QUANTA TECHNOLOGY

ESIG Fall Workshop
October 2025

Understanding LLM–Induced Load Profiles and Complications

Amin Zamani
Sr. Director, Technology Integration

Agenda

- Introduction to Large Language Models (LLMs)
- LLM-Induced Load Characteristics
- Challenges
- Q&A

Classification of LLM Loads

In general, there are two main operation stages for LLM loads:

1 Training

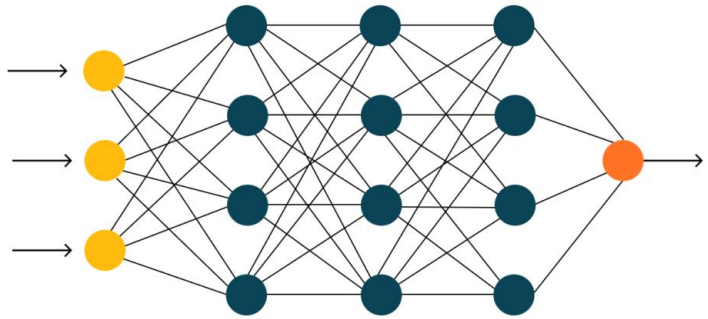
- The model trains on a vast dataset.
- It is the most power-intensive phase with sustained high GPU utilization.
- GPUs are ideal for this task because they can perform many operations simultaneously.

2 Inference

- It involves the application of trained models to new data.
  - e.g., using Chat GPT to answer a question
- It is the least power-intensive phase.
- It requires variable, user-driven power usage.
- The variability leads to short bursts of power consumption.

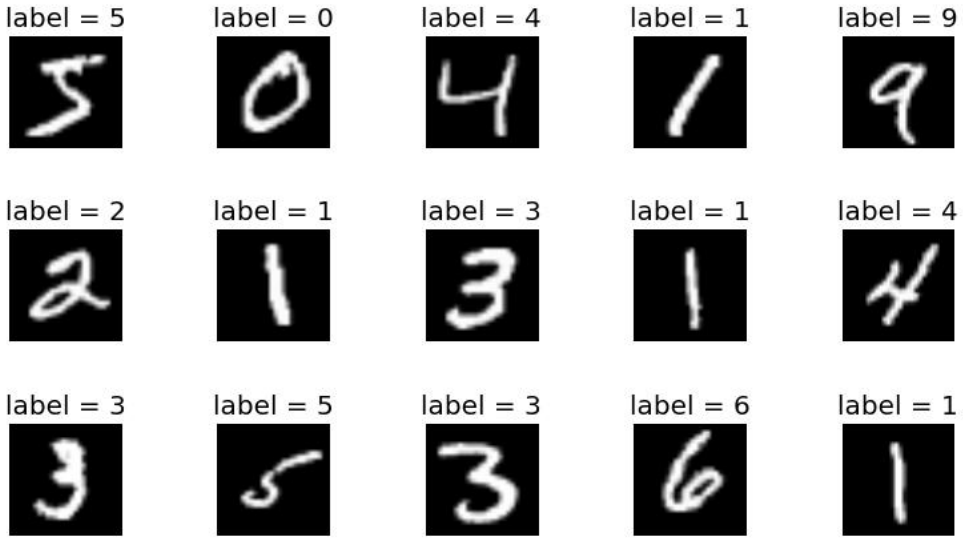
Introduction to Neural Networks

- Neural networks are the basis of modern large language models (LLMs). They consist of layers of interconnected nodes (“neurons”) that process data and learn patterns.
- Different neural network architectures are suited for specific tasks (from simple image classification to modern-day AI).



