SSDの基礎
HDDとSSDの特性差

株式会社 東芝 セミコンダクター社
メモリ事業部
SSD応用技術部
永井　宏一
多部　光一

JDSF～JAVCOM2010年秋季技術交流セミナー
『SSDの最新情報から映像まで』
主催: JDSF技術交流WG
共催: JAVCOM技術研究委員会

2010/10/27

TOSHIBA
Leading Innovation

2010年秋季JDSF～JAVCOM 技術交流セミナー
「 SSD の最新情報から映像まで 」

SSDの基礎
HDDとSSDの特性差

2010年10月27日
於： 東放学園大学

株式会社 東芝 セミコンダクター社
メモリ事業部
SSD応用技術部

永井 宏一　　多部 光一

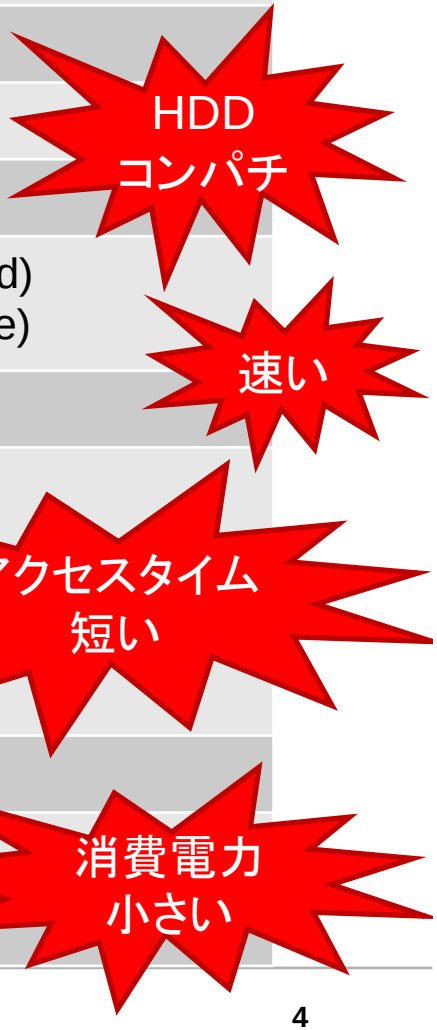
Agenda

1. はじめに　HDDとSSDの仕様比較
2. HDD及びSSDの原理と基本特性
3. SSDの基本的動作の仕組みと、SSD固有の振る舞い
4. おわりに

はじめに
HDDとSSDの仕様比較

PC用HDDとSSD (Solid State Drive)の仕様比較例

	HDD	SSD
容量	640GB	512GB
サイズ	2.5” FF （高さ9.5mm 幅69.85mm 奥行100.0mm）	
Interface 転送レート	SATA (3Gbps)	
セクター長	512Byte	
回転数	7200rpm	N/A
データ転送速度（最大）	約115MiB/s	220MiB/s（Read) 180MiB/s（Write)
平均回転待ち時間	4.16ms	N/A
平均シーク時間	11 msec（Read) 12 msec(Write)	N/A
トラック間シーク時間	1ms	N/A
最大シーク時間	19ms	N/A
電圧	5V	
消費電力（最大）	5.5W	3.4W
質量	110g	58g



PC用HDDとSSD (Solid State Drive)の仕様比較例

	HDD	SSD
容量	640GB	512GB
サイズ	2.5” FF（高さ9.5mm 幅69.85mm 奥行100.0mm）	
Interface 転送レート		SATA (3Gbps)
セクター長		512Byte
回転数	7200rpm	N/A
データ転送速度（最大）	約1Gbps	220MiB/s (Read) 180MiB/s (Write)
平均回転待ち時間	4.17ms	N/A
平均シーク時間	11ms(Read) 13ms(Write)	N/A
トラック間シーク時間	1ms	N/A
最大シーク時間	19ms	N/A
電圧	5V	
消費電力（最大）	5.5W	3.4W
質量	110g	58g

SSDを使用すれば、
どんな用途でも
簡単に性能UP!!？

HDD コンパチ

速い

アクセスタイム 短い

消費電力 小さい

PC用HDDとSSD (Solid State Drive)の仕様比較例

	HDD	SSD
容量	640GB	512GB
サイズ	2.5” FF（高さ9.5mm 幅69.85mm 奥行100.0mm）	
Interface 転送レート	SATA (3Gbps)	
セクター長	512Byte	
回転数	7200rpm	N/A
データ転送速度	約64MB/s	220MB/s(Read)
		te)
トラック間シーク時間	1ms	N/A
最大シーク時間	19ms	N/A
電圧	5V	
消費電力(最大)	5.5W	3.4W
質量	110g	58g

SSDを使用すれば、

というほど、話は簡単ではありません

HDD コンパチ

速い

アクセスタイム 短い

消費電力 小さい

SSDに関する巷の情報

- プチフリって聞くが？
- TRIMがないとSSDは性能が落ち、寿命が短くなる？
- 書換え寿命に達するとすぐ壊れる？
- 使っていくと性能が落ちる？
- デフラグはSSDの寿命を縮めるだけ？

HDD及びSSDの原理と基本特性

HDD及びSSDの原理と基本特性

	HDD	SSD
記録原理	磁気 ダイレクトオーバーライト	電荷 消去してから記録
記録媒体	磁気ディスク	NANDフラッシュメモリー
記録による媒体劣化	無	有
読み出しによる情報劣化	無	有
年月、温度による情報劣化	無視できる	考慮が必要
アクセス場所の決定方法	磁気ディスクの面、半径位置、周方向位置。スピンドルモータとシーク機構が必要	電子スイッチ
LBAと物理記録位置	原則として固定	対応は可変
記録再生単位	セクターサイズ単位（Large Sectorでも4KiB程度）	セクターサイズよりかなり大きい。 消去単位は、さらに大きい

HDDの速度

■速度の要因

- 物理的な因子　：　これが主因
  - LBAと、物理的記録位置が原則として固定
  - 機構的な仕組みでアクセス。LBAが離れれば、アクセス時間が延びる。
  - ディスクは一定速度で回転し、記録密度はディスク全体でほぼ均一。
    したがって、外周ほど転送レートが速く、内周は遅い。
- ファームウェア処理：　副因　（HDDの高密度化に伴い増えてきたが）
  - スペアセクターによる交替処理、キャリブレーション処理など

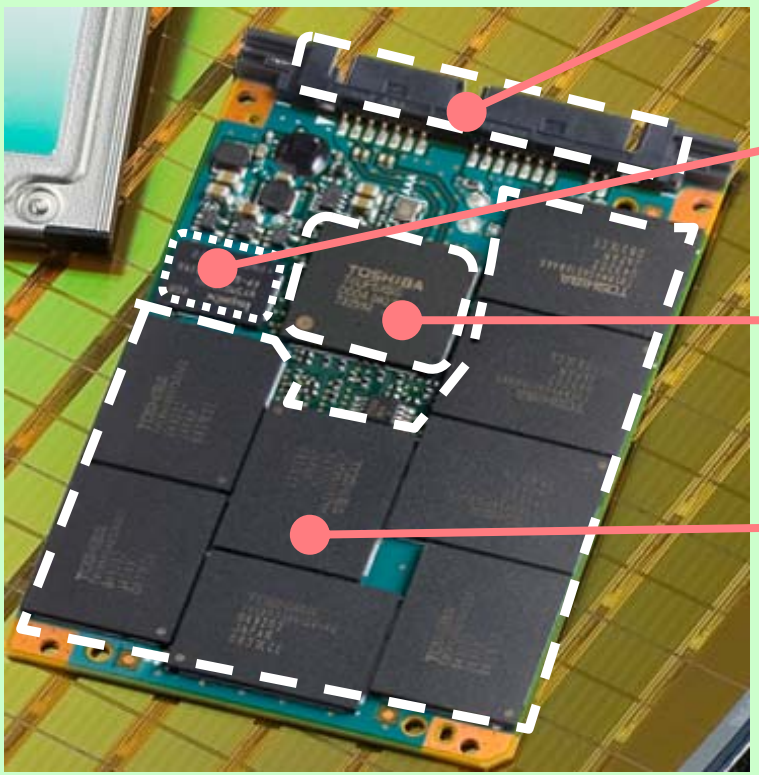
■物理的な要因は、LBAによってコントロール可能

- 性能を上げるためには
  - スピードが必要なデータは外周に記録
  - 同一ファイルはLBAを連続させる
  - 関連するファイルはLBA空間内で近くに配置
- ファイルシステムは、これらのHDDの特性を考慮して設計されている

