解 説 高速プロセッシングデータバス技術

3. スーパーコンピュータSX-4におけるデータ供給能力

Data Transfer Capability in Supercomputer SX-4 by Takashi NISHIKAWA, Takashi HAGIWARA (Computer Division, 1st Computers Operations Unit, NEC Corporation), Noriyuki ANDO (Computer Division, NEC Engineering Ltd.) and Yoko ISOBE (2nd Computer Engineering Department, NEC Kofu Ltd.).

西川 岳[1] 萩原 孝[1] 安藤憲行[2] 磯部洋子[3]

1 日本電気(株)
2 NECエンジニアリング(株)
3 甲府日本電気(株)

1. はじめに

スーパーコンピュータのことを“Number Cruncher”と呼ぶことがある．数値データを「バリバリ」と貪り食うように処理することからつけられた呼称のようである．このスーパーコンピュータのシステム最大演算能力は数百GFLOPSを超え，最近ではTFLOPSを超えるシステムも提供されてきている．

これらのシステムでは時代の最先端テクノロジーを駆使しマシンサイクルを短縮させ，最新のアーキテクチャを採用してマシン内部レベルから装置レベルまで各階層の並列処理を実装して同時並列処理を強化するなどさまざまな工夫が施されている．

この貪欲なまでの胃袋を満たすための能力，すなわち高速演算処理能力に見合うだけのデータをメモリから演算器に供給するためのデータ供給能力が計算機の処理能力を決定する重要な要因となっている．

計算機の利用者が享受する計算機の能力は，単にカタログに記載された演算ピーク性能だけではなく，むしろその演算器へのデータ供給能力に大きく左右されるため(科学技術計算分野ではその能力の方が支配的である場合がある)，計算機の処理能力を考える場合，演算性能だけではなくデータ供給能力についても十分に配慮する必要がある．

本稿ではスーパーコンピュータにおけるデータ供給能力について，性能の見地から演算能力との関係を考察したのち，高い実効性能を実現したSX-4で採用している高速データパス技術，およびその性能実測値を紹介する．

2. スーパーコンピュータにおける性能

スーパーコンピュータの性能を表す指標として，『ピーク性能(演算能力)』『メモリ性能(転送能力)』『入出力性能』などの性能指標がよく使用される．しかし，実際に計算機を利用する場合にはそうした個々の性能指標よりも，利用者が実際に体感する計算機の性能，すなわち計算機の総合力を示す『実効性能』が重要な指標となる．プログラムの中にはある特定の性能がとくに要求される場合もあるが，多くのプログラムで高い実効性能を実現させるには上記各能力がバランスよく釣り合っていることがきわめて重要となる．

SX-4は，多重演算パイプライン構成のベクトル型計算機であり，演算能力という観点からは，1マシンサイクル(8ns)あたり16個の浮動小数点演算を実行する(図-1参照)ことによってピーク性能『2GFLOPS』(16op/8ns = 2GFLOPS)を実現している．そしてこの演算能力を十分に引き出し，システムとして高速な処理を実現するため

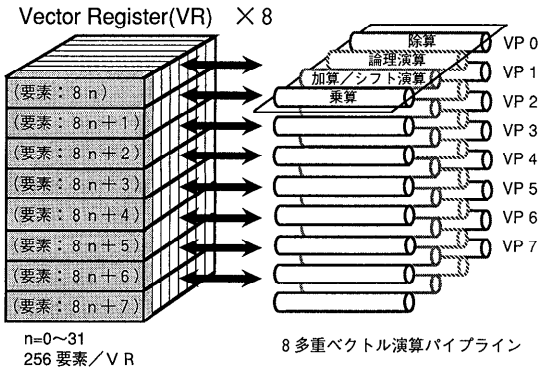


図-1 SX-4ベクトル演算器構成

表-1 LFK14 におけるロード命令数と演算命令数の割合

		ロード	演算	ロード：演算
Kernel 1	流体	3	5	0.60 ： 1
Kernel 2	MLR 内積	10	10	1 ： 1
Kernel 3	内積	2	2	1 ： 1
Kernel 4	帯型連立 1 次方程式	2	2	1 ： 1
Kernel 5	三角化消去上三角	6	6	1 ： 1
Kernel 6	三角化消去下三角	6	6	1 ： 1
Kernel 7	状態方程式	9	16	0.56 ： 1
Kernel 8	P.D.E.積分	15	36	0.42 ： 1
Kernel 9	整数予測	10	17	0.59 ： 1
Kernel 10	差分予測	10	9	1.11 ： 1
Kernel 11	総和	2	1	2 ： 1
Kernel 12	差分	2	1	2 ： 1
Kernel 13	2 次元粒子推進	15	9	1.67 ： 1
Kernel 14	1 次元粒子推進	8	14	0.57 ： 1
平 均				1.04 ： 1

