Intelligent Power Smoothing at the Vera-Rubin Rack

Gabriele Gorla, Rouslan Dimitrov

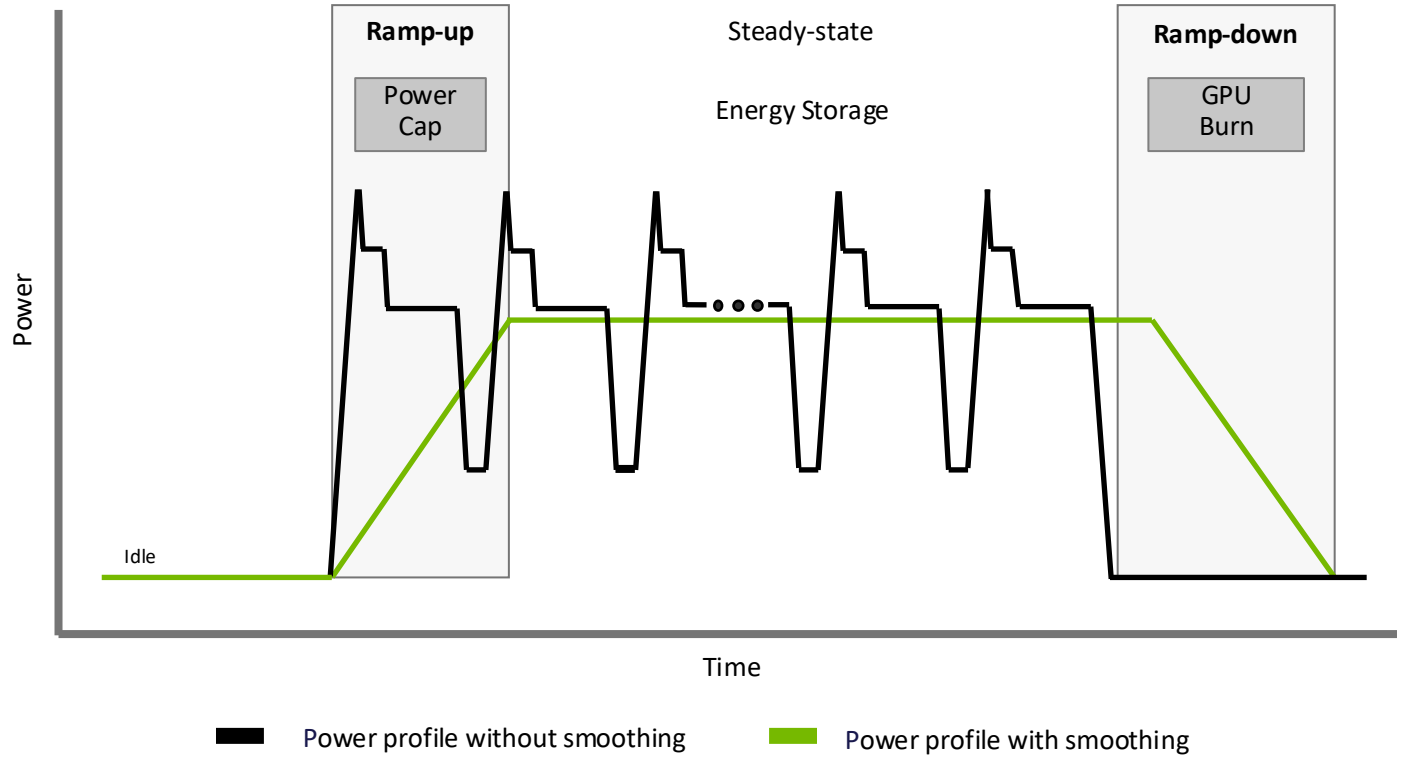
NVIDIA

Intelligent Power Smoothing Reclaims Squandered Power

Make the rack an ideal grid load:

- controlled ramp up and ramp down
- smooth steady state

Decrease the power peaks for more tokens per provisioned power

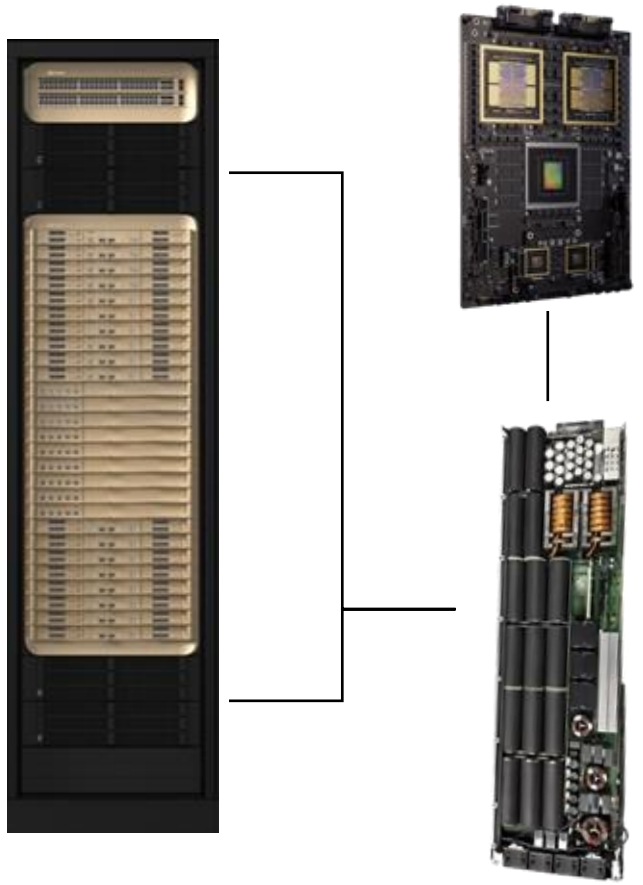


Energy Storage and GPUs Work Together

400 J / GPU of energy storage in the rack

Multiple closed-loop controllers communicate across trays

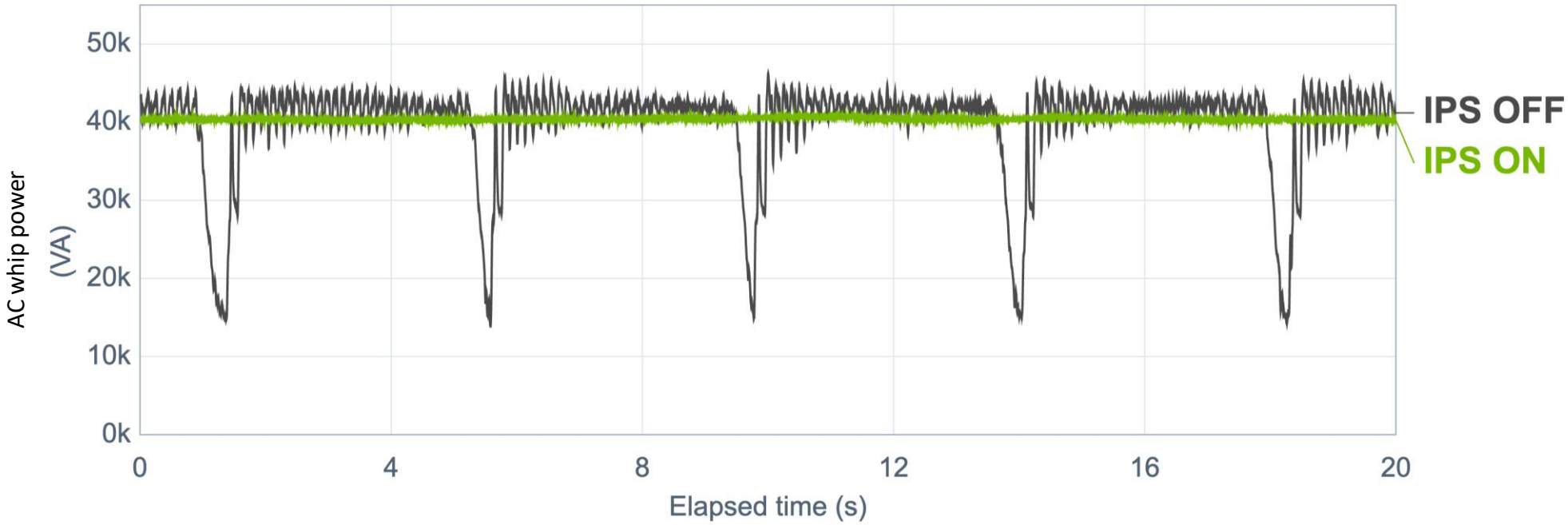
State of charge level dependent on:
- allowed AC ramp rate
- workload power profile
- GPU efficiency optimizing controller



