PHOTO: DARCY PETERKA

Characterizing Training Performance and Energy for Foundation Models and Image Classifiers on Multi-Instance GPUs

Connor Espenshade, Rachel Peng, Eumin Hong (Columbia University)
Max Calman, Yue Zhu, Pritish Prada, Eun Kyung Lee (IBM Research)
Martha A. Kim (Columbia University)

Scaling Up & Scaling Down

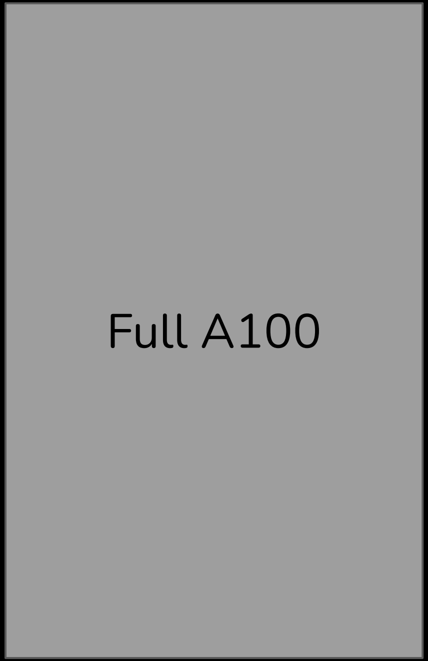
Large Workloads Require Scaling Up: Distributed Training

Distributing one job across many GPUs, pooling resources
GPT-2 pre-training requires minimum 8 A100s
GPT-3 training in 11 minutes with 3584 H100s

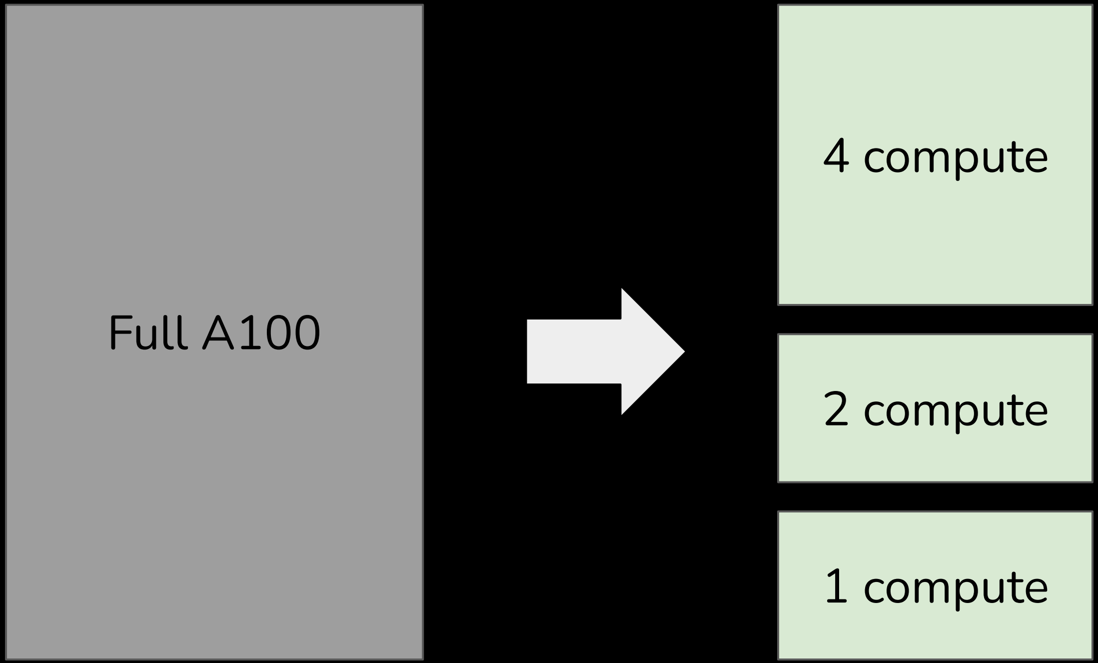
Can we Scale Down Smaller Workloads?

