計算機アーキテクチャ 101－6
（1993. 8. 20）

超並列テラフロップスマシン TS/1 の構想

田邊 昇　菅野 伸一　鈴木 真樹　小柳 滋

(株) 東芝 研究開発センター

1993 年 4 月より RWC(Real World Computing) プロジェクトの一環として RWC 東芝超並列研究室はピーク性能 20TFLOPS を実現可能な超並列計算機 **TS/1** の開発を開始した。本稿では **TS/1** のアーキテクチャの全体像について概要を報告する。**TS/1** は三次元実装によって接続される最大構成時 65,536 台の R4000 タイプのマイクロプロセッサと **TSC1** コプロセッサと 64MB 同期 DRAM により構成されるノードからなる。**TSC1** は (1) ピーク速度 250MFLOPS のマルチスレッデッドベクトルプロセッサ、(2) 遠隔の FIFO 型ベクトルレジスタ間のチェイニング機構 (プロセッサ間チェイニング機構)、(3)1GB/s/node のメモリバンド幅を実現する同期型 DRAM のためのブロック化メモリアクセス機構、(4) 仮想記憶をサポートした分散共有メモリアクセス機構、(5)3GB/s/node の結合網バンド幅を実現する三次元トーラス用フォールトトレラントな wormhole 型ルータを内蔵する。

Basic Features of a Massively Parallel Teraflops Machine TS/1

Noboru Tanabe　Shin-ichi Kanno　Masaki Suzuki　Shigeru Oyanagi

TOSHIBA Research and Development Center

1, Komukai Toshiba-cho, Saiwaiku, Kawasaki, 210 Japan

The massively parallel Teraflops machine **TS/1** is going to be developed by Toshiba under the project of Real World Computing(RWC) Project. **TS/1** will have up to 65,536 nodes consisted by R4000 type microprocessor, **TSC1** co-processor and 64MB Synchronus DRAM. Each nodes are connected to 3D torus by using 3D packaging technologies. **TSC1** includes (1)250MFLOPS multi-threaded vector processors, (2)mechanisms for chaining between remote FIFO vector registers (*interprocessor chaining*), (3)mechanism for new generation high bandwidth synchronus DRAMs which realize 1GB/s/node memory bandwidth, (4)mechanisms for distributed shared virtual memory and (5)3D torus fault tolerant wormhole router which realizes 3GB/s/node network bandwidth.

1 はじめに

計算機の高速化に対する要求はとどまるところを知らず、先端的な研究開発や地球環境の予測などを行っていくには 1TFLOPS 程度は必要といわれる。しかし、デバイスの高速化を頼りにした少数精鋭的なベクトル型アーキテクチャでは、この要求には対応しきれない。超並列アーキテクチャは従来型では到達し得ない性能領域を実現する最も現実的な方式と見られている。

筆者らは超並列マシン以外では実現困難な高い演算能力を要求する多くの応用で 1TFLOPS を大幅に越える高い実効性能を実現する計算機のアーキテクチャを追究し、実際に開発することを前提に実現可能性を調査・検討してきた。その中間結果として SWoPP'92 では超並列マシンの課題とその解決に対する設計思想の方向性を明らかにし、開発予定マシンのアーキテクチャの特徴的項目名のみ列挙した。[1] さらに JSPP'93 において提案マシンの最も特徴的な一部のハードウェア機構を明らかにし、これを用いた並列処理の性能見積りを行った。[2]

以上のような先行的準備研究を終え、1993 年 4 月より RWC(Real World Computing) プロジェクトの一環として、最大構成時 65,536 台において物理的・経済的現実性を備えつつ、周波数 62.5MHz 時に最大構成時ピーク性能は倍精度時に 12.3TFLOPS、単精度時に 20.5TFLOPS、65TB/s のメモリバンド幅と、三次元実装による Petabit/s オーダーの結合網バンド幅を実現可能な超並列計算機 **TS/1** の 具体的な設計を開始した。本報告では **TS/1** のアーキテクチャの全体像について、機構の詳細ではなく設計思想を中心に報告する。

