MIC Associates, Inc.

SATAドライブの有効な使用法について

ー待ち行列理論的視点での考察ー

エムアイシー・アソシエーツ株式会社

2010/03/25

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
1.0	2010/03/25	-	7

本書作成、編集、管理

MIC Associates, Inc.

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

目 次

2010/03/25

はじめに

現在、デジタル化の進行するなかで、膨大なデータに対するより高いアクセス性能への要求はますます強まって来ています。確かに、大容量メモリを沢山搭載したり、高性能RAIDコントローラを使用してストレージのアクセス性能を向上させることはストレージ製品、ストレージシステムに携わるITの仕事に携わる人間にとって常識です。

15,000rpmの高回転タイプのハードディスクで、アクセス性能と信頼性を実現しているのはSAS(Sirial Attached SCSI)ドライブですが、SAS ドライブの最大容量は1TB、2TBのSATA ドライブに比べ、半分しかありません。そこで、経済性にすぐれるSATAドライブをミッションクリティカルな業務アプルケーション用ストレージとして選択してしまうのは、現下の経済情勢での限られたIT予算からやむを得ない状況にあります。

その反面、経済性のあるSATAを選択したが故に、SASドライブが持つ高速なアクセス性能を犠牲にしていることも事実です。

そこで本稿では、どのようにしたら経済性に優れたSATAドライブを、できるだけ高い性能で、ストレージシステムに対し性能の維持ができるのかという観点から、ストレージにおける待ち行列の問題に触れてみたいと思います。

SATAドライブの物理仕様

まず、SATAのハードディスクの基本性能を確認しましょう。SATAドライブベンダとしては日立グローバルストレージ社、シーゲート社、他数社あり、それぞれがRAID装置用として信頼性の高いエンタープライズ仕様のSATAドライブとして生産しています。

・ドライブ容量：１TB～2TB
・ドライブ回転数：1分間当たり7,200回転（8.3ミリ秒／１回転）
・ヘッドのシーク時間：約10.2ミリ秒
　　フルストロークシーク：約10.2ミリ秒
　　シリンダー間シーク：約1ミリ秒
・コマンド処理時間：0.1ミリ秒から0.5ミリ秒

上記が現在RAID装置に使用されるSATAハードディスクの代表的な性能です。ここでSATAドライブの概要仕様として容量以下に回転数、ヘッドのシーク時間、ハードディスクのコマンドの処理時間をあげた理由は、ハードディスクにリードライトのリクエストをしてからデータその処理が終了するまでの最短時間と最大時間を予測するためです。

以上の仕様をじっくり眺めると、経済的なSATAドライブのアクセス性能を最大化させるにはどうすれば良いかということが見えてきます。

SATAドライブの性能について

まず、ドライブのシーク（シリンダー間のヘッドの移動）とサーチ(メディアの回転待ち）の時間を見てみましょう。7200回転/分で回転するメディア上のデータを読み出すには平均で4.1ミリ秒かかります。また、ヘッドがメディア上を移動して所定のシリンダーに到達する時間は平均で9ミリ秒です。

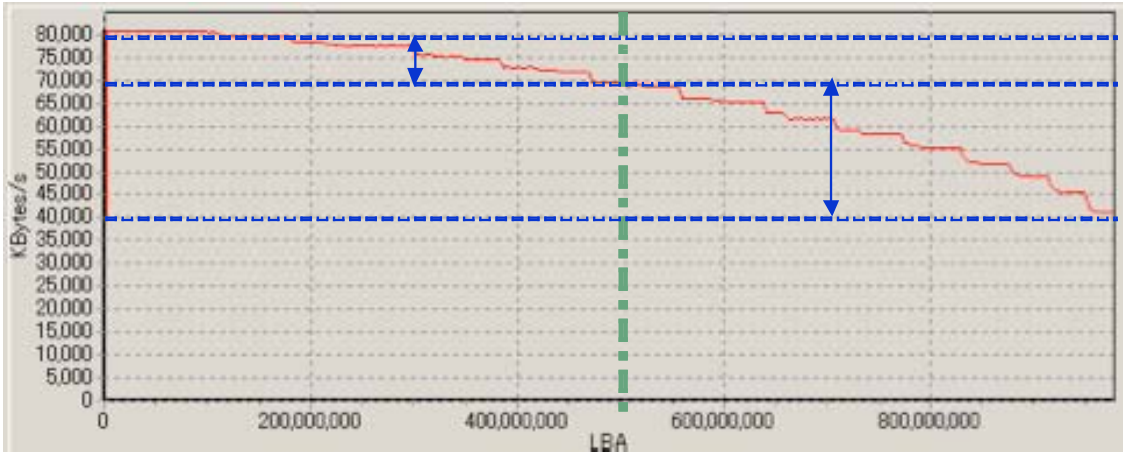
この結果、読み書きのコマンドを受けて処理が終わるまでの時間は以下の様になります。

