
ソリッドステートストレージのメリット

このデルテクニカルホワイトペーパーは、各種のソリッドステートストレージとそれぞれの強みについて説明します。

Jeff Armstrong
Gary Kotzur
Rahul Deshmukh



目次

はじめに.....	3
PCIe SSS	3
SAS SSD/SATA SSD と、PCIe SSS との相違点	3
PCIe SSS の標準化.....	4
フォームファクタ.....	4
ソリューションの選択	4
Dell Express Flash PCIe SSD の強み.....	5
ハイパフォーマンス	5
耐久性.....	5
柔軟性.....	5
PCIe SSD の Q&A.....	5
PCIe SSD からブートできますか？	5
PCIe SSD は RAID に対応しますか？	6
PCIe SSD を使ったストレージの階層化はどのような用途に向いていますか？	6
なぜソリッドステートドライブは磨滅してしまうのですか？	6
まとめ.....	7

本書は、情報提供のみを目的に執筆されており、誤字脱字や技術上の誤りには責任を負いません。

本書の内容は執筆時現在のものであり、明示的、暗示的を問わず、いかなる内容も保証いたしません。

© 2012 Dell Inc. ©2011 デル株式会社 All rights reserved. (著作権所有)

デルとその関連会社は、誤字、脱字、誤植や、図、写真の誤りや不備について一切の責任を負いません。Dell、DELL のロゴマーク、PowerEdge は、米国 Dell Inc. の商標です。Intel、インテル、Xeon は、アメリカ合衆国およびその他の国におけるインテルコーポレーションおよび子会社の登録商標または商標です。Microsoft、Windows、Windows Server は、米国やその他の国々における Microsoft Corporation の登録商標または商標です。本書では、マークや名前を届け出た実在のもの、もしくは、その製品のいずれかを参照するため、その他の商標、商号を使用している可能性があります。デルは、その他のマークや名称について、商標上の利権に対する要求に一切に応じません。

はじめに

従来型ストレージ技術とそのインターコネクト技術は、長年にわたって進化を遂げてきました。しかし、機械装置であるハードディスクドライブ (HDD) は動作速度に限界があり、ディスク回転速度は 15,000 RPM で頭打ちとなっています。このためデータベースや I/O 集中型のアプリケーションでは、バックエンドストレージがボトルネックとなる傾向があります。

こうしたサーバとストレージ間の性能差を縮める技術として有望視されているのが、ソリッドステートストレージ (SSS) です。ソリッドステートストレージとは不揮発性記録媒体の一種で、回転式の磁気媒体を使わず、集積回路を使って情報を保存します。現在では、様々な形態 (フォームファクタ) の製品が利用可能となっており、たとえば、次のような製品があります。

- **SSD** : Solid-State Drive (ソリッドステートドライブ)
- **SSC** : Solid-State Card (ソリッドステートカード)
- **SSM** : Solid-State Module (ソリッドステートモジュール)

提供されているデバイスインタフェースには、従来どおりのストレージインタフェース (SATA、SAS、FC 等) も含まれます。最近では、既存の PCIe スロットに直接差し込むだけでシステムに接続できるカード式の SSS も出荷されており、この場合、HBA の調停は必要ありません。さらに、従来の外部ストレージネットワーク (FC、SAS、iSCSI 等) を通すことで、サーバでもソリッドステートストレージデバイスが利用できるようになりました。

PCIe SSS

PCI Express ソリッドステートストレージ (PCIe SSS) とは、サーバの PCIe バスに直接接続するタイプのストレージデバイスです。最近の PCI SSS デバイスは、標準の PCIe カードと同じフォームファクタとなっており、通常は、AHCI 対応インタフェースを実装しています。基盤となるストレージ媒体には、「NAND 型フラッシュ」と呼ばれるソリッドステートメモリデバイスが使われます。この NAND フラッシュは、省電力でコスト効果に優れた不揮発性メモリです。

