



**CONFERENCE**

—NORTH AMERICA 2026—

# Scaling Foundation Models: From Broken to Near-Linear Multi-Node Training

Sheng Huang, Pinterest



Copyright 2026 Pinterest, Inc.

# What Are We Scaling?

Pinterest • 640M monthly active users • Visual discovery & recommendations

## Pinterest Foundation Model

### USE CASE

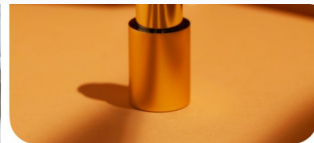
Home Feed & Related Pins ranking

### MODEL ARCHITECTURE

99% of parameters are embeddings

### DISTRIBUTED TRAINING

TorchRec-sharded embeddings · All-to-All communication



...

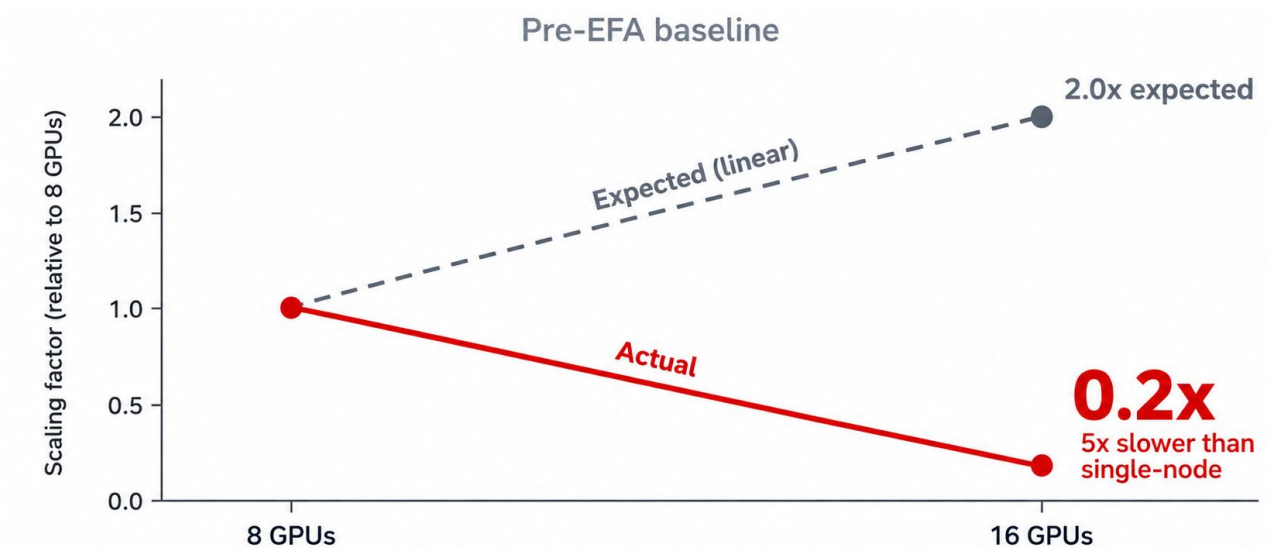


...



# The Second Node Made Training 5× Slower

1 node · 8 GPUs · 1.0× baseline → 2 nodes · 16 GPUs · **0.2× actual** · Expected: ~2.0×



**Instead of speeding up**, adding a second node caused an immediate collapse in throughput.

# Why Sparse Training Gets Hard at Scale

## 1. Network Fabric

Inter-node bandwidth and latency can dominate sparse collectives.

## 2. Data Movement

Distributed embedding lookups create heavy All-to-All traffic.

## 3. Load Balance

Uneven work across ranks makes the slowest rank set step time.

## 4. Topology & Overlap

Where communication runs, and whether it can be hidden, determines the critical path.

