Qualcomm

Scaling LLM Training Using RDMA over Converged Ethernet

Adrián Pérez Diéguez
Staff Engineer, Qualcomm AI Research

SC INDIS Workshop
November 16th, 2025, St. Louis, Missouri, USA

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

Training LLMs at Scale: Motivation

Objective: Demonstrate LLM training scalability with RoCE.

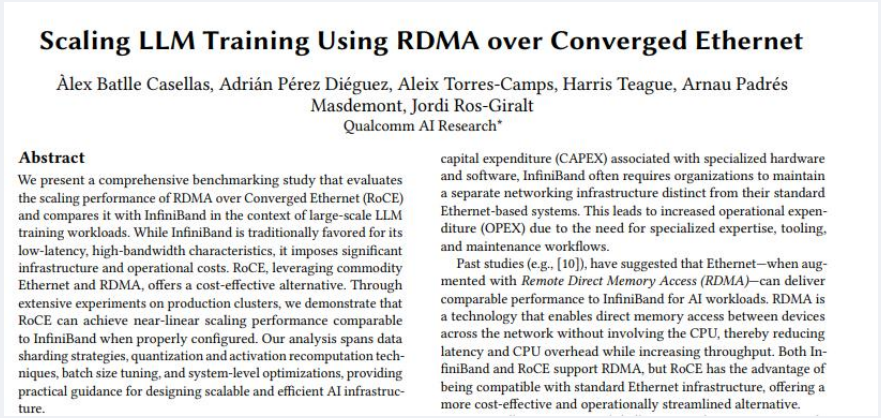
- InfiniBand requires a separated networking infra, coming with significant cost implications.

- TCP is not optimal for massive communication during LLM training.

so...

Is RoCE a good alternative for LLM training? What techniques can we use to scale with nodes?

More details in paper:



Scaling LLM Training Using RDMA over Converged Ethernet

Àlex Batlle Casellas, Adrián Pérez Diéguez, Aleix Torres-Camps, Harris Teague, Arnau Padrés Masdemont, Jordi Ros-Giralt
Qualcomm AI Research*

Abstract

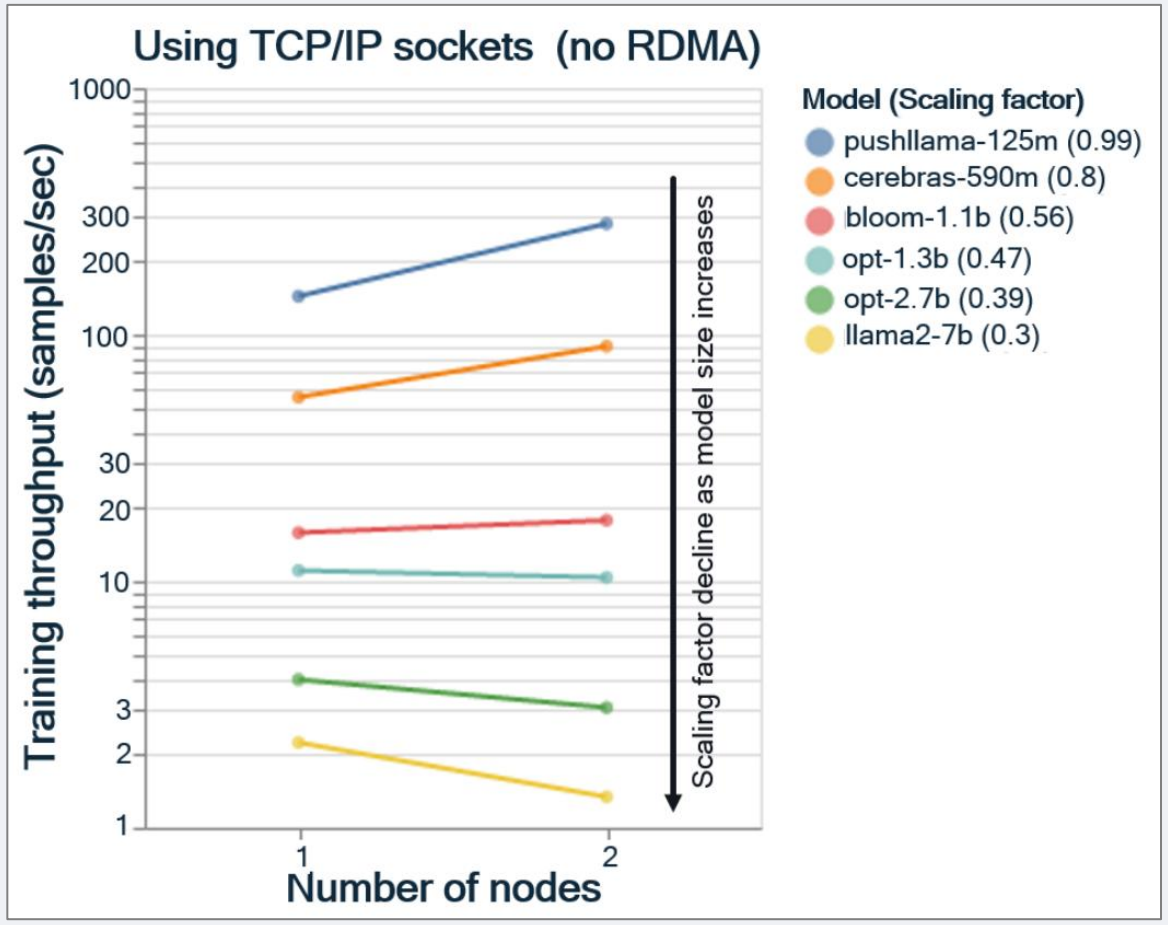
We present a comprehensive benchmarking study that evaluates the scaling performance of RDMA over Converged Ethernet (RoCE) and compares it with InfiniBand in the context of large-scale LLM training workloads. While InfiniBand is traditionally favored for its low-latency, high-bandwidth characteristics, it imposes significant infrastructure and operational costs. RoCE, leveraging commodity Ethernet and RDMA, offers a cost-effective alternative. Through extensive experiments on production clusters, we demonstrate that RoCE can achieve near-linear scaling performance comparable to InfiniBand when properly configured. Our analysis spans data sharding strategies, quantization and activation recomputation techniques, batch size tuning, and system-level optimizations, providing practical guidance for designing scalable and efficient AI infrastructure.

