解 説 高速プロセッシングデータバス技術

2. SMPサーバのデータ・バス技術

Data Bus Technology for SMP Server System by Hiroo HAYASHI (Computer on Silicon Development Center, TOSHIBA Corp.), Tsutomu INABA (Information & Communication Systems Laboratory, TOSHIBA Corp.) and Tetsuo HATAKEYAMA (Computer on Silicon Development Center, TOSHIBA Corp.).

林 宏雄[1] 稲葉 勉[2] 畠山哲夫[1]

1 (株)東芝COS開発センタ
2 (株)東芝情報・通信システム技術研究所

1. はじめに

SMP(Symmetric Multi Processor)構成のサーバ・コンピュータで採用されているバス技術について説明を行う．SMPサーバのバスには，複数プロセッサの能力を十分に発揮するための高性能と，サーバ・マシンとしての高信頼性が要求される．ここでは高性能・高信頼性を確立するための技術に着目し，代表例も交え解説を行う．

2. SMPサーバで用いられるバス技術

高性能化のために，バス幅の拡大，バス処理の並列化，バス動作周波数の引き上げなどの手段がとられる．バス幅に関しては64bit，128bit，256bitと拡張されてきている．本章で紹介する技術は，**表-1**に示すように，バス処理の並列化，バス動作周波数の引き上げを目的とするものである．また信頼性の確立のための技術も紹介する．

2.1 スプリット・バス

従来のバスは，メモリ・アクセス命令がバスに出力されてからそれに対応するデータがバスに出力されるまでの間，そのバス・トランザクションに占有されていた．プロセッサおよびバスの周波数が低く，メモリ・アクセス時間が数バス・クロック・サイクルであった時は，これは問題とはならなかった．しかしプロセッサの動作周波数とともにバスの動作周波数も上がり，メモリ・アクセス時間が相対的に大きくなってきている．このため1バス・トランザクションあたりのバスの占有時間を短くすることが重要となった．SMPサーバでは，あるプロセッサがバスを使用している間，ほかのプロセッサがバスを使用できないことだけでも問題となるが，さらに個々のプロセッサの処理の多重度が上がり，1つのプロセッサが同時に複数のメモリ・アクセスを発行できるようになったことによりバス・トランザクションのバス占有時間を短くすることがますます重要となる．

そこで，バス命令を発行した時にいったんバスを解放し，データが出力できるようになった時点で再度バスを使用し，その間はほかのバス・トランザクションがバスを使用できるようにしたものをスプリット・バス☆と呼ぶ．**図-1**にIBMの6XXバスの動作タイミングを示す．サイクル4で発行されたバス命令に対して，サイクル15でデータ・バス要求(DBG_)が出され，サイクル16から19でデータ転送が行われている．この間別のバス・トランザクションの処理が行われる．

またバス命令が発行された順序で，対応するデータ転送が行われなければならないものをin-orderバス，この制約がないものをout-of-orderバスと呼ぶ．後者はI/Oアクセスなど，応答の遅いアクセスがあっても，ほかのトランザクションが影響を受けないという利点がある反面，制御が複雑となる．先に示した図-1の例はout-of-orderバスの例でもあり，後に発行されたバス・トランザクションBに対するデータ応答が，先に発行

表-1 高速化手法

	処理の並列化	動作周波数の引き上げ
スプリット・バス	○	
パイプライン・バス	○	
クロスバー・スイッチ	○	○
階層構成	○	○
GTLドライバ		○

☆ これに対して従来のバスをスイッチ・バスなどと呼ぶことがある．

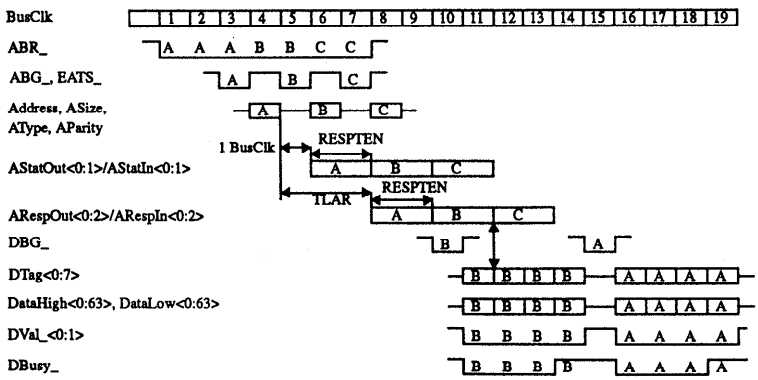


図-1 6XXバスの動作シーケンス(スプリット・バス，パイプライン・バス)[4]

