Article
Tomohiro Iwamoto · Jun 29, 2020 32m read

InterSystems データプラットフォームとパフォーマンス - パート8 ハイパーコンバージドインフラストラクチャのキャパシティプランニングとパフォーマンス

ここ数年の間、ハイパーコンバージドインフラストラクチャ（HCI）ソリューションが勢いを増しており、導入件数が急速に増加しています。 IT部門の意思決定者は、VMware上ですでに仮想化されているアプリケーションなどに対し、新規導入やハードウェアの更新を検討する際にHCIを考慮に入れています。 HCIを選択する理由は、単一ベンダーと取引できること、すべてのハードウェアおよびソフトウェアコンポーネント間の相互運用性が検証済みであること、IO面を中心とした高いパフォーマンス、単純にホストを追加するだけで拡張できること、導入や管理の手順が単純であることが挙げられます。

この記事はHCIソリューションの一般的な機能を取り上げ、HCIを初めて使用する読者に紹介するために執筆しました。 その後はデータベースアプリケーションの具体的な例を使用し、InterSystems データプラットフォーム上に構築されたアプリケーションを配置する際の、キャパシティプランニングとパフォーマンスに関する構成の選択肢と推奨事項を確認します。 HCIソリューションはパフォーマンスを向上させるためにフラッシュストレージを利用しているため、選択されたフラッシュストレージオプションの特性と使用例に関するセクションも含めています。

この記事のキャパシティ計画とパフォーマンスに関する推奨事項は、VMWare vSANに特化しています。 ただし、HCI市場で成長しているのはvSANだけではなく、同じく導入件数が増加しているNutanixをはじめとする他のHCIベンダーも存在します。
選択したHCIベンダーにかかわらず多くの機能は共通しているため、この記事の推奨事項は広く適用できます。 ただし、どの場合もアプリケーション固有の要件を考慮しながらHCIベンダーとこの記事の推奨事項について話し合うことが得策です。

InterSystems データプラットフォームとパフォーマンスに関する他の連載記事のリストはこちらにあります。

HCIとは？

厳密に言えばコンバージ
ドソリューションは以前から長らく存在しますが
、この記事ではWikipedia
に「ハイパーコンバージェンスはパッケージ化された複数の個別システムから、市販されている商用x86ラックサーバーですべてが実行されるソフトウェア定義のインテリジェントな環境に進化している…」と記載されているような最新のHCIソリューションを取り上げています。

では、HCIは単独で存在するものなのでしょうか？

違います。 ベンダーに相談する際は、HCIには多くの置き換え可能な要素があることを覚えておく必要があります。コンバージドとハイパーコンバージドは具体的な青写真や標準ではなく、どちらかといえば一種のアーキテクチャなのです。 HCIハードウェアには商品性があるため、市場では複数のベンダーがソフトウェアレイヤーのほか、コンピューティング機能、ネットワーク機能、ストレージ機能、管理機能を組み合わせたその他の革新的な方法を使って差別化を図っています。

ここではあまり深追いしませんが、HCIの名が付いたあるソリューションでは、クラスター内の複数サーバー内にストレージを配置したり、サーバーのクラスターと独立したSANストレージ（複数の異なるベンダー製のものである可能性もあります）を使ったより普通の構成を作ることができます。これらはまた、相互運用性がテストおよび検証されており、単一のコントロールプレーンから管理できます。 キャパシティとパフォーマンスを計画する

際は、ストレージがSANファブリック（ファイバーチャネルやイーサネットなど）を介して接続されたアレイにあり、ストレージプールがソフトウェアで定義され、サーバーノードの各クラスター内に配置され、複数のサーバー上でストレージが処理される場合とは異なるパフォーマンスと要件を備えたソリューションを検討する必要があります。

では、改めてHCIとは何でしょうか？

この記事ではHCI、特にストレージが物理的にホストサーバー内にあるVMware vSANに焦点を当てています。 このようなソリューションでは、HCIソフトウェアレイヤーが処理を実行するクラスター内の複数のノードのそれぞれの内部ストレージを1つの共有ストレージシステムのように機能させています。 そのため、HCIソフトウェアにコストがかかったとしても、エンタープライズストレージアレイを使用するソリューションと比較して、HCIを使用すると大幅にコストを節約できる可能性があることがHCIを採用するもう一つの要因となっています。

> この記事では、HCIがコンピューティング、メモリ、ストレージ、ネットワーク、および管理ソフトウェアを仮想化したx86サーバーのクラスターに統合するソリューションについてご紹介します。

一般的なHCIの特徴

上記のとおり、HCIソリューションの一例にはVMWare vSANとNutanixがあります。
どちらもHCIに対して類似した高レベルのアプローチを採用しており、良い典型例だと言えます。

