超並列テラフロップスマシン TS/1 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——

田邊 昇†

数万台規模のプロセッサ数で数 TFLOPS の実効性能を実現する超並列マシンを構築するためには，高集積なプロセッサと DRAM をベースとした安価な構成により，演算性能・通信性能・同期性能・メモリバンド幅・メモリ容量をハイレベルでバランスさせうるアーキテクチャが重要である．将来の更なるプロセッサ数増加に対応するためには大規模超並列マシンで小さめの問題を扱った時に見られる細粒度処理の高速化も重要となる．このような課題を踏まえ，著者らは Real World Computing Partnership の一貫として，遠隔 FIFO 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）と次世代 DRAM 向けアクセスブロック化機構を備えた 250 MFLOPS の 1 チップマルチスレッデッドベクトルプロセッサを 3 次元実装に基づく高バンド幅結合網で最大 65,536 台結合する超並列マシン **TS/1** を開発中である．本論文では，上記の機構の効果を推定するために，典型的な応用に対する性能評価を行った．その結果として，演算性能だけを上げた時に通信ネックとなる従来の超並列マシンが苦手としていた細粒度型応用（中小規模行列の乗算・軸選択付き LU 分解）で，商用スーパーコンピュータを桁違いに上回る性能が実現できることを示す．さらに演算性能だけを上げた時にメモリバンド幅ネックとなる粗粒度型応用 (2-colored SOR) で，高性能な浮動小数点演算器がピーク性能に近い性能で動作しうることを示す．

Parallel Processing on a Massively Parallel Teraflops Machine TS/1

—Interprocessor Chaining and Its Applications—

NOBORU TANABE †

This paper shows a massively parallel architecture for over Teraflops sustained performance based on cheaper configurations with highly integrated micro vectorprocessors and DRAMs. The architecture will be applied to a massively parallel Teraflops machine **TS/1** which is going to be developed by Toshiba under the project of Real World Computing Partnership (RWCP). **TS/1** will have 65,536 of 250 MFLOPS single chip multi-threaded vector processors with the mechanisms for chaining between remote FIFO vector registers (*Interprocessor chaining*) and for new generation high bandwidth DRAMs, which are connected by high bandwidth network using 3D packaging technologies. The *interprocessor chaining* will be important in the future massively parallel computers with larger number of processing elements which need high performance fine-grain processing. The effect of the *interprocessor chaining* is evaluated by estimating the performance of **TS/1** with typical applications. The result shows that **TS/1** has extremely higher performance than commercial supercomputers for communication constrained fine-grain-applications (Multiplication and LU decomposition with pivoting for small matrices) and memory bandwidth constrained coarse-grain-application (2-colored SOR).

1. はじめに

科学技術計算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等を行っていくには 1 TFLOPS 程度は必要といわれる[1]．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．

つまり，超並列マシンを将来必要不可欠なものと認識させている応用は科学技術計算であり，一方では柔らかな情報処理の担い手として有望視されるニューロ処理も密行列計算に帰着され，超並列マシンの実用化においては数値計算の高速化は避けて通れない．ニューロ処理と知識処理の融合を含む柔らかな情報処理や

† 新情報処理開発機構 RWCP 超並列東芝研究室（東芝研究開発センター内）
Real World Computing Partnership Massively Parallel Systems TOSHIBA Laboratory

モンテカルロ法への適用性を考慮するならばMIMD型をベースとすべきで，それ以外の重要な応用においてMIMD型の欠点（同期と通信のオーバーヘッド）を軽減することが重要である．

従来の商用超並列計算機は，プロセシングエレメント（PE）数の数倍以下の要素数を持つ行列乗算や密行列の軸選択付きLU分解の際などに現れるグローバル演算や放送を，通信バンド幅の不足や通信オーバーヘッドが大きいという理由で不得意としていた．このような処理ではPE数の少ないベクトル型スーパーコンピュータのほうがはるかに効率が良く，特に中小規模の問題では絶対的性能の上からもベクトル型スーパーコンピュータのほうが速いということもあった[2]．この問題点はPE数を増やすほど，またPE単体性能を上げるほど厳しくなる問題点であるだけに，単にPE数を上げたり演算器の性能を上げたりするだけでは解決しない．

実験レベルの従来のMIMD型並列マシンの中にはJ-machine[3]やEM-4[4]のように細粒度処理を意識して，命令による通信の起動と，メッセージによる命令列の起動と，高い通信バンド幅を持つものもあるが，これらに浮動小数演算器を導入して演算を高速化した際には，上記のような処理においては十分な粒度が確保できず，PEをまたがるたびに積み重なる命令起動オーバーヘッドが陽に現れてきてしまい性能低下を免れない．

一方，超並列マシンが1〜10 TFLOPSクラスの実効性能を現実的なコストで実現する上で重要な第二の課題は，特にPEにおけるベクトル処理が有効な粗粒度な処理において演算能力に見合ったメモリバンド幅をいかに低コストで実現できるかという点である．半導体技術の進歩に伴い演算性能は順調に高速化し続けると思われるが，高い浮動小数点演算能力を持つマイクロプロセッサは大容量の外付けのキャッシュの使用や，インタリーブドメモリや高速SRAMによる主記憶を前提としているために，これらを用いた数万台以上のPEを持つ超並列マシンはコスト的現実性を欠いていた．

本論文では，まず科学技術計算において重要な基本処理を示し（第2章），その多様性を考慮して設計された超並列テラフロップスマシン**TS/1**におけるマルチパラダイムサポートアーキテクチャ群の一部であるベクトル処理パラダイム（粗粒度）支援機構とウェーブフロントアレイ処理パラダイム（細粒度）支援機構について説明する（第3章）．

第4章では第3章で示された機構を用いて第2章で示された応用の中から細粒度処理の高速化が重要な応用（中小規模行列の乗算と軸選択付きLU分解）や，粗粒度処理時のメモリバンド幅確保が重要な応用（2-colored SOR）で，**TS/1**が商用超並列マシンや商用スーパーコンピュータを桁違いに上回る性能が実現できる見通しを示す．

