ハイパフォーマンス
コンピューティング 48−12
（1993. 8. 20）

メッセージ交換型マルチコンピュータの通信性能
— AP1000に関する通信性能モデリングおよび性能評価 —

國貞 勝弘　　村上和彰

九州大学 大学院総合理工学研究科 情報システム学専攻

〒816 福岡県春日市春日公園 6-1

E-mail: {kunisada, murakami}@is.kyushu-u.ac.jp

並列コンピュータの性能を有効に引き出すためには，モデリングや，シミュレーションなどによる性能の評価，解析を欠かすことはできない．そのような性能評価やモデリングは現在までさまざまな対象，手法について行われてきた．本稿では，並列コンピュータに関して過去にどのような評価がなされてきたかということについていくつか例を挙げ，またメッセージ交換型マルチコンピュータの通信性能の解析のためのモデリングの一例を挙げる．実際のメッセージ交換型マルチコンピュータ AP1000 において，ホットスポットがどのような影響を与えるかを実測し，同ホップ数ランダム通信においてホップ数の変化が通信時間にどのような影響を与えるかを実測した．

Communications Performance of Message Passing Multicomputer
— Performance Modeling and Evaluation on AP1000 —

Katsuhiro Kunisada　Kazuaki Murakami

Department of Information Systems
Interdisciplinary Graduate School of Engineering Sciences
Kyushu University
Kasuga-shi, Fukuoka 816 Japan

E-mail: {kunisada, murakami}@is.kyushu-u.ac.jp

In this paper, we give some examples of past evaluation method on a parallel computer. And we present an example of modeling to analize communications performance of a message passing multicomputer. We make a survey of an influence of Hot Spot on AP1000, which has a 2D torus network with wormhole routing. We also make a survey of an influence of changing the number of hop on performance when the same-hop-number-communication is used.

1 はじめに

並列マシンの性能は，その通信性能に大きく依存する．並列マシンは，そのプロセッサ間通信モデルとして共有メモリ・モデルとメッセージ交換モデルのいずれを採用するかで，共有メモリ型マルチプロセッサ（*shared-memory multiprocessor*）とメッセージ交換型マルチコンピュータ（*message-passing multicomputer*）に2分される．本稿では，メッセージ交換型マルチコンピュータを対象とする．

1980年代に多数登場したメッセージ交換型マルチコンピュータの大部分は，その相互結合網のトポロジーとしてハイパーキューブ（直接2進nキューブ網）を採用していた[8]．これは専ら，その直径の小ささ（ノード数Nに対して，直径は$\log_2 N$）に理由があった．ところが，最近では，メッシュを採用したメッセージ交換型マルチコンピュータが増えてきている．これは，以下の理由による[3]．

- 相互結合網の次数（ノード当りのリンク数）がノード数Nに対して，2次元メッシュでは4と定数で済むが，ハイパーキューブの場合は$\log_2 N$となる．すなわち，ハイパーキューブはマシン・サイズに関してスケーラブルではない．

- ハイパーキューブおよびメッシュはいずれも，k-ary n-cube網の特殊なケースに相当する．すなわち，ノード数N，基数k，および，次元nが以下のような関係にある．

  - k-ary n-cube：$N = k^n$，
    $k = \sqrt[n]{N}, n = \log_k N$
  - ハイパーキューブ（binary n-cube）：
    $N = k^n = 2^n$，
    $k = 2, n = \log_k N = \log_2 N$
  - 2次元メッシュ：$N = k^n = k^2$，
    $k = \sqrt[n]{N} = \sqrt{N}, n = 2$

  ワームホール・ルーティングを用いたk-ary n-cubeの性能評価の結果，nの最適値は2ないし3であると報告されている[13]．すなわち，メッシュの直径は$\sqrt{N}$と大きいものの，ワームホール・ルーティングやバーチャル・カットスルーといった新しいフロー制御方式の登場により，直径の大きさによるネットワーク・レイテンシが低減可能となった．これにより，直径が小さいというハイパーキューブの利点が相対的に減ってしまった．

