10

日立スーパーテクニカルサーバー HITACHI SR8000の御紹介と今後の方向

(株)日立製作所 エンタープライズサーバ事業部
深川 正一

Hitachi Super Technical Server SR8000: Present and Future

Masakazu Fukagawa, Hitachi, Ltd., Enterprise Server Division

ABSTRACT

In the last ten years, a wide variety of large-scale technical computer systems have been installed. Under this circumstance, the Hitachi has developed a Super Technical Server SR8000 and the SR8000 was first shipped in 1998. The SR8000 has two main advantages. One is high-sustained performance of vector supercomputers. The other is high scalability of massively parallel processors. The SR8000 has realized these advantages by CO-operative MicroProcessors in single Address Space (COMPAS), Pseudo-Vector Processing (PVP) and remote DMA communication. Near future, two types of gap will become a serious problem for both high_end technical computer vendors and users. One is the gap between node performance and memory bandwidth and the other is the gap between node performance and inter-node communication throughput. We propose the utilization of Cache Memory-Layer and optical interconnection as a solution.

1. はじめに

近年、国内、海外において大規模な数値計算環境が構築されてきている。このような計算需要を背景に、従来のベクトル型スーパーコンピュータやスカラ並列コンピュータ以上に大きな処理能力を持った高性能コンピュータが求められるようになってきている。

このような状況の中、1998年5月、従来のベクトル型スーパーコンピュータの持つ実効性能の高さと、並列コンピュータの持つ高いスケーラビリティを併せもった、スーパーテクニカルサーバー「SR8000」を発表した。現在 SR8000 は、ノードあたり 8GFLOPS、9.6GFLOPS、12GFLOPSと3種の性能のノードを最小4ノードから最大512ノードまでに構成する事が可能となっている。

2. SR8000開発の背景

スーパーコンピュータとしてはベクトルタイプが長い間主流であった。これに対し、スーパーテクニカルサーバーSR8000 は複数の RISC プロセッサをメモリ共有で接続したSMPをエンジンの基調にしたものである。SR8000 を開発するに至った背景として2つの大きな流れがある。1 つ目はメモリの作成法である。ベクトルプロセッサは商用化の波にのり、その性能を向上させるべく、ベクトルパイプライン自体の複数

パイプライン化や、ベクトルエンジンの、メモリを共有した複数エンジン化を行って来た。メインメモリの作りに大きく左右されるベクトルマシンではエンジン部分の高性能化はメモリシステムの、エンジンの性能改良を上回るさらなる改良が要求される。より高性能のメモリシステムを作成する際のコストや物量が問題になり始めたのである。2 つ目の大きな流れは Cache 容量が程々に大きくなり、かつ、高い周波数で動作する RISC や CISC のマイクロプロセッサが市場に安価で供給されるようになったことである。メモリに引きずられてエンジン性能を引き出すのが難しくなってきたベクトルの世界に Cache を介して高速性能を確保するマイクロプロセッサを導入することに着目したのが2点目の背景である。

3. SR8000 の特徴

ベクトル機の高い実行性能効率に対し、スカラ並列機の効率が低いことが指摘されているが、これを SMP 並列の SR8000 ではどう克服しているかについて述べる。

(1) 協調型マイクロプロセッサ機構

ベクトルプロセッサにおいては演算パイプラインやロード/ストアパイプラインと称して、ベクトル処理を行うパイプラインが複数本実装されているのが最近の傾向である。ベクトルプロセッサにおいてはこの複数のパイプラインのロスを少なくして同時実行開始を狙ってハードウエアを組んでいる。SR8000 においても RISC マイクロプロセッサベースのアーキテクチャのもとで、ベクトルプロセッサと同等の高速性を実現するために、協調型マイクロプロセッサ機構 COMPAS(CO-operative Micro-Processors in single Address Space)をノードに採用して、複数のマイクロプロセッサを効率良く使用する機能を実現している。

図 1 に協調型マイクロプロセッサ機構の概念を示す。ノードを構成する各マイクロプロセッサ(図中の IP)は、演算処理を、それぞれ、別のベクトル要素に対して実行する。1 つのマイクロプロセッサが、「スタート命令」の発行により、他の全マイクロプロセッサに一斉に起動をかけ、全マイクロプロセッサは演算を開始する。各マイクロプロセッサは、演算が終了すると、「エンド命令」の発行により、演算を終結させる。上記起動(スタート)／終結(エンド)処理を高速に行う為に専用ハードウエアを用意してある。また、コンパイラが自動的に、ノード内の並列処理を行うので、ユーザーは、ハードウエアを意識することなく、コーディングできる。この点はユーザーがベクトルプロセッサにおいて、ベクトルパイプラインを複数本使用す