Older models
Smaller parameter counts
Image classifiers
Fine-tuning
Inference

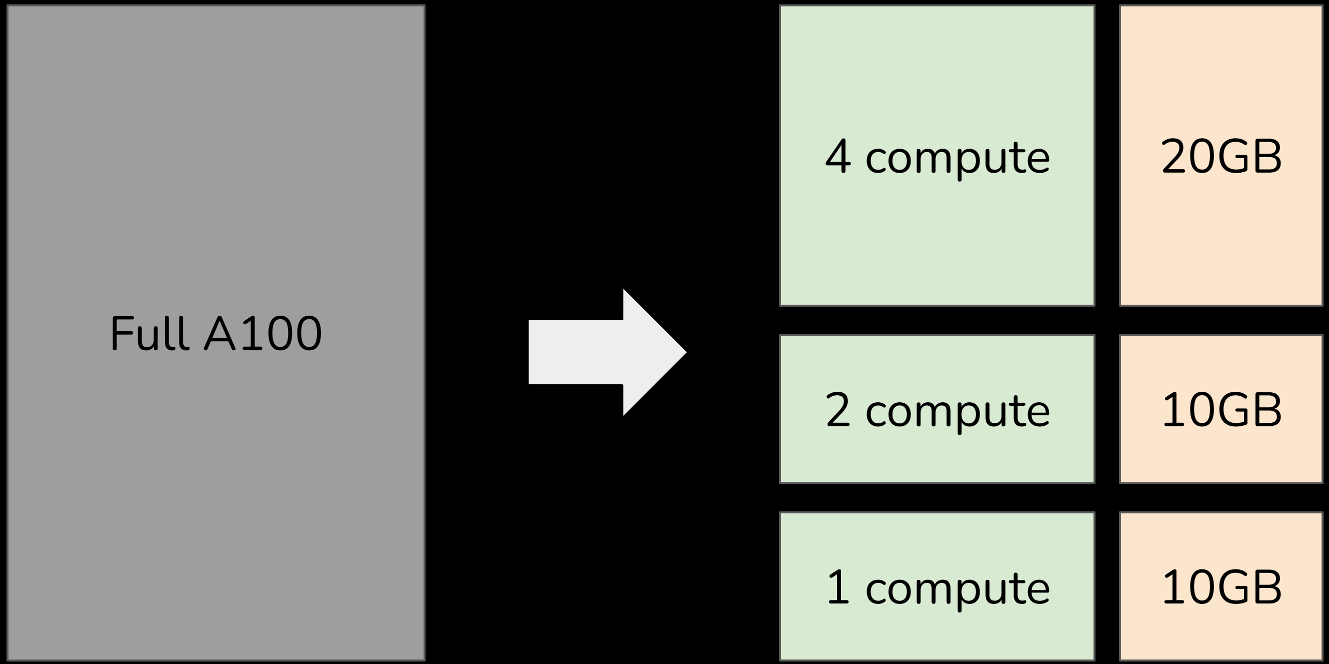
Background on Multi-Instance GPUs



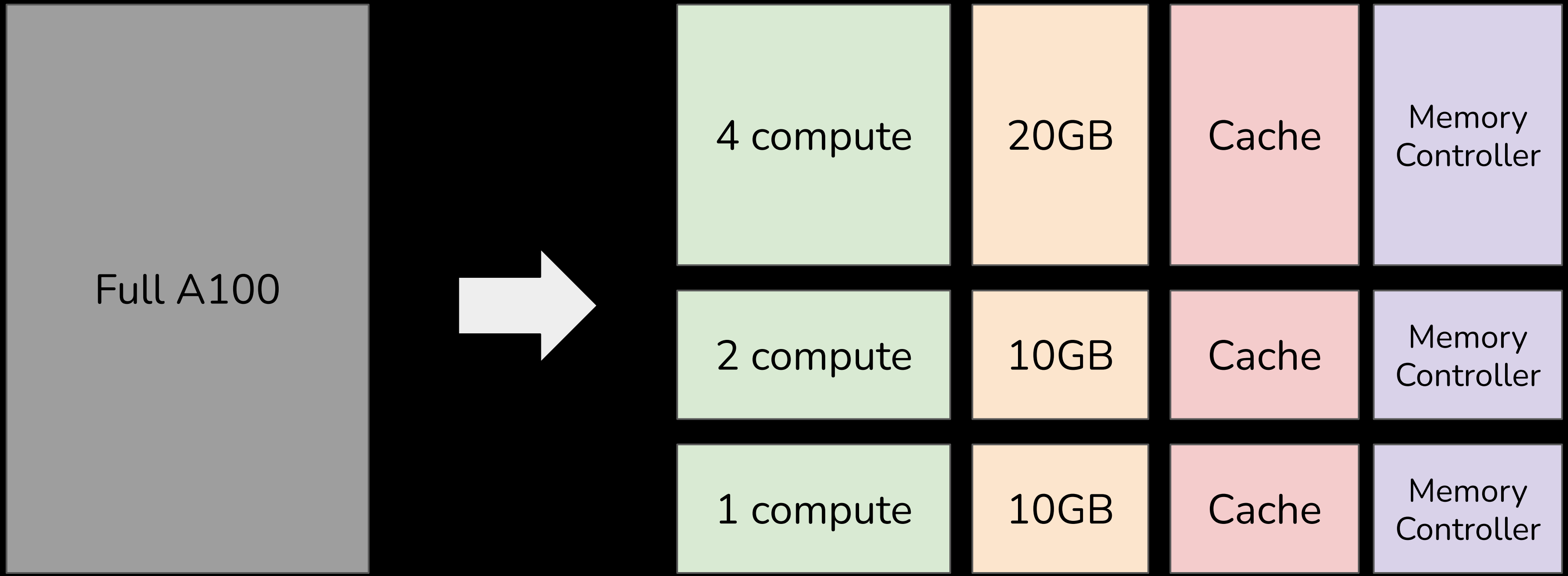
Background on Multi-Instance GPUs



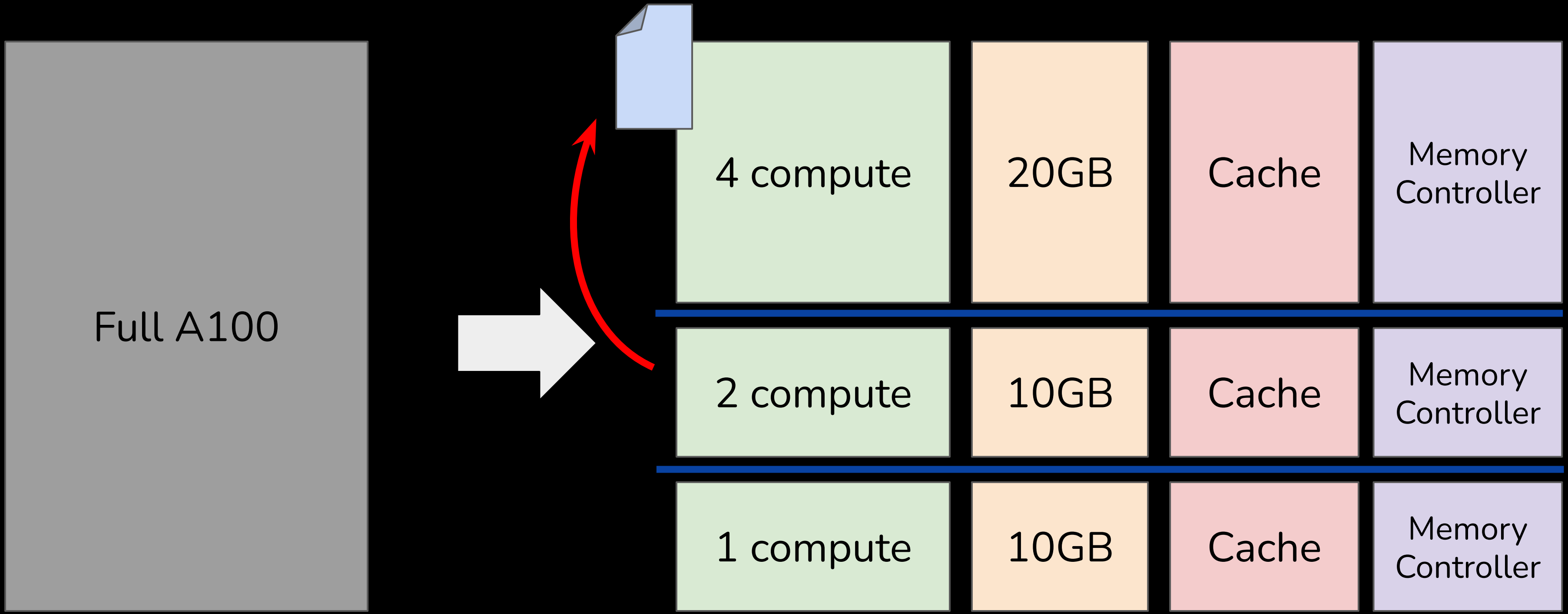
Background on Multi-Instance GPUs



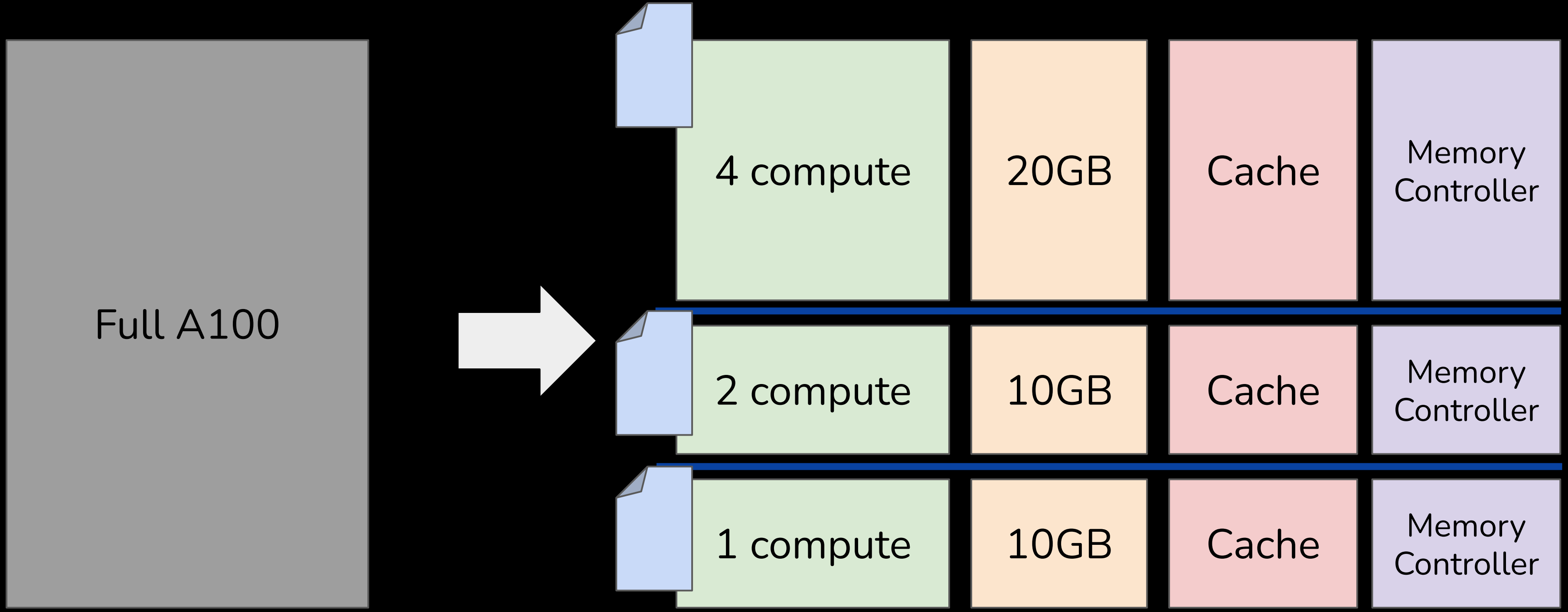
Background on Multi-Instance GPUs



