2005－EVA－14（2）
2005／8／5

SAN ブート環境におけるブレードサーバのディスク I/O 性能評価

上野　仁[†]、 中代　浩樹[†]

SAN（Storage Area Network）接続のストレージ装置1台の中に多数のシステムディスク用 LU（Logical Unit）を設ける SAN ブート環境を構築することにより、多数のサーバをいつでも任意の用途に変更して起動することができる。これはユーティリティコンピューティング環境実現に向け必要とされる機能の一つである。しかしながら、SAN ブート環境ではストレージへのアクセスパス中に共用部分が多く、構成によっては十分な性能がないためにシステム障害を起こす恐れがある。本報告では8台のサーバが1台の SAN 接続ストレージを共用する構成での性能実測方法を提案するとともに、ストレージ装置設定指針の考え方を示す。

Performance Measurement of Disk I/O for Blade Servers under SAN Boot Environment

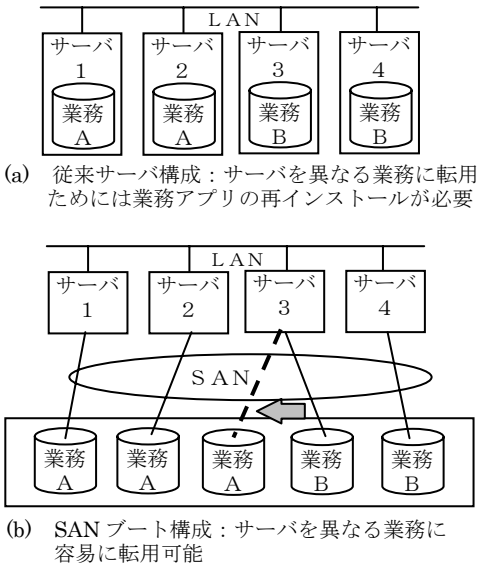
Hitoshi Ueno, Hiroki Nakashiro

An environment of boot from SAN enables easy change of server roles. The SAN boot function is made by preparing many LUs for system disks in a SAN storage unit, and it is one of necessary function for implementing utility computing environment. However, there are many shared parts on SAN boot environment, it may cause performance problems. In this report, we propose a method of I/O performance measurement and analysis for the SAN boot environment, by using an example system which has eight servers and one shared SAN storage unit.

1. はじめに

サーバベンダ各社より設置面積削減や運用の容易化を特長とするブレードサーバが製品化されサーバの実装密度が年々高まるとともに、ユーザサイトで利用可能なサーバ台数が増加してきている。一方ユーティリティコンピューティングの考え方が進展しつつあり、必要なときに必要な台数のサーバを必要とされる業務に割り当てる運用方法が求められている。

SAN ブート構成は従来サーバ構成（図1(a)）と比較して、任意のサーバを容易に任意のアプリケーションシステム用に変更、再起動できる点が特長である。（図1(b)）



(a) 従来サーバ構成：サーバを異なる業務に転用するためには業務アプリの再インストールが必要

(b) SAN ブート構成：サーバを異なる業務に容易に転用可能

図１　SAN ブート構成の一効果

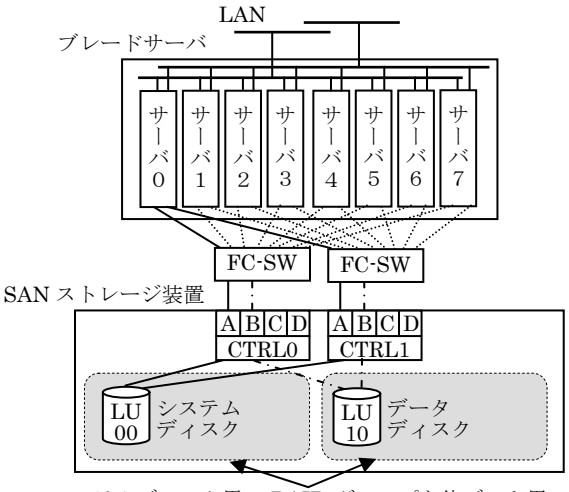
† 株式会社日立製作所システム開発研究所
Systems Development Laboratory, Hitachi, Ltd.

