テクニカル レポート

NetApp AFF AシリーズとCシリーズでのOracleデータベースのパフォーマンス

NetApp
Joe Carter / Jeffrey Steiner
2023年5月| TR-4969

重要

本レポートに指定された環境、構成、バージョンがお客様の環境に対応しているかどうかは、Interoperability Matrix Tool（IMT）を参照してください。

<<本レポートは機械翻訳による参考訳です。公式な内容はオリジナルである英語版をご確認ください。>>

目次

はじめに

NetApp® ONTAP®は、インライン圧縮、ハードウェアの無停止アップグレード、他社製ストレージ アレイからのLUNインポートなど、さまざまな機能を標準搭載した強力なデータ管理プラットフォームです。最大24ノードのクラスタ構成が可能で、Network File System（NFS）、Server Message Block（SMB；サーバメッセージブロック）、iSCSI、Fibre Channel（FC；ファイバチャネル）、Nonvolatile Memory Express（NVMe）の各プロトコルを通じてデータを同時に提供できます。また、NetApp Snapshot®テクノロジをベースに、何万ものオンライン バックアップや完全に動作可能なデータベース クローンを作成することもできます。

ONTAPの豊富な機能セットに加えて、ユーザにはデータベースのサイズ、パフォーマンス要件、データ保護のニーズなど、さまざまな要件が存在します。NetAppストレージは、VMware ESXの仮想環境で稼働する約6,000のデータベースから、996TBのシングル インスタンス データ ウェアハウス（規模は拡大中）まで、あらゆる環境に導入されています。

本ドキュメントでは、NetApp AFFストレージシステム（AシリーズとCシリーズの両方を含む）を使用したベアメタルデータベースのパフォーマンスについて説明し、2つのAFFオプションの最大値と実際の違いについて説明します。

詳細については、次のリソースを参照してください。

- TR-4591：『Database Data Protection』
- TR-4592：『Oracle on MetroCluster』
- TR-4534：『Migration of Oracle Databases to NetApp Storage Systems』

ONTAP AFFプラットフォーム

ONTAPは高度なデータ保護と管理の基盤をなすものですが、ただし、ONTAPはソフトウェアだけを指しているわけではありません。フラッシュメディア、回転式ドライブ、仮想ストレージなど、さまざまなストレージテクノロジに依存するONTAPハードウェアプラットフォームから選択できます。現在導入されているほぼすべてのデータベースがソリッドステートストレージでホストされており、その傾向は加速しています。

NetAppは、AシリーズとCシリーズの2つのAFFプラットフォームを提供しています。どちらもオールフラッシュのソリッドステートストレージソリューションですが、Aシリーズは超レイテンシの影響を受けやすいワークロードをターゲットとしており、Cシリーズはコストと容量の最適化を優先するソリューションをターゲットとしています。違いはほとんどすべてメディアです。TLCフラッシュメディアは、ソリッドステートドライブテクノロジのエンタープライズパフォーマンスマーケットリーダーとして台頭していますが、QLCのコストは大幅に削減されています。

AシリーズとCシリーズのシステムは同じコントローラをベースにしているため、CPU、ホスト、ネットワーク接続、RAMの観点から、ワークロードに合わせて解決策のサイズを最適化できます。AシリーズとCシリーズを同じクラスタ内に混在させることもできるため、階層化アーキテクチャを構築できます。最後に、AシリーズとCシリーズは同じNVRAMとWAFL®テクノロジを使用しています。このテクノロジでは、書き込みI/OがミラーリングされたNVRAMにコミットされ、フルRAIDストライプでメディアに書き込まれるため、書き込みレイテンシはマイクロ秒単位で測定されます。

データベースのストレージパフォーマンス

データベース用のプラットフォームを選択するための最も重要な要件は、実際のニーズを理解することです。オールフラッシュ解決策が手頃な価格になった瞬間に、多くのお客様が回転式メディアから100%ソリッドステートストレージに移行しましたが、すべてのお客様が明確なメリットを享受できるわけではありません。データベースの中には、そもそもレイテンシによる制約がなかったものもありました。帯域幅による制約もあり、ソリッドステートメディアと回転式メディアのパフォーマンスはほぼ同じでした。他のケースでは、データベースはストレージパフォーマンスにまったく制限されず、データベースクエリロジックやデータベースサーバのCPUリソースによって制限されていました。

次のセクションでは、AFFストレージプラットフォームを選択する際の考慮事項について、いくつか詳しく説明します。

読み取りレイテンシ-ストレージ

手頃な価格のオールフラッシュストレージが登場する以前は、ストレージのレイテンシが、通常どおりデータベースのパフォーマンスに関する第1位の問題とみなされていました。これには2つの理由があります。まず、回転式メディアからデータベースブロックを読み取るのに約8~10ミリ秒かかりました。データベースが1、000、000個々のブロックをシリアルに読み取る必要がある場合、それぞれ10ミリ秒の読み取りレイテンシが加算されて多くの時間がかかります。

回転式メディアのレイテンシが課題だった2つ目の理由は、1つのドライブで同時に処理できるI/O処理の最大数が原因でした。通常は毎秒120回程度であった。120 IOPSを超える処理を試みた結果、レイテンシが急増しました。当時の解決策は、ドライブI/Oを完全に回避するために、ストレージ解決策にドライブを追加するか、キャッシュを追加した大容量のコントローラを使用するしかありませんでした。

NetApp AFFストレージは、次の2つのレイテンシ制限に対応します。

- TLCメディアを搭載したAシリーズコントローラは、100 µ sに近い読み取りレイテンシを実現できます。
- QLCメディアを搭載したCシリーズコントローラでは、約2ミリ秒の読み取りレイテンシを実現できます。
- AシリーズとCシリーズのどちらのドライブも、回転式メディアに比べてIOPSが大幅に向上しています。その結果、レイテンシが向上するだけでなく、一貫した予測可能なレイテンシが向上します。

