SDS、ハードウェア RAID、ソフトウェア RAID の比較

今後選ぶべきストレージはどれでしょうか?

このホワイトペーパーでは、NVMe ストレージのめざましい性能の進化と、SATA/SAS テクノロジーが組み合わさることで、高度なプロビジョニングの可能性が次々に生まれている状況を解説します。

#KingstonIsWithYou

変化の進捗状況を知る:業界における、現在と将来のストレージの動向

この 10 年以上におよび、以前は独立したハードウェアとして個別に管理されていた一連の IT サービスを、「ソフトウェア定義」の概念が仮想化に向けて推進しています。

最初は SDN (software-defined networking) から始まり、まもなく SDC (Software Defined Compute)や SDS (software-defined storage) へ発展しました。さらには、HCI (hyper-converged infrastructure) 内の要素として機能するよう、データセンターそのもの（ソフトウェア定義データセンター - SDDC）すら仮想化されるまでにいたっています。

このコネクティビティのすべてを実現したのは、I/O 仮想化をサポートするために CPU、GPU、ストレージ、ネットワーキングの統合から進化した PCIe (ピーシーアイエクスプレス) と呼ばれるハイスピードバスです。さらに PCIe は、複数のパッケージサイズで用意された新しいタイプの SSD と最適に接続できるよう設計された、NVMe (不揮発メモリエクスプレス) インターフェイスにも対応しています。

SAS や SATA と比べて、NVMe SSD は、帯域幅、遅延性、消費電力面ではるかに進化しています。2011 年の標準化以降、NVMe ストレージインターフェイスのスループットゲインは、PCIe の世代交代ごとに倍増し、PCIe 2.0 の 5GT/s から PCIe 4.0 の 16GT/s にまで増えました。32GT/s のスループットを実現した PCIe 5.0 が登場していますが、まだ主流ではありません。

現在、計画上では、すでに転送速度 64GT/s の PCIe 6.0 の仕様が実現しており、前の世代との互換性を維持しながら 128GT/s を目指した PCIe 7.0 仕様は 2025 年にリリースの予定です。以上のように、NVMe ストレージは、これらすべてのメリットを引き出すことができるでしょう。

接続を実現

最近では、PCIe 3.0 や PCIe 4.0 をサポートするハードウェアが多く存在し、NVMe SSD への移行は、それほど難しくはないようです。たとえば、アドインカード (AIC) は、PCIe スロットに直接差し込む使い方ができます。U.3 トライコネクターインターフェイスを備えた NVMe ドライブは、SATA ストレージや SAS ストレージと並行してサーバーアレイに簡単に組み込むことができます。多くのデータセンター運営にとって、アクセス面で不利な AIC と比較したとき、ドライブベイが前面にあって整備性の良い U.2 の実装は現実的な選択です。

あるいは、M.2 インターフェイスは、適切な M タイプコネクターのキーイングに対応したマザーボードや AIC を使用すれば、PCIe x4 のフルパフォーマンスを発揮できる NVMe ドライブと接続できるようになります。B タイプのキーイングは、SATA 3 や PCIe x2 の転送速度までしか対応できません。B+M キーイングを備えた SSD を見かけたとしても、ほぼ間違いなく、両方のソケットタイプとの互換性を備えた SATA 3 デバイスです。

これらのフォームファクターで NVMe ストレージをシステムへ物理的にインストールできたとしても、本当に肝心なのはそれらを支える基幹インターフェースであり、NVMe エコシステムでストレージを最適に管理する方法をゼロから考え直す必要が出てきます。

