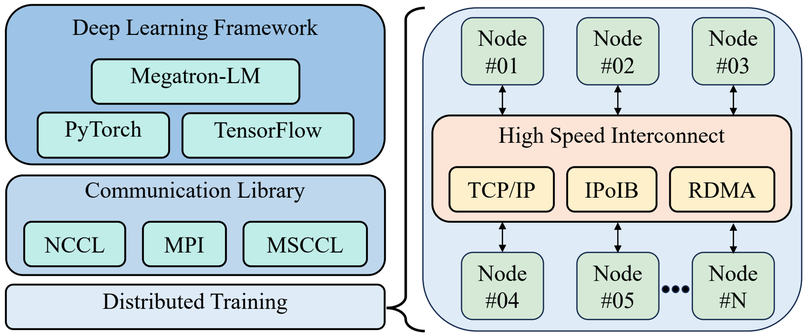
Early Experience in Characterizing Training Large Language Models on Modern HPC Clusters

Hao Qi, Liuyao Dai, Weicong Chen, and Xiaoyi Lu
University of California, Merced

ABSTRACT

- Large Language Models (LLMs) are powerful tools in natural language processing, aiding in tasks like language translation, text generation, and sentiment analysis.
- The immense size and complexity of LLMs pose significant challenges.
- This study explores high-performance interconnects in the distributed training of various LLMs.
- High-performance network protocols, especially RDMA, are found to significantly outperform IPoIB and TCP/IP in training performance.
  - RDMA offers improvements by factors of 2.51x and 4.79x respectively.
- Despite the significant findings, there is potential for further optimization in overall interconnect utilization.
- The research provides deeper insights into the performance characteristics of LLMs over high-speed interconnects and paves the way for developing more efficient training methodologies for LLMs.

INTRODUCTION

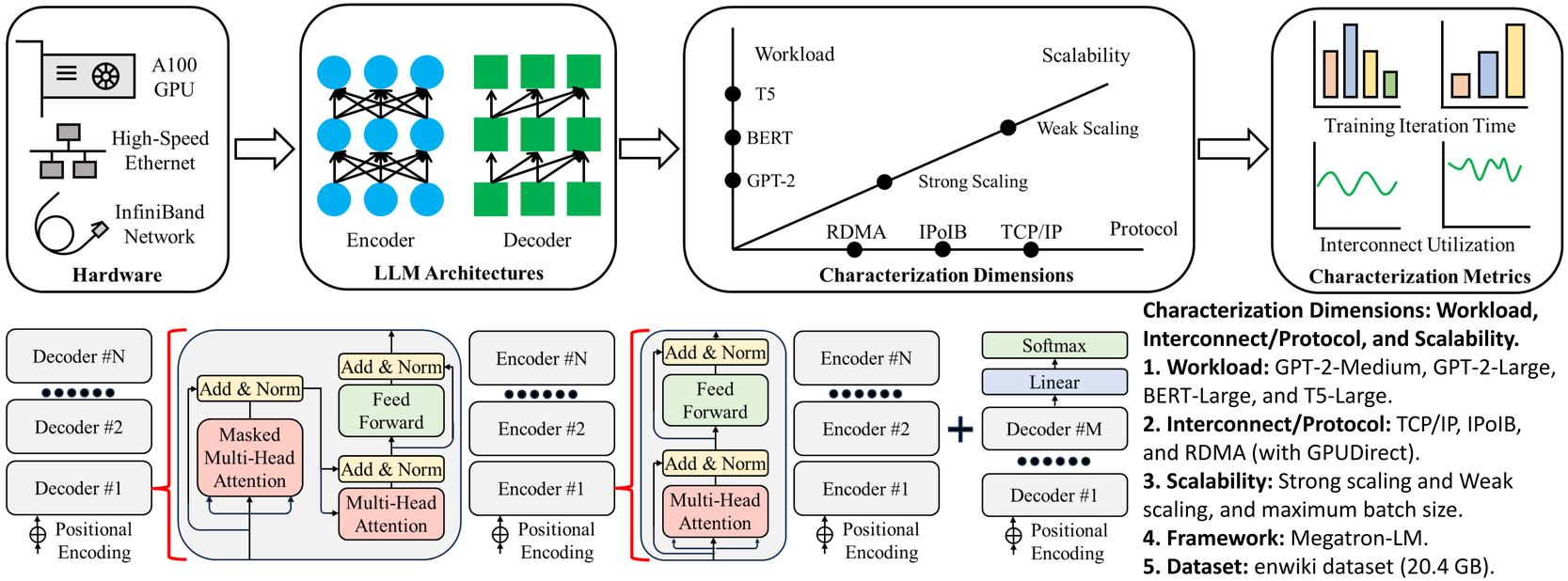


Challenges in Distributed Training for LLMs:
- Need for communication & coordination among nodes and GPUs.
- Vast amount of training data.
- Requirement for GPU-enabled distributed training.
- Emphasized need for high-performance interconnects.