- アプリケーション，特に科学技術計算分野においては，その通信トポロジー（通信トラフィック・パターン）自身がメッシュであることが多い．

相互結合網のトポロジーとしてメッシュを採用した商用のメッセージ交換型マルチコンピュータには，AP1000，Paragon XP/S，T3D，等がある．AP1000[1, 4, 5]は，最小16ノード～最大1024ノード構成で，トーラスネット（T-Net），ブロードキャスト・ネット（B-Net），および，同期ネット（S-Net）の3種類の独立したネットワークを備える．一般的な意味でのプロセッサ間通信に使用する相互結合網は，T-Netである．T-Netは2次元トーラスのトポロジを持ち，フロー制御方式として構造化バッファ・プール＋ワームホール・ルーティング方式を，また，ルーティング・アルゴリズムとして静的X-Yルーティング・アルゴリズムを採用している．本稿では，64（8×8）ノード構成のAP1000を用いて，そのT-Netの通信性能を実測により評価する．

まず，2章で通信性能の評価方法を整理する．4章および5章で，ホットスポット・トラフィックが非ホットスポット・トラフィックに与える影響，および，同一ホップ数ランダム1対1通信トラフィックの性能を評価する．最後に，6章で本稿のまとめとする．

2 通信性能の評価項目および評価尺度

通信性能の評価項目および評価尺度は多岐にわたるが，主として以下のものが用いられている[2]．

1. 評価項目：

   (a) アーキテクチャ関連：

   - 相互結合網のトポロジー
   - マシン・サイズ：ノード数，次元数，等
   - 通信方式：ルーティング・アルゴリズム，フロー制御方式，単方向／両方向チャネル，等
   - ノードの能力：バッファ・サイズ，許容できる未完了メッセージ数，送受信遅延時間，中継遅延時間，等
   - 相互結合網の能力：転送バンド巾（転送スループット），転送レイテンシ，等

   (b) メッセージ関連：

   - メッセージ・サイズ
   - メッセージ発生間隔，発生率
   - 通信距離，ホップ数
   - 通信トラフィック・パターン，相互結合網の負荷

2. 評価尺度（評価指標）：

   (a) 時間（time）：

   - プログラム総実行時間
   - ネットワーク・レイテンシ（メッセージの相互結合網内滞在時間）
   - 通信レイテンシ，プロセッサ待ち時間，応答時間

   (b) 速度（rate）：

   - 実効転送バンド巾（転送スループット）

   (c) 比率（ratio）：

   - プロセッサ使用効率
   - チャネル使用効率

いくつかの文献で採用されている評価項目と評価尺度との組合せを表2にまとめる．

本稿では，以下の評価項目および評価尺度を用いてAP1000の（トーラスネットT-Netの）通信性能を実測により評価する．

1. 評価目的：ホットスポット・トラフィックが非ホットスポット・トラフィックの通信性能に与える影響の調査

   (a) 評価項目：

   - メッセージ・サイズ
   - ホットスポットに対してメッセージを送信するノード数
   - ホットスポット以外のノードに向かうメッセージのとる経路とホットスポットとの間の距離

   (b) 評価尺度：

   - 平均ネットワーク・レイテンシ

2. 評価目的：同一通信距離（ホップ数）で送信先がランダムな1対1通信トラフィックにおける通信性能の調査

   (a) 評価項目：

   - メッセージ・サイズ
   - 通信距離（ホップ数）

   (b) 評価尺度：

   - 平均ネットワーク・レイテンシ

3 性能モデリング

AP1000の2次元トーラス網T-Netを用いてノード間通信を行なう場合の通信レイテンシを定式化する[6]．

まず，1個のノードiがX方向（またはY方向）にh個離れたノードに長さmフリットのメッセージを送る場合の通信レイテンシ$T^i(1,h,m)$は，下式で表せる．

$$T^i(1,h,m) = T_{send} + hT_u + (h-1)T_{hop} + (m-1)(T_{hop}+T_u) + T_{receive}$$

ここで，

- T_{send}：送信オーバーヘッド，すなわち，送信ノードが送信要求を出してから最初のフリットが相互結合網に送り出されるまでの時間，
- T_{hop}：中継オーバーヘッド，

