計算機アーキテクチャ 111－6
（１９９５．３．１０）

分散独立型入出力システムのための結合網の構成と評価

大西一正　北村徹　大上靖弘　清水雅久

RWCP[1]超並列三洋研究室[2]

分散独立型入出力システムのための結合網のシミュレーションによる性能評価を行う。超並列計算機で実行されるアプリケーションの多くは大量のデータを使用するものであり、超並列計算機には大規模な二次記憶を有することが要求される。また画像、音声データなど種々のデータを扱うことも要求されており、画像や音声の入出力装置が接続されるため、プロセッサ群と二次記憶装置および画像音声入出力装置を接続するための結合網が必要である。本論文では入出力結合網のために、クロスバースイッチを複数組み合わせた階層型のネットワークを提案し、種々のアクセスパターンに対し、ソフトシミュレーションによりアクセス性能の評価を行う。

Network Architecture and Performance Evaluation for a Dispersive Independent I/O System

Kazumasa OHNISHI, Toru KITAMURA, Yasuhiro OUE and Masahisa SHIMIZU

Massively Parallel Systems Sanyo Laboratory, RWCP.

This paper describes network architecture and performance evaluation for a dispersive independent I/O system. The almost applications executed on the massively parallel computer treat massive amount of data. Massively parallel computer must have enormous amount of secondary storage. And also, the data treated on a massively parallel computer include video or audio data frequently. So a massively parallel computer should be connected with multimedia devices. We had proposed the architecture which I/O devices separated from processors. We propose the multi-layer network which consists of some crossbar switches, and evaluate this I/O network by software simulation.

[1]RWCP:Real World Computing Partnership(新情報処理開発機構)
[2]三洋電機(株)東京情報通信研究所内

1 はじめに

超並列計算機が扱う問題は大規模であり、それらの問題では膨大なデータが使用される。大規模な問題では、例えば、Grand Challenge のように 100GB から 1TB のデータを必要とするアプリケーションもある [1]。超並列計算機の二次記憶に対しては、このような大量のデータを蓄え、かつ高速にアクセスする能力が必要とされる。このため大規模な外部ディスクが必要な場合が多い。また、超並列計算機上では膨大な数のアクセス要求が発生するため、プロセッサの数やアクセス性能に見合う高い並列度とバンド幅が要求される。

このような大規模な二次記憶を実現するために、我々は 1000 台規模の二次記憶装置を並列に接続し、効率的な並列アクセスが可能な二次記憶の基本アーキテクチャを提案した [2][3]。我々はマルチメディアデータなどさまざまなタイプのデータの処理が可能な汎用超並列計算機のための入出力システムを構築することを目指している。

特に本アーキテクチャでは、ファイルデータを蓄える複数の二次記憶装置、及び画像や音声の入出力装置等をプロセッサの各クラスタから分離し、入出力結合網を介してプロセッサと接続する構成をとることにより、スケーラブルな二次記憶の実現を容易にした。また、二次記憶もクラスタに分割し、クラスタごとに入出力結合網と接続することにより、入出力結合網の規模を実現可能なノード数に削減した。

これまでも１０００台規模のプロセッサを前提としたプロセッサ間相互結合網については多数の研究がなされてきた [4]。それに対して本論文では入出力のためのネットワークを分離したアーキテクチャーに着目した。以降ではファイルデータ、画像データの混在する入出力結合網の実現について述べる。

1.1 想定する超並列計算機のアーキテクチャの概要

まず、我々が提案する分散独立型入出力システムを実現する前提となる超並列計算機の概略構成を述べる。図 1は、超並列計算機の全体構成を示したものである。超並列計算機は 1024 台のプロセッサを有し、128 のクラスタに分割される [5]。各クラスタは入出力結合網を介して並列二次記憶や画像音声入出力デバイスと接続される。従って、プロセッサとの接続のために必要となる入出力結合網のノード数は 128 となる。

プロセッサ側とのバランスをとるために、入出力結合網は並列二次記憶及び画像音声入出力に対してもプロセッサ側と同様に 128 のノードを有し、クラスタ側と同様の転送性能を持つ。

