解　説　高速プロセッシングデータバス技術

4. 高並列スーパコンピュータ VPP700E におけるデータ供給能力

Data Transfer Scheme of the VPP700E Parallel Supercomputer by Kenichi SAKAI (High Performance Computing Group, 1st Development Division Hardware System Department, FUJITSU LIMITED).

坂　井　賢　一[1]

1 富士通(株)HPC 本部第 1 開発統括部技術部

1. はじめに

スーパコンピュータは，ベクトル型計算機アーキテクチャを基本として，マシンサイクルタイムの高速化，および単一プロセッサの設計改良により発展してきた．一層性能を向上させるため，複数のプロセッサを並列に動作させるマルチプロセッサ構成のスーパコンピュータも開発された．

従来技術では達成できない高性能を実現するため，富士通は分散メモリ型ベクトル並列計算機 VPP500 を開発した[1)~3)]．VPP500 は，GaAs および BICMOS LSI を使用し，単一 PE (Processing Element) 性能 1.6GFLOPS，システム最大性能 355GFLOPS を提供する．VPP500 は，科技庁航空宇宙技術研究所と富士通の間で 1989 年に開始された実現検討および共同研究の成果に基づき商用化したシステムであり，この共同研究は数値風洞(NWT)の開発につながった．

一方，CMOS テクノロジーの進歩とともに，マイクロプロセッサの性能は著しく向上し，それにともないワークステーションの価格性能比は向上してきた．また，設備投資の効率化要求から，スーパコンピュータも，高性能だけでなく，設置性，および運用性の向上もますます求められるようになってきた．

このような状況の中で，VX-E/VPP300E/VPP700E シリーズは，VPP500 のベクトルパラレルアーキテクチャを踏襲し，また論理 LSI には CMOS を採用，メモリ素子には SDRAM を採用することにより，スーパコンピュータに必要とされる性能を維持するとともに，価格性能比，使いやすさの向上を主眼に開発された．VX-E はオフィスに設置が可能な高性能計算サーバを，VPP300E は最大 38.4GFLOPS の高性能センターマシンを，VPP700E は最大 1.2TFLOPS の高性能をそれぞれ提供する[4)]．

本稿では，本 VX-E/VPP300E/VPP700E シリーズ(以下，VPP700E と称す)におけるデータ転送技術について紹介する．

2. VPP700E のアーキテクチャ

従来，並列処理方式として，ベクトルプロセッサ数台～数十台を共有メモリで結合した共有メモリ結合型並列方式，および RISC マイクロプロセッサ(スカラプロセッサ)を百台ないし千台程度ネットワークで結合したメモリ分散型並列方式 (MPP:Massively Parallel Processor) が開発されてきた．

共有メモリ結合型並列方式では，演算性能に見合うメモリスループットを確保するためのハードウェア実現技術に限界がある．台数効果を維持して接続が可能なプロセッサ数は数十台程度に抑えられる．すなわち，メモリ系実現の壁により性能のスケーラビリティが頭打ちになる．

一方，MPP では以下に示す原因により，その実効性能はピーク性能の約 2 %～ 30 %程度に留まっている．

- RISC マイクロプロセッサの性能はキャッシュ動作に依存しており，スーパコンピュータで頻発する広域データアクセスに対してキャッシュミスにより強固な性能を発揮できない．
- 千台規模のプロセッサを結合する必要上採用しているネットワークが低性能であり，隣接以外のプロセッサ間通信に対して高い転送効率および使用率を実現できない．

