NVMFS: A New File System Designed Specifically to Take Advantage of Nonvolatile Memory

Dhananjoy Das, Sr. Systems Architect

SanDisk Corp.

Forward-Looking Statements

During our meeting today we will make forward-looking statements.

Any statement that refers to expectations, projections or other characterizations of future events or circumstances is a forward-looking statement, including those relating to market growth, products and their capabilities, performance and compatibility, cost savings and other benefits to customers. Information in this presentation may also include or be based upon information from third parties, which reflects their expectations and projections as of the date of issuance.

We undertake no obligation to update these forward-looking statements, which speak only as of the date hereof.

SanDisk®

Agenda: Applications are KING!

- Storage landscape (Flash / NVM)
- Non Volatile Memory File System
  - [Use Case 1] MySQL Atomics
  - [Use Case 2] MySQL NVM-Compressed
  - [Use Case 3] Extended memory, DB Acceleration

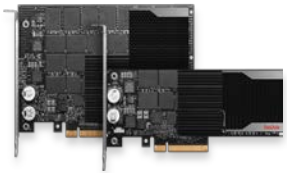
Non-Volatile Memory (NVM)

Current:

NAND Flash

- Capacity: 100s of GB to 100s of TB per device
- Trends: Higher capacity, lower cost/GB, lower write cycles, SLC->MLC->3BPC
- IOPS: 100K to millions, GB/s of bandwidth

SAS and SATA attach SSDs



PCIe attached

Non-Volatile Memory

- DDR/PCI-e attached NVDIMM / Capacitance backed power-safe buffers + FLASH
- orders of magnitude performance improvement

Fabric attached

NVDIMMs

Future: Non-Volatile Memory technologies (Phase Change Memory, MRAM, STT-RAM, etc.)

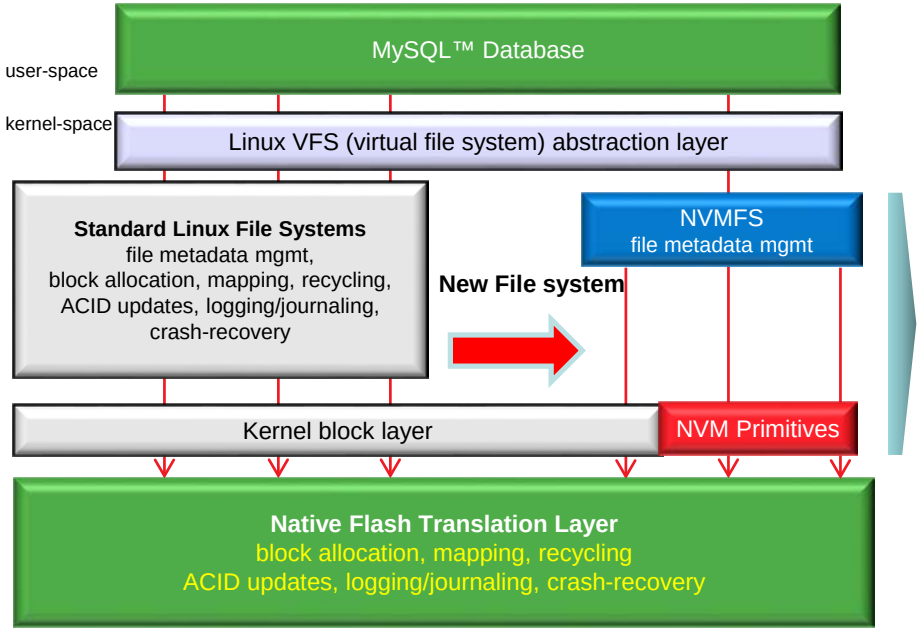
Why Do Applications Need Optimization for Flash?

	HDD	Flash
Read/Write Performance	Largely symmetrical	Heavily asymmetrical
Wear out/ Background ops	Largely unlimited / Rare	Limited write cycles / Regularly
IOPS/ Latency	100 to 1,000 / 10's milli sec	100Ks to Millions / 10's-100's micro sec
Addressing	Sequence, Sector	Direct, addressable

Managed writes = greater device lifetime (wear leveling, endurance)
Improved system efficiency (TCO and TCA)

Becoming “Flash Aware”: SanDisk NVMFS

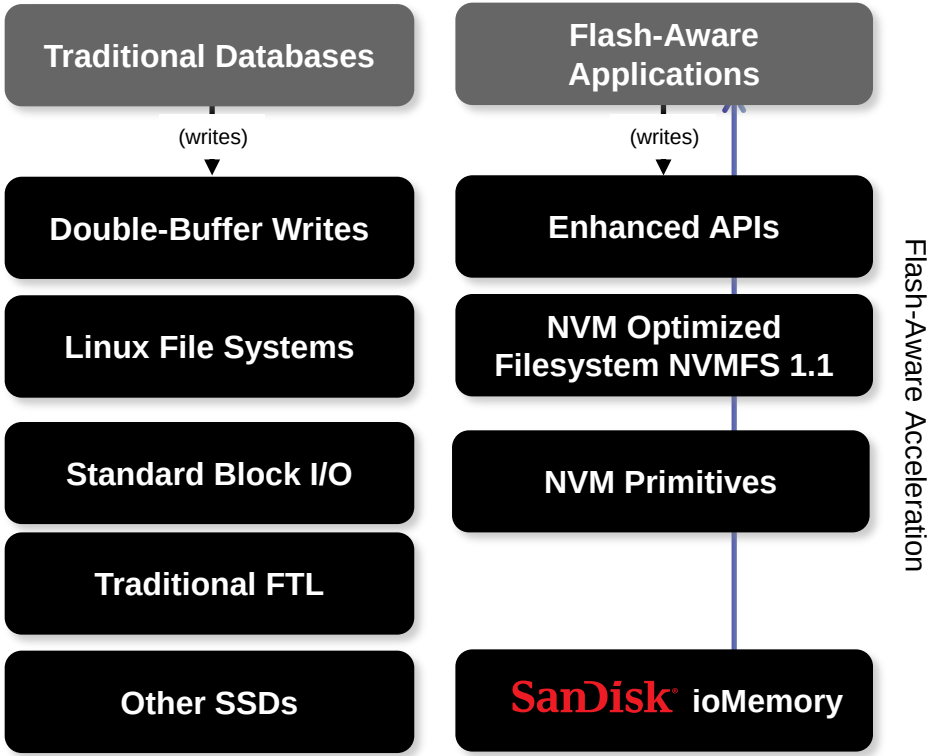
Non-Volatile Memory File System – Optimized for Flash and Persistent Memory



