ホワイトペーパー
FUJITSU PRIMERGY サーバ
ディスク I/O パフォーマンスの基本

本書は、Fujitsu PRIMERGY サーバのディスク I/O パフォーマンスの担当者を対象としています。本書では、ディスク I/O パフォーマンスの測定方法や性能データについての情報を提供しています。お客様の要件に沿った、適切な内部ディスクサブシステムのサイジングおよび構成を決定する際の参考としてください。

バージョン

1.0
2011-05-09

目次

ドキュメントの履歴 2
ディスクサブシステムのパフォーマンス指標 3
パフォーマンスに影響を与える要因 3
　ブロックサイズ 3
　ディスクサブシステムへの同時アクセス 4
　オペレーティングシステムおよびアプリケーション 5
　コントローラー 5
　ストレージ媒体 6
ディスク I/O パフォーマンス測定 7
　Iometer 測定ツール 7
　ベンチマーク環境 8
　負荷プロファイル 8
　測定手順 9
　測定結果 10
ディスクサブシステムの分析 11
　計画 11
　パフォーマンス問題が発生した場合の分析 12
関連資料 15
お問い合わせ先 15

ドキュメントの履歴

バージョン 1.0

ディスクサブシステムのパフォーマンス指標

不揮発性ストレージであるハードディスクドライブと SSD（ソリッドステートドライブ）は、サーバ環境においては安全性とパフォーマンスが特に重要視されるコンポーネントです。これらのストレージは、プロセッサやメインメモリなどのサーバコンポーネントと比べて処理速度が非常に遅いため、ディスクサブシステムのサイジングと構成は特に重要になります。また、アプリケーションシナリオが多岐にわたるため、ディスクサブシステムの構成オプションは膨大な数になります。そのため、一つのパフォーマンス指標で、ディスクサブシステムのあらゆる側面を評価することはできません。I/O パフォーマンスの主要指標は次のとおりです。

- データスループット　　単位時間あたりのデータ転送量
- リクエスト数　　単位時間あたりの I/O オペレーション数（トランザクション）
- 平均応答時間　　リクエストの平均処理時間

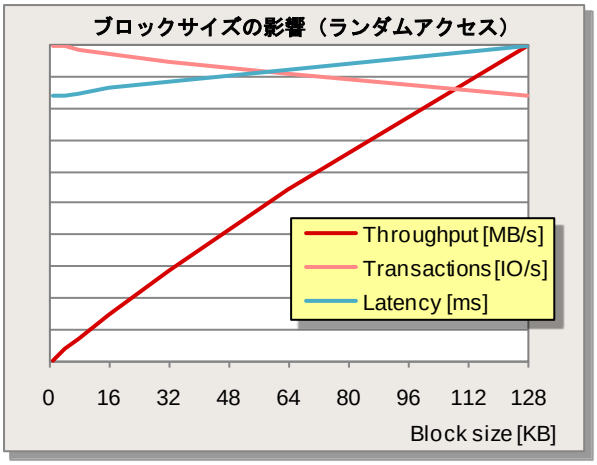
パフォーマンスに影響を与える要因

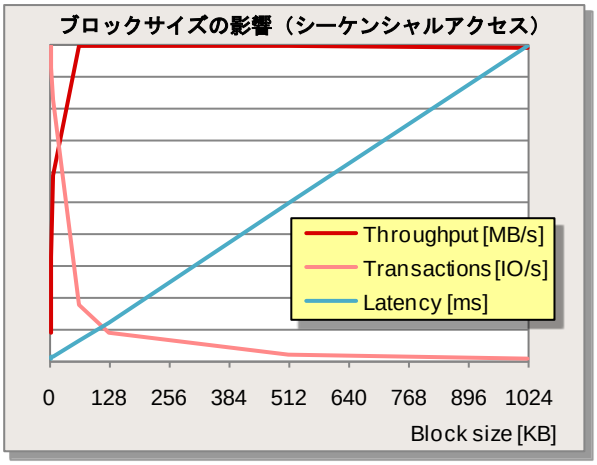
パフォーマンスに影響する要因は、次の 5 種類に分類できます。

- ブロックサイズ
- ディスクサブシステムへの同時アクセス
- オペレーティングシステムおよびアプリケーション
- コントローラー
- ストレージ媒体

ブロックサイズ

ディスクサブシステムにアクセスする際のデータ転送は、常にブロック単位で行われます。データ転送時のブロックサイズは、オペレーティングシステムやアプリケーションによって決まっており、ユーザーが調整することはできません。