2 基本方針

TS/1 の設計における基本方針を以下に示す。

- 少数精鋭型スーパーコンピュータの延長では対応できない超並列マシンが必要不可欠とみられる多くの応用を高速に処理できること
- 科学技術計算に限らず、より多くの人々の所持する応用の特性や、プログラミングスタイルの多様性を受容できる汎用性があること
- 消費電力×コストあたりの実効演算性能 (特に FLOPS 値) を最大化すること
- むやみに演算能力を上げてデータ供給に顕著なボトルネックを生じさせないこと
- むやみに周波数を上げてシステム全体の消費電力を上げないこと
- むやみに製造コストを上げて現実的なコストで実現できる FLOPS 値を下げないこと
- 高速にアクセスできる全体でのメモリ容量を演算性能に見合う分量だけ確保すること
- キャッシュによる見せかけのバンド幅ではなく主記憶からの実際のバンド幅が必要な大規模科学技術計算にもデータ供給能力がネックとならないこと（具体的には一浮動小数演算あたり最低一語のメモリアクセスバンド幅）
- 開発チップに使用するプロセスは過去に使用実績のあるものを用い、歩留まりが下がらないようにすること
- 実効性能で 1TFLOPS 以上を現実的なコスト（20〜30 億円程度）で実現できること
- 複数ユーザーでの使用を前提としてプロテクションは十分に行うこと
- 複数ユーザーでの使用時にもある程度効率的な使用を可能とすること
- 数万台規模のマシンとなることを考慮した信頼性・可用性維持対策を行うこと
- 少数精鋭型スーパーコンピュータが得意とし従来の超並列マシンが苦手とする処理の高速化の道を与え、そのようないくつかの応用においても超並列マシンの優位性を示せること
- 設計期間と設計コストを抑えるためには新規に開発するものを最小限に抑え、保有技術や既存部品を徹底的に活用すること
- 最低限の機能を持った並列計算機として動き出すまでのハードウェア・ソフトウェア設計コストと構築期間を抑え、開発チップの陳腐化を防ぐために、比較的安価な汎用マイクロプロセッサにより代替しても致命的にはならない用途にはこれを活用する。

3 TS/1 のアーキテクチャの特徴

TS/1 はマルチパラダイムの支援を意識して設計された超並列マシンであり、三次元実装によって接続される最大構成時 65,536 台の R4000 タイプのマイクロプロセッサと **TSC1** コプロセッサと 16

~64MB 同期 DRAM により構成されるノードからなる。**TSC1** は図 1に示すように (1) ピーク速度 250MFLOPS のマルチスレッデッドベクトルプロセッサ、(2) 遠隔の FIFO 型ベクトルレジスタ間のチェイニング機構 (プロセッサ間チェイニング機構)、(3)1GB/s/node のメモリバンド幅を実現する同期型 DRAM のためのブロック化メモリアクセス機構、(4) 仮想記憶をサポートした分散共有メモリアクセス機構、(5)3GB/s/node の結合網バンド幅を実現する三次元トーラス用のフォールトトレラントな wormhole 型ルータなどを内蔵する。

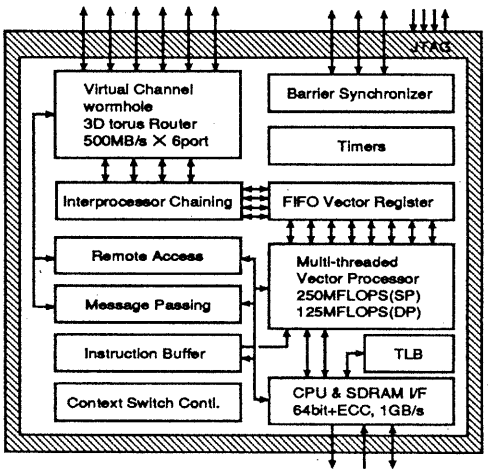

図 1: **TS/1** 構築用コプロセッサ **TSC1** の構成

3.1 マルチパラダイム支援

多くの人々によって共有されるべきテラフロップスマシンには、多くの人々の所持する応用の特性や、プログラミングスタイルの多様性を受容できる汎用性があることが望ましい。ただし普及型のテラフロップスマシンに求められる汎用性はテラフロップスの性能を要求する人々が必要または便利とする汎用性を主眼にすえるべきであって、低並列・高並列むけ応用や、テラフロップスの性能を要求する明確な応用の見えない不規則ランダム細粒度通信を伴う応用はコストの大幅な上昇を伴わない範囲でサポートすべきである。

