Vol.2010-ARC-190 No.16
2010/8/4

InfiniBand と Ethernet の混在環境でのクラスタノード間通信に関する提案

中浜徹也[†1]　西川由理[†1]
吉見真聡[†2]　天野英晴[†1]

近年の高性能な PC クラスタでは、GPU と汎用 CPU などの異種混合型が主流となりつつある。本研究報告では、まず、現在我々が構築中の異種混合型の Cell PC クラスタ構想を述べる。本 PC クラスタは、312 個の Cell/B.E. を搭載し、そのインターコネクトには Ethernet, InfiniBand, Fibre Channel の 3 つが採用されている。次に、Cell/B.E. 間を Ethernet および InfiniBand で接続した場合の通信性能の予備評価を行う。そして、High Performance Linpack (HPL) を用いた 4 ノードまでの性能測定を行い、本クラスタが構築できた場合の性能を見積る。

The proposal of inter-node communication in heterogeneous interconnect cluster using InfiniBand and Ethernet

Tetsuya Nakahama,[†1] Yuri Nishikawa,[†1]
Masato Yoshimi[†2] and Hideharu Amano[†1]

In high performance PC clusters, heterogeneous type of cluster is increasing in recent years. Our research group is designing a heterogeneous PC-cluster which equips X86 and Cell/B.E. as computing nodes, and Ethernet, Optical fibers, InfiniBand and Fibre Channel as interconnects. This report evaluates basic performance such as network delay and bandwidth using an environment with multiple Cell/B.E.s connected by InfiniBand or Ethernet. In addition, operation performances and their tuning technique is discussed through preliminary evaluation of High Performance Linpack(HPL).

†1 慶應義塾大学 理工学部
Faculty of Science and Technology, Keio University
†2 同志社大学 理工学部 インテリジェント情報工学科
Department of Intelligent Information Engineering and Science, Doshisha University

1. はじめに

近年，ハイパフォーマンスコンピューティング分野では，複数の種類のプロセッサやネットワークが混在した異種混合型のスーパコンピュータが主流になりつつある．例えば，プロセッサが異種混合型であるスパコンの例としては Intel Westmere-EP CPU に加え，NVIDIA Fermi アーキテクチャの GPU である Tesla M2050 を搭載した TSUBAME2.0 や Cell B.E を用いた Roadrunner が挙げられる[1)2)]．これらは，一部の重い計算処理を汎用プロセッサではなく GPU や Cell B.E. に代表されるようなアクセラレータにより実行することで高性能を得ることが可能である．

また，インターコネクトに関しては，帯域が 10Gbps 以下の場合，1000base-T GbE，あるいは InfiniBand 等の電気パケット交換方式が一般的であり，ファイルシステム (ストレージ) は GbE で接続し，MPI 通信では InfiniBand を用いる，などの混合型が多い．その他の例として，IBM Blue Gene では集合通信用のツリートポロジ，隣接通信用の 3 次元トーラスなどの通信用途に応じて異なる形状のネットワークが用いられている．

しかし，近い将来，大規模なシステムを高バイセクションバンド幅を持つフルクロスバ，Fat Tree 等のトポロジでフラットに接続することは，コスト・性能面から難しくなると考えられる．そこで，GbE および光ネットワークを混在したネットワーク環境[3)] や，InfiniBand ネットワークの一部を光ネットワークに置き換えて補助的に利用する手法が研究されている[4)]．また，ネットワーク速度が 100Gbps を超える exascale の次世代スパコンでは，光・InfiniBand・カッパーケーブル等のインターコネクトの混在が検討事項として挙げられている[5)]．電気ケーブルは安価であるが，100Gbps を越えると，30cm 程度で性能が大きく減衰することが知られている．よって，物理的に極めて近距離はカッパーケーブルを用い，それ以外を光ケーブルに置き換えることが提案されている．

現在，我々もプロセッサとして 33 台の X86 ブレード，156 個の Cell/B.E. ブレードを搭載し，そのインターコネクトには Ethernet, InfiniBand, Fibre Channel の 3 つを用いた異種混合型 PC クラスタであるセルクラスタを構築中である．