2. 科学技術計算において重要な基本処理

科学技術計算は最終的には以下の基本処理に帰着されることが多く，これらの処理を多数回繰り返し行うことになる[5]．つまり，以下の処理の大半が高速で行えるということが望ましい．

（1）規則的疎行列を係数とする連立一次方程式の求解（反復法）：差分法による流体，熱，デバイスSim.，量子色力学など

（2）帯行列/密行列を係数とする連立一次方程式の求解（直接法）：有限要素法や境界要素法による電磁界解析，構造解析など

（3）行列加減算，乗算などの基本線形代数：ニューラルネットワークシミュレーション，画像処理など

（4）高速フーリエ変換（FFT）：分子化学，粒子，信号解析，画像処理，流体など

（5）固有値計算：分子化学など

（6）乱数を用いた試行（モンテカルロ法）：希薄気体，中性子輸送，薄膜成長，デバイスSim.など

（7）ランダム疎行列を係数とする連立一次方程式の求解（直接法）：回路シミュレーションなど

本論文では上記の（1）〜（3）についての並列処理方針とその性能に関する予備検討について述べる．これらは通信パターンや同期位置が単純かつ静的に読み切れるので，机上での計算でかなりの信頼性のある予測を行うことができる．

3. 超並列マシン TS/1 のハード機構

1993年4月より通産省大型プロジェクトReal World Computing Partnership（RWCP）の下で東芝の研究テーマとして超並列マシン**TS/1**の開発を開始した．その全体的なアーキテクチャ[6]はSWoPP'92で示した方向性[7]をベースに，SWoPP'93において発表した．周波数62.5 MHz時に64 K台の最大構成ピーク性能は12.3 TFLOPS（倍精度）20.5 TFLOPS（単精度）で，65 TB/sのメモリバンド幅と，三次元実装によるPetabit/sオーダーの超高速結合網により，汎用性を維持しつつ，コストと消費電力あたりの実効性能の最大化の現実的なコストでの達成を目標としている．

TS/1 では第2章で述べた科学技術計算における応用の多様性に柔軟に対応するハードウェア機構を提供する．本章では今回の評価に関係のあるベクトル処理パラダイム支援機構（次世代DRAM用アクセスブロック化機構，マルチ・スレッデッド・ベクトル演算機構）とウェーブフロントアレイ処理パラダイム支援機構（プロセッサ間チェイニング機構）[16]について概略を説明する．

3.1 次世代DRAM用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が不可欠であり，現在のハイエンドCMOSプロセッサが150 MFLOPS程度なので，近い将来に10 TFLOPS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．科学技術計算向け超並列マシンにおいては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10 TFLOPSに見合った主記憶バンド幅と容量を確保できない．

そのような課題の解決のために **TS/1** では「次世代高バンド幅DRAM用アクセスブロック化機構を備えたベクトルプロセッサ」を導入する．次世代DRAMであるRambus型DRAM（R-DRAM）[8]や同期型DRAM（S-DRAM）[9]は大きなブロック長の連続アクセスほど高い性能が実現できるので，連続アクセスが必然的に起こりやすいプロセッサアーキテクチャを採用することが望ましいが，連続アクセスとなりやすく，かつ1つのベクトルレジスタのサイズは通常のスカラ型マイクロプロセッサの内蔵キャッシュのラインサイズやRambusが有効に働くアクセス長と比較して大きいために，ベクトル演算機構はR-DRAM等の次世代DRAMの高い性能を引き出す必然性を備えている．

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** では主記憶にR-DRAMまたはS-DRAMを想定し，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高バンド幅を引き出す．

★ Rambus型DRAMによるメモリバンド幅

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1 GB/sのメモリバンド幅が得られる．このメモリバンド幅は125 MFLOPS倍精度（8バイト）または250 MFLOPS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[10]と一致する．実際はR-DRAMのアクセス遅延（読みだしセンスアンプキャッシュヒット時48 ns，読みだしセンスアンプキャッシュミスヒット時220 ns[8]）のため，長いベクトル長に対して見積もられるメモリバンド幅は以下のようになる．

連続アクセス時：2 port ×1024 byte/(2 ns × 1024+220 ns+48 ns×3)=849 MB/s

等間隔アクセス時

最良：2 port×512 byte/(2 ns×512+220 ns+48 ns×63)=240 MB/s

最悪：2 port×8 byte/(2 ns×8+220 ns)=68 MB/s

リストアクセス時

最良：2 port×512 byte/(2 ns×512+220 ns+48 ns×63+256 byte/0.849 GB/s)=224 MB/s

最悪：2 port ×8 byte/(2 ns ×8+220 ns+4 byte/0.849 GB/s)=66 MB/s

非連続アクセスではインタリーブドメモリに比べて性能低下は免れないが，上記の性能が最低2チップのメモリチップのみで実現できる点で優れるので，超並列マシンの要素プロセッサ用メモリアーキテクチャとしては許容範囲と言えよう．

★同期型DRAMによるメモリバンド幅

100 MHz版S-DRAMを用いて64 bit幅バスを構築した場合は最大で800 MB/sであり，2ポートのRambusより若干性能が低いが，Rambusより従来のメモリに近い特性を持ち，ページアクセスモードを持

つ．典型的な 100 MHz 8 bit 幅 16 MbitS-DRAM においてはクロックサイクルタイム 10 ns，RAS アクセスタイム 60 ns, RAS サイクルタイム 100 ns, ページモードサイクルタイム 20 ns, ページサイズ 512 ワードであるので，長いベクトル長に対して見積もられるメモリバンド幅は以下のようになる．

連続アクセス時：8 byte ×∞/(60 ns+(∞−1)×20 ns)＝800 MB/s

等間隔アクセス時

最良：8 byte×256/(60 ns+255×20 ns)＝397 MB/s

最悪：8 byte/100 ns＝80 MB/s

リストアクセス時

最良：8 byte×512/(60 ns+512×20 ns+4 byte×512/0.8 GB/s)＝319 MB/s

最悪：8 byte/(100 ns+4 byte/0.8 GB/s)＝76 MB/s

上記のように S-DRAM は R-DRAM と比較して連続アクセス時に 7%の性能低下があるものの，等間隔アクセス時やリストアクセス時に 15〜65%の性能向上が見込め，総合的には S-DRAM が優れると考えられる．さらに学会発表レベルでは 100 MHz を越える S-DRAM も試作されているので周波数の向上も今後期待できる．以上のような理由から，**TS/1** では S-DRAM を採用する．

