Western Digital.

ホワイトペーパー 2019年10月

エンタープライズSSDに関する主な考慮事項

目次

はじめに

エンタープライズアプリケーション向けソリッドステートドライブ（SSD）の評価は非常に難しい場合があります。フォームファクターを間違えると、利用するサーバーに合わないこともあります。間違ったインターフェイスタイプを選択すると、最大パフォーマンスを無駄に制限し、アプリケーションが低速になる可能性があります。誤った耐久性評価を選ぶと、数か月後にSSDを交換することになるかもしれません。こうした基本的な要素を選んだ後で、その他の一連の仕様がSSDの実装の成功を左右する場合があります。残念ながらメーカーのデータシートには、さまざまなテスト手法から得られたデータやユーザーを惑わす膨大なデータポイントが含まれており、適切な意思決定の妨げになっている場合がよくあります。

このホワイトペーパーを読めば、より的確な情報に基づき用途に応じて適切なSSDを選択できるようになります。ここでは物理的なフォームファクター、論理インターフェイス、フラッシュテクノロジー、さまざまな書き込み耐久性の差別化要因などに基づいて、SSDを系統的に調査します。これらの各要因の理解を深めることで、エンタープライズSSDを選択する難しさが軽減し、より適切な運用が可能になります。

概要：エンタープライズSSDを選択する際の主な8つの考慮事項

フォームファクター
主なオプションに2.5インチドライブ、M.2、アドインカード、EDSFFがある

インターフェイス
主なオプションにSAS、SATA、NVMeがある

耐久性
高耐久性、中耐久性、読み取り集中型

エラー処理とデータ保護
信頼性を向上させる組み込みテクノロジー

NAND
SLC、MLC、TLC、QLC
2Dまたは3Dの製造を使用

パフォーマンス
マクロおよびミクロレベルでのIOPS、スループット、待ち時間、QoS

電力
設定可能な電力モードで大規模な導入を最適化

監視と管理
大規模な導入を簡単に維持および管理

フォームファクター

これが重要な理由：フォームファクターにより、SSDがどこに収納されるのか、サーバーの電源を切らずに交換できるかどうか、筐体に何台のSSDを設置できるのかが決まります。さまざまなフォームファクターが用意されており、それぞれに、ニーズに応じて変化する特定の強みと弱みがあります。

SSDには可動部品がないため、ハードディスクドライブの同等なフォームファクターに比べ、よりユーザーの状況に合った物理フォームファクターで利用できます。同じSSDであっても、異なるフォームファクターで同様のパフォーマンスを発揮するさまざまなバージョンが用意されており、お使いのインフラストラクチャに最適なSSDを選択できます。

2.5インチドライブ

最も一般的なフォームファクターは2.5インチドライブで、SFF（Small Form Factor）またはU.2としても知られています。このフォームファクターでSSDの幅と長さは決まりますが、複数の高さがあります。ノートパソコンやコンシューマー向けのSSDの高さは、多くの場合7mmです。エンタープライズSSDの高さは7mm、9mm、15mmです。サーバーを設置する状況では、サーバーの大部分でこれらの高さのいずれかがサポートされているので、高さはそれほど重要にはなりません。縦方向の寸法つまりZ高が高いほど、フラッシュチップとコントローラーの容積は大きくなり、パフォーマンス、冷却能力が向上し、容量が増加する可能性があります。2.5インチフォームファクターでは、SATA、SAS、さらにNVMe™インターフェイスのテクノロジー（x4または2×2、次のセクションで説明）をサポートできます。

表1に示す2.5インチフォームファクターは、多くのサーバーおよびJBODフロントパネルに対応し、SSDのホットスワップが可能なため、サーバーの電源を切る必要はありません。1Uサーバーでは、フロントパネルから最大10台のドライブにアクセスでき、ほとんどの2Uサーバーで24個のスロットを使用できます。

アドインカード（AIC）

表1に示すもう1つの一般的なフォームファクターは、HH-HL（ハーフハイト、ハーフレングス）とも呼ばれるアドインカード（AIC）のフォームファクターです。サーバー筐体内のPCIeスロットに接続されるボードで構成されています。このフォームファクターで使用できるのは、PCIeを使用してネイティブに通信するNVMe SSDのみです。サーバー筐体内に設置するため、AICフォームファクターのSSDはホットスワップには対応しません。ただし、通信パスが広く（x8またはx16）、PCIeスロットの電力供給能力が一般的に高いため、同じSSDの2.5インチフォームファクターバージョンよりも広い帯域幅と高い電力プロファイルを備えています。

M.2

データセンター環境で広く利用されているもう1つのSSDフォームファクターにM.2があります。これはマザーボードに直接接続される長くて薄いベアカードのフォームファクターであり、通常は通信にNVMeまたはSATAを使用します。その長さは可変であり、片面または両面にコンポーネントがあるものもあります。サイズ識別子M2-22XXの「XX」部分は、ミリメートルでの長さを表し、一般的なサイズは42、60、80、110mmです。表1にはM2-2280 M.2 SSDのフォームファクターも含まれています。接続が難しい場合も多いため、これらのSSDはホットスワップには対応しません。NVMeバージョンはx2またはx4 PCIeレーンを使用し、SATAベースのM.2 SSDは標準SATA-III信号を使用します。サイズが小さく、取り付け困難な場所にある場合が多いため、M.2ドライブのパフォーマンスを持続するには温度管理が大きな問題になる場合があります。こうした小型のドライブは、放熱のための表面積が小さい場合が多く、サーマルスロットリングやパフォーマンスの長期的な安定性の低下を引き起こします。M.2ドライブの一般的なユースケースにはブートデバイスがあります。この場合、主にホストがデバイスから読み取りを実行します。

