HDD 使用による
500MB/s 級、
1000MB/s(1GB/s)級
ストレージサーバ

エムアイシー･アソシエーツ株式会社

2009/03/18

ここに記載された内容は更新される可能性があります。この文書に記載されている内容はこの文書の発行時点におけるエムアイシー・アソシエーツ株式会社の見 解を述べたものです。エムアイシー・アソシエーツ株式会社が、この文書に記載された内容の実現に関して確約するものではありません。また発行日以降につい ては、この文書に記載された内容の正確さは保証しません。

この文書は情報の提供のみを目的としており、明示的または黙示的に関わらず、この文書の内容についてエムアイシー・アソシエーツ株式会社はいかなる保証をするものでもありません。

バージョン	作成日付	旧バージョンからの変更点	総ページ数
1.00	2009/03/18	初版発行	6

本書作成、編集、管理

エムアイシー・アソシエーツ株式会社
〒103-0004 東京都中央区東日本橋 3-12-12
櫻正宗東日本橋ビル 9F
Tel 03-5614-3757 Fax 03-5614-3752

2009/03/18

目　次

2009/03/18

はじめに

WEBサーバ、Eコマースサーバ、コンテンツ配信サーバなどにおいて非常に高速のランダムI/Oを要求するニーズは従来から存在していました。一方で非圧縮もしくは低圧縮（高品位、高レート）の映像を扱う分野や、高速で生成されるデータを記録しなければならない科学技術分野においてもシーケンシャル I/O で高いスループットを要求するニーズがありました。これらは一見高い性能を要求する点で似ていますが、ランダム I/O と シーケンシャル I/O と異なった性格の使用方法になります。

従来、一般的に入手可能なハードディスクによる記録装置ではディスク単体の性能から来るシステム上のボトルネックと対ホスト間I/Fのボトルネックが存在し、数100MB/sと言った帯域が実態である為に、これを上回る帯域を要求される場合には大量の DRAM を使用した半導体ディスク装置を選択する事が解決策として選択されて来ました。 HDD にはランダム I/O 性能がシーケンシャル I/O 性能より落ちる宿命がありますが、メモリを使用した記憶装置ではランダム I/O でもシーケンシャル I/O でも同じレベルの性能が得られます。しかし、これらの半導体ディスクは高速である反面、大容量化が困難であったり、価格面で苦しい立場に立たされるケースも多々見られました。

2008 年後半から民生用コンピュータ部品として急速な成長を見せているパーツとして SSD(Solid State Disk)があります。これは不揮発性メモリを搭載し、HDD と物理互換、容量的に HDD に迫る容量を備えたパーツであり、従来問題とされてきた繰り返し書き込みで発生するエラー率の上昇への対処（書込み位置の平均化）と価格の低減を図ったもので、以下の様なメリットを持っています。

- モーターを搭載しない為、省電力化と衝撃への耐性が高い
- メモリ使用のため高いランダム I/O 性能が期待できる
- Read 性能が高く、Write 性能についても次々と改良製品が発表されている

当初はバッテリー駆動を行うノート型 PC において、省電力化と耐衝撃性が注目され、その後デスクトップ、サーバ、ブレードサーバと適用製品は拡大を続けています。繰り返し書込みでのエラー率の懸念については、2009 年 1Q にブレードサーバメーカーが相次いで SSD 搭載モデルを発表したことで、PC 側の実用寿命年数をはるかに超えて使用を続けないとのエラー上昇は起こらないとする SSD メーカーの主張が受け入れられ始めたと見ることができます。

これに続き、2009 年 3 月には 米 FUSIONIO 社（http://www.fusionio.com/)がメモリベースの、 また OCZ Technology 社（http://www.ocztechnology.com/ ）が RAID コントローラと複数自社製 SSD を一体化した、PCI-Express バス直結の高速なディスク装置を相次いで発表しています。 これらは容量として1～1.5TB 、スループットとして 800MB/s 前後を公称しています。
コンシューマ向け/サーバ向けの 2.5”/3.5” HDD 互換 SSD は SATA/SAS の I/F（RAID コントローラ）を介して接続されますが、これらの新しい発想に基づく製品は PCI-Express スロットに直接実装することで、I/Fのボトルネックに縛られること無く、高スループットかつ低遅延のランダムI/Oに強いディスク装置が利用できる環境を提供することになります。

