GPU アクセラレータと不揮発性メモリを考慮した I/O 性能の予備評価

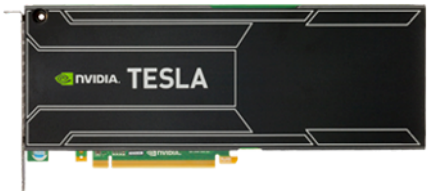
白幡 晃一[1,2]、佐藤 仁[1,2]、松岡 聡[1]

1: 東京工業大学

2: JST CREST

GPU と不揮発性メモリを用いた大規模データ処理

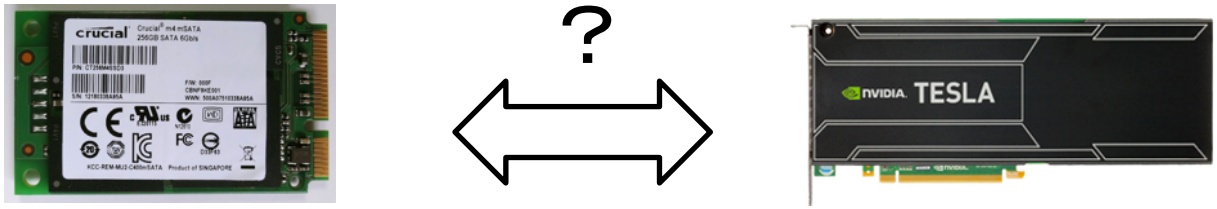
- 大規模データ処理
  - センサーネットワーク、遺伝子情報、SNS など
  - ペタ～ヨッタバイト級 → 高速処理が必要
- スーパーコンピュータ上での大規模データ処理
  - GPU
    - 高性能、高バンド幅
      - 例) Tesla K20X 3.95Tflops、250 GB/s
    - メモリ容量は ～5GB 程度 → データの退避が必要



  - 不揮発性メモリ
    - SSD, PCI-E接続型フラッシュメモリなど
    - 高バンド幅（数 GB/s）、高速 I/O（～1M IOPS）
    - 安価で高性能な mSATA SSD (mini SATA SSD) の出現

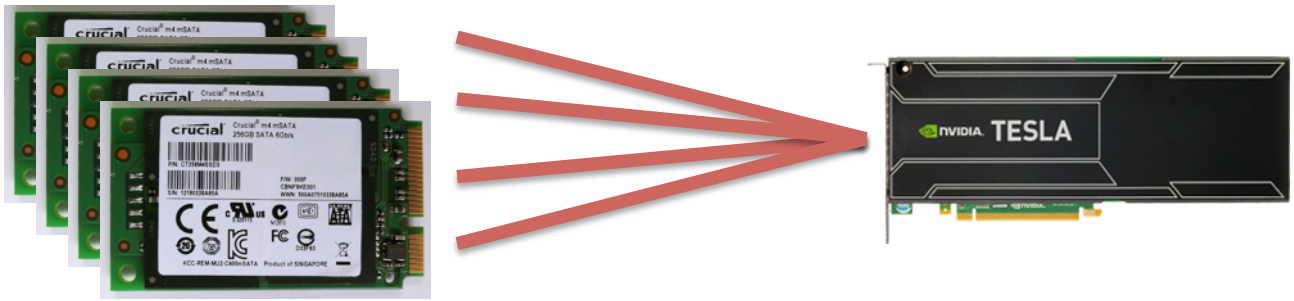
問題点

- スーパーコンピュータ上でのローカルディスクの最適な構成方法は明らかではない
  - ローカルディスクの構成方法
    - 最適な不揮発性メモリの選択
    - 不揮発性メモリを用いたマシンの構成方法
  - ローカルディスクから GPU への最適な I/O 手法
    - ローカルディスクの設定、GPU への転送手法、転送粒度



解決手法

- 不揮発性メモリから GPU への I/O 手法の比較
  - 複数 mSATA SSD を用いたプロトタイプマシンの設計
    - mSATA SSD のバンド幅を最大限に引き出す設計
    - 既存の不揮発性メモリとの性能比較
  - 複数 mSATA SSD から GPU への I/O 性能評価
    - 複数 mSATA SSD の構成、設定、I/O手法、GPU への転送手法、転送粒度



目的と成果

- 目的
  - 不揮発性メモリから GPU への最適な I/O 手法を把握
- 成果
  - **16 枚の mSATA SSD を用いたプロトタイプマシンの設計**
  - **複数 mSATA SSD の I/O 基本性能の評価**
    - **16枚の mSATA SSD で 7.39 GB/s (理論ピークの 92.4%)**
    - **8枚の mSATA SSD で PCI-E 接続型フラッシュメモリに対して 3.20～7.60 倍 の Read 性能**
  - **複数 mSATA SSD から GPU への I/O 性能の予備評価**
    - **8枚の mSATA SSD から GPU へ 3.06GB/s のスループット**

発表の流れ

1. 背景
2. 複数 mSATA SSD を用いた予備評価
   1. プロトタイプマシンの設計
   2. I/O ベンチマークを用いた評価
   3. 既存の不揮発性メモリとの性能比較
3. 複数 mSATA SSD と GPU を用いた予備評価
   1. プロトタイプマシンの設計
   2. ベンチマークアプリケーションの実装
   3. 予備評価
4. 関連研究
5. まとめ

発表の流れ

1. 背景
2. 複数 mSATA SSD を用いた予備評価
    1. プロトタイプマシンの設計
    2. I/O ベンチマークを用いた評価
    3. 既存の不揮発性メモリとの性能比較
3. 複数 mSATA SSD と GPU を用いた予備評価
    1. プロトタイプマシンの設計
    2. ベンチマークアプリケーションの実装
    3. 予備評価
4. 関連研究
5. まとめ

複数枚の mSATA SSD を用いた I/O

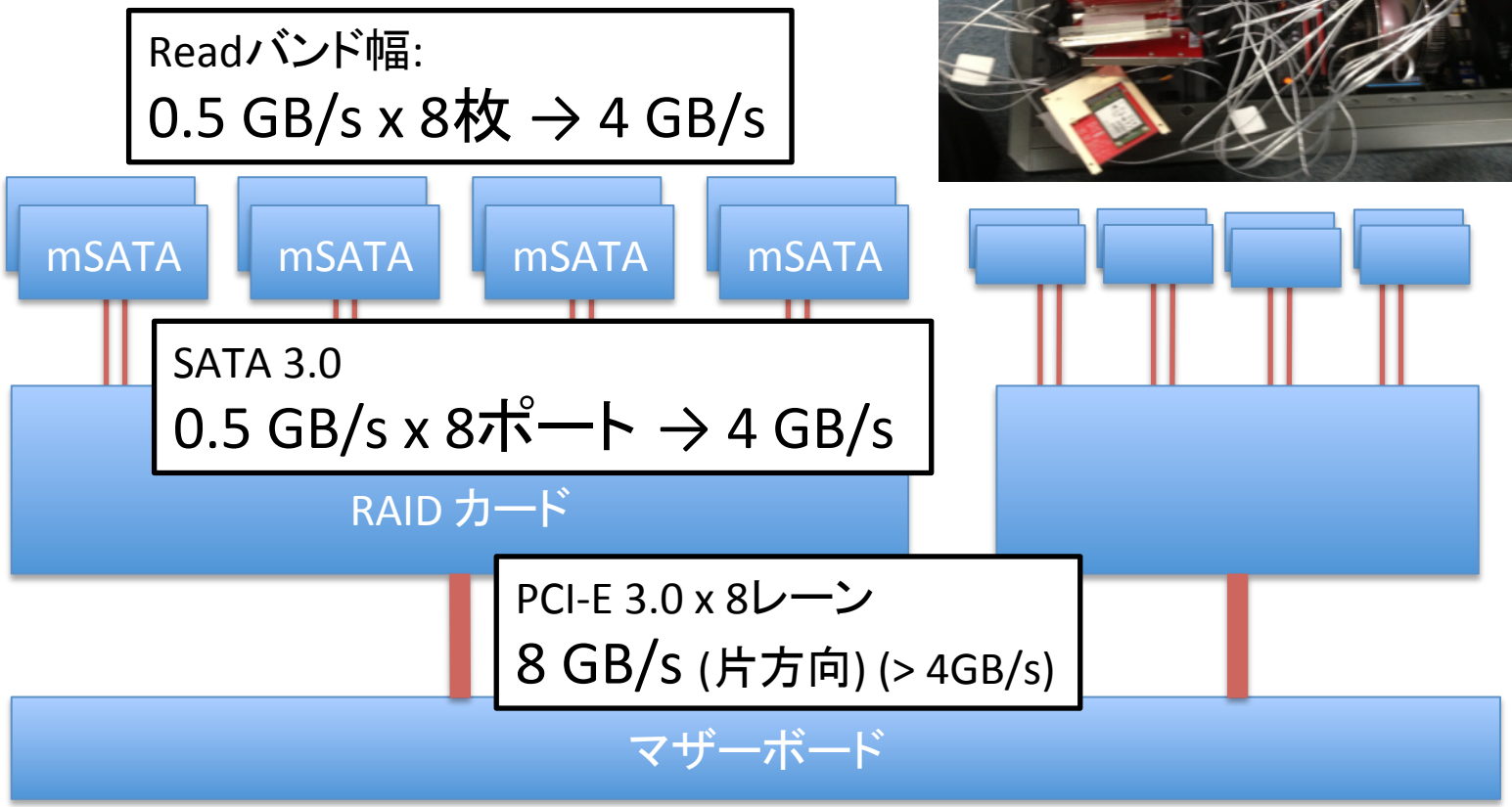
- mini SATA SSD (mSATA SSD)
  - mSATA: SATA 規格コネクターの仕様の一つ
  - mSATA SSD: mSATA 接続の SSD
    - 通常の SSD に比べ面積が小さい
- 複数枚の mSATA SSD を組み合わせて使用
  - 高いコストパフォーマンスを実現可能
    - 例) crucial® m4 msata SSD 256 GB:
      - Read: 500 MB/s、Write: 260 MB/s
      - 平均アクティブ時消費電力: <200mW
      - $260 ～ $300
    - SSD に比べ、設置面積・消費電力で優位
    - PCI-E 接続型フラッシュメモリに比べ、価格・バンド幅で優位

複数 mSATA SSD を用いた プロトタイプマシンの設計

mSATA SSD:

容量: 256GB x 16枚 → 4TB

Readバンド幅: 0.5GB/s x 16枚 → 8 GB/s



複数 mSATA SSD プロトタイプマシン上での予備評価

- 複数 mSATA SSD の基本 I/O 性能評価
  - ハードウェア RAID の有無
    - Raw デバイス (ハードウェアRAIDを組まない)
      - 1枚毎にマウントし、それぞれ Ext4 でファイルシステムを作成
      - ブロックサイズ(1枚当たり): 4KB
    - RAID 0
      - ストライプサイズ: 64KB, 1MB
      - キャッシュ機能の ON・OFF
  - 複数 mSATA SSD のスケーラビリティ
    - 枚数: 1, 2, 4, 8, 16
- 他の不揮発性メモリとの比較
  - SSD
  - PCI-E 接続型フラッシュメモリ

