Vol.2009-HPC-120 No.9
2009/6/12

高性能・耐故障マルチリンク Ethernet 結合システムの性能評価

三浦信一[†1] 米元大我[†2] 塙敏博[†1,†2]
朴泰祐[†1,†2] 佐藤三久[†1,†2]

コストパフォーマンスが高い Gigabit Ethernet は，比較的中規模な PC クラスタで多く用いられている．この Gigabit Ethernet の高性能化を実現するために，Ethernet Trunking 技術が多く用いられている．しかし，これらの技術は既存の TCP/IP を用いる場合，遅延時間やバンド幅などの性能が低い問題があった．本問題を解決するためには，我々は Linux Channel Bonding と非常に近い実装で RI2N/DRV（Redundant Interconnection with Inexpensive Network with Driver）を開発している．本稿では，先行研究で開発した RI2N/DRV を実際の高性能アプリケーションに適用し評価を行った．本評価結果より RI2N/DRV が，既存手法と比較して高い性能が得られることを確認した．

Performance evaluation on Ethernet Multilink Bonding System for High performance and Fault-tolerance

Shin'ichi Miura ,[†1] Taiga Yonemoto,[†2]
Toshihiro Hanawa ,[†1,†2] Taisuke Boku [†1,†2]
and Mitsuhisa Sato[†1,†2]

Although recent high-end interconnection network devices and switches provide a high performance/cost ratio, most of the small to medium sized PC clusters are still built on the commodity network, Ethernet. To enhance performance on commonly used Gigabit Ethernet networks, link aggregation or binding technology is used. However, Theses study has the problem of mismatching with the commonly used TCP protocol, which consequently implies several problems of both large latency and instability on bandwidth improvement. The fault-tolerant feature is also supported, but the usability is not sufficient. We have developed a new implementation similar to LCB named RI2N/DRV (Redundant Interconnection with Inexpensive Network with Driver) that is very compatible with the TCP protocol. We confirmed that this system improves the performance and reliability of the network without any modification of other modules.

1. はじめに

高性能 PC サーバを相互結合した PC クラスタは，HPC 分野の様々な局面で多用されている．これらの HPC 向け PC クラスタは，比較的安価でありながら高性能であるため幅広い分野で利用されている．しかし HPC 分野では，ノード間ネットワークに対しバンド幅・耐故障性そして拡張性に関して要求が厳しい．そのため主要部品の多くにコモディティ製品を活用しつつも，ネットワークだけは専用ネットワークを選択することが多い．特に現在，クラスタの規模は大きくなり，ノード数の増加と高密度化が進んでいる．そのため，これらのシステムの持続的な安定動作のためには，それらの計算機のみならずネットワークを含めたシステム全体の耐故障性も同時に高める必要がある．

現在，最も多くのシステムで用いられているネットワーク環境は Ethernet である．特に多くのサーバシステムでは，そのコストパフォーマンスの高さから Gigabit Ethernet（以下，GbE）が幅広く用いられている．しかしより性能を求める環境においては，Gigabit Ethernet よりも高バンド幅を持つ 10 Gigabit Ethernet（以下，10GbE）など既存の Ethernet 技術を拡張した高速ネットワークも用いられている．一方で，InfiniBand[1] や Myrinet[2]，高速な専用ネットワークが使用される．しかし，このようなバンド幅が高いネットワークは，コストパフォーマンスが良くなりつつあるが，現在最もコストパフォーマンスが高い GbE と比較して高価であり，ネットワーク構築のコストが問題となる．加えて，これらのネットワークシステムは単純に構成しただけでは，故障に対して脆弱であり，何らかの冗長性を確保する必要がある．

このような中，我々は高いバンド幅と耐故障性を持つ高性能クラスタ向けネットワーク RI2N[3][4][5] を研究・開発してきた．RI2N はマルチリンクの Ethernet を用いて，平常時にはそれらの複数の Ethernet を同時に利用することにより高いバンド幅を実現する．また，ネットワークの異常時においては，ネットワークの冗長性を利用し正常なネットワークのみを用いて通信が継続することを可能にする．我々は，この RI2N のコンセプトを元にユーザ透過に本機能を提供可能な RI2N/DRV を開発している[5]．本稿では，先行研究で開発してきた

†1 筑波大学 計算科学研究センター
Center for Computational Sciences, University of Tsukuba
†2 筑波大学大学院 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

