Quentin Anthony

614-906-5623 • qubitquentin@gmail.com • LinkedIn • Github • Google Scholar

Research Interests

Broadly, my research is focused on the intersection of High Performance Computing (HPC) and Deep/Machine Learning (DL/ML). Specifically, I work to resolve bottlenecks in applying HPC systems to DL applications, such as checkpointing, model/optimizer compression, and DL/ML framework co-design.

Education

2019-Present • Ph.D. Computer Science and Engineering • The Ohio State University
• Advisor: D.K. Panda
2017-2019 • B.S. Engineering Physics • The Ohio State University
• Magna Cum Laude

Awards

2019 • Graduate University Fellowship (Full Funding for First PhD Year) • The Ohio State University
2019 • Magna Cum Laude • The Ohio State University
2019 • Hazel Brown Senior Award for Excellence in Physics • The Ohio State University
2018 • Helen Cowan Book Award • The Ohio State University
2017 • Maximus Merit Scholarship • The Ohio State University
2017 • Valentino Physics Scholarship Runner-up (1/2 award) • The Ohio State University

Select Publications

For full list, please see my Google Scholar page.

1. **Q. Anthony**, et al, ***MCR-DL: Mix-and-Match Communication Runtime for Deep Learning***, 37th IEEE International Parallel & Distributed Processing Symposium (IPDPS '23), May 2023
2. **Q. Anthony**, L. Xu, A. Shafi, H. Subramoni, and DK Panda, ***ScaMP: Scalable Meta-Parallelism for Deep Learning Search***, The 23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID '23), May 2023
3. **Q. Anthony**, D. Dai, ***Evaluating Multi-Level Checkpointing for Distributed Deep Neural Network Training***, Second International Symposium on Checkpointing for Supercomputing (SuperCheck '21), Nov 2021
4. **Q. Anthony**, L. Xu, H. Subramoni, and DK Panda, ***Scaling Single-Image Super-Resolution Training on Modern HPC Clusters: Early Experiences,*** Scalable Deep Learning over Parallel and Distributed Infrastructures (ScaDL '21), May 2021

EMAIL
qubitquentin@gmail.com

TELEPHONE
614-906-5623

5. **Q. Anthony**, A. A. Awan, A. Jain, H. Subramoni, and DK Panda, ***Efficient Training of Semantic Image Segmentation on Summit using Horovod and MVAPICH2-GDR***, Scalable Deep Learning over Parallel and Distributed Infrastructures, (ScaDL '20), May 2020
6. M. Ghazimirsaeed, **Q. Anthony**, H. Subramoni, and DK Panda, ***Accelerating GPU-based Machine Learning in Python using MPI Library: A Case Study with MVAPICH2-GDR***, Machine Learning in HPC Environments, (MLHPC '20), Nov 2020

Experience

[May 2022 – July 2022] and [May 2023 – Aug 2023] • Graduate Research Intern (Microsoft Research)
• [**2022**] Design and implement dedicated communication module within DeepSpeed, supporting advanced MPI, NCCL, and MSCCL backends directly
• [**2022**] Add communication logging, benchmarking, and live monitoring utilities to DeepSpeed
• [**2022**] Bring EleutherAI's DeeperSpeed up to date with the latest DeepSpeed, and integrate DeeperSpeed improvements into Microsoft DeepSpeed
• [**2022**] Publication on project accepted at IPDPS 2023 conference. Published DeepSpeed tutorials for project on communication logging and live monitoring
• [**2023**] Developed distributed GPU-aware KV-store with compression support and for BingAds

Jan 2022 – Present • Head of HPC (EleutherAI)
• Lead on DeeperSpeed and GPT-NeoX.
• Performed system tuning across software stack on partner cloud provider (CoreWeave)
• Performed systems optimization for Pythia suite and GPT-NeoX-20B
• Core contributor on joint INCITE ORNL grant with EleutherAI and MILA, leading to RedPajama-INCITE models, and a continual pretraining paper
• Lead author on Transformers Math 101 blog post

Jan 2022 – Present • Independent Consultant (StabilityAI, MILA, etc)
• Consulted on HPC, communication/system optimizations, and distributed LLM training
• Supported client research and development

May 2019 – Present • Graduate Research Assistant (NOWLAB)
• Investigate collective communication designs and implementations for CUDA-Aware MPI libraries like *MVAPICH2* and *MVAPICH2-GDR*
• Co-design MPI libraries like MVAPICH2 and Deep Learning frameworks like Pytorch and Tensorflow to enable efficient and scalable distributed Deep Learning on modern GPU clusters
• Led release of mpi4cuML, an MPI-Aware implementation of NVIDIA RAPIDS cuML

Aug 2020 – May 2022 • Software Engineer (X-Scale Solutions)
• Design, implement, and test distributed Deep Learning checkpointing tool to efficiently load and store massive DNN models at scale
• Develop and rigorously test X-ScaleAI, a distributed deep learning profiling tool, on HPC systems

Technical Skills

• Python, Java, C/C++, CUDA, MPI

• Strong communication and presentation ability
• Machine Learning (cuML, scikit-learn), Deep Learning Frameworks (Tensorflow, Pytorch, MXNet), and distributed Deep Learning Frameworks (DeepSpeed, Horovod, etc)
• Proficient in HPC and systems tools (Git, Linux kernel, debugging/build tools)

Professional Service

MEMBERSHIPS

- ACM Student Member
- IEEE Student Member

REVIEWER

- 28th IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC '21)
- ExaMPI21: Workshop on Exascale MPI [Held in conjunction with SC '20] (ExaMPI '21)
- IEEE TPDS Special Section: Innovative R&D toward the Exascale Era (2021)
- The 21st IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (Cluster '21)
- 50th International Conference on Parallel Processing (ICPP '21)
- Practice & Experience in Advanced Research Computing (PEARC '21)
- Scalable Deep Learning over Parallel And Distributed Infrastructures (ScaDL '21)
- 34th IEEE International Parallel & Distributed Processing Symposium (IPDPS '20)
- 38th IEEE International Conference on Computer Design (ICCD '19)

VOLUNTEER

- MVAPICH Users Group Meeting (MUG '19-'23)
- The International Conference for High Performance Computing, Networking,Storage, and Analysis (SC '21)
- 52nd IEEE/ACM International Symposium on Microarchitecture (MICRO '19)

References

Dhabaleswar Kumar (DK) Panda, Professor.
Dept. of Computer Science and Engineering
Tel: (614) 292-5199
Email: panda@cse.ohio-state.edu
Website: http://web.cse.ohio-state.edu/~panda.2/

Hari Subramoni, Professor.
Dept. of Computer Science and Engineering
Tel: (614) 688-8320
Email: subramoni.1@osu.edu

Ammar Awan, Senior Researcher.
Microsoft
Email: Ammar.Awan@microsoft.com