計算機アーキテクチャ 102－9
（1993．10．21）

相互結合網シミュレータ INSIGHT による超並列マシン向き相互結合網の性能評価

柴村英智† 久我守弘‡ 末吉敏則†‡

† 九州工業大学 情報工学部 知能情報工学科
‡ 九州工業大学 マイクロ化総合技術センター

sibamura@mickey.ai.kyutech.ac.jp
kuga@cms.kyutech.ac.jp
sueyoshi@ai.kyutech.ac.jp

超並列計算機の相互結合網は数千から数万台のプロセッサ要素間の通信を効率良く実現しなければならず．トポロジ，チャネル幅およびフロー制御方式などの相互結合網に関する適切なパラメータの設定が重要である．本稿では，相互結合網シミュレータ INSIGHT の概要について述べた後，INSIGHT を用いて 2 次元トーラス網，3 次元トーラス網，12 次元ハイパーキューブ網ならびに RDT 網におけるフロー制御方式，チャネル幅，ネットワークの動作周波数の変化による通信性能への影響について調査する．その結果，2 次元トーラス網に上位トーラス網を再帰的に構築した RDT 網は，2 次元トーラス網と比較して大幅な通信性能の向上が確認された．

Performance Evaluation of Interconnection Networks for Massively Parallel Computers Utilizing An Interconnection Network Simulator INSIGHT

Hidetomo Shibamura† Morihiro Kuga‡ Toshinori Sueyoshi†‡

† Department of Artificial Intelligence,
Kyushu Institute of Technology
Iizuka, 820 Japan

‡ Center for Microelectronic Systems,
Kyushu Institute of Technology
Iizuka, 820 Japan

Interconnection networks for massively parallel computers must be realized efficient inter-processor communication between tens of thousands of processor elements. Some suitable parameters related to interconnection network such as topology, channel width, flow control, and so on, should be set up. In this paper, we describe framework of interconnection network simulator, INSIGHT, and examine the effects on the communication performances of 2-dimensional torus, 3-dimensional torus, 12-dimensional hypercube, and RDT (Recursive Diagonal Torus) , with different flow controls, channel widths, and data transfer frequency utilizing INSIGHT. As a result, we found out that the RDT has achieved great improvement of the communication performance compared with 2-dimensional torus.

1 はじめに

近年，大規模並列処理の需要に応へるべく，各所で超並列計算機に関する研究ならびにシステムの設計や開発が行われている．超並列計算機の構成要素の一つである相互結合網は，数千から数万台のプロセッサ要素間の通信を効率良く実現しなければならない．従って，高バンド幅，高スループット，低レイテンシおよびハードウェアの拡張容易性 (スケーラビリティ) が重要視されている．所望する性能を発揮する相互結合網を設計する際には，通信性能を把握するために，予めシミュレーション等による評価ならびに解析が必須となる．また，実現可能な範囲内で高い通信性能を得るために，回路規模を検討すると共に，フロー制御方式，チャネル幅，パケット長ならびにルータ間のデータ転送周波数などに関したパラメータを適切に設定しなければならない．このように，開発段階における設計支援のためのネットワーク・シミュレーションは重要な課題である．

本研究では，文部省重点領域研究において開発が進められている超並列プロトタイプ計算機 (仮称：D-machine) の相互結合網に採用予定である RDT (Recursive Diagonal Torus) 網[1][2]に重点を置き，基礎的な通信性能の評価を行う．RDT は，2 次元トーラス網上に再帰的に 2 次元トーラス網を構築する．その結果，大規模なノード構成時においても，他の相互結合網と比較して通信直径を小さく抑えることができ，高い通信処理性能の実現を期待できる．今日までに，メッシュ(トーラス) 網に関する様々な性能評価[3][4]が行なわれているが，ハードウェアの性能を考慮したシステム開発の観点からの評価事例は少ない．従って，システムの規模やハードウェアの仕様を踏まえた性能評価が必要である．

本稿では，RDT に加え 2 次元トーラス網，3 次元トーラス網およびハイパーキューブ網について，それぞれフロー制御方式，チャネル幅ならびにネットワークのデータ転送周波数を変化させた場合の通信性能について比較検討する．