capital expenditure (CAPEX) associated with specialized hardware and software, InfiniBand often requires organizations to maintain a separate networking infrastructure distinct from their standard Ethernet-based systems. This leads to increased operational expenditure (OPEX) due to the need for specialized expertise, tooling, and maintenance workflows.

Past studies (e.g., [10]), have suggested that Ethernet—when augmented with *Remote Direct Memory Access (RDMA)*—can deliver comparable performance to InfiniBand for AI workloads. RDMA is a technology that enables direct memory access between devices across the network without involving the CPU, thereby reducing latency and CPU overhead while increasing throughput. Both InfiniBand and RoCE support RDMA, but RoCE has the advantage of being compatible with standard Ethernet infrastructure, offering a more cost-effective and operationally streamlined alternative.

Training LLMs at Scale: Motivation

As model size increases, network performance collapses when using traditional TCP/IP sockets

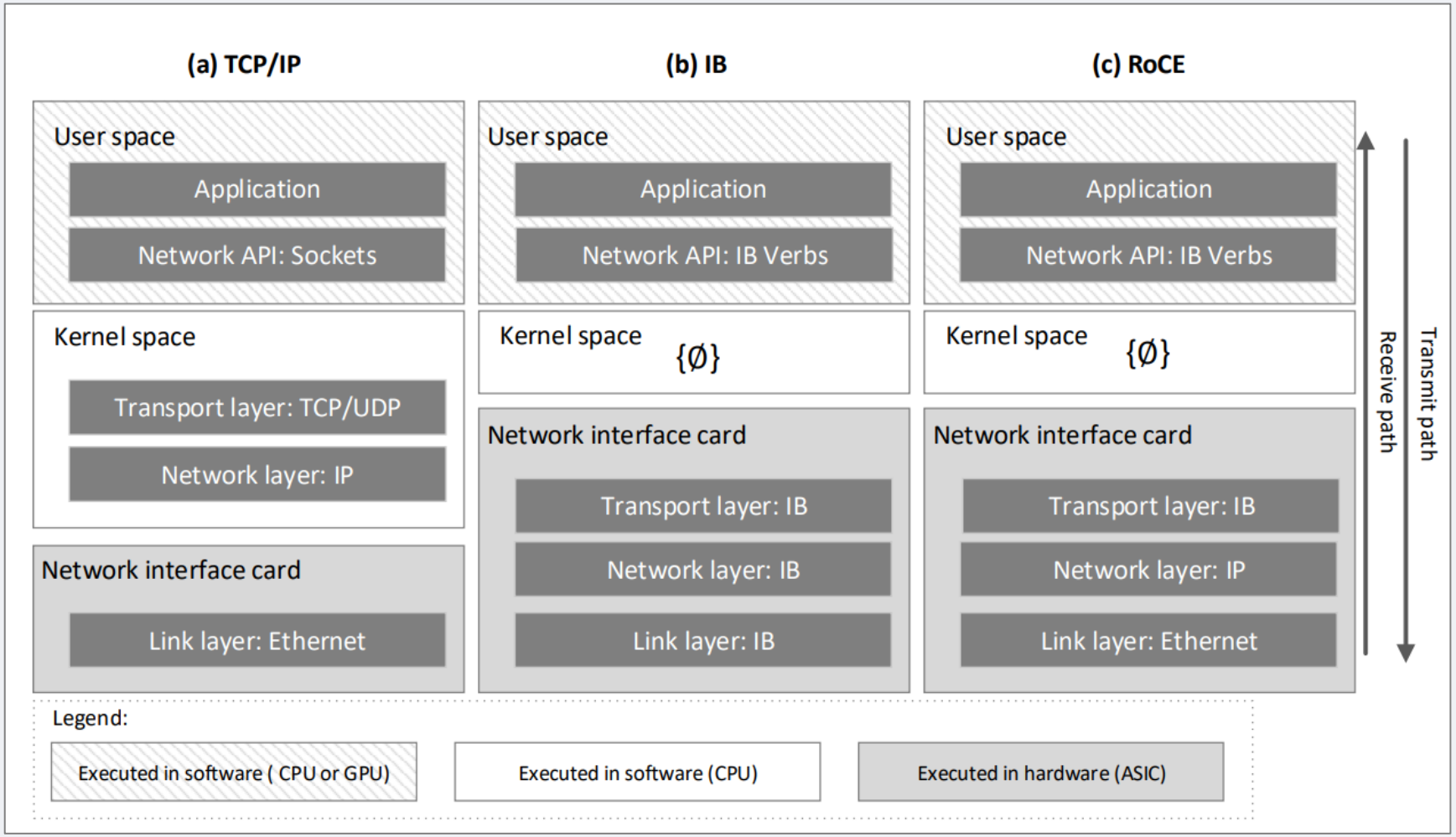


In our production clusters:

>1B models do not scale with 2 nodes using TCP

