スーパーコンピュータ「京」のインターコネクトTofu

Tofu: Interconnect for the K computer

● 安島雄一郎　● 井上智宏　● 平本新哉　● 清水俊幸

あらまし

Tofu（Torus fusion）インターコネクトは，8万ノード以上を接続するスーパーコンピュータ「京」への搭載に向けて開発され，10万ノードのスケーラビリティと高性能，高信頼性，高可用性を兼ね備える。ネットワークトポロジーは高スケーラビリティの6次元メッシュ/トーラス，リンクあたりスループットは片方向5 Gバイト/秒であり，各ノードは4方向同時送受信が可能である。また，3次元トーラス・ランクマッピング機能により可用性を向上し，Tofuバリア・インターフェースにより低遅延で集団通信を実行する。Tofuインターコネクトを構成するネットワークインタフェースとルータは専用に開発されたLSI，インターコネクトコントローラ（ICC）に統合実装される。

本稿では，Tofuインターコネクトについて，ICC，6次元メッシュ/トーラストポロジー，高性能・高信頼通信機能，Tofuバリア・インターフェースの概要と特徴を紹介する。

Abstract

Torus fusion (Tofu) is an interconnect for massive parallel computers, and it has been developed to build the K computer that interconnects more than 80 000 nodes. The Tofu interconnect achieves high scalability beyond 100 000 nodes, high performance, high reliability, and high availability. The network topology is a highly scalable six-dimensional mesh/torus. The link throughput is 5 GB/s in each direction. Each node can communicate in four directions simultaneously. The three-dimensional torus rank-mapping scheme improves the system availability and the Tofu barrier interface (TBI) processes collective communications with low latency. Network interfaces and a router of the Tofu interconnect are integrated into a newly developed chip called InterConnect Controller (ICC). This paper describes overviews and characteristics of the ICC chip, the six-dimensional mesh/torus network, high-performance and highly reliable communication functions and the TBI.

まえがき

スーパーコンピュータ「京」(注)に搭載するTofu（Torus fusion）は，間接網を使用する従来の並列計算機と比較して2桁大きい，10万ノードのスケーラビリティを目標に開発されたインターコネクトである。Tofuインターコネクトでは，前例のない大規模な超並列計算機において高性能，高信頼性，高可用性を実現するために，多数の技術が開発された。

本稿では，Tofuインターコネクトの専用LSI，ネットワーク，通信機能について，概要と特徴を紹介する。はじめに専用LSIであるインターコネクトコントローラ（ICC）の概要を説明する。次に6次元メッシュ/トーラスについて述べ，さらに高性能・高信頼通信機能を紹介する。最後に特徴的な技術であるTofuバリア・インターフェースについて説明する。

インターコネクトコントローラ（ICC）

ICCはTofuインターコネクトを実装したLSIであり，SPARC64プロセッサに1対1で接続する。ICCはTofuネットワーク・ルータ（TNR），四つのTofuネットワーク・インターフェース（TNI），Tofuバリア・インターフェース（TBI），PCI Express，Processor Busで構成される（**図-1**）。TNRはTofuインターコネクトのパケットを転送し，TNIはプロセッサからネットワークへのパケット送受信を行い，TBIは集団通信を処理する。PCI Expressは拡張I/Oカードを接続し，I/Oノードでのみ使用される。TNRは10ポートのTofuリンクを備え，ICCはTofuリンクにより，ほかのノードに搭載される最大10個のICCと相互接続する。ICCの諸元を**表-1**に示す。