3.2 マルチスレッデッドベクトル機構

TS/1 では R-DRAM を採用した場合連続アクセス時に期待できる転送速度は 849 MB/s，S-DRAM 採用時に 800 MB/s が得られるに過ぎない．しかし，**図 1** 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 2 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[11]．

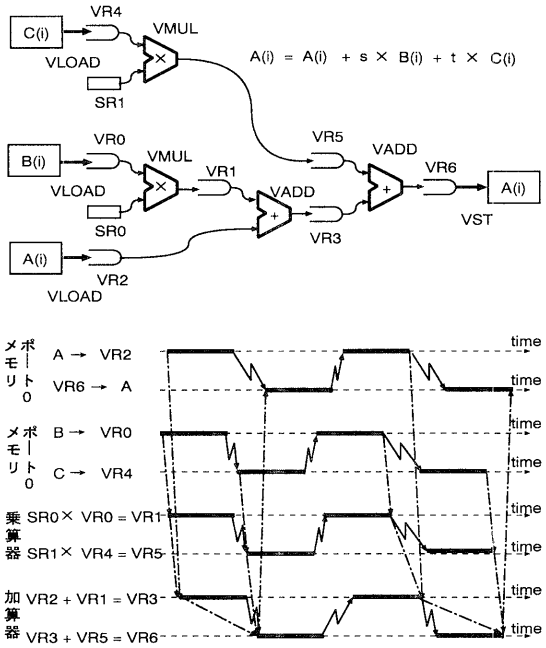

図 1 マルチスレッデッドベクトル機構の動作
Fig. 1 Behavior of multi-threaded vector mechanism.

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 125 MFLOPS 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．

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 80 MHz 動作時に単精度 320 MFLOPS，倍精度 160 MFLOPS の浮動小数演算器[12]に若干の改良を加えて実現される．

3.3 プロセッサ間チェイニング機構

各プロセッサに要素がばらまかれた行列間の乗算のように短いメッセージによる激しいデータ移動を伴う処理においては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．またメインの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[13]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．

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 2** 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

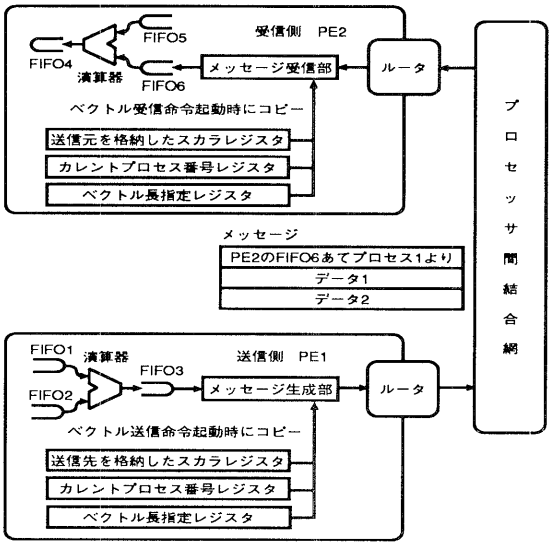

図2 プロセッサ間チェイニング機構
Fig. 2 Interprocessor chaining mechanism.

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．

上記の機構の効率的動作のためには通信バンド幅の確保が必要だが，**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（その実現性は実装プロトタイプの作成により確認済）を実現する予定であり，適度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．また，通信性能と信頼性・アベイラビリティの両立をはかるフォールトトレラントな結合網のアーキテクチャ[17),18)]を採用する．

4. 基本処理の並列化と性能見積もり

4.1 行列乗算

行列の乗算は行列の加減算の場合と異なり，行列の各要素がプロセッサアレイ上に分散して保持されている状態から行列積を計算する場合，片方の行列のi行を形成する行ベクトルともう片方の行列のj列を形成する列ベクトルの内積が結果行列の要素(i, j)となるため，必然的に激しい通信を必要とする．

商用超並列マシンではライブラリとして提供される放送を用いる行列乗算が良く用いられ，SIMD型のCM-2ではPE数16K，256元行列で38 MFLOPS，MP-1ではPE数1K，32元行列で16 MFLOPSのように低速となっている．これは通信バンド幅が足りないことが主な原因と考えられる．

一方，ベクトル型スーパーコンピュータ上で行列乗算を行う場合は最内側ループはSAXPY演算（V＝V＋S×V）となるが，そのベンチマーク結果[14)]は1CPUでベクトル長1000の場合CRAY-YMPが262 MFLOPS，CRAY-C 90が712 MFLOPS，SX-3が1402 MFLOPSとなっておりSIMD型商用超並列マシンよりはるかに高速である．ただしパイプライン数が多い機種ほどピーク性能と比べた効率が低く，ベクトル長と効率の関係からむやみにパイプライン数を増やしても高速化しない．

このように行列の乗算は通信がネックとなるため従来の超並列マシンではベクトルマシンに比べ，あまり高速化できなかったものである．行列乗算の並列処理法には(1)通信フェーズと演算フェーズに分け，通信に放送を用いる方式，(2)行方向および列方向にデータを隣のPEにシフトさせつつ乗算・加算を繰り返す方式，の二つの方針が考えられる．**TS/1**では(2)のアルゴリズムを用いれば通信と演算はオーバーラップされオーバーヘッドの削減が期待できるので，プロセッサ間チェイニング機構を用いて(2)の方式の実現を考える．

本節ではまず4×4構成のプロセッサアレイにより8元の正方行列積を実行する場合を説明し，次に64×64構成のプロセッサアレイで64n元正方行列積を実行する場合の性能を求める．

★16(4×4構成)PEによる8元行列積

TS/1で用いられる浮動小数点演算器[12)]は加算の遅延が3クロックある．このようにパイプライン型の演算器を用いて累算を行う場合には，累算演算が3クロックに1回しか実行されない危険性があり，通信が速いマシンではこのロスをなくす工夫が必要になる．行列A_{ij}，B_{ij}，C_{ij}は**図3**のように4×4の部分正方行列{A0, A1, A2, A3}{B0, B1, B2, B3}{C0, C1, C2, C3}に4分割され，要素プロセッサPE_{ij}のメモリに**図4**のように自然なマッピングで格納されているとする．

このような分割において部分行列はC0＝A0×B0＋A1×B2，C1＝A0×B1＋A1×B3，C2＝A2×B0＋A3×B2，C3＝A2×B1＋A3×B3となる．8個の部分行列積の間には依存関係はなく，これらを演算パイプラインのステージ間の並列処理に用いる．

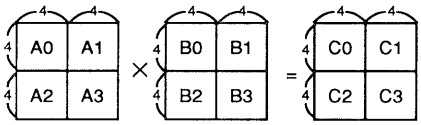

図3 プロセッサアレイより大きな行列の分割
Fig. 3 Decomposition scheme of larger matrices than processor array.