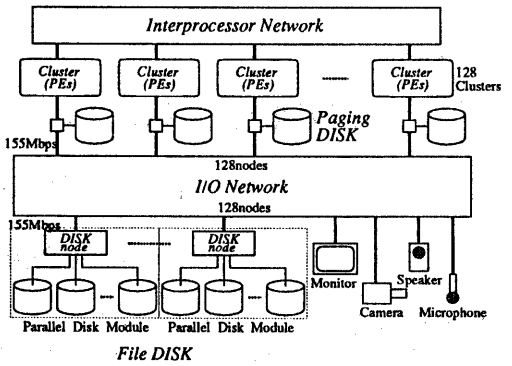

図 1: 並列二次記憶のアーキテクチャ

2 入出力結合網の要件

次にこの入出力結合網に対する要件について述べる。

2.1 画像データの転送

プロセッサとディスク装置もしくは画像入出力装置を接続することから、ファイルデータの転送と画像データの転送が混在する状況が想定され、この場合の性能が高いことが求められる。特に画像データについてはデータ量も多いのでその特質について考察したい。

画像データを転送する時、NTSC 画像の場合で約１００Mbps、ハイビジョン画像の場合で約 1.2Gbps の転送レートが必要である。このレートの信号を伝送するには画像圧縮によりデータ量を削減するか 、もしくは専用の通信路を確保しなければならない。また画像を構成するフレームを１／３０秒に一枚再生する必要があり、レイテンシに関しても厳しい要求が課せられる。

以降ではハイビジョンの場合を例にとり述べる。MPEG 2を用いてデータ圧縮することによりデータ量は１／３０から１／８０に削減することが可能である。これによりデータレートは４０Mbps- １５Mbps 程度に低下する。しかしこれは平均値であり、シーンチェンジの場合など急激に画面が変化する場合には突発的に大量のデータが発生する。この際 MPEG 2ではＩフレームとよばれる前後画面情報なしに生成可能な DCT のみによる符号化画面を生成する。このＩフレームは１／１０から１／２０のデータ量に削減することしかできない。したがってＩフレームを送信する１／３０ｓの間には 1.2Gbps/10 = 120Mbps 程度の通信バンド幅が必要になる。急激なシーンチェンジが連続する可能性は高くないが、やはり数枚程度のＩフレームを送信可能な性能が入出力結合網には求められ

る。またこの際に入出力結合網上で混在しているファイルデータの通信性能を極力損なわないような配慮が必要である。

2.2 その他の要件

計算を行うプロセッサは比較的近距離に配置され、同一筐体に収められることが多い。しかしながら入出力装置などは実装上の問題からこの筐体に収められずに比較的遠距離に配置される場合が多い。そこで、プロセッサと入出力装置を結ぶ入出力結合網は遠距離の通信に適した形態が求められる。

入出力装置に関してはディスクドライブ等は稼働部があるために、プロセッサ部等に比べて故障率が高い。また遠距離の通信なので通信路に関しても信頼性が低下することが予想される。

上記の内容も含めて 入出力結合網の要件をまとめると、次のようになる。

- 多ノード、遠距離通信に対応した結合網の実現
- 多様なアクセスの混在
- 画像データ通信への対応
- フォールトトレラント性

3 解決の指針

3.1 結合網の実現

半導体技術の進歩により現在では5 Gbps程度の交換性能をもつスイッチも実現されつつある。このスイッチ内であればデータは宛先が重複しない限りブロッキングを起こさずに宛先のポートに到達することができる。1ポートのバンド幅を SONET(SDH) の155 Mbpsとすれば5 Gbps/155Mbps = 32.258となり 32ポートの完全結合網を構成することが可能になる。そこで我々はこの32ポートの完全結合網をベースにスイッチを複数組み合わせることで効率的な結合網を構築していくことが現実的な実現方法であると考えた。すなわち下位の階層には完全結合網を採用し、上位の階層には別の結合網を採用し、二つのトポロジーを階層的に組み合わせた構成により結合網全体を構成する。当然この上位層での結合においては完全結合を実現することはできなので、この部分のトポロジーの選択が重要である。

