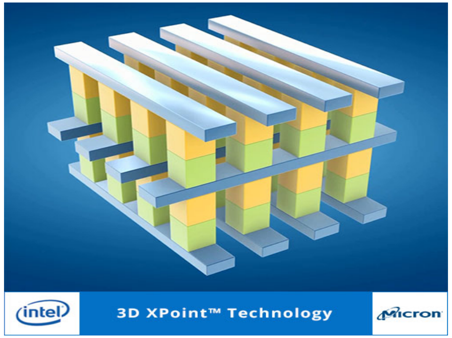
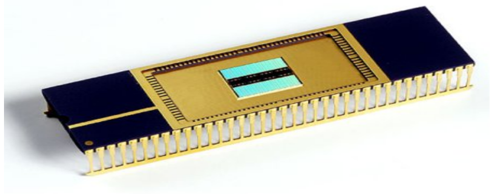





Larger-than-Memory Data Management on Modern Storage Hardware for In-Memory OLTP Database Systems

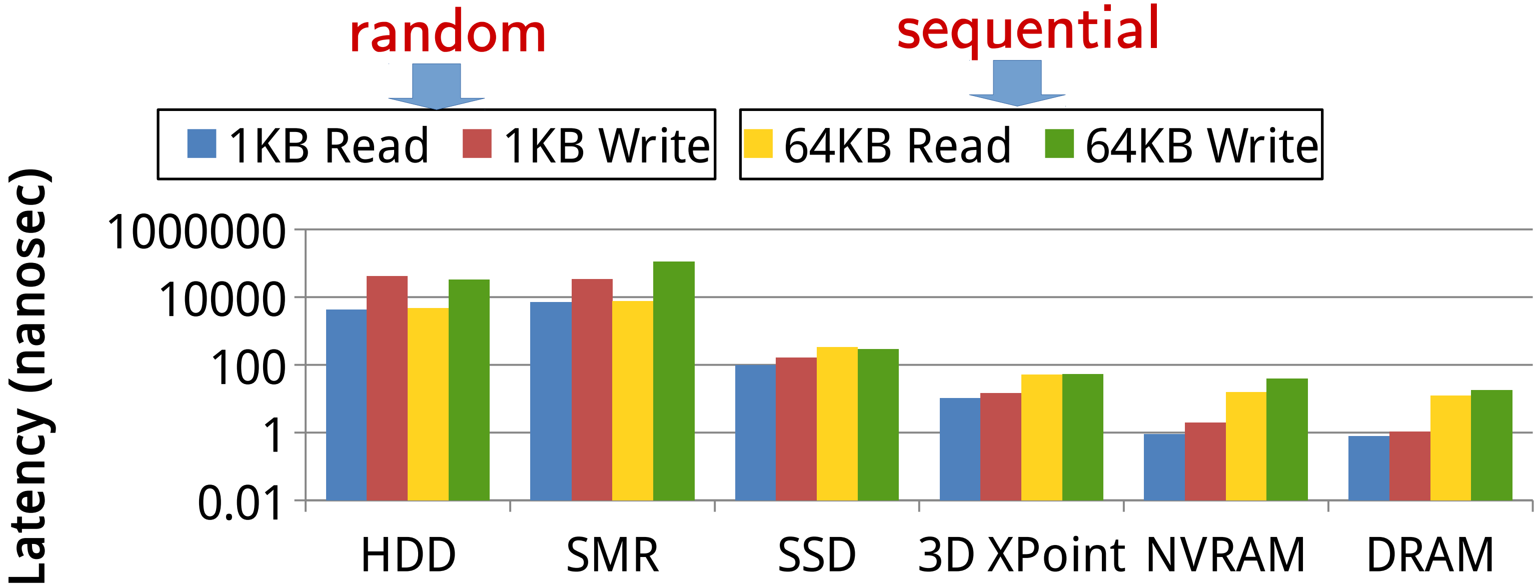
Lin Ma, Joy Arulraj, Sam Zhao, Andrew Pavlo, Subramanya R. Dulloor, Michael J. Giardino, Jeff Parkhurst, Jason L. Gardner, Kshitij Doshi, Col. Stanley Zdonik

MOTIVATION

- Allow an in-memory DBMS to store/access data on disk without bringing back all the slow parts of a disk-oriented DBMS.
- Different properties of storage devices may affect important design decisions.