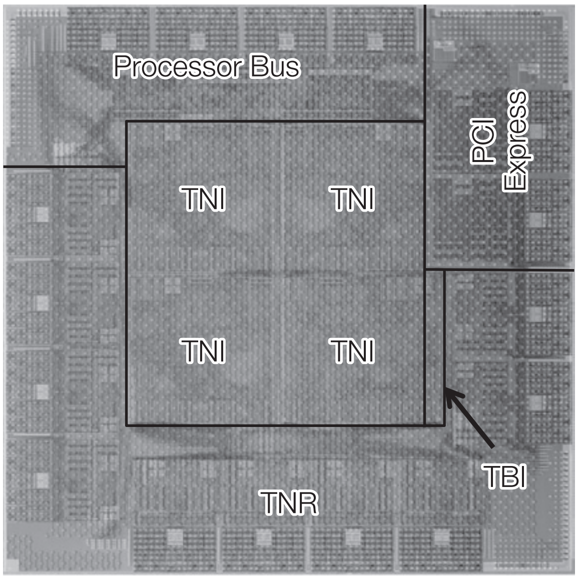

図-1　ICCチップ

6次元メッシュ/トーラスネットワーク

● ネットワークの構成

6次元メッシュ/トーラスネットワーク上の座標はX，Y，Z，A，B，Cの6次元で与えられる。X，Yは筐体間を接続する座標軸であり，X軸，Y軸の長さはシステム規模に対応する。Z，Bはシステムボード間を接続する座標軸であり，Z座標0にI/Oノード，Z座標1以上には計算ノードが配置される。B軸は3枚のシステムボードをリング接続し，冗長性を確保する。A，Cはシステムボード上のプロセッサを接続する長さ2の座標軸である。

全体のネットワークトポロジーは，大きさ2×3×2のABC3次元メッシュ/トーラスがXYZ3次元メッシュ/トーラスで結合された構造となる。このトポロジーを表すモデルを**図-2**に示す。

● 高スケーラビリティ

Tofuインターコネクトはケーブルを接続するだけでノード数を拡張でき，最大システムで10万ノー

表-1　ICC諸元

項目	諸元
同時通信数	4送信＋4受信
動作周波数	312.5 MHz
スイッチ容量	100 Gバイト/秒
リンク速度	5 Gバイト/秒×双方向
ポート数	10
プロセステクノロジー	65 nm CMOS
ダイサイズ	18.2 mm×18.1 mm
論理ゲート数	4800万ゲート
SRAMセル数	1200万ビット
差動入出力信号 Tofuリンク Processor Bus PCI Express	 6.25 Gbps，80レーン 6.25 Gbps，32レーン 5 Gbps，16レーン

(注)　理化学研究所が2010年7月に決定したスーパーコンピュータの愛称。

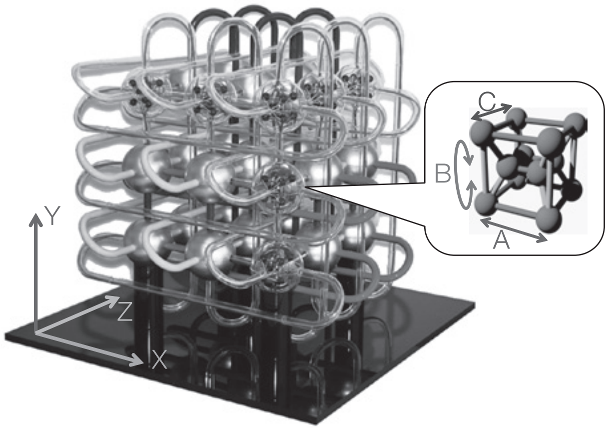

図-2　6次元メッシュ/トーラスのトポロジーモデル

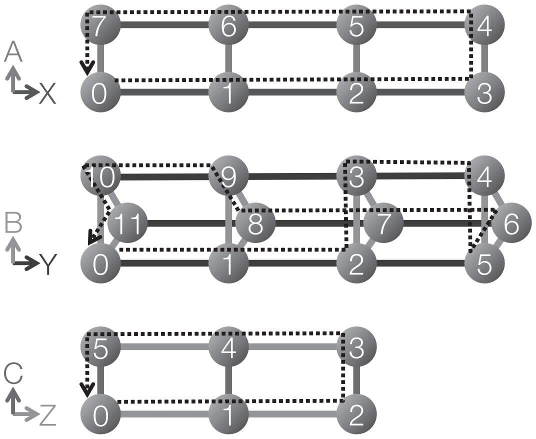

図-3　3次元トーラス・ランクマッピングの例

ドを超える高いスケーラビリティを実現する。6次元メッシュ/トーラスネットワークは，外部スイッチを使用しない「直接網」と呼ばれるネットワークに分類され，システム規模が変わってもノードあたりの平均ハードウェア量が大きく変わらない特徴がある。Tofuインターコネクトのノードあたりの平均ハードウェア量は，ICC1チップとケーブル約2.2本である。

