

スループットを用いた「京」における MPI 通信性能の評価

北澤 好人¹ 黒田 明義^{2,a)} 志田 直之¹ 安達 知也¹ 南 一生²

受付日 2017年4月21日, 採録日 2017年8月2日

概要: スーパーコンピュータ「京」は 82,944 ノードから構成される超並列のシステムであり, アプリケーションで高い実効性能を得るためには, MPI 通信関数の性能が重要である. 本稿では, スーパーコンピュータ「京」における MPI 通信性能を測定し, 実測したスループットのピーク性能と半性能長から, 通信バンド幅とレイテンシを算出した. また通信モデルから通信時間の見積りを理論的に導出し, 測定値と比較することで性能の評価を行った. その結果, 隣接通信ではピーク性能と通信バンド幅がほぼ一致して, 遅延時間は通信のレイテンシによる値とおおむね合っていることが分かった. 集団通信では Bcast, Allreduce, Allgather について同様の評価を行い, ピーク性能は最大で約 20%, 遅延時間は約 50%の違いがあるものの, 測定結果が通信時間見積りと同様の傾向であることが分かった.

キーワード: スーパーコンピュータ「京」, MPI 通信性能評価, Intel MPI ベンチマーク, 通信バンド幅, レイテンシ

Evaluation of MPI Communication Performance Using Throughput on the K computer

YOSHITO KITAZAWA¹ AKIYOSHI KURODA^{2,a)} NAOYUKI SHIDA¹ TOMOYA ADACHI¹ KAZUO MINAMI²

Received: April 21, 2017, Accepted: August 2, 2017

Abstract: The K computer is a massively parallel system with 82,944 nodes, and the performance of the MPI communication function is extremely important to obtain high effective performance for the application. In this paper, we measured the MPI communication performance on the K computer and calculated the communication bandwidth and latency from the peak performance and the half-performance length of measured throughput. Furthermore, we derived the estimation formula of the communication time theoretically from the communication model, and evaluated the communication performance by comparing with the measured value. As a result, in nearest neighbor communication, the peak performance almost matched the communication bandwidth, and the delay time roughly matched the communication latency. In collective communication, we evaluated Bcast, Allreduce and Allgather similarly. The peak performance was up to about 20% different and delay time was up to about 50% different, but measurement results were the same trend as estimation results.

Keywords: K computer, performance of MPI communication, Intel MPI Benchmarks, band width, latency

1. はじめに

スーパーコンピュータ「京」(以下,「京」と記す)では, 様々な分野のアプリケーションプログラムが実行されている. 並列規模は数千~数万であり, 並列化のための通信処

理として MPI 通信が必須である.

「京」の開発時には, 通信 LSI や伝送経路などのハードウェアや通信ライブラリなどのミドルウェアが, 設計どおりの性能が出ているかを検証する必要があった. このため目標性能の設定や, 実測との比較が行われ, ライブラリやシステム開発などにフィードバックしていた. 目標性能の設定には, 通信時間の見積りを導出し, どの程度の性能が実現しうるかの評価を行った. ここでの主な評価ポイントは, ハードウェアの限界まで性能を引き出しているかに重

¹ 富士通株式会社
FUJITSU, LTD., Kawasaki, Kanagawa 211-8588, Japan

² 理化学研究所
RIKEN, Kobe, Hyogo 650-0047, Japan

^{a)} kro@riken.jp

きがおかれていたため、一部のメッセージサイズでの評価に偏りがちであった。また見積式の導出では、集団通信としては初めての試みであったといえるが、実際の開発見込みのアルゴリズムを十分に反映したものではなく、使用されるパラメータの評価も十分ではなかった。このためすべてのメッセージサイズの通信性能を評価するのは難しかった。

現在「京」を運用するにあたり、ユーザへの利便性の向上を目的に、ライブラリのバージョンごとに通信性能を測定して公開している [1]。実際にユーザが使用する条件で、「京」での通信が想定される性能を実現できているか検証する必要があった。本稿では、Intel MPI Benchmarks [2] (以下、IMB と記す) を用いて「京」における主要な MPI 通信の性能測定を行い [1]、実測されたスループットのピーク性能と半性能長から、通信バンド幅とレイテンシを算出した。また実際に開発され提供された通信アルゴリズムを考慮して精細化した通信性能の見積式を新たに導出し、広い範囲のメッセージサイズでの測定性能との比較を行ったのでその結果を報告する。

2 章では「京」の通信の概要、3 章ではスループットの概要を説明し、4 章で通信性能を検証した内容を報告する。5 章では展開と今後の課題をまとめる。

2. 「京」の通信性能の概要

本章では、「京」における計算ノード間の通信機構の概要を説明する。

2.1 通信性能の概要

「京」では、計算ノード間の通信ネットワークを構成するために、Tofu (Torus fusion) インタコネクト [3], [4] が採用されている。4 つの Tofu ネットワークインタフェース (TNI) により、最大で 4 方向の同時通信が可能である。また Tofu ネットワークルータ (TNR) は 10 本のリンクを持ち、各リンクの理論上の通信バンド幅は 5 [GB/s] (双方向) である [3]。ただし、同時通信についてはハードウェアのバスの性能制限により、1 ノードあたりの総通信バンド幅が約 15 [GB/s] に律速される [5], [6]。「京」ではこの計算ノードが 6 次元メッシュトラスの直接網で接続されており、ジョブごとに 3 次元トラスを切り出せる。

2.2 MPI 環境

「京」の MPI 環境は Open MPI Version 1.6.3 をベースに開発された富士通 MPI である。MPI の集団通信関数については、Open MPI 由来の通信アルゴリズムのほかに、Tofu インタコネクト向けに最適化された通信アルゴリズム (Tofu 専用アルゴリズム) を開発し、高い通信性能を実現している。たとえば Bcast 通信では、3 次元ノード形状の場合、メッセージを 3 分割して 3 方向同時通信をする通信

アルゴリズム Trinaryx3 を新規に開発した [5], [7]。これは直接網ネットワークポロジや同時通信の機構を活用したものとなっている。Open MPI のアルゴリズムと Tofu 専用アルゴリズムは、通信時のメッセージサイズに応じて、最適なアルゴリズムが自動的に選択される。測定に使用したセグメントサイズなどの条件は、Open MPI の MCA parameter を用いて設定した。