- NVMFS is POSIX compliant
- Leverages the functionality of underlying Flash Translation Layer (FTL)
- Namespace Management
- Enables Direct flash/memory access and crash recovery
- NVM primitives are exposed through standard system file interface

Eliminating Duplicate Logic and Leveraging New Primitives for Optimal Flash Performance and Efficiency

Flash Beyond Disk: Adapting the Software Stack

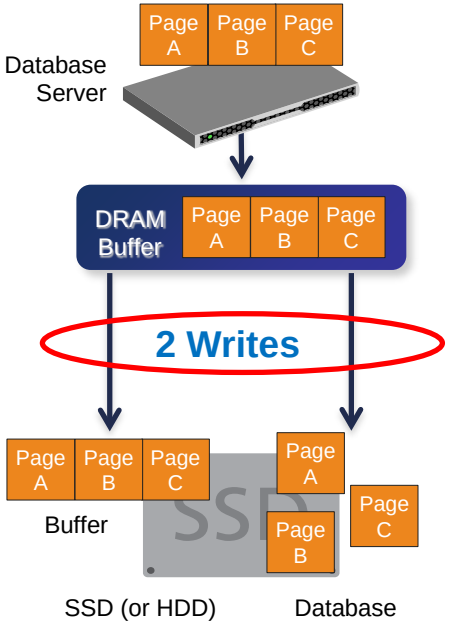


- **Flash-Aware Acceleration**
  - Changes to MySQL are “aware” of flash and automatically leverage optimized API
- **NVMFS 1.1 *New!***
  - NVM optimized filesystem
  - Standard file namespace, all existing customer management tools work
  - Raw performance of NVM
  - Flash-aware interfaces direct to applications

Legacy MySQL Challenges

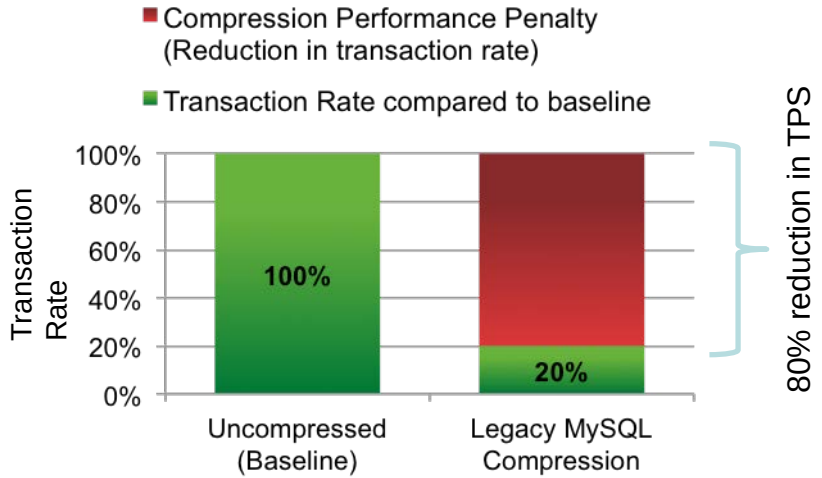
Double-Write and Compression Penalties

1 Every MySQL write translates to **2 writes** to storage device



1. Application initiates updates to pages A, B, and C.
2. MySQL copies updated pages to memory buffer.
3. MySQL writes to double-write buffer on the media.
4. Once step 3 is acknowledged, MySQL writes the updates to the actual tablespace.

2 **80% performance penalty** with legacy MySQL compression enabled



Results and performance may vary according to configurations and systems, including drive capacity, system architecture and applications.

Solving the Double-Write Problem

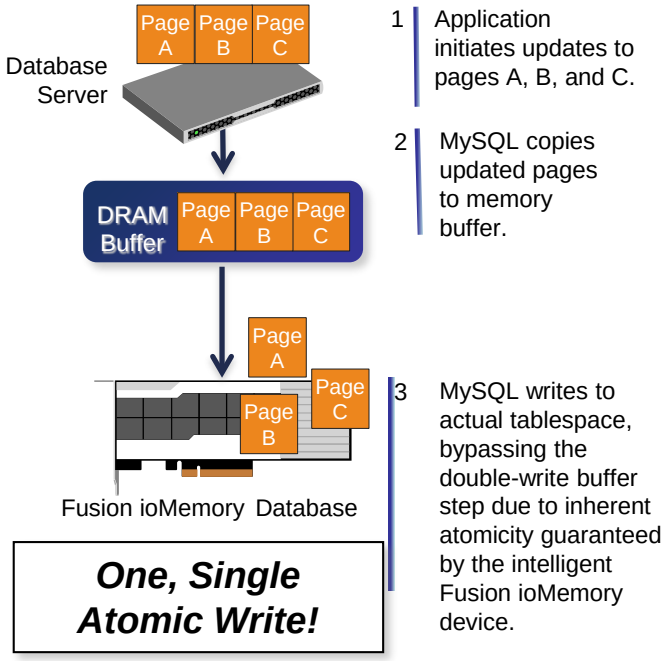
SanDisk NVMFS with Atomic Write

1

- Enhanced Life Expectancy of Fusion ioMemory Devices:
  - Reduce Writes to flash by half at similar throughput
- Improved performance consistency
- Reduced latency, increased transaction/sec
- Higher performance
  - Especially workloads with datasets that are bigger than DRAM