現在使用されている NAND には、シングルレベルセル (SLC) とマルチレベルセル (MLC) の二種類があります。1 セルあたり 1 ビットを含む SLC は、書き込み耐用回数に優れ、高い読み書き性能を必要とする、書き込み集中型のアプリケーションに向いています。一方、1 セルあたり通常 2 ビットを含む MLC は、書き込み耐用回数をさほど重視せず、大容量を優先し、コストも決定要因となり得るアプリケーションに向いています。

SAS SSD/SATA SSD と、PCIe SSS との相違点

SAS/SATA SSD は、ハードディスクドライブ (HDD) と同じホストインタフェースを使用するという点で HDD と似ており、それゆえ、システムからも同じ OS ドライバスタックを通してアクセスされます。SAS/SATA ソリッドステートストレージデバイスそのものは、磁気媒体が回転する通常の HDD よりはるかに高性能ですが、このドライバスタックにつきもののレイテンシやボトルネックのせいで性能が制限されかねません。一方 PCIe SSS は、PCIe バスを通してプロセッサに直接接続できるため、より高い性能を達成します。

PCIe SSS デバイスは、ストレージキャッシュとしても超高速なストレージ階層 (ティア) としても利用可能です。ストレージキャッシュとして導入する場合は、ページファイルのように利用できます。また、ストレージ階層として導入する場合は、頻繁にアクセスされるデータ (ホットデータ) の保存に活用できます。プロセッサに近接し、直接接続できる SSS のアクセス方式はレイテンシが極めて低くなるため、このような用途に理想的です。今後登場する新世代の PCIe は、現行製品よりバンド幅が広がるため、実質的なアクセスレイテンシはさらに低くなるものと期待されています。

PCIe SSS の標準化

フォームファクタ

現在、PCIe SSS デバイスのフォームファクタは PCIe カード形となっており、ハーフハイト/ハーフレングス (ロープロファイル、フルハイト/フルレングス、フルハイト/フルレングスで標準化されています。しかし、これらのカードは、特に高可用性 (HA) システムの場合、保守性に優れているとは言えません。そこで、この難点を解決すべく発足されたのが、「SSD Form Factor Working Group¹」(SSD フォームファクタワーキンググループ) です。同グループの創設理念は、SSS のスムーズな統合、相互運用性、スケーラビリティ、保守のしやすさ、可用性を向上することで、PCIe SSD を普及させることにあります。そのため、「デバイスのフォームファクタ」「コネクタ」「ホットプラグ動作」という三分野を中心に取り組んできました。

これまでのところ、2つの仕様が策定されており、1つはコネクタとそのピン配列の仕様、もう1つは、2.5 インチ HDD に倣ったデバイスのフォームファクタ仕様です。この仕様は、フォームファクタのみならず、ユーザが求める高い信頼性、保守性、可用性を満たすのに必要なデバイス動作についても定義しています。また、コネクタについては、SAS および SATA との下位互換をサポートするドライブコネクタと、最大 4 レーンの PCIe をサポートするための仕様を定義しています。今や同グループは、「SSD FF WG」という通称で広く知られるようになり、創設メンバーには Dell、EMC、富士通、IBM、インテル[®]、その他、業界を代表する企業陣が名を連ねます。現在、これらの企業は、策定した仕様の批准に向けて取り組んでおり、いずれは SFF 標準化団体の認可済み業界標準とすることを目指しています。

ソリューションの選択

「SAS/SAT SSD」と「PCIe SSS」というどちらのタイプのソリューションも、ユーザのニーズ/要件に応じて独自の強みがあります。高性能な PCIe SSS に基づく PCIe SSD は、最高の IOPS 性能が達成可能なうえ、低レイテンシのため、超高速を必要とするアプリケーションに向いており、キャッシュやブライマリストレージとして活用できます。ユーザの期待に応えるべく生まれたドライブ型フォームファクタの PCIe SSD 製品は、保守が容易なため、高可用性/基幹系アプリケーションでも採用できますし、サーバやストレージプラットフォームでドライブを拡張することもできます。一方、お使いのソリューションに厳しいレイテンシ要件がなければ、SAS SSD/SATA SSD を従来のホストストレージ内や外付ストレージインフラに導入することができ、この場合、データ保護対策としてソフトウェア RAID かハードウェア RAID を構成することも可能です。