我々は，以上のような共有メモリ結合型並列方

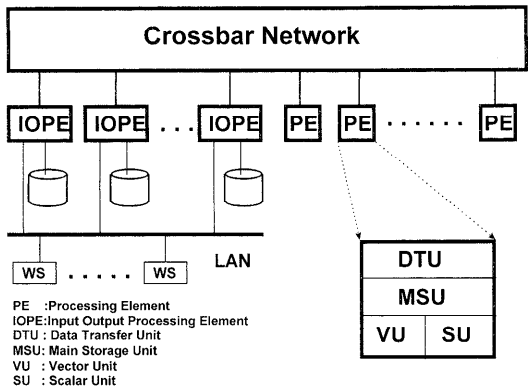


図-1 VPP700Eのシステム構成

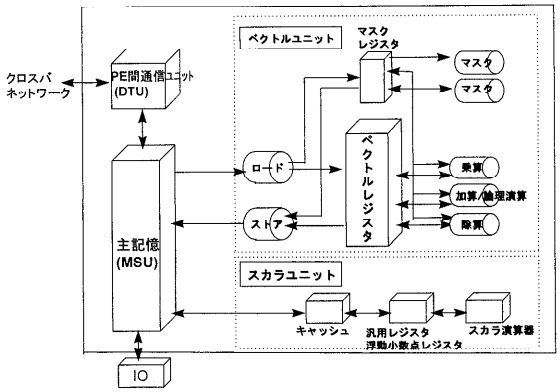


図-2 PEの構成

式の限界およびMPPの問題点を解決するため，分散メモリ型ベクトル並列アーキテクチャ(ベクトルパラレルアーキテクチャ)を創出した．これがベクトル並列処理方式(VPP)であり，その概念の中心は以下にある．

(1)要素プロセッサにはベクトルプロセッサを配置する．これにより，要素プロセッサの性能をMPPにおける要素プロセッサの性能の約20倍強にできる．また，ベクトル処理方式により，スーパコンピュータで頻発する広域アクセスを扱うプログラムに対しても高い実効性能を発揮できる．

(2)相互結合する要素プロセッサ台数をMPPの約1/20に抑えられるため，相互結合ネットワークをクロスバで構築できる．クロスバは，相手のプロセッサが通信中でないかぎり必ず通信ができ，任意のプロセッサ間の距離がすべて等しく，OSによる動的なPE割つけが容易にできる．各PEグループ内のPE間結合は，他グループとは独立なので他グループにおけるジョブの実行に妨害されない．

(3)従来のスーパコンピュータで走行していた逐次プログラムがPE上でそのまま走行できるため，並列化プログラム作成の前段階の立ち上がり促進を果たす．さらには，アルゴリズムおよびプログラム構造の延長線上で並列化を考えることが可能な場合が多い．また，複数ユーザが同時に使用するシステムの運用にも効率よく適応できる．

表-1 VX-E/VPP300E/VPP700Eシリーズ諸元

	VX-E	VPP300E	VPP700E
PE数	1～4	1～16	16～512
Peak性能	2.4～9.6GF	2.4～38.4GF	38.4～1228GF
主記憶容量	512MB～8GB	512MB～32GB	8GB～1024GB
主記憶throughput	19.6～78.4GB/s	19.6～313GB/s	313～10035GB/s
I/O throughput	240～480MB/s	480～1920MB/s	0.48～19.2GB/s
PE間転送throughtput	615MB/s×2/PE	615MB/s×2/PE	615MB/s×2/PE

表-2 PE主要諸元

ピーク性能		2.4GFLOPS/PE
ベクトルパイプライン数		7本
レジスタ	ベクトルレジスタ	(64bit×64)×256個
	汎用レジスタ	(32bit)×32個
	浮動小数点レジスタ	(64bit)×32個
キャッシュ		32KB×2個
主記憶	記憶素子	16Mb-SDRAM
	メモリ容量	512MB/1GB/2GB
	スループット	19.6GB/s

3. VPP700Eの概要とデータ転送系

3.1 VPP700Eの概要

図-1にVPP700Eのシステム構成を，**表-1**に各シリーズの諸元を示す．システムは，PEとそれを結合するクロスバネットワーク，およびIOPEに接続されるチャネルおよび各種I/O機器を接続するコントローラ・アダプタから構成される．

図-2にPEの構成を，**表-2**にPEの主要諸元を示す．PEは以下のユニットにより構成される．

(1)スカラユニット(SU)

LIW(64ビット)アーキテクチャを用いた1チッププロセッサを搭載しており，スカラ命令の実行，および各種割り込み処理などを行う．各命令語には1～3個のスカラ操作または1個のベク

トル操作を割りあて可能である．また，データ依存関係を保つ範囲内において，メモリアクセス命令，浮動小数点数演算命令，およびベクトル命令の非同期操作間の実行順序が変更できる．

(2)ベクトルユニット(VU)

