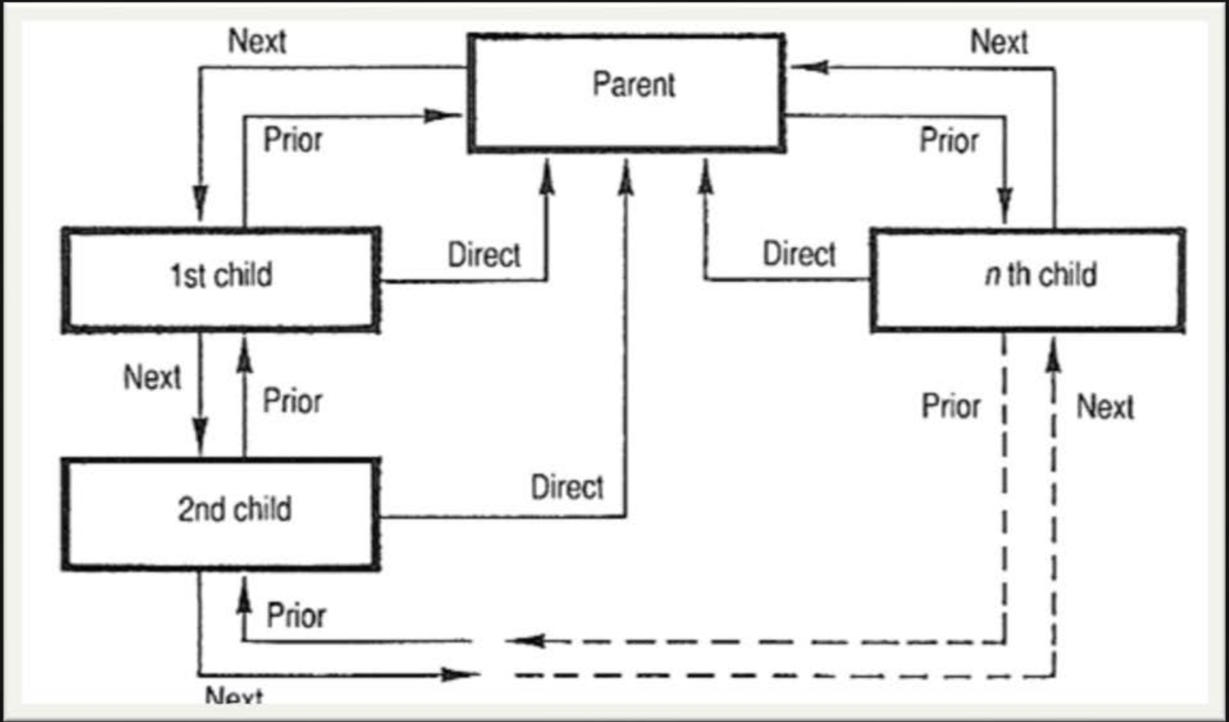
WHAT NON-VOLATILE MEMORY MEANS FOR THE FUTURE OF DATABASE SYSTEMS

CARNEGIE MELLON DATABASE GROUP

@ANDY_PAVLO

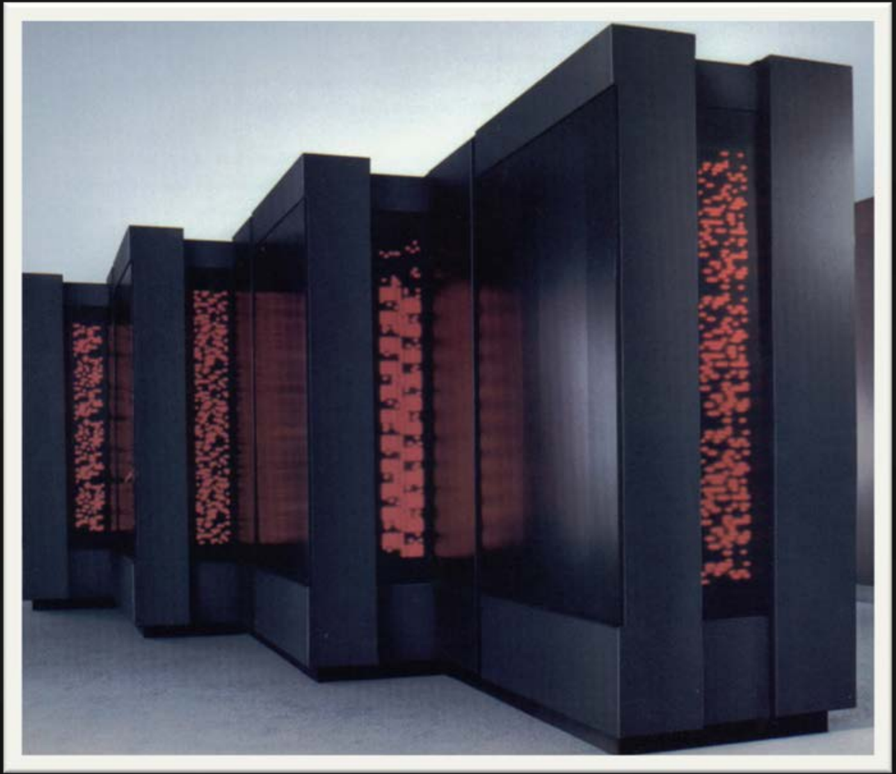

Next
Parent
Next
Prior
Prior
1st child
Direct
Direct
n th child
Next
Prior
Prior
Next
2nd child
Direct
Prior
Next