SSDの基本的動作の仕組みと、SSD固有の振る舞い

SSDの基本的構造

SSDの構造例



●コネクタ
HDDと同様のSATA、micro SATAの他 mSATA等も。

●DRAM
各種管理データの保持、Cache用。コントローラ内部のSRAMで済ませて不要の場合もあり。

●コントローラ
SSDの心臓部となる。このコントローラでの制御により、高速化/書換寿命の長期化/高信頼性化を実現。

●NANDフラッシュメモリ
データを蓄積するNANDフラッシュメモリ。MLC（現在は4値）技術により低価格/大容量化を実現。
（※SLC(2値)NANDが使用される場合もある）
SSD用では1パッケージに8chipまで積層。この例ではSSD1台に64チップまで搭載可能。

●形状

・2.5inch case	（コネクタ：SATA）
・1.8inch case	（コネクタ：Micro SATA）
・1.8inch case less <筐体なし>	（コネクタ：Micro SATA）
• Half Slim <筐体なし>	（コネクタ：SATA）
• mSATA	（コネクタ：mSATA）

など

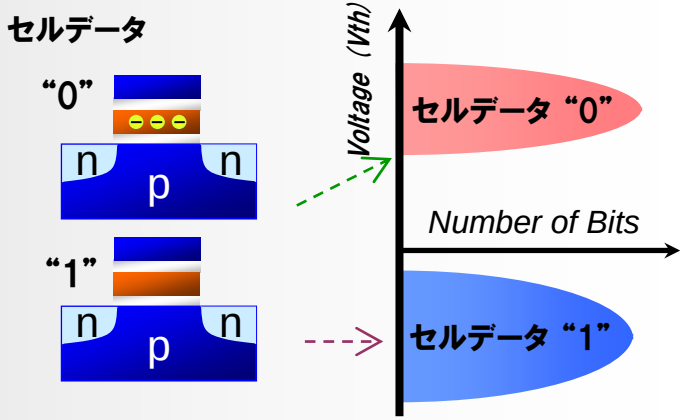
NANDのSLCとMLC

SLC（Single Level Cell）：2値NAND技術

・1セルに1ビット分のデータの格納が可能（1 bit/Cell）

●1セルに格納されているデータ

1 OR 0



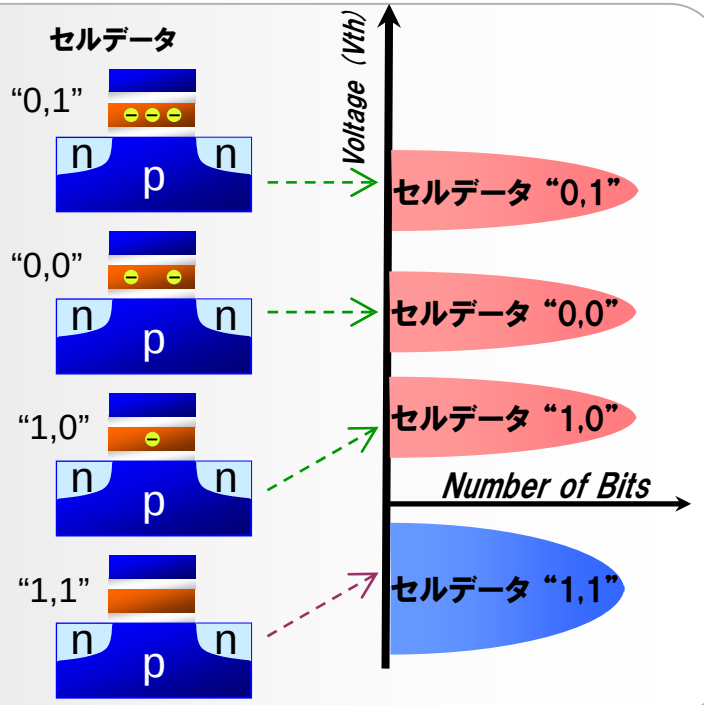
MLC（Multi Level Cell）：多値NAND技術

・1セルに2ビット分のデータの格納が可能(2bit/Cell)
これにより、1セルあたりに格納できるデータ量がSLCの2倍になる
・最近は、1セルあたり、3ビットを格納できるMLCもでてきた。

SLCの2倍のデータ量

●1セルに格納されているデータ

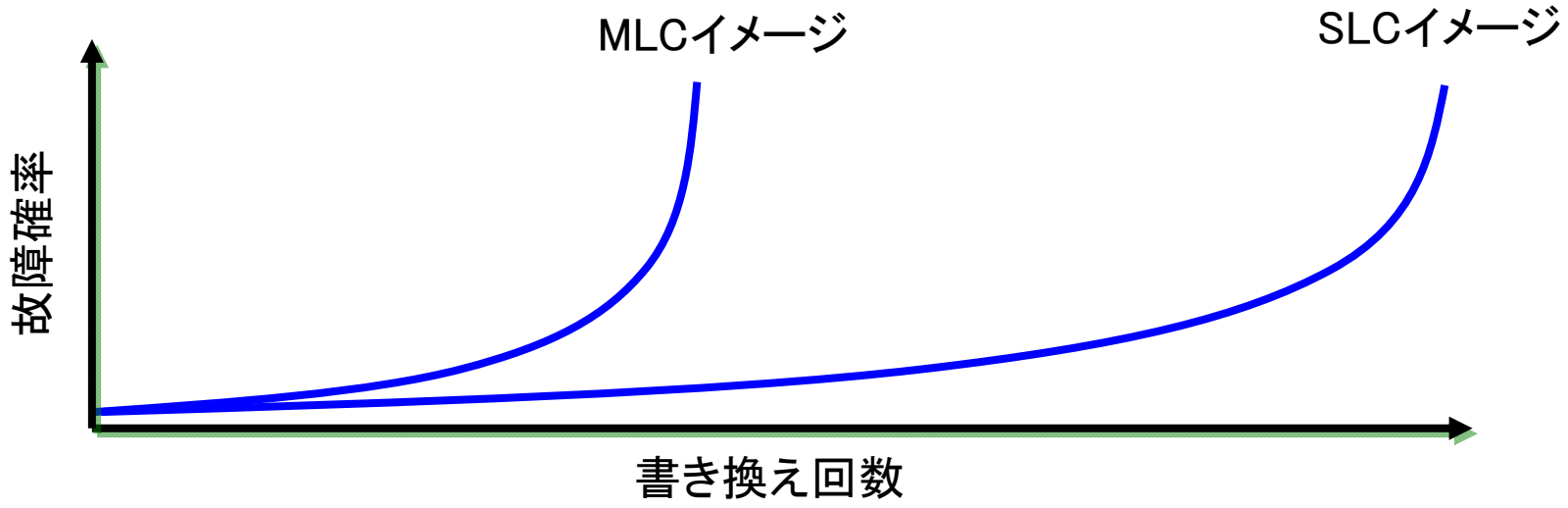
11 OR 10 OR 00 OR 01



NANDの寿命に対する誤解

- MLC NANDは、SLC NANDに比べ 書き換え耐性が低い　Truth
- NANDの書き換え回数が規定値を超えると、NANDはすぐ壊れる　Myth

NANDは書き換え回数が増えるほど、”故障確率”が高くなるデバイスである



NANDの故障モードは幾つかあり、SSDとして予防・救済出来るものもある

NANDとSSDの論理構造

NAND

- 読み書きの単位： ページ
  - ページにはオーバーライトできない.
- 消去の単位：ブロック
  - ブロックにはページ単位で順番に追記（ブロック内ランダム記録不可）