本報告では SAN 接続のストレージシステムを利用し、システムディスクとしてインストールした LU を各々のサーバに接続することにより、必要な時点で必要なサーバを立ち上げることができるシステムについて、課題を検討する。

2. SAN ブート構成の必要性と課題

ユーティリティコンピューティング環境では、時間帯によって1台のサーバを異なる用途で使用する、構成変更の容易性が望まれる。このためには時間帯によって異なるシステムディスクからサーバをブートする機能が必要となる。現状、サーバとしての性能を確保し、かつ、自由にブートディスクを変更できるストレージシステムとしては、SAN 接続のストレージシステムを挙げることができる。

一方システムディスクを SAN 接続のストレージに置く場合、複数のサーバ間で SAN 上の資源を共有するために性能問題が発生する危険性があると言われている。特にシステムディスク上に配置される swap ファイルへのアクセスが、十分なレスポンスを維持できない場合、オペレーティングシステムの挙動が不安定になったり、システムダウンが発生する恐れがあると言われている[1]。



・システムディスク用の RAID グループと他データ用の RAID グループを分離。
・システムディスクアクセスパス（──）データディスクアクセスパス（-･-･-）は各々二重化するが、負荷の分離のため、CTRL0 はシステム優先、CTRL1 はデータ優先。

図２　推奨 SAN ブート構成例

このような性能問題の発生を回避するための構成上の配慮として swap ファイルが存在するシステムディスクと長大なデータ転送の可能性があるデータディスクを分離するシステム構成方法を採用することが望ましい。

図2に2台の FC-SW（fiber channel switch）と1台の SAN 接続ストレージ装置装置を用いて8台のサーバの SAN ブートを可能とする構成例を示す。LU00 はサーバ0のシステムディスク、LU10 はサーバ0のデータディスクとして使用する。サーバ0からは2台の FC-SW を経由して LU00 と LU10 両方の LU にアクセスが可能である。システムディスクとデータディスクのトラフィックを分離するために、CTRL0（controller 0）に接続されるパスでは LU00 への read/write を優先し、CTRL1 に接続されるパスでは LU10 への read/write を優先する設定とする。サーバ1から7に対しても同様な LU 構成としてシステムを構築する。

このような場合、1台のストレージシステムに対して何台までのサーバを接続可能かが、システム構築コストを決めるひとつの要因となる。以下に、1台の SAN ストレージ装置に対して接続可能なサーバ台数の基準をもとめるシステム評価方法の例を報告する。

3. 性能測定構成および測定方法

性能測定は図3に示す構成で行った。図2に示す推奨構成例のシステムディスクアクセスに関する部分だけを実装した構成である。

8台のブレードサーバから1台の SAN 接続ストレージを接続し、それぞれのサーバに LU0 から LU8 までの LU を1台ずつ接続する。LU の接続設定は SAN ストレージ装置には標準的に装備されている LUN セキュリティ機能を用いる。また、ディスクコントローラ CTRL0 には4GB のディスクキャッシュを搭載している。