こうすると通信が無限に速ければ若干の立ち上がり時間の後にほぼピーク性能で8個の部分行列積を計算し出力する．以下にその手順を示す．

(Step. 1) 前処理

- 図5のように A_{ij} が流れる FIFO (VR 8) に A_{ij}, B_{ij} が流れる FIFO (VR 9) に B_{ij} を8要素書き込む
- C_{ij} の部分和が累算される FIFO に0を8回書き込む
- 東，西，南，北方向からそれぞれ別の FIFO にデータを受信する命令を起動する
- バリア同期を1回とり上記の処理の完了を確認する

(Step. 2) 本処理 (注：アレイの端部は若干処理が違う)

- 東，西，南，北方向にそれぞれ別の FIFO からデータを送信する命令を起動する
- 北，西から受信した FIFO からそれぞれ2本の FIFO にデータをコピーする命令を起動する
- 2本の FIFO からデータを乗算し結果を FIFO に格納する命令を起動する
- 2本の FIFO からデータを加算し結果を FIFO に格納する命令を起動する

以上で図6のようにパイプラインはチェイニングされる

(Step. 3) 後処理

- ベクトル命令がすべて終了したことを PE ごとに確認する
- C_{ij} の部分和が貯まっている FIFO 中のデータを2個取り出し加算し C_{ij} とする
- 上記を4回繰り返し C_{ij} の4要素をメモリに格納する
- バリア同期を1回とり上記の処理の完了を確認する

★ $\frac{N}{2}\times\frac{N}{2}$ 構成プロセッサアレイによる N 元行列積

TS/1 のようにクロックごとに1加算と1乗算が可能な演算パイプラインと，同時に4送信4受信が可能なプロセッサ間チェイニング機構を備えた $N/2\times N/2$ 構成二次元プロセッサアレイによる，N 元の正方行列乗算の演算速度[GFLOPS]は以下の式で表すことができる．

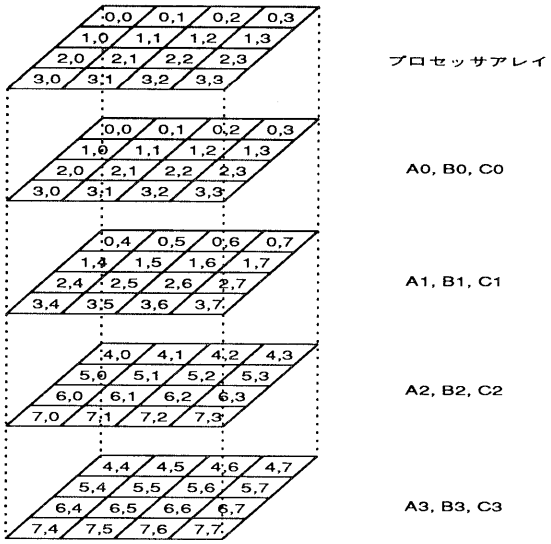

図4 部分行列のマッピング
Fig. 4 Mapping scheme of submatrices.

$$\frac{2fN^3}{10^3\left(4NT_c+\frac{NT_d}{2}+T_m\right)}$$

T_c：1浮動小数を転送するクロック数
T_d：ノード通過遅延クロック数
T_m：FIFO 初期化，バリア同期，加算4回
f ：クロック周波数[MHz]

上式によれば T_c, T_d, T_m の順に性能に敏感で，同時に4送信4受信が可能であってもリンク当たりのバンド幅 T_c が性能を大きく左右する．

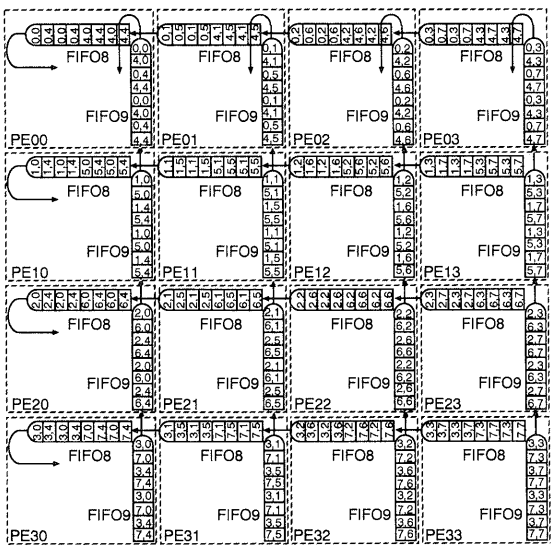

図5 行列乗算時の FIFO の初期化
Fig. 5 FIFO initialization for matrix multiplication.

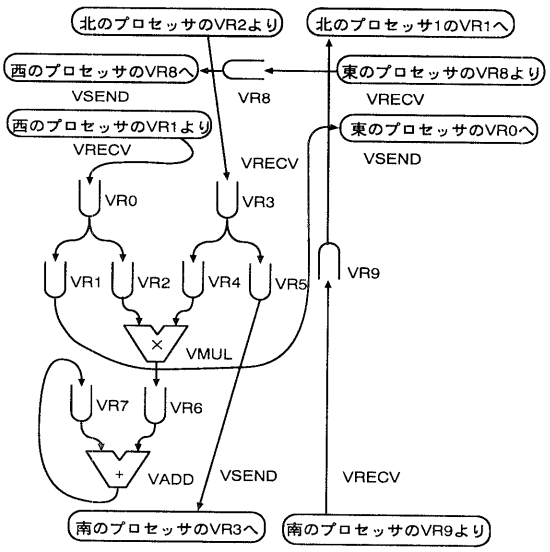

図6 行列乗算時の仮想パイプライン
Fig. 6 Virtual pipelines for matrix multiplication.