- $T_{receive}$：受信オーバーヘッド，すなわち，受信ノードが最後のフリットを受信してから，当該メッセージの処理にかかるまでの時間，および，
- T_u：1個のフリットが1つのリンクを通過するのに要する時間，

である．

次に，同一X軸（またはY軸）上の全k個のノードがX方向（またはY方向）にh個離れたノードに長さmフリットのメッセージを送る場合のある送信ノードiの通信レイテンシ$T^i(k,h,m)$は，下式で表せる．

$$T^i(k,h,m) = T_{send} + \sum_{j \in L_x(i,h)} m_j T_u + (h-1)T_{hop} + (m-1)(\max_{j \in L_x(i,h)} m_j + T_{hop}) + T_{receive}$$

ここで，

- $L_x(i,h)$：ノードiからX方向にh個離れたノードに至るまでのリンクjを要素とする集合，および，
- m_j：リンクjにおけるメッセージの重なり度，

である．

文献[7]では，バーチャル・チャネル＋ワームホール・ルーティングを用いた2次元トーラスにおける平均通信レイテンシをキューイング・ネットワークを用いて定式化している．

4 実験1: ホットスポット・トラフィックの影響

ホットスポット競合（hot spot contention）とは図1に示すようにある特定のノードにトラフィックが集中する現象を言う．当該ノードをホットスポットと，当該トラフィックをホットスポット・トラフィックと呼ぶ．また，ホットスポットを送信先としない通信トラフィックを非ホットスポット・トラフィックと呼ぶ．ホットスポットが存在する場合，多段結合網などの間接網ではツリー・サチュレーションが生じて非ホットスポット・トラフィックを含めた通信性能全体に悪影響を及ぼすことが知られている[16]．

そこで，本稿では，2次元トーラス網におけるホットスポット・トラフィックの非ホットスポット・トラフィックの通信性能に与える影響を調査することにした．具体的には，

- ホットスポットに対してメッセージを送信するノード数（ホットスポット・トラフィック・サイズ）の大小によって，当該ホットスポット上を通過する非ホットスポット・トラフィックの通信レイテンシがどのように影響されるか？
- ホットスポット以外のノードに向かうメッセージ（非ホットスポット・トラフィック）のとる経路とホットスポットとの間の距離の違いが，当該メッセージの通信レイテンシにどのような影響を与えるか？

をメッセージ・サイズを変化させながら測定した．

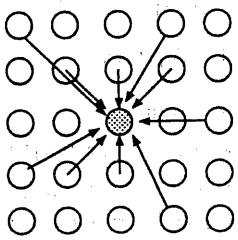

図1: ホットスポット

最初の実験は以下のように行なった．

1. ホットスポットとなるノードを1個選ぶ．
2. ホットスポット上を通過する通信（測定対象）を行なう送受信ノードを1対選ぶ．
3. 上記以外のノードの中から，ホットスポットに対して送信を行なうノードをランダムに一定個数（20，40，61）選ぶ．
4. 以上で定まったホットスポット・トラフィックと非ホットスポット・トラフィックを同時に実行し，測定対象のネットワーク・レイテンシを測定する．

5. ノード選択をランダムに変えて上記測定を一定回数繰り返し，測定結果の平均をとる．

また，2番目の実験では，図2に示すように，ホットスポットを原点とした場合のX軸あるいはY軸を非ホットスポット・トラフィックの経路が横切る際の座標値を0，2，4と変化させて測定を行なった．