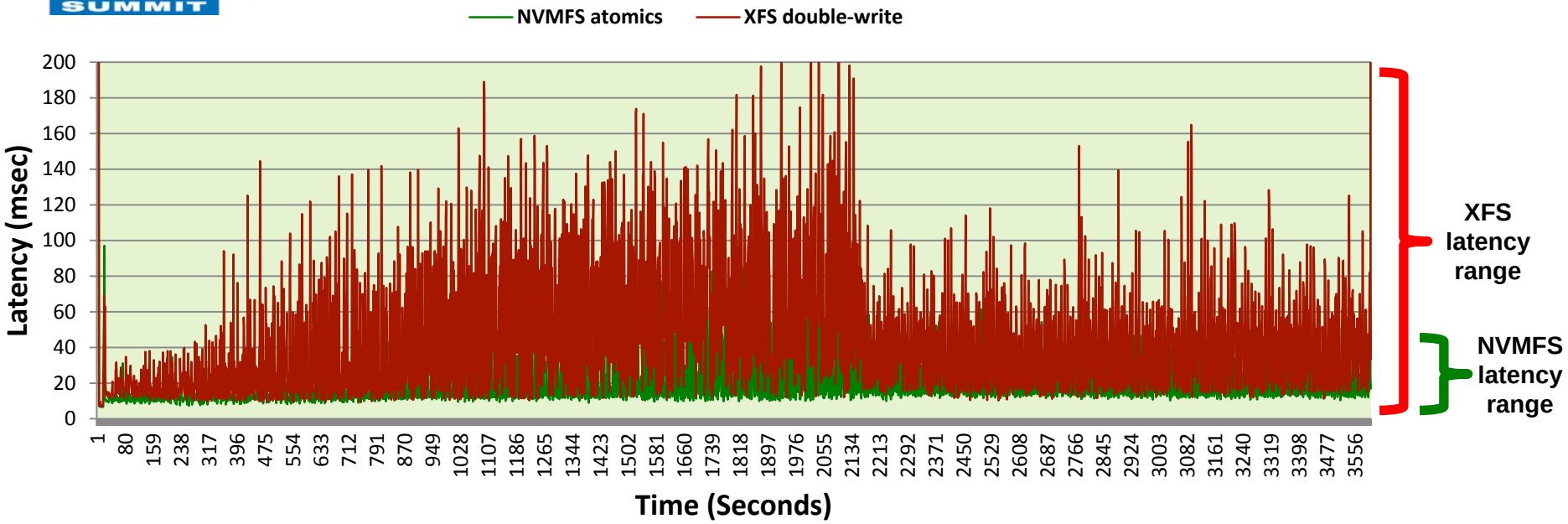
Perfect Fit for ACID-compliant MySQL

- MySQL with Atomic Write



SanDisk NVMFS Improves Latency Consistency

Lower Latency with Greater Stability



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

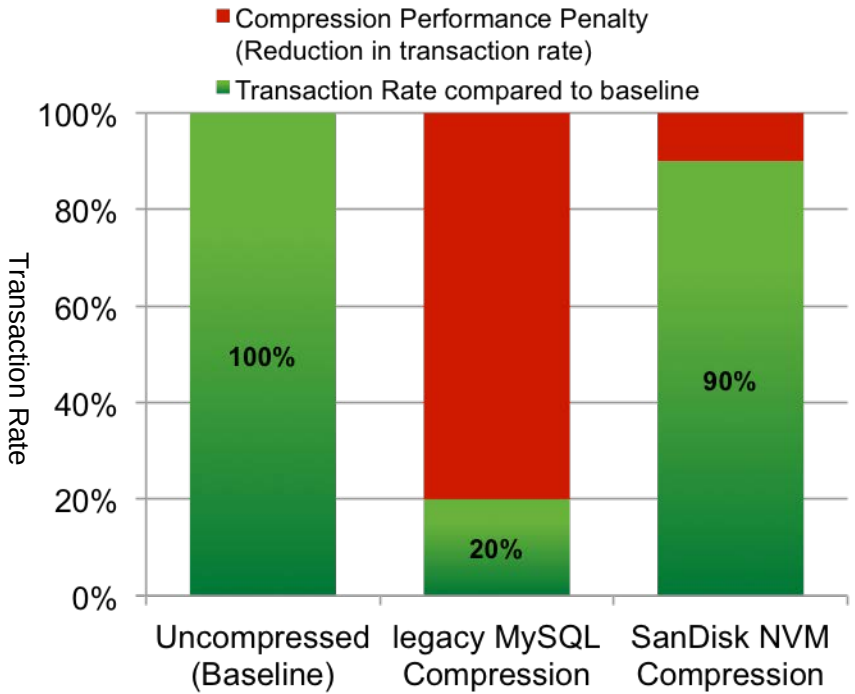
NVMFS Atomic Write **Significantly Reduces Latency** while Increasing Performance Consistency

Improving MySQL Compression

SanDisk Contribution to MySQL Community

2

- Benefits of compression without severe performance penalty
  - Within 10% of uncompressed
- Up to 50% improvement in capacity utilization[1]
- Enhanced life expectancy of flash devices[2]
  - Up to 4x fewer writes to storage with Compression and Atomic Write



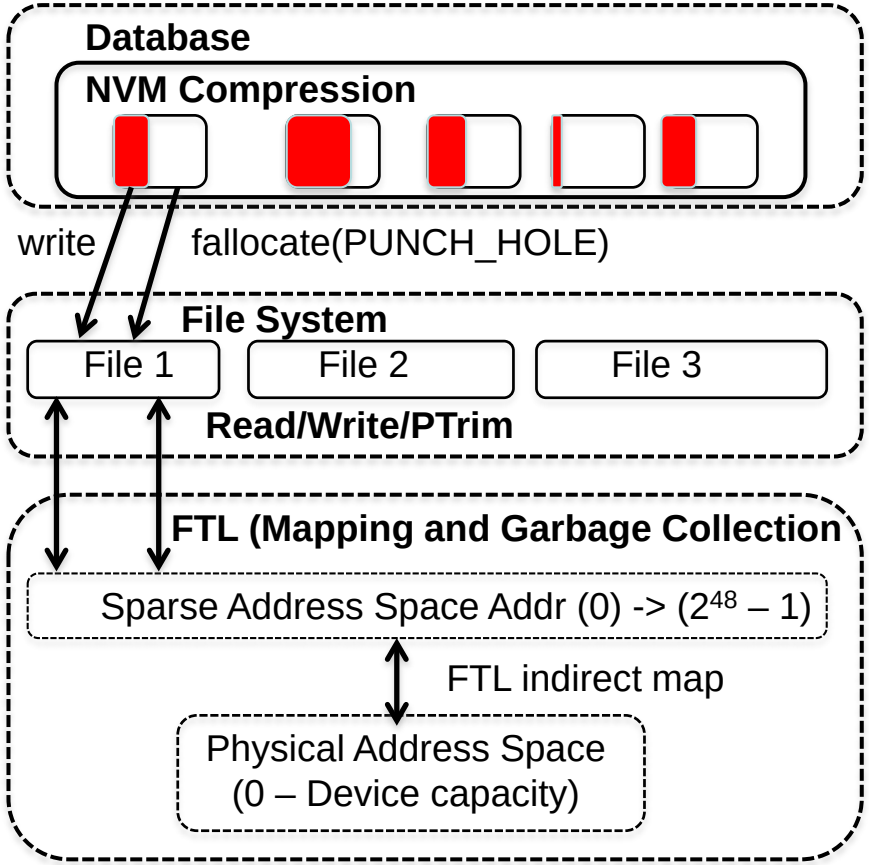
Compression with almost **no performance penalty**