1973

1974

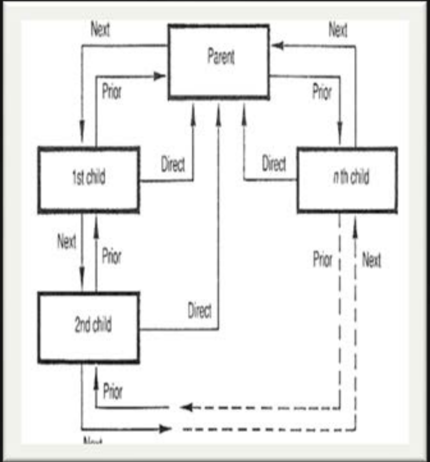
1978

1986



1994

2010


Next
Parent
Next
Prior
Prior
1st child
Direct
Direct
n th child
Next
Prior
Prior
Next
2nd child
Direct
Prior




2016

The Future

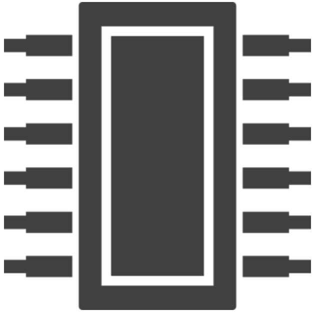
Non-Volatile Memory

- Persistent storage with byte-addressable operations.
- Fast read/write latencies.
- No difference between random vs. sequential access.

What does NVM mean for DBMSs?

- Thinking of NVM as just a faster SSD is not interesting.
- We want to use NVM as permanent storage for the database, but this has major implications.
  - *Operating System Support*
  - *Cloud Provider Provisioning*
  - *Database Management System Architectures*

Existing Systems



NVM-Only Storage

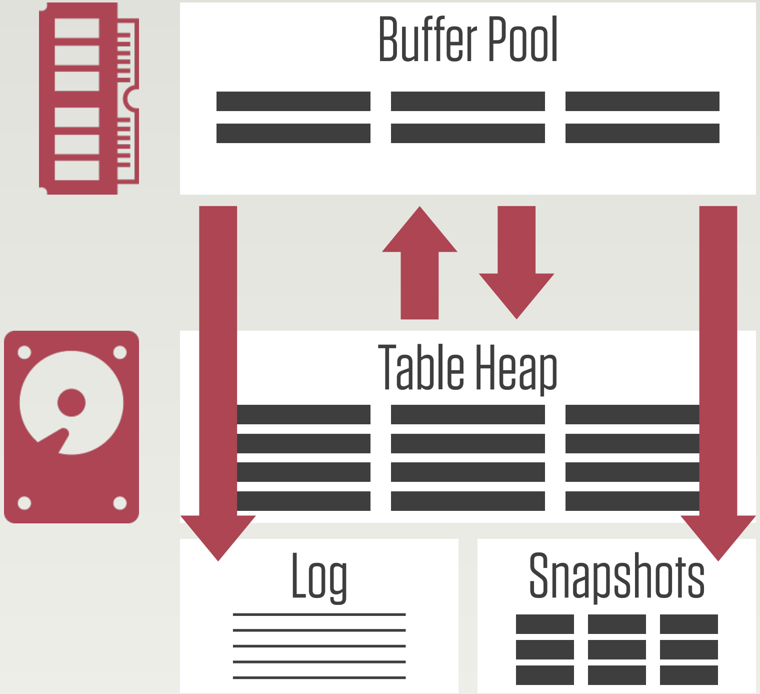
Hybrid DBMS

Chapter I – Existing Systems

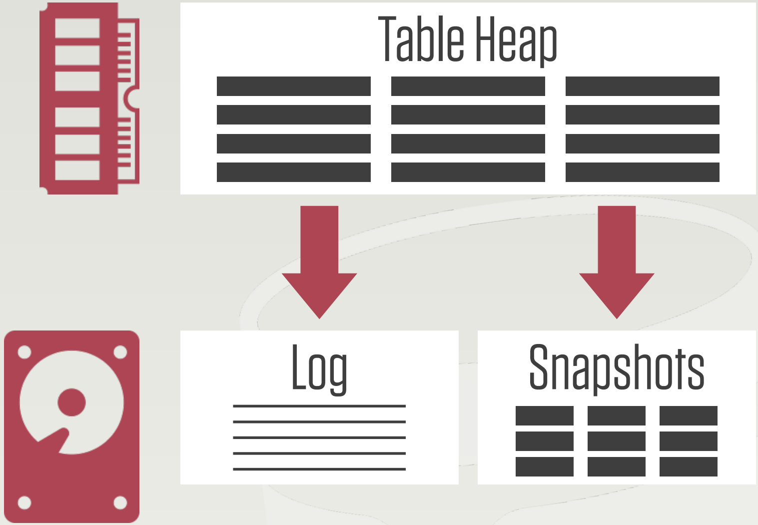
- Investigate how existing systems perform with NVM for write-heavy transaction processing (OLTP) workloads.
- Evaluate two types of DBMS architectures.
  - *Disk-oriented (**MySQL**)*
  - *In-Memory (**H-Store**)*