**Key Takeaway:** As one bottleneck is removed, another can become dominant.

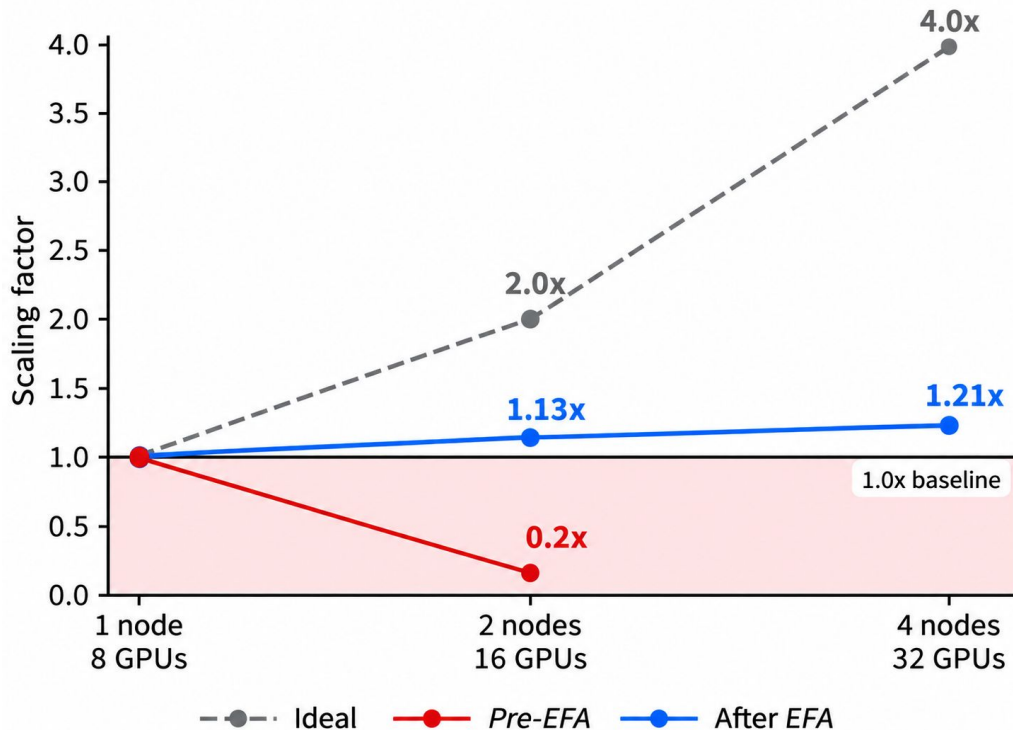
# EFA Fixed the Floor, But Scaling Was Still Flat

## First Fix (Elastic Fabric Adapter)

- Before EFA: 2 nodes → **0.2x**
- After EFA: 2 nodes → **1.13x**
- After EFA: 4 nodes → **1.21x**

Multi-node training became viable, but scaling remained far from ideal.

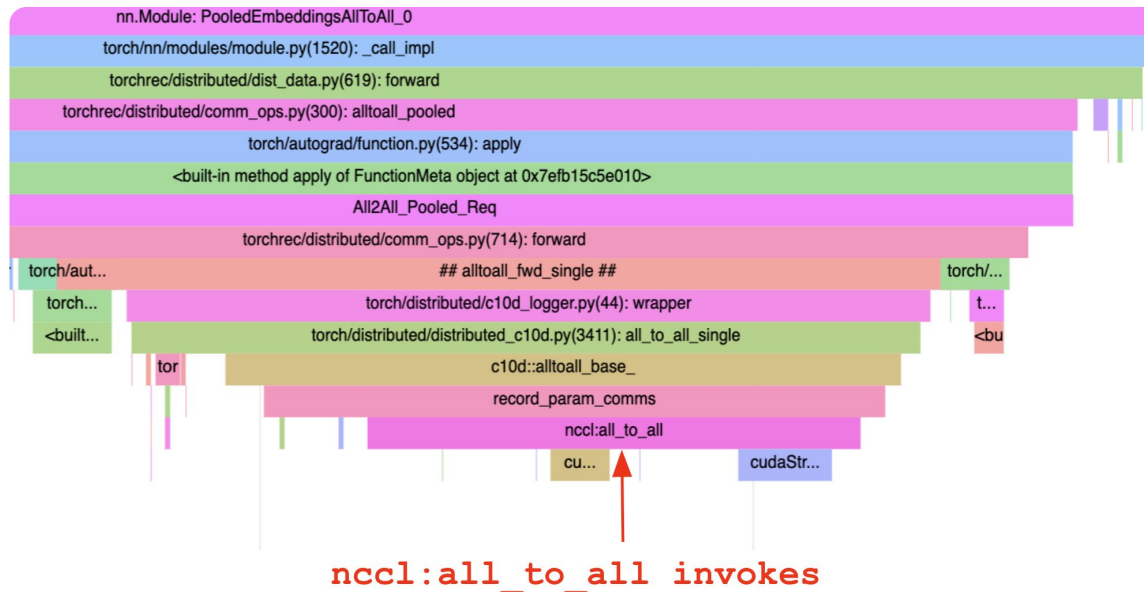
**Takeaway:** EFA removed the networking floor, but not the end-to-end bottleneck.



