Vol.2009-EVA-30 No.2
2009/11/16

I/O レイテンシに着目した サーバ性能推定モデルの提案と評価

上原敬太郎† 馬場貴成† 對馬雄次†

近年 I/O 仮想化によるサーバ統合が注目されている．しかし I/O 仮想化導入に伴うプロトコル変換等のオーバヘッドが，アプリケーションに与える影響を定量的に見積もる評価手法は従来存在しなかった．本研究では I/O レイテンシに着目したサーバ性能推定モデルを提案する．筆者らは I/O レイテンシを変化させることのできる Advanced Switch Interconnect(ASI)実験環境を用いてアプリケーションの性能に対する I/O レイテンシ感度評価を行った．データベースアプリケーションを用いた評価の結果，レイテンシの増加が 1500ns を超えると性能が急激に低下する待ち行列的な性質を持つことがわかった．これらの結果から，平均的な I/O スイッチ 2 階層構成までであれば，5%以下の実用的な性能低下で I/O 仮想化が導入できることが明らかとなった．

A Proposal and Evaluation of Server Application Performance Estimation Model with I/O Latency

Keitaro Uehara†, Takashige Baba† and Yuji Tsushima†

In today's IT systems, I/O virtualization technology is emerging for server integration. Although I/O virtualization involves some overheads such as protocol conversion, there are no evaluation methods or models to estimate the quantitative effects of such overheads for server applications. This paper proposes a server application performance estimation model with I/O latency. We use an experimental environment with Advanced Switch Interconnect, which enables I/O latency to be delayed gradually, to examine effects for the server application performance with delayed I/O latency. The results show that the performance of the database application degrades suddenly when I/O latency exceeds over 1500ns, while only degrades 5% or less within the 1500ns I/O latency in our environment. The results imply that two layered I/O switches are key to the realization of I/O virtualization.

1. 背景と目的

半導体のプロセス進化に伴う CPU マルチコア化の進行，およびブレードサーバや仮想化によるサーバ統合ニーズの高まりにより，サーバの体積当たりの計算量は年々増加している．一方，本質的に外部との入出力チャネルを必要とする I/O に関してはおのずと集約に限界があるため，結果としてサーバの計算量に対する I/O チャネル不足が顕在化しつつある．

この問題に対する解決案として，複数の I/O プロトコルをスイッチファブリック上に混在させる I/O 仮想化を用いた I/O 集約技術が着目されている 1)2)3)．しかし I/O 仮想化は，プロトコル変換やスイッチ導入のためのオーバヘッドを伴うが，これらのオーバヘッドがサーバアプリケーションに与える影響を定量的に見積もる評価手法は従来存在しなかった．

本研究の目的は，実サーバアプリケーションに対する I/O レイテンシの影響を定量的に評価し，性能推定モデルを構築することである．そのための最初のステップとして，I/O レイテンシを変化させることができる Advanced Switch Interconnect(ASI)4)実験環境を使って I/O レイテンシを可変とし，I/O レイテンシを増加させることによるデータベースアプリケーションの性能への与える影響を定量的に見積もることを目標とする．

2. I/O レイテンシ性能推定モデルの構築

サーバ集約のために I/O 仮想化が導入されたサーバ環境におけるサーバアプリケーションの性能を推定するに当たっては，I/O レイテンシが重要なファクターとなる．従来の I/O 性能評価では，I/O からメモリまでのレイテンシと DMA[I]の同時発行可能トランザクション(Tx)数で帯域の何%を利用できるか算出する I/O 帯域重視の手法が用いられてきた．しかしシステムの大規模化や I/O インタフェースの高速化に伴い，I/O 帯域が十分大きくなった昨今のサーバプラットフォームにおいては，I/O レイテンシの I/O 帯域に対する影響を評価するだけでは不十分であり，I/O レイテンシのアプリケーション性能に与える影響をより直接的に評価する手法が必要となる．

そこで，I/O レイテンシ増加のアプリケーションに対する影響を直接測定し定量化した性能推定モデルを構築することが本研究の課題である．本モデルでは，基準となるアプリケーション性能値（スループット）P_{Local} に対して，I/O レイテンシを x [ns] 増加させた時のアプリケーション性能値 $P_{Lateny[x]}$を計測し，この時の性能低下率 PD[x] を次のように定義する：