VUはSUからベクトル命令を受け取り，パイプライン演算器で高速に実行する装置である．

命令実行パイプラインは，乗算，加算／論理，除算，マスク×2，ロード，ストアの7本あり，そのうち6本が並列実行可能(演算系は3本のうち2本が並列実行可能)である．それぞれの命令実行パイプラインは1システムクロックサイクルで8エレメント(エレメント：8バイトデータ)が並列実行可能である．

(3)主記憶装置(MSU)

各PEごとのプログラムやデータの格納を行う装置．VU/DTU/SU/IOが要求する大量のメモリアクセスを高速に処理する．

(4)データ転送ユニット(DTU)

クロスバネットワークを介して，PE－PE間のデータ転送，および同期処理を行う．

3.2 VPP700Eのデータ転送系

VPP700Eのデータ転送系としては，大きく分けて，ベクトルメモリアクセス系，PE間通信系，I/O系の3点が重要である．

(1)ベクトルメモリアクセス系

ベクトルレジスタと主記憶との間のメモリアクセスパイプラインは，ベクトルロードパイプラインが1本，ベクトルストアパイプラインが1本ある．これらのパイプラインは，8エレメントが並列実行可能であり，連続アクセスの場合は上記2本のパイプラインが同時に実行可能である．主記憶に対するスループットは，合計16エレメントが同時にアクセス可能な19.6GB/sを確保した．

(2)PE間通信機構とクロスバネットワーク

[DTU]

DTUは各PE内に搭載され，高い並列処理効率を実現するために，PE内部のスカラ・ベクトル演算とは独立にクロスバネットワークを介してPE間通信を実行する．DTUはデータ転送処理部とPE間同期処理部からなり，以下の特徴をもつ．

ーデータ転送処理部：

- データの送信／受信は並列に実行可能であり，それぞれ615MB/sの高速データ転送を行う．DTUとクロスバネットワークとのバス幅は送信／受信ごとに8バイトである．
- OSを介さずに，ユーザモードでの転送指示を可能とした．
- データ転送時のメモリアクセスとして，連続，ストライド，サブアレイ，間接アドレスの各パターンがあり，PE間転送処理効率を向上させている．
- 転送データのメモリアドレスおよび，転送受信先PEアドレスを変換するアドレス変換機構をもち，PE番号およびメモリアドレスの仮想化を実現した．

ーPE間同期処理部：

- 複数のPEを同期化するために，ハードウェア同期機構をもつ．各PEのプログラム進行状態を示す情報を指定のPEにブロードキャストする機構と，PE内でこの情報を受信して同期の完了を検出する機構からなる．
- 各PE内に同期をとるべきPEグループを示すマスクレジスタをもつことにより，プログラムを任意のPEからなるグループで走行することを可能としている．これにより，複数の並列プログラムを効率よく実行可能としている．

[クロスバネットワーク]

クロスバネットワークは，すべてのPEのDTUと接続され，PE間のデータ転送を実行する．VPP700Eでは，PEとは別フレームにクロスバを収容し，VX-E/VPP300Eでは各PE内部にクロスバを分散してもっている．

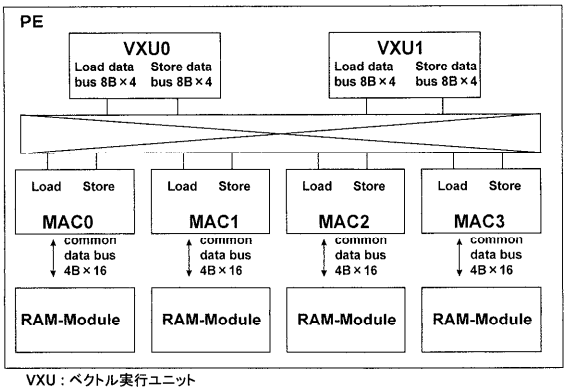
クロスバネットワークは任意のPE間で送信・受信それぞれ615MB/sの転送スループットを提供する．VPP700Eでは最大512PEを接続するクロスバネットワーク全体で，314GB/sの巨大なバンド幅をもつ高速ネットワークを実現した．