ルーティングの管理に関してはノード数が増えた場合を考慮して集中管理ではなく分散管理を行う。すなわちルーティングは固定的に行うか、もしくは、各クロスバースイッチで決定して行う。このことによりクロスバースイッチ間での余分なデータ通信はなくなりポート数が増えた場合にも対応が可能になる。

入出力結合網の結線の量を考えると、多ビット幅の接続を実現することは困難である。そこでシリアル転送を基本に考え、高い転送速度を確保するために光ファイバーでの通信を行う。

3.2 多様なアクセスへの対応

データ通信方式について考えると、回線交換は発信端末と着信端末で物理的な回線を作る方式であり、相手側の端末に即座に転送が可能である。しかし転送するデータがない場合には回線が使われないので無駄が生じる可能性がある。一方、パケット交換方式は、ディジタル情報を決められた大きさのブロックに区切り その各ブロックにアドレス情報などをヘッダとして付加し、パケットとして送る方法である。情報が発生した場合のみパケットをつくり 送信することにより、間欠的に発生する情報の伝達に適している。

本論文における入出力結合網には、ファイルデータ、画像データ、音声データなどの、さまざまなタイプのデータ転送が混在しており、それらのデータ転送にはパケット通信が適していると考えられる。

また、パケット長は固定として、データ交換をハードウェアで高速に処理することによって、遅延時間を極めて小さくすることが可能である。

3.3 画像データの通信

取り扱う画像データとしてはハイビジョンクラスの映像をMPEG 2クラスの符号化を行うことによって転送することを想定する。シーンチェンジなどのバースト転送時の輻輳を回避するためにバッファを設けて対処する。またこのようなバッファを設けることはファイルデータの通信の場合にもホットスポット衝突の時の対策ともなる。

またパケットに優先ビットを付加することにより、画像データなどには優先的な送信を保証し、レイテンシの削減、必要なスループットの確保を図る。

3.4 フォールトトレラント性

入出力装置本体の耐故障性に関しては既にRAIDディスクなどの対処がなされている。入出力結合網に関してはどのノード間であっても複数のルートを用意し、ダイナミックなルーティングを行うことでフォールトトレラント性を高める。

我々の提案する入出力結合網の特徴を以下にまとめる。

- 階層的な結合網
- 固定長パケットによる通信
- ルーティングの分散管理

- シリアル光通信によるプロセッサ、入出力結合網、入出力装置の接続

4 上位トポロジーの選択（シミュレーションによる評価）

以上の方針に基づき、ネットワークのトポロジー、ルーテイングを決定するために3つの構成案を比較してシミュレーションにより評価を行った。

1024台のプロセッサを8台ごとの128個のクラスタに分割し、そのクラスタと入出力装置を接続する場合を想定した。また入出力装置のバンド幅はクラスタ側のバンド幅に見合うように１２８ノード用意した。このように合計２５６ノードのネットワークについて検討した。

また、各方式を比較評価するためにクロスバースイッチは１６台に固定し、ハードウェア量等の条件を同じにしてシミュレーションを行った。図2にスイッチエレメント（以下SE）内部の構成図を示す。各ポートは宛先が重ならない限りすべてのポートと任意に通信することが可能である。出力側にバッファを持ち宛先が重なった場合のパケットを保存する。

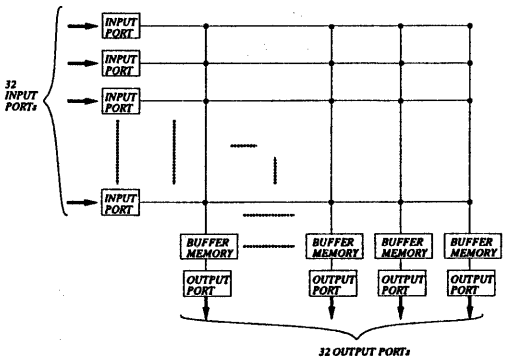

図 2: スイッチエレメント (SE)