† （株）日立製作所 中央研究所
Hitachi Ltd., Central Research Laboratory

[I] DMA=Direct Memory Access

$$PD[x] = (P_{Local} - P_{Latency[x]}) / P_{Local} \times 100\ [\%]$$

PD[x]を計測することにより，アプリケーションごとの I/O レイテンシと性能低下率の相関関係をモデル化できる．以下，x と PD の相関関係のグラフを性能低下曲線と呼ぶ．アプリケーション性能推定モデルを構築するに当たっては，あるサーバ環境に対する性能低下曲線を求めることが第一のステップとなる．

さらに得られた結果を，実験環境と異なるサーバのアプリケーション推定に適用するためには，サーバ間のアーキテクチャの違い（CPU 性能，メモリレイテンシ，I/O 帯域等）を考慮し，外挿推定を行う必要がある．ここでは単純化のために CPU 性能にのみ着目する. I/O デバイスからの DMA 転送は CPU の実行と非同期に行われるため，I/O レイテンシの影響は直接 CPU のストール時間にマッピングできない．このため CPU コア性能の向上により，I/O レイテンシと性能低下の相関関係（性能低下曲線）がどう変化するかについては，以下のように拡大・縮小の両方の要因が考えられる．

- 性能低下が拡大する要因：I/O レイテンシの伸びにより，CPU の I/O 処理待ちが発生している場合，CPU コア性能が向上すると，同じ待ち時間でより多くのサイクル数を消費することになるため，性能低下は拡大する．
- 性能低下が縮小する要因:CPU の処理ネックにより DMA の発行数が不足している場合，CPU コア性能が向上すると，より多くの DMA を発行できるようになるため，個々の DMA の I/O レイテンシは隠蔽されて，I/O レイテンシを伸ばした時の性能低下は縮小する．

従って，アプリケーション性能推定モデルを，世代の異なる CPU を搭載したサーバの性能予測に適用するためには，第二ステップとして異なる CPU 性能を持つサーバの性能低下曲線を測定し，CPU 性能の向上が性能低下に与える影響の傾向を把握する必要がある. 以上の 2 つのステップで得られたデータを元に, 求めるサーバにおける CPU 性能を外挿することにより，I/O レイテンシに対するアプリケーションの性能低下率を推定することが可能となる．

3. 評価環境と評価方法

3.1 ASI 実験環境と評価対象の概要

ASI(Advanced Switching Interconnect)とは，標準化団体 ASI SIG によって推進されていた，PCI-Express と互換性を持ったスイッチの拡張規格である．PI(Protocol Interface)と呼ばれるさまざまなプロトコルをカプセル化したインタフェースでファブリック上に混在させることができるため，スイッチによる自律的な管理や帯域保証，ホスト間での I/O 共有，共有メモリなどを実現可能としている．本実験では，ASI をレイテンシ遅延装置として用いるために, PCI-Express パケットをトンネリングするプロトコルである PI-8 のみを使用する．

表 3-1：実験環境のスペック

DB サーバ (1 台)	CPU	Pentium 4 3.0GHz (FSB 800MHz)×1
	チップセット	Intel E7230
	メモリ	DDR2-533MHz 1024MB
	OS	Red Hat Enterprise Linux 4
	I/O Interface	PCI-Express Gen1 x4
	Host Bus Adapter	QLogic QLE2360
	Network Interface Card	Intel PRO1000 (e1000) or Broadcom BCM5721 (bcm5700)
	Database	Oracle 10g
	Storage	Data 領域：36GB×2
		Redo 領域：36GB×1
クライアント (1 台)	CPU	Pentium 4 3.0GHz (FSB 800MHz)×1
	チップセット	Intel E7230
	メモリ	DDR2-533MHz 1024MB
	OS	Red Hat Enterprise Linux 4
	I/O Interface	PCI-Express Gen1 x4
	Network Interface Card	Broadcom BCM5721 (bcm5700)