複数 mSATA SSD プロトタイプマシン上での予備評価

- 複数 mSATA SSD の基本 I/O 性能評価
  - ハードウェア RAID の有無
    - Raw デバイス (ハードウェアRAIDを組まない)
      - 1枚毎にマウントし、それぞれ Ext4 でファイルシステムを作成
      - ブロックサイズ(1枚当たり): 4KB
    - RAID 0
      - ストライプサイズ: 64KB, 1MB
      - キャッシュ機能の ON・OFF
  - 複数 mSATA SSD のスケーラビリティ
    - 枚数: 1, 2, 4, 8, 16
- 他の不揮発性メモリとの比較
  - SSD
  - PCI-E 接続型フラッシュメモリ

複数 mSATA SSD の基本 I/O 性能評価

- 目的: 複数 mSATA SSD の基本 I/O 性能を確認
- シーケンシャル Read, Write の測定
- fio の設定
  - I/O エンジン: libaio
  - I/O queue depth: 1
  - I/O ブロックサイズ
    - シーケンシャル Read: 4MB
    - シーケンシャル Write: 4MB
  - 使用したデータサイズ: 200GB (1 mSATA SSD 当たり)

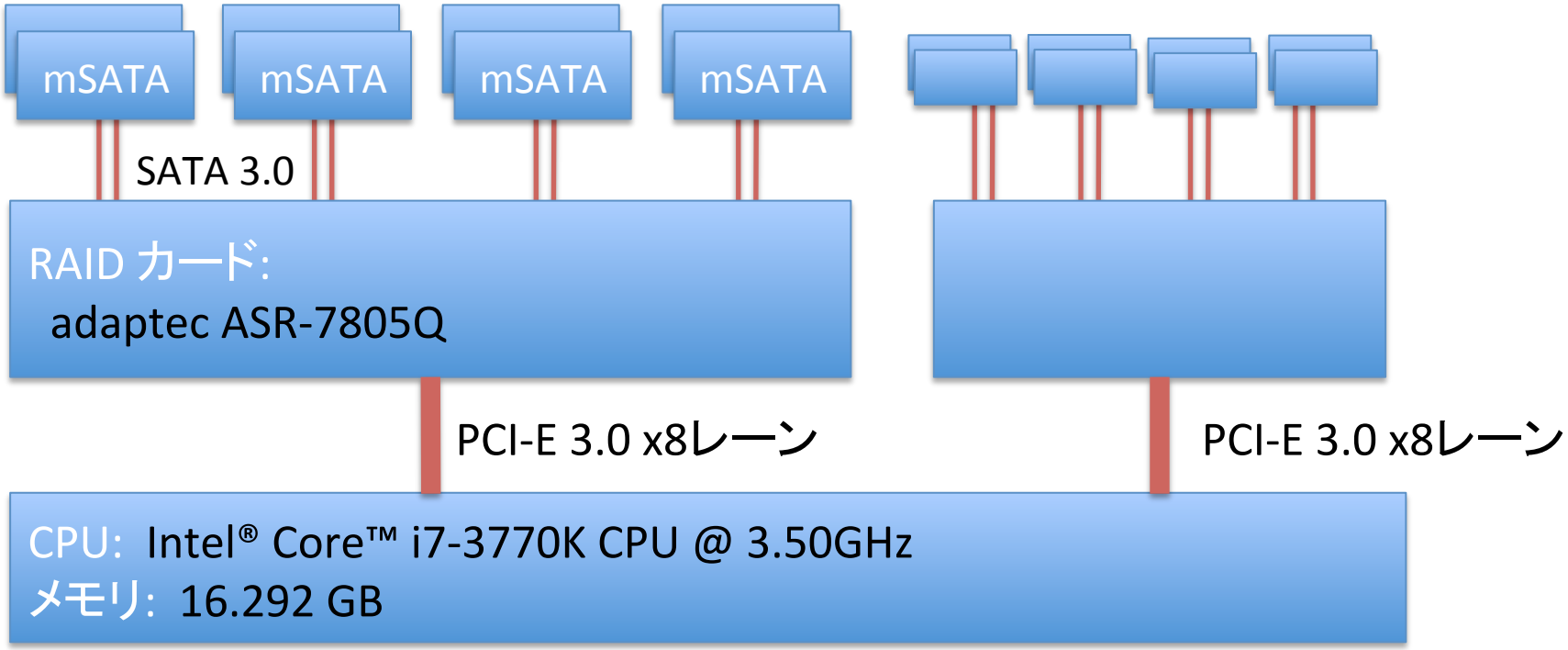
評価環境

mSATA SSD:

crucial® m4 msata 256GB SATA 6Gbps

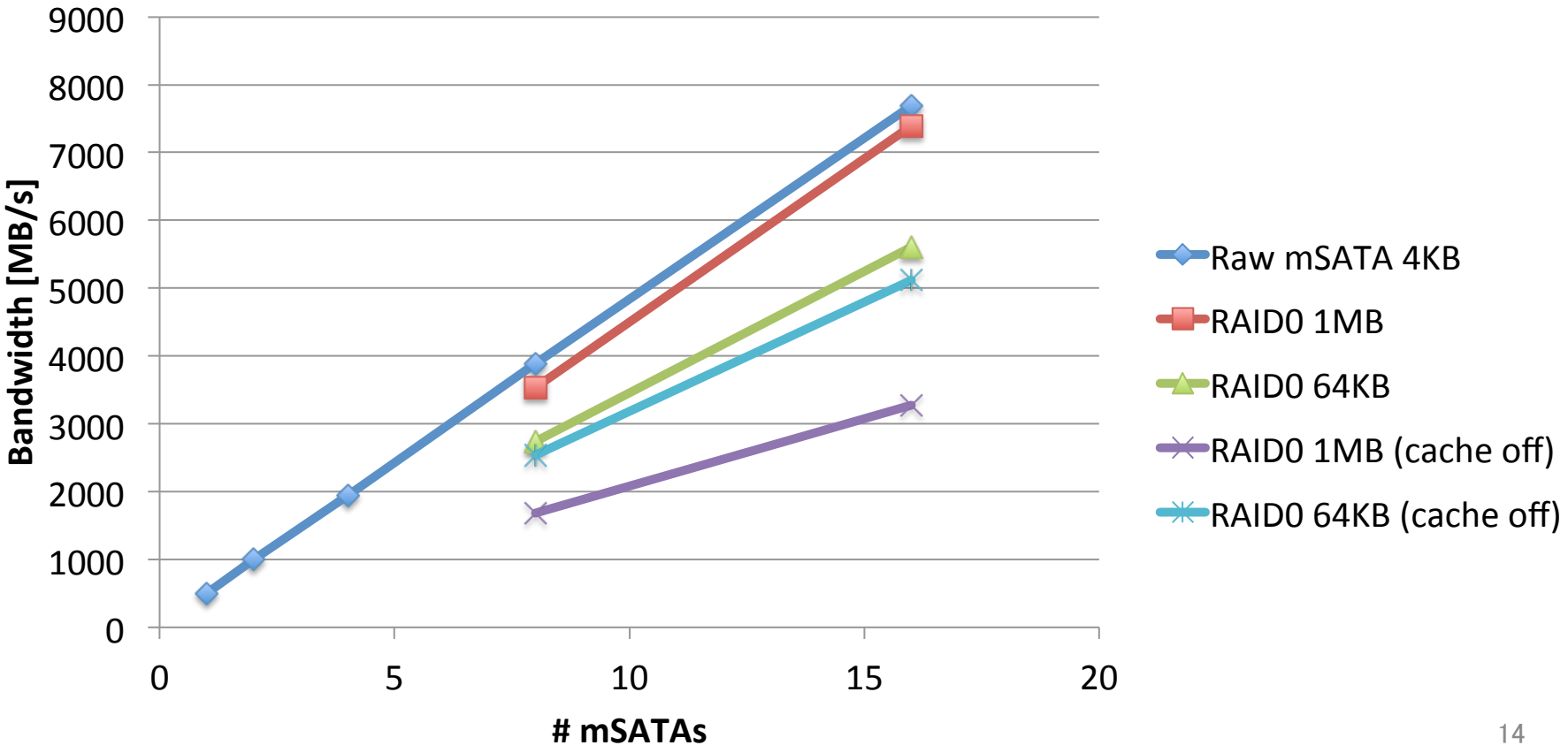
Readバンド幅: 0.5GB/s x 16枚 = 8 GB/s

Writeバンド幅: 0.26 GB/s x 16枚 = 4.16 GB/s



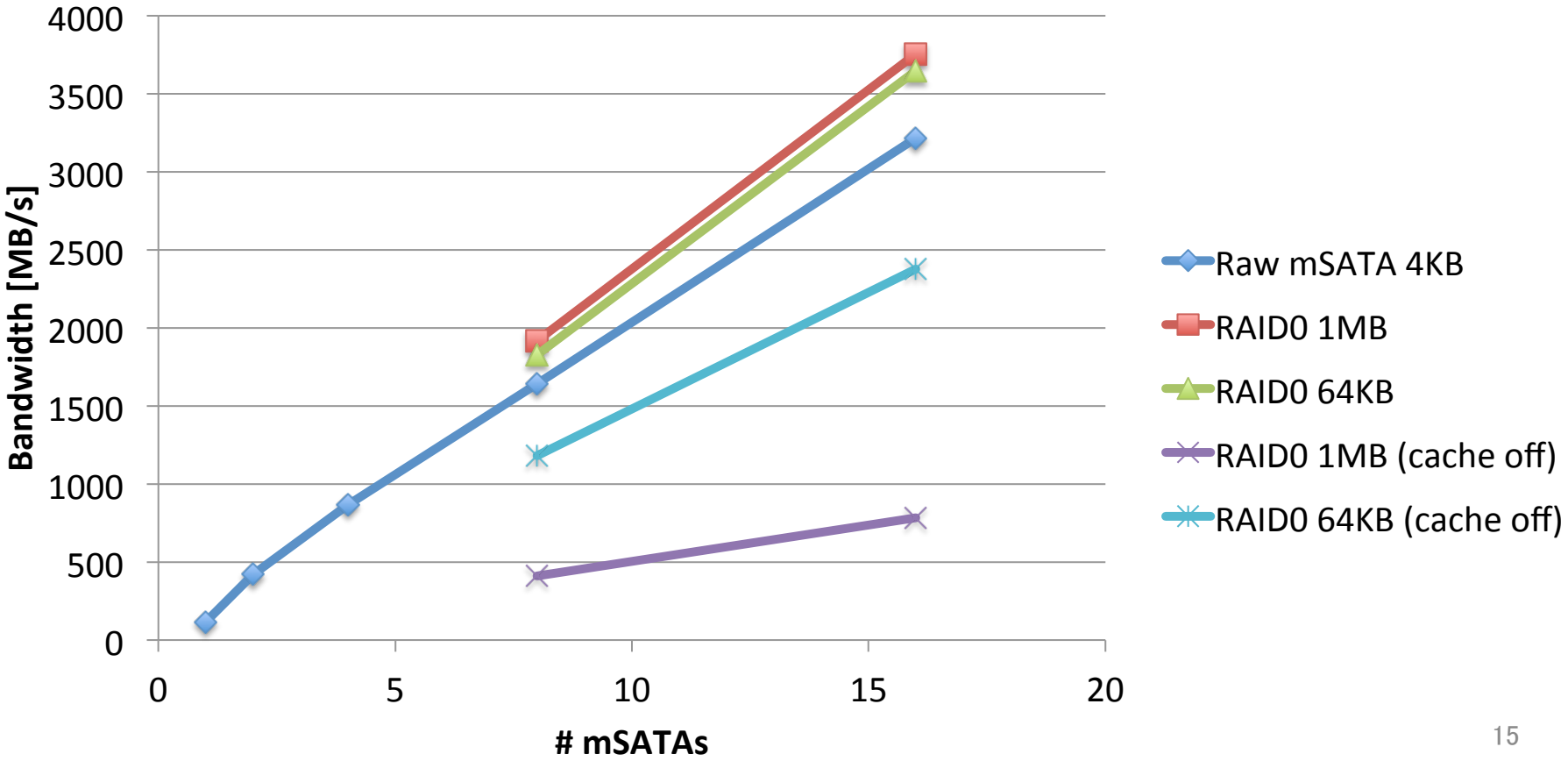
Read バンド幅の測定結果

- Raw デバイスは RAID0 に対して 10% 程度高速
  - 7.69 GB/s, 理論ピークの 96.2%
- RAID0 でもストライプサイズを大きくすれば高性能
  - 1MB では 7.39 GB/s (理論ピークの 92.4%)