テラフロップスの性能を要求する人々が抱える応用には乱流解析、量子色力学、地球気候環境予測、原子炉モンテカルロシミュレーション、超高精細動画像リアルタイム処理などに見られるように、非常に多くのデータ並列性を内在するケースが多い。

これらの中で乱流解析のように精度の確保のために空間メッシュの高精細化に伴うベクトル長が増大するタイプの応用にとっては、CM5 のごとくベクトル演算器が沢山並んでいることが望ましい。

原子炉中性子輸送モンテカルロシミュレーションのように精度確保のために試行回数が増大するタイプの応用にとっては、試行ごとの分岐種類が多いため高性能なスカラプロセッサが沢山並んでいて、メッセージ交換が行えることが望ましい。

またニューラルネットワークシミュレータや境界要素法を用いた応用では密行列演算が高速であることが望まれる。行列の規模とプロセッサ台数の差が少ない時は一浮動小数演算あたり数回の通信が発生するので、シストリックアレイまたはウェーブフロントアレイ的な超高速演算機構が備わっていることが望ましい。

また現在所持しているプログラムになるべく手を入れたくないというユーザーには共有メモリ機構によるプログラミングまたはコンパイラへの負担軽減がなされることが望ましい。

以上の見地から **TS/1** では分散メモリ型並列ベクトル、メッセージ交換、ウェーブフロントアレイ、分散共有メモリといった複数のパラダイムを支援する機構を同一マシン上に実現する。

1TFLOPS を越えるクラスの超並列マシンにおける現実的なコストで性能的に全く問題のないマルチパラダイム支援は困難と思われるが、妥協できる範囲でのマルチパラダイム支援の実現形態としては、以下の三つのアプローチが提案されている。

- J-machine[3]、RWC-1[4] などのアプローチ
  汎用マイクロプロセッサの活用を諦め、全面的に新規なプロセッサを開発し、結合網やメモリのアクセスレイテンシ・バンド幅等を高めに設定した作りとし、命令から起動される不規則ランダム細粒度処理支援機構により、全ての応用プログラムをその実行モデルに落し込む。粒度が極めて小さい応用の実行時や、バンド幅やプロセッサ内並列処理資源等が十分に確保できない場合は性能の低下を許す。
- DASH[5]、D-machine[6] などのアプローチ
  汎用マイクロプロセッサを全面的に用い、分散共有メモリ機構により全てのプロセッサ間

通信をメモリアクセスで扱う実行モデルに落し込む。その変形としてはコヒーレンス管理だけでなくページ管理等のOS機能の実行を代行するコプロセッサを設ける場合もある。汎用マイクロプロセッサと分散共有メモリでは究極的な性能が得にくい場合を残す。

- **TS/1のアプローチ**
  汎用マイクロプロセッサを部分的に用い、数値演算コプロセッサを併用して演算性能を補い、このコプロセッサに入る範囲内で分散共有アクセス機構やプロセッサ間チェイニング機構、メッセージ交換支援機構といった多様な通信機構を設け、複数のプロセッサ間通信モデルを適材適所で使い分け、これらでカバーできない場合を残す。

3.2 S-DRAMアクセスブロック化機構

消費電力あたりのコストパフォーマンスの最大化にはCMOSのVLSIが不可欠であり、現在のハイエンドCMOSプロセッサが150MFLOPS程度なので、近い将来に10TFLOPSを実現する超並列マシンの要素プロセッサ(PE)数は数万が妥当であると考えられる。この規模を現実的なコストにおさめるためにはPEあたり10〜20万円以下で作る必要があると考えられ、少数のDRAMとCMOSプロセッサによる簡素な構成とせざるをえない。

EWSではDRAMによる脆弱な主記憶を二次キャッシュでカバーすることでコストパフォーマンスを高めているが、主記憶の脆弱なEWSほど顕著に観測されるように大規模科学技術計算ではキャッシュに乗り切らない配列を薄くアクセスすることが多いためキャッシュの有効性が低い。科学技術計算の高速化を重要視する超並列マシンにおいて、二次キャッシュはコストと実装面積と消費電力を増加させるために避けるべきで、二次キャッシュに基盤を置く方式は消費電力あたりのコストパフォーマンスを最大化する必要の無い高並列向けの方式と言わざるをえない。