表 3-1 に実験環境に用いたデータベース（DB）サーバおよびクライアントの仕様を，図 3-1 に本実験環境の構成を示す．図 3-1(a)に示すように，DB サーバに搭載された AS アダプタから，2 段の Advanced Switch を経由して，I/O 拡張筐体へと接続する．また，2 台の Advanced Switch 間はケーブルで多重に接続され，Advanced Switch 間の往復回数を変えることで I/O レイテンシを変化させることができるように構築している．I/O 拡張筐体には NIC および HBA が搭載され，それぞれクライアントと FC スイッチへと接続される．拡張 I/O 筐体のスロットは通常の PCI-Express であるため，DB サーバ上で動く OS およびアプリケーションからは，点線で囲まれた範囲までが DB サーバの筐体と同じように認識される．また，比較のために DB サーバに直接 NIC と HBA を載せた構成（以降 Local と呼ぶ）も用いる（図 3-1(b)）. さらにアダプタと PCI-Express スロットの間にプロトコルアナライザを挿入することにより, PCI-Express のパケット

トレース列を採取することができる．

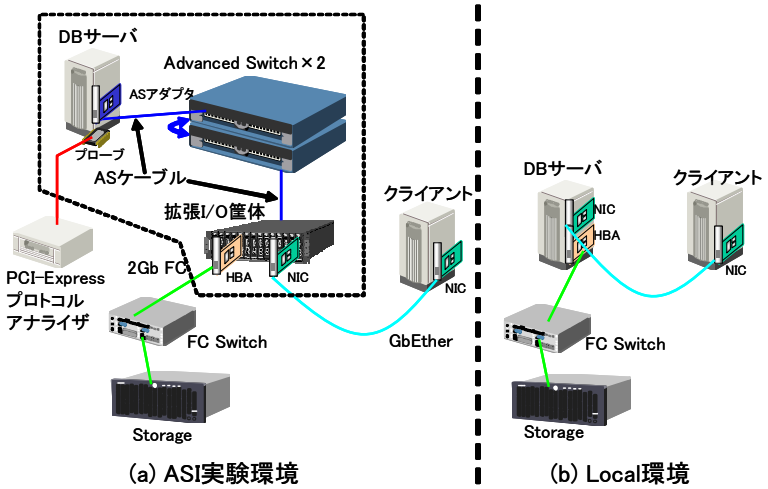

図 3-1：ASI 実験環境および Local 環境の接続構成

以上の実験環境の下で，評価対称としては以下のベンチマークを用いる．まずレイテンシを変化させる ASI 実験環境が，意図通りに設定されているかを確認するために，スループットを計測するベンチマークとしてディスク性能を測る iozone，およびネットワーク性能を測る netperf を用いる．さらに I/O 仮想化によるサーバ集約が期待される分野として，DB サーバにおける OLTP 性能を測る TPC-C を模したベンチマークを用いる．評価する I/O アダプタは表 3-1 に載せた GbEther NIC(2 種類)と FC HBA(1 種類)である．また，TPC-C のパラメータは表 3-2 に示す通りである．

表 3-2：TPC-C のパラメータ

ScaleFactor	32
端末数	16
バッファサイズ	1GB
シンクタイム	無し(BATCH モード)
Warmup 時間	30 分
計測時間	60 分

3.2 評価方法

上記 ASI 実験環境を用いて I/O レイテンシを変化させ，アプリケーション性能を測定し，Local と比較した場合の性能低下率を求める．ASI は 1 段通過ごとに片道 150ns・往復 300ns レイテンシが増加し，今回用いた実験環境では最大 7 段まで設定可能であるため，最大で往復 2100ns 分の I/O レイテンシの増加を評価可能である．

表 3-3 に現行製品である各種 I/O スイッチの通過レイテンシを比較した表を示す．この中では Cisco の FCoE スイッチが他と比べて一桁レイテンシが大きいが，一般的にネットワークスイッチはレイテンシよりもスループット重視であるため，FC スイッチや InfiniBand スイッチに比べてレイテンシが大きくなる傾向がある．この表の結果より，ネットワークスイッチを除く一般的な I/O スイッチの通過レイテンシは片道 150～400ns 程度であると考えられる．

表 3-3：各種 I/O スイッチの通過レイテンシ比較

ベンダ	StarGen4)	Mellanox5)	QLogic6)	Cisco7)
製品名	Advanced SW	Infini ScaleIII	SANbox 5200	Nexus5000
プロトコル種別	ASI/ PCI-Express	InfiniBand	FibreChannel	FCoE/DCE[II]
通過レイテンシ	150ns	200ns	400ns	3200ns