3. 通信のスループットの概要

通信性能を評価する指標の 1 つとして、スループットがある。スループットとは単位時間あたりに転送するメッセージサイズで、メッセージサイズを通信時間で割った値である。通信では、転送の初期にメッセージが到着するまでに遅延時間 t_L が発生し、その後、メッセージが転送時間 t_M で順次転送される。転送するメッセージサイズを M とし、通信バンド幅を B として、全体の通信時間 t は、以下のように M に関して線形の関係式で書き下せる。

$$t = t_L + t_M = t_L + \frac{M}{B} \quad (1)$$

スループット T_p は単位時間 t あたりのメッセージサイズ M であるので、以下となる。

$$T_p = \frac{M}{t} = \frac{M}{t_L + M/B} = \frac{B}{1 + (Bt_L)/M} \quad (2)$$

$T_p = B/2$ となる通信量を $M_{1/2}$ としたとき、これを半性能長と呼ぶ。半性能長を使ってスループットを表すと、以下ようになる。

$$T_p = B / \{1 + 1/(M/M_{1/2})\} \quad (3)$$

$$M_{1/2} = Bt_L \quad (4)$$

これを、ベクトル演算の性能で用いられた Hockney の pipe 関数 [8] を用い、スループットは以下のように書ける。

$$T_p = B \text{ pipe}(M/M_{1/2}) \quad (5)$$

$$\text{pipe}(x) \equiv 1/(1 + 1/x) \quad (6)$$

pipe 関数は、 $x = 1$ のときに $1/2$ となり、 x が大きい極限で 1、0 の極限で 0 である。式 (5) からスループット T_p はメッセージサイズ M が大きい極限でピーク性能である通信バンド幅 B に近づき、メッセージサイズが小さい場合、遅延時間 t_L の比率が大きくなる。半性能長 $M_{1/2}$ は通信バンド幅 B と通信レイテンシ t_L がバランスする点であり、半性能長 $M_{1/2}$ から、通信レイテンシ t_L を見積もることが可能である。

$$B = \lim_{M \rightarrow \infty} (T_p) \quad (7)$$

$$t_L = M_{1/2}/B \quad (8)$$

表 1 通信時間の実測値ならびに見積式から得られた通信性能

Table 1 Communication performance parameter by measurement and estimation.

	Parameter	PingPong	Bcast 1D/16KiB	Bcast 2D/16KiB	Bcast 3D/16KiB	Allreduce 3D/16KiB	Allgather 3D/～32KiB	Allgather 3D/64KiB～
Fitting	B [MB/s]	4,685	3,616	7,467	10,400	5,266	149,800	13,210
	$M_{1/2}$ [byte]	4.894E+04	1.010E+07	4.092E+06	1.756E+06	1.024E+06	6.206E+05	6.917E+03
	T_L [μ s]	10.45	2,795	548.0	168.9	388.8	1,590	201.1
Estimation	B [MB/s]	5,000	4,226	8,452	12,680	5,129	1,920,000	15,040
	$M_{1/2}$ [byte]	3.483E+04	7.919E+06	3.477E+06	1.818E+06	1.315E+06	7.718E+06	9.852E+03
	T_L [μ s]	6.965	1,874	411.4	143.4	512.8	1,544	251.6

4. スループットによる性能評価

通信のスループットを用いて、IMB の 1 対 1 通信の PingPong と集団通信の Bcast, Allreduce, Allgather の性能を評価した。測定に用いたベンチマークプログラムは IMB の Ver.3.2.4 で、通信ライブラリを含む「京」の言語環境のバージョンは、2016 年 5 月から 2017 年 1 月までの間、最新版として提供された K-1.2.0-20-1 である。

IMB を用いた測定は 1 回行った。ただし、IMB 内部では、通信サイズが 2^{15} [byte] 以下の小さいものについては、1,000 回実行されており、通信時間の最大値を用いて評価が行われている。通信サイズが 2^{25} [byte] 以上の大きいものについては、1 回の測定値となっている。

測定結果に対して、前述のスループットの式 (3) を最小二乗法によりフィッティングして、通信性能の指標となるピーク性能 B と半性能長 $M_{1/2}$ (または遅延時間 t_L) を算出した。また、通信モデルから通信時間の見積式を理論的に導出し、そこから算出した通信性能を比較し、評価を行った。この評価式は、システム設計時の通信性能予測や実際の完成したシステム性能評価を行うことを目的に、「京」開発時に通信性能を評価するために導出されたモデル式を、Tofu 専用アルゴリズムを考慮して拡張したものである。今回の評価で得られた各通信方式の通信性能を表 1 にまとめた。上段が実測値から得られた値で、下段が後述する見積式から算出した値である。

4.1 1 対 1 通信の性能評価 (PingPong)

本節では、1 対 1 通信として IMB の PingPong の性能を説明する。PingPong は、2 つのランク間において、一方のランクからデータを Send 通信により送信し、もう一方のランクで受信した後に Recv 通信によりデータを元のランクへ送り返すことで、データを往復させる。

(1) 測定結果のフィッティング

PingPong の測定には隣接の 2 ノードを用いた通信時間の測定を行った。また「京」の MPI による 1 対 1 通信は 13 [KiB] から通信アルゴリズムが変わる [5] ため、測定結

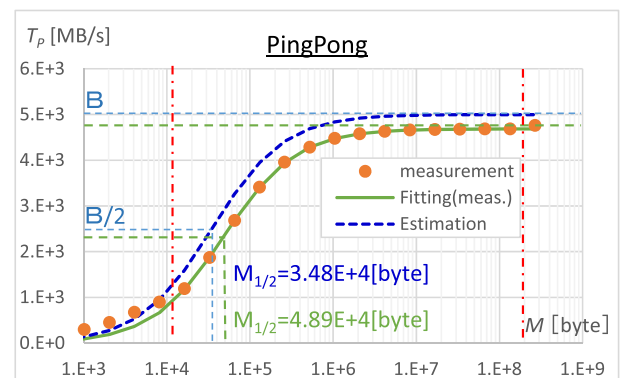


図 1 通信時間の測定結果と見積式から算出したスループットの比較 (PingPong)

Fig. 1 Comparison of throughput calculated from measurement and estimation (PingPong).

果の 13 [KiB] 以上のデータを用いて、最小二乗法によるフィッティングを行った。

求めた値は表 1 のとおりである。ピーク性能 B は、4,685 [MB/s] であり、通信バンド幅の理論ピーク 5,000 [MB/s] と比べて 6.3% 低い。ハードウェアが出しうる限界性能は、理論ピーク性能の約 95% となる 4,760 [MB/s] という報告があり [9]、今回測定された PingPong のピーク性能は、この性能に近い値であるといえる。スループットの測定結果とフィッティングの比較を図 1 に示す。丸印がスループットの測定結果で実線がフィッティング結果である。破線の範囲はフィッティングを行ったメッセージサイズである。

(2) 見積式からの通信時間の算出

通信時間はメッセージの転送時間と遅延時間の合計であり式 (1) で記述できる。遅延時間 t_L には、MPI レイテンシと 1 対 1 通信のための制御通信の時間が含まれる。MPI レイテンシは、「京」において隣接ノード間の片側転送の実測結果から、1.6 [μ s] が見積もられている。またこのほかに 13 [KiB] 以上の 1 対 1 通信では、データ転送の前に格納先のアドレスなどの情報を転送する制御通信が行われる [5]。「京」での制御通信は、3 種類の情報 (40 [byte], 64 [byte], 80 [byte]) が通信前後に転送される。それぞれ制御通信を

表 2 制御通信時間

Table 2 Time of control communication.

