ホワイトペーパー
FUJITSU Server PRIMERGY & PRIMEQUEST
パフォーマンスレポート PCIe-SSDs P3700

この資料では、一部の PRIMERGY サーバおよび PRIMEQUEST サーバで使用できる P3700 シリーズの PCIe-SSD のディスク I/O パフォーマンスについて詳しく説明します。測定結果とともに、測定方法と測定の実施環境についても簡単に説明します。

バージョン
1.0b
2016-03-24

目次

ドキュメントの履歴

バージョン 1.0

初回報告版

バージョン 1.0a

マイナー修正

バージョン 1.0b

マイナー修正

基本情報と製品データ

PCIe-SSD は、PRIMERGY および PRIMEQUEST サーバ用の革新的な不揮発性ストレージ媒体です。特定のケースにおいて、SAS または SATA インターフェースを持つ「ハードディスクドライブ」（HDD）または「ソリッドステートドライブ」（SSD）の代わりに論理ハードディスクドライブとして使用できます。PCIe-SSD ストレージ媒体は PCIe バスと直接結合されているため、非常に高いトランザクションレートを発揮し、遅延を削減できます。このストレージ媒体は、以下のような不揮発性の環境に保存されているデータへの高速アクセスを可能します。

データベース

- Web 2.0 サーバ
- ページファイル（特にページファイルがメモリ管理コンセプトで重要な役割を果たす仮想環境で有効です）

また、ストレージメディアファミリに新たに追加されたフォームファクタについても、以下で説明します。PCIe-SSD は、サーバで PCIe スロットの 1 つのバージョンとして引き続き存在しています。さらに、この目的で明示的に提供されているサーバの 2.5″ HDD ベイで便宜上外部接続できるバージョンもあります。

製品データ

現在発売されている PCIe-SSD は、マルチレベルセル（MLC）NAND タイプのメモリをベースとしたフラッシュメモリです。PCIe-SSD は、5 年間の 10 DWPD（drive writes per day）が可能であることから、書き込み耐久性クラスは「主流の耐久性」[1]に該当します。現在、サーバモデルに応じた 3 種類のストレージ容量があり、それぞれ 2 種類のフォームファクタのいずれかで発注できます。

ストレージ容量	発注コード	
	SFF バージョン（スモールフォームファクタ = 2.5″）	AIC バージョン（アドインカード）
800 GB	SSD PCIe3 800GB Main 2.5' H-P EP	PCIe-SSD 800GB P3700
1.6 TB	SSD PCIe3 1.6TB Main 2.5' H-P EP	PCIe-SSD 1.6TB P3700
2 TB	SSD PCIe3 2TB Main 2.5' H-P EP	PCIe-SSD 2TB P3700

2 つのフォームファクタではパフォーマンスは異なりません。

ストレージ媒体自体に強力なコントローラーが内蔵されているため、RAID コントローラーを別途用意する必要はありません。サーバへのバスインターフェースは PCIe 3.0、x4 タイプで、最大約 3380 MB/s の実効スループットが得られます。SFF バージョンはホットプラグ対応です。PCIe 3.0 x16 スロットを占有する特殊な PCIe スイッチは、4 台の PCIe-SSD に 1 個の割合で、RX/TX サーバの SFF バージョンに使用されます。

他のほとんどのストレージメディアに比べて、この世代の PCIe-SSD は、熱特性の影響をより多く受けます。ライト負荷がきわめて高い負荷プロファイルの場合、ここで取り上げている PCIe-SSD では、最大で 25 W もの廃熱が発生します。2.5″ HDD ベイの特殊な冷却機能により、この廃熱は SFF バージョンに大きな影響があります。ストレージメディア自体の温度は、（パフォーマンスの低下を伴う）適切なタイミングでのスロットリングにより、上限の 71 °C を超えることはありません。このような廃熱によるパフォーマンスの低下を避けるためには、リリースされているサーバおよびストレージメディアの構成および設置環境の条件を守ることが不可欠です。これは、特に以下のことを意味します。

- サーバの iRMC ファームウェアには、最新リリースのバージョンを使用する必要があります。これは、ファンの規定が PCIe-SSD に必ず対応するようにするための唯一の方法です。
- 通気に影響するような改造など、リリース時のサーバの状態を変更することは、禁じられています（例えば、ダミーカバーの取り外しなど）。
- 環境条件によって冷却に支障が生じることがないようにします（例えば、高温、通気を妨げる障害物など）。

[1] SSD の特殊な機能の詳細な説明はホワイトペーパー『ソリッドステートドライブ - FAQ』を参照ください。

本書で取り上げている PCIe-SSD は、PRIMERGY および PRIMEQUEST サーバ用ストレージメディアの第 3 世代です。第 2 世代（ioDrive®2 PCIe-SSD）については、ホワイトペーパー『パフォーマンスレポート PCIe-SSDs ioDrive®2』のパフォーマンスに関する記述で説明しています。

リリースマトリックス

以下の表に記載されている発注可能な PCIe-SSD に対するサーバモデルの割り当ては、本書の発行時点で有効な組み合わせです。ここで PCIe-SSD ブートドライブとしてはリリースされていません。ただし、以下の表の注釈で示されているものは除きます。

サーバ	AIC	SFF
	#PCIe-SSDs P3700	#PCIe-SSDs P3700
PRIMERGY BX2560 M1	-	2 *)
PRIMERGY CX2550 M1	-	2
PRIMERGY CX2570 M1	-	2
PRIMEQUEST 2x00x2	8 - 32	-
PRIMERGY RX2530 M1	-	4
PRIMERGY RX2540 M1	-	8
PRIMERGY RX2560 M1	-	8
PRIMERGY RX4770 M1	-	4
PRIMERGY RX4770 M2	-	4
PRIMERGY TX2560 M1	-	8

*) *特定の OS で UEFI モードでブートすることができます。*

国または販売地域によっては、一部のコンポーネントが利用できない場合があります。

管理ソフトウェア

Intel® Solid-State Drive Data Center Tool をインストールしてから isdct コマンドでツールにアクセスすると、PCIe-SSD の（温度などの）多くの特性をコマンドラインレベルでモニタリングできるようになります。ただし、ServerView RAID Manager による PCIe-SSD のモニタリングの方がはるかに便利です。

測定方法

測定方法とディスク I/O パフォーマンスの基本については、「ディスク I/O パフォーマンスの基本」ホワイトペーパーで説明しています。

標準では、PRIMERGY および PRIMEQUEST サーバのディスクサブシステムのパフォーマンス測定は、既定の測定方法で行われます。この測定方法では、仕様に基づいて、実際のアプリケーションシナリオのハードディスクアクセスをモデル化します。

必要な仕様は次のとおりです。

- ランダムアクセス／シーケンシャルアクセスの比率
- リード／ライトアクセスタイプの比率
- ブロックサイズ（KB）
- パラレルアクセスの数（処理待ち I/O の数）

これらの仕様値を組み合わせたものを「負荷プロファイル」と呼びます。以下の 5 種類の標準負荷プロファイルを一般的なアプリケーションシナリオに割り当てることができます。

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

異なる負荷強度でパラレルにアクセスするアプリケーションをモデル化するには、「処理待ち I/O の数」を 1、3、8 と増やしていき、最終的に 512 まで引き上げます（8 以降は 2 の乗数で増加します）。

この資料の測定は、これらの標準負荷プロファイルに基づいています。

測定の主な結果は以下のとおりです。

- 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 [IO/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）

通常、シーケンシャルな負荷プロファイルでは「データスループット」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「トランザクションレート」が使用されます。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]*

本書では、ストレージメディア容量を表す場合のみ 10 のべき乗（1 TB = 10^{12} バイト）で表記しており、その他の容量やファイルサイズ、ブロックサイズ、スループットを表す場合は 2 のべき乗（1 MB/s = 2^{20} バイト/s）で表記しています。

ベンチマーク環境

本書で示すすべての測定は、次の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行いました。

SUT（System Under Test：テスト対象システム）	
ハードウェア	
モデル	PRIMERGY RX2530 M1 × 1 PRIMERGY RX2540 M1 × 1
プロセッサ	Xeon E5-2637 v3（3.5 GHz）× 2 Xeon E5-2698 v3（2.30 GHz）× 2 Xeon E5-2603 v3（1.60 GHz）× 2
ストレージ媒体	PCIe-SSD 800GB P3700 × 4 PCIe-SSD 1.6TB P3700 × 1 PCIe-SSD 2TB P3700 × 1 SSD PCIe3 800GB Main 2.5' H-P EP × 1 SSD PCIe3 1.6TB Main 2.5' H-P EP × 1 SSD PCIe3 2TB Main 2.5' H-P EP × 1
ソフトウェア	
BIOS	PRIMERGY RX2530 M1: R1.9.0 PRIMERGY RX2540 M1: R1.24.0
BIOS 設定	「パフォーマンス」タイプの測定： Intel Virtualization Technology = Disabled VT-d = Disabled Energy Performance = Performance Utilization Profile = Unbalanced CPU C6 Report = Disabled 「デフォルト」タイプの測定（BIOS のデフォルト設定）： Intel Virtualization Technology = Enabled VT-d = Enabled Energy Performance = Balanced Performance Utilization Profile = Even CPU C6 Report = Enabled
ファームウェア	システム： PRIMERGY RX2530 M1: iRMC S4 7.82F PRIMERGY RX2540 M1: iRMC S4 8.00d PCIe-SSD: 8DV1FJP5
オペレーティングシステム	Microsoft Windows Server 2012
オペレーティングシステム設定	電源計画： 「パフォーマンス」タイプの測定： Select a power plan = High performance; 「デフォルト」タイプの測定： Select a power plan = Balanced;
ドライバ	PCIe-SSD: IaNVMe 1.1.0.1004
管理ソフトウェア	Intel® Solid-State Drive Data Center Tool 2.2.2 ServerView RAID Manager 6.1.4
RAID アレイの初期化	-
ファイルシステム	NTFS
測定ツール	Iometer 2006.07.27
測定データ	32 GB の測定ファイル 測定ファイルの表記の一部には、フルサイズのパーティションが記載されています。
Iometer アクセスの調整	4096 バイトの整数倍に調整

