FUJITSU

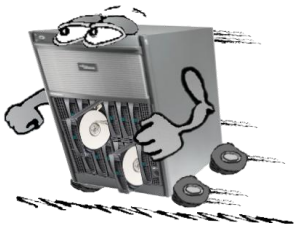
WHITE PAPER

バージョン 1.1
2008 年 3 月

パフォーマンスレポート
PRIMERGY 用モジュラー RAID

ページ数 15

要約



このテクニカルドキュメントは、PRIMERGY に接続されたストレージシステムの RAID コントローラーおよび RAID テクノロジー の選択に携わる人向けに作成されました。このドキュメントを参照することにより、計画しているソリューションに対して、初期段階で適切な RAID 構成を見つけることができます。まず、PRIMERGY の「モジュラーRAID」コントローラーに関連して、さまざまな RAID レベルについて説明します。次に「モジュラー RAID」コントローラーのパフォーマンスデータを、RAID アレイごとに示し、各種ハードディスクを用いた最も一般的な RAID 構成環境でのデータスループットに対するキャッシュ設定の影響を検討します。また、最後に、「モジュラー RAID」コントローラーのパフォーマンスデータを比較しています。

目次

基本原理 2
- コントローラー 2
  - コントローラーキャッシュ 2
  - ディスクキャッシュ 3
- RAID レベル 3
- 測定方法 4
  - 測定ツール 4
  - 負荷プロファイル 4
  - 測定シナリオ 5
  - 測定環境 5

パフォーマンス分析 6
- ディスク 6
- コントローラー 8
  - LSI MegaRAID SAS 1078 コントローラー 8
  - LSI MegaRAID SAS 1064/1068 コントローラー 12
  - コントローラーの比較 13

結論 14

関連資料 15

お問い合わせ先 15

基本原理

各種PRIMERGY は、さまざまな機器構成、種々のハードディスクおよび RAID コントローラー構成で利用することが出来ます。ソリューションの種類を減らすために、通常は個々のサーバで個別に使用していた RAID ソリューションのすべてを、PRIMERGY ファミリーでは「モジュラー RAID」コンセプトで置き換えます。包括的な RAID ソリューションの提供により、ユーザーはアプリケーションシナリオに合った適切なコントローラーを選択できます。

コントローラー

「モジュラー RAID」コンセプトの一部として、3 種類のコントローラーが提供されています。

1. RAID コントローラー LSI MegaRAID SAS 1068

   このコントローラーは、PCI Express カードとして供給されています。このコントローラーに接続できる SATA および SAS ハードディスクの最大数は 8 台です。RAID レベルは、 0、1、1E に対応しています。このコントローラーには、コントローラーキャッシュ設定はありません。

2. RAID コントローラー LSI MegaRAID SAS 1064

   このコントローラーの機能とパフォーマンスは、LSI MegaRAID SAS 1068 コントローラーと同じです。ただし、このコントローラーに接続できるハードディスクは 4 台だけです。LSI MegaRAID SAS コントローラーは、PCI Express カードとして供給され、多数の PRIMERGY モデルではオンボードとしても供給されています。本ドキュメントで示される、最大 4 台のハードディスクに接続した LSI MegaRAID SAS 1068 を使用して測定された値は、LSI MegaRAID SAS 1064 コントローラーでも有効です。

3. RAID コントローラー LSI MegaRAID SAS 1078

   このコントローラーは PCI Express カードとして供給され、RAID ソリューション一式を提供します。SATA と SAS の両方のハードディスクを接続できます。RAID レベルは、0、1、5、6、10、50、60 に対応しています。このコントローラーには、256 MB または 512 MB のキャッシュを備えた 2 つのタイプがあります。電源障害時のデータ損失に対しては、オプションのバッテリーバックアップユニット（BBU）により保護できます。コントローラーは 240 台までのハードディスクをサポートします。PRIMERGY モデルおよびディスクキャビネットの構成によっては、サポートできるハードディスクの数は少なくなります。

コントローラーキャッシュ

LSI MegaRAID SAS 1068 コントローラー とは異なり、LSI MegaRAID SAS 1078 コントローラーには、2 つのバージョンのコントローラーキャッシュがあります。このコントローラーキャッシュは、オプションとして電源障害に対して BBU により保護することができます。コントローラーキャッシュを使用すると、リード／ライトのパフォーマンスを改善することができます。コントローラーキャッシュは、3 種類の設定パラメーターの影響を受けます。

Writeモード

「Write モード」は、コントローラーキャッシュの設定オプションを簡潔に表した用語です。設定できるライトキャッシュ設定には、「write-through」、「write-back」、「write cache bad BBU」の 3 種類があります。「write-through」オプションでは、データが実際にハードディスクで書き込まれたときにのみ、各書き込み要求がコントローラーに認知されるようにします。「write-back」と「write cache bad BBU」オプションでは、要求はコントローラーキャッシュにバッファされ、終了するとユーザーに通知されます。実際には、データはハードディスクにはまったく存在せず、後でハードディスクに書き込まれます。この手順により、コントローラーリソースを最適に利用することができ、書き込み要求のシーケンスが速くなり、その結果スループットが向上します。電源障害には、オプションの BBU により対応できるので、コントローラーキャッシュのデータ整合性が保証されます。「write cache bad BBU」オプションを使用すると、BBU のバッテリーが空になったり、BBU が搭載されていない場合でもキャッシュへの書き込みが有効になります。また、コントローラーキャッシュがバッテリーバッファリングなしの場合は、「write-back」オプションは自動的に「write-through」に切り替わります。

Read モード

「Read モード」パラメーターでは、読み込み中のキャッシュの動作を変えることができます。3 種類のオプション、「No read ahead」、「Read ahead」、「Adaptive」が用意されています。「No read ahead」では、読み取り中にキャッシュは発生しません。オペレーティングシステムによってデータブロックが要求された場合、「Read ahead」の設定では、オペレーティングシステムが後続の要求で他のシーケンシャルデータブロックを要求することを予測して、これらがハードディスクからコントローラーキャッシュに事前に読み込まれます。「Adaptive」に設定した場合は、コントローラー自体が「Read-ahead」が適切かどうかを判断します。

キャッシュモード

「キャッシュモード」パラメーター(Web BIOS では「I/O キャッシュ」として参照されます)は、コントローラーキャッシュの読み取り動作にも影響します。「Direct」オプションにより、読み取るデータをハードディスクから直接読み取るのかどうか、また、コントロールキャッシュに保存しないのかどうかが決まります。別のオプション「Cached」に設定すると、ハードディスクにアクセスする前にまずコントローラーキャッシュ内に読み取り要求を満たすデータがないかの検索が行われ、後続の読み取り要求で使用できるように、すべてのデータがコントローラーキャッシュに書き込まれます。

ディスクキャッシュ

ほとんどの場合、ディスクキャッシュを有効にすると、書き込みアクセスのスループットが向上します。ただし、ディスクキャッシュを有効にすると、パフォーマンスは向上しますが、デメリットも生じます。電源装置で障害が発生した場合、ディスクキャッシュからハードディスクへの書き込みを終えていない重要なデータを永久に損失してしまう可能性があります。そのため、無停電電源装置（UPS）を有効にしてハードディスクに継続的に電源を供給することを推奨しています。システムが UPS で保護されている場合は、パフォーマンスの向上のためにディスクキャッシュを有効にすることをお勧めします。

RAID レベル

RAID 0 を使用すれば、最高のスループットを実現することができます。アレイのハードディスクの数が増加すると、スループットも向上します。スループットの向上は、ハードディスクへの並列アクセスによって達成できます。RAID 0 では、ユーザーはハードディスクの容量全体を使用できます（オーバーヘッド 0%）。