には，これらの演算器に滞りなくデータを供給する必要がある．

この意味から演算を実行するプロセッサとデータを格納する主記憶の間を結ぶパス＝メモリネットワークは CPU の演算能力を引き出す重要な鍵になる部分であり，この能力が十分でない場合，主記憶装置から十分なデータ供給が行われないため，演算器が十分に動作せず，高い実効性能を引き出すことができない．

ここで演算とメモリのデータ供給能力のバランスに関して考察する．各種ベンチマークプログラムにおけるロード命令数と演算命令数の比を調査したところ，その比は約 1 ： 1 であった．これは 1 演算あたり 1 オペランドを供給できればバランスのとれた性能を実現できることを意味する．たとえば，科学技術計算処理で頻繁に使用される DO ループを集めた著名なベンチマークプログラムである Livermore Fortran Kernel（LFK）では，ロード命令と演算命令の比は約 1 ： 1 となっている（**表-1**）．

データ供給能力と演算能力の比を 0.5 ： 1 にすれば装置を比較的少ない金物量で安価に実現することができるが，そうした場合すべてのプログラムにおいて高い実効性能を得ることは期待できない．

図-2，**図-3** は LFK のあるループの動作におけるデータ供給能力と演算能力の関係を示したタイミングチャートである．図-2 はデータ供給能力と演算能力の比が 1 ： 1 のケース，図-3 は比が 0.5 ： 1 のケースのタイミングチャートである．図-3 の比が 0.5 ： 1 のケースでは演算器がメモリからのデータ供給を待つ形となるため，演算と演算の間にすき間が多く（図-3 内の網かけ部）演算能力を十分に発揮できないことがわかる．

このような分析結果から SX-4 では，ロードと演算の能力バランスが 1 ： 1 になるように設計を行った．これにより，幅広い実用プログラムでも高い実効性能を得ることができる．

3. SX-4 の高速データ転送能力

3.1 SX-4 の共有メモリシステム

SX-4 共有メモリシステム（プロセッサ（CPU）－主記憶装置（MMU）間ネットワークと主記憶装置）のハードウェア構成について説明する．

前章で述べたように SX-4 は，演算（浮動小数点演算： FLOP）あたり 1 データ（8 バイト）のデータ供給能力（8 バイト／ FLOP）を実現している．SX-4 の演算性能は，1CPU あたり 2GFLOPS である．したがって，共有メモリシステムは，各

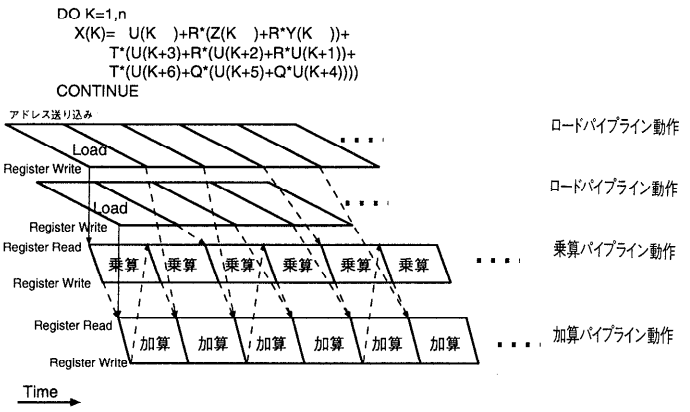


図-2 データ供給能力：演算能力＝ 1 ： 1(2 ロード： 2 演算)

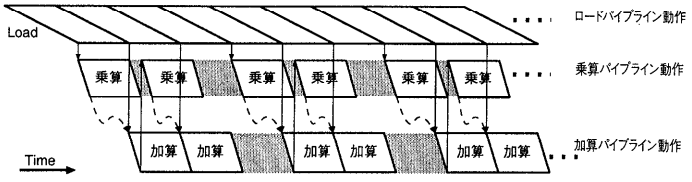


図-3 データ供給能力：演算能力＝ 0.5 ： 1(1 ロード： 2 演算)