たとえば、U.2 NVMe SSD や M.2 NVMe SSD を単独で取り付ける場合、最適なスループットを得るには、デバイスごとに 4 本の PCIe レーン (PCIe x4) を使用します。しかし、たとえば PCIe レーンが 24 本の古いCPU システム場合、GPU で lPCIe x16 レーンを使用すれば、すぐに余裕はなくなります。しかし最近のハードウェアでは、最高 128 本の PCIe レーンを持つマルチコアと、レーン数を拡張できる PCIe スイッチングを利用できるため、これは大きな問題ではなくなります。そうは言っても、どんなインフラのアップグレード計画であれ、PCIe レーンの割り当てで最適な性能が得られるよう、配分を考慮しなければなりません。

NVMe ストレージでは、これまでとは別の方法で実装することが求められます。ドライブの構成では、ソフトウェア RAID やハードウェア RAID など使い慣れた方法を使用できます。それらの方法もまた、SATA SSD や SAS SSD を越える NVMe ストレージの能力を活かせるよう進化しています。

ソフトウェア RAID

NVMe ストレージのわかりやすいメリットは、主流の OS がすべて、NVMe ストレージをサポートするために NVMe ドライバーに対応している点です。ホストが Windows、Linux、macOS、Solaris などのいずれであっても、NVMe SSD を追加してデバイスにアクセスできます。VMware の仮想化環境は NVMe ドライバーをサポートしており、ソフトウェア定義ストレージのアプリケーションに対応する多くのオプションが使えます。

このように NVMe ストレージデバイスが使いやすくなったことで、主な OS で標準機能となったソフトウェア RAID アプリケーションが補強されています。シンプルで実質的に無料なソフトウェア RAID 機能は、一般エンドユーザー、ゲーミング愛好家、コンテンツクリエーター、本格的な企業向け実装など、すべての形態で利用できるため、堅牢なストレージ管理の基本機能を初めて使い始める際に便利です。

基本のソフトウェア RAID アプリケーションで利用できるのは、パフォーマンス重視の RAID 0 (ストライプ) と安全性重視の RAID 1 (ミラー) のみです。**確かに、ハードウェア RAID は、ソフトウェア RAID よりも多くの RAID レベルに対応しています。**そうであっても、Linux のデフォルトソフトウェア RAID である mdraid のようなアプリケーションですら、パフォーマンスとデータ保護のバランスをとる組み合わせとして、RAID 4/5/6/10 を提供しています。

SSD の容量はいまだに従来のハードドライブの容量に追いついていないため、指定された台数のドライブに対するストレージ要件も、RAID アレイ構築時の重要な検討事項です。さらに、RAID ストレージ環境内でソフトウェアを使用してデータ配分とパリティチェック機能を管理すると、これらのルーチンを実行するホスト CPU に影響が及びます。アルゴリズム操作は、たとえば読み取りよりも書き込みの計算に負荷が掛かるなど、多様な複雑さを持ちます。データのスループットが膨大で、RAID 構成の冗長性が高い場合、これらのタスクがシステム全体のパフォーマンスに影響をおよぼすおそれがあります。

ソフトウェアのライセンスがコア単位で請求される場合、システムにストレージのタスクを任せるのは合理的と言えるでしょうか?これが、長い間ハードウェア RAID が好まれてきた理由ですが、現在はもう SATA/SAS 環境の時代ではありません。ソフトウェア RAID につきものの低パフォーマンスは、NVMe が持つ低遅延と高スループットと、PCIe バスへの直接アクセスである程度は解消されます。

設計段階から改良

さらに、SATA インターフェイスはハードドライブを念頭に設計されており、SSD に使用する際は、常に妥協が求められます。HDD に優るSATA SSD の転送速度は非常に便利ですが、それはフラッシュストレージが持つ優位性のごく一部に過ぎません。SATA で使用する AHCI （高度ホストコントローラーインターフェース） は、回転するディスクの物理的制約を補うために 120 個のコマンドを備えてレガシー技術の特殊性に対処し、フラッシュストレージへのアップグレードを可能にしていますが、最終的にはボトルネックになっています。これに対して NVMe は、管理用に 10 個、I/O (読み取り、書き込み、消去) 用に 3 個の、計 13 個のコマンドがあれば最低限機能します。