以下，第 2 章では様々な相互結合網の性能評価を遂行するために開発した相互結合網シミュレータについて概説する．次に，第 3 章ではシミュレーションを行った相互結合網の仕様構成について述べる．さらに，第 4 章ではシミュレーション結果について考察し，最後に，第 5 章で簡単なまとめを述べる．

2 相互結合網シミュレータ：INSIGHT

超並列計算機の開発に向けて相互結合網の性能評価ならびに設計支援を様々な側面から行うために，相互結合網シミュレータ (名称：INSIGHT) を開発した[5][6]．超並列計算機向き相互結合網のシミュレータに求められる要求仕様として次に示すものが挙げられる．

```
;;; 64x64 RDT(2,4,1) NDL program
(FlowControlMethod VirtualCutThrough)
(NetworkChannelWidth 32bit)
(NetworkFrequency 50MHz)
(NetworkPacketSize 256bit)
(NetworkDataSize 192bit)
(NodeHardwareDelay 60nsec)
(NodeBufferSize 256bit)
(MessageCreationMax 10000)
(MessageCreationInterval 2usec)
(MessageCreationRate 1.0%)
(MessageLength 8byte to 32byte)
(MessageMaxCount 10000)
(SimulationInterval 1usec)
;(CommunicationPatternFile "cpf/pde.cpf")
;(CommunicationPatternFile "cpf/sort.cpf")
;(CommunicationPatternFile "cpf/matrix.cpf")
;(ProcessorPerformanceExpandRate 50%)
(DataOutputFile "rdt.4096.vct.dof")
(ScreenOutputFile "rdt.4096.vct.sof")
```

図 1：ネットワーク記述言語 (一部)

(1) 数千から数万台のプロセッサ要素を結合する大規模な相互結合網をシミュレートできる．

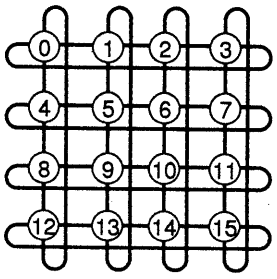
(2) 相互結合網の仕様の設定および変更を容易に行うことができる．

(3) ランダムな通信パターンに加え，実際のアプリケーションの通信パターンも踏まえた性能評価が行える．

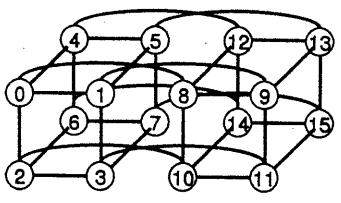
INSIGHT は，これらの要求仕様を満たすように設計し，作成した．特に上記 (2) に関しては，トポロジ，フロー制御方式，チャネル幅，ネットワークのデータ転送周波数などの様々な構成要素のパラメータおよびハードウェアの性能を，図 1 に示すようなネットワーク記述言語と呼ぶ本シミュレータ専用の言語仕様に基づき記述することで対処している．そのため，図 2に示すように所望とする相互結合網を柔軟に表現でき，様々な観点からシミュレーションを行うことができる．

また，INSIGHT を用いて得られるシミュレーション結果から以下に示す相互結合網の性能を把握することができる．

- ネットワーク遅延・通信遅延
  - ネットワーク全体におけるメッセージおよびパケットのネットワーク (通信) 遅延
  - 距離別ごとのメッセージおよびパケットのネットワーク (通信) 遅延
  - 特定ノード間のメッセージおよびパケットのネットワーク (通信) 遅延



(a) トーラス網



(b) ハイパーキューブ網

図2:ネットワーク記述言語による相互結合網の記述例

- **転送処理能力**
  - 単位時間あたりにネットワークへ注入されるメッセージ長の総計およびメッセージの総数
  - 単位時間あたりにネットワーク内で残留しているメッセージ長の総計およびメッセージの総数
  - 単位時間あたりにノードへ到着したメッセージ長の総計およびメッセージの総数
- **負荷状態**
  - 各ノードあたりのメッセージおよびパケットの衝突回数
  - 1メッセージおよび1パケットあたりの衝突回数
  - ノードにおけるメッセージの送受信回数

