PM/InfiniBand-FJ：InfiniBandを用いた大規模PCクラスタ向け高性能通信機構の設計

住元真司[†] 成瀬彰[†] 久門耕一[†]
細江広治[††] 清水俊幸[††]

本論文では，10 Gbps クラスのネットワークである InfiniBand を用いた大規模 PC クラスタ向けの高性能通信機構 PM/InfiniBand-FJ の設計について述べる．PM/InfiniBand-FJ は PC クラスタで商用スーパコンピュータを凌駕する性能とそれに匹敵する信頼性を実現するため開発された．InfiniBand は，サーバ間通信や I/O 通信を主要ターゲットとして設計されている．この InfiniBand を 1,000 台を超える高性能計算用途の大規模 PC クラスタ上で高い通信性能を実現し，信頼性の高い運用を実現するためには，性能面，運用面で問題がある．このため，PM/InfiniBand-FJ では，オリジナルの InfiniBand 仕様を拡張して問題を解決している．PM/InfiniBand-FJ を SCore クラスタシステムソフトウェア上に実装し性能評価を行った．評価の結果，Xeon 2.8 GHz プロセッサ，ServerWorks GC-LE チップセット搭載の PC クラスタで，913.2 MB/s の通信バンド幅性能と，15.6 μs のラウンドトリップ時間を実現している．また，Xeon 3.06 GHz を搭載した富士通 PRIMERGY RX200 256 ノードの PC クラスタで姫野ベンチマークの結果では，242.8 GFlops と HP Alphaserver SC45（1 GHz 256 CPU）に比べ 2.4 倍高い性能を実現している．

PM/InfiniBand-FJ: A Design of High Performance Communication Facility Using InfiniBand for Large Scale PC Clusters

Shinji Sumimoto,[†] Akira Naruse,[†] Kouichi Kumon,[†]
Kouji Hosoe[††] and Toshiyuki Shimizu[††]

This paper describes a design of high performance communication facility called the PM/InfiniBand-FJ using InfiniBand interconnect for large scale PC clusters. The PM/InfiniBand-FJ has developed to realize higher application performance than commercial supercomputers and comparable availability to them. The specification of InfiniBand interconnect is mainly designed for communication among servers and I/O communication, so there are some issues to use InfiniBand for high performance computation on over 1,000 node PC clusters in the point of performance and system operation. Therefore, the PM/InfiniBand-FJ solves the issues by expanding the original specification of InfiniBand. We have implemented the PM/InfiniBand-FJ on SCore cluster system software, and evaluated the communication and application performance. A 913.2 MB/s of bandwidth and 15.6 μs round trip time have been achieved on Xeon 2.8 GHz PC with ServerWorks GC-LE chipset. The result of Himemo benchmark achieves good scalability, the result is 242.8 GFlops on 256 node PC cluster (Fujitsu PRIMERGY RX200 with Xeon 3.06 GHz) which is 2.4 times faster than HP Alphaserver SC45 (1 GHz 256 CPU).

1. はじめに

高性能計算分野において，ベクタ型計算機や超並列計算機に代表される従来型スーパコンピュータは長年にわたる実績を背景に企業や計算機センタユーザを中心に多くのユーザに利用されている．従来型スーパコンピュータは専用の計算機ハードウェアと 10 Gbps クラスの専用の高性能ネットワーク，そして，バッチシステムや運用管理機構を備えている．長期間ジョブを実行するため，システム保守時にもそれまで実行していたジョブの実行イメージをいったんディスクシステムに格納し，保守終了時に，ディスクシステム上のイメージからジョブを再実行させるチェックポイントリスタート機構を備えているものが多い．

従来型スーパコンピュータに加え，多数の PC や PC サーバをクラスタ結合した PC クラスタは高性能

† 富士通研究所
Fujitsu Laboratories Limited
†† 富士通
Fujitsu Limited

計算の重要なプラットホームの 1 つとなっている．これまで研究機関や大学での利用が主であったが，最近では企業の製品開発におけるシミュレーションにも利用されており，価格対性能比の高さを背景に導入事例はますます増えている．

PC クラスタが利用されるようになった 1990 年代前半においては，計算性能は当時の最新鋭の従来型スーパコンピュータには及ばなかった．しかし，PC の技術革新とともに高速化を続け，現在では，2003 年 11 月の TOP 500[1] に象徴されるように上位 10 システム中の半分以上（6 システム）を PC クラスタシステムが占めるまでになった．これは，PC クラスタが最新鋭の従来型スーパコンピュータを凌駕する実行性能を達成できることを示している．

さて，PC クラスタがここまで広く使われるようになったのは，マイクロプロセッサの性能向上を主とした PC のハードウェア技術の進歩と Gbit クラスの高速ネットワークの普及が大きい．現状は，Gbit クラスの高速ネットワークには Myrinet[2] や QsNet (ELAN)[3] といったクラスタ専用インタコネクトや Gigabit Ethernet が使われている．

しかし，近年，InfiniBand[4] や 10 Gb Ethernet[5] といった 10 Gbps クラスのネットワークが登場し，従来型スーパコンピュータに匹敵する通信性能が得られるようになった．我々は 10 Gbps クラスのネットワークと十分な運用管理機構を搭載した PC クラスタを実現することにより，PC クラスタのユーザ層をよりいっそう広げられると考えている．

そこで我々は大規模 PC クラスタにおいて高い通信性能を実現し，従来型スーパコンピュータのユーザの利用に耐えうる PC クラスタを実現するため，10 Gbps クラスのネットワークである InfiniBand を用いて，チェックポイントリスタート機能を実現する PM/InfiniBand-FJ を開発した．

本論文では，10 Gbps クラスのネットワークである InfiniBand を用いた大規模 PC クラスタ向けの高性能通信機構 PM/InfiniBand-FJ の設計について述べる．