コマンドキューの面では、AHCI/SATA テクノロジーの場合、32 個のコマンドを送信できるキューが 1 本あるだけです。一方、NVMe には、64,000 本の I/O キューがあり、キューごとに 64,000 個のコマンドを送信できるので、使用する CPU サイクル数が非常に少なくてすみます。

NVMe ストレージで使用される整合化された PCIe データパスは、その巨大なスループットと効率性と合わせ、この分野でソフトウェア RAID に対する捉え方を一変させました。この領域では、ソフトウェア RAID は、制約の存在よりも、効率性に目が向けられています。実際、多くの人にとって、ソフトウェア RAID 以外に選択肢はありません。従来の方式のハードウェア RAID は、NVMe ストレージの規模を拡張できるよう、機能を進化させる必要があったからです。

ハードウェア RAID

ハードウェア RAID PCIe カードには専用のコントローラーチップがあり、このチップは対象となるストレージハードウェアから RAID アレイを作成して管理するために必要な計算機能を実行します。この処理はすべて RAID カードで行われるため、ハードウェア RAID は、ホストプラットフォームには何ら処理の負担を課すことなく、さまざまな複雑さの RAID レベルを幅広く提供できます。

RAID 処理アルゴリズムに貴重なホスト CPU リソースがとられないため、読み書きの速度が最適化され、ドライブのホットスワッピングが可能になります。ソフトウェア RAID は専用の演算リソースを持たないため、大容量 SAS/SATA 環境では、遅延とスループットが増加します。ハードウェア RAID と違って、ドライブの交換では取り外し前に RAID 管理作業が必要になることが多く、しばしば再起動も必要になります。

PCIe ハードウェア RAID カードは高価ですが、遅延が短く、データ保護とキャッシングの機能を持ち、ドライブアレイを拡張できるため、企業向けストレージ管理で必須となっています。また、PCIe ハードウェア RAID カードはさらに進化を遂げています。NVMe 専用の RAID カードは市場で普及し始めたばかりですが、Broadcom、Marvell、Microchip などのメーカーが提供しているトライモード PCIe 第 4 世代 RAID-on-chip (ROC) カードは、SATA、SAS、NVMe の組み合わせをサポートしています。

こうしたハードウェア RAID カードは、複合ストレージ環境で NVMe SSD を共存させるための手近な方法となります。U.2 バックプレーンは、基本的なケーブル配線を行なうだけで、U.2 フォームファクターの SATA/SAS と NVMe SSD を組み合わせて使用できるようになります。

U.3 規格が出現し、このフォームファクターがさらに進化しました。ケーブル配線を一元化して複雑さを軽減したため、真のトライモードバックプレーンを利用できるようになっています。ただし、欠点もあります。U.3 の物理的ドライブインターフェイスは同じままですが、ピン構成が変更されているのです。そのため、U.3 ドライブは U.2 バックプレーンで使用できますが、U.2 ドライブは U.3 バックプレーンで使用できません。

U.3 は混合と適応の機能で十分な目標を果たしたように見えますが、このような構成が広く受け入れられるかどうかは別の問題です。

状況の認識

UBM（汎用バックプレーン管理）規格の登場で、ストレージの混合実装が可能になり、U.2 設計と U.3 設計両方との互換性が実現しています。20 社を越える業界大手のストレージハードウェアベンダーで構成されたコンソーシアムから支持される UBM は、ホストデバイスとコントローラーデバイスがバックプレーンの機能を認識できるようにしています。また、ドライブベイが 1 つだけの場合でも、さまざまなドライブタイプ (SATA、SAS、NVMe) を検出して監視する機能をサポートしています。さらに、UBM は SATA/SAS エクスパンダーと PCIe スイッチでも機能し、U.2 と U.3 のシステムアーキテクチャを強化する実用的なバックプレーン管理機能を多数備えています。