国または販売地域によっては、一部のコンポーネントが利用できない場合があります。

測定結果

以下で「PCIe-SSD」という用語が使われている場合、「基本情報と製品データ」の項で記述されているモデルのみを指しています。PCIe-SSD P3700 シリーズの AIC と SFF の 2 つのフォームファクタで、パフォーマンスは異なりません。従って、このホワイトペーパーでの測定結果は、両方のフォームファクタに適用されませう。

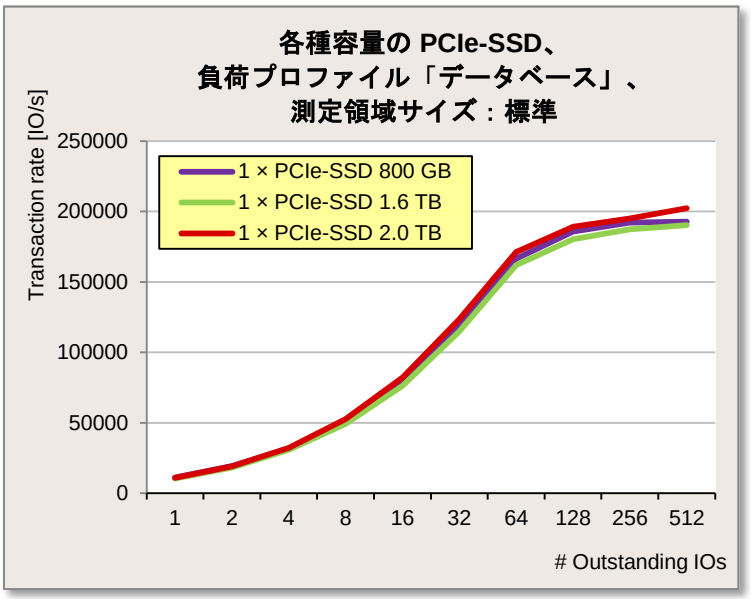
ここでは、「データベース」、「ファイルサーバ」、「ファイルコピー」、「ストリーミング」、「リストア」という 5 つの標準負荷プロファイルを使用します。プロファイルについては、PCIe-SSD のパフォーマンスを検証するための「測定方法」を参照してください。ストレージ媒体へのアクセス図でさまざまな負荷強度を考慮する場合、富士通では、「処理待ち IO の数」で負荷強度を指定します。負荷強度が低いアプリケーションの場合は処理待ち IO の数を 1 に設定し、負荷強度が高い場合は処理待ち IO の数を 512 に設定します。データメディアのパフォーマンスを具体的に表すには、サーバの最適な条件をベースとします。例えば、最大公称周波数（3.5 GHz）の CPU などです。パフォーマンスと CPU 周波数の関連については、適所に記載されます。

単体の PCIe-SSD

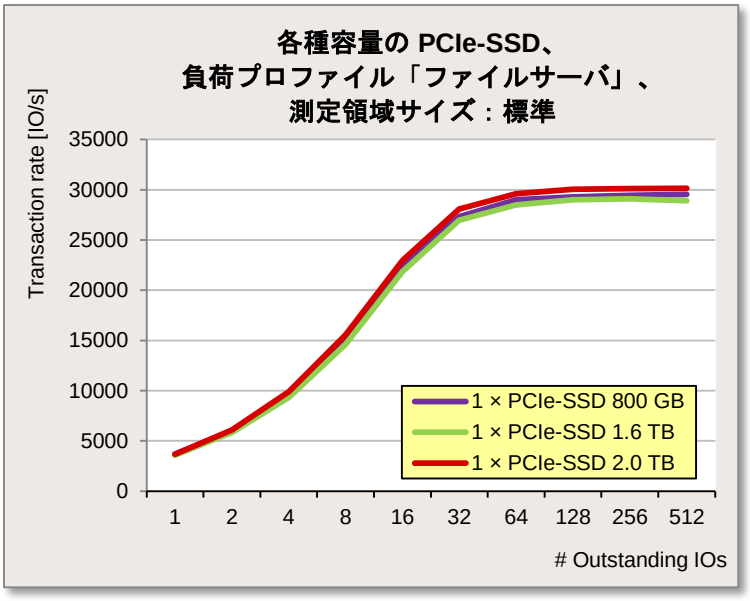
最初にランダムアクセスの負荷プロファイルを検証し、その後、シーケンシャルアクセスの負荷プロファイルを検証します。

ランダムアクセス

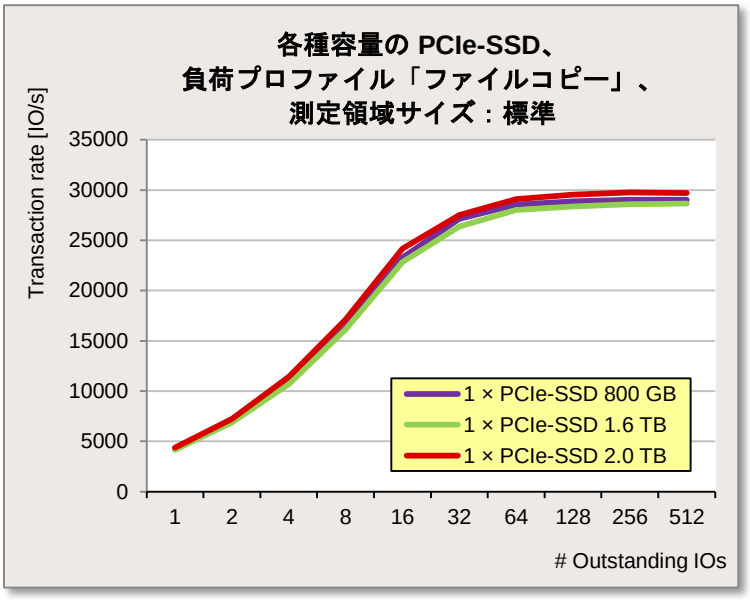
ランダムアクセスの場合、パフォーマンスの測定基準として IO/s 単位でトランザクションレートを示します。次の図は「データベース」負荷プロファイル（ランダムアクセス、67 %リード、8 KB ブロックサイズ）のトランザクションレートを示したものです。負荷強度が低い場合、PCIe-SSD のトランザクションレートは約 10000 IO/s です。負荷強度が高まるにつれ、トランザクションレートは非常に高い負荷強度において 190000～200000 IO/s 弱に到達するまで均等に上昇します。このような負荷プロファイルの場合、3 つの容量バージョンはパフォーマンスに関してほんのわずかしか異なりません。



次の図は「ファイルサーバ」負荷プロファイル（ランダムアクセス、67 %リード、64 KB ブロックサイズ）のトランザクションレートを示したものです。負荷強度が低い場合、PCIe-SSD のトランザクションレートは約 3600 IO/s です。負荷強度が高まるにつれ、トランザクションレートは非常に高い負荷強度において約 30000 IO/s に到達するまで均等に上昇します。また、このような負荷プロファイルの場合も、3 つの容量バージョンはパフォーマンスに関してほんのわずかしか異なりません。

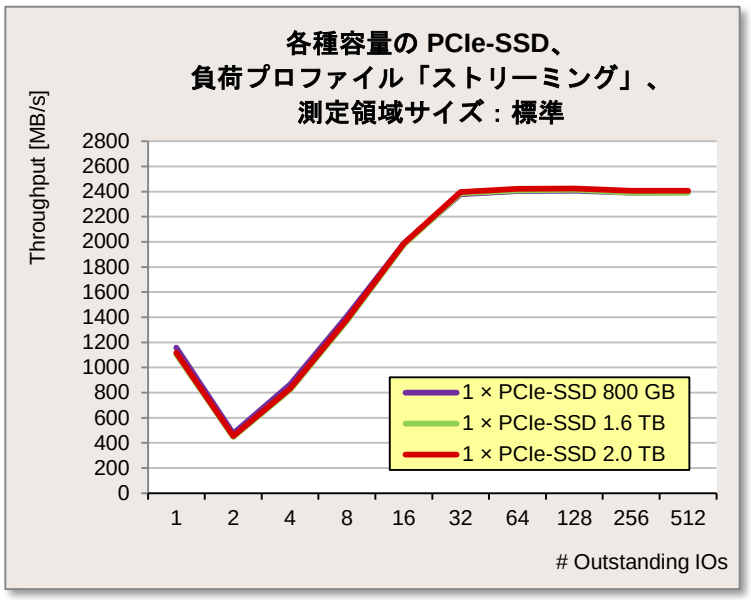


次の図は「ファイルコピー」負荷プロファイル（ランダムアクセス、50 %リード、64 KB ブロックサイズ）のトランザクションレートを示したものです。負荷強度が低い場合、PCIe-SSD のトランザクションレートは約 4200 IO/s です。負荷強度が高まるにつれ、トランザクションレートは非常に高い負荷強度において約 29000 IO/s に到達するまで均等に上昇します。このような負荷プロファイルの場合、3 つの容量バージョンはパフォーマンスに関してほんのわずかしか異なりません。



