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

Hao Qi, Liuyao Dai, Weicong Chen, Xiaoyi Lu
{hqi6,ldai8,wchen97,xiaoyi.lu}@ucmerced.edu
University of California, Merced
Merced, California, USA

ABSTRACT

In the realm of natural language processing, Large Language Models (LLMs) have emerged as powerful tools for tasks such as language translation, text generation, and sentiment analysis. However, the immense parameter size and complexity of LLMs present significant challenges. This work delves into the exploration and characterization of high-performance interconnects in the distributed training of various LLMs. Our findings reveal that high-performance network protocols, notably RDMA, significantly outperform other protocols like IPoIB and TCP/IP in training performance, offering improvements by factors of 2.51x and 4.79x respectively. Additionally, we observe that LLMs with larger parameters tend to demand higher interconnect utilization. Despite these findings, our study suggests potential for further optimization in overall interconnect utilization. This research contributes to a deeper understanding of the performance characteristics of LLMs over high-speed interconnects, paving the way for more efficient training methodologies.

1 INTRODUCTION

Generative Artificial Intelligence (AI) has recently become a hot topic. Transformer [10]-based large language foundation models, such as BERT[5] and GPT series [2, 4, 8], have emerged as potent tools for a variety of natural language processing (NLP) tasks. These tasks encompass language translation, text generation, and sentiment analysis. These models possess an extraordinary ability to comprehend and generate text that mirrors human language, making them indispensable across various sectors, like healthcare [6], finance [3], and marketing [9].

However, the development of effective Large Language Models (LLMs) is fraught with challenges, primarily due to their inherent complexity and the enormity of their parameter size, which often extends to millions, billions, or even trillions. The training of LLMs necessitates significant computational power and memory capacity to handle the extensive model weights. Consequently, LLM training is resource-intensive, exerting considerable pressure on the underlying infrastructure. To mitigate resource constraints and expedite the training process, distributed training is employed. This approach involves the division of the model and/or training data across multiple compute nodes, typically equipped with GPUs and high-speed interconnects. Distributed training for LLMs presents its own set of challenges, particularly concerning communication and coordination among nodes and GPUs. The substantial volume of training data and the requirement for distributed GPU-enabled training of LLMs further amplify the need for high-performance interconnects. High-speed interconnects facilitate efficient data transfer and synchronization during LLM training, which is essential for achieving rapid and scalable communication among nodes and GPUs, minimizing communication overhead, and optimizing overall system performance. The computing capability of modern HPC systems has scaled at a rate that more than doubles the pace of interconnect bandwidth across generations. This trend presents many potential research problems for achieving efficient and scalable LLM training.

In this work, we aim to investigate the following questions: 1. Will high-speed interconnects become a bottleneck for communication, and what proportion of the training process is occupied by communication for various types and configurations of LLMs? 2. Are the current high-performance interconnects utilized well during different distributed training scenarios? 3. What kind of quantitative performance impact will different networking technologies and protocols (such as RDMA, IPoIB, TCP/IP) have on various LLMs training?

2 CHARACTERIZATION METHODOLOGY

Our characterization focuses on three pivotal dimensions: workload, interconnect/protocol, and scalability, as shown in Figure 1.

Workloads: We consider some popular open-source LLMs, including GPT-2-Medium, GPT-2-Large, BERT-Large, and T5-Large, representing a range of model sizes, architectures, and application domains. These models have been widely adopted in various NLP tasks and exhibit different computational requirements.

Interconnect/Protocol: We consider different interconnect technologies such as TCP/IP, IPoIB, and RDMA (with GPUDirect). By exploring the influence of these interconnect options on LLMs, we can understand how they affect communication patterns, data transfer rates, and overall performance.

Scalability: We evaluate both strong scaling and weak scaling aspects. By examining the scalability of LLMs across different interconnect technologies under the data parallelism training architecture, we can identify potential bottlenecks, scalability limits, and the overall efficiency of the distributed training process.

Framework: As for framework, we leverage the Megatron-LM [7] for our characterization methodology as our primary distributed training framework. It provides efficient and scalable implementations of distributed training algorithms, making it an ideal choice for our investigation.

SC'23, November 12–17, 2023, Denver, CO

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

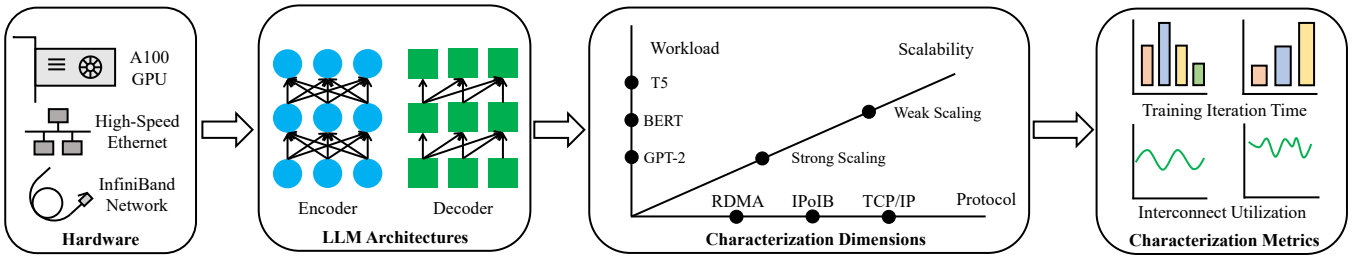

Figure 1: Characterization Methodology.