すべての PCIe SSS が同じアーキテクチャを採用しているわけではないため、実際の購入時やシステム構成のプランニング時は、いくつもの要因を鑑みて決定しなければなりません。NAND フラッシュのアーキテクチャには、大きく分けて SLC と MLC という二種類があります。NAND フラッシュを使ったデバイスは、残念ながら、SLC であろうと MLC であろうと、どちらも消耗品であり、「消去→書き込み」サイクル数には限度があります。そこで各メーカーは NAND フラッシュの寿命を延ばす (= 書き込み回数を増やす) ために、様々なメカニズムを開発してきました。実際これからは、1つのデバイスに複数のメカニズムが組み込まれるようになります。これらのアルゴリズムはすべて、「フラッシュマネジメント」と呼ばれる共通分野の技術です。フラッシュマネジメントの実装法は多々ありますが、ハードウェア側の仕組みを極力シンプルにしたければ、OS 側に実装する方法、つまり、デバイスドライバスタックの一部としてフラッシュマネジメントを処理させる方法があります。一方、もっと洗練されたデバイスハードウェアを実装する場合は、デバイス自身にフラッシュマネジメント処理を実行させます。本来、フラッシュマネジメントに必要な機能群を実行するには、相当量のプロセッサリソースが必要になることから、決して負荷の軽いタスクではありません。そこで、これらの処理をプロセッサからデバイス側にオフロードすれば、ソリューションのスケーラビリティとパフォーマンスが飛躍的に向上します。

¹ SSD Form Factor Working Group の詳細は、<http://www.ssdformfactor.org/aboutus.html> をご覧ください。

Dell Express Flash PCIe SSD の強み

Dell PowerEdge™ Express Flash PCIe SSD とは、PCIe SSC ソリューションを凌ぐ独自の強みを備えた高性能ストレージであり、本製品に搭載されたエンタープライズクラスの SLC NAND フラッシュは、低レイテンシ、高性能 IOPS (入出力処理件数/秒)、エンタープライズストレージの高い信頼性・保守性を必要とするソリューションに最適です。本デバイスは、PCIe Gen2 に準拠し、負荷の高いエンタープライズ環境 (エンタープライズブレードおよびラックサーバ、ビデオオンデマンドサービス、Web アクセラレータ、仮想化アプライアンスなど) のプライマリストレージデバイス、または、ストレージキャッシュとして構成できます。以降のセクションでは、Express Flash PCIe SSD の特長について詳しく説明します。

ハイパフォーマンス

PowerEdge Express Flash PCIe SSD は極めて高い IOPS 性能を発揮し、読み取り/書き込み処理では、最大 1.8/1.2GBps のシーケンシャルスループットを達成する設計となっています。さらに本製品なら、サーバのプロセッサとメモリに NAND フラッシュマネジメントのオーバーヘッドがかかることはありません。フラッシュマネジメントはデバイスのコントローラ側が行い、このコントローラは、PCIe SSD 内に格納された高性能 ASIC 上のファームウェアを実行します。このようなアーキテクチャにしたことで、サーバ側のリソースを「借りる」ことなく、優れた IOPS とスループット性能が達成できます。

耐久性

PowerEdge Express Flash PCIe SSD は、ハードウェア以外にも、数々の性能強化が施されています。たとえば、特別設計された NAND マネジメントテクノロジーの採用により、製品ライフ全般にわたる堅牢なウェアレベリング機能とデータ保護機能を提供します。また、NAND SSD は消去&プログラム (書き込み) サイクル数が有限であることから、Dell は、SSD への総書き込みバイト数に基づく最大データ量で PowerEdge Express Flash PCIe SSD を保証します。SSD はこれらの回数をモニタリングすると同時に、Dell の管理アプリケーションソフトウェアは、保証の上限に達したことをユーザに通知することができます。