- VMware vSAN にはVMware vSphereが必要であり、複数のベンダーのハードウェアで利用できます。 利用可能なハードウェアの選択肢は多数ありますが、これらはVMwareのvSANハードウェア互換性リスト(HCL)に厳密に依存しています。 ソリューションは、パッケージ化および事前構成されたEMC VxRailなどで購入できます。または、HCLでコンポーネントを購入して独自に構築することもできます。
- Nutanixは、最大で2U、4ノードのアプライアンスを構成済みブロック化したハードウェアを含むオールインワンソリューションとして購入および導入することもできます。 Nutanixソリューションは、他のベンダーのハードウェアで検証された独自のソフトウェアソリューションとしても利用できます。

実装にはいくつかのバリエーションがありますが、一般的にHCIにはパフォーマンスとキャパシティの計画に関して、あなたが知っておくべき一般的な特徴があります。

- 仮想マシン(VM)はVMware ESXiなどのハイパーバイザーで実行されますが、Hyper-VやNutanix Acropolis Hypervisor(AHV)などのハイパーバイザーでも実行されます。
  NutanixはESXiを使用して導入することもできます。
- ホストサーバーは多くの場合、コンピューティング、ストレージ、ネットワークのブロックに統合されます。 例えば、4つのノードを持つ2Uアプライアンスがあります。

- 管理と可用性のために、複数のホストサーバーがクラスターに統合されます。
- ストレージは階層化されており、オールフラッシュ、またはフラッシュキャッシュ層とキャパシティ層として利用する回転式ディスクとのハイブリッドになっています。
- ストレージは、容量、パフォーマンス、可用性を確保するためのデータ配置とポリシーを含むソフトウェア定義のプールとして表現されます。
- 容量とIOのパフォーマンスは、クラスターにホストを追加することでスケーリングされます。
- データは複数のクラスターノード上のストレージに同期的に書き込まれるため、クラスターはデータを失うことなくホストやコンポーネントの障害に耐えることができます。

- VMの可用性とロードバランシングは、vMotion、VMware HA、DRSなどのハイパーバイザーによって提供されます。

上記の通り、外部ストレージアレイ、ストレージ専用ノードのサポートなど、このリストに変更を加えた他のHCIソリューションもあります。このリストはベンダーのリストと同じく長いです。

HCIの採用が加速し、ベンダー間の競争がイノベーションとパフォーマンスの向上を推進しています。 HCIがクラ

ウドを導入するための基本的要素になっていることも注目に値します。

InterSystemsの製品はHCIでサポートされていますか？

オペレーティングシステムが仮想化されている場合を含め、さまざまなプロセッサのタイプとオペレーティングシステムに対してInterSystemsの製品を検証およびリリースするのは、InterSystemsのポリシーであり、決まりでもあります。 詳細については、InterSystemsサポートポリシーおよびリリース情報を参照してください。

例えばx86ホスト上のvSANの場合、Caché 2016.1をRed Hat
7.2オペレーティングシステム上で実行することができます。

注意：独自のアプリケーションを作成しない場合は、アプリケーションベンダーのサポートポリシーも確認する必要があります。

---

vSANキャパシティプランニング

このセクションでは、Caché、Ensemble、HealthShareなどのInterSystemsデータプラットフォーム上のデータベースアプリケーションにVMware vSANを導入する場合の考慮事項と推奨事項について説明します。 ただし、これらの推奨事項はHCIベンダーと検討するための一般的な構成関連の質問リストとして使用することもできます。

VM vCPUとメモリ

まず、複数の同じプロセッサを持つVMware ESXiにアプリケーションを導入するためにすでに使用中のものと同じキャパシティプランニングのルールをデータベースVMのvCPUとメモリに使用します。

Cachéの一般
的なCPUとメモリのサイ
ジングについて復習するには、この連載の他の記事
のリスト「キャパシティ計画とパフォーマンスに関する連載の索引」を参照してください。

HCIシステムの特徴の1つに、非常に低いストレージIOレイテンシと高いIOPS機能があります。
この連載の第2回目の投稿にあった、CPU
/ メモリ / ストレージ / ネットワークを示すハードウェアの食品群の図を覚えていらっしゃるかもしれません。 その記事ではこれらのコンポーネントがすべて相互に関連しているため、1つのコンポーネントに対する変更が別のコンポーネントに影響する可能性があり、予期しない結果が生じることがあることを指摘していました。 例えば、私はストレージアレイの特にひどいIOボトルネックを解消するとCPU使用率が100%に跳ね上がり、ユーザーエクスペリエンスがさらに悪化した事例を見たことがあります。これは、システムが突然自由に作業量を増やせるようになったものの、ユーザー活動とスループットの増加に対応するためのCPUリソースがなかったために発生したものです。 新しいシステムを計画する際、サイジングモデルがパフォーマンスの低いハードウェアのパフォーマンスメトリックに基づいている場合はこのような影響を考慮する必要があります。 新しいプロセッサを搭載した新しいサーバーにアップグレードする場合でも、新しいプラットフォームでのIOレイテンシが低いために適切なサイズにする必要がある場合は、データベースVMの動作を注意深く監視する必要があります。