Dataset: In our methodology, we utilize the enwiki [1] dataset as a representative example of a large-scale dataset. The enwiki dataset (20.4 GB) is derived from English Wikipedia and contains vast text documents spanning diverse topics and genres.

3 PERFORMANCE EVALUATION

We used NSF-funded Pinnacles cluster at UC Merced that is equipped with 8 GPU nodes. Each node has two Intel 28-Core Xeon Gold 6330 CPUs (2.0GHz), 256GB DRAM, 2x NVIDIA Tesla A100 40GB GPUs with PCIe, and interconnected via 100 Gbps EDR InfiniBand with RDMA and 10 Gbps Ethernet. We use up to 4 GPU nodes in the evaluation. All used software for four models includes CUDA 11.8.0, PyTorch 2.0.0, NCCL 2.14.3, NVIDIA Apex 22.03, and Megatron-LM v3.0.2. We use data parallelism to emphasize the influence of interconnects on experiments in this section. We use FP16 precision training and set global batch size = 16 for strong scaling and micro batch size = 4 for weak scaling. The number of GPUs and batch size have such a relation: #GPU × micro batch size = global batch size.

For data parallelism, the communication burden is mainly from the backward parameter synchronization (i.e., AllReduce by NCCL). Our key observations are the following: 1. In strong scaling, AllReduce time takes up most training time in each iteration, with 53.4%, 82.48%, and 91.72% for RDMA, IPoIB, and TCP/IP, respectively. 2. In weak scaling, AllReduce time takes up to 50.5%, 80.78%, and 91.12% of iteration time for RDMA, IPoIB, and TCP/IP. 3. RDMA-100 Gbps outperforms IPoIB-100 Gbps and TCP/IP-10 Gbps by an average of 2.51x and 4.79x regarding the 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. Besides, we observe that GPT-2-Large consistently achieves higher RX and TX speeds (30.47 Gbps) within the models tested than other models, like GPT-2-Medium (19.93 Gbps), BERT-Large (26.48 Gbps), and T5-Large (24.19 Gbps).

As illustrated in Figure 2, we also evaluate these models with increased batch sizes until out-of-memory (OOM) and observe a similar trend where communication takes a large portion of the iteration time. Notably, this figure demonstrates that even though increasing the batch sizes can result in a reduced proportion of communication time in the overall iteration time (amortized by the prolonged computation time), the communication time can still occupy at least 34% of iteration time except for BERT-Large, as it allows for a larger batch size. But even in this case, it still requires 30% of iteration time for communication. This observation

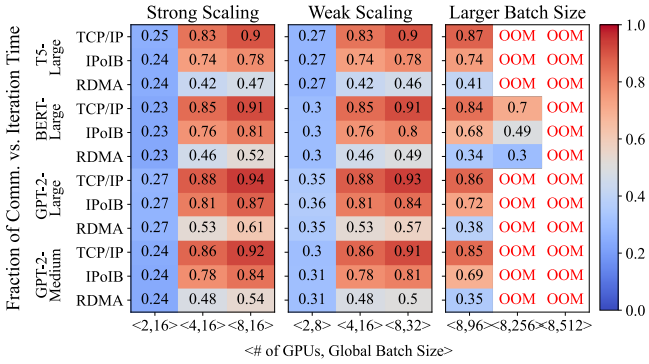

Figure 2: Fraction of Comm. vs. Iteration Time with Larger Batch Sizes

highlights a lower bound of the communication time proportion since it investigates until the maximum batch size a model can train with at the given scale.

4 CONCLUSION

Overall, this work makes a contribution to the understanding of the performance dynamics of large language models over high-speed interconnects. We have conducted a thorough exploration and summary of the crucial role that communication plays in the distributed training of LLM. The results can inform the design and deployment of efficient systems to support the growing demand for LLM applications. Our evaluation outcomes reveal that both strong scaling and weak scaling experiments display similar patterns, thus underlining the impact of the interconnect/protocol on distributed training and the necessity of efficient interconnect utilization. Some of the future work may encompass investigating alternative parallelism strategies like model parallelism, delving into the behavior of even larger models at larger scales, and developing methods to further optimize interconnect utilization.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] [n. d.]. English Wikipedia Dump. https://dumps.wikimedia.org/enwiki/20230501/.

[2] [n. d.]. GPT-4 Technical Report. https://openai.com/research/gpt-4.

[3] Dogu Araci. 2019. Finbert: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does CHATGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education* 9, 1 (2023), e45312.

[7] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. 2021. Efficient Large-scale Language Model Training on GPU Clusters Using Megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–15.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9.

[9] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. 2019. BERT for Stock Market Sentiment Analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 1597–1601.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in neural information processing systems* 30 (2017).