柔軟性

Express Flash PCIe SSD テクノロジーは多彩な用途にお使いいただけます。さらに、サーバのシステムバスやカーネルレベルで統合されるため、新しいフラッシュストレージキャッシュ/ティアが形成でき、ひいては、読み取りが集中するブロックの二次データベースバッファキャッシュとして活用することも、また、低レイテンシ&高速 I/O を必要とするデータベースオブジェクトの格納先として活用することも可能です。1 台のシステムに複数の Express Flash PCIe SSD を導入すれば、キャッシュとプライマリストレージ容量の両方に活用可能となり、これらの処理を従来の HDD と併用して行うこともできます。

PCIe SSD の Q&A

このセクションでは、PCIe SSD テクノロジーと Dell Express Flash ドライブについて想定される質疑応答例を紹介します。

PCIe SSD からブートできますか？

初期リリース版の Dell PCIe SSD は、ブート対応となっていません。しかし、将来は SSD からの起動をサポートする予定です。

PCIe SSD は RAID に対応しますか？

標準的なハードウェア RAID は、SAS、SATA、FC、iSCSI などの標準ストレージ周辺インタフェースに提供される機能であり、サーバ内のストレージ、または、サーバに接続されたストレージで使用されます。しかし、PCIe SSD は PCIe バス上に搭載されるため、これらのハードウェアコントローラは、PCIe SSD に RAID を提供することができません。ただし、PCIe SSD から OS の RAID 機能を利用することは可能で、次のように様々なオペレーティングシステムが RAID をサポートします。

- Hyper-V 仮想化対応の Microsoft Windows Server® 2008 R2 以降 (プロセッサ：x64/EM64T)
- Red Hat® Enterprise Linux® 6.1 64 ビット以降 (プロセッサ：x64/EM64T)
- SUSE® Linux Enterprise Server 11 SP2 64 ビット以降 (プロセッサ：x64/EM64T)

PCIe SSD を使ったストレージの階層化はどのような用途に向いていますか？

2.5 インチの Dell Express Flash PCIe SSD は、多くのアプリケーションでその強みを発揮します。下記はその一例です。

- オンライントランザクション処理 (OLTP)
- オンライン分析処理 (OLAP)
- コラボレーション (Microsoft SharePoint®)
- 仮想化 (VMware®, Hyper-V)

2.5 インチの Dell Express Flash PCIe SSD は、サーバのシステムバスやカーネルレベルとの統合により、新しいフラッシュメモリ層が形成されるため、高速 I/O と低レイテンシを必要とするデータベースオブジェクトの格納先として、また、二次データベースバッファキャッシュとして活用可能です。

OLTP と OLAP 環境では、飛躍的な IOPS 性能の向上、サポートユーザ数の増加、クエリー応答時間の短縮が望めます。さらに、2.5 インチの Express Flash PCIe SSD は非常に柔軟に導入できるため、データベースアプリケーションの性能強化に活用でき、たとえば、Express Flash ドライブ内にデータベース全体をそっくりそのまま収めても良いですし、データベースのうち頻繁にアクセスされるコンポーネントのみを格納しても構いません。

なぜソリッドステートドライブは磨滅してしまうのですか？

SSD の書き込み回数制限について説明するには、まず、フラッシュの仕組み、フラッシュタイプによる違い、機能強化のために採用されているテクノロジーについて説明する必要があります。

フラッシュメモリデバイスは、読み取り/書き込みをページ単位で実行します。読み取り処理は比較的単純で、ホストからアドレス付きのリード (Read) コマンドが発行されたら、それに対応するデータを応答すれば完了です。しかし書き込みは、ページを消去してからでないと実行できません。したがって、ホストからライト (Write) コマンドが発行されると、まず、フラッシュの消去サイクルが呼び出され、次に、フラッシュへの書き込みが発生します。この消去/書き込みサイクルがセルを摩耗させ、その結果、書き込み回数に制限が生まれるのです。