また、後述するように物理ホストのCPUリソースとメモリリソースをサイジングする際はソフトウェア定義のストレージIO処理も考慮する必要があります。

ストレージキャパシティプランニング

ストレージキャパシティプランニングを理解し、データベースの推奨事項を理解するには、まずvSANと従来のESXiストレージの基本的な違いを理解する必要があります。 最初にこれらの違いを説明し、次にCachéデータベースに関するすべてのベストプラクティスの推奨事項を詳しく説明します。

vSANストレージモデル

---

vSANおよびHCIでは、一般的にソフトウェア定義ストレージ(SDS)が重要な役割を果たしています。 データの保存方法と管理方法は、ESXiサーバーのクラスターと共有ストレージアレイを使用する場合とは大きく異なります。 HCIの利点の1つはLUNがないことです。その代わり、必要に応じてVMに割り当てられるストレージのプールがあり、VMDKごとに可用性、容量、およびパフォーマンスの機能を表すポリシーが適用されています。

例えば、パフォーマンスと可用性の要件に応じて、ディスクの数やタイプが異なるさまざまなサイズのディスクグループやディスクプールにまとめられた回転式ディスクのシェルフで構成された従来のストレージアレイを想像してください。 その後、ディスクグループは多数の論理ディスク(ストレージアレイボリュームまたはLUN)として表現され、データストアとしてESXiホストに提示され、VMFSボリュームとしてフォーマットされます。
VMはデータストア内のファイルとして表現されます。 可用性とパフォーマンスに関するデータベースのベストプラクティスでは、データベース(ランダムアクセス)、ジャーナル(シーケンシャル)、およびその他(バックアップや非本番システムなど)用に、少なくとも独立したディスクグループとLUNを使用することを推奨しています。

vSANの場合はそうではありません。vSANのストレージは、ストレージポリシーベースの管理(SPBM)を使用して割り当てられます。 ポリシーは、以下を含む機能を組み合わせて作成できます(ただし、これ以外の機能もあります)。

- 冗長なデータのコピー数を決める許容障害数(FTT)。
- 容量を節約するイレイジャーコーディング(RAID-5またはRAID-6)。
- パフォーマンスを向上させるディスクストライプ。
- シックまたはシンディスクプロビジョニング(vSANではデフォルトでシン)。
- その他...

VMDK(個々のVMディスク)は適切なポリシーを選択することにより、vSANストレージプールから作成されます。したがって、属性を設定してアレイ上にディスクグループとLUNを作成するのではなく、SPBMを使用してストレージの機能をvSANのポリシーとして定義します。例えば、「データベース」は「ジャーナル」やその他の必要なものとは異なります。 VM用のディスクを作成する際は容量を設定し、適切なポリシーを選択します。

もう一つ重要な概念があります。VMはVMDKデータストア上のファイルのセットではなく、ストレージオブジェクトのセットとして保存されます。 例えば、データベースVMはVMDK、スワップ、スナップショットなどを含む複数のオブジェクトとコンポーネントで構成されます。vSAN
SDSは選択したポリシーの要件を満たすため、すべてのオブジェクト配置メカニズムを管理します。

ストレージ階層とIOパフォーマンスのプランニング

高いパフォーマンスを確保するため、2つの階層のストレージがあります。

- キャッシュ層 - 高耐久性フラッシュである必要があります。
- キャパシティ層 - フラッシュ、またはハイブリッドの場合は回転式ディスクを使用します。

下の図に示すように、ストレージは複数の階層とディスクグループに分かれています。 vSAN 6.5では、各ディスクグループに単一のキャッシュデバイスと最大7台の回転式ディスクまたはフラッシュデバイスが含まれます。
最大5つのディスクグループを使用できるため、ホストごとに最大35台のデバイスを使用できます。 次の図は、4つのホストを持つオールフラッシュvSANクラスターを示しています。各ホストには2つのディスクグループがあり、それぞれに1台のNVMeキャッシュディスクと3台のSATAキャパシティディスクがあります。