TS/1 においては通信リンクは双方向通信に安定した性能を実現できる全二重通信路で，単方向分が1クロックあたり32 bitを転送する能力を持ち，64 bit浮動小数を転送する際にヘッダー1語＋データ2語の合計3語を転送する．(注：ヘッダー付加はフォールトトレランスとユーザー間プロテクションのための投資である)

つまりすべての64 bit浮動小数にヘッダーを付加したとすると $T_c=3$ であり，さらに $T_d=2$，$T_m=256$，$f=62.5$ とすると，**TS/1** を64×64構成(4096 PE)のプロセッサアレイとして動作させた時に128元行列乗算が約137 GFLOPSの速度で実行されることがわかる．この速度は現在のSIMD型商用超並列マシンの数千倍，ベクトル型スーパーコンピュータの約百倍高速であることを示しており，プロセッサ数に比較してこのように小さな行列乗算が極めて高い性能で実現される点で画期的である．

もし，ヘッダー転送を省略したり，転送クロック周波数を演算器のクロックをあげたり，セルフタイミング非同期転送プロトコルを使うなどによって，3倍の転送レートを実現できるならば，512 GFLOPS (125 MFLOPS×64×64) のピーク速度のマシン(最大構成の1/16) において約293 GFLOPSの実効性能が実現できる．

$N=64n$ 元 $(n>2)$ の行列を64×64のプロセッサアレイにマッピングする場合は，64元の部分行列が n^2 個に分かれ，行列Cの部分行列はそれぞれ n 個の行列A, Bの部分行列積の和により計算できる．つまり全体では n^3 個の部分行列積を並列処理できる．

しかし，FIFOの長さは有限なのでその範囲に収まるように分割して処理する．例えばFIFOの長さを256バイト＝32語とし，1回の分割処理でFIFOに初期化するベクトル長が8ではなく32とすることができるようにすると，プロセッサアレイを通り抜けるベクトル長は32にプロセッサアレイの一辺の長さをかけた64×32＝2048となる．その場合は128元の例よりもベクトル長(上式の分子と T_c にかかる項)が4倍に増えるので T_d，T_m の項の影響が薄まり，より高効率($T_c=1$ の場合431 GFLOPS) で動作することになる．

4.2 反復法 (2-color SOR) による連立一次方程式求解

差分法で二次元空間(長方形領域)を離散化した方程式 $A(i,j)\times X(j)=B(i)$ は大規模な問題では領域の格子数の数倍の記憶容量で計算できる反復法が用いられる．その一種である2-color SOR法[15]の並列処理について評価する．これは**図7**に示すように対象空間を偶点と奇点の2色に色わけし，偶点どうしの計算および奇点どうしの計算が並列に実行できる．2-color SORの並列処理法としては(1)すべての点の演算に対して通信を行うことにより単純化する方式，(2)処理が複雑になるが通信を抑制するようにする方式，の

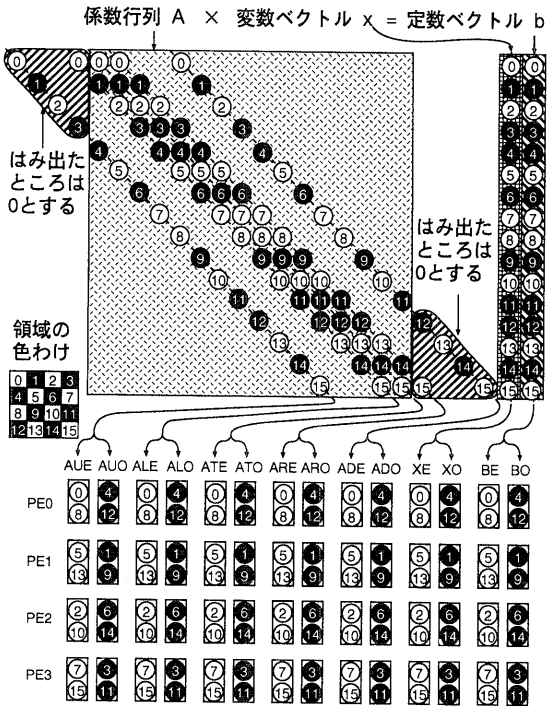

図7 2color SORのデータマッピング
Fig. 7 Mapping scheme of matrix for 2-colored SOR.

2つの方針が考えられる．

通常の商用超並列マシンは通信能力が低く(1)の方針は大幅な性能低下をもたらす．一方，**TS/1** は演算器が結果を出すレートと同程度の通信レートと，プロセッサ間チェイニング機構により通信オーバーヘッドは低いので(1)の方針は必ずしも性能低下をもたらさない．逆に(2)の方針はベクトル処理効率を低下させる要因となる．そこで従来型のマシンではうまく実現できなかった(1)の方式についてどのように処理され，予測される性能について説明する．**TS/1** での高速実行を行う上ではメモリポートが効率的に動くように連続アクセスになるよう工夫すると良い．

この方式の場合は，空間の次元数−1の次元に関して並列化し，残った一つの次元に関してベクトル化(パイプライン処理)する．例えば，二次元の問題では一次元プロセッサアレイにより x 方向を並列処理し，プロセッサ内部で y 方向をベクトル処理する．

演算は偶点および奇点の二つのフェーズに分けられるので，メモリアドレスに交互に偶点と奇点が格納される通常のメモリ割り当てをした場合は，等間隔メモリアクセスとなるため性能が低下するので，予め偶点のみ集めた配列と奇点のみ集めた配列に分けて格納する．つまり図7に示すように7本の配列を奇点と偶点に対応する配列に2分割し14本の配列に置き換える．連続アクセス用に最適化された 2-color SOR のプログラムを図8に示し，ループ20および30の内部における異なる i に対するステートメントは並列に実行できる．