AシリーズとCシリーズのパフォーマンスに違いがあるかどうかは、ワークロードの種類によって異なります。データベースタスクの多くは、数十億個の読み取りを連続して実行する必要がありますが、データベースタスクの多くはユーザアクティビティによって実行されます。たとえば、日中に実行された数十万件の銀行取引をまとめたレポートを待っている場合は、Aシリーズコントローラの読み取りレイテンシが改善されることでメリットが得られます。データベースがオンライン注文エントリシステムをホストしている場合、レイテンシの影響を受けない可能性があります。エンドユーザーは、 **Submit** ボタンをクリックしてから **Order Accepted**という単語が表示されるまでに、わずか数ミリ秒の遅延に気付くことはありません。

読み取りレイテンシ-キャッシュ

IOPSに制限があり、回転式メディアI/Oのレイテンシが高いため、回転式ディスクをベースにストレージ解決策をサイジングする際には、コントローラのRAM容量が重要な考慮事項となることがよくあります。優れたパフォーマンスは、回転式メディアの比較的低いパフォーマンスを相殺するRAMに依存していました。

ほとんどのオールフラッシュストレージソリューションでは、ストレージからの読み取りI/Oのサービス時間はRAMと比較して同等であるため、コントローラのRAM容量が重要になることはほとんどありません。この場合、「サービス時間」とは、データベースがI/O処理を実行してから、そのI/Oへの応答を受信するまでの経過時間を意味します。RAMからブロックを読み取るのに必要な実際の時間は、NVMeドライブからブロックを読み取るのに必要な時間よりも明らかにはるかに短いですが、ホスト、プロトコル、ネットワークの各レイヤに必要な時間が含まれると、キャッシュ読み取りとドライブ読み取りのレイテンシは同等になります。

これは、QLCメディアのレイテンシが高いため、Cシリーズでは少し変化します。AシリーズおよびCシリーズのキャッシュで処理できる読み取りI/Oは、100 µ sに近いレイテンシで処理できますが、QLCではドライブ読み取りのレイテンシが著しく高くなります。

AシリーズとCシリーズに違いがあるかどうかは、選択したコントローラモデルによって異なる場合があります。多くの場合、データベースのワーキングセットは非常に小さくなります。たとえば、NetAppは、約250TBのデータベースを所有しているお客様を知っていますが、アクティブなデータベースは約1TBしかありません。作業セットはコントローラのRAMに格納されるため、A800とC800ではほぼ同等のパフォーマンスが得られます。

これは、AシリーズとCシリーズのどちらを選択する場合に重要な考慮事項です。データベースのワーキングセットのサイズはどれくらいですか。それを直接定量化するのは非常に難しいですが、時には推定することもできます。たとえば、サイズが50TBの5年間のデータを含むコールセンターデータベースでは、ほとんどのアクティビティが最近のお客様の連絡先からのものであるため、ワーキングセットは100GBしかありません。お客様が数か月以上前の請求書について問い合わせに来ることはめったにありません。その場合、必要なデータを取得する際に多少の遅延が発生しても、原因の問題が発生することはほとんどありません。

書き込みレイテンシ

高い書き込みレイテンシよりもデータベースのパフォーマンスに影響することはありません。データベースによって変更がコミットされるたびに、トランザクションログへの1つ以上の書き込みが完了し、ストレージシステムから確認応答を受け取る必要があります。

ただし、データベースが書き込みレイテンシによって制限されることはほとんどありません。主な理由は、ほとんどの最新のストレージシステムと同様に、NetAppアレイはバックエンドメディアに直接書き込みをコミットしないためです。ONTAPでは、インバウンドの書き込みはミラーされたNVRAMにジャーナルされ、ホストに確認応答されます。ドライブの更新は、書き込みプロセスのかなり後の段階で行われます。そのため、ONTAPは回転式ドライブストレージシステムを使用していても、書き込みレイテンシをマイクロ秒単位で達成できます。

さらに、ONTAP WAFLテクノロジは、競合する多くのストレージシステムの書き込みに影響するRAIDパリティの問題を回避します。WAFLを使用していない場合は、パリティが原因でRAID-4 / 5 / 6実装を使用する必要はありません。データベースの書き込みごとに、パリティを再計算するためにストレージからの複数の読み取りI/Oが必要になります。書き込みを完了するには追加の読み取りI/Oが必要なため、これはRAIDペナルティと呼ばれることもありました。WAFLのため、ONTAPにはこの制限はありません。インバウンド書き込みI/OはNVRAMにジャーナルされ、フルRAIDストライプに編成されて1つのユニットとして書き込むことができます。書き込みを完了するために読み取りは必要ありません。

AシリーズとCシリーズは、どちらも同じNVRAM書き込みテクノロジを搭載したONTAPとWAFLを使用しているため、超低レイテンシの書き込みI/Oを提供します。従来のデータベースに関するドキュメントや、RAID、メディア、およびデータベースストレージのサイジングに関するその他の要素の使用に関するガイドラインの多くは、NetAppストレージには適用されません。ONTAPは、このような推奨事項で解決しようとする問題に影響を受けません。

帯域幅

帯域幅を大量に消費するデータベースもあります。これは、データウェアハウスと呼ばれるデータベースや、バッチレポートなどのタスクでよく発生します。実際のI/Oパターンは多くの要因に左右されますが、大量のラージブロックシーケンシャルI/O処理が含まれます。このようなワークロードは大量のI/Oを必要としますが、実際に使用するメディアタイプによって違いが生じることはほとんどありません。これは、ラージブロックのシーケンシャルI/O処理は本質的に非常に効率的な処理であるためです。ONTAPストレージシステムでは、実行中のシーケンシャルI/O処理を検出し、ホストが要求を発行する前に、ドライブからの読み取りと必要なデータのアセンブルをプロアクティブに開始できます。また、データはより大きなブロックでより効率的に読み取られます。

したがって、これらのタイプのワークロードはレイテンシの影響を受けないため、AシリーズとCシリーズでほぼ同じように実行する必要があります。シーケンシャルI/Oのパフォーマンスによって制限されることがわかっているデータベースでは、通常、ホストの構成エラーまたはポートが最大回線速度に達したネットワーク自体の制限が原因で問題が発生します。