A PROLEGOMENON ON OLTP DATABASE SYSTEMS
FOR NON-VOLATILE MEMORY
ADMS@VLDB 2015

DISK-ORIENTED

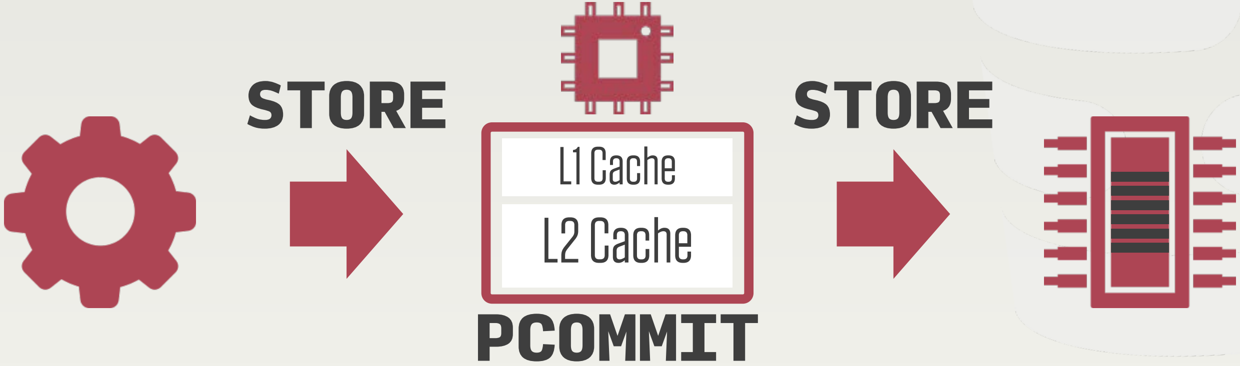


IN-MEMORY



Intel Labs NVM Emulator

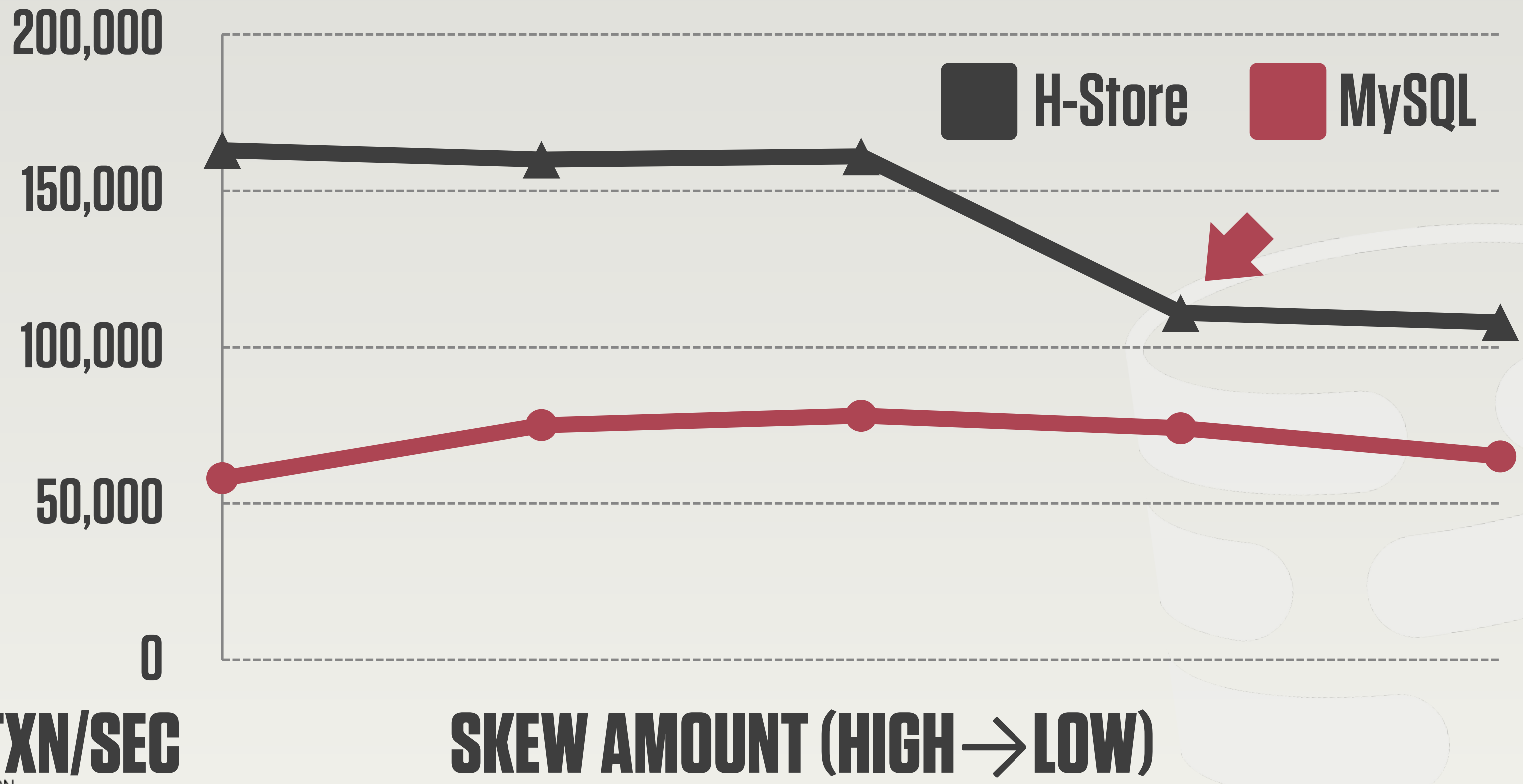
- Instrumented motherboard that slows down access to the memory controller with tunable latencies.
- Special assembly to emulate upcoming Xeon instructions for flushing cache lines.