InfiniBand は，元々ビジネス系システムにおけるサーバ間通信や I/O 通信を主要ターゲットとして設計されているが，これをそのまま 1,000 台を超える高性能計算用途の大規模クラスタで信頼性の高い大規模並列計算を行うには問題がある．このために，我々は，オリジナルの InfiniBand の仕様を拡張しこの問題を解決している．

PM/InfiniBand-FJ を SCore クラスタシステムソフトウェア[6] 上に実装し，Xeon 2.8 GHz プロセッサを搭載の PC クラスタで評価した結果，913.2 MB/s の通信バンド幅性能と，15.6 μs のラウンドトリップ時間を実現している．また，Xeon 3.06 GHz を搭載した富士通製 PRIMERGY RX200 256 ノードの PC クラスタで姫野ベンチマークを評価した結果，242.8 GFlops と従来型スーパコンピュータに比べ 2.4 倍高い性能を実現している．

本論文は次のように構成されている．2 章では，従来型スーパコンピュータのユーザの利用に耐えうる PC クラスタシステムの実現に必要な通信機構の条件をまとめる．3 章では，InfiniBand の特徴についてまとめる．4 章では，PM/InfninBand-FJ 実現上の課題を述べ，5 章でその設計について，6 章では実装の概要を述べる．7 章では PM/InfiniBand-FJ の遅延，バンド幅性能とアプリケーション性能について評価する．8 章に高速通信ライブラリに関する関連研究をまとめ，9 章に結論を述べる．

2. 大規模 PC クラスタシステムにおける通信機構の要件

従来型スーパコンピュータのユーザの利用に耐えうる 1,000 台を超える大規模 PC クラスタシステムを実現するための通信機構の要件を次に示す．

ハードウェア性能を最大限引き出せること： 10 Gbps クラスのネットワークハードウェアのバンド幅性能を最大限に引き出し，かつ，低遅延であること．

スケーラブルであること： 計算ノード台数に比例して，計算性能が向上するとともに，計算ノード数が増えてもメモリ記憶などの計算機資源消費の増加を最小限に抑えられること．

長時間ジョブにおける耐故障性を有すること： 数週間にまたがる長時間計算において，仮に計算機故障が発生しても，その被害を最小限に抑えられること．

3. InfiniBand

InfiniBand は，InfiniBand Trade Association で仕様が策定されている高性能汎用インタコネクトである．仕様で定義されているネットワークバンド幅性能は，2 Gbps（1×），8 Gbps（4×），24 Gbps（12×）であり，現在 8 Gbps（4×）の仕様の製品が入手可能である．InfiniBand の仕様は，サーバ間通信だけでなく I/O 通信をもターゲットとして設計されており，Windows やベンダ製 Unix システム上で最適な API を実現するために API 自体は定義されておらず，Verb と

呼ばれる動作仕様で記述されている．

InfiniBand では，コネクションベース通信，データグラム通信のほか，Ethernet や IPv6 のパケットを encapsulate してそのまま送受するための RAW レベル通信をサポートしている．さらに，コネクションベース通信，データグラム通信には Reliable なプロトコルと Reliable ではないプロトコルがある．それぞれのプロトコルには，メッセージ送受信，Remote Direct Memory Access（RDMA）Write/Read，そして高機能な Atomic 通信をサポートしている．**表 1** に，InfiniBand の通信がサポートしている機能をまとめる．表中，Send はメッセージ通信，Write は RDMA Write，Read は RDMA Read，Atomic は Atomic オペレーションを示している．

InfiniBand において通信は，QueuePair（QP）と呼ばれる送信キューと受信キューが対になった通信コンテキストを使う．QP を利用した通信は，CreateQP オペレーションで QP を生成し，送信用 Work Queue（WQ）と呼ばれるキューに Work Request（処理要求）を書き込み InfiniBand のハードウェアに通知することで実行される．受信バッファも同様に受信用 WQ により供給する．通信の終了時には，DestroyQP オペレーションにより QP を破壊する．また，送受信オペレーションの完了通知のため，Completion Queue（CQ）と呼ばれるキューを利用する．これらの通信において，送信用 WQ と受信用 WQ をユーザプロセス空間にマップして，Doorbell [7] という仕組みを使ってユーザレベルで発行できるように実現されている．

InfiniBand がサポートする通信の中で，クラスタ通信で使える Reliable 通信で RDMA 通信をサポートしているものに，Reliable Connection（RC）と Reliable Datagram（RD）通信があり，以下の特徴がある．

RC 通信 コネクション型の通信である．相手のノード 1 台ごとに個別の QP を割り当てる必要がある．送信用 WQ 内の一連の Work Request（WR）の列は，順番に実行されるが，先行する WR の送信完了（ACK）の到着を待たずに，後続の WR 実行を開始できる．

表 1 InfiniBand がサポートする通信
Table 1 Transport service types and operations of InfiniBand.

	Send	Write	Read	Atomic
Reliable Connection	x	x	x	x
Reliable Datagram	x	x	x	x
Unreliable Connection	x	x		
Unreliable Datagram	x			
RAW	x			

RD 通信 データグラム型の通信である．1 つの QP で複数のノードと通信が可能である．送信 WQ 内の一連の送信コマンドは，順番に実行され，先行する WR の ACK が到着してはじめて後続の WR の実行を開始できる．

4. PM/InfiniBand-FJ の課題

InfiniBand を用い 2 章で述べた要件を満たす通信機構 PM/InfiniBand-FJ の実現すべき課題として次の 3 つがある．

高い通信性能： 10 Gbps クラスの高性能ネットワーク上で，高バンド幅，かつ，低遅延通信を実現する通信方式を開発する．

高いスケーラビリティ： メモリ使用量を最小限に抑える実現方式を採用する必要がある．

長時間計算への耐故障性機能： 耐故障性実現のためのチェックポイントリスタート機構を開発する．

5. 設　　計

5.1 設計方針

PM/InfiniBand-FJ の設計方針として，課題を解決するために InfiniBand の仕様拡張や変更が必要であれば実施することとする．しかし，その変更はオリジナルの変更による影響を最小限にすることとする．