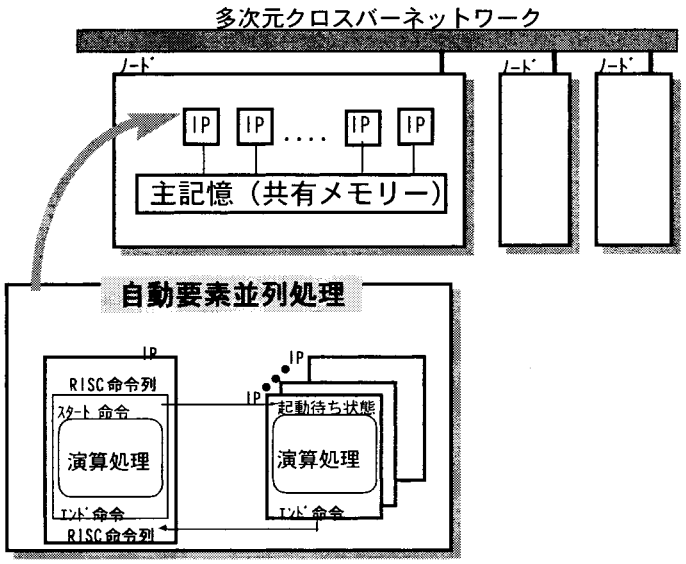

図 1 協調型マイクロプロセッサ機構

ることを意識しないのと同等で、特にSR8000において、プログラム上並列処理を意識する必要は無い。

ベクトルプロセッサにおいては同一の命令処理に関し歩調を合わせて、複数パイプラインを使用して同じ処理を行う。SR8000 においても同様の同一の命令処理を行えるが、SR8000 ではさらに、各 RISC プロセッサが別々の命令処理を行い、複数 RISC プロセッサにおいて全く異なる命令系列を並行して行うことが可能となっている。この機構は SMP 型の WS では可能な枠組ではあるが、SR8000 においてはメモリスループットをベクトルプロセッサ並みに強化しているので、各 RISC プロセッサからのメモリ要求にも応えることが可能となっており、ブロッキングを意識し、複数プロセッサからのメモリへの要求を減らす努力は不要となっている。この複数命令系列を並行して処理する機構により、細かいサブルーチン群がいくつもあり、ベクトル化率があげられなかったり、ベクトル化しても性能向上が得られない時に、各 RISC プロセッサに細かい処理を分担させることで実効性能を確保することが可能となる枠組を提供している。尚、ノード間の並列処理については、MPIによるコーディング等が必要となっている。

(2) 擬似ベクトル処理機構

擬似ベクトル処理機構(Pseudo Vector Processing)は、ベクトル型スーパーコンピュータにおけるベクトルロードに相当する機構である。

図2に擬似ベクトル処理機構の概要を示す。

擬似ベクトル処理機構では、①多数の浮動小数点レジスターと大容量の L1_Cache、②パイプライン動作可能なメモリー構成、③後続命令を止めない RISC プロセッサ制御により、演算器にデータを連続的に供給し、ベクトル型スーパーコンピュータ流のベクトル処理を実現している。

演算器にデータを連続的に供給するための手段として、(a)プリロード命令による主記憶から浮動小数点レジスターへの先読み、(b)プリフェッチ命令による主記憶からキャッシュへの先読みの 2 つを備えている。(a)のプリロード命令によるデータフェッチではキャッシュを介することなく、浮動小数点レジスタに直接データを持ってくるので、ストライドデータ等の非連続データを扱う際に有効である。通常の RISC プロセッサは Cache ベースのみでの動作となる為、とびとびの値を処理しようとすると常に Cache に転送するサイズの、ある固まった量のデータをハンドリ