相互結合網の評価指標としては上記の項目を選択できる．通信性能について評価を行う場合，ネットワーク遅延を用いることが多い．一方，通信遅延時間を評価指標として用いる場合，メッセージ送受信処理に起因するソフトウェアやハードウェアの複雑な制約に対応しなければならない．すなわち，シミュレーションによる性能解析においても相互結合網の仕様に加えプロセッサを中心としたノード内部の構成仕様も考慮しなければならない．従って，プロセッサ・ノードに関する機能レベルのシミュレーションが必要になる．

3 相互結合網の仕様

本章では，INSIGHTを用いて性能評価を行った相互結合網の構成およびシミュレーションのパラメータについて述べる．

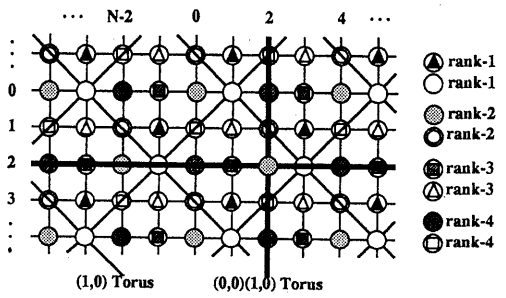


図3:RDT(2,4,1)/αの構成

3.1 RDT

D-machineに採用する相互結合網への要求の一つとして，メッシュ(トーラス)網の包含あるいはエミュレーションの実現が挙げられている[7]．RDTは，D-machineの相互結合網に対する要件を満たすために設計され，図3に示すように，2次元トーラス上に再帰的に2次元トーラスを構成する[8]．RDTにおいて，基数を2，最大の上位トーラスをランク4，ノードの多重度を1と設定したRDT(2,4,1)の直径は，4,096ノード構成の場合は8であり，16,384ノード構成の場合では10である．同じハードウェア規模の相互結合網と比較して直径が小さいため，細粒度並列処理に耐える高い通信性能が期待できる．

本研究では，64×64のノード構成をとるRDT(2,4,1)についてシミュレーションを行う．トポロジの通信直径は8，ノードの次数は8である．ルーティング・アルゴリズムにはRDT専用に設計されたベクトル・ルーティングを用いる．なお，デッドロック防止に関しては，Virtual channel[9]を考慮したベクトル・ルーティングが現在検討されている．従って本稿では，フロー制御方式にWormholeおよびVirtual cut-throughを用いる場合には，Virtual channelバッファの多重度を1と設定し，シミュレーション時のデッドロックについてはメッセージをドロッピングすることにより切捨る．

3.2 2次元トーラス網

64×64ノード構成の2次元トーラス網についてシミュレーションを行う．トポロジの通信直径は64，ノードの次数は4である．

一般的なルーティング・アルゴリズムとしては，基本的なx-first y-nextによるものや，デッドロック回避のためにVirtual channelフロー制御と併用するe-cubeルーティング[10]がある．e-cubeルーティングを用いた場合，プロセッサ要素間の通信パターンの局所性ならびにプロセッサ要素の性能によりネットワーク遅延が著しく変化することが明らかになっている[8]．また，2次元トーラス網に対して上位トーラス網を持つRDTの性能向上について調査を行うため，ルーティング・ア

ルゴリズムはx-first y-nextによるものとする．従って，RDTの場合と同様にフロー制御方式にWormholeおよびVirtual cut-throughを用いる場合には，Virtual channelバッファの多重度を1と設定し，デッドロックについてはメッセージをドロッピングする．

3.3 3次元トーラス網

16×16×16ノード構成の3次元トーラス網についてシミュレーションを行う．トポロジの通信直径は24，ノードの次数は6である．

2次元トーラス網と仕様の設定基準を同じにするために，ルーティング・アルゴリズムには*e*-cubeルーティングを用いず，x, y, z座標の順序で静的なルーティングを行う．また，Virtual channelバッファの多重度も1と設定し，デッドロックについてはメッセージをドロッピングする．

3.4 ハイパーキューブ網

12次元構成のハイパーキューブについてシミュレーションを行う．トポロジの通信直径は12，ノードの次数は12である．

ルーティング・アルゴリズムには，ハミング距離1の隣接ノードへ各次元毎にアドレスを一致させる一般的な手法を用いる．また，上記と同様に，Virtual channelバッファの多重度は1と設定し，デッドロックについてはメッセージをドロッピングする．