32Gibitチップの例

8KiB/page * 128 page/block * 4096 block

- 物理容量は、これにECCのための冗長、予備ブロックの容量が含まれる。

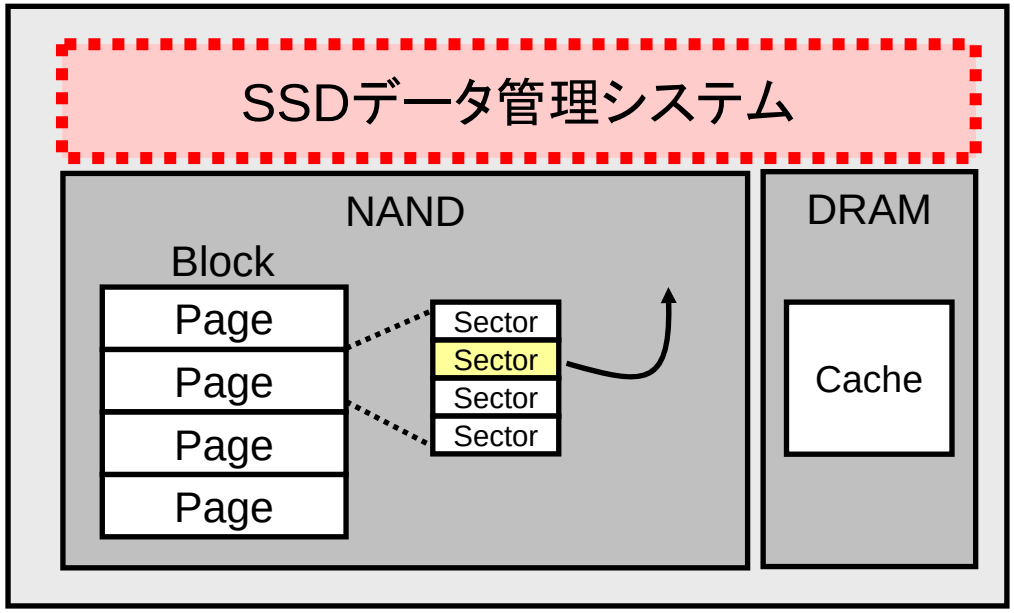
SSD

- SSDは、高速化のため、論理的に複数チップを1つとして扱う場合がある。
  ー＞論理ページ＆ブロックサイズは、さらに拡大
- SSDは、1セクター(512Byte)単位での読書きが必要

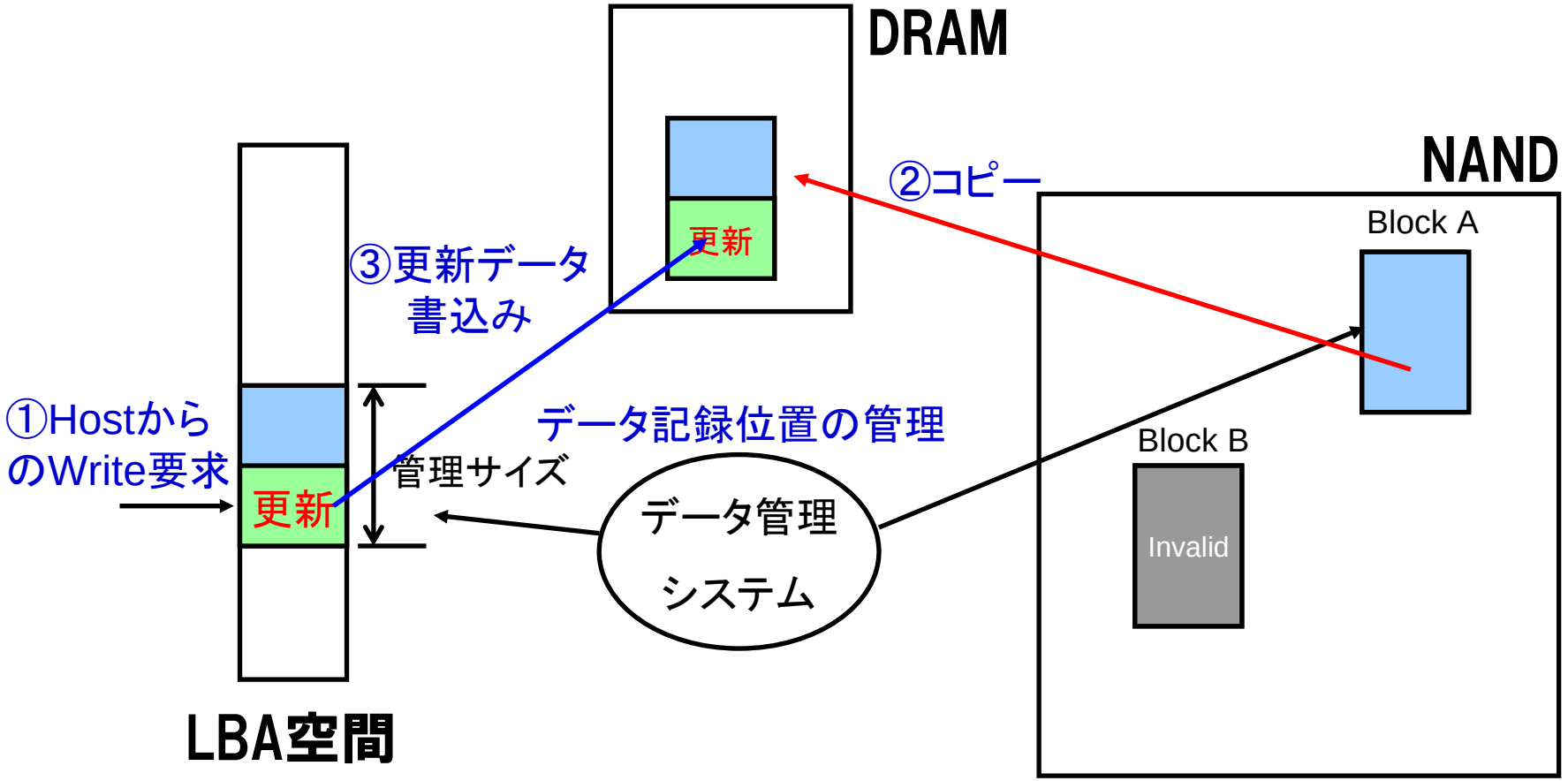
NANDの論理ページ/ブロックサイズと、セクターサイズの差や、
ランダムアクセスの制限を解決する手段、データ管理システムが必要！！

SSDのデータ管理システム

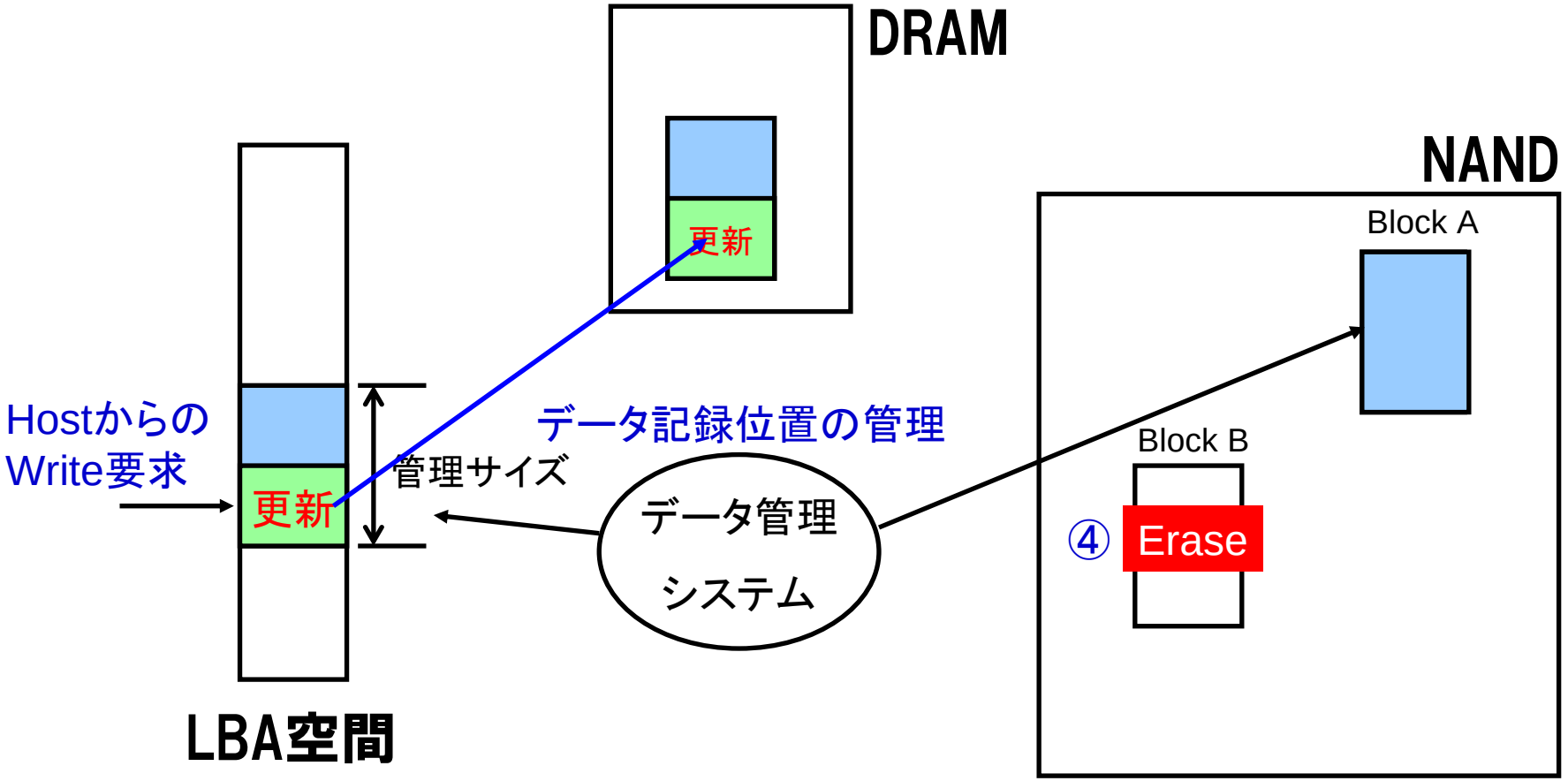
- NANDの制約を吸収
- LBA空間とNAND内データの対応管理
  - セクタ単位でのランダムアクセスを実現
- NAND・DRAMキャッシュ管理
  - 書き込み効率向上
- 消去回数の均一化
  - 寿命を延ばす



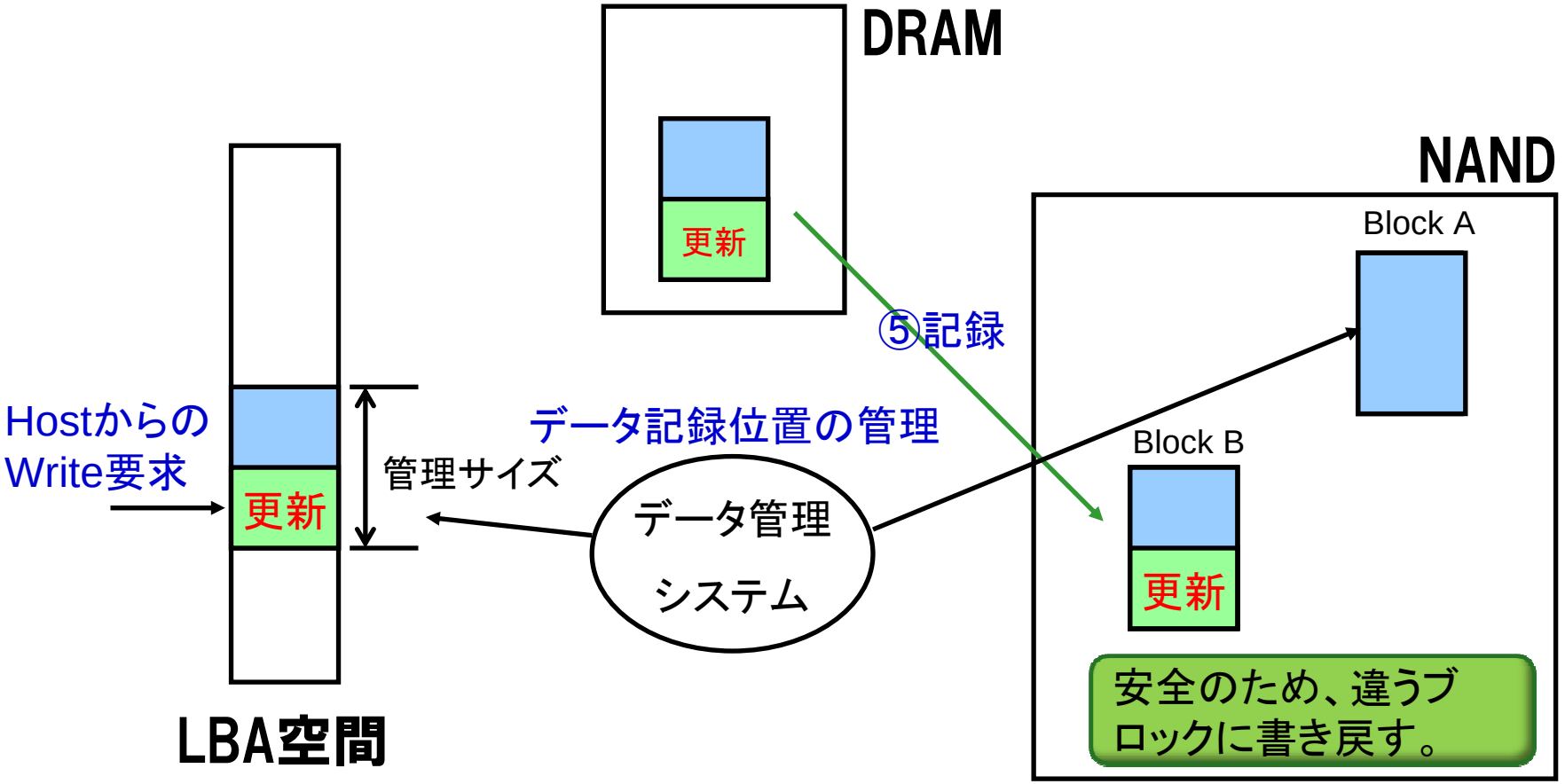
データ更新方法の原理 （Read-Modify-Write）



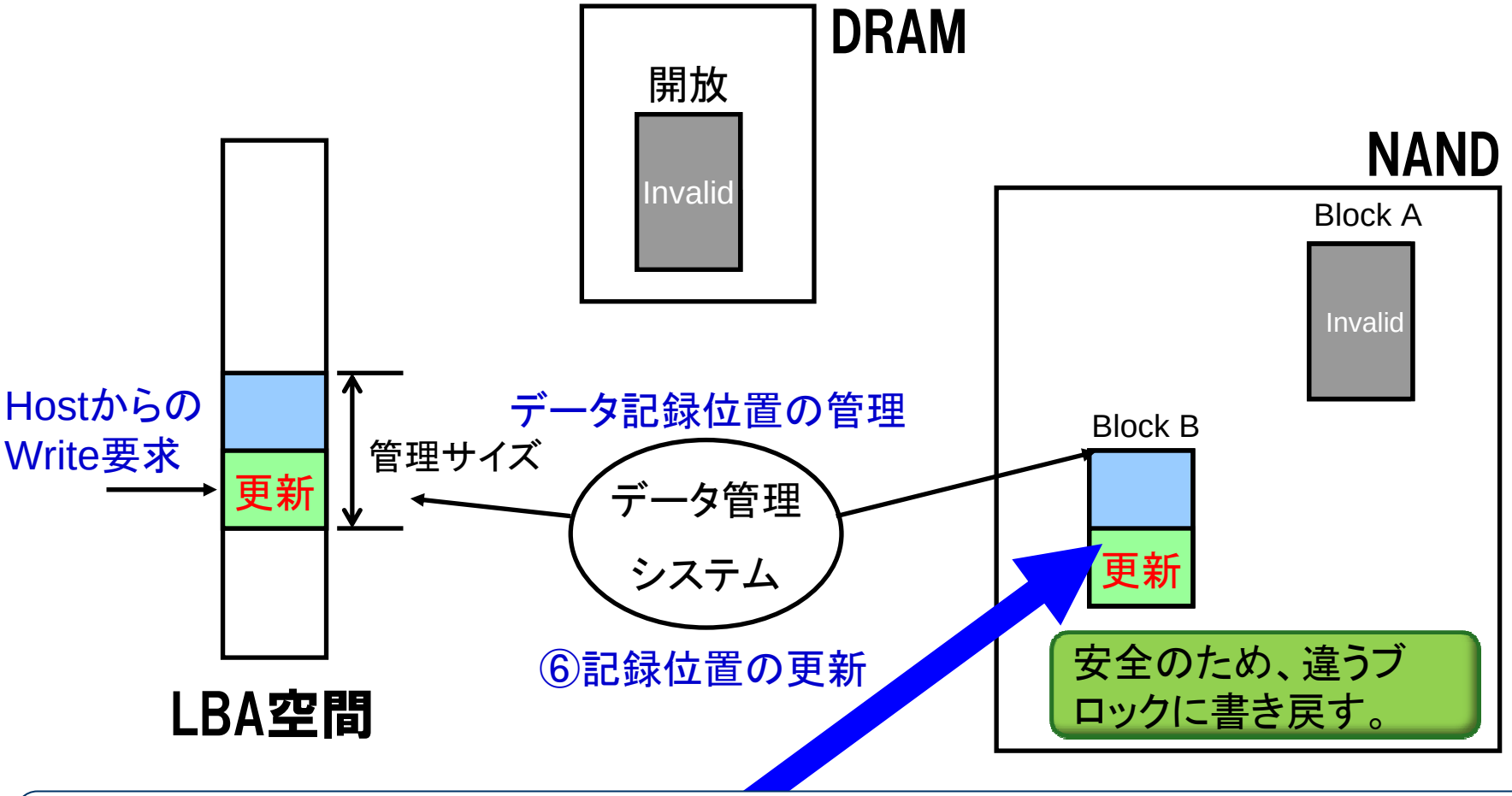
データ更新方法の原理 （Read-Modify-Write）



データ更新方法の原理 （Read-Modify-Write）



データ更新方法の原理 （Read-Modify-Write）

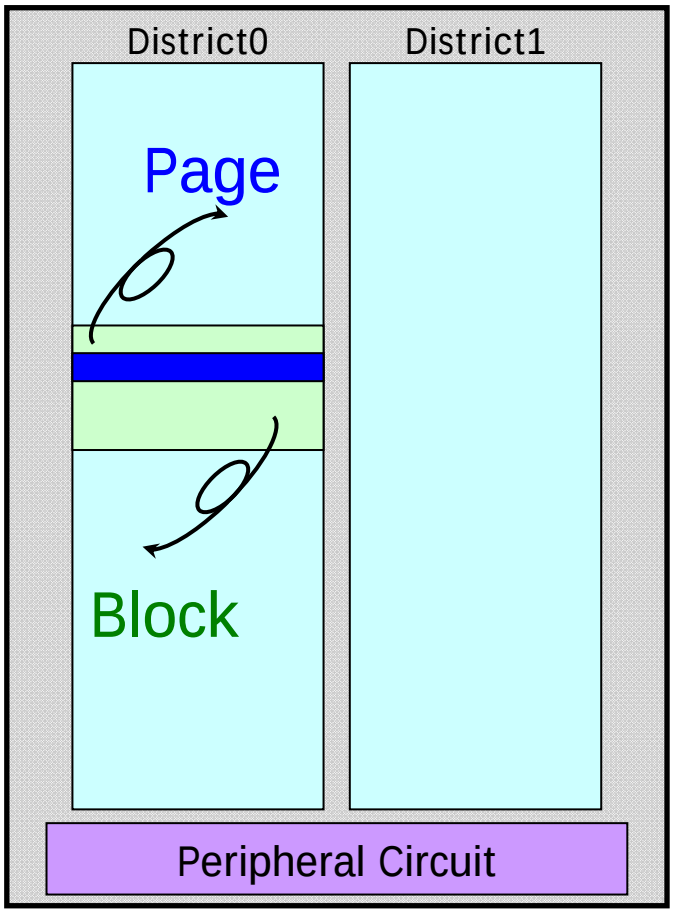


更新したいデータ量と管理サイズが書き込み効率へ影響
例．管理サイズ8KiBに対して4KiBの書き込みを行った場合には書き込み量は2倍

データ管理サイズの選択

■ページ管理
- テーブルが大きくなり、大容量DRAMが必要
  - 例、8KiBページで512GiBを管理
    64Miエントリ*4B=256MiB
- 書き込み効率は良い
  - 例、8KiB管理時に4KiB書き込み
    書き込み効率2倍
- ガーベージコレクション処理が必要
  - 応答悪化要因、記録に追いつかないと、**極端に悪化する場合がある。**
  - 休ませると、処理が進み回復する。

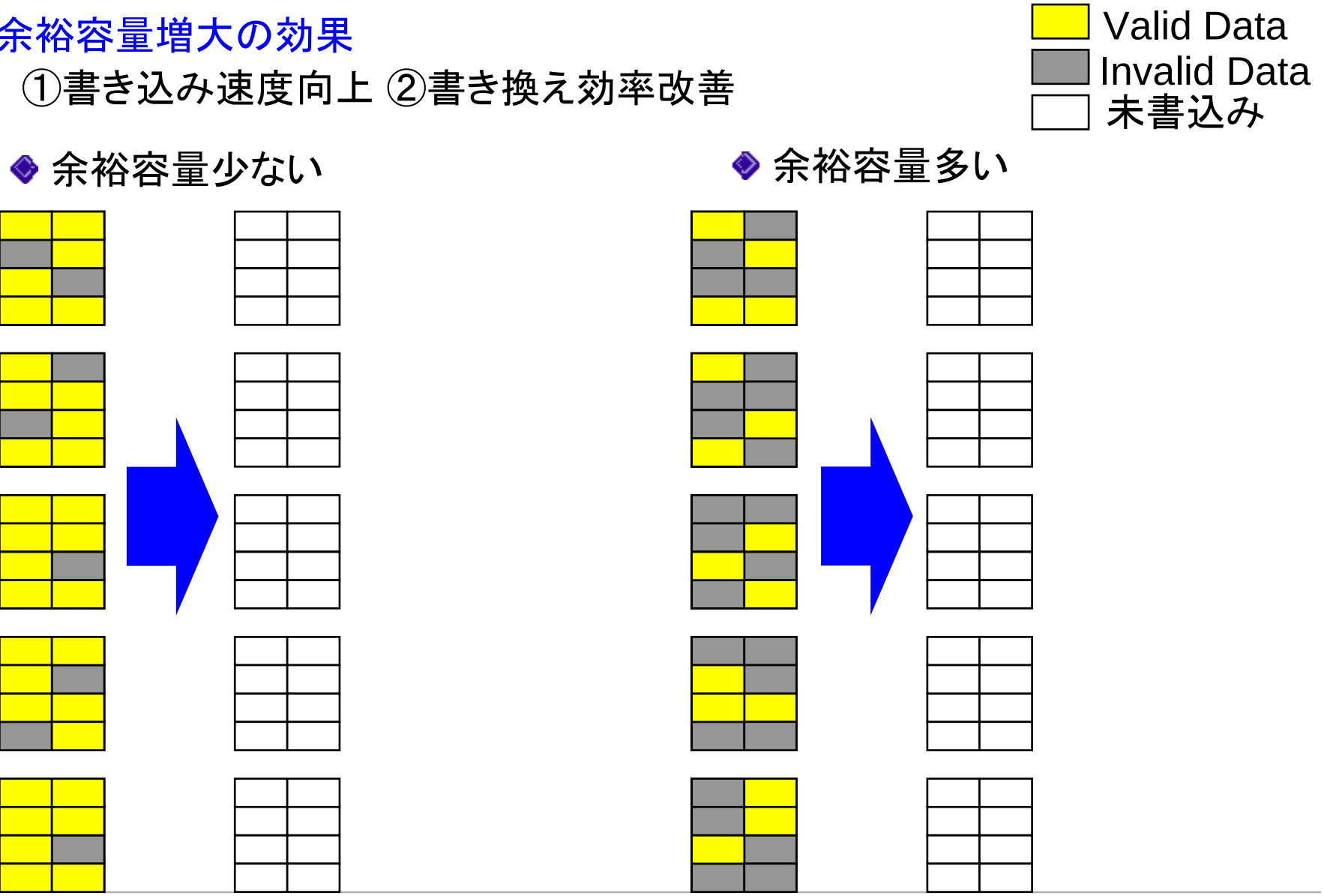
■ブロック管理
- テーブルが小さく済む
  - 例、1MBブロックで512GBを管理
    512Kiエントリ*4B=2MiB
- 書き込み効率が悪い->記録速度低下要因
  - 例、1MiB管理時に4KiB書き込み
    書き込み効率256倍



どうやって管理するかが
各社のノウハウ

ページ管理における余裕容量と性能

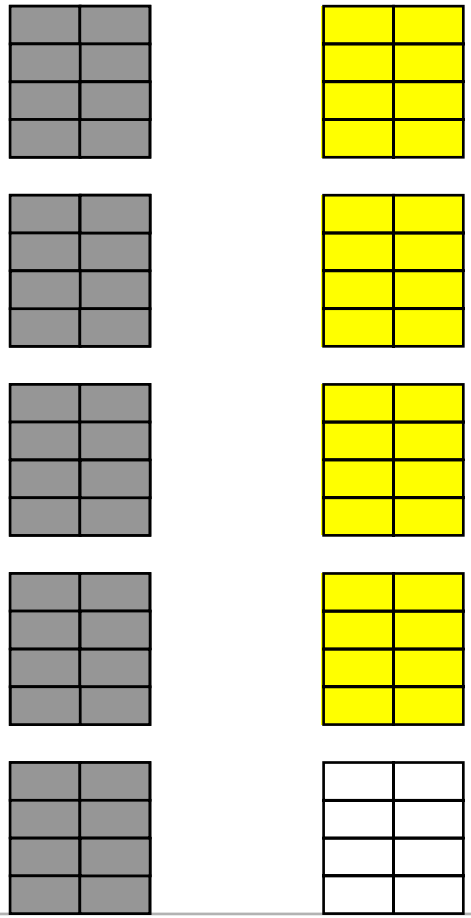
- 余裕容量増大の効果
  ①書き込み速度向上 ②書き換え効率改善



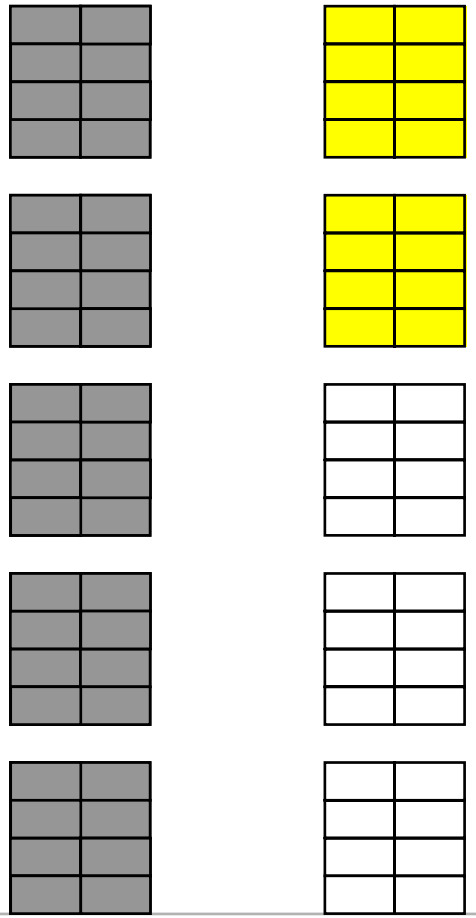
ページ管理における余裕容量と性能

- 余裕容量増大の効果
  ①書き込み速度向上 ②書き換え効率改善

◆ 余裕容量少ない

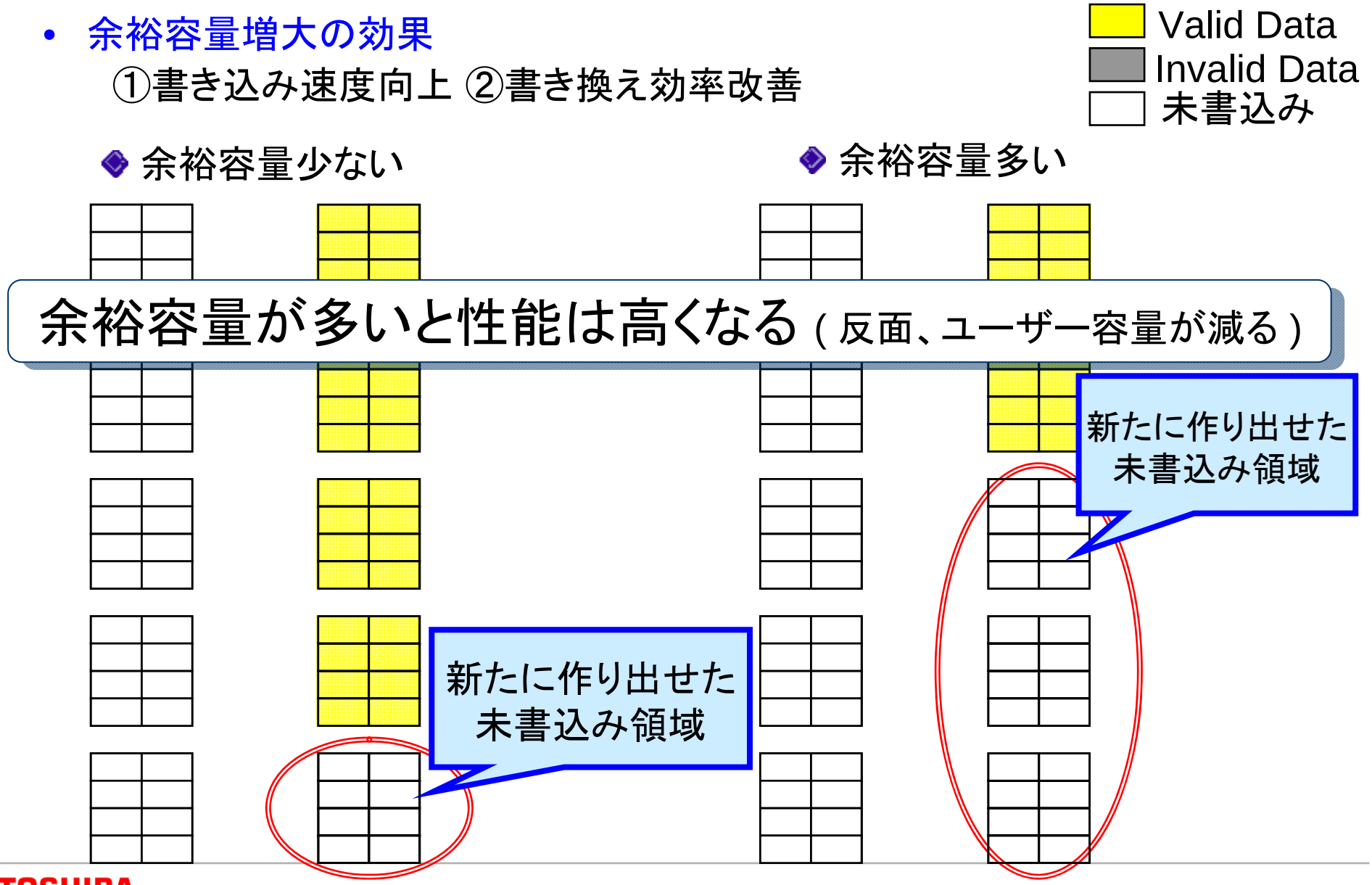


◆ 余裕容量多い



ページ管理における余裕容量と性能

- 余裕容量増大の効果
  ①書き込み速度向上 ②書き換え効率改善

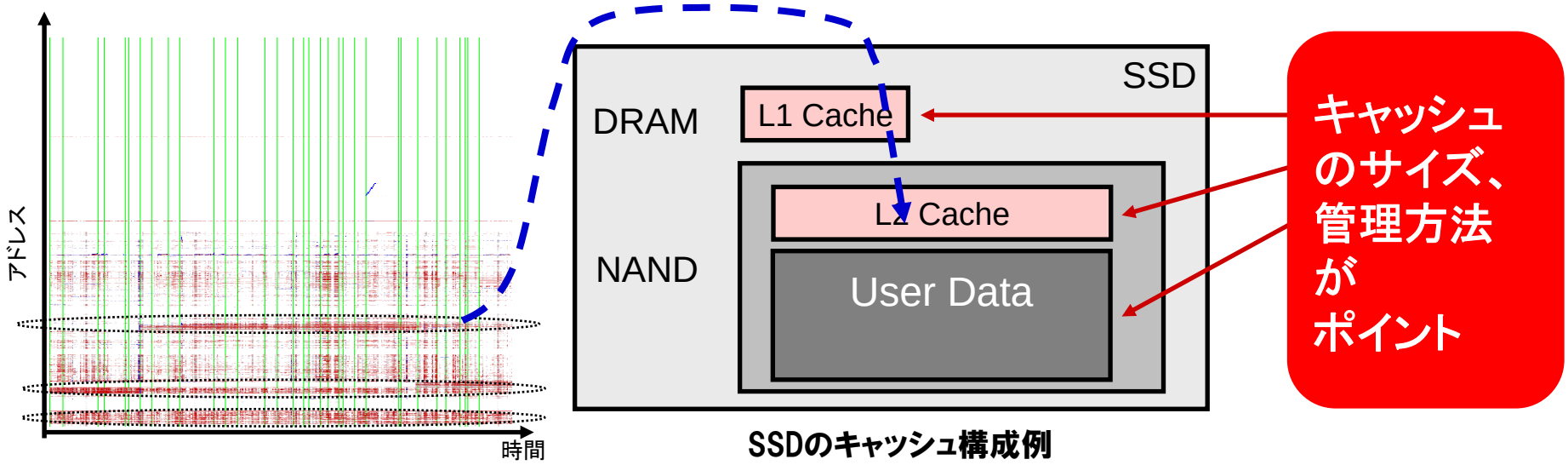


SSDのキャッシュ

- ランダムアクセスにキャッシュが効かないのは当然
  - ただし、フットプリントがキャッシュ容量に収まれば別
- DRAMでは収まりきらない
  - 電源切れると忘れてしまう
  - アクセスには、日を超えて局所性がある

⇒ NAND内キャッシュも使用（多段キャッシュ構成）

書き換え頻度が高いデータをうまく扱えるキャッシュを持っているSSDが速い



SSDのキャッシュ構成例

DRAMキャッシュが大きければ高性能　というわけではない。

TRIMコマンド

- TRIMコマンドとは、SSDのために作られたコマンド
- ドライブに対し、無効となったセクターを知らせる
  - 例えば、ファイルシステム上でファイルが消去された場合、ファイルの実体が記録されていたセクターを、TRIMで無効化になったことを知らせる。
- 受け取ったSSDが、何をするかはSSD依存
  - 場合によっては何もしなくてもよい。（データが消えるとは限らない）
- SSD内のInvalid　Dataを保持するページが増大
  - ページ管理部分にのみ効果あり
  - 余裕容量が増えるのと、同等の効果
  - 無効なデータをコピーしなくて済むので、速度低下が減り、寿命が延びる
- ブロック管理部分にはほとんど効果無
- ページ管理が主体のSSDには効果があり
- 残り容量が少ない場合は、デフラグの方が有効な場合あり。

TRIMコマンドの効果はSSD依存

SSDの故障モード

1. NANDのブロック故障が増加して、動作不能になる

- Read Only Modeに移行する実装もある

2. エラー訂正に失敗して、データが失われる

- ホストにはUNC(Uncorrectable Error)応答
  - NANDの消耗、温度、記録後の経過時間の影響で確率が変動する

3. ATAコマンドも処理できない

- NANDチップ丸ごと故障
- コントローラ故障
- DRAM故障

など

SSDの故障確率

■ SSDの故障確率は、以下の変数による関数

1. Drive全体の容量
2. 累積書込み量
   平均的なClient PCでは、4GiB/day
3. 書き込み効率
   ホストから“1”書いたときの、SSD内部での実書き込み量
   使い方で大きく変動

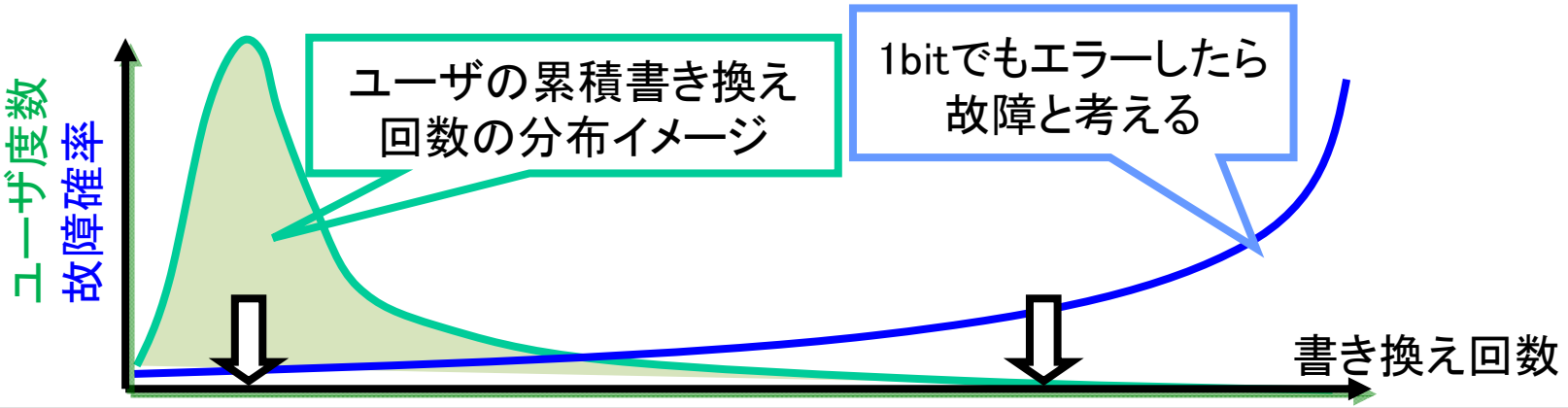
■ NANDの書換え回数を求める

（1日4GiB/day、5年使用，書き込み効率6として）

$$\text{書き換え回数 342 回} = \frac{\text{2 4GiB} * \text{1825日（5年）} \quad \text{3 6（書き込み効率）}}{\text{1 128GiB}}$$

現在のNANDの実力に対して十分すぎる余裕があり

■ SSDの故障確率は、ユーザによって異なってくる



SSDの寿命を延ばす技術(1/2) ウエアレベリング

■消去回数を均一化する技術

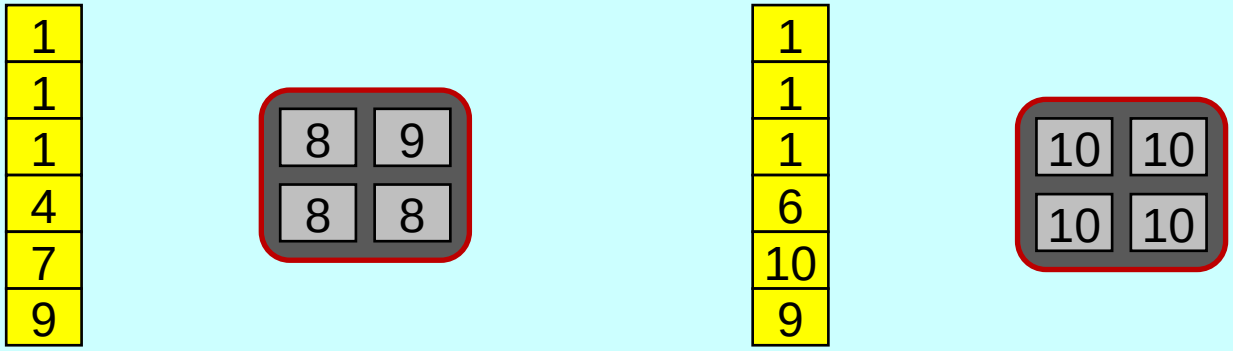
レベル1： 通常動作

- 無効ブロックリストから消去回数の少ないブロックを選んで記録
- 副作用無

レベル2： 有効ブロックリストに消去回数が非常に少ないブロックがある場合

- 書換えられないままの領域がある場合に発生する
- 有効ブロックのデータを無効ブロックにコピーし、元のブロックに記録
- 副作用あり （書込み効率悪化、消去回数の増大、書込み速度低下）

ウエアレベリングの原理



SSDの寿命を延ばす技術(1/2) ウエアレベリング

■消去回数を均一化する技術

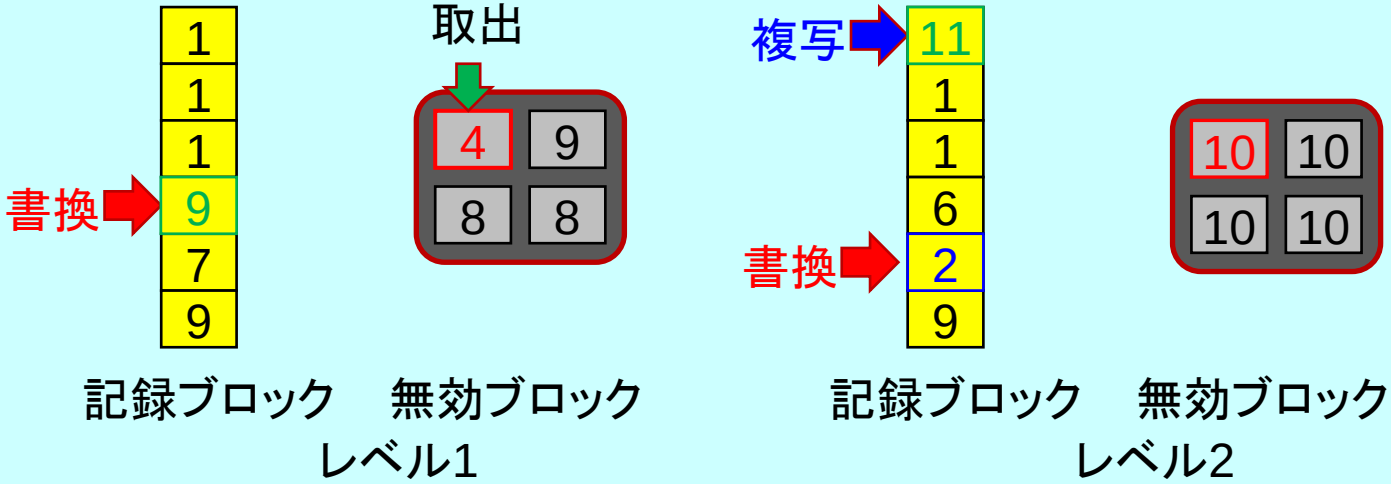
レベル1： 通常動作

- 無効ブロックリストから消去回数の少ないブロックを選んで記録
- 副作用無

レベル2： 有効ブロックリストに消去回数が非常に少ないブロックがある場合

- 書換えられないままの領域がある場合に発生する
- 有効ブロックのデータを無効ブロックにコピーし、元のブロックに記録
- 副作用あり （書込み効率悪化、消去回数の増大、書込み速度低下）

ウエアレベリングの原理



SSDの寿命を延ばす技術(2/2)　ECCとリフレッシュ

■NANDのエラーレート

- 書換え回数が増えるにつれ悪化
- 記録後、時間の経過とともに悪化（高温だとより加速される）
- 同じデータの繰り返し読み出しで悪化
- 隣接データの繰り返し読み出しで悪化
- セル以外の故障に起因する増加

■ECC等によるデータエラー修復

- 同じNANDチップでも、寿命はエラー訂正能力で変わる
- 訂正方法によっては、エラーレートが悪化すると、訂正時間がホストから見えるようになる　－＞応答性能の悪化

■ECCの訂正限界を超える前に、書き直す　－＞　リフレッシュ

- 副作用

  消去回数増加、応答悪化

HDDとSSDの消費電力と電力量

- 最大消費電力は、HDDでは起動時なのに対し、SSDはWrite 時。Staggered　Spin-up　では回避できない。
- SSD内部での裏処理のWriteもあるので、外部からタイミングをコントロールするのは難しい。RAID構築時、電源容量注意。
- SSDは、Read/Write時以外は、消費電力が非常に小さいので、電力量は削減できる。

	HDD(640GB)	SSD(512GB)
Startup	5.5W	N/A
Read/Write	2.1W/2.1W	1.8W/3.4W
Seek	2.3W	N/A
Idle	1100mw（Active Idle)	210mW (PhyRdy)
Idle	800mW(Low Power Idle)	53mW (Slumber)
Standby	160mW	53mW (Slumber)
Sleep	130mW	53mW (Slumber)

使用に伴う長期的性能劣化

使用に伴い

- リフレッシュ頻度増加
- エラー訂正時間増加
- 記録単位の断片化（ファイルシステム側で、空き領域をデフラグすると応答改善に効く場合がある。）

などの劣化要因の反面

- 記録時間減少

という性能向上効果もある

どの要因が表れるかはSSDと使い方次第

偽劣化

HDD用ベンチマークプログラムは、未記録のセクターでも読むものがある。

- SSDでは、未記録のセクターは、読むNANDが無く、読まずに0を返す実装がある。
- このようなプログラムで測定すると、記録領域が増えるにつれ、読み出し速度が劣化するように見える。

おわりに

おわりに

■SSDは、

- 機械的、電気的、論理的インターフェースは、HDDと同じで互換あり
- 中身は、HDDとまったく別物
- Client　PC用SSDは、その用途に使う分には、何も考えなくても不都合無いように設計されている（筈）
- 設計の想定外の用途に使う場合に、期待通り動くかはわからない

■応答遅れの原因

- ガーベージコレクションなどのデータ整理
- リフレッシュ
- ウエアレベリング
- エラー訂正
- その他（エラーリカバリ、管理情報記録）

■応答遅れ低減方法

- TRIMが効くSSDと効かないSSDがあるように、アーキテクチャと実装に強く依存
  - リアルタイム性を重視される場合は、注意が必要