STORAGE TECHNOLOGIES

- 10m Tuples – 1KB each
- Synchronization Enabled



DESIGN DECISIONS

- Hardware independent policies
  - Cold Tuple Identification
  - Evicted Tuple Meta-data

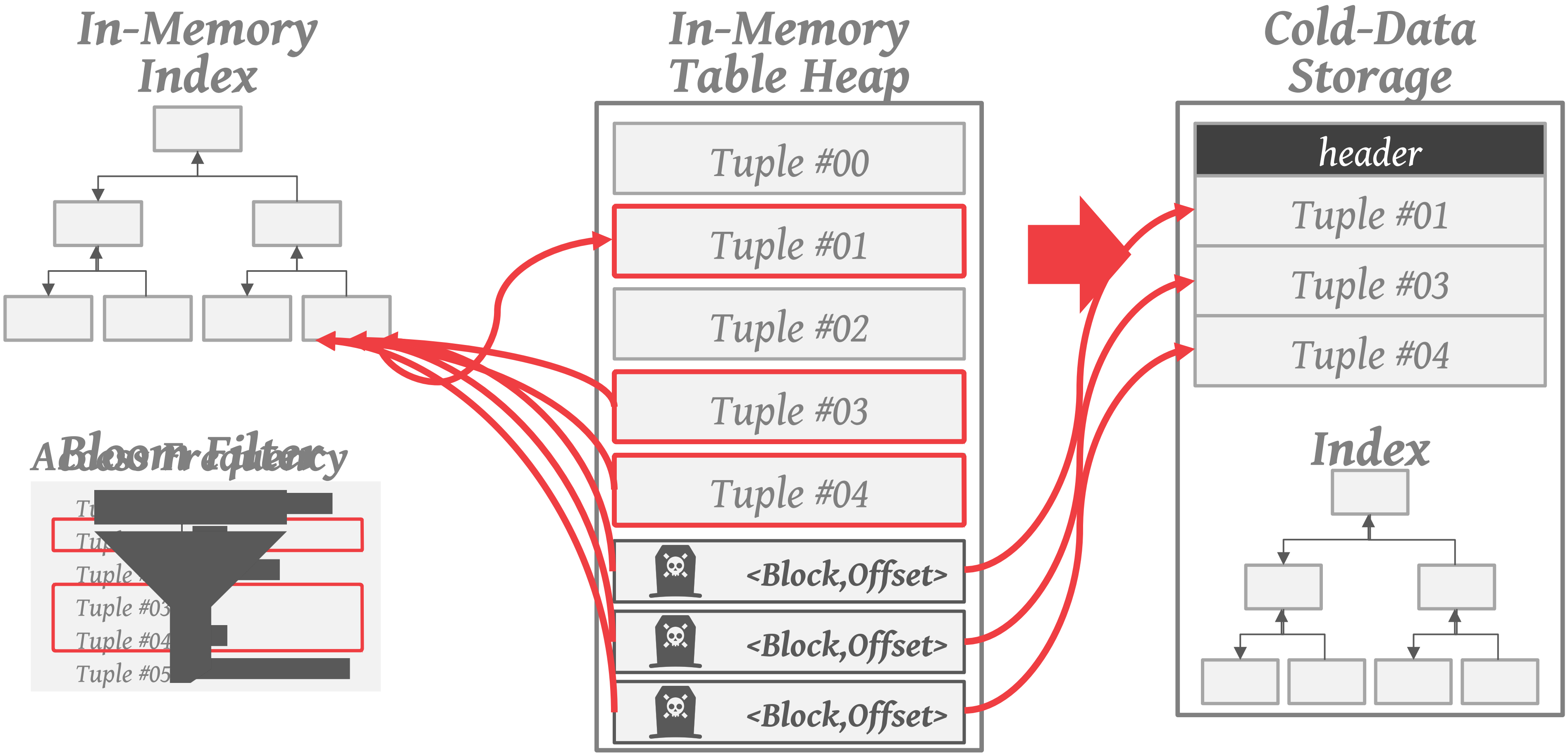
- Hardware dependent policies
  - Cold Tuple Retrieval
  - Merging Threshold
  - Access Methods

HARDWARE INDEPENDENT POLICIES

INDEPENDENT POLICIES

- Cold Tuple Identification
  - Option #1: On-line identification
  - Option #2: Off-line identification
- Evicted Tuple Meta-data
  - Option #1: Marker to represent the on-disk position
  - Option #2: Bloom filter
  - Option #3: Rely on virtual paging

