Research Summary

Hongyu Miao

I'm broadly interested in computer systems, including operating systems, networking, compiler, architecture, and runtime systems for emerging applications.

During my PhD, I have been focusing on systems support for real-time data analytics (stream processing and machine learning) by exploiting modern hardware (multicore, hybrid memory, and microcontrollers):

(1) StreamBox: modern stream processing on a multicore machine [1]

StreamBox exploits the parallelism and memory hierarchy of modern multicore hardware. StreamBox executes a pipeline of transforms over records that may arrive out-of-order. The key contribution is to produce and manage abundant parallelism by generalizing out-of-order record processing to out-of-order epoch processing, and by dynamically prioritizing epochs to optimize latency.

(2) StreamBox-HBM: stream analytics on high bandwidth hybrid memory [2]

StreamBox-HBM achieves scalable high performance. It performs data grouping with sequential access sorting algorithms in HBM, in contrast to random access hashing algorithms commonly used in DRAM. It dynamically creates and manages prodigious data and pipeline parallelism, choosing when to allocate HBM. It dynamically optimizes for both the high bandwidth and limited capacity of HBM, and the limited bandwidth and high capacity of standard DRAM.

(3) Enabling large neural networks on tiny microcontrollers with swapping [3]

To run neural networks (NNs) on microcontroller units (MCUs), memory size is the major constraint. While algorithm-level techniques exist to reduce NN memory footprints, the resultant losses in NN accuracy and generality disqualify MCUs for many important use cases. We investigate a system solution for MCUs to execute NNs out-of-core: dynamically swapping NN data chunks between an MCU's tiny SRAM and its large, low-cost external flash. Out-of-core NNs on MCUs raise multiple concerns: execution slowdown, storage wear out, energy consumption, and data security. We present a study showing that none is a showstopper; the key benefit - MCUs being able to run large NNs with full accuracy/generality - triumphs the overheads. Our findings suggest that MCUs can play a much greater role in edge intelligence.

References

[1] Hongyu Miao, Heejin Park, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S. McKinley, and Felix Xiaozhu Lin. StreamBox: Modern stream processing on a multicore machine. USENIX Annual Technical Conference (USENIX ATC 2017), pp. 617-629, Santa Clara, CA, July 2017

[2] Hongyu Miao, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S. McKinley, and Felix Xiaozhu Lin. StreamBox-HBM: Stream Analytics on High Bandwidth Hybrid Memory. The 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2019), pp. 167-181, Providence, RI, April 2019

[3] Hongyu Miao and Felix Xiaozhu Lin. Enabling Large Neural Networks on Tiny Microcontrollers with Swapping. (Under Review)