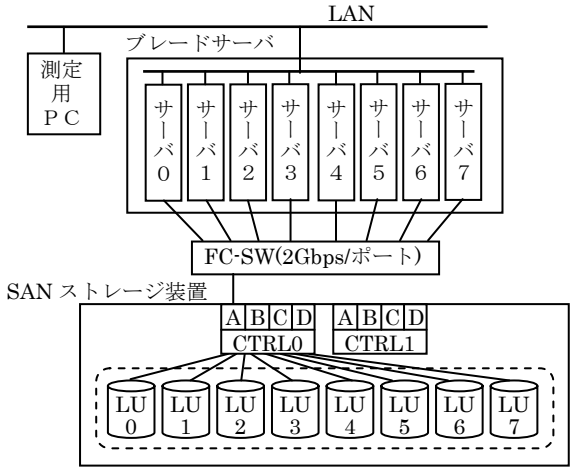


図３　性能実測構成

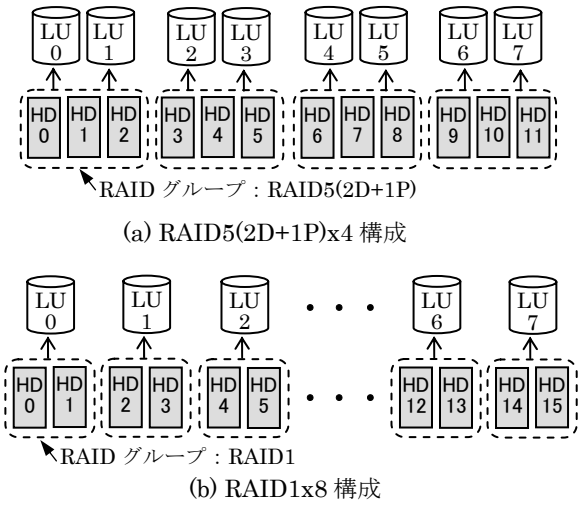


図４　性能評価時の RAID 構成

この構成において、各 LU はそれぞれ対応するサーバのシステムディスクとして動作し、さらに測定対象としてのファイル配置場所としても用いる。各サーバ上では、I/O 負荷を発生するツールプログラム[2]を稼動し、LAN 経由で接続した測定用 PC から負荷パタンを制御することにより、負荷発生源であるサーバの台数を1台から8台まで順次増加した場合の I/O 処理性能変化を測定する。

4. 性能測定結果

測定用の I/O 負荷は以下の条件とした。

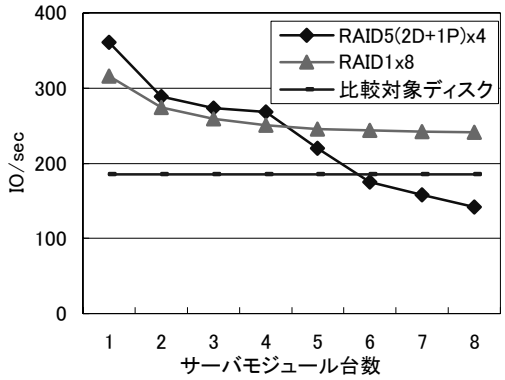
- read/write 時のブロックサイズ：4KB
- read/write 対象ファイルのサイズ：3GB
- アクセスパタン：ランダムアクセス（アクセス位置を、毎回ランダムに選択）、シーケンシャルアクセス（アクセス位置として前回のブロック位置の次のブロックを選択）
- read／write 比率：read 100%、write 100%、read 75%-write 25%

また、LU 構成による性能傾向の相違を見るために、RAID5(2D+1P)構成（図4(a)）および RAID1 構成（図4(b)）各々について上記 I/O 負荷を測定した。

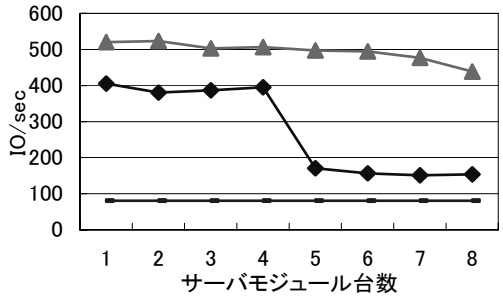
測定結果を図5に示す。横軸は SAN ストレージ装置に負荷をかけたサーバの台数を示し、縦軸は1サーバあたりの平均秒間 I/O 回数を示す。各々の図中には SAN ブート構成でない、内蔵ディスクを用いるサーバにおいて同様な測定を実施した結果を「比較対象ディスク」として示した。内蔵ディスクにシステムディスクを置く場合の I/O 性能との比較をするためである。

ランダムアクセス、read 100%における測定結果（図5(a)）では、RAID5 構成においてサーバ台数が6台以上の場合に平均 I/O 回数が比較対象ディスク以下となり内蔵ディスクを備えたサーバよりも性能が悪くなることが分かる。これはサーバ台数5台～8台の場合にはひとつの RAID グループに2台分のサーバから I/O 負荷がかかるためである。これに対して、RAID1構成では、サーバが8台の場合でもそれほど性能の低下が起きていない。これは、ひとつの RAID グループに1台のサーバからのみ I/O 負荷がかかるためである。

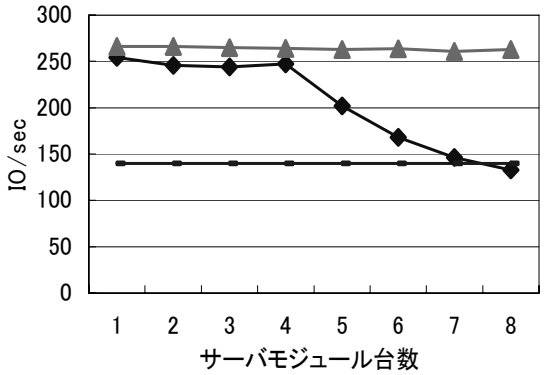
ランダムアクセス、write 100%における測定結果（図5(b)）では、同様に RAID5 構成のサーバ5台～8台のケースでサーバあたりの I/O 性能低下が見られるが、比較対象ディスクにおける性能を下回るケースはない。