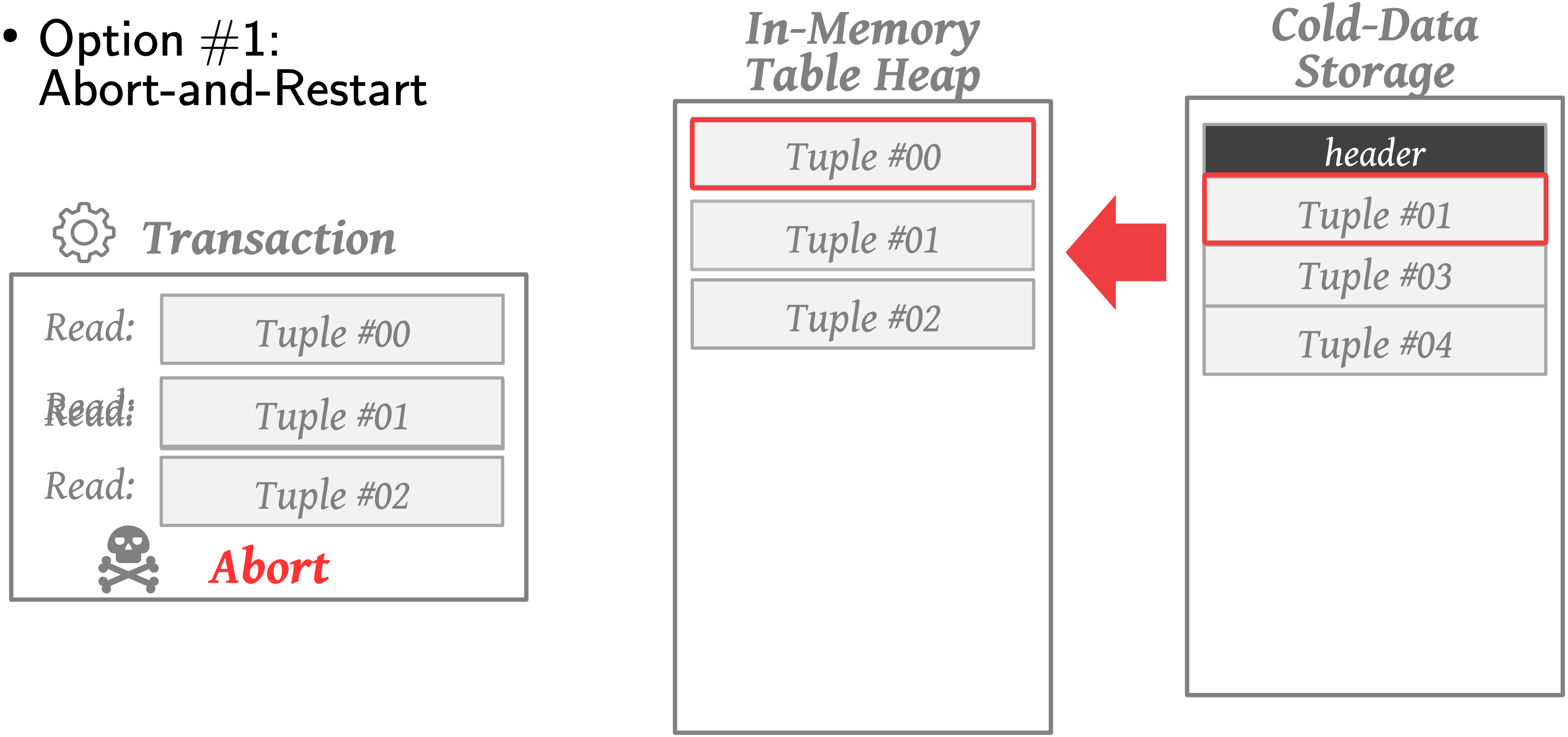
EVICTED TUPLE META-DATA



HARDWARE DEPENDENT POLICIES

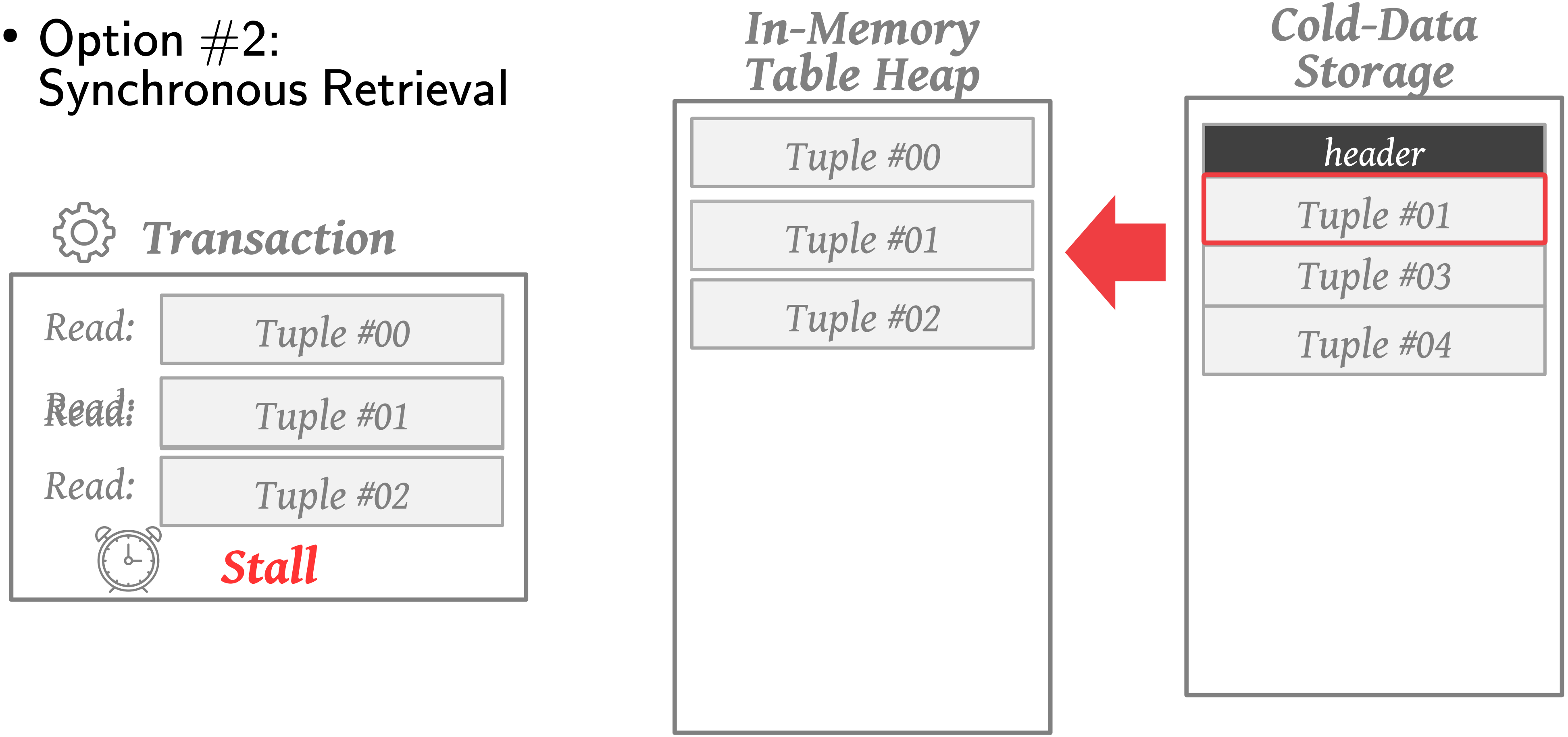
COLD TUPLE RETRIEVAL

- Option #1: Abort-and-Restart



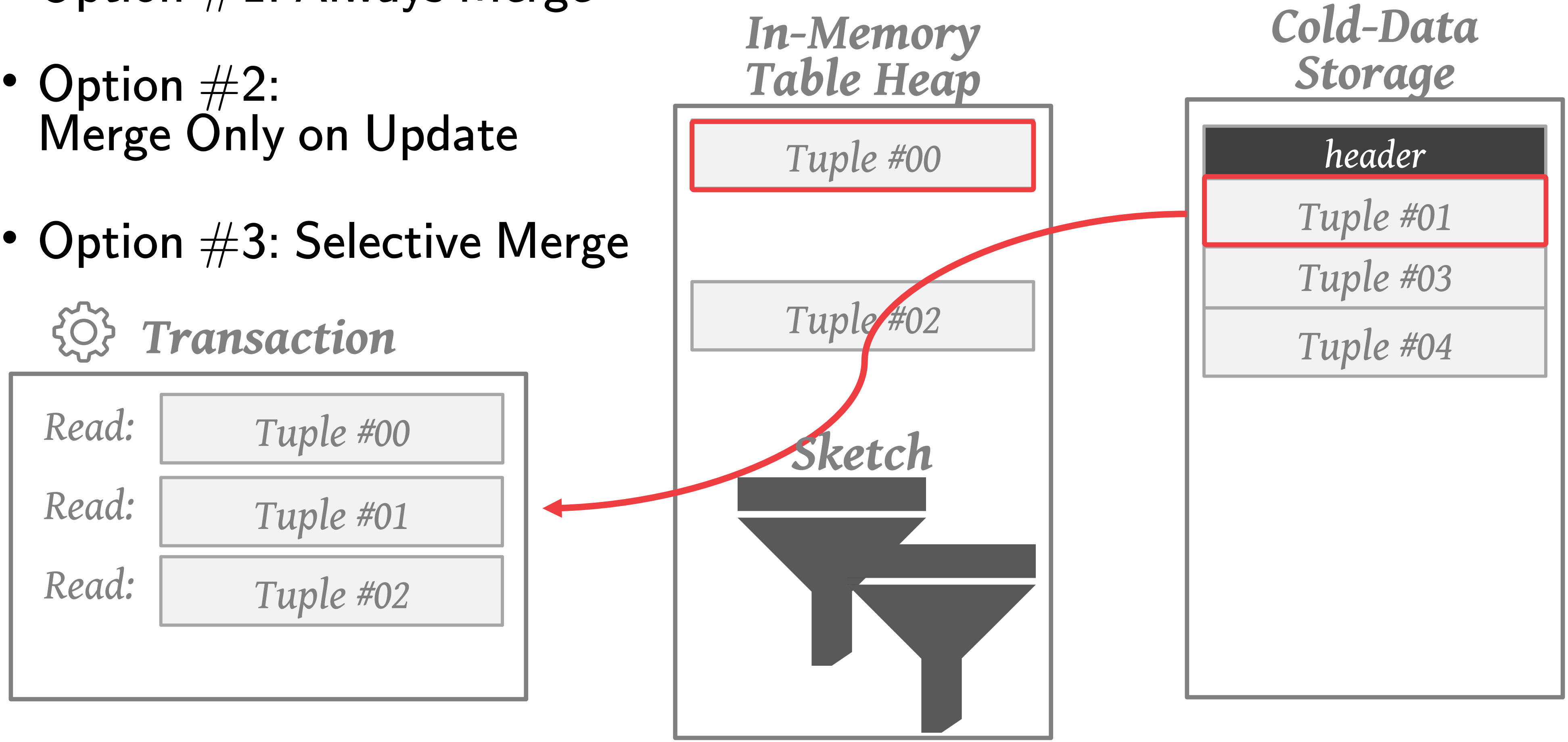
