Vol.2012-ARC-202 No.7
Vol.2012-HPC-137 No.7
2012/12/13

大規模並列システムのノード間通信を考慮した性能モデルに関する一検討

安田 一平[1]　小松 一彦[2,3]　江川 隆輔[2,3]　小林 広明[2]

概要：近年，大規模並列システムのノード数が増大するのに伴い，その高い演算性能を引き出すためには各ノードの演算性能ばかりではなく，ノード間の通信性能を考慮する必要がある．そのため，大規模化したシステムにおいて，容易にアプリケーションの性能解析を示すことができる手法が求められている．アプリケーションの性能解析や，最適化指針を与える方法として，性能モデルを用いたボトルネック解析が挙げられる．しかしながら，ノード間の通信を考慮した性能モデルや性能モデルに基づく解析・最適化手法は確立されていない．本報告ではノード間の通信を考慮したシステムの性能モデルを提案し，SX-9, Nehalem EX クラスタ，FX1，FX10，SR16000 の 5 つの大規模並列システムを用いて提案するモデルの妥当性を調査する．

1. はじめに

大規模並列システムは，科学，地球全体の気象・海流・温度のシミュレーション，航空機の周囲に流れる空気の流体計算シミュレーションなど，様々な分野で利用されている．より大規模かつ高精度なシミュレーションを行うためには，並列システムの更なる性能向上が求められている．

システムの演算性能を向上させるためには，ノードあたりの演算性能の向上と，ノード数の増加が考えられる．ノードあたりの演算性能を向上させるためには，マルチコアプロセッサのコア数の増加，コアあたりの性能向上，アクセラレータの採用など様々な取り組みが積極的になされている [1]．また，ノード数を増加させることで，システム全体の性能を向上させている．2002 年当時の最大規模のシステムは数千ノード程度であったが，現在，京などのシステムでは約十万ノードにも達する [2][3]．今後は数十万ノードや数百万ノードの構成を持つシステムの登場が予想されており，ノード数がさらに増加すると考えられる [4]．

一方，演算性能の向上速度に対して，通信性能の向上速度は遅いため，近年，大規模計算システムの演算性能と通信性能の性能差が拡大している．また，ノード数が増加するにつれて，ノード間の通信量も増加するため，実行時間のうちノード間通信が占める時間の割合が大きくなっている [4]．そのため，大規模並列システムの高い性能を引き出すためには，ノード間通信を考慮したアプリケーションの最適化が必要になる．アプリケーションの最適化では，性能向上を阻害する要因を特定するボトルネック解析に基づいた，最適化の指針が重要となる．

本研究では，通信がボトルネックの解析が容易な，大規模並列システムのための性能モデルを提案する．提案する性能モデルの妥当性を複数の大規模並列システムとベンチマークプログラムを用いて調査する．

2. ルーフラインモデル

CPU とメモリからなるノードを表す性能モデルとして，ルーフラインモデルが提案されている [5]．ルーフラインモデルは，性能が演算性能のみならず，オフチップのメモリ性能にも影響されると仮定し [6]，最大演算性能を視覚的に示すモデルである．

ルーフラインモデルでは，理論演算性能，メモリバンド幅及び，演算と演算に必要なデータの比である演算密度を利用し，二次元の両対数グラフに最大演算性能を示す．ルーフラインモデルでは，グラフの縦軸を最大演算性能 (GFlops/s)，横軸を演算密度 (Flops/Byte) とする．演算密度は，演算数とメモリからのデータ転送量の比を示し，以下の式で示される．

$$\frac{Flops}{Byte} = \frac{\text{演算数}\,(Flops)}{\text{データ転送量}\,(Bytes)} \qquad (1)$$

[1] 東北大学大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University
[2] 東北大学サイバーサイエンスセンター
Cyberscience Center, Tohoku University
[3] 科学技術振興機構戦略的創造研究推進事業
Japan Science and Technology Agency, Core Research for Evolutional Science and Technology

