Titan: A Scheduler for Foundation Model Fine-tuning Workloads

Gao Wei[1,2] , Sun Peng[3], Wen Yonggang[1], Zhang Tianwei[1]

ACM SoCC 2022

Outline

- Background about Foundation Models
- Limitations of Existing Solutions
- Proposed Solution: Titan

Outline

- **Background about Foundation Models**
- Limitations of Existing Solutions
- Proposed Solution: Titan

Foundation Model (FM)

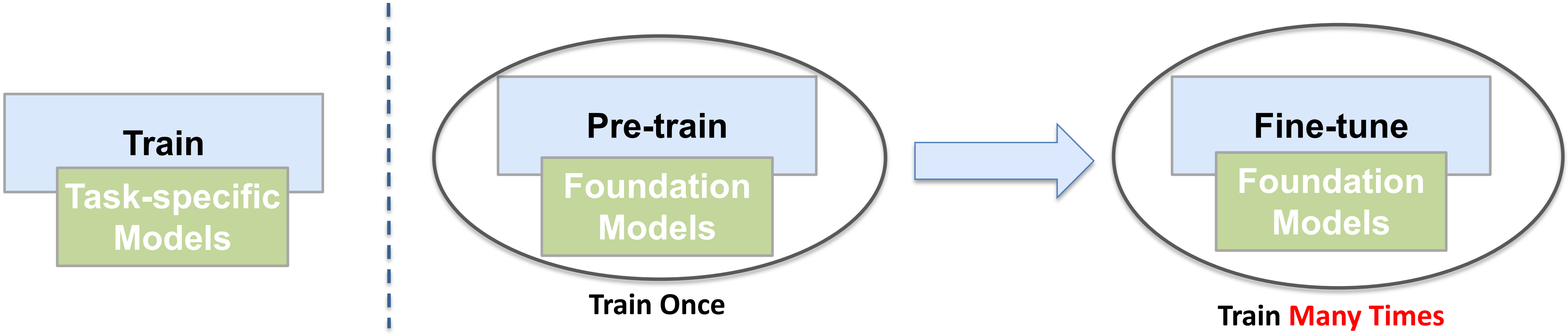
An extremely large model trained on an extremely large dataset

Model	Parameter Size	Dataset Description	Super GLUE Score
Bert	340 M	3300M words	69.0%
GPT-3	175 B	400 billion tokens	71.8 %
PaLM	540 B	Wikipedia + Social Media + News articles	90.4 %

Foundation models achieve impressive performance over many AI tasks

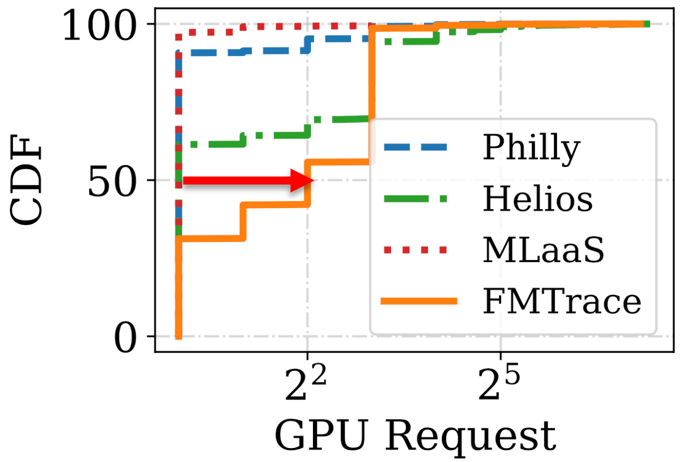
Fine-tuning FMs will become important workloads in GPU datacenters

- **The "pretrain-then-finetune" technique emerges as a new paradigm for building AI systems**
  - OpenAI releases fine-tuning GPT-3 as a paid service for language understanding
  - AliCloud provides a service of fine-tuning M6 which supports various down-stream tasks, e.g., image-text matching, visual question answer

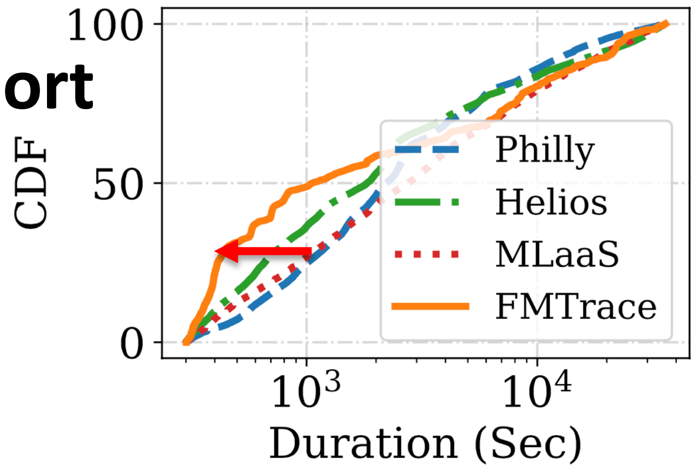


FM fine-tuning workloads demand extensive GPUs for a short time

- **FM fine-tuning workloads tends to request more GPUs**
  - Single GPU device cannot hold the foundation models