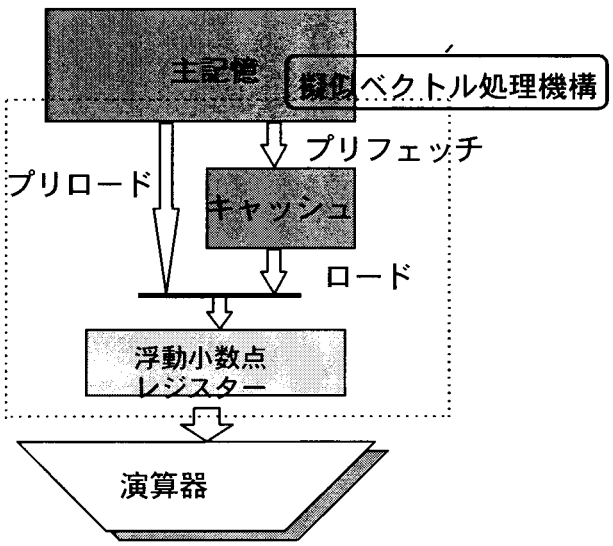

図 2 擬似ベクトル処理機構

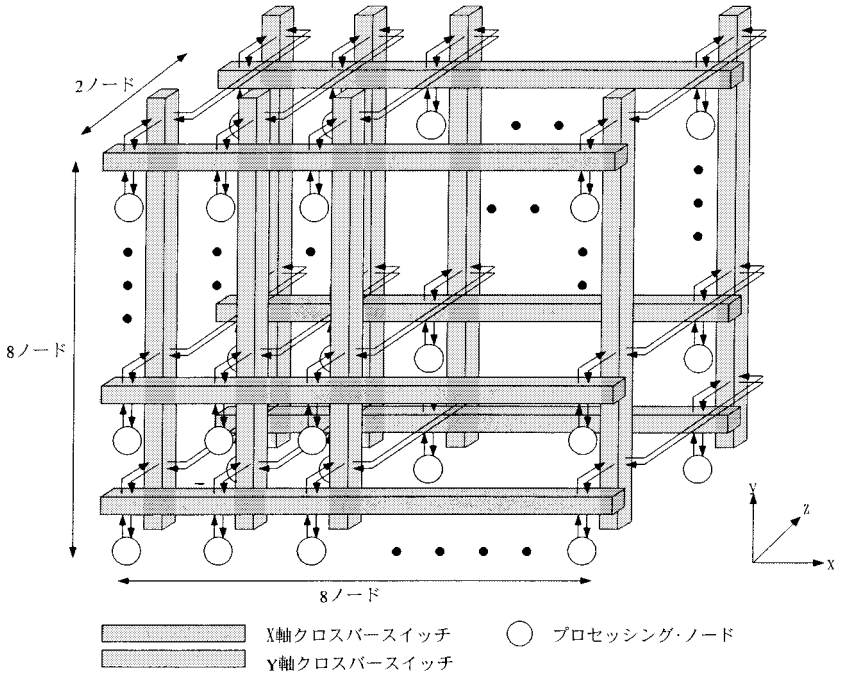

図3 3次元クロスバ(128ノード構成時)

ングすることになり、ただでさえ充分でないメモリスループットをいたずらに消費することになり、効率を著しく低下させる原因となっている。(b)のプリフェッチ命令によるデータフェッチでは連続した固まりのデータをベクトルレジスタよりも大容量のL1_Cacheに連続的に持ってくることができる。SR8000/L1_Cache容量の、S-3800(当社ベクトル型スーパーコンピュータ)のベクトルレジスタ容量との比は同数のパイプライン本数同士で比較した場合、8倍にもなっている。

(3) 多次元クロスバネットワーク

SR8000のネットワーク構成例(128ノード構成時)を図3に示す。ノードは3次元に配列され、ノード間をクロスバスイッチが接続している。ここで、128ノード構成は、8(X)× 8(Y)× 2(Z)構成となっている。そのため、Z軸方向は、ノード台数が2になるので、直接、Y軸クロスバスイッチの入出力同士を接続する構造になっている。