4. 評価結果と考察

4.1 I/O アダプタごとの I/O レイテンシ感度特性の評価

図 4-1 に iozone および netperf の測定結果を示す．横軸に Local を基準とした I/O レイテンシ増加分（往復），縦軸に Local と比較した性能低下率を取る．

[II] FCoE=FibreChaneel over Ethernet, DCE=DataCenter Ethernet

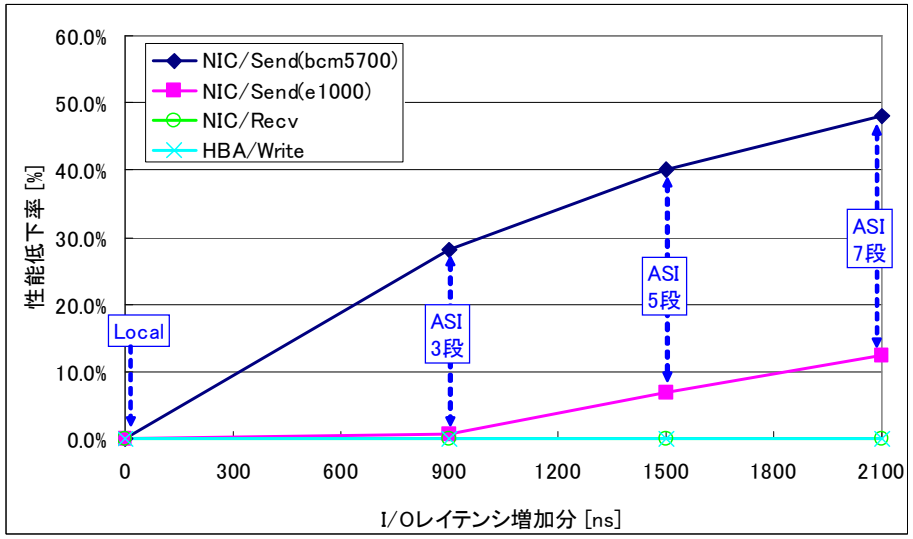

図 4-1：I/O レイテンシと性能低下率の関係

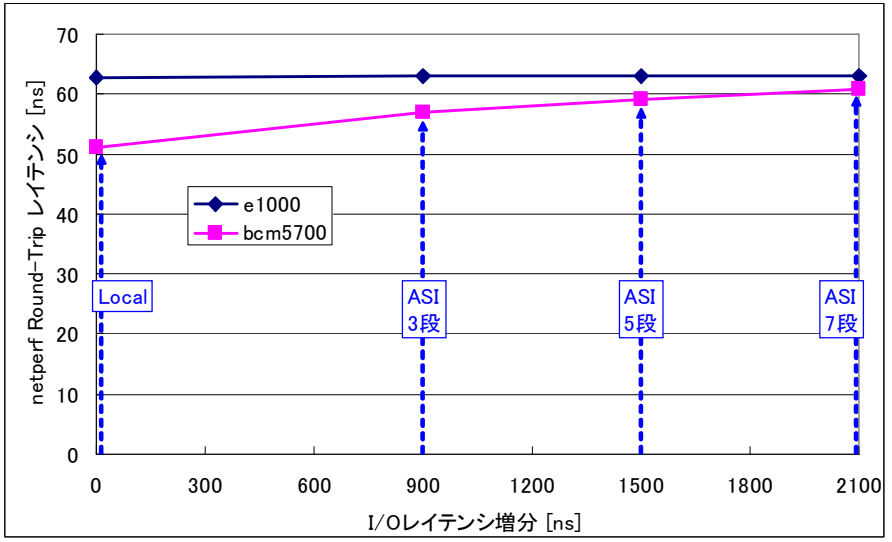

図 4-2：netperf/Round-Trip レイテンシ結果

この結果より，以下が推測できる：

- HBA/Write はレイテンシの増加の影響を受けていない．これは，HBA は同時発行 Tx 数が十分に多いために，レイテンシ増加の影響を受けていないためと推定される．
- NIC/Recv もレイテンシ増加の影響は見られない．NIC の受信は DMA Write を伴う．DMA Write は PCI-Express では完了を待つ必要の無い Tx（Posted Tx）であるため，I/O アダプタからスイッチに対して発行された時点でバッファを解放できる．このため，その後のレイテンシがいくら延びても同時発行 Tx 数には影響を与えないと推定される．
- NIC/Send はレイテンシ増加の影響を受け，その影響は使用する NIC 種別によって大きく違い，Intel PRO1000(以下 e1000)では ASI×7 段(＋2100ns)でも 10%強程度だが，BCM 5721(以下 bcm5700)では 50%弱もの性能低下が起きる．NIC の送信は DMA Read を伴い，DMA Read は完了(CplD)を待つ必要のある Tx(Non-Posted Tx)であるため，同時発行数の影響を受け，レイテンシ増加の影響を受けたと推定される．

