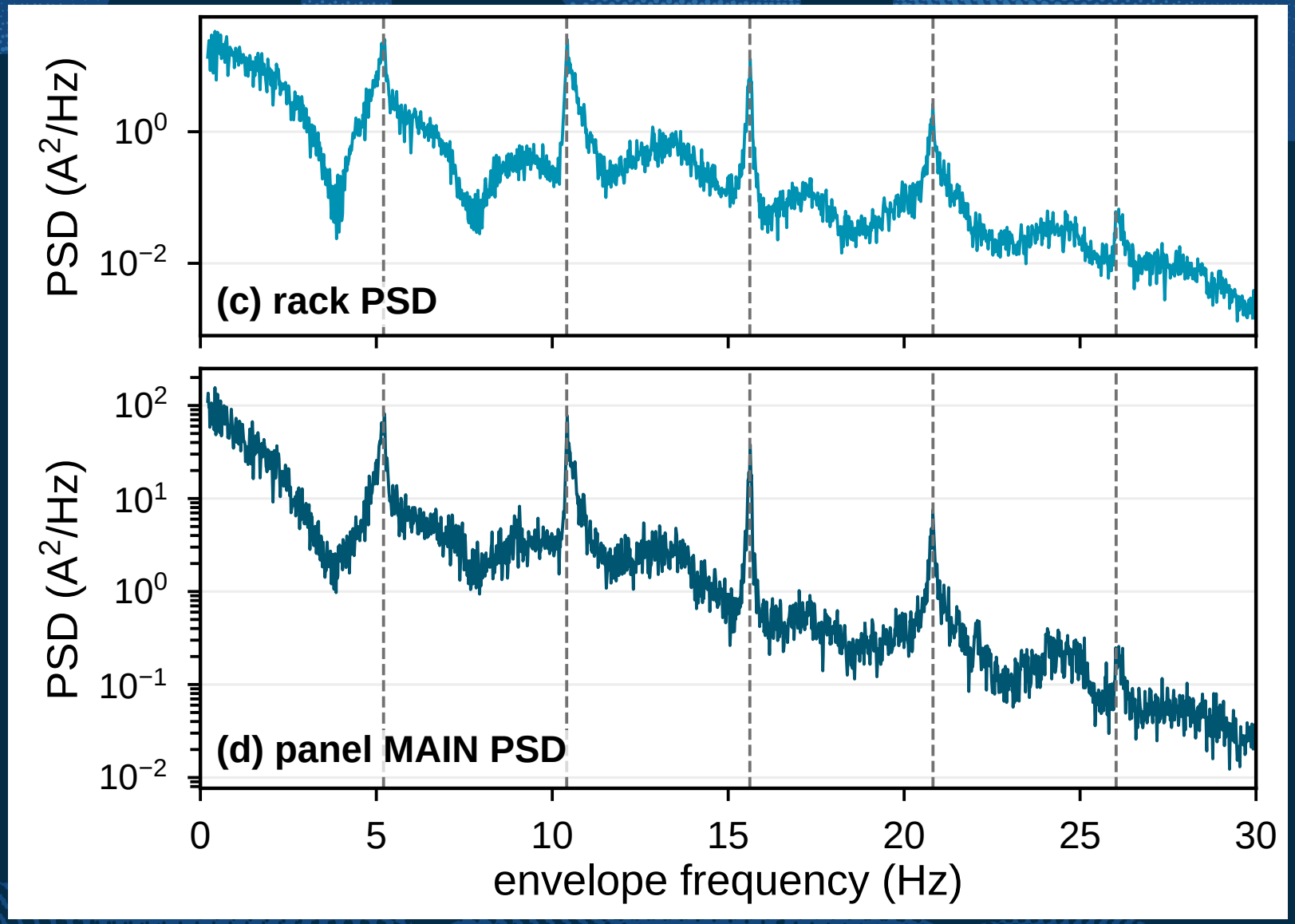
The same bursts, as the GPUs' own meters report them