メモリディスクの得失

これらのメモリディスク装置は量産効果により、比較的低価格で高速の I/O デバイスとして提供される見込みです。旧来型の半導体ディスクや SATA SSD ベースの RAID に比べるとメリットが目立ちますが、全ての用途で万能ではありません。

○ ランダムアクセス時の遅延が小さい
○ シーケンシャル、ランダムアクセス共に性能が変わらない
○ 容量とニーズに合えば安価に入手できる
× PCI-Express バスに実装されているため、増設限界が早い
× 最大容量が現在の HDD に比べ小さい（転送レートに比べ記録可能時間が短い）

これらの特徴から導き出される最適のメモリディスク装置の用途は、小さいサイズのランダム I/O が大量に発生する（高い IOPS が要求される）WEB サーバ、E コマースサーバ、データベースサーバとなります。一方で、スループット上限近くでデータの記録/再生をする場合には依然として短時間の容量しか取れない弱点があります。

高スループットとデータ量

高速で大量のデータを扱う一例として映画、CG のマスターデータがあります。エフェクト処理、編集等が終わった後の上映版ﾃﾞｰﾀでは圧縮をかけるため、データは小さく（要求スループットも低く）なっていますが、圧縮データでは合成がきれいにできないため、マスター画像としては非圧縮を支持する声が依然として多くあります。高精細の映画フォーマットとして利用される 4Kx2K 10bit のデータは 796MB/s の帯域を必要とします。映画の上映時間を2時間としても、元データはその数倍から十倍程度は存在しているため、数 10TB の容量が必要となります。これにワークエリアを考えるとさらに容量が必要となります。
従来の HDD ベース装置では容量的な問題は無い反面、スループットが追いつかない為に、ダイレクトに HDD ベース装置からリアルタイムの記録・再生ができない問題がありました。メモリによる記録装置でも帯域的な余裕が無く、さらに容量が追いつかないと言う問題が発生します。

表1 4Kx2K 映像のデータ量

	1.5TB	10TB	20TB	40TB	80TB
4Kx2K 24P 10bit 796MB/s	31 分 (0.5 時間)	209 分 (3.49 時間)	418 分 (6.9 時間)	837 分 (13 時間)	1675 分 (27.9 時間)

映像配信を行うコンテンツメディアサーバは、多くのクライアントリクエストに答える為にランダム I/O になると思われがちですが、ある程度の大きさで連続的に読み出す動作が複数重なるため、かなりシーケンシャル I/O に近い性能が出る傾向が有ります。この場合には容量の上限からメモリベースの装置は不利となり HDD ベースの装置が使用されます。トータルスループットが足りない場合には（冗長性確保の面からも）サーバを複数立ち上げ、負荷分散を行うシステム構成が取られています。

