PURETENSOR
INTELLIGENCE DISTILLED

Distributed Inference on Workstation Blackwell

Part 1 — Cluster Bring-Up and Cross-Node Fabric Characterisation

Paper: *PT-R-2026-004 v1.0*
Author: *PureTensor Ltd*
Date: *17 April 2026*
Status: *Published, Part 1 of a multi-part programme*

Abstract

Workstation-class NVIDIA Blackwell GPUs connected by RDMA over Converged Ethernet (RoCE) are evaluated as a deployment target for frontier open-weight large language models in the absence of NVLink or InfiniBand. A three-GPU, two-node cluster of NVIDIA RTX PRO 6000 Blackwell GPUs connected by a 200 Gbps RoCE fabric is brought up, characterised, and exercised with a first cross-node inference run of Llama 3.1 405B Instruct quantised to AWQ-INT4, under pipeline-parallel vLLM orchestrated by Ray. NCCL all-reduce bus bandwidth reaches 23.0 GB/s on the inter-node path, approximately 92% of the 200 Gbps line rate; `iperf3` measures 198 Gbps on the same link, approximately 99% of line rate. GPUDirect RDMA is verified active through NCCL debug logs. The 405-billion parameter model loads and generates across three GPUs split as three pipeline stages of 42 decoder layers each. This paper is Part 1 of an ongoing research programme characterising distributed inference of frontier open-weight models on workstation-class hardware.

1. Introduction

Frontier open-weight large language models are released at parameter counts and memory footprints that presuppose deployment on hyperscaler-grade infrastructure: DGX systems, HGX B200 platforms, dense NVLink fabrics, InfiniBand interconnect. The reference deployments and vendor benchmarks for Llama 3.1 405B, Nemotron Super 120B, Qwen3 235B and the forthcoming Nemotron Ultra 500B are all specified in that domain.

A distinct and commercially relevant segment, however, cannot or will not procure HGX-class infrastructure. Research laboratories, regulated verticals, sovereign deployments, and mid-size enterprises face capital, operational, data-residency or regulatory constraints that rule out hyperscaler-tier gear. What this segment can procure is workstation-class professional GPUs in tier quantities and commodity high-speed Ethernet fabric. Whether frontier models are practically deployable at useful throughput on such topologies, in the explicit absence of NVLink and InfiniBand, is an open empirical question.

This paper reports Part 1 of a research programme that answers the question empirically. The scope of Part 1 is to (a) bring up a three-GPU, two-node cluster of NVIDIA RTX PRO 6000 Blackwell GPUs interconnected by a 200 Gbps RoCE fabric, (b) characterise that fabric under both synthetic throughput tests and NCCL collective benchmarks, (c) verify GPUDirect RDMA

end-to-end, and (d) demonstrate first cross-node inference of a frontier open-weight model under a production-grade serving stack.

2. System Under Test

All measurements reported in this paper were produced on the PureTensor Trinity cluster, a two-node deployment built from workstation-class and commodity components. The component selection is deliberate: results must be transferable to the buyer profile the programme addresses, which precludes reliance on datacenter-exclusive hardware.

Component	Specification
Node A (local)	AMD Ryzen Threadripper PRO 9975WX (32 core, single socket), 512 GB DDR5 ECC, two NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPUs, PCIe Gen 5 x16 per GPU, 200 Gbps Mellanox ConnectX-6 NIC
Node B (remote)	AMD Ryzen Threadripper 7970X (32 core, single socket), 256 GB DDR5, one NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU, PCIe Gen 5 x16, 200 Gbps Mellanox ConnectX-6 NIC
Switch	MikroTik CRS812-4XS-RM with two QSFP-DD 400 Gbps ports and two QSFP56 200 Gbps ports. No NVLink, no InfiniBand.
Transport	RDMA over Converged Ethernet v2 (RoCEv2), NCCL with GPUDirect RDMA
Storage	Consumer NVMe Gen 4 working set per node, distributed object store for shared state
Software	Linux 6.x, current NVIDIA driver and CUDA toolkit, Ray 2.x, vLLM with AWQ Marlin kernels, Prometheus, Grafana, DCGM exporter, nccl-tests, iperf3

3. Methodology

3.1 Fabric characterisation

Two independent and separately reported measurements are taken on the inter-node link.

- Raw TCP throughput using `iperf3` with eight parallel streams and default window sizing. Reported result is line-rate host-to-host throughput and does not involve the GPU data path.
- NCCL collective bus bandwidth via `all_reduce_perf` from NVIDIA's nccl-tests suite across message sizes from 1 MiB to 4 GiB. Reported result is NCCL bus bandwidth in the standard definition: effective data throughput in the steady state of the ring or tree collective, which includes NCCL protocol overhead and, when GPUDirect RDMA is active, direct NIC-to-GPU memory transfers.

These two measurements describe different properties of the fabric and should not be merged into a single wire-rate percentage. `iperf3` bounds what the Ethernet link can carry; NCCL bus bandwidth bounds what a collective inference or training run will see in practice. This paper

reports both, separately.

3.2 GPUDirect RDMA verification

NCCL is configured for RoCE with GPUDirect RDMA through the following environment:

```
NCCL_IB_DISABLE=0, NCCL_NET_GDR_LEVEL=5, NCCL_IB_HCA=mlx5_0,
NCCL_IB_GID_INDEX=3, NCCL_SOCKET_IFNAME=<interface names>,
NCCL_BUFFSIZE=8388608, NCCL_DEBUG=INFO
```

Debug output is inspected for the "GPU Direct RDMA Enabled" line. Runs lacking this confirmation are discarded: without it, NCCL silently falls back to host-staged transfers and bus bandwidth drops by roughly an order of magnitude.