# Profiler Found the Next Bottleneck

After EFA: 2 nodes = 1.13x · 4 nodes = 1.21x

Viable, but still far from scalable.



Primary Bottleneck

**Embedding All-to-All  
dominated the two-node  
forward path**

Forward pass 400 ms → 700 ms

**+75%**

Most of the Increase came from NCCL  
All-to-All communication

The GPUs looked busy. The profiler showed where the time went.

# What It Took to Reach Near-Linear Scaling

## 01 • Network Floor

### AWS EFA

Made multi-node training viable

## 02 • Communication Volume

### FBGEMM Quantized Comms (QComms)

FP32 → FP8 wire payload

## 03 • Load Balance

### Balanced Sharding

More even work across GPUs

## 04 • Per-Lookup Payload

### Bandwidth-Aware Embedding

Reduced bytes per lookup

## 05 • World-Size Portability

### PyTorch Distributed Checkpoint (DCP)

Load-time resharding across world sizes

## 06 • Topology

### 2D Parallel Topology

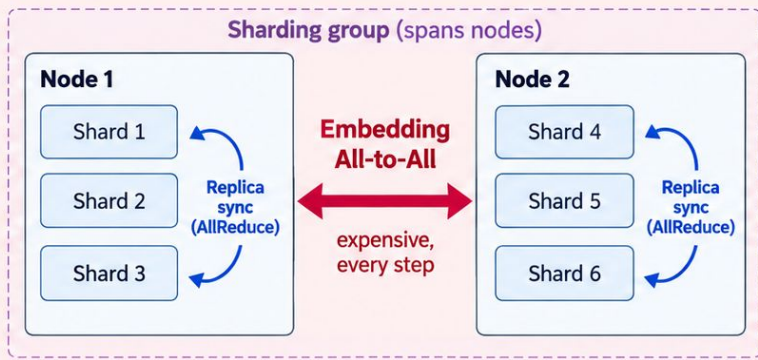
Kept embedding All-to-All on fast local links

No single optimization was enough. We followed the critical path as it moved.

# Put the Expensive Communication on the Fastest Link

## Option 1: A2A crosses nodes

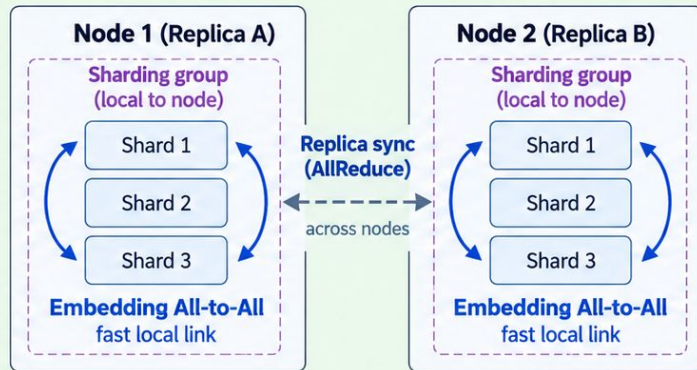
Good for replica sync, bad for embedding-heavy workloads