Let us review a basic example:

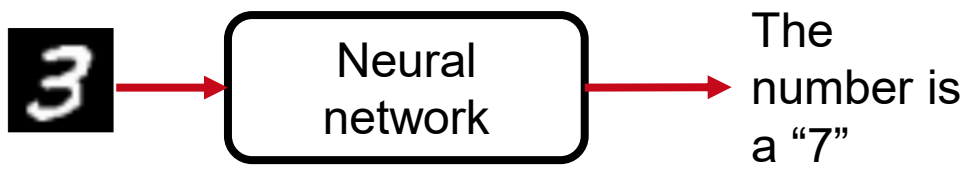
Goal: Determine what number is written in each handwritten image



Introduction to Training

A neural network learns by repeatedly receiving inputs, recognizing patterns within the data, and gradually improving its accuracy over time. There are two main phases in training:

Forward pass

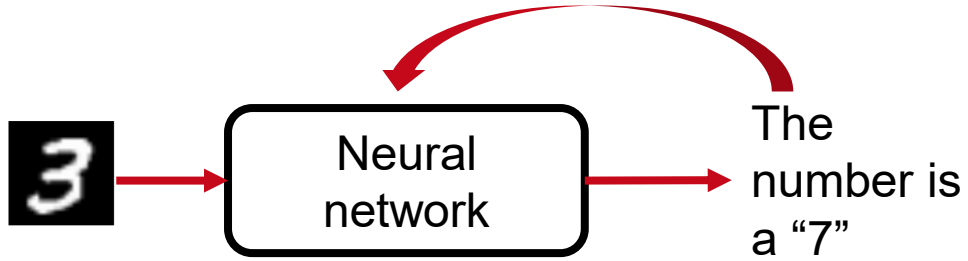


First, each input passes through the network in a **forward pass**, producing a classification.

Then, the difference between this prediction and the correct label is used to compute an error.

Backward pass

Prediction was wrong, adjust weights to learn from the mistake.



Using this error, the network goes into a **backward pass** adjusting its internal parameters.

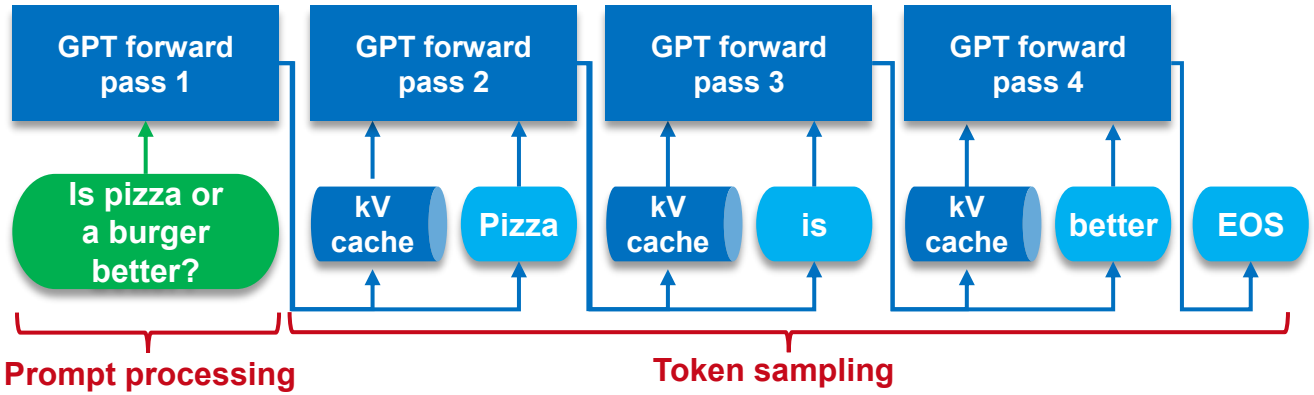
Repeating this process over many inputs helps the network learn patterns and improve accuracy.

Conventional Transformer-based LLMs – Inference Process

In general, when a request is made to an LLM (ChatGPT, Copilot, etc.), the model will go through two phases:

1. **Prompt phase:** At the start of each request, the system enters a compute-intensive phase where the prompt is processed.
2. **Token phase:** Once the prompt phase is processed, the system shifts into a more stable, lower-power state where it performs sequential calculations.

In data centers, inference requests are stochastic. This means it is unlikely that multiple GPU racks serving inference will start or stop heavy compute at the same time.