Write バンド幅の測定結果

- RAID0 1MB が最も高速 (3.75 GB/s, 理論ピークの 90.2%)
  - RAIDカードが遅延書き込みの最適化を行っている可能性

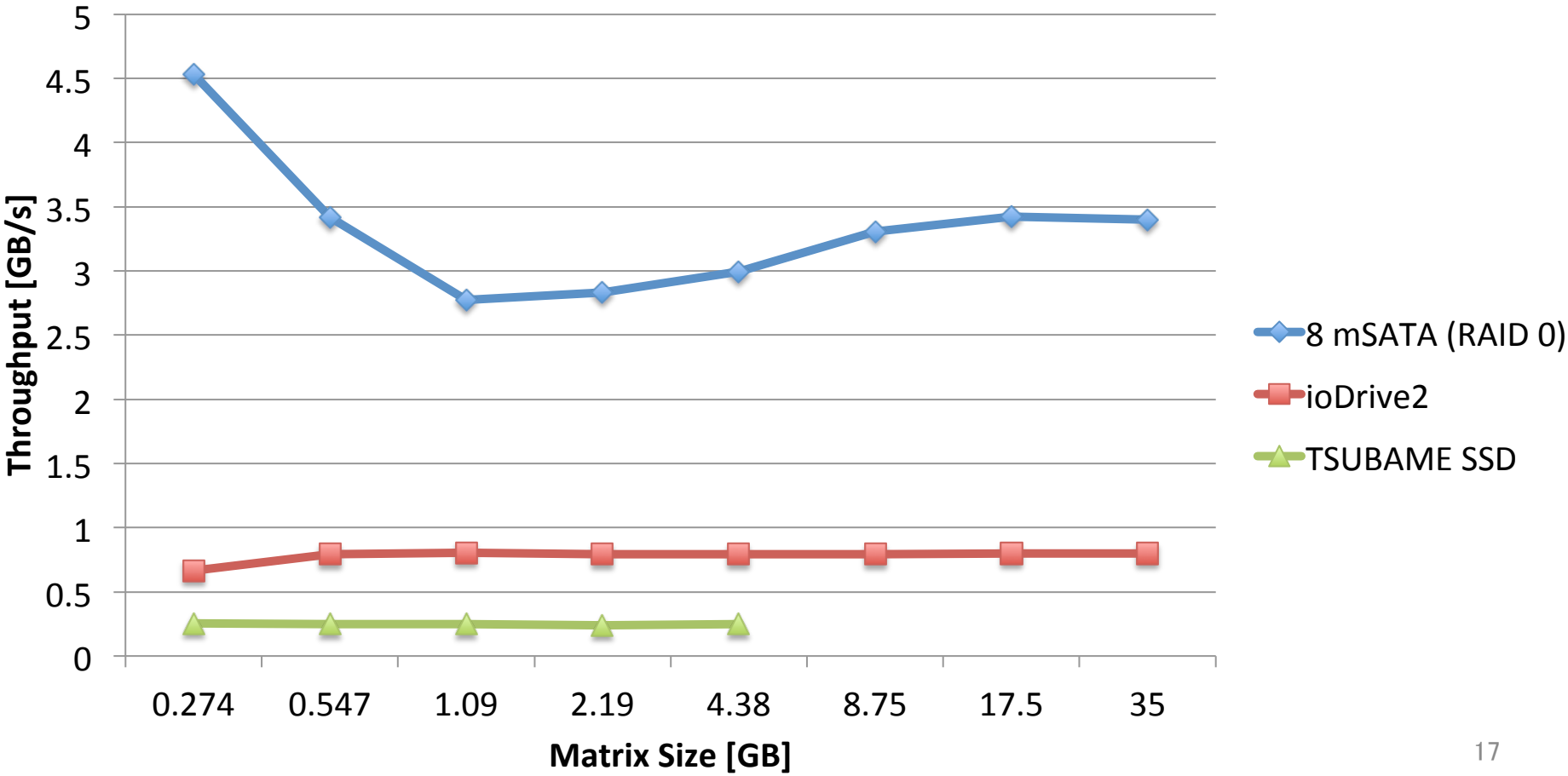


他の不揮発性メモリとの性能比較

- 目的: 他の不揮発性メモリとの性能差を把握
- 比較対象
  - SSD
    - TSUBAME 2.0 の計算ノードに搭載されているローカル SSD を使用
  - PCI-E 接続型フラッシュメモリ
    - Fusion IO 社の ioDrive2 を使用
    - Readバンド幅: 1.4 GB/s
- CPU 上での密行列ベクトル積を用いて比較
  - シーケンシャル Read が実行時間の多くを占める

他のデバイスとの比較結果

- mSATA SSD 8枚の方が ioDrive2 より 3.20～7.60倍 高速
- mSATA SSD はシーケンシャルI/O性能に優れる



複数 mSATA SSD を用いた実験のまとめ

- Read, Write ともに RAID0 を使用した場合に良好な性能
  - 理論ピークの 90% 以上の性能
  - 複数 mSATA SSD でスケールを確認
  - ストライプサイズを大きく設定することにより、複数 mSATA SSD のバンド幅を活かせている
- 複数 mSATA SSD を用いることにより既存の不揮発性メモリに対して高いスループット
  - ioDrive2 より 3.20～7.60 倍高速
  - PCI-E 接続型フラッシュメモリは IOPS に特化しているため

複数 mSATA SSD を用いた実験のまとめ

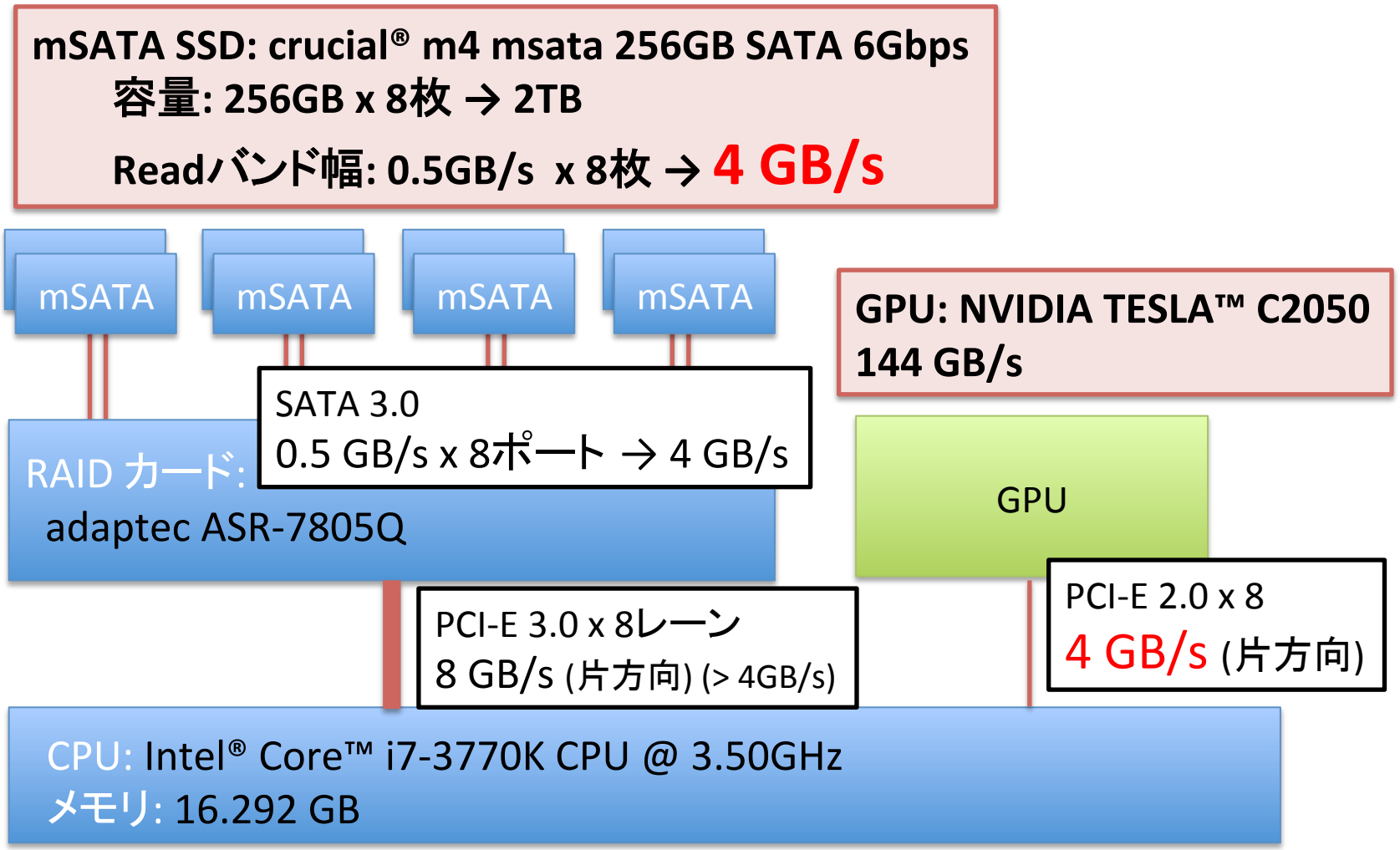
- Read, Write ともに RAID0 を使用した場合に良好な性能
  - 理論ピークの 90% 以上の性能
  - 複数 mSATA SSD でスケールを確認
  - ストライプサイズを大きく設定することにより、複数 mSATA SSD のバンド幅を活かせている
- 複数 mSATA SSD を用いることにより既存の不揮発性メモリに対して高いスループット
  - ioDrive2 より 3.20～7.60 倍高速
  - PCI-E 接続型フラッシュメモリは IOPS に特化しているため