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

Flash Beyond Disk: NVM Compression

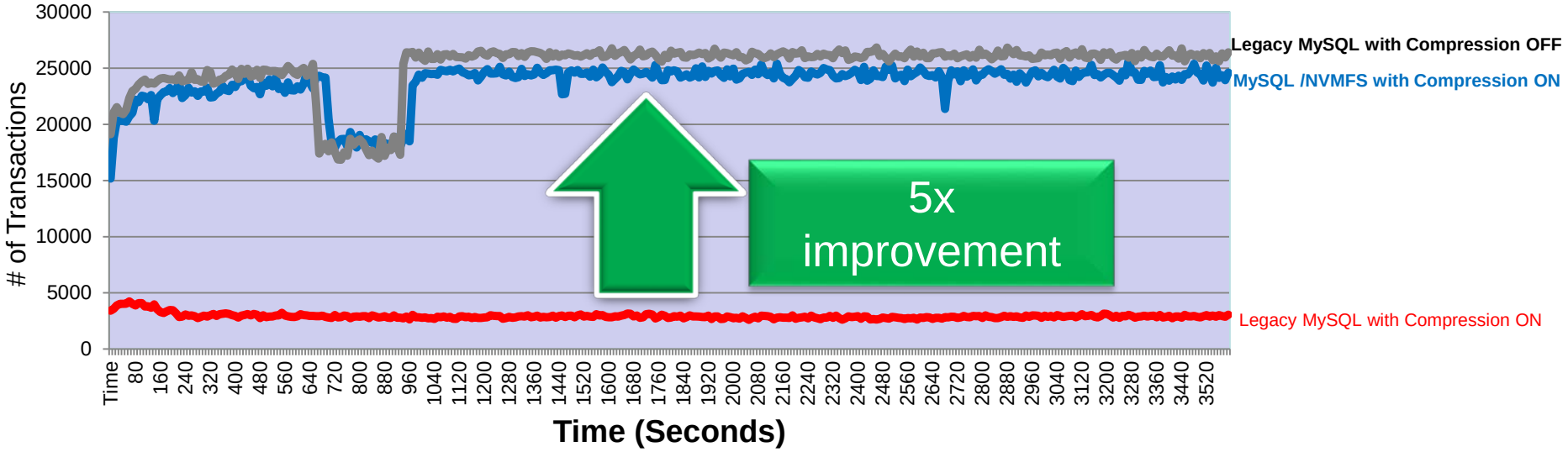
2 High level approach

- Application operates on sparse address space which is always the size of uncompressed.
- Compressed data block is written in place at same virtual address as the un-compressed. Leaving a hole, empty space in the remainder of space allocated.
- FTL garbage collection naturally coalesces the addresses in physical space, allowing for re-provisioning of physical space.



