社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

大規模クラスタにおける ソケットダイレクトプロトコルSDPの性能評価

緑川　博子† 　渡辺　義人† 　黒川　原佳†† 　姫野龍太郎††

† 成蹊大学 理工学部 東京都 武蔵野市吉祥寺北町３－３－１
†† 理化学研究所 情報基盤センター　埼玉県 和光市広沢２－１

E-mail: †{midori, wata}@is.seikei.ac.jp, ††{motoyoshi, himeno}@riken.jp

あらまし　ソケットダイレクトプロトコルSDPは，伝統的なソケットストリームAPIを提供しつつ，RDMAプロトコルに直接マッピングすることにより高性能な通信を可能にするプロトコルである．本報告ではInfiniBand上に実装されたSDPの通信性能を，実際の大規模クラスタにおいて評価し，ギガビットイーサネットTCPとIPoIBとも比較した．1バイトメッセージの通信レイテンシは，SDPは24.8usec，IPoIBは33.2usec，ギガビットイーサネットは21.6usecで，SDPはギガビットイーサネットに対して優位性はない．しかし64KBメッセージの一方向ストリーム通信のスループットは，IPoIBが1763Mbit/sec，　ギガビットイーサネットが941Mbit/secに対して，SDPは5378Mbit/secで，InfiniBandの理論転送レートの70%近くを達成している．さらにSDPは，64のクライアントからのマルチストリーム片方向通信に対し，総スループットとして7853Mbit/secを達成し，ギガビットイーサネットやIPoIBに比べ約7.8倍と2.6倍の高バンド幅を利用できる．SDPのCPU使用率は，IPoIBに比べ低いものの，ギガビットイーサネットに対しては優位性が観測されなかった．SDPは小さなメッセージ通信に関しては利点が少ないが，大きいメッセージの通信にはレイテンシ，スループット共に有利であることが確認された．

キーワード　ソケットダイレクトプロトコル, SDP, クラスタ，InfiniBand, IPoIB, RDMA

Performance Evaluation of Socket Direct Protocol on a Large Scale Cluster

Hiroko MIDORIKAWA†, Yoshihito WATANABE†, Motoyoshi KUROKAWA††, and Ryutaro HIMENO††

† Department of Computer and information science, Seikei University, 3-3-1, Kichijojikita-machi, Musashino-shi, Tokyo, 180-8633, Japan
†† Advanced Center for Computing and Communication, RIKEN The Institute of Physical and Chemical Research, 2-1, Hirosawa, Wako-shi, Saitama, 351-01981, Japan

E-mail: †{midori, wata}@is.seikei.ac.jp, ††{motoyoshi, himeno}@riken.jp

Abstract　The communication performance of Socket Direct Protocol (SDP) on InfiniBand (IB) is evaluated. We also compared SDP with IPoIB and TCP on gigabit Ethernet in terms of throughput and latency on a large-scaled cluster. The one-way latencies for a 1byte-message achieved by SDP, IPoIB and gigabit Ethernet are 24.8usec, 33.2usec and 21.6usec individually. It shows SDP has no advantage in a latency compared to gigabit Ethernet. The uni-directional bandwidths for 64Kbyte messages achieved by SDP, IPoIB and gigabit Ethernet are 5378Mbit/sec, 1763Mbit/sec and 941Mbit/sec. SDP achieves 70% of the theoretical bandwidth of IB(x4). Moreover 64 multiple stream aggregate bandwidth of SDP shows around 7800Mbit/sec, which is 2.6 times better than IPoIB. The CPU utilization of SDP is lower than IPoIB, but higher than gigabit Ethernet. The result shows that SDP has an advantage in the latency and the bandwidth of large-size message communication, but it has little advantage in small-size communication.