Size	Time [μ s]
40 [byte]	1.751
64 [byte]	1.785
80 [byte]	1.829
Total	5.365

含まない 13 [Kib] 以下の PingPong の通信時間を用いて評価すると表 2 となる。

以上から、遅延時間は合計で 6.965 [μ s] となる。ピーク性能 B は通信バンド幅の理論ピーク性能を用いた。これらの数字を式 (4) に代入すると、半性能長 $M_{1/2}$ が算出できる (表 1)。通信時間の見積式と測定結果のフィッティングによるスループットの比較は図 1 のとおりである。点線が見積式による結果である。見積式による遅延時間は約 7.0 [μ s] であり、測定結果による遅延時間の約 10.5 [μ s] と比べて短い。遅延時間の内訳は、制御通信時間と MPI レイテンシを用いて、ある程度説明可能であることが分かる。

4.2 集団通信の性能評価

本節では、IMB の集団通信の Bcast, Allreduce, Allgather の通信性能を説明する。Bcast はルートランクから全ランクにデータを転送する。実行時のノード形状は 1 次元、2 次元、3 次元の 3 種類について通信性能を評価する。Allreduce は全ランクのデータを集めて集合演算を行い、その結果を全ランクに転送する。Allgather は全ランクが持つデータを集めて全ランクへ転送する。Allreduce と Allgather については、実行時のノード形状が 3 次元での通信性能を評価する。

4.2.1 Bcast (1 次元)

(1) 測定結果のフィッティング

384 ノード (1 次元ノード割当て) における Bcast の測定結果について、PingPong と同様に評価した。「京」の Bcast 通信は Tofu 専用アルゴリズムの Trinaryx3 が適用され、パイプライン方式によりデータが転送される。パイプライン転送のイメージを図 2 に示す。パイプライン方式ではデータをセグメントに分割して転送を行い、実際の通信ではメッセージサイズによって最適なセグメントサイズが選択される。ここではセグメントサイズを 16 [KiB] に固定して測定を行った。Tofu 専用アルゴリズムの Trinaryx3 は、その次元によらずコミュニケーターの形状が連続な直方体形状であれば、同時転送する分割数を変更して適用することができる。Trinaryx3 は 128 [KiB] 以上のメッセージサイズから適用されるため、128 [KiB] 以降のデータを用いて最小二乗法によるフィッティングを行った。表 1 に求められた通信性能を、図 3 にスループットの測定結果とフィッティング結果を示す。丸印の点が測定結果で実線が

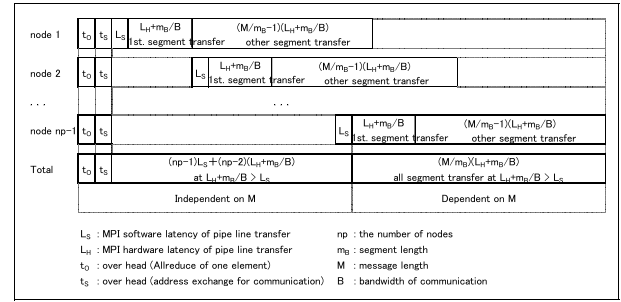


図 2 Bcast 通信のパイプライン転送のイメージ

Fig. 2 Concept of pipeline transfer.

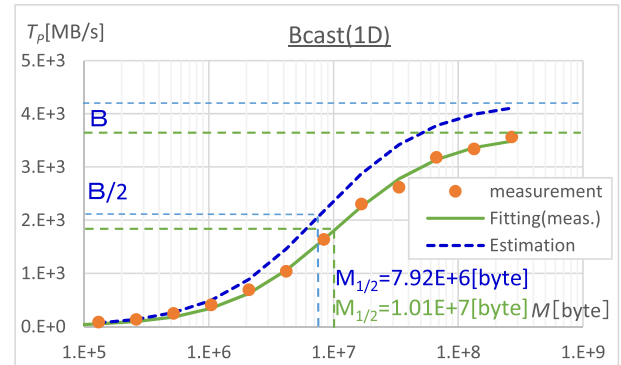


図 3 通信時間の測定結果と見積式から算出したスループットの比較 (Bcast/1D/16 KiB)

Fig. 3 Comparison of throughput calculated from measurement and estimation (Bcast/1D/16 KiB).

フィッティング結果である。

(2) 見積式からの通信時間の算出

Trinaryx3 はデータをセグメントに分割して、パイプライン方式により次のランクへ順次転送する。図 2 からパイプライン方式の通信時間は、メッセージサイズ M に依存しない時間と依存する時間の合計である。メッセージサイズ M に依存しない時間には、アルゴリズム判定処理のための 1 要素の allreduce などによるオーバーヘッド t_O に加え、通信領域のアドレス交換の時間 t_S が考えられる。パイプライン転送については、隣接ノード間あたりに MPI レイテンシと 1 セグメントのデータの転送時間から構成される。このうち MPI レイテンシにはソフトウェアとハードウェアそれぞれの通信の準備が含まれ、ソフトウェア部分 L_S とハードウェア部分 L_H に分けることができる。パイプライン転送中のソフトウェアの処理は、前のハードウェアの通信処理と重ねることができるため、通信時間としては L_S と $L_H + m_B/B$ の大きい方が支配的になる。また転送ノード総数を np とすると、パイプラインのステップ数は $np - 1$ となる。1 セグメントのサイズを m_B とおき、式 (1) を書き下すと通信時間 t は以下のように書き下せる。

$$t = t_O + t_S + (np - 1)(L_S + L_H + m_B/B) + (M/m_B - 1) \max(L_S, L_H + m_B/B) \quad (9)$$

表 3 見積式に用いた評価パラメータ (Bcast/1D/16 KiB)
Table 3 Evaluation parameter (Bcast/1D/16 KiB).

Name	Value
np	384 [node]
B	5,000 [MB/s]
m_B	16,384 [byte]
t_O	8.370 [μ s]
t_S	1.396 [μ s]
L_S	1.000 [μ s]
L_H	0.600 [μ s]

この式から、今回測定に用いたセグメントサイズの条件 $L_S < L_H + m_B/B$ について、メッセージサイズ M に依存しない項ならびに依存する項を抽出すると、遅延時間 t_L ならびに実効のピーク性能 B' は、以下となる。

$$\begin{aligned}
 t_L &= t_O + t_S + (np - 1)(L_S + L_H + m_B/B) \\
 &\quad - \max(L_S, L_H + m_B/B) \\
 &= t_O + t_S + (np - 1)(L_S + L_H + m_B/B) \\
 &\quad - (L_H + m_B/B)
 \end{aligned} \tag{10}$$

$$\begin{aligned}
 B' &= m_B / \max(L_S, L_H + m_B/B) \\
 &= m_B / (L_H + m_B/B)
 \end{aligned} \tag{11}$$

評価に用いた各パラメータの値を表 3 に示す。 t_O の値は、評価で利用したノード形状を用いて実測したものである。また t_S の値は、メッセージサイズ 8 [byte] の PingPong 性能の実測値を用いた。MPI レイテンシのソフトウェア部分 L_S は、4.2.5 項で述べる 4 [byte] の Allgather の通信時間が近似的に $4L_S(np - 1)$ となることから算出し、MPI レイテンシ 1.6 [μ s] から引いた値をハードウェア部分 L_H とした。通信時間の見積式から算出された通信性能は表 1 のとおりである。通信時間の見積式と測定のフィッティングによるスループットの比較は図 3 のとおりである。点線が見積式による結果である。

ピーク性能 B は、通信バンド幅のピーク性能 5,000 [MB/s] に対して、パイプライン転送にともなうオーバーヘッドの影響を受ける。このため通信時間の見積式でピーク性能の実効値は 4,226 [MB/s] となるのに比べ、測定結果では 3,616 [MB/s] と約 14% 低い。遅延時間 t_L は、通信時間の見積式から求めた遅延時間 1,874 [μ s] と比べて、測定結果での値は 2,795 [μ s] と約 49% の違いがあり、PingPong の遅延時間 10.5 [μ s] と比較すると大きく異なる。両者の通信時間の見積式を比べると、Bcast ではノード数に比例して増加するパイプラインの立ち上げ時間が含まれることが主要因であると考えられる。ノード数が増えることでパイプラインの立ち上げ時間が増加して、半性能長と遅延時間が増加する。