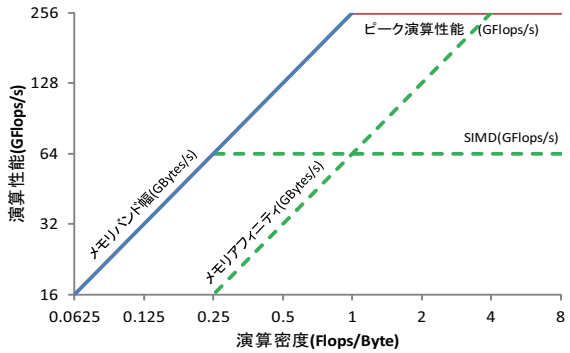

図 1 ルーフラインモデルの例

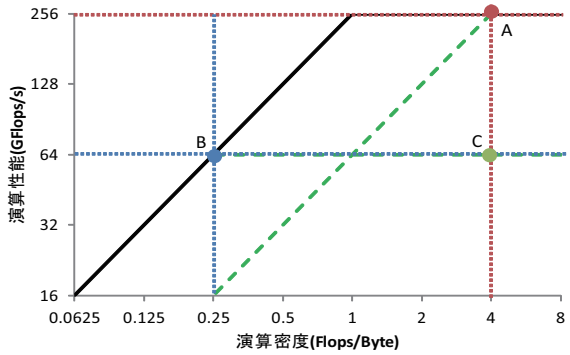

図 2 ルーフラインモデルを用いたボトルネック解析

図 1 に，理論演算性能が 256GFlops/s, メモリバンド幅が 256GBytes/s のノードのルーフラインモデルを示す．まず，ノードの最大演算性能が 256GFlops/s となるため，図 1 の赤色の線を描画する．次に，1 ノードあたりのメモリバンド幅は 256GBytes/s であるため，メモリバンド幅と，横軸の通信演算密度の値に応じ，$\frac{\text{メモリバンド幅 } Bytes/s}{\text{演算密度 Flops/Byte}}$ より求められる値より図 1 の青色の線を描画する．理論演算性能とメモリバンド幅による赤色と青色の線で示されるルーフが各システムで達成できる最大演算性能を表す．また，2 つの線の交点をリッジポイントと呼ぶ．最大演算性能を式で表すと以下の通りとなる．

$$\text{最大演算性能} = Min\left(\text{理論演算性能}\,(Flops/s)\; or \frac{\text{メモリバンド幅}\,(Byte/s)}{\text{演算密度}\,(\text{Flops/Byte})}\right) \quad (2)$$

ルーフラインモデルでは，理論演算性能とメモリバンド幅のルーフラインのほかに，Fused Multiply Add(FMA) や Single Instruction Multiple Data(SIMD) が利用できない場合の演算性能や，アフィニティの違いによるメモリバンド幅を補助線として表すことができる．これらの線は，図 1 の緑色の点線のように，理論演算性能と理論メモリバンド幅によるルーフの内側に描画される．

ルーフラインモデルを用いて，アプリケーションのボトルネックを容易に解析することが可能となる．また，この解析結果に基づき，アプリケーションの最適化指針を立てることができる．図 2 にボトルネック解析の例を示す．図 2 の点 A のように，アプリケーションの実効性能が，理論演算性能によるルーフ近辺に位置する場合，システムの演算性能を最大限に利用していると考えられる．そのため，既存の演算器ではこれ以上の性能を引き出すことができない．さらなる高速化を行うための手法として，演算器自体の性能向上があげられる．

図 2 の点 B のように，アプリケーションの実効性能が，メモリバンド幅によるルーフ近辺に位置する場合，ノードのメモリバンド幅に実効性能が制限されている．メモリボトルネックの場合，より高い実効性能を実現するため，メモリから CPU へのデータ転送回数の削減を行う最適化が有効だとの指針を示すことができる．このような場合の最適化手法としては，レジスタ内のデータの再利用性向上，ループ分散やキャッシュブロッキングによるキャッシュメモリ上のデータの再利用性向上があげられる．