トライモードの RAID や HBA (ハードウェアバスアダプター) カードでは、x8 または x16 の PCIe ホストレーンを使用し、PCIe スイッチングを通じてレーン数を飛躍的に増やして、実質的な帯域幅を増加させます。カードの仕様には、たとえば最大 32 の NVMe デバイスをサポートすると記載されますが、これは x4 のフル転送速度で 8 つの NVMe SSD をサポートするのとは違います。その場合は、32 本の PCIe レーンが必要になります。理論的には、転送速度 x1 の物理 NVMe ドライブを32 台稼働できます。たとえ PCIe 3.0 環境であっても各ドライブは 1000MB/s で動作し、これは SATA の 600MB/s スループットよりも 66% 高速化されます。しかし、そのような構成は、PCIe レーンの並行利用で NVMe SSD が持つ最高水準のパフォーマンスがさらに強化されていることを考慮すれば、NVMe SSD ストレージの最善な使用方法とは言えません。複合的な使用方法では、トライモードコントローラーは NVMe ストレージ専用に、x8 または x16 レーンしか割り当てられず、ドライブ数を減らすか、スループットを低下させるかを選択しなければなりません。

単一のバックプレーンに各種のドライブをシームレスに統合した状態で、ホット (NVMe)、ウォーム (SAS/SATA)、コールド (SATA/HDD) ストレージ需要を 1 台のシャーシで管理できる上級システムの構築が普及するか否かは、まだわかりません。

結局のところ、レガシーストレージデバイスとの互換性を維持するために PCIe レーンの割り当てを分割するのは妥協にすぎず、NVMe は利用可能になりますが、制約と負担がともないます。現在、既存の SAS/SATA ストレージ実装で対応できている多くの運用では、信頼性を確保して容量を増強できる、ドライブの入れ替えにしか興味がありません。U.2 ストレージはしばらくの間利用され続けるでしょうが、既存の SAS/SATA ストレージ資産と、低価格の専用コントローラーやエクスパンダーを活用する場合は、1 種類のデバイスレーンを使用する構成が普通になりそうです。同じく、パフォーマンスの強化と容量を最大化するには、NVMe SSD のみに絞り込むのがベストです。

NVMe ストレージの採用が進むかどうかは、ワークロードの負荷と、既存システムをどの程度強化できるかによって決まります。NVMe に絞った実装へ大幅に投資しているクラウドサービスのプロバイダーは、さまざまな顧客のニーズに対応できる、階層化された新しいサービスを提供するために帯域幅を大きく増やし、すでにそのメリットを実現しています。