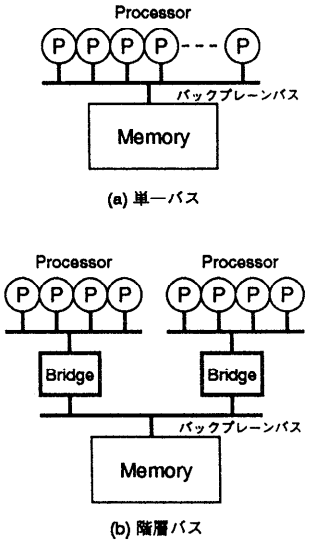


図-2 単一バスと階層バス

されたバス・トランザクションAに対するデータ応答よりも先に行われている．

2.2 パイプライン・バス

パイプライン・バスはスプリット・バスと同様に，1バス・トランザクションあたりのバスの占有時間を短くするための手法であり，スプリット・バス手法とともに用いられることが多い．

SMPサーバでの個々のバス・トランザクションは以下に示すいくつかの段階を踏んで行われる．(スプリット・バスを前提とする．)

(1)アドレス・バス・アービトレーション
(2)トランザクション(アドレス，バス命令の種類，そのほかの属性)発行
(3)スヌープ応答
(4)データ・バス・アービトレーション
(5)データ発行

スプリット・バスの説明で示した図-1は，パイプライン・バスの例でもある．サイクル8，9では，バス・トランザクションCのバス命令発行，バス・トランザクションBのStatus応答(パリティ・エラー通知など)，バス・トランザクションAのスヌープ応答(キャッシュ・コヒーレンシ保持のための応答)が行われている．このように並列に処理を行うことによって2サイクルに1つのバス・トランザクションの発行が可能となり，バスの実効バンド幅が大きく改善されている．

2.3 クロスバー・スイッチ

データ・バスの接続形態(構成)として，広く用いられている共有バス構成のほかに，従来メインフレームやスーパー・コンピュータなどで採用されていたクロスバー・スイッチを用いた構成がある．

共有バス構成は，一組の信号線に複数のデバイスを接続すればよい．このため低コストで実現でき，データ転送のレイテンシが小さいという利点がある．しかし共有バス構成では，電気的問題のために動作周波数を上げることが難しく，また同時に一組のデータ転送しか行うことができない．

これに対してクロスバー・スイッチを用いた場合，信号が1対1接続となるため動作周波数を高くすることが容易であり，同時に複数のデータ転送を行うことも可能となる．一方クロスバー・デバイスとその実装領域を用意しなければならず，コスト高となるが，多ピンのパッケージの登場などによってコストを抑えられるようになってきている．またクロスバー・デバイスを経由するために，データ転送のレイテンシが共有バス構成に比べ数サイクル長くなる，アドレス・バスのスヌープ処理に向かない，などの欠点をもつ．

2.4 階層構成(バックプレーン・バスの負荷制御)

プロセッサ数の増加にともない，バス・トラフィックが増大する．1つのバスで処理できるバス・トラフィックには限界があり，このためプロセッサ数が増えても，それに応じたシステム性能の向上が得られなくなる．

これに対応するために図-2に示すように，1つのバスにすべてのバス・デバイスを接続するかわりに，バスを階層的に構成する手法がある．

これにより階層ごとにバス・トラフィックを1つにまとめてしまうことができる．すなわち，

各階層バス内で処理できるバス・トランザクションはバックプレーン・バスには発行されないため，各バスでのバス・トランザクション数が減る．また階層バスとバックプレーン・バスをつなぐバス・ブリッジにキャッシュ・メモリをもたせることもある．