CPU

NetAppサポートセンターに報告されるデータベースパフォーマンスの問題のほとんどは、実際にはデータベースレベルの処理の結果です。ほとんどの場合、ルート原因はデータベースサーバ自体に対するCPUの計算処理です。たとえば、データベースの時間の85%がDB CPUの処理に費やされているとしたら、ストレージパフォーマンスの最適化を試みる価値はほとんどありません。改善が目立つことはほとんどありません。パフォーマンスを向上させるには、SQLクエリ自体を最適化して効率を高める必要があります。他のケースでは、データベースの競合によってパフォーマンスが制限されます。クエリは、他の処理によってデータのロックが解放されるのを待機しているために遅延されます。

Oracleバッファ キャッシュ

I/Oを最適化するよりも、I/Oを完全に回避することを推奨します。RAMのコストは継続的に低下しますが、多

くのDBAはOracleバッファキャッシュのサイズを大きくする機会を容易に得ておらず、実行されるストレージI/Oの量が削減されています。Aシリーズコントローラは高速ですが、NVMeドライブを使用した場合でも、達成可能な最高のレイテンシは約100 µ sです。多くの場合、サーバRAMへの小規模な投資（またはデータベースサーバにすでに存在している見落とされたRAMを使用）によって、これらの100 µ sのI/O処理が、ナノ秒単位のレイテンシでローカルメモリ読み取りに変換される可能性があります。

データベースキャッシュが増加すると、ストレージシステムに必要なIOPSも削減されます。サーバRAMへのわずかな投資で、より安価なストレージ解決策を使用できる可能性があります。

NetApp AFF AシリーズとCシリーズ

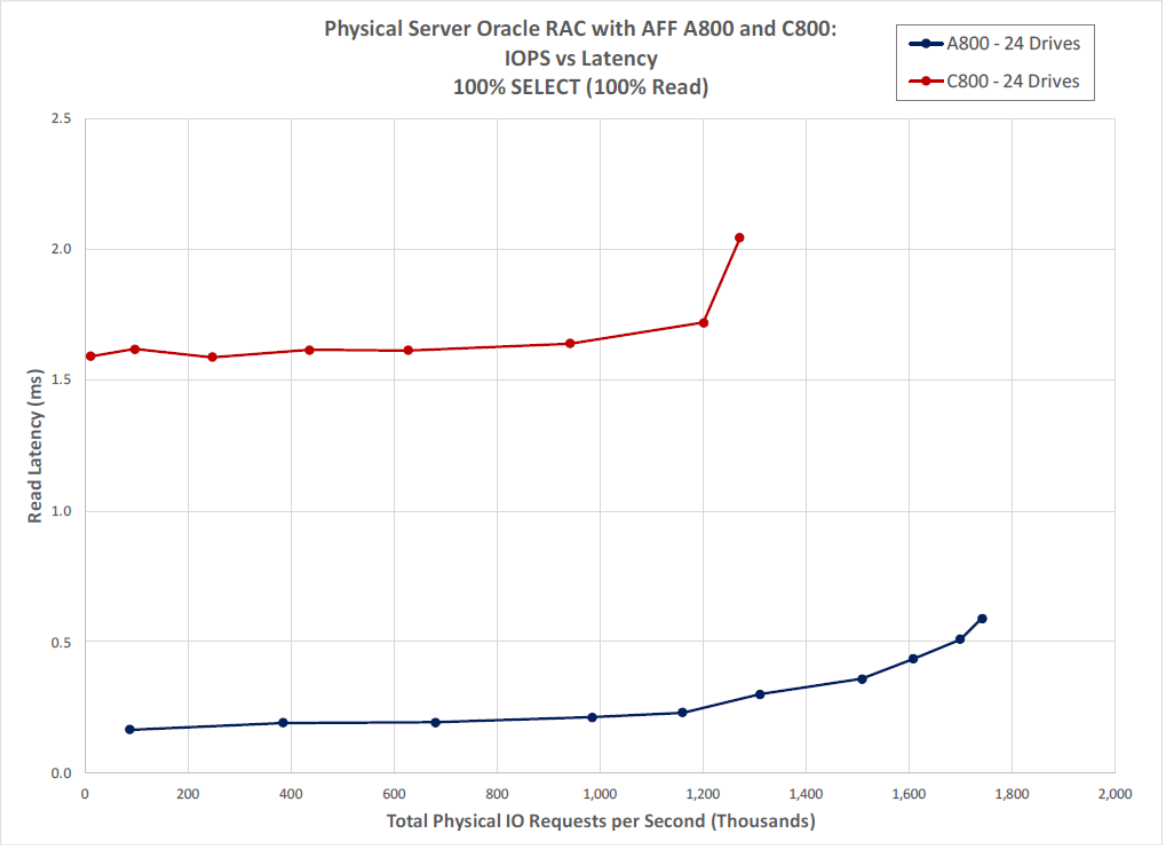
次のグラフは、さまざまな構成におけるAFFプラットフォームのパフォーマンス機能を示しています。このテストでは、4ノードのOracle 19C RACクラスタを使用してワークロードを生成し、代表的な2ノードのA800 HAペアと2ノードのC800 HAペアでテストを実施しました。

注：

- 今回のテストでは最先端のSANプロトコルであるNVMe/FCプロトコルを選択しました。従来のFCP SANでも同様のレイテンシが見られますが、最大IOPSはわずかに低下します。
- NFS、iSCSI、NVMe/TCPを使用した同様のテストでは、レイテンシが約100 µ s増加しています。Cシリーズコントローラでは一般にレイテンシが高くなるため、この増加はほとんど検出できません。Aシリーズでは100 µ sのレイテンシ増加が検出されますが、250 µ sはまだ十分に高速であり、ストレージがパフォーマンス制限になる可能性は低いため、ユーザが気づくことはほとんどありません。これらのプロトコルを使用した場合のIOPSの最大値は約30%低くなります。これは、ストレージシステムと接続されているホストのTCP/IPプロトコルのオーバーヘッドが原因です。この削減によって60万IOPSを達成しても、ほとんどのデータベースフットプリントで必要とされるIOPSをはるかに上回ります。また、NFSやIPネットワークの使用によるコストや管理性のメリットは、実際のパフォーマンスへの想定上の影響を上回ることがよくあります。

