White paper
スーパーコンピュータ PRIMEHPC FX10 の先進技術

富士通株式会社
次世代テクニカルコンピューティング開発本部

PRIMEHPC FX10

目次

スーパーコンピュータ PRIMEHPC FX10 の概要	2
高性能と省電力を両立した自社開発プロセッサ：SPARC64™ IXfx	3
科学技術計算命令拡張 HPC-ACE	4
ハイブリッド並列を実現するハードウェア	5
高性能・高信頼 Tofu インターコネクト	6
スケーラブル・高可用性 6 次元メッシュ／トーラス	7
低遅延集団通信 Tofu バリア・インターフェース	8

スーパーコンピュータ PRIMEHPC FX10 の概要

はじめに

富士通は 1977 年に我が国初のスーパーコンピュータを開発して以来、30 年以上にわたり最先端技術を投入したスーパーコンピュータを開発してきました。PRIMEHPC FX10 は、スーパーコンピュータ「京」(*1) で到達したペタスケールコンピューティングの世界をより身近なものにする、最新鋭のスーパーコンピュータです。

高性能 HPC 向け専用設計

PRIMEHPC FX10 は富士通が HPC 専用に設計した SPARC64™ IXfx プロセッサ、Tofu（Torus fusion）インターコネクトを搭載する超並列計算機です。SPARC64™ IXfx プロセッサは HPC 向け拡張命令と高効率マルチコア並列処理機構、8 チャネルの高メモリ帯域を有します。「Tofu インターコネクト」(*2) は 5GB/s の高速リンクでノード間を結合し、拡張性の高い「6 次元メッシュ／トーラス」(*3) 構成でシステムを構築します。

高効率 専用ノード構成

PRIMEHPC FX10 は計算ノードと I/O ノードの 2 種類の専用ノードからなります。計算ノードは計算処理に専従し、並列処理の効率を向上します。I/O ノードは PCI Express インターフェースを搭載し、ローカルファイルシステム、外部ファイルシステム、外部ネットワークとの高帯域な接続を実現します。

低消費電力・高信頼 直接水冷

プロセッサ、インターコネクト LSI を冷やすコールドプレートに冷却水を循環することにより、チップ温度の上昇を防ぎます。チップ温度を低く保つことにより半導体のリーク電流を削減するとともに故障率を低く抑え、低消費電力と高信頼性を実現しています。

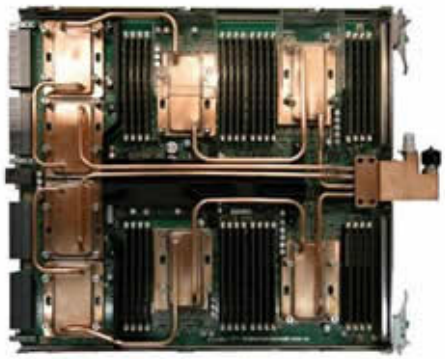

図 1 PRIMEHPC FX10 システムボード上の水冷配管

グリーン 間接水冷

筐体内のメモリや電源、ストレージから発生する熱を背面ドアの水冷ラジエータで吸収し、筐体外への熱の拡散を防ぎます。空調設備への負荷を軽減し、トータルの冷却コストを最適化します。

筐体構成

4 つの計算ノードを搭載するシステムボードを筐体あたり 24 枚、I/O ノード 1 台を搭載する I/O システムボードを筐体あたり 6 枚搭載します。また、筐体あたりシステム用 RAID 装置 1 台、電源ユニット 12 台、システム監視用サービスプロセッサ 2 台を搭載します。電源ユニットとシステム監視用サービスプロセッサは冗長構成により信頼性を確保します。

筐体あたりの計算ノード数は 96 台、I/O ノード数は 6 台です。PCI Express インターフェースは I/O ノードあたり 4 スロット、筐体あたり合計 24 スロットを搭載します。

図 2 PRIMEHPC FX10 システム外観

最大システム構成

最大で 1,024 筐体のシステム構成が可能です。最大構成は計算ノード 98,304 台、I/O ノード 6,144 台になります。

表 1 PRIMEHPC FX10 システム諸元

	64 ラック構成	最大構成
筐体数	64	1,024
計算ノード数	6,144	98,304
ピーク性能	1.4 ペタフロップス	23 ペタフロップス
メモリ容量	384 TB	6.0 PB
メモリ帯域	522 TB/s	8.3 PB/s
インターコネクト帯域	245 TB/s	3.9 PB/s
バイセクション帯域	7.6 TB/s	30.7 TB/s
I/O ノード数	384	6,144
拡張スロット数	1,536	24,576
消費電力	1.4 MW	23 MW