RI2N/DRV を，並列プログラムベンチマークを用いて評価する．また，HPC クラスタの利用のみならず，RI2N/DRV の高いユーザ透過性を利用し，より一般的な UNIX 環境のネットワークに適用し，その有効性についても検証する．具体的な応用例として，一般的なネットワークファイルシステムである NFS のネットワークに適用し，RI2N/DRV によって得られるバンド幅について評価する．

2. RI2N

2.1 概　　要

我々は高性能クラスタ向けに複数リンクの Ethernet を同時に利用することによって高いバンド幅と耐故障性を同時に実現する RI2N（Redundant Interconnection with Inexpensive Network）というコンセプトを提唱し，それを実現するシステムを提案・実装している[3)4)5)]．RI2N とは，安価な複数リンクの Ethernet とソフトウェアの拡張のみで高バンド幅化と信頼性の向上を同時に実現することを目指すものである．具体的には各ノード間に複数リンクの Ethernet ネットワークを設置し，正常時にはデータのストライピングによってスループットを向上させる．そしてリンクが故障した場合には，冗長なリンクを利用して通信を継続させる．

このような機能は一般的には Ethernet トランキングと呼ばれるものであり，いくつかの先行研究が挙げられる．まず高性能クラスタでの利用に特化した軽量通信ライブラリ PM/Ethernet[6)7)] がある．PM/Ethernet は複数のネットワークを同時に利用する機能として，PM/Ethernet Network Trunking を持っており，遅延時間，スループットにおいて高い性能を得ることができる．一方で PM/Ethernet は既存の UNIX Socket とは互換性のない専用の通信体系と API を用いるため，プログラムの可搬性や相互運用性に大きな問題がある．ハードウェアもしくはソフトウェア機能により，マルチリンク Ethernet を用いるシステムとして，すでに IEEE 802.3ad[8)] によって規格されている Link Aggregation Control Protocol がある．本技術は主に 2 台のスイッチ間もしくは，ノードとスイッチ間に 2 つ以上のネットワークリンクを用意し，高バンド幅と耐故障性を同時に実現する．しかし，本技術はスイッチの冗長化は考慮にいれておらず，スイッチ自体の故障に対処することは難しい．また，専用スイッチの導入も必要とする．最も RI2N コンセプトに近いものとして，Ethernet におけるマルチリンクの利用をノード間にも適用するドライバソフトウェアとして Linux Channel Bonding[9)]（以後，LCB）がある．LCB はいくつかの通信モードがあり，経路の多重化との 2 ノード間の高バンド幅を同時に実現するモードとして balance–rr モードがある．しかし本機能はパケット到着順序の入れ替わりが頻繁に発生し，TCP/IP では性能を十分に発揮することが難しいことが示されている[9)]．また，LCB では ARP 機能を用いてリンクの故障判断を行っているが，検出可能な範囲が限定され，耐故障性については十分に配慮されていない．

(a) IEEE 802.3ad　　(b) RI2N/DRV

図 1　IEEE 802.3ad と RI2N/DRV のネットワーク構成

これらの問題を解決するために，我々は RI2N のコンセプトを実現する 1 つの実装として RI2N/DRV を開発している．ここでは，RI2N/DRV について，本稿を理解するための最小限の説明を行う．詳細については文献[5)] を参照されたい．

2.2 RI2N/DRV の実装

RI2N/DRV は RI2N のコンセプトに基づき，LCB と同様に仮想的な Ethernet デバイスとして実装されている．RI2N/DRV は Linux のローダブルモジュールで実装されており，OS に対して既存の Ethernet デバイスドライバのように振る舞う．そして既存の Ethernet 環境の上位プロトコルである TCP/IP，UDP/IP もしくは ARP などのプロトコルを，一切の変更なく利用することを可能にする．

図 1(a) に示すような IEEE 802.3ad を用いたネットワーク構成と異なり，RI2N/DRV では，スイッチの故障にも対応するために，**図 1(b)** のようにスイッチについても多重化し，2 ノード間で全く異なる 2 つ以上の経路を用意することで 1 台のスイッチが故障した場合にも通信が継続できるようにする．