100%読み取り

図1）AシリーズとCシリーズの100%読み取り



このチャートの主なポイントは次のとおりです。

レイテンシ

Aシリーズのレイテンシは一貫して約250 µ sで、コントローラが飽和状態に近づくと500 µ sにしか達しません。これは、より高速なTLCメディアの結果です。お客様が必要とするデータベースのほとんどは、レイテンシのカットオフが1ミリ秒であることに基づいており、A800はパフォーマンス曲線全体でこの要件を満たしています。このため、NetApp Aシリーズコントローラでは通常、データベースパフォーマンスのボトルネックとしてストレージが消去されます。データベースは、データベースサーバ自体のクエリロジックとCPU処理によってほぼ完全に制限されます。

CシリーズのレイテンシはQLCメディアを使用したために高くなりますが、従来の回転式ドライブシステムのレイテンシよりもはるかに優れています。高度なデータベースの多くは、Aシリーズコントローラのパフォーマンス機能を必要としますが、すべてではありません。現在使用されている大規模データベースの多くは、オールフラッシュストレージが登場する15年前と同じ機能を実行しています。

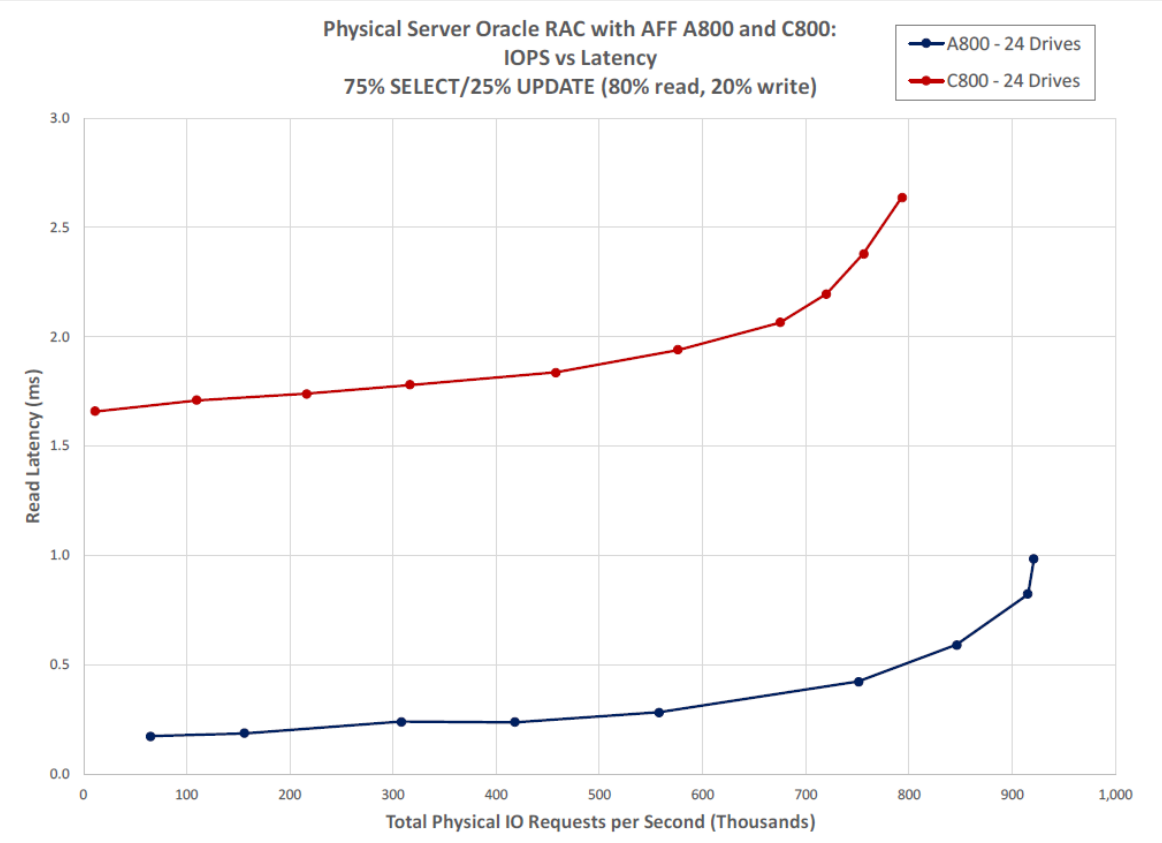
IOPS

グラフには、構成が飽和状態になるまでのパフォーマンス容量が表示されます。

Cシリーズ構成ではパフォーマンス制限がほとんどドライブであるため、限界点に徐 々 に達しています。レイテンシは、ドライブが制限に達するにつれて増加します。一方、Aシリーズの構成は通常、コントローラのCPUによって制限されます。通常、ドライブ自体は 、コントローラが抽出できるよりも多くのパフォーマンスを提供します。その結果、コントローラCPUの容量が100%に達するまで、一貫したパフォーマンスが実現します。

読み取り / 書き込み

図2）読み取り80%、書き込み20%



上のグラフは75% SELECTテストを示しており、その結果、読み取り比率は約80%になります。これは、25%の更新処理を使用したテストでは、更新される各ブロックの読み取りが作成され、読み取りの割合がわずかに増加するためです。

レイテンシ

レイテンシは100%読み取りテストよりも少し高くなっています。これは、書き込みI/Oによってコントローラに追加の負荷が発生しているためです。書き込みパスが同じであるため、AシリーズとCシリーズはどちらも書き込みI/Oの影響を受けます。ストレージOSは引き続きONTAPです。つまり、変更がNVRAMにコミットされ、RAMでステージングされてからドライブに書き込まれる場合は、同じ書き込みパスが使用されます。