一方、従来のスーパーコンピュータでは集積度の低いSRAMやインタリーブ構成にした大量のメモリチップを使用することにより主記憶そのもののバンド幅を確保してきたが、このアプローチでは厳しいコスト制約のもとで10TFLOPSに見合った主記憶バンド幅と容量を確保できない。

従来のベクトルプロセッサではワード単位にアドレスが計算され、細切れのアクセス要求がインタリーブドメモリに供給されるが、**TS/1**では、連続アクセスを行うベクトルロードストア命令ではアクセスはブロック化され、アドレス計算の負担を軽減させつつ、長いブロック長による次世代DRAMの高いメモリバンド幅を引き出す。

次世代高バンド幅DRAMとしてはRambus型DRAM(R-DRAM)[7]と同期型DRAM(S-DRAM)[8]が市場に送り出されつつある。Rambusは必要なピン数が少ないので複数ポートを1チップに持たせることは通常のバスに比べれば容易であると考えられ、現実的な妥協点として2ポートのRambusをチップに持たせると最大1GB/sのメモリバンド幅が得られる。

このメモリバンド幅は125MFLOPS倍精度（8バイト）または250MFLOPS単精度（4バイト）のデータを演算1回につき1語メモリアクセスする能力にあたり、ベクトルプロセッサにおいて必要なメモリバンド幅対演算性能の比率[9]と一致する。

ところが、R-DRAMはデータが500MHzで入出力されるため、基板上配線パターンやパッケージ形状に厳しい制約があり、現時点で使用実績のある最高ピン数を持つ820ピンButt-PGAパッケージを使用できない可能性が高いことがわかった。

これに伴い再検討した結果、プロセッサ間通信バンド幅の維持は後述するウェーブフロントアレイ処理では重要であり、代替案であるS-DRAMはR-DRAMより従来のメモリに近い特性を持つので性能の変動幅が小さく、特に降順連続アクセスなどの性能がR-DRAMより優れることもあり、**TS/1**では同期型DRAMを採用することとした。

S-DRAMは通常のDRAMと類似した信号構成をしており多くのアドレス線、制御線が必要であり、データ転送周波数が現在の最大で100MHzとR-DRAMの1/5であるので、R-DRAMより多くのピン数を必要とする。しかし、データバス幅を64bitとして汎用マイクロプロセッサのバス幅に合わせて共用とすることによってMPUとベクトルプロセッサの並列処理時の性能低下を許容すれば、R-DRAM採用時とほぼ同数のピン数に納めることが可能である。

64bit幅のデータバスでS-DRAMにより1GB/sのバンド幅を実現しようとすると125MHzバージョンのS-DRAMの出現を待つ必要があるが、現時点で

100MHz のバージョンがサンプル出荷され、125MHz のバージョンが学会発表されている現状を考えると、数年後には 125MHz バージョンの出荷の可能性は高いと思われる。よって TS/1 ではメモリポートは 125MHz の S-DRAM と 62.5MHz の MPU が共通バスに接続可能なチップを開発する。

3.3 マルチスレッデッドベクトル機構

TS/1 の実際の使用では S-DRAM 内部バンクコンフリクトの影響で 1GB/s よりは若干性能低下が見込まれる。しかし、図2に示すように実パイプライン数を越える複数のベクトル演算命令を起動して時分割にパイプラインを割り当て見かけ上多くのパイプラインで並行処理されているように制御するマルチスレッド制御や、FIFO 型のベクトルレジスタをベクトル演算機構に導入することにより、メモリバンド幅が有効利用され実質的に約二倍のメモリバンド幅があるものと同等の処理性能が得られることが知られている。[10]

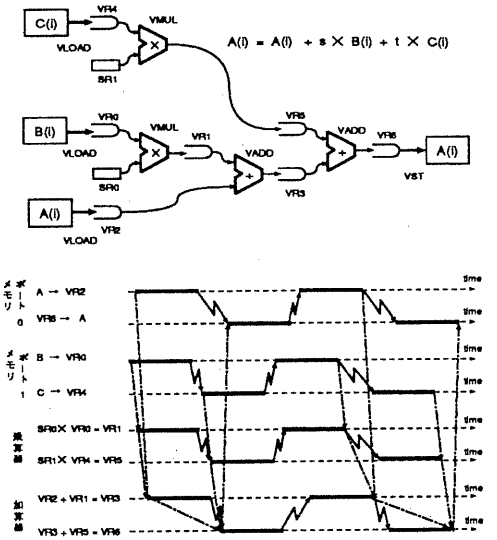

図 2: マルチスレッデッドベクトル機構の動作

これにより単体性能としては連続アクセス時には乗算と加減算の比率が等しい（例えば Livermore Loop #3（内積）や#7（状態方程式）など）場合、倍精度で 125MFLOPS のピーク性能に近い実効処理性能が得られることが見込まれる。