Background on Multi-Instance GPUs

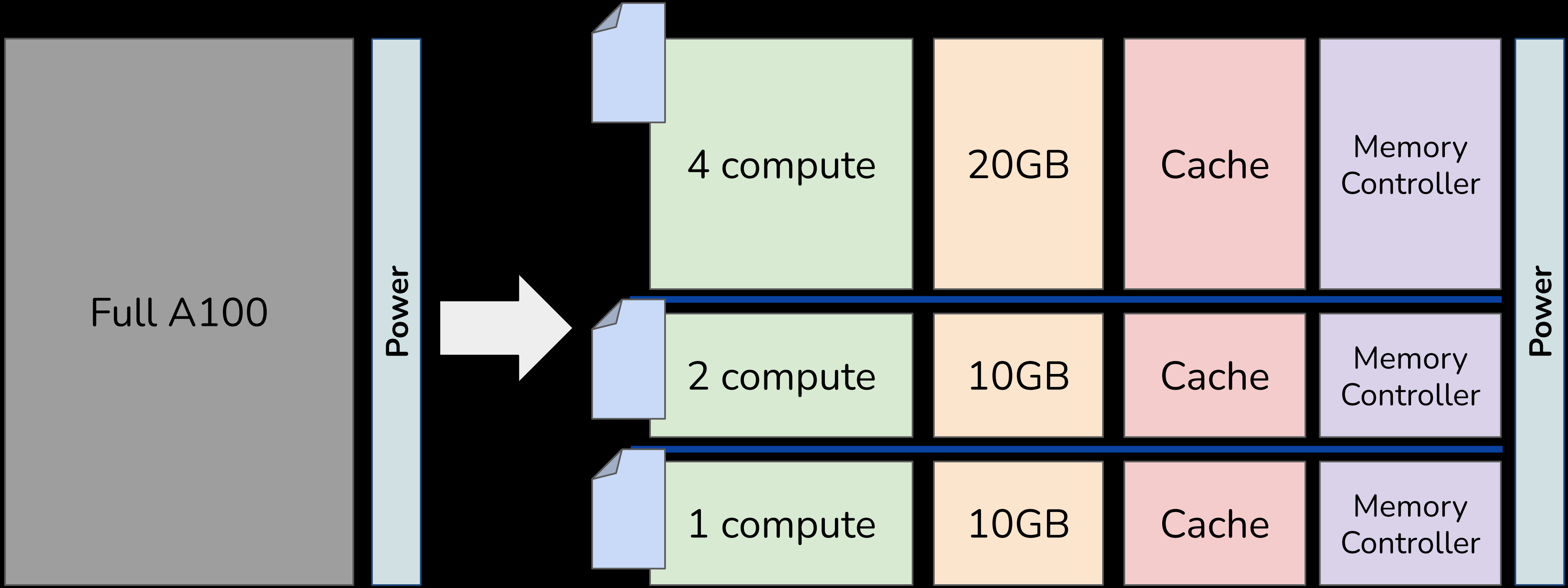


Background on Multi-Instance GPUs



Multiplexed jobs running in parallel, inaccessible from other slices, with resources self-contained

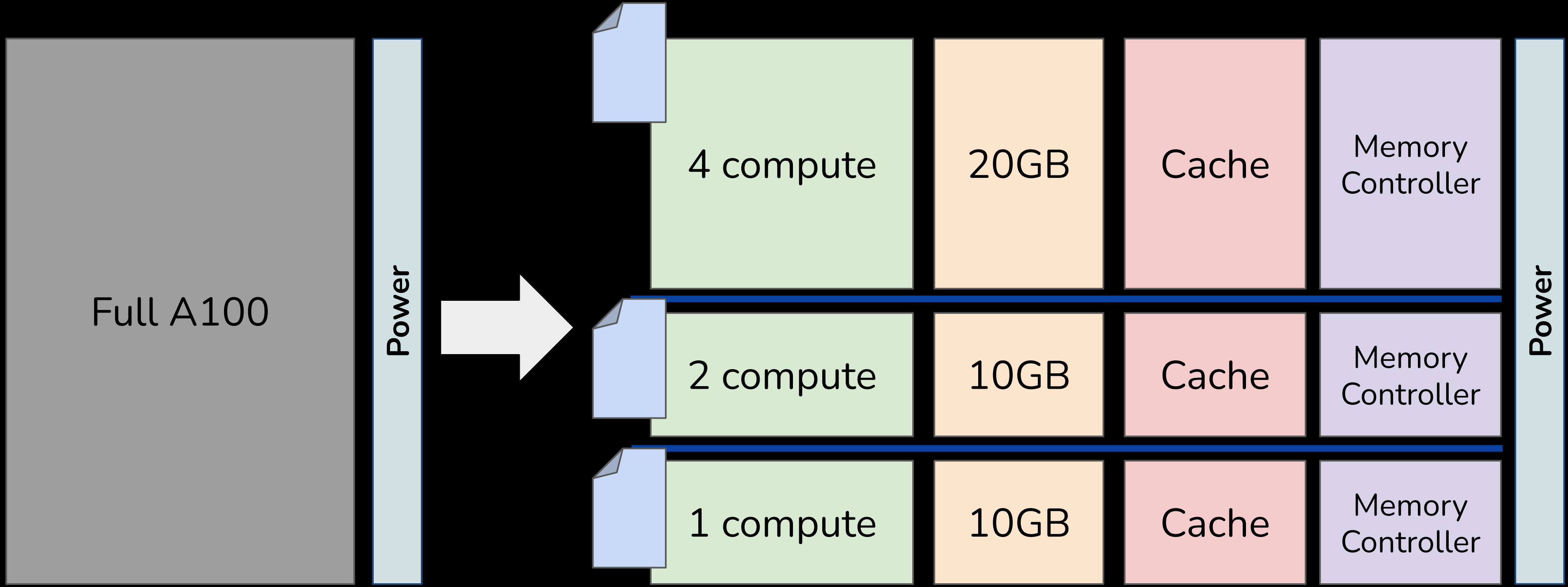
Background on Multi-Instance GPUs



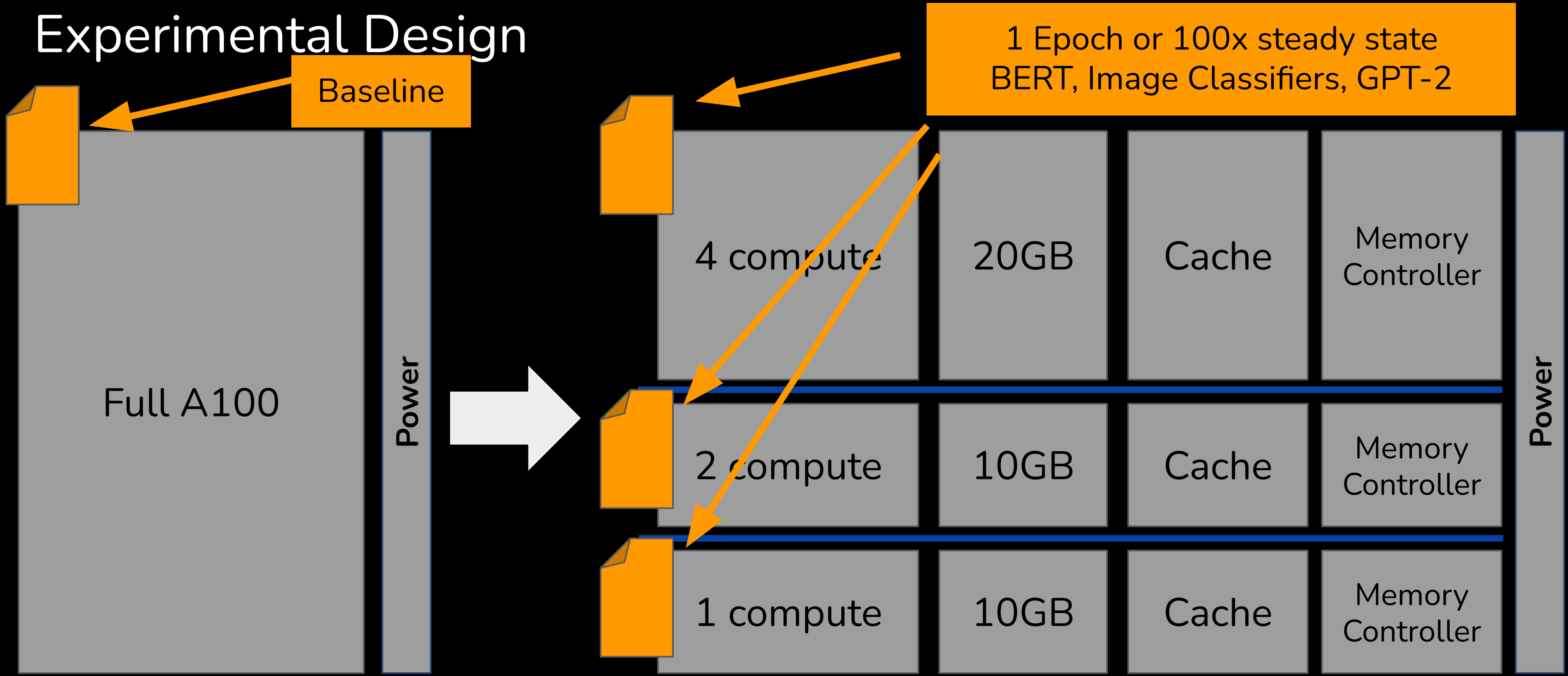
Power capping throttles GPU by capping power available for all slices