図 2 の点 C のように，アプリケーションの実効性能が，どちらのルーフ付近にもない場合は，演算性能かメモリバンド幅のどちらがボトルネックかを判断することが難しい．そのため，データ転送量を削減する最適化と，演算性能を向上させる最適化の両方の最適化が必要である．演算性能を向上させる最適化としては，ループアンローリングやループ融合による分岐の削減があげられる．

3. ノード間通信を考慮したルーフラインモデル

多数のノードから構成される大規模並列システムでは，システム全体の性能向上にノード間通信性能の性能向上が追い付かず，ノード間通信が容易にボトルネックとなりえるため，ノード間通信の振る舞いを考慮することが重要である．

従来のルーフラインモデルは，複数のノードから構成される大規模計算システムにおけるボトルネック解析が困難である．そのため，本報告ではノード間通信を考慮したルーフラインモデルを提案し，システム全体のボトルネック解析を行う．

提案するルーフラインモデルでは，1 ノードの最大演算性能が，ノード内の演算性能または，ノード間の通信性能に律速されると仮定する．理論演算性能，ノード間通信バンド幅と，演算と演算に必要なノード間通信量の比である通信演算密度を利用して，二次元の両対数グラフに 1 ノードごとに最大演算性能を示す．縦軸は従来のルーフラインモデルと同じく最大演算性能 (GFlops/s)，横軸は通信演算密度 (Flops/Byte) とする．通信演算密度とは，演算数とノード間通信量の比であり，以下の式で示される．

$$\frac{Flops}{Byte} = \frac{\text{演算数}\,(Flops)}{\text{ノード間データ転送量}\,(Bytes)} \quad (3)$$

図 3 に，ノード間通信を考慮したルーフラインモデルを

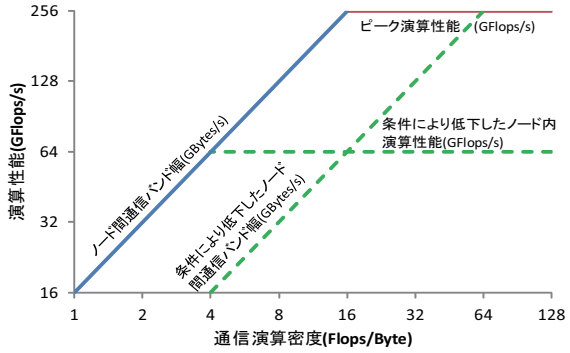

図 3 ノード間通信を考慮したルーフラインモデルの例

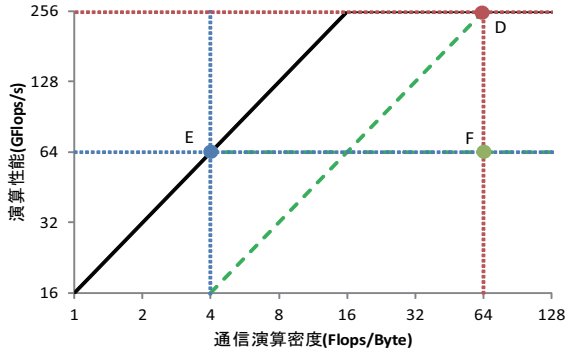

図 4 ノード間通信を考慮したルーフラインモデルを用いたボトルネック解析

示す．図 1 で示す従来のルーフラインモデルと異なる点として，演算密度の代わりに通信演算密度を利用し，ノード間通信を表現していることがあげられる．提案するルーフラインモデルにおいても，理論演算性能とノード間通信バンド幅のルーフに加え，ノード内での最適化が不十分などの条件により低下したノード内演算性能や，通信粒度が細かいなどの条件により低下したノード間バンド幅を補助線として表すことができる．これらの補助線を利用することにより，ルーフ以外のボトルネックによりシステムの性能が発揮されない場合を容易に解析することが可能になる．