インターコネクトには，GbE に加え，10GbE 光インターコネクト，InfiniBand，Fiber Channel が混在する．また各計算ノードには，Cell/B.E. プロセッサを 2 個搭載した QS21 ブレードを用いる．クラスタは，今年度の 4 月に機材が納入され，2010 年度 11 月の稼働

を目標に，一部の機材を用いた試験的な運用を開始している．

本研究報告では部分的に稼働しているセルクラスタを用いて，各種の予備評価を行う．

まず，Cell/B.E. 間を Ethernet および InfiniBand で接続した場合の通信性能の予備評価を行う．次に，High Performance Linpack (HPL) を用いた 4 ノードまでの性能測定を行い，本クラスタが構築できた場合の性能を見積る．

以後の構成は以下の通りである．2 章では，Ethernet，InfiniBand, Fiber Channel について述べる．次に 3 章で Cell B.E. および QS21 Cell Blade に関して簡潔に述べる．4 章で構築中であるセルクラスタの構成に関して述べる．5 章で評価結果を示し，6 章で本研究報告をまとめ，今後の展望についても述べる．

2. 混在ネットワーク環境

本章では，混在ネットワーク環境にて主に使用されているインターコネクトである，Ethernet, InfiniBand, Fibre Channel について述べる．

2.1 Ethernet

Ethernet の特徴として最も重要な点は，スパコンの分野に限らずデファクトスタンダードのため安価であることが挙げられる．また，ノウハウも充実しており，管理運用も比較的容易である．近年では，10G Ethernet の普及でスパコンのインターコネクトとしても注目を集めている．これらの長所から，2010 年 6 月の Top500 では 48%以上と，約半数にものぼる数のスパコンが Ethernet を採用している[6]．

2.2 InfiniBand

InfiniBand の長所として，まず，帯域が広いことが挙げられる．InfiniBand の帯域はレーンとデータレートから決まり，組み合わせ次第では最大で 100Gbps 以上の帯域を持つことが可能である．さらに，レイテンシに関しても一般的に Ethernet と比較して性能がよいとされている．その理由は，処理をハードウェアにオフロードしている点，カーネル内へのバッファリングが存在しない点，TCP を使用していない点，RDMA をサポートしている点などが挙げられる．

2.3 Fibre Channel

物理的なケーブルとしては，Gigabit Ethernet 用の LAN ケーブルや 10G Ethernet の光ケーブルを使用する．ただし，Ethernet と異なり，Ethernet では OSI 参照モデルの第 2 層までをハードウェアで処理しているのに対し，FC アダプタや FC スイッチのハードウェアで OSI 参照モデルの第 5 層まで処理している点である．

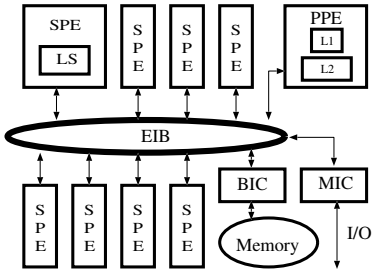

図 1 Cell/B.E. の構成
Fig. 1 Structure of Cell/B.E.

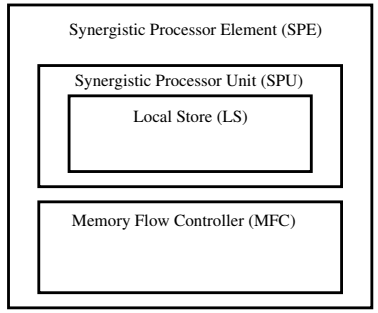

図 2 SPE の構成
Fig. 2 Structure of SPE

また，特徴としてブロックアクセスに特化しており，サーバ-ストレージ間の接続に用いられる．

3. Cell Broadband Engine

Cell Broadband Engine(Cell/B.E.) は，IBM Power Architecture ベースの汎用コアである PPE (PowerPC Processor Element) 1 基，及びマルチメディア演算に特化した SPE (Synergistic Processor Element) 8 基からなるヘテロジニアス (Heterogeneous:非対称) マルチコアプロセッサである．各プロセッサは Element Interconnect Bus (EIB) によって接続される．本章では，Cell/B.E. のアーキテクチャ及び性能を引き出すためのプログラミングの手法について述べる．

