Asian Deans' Forum 2026

THE RISING STARS

Women in Engineering Workshop

SEP 13 – 15 | HKUST, Hong Kong

Understanding Communication Predictability of Distributed LLM Training

Traffic patterns • Communication overhead • ConfigTuner

Dense GPT: predictable • MoE: semi-predictable

Wenxue Li et al. | HKUST · MIT · Peking University · Meta

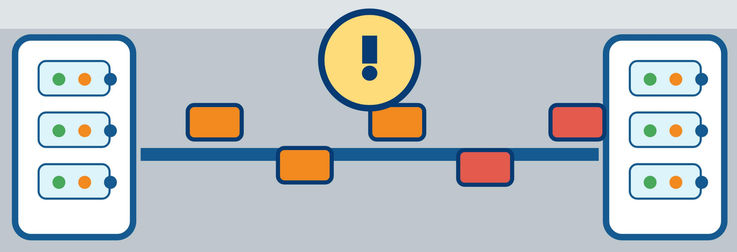
IEEE/ACM Transactions on Networking

1. THE PROBLEM

Communication overhead is significant.

Large models are trained across many GPUs using multiple parallelism strategies that introduce intensive communication.

Data Parallelism; Pipeline Parallelism; Tensor Parallelism; Expert Parallelism; and more...



COMMUNICATION BOTTLENECK

THE CHALLENGE

01 Production traces show that communication can occupy a substantial fraction of an iteration.

02 Prior systems mainly exploit repetition through online profiling.

03 A systematic understanding of predictability is still missing.

WHY IT MATTERS

TIME — ITERATION TIME
Communication directly affects training performance

CFG — TRAINING CONFIGURATION
Parallelism and micro-batch choices change the overhead

GPU — DATA-PARALLEL SCALE
Adding GPUs eventually produces diminishing returns

FROM ONLINE PROFILING TO ANALYTICAL ESTIMATION

COMMON APPROACH

Run → Profile → Reuse

Collect runtime statistics from completed iterations

THIS WORK

Model + config → Estimate → Tune

Determine communication before full execution.

Analytical formulation + a small set of hardware reference values

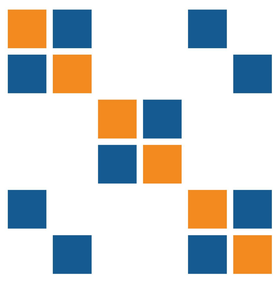
2. THE BIG IDEA

Communication exhibits predictable structure.

The paper analyzes both spatial traffic patterns and temporal communication behavior.

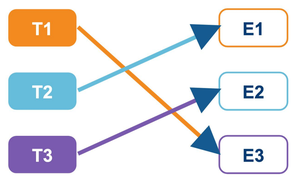
1 Dense GPT: predictable

The parallel configuration determines the communication matrix; the model architecture determines traffic volume; communication blocks repeat over time.



2 MoE: semi-predictable

AllToAll traffic depends on runtime expert selection, but four communications at each expert layer are identical or transposed, and routing becomes more uniform during training.



3 Predict the communication cost

The formulation dissects iteration time into computation, TP, PP, DP, and pipeline bubble time, using GPU utilization and effective bandwidth as reference variables.

GPU + REAL SPEED

THE COMMUNICATION FINGERPRINT

1	2	3
WHO	HOW MUCH	WHEN
GPU pairs	data volume	timing

WHAT SHAPES COMMUNICATION TIME?

GPU	GPU utilization	Stable when per-GPU model size and micro-batch size are fixed.
NET	Effective bandwidth	Its upper bound is set by the GPU interconnect capacity.
MSG	Practical reductions	Message size, collective scale, and synchronization reduce bandwidth.

3. FROM INSIGHT TO ACTION

ConfigTuner predicts training performance.

It applies the analytical formulation to compare candidate configurations with minimal profiling.

HOW IT WORKS

GPT SETTING | TRAINING CONFIG | PLATFORM

↓

Analytical model

↓

ConfigTuner

Output: predicted iteration time for each candidate configuration

WHAT CHANGES?

1.36x throughput — UP TO
Higher throughput than Megatron-LM configurations

~16x less profiling — COMPARED WITH ALPA
Same recommendation with far less profiling

WHAT THIS ENABLES

01 **Tune micro-batch size**
Balance GPU utilization, communication, and pipeline bubbles.

02 **Tune model parallelism**
Compare tensor- and pipeline-parallel configurations.

03 **Analyze data-parallel degree**
Evaluate scaling benefit, training time, and cost.

VALIDATED WITH REAL TRAINING RUNS

The analytical formulation closely matches measured computation time, communication time, bubble ratio, and communication ratio.

We offer new insights into training optimization! :)

Organizers

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY · UNSW SYDNEY · SEOUL NATIONAL UNIVERSITY · NUS National University of Singapore · 臺灣大學 NTU · UTokyo · 清華大學 Tsinghua University