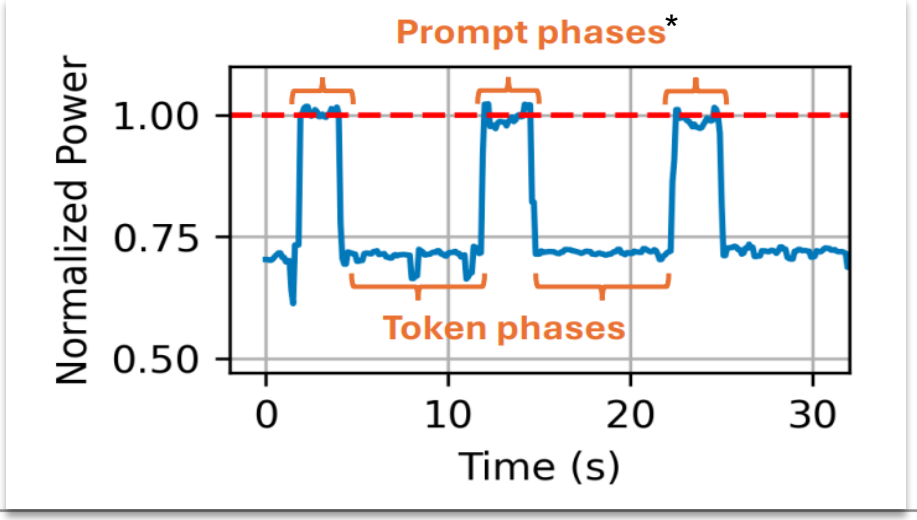


*Image taken from P. Patel et al. "Characterizing Power Management Opportunities for LLMs in the Cloud", Association for Computing Machinery, April 2024.

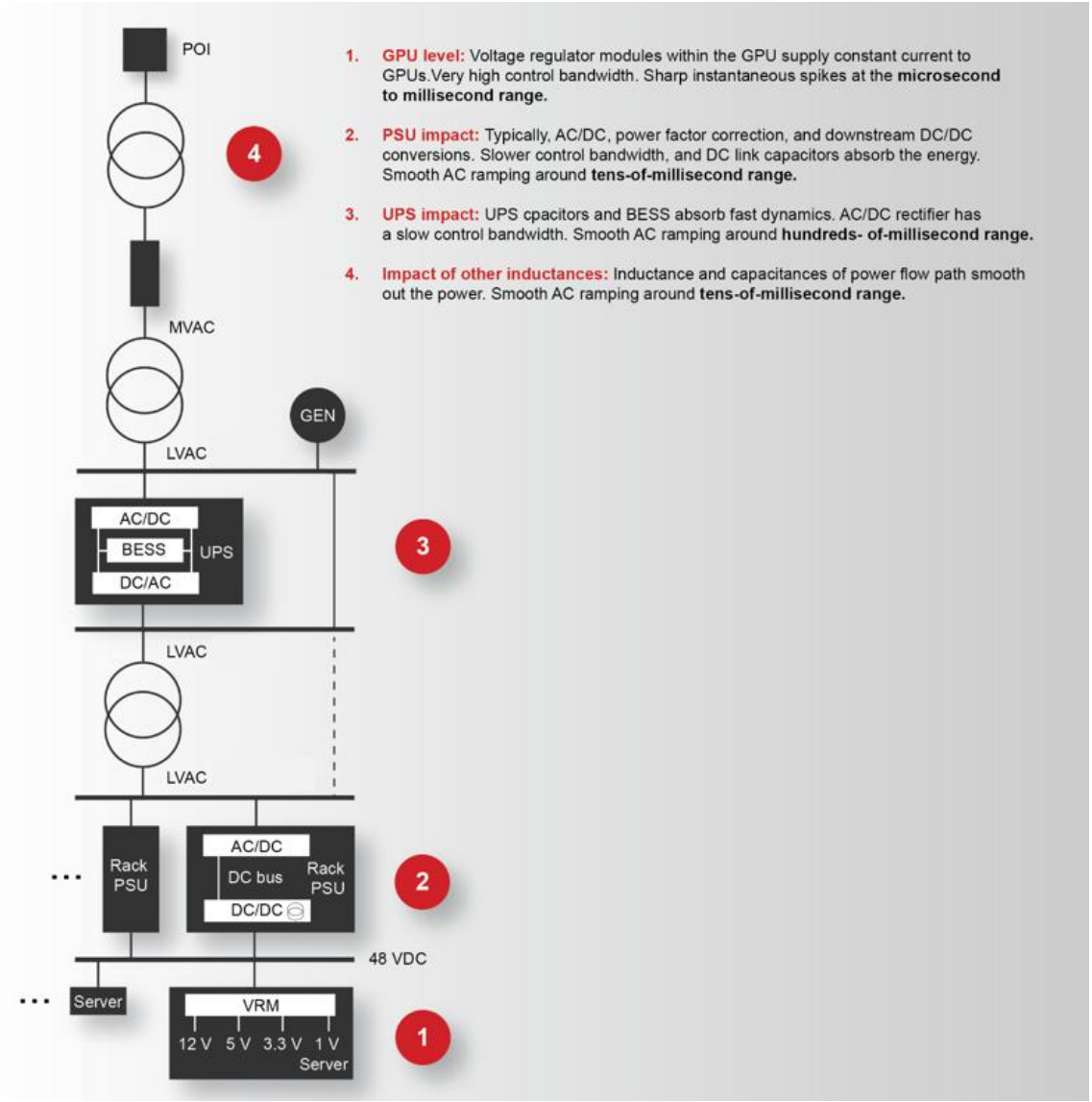
Power Transients During LLM Training and Inference