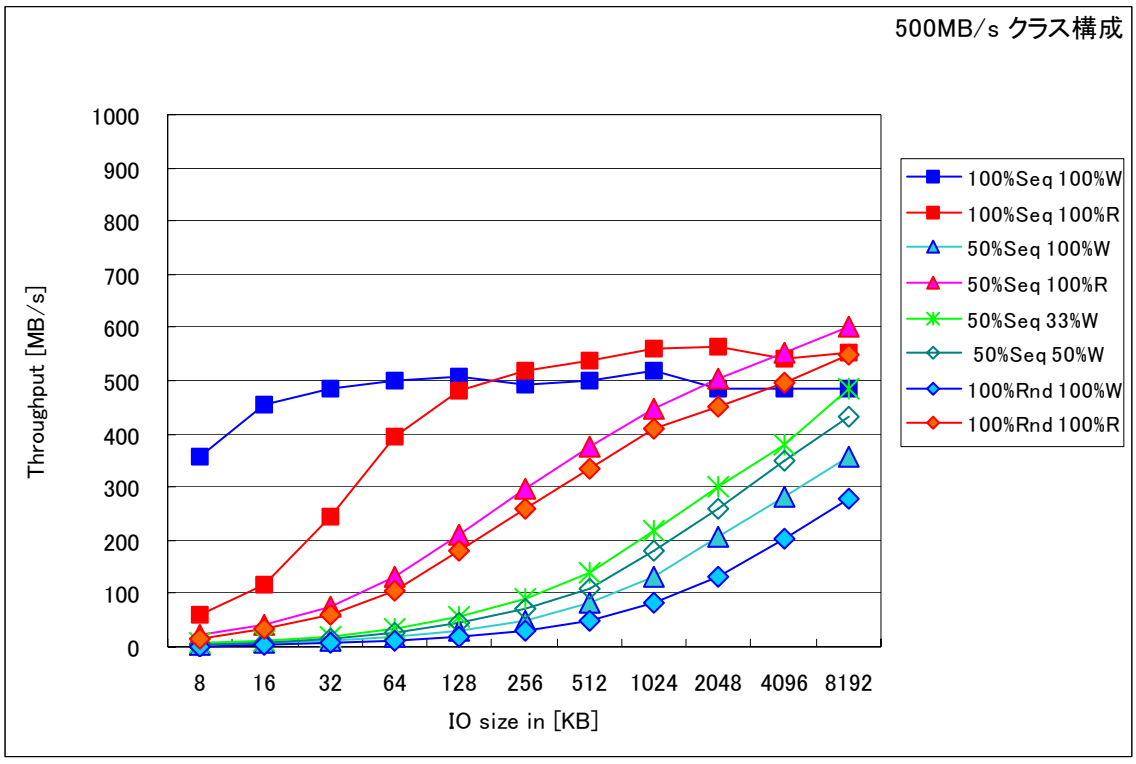
HDD で作る高スループット構成

メモリディスクと HDD はその特質にあった使い方をする事で、お互いを補完しあうべきです。 とは言え、メモリディスクに比べ非常に低速な HDD 装置ではバランスが取れません。
特に帯域上限近くでの大量データを扱う場合にはシーケンシャル I/O となるため、HDD でも実現が可能です。弊社ではその実現性を探るため、いくつかの実験を行いました。

500MB/s クラス構成

HS1235E-M 標準構成

Intel 製 Quad Core Xeon® をデュアル(max)搭載する強力なサーバ用マザーボードと、OEM ストレージサプライヤとして出荷数世界第 2 位(IDC 2007 年)の Xyratex 社製ストレージが一体になった HS1235-M に 1TB の SATA HDD を搭載した構成で、下記に示す様に 500MB/s の帯域をカバーする事が確認されました。ここで注目すべきは、このテストが高速 HDD の SAS モデルではなく、7500rpm の SATA HDD によって構成されている点にあります。
SAS より安価に大容量が確保できる SATA HDD は、回転数の違いから一般に SAS HDD よりも性能が低いと思われがちですが、用途と構成を間違えなければ（何でも SATA で代用可能とは言えません）SAS 構成に迫る性能を出すことができます。
拡張筐体を使用することで、標準構成の数倍の容量構成も可能です。