処理時間=コマンド処理時間+シーク時間+サーチ時間=約13.4ミリ秒

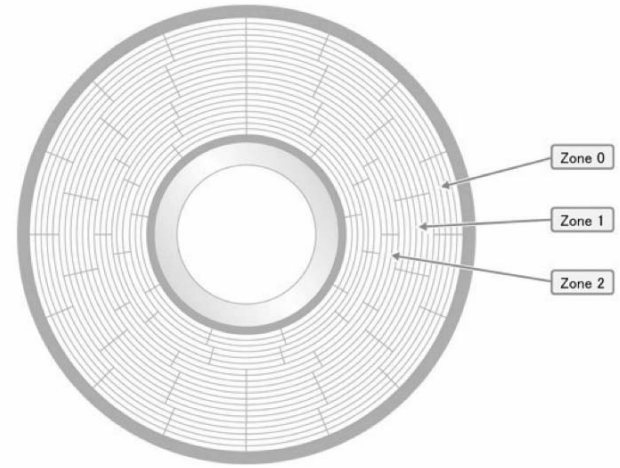
この時間は、パワフルなサーバのメモリ応答時間が数ナノ秒程度だと考えると、数万倍の時間がかかることになり、CPUはこの時間を待たされることになります。

しかし、上記の例はホストからストレージに対するコマンドが１つだけ出され、それを処理する時間のことです。実際は、このリード、ライトのコマンドが１秒間に数十から数千個出されストレージに対し、データが出し入れされます。結果的にストレージシステムのあちらこちらに処理の待ち時間が発生し、ストレージへのデータ書込みや、読み出しに多くの時間を要することになります。

ここで今暫くドライブの内部性能に関して見て行きましょう。先程述べたように、典型的な１TBのドライブの最外周のセクター数と最内周のセクター数は倍程度の開きがあります。 また、以上の仕様の他に、ドライブメディア上で記録されるデータの密度は一定ですので、ドライブメディアの最外周のシリンダーと最内周のシリンダーではセクターの数が少なくなります。例えば、典型的な1TBのドライブのセクター数は最外周部分で1680個、最内周で840個と2対1の比率になります。この最外周のシリンダーと最内周のシリンダーでのセクター数の差は下図の様に大幅なデータ転送性能の差を意味します。



図の横軸は最外周のシリンダーに配置された論理ブロックから最内周の論理ブロックに向けてIOを実行した時のデータ転送レート(秒当りのデータ転送量)が次第に減衰して行くグラフを示しています。LBA 0の部分が最外周シリンダーに当り、右端が最内周のシリンダーの位置です。最外周では約80MB/秒ですが、最内周ではその半分の40MB/秒であることが判ります。更に、全体の容量の50%を超えた辺りからデータの転送レートの減衰が大きくなります。全般の50%の容量ではデータレートの減衰は10MB/秒ですが、50%を過ぎた辺りから急激に減衰し、30MB/秒程度遅くなります。



これはハードディスクでは周速と呼びますが、ヘッドの下を通過するビットの量が最外周と最内周では2:1の差があり、一定のビット密度でヘッドの下を通過するデータ量はメディア内周に進むに従って減って行きます。これは、Zone-Bit(右図参照）とよばれるハードディスクの論理セクターのフォーマット方式により、最外周部分には多くのセクターを最内周部分には少ないセクターを論理的配置し、メディア外周部分の