$n\times m$ のメッシュで切られた二次元の問題を n 台の一次元ベクトルプロセッサアレイで処理する場合，各プロセッサが持つべき解ベクトルの2本の一次元配列 XO(i)，XE(i)のサイズ(ベクトル長)は $m/2$ となる．ここでX方向に隣接する点は必ず隣のプロセッサにマッピングさせる．各プロセッサの動作は右端のPE，偶数番目のPE，奇数番目のPE，左端のPEの4種類が必要になる．偶数番目のPEは偶数行目の偶点と奇数行目の奇点を担当し，奇数番目のPEは奇数行目の偶点と偶数行目の奇点を担当する．そのようにマッピングでは，XE，XOの上下は以下のように対応する．

偶数番目のPE

XO(上)＝XO(i−1)　XE(上)＝XE(i)

XO(下)＝XO(i)　　XE(下)＝XE(i+1)

奇数番目のPE

XO(上)＝XO(i)　　XE(上)＝XE(i−1)

XO(下)＝XO(i+1)　XE(下)＝XE(i)

また右，左が付いているXO，XEは右，左のプロセッサから受信されたデータを用いることになる．つまり各プロセッサは奇点を計算しているフェーズではXE(i)を，偶点を計算しているフェーズではXO(i)を両隣のプロセッサに送信する必要がある．

```
10 CONTINUE
   フラグ=1
   DO 20 i=1,n*m/2
     誤差=BO(i)-(AUO(i)*XE(上)+ALO(i)*XE(左)
     +ATO(i)*XO(i)+ARO(i)*XE(右)+ADO(i)*XE(下))
     XO(i)=XO(i)+加速係数*誤差*対角要素の逆数(i)
     IF(|誤差|.GE. 許容誤差) フラグ=0
20 CONTINUE
   DO 30 i=1,n*m/2
     誤差=BE(i)-(AUE(i)*XO(上)+ALE(i)*XO(左)
     +ATE(i)*XE(i)+ARE(i)*XO(右)+ADE(i)*XO(下))
     XE(i)=XE(i)+加速係数*誤差*対角要素の逆数(i)
     IF(|誤差|.GE. 許容誤差) フラグ=0
30 CONTINUE
     IF(フラグ.EQ.0) GOTO 10
```

図8 2-color SOR のプログラム
Fig. 8 Program for 2-colored SOR.

上記の方針でハンドコーディングしたアセンブリプログラム中のベクトル命令種類ごとの回数はメモリアクセス10回，リンク当り通信1回，演算13回(乗算7，加減算6)となる．**TS/1** は演算1回当りメモリアクセス0.8回，リンク当り通信を最悪1/6回が可能とする予定であり，上記の命令の頻度と照合しても演算，メモリアクセス，通信はほぼバランスが取れている．

例えば1000×1000の二次元空間の問題を1000台のPEで実行するような場合，ベクトルプロセッサの立ち上がり時間が無視できるようになり，加減算は乗算より1回少ないので，このループはピーク速度の13/14の実効性能で動作すると考えられる．つまり1000PEならばピークで125 GFLOPSのところを116 GFLOPSの性能が得られることになる．

ところが，もしプロセッサ間チェイニングによらない通信で同様の処理を実行する場合は，通信オーバーヘッドを0にしてもメッセージの送受信のために最低4回のメモリアクセスが増加するのでメモリバンド幅がネックとなるため上記の性能は達成し得ない．すなわちプロセッサ間チェイニングはメモリバンド幅を消費しない通信方式と言える．

なお上記の説明においては二次元空間の問題を例に説明したが，三次元空間の問題であれば二次元メッシュ結合のプロセッサアレイにより同様な処理が実現でき，同等のループ内で送信，受信，加算，乗算，ロードが二命令ずつ増加する．つまり，命令種類間のバラン

スとマシン能力の議論は二次元の場合と大差がない．

なお，問題の規模によってはベクトル長が百程度しか無い場合があり，この場合立ち上がりロスが観測されると考えられる．上述のようにSORにおいてはメモリバンド幅ネックとならず，乗算と加算の比率が7：8とほぼ1：1に近いので，例えば定数倍したベクトルと別のベクトルの加算（V＝V＋S＊V）のベンチマーク値が参考にできる．文献14)によればCray Y-MPではベクトル長100の場合は187.9 MFLOPSとなっており，すでにピーク性能333.3 MFLOPSの6割に近い56.3%となっている．FIFO型ベクトルレジスタは(1)ベクトルレジスタ全体にデータがロードされる前でも必要なデータが一組でも揃えば演算の実行が開始される，(2)ストリップマイニング処理が不要，という二つの理由から立ち上がりロスが通常のベクトルレジスタより少なく，さらに**TS/1**ではCray Y-MPよりクロック周期が16/6倍遅いのでパイプラインの段数を浅くできるため，Cray Y-MPより短いベクトル長に対してかなり強いことが予想される．よって例えば64×256のプロセッサアレイで256×256×256の三次元問題を実行した場合，ベクトル長は128となるので，ピーク速度が2.0 TFLOPSのところを理論的最大性能がその14/15の1.9 TFLOPSで，その2/3程度の1.3 TFLOPS程度は実現可能と予測される．

また，乱流のシミュレーションにおいては1024×1024×1024もの巨大な三次元空間メッシュに切らなければならない問題も存在する．このような巨大問題は最大構成の65536台（64×64×16構成）の全体を1ユーザーで用いることが推奨される．解法としてSORを採用したと仮定し，マシンを64×1024の二次元プロセッサアレイとして用い上記の手法により並列化した場合，ベクトル長は長くなるため立ち上がりロスはほぼ隠され，ベクトル演算部のピーク性能の14/15にあたる理論的最大性能である7.6 TFLOPSに近いSOR部の実効性能が達成されると予測される．