Compression without Performance Penalty

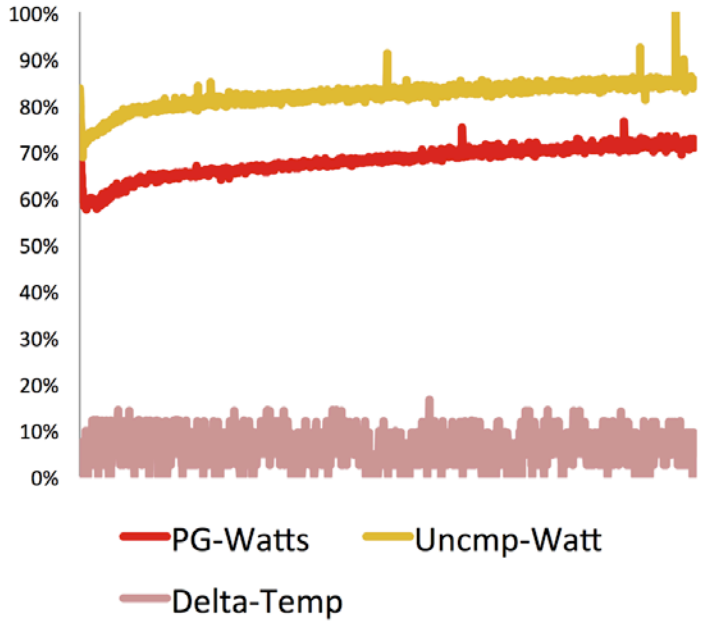
Improved Flash Utilization

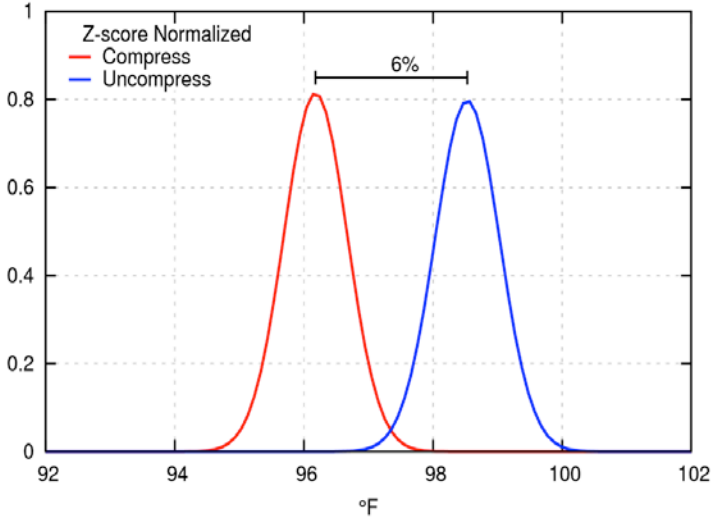


TPC-C-Like benchmark, 1000 warehouses - 75GB Buffer pool, MariaDB 10.0.15

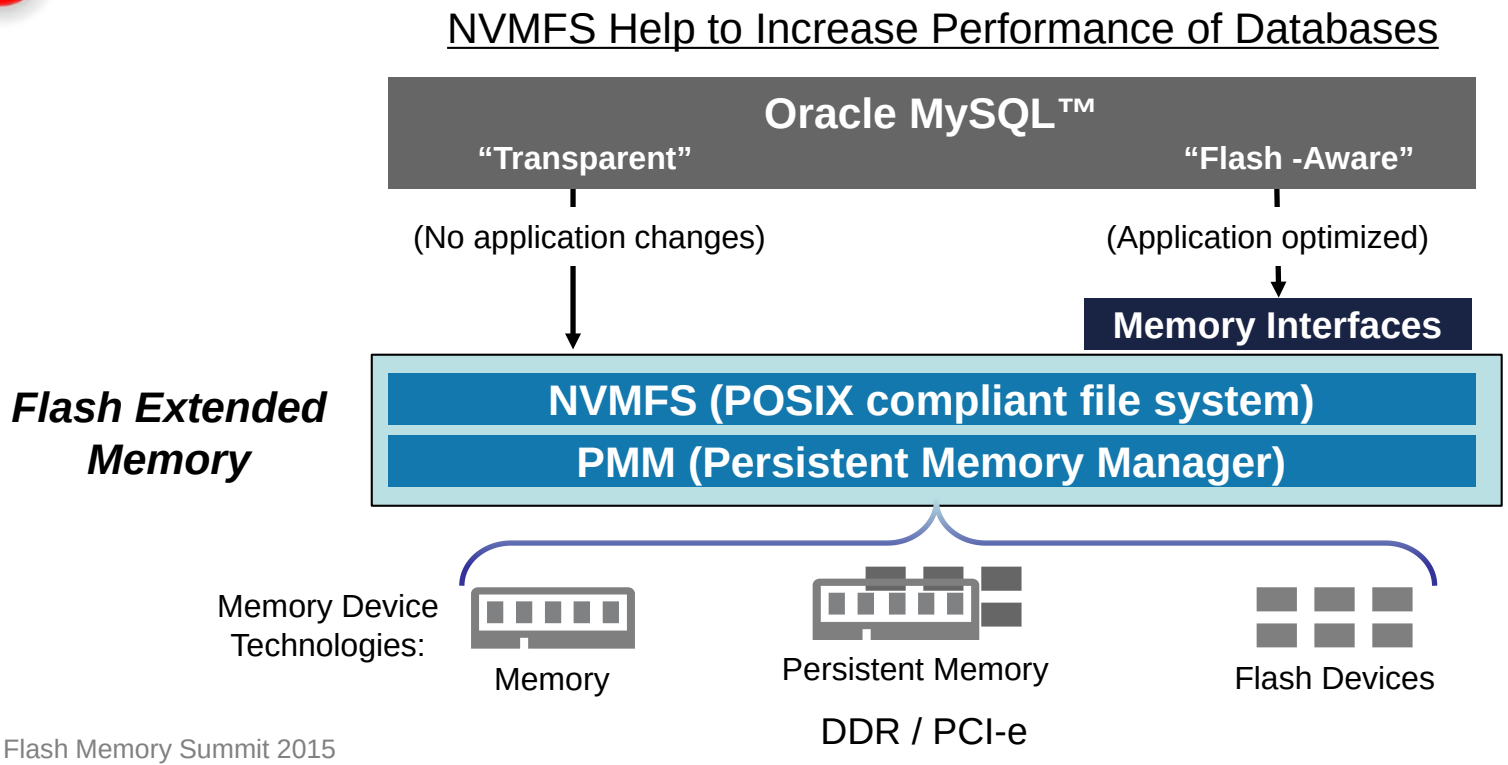
NVM Compression Eliminates Legacy MySQL's Compression Penalty

Power/Cooling





Database Acceleration using Flash Extended Memory



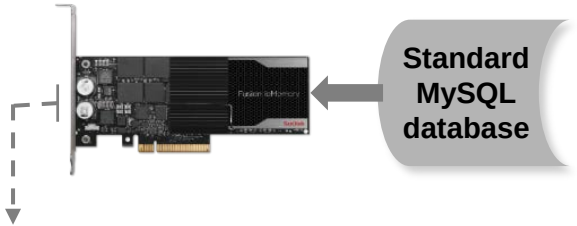
MySQL Configuration

NVM attach points: **DDR / PCI-e**

3

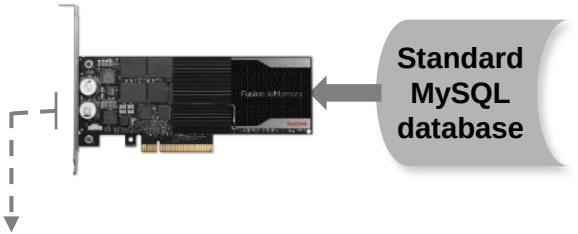
“Baseline”

Fusion ioMemory™ PCIe-based flash



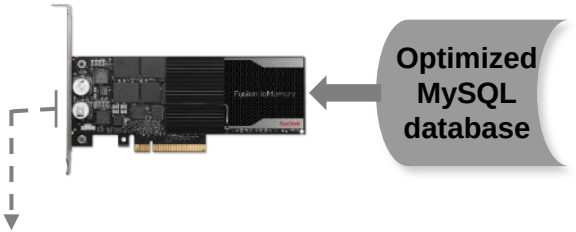
“Transparent”

Fusion ioMemory™ and persistent memory with NVMFS



“Flash Aware”

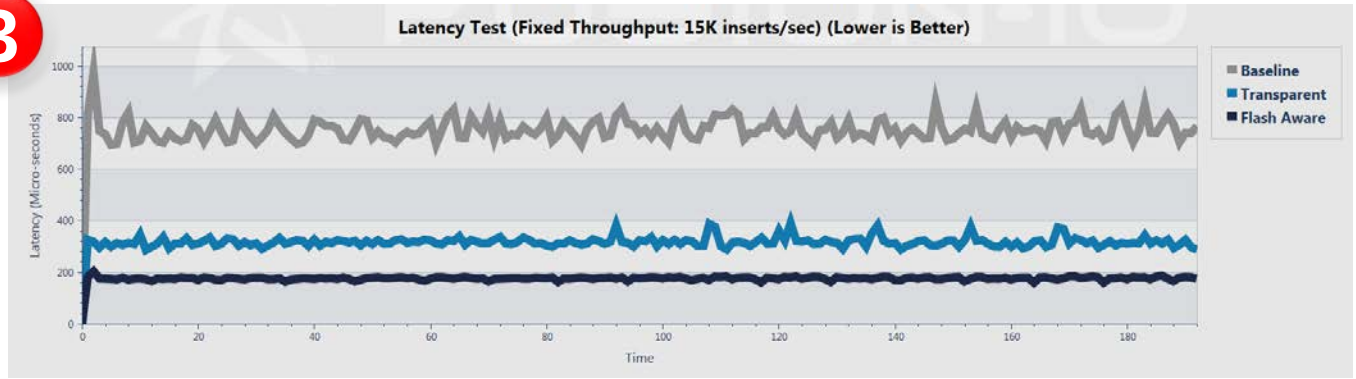
Fusion ioMemory™ and persistent memory with NVMFS