システム全体の基本的な仕組みは LCB で balance–rr モードを用いた場合と類似する．しかし，LCB や IEEE 802.3ad でのリンク結合のように非常に単純な処理のみを行うのではなく，RI2N/DRV では順序入れ替えのような比較的複雑な処理を行い，ある程度のオーバヘッドを許容した上で性能改善を目指している．RI2N/DRV は，基本的に上位レイヤにどのようなプロトコルが使われても対応可能であるが，現在最も利用されている TCP/IP の挙動を主眼においた実装となっている．以後の説明では，TCP/IP の使用を想定し説明する．RI2N/DRV

の重要な機能として，**パケット到着順序制御機構**と**故障/回復検出機構**の 2 つがある．

マルチリンクのネットワークを用いる場合，物理的な距離やネットワークの混雑状態により，パケット到着順序に不整合が生じる場合がある．下位レイヤでパケットが比較的順序正しく到着することを期待する上位プロトコルでは，これは性能低下の原因になる．通信プロトコルである TCP/IP は，パケット到着順序の狂いを許容できる仕組みを持つ．しかし，定常的に到着の順序の狂いが発生するマルチリンクの環境では，この仕組みを用いてもスループットの低下を招き，またこれを回避するためにはプロセッサに高負荷が生じる．この順序の不整合の最も大きな原因は，NIC が提供する interrupt coalescing 機能[10] などによる受信バッファ蓄積である．基本的にこの順不整合を，送信や受信の工夫のみで取り除くことは難しい．RI2N/DRV ではこの問題を取り除くために，新たに RI2N ヘッダを導入し，受信側で順序並び替えを行うことでこの不整合を取り除く．このヘッダは Ethernet ヘッダ以降のペイロード部に格納されるため，使用できるスイッチや NIC などのハードウェアに制限がない．RI2N/DRV では，このヘッダを用いて簡略された順序制御を行うが，再送制御や輻輳制御などは行わない．これらの機能は，上位層のプロトコルである TCP/IP などに任せる．

マルチリンクのネットワークで 1 系統のネットワークに故障が発生すると $1/n$（n はネットワークの冗長度）の確率でパケットが損失される．TCP/IP では，パケット損失が確認されると混雑が原因であると推測し，congestion control により window サイズが低下する．しかし，パケット損失の原因が混雑ではなく故障である場合には，window サイズが低下したとしても一定の確率でパケット損失が継続する．それによって，いつかは window サイズがきわめて 0 に近づきネットワークが停止と同様の状態になる．これを回避するために，なるべく早期にネットワークの故障を検出する必要がある．そこで RI2N/DRV では，バースト転送時にはパケット到着数の偏りを用いて故障の検出を早めている．また，バースト転送状態以外においても故障を検出する機構として，ハートビートも併用している．RI2N/DRV では，このパケット到着の偏りとハートビートの 2 種類の故障検出機構を利用することで，様々な故障のパターンに対応できる．回復の検出には検出に比較的長い時間を要するハートビートを用いるが，回復には一般的に人手を要するため問題ない．

3. 性 能 評 価

本章ではマルチリンク Ethernet 環境での RI2N/DRV の有効性について評価する．すでに先行研究[5] では，遅延時間・スループット等の基礎的な性能評価を行い，それらの結果により RI2N の有効性を示している．本稿では先行研究の基礎的な性能評価結果を踏まえて，新たに双方向通信における性能評価を行う．また，既存の MPI 実装を用いた高性能計算のアプリケーションの例として，NAS Parallel Benchmarks[11]（以下，NPB）を評価する．これらに加えて，RI2N の応用アプリケーションの例として，UNIX のネットワークサービスの代表例の 1 つである NFS サービスに適用し，その性能を評価する．

表 1 PC-A

Item	Specification
CPU	Intel Xeon E5110 1.6GHz dual-core
Memory	PC2-4200 2048MB
NIC	Intel PRO1000PT dual port 1000base-T

表 2 PC-B

Item	Specification
CPU	Intel Xeon E3110 3.00GHz dual-core
Memory	PC2-6400 8192MB
NIC	Intel PRO1000PT dual port 1000base-T

表 3 ソフトウェア環境

Item	Specification
OS (Kernel)	Linux 2.6.27.21
NIC driver	Intel PRO/1000 Network Driver
	e1000e 0.3.3.3-k6
Bonding driver	Ethernet Channel Bonding Driver, v3.3.0
MPI	Open MPI v1.2.4