Inter-host A2A / Intra-host AllReduce

## Option 2: A2A stays local ✓

Chosen for this workload

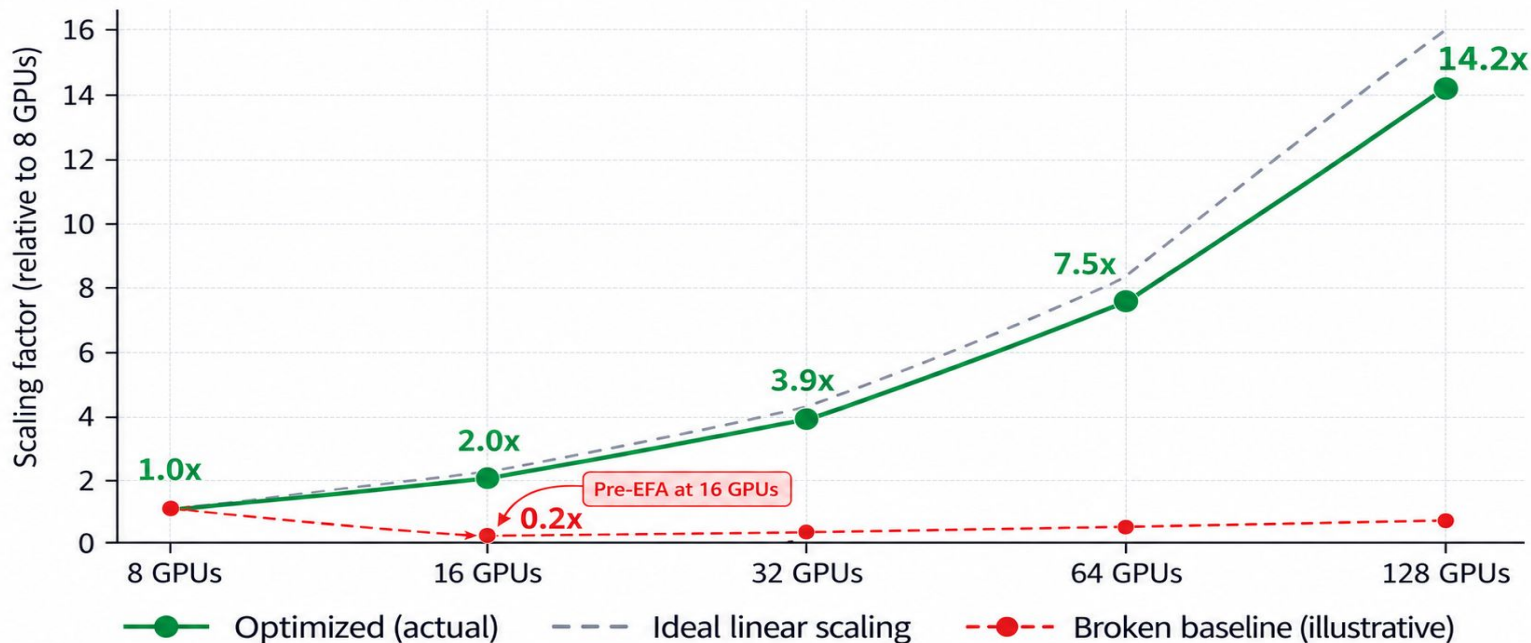


Intra-host A2A / Inter-host AllReduce

For embedding-heavy training, keep the expensive A2A on the fastest local link and let cheaper replica synchronization cross nodes.

# Near-Linear Scaling Beyond 100 GPUs

Pinterest Foundation Model · Multi-node training



# From One Training Job to a Reusable Scaling Playbook

## A Repeatable Systems Loop



## Extending the Loop

The same loop, now driven by an agent.



### Playbook in Experiment Bank

Reuse past results to guide new experiments

Already delivered wins on real training workloads

# Acknowledgments

## Home Feed, Ads, Related Pins

- Abhinav Naikawadi
- Yi Wu
- Zhihua Zhang
- Matthew Lawhon
- Weiguo Ye
- Jenny Jiang
- Sameer Jain
- Eesha Shetty
- Lucy Song
- Usha Amrutha Nookala
- Paulo Soares
- Filip Ryzner
- Konish Dutta
- Nan Li
- Archer Liu
- Dhruvil Deven Badani

## AI Platform

- Saurabh Vishwas Joshi
- Eli Lucherini
- Karthik Anantha Padmanabhan
- Chia-Wei Chen
- Shun Yao Li
- Zirui Li
- Xiangchen Zhao
- Nikhil Kagita
- Alborz Alavian
- Martin Coronel
- Qingxian Lai
- Gaurav Arora
- Lu Niu
- Charles-A. Francisco
- Luyan Wu

## Sponsors

- Bo Liu
- Roger Wang

## Applied Science / ATG

- Pong Eksombatchai
- Chen Yang
- Xihuan Zeng
- Matthew Poska
- Yi-Ping Hsu
- Kousik Rajesh
- Xiangyi Chen
- Hanyu Li
- Jaewon Yang

## Pytorch / TorchRec Team

## SCALING FOUNDATION MODELS

***The real breakthrough wasn't linear scaling.  
It was knowing how to do it again.***