Key words　Socket Direct Protcol, SDP, Clusters,InfiniBand,IPoIB,RDMA

1. はじめに

クラスタは，今や価格性能比の優れた並列システムとしてばかりでなく，スーパーコンピュータにとってもますます重要度を増してきている．中でもコンピュータノード間通信の高速化は性能上の大きな鍵で，VIA, Myrinet, InfiniBand, Quadrics など様々な通信媒体，方式が提案開発されている．また従来のTCP/IP 通信における OS によるプロトコルスタック処理やデータの多量のデータコピーなどのオーバヘッドを改良するために，FM，GM，PM など様々なユーザレベルのプロトコルが提案されてきた．最近では，既存のイーサネット NIC に TCP Offload Engine (TOE) を組み込むなど，ノード間通信における CPU の介在を最小限に抑えるためのプロトコルオフロード技術，RDMA 技術に関する研究開発 [1] [5]，さらには LinuxOS におけるプロトコルスタックの改良開発，拡張型ソケット API の提案 [9] なども盛んに行われてきている．現在では，このような通信高速化のために工夫されたハードウエアやプロトコルをできうるかぎり駆使して，新しくプログラムを作るという場合には，大きな恩恵を被ることができる．

しかし，これらの性能を最大限に引き出すためのプロトコルや特殊な API に合わせて，応用プログラムを毎回開発することは高コストで，さらに既存の多くのネットワークプログラムが，従来のソケットベースの API で書かれていることを考えると，各通信リンク特有な API とは別に，従来からあるソケットと同様の擬似的なソケット API を提供することは，非常に重要であると認識されてきている．

このような背景をもとに，2002 年には InfiniBand Trade Association (IBTA) [3] からソケットダイレクトプロトコル SDP の仕様 [4] がオープンプロトコルとしてリリースされた．並行して，2002 年 5 月には，Adaptec, HP, IBM, Intel, Microsoft などの 7 社が RDMA Consortium (RDMAC) を組織した．2002 年 10 月から 2003 年 4 月にかけて，RDMAC は，Verb 仕様 V.1 や SCSI Extension（iSER) プロトコルと共に，TCP/IP 上の新しい RDMA へマッピングする SDP の仕様 Ver.1 をまとめた．その後 RDMAC には 50 以上の企業が所属している．

2004 年 6 月には InfiBand に対応する Linux ベースのソフトウエアスタックのオープンソース共同開発を進めるための intel, Dell, Sun などの PC/WS メーカ，Mellanox, Topspin など InfiniBand ベンダー，Oracle，ARNL，LLNL 研究所の集まりである OpenIB Alliance [2] も設立され，米エネルギー省も資金援助を表明している．

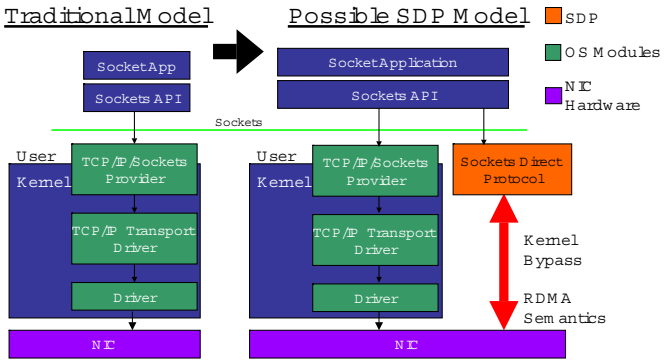

図 1　SDP の目指す実装モデル (Jim Pinkerton の発表資料より引用)

今後，ハードウエアの進化や通信周りの OS 改良，Linux との融合性 [9] などが高められることにより，さらに性能向上が期待される．すでに Linux の 2.6 カーネルには InfiniBand のドライバも組み込まれ，OS，NIC, プロトコルを含んだ，高速通信のための動きは，非常に活発になってきている．

本報告では，現状の SDP の実装でどの程度の性能があるのか，実際に大規模クラスタにおいて InfiniBand SDP (Topspin) の性能評価を行った．ここで扱うソケットダイレクトプロトコル SDP は，IBTA が策定したもので，伝統的なソケットストリーム API を提供しつつ，InfiniBand の高バンド幅などの利点を使用できるようにしたプロトコルである．また，本報告ではギガビットイーサネットにおける TCP と IPoIB（IP over InfiniBand) との比較も行った．