EDSFF

SSDフォームファクターリストにエンタープライズおよびデータセンター向け小型フォームファクター（EDSFF）が新たに追加されました。EDSFFの3つの主なバリアントには E1.L、E1.S、E3があります。E1.Sフォームファクターは M.2に比べ若干長さと幅が大きく、1Uのコンピューティングに最適化されたサーバーの設計を対象にしています。E1.Lは1Uストレージサーバー向けに最適化されており、小さいサーバーフットプリントで高密度化を可能にし、クラウドインフラストラクチャでサーバーラックストレージあたりのペタバイト数の増加に対応します。EDSFF 3インチバージョンのE3には、4つのサブバリアント、E3 Short-Thick、E3 Short-Thin、E3 Long-Thick、E3 Long-Thinがあります。E3 FFは2Uサーバーおよびストレージ設計向けに作られています。EDSFFではすべてのフォームファクターで同じコネクター規格（SFF-TA-1002）を使用し、PCIe NVMeプロトコルを使用します。

EDSFFの主な利点：

- データセンターの平方フィートあたりの容量と密度がU.2またはM.2よりも高い
- 汎用システムの将来の拡張性に向け、共通コネクターが用意されている
- サーマルスロットリングがない優れたエアフローと放熱機能がある
- システムの温度とパフォーマンスのバランスを最適化するさまざまな電力レベルをサポートしている

詳細については、*http://www.snia.org/sff/specifications*をご覧ください。


フォームファクター：	2.5インチドライブ（U.2）	アドインカード	M.2	EDSFF（E1.L）
外寸	70×100mm 高さ7～15mm	65×170mm （HH-HL）	22×30～110mm	38.4mmx318.75mmx 9.5mm（25W）/38.4mmx 318.75mmx18mm（40W）
標準電力	11–20W	最大25W	< 8W	< 8W
ホットスワップ対応	対応	なし	なし	対応
フロント処理可能	対応	おそらく	おそらく	対応
標準ドライブロード	最大24台	4～6台（CPUのPCIeレーンにより異なる）	1～2台（CPUのPCIeレーンにより異なる）	最大32台のSSD

表1：エンタープライズSSDはさまざまなファームファクターで利用できます。

インターフェイスオプション

これが重要な理由：インターフェイスはSSDとCPU間の電気的および論理的な信号です。SSDの最大帯域幅、最小待ち時間、拡張性、ホットスワップ機能を定義します。

インターフェイスはSSDがホストとの通信に使用する論理プロトコルです。現在、SSDには3つの基本的なインターフェイスオプションとして

SATA（シリアルATA）、SAS（シリアル接続SCSI）、NVMe（PCIe）があります。通常、2.5インチフォームファクターでのみ使用可能なエンタープライズSSDのSATAおよびSASインターフェイスとは異なり、NVMeは2.5インチ、アドインカード、M.2、EDSFFのフォームファクターで使用できます。各インターフェイスは、具体的な対象者を念頭に置いて開発されました。SATAはコスト重視のホームユーザー向けに、SASは高可用性アクセスをサポートするためにマルチパスなどの機能を必要とするエンタープライズユーザー向けに、NVMeは待ち時間が短く、最も広い帯域幅を重視するパフォーマンスアプリケーション向けに開発されました。SATAとSASはSSDとHDDの両方をサポートできますが、NVMeは通常、SSDのみのプロトコルです。

SATAインターフェイス

一般的にSATAは、最も安価、低い拡張性、最も低い可用性、最も低いパフォーマンスのSSD向けインターフェイスです。SATAの最新世代であるSATA-IIIは、最大約600 MB/秒の転送速度を備え、回転式メディア向けに最適化されたレガシープロトコルのために待ち時間が長くなります。インターフェイス上で高可用性を実現できないため、リ

ンクまたはコントローラーの障害に耐える必要があるユーザーは、非常に低いパフォーマンスのアプリケーションレベルのレプリケーションやその他の戦略を使わざるを得ません。また、ほとんどのサーバーは6台未満のSATAデバイスをサポートする能力がありますが、SATAは通常、高レベルの拡張をサポートしていません。ただし、その低コスト性とマザーボードとチップセットによる完全サポートを考慮すると、SATAはブートデバイスや、レプリケーションでデータの可用性を確保するために必要なアプリケーションロジックをすでに実装しているスケールアウトNoSQLデータベースなどに非常に有効です。RAIDの形式でのデータ保護は、ほとんどの主要なオペレーティングシステムによりソフトウェアレベルで実装可能です。