Flash Extended Memory Enabled

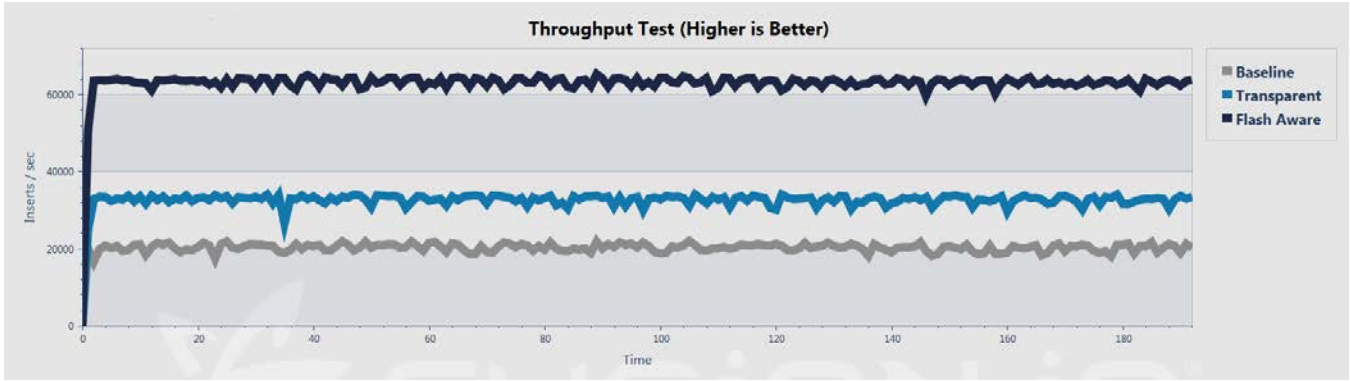
Performance Results

PCI-e attached NVM



Latency
(lower is better)

Throughput
(higher is better)

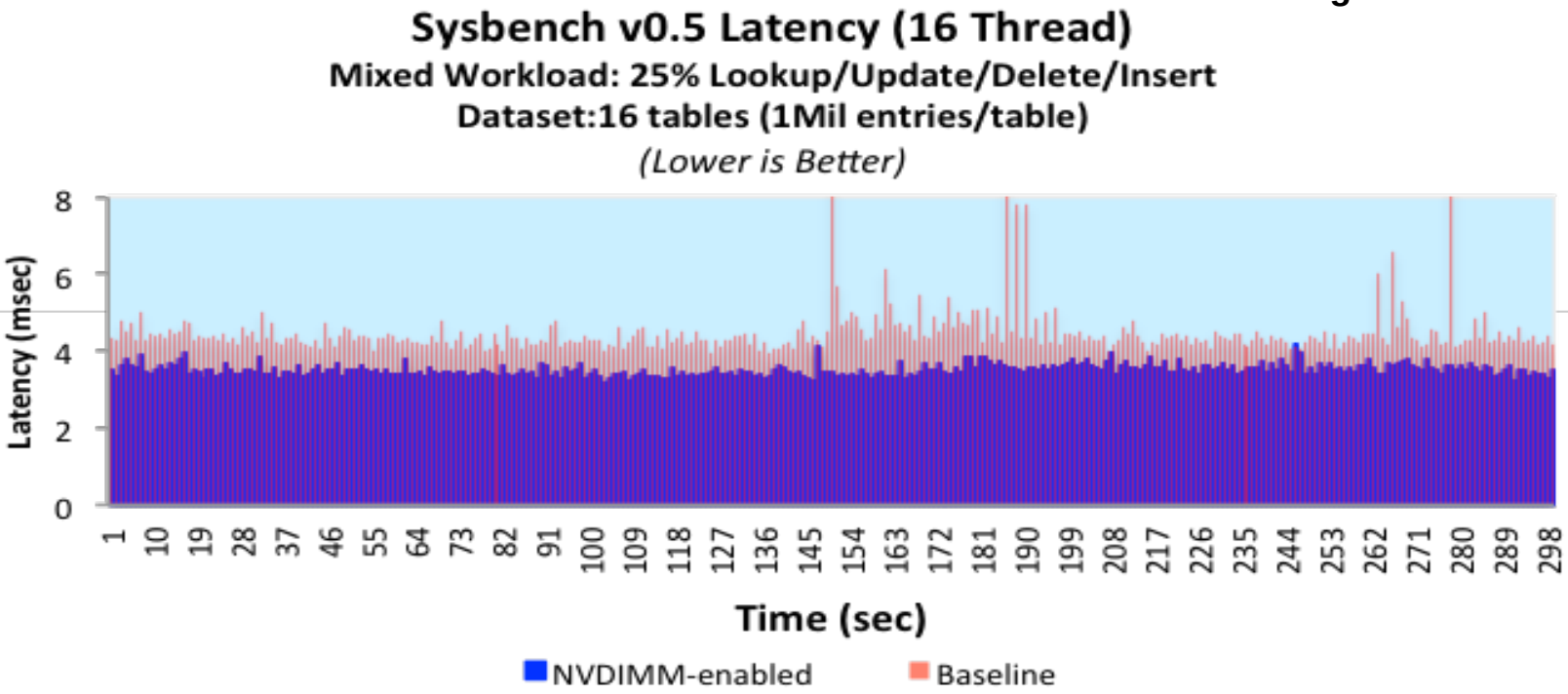


Source: Based on internal testing by SanDisk

MySQL :: Transparent Acceleration (latency)

Performance Overview:: Comparison Between Baseline and NVDIMM

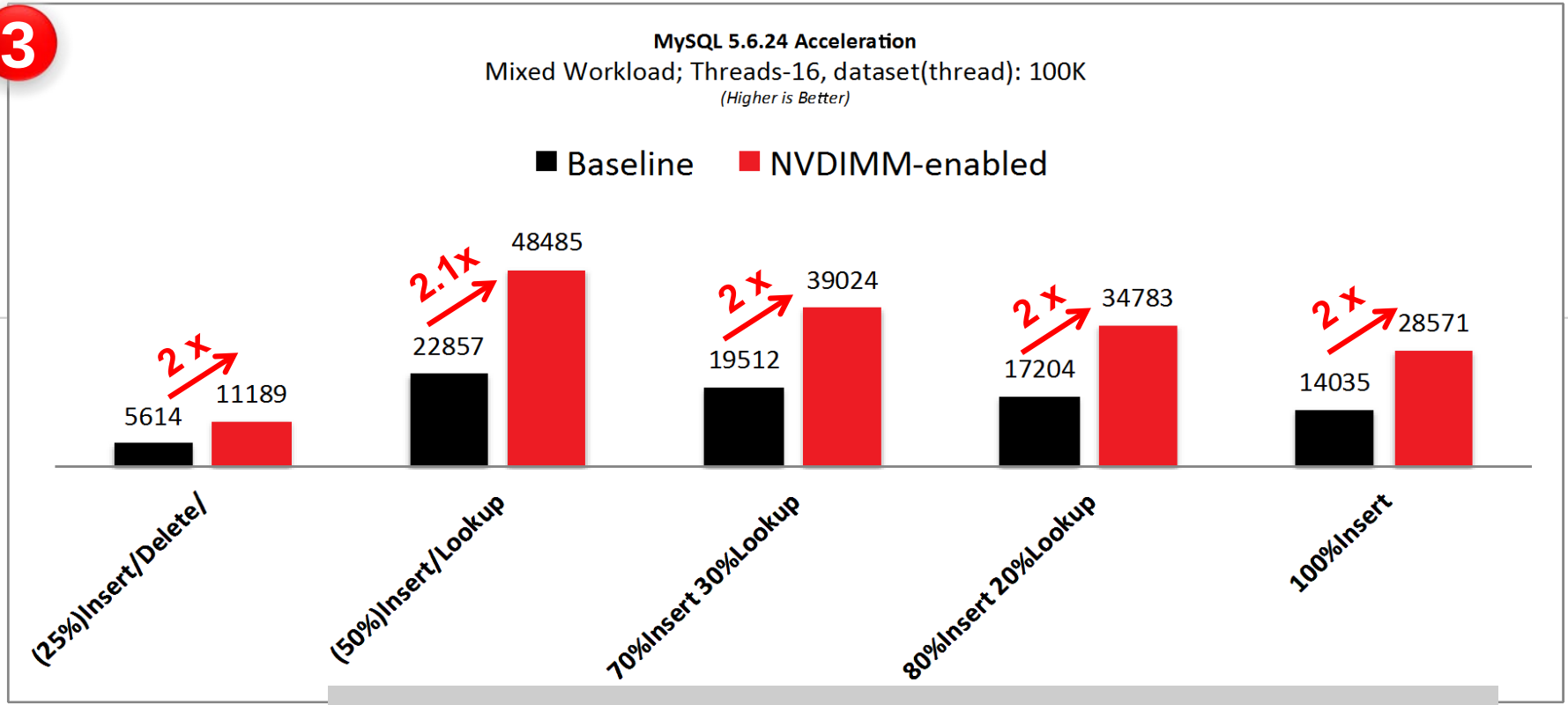
Using DDR NVDIMM



MySQL :: Transparent Acceleration

Performance Overview:: Comparison Between Baseline and NVDIMM

3



Up to 2x higher transactions per second using DDR attached NVDIMM

Atomic Writes Summary

1

- Transaction throughput improvements of roughly **2.4x** over conventional flash SSDs
- Half as many writes per transaction means potential for as much as **2x** flash endurance for write intensive workloads:

  more cost effective flash storage

- Standardized: Approved SNIA standard,SBC-4 SPC-5 Atomic-Write http://www.t10.org/cgi-bin/ac.pl?t=d&f=11-229r6.pdf
- Uses unique flash-aware optimizations from SanDisk
- Available commercially and fully supported from SanDisk (NVMFS 1.1) and Oracle MySQL 5.7.4, Percona Server 5.5, 5.6 and MariaDB 10

The performance results discussed herein are based on internal testing and use of Fusion ioMemory products. Results and performance may vary according to configurations and systems, including drive capacity, system architecture and applications.

NVM Compression Summary

- Performance within 10% of uncompressed (and sometimes greater) for Linkbench and TPC-C *like* workloads. **5x** acceleration of TPC-C *like* as compared to Row Compression
- Storage savings of ~**2x** (data dependent) with as much or more compressibility as MySQL row compression

- Upto **4x** better flash endurance when combined with Atomic Writes
- Addition power/cooling benefits and scalability benefits
- Available commercially and fully supported from SanDisk (NVMFS 1.0) and MariaDB 10, coming soon from Oracle and Percona distributions

Extended Memory: Advantages & Benefits

3

Workload: Insert Heavy	**Transparent (no software mods)**	**Flash-Aware (modified software)**
Throughput	**1.8x - 2x**	**4x**
Latency	**2x**	**4x**

- Uses “Flash-as-Memory” byte-addressable architecture and interface
- Seamless deployment – add ioMemory and NVMFS/PMM software to Linux
- Increase performance and capacity in flexible configurations

Who is NVMFS for?

- **NVMFS will optimize customer database flash storage by improving**
  - Transactional performance such as, latency and throughput
  - Enhanced lifespan of flash devices
  - Practical capacity
- **Enterprise environment**
  - OLTP databases running in a Linux OS environment
  - Insert heavy workloads needing to persist large amounts of data
  - Latency sensitive OLTP workloads
  - Concerned about flash endurance
- **Hyperscale environment**
  - To improve CPU utilization per node
  - Clusters of MySQL nodes by being able to store more data

Questions?

Thank You!

@BigDataFlash
#bigdataflash
ITblog.sandisk.com
http://bigdataflash.sandisk.com

Backup

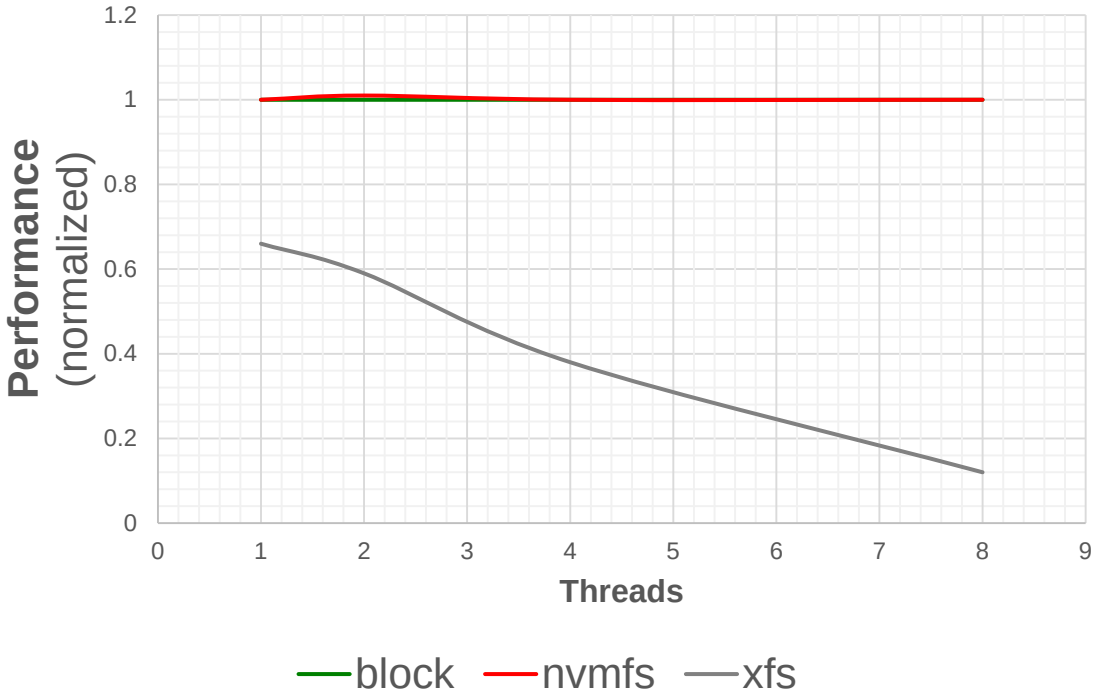
Performance - Datapath

Micro-benchmark

- Parallel, direct I/O on a single file on a very fast device

Applications

- Databases
- Virtual Machines



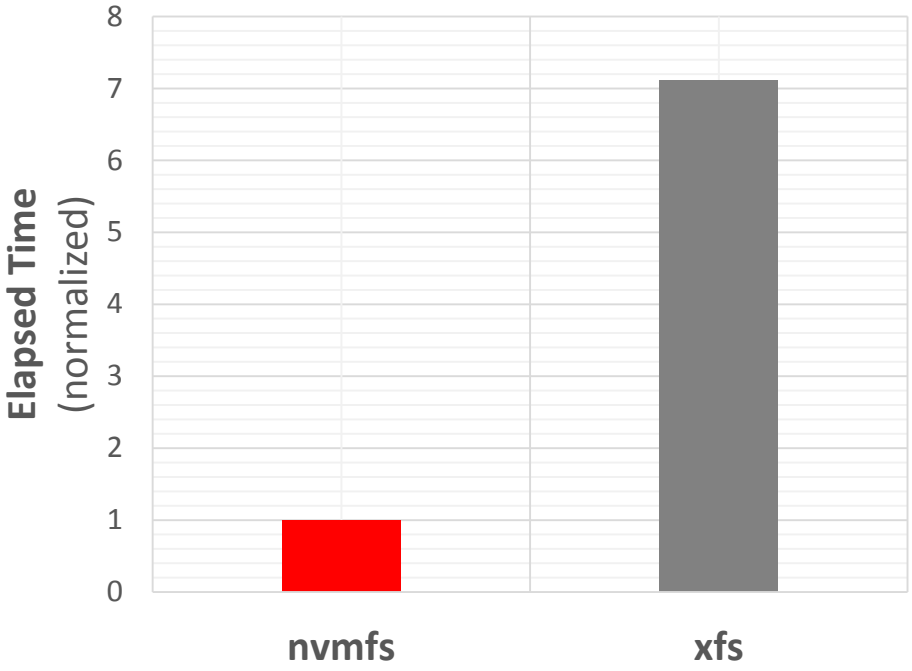
Performance - TRIM Handling

Micro-benchmark

- Trim after write
- 16 KiB Direct Write + 4 KiB TRIM

Applications

- MySQL Page-compression



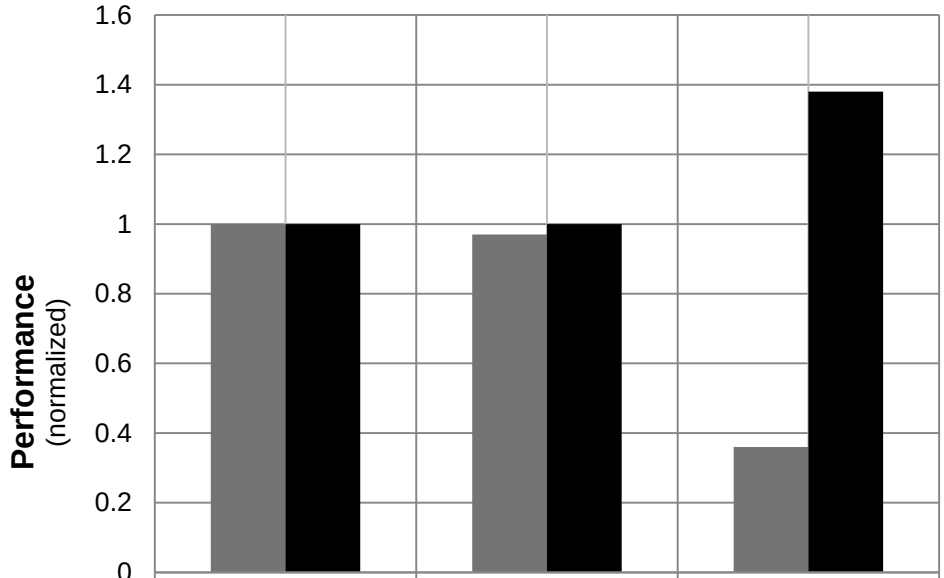
Performance - File Logging

Micro-benchmark

- Append data at end of file
- 4 KiB write(2) + fdatasync(2)

Applications

- Databases
- HFT
- Log Structured Systems



	block device	nvmfs	xfs
IOPS	1	0.97	0.36
Physical Bytes Written	1	1	1.38

Acceleration for Cassandra

Performance Overview:: Comparison Between Baseline and NVDIMM

- **Up to 3.2x reduction in Writes to flash resulting in a longer device lifetime**
  - Utilize flash hardware longer