見通しの見直し

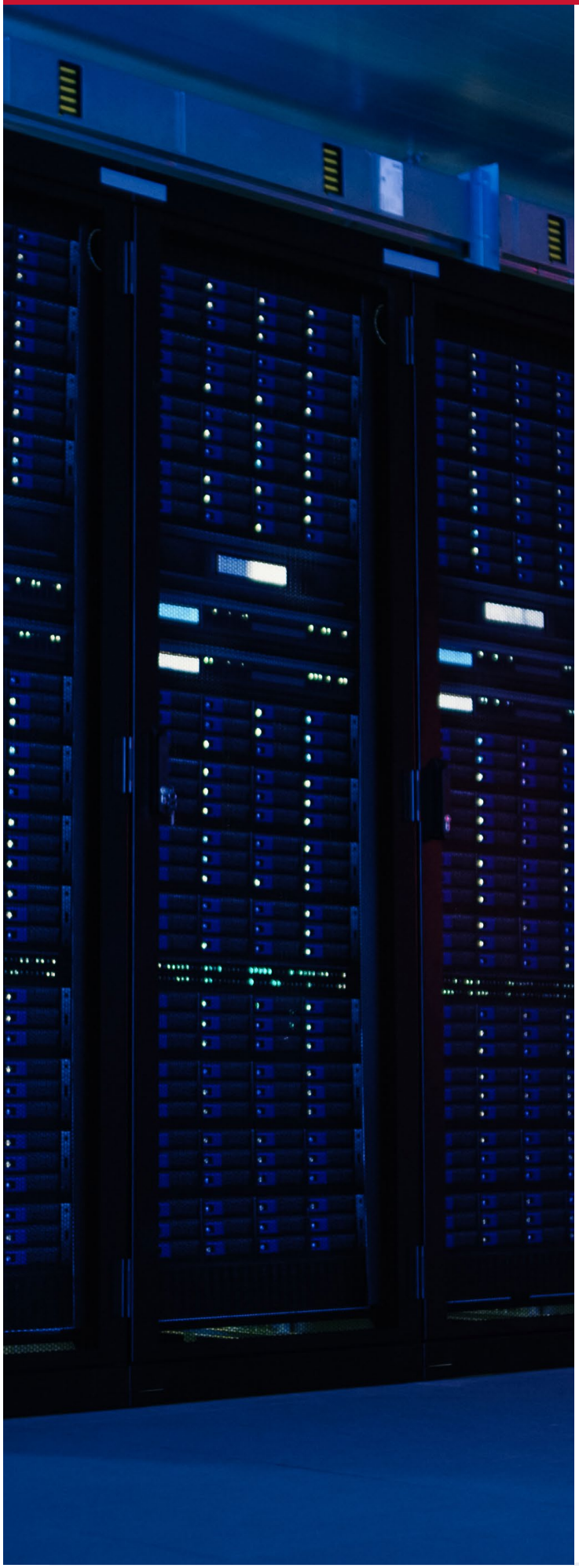
NVMe を利用した機動性の高いクラウドプロバイダーと、従来型のデータセンターの両極端の間に、機能数、効率性、拡張性の改善を求める一般企業が存在します。

こうした企業は、経費、メリット、統合、最適化を調査して段階的に展開を進める、より的を絞ったアプローチで NVMe を採用しています。

NVMe ストレージをキャッシングに使用すれば、アプリケーションでコーディングの質が低いためにレイテンシやスループットの改善効果が減るなど、システム上の欠陥がすぐに明らかになります。その他のボトルネックも自ずと明らかになるため、PCIe/NVMe エコシステムでパフォーマンス上のメリットを活かすにはそうした問題を解決しなければなりません。

これは、同等品との交換ではなく、自転車を新幹線に置き換えるようなアップグレードです。このため、SSD 仕様も見直す必要があります。サービスレベル契約で、NVMe ストレージの真の威力を認識していない運用判断に頼っている可能性があります。

その一例が、保証期間内のフラッシュストレージの耐久性を決定する DWPD（1日あたりのドライブ書き込み）です。フラッシュドライブには、ライトアンプリフィケーション（書き込みの増幅）という問題があります。これは、メモリセルにデータを保存する手法が理由で、SSD の摩耗が増加する現象です。短く言うと、データはセルに直接保存されるわけではありません。データを上書きする前に以前のデータを消去する必要があります。この複雑な処理は、時間をかけてストレージの劣化を招きます。オーバープロビジョニングは、SSD の予備容量の一種です。この問題の解決策として使用されており、ガベージコレクションなど、ドライブでハウスキーピングのルーチンを実行します。これは、ライトアンプリフィケーションの主な原因であるストレージブロックを解放するためにデータを再割り当てするプロセスです (解放されたストレージブロックは書き込みに備えて消去されます)。



区分分け

NVMe 2.0 仕様に最近加えられたのは、ZNS（ゾーン分割ネームスペース）です。これは、NVMe SSD の読み書き処理の新しいアプローチです。ゾーンブロック管理インターフェイスは、ホストと NVMe SSD の間に配置されます。ゾーニングは、ディスクパーティショニングと一部似ていますが、それはホストアプリケーションレベルの話です。ZNS は、SSD とホスト間の通信を可能にして、パフォーマンス水準を説明するか、示唆します。たとえば、ZNS が書き込みと消去を連続して実行する際に、データ配置の最適なパターンやレイアウトの詳細を提供します。

