Xu, RuQing

Deep learning algorithm engineer focusing on fast CUDA kernels, op-fusion, and world foundation models.

https://github.com/xrq-phys/ +81 70 3760 4882
https://rxu.purineko.me/ rxu.sol@gmail.com

Education

The University of Tokyo 01-OCT-2021 – 24-SEP-2024
Ph.D. *in Physics received on 24-SEP-2024* *Tokyo, JP*
Focus: *Statistical Physics (Computer Simulations: Algorithmic and Platform Optimizations)*

The University of Tokyo 20-SEP-2019 – 30-SEP-2021
M.S. *in Physics received on 24-SEP-2021* *Tokyo, JP*
Focus: *Statistical Physics: (Computer Simulations & HPC Basics)*

University of Science and Technology of China 01-SEP-2015 – 01-JUL-2019
B.S. *in Physics received on 01-JUL2019* *Hefei, CN*
Dept: *Theoretical Physics*

Employment History

Snr. Deep Learning Algorithm Engineer 29-JUL-2024 – Present
NVIDIA *Tokyo. JP*

- Developing multiple variants of high-throughput attention kernels for NVIDIA's DiT-based research workloads.
- Ensuring throughput-critical kernels work as fast as possible at our current and next-generation GPU models.
- Taking charge of the optimizations before launch of Cosmos DiT models as a NVIDIA Inference Microservice.
- Deliver Tensor-Core-emulated SGEMM & CGEMM kernels to cuBLAS & cuTENSOR for NVIDIA Blackwell architectures.

Deep Learning Algorithm Engineering Intern (Deep Learning) 31-JAN-2023 – 28-JUL-2024
NVIDIA (Intern) *Tokyo, JP*

- Helps development & performance verification for workload-specific kernels.
- Synchronize with research teams to make their workloads run as fast as possible.
- Integrate algorithms for Tensor-Core-based fast matrix multiplication from published materials into cuTENSOR. Polish the implementation so as to align with the best that the latest hardware (Hopper GMMA) can provide.

Deep Learning Algorithm Engineering Intern (cuTENSOR) 27-SEP-2021 – 31-JAN-2022
NVIDIA (Intern) *Tokyo, JP*

- Special techniques for medium-size performance improvements.
- Tackled multiple unusual underperforming cases.
- In-depth L2 bandwidth analysis & optimizations.

Publications

SCA/HPCAsia 2026, Angelika Schwarz, Anton Anders, Cole Brower, Harun Bayraktar, John Gunnels, Kate Clark, **RuQing G. Xu**, Samuel Rodriguez, et. al., *Guaranteed DGEMM Accuracy While Using Reduced Precision Tensor Cores Through Extensions of the Ozaki Scheme*

SIAM SISC (in review), Ishna Satyarth, Chao Yin, Devin A. Matthews, Maggie Myers, Robert van de Geijn, **RuQing G. Xu**, *Performant Tridiagonal Factorization of Skew-symmetric Matrices*

arXiv:2311.10700, Robert van de Geijn, Maggie Myers, **RuQing G. Xu**, Devin Matthews, *Deriving Algorithms for Triangular Tridiagonalization a (Skew-)Symmetric Matrix*

ICS '23: Proc. 37th Intl. Conf. Supercomputing: pp. 111–121, **RuQing G. Xu**, Field G. Van Zee, Robert A. van de Geijn, *Towards a Unified Implementation of GEMM in BLIS*

*Comput. Phys. Commun. **277**, 108375*, **RuQing G. Xu**, Tsuyoshi Okubo, Synge Todo, Masatoshi Imada, *Optimized Implementation for Calculation and Fast-Update of Pfaffians Installed to the Open-Source Fermionic Variational Solver mVMC*

*Phys. Rev. Research **3**, 023048*, Xinliang Lyu, **RuQing G. Xu**, Naoki Kawashima, *Scaling dimensions from linearized tensor renormalization group transformations*

*J. Chem. Theory Comput. 2019, **15**, 3, 1728-1742*, James S. Spencer, Nick S. Blunt, ..., William A. Vigor, **RuQing Xu**, Alex J. W. Thom, *The HANDE-QMC project: open-source stochastic quantum chemistry from the ground state up*

Skills

Programming Languages: C, C++, CUDA, NVPTX, x86 Assembly, Arm64 Assembly, Julia, Python, OpenAI Triton
Natural Languages: Chinese (native), English (GRE: 330; TOEFL: 108), Japanese (JLPT N1)