またマルチスレッデッドな制御を行うことにより、論理的に同時に実行できる命令の個数を増やすならば、以下に述べるプロセッサ間チェイニング機構を用いた並列処理を効率的に適用できる応用の範囲を広げることができる。例えばプロセッサ間チェイニングにより左右のプロセッサから1本ずつのベクトルデータをもらって、自分のプロセッサが持っているベクトルに加えるならば1結果データあたり2回加算を行わなければならないが、物理的な加算器が1個しかなければこれらをチェイニングにより実行できないため、受信した2本のベクトルを加算しメモリに部分和のベクトルを形成してから、部分和ベクトルをロードしなおして最終結果を得るベクトル加算を行う等のオーバーヘッドが入り込む。一方マルチスレッデッドな制御を行うならば両方のベクトル加算がチェイニングにより実行できるのでオーバーヘッドが少なく、メモリアクセスは最小限に抑えられる。

なお、TS/1 のベクトルプロセッサはスーパースカラプロセッサ用に開発済みの 80MHz 動作時に単精度 320MFLOPS、倍精度 160MFLOPS の浮動小数演算器 [11] にベクトルレジスタ間のデータ経路増設のための改良を加えて実現される。

3.4 プロセッサ間チェイニング機構

要素数がプロセッサ数と桁違いにならない中小規模の行列に対して要素がばらまかれた状態から行列間の乗算を行うような場合を想定すると、短いメッセージによる激しいデータ移動を伴うため、同期と通信のオーバーヘッドの軽減が必要になる。SIMD 型マシンにおいてはグローバルにクロックを厳密に合わせるという制約と引き換えに同期に対するオーバーヘッドはないために通信オーバーヘッド（主に通信バンド幅）のみがこの種の処理においては問題になるが、MIMD 型並列マシンではデータの同期に関しても十分な配慮が必要になる。また計算量の観点からメインとなる処理が並列化によって高速化された場合、総和などの大域演算が実行時間において顕在化してくる。この大域演算も通常、規則的な細粒度通信を伴う処理であり、高速化が望まれる。

そこで TS/1 では MIMD 型におけるそのような問題の解決のために、プロセッサ間チェイニング機構を利用したウェーブフロントアレイ [12] 動作の実現をする。ウェーブフロントアレイとはデータ転送と演算がパイプライン的に処理され、演算に必

要なデータの到着により演算が開始されるデータフロー的な低コストな同期原理に則った方式である。1クロックに複数の同期と複数の送受信が可能であるため並列演算パイプラインや多数の通信リンクを有効活用できる。このことはLSIパッケージピンネックの観点からプロセッサとルータを同一VLSI上に集積することによってはじめて現実的となる。

TS/1 におけるプロセッサ間チェイニング機構は、FIFO型ベクトルレジスタのデータ存在信号を用いたマルチスレッデッドベクトル演算機構に、図3に示すようにFIFO直接メッセージ送受信機構を付加することにより実現され、ベクトル送信命令により指定されたFIFOレジスタ上のデータに予め決められたメッセージヘッダーを付加して結合網へ送信し、そのヘッダー情報に基づき指定されたリモートプロセッサの指定されたFIFOにハードウェアがデータ転送を行う。よって通信実行時のソフトウェアオーバーヘッドが排除される。受信側では通常のベクトル演算命令間のチェイニングの場合と同様に、FIFOのデータの存在によって演算が実行される。