3.1 Cell/B.E. の構成

Cell/B.E. の構成を図 1 に，本論文で提案する機構によって実際に操作を行う SPE の構成を図 2 示す． Cell/B.E. は 1 基で 200GFLOPS を超える高い単精度浮動小数点演算能力をもつ．しかしながら，ヘテロジニアスマルチコアプロセッサであるという特徴から，従来とは異なるプログラミング手法が必要となる．

PPE は，メインメモリや外部デバイス等の制御をおこなう汎用プロセッサである．PPE は命令 · データ用にそれぞれ 32KB の一次キャッシュ，512KB の二次キャッシュを備えている．また PPU は PowerPC アーキテクチャをベースとした命令セットを持ち，128 ビット SIMD ユニットである VMX を搭載している．但し，PPE における VMX 命令は倍精度浮動小数点演算には対応していない．PPE は，主に計算のみを行う SPE と比べ汎用性が高

く，SPE への命令実行，及び OS の管理に用いることが一般的であるとされている．

SPE は Synergistic Processor Unit (SPU)，Local Store (LS)，Memory Flow Controller (MFC) からなる 128 ビット SIMD 型のプロセッサである．SPE は libspe2 ライブラリを用いて C などの高級言語によって PPE から制御することが可能である．SPU は 128 ビット長のレジスタを搭載した SIMD 命令を持つ演算機である．1 サイクルあたり 4 並列で演算を行うことが可能であるが，SPE における単精度浮動小数点演算において丸め誤差は切り捨てられ，IEEE754 に準拠しない．SPE は専用のメモリ (LS) を搭載し，容量は 256KB である．LS は 128bit/cycle アクセスが可能である．

3.2 IBM BladeCenter QS21 の構成

IBM BladeCenter QS21 は Cell/B.E. を 2 基搭載したブレードサーバである．その構成図を図 3 に示す．

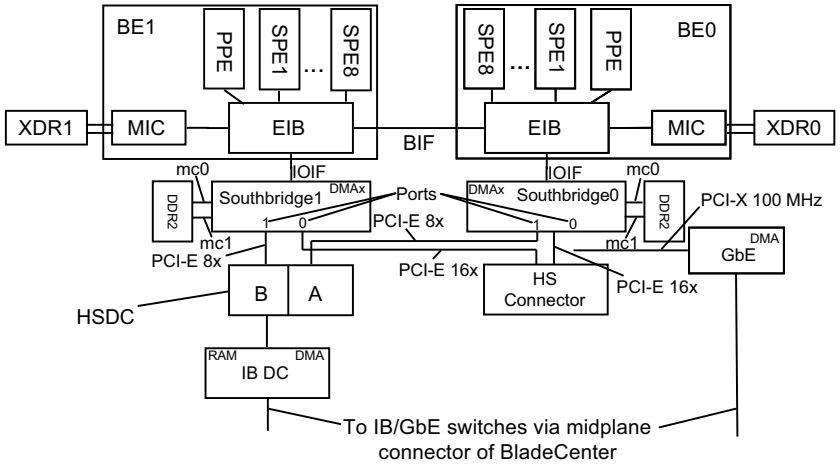

図 3 QS21 の内部アーキテクチャ

QS21 の構成は主に次の二つの部分からなる．

- 2 つの Cell B.E. と XDR DRAM からなるプロセッサ部分
- DDR2 メモリや InfiniBand 用コントローラ，Gbit Ethernet コントローラなどからなる I/O サブシステム

上記のうち，プロセッサ部分では SPE からの DMAget 実行時にスヌープキャッシュプロトコルのために Cell B.E. 間を結ぶ BIF がボトルネックになりうる．しかし，実際のアプリケーションにてデータローカリティが存在すれば深刻な問題にはならない．

一方，I/O サブシステム部分では，PCIe x16 でサウスブリッジに接続されている InfiniBand 用 DMA コントローラがある．なお，この内部コントローラを用いて InfiniBand 通信を行うには，ブレード毎に拡張カードが必要だが，本研究で用いる Cell クラスタは拡張カードを使用しない．かわりに，一部のブレードサーバに PCIe 拡張スロットを持つ PEU3 を接続し，InfiniBand PCIe アダプタを接続し，シャーシ間の InfiniBand 接続を行う．

以上のような構成を持つ QS21 ブレードサーバ上で Single Precision Linpack を実行すると，一つのブレードにおいて単精度演算では，理論性能が 409Gflops であるのに対し 342Gflops 程度の性能が確認されている[7]．