SASインターフェイス

SASはデュアルポート、エキスパンダー機能、より高いデータレートなど、非常に堅牢なエンタープライズ機能セットを提供します。最先端のSAS12Gインターフェイスは、2つの各リンクで1ギガバイト/秒以上をサポートできます。これらのリンクは、さまざまなケーブル、コントローラー、さらにサーバーに接続でき、プライマリサーバーに障害が発生した場合に、単一のハードディスクドライブを別のサーバーにフェールオーバーできます。エンタープライズSANおよびNASアレイでは、多くの場合この機能が必要になります。ほとんどのベンダーの製品にSASドライブがあるのはこのためです。SASドライブには、多くの場合、ドライブをサポートする特定のホストバスアダプター（HBA）またはRAIDカードが必要になります。HBAは単純にプロトコルのサポートを提供しますが、RAIDカードは多くの場合、読み取りおよびバッテリーバックアップ式書き込みキャッシュ、ハードウェアRAIDおよびRAIDリカバリオフロードを実装しています。SASプロトコルのもう1つの便利な機能に、大規模な拡張のサポートがあります。通常、24個以上のドライブスレッドを備えたJBODが利用可能であり、単一のサーバーに大量のSSDを接続できます。

NVMeインターフェイス

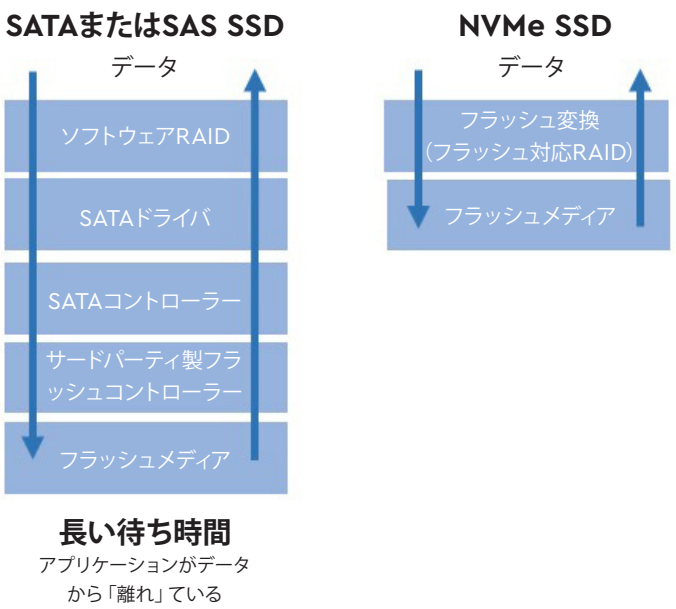

図1：NVMe vs SATAまたはSASプロトコルレイヤー

NVMeはすべてのモダンサーバーシステムに採用されているPCI Expressに基づいています。これは、「x1」、「x2」、「x4」、「x8」、「x16」として識別される、インターフェイス内の可変数のデータ「レーン」を備えたシリアルのポイントツーポイントバスです。これらのレーンはロジックを介在させずにプロセッサに直接接続するか、チップセットまたはPCIスイッチを介して接続します（適切に選択しないと、待ち時間が増し、最大帯域幅が狭くなります）。NVMeはPCI Expressベースであるため、通常はサーバー筐体内でのみ使用されますが、NVMeoF™（NVMe over Fabric）などのサーバー外のNVMeデバイスを実現する取り組みが進行中です NVMeはハードディスクドライブではなく、当初からメモリ速度のデバイスの超高速接続インターフェイスとして設計されました。SASおよびSATAプロトコルスタックに存在する複雑さとレイヤーの多くが完全に排除されます。キャッシュされていないメモリアクセスの数（実質的にプロセッサを遅くする）なども最小限に抑えられます。このテクノロジーにより、図1に示すように、ストレージベースのプロトコルよりも数倍高速で、前例がないほどに短いインターフェイス待ち時間（1～2マイクロ秒）が実現します。

NVMeはPCI Expressに基づいているため、PCIeの「世代」の概念を理解することが重要になります。すべてのPCIeスロットは「Gen X」として標記されます。Xは世代です。今日のほとんどのモダンサーバーは、レーンごとに最大1GB/sの帯域幅を提供する「Gen 3」インターフェイスを備えています（「x4 Gen 3」スロットはNVMe SSDで4 GB/sを実現できる可能性があります）。世代間の主な違いは帯域幅です。世代ごとに実質的に2倍になります。つまり「x4 Gen 3スロット」が「x8 Gen 2スロット」と同じ帯域幅を提供できることになります。この増加により、フロントローディングU.2 NVMeベースのSSDが広範に普及するようになりました。

高可用性NVMe「2×2」モード

一部のU.2 NVMe SSDは、「2×2」とも呼ばれる「デュアルポート」モードもサポートできます。x4の物理リンクは、それぞれ幅が半分の2つの個別論理チャネルに分割されます。2×2モードでは、インターフェイスで使用可能な最大帯域幅が2GB/sに制限されますが、SASドライブがデュアルポート（すなわち、別のコントローラーを介したストレージへの冗長リンク）で提供するのと同じ機能を提供します。この機能は、デュアルコントローラーを備えた高可用性NVMe JBOFデバイスでよく使用され、NVMeリンクまたはコントローラーがダウンした場合にシームレスなフェールオーバーを可能にします。