COLD TUPLE RETRIEVAL

- Option #2:
  Synchronous Retrieval



MERGING THRESHOLD

- Option #1: Always Merge
- Option #2:
  Merge Only on Update
- Option #3: Selective Merge



ACCESS METHODS

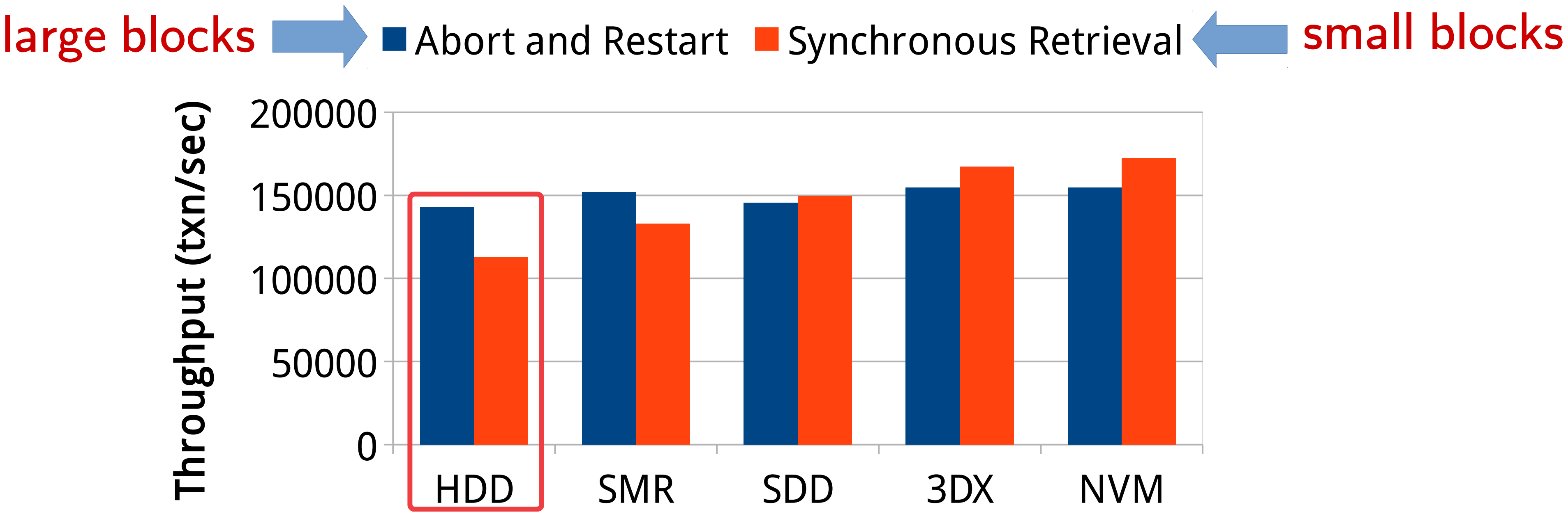
- Option #1: Block-addressable
  - Block-level access through file system
- Option #2: Byte-addressable (NVRAM)
  - Use mmap through a filesystem designed for byte-addressable NVRAM (PMFS)
  - Directly operate on NVRAM-resident data as if it existed in DRAM