(3)I/O系インタフェースと複数IOPE

VPP700Eでは，PEとチャネル間のデータ供給能力として，PEあたり480MB/sの転送スループットをもつ．各PEには，8個のVMEチャネルと，2個のSBusチャネルが直接接続可能であり，各チャネルの配下には標準インタフェースをもつ各種コントローラ／アダプタが接続される．

表-3 VPP500とVPP700Eの主記憶

	VPP500	VPP700E
素　子	4Mb-SSRAM	16Mb-SDRAM
主記憶容量	1GB/PE	2GB/PE
cycle time	18ns	100ns
bank busy	2マシンサイクル	18マシンサイクル
interleave	32 bank (8byte)	512 bank (8byte)
throughput	12.8GB/s	19.6GB/s



VXU：ベクトル実行ユニット
MAC：メモリ制御ユニット

図-3 メモリ系構成

また，VPP700EではI/O処理を行うIOPEを複数台接続し，並列動作する“複数IOPE”機構を実現した．これにより，I/O性能も演算性能と同様にスケーラブルな向上が可能となった．

4. メモリ系データ供給技術

4.1 SDRAM採用の目的

スーパコンピュータのメモリは強力なメモリスループットが要求されるため，アクセス時間が高速なSSRAMが採用されてきた．しかし，科学技術計算分野におけるシミュレーションが大規模化するにともない，大容量メモリが必要となってきており，メモリ分散アーキテクチャのVPP500においても単一PE上での大容量メモリが要求されていた．一方，VPP700Eを開発するにあたり，論理素子としてCMOS-LSIを採用したことによる低コスト化，および実装構造のコンパクト化をさらに進めるため，そしてPE台数の増加にともない主記憶容量もスケーラブルに増加させるため，PEボード上にSU/VU/DTUのほか，主記憶も高密度実装することが必須であった．このような要件から，主記憶にSDRAM(16Mbit)を採用した．サイクルタイム(バンクビジー時間)がSSRAMと比較して大きいことに起因する性能劣化は，ある程度覚悟せざるを得ないが，大容量性および価格性能比の優位性を考慮するとSDRAMの採用が合理的である．

4.2 インプリメンテーション

(1)素　子

記憶素子としては，16Mbit(8bit × 2Mword)，動作周波数100MHz，アクセスタイム60ns，サイクルタイム100nsのSDRAMを1280個使用した．SDRAMを使用することにより，特別スペック品を開発することはなく，標準品を使用することが可能となった．

(2)実　装

SDRAMは，1枚のRAMモジュール基板に両面で20個搭載され，1PEあたり64枚のRAMモジュールを，CMOS-LSIの搭載されるPEボード基板上に高密度実装した．これによりシステムクロックに同期させた制御が容易となった．また，PE数の増加にともない，主記憶容量もスケーラブルに増加可能となった．

(3)インタリーブ

PEのシステムクロック(τ)でみたSDRAMのバンクビジー時間は18 τである．トータルインタリーブ数は，ベクトルロード8エレメントとベクトルストア8エレメントのトータル16エレメント(8バイト)の連続アクセスが続いた場合に，常に動作可能となる十分な幅が必要である．したがって，インタリーブ数は以下の式を満足する必要がある．

16エレメント×18 $\tau \leqq$インタリーブ数

VPP700Eでは，上記式を満足する値として512バンク(8バイト)のインタリーブ数を確保できるよう設計した．

表-3に，VPP500とVPP700Eの主記憶について示す．VPP500では，バンクビジー時間が2 τのため，32バンクのインタリーブ数で高速なメモリ系となっている．512バンクを実現するため，SDRAM固有の特徴を下記のように利用している．

- SDRAM内部は2バンクから構成されるが，この2バンクはインタリーブ方向にみえるようアドレス割つけを決定した．
- SDRAMには，連続するアドレスのデータ入出力(バースト転送モード)を10 ～ 20nsのサイクル時間で連続して実行できる機能がある．