記録密度を向上させているためです。

以上の様に、同じハードディスクでもその最外周部分と最内周部分ではデータの転送レートがかなりことなることがお判りいただけたと思います。新しいハードディスクを購入してデータ転送の性能に満足したが、やがてディスクの性能が遅くなるのは一つは内部のデータの配置場所がバラバラになるフラグメンテーションと呼ばれる現象が原因ですが、この最内周と最外周の性能差も大きく影響しています。

コマンド待ち行列的考察とストレージのアクセス性能について

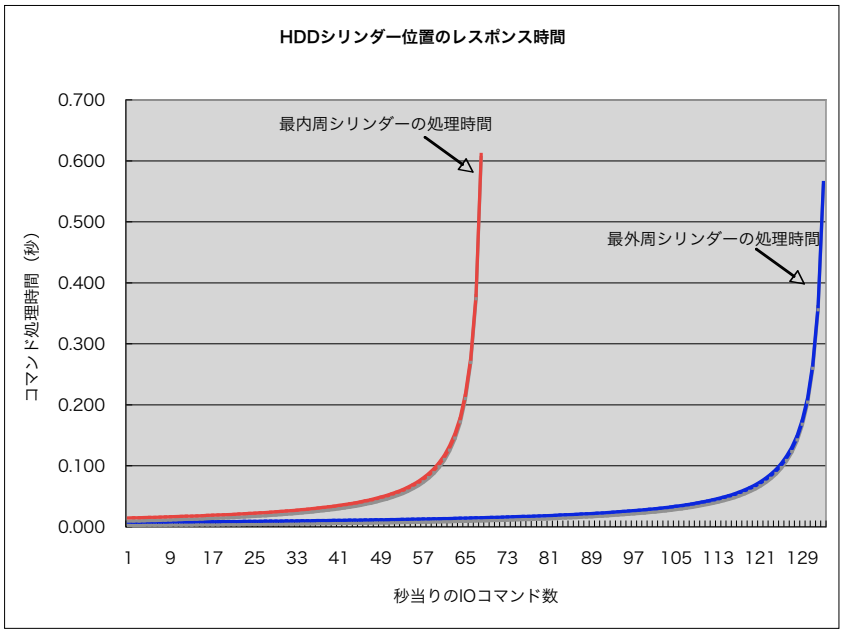
通常、コンピュータにおける待ちはCPU、メモリ、IOバス、IOデバイス等で発生します。今回のお話はストレージですので、コンピュータのIO要求に対してストレージがどの位の性能でデータを処理し、コンピュータにその処理が終了したことを伝えるまでの時間が大きければ、待ち時間が長いと言い、短ければ待ち時間が小さいと言います。この待ちの時間の原因はまずコンピュータがコマンドをストレージに投げて、ドライブメディアの特定の場所でデータを読み出し、または、書き出しし、データをコンピュータに返しその処理が終了します。この中で、コンピュータ側での時間は1ミリ秒以下で処理が終わりますが、ドライブメディア上でデータを読み出したり、書込んだりする時間はその数十倍から数千倍になります。例えば、コンピュータがハードディスクに512バイトのデータの書込みを最外周の位置と最内周の位置に行なう場合ですと、先に使用した1TB SATAドライブの場合ですと最外周部分では約5ミリ秒かかります。一方、最内周では15ミリ秒かかります。実は最外周では512Kバイトのデータを1つのシリンダーに書込むことができますが、最内周部分では512バイトのセクターが840個とすると、総容量が430KBとなり、不足した分は別のシリンダーに書込む必要が起き、最少のトラック間のヘッドシークが発生します。この分を合わせるとドライブメディアは約1.5回転する必要があり、同時にトラック間シークが起きます。この前提でディスクへの書込みの所用時間を計算すると以下のようになります。（コマンドの処理時間は0.25ミリ秒、最外周のセクター数1680個、最内周840個、セクターサイズ512バイト、トラック間シーク時間を1ミリ秒と想定。トラック間シーク実行後に所定のブロックに対する書込みを開始するのに1ミリ秒必要とするとした場合。）