2. SDP と IPoIB

ソケットダイレクトプロトコル SDP は，広く使われているソケット API を，InfiniBand や iWARP といった RDMA プロトコルに直接マップするためのオープンプロトコルで，現在 2 つの定義がある．一つは InfiniBand 用に IBTA [3] の software working group (SWG) が策定したもの [4] で，もう一つは，RDMAC [5] によって iWARP 用に策定されたもの [6] である．

SDP は，手順に則ったソケットのクローズ手法，TCP ポート空間，IP アドレッシングが利用可能，コネクト/アクセプトによる結合モデル，帯域外データ転送，一般的なソケットオプションのサポートなどを特徴とする TCP/IP のような SOCK_STREAM モデルを前提にしている．UDP/IP のような SOCK_DGRAM は対象にしていない．図 1

表 1　測 定 環 境

Cluster		Express5800/420Ma BladeServer(NEC) 128nodes, 256CPUs
Node	CPU Chipset Bus Memory	Pentium Xeon3.06GHz x 2 (cache512KB) ServerWorks GC-LE 64bit 133MHz PCI-X 2GB
OS		Linux 2.4.21-27.0.2.ELscore
Network	InfiniBand	
	Card Switch	Topspin PCI-X HCA Topspin 90 Server Switch Topspin 270 Server Switch
	1gigabit Ethernet	
	Card Switch	Intel E1000/Pro Edge Switch :GeoStream SH4322G CenterSwitch: Cisco 6509

図 2　2 ノード間ラウンドトリップ通信レート

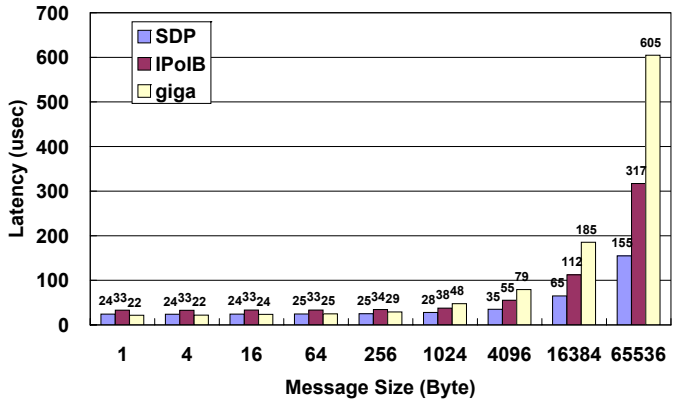

図 3　レイテンシ

に示すように，SDP の目指すモデルは，プロトコルスタック処理や割り込み/コンテキストスイッチなどのオーバヘッドやメモリ間データコピーなど，CPU 負荷やメモリバンド幅の過剰利用を防ぐ，カーネルバイパス，ゼロコピーなどを RDMA ハードウエアで行うことにより高速化を可能にしようとするものである．InfiniBand を使用した SDP であれば，InfiniBand の持つ RDMA read/write 機能や，send 機能などを用いて実装されることになるが，実装や API の詳細は指定していない．

一方，IPoIB [7] は，InfiniBand と既存のイーサネットをつなぐプロトコルとして，IP を InfiniBand にマップするためのプロトコルで，Internet Engineering Task Force (IETF) [8] の IPoIB グループが策定し，Linux InfiniBand プロジェクト [9] などが Linux にドライバを実装している．データリンク層，ネットワーク層における IP パケットのブリッジやルーティング機能をサポートするが，RDMA ハードウエアによる高速化効果は組み込まれない．

3. 測 定 環 境

今回の SDP 通信性能評価実験には RSCC(RIKEN Super Combined Cluster) の一部である 128 ノード (256CPU) の Cluster2D を使用した．クラスタとネットワーク，OS の仕様を表 1 に示す．ギガビットイーサネット は，16 ノードが Edge Switch 1 つに結合され，これが 8 セット CenterSwitch に結合される多段構成になっている．OS カーネルは，Linux 2.4.21-27.0.2.ELscore である．通信性能の測定には Netperf2.3 [10] を用いた．

4. SDP の性能評価

4.1 レイテンシ

2 ノード間でメッセージサイズを変えて，クライアントからの要求とサーバからの応答メッセージのラウンドトリップ通信の通信レートを計測した．図 4.1 は，1〜16KB のメッセージの 1 秒間の平均通信レートを示す．1〜64B 程度のメッセージでは SDP よりもギガビットイーサネット TCP のほうが高い通信レートで通信できることがわかる．また，図 4.1 は，各メッセージサイズにおける片方向通信遅延時間（レイテンシ）の値である．1B のレイテンシは，SDP は 24.8usec，IPoIB では 33.2usec，ギガビットイーサネットでは 21.6usec で，小サイズメッセージのレイテンシに関しては，InfiniBand はギガビットイーサネットに対して優位性はない．

図 4.1 は，この実験における CPU 使用率 (%) を示している．本クラスタのノードは 2CPU であるため，そのトータルを 100%としたものである．実験で用いた Netperf2.3 では，CPU 使用率の計測に，OS に依存して異なる方法を用いており，LINUX の場合には/proc/stat の CPU アイドルクロック数などの情報

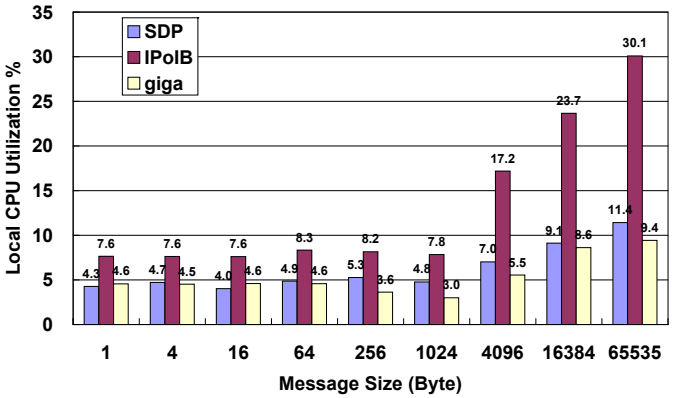

図 4　要求・応答ラウンドトリップ通信におけるクライアント CPU 使用率

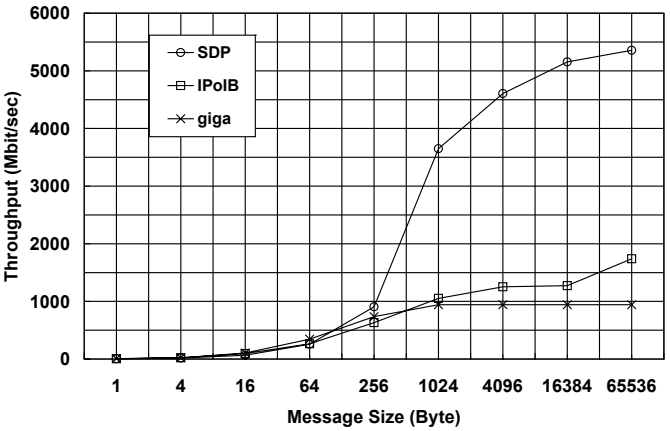

図 5　スループット

を元に計算している．これによると，IPoIB が SDP に比べ CPU を消費しているものの，ギガビットイーサネットによる TCP 通信での CPU 消費は SDP に比べ同等，あるいはむしろ少ないといえる．256B 以上のメッセージに関しては，ギガビットイーサネットではもともと低い通信性能であるため，メッセージの送受信回数が SDP に比べ少なくなり，CPU を使わないため使用率が下がると考えられる．

4.2　スループット

図 4.2 は，クライアントからサーバへ向けての片方向のメッセージストリーム通信のスループットを示す．64KB メッセージの通信では，IPoIB は 1763Mbit/sec，ギガビットイーサネットは941Mbit/sec に対して，SDP は最高 5378Mbit/sec を達成している．InfiniBand(4x) は，物理仕様としては 1 方向 10Gbit/sec であるが，802.3.z 8B/10B コーデイングを用いており，実効転送レートはその 80%で 8Gbit/sec になる．実際にはプロトコルのオーバヘッドが加わるので，データ転送レートはさらにこれよりも低下する．今回の結果は，転送理論値の 8Gbit/sec

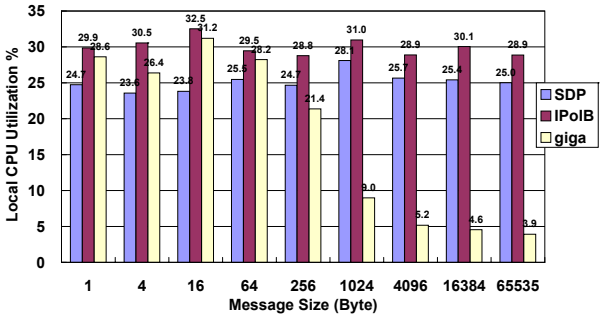

図 6　片方向ストリーム通信時のクライアント送信側 CPU 使用率

の約 70%近くを単一のストリームの SDP で達成できており，1KB 以上のサイズのメッセージ通信には SDP が威力を発揮する．しかし，64B 以下のメッセージ転送では，SDP は IPoIB やギガビットイーサネットに比べてスループットが劣っている．

この時のクライアント送信側とサーバ受信側の CPU 使用率をそれぞれ図 4.3，図 7 に示す．送信側では，メッセージサイズに依らずほぼ 30%程度の CPU 使用率であるが，ギガビットイーサネットはメッセージサイズが大きくなるにつれ，4%程度まで下がる．これは，ギガビットイーサネットの帯域が飽和して送信が待たされているためだと思われる．このように CPU 使用率はその時の通信バンド幅にも依存するので，CPU 使用率を達成バンド幅で正規化した指標であるサービスデマンド（1KB のデータ転送に必要な CPU 時間）を図 4.3 に示す．これによると，図 4.3 では 64B 以下のメッセージ通信で SDP の CPU 使用率が少ないように見えるが，実際にはバンド幅も落ちているので，CPU 消費はギガビットや IPoIB に比べ高いという結果になる．

一方，受信側の CPU 使用率は，メッセージサイズが大きくなるにつれ増加していく．SDP，IPoIB，ギガビットイーサネットの順に使用率は高い．

4.3　マルチクライアント・シングルサーバにおけるレイテンシ

大規模クラスタの利点を生かし，65 ノードを使用して，1 ノードから 64 ノードまでの複数のクライアントから，1 つのサーバへの要求送信・応答受信のラウンドトリップ通信を行った際の，それぞれのクライアントにおけるレイテンシを調べた．図 9 は，各プロトコルにおける各メッセージのレイテンシを示す．サーバへの要求が集中する環境下におけるレイテンシは，メッセージが 8KB 以上のマルチクライアントからの要求になると，レイテンシは急激に長くなっていく．図 10，図 11 に，メッセージサイズが 32B と

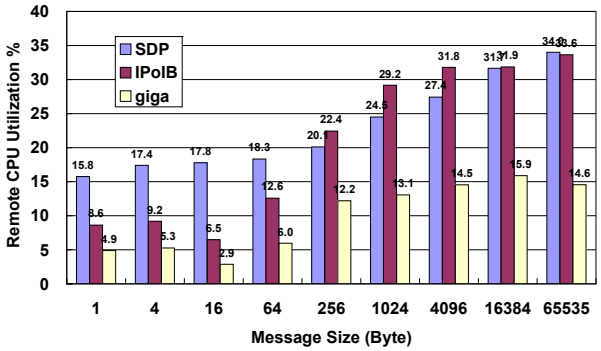

図 7　片方向ストリーム通信時の
サーバ受信側 CPU 使用率

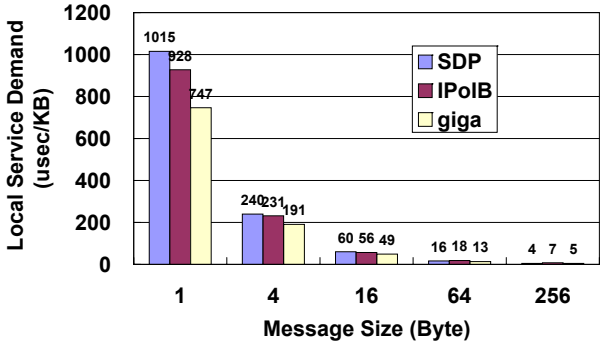

図 8　片方向ストリーム通信時の
クライアント送信側サービスデマンド

8KB におけるレイテンシを示す．32B 程度であれば，たとえ 64 のマルチクライアントからの要求であっても SDP とギガビットイーサネットとの間でレイテンシに差は生まれない．一方 8KB 以上になると，SDP のレイテンシは他に比べ小さい．しかしギガビットイーサネットとのレイテンシの比は，1 ストリームであっても 64 ストリームであっても，2.3〜2.5 で大きな差はない．従って，ギガビットイーサネットと SDP のマルチストリーム通信のレイテンシは，シングルストリームと基本的に同じ傾向を示す．

これに対し，IPoIB はマルチストリームになるとむしろ，ギガビットイーサネットに比べ，相対的にレイテンシが長くなり利点がない．

4.4　マルチクライアント・シングルサーバにおけるスループット

複数クライアントからサーバへ向けての送信マルチストリーム通信の性能を調べた．図 12，図 13，図 14 は，それぞれ SDP，IPoIB，ギガビットイーサネットにおいて，1〜64 ノードのクライアントから同一サーバにむけてのストリーム通信をした際の，総スループット（折れ線グラフ）とその時のサーバの CPU 使用率（棒グラフ）を示す．メッセージサイズは 32B〜64KB で，図 12，図 13，図 14 とも図 14 の

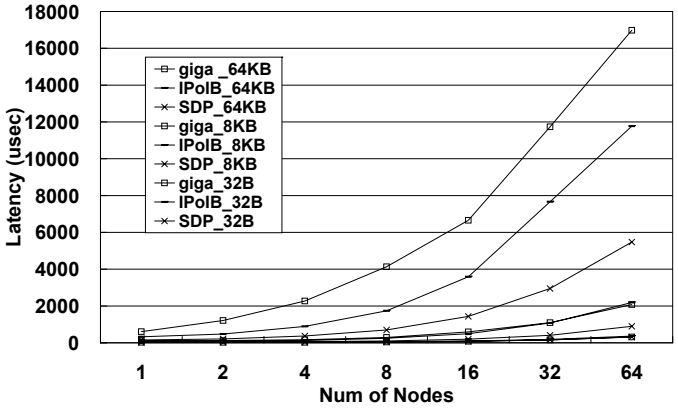

図 9　多対 1 通信のレイテンシ

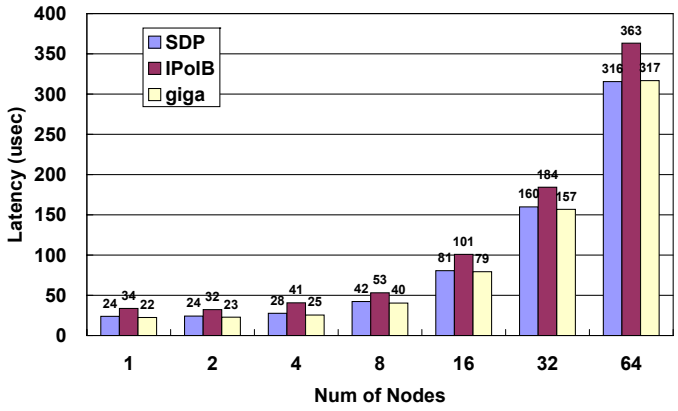

図 10　多対 1 通信のレイテンシ (32B Msg)

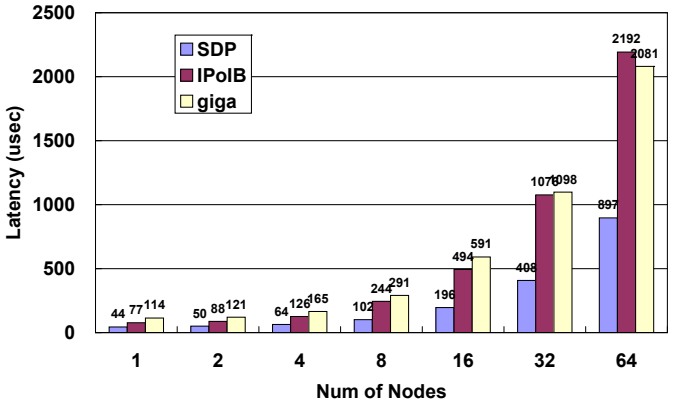

図 11　多対 1 通信のレイテンシ (8KB Msg)

凡例と同じになっている．図 12 の SDP では，小さいサイズのメッセージ通信では，クライアント数が増えるにつれて，総スループットは上昇し，十分な帯域を使えることがわかる．すなわち SDP による小さいメッセージの通信には，レイテンシは悪いがマルチストリームでは高スループットで通信ができるという利点がある．1KB 以上のメッセージ転送では，4〜64 クライアント時にはほぼ帯域は飽和しており，6600〜7600Mbit/sec 程度を示している．64 ストリー