- **The duration of FM fine-tuning workloads is relative short**
  - Fine-tuning workloads converge relatively fast

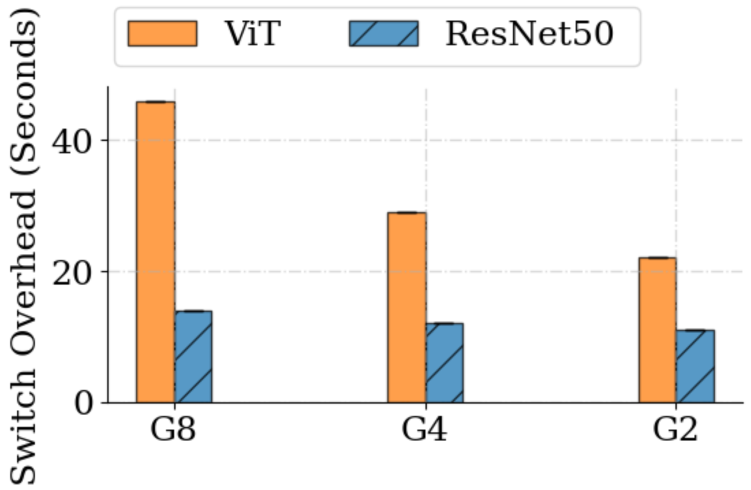


Outline

- Background about Foundation Models
- **Limitations of Existing Solutions**
- Proposed Solution: Titan

Existing Deep Learning (DL) schedulers cannot mitigate the significant context-switch overhead

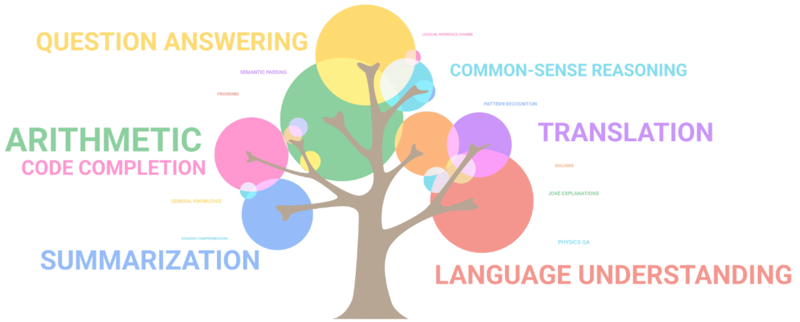
- **Most DL schedulers assume the context-switch overhead is acceptable**
  - *Gandiva [OSDI'18] & Gavel [OSDI'20] frequently make resource re-allocations*
- **However, this is not applicable to FM fine-tuning workloads**
  - The frequent preemption might delay the job progress of FM fine-tuning workloads



Existing DL schedulers manage each job separately

- **Existing schedulers do not consider the *multi-task adaptivity* of FMs**

- **Applying multi-task learning on foundation models can accelerate the convergence of fine-tuned tasks**
  - Jointly fine-tuning FashionMnist and cifar100 can reduce the 1.55x time-to-accuracy



The animation is borrowed from https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

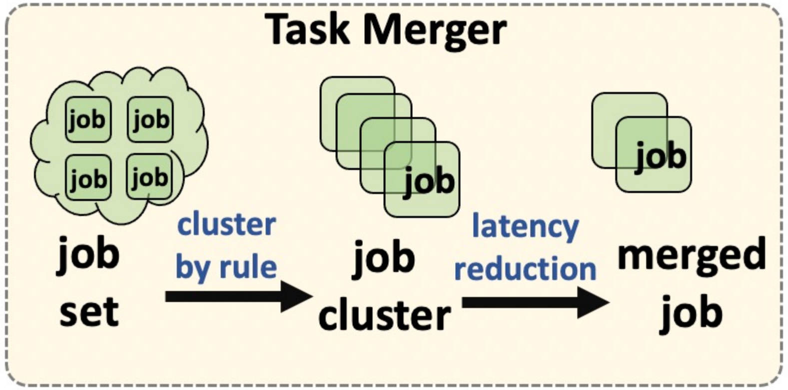
Outline

- Background about Foundation Models
- Limitations of Existing Solutions
- **Proposed Solution: Titan**