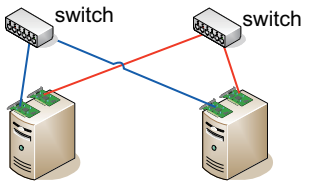


図 2 評価システムのネットワーク構成

評価には，**表 1** および**表 2** に示すハードウェアを有する，2 種類のノードを用いる． それらのソフトウェア環境は共通であり**表 3** に示す構成となっており，**図 2** に示すようなネットワーク構成で評価した． すべてのノードをそれぞれ 1 台のスイッチ（Dell Power Connect 5324）によって構成された物理的に分けられた 2 系統のネットワークに接続し，次に示す 3 種類のネットワーク環境について比較する．

single 1 系統のネットワークのみを使用し，通常のシングルリンクの GbE として使用する
LCB 2 系統のネットワークを Linux Channel Bonding を用いて同時に使用する
RI2N 2 系統のネットワークを RI2N/DRV を用いて同時に使用する

3.1 双方向通信

2 ノード間で送信・受信を同時に行う，双方向通信時の性能特性について評価する．評価には表 2 に示す PC-B のノードを 2 台用いた．1 台のノードがデータをバースト送信し，もう 1 台のノードはそのデータを受信の後，送信側に再送信する．送信側がこれを受信することで，双方向の通信状態を作る．このとき初めにデータを送信するノードで観測された 120 秒間の受信スループットの変化を**図 3** に示す．

評価結果では，RI2N は約 214 MB/sec の性能を得られた．single では約 117 MB/sec の性

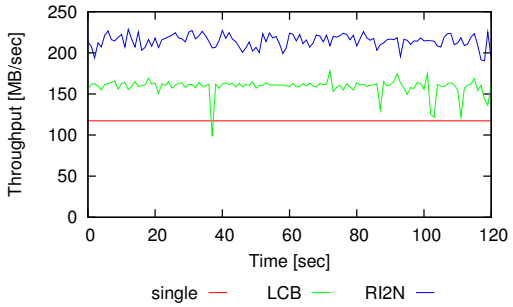


図 3 双方向通信時のスループットの変化

表 4 1.0 GB 送信時の送受信パケット数

	TX Packets	RX Packets
single	741,535	370,776
LCB	743,439	741,530
RI2N	754,616	376,004

能を得られていることから，約 83%の性能向上となる．一方，LCB では約 159 MB/sec の性能が得られた．LCB は single よりも性能は向上しているが，同じマルチリンクを用いる RI2N と比較して性能向上率は低い．この原因として，2.2 節で示した，LCB の受信側でパケットの到着順序不整合の発生が考えられる．TCP/IP を用いた通信機構では，この順序不整合によって受信側より送信側に対し ACK パケットを送信することになる．この現象を確認するために，2 ノード間で TCP/IP を用いてデータサイズ 1 GB の片方向バースト転送を行い，送信側の送受信パケット数を観測する．結果を**表 4** に示す．結果では，指標となる single に対して，RI2N では送受信パケット数に大きな差はない．一方，LCB では送信パケット数は single と同じであるが，受信パケット数は single の約 2 倍になっており，この受信パケット数の差は順序不整合により発生した ACK パケットである．現在の Linux Kernel で実装され，本評価で用いた TCP の輻輳制御である CUBIC[12] では，送信側受信側から送られてくる ACK パケットの原因を推測する機構がある．これにより，順次不整合による ACK パケットを受信した場合には，再送処理を行わない．これにより，片方向通信時においてスループットが向上するが，この推測処理により CPU の処理量が多くなる．加えて，受信側からの ACK パケット数は減らないため，この受信処理も必要になる．双方向にデータを交換する通信では，大量の ACK パケットも双方向に送受信されることになる．大量の双方向の ACK パケットがデータストリームに大きな影響を与える．結果として図 3 の評価結果のように，マルチリンクを用いたとしても性能向上率は低くなる．一方，RI2N では TCP/IP の処理前にパケット到着順序制御機構を有する．そのため，パケットの到着順序の不整合がなくなり，受信側から送信側への ACK パケットが少なくなる．その結果，RI2N では LCB と比較して逆方向のデータストリームに悪影響を及ぼす ACK パケットが少なく，高いスループットを得ることができる．