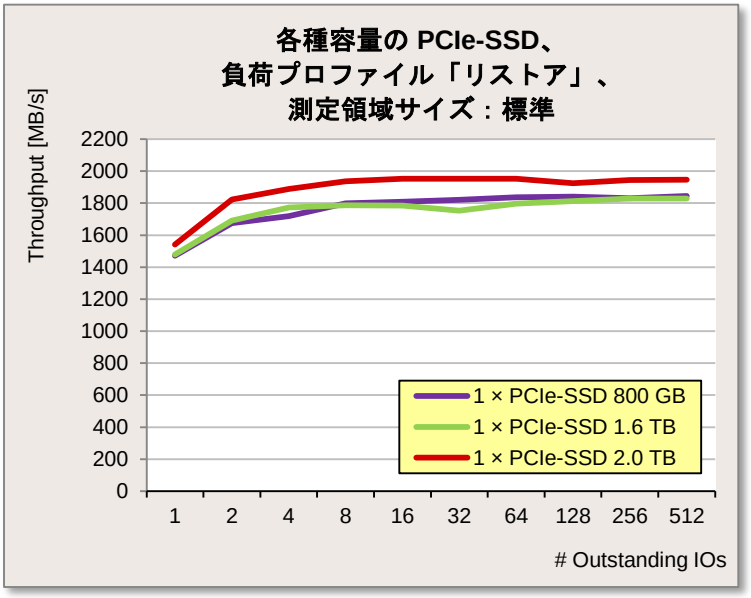
シーケンシャルアクセス

PCIe-SSD は、実稼動アプリケーションのシーケンシャルアクセス用途ではほとんど使用されません。これは、通常のハードディスクがこの用途で既に高いパフォーマンスを達成しているからです。しかし、完全を期すために、この資料ではこの負荷プロファイルについても説明します。シーケンシャルアクセスでは、トランザクションレートはパフォーマンスの測定基準として使用されず、スループットが MB/s 単位で示されます。



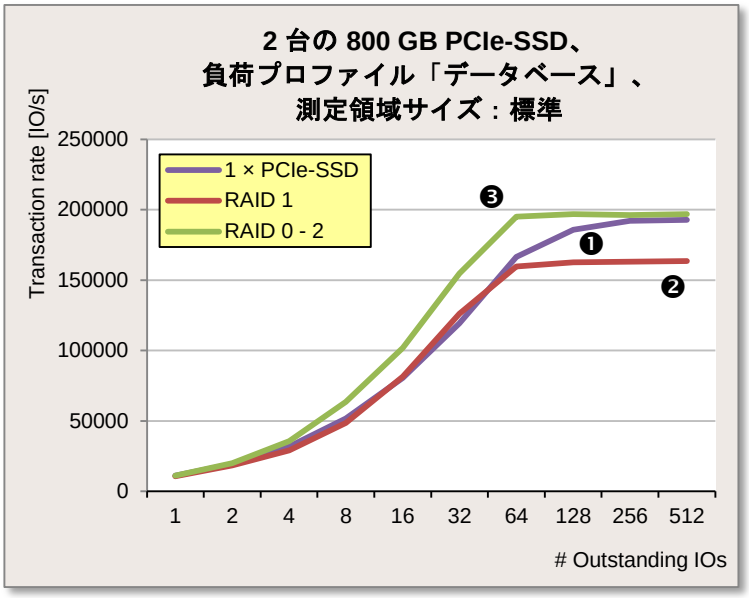
反対側の図は「ストリーミング」負荷プロファイル（シーケンシャルアクセス、100 %リード、64 KB ブロックサイズ）のスループットを示したものです。同期リード（処理待ち I/O = 1）の場合、PCIe-SSD のスループットは約 1100 MB/s に達します。また、処理待ち I/O = 2 の場合、スループットは約 450 MB/s になります。負荷強度が処理待ち I/O = 32 まで上昇すると、スループットは 2400 MB/s まで上昇しますが、負荷強度がさらに高くなってもスループット値は変わりません。このような負荷プロファイルの場合、3 つ容量のバージョンはパフォーマンスに関しても異なりません。

次の図は「リストア」負荷プロファイル（シーケンシャルアクセス、100 % ライト、64 KB ブロックサイズ）のスループットを示したものです。負荷強度が低い場合、容量の少ない 2 つのバージョンでは、約 1470 MB/s のデータスループットに達し、負荷強度が高くなるにつれて約 1830 MB/s まで徐々に上昇します。容量の最も多いバージョンは、この負荷プロファイルではややスループットが上昇します。負荷強度が低い場合、スループットは 約 1540 MB/s で、負荷強度が高くなるにつれて約 1950 MB/s まで上昇します。

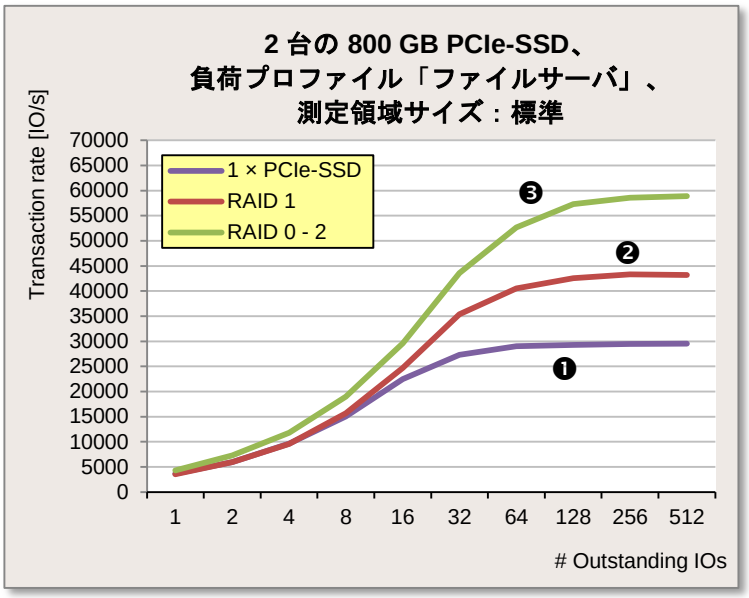


2 つの PCIe-SSD

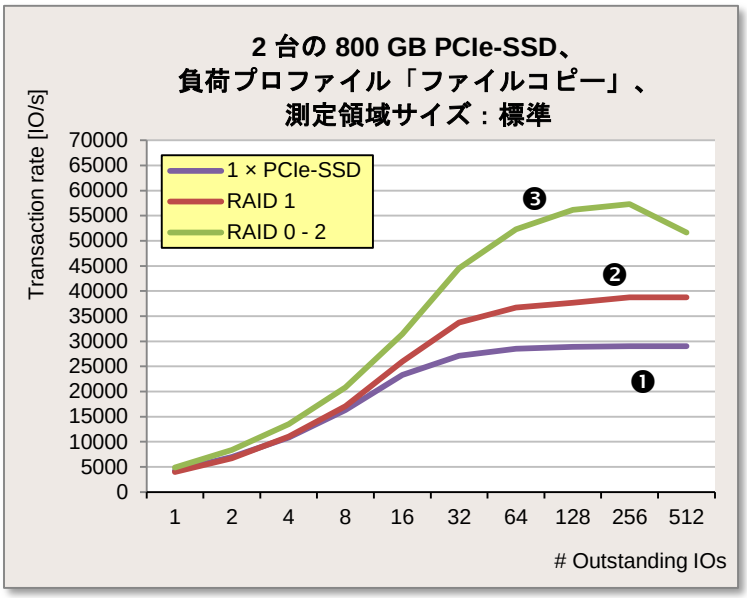
他の論理ドライブと同様に、RAID アレイは、オペレーティングシステムレベルで PCIe-SSD から構築することもできます。フェールセーフを高めるには、RAID 1 を使用します。ライトプロファイルにそれほど特化していない負荷プロファイルでは、この RAID レベルでアクセスのリード比率のパフォーマンスを高めることもできます。フェールセーフよりもパフォーマンスの高さを重視するアプリケーションでは、2 つの PCIe-SSD から RAID 0 アレイを構築することもできます。ここからは、単体の PCIe-SSD（ここでは 1.2 TB バージョンを例として使用）におけるさまざまな負荷強度に対応する RAID 構成を、5 つの標準負荷プロファイルでそれぞれ比較します。



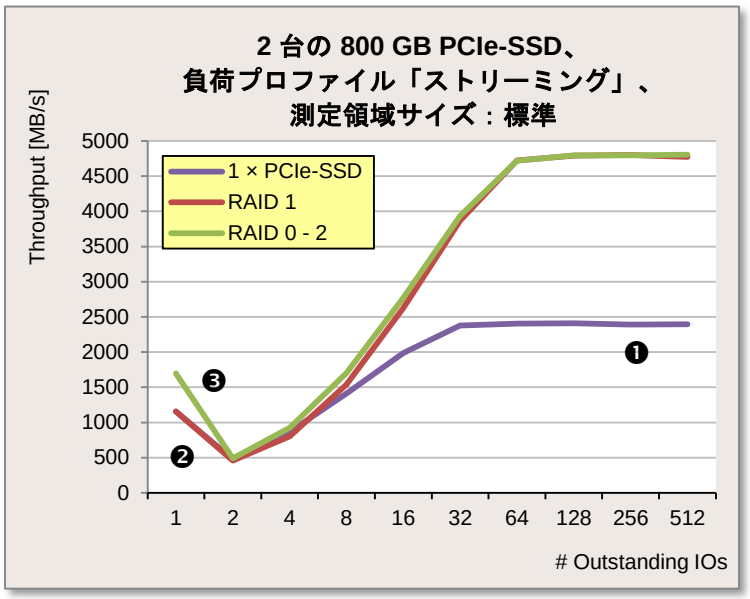
「データベース」負荷プロファイル（ランダムアクセス、67 % リード、8 KB ブロックサイズ）の場合、最低の負荷強度では、RAID 1 アレイ（❷）と RAID 0 アレイ（❸）は単体 PCIe-SSD（❶）と同じトランザクションレート、約 10000 IO/s になります。負荷強度が処理待ち IO = 512 まで上昇すると、トランザクションレートは次第に離れた値になります。単体 PCIe-SSD（❶）は約 193000 IO/s、RAID 1 アレイ（❷）は約 163000 IO/s、RAID 0 アレイ（❸）は約 197000 IO/s まで上昇します。