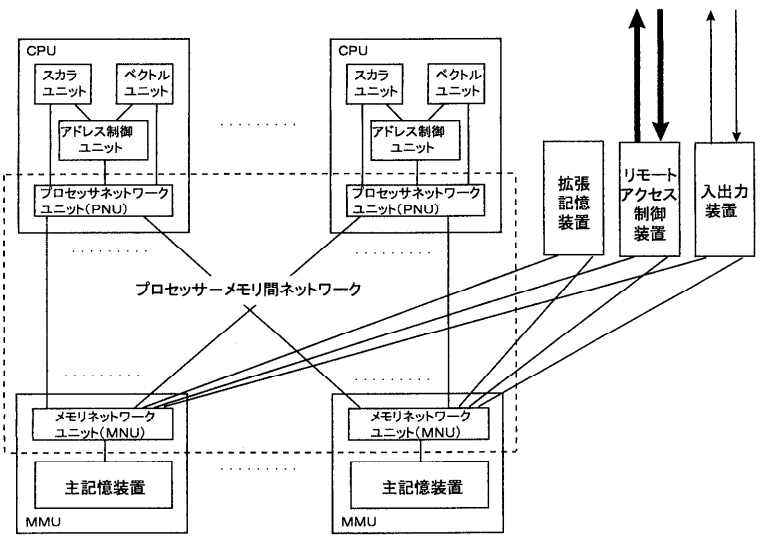


図-4 プロセッサーメモリ間ネットワーク構成

CPUに対して16GB／秒，最大構成である32CPU構成時は512GB／秒のデータ供給能力を提供しなければならない．また，この広いメモリバンド幅を実現するとともにメモリシステムとして，CPU台数(演算性能)に比例したハードウェアのスケーラビリティを確保する必要がある．SX-4では，最小構成で1CPU－1MMU，以降2MMU，4MMU，8MMU，…のように最大構成で32CPU－32MMU構成の共有メモリシステムを実現している．

3.2 プロセッサーメモリ間ネットワーク

プロセッサーメモリ間ネットワークは，プロセッサネットワークユニット(PNU)とメモリネットワークユニット(MNU)から構成され，PNUはCPU側のネットワーク制御，MNUはMMU側のネットワーク制御を行う．PNUは各CPU内に，MNUは各MMU内に分散配置されている．**図-4**にプロセッサーメモリ間ネットワークの構成を示す．

CPUおよびMMUは相互に接続するためのインタフェースポートを各32ポートずつもつ．CPUとMMU間の接続は，32CPU-32MMU構成の場合，各CPUとMMUの1つのインタフェースポート同士が接続される．すなわち，CPUの32個のインタフェースポートのそれぞれと，32個のMMU間で1本ずつのパスが張られる．

1パスは4バイトデータ幅のロード／ストア兼用パイプラインであり，ロード／ストアとも，1メモリアクセスリクエストの8バイトデータ転送には2クロック必要とする．したがって，CPUあたりのデータ転送能力(メモリバンド幅)は16GB／秒(＝256B/16ns(32要素／2クロック))，32CPUで512GB／秒(＝8192B/16ns(1024要素／2クロック))となる．

各CPU内のPNUは，8バイト幅入力16ポートと4バイト幅出力32ポートをもつクロスバースイッチであり，入力ポートはベクトルユニットに，出力ポートはMNUに接続される．ベクトルロード／ストア命令が発行されると，入力16ポートに対し16要素のベクトルリクエストが並列入力される．出力側は出力32ポートに対しメモリアドレスがインタリーブ式に振られているので，8バイト幅連続，または，奇数飛びの要素間距離のベクトルメモリアクセスではPNUにて競合が発生しない．偶数飛びでは8バイト×2飛びで1/2に，8バイト×4飛びで1/4に，…，8バイト×64飛びで1/64にスループットが低下する．A(ID(K))のようなリストベクトル(間接指標アドレス)アクセスにおいてもアドレスに規則性がないため，一般的にPNUにて競合が発生する．