*1 「京」は 2010 年 7 月に理化学研究所が発表した次世代スーパーコンピュータの愛称です。

*2 「Tofu インターコネクト」は、富士通が独自に開発した高速インターコネクトです。

*3 「6 次元メッシュ／トーラス」は、Tofu インターコネクトのネットワーク接続形態です。詳しくは、「スケーラブル・高可用性 6 次元メッシュ／トーラス」を参照してください。

高性能と省電力を両立した自社開発プロセッサ：SPARC64™ IXfx

自社開発プロセッサ

SPARC64™ VIIIfx をベースにコア数を倍増し、性能を向上させた SPARC64™ IXfx を新たに開発して PRIMEHPC FX10 に搭載しました。

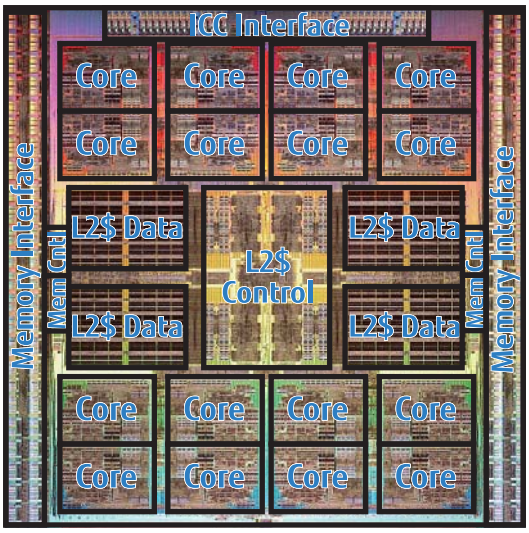

図 3 SPARC64™ IXfx

SPARC64™ IXfx の概要

SPARC64™ IXfx は 16 個のコア、コア間で共有される 12MB のレベル 2 キャッシュ、およびメモリコントローラーなどから構成されています。半導体には最先端 40nm テクノロジーを採用しています。

また、処理性能と消費電力とのバランスを重視し、高クロック化ではなくコア数を増やすことで、消費電力の増加を最小限に抑えつつ性能向上を実現しました。

各コアは IU（Instruction control Unit）、EU（Execution Unit）、SU（Storage Unit）の 3 つのユニットにわかれます。IU は命令のフェッチ、発行および完了を制御します。EU は 2 つの整数演算ユニット、ロード・ストア命令用の 2 つのアドレス計算ユニット、および 4 つの浮動小数点積和演算ユニット（FMA: Floating-point Multiply and Add）から構成され、整数演算、および浮動小数点演算命令を実行します。1 つの FMA は 1 サイクルあたり 2 つの浮動小数点演算（加算と乗算）を実行可能です。次ページで述べる SIMD 技術により、1 つの SIMD 演算命令で 2 つの FMA を動作させます。一方、各コアは 1 サイクルあたり 2 つの SIMD 演算命令を実行します。したがって、各コアで 1 サイクルあたり 8 個、チップ全体で 128 個の浮動小数点演算が実行可能となります。SU はロード・ストア命令を実行します。コアごとに 32KB のレベル 1 命令キャッシュとデータキャッシュをそれぞれ内蔵しています。

メモリコントローラー内蔵

近年のプロセッサは処理能力が急激に向上したため、演算に必要なメモリのデータ供給能力が相対的に不足してきています。SPARC64™ IXfx では、メモリコントローラーをプロセッサに内蔵することでこの問題に対処しました。これにより、チップセットを経由することなくメモリへのアクセスが可能となり、低レイテンシー化と高スループット化を実現しました。

高信頼性

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、宇宙線の衝突などにより信号が変化してしまう可能性がありますが、SPARC64™ IXfx では誤動作することなく処理を続けます。命令実行中にエラーが発生した場合は、その命令をハードウェアで自動的に再実行して処理を継続する命令リトライ機構を備えています。またプロセッサ内の大部分の回路をエラー訂正コードで保護します。プログラム実行に関連する部分については、すべてエラー検出コードで保護することで、データ保全性を確保しています。

HPC 向け命令セット拡張 HPC-ACE

SPARC64™ IXfx は、SPARC-V9 命令セットアーキテクチャーを拡張することにより、低消費電力かつ高い処理性能を実現しました。この拡張命令セットを HPC-ACE（High Performance Computing - Arithmetic Computational Extensions）と呼んでいます。HPC-ACE については次ページで詳しく解説します。

表 2 SPARC64™ IXfx 諸元

コア数	16
共有 L2 キャッシュ	12 MB
動作周波数	1.848 GHz
ピーク性能	236 Gflops
メモリ帯域	85 GB/s（ピーク値）
プロセス・テクノロジー	40 nm CMOS
ダイサイズ	21.9 mm × 22.1 mm
トランジスタ数	約 18 億 7000 万個
信号数	1,442
消費電力	110 W（プロセス条件：TYP）

科学技術計算命令拡張 HPC-ACE

HPC-ACE の概要

HPC-ACE は SPARC-V9 命令セットアーキテクチャーに対する HPC 向けの拡張命令セットです。HPC-ACE は高性能と省電力を両立します。

レジスタ数の拡張

SPARC-V9 における浮動小数点演算レジスタの数は 32 本です。この本数は他社汎用プロセッサと同程度ですが、スーパーコンピュータ用アプリケーションの性能を最大限に引き出すためには必ずしも十分ではありません。その対策として HPC-ACE では浮動小数点レジスタ本数を SPARC-V9 の 8 倍の 256 本に強化しました。コンパイラはこの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します。

HPC-ACE では SXAR（Set eXtended Arithmetic Register）と呼ばれる前置命令を新たに定義することで、256 本という多数のレジスタを指定可能としました。SPARC-V9 では命令長が 32bit に固定されており、1 つの命令中に拡張したレジスタ番号を指定するフィールドはありませんが、SXAR 命令でレジスタ番号の拡張部分上位 3bit、後続の 1～2 命令で従来通りのレジスタ番号 5bit を指定することで、合計 8bit＝256 本のレジスタを指定します。

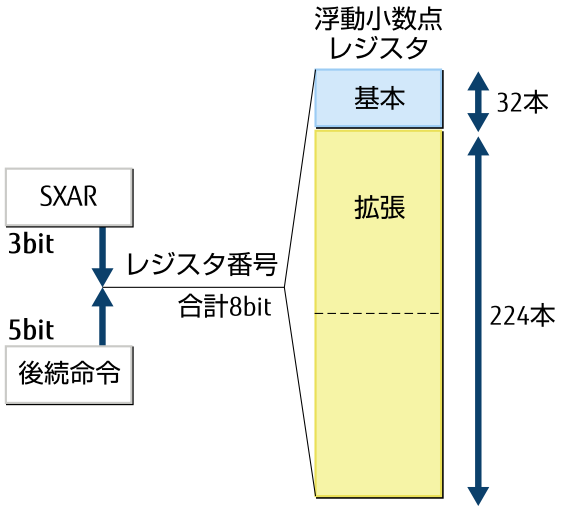

図 4 レジスタ数の拡張

SIMD 演算

SIMD（Single Instruction Multiple Data）は、1 つの命令で複数のデータに対する演算を実行させる技術です。HPC-ACE は SIMD 技術を採用することにより、1 つの命令で 2 つの浮動小数点積和演算を実行することが可能となりました。また複素数の乗算高速化のための SIMD 演算もサポートしています。

ソフトウェア制御可能キャッシュ（セクタキャッシュ）

プロセッサの処理速度とプロセッサにデータを供給する主メモリの速度の乖離が課題となっています（メモリウォール問題）。

メモリウォール問題の解決方法としては、キャッシュとローカルメモリがよく知られています。キャッシュはハードウェアが制御するのに対し、ローカルメモリはソフトウェアがデータアクセスを制御するためプログラムの大幅な書き換えが必要でした。

HPC-ACE では、従来のキャッシュとローカルメモリの長所を兼ね備えた、ソフトウェア制御可能なキャッシュ（セクタキャッシュ）を新たに開発しました。従来のキャッシュにおけるハードウェアによる制御では、再利用頻度の低いデータが、再利用頻度の高いデータをキャッシュメモリから追い出してしまい、性能向上の妨げになる場合がありました。セクタキャッシュでは、ソフトウェアがキャッシュ上のデータをグループ分けし、再利用頻度の低いデータを別の領域（セクタ）を割り当てることにより、再利用頻度の高いデータをキャッシュメモリに保持します。セクタキャッシュでは、従来のキャッシュの使いやすさを踏襲しつつ、必要に応じてソフトウェアによる制御を行うことで性能向上が可能です。

三角関数 sin, cos の高速化

三角関数の sin, cos 関数を高速化するため、専用命令を追加しました。従来は、多数の命令を多数組み合わせて処理を行っていましたが、専用命令化により命令数を削減し高速化しました。

条件付き実行

if 文を含むループの実行を効率化するために、条件付き実行の命令を追加しました。具体的には、まず比較命令で条件判定結果を浮動小数点レジスタに書き込みます。次に比較結果を用いて、選択的に浮動小数点レジスタ間のデータ転送、または浮動小数点レジスタからメモリストアを行います。これによりコンパイラは if 文を含むループについても、ソフトウェアパイプライニングによる最適化を行います。

除算、平方根近似

逆数近似計算を行う命令を追加しました。これにより、従来パイプライン処理ができなかった除算、平方根のパイプライン処理を可能にして、処理スループットを向上させています。

ハイブリッド並列を実現するハードウェア

ハイブリッド並列

VISIMPACT（Virtual Single Processor by Integrated Multicore Architecture）は、自動並列化をサポートするハードウェア機構と、高機能な自動並列化コンパイラにより、1 プロセスを複数コアで効率的に並列処理します。容易なプロセス／スレッドのハイブリッド並列化により、超高並列においてボトルネックとなるプロセス間通信時間の削減を実現します。

VISIMPACT を実現するハードウェア機構として、SPARC64™ IXfx は共有レベル 2 キャッシュとハードウェアバリアを備えています。

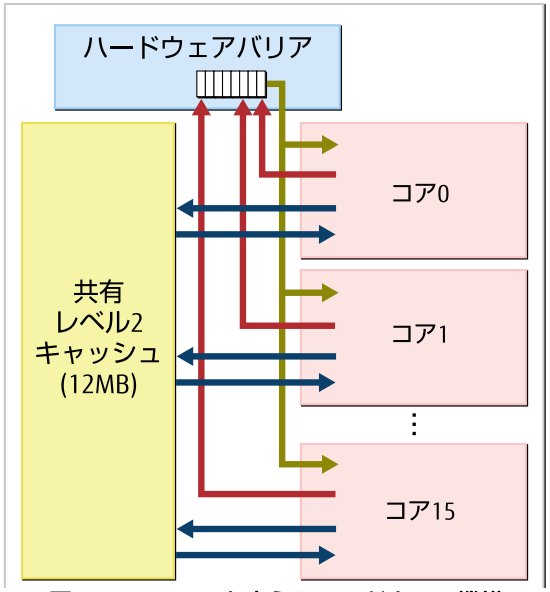

図 5 VISIMPACT を支えるハードウェア機構

共有レベル 2 キャッシュ

SPARC64™ IXfx は、12MB の共有レベル 2 キャッシュを持っています。プロセッサ内の全 16 コアでレベル 2 キャッシュを共有することでコア間でのデータ共有を容易にし、さらにキャッシュ上で隣接するデータを別々のコアで処理した場合の性能低下を防いでいます。レベル 2 キャッシュはラインと呼ばれる 128byte 単位でデータを管理しますが、同じラインに属する異なるデータを複数のコアが必要とした場合、このラインを複数のコアで取り合うことになります（false sharing）。

SPARC64™ IXfx はレベル 2 キャッシュを全 16 コアで共有する構造のため、このようなケースも効率的に扱えるようになっています。これは 1 つのプロセスを複数のコアで効率よく並列実行させるために不可欠です。

ハードウェアバリア機構

さらに SPARC64™ IXfx はハードウェアバリア機構を持っています。1 つのプロセスを複数のコアで並列実行させた場合、コア間で待ち合わせを行うケースが生じます。通常のプロセッサではこの待ち合わせ処理（同期処理）をソフトウェアで実現していますが、SPARC64™ IXfx は専用のハードウェアによりこの処理を 10 倍以上高速化しています。

ハードウェアバリア機構によって同期処理のオーバーヘッドが大きく削減されたことから、粒度の小さい演算ループに対しても、複数コアによる並列化処理を効率的に行うことが可能です。

自動並列化

SPARC64™ IXfx 上で複数コアによる並列実行を実現するために、特別なプログラムの書き方は不要です。富士通の自動並列化コンパイラは、プログラム上のループを分割し、複数コアで並列処理させることができます。

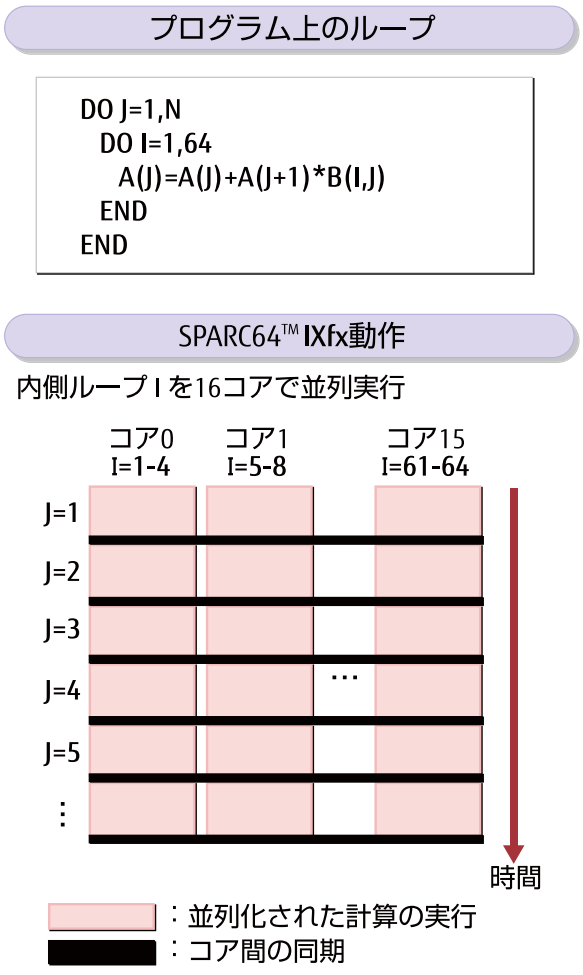

図 6 複数コアによるループ並列処理

高性能・高信頼 Tofu インターコネクト

インターコネクト・コントローラー ICC

PRIMEHPC FX10 のプロセッサ SPARC64™ IXfx には専用のインターコネクト・コントローラーICC が 1 対 1 で接続されます。ICC は PCI Express ルート・コンプレックスと Tofu インターコネクトを統合した LSI です。Tofu インターコネクトは ICC 間のパケット転送を行う Tofu ネットワーク・ルータ（TNR）と、プロセッサからのパケット送受信を行う Tofu ネットワーク・インターフェース（TNI）、集団通信を処理する Tofu バリア・インターフェース（TBI）で構成されます。TNI は ICC に 4 つ実装され、TNR は 10 ポートの Tofu リンクを備えます。ICC は Tofu リンクにより、最大 10 個の ICC と相互接続します。

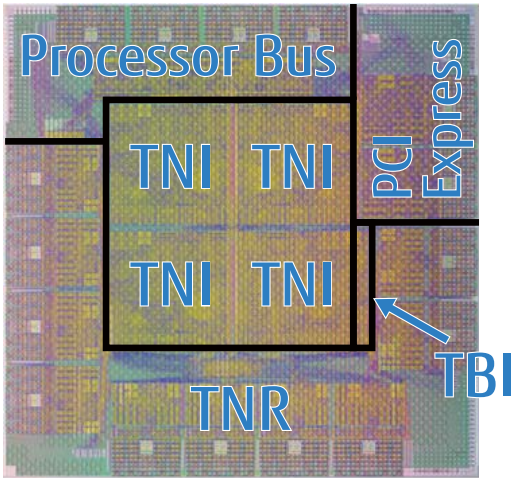

図 7 ICC のフロアプラン

RDMA 通信

TNI は RDMA 通信機能を備えます。RDMA 通信は、宛先ノードのソフトウェアを介在せずに、データの読み出しや書き込みを行う通信です。TNI は RDMA 通信コマンドの連続実行が可能です。1 コマンドあたり最大 16MB のデータを転送します。データは最大 2KB のパケットに分割して転送します。TNI は独自の 2 レベルアドレス変換機構により、仮想アドレス・物理アドレス変換とメモリ保護を行います。アドレス変換機構はハードウェアにより主記憶上のアドレス変換テーブルを検索します。また、キャッシュ機能により変換オーバーヘッドを最小化します。

低遅延パケット転送

TNR は Virtual Cut-Through 方式により、パケットを受信し切る前に次の TNR への転送を開始します。パケットを完全に受信してから次の TNR への転送を開始する Store and Forward 方式に比べ、低遅延を実現します。

4 方向同時通信

ICC は 4 つの TNI により 4 方向送信と 4 方向受信を同時に行います。非同期通信は複数並行転送、同期通信はデータ分割多重転送、集団通信は多次元パイプライン転送により、転送時間を短縮します。

仮想チャネル

Tofu インターコネクトはルーティングのデッドロック回避に 2 チャネル、要求／応答の用途別に 2 チャネルで合計 4 つの仮想チャネルを備えます。用途別仮想チャネルは要求パケットの輻輳による応答パケットの遅延を緩和します。Tofu インターコネクトの各受信ポートは 32KB の仮想チャネルバッファを備えます。

リンクレベル再送信

Tofu リンクは 1 ホップごとにエラーを修復する、リンクレベル再送信機能を備えます。TCP/IP や InfiniBand の送受信ノード間再送処理に比べ、Tofu インターコネクトのリンクレベル再送信は高速伝送路のビットエラーによる性能劣化を大幅に低減します。リンクレベル再送信のために、Tofu リンクの各送信ポートは 8KB の再送信バッファを備えます。

高信頼設計

ICC の SRAM および全データパス信号はエラー訂正コードで保護され、二次宇宙線などに起因するソフトエラーに耐性があります。デバッグ系を除く全制御信号はパリティで保護されます。ICC は機能ごとにモジュール設計されており、モジュール間を異常が伝播しない設計になっています。

表 3 ICC 諸元

同時通信数	4 送信＋4 受信
動作周波数	312.5MHz
スイッチ容量	100GB/s
（リンク速度×ポート数）	5GB/s×双方向X10 ポート
プロセス・テクノロジー	65nm CMOS
ダイサイズ	18.2mm×18.1mm
論理ゲート数	4,800 万ゲート
SRAM セル数	1,200 万ビット
差動入出力信号	
Tofu リンク	6.25Gbps、80 レーン
プロセッサバス	6.25Gbps、32 レーン
PCI Express	5Gbps、16 レーン

スケーラブル・高可用性 6 次元メッシュ／トーラス

6 次元メッシュ／トーラスの構成

6 次元メッシュ／トーラス・ネットワークの X 軸・Y 軸は筐体間を、Z 軸・B 軸はシステムボード間を、A 軸・C 軸はシステムボード上を接続します。各次元の軸を X, Y, Z, A, B, C と呼びます。X 軸・Y 軸はシステム規模により長さが可変です。Z 軸は座標 0 に I/O ノード、座標 1 以上に計算ノードが配置されます。B 軸は、3 つのシステムボードをリング接続して冗長性を確保します。A 軸・C 軸はシステムボード上の 4 ノードを接続します。

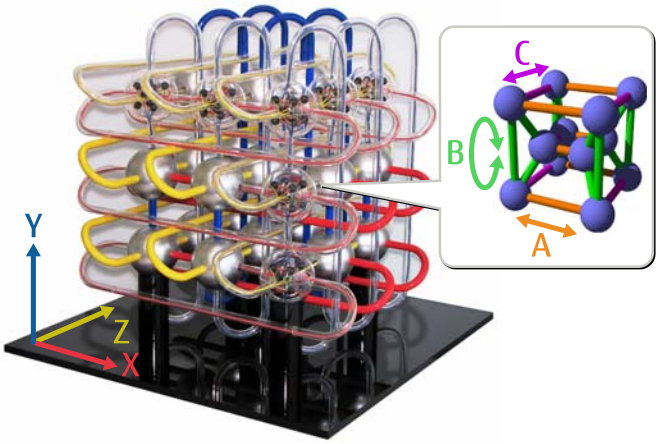

図 8 Tofu インターコネクト トポロジーモデル

高スケーラビリティ

6 次元メッシュ／トーラス・ネットワークは直接網と呼ばれるネットワークに属し、ケーブルを接続するだけでノード数の拡張が可能です。スイッチのポート数に制約されないため、最大システムで 10 万ノードを超える高いスケーラビリティを実現します。また外部スイッチが不要であるため、規模によらずノードあたりのコストは一定です。Tofu インターコネクトに必要な筐体間ケーブル数は、システム規模にかかわらずノードあたり 2 本です。

拡張次元オーダー・ルーティング

Tofu インターコネクトの拡張次元オーダー・ルーティングは、パケットを ABC 軸、XYZ 軸、ABC 軸の順に、ABC 軸を 2 回ルーティングします。通信ライブラリは送信コマンドごとに ABC 軸ルーティング経路を指定し、データ分割多重転送の経路分散や、故障回避経路の選択を行います。システムはジョブ開始時に故障ノードの位置情報を通信ライブラリに通知し、通信ライブラリによる故障回避を可能にします。

3 次元トーラス・ランクマッピング

Tofu インターコネクトは隣接通信を用いた通信パターンの最適化を容易にするため、ユーザーが指定する大きさの 1 次元／2 次元／3 次元トーラス空間をユーザービューとして提供します。ユーザー指定トーラス空間上の位置はランク番号で識別されます。

3 次元トーラスが指定された場合、システムは XYZ の 1 軸と ABC の 1 軸の組合せによる 3 つの空間を形成します。そして、各空間で一筆書きの隣接関係を保証するようにランク番号を与えます。図 9 に 8×12×6 サイズ 3 次元トーラス指定時の、ランク番号割当ての一例を示します。

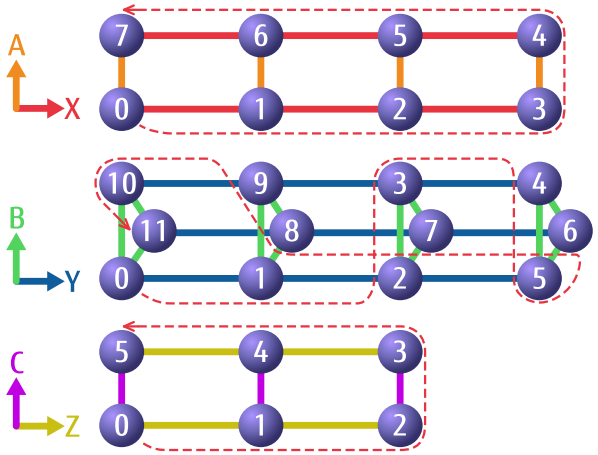

図 9 トーラス・ランクマッピングの例

高可用性

3 次元トーラス・ランクマッピングにおいて B 軸を含む空間は、B 軸が 3 ノードのリングであるため、1 ノードを避けて一筆書きを行うことが可能です。このため故障システムボードの保守交換時も周辺システムボードの運用継続が可能であり、システムの可用性を高めます。

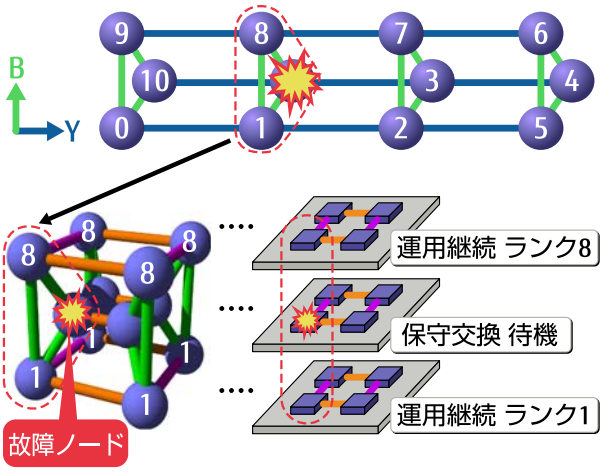

図 10 保守交換時のシステム運用継続

低遅延集団通信 Tofu バリア・インターフェース

Tofu バリア・インターフェース

集団通信処理は各ノードがパケット受信・演算・パケット送信を繰り返すことで実現されます。Tofu バリア・インターフェース（TBI）はプロセッサの代わりに集団通信を処理し、集団通信の遅延を短縮するハードウェアです。

TBI は受信・演算・送信動作を行うバリアゲートを 64 個搭載します。バリアゲートは 1 パケット分の受信バッファを持ちます。複数のバリアゲートによる複数回のパケット送受信は多様な通信アルゴリズムを実現します。

対応する集団通信の種類

TBI が対応する集団通信は Barrier と AllReduce です。AllReduce は Broadcast、Reduce としても利用できます。AllReduce 集団通信を実行するため、TBI はバリアパケットに演算タイプと 1 要素のデータを格納します。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す。64 ビット整数は AND, OR, XOR, MAX, SUM 演算に、独自浮動小数点数は SUM 演算に対応します。

独自浮動小数点数は通信アルゴリズムにかかわらず等値の結果を低遅延で得られるように設計されています。フォーマットは低オーバーヘッドで IEEE754 浮動小数点数との相互変換が可能な、独自の基数 2^{80} 半固定小数点です。計算順序によって SUM の結果を変えないために、加算は 154 ビット仮数を 2 つ使用する独自アルゴリズムで行います。

多様な通信アルゴリズム

バリアゲートは組合せにより様々な通信アルゴリズムを実現可能です。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します。

Recursive Doubling アルゴリズムは $\log_2 N$ ステップで N ノードの AllReduce を完了するので、低遅延を重視する用途に最適です。しかし、全ノードが $\log_2 N$ ステップ分の受信バッファを必要とするので、多数のバリアゲートを消費します。

Ring アルゴリズムは N ノードの AllReduce の完了に 2N ステップかかりますが、各ノードは 2 ステップ分の受信バッファだけを使うため、バリアゲートの節約に最適です。

縮約二分木と広報二分木で構成される Tree アルゴリズムは N ノードの AllReduce を $2\log_2 N$ ステップで完了し、各ノードのバリアゲート消費も 5 つに抑えられるため、低遅延とバリアゲート節約のバランスに適しています。

ハードウェアによる集団通信処理の利点

ソフトウェアで集団通信を処理する場合、受信データ・送信データとも主記憶を経由するために遅延が増加します。通信ハードウェアが集団通信処理を行う場合、主記憶参照が不要になり遅延が短縮されます。

またソフトウェアによる集団通信処理は OS ジッタの影響を受けます。OS ジッタとは、OS のプロセス・スイッチによってユーザープロセスの処理が数十から数百μ秒、最悪ケースでは 10m 秒中断されることにより、ユーザー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す。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ります。OS ジッタによる性能劣化は並列度が高いほど深刻になります。これに対し、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ります。

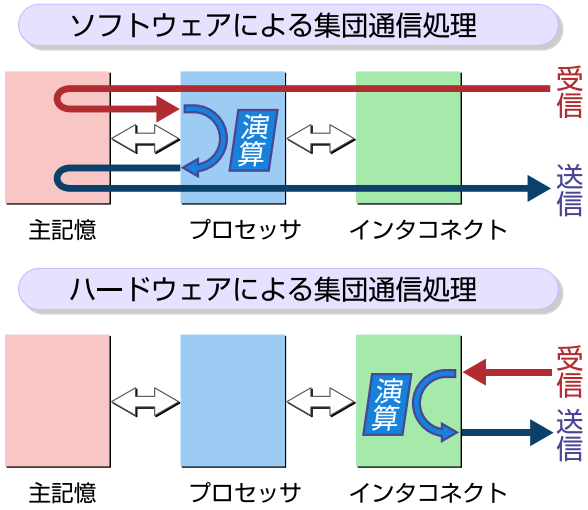

図 11 ソフトウェア集団通信処理とハードウェア集団通信処理の違い

参考情報
PRIMEHPC FX10 に関する情報は、当社営業までお問い合わせいただくか、以下の Web サイトをご参照ください。

http://jp.fujitsu.com/solutions/hpc/products/primehpc/

スーパーコンピュータ PRIMEHPC FX10 の先進技術
富士通株式会社
2011 年 11 月 7 日初版
2011-11-07-JP

・ SPARC64 およびすべての SPARC 商標は、米国 SPARC International, Inc.のライセンスを受けて使用している、同社の米国およびその他の国における商標または登録商標です。
・ その他、会社名と製品名はそれぞれ各社の商標、または登録商標です。
・ 本資料に掲載されているシステム名、製品名などには、必ずしも商標表示（TM 、（R））を付記しておりません。

本書を無断で複製・転載しないようにお願いします。
All Rights Reserved, Copyright © 富士通株式会社 2011