Nexa LLM Acceleration Project

Weekly Progress Update: Model Parallelism & Pipeline Optimization

Dr. Sarah Jenkins Marcus Chen

Vertex Intelligence Lab ▪ Synapse Systems Group

August 2026

Agenda

Meeting Roadmap

Project Overview

Experimental Results

Challenges & Blockers

Next Steps

Weekly Progress Overview

Milestones and Current Status

The **Vertex Intelligence Lab** team focused on accelerating the Nexa LLM pipeline.

- **Model Training:** Completed Nexa-Base-v3 baseline training run.
- **Tensor Slicing:** Implemented dynamic slicing across 64 nodes.
- **Latency Reductions:** Reduced gradient sync latency by 18%.

Task Descriptor	Owner	Priority	Status	Key Commentary
Pipeline Parallelism	M. Chen	High	Done	Merged into main branch.
FP8 Quantization	E. Rostova	Med	Active	Initial speedup of 1.4x verified.
Multi-node memory leakage	S. Jenkins	High	Blocked	Blocked on GPU driver compatibility.

Performance Benchmarking

Throughput Scaling under Model Parallelism

We evaluated the scaling behavior of Nexa-Base-v3 using different tensor and pipeline parallel configurations [1].

Key Takeaways:

- **Tensor Parallelism (TP):** High intra-node performance, but scaling is constrained by NVLink limits.
- **Pipeline Parallelism (PP):** Effective across nodes; communication overhead is hidden.
- **Peak Speed:** Obtained 240 TFLOPS/GPU using TP=2, PP=4.
- **Efficiency:** Sustained 78% scaling efficiency up to 128 GPUs.

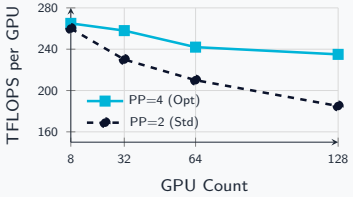

Figure 1: Scaling performance comparison.

Current Blockers & Mitigation Strategy

Technical Impediments

The following items require active resolution before scaling the training run to the full cluster.

Blocker: Memory Leakage in Distributed Optimizer (S. Jenkins)

Out-of-memory (OOM) errors occur when scaling past 64 GPUs.

- **Issue:** VRAM grows linearly by 420 MB per step.
- **Action:** Profiling autograd tensors with PyTorch team.

Risk: Interconnect Bandwidth Bottlenecks (M. Chen)

Nodes lacking direct NVLink connections suffer 15% latency overhead.

- **Mitigation:** Standardize batch sizes to hide pipeline bubbles.

Upcoming Plan & Target Goals

Sprint Deliverables

For the next sprint (ending August 15), we aim to resolve all distribution bottlenecks and finalize the FP8 scaling analysis [2].

Target Deliverables

- Resolve distributed memory leak.
- Release Nexa-Base-v3 model checkpoint.
- Deploy FP8 training container.

Task Assignments:

- **M. Chen:** Integrate communication primitives.
- **E. Rostova:** Benchmark fp8 kernels.
- **S. Jenkins:** Work with IT on cluster setup.

References I

[1] Sarah Jenkins and Marcus Chen. “Optimizing Pipeline Parallelism for Billion-Parameter Models on Heterogeneous clusters”. In: *Vertex Transactions on Machine Learning* 14.2 (2025), pp. 89–104.

[2] Marcus Chen and Elena Rostova. “Dynamic Load Balancing in Multi-Tenant LLM Inference Systems”. In: *Proceedings of the Apex Conference on Distributed Systems (ACDS)*. Meridian Computer Society, 2026, pp. 215–228.