ＳＸ－４　マルチノードシステムのハードウェア構成

幅田　伸一

１．まえがき

ＳＸ－４　マルチノードシステムは、ＳＸ－４シリーズの最上位に位置するシステムです。ＳＸ－４シリーズの最先端アーキテクチャ、国際標準の超高速インタフェース、最先端ＣＭＯＳ　ＬＳＩ技術、高密度実装技術を採用することにより、プロセッサ当たり２ギガＦＬＯＰＳ、システムの最大演算性能１ＴＦＬＯＰＳを実現しております。ＳＸ－４シリーズは、超高速演算性能を実現するために、

（１）ベクトル型スーパーコンピュータとして、世界で初めて、共有メモリ型システム（シングルノードシステム）を最大１６台接続するクラスタ型システム構成
（２）共有メモリ型のベクトル型スーパーコンピュータとして、最大３２台のマルチプロセッサ構成
（３）種々の高速化技術を採用した高速スカラユニット
（４）３２本のベクトル演算パイプラインの同時動作が可能な強力なベクトルユニット
（５）ＡＮＳＩ準拠の超高速ＨＩＰＰＩチャネル
（６）高速入出力処理装置
（７）高性能主記憶装置
（８）大容量高速拡張記憶装置
（９）超高速ノード間接続装置

などを備えております。

以下に、これらのハードウェアについて紹介いたします。

２．ハードウェア構成の要素

ＳＸ－４　マルチノードシステムの構成および構成要素のシステム内での位置づけを図１に示します。

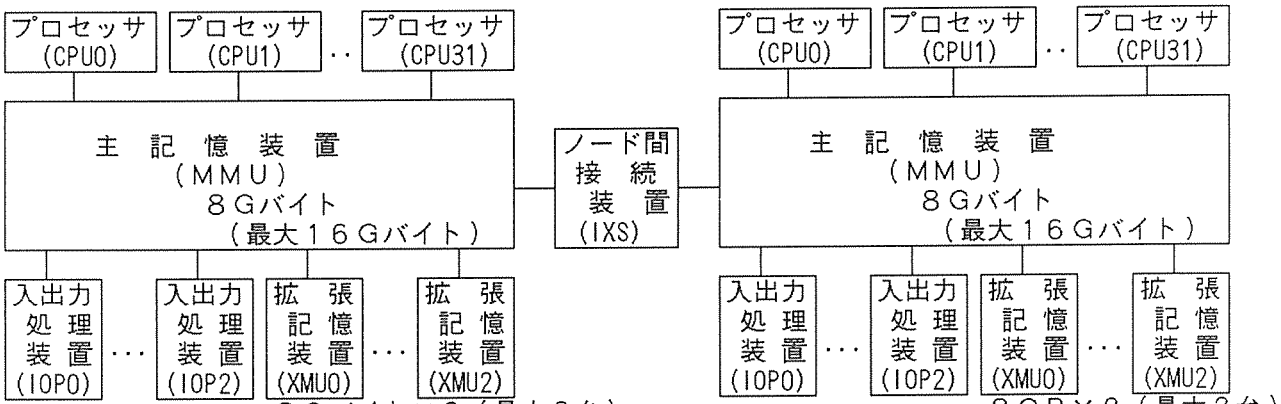


図１　ＳＸ－４／６４Ｍ２の構成

ＳＸ－４　マルチノードシステムは、ＳＸ－４シリーズの最上位システムであり、ＳＸ－４　シングルノードシステムを最大１６台接続するクラスタ型システムを構成します。ＳＸ－４　シングルノードシステムは、最大３２台のプロセッサ（ＣＰＵ：Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）を接続できる共有メモリ型マルチプロセッサシステムであり、プロセッサ、入出力処理装置、主記憶装置、拡張記憶装置から構成されます。

プロセッサは機能的にスカラユニット（ＳＵ：　Ｓｃａｌａｒ　Ｕｎｉｔ）とベクトルユニット（ＶＵ：　Ｖｅｃｔｏｒ　Ｕｎｉｔ）とに分けられます。スカラユニットはプロセッサ上で実行される命令の解読と実行制御を行うとともに、スカラ演算パイプラインによりスカラ演算を高速実行します。ベクトルユニットは８セットのベクトル演算パイプラインと１４４Ｋバイトのベクトルレジスタを備え、ベクトル演算を高速実行します。