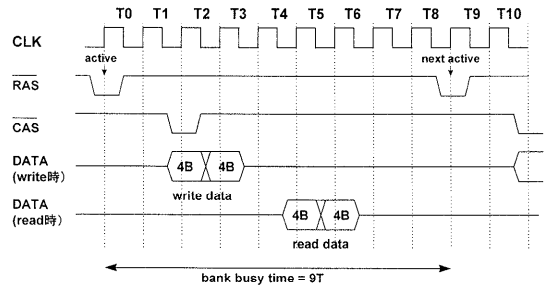


図-4 単体メモリアクセス基本タイムチャート

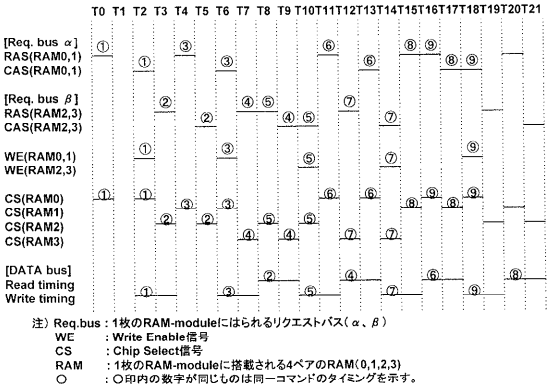


図-5 DATA bus 100 %稼働時のタイムチャート

VPP700Eでは，RAMモジュールに接続されるデータバスを4バイトとし，バースト転送＝2でアクセスすることにより，RAMの深さ方向を使用して8バイトデータを縮退して格納した．

これにより，SDRAM5個で2バンク(8バイト)，RAMモジュール1枚(SDRAM×20個)で8バンク，RAMモジュール64枚で，512バンクを実現した．

(4)バ　ス

図-3に，メモリ系構成を示す．VXU(ベクトル実行ユニット)はベクトルレジスタ，ベクトル演算器，ロード／ストアパイプラインから構成され，1ユニットあたり4エレメント分の並列度をもち，トータル2ユニットで構成される．MAC(メモリ制御ユニット)は，メモリリクエストとストアデータを受け取り，SDRAMとの実アクセス制御を行い，ロードデータを返す．MACは1ユニットあたり，RAMモジュール16枚分(128バンク分)と対応したSDRAMとのインタフェースをもつ．

VXUとMAC間は以下のバスで接続されている．

ロードデータバス：8B×8本

ストアデータバス：8B×8本

一方，MACとSDRAM間は以下のバスで接続されている．

コモンデータバス：4B×64本

最大16エレメントのメモリスループットを得るためには，64本のコモンデータバスが100 %稼動することが必要であり，そのため以下の制御を実施した．

- コモンデータバスを100 %使用するため，ロード／ストアのバス切り替えにかかる時間を設けず，休止サイクルなしで制御した．
- 連続アクセスのロード／ストアが100 %動作したときに，1本のコモンデータバスがロードとストアで交互に使用されるよう，アドレス割つけを決定した．

(5)タイミング制御

図-4に，SDRAM単体アクセスの基本タイムチャートを示す．TはSDRAM制御クロックサイクルを示す．また，tRCD(Ras Cas Delay-time)は2，CASレイテンシは3である．リクエストは，RAS(Row Address Strobe)/CAS(Colum Address Strobe)の2回のコマンドにより行われる．RASコマンドをトリガに，writeデータは4 τ後に，readデータは10 τ後にそれぞれ4 τ間データバスをビジーにする．18 τ後に次のRASコマンドを受け付け可能であり，バンクビジーは18 τとなる．

図-5に，データバスが100 %連続に動作した場合に，RAMモジュール1枚に着目したときのタイムチャートを示す．1本のコモンデータバスがロード／ストア交互に100 %稼動することがわかる．

4.3 実効性能

ベクトルロードアクセスにおいて，ストライド値を1から512まで変化させたときのVPP500とVPP700Eのスループットをそれぞれ**図-6**，**図-7**に示す．ストライド値が1，すなわち連続アクセスの場合は，VPP500，VPP700Eとも，ほぼ8エレメント/τのスループットが得られている．また，ストライド値1から512までの平均スループットはVPP500：3.3エレメント/τ，VPP700E：2.9エレメント/τであり，VPP700E