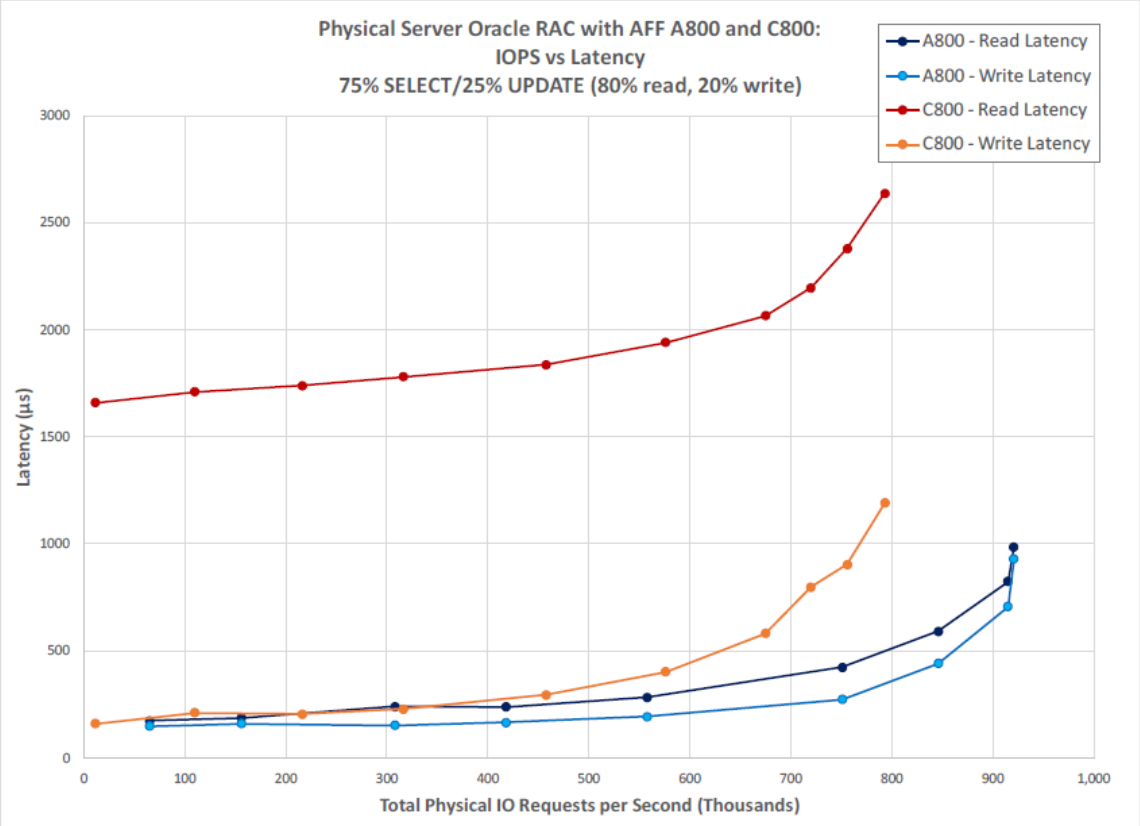
IOPS

どちらのプラットフォームでも、書き込みアクティビティが増加した結果、IOPSの最大値が同様に影響を受けます。書き込みI/Oは処理に多くのCPU処理を必要とし、読み取りI/Oと書き込みI/Oが混在している場合は、純粋な読み取りI/Oよりもドライブの処理速度がわずかに遅くなります。

書き込みレイテンシ

書き込みレイテンシは、データベースレベルでさまざまな方法で測定できます。ブロックサイズ、並列処理、および書き込みに使用されるシステムコールはさまざまです。NetAppは通常、 `log file parallel write` レイテンシに重点を置いています。これは、データベースのRedoロギング処理のストレージI/Oコンポーネントであり、通常はデータベースパフォーマンスの中で最もレイテンシの影響を受けやすい側面です。DBAは `log file sync` レイテンシに重点を置くこともありますが、これは完全にストレージ操作ではありません。`log file parallel write` イベントはトリガーされますが、REDOログのコミットの最終的なストレージI/Oコンポーネントを遅らせる可能性のあるその他のデータベースレベルの処理も含まれます。

図3）AシリーズとCシリーズの書き込みレイテンシの比較



前述のように、AシリーズとCシリーズの書き込みパスは基本的に同じです。AシリーズとCシリーズのどちらを選択した場合でも、I/OがNVRAMにジャーナルされると、ホスト側からは書き込み処理は完了します。Cシリーズの書き込みレイテンシは、全体的なI/O負荷が増大するにつれてわずかに増加します。これは、ドライブが一般的にビジー状態になったことによるものです。全体的に、書き込みパフォーマンスは各プラットフォームで同じです。

AシリーズとCシリーズの比較

データベースのサイジング作業で最も重要な質問は、「何が必要ですか？」です。

たとえば、ワークロードがIOPS機能やレイテンシのメリットに依存していなければ、コストをかける理由はありません。特定のテストで、どの構成でIOPSの可能性が最も高いことがわかるかを基準とするPOCは、特定のビジネスニーズに対する制限要因が純粋に物理IOPSであることがわかっていないかぎり、役に立ちません。また、純粋なIOPSテストでは、レイテンシなどのその他の重要な要素は無視されます。

すべてのワークロードを数値で定量化できるわけではありません。時々唯一の選択肢は、ビジネスニーズを理解するために時間をかけることです。たとえば、データベースを使用しているのは誰ですか。更新を行っているだけなのか、それとも数百万個のI/Oを必要とする大量のレポートを実行しているのか。

Oracleデータベースを使用する場合、ゴールドスタンダードはAWRレポートです。ほんの少しの時間で、`awrrpt.sql` I/O負荷が高いことがわかっている時間、ユーザからの苦情、重要なプロセスが実行されている時間の詳細なパフォーマンス内訳が生成されます。実際に必要なIOPSのレベル、現在のレイテンシ、そもそもストレージにパフォーマンスの問題があるかどうかがわかります。

ニーズが何であるかがわかったら、適切なコントローラを選択できます。

レイテンシの影響を受けやすいワークロード

読み取りレイテンシが重要な場合、Aシリーズコントローラは間違いなく最適なオプションです。AシリーズとCシリーズはどちらも同等の書き込みレイテンシを実現し、Redoロギングなどの重要なプロセスが常に最高のパフォーマンスで実行されるようにします。同様に、すでにRAMにキャッシュされているホットデータの読み取りでも、応答時間は非常に短くなります。読み取りI/O処理をドライブで処理する必要がある場合は、AシリーズとCシリーズの違いが関係します。