多次元クロスバネットワークは、論理的に他のトポロジーの結合網を大部分包含でき、しかも広範囲の転

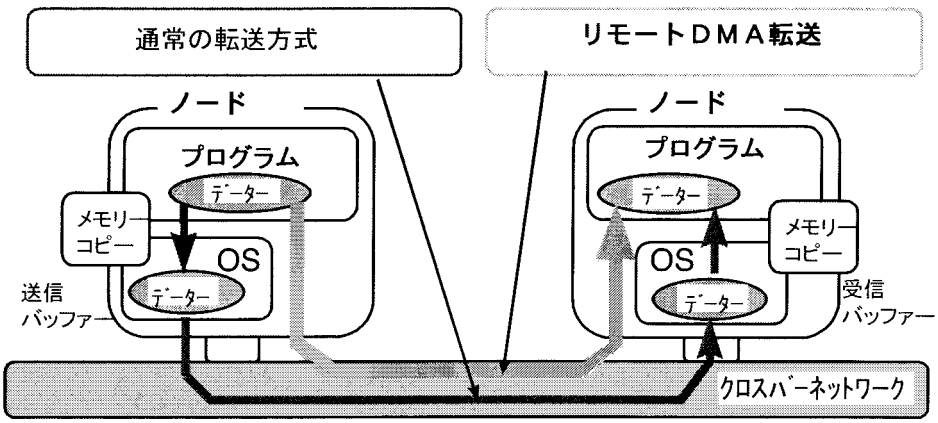

図4 リモートDMA転送の概略

送パターンにおいて、結合網での制約(衝突)による性能の低下を起こすことなく、通信が可能となるネットワークである。

(4) リモート DMA 転送

図4にリモート DMA 転送の概略を示す。

通常のノード間の通信はOSを介した通信であり、送信側ノードは、ユーザー空間内のデータを OS の送信バッファーにコピーし、受信側ノードに転送する。受信側ノードは、 OSの受信バッファーにこれを受け、受信するユーザープロセスから要求を受けると、データをユーザー空間にコピーしている。

これに対し、リモート DMA 転送は、OS を介さず、ユーザー空間からユーザー空間へ直接データを転送する。これによって、通信処理オーバーヘッドを低減でき、ネットワークの高速性を最大限に活用し、大量のデータをノード間で高速に転送することができるだけでなく、小さいサイズのデータ転送も立ち上がりレイテンシを抑えてあるので効率良く転送することが可能となっている。

4. SR8000 の今後の方向

SR8000 は大規模 JOB をランさせるべくTflops クラスのスーパーコンピュータを現実的なものとして実現させることが目標である。現実的なものの指標としては設置面積、消費電力、マシンの価格があげられる。次世代も上記の考え方を踏襲する方向で開発したいが、次の世代の Tflops マシンを作る上で課題が2つある。メモリの問題とノード間の接続の問題である。Cache を利用してマイクロプロセッサの性能を引き出すとしてもメモリからの転送性能を全く考慮しなくて良いわけではない。FFT などメモリを一通り読み書きするだけで Cache 内のデータを再利用しないものではメモリ性能が相変わらず重要である。だが一方でエンジンの性能は今以上に向上するので、向上の仕方が緩慢なメモリ用 RAM をどう構成してどう使用するかが鍵となり、2つの問題が存在する。一つはエンジン性能を支えるだけを考えてメモリインターリーブ数をそのまま増やすとたいへんなメモリバンク数になってしまうというスループット確保の問題がある。(図5) もう一つは計算エンジン部とメモリ自体の