この協調的なやり取りを通じて、ストレージ管理機能の一部がホストアプリケーションに移されます。これによって、オーバープロビジョニングの必要性が減り、最大 20% も余分にストレージ容量を確保できるというメリットがあります。ZNS を実装すると、I/O のレイテンシが改善され、ドライブアンプリフィケーションが 4 分の 1 または 5 分の 1 へ軽減されます。さらに、パフォーマンスパターンを予測しやすいよう、ゾーンごとに特定のワークロードやデータタイプを割り当てることができます。

ZNS（ゾーン分割ネームスペース）の普及は始まったばかりですが、ZNS はすでに Linux カーネル 5.9 の機能として組み込まれています。さらに、ZNS の研究には、大規模なハイパースケーリング運用着目した Microsoft、Alibaba、NetApp が資金を提供しています。これは、産業レベルで ZNS が採用されるのは、時間の問題だということを表しています。

ZNS が進化し、ZNS を採用する NVMe ドライバー数が増加するにつれ、アプリケーションは、この機能をフルに利用するためにアップデートｓる必要があります。一部の事例では、既存の NVMe SSD による実装で必要なのは、ファームウェアのアップデートのみになっています。

システムアーキテクトにとって、厳密な仕様に対応するために、DWPD の実際の意味を根本から見直す時が来ました。ZNS を実装すると、ライトアンプリフィケーションの量が大幅に削減され、ドライブの耐久性が大幅に改善されます。結果として、必要なドライブ数は何台になるでしょうか?オーバープロビジョニングの大幅な軽減により、ドライブ容量はかなり増えます。データ管理の将来という意味では、NVMe SSD と ZNS がインターフェースされることで、少ない投資で多くの成果が得られます。

ソフトウェア定義ストレージ

NVMe は、M.2 ドライブと PCIe アドインカードの組み合わせから U.2 ストレージや U.3 ストレージに至るまで正しい採用の道筋を示してくれます。新たに登場した EDSFF（エンタープライズとデータセンターSSDフォームファクター）は、NVMe エコシステムのために設計された、ロングとショート (L と S) の長さで 2 つの幅 (E.1 と E.3) の構成が用意されたドライブを使用する、もう 1 つのストレージフォーマットです。E.1L ドライブは、1U シャーシに高密度のストレージを実装でき、拡張性に富みより柔軟な E.1S サイズは熱効率面でメリットがあります。U.2 2.5 インチ SSD の交換用にパッケージ化された E.3 ドライブは、より従来型の 2U サーバーとドライブアレイシャーシにも格納でき、ストレージ密度向上のために、ドライブ当たりのフラッシュメモリチップ数が増えた設計になっています。

NVMe を 1 つの共通標準として利用し、すべての主流 OS でドライバーがサポートされれば、間違いなく、上記のすべてのオプションを簡単に実装できます。これを選択できるかどうかは、ワークロードと冗長性の要件に最も適合性の高いストレージ特性と構成次第です。そのためには、高負荷の運用を軽減する SAS/SATA ハードウェアともに、NVMe ストレージのエッジサーバーへの統合が考えられます。ハードディスクドライブや、テープバックアップも、ストレージインフラの一部に組み込まれる可能性があります。エンタープライズストレージ管理では多数の独自プラットフォームが存在しており、これらの異種ストレージシステムを統制しようとすると、すぐに複雑になりがちです。その場合は、複合ストレージシステムの運用を調和して利用を最適化する手段として、SDS (ソフトウェア定義ストレージ) が利用できます。