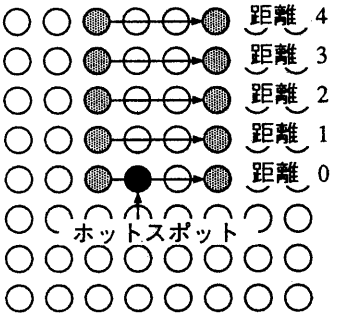

図 2: 性能の測定方法

1番目の実験結果を図3に，2番目の実験結果を図4にそれぞれ示す．

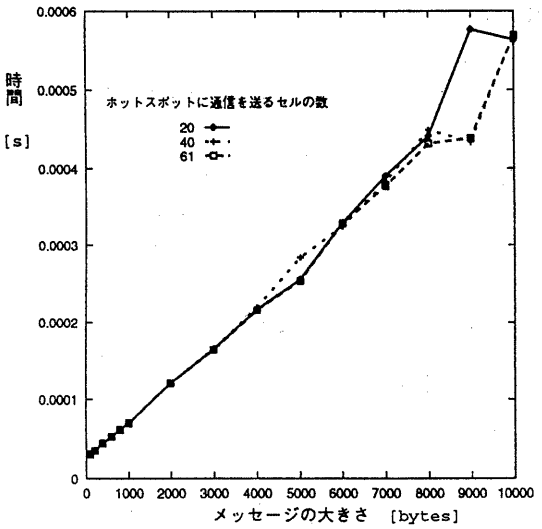

図 3: ホットスポット・トラフィック・サイズの影響

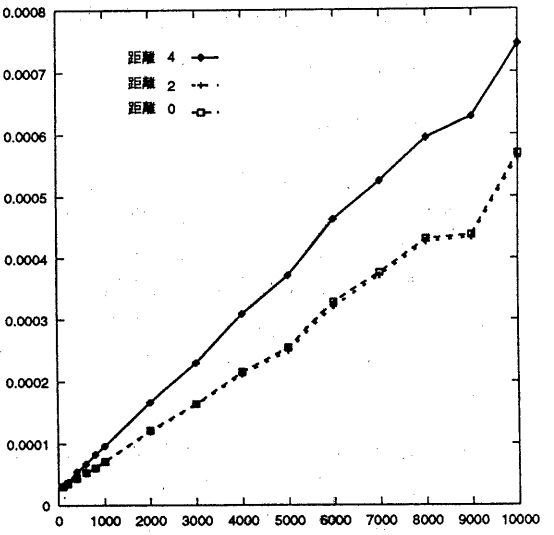

図 4: ホットスポットからの距離による影響

図3から，メッセージの大きさが9000バイトの部分で値のばらつきがあるが，ほぼグラフは直線であることがわかる．したがって，64ノードのAP1000では，ホットスポットに対して送信を行なうノード数（ホットスポット・トラフィック・サイズ）の違いによる性能への影響は見受けられない．なお，文献[7]では，バーチャルチャネル＋ワームホール・ルーティングを用いた2次元トーラス網におけるホットスポットの影響をシミュレートしているが，そこではホットスポット・トラフィック・サイズの影響が若干認められる．

図4から，距離0の場合（非ホットスポット・トラフィックの経路がホットスポット上を通過する場合）の方が距離4の場合（非ホットスポット・トラフィックの経路がホットスポットから最も遠いところを通る場合）よりも平均ネットワーク・レイテンシが小さいことがわかる．

今回はホットスポット・トラフィック自体の通信性能は評価しなかったが，非ホットスポット・トラフィック以上に通信性能が悪くなるものと予想される．

5 実験 2: 同一ホップ数ランダム 1 対 1 通信トラフィックの性能

図 5に示すように通信距離 (ホップ数) が同じで通信先がランダムな 1 対 1 通信を全ノードが同時に行う場合の性能を測定する. この実験により, 通信経路の重なり具合が通信性能に与える影響を識別できるものと期待される.

具体的には, 次のような方法で測定を行なった.

1. 予めホップ数が決められた数になるように, 通信を行うノードの組をノードが重ならないようにランダムに選び出す.

2. 通信するメッセージの大きさを変えながら, 通信, 測定を繰り返す.

これらをホップ数を変えながら繰り返し, 通信時間の測定, 平均を行なう.

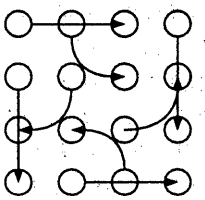

図 5: 同一ホップ数ランダム 1 対 1 通信トラフィック

メッセージ・サイズが 10000 バイトの時の実験結果を図 6に示す.

同ホップ数で通信のソースとデスティネーションが異なるので, 必ず通信経路の重なる部分がでてくる. ホップ数が増えるとこの重なりが増えるために通信時間が遅くなると考えられる.

図 7のようにして, ノードにおいて通信経路の重なる数を数え平均をとると, 表 1 のようになった.