Causes of Electric Power Fluctuations

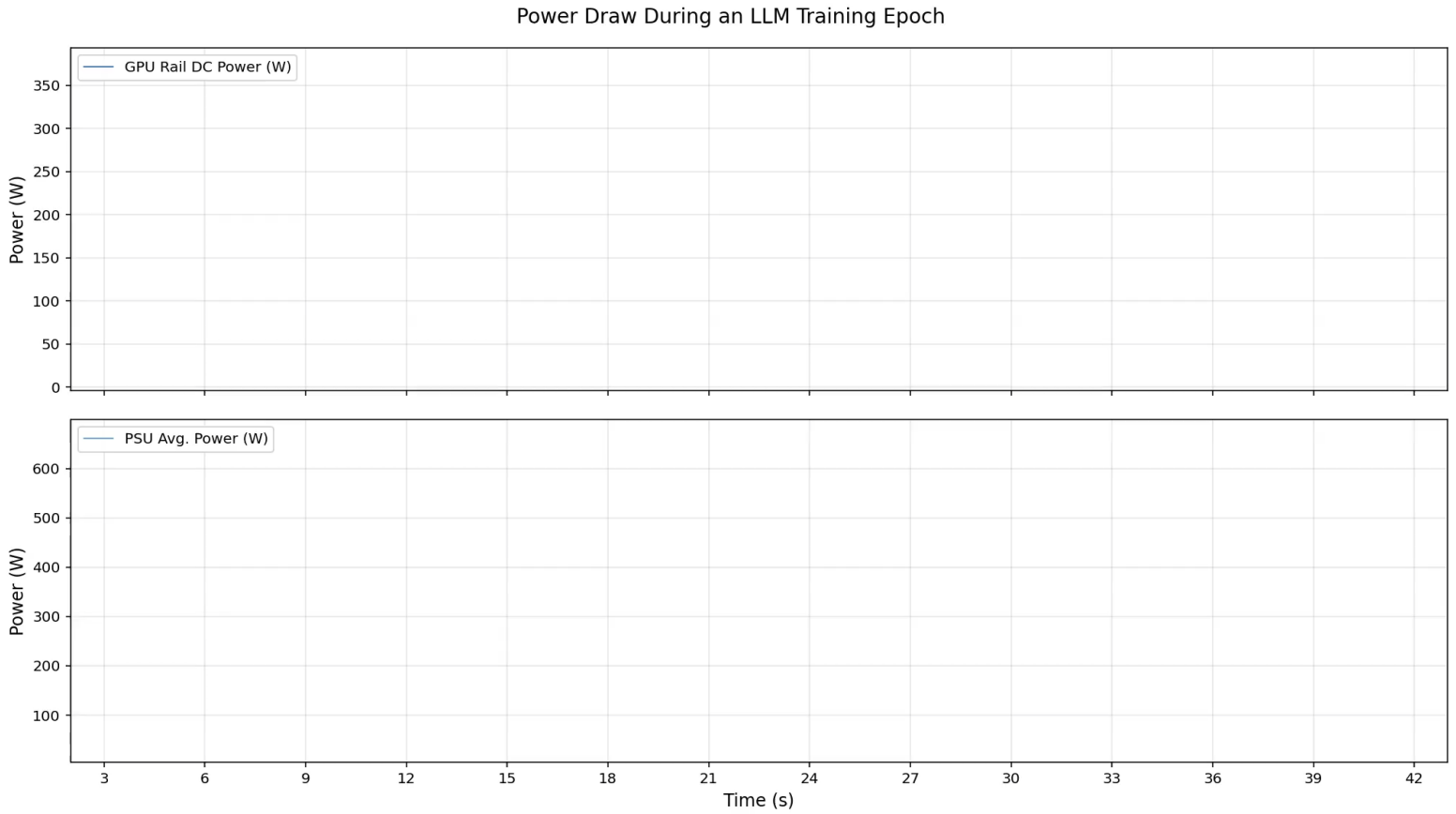
Training	**Model initialization**	Model configuration and instantiation can trigger an initial power surge, which increases as memory is allocated.
Training	**Evaluation intervals**	During training, model evaluation and checkpointing decrease GPU utilization to near idle, causing power dips.
Training	**End of iterations**	During training, GPUs synchronize before the next iteration, causing a large power dip.
Training	**Forward and backward pass**	Each training iteration includes forward and backward passes, with thread synchronization in between, reducing GPU utilization and causing a power dip.
Inference	**Prompt vs. token sampling**	Inference requests trigger power spikes at state due to prompt-related GPU computation.