●　拡張次元オーダー・ルーティング

パケットは座標軸をB，C，A，X，Y，Z，A，C，Bの順にルーティングされる。最初のABC軸ルーティングは宛先が12通りあり，送信コマンドごとに指定可能である。通信ライブラリはABC軸経路指定により故障ノードを回避して通信する。故障ノードの位置情報は，ジョブ開始時にシステムから通信ライブラリに通知される。

●　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ビューとして提供する。実行されるユーザプログラムはプロセスごとにランク番号が与えられる。各プロセスのユーザ指定トーラス空間上の位置は，ランク番号で識別される。3次元トーラスが指定された場合，システムはXYZから1軸とABCから1軸を組み合わせた空間を3組形成する。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る。**図-3**はユーザが8×12×6サイズの3次元トーラスを指定した場合の，ランク番号割当ての例である。

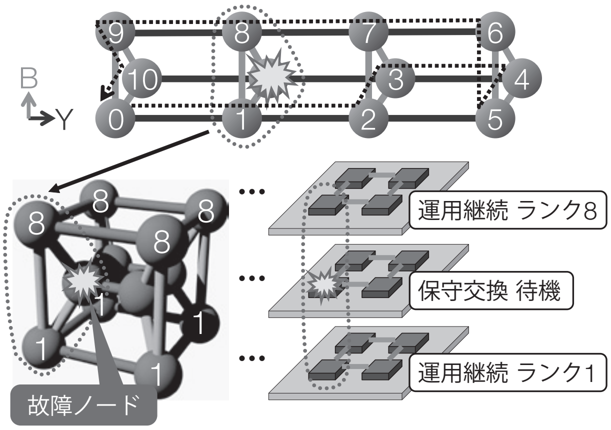

図-4　保守交換時のランク番号割当て

●　高可用性

Tofuインターコネクトは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，システムの高可用性を実現する。具体的には3次元トーラス・ランクマッピングにおいてB軸を利用する。B軸は3ノードのリングであり，B軸を含む空間は1ノードを避けて一筆書きを行うことが可能である。(1) 保守交換対象の座標を避けるランク番号割当て例を**図-4**に示す。

高性能・高信頼通信

●　RDMA通信

Tofuインターコネクトは RDMA（Remote Direct Memory Access）通信機能を備える。RDMAは宛先ノードに対してアドレス指定でメモリ参照する通信であり，宛先ノードにおける受信処理負荷が小さいことを特徴とする。Tofuインターコネクト

はユーザプロセスのメモリ空間に対してRDMA通信を行うために，仮想アドレス・物理アドレス変換機構を備える。アドレス変換機構はメモリ保護機能，変換キャッシュ機能，およびアドレス変換テーブル検索機能を有する。

● 低遅延・高効率転送

Tofuインターコネクトはvirtual Cut-Through方式による1ホップあたり約0.1μsの低遅延パケット転送と，最大2 Kバイトのパケット長による理論帯域比90%以上の高効率データ転送を両立した。Virtual Cut-Through方式はパケット・ヘッダを受信した時点で次ホップの送信を開始し，1ホップあたりの遅延をパケット長によらず最小化する。

● 4方向同時通信

TofuインターコネクトのTNIは，送信と受信を同時並行に処理する。四つのTNIは独立に動作するので，各ノードは4方向送信と4方向受信を並行して行うことが可能である。

通信ライブラリは複数の非同期転送を処理する場合，宛先別にTNIを割り当てる。同期転送を処理する場合，データを分割して複数のTNIに割り当て，異なる経路で並行転送する。集団通信では複数TNIにより，ツリーなどの分岐のある仮想トポロジー上でのパイプライン転送を実現する。

● 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・応答のデッドロック回避に2チャネルの仮想チャネルを備える。各受信ポートは仮想チャネルあたり8 Kバイト，合計32 Kバイトのバッファを備え，新開発の仮想チャネル・スケジューリング・アルゴリズムにより輻輳時のスループット低下を緩和する。(2)，(3)