PM/InfiniBand-FJ では，InfiniBand 4× 仕様の富士通製 Host Channel Adapter（IB HCA）を用いる．IB HCA は，I/O バスとして PCI-X バス 133 MHz に対応し，送信用（SQ），受信用（RQ），そして制御用（Schedler）の専用のマルチスレッドプロセッサを搭載している（**図 1**）．このプロセッサのクロックは 266 MHz である．InfiniBand 処理の一部機能をこのプロセッサ上で処理している．

また，PM/InfiniBand-FJ では，クラスタシステムソフトウェアとして SCore [6] を採用する．SCore ク

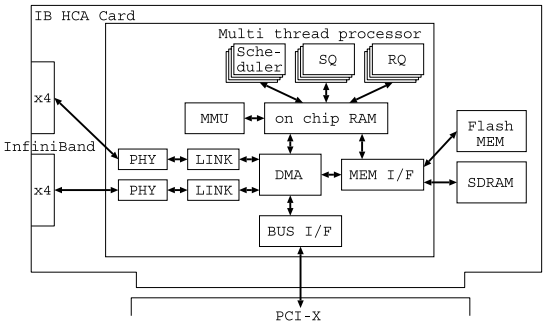

図 1 富士通製 Host Channel Adapter（IB HCA）の構成
Fig. 1 Architecture of Fujitsu's Host Channel Adapter.

ラスタシステムソフトウェア（SCore）は，新情報処理開発機構（RWCP）で開発された大規模PCクラスタ向けクラスタシステムソフトウェアである．SCoreは，高性能通信ライブラリPMv2，ユーザレベルギャングスケジューラによるマルチユーザ利用とチェックポイントリスタート機構を備えるSCore-D，その他のコンポーネントから構成される．PMv2は，Myrinet，Ethernet，Shmem，SCI，UDP/IP上に実装されている．PM/InfiniBand-FJではPM API仕様[8]をIB HCA向けに実現することにより，SCoreの持つ通信機能，ギャングスケジューラによるマルチユーザ利用，チェックポイントリスタート機能を実現する．

これ以降，IB HCAとSCoreを用い，課題を解決するPM/InfiniBand-FJの設計について述べる．

5.2 通信プロトコルの選定

本節では，PM/InfiniBand-FJで採用するInfiniBandの通信プロトコルについて整理し選定する．

まず，InfiniBandで定義されるRD通信とRC通信の2つの通信の得失を，通信性能と計算機資源消費の点から整理する．

通信性能： RD通信は1対1の連続送信においてさえも1メッセージごとにACKを待つので，Work Requestごとにラウンドトリップ時間（RTT）分のオーバヘッドが入る．これは，小さなメッセージのときには無視できないため，性能低下につながる．これに対し，RC通信ではACKを待たずに後続のWork Requestの処理を開始できるため，ACKを待つオーバヘッドが入らず，メッセージ長が短い場合でも性能低下を抑えられる．

計算機資源消費： RC通信は1対1のコネクションを張る必要があるので，ノード数が多い大規模クラスタでは，必要なQP数が増え，QPのコンテキスト情報や受信バッファなど必要なメモリ量がノード数に比例して必要である．これに対し，RD通信はデータグラム通信であるため，メモリ量の増大を抑えることが可能である．

以上の議論より，通信性能を重視しRC通信を採用する．しかし，RC通信は受信バッファがノード数に比例して必要になる．受信バッファ利用量を抑えるための方式を検討する．

5.3 高い通信性能の実現

本節では，高い通信性能（高バンド幅，低遅延）を実現するための通信機構の設計について述べる．

高バンド幅通信

IB HCAでは，4× の仕様で8 Gbps（双方向）転送をサポートしている．このような高速ネットワークを用いたデータ転送は，I/Oバス，メモリ（バス）システムに大きな負担を与える．また，実際の転送能力はチップセットの種類やメモリの種類で異なる[9]．このため，I/Oバスとメモリシステムの負担を減らすための通信方式の採用が重要になる．

メモリシステムへの負担を増大させる要因として，ホストプロセッサによるメモリコピーがある．ホストプロセッサによるメモリコピーを行わない通信方式としてリモートメモリアクセス通信（RDMA）通信がある．RDMA通信はさらに通信先のホストプロセッサを介在しないで通信できるため，ホストプロセッサ負荷を減らすことができる．PM API仕様[8]では，メッセージ通信とこのRDMA通信をサポートしている．

PM/InfiniBand-FJは，InfiniBandがサポートする通信（表1）の中でSendをメッセージ通信に，Write，ReadをRDMA通信に使い実現する．

低遅延通信

高バンド幅，低遅延通信の実現には，ホストプロセッサとネットワークインタフェース（NIC）間の情報交換を最小限に抑える必要がある[10]．

この情報交換には，ホストプロセッサによるI/Oバス経由のNICへのアクセスとPCI DMAによるものがあるが，PCI DMAは起動オーバヘッドが大きいため，1メッセージ通信あたりのPCI DMAの起動回数は最小限に抑えるべきである．特に小さいメッセージ長では，このオーバヘッドが相対的に大きくなる．

IB HCAにおける送受信用WQの実装は，WQ自体がホストメモリ上に実現され，ホスト上のバッファへのポインタとサイズをWQを構成するWQE（Work Queue Element）に書き込み，Work RequestとしてHCAに処理要求を行う．この実現手法では，1回のメッセージ転送にPCI DMAが2回必要になる．

このため，WQE自体の中に送信データ自体を埋め込んでデータを送信するData on WQE（DoWQE）を実現する．これによりメッセージ長が短い場合にはPCI DMAの回数を1回にできる．