耐久性に関する考慮事項

これが重要な理由：基盤となるフラッシュがサポートするのは、有限数の消去および書き込みサイクルに限られるので、各SSDの製品保証では、その有効使用寿命にわたる限られた書き込みデータの量を許容しています。読み込みがほとんどのアプリケーション向けに耐久性が高すぎるSSDを選択すると、コストが不必要に上がります。一方、書き込み頻度の高いワークロード向けに耐久性が低すぎるSSDを選択すると、早い段階でアプリケーション障害が発生することになります。

ハードディスクドライブは磁気ドメインを使用して1と0を記録し、実質的に無制限数の書き込みを提供します。対照的に、フラッシュセル（データが記録される場所）は、実際に絶縁体内外で電荷を移動させてビットを格納し、有限回数だけ書き込むことができます。ビットを消去してフラッシュに書き込む物理プロセスは、100～10,000回以上実行可能で、実質的にデバイスを破壊します。これがフラッシュドライブに異なる耐久性評価が採用されている理由であり、フラッシュエラー訂正テクノロジーが重要になる理由です。

SSDの寿命計算

通常、耐久性として知られるSSD寿命は、一般的にDW/DまたはDrive Writes per Day（DWPD）またはTerabytes Written（TBW）で規定されます。これらの数値はデバイスがデバイスの寿命中に書き込み可能なユーザーデータの量を表します。

Drive Writes per Day（ドライブ書き込み/日）は、SSDの耐久性の仕様を表現するために使用される最も一般的な指標です。この数値は0.1未満から10以上とさまざまで、保証期間中に1日に書き込み可能なデータ量を示しています。

たとえば、3 DW/Dの1TB SSDで5年間の保証を受ける場合、1TB * 3 DW/D *（365 * 5）=～5.5PBの書き込みができます。

TBW = DW/D *保証年数* 365 *容量

注意：必要に応じて、容量をTBに変換してください。

異なるDW/D仕様のSSDの比較は複雑になる場合があります。同じDW/D指標であっても、ドライブ容量に応じて2つのSSDの合計書き込み寿命が大幅に異なる可能性があります。場合によっては、DW/Dが低いSSDが、DW/D評価が高いSSDよりも、実際にははるかに多くのデータ書き込みに対応できる場合もあります。たとえば、DW/Dが比較的高い10 DW/Dで4年間の保証が付帯する256 GBのSSDは、256 GB * 10 * 365 * 4 = 3.65PBを書き込むことができますが、DW/Dがはるかに低い1 DW/Dの6.4TB SSDでは、同じ4年間保証であってもこの数値はほぼ3倍となり、6.4TB * 1 *365 * 4 = 9.3PBの書き込みが可能になります。

SSDの寿命は「Terabytes Written」（テラバイト書き込み）として規定され、あらかじめ計算されています。通常、これにより寿命を直接比較できます。簡単に言うと、1000TBの書き込み仕様のドライブは、500TBの書き込みとして規定されたドライブに比べ2倍のデータ量を書き込むことができます。

SSD耐性性レベル

SSDの耐久性レベルにはさまざまな組み合わせがあり、場合によって単一の製品ライン内であっても複数の組み合わせが存在します。業界リーダーの多くは各レベルをHigh Endurance（HE）、Medium Endurance（ME）、Read-Intensive（RI）、Very Read-Intensive（VRI）と呼んでいます。

SSDでは、その使用目的に応じて耐久性の評価方法を選択します。測定インフラストラクチャがある場合は、自社のアプリケーションを監視して正確な値を取得できます（ただし、予想される拡張を考慮する必要があります）。そうでない場合には、表2の汎用的な経験値を参照し、適切な評価を選択してください。

エラー処理、電源保護、エンドツーエンドのデータ保護

これが重要な理由：コンシューマーSSDとエンタープライズSSDの真の差別化要因は、エラーケース処理にあります。予期しない電源障害、コントローラーまたはデータパスのランダムビット反転、その他のフラッシュエラーはすべてデータ破損を引き起こす可能性があり、こうした状況に対応できるとしても、その効果には大きなばらつきがあります。

SSDは高速ですが、適切なエンタープライズレベルのデータ保護を使用しないと、データ損失を起こすリスクがあります。エンタープライズSSDのデータ保護では3つの主要な領域、NANDエラー処理、電源障害保護、エンドツーエンドのデータパス保護に対応します。これらの領域では、SSDの処理パイプラインのさまざまな段階におけるデータ損失または誤ったデータ取得からの保護に対応します。

エラー訂正コードと信号処理

最も基本的なNANDエラー保護に、特定の領域で使用され、反転したビット数を検出して修正するエラー訂正コード（ECC）があります。訂正可能なビット数で測定されるこのECCの強度は、デバイスの信頼性と寿命に直接影響します。ECCを使用すると、古い「ノイズの多い」NANDであっても、そのデバイス寿命全体にわたり有効なデータを提供し続けることができます。SSD内に実装される実際のECCは、NANDの世代とジオメトリによって異なりますが、通常、セルあたりのレベルが高いほどECC要件が高くなります。そのため、TLCには通常MLCよりもはるかに高いECC訂正機能が必要になります。