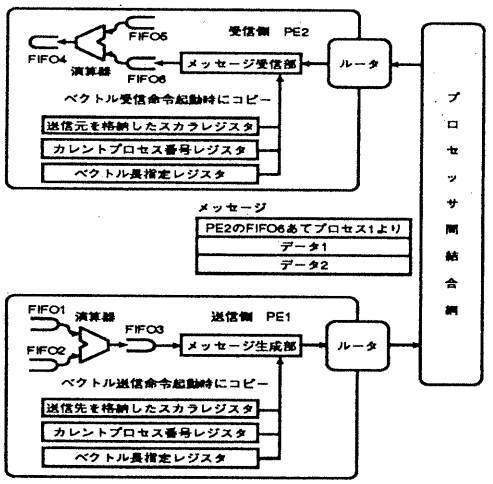

図3: プロセッサ間チェイニング機構

3.5 仮想記憶と分散共有アクセス機構

TS/1 における分散共有アクセス機構の導入の目的を以下に示す。

- プロセッサ間チェイニングだけではカバーしきれない種類の細粒度通信における通信オーバーヘッドを短縮すること
- プログラマが陽に送受信の記述をしなくても良い簡便な通信手段を提供すること
- ローカルに実装されているメモリ量を越える記憶空間を提供すること

超並列マシンではプロセッサあたりの主記憶容量に限りがあるため、ローカルメモリに対する物理アドレスによるアクセスしか許さなければユーザプログラム領域、ユーザデータ領域、システム領域を合わせて単体プロセッサの実メモリ容量に制約されてしまう。この制約によりプログラムに余計な通信指示の記述を強要したり、場合によっては対処不能な状況に陥ったりすることがあり、大規模なデータを扱うことが多いテラフロップスの性能を要求するような問題を扱うに当たって仮想記憶は重要である。

また多くのユーザにより共同利用されるべきテラフロップスマシンにおいてはマルチユーザーマルチタスク環境となり、資源のプロテクションが必要で、空間分割だけでなく時分割利用の場合は特にその柔軟な実現においても仮想記憶が有効である。

TS/1 では仮想記憶の管理は汎用マイクロプロセッサ上で走行するOSによりインプリメントされる。OSは汎用マイクロプロセッサ内蔵のTLBとベクトルプロセッサの搭載されるコプロセッサLSI上のTLB(CTLB)の双方を管理する。汎用マイクロプロセッサとしては64bitのRISCプロセッサであるR4000系のチップ(R4400PCなど)を採用する。R4000は仮想空間が1TBと広く、物理アドレスが36bitありSPARCやAlphaなどに比べて大きいという技術的な理由と、東芝社内でも製造販売を行っているという経済的な理由による。

各PE上に実装されているメモリはページングされ、これらはローカル領域と分散共有領域に区別され、そのうち分散共有領域として割り当てられたページはキャッシュ不可能領域という属性でR4000の各プロセスの仮想空間(1TB)内に寄せ集められてマップされる。R4000に内蔵されるわずかばかりの一次キャッシュは主にユーザプログラムのR4000用命令およびOSの命令およびデータ領域の高速化のために利用される。

分散共有アクセス機構はR4000の外部にあるコプロセッサLSIに実装されるので、メモリアクセス時のアドレス出力としてアドレスを分散共有アクセス機構に伝えるには物理アドレス空間の大きさが問題となる。R4000は物理アドレスが36bitに過ぎ

ないので64GBしかアドレスすることができず、ソフトウェアの介在(すなわちデータとしてのアドレスの分散共有アクセス機構への伝達)を全く無しに64GBを越える空間に対するあらゆるリモートアクセスを実現することは不可能である。

このように既存の汎用マイクロプロセッサを使う上では上記のような問題があり、例えばVR3000を用いたCenju2[13]でも同様な問題に遭遇している。Cenju2ではリモートアクセスのたびに割込み処理となる高オーバーヘッドなインプリメントを行っているために分散共有アクセス機構を使ったのでは場合によってはアプリケーションが高速化できず、結局DMAを用いたブロック転送に基づくメッセージ交換式にプログラムを書き換えて高速化を達成したという報告もなされている。このように常時割込み処理方式はプロセッサ間チェイニングではカバーしきれない細粒度通信を高速化するという**TS/1**における分散共有アクセス機構の導入の目的から外れるため、採用しない。

TS/1ではこのような問題を避けるためにやむをえずセグメンテーションの機構を導入し、R4000外部に出力される36bitの物理アドレスにより極力リモートアクセスを済ませる。セグメント内でのアクセスを続けている間はソフトウェアオーバーヘッド無しでリモートアクセスを行い、別のセグメントをアクセスするときはコプロセッサ内部のアドレス拡張用レジスタを書き換えてから行う。主記憶の合計が64GB以下に納まる中小規模の構成の**TS/1**ではセグメント導入による性能低下は防止できる。

