Research Summary

Hongyu Miao

I'm broadly interested in computer systems, including operating systems, networking, compiler, computer architecture, and runtime systems for emerging applications. During my Ph.D. study, 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 [3] Stream analytics on real-time events has an insatiable demand for throughput and latency. Most existing distributed stream engines (e.g., Spark Streaming and Apache Beam) are only optimized for tens/hundreds of machines, but leave every single multicore machine underutilized, which wastes resources.

StreamBox exploits the parallelism and memory hierarchy of modern multicore hardware on single machines. 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. On a 56-core machine, StreamBox processes records up to 38 GB/sec (38M Records/sec) with 50 ms latency, which is an order of magnitude faster than existing stream engines, e.g., Spark and Beam.

(2) StreamBox-HBM: stream analytics on high bandwidth hybrid memory [1] Stream analytics has an insatiable demand for memory and performance. Emerging hybrid memories combine commodity DDR4 DRAM with 3D-stacked High Bandwidth Memory (HBM) DRAM to meet such demands. However, achieving this promise is challenging because (1) HBM is capacitylimited and (2) HBM boosts performance best for sequential access and high parallelism workloads. At first glance, stream analytics appears a particularly poor match for HBM because they have high capacity demands and data grouping operations, their most demanding computations, use random access.

StreamBox-HBM exploits hybrid memories to 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 solely uses HBM to store Key Pointer Array (KPA) data structures that contain only partial records (keys and pointers to full records) for grouping operations. 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.

StreamBox-HBM is the first stream engine optimized for hybrid memories. It achieves 110 million records per second and 238 GB/s memory bandwidth while effectively utilizing all 64 cores of Intel's Knights Landing, a commercial server with hybrid memory. It outperforms stream engines with sequential access algorithms without KPAs by 7× and stream engines with random access algorithms by an order of magnitude in throughput.

(3) Enabling large neural networks on tiny microcontrollers with swapping [2] 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, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S McKinley, and Felix Xiaozhu Lin. Streambox-hbm: Stream analytics on high bandwidth hybrid memory. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 167–181, 2019.

[2] Hongyu Miao and Felix Xiaozhu Lin. Enabling large neural networks on tiny microcontrollers with swapping. *arXiv:2101.08744*, 2021.

[3] Hongyu Miao, Heejin Park, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S McKinley, and Felix Xiaozhu Lin. Streambox: Modern stream processing on a multicore machine. In *2017 USENIX Annual Technical Conference (USENIX ATC 17)*, pages 617–629, 2017.