高度な信号処理テクノロジーにより、SSDの寿命と信頼性を向上することもできます。これらがSSDの寿命にわたってNANDの動作を動的に管理します。クラス最高のコントローラーの一部では、デバイスの経年変化に応じてNANDセルのプログラミングと読み取りアルゴリズムを動的に変更し、エラー訂正コードで修正する必要がある未加工の読み取り/書き込みエラーレートを大幅に低減させています。Western Digitalのエンタープライズクラス Ultrastar SSDでは、高度なエラー訂正テクノロジーを実装しています。

電源障害保護

電源障害保護は、トランザクションデータベースを格納するデバイスや、書き込まれたデータの損失がないことを保証する必要があるその他の用途で重要になります。フラッシュプログラミングのブロックベースの性質により、すべてのSSDには、データが以前に消去されたフラッシュブロックに書き込まれる前に保存される、小さなRAMキャッシュがあります。通常の動作では、SSDは書き込みを「完了」したという信号をアプリケーションに返しますが、実際には、書き込まれたデータは一般的にRAMキャッシュにのみ存在し、実際のフラッシュ書き込みはまだ継続しています。このフラッシュ更新が完了する前にサーバーの電力が失われると、フラッシュへの書き込みが行われない場合があります。電源復旧時に、アプリケーションは失われたトランザクションをログから回復しようとしますが、不完全なフラッシュ更新を見つけることができないと、データ損失またはデータ破損につながります。

エンタープライズSSDでは、SSDに必要な蓄電器（通常はPCB上に搭載された複数のコンデンサー）を装備することでこれを防ぎます。これらのコンデンサーはメインサーバーから充電されます。予期しない電源障害では、SSDを動作させ、放電する前にRAMに残っている未完了の書き込みを完了するのに十分な電力が保証されます。高性能のエンタープライズレベルSSDでは、サーバーの起動時にこうしたコンデンサーのパフォーマンスと寿命も検証されます。これは従来のRAIDバッテリーバックアップユニットがバッテリーバックアップユニット（BBU）の寿命テストを定期的に実行する方法に似ています。

エンドツーエンドのデータ保護

最後に、エンドツーエンドのデータパス保護では、SSDに転送されるすべてのユーザーデータを一時的なエラー（ランダムな「ビット反転」）から保護します。サーバーのメインメモリがECCメモリを使用し、SSDは、内部プロセッサのデータパスにパリティ情報を含めるのと同じ方法ですべてのユーザーデータにチェックビットを追加し、オペレーションを実行する前に状態を確認します。この保護がないと、エラーとして検出しにくいビット反転がコントローラーを介して伝播し、最終的に破損したデータがフラッシュに書き込まれる場合があります。

NAND型

これが重要な理由：SSDは現在、フラッシュセル上に形成されており、さまざまに実装されています。これらは、シングルレイヤー、セルあたり1ビットの構成から、各フラッシュセルが最大16の異なる電荷レベルを格納する3次元スタックグループまでの範囲におよびます。各NAND型の強みと弱みが分かれば、特定の用途に適した寿命と信頼性のSSDを選択できるようになります。

NANDセルはSSDの最も基本的なストレージコンポーネントです。概略レベルでは、性能を示す重要な指標として、セルごとに格納されるビット数があります。この測定は、耐久性とNANDセルアレイのレイアウトに大きな影響を与え、密度とコストにも大きく影響します。

NANDフラッシュ型

NANDストレージとその他のストレージタイプの主な違いは、NANDストレージでは、書き込みおよび読み取りアルゴリズムを入念に調整することで、各NAND素子、つまりセルに複数ビットのデータを格納できることです。これにより、SSDあたりの使用可能なフラッシュビット数が大幅に増加しても、コストは安価に保たれます。SLC（シングルレベルセル）テクノロジーでは、NANDセルに格納できる値は0または1の値のみです。このメソッドは初期のSSDで使用されていましたが、パフォーマンスと信頼性に優れる一方でコストも高くなるため、今日では一般的には使用されていません。MLC（マルチレベルセル）テクノロジーは、2ビット（00、01、10、11）で示される4つの異なる電荷レベルのいずれかを使用します。

このテクノロジーは、SLCと比較して単一のNANDチップの容量を本質的に2倍にし、今日のSSDに見られる劇的なコスト削減の一部を担っています。エンタープライズクラスSSDの大部分はMLCフラッシュを使用しており、優れたパフォーマンスと耐久性を提供します。

最新の商用テクノロジーにTLC（トリプルレベルセル）NANDがあります。このテクノロジーでは、8つの異なる電荷レベル（000、001、010、011、100、101、110、111）でセルごとに3ビットを格納します。多くの場合、読み取りパフォーマンスは同等ですが、このテクノロジーの基本的な性質により、TLCの書き込みパフォーマンスは一般的にMLCと比較して低くなります。

QLC（クアッドレベルセル）NANDは、単一のNANDセルに4ビットのデータを格納し、Western Digitalを含む複数のベンダーからも発表されています。このテクノロジーでは、16個の異なる電荷レベル（0000、0001、0010、0011、0100、0101、0110、0111、1000、1001、1010、1011、1100、1101、1110、1111）を任意のセルに正確に格納するためにフラッシュが必要となるため、書き込み寿命が非常に短くなります。このため、想定される用途は更新が頻繁に行われないフラッシュアーカイブストレージに限られます。

