4B－2

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～

田邊 昇　菅野 伸一　鈴木 真樹　小柳 滋
RWCP* 超並列東芝研究室†

1　はじめに

超並列 Teraflops マシン TS/1 では遠隔ベクトルプロセッサ上の FIFO 型ベクトルレジスタ間のデータ転送機構（プロセッサ間チェイニング機構 [1]）による Wavefront Array[2] 動作が可能である。本機構によれば行列計算などにおいて、PE 数や演算性能が高い超並列マシンほど深刻となる粒度低下による処理効率の低下を抑制できる。

そのインプリメントに際し、PE 数の数倍の要素数しか持たない行列の乗算のような超細粒度処理においても高効率な動作を可能とするには、プロセッサのスループットの数倍の通信バンド幅の実現が望まれる。またプロセッサ数や通信速度を高く設定した Wavefront Array 型超並列マシンにおいてはフォールトトレランスや高信頼化対策と実効通信バンド幅の両立も重要である。さらに大規模な Wavefront Array を実現する上では正方形に近い二次元メッシュの埋め込み能力も要求される。

本稿では、故障 PE の存在を想定した TS/1 における大規模にして高効率で高信頼な Wavefront Array 動作実現のための結合網アーキテクチャを提案する。

2　超細粒度処理高速化のポイント

J-machine[3], EM-4[4] などの細粒度高並列計算機と呼ばれているマシンは 10 命令に 1 回程度以下の通信を行う粒度を想定した作りになっている。そこで、演算 1 回につき 1 回以上通信が必要な処理を超細粒度処理と名付ける。これは百万 PE の時代には必須の処理であり、数万 PE で数万ニューロンのシミュレーションを行う時などにもこのような状況が発生する。

このような問題意識から、TS/1 ではプロセッサ間チェイニング機構を用いた Wavefront Array 動作による規則的超細粒度処理の究極的高速化を目標とした。そこで超細粒度処理高速化のポイントの把握のために行列乗算などを例に検討を行った。[1]

TS/1 のようにクロック毎に 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 × 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 [TFLOPS] は以下の式で表すことができる。

$$\frac{2fN^3}{10^6(4NT_c + \frac{NT_d}{2} + T_m)}$$

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ため、TS/1 の設計においてはこの点に特に留意した。

3　結合トポロジー

TS/1 の設計に際し、分散共有アクセスやメッセージ交換による通信の存在も踏まえた上で、大規模にして高効率で高信頼な Wavefront Array 動作を実現するための結合トポロジー選択において重点を置いた性質を以下に列挙する。

- 遠隔基板間配線排除によるノード数拡張容易性
- 三次元実装との親和性による高い通信バンド幅
- 故障回避時の局所的通信バンド幅低下の防止
- 基板間配線を殆ど増加させない故障回避用配線
- 大規模正方 2D-mesh 埋め込み時の高いバンド幅
- 高い二分割バンド幅による高ランダム通信性能

以上のような観点から TS/1 の結合トポロジーは最大構成で 64 × 64 × 16 の 3D-torus を基本とし、さらに基板内の 16 個の PE の各 1 ポートと 1 個の代替 PE の 2 ポートを 18 × 18 クロスバ網により結合する。代替 PE どうしは 2D-torus で結合する。

TS/1 では直径には重点を置いていないがクロスバにより直径短縮、ランダム通信性能向上などの二次的な効果もある。なお代替 PE には I/O デバイスのためのインタフェースを設け、代替 PE を有効活用する。

Massively Parallel Teraflops Machine "TS/1", – Network Architecture for Wavefront Array –
Noboru TANABE, Shin-ichi KANNO, Masaki SUZUKI, Shigeru OYANAGI
*Real World Computing Partnership(新情報処理開発機構)
†(株) 東芝 研究開発センター 内

4 耐故障対策

4.1 耐故障対策の概要

TS/1 の結合網の耐故障対策を以下に列挙する。

- 単発的データ誤りの訂正:32bit データにつき 2bit 誤り訂正機構付き通信リンク
- 単発的制御誤動作の検出:順序検査 bit 列による到着順序違反検出
- 永久的故障の回避 (性能低下なし):3D-torus の 1 方向にクロスバを用いた代替 PE との結合
- 永久的故障の回避 (性能低下あり):仮想ネットワークを用いた迂回トーラスルーティング