3.5 シミュレーション・パラメータ

評価のためにフロー制御方式，チャネル幅，ネットワークのデータ転送周波数およびメッセージの発生間隔をそれぞれ変化させ，それに対する相互結合網のネットワーク遅延を測定した．評価パラメータとして，以下に示すように現実に即した設定値を種々組合せてシミュレーションを行った．

- **トポロジ**：2次元トーラス網，3次元トーラス網，RDT，12次元ハイパーキューブ
- **ノード数**：4096
- **フロー制御方式**：Store and forward, Wormhole, Virtual cut-through
- **ルーティング・アルゴリズム**：前述のアルゴリズムを用いて静的なルーティングを行う．
- **チャネル幅**：16 bits, 32 bits, 64 bits (双方向通信)
- **パケット長**：256 bits (ヘッダ部：64 bits，データ部：192 bits)
- **データ転送周波数**：25 MHz, 50 MHz, 100 MHz
- **ハードウェア遅延**：2 clock (ルーティング処理：1 clock，スイッチ通過：1 clock)
- **メッセージ発生間隔**：2 μsec., 10 μsec., 50 μsec., 100 μsec.
- **メッセージ発生率**：1 clockにつき1.0 %の割合．すなわち各ノードにおいて100 clock毎に1回の割合でメッセージを発生．

また，本研究では相互結合網の基本的な性能を調査するために，評価基準としてネットワーク遅延を用いる．

4 性能評価

本章では，シミュレーション結果からフロー制御方式，チャネル幅，およびデータ転送周波数の違いが相互結合網の性能に対して，どのような影響を及ぼすかを考察する．

4.1 フロー制御方式の差異による影響

チャネル幅を32 bits，パケット長を256 bits，ネットワークのデータ転送周波数を50 MHzと設定し，フロー制御方式をStore and forward, Wormhole, Virtual cut-throughと変化させた場合について測定した．

Store and forwardを用いた場合の平均ネットワーク遅延を図4に示す．図4から，2次元トーラス網の遅延が非常に大きいことが分かる．一方，2次元トーラス網上に上位トーラス網を構成したRDTは，2次元トーラス網と比較して通信直径が64から8に減少し，ノードあたりのリンク数が4から8へと増えるため，ネットワーク遅延の大幅な減少が達成されている．

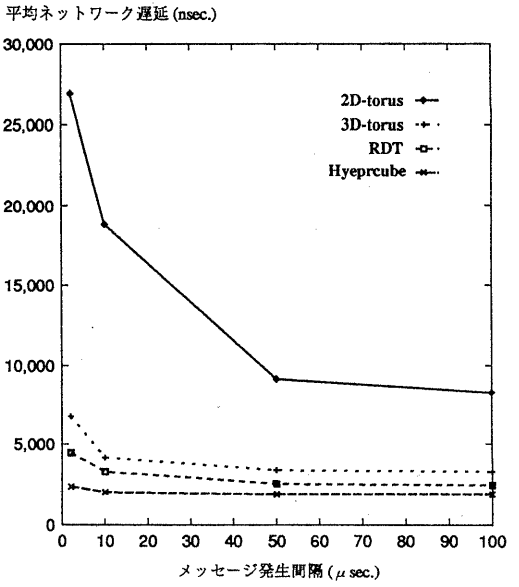


図4: Store and forwardによる平均ネットワーク遅延

次に，Wormholeを用いた場合の平均ネットワーク遅延を図5に示す．前述のStore and forwardと比較して，2次元トーラス網のネットワーク遅延は大幅に減少している．一方，3次元トーラス，RDTおよび12次元ハイパーキューブについては大きな変化が見られない．これは，次数が低い2次元トーラス網では，フロー制御方式の変化によりデータの転送効率が向上するが，次数が高い他の相互結合網は潜在的に転送効率が良いため，フロー制御方式の変化によるデータの転送効率の向上があまり見られないためである．

メッセージの発生間隔が短い場合には，3次元トーラスとRDTのネットワーク遅延の逆転が見られる．これは，RDTにおいて網内に注入されるメッセージが増加するにつれ，上位トーラスを経由するメッセージも増加するので，遠距離ノード同士のリンクを持つ上位トーラス網にメッセージが集中し衝突が頻繁に発生するため，3次元トーラスより性能が低下すると考えられる．

