

StreamBox-HBM: Stream Analytics on High Bandwidth Hybrid Memory

The first stream analytics engine optimized for 3D Memory + DRAM on real hardware

Hongyu Miao, *Purdue ECE*; Myeongjae Jeon, *UNIST*; Gennady Pekhimenko, *UToronto*; Kathryn S. McKinley, *Google*; Felix Xiaozhu Lin, *Purdue ECE*

1. Motivation

- Performance demands of Stream analytics**
High throughput: hundreds MRec/s
Low latency: sub second delay

- Hybrid Memory: 3D memory + DRAM**
3D memory: high bandwidth, small capacity;
no latency benefit (unlike: SRAM+DRAM); same as DRAM w/o high parallelism / sequential access

- How can stream analytics use hybrid memory?**

2. Parallel Sort for Grouping

- Grouping performs poorly on 3D memory
- Known duals of Grouping: Hash vs. Sort**
DRAM: Hash is the best
3D mem: Sort outperforms Hash with
 - Abundant bandwidth
 - High task parallelism
 - Wide SIMD (AVX-512)
 - Sequential access

3. Only use 3D mem for in-mem index

- 3D mem is capacity limited
- Minimize the use of precious capacity while exploiting the high bandwidth

4. Balance two limited resources

- 3D mem capacity and DRAM bandwidth
- Dynamically place data based on pressures

5. Key results

- Comparing to Flink**

- Validation of key designs**

Intel Xeon Phi Knights Landing: 16GB 3D memory, 96 GB DRAM, 64 cores@1.3GHz. \$5,000
x56: Intel Xeon E7-4830v4. 4x14 cores @2.0GHz. 256GB. \$23,000

6. Lessons

- If you want to use 3D memory to speed up applications, consider all of the following:

Apps	High task parallelism	Wide SIMD (avx512)	Sequential mem access	Packed data structure
Runtime	Thread pool + custom task scheduler		Custom mem allocator	
OS kernel	Cheap VM (huge page)		RDMA network bypass kernel, free CPU	
Hybrid Memory				

Acknowledgement: The authors affiliated with Purdue ECE were supported in part by NSF Award 1718720, NSF Award 1619075, and a Google Faculty Award