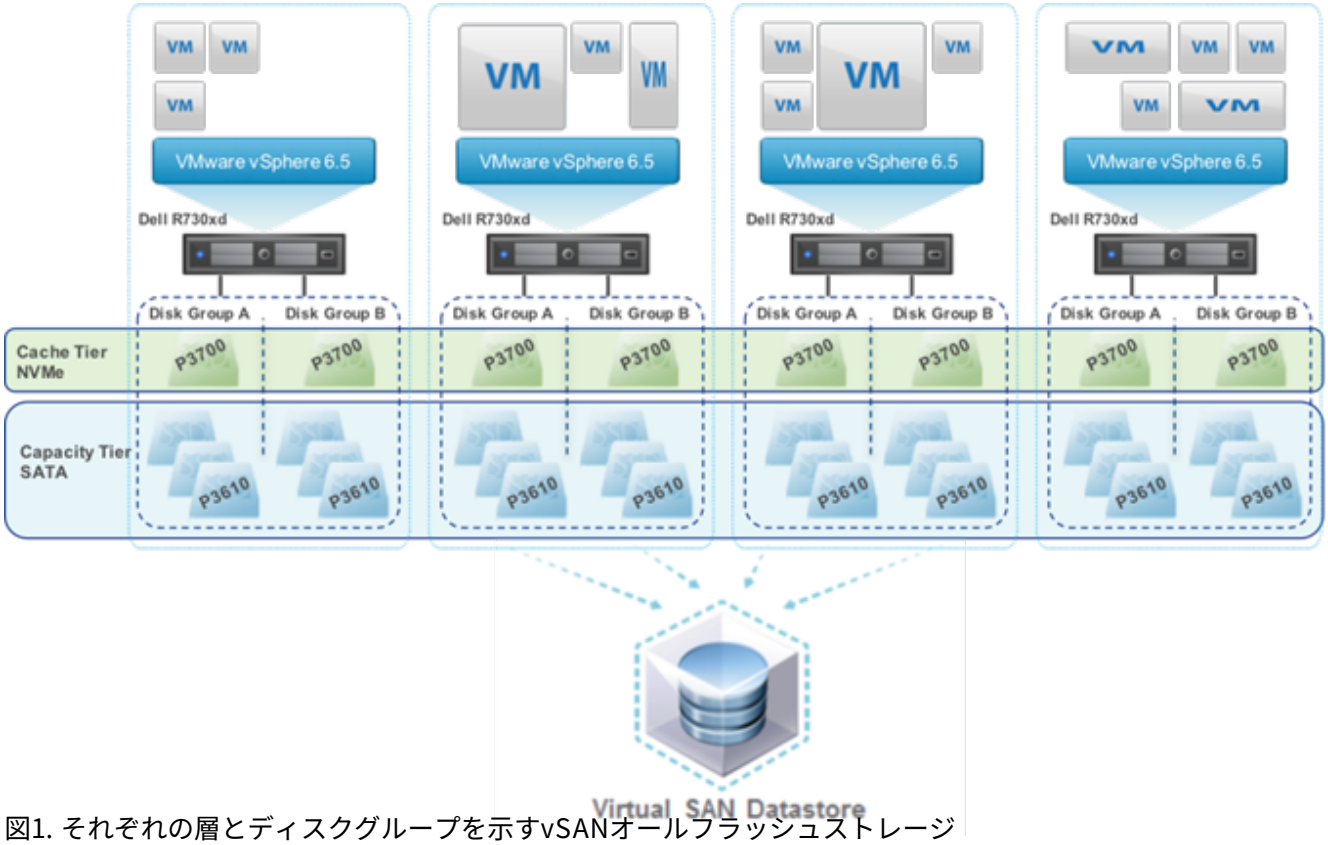

図1. それぞれの層とディスクグループを示すvSANオールフラッシュストレージ

それぞれの層の設定方法やキャッシュ層とキャパシティ層に使うフラッシュのタイプを検討する場合、IOパスを考慮する必要があります。レイテンシを最小にしてパフォーマンスを最大にするため、書き込みはキャッシュ層に移され、ソフトウェアがその書き込みをまとめてキャパシティ層に移します。 キャッシュの使用状況は導入モデルに依存します。例えばvSANハイブリッド構成の場合はキャッシュ層の30%が書き込みキャッシュですが、オールフラッシュの場合はキャッシュ層の100%が書き込みキャッシュで、読み込みは低レイテンシなフラッシュキャパシティ層から行われます。

オールフラッシュを使用すると、パフォーマンスが向上します。 大容量で耐久性のあるフラッシュドライブが利用できるようになった今、回転式ディスクが必要かどうかを検討する必要があります。 近年のビジネス事例では回転式ディスクの代わりにフラッシュが採用されており、はるかに低いIOPS単位のコスト、パフォーマンス(低レイテンシ)、高信頼性(可動部品が故障せず、必要なIOPSで故障するディスクが少ない)、低電力かつ低発熱なプロファイル、小さな設置面積などを特徴としています。 また、その他のHCI機能のメリットも得られます。例えばvSANでは、オールフラッシュ構成でのみ重複排除と圧縮が許可されます。

- 推奨: 最高のパフォーマンスとTCOの削減のため、オールフラッシュを検討してください。

最高のパフォーマンスを得るには、特にvSANの場合はディスクグループごとにキャッシュデバイスが1つしかないため、キャッシュ層のレイテンシを最低にする必要があります。

- 推奨: SASでも問題ありませんが、可能であればキャッシュ層にNVMe SSDを選択してください。
- 推奨:
  キャッシュ層で高耐久性フラッシュデバイスを選択し、高負荷なI/Oを処理するようにしてください。

キャパシティ層のSSDについては、SAS SSDとSATA SSDの性能の差はごくわずかです。
データベースアプリケーションについては、キャパシティ層でNVMe
SSDのコストを負担する必要はありません。
ただし、いかなる場合も、電源障害保護などの機能を備えたエンタープライズクラスのSATA
SSDを使用するようにしてください。

- 推奨: キャパシティ層には大容量のSATA SSDを選択してください。
- 推奨: 電源障害保護機能を備えたエンタープライズSSDを選択してください。

スケジュールによってはIOPSが高い3D
Xpointなどの新しいテクノロジーを使用し、レイテンシを下げ、容量を増やし、耐久性を高めることができます。
この記事の最後に、フラッシュストレージの構成を記しています。