Aシリーズでの150マイクロ秒の読み取り処理とCシリーズでの2ミリ秒の読み取り処理の違いは、かなり大きいように思えるかもしれません。ただし、数テラバイトのストレージと高スループットを必要とし、ランダムリードでは8~12ミリ秒のレイテンシを許容する、多くのミッションクリティカルなデータベースで現在も使用されている回転式ディスクソリューションのレイテンシは、2ミリ秒に比べて大幅に向上します。

バッファキャッシュ

nec前述 のとおり、I/Oを最適化するよりも、I/Oを完全に回避することを推奨します。データベースのバッファキャッシュに割り当てられるRAMを増やすと、ストレージからの物理読み取りがキャッシュからの論理読み取りに変わるため、レイテンシが実質的に短縮されます。

一方、バッファキャッシュのサイズを大きくすると、 ストレージの平均レイテンシが長くなることがよくありますが、これはストレージシステム内のキャッシュヒットをデータベース内のキャッシュヒットに変換した結果です。低レイテンシのストレージキャッシュヒットが解消されると、ドライブアクセスを必要とするI/O処理だけが残ります。これにより平均レイテンシは高くなりますが、実行されるストレージI/O処理が少ないためパフォーマンスは向上しています。データベースがストレージI/Oに費やす時間が短縮されます。

Cシリーズコントローラで特定のデータベースのレイテンシが十分に低くない場合は、データベースのバッファキャッシュに割り当てられたRAMを増やすだけで効果的なレイテンシが向上し、Aシリーズシステムでデータベースを再ホストするよりも望ましい場合があります。

IOPS

必要なIOPSレベルは、ほとんどがストレージコントローラで制御されます。NetAppのアカウントチームとパートナーには、適切なコントローラの選択に役立つサイジングツールが用意されています。AFFシステムのドライブ数は、回転式メディアほど最大IOPSには影響しませんが、多少の影響があります。経験則として、AFFシステムにはHAペアあたり最低24本のドライブを搭載する必要があります（コントローラあたり12本）。

データ保持性

QLCドライブと他のフラッシュテクノロジの（レイテンシ特性に加えて）2つ目の違いは、摩耗容量です。多くのQLCドライブのメーカー仕様では、TLCドライブに比べて上書き機能が低下していますが、ONTAPストレージシステムへの影響は最小限に抑えられています。まず、ONTAP RAIDは、メディア障害からデータを保護し、ダブルパリティとトリプルパリティの両方のドライブオプションを備えています。さらに、ONTAP WAFLテクノロジは、インバウンドの書き込みデータを空きブロックに複数のドライブに分散します。これにより、ドライブ内の個々のセルの上書きが最小限に抑えられ、ドライブの耐用年数が最大になります。最後に、ドライブ障害に対応するNetAppサポート契約には、書き込みサイクルを使い果たしたSSDのドライブ交換も含まれます。

階層化

AシリーズとCシリーズのどちらを選択するかは、どちらか一方ではありません。たとえば、レイテンシの影響を受けやすいデータベース用にA800コントローラを2台、それ以外のデータベース用にC800コントローラを2台含む4ノードクラスタを構築できます。ニーズの変化に応じて、システムを停止することなく簡単に階層間でデータベースを移行できます。C250コントローラは、Oracleデータベースのバックアップなどのコールドデータ用に追加できます。

圧縮

AシリーズではTemperature-densitive Storage Efficiency（TSSE）を使用でき、Cシリーズでは常に有効になっています。この機能により、アクセス頻度の低いデータが検出され、より大きな圧縮ブロックで再圧縮されるため、ストレージ効率が向上します。これにより、ストレージ要件がさらに削減されます。実際には、フルテーブルスキャン、バックアップ、インデックスの再作成、アップグレード、およびデータベース内のほとんどまたはすべてのブロックに影響を与えるその他のアクティビティなど、通常十分なルーチンOracleアクティビティがあり、TSSEが最初から有効にならないようにします。

特定のデータベースボリュームセットに対して14日間の冷却ポリシーを持つTSSEが有効になっていて、データベースが30日間完全にシャットダウンされた場合、基盤となるブロックはTSSEによって100%再圧縮されます。その後データベースが再起動され、大量のランダムリード/ライトI/Oが発生すると、パフォーマンスが低下します。パフォーマンスは最終的にTSSE以前の通常のレベルに戻ります。

これは必ずしもレイテンシの増加が問題になるわけではありません。特に、ドライブのレイテンシがAシリーズのドライブよりもすでに高くなっているCシリーズシステムでは顕著です。さらに、コールドデータの削減量が増えることで、このような増加が相殺される可能性があります。

回転式アレイとハイブリッドアレイのアップグレード

従来の回転式ディスク解決策で同様のテストを実行した場合、平均レイテンシは約8ミリ秒から始まります。これは、電磁ドライブヘッドが適切な位置を探してブロックを転送するのに必要な時間であるためです。Cシリーズは、ほぼ4倍の性能向上を実現しています。

また、飽和点の問題により、回転式ディスクソリューションからCシリーズへの移行も促進されます。多くのワークロードはレイテンシの影響を受けませんが、高いIOPSレベルが必要です。回転式ディスクソリューション時代の一般的なハイパフォーマンスデータベースには、ドライブ1台あたり約120 IOPSしか処理できないため、最大1000本のドライブが使用されます。ドライブが1、000台の場合、約12万IOPSになります。

