Databases Acceleration with Non Volatile Memory File System (NVMFS)

Saeed Raja

SanDisk Inc.

MySQL?

- Widely used **Open Source** Relational Database Management System (**RDBMS**)
- Popular choice of database for use in web applications**,** OLTP, embedded database

- **Oracle MySQL (Stockholm, Sweden)**
  - All other MySQL forks are based of Oracles MySQL releases
  - InnoDB storage engine

- **Percona Server** (Pleasanton, California, USA)
  - Developer of XtraDB storage engine

- **MariaDB** (Helsinki, Finland)
  - Founded by Monty, MySQL original author, MariaDB foundation
  - Uses XtraDB, joint development with Percona

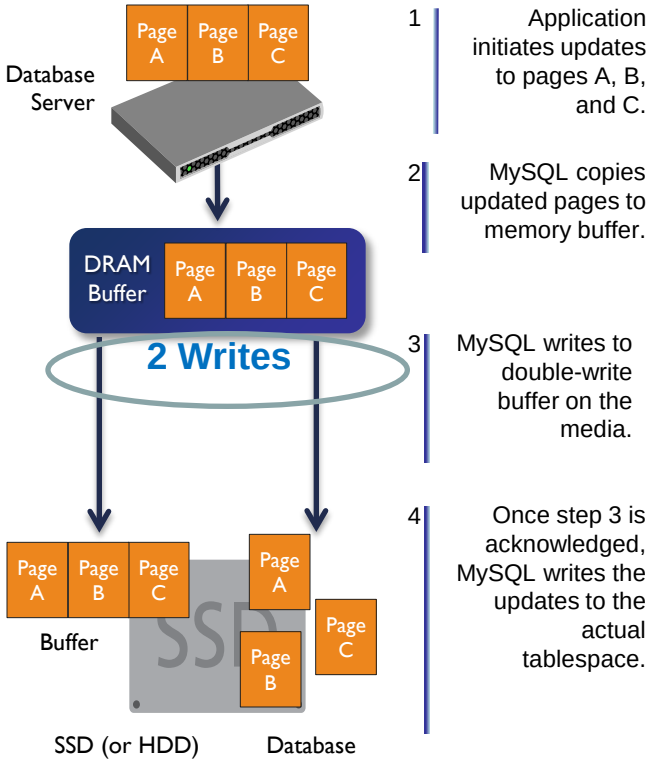
MySQL Has Strong Momentum!!!

- **Leading open source database for Web applications**
- **#1 Open Source Database in the Cloud[1]**
  - dbPaaS market is gaining momentum[2]
  - Amazon RDS offer Oracle MySQL RDBMS engine[2]
  - Rackspace Cloud Databases offer fully managed instances of MariaDB, MySQL and Percona, with container-based virtualization[2]
- **Integrated with Hadoop in big data platforms**

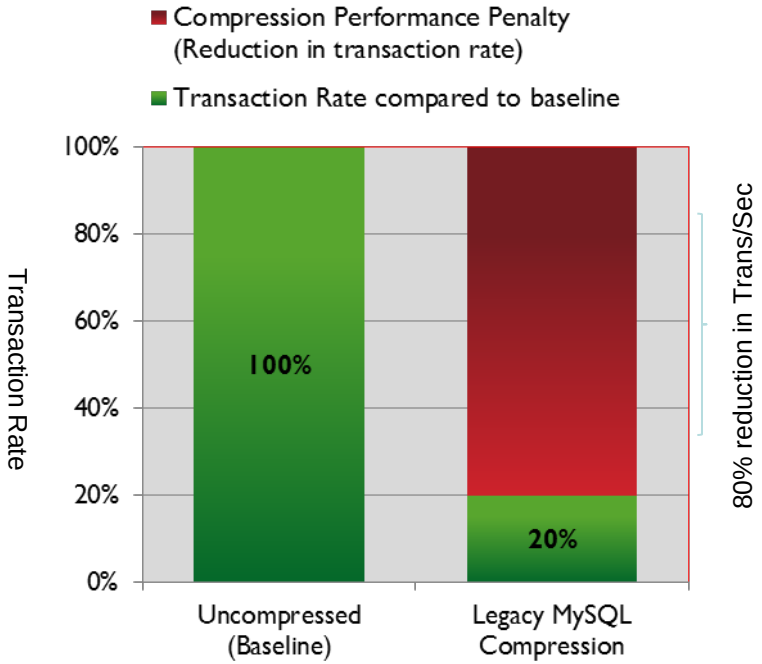
[1]Oracle: "State of the Dolphin" Keynote - MySQL Central @ OpenWorld 2014
[2]Gartner: Market Guide for Database Platform as a Service