図 4-2 に，netperf で計測した Round-Trip レイテンシ（1 回分の往復レイテンシ）を，ASI 段数を変えて計測した結果を示す．Local 時は e1000 よりも bcm5700 の方がレイテンシは短い．しかし ASI 段数が増えても e1000 はほとんどレイテンシが増えないにもかかわらず，bcm5700 はレイテンシが増加している．このことから，bcm5700 はより少ないバッファで DMA を処理しており，単発の処理は早く処理できるが，レイテンシ増加の影響を受け易いと推定される．一方 e1000 は大きなバッファで DMA を処理しているため，単発の処理には時間がかかるが，その分レイテンシ増加の影響を受けにくいと推定される．次節では，プロトコルアナライザのトレース列の解析によりこの仮説を検証する．

4.2 プロトコルアナライザを用いたアクセス特性の解析

図 4-3 に netperf 実行時の NIC/Send のアクセスパターン(e1000 と bcm5700 の比較)を示す．左のカラムから方向（↑が Upstream，すなわち I/O からホスト方向，↓が Downstream，すなわちホストから I/O 方向を示す），PCI-Express Tx 種別，データサイズ，推定処理内容，を示す．どちらも DMA Read 列と DMA Ring 更新のための INT（割り込み）～レジスタアクセス列の繰り返しで構成される点は同一である．しかし e1000 が 512B×3 を単位として DMA Read を発行するのに対して，bcm5700 では 512B 単位

での DMA Read となっている．このため，e1000 の方がレイテンシ増加の影響を直接受けにくいと推定される．この推定に基づき，最後のパケットが到着するまでの時間の伸びを元に推定したスループット低下率は，実測によるスループット低下率の値とほぼ一致し（図 4-4 および表 4-1 参照），上記推定が正しいことが検証された．