4. セルクラスタ

この章では，現在構築中の Cell blade を用いた異種混在クラスタに関して述べる．

4.1 セルクラスタの構築

予定しているセルクラスタの構成を図 4 に示す．

このシステムの特徴として，一部の QS21 Cell blade のみに InfiniBand が接続されており，InfinIBand と GbE が混在した環境で MPI 通信が行われることが挙げられる．InfiniBand の HCA (Host Channel Adapter) を使用するために必要な QS21 Cell blade の PCIe 拡張スロット PEU3 は，Cell blade 1 基と同数のスロットを占有してしまう．そのため，PEU3 は部分的に採用するに留め，シャーシごとに十分なブレード数を確保することとした．こうして，図 4 のように，シャーシ中に 12 台あるブレードの中で，PCIe 拡張スロットを取り付けた 2 台は主に通信用のノードとして用いる構成とした．

さらに，表 1 に主な機材と台数を示す．

今回は InfiniBand は SDRx4 レーン (10Gbps) を用いる．また，GbE と InfiniBand のソフトウェアの互換性を重視し, InfiniBand では IPoIB(IP over InfiniBand) を用いる．なお，Fibre Channel は，今後使用する予定である．

4.2 Cell blade クラスタの実効演算性能試算

ここで，構築を予定しているクラスタで HPL を実行する場合について検討する．

QS21 では前節で述べたように倍精度小数点演算性能が低く，LU 分解中の前進・後退代入部分など，HPL において計算負荷の高い箇所は単精度で計算し，高精度が求められる一

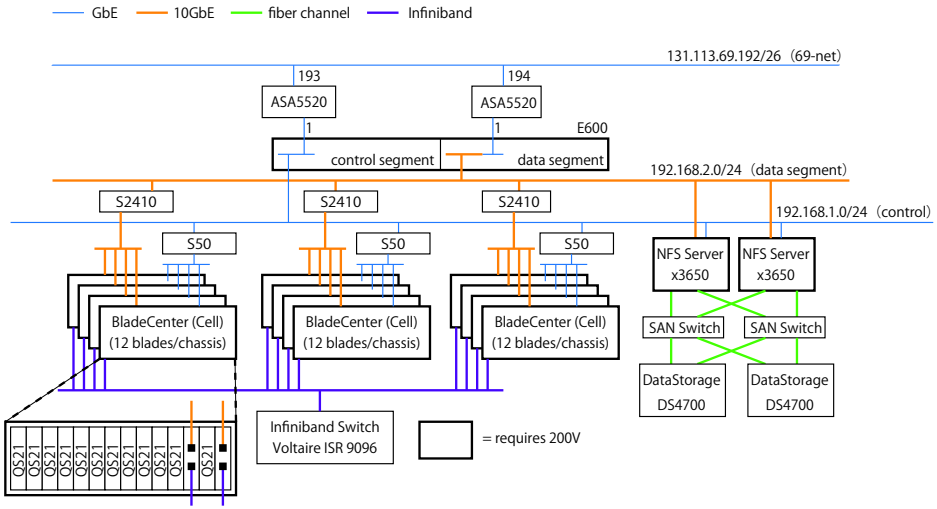

図 4 Cell B.E. クラスタ構成図 (2010 年 11 月完成予定)

表 1 評価環境

IBM Bladecenter H chasiss	15 台
QS21 Cell blade	156 台
HS20 x86 blade	29 台
HS20 x86 blade	4 台
HS Expansion Card (PEU3)	42 台
Voltaire InfiniBand switch	1 台
10G Ethernet switch	3 台
Gbit Ethernet switch	計 11 台

部の処理のみ倍精度で計算する混合精度演算方式 (Mixed precision) を用いる．文献 8) によると，Mixed precision 方式の Linpack では Cell/B.E. の単精度演算性能のうち 74%程度が引き出されることが報告されている．

前節より，ブレード 1 基で Single Precision Linpack を実行した際の性能を 342Gflops とすると，1 シャーシあたりブレード 12 枚なので，シャーシあたりの単精度演算性能は 4.104Tflops となる．ここで，Mixed precision を用いる場合，シャーシあたりの演算性能は最大 3.037Tflops となる．