Goal: Understand how multiplexing concurrent workloads on MIG change performance and energy with various workloads

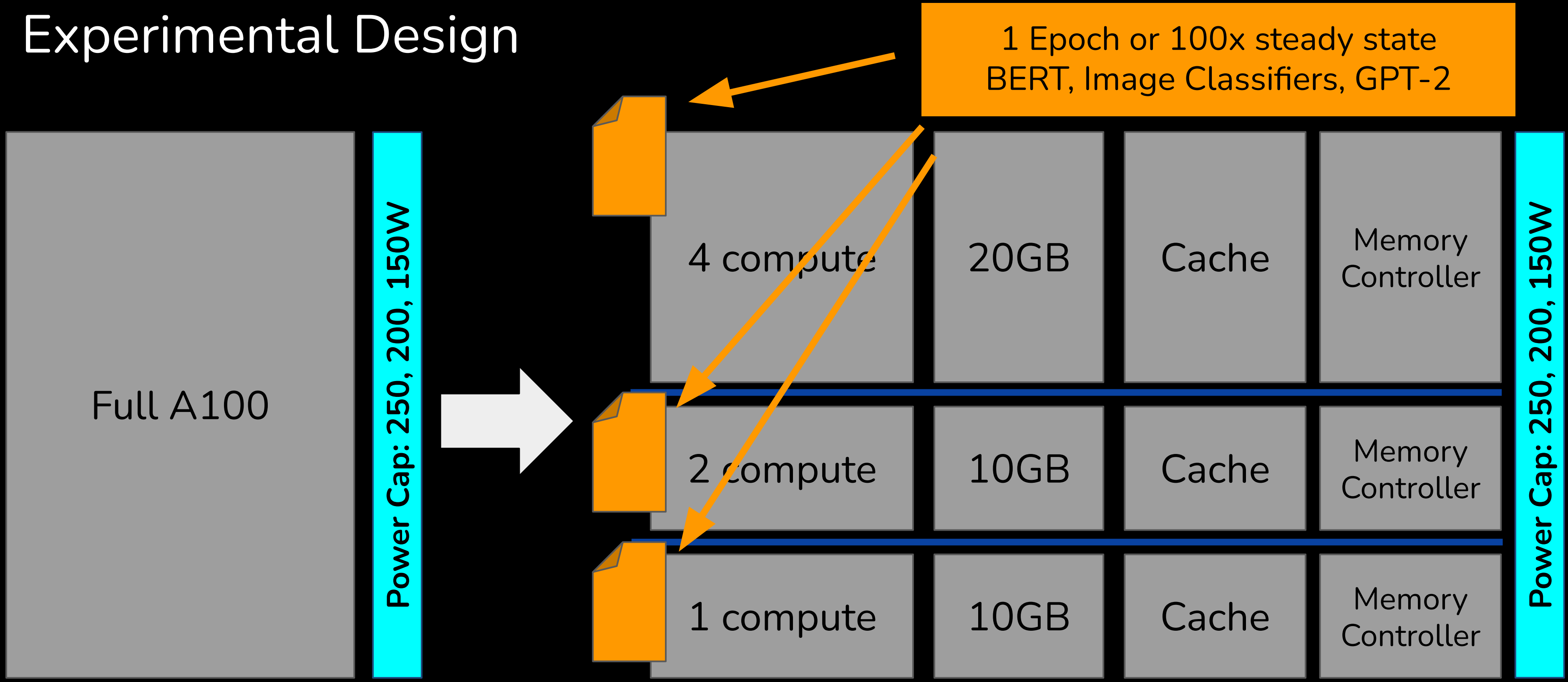
Experimental Design



Experimental Design



Experimental Design



Methodology & Design of Experiments

1. **Run identical copies of given workload on each slice**
2. **Train using maximum batch size permitted by a slice's memory**
3. **Query nvidia-smi power every 250ms & record time to complete iterations**
4. **Divide by number of samples, normalize to full GPU, full power**

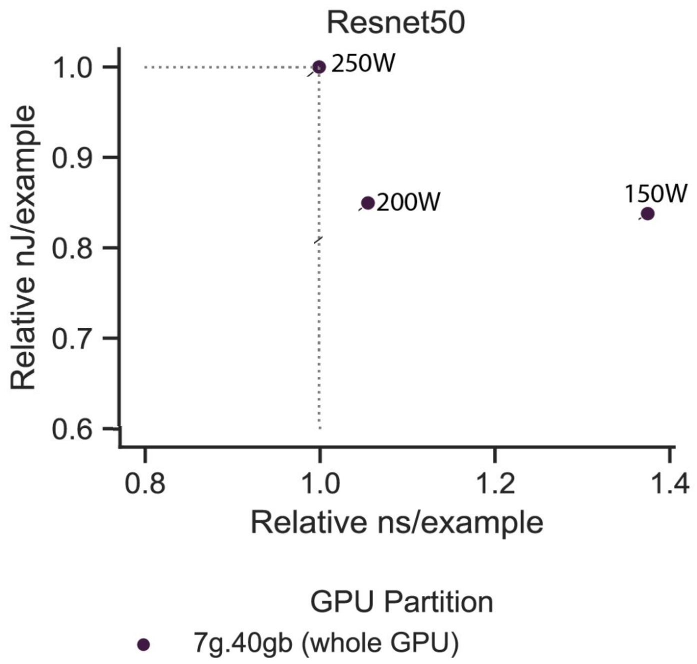
Profiling Metrics

GPU Provider Training Time (ns / example)
Rate examples are processed across all slices in a given configuration

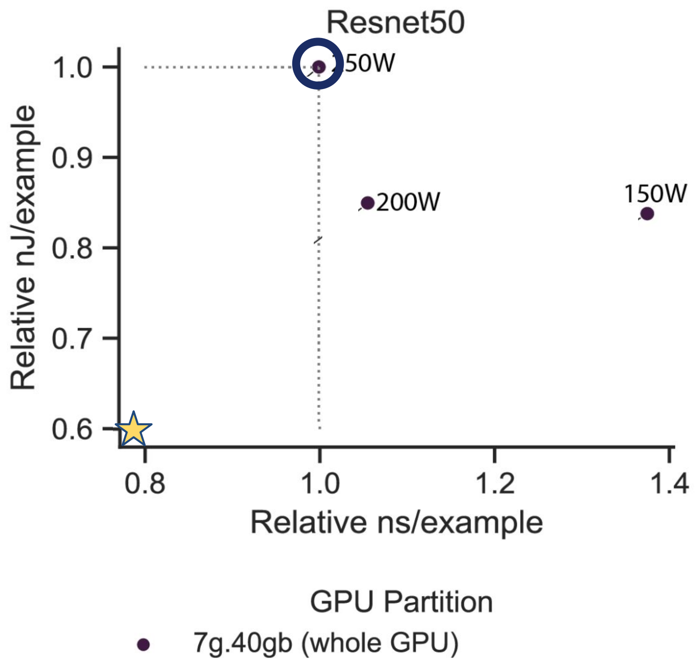
GPU Provider Energy (nJ / example)
Average energy per example while all slices running

Client Training Time (ns / example)
Rate examples are processed for an individual client's workload in a given configuration