EVALUATION

- Compare design decisions in H-Store with anti-caching.
- Storage Devices:
  - Hard-Disk Drive (HDD)
  - Shingled Magnetic Recording Drive (SMR)
  - Solid-State Drive (SSD)
  - 3D XPoint (3DX)
  - Non-volatile Memory (NVRAM)

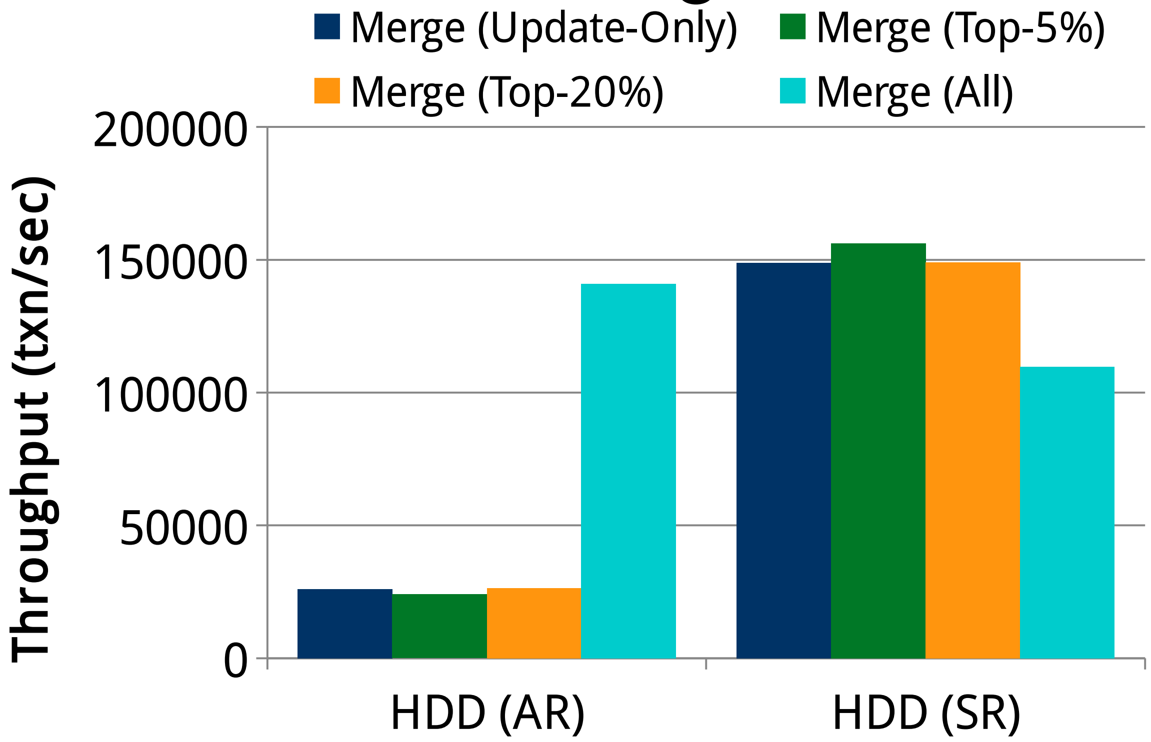
COLD TUPLE RETRIEVAL

- YCSB Workload – 90% Reads / 10% Writes
- 10GB Database using 1.25GB Memory



MERGING THRESHOLD

- YCSB Workload – 90% Reads / 10% Writes
- 10GB Database using 1.25GB Memory



Improvement:

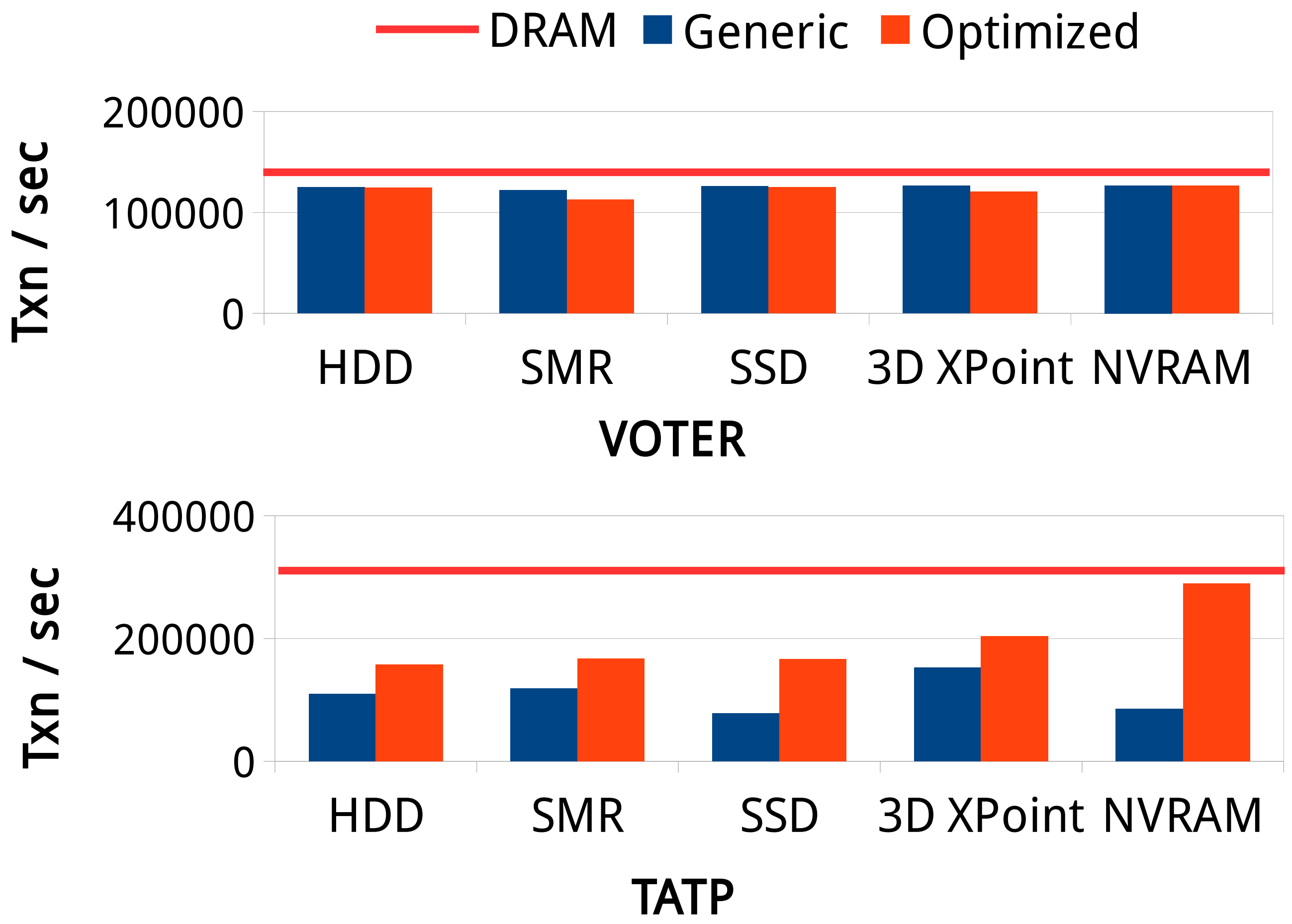
- SMR: 30%
- SSD: 17%
- 3DX: 21%
- NVRAM: 10%