主記憶装置（ＭＭＵ：　Ｍａｉｎ　Ｍｅｍｏｒｙ　Ｕｎｉｔ）はプログラムおよびそれらが使用するデータ類を格納し、ノードあたり最大１６Ｇバイトの記憶容量があります。

入出力処理装置（ＩＯＰ：　Ｉｎｐｕｔ　Ｏｕｔｐｕｔ　Ｐｒｏｃｅｓｓｏｒ）は主記憶装置と周辺処理装置の間に位置し、プロセッサの指示により、プロセッサと独立して入出力処理を高速実行し、ノードあたり最大３台まで接続できます。ＡＮＳＩ準拠の１００Ｍバイト／秒の超高速チャネルであるＨＩＰＰＩ（Ｈｉｇｈ　Ｐｅｒｆｏｒｍａｎｃｅ　Ｐａｒａｌｌｅｌ　Ｉｎｔｅｒｆａｃｅ）を最大４８本接続でき、超高速スター型のＨＩＰＰＩネットワーク、ＵｌｔｒａＮｅｔ、超高速アレイディスクを接続できます。この他、１６ビットＳＣＳＩ－２　ｄｉｆｆｅｒｅｎｔｉａｌ、ＦＤＤＩなどをノードあたり最大１６チャネル接続できます。

拡張記憶装置（ＸＭＵ：　Ｅｘｔｅｎｄｅｄ　Ｍｅｍｏｒｙ　Ｕｎｉｔ）は磁気ディスク装置と主記憶装置の間に存在するアクセスギャップ（アクセス時間の差）を埋めるもので、ノードあたり最大２４Ｇバイトの記憶容量があります。

ノード間接続装置（ＩＸＳ：　Ｉｎｔｅｒｎｏｄｅ　Ｃｒｏｓｓｂａｒ　Ｓｗｉｔｃｈ）はＳＸ－４シングルノードシステムを最大１６台接続することができるパケット交換型スイッチネットワーク装置です。シングルノードシステム側に実装されるノード間接続アダプタ（ＩＸＳアダプタ）が提供するデータムーバ機能により、他のノードの主記憶装置上のデータを他のノードのプロセッサを使用せずに、アクセスすることができます。

３．プロセッサ（ＣＰＵ）

ＳＸ－４のプロセッサは、ＳＸ－２およびＳＸ－３で培われてきた超高速演算プロセッサ技術をさらに発展させ、最新のアーキテクチャおよびテクノロジを採用することにより、高速なプロセッサを実現しております。

テクノロジとして、およそ４００万トランジスタを収容可能な超高集積度のＣＭＯＳ　ＬＳＩと空冷高密度実装技術を採用し、プロセッサを１枚のカードに実装した１カードプロセッサを実現しました。この最新テクノロジと高度なパイプライン処理により、２ＧＦＬＯＰＳの演算性能を備えた１カードプロセッサを実現しました。

ＳＸ－３で培われてきた共有メモリ型マルチプロセッサ方式を発展させ、ＳＸ－４　シングルノードシステムは最大３２台のマルチプロセッサシステムを構築可能とすることにより、最大６４ＧＦＬＯＰＳの性能を実現しています。

さらに、共有メモリ型マルチプロセッサ方式の限界を乗り越えるべく、ベクトル型スーパーコンピュータとしては初めて、クラスタ型システム方式を採用し、最大１６台のシングルノードシステムを接続するマルチノードシステムにより、最大１ＴＦＬＯＰＳの性能を実現しました。