Legacy MySQL Challenges

1 Every MySQL write translates to **2 writes** to SSD

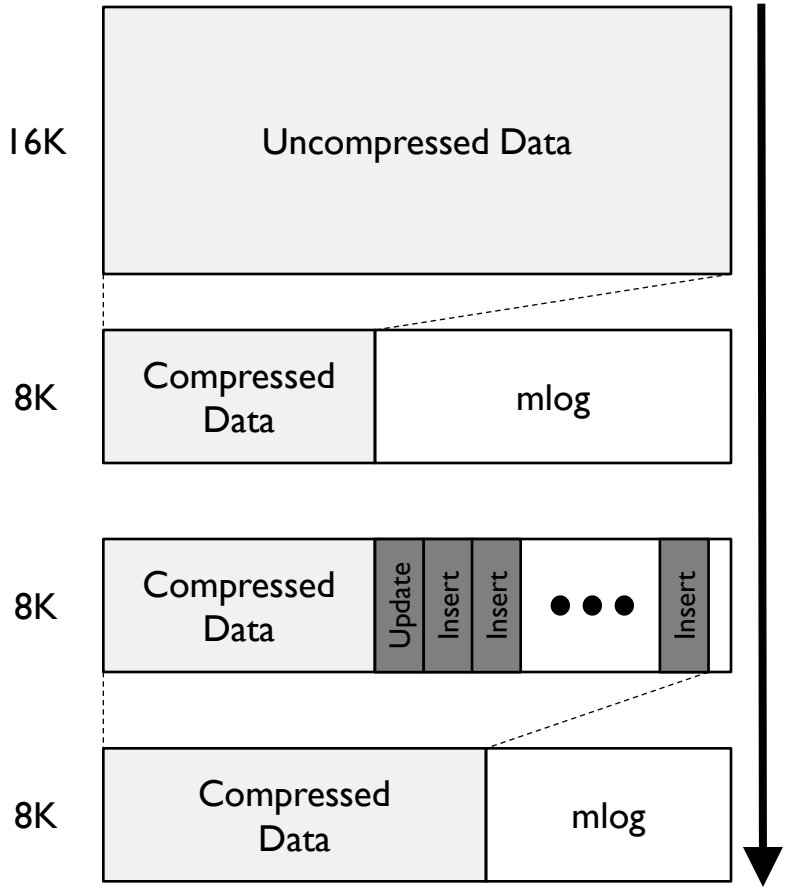


2 80% Performance penalty with legacy MySQL compression on



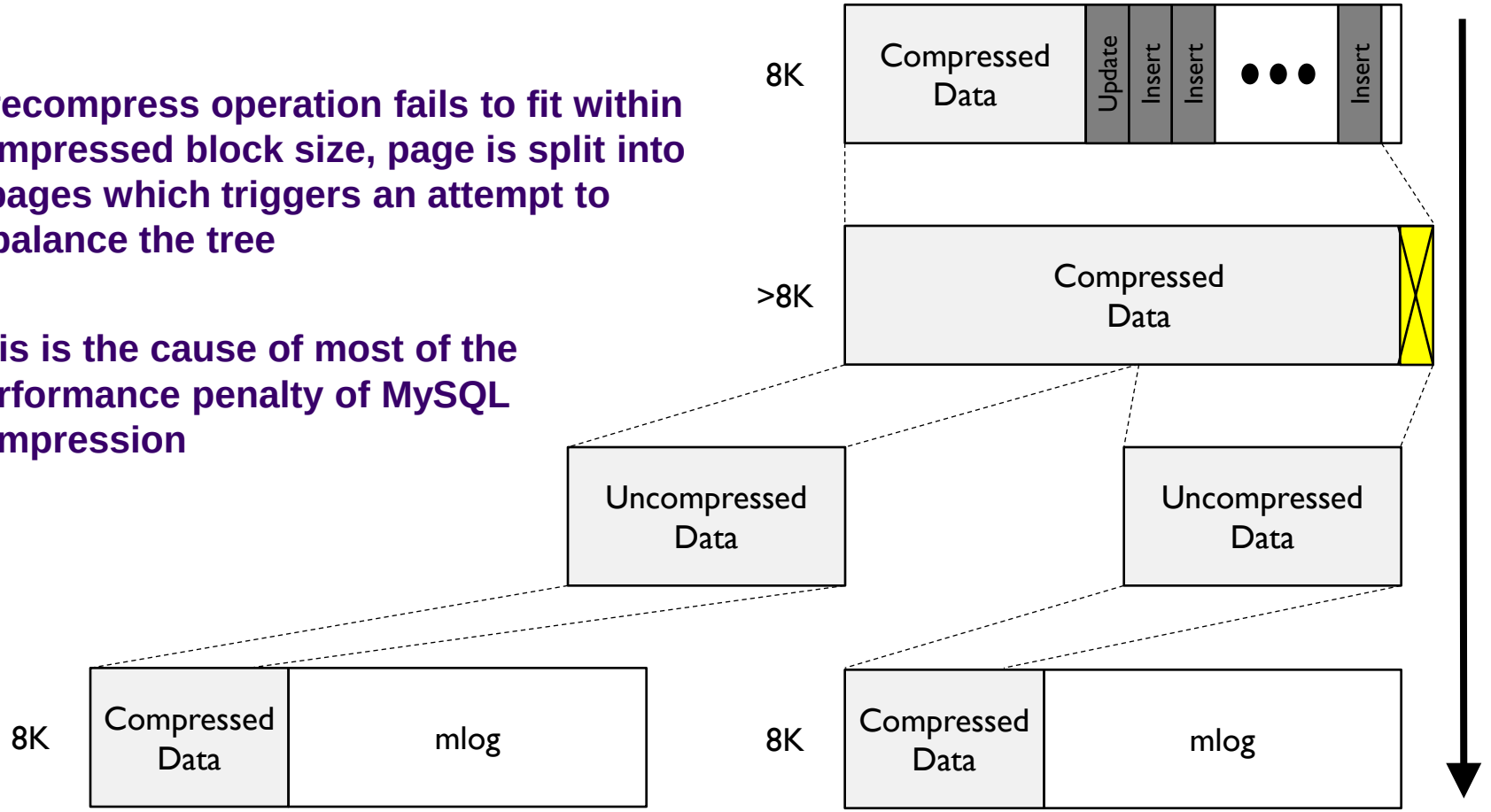
Legacy MySQL Compression

- MySQL stores uncompressed data in 16K pages
- 16K pages are compressed into a fixed compressed page size of 1K, 2K, 4K, 8K
- Compressed page size is chosen at table creation
- Compression is performed using regular software compression libraries (zlib)
- Table updates appended to Page Modification Log (mlog) at the end of the compressed (8K) page
- When mlog gets full, page is recompressed



Fail – Split – Rebalance – Recompress

- **If recompress operation fails to fit within compressed block size, page is split into 2 pages which triggers an attempt to rebalance the tree**
- **This is the cause of most of the performance penalty of MySQL compression**