Training LLMs at Scale: Communication

Transport and Network layers offloaded to HW with RDMA: eliminates data copies and CPU intervention



Benchmark Methodology: Configuration Space and Performance Metrics

Training requires tuning, evaluating performance for multiple Configuration Spaces

Dimensions explored in our benchmarking framework:

- Network Stack: TCP / RDMA
- Batch size
- Parallelism Strategy: Data parallelism and sharding (ZeRO stages 1, 2 and 3)
- Quantization technique: Weight and gradient quantization
- Activation recomputation

Metrics used to evaluate performance:

- **Training throughput** in samples per sec: **$(b \cdot s) / T$**
- **Scaling factor**: ***tpg.M / tpg.m***, where *tpg.M* denotes the throughput per GPU for the larger configuration, and *tpg.m* corresponds to the smaller configuration, respectively.

Platform and Model Setup

Platforms

- **IB Cluster**: IB interconnect, 16 nodes (8x NVIDIA 80GB A100 each). Folded-Clos 1.6 Tbps.
- **RoCE Cluster**: RoCE interconnect, 64 nodes (8x NVIDIA 80GB H100 each). Folded-Clos 1.6 Tbps.

Models

From small to mid-sized architectures.

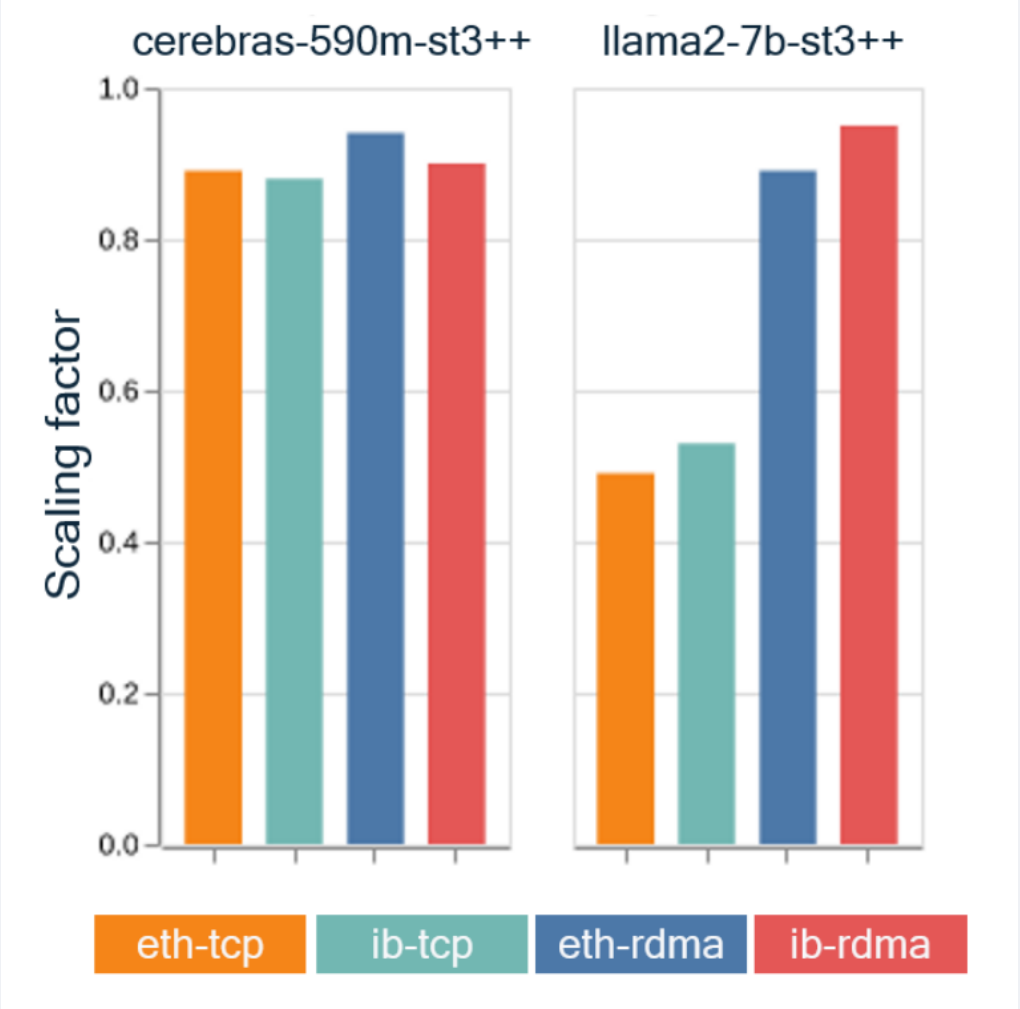
Models evaluated:

- **opt-125m**
- **cerebras-590m**
- **opt-1.3b**
- **cerebras-2.7b**
- **llama2-7b**
- **opt-30b**
- **llama2-70b**

Benchmark: On the Effect of RDMA

RDMA is able to achieve close to linear scaling, TCP is not.

- From 1 to 2 nodes:

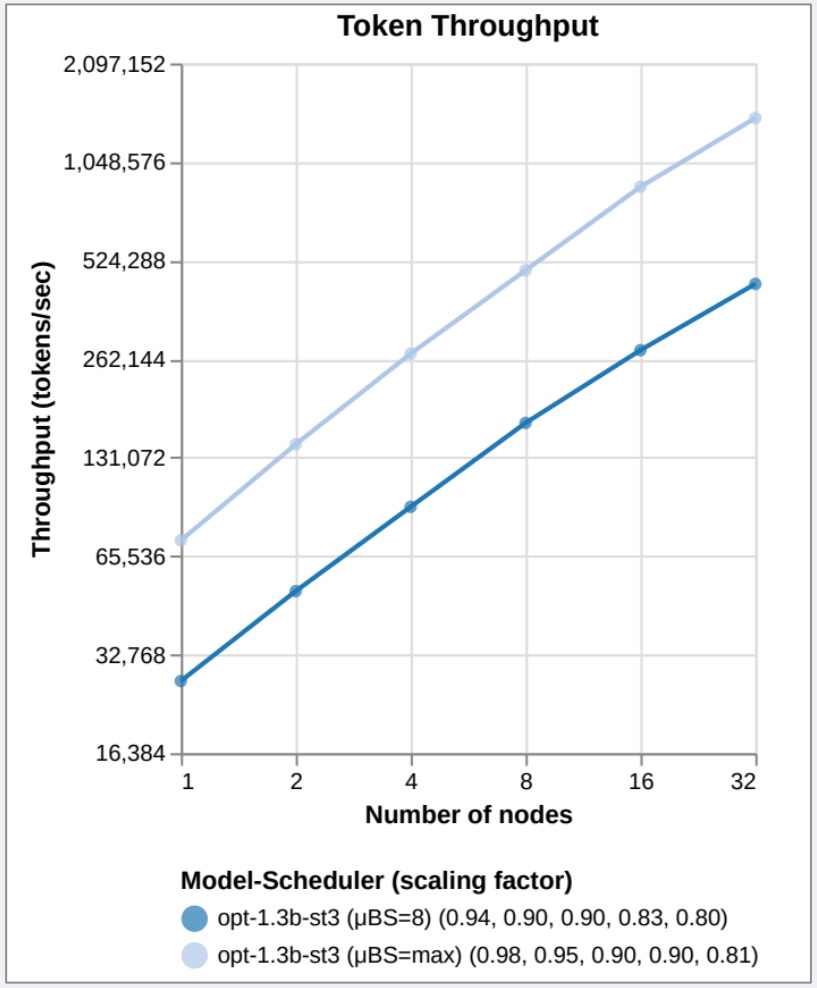


Tested for both:

- Commodity Ethernet (eth)
- InfiniBand (ib)

Benchmark: On the Effect of Batch Size

By maximizing the batch size, training throughput can be significantly increased



ZeRO reduces memory footprint by adding communication

- Reducing memory enables larger batch sizes

# Nodes	1	2	4	8	16	32	48
Max-BS	208	432	864	1664	3456	6912	10752
LS-BS	208	416	832	1664	3328	6656	9984
Z-gain (%)	0.0	3.8	3.8	0.0	3.8	3.8	7.7
Max-μBS	26.0	27.0	27.0	26.0	27.0	27.0	28.0