RAID 0 のデメリットは、冗長性をまったく備えていないことです。RAID 0 ハードディスクが故障すると、すべてのデータが失われてしまいます。RAID 0 は通常、データセキュリティがあまり重要でない場合や、データがバックアップされている場合に使用します。

RAID 1 では、2 台のハードディスクを使用してデータの冗長性が最大限保証されます。最高の状態では、読み取りのスループットは、2 台のハードディスクの合計スループットと同等で、 書き込みのスループットは、アレイ内の 1 台のハードディスクと同等です。デメリットは、ユーザーが利用できる容量がアレイ全体の半分になってしまう点です（オーバーヘッド 50%）。

RAID 1E では、2 台以上のハードディスク上で最大限のデータの冗長性が保証されます。読み取りスループットは RAID 1 と同じです。メリットは、より高い RAID レベルには劣りますが、構成の柔軟性が向上する点です。デメリットは、RAID 1 と同様、ユーザーが利用できる容量が全体の半分になってしまう点です（オーバーヘッド 50%）。

RAID 5 は、最低 3 台のハードディスクで構成されます。データおよび追加で計算されたパリティ情報は、全てのハードディスクに分散して書き込まれます。RAID 5 は、高度なデータセキュリティを提供しますが、特に書き込みアクセスのスループットの低い点がデメリットです。アレイ内のハードディスク 1 台分が容量オーバーヘッドとなり、$^{100}/_{\text{RAID 5 アレイ内のハードディスクの台数}}$［%］に相当します。

RAID 6 は、RAID 5 の拡張版で、同時に 2 台のハードディスクが故障しても、データの損失を防ぐことができます。RAID 6 は、高度なセキュリティを提供しますが、RAID 5 と比較するとスループットは低くなります。容量オーバーヘッドは、$^{200}/_{\text{RAID 6 アレイ内のハードディスクの台数}}$［%］です。

RAID 10 は、最低 2 つの RAID 1 を組み合わせた RAID 0 で構成されています。最適なパフォーマンスと、最大限のシステムの信頼性を提供します。RAID 10 では、全体の容量の半分しか使用できません（容量オーバーヘッド 50%）。

RAID 50 は、最低 2 つの RAID 5 を組み合わせた RAID 0 で構成されています。その結果、単なる RAID 5 に比べ、書き込みスループットが向上します。容量オーバーヘッドは、RAID 5 の 2 倍です。

RAID 60 は、RAID 6 と RAID 0 を組み合わせた技術です。RAID 6 の特徴はそのままで、RAID 6 や RAID 50 と比べ耐障害性が向上します。ただし、スループットは RAID 50 より低く、容量オーバーヘッドは RAID 6 の 2 倍です。

測定方法

ディスクサブシステムの能力を評価するために、富士通テクノロジー・ソリューションズは StorageBench というベンチマークを開発しました。StorageBench は、サーバに接続されている異なるストレージシステムを比較することができます。このベンチマークでは、インテルで開発された Iometer という測定ツールと、実際の顧客アプリケーションで発生する負荷プロファイルを組み合わせ、測定シナリオを定義しました。

測定ツール

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 のユーザーインターフェースと、さまざまなプラットフォームで利用できる、いわゆる「dynamo」で構成されています。この数年で、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るようになりました。

Iometer は、IO 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することができます。使用するブロックサイズ、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、アクセスの種類を設定可能です。その結果、Iometer は 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、各アクセスパターンの平均応答時間などの基本的なパラメーターを含むカンマで区切られたテキストファイル(.csv)を生成します。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サブシステムの性能を比較できます。Iometer は、ファイルシステムを使用して 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。

Iometer では、さまざまなアプリケーションのアクセスパターンをシミュレートおよび測定できますが、オペレーティングシステムのファイル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で行われます。

負荷プロファイル

アプリケーションがストレージサブシステムにアクセスする方法は、ストレージシステムのパフォーマンスに多大な影響を及ぼします。各種アプリケーションのさまざまなアクセスパターンの例：

アプリケーション	アクセスパターン
データベース（データ転送）	ランダム、67 %リード、33 %ライト、8 KB（SQL Server）
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 %リード、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
ファイルサーバ	ランダム、67 %リード、33 %ライト、64 KB ブロック
Web サーバ	ランダム、100 %リード、64 KB ブロック
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	ランダム、50 %リード、50 %ライト、64 KB ブロック

これから次の 4 つの独特なプロファイルが導き出されました。

負荷プロファイル	アクセス	アクセスパターン		ブロックサイズ	負荷ツール
		リード	ライト		
ストリーミング	シーケンシャル	100 %		64 KB	Iometer
リストア	シーケンシャル		100 %	64 KB	Iometer
データベース	ランダム	67 %	33 %	8 KB	Iometer
ファイルサーバ	ランダム	67 %	33 %	64 KB	Iometer

4 つのプロファイルはすべて Iometer で生成されました。

測定シナリオ

比較できる測定結果を得るためには、再現可能な同一の環境ですべての測定を実行することが重要です。そのため StorageBench は上記の負荷プロファイルに加えて次の規則に基づいています。

- 実際の顧客構成で RAW デバイスを使用するのは例外的な状況のみであるため、内蔵ディスクのパフォーマンス測定は常にファイルシステムを使用したディスク上で実行されます。高いパフォーマンスが他のファイルシステムや RAW デバイスで実現できる場合でも、Windows では NTFS が使用され、Linux では ext3 が使用されます。
- 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。ハードディスクの故障によるデータの損失をなくすためにサーバシステムで RAID コントローラーが使用される理由はここにあります。ここでは、複数のハードディスクを組み合わせて「Redundant Array of Independent Disks」（RAID）を形成し、1 つのハードディスクが故障した場合でもすべてのデータが維持されるようにすべてのデータを複数のハードディスクに分散させます。ハードディスクをアレイで編成する一般的な RAID レベルは、RAID 0、RAID 1、RAID 1E、RAID 5、RAID 6、RAID 10、RAID 50、RAID 60 です。
- ハードディスクのサイズに関係なく、サイズが 8 GB の測定ファイルを常に測定に使用しています。
- I/O サブシステムの効率の評価では、プロセッサパフォーマンスおよびメモリ構成は、今日のシステムでは大きな要因ではありません。通常、考えられるボトルネックは CPU やメモリではなく、ハードディスクや RAID コントローラーに影響を及ぼします。したがって、CPU やメモリの構成を数々変えながら StorageBench で解析する必要はありません。

測定環境

このドキュメントで説明したすべての測定は、下記の一覧で示したハードウェアとソフトウェアのコンポーネントを使用して実行されました。

コンポーネント	詳細
サーバ	PRIMERGY TX200 S4 PRIMERGY RX300 S4
コントローラー LSI MegaRAID SAS 1068	ドライバ名：lsi_sas.sys、ドライバのバージョン：1.25.05.00、ファームウェアのバージョン：01.18.41.00、BIOS のバージョン：06.12.00.00
コントローラー LSI MegaRAID SAS 1078（256 MB または 512 MB のキャッシュを搭載）	ドライバ名：msas2kr.sys、ドライバのバージョン：2.17.0.32、ファームウェアのパッケージのバージョン：6.0.1-0081、ファームウェアのバージョン：1.11.72-0356、BIOS のバージョン：NT10
ハードディスク SATA、3.5 インチ、7.2 krpm	Western Digital WD1600AAJS、160 GB
ハードディスク SAS、2.5 インチ、10 krpm	Seagate ST973402SS、73 GB
ハードディスク SAS、2.5 インチ、15 krpm	Seagate ST973451SS、73 GB
ハードディスク SAS、3.5 インチ、10 krpm	Seagate ST373355SS、73 GB
ハードディスク SAS、3.5 インチ、15 krpm	Seagate ST373455SS、73 GB
オペレーティングシステム	Windows Server 2003、Enterprise Edition、Service Pack 1
ファイルシステム	NTFS
テストツール	Iometer 2006.07.27
テストデータ	8 GB の測定ファイル

パフォーマンス分析

ディスク

「モジュラー RAID」ファミリーのコントローラーを搭載することができる PRIMERGY モデルでは、さまざまなハードディスクが使用可能であり、SATA および SAS ハードディスクで RAID を構成できます。必要なパフォーマンスに応じて、適切なディスクサブシステムを選択できます。使用している PRIMERGY に関係なく、各種ハードディスクタイプのパフォーマンスを比較したデータを以下に示します。「モジュラー RAID」コントローラー搭載のすべての PRIMERGY モデルが全種類のハードディスクに対応しているわけではありません。さらに、個々のモデルの構成オプションは異なります。詳細については、PRIMERGY のデータシートを参照してください。

SAS ディスク

SAS ハードディスクの回転数は、SATA ハードディスクと比べると速いため、SAS ハードディスクの方がアクセス時間は短く、スループットも高くなります。回転数が速いことの唯一のデメリットは、ノイズや熱が発生しやすくなるため、追加の冷却が必要になることです。2.5 インチのドライブは、消費電力と熱の発生を抑えることができ、デバイスの冷却コストを削減できるという大きなメリットがあります。また、スペースを有効利用できる点もメリットです。ただし、2.5 インチハードディスクにも、容量やスループットが低いなどのデメリットがあります。

SATA ディスク

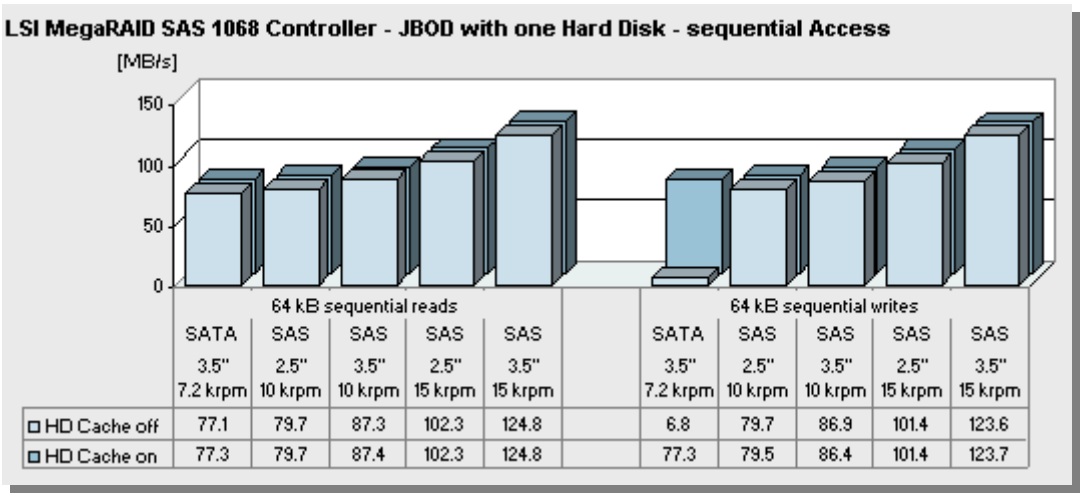
SATA ハードディスクの回転数は、SAS ハードディスクと比べると遅いため、SATA ハードディスクの方がアクセス時間は長く、スループットも低くなります。ただし、SATA ハードディスクは、SAS と比べて安価に、テラバイト級の大容量のシステムを構築できます。特に、セカンダリストレージやバックアップシステムでの使用に適しています。

ハードディスクの比較

テストでは、1 台のハードディスクだけを LSI MegaRAID SAS 1068 コントローラーに接続して、JBOD として構成しました。測定では、PRIMERGY で現在使用可能なすべてのハードディスク、例えば、3.5 インチ、2.5 インチ SAS ハードディスク（回転数 10 krpm または 15 krpm）または 3.5 インチ SATA ハードディスク（回転数 7.2 krpm）について分析を行いました。個々のハードディスクのスループットを、さまざまなアクセスパターンで比較します。

ハードディスクキャッシュは ディスク I/O パフォーマンスに影響を及ぼします。残念ながら、多くの場合、この機能は電源障害時の安全上の問題により無効化されています。しかし、ハードディスクの製造元は、書き込みパフォーマンスの向上のためにディスクキャッシュを組み込んでいます。 NCQ（Native Command Queuing：ネイティブ・コマンド・キューイング）などの機能は、ディスクキャッシュが有効なときにしか機能しません。特に SAS ハードディスクに比べて回転数が遅い SATA ハードディスクを用いる場合には、パフォーマンスを向上させるため、ディスクキャッシュを有効にしてください。I/O アクセス用のキャッシュは圧倒的に大きく、電源障害時の潜在的なリスク（データの損失）がどんな場合でもメインメモリには存在します。これは、オペレーティングシステムによって管理されます。データの損失を防止するには、システムに UPS を装備することを推奨します。ハードディスク比較の測定は、それぞれディスクキャッシュあり・なしで実施しました。

次の図は、64 KB ブロックサイズを使用してシーケンシャルリード／ライトを行った場合のスループットが、回転数の増加に伴って向上することを示しています。



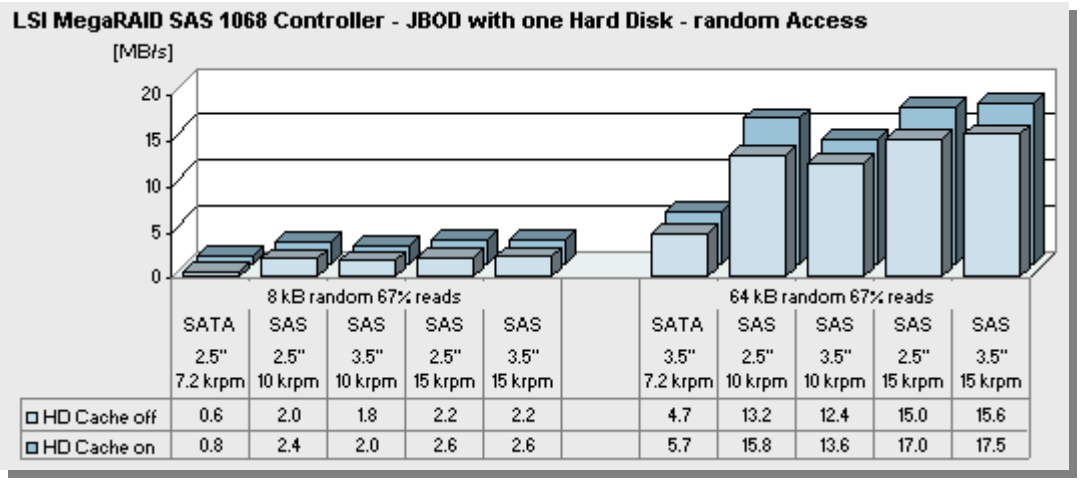
	SATA 3.5" 7.2 krpm	SAS 2.5" 10 krpm	SAS 3.5" 10 krpm	SAS 2.5" 15 krpm	SAS 3.5" 15 krpm	SATA 3.5" 7.2 krpm	SAS 2.5" 10 krpm	SAS 3.5" 10 krpm	SAS 2.5" 15 krpm	SAS 3.5" 15 krpm
HD Cache off	77.1	79.7	87.3	102.3	124.8	6.8	79.7	86.9	101.4	123.6
HD Cache on	77.3	79.7	87.4	102.3	124.8	77.3	79.5	86.4	101.4	123.7

回転数 10 krpm のハードディスクの代わりに回転数 15 krpm のものを使用した場合、2.5 インチハードディスクでは約 28 %、3.5 インチハードディスクでは 42 %スループットが向上しました。回転数が 10 krpm の 2.5 インチおよび 3.5 インチハードディスクのスループットを比較した場合、3.5 インチハードディスクのスループットは 2.5 インチハードディスクよりも約 9 %大きいことがわかります。回転数が 15 krpm の場合、3.5 インチハードディスクのスループットは 2.5 インチハードディスクよりもさらに大きく、その差は約 22 %であることがわかります。

3.5 インチ SAS ハードディスクと 3.5 インチ SATA ハードディスクをディスクキャッシュを有効にしてシーケンシャルリード／ライトで比較すると、回転数 10 krpm の SAS ハードディスクのスループットは回転数 7.2 krpm の SATA ハードディスクよりも約 12 %高く、回転数 15 krpm の 3.5 インチ SAS ハードディスクと SATA ハードディスクを比較すると、回転数 15 krpm の 3.5 インチ SAS ハードディスクのスループットは SATA ハードディスクよりも約 60 %高いことがわかります。

ディスクキャッシュを有効にした SATA ハードディスクでは、シーケンシャルライトのスループットが特に向上し、最大で 11 倍にもなります。ただし、SAS ハードディスクの場合は、ディスクキャッシュを有効にしても、シーケンシャルリード／ライトでのパフォーマンスに顕著な向上は見られません。

67 %リードのランダムアクセスでは、SAS ハードディスクのディスクキャッシュは、基本的に、シーケンシャルリード／ライトの場合よりも、スループットの向上に大きく寄与していることがわかります。増加率は最大 20 %です。SATA ハードディスクでは、スループットは最大 33 %増加しました。回転数 10 krpm のハードディスクと 15 krpm のハードディスクを比較すると、15 krmp の方が、2.5 インチのハードディスクでは約 8 %、3.5 インチのハードディスクでは約 30 %パフォーマンスが向上しています。



	SATA 2.5" 7.2 krpm	SAS 2.5" 10 krpm	SAS 3.5" 10 krpm	SAS 2.5" 15 krpm	SAS 3.5" 15 krpm	SATA 3.5" 7.2 krpm	SAS 2.5" 10 krpm	SAS 3.5" 10 krpm	SAS 2.5" 15 krpm	SAS 3.5" 15 krpm
HD Cache off	0.6	2.0	1.8	2.2	2.2	4.7	13.2	12.4	15.0	15.6
HD Cache on	0.8	2.4	2.0	2.6	2.6	5.7	15.8	13.6	17.0	17.5

ディスクキャッシュを有効にした場合のランダムアクセスに関して、3.5 インチの SAS ハードディスクを 3.5 インチの SATA ハードディスクと比較すると、10 krpm の SAS ハードディスクのスループットは、7.2 krpm の SATA ハードディスクよりも約 3 倍高いことがわかります。回転数 15 krpm の 3.5 インチ SAS ハードディスクと、7.2 krpm の 3.5 インチ SATA ハードディスクを比較すると、3.5 インチ SAS ハードディスクのスループットは SATA ハードディスクよりも約 3.35 倍高いことがわかります。

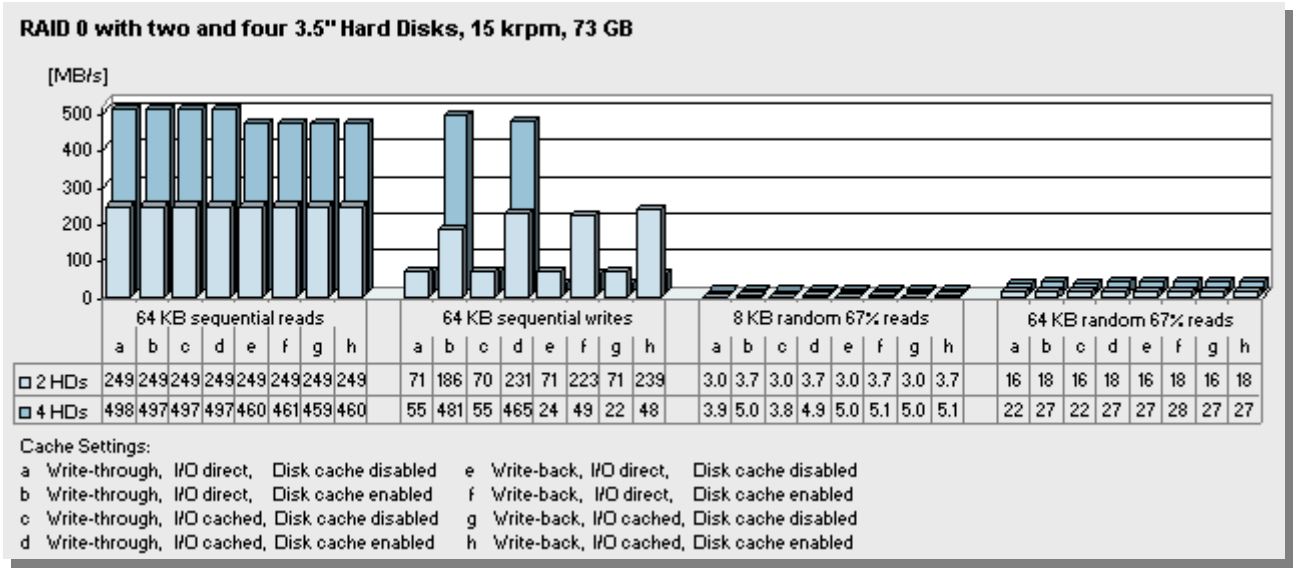
コントローラー

可用性の観点からどのようにデータが扱われるかは、RAIDアレイによって決まります。各 RAID アレイ内でデータが転送される速さは、ハードディスクのデータスループットによって大きく異なります。RAID アレイで測定用に構成されるハードディスクの数は、RAID レベルに応じて決定されました。使用されたハードディスクの台数は、2〜6 台です。さまざまなキャッシュ設定でのコントローラーの性能を測定する際に、ハードディスクがボトルネックにならないように、可能な限り、回転速度が 15 krpm の 3.5 インチハードディスクを用いて測定を実施しました。ただし、「モジュラー RAID」コントローラーを搭載できる各種 PRIMERGY モデルは、システムに応じて各種ハードディスクモデルを搭載できるので、実際に達成できるスループットは低くなる可能性があります。

LSI MegaRAID SAS 1078 コントローラー

LSI MegaRAID SAS 1078 コントローラーには、各種 RAID レベルが用意されており、それぞれのパフォーマンスは以下のように分析されます。キャッシュ設定によって、スループットが大幅に向上する場合があるので、さまざまなキャッシュ設定で比較されています。ただし、このようなスループットの増加は、データの構造とアクセスのパターンによって異なります。すべての測定は、512 MB のキャッシュを搭載したLSI MegaRAID SAS 1078 コントローラーを使用して行われました。

次の図は、2 台または 4 台の 3.5 インチハードディスクを使用した RAID 0 のアレイにおける、各種負荷プロファイルおよびキャッシュ設定下でのスループットを示しています。



LSI MegaRAID SAS 1078（512 MB のキャッシュを搭載）

キャッシュを有効にすれば、書き込みスループットを数倍向上させることができます。この場合、ディスクキャッシュを有効にすると、スループットは大きく向上します。ただし、最大書き込みスループットを達成するには、キャッシュ設定「Disk-Cache enabled」、「Write-back」、「I/O cached」を組み合わせる必要があります（上図内のCache Settings：h の項を参照）。これは、シーケンシャルライトアクセスの最適キャッシュ設定と同じです。キャッシュが無効の書き込みスループットと比較すると、この方法では 3 倍のスループットを達成できます。

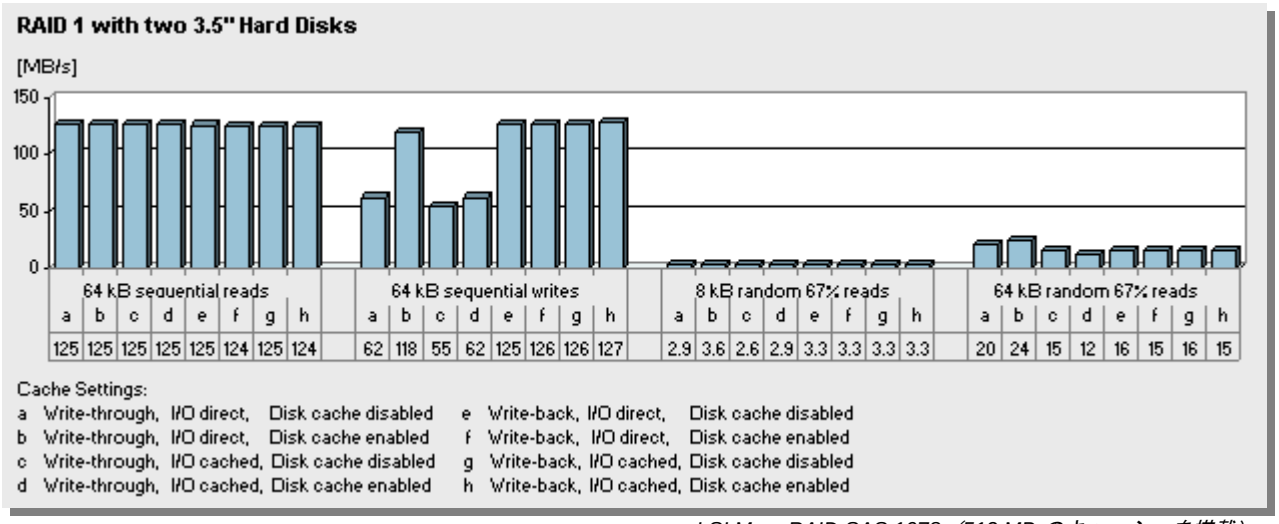
また、ハードディスク 2 台の RAID 0 アレイの読み取りスループットは、キャッシュ設定の影響を受けません。読み取りスループットの最大値は、キャッシュ設定に関係なく約 250 MB/s（2 × 125 MB/s）です。最適なキャッシュ設定を行った場合のスループットは、読み取りが 67 %を占める 8 KB ブロックでのランダムアクセスでは約15 %、読み取りが 67 %を占める 64 KB ブロックのランダムアクセスでは約 20 %増加します。

4 台の 3.5 インチハードディスクを使用した RAID 0 アレイでは、「Write-back」に設定するとスループットが低下します。この場合、最大可能値と比較して、読み取りスループットは約 8 %低下します。

最適なキャッシュ設定を行った場合の書き込みスループットは、最大可能書き込みスループットをわずかに下回るだけです。キャッシュを無効にした場合のスループットと比べると、スループットは、約 8.7 倍に向上します。

ディスクキャッシュを有効にすると、67 %読み取りのランダムアクセスでは大きなメリットがあり、スループットは 8 KB ブロックでは約 25 %、64 KB ブロックでは約 35 %向上します。2 台と 4 台の 3.5 インチハードディスクの RAID 0 アレイを直接比較した場合、ハードディスクの台数を増やして、シーケンシャルリード／ライトアクセスのためのキャッシュ設定を最適にすると、スループットは 2 倍に向上します。

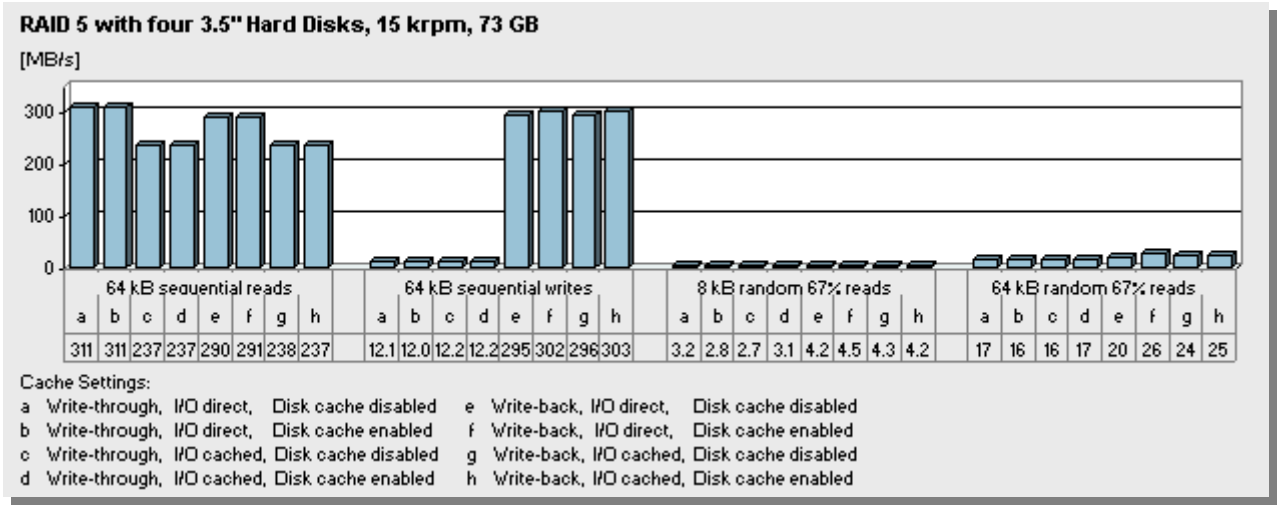
次の図は、3.5 インチハードディスク2 台を使用した RAID 1 アレイでのスループットを示しています。読み取りスループットは、可能な最大スループット値、125 MB/s に達しています。キャッシュ設定が読み取りスループットに与える影



LSI MegaRAID SAS 1078（512 MB のキャッシュを搭載）

響はわずかです。対照的に、書き込みスループットは、キャッシュ設定の影響を大きく受けます。最善のパフォーマンスを実現するには、最適なキャッシュ設定「Write-back」、「I/O direct」、「Disk-Cache enabled」を使用する必要があります。この方法では、約 60 %パフォーマンスが向上します。同じように、64 KB ブロックのランダムアクセスでも、キャッシュ設定によってスループットが大きく異なります。最適なキャッシュ設定により、スループットは約 50 %向上します。

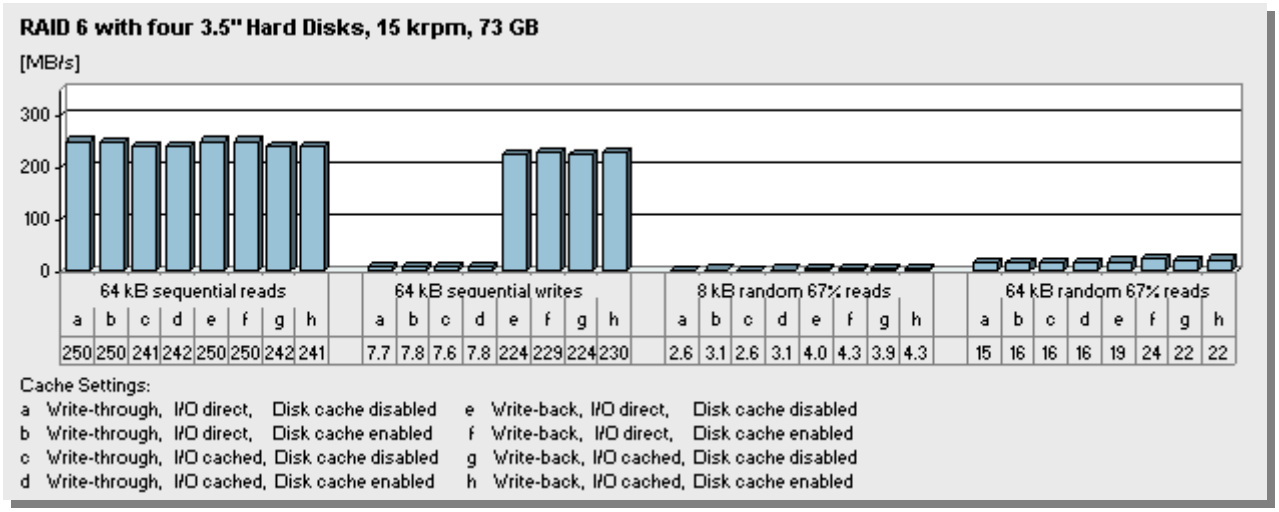
最適なキャッシュ設定を行うことの重要性は、RAID 5 で特に明らかです。次の図は、オプションを「Write-back」に設定してコントローラーキャッシュを有効にするとシーケンシャルライトスループットが大きく向上し、書き込みアクセスでは追加でパリティブロックを計算して書き込む必要があるにもかかわらず、シーケンシャルリードよりも高い値になることを示しています。一方、シーケンシャルリードのスループットに対しては、キャッシュ設定の影響はあまりありません。興味深いことに、読み取りスループットには、I/O キャッシュが逆効果であることがわかります。



LSI MegaRAID SAS 1078（512 MB のキャッシュを搭載）

64 KB ブロックのランダムアクセスでも、キャッシュ設定によってスループットが大きく異なります。最適なキャッシュ設定により、スループットは約 40 %向上します。

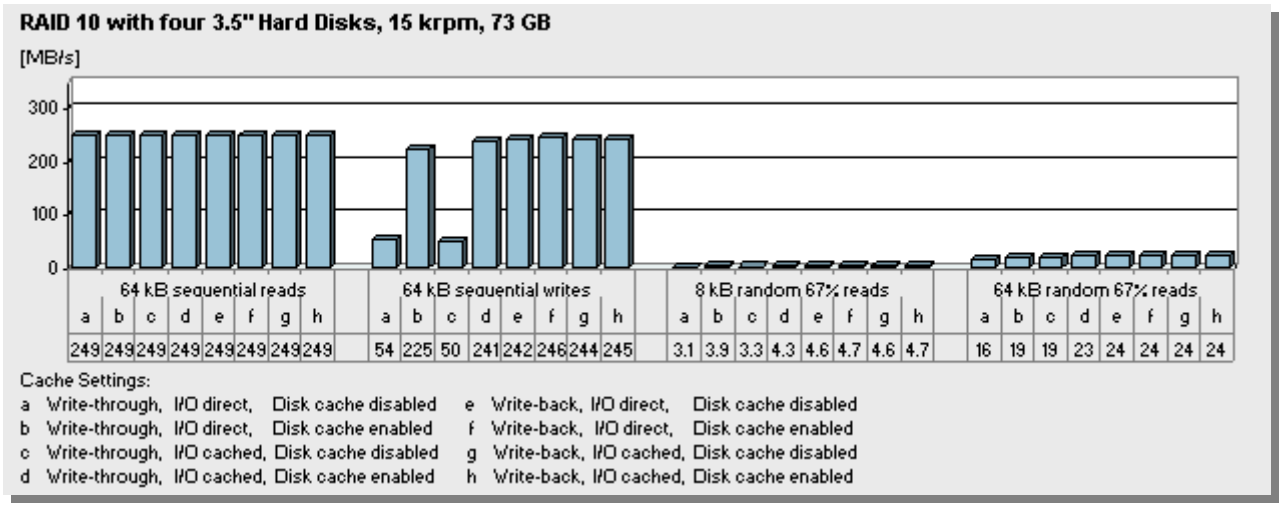
RAID 6 は、RAID 5 よりもシステムの信頼性が向上しており、 アレイ内で 2 台のハードディスクが故障しても、データが失われることはありません。RAID 6 においてキャッシュ設定がスループットに及ぼす影響は、下図が示しているよう



LSI MegaRAID SAS 1078（512 MB のキャッシュを搭載）

に、RAID 5 と非常に似ています。ただし、I/O キャッシュを有効にした場合の影響は、RAID 6 では RAID 5 の場合のようには明確ではありません。追加のパリティブロックを書き込む影響から、RAID 6 の書き込みスループットは、RAID 5 アレイよりもやや低くなります。ランダムアクセスのスループットは、どちらのアレイでもほぼ同じですが、シーケンシャルリード／ライトのスループットでは違いが大きくなります。キャッシュが最適に設定された場合、読み取りスループットで約 24 %、書き込みスループットではさらに大きく40 % RAID 5 の方がRAID 6 より高い値を示します。

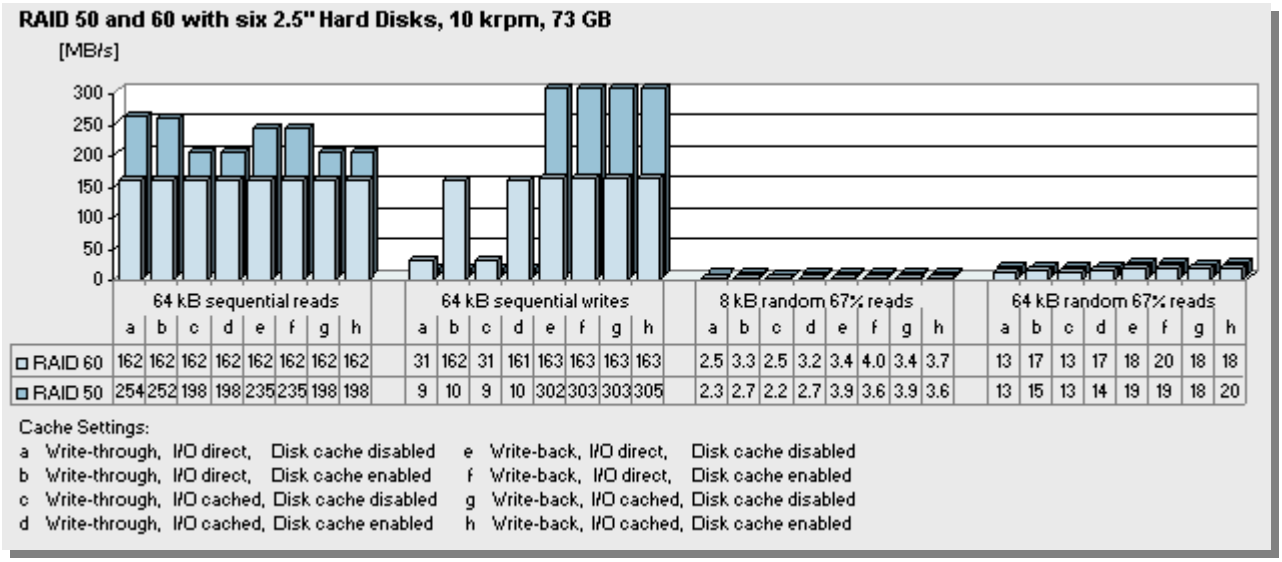
さらに高度なデータセキュリティとデータスループットを確保できるソリューションは、RAID 10 です。ただし、容量オーバーヘッドは 50 %です。次の図は、4 台のハードディスクを使用した RAID 10 のスループットの比較を示しています。