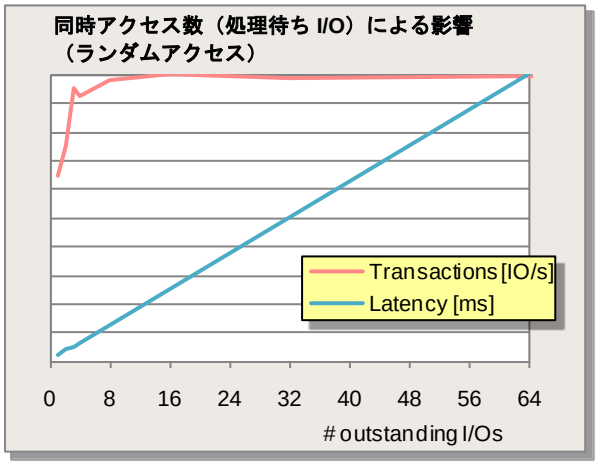
上記の左側のグラフは、ランダムアクセス時の測定結果です。スループットおよび応答時間（遅延）はブロックサイズの増加に伴って直線的に増加し、トランザクション数は逆に減少しています。一般的に、ディスクサブシステムの理論上での最大スループットは、ここでは達成されません。右側のグラフは、シーケンシャルアクセス時の測定結果です。ここでは、ブロックサイズの増加に伴って直線的に増加するのは応答時間（遅延）だけです。スループットは、最初はブロックサイズの増加に伴って増加しますが、ブロックサイズ 64 KB でディスクサブシステムの理論上の最大値に達し、その後はそのままで推移します。以上のことから、スループットは、アプリケーションのディスクアクセスパターンに大きく依存するといえます。

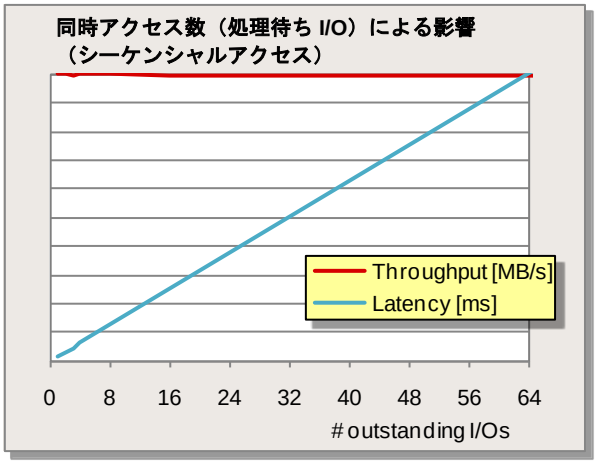
なお、各種アプリケーションの典型的なアクセスパターンは次のとおりです。

アプリケーション	アクセスパターン
オペレーティングシステム	ランダム、40 %リード、60 %ライト、ブロック ≥ 4 KB
ファイルコピー（SMB）	ランダム、50 %リード、50 %ライト、64 KB ブロック
ファイルサーバ（SMB）	ランダム、67 %リード、33 %ライト、64 KB ブロック
メールサーバ	ランダム、67 %リード、33 %ライト、8 KB ブロック
データベース（トランザクション処理中）	ランダム、67 %リード、33 %ライト、8 KB ブロック
Web サーバ	ランダム、100 %ライト、64 KB ブロック
データベース（ログファイル）	シーケンシャル、100 %ライト、64 KB ブロック
バックアップ	シーケンシャル、100 %ライト、64 KB ブロック
リストア	シーケンシャル、100 %ライト、64 KB ブロック
ビデオストリーミング	シーケンシャル、100 %リード、ブロック ≥ 64 KB

ディスクサブシステムへの同時アクセス

通常、サーバへは同時に多数のクライアントがアクセスします。また、クライアントから 1 台のサーバに、応答を待たずに複数のリクエストを送信することもあります。その結果、1 基のコントローラーやストレージに対して、同時アクセスが発生します。このため、多くのコントローラーおよびストレージは、待ち行列（キューイング）の機能を備えています。これにより、特定の条件下では、同時アクセスを処理する際に、より少数の同時アクセスや単一アクセスを処理するときよりも高いパフォーマンスが得られます。しかしその代わり、応答時間はより長くなります。多数の同時アクセスにより応答時間が長くなり、スループットの限界に達した場合は、ディスクサブシステムは過負荷ということになります。