またバスの階層化には，バス動作周波数を上げることを容易にする効果もある．単一バス構成では，同一の信号線に多数のデバイスが接続されるため，電気的問題により高速化が困難である．これに対しバスを階層化すると，各信号線に接続されるデバイス数が減り，バス動作周波数を上げることが容易となる．

一方，バスを階層化すると，バス・ブリッジを経由するためレイテンシが増加する，キャッシュ・メモリの一貫性保持のための制御が複雑になるという欠点もある．

2.5 GTL ドライバ

これまで高速な信号のインタフェースには，ECL(Emitter Coupled Logic)がよく用いられてきたが，ECL はバイポーラ回路であり，消費電力・発熱が大きい，集積度が低いなどといった欠点があった．

ECL にかわる高速の小振幅インタフェースが開発されている．GTL(Gunning Transceiver Logic)，CTT(Center Tap Terminated)，HSTL(High Speed Transmission Logic)などである．このうち GTL ドライバおよび GTL ドライバを基に変更を加えたものが SMP システムのバスに使用されている．

GTL ドライバは，Xerox 社で開発され，開発者である Bill Gunning 氏の名をとって GTL(Gunning Transceiver Logic)と名づけられた．GTL は，CMOS のオープンドレイン構造をとるので低消費電力化，高集積化が容易である．また，高速な信号のインタフェースの要件(小振幅であること，ドライバ自身が高速であること，進行波伝送であること)を満たす．またオープンドレイン構造であるため信号波形が乱れやすいが，信号の立ち上がり時間の制御を行うことにより，高速化を実現している．

2.6 信頼性の向上

SMP サーバは信頼性を要求されるシステムで使用されることが多いため，バスの信頼性も重要となる．

表-2 各システムの諸元

システム名	Power Challenge	RS/6000 J30	UltraEnterprise
バス	POWERpath-2	6XX バス	Gigaplane
ベンダ	SGI	IBM	Sun
実効バンド幅	1.2GB/sec	800MB/sec	2.67GB/sec
バス動作周波数	47.6 MHz	75 MHz	83.5 MHz
データ・バス幅	256 bit	64 bit	256 bit
最大 CPU 数	36	8	30
最大 transaction 数	8	256	112
スプリット・バス	○	○	○
パイプライン・バス	×	○	○
in-order/ out-of-order	out-of-order	out-of-order	out-of-order
信号 level	NTL	不明	open drain LVCMOS
データ保護	parity	parity	ECC
活線挿抜	不明	対応	対応

SMP サーバでは，メモリのデータ保護のために 2bit 誤り検出，1bit 誤り訂正が可能な ECC 保護が広く採用されている．このうちいくつかのシステムではデータ・バス上のデータ保護にも ECC を採用している．

またシステムを停止することなく，システム構成の変更，故障したバス・デバイスの交換を可能とするため，システムの動作中にバス・デバイスの挿入・抜去を行う活線挿抜に対応したものもある．

さらに，バスを複数もち，通常はこれらを同時に使用することで高いバス・バンド幅を確保し，そのうちのいくつかのバスが故障した時に，残ったバスで縮退運転を行うことにより，処理を継続することが可能なものもある．

3. SMP サーバの例

本章では，特徴的な次の 3 つの SMP サーバのバスを例にあげ，そこで用いられている技術について紹介する．

- SGI Power Challenge(POWERpath-2 Coherent Interconnect)[1), 2)]
- IBM RS/6000 J30(6XX bus)[3), 4)]
- Sun UltraEnterprise x000(Gigaplane)[5)]

表-2 にここで例にあげたシステムのバスの諸元を示す．なおこれらは異なる時期に発表されたものであり，これらを比較して優劣をつけるのは意味をもたない．

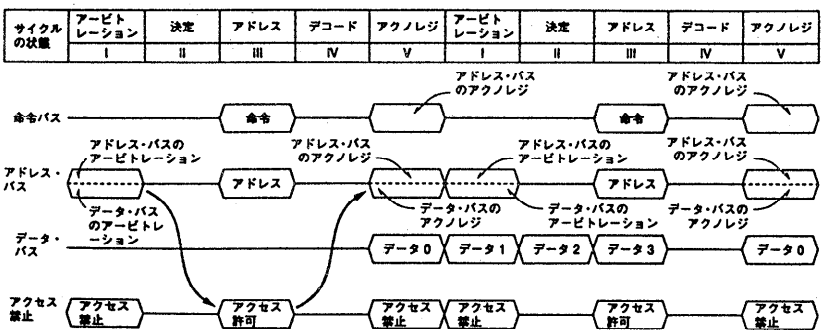


図-3 Powerpath2 のデータ転送シーケンス[2)]