NAND型	セルあたりのビット	電荷レベル	書き込み寿命	一般的なユースケース
SLC	1	2	非常に高い	一般的には見られないが、SSDの初期には最も高い読み取り/書き込みパフォーマンスの用途で使用。
MLC	2	4	高～中	さまざまな書き込み寿命での一般的な使用。
TLC	3	8	中～低	大規模なコンシューマーによる採用と企業での採用が増加。多くの場合、コストと耐久性が最適な比率の中規模なパフォーマンス用途に最適。
QLC	4	16	非常に低い	まだ一般には入手できないが、WORM（Write-Once-Read-Many）型のアーカイブ用途を想定。

表2: NANDフラッシュ型

業界の3D NANDへの移行

NANDセルの形成にはさまざまな方法がありますが、ユーザーの観点からするとその最大の違いは2D NANDと3D NANDの比較にあります。2D NANDでは、DRAMやプロセッサの形成方法と同様に、シリコンウエハー上にセルの単一層が形成されています。これはNANDデバイスを形成するために使われてきた従来からの方法です。2D NANDの容量を増やすために、メーカーはNANDデバイスのサイズを縮小してきました。しかし、NANDセルを微細化し、適切な読み取りに必要な電子数を確実に格納するには物理的な限界があります。そのため業界では、2次元NANDを3次元NANDに進化させ、より大きな容量を実現しています。

メーカーは、個々のNANDセルが最新世代の2D NANDよりも実際には大きい3Dプロセスに移行し、半導体リソグラフィーの制約を緩和して、電荷を安全に格納するのに十分な容積を確保しています。このサイズの増加を実現するために、NAND電荷蓄積層を高層ビルのように垂直に形成しています。層が追加されると、単一のダイで格納できるデータの合計量が増えますが、これらのすべてのNANDセルを積層して位置合わせする複雑さも増します。

ベンチマーク

これが重要な理由：多くの場合、設計者がSSDに注目する最初の理由はパフォーマンスの向上です。ただし、実際のパフォーマンスのベンチマークは単純でも簡単でもありません。適切な性能指数は、事前に調整された典型的なワークロードの下で取得する必要があります。これは、特定の用途でSSDのパフォーマンスを理解するために必要な手順です。

パフォーマンスの測定は、多くの場合、非常に難しくなります。SSDの実際のアプリケーションパフォーマンスを、アプリケーションを実行しないで決定する万能のアプローチはありません。ただし、特定のSSDでアプリケーションをテストできる場合でも、あらゆる入力の可能性を網羅する（たとえば、Webストアでのブラックフライデーや独身の日の間、または会計データベースの年末調整の間のピーク負荷）テスト方法が必要になります。そうしないと、その結果の信頼性が完全に失われる可能性があります。

SSDパフォーマンスの指標

このように、実際のアプリケーションワークロードでテストをするのは困難なため、ほとんどのSSDは特定の条件下で統合的なベンチマークを使用して測定されます。SSDの典型的な性能指数には、IOPS（Input/Output Operations per Second）、スループット、ロードおよびアンロードの待ち時間、QoS（Quality of Service）、外れ値があります。

IOPSはSSDが1秒間に実行できるI/O操作の数です。通常、これらの操作はすべて同じブロックサイズ（4KBと128KBは一般的なサイズですが、サイズはデータシートで指定する必要があります）で実行されます。読み取りと書き込みの混在を指定する必要もあります。これは、純粋な読み取りまたは書き込みタスクであるワークロードが非常に少ないためです。SSDでは混合ワークロードは多くの場合さらに処理が難しくなります。

スループットとは、単位時間内にSSDとの間で相互転送できるデータ量です。通常、MBPS（1秒あたりのメガバイト）またはGBPS（1秒あたりのギガバイト）で測定されます。データ転送の方向（純粋な書き込み、純粋な読み取り、または書き込み・読み取り混合ワークロード）もここで指定する必要があります。

待ち時間とは、ロードされたアプリケーションからSSDに操作が移動し、受信確認または要求された読み取りデータのいずれかが戻されるまでにかかる時間です。これは実質的に、SSDからのデータ転送の往復時間になります。待ち時間は通常ミリ秒またはマイ

クロ秒で測定されますが、今後でてくる永続的なメモリデバイスではこれがナノ秒に短縮される可能性があります。待ち時間を指定する場合、未処理のI/O数を含めることが重要です。「アンロード待ち時間」はシステムで他の作業が実行されていないI/O操作の待ち時間であり、「キューの深さXのロード待ち時間」は、SSDリソースをI/Oの合計Xで同時に共有する必要がある場合のI/Oの待ち時間です。通常、未処理のI/Oの数が増えると、ロードされる待ち時間が長くなるため、予想されるワークロードレベルでこれを測定することが重要になります。

QoS（Quality of Service）では、一定の信頼性レベルでのパフォーマンスの一貫性を指定された時間間隔で測定します。この種のレポートには大きなばらつきがありますが、一般的に「マクロQoS」では時間の経過に伴う平均IOPSの一貫性をレポートし、「ミクロQoS」では個々のI/Oの待ち時間をプロットし、超過などの測定を判断します。