Power Capping Has Trade-Offs

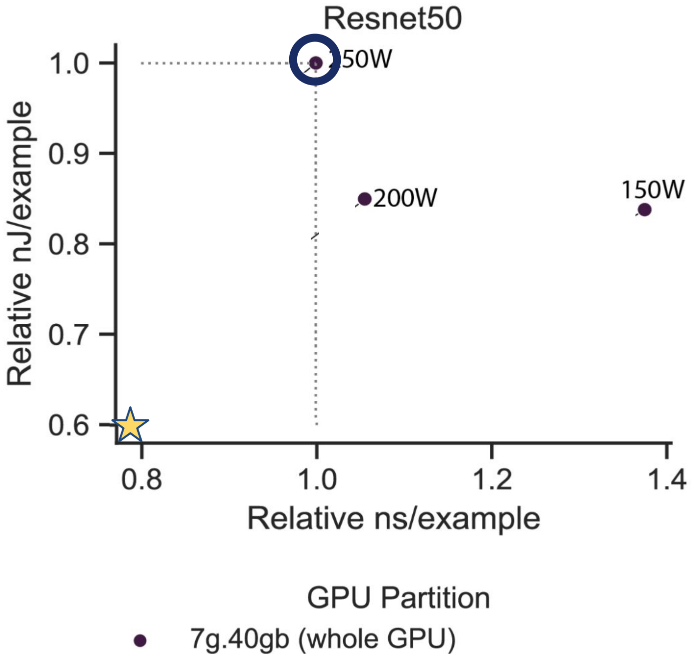


Power Capping Has Trade-Offs



Power Capping Has Trade-Offs

Power capping acts as a dial between speed & energy



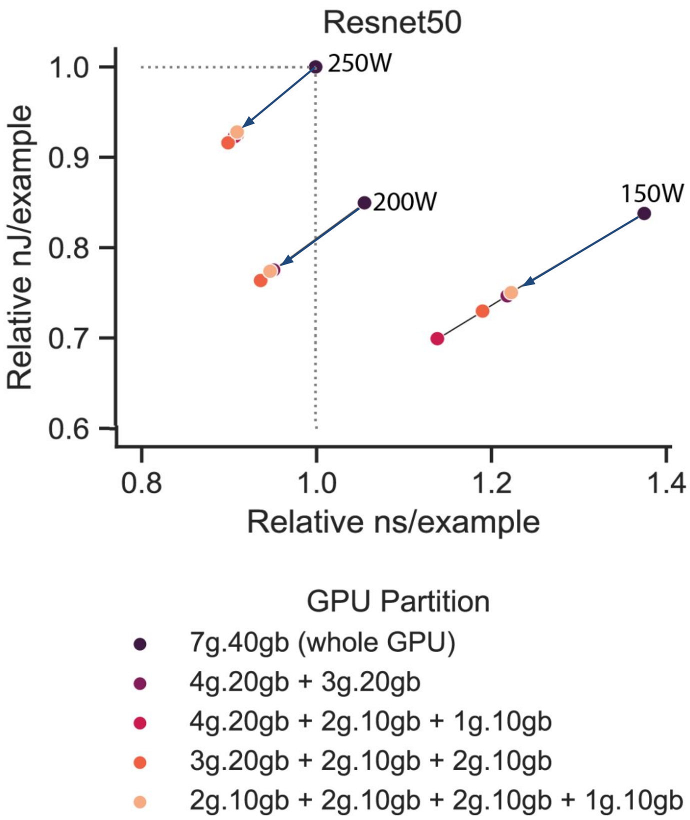
MIG Enables More Jobs in Less Time at Lower Energy

Energy vertical, training time horizontal per example

Down & Left inside box: faster speeds, less energy

Power capping: dial between speed & energy

At every power cap, every multiplexed MIG configuration trains at less time, less energy



MIG Enables More Jobs in Less Time at Lower Energy

Energy vertical, training time horizontal per example

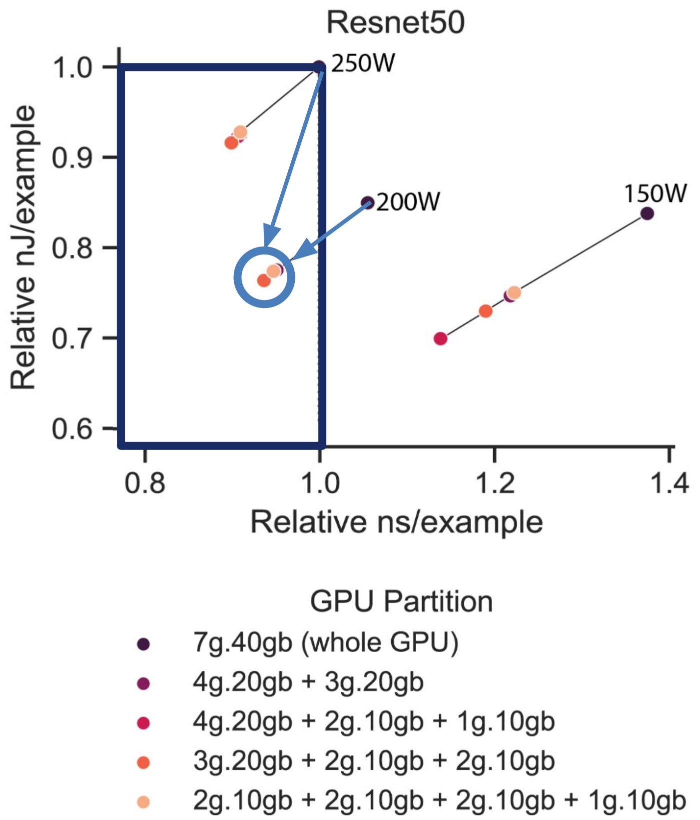
Down & Left inside box: faster speeds, less energy

Power capping: dial between speed & energy

At every power cap, every multiplexed MIG configuration trains at less time, less energy

Throttled MIG at 200W Beats Unthrottled Full GPU by 5% Performance, 20% Energy

MIG makes power capping more effective

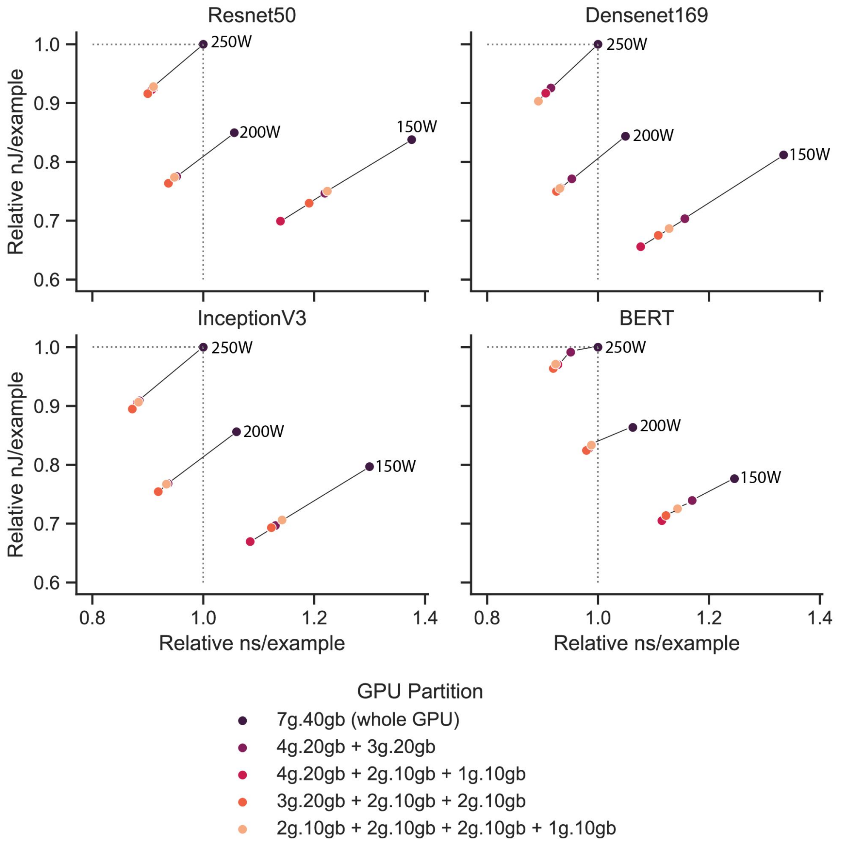


MIG Effectiveness Consistent Across Image Classifiers

Down & Left inside box: faster speeds, less energy

Every MIG trains at less time, less energy

Results hold across image classifiers of similar parameter counts regardless of other model differences