上の 2 つのグラフは、ランダムアクセスおよびシーケンシャルアクセスのパフォーマンスの例を示しています。どちらの場合もブロックサイズは一定で、同時アクセス数（処理待ち I/O）を 1～64 の範囲で順に増やしています。上記のグラフでは、トランザクション数とスループットは一致しています。ブロックサイズが一定であるため、トランザクション数とスループットの比率は（同時アクセス数にかかわらず）一定だからです。ランダムアクセス（左側のグラフ）の場合、スループットは同時アクセス数の増加に伴ってすぐに最高値に達し、そのまま維持されます。シーケンシャルアクセス（右側のグラフ）では、同時アクセス数に関係なく、常に理論的な最大値に近いスループットに達しています。応答時間（遅延）は、ランダムアクセス、シーケンシャルアクセスのどちらでも、同時アクセス数の増加に伴って直線的に増加します。このため、シーケンシャルアクセスの場合は、ディスクサブシステムを拡張して安全に応答時間を短縮できますが、ランダムアクセスの場合は、スループットのパフォーマンスに注意しなければなりません。

また、サーバの応答速度を無視できないアプリケーションシナリオでは、スループットと応答時間のどちらを最適化するか選択する必要があります。その上で、個々の要件に従って同時アクセスを処理できるように、サーバのサイジングと構成を行ってください。

オペレーティングシステムおよびアプリケーション

アプリケーションによる大容量ストレージシステムへのアクセスパターンは、ディスクサブシステムのパフォーマンスに大きく影響します。また、オペレーティングシステムによるパフォーマンスへの影響として、仮想化層、I/O スケジューリング機能、ファイルシステム、ファイルキャッシュ、ストレージの構成（パーティショニングやソフトウェア RAID など）などがあります。

コントローラー

ソフトウェア RAID を使用している場合を除き、ストレージのコントローラーはスループットのパフォーマンスに大きく影響します。サーバシステムでは、オンボードコントローラーの他に、内部または外部のストレージに接続するためのさまざまな拡張コントローラーを使用できます。ただし、コントローラーに接続できるストレージの数には制限があります。制限以上のストレージを接続すると、コントローラーはパフォーマンスを阻害する要因となります。

RAID（アレイ）と JBOD

ハードディスクは、コンピュータシステムで最もエラーが発生しやすいコンポーネントです。そのためサーバシステムでは、ハードディスクの故障によるデータの損失を防ぐため、RAID コントローラーを使用します。RAID コントローラーは、複数のハードディスクを組み合わせて RAID（Redundant Array of Independent Disks、アレイ）を構成し、1 台のハードディスクが故障してもすべてのデータを復元できるように、複数のハードディスクにデータを分散して保存します。ただし、JBOD（Just a Bunch of Disks）と RAID 0 は例外で、これらは、複数のハードディスクを組み合わせて構成しますが、冗長性はありません。複数のハードディスクによる一般的な構成の方法は、JBOD、RAID 0、RAID 1、RAID 5、RAID 6、RAID 10、RAID 50、RAID 60 です。アレイの種類や、アレイを構成するストレージの数は、ディスクサブシステムのパフォーマンスに大きく影響します。

LUN（論理ユニット番号）

LUN は「Logical Unit Number（論理ユニット番号）」のことで、元々は SCSI ハードディスクの識別番号として使用されていたものです。オペレーティングシステムの観点では、LUN は通常 1 台の仮想的なハードディスクを指します。この仮想ハードディスクは、物理的なハードディスクと一致する場合もあれば、ハードディスクアレイ（JBOD や RAID）を指す場合もあります。

ストライプサイズ

アレイでは、データは「チャンク」と呼ばれる断片に分割され、複数のストレージに適切に分散して保存されます。各ストレージに分散して格納されたチャンクの全構成を、ストライプセットといいます。ストライプセットからパリティチャンクを除いた容量を、ストライプサイズといいます。このストライプサイズは、アレイの作成時に指定する必要があり、スループットおよび応答時間の両方に影響します。

キャッシュ

多くのコントローラーにはキャッシュがあり、主に次の 3 つの要因によってスループットに影響することがあります。これらの要因は、多くの場合、ストレージの使用時に個別に調整可能です。

- ライトデータのキャッシュ。ライトデータをキャッシュに一時保存すると、ユーザーに対するデータ書き込み終了のレスポンスは速くなりますが、実際にはデータはまだストレージに格納されていません。実際の書き込み処理は、後でまとめて実行されます。この方法により、コントローラーのリソース利用が最適化され、ライトリクエストの処理が速くなり、スループットが向上します。なお、オプションのバッテリーバックアップユニット（BBU）を使うことで、システム停電時のデータ破損を防止できます。

- 純粋にシーケンシャルなリードアクセスを行うアプリケーションシナリオでのリードデータのキャッシュ。一部のコントローラーでは、完全なシーケンシャルリードでないアクセスにも有効です。
- リクエストキューの設定。複数のリクエストを最も効率のよい順序に並び替えることで、ハードディスクのリード／ライトヘッドの動きを最適化できます。ただし、そのためには、キュー（待ち行列）を形成できるだけの十分なリクエストがコントローラーに送信されていることが必要です。