4.1 上位トポロジーの決定

以下に上位階層の候補となる3つの構成案を示す。以下説明のために各SEと各ポートをペアにして[SE No., Port No.]の書式で表す。（SE No.はSN、Port No.はPN)

4.1.1 構成案1

図3に示すように、１２台のクロスバースイッチ(SN 0～11)について、２２個のポート（PN 0～21）を外部接続用に用い、残りの１０ポート（PN 22～31）を内部接続用に用いる。そして残り4台のクロスバースイッチ(SN 12～15)を内部接続専用に使い、TREE状の結合網を構成する。

ルーティングは以下のとおりである。第1段目のSEに入ったパケットは、最終宛先ポートが同一SE内の場合はそのまま目的のポートに出力される。もし同一のSEでない場合は最終宛先ポートのPN(FPN)によって$(FPN/5.4) + 12$で表されるSEに送出される。第2段目のSEに入ったパケットは、既にその時点で最終段のSEは決定されている。そこで接続されている結線の数に応じてポートに振り分けられ、送出される。

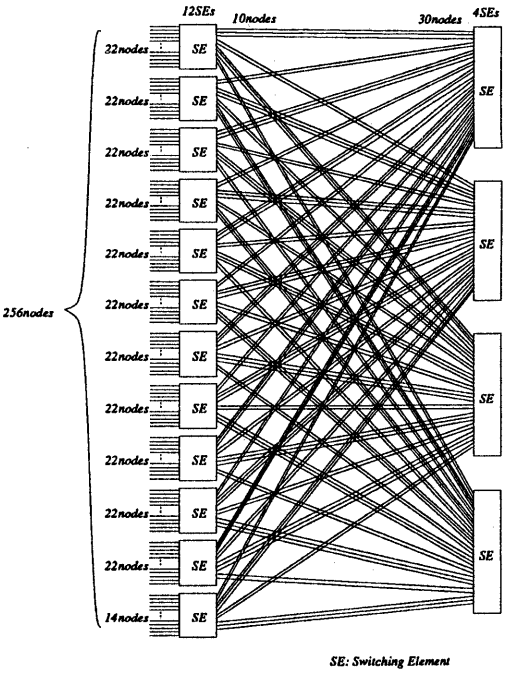

図 3: クロスバー・TREE

4.1.2 構成案2

図4に示すように、１６台のクロスバースイッチについて１６個のポート（PN0～15）を外部接続用に用い、残りの１６ポート（PN16～31）を内部接続用に用いる。そしてその内部接続用ポートを用いてクロスバースイッチをハイパーキューブ状に接続する。各SEはx、y、z、w、のそれぞれの方向のSEとは4本のポートで繋がれている。（PN16-19がx方向、PN20-23がy方向、PN24-27がz方向、PN28-31がw方向に接続されている。）

ルーティングは以下のようなものである。第1段目のSEに入ったパケットは、最終宛先ポートが同一SE内の場合はそのまま目的のポートに出力される。もし同一のSEでない場合は、SNを2進数（x,y,z,w）で

表した形で、最終目的ポートのSNと現在のSNを比べて、x y z wの順番で次元を一致させる方向に送出される。最終宛先ポートのPN(FPN)によって $mod(FPN,4)+16+k$、

k=0 (if x方向)
k=4 (if y方向)
k=8 (if z方向)
k=12(if w方向)
で表されるポートに送出される。

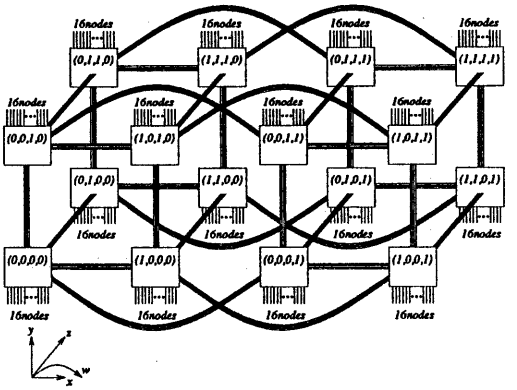

図 4: クロスバー・HC

4.1.3 構成案3

図5に示すように、16台のクロスバースイッチについて16個のポート（PN0～15）を外部接続用に用い、残りの16ポート（PN16～31）を内部接続用に用いる。そしてその内部接続用ポートを用いてクロスバースイッチをTORUS状に接続する。各SEは右、左、上、下のそれぞれの方向のSEとは4本のポートで繋がれている。（PN16-19が右方向、PN20-23が左方向、PN24-27が上方向、PN28-31が下方向に接続されている。）