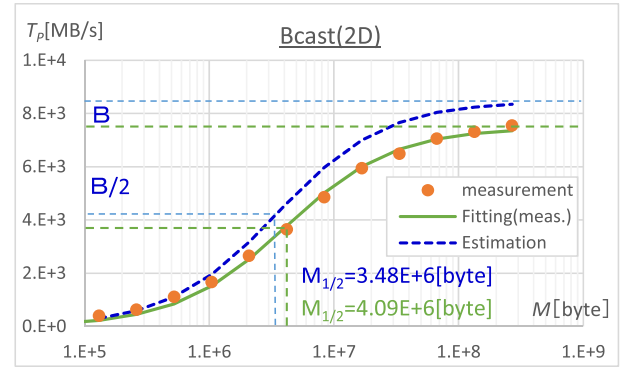


図 4 通信時間の測定結果と見積式から算出したスループットの比較 (Bcast/2D/16 KiB)

Fig. 4 Comparison of throughput calculated from measurement and estimation (Bcast/2D/16 KiB).

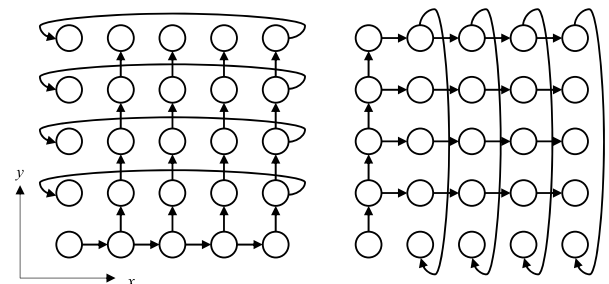


図 5 2 次元での Bcast の転送経路
Fig. 5 Concept of data flow (Bcast/2D).

4.2.2 Bcast (2 次元)

(1) 測定結果のフィッティング

$64 \times 6 = 384$ ノード (2 次元ノード割当て) における Bcast の測定結果について、1 次元と同様にセグメントサイズを 16 [KiB] と固定して、128 [KiB] 以降のデータを用いて、最小二乗法によるフィッティングを行った。表 1 に求められた通信性能を、図 4 にスループットの測定結果とフィッティング結果を示した。丸印が測定結果で実線がフィッティング結果である。

2 次元ノード割当ての場合、データを 2 分割して 2 方向に同時通信を行う。このため、1 次元に比べて、ピーク性能 B は約 2 倍となる。

(2) 見積式からの通信時間の算出

2 次元ではノード割当てが $np_1 \times np_2 (= np)$ となり、1 次元と比べて、データを 2 分割して 2 方向に同時通信をするため、1 方向あたりの通信では、 $M \rightarrow (M/2)$, $B \rightarrow (B/2)$ となる。2 次元での bcast の転送経路を図 5 に示す。左右は 2 分割したメッセージそれぞれの転送経路であり通信経路の重なりはない。最長となる転送経路のステップ数を勘定すると、パイプラインのステップ数は、2 次元ノード割当ての場合、1 次元方向の転送で $np_1 - 1$ 回、2 次元方向の転送で $np_2 - 1$ 回。最後に転送ができていない残りのノードへの転送に 1 回の転送を行う。またパイプラインの MPI

表 4 見積式に用いた評価パラメータ (Bcast/2D/16 KiB)
Table 4 Evaluation parameter (Bcast/2D/16 KiB).

Name	Value
$np1$	64 [node]
$np2$	6 [node]
B	10,000 [MB/s]
m_B	16,384 [byte]
t_O	8.370 [μ s]
t_S	1.396 [μ s]
L_S	1.000 [μ s]
L_H	0.600 [μ s]

レイテンシは、2つの TNI を用いた 2 方向同時通信を行うことから、MPI レイテンシのうちソフトウェアの処理回数が 2 倍になる。式 (9) から通信時間の見積式は以下のとおりである。

$$\begin{aligned}
 t &= t_O + t_S \\
 &+ ((np1 - 1) + (np2 - 1) + 1) \\
 &\times (2L_S + L_H + m_B/(B/2)) \\
 &+ ((M/2)/m_B - 1) \max(2L_S, L_H + m_B/(B/2))
 \end{aligned} \quad (12)$$

今回測定に用いたセグメントサイズの条件 $2L_S < L_H + m_B/(B/2)$ で、メッセージサイズ M に依存しない項ならびに依存する項を抽出すると、遅延時間 t_L ならびに実効のピーク性能 B' は、以下となる。

$$\begin{aligned}
 t_L &= t_O + t_S + (np1 + np2 - 1) \\
 &\times (2L_S + L_H + m_B/(B/2)) \\
 &- \max(2L_S, L_H + m_B/(B/2)) \\
 &= t_O + t_S + (np1 + np2 - 1) \\
 &\times (2L_S + L_H + m_B/(B/2)) \\
 &- (L_H + m_B/(B/2))
 \end{aligned} \quad (13)$$

$$\begin{aligned}
 B' &= 2m_B / \max(2L_S, L_H + m_B/(B/2)) \\
 &= 2m_B / (L_H + m_B/(B/2))
 \end{aligned} \quad (14)$$

なお、Trinaryx3 の場合、パイプラインのステップ数 $((np1 - 1) + (np2 - 1) + 1 = np1 + np2 - 1)$ が、1 次元のステップ数 $(np - 1)$ に比べて小さいため、 $np1 > 1$, $np2 > 1$ の条件で、遅延時間 t_L は 1 次元に比べて短くなる。

$$np1 + np2 - 1 \leq np1 np2 - 1 = np - 1 \quad (15)$$

評価に用いた各パラメータを表 4 に示す。通信時間の見積式から算出した通信性能は表 1 のとおりである、通信時間の見積式と測定のフィッティングによるスループットの比較は図 4 のとおりである。点線が見積式による結果である。

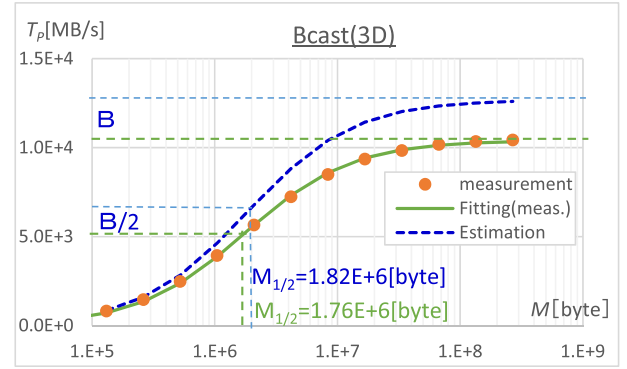


図 6 通信時間の測定結果と見積式から算出したスループットの比較 (Bcast/3D/16 KiB)

Fig. 6 Comparison of throughput calculated from measurement and estimation (Bcast/3D/16 KiB).