Virtual cut-throughを用いた場合の平均ネットワーク遅延を図6に示す．Wormholeの場合と比較して，メッセージの発生間隔が短くなると，2次元トーラス網，3次元トーラスおよびRDTはVirtual cut-throughにおけるパケット長のバッファが有効になり，ネットワーク遅延がさらに減少していることが分かる．また，RDTと3次元トーラスを比較すると，バッファによるフリット同士の衝突の緩和により，Wormholeの場合に見られたネットワーク遅延の逆転が抑えられていると考えられる．一方，12次元ハイパーキューブについては，ほとんどネットワーク遅延の減少が見られない．これは，12次元ハイパーキューブの転送効率が他の3つの相互結合網と比較して非常に高いことから，バッファの効果が低いためと考えられる．

4.2 チャネル幅の差異による影響

フロー制御方式をStore and forward，パケット長を256 bits，ネットワークのデータ転送周波数を50 MHzと設定し，チャネル幅を16 bits，32 bits，64 bitsと変化させた場合について測定を行った．

チャネル幅を16 bitsとした場合の平均ネットワーク遅延を図7に示す．図7から，2次元トーラス網はメッセージ発生間隔が50μ秒あたりからのネットワーク遅延の増加が大きい．また，2次元トーラス網において，相互結合網内のメッセージが混雑し始めると急激にネットワーク遅延が大きくなることが分かる．一方，3次元トーラス網，RDTおよび12次元ハイパーキューブは2次元トーラス網と比較して大きな次数ならびに小さな通信直径を持つため，メッセージの発生が増加してもネットワーク遅延が急激に増加しない．

次に，チャネル幅を32 bitsとした場合の平均ネットワーク遅延を図8に示す．図8から，チャネル幅を16 bitsとした場合と比較して，ネットワーク遅延の増加率が減少し始めている．また，チャネル幅が16 bitsから32 bitsへと，2倍になるとネットワーク遅延もほぼ