ルーティングは以下のようなものである。第1段目のSEに入ったパケットは、最終宛先ポートが同一SE内の場合はそのまま目的のポートに出力される。もし同一のSEでない場合は、最初に左右方向を一致させるように、次に上下方向を一致させる方向に送出される。最終宛先ポートのPN(FPN)によって $mod(FPN,4)+16+k$、

k=0 (if x方向)
k=4 (if y方向)
k=8 (if z方向)
k=12(if w方向)
で表されるポートに送出される。

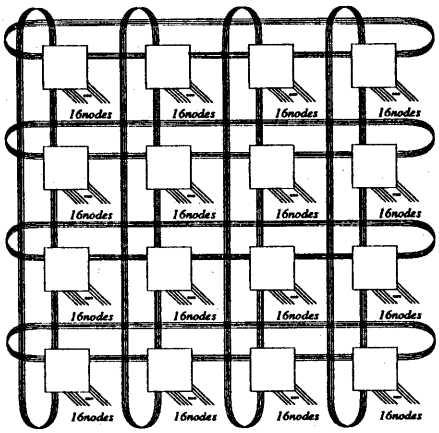

図 5: クロスバー・TORUS

4.2 ランダム通信の場合の評価

各構成に関して、ランダム通信を行った場合のレイテンシを比較した。通信パターンとしては、256の各ノード全てに最大転送容量の50%のスループットでデータを入力し、各ノードから他の255ノードに対して一様な確率でデータを転送する。図6に結果を示す。

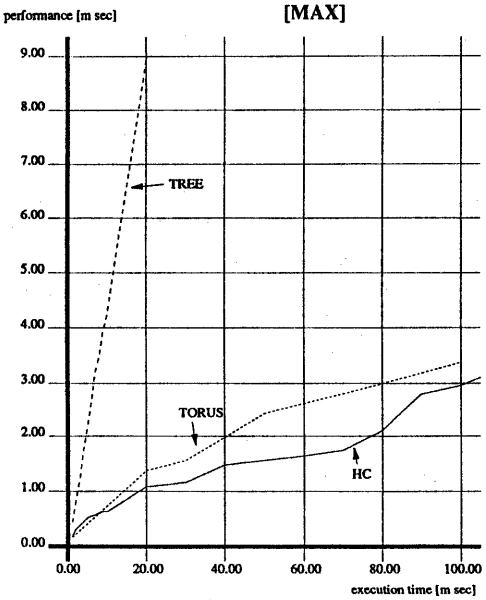

図 6: ランダム通信の場合のレイテンシ

特性は良い順にHC、TORUS、TREEである。特にTREEは劣っている。

TREEは第1層のSEにおいて、外部ノードに接続されるポート数と、内部接続用SEに接続されるポート数の比が22対10でありアンバランスになっている。このため[0～11,22～31]で表されるポートのバッファにパケットが溜り、これが性能劣化の原因の一つになっていると考えられる。また第1層のSEと内部接続用SEとの間の接続が2本の場合と3本の場合があり、特に2本で接続されているポートに他のポートと比べて30倍程度のパケットが溜っていた。

これに比べてHC、TORUSは外部接続されるポート数と内部接続用のポート数が16対16でありバランスがとれている。但し各SEには通過するだけのパケットも多く転送されるため、内部接続用のポート[*,16～31]には外部接続用のポート[*,0～15]に比べてパケットが溜る傾向がある。