ピーク性能 B は、通信バンド幅の 2 方向分のピーク値 10,000 [MB/s] に対して、パイプライン転送にともなうオーバーヘッドの影響を受ける。このため通信時間の見積式でピーク性能の実効値は 8,452 [MB/s] となるのに比べ、測定結果では 7,467 [MB/s] と約 12% 低く 1 次元と同じ傾向である。遅延時間 t_L は、通信時間の見積式から求めた値が 411 [μ s] と比べて、測定結果の値は 548 [μ s] と約 33% 増加している。1 次元の値と比較すると、パイプラインのステップ数が 383 から 69 へと小さくなるため、遅延時間も 1 次元の 2,795 [μ s] から 548 [μ s] へ減少する。

4.2.3 Bcast (3 次元)

(1) 測定結果のフィッティング

$8 \times 6 \times 8 = 384$ ノード (3 次元ノード割当て) における Bcast の測定結果について、1 次元と同様にセグメントサイズを 16 [KiB] と固定して、128 [KiB] 以降のデータを用いて、最小二乗法によるフィッティングを行った。表 1 に求められた通信性能を、図 6 にスループットの測定結果とフィッティング結果を示した。丸印が測定結果で実線がフィッティングした結果である。

3 次元ノード割当ての場合、データを 3 分割して 3 方向に同時通信を行うので、1 次元に比べて、ピーク性能 B は約 3 倍となる。

(2) 見積式からの通信時間の算出

3 次元ではノード割当てが $np1 \times np2 \times np3 (= np)$ となり、データを 3 分割して 3 方向に同時通信する。このため、通信時間の内訳は 1 次元と比べて、1 方向あたりの通信では、 $M \rightarrow (M/3)$, $B \rightarrow (B/3)$ となる。最長となる転送経路のステップ数を勘定すると、パイプラインのステップ数は、2 次元ノード割当ての場合と同様に、1 次元方向の転送で $np1 - 1$ 回、2 次元方向の転送で $np2 - 1$ 回、3 次元方向の転送で $np3 - 1$ 回、最後に転送ができてない残りのノードへの転送に 1 回の転送を行う。またパイプライン転送における MPI の遅延時間は、3 つの TNI を用いた 3 方向同時通信を行うことから、MPI レイテンシの内ソフウェ

表 5 見積式に用いた評価パラメータ (Bcast/3D/16 KiB)
Table 5 Evaluation parameter (Bcast/3D/16 KiB).

Name	Value
$np1$	8 [node]
$np2$	6 [node]
$np3$	8 [node]
B	15,000 [MB/s]
m_B	16,384 [byte]
t_O	8.370 [μ s]
t_S	1.396 [μ s]
L_S	1.000 [μ s]
L_H	0.600 [μ s]

アの処理回数が3倍になる．通信時間の見積式は式 (9) から以下ようになる．

$$t = t_O + t_S + ((np1 - 1) + (np2 - 1) + (np3 - 1) + 1) + (3L_S + L_H + m_B/(B/3)) + ((M/3)/m_B - 1) \times \max(3L_S, L_H + m_B/(B/3)) \quad (16)$$

今回測定に用いたセグメントサイズの条件 $3L_S < L_H + m_B/(B/3)$ で，メッセージサイズ M に依存しない項ならびに依存する項を抽出すると，遅延時間 t_L ならびにピーク性能の実効値 B' は，以下となる．

$$\begin{aligned} t_L &= t_O + t_S \\ &+ (np1 + np2 + np3 - 2)(3L_S + L_H + m_B/(B/3)) \\ &- \max(3L_S, L_H + m_B/(B/3)) \\ &= t_O + t_S \\ &+ (np1 + np2 + np3 - 2)(3L_S + L_H + m_B/(B/3)) \\ &- (L_H + m_B/(B/3)) \end{aligned} \quad (17)$$

$$\begin{aligned} B' &= 3m_B / \max(3L_S, L_H + m_B/(B/3)) \\ &= 3m_B / (L_H + m_B/(B/3)) \end{aligned} \quad (18)$$

なお，Trinaryx3 の場合，パイプラインのステップ数 $((np1-1)+(np2-1)+(np3-1)+1 = np1+np2+np3-2)$ が，1次元のステップ数 $(np-1)$ に比べて小さいため， $np1 > 1$ ， $np2 > 1$ ， $np3 > 1$ の条件で，遅延時間 t_L は1次元に比べて短くなる．

$$\begin{aligned} np1 + np2 + np3 - 2 &\leq np1 np2 np3 - 2 \\ &< np - 1 \end{aligned} \quad (19)$$

評価に用いた各パラメータを表 5 に示す．通信時間の見積式から算出された通信性能は表 1 のとおりである．通信時間の見積式と測定のフィッティングによるスループットの比較は図 6 のとおりである．点線が見積式による結果である．

ピーク性能 B は，通信バンド幅の3方向分のピーク値

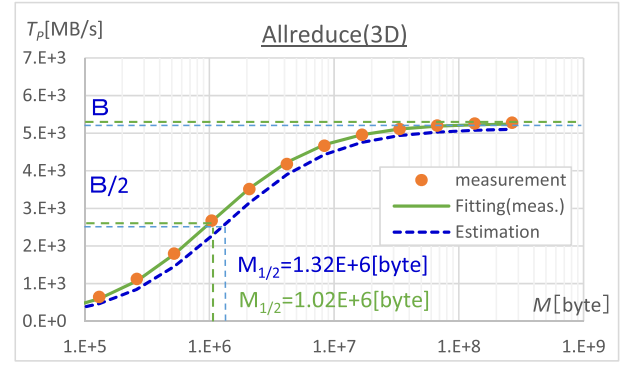


図 7 通信時間の測定結果と見積式から算出したスループットの比較 (Allreduce/3D/16 KiB)

Fig. 7 Comparison of throughput calculated from measurement and estimation (Allreduce/3D/16 KiB).

15,000 [MB/s] に対して，パイプライン転送にともなうオーバーヘッドの影響を受ける．このため通信時間の見積式でピーク性能の実効値は 12,680 [MB/s] となるのに比べ，測定結果では 10,400 [MB/s] と約 18% 低く 1 次元および 2 次元と同様の傾向である．遅延時間 t_L は，通信時間の見積式から求めた値が 143 [μ s] に比べ測定結果からの値は 169 [μ s] と約 18% 増加している．3 次元では 1 次元および 2 次元と比べると，パイプラインのステップ数が 383 (1 次元)，69 (2 次元) から 20 (3 次元) へとさらに減少するため，遅延時間が 2,795 [μ s] (1 次元)，548 [μ s] (2 次元) から 169 [μ s] (3 次元) へ減少した．

4.2.4 Allreduce (3 次元)

(1) 測定結果のフィッティング

Allreduce について Trinaryx3 の効果がある 3 次元ノード割当てについて評価を行った． $8 \times 6 \times 8 = 384$ ノード (3 次元ノード割当て) において，Bcast と同様にパイプライン転送のセグメントサイズを 16 [KiB] と固定として Allreduce の測定を行い，128 [KiB] 以降のデータを用いて，最小二乗法によるフィッティングを行った．Allreduce は内部で Reduce 処理と Bcast 処理を行うため，スループットの算出には，実際に通信されるメッセージサイズ M が 2 倍であることを用いた．表 1 に求められた通信性能を，図 7 にスループットの測定結果とフィッティング結果を示す．丸印が測定結果で実線がフィッティング結果である．