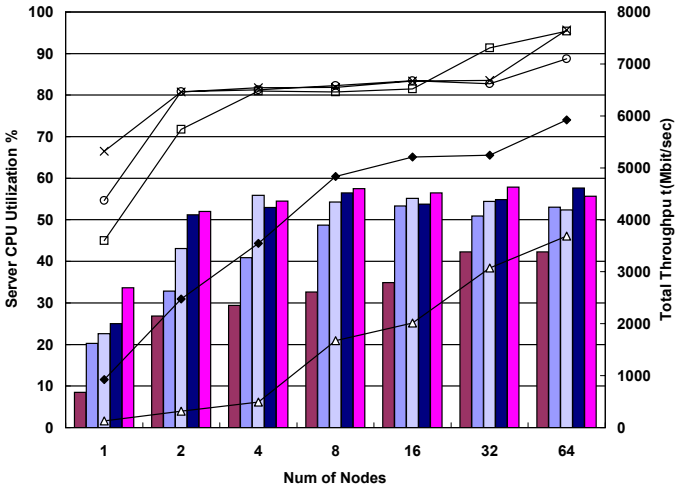

図 12　マルチクライアント通信の SDP スループット

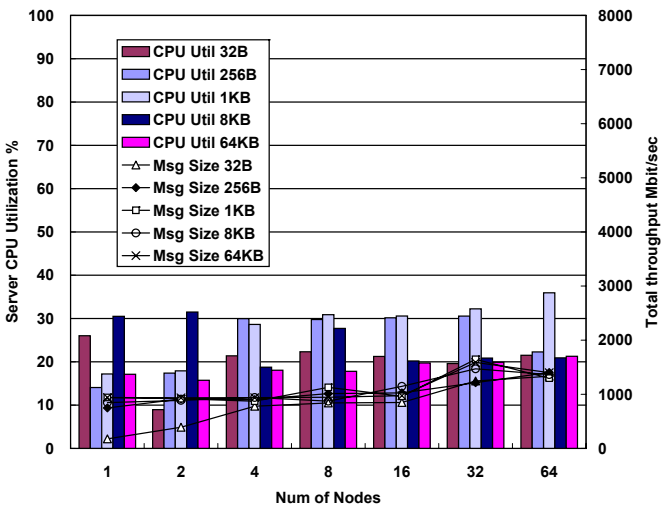

図 14　マルチクライアント通信の GEther スループット

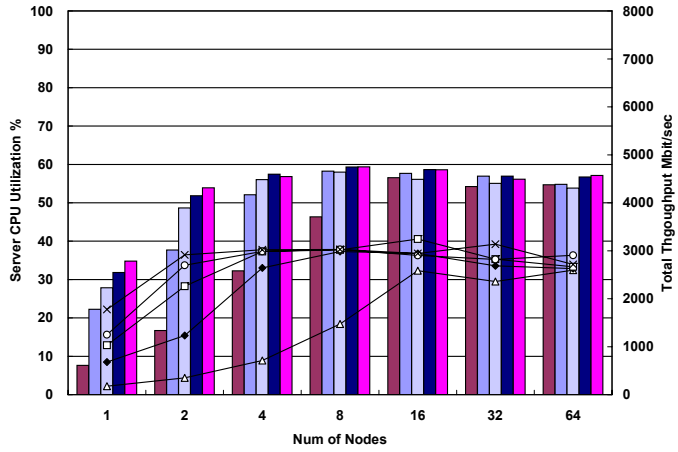

図 13　マルチクライアント通信の IPoIB スループット

ム 64KB メッセージの転送では最高 7653Mbit/sec に達している．

一方 IPoIB では，3100〜3200Mbitr/sec 程度で総スループットは飽和する．ギガビットイーサネットでは，総スループットとしては，1Gbit/sec を超える現象が見られたが，現段階では原因が明らかでない．

サーバの CPU 使用率についてみると，ギガビットイーサネットは転送速度の制限でメッセージ受信頻度が低くなるためか，サーバの CPU 使用率は SDP と IPoIB に比べ低い値にとどまっている．

5. おわりに

従来のネットワークプログラムに変更を加えずに用いることのできる SDP に関し，現状の基本性能を評価した．SDP の実装には，まだ改善の余地があるものと思われるが，現状においても，大きなメッセージ通信によるメリットは大きく，大容量ファイルやデータの転送を多く用いる応用に関しては効果があると考えられる．しかし小さいサイズのメッセージ通信に関しては利点が少ない．CPU 使用率に関しては，IPoIB に対し SDP が CPU 使用率で優位であるという報告 [11] があるが，この現象は確認できたものの，実際には広く使用されている安価なギガビットイーサネットの NIC に比べ，十分な優位性がみられなかったのには疑問が残る．

SDP はストリーム指向の設計を前提にしているため，TCP の代用として用いることができるが，パケットベースの UDP の代用として用いることができない．その意味で，IPoIB は互換性を維持するためには重要であるが，多くの点でギガビットイーサに比べ，利点は少ない．

文　献

[1] Proc. of First Workshop on Remote Direct Memory Access(RDMA): Applications, Inplementasions and Technologies,RAIT2004, Sep.2004
[2] http://www.openib.org
[3] http://www.infinibandta.org
[4] SDP 仕様: IBTA 文書, InfiniBand Architecture Specification Release 1.2, Vol.1 ANNEX A4:Socket Direct Protcol, October 2004
[5] http://www.rdmaconsortium.org
[6] J.Pinkerton et al."SDP for iWARP over TCP (v1.0)", RDMA Consortium Released Specification draft-pinkerton-iwarp-v1.0, October 2003
[7] J.Chu,et al., "Transmission of IP over InfiniBand", IETF internet draft dra http://www.ietf.org/internet-drafts/draft-ietf-ipoib-ip-over-infiniband-09.txt, January 2005
[8] Internet Engineering Task Force http://www.ietf.org
[9] http://infiniband.sourceforge.net
[10] http://www.netperf.org
[11] Pavan Balaji et.al: Socket Direct Protcol over InfiniBand in clusters: Is it Beneficial?, Ohio University Technical Report OSU-CISRC-10/03-TR54.