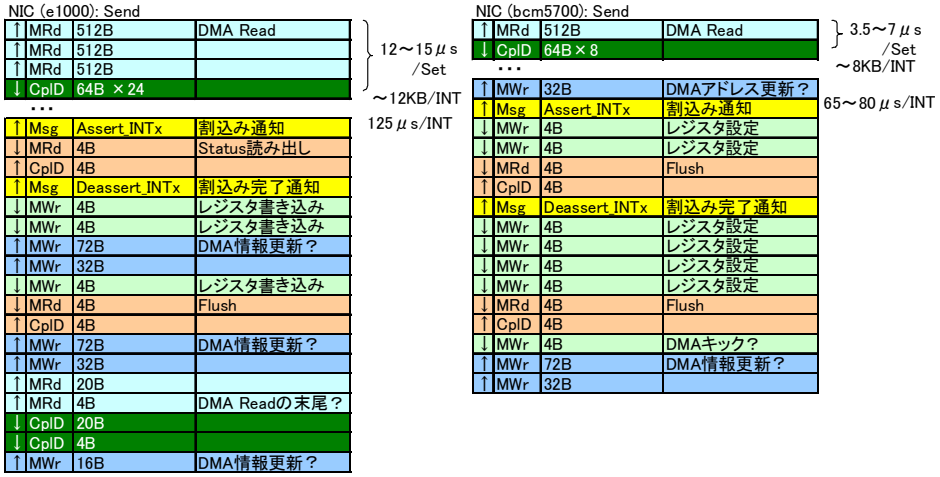

図 4-3：NIC/Send 時のアクセスパターン

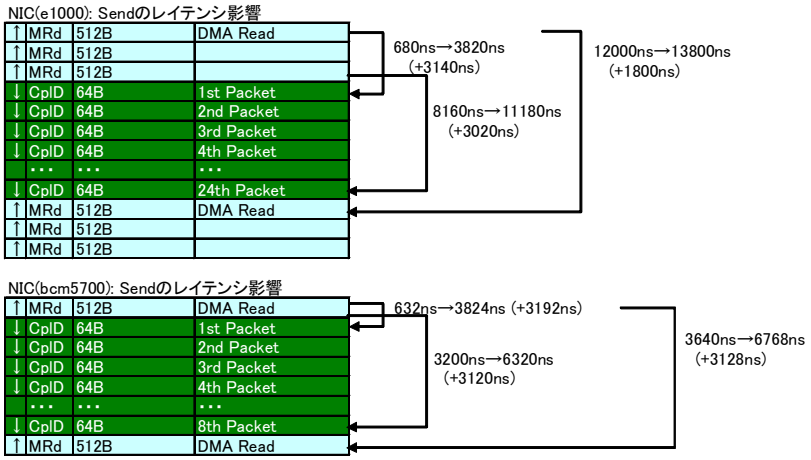

図 4-4：MRd 発行間隔とレイテンシへの影響

表 4-1：MRd 発行間隔から求めた性能低下率

NIC 種別	転送 Byte	MRd 発行間隔 [ns]			推定スループット [MB/s]			実測低下率
		Local	ASI×7	増分	Local	ASI×7	低下率	
e1000	1,536	12,000	13,800	1,800	128.0	111.3	-13%	-12.4%
bcm5700	512	3,640	6,768	3,128	140.7	75.7	-46%	-48.1%

4.3 データベースアプリケーションを用いた I/O レイテンシ感度評価

本節では，より現実的なサーバアプリケーションとして，データベースを介した OLTP[III]である TPC-C を模擬したベンチマークを用いて評価を行った．実験に用いた環境・パラメータは表 3-1 に示す通りである．ベンチマーク値の単位は tpmC である．

ASI 段数を 3 段～7 段および Local とした場合の TPC-C 結果および性能低下率を図 4-5 に示す．ASI 段数が 5 段を超えたあたりに変曲点があり，スループットが急激に悪化している．このような急激な性能低下の増加は待ち行列による性能低下の際に見られる傾向と似ているため，I/O 処理に依存する何らかの処理待ちが発生していると推定される．詳細な原因の分析には，CPU パフォーマンスモニタを採取して比較する等の解析が必要となる．