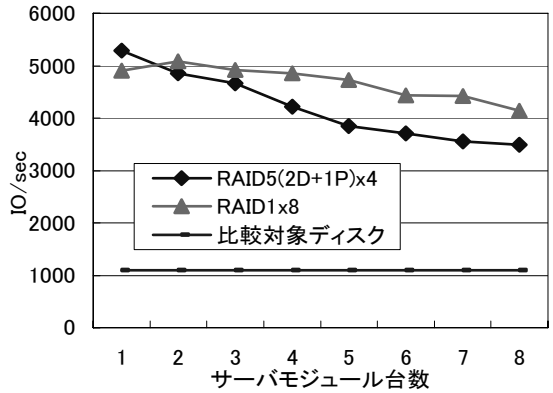
(a) ランダムアクセス、read100%



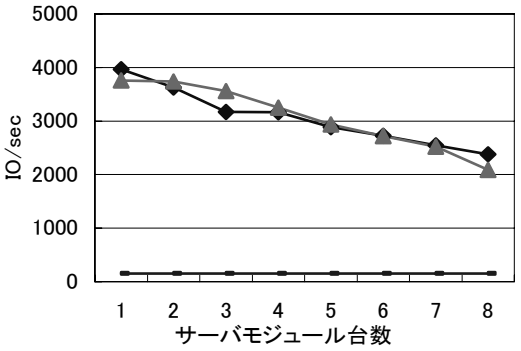
(b) ランダムアクセス、write100%



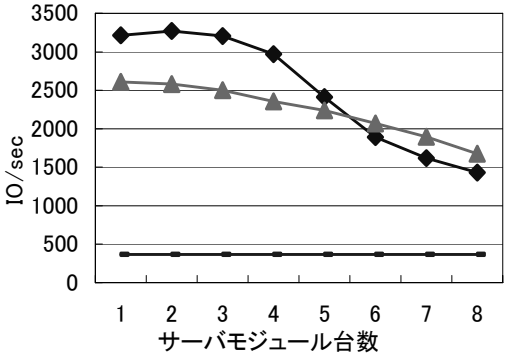
(c) ランダムアクセス、read75%-write25%



(d) シーケンシャルアクセス、read100%



(e) シーケンシャルアクセス、write100%



(f) シーケンシャルアクセス、read75%-write25%

図５　I/O 性能測定結果

シーケンシャルアクセス、read 100%および write 100%における測定結果（図5(d)、(e)）ではRAID5、RAID1 構成ともに比較対象ディスクの性能を大幅に上回っていることが分かる。これはシーケンシャルアクセスであるため、ディスクキャッシュへのプリフェッチ効果が大きいものと考えられる。

以上の測定結果から、サーバ8台までの SAN ストレージ装置の共有ではシーケンシャルアクセス性能に問題はないが、ランダムアクセス性能では RAID グループ構成方法に注意する必要があることが分かる。すなわち、いかなる負荷環境においても内蔵ディスクによるシステムと同等以上の性能を確保することを目的とするなら、「システムディスクとして使用する LU は、できるだけ他の LU とは別の RAID グループに配置することが望ましい」ことが分かる。

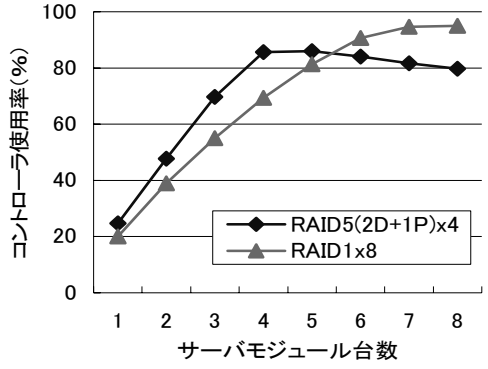
しかし、例えば「ランダムアクセス read 100%にお

けるサーバ8台のときの I/O 性能が比較対象ディスクに対して 20%程度低い」という特性が実運用システムに対してどの程度インパクトを与えるかは、ソフトウェアとハードウェアの構成に依存する。システムを不安定にする要因とされる swap ファイルへの I/O は、一般にはサーバシステムではほとんど発生しないようにメモリ設計をすることが多く、瞬間的な性能ボトルネックが生じたとしてもシステム障害にまで発展する危険性はないためである。