Titan contains three key designs

Task merger leverages the multi-task adaptivity

- **Objective**
  - Retain the accuracy
  - Reduce the latency

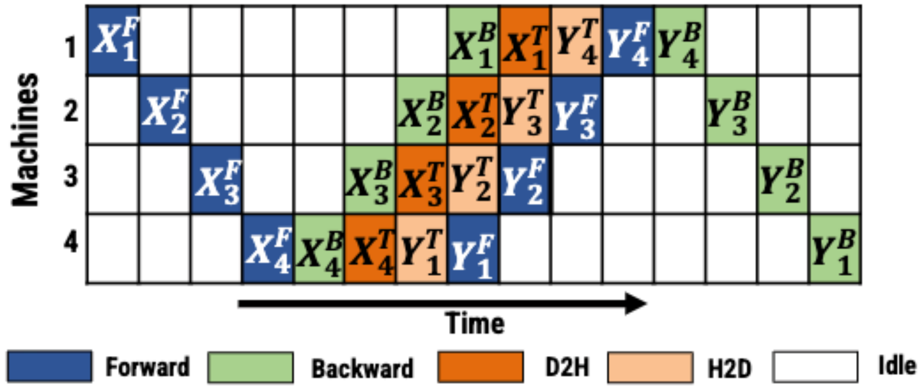


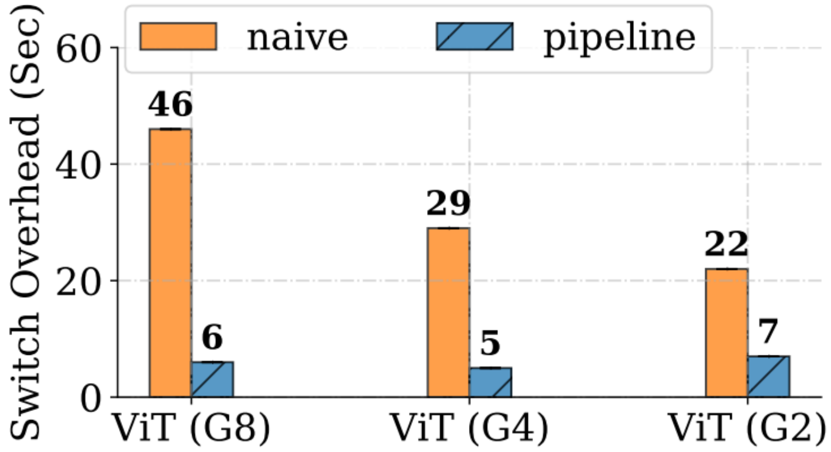
- **Method Overview**
  - A rule-based method to determine whether tasks can be merged without accuracy loss
    calculate classes similarity by internal semantic hierarchy
    similarity('cat', 'dog') > similarity('cat', 'car')
  - Formulate task combination as an Integer Linear Programming (ILP) Problem

	A	B	Task Merger	SRTF
Case 1	20	20	25	30
Case 2	10	20	21	20
Case 3	5	100	102	55

Pipeline switch can address the significant overhead of context switch

- **Objective**
  - Reduce the overhead of context switch
- **Method Overview**
  - Overlap parameter transfer and gradient computation
  - Reversed parameter load





Titan achieves significant performance improvement

Scheduler Policy	Average JCT	Makespan
SRTF	1.68h(ours)	33.09h
Tiresias	1.67h	33.09h
TITAN (w/o task merger)	1.23h	33.11h
TITAN (w/o pipeline switch)	1.16h	**29.01h**
TITAN	**1.04h**	**29.01h**

Table 3: Summary of evaluation results.

- Titan can reduce up to 38% average JCT and 12% makespan compared to baseline schedulers

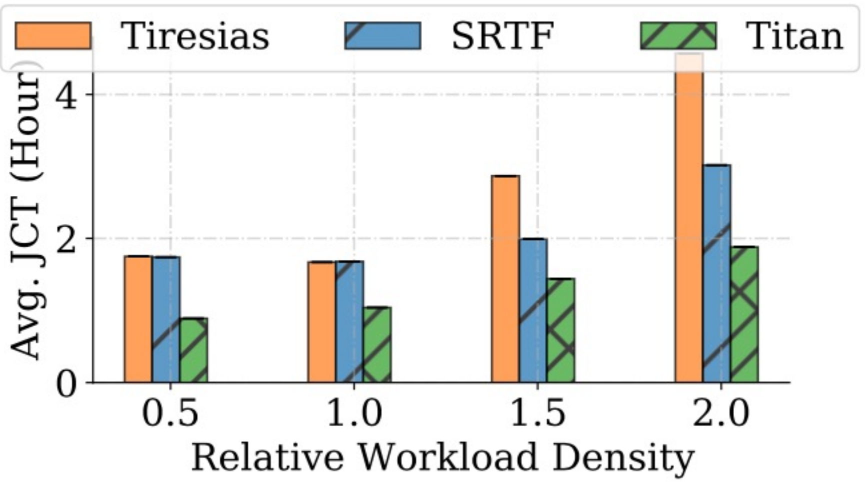


Figure 5: Performance across various workload density.