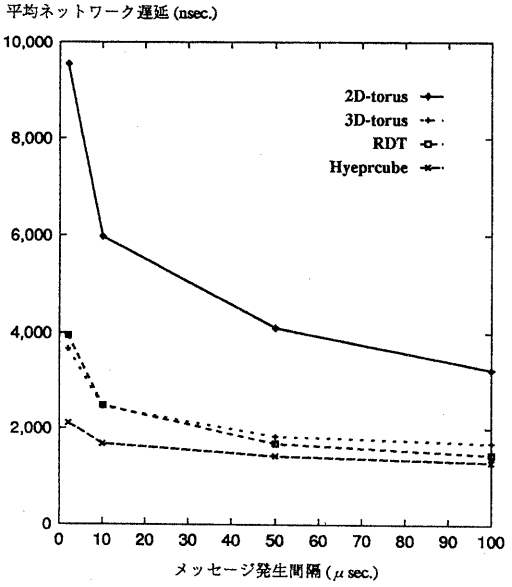


図5：Wormholeによる平均ネットワーク遅延

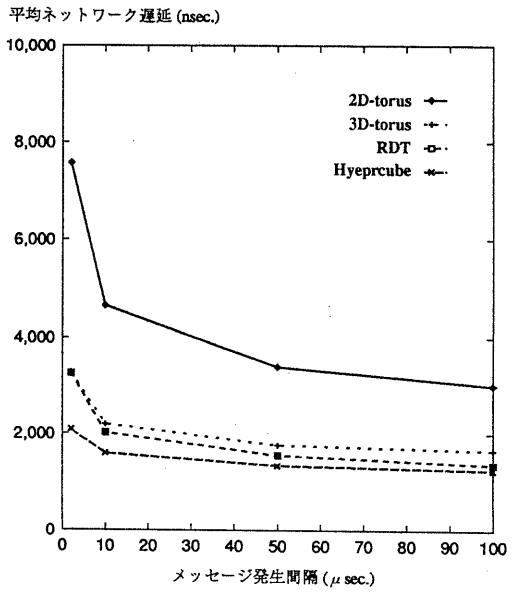


図6：Virtual cut-throughによる平均ネットワーク遅延

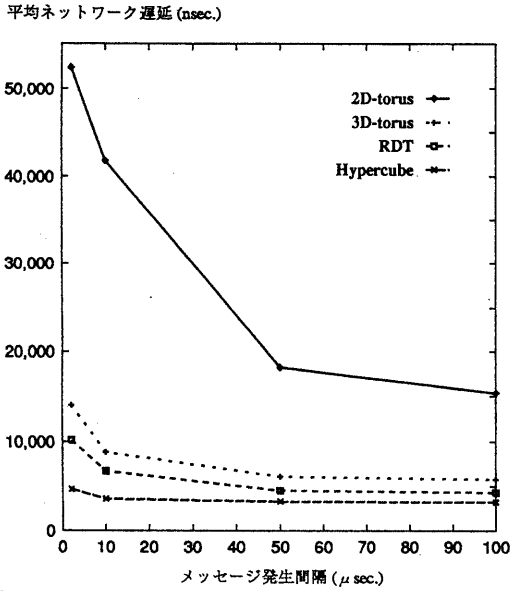


図7:チャネル幅が16 bitsの場合の平均ネットワーク遅延 (Store and forward)

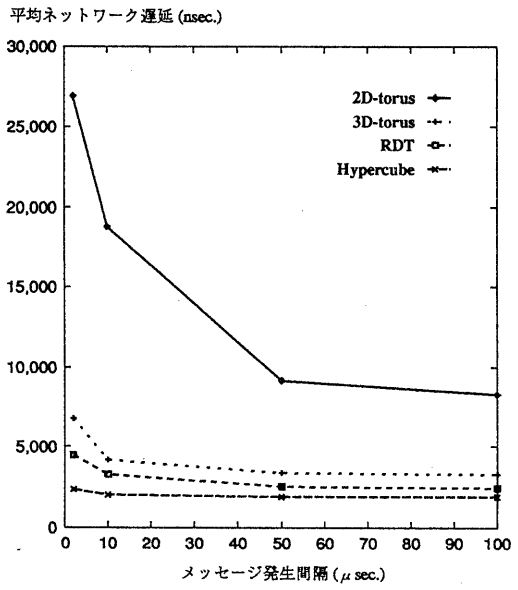


図8:チャネル幅が32 bitsの場合の平均ネットワーク遅延 (Store and forward)

1/2に減少していることが分かる.

さらに,チャネル幅を64 bitsとした場合の平均ネットワーク遅延を図9に示す.図9から,チャネル幅が比

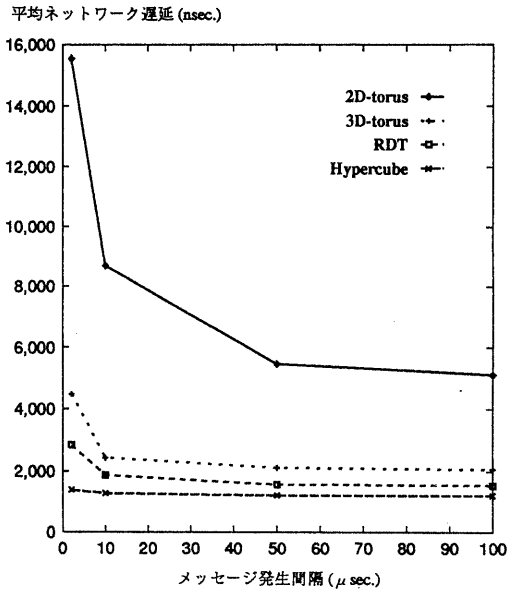


図9:チャネル幅が64 bitsの場合の平均ネットワーク遅延 (Store and forward)

較的広いためデータ転送効率が良くなり,メッセージの発生に対する転送処理能力が向上していることが分かる.前述と同様に,チャネル幅が32 bitsから64 bitsへと,2倍になるとネットワーク遅延もほぼ半減している.

4.3 データ転送周波数の差異による影響

フロー制御方式をStore and forward,チャネル幅を32 bits,パケット長を256 bitsと設定し,ネットワークのデータ転送周波数を25 MHz,50 MHz,100 MHzと変化させた場合について測定を行った.

データ転送周波数を25 MHzとした場合の平均ネットワーク遅延を図10に示す.図10から,メッセージの発生間隔が短い場合には3次元トーラスとRDTの通信性能が近似していることが分かる.

次に,データ転送周波数を50 MHzとした場合の平均ネットワーク遅延を図11に示す.25 MHzの場合と比較して,データ転送周波数が2倍になるとネットワーク遅延は1/2以下になることが分かる.