3.6 三次元実装による高バンド幅結合網

プロセッサ間チェイニング機構を効率的に動作させるには通信バンド幅の確保が必要である。つまり実行時の同期や通信のオーバーヘッドが排除されているので演算器が十分に高性能であれば、中小規模の行列乗算などでは通信バンド幅がウェーブフロントアレイ動作の実効性能を決定する。[2] ここでいう通信バンド幅は単にルータに出入りするプロセッサ間通信リンクのバンド幅のみならず、ルータと演算器との間のバンド幅も含み、これらにネックがないことが究極的な性能を実現するには必要となる。

HPFなどのデータパラレル言語は超並列マシンによる科学技術計算における主流となる兆しがあるが、これらをベースとしたプログラミングではメッシュを仮想的なハードウェアイメージとしてプログラムすることが多く、メッシュ系の通信が全ての仮想プロセッサで発生した時に最も遅い部分のバンド幅が処理性能を決定する危険がある。ゆえにデータパラレルという超並列マシンにとって極めて重要なプログラミングパラダイムをサポートするためにはメッシュのエミュレーションが高速に行われる必要がある。このことは結合トポロジーに向き不向きがあり、クラスタ化された結合トポロジーをとった場合クラスタ間通信バンド幅を十分に確保しなければならないことを意味する。

また分散共有アクセス機構やメッセージ交換支援機構を用いた知識処理や並列I/Oに対するアクセスでは、ランダム通信が発生しうる。細粒度なものは分散共有アクセス機構、疎粒度なものはメッセージ交換支援機構によって通信が起こるが、プロセッサ能力の強化による処理時間の短縮と通信ハードウェアの工夫による通信オーバーヘッド削減により、結合網にメッセージが押し出される頻度が上昇するため、結合網におけるランダム通信時の実質バンド幅が高いことが要求される。

以上の見地から**TS/1**では三次元実装によりPEあたり六方向に各々約350本の長距離配線を排除した拡張性のある三次元トーラス型基板間配線を実現し、演算器と同等な周波数で演算レートと同程度の通信リンクバンド幅を実現する。350本のうち半分はグランドとして利用され、残りの半分の半分は一個飛び基板間配線のために用いられる。よって各方向あたりの実質的な信号数は80程度となり、64bit幅データ8bitECC付き半二重通信路で構成される三次元トーラス結合網または同等スペックの全二重通信路で構成される三次元メッシュ結合網の実装が可能である。ただし現在当社で使用実績のある最大ピン数のパッケージである820ピン表面実装型PGAを用いたとしてもピンネックとなるため全二重のメッシュとはしない。なお上記の実装方式はプロトタイプによる実験を既に行ってある。

三次元トーラスは二次元トーラスをエミュレーションできるとともに、結合網の直径が少ないため通信レイテンシを少なくでき、二分割バンド幅を向上できるのでランダム通信性能も優れるが、base-m n-cube[15]やbinary n-cubeに比べてトポロジー的には直径が大きくなることは否定できない。

しかし同一の二分割バンド幅を仮定した場合binary n-cubeは低次元トーラスよりレイテンシが大

きい。[14] 数万ノード規模の拡張性を考慮した場合 binary n-cube ではパラレル通信路による実装が困難で、パラレルシリアル変換におけるレイテンシの増加もあり好ましくない。

base-m n-cube ならば例えば base-16 4-cube のように 65,536 台のノード数とパラレル通信路の実装が可能と考えられるが、次数 n が 2 を越えると大量の長距離配線 (ケーブルとドライバ) とクロスバスイッチが発生するため、同一の実装コストにおける通信バンド幅は三次元実装に基づく三次元トーラスには及ばないものと考えられる。

大きな直径による静的なレイテンシの増加は通信時期の最適化 (データ先取り) やマルチスレッディングにより解決できることがある [16] が、バンド幅の不足に伴う性能劣化にはアルゴリズムを変更して通信量を抑える以外の解決策がない。よって **TS/1** では静的レイテンシ短縮よりバンド幅の強化に重点を置く三次元トーラスを採用する。