5.4 受信バッファ利用量を抑える方式の実現

5.2節の議論より，高い通信性能を得るためRC通信を採用する．しかし，RC通信はノード数に比例した計算機資源が必要である．さらにそれぞれのRCのQP（RCQP）の受信バッファは以下の条件を満たす必要がある．

(1) RQに投入されたWork Requestごとでの受信となるため，Work Requestごとに最大転送メッセージ長（MTU）分のバッファを準備する必要がある．

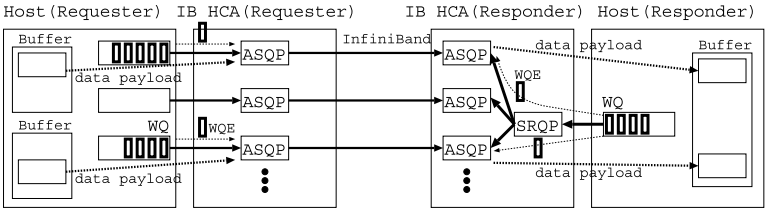

図 2 ASQP，SRQP によるメッセージ送受信の様子
Fig. 2 Message transfer using ASQPs and an SRQP.

(2) 受信バッファの枯渇は性能に大きな影響を与える．受信バッファが枯渇しないだけの十分な数の Work Request を受信用 WQ に供給する必要がある．

MTU については，1 GB/s クラスのネットワークで高い通信性能を実現するためには 8 KB 以上は必要である．また，メッセージ数としては受信バッファの枯渇を抑え，高い通信性能を実現するためには 1 つの RCQP あたり最低 128 メッセージ分は必要である．このため，1 つの RCQP だけで 1 MB の受信バッファが必要となり，1024 ノードのクラスタでは，1 GB の受信バッファが必要になる．これは，ホストメモリが数 GB のシステムでは問題である．

これを解決するために，RC 通信を用いながら受信バッファだけを共有する方式を採用する．このために，新たな QP タイプ ASQP（ASsociated QP），SRQP（Shared Receive QP）の 2 つの QP を導入する．ASQP は従来の RCQP から受信キューである RQ 機能のみを除いた QP で，SRQP は RQ 機能を提供する QP で ASQP に対して受信バッファを提供する．

図 2 に ASQP，SRQP によるメッセージ送受信の様子を示す．図 2 において，送信要求は ASQP に，ASQP 受信バッファの供給は SRQP に対して行う．メッセージが ASQP に受信されると，ASQP は SRQP から受信バッファを獲得してメッセージを受信する．受信バッファの獲得は単純に排他制御を行い SRQP に登録済みのバッファを FIFO 順に取り出すことで行われる．

以上のように ASQP と SRQP の導入により，通信性能を RC 通信と同等にしながら，ノード数に比例しない受信バッファの実現が可能となる．

5.5 チェックポイントリスタート機能（IbCPR）の実現

PM/InfiniBand-FJ のチェックポイントリスタート機能は SCore の機能を利用する．SCore のチェックポイントリスタート機能は，SCore-D グローバルオペレーティングシステムと PMv2 通信機構でユーザレベルでのチェックポイントリスタートを実現している[11]．PMv2 通信ライブラリは SCore-D からの要求で要求時点でのネットワーク状態（コンテキスト）を確定する *pmSendStable()* 機能とメモリ上に Save（Restore）する機能 *pmSaveContext()*（*pmRestoreContext()*）を用いて実装されている．この機能はギャングスケジューラの実現にも利用されている．

さて，以上の機能を InfniniBand に効率的に実装するのは次の理由から困難である．

- InfiniBand の仕様では，QP コンテキストの状態を Save-Restore する機能がない．
- *pmSendStable()* 機能は，InfiniBand の仕様に定義されている通信停止状態を用いることにより実現できる．しかし，QP 数に比例した時間がかかる．ギャングスケジューリング実行時には，デフォルトで 100 ms 単位に *pmSendStable()* が呼ばれるため，この時間を最小にする必要がある．

以上の問題を解決するため，5.4 節で導入した ASQP と SRQP に次に述べる機能を拡張する．

QP 内の状態を Save-Restore する機能の導入： QP の通信のコンテキストを *pmSaveContext()*（*pmRestoreContext()*）実行時に送受信バッファ，WQE とともに HCA 内部の QP，CQ コンテキストを Save（Restore）する機能を追加する．QP のコンテキスト Save 時には 1 度の要求で複数 QP のコンテキストの Save ができるようにする．

QP への一括状態変更の導入： 通信停止状態と復帰を複数 QP を範囲指定することにより一度の要求で QP の状態変化ができるようにする．

以上の機能により，ギャングスケジューリング時の QP の破壊と生成のオーバヘッドを削減し，効率の良いチェックポイントリスタート機能を実現する．

6. 実　　装

PM/InfiniBand-FJ は，SCore クラスタシステムソフトウェアの通信層である PMv2 の API を用いて実

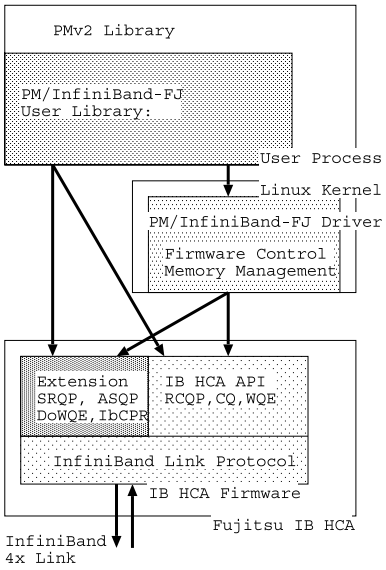

図 3 PM/InfiniBand-FJ の構成
Fig. 3 The architecture of PM/InfiniBand-FJ.