- 推奨: キャッシュ層とキャパシティ層には、3D
  Xpointなどの新しいテクノロジーを組み込むことを検討してください。

前述したように、ホストごとに最大5つのディスクグループを作成でき、ディスクグループのキャパシティ層は1台のフラッシュデバイスと最大7台のデバイスで構成されます。 1台のフラッシュデバイスかつ必要な容量の単一のディスクグループ、またはホストごとに複数のディスクグループを作成できます。
ホストごとに複数のディスクグループを持たせると、次のようなメリットがあります。

- パフォーマンス：階層内に複数のフラッシュデバイスがあると、ホストごとのIOPSが増加します。
- 障害ドメイン：キャッシュディスクの障害はディスクグループ全体に影響しますが、vSANが自動的に再構築されるため、可用性は維持されます。

可用性、パフォーマンス、容量のバランスをとる必要がありますが、一般的にはホストごとに複数のディスクグループを用意することをお勧めします。

- 推奨:
  ストレージの要件を確認し、ホストごとに複数のディスクグループを用意することを検討してください。

どのようなパフォーマンスを期待できますか？

アプリケーションのユーザーエクスペリエンスを向上させるには、ストレージのレイテンシを下げることが重要です。通常
は、データベース
の読み取りIOのレイテンシを10ミリ
秒未満にすることが推奨されています。詳細については、この連載のパート6の表を参照してください。

既定のvSANストレージポリシーとCaché RANREADユーティリティ
を使用してCachéデータベースのワークロードをテストした結果、キャパシティ層でIntel S3610 SATA SSDを使用したオールフラッシュvSANのレイテンシは1ミリ秒未満で、3万IOPS超のランダムな読み取りIOが100%持続されることを確認しました。
Cachéデータベースが基本的に可能な限り多くのデータベースIOにメモリを使用する
ようにインスタンスをサイジングすることを考慮すれば、オールフラッシュのレイテンシとIOPS能力はほとんどのアプリケーションに十分な余裕を与えるものです。
メモリのアクセス時間は、NVMeフラッシュストレージよりも桁違いに短いことを覚えておいてください。

いつものことですが、何が最適なのかは状況によって違います。ストレージポリシー、ディスクグループの数、ディスクの数とタイプなどがパフォーマンスに影響するため、ご自身のシステムで検証してください！

キャパシティとパフォーマンスのプランニング

vSANストレージプールの物理容量(TB)は、キャパシティ層のディスクの合計サイズとして大まかに計算できます。 図1の構成例では、合計24個のINTEL S3610 1.6 TB SSDがあります。

クラスターの物理容量：24 x 1.6TB = 38.4 TB

ただし、利用可能な容量は選択する構成によって大きく異なり、計算が煩雑になります。例えば、使用されるポリ

シー(データのコピー数を指定するFTTなど)のほか、重複排除や圧縮が有効になっているかどうかによって決まります。

ここでは選択されたポリシーを段階的に追い、その容量とパフォーマンスへの影響とデータベースのワークロードに関する推奨事項について説明します。

私が目にするあらゆるESXiの導入環境は、複数のVMで構成されています。例えば、統合医療情報システムであるTrakCareはInterSystemsの医療情報プラットフォーム上に構築されており、HealthShareの中核には「ティア1ビジネスクリティカルアプリケーション」の説明に完全に適合する少なくとも1台の大規模(モンスター)データベースサーバーVMがあります。 ただし、導入環境には本番ウェブサーバー、プリントサーバーなど、単一の目的を持つ他のVMも混在しています。 テスト用、トレーニング用、および本番用ではないその他のVMもです。
通常、すべてが単一のESXiクラスターに導入されます。 ここではデータベースVMの要件に焦点を当てていますが、SPBMはすべてのVMに対してVMDKごとに調整できることを覚えておいてください。

重複排除と圧縮

vSANの場合、重複排除と圧縮はクラスター全体でオン/オフします。 重複排除と圧縮は、オールフラッシュ構成を使用している場合にのみ有効にできます。 両方の機能を同時に有効にできます。

一見すると、重複排除と圧縮は良い考えのように思えます。キャパシティ層で(より高価な)フラッシュデバイスを使用している場合は特に容量を節約したいものです。 重複排除と圧縮を有効にすると容量を節約できますが、大規模な本番データベースや常にデータが上書きされているクラスターではこの機能を有効にしないことをお勧めします。

重複排除と圧縮によってホストの処理負荷が1桁の%CPU使用率の範囲で増加する可能性がありますが、それはデータベースに推奨されない主な理由ではありません。

要約すると、vSANは4Kブロックを使用する単一ディスクグループ内のキャパシティ層にデータが書き込まれるときにデータの重複排除を試みます。 したがって、図1の例では重複排除するデータオブジェクトは、同じディスクグループのキャパシティ層に存在していなければなりません。 一意のポインタやコンテンツなどを含む8Kのデータベースブロックで埋められた基本的に巨大なファイルであるCachéデータベースファイルの容量が大幅に減るとは思えません。 また、vSANは重複ブロックの圧縮のみを試み、圧縮率が50%以上に達した場合にのみブロックが圧縮されたと見なします。 重複排除されたブロックが2Kに圧縮されない場合は、圧縮されずに書き込まれます。
オペレーティングシステムや他のファイルに重複がある場合がありますが、重複排除と圧縮の本当のメリットはVDI用に導入されたクラスターにあります。

また、重複排除と圧縮が有効になっている場合、ディスクグループ内の1台のデバイスの(まれではありますが)障害がグループ全体に影響を及ぼすことにも注意すべきです。 ディスクグループ全体が「異常」とマークされると、クラスター全体に影響があります。グループが異常とマークされると、ディスクグループ上のすべてのデータがそのグループから他の場所に退避され、その後はデバイスを交換しなければならず、vSANはリバランスするためにオブジェクトを再同期します。

- 推奨: データベースの導入では、圧縮と重複排除を有効にしないでください。

  補足：InterSystemsのデータベースミラーリングについて。

  最高の可用性を必要とするミッションクリティカルなティア1のCachéデータベースアプリケーションインスタンスの場合、
  仮想化されている場合でもInterSystemsの同期データベースミラーリングをお勧めします。
  仮想化ソリューションにはHAが組み込まれています。例えばVMWare
  HAの場合、ミラーリングを使用すると次のようなメリットもあります。

  - 最新データの独立したコピーが存在します。
  - 秒単位でフェイルオーバーできます(VMを再起動してからオペレーティングシステムを起動し、Caché をリカバリするよりも高速です)。
  - アプリケーション/Cachéに障害が発生した場合(VMwareでは検出されません)にフェイルオーバー

できます。

同じクラスターでデータベースをミラーリングしている場合に重複排除を有効にすると、問題があることに気付きましたか？ ミラーデータの重複排除を試すことは
一般的には賢明ではなく、処理のオーバーヘッドも発生します。

HCIでデータベースをミラーリングするかどうかを決定する際は、必要な合計ストレージ容量を考慮する必要があります。 vSANは可用性を確保するためにデータのコピーを複数作成します。このデータストレージもミラーリングによって複製されます。 ストレージの追加コストと、VMware HAによるわずかな稼働時間の増加を天秤に掛ける必要があります。

稼働時間を最大にするため、2つのクラスターを作成し、データベースミラーの各ノードを完全に独立した障害ドメインに配置できます。
ただし、このレベルの稼働時間を提供するには、サーバーとストレージの合計容量に注意してください。

暗号化

また、保管データの暗号化方法も考慮する必要があります。
IOスタックには、次のようないくつかの選択肢があります。

- Cachéのデータベース暗号化を使用する(データベースの暗号化のみ)。
- ストレージで暗号化する(SSDでのハードウェアディスク暗号化など)。

暗号化がパフォーマンスに与える影響はごくわずかですが、HCIで重複排除または圧縮を有効にすると容量に大きな影響を与える可能性があります。 重複排除や圧縮を選択した場合、暗号化されたデータは設計上ランダムであり、十分に圧縮されないため、Cachéのデータベース暗号化を使用することは望ましくありません。 保護対象の場所や回避したいリスク(ファイルの盗難とデバイスの盗難のどちらを危険視するかなど)を検討してください。

- 推奨: 最低限の暗号化を行うには、可能な限り最下層のIOスタックで暗号化してください。
  ただし、回避したいリスクが多くなるほど、スタックの階層は高くなります。

許容障害数(Failures To Tolerate, FTT)

FTT は、ストレージオブジェクトにクラスター内で少なくともn件のホスト、ネットワーク、またはディスクの障害が同時に発生してもオブジェクトの可用性を確保するようストレージに要件を設定します。 デフォルトは1(RAID-1)です。VMのストレージオブジェクト(VMDKなど)はESXiホスト間でミラーリングされます。

したがって、vSAN構成には少なくともn +
1個の複製(データのコピー)が含まれている必要があります。これは、クラスター内に2n +
1台のホストがあることも意味します。

例えば許容障害数が1のポリシーに従うには、たとえ1台のホストに障害が発生したとしても常に最低3台のホストを稼働させておく必要があります。 したがって、1台のホストがオフラインになったときのメンテナンスやその他の時間を考慮に入れるには、4台のホストが必要です。

- 推奨: vSANクラスターの場合、可用性を確保するには少なくとも4台のホストが必要です。

ただし、2台のホストと1台のリモート監視VMを想定したリモートオフィスブランチオフィス(ROBO)構成のような例外もあります。

イレイジャーコーディング

vSANのデフォルトのストレージ方式はRAID-1-データレプリケーションまたはミラーリングです。 イレイジャーコーディングは、ストレージオブジェクト/コンポーネントがクラスター内のストレージノードに分散されるRAID-5またはRAID-6です。 イレイジャーコーディングの主なメリットは、データ保護レベルを維持したまま容量効率を上げられることです。