Impact of PSU Filtering

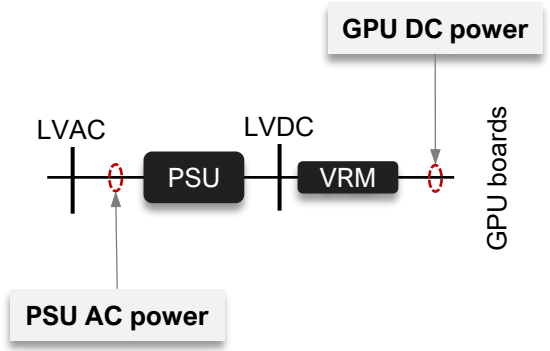
1) Voltage Regulator Modules supply stable current to GPUs. Their high control bandwidth causes sharp instantaneous spikes.
2) Slower control bandwidth of PSUs, along with the DC link capacitors, smooth AC ramping around tens-of-millisecond range.
3) UPS capacitors and BESS absorb fast dynamics. The slow control bandwidth of AC/DC rectifiers smooth AC ramping around hundreds-of-millisecond range.
4) Inductances help with load smoothing.



LLM Load Examples – Training Epoch



- Data is collected from an NVIDIA RTX 4090 GPU, 1000 W MPG PCIe 5.0 Gold (80+ Gold) power supply unit[1].
- Training performed on LLaMa-3.1 (8B) local Meta model.



1. Data obtained from Y. Li, Y. Li , “AI Load Dynamics – A Power Electronics Perspective.” Jan. 2025.

LLM Load Examples – Training

Transient shape and behavior are dictated by hardware and LLM architecture**.** More powerful AI-accelerators and State-of-the-Art models (ChatGPT, Claude, Gemini) may result in more aggressive transients, longer sustained periods of utilization.