Llama 3.1 8B Training (TP4/DP8): Rack vs Panel MAIN



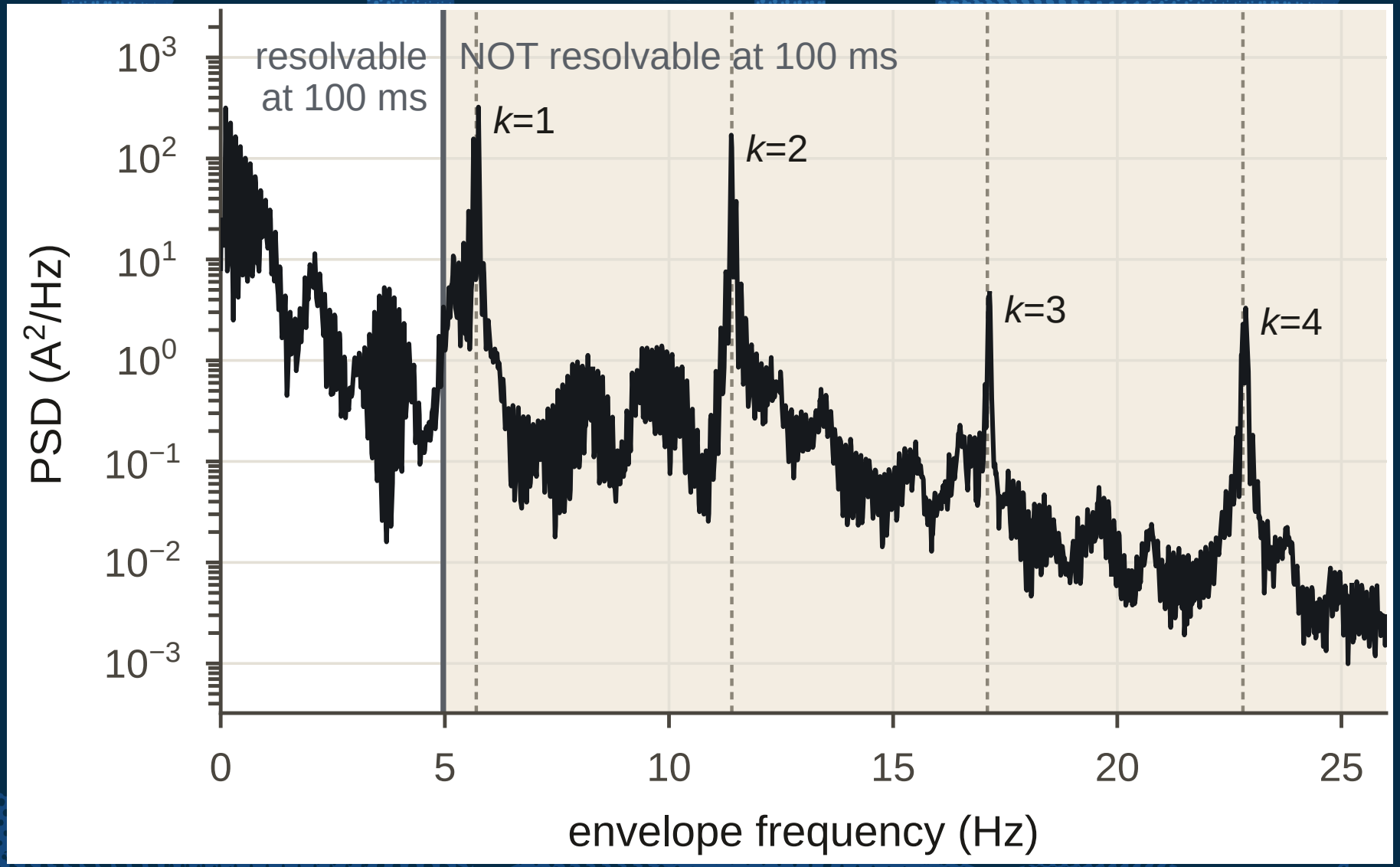
This job is ~10% of the panel MAIN's current, and its all-reduce dip still shows

Llama 3.1 8B Training (TP4/DP8): Rack vs Panel MAIN



The job's comb lines also appear in the panel MAIN's power spectral density (PSD)

Llama 3.1 8B Training (TP4/DP10)



All comb frequencies are above the Nyquist frequency of DCGM meters

What do 10 kHz meters reveal for GPU workloads?

- Power oscillates faster than 10 Hz onboard meters can resolve
- Real ramp rates are an order of magnitude faster than DCGM shows
- Racks running the same training job ramp in lockstep
- One rack’s signature is visible at the panel main
- Chunked inference prefill produces periodic power swings

OPERATIONS
ANALYSIS
DATA
TELEMETRY

TELEMETRY
DATA
ANALYSIS
OPERATIONS

Thank you! Questions?

This work was authored by the National Laboratory of the Rockies for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy National Laboratory of the Rockies. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

National Laboratory of the Rockies

Photo from Getty Images 181828180