適切なテストパラメータの選択

シミュレートされるワークロードは、ブロックサイズ、アクセスパターン、キューまたはI/Oの深さ、読み取り/書き込み比によって特徴付けられます。ブロックサイズは、単純に直近のタスクの未処理のI/Oサイズになります。データベースの場合、これは4KB、8KB、16KBになりますが、ストリーミングメディアの場合は128Kの大きなサイズになる場合があります。アクセスパターンは、シーケンシャル（SSDの連続した範囲が順番にアクセスされる）、またはランダム（各I/O操作の位置は以前のI/Oに依存しない）として定義されます。キューの深さ、つまりI/Oの深さは、I/O操作の並列性を示し、特定の時間に実行されるI/O転送数を反映します。

ほとんどのアプリケーションでは複数のスレッドを使用し、それぞれ異なるI/Oストリームを読み書きできるため、こうした構成を再現するには、多くの場合、16～256までのキューの深さが採用されます。シングルスレッドアプリケーションでさえ、仮想化環境またはコンテナ化環境で実行すると、キューの深さが高くなります。これは、単一のキューの深さを使用する複数のアプリケーションストリームを集約することによって実行されます。

読み取り/書き込み比（R:WまたはR/W）は、既存のデータを読み取るI/O操作と、新規または更新されたデータを書き込むI/O操作の割合を示します。多くのSSDデータシートには100%の読み取りまたは100%の書き込みパフォーマンスと記載されていますが、現実の世界では、このような純粋な読み取りまたは書き込みのワークロードは非常にまれです。SSDはこれらの純粋なワークロードに簡単に最適化できるため、報告される結果は実際のアプリケーションが達成できるレベルを超える場合があります。そのため、ベンチマークにいくつかの混合ワークロードを含めることが非常に重要になります。より現実的な60:40または70:30の読み取り/書き込み比率は、OLTPやOLAPデータベースのテストに有効な場合がありますが、90:10の比率はアクセス頻度の低いデータベースやログに適していると言えます。

SSDの事前調整の重要性

ドライブの事前調整の状態もすべてのテストで考慮される必要があります。ほとんどのSSDのパフォーマンスプロファイルは、「購入直後」（FOB）と「定常状態」（長期にわたるI/O操作によって全面的に書き込みが行われている）で大きく異なります。SSDのパフォーマンスを表した図2を見ながら、数時間にわたってランダムな4KBの書き込みワークロードが実行される新品のSSDから始めます。FOBのパフォーマンスはドライブの初期パフォーマンスを見るのに良い判断材料になる場合がありますが、数日または数か月間使用すると、SSDは通常「定常状態」モードになり、パフォーマンスレベルが低下します。多くのエンタープライズSSDは非常に高いドライブ使用率で数年間使用されると予想されるため、定常状態のほうが、FOBよりもアプリケーションパフォーマンスの実態をより正確に表していると考えられます。このため、すべてのテストで定常状態を考慮する必要があります。

SSDを定常状態のパフォーマンスで実行するには、事前の調整が必要になります。ドライブを複数回完全に書き込む（ブロックが複数回上書きされます）のが最適なテスト方法となります。監視対象のパフォーマンスが定常状態に低下するまで、選択したブロックサイズのランダム書き込みが、数時間または数日間実行されます。そのため、16KBの書き込みをテストするワークロードの場合、高いキューの深さ、16KB、100%書き込みパターンでドライブが繰り返し完全に書き込まれます。しかし、ほとんどのテストには複数のブロックサイズが含まれており、各事前調整段階に時間がかかる場合があるため、このプラクティスは煩わしいものとなります。精度とテスト時間の間の最善な関係を得るには、高いキューの深さのランダムな4KBワークロードを一連のテストの開始時に単純に実行するようにします。