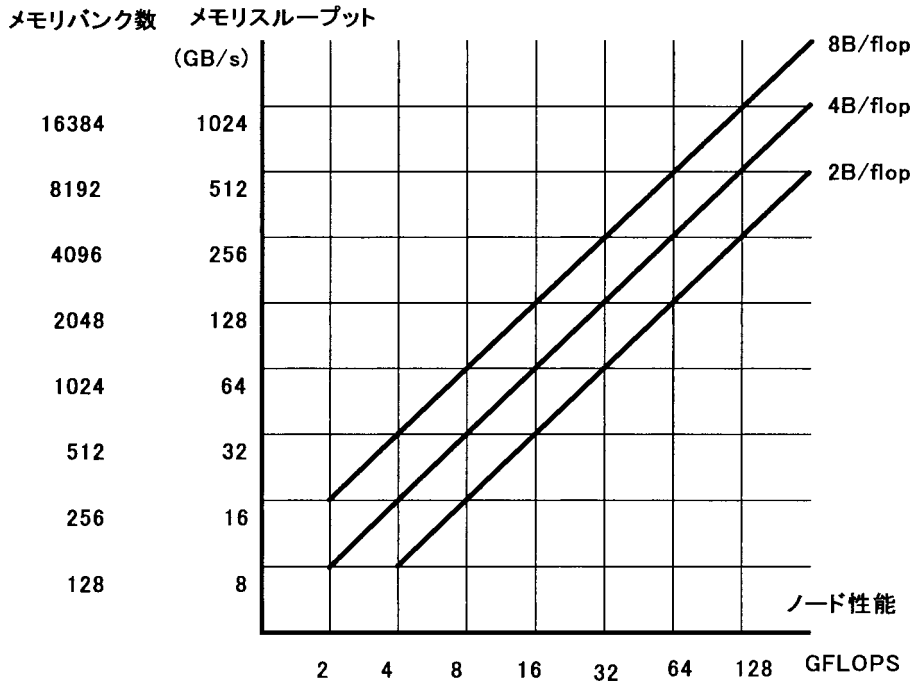

図5 メモリスループット/ノード性能の関係

物理的距離が縮まらない限り、計算エンジン部でのクロックでその物理的距離が刻まれてしまうので相対的なサイクル数がどんどん延びていってしまうというメモリレイテンシの問題が存在する。これらのことを考えると WS 等に導入されている階層化されたCacheメモリレイヤの導入は不可避と思われ、かつ、メモリからのスループットを引き出す為に固め読みのような制御方法を工夫する必要性が大きくなる。

もう一つの問題は高性能エンジンである固まり(ノード)を接続する計算機内ネットワークのデータ転送性能である。エンジンが高性能化するのに対し、それを接続するネットワークも同様の向上度で改良されなければエンジン性能とネットワーク性能とが乖離して行ってしまう。(図6) エンジン性能にあわせてネットワーク性能を向上させようとすると従来通りの電気信号による接続ではネットワークを構成する信号本数が爆発し始めることになる。また、転送距離の問題も生じてくるので近く同士を接続するトポロジ形態も視野に入れる必要が出てくる。

システム全体でエンジン同士をある程度のスループットで接続し、かつ、その接続に制約を加えないという意味ではもはや光による接続しか解がないように見える。通信の世界で社会基盤等への光ケーブルの利用で光ケーブルが大量供給されることでコストが下がることもそう遠い先ではなさそうであることが期待の一端にある。

5. まとめ

Pflops へ向けて高性能科学技術計算機の開発が続けられていくがメモリへ依存したマシンへの限界が見え始めてくること、エンジン性能が上がってきた時にエンジン間を結合するネットワークの性能を引き上げるのは光による接続が不可避であること、について触れた。大規模マシンを開発するにあたり、価格、電力、設置面積、も重要な要素になるのでこれらも考慮に入れて考えることを今まで以上に要請されている。

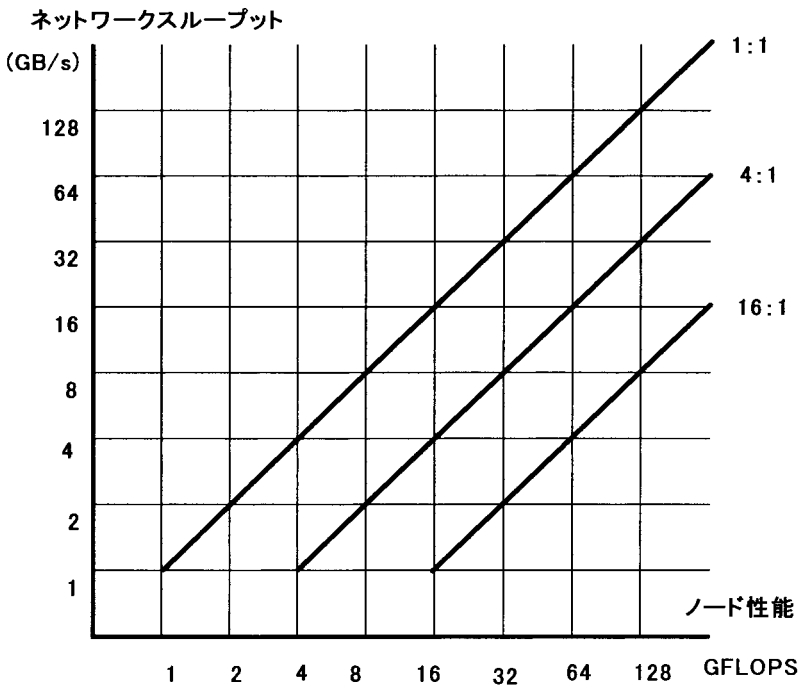

図6 ノード間スループット/ノード性能の関係