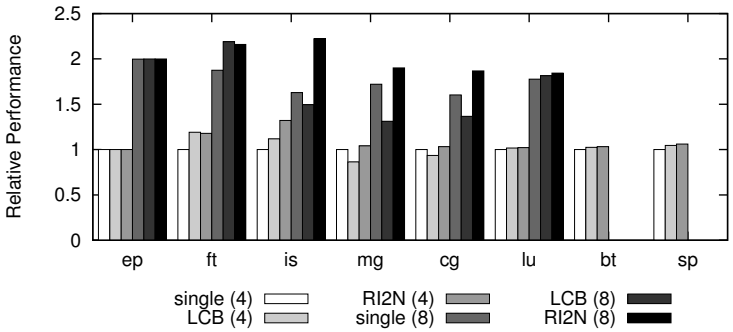


図 4 NAS Parallel Benchmarks 評価結果．() 内はプロセス数を示す

このような RI2N と LCB が示す双方向通信の性能差は，実アプリケーションに大きな影響を与える可能性がある．特に MPI を用いた高性能計算のアプリケーションでは，*MPI_Isend* や *MPI_Irecv* を用いて双方向通信を行い，通信部分の高性能化を行っている．そのような通信を用いるアプリケーションでは，双方向通信に安定した高いスループットを必要とする．また，高性能計算に関わらず，アプリケーションによって意図しない双方向通信が発生する場合もある．3.3 節では，このようなシステムアプリケーションの例として NFS を取り上げ，双方向通信の性能が NFS システムの性能へ与える影響について評価する．

3.2 NAS Parallel Benchmarks

次に，実際の高性能計算アプリケーションを想定して NPB[11] を評価する．評価では表 1 に示した PC-A を用いて，合計 8 ノードのクラスタを構築する．本評価では NPB ver.3.3 CLASS=B を用いる．MPI のプロセス数は 4 および 8 プロセスの 2 種類を評価し，さらにメモリバンド幅等の影響を避けるため，dual-core CPU ではあるが 1 ノードにつき 1 プロセスとする．Kernel (ep, ft, is, cg, mg) と Application (lu, sp, bt) をベンチマークの対象とし，4 ノードにおける single との相対性能を評価する．なお sp と bt は問題の性質上，n^2 のノード数が必要で 8 ノードでは実行できないため，4 ノードでのみ評価する．**図 4** に評価結果を示す．

まず，ep では計算中の通信はほとんどないため，ネットワーク性能を必要としない．そのため，どのネットワーク環境においても性能は変わらない．

ft の結果では，LCB と RI2N は single と比較して性能が向上している．しかし，RI2N は

4 ノードおよび 8 ノードのいずれの場合でも，LCB の性能に対して性能が低くなる結果が得られた．この理由として，ft の通信では比較的大きなデータを転送することが考えられる．現在の TCP アルゴリズムでは，順序の不整合が続くとそれを学習する仕組みがある．それによって LCB のように順序不整合が発生する環境を用いた場合でも，転送するデータサイズが大きくなることで，高いスループットを得ることができる．一方で RI2N の場合では受信側で順序制御などの処理を行うため，この処理がオーバヘッドとして現れ，これが性能に悪影響を与えたと考えられる．

is の結果では，RI2N は single や LCB に対し大きな性能向上がみられている．前述の ft の主な通信は *MPI_Alltoall()*，is で行われる通信は *MPI_Alltoallv()* であり，通信パターンに大きな差はない．一方でこのような異なる結果が得られた理由として，ft と is の通信するデータサイズが異なることが考えられる．is は ft と比べて小さいメッセージサイズのデータ送受信を繰り返す．そのため，各メッセージサイズに対して得られるスループット性能の影響を受けやすい．先行研究[5] の評価では MPI におけるスループットは，LCB と比較して RI2N は通信性能の立ち上がりが早い結果を示している．この性能差が is における LCB と RI2N の性能差として表れていると考えられる．

mg および cg では，LCB が single よりも性能が低下しているのにも関わらず，RI2N では性能向上を示した．mg および cg の通信の大部分は *MPI_Send()* と *MPI_Irecv()* を用いたデータの交換である．これは 3.1 節の評価で示した双方向通信となる．RI2N は single に対して安定して高いスループットを得られるが，LCB では安定して良い性能を得られていない．この結果 LCB では single よりも性能が低くなったと推測する．特に MPI のプロセス数が 4 から 8 へと増加すると，LCB の各ノード数での single に対する相対性能がそれぞれ mg では 14%から 24%，cg では 6%から 15%と性能低下し，その割合が大きくなる．一方で RI2N は mg では 4%から 10%，cg では 3%から 16%と性能向上の割合が大きくなり，RI2N と LCB の性能差がノード数の増加に伴って，顕著な差となって現れる．ノード数が増加することによって，計算に対する通信回数が増加し，それにより通信が与える影響が大きくなったためと考えられる．