New Primitives for a New Type of Media

Application and File Systems Lagging Behind

Tape	Open, read, write, rewind, close.
Disk	Open, read, write, seek, close.
SSD	Open, read, write, seek, close.
Fusion ioMemory NVM	Open, read, write, seek, close. *Plus, new primitives to exploit characteristics of non-volatile memory* Basic write + atomic write, Transactional write. Persistent Trim

SanDisk NVMFS

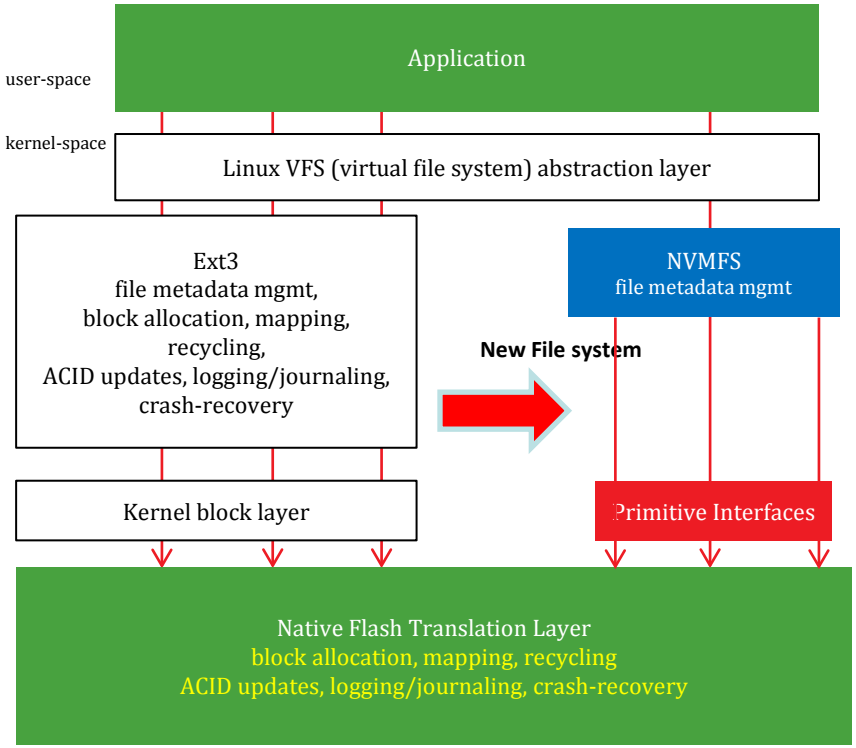
Eliminating Duplicate Logic & Leverage New Primitives for Optimal Flash Performance & Efficiency

Value

- Increase life expectancy of flash devices
- Consistent low latency
- Consistent high performance

How?

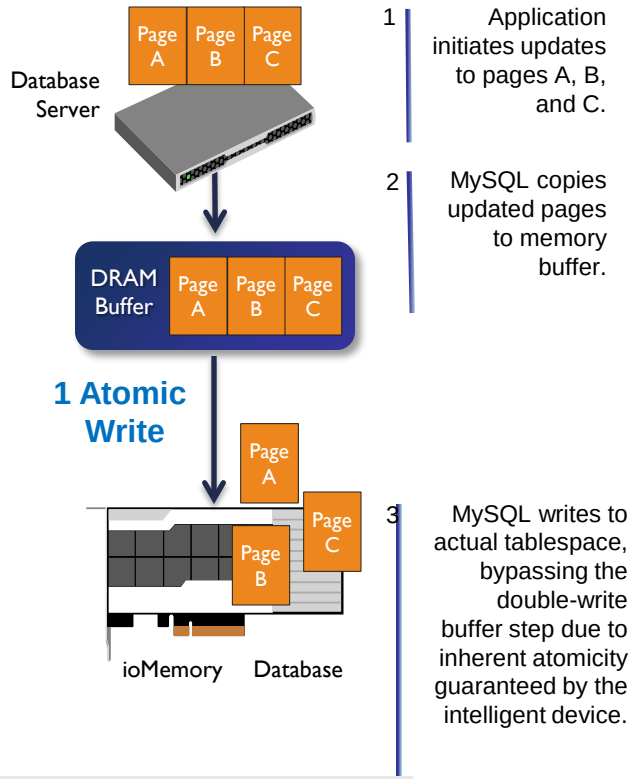
- Reducing Writes to flash
- Optimize IO Write path for flash
- Applications leverage enhanced I/O interface