前のセクションのFTTの計算を例として使用します。VMが2つの障害を許容する場合、RAID-1を使用するに、ストレージオブジェクトのコピーが3つ必要です。つまり、VMDKはベースVMDKのサイズの300%を消費します。RAID-6ではVMが2つの障害に耐えることができ、VMDKのサイズの150%しか消費しません。

ここでは、パフォーマンスと容量のどちらかを選択する必要があります。 容量を節約できるのは素晴らしいことですが、イレイジャーコーディングを有効にする前にデータベースのIOパターンを考慮する必要があります。 容量効率が向上する代わりに、I/O処理が増加することになります。コンポーネントの障害が発生している間はさらに負荷が高くなるため、最高のデータベースパフォーマンスを確保するにはRAID-1を使用してください。

- 推奨: 本番データベースではイレイジャーコーディングを有効にしないでください。
  本番環境以外で有効にしてください。

イレージャーコーディングはクラスターの必要なホスト数にも影響します。
例えばRAID-5の場合はクラスター内に最低4台のノードが必要で、RAID-6の場合は最低6台のノードが必要です。

- 推奨: イレイジャーコーディングの構成を計画する前に、追加ホストのコストを検討してください。

ストライピング

ストライピングはパフォーマンスを向上させるのに役立ちますが、役に立ちそうなのはハイブリッド構成だけだと思われます。

- 推奨: 本番データベースではストライピングを有効にしないでください。

オブジェクトスペースの予約(シンまたはシックプロビジョニング)

この設定の名前は、オブジェクトを使用してVM(VMDKなど)のコンポーネントを格納するvSANに由来しています。 デフォルトでは、VSANデータストアにプロビジョニングされるすべてのVMのオブジェクトスペースの予約は0%(シンプロビジョニング)になっています。この設定は容量を節約し、vSANのデータをより自由に配置できるようにします。 ただし、本番データベースでは予約値に100%(シックプロビジョニング)を使用し、作成時に容量を割り当てるのが最適です。 vSANの場合は、各ブロックへ初めて書き込まれるときに0が書き込まれるLazy Zeroedになります。 本番データベースの予約値に100%を選択する理由としては、データベースが拡張される際の遅延が少なくなり、必要なときにストレージを使用できることが保証されることが挙げられます。

- 推奨: 本番データベースのディスクの予約値には100%を使用してください。
- 推奨: 本番以外のインスタンスの場合、ストレージはシンプロビジョニングのままにしてください。

それぞれの機能をいつ有効化すべきですか？

通常はシステムをしばらく使用した後(システム上にアクティブなVMとユーザーが存在する状態)に可用性と容量節約の機能を有効化できます。 ただし、パフォーマンスと容量に影響します。
元のデータに加えてデータの複製が必要になるため、データを同期中は追加の容量が必要になります。 経験上、大規模なデータベースを使用するクラスターでこの種の機能を有効にすると、処理時間が非常に長くなり、可用性が低下する可能性があります。

- 推奨: 本番稼働を開始する前に、そして大規模なデータベースを読み込む前には必ず、事前に時間をかけて重複排除や圧縮などのストレージの機能を理解して構成するようにしてください。

ディスクバランシングや障害発生時用の空き領域を残すなど、他にも考慮事項があります。 つまり、この記事の推奨事項とベンダー固有の選択肢を考慮して物理ディスクの要件を把握する必要があります。

- 推奨: 多くの機能と置き換え可能な要素があります。 まずは合計GB容量の要件を見積もり、この記事の推奨事項を(アプリケーションベンダーと一緒に)確認してからHCIベンダーに相談してください。

ストレージ処理のオーバーヘッド

ホスト上でのストレージ処理のオーバーヘッドを考慮する必要があります。 かつてはエンタープライズストレージアレイのプロセッサが処理していたストレージの処理が、クラスター内の各ホストで実行されるようになっています。

各ホストのオーバーヘッドの大きさは、ワークロードと有効になっているストレージの機能によって決まります。
vSAN上のCachéで行ったテストの結果を見る限り、特に現在のサーバーで使用可能なコアの数を考慮すると、過度な処理要件はないと言えます。
VMwareはホストのCPU使用率が5〜10%になるよう計画することを推奨しています。

上記はサイジングを行う際にまず考慮すべき内容ですが、何が最適なのかは状況によって異なるため、確認が必要です。

- 推奨: CPU使用率が10%となる最悪のケースを想定し、実際のワークロードを監視してください。

ネットワーク

ベンダーの要件を確認してください。最小10GbEのNICを採用し、ストレージトラフィックや管理(例: vMotion)などに複数のNICを使うことを想定してください。 私は自分の苦い経験から、クラスターの最適な運用にはエンタープライズクラスのネットワークスイッチが必要であることを皆さんに伝えることができます。結局、可用性を確保するためにどんな書き込みもネットワーク経由で同期的に送信されるからです。