装され，オペレーティングシステムは Linux である．

PM/InfiniBand-FJ は，PMv2 の API を実装しており，ユーザライブラリ，デバイスドライバ，そして IB HCA のファームウェアで構成されている．**図 3** に PM/InfiniBand-FJ の構成を示す．

全体の役割分担は次のようになっている．

PM/InfinBand-FJ User Libary： PMv2 API と通信処理を実現，ユーザレベル通信を実現

PM/InfinBand-FJ Device Driver： QP の生成破壊などの HCA 制御，メモリ管理を実装

IB HCA Firmware： SRQP, ASQP, DoWQE, IbCPR 拡張機能を実装

以下に，拡張機能の実装についてまとめる．

ASQP，SRQP の実装： RCQP の中で，受信バッファのみ SRQP から取るように実装．バッファ量は可変であるが，現状，送信バッファは 4 MB，受信バッファは 8 MB である．

DoWQE の実装： WQE 64 バイトのうち，48 バイトを送信バッファとして利用して実現している．

チェックポイントリスタート機能（IbCPR）の実装： PMv2 の複数の API 関数で実現している．

- *pmSendStable()*, *pmControlSend()* の実装：*pmSendStable()* 時に，通信に利用している QP 全体を通信停止状態に移行させる．*pmControlSend()* の送信開始時には，QP 全体の通信を 1 度の要求で再開させる．
- *pmSaveContext()* の実装：複数の QP コンテキストを一括してメモリ上に Save する．
- *pmRestoreContext()* の実装：RCQP, SRQP 用の CreateQP を実現し，QP 生成時に QP コンテキストを Restore する．

以上の実装により，PM/InfiniBand-FJ は実現されており，NAS 並列ベンチマーク[12] などで，SCore のギャングスケジューラがマルチ並列プロセスで正常に実行できることを確認している．

図 4 にチェックポイントリスタート時の出力を示す．チェックポイント実行後正しく再開されていることが分かる．

```
SCore-D 5.6.0 connected (jid=21,reconnect=33054).
<0:0> SCORE: 16 nodes (8x2) ready.


 NAS Parallel Benchmarks 2.2 -- LU Benchmark

 Size: 102x102x102
 Iterations: 250
 Number of processes:    16

 Time step    1
 Time step   20
 Time step   40

SCORE: Checkpointing ... done.                    ***注 チェックポイント実行
 Time step   60
 Time step   80
FEP:WARNING SCore-D unexpectedly terminated.      ***注 プログラム中断
FEP: [29/Jan/2004 14:52:44] Waiting for SCore-D restarted ...
FEP: [29/Jan/2004 14:53:04] SCore-D restarted.    ***注 リスタート処理開始
SCore-D 5.7.0 connected (jid=21,reconnect=33054).
SCORE: Execution restarted from checkpoint.       ***注 プログラム開始
 Time step   60                                   ***注 同じ Time step から再開
 Time step   80
```

図 4 チェックポイントリスタートの様子
Fig. 4 The console outputs of running checkpoint-restart function.

7. 評　　価

本章では，PM/InfiniBand-FJ について，その通信性能とアプリケーション性能について評価する．**表 2** に評価環境を示す．通信性能はチップセットの影響を確認するため，ServerWorks GC LE（以下 GC LE）と Intel E7501（以下 E7501）の 2 種類のチップセット上で測定する．

アプリケーションの評価として，NAS 並列ベンチマーク[12] を Myrinet XP，Gigabit Ethernet と比較する．また，姫野ベンチマーク[13] の結果を従来型スーパコンピュータの結果と比較する．

Myrinet XP 搭載のクラスタは 128 ノード，InfiniBand 搭載のクラスタは 256 ノードである．Gigabit Ethernet のクラスタは Myrinet XP 搭載のクラスタを用いたが，すべてのネットワークをフルバイセクションの環境で測定するため，16 ノード以上はスイッチ間リンクが細いため 16 ノードまでの測定とした．スイッチ結合は，MyrinetXP と Ethernet は 1 台のスイッチでの結合，InfiniBand は 24 台のスイッチの結合である．

表 2 評価環境
Table 2 Measurement environments.

	GC LE 自作クラスタ環境
ノード計算機	Self made DUAL Xeon 2.8 GHz 搭載 (ServerWorks GC LE chipset, 1 GB DDR SDRAM, 64 bit 133 MHz PCI-X Bus)
Ethernet (1 Gbps)	Intel 社 E1000 (Gigabit Ethernet) NIC DELL PowerConnect 5224 Switch
InfiniBand (8 Gbps)	IB HCA (PCI-X 133 MHz) RedSwitch 社 8 ポート Switch
ホスト OS	Redhat 8.0 Linux (2.4.21 kernel) SCore5.6

	富士通 PRIMERGY RX200 クラスタ環境
ノード計算機	RX200 DUAL Xeon 3.06 GHz 搭載 (Intel E7501 chipset, 2 GB DDR SDRAM, 64 bit 133 MHz PCI-X Bus)
Ethernet (1 Gbps)	Intel 社 E1000 (Gigabit Ethernet) NIC 富士通 SH2422 Switch
Myrinet (2 Gbps)	Myricom 社 M3F-PCIXD PCI-X 133 MHz Myricom 社 M3F-E128
InfiniBand (8 Gbps)	IB HCA (PCI-X 133 MHz) InfiniCom 社 32 ポート Switch
ホスト OS	Redhat 8.0 Linux (2.4.21 kernel), SCore5.6

表 3 PM レベルのラウンドトリップ時間
Table 3 PM level communication round trip time.