TS/1 のルータは複数の仮想チャネルを有効に利用してデッドロックフリーな故障迂回と通信性能向上をはかるものであるが、詳細は別途報告する。

4 おわりに

本報告では **TS/1** の設計における基本方針を示し、これらを考慮して設計された **TS/1** における特徴的なアーキテクチャについて概略を説明した。

TS/1 は三次元実装によって接続される最大構成時 65,536 台の R4000 タイプのマイクロプロセッサと **TSC1** コプロセッサと 16〜64MB 同期 DRAM により構成されるノードからなる。**TSC1** は (1) ピーク速度 250MFLOPS のマルチスレッデッドベクトルプロセッサ、(2) 遠隔の FIFO 型ベクトルレジスタ間のチェイニング機構 (プロセッサ間チェイニング機構)、(3)1GB/s/node のメモリバンド幅を実現する同期型 DRAM のためのブロック化メモリアクセス機構、(4) 仮想記憶をサポートした分散共有メモリアクセス機構、(5)3GB/s/node の結合網バンド幅を実現する三次元トーラス用フォールトトレラントな wormhole 型ルータなどを内蔵する。

なおメッセージ交換支援機構、コンテクストスイッチ支援機構、同期機構、入出力機構、高信頼化技術、および超並列ソフトウェアに関しては別途報告する予定である。

参考文献

[1] 田邊, 小柳:"三次元実装に基づくマルチパラダイム超並列テラフロップスマシンのアーキテクチャ", 電子情報通信学会技術報告 Vol.92,No.173(SWoPP 日向灘'92), CPSY92-24, pp,33-40(1992.8)

[2] 田邊:"マルチパラダイム超並列 TFLOPS マシンにおける並列処理 〜プロセッサ間チェイニングとその応用 〜", 並列処理シンポジウム JSPP'93, pp,79-86(1993.5)

[3] W. J. Dally et al. : "The J-Machine : A Fine-Grain Concurent Computer", Proc. of IFIP Congress, pp.1147-1153 (1989.8)

[4] 坂井 他:"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, pp,87-94(1993.5)

[5] D. Lenoski et al. : "The Stanford Dash Multiprocessor", IEEE computer, Vol.25, No.3, pp.63-79 (1992.3)

[6] 松本, 平木:"Memory-Based Processor による分散共有メモリ", 並列処理シンポジウム JSPP'93, pp,245-252(1993.5)

[7] 串山, 大島, 古山:"500M バイト／秒 Rambus 仕様 4.5M ビット DRAM", 東芝レビュー, Vol.47, No.7, pp575-578 (1992.7)

[8] 安保:"見えてきたシンクロナス DRAM の仕様, 100MHz 動作品が 1993 年に市場へ", 日経エレクトロニクス, 1992 年 5 月 11 日号, pp.143-147(1992.5)

[9] 古勝, 渡辺, 近藤:"最大性能 1.3GFLOPS, マシン・サイクル 6ns のスーパーコンピュータ SX システム", 日経エレクトロニクス, 1984 年 11 月 19 日号,pp.237-272 (1984.11)

[10] 橋本, 村上, 弘中, 安浦:"マイクロベクトルプロセッサ・アーキテクチャ〜演算スループットとメモリ・バンド巾との関係〜", 電子情報通信学会技術報告 Vol.92, No.173, CSPY92-21, pp.9-16(1992.8)

[11] N,Ide et al. : "A 320MFLOPS CMOS Floating-Point Processing Unit for Superscalar Processors", Proc. of Custom Integrated Circuit Conference (CICC) '92, p30.2 (1992.5)

[12] S.Y.Kung : "On Supercomputing with Systolic / Wavefront Array Processors", Proc. of the IEEE, Vol.72, No.7, pp.867-884 (1984.7)

[13] 加納, 中田, 奥村, 大竹, 小池:"並列マシン Cenju2 上の有限要素法による非線形変形解析", 並列処理シンポジウム JSPP'93, pp,379-386(1993.5)

[14] W. J. Dally : "Performance Analysis of k-ary n-cube Interconnection Networks", IEEE Trans. Computer, Vol.39, No.6, pp.775-785 (1990.6)

[15] N. Tanabe et al. : "Base-m n-cube : High Paerformance Interconnection Networks for Highly Parallel Computer PRODIGY", 1991 Int'l Conf. on Parallel Processing, pp. I 509-516 (1991.8)

[16] 平木, 島田, 関口:"細粒度並列処理におけるレイテンシ隠蔽効果の評価", 並列処理シンポジウム JSPP'93, pp,15-22(1993.5)