LSI MegaRAID SAS 1078（512 MB のキャッシュを搭載）

シーケンシャルリードの読み取りスループットは、3.5 インチハードディスクの最大スループットとほぼ同じです。「Write-back」オプションで最適なキャッシュ設定を行った場合は、シーケンシャルライトのスループットも 3.5 インチハードディスクの最大スループットとほぼ同じになります。また、67 %リードのランダムアクセスでは、平均で約 20 %スループットが向上します。一見、RAID 5 および RAID 6 の方が、RAID 10 よりも効率的であるように見えます。測定データを詳しく見てみると、確かに純粋なシーケンシャルアクセスパターンの場合には当てはまります。しかし、このような純粋なシーケンシャルアクセスパターンが実際に発生することはほとんどありません。リード／ライトアクセスが混合したより現実的なアクセスプロファイルでは RAID 5 および RAID 6 より、RAID 10 のスループットの方が最大 61 %高くなります。

次の図では、回転数 10 krpm の 6 台の 2.5 インチハードディスクを使用した RAID アレイを例として、RAID 50 と RAID 60 のスループットを比較しています。最適なキャッシュ設定が行われている場合、RAID 50 のシーケンシャルリ

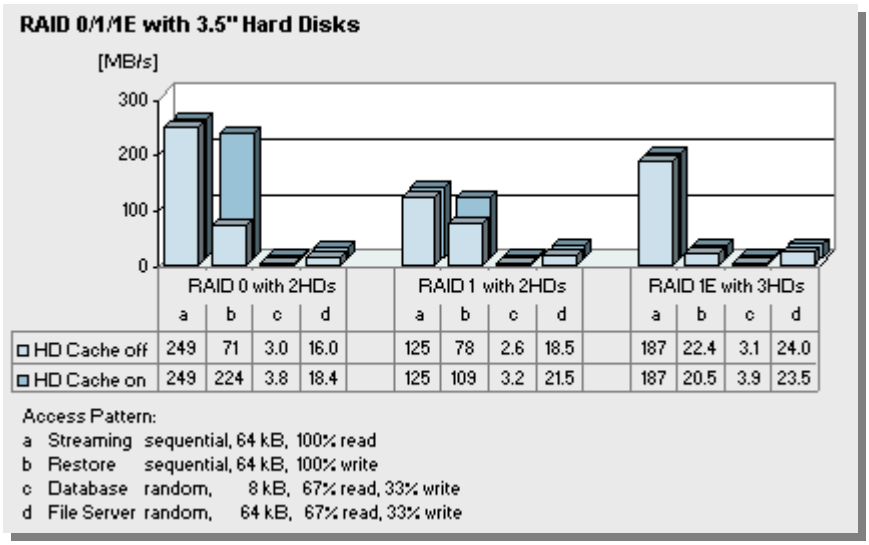


LSI MegaRAID SAS 1078（512 MB のキャッシュを搭載）

ードのスループットは、RAID 60 よりも約 55 %高くなります。シーケンシャルライトでは、すべてのキャッシュを有効にした場合、スループットは 88 %も高くなります。67 %読み取りのランダムアクセスでは、2 つの RAID アレイの平均値はほぼ同じです。RAID 50 のパリティオーバーヘッドは、RAID 60 よりも低くなります。容量オーバーヘッド（%）は、RAID 50 では $2\times {}^{100}/_{\text{RAID 50 アレイのハードディスクの台数}}$、RAID 60 では、$2\times {}^{200}/_{\text{RAID 60 アレイのハードディスクの台数}}$となります。RAID 60 は、データセキュリティが高い分、スループットとディスク容量が低くなります。各 RAID 6 サブセットのハードディスクが最大 2 台ずつ同時に故障しても、データは保護されます。RAID 50 では、各 RAID 5 サブセットのハードディスクが最大 1 台ずつ同時に故障しても、データは保護されます。

LSI MegaRAID SAS 1064/1068 コントローラー

LSI MegaRAID 1064 および 1068 SAS コントローラーには、コントローラーキャッシュはありません。RAID レベル 0、1、1E のみに対応しています。使用している PRIMERGY のシステム構成によって、最大で 4 台（LSI SAS 1064 コントローラー）と 8 台（LSI SAS 1068 コントローラー）の SAS または SATA ハードディスクをコントローラーに接続することができます。一部の PRIMERGY モデルでは、LSI MegaRAID 1064 SAS コントローラーは、オンボードで提供されます。



	RAID 0 with 2HDs				RAID 1 with 2HDs				RAID 1E with 3HDs			
	a	b	c	d	a	b	c	d	a	b	c	d
HD Cache off	249	71	3.0	16.0	125	78	2.6	18.5	187	22.4	3.1	24.0
HD Cache on	249	224	3.8	18.4	125	109	3.2	21.5	187	20.5	3.9	23.5

Access Pattern:
a Streaming sequential, 64 kB, 100% read
b Restore sequential, 64 kB, 100% write
c Database random, 8 kB, 67% read, 33% write
d File Server random, 64 kB, 67% read, 33% write

左図では、サポートされている各 RAID アレイのスループットを比較しています。シーケンシャルリードでは、RAID アレイの種類、ディスクキャッシュが有効か無効かに関係なく、すべてのハードディスクで最大可能スループットを達成していることがわかります。

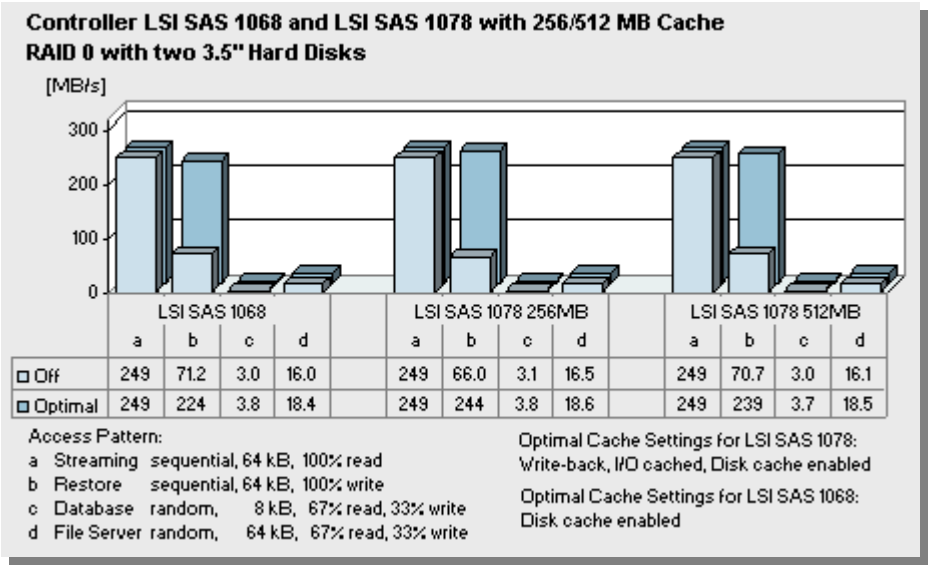
RAID 0 アレイのシーケンシャルライトのスループットは、ハードディスクのディスクキャッシュを無効にすると、読み取りスループットよりも大幅に低くなります。ディスクキャッシュを有効にすれば、シーケンシャルライトを 3 倍 以上に向上させることができます。RAID 0 のランダムアクセスでも、ディスクキャッシュを有効にすることでスループットを、8 KB ブロックのランダムアクセスでは 26 %、64 KB ブロックのランダムアクセスでは 15 %向上させることができます。