4.3 直接法（軸選択付きLU分解）による連立一次方程式求解

連立一次方程式を直接法で解く場合は大規模な行列になればなるほど演算精度的に問題が生じやすく，このため倍精度，倍々精度浮動小数の使用や，スケーリングや軸選択などの精度を保つための手法の必要性が高い．軸選択付きのLU分解はLinpackベンチマークの中でも用いられており，スーパーコンピュータの性能指標として様々なマシンにおける実測性能データ[2]が揃っている．軸選択をする場合にはデータの流れは定常的ではないので古典的なシストリック手法が使えず，軸選択（最大値検出）の結果を受けてピボット（対角要素）やピボット行やピボット列の要素を同じ列または行のプロセッサに放送する必要がある．以下に$N\times N$行列の軸選択付き外積型LU分解の$P\times P$構成二次元プロセッサアレイによる並列処理の流れを示す．

(STEP.0)：STEP.1～4をN回繰り返す

(STEP.1)：Lベクトル（ピボット列の分解が完了していない要素）をx方向に放送

(STEP.2)：Lベクトル内最大値と行番号を求めピボットとし，逆数と行番号をy方向に放送

(STEP.3)：ピボットの逆数をUベクトル（ピボット行の分解が完了していない要素）に乗じピボットで除算されたUベクトルをy方向に放送

(STEP.4)：受信したLベクトル，Uベクトルの対応する要素間で$a=a-l\times u$を計算する

上記の各ステップにおける通常のメッセージ交換方式の実行時間に反映される浮動小数演算量と通信回数とメッセージ長と同期回数の概略を**表1**に示す．

TS/1の演算器[12]のパラメータに基づき除算は17クロック，乗算と加算は同時実行可能で1クロックずつで実行できるとする．STEP.4における演算は全演算の大半を占めるが，STEP内での配列要素間の演算にはデータ依存関係がなく並列実行可能であり，ベクトル処理することができる．この時分解終了行および終了列に対応するデータの演算結果の格納はマスクされる．

上記より，Pがある程度大きく（高並列・超並列），N/Pがあまり大きくない場合，短いメッセージが多数回飛ぶことになり，演算量が少なくなる（つまり細粒度処理になる）のでメッセージ通信起動オーバーヘッドがネックとなるため効率が著しく低下するのが一般的である．よって通信オーバーヘッドをいかに少なくできるかという点が問題になる．

具体的にはLINPACKベンチマークのように1000×1000の行列をLU分解することを想定すると64×64構成の二次元プロセッサアレイを用いた場合，メッセージ長は16語の通信回数が12回，メッセージ長1の通信が12回，同期4回に対し，演算量が311クロック分となり通常のOSが関与する通常のメッセージ交

表1 軸選択付きLU分解の処理内容の内訳
Table 1 Contents of LU decomposition with pivoting.

STEP番号	#1	#2	#3	#4
演算時間	0	$N/P+\log_2 P+17$	N/P	N^2/P^2
通信回数	$\log_2 P$	$2\log_2 P$	$\log_2 P$	0
メッセージ長	N/P	1	N/P	0
同期回数	1	1	1	1

換(1回あたり数百～千クロック程度必要)を行っていたのでは完全に通信ネックとなり性能低下は免れない.

通信リンクあたり 200 MB/s という高速な通信バンド幅を持ちながら,通信に OS が介してしまうために千クロック程度のオーバーヘッドを有する二次元メッシュマシンである Touchstone DELTA が 512 台構成の時,単体実効処理性能(注:単体ピーク性能ではない)に台数を掛けた性能の 3%の効率しかでないという 1000×1000 の LINPACK ベンチマーク結果[2]が公表されている.

一方行列サイズを 7500×7500 まで大きくすれば通信の比率が減るために DELTA でもピーク性能の半分の性能(10 GFLOPS)が得られる.このことはこの程度の規模の行列を扱う場合は通信バンド幅だけでなく通信オーバーヘッドを減らすことが重要であることの一つの証拠となっている.

TS/1 ではグローバル演算や放送はプロセッサ間チェイニングによって実現することを想定しており,通信オーバーヘッドを激減させ,上記のような問題点を解決する.LU 分解においてネックとなる行方向,列方向の放送や列方向のグローバル MAX 演算は,トリー状の放送過程や,トリー状の収集過程を用いることによって,プロセッサ間チェイニングという極めて低オーバーヘッドな通信を利用しながら高速実行が可能である.

通常の OS を介する通信が数百～千クロック程度の大きなオーバーヘッドがかかるのに対し,プロセッサ間チェイニングであれば一回に換算すれば数クロック程度のオーバーヘッドで済むことになり,これを 10 クロックとしても 24 回の通信では 240 クロックで済むので演算量の 311 クロックより少なくてすみ,やっと同期オーバーヘッドが陽に現れて来る程度の状態となる.

つまり 1000×1000 の行列を 64×64 構成の二次元プロセッサアレイで LU 分解する場合,本方式を採用せず一回の通信に数百から千クロックかかってしまう場合がピーク性能の 1～2%程度しか期待できないのに対し,本方式では同期オーバーヘッドを 100 クロック程度に抑えれば通信時間と同期オーバーヘッドを合計した時間が 340 クロック程度となり,オーバーヘッドの合計は演算時間とほぼ等しくなるので,ピーク性能の半分弱の性能は期待できる.

最大構成の 16 分の 1 にあたる 64×64 のプロセッサアレイのピークベクトル演算速度が 512 GFLOPS であるので,そのマシンで 200 GFLOPS 程度を実現できることを意味する.SX 3 (4 プロセッサ)や CRAY Y-MP C 90 (16 プロセッサ)を代表とするこの程度の小規模な問題でも高い効率を示す既存のマルチプロセッサ型ベクトルスーパーコンピュータは最大構成でも 10 GFLOPS 程度[2]であるので,その約 20 倍の性能が期待できることになる.

超並列マシンではより大きな行列に対しては演算の比率が高まり,通信や同期のオーバーヘッドが隠されるようになるのに対し,ベクトル型スーパーコンピュータではある程度以上ベクトル長が増えてもその並列性を性能向上に生かしきれなくなるので,サイズが大きいほどベクトル型スーパーコンピュータと本超並列マシンとの差は広がる.

また超並列のアプローチは極めて大きな行列に対しては更にプロセッサ数を増加させることで性能・メモリ容量に対する要求に対応でき,最大構成であと 16 倍まで性能を向上可能である.ゆえに計算時間や,メモリ容量によって従来のベクトル型スーパーコンピュータで計算できなかったような巨大な密行列に対して LU 分解を行うことが可能になる.