CONFIGURATION COMPARISON

- **Generic Configuration (2013 Anti-caching)**
  - Abort-and-Restart Retrieval
  - Merge (All) Threshold
  - 1024 KB Block Size

- **Optimized Configuration**
  - Synchronous Retrieval
  - Top-5% Merge Threshold
  - Block Sizes (HDD/SMR-1024 KB) (SSD/3DX-16 KB)
  - Byte-addressable access for NVRAM

GENERIC VS OPTIMIZED



CONCLUSION

- Low-latency storage devices: Smaller block sizes and synchronous retrieval
- Constraints on merge frequency improve performance
- The performance of NVRAM is as good as pure DRAM if treated correctly

END

lin.ma@cs.cmu.edu

CARNEGIE MELLON
DATABASE GROUP

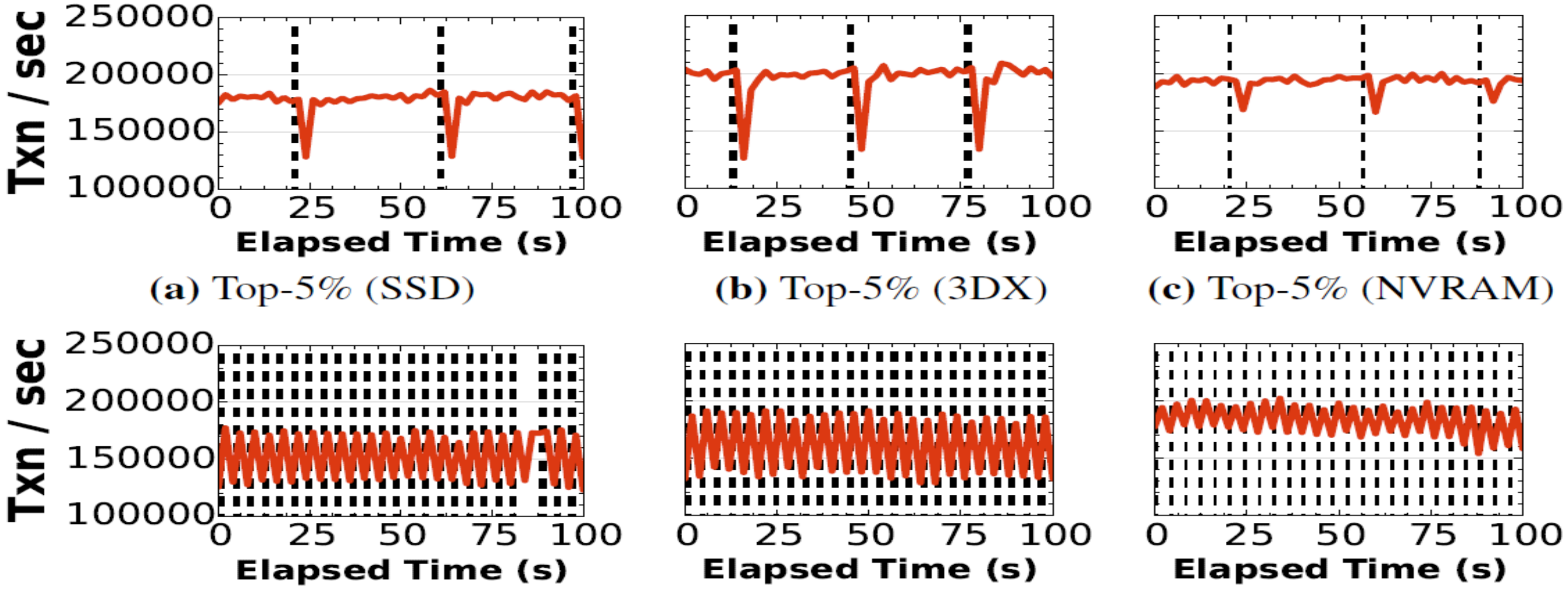
1-844-88-CMUDB

REAL-WORLD IMPLEMENTATIONS

- H-Store – Anti-Caching
- Microsoft Hekaton –Project Siberia
- EPFL's VoltDB Prototype
- Apache Geode – Overflow Tables
- MemSQL – Columnar Tables
- SolidDB
- P*TIME

MERGING THRESHOLD

- YCSB Workload – 90% Reads / 10% Writes
- 10GB Database using 1.25GB Memory



(a) Top-5% (SSD) (b) Top-5% (3DX) (c) Top-5% (NVRAM)

(d) Merge All (SSD) (e) Merge All (3DX) (f) Merge All (NVRAM)