各MMU内のMNUは，4バイト幅入力36ポートと4バイト幅出力32ポートをもつクロスバースイッチであり，入力36のうち32ポートはCPU，4ポートは拡張記憶装置／入出力装置／リモートアクセス制御装置に，出力ポートは主記憶装置に接続される．MNUの入力ポートは異なるCPUが接続されるため，入力ポート間のアクセスパターンの相関性がなく競合が発生しやすい．したがって，高いスループットを確保するためには競合に強いスイッチ構成を実現する必要がある．

3.3 主記憶装置

主記憶装置は，最小256MBから最大16GBの主記憶容量をサポートしている．主記憶装置は1枚から32枚のMMUから構成され，MMUあ

表-2 主記憶装置諸元

項　目	諸　元
記憶容量	256MB ~ 16GB
インタリーブ数	32way ~ 1024way
記憶素子	2M/4Mbit SSRAM
最大転送速度	16GB/s ~ 512GB/s
バンクサイクル	2クロック

たり32way(バンク)にインタリーブされている．したがって，最大構成時には1024wayのインタリーブ数を実現し，共有メモリシステムの問題点としてよく指摘されるメモリ競合による性能低下を軽減している．

また，メモリ素子としてアクセス時間15nsの高速SSRAM(Synchronous Static Random Access Memory)を使用することにより，バンクサイクル待ち☆の発生をなくしている(バンクビジーレス)．すなわち，バンクサイクルは2クロックであり，MNUからメモリバンクへデータを転送するのにも同様に2クロックを要するため，メモリバンクにおけるバンクサイクル待ちを考える必要はない．メモリアクセスの競合はすべて，PNU，もしくはMNUのポート競合として現れることになる．

主記憶装置の諸元を**表-2**に示す

4. SX-4における実効転送能力にかかわる技術

SX-4は多重演算パイプライン構成であり，ベクトル演算能力を引きだすためには，多重要素のベクトルロード命令に対し滞りなくデータを供給する必要がある．また，最大32CPUの共有メモリ型密結合方式のマルチプロセッサ構成であり，台数効果を得るためには，複数のCPUからのメモリアクセス競合に対しても，効率よく調停処理を行い，極端な性能低下の起きないネットワーク構成にする必要がある．ただし，単純にネットワークに対し物量を投じることにより性能向上をはかるのではなく，同一の物量コストにおいて実効性能を最大限に引き出すことが重要となる．

前章ではSX-4プロセッサーメモリ間ネットワークのネットワーク構成(トポロジー)について述べたので，本章では，SX-4が採用したネットワーク制御方式，およびスイッチ構成方式の特徴について述べていく．

4.1 ネットワーク制御方式

ネットワークの制御方式として，PNU/MNUの各クロスバースイッチに競合調停回路を内蔵させ，各スイッチが独立にルーティング処理を行う分散制御方式を採用している．また，8バイト幅メモリアクセスを1リクエスト(パケット)とするパケット交換方式を採用している．たとえば，256要素長のベクトルメモリアクセス命令ならば，256個のリクエストに分解され，各リクエストが独立にセルフルーティングを行う．

SX-4はマルチプロセッサ構成であると同時に，各CPU，およびMMUが物理的に分散配置構造をとるため，集中制御方式を採用すると競合調停制御が複雑になると同時に，制御部の物量コストが非常に大きくなる．そこで，スイッチを小規模化し，各スイッチでの分散制御とすることにより，ネットワーク全体のコストを低減する．

しかし，分散制御方式を採用すると以下に示すデメリットが生じてしまう．これらの問題に対するSX-4の方策をあわせて以下に示す．

(1)リクエストのランダムパターン化

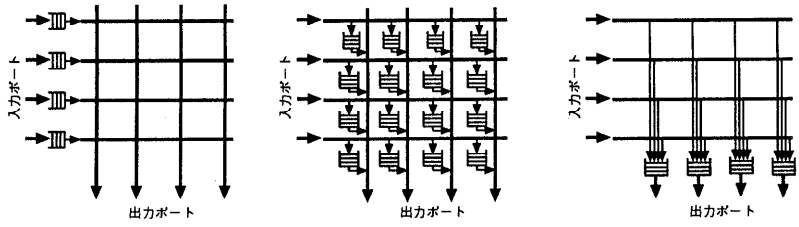
各リクエストが独立にネットワーク上を流れることにより，ネットワークルーティングパターン，バンクアクセスパターンがランダムになってしまう．したがって，ベクトルアクセスの特徴である規則的な飛びアクセスパターンを生かし，競合を回避する制御方法をとることができない．

