新世代HPCサーバPRIMEPOWER HPC2500

New generation HPC server PRIMEPOWER HPC2500

新庄直樹（富士通）
Naoki SHINJO (Fujitsu)

Abstact

The PRIMEPOWER HPC is a new Fujitsu HPC server, which combines the vector parallel processing technology of the VPP Series with the symmetric multi-processor (SMP) technology of the PRIMEPOWER UNIX server. The PRIMEPOWER HPC enables an SMP node configuration consisting of up to 128 high-speed (5.2 GFLOPS) scalar CPUs to be built. As up to 128 processing nodes can be crossbar-connected via a high-speed optical interconnection unit, the maximum configuration of the parallel multi-node system attains a logical performance of 85.2 TFLOPS and 64 Tbytes of main memory. Features for high performance include high-speed CPUs, high-speed (133 gigabytes per second) snoop performance, and a barrier synchronization unit, while features for high reliability include redundancy of the node crossbar switch units and partitioning. This paper describes the parallel processing method and hardware features of the PRIMEPOWER HPC.

1. はじめに

高性能計算機を用いた計算科学は，航空宇宙，バイオインフォマティクス，気象予測，構造解析など，学術・産業の幅広い分野で大きな成果を上げている。より高精度・大規模な計算を実現するために，より高性能な計算機が求められている。富士通は，このニーズに応えるため，次世代のHPC（High Performance Computing）サーバPRIMEPOWER HPCを開発した。

PRIMEPOWER HPCは，HPCサーバVPPシリーズのベクトルパラレル処理技術（以下，VPP技術）と，UNIXサーバPRIMEPOWERのSMP（Symmetric Multi Processor）技術を融合したHPCサーバである。

富士通が世界に先駆けて開発したVPP技術は，プログラミングの容易さ・並列処理効率の高さ・適用アプリケーションの広さで定評のある並列処理技術である。PRIMEPOWERは，富士通が持つメインフレームコンピュータ技術を駆使した，世界最高レベルの高性能・高信頼性・拡張性を実現するUNIXサーバである。PRIMEPOWER HPCは，これら富士通が持つ最先端の技術を駆使して開発された。

PRIMEPOWER HPCは，128個のスカラCPUから成るSMPノードを，高速光インタコネクト装置を用いて最大128台結合することで，世界最大級の並列マルチノードシステムを構築する。

本稿では，まずPRIMEPOWER HPCシステムの構成と，並列処理方式について述べ，さらにPRIMEPOWER HPCのハードウェアの特徴を紹介する。

2. PRIMEPOWER HPCシステムの構成

PRIMEPOWER HPCシステムの構成を図-1に示す。

8個のスカラCPU，メインメモリ，および基本IOアダプタを搭載したシステムボードと，ノード間データ転送を行うためのDTU（Data Transfer Unit）ボードを，高速なクロスバスイッチで接続してノードを構成する。ノードには最大16枚のシステ

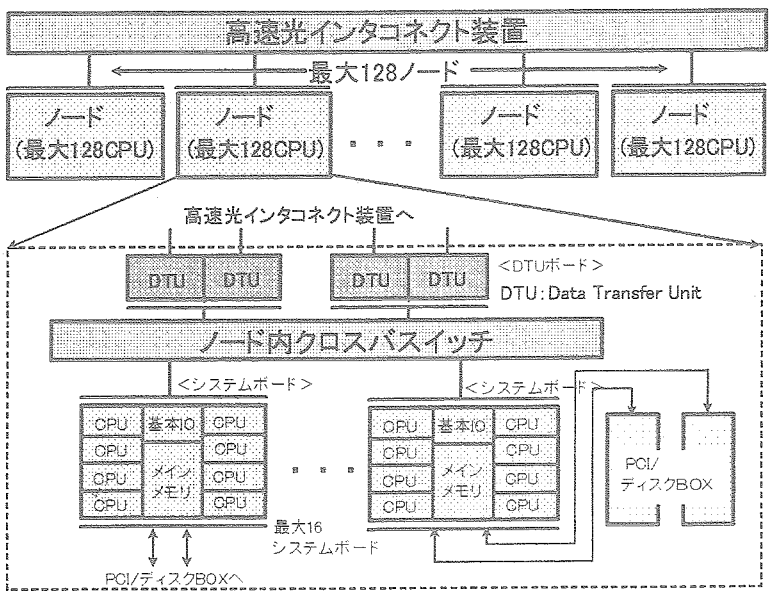

図-1 PRIMEPOWER HPCシステムの構成
Fig.1-System configuration of PRIMEPOWER HPC.

表-1 PRIMEPOWER HPCノード諸元

項目	諸元
CPU	SPARC64 V
CPU周波数	1.3 GHz
最大CPU数	128
アドレススヌープ性能	133 Gバイト/秒
最大メインメモリ容量	512 Gバイト
最大メインメモリインタリーブ数	512ウェイ
最大PCIスロット数	320

表-2 PRIMEPOWER HPCシステム諸元

項目	諸元
最大ノード数	128
最大CPU数	16,384
最大論理性能	85.2 TFLOPS*
最大メインメモリ容量	64 Tバイト
ノード間結合方式	クロスバ
ノード間転送性能	1ノードあたり 最大16 Gバイト/秒×2（入力/出力）

☆：Tera FLoating point Operation Per Second，Tera：10^{12}

ムボードと最大2枚のDTUボードを接続することができ，最大構成時には128CPU，メインメモリ容量が512 GバイトのSMPとなる。

各システムボードには基本IOとしてLANポート，シリアルポート，SCSIポートが装備されている。また，ケーブルを用いて外付けPCI（Peripheral Component Interconnect）/ディスクBOXを接続すると，ノードあたり320個のPCIスロットを利用できる。ノードの諸元を表-1に示す。

高速光インタコネクト装置を用いて128台のノードを接続した場合，16,384 CPU，論理性能85.2 TFLOPS，メインメモリ64 Tバイトのシステムとなる。PRIMEPOWER HPCシステムの諸元を表-2に示す。

3. PRIMEPOWER HPCの並列処理方式

PRIMEPOWER HPCでは，ベクトルパラレル処理とSMP並列処理を融合した並列処理方式を採る。

VPPシリーズでは，ベクトル演算ユニットと自動ベクトル化コンパイラの組合せで，PE（Processing Element）内で演算レベルの並列（ベクトル）処理を行う。さらに，PE間データ転送機構を用いて演算の中間結果をPE同士で高速に通信し合い，PE間バリア同期機構を用いてPE間の演算終了同期やデータ転送終了同期を高速に行うことで，PE内，PE間の2階層それぞれにおいて効率良い並列処理を実現してきた。

PRIMEPOWER HPCでは，VPPシリーズのPEを1ノードに対応させている。ノード内は，VPPシリーズのベクトル処理の代わりに，図-2に示すようにスカラCPUによる高速処理とノード内並列処理の組合せで高速化を実現する。スカラCPUによる高速処理は次に挙げる特徴を持つ。

(1) ソフトウェアパイプライニングによるループ高速化（図-2③）

(2) コンパイラとハードウェア協調によるプリフェッチの高速化，キャッシュの有効利用（図-2④）

また，ノード内並列処理は次の特徴を持つ。

(1) 高度なループ解析能力を持つ先進のコンパイラによるノード内自動並列化（図-2①）

(2) 多数のCPUを用いたスレッド間並列処理による高速化（図-2②）

(3) ハードウェアバリア同期機構を用いた高速スレッド間同期による並列処理オーバヘッド削減（図-2⑤）

さらに，VPPシリーズのPE間データ転送およびPE間バリア同期機構と同様に，PRIMEPOWER HPCにはノード間データ転送およびノード間バリア同期機構が設けられている。

以上のように，VPPシリーズにおけるPE内ベクトル処理，PE間データ転送処理，PE間バリア同期処理を，PRIMEPOWER HPCでは，スカラCPUによる高速処理とノード内並列処理，ノード間データ転送処理，ノード間バリア同期処理にそれぞれ置き換えて並列処理を行うことで，図-3に示すようにユーザから見たプログラミングモデルは同一になる。その結果，VPPシリーズで蓄積した豊富なHPCアプリケーション，チューニングツール，ノウハウをそのままPRIMEPOWER HPCでも利用できる。

さらに，PRIMEPOWER HPCでは，VPPシリーズと比較して，自動並列化・高速化コンパイラの適用範囲が大きく拡大した。VPP5000では1PEのメインメモリサイズは最大16 Gバイトで，データサイズがこれを超えなければ自動ベクトル化コンパイラを用いて容易に高速処理を行える。しかし，より大きなデータサイズの処理を実行するためには，PE間並列処理のためのプログラム書き換え・並列化指示行の追加が必要となる。

PRIMEPOWER HPCにおいて自動並列化・高速化の対象となるノードのピーク性能はVPP5000の1PEの約70倍，メインメモリ容量は32倍である。コンパイラによる自動並列化・高速化の適用範囲は，性能，データサイズの両面で飛躍的に拡大している。プログラムの並列化書き換え・並列化指示行の追加作業がネックとなり大規模高速処理を断念していたユーザでも，先進のノード内自動並列化コンパイラを用いて容易に大規模高速処理を行うことができる。

4. PRIMEPOWER HPCノードの特徴

PRIMEPOWER HPCのノードの特徴を以下に述べる。

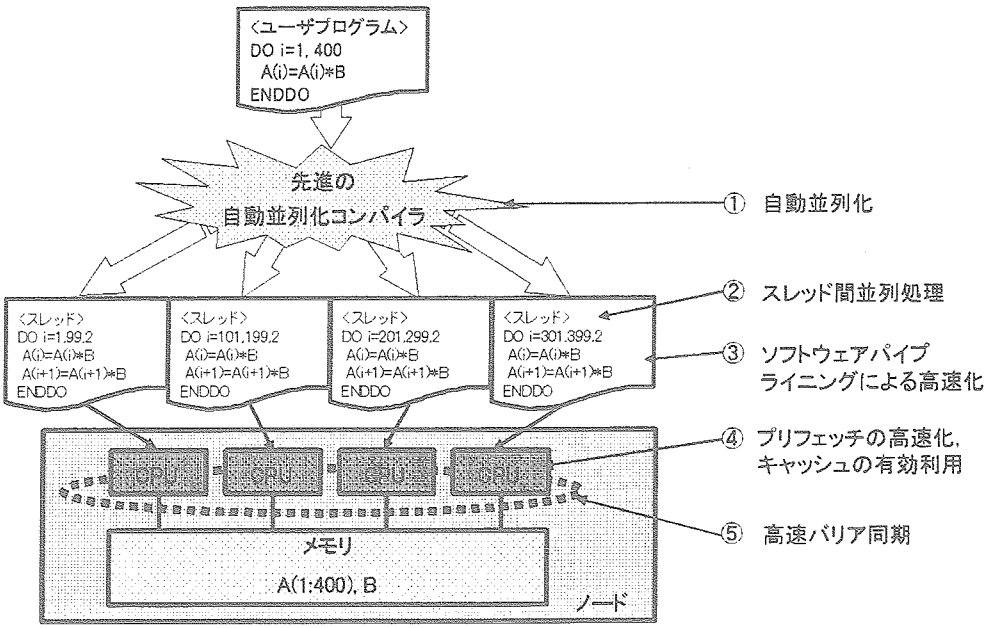

図-2 ノード内並列処理
Fig.2-Parallel execution method in node.

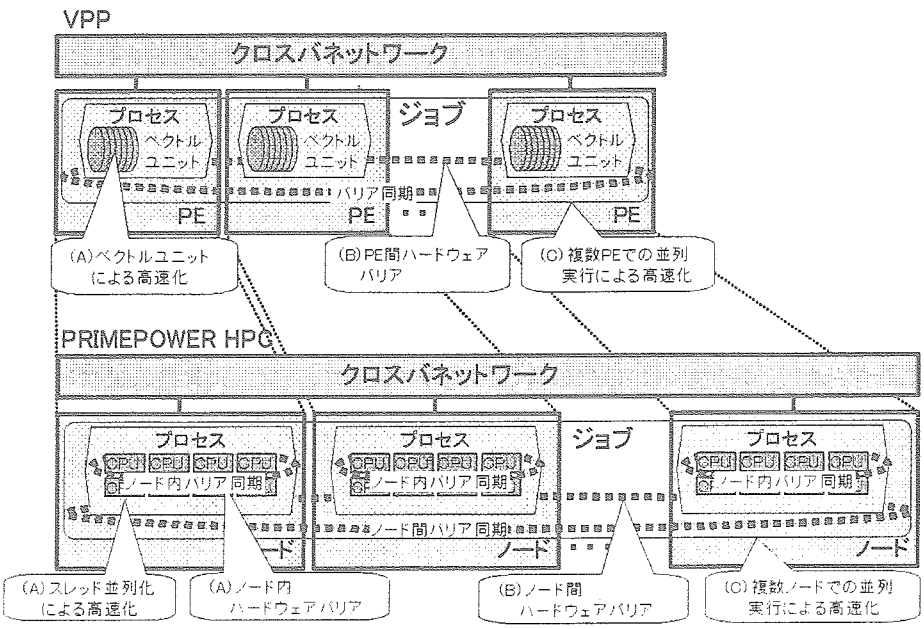

図-3 プログラミングモデル
Fig.3-Programming model.

● 高性能・高信頼SPARC64 V CPU

富士通のメインフレーム技術を取り込んだ高性能・高信頼なSPARC64 VをCPUに採用した。本プロセッサは内部周波数1.3 GHzで動作し，4命令同時発行，アウトオブオーダ実行，メモリアクセスの並列実行数増加など高速化のための機能を持つ。1次，2次キャッシュのECC (Error Checking and Correcting) によるデータ保護，命令リトライ，演算器のパリティ検査など，高信頼化の機能も持つ。

● 高速メモリシステム

HPCプログラムで高い実効性能を得るためにはプリフェッチを多重に発行し，演算器に次々にデータを供給することが重要である。このときメモリシステムには高いスループットと低いレイテンシが要求される。

まず，PRIMEPOWER HPCでは，SMPシステムの基本性能を示すスヌープ性能を133 Gバイト/秒にまで高速化した。また，システムボード間を接続するノード内データクロスバスイッチは，低レイテンシと高い実効性能を実現するために1階層クロスバ方式とした。システムボードとの接続は8.3 Gバイト/秒×2（入力/出力）のピークスループットである。

これら高性能なノード内クロスバスイッチを実現するため，クロスバスイッチボードとシステムボード間のバスにはソース同期転送方式を採用し，バスクロック周波数520 MHzの高速伝送を実現している。

メインメモリは，16システムボード搭載時に，ノード内で512ウェイインタリーブ構成となる。最大512 Gバイトの巨大な空間に対してノード内のどのCPUからも等距離・高速にアクセス可能であり，アクセスパターンに局所性を持たない大規模計算処理も高速処理が可能である。また，DDR-SDRAM（Double Data Rate Synchronous DRAM）素子使用によるデータバス使用効率の改善，SDRAM制御の最適化によるバンクビジーの短縮，さらに，プロセスがアクセスするメモリ領域のアドレスが局所的な場合でも一つのシステムボード上のメモリにアクセスが集中しないアドレス割り付けにより，高い実効メモリスループットを提供する。

● ノード内バリア同期機構

高い並列化効率を実現するためには，スレッド間並列処理のオーバヘッド削減は重要である。このため，ノード内バリア同期機構を導入した。バリア同期のための専用ハードウェアをノード内に設けるとともに，CPUの内部にもバリア同期機構専用のレジスタを設けており，高速なバリア同期が可能である。

● 大規模高速IO

ノード筐体の外にPCI/ディスクBOXを収納するIOラックを取り付けることで，ノードあたり最大32台のPCI/ディスクBOXを接続することができる。PCI/ディスクBOX内部には3個の64ビット/66 MHzのPCIスロットと，7個の64ビット/33 MHzのPCIスロットが実装されている。最大構成ではノードあたり320個のPCIスロットを搭載でき，大規模計算に不可欠な大容量入出力処理を高速に実行できる。

● 高密度実装

システムボードは高密度実装を採用し，8CPU，システムコントローラLSI，32枚のDIMM（Dual In-Line Memory Modules），システムボード内電源ユニットを80×475×585 mm（幅×奥行×高さ）のサイズにパッケージングした。システムボードの構造を図-4に示す。

一例として，16システムボード（128CPU，512 Gバイトメインメモリ），160PCIスロットから成るノードの構造を図-5に示す。IOラックを除いたノードは1,066×1,788×1,800 mm（幅×奥行×高さ）のキャビネットに収納する。

● ECCによるデータ保護

データバス，メインメモリ，およびCPU内キャッシュメモリはECCによる1ビットエラー訂正，2ビットエラー検出機能を持つ。これらの箇所において1ビット故障が発生しても，エラー訂正機能がデータを完全に訂正でき，運用を継続することが可能である。

● 冗長構成

ノード内クロスバスイッチは二重化されている。クロスバスイッチボードが故障した場合でも，故障クロスバスイッチボードを構成から外してリブートすることで，運用を再開できる。

ノードおよび高速光インタコネクト装置内の電源はすべてn+1冗長化されており，一つの電源ユニットが故障した場合でも残りの電源ユニットだけで運用を継続できる。

また，電源ユニット，ファントレイ，ディスク装置，システム監視機構が故障しても，業務を停止することなくこれら部品の交換が可能であり，高可用性を実現している。

● パーティショニング機能

パーティショニング機能により，ノード内のハードウェアリソースを論理的に分割し，各分割片（パーティション）でOSを動作させて独立したシステムとして運用することができる。運用コストや設置面積の削減が可能である。

● Solaris Operating Environmentを採用

OSにはSolaris Operating Environmentを採用しているため，ISV（Independent Software Vendor）アプリケーションをはじめとした豊富な既存アプリケーションをPRIMEPOWER HPCで利用できる。

5. 高速光インタコネクト装置の特徴

高速光インタコネクト装置は，VPPシリーズで培ったPE間高速データ転送技術をベースに，最先端の光伝送技術を用いて開発した超高速ノード間データ転送装置である。以下に本装置の特徴を述べる。

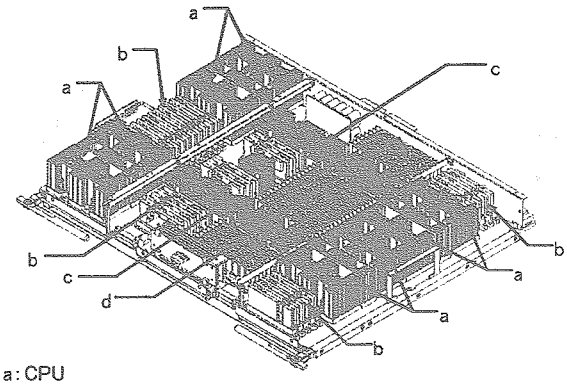

図-4　システムボードの構造図
Fig.4-Packaging structure of system board.

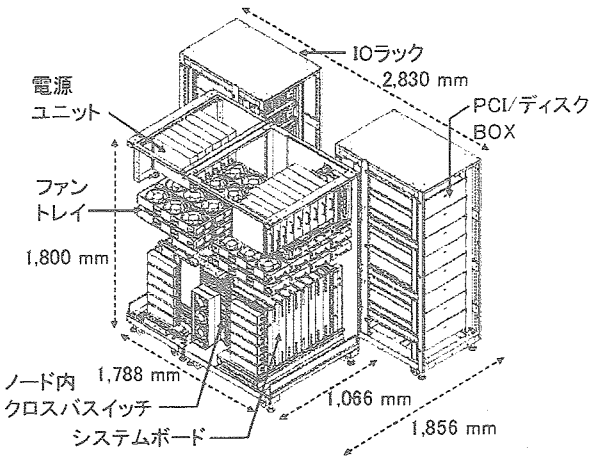

図-5　ノードの構造図
Fig.5-Packaging structure of node.

● ノード間データ転送機構

高速光インタコネクト装置は，最大128ノードをクロスバ結合して高速なノード間データ通信を提供する。高速光インタコネクト装置とノードの間は光ケーブルを用いて接続し，ピークスルー

プットは16 Gバイト/秒×2（入力/出力）である。

ノード間データ転送はノード内のDTUが行う。DTUは，連続メモリアクセス，ストライドメモリアクセス，メッセージ転送，返信要求転送など豊富なDMA（Direct Memory Access）モードを持ち，アプリケーションに応じて最適なモードでノード間データ転送を行える。また，アプリケーションプログラムがDTUを利用する場合，OSを介さずに直接DTUの起動が可能であり，ノード間データ転送のレイテンシを短縮している。

また，高速光インタコネクト装置には，エラー検出時の自動リトライ機能，および故障ビットの自動検出・交替機能を導入して高信頼化している。

● ノード間バリア同期機構

ノード間並列処理においても，高い並列化効率を得るためには，高速なノード間バリア同期機構が必要である。高速光インタコネクト装置はノード間バリア同期機構を備えており，異なるノード上で走行するプロセス同士でも高速に同期処理を行える。

6. むすび

本稿では，HPCサーバPRIMEPOWER HPCで採用した並列処理方式，高速化機能，および高信頼性・高可用性のための機能について紹介した。PRIMEPOWER HPCは富士通が持つ世界最先端のHPC技術と大規模UNIXサーバ技術を融合した，高性能・高信頼な大規模並列マルチノードシステムである。今後とも継続してHPCサーバの高性能化，高信頼化を進め，より広範囲のユーザニーズに応えることができるHPCサーバの開発を推進する。