RAID 1 で、ディスクキャッシュを無効にした場合のシーケンシャルライトのスループットは、RAID 0 のスループット値とほぼ同じになります。ディスクキャッシュを有効にすれば、約 40 %スループットを向上させることができます。

RAID 1E のシーケンシャルライトのスループットは値が低く、RAID 1 よりも低くなります。書き込みスループットは、ディスクキャッシュを有効にすることで、少し向上させることができます。ただし、ランダムアクセスでは、ディスクキャッシュを有効にすることで、スループットは約 26 %向上し、得られたスループットは、RAID 0 および RAID 1 と同じレベルです。ランダムアクセスでは、RAID 1E のスループットは、RAID 1 よりも約 30 %高く、RAID 0 よりも 50 %高くなります。

コントローラーの比較

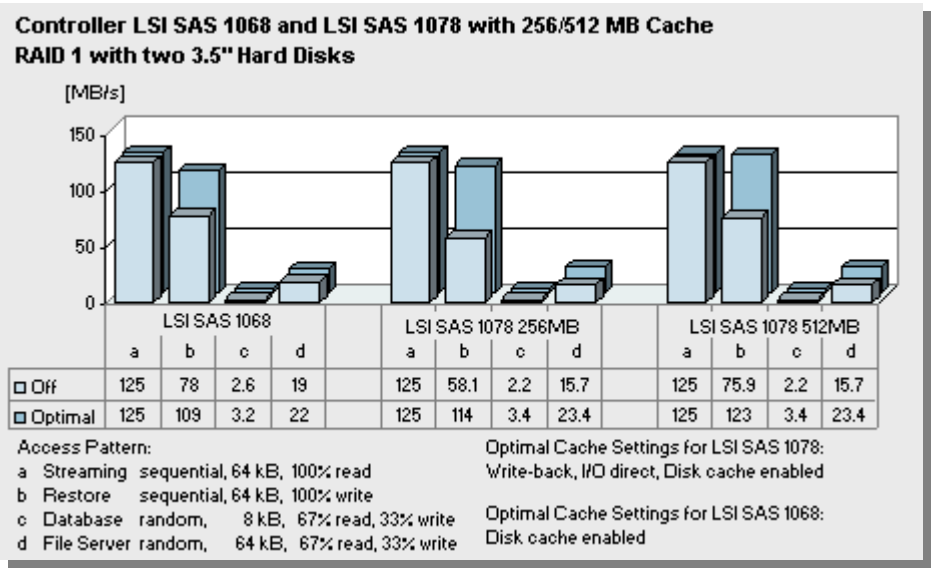
ここでは、さまざまなコントローラーのスループットを比較します。測定は、同じ RAID アレイで同じ種類のハードディスクを用いて行われました。下図では、キャッシュを無効にした場合（Off）と、最適なキャッシュ設定を行った場合（Optimal）に得られるスループットを示しています。



2 台のハードディスクを使用した RAID 0 アレイでは、キャッシュの設定は、シーケンシャルリードのスループットに対して、まったく影響しないことがわかります。得られたスループットは、最大可能値に非常に近くなります。3 台すべてのコントローラーは、64 KB ブロックのシーケンシャルリードでは同等の性能を示します。一方、64 KB ブロックのシーケンシャルライトでは、最適なキャッシュ設定を行うことによって、スループットを向上させることができます。LSI MegaRAID SAS 1068 コントローラーでは、スループットは 3.1 倍に向上します。LSI MegaRAID SAS 1078 コントローラー（512 MB キャッシュ搭載）では 3.4 倍、LSI MegaRAID SAS 1078 コントローラー（256 MB キャッシュ搭載）では 3.7 倍に向上します。LSI MegaRAID SAS 1078 コントローラーでは、使用されているコントローラーキャッシュのサイズに関係なく、ほぼ同じスループットを得ることができ、LSI MegaRAID SAS 1068 コントローラーよりも最大 9 %パフォーマンスが向上します。

67 %読み取りのランダムアクセスでは、最適なキャッシュ設定を行うことである程度スループットを向上させることができますが、2 台のハードディスクを使用した RAID 0 のスループットは本質的にほぼ同じです。

右図は RAID 1 での測定結果です。RAID 0 アレイと同様、キャッシュ設定は、コントローラーに関係なく、シーケンシャルリードのスループットにはまったく、あるいはほとんど影響しません。得られたスループット値は、最大可能値と同じです。シーケンシャルライトでも、最適なキャッシュの設定を行うことで、スループットを向上させることができます。ただし、RAID 0 アレイほど大きく向上するわけではありません。LSI MegaRAID SAS 1068 コントローラーでは、スループットは約 40 %向上します。最大スループット値は、LSI MegaRAID SAS 1078 コントローラー（512 MB コントローラーキャッシュ搭載）で達成されました。他と比較すると、LSI MegaRAID 1078 コントローラー（256 MB コントローラーキャッシュ搭載）よりも約 8 %高く、 LSI MegaRAID SAS 1068 コントローラーよりも、約 13 %高い値を示しました。



結論

PRIMERGY の「モジュラー RAID」のコンセプトによって、さまざまなアプリケーションシナリオの多様な要件を満たすことができます。

LSI MegaRAID SAS 1068 コントローラーに代表されるエントリーレベルのコントローラーでは、基本的な RAID ソリューション RAID 0、RAID 1 および RAID 1E が提供されており、それぞれのRAID レベルにおいて最善のパフォーマンスがサポートされています。

LSI MegaRAID SAS 1078 コントローラーに代表されるハイエンドコントローラーでは、現在のすべての RAID ソリューション、RAID レベル 0、1、5、6、10、50 および 60 が実現できます。このコントローラーには、256 MB または 512 MB のコントローラーキャッシュが搭載され、オプションとして、BBU を使用したデータの保護が可能です。キャッシュに関するさまざまな設定を行うことで、使用する RAID レベルに合わせた最適なパフォーマンスを柔軟に引き出すことができます。

RAID 5 または RAID 6 を使用すると、既存のハードディスクの容量を経済的に活用して、優れたパフォーマンスを実現できます。ただし、最善のパフォーマンスとセキュリティのためには、RAID 10 をお勧めします。

「モジュラー RAID」は、ハードウェア構成や、コントローラーおよびハードディスクの構成オプションが異なるさまざまな PRIMERGY モデルで使用されています。PRIMERGY の機種によって、SATAハードディスク 、SAS 2.5 インチハードディスク、SAS 3.5 インチハードディスクから選択できます。また、SAS ディスクの回転数は 10 krpm または 15 krpm から選べます。使用するハードディスクは、必要なパフォーマンスに応じて、回転速度も含めて決定する必要があります。15 krpm のハードディスクでは、約 42 % のパフォーマンスの向上が可能です。RAID レベルによりますが、2.5 インチハードディスクを使用することでハードディスクを増やし、より高いレベルでの並列処理を実現できます。

最高のパフォーマンスを達成するためには、特に SATA ハードディスクを使用する場合やコントローラーキャッシュを持たないコントローラーを使用する場合、ハードディスクのキャッシュを有効にします。使用するディスクの種類とアクセスパターンによっては、パフォーマンスは最大 11 倍に向上します。ハードディスクのキャッシュを有効にする場合には、UPS の使用を推奨します。

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

PRIMERGY パフォーマンスとベンチマーク
mailto:primergy.benchmark@ts.fujitsu.com

発行部門：

Enterprise Products
PRIMERGY Server
PRIMERGY Performance Lab
mailto:primergy.benchmark@ts.fujitsu.com

インターネット：
http://ts.fujitsu.com/primergy

エクストラネット：
http://partners.ts.fujitsu.com/com/products/servers/primergy