Analyzing and Modeling GPU Power Behavior in AI Workloads for Energy-Efficient Data Centers

Nagham Mousa, Computer Systems Engineering
Mentor: Meng Wu, Assistant Professor
School of Electrical, Computer and Energy Engineering

Research Question

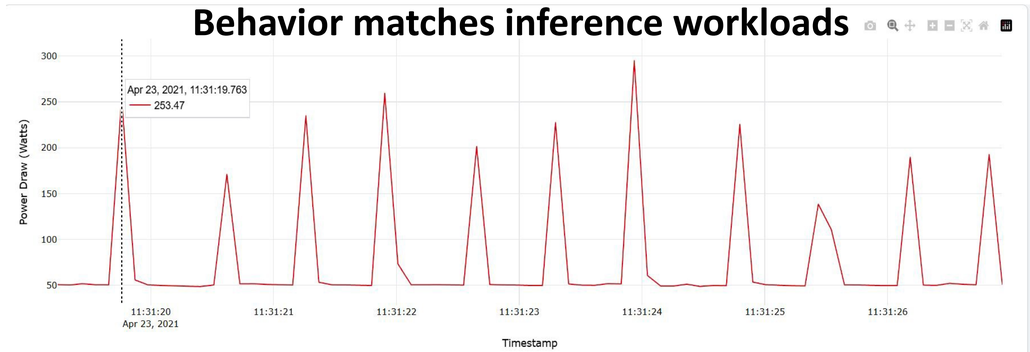
- GPU power is highly variable and transient
- Traditional averages hide important spikes
- Understanding power behavior is important:
  - Energy efficiency
  - System stability

How can we model GPU power behavior to better understand AI workloads and improve data center efficiency?

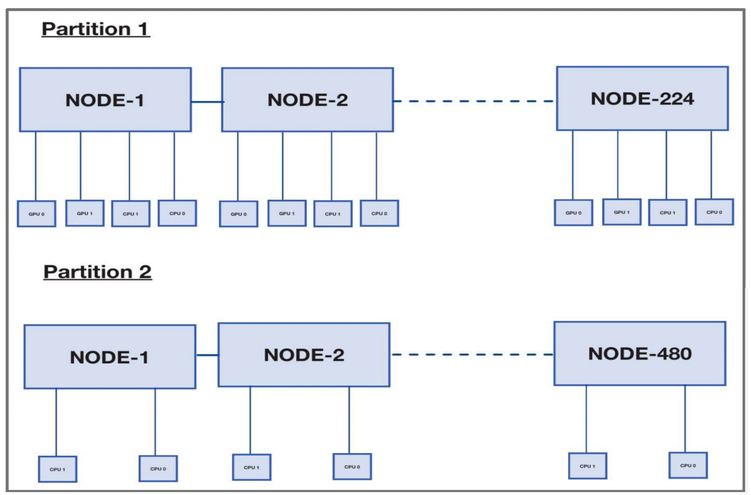
Methods

- Process large-scale GPU data using Python
- Extract features:
  - Duty cycle
  - Peak-to-idle ratio
  - Variability
- Compare power behavior patterns
- Apply frequency analysis (FFT)

Case Study Analysis



Dataset



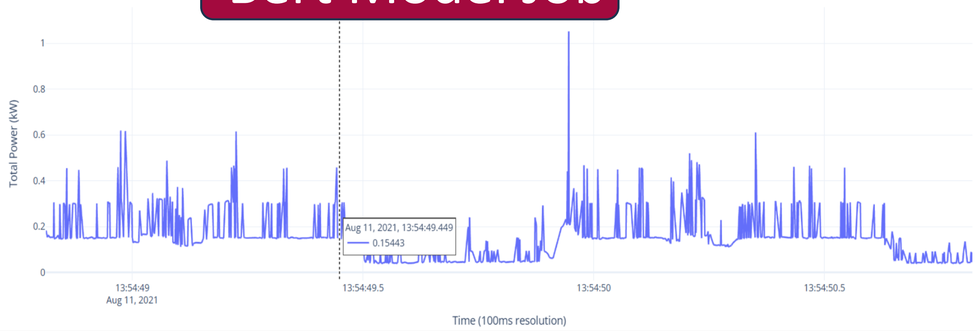
- MIT Supercloud GPU dataset (~2TB)
- High resolution (~100 ms)
- Most jobs use only 1 GPU (~85%)

Workload Types

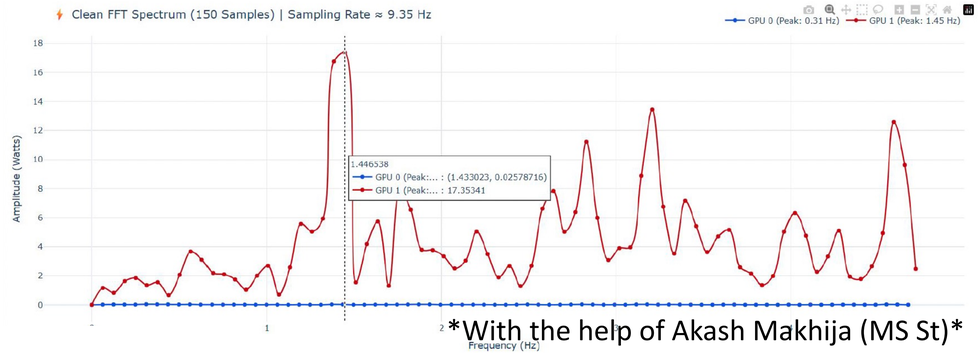
Training → steady high power (70-80%)
Fine-tuning → moderate fluctuations (30-70%)
Inference → short, bursty spikes (<30%)

$$\text{Duty Cycle (\%)} = \left(\frac{\text{Number of Power Samples} > (\text{Mean} + \text{Standard Deviation})}{\text{Total Number of Samples}} \right) \times 100$$

Bert Model Job



Frequency Analysis (FFT)



With the help of Akash Makhija (MS St)

Progress & Future Work

Currently analyzing individual jobs
Expanding to BERT and other models
Improving workload classification
Exploring ML-based prediction
Goal: optimize energy usage in data centers

FURI

Ira A. Fulton Schools of Engineering
Arizona State University