ストレージ媒体

ストレージの種別は、パフォーマンスに大きく影響します。ストレージには、回転磁気ストレージであるハードディスクと、記憶装置に半導体メモリを使用し非常にパフォーマンスが高い SSD（ソリッドステートドライブ）があり、それぞれ異なる特徴を持っています。パフォーマンスについては、SSD がハードディスクの数倍優れています。しかし、SSD はハードディスクに比べて寿命が短く、非常に高価です。また、ハードディスクとは異なり、SSD では空のメモリセルに書き込む際に比べて、既存のメモリコンテンツに上書きする際にパフォーマンスが低下します。上書きする場合は、まず古いデータを削除する必要があるからです。そのため、データ書き換えの頻度が上がると、SSD の書き込み速度は急速に低下します。とはいえ一般的には、パフォーマンスについても、SSD の方がハードディスクより優れています。

ストレージの転送プロトコルやキャッシュも、パフォーマンスにおいて重要な役割を果たします。

- ハードディスクインターフェースの最大転送速度：
  SATA 3.0 Gbit/s　相方向で 286 MB/s の実効スループット
  SAS：　相方向で 286 MB/s の実効スループット
  SAS II：　相方向で 572 MB/s の実効スループット
- キャッシュ
  次の 2 つの要因が、パフォーマンスに影響します。
  - リクエストキューの設定。複数のリクエストを最も効率のよい順序に並び替えることで、ハードディスクのリード／ライトヘッドの動きを最適化できます。ただし、そのためには、ハードディスクキャッシュを有効にする必要があります。また、キュー（待ち行列）を作れるだけの十分なリクエストがハードディスクに送信されていることが必要です。
  - データのキャッシュ：
    通常、リードリクエストがあると、対象のセクターだけでなく、同一トラック上にある続きのセクターのデータも読み出されます。これらのデータは、リクエストされる場合に備えて、キャッシュに一時保存されます。また、ライトリクエストをハードディスクキャッシュに一時保存することでパフォーマンスを上げることもできます。リードリクエストはアプリケーションが待機しているためできるだけ速く処理する必要がありますが、ライトリクエストは通常、少し後で処理しても問題ないからです。これは、SSD でも同様です。

ハードディスクの回転速度、および、データ領域のサイズを指定している場合はディスク容量も、パフォーマンスに影響します。

- 回転速度：
  回転速度が速くなるほど、リード／ライトヘッドのアクセスも高速になります。SATA ハードディスクの回転速度は、5400 rpm または 7200 rpm です。SAS ハードディスクではより速く、回転速度は 10000 rpm または 15000 rpm です。
- 容量：
  ハードディスクでは、1 分あたりの回転数と記録密度は、円盤状のディスク全体にわたって一定です。つまり、トラックあたりのデータ量は、内側から外側に向かって増加します。そのため、アクセス速度は最外周で最高速になります。データ領域として指定したサイズは、大容量ハードディスクのより外側に向かって確保されるため、ハードディスク容量はパフォーマンスに大きく影響します。

ディスク I/O パフォーマンス測定

富士通では、すべての PRIMERGY サーバに対して、PRIMERGY Performance Lab でディスク I/O パフォーマンス測定を行っています。アプリケーションのベンチマークとは異なり、ディスク I/O パフォーマンスでは、通常、サーバ全体ではなく、ディスクサブシステム（ストレージとそのコントローラー）のみのパフォーマンスを測定します。そのため、プロセッサやメインメモリなどのサーバコンポーネントが測定時のボトルネックにならないように考慮して、ディスクサブシステムのサイズを決定します。なお、十分な容量のディスクサブシステムを使用して、サーバ構成全体の最大スループット性能を測定することも十分可能ですが、それは本書で説明するディスク I/O パフォーマンス測定の目標ではありません。測定結果は、PRIMERGY サーバのパフォーマンスレポートに記載されています。レポートは次のリンクから入手できます。http://ts.fujitsu.com/products/standard_servers/primergy_bov.html

Iometer 測定ツール

PRIMERGY Performance Lab では、Iometer という、Intel 社によって開発されたツールを使って、ディスク I/O パフォーマンスを測定しています。2001 年末以降、Iometer は http://SourceForge.net のプロジェクトとなり、さまざまなプラットフォームに移植され、国際的な開発者グループによって強化されています。Iometer は、Windows 用のユーザーインターフェースと、各種プラットフォームで利用できる「dynamo」というコンポーネントで構成されています。これら 2 つのコンポーネントは、http://www.iometer.org/ または http://sourceforge.net/projects/iometer から「インテルオープンソースライセンス」でダウンロードできます。