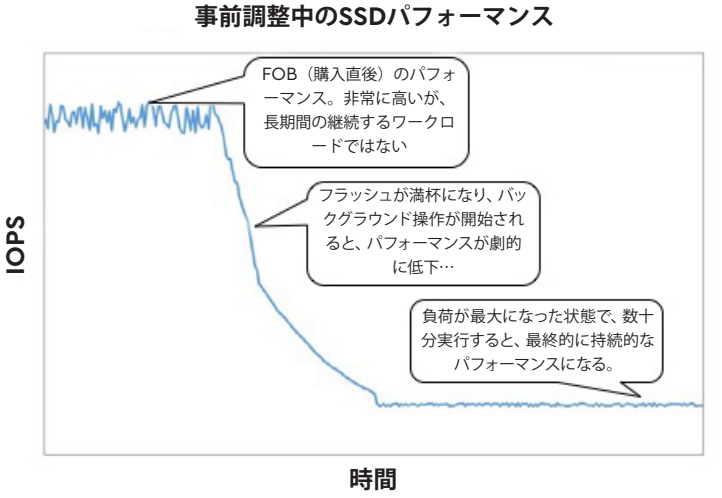

図2：FOBから定常状態へのパフォーマンスの変化

外れ値とアプリケーションへの影響

応答性

見落とされがちなSSDのパフォーマンス特性の1つに、QoS（Quality of Service）があります。これは個々のI/O操作によって、待ち時間にどれほどの違いが生じるかを表します。ここでの懸念事項は「ロングテールI/O」（平均よりも大幅に時間がかかるI/O）が最終的にアプリケーションのSLA違反を引き起こす可能性があることです。たとえば、I/O操作の平均待ち時間が100マイクロ秒のSSDであっても、5%のI/Oでこの待ち時間が最大1,000マイクロ秒（1ミリ秒）に急上昇する場合があります。ここで、操作あたり500マイクロ秒というような、99%の参照時間を保証する必要があるNoSQLデータベースを使用しているとします。SSDの平均アクセス待ち時間は非常に優れていましたが、上位レベルのアプリケーションSLAを満たすことができませんでした。この場合、平均待ち時間は150ミリ秒と若干高いが、1ミリ秒以上の待ち時間がわずか1%のSSDが、SLAを満たすより良い選択肢になります。SSDの外れ値のパフォーマンスは大きく変化する可能性があるため、特定の読み取り/書き込み比とブロックサイズの外れ値を調べることも重要です。

電力とオーバープロビジョニング

これが重要な理由：SSDでは、ほとんどの場合、その場で調整して、電力またはパフォーマンスエンベロープを最適化できます。これらのオプションをインテリジェントに活用することにより、データセンター全体で大幅な省電力またはパフォーマンスの向上を実現できます。

すべてのエンタープライズSSDは、通常インターフェイス要件内のデフォルトの電力エンベロープで規定されます（NVMe SSDの場合は最大約25ワット、SASの場合は約10～14ワット、SATAの場合は約5～10ワット）。電力または冷却の制限を受ける大規模なラックに配置する場合など、電力消費に厳しい制限がある特殊なワークロード向けに、一部のエンタープライズクラスSSDでは、電力消費制限を低く設定するように構成できます。この場合、SSDは電力を調整し、多くの場合パフォーマンスを犠牲にして、最大エネルギー消費を削減します。ワークロードで電力スロットリング設定を使用する場合は、電力スロットリング設定を有効にしてSSDをテストし、実際のパフォーマンスへの影響を確認する必要があります。

一部のエンタープライズSSDで、データセンター導入向けに公開されているもう1つの構成オプションに、可変オーバープロビジョニングがあります。通常、エンタープライズSSDは、ユーザーのアクセス可能なスペースがカードのフラッシュ総量よりも1%～20%少ない状態で出荷されます。この余分なスペースはパフォーマンスやデバイスの寿命を増やすために使用されます。一部のSSDではこの割合を変更できます。通常、オーバープロビジョニングを増やす（使用可能なスペースを減らす）と、使用可能な書き込みパフォーマンスも向上します（ただし、読み取りパフォーマンスは向上しません）。このオーバープロビジョニングの変更は使用可能なサイズに影響し、データの破壊をもたらすため、ファイルシステムまたはアプリケーションを使用する前に設定を完了する必要があります。この構成オプションを使用する場合は、必ず事前調整手順を完了し、一連のパフォーマンステスト全体を再実行して、実際の影響を確認してください。

監視と管理

これが重要な理由：SSDの導入は比較的簡単です。ただし、多数のSSDを設置する際は、一元化されたプラットフォームから健全性、パフォーマンス、使用率を監視できるツールを使用すると、時間を節約し、ストレスを軽減できます。

さまざまなインターフェイステクノロジーがさまざまな監視テクノロジーをサポートしています。SATA、SAS、NVMeインターフェイスで利用可能な基本的な監視テクノロジーは、SMART（自己監視、分析、レポートテクノロジー）と呼ばれています。これには単一のドライブの基本的な健全性とパフォーマンスを確認するための監視ツールが用意されています。SASインターフェイスは、これを基に、数十のエラーログページを追加し、ドライブの健全性をより詳細に表示できるように構築されています。最近のベンダーは、同じような詳細度で監視およびログインフラストラクチャを定義し、NVMeドライブの健全性を表示しています。より先進的なSSDメーカーは、データセンターのSSDポートフォリオ全体を統合および管理できる監視ツールを提供することができます。また、ファームウェアのフォーマット、サニタイズ、サイズ変更、アップデートなどのデバイス固有またはエンタープライズ全体の機能を実行する能力だけでなく、Active Directory/LDAP統合、自動電子メールアラート、耐久性レポートも提供する場合があります。

結論

エンタープライズSSDはデータセンターに大きな革命をもたらしました。大量のアーカイブデータとバルクデータを保存するデータセンターのハードディスクドライブにはまだ明るい未来がありますが、高速データベースやその他のアプリケーション向けのSSDの他に類を見ない待ち時間と帯域幅の利点を否定することはできません。

エンタープライズクラスのSSDには、さまざまな価格、パフォーマンス、フォームファクター、耐久性、容量にわたる幅広い種類があるため、特定の導入に適したSSDを選択するのは難しい場合があります。エンタープライズSSDを評価する際には、パフォーマンスをニーズに一致させるために、単なるIOPSや帯域幅の数値以外の指標も確認する必要があります。アプリケーションSLAを確実に満たすことができるQoS、実際のワークロードに適合する混合ワークロードパフォーマンス、ホットスワップやフェールインプレース対応アーキテクチャをサポートするフォームファクターを考慮するようにしてください。

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
米国（フリーダイヤル）：800.275.4932
その他の国： 408.717.6000

www.westerndigital.com