5. サーバ台数増加時の影響予測方法

前節に述べた性能測定ではサーバを8台用意し、実測することによりI/O性能を評価した。しかし、既存の資源による測定により、さらに大きな構成を採用する場合の性能を予測しなければならない場合がある。例えばサーバ台数を9台以上に増加する場合の性能を、実測済みの I/O 処理性能の傾向から単純に推定したいという場合である。このような場合には、性能の傾向の真のボトルネックが何であるかを知るために、ストレージ内部の性能をモニタする必要がある。

図6はストレージ装置内部の性能モニタ機能により、シーケンシャルアクセス read 75%-write 25% 負荷におけるディスクコントローラ内のプロセッサ使用率を測定した結果である。

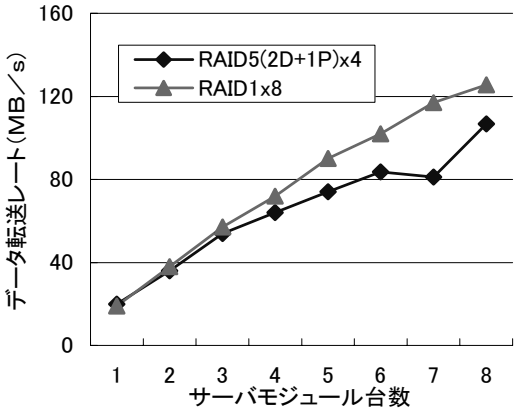


シーケンシャルアクセス、read75%-write25%

図６　コントローラ使用率

この負荷における I/O 処理性能(図5(f))の傾向からは、RAID1 構成の I/O 処理性能はサーバ台数が9台以上になっても比較対象ディスクに対して問題になるほどの性能低下は起こさないように見える。しかし図6によれば同構成のコントローラ使用率はサーバ台数8台のときには丁度限界に達しており、9台以上では急激に性能低下を起こす可能性を示唆している。

また、図7はシーケンシャルアクセスread 100%負荷におけるディスクコントローラから FC-SW に接続するポートのデータ転送レート(全サーバに対する転送レート合計)を示す。



シーケンシャルアクセス、read100%

図７　データ転送レート

同構成における I/O 処理性能(図5(d))の傾向からは、RAID1 構成の I/O 処理性能はサーバ台数が9台以上になっても比較対象ディスクに対して問題になるほどの性能低下は起こさないように見える。しかし図7によれば同構成のデータ転送レートはサーバ8台において120MB/Sに達しており、2Gbps の FC ケーブルを用いたシステムにおいては、エンコーディングオーバヘッドとプロトコルオーバヘッドを考慮すると、限界に近づいており、サーバ台数が9台以上に増加すると早期に急激な性能低下が発生すると推定できる。

6. まとめ

サーバのシステムディスクをSAN接続のストレージ装置に置くSANブート環境において、システムを安定的に動作させる十分な性能を確保するためには、RAIDグループ構成に注意する必要があることを示した。具体的にはひとつのシステムディスクは1個の独立したRAIDグループに配置することが望ましいことを示した。

また、既設機器による実測結果からサーバ台数増加時の影響を推定する場合には、ストレージ内部の性能要素に注意が必要であることを示した。

本報告の測定で示した上記の基準値は、現実的なシステム構築においてコスト的に厳しい条件である。したがって、システム障害を起こさない範囲で1個のRAIDグループ内に最大何台のシステムディスクを配置可能か、評価する方法を検討していく必要がある。

また、今後はユーティリティコンピューティング環境における各種資源割り当ての際の性能評価方法に関する研究を進めていく。

文　献

［1］ “Boot from SAN in Windows Server 2003 and Windows 2000 Server” Microsoft Corporation, December 2003.

［2］ Iometer Project、http://www.iometer.org/

［3］ Joe Carlisle, “Exchange 2000 RAID 1+0 versus RAID 5 Performance on a Hitachi Thunder Series 9570V System”, CMG Symposium 2003, May 2003.

［4］ 茂木、喜連川、”ストライプの動的再編成を伴うRAID5型ディスクアレイに於けるアクセスローカリティが存在する場合の更新処理の性能評価”計算機アーキテクチャ研究会107-25、情報処理学会、July 1994.

［5］ 松本、”NICを活用したネットワークRAID方式の提案”計算機アーキテクチャ研究会139-14、情報処理学会、August 2000.