	RTT	Ratio
PM/InfiniBand-FJ*	15.6 μs	100%
PM/InfiniBand-FJ**	16.8 μs	108%
PM/MyrinetXP**	8.3 μs	53%
PM/Ethernet**	29.6 μs	190%

注）Switch の遅延を含む，*：GC-LE，**：E7501

7.1 通信性能

表 3 に PM/InfiniBand-FJ，PM/MyrinetXP，PM/Ethernet の pmtest プログラムによるラウンドトリップ時間（RTT）と GC LE 上の PM/InfiniBand-FJ の結果との比を示す．

表 3 の結果より，GC LE 上と E7501 との差は 8%で，チップセットにより差があることが分かる．E7501 上の PM/MyrinetXP では，GC LE 上の PM/InfiniBand-FJ に比べ 47% RTT が小さい．また，PM/Ethernet の結果は，PM/InfiniBand-FJ の結果に比べ RTT が 90%大きい．同じユーザレベル通信を採用している PM/MyrinetXP に比べ，PM/InfiniBand-FJ の RTT 時間が大きいのは，HCA 上での InfiniBand プロトコル処理のオーバヘッドが大きいからである．

表 4 に，DoWQE 機構の効果を示す．結果より DoWQE の効果は最大 12.5%の効果があった．PCI DMA 回数を減らす効果は大きいといえる．

図 5 に pmtest のメッセージバンド幅性能，**図 6** に同様に pmtest の RDMA write バンド幅性能を示す．

表 4 DoWQE 機構の効果
Table 4 Effects of DoWQE extension.

	RTT	w/o DoWQE	Effects
GC LE	15.6 μs	17.9 μs	12.5%
E7501	16.8 μs	18.9 μs	11.1%

* Switch の遅延を含む

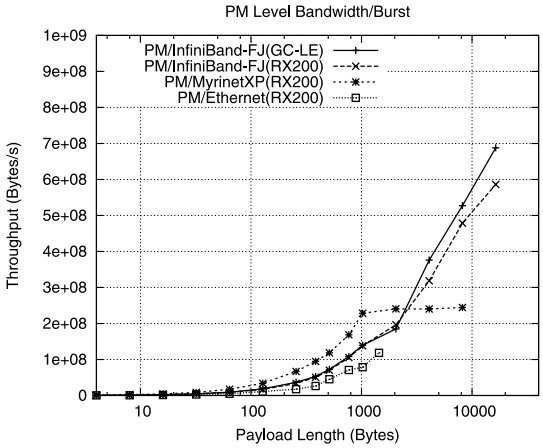

図 5 PM レベルの転送バンド幅性能（メッセージ通信）
Fig. 5 PM level communication bandwidth (Message).

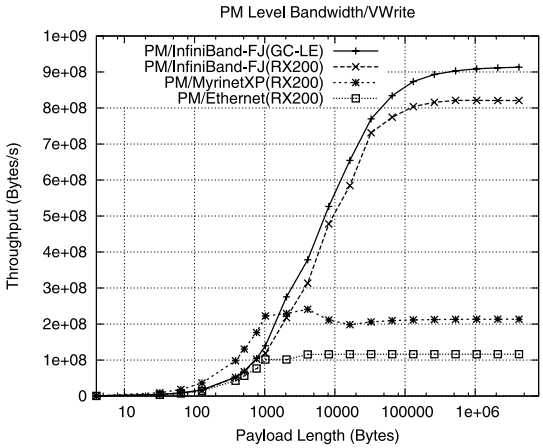

図 6 PM レベルの転送バンド幅性能（RDMA 通信）
Fig. 6 PM level communication bandwidth (RDMA).

図 5，図 6 を元に，**表 5** にメッセージと RDMA の通信バンド幅と，MyrinetXP との比をまとめる．表 5 より，PM/InfiniBand-FJ の GC LE と E7501 の RDMA の差は 92.3 MB/s であり，チップセットの選択により通信性能に大きな差があることが分かる．

また，PM/InfiniBand-FJ（GC LE）の結果は，MyrinetXP 上での RDMA 通信バンド幅より 3.8 倍高い性能を実現している．なお，PM/Myrinet に比べ，PM/InfiniBand-FJ の通信バンド幅性能の立ち上

表 5 通信バンド幅性能比較
Table 5 Performance comparison of PM communication bandwidth.

	メッセージ	RDMA
PM/InfiniBand-FJ*	688.0 (2.82)	913.2 (3.80)
PM/InfiniBand-FJ**	586.1 (2.40)	820.9 (3.40)
PM/MyrinetXP**	244.4 (1.00)	241.0 (1.00)
PM/Ethernet**	118.4 (0.48)	116.4 (0.48)

注）単位 MB/s（Ratio），*：GC-LE，**：E7501

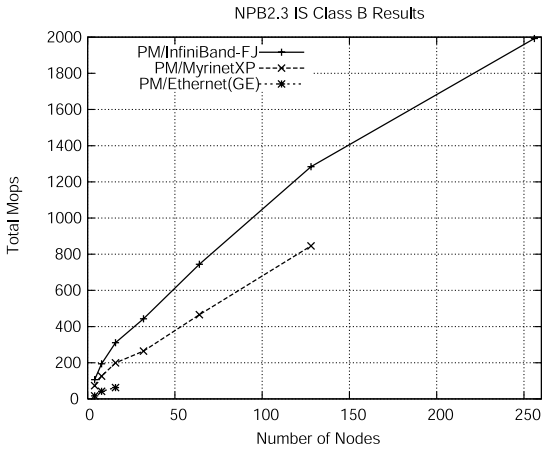

図 7 IS クラス B の実行性能
Fig. 7 Results of IS class B.