1 SanDisk NVMFS – Solve Double Write Problem with Atomic Write Feature

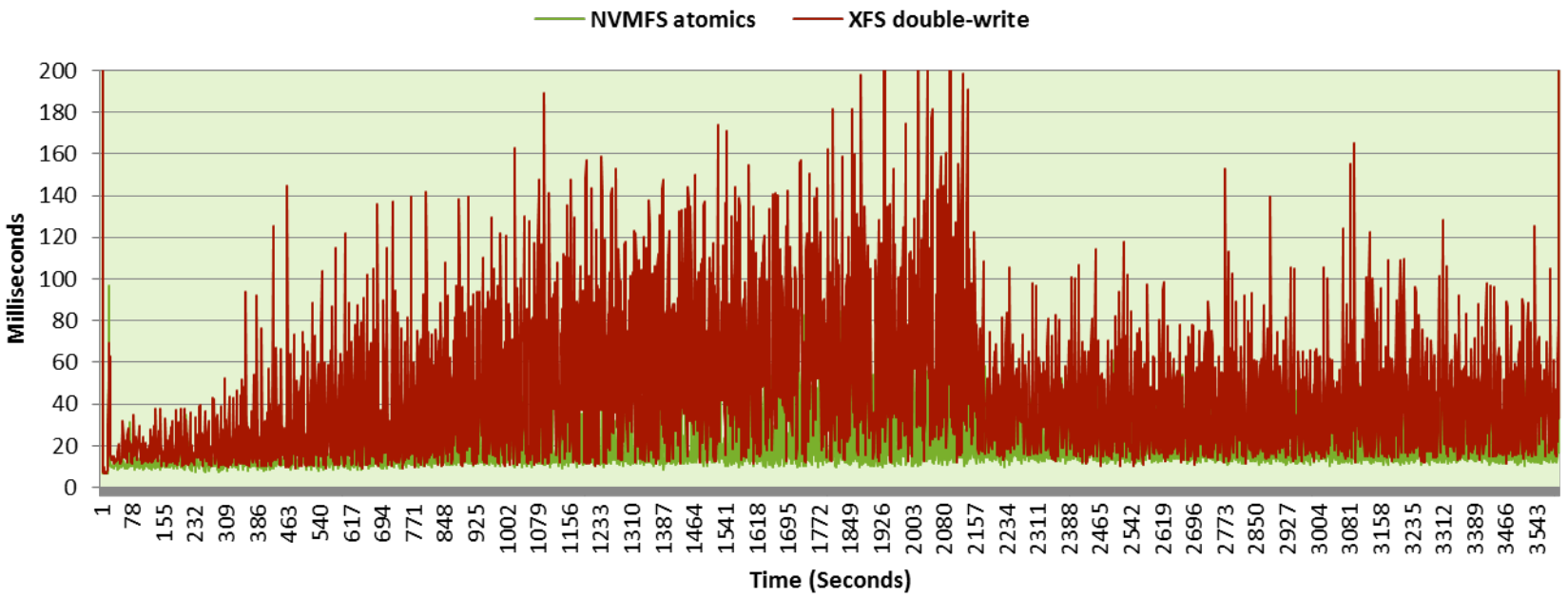
- Enhanced Life Expectancy of Flash Devices
  - Reduce Writes to flash by half at similar throughput
- Consistent low latency
- Higher performance especially for workloads with datasets that are bigger than DRAM

