Flash-aware MySQL

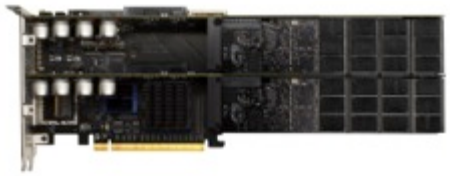
～フラッシュがMySQLを変える～

Takeshi Hasegawa | Senior Sales Engineer APAC Japan | Fusion-io

April 2014

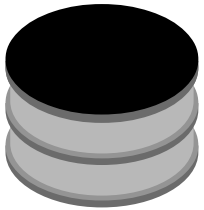
不揮発メモリ(NVM)の登場

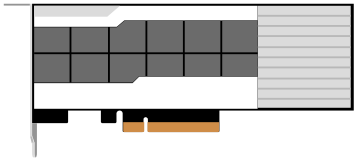
- フラッシュ(NAND)
  - デバイスあたり数百GB～ 10TBの容量
  - フラッシュ技術のトレンド
    - 大容量化
    - GB単価コスト↓
    - 書き込み回数の減少
    - セルの多値化（SLC→MLC→3BPC）
  - 10万～100万IOPS, GB/s級の帯域幅
- その他の不揮発メモリ技術（PCM/MRAM/STT）
  - 現時点では開発中のメモリ技術



なぜフラッシュを使うのか?

- I/O特性がデータベース用途に適している
  - 低レイテンシ、QDの低いI/Oでも高性能
  - ワークロードの種類を問わず性能が高い
    - ✓シーケンシャル ワークロード
    - ✓ランダム ワークロード
    - ✓様々なブロックサイズ

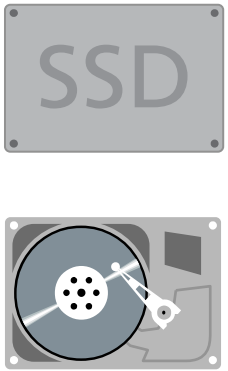




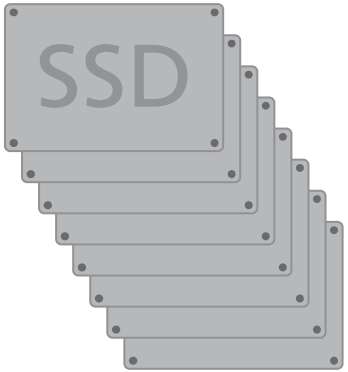
容量	4TB	3TB
IOPS	150	200,000
IO単価	$$$$	¢¢¢¢

フラッシュの利用方法の遷移

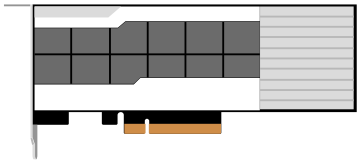
フラッシュ ＋ ディスク



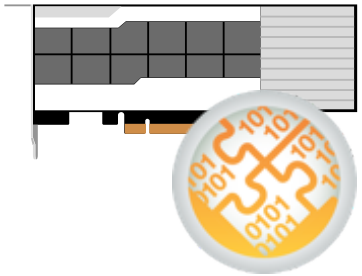
ディスクとしてのフラッシュ



フラッシュとしてのフラッシュ



メモリとしてのフラッシュ



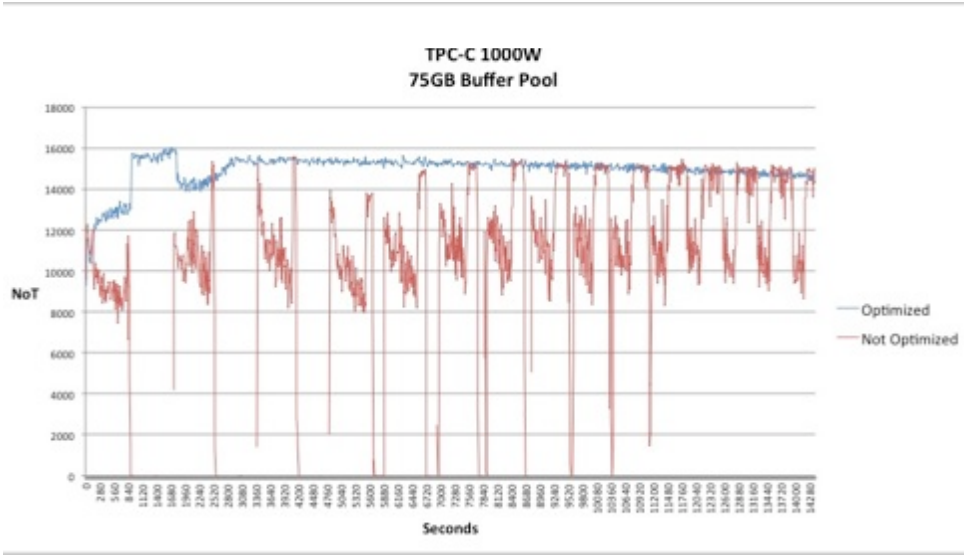
より低消費電力、低コストなトランザクションを実現

フラッシュの特性を意識した実装

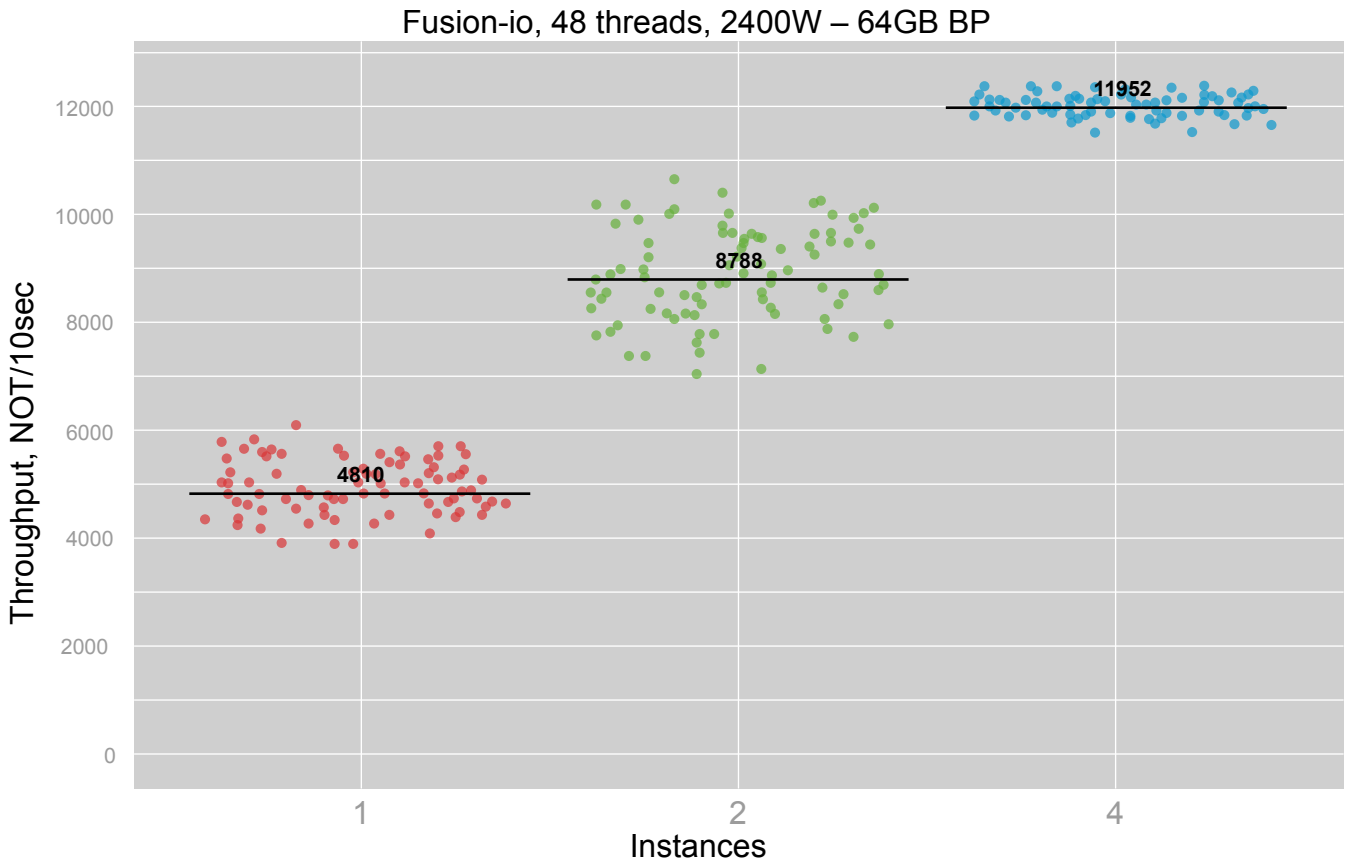
FUSION-IO

ディスクとしてのフラッシュ: そのスピードにチューニング

- 過去数年間の取り組みにより大幅な性能向上を達成
- データ配置の最適化、NOOPスケジューラ、シークなしメディアへの最適化、並列度の最適化
- ブロックI/Oサブシステムの高速化
- 高速なファイルシステムの探求



マルチインスタンス MySQL: IOPS性能を絞り出す



FUSION-IO

フラッシュとしてのフラッシュ: ただのディスクとは違う

メトリック	ハードディスク	フラッシュメモリ
リード／ライト性能	リード／ライト、ほぼ対称	リード／ライト性能が非対称。 イレースという新たな操作が 登場
シーケンシャル／ランダムの 性能傾向	100倍の性能差。 ヘッドの動作を想定した I/Oスケジューリング	～10倍の性能差。 メモリ素子にはヘッド動作なし
ブロックのリマッピング、 バックグラウンドでの処理	極めて少ない	ログ構造のファイルシステムの ように、定常的に発生
書き込み量の限界	ほぼ無し	制限あり
秒間あたりのI/O回数（IOPS）	100回～1,000回／秒	10万回～100万回／秒
レイテンシー（応答遅延）	10ミリ秒台	10～100マイクロ秒台

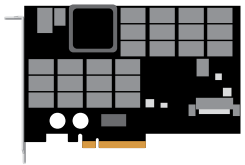
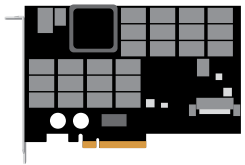
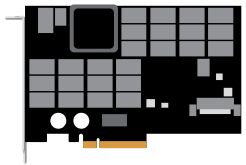
“Flash-aware” API によるMySQLの強化

“Flash-aware” スタックの構成

MySQL – アトミックライト, and NVM コンプレッション

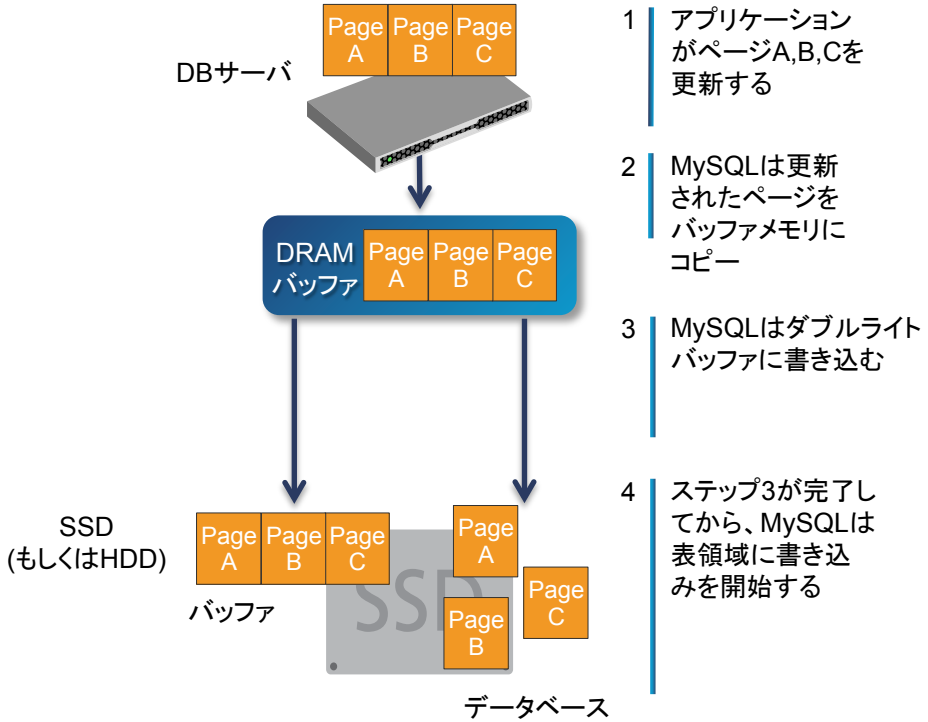
ファイルシステム(XFS, Ext4, Btrfs, NVMFS)

フラッシュストレージ –I/O と 新たなプリミティブ
(アトミックライト、PTRIMなど)

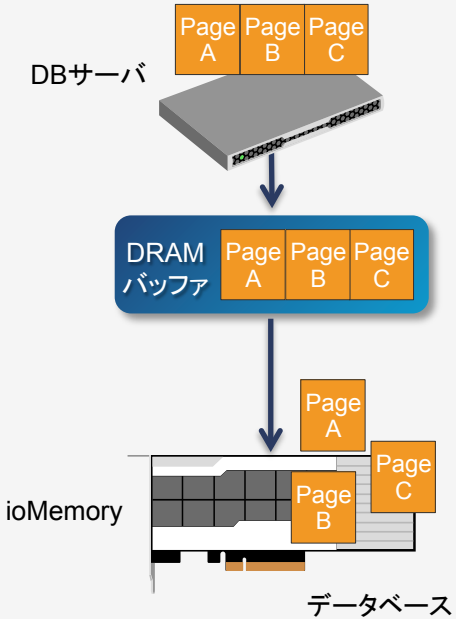
ダブルライト／アトミックライトの比較

従来のMySQLのライト処理



1. アプリケーションがページA,B,Cを更新する
2. MySQLは更新されたページをバッファメモリにコピー
3. MySQLはダブルライトバッファに書き込む
4. ステップ3が完了してから、MySQLは表領域に書き込みを開始する

アトミックライト対応版MySQLのライト処理

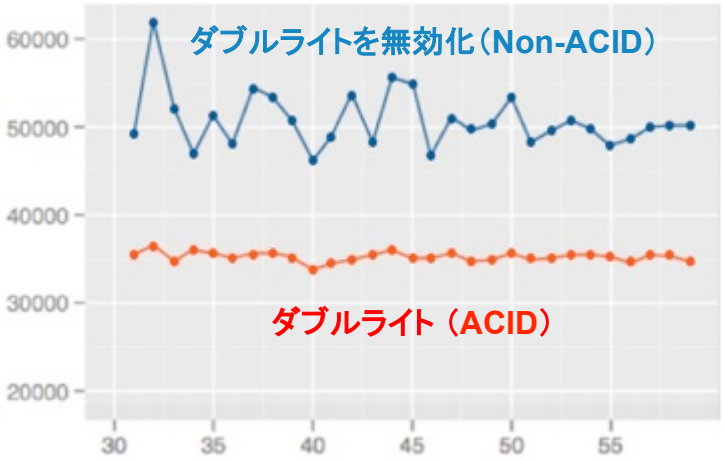


1. アプリケーションがページA,B,Cを更新する
2. MySQLは更新されたページをバッファメモリにコピー
3. MySQLはダブルライトを省略し、表領域に直接書き込み

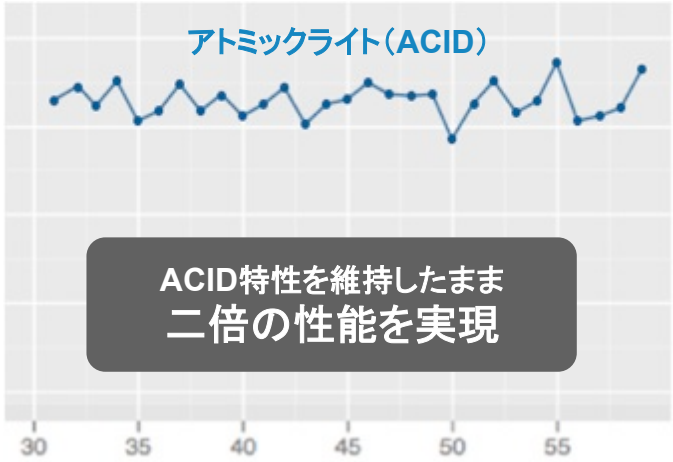
   データの整合性は下位デバイスの機能により担保される

FUSION-IO

MySQL + アトミックライトの利点



従来の場合

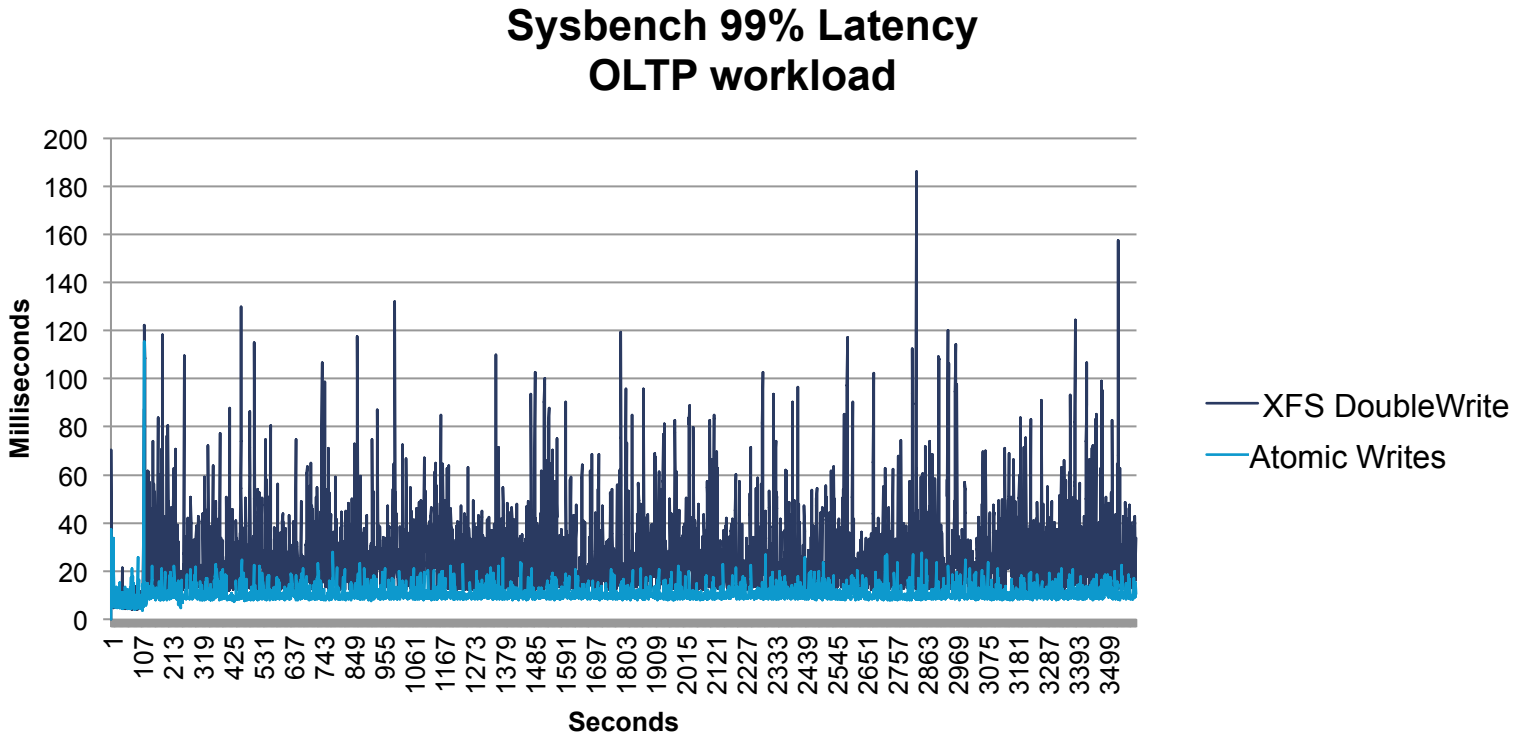


アトミックライトを利用する場合

- アトミックライトによりデバイス性能の99%を利用可能
- デバイスの書き込み耐用期間が2倍に

アトミックライト: トランザクションのレイテンシー改善

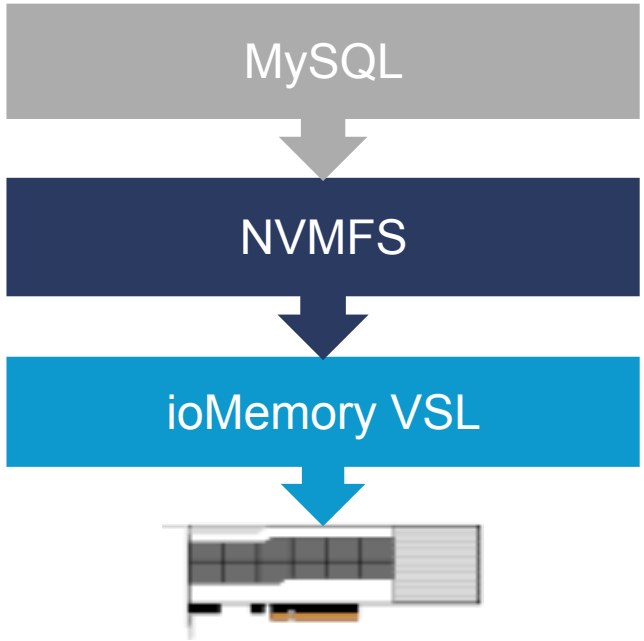
トランザクションのレイテンシーが 2分の1～4分の1まで短縮



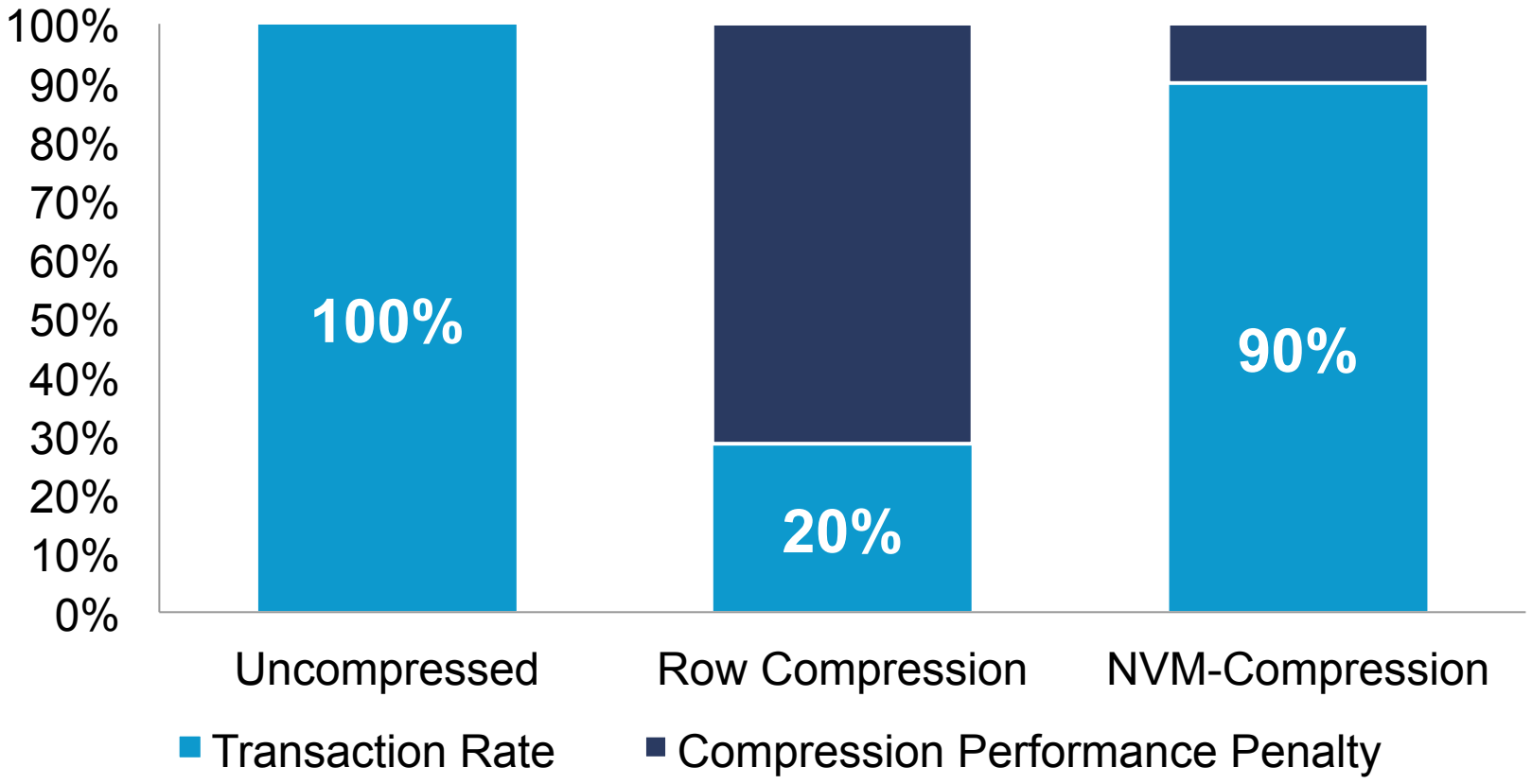
FUSION-IO

NVM コンプレッション

- フラッシュデバイスが持つ、内部的な「シンプロビジョニング」動作を活用
- データファイル上の不要ブロックをTRIM(UMMAP)しホール（スパース）化
- フラッシュ処理のマルチスレッド化、アトミックライトによりレイテンシーを削減
- プラグイン式で置き換え可能な圧縮アルゴリズム

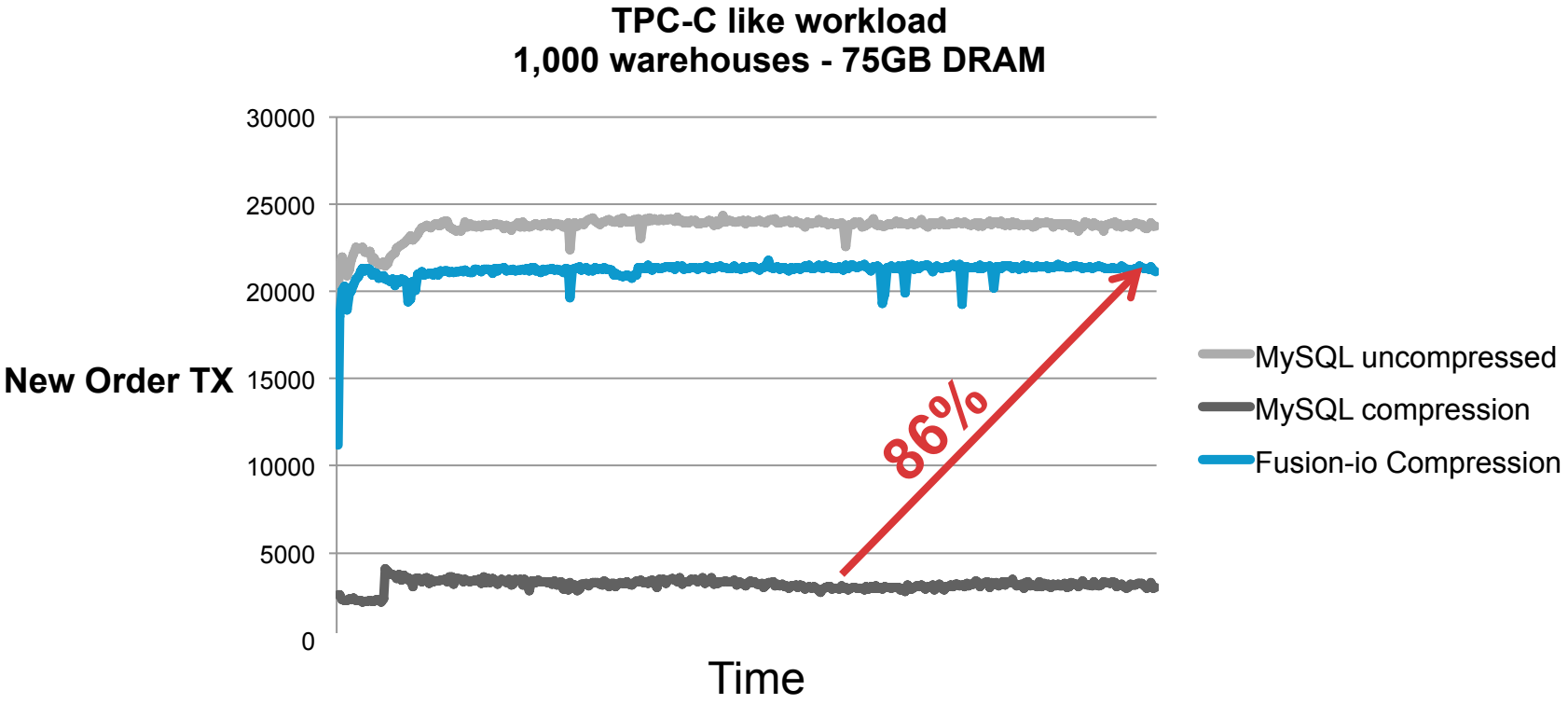


NVMコンプレッションの性能オーバーヘッドはごく僅か



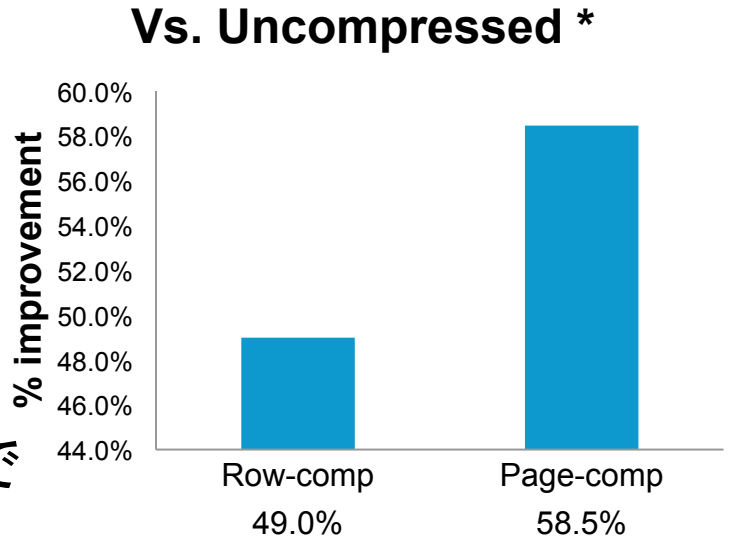
FUSION-IO

NVMコンプレッションの性能オーバーヘッドはごく僅か



圧縮により書き込み減少→容量の有効利用、長寿命化

- 従来のInnoDBストレージエンジンの行ベース圧縮を超える高圧縮率
- デバイスの耐用期間がアトミックライトと組み合わせで最大4倍に



*For LinkBench with lz77. Comparable results with lz4.

FUSION-IO

ファイルシステムからのミドルウェア高速化

NVM コンプレッションは、POSIXインターフェイスで実現

POSIXインターフェイス	動作
fallocate(offset, len)	既存ファイル／テーブルスペースの容量追加、プリアロケーション
fallocate(PUNCH_HOLE)	アンマップ（Punch Hole）操作。 デバイスに対しPersistent TRIMコマンド発行
io_submit()	非同期I/Oで透過的にアトミックライトを実現

新ファイルシステム”NVMFS”がNVMコンプレッションを高速化

NVMFS ― フラッシュメモリのためのファイルシステム

- **Non Volatile Memory FileSystem**
  （不揮発メモリ用ファイルシステム）
- Fusion-ioが開発した、POSIX準拠のファイルシステム

- **利点**
  - 大きなファイルのプリアロケーションを効率的に実現
  - ファイルシステムを使い続けても、”断片化”は発生しない
  - ファイルシステム経由でもデバイスのI/O性能が落ちづらい
  - アトミックライトやファイル内TRIMなどの機能を利用可能に

https://opennvm.github.io

http://www.opencompute.org/projects/storage

MySQL 5.7: InnoDB Compression

Thank you, Fusion-io

labs.mysql.com

- Transparent Page Level Compression
  - Happens transparently in background threads
  - Managed entirely within the IO layer
  - Uses sparse file and "hole punching" in OS kernels and File Systems
- Reduces IO
  - Improves performance
  - Reduces write cycles, thus increasing SSD lifespan
- Applies to all tables, including the system tablespace and UNDO logs

“Flash-aware MySQL” by Oracle

- アトミックライト対応
  - Oracle MySQL >= 5.7.4
- NVM コンプレッション対応
  - Oracle MySQL – labs release (http://labs.mysql.com/)
- NVMFS のアーリーアクセスがスタート！（クローズドベータ）


Thank You
fusionio.com | DELIVERING THE WORLD'S DATA. FASTER.