→ランダムアクセスに強いメモリ／ネットワーク構成とする．たとえば，SSRAM採用によるバンクビジーレスとなるメモリ構成．また，高スループットが得られるクロスバースイッチの採用．

(2)リクエストのルーティングタイミング不定

分散制御方式では各スイッチが独立に処理を行うため，スイッチでのルーティングタイミングや，CPUへのロードリプライ返却タイミングを予測することはできない．そのため，各スイッチにおいてリクエスト到着後に競合調停を開始せざるを得なく，競合調停とデータルーティングが逐次的になる．さらに，CPUへのリプライの返却タイミングを予測することができず，チェイニングの前準備ができない．これらはメモリアクセスのレ

☆ メモリアクセスリクエストの発行間隔に対して，バンク(メモリ素子)のアクセスサイクルが長い場合，同一バンクにアクセスが連続して発行されるとバンクのアクセスサイクル待ちが発生する．



(1)入力バッファ方式　(2)クロスポイントバッファ方式　(3)出力バッファ方式

図-5 クロスバースイッチのバッファ構成方式

イテンシを増加させる要因となる．

→ネットワークをデータ系統と制御系統の2系統もたせ，制御系統は各PNU/MNUスイッチの調停制御のみ実行し，アドレス変換ステージやメモリアクセスステージを省く．これにより，制御系統はデータ系統に対し先行して競合調停制御の実行ができ，この結果をデータ系統のルーティング情報とすることにより，データ系統での競合調停時間をなくすことができる．また，制御系統を通じCPUへのリプライ返却タイミングをリプライデータ到着前に知ることができる．

(3)メモリアクセス順序不定

あるプロセッサが発行したメモリアクセス命令が，ほかのプロセッサからみると，その命令発行順どおりにみえない．たとえば，プロセッサAが共有データ書き込み後，書き込み完了フラグのセットを行い，プロセッサBがこのフラグのセットを確認した後で，共有データを読みだしたとしても，プロセッサAが書き込んだ共有データが読みだされる保証はない．これは，共有データ書き込みと書き込み完了フラグセットとのアクセス順序の保証ができないためである．

→メモリアクセス順序の保証方式として，ウィークコンシステンシモデルを採用し，ソフトウェアに対し，このモデルに従った共有データ受け渡しを行う制約を課す．

すなわち，先行するすべてのメモリアクセス命令の完了を待って実行する同期命令を用意し，共有データ受け渡しを行う場合には，共有データ書き込み後，同期命令を用いた書き込み完了フラグのセットを行うことにより，順序関係を保証する．共有データ読み出し側は，このフラグのセットを確認してから共有データの読み出し処理を開始する．

4.2 クロスバースイッチ構成

クロスバースイッチの性能特性はスイッチ内のバッファ構成に強く依存する．クロスバースイッチはバッファの配置構成により，

(1)入力バッファ型

(2)クロスポイントバッファ型

(3)出力バッファ型

に分類できる．これらの構成についてはATM交換機の分野で広く研究応用されている[5),6)]．図-5に各バッファ構成を簡単に示す．

クロスポイントバッファ型，および出力バッファ型はランダムな入力アクセスパターンに対して，ほぼ100％の実効スループット(バンド幅)が得られる．とくに出力バッファ型は，入力アクセスパターンの偏りに対して強いタイプである(たとえば，あるプロセッサのみメモリアクセス頻度が高いケースに相当する)．しかし，両バッファ構成ともバッファサイズが入出力ポートの2乗オーダとなり，バッファ収容性の点からみてLSI化が非常に難しい．

入力バッファ型は，入力バッファ構成によりFIFO(First In First Out)タイプとRIRO(Random In Rondom Out)タイプの2種類に分けられる．FIFOタイプは競合調停で敗れたリクエストがFIFOの先頭で滞り，他出力ポートに向かう後続リクエストをブロックするHOL(Head Of Line)ブロッキングが生じ，実効スループットが約60％に低下してしまう特性がある．RIROタイプはHOLブロッキングを回避することができ，ほぼ100％の実効スループットが得られる．しかし，バッファ制御は複雑となり高速化に向いていない．