- MySQL with Atomic Write



A perfect fit for ACID compliant MySQL

Consistent Low Latency

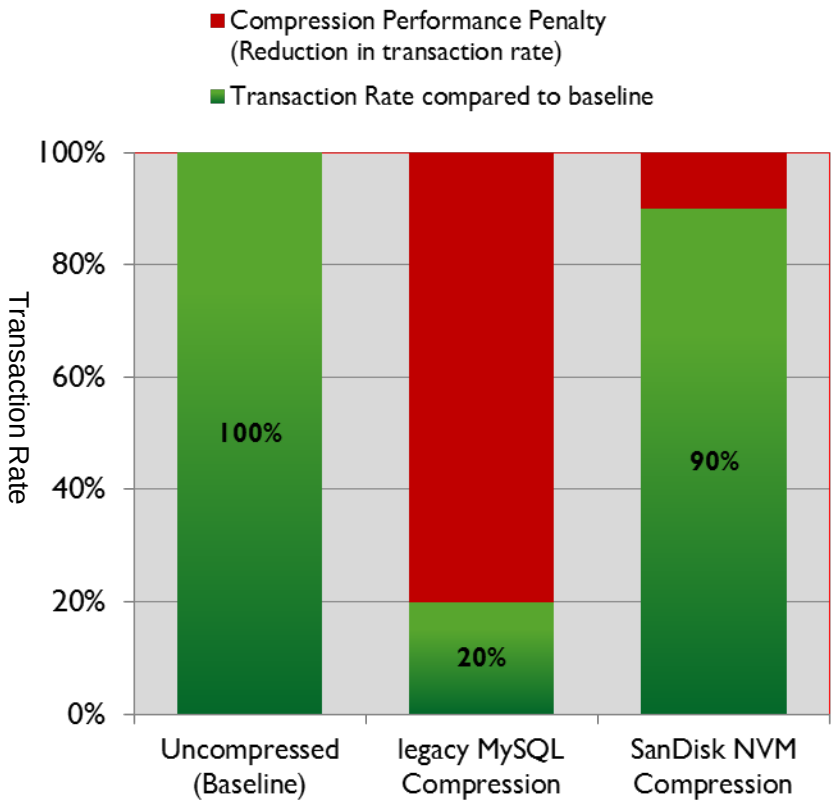


Sysbench - MariaDB 10.0.15, 4000 OLTP TXN injection/second, 99% latency, 220GB data - 10GB buffer pool

Significantly Lower Latency with SanDisk NVMFS Atomic Write
(compared to traditional double-write)

2 SanDisk NVMFS – Significantly Improved Compression

- SanDisk Accelerated Compression:
  - Within 10% of uncompressed performance
  - 50% improvement in capacity[1]
  - Enhanced life expectancy of flash devices[2]
    - Up to 4x fewer writes to storage
    - With compression and Atomic Write