本システムでこのうちの 80%の性能が引き出せると仮定すると，シャーシあたりの実効

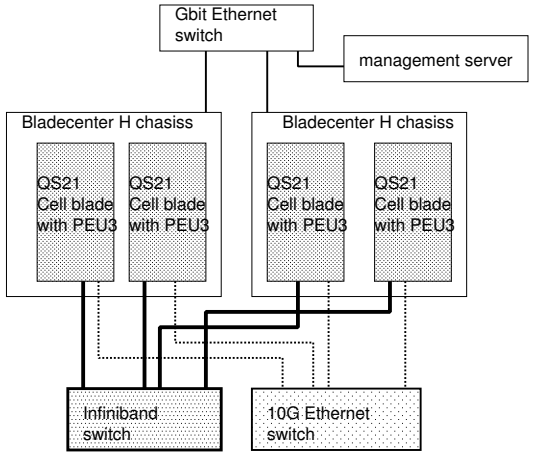

図 5 Cell B.E. クラスタ 評価環境

性能は 2.458Tflops と予測される．稼働予定のシャーシは全部で 15 基なので，合計の実効性能は 36.88Tflops 程度と予測される．この数値は Top500 のサイトによると，2010 年 6 月の 500 位が 24.67Tflops なので，Top500 ランクインが可能な性能である．

5. 評　　価

5.1 評価環境

評価環境を図 5 に示す．まず，QS21 Cell blade に PCIe 拡張スロットである PEU3 を取り付け，Infiniband カード，10G Ethernet NIC をそれぞれ接続した．それらを図 5 のようなネットワーク構成で接続した．なお，シャーシとブレードは帯域 1Gbps の内部バスで接続されている．また，Infiniband は SDR, x4 レーン (10Gbps) で動作する．

評価環境の構成を表 2 に示す．基礎評価のために MPI 実装として MPICH2 を使用した．

5.2 基礎評価

基礎評価として，Gbit Ethernet, 10G Ethernet, Infiniband をそれぞれインターコネクトとして用いた際の，ping-pong レイテンシおよびスループットを測定した．

測定は，MPI で送信するメッセージサイズを変更して行った．

5.2.1 レイテンシ

図 6 にそれぞれのインターコネクトにおける ping-pong レイテンシを示す．縦軸がレイ

表 2 評価環境

Chasiss	IBM Bladecenter H
Prossesor	QS21 Cell Blade
Memory size	2 GB
Infiniband HCA	Mellanox MHGS18-XTC
Ethernet NIC	Myricom 10G-PCIE-8A
Infiniband switch	voltaire ISR 9096
Ethernet switch	Force10 S2410
OS	Fedora 7
kernel	2.6.22-5.20070920bsc
MPI	mpich2

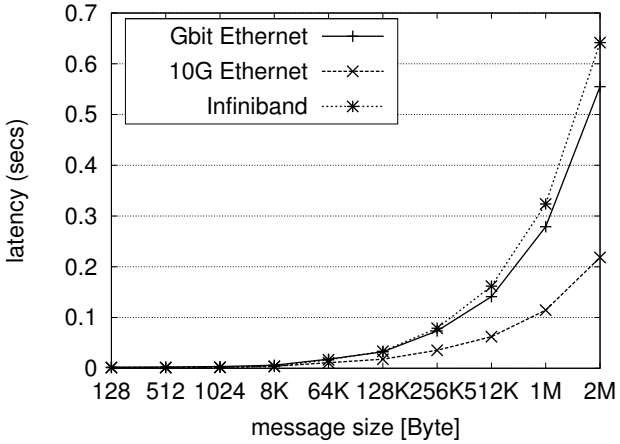

図 6 インターコネクトごとのレイテンシ

テンシ，横軸が送信したメッセージサイズである．

2 ノード上でベンチマークを動かし，ベンチマーク内では一方から他方へ MPI_SEND をした後，他方からこちら側へメッセージを送り返している．

結果から，Infiniband よりも 10G Ethernet の方がレイテンシが小さいことが示されている．これは互換性を重視して IPoIB を用いたことにより，InfiniBand の性能を引き出せなかったことが原因と考えられる．