3.1 SGI Power Challenge(POWERpath-2 Coherent Interconnect)

POWERpath-2 は，40 ビット幅のアドレス・バスと 256 ビット幅のデータ・バスから構成される SMP システム用のバスである．スプリット・バス方式により，動作周波数 47.6MHz で実効データ転送速度 1.2G バイト／秒を実現する．

POWERpath-2 でのバス・プロトコル処理は，RISC の思想を採り入れている．RISC は，命令数を減らし，命令語長を統一，各命令を同じサイクル数で実行可能とすることによって，制御の容易化や処理速度の高速化を狙う．POWERpath-2 (**図-3**)ではバス・プロトコルでのトランザクション処理において，スプリット・バス方式によるアドレス／データ・バスでのトランザクションの種類を 13 に抑え，各バスでのトランザクション処理動作を 5 サイクル固定で実行できるようにしている．データ・バスでは 1 トランザクションごとにキャッシュ・ライン・サイズ 128 バイト(256 ビット×4)を転送，47.6MHz で動作することにより，毎秒 950 万トランザクションを処理，1.2G バイト／秒の転送速度を実現する．スプリット・バス方式によって，アドレス・バスとデータ・バスを各々異なるプロセッサの処理に割り当てることができる．また，応答が必要なトランザクションを記憶することにより，out-of-order バスとしての処理を実現，バスの使用効率を上げている．POWERpath-2 に接続される各ボード(ノード)は最大 8 つのトラザクションを記憶する．応答のためのトランザクションは，バス・アービトレーション処理により優先的に処理されるように制御される．POWERpath-2 では分散型アービトレーション方式を採用，各ボードでアービトレーション処理を行うことによって，他デバイスからの通知を待たずに次の動作を開始できる．

POWERpath-2 のバス・プロトコルは，バス・スヌープによるライト・インバリデート方式のライトバック・キャッシュ処理をサポートする．各キャッシュ・ラインは Invalid，Shared，Exclusive，Modified の 4 状態で管理される．バス制御回路はプロセッサのキャッシュ・メモリの状態を保持するタグ・メモリを内蔵し，高速なスヌープ応答を実現している．キャッシュ処理による転送動作として，メモリからキャッシュ，キャッシュからメモリ，キャッシュからキャッシュへの 1 対 1 転送のほかに，キャッシュからキャッシュとメモリ，メモリから複数のキャッシュへの転送を行うことによって，バスでのトランザクション数を減らし，リード・アクセスのレイテンシを小さくする．

POWERpath-2 では信号入出力インタフェースとして，CMOS 小振幅インタフェースである GTL を変更した NTL(nMOS Transceiver Logic)を使用し，バックプレーン長 12.8 インチで 15 枚のボード(ノード)が接続できる．NTL ドライバは，GTL よりさらに大きな 48mA の駆動電流を用いる．15 枚のボードを実装したバックプレーンにおいて，動作周波数 47.6MHz での信号入出力を可能にしている．

POWERpath-2 に接続された各ボード内(キャッシュ／メモリ)では，ECC によって格納しているデータを保護している．各ボードからバックプレーン・バスに入出力されるアドレス／データ信号の値は，パリティ・ビット信号を付加することによって保護されている．