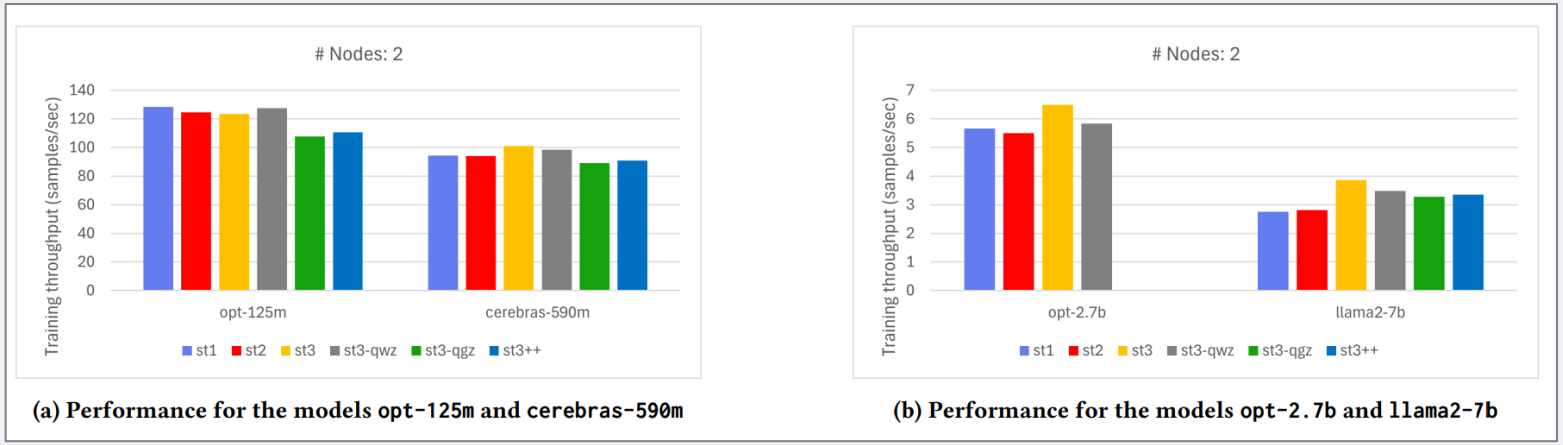
Max-BS: Maximum batch size achieved
LS-BS: Expected batch size assuming linear scaling
Z-gain %: Gain obtained from ZeRO memory reduction
Max-μBS: Maximum micro-batch size (per GPU)

Benchmark: On the Effect of Sharding and Quantization

Selecting the best strategy depends on several factors, tailored to model size.

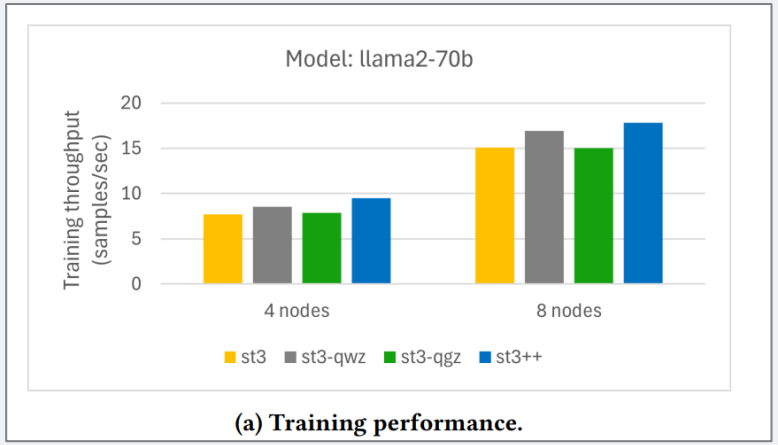
These optimizations are relevant for improving memory/communication overhead.

On the IB Cluster:



(a) Performance for the models `opt-125m` and `cerebras-590m`

(b) Performance for the models `opt-2.7b` and `llama2-7b`

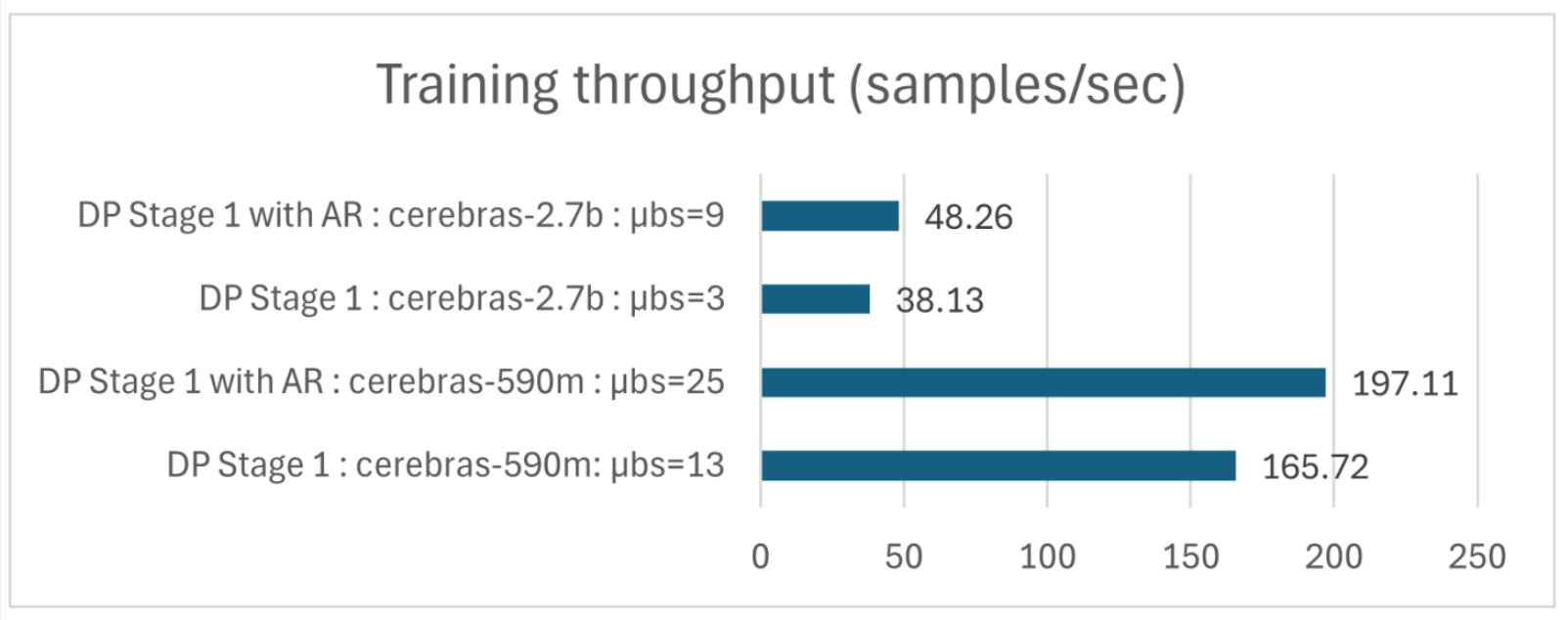


(a) Training performance.

Benchmark: On the Effect of Activation Recomputation

Activation recomputation allows to increase the batch size, yielding higher training throughput

- On the IB Cluster:

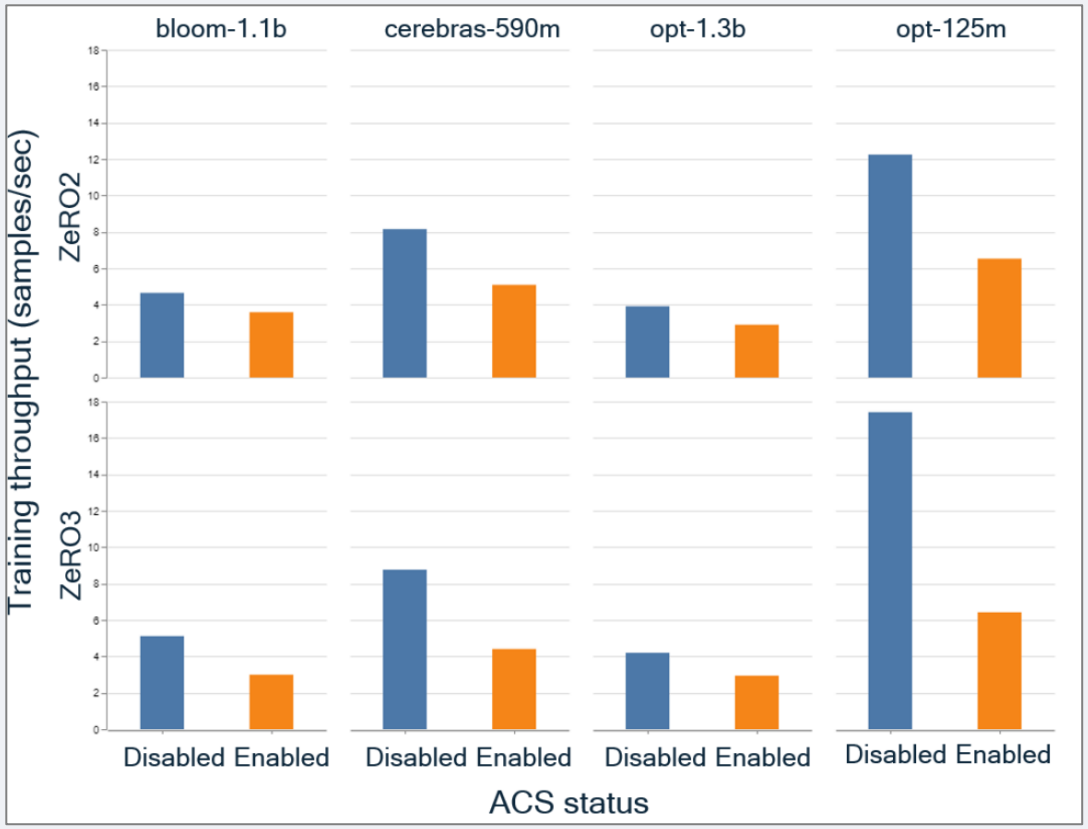


Benchmark: On the Effect of Direct GPU-to-GPU Communication

Warning! Enabling Access Control Services (ACS) leads to poor performance.