ソフトウェア定義ストレージの領域では、利用できるストレージリソースは、ストレージハードウェアから抽出され、仮想化されます。産業標準のプロトコルを使用すれば、独自ハードウェアでさえ SDS 仮想化でアクセスでき、モノリシックなストレージアプライアンスは解放され、市販サーバーで構成された、新しい低コストで拡張性の高いストレージを利用した大きなプールを構成できます。こうして結合を解除することで、ストレージハードウェアの交換、アップグレード、拡張時にも混乱を避けることができます。

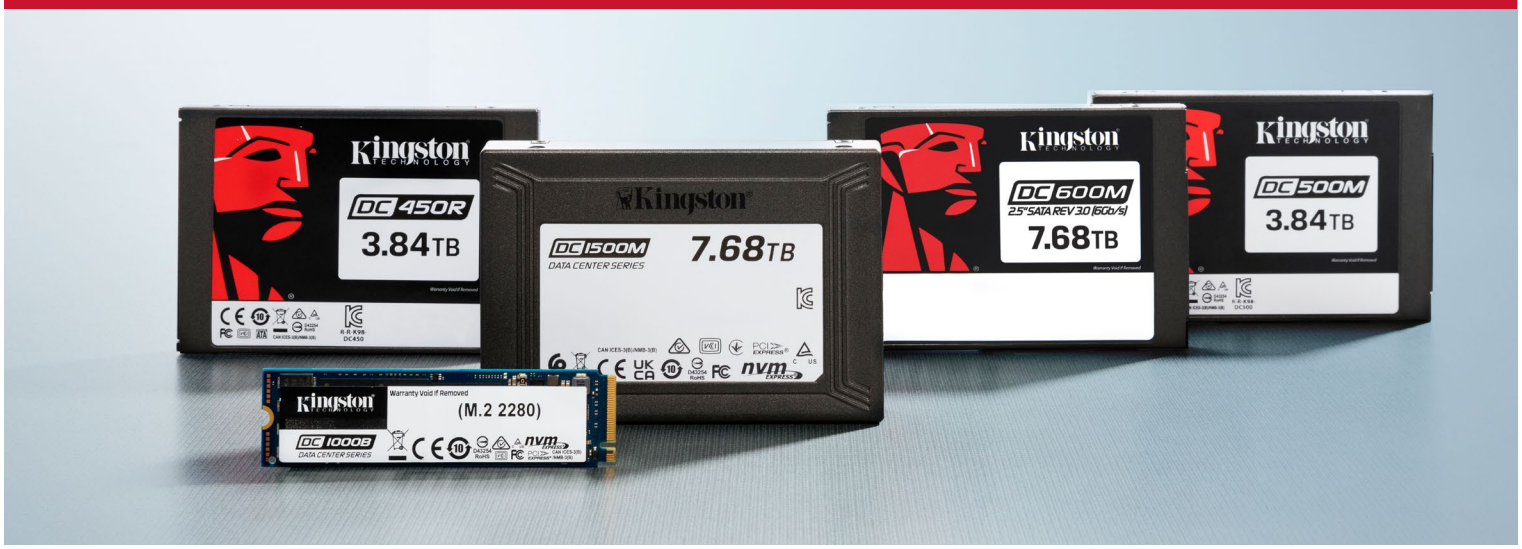
利用できるすべてのストレージを仮想プールに統合した状態では、プロビジョニングを決定をする必要があるため、自動化を含めて、割り当てを支援する幅広い機能が存在します。SDS ダッシュボードでは、各種プールにあるストレージのハードウェアプロファイルに基づいてホット、ウォーム、コールドストレージを特定します。スクリプトを利用して、これらのリポジトリに最もよく適合したデータ負荷を割当て配分するよう、タスクを実行できます。

SDS はその仮想ストレージレイヤーを通じて柔軟性と拡張性をともに提供し、企業の需要とクライアントの変化しやすいニーズに応じたストレージ環境の作成と実装を、VM (仮想マシン) のキャッシングやプロビジョニングからミラーリングやレプリケーションに至るまで管理します。