3.2 IBM RS/6000 J30(6XX bus)

J30 は PowerPC 601，604 を用いた SMP システムであり，そのバスでは，電気的に無理をせずに，高い性能を得る工夫がされている．

アドレス・バスは，バス接続を採用することにより，スヌープ処理が可能となりキャッシュ・メモリの一貫性維持の処理を容易にしている．集中

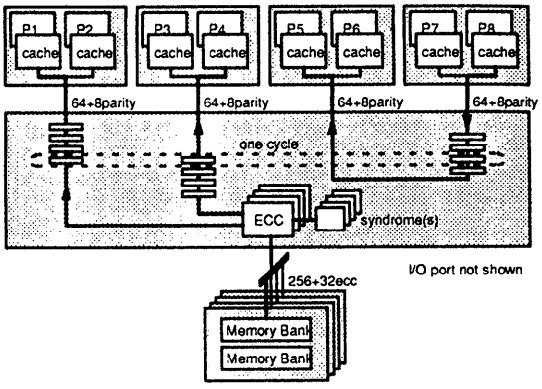


図-4 RS/6000 J30 のデータ・クロスバー・スイッチ[3]

型アービトレーション方式を採用し，バスのアービトレーション信号，バス応答信号などは，いったんアービタ回路に集められ，結果が各ボードに返される．これらの信号は1対1接続であり，75MHz で動作する．これに対し，ほかの共有バス接続されている信号は2サイクルごとに変化し，実質37.5MHz で動作する．パイプライン・バスを導入することにより，2バス・クロックごとに1つのバス・トランザクションの発行が可能となっている(図-1)．

データ・バスにはクロスバー・スイッチが採用されている．クロスバー・スイッチはプロセッサ・ポート4組，I/O ポート1組，メモリ・ポート1組の合計6ポートからなる．各プロセッサ・ポートには75MHz 動作が可能な範囲である2組のプロセッサがバス接続されることにより，クロスバー・スイッチのポート数を抑えながらも，最大8プロセッサのシステム構成が可能となる．プロセッサ・ポートは64bit データ＋8bit パリティ，メモリ・ポートは256bit データ＋32bit ECC という構成となっている☆☆．図-4 に，2つのメモリ読み出し，1つのキャッシュ・メモリ間データ転送が並行して行われている様子を示す．

3.3 Sun UltraEnterprise x000 (Gigaplane)

共有バス接続で83.5MHz と高い動作周波数を実現している．

最大16ボードを，センタプレーン・ボードの表裏両面に8ボード・スロットずつ装着することにより，実際のバス長を16.0インチに抑えている．

専用に設計したオープン・ドレイン型低電圧振幅の CMOS のドライバを用いている．センタプレーンのインピーダンスを27オーム程度にすることにより通常の50オーム程度の場合と比べて46％バス負荷遅延を低減している．並列終端と直列終端を併用することにより，信号波形の乱れを抑えた．これにより空きサイクルなしに，連続して異なるバス・ドライバがバスを駆動することを可能としている☆3．アドレス・バス，データ・バスともに1バス・トランザクションごとに2サイクル使用するため，バンド幅が33％向上している．

バスのアービトレーションは分散型アービトレーション方式を採用している．アドレス・バスのアービトレーションにはバス・パーキング手法と“fast-arb”と呼ばれる手法のいずれかを選択して使用する．バス・パーキング手法は従来から広く用いられている方法で，直前に権利をもったバス・デバイスがアービトレーションを行うことなく権利をもつ方法である．これに対して“fast-arb”手法は，どのボードもアービトレーションの結果を待たずにバス・トランザクションの発行とアービトレーションの開始を並行して行い，もし衝突(collision)が検出された時にはアービトレーションの結果に従い再発行を行う方法である．この手法はオープン・ドレイン型のドライバを採用しているために可能となっている．データ・バスでは，先行してバス要求を出すことができるため，これらの方法は使用しない．

図-5 にバス信号のタイミングの例を示す．サイクル0でバス命令，アドレスが出力され，サイクル1でアドレス・バス・トランザクションとそのデータ応答を対応づけるための ID を出力する．サイクル5，6でキャッシュ・メモリ状態のスヌープの結果が出力される．バス制御回路はプロセッサのキャッシュ・メモリの状態(Modified, Owned, Exclusive, Shared, Invalidate の5状態)を Owned, Shared, Invalid の3状態で保持する DTag(Duplicate Tag)をもち，プロセッサに問い合わせることなく，このタイミ