5.2.2 スループット

図 7 にそれぞれのインターコネクトにおけるスループットを示す．縦軸がスループット，

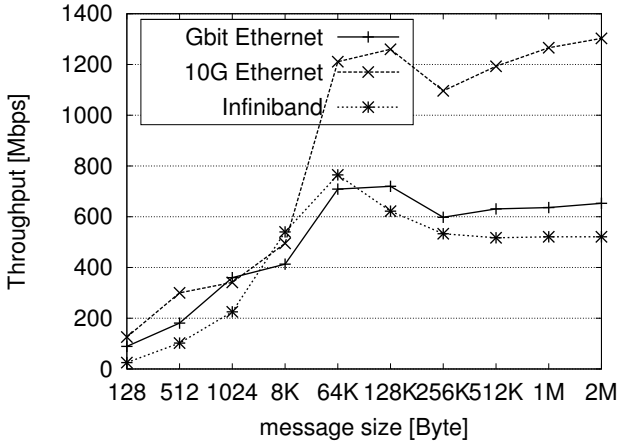

図 7 インターコネクトごとのスループット

横軸がメッセージサイズである．

図 7 のスループットもレイテンシと同様に，2 ノード上で MPI_SEND を十分な回数繰り返した場合の平均値を使用している．

5.3 High Performanc Linpack

アプリケーションを用いた評価として，それぞれのインターコネクト上で HPL を実行して評価した．HPL は TOP500 ランキングで用いられているベンチマークである．ユーザが問題サイズやグリッドサイズなどの各種パラメータを設定でき，以下の特徴がある[9)10)]．

表 3 に実行時に与えた主なパラメータを示す．なお，HPL のバージョンは 2.0，使用した線形代数ライブラリは ATLAS である．文献 11) によると

- 計算時間の多くが行列積の計算が占める
- 問題サイズ N は，メモリに収まる最大の大きさにする
- ブロックサイズ NB に関して
  - NB が大きいと，通信頻度は低くなるが，各ノードの処理量はバランスを欠くようになる
  - NB が小さいと，通信頻度が多くなるが，各ノードの処理量は均一化される方向に近づく

このうち，ネットワークの性能に対し特に影響を与えるのはブロックサイズ NB である．ク

表 3 主な HPL パラメータ

N	6400
NB	32, 64, 128
(P, Q)	(1, 8), (1, 4), (1, 1)
BCAST	1 ring

ラスタで使用予定の Infiniband と，優れた通信性能を示した 10Ethernet の 2 種類のインターコネクトに対して，NB の値を変更して HPL 実行結果を測定した．このとき，HPL を実行する際の評価環境のノード数を 4 ノード，2 ノード，1 ノードの 3 通りの構成で評価した．測定結果を図 8 および図 9 に示す．

まず，ブロックサイズ NB を変更して測定した場合に関して考察する．NB を増加していくと，基本的には実行性能も高くなる．この結果から，今回の評価環境ではロードバランスの良し悪しよりも通信発生回数の多さの方が，性能に影響があることが示された．

続いて，ノード数の変化にともなう 10G Ethernet と Infiniband の性能差に関して考察する．まず，1 ノードでの実行時はネットワークを介した通信は発生していないこともあり，性能差は最大でも 0.05Gflops 程度に収まっていることから測定時の変動の範囲内であると考えられる．一方，2 ノードでの実行時は，どの NB においても 10G Ethernet の方が 0.2Gflops 程度性能が高いことが示されている．また，4 ノードでの実行時は，10G Ethernet の方が約 0.4Gflops 高い性能が示された．これらのことから，ネットワークの種類に対する性能差は図 6 および図 7 における結果と同様に，10G Ethernet の方が性能面で有利である．しかし，Infiniband との性能差が最も大きい 4 ノードの場合でさえ，10Gflop 程度のうちの 0.4Gflops であり，高々4%程度である．このことから，ノード数が 4 程度の小型のシステムにおいて HPL を実行した場合は，ネットワーク面よりも，それ以外の処理のほうが支配的であることが示された．

6. まとめと今後の課題

本研究報告では，まず現在我々が構築中の異種混合型の Cell blade クラスタについて述べた．本クラスタは，156 台の QS21 ブレード，すなわち 312 個の Cell/B.E. を搭載し，そのインターコネクトには Ethernet, InfiniBand, Fibre Channel を採用している．