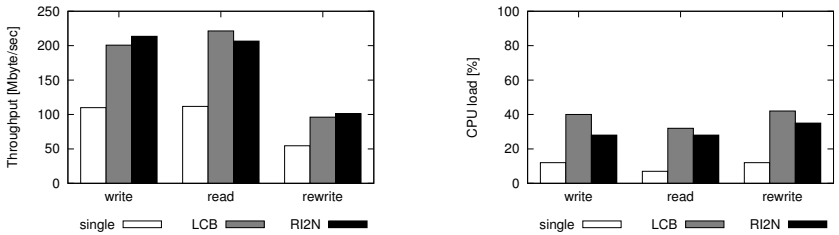
lu，bt および sp では RI2N，LCB 共に single に対して僅かに性能が向上した．しかし，single に対して LCB および RI2N の性能向上率は小さい．これは計算時間に対して通信に占める時間が小さく，通信バンド幅の向上の効果が得られないためと考えられる．

各ベンチマークの結果から，RI2N を用いることで single と比較して概ね性能向上が得られた．しかし，single と比較して，潜在的なバンド幅を約 2 倍持つ RI2N であるが，それに見合う性能向上は得られなかった．RI2N は既存の Ethernet と比較して高いバンド幅を得られるが，遅延時間については改善されず，むしろ僅かに増加する．NPB においては，バンド幅よりも遅延時間に左右されるようなベンチマークも多く，すべてのアプリケーションで RI2N の絶対性能を生かすことは困難である．しかし 4 ノードでの single と比較すると，8 ノードでの RI2N では 1.8 倍から 2.2 倍の性能向上が得られている．RI2N によって拡大された通信バンド幅が，ノード数が増加した際に性能向上のスケーラビリティを支えていると考えることができる．

3.3 NFS

RI2N はユーザ透過な実装であり，どのような UNIX アプリケーションにも適用できる．我々はこの高い透過性を利用し，高性能計算などのノード間通信以外への応用として，UNIX の標準的なネットワークサービスの 1 つである NFS サービスに RI2N を適用した．ここで，RI2N を用いて NFS サーバ・クライアントの通信環境を構築し，同システム上で性能を評価する．評価環境として，表 2 に示す PC-B を 2 ノード用意し，2 系統の GbE で構築されたネットワークを用いた．そのため，200MB/sec 以上の最大性能が得られる．一般に，これに見合う十分な I/O 性能を備えたストレージは高機能な RAID システムとなる．本評価で実験機材の都合上，そのような高性能ストレージの代わりに，NFS サーバ上にメモリファイルシステムである tmpfs を用いて 4.0 GB のファイルシステムを構築し，NFS によって export する領域として用いることで，ストレージが性能に与える影響を最小限にしている．本環境でファイルシステムの直接 I/O 性能を調べたところ，約 1.4 GB/sec の write 性能が得られ，今回の実験で用いる NFS サーバのネストレージ性能として十分な I/O 性能が得られた．このような NFS の環境において，NFS サーバのスループットについて評価する．評価では bonnie++ v1.03[13] を用いた．ファイルキャッシュの影響を最低限に抑えるため，クライアント側のメモリサイズに対して，倍のファイルサイズの read/write を行う．そのため，クライアント側はメインメモリの容量を Linux カーネルの起動パラメータで 1GB に制限し，read/write を行うファイルサイズを 2GB に設定した．**図 5(a)** と**図 5(b)** に，bonnie++で得られた各ベンチマークのスループットと，その時の CPU ロード値を示す．

write は *write()* システムコールを用いてブロック単位で書き込みを行うベンチマークである．評価結果では，single，LCB および RI2N でそれぞれ 110 MB/sec，201 MB/sec，213 MB/sec となった．single の性能に対して LCB，RI2N は 82%，94%の性能向上を示しており，2 本の GbE を用いることにより性能が大きく向上する．一方，図 5(b) に示す CPU のロード値は，LCB が 40%に対して RI2N が 28%と，RI2N は LCB に対して 12%も CPU の



