Dear Selection Committee,

I am writing to apply for the 2021 Summer UTEA. I have completed three years of the Computer Science Specialist program and am currently doing a PEY placement at SOTI. My primary research interests are in the fields of heterogeneous computing and compiler design and development, which I hope to explore in working with Professor Gennady Pekhimenko's EcoSystems research group. I would particularly like to work on their "Horizontally Fused Training Array (HFTA)" project.

My computer science education has given me a solid theoretical understanding of computer architecture and concurrency, particularly in a heterogeneous setting. I am passionate about applying this background to practical problems in systems development and design. I have extensive experience developing high-performance low-level software, including highly-optimized multithreaded code, operating systems components, distributed and fault-tolerant networked systems, compilers, and GPU kernels. I also have an interest in scientific computing and its applications, especially in machine learning. Recently, I have participated in the development of RAIN, an RVSDG-based research compiler built on top of LLVM. I was responsible for writing the code-generation module in the backend, which was to generate LLVM IR from the compiler's internal data structures. In the field of machine learning, I have applied convolutional neural networks to the classification of medical images as part of the Image-Guided Radiation Therapy Net project at the Fields Institute.

I believe the HFTA project aligns significantly with my skills and experience in machine learning and compiler development. I am especially interested in the potential gains in throughput (according to HFTA's project description, up to 15.1x) obtained by merging shared operator structure in common deep learning workflows. One limitation of current approaches, according to unpublished work, is that they cannot fuse operators having different shapes, which may occur, e.g., when considering networks having different batch sizes. I am interested in applying persistent CUDA threads, as described in Rammer [1], as one possible method to remove this limitation. I also believe that extending HFTA with a modified version of the scheduling techniques in Rammer could yield a promising direction for further research.

I believe that I can make a meaningful contribution to the work of Professor Pekhimenko's research group. I hope to take advantage of this opportunity to obtain invaluable research experience, connections, and knowledge, which I hope to apply in my career as a systems researcher.

Thank you very much for your time and consideration,

Sincerely,

Qingyuan Qie

References

[1] Lingxiao Ma et al. "Rammer: Enabling Holistic Deep Learning Compiler Optimizations with rTasks". In: *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*. USENIX Association, Nov. 2020, pp. 881–897. ISBN: 978-1-939133-19-9. URL: https://www.usenix.org/conference/osdi20/presentation/ma.