プロセッサの構成を図２に示します。

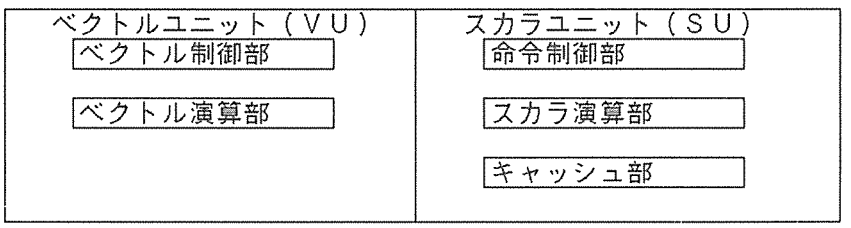


図２　プロセッサ（ＣＰＵ）の構成

プロセッサは大きく分けて、スカラユニットとベクトルユニットの２つのユニットによって構成されます。

スカラユニットは、プロセッサ上で実行される命令の解読と実行制御を行うとともに、スカラ演算パイプラインによりスカラ演算を高速実行します。

ベクトルユニットは８セットのベクトル演算パイプラインと１４４Ｋバイトのベクトルレジスタを備え、ベクトル演算を高速実行します。

３．１　プロセッサ（ＣＰＵ）のアーキテクチャ

ＳＸ－４のプロセッサは、ＳＸ－３のアーキテクチャを基に、パイプライン処理およびマルチプロセッサ機能の改善、マルチノードシステムにおける他ノードとのデータ転送／通信機能の強化を行っております。

スカラでは、ＳＸ－３のＲＩＳＣアーキテクチャ、スカラ演算のパイプライン処理、ハードウェアによる命令の実行順序の並べ換え制御を改善するとともに、２セットの命令処理パイプラインを備えるスーパースカラアーキテクチャを採用し、高速スカラ処理を実現しています。

ベクトルでは、演算パイプラインの多重並列化により、最大８個の要素に対する浮動小数点加算と最大８個の要素に対する浮動小数点乗算を同時に実行可能とし、さらに、ＧＦＬＯＰＳあたりのベクトルレジスタと主記憶装置間のデータ転送性能を改善し、超高速ベクトル処理を実現しています。

マルチプロセッサをサポートする機能としては、プロセッサ間の高速な通信を可能とする通信レジスタの容量を拡大し、マルチタスク方式による高速並列処理を実現しています。シングルノードシステム内の並列処理に使用する通信レジスタは主記憶装置に実装されており、マルチノードシステムにおけるノード間にまたがる並列処理に使用する通信レジスタはノード間接続装置に実装されています。

ＳＸ－４　マルチノードシステムでは、ＳＸ－３に対して以下のような命令を追加して、ノード間にまたがる並列処理を実現しています。

・ノード間データ転送命令

・グローバル通信レジスタアクセス命令

３．２　スカラユニット

スカラユニット（ＳＵ）は、命令制御部、スカラ演算部、キャッシュ部から構成されています。

（１）　命令制御部

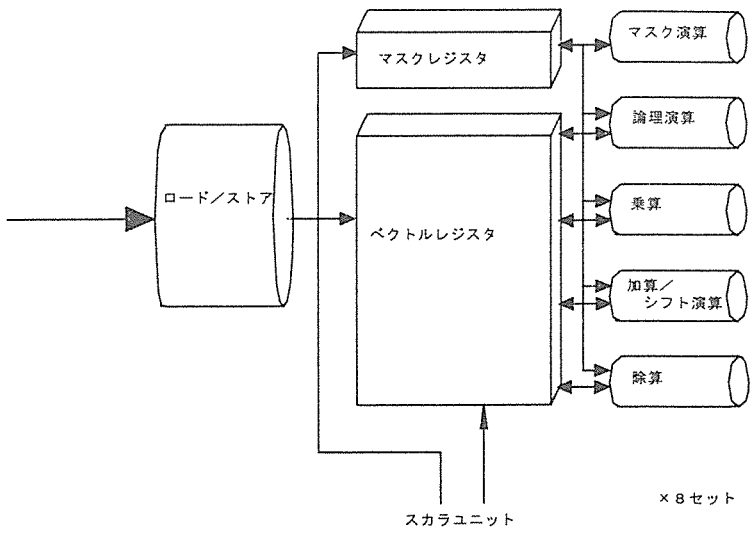
命令制御部は、高速大容量の命令バッファを持ち、プロセッサが実行すべき命令を解読し、スカラユニットおよびベクトルユニットの各機能を有効に使用するように、命令の実行順序を制御する機能を持っています。

１）　命令制御回路

命令制御回路はパイプライン化され、命令を解読し、使用するオペランドの準備状態、スカラレジスタへの書き込みパスの競合、ベクトル命令においては、使用するベクトルレジスタ、演算パイプライン、チェイニングのタイミングなどを管理し、スカラ性能およびベクトル性能を最大限に引き出す命令実行順序の制御を行います。

さらに、２セットの命令処理パイプラインからなるスーパースカラアーキテクチャの採用により、1マシンサイクルに２命令の解読、最大４命令の実行起動を可能としています。

図３にスカラユニットの構成を示します。



この他に、命令制御部は以下のようなプログラムの実行状態を計数するカウンタを備え、プログラムの実行を高速化するために必要な支援情報を提供できるようにしています。

・タイマレジスタ
・命令実行数カウンタ
・ベクトル命令実行数カウンタ
・ベクトル要素数カウンタ
・ベクトル命令実行時間計測カウンタ
・浮動小数点演算カウンタ

２） 命令バッファ

命令バッファは１２８バイト×６４ブロック（８Ｋバイト）の容量を持ち、４バイト命令を２０４８個分格納することができます。

繰り返しループ中の命令などは、命令バッファに格納されているる確率が高く、主記憶装置から命令を読み出すことなく、命令バッファから命令を取り出して実行することができます。

命令バッファの中に実行すべき命令がない場合には、その命令を含む１２８バイトのブロックを命令キャッシュまたは主記憶装置からロードしますが、その場合、必要になっている命令語からロードを始め、その命令がロードされた時点でただちに解読処理を開始することにより、高速な命令実行処理を可能としています。

３） 高速分岐機構

分岐命令は、分岐する側としない側のいずれかにかたよる傾向があります。命令バッファは、分岐命令が前回どちら側の命令を実行したかを記憶する分岐履歴バッファを備え、この内容に基づいて前回と同じ側の命令を先行して取り出し、解読を進めることにより、分岐命令の高速化を図っています。

４） 大容量スカラレジスタ

スカラユニットは１２８個のスカラレジスタを備えています。このレジスタは汎用のレジスタであり、アドレス計算用のベース／インデックスレジスタ、固定／浮動小数点演算用の演算レジスタとして利用できます。

１２８個の大容量レジスタを備えることによって、レジスタ競合による演算パイプラインの乱れを少なくしたり、変数や中間結果のセーブ／リストアする頻度を下げ、メモリアクセス回数を減少させています。さらに、大容量のスカラレジスタにより、コンパイラのレジスタ割り当ての自由度が増し、ハードウェアを効率よく使用するオブジェクトコードの生成が可能となり、高速なスカラ演算を実現しています。

（２） スカラ演算部

スカラ演算部は、８バイトデータ演算を基本とし、２つの固定小数点算術／論理演算器、浮動小数点加算器、乗算器、除算器の５つの演算器から構成されています。これらの演算器は、ベクトル演算器と同様にパイプライン化されており、スカラ演算命令を連続、並列に実行することができます。

（３） キャッシュ部

プロセッサの性能向上には、演算器の性能向上とともに、メモリアクセス性能の向上が重要な要素となります。キャッシュ部には、主記憶装置上の使用頻度が高い命令の写しを保持する６４Ｋバイトの容量をもつ命令キャッシュと、データの写しを保持する６４Ｋバイトの容量をもつオペランドキャッシュがあり、命令制御部からの命令語あるいはオペランドのメモリアクセス要求に対し、高速に応答します。