複数 mSATA SSD マシンの有効性を確認

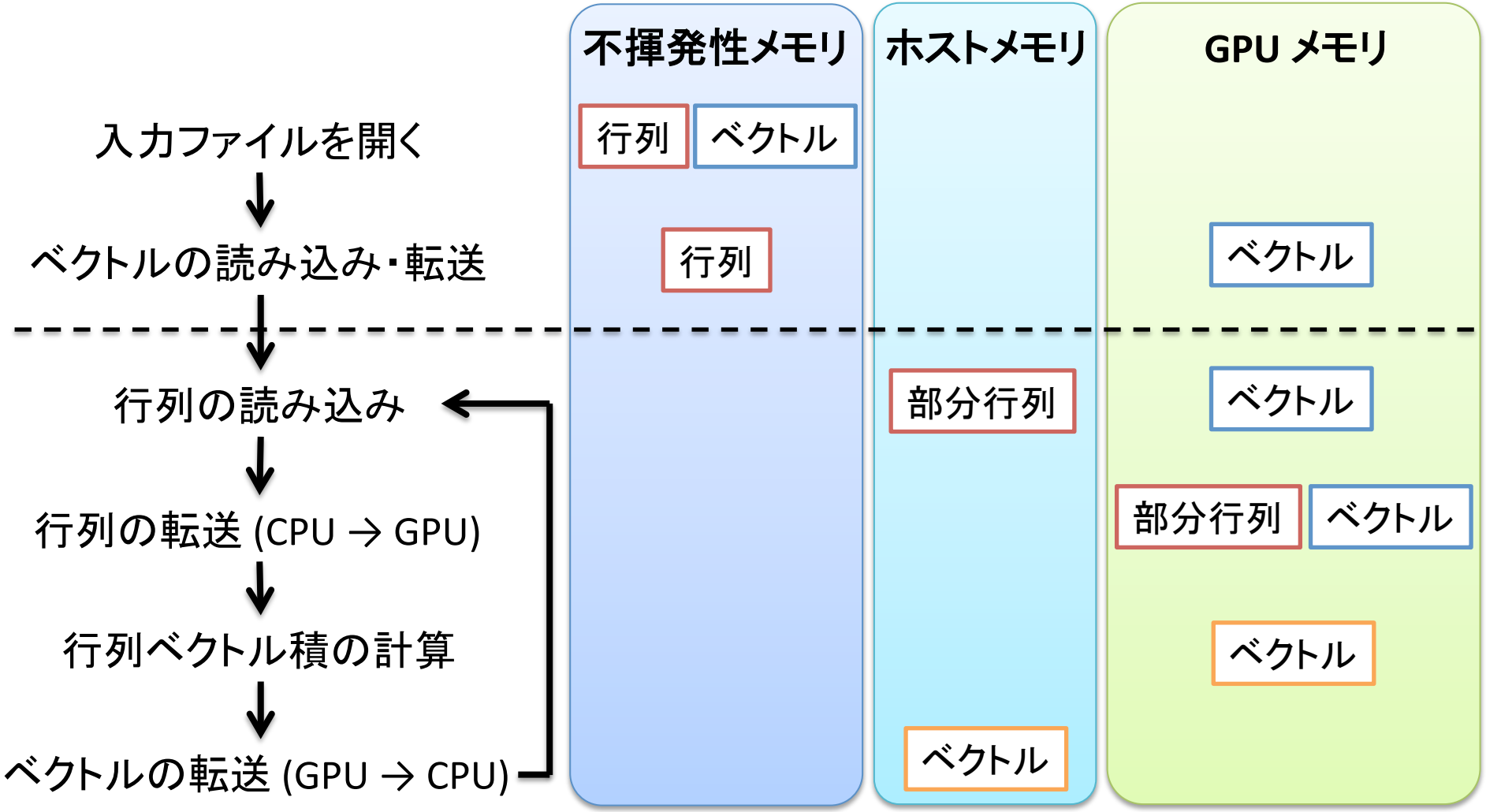
発表の流れ

1. 背景
2. 複数 mSATA SSD を用いた予備評価
   1. プロトタイプマシンの設計
   2. I/O ベンチマークを用いた評価
   3. 既存の不揮発性メモリとの性能比較
3. 複数 mSATA SSD と GPU を用いた予備評価
   1. プロトタイプマシンの設計
   2. ベンチマークアプリケーションの実装
   3. 予備評価
4. 関連研究
5. まとめ

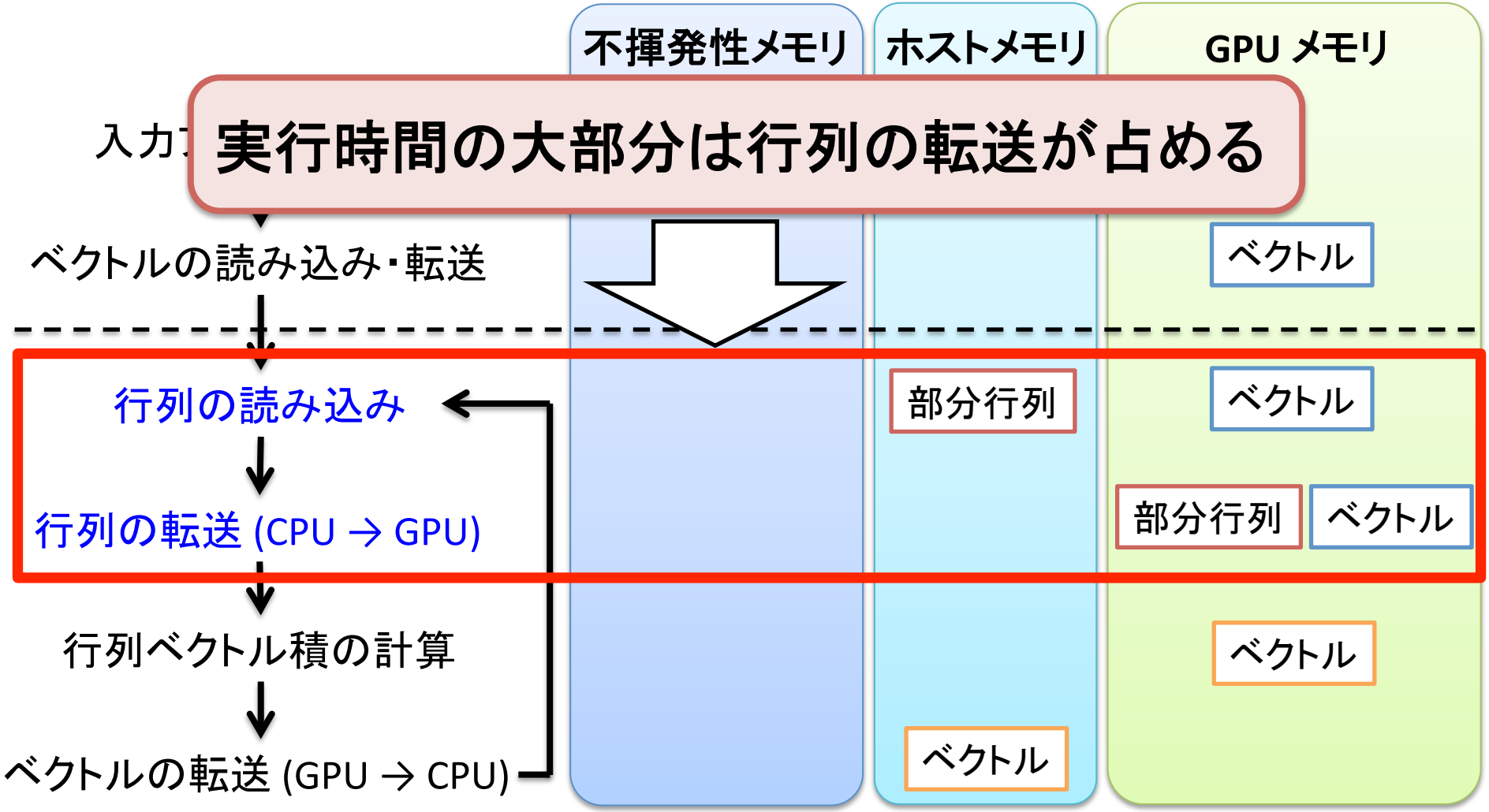
複数 mSATA SSD と GPU を用いた プロトタイプマシンの設計



不揮発性メモリと GPU を用いた密行列ベクトル積の実装



不揮発性メモリとGPUを用いた密行列ベクトル積の実装

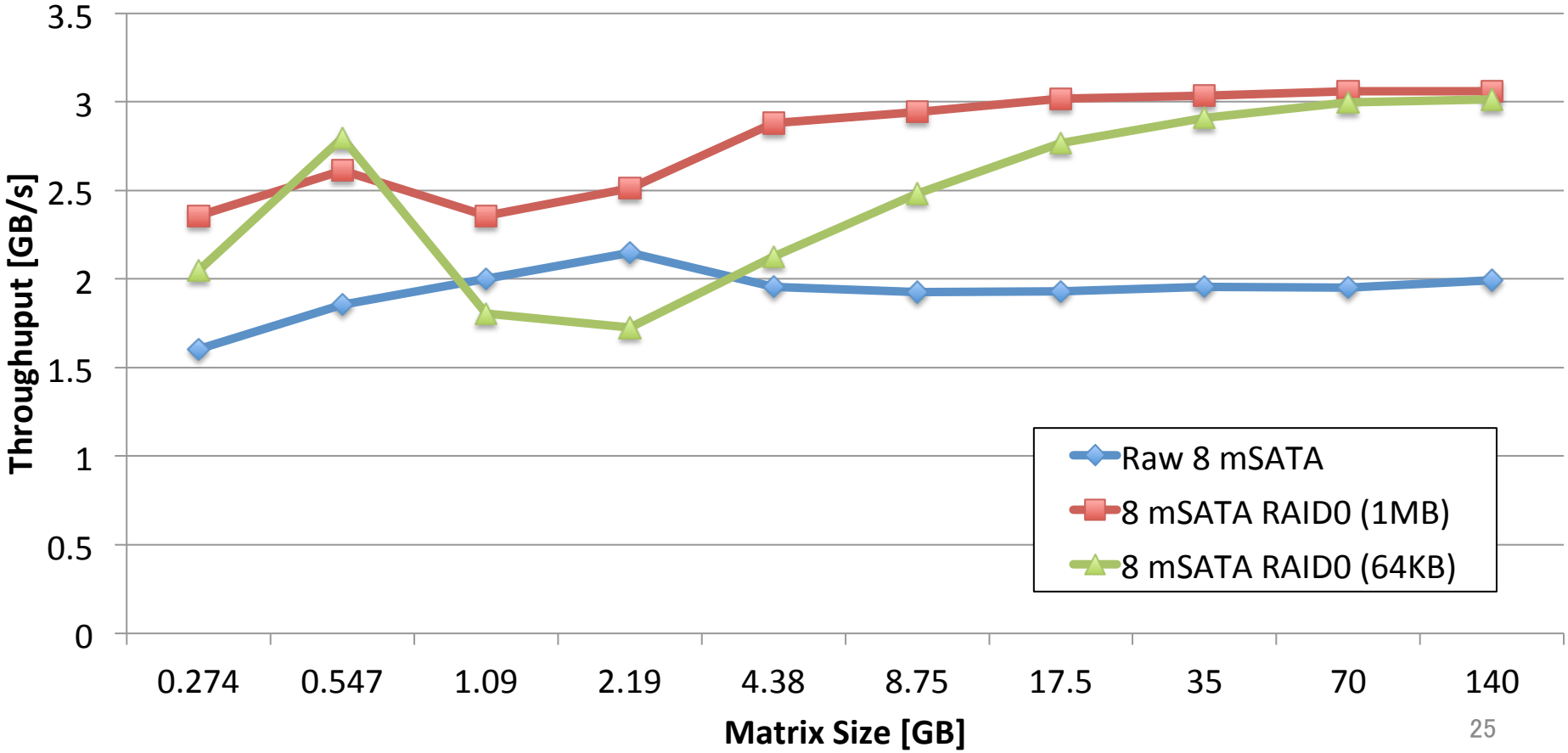


複数 mSATA SSD から GPU への I/O 性能の予備評価

- 目的: 複数 mSATA SSD から GPU への基本 I/O 性能を確認
- 評価方法
  - 密行列ベクトル積ベンチマークを使用
  - 行列データサイズ: 280 MB ～ 140 GB まで変化させて実験
- 比較内容
  - ハードウェアRAIDの有無
    - RAID0 を使用するか、使用する場合のストライプサイズ
  - mSATA SSD からの読み込み手法
    - mmap, pread
  - ホストメモリから GPU への DMA 転送の有無
    - Pinned メモリの使用の有無
  - データ転送粒度
    - 35, 70, 140, 280, 560 MB
  - GPU の使用の有無
    - ホストメモリ上で CPU が計算する場合との比較

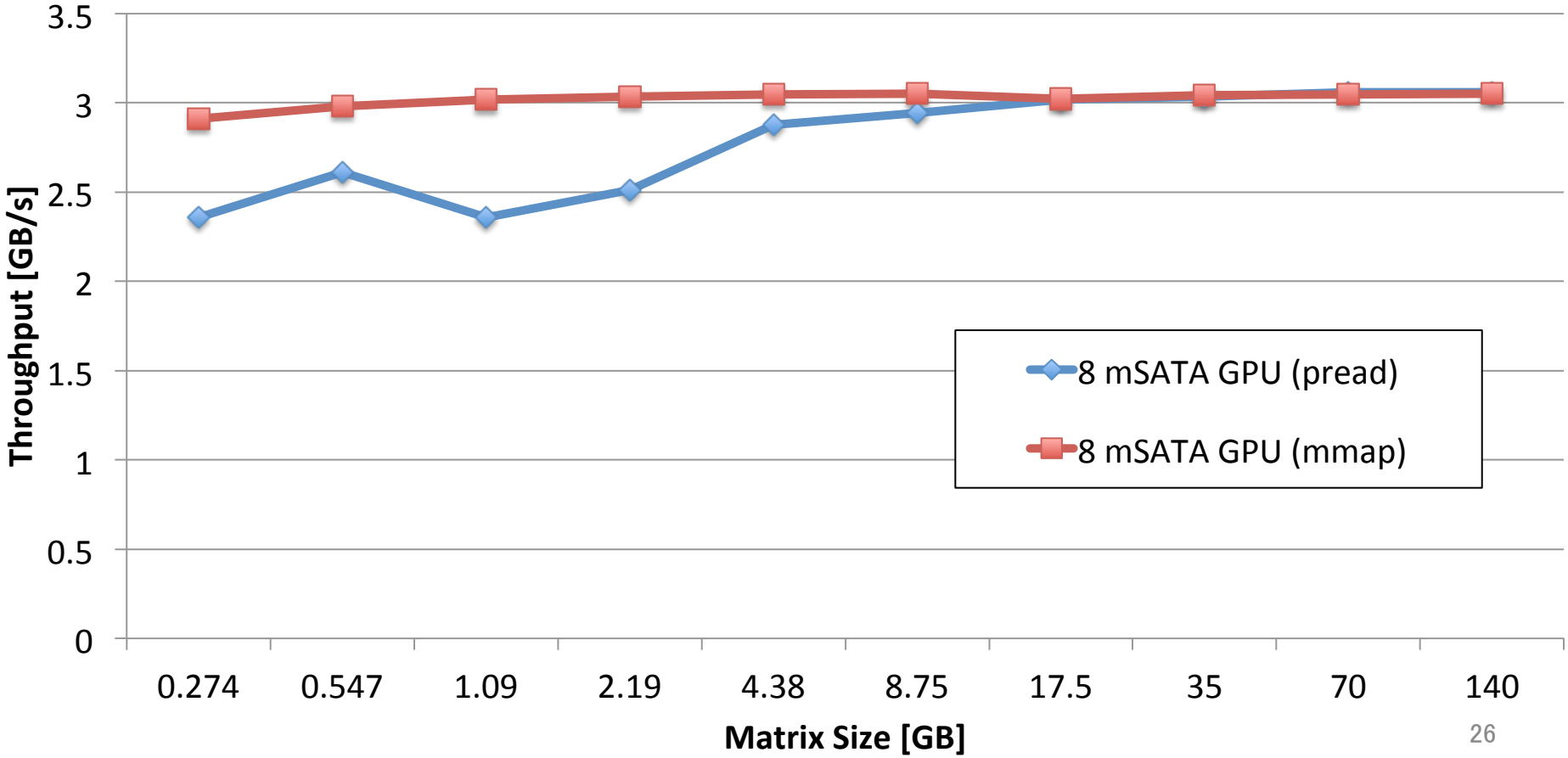
RAID0 と Raw デバイスの比較

- Raw デバイスの場合は OpenMP で並列読み込み
- 行列の転送粒度は 70MB
- RAID0 1MB が最も高速
  - OpenMP による並列読み込みが最適化されていない可能性



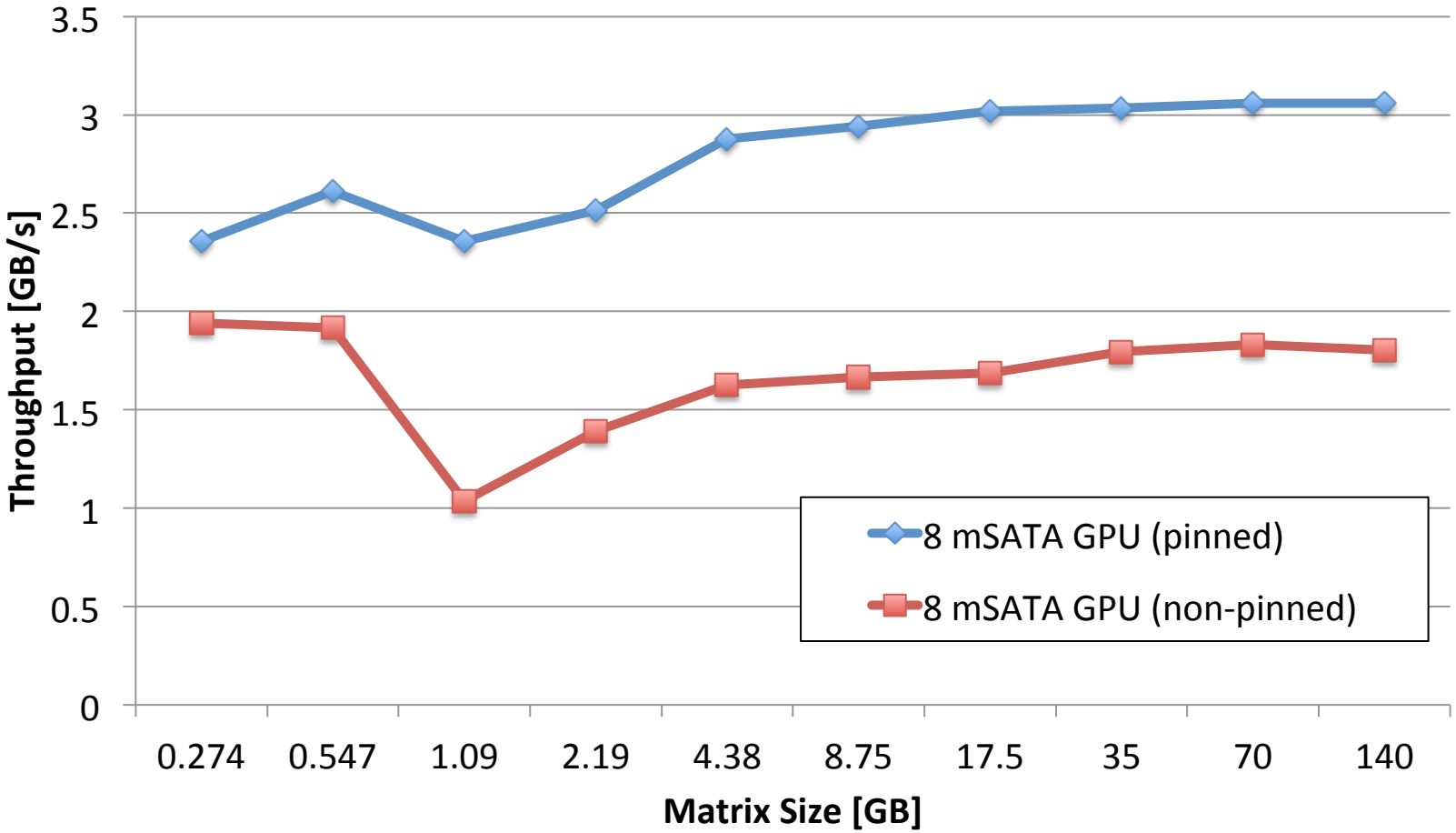
読み込み手法による比較

- mmap, pread を比較
- 行列サイズが大きい場合は同等の性能
- 行列サイズが小さい場合は mmap の方が高速
  - pread の粒度を 70MB としたため、オーバーラップが不十分のため



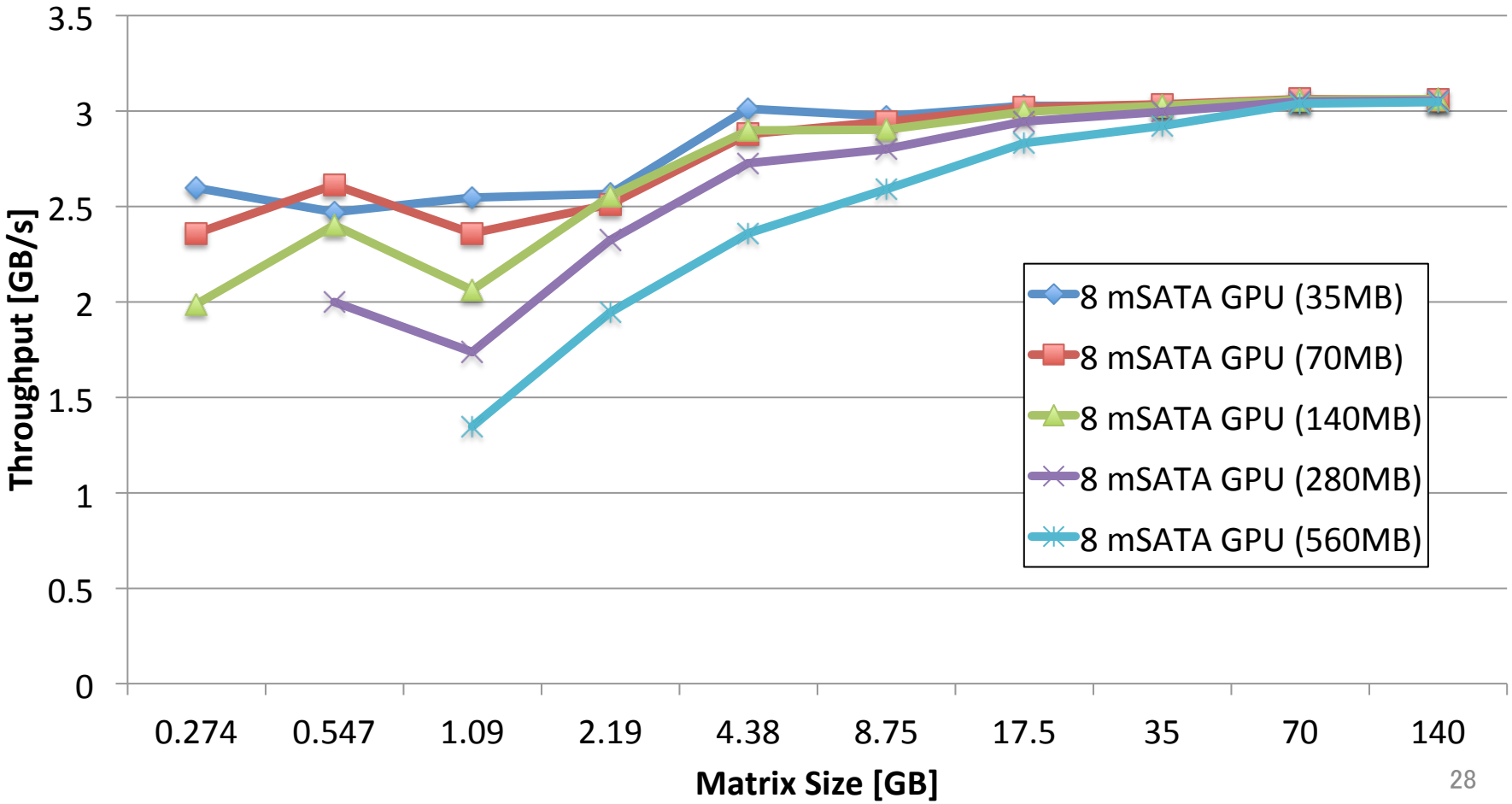
ホストメモリから GPU メモリへの DMA 転送の有無による比較

- Pinned メモリを使用した方が 1.21～2.28倍 高速
  - DMA 転送による効果



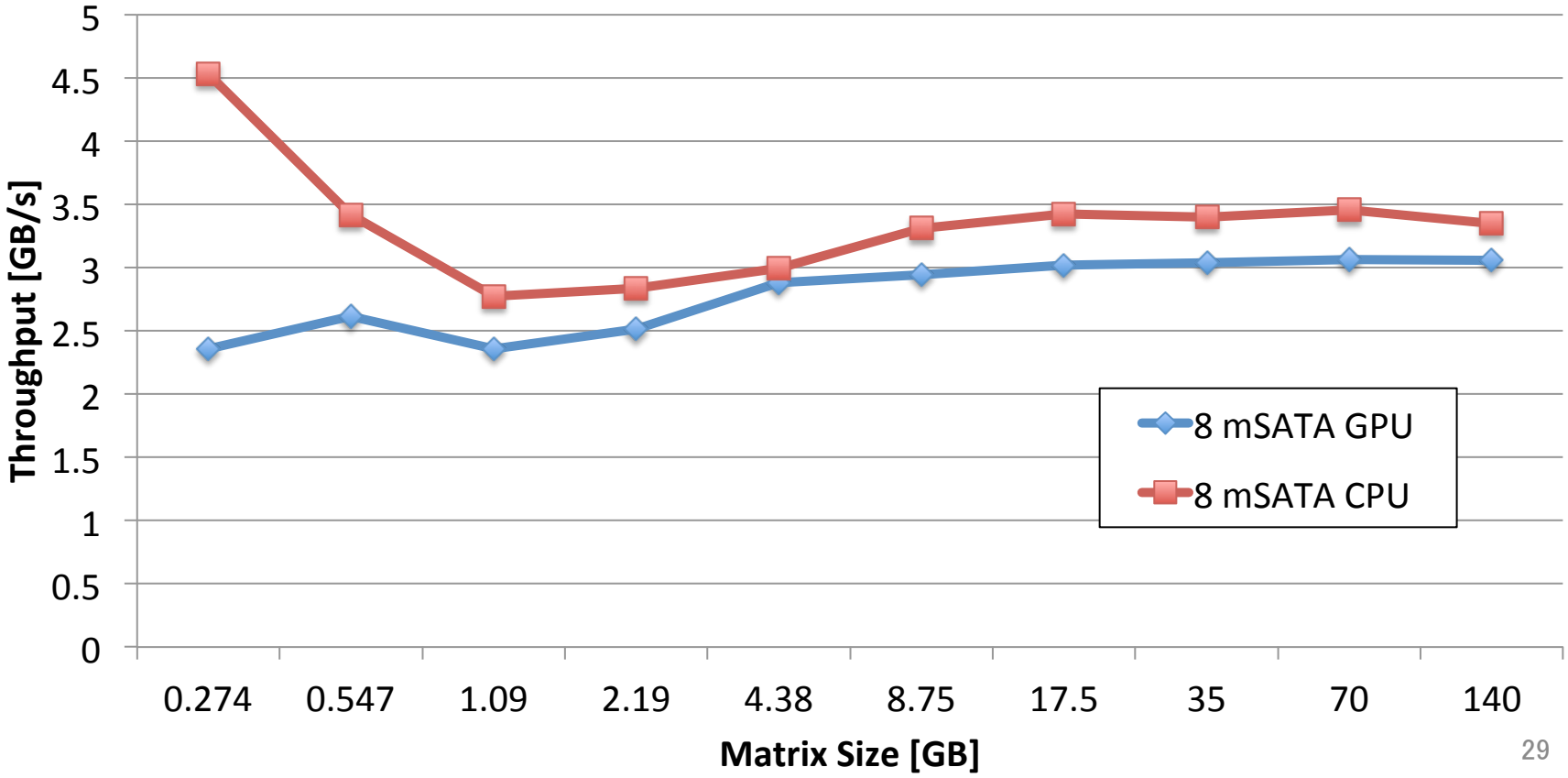
行列の転送粒度による比較

- 35～70MB の粒度が最も高速
  - 粒度が大きいと オーバーラップ領域が小さくなるため



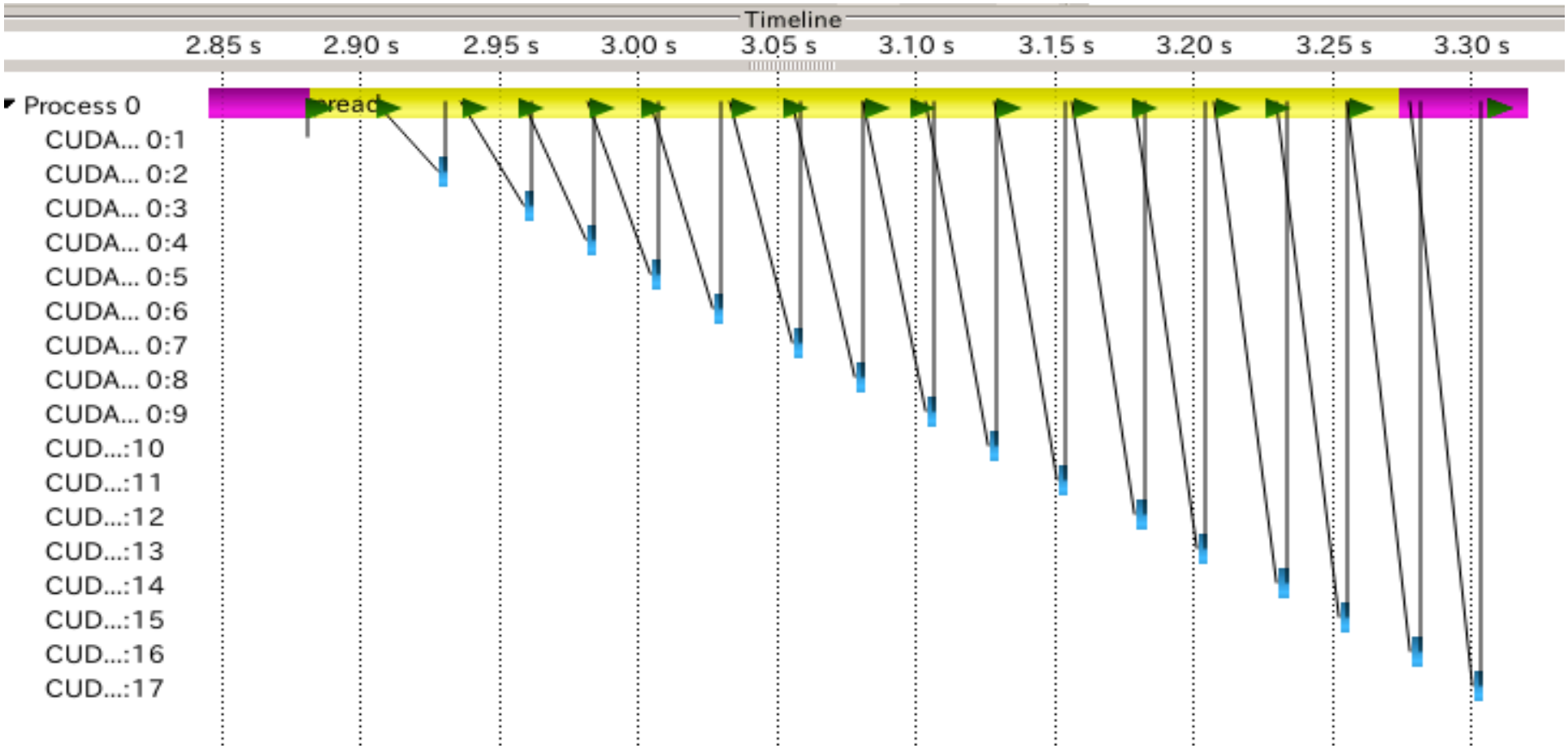
GPU を使用しない場合との比較

- CPU を使用した方が高速
  - GPU を使用した場合は PCI-E のバンド幅 (3.06 GB/s) に律速
  - PCI-E バンド幅の上限が上がれば、CPUと同等のスループットになると考えられる



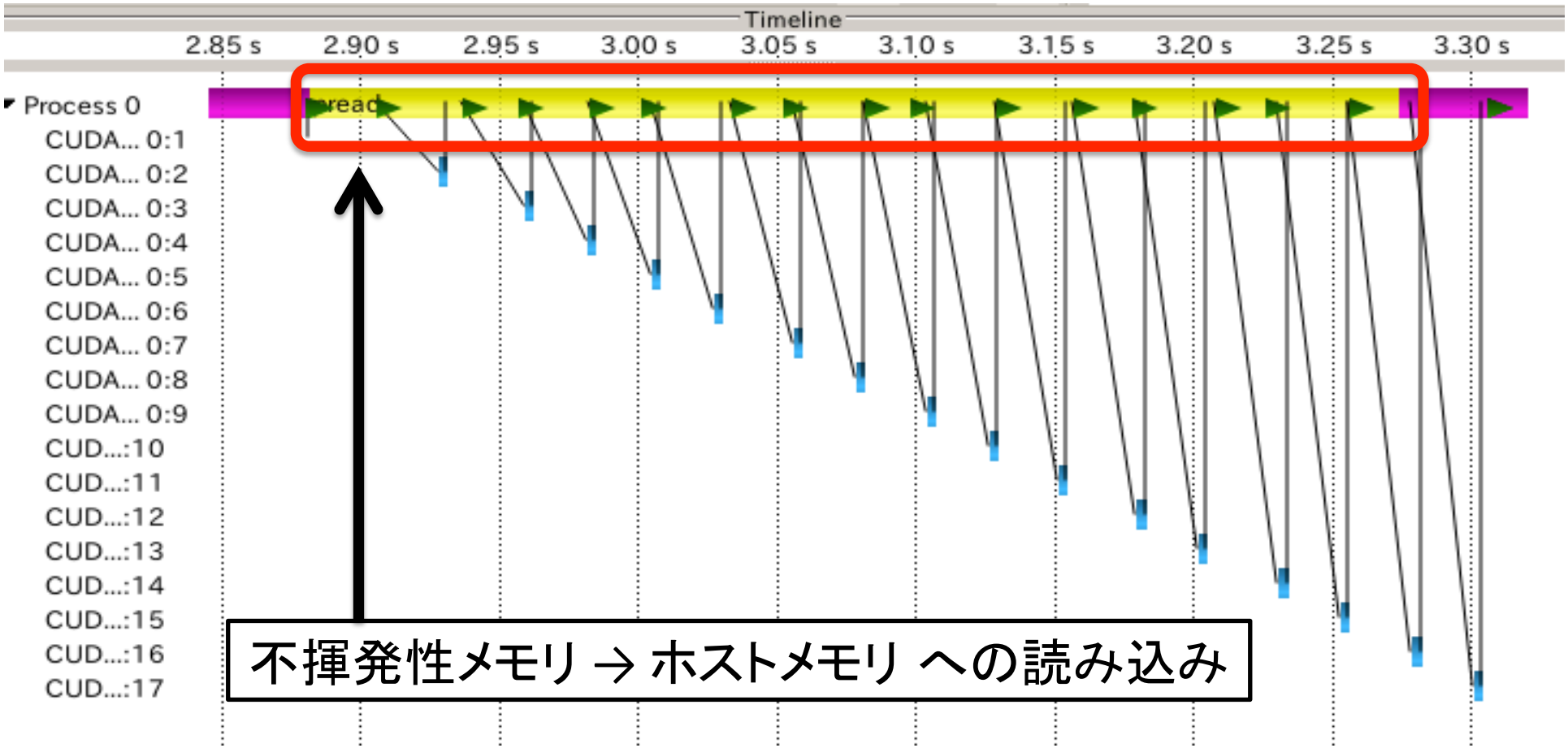
ブレイクダウン

- Vampir Trace を使用
- 行列サイズが 1.12 GB の場合



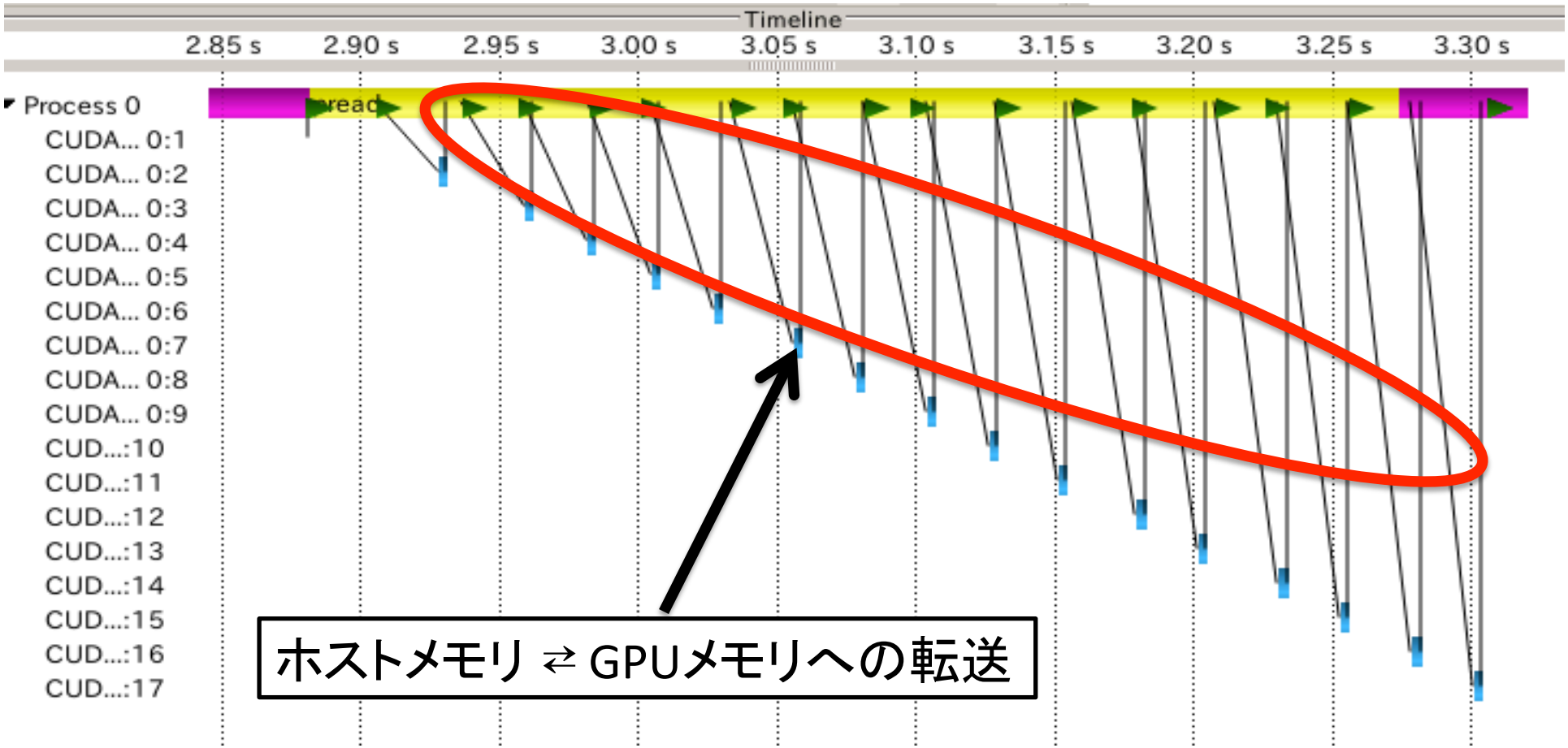
ブレイクダウン

- Vampir Trace を使用
- 行列サイズが 1.12 GB の場合



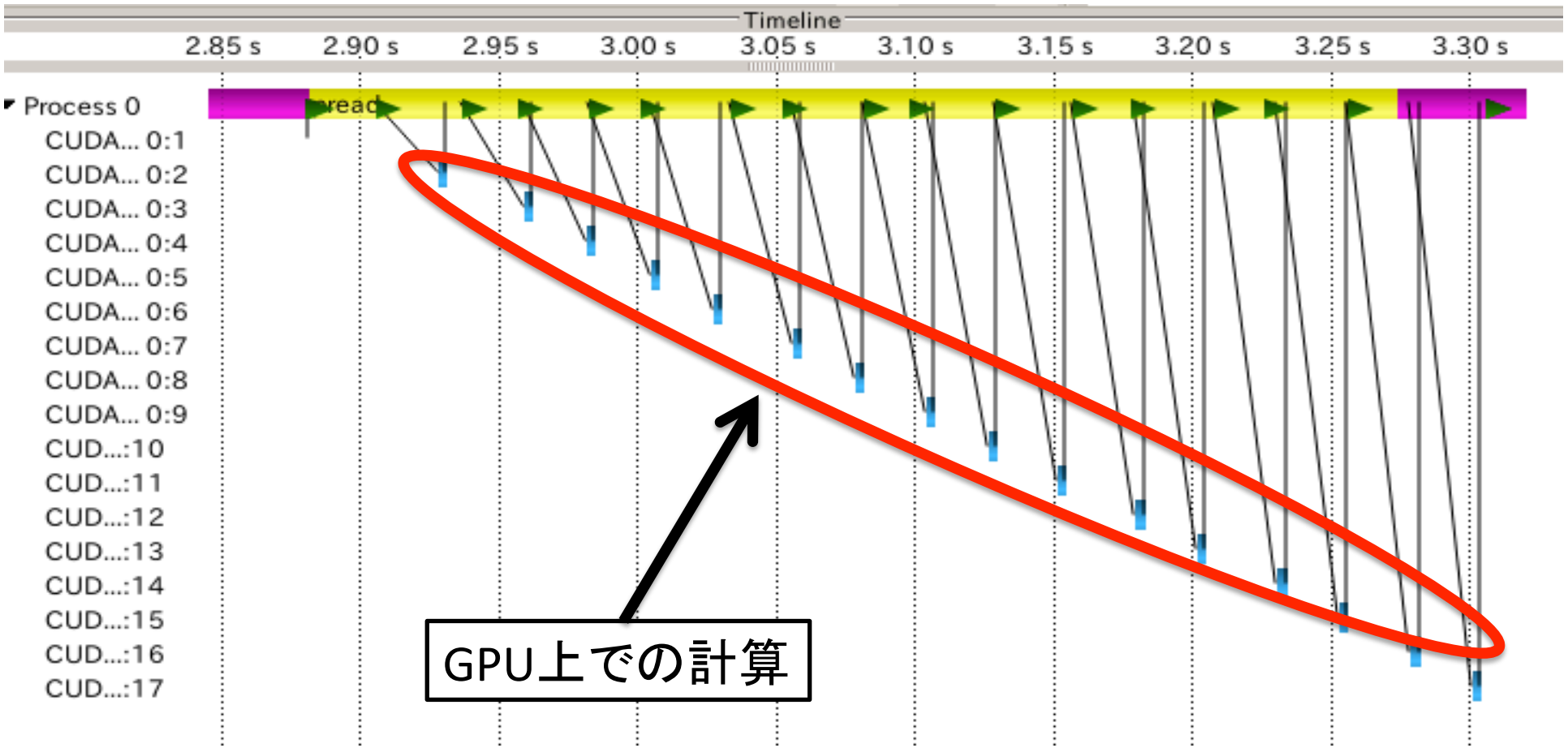
ブレイクダウン

- Vampir Trace を使用
- 行列サイズが 1.12 GB の場合



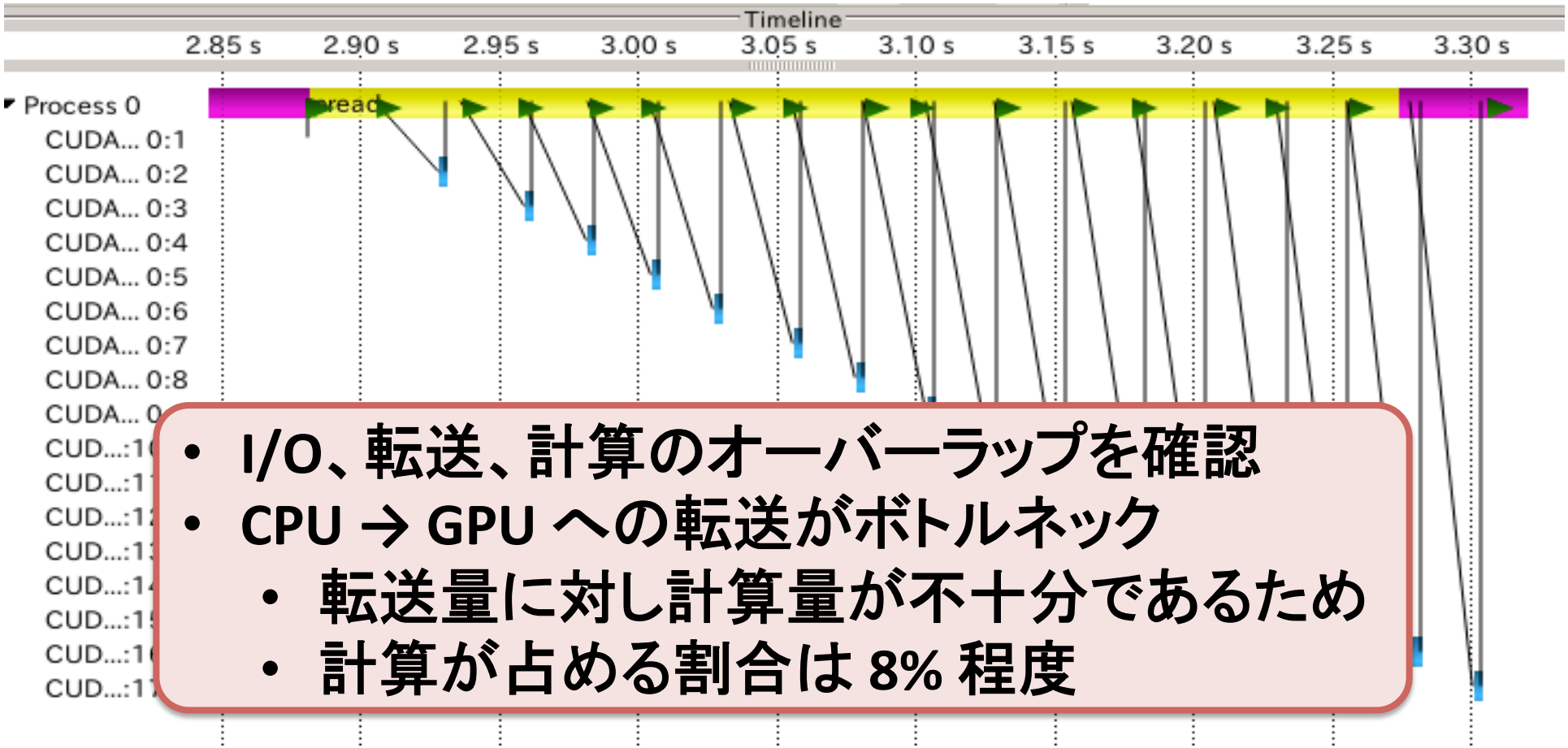
ブレイクダウン

- Vampir Trace を使用
- 行列サイズが 1.12 GB の場合



ブレイクダウン

- Vampir Trace を使用
- 行列サイズが 1.12 GB の場合



- **I/O、転送、計算のオーバーラップを確認**
- **CPU → GPU への転送がボトルネック**
  - **転送量に対し計算量が不十分であるため**
  - **計算が占める割合は 8% 程度**

複数 mSATA SSD と GPU を用いた実験のまとめ

- 最適な I/O 手法の選択により，複数 mSATA SSD から GPU へ PCI-E 上限(3.06 GB/s)のスループット
  - RAID0 を使用し、ストライプサイズを 1MB に設定
  - Pinned メモリへ 35～70MB 程度の粒度で転送
  - DMA 転送を用いると 1.21～2.28倍高速
  - 十分にオーバーラップされる場合は pread と mmap に大きな性能差は見られなかった
- GPU への転送を隠蔽できる計算量を持つアプリケーションが必要
  - 密行列ベクトル積では計算が占める割合は 8% 程度

複数 mSATA SSD と GPU を用いた実験のまとめ

- 最適な I/O 手法の選択により，複数 mSATA SSD から GPU へ PCI-E 上限(3.06 GB/s)のスループット
  - RAID0 を使用し、ストライプサイズを 1MB に設定
  - Pinned メモリへ 35～70MB 程度の粒度で転送
  - DMA 転送を用いると 1.21～2.28倍高速
  - 十分にオーバーラップされる場合は pread と mmap に大きな性能差は見られなかった
- GPU への転送を隠蔽できる計算量を持つアプリケーションが必要
  - 密行列ベクトル積では計算が占める割合は 8% 程度

複数 mSATA SSD から GPU への最適な転送方法を確認

発表の流れ

1. 背景
2. 複数 mSATA SSD を用いた予備評価
   1. プロトタイプマシンの設計
   2. I/O ベンチマークを用いた評価
   3. 既存の不揮発性メモリとの性能比較
3. 複数 mSATA SSD と GPU を用いた予備評価
   1. プロトタイプマシンの設計
   2. ベンチマークアプリケーションの実装
   3. 予備評価
4. 関連研究
5. まとめ

関連研究

- 不揮発性メモリ同士の性能比較*1
  - SATA 接続型と PCI-E 接続型不揮発性メモリの性能比較
- RAID カード上での不揮発性メモリの性能調査*2
  - パラメータ探索による性能調査
- オーバーヘッドの削減による mmap の最適化*3
  - プロセッサ間割り込みの削減による最適化
- GPUからファイルシステムへのI/Oインターフェイス*4
  - GPU メモリ上のバッファキャッシュの最適化

*1: Master et al.: “Performance Analysis of Commodity and Enterprise Class Flash Devices”, PSDW 2010
*2: He et al.: “DASH-IO: an empirical study of flash-based IO for HPC”, TG 2010
*3: Song et al.: “Low-latency memory-mapped I/O for Data-intensive Applications on Fast Storage Devices”, DISCS 2012
*4: Silberstein et al.: “GPUfs: Integrating a File System with GPUs”, ASPLOS 2013

まとめと今後の課題

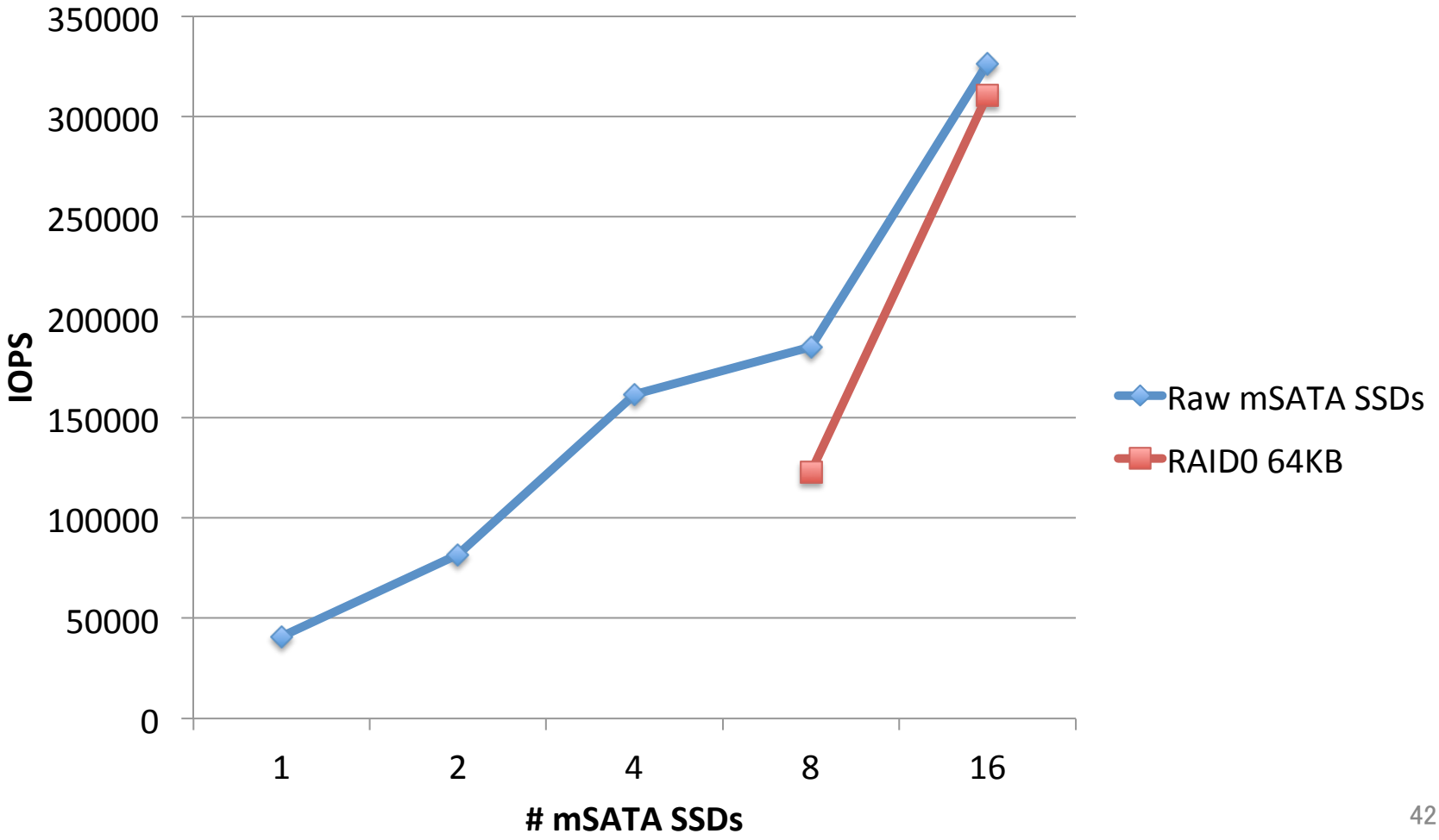
- まとめ
  - 不揮発性メモリから GPU への最適な I/O 手法を把握
  - 16 枚の mSATA SSD を用いたプロトタイプマシンの設計
  - 複数 mSATA SSD の I/O 基本性能の評価
    - 16枚の mSATA SSD で 7.39 GB/s (理論ピークの 92.4%)
    - 8 mSATA SSD で ioDrive2 に対し 3.20～7.60 倍 の Read 性能
  - 複数 mSATA SSD から GPU への I/O 性能の予備評価
    - 8 mSATA SSD から GPU へ 3.06 GB/s のスループット
    - RAID0 (ストライプサイズ 1MB) を組み、35～70MB の粒度で DMA 転送
- 今後の課題
  - ランダム I/O の評価
    - 大規模グラフ探索など
  - 実アプリケーションを用いた評価
    - スペクトラルクラスタリングなど

- Backup

mini SATA SSD (mSATA SSD)

- mSATA SSD
  - 大容量、高性能、低消費電力、低価格
  - 例) CT256M4SSD3:
    - 256 GB
    - Read: 500 MB/s、Write: 260 MB/s
    - 平均アクティブ時消費電力: <200mW
    - アイドル時消費電力: <85mW
    - $260-$300

複数 mSATA SSD の IOPS



Implementation using multi-mSATA without RAID 0