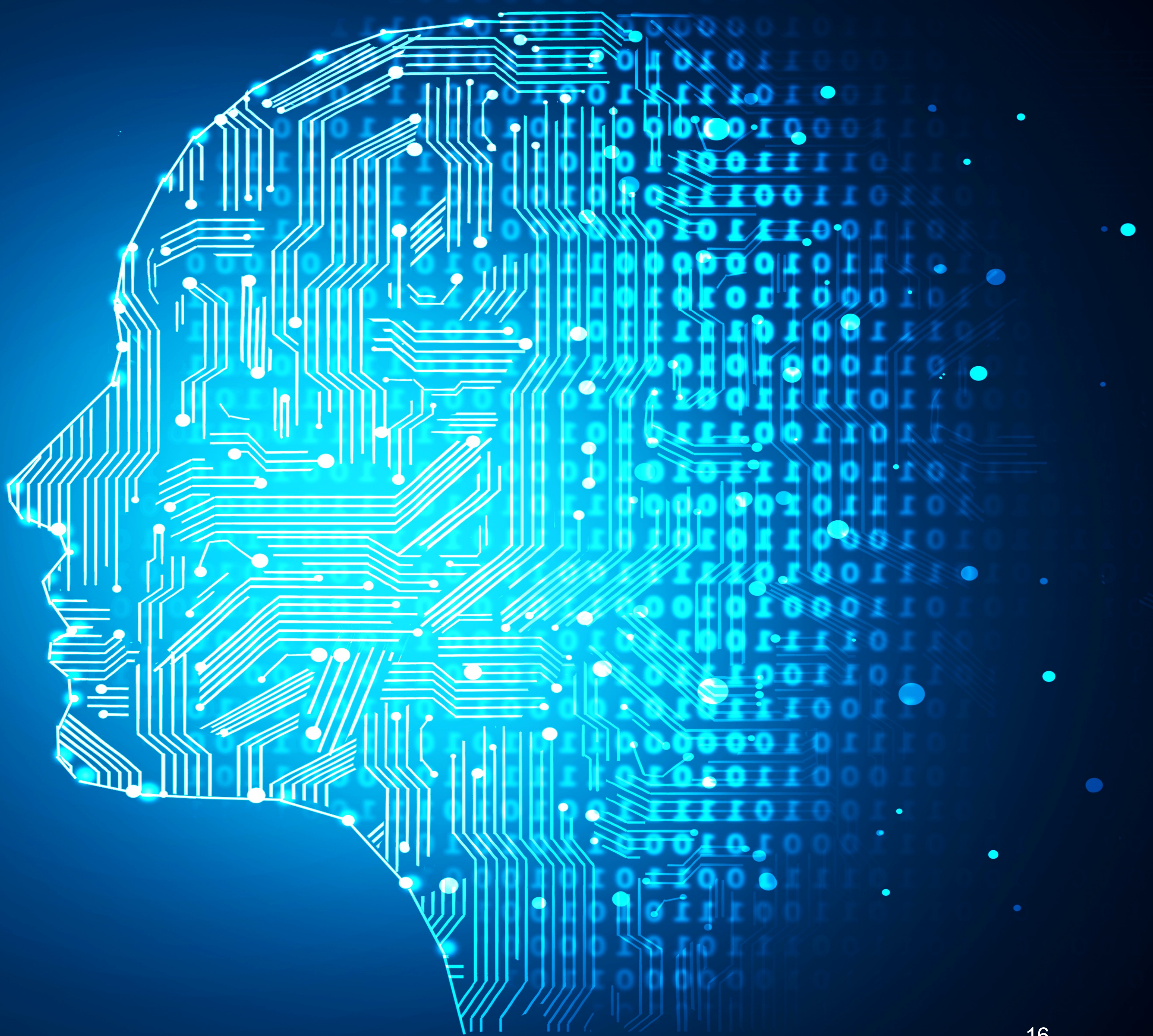
- Titan can maintain its competitive advantage over baseline schedulers with the job density increasing

Conclusion and Future Works

- We present a scheduling system tailored for FM fine-tuning workloads in GPU datacenters
- We need to conduct thorough analysis about the multi-task adaptivity of FM fine-tuning workloads
- We need to extend the pipeline switch to support the single-GPU training

THANK YOU!

Q&A



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE