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

@JOY_ARULRAJ

CARNEGIE MELLON DATABASE GROUP

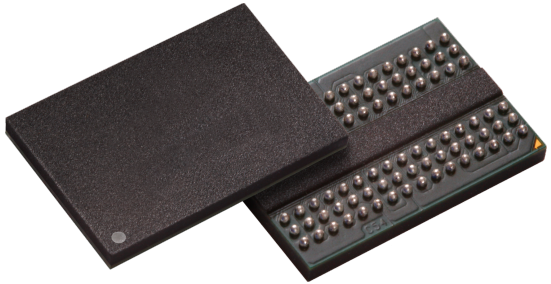
STORAGE LATENCY

RAMAC 350 (600 ms)	10^5x ↓	NAND SSD (60 us)
1956		2016

COMPUTE LATENCY

RAMAC 305 (100 Hz)	10^8x ↓	CORE I7 (1 GHZ)
1956	1000x	2016

NON-VOLATILE MEMORY



3D XPOINT
(60 ns)

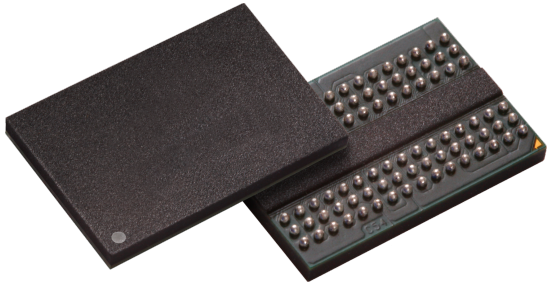
2017

1000x faster than NAND

10x denser than DRAM

1000x endurance of NAND

NON-VOLATILE MEMORY



3D XPOINT
(60 ns)

2017

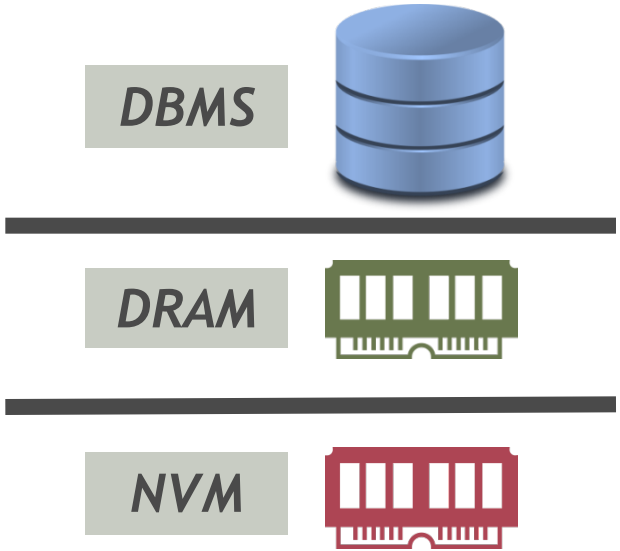
Support in Linux 4.3+

New assembly instructions

SNIA programming model

WHAT NVM MEANS FOR DATABASES?

- Option 1: Treat NVM like a faster SSD

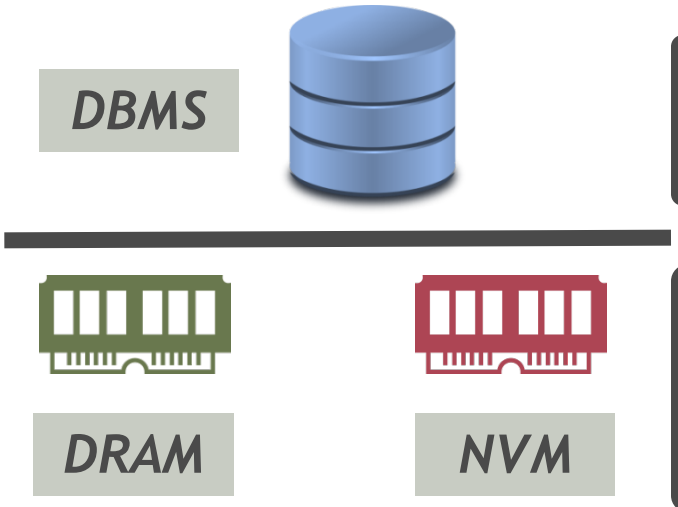


Use NVM as a logging and backing store

Improves performance

WHAT NVM MEANS FOR DATABASES?

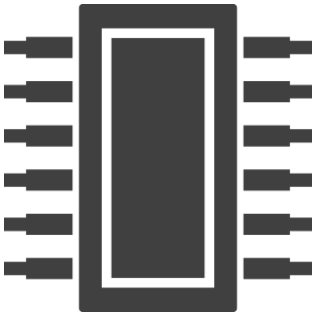
- Option 2: Treat NVM as extended memory



Redesign the key algorithms in a database system

Improves not only performance, but also availability and device utilization

PAST:
EXISTING
SYSTEMS

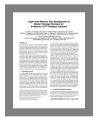


PRESENT:
NVM-AWARE
DBMS

FUTURE:
ANALYTICS
ON NVM

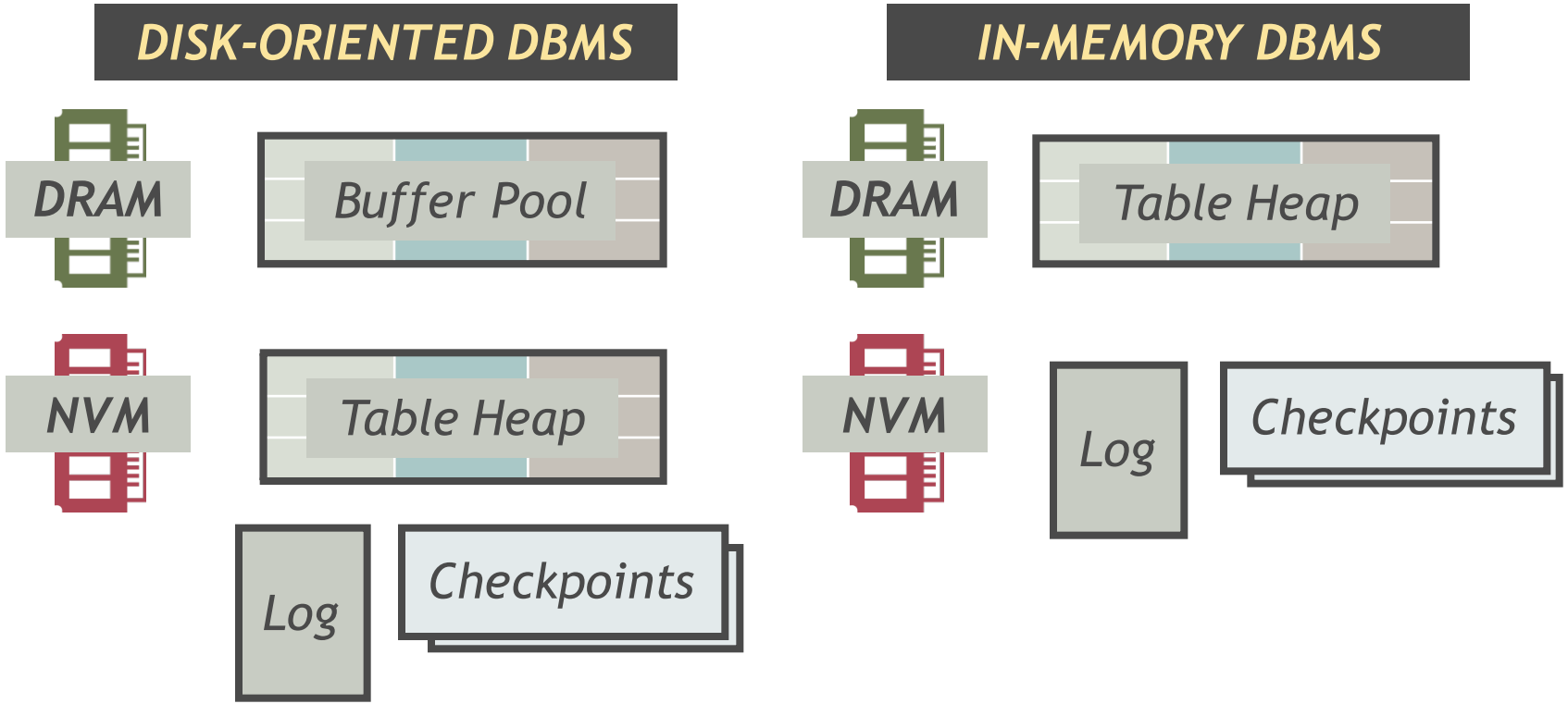
PAST — EXISTING SYSTEMS

- Investigate how existing systems perform on NVM
  - *Option 1: Treat NVM like a faster SSD*
- Evaluate two types of DBMSs
  - *Disk-oriented DBMS*
  - *In-memory DBMS*



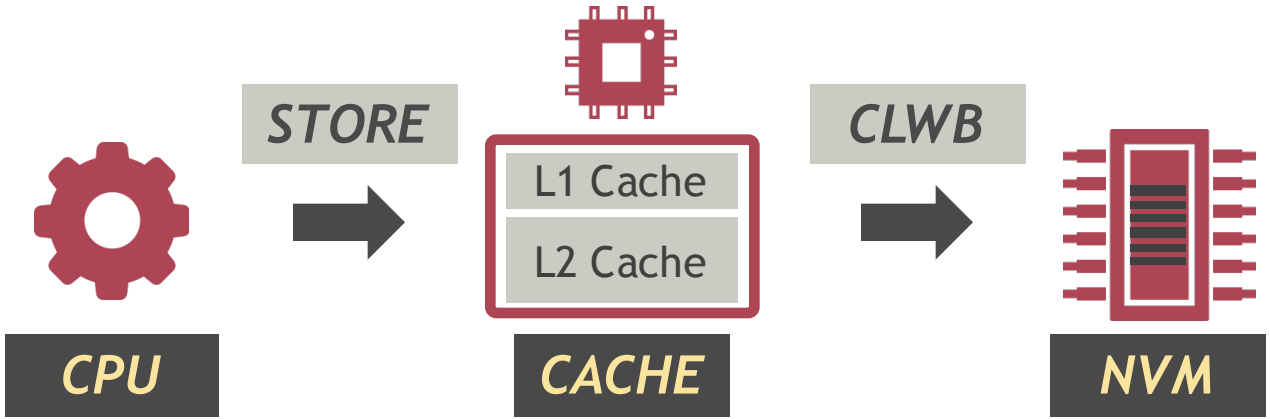
A PROLEGOMENON ON OLTP DATABASE SYSTEMS
FOR NON-VOLATILE MEMORY
ADMS 2014

DBMS ARCHITECTURES



NVM HARDWARE EMULATOR

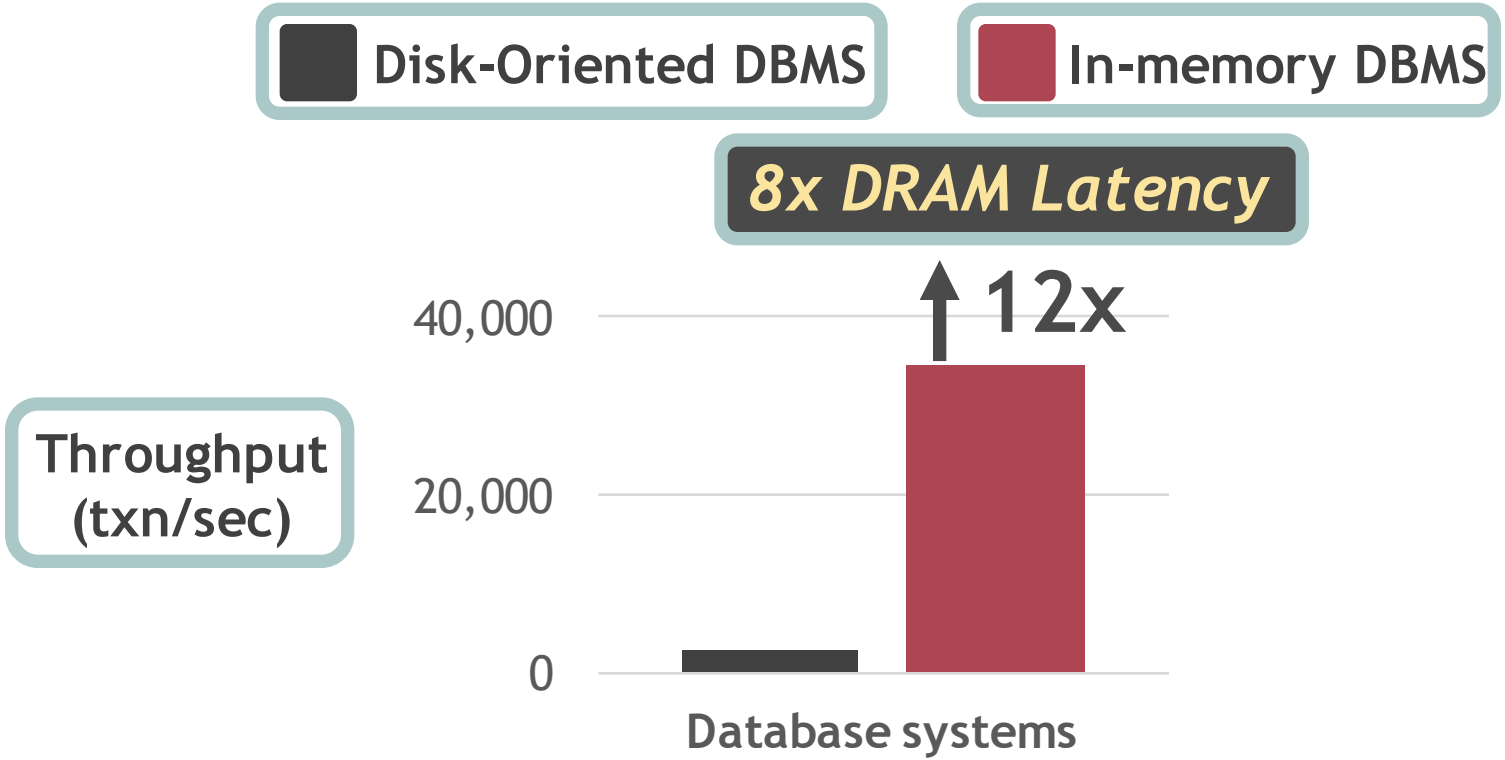
- Tunable DRAM latency for emulating NVM
- Special assembly instructions for NVM
  - *Cache line write-back*



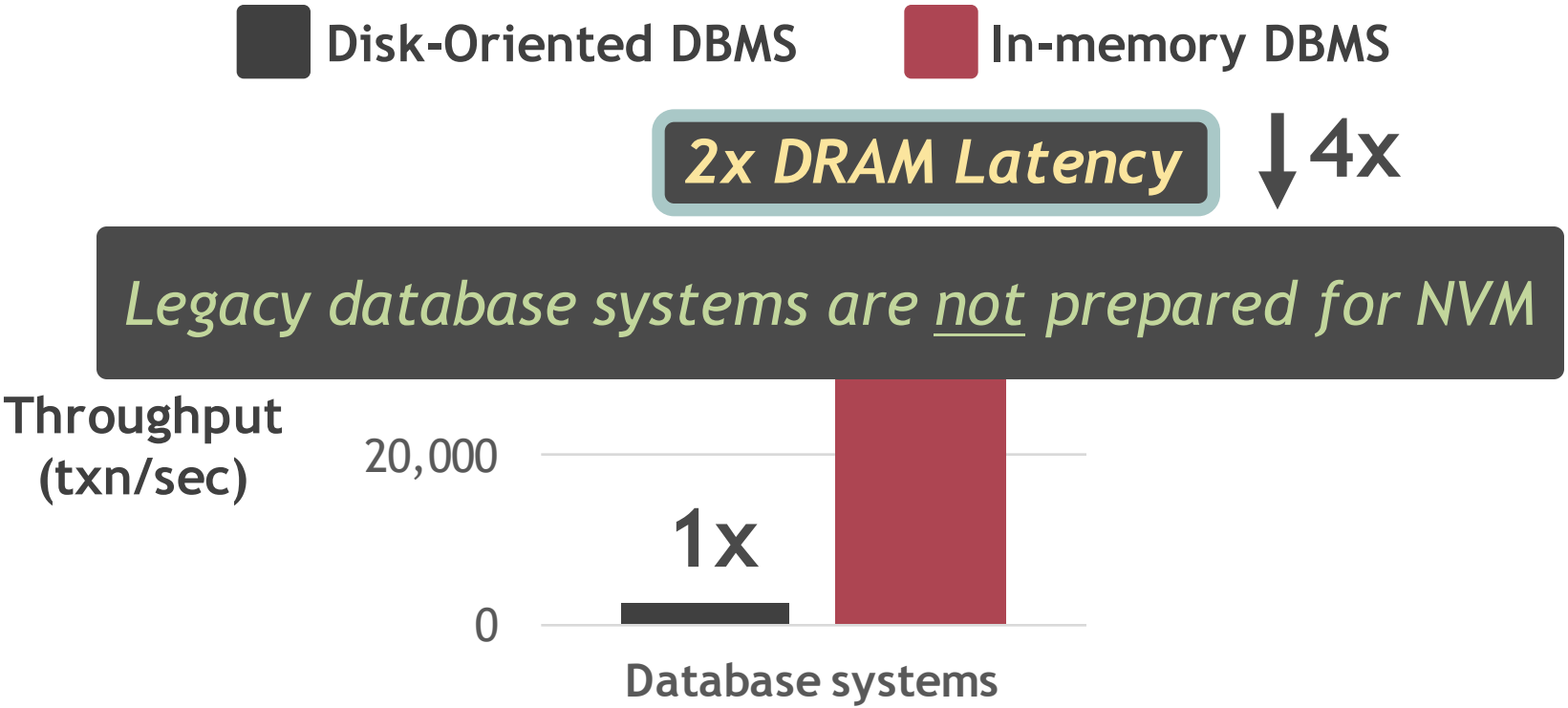
EVALUATION

- Compare existing DBMSs on NVM emulator
  - *MySQL (Disk-oriented DBMS)*
  - *H-Store (In-memory DBMS)*
- TPC-C benchmark
  - *1/8th of database fits in DRAM, rest on NVM*

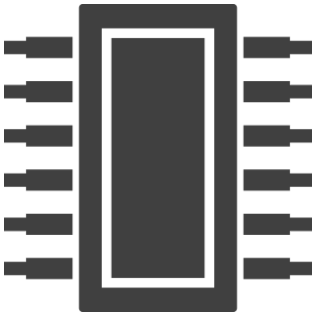
PERFORMANCE



PERFORMANCE



PAST:
EXISTING
SYSTEMS

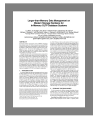


PRESENT:
NVM-AWARE
DBMS

FUTURE:
ANALYTICS
ON NVM

PRESENT — NVM-AWARE DBMS

- Understand changes required in DBMSs to leverage NVM
  - *Option 2: Treat NVM as extended memory*
- Rethink key algorithms in database systems
  - *Logging and recovery protocol*
- Designed for real-time analytics
  - *Multi-versioned database*



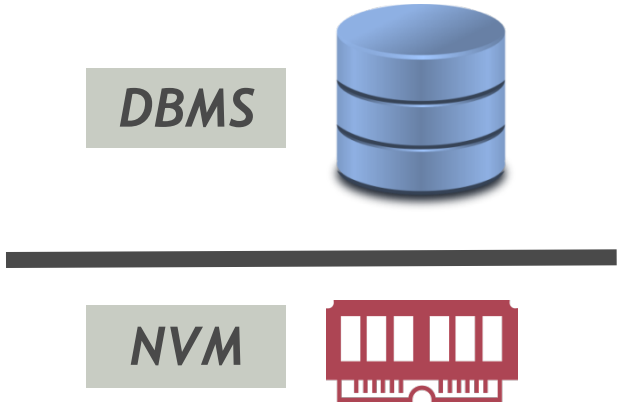
LET'S TALK ABOUT STORAGE AND RECOVERY METHODS FOR NON-VOLATILE MEMORY DATABASE SYSTEMS
SIGMOD 2015

MULTI-VERSIONED DBMS

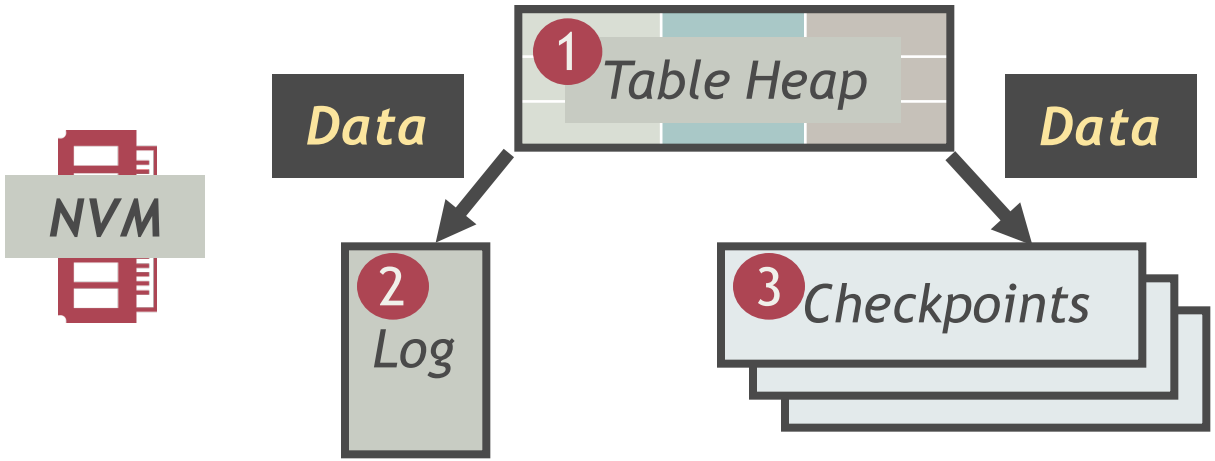
TUPLE ID	BEGIN TIMESTAMP	END TIMESTAMP	PREVIOUS VERSION	TUPLE DATA
1	10	∞	—	X
2	10	20	—	Y
3	20	∞	2	Y’

THOUGHT EXPERIMENT

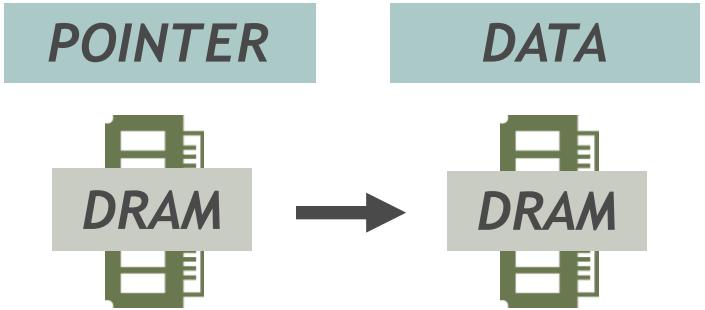
- NVM-only storage hierarchy
  - *No volatile DRAM*

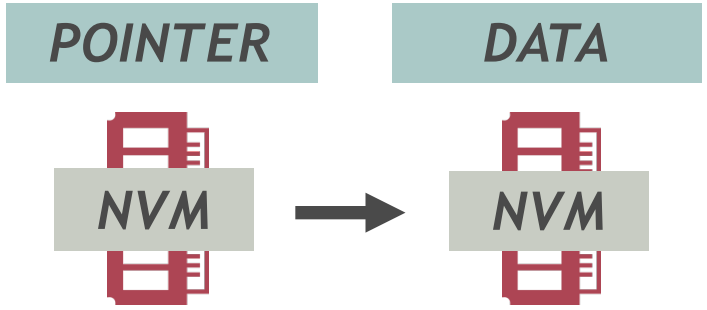


TRADITIONAL STORAGE ENGINE



NON-VOLATILE POINTER





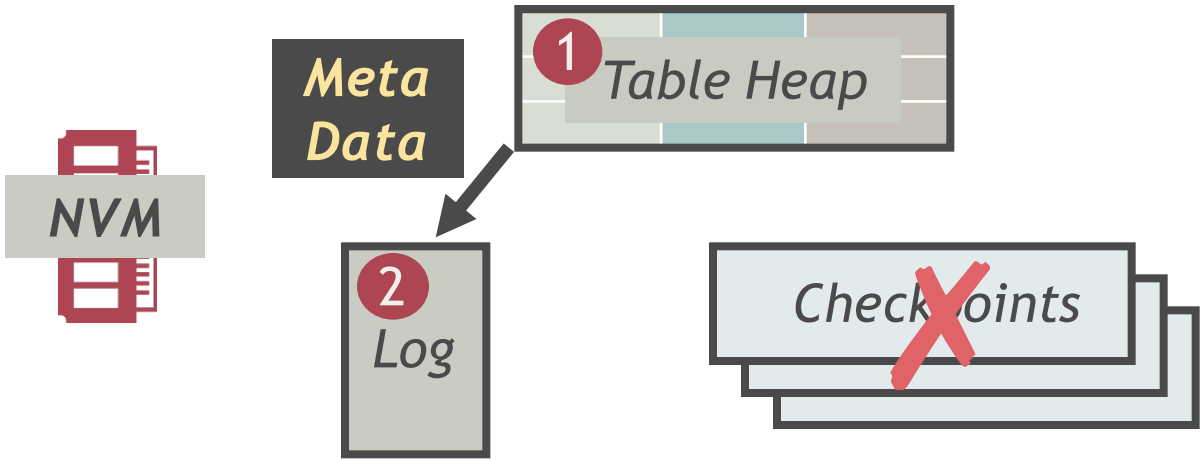
NVM-AWARE STORAGE ENGINE

- Instead of duplicating tuple data in log
  - *Store non-volatile tuple pointers in log records*

TRADITIONAL LOG
INSERT TUPLE 1 (DATA)
UPDATE TUPLE 1 (DATA)

NVM-AWARE LOG
INSERT TUPLE 1 (POINTER)
UPDATE TUPLE 1 (POINTER)

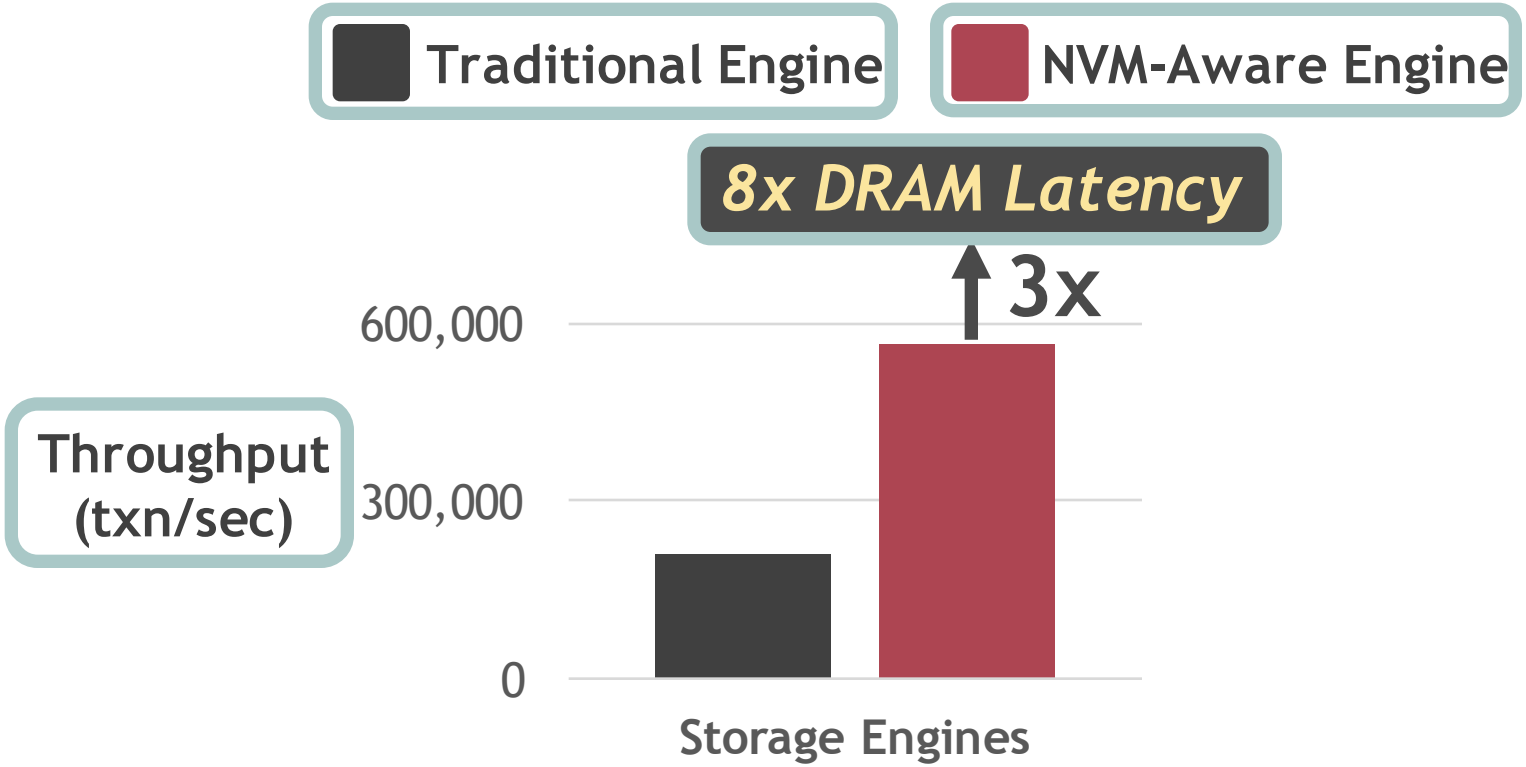
NVM-AWARE STORAGE ENGINE



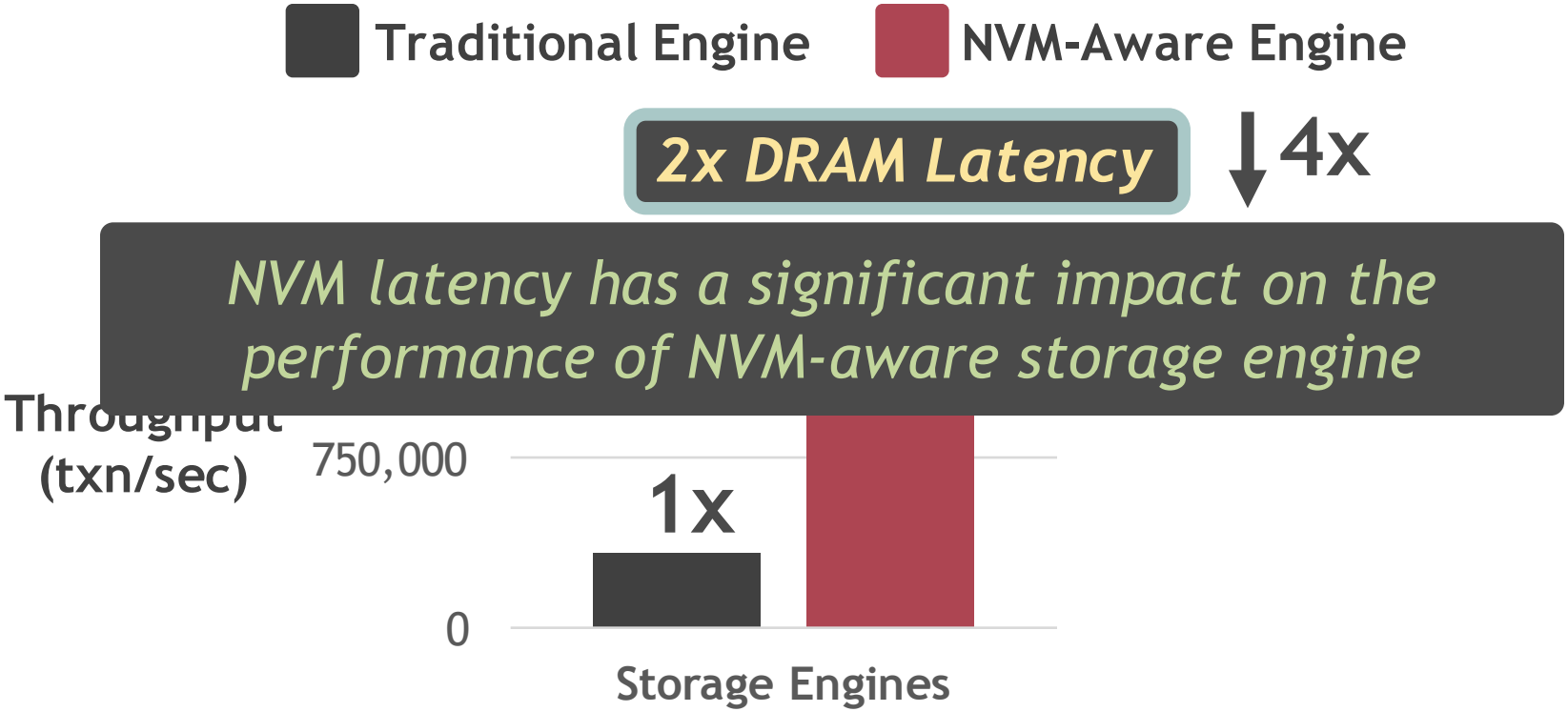
EVALUATION

- Compare storage engines on NVM emulator
  - *Traditional engine*
  - *NVM-aware engine*
- Yahoo! Cloud Serving Benchmark
  - *Database fits in NVM*

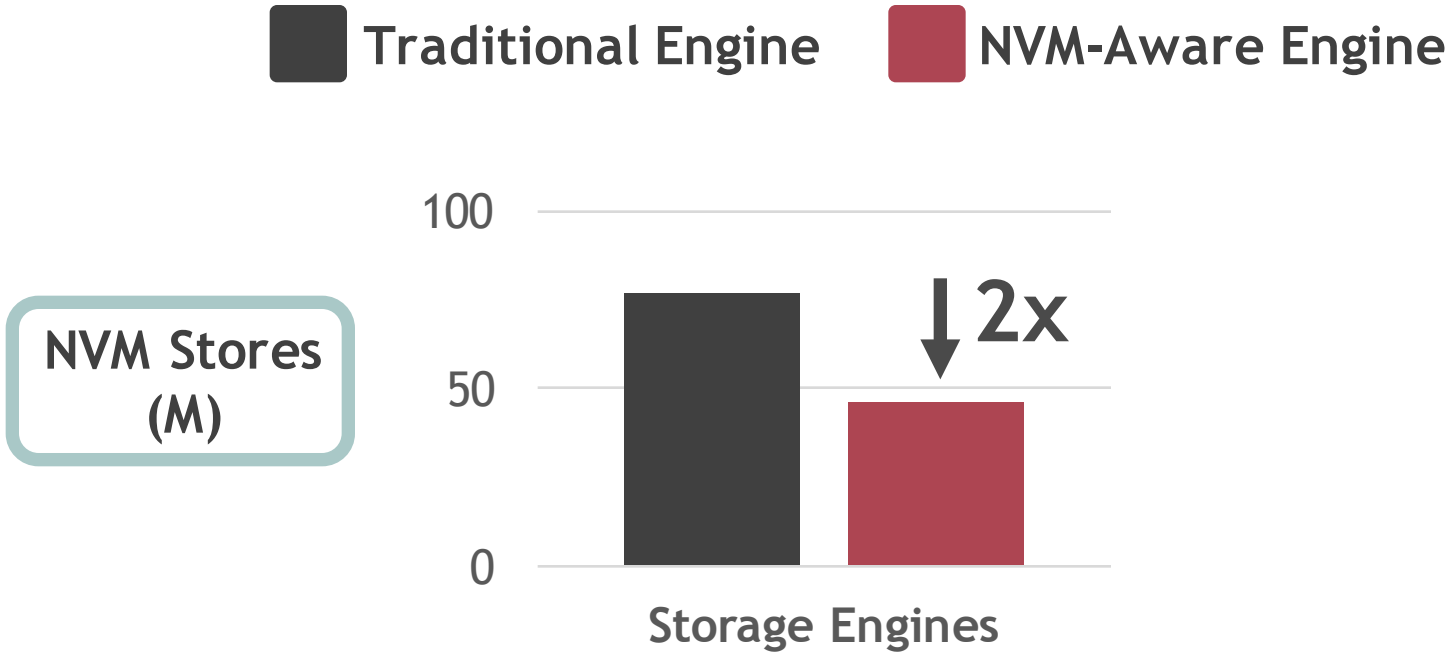
PERFORMANCE



PERFORMANCE

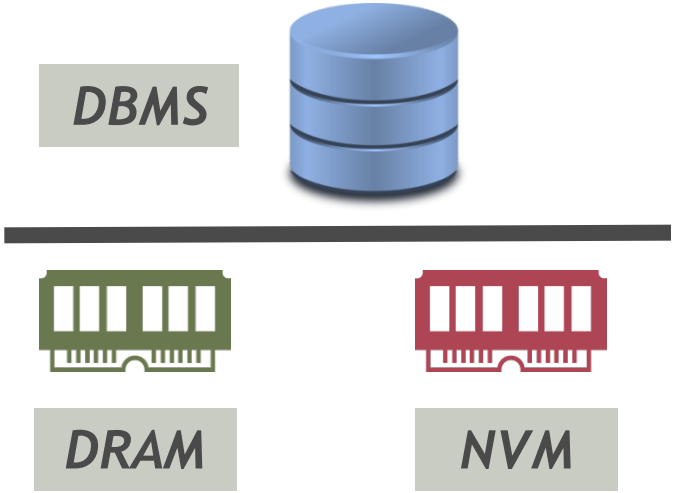


DEVICE LIFETIME



THE PELOTON DBMS

- A new database system for NVM research
  - *Designed for a two-tier storage hierarchy*

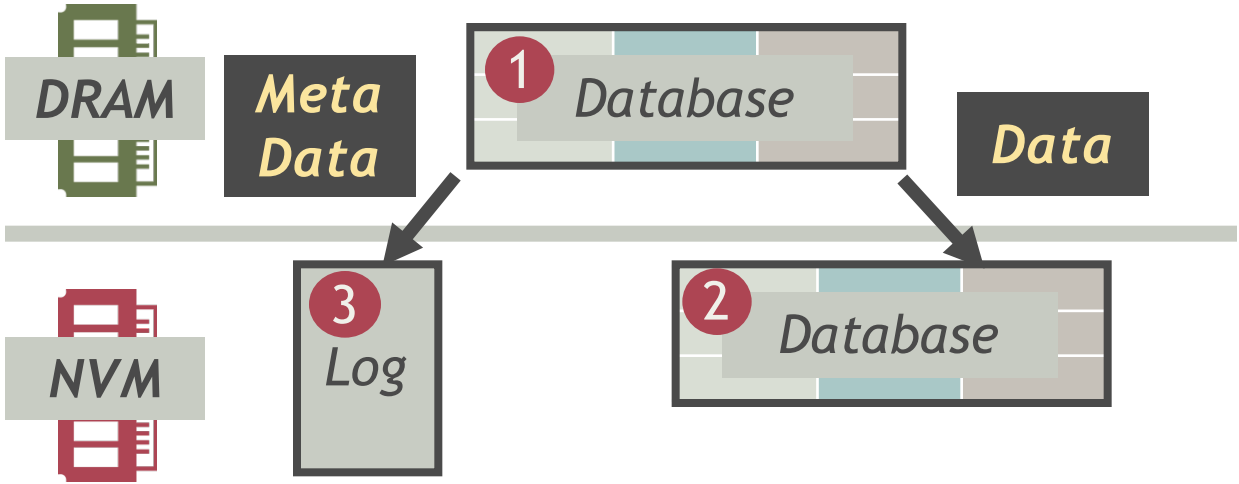


WRITE-BEHIND LOGGING
UNDER REVIEW

WRITE-BEHIND LOGGING

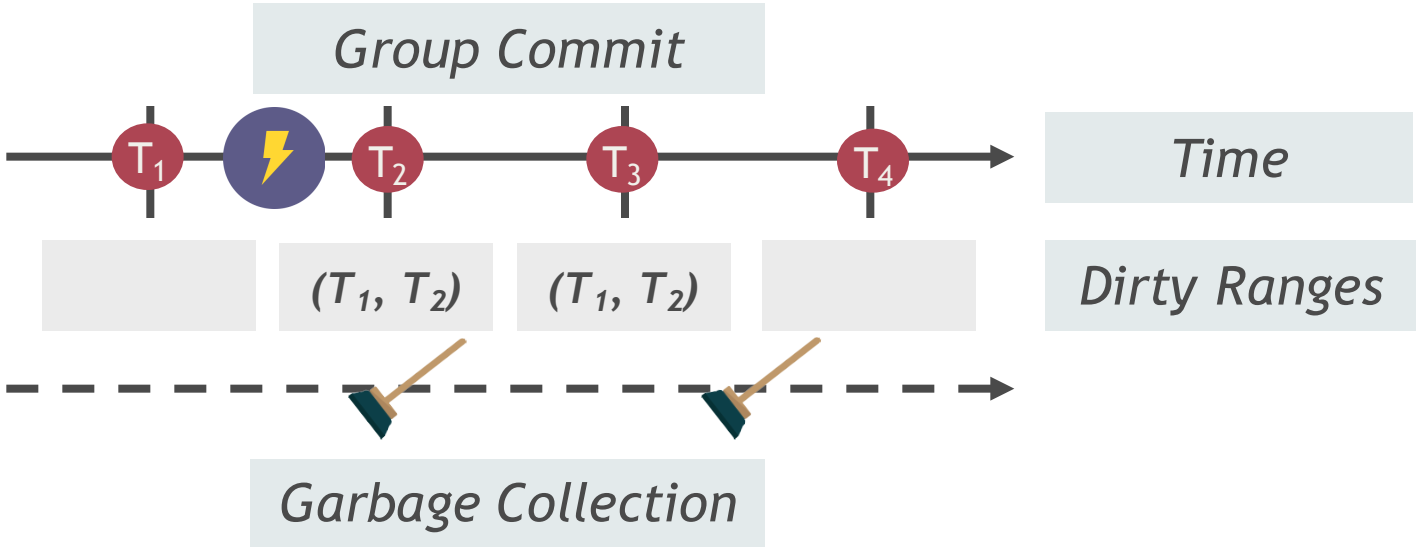
- Write-ahead log serves two purposes
  - *Transform random database writes into sequential log writes*
  - *Support transaction rollback*
- NVM supports fast random writes
  - *Directly write data to the database*
  - *Later, record metadata in write-behind log*

WRITE-BEHIND LOGGING



METADATA FOR INSTANT RECOVERY

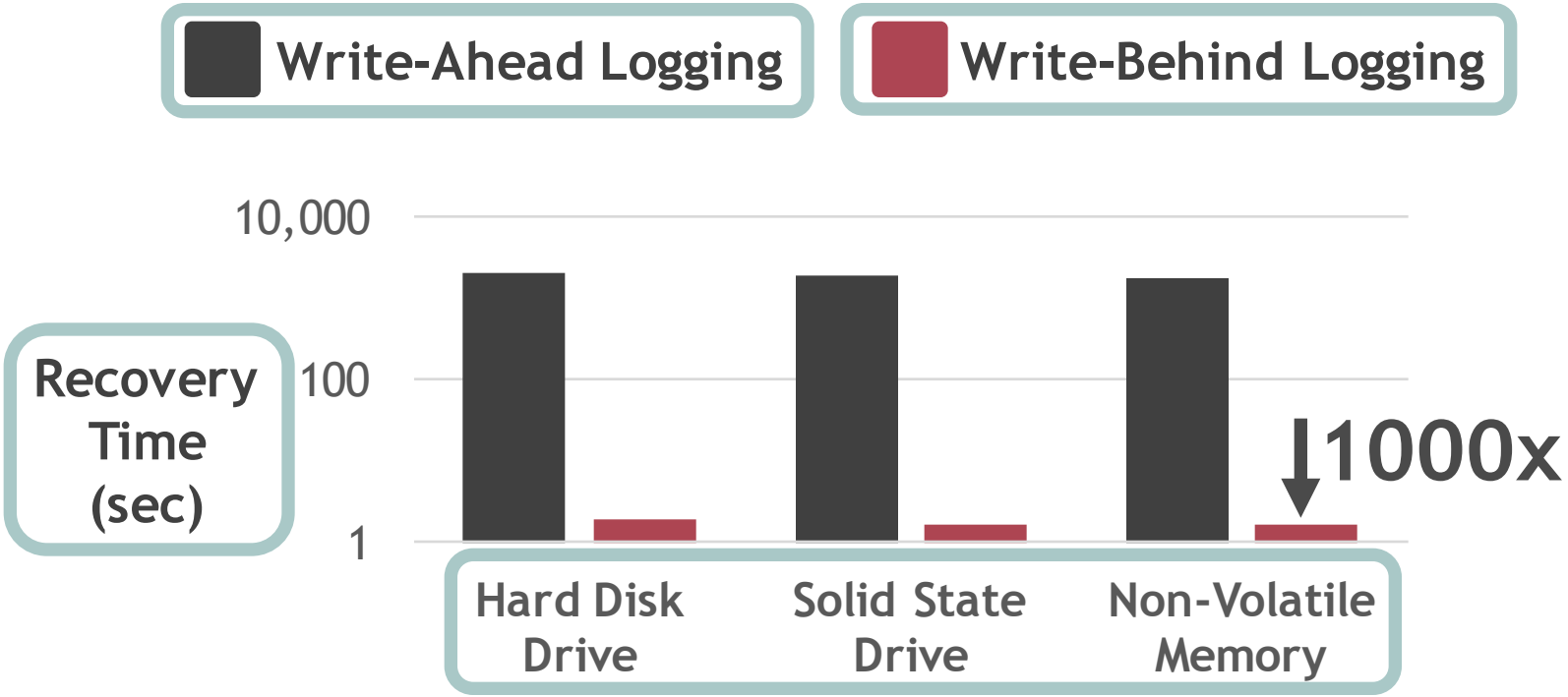
- Record failed timestamp ranges in log
  - *Use it to ignore effects of uncommitted transactions*



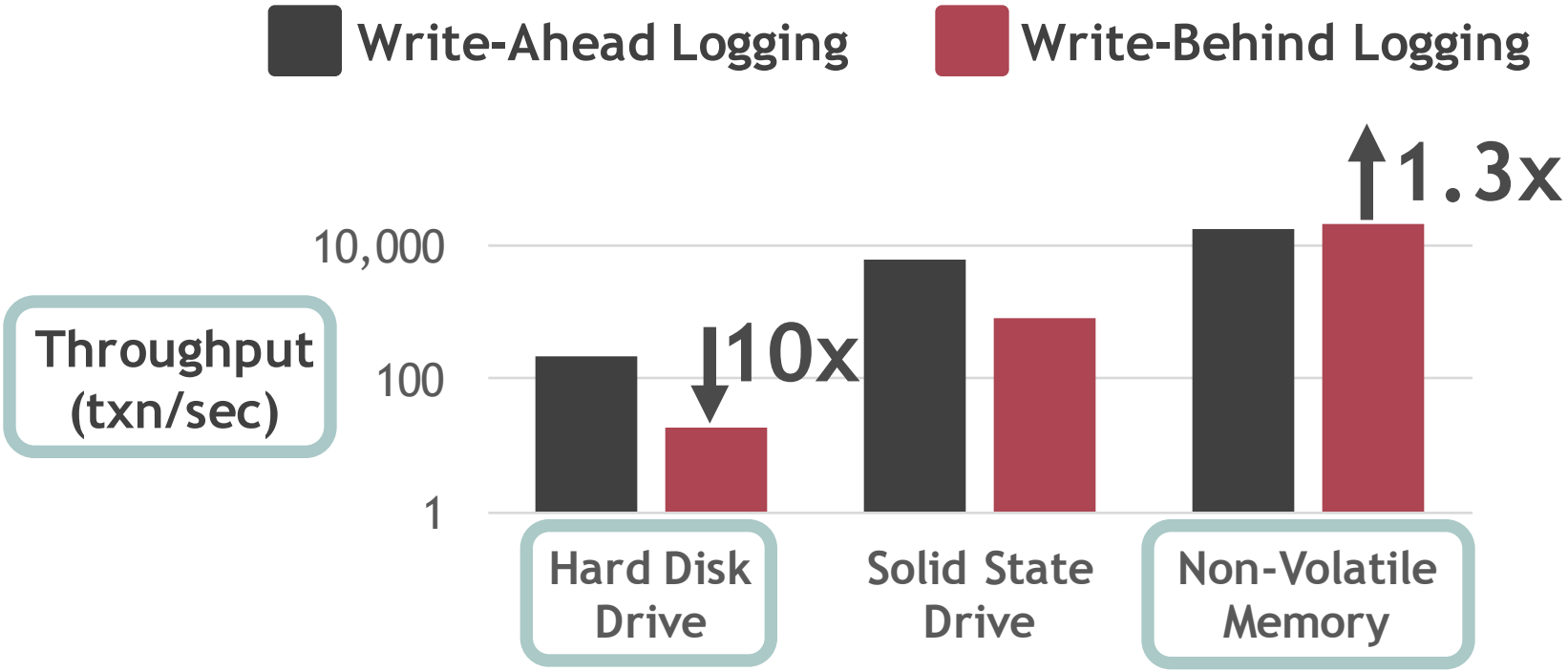
EVALUATION

- Compare logging protocols in Peloton
  - *Write-Ahead logging*
  - *Write-Behind logging*
- Storage devices in Intel's NVM hardware emulator
  - *Hard-disk drive*
  - *Solid-state drive*
  - *Non-volatile memory emulator*
- TPC-C benchmark

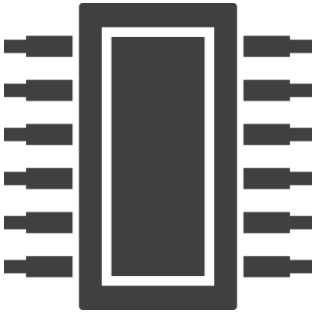
APPLICATION AVAILABILITY



PERFORMANCE



PAST:
EXISTING
SYSTEMS



PRESENT:
NVM-AWARE
DBMS

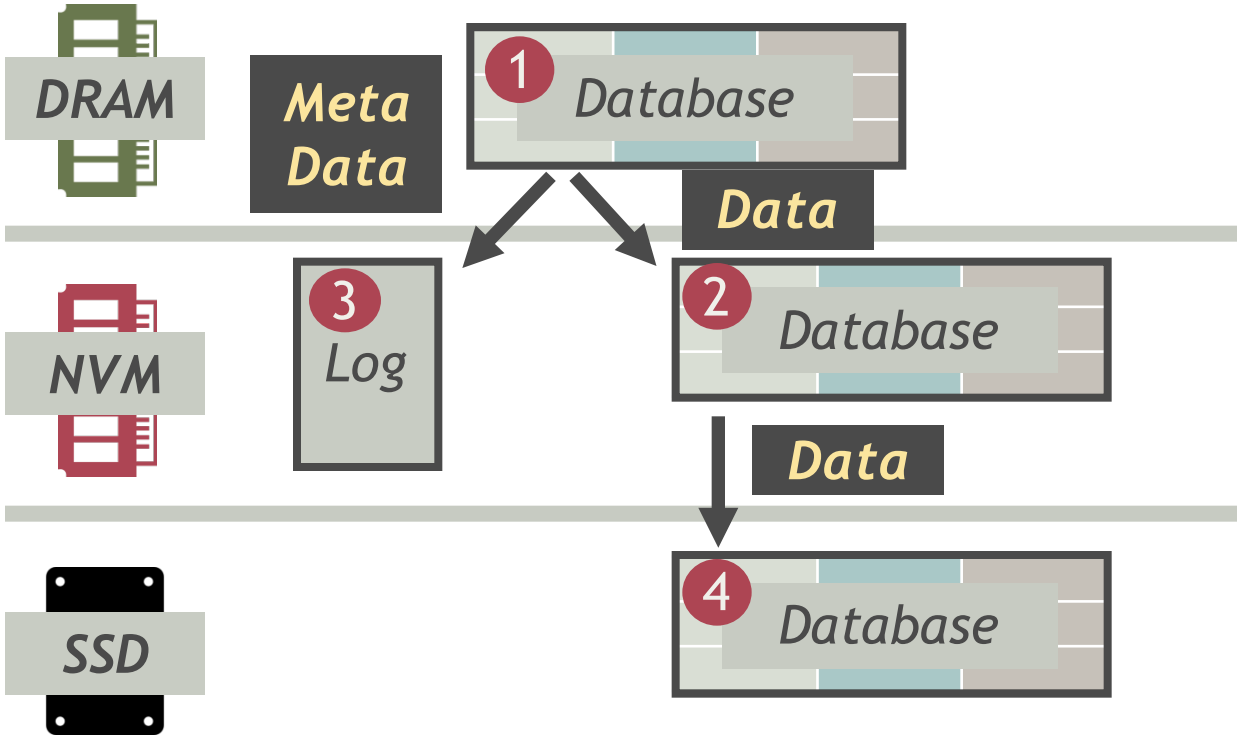
FUTURE:
ANALYTICS
ON NVM

FUTURE — REAL-TIME ANALYTICS

- Support analytics on larger-than-NVM databases
  - *Cost of first-generation NVM devices*
- Three-tier storage hierarchy
  - *DRAM + NVM + SSD*
- When to migrate data between different layers?

REAL-TIME ANALYTICS IN NON-VOLATILE MEMORY DATABASE SYSTEMS
WORK IN PROGRESS

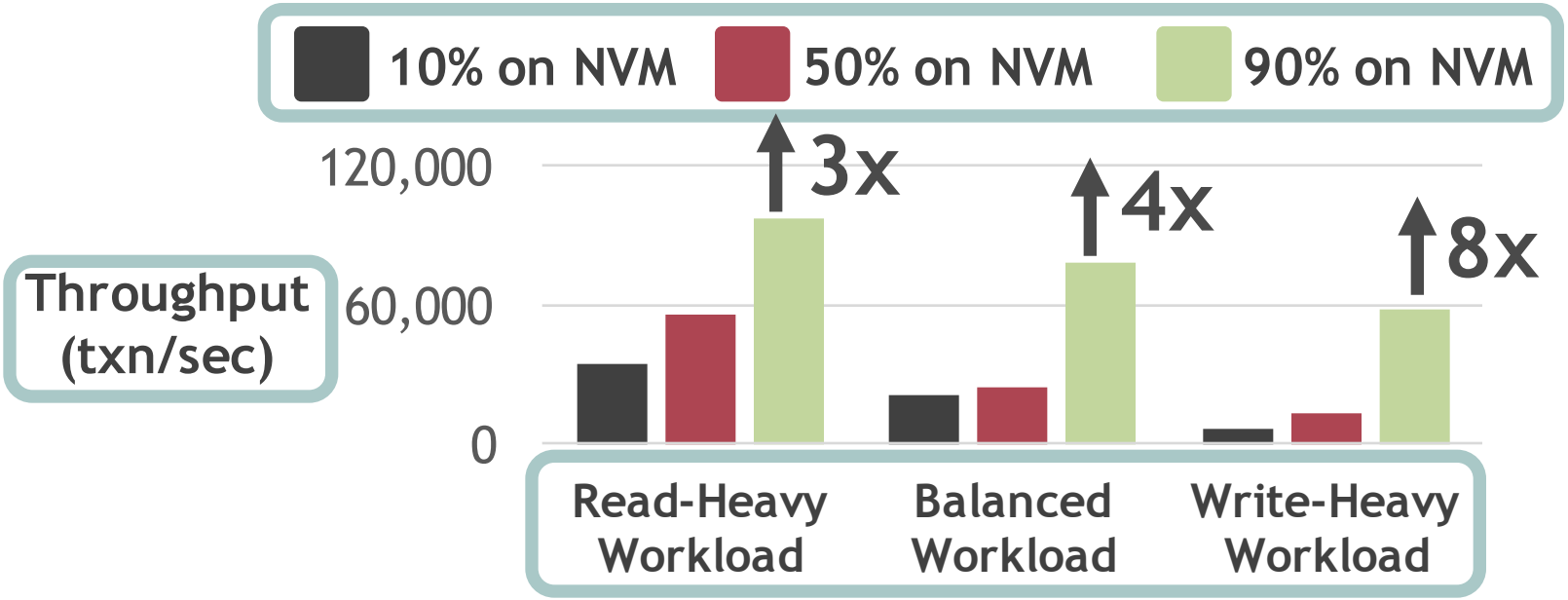
THREE-TIER STORAGE HIERARCHY



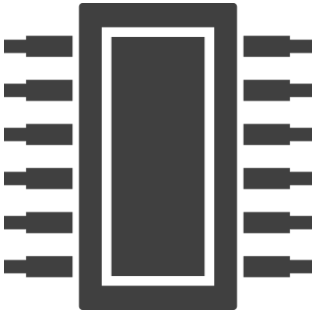
DATA MIGRATION

- Can directly read warm data from NVM
  - *No need to copy data over to DRAM for reading*
- Keep hottest data in DRAM
- Dynamically migrate cold data to SSD

THREE-TIER STORAGE HIERARCHY



PAST:
EXISTING
SYSTEMS



PRESENT:
NVM-AWARE
DBMS

FUTURE:
ANALYTICS
ON NVM

LESSONS LEARNED

- NVM outperforms SSD when correctly used
- Rethink key algorithms in database systems
  - *Write-behind logging enables instant recovery*
  - *Improves device utilization and extends its lifetime*
- Ongoing work
  - *Data placement policy*
  - *Replication*

PELOTON

http://pelotondb.org

NVM Ready

Autonomous

Apache Licensed

END

@joy_arulraj