ACS adds a security layer in multi-tenant platforms, mitigating unauthorized P2P communication: a performance killer in HPC platforms!

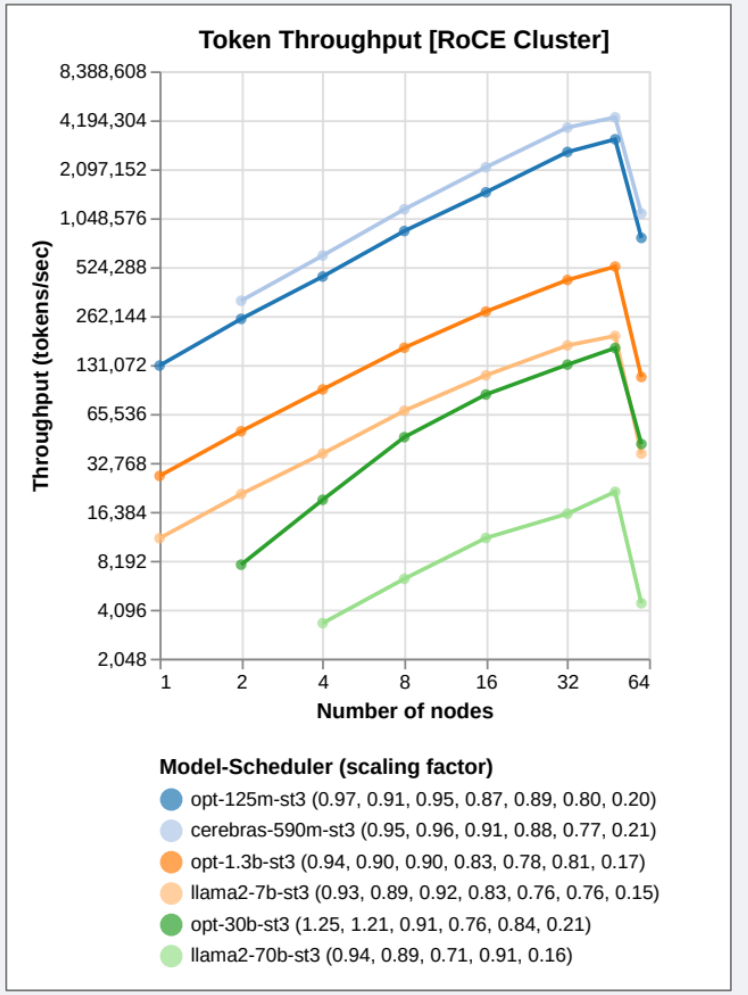
- On the IB Cluster:



Benchmark: On the Effect of Straggler Nodes

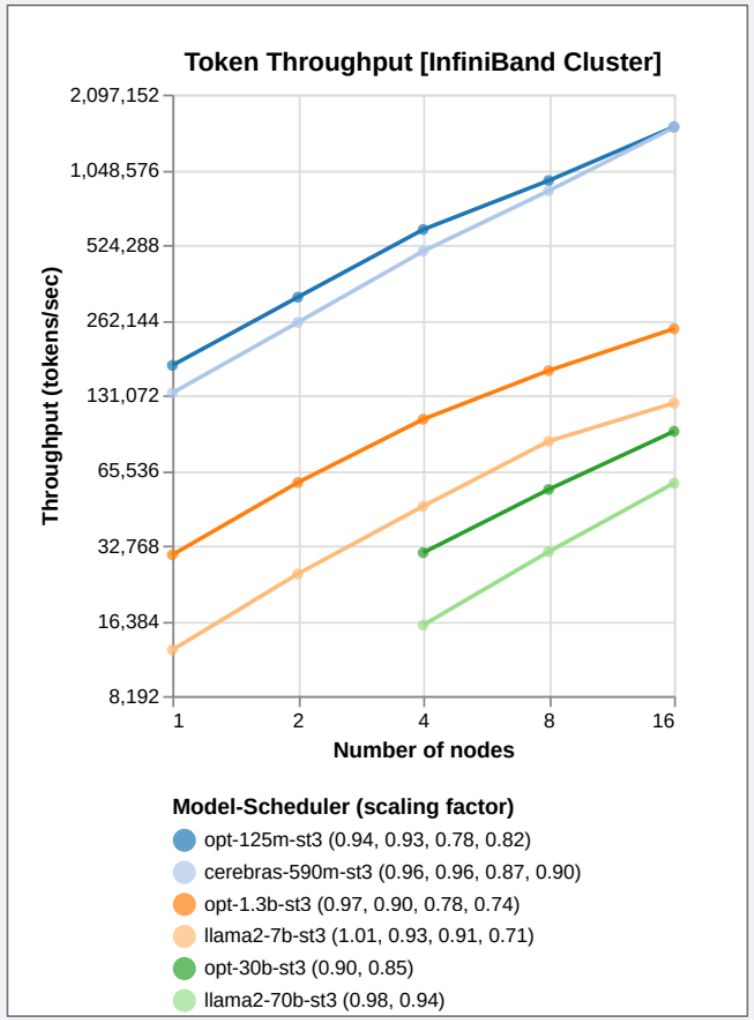
Malfunctioning of a single GPU can lead to significant performance degradation in the cluster (aka *straggler effect*).