[III] OLTP=OnLine Transaction Processing

Vol.2009-EVA-30 No.2
2009/11/16

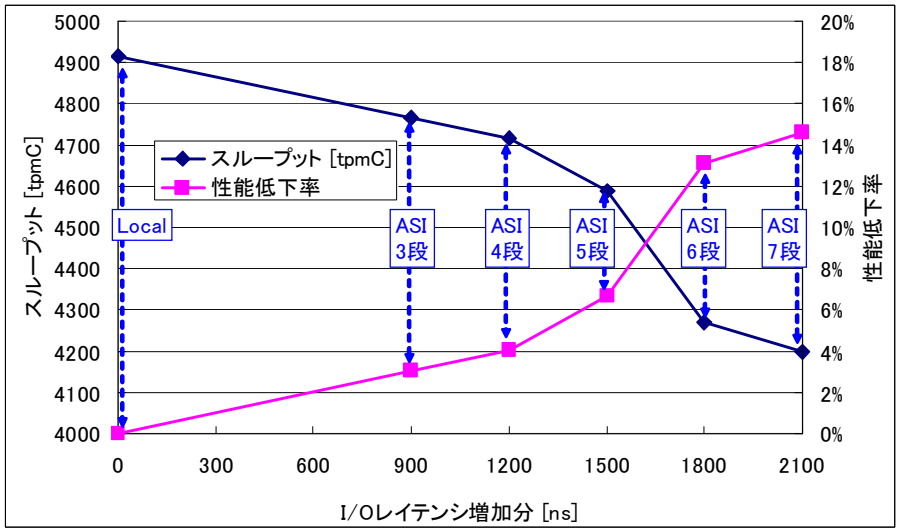

図 4-5：レイテンシ増加分と性能低下率の関係（性能低下曲線）

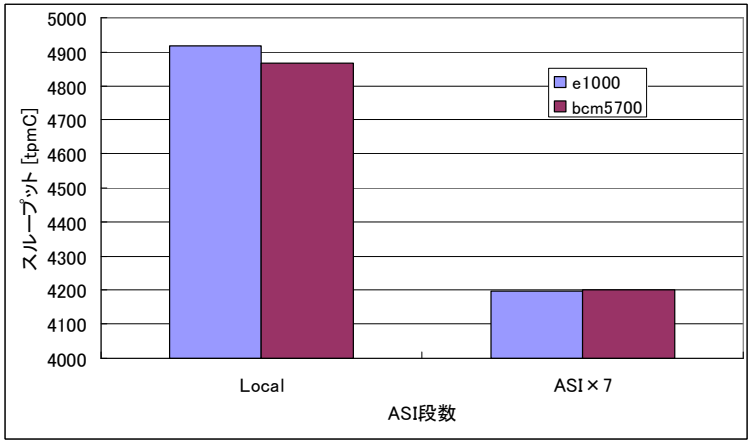

図 4-6：NIC 種別による TPC-C 性能の違い

図 4-6 に NIC の種別によるスループットの比較を示す．前節の netperf/Send における測定では Intel PRO1000/Brocade bcm5700 でレイテンシを伸ばした時のスループット低下率に大きな差が見られたが，TPC-C では両者で有意な差は見られない．これは TPC-C 実行時のネットワーク負荷が低いため，レイテンシ増加による同時発行数不足の問題が顕在化しないためと推定される．

4.4 実用的な I/O スイッチ段数の見積もり

前節で得られた性能低下曲線を元に，実用的な性能低下率である 5%以下となる I/O スイッチ段数を見積もる．3.2 節で示した典型的な I/O スイッチの通過レイテンシ 150ns ～400ns の範囲で変化させた時に，段数に応じた性能低下率をプロットした結果を図 4-7 に示す．

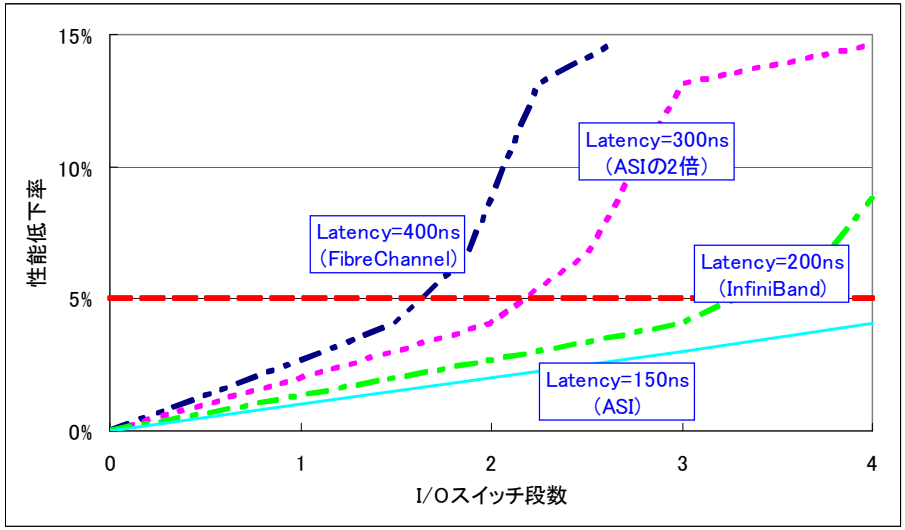

図 4-7：I/O スイッチ段数と性能低下率の関係

この結果より，最も早い ASI のレイテンシ 150ns からその倍の 300ns までの範囲の I/O スイッチであれば，実用的な 5%未満の性能低下率で 2 階層までの I/O スイッチ構成を組むことが可能であることがわかる．しかし FibreChannel スイッチのレイテンシである 400ns を仮定した場合には，2 階層構成とすると性能低下率が 10%弱にまで悪化する．これらのことから，I/O 仮想化によるサーバ統合を行う際には，I/O スイッチ導入やプロトコル変換に伴うレイテンシ増加の影響を十分に考慮した上で，システムの全体構成を検討することが重要であることがわかる．