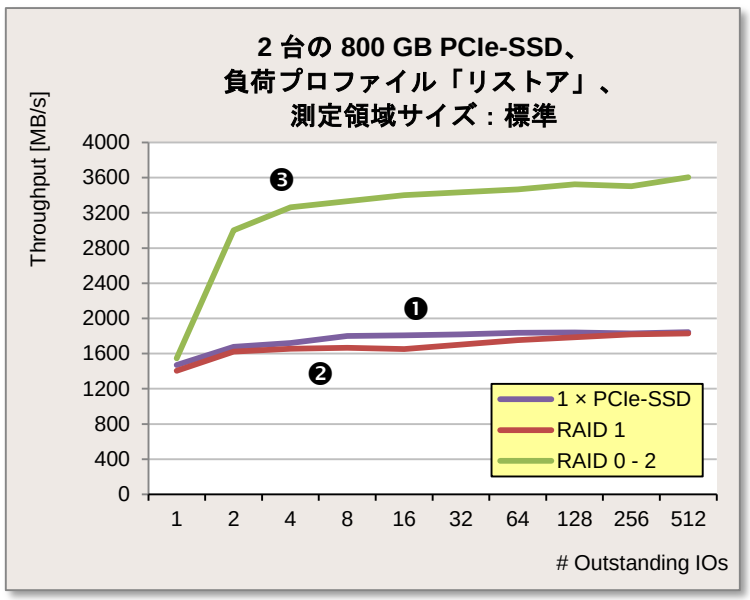
「ファイルサーバ」負荷プロファイル（ランダムアクセス、67 % リード、64 KB ブロックサイズ）の場合、RAID 1 アレイと単体 PCIe-SSD（❷ と❶）のトランザクションレートは、最低の負荷強度において約 3600 IO/s で、RAID 0 アレイ（❸）では 4300 IO/s に達します。負荷強度を高くすると、すべてのトランザクションレートは上昇し続けます。最大の負荷強度では、RAID 0 アレイ（❸）は約 59000 IO/s、RAID 1 アレイ（❷）は約 43000 IO/s、単体 PCIe-SSD（❶）は約 29000 IO/s に到達します。RAID 1 のトランザクションレートは、ほぼすべての負荷強度において、単体 PCIe-SSD と RAID 0 のトランザクションレートの中間に位置します。



「ファイルコピー」負荷プロファイル（ランダムアクセス、50 %リード、64 KB ブロックサイズ）の場合も、「ファイルサーバ」負荷プロファイルと同様の結果を示します。RAID 1 アレイ（❷）は最低の負荷強度において約 4000 IO/s に達し、単体 PCIe-SSD と RAID 0 アレイ（❶ と ❸）はそれぞれ 4300 IO/s と 4900 IO/s に達します。負荷強度を高くすると、これらの 3 つの場合のすべてで、大部分のトランザクションレートが上昇し続けます。負荷強度が高い場合、RAID 0 アレイ（❸）は約 57000 IO/s、RAID 1 アレイ（❷）は約 39000 IO/s、単体 PCIe-SSD（❶）は約 29000 IO/s に到達します。RAID 1 のトランザクションレートは、ほぼすべての負荷強度において、単体 PCIe-SSD と RAID 0 のトランザクションレートの中間に位置します。



「ストリーミング」負荷プロファイル（シーケンシャルアクセス、100 %リード、64 KB ブロックサイズ）の場合、両方の RAID アレイ（❷ と❸）および単体 PCIe-SSD（❶）は、スループットが最低の状態で約 500 MB/s（処理待ち I/O = 2）になります。負荷強度が処理待ち I/O = 512 まで上昇すると、RAID 0 は最大 4800 MB/s に達し、単体 PCIe-SSD は最大 2400 MB/s に達します。RAID 1 のスループットは、処理待ち I/O = 8 まで単体 PCIe-SSD と同様の動きを見せ、処理待ち I/O = 約 16 からは RAID 0 とほぼ同じスループットとなります。

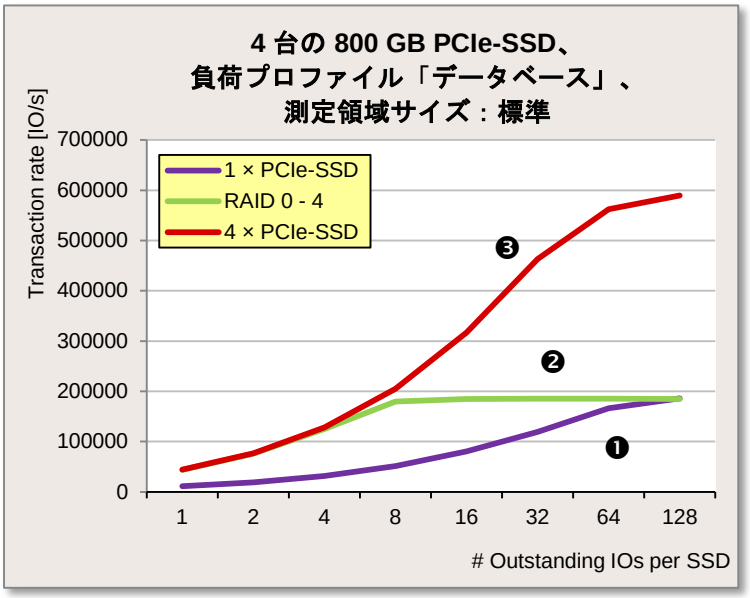


「リストア」負荷プロファイル（シーケンシャルアクセス、100 %ライト、64 KB ブロックサイズ）の場合、単体 PCIe-SSD（❶）、RAID 0、および RAID 1 の 3 つの場合のスループットは、処理待ち I/O = 1 で 1400～1550 MB/s になり、ほとんど差がありません。負荷強度がこれより高くなると、RAID 0（❸）のスループットは他の 2 つの場合とは明らかに異なります。RAID 0 の場合、処理待ち I/O = 4 ではすでに 3200 MB/s を上回り、最大値は約 3600 MB/s になります。他の 2 つ（❶ と ❷）の場合、スループットは処理待ち I/O = 1 の場合と比較して若干上昇します。最大値はいずれの場合も約 1800 MB/s です。

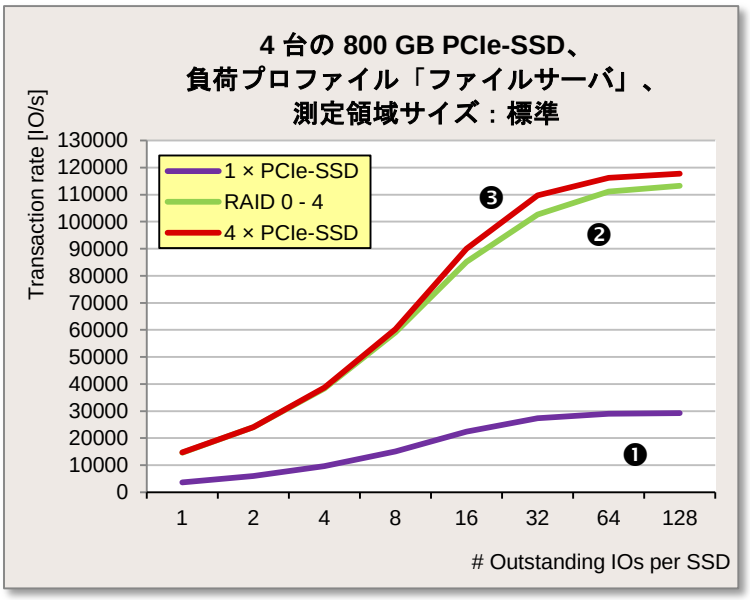
4 つの PCIe-SSD

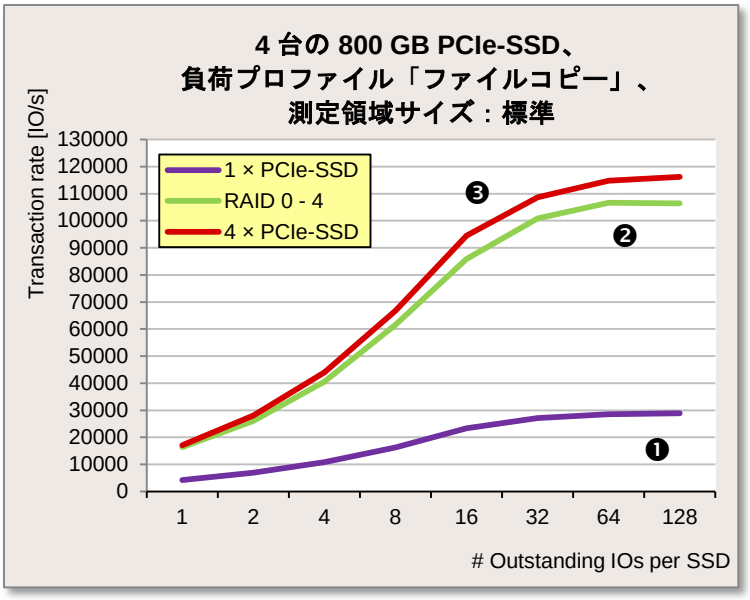
現在の PRIMERGY サーバおよび PRIMEQUEST サーバ（例えば、PRIMERGY RX2540 M1 など）のほとんどには、4 台以上の PCIe-SSD を搭載できます。このような場合は、すべての PCIe-SSD から RAID アレイが必ず構築されるわけではありません。ディスク I/O を作成する複数のプロセスまたはアプリケーションがあり、それらを個別の PCIe-SSD で実行することを希望する場合がよくあります。このような場合、利用可能なディスク I/O パフォーマンスは、サーバ内の PCIe-SSD の数に応じて異なります。一例として、以下の 5 つの図では、単体 PCIe-SSD と 4 台の PCIe-SSD のパフォーマンスを比較しています。4 台の PCIe-SSD は、別個の負荷として考えた場合、4 台の単体メディアであると見なされるのに対し、RAID 0 アレイはオペレーティングシステム RAID を介したものであると見なされます。後者の 2 つの場合を公平に比較できるよう、（この項の図の右軸に示された「SSD あたりの処理待ち I/O 数」という見出しのように）それぞれの負荷強度を同じ状態にして、関連する各 PCIe-SSD を比較しています。

「データベース」負荷プロファイル（ランダムアクセス、67 %リード、8 KB ブロックサイズ）の場合、単体 PCIe-SSD（❶）のトランザクションレートは、最低の負荷強度において約 11000 IO/s で、最大の負荷強度におけるトランザクションレートは約 186000 IO/s です。4 台の PCIe-SSD（❸）の負荷がそれぞれ異なる場合、対応するトランザクションレートは 44600 IO/s（倍率 4.0）および 589000 IO/s（倍率 3.17）です。4 台の PCIe-SSD（❷）で構成されるオペレーティングシステム RAID の場合、最低の負荷強度におけるトランザクションレートは、負荷がそれぞれ異なる 4 台の PCIe-SSD（□）とほぼ同じ 43800 IO/s です。最大の負荷強度におけるトランザクションレートは、単体 PCIe-SSD（❸）とほぼ同じ 185000 IO/s です。そのため、4 台の PCIe-SSD で構成されるオペレーティングシステム RAID は（2 台の PCIe-SSD の場合と同様）このようにブロックサイズの小さい単体 PCIe-SSD（❶）のトランザクションレートの最大値を上回ることはありません。

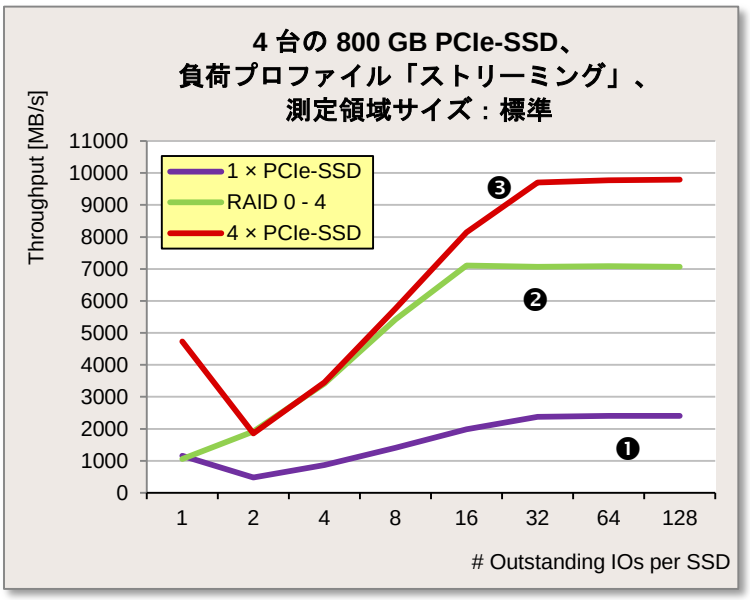


「ファイルサーバ」負荷プロファイル（ランダムアクセス、67 %リード、64 KB ブロックサイズ）の場合、負荷がそれぞれ異なる 4 台の PCIe-SSD（❸）の最低の負荷強度におけるトランザクションレートは、約 14700 IO/s（倍率 3.99）で、最大の負荷強度におけるトランザクションレートは約 118000 IO/s（倍率 4.02）です。4 台の PCIe-SSD（❷）で構成される RAID アレイでは、負荷強度が低い場合のトランザクションレートは 14600 IO/s ですが、負荷強度が最大になると 113000 IO/s まで上昇し、負荷がそれぞれ異なる 4 台の PCIe-SSD の場合とほぼ同じになります。

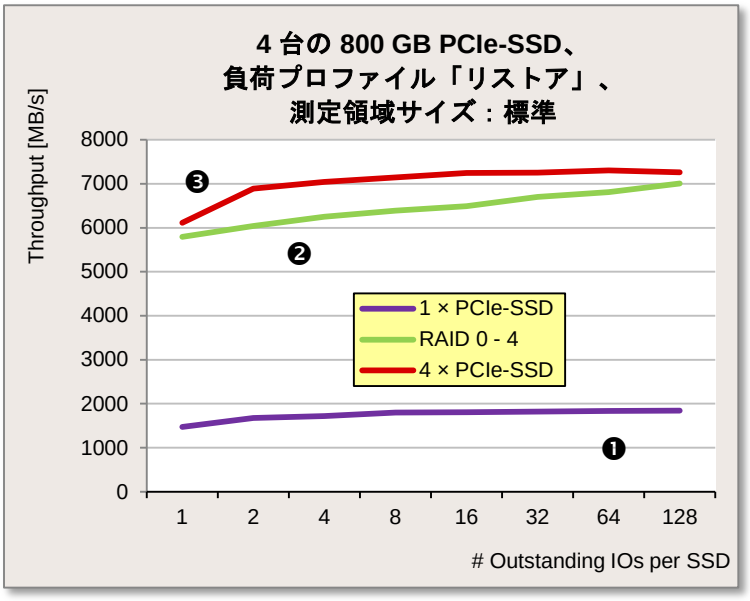




「ファイルコピー」負荷プロファイル（ランダムアクセス、50 %リード、64 KB ブロックサイズ）の場合、負荷がそれぞれ異なる 4 台の PCIe-SSD（❸）の最低の負荷強度におけるトランザクションレートは、約 17000 IO/s（倍率 4.00）で、最大の負荷強度におけるトランザクションレートは約 116000 IO/s（倍率 4.02）です。4 台の PCIe-SSD（❷）で構成される RAID アレイでは、負荷強度が低い場合のトランザクションレートは約 16400 IO/s ですが、「ファイルサーバ」負荷プロファイルの場合のように、負荷強度が最大になると 106000 IO/s まで上昇し、負荷がそれぞれ異なる 4 台の PCIe-SSD の場合とほぼ同じになります。



「ストリーミング」負荷プロファイル（シーケンシャルアクセス、100 %リード、64 KB ブロックサイズ）の場合、負荷がそれぞれ異なる 4 台の PCIe-SSD（❸）では、再びほぼ 100 %に近い倍率になります。SSD あたりの処理待ち I/O = 1 のリードの場合、4 台の PCIe-SSD のデータスループットは約 4700 MB/s（倍率 4.08）に達します。SSD あたりの処理待ち I/O = 2 の場合、4 台の PCIe-SSD のデータスループットは約 1860 MB/s（倍率 3.90）です。SSD あたりの処理待ち I/O = 32 まで負荷強度を高くすると、データスループットは 9700 MB/s まで上昇し、負荷強度がさらに高くなってもスループット値は変わりません（SSD あたりの処理待ち I/O = 128 で倍率 4.07）。オペレーティングシステム RAID（❷）のスループット曲線は、この中間あたりに位置し、最大値は 7100 MB/s です。



「リストア」負荷プロファイル（シーケンシャルアクセス、100 %ライト、64 KB ブロックサイズ）の場合、負荷がそれぞれ異なる 4 台の PCIe-SSD（❸）のスループットは、SSD あたりの処理待ち I/O = 1 で約 6100 MB/s に達し、SSD あたりの処理待ち I/O = 128 で 7264 MB/s まで上昇します。倍率はほとんどの場合、約 4.0 です。オペレーティングシステム RAID（❷）のスループット曲線は、他の 2 つの曲線の中間に位置し、最大で約 7000 MB/s に達します。

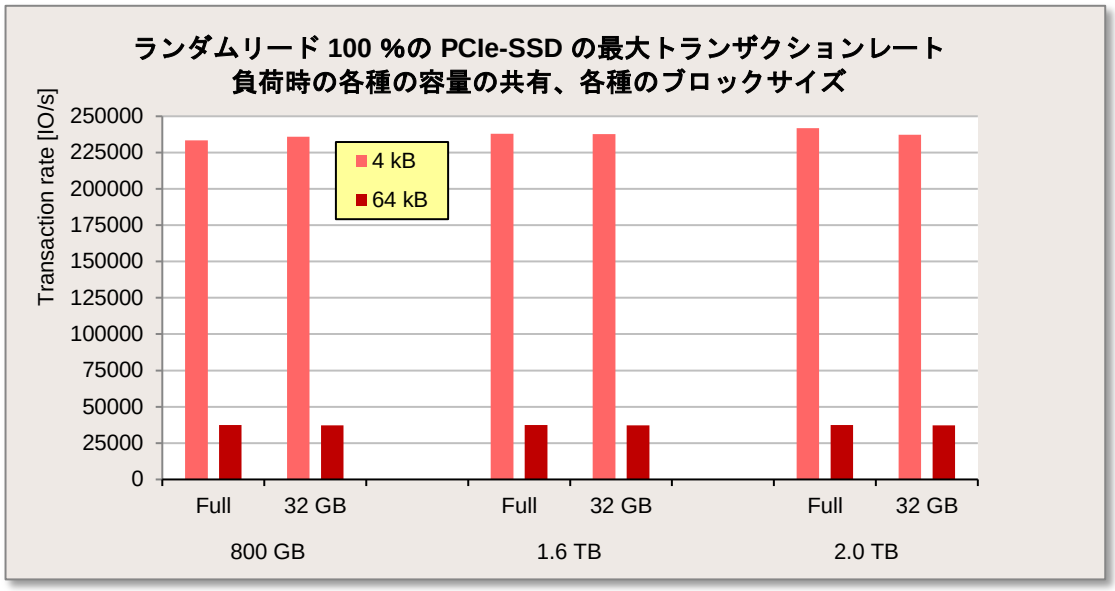
負荷時のストレージ容量の共有

ほとんどの場合、一定時間内に発生する PCIe-SSD へのアクセスの大半は、利用可能なストレージアドレス（ホットスポット）のうち比較的小さい下部領域に集中します。例：衣料品の通販会社では、注文処理データベースへのアクセスの多くが季節のファッションに関する記事を参照するためのものです。このようなアクセスの集中は、測定方法（「測定方法」の項を参照）に従って使用され、その結果として前項での基本前提にもなったサイズの小さい測定ファイルによってモデル化されます。

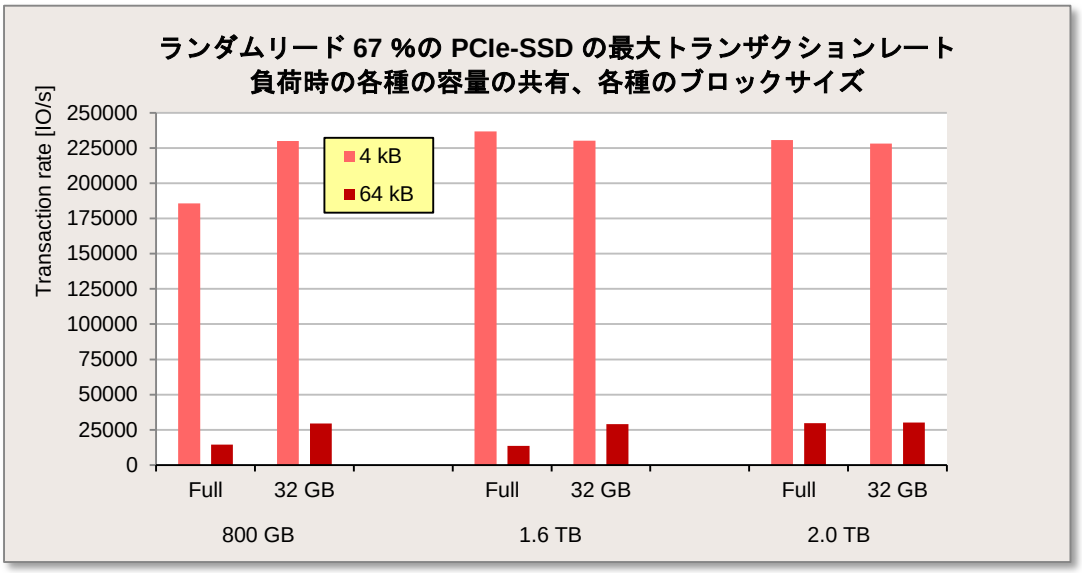
ただし、SSD の利用可能なすべてのストレージ容量も同様にアクセスの集中による負荷が発生するような場合は、パフォーマンスという観点から見て最悪の状況であると言えます。このような状況は、通常、データシートで対応します。PCIe-SSD のパフォーマンスに影響を与える「負荷時のストレージ容量の共有」という可変要因の重要性は、各種の負荷プロファイルに関する以下の 3 つの図で表されています。

シーケンシャル負荷プロファイルの場合、負荷時のストレージ容量の共有のサイズが変わっても、パフォーマンスに大きな違いは生まれません。そのため、以下の 3 つの図では、各種のライト共有のランダム負荷プロファイルの比較に限定しています。いずれの場合にも、これらの図ではストレージ容量全体で負荷が均等な PCIe-SSD と、測定方法によると負荷が 32 GB という小さい下部領域の PCIe-SSD を比較しています。

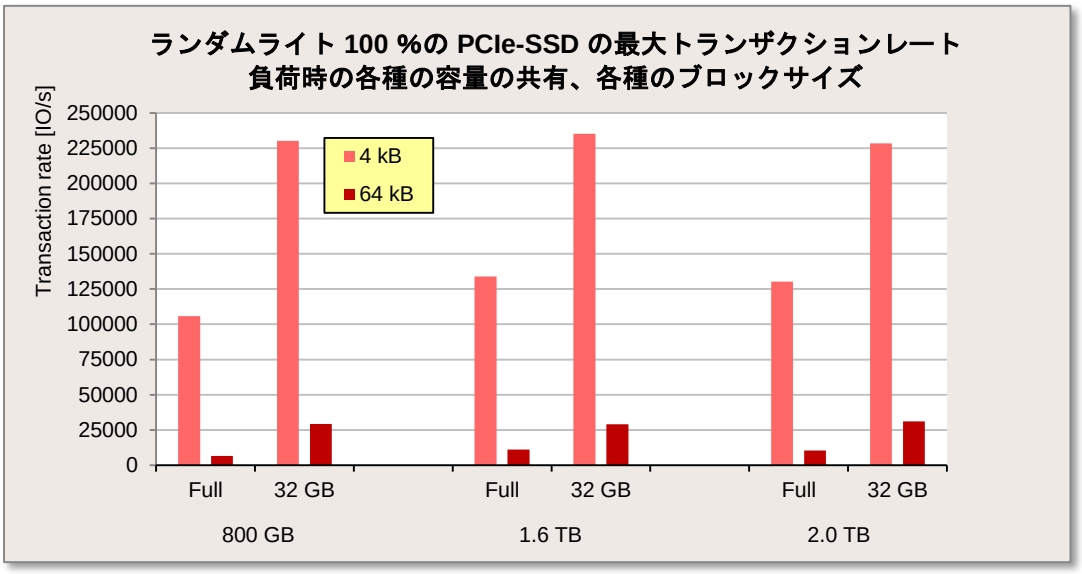
最初の図は、「ランダムアクセス、100 % リード」の例です。この場合、3 つのすべてのストレージ容量とすべてのブロックサイズにおいて、負荷時のストレージ容量の共有サイズが変わっても、パフォーマンスには顕著な違いは見られません。



次の図は、「ランダムアクセス、67 %リード」の例です。この例の 800 GB バージョンの場合、すべてのブロックサイズにおける負荷時の下部領域のサイズによって大きな違いが見られます。負荷時の領域のサイズによる顕著な違いは、1.6 TB バージョンのブロックサイズが大きい場合にのみ見られます（この例では 64 KB）。



最後の図は、「ランダムアクセス、100 %ライト」の例です。負荷時の領域のサイズによる大きな違いは、この例のようにすべての容量バージョンで見られます。



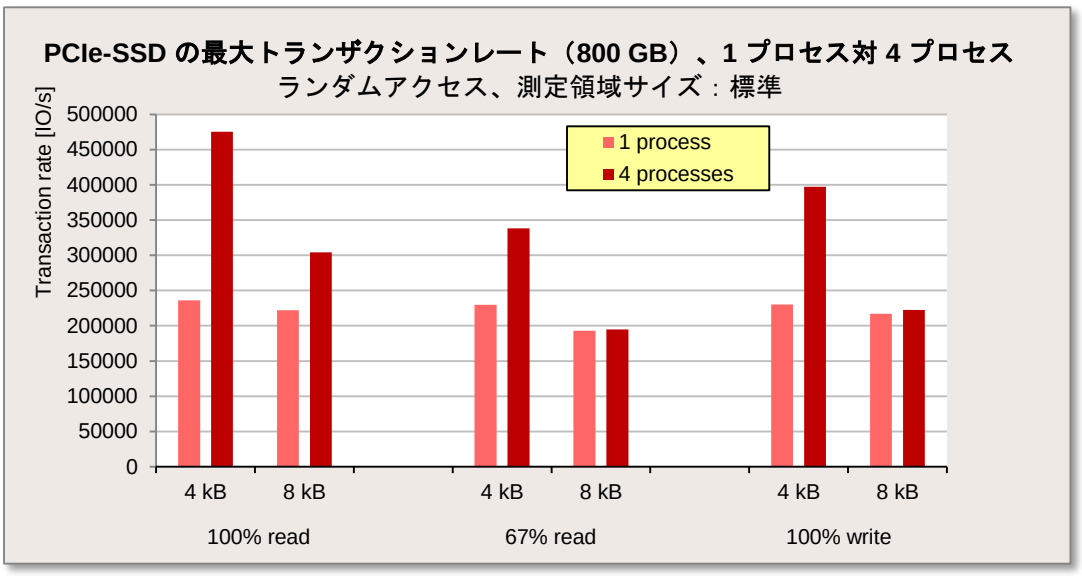
要約：

負荷時の SSD 下部領域のサイズが小さい場合、PCIe-SSD のライト共有によるランダムアクセスにおけるトランザクションレートは最高になります。ストレージ領域全体で負荷が均一の場合、PCIe-SSD のパフォーマンスの低下が大きくなるほど、ライト共有率が高くなり、アクセスのブロックサイズが大きくなります。このようなデメリットは、大容量のバージョンではあまりはっきり見られません。