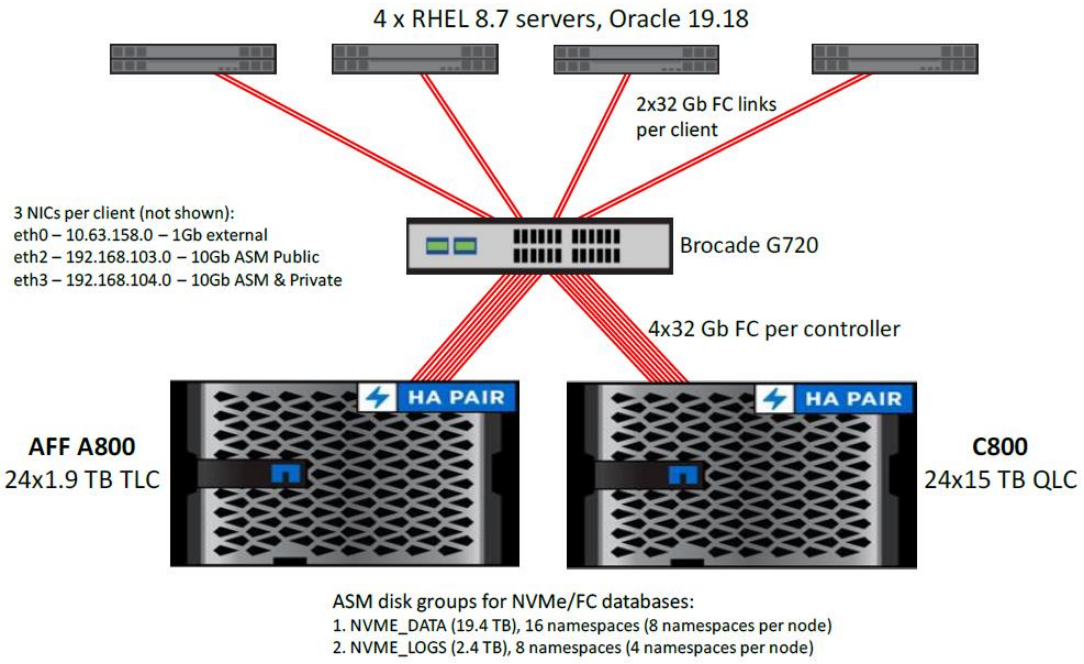
これはもはや実行不可能です。回転式ドライブの信頼性の低下、消費電力の増加、発熱量の増加を無視して、このようなソリューションを構築することは不可能です。回転式ディスク市場では、ドライブのサイズが膨大であるためです。ストレージソリューションでは、少数のドライブを使用した場合、バイト単位で最大容量に達します。IOPS制限は深刻になります。回転式ドライブが50本しかない解決策では、約6、000 IOPSしかサポートされません。

Cシリーズでは、AシリーズコントローラのTLCメディアにコストをかけることなく、この問題を解決します。消費電力と冷却の要件を大幅に抑えながら、設置面積を大幅に削減し、IOPS性能と信頼性を向上させることができます。これだけでも大きなメリットがありますが、レイテンシも4倍以上に向上します。

テスト構成

テストは次のように実行されました。

図4）OracleデータベースとAFFの構成



データベース サーバ

データベースサーバは、Red Hat Enterprise Linux 8.7で実行される4ノードクラスタとして構成しました。

Oracleのバージョン

テストには、Oracle 19.18 RAC Grid InfrastructureとOracle 19.18 Databaseを使用しました。

ネットワーク

FCネットワークは、各RACノードに2つの32Gb FC接続、各ストレージコントローラに4つの32Gb FC接続で構成しました。通常、コントローラはHAペア構成で導入されるため、Aシリーズシステムには合計8つのFC接続、Cシリーズコントローラには8つのFC接続を使用しました。テストはAシリーズとCシリーズのどちらかのシステムで（同時にはなく）実行され、それぞれのテストでRACクラスタからターゲットストレージシステムまで128GBの帯域幅を使用しました。

IPネットワークには10Gb NICを使用しましたが、ワークロードジェネレータで必要なRACノード間の通信は最小限で済み、IPネットワークは結果に影響しません。

Oracle ASM構成

このストレージは、NVMeを基盤とするOracle向けのNetAppのベストプラクティスに基づいて構成されており、単一のストレージシステムの総合的なパフォーマンス機能を使用することを目的としたデータベースが含まれています。

データファイル用のASMディスクグループを16個のネームスペースに、ログディスクグループを8個のネームスペースにそれぞれ構築しました。ASMディスクグループで推奨される最小ネームスペースは、ストレージ、ネットワーク、およびクライアントのNVMe接続のさまざまな側面から求められます。

単一のネームスペースで最大20万のデータベースIOPSをサポートできます。コントローラごとに8つのデータファイルネームスペースを使用するASMディスク構成を使用すると、1つのデータベースでコントローラの容量をほぼ100%まで増やすことができます。ASMディスクグループをこのレベル以上に増やすメリットはほとんどなく、システム上のストレージオブジェクトの数が不必要に増えて管理が複雑になります。

並列処理の要件が限られているため、制御ファイル、アーカイブログ、REDOログなど、ログIOに使用されるネームスペースの数はそれほど重要ではありません。コントローラあたり4つ以上のネームスペースを使用する必要があります。

各ネームスペースを専用ボリュームに配置しました。通常は、管理を簡易化するために、単一のASMディスクグループのネームスペースは1つのボリュームに同じ場所に配置されますが、ネームスペースを複数のボリュームに分けると、ONTAPレイヤでの並列処理が向上するため、書き込みレイテンシがわずかに改善されます。

Oracle SLOB

Silly Little Oracle Benchmark（SLOB）は、Oracleデータベースを使用したI/Oベンチマークのための主要なツールです。HammerDBやSwingbenchなどの他のツールは、セットアップが複雑であるか、サーバハードウェア、Oracleのバージョン、Oracleの構成に依存しています。SLOBは、データベースをドライブしてストレージI/Oを実行するのに理想的なツールです。その結果、完全なストレージパス、つまりデータベースのストレージ関連パラメータ、ネットワーク特性と制限、そしてもちろんストレージのIOPSと応答時間の能力が得られます。

すべてのテストは、非常に小さいOracleバッファキャッシュを使用して実行しました。これにより、サーバRAMのキャッシュヒットが最小限に抑えられ、アレイ上のストレージI/Oが最大化されます。

最後に、を使用してSLOBテーブルにデータを入力しました `OBFUSCATE_COLUMNS=TRUE`。この設定は、圧縮や重複排除などの効率化機能を備えたONTAPなどの最新のストレージシステムで正確な結果を得るために重要です。この設定を使用しないと、SLOBによって作成されるデータは、現実的な方法で圧縮や重複排除を行うことができなくなり、ストレージシステムに過剰なキャッシュが発生する可能性があります。これらのテストの目的は、Oracleデータベースとドライブ間の完全なコードパスを強調することであり、キャッシュされたIOパフォーマンスをテストすることではありません。