Role of High-speed Interconnects:
- Ensure efficient data transfer and synchronization.
- Crucial for rapid and scalable communication among nodes and GPUs.

Research Questions:
- Potential bottlenecks in communication and the time proportion for communication in various LLM training configurations.
- Efficiency of current high-performance interconnects in different distributed training scenarios.
- Quantitative performance impact of networking technologies & protocols (e.g., RDMA, IPoIB, TCP/IP) on LLM training.

METHODOLOGY



Characterization Dimensions: Workload, Interconnect/Protocol, and Scalability.

1. Workload: GPT-2-Medium, GPT-2-Large, BERT-Large, and T5-Large.

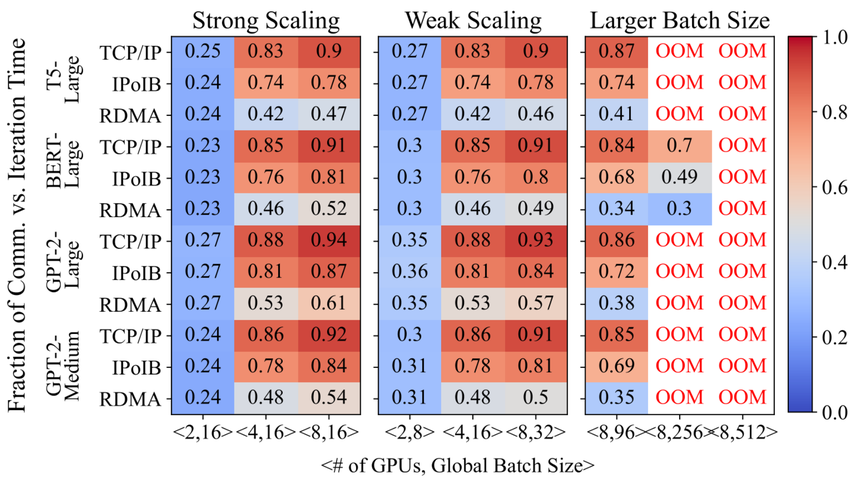
2. Interconnect/Protocol: TCP/IP, IPoIB, and RDMA (with GPUDirect).

3. Scalability: Strong scaling and Weak scaling, and maximum batch size.

4. Framework: Megatron-LM.

5. Dataset: enwiki dataset (20.4 GB).

EVALUATION



RDMA-100 Gbps outperforms IPoIB-100 Gbps and TCP/IP-10 Gbps by an average of 2.51x and 4.79x regarding training iteration time, and scores the highest interconnect utilization (up to 60 Gbps) in both strong and weak scaling, compared to IPoIB with up to 20 Gbps and TCP/IP with up to 9 Gbps, leading to the shortest training time.

CONCLUSION

Main Contribution:
- Enhanced understanding of performance dynamics of LLMs over high-speed interconnects.

Research Efforts:
- Detailed exploration of communication's role in distributed LLM training.
- Findings can guide the design and setup of systems for growing LLM application demands.

Evaluation Outcomes:
- Both strong scaling and weak scaling experiments show similar patterns.
- Highlighted impact of interconnect/protocol on distributed training.
- Emphasized the importance of efficient interconnect utilization.

Directions for Future Work:
- Explore alternative parallelism strategies, e.g., model parallelism.
- Study behavior of even larger models at greater scales.
- Develop methods to further enhance interconnect utilization.

ACKNOWLEDGEMENT

This work was supported in part by an NSF research grant OAC #2321123, a DOE research grant DE-SC0024207, and an Amazon Research Award.Part of this research was conducted using Pinnacles (NSF MRI, #2019144) at the Cyberinfrastructure and Research Technologies (CIRT) at the University of California, Merced.