600MB/s クラス構成

F5404E による構成

先の HS1235E-M による構成とは別の機材を使用し、より多くの容量を積み上げられる事を重視した構成です。

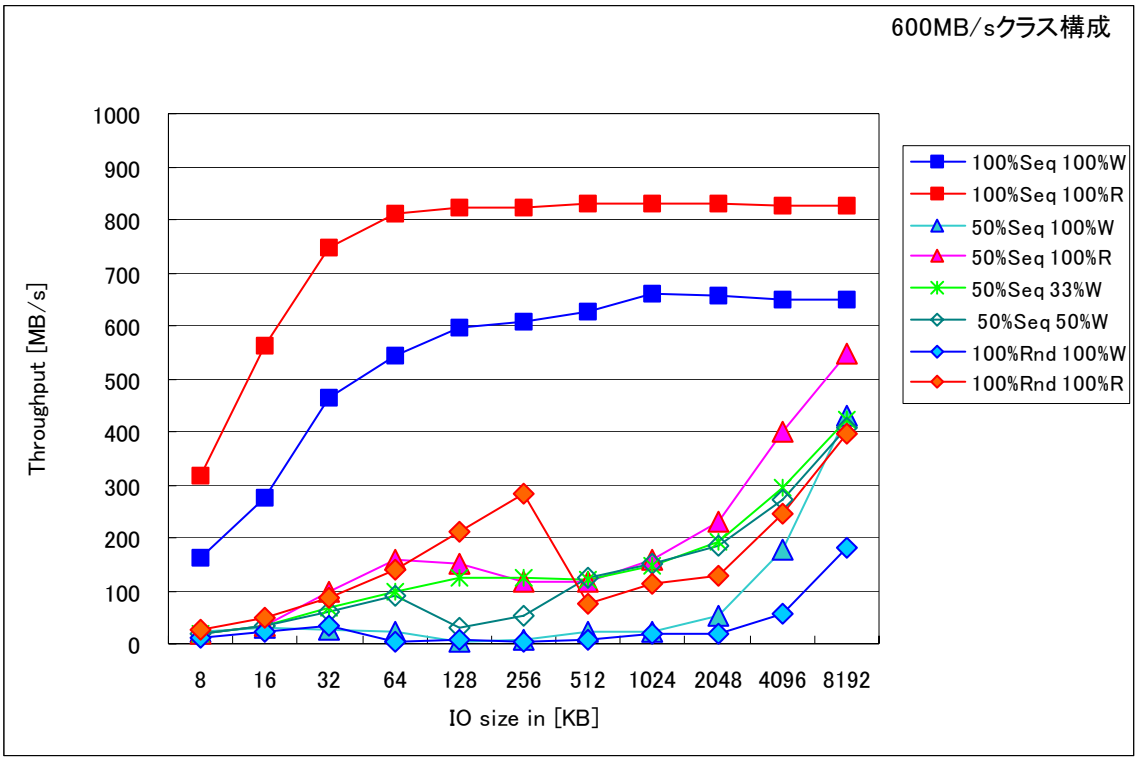
ストレージ部分に Xyratex 社製 F5404E FC-SATA RAID 装置を使用し、上記の写真構成で 1TB SATA x 48 台を内臓しています。拡張筐体を使用することで 96 台構成にする事も可能で、FC SW を併用することで 1TB x 96 台の構成を複数接続することも可能です。

（注意：サーバに使用する OS とファイルシステムにより、1 ボリュームあたりの最大容量は制限されますので、全てのケースで単一ボリュームを構成できるものではありません。）

RAID を構成し、ファイルシステムを作ることでユーザーが使用できる容量は案フォーマット物理容量よりも小さくなります。計算を簡単にするため、上記 48 台が搭載される筐体を単位に構成、40TB がユーザーにより使用できる容量であったと仮定し、500MB/s のデータを連続記録した場合の記録時間は以下の様になります。