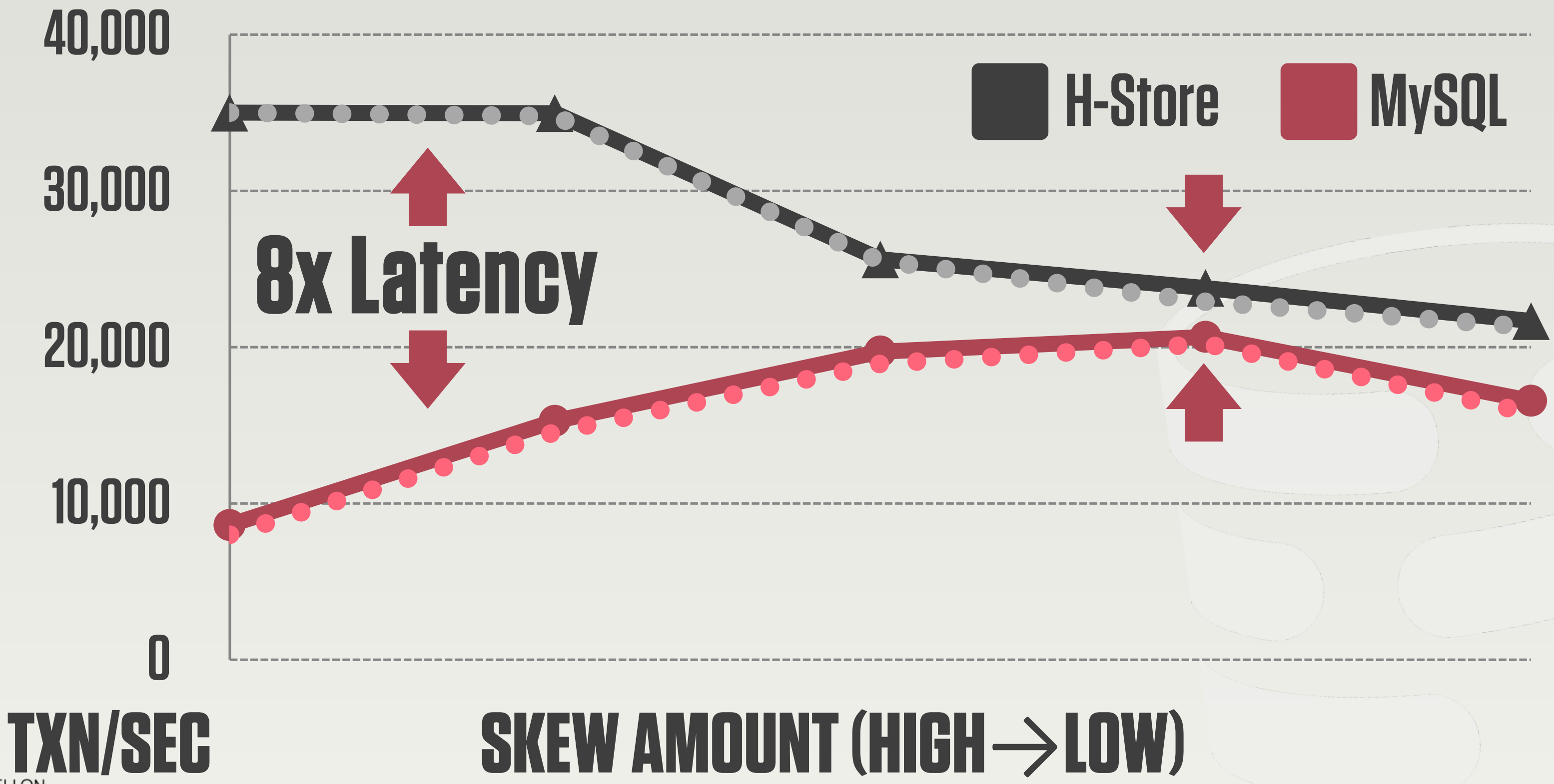
Experimental Evaluation

- Compare architectures on Intel Labs NVM emulator.
- Yahoo! Cloud Serving Benchmark:
  - *10 million records (~10GB)*
  - *8x database / memory*
  - *Variable skew*

YCSB // *Read-Only Workload 2x Latency Relative to DRAM*



YCSB // *50% Reads / 50% Writes Workload* *2x Latency Relative to DRAM*



LESSONS

1. NVM Latency does not have a large impact.
2. Logging is a major performance bottleneck.
3. Legacy DBMSs are **not** prepared to run on NVM.

What would Larry Ellison do?


BNP PARIBAS
BNP PARIBAS OPEN
BNP PARIBAS OPEN

Chapter II – NVM-only Storage

- Evaluate storage and recovery methods for a system that only has NVM.
- Testbed DBMS with a pluggable storage engines.
- We had to build our own NVM-aware memory allocator.

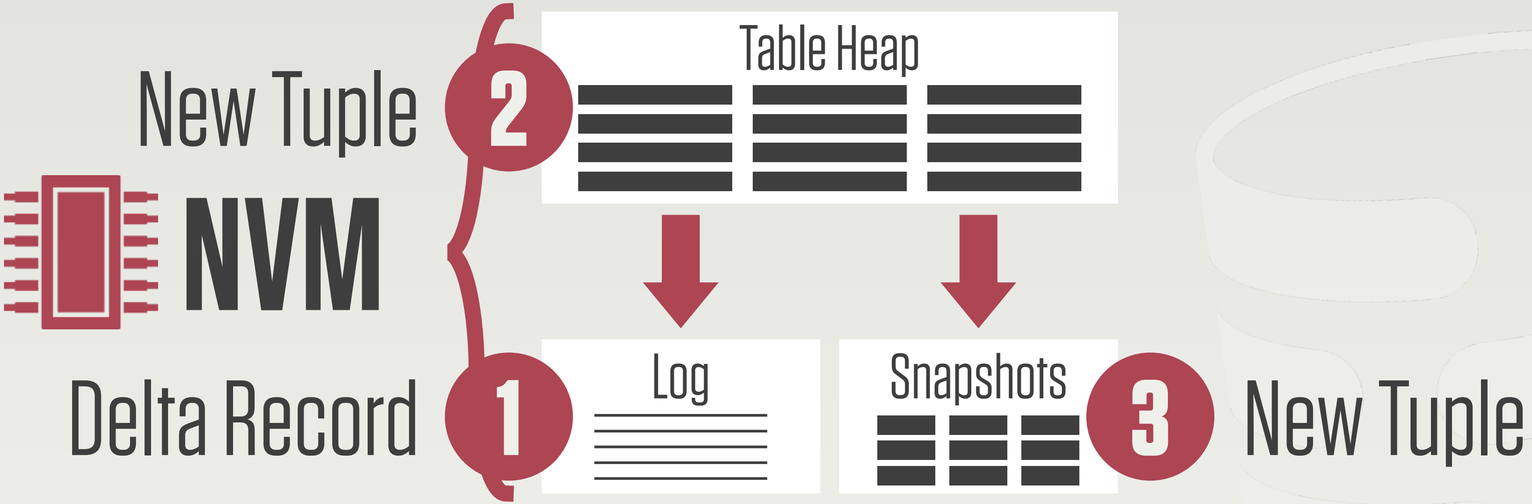
LET'S TALK ABOUT STORAGE & RECOVERY METHODS FOR NON-VOLATILE MEMORY DATABASE SYSTEMS
SIGMOD 2015

DBMS Architectures

In-Place	Copy-on-Write	Log-Structured
Table Heap	Table Heap	No Table Heap
Log + Snapshots	No Logging	Log-only Storage
VOLTDB	LMDB	RocksDB

In-Place Engine

```
UPDATE table SET val=ABC
 WHERE id=123
```

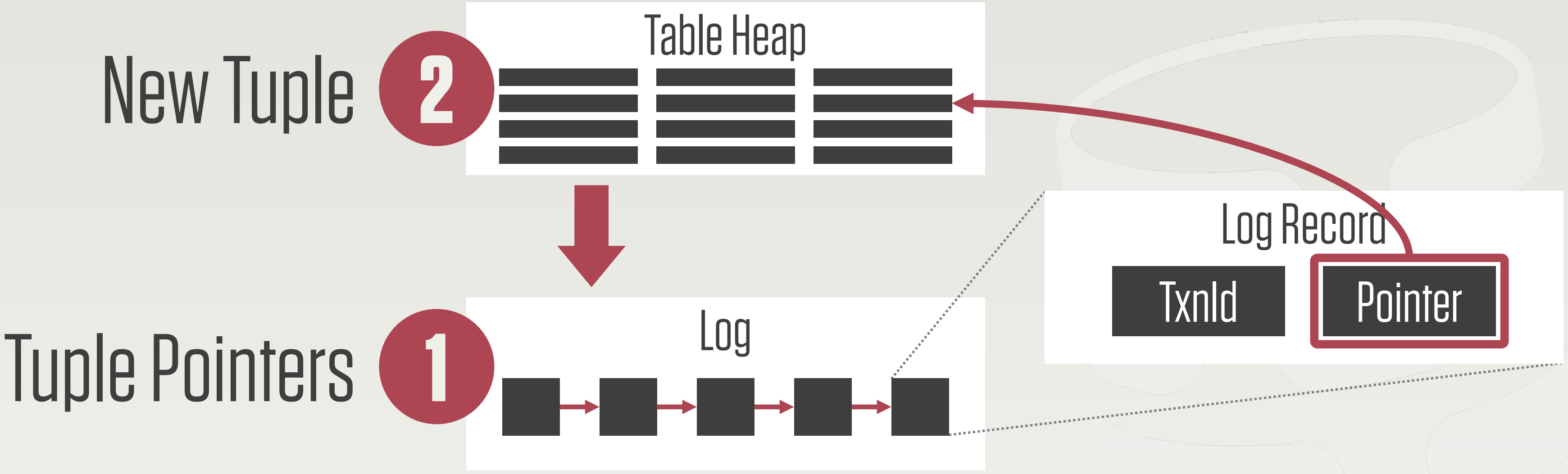


NVM-Optimized Architectures

- Use non-volatile pointers to only record <u>what</u> changed rather than <u>how</u> it changed.
- Be careful about how & when things get flushed from CPU caches to NVM.