4.3 画像データ混在通信の場合の評価

4.3.1 レイテンシの評価

各構成に関して、画像データ混在通信を想定した場合のレイテンシを比較した。

特定の1ノードに最大転送容量の80%のスループットで入力し、そこから1ノードを選択して全データを転送する。この際、距離の最も大きい2ノード間を選択した。また、その他のノードに対しては最大転送容量の50%のスループットでデータを入力し、各ノードから他の255ノードに対して一様な確率でデータを転送する。結果を図7に示す。

画像データの転送性能に関してはHC、TORUS、TREEの間で大きな差はない。

一方ファイルデータの性能に関してはランダム通信の場合と同様に、HCとTORUSが良くTREEが劣る。しかしTREEとHCとの性能差がランダム通信の場合は8.2倍程度であったのが、混在通信では5.4倍にまで縮まっている。これはHC、TORUSの直径が4であるのに対してTREEが2であり、ファイルデータの通信を阻害することが少なかったためと考えられる。

4.3.2 蓄積パケット量の評価

各構成に関して、画像データ混在通信を想定した場合のバッファに溜ったパケットの量を比較した。結果を図8に示す。

画像データ通信時の蓄積されたパケット量に関してはHC、TORUS、TREEの間で大きな差はない。どの場合もパケットサイズを48バイトと考えると、Iフレームを転送する際（33 ms間）にバッファに溜った量は478KBとなる。

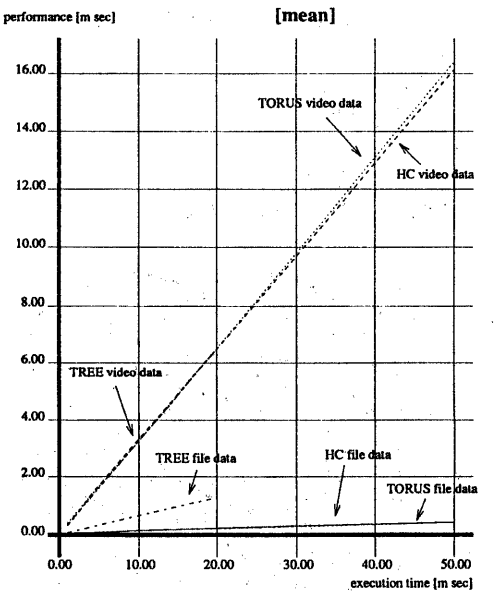

図 7: 画像データ混在時のレイテンシ

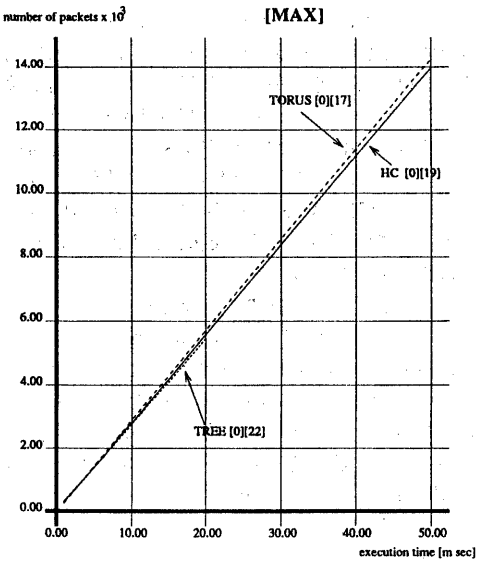

図 8: 画像データ混在時の蓄積パケット量

4.4 ファイルデータ通信の場合の評価

各構成に関して、ファイルデータ通信を想定した場合のレイテンシを比較した。通信パターンは以下の通りである。256の各ノード全てに最大転送容量の50%のスループットでデータを入力し、各ノードから送り先の255ノードのうち、特定の１２ノードに対してのみファイルデータを想定して通信量を多くし(データ量の80%)、他の244ノードに対しては残りのデータ(データ量の20%)を一様な確率で転送する。結果を図9に示す。