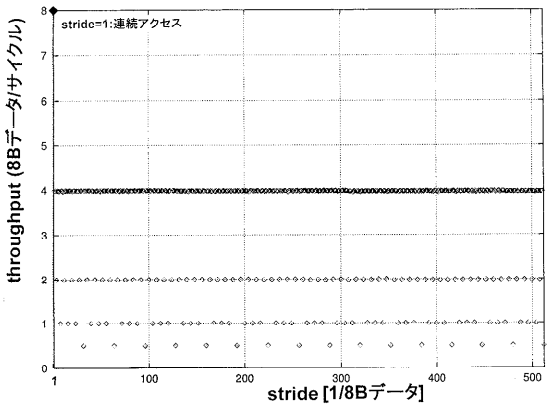


図-6 Load Access (stride) on the VPP500

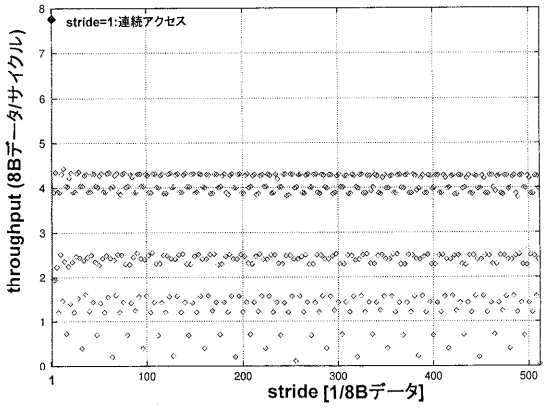


図-7 Load Access (stride) on the VPP700E

のストライドアクセス性能はVPP500と比較して，約10％程度の性能低下に抑えている．

一方，ランダムアクセス(インダイレクトアクセス)では，バンクビジー時間が18 τと長いため，バンク競合が発生する確率を回避するのが困難になり，実効性能は低下する．

連続アクセスであっても先行命令の終了付近のアドレスと後続命令の先頭アドレスの関係によっては，バンク競合が発生する確率が高くなる．また，ほかのアクセス(I/O，スカラ，DTU)と競合し，アクセスが乱れた場合は，バスが100％稼動状態に戻るまでに時間がかかる欠点もある．

以上のように，SDRAMを採用したVPP700Eのメモリアクセス性能は，ランダムアクセスにおいては実効性能低下があるが，頻度の高い連続アクセスではVPP500とほぼ同等，ストライドアクセスでも平均で約10％程度の低下に抑えている．

5. おわりに

本稿では，VPP700Eのデータ転送技術について紹介した．

主記憶を中心とした，ベクトルメモリアクセス系，PE間通信系，I/O系それぞれについて，その機能と転送スループットを示した．これらはシステム全体の性能を最大限引き出すために，コスト的にバランスよくインプリメントしていくことが重要である．

また，メモリ系データ供給技術では，主記憶にSDRAMを採用したことでインプリメントを工夫し，頻度の低いランダムアクセスについては性能が低下するものの，頻度の高い連続アクセス・ストライドアクセスについては高性能を維持していることを示した．PEボード上に2GBの主記憶が実装でき，低コストで，コンパクトな実装で，1TBまでスケーラブルに主記憶容量の増加が可能となった．

参考文献

1) Miura, K., Takamura, M., Sakamoto, Y. and Okada, S.: Overview of the Fujitsu VPP500 Super-computer, Digest of Papers, COMPCON Spring 93 (1993).
2) Utsumi, T., Ikeda, M. and Takamura, M.: Architecture of the VPP500 Parallel Supercomputer, Proceedings of Supercomputing '94, pp.478-487, Washington D.C. (Nov. 1994).
3) Takamura, M. and Utsumi, T.: Why Vector Parallel?, The Proceedings of the HPC Conference '94, Singapore, pp.394-398 (Sep. 1994).
4) 内田信男：ベクトルパラレルスーパーコンピュータVX/VPP300/VPP700シリーズのハードウェア，FUJITSU, Vol.47, No.6, pp.434-441 (Nov. 1996).

(平成9年4月18日受付)

坂井 賢一

1962年生．1984年東北大学工学部通信工学科卒業．1986年同大学院工学研究科情報工学専攻修士課程修了．同年富士通(株)入社．以来，スーパコンピュータVPシリーズ，VPPシリーズの開発に従事．e-mail:sakai@ayame.mfd.cs.fujitsu.co.jp