4.2 クロスバを用いた故障回避

遅延時間の局所的な若干の増加は Wavefront Array の処理性能に殆ど影響を与えないが、バンド幅の低下は大きな影響を与える。TS/1 のクロスバを用いた故障回避は、基板内に 1 つの故障 PE を局所的バンド幅低下なしに基板内の代替 PE と置き換える。

3D-torus 上で故障 PE に隣接する 6 個の PE のうち基板内の 2 個はクロスバにより代替 PE と接続する。残りの別基板上の 4PE は、各々の基板上の代替 PE を経由すれば、その隣が目的とする代替 PE であり、この経路が健全ならばバンド幅の低下はない。

5 大規模正方 2D-mesh の埋込み

64 × 64 × 16 の 3D-torus 上に 256 × 256 の 2D-mesh をバンド幅の低下無しに埋め込むことは不可能だが、クロスバの併用によりそれが可能である。

例えば 16 枚の 64 × 64 の 2D-mesh のタイルを図 1 のように 4 × 4 の 2D-mesh 状にならべる。このようにするとタイルの境界は互いに Z 座標のみが異なり、横方向には Z 座標が連続しているので Z 方向 torus リンクで、縦方向にはクロスバで 256 × 256 の 2D-mesh の論理的リンクを代行できる。

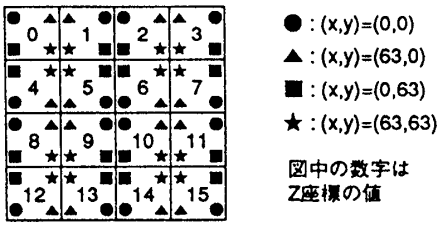

図 1: 256 × 256 の 2D-mesh の埋込み法

TS/1 のプロセッサ間チェイニング機構ではベクトル送信命令起動時にスカラレジスタ上に設定しておくことでチェイニング先の PE 番号を指定できるが、上記の規則でソフト的に論理的な東西南北の PE 番号設定することで 256 × 256 の 2D-mesh の高バンド幅な埋め込みが実現される。

6 通信バンド幅向上対策

TS/1 の結合網の通信バンド幅向上対策を以下に列挙する。特にプロセッサ間チェイニング機構を用いた Wavefront Array 動作においては継続的に同じ PE あての同一ヘッダーを持つパケットが飛ぶため、一度通った経路を他のヘッダーを持ったパケットが通らない限りヘッダーの転送を省略する機構の効果が高い。

- 基板渡り遠隔リンクが発生しない torus の実装
- ケーブルレス三次元実装
- 多ピン LSI パッケージの使用
- 送信クロックの両エッジでのデータ転送
- 直前と同一のヘッダー転送の省略
- 近傍通信に絞った簡略化ヘッダーの併用
- 多重仮想 network によるリンクの利用率向上

7 TS/1 の超細粒度行列乗算性能

TS/1 では上記の大規模正方二次元メッシュ埋め込み能力により $N/2 = 256$ とすることが可能であり、上記の通信バンド幅向上策により倍精度時 $T_c = 2$, 単精度時 $T_c = 1$ が得られる。ここで $T_d = 2$, $T_m = 256$, $f = 62.5$ とすると最大構成の TS/1 による 512 元行列乗算性能が概算でき、倍精度時に約 3.4TFLOPS、単精度時に約 6.0TFLOPS となる。以下に Cray Y-MP(1CPU) のベンチマーク値との比較結果を示す。

マシン	CPU 数	実効性能	性能比
CrayY-MP	1	260MFLOPS	1
TS/1	65536	3.4TFLOPS	13100

8 おわりに

本稿では、故障 PE の存在を想定した TS/1 における大規模にして高効率で高信頼な Wavefront Array 動作実現のための結合網アーキテクチャを提案した。

TS/1 ではプロセッサ間チェイニング機構と本結合網アーキテクチャを組み合わせることによって画期的な規則的超細粒度処理性能が実現される。他の通信パラダイムとの連携によりニューロと記号処理を統合した柔軟な情報処理などへの応用が期待される。

参考文献

[1] 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〜」, JSPP'93, pp.79-86, (1993.5)

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

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

[4] 児玉 他:「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, (1991.7)