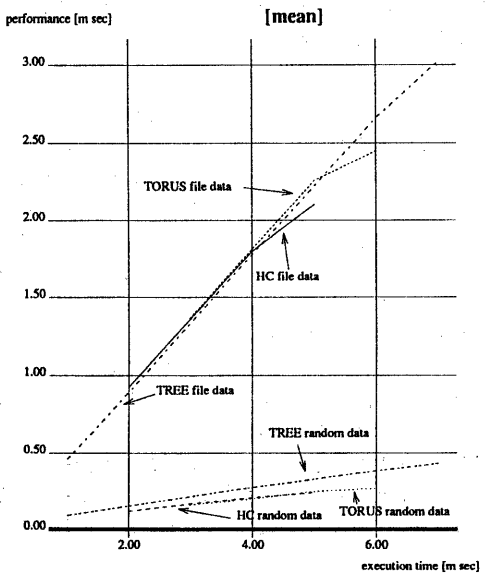

図 9: ファイルデータ通信時のレイテンシ

ファイルデータのレイテンシに関してはHC、TORUS、TREEの間で大きな差はない。しかしランダムデータの場合の性能はやはりTREEが劣っている。

4.5 上位トポロジーの決定

今回のシミュレーションで提案した階層型入出力結合網においては １０２４台プロセッサ規模の超並列計算機の場合、上位階層はHCとTORUSが性能が良く、TREEは性能が劣ることがわかった。HCとTORUSはほぼ同様な性能を有しているがランダム通信の性能についてはHCが優る。

本稿では上位階層にHCを採用することにし、これ以降はHCを用いた場合の特性について検討を加える。

5 画像データ通信の場合の改良

5.1 実際の画像データ通信

画像データとファイルデータが混在している場合について、より実際に近い状態を想定する。先にも述べたようにMPEG 2により符号化されたデータは前後関係のないIフレームと前後のデータの差分から生成されるBフレーム、Pフレームからなる。そこでIフレーム1枚に対してB、Pフレーム9枚というパターンの映像が連続する場合を想定して、シミュレーションを行う。連続画像データ時の入力データパターンを図10に示す。

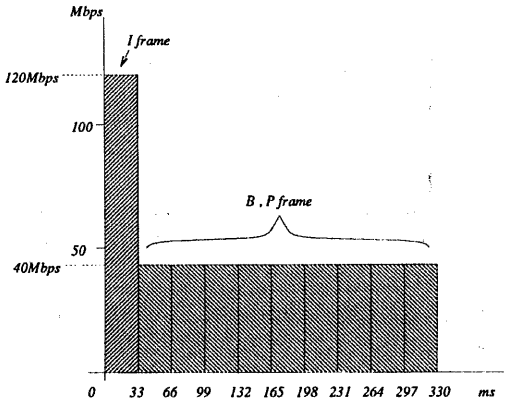

図 10: 連続画像データ時の入力データパターン

図11にレイテンシとバッファサイズの変化を示す。B及びPフレームの転送の場合もレイテンシは増え続けておりバッファに溜ったパケットも減少していない。このまま連続的にデータを送信し続けると、バッファ溢れがおこりパケットの喪失が発生してしまう。最初のSEから次のSEに向かうポートのバッファにとくにデータが溜っていることが分かった。

5.2 出力バッファの共有化

これまでは、スイッチ間のポートを選ぶ際に、宛先のアドレスから一意的に出力ポートを決定していた。しかしこの方式では画像データの流れるポートにデータが集中してしまう。そこで同一SEに向かう4つのポートのバッファに溜るパケットの量に偏りができないように、出力バッファを共有化して4つのポートに均等に振り分けることにした。4つのポートのどれを選択しても、送り先のSEは変わらない。また、この動作は各SEの内部情報を元に行われるので、SE間の通信量が増えることはない。