NVMe SSD の場合、SDS プラットフォームは、NVMe パススルーという機能を利用し、PCIe バスを通じてストレージに直接アクセスできます。たとえば VMware には、ESXi/vSAN SDS プラットフォームに対応した専用の NVMe ストレージドライバーがあります。これは VMDirectPath I/O と呼ばれる機能で、仮想マシンに対して NVMe ストレージを直接割り当てることができます。ホスト CPU の構成に応じて、仮想マシン当たり最大 16 台のパススルーデバイスをサポートできます。

NVMe パススルーを有効にすると、全体として、ホストからの干渉が最小限に抑えられ、パフォーマンスが改善され、VM インスタンスやその他のサービスのための NVMe SSD の構成がシンプルになります。SDS はNVMe ソフトウェア RAID を直接構成できるため、サードパーティのソフトウェア RAID コントローラーやハードウェア RAID コントローラーが NVMe RAID 機能をサポートしているかどうかは、あまり問題にならなくなります。

SDS にはデータ管理の万能薬になる可能性がありますが、高価について初期設定が複雑なため、よりシンプルな要件を持つ企業は採用を躊躇することもあります。ただし、ストレージそのものと同じく、費用は規模に合わせて調整でき、小規模なハードウェア実装に合わせてさまざまなバージョンが用意されています。



変化に合わせた対応

ストレージは進化していますが、既存の資産は廃棄計画の戦略に組み込まれていることが多いため、一夜にして交換されることはめったにありません。そのため、ハードウェアディスクや SATA SSD などの技術を使用したストレージの開発も継続されます。それぞれのストレージに相応しい用途があり、ストレージアレイで役に立つサービスを提供しています。たとえば Kingston の DC600M 2.5 インチ混合型エンタープライズ SATA SSD は、容量を 7.68TB に倍増しました。

ハードウェア RAID とバスアダプターは、世界各地のデータセンターで引き続き優勢であり、ベンダーはますます拡大を続ける IT 業界の需要に応じて開発を続けています。

Broadcom や Microchip との提携を通じて、Kingston SSD は、今日のデータ主導型テクノロジーの厳しい要求に合わせるため、徹底的に試行されています。

これら大手ベンダーのストレージアダプターを利用し、テストプログラムでは苛酷なワークロードと困難な構成を取り入れることで、Kingston エンタープライズ SSD が必要なパフォーマンス、耐久性、信頼性を備えていることを確認しています。もちろん、Kingston の U.2 DC1500M PCIe NVMe Gen3x4 エンタープライズ SSD にも、このテストをすべて行っています。最大容量 7.68TB でと1 DWPD 化を備え、最新世代のサーバーとストレージアレイで十分以上に活躍できます。

ソフトウェア定義ストレージのシステムは市販のハードウェアをベースに構築できますが、SSD の選択は、規模が大きいほど重要性を増します。市販グレードの SSD は低価格ですが、耐久性と長期間の高帯域負荷に対応して構築されたエンタープライズ SSD と比較すれば、経済的であるとは言えません。ワークロードを効率的に管理でき、顧客の期待に応えられるよう、ハイパーコンバージドインフラが提供するサービスには、パフォーマンスが予測できることが求められます。Kingston エンタープライズ SSD は VMWare ストレージアプリケーションの動作要件を満たしており、ソフトウェア定義ストレージの仮想世界であっても、現実の目標を達成できることが確認されています。

ストレージのプロビジョニングが変化しているのは確かですが、その速度はビジネスモデルによって異なります。レガシーインターフェイスの継続性から、NVMe のイノベーションまで、あらゆるレベルでさまざまな改善が行なわれています。アップデートが困難な場合は、Kingston の専門家に照会サービスがお手伝いします。業務内容や予算に応じて重要な決定を行なえるよう、無料のサポートをご提供いたします。御社がどの段階にあろうとも、Kingston はお客様と共にあります。

#KingstonIsWithYou