Quentin Anthony

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

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**, D. Dai, Evaluating Multi-Level Checkpointing for Distributed Deep Neural Network Training, Second International Symposium on Checkpointing for Supercomputing (SuperCheck '21), Nov 2021
2. **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
3. **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
4. 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

EMAIL
qubitquentin@gmail.com

TELEPHONE
614-906-5623

5. A. A. Awan, A. Jain, **Q. Anthony**, H. Subramoni, and DK Panda, *HyPar-Flow: Exploiting MPI and Keras for Scalable Hybrid-Parallel DNN Training using TensorFlow*, ISC High-Performance (ISC '20), June 2020
6. K. Khorassani , C. Chu , **Q. Anthony**, H. Subramoni , and DK Panda Adaptive and Hierarchical Large Message All-to-all Communication Algorithms for Large-scale Dense GPU Systems, 21st IEEE International Conference on Cluster Computing, (Cluster '19), Sep 2019

Experience

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
 • Evaluate designs on diverse HPC systems using existing benchmark suites like OSU Microbenchmarks (OMB), Intel MPI benchmarks (IMB) and test suites like MPICH tests, Intel tests, etc.
 • Regularly contribute to NOWLAB projects such as MVAPICH2, OMB, and mpi4cuML

Aug 2020 – Present • 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 emerging HPC systems
 • Manage production HPC licensing software for MVAPICH2-DPU, a derivative of the MVAPICH2 MPI library optimized for NVIDIA Bluefield Data Processing Units (DPUs)
 • Providing support to end-users for all X-Scale products

MAY 2018 – MAY 2019 • UNDERGRADUATE RESEARCH ASSISTANT (QUANTINFO LAB)
 • Implemented chaotic Boolean networks on FPGAs for random number generation
 • Contributed theory and setup on evaluating Boson Sampling problem in quantum information via coherent states
 • Honors senior thesis in simulating single photon states with phase-randomized weak coherent states for quantum networks

AUG 2017 – MAY 2019 • TECHNICAL INTERN – ARTIFICIAL INTELLIGENCE (NCI INC.)
 • Performed research in simulated quantum computing and noisy near-term quantum computers for use in machine learning, graph theory, and molecular simulation
 • Assisted in creating NCI Shai: An AI solutions framework built around robotic process automation (RPA) and Machine Learning (ML). Responsible for Natural Language Processing (NLP) and Optical Character Recognition (OCR) subsystems

JAN 2018 – MAY 2018 • STUDENT INSTRUCTIONAL ASSISTANT (OSU CSE DEPARTMENT)
 • Course covers logic, graph algorithms, asymptotic analysis of algorithms, and discrete math.

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)

Select Projects

June 2021 – Present • [EleutherAI](#)
• Performed system tuning across software stack on partner cloud provider ([CoreWeave](#))
• Contributed to [DeeperSpeed](#) and [GPT-NeoX](#). Author on [GPT-NeoX](#).
• Led integration of quantized optimizers into GPT-NeoX, demonstrating 20% smaller memory footprint for 2.7B parameter model ([announcement](#))

July 2020 – May 2021 • [mpi4cuML](#)
 • Co-Designed and implemented MPI-Aware implementation of NVIDIA RAPIDS cuML
 • Led subsequent release and customer support

Aug 2018 – May 2019 • Senior Capstone (Team Lead)
• Implementing computational real-time Holography on the Ohio Supercomputer Center
• Developed undergraduate optics lab with demonstrations of Poisson's spot, Talbot carpet, and polarization-dependent single slit diffraction

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

Volunteer Experience

• FIRST robotics mentor (Officially August 2018 – May 2020) teaching SolidWorks, electronics, Java to team Digital Fusion
• Volunteer for Ohio Science Olympiad at Ohio State University (2020)

References

Dhabaleswar Kumar (DK) Panda, Professor.
Dept. of Computer Science and Engineering
The Ohio State University
2015 Neil Avenue
Columbus, OH-43210, USA
Tel: (614) 292-5199
Email: panda@cse.ohio-state.edu
Website: http://web.cse.ohio-state.edu/~panda.2/

Hari Subramoni, Research Scientist.
Dept. of Computer Science and Engineering
The Ohio State University
2015 Neil Avenue
Columbus, OH-43210, USA
Tel: (614) 688-8320
Email: subramoni.1@osu.edu
Website: http://web.cse.ohio-state.edu/~subramoni.1/

Donglai Dai, Chief Engineer.
X-ScaleSolutions
Email: d.dai@x-scalesolutions.com