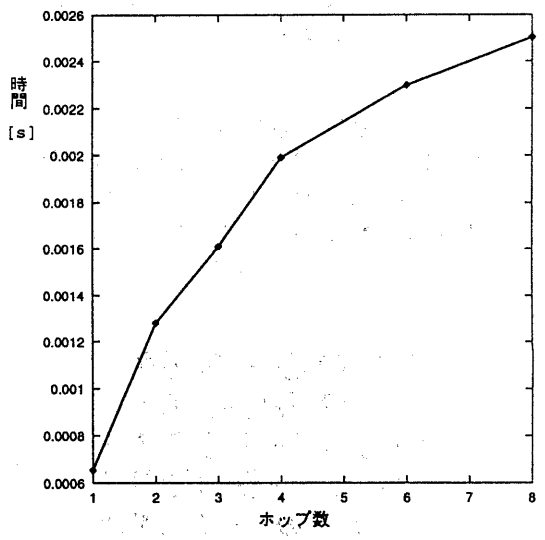

図 6: 同一ホップ数ランダム 1 対 1 通信トラフィックの性能

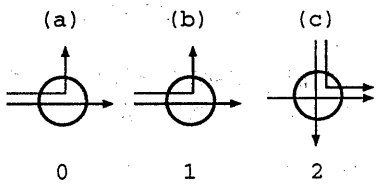

図 7: ノードでの通信経路の重なる数の例

表 1において, ノードでの通信経路の重なる数の平方根を定数倍し, 図 6と重ね合わせ

表 1: 重なる数

ホップ数	重なる数	平方根	通信時間
1	0	0	654μs
2	32	5.65	1280
3	64	8	1610
4	95	9.74	1991
6	162	12.72	2301
8	224	14.96	2503

ると，図8のようになった．

この図から，全てのノードを通信に用いた同一ホップ数ランダム1対1通信では，通信時間は，ノードでの通信経路の重なる数の平方根に比例することが予想される．

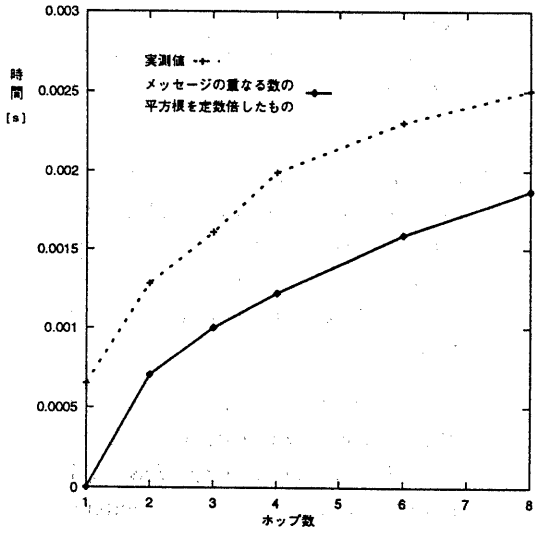

図 8: ノードで重なる数と通信時間

6 おわりに

本稿では，相互結合網を用いた並列マシンの評価に関して，現在どのような評価項目および評価尺度が用いられているかを述べ，通信性能のモデリングを行なった．さらにAP1000上での実験として，ホットスポット・トラフィックの影響と同一ホップ数ランダム1対1通信トラフィックの影響を測定した．その結果，ホットスポットはその他の通信に対してほとんど影響を与えないことがわかった．また同一ホップ数ランダム1対1通信では，ホップ数が大きくなるにしたがって，通信時間が長くなり，それがノードでの通信経路の重なる数の平方根に比例することがわかった．

今後はよりしっかりしたモデリングと，実際のアプリケーションで用いられているさまざまな通信パターンを解析，評価し，並列アーキテクチャやライブラリなどにフィードバックしていきたい．

表 2: 評価項目および評価尺度

		時間		速度		率	
		プログラム総実行時間	ネットワーク・レイテンシ	通信レイテンシ	実効転送バンド幅	プロセッサ使用効率	チャネル使用効率
アーキテクチャ関連	トポロジー	[5]	[14]				
	マシン・サイズ	[5]	[14, 13, 9]	[7, 11]	[15, 9]	[7]	
	通信方式			[7]		[7]	
	ノード能力		[7, 9]	[7]	[11, 9]	[7]	
メッセージ関連	メッセージ・サイズ	[5]	[12]				
	メッセージ発生間隔		[14, 7, 12]	[7]	[14, 15, 11]	[7]	[15]
	通信距離				[4]		
	通信トラフィック		[13, 10]			[7]	

