Heyang Qin

Senior Researcher (L64)
Microsoft

Telephone: (775) 376-3809
Email: heyangqin@microsoft.com
https://heyangqin.github.io

Experience

Senior Researcher 06/2024 - Present
Azure OpenAI, Microsoft Corporation

Researcher 01/2023 - 05/2024
DeepSpeed Team, Microsoft Corporation

Researcher Intern 05/2021 - 12/2022
DeepSpeed Team, Microsoft Corporation

Ph.D. in Computer Science 08/2017 - 12/2022
University of Nevada, Reno (Advisor: Dr. Feng Yan and Dr. Lei Yang)
First author publications in NeurIPS, ICLR, SC, etc.

Engineering Projects

1. Azure OpenAI: EngineV2 / V3, Model Support (Current)
    - Working on the OpenAI inference stack, extending it for Azure deployment; with the focus on model support including CLIP, GPT4.1-mini, GPT5.2 on EngineV3.
    - Implemented EngineV2 resharding on AMD GPUs, fixing attention head partitioning, embedding dimension mismatches, and cross-rank sync issues for stable multi-GPU/multi-node execution.
    - Added CLIP model support on AMD EngineV3, fixing sharding implementation and resolving accuracy issues in embedding-scatter kernel.
    - Deployed GPT-4.1-mini with disaggregated prefill–decode pipelines, directly used in GitHub Copilot and impacting millions of users worldwide.
2. Efficient Model Serving for MS Internal Foundation Models (Phi, MAI)
    - Added multi-LoRA support for Phi into vLLM to meet Azure business needs
    - Implemented MAI model support from scratch on SGLang, delivering up to 50% reduction in time-to-first-token (TTFT) and improved parallelism support
    - Optimized execution graphs and parallel strategies to improve latency–throughput trade-offs in real-world serving workloads
3. DeepSpeed-FastGen / DeepSpeed-ZeRO3
    - Core contributor to DeepSpeed ZeRO3 and FastGen, focusing on scalable, reliable training and inference systems
    - Designed a custom memory allocator and leveraged CUDA Graph execution for FastGen, achieving 47% end-to-end inference speedup compared to baseline

- Advanced quantized/mixed-precision communication (INT4, HPZ) in ZeRO3, improving efficiency without accuracy loss
- Enhanced correctness and robustness at scale by fixing race conditions, failure modes, and distributed parameter coordination

Research Projects

1. DeepSpeed-ZeRO++/MixZ++: Extremely Efficient Collective Communication for Large Model Training

   - Designed block-quantized all-gather, data remapping, and quantized all-to-all gradient averaging, preserving accuracy with low-precision communication.
   - Reduced ZeRO communication volume by 4×, achieving 2.16× throughput improvement at 384 GPUs; LinkedIn adopted ZeRO++ in DragonKnight, reporting 2.4× end-to-end speedup in production.

2. DeepSpeed-SimiGrad: Fine-Grained Adaptive Batching for Large Scale Training using Gradient Similarity Measurement

   - Developed fully automated adaptive batching adjusting mini-batch sizes dynamically based on gradient similarity metrics.
   - Achieved state-of-the-art batch size of 78k for BERT-Large pretraining (SQuAD F1 = 90.69), surpassing previous 59k batch baseline; lightweight, efficient, and compatible with ZeRO-based optimizations.

3. Region Based Reinforcement Learning Scheduling Framework for Model Inference

   - Achieves faster convergence than standard RL algorithms, supported by mathematical proof of regional decomposition and policy update rules.
   - Developed general region-based RL framework for model-parallel scheduling; partitions models into regions and learns near-optimal policies, improving latency, throughput, and resource utilization.

Selected Publication

Heyang Qin*, Guanhua Wang*, Sam Ade Jacobs, Xiaoxia Wu, Connor Holmes, Zhewei Yao, Samyam Rajbhandari, Olatunji Ruwase, Feng Yan, Lei Yang, Yuxiong He, ZeRO++: Extremely Efficient Collective Communication for Large Model Training., *The Twelfth International Conference on Learning Representations. 2023 (ICLR 2023).*

Heyang Qin, Samyam Rajbhandari, Olatunji Ruwase, Feng Yan, Lei Yang, Yuxiong He, SimiGrad: Fine-Grained Adaptive Batching for Large Scale Training using Gradient Similarity Measurement, *in Proceedings of the Neural Information Processing Systems 2021 (NeurIPS 2021)*, Virtual, December, 2021 (Acceptance rate: 2371/9122=26%).

Heyang Qin, Syed Zawad, Yanqi Zhou, Lei Yang, Dongfang Zhao, Feng Yan, Swift Machine Learning Model Serving Scheduling: A Region Based Reinforcement Learning Approach, *in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2019)*, Denver, CO, USA, Nov, 2019 (Acceptance rate: 78/344=22%).

Technical Skills

Languages: Python, C++, CUDA, Triton.

Frameworks: OAI EngineV2/3, DeepSpeed, PyTorch, vLLM, SGLang, Megatron-LM.

Infrastructure: NCCL, Infiniband, MPI, Docker, Kubernetes.