Mobius: Fine Tuning Large-Scale Models on Commodity GPU Servers

Yangyang Feng, Minhui Xie, Zijie Tian, Shuo Wang, Youyou Lu, and Jiwu Shu
Tsinghua University

http://storage.cs.tsinghua.edu.cn
Explosive Growth of Model Size

3640 petaflop/s-day ≈ A100 x 30 years
≈ $4.5 million

Pre-training and Then Fine Tuning

Large-scale text corpus → Pre-training → Pre-trained models → Fine-tuning → Models for downstream tasks

Much less computing power

Task datasets
Commodity GPU Servers

Data Center GPU Server

- $100,000
- 432 tensor cores
- 900 GB/s inter-GPUs comm.
- GPU Direct P2P

Commodity GPU Server

- $10,000
- 336 tensor cores
- 16 GB/s inter-GPUs comm.
- NO GPU Direct P2P
70% of training time is spent on communication
Mobius Overall

Mobius Partition

Cross Mapping

PCIe Switch

Mobius Pipeline
Mobius Pipeline

Stage_1 \rightarrow Stage_2 \rightarrow Stage_3 \rightarrow Stage_4 \rightarrow \text{Stage}_5 \rightarrow \text{Stage}_6 \rightarrow \text{Stage}_7 \rightarrow \text{Stage}_8
Mobius Pipeline

Stage₁ → Stage₂ → Stage₃ → Stage₄ → Stage₅ → Stage₆ → Stage₇ → Stage₈

DRAM

CPU Switch

GPU Memory

GPU Memory

GPU Memory

CPU Switch

GPU Memory

GPU Memory
Mobius Pipeline

![Diagram of Mobius Pipeline]

- Stage 1, Stage 2, Stage 3, Stage 4, Stage 5, Stage 6, Stage 7, Stage 8
- DRAM
- PCIe Switch
- GPU Memory
Mobius Pipeline

Stage_i’s execution on j^{th} microbatch
Mobius Pipeline

$F_{i, j}$: Stage $i$’s execution on $j$th microbatch

Communication without contention

Communication with contention
Two Partition Questions

- How many stages are in each GPU?
- How many layers are in each stage?
Mobius Partition

- Profile each layer’s memory footprint and computation overhead
- Profile hardware performance, i.e. bandwidth
- Use mixed integer program (MIP) to fine the optimal partition scheme

A → B → C → D → E

MIP
Mobius Partition

minimize \text{Training time of one step}

subject to \text{Memory constraints}
  
  • Memory required by computation
  • Memory required by prefetching

Pipeline order constraints
  
  • Stage execution order
  • Microbatch execution order

More in the paper
Communication Contention

Computation bubble

PCIe Switch

\[
\begin{align*}
F_{1,1} & \quad F_{1,2} & \quad F_{1,3} & \quad F_{1,4} & \quad F_{5,1} & \quad F_{5,2} & \quad F_{5,3} & \quad F_{5,4} \\
F_{2,1} & \quad F_{2,2} & \quad F_{2,3} & \quad F_{2,4} & \quad F_{6,1} & \quad F_{6,2} & \quad F_{6,3} & \quad F_{6,4} \\
F_{3,1} & \quad F_{3,2} & \quad F_{3,3} & \quad F_{3,4} & \quad F_{7,1} & \quad F_{7,2} & \quad F_{7,3} & \quad F_{7,4} \\
F_{4,1} & \quad F_{4,2} & \quad F_{4,3} & \quad F_{4,4} & \quad F_{8,1} & \quad F_{8,2} & \quad F_{8,3} & \quad F_{8,4}
\end{align*}
\]
Cross Mapping

Map adjacent stages to GPUs not under the same CPU root complex
Cross Mapping

Number of GPUs under the same CPU root complex

$$contention(stage_i, stage_j) = \frac{\text{shared}(i,j)}{|i - j|}$$

Time difference to upload the two stages’ data
Experimental Setup

TOPO 2+2

TOPO 3+1

CPU
PCIe
3090 Ti

GPU Topologies
Overall Results

- Mobius and DeepSpeed with heterogeneous memory mode are able to **train larger models**
- Mobius **decreases per-step training time**
- Mobius brings more **significant performance improvement** when the GPU topology has **more severe communication contention**
Communication Analysis

• DeepSpeed with heterogeneous memory mode requires frequent GPU all-to-all collective communications, while Mobius pipeline only transfers **small activations and activation gradients**

• **More than half of the data** is transferred at a bandwidth of more than 12 GB/s in Mobius
Conclusion

- Commodity GPU server is an **affordable** option for fine-tuning large-scale models. However, communication resources on commodity GPU servers are scarce.
- We propose **Mobius** to reduce communication traffic and mitigate communication contention problem.
  - **Mobius pipeline**: heterogeneous memory-based pipeline training scheme
  - **Mobius partition**: find the optimal partition scheme
  - **Cross mapping**: mitigate communication contention
- Mobius significantly reduces the training time by **3.8-5.1 times** compared with the prior art.
Thanks

Mobius: Fine Tuning Large-Scale Models on Commodity GPU Servers

Yangyang Feng, Minhui Xie, Zijie Tian, Shuo Wang, Youyou Lu, and Jiwu Shu

Tsinghua University

http://storage.cs.tsinghua.edu.cn
Email: fyy21@mails.tsinghua.edu.cn