謝辞

日頃ご討論頂く，九州大学 大学院総合理工学研究科 安浦寛人 教授，ならびに，岩下茂信

氏をはじめ安浦研究室の諸氏に感謝します．

本研究は一部，文部省科学研究費補助金 重点領域研究「超並列原理に基づく情報処理体系」，および，EAGL事業推進機構 育成研究助成金による．

参考文献

[1] 清水, 堀江, 石畑, "高速メッセージハンドリング機構 — AP1000における実現—," 情処論, vol.34, no.4, pp.638–647, 1993年4月.

[2] 柴村, 久我, 末吉, "超並列計算機のための相互結合網シミュレータの開発," 情処研報, ARC–97–16, 1992年12月.

[3] 新世代コンピュータ開発機構, 並列アーキテクチャに関する技術開発動向調査研究報告書, 3.2節, pp.32–50, 1993年3月.

[4] 堀江, 石畑, 池坂, "並列計算機AP1000における相互結合網のルーチング方式," 信学論, vol.J75-D-I, no.8, pp.600–606, 1992年8月.

[5] 堀江, 林, "トーラスネットワークにおける最適全対全通信方式," 情処論, vol.34, no.4, pp.628–637, 1993年4月.

[6] 國貞 勝弘, AP1000におけるセル間通信性能の測定, 九州大学工学部卒業論文, 1993年2月.

[7] Vikram S. Adve, Mary K. Vernon, *Performance Analysis of Multiprocessor Mesh Interconnection Networks with Wormhole Routing*, Computer Sciences Technical Report #1001, Univ. of Wisconsin, Madison, Feb. 1991.

[8] Athas, W. C. and Seitz, C. L., "Multicomputers: Message-Passing Concurrent Computers," *IEEE Computer*, vol.21, no.8, pp.9–24, Aug. 1988.

[9] Hyunsoo Yoon, Kyungsook Y. Lee, Ming T. Liu, "Performance Analysis of Multibuffered Packet–Switching Networks in Multiprocessor Systems," *IEEE transaction on computers*, vol. 39, No.3, pp.319–327, Mar. 1990.

[10] Luis Gravano, Gustavo D. Pifarre, Gustavo Denicolay, Jorge L. C. Sanz, "Adaptive Deadlock–free Wormhole Routing in Hypercubes," *Proc. 6th Int'l. Paral. Processing Symp.*, pp.512–515, Mar. 1992.

[11] Jianxun Ding, Laxmi N. Bhuyan, "Performance Evaluation of Multistage Interconnection Networks with Finite Buffers," *Proc. International Conference on Parallel Processing*, pp. I-592–I-599, Aug. 1991.

[12] Jiun-Ming Hsu, Pnthviraj Banerjee, "Performance Evaluation of Hardware Support for Message Passing in Distributed Memory Multicomputers," *Proc. International Conference on Parallel Processing*, pp. I-604–I-607, Aug. 1991.

[13] William J. Dally, "Perfomance Analysis of k–ary n–cube Interconnection Networks," *IEEE Trans. on Comp.*, vol 39, No. 6, pp.775–785 June 1990.

[14] Seth Abraham, Krishnan Padmanabhan, "Performance of the Direct Binary n–cube Network for Multiprocessors," *IEEE Trans. on Comp.*, vol 38, No.7, pp.1000–1011, July 1989.

[15] Imrich Chlamtac, Ora Ganz, "Performance Models of Asynchronous Multitrunk HYPERchannel Networks, " *IEEE Trans. on Comp.*, vol C–36, No. 2, pp.138–146 Feb. 1987.

[16] Pfister, G. P. and Norton, V. A., "Hot Sopt Contention and Combining in Multistage Interconnection Networks," *IEEE Trans. Comput.*, vol.C-34, no.10, pp.943–948, Oct. 1985.