表2 500MB/s データの記録時間

	40TB	80TB	120TB	160TB
500MB/s	22 時間	44 時間	66 時間	88 時間



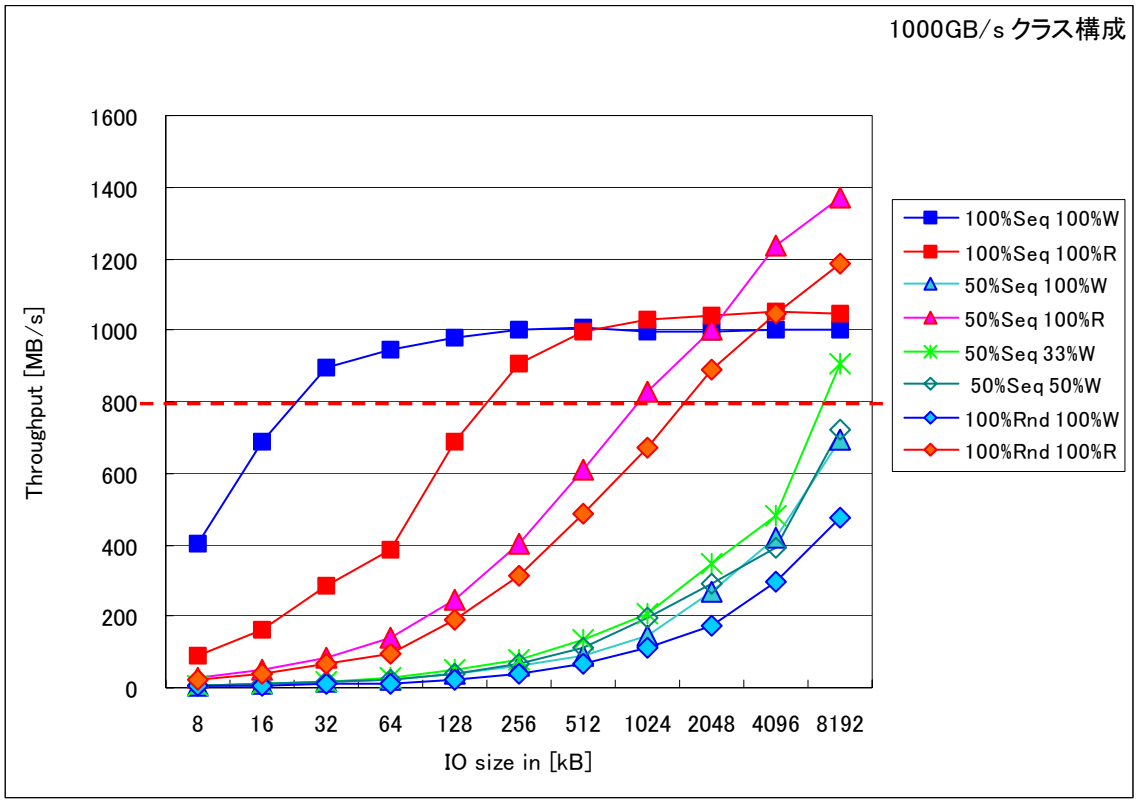
1000MB/s クラス構成

HS1235E-M 拡張構成

500MB/s クラスの HS1235E-M をベースにドライブ周りを強化した構成で SATA HDD 使用でありながら 1000GB/s の安定したスループットを得る事ができました。

PCI-Expres x4 スロット(2) / PCI-X (2) が未使用の為、ここを外部 I/F に利用する事により他機材との I/F や 10Gbit Ethernet NIC の増設が可能になります。
またサーバ部は Intel Quad Core Xeon (3GHz) のデュアル搭載が可能なため、CPU の処理能力が非常に高い点も特徴であり、様々なアプリケーションソフトウェアと組み合わせたシステム構築の素材となり得ます。
今回のテストは 20TB 構成で行いましたが、容量的にはさらに拡張が可能であり、倍の 40TB はもとより、それを超えるレンジをカバーすることが可能な能力を持っています。

この 1000MB/s の動作確認が SATA HDD での構成により行われた事はコストを抑えながら大容量化が可能である事を意味します。



まとめ

システム構築の際には SSD と HDD を適材適所で使い分ける必要があります。
単純に「速そうだから」とメモリを選択する事も常に正解とは言えません。PCI-Express x8 は片道 2GB/s の帯域を持っていますが、メモリによる PCI-Express 直結のディスク装置はまだその帯域まで届いておらず、逆に HDD でもスループット的には同等になっています。

ランダム I/O の場合はメモリ、シーケンシャル I/O で大量記録の場合は HDD との使い分けが　　現実の解となり得ます。シーケンシャル I/O かつ大量記録で応答も速くとの要求には SATA に代えて SAS HDD を使用することで折り合いをつける事も考えられます。
またデータアクセスのパターンによっては、メモリと HDD 両方を使い、ランダムアクセス部分をメモリでこなし、背後で大容量の HDD がアーカイブする構成も考えられます。

適用しようとするシステムとデータアクセスの性格を正しく評価し、最適なデバイスを選択する必要があります。