● リンクレベル再送信

Tofuリンクはリンクレベル再送信機能により，高速伝送路のビットエラーを1ホップごとに修復する。TCP/IPやInfiniBandの送受信ノード間再送処理に比べ，リンクレベル再送信はビットエラーによる性能劣化を大幅に低減する。Tofuリンクの送信ポートはリンクレベル再送信のために8 Kバイトの再送信バッファを備える。

● 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，2次宇宙線などに起因するソフトエラーに耐性がある。また，デバッグ系を除く全制御信号をパリティで保護し，重要な制御信号は，異常な状態遷移を検出する監視ステートマシンにより，故障検出を強化した。ICCの各機能モジュールは故障を検出すると出力信号を閉塞し，異常を伝播しない設計になっている。

Tofuバリア・インターフェース

TBIは各ノードにおけるBarrier，Broadcast，Reduce，AllReduce集団通信の処理を，ソフトウェアに代わって実行するハードウェアである。TBIは遅延が小さいだけでなく，多くのアルゴリズムの実装が可能な柔軟な仕様となっている。TBIは八つのバリアチャネルを有し，バリア同期を並行実行できる。1チャネルは同期スケジューリングのためにシステムが予約しており，残り7チャネルは通信ライブラリが使用する。

● 低遅延処理

ソフトウェアによる集団通信処理とハードウェア（TBI）による集団通信処理の違いを**図-5**に示す。ソフトウェア処理では受信データ・送信データとも主記憶を経由するため遅延が大きい。TBIの通信処理は主記憶参照が不要であり，遅延が小さい。

● 通信アルゴリズム

各ノードは受信・演算・送信動作を行うバリアゲートを64個搭載する。X個のバリアゲートを使用すると，各ノードでX回の送受信を行う通信アルゴリズムを実現できる。通信ライブラリは用途に応

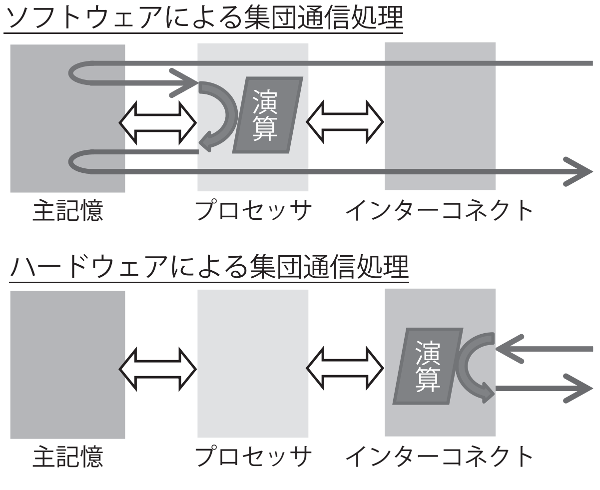

図-5　ソフトウェアとハードウェアの集団通信処理の違い

じて，適切な通信アルゴリズムを使用する。例えばNノードのRecursive Doublingアルゴリズムは遅延が$\log_2 N$に比例するので低遅延だが，各ノードで$\log_2 N$回の送受信を行うのでバリアゲート消費が多い。Double RingアルゴリズムはNに比例する大きな遅延がかかるが，送受信回数は2回と少ない。Treeアルゴリズムは遅延がRecursive Doublingの約2倍，送受信回数は5回であり，低遅延とバリアゲート節約のバランスに優れる。

● Reduce演算の種類

TBIが対応するReduce演算タイプは，64ビット整数に対してはAND，OR，XOR，MAX，SUM演算，浮動小数点数に対してはSUM演算である。浮動小数点のSUMは演算の順序にかかわらず等値の結果を低遅延で得られるように，中間結果を二つの160ビット浮動小数点数で表現する独自の演算方式を採用している。なお，TBIが対応するメッセージ長は1要素（スカラデータ）である。

● OSジッタ

ソフトウェア処理による集団通信はOSジッタにより性能が劣化する。OSジッタとは並列計算における計算プロセス間の処理のばらつきであり，デーモンプロセスへのプロセススイッチなどによる計算プロセス処理の中断により生じる。