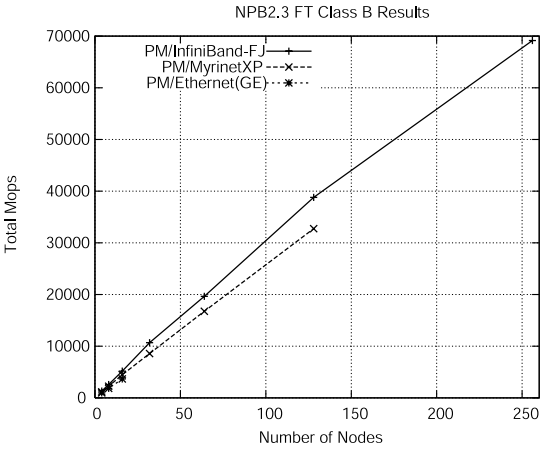

図 8 FT クラス B の実行性能
Fig. 8 Results of FT class B.

表 6 姫野ベンチマーク性能比較
Table 6 Performance comparison of the Himeno benchmark.

	PRIMERGY RX200 クラスタ	Alphaserver SC45
128 CPUs	123.0 Gflops	-
256 CPUs	242.8 Gflops	102.7 Gflops

がりが遅いのは，通信処理時間が大きいためである．特に現状の IB HCA のファームウェアは，メッセージ送信時の ACK をメッセージごとに要求するのに比べて PM/Myrinet は複数メッセージの ACK を 1 つにまとめているため，ACK 送受信処理が軽くなりメッセージ長が小さなときのバンド幅が高くなっている．

7.2 NAS 並列ベンチマーク性能

図 7，**図 8** に NAS 並列ベンチマーク CLASS B IS，FT の結果を示す．IS は整数ソートプログラム，FT はフーリエ変換を行うプログラムである．コンパイラは Intel コンパイラを利用し，1 計算ノードあたり 1 CPU を利用した結果である．

図 7 の IS の結果より，PM/InfiniBand-FJ は，256 ノードで 2.0 Gflops と高い台数効果を実現していることが分かる．また，128 ノードで 1.28 Gflops と MyrinetXP の 0.85 Gflops に比べ 52%高い性能を実現している．16 ノードでは，311 Mflops と Ethernet の 63 Mflops に比べ 5 倍の性能となっている．IS は all-to-all 通信を短時間で行うため通信バンド幅性能の差が出ている．

図 8 の FT の結果より，PM/InfiniBand-FJ は，256 ノードで 69.1 Gflops と高い台数効果を実現していることが分かる．128 ノードの結果は 38.8 Gflops と MyrinetXP の 32.7 Gflops に比べ 18%高い性能を実現，16 ノードでは，5.2 Gflops と Ethernet の 3.6 Gflops に比べ 41%高い性能を実現している．FT は IS と比較して単位時間に必要なバンド幅性能が小さいので，性能差が小さくなっている．

7.3 姫野ベンチマークによる従来型スーパコンピュータとの比較

表 6 に姫野ベンチマークの結果を示す．比較として従来型スーパコンピュータで公開されている中での最速値を実現している HP Alphaserver SC45（Alpha 21264 1 GHz，256 CPU，QsNet）の結果を載せる[14]．PRIMERGY RX200 クラスタの結果は，富士通コンパイラを利用しており，1 計算ノードあたり 1 CPU を利用したものである．

表 6 より，PM/InfiniBand-FJ を用いた PRIMERGY RX200 クラスタは，256 ノード時に 242.8 GFlops と高い性能を実現していることが分かる．これは同じ CPU 数の HP Alphaserver SC45 の 102.68 GFlops に比べ，同一 CPU 数で 2.4 倍高性能であり，従来型スーパコンピュータを凌駕する性能を実現している．

8. 関連研究

RDMA 通信を実現した高性能通信機構は，Myrinet

を[2] 用いた GM[15]，BIP[16]，VMMC-2[17]，PM[18] があるが，この中でチェックポイントリスタート機能を備えているのは PM[18] だけである．

PM/Ehernet-kRMA[9] では，カーネルレベルで Remote Memory Access を実現しており，Gigabit Ethernet 4 系統で 485 MB/s の通信バンド幅を実現している．しかし，PM/Ehernet-kRMA ではカーネルレベルでホストプロセッサのコピーを行うため，10 Gbps クラスのネットワークではオーバヘッドが大きい．

InfiniBand を用いた MPI の実装として文献 19) がある．Mellanox 社の InfiniBand HCA 上で RDMA を用いて高速な MPI を実現することに主眼を置いており，MPI レベルで 871 MB/s の通信性能を実現している．また，InfiniBand 上での低レベル通信ライブラリの仕様として DAPL[20] がある．しかしこれらは，オリジナルの InfiniBand の仕様ベースであるため，大規模 PC クラスタ上では受信バッファの問題があるうえ，チェックポイントリスタート機能を持たない．

9. 結　　論

本論文では，10 Gbps クラスのネットワークである InfiniBand を用いた大規模 PC クラスタ向けの高性能通信機構 PM/InfiniBand-FJ の設計について述べた．

PM/InfiniBand-FJ は，大規模クラスタで高い通信性能を実現し，計算機センタでの利用に耐えうる PC クラスタを実現するため，オリジナルの InfiniBand の RCQP 仕様に対して次の拡張を行っている．

ASQP，SRQP の導入： ノード数に比例した受信バッファが必要である RCQP 通信の問題を解決し，ノード数に依存しない受信バッファ量を実現．

Data on WQE（DoWQE）の導入： WQE にデータを格納することで PCI DMA の回数を削減し通信遅延を最大 12.5%短縮．

チェックポイントリスタート機能（IbCPR）の実現： QP コンテキストの Save/Resore 機能を実現．

これらの拡張の結果，計算機資源利用を最小限に抑えながら高い通信性能を実現し，さらに，長時間計算での計算ノード故障に対する耐故障性を実現している．

PM/InfiniBand-FJ を SCore クラスタシステムソフトウェア上に実装し，Xeon 2.8 GHz プロセッサを搭載の PC クラスタで評価した結果，913.2 MB/s の通信バンド幅性能と，15.6 μs のラウンドトリップ時間を実現している．また，Xeon 3.06 GHz を搭載した 256 ノードの富士通 PRIMERGY RX200 PC クラスタで姫野ベンチマークを評価した結果，242.8 GFlops と同じ CPU 数の HP Alphaserver SC45 の 102.68 GFlops に比べ，同一 CPU 数で 2.4 倍高い性能を実現した．