最外周への書込み：(512x1000/1680x512)x8.3+0.25=5.19ミリ秒

最内周への書込み：(512x1000/840x512)x8.3+１+1+0.25=12.13ミリ秒

以上の通り、最内周部分に対する書込みは2.3倍以上の時間が必要になります。
さて、例えば映像データ等は512KBやそれ以上のサイズでデータをハードディスクに連続的に書込みます。また、データベース等は4KBや16KB程度の小さいデータのIOを繰返します。これらのユーザアプリケーションにおいてはこの2.3倍程度のレスポンスの差は現実的には大変大きな差になって感じられます。実際に上記の例で、1秒間にデータを70%シーケンシャル、30%ランダムと想定してこのファクターを加重平均して実際の待ち時間の差を計算してみます。シーケンシャルの場合は前のIOの後に次のIOが処理されると想定するとヘッドの移動は最少の1m秒ですみます。また、ランダムの場合は平均のヘッドシーク時間の5.1m秒(フルストロークの半分)が必要と仮定します。以上の内容で待ち行列の式を用いて1秒間でのコマンド数を1つから次第に増やして行くとどのようにコマンドの処理時間が変化するかを以下の表で表します。待ち行列の数式は以下計算式と最内周、最外周の処理時間をパラメータにします。

・平均待ち時間＝ρ／(1-ρ)xTs
・$\rho=\lambda$ xTs
・最外周Ts=5.19ms、最内周Ts=12.13ms



上図の様に、最内周部分では1秒間に50個のIOが発生すると処理時間が急激に悪化し、コンピュータはデータ転送が終了するための待ちが生ずる事になります。一方、最外周では１秒間に100を超えた時点で悪化しはじめますので、単純に見て最内周より倍の量のデータの読み書きが単位時間当り実行することができます。

また、今回のシミュレーションはシーケンシャルアクセスが70%、ランダムアクセスが30%という想定ですが、近年の仮想化サーバによる共有ストレージに対するアクセスの様に、それぞれが異なるモードで異なるデータストリームに同時にアクセスするような場合はほぼ100%に近いランダムなアクセスが起こり、ストレージのアクセス性能の劣化要因になってきています。

まとめ

以上のことから、大容量SATAドライブを使用する場合、使用可能なシリンダーの一杯に使用してパーティションを構成するとデータ処理時間に大きなロスを生じることが確認できます。実際のSATAハードディスクにおいてはこれらの物理的、機械的な処理時間を短縮するため、内部にリードキャッシュや、ライトバックキャッシュを設け、実際のディスク上での読み書きより、早く処理が終わる様に見せる技術が実装されていたり、実際のブロックに対するアクセスを効率良く行なことができるようにコマンドをバッファーにため、ドライブの回転や、ヘッドのシークを出来るだけ減らす方法がとられています。（弊社WP: 「Fibre Channel, SCSI Hard disk Drive vs. Serial ATA Hard disk drive」を参照ください。）また、SATAドライブを搭載したRAID装置にはコマンドを複数まとめて一度にアレーに効率良く読み書きさせる方法（コマンドクラスタ）や、実際のホストからの書込みに継ぐ同一データへのリードに関しては、ハードディスクに書込む前にキャッシュから直接データを送る方法等、効率化の為の手法が採用されています。(弊社WP: 「StorView "Statistics モニターの活用法"」をご参照ください。）

しかし、ディスクが回転するSATAハードディスクをストレージとして使用する限り、ドライブメディアの回転待ち、ヘッドのシーク待ちの時間は避けて通ることはできません。その為、 経済性に優れ、大容量なSATAハードディスクをミッションクリティカルなアプリケーション用ストレージとして使用する上で、出来るだけ効率的なIOが実現できるようにパーティションの構成を行なうことで、ストレージの回転待ち時間、ヘッドのシーク時間を軽減することは、システムの投資効果を最大化させなくてはならないITエンジニアにとっての課題といえます。