NVM-Aware In-Place Engine

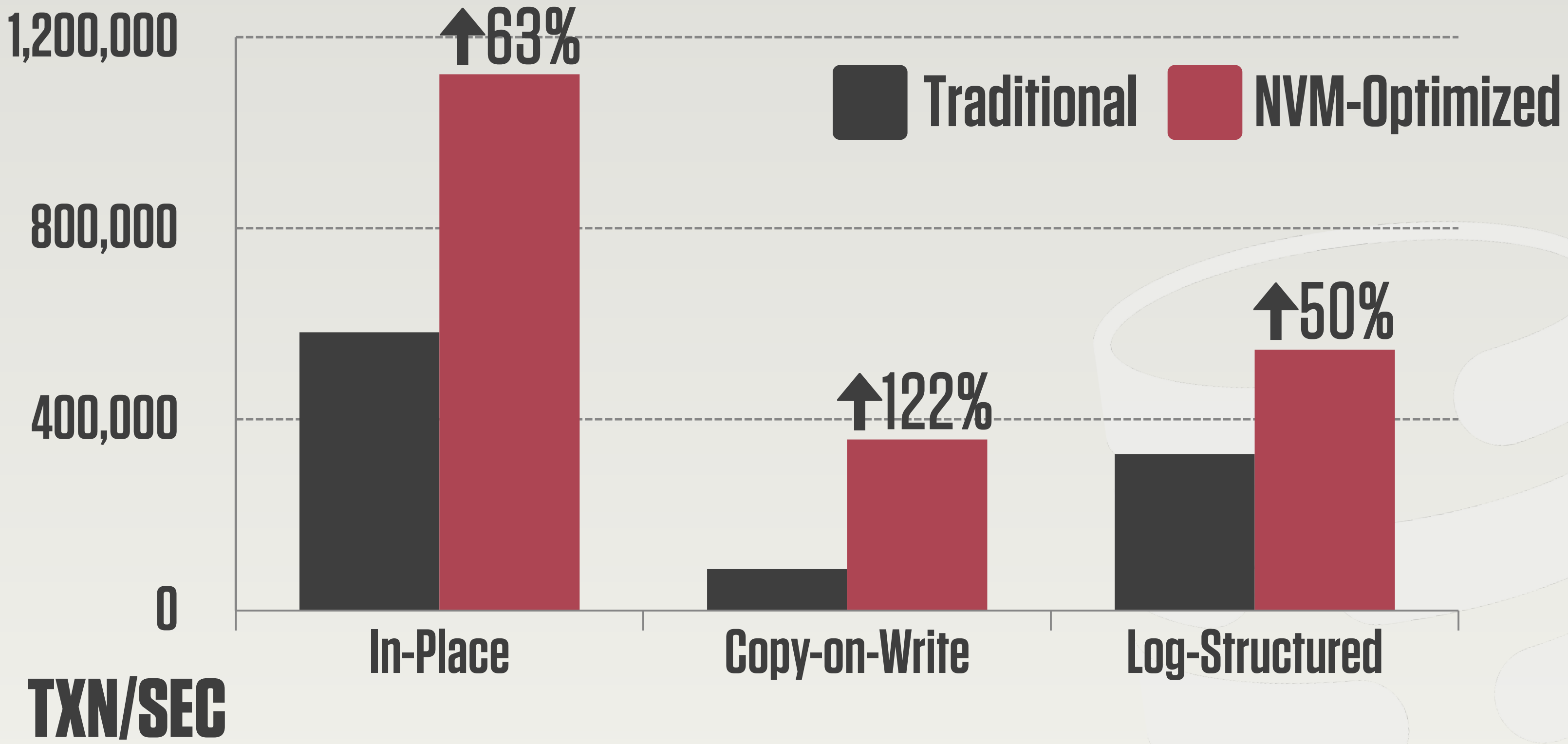
```
UPDATE table SET val=ABC
 WHERE id=123
```