NAND フラッシュデバイスには、シングルレベルセル (SLC) とマルチレベルセル (MLC) という種類があります。このうち SLC は、セル内に 1 ビットの情報しか保存できないため、電圧レベルも「0」か「1」を表す二段階しかありません。この NAND 実装法は最もシンプルなので耐久性も高まり、約 100,000 サイクルをサポートします。しかし、今後登場する新世代フラッシュは小型化が進むため耐久性が落ち、既に 50,000 回という製品も見受けられるようになりました。MLC は、1 つのセルに 2 ビットの情報を格納する方式が一般的で、「00」「01」「10」「11」を表すために 4 レベルの電圧が必要です。セルの摩耗そのものは SLC と MLC で大きな違いはありませんが、より多数の電圧レベルを感知しなければならない分、MLC の耐久性は大きく落ち込み、通常 10,000 サイクル前後です。しかし、新世代の製品では、3,000～5,000 サイクルまで低下しているものもあります。

ソリッドステートストレージのメリット

大容量を達成するために、1つのフラッシュ SSD には多数のフラッシュダイが搭載されています (ダイ = シリコンウェハーの断片)。また、SSD の書き込み回数を増やすために、どちらの NAND タイプにも適用できる各種の技術が使用されています。ホストの書き込み処理はどの場所にも起こり得ますが、書き込みが一個所に集中してしまうと「ホットスポット」が生じ、この場所だけ摩耗が進んでしまいます。このホットスポットを回避するために使われるのが「ウェアレベリング」技術です。この技術は、その名 (ウェアレベリング = 摩耗を均す) のとおり、書き込みアクセスを SSD 容量全体にわたってほぼ均等に分散させます。また、「Write Amplification」は、ホストが一回書き込みを実行すると、実際にはフラッシュ内部で何回の書き込み操作が発生するのかわかる指標 (比率) です。たとえば、ホストが書き込みを一回実行したとき、フラッシュ内部では実際に 2 回の書き込み処理が行われると、Write Amplification は「2」となります。Write Amplification の削減策としては、「オーバープロビジョニング」と呼ばれる手法が使われており、これで「ガベージコレクション」の効率性が改善されるため、結果的にフラッシュ内部の書き込み処理も減らせます。最後に MLC で良く利用される改善策の 1 つとして、書き込みサイクル中に低電圧レベルを使用し、セルの摩耗を緩和させることで、書き込み耐用回数を向上させるという手法があります。

SSD の寿命を正確に判断することは困難ですが、予測するためのガイドラインはあります。SSD は、JEDEC が開発した「TBW (terabytes written、書き込みテラバイト数)」と呼ばれる判断基準を使用します。実際の書き込み回数は、ワークロード特性 (例：ランダム/シーケンシャル、ブロックサイズ、書き込みアクティビティなど) によって変わってきますが、この TBW を使えば試算することは可能です (繰り返しますが、実際の寿命は変わってきます)。期待される耐用年数を試算するには、TBW を、ドライブ書き込み時に予想される平均 BW 数で割ってください。負荷が特別高いアプリケーションは例外ですが、通常の SSD は 3 年以上の利用が期待できます。

まとめ

Dell PowerEdge Express Flash のような PCIe ソリッドステートデバイスを採用すると、飛躍的な性能向上が望め、レイテンシに敏感なワークロードに多大なメリットをもたらします。Express Flash は標準ベースのサーバストレージソリューションとして HDD ストレージの補完、ストレージキャッシュ、超高速ストレージ階層の追加、その他に幅広く利用でき、IOPS とスループット性能の大幅な強化策となります。Dell PowerEdge Express Flash PCIe SSD は 2.5 インチというフォームファクタに加え、ホットプラグもサポートするため、保守が極めて容易になり、可用性のレベルが一段と高まります。