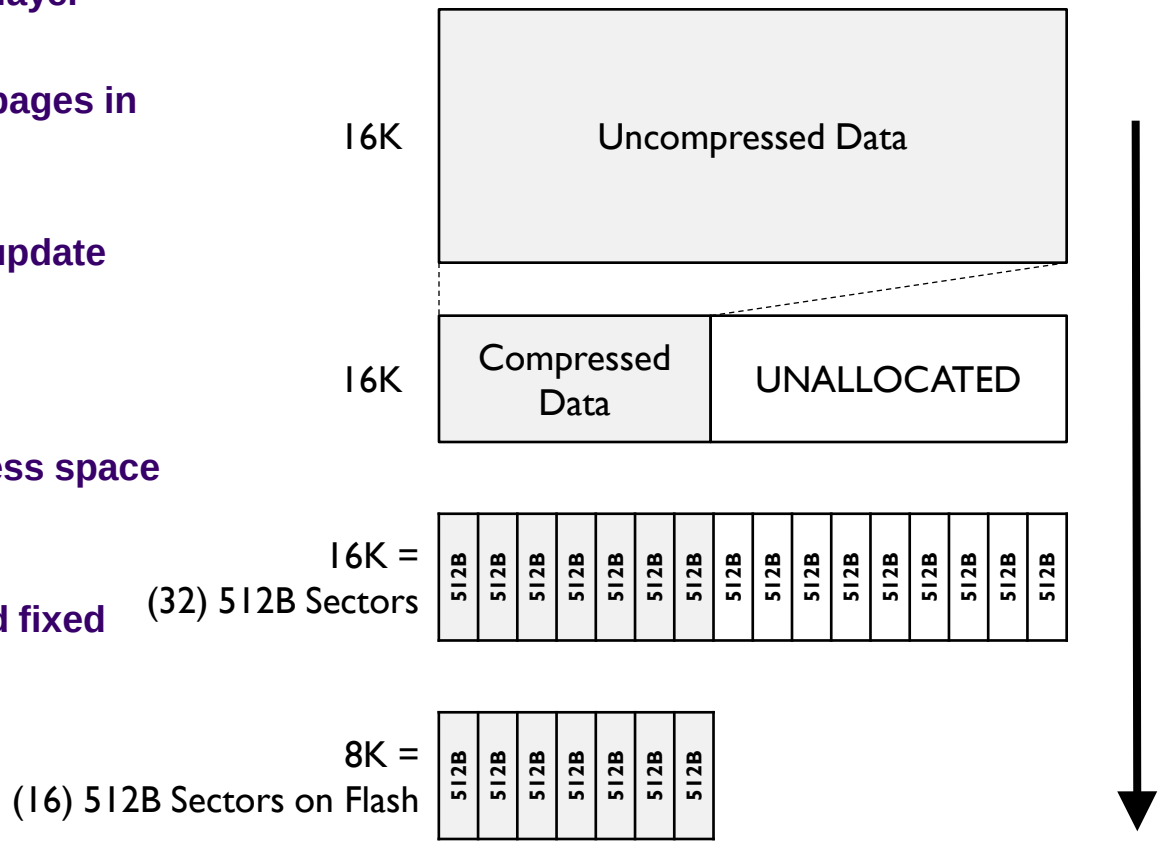


Compression with almost **no performance hit**

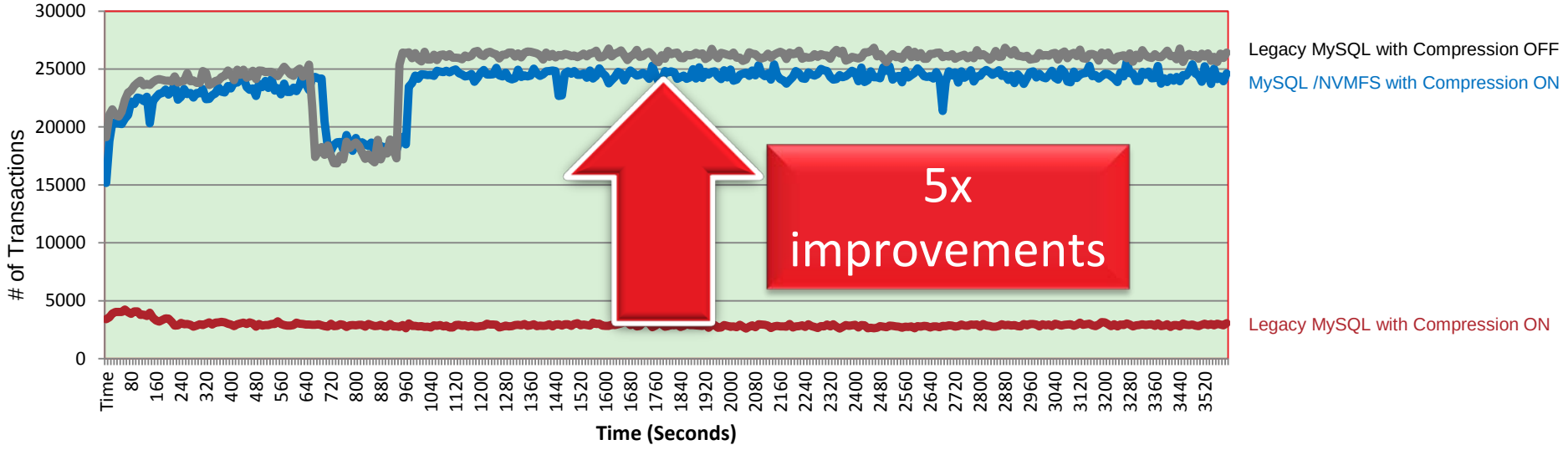
[1]For workloads that compress well. Improvement will vary
[2]At Similar Throughput (assuming same load)