☆☆ 6XX bus は，データ・バスが128bit 幅のバス接続の形態も定義している．

☆3 通常のバスでは，バスを駆動するデバイスが変わる時には1サイクル程度の空きサイクルが必要である．

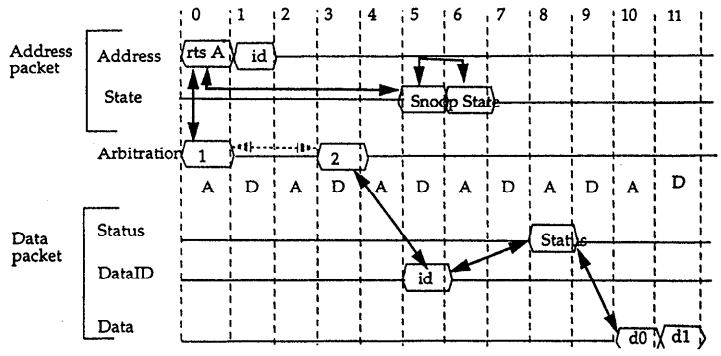


図-5 Gigaplane バスの動作シーケンス[5]

ングでスヌープの結果を出力可能としている．

一方メモリ制御回路は，スヌープ応答の結果を待つことなくDRAMアクセスを開始し，サイクル3☆4でデータ・バス要求を行う．サイクル5でIDを出力し，もしスヌープ応答の結果，メモリから読み出したデータが不要であることが判明した場合は，サイクル8のStatus信号でデータ転送をキャンセルをする．

システムの動作中にボードの抜き差しを行う活線挿抜をサポートする．たとえばボードの挿入時にはコネクタの特別な長いピンの接触により，ボードの挿入を検出し，バス制御回路はいったん処理を停止し，特別な短いピンの接触により処理を再開する．ボードの抜去時にもこれと同様の処理を行う．

4. おわりに

SMPサーバで用いられるバス技術をいくつか紹介した．今後ますますプロセッサの動作周波数が向上することが予想される．一方従来の単一バス接続では動作周波数100MHz程度が限界であり，複数の高性能プロセッサのデータ要求を満たすことは困難である．今後クロスバー・スイッチ接続や階層構成などの技術が広く用いられるようになると考えられる．

参考文献

1) Silicon Graphics: POWER CHALLENGE™ Technical Report, http://www.sgi.com/Products/software/PDF/pwr-chlg/index.html

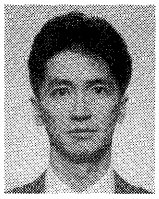
2) Galles, M. and Williams, E.：実効データ転送1.2Gバイト／秒のPOWERpath-2：日経エレクトロニクス，pp.106-114(Oct. 1993)．

3) Nicholson, J. O.: The RISC System/6000 SMP System: Digest of COMPCON '95, pp.102-109 (1995)．

4) Allen, M. S. and Lwechuk, W. Kurt: A Pipelined, Weakly Ordered Bus for Multiprocessing Systems: Digest of COMPCON '95, pp.292-298(1995)．

5) Singhal, A. et al.: Gigaplane™: A High Per-formance Bus for Large SMPs: Hot Interconnect '96, pp.41-52(1996)．

(平成9年4月17日受付)



林 宏雄

1962年生．1988年京都大学大学院工学研究科電気工学専攻修士課程修了．同年(株)東芝入社．以来，並列・分散処理コンピュータ・アーキテクチャの研究開発に従事．現在，(株)東芝COS開発センター開発第三部アーキテクチャ第一担当主務．IEEE会員．

稲葉 勉

1964年生．1988年電気通信大学電子情報学科卒業．同年(株)東芝入社．以来，OAシステムコンピュータの中央処理装置の開発を経て，サーバ・コンピュータの高速データ伝送に関する技術開発に従事．現在，(株)東芝情報・通信システム技術研究所開発第一担当．

畠山 哲夫

1966年生．1989年横浜国立大学工学部電子情報工学科卒業．1991年同大学院工学研究科修士課程電子情報工学専攻修了．同年(株)東芝入社．以来，並列・分散処理コンピュータ・アーキテクチャの研究開発に従事．現在，(株)東芝COS開発センター開発第三部アーキテクチャ第一担当．

☆4 これは最も早い場合である．