３．３ ベクトルユニット（ＶＵ）

ベクトルユニットは、ベクトル演算部とベクトル制御部から構成されます。図４にベクトルユニットの構成を示します。

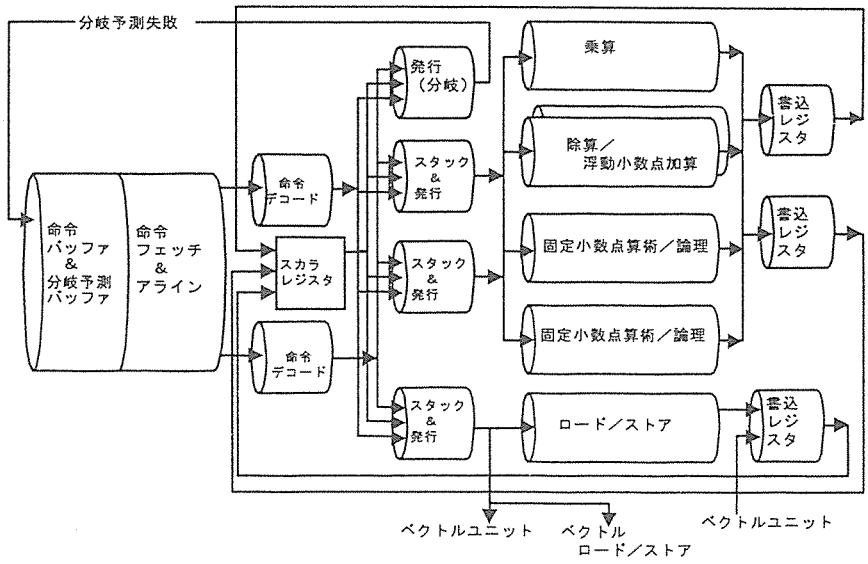


図４　ベクトルユニット（ＶＵ）の構成

（１）ベクトル演算部

ベクトル演算部は、加算器／シフタ、乗算器、除算器、論理演算器の４本のベクトル演算パイプラインからなるベクトル演算パイプラインセットを８セット、１４４Ｋバイトの容量をもつ７２個のベクトルレジスタ、２５６ビット長のデータを保持する１６個のベクトルマスクレジスタ、および、これらを制御する回路から構成されます。

１）ベクトル演算パイプライン

加算器／シフタ、乗算器、除算器、論理演算器からなるベクトル演算パイプラインセットは、合計４本のベクトル演算パイプラインを独立して動作できます。プロセッサは、ベクトル演算パイプラインセットを８セット並列にもつことにより、合計３２本のベクトル演算パイプラインの同時動作を可能とし、ベクトル演算の高速化を図っています。

各々の演算パイプラインは、ＳＸ－３の標準形式と拡張指数データ形式の２つの浮動小数点データ形式に加え、新たにＩＥＥＥの浮動小数点データ形式をサポートしています。

２）ベクトルレジスタ

それぞれが最大２５６要素まで保持することができるベクトルレジスタを７２個用意しています。ベクトルレジスタの総容量は１４４Ｋバイトとなります。主記憶装置からベクトルレジスタへのベクトルデータのロード、またはベクトルレジスタから主記憶装置へのベクトルデータのストアは、１セットのベクトル演算パイプラインあたり１マシンサイクルに２語（８バイト／語）の転送能力を持ち、８セットのパイプラインが同時に動作できるように構成しています。さらに、先行するベクトル命令によるベクトルデータの処理が完了するのを待たずに、最初の要素の演算結果がベクトルレジスタに書き込まれた時点で、その結果を入力とする後続ベクトル命令の実行を開始することができる自動チェイニング機構、データ転送前後のベクトルデータの格納読み出しを最適なタイミングで行う制御などを行っています。

３）ベクトルマスクレジスタ

ベクトルマスクレジスタを１６個備えており、それぞれ２５６ビット長のデータを保持します。マスク付き演算、ベクトルの圧縮／伸張／マージなどのベクトル命令の制御に使用するデータを保持し、保持するデータの各ビットがベクトルレジスタの保持する各ベクトル要素に対応します。

４）ベクトル制御部

ベクトル制御部は、スカラユニット内の命令制御部から送られてくるベクトル命令の解読、８セットの並列演算パイプラインをインタレース方式で動作させるための制御、８セットの並列演算パイプラインに対応するベクトルデータおよびマスクビットのアライン制御、ベクトルデータ長の制御を行います。

スカラレジスタとベクトルマスクレジスタ間のマスクビットの設定および読み出しも並列演算パイプラインを意識することなく実行できるようにしています。

４．入出力処理装置（ＩＯＰ）

入出力処理装置は、主記憶装置と周辺処理装置の間に位置し、プロセッサの指示により、プロセッサと独立して入出力処理を高速実行します。入出力チャネルとしては、ＡＮＳＩ　Ｘ３Ｔ９．３に準拠した超高速チャネルＨＩＰＰＩ、１６ビットＳＣＳＩ－２　ｄｉｆｆｅｒｅｎｔｉａｌ、ＦＤＤＩなどを提供します。

ＨＩＰＰＩはスーパーコンピュータの超高速周辺装置接続手段および超高速ネットワーク構築手段として広く使用されており、ＨＩＰＰＩ用超高速周辺装置としてアレイディスク、ネットワークとしてＨＩＰＰＩスイッチを使用したスター型ネットワーク、ＵｌｔｒａＮｅｔなどが接続されます。ＨＩＰＰＩの特徴は以下の通りです。

① １００Ｍバイト／秒の単方向インタフェース（入出力には２チャネル必要）

② ４バイト幅のＥＣＬレベルの対信号を使用し、最大２５ｍのケーブルで１：１接続される。

③ １Ｋバイト単位のブロック転送（バーストと呼ぶ）を基本に、大量のデータを高速に転送する。

④ 垂直パリティおよび水平パリティにより、データの正当性を保証している。

入出力処理装置は、ＨＩＰＰＩを基本チャネルとし、最大１６本のＨＩＰＰＩチャネルを接続できます。ノードあたり最大３台の入出力処理装置が接続ができ、ＨＩＰＰＩチャネルの最大本数は、ノードあたり４８本となります。また、チャネル拡張機構を接続することにより、１６ビットＳＣＳＩ－２ｄｉｆｆｅｒｅｎｔｉａｌ、ＦＤＤＩなどのチャネルをノードあたり最大１６チャネルまで接続することができます。

５．主記憶装置（ＭＭＵ）

プロセッサ上で実行するプログラムおよびそれらが使用するデータを格納します。

表１に主記憶装置の諸元、図５に主記憶装置の構成を示します。

表１　主記憶装置の諸元

項　　目	諸　　元
記憶容量	２５６Ｍバイト　～　１６Ｇバイト
インタレース	３２ウェイ　～　１、０２４ウェイ
記憶素子	４Ｍビット同期型スタティックＲＡＭ
論理素子	ＣＭＯＳ－ＬＳＩ
誤り訂正	１ブロック誤り訂正と２ブロック誤り検出
データ転送速度	１６Ｇバイト／秒　～　５１２Ｇバイト／秒

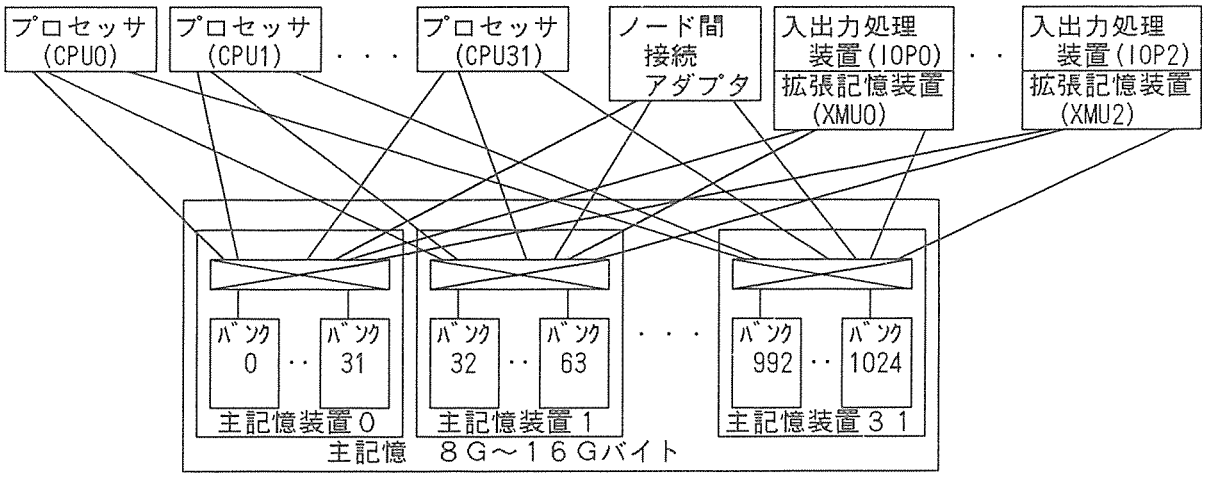


図５　主記憶装置（ＭＭＵ）の構成

主記憶装置は、３６ポート、３２バンク、最大５１２Ｍバイトの記憶容量をもち、ノードあたり最大３２台の主記憶装置を使用し、最大記憶容量１６Ｇバイト、最大インタレース数１、０２４ウェイ、最大ポート数１、１５２ポートの主記憶を構成します。１台のプロセッサは主記憶の３２個のポートと接続され、１マシンサイクルに最大１６語のベクトルデータをロードまたはストアすることができ、最大１６Ｇバイト／秒の極めて高いデータ転送能力を実現しています。

ノード間接続アダプタ、入出力処理装置／拡張記憶装置も、プロセッサと同じように、主記憶の３２個のポートに接続されます。

大容量高性能の主記憶装置を実現するため、記憶素子としてアクセス時間１５ナノ秒、４Ｍビットのシンクロナスｓｒａｍを使用しています。

データに対しては、1ブロックの誤り訂正と2ブロックの誤り検出符号を採用し、信頼性の向上を図っています。

６．拡張記憶装置（ＸＭＵ）

拡張記憶装置は、主記憶装置と磁気ディスク装置の間に位置する大容量半導体記憶装置です。拡張記憶装置は、最大８Ｇバイトの記憶容量を有し、４Ｇバイト／秒のデータ転送性能をもっています。ノードあたり最大３台の拡張記憶装置を接続することができ、最大４８Ｇバイト、１２Ｇバイト／秒のデータ転送速度を備える拡張記憶を構成することができます。

表２に拡張記憶装置の諸元を示します。

表２　拡張記憶装置の諸元

項　　目	諸　　元
記憶容量	２Ｇバイト　～　８Ｇバイト
増設単位	２Ｇバイト
記憶素子	１６Ｍビット・ダイナミックＲＡＭ
論理素子	ＣＭＯＳ－ＬＳＩ
誤り訂正	1ブロック誤り訂正と2ブロック誤り検出
最大データ転送速度	４Ｇバイト／秒

拡張記憶装置の特徴を以下に示します。

① 記憶素子に１６ＭビットのダイナミックＲＡＭを使用し、高密度実装とともに、装置の大容量／小型化を実現しています。

② 論理素子にプロセッサと同じＣＭＯＳ　ＬＳＩを採用し、処理速度の高速化を実現しています。

③ 電源供給モジュールを冗長構成とし、電源供給モジュールの障害が発生した場合でも、装置の電源断とならない方式を採用し、信頼性の向上を実現しています。

④ 内部処理データ幅を４バイト単位で複数のブロックに分割し、各々のブロックに対してＥＣＣ（Ｅｒｒｏｒ　Ｃｅｃｋｉｎｇ　ａｎｄ　Ｃｏｒｒｅｃｔｉｏｎ）による1ブロックの誤り訂正と2ブロックの誤り検出を行うことにより、ブロック間にまたがる複数ブロックエラーの自動訂正を可能とし、信頼性の向上を実現しています。

⑤ プロセッサに同期したクロックで動作することにより、データ転送速度の高速化を実現しています。

７．ノード間接続装置（ＩＸＳ）

ノード間接続装置は、ＳＸ－４シングルノードシステムを最大１６台接続できるパケット交換型スイッチネットワークです。シングルノード側に実装されるノード間接続アダプタが提供するデータムーバ機能とノード間接続装置のネットワーク機能により、他ノードのプロセッサを使用せずに、他ノードの主記憶装置上のデータをアクセスすることができるノード間主記憶データ転送機能を提供します。ノード間主記憶データ転送機能には、以下のものがあります。

・ブロック転送機能
・２ディスタンス転送
・間接アドレス転送（リストベクトル転送）

ノード間にまたがるマルチタスク方式の並列処理を高速化するため、上記ノード間主記憶データ転送機能の他に、ノード間接続装置はグローバル通信レジスタ（ＧＣＲ：Ｇｌｏｂａｌ　Ｃｏｍｍｕｎｉｃａｔｉｏｎ　Ｒｅｇｉｓｔｅｒ）を備えております。グローバル通信レジスタに対する操作には、以下のようなものがあります。

・ロード／ストア
・テスト＆セット
・フェッチ＆インクリメント
・フェッチ＆ディクリメント

ノード間接続装置とノード間接続アダプタ間は、超高速光インタフェースを使用し、８Ｇバイト／秒の極めて高いデータ転送性能を実現しています。光を使用することにより、ノード間接続装置とシングルノード間を接続する光ケーブル長は最大１００ｍまで延長することができます。表３にノード間接続装置の諸元を示します。

表３　ノード間接続装置の諸元

項　　目	諸　　元
最大データ転送速度	８Ｇバイト／秒
最大ケーブル長	１００ｍ
ＧＣＲの容量	４Ｋ語（３２Ｋバイト）
論　理　素　子	ＣＭＯＳ－ＬＳＩ
光　素　子	１．２５Ｇｂｐｓ　光モジュール

［執筆者紹介］　ＮＥＣ　コンピュータ事業部第四技術部