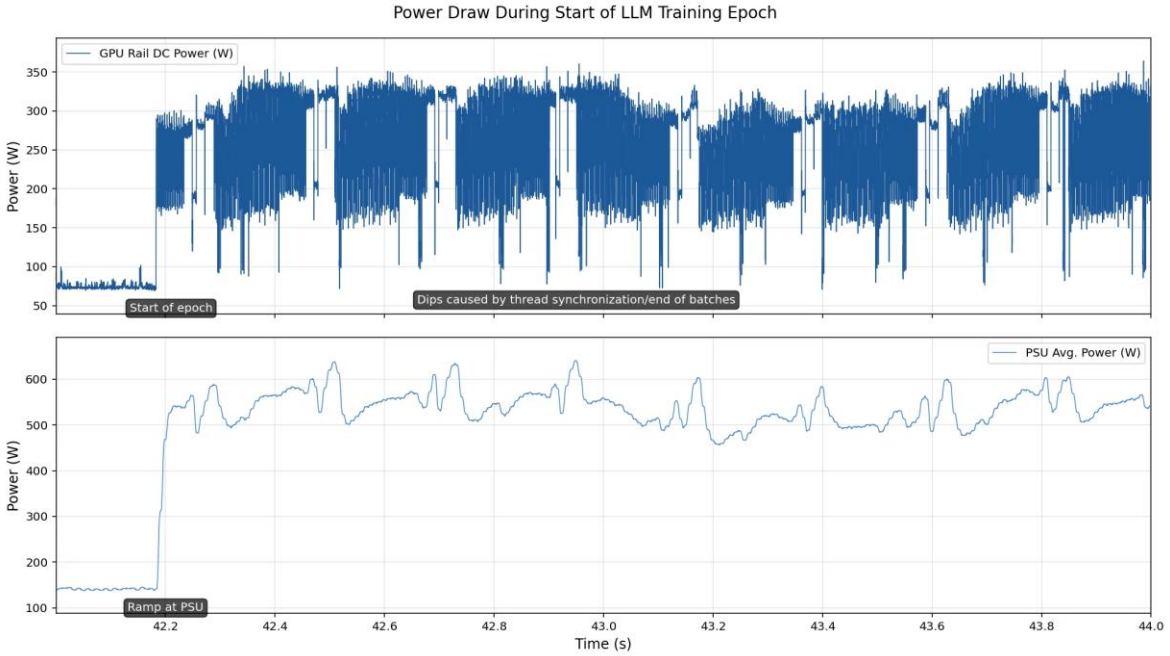
GPU (DC-side active power profile):

- Maximum ramp rate: **309 kW/s**
- Maximum de-ramp rate: **301 kW/s**

PSU (AC-side active power profile):