(2) 見積式からの通信時間の算出

通信時間 t の内訳は，Bcast 処理の時間 t_B ，Reduce 処理の時間 t_R であるとする． t_B は，式 (16) の Bcast (3 次元) の見積式で記述できる． t_R は，Bcast 処理の t_B の見積式に，Reduce 処理の演算スループット T_C (= 演算データ容量/処理時間) を考慮することで算出可能である．通信時間の式を以下に示す．通信領域のアドレス交換の時間 t_S は各通信段階で必要であり，1 要素の allreduce などによるオーバーヘッド t_O は，Reduce 処理の前に 1 回だけ必要である．なお，セグメントサイズを，Bcast 処理で m_B ，

Reduce 処理用で m_R とする.

$$t = t_B + t_R \quad (20)$$

$$\begin{aligned} t_R = & t_O + t_S + ((np1 - 1) + (np2 - 1) + (np3 - 1) + 1) \\ & \times (3L_S + L_H + m_R/(B/3) + m_R/(T_C/3)) \\ & + ((M/3)/m_R - 1) \\ & \times \max(3L_S + m_R/(T_C/3), L_H + m_R/(B/3)) \end{aligned} \quad (21)$$

$$\begin{aligned} t_B = & t_S + ((np1 - 1) + (np2 - 1) + (np3 - 1) + 1) \\ & \times (3L_S + L_H + m_B/(B/3)) + ((M/3)/m_B - 1) \\ & \times \max(3L_S, L_H + m_B/(B/3)) \end{aligned} \quad (22)$$

今回測定に用いたセグメントサイズの条件 $3L_S < L_H + m_B/(B/3) < 3L_S + m_R/(T_C/3)$ で, これらの式からメッセージサイズ M に依存しない項を抽出すると, 遅延時間 t_L は, 以下となる.

$$\begin{aligned} t_L = & t_O + 2t_S \\ & + (np1 + np2 + np3 - 2)(3L_S + L_H + m_B/(B/3)) \\ & + (np1 + np2 + np3 - 2) \\ & \times (3L_S + L_H + m_R/(B/3) + m_R/(T_C/3)) \\ & - \max(3L_S, L_H + m_B/(B/3)) \\ & - \max(3L_S + m_R/(T_C/3), L_H + m_R/(B/3)) \\ = & t_O + 2t_S \\ & + (np1 + np2 + np3 - 2)(3L_S + L_H + m_B/(B/3)) \\ & + (np1 + np2 + np3 - 2) \\ & \times (3L_S + L_H + m_R/(B/3) + m_R/(T_C/3)) \\ & - (L_H + m_B/(B/3) + 3L_S + m_R/(T_C/3)) \end{aligned} \quad (23)$$

一方, Reduce と Bcast 処理によりデータ転送量は 2 倍となるため, 実効のピーク性能 B' は以下のように書き下せる.

$$t = t_L + 2M/B' \quad (24)$$

$$\begin{aligned} B' = & 2/(\max(L_S/m_R + 1/T_C, L_H/(3m_R) + 1/B) \\ & + \max(L_S/m_B, L_H/(3m_B) + 1/B)) \\ = & 2/(L_S/m_R + 1/T_C + L_H/(3m_B) + 1/B) \end{aligned} \quad (25)$$

よってスループット T_p は以下となる.

$$\begin{aligned} T_p = & 2M/t = 2M/(t_L + 2M/B') \\ = & B'/(1 + (B't_L)/2M) \end{aligned} \quad (26)$$

スループットは式 (3) で記述されるため, $M_{1/2}$ は B' を用いて,

$$M_{1/2} = B't_L/2. \quad (27)$$

ここで, 演算スループット T_C は「京」においては, メモ

表 6 見積りに用いた評価パラメータ (Allreduce/3D/16 KiB)

Table 6 Evaluation parameter (Allreduce/3D/16 KiB).

Name	Value
$np1$	8 [node]
$np2$	6 [node]
$np3$	8 [node]
B	15,000 [MB/s]
m_B	16,384 [byte]
m_R	16,384 [byte]
T_C	4,000 [MB/s]
t_O	8.370 [μ s]
t_S	1.396 [μ s]
L_S	1.000 [μ s]
L_H	0.600 [μ s]

リバンド幅 B_M (実効値は 46,000 [MB/s]) から通信処理で利用する分を除いて算出され以下の式と値となる.

$$T_C = (B_M - 2B)/4 = 4,000 [\text{MB/s}] \quad (28)$$

ここで係数 2 は受信と送信による効果であり, 係数 4 は変数バッファの Load/Store 数である. 評価に用いた各パラメータを表 6 に示す. 通信時間の見積りから算出した通信性能は表 1 のとおりである, 通信時間の見積りと測定結果のフィッティングによるスループットの比較は図 7 のとおりである. 点線が見積りによる結果である.

ピーク性能 B は, 通信バンド幅の 3 方向分のピーク値 15,000 [MB/s] に対して, Reduce 処理と Bcast 処理の両方を実行するため, 通信時間の見積りで実効 5,129 [MB/s] となる. 測定結果は 5,266 [MB/s] であり, 見積りに比べ約 2.7% の違いがある. 遅延時間 t_L は, 見積りから求めた値が 513 [μ s] に比べ, 測定結果の値は 389 [μ s] と約 25% の違いがある.

4.2.5 Allgather (3 次元)

(1) 測定結果のフィッティング

Allgather 通信について, Tofu 専用アルゴリズム 3D-multiring [5] が利用可能である. 3D-multiring は, 3 次元連続な直方体のコミュニケータでのみ適用可能なため, 本評価は, 3 次元ノード割当てについて評価した. $8 \times 6 \times 8 = 384$ ノード (3 次元ノード割当て) において, Allgather の測定を行い, 最小二乗法によるフィッティングを行った. Allgather は各ランクが持つデータを全ランクへ転送するため, スループットの算出には, 実際のメッセージサイズがノード数倍であることを用いた. また通信サイズにより通信時間の見積りが変わることが想定されるため, 32 [KiB] までと 64 [KiB] 以降の 2 つの領域についてそれぞれフィッティングを行った. 表 1 に求められた通信性能を, 図 8 にスループットの測定結果とフィッティング結果のグラフを示す. 丸印が測定結果で実線がフィッティング結果である. 破線は, フィッティング領域の境界である.

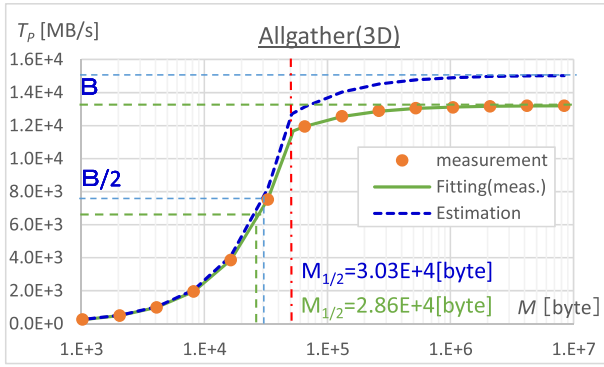


図 8 通信時間の測定結果と見積式から算出したスループットの比較 (Allgather/3D)

Fig. 8 Comparison of throughput calculated from measurement and estimation (Allgather/3D).