MIG Effectiveness Consistent Across Image Classifiers

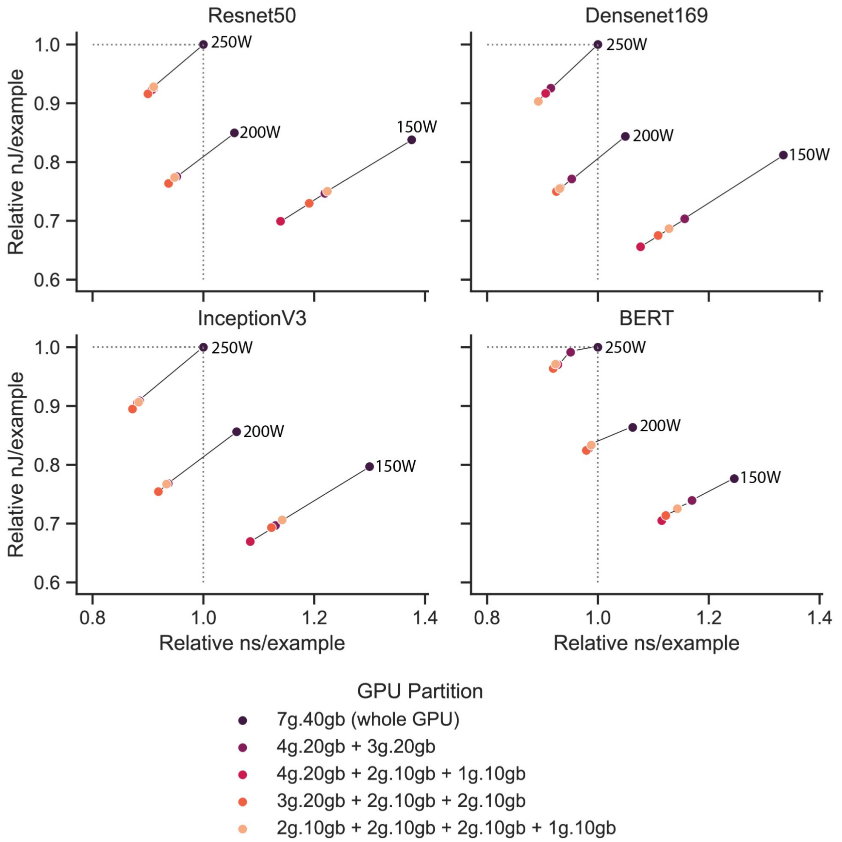
Down & Left inside box: faster speeds, less energy

Every MIG trains at less time, less energy

Results hold across image classifiers of similar parameter counts regardless of other model differences

Results consistent between models within ±1.5% energy, ±0.7% training time on 250W