SanDisk Acceleration

- Move compression to the lowest layer
- Only store uncompressed 16KB pages in memory. Keep code ‘as is’
- Tables recompressed with each update
- Use TRIM to free unused space
- NVMFS file system reports that less space is used on media
- No limitations due to pre-selected fixed compressed page size
- Very simple



Compression With Almost No Performance Hit Write-Heavy Applications

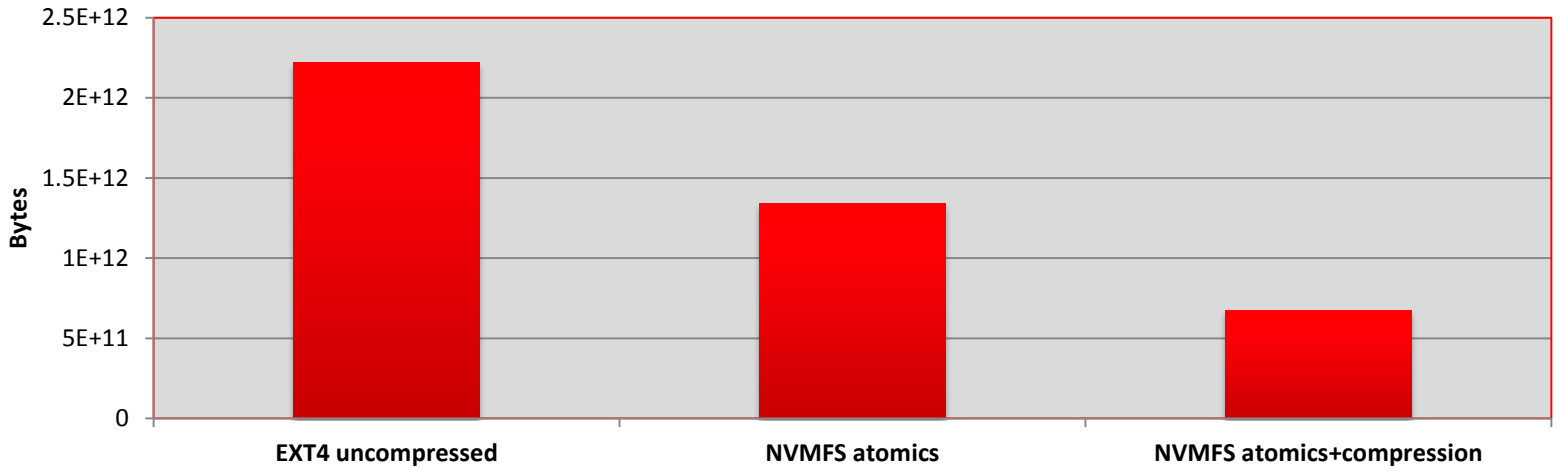


TPC-C like benchmark, 1000 warehouses - 75GB Buffer pool, MariaDB 10.0.15

- Enabling NVM compression has little impact on the MySQL transaction rate
- Enabling legacy MySQL compression has 80% penalty

Combining Atomic Write with NVM Compression

Reduces MySQL Write Operations to Flash by 70%



Summary

- **NVMFS is designed from grounds up for flash storage**
- **Achieve optimum flash performance and efficiency**
- **Customers will benefit:**
  - **Increase life expectancy of flash devices**
  - **Consistent low latency**
  - **Consistent high performance**