FIFOタイプは論理構成も簡単であり，バッファ物量も大きくない．しかし，高いネットワーク性能を得るためには，実効性能が60％に低下す

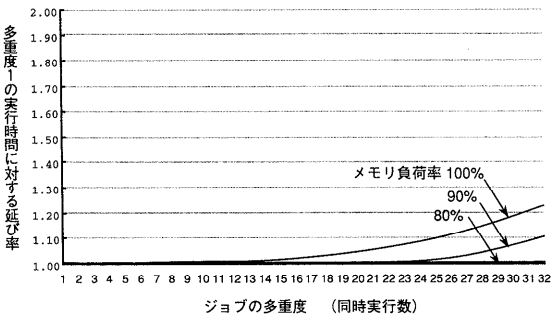


図-6 メモリアクセス競合による性能低下

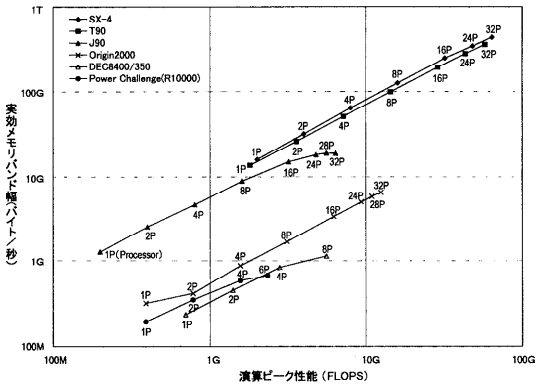


図-7 Stream(Triad:A(I)=B(I)+q*C(I))ベンチマークにおける演算ピーク性能と実効メモリバンド幅の関係

るFIFOタイプを採用することはできない．6割のスループットしか得られないならば，ネットワークに高価な物量を投じる意味が半減するからである．

SX-4では，プロセッサーメモリ間ネットワークを2クロックサイクルで1リクエストを転送する4バイト幅パス構成とし，さらにクロスバースイッチをFIFOタイプの入力バッファ型を採用するが，2クロックサイクルで入力ポートに到着するリクエストに対し，1クロックサイクルごとに競合調停しリクエストを通過するバッファ構成としている．このクロスバースイッチは2クロックサイクルに2リクエスト通過できると考えることができ，通常の入力バッファ型の倍の転送能力をもつことになる．このタイプはランダムな入力アクセスパターンに対して約90％の実効スループットが得られ，他タイプの複雑なバッファ構成をとることなく高いスループットを実現することができる．

表-3 Stream(Triad)ベンチマークにおける1演算あたりのデータ転送能力

	CPU数	ピーク演算性能(Gflops)	実測メモリバンド幅(GB/秒)	1演算あたりのデータ転送能力(w/flop)
SX-4 *	32	64.00	436.95	0.85
T90	32	57.60	359.27	0.78
J90	32	6.40	18.87	0.37
Origin2000	32	12.48	6.54	0.065
Power challenge (R10K)	6	2.34	0.67	0.036
DEC 8400 5/350	8	5.60	1.13	0.025

* SX-4以外は，登録値

5. 実測性能

SX-4の共有メモリシステム性能に関する実測結果を紹介する．

5.1 ジョブスループット性能

まず，プロセッサ間のメモリアクセス競合による性能低下について，シングルプロセッサでの実行時間を1としたときに対して，多重度(同時実行ジョブ数)を変化させた場合の実測結果を図-6に示す．

ここで使用しているメモリ負荷率とは，ピーク演算性能を得られるような演算(加算＋乗算)を仮定したとき，評価式の演算数とメモリアクセス数の比率であり，負荷率100％の評価式は以下のとおりである．

```
DO 10 I=1,3*N,3
W(I)=(((((X(I)+C0)*Y(I)+C0)*Z(I)+
      X(I+1))*Y(I+1)+Z(I+1))*X(I+2)+
      Y(I+2))*Z(I+2)
10 CONTINUE
```

負荷率90％場合は，右辺の配列変数の1つをスカラ変数(C0)に，80％の場合は2つをスカラ変数にしている．32CPU－32MMUの最大構成下で，多重度32(32ジョブ同時実行)のメモリ負荷率100％の場合で，多重度1の実行時間の1.23倍，90％の場合で1.12倍，80％の場合では1.0倍の実行時間であった．多重度32で負荷率100％とは，ジョブの実行期間中に全プロセッサが16GB／秒の最大データ転送性能でメモリにアクセスしている状態であり，実アプリケーションでは，かなり希なケースである．事実，SX-4で行った各種の実アプリケーションを用いたスループットテストでも，ほぼ1％～3％程度

の実行時間の延びしか観測されなかった.