Iometer では、豊富なパラメーターにより詳細な設定が可能なため、ディスクサブシステムへのアクセスについて実際のアプリケーションの動作を再現できます。初めに、測定中にアクセスするデータ領域を定義します。データ領域を作成するには、次のパラメーターを使用します。

- Maximum Disk Size
- Starting Disk Size

詳細なアクセスシナリオを定義するには、次のパラメーターを使用します。

- # of Worker Threads
- # of Outstanding I/Os
- Test Connection Rate
- Transfer Request Size (block size)
- Percent of Access Specification
- Percent Read/Write Distribution
- Percent Random/Sequential Distribution
- Transfer Delay
- Burst Length
- Align I/Os
- Reply Size

このように、使用するブロックサイズや同時アクセス数、シーケンシャルリード／ライト、ランダムリード／ライト、およびこれらの組み合わせなどを設定することで、さまざまなアプリケーションシナリオを再現できます。

測定結果は、アクセスパターンごとに、CSV（カンマ区切り）ファイルに出力されます。主要な指標は次のとおりです。

- 1 秒あたりのスループット
- 1 秒あたりのトランザクション数
- 平均応答時間

この方法により、特定のアクセスパターンを使ってさまざまなディスクサブシステムのパフォーマンスを比較できます。Iometer は、ファイルシステムを使用してディスクサブシステムにアクセスできるばかりでなく、いわゆる RAW デバイスにもアクセスできます。ただし、どちらの場合も、オペレーティングシステムのキャッシュは考慮されません。また、オペレーションは単一のテストファイルに対してブロック単位で行われます。

PRIMERGY Performance Lab では、標準で Iometer「dynamo」の Windows バージョンを使用します。これには、データ領域、実際に発生する負荷プロファイルの記録、およびディスク I/O パフォーマンス測定の測定シナリオが定義されています。これらの定義が測定結果の再現性の基盤となっているため、さまざまなディスクサブシステムのパフォーマンスを客観的に比較できます。

ベンチマーク環境

PRIMERGY サーバのディスク I/O パフォーマンス測定は、内部ディスクサブシステム、およびストレージブレード（ブレードサーバの場合）を対象として行います。測定前に、まず RAID アレイの初期化を行います。測定時のオペレーティングシステムは、通常、Windows Server 2008 Enterprise Edition を使用します。測定するストレージは、NTFS（クイックフォーマットなし、圧縮なし）でフォーマットします。他のファイルシステムや RAW デバイスを使用した方がパフォーマンスが高い場合でも、このようにフォーマットします。また、測定するドライブについて、ドライブのプロパティで「検索を速くするため、このドライブにインデックスを付ける」を無効にします。測定ファイルの数は、仮想的なハードディスクの数に対応しています。したがって、オペレーションは通常、単一の測定ファイルで実行されます。測定ファイルのサイズは、アレイを構成するストレージの数に対応して変えていきます（ストレージの容量には関連しません）。

データメディアの数	測定ファイルのサイズ
1～8	32 GB
9～16	64 GB
17～24	96 GB

負荷プロファイル

ディスク I/O パフォーマンス測定の負荷プロファイルには、標準で、次のような大容量ストレージへのさまざまなアクセスパターンを使用します。

アクセス方法	アクセスの種類		転送リクエストのサイズ [KB]（ブロックサイズ）	同時アクセス数（処理待ち I/O）
	リード	ライト		
シーケンシャル	100 %	0 %	1、4、8、64、128、512、1024	1、3、8、16、32、64、128、256、512
シーケンシャル	0 %	100 %	1、4、8、64、128、512、1024	1、3、8、16、32、64、128、256、512
ランダム	100 %	0 %	1、4、8、64、256、1024	1、3、8、16、32、64、128、256、512
ランダム	0 %	100 %	1、4、8、64、256、1024	1、3、8、16、32、64、128、256、512
ランダム	67 %	33 %	1、4、8、16、32、64、128	1、3、8、16、32、64、128、256、512
ランダム	50 %	50 %	64	1、3、8、16、32、64、128、256、512

また、どの負荷プロファイルを使用する場合でも、1 基のコントローラーを測定する際は、次の標準設定を適用します。

- # of Worker Threads＝1
- Test Connection Rate＝off
- Transfer Delay＝0
- Burst Length＝1
- Align I/Os＝Sector Boundaries
- Reply Size＝No Reply

これらの負荷プロファイルの一部は、典型的なアプリケーションによる負荷プロファイルに相当します。

標準負荷プロファイル	アクセス方法	アクセスの種類		ブロックサイズ [KB]	アプリケーション
		リード	ライト		
ファイルコピー	ランダム	50 %	50 %	64	ファイルのコピー
ファイルサーバ	ランダム	67 %	33 %	64	ファイルサーバ
データベース	ランダム	67 %	33 %	8	データベース（データ転送） メールサーバ
ストリーミング	シーケンシャル	100 %	0 %	64	データベース（ログファイル） データバックアップ ビデオストリーミング（一部）
リストア	シーケンシャル	0 %	100 %	64	ファイルのリストア

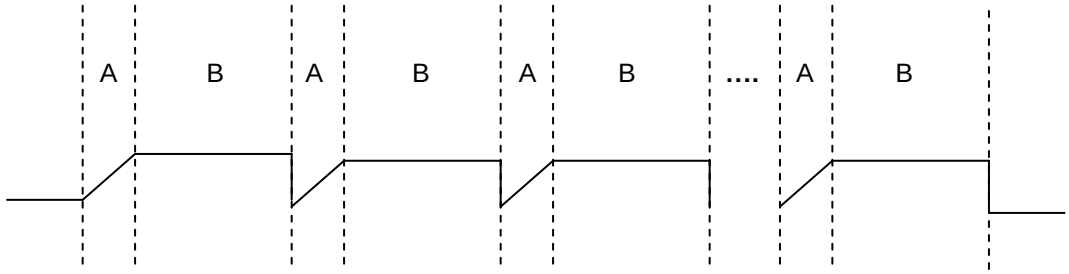
測定手順

定義したアクセスパターンごとに、40 秒間の測定を行います。ただし、最初の 10 秒間（起動段階）は測定データを収集せず、その後の 30 秒間（定常状態段階）のみ、測定データを収集します。

次の図は、測定手順の概略を示しています。

測定フェーズ：

A＝起動段階（10 秒）
B＝定常状態（30 秒）



測定結果

Iometer の測定結果は、負荷プロファイルごとに、さまざまな指標で出力されます。主要な指標は次のとおりです。

- スループット [MB/s] 1 秒あたりのデータ転送量（メガバイト単位）
- トランザクション [MB/s] 1 秒あたりの I/O 処理数
- 遅延 [ms] 平均応答時間（ミリ秒単位）

スループットとトランザクションは互いに正比例の関係にあるので、次の計算式で相互に算出できます。

スループット [MB/s]	*= トランザクション [IO/s] × ブロックサイズ [MB]*
トランザクション [IO/s]	*= スループット [MB/s]／ブロックサイズ [MB]*

通常、シーケンシャルな負荷プロファイルでは「スループット」が使用され、小規模なブロックサイズを使用するランダムな負荷プロファイルでは「トランザクション」が使用されます。

また、負荷プロファイルとは別に、平均応答時間も重要です。平均応答時間は、トランザクションと同時アクセス数に依存します。平均応答時間は、次の計算式で算出できます。

平均応答時間 [ms]	*= 10^3 × ワーカースレッドの数 × 並列 IO／トランザクション [IO/s]*

ディスクサブシステムの分析

計画

ディスクサブシステムのスループットパフォーマンスに重大な影響を与える要因は多岐にわたるため、ディスクサブシステムのサイジングや構成を決定するには、アプリケーションに関する詳細情報が必要です。

特に次の情報が重要です。

- 使用されるアクセスパターン
- 必要なトランザクションレート
- 必要な容量（GB）
- バックアップに許容できる時間
- その時間内にすべてのデータをバックアップ可能か
- リストア中にデータが使用不可になる最長時間
  （バックアップメディアディスクからのリストアだけでなく、トランザクションログからのリストアも含む）

なお、過去の事例から、パフォーマンスの問題を回避するには、ディスクサブシステムを設計する際に次の経験則に従う必要があります。

- アクセスパターンが異なるデータは、別のアレイに配置します。例えば、トランザクションログによるシーケンシャルアクセスとデータベースによるランダムアクセスを同一のドライブに対して行うと、パフォーマンスの問題が発生します。必要なトランザクションレートが確保できるのであれば、1 つのアレイに複数のデータベースを保存する方が危険性は低くなります。
- 大規模なデータベースシステムでは、不適切なトランザクションレートがしばしばボトルネックになります。1 秒あたりに処理可能な I/O 数（トランザクション）を増加させるには、大容量ストレージを少数使用するのではなく、小容量ストレージを多数使用します。
- RAID コントローラーの設定を適切に行います。高パフォーマンスを得られるように RAID コントローラーを設定するには、「ServerView RAID Manager」ユーティリティで、初期設定の「Data Protection」オプションの代わりに「Performance」オプションを使うと便利です。これらの 2 つのオプションには、RAID コントローラーのパラメーターの値があらかじめ設定されているので、すべてのパラメーターを最適に設定できます。また、これらのオプションを使わずに、個別にさまざまな設定を行うこともできます。RAID コントローラーのキャッシュを使用する場合は、停電時のデータ損失を防ぐためにバッテリーバックアップユニット（BBU）を使用する必要があります。
- また、可能であれば、ストレージのライトキャッシュを有効にします。ただし、その場合は、停電時のデータ損失を防ぐために無停電電源装置（UPS）を使用する必要があります。