今後は PM/InfiniBand-FJ の通信遅延を削減し，より高性能な通信機構を実現するとともに，ユーザの利用実績を積みより高い信頼性を実現していく予定である．

参 考 文 献

1) Super Computer TOP500. http://www.top500.org/
2) Boden, N.J., Cohen, D., Felderman, R.E., Kulawik, A.E., Seitz, C.L., Seizovic, J.N. and Su, W.-K.: Myrinet — A gigabit-per-second local-area network, *IEEE MICRO*, Vol.15, No.1, pp.29–36 (1995).
3) ELAN (QSNET). http://www.quadrics.com/
4) InfiniBand Trade Association. http://www.infinibandta.org/
5) 10 Gigabit Ethernet Alliance. http://www.10gea.org/
6) SCore Cluster System Software. http://www.pccluster.org/
7) Compaq Computer Corporation, Intel, and Microsoft: Virtual Interface Architecture Specification, Version 1.0 (Dec. 1997). http://www.viarch.org/
8) PM 2.1 API. http://www.pccluster.org/score/dist/score/html/en/man/man3/PM.html
9) Sumimoto, S. and Kumon, K.: PM/Ehernet-kRMA: A High Performance Remote Memory Access Facility Using Multiple Gigabit Ethernet Cards, *3rd International Symposium on Cluster Computing and the Grid*, pp.326–334. IEEE (May 2003).
10) Sumimoto, S., Tezuka, H., Hori, A., Harada, H., Takahashi, T. and Ishikawa, Y.: The Design and Evaluation of High Performance Communication using a Gigabit Ethernet, *International Conference on Supercomputing '99*, pp.243–250, ACM SIGARCH (June 1999).
11) 西岡利博，堀　敦史，手塚宏史，石川　裕：クラスタにおけるコンシステントチェックポイントの実現，並列処理シンポジウム JSPP'99，pp.229–236 (June 1999).
12) The NAS Parallel Benchmarks (NPB). http://www.nas.nasa.gov/Software/NPB/
13) Himeno Benchmark. http://w3cic.riken.go.jp/HPC/HimenoBMT/index.html.
14) Himeno Benchmarck Results. http://w3cic.riken.go.jp/HPC/HimenoBMT/himenoDB1.pdf
15) The GM API.

http://www.myri.com/GM/doc/gm_toc.html
16) Prylli, L. and Tourancheau, B.: BIP: A new protocol designed for high performance, *PC-NOW Workshop, held in parallel with IPPS/SPDP98*, Orlando, USA (March 30–April 3 1998).
17) Dubnicki, C., Bilas, A., Chen, Y., Damianakis, S. and Li, K.: VMMC-2: Efficient Support for Reliable, Connection-Oriented Communication, *Hot Interconnect'97* (Aug. 1997).
18) Tezuka, H., Hori, A., Ishikawa, Y. and Sato, M.: PM: An Operating System Coordinated High Performance Communication Library, *High-Performance Computing and Networking*, Sloot, P. and Hertzberger, B. (Eds.), Vol.1225 of Lecture Notes in Computer Science, pp.708–717, Springer-Verlag (Apr. 1997).
19) Kini, S.P., Wyckoff, P., Liu, J., Wu, J. and Panda, D.K.: High Performance RDMA-Based MPI Implementation over InfiniBand, *Proc. 17th Annual ACM International Conference on Supercomputing*, San Francisco Bay Area, IEEE (2003).
20) DAT Collaborative. http://www.datcollaborative.org/

(平成 16 年 1 月 31 日受付)
(平成 16 年 6 月 17 日採録)

住元　真司（正会員）

1986 年同志社大学工学部電子工学科卒業．同年（株）富士通入社．（株）富士通研究所にて並列オペレーティングシステム，並列分散システムソフトウェアの研究開発に従事．1997 年より新情報処理開発機構に出向．コモディティネットワークを用いた高速通信機構の研究開発に従事．2002 年より（株）富士通研究所にて高速通信機構の研究開発に従事，並列分散システムのアーキテクチャ，システムソフトウェア等に興味を持つ．平成 12 年度情報処理学会論文賞受賞，工学博士（慶應義塾大学大学院理工学研究科）．

成瀬　彰（正会員）

1996 年名古屋大学大学院工学研究科修了（情報工学専攻）．同年富士通（株）入社．（株）富士通研究所にて IA サーバに関わる研究・開発に従事．並列処理，計算機アーキテクチャに興味を持つ．

久門　耕一（正会員）

1979 年東京大学工学部電気工学科卒業．1981 年同大学大学院電子工学専門課程修士課程修了．1984 年同課程博士課程中退．同年（株）富士通研究所入社．現在，同社 IT コア研究所に所属．CPU，メモリ，並列計算機アーキテクチャに関する研究に従事．GCC，Linux カーネル等の改良にも興味を持つ．日本ソフトウェア科学会会員．

細江　広治

1994 年名古屋大学工学部卒業．1996 年同大学大学院電子情報工学研究科修士課程修了，同年（株）富士通入社．並列計算機システムの開発に従事．現在，富士通（株）サーバシステム事業本部に勤務．サーバハードウェアの開発に従事．

清水　俊幸（正会員）

1986 年東京工業大学工学部卒業．1988 年同大学大学院理工学研究科修士課程修了．同年（株）富士通研究所入社．並列計算機のアーキテクチャの研究に従事．現在，富士通（株）サーバシステム事業本部に勤務．サーバハードウェアの開発に従事．信号処理学会会員．