さらに,データ転送周波数を100 MHzとした場合の平均ネットワーク遅延を図12に示す.RDTと12次元ハイパーキューブを比較すると,平均ネットデータ転送周波数が25 MHz,50 MHz,100 MHzと変化するにつれてRDTはネットワーク遅延の傾きが減少している.すなわち,データ転送周波数が向上するとともに通信性能が向上している.これは,データ転送周波

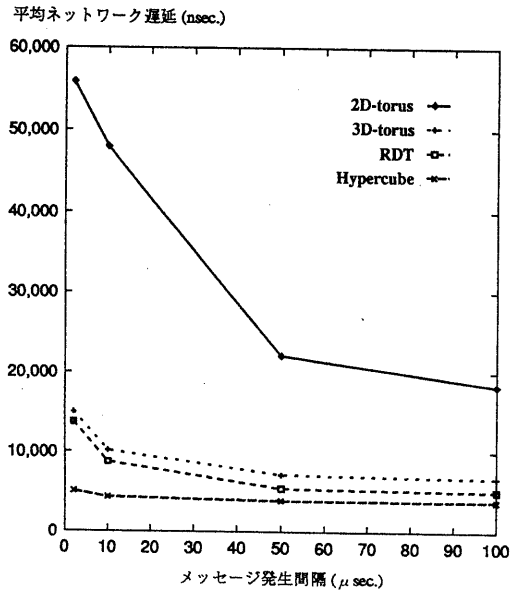


図10: データ転送周波数が25 MHzの場合の平均ネットワーク遅延 (Store and forward)

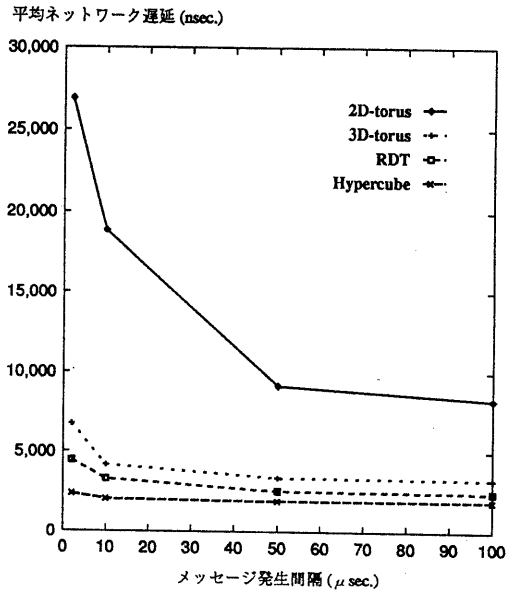


図11: データ転送周波数が50 MHzの場合の平均ネットワーク遅延 (Store and forward)

数が低い時，相互結合網内でメッセージが混雑し，多くのメッセージがRDTの上位トーラスを経由するため衝突が頻繁に発生していたが，データ転送周波数が高くなるにつれて衝突が緩和されるためである．

4.1節から，2次元トーラス網は比較的にデータ転送効率が良くないが，RDTでは再帰的に上位トーラス網を構築し，ベクトル・ルーティングを使用した結果，大幅なデータ転送効率の向上が確認できた．一方，メッセージの発生間隔が短い場合，フロー制御方式の変更によりネットワーク遅延が3次元トーラス網のものと逆転する傾向が一部見られた．遠距離ノード同士の通信リンクを持つ一部の上位トーラス網にメッセージが集中することにより通信性能の低下を招く恐れがある．

また，4.1~4.3節の結果から，Store and forwardにおいては，チャネル幅を広げるよりもデータ転送周波数をあげる方が相互結合網のデータ転送効率を向上させることができることが判明した．

5 おわりに

本稿では相互結合網シミュレータINSIGHTを利用して超並列計算機向きの相互結合網の性能評価について述べた．相互結合網を特徴付けるフロー制御方式，チャネル幅，およびデータ転送周波数の差異は，ネットワークにおけるネットワーク遅延に影響を与え，相互結合網の通信性能を大きく左右することが明らかになった．