LLM Training Result #1

LLM training, Vera-Rubin NVL72 rack

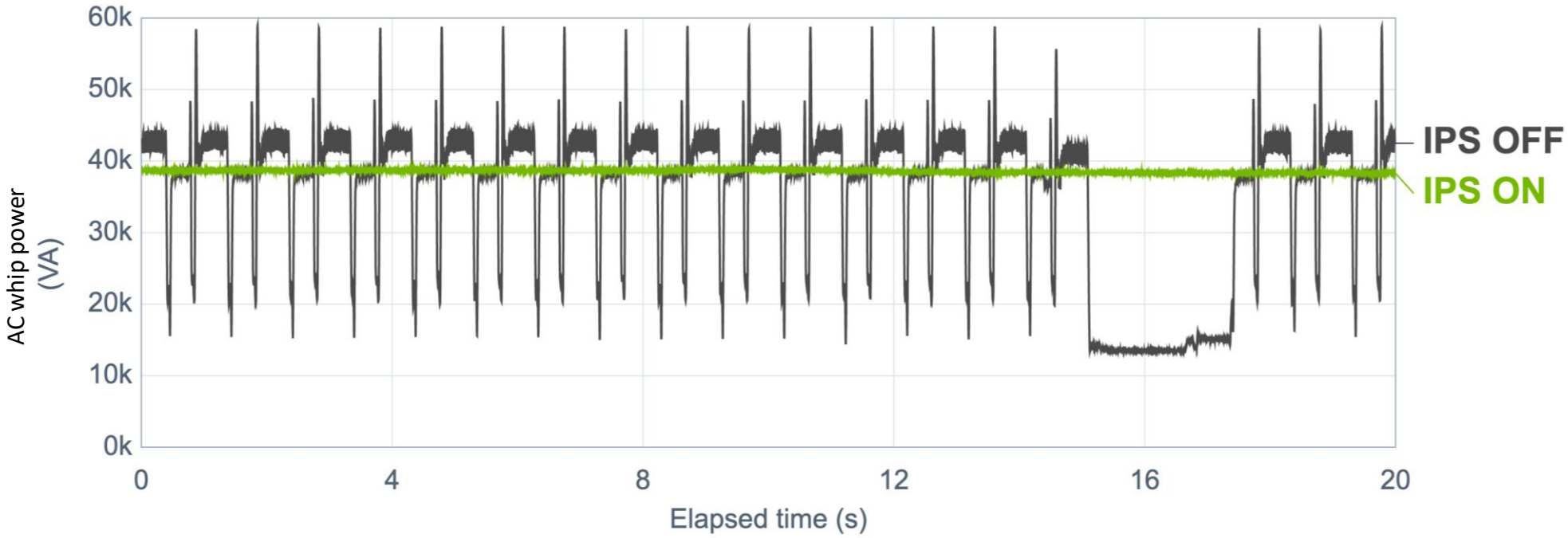
13% peak power reduction



LLM Training Result #2

LLM training, Vera-Rubin NVL72 rack

51% peak power reduction



IPS Hardware Requirements

NVIDIA Approved Vendor List (AVL) Power Shelves:

- 110 kW AC/DC Power Shelves
- 800V Power Rack + 90kW DC/DC Power Shelves

DELTA LITEON®

More vendors under final validation

Compute Trays:

- Rubin GPU equipped systems (NVL72 and future)



IPS Software Enablement

Single script configures the entire rack (power shelves, HPMs, and GPUs).

Configure — config.json	Commands

Add power-shelf PMC and compute BMC addresses.
 	
Choose the IPS settings below.
 	
"power_smoothing": {
"psu_max_ac_ramp_rate_w_per_s": 400,
"gpu_ramp_up_w_per_s": 20,
"gpu_ramp_down_w_per_s": 20,
"gpu_idle_hysteresis_seconds": 10
}
	Apply settings and verify readback:
	./ips_power_smoothing.sh apply config.json
	
	Check current settings:
	./ips_power_smoothing.sh status config.json
	
	Disable IPS:
	./ips_power_smoothing.sh disable config.json

Script and documentation at nvonline ID 1153760.

Open Discussion