3.3 First cross-node model run

Llama 3.1 405B Instruct quantised to AWQ-INT4 (approximately 215 GB on disk) is deployed across the three-GPU cluster in a pipeline-parallel configuration. The model has 126 transformer decoder layers and 128 attention heads. Tensor parallelism across three GPUs is precluded because 128 is not evenly divisible by 3; pipeline parallelism with three stages of 42 layers each is used instead. The first two stages reside on Node A (intra-node p2p over PCIe); the third stage resides on Node B (cross-node p2p over RoCE). Serving is handled by vLLM with Ray as the distributed executor backend.

This configuration places the primary inter-stage communication on NCCL p2p send/recv. Pipeline parallelism does not use all-reduce in the steady state; the fabric-level all-reduce measurements in section 4.1 are therefore a separate fabric-validation exercise, not the communication primitive of the inference run itself. A full pipeline-bubble and p2p-latency analysis of this serving configuration is deferred to Part 2 of the programme.

4. Results

4.1 Fabric measurements

Measurement	Tool	Result	Relative to 200 Gbps
TCP throughput	iperf3, 8 streams	198 Gbps	approximately 99%
NCCL bus bandwidth	all_reduce_perf	23.0 GB/s (approx. 184 Gbps)	approximately 92%
GPUDirect RDMA	NCCL debug log	Enabled	verified

Ring and tree all-reduce patterns were both exercised. Bus bandwidth approaches the reported figure on large messages (256 MiB and above) and declines on smaller messages as collective protocol overhead dominates. This behaviour is consistent with published characterisations of NCCL over RoCE on Mellanox hardware.

4.2 Cross-node inference

Llama 3.1 405B AWQ-INT4 loads across the three-GPU pipeline without incident. Per-GPU VRAM occupancy is approximately 72 GB for model weights, leaving approximately 24 GB of headroom per GPU for KV cache and activation buffers at a bounded maximum context length. Generated output is coherent and consistent against held-out reference prompts.

Detailed tokens-per-second, time-to-first-token and pipeline-bubble measurements for this configuration are deferred to Part 2. The purpose of the Part 1 run is to demonstrate that the frontier-scale model loads and generates correctly end-to-end across the workstation-class topology, which it does.

5. Discussion

The 92% NCCL-over-RoCE efficiency observed on the 200 Gbps link is broadly in line with public characterisations of equivalent Mellanox-based fabrics. The gap between `iperf3` line rate (99%) and NCCL bus bandwidth (92%) reflects NCCL protocol overhead, synchronisation and ring-descriptor exchange rather than fabric saturation. The practical consequence for workstation-class deployments is that at 200 Gbps, cross-node NCCL collectives are not fabric-bound. Whether this remains true at 400 Gbps, and whether pipeline-parallel p2p send/recv is similarly well-served, are questions for Part 2 and Part 3.

NCCL configuration for non-NVLink topologies is not automatic. The settings above, in particular `NCCL_NET_GDR_LEVEL=5` and an explicit `NCCL_IB_GID_INDEX`, are preconditions for GPUDirect RDMA over RoCE. Omitting them does not produce an error: NCCL silently falls back to host-staged transfers and the cluster appears to function, but bus bandwidth collapses into the single-digit GB/s range. Operators migrating from NVLink-native references to RoCE should treat GPUDirect RDMA activation as a defensible configuration step requiring log-level verification, not a default behaviour.

6. Related Work

Large-language-model deployments on NVLink-rich hyperscaler infrastructure are extensively characterised in vendor reference architectures and academic systems papers. At the other extreme, consumer-grade multi-GPU configurations over PCIe are documented in hobbyist and small-scale research settings. The middle tier, workstation-class professional GPUs connected by Ethernet-based RDMA fabric at 200 Gbps and above, is comparatively undercharacterised. Vendor materials on NVIDIA GPUDirect RDMA and NCCL over RoCE provide the foundational protocol description; this programme contributes an empirical, end-to-end characterisation of a concrete deployment at frontier model scale.

7. Conclusion and Roadmap

Part 1 establishes the baseline. The fabric is healthy, GPUDirect RDMA is verified active, NCCL all-reduce bus bandwidth reaches approximately 92% of the 200 Gbps line rate, `iperf3` reaches approximately 99%, and a 405-billion parameter open-weight model loads and generates correctly across a three-GPU pipeline spanning two nodes.

The remainder of the programme is staged to produce a self-contained deliverable at each step. Forthcoming parts are scoped as follows:

- **Part 2.** Pipeline-parallel throughput and latency analysis, including pipeline-bubble utilisation and p2p send/recv latency at 200 Gbps, across Llama 3.1 405B, Nemotron Super 120B and Qwen3 235B.
- **Part 3.** Fabric upgrade to 400 Gbps RoCE (ConnectX-7) and characterisation of the 200 Gbps to 400 Gbps delta for distributed inference.
- **Part 4.** Asymmetric-topology parallelism: mixing full Workstation Edition and Max-Q Workstation Edition SKUs in the same distributed configuration.
- **Part 5.** Flagship measurement on Nemotron Ultra 500B once released.
- **Part 6.** Consolidated whitepaper, open benchmark dataset, and conference submission.

All raw benchmark data, harness code, and analysis scripts for the programme will be released under a permissive licence. Third parties running the same harness on comparable hardware produce directly comparable result files.

8. Acknowledgements

The cluster described in this paper was assembled under the NVIDIA Inception programme, which grants member startups access to professional NVIDIA hardware through NVIDIA's distributor network. Specific thanks are due to the NVIDIA Inception team for the reseller introduction that enabled procurement of the third Blackwell GPU used in these measurements.