5. おわりに

本論文では次世代高バンド幅 DRAM 用アクセスブロック化機構と FIFO 型ベクトルレジスタを有するマルチスレッデッドベクトルプロセッサと,三次元実装による高い通信バンド幅の下で有効に機能するプロセッサ間チェイニング機構の組み合わせにより,超並列計算機が苦手としてきた科学技術計算において重要ないくつかの問題においても,現在開発中の MIMD ベースの超並列マシン **TS/1** が絶対性能の上でベクトル型スーパーコンピュータを数十倍から数百倍上回る性能が実現できる見通しを示した.

つまり,高速な演算器に見合ったメモリバンド幅を与える連続アクセスが高速な Rambus 型 DRAM や同期型 DRAM の特質を生かしたプロセッサアーキテクチャにより,単体において達成される高水準な浮動小数演算性能が,プロセッサ間チェイニング機構によって超並列処理時においてもいかんなく発揮され,厳しい通信が要求される細粒度処理において,単体演算性能が高くても並列処理効率が高いという理想的な状態を実現している.プロセッサ間チェイニングによりプロセッサ数に対して相対的に小さな行列の乗算の超高速処理が示されたことは,百万プロセッサの超並列マシンに向けた一つの明るい展望といえよう.

また,反復法の一種の SOR を従来の超並列計算機では禁止的だった通信を抑えないスタイルのプログラムでも,極めて高い処理速度が実現できることを示し

た．このことは**TS/1**が他の応用に際しても性能確保のためのプログラマの負担を軽減することを予感させる．プロセッサ間チェイニングはメモリバンド幅を消費しない通信方式であり，SORの例を通してメモリバンド幅が不足しがちな応用に対しても効果があることが示された．今後はSORより収束性の優れた他の反復法などへの適用についても検討したい．

本論文で触れた機構はライブラリのプログラマや，ボトルネックとなっている箇所の高速化を切望する先進的プログラマに対し，究極的高速化のための新しい手段を提供していると言える．コンパイラがこれらの機構をコード生成時に利用すれば，一般ユーザにも高い性能を提供できる可能性がある．今後はそのような高度なコード生成を行う高級言語コンパイラの開発を行う予定である．

本論文ではプロセッサ間チェイニングは数値計算に見られる規則的な細粒度処理には有効であることが示されているが，不規則な通信には向かない一面がある．ゆえにすべての通信をプロセッサ間チェイニングに頼るべきではない．このような観点から**TS/1**ではこのほかに分散共有アクセス機構[19]やメッセージ交換支援機構といった複数の通信モードに対応するハードウェアを導入する予定である．

参考文献

1) 有馬，村上，金田：アドバンスト・コンピューティング—21世紀の科学技術基盤—，培風館（1992）．
2) Dongarra, J. J.: Linpack Benchmark: Performance of Various Computers Using Standard Linear Equations Software, *Supercomputing Review,* March 1992, pp. 54-63 (1992).
3) Dally, W. J. et al.: The J-Machine: A Fine-Grain Concurrent Computer, *Proc. of IFIP Congress,* pp. 1147-1153 (1989).
4) 児玉，坂井，山口：データ駆動型シングルチッププロセッサEMC-Rの動作原理と実装，情報処理学会論文誌，Vol. 32, No. 7, pp. 849-858 (1991)．
5) 村田，小国，唐木：スーパーコンピューター科学技術計算への適用—，丸善（1985）．
6) 田邊，菅野，鈴木，小柳：マルチパラダイム超並列テラフロップスマシンTS/1の構想，情報処理学会技術報告，Vol. 93, No. 71 (SWoPP鞆の浦'93), 101-6, pp. 41-48 (1993)．
7) 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), CPSY 92-24, pp. 33-40 (1992)．
8) 串山，大島，古山：500 Mバイト/秒Rambus仕様4.5 MビットDRAM，東芝レビュー，Vol. 47, No. 7, pp. 575-578 (1992)．
9) 安保：見えてきたシンクロナスDRAMの仕様，100 MHz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)．
10) 古勝，渡辺，近藤：最大性能1.3 GFLOPS，マシン・サイクル6 ns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) 橋本，村上，弘中，安浦：マイクロベクトルプロセッサ・アーキテクチャー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 (SWoPP日向灘'92), CSPY 92-21, pp. 9-16 (1992)．
12) Ide, N. et al.: A 320 MFLOPS CMOS Floating-Point Processing Unit for Superscalar Processors, *Proc. of Custom Integrated Circuit Conference (CICC) '92,* p. 30.2 (1992).
13) Kung, S. Y.: On Supercomputing with Systolic/Wavefront Array Processors, *Proc. of the IEEE,* Vol. 72, No. 7, pp. 867-884 (1984).
14) Margaret, L. S. et al.: A Performance Comparison of Four Supercomputers, *Comm. ACM,* Vol. 35, No. 8, pp. 117-124 (1992).
15) Adams, L. and Ortega, J.: A Multi-Color SOR Method for Parallel Computation, *1982 Int'l Conf. on Parallel Processing,* pp. 53-56 (1982).
16) 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)．
17) 田邊，菅野，鈴木，小柳：超並列TeraflopsマシンTS/1—Wavefront Arrayのための結合網アーキテクチャ—，第48回情報処理学会全国大会論文集，Vol. 6, pp. 49-50 (1994)．
18) 菅野，田邊，鈴木，小柳：超並列TeraflopsマシンTS/1—フォールトトレラントルーティング—，第48回情報処理学会全国大会論文集，Vol. 6, pp. 51-52 (1994)．
19) 鈴木，田邊，菅野，小柳：超並列TeraflopsマシンTS/1—分散共有メモリアーキテクチャー，第48回情報処理学会全国大会論文集，Vol. 6, pp. 53-54 (1994)．

（平成5年9月14日受付）
（平成6年12月5日採録）

田邊　昇（正会員）

昭和38年生．昭和60年横浜国立大学工学部電気工学科卒業．昭和62年同大大学院修士課程修了．同年(株)東芝入社．現在，同社研究開発センター勤務．この間疎行列LU分解専用計算機，高並列AIマシン，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の研究開発に従事．電子情報通信学会会員．