(2) 見積式からの通信時間の算出

3D-multirring は、データを 4 分割して 4TNI を使った 4 方向の同時通信を行い、3 次元方向を 3 段階に分けてリング状に転送する方式である [5]。まず 1 段階目で 3 次元の各軸方向に転送して、2 段階目で 1 方向分のデータを次の軸方向に転送し、3 段階目で 2 方向分のデータを次の軸方向に転送する。1 方向あたりの通信では、メッセージサイズが $M/4$ 、通信バンド幅が $B/4$ となり、データ転送の時間は $(M/4)/(B/4) = M/B$ となる。また 3D-multirring でのパイプラインの MPI 遅延時間は、4 つの TNI を用いた 4 方向同時通信を行うことから、MPI レイテンシの内ソフトウェアの処理回数が 4 倍になる。各段階における転送方向のノード数を $np1, np2, np3$ (総ノード数 $np = np1 \times np2 \times np3$) とおくと、通信時間の見積式は以下となる。

$$\begin{aligned} t &= t_O + t_S + (4L_S + L_H + M/B) \\ &\quad + \max(4L_S, L_H + M/B)(np1 - 1 - 1) \\ &\quad + (4L_S + L_H + M/B) \\ &\quad + \max(4L_S, L_H + M/B)(np1 \times np2 - np1 - 1) \\ &\quad + (4L_S + L_H + M/B) \\ &\quad + \max(4L_S, L_H + M/B)(np - np1 \times np2 - 1) \\ &= t_O + t_S + 3(4L_S + L_H + M/B) \\ &\quad + \max(4L_S, L_H + M/B)(np - 4) \end{aligned} \quad (29)$$

この式をメッセージサイズ M に依存する項と依存しない項に分けると実効のピーク性能 B' ならびに遅延時間 t_L は以下となる。

1) $4L_S > L_H + M/B$ ($M < B(4L_S - L_H)$): 短メッセージ時

$$t_L = t_O + t_S + 3(4L_S + L_H) + 4L_S(np - 4) \quad (30)$$

$$B' = B np / 3 \quad (31)$$

2) $4L_S \leq L_H + M/B$ ($M \geq B(4L_S - L_H)$): 長メッセー

表 7 見積式に用いた評価パラメータ (Allgather/3D)

Table 7 Evaluation parameter (Allgather/3D/16 KiB).

Name	Value
$np1$	8 [node]
$np2$	6 [node]
$np3$	8 [node]
B	15,000 [MB/s]
t_O	8.370 [μ s]
t_S	1.396 [μ s]
L_S	1.000 [μ s]
L_H	0.600 [μ s]

ジ時

$$t_L = t_O + t_S + 3(4L_S + L_H) + L_H(np - 4) \quad (32)$$

$$B' = B np / (np - 1) \quad (33)$$

ただし、実効のピーク性能 B' の導出には、実際に送受信されるメッセージサイズが np 倍であることを用いて算出した。ピーク性能となる最大スループット $B_{\max}$ は長メッセージ領域の見積式で記述され、以下となる。

$$B_{\max} = B np / (np - 1) \quad (34)$$

「京」における 4TNI 分の総通信バンド幅は最大 15,000 [MB/s] である [5]。評価に用いた各パラメータを表 7 に示す。また各メッセージサイズ領域で求めた通信性能は表 1 のとおりである、通信時間の見積式と測定結果のフィッティングによるスループットの比較は図 8 のとおりである。点線が見積式による結果である。

スループット T_p はメッセージサイズ M により 2 種類の評価式に従う。評価パラメータからこのメッセージサイズを見積もると約 49.8 [KiB] となり、32 [KiB] と 64 [KiB] の間で評価式が切り替わることが分かる。

ピーク性能 B は、短メッセージ領域での実測値と見積式での値とに大きな差があるが、 B は長メッセージの極限で達成される外挿値であるため、短メッセージ領域の評価に大きな問題はない。長メッセージ領域での通信時間の見積式で算出された値は、15,040 [MB/s] であるのに対して、測定結果では 13,210 [MB/s] であり、見積式に比べ約 12% の違いである。遅延時間は、見積式から求めた値は、各メッセージ長領域で 1,544 [μ s] と 252 [μ s] であるのに対し、測定結果での値は 1,590 [μ s] と 201 [μ s] で約 12%~20% の違いである。

4.3 通信性能のまとめ

今回は、「京」での MPI 通信性能のスループットの実測値について、通信モデルから通信時間の見積式を理論的に導出し、ピーク性能と遅延時間を用いた評価を行った。

PingPong と 1 次元の Bcast では、実効のピーク性能 B'

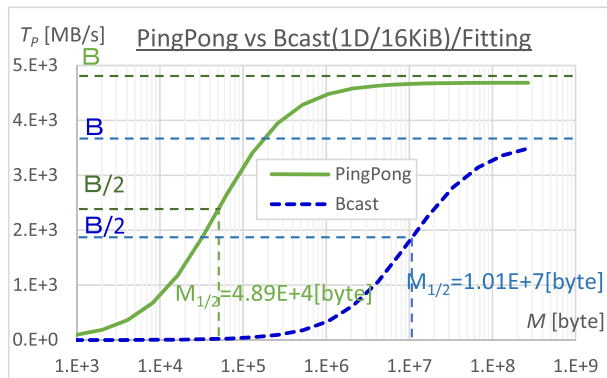


図 9 PingPong と Bcast (1D/16 KiB) のスループット

Fig. 9 Throughput of PingPong and Bcast (1D/16 KiB).

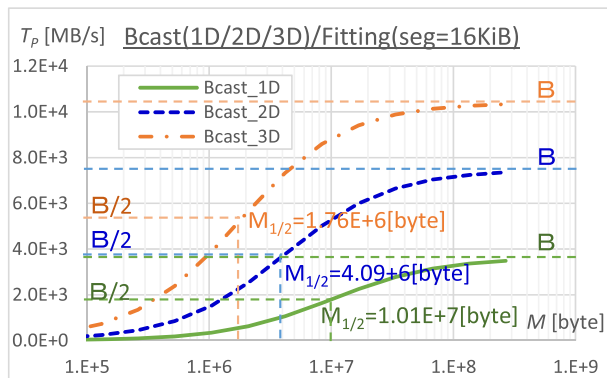


図 10 Bcast (1D, 2D, 3D/16 KiB) のスループット

Fig. 10 Throughput of Bcast (1D, 2D, 3D/16 KiB).

について、パイプライン転送のオーバーヘッドなどにより、Bcast が 23%ほど低く、Allreduce でも同様の傾向であった。また遅延時間 t_L についてはパイプライン転送の立ち上げ時間が余計にかかるため Bcast の方が大きくなる。図 9 に PingPong と Bcast の測定結果から得られたスループットの比較を示す。実線が PingPong で点線が Bcast である。

図 10 は、Bcast について 1 次元、2 次元、3 次元での測定結果のフィッティングしたスループットを比較したものである。実線が 1 次元、点線が 2 次元、破線が 3 次元である。ジョブ形状の次元が上がると多方向の同時通信が可能となり、ピーク性能 B が 1 次元に比べて 2 次元で約 2 倍、3 次元で約 3 倍となる。また、次元が上がると、パイプラインのステップ数が小さくなるため立ち上げに必要な時間が小さくなり、遅延時間 t_L が小さくなる。つまりジョブのノード割当てを、より高次元で使用した方が、通信性能が高いことが分かる。しかし実際には、短いメッセージでの通信において、通信時間の最適化のためセグメントの長さが変化し、かつ分割のレイテンシも増大するため、より汎用的な Open MPI アルゴリズムが選択されることが想定される。このため通信性能に差異がなくなるか、逆転が起こりうる可能性もある。