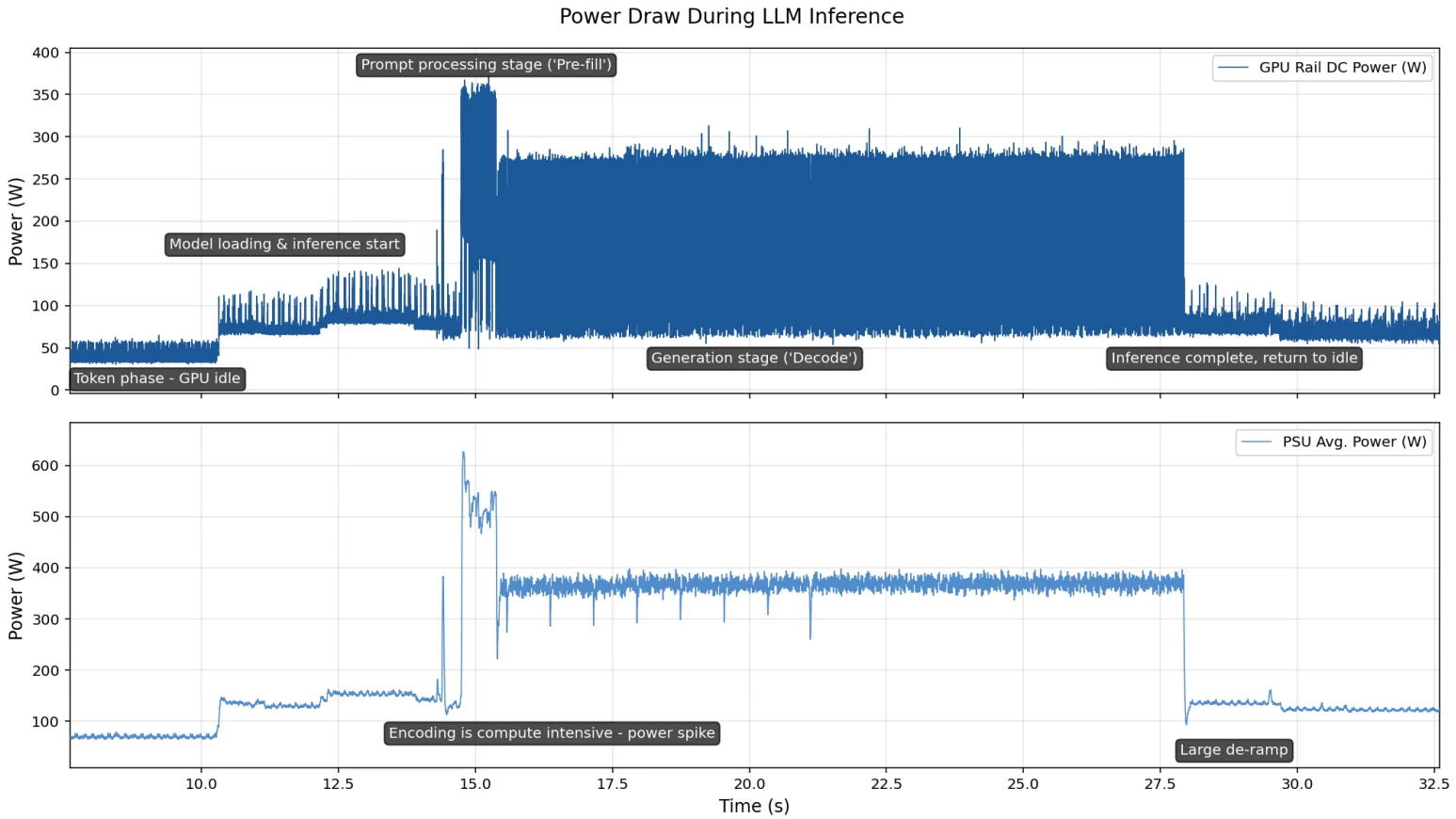
- Maximum ramp rate**: 61 kW/s**
- Maximum de-ramp rate: **58 kW/s**

Note that PSU sustained power is higher due to supplying the CPU as well.



Data obtained from Y. Li, Y. Li , “AI Load Dynamics – A Power Electronics Perspective.” Jan. 2025

LLM Load Examples – Inference



Inference on **Gemma (27B)** Google local LLM model.

GPU:

- Maximum ramp rate: **407 kW/s**
- Maximum de-ramp rate: **377 kW/s**

Observed during 'prompt processing' transients.

PSU:

- Maximum ramp rate**: 72 kW/s**
- Maximum de-ramp rate: **63 kW/s**

Observed during prompt processing.

Data obtained from Y. Li, Y. Li , "AI Load Dynamics – A Power Electronics Perspective." Jan. 2025

AI Load Challenges

The AI load can also introduce several challenges to the electric grid:

Load-side impacts:

- Low-voltage generator oscillation
- Reduced lifetime of UPS and battery backup units
- Power quality challenges (harmonics & flicker)
- Strain on distribution infrastructure and sensitive IT equipment
- Higher cooling demand
- Reduced energy efficiency (PUE)

Grid-side impacts:

- Voltage flicker and TOV
- Frequency variation and reduced damping of interarea oscillations
- Sub-synchronous oscillations, affecting synchronous generators (SSR) and inverter-based resources (SSTI/SSCI)
- Ride-through limitation
- Harmonics from HF switching
- Thermal stress on primary power equipment

Conclusion

AI-driven workloads introduce rapid power swings and oscillations, impacting both data centers and the grid.

Collaboration among utilities, system operators, facility owners/developers, technology providers, and experienced partners is critical to fully understand facility design and develop accurate load models for power system studies and for mitigation strategies.

AI load characterization and modeling (transients, harmonics, ride-through, etc.) are essential to evaluate grid interactions and design commercially and technically viable solutions.

Technology assessment through transient studies, lab testing, and HIL validation ensures robust performance under real-world conditions and compliance requirements.

Achieved by implementing tested, validated, and collaborative solutions, tailored to site-specific operating profiles and integrated across both BTM and FTM.

Questions and Discussion