提案する性能モデルを用いることで，ノード間の通信がボトルネックか否かの解析を行い，ボトルネックを解消するためにどのような最適化を行うべきかの判断することができる．

図 4 に提案するルーフラインモデルを利用したボトルネック解析の例を示す．点 D のように，アプリケーションの実効性能が，理論演算性能によるルーフ近辺に位置する場合，従来のルーフラインと同様，ノード内の演算性能を最大限に活用していると判断できる．

点 E のように，アプリケーションの実効性能が，ノード間通信バンド幅を示すルーフ近辺に位置する場合，ノード間通信バンド幅に演算性能が制限されている．ノード間通信バンド幅がボトルネックの場合，より高い実効性能を実現するためには，ノード間データ転送量を削減する最適化

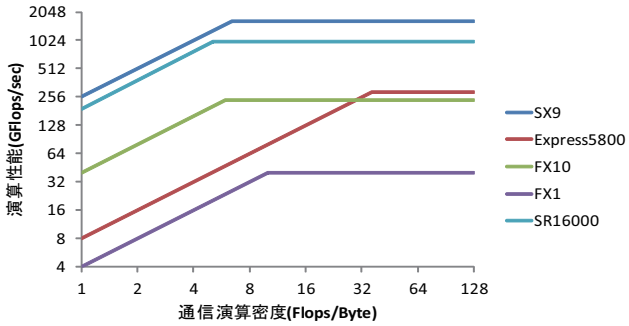

図 5 各システムのルーフライン

が必要である．このような場合の最適化手法としては，タスク配置の最適化や余分な通信の削減があげられる．

点 F のように，アプリケーションの実効性能が，どちらのルーフ近辺にも位置せず，ある条件下でのノード内演算性能を示す補助線付近に位置する場合，補助線が示す条件の影響でノード内演算性能が低下している可能性があると判断できる．このように，補助線が表している条件をアプリケーションのボトルネック解析の候補とすることができる．

4. 性能評価

本節では，複数の大規模並列システムを用いた評価を通じて，提案するノード間通信を考慮したルーフラインモデルを検証する．まず，提案する性能モデルの評価に用いた大規模並列システムとベンチマークの概要を説明する．次に評価結果を利用し，提案するルーフラインモデルについて考察を行う．

4.1 評価に用いた大規模並列システム

表 1 に示す 5 つの大規模並列システムを用いて，提案する性能モデルの評価を行った．図 5 に示すように，これらのシステムは，演算性能，ノード間通信バンド幅，ノード間ネットワークトポロジーがそれぞれ異なるため，異なるルーフラインを持つ．

NEC SX-9 は 1676.8GFlops/s のノード理論演算性能を持つシステムである [7]．ノード間は IXS と呼ばれる片方向 128GBytes/s の転送性能を持つネットワークで接続されており，1 ノードあたりのリッジポイントにおける横軸の値は 6.4 である．IXS のネットワークトポロジーはフルクロスバー型である．

Nehalem EX クラスタは 289.92GFlops/s のノード理論演算性能を持つシステムである [8]．ノード間は片方向 4GB/s の転送性能の Infiniband QDR x4 により接続されており，1 ノードあたりの通信演算密度は 38.2 である．Nehalem EX クラスタのネットワークトポロジーはスター型である．

FX1 は 40GFlops/s のノード理論演算性能を持つシステ

表 1　対象とする HPC システムの各種性能

システム名	理論演算性能 (GFlops/s)	理論通信バンド幅 (GBytes/s)	ネットワークトポロジー	通信演算密度	利用ノード数
NEC SX-9	1676.8	128	スター	6.4	16
Intel Nehalem EX クラスタ	289.92	4	スター	36.2	4
Fujitsu FX1	40	2	ファットツリー	10	128
Fujitsu FX10	236	5-50(通信先ノードにより変化)	6 次元トーラスメッシュ	2.36-23.6	12
Hitachi SR16000 M1	980.48	96-24(利用ノード数により変化)	多階層完全結合	5.1-20.41	64