Results widen with power capping: ±8% energy, ±10% training time on 150W



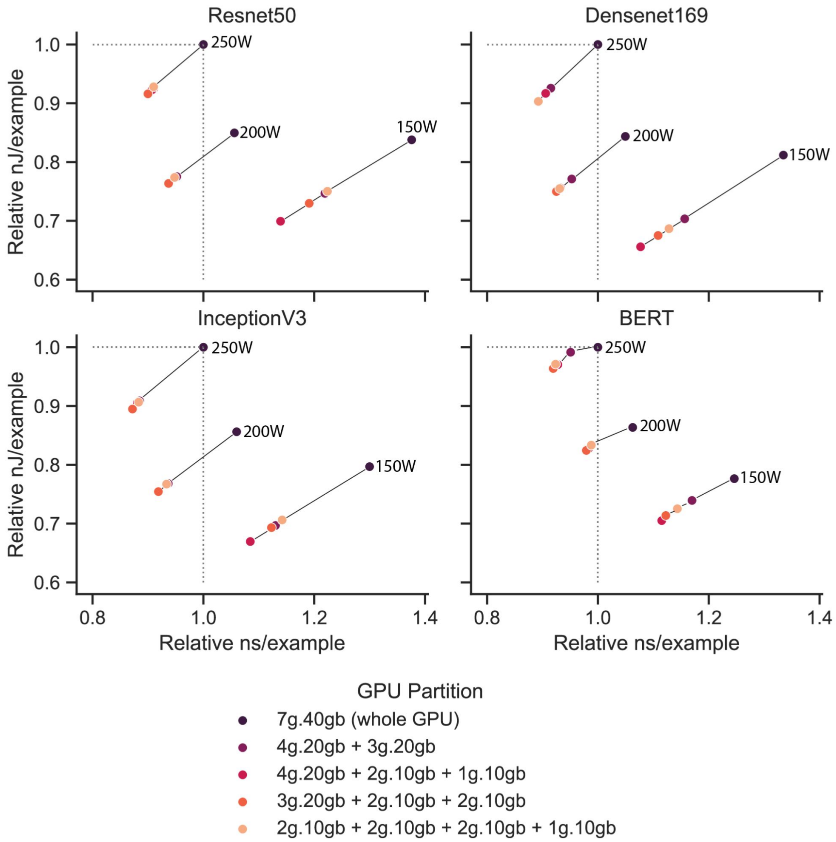
MIG Effectiveness Drops with BERT

Down & Left inside box: faster speeds, less energy

Every MIG trains at less time, less energy

Results hold across image classifiers of similar parameter counts

Worse effectiveness on BERT transformer: 7% faster at 4% less energy



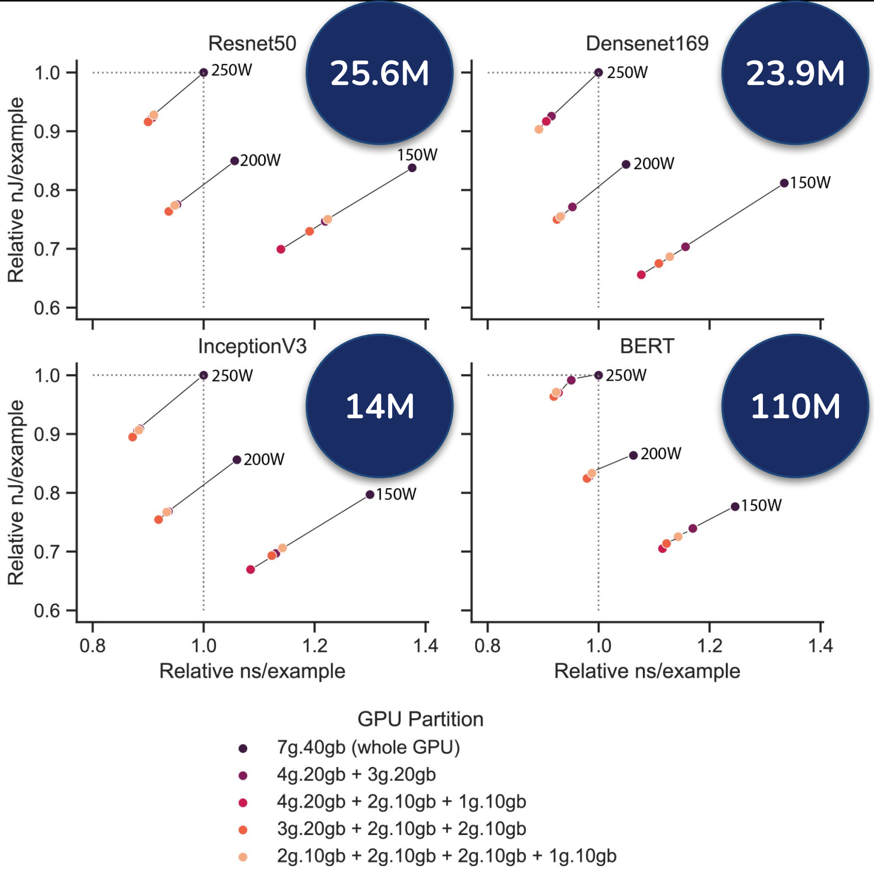
MIG Effectiveness Drops with BERT

Down & Left inside box: faster speeds, less energy

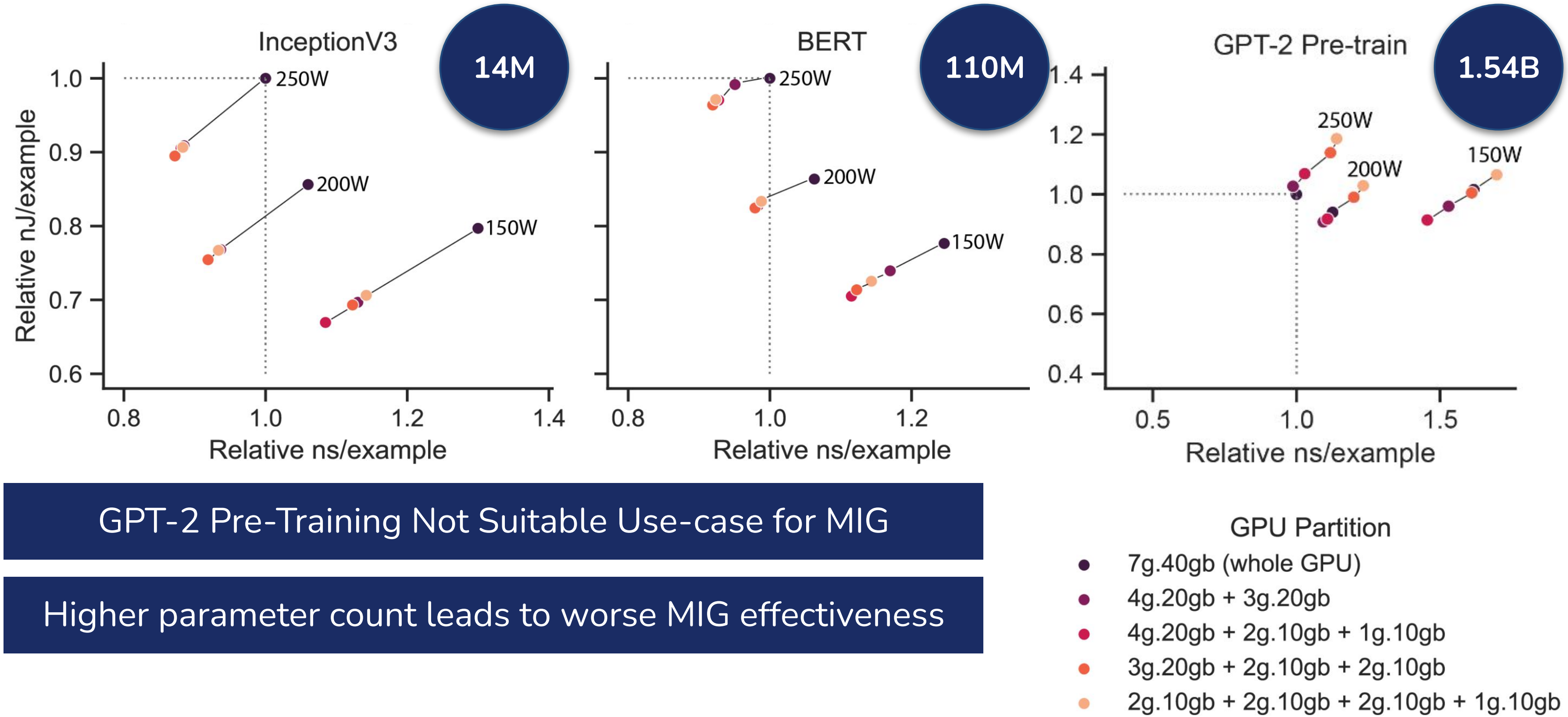
Every MIG trains at less time, less energy

Results hold across image classifiers of similar parameter counts

Worse effectiveness on BERT transformer: 7% faster at 4% less energy



MIG Effectiveness Continues to Decline on GPT-2 Pre-Training



GPT-2 Pre-Training Not Suitable Use-case for MIG

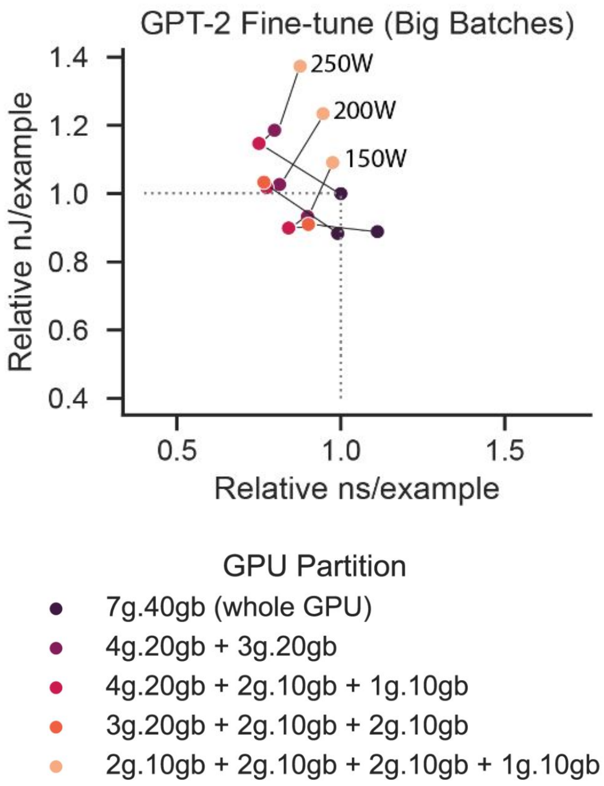
Higher parameter count leads to worse MIG effectiveness

MIG on GPT-2 Fine-Tuning with Largest Batch Possible

GPT-2 Pre-Training Not Suitable Use-case for MIG

Initially, fine-tuning GPT-2 Medium (350M) unpromising

Fine-tuning needs smaller batch sizes: less samples lead to overfitting and longer train times



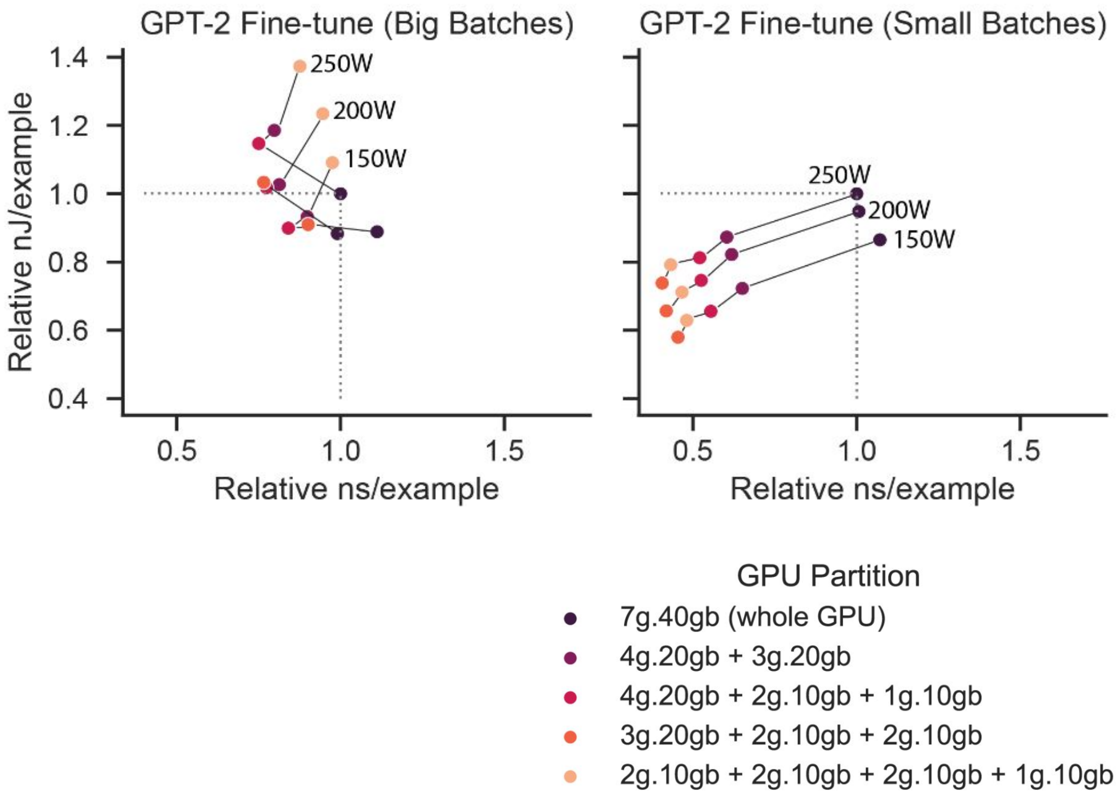
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Challenging the Batch Size Assumption on Fine-Tuning