パフォーマンス問題が発生した場合の分析

ディスクサブシステムのパフォーマンスの分析では、最適化の余地がある領域を特定するために、詳細な情報が必要です。また、異なる構成間で比較する場合は、ディスクサブシステム以外のサーバコンポーネントが重要な場合もあります。例えば、プロセッサ、メモリなどに関連する構成の違いが、不適切な負荷を生成する原因となることがあります。

サーバハードウェア	
サーバ	
CPU	
CPU 数	
メモリ	
メモリの容量	
PCI コントローラー	
サーバソフトウェア	
ハイパーバイザー（使用している場合）	
オペレーティングシステム	
パーティション、ボリューム	
ソフトウェア RAID	
ファイルシステム	
オペレーティングシステム固有のパラメーター設定	
アプリケーション	
ストレージハードウェア	
各コントローラーの情報：	
コントローラータイプ	
バッテリーバックアップユニット（BBU）	
キャッシュサイズ	
キャッシュ設定	
各 RAID アレイの情報：	
RAID レベル	
ドライブ数	
ストライプサイズ	
各ドライブの情報：	
ドライブタイプ	
キャッシュ設定	

ツール

Iometer 以外にもさまざまなツールを使用して、ストレージシステムのパフォーマンスを分析できます。よく使用されるツールの概要を示します。

- Linux
  - sar コマンドを使うと、システム情報を収集、評価、保存できます。
  - strace コマンドを使うと、システムコールとシグナルのログを記録できます。
- Windows
  - パフォーマンスモニターを使うと、Windows システム内の各部に用意されたさまざまなパフォーマンスカウンタを記録し、評価できます。
  - Process Monitor（http://sysinternals.com で入手可能）を使うと、ファイルシステムの動作（ファイルアクセス、レジストリアクセス、ネットワークアクセスなど）の情報を表示して分析できます。
- 外部ディスクサブシステム：
  一部の外部ディスクサブシステムの I/O 動作を分析するためのツールがあります。

これらのツールの詳細については、本書では説明していません。これらのツールを使用する前に、オンラインヘルプやマニュアルをご覧になり、使用方法を確認してください。

ヒント

何らかの理由により、十分なパフォーマンスが得られない原因としてディスクサブシステムが疑われる場合は、関連するアプリケーションの I/O 動作を理解して、パフォーマンスカウンタ（Windows のパフォーマンスモニターなど）を正しく分析する必要があります。ここでいうアプリケーションはサーバ環境でのアプリケーションのことで、通常、エンドユーザーに見えるプログラムではなく、ファイルサーバ、Web サーバ、SQL Server、Exchange Server などを指します。なお、最適化の戦略は、アプリケーションとディスクサブシステムとの間のあらゆるソフトウェア層（ファイルシステムとそのキャッシュ機能、ボリュームマネージャー、I/O ドライバなど）に適用できますが、それによってあらゆる状況でシステム全体が最適なパフォーマンスを発揮できるとは限らないことに留意してください。

また、実際の環境では、各種要因がスループットに与える影響は常に一定ではなく、時間の経過や使用される LUN によって変化します。

各種ツールのパフォーマンスカウンタからパフォーマンスの問題を分析する方法について、いくつか例を挙げて説明します。

- リードリクエストとライトリクエストの比率
  リクエストのリード／ライト比率を取得するには、関連する論理ドライブで実行されている I/O を、オペレーティングシステムやストレージシステムにより提供されるツール（Windows のパフォーマンスモニター、Linux の strace など）を使用して測定します。この測定結果を、アプリケーションに関する知識に基づいて分析し、ストレージシステムが期待どおりに動作しているかどうかを判断します。例えば、主にデータ検索に使用しているファイルサーバで集中的な書き込みアクセスが測定された場合などは、このサーバについてさらに詳細な分析が必要だと判断できます。
- トランザクションのブロックサイズ
  特定のブロックサイズのリード／ライトリクエスト数によって、潜在的なパフォーマンスの問題を明らかにすることができます。例えば、使用しているアプリケーションが 16 KB のブロックサイズで動作する場合、リクエストの大多数はこのサイズであると予測されます。そうでない場合、ボリュームマネージャーまたは I/O ドライバが、リクエストを結合または分割して調整することになります。このような分析を行う際は、Windows のパフォーマンスモニターで提供される平均値（「Avg. Disk Bytes/Read」）などには、ブロックサイズの正確な分布は反映されていないことに注意が必要です。一方、「Process Monitor」を使うと、アプリケーションからファイルシステムに送信されたリクエストを記録できますが、最終的にディスクサブシステムのインターフェース部分で発生したリクエストを直接測定することはできません。外部ディスクサブシステムの分析ツールには、他にもオプションが用意されています。
- アクセスの局所性
  データへのアクセスがデータストック全体に分散せずに、特定の領域で発生することが多い場合、