ムである [9]．ノード間は片方向 2GBytes/s の転送性能の Infiniband DDR x4 により接続されており，1 ノードあたりのリッジポイントにおける横軸の値は 10 である．FX1 のネットワークトポロジーはファットツリー型である．

FX10 は 236GFlops/s のノード理論演算性能を持つシステムである [10]．ノード間は片方向 5GBytes/s の転送性能をもつ Tofu ネットワークルータにより接続されている．FX10 は，ノード間通信に複数の経路を持つ．FX10 は 10 ポートのインターコネクトを持つため，隣接するすべてのノードにデータを転送する場合は，片方向 50GBytes/s のノード間通信バンド幅を持ち，リッジポイントにおける横軸の値は 2.36 となる．1 つのノードにのみデータを転送する場合は，1 ポートのみを用いるため，ノード間通信バンド幅は 5GBytes/s となり，リッジポイントにおける横軸の値は 23.6 となる．FX10 のネットワークトポロジーは 6D トーラスメッシュ型であり，通信先によって通信バンド幅や通信レイテンシが異なる．

SR16000 は 980GFlops/s の理論演算性能を持つシステムである [11]．SR16000 のネットワークトポロジーは多階層の完全結合ネットワークである．SR16000 では各 CPU とネットワーク Hub と呼ばれるチップが片方向 96GBytes/s で接続されている．この部分がボトルネックになる場合は，リッジポイントにおける横軸の値は 5.1 となる．また，8 ノードの組を drawer と呼び，drawer 内の各ノードは片方向 24GBytes/s の転送性能となる．この場合，1 ノードとの通信によるリッジポイントにおける横軸の値は 20.4 となる．各 drawer 間は片方向 10GBytes/s の接続が 2 本ずつ．片方向合計 20GBytes/s の接続により結合されている．drawer 間の通信では，異なる drawer への通信を組み合わせ，間接通信を行うことができる．間接通信を行う場合，drawer 間の通信経路は，最大 46 本となり，drawer 間通信バンド幅は最大 460GBytes/s となる．この場合，リッジポイントにおける横軸の値は 8.1 となる．

4.2　評価用ベンチマークの概要

提案するノード間通信を考慮したルーフラインモデルを検証するためには，通信演算密度を変化させた際の最大実効性能がルーフラインの内部に収まるかどうかをまず確認する必要がある．そのため，通信演算密度を変化させることが出来るように，通信量と演算数をそれぞれ独立して変更することができるベンチマークを作成した．このベンチマークは，行列積計算を行いながら，計算と無関係な通信を同時に行う．また，同じノード間の通信が同時に起こらない，理想的な状態にてノード間の通信を行う．

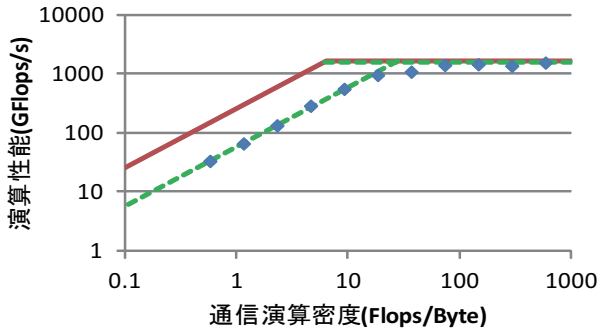

図 6　SX-9 16 ノードでの評価結果

ベンチマークの演算数が通信量より大きく通信演算密度が大きい場合は，ノード間の通信に必要な時間よりも，行列積演算に要する時間が長く，演算性能がボトルネックになる．通信量が演算数より大きく通信演算密度が小さい場合，行列積演算にかかる時間よりも，ノード間通信に要する時間のほうが長いため，ノード間通信がボトルネックになる．