複数のプロセスによる PCIe-SSD へのアクセス

「負荷時のストレージ容量の共有」の項では、PCIe-SSD のパフォーマンスレベルが最悪の状態（単一のプロセスによってストレージ容量全体の負荷が均一な状態）を示しています。これとは対照的に、本書で使用されている標準の測定方法では、小さい下部領域へのアクセスの集中をモデル化することで、「均一負荷」の場合に比べて、パフォーマンス値が大幅に向上します。これらのパフォーマンス値は、それだけでも十分に高いと言えますが、ストレージ容量のごく一部に負荷を集中させることに加え、複数のサーバプロセス（例えば、複数のアプリケーションなど）から PCIe-SSD の個々の下部領域にアクセスすることで、小さいブロックサイズではさらに大幅に向上します。

以下の例では、複数のアプリケーションまたはアプリケーションインスタンス（図中では通常、「プロセス」と呼ばれる）が PCIe-SSD の専用の個別パーティションにアクセスするアプリケーションシナリオをモデル化しています。標準の測定方法では、各パーティションの小さい下部領域（32 GB）へのアクセスの集中もモデル化しています。アプリケーションのアクセスは、小さいブロックサイズでのランダムアクセスです。さらに、以下の図に示されているトランザクションレートに到達するには、高周波数 CPU（この場合は、3.5 GHz 公称周波数）が必要です。この例は誇張されており、実際には達成することが困難なトランザクションレートを示しています。それは、多くの I/O に必要な特権的 CPU 時間だけでもサーバ CPU のパフォーマンス全体のかなりの部分を占めることになるからです。それでもなお、この例では対象となる PCIe-SSD の重量なオプションを示しています。ここに示されたトランザクションレートは、それぞれのアクセスパターンのすべての処理待ち I/O における最大値です。



この例では、小さいブロックでのランダムアクセスに 1 プロセスではなく 4 プロセスを使用することで、パフォーマンスの大幅な向上が可能であることを示しています。4 KB では最大で約 100 %の向上（100 %リードで最も顕著に見られる）、8 KB でも 100 %リードで約 37 %の向上。

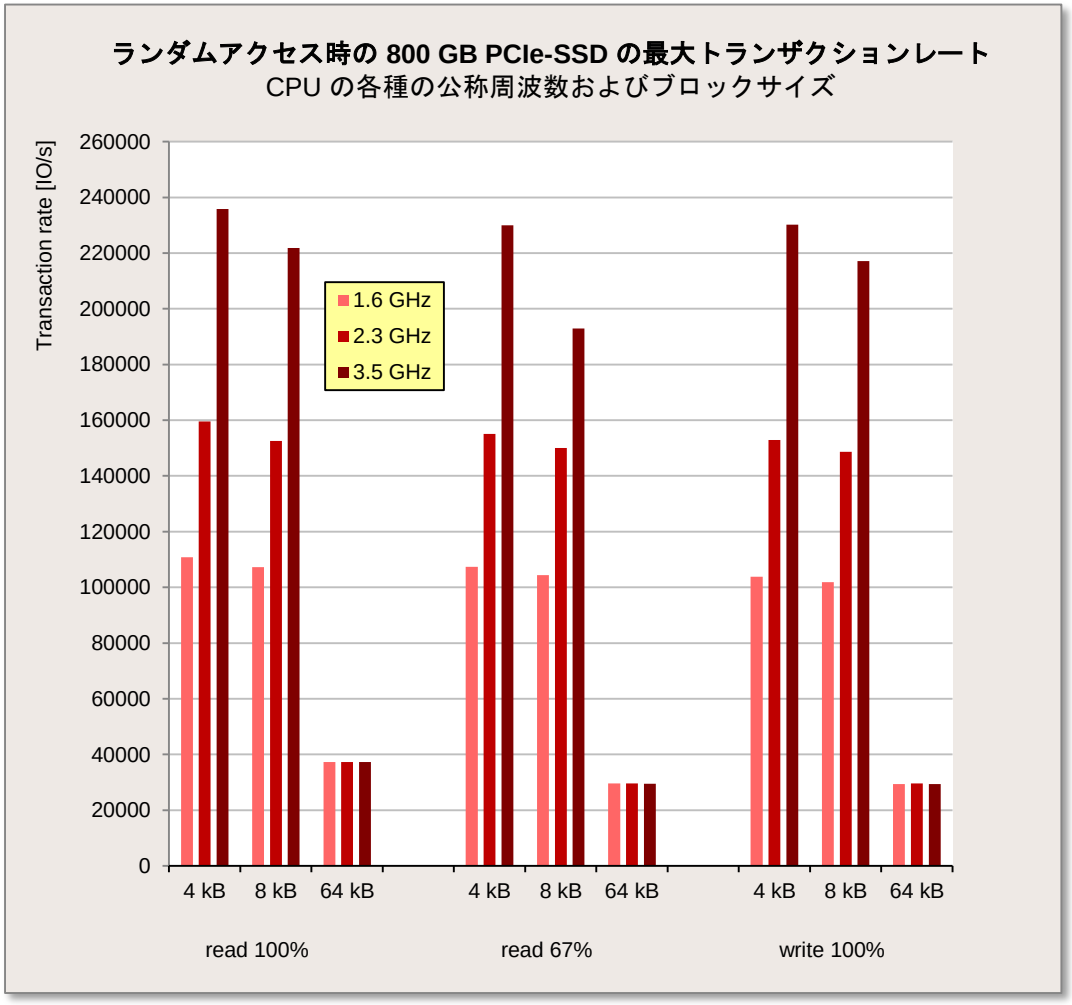
ベストプラクティス

サーバのすべての I/O コンポーネントと同様に、PCIe-SSD もサーバのハードウェア条件に依存します。特にプロセッサ、メインメモリ、および PCIe バスはここで重要な役割を果たします。サーバの選択、構成、および設定が適切な場合は、PCIe-SSD のパフォーマンスが改善されます。

現在のプロセッサ周波数の最適化

実際のプロセッサ周波数は、PCIe-SSD のパフォーマンスに影響を与えます。PRIMERGY または PRIMEQUEST システムにできる限り高い公称周波数のプロセッサタイプが搭載されていることが理想的です。しかし、最近のプロセッサはターボモードや省電力機能などのテクノロジーを使用して、公称周波数と比較して実際の周波数を一時的に増減できるため、公称周波数自体はそれほど重要ではありません。削減は、演算処理要件の低いアプリケーションなどで発生します。アプリケーションで省電力よりもパフォーマンスが重要である場合は、この周波数の低減は設定を変更することで防止できます。次の例は、CPU 周波数の最適化により実現できるパフォーマンス効果を説明しています。

測定する最初の点は、高い公称周波数のプロセッサを選択することによる影響です。このような目的で、以下の図では現在のデュアルソケットサーバの PCIe-SSD は、800 GB であると見なされています。サーバには Xeon E5-2600 v3 ファミリの 2 台の CPU が搭載されており、適切な BIOS およびオペレーティングシステム設定（「ベンチマーク環境」の項を参照）によって最大のパフォーマンスに調整されます。この図では、選択した負荷プロファイルに対する、1.6 GHz、2.3 GHz、および 3.5 GHz のそれぞれの CPU のトランザクションレートを比較しています。

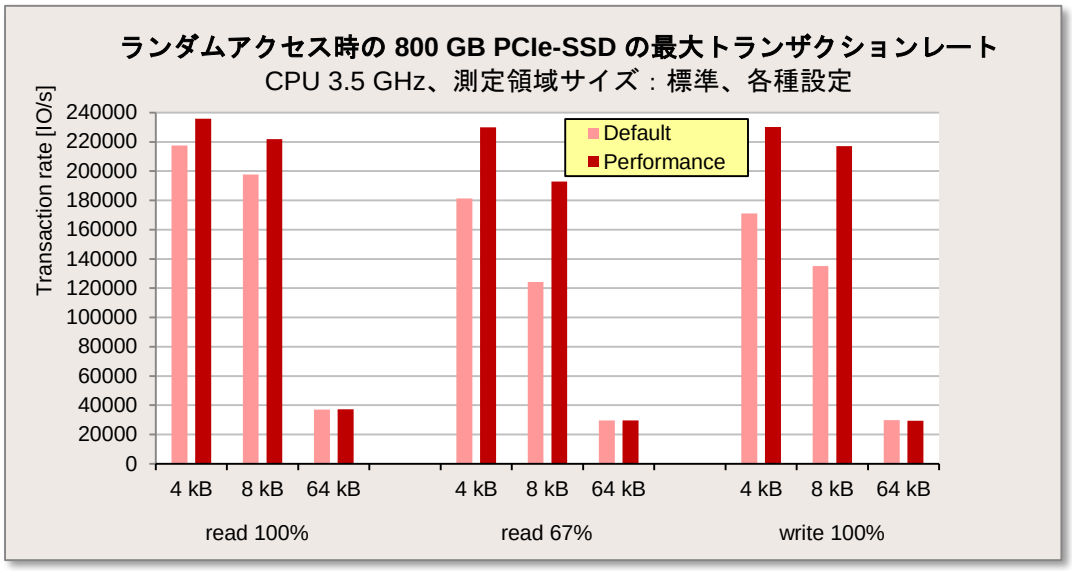


図は 8 KB までのブロックサイズでは、高い公称周波数によって大幅な改善が見られることを示しています。CPU の公称周波数は、ブロックサイズの大きい負荷プロファイルには影響しません。