これをアクセスの局所性と呼びます。アクセスの局所性についての情報は、キャッシュに関する統計情報から得られます。例えば、キャッシュのヒット率が高い場合は、少なくともアクセスの一部が特定のデータ領域内で発生していることを示します。「Process Monitor」または「strace」を使用して、ファイルの処理に使用された領域が分かれば、リード済みまたはライト済みデータのバイト数と共に、例えば 80 GB のファイルへのアクセスが、80 GB の領域全体に完全にランダムに分散されているのか、または特定のフェーズで数ギガバイトのみが処理されているのかを把握できます。後者の場合は、リードキャッシュを有効にすると、パフォーマンスが向上します。前者の場合は、キャッシュを有効にしても、要求するデータがキャッシュに保存されていないことが多いため、効果がありません。

- 論理ドライブへの同時リクエスト数
  論理ドライブのキューで処理を待つリクエスト数（「Avg. Disk Queue Length」）と、特定の状況ではボリュームの利用率（「% Disk Time」）によって、I/O の集中度を測定することができます。論理ドライブを拡張してストレージを追加することで、各ストレージの並列処理に影響を与えて、トランザクションと応答時間を最適化できます。ただし、論理ドライブを拡張する際には、関連するデータベースの完全バックアップとリストアが必要です。
- 応答時間
  論理ドライブの応答時間（「Avg. Disk sec/Transfer」）は、与えられた負荷に対してストレージシステムがどのように反応するかを示します。応答時間は、リクエスト数とブロックサイズだけでなく、リクエストの種類（リードかライトか）により異なり、また、両者混合の場合はその比率によっても異なることに注意してください。
- 論理ドライブ全体に対する I/O 分散の時間による変化
  ディスクサブシステムが高負荷で動作している場合、I/O の集中度が LUN 全体でできるだけ均等になるように注意する必要があります。ただし、実際には I/O の集中度と LUN への分散は、どちらも時間の経過とともに変化するため、均等に分散させることは容易ではありません。例えば、月末、四半期末、年末の負荷は、日常業務で発生する I/O 負荷とはまったく異なります。また、定期的に行われるバックアップや大規模なデータベースクエリがボトルネックになる可能性もあります。ユーザーのログオンまたは休憩時間も、I/O の集中度とボリュームへの分散に影響を与えます。
  そのため、論理ドライブの I/O 負荷を分析する場合は、同時に次の事項についても確認してください。
  - 重要な時間帯の I/O 負荷
  - 負荷が最も高いドライブ
  - 負荷が低いドライブ

  これらの情報は、バランスよくボリュームを利用するために、データを別の論理ドライブに移動するかどうかなどを決める際に役立ちます。バックアップなどの定期的な作業を延期することで、若干のボトルネックを取り除くこともできます。ただし、このような変更を行った後は、別の場所で新しいボトルネックが発生している可能性があるため、ストレージシステムの監視を続ける必要があります。
- RAID レベルとディスク数の最適化
  データベースなどトランザクションが集中するアプリケーションでボトルネックが発生した場合は、RAID レベルを変更する、またはストレージの数を増やすと効果があります。ただし、これらの変更を行う際には、関連するデータベースの完全バックアップとリストアが必要です。現在の RAID コントローラーには、RAID アレイをオンラインで拡張するオプション（「Online capacity expansion」）があります。ただし、このオプションを使用する場合、保存済みの全データの再編成、つまり拡張した アレイへの再分散が必要なため、拡張作業は非常に時間がかかることに留意してください。

以上の分析により、十分なパフォーマンスが得られない原因はストレージシステムではなく、アプリケーション自体、さらにはアプリケーションによるストレージシステムの使用方法にあることが明らかになる場合もあります。

関連資料

PRIMERGY システム

http://ts.fujitsu.com/primergy

PRIMERGY のパフォーマンス

http://ts.fujitsu.com/products/standard_servers/primergy_bov.html

Iometer についての情報

http://www.iometer.org

PC サーバ PRIMERGY（プライマジー）

http://primeserver.fujitsu.com/primergy

お問い合わせ先

富士通テクノロジー・ソリューションズ

Web サイト：http://ts.fujitsu.com

PRIMERGY のパフォーマンスとベンチマーク

mailto:primergy.benchmark@ts.fujitsu.com