4.3　ノード間通信を考慮したルーフラインモデルの評価

ベンチマークを用いて，ノード間通信を考慮したルーフラインモデルの検証を行う．ベンチマークのボトルネック解析には，理論通信性能や理論ノード間通信バンド幅のルーフの他に，補助線を利用することでより詳細な解析を行う．補助線として，ベンチマークの 1 ノードで実行した際の実効演算性能と，評価用ベンチマークの通信部分のみを用いて測定したノード間通信バンド幅を用いる．

図 6 に，SX-9 の評価結果を示す．通信演算密度が 75 以上と大きい場合，各点が理論演算性能を示す直線近辺に位置しているのがわかる．そのため，通信演算密度が大きい場合はベンチマークがノード内の演算性能を最大限に利用していると考えられる．一方，通信演算密度が 18 以下と小さい場合は各点が理論ノード間通信バンド幅以内を示すルーフの内側であり，かつ実測ノード間通信バンド幅を示すルーフの近辺に位置する．このように通信演算密度が小さい場合は，通信ボトルネックと考えられる．

図 7，図 8 は，Nelahem EX クラスタと FX1 の評価結果をそれぞれ示す．図 7，図 8 において，通信演算密度が 140

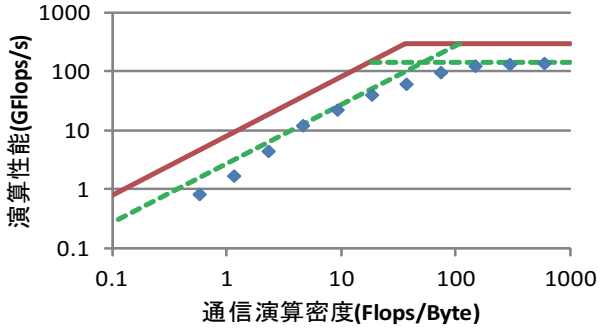

図 7 Nehalem EX クラスタ 4 ノードでの評価結果

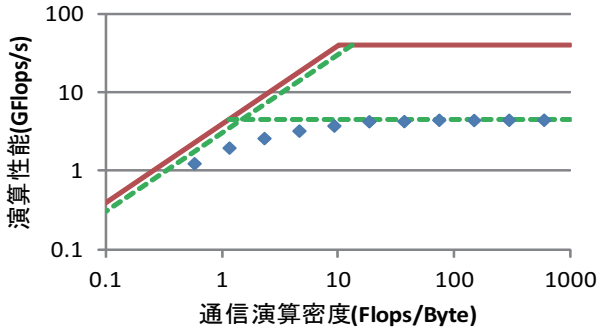

図 8 FX1 128 ノードでの評価結果

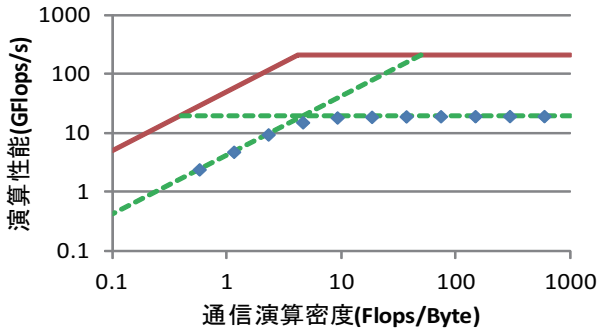

図 9 FX10 12 ノードでの評価結果

以上と大きい場合，各点が理論演算性能ではなく，1 ノードあたりの演算性能を示すルーフ近辺に位置している．これは，Nehalem EX クラスタや FX1 の 1 ノードで行列積演算を行った際の実効性能が，理論演算性能より小さいため，1 ノードあたりの実効性能を，当ベンチマークを利用した際の最大演算性能とみなすことができるためである．そのため，Nehalem EX クラスタや FX1 では演算性能を制限する要素として，1 ノードあたりの性能を用いることが適当である．これにより，SX-9，Nehalem EX クラスタ，FX1 では，提案するルーフラインモデルによって描画されるルーフによって，最大演算性能が表されていることを確認できた．