Cell blade によるクラスタを構築する事前の評価として，Blade 間を Ethernet および InfiniBand で接続した場合の通信性能評価を行った．また，High Performance Linpack (HPL) を 4 ノードの QS21 上で実行することで性能測定を行い，本クラスタが稼働した場合の性能を見積った．

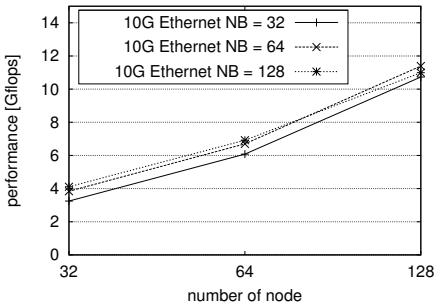

図 8 10G Ethernet 上での HPL 実行結果

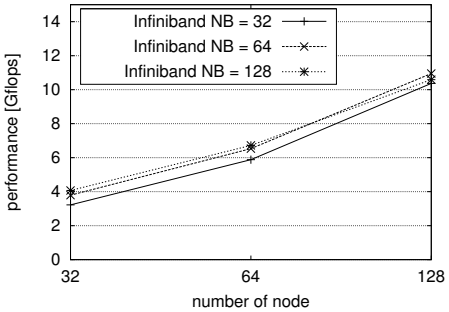

図 9 Infiniband 上での HPL 実行結果

今後の予定としては，今年度 11 月完成を目標にクラスタの構築を続けるとともに，通信に Fibre Channel も使用した場合の性能を評価する予定である．また，Cell/B.E. の単精度浮動小数点演算性能を活かすため，HPL を Mixed Precision 方式でチューニングする予定である．

謝 辞 本研究にて使用した Cell Blade クラスタは株式会社ソニー・コンピュータエンタテインメントから寄付を受けたものである．また，本研究を行うにあたり，数多くの貴重なアドバイスを頂きました，国立情報学研究所の鯉渕道紘准教授，ならびに東芝セミコンダクタ社の渡邊幸之介氏に感謝いたします．

参 考 文 献

1) 東京工業大学学術国際情報センター：TSUBAME2, http://www.gsic.titech.ac.jp/tsubame2.
2) Barker, K.J., Davis, K., Hoisie, A., Kerbyson, D.J., Lang, M., Pakin, S. and Sancho, J.C.: Entering the petaflop era: the architecture and performance of Roadrunner, *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, Piscataway, NJ, USA, IEEE Press, pp.1–11 (2008).
3) Barker, K.J. and Kerbyson, D.J.: Performance Analysis of an Optical Circuit Switched Network for Peta-Scale Systems, *Euro-Par 2007 Paralle Processing*, Vol.4641/2007, pp. pp.858–867 (2007).
4) 滝澤真一朗，遠藤敏夫，　松岡聡：光サーキットネットワークの補助的利用による HPC アプリケーション性能向上，2009 年ハイパフォーマンスコンピューティングと計算科学シンポジウム (HPCS2009)，pp.pp.65–72 (2009).
5) Project(IESP), I. E.S.: Roadmap 1.0, http://www.exascale.org/mediawiki/images/4/

42/IESP-roadmap-1.0.pdf.
6) TOP500: Supercomputing Sites, http://www.top500.org.
7) Altevogt, P., Boettiger, H., Boettiger, H. and Krnjajic, Z.: IBM BladeCenter QS21 Hardware Performance, http://www.ibm.com/developerworks/power/library/pa-qs21perf/index.html.
8) Kurzak, J. and Dongarra, J.: Implementation of a Mixed-Precision in Solving Systems of Linear Equations on the CELL Processor, *LAPACK Working Note 177* (2006).
9) 廣安知之，三木光範，荒久田博士：テラフロップスクラスタの構築と Benchmark による性能評価，同志社大学理工学研究報告 45(4)，pp.pp.187–198 (2005-1).
10) 笹生健，松岡聡：HPL のパラメータチューニングの解析，ハイパフォーマンスコンピューティング 2002(80)，pp.pp.125–130 (2002).
11) Hiroyasu, T., Miki, M. and Kugii, Y.: Evaluation of Linux-based High Performance Computing Cluster using LINPACK Benchmark, ISDL ジャーナル (2003).