また、その他の限界条件によって期待される IO/s 値に達することができる場合、高い CPU 周波数によるトランザクションレートの場合だけメリットがあります。例えば、「負荷時のストレージ容量の共有」の項で説明している点は重要です。

対象の CPU で検討すべき 2 番目の問題は、BIOS およびオペレーティングシステムの設定オプションと、現在の CPU 周波数への影響です。パフォーマンスまたは省エネに関連した設定オプションの詳細については、ホワイトペーパー『Xeon E5-2600 v3 搭載システムのための BIOS 最適化』を参照してください。前の図では、測定用の BIOS およびオペレーティングシステムの設定は、すべてパフォーマンスに設定されています。このような効果をもたらす設定セットは、ここでは集合的に「パフォーマンス」と呼ばれ、「ベンチマーク環境」の項で説明しています。この設定セットは、消費電力の増加と関連するため、デフォルトの BIOS およびオペレーティングシステム設定（この設定セットは、ここでは「デフォルト」と呼ばれます）によるパフォーマンスには問題があります。

次の図では、前述の 2 種類の設定での 800 GB PCIe-SSD の最大トランザクションレート（すべての処理待ち I/O での最大値）を比較することで、この質問に回答しています。



この図は、「パフォーマンス」設定セットでは、最大 8 KB のブロックサイズでかなりのライト共有を伴う負荷プロファイルでのトランザクションレートが「デフォルト」設定セットより高くなることを示しています。この前提条件は、120000 IO/s を上回るトランザクションレートがその他の限界条件（アクセスの集中）に関する限り可能な原則に基づいていることです。BIOS のデフォルト設定を使用する場合、高トランザクションレートを実現するには、Windows の電源プランで「高パフォーマンス」を選択するだけで十分です。

この項の 2 つの図のいずれの場合も、ランダムアクセスは単なる例として使用されているだけで、記述はシーケンシャルアクセスにも適用されます。さらに、PCIe-SSD の測定領域における CPU に関連した違いについて詳述する場合、いずれの図も高トランザクションレートの実現に必要とされる最適な前提条件を作成するのに不可欠です。そのため、2 つの図はいずれも「標準」の測定領域サイズに基づいているため、I/O アクセスはストレージ容量の小さい下部領域に集中しています。

他のストレージ媒体との比較

以下は、PCIe-SSD P3700 と従来のモデルとの比較のほか、SAS-12G インターフェースが搭載された現在利用可能な HDD および SSD との比較を示しています。

以下の表は、これまで使用してきた 5 つの標準負荷プロファイルに基づいて、4 つのストレージ媒体のパフォーマンス値を比較したものです。

処理待ち IO = 1～512 における最大パフォーマンス							
負荷プロファイル	SAS-12G-HDD 450 GB、15 krpm、2.5″ ディスクキャッシュ有効	SAS-12G-SSD MLC 400 GB 2.5″ ディスクキャッシュ有効	PCIe-SSD ioDrive®2 1.2 TB フォーマット「出荷時容量」	PCIe-SSD P3700 800 GB 1 プロセス、測定領域 32 GB	比率 PCIe-SSD P3700 ／ SAS-12G-HDD	比率 PCIe-SSD P3700 ／ SAS-12G-SSD	比率 PCIe-SSD P3700 ／ PCIe-SSD ioDrive®2
データベース	864 IO/s	64416 IO/s	49919 IO/s	192881 IO/s	223.2	3.0	3.9
ファイルサーバ	766 IO/s	8977 IO/s	19005 IO/s	29530 IO/s	38.6	3.3	1.6
ファイルコピー	789 IO/s	8157 IO/s	18018 IO/s	29033 IO/s	36.8	3.6	1.6
ストリーミング	237 MB/s	790 MB/s	1390 MB/s	2395 MB/s	10.1.	3.0	1.7
リストア	237 MB/s	419 MB/s	1037 MB/s	1844 MB/s	7.8	4.4	1.8

この表に示しているように、PCIe-SSD のパフォーマンス値は、負荷プロファイルによっては、先行世代の ioDrive®2 の約 3.9 倍に達し、SAS-12G-SSD のほぼ 4.4 倍、SAS-12G-HDD の 220 倍以上に達します。

次の表は、小さいブロックの読み取りおよび書き込み時の最小レイテンシ時間を比較しています。

負荷プロファイル	SAS-12G-HDD 450 GB、15 krpm、2.5″ ディスクキャッシュ有効	SAS SAS-12G-SSD MLC 400 GB 2.5″ ディスクキャッシュ有効	PCIe-SSD ioDrive®2 1.2 TB フォーマット「出荷時容量」	PCIe-SSD P3700 800 GB 1 プロセス、測定領域 32 GB
読み取りレイテンシ（1 KB シーケンシャル）	0.105 ms	0.21 ms	0.066 ms	0.041 ms
書き込みレイテンシ（1 KB シーケンシャル）	0.51 ms	0.11 ms	0.018 ms	0.016 ms

そのため、最小待ち時間は先行世代の ioDrive®2 と比べてさらに減少します。

結論

PCIe-SSDs P3700 は高度な不揮発性ストレージメディアで、小さいスペースにおいて、特にトランザクション数の多い I/O アクセスで非常に優れたパフォーマンスを実現できます。もう 1 つの革新は、SFF フォームファクタとしての可用性です。先行世代と比較して、最大トランザクションレートおよび最大データスループットのどちらにおいてもパフォーマンスは大幅に向上しています。1 つのアプリケーションインスタンス（プロセス）を使用して、メディアの 1 つの小さな部分にアクセスを集中させた場合、PCIe-SSD P3700 は、例えば、通常のデータベースのアクセスにおけるトランザクションレートは最大約 200000 IO/s、ビデオストリーミングなどのシーケンシャルリードアクセスにおけるデータスループットは最大 2490 MB/s を実現します。

小さいブロックサイズによる個々の例で実現できたトランザクションレートは、アクセスを集中させる PCIe-SSD の領域のサイズによって大きく左右されます。さらに、サーバの現在の CPU 周波数も最高のトランザクションレートの実現に大きく影響します。

各種のプロセスがメディアの個別の下部領域にアクセスすると、PCIe-SSD のパフォーマンス全体が大幅に向上する可能性もあります。例えば、4 つのプロセスが個別の下部領域にランダムアクセスをすると、最大 475000 IO/s のトランザクションレートが可能です。

オペレーティングシステムにおける RAID アレイの構成、またはこれらのストレージメディアの独立した稼働により、サーバごとのディスク I/O パフォーマンスはさらに向上することが可能です。複数の PRIMERGY モデルで、8 台の PCIe-SSD を稼働できます。

関連資料

PRIMERGY & PRIMEQUEST サーバ

http://www.fujitsu.com/jp/products/computing/servers/

PRIMEQUEST のパフォーマンス

http://www.fujitsu.com/jp/products/computing/servers/primequest/products/2000/benchmark/

PRIMERGY のパフォーマンス

http://jp.fujitsu.com/platform/server/primergy/performance/

コンポーネント別性能情報

このホワイトペーパー：

http://docs.ts.fujitsu.com/dl.aspx?id=e123e7ef-20e4-4a40-9083-876a2c106494
http://docs.ts.fujitsu.com/dl.aspx?id=daabec11-a857-4781-af72-64b7d86a194c
http://docs.ts.fujitsu.com/dl.aspx?id=5c4d91e4-3fa0-49f9-a92e-6d51dbcbd7bd

PCIe-SSD P3700 シリーズ

Datenblatt (EN)
http://docs.ts.fujitsu.com/dl.aspx?id=b332a27a-96a5-4dfc-8af2-5fc30de74226:

ソリッドステートドライブ - FAQ

http://docs.ts.fujitsu.com/dl.aspx?id=1d8b7d65-e5f4-4a99-8e7b-f47c74ccc85e

Xeon E5-2600 v3 搭載システムのための BIOS 最適化

http://docs.ts.fujitsu.com/dl.aspx?id=2009eb5b-f273-4f1f-94ef-07f1d0304255

ディスク I/O パフォーマンスの基本

http://docs.ts.fujitsu.com/dl.aspx?id=35801735-a223-491a-a879-43f506444366

Fusion-io® ioDrive®2 ソリッドステートストレージデバイス（先行世代）

データシート（英語）
http://docs.ts.fujitsu.com/dl.aspx?id=cb40337c-4292-44fe-ac8f-50a96bc653e1

パフォーマンスレポート PCIe-SSDs ioDrive®2
http://docs.ts.fujitsu.com/dl.aspx?id=91b93b35-7511-45e4-a5be-ccde9219772f

Iometer についての情報

http://www.iometer.org/

お問い合わせ先

富士通

Web サイト：http://www.fujitsu.com/jp/

PRIMERGY のパフォーマンスとベンチマーク

mailto:primergy.benchmark@ts.fujitsu.com

ioDrive®2 は SanDisk の登録商標です。

2016-03-24 WW JA