図 9 には，FX10 の評価結果を示す．FX10 では，通信演算密度が 10 以上と大きい場合は，他システムと同様の

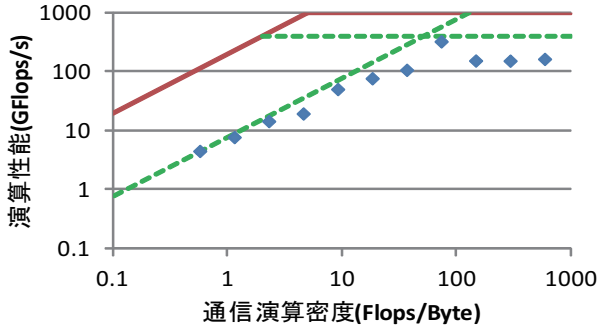

図 10 SR16000 8 ノードでの評価結果

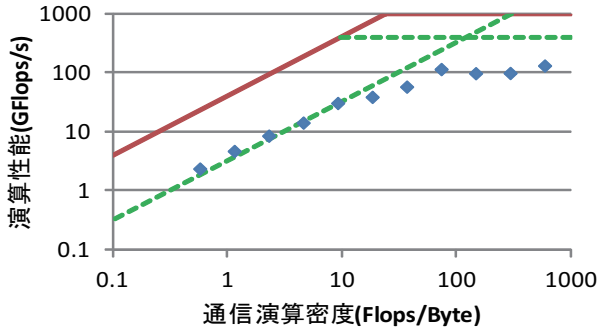

図 11 SR16000 64 ノードでの評価結果

結果を示す．一方，通信演算密度が 3 以下と小さい場合，各点はルーフが示す最大演算性能の 1/10 程度と，かなり低い位置になっている．これは，理論通信バンド幅をすべてのポートを利用した場合の 50GBytes/s とし，ルーフを描画したためである．しかしながら，FX10 でのノード間通信には，複数の経路が存在するため，すべてのポートの通信バンド幅を有効に利用してノード間通信を行うことは困難である．そのため，実効ノード間通信バンド幅をより正確に測定し，ルーフを描画する必要がある．

図 10 に SR16000 を用いて，8 ノード，1drawer で実行した際の評価結果を示す．SR16000 の結果では，Nehalem EX や FX1 と異なり，通信演算密度が 140 以上と大きい場合は，実効性能が 1 ノード実行時の性能より小さい場合が多い．これについては，後日解析する必要がある．しかしながら，通信演算密度が 9 以下と小さい場合は各点が実測通信バンド幅を示すルーフの近辺に位置している．そのため，通信演算密度が小さい場合は，提案するルーフラインによる，モデルによって示される通信のルーフによって，最大演算性能が制約されていることを確認できる．

また，図 11 に 64 ノード，8drawer で実行した際の評価結果を示す．64 ノードで実行した際にも，8 ノードとほぼ同様の傾向がみられる．よって，drawer 内と drawer 外で異なる通信バンド幅を持つネットワークであるにもかかわらず，それぞれのボトルネックを表現できることが確認できる．

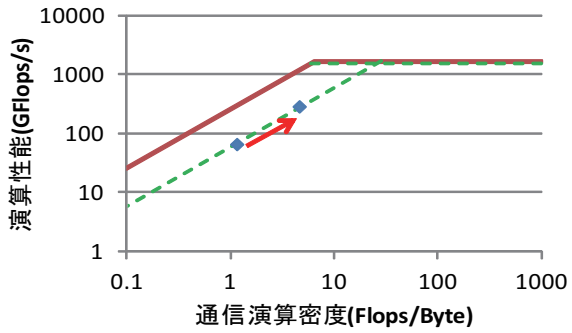

図 12 通信削減後の予想性能

以上のように本評価では，異なるネットワークトポロジーを持つすべてのシステムにおいて，適切な通信バンド幅を用いることにより，通信のルーフで最大演算性能が制限されていることを確認することができる．