(a) スループット　　(b) CPU 使用率

図 5　bonnie++ベンチマーク結果

負荷が小さい．この結果より，LCB に対して小さな CPU 負荷でほぼ同等の性能を RI2N で実現できることが分かる．

read は *read()* システムコールを用いてブロック単位で読み込みを行うベンチマークである．single，LCB および RI2N はそれぞれ 111 MB/sec と 221 MB/sec，207 MB/sec の性能を示している．RI2N は LCB に対して 6%性能が低い結果となっている．本性能低下の原因については，より詳細な解析が必要である．

rewrite は *read()* と *write()* の処理を繰り返して行う．single，LCB そして RI2N のスループットはそれぞれ 55 MB/sec，96 MB/sec そして 101 MB/sec のスループットを示した．差は小さいが，LCB に対して RI2N のスループットの向上を得られる．また CPU 負荷についても，LCB が 42%にたいして RI2N が 35%と 17%も低く，RI2N が有効に機能していることが分かる．

4. おわりに

本稿では，PC クラスタ向けに開発されたマルチリンク Ethernet 環境を用いた高バンド幅かつ耐故障性を持つネットワークシステムである RI2N/DRV を，既存の MPI 実装を用いた高性能計算のアプリケーションの例として，NAS Parallel Benchmarks に適用し評価した．その結果，計算の種類によっては LCB と比較して RI2N/DRV は高い性能向上を示した．RI2N/DRV は既存技術である LCB と比較して，ネットワークの拡張性が高く，現実的なネットワークシステムとして既存の UNIX サービスに適用することが可能である．そこで，RI2N/DRV を UNIX のネットワークサービスへの応用についても検討を行い，本システムをネットワークファイルシステムの 1 つである NFS に RI2N 適用した．それにより，既存のシングルリンクの GbE をそのまま利用する場合と比較して，LCB よりも低い CPU 負荷でほぼ同等の性能を示した．これらの結果より，既存のマルチリンクを用いるネットワーク環境である LCB と比較し，RI2N/DRV が高い性能が得られる．

謝辞　本研究の一部は，科学技術振興機構戦略的創造研究推進事業（CREST）研究領域「実用化を目指した組込みシステム用ディペンダブル・オペレーティングシステム」，研究課題「省電力高信頼組込み並列プラットフォーム」による．

参 考 文 献

1) InfiniBand Trade Association: InfiniBand.
2) Myricom: Myrinet.
3) Miura, S. et al.: RI2N - Interconnection Network System for Clusters with Wide-Bandwidth and Fault-Tolerancy Based on Multiple Links, *ISHPC-V*, Lecture Notes in Computer Science, Vol.2858, Springer, pp.342–351 (2003).
4) 岡本高幸ほか：Ethernet マルチリンクによる PC クラスタ向け高バンド幅・耐故障ネットワーク RI2N/UDP，情報処理学会論文誌. コンピューティングシステム， Vol.48, No.8, pp.153–164 (2007).
5) 岡本高幸ほか：ユーザ透過に利用可能な高性能・耐故障マルチリンク Ethernet 結合システム，情報処理学会論文誌. コンピューティングシステム， Vol.Vol.1 No.1, No.8, pp. 12–27 (2008).
6) Sumimoto, S. et al.: PM/Ehernet-kRMA: A High Performance Remote Memory Access Facility Using Multiple Gigabit Ethernet Cards, *CCGrid 2003*, pp.326–333 (2003).
7) Sumimoto, S. et al.: A scalable communication layer for multi-dimensional hyper crossbar network using multiple gigabit ethernet, *ICS '06: Proceedings of the 20th annual international conference on Supercomputing*, pp.107–115 (2006).
8) IEEE: IEEE 802.3ad – Link Aggregation (2000).
9) Davis, T.: Linux Ethernet Bonding Driver.
10) Intel Corporation: *Intel PRO Network Connections User Guides*.
11) Bailey, D.H. et al.: The NAS parallel benchmarks–summary and preliminary results, *Supercomputing '91: Proceedings of the 1991 ACM/IEEE conference on Supercomputing*, New York, NY, USA, ACM, pp.158–165 (1991).
12) Ha, S. et al.: CUBIC: a new TCP-friendly high-speed TCP variant, *SIGOPS Oper. Syst. Rev.*, Vol.42, No.5, pp.64–74 (2008).
13) Coker, R.: Bonnie++ : benchmark suite of hard drive and file system performance.