INSIGHTを用いることにより目的とする相互結合

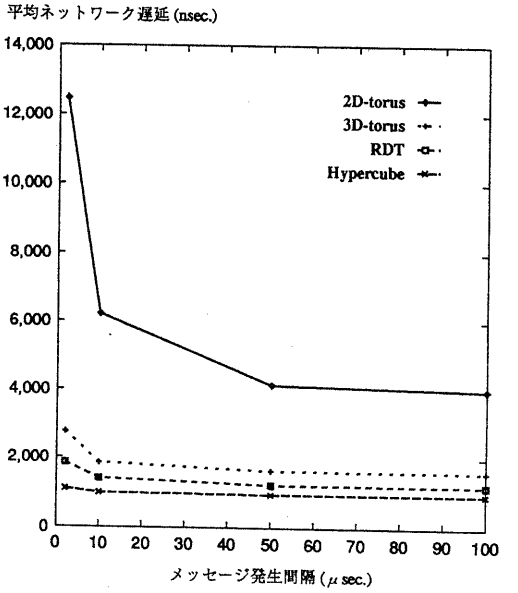


図12: データ転送周波数が100 MHzの場合の平均ネットワーク遅延 (Store and forward)

網の具体的な性能評価を容易に遂行できた．現在，研究の進展に伴って要求が生じたネットワーク記述言語の拡張を行っており，ノード内部の詳細な仕様をノー

ド単位に記述し，異種ノード間の接続記述ができるようにしている．また，トポロジを表現する際に，バス(クラスタ)構成ならびに単方向リンクの記述も可能となる予定である．なお，さらに大規模な相互結合網も短時間で評価できるように，シミュレータの高速化ならびにメモリの使用効率の向上を図っている．

今後，種々のパラメータの設定に従って様々な相互結合網の性能評価を進めていく予定である．

謝辞

RDT 網の性能評価にあたり，RDT およびルーティング・アルゴリズムについて日頃から貴重な御助言をいただく慶応義塾大学理工学研究科の天野英晴先生ならびに同学理工学研究科院生の加藤和彦氏に感謝します．また，シミュレータの開発にあたり，御討論いただいた本学情報工学研究科の手塚忠則氏(現在，松下電器産業株式会社)ならびに了戒清氏に感謝の意を表す．なお，本研究の一部は文部省科学研究費(重点領域研究(1)課題番号 04235103「超並列ハードウェア・アーキテクチャの研究」)の補助を受けたことを付記する．

参考文献

[1] 楊 愚魯，天野 英晴，柴村 英智，末吉 敏則：超並列向き結合網 Recursive Diagonal Torus の諸特性，電子情報通信学会 CPSY 93-13~27 (SWoPP '93)，pp.105–112, Aug. 1993.

[2] Y. Yang, H. Amano, H. Shibamura, T. Sueyoshi: Recursive Diagonal Torus : An Interconnection Network for Massively Parallel Computers, *Proc. of IEEE Symposium on Parallel and Distributed Processing Symposium*, to be appear.

[3] William J. Dally : Performance Analysis of k-ary n-cube Interconnection Networks, *IEEE Trans. Comput.*, vol. 39, no.6, pp.775–785, June 1990.

[4] Kirk L. Johnson : The Impact of Communication Locality on Large-Scale Multiprocessor Performance, *Proc. of 19th Annual International Symposium on Computer Architecture*, pp.392–402, May 1992.

[5] 柴村 英智, 久我 守弘, 末吉 敏則：超並列計算機のための相互結合網シミュレータの開発, 情報処理学会研究報告, 92-ARC-97, pp.121-128, Dec. 1992

[6] 柴村 英智, 久我 守弘, 末吉 敏則：超並列計算機のための相互結合網シミュレータ，並列処理シンポジウム JSPP '93 論文集，pp.159–166，May. 1993.

[7] 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 2 回シンポジウム予稿集，p. 197, Mar. 1993.

[8] 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 3 回シンポジウム予稿集，p. 316, Sep. 1993.

[9] William J. Dally : Virtual-Channel Flow Control, *IEEE Trans. Parallel and Distributed Systems*, vol.3, no.2, pp.192–205, Mar. 1990.

[10] William J. Dally, Charles L. Seitz : Deadlock-Free Message Routing in Multiprocessor Interconnection Networks, *IEEE Trans. Comput.*, vol. 36, no.5, pp.547–553, May 1987.