シミュレーション結果を図12に示す。Iフレームを転送したためにバッファに溜ったパケットはほぼ６６msの時点でなくなった。この時点で次のIフレーム

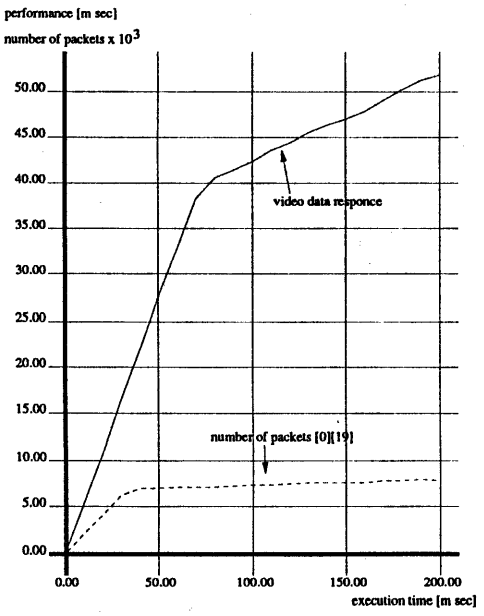

図 11: 連続画像データ時のレイテンシと蓄積パケット量

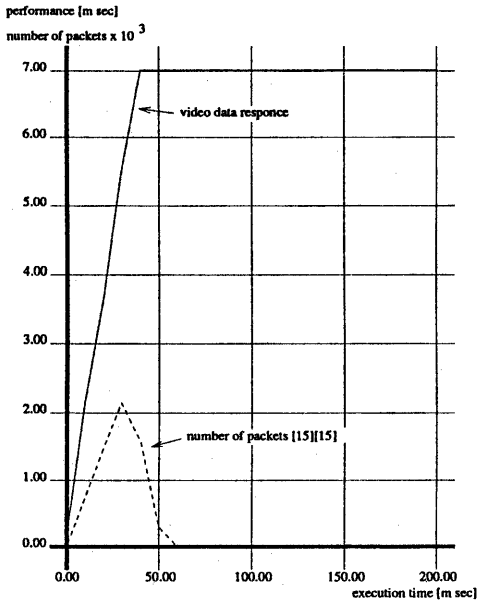

図 12: 出力バッファ共有化時のレイテンシと蓄積パケット量

を転送することが可能な状態になっている。通常の映像では、６６ msごとにシーンチェンジがある可能性は低く、バッファの共有化を行うことにより、動画像データの転送が可能になる。

6 まとめ

今回の評価により、１０２４プロセッサ規模の超並列計算機の分散型入出力システムにおいては、階層型ネットワークを採用した場合、上位階層に HC を採用した場合の性能が優れていることが分かった。

また、同一 SE に送出するポートのバッファを共有することにより、圧縮されたハイビジョン画像を転送することが可能になる。

7 今後の課題

今回の評価では、同一 SE へ接続されたポート間でバッファの共有化を行ったのみで、ルーティングに関しては固定にしている。しかし今後は性能向上のためにも、またフォールトトレラント性の向上のためにもダイナミックなルーティングの検討を行っていく予定である。また、パケットのサイズに関しても最適化を図っていきたい。

本方式を超並列計算機上に構築することにより、実機上での有効性を検証していくことも重要である。そのために入出力結合網のプロトタイプを構築することを予定している。プロトタイプの入出力結合網は 192 のノードを有し、プロセッサクラスタ 128 ノードとディスククラスタ（PDM）32 ノードを接続し、残りのノードに画像音声入出力ユニットを接続する予定である。

参考文献

[1] J.M.del Rosario, A.N.Choudhary. High-Performance I/O for Massively Parallel Computers. IEEE Computer March 1994, (1994), pp.59-68.

[2] 大西、北村、大上、清水. 超並列計算機における並列二次記憶の基本アーキテクチャ. 情報処理学会第 48 回全国大会論文集, 3B-5, (1994).

[3] 大上、北村、大西、清水. プロセス間のアクセス競合を低減する並列二次記憶システムの構想. 情報処理学会研究報告 ARC-106-7, (1994), pp.49-55.

[4] 横田、松岡、岡本、廣野、坂井 超並列向け相互結合網 MDCE の提案と評価. 並列処理シンポジウム JSPP'94, (1994), pp.121-128.

[5] 坂井、岡本、松岡、広野、児玉、佐藤、横田. 超並列計算機 RWC-1 の基本構想. 並列処理シンポジウム JSPP'93, (1993), pp.87-94.