5. 展開と今後の課題

今回評価した MPI 集団通信の実測値と通信時間の見積りによるスループットの傾向について、ピーク性能で 20%程度、遅延時間で 50%までの精度で説明することができた。特に Tofu インタコネクに最適化されたアルゴリズムが、想定される性能をある程度達成できていることを確認することができたといえる。これら性能の差異の原因として、ハードウェアに起因するパラメータの見積りが不十分である可能性や、通信ライブラリの実装方法による性能の変化などが可能性として考えられる。

ピーク性能 B の誤差の原因としては、ハードウェアの限界性能によるもの、双方向通信の効果によるものなど、大きく 2 点があげられる。1 次元の Bcast で、パイプライン転送のオーバーヘッドを除いた実効のピーク性能を逆算すると 4,167 [MB/s] であり、PingPong のピーク性能 4,684 [MB/s] と比べて 11%程度低く実測されている。PingPong のピーク性能は、ハードウェアの限界性能に近い性能であると想定されるため、この差異は、データを分割して双方向の同時通信している影響 [5] と推察され、Tofu 専用アルゴリズムを用いる集団通信全般に共通な性能といえる。

遅延時間 t_L の誤差の原因としては、MPI レイテンシ L_S, L_H の見積りにおける過小評価。ピーク性能 B の見積りにおける過大評価の 2 点があげられる。MPI レイテンシの過小評価としては、Tofu 専用アルゴリズムの実装において、転送データが L_2 に乗っている場合とそうでない場合で、レイテンシに影響が考えられるためその効果を積み上げることで評価精度が向上することが期待できる。ピーク性能 B が遅延時間 t_L に与える影響については、パイプライン転送が行われる集団通信では、遅延時間の値にパイプライン転送の準備段階の通信時間が含まれ、ピーク性能の大小によってその時間は変わる。実際この立ち上げの転送時間は、実効のバンド幅で通信が行われるため、理論ピーク性能を用いた見積りでは、遅延時間が短めに算出されることになる。実測されたピーク性能を用いて遅延時間を再見積りすると、1 次元の Bcast の t_L は 1,874 [μ s] $\rightarrow$ 2,124 [μ s] となり、その誤差は 49.2% $\rightarrow$ 31.6%へと小さくなる。

今後、さらなる検証を行うことで、通信性能向上に向けた課題を見だし、通信アルゴリズムやライブラリ開発などへフィードバック可能であると期待される。また、今回報告した以外の集団通信や他の通信アルゴリズムの性能についても、Bcast, Allreduce, Allgather の見積りを拡張して評価することで、より多くの知見を得ることができると考えられる。

謝辞 本報告に際し、理化学研究所計算科学研究機構運用技術部門、富士通株式会社 SE の諸氏に感謝します。本稿の結果は、理化学研究所計算科学研究機構が保有するスーパーコンピュータ「京」によるものです。

参考文献

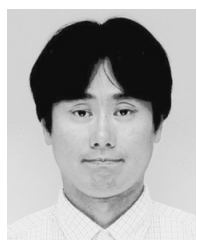
- [1] 北澤好人, 黒田明義, 南 一生, 庄司文由: スーパーコンピュータ「京」における MPI 通信性能の評価, 研究報告ハイパフォーマンスコンピューティング (HPC), 2015-HPC-153(34), pp.1-7 (2016).
- [2] Intel® MPI Benchmarks User Guide and Methodology Description, available from (<https://software.intel.com/en-us/articles/intel-mpi-benchmarks>).
- [3] Toyoshima, T.: ICC: An interconnect controller for the Tofu interconnect architecture, *Hot Chips 22* (2010).
- [4] 安島雄一郎, 井上智宏, 平本新哉, 清水俊幸: スーパーコンピュータ「京」のインターコネクト Tofu, *FUJITSU.63,3*, pp.260-264 (2012).
入手先 (<http://www.fujitsu.com/downloads/JP/archive/imgjp/jmag/vol63-3/paper05.pdf>).
- [5] Adachi, T., Shida, N., Miura, K., Sumimoto, S., Uno, A., Kurokawa, M., Shoji, F. and Yokokawa, M.: The Design of Ultra Scalable MPI Collective Communication on the K Computer, *2013 Comput. Sci.*, Vol.28, No.2-3, pp.147-155 (May 2013).
- [6] Ajima, Y., Takagi, Y., Inoue, T., Hiramoto, S. and Shimizu, T.: The Tofu interconnect, *Proc. HotI 2011*, pp.87-94 (Aug. 2011).
- [7] 松本 幸, 安達知也, 住元真司, 南里豪志, 曾我武史, 宇野篤也, 黒川原佳, 庄司文由, 横川三津夫: MPI_Allreduce の「京」上での実装と評価, 情報処理学会誌, コンピューティングシステム, Vol.5, No.5, pp.152-162 (2012).
- [8] Hockney, R.W.: *The Science of Computer Benchmarking*, ISBN-13: 978-0898713633, section 3.3, pp.42-47 (1996).
- [9] 志田直之, 住元真司, 宇野篤也: スーパーコンピュータ「京」の MPI と低レベル通信, *Fujitsu*, Vol.63, No.3, pp.299-304 (2012).



北澤 好人

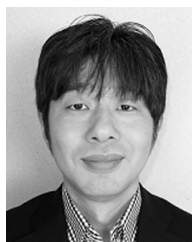
1990 年信州大学理学部物理学科卒業。2009 年から (株) 富士通長野システムエンジニアリング (現, 富士通株式会社) で, 「京」コンピュータのソフトウェア高度化に従事。2014 年から理化学研究所計算科学研究機構に出向。

2017 年から富士通株式会社で「京」コンピュータのソフトウェア高度化に従事。



黒田 明義

1998 年京都大学大学院人間・環境学研究科博士後期課程修了。専門は統計力学, 計算物理学。2006 年から理化学研究所次世代スーパーコンピュータ開発実施本部ならびに計算科学計算機構で, アプリケーション開発の立場から「京」コンピュータの開発ならびにソフトウェアの高度化に従事。博士 (人間・環境学)。



志田 直之

1992 年徳島大学工学部知能情報工学科を卒業し, (株) 富士通静岡エンジニアリングに入社。2003 年に富士通株式会社へ転籍。2007 年からスーパーコンピュータ「京」の通信ライブラリ

の開発に従事。現在, フラッグシップ 2020 プロジェクトにおいてポスト「京」の通信ライブラリ



安達 知也 (正会員)

2010 年東京大学大学院情報理工学系研究科博士前期課程修了。富士通株式会社でスーパーコンピュータ「京」およびフラッグシップ 2020 プロジェクトにおいてポスト「京」の通信ソフトウェアの開発に従事。



南 一生 (正会員)

1981 年日本大学理工学部物理学科卒業。富士通株式会社入社。2000 年財団法人高度情報科学技術研究機構入社, 地球シミュレータ用ソフトウェア性能最適化研究に従事。2008 年理化学研究所次世代スーパーコンピュータ

開発実施本部アプリケーション開発チームリーダー。2017 年理化学研究所計算科学研究機構運用技術部門チューニング技術チームヘッド。2011 年ゴードンベル賞受賞。博士 (工学)。