ARIA
APPLIED RESEARCH IN ACTION

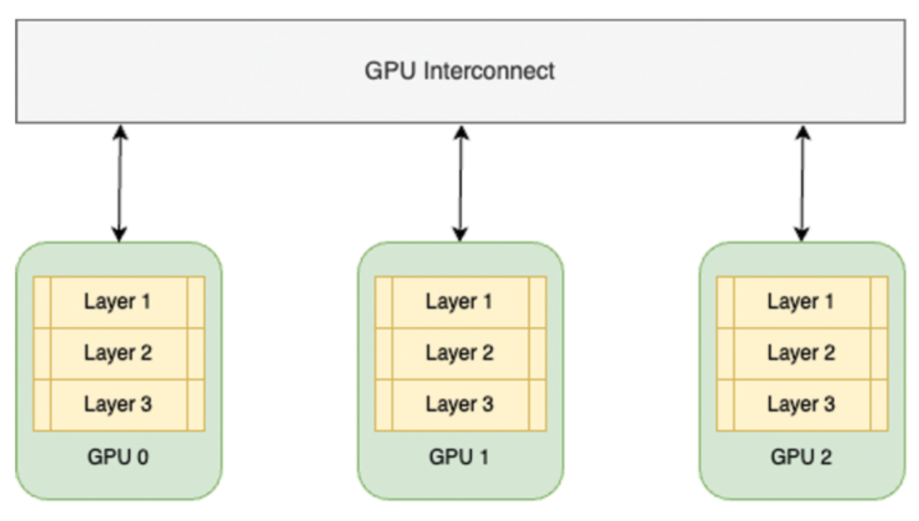
Efficiently Training Large Language Models with Advanced Distributed Training Techniques

Unlocking Peak Efficiency: Scaling Large Language Model Training with DeepSpeed's ZeRO Optimization Strategies and Vertex AI's Distributed Training.