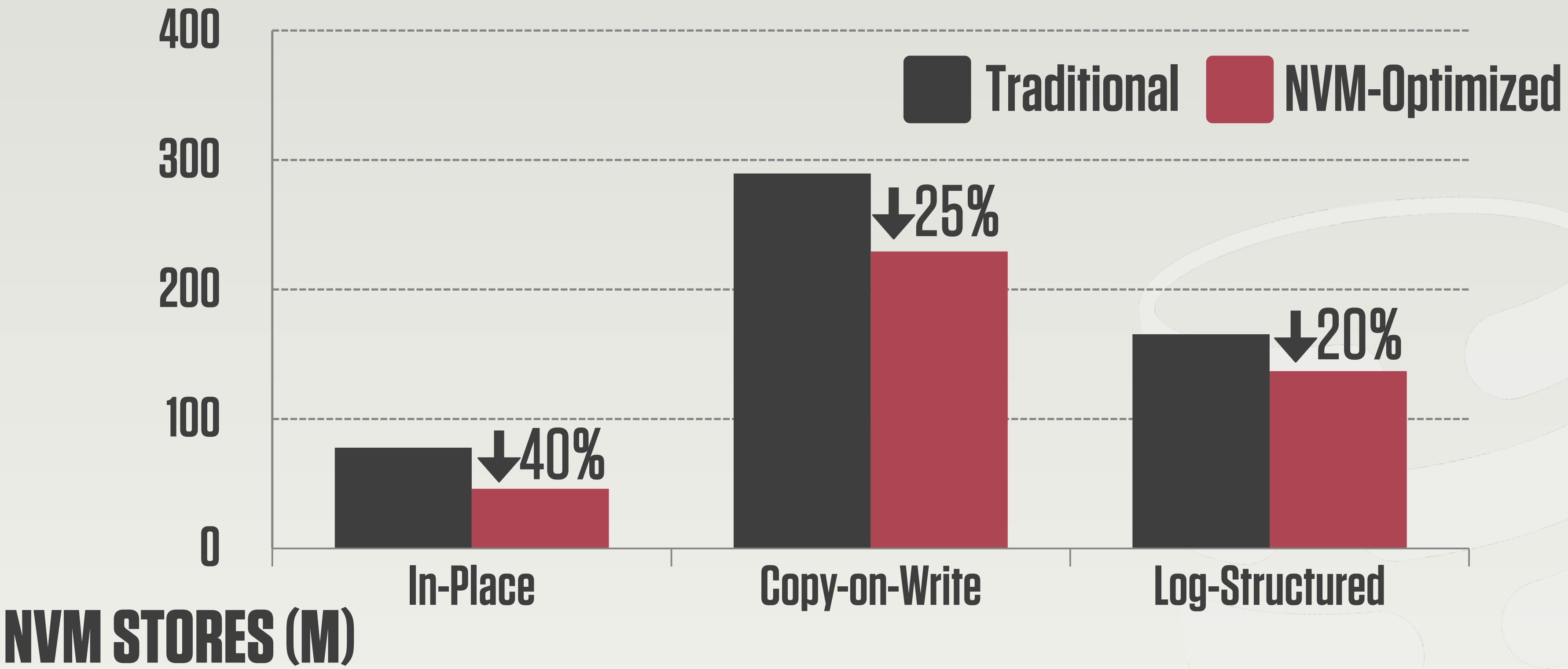
Evaluation

- Testbed system using the Intel NVM hardware emulator.
- Yahoo! Cloud Serving Benchmark
  - *2 million records + 1 million transactions*
  - *High-skew setting*

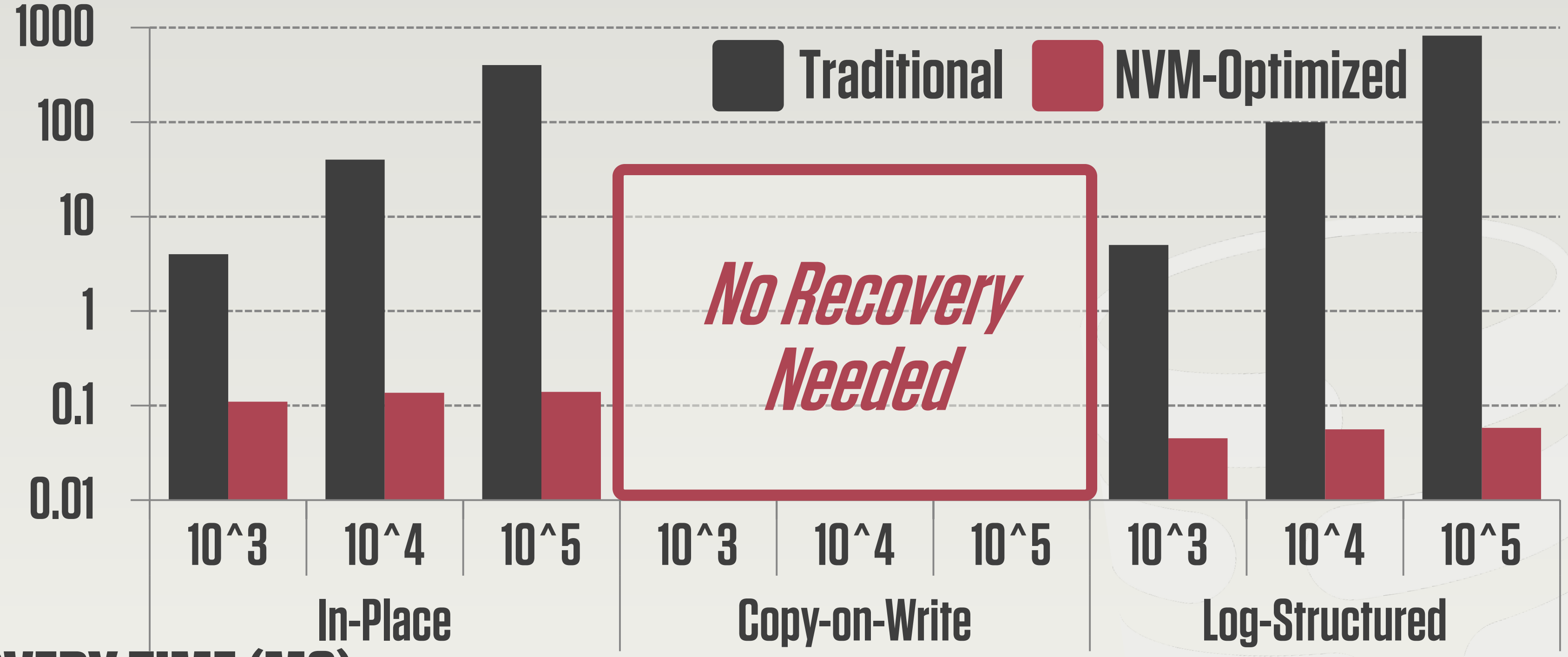
YCSB // *10% Reads / 90% Writes Workload* *2x Latency Relative to DRAM*