GPT-2 Pre-Training,

Fine-Tuning Large Batch Poor

Constant, small batches result in strongest MIG gains



Optimal MIG Effectiveness Unlocked on Fine-Tuning

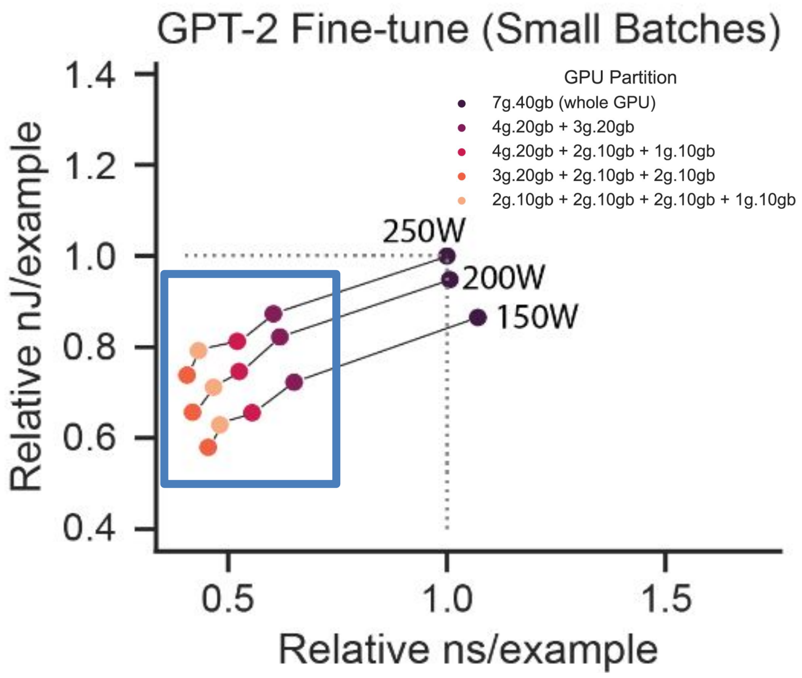
GPT-2 Pre-Training, Fine-Tuning Large Batch

Poor

Constant, small batches result in strongest MIG gains

250W: 59.4% faster at 26.2% less energy

150W: 54.6% faster at 42% less energy



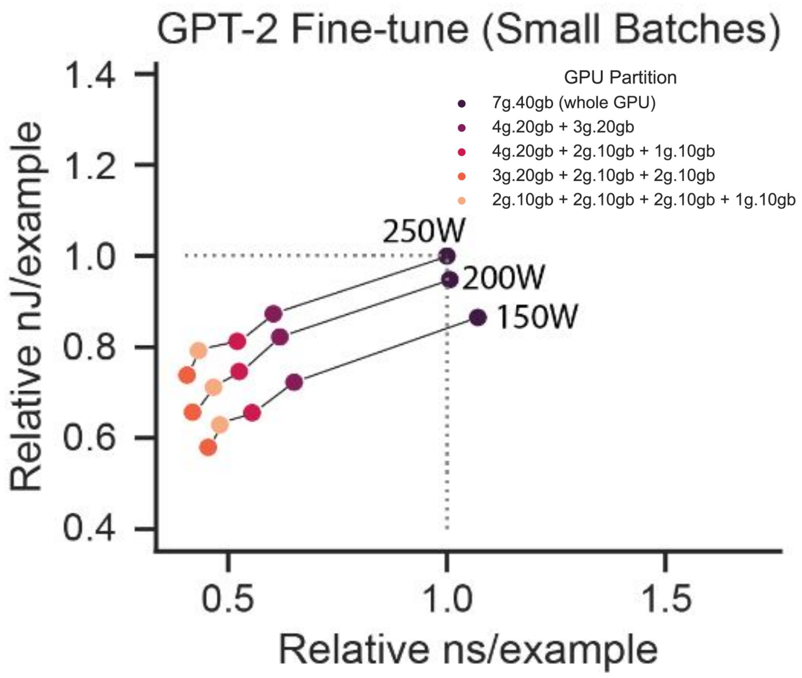
Optimal MIG Effectiveness Unlocked on Fine-Tuning

GPT-2 Pre-Training, Fine-Tuning Large Batch Poor

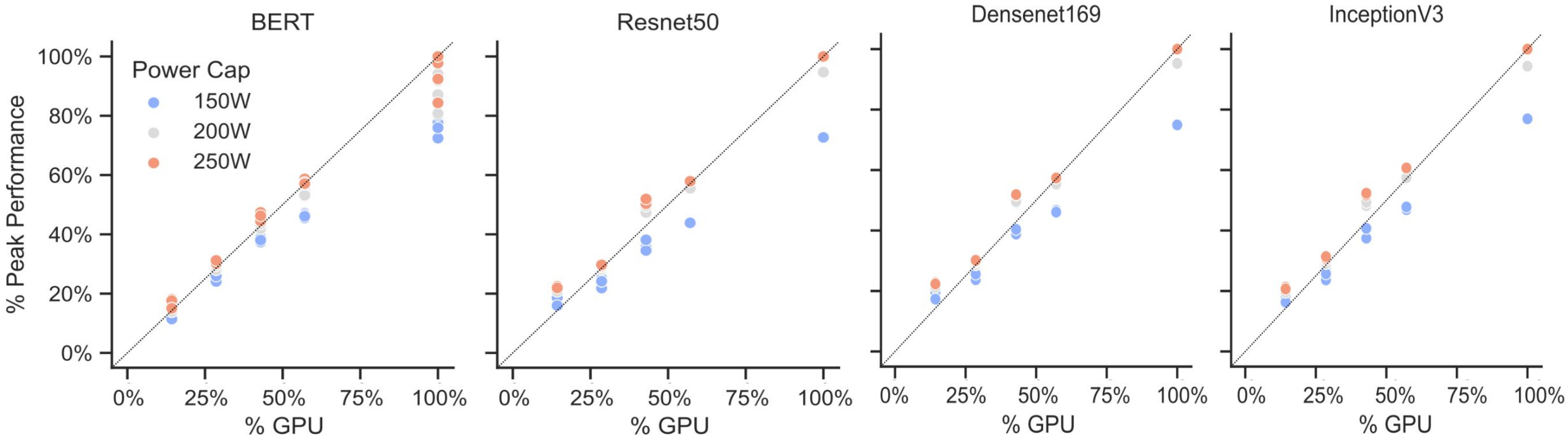
Constant, small batches result in unprecedented MIG gains

Theory: Fine-tuning requires memory more than compute resources

Scaling down GPU most effective with fine-tuning



Latency on Individual Job Impacts Roughly 1:1 with Partition Size



On dashed line: performance to portion of GPU is 1:1

Provider throughput gains do cause client jobs slowdown

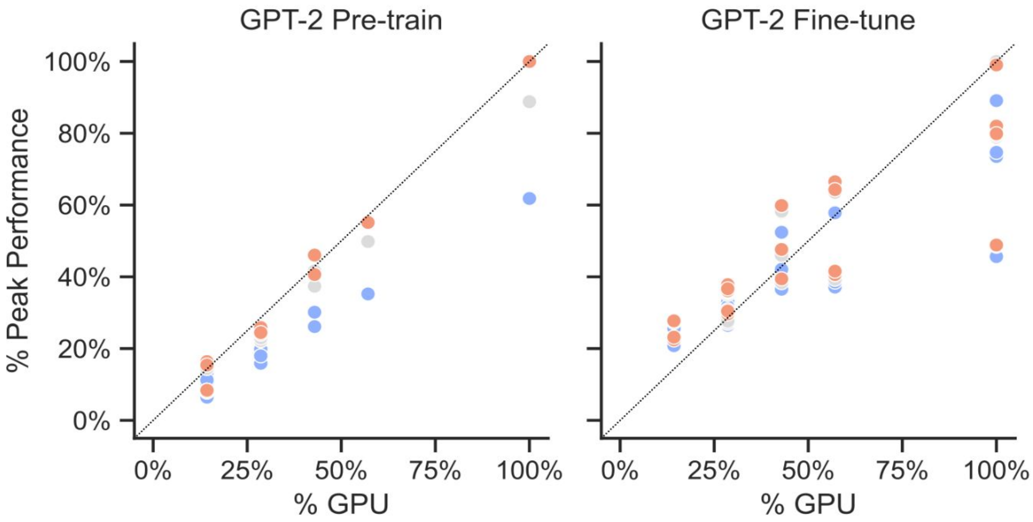
Slowdown roughly proportional to size of resource: 1/7th GPU has 1/7th performance

Latency Improves for Fine-Tuning, Decreases for Larger Models

Slowdown roughly proportional to size of resource: 1/7th GPU has 1/7th performance

GPT-2 Pre-Train experiences worse-than-linear slowdowns

GPT-2 Fine-Tuning experiences better-than-linear slowdowns



Smaller parameter models & fine-tuning provide optimal MIG performance for both provider & client

Conclusions

MIG’s competitive benefits provide increased throughput at decreased power for nearly all models

Partitioning (MIG) benefit decreases as model parameter size increases

Greatest success on LLM fine-tuning, worst performance on LLM pre-training

Clients may experience delays for provider gains, yet also experience better-than-linear slowdowns

Thank you! Questions?

connor.espenshade@columbia.edu