- 推奨: ストレージトラフィック用に最低10GbEの帯域幅を持つネットワークスイッチを使用してください。ベストプラクティスに従い、ホストごとに複数のNICを用意してください。

フラッシュストレージの概要

HCIにはフラッシュストレージが必須であるため、フラッシュストレージの現状と今後の展望を確認することをお勧めします。

HCIを使用するかどうかにかかわらず、現時点でフラッシュを搭載したストレージを使用してアプリケーションを導入していないのであれば、次に購入するストレージにはフラッシュが搭載されている可能性があります。

ストレージの現状と未来

一般的に導入されているストレージソリューションの機能を確認し、用語を明確にしましょう。

回転式ディスク

- 昔ながらの SASまたはSATAインターフェースを備えた7,200回転、10,000回転、または15,000回転の回転式物理ディスクです。 ディスクあたりのIOPSは低いです。
  このディスクは大容量にできますが、GBあたりのIOPSは減少します。 パフォーマンスを確保するため、通常は複数のディスクにデータを分散して「十分な」IOPSと大容量を実現します。

SSDディスク - SATAおよびSAS

- 現在、フラッシュは一般的にNANDフラッシュを使用する、SASまたはSATAインターフェースを持つSSDとして導入されています。
  また、SSDにはいくらかのDRAMが書き込み可能なバッファメモリとして搭載されています。 エンタープライズSSDには停電保護機能が搭載されており、停電時にはDRAMの内容がNANDに書き込まれます。

SSDディスク - NVMe

- SSDディスクと同様ですが、NVMeプロトコル(SASまたはSATAではない)とNANDフラッシュを使用します。 NVMeメディアはPCI Express(PCIe)バス経由で接続されるため、システムはホストバスアダプターやストレージファブリックのオーバーヘッドなしで直接通信でき、レイテンシが大幅に短縮されます。

ストレージアレイ

- エンタープライズアレイは保護機能と拡張機能を提供します。
  現在、ストレージの構成はハイブリッドアレイまたはオールフラッシュが一般的です。 ハイブリッドアレイにはNANDフラッシュのキャッシュ層のほか、7,200回転、10,000回転、または15,000回転の回転式ディスクを使用する1つ以上のキャパシティ層が含まれています。
  NVMeアレイも利用できるようになっています。

ブロックモードNVDIMM

- このデバイスは出荷が始まったばかりであり、非常に低いレイテンシが必要な場合に使用されます。
  NVDIMMはDDRメモリソケットに装着され、レイテンシは約30ナノ秒です。 現在は8GBモジュールで出荷されているため、レガシーなデータベースアプリケーションには使用されない可能性が高いですが、新しいスケールアウトアプリケーションではこのパフォーマンスを利用できます。

3D XPoint

これは未来の技術であり、2016年11月時点では利用できません。

- MicronおよびIntelによって開発されました。 また、**Optane**(Intel)および**QuantX**(Micron)としても知られています。
- 少なくとも2017年までは利用できませんが、NANDと比較して容量が大きく、IOPSが10倍以上、レイテンシが10倍以上低くなるため、非常に高い耐久性と一貫したパフォーマンスが約束されます。

- 最初はNVMeプロトコルが採用される予定です。

SSDデバイスの耐久性

キャッシュ層とキャパシティ層のドライブを選択する際は、SSDデバイスの耐久性を考慮することが重要です。
フラッシュストレージの寿命は有限です。 SSDフラッシュのセルは、決まった回数だけ削除および再書き込みできます(読み取りに関しては制限はありません)。 デバイスのファームウェアはSSDの寿命を最大化するため、ドライブ全体に書き込みを分散するように調整します。 また、エンタープライズSSDは通常見た目よりも実際にはフラッシュの容量が大きいため(オーバープロビジョニング)、800 GBのドライブには1
TBを超えるフラッシュが搭載されている場合があります。

ストレージベンダーと話し合うために注目すべき指標は、一定年数で保証されている1日あたりのドライブ全体の書き換え可能回数(Drive Writes Per Day, DWPD)です。 例えば、1 DWPD(5年対応)の800 GB SSDは、1日に800 GBを5年間書き込むことができます。 したがって、DWPD(および年数)が多いほど耐久性が高くなります。
測定基準の計算方法を変え、指定されたSSDデバイスをテラバイト書き込み量(Terabytes
Written, TBW)で表すこともできます。同じ例のTBWは、1,460 TB(800GB * 365日 * 5年)です。
どちらの方法でも、予想されるIOに基づいてSSDの寿命を知ることができます。

要約

この記事では、HCI、特にVMWare
vSANバージョン6.5を導入する際に考慮すべき最も重要な機能について説明しました。 説明していないvSANの機能もあります。ここで言及していない機能については、デフォルト値を使用すべきだと考えてください。
ただし、ご質問やご意見がありましたら、コメントセクションで回答いたします。

今後の投稿ではHCIに戻る予定です。HCIは確実に発展していくアーキテクチャであるため、HCIに導入するインターシステムズのお客様が増えることを期待しています。

#System Administration #Performance #Caché #InterSystems IRIS