YCSB // 10% Reads / 90% Writes Workload
2x Latency Relative to DRAM



YCSB // *Elapsed time to replay log with varying log sizes*
2x Latency Relative to DRAM



RECOVERY TIME (MS)

LESSONS

1. Using NVM correctly improves throughput & reduces weadown.
2. Avoid block-oriented components.
3. NVM-only systems are 15-20 years away

What would **Nikita Kahn** do?


nikita kahn RHINO RESCUE CENTER

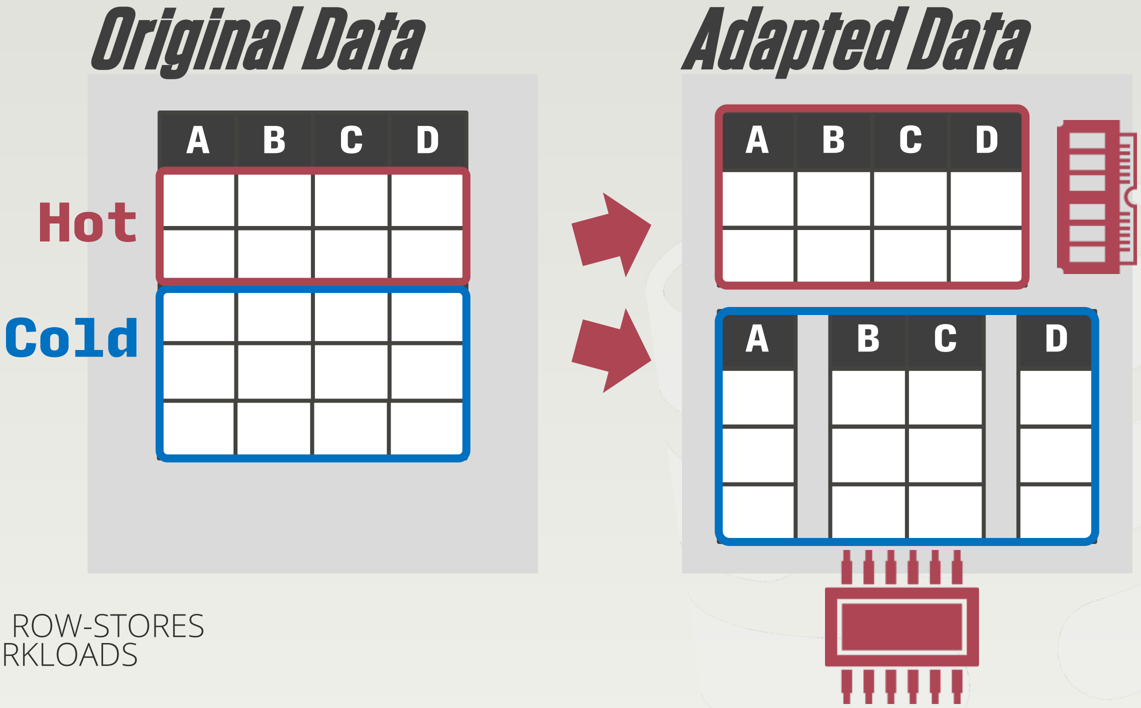
Chapter III – Hybrid DBMS

- Design and build a new in-memory DBMS that will be ready for NVM when it becomes available.
- Hybrid Storage + Hybrid Workloads
  - *DRAM + NVM oriented architecture*
  - *Fast Transactions + Real-time Analytics*

Adaptive Storage

```
UPDATE myTable
   SET A = 123,
       B = 456,
       C = 789
 WHERE D = “xxx”
```

```
SELECT AVG(B)
  FROM myTable
 WHERE C < “yyy”
```



BRIDGING THE ARCHIPELAGO BETWEEN ROW-STORES AND COLUMN-STORES FOR HYBRID WORKLOADS
SIGMOD 2016

LESSONS

Peloton

The Self-Driving Database Management System

NVM Ready

Query Compilation

Vectorized Execution

Autonomous

Apache Licensed

Anthony **Tomasic**

Todd **Mowry**

Joy **Arulraj**

Prashanth **Menon**

Lin **Ma**

Dana **Van Aken**

Michael **Zhang**

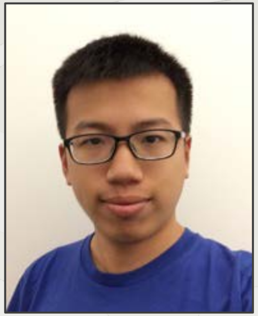
Matthew **Perron**

Yingjun **Wu**

Ran **Xian**

Runshen **Zhu**

Jiexi **Lin**



Jianhong **Li**

Ziqi **Wang**

http://pelotondb.org

END

@ANDY_PAVLO