5.2 Stream ベンチマーク[7]

Stream ベンチマークは，コンピュータシステムのメモリアクセスパスの実効バンド幅(MB／秒)を計測することを目的としたベンチマークであり，Copy(A(i)=B(i))，Scale(A(i)=q*B(i))，Sum(A(i)=B(i)+C(i))，Triad(A(i)=B(i)+q*C(i))の4つの式から構成されるベンチマークである．**図-7** は，Triad について，演算ピーク性能と実効メモリバンド幅の関係を SX-4 などのベクトル型コンピュータシステムとマイクロプロセッサベース(キャッシュベース)システムの結果である．同図より，同一演算ピーク性能あたりの実効メモリバンド幅は，ベクトル型コンピュータシステムがマイクロプロセッサベースシステムに対して，およそ6～15倍強力であることがわかる．また，実測結果を基にした FLOP あたりの転送ワード数(1word＝8バイト)を**表-3** に示す.

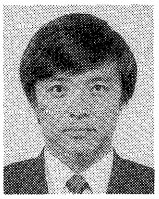
6. おわりに

本稿では計算機の性能は単に演算性能だけではなくデータ供給能力が重要であることを述べた．しかしながら計算機に対し湯水のごとく費用をかけることができない社会情勢下で，いかに低コストで計算機のもつ能力を引き出すかは，計算機を構築する側(ハードウェア)からも，それを利用する側(ソフトウェア／アプリケーション)からも重要な課題であり，今後は従来以上にいかに両者が一体となってバランスのとれた計算機を構築していくかが「計算科学」の発展にとって重要な鍵になると考える.

参考文献

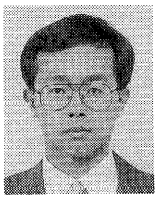
1) NEC 技報，スーパーコンピュータ SX-4 シリーズ特集，Vol.48，No.11(Nov. 1995).
2) Watanabe, T.: Architecture and Performance of NEC Supercomputer SX System, PARALLEL COMPUTING, Vol.5, No.1&2, pp.247-255(July 1987).
3) Kinoshita, K. and Takenaga, S.: SX SUPER-COMPUTERS, NEC Reserach & Development, No 96, Section II.1.4, pp 104-109(Mar. 1990).
4) Nishi, N. et al.: SX-4 Architecture for Scalable Parallel Vector Processing, Proc. of International Symposium on Parallel and Distributed Supercomputing, pp.45-50(Sep. 1995).
5) 鈴木 洋他：ATM 交換機アーキテクチャの一検討，信学技報 SSE88-60(July 1988).
6) 尾家祐二他：B-ISDN 用 ATM 交換機のアーキテクチャ，情報処理，Vol.33，No.2，pp.134-142(Feb. 1992).
7) http://www.cs.virginia.edu/stream/

(平成9年4月1日受付)



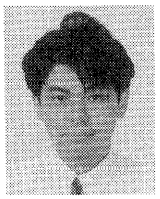
西川 岳(正会員)

1955年生．1982年大阪大学大学院工学研究科応用物理学科応用物理学専攻修士課程修了．同年 NEC 入社．以来スーパーコンピュータ SX シリーズのハードウェア開発・設計・科学技術計算性能評価に従事．現在コンピュータ事業部第四技術部技術課長．1984年本会論文賞受賞.



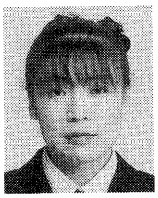
萩原 孝

1964年生．1989年早稲田大学大学院理工学研究科電気工学専攻修士課程修了．同年 NEC 入社．以来スーパーコンピュータ SX シリーズのハードウェア開発，性能評価に従事．現在，コンピュータ事業部第四技術部主任.



安藤 憲行(正会員)

1964年生．1989年東北大学大学院工学研究科情報工学専攻修士課程修了．同年 NEC 入社．以来スーパーコンピュータ SX シリーズの研究開発，ハードウェア開発に従事．現在，NEC エンジニアリング(株)コンピュータ事業部第1コンピュータ技術部主任.



磯部 洋子

1965年生．1987年山梨大学工学部環境整備工学科卒業．同年 NEC 甲府(株)入社．以来スーパーコンピュータ SX シリーズのハードウェア開発・科学技術計算性能評価に従事.