Committee: Kai Li (advisor), Tri Dao, & Ravi Netravli

Date: May 14 at 10am (room CS 302, pending)

Zoom link: https://princeton.zoom.us/j/8780770538

—-

Title: **Laying Down the Sequence Train(ing) Tracks: Minimizing Communication and Reducing HBM Footprint via Local Computations and Orchestrated Dataflow**

Abstract:

The remarkable sequence prediction abilities of current AI systems has resulted in a flurry of world-changing applications. However, pretraining LLMs requires billions of example sequences and thus demands a massive amount of raw computational power (FLOPs), necessitating distributed computations. Achieving high compute utilization across such a large quantity of processors has proven to be challenging due to the communication volume. In addition to the very large scale, data-dominated settings, many training scenarios (post-training, fine-tuning) require significantly less FLOPs but still operate over large models which consume significant fractions of expensive on-device memory – in these settings distributed processing is commonly commissioned simply to satisfy the memory needs.

I will present a scheme that maintains near peak computational throughput even in highly constrained communication or on-device memory environments. The key insights are (1) constructing a model pipeline via cyclic sharding of layers, forming a ring-track for data to flow along (2) employing “Temporal” parallelism wherein chunks form a sequential train moving along the pipeline (3) continual prefetching/checkpointing in/out of device memory by taking advantage of cheaper, larger capacity local system DRAM. This division of labor among workers and tasks opens doors for AI to process richer information sources (longer sequences) at full compute utilization.

References

Relevant Textbook:

[1] David B. Kirk and Wen-mei W. Hwu. 2010. Programming Massively Parallel Processors: A Hands-on Approach (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Historical Background:

[2] F. J. Corbató and V. A. Vyssotsky. 1965. Introduction and Overview of the Multics System. In Proceedings of the November 30-December 1, 1965, Fall Joint Computer Conference, Part I (AFIPS '65 Fall, part I). Association for Computing Machinery, New York, NY, USA, 185-196.

[3] J. B. Dennis. 1974. First Version of a Data Flow Procedure Language. In Programming Symposium, Proceedings Colloque Sur La Programmation. Springer-Verlag, Berlin, Heidelberg, 362-376.

[4] T. Kilburn, D. B. G. Edwards, M. J. Lanigan, and F. H. Sumner. 1962. One-Level Storage System. IRE Transactions on Electronic Computers EC-11, 2 (1962), 223-235.

[5] U.J. Kapasi, W.J. Dally, S. Rixner, J.D. Owens, and B. Khailany. 2002. The Imagine Stream Processor. In Proceedings of the IEEE International Conference on Computer Design: VLSI in Computers and Processors (ICCD). 282-288.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000-6010.

ML Systems for Training (Emphasis on Long-Context):

[7] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL]. https://arxiv.org/abs/2412.19437

[8] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Curran Associates, Inc.

[9] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Reza Yazdani Aminadabi, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2024. System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models. In Proceedings of the 43rd ACM Symposium on Principles of Distributed Computing (PODC '24). Association for Computing Machinery, New York, NY, USA, 121-130.

[10] Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context. arXiv: 2310.01889 [cs.CL]. https://arxiv.org/abs/2310.01889

[11] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. 2019. PipeDream: Generalized Pipeline Parallelism for DNN Training. In Proceedings of the 27th ACM Symposium on Operating Systems Principles (SOSP '19). Association for Computing Machinery, New York, NY, USA, 1-15.

[12] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 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 (SC 21). Association for Computing Machinery, New York, NY, USA, Article 58.

[13] Jinghan Yao, Sam Ade Jacobs, Masahiro Tanaka, Olatunji Ruwase, Aamir Shafi, Hari Subramoni, and Dhabaleswar K. Panda. 2024. Training Ultra Long Context Language Model with Fully Pipelined Distributed Transformer. arXiv: 2408.16978 [cs.DC]. https://arxiv.org/abs/2408.16978

[14] Pinxue Zhao, Hailin Zhang, Fangcheng Fu, Xiaonan Nie, Qibin Liu, Fang Yang, Yuanbo Peng, Dian Jiao, Shuaipeng Li, Jinbao Xue, Yangyu Tao, and Bin Cui. 2025. MEMO: Fine-grained Tensor Management For Ultra-long Context LLM Training. Proceedings of the ACM on Management of Data (Proc. ACM Manag. Data) 3, 1, Article 53 (2025).

[15] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. Proceedings of the VLDB Endowment (Proc. VLDB Endow.) 16, 12 (2023), 3848-3860.

Relevant For Future Directions:

[16] Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Dan Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, Brennan Saeta, Parker Schuh, Ryan Sepassi, Laurent El Shafey, Chandramohan A. Thekkath, and Yonghui Wu. 2022. Pathways: Asynchronous Distributed Dataflow for ML, In Proceedings of the Fifth Conference on Machine Learning and Systems (MLSys 2022). mlsys.org, Santa Clara, CA, USA.

[17] NVIDIA. 2025. Cosmos World Foundation Model Platform for Physical AI. arXiv:2501.03575 [cs.CV]. https://arxiv.org/abs/2501.03575

[18] Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Y. X. Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. 2025. Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. arXiv:2502.11089 [cs.CL]. https://arxiv.org/abs/2502.11089