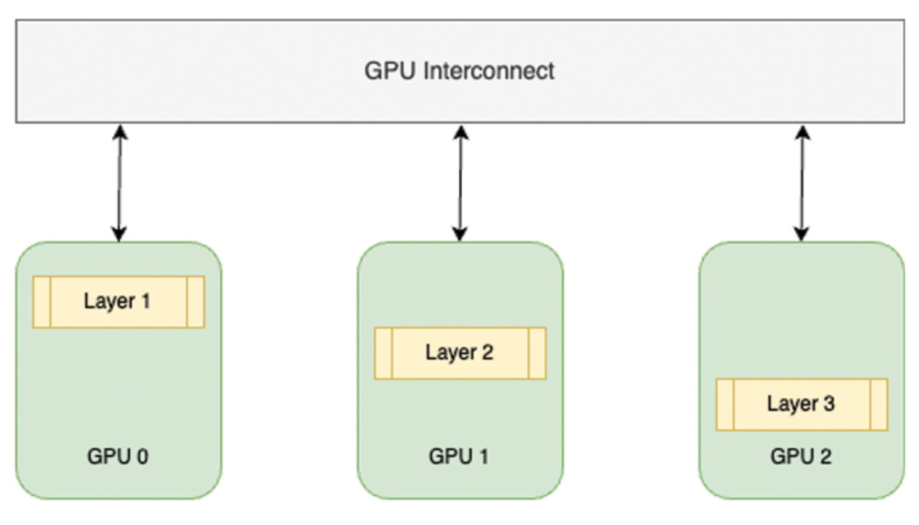
Tirth Patel

Nick Koudas
ACADEMIC SUPERVISOR

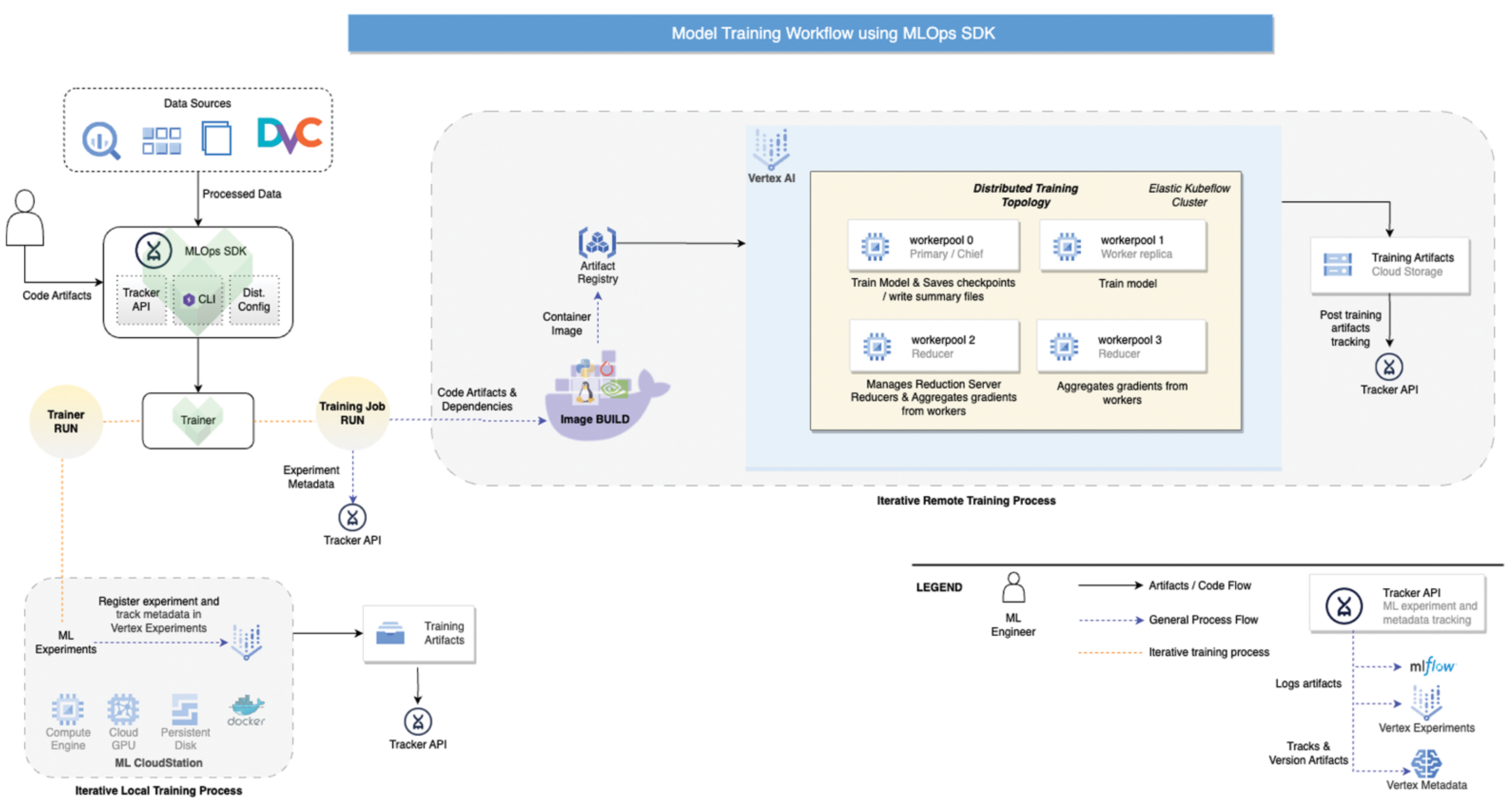
Ramtin Rassoli
INDUSTRY SUPERVISOR



Distributed Data Parallel Training



ZeRO Training



PROJECT SUMMARY

In today's highly competitive business landscape, integrating Large Language Models (LLMs) into products is crucial. While open-source LLM APIs like ChatGPT are accessible, they can be costly. For businesses prioritizing data privacy, proprietary data, and its use-cases, developing an in-house Large Language Model Operations (LLMOps) solution becomes necessary.

Our methodology outlines an approach to cost-effective LLM training using Google Cloud Platform and the Vertex AI service. We employ Distributed Data Parallel (DDP) techniques, including standard DDP for smaller models and Zero Redundancy Optimizer (ZeRO) for larger ones. ZeRO optimally shards model optimizer states, gradients, and parameters across multiple devices, improving memory efficiency. Moreover, it can also offload the model states and computation from the GPU to the host CPU, resulting in memory-efficient training. Additionally, we explore the LoRA technique for quantizing and fine-tuning larger LLMs on a single GPU.

Notably, we successfully trained a ~3 billion-parameter model (GPT-2 architecture) using 8 Nvidia Tesla V100 GPUs (16GB) and a general-purpose n1-standard-64 machine type. Establishing an in-house distributed system with tools like DeepSpeed, Accelerate, or Megatron-LM has the potential to transform LLM training and fine-tuning. This transformation could provide businesses with a competitive advantage, enabling them to harness the full potential of LLMs while optimizing resources and staying agile in a rapidly evolving market.

REFERENCES

1. Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). ZeRO: Memory optimizations Toward Training Trillion Parameter Models. SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, 1–16. https://doi.org/10.1109/SC41405.2020.00024
2. Rajbhandari, S., Ruwase, O., Rasley, J., Smith, S., & He, Y. (2021). ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning. https://doi.org/10.48550/arxiv.2104.07857
3. Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., & Chintala, S. (2020). PyTorch Distributed: Experiences on Accelerating Data Parallel Training. https://doi.org/10.48550/arxiv.2006.15704
4. Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.-C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., Desmaison, A., Balioglu, C., Damania, P., Nguyen, B., Chauhan, G., Hao, Y., Mathews, A., & Li, S. (2023). PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. https://doi.org/10.48550/arxiv.2304.11277

BenchSci

Computer Science
UNIVERSITY OF TORONTO

Master of Science in
Applied Computing