提案するルーフラインモデルを用いて解析した結果，通信がボトルネックと判明した場合，通信演算密度を向上させる最適化が有効である．図 12 は，SX-9 の評価結果の一部を抜き出したグラフである．通信の最適化により，通信量を 1/4 に削減できるとすると，通信演算密度が 4.59 となり，理論演算性能は 1175GFlops/s となる．そのため，通信演算密度が低く通信バンド幅がボトルネックとなる場合，通信の削減による最適化が効果的であることがわかる．

5. おわりに

大規模並列システムでは，ノード間通信が実効性能を決める大きな要因となりつつあるため，容易にノード間通信がボトルネックかどうかを判断する必要がある．本報告ではノード間通信を考慮したルーフラインモデルを提案し，ベンチマークを利用して提案する性能モデルを検証した．今後の課題としては，トーラスメッシュ型などの 1 ノードが複数のインターフェイスを持ち，他ノードへの通信に複数の経路が存在するネットワークにおけるノード間通信に基づくルーフラインモデルの検討や，実アプリケーションを用いた検証が必要である．

謝辞 本研究は，北海道大学情報基盤センター，東北大学サイバーサイエンスセンター，東京大学情報基盤センター，名古屋大学情報基盤センターのスーパーコンピュータを利用することで実現することができた．本研究の一部は，文部科学省科研費研究 (S)(21226018) と科学技術振興機構 (JST) 戦略的創造研究推進事業 (CREST) 研究領域「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」研究課題「進化的アプローチによる超並列複合システム向け開発環境の創出」の助成を受けている．

参考文献

[1] Steen, A. J. V. D.: Overview of recent supercomputers, Technical report, NCF (2011).
[2] Top 500 Supercomputer sites, http://www.top500.org/.
[3] Yokokawa, M., Shoji, F., Uno, A., Kurokawa, M. and Watanabe, T.: The K computer: Japanese next-generation supercomputer development project, *Low Power Electronics and Design (ISLPED) 2011 International Symposium on*, pp. 371 –372 (online), DOI: 10.1109/ISLPED.2011.5993668 (2011).
[4] Bergman, K., Borkar, S., Campbell, D., Carlson, W., Dally, W., Denneau, M., Franzon, P., Harrod, W., Hiller, J., Karp, S., Keckler, S., Klein, D., Lucas, R., Richards, M., Scarpelli, A., Scott, S., Snavely, A., Sterling, T., Williams, R. S., Yelick, K., Bergman, K., Borkar, S., Campbell, D., Carlson, W., Dally, W., Denneau, M., Franzon, P., Harrod, W., Hiller, J., Keckler, S., Klein, D., Kogge, P., Williams, R. S. and Yelick, K.ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems (2008).
[5] Williams, S., Waterman, A. and Patterson, D.: Roofline: an insightful visual performance model for multicore architectures, *Commun. ACM*, Vol. 52, pp. 65–76 (online), DOI: 10.1145/1498765.1498785 (2009).
[6] McKee, S. A.: Reflections on the memory wall, *Proceedings of the 1st conference on Computing frontiers*, CF '04, New York, NY, USA, ACM, pp. 162– (online), DOI: 10.1145/977091.977115 (2004).
[7] SX-9 装置緒元 : HPC ソリューション — NEC:, http://www.nec.co.jp/solution/hpc/sx9/product/spec.html.
[8] Express5800/A1080a - Nec:, http://www.nec-itplatform.com/-Express5800-A1080a-.html.
[9] HPC ハイエンドテクニカルコンピューティングサーバ FX1 : 富士通:, http://jp.fujitsu.com/solutions/hpc/products/fx1.html.
[10] Specifications : PRIMEHPC FX10 : Fujitsu Global:, http://www.fujitsu.com/global/services/solutions/tc/hpc/products/primehpc/spec/.
[11] SR16000 : 仕様 : 技術計算向けサーバ : 日立:, http://www.hitachi.co.jp/Prod/comp/hpc/SR_series/sr16000/spec.html.