3 ハードウェア
―ラック，冷却，プロセッサ，インターコネクト―

吉田利雄　池田吉朗　安島雄一郎
富士通（株）

本稿では「京」の主要ハードウェア技術であるSPARC64™ VIIIfx プロセッサ，Tofu インターコネクト，および水冷による高信頼性を解説する．「京」では水冷により LSI の温度を約 50℃下げ，プロセッサ LSI およびインターコネクト・コントローラ LSI の信頼性を数十倍高める．SPARC64™ VIIIfx は拡張命令セット HPC-ACE により 128GFLOPS の高性能を実現しながら，消費電力を 58W に抑える．Tofu インターコネクトは 6 次元メッシュ／トーラス・ネットワークにより高いスケーラビリティを実現する．

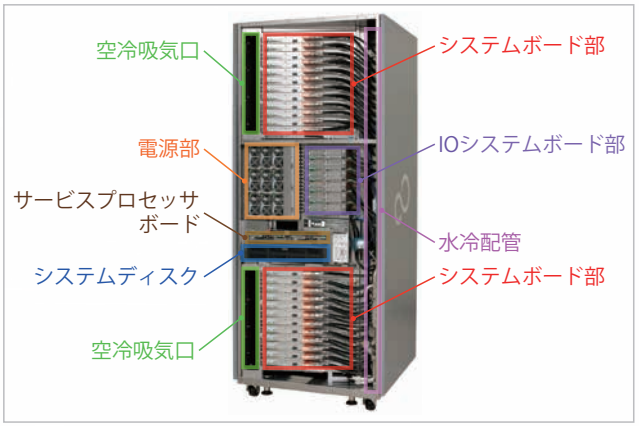

図 -1　計算ラックの構成

計算ラックと冷却方式

図 -1 に「京」の計算ラックの構成を示す．計算ラックの上段部と下段部は計算ノードを搭載するシステムボードを収納する．計算ラックの中央部は IO ノードを搭載する IO システムボード，電源ユニット，ブートのためのシステムディスク，システム監視のためのサービスプロセッサボードを収納する．計算ラック正面の左側に空冷吸気口，右側に水冷配管が配置される．各計算ラックは 24 枚のシステムボードと 6 枚の IO システムボードを搭載する．各システムボードは 4 つ，各 IO システムボードは 1 つのプロセッサを搭載する．計算ラックの設置寸法は 80cm × 80cm で 19 インチラックより幅が広く，奥行きは短い．計算ラック単体重量は約 1 トン，設置床耐荷重はケーブル等を含み約 1.5 トンである．計算ラックあたり 400 本以上のケーブルが接続され，ラック間を接続するケーブルは床下およびラック上部の架台に収納される．

システムの信頼性を確保するため，「京」では LSI に機能を統合して部品点数を抑えた．プロセッサ LSI にはプロセッサ・コアとメモリコントローラ，インターコネクト･コントローラにはネットワーク･インタフェースとネットワーク・ルータを統合した．それでも「京」が使用する LSI は合計 17 万個以上であり，24 時間の連続稼働には LSI 1 個の平均故障間隔を 480 年以上にする必要がある．これは非常に高い目標であるため，「京」では各 LSI に徹底した高信頼設計を施すとともに水冷方式を導入した．

●ジャンクション温度を下げる水冷方式

計算ラックの冷却系は空冷・水冷ハイブリッドである．水冷導入の目的は，LSI のジャンクション温度を下げることによる信頼性向上である．空冷における通常のジャンクション温度は 85℃程度であるが，「京」では水冷によりプロセッサおよびインターコネクト・コントローラのジャンクション温度

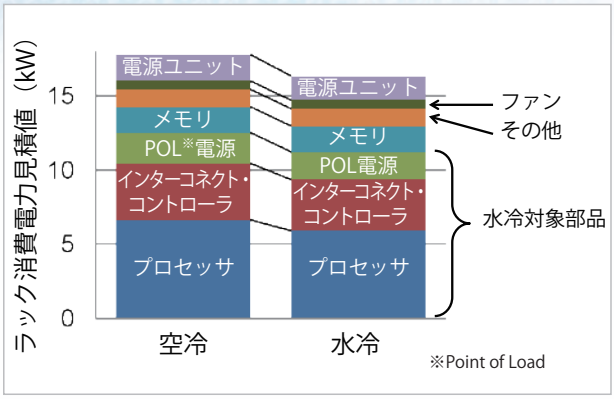

図 -2　計算ラックあたり消費電力の見積値

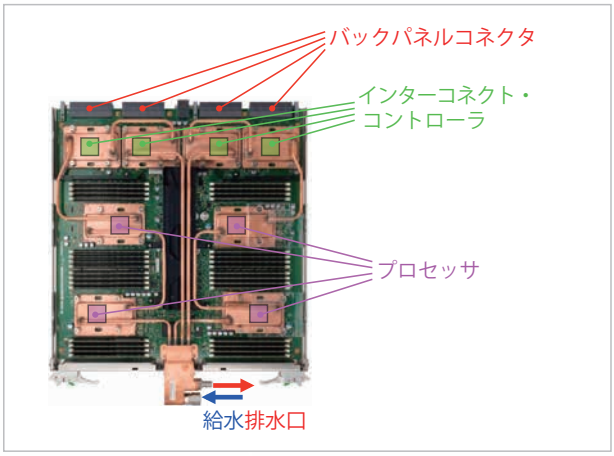

図 -3　システムボードの上面写真

を 30℃近辺にまで下げる．アレニウスの法則に従った経験則により 10℃の温度低下で LSI の寿命が約 2 倍になるので，LSI の平均故障間隔は水冷によって数十倍延びると予想される．また，低いジャンクション温度にはリーク電流を削減する効果もある．図 -2 に LINPACK ベンチマーク実行時の計算ラックあたり消費電力の見積値を示す．水冷の導入により「京」の消費電力は約 9% 削減された．これは 10PFLOPS の LINPACK ベンチマーク実行時で 1.3MW に相当する．

図 -3 にシステムボードの上面写真を示す．システムボード上には銅製のコールドプレートおよび水冷配管が据え付けられる．給水排水口に近い 4 つのコールドプレートの下にはプロセッサ，バックパネルコネクタ側の 4 つのコールドプレートの下にはインターコネクト・コントローラの LSI が実装される．

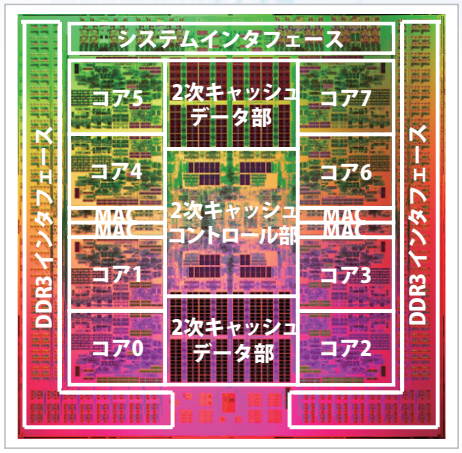

図 -4
SPARC64™VIIIfx チップ

SPARC64™ VIIIfx プロセッサ

SPARC64™ VIIIfx[1]（図 -4）は高性能，低消費電力，高信頼性を兼ね備えたスーパーコンピュータ向けのプロセッサである．富士通セミコンダクターの 45nm CMOS プロセスを採用し，8 つのコアと共有 2 次キャッシュ，メモリコントローラ（MAC），そして高速シリアル伝送のシステムインタフェースを内蔵する．動作周波数 2GHz でピーク性能は 128GFLOPS である．実アプリケーションで高い実行性能を発揮するために，SPARC-V9 アーキテクチャ☆1 の拡張を行い，科学技術計算を効率良く実行可能な命令セット HPC-ACE（HPC-Arithmetic Computational Extensions）☆2 を開発した．これに関しては後述する．チップ上の 8 つのコアによる並列処理を高速化するため，すべてのコアで 2 次キャッシュを共有し，さらにコア間の同期処理をハードウェアで行う機能を備える．これに富士通の自動並列コンパイラを組み合わせることで，ユーザはプログラミングの際に複数コアであることを特に意識せず，複数コアをあたかも高速な 1 つの CPU として扱うことが可能となる．これを富士通は VISIMPACT（Virtual Single Processor by Integrated Multi-core Parallel Architecture）と呼んでおり，SPARC64™ VII から継承している．

システム全体の電力制限からプロセッサの消費電

☆1 http://www.sparc.org/standards/SPARCV9.pdf
☆2 http://jp.fujitsu.com/solutions/hpc/brochures/

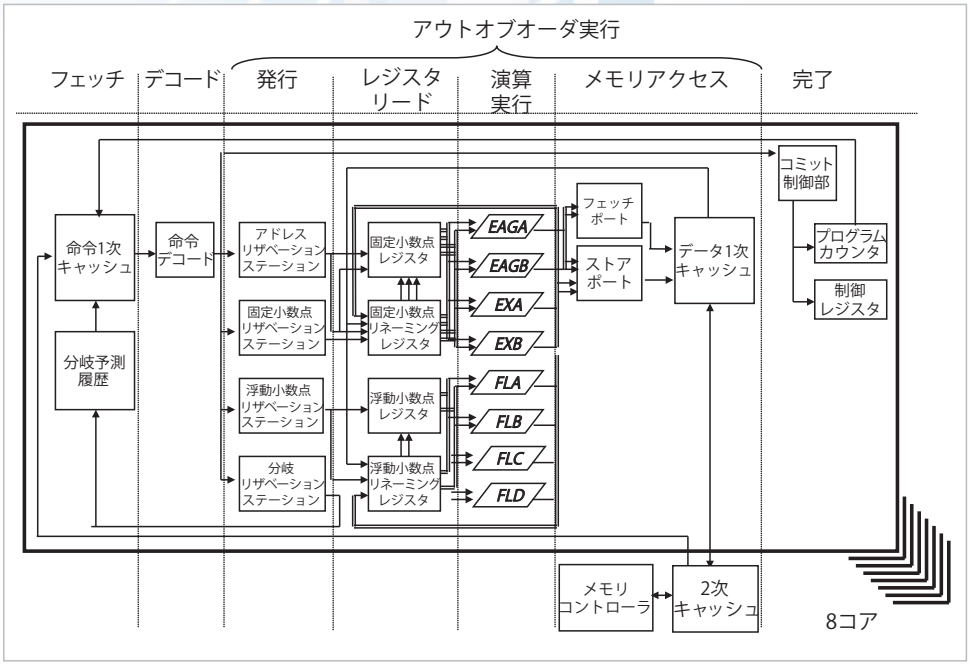

図 -5
SPARC64™ VIIIfx パイプライン

力を 58 W 以下に設計した．そのため低リークのトランジスタの使用や，水冷による冷却方式によりジャンクション温度を 30℃まで低下させてリーク電力をチップ全体の電力の 10% に抑えている．またラッチへのクロック供給を停止するクロックゲーティングを徹底して行い，電力削減に効果的な回路の組み方や制御方式に変更して動作時に消費するダイナミック電力を削減した．その結果 128 GFLOPS という高性能ながら，チップばらつきの平均で 58 W という低消費電力を実現した．これは電力あたりの性能で当社前機種の SPARC プロセッサに対し 6 倍以上にまで達する．

また，メインフレーム，UNIX サーバで用いる高信頼性技術を継承し，システムの安定稼働を実現する．プロセッサは非常に微細なトランジスタで構成されており，宇宙線の衝突などで信号が変化する可能性がある．このような間欠エラーの場合も誤動作することなく処理を続けるために，エラーが発生した命令をハードウェアで自動的に再実行する命令リトライ機構を備えている．またプロセッサ内のすべての RAM および固定小数点，浮動小数点レジスタの 1 ビットエラーはハードウェアで訂正処理を行う．プログラム実行に関連する部分についてはエラー検出コードで保護し，データ保全性を確保している．これらの技術によりプロセッサを 8 万個以上接続したシステムの安定稼働を実現する．

項　目	諸　元
ピーク演算性能	128GFLOPS
コア数	8
動作周波数	2GHz
浮動小数点演算器（コアあたり）	積和演算器 4
レジスタ数（コアあたり）	浮動小数点レジスタ（64 ビット）256 汎用レジスタ（64 ビット）188
1 次キャッシュ（コアあたり）	命令キャッシュ　32KB 2 ウェイ データキャッシュ 32KB 2 ウェイ
2 次キャッシュ（チップ）	6MB 12 ウェイ
プロセステクノロジー	富士通セミコンダクター（FSL） 45nm CMOS
ダイサイズ	22.7mm x 22.6mm
トランジスタ数	約 7 億 6,000 万個
メモリ帯域	64GB/s（理論ピーク値）
消費電力	58W（プロセス条件 TYP）

表 -1　SPARC64™ VIIIfx 諸元

●SPARC64™ VIIIfx のマイクロアーキテクチャ

SPARC64™ VIIIfx のパイプラインを図 -5 に，諸元を表 -1 に示す．コアは，命令制御部，演算処理部，1 次キャッシュ部からなる．命令制御部は命令フェッチ，命令デコード，命令のアウトオブオーダ処理制御，そして命令完了の制御を行う．演算部は，2 つの固定小数点演算器（EXA/B），2 つのロ

ード，ストアのアドレス計算を行う演算器（EAGA/B），および浮動小数点積和演算器（FMA：Floating-point Multiply-and-Add）を 4 つ（FLA/B/C/D）備える．FMA 演算器は SIMD（Single Instruction Multiple Data）構成を採り，1 つの命令で 2 つの演算を並列に行う．1 つの FMA 演算器は毎サイクル浮動小数点の乗算と加算を実行することが可能であり，各コアで毎サイクル 8 個，チップでは 64 個の倍精度浮動小数点演算が実行可能である．動作周波数は 2 GHz であり，ピーク性能は 128 GFLOPS となる．レジスタは固定小数点系で 192 本，浮動小数点系では 256 本である．

1 次キャッシュ部は，ロード，ストア命令を処理する．コアごとに 32 KB 2 ウェイの命令キャッシュとデータキャッシュをそれぞれ有する．データキャッシュは，2 つ同時にロードアクセスが可能なデュアルポート構成であり，16 バイトの SIMD ロードを 2 つ，または 16 バイトの SIMD ストアを 1 つ実行する．2 次キャッシュ部は 8 つのコアで共有され，各コアを含めたキャッシュコヒーレンスを保証する．また先に述べたように，コア間の高速同期処理のためコア間のハードウェアバリア機構を有する．メモリアクセスの低レイテンシ化，高スループット化のためメモリコントローラを内蔵した．メモリ帯域は理論ピーク値 64 GB/s である．また「京」専用のインターコネクトチップと高速シリアル IO で結合し，チップ間通信のスループットを確保している．

●命令拡張　HPC-ACE

HPC-ACE は SPARC-V9 アーキテクチャに対する科学技術計算向けの拡張命令セットである．レジスタ数の拡張，SIMD 演算，セクタキャッシュ機構，条件付き実行，三角関数の高速化命令，除算・平方根近似の機能を有する．以上の機能はいずれも周波数を上げることなく性能向上を可能とし，SPARC64™ VIIIfx の電力あたりの性能の向上に大きく寄与している．以下に各機能について説明する．

SPARC-V9 における浮動小数点レジスタの数は 32 本であり，HPC アプリケーションには十分な本数ではない．しかしレジスタ数を増やすにも 32 ビット長の SPARC アーキテクチャでは命令長が不足し不可能であった．この課題解決のため HPC-ACE では SXAR（Set eXtended Arithmetic Register）という前置命令を新設した．SXAR 命令は直後の最大 2 命令に対して，レジスタのアドレッシングの拡張などを行う．レジスタアドレスを 3 ビット拡張して浮動小数点レジスタ本数を SPARC-V9 の 8 倍，256 本まで指定可能にした（図 -6）．コンパイラはこの大容量レジスタを用いてソフトウェアパイプライニングなどの最適化を行い，アプリケーションが持つ命令レベルの並列性を最大限に引き出す．HPC の代表的なベンチマークの 1 つである姫野ベンチマークでは 1.65 倍の性能向上となっている．

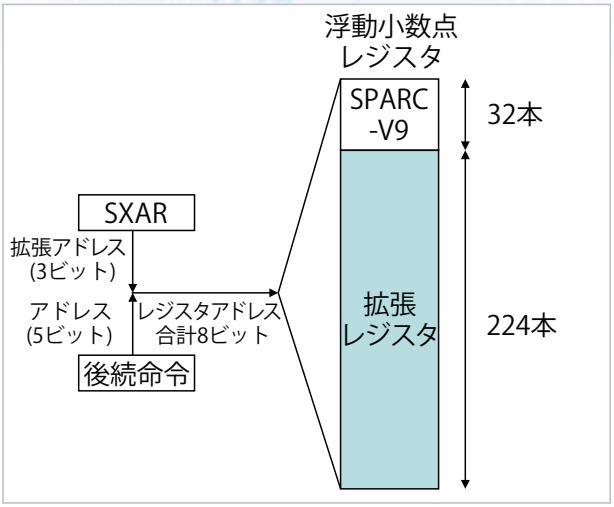

図 -6　SXAR 命令によるレジスタアドレス拡張

SIMD は 1 つの命令で複数のデータ処理を並列実行する技術である．HPC-ACE は SIMD 技術を採用し，1 つの命令で 2 つの FMA 演算を実行する．さらに複素数の乗算を高速化するための SIMD 演算もサポートしている．またロード命令とストア命令も SIMD 実行が可能となっている．ロード命令は倍精度のとき 8 バイトアライン，単精度のとき 4 バイトアラインでペナルティなく SIMD 処理を行う．

セクタキャッシュ機構はユーザが再利用頻度の高いデータをキャッシュに保持し続けるよう制御することを可能にする．条件付き実行命令は if 文を含むループを効率良く処理するために条件分岐命令を

削除する．具体的には新規の比較命令で比較結果を浮動小数点レジスタに書き込み，その比較結果に基づいて条件付き実行命令を処理する．条件付き実行命令には，浮動小数点レジスタ間のデータ転送と浮動小数点レジスタからメモリへのストアを用意した．これらの命令を組み合わせて条件分岐命令を取り除くことで，コンパイラは if 文を含むループに対してソフトウェアパイプライニングなどによる最適化が可能になる．また，HPC-ACE では三角関数の sin，cos の高速化命令を追加した．従来は多数の命令を組み合わせて処理を行っているが，専用命令化により命令数を削減したことで 5 倍以上高速化する．さらに逆数近似値を求める命令も追加した．これにより除算，平方根のパイプライン処理を可能にしており，レジスタ数の拡張と合わせた効果でおよそ 4 倍のスループット向上となっている．

Tofu インターコネクト

Tofu インターコネクト[2), 3)] は，10 万ノードのスケーラビリティと広帯域，低遅延を兼ね備えたインターコネクトである．Tofu インターコネクトの全機能は，単一のコントローラチップ，ICC（InterConnect Controller）に実装される．図 -7 に ICC チップの写真を示す．ICC は SPARC64 プロセッサのコンパニオンチップであり，チップ写真上部のバスインタフェースでプロセッサと接続する．赤いブロックは Tofu ネットワーク・インタフェース（TNI）である．TNI は ICC に 4 つ実装される．11 個の青いブロックは Tofu ネットワーク・ルータ（TNR）である．TNR はクロスバースイッチと 10 ポートのリンクモジュールで構成される．サーバの構成部品で例えると，TNI はネットワークインタフェースカード，TNR はスイッチボックスに相当する．また，ICC は第 2 世代の PCI Express を 2 ポート備える．PCI Express ポートは IO ノードでのみ使用される．

ICC は伝送速度 6.25Gbps のレーン 8 本からなるポートを 16 個備え，消費電力は 28W である．チップサイズは 18.2 ミリ× 18.1 ミリで，富士通セミコンダクターの 65nm CMOS プロセスで製造される．トランジスタ数は約 2 億で，312.5MHz で動作する．また，高信頼性のためにエラー訂正コードと耐ソフトエラー Flip-Flop が使用される．ICC 同士は電気的に直接相互接続する．リンク延長モジュールやスイッチチップは不要で，部品点数を削減する．また，柔軟な RAS 機能を実現するため，TNR はプロセッサや TNI とは独立に動作可能な設計になっている．

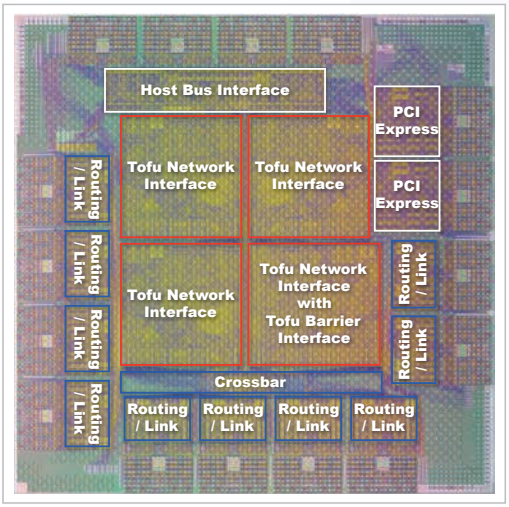

図 -7
ICC チップ

TNI は他ノードのメモリを直接参照する RDMA（Remote Direct Memory Access）通信機能を有する．Tofu インターコネクトの RDMA 通信では仮想アドレスによってメモリを保護する．TNI は主記憶上の仮想アドレス変換テーブルを検索するが，この際キャッシュ機能により変換オーバヘッドを最小化する．また，TNI は SPARC64 プロセッサのレジスタから直接，通信コマンドを受け取ることができる．この機能により，主記憶上のコマンドキューを使用する通常の場合に比べ，通信遅延が約 0.2 μ 秒短縮される．また 32 バイト以下のデータは通信コマンドに便乗でき，パケット組み立ての遅延が短縮される．

TNR はパケットを受信しきる前に次の TNR への転送を開始する Virtual Cut-Through 方式でパケットを転送する．この方式はパケットをすべて受信してから転送する Store and Forward 方式に比べ低遅延である．また，Tofu インターコネクトでは高いス

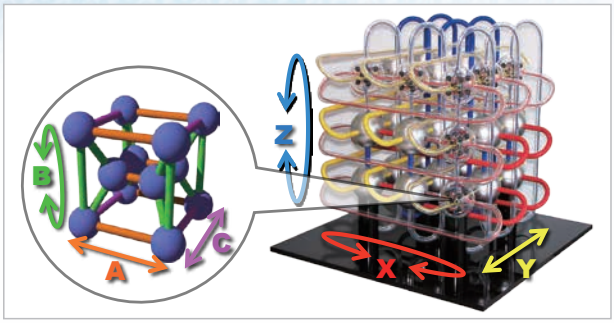

図 -8　6 次元メッシュ・トーラスの概念モデル

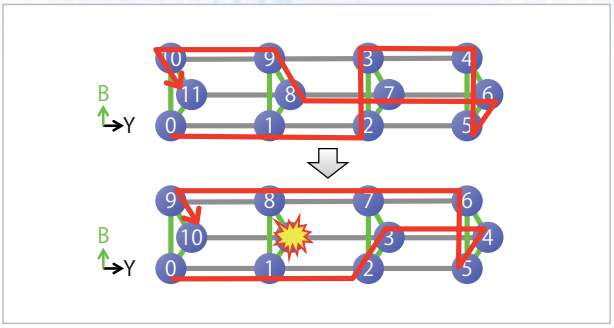

図 -9　仮想トーラスの構成と故障ノードの隔離

ケーラビリティを実現するため，ロスレスなパケット転送を行い，宛先ノードではタイムアウト検出を行わない．ロスレス転送を実現するため，TNR の各送信ポートは再送信バッファを備え，1 ホップごとに伝送エラーを修復する．

●6 次元メッシュ／トーラス・ネットワーク

Tofu インターコネクトのネットワークトポロジは 6 次元メッシュ／トーラスである．3 次元トーラスよりさらに次元数が多く，10 万ノードに達するスケーラビリティを実現する．ノードは 6 次元のネットワークで接続されるので，各ノードには（X, Y, Z, A, B, C）の 6 次元座標が与えられる．ここで，トポロジ全体の概念モデルを図 -8 に示す．システム全体は 3 次元メッシュ／トーラスであり，XYZ の座標軸が与えられる．XYZ の格子点には 12 ノードのグループが接続される．XYZ の格子点間はノードあたり 1 本，合計 12 本のリンクで相互接続される．各 XYZ 格子点の 12 ノードは 3 次元メッシュ／トーラスで接続され，ABC の座標軸が与えられる．

システム実装においては，X，Y 軸のリンクはラック間を接続し，長さはシステムのラック数に従って変わる．Z，B 軸のリンクはシステムボード間を接続し，Z 軸の長さはシステムモデルに従って変わり，B 軸の長さは 3 に固定される．A，C 軸のリンクはシステムボード上の 4 ノードを大きさ 2 × 2 の 2 次元正方形で接続する．4 ノードの接続トポロジはリングとも解釈できるが，後述のルーティング・アルゴリズムで 2 次元として扱うので 2 つの座標軸を与える．ICC あたり 10 ポートの内訳は XYZ 軸と B 軸が 2 ポートずつ，AC 軸が 1 ポートずつである．消費電力の制限から ICC あたりの合計入出力帯域は 100GB/s 程度だが，ポート数を 10 とすることでリンクあたり片方向 5GB/s，双方向 10GB/s の広帯域を確保した．

ルーティング・アルゴリズムは新規に開発した拡張次元オーダである．パケットを ABC 軸，XYZ 軸，ABC 軸の順に，ABC 軸を 2 回ルーティングすることを特徴とする．このアルゴリズムにより同一宛先への経路は合計 12 通り利用可能であり，データ転送経路の分散や，故障回避通信を可能にする．

6 次元座標系は通信パターンの最適化が難しいため，Tofu インターコネクトでは 1 次元／2 次元／3 次元の仮想トーラスを提供する．仮想トーラスは通信ライブラリの機能として実装される．ユーザは 3 次元以下の仮想座標で宛先を指定し，通信ライブラリは仮想座標と物理座標を相互に変換する．図 -9 に物理 Y，B 座標から仮想座標を生成する例を示す．上の図の例では Y 座標の長さ 4，B 座標の長さ 3 の領域に，0 から 11 の仮想座標を割り当てる．仮想座標で隣接するノードは物理的にも隣接する．仮想座標軸がリングになるように，仮想座標 0 と 11 も物理的に隣接する．図 -9 の下側の例では 1 つの座標に故障がある．この場合でも残り 11 個の座標を使用した仮想座標を生成できる．故障個所が 1 カ所であれば必ずリングの仮想座標を構築可能であり，仮想トーラスはシステムの可用性向上にも役立つ．

●Tofu バリア

Tofu バリア・インタフェース（TBI）はバリア同

期と縮約計算（AllReduce）の集団通信をハードウェアで処理し，遅延を削減して並列処理の効率を向上する．TBI のバリアパケットは演算タイプと 1 要素のデータを格納する．格納可能なデータ型は 64 ビット整数と独自浮動小数点数である．64 ビット整数は AND，OR，XOR，MAX，SUM 演算に対応する．独自浮動小数点数は SUM 演算に対応する．独自浮動小数点数のフォーマットは低オーバヘッドで IEEE754 浮動小数点数との相互変換が可能であり，計算順序によって SUM の結果を変えないために，154 ビットの仮数を 2 つ保持する．

集団通信は，各ノードがパケット受信，演算，パケット送信のステップを複数回行う通信アルゴリズムにより実現される．TBI ではバリアパケットの受信バッファと送信先設定を組にしたバリアゲートで集団通信を処理する．TBI は合計 64 個のバリアゲートを搭載し，任意の数のバリアゲートを使用することで多様な通信アルゴリズムを実行する．たとえば Recursive Doubling アルゴリズムを実行できる．Recursive Doubling アルゴリズムは N ノードを $\log_2 N$ ステップで同期するので低遅延だが，ノードあたり $\log_2 N$ 個のバリアゲートを使用する．Ring アルゴリズムはノードあたり 2 個のバリアゲートで実行できる．ただし Ring アルゴリズムは N ノードの同期完了に $2N$ ステップかかり，遅延が大きい．縮約と広報を二分木で行う Tree アルゴリズムは，ノードあたりの 5 個のバリアゲートで実行できる．Tree アルゴリズムは $2\log_2 N$ ステップで同期を完了するので，遅延とバリアゲート消費数のトレードオフは良好である．通信ライブラリは用途に応じて，適切な通信アルゴリズムを使用する．

従来の集団通信はソフトウェア処理のため，受信データ，送信データとも主記憶を経由し遅延が大きい．TBI はバリアパケットを ICC でバッファするため，主記憶参照が不要で遅延が小さい．さらに，ハードウェア処理による集団通信は OS ジッタの影響を受けない利点がある．OS ジッタとは，OS のプロセス・スイッチによってユーザプロセスの処理が数十から数百 μ 秒，最悪ケースでは 10m 秒程度中断されることにより，ユーザプロセスが使用できるプロセッサ時間に揺らぎが生じる現象である．集団通信処理では他のノードがデータを待ち合わせているため，OS ジッタの影響が広範囲に広がる．OS ジッタによる性能劣化は並列度が高いほど深刻になる．

まとめ

本稿では「京」の計算ラック，SPARC64^{TM} VIIIfx プロセッサ，Tofu インターコネクトの概要を解説した．10 万ノード級のスケーラビリティを有する Tofu インターコネクトにより，「京」は 88,128 個の SPARC64^{TM} VIIIfx プロセッサを接続した．SPARC64^{TM} VIIIfx は高性能，低消費電力，高信頼性を兼ね備えたプロセッサであり，拡張命令セット HPC-ACE により科学技術計算を加速する．Tofu インターコネクトは高スケーラビリティと広帯域，低遅延を兼ね備えたインターコネクトであり，6 次元メッシュ／トーラス・ネットワークはシステムの可用性も向上する．今後は将来のエクサスケールに向けて各技術の性能を高めるとともに，さらなる機能統合による高性能，低消費電力化を進める．

参考文献

1) Maruyama, T., Yoshida, T., Kan, R., Yamazaki, I., Yamamura, S., Takahashi, N., Hondou, M. and Okano, H. : SPARC64VIIIfx : A New-Generation Octocore Processor for Petascale Computing, IEEE Micro, Vol.30, No.2, pp.30-40 (2010).
2) Ajima, Y., Sumimoto, S. and Shimizu, T. : Tofu : A 6D Mesh/Torus Interconnect for Exascale Computers, IEEE Computer, Vol.42, No.11, pp.36-40 (2009).
3) Ajima, Y., Inoue, T., Hiramoto, S., Takagi, Y. and Shimizu, T. : The Tofu Interconnect, IEEE Micro, Vol.32, No,1, pp.21-31 (2012).

(2012 年 4 月 27 日受付)

吉田利雄 yoshida.toshio@jp.fujitsu.com
1999 年東京大学大学院理学系研究科物理学専攻修士課程修了．同年富士通（株）に入社．プロセッサ開発に従事．

池田吉朗（正会員） ikeda.yoshir-02@jp.fujitsu.com
1999 年北陸先端科学技術大学院大学情報科学研究科修士課程修了．（株）富士通研究所を経て，2007 年より富士通（株）勤務．プロセッサ開発に従事．

安島雄一郎（正会員） aji@jp.fujitsu.com
2002 年東京大学大学院工学系研究科博士課程修了．博士（工学）．（株）富士通研究所を経て，2007 年より富士通（株）勤務．計算機アーキテクチャの研究開発に従事．