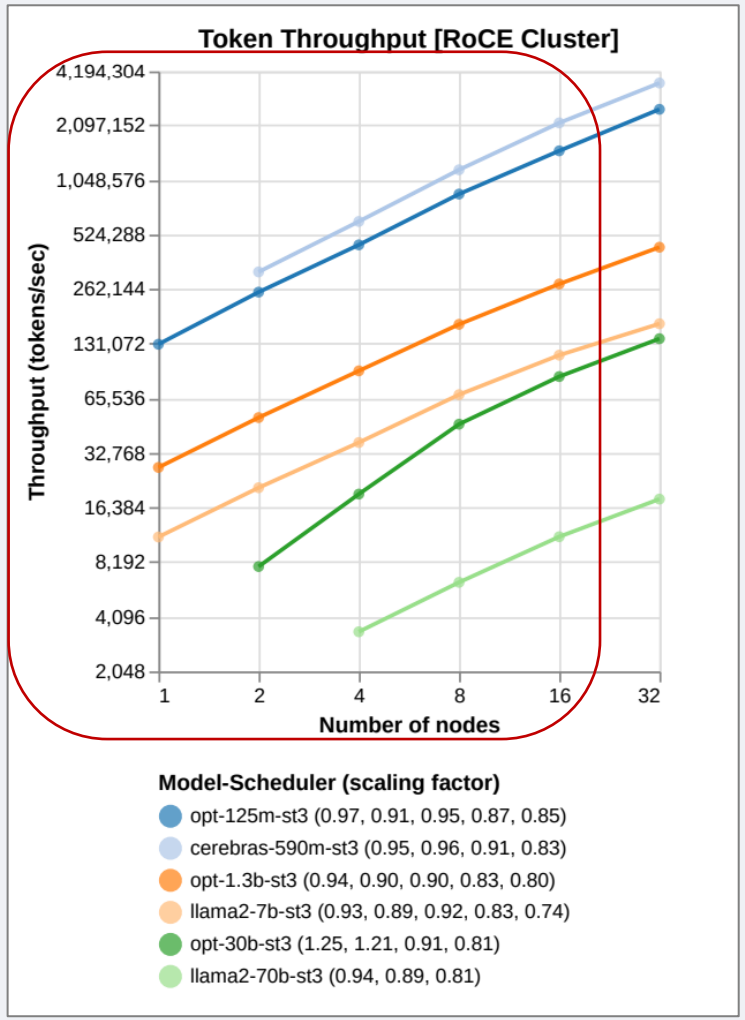
- The overall performance is bottlenecked by the slowest worker.
- Traditional health indicators and monitoring tools didn't detect abnormalities.



On Achieving Linear Scaling for Multi-Node Training

RoCE: All scaling factors exceeding *0.8* with ZeRo St3

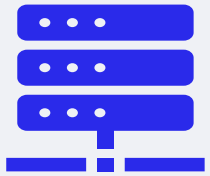




Conclusions

RoCE can deliver scaling performance comparable to InfiniBand for LLM training.

Comprehensive benchmarking of large-scale LLM training using RDMA over RoCE and InfiniBand.

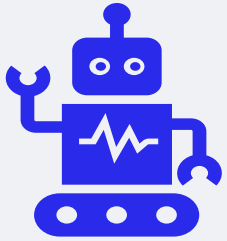


RoCE-based clusters on commodity Ethernet can achieve near-linear scaling, comparable to InfiniBand.

RDMA is critical for efficient multi-node training.

Performance depends on batch size tuning, sharding strategies, and system-level settings (e.g., GPU-to-GPU communication).

Identification of straggler nodes to maintain linear scalability.



Insights provide guidance for cost-effective, production-grade AI training infrastructure.

Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated.
Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patents are licensed by Qualcomm Incorporated.

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog