

製品寿命への影響

多値化や微細化は、低価格化と大容量化に大きく貢献しました。しかし製品寿命（書き換え可能回数とデータ保持可能期間）は、大きく短くなってしまいました。なぜでしょうか？

微細化は、トンネル酸化膜の面積が縮小するため、同じ回数の書き換え処理を行っても、痛みの度合いは面積の小さいほうが早く進みます。多値化は、蓄えられている電子量を判別するための「閾値」を増やすということです。同じ酸化膜の通過頻度が増加し、損傷速度が速くなります。

以上の理由から、SSDは長期保存用デバイス、所謂「アーカイブデバイス」にはなり得ません。これは同じNANDフラッシュメモリを使用している、USBメモリやCFカードといった製品にも当てはまります。^{*4}

2.14 微細化の影響

2D NANDにおける微細化の影響は、正確には、主に下記の2点に集約されません。

- 電荷蓄積量の減少による「状態間マージンの縮小」
- セル間距離短縮による「容量結合（セル間干渉）の増大」

微細化によって、1つの記憶素子（セル）が保持できる電子数が減少します。その結果、セルの閾値電圧（ V_{th} ）の状態間差（例：00 \leftrightarrow 01）が小さくなり、状態間マージンが縮小します。このため判定誤り（read error）が発生しやすくなり、更に電子リーク（電子漏出）やデータ保持特性の劣化が顕在化しやすくなります（第20章参照）。

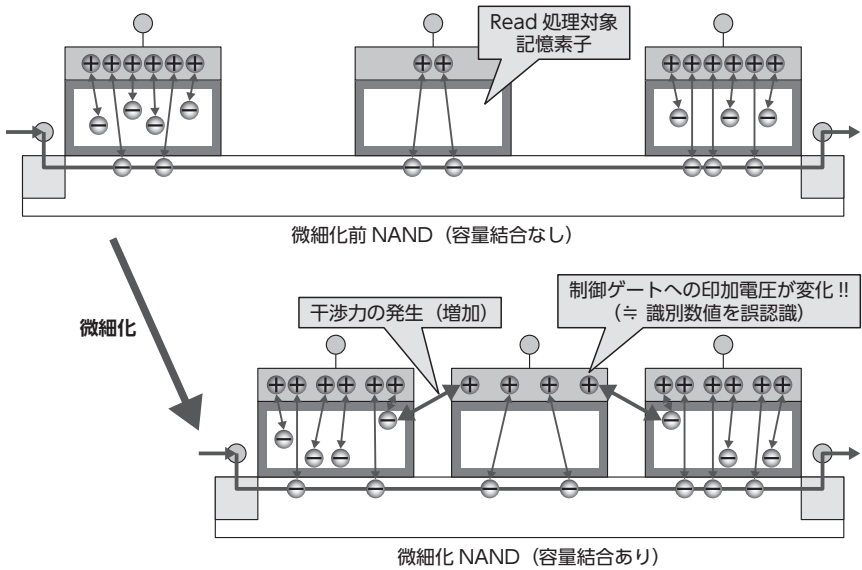
また、セル間の物理的距離が短くなることで、電荷蓄積層間の容量結合（capacitance coupling）が強くなります。容量結合とは、対象セルの「実際の蓄積電荷量」ではなく、隣接セルとの電気的結合によって「見かけ上の電位」が変化してしまう現象です。つまり、隣接セルに蓄積された電子の影響を「擬似的に

*4 昔（セミナー講師を務めるようになった初期の頃）、某国立図書館の方が、セミナーに聴講にこられ終了後に、「本日の内容は大変参考になりました。やはりSSDでデータ保管は駄目ですか？」と聞かれたので、「駄目です」とお答えした記憶があります。

自分のセルの電荷」として感じ取ってしまうことを指します (図2.16)。

但し、現行の電荷トラップ方式による3D NANDでは、容量結合の影響は一旦は大きく緩和されています。そのため当面問題となるのは、多値化と多層化 (薄層化) による電荷蓄積量の減少と考えられています。

図2.16 セル間距離短縮による「容量結合の増大」



Column

技術公開について筆者が感じること

個人的には、搭載技術を明らかにしなくなってきたことは問題だと考えています。

2024年12月に、Micron社「Crucial MX500」の製造終了が発表されていますが、この製品はConsumerモデルとしては珍しく、RAIN (Redundant Array of Independent NAND) 技術^{*1}の搭載を開示していた数少ないモデルの一つです。製造終了のニュースが発表された際に、筆者がこのことをSNS上で何気なく書いたら、思いのほか反響があったので驚いた記憶があります。

邪推だとは思いますが、こういった信頼性のための機能を詳細に説明することは、各メーカーがSSDの販売にとって「マイナス効果がある」と思って、業界全体として避けているのではないかとさえ思っています。蛇足情報ですが、過去に一世を風靡したSandForce社が、2014年に当時の親会社であったLSI社からSeagateに売却されたりして年々影が薄くなっていったのは、LDPCの実装開発の遅れによるものです。

表3.2 主な搭載技術一覧

No.	技術名称例	主な実現方法	搭載可能性	
			Enterprise	Consumer
①	RAID (RAID5/6相当)	NANDチップによるRAID構成	高	低～中
②	Adaptive Read Management	Read処理時のThreshold voltage (閾値電圧)の変更	高	中～高
③	Erase/Program voltage management	eMLC、HET (High Endurance Technology)	高	中～高
④	pSLC	疑似 (pseudo)SLC	Write キャッシュ	同左& DRAM代替
⑤	Error correction	ECC回路等	高	高
⑥	Patrol Read +強制リフレッシュ	同左	高	中～高
⑦	End-to-End Data protection	伝送路でのエラー発生補正	高	低
⑧	Power Loss Protection	瞬停対策	高	低

*1 RAINとは、NANDチップによるRAID5に近い技術のこと。

⑨	Data Retention Management	Write 処理時間からの時間計測とリフレッシュ処理実行	高	低～中
⑩	Read Disturb Management	Read 処理回数の計測とリフレッシュ処理実行	高	低～中
⑪	Write Disturb Management	Write 処理回数の計測とリフレッシュ処理実行	高	低～中
⑫	Over-provisioning	余剰領域	高	低～中

※実際の搭載有無は製品によって異なります。

表 3.3 搭載技術と影響内容 (速度とデータ信頼性)

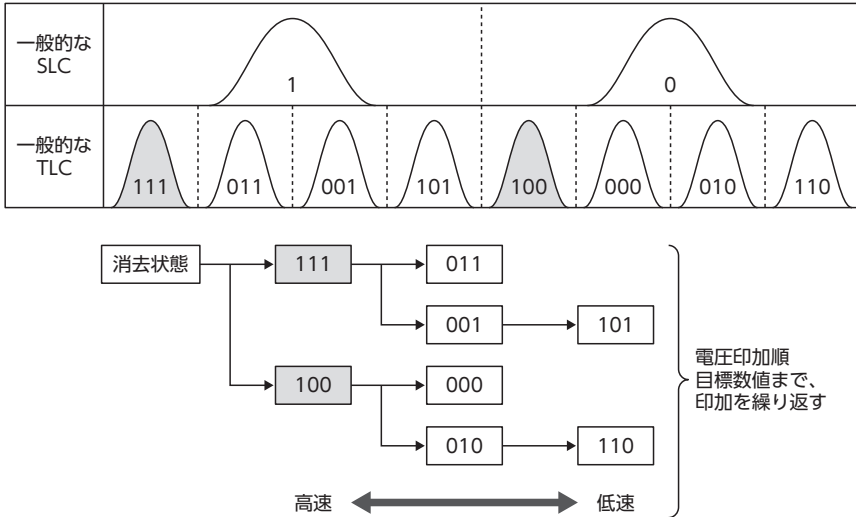
No.	技術名称例	内容	速度		データ	
			平均	Tail	信頼性	保持期間
①	RAID (RAID5/6相当)	NAND チップによる RAID 構成	○	▼	○ (チップ故障)	—
②	Adaptive Read Management	Adaptive Read Management	—	▼	○ (bit エラー抑制)	△
③	Erase/Program voltage management	eMLC、HET (High Endurance Technology)	▼	—	△	×
④	pSLC	疑似 (pseudo) SLC	○	○	?	?
⑤	Error correction	ECC 回路等	▼	▼	○	△
⑥	Patrol Read + 強制リフレッシュ	(Data scrubbing)	▼	▼	○	△
⑦	End-to-End Data protection	伝送路でのエラー発生補正	△	—	○	—
⑧	Power Loss Protection	瞬停対策	—	—	○	—
⑨	Data Retention Management	Write 処理からの時間計測 & リフレッシュ処理	—	?	○	△
⑩	Read Disturb Management	Read 処理回数計測 & リフレッシュ処理	—	?	○	△
⑪	Write Disturb Management	Write 処理回数計測 & リフレッシュ処理	—	?	○	△
⑫	Over-provisioning	余剰領域	○	○	—	—

※○=主目的、△=副次的プラス効果、▼=副次的マイナス効果、×=マイナス効果、?=状況/実装次第、—=大きな影響無し

※「Tail」とは、Tail Latency 発生への影響度を示しています。Tail Latency については第9章で改めて説明します。

TLC以降の多値化NANDを、疑似SLCとして利用する場合の速度は、内部の電圧印加方法と、どの「値」を使うかによって異なります。なお、TLCよりSLCの方が、印加電圧許容範囲が広いため高速です(図3.6)。

図3.6 TLCをpSLCとして利用する場合の例(下段の「ツリー図」は、Writeの順番を表わす)



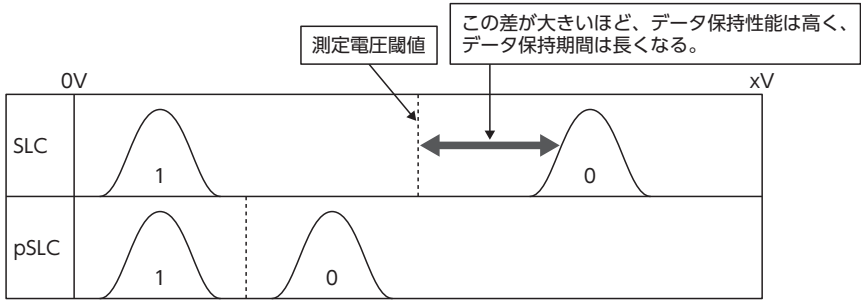
3.8

④ pSLC (データ保持性能)

NANDは、識別閾値電圧と注入電圧(電子量)の差が大きいほどRead速度は向上し、RBER(Raw Bit Error Rate: 訂正前bitエラー率)は下がり、データ保持性能は長くなります。よってWrite速度重視のpSLCは、SLCと比較して、データ保持性能は高くない場合がほとんどです(図3.7)。

ちなみにpSLC技術は、Z-NAND(Samsung)やXL-Flash(東芝)で利用されており、速度重視は理解していますが、データ保持性能への影響度合いは公開情報もなく筆者が不勉強なため理解できていません。

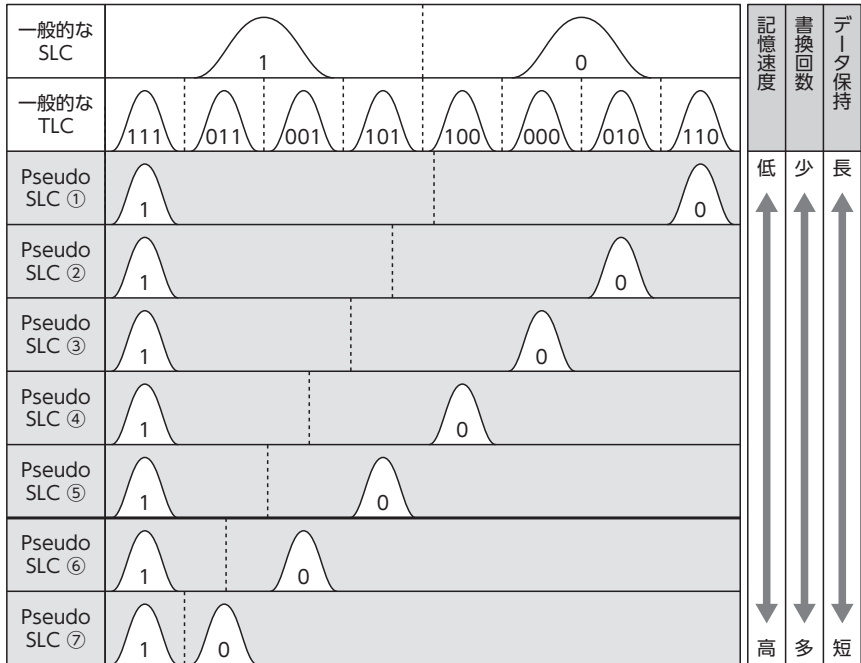
図3.7 電圧差とデータ保持性能



3.9 ④ pSLC (実装の違い)

TLC NANDを疑似SLCとして使う例を先に示しましたが、実装方法で効果に微妙に違いがあります(図3.8)。

図3.8 pSLCの実装別性能への影響



➡ ZNS SSDのデメリット

ZNS規格のSSDは、ユースケースを絞り込み（決め打ちし）、ユーザ開発負荷を下げたものの、「縛り」が増えました。

1つ目は順次書き込みの縛りです。ZNSでは、各アプリケーションは、予め指定されたゾーンしか利用できません。Host-Managed SMR方式のHDD同様、Read処理は任意の順序で処理できますが、Write処理はシーケンシャル（順次）で実行する制限を採用しています。つまりホスト側でバッファリング（ランダムなデータをメモリ上でまとめ、シーケンシャルなデータに変換してからSSDに送る）といった処理を行わない限り、ランダムWrite処理はできません。更にNANDは上書きができないため、結果として追記処理となります。

2つ目はゾーン管理の縛りです。各アプリケーションを特定のゾーン別に保管するために、アプリケーションの修正が必要です。つまり、ユーザ責任での修正は相当残っています（後述のFDP規格が宣伝する「下位互換性」がない）。

➡ ZNSからFDPへ

繰り返しになりますが、耐久性能（TBW）を良化するためには、WAFの良化（低下）が必要で、そのためには、NAND上の「データ配置」が鍵となります。

FDP（Flexible Data Placement）規格は、GoogleとMetaが上記の認識を元に、この課題の解決のために標準化を進めていた技術（実環境での実証を並行実施）となります。元々は、Googleが「SMART FTL Proposal」、Metaが「Direct Placement Mode Proposal」という名前で別々に提案していましたが、2022年に両方の提案を統合する形で、NVMe仕様「Flexible Data Placement (FDP)」として策定・公開されました。

19.9 FDP vs. 他のデータ配置技術

ここからはFDPに注目して解説を進めます。

FDPが過去の技術や規格と最も大きく異なる点は、一般的なSSDとの下位（後方）互換性を持っていることです。つまり同じSSDが、普通のSSDとしても、FDP対応SSDとしても併用可能な点です。一方、Open-Channel SSD、ZNS SSDは、ホスト側の機能追加・開発負荷が高い上に、専用SSDとしてのみ機能します。過去の技術や規格で目指した方向性と同じですが、後方互換性を採用して、導入

のための「ハードル」を下げた点が大きな優位性となっています（表19.3）。

表19.3 データ配置技術別「管理担当」の違い

	一般的なSSD	Open-Channel	ZNS	FDP
ユーザアプリケーション	ホスト	ホスト	ホスト	ホスト
ファイルシステム&ブロック層	ホスト	ホスト	ホスト	ホスト
ブロック・マッピング	ホスト	ホスト	ホスト	ホスト
ファイル・インターフェイス	ホスト	ホスト	ホスト	ホスト
空き領域管理	ホスト	ホスト	ホスト	ホスト
フォーマット&復旧	ホスト	ホスト	ホスト	ホスト
アドレス・マッピング	SSD	ホスト	ホスト	SSD
バッファリング	SSD	ホスト	ホスト	SSD
ガベージコレクション	SSD	ホスト	SSD	SSD
ウェアレベリング	SSD	ホスト	SSD	SSD
ECC (Error Correction)	SSD	SSD	SSD	SSD

19.10 データ配置：FDP vs. ZNS

FDPとZNSをデータ配置の点で比較してみます。図19.7のようにSSD内部を所定の領域に分割しています。分割単位は、ZNS（図19.7の左）は「Zone」、FDP（図19.7の右）は「Reclaim Unit（再生ユニット）」と呼称します。

ZNSが、アプリケーションを修正して、アプリケーション別に別「Zone」への書き込みを強制するのに対して、FDPでは後方互換性実現のために、データの混在を許諾する「Reclaim Unit」も提供し、FDP非対応アプリケーションからでも普通のSSDとして利用可能にしています。

昨今のSSDでは、前節の「誤り訂正符号」の強化の他に、経過時間やRead/Write disturb回数を管理し、規定時間や回数の達した記憶領域に対しては強制的コピー処理（Refresh処理）を実施することで、Read処理失敗の可能性を低く抑えます。

USBメモリはこの機能を実装していません。

24.13 Adaptive Read Management

NANDフラッシュメモリでは、前節の「電子リーク、Read/Write disturb状態管理機能」を搭載していても、NAND品質や摩耗状態により、Write時に印加された電圧（電子量）が予想値より変化することで、Read処理が失敗、データ破損に陥る場合があります。

SSDでは、Adaptive Read Management機能を搭載することで、Read処理の成功確率を向上させています。USBメモリにも同様の機能は搭載されていますが、検査電圧の変更範囲が小さかったり、変更精度が低かったりするために、SSDと比較して再試行時の成功率は低くなります。

また、リムーバブルメディアが前提であるUSBメモリには、長く**電源ON**かつホストからのアクセスがない状態は想定されておらずPatrol Read、Auto Refresh機能等も搭載されていません。

24.14 USBメモリ vs. SSD (搭載技術等比較)

ここまで説明してきたUSBメモリとSSDの搭載技術等を比較した表を作りました（表24.1）。なお、SSDはモデル／製品によって機能は搭載されていても機能レベルは異なります。

表24.1 搭載機能比較

No.	技術名称例	備考	USBメモリ	SSD	参照
1	搭載ウェアレベリング種別	Dynamic、Global (Static含む) 等の違い	Dynamic	Global	本章
2	ガベージコレクション起動タイミング	アイドル時間等での実行有無	Write処理時のみ	様々なタイミングで実行	本章
3	マッピング情報サイズ	マッピング情報のサイズ	極小	様々	本章
4	RAID	NANDチップによるRAID構成	非搭載	RAID0～RAID5/6	—
5	Read処理失敗時の再試行	Read処理時の閾値電圧の変更(再試行パラメータの扱い)	最低限	静的～動的	第21章参照
6	ビットエラー訂正(ECC回路等)	ECC回路等	最低限	中～高	—
7	Patrol Read処理	同左	非搭載	非搭載／搭載	第21章参照
8	End-to-End Data protection	伝送路でのエラー発生補正	非搭載	非搭載／搭載	第22章参照
9	瞬停対策	回路保護～コンデンサー搭載	回路保護のみ	様々	本章
10	データ保持状態管理	Write処理からの経過計測とリフレッシュ処理実行	非搭載	低～高	—
11	Read/Write disturb管理	Read/Write処理回数計測とリフレッシュ処理実行	非搭載	低～高	—
12	NAND品質	多値化とは別(センター品 vs. 非センター品)	低～中	低～高	本章

Command TimeoutメッセージがSSDから返信された場合は次のようになります。

- **非RAID環境：**

OSはエラーメッセージをログに保存した上で、コマンドを再発行します（コマンド再発行回数は設定次第）。なお、起動デバイスにおいて再発行後も状況が改善しない場合は、悪名高い「ブルースクリーン（Windowsの場合）」が発生します。

- **RAID環境：**

上記と基本的な挙動は同じですが、メッセージが頻繁に到着した場合はSSDに障害が発生していると考え、切り離します。

あるRAIDコントローラの設定閾値

下記の閾値は、あるIAサーバー製品マニュアルに掲載されていたものです。RAIDコントローラ製品や、サーバーモデルで異なります。

- **HDDの場合**

- 5分間にエラーイベントが110個発生した際に、切り離し処理を実行。

- **SSDの場合**

- 5分間にエラーイベントが20個発生した際に、切り離し処理を実行。

- **主な対象エラーイベント**

- Command Timeout（応答が一定時間以内でない）
- Unexpected Sense（予期しない挙動）

26.4 SSD障害事例

ユーザ障害事例

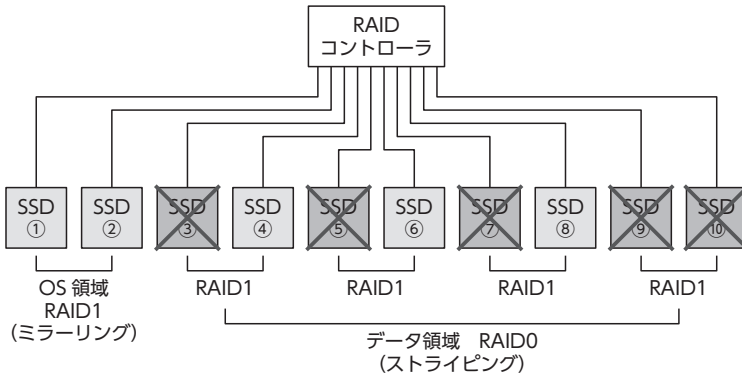
本節では、ユーザ事例を用いてSSDの障害を解説します。一言でCommand Timeoutと言っても発現のタイミングがシステム環境によって変化する事例として興味深いものかと思えます。

或る日突然、「データ領域が全く見えなくなる」現象が発生しました。ユーザ領域は、障害耐性と高速性能への必要性から「RAID10」を採用しています（図26.1）。

詳細な障害原因調査を行ったところ、以下のような障害履歴が判明します。

- 障害が発生する以前に、数か月の間に次々と4台のSSD (図26.1の③、⑤、⑦、⑨) が切り離されていた (ユーザ側の監視漏れにより、SSD障害を放置した状態で運用を継続)。
- 5台目のSSD (図26.1の⑩) が切り離されたタイミングで、データ領域が見えなくなった。

図26.1 システム構成と障害発生SSD



障害原因を分析した結果、切り離されたSSDは、全て「Secure Erase」処理により、継続利用が可能であることを確認できました (つまり経年低下であってSSD故障ではない)。しかし、いずれのSSDも長期間稼働の中で、経年低下が発生していました。このことからCommand Timeoutによる「切り離し」によって障害が発生していたと判明しました。

ところで、なぜ「RAID1」構成の「片側SSD」が先行して切り離されたのでしょうか？ 更に、「片側SSD (RAID1構成)」が切り離された後、残りのSSDが切り離されるまでにタイムラグ (時間的ずれ) が生じたのはなぜでしょうか？ これらの点については次節の参考資料の中で考察します。