5. 関連研究

PCI-Express に関する I/O 性能評価の論文としては，InfiniBand の HCA(Host Channel Adapter)の性能が PCI-X から PCI-Express になったことでどの程度改善したかを評価した 8)がある．ASI を用いたレイテンシおよびスループットの評価としては，StarGen 社の 4)や日立の 9)がある．また，PCI-Express のパケットをカプセル化し，Ethernet を介して通信する ExpressEther という規格を NEC が提唱している 10).

6. まとめと今後の課題

6.1 まとめ

仮想化によるサーバ統合が進み，CPU・メモリの集約から，I/O の集約へとフェーズが移行しつつある．このような中，I/O 仮想化に向けた I/O スイッチ導入による I/O レイテンシ増加が実サーバアプリケーションに与える影響を評価するため，レイテンシを変化できる Advanced Switch 実験環境を用いた実測による I/O レイテンシ感度評価を行った．

基幹向けデータベースサーバ上の実アプリケーションとして TPC-C を模擬したベンチマークを用いた測定では，ASI 段数 5 段（1500ns）を超えたあたりに変曲点があり，5%から 15%程度に急速に性能が低下することがわかった．スループット中心のベンチマークでは性能低下が見られない場合でも，実アプリケーションでは有意な性能差が生じるケースがあることが確かめられた点，およびオーバヘッドの増加が I/O レイテンシに対して待ち行列的な性質を持つことを確かめられた点は，今回の研究による重要な知見である．また，通過レイテンシが 300ns 以下の I/O スイッチであれば，I/O スイッチを 2 階層までに抑えることで，実用的な性能低下率である 5%以下で I/O 仮想化を導入できることがわかった．

ただし，今回のデータは 1 つの CPU スペックのサーバを用いたデータのみであり，今後別の CPU を搭載したサーバに今回の結果を適用するためには，異なる CPU を搭載したサーバを用いたデータを取得し，外挿による推定を行う必要がある．

6.2 今後の課題

今回の I/O レイテンシ感度評価の結果を実システムに適用するためには，異なる周波数・μ アーキテクチャの CPU を搭載したサーバのデータを集め，近似の精度を向上させる必要がある．さらに I/O レイテンシがアプリケーション性能に与える影響は見積もれたが, I/O レイテンシを伸ばした時にタイムアウトとならずにブートできるか，といった性能以外の指標による評価を今後も継続して行う必要があると考えている．

参考文献

1) Cisco & VMware: Data Center 3.0:データセンターの仮想化を促進するソリューション，ホワイトペーパー，2008 年 12 月．
2) Brocade Technical Brief: サーバ仮想化と Brocade DCF (Data Center Fabric) アーキテクチャによるデータセンター統合の最適化, 2008 年.
3) PCI-SIG I/O Virtualization home page, http://www.pcisig.com/specifications/iov/
4) Venkata Krishnan et al.: A Case Study in I/O Disaggregation using PCI Express Advanced Switching Interconnect (ASI), 14th IEEE Symposium on High-Performance Interconnects (HOTI'06), pp.15-24, Aug 2006.
5) Mellanox: IPoIB Stateless Offloads and More, Nov 2007, SC07 OpenFabrics Developer Summit.
6) Steven Schuchart Jr.: High on Fibre, Network Computing, Dec 2007.
7) Cisco Systems Inc.: Unified Fabric: Benefits and Architecture of Virtual I/O, 2005.
8) Jiuxing Liu et al.: Evaluating InfiniBand performance with PCI Express, IEEE Micro Vol.25, Issue 1, pp.20-29, Jan-Feb 2005.
9) 沖津潤 他: I/O 拡張を実現するリモート I/O システムの性能評価，電子情報通信学会技術研究報告 コンピュータシステム(CPSY), Vol.106, No.436, pp.43-48，2006 年 12 月．
10) Nobuyuki Enomoto et al.: High-speed, Short-latency Multipath Ethernet Transport for Interconnections, 16th IEEE Symposium on High-Performance Interconnects (HOTI'08), pp.75-84, Aug 2008.