ワーキングセット

ワーキングセットは、頻繁にアクセスされるデータセットの部分です。ほとんどのOracleデータベースでは、ランダムI/Oがデータベースの合計サイズのごく一部に集中しています。競合他社から報告されたPOCとOracleのパフォーマンス結果の多くは、小規模なワーキングセットを使用しています。そのため、バックエンドドライブではなく、ほぼすべてのI/OをRAMから直接処理しています。これにより結果が歪むことがあります。このテストでは、実際のドライブ読み取りに関連するI/Oのほとんどが確実に行われるようにすることで、この歪みを回避したいと考えていました。

また、テスト中にSLOBユーザの数を変更することもありませんでした。理由を理解するには、SLOBユーザとSLOBスレッドの違いを理解する必要があります。SLOBテスト用のテーブルを作成する場合は、ユーザ数（スキーマとも呼ばれます）を指定する必要があります。サイズが1TBでユーザが16人のSLOBデータベースを作成すると、1TBのスペースのパーティションが16個作成され、各パーティションのサイズは64GBになります。単一のユーザまたはスキーマのみを使用したテストでは、64GBのスペースしか消費されません。

そのため、ユーザまたはスキーマの数を増やしてシステムの全体的なIOPSを増やすテストでは、作業セットのサイズも拡張されます。ユーザやスキーマの数が少ないテストでは、ストレージシステムのキャッシュヒットの割合がはるかに高くなるため、より現実的な構成よりもレイテンシが優れているように見えます。すべてのデータベースがキャッシュヒットの恩恵を受けますが、キャッシュに完全に収まるデータベースとしてA800/C800ほどの大容量のシステムを購入するお客様はほとんどいません。

より現実的なアプローチは、常にすべてのSLOBスキーマを使用し、スキーマごとにスレッドを変更することです。これはSLOBでサポートされています。

そのため、バランスのとれた構成を作成しました。SLOB用に10TBのテーブルを構築しました。SLOBは、ストレージコントローラで使用可能なRAMの約10倍のサイズです。SLOB scaleパラメータを655360Mに設定し、16のスキーマを使用しました。すべてのテストで16のスキーマすべてを使用し、10TBデータベース全体がアクセスされていることを確認しました。

曲線の生成

パフォーマンスグラフは、飽和点に達するまで負荷が増加した状態で20分間のテストを実行して作成しました。各テストを3回実行し、結果が一貫していることを検証しました。

データの選択

IOPSとレイテンシを基準にした結果が報告されました。SLOBスレッドは単なる負荷インジェクタであり、定義された数のSLOBスレッドによって生成されるIOPSは、RAC構成、Oracleのバージョン、およびその他の要因によって異なります。SLOBユーザを現実世界で役立つものに外挿する方法はありません。ただし重要なのは、IOPSとそれに伴うレイテンシです。

グラフのX軸は、データファイルの読み取りと書き込みの合計を示します。0%を超える書き込みを含むテストにもRedoロギングアクティビティがあります。Redoロギングアクティビティは大量のI/Oを生成する可能性がありますが、ブロックサイズとI/Oの非同期または同期の性質はどちらも大きく異なるため、IOPSで表すことは簡単ではありません。そのため、データファイルのランダムリードIOPSとデータファイルのランダムオーバーライトという最も要件の厳しいタイプのI/Oに注目しました。

最初のグラフのY軸はランダムリードのレイテンシを示しています。ランダムリードのレイテンシはデータベースの主なパフォーマンス制限であり、ストレージI/Oが特定のデータベースのパフォーマンス制限の原因であると仮定した場合、ランダムリードのレイテンシが最も重要な値です。この数値はOracle `db file sequential read` の統計から取ったものです。名前にもかかわらず、これはシーケンシャルI/Oではありません。インデックス付けされたブロックシーケンスに対するランダムI/O処理、またはブロック読み取りのシーケンスとして定義されることもあります。

また、一部のグラフでは、ランダムブロック上書きのレイテンシを設定しました。これはオラクルの `db file parallel write` 統計から取ったものですこの種のI/Oはほとんどの場合バックグラウンド処理であるため、データベースの制限要因になることはほとんどありません。データベースに対する変更はREDOログにコミットされ、データファイルはあとで更新されます。ランダムライトのレイテンシは、書き込みパスが同じであるため、AシリーズとCシリーズで書き込み動作がほぼ同じであることを実証するためだけに実施しました。書き込みはバックエンドディスクではなくNVRAMにコミットされます。

また、ストレージの書き込みレイテンシはこの数値よりも高くなっています。このタイプのI/Oは、通常、I/Oの非同期バッチとしてデータベースによって実行されます。各ブロックの個々のI/O処理は、その詳細レベルをデータベースレベルから測定できればレイテンシが低くなります。

詳細情報の入手方法

このドキュメントに記載されている情報の詳細については、以下のドキュメントやWebサイトを確認してください。

- NetApp Aシリーズ
  https://www.netapp.com/data-storage/aff-a-series/
- NetApp Cシリーズ
  https://www.netapp.com/data-storage/aff-c-series/
- NetAppの製品ドキュメント
  https://docs.netapp.com

バージョン履歴

オプションとして、NetApp Tableスタイルを使用し、バージョン履歴テーブルを作成することもできます。テーブル番号またはキャプションは追加しないでください。

バージョン	日付	ドキュメント バージョン履歴
バージョン1.0	2023年5月	初版リリース

バージョン	日付	ドキュメント バージョン履歴
バージョン1.0.1	2023年5月	マイナーな誤植修正

本ドキュメントに記載されている製品や機能のバージョンがお客様の環境でサポートされるかどうかについては、NetApp サポート サイトで Interoperability Matrix Tool (IMT)を参照してください。NetApp IMT には、NetApp がサポートする構成を構築するために使用できる製品コンポーネントやバージョンが定義されています。サポートの可否は、お客様の実際のインストール環境が公表されている仕様に従っているかどうかによって異なります。

機械翻訳に関する免責事項

原文は英語で作成されました。英語と日本語訳の間に不一致がある場合には、英語の内容が優先されます。公式な情報については、本資料の英語版を参照してください。翻訳によって生じた矛盾や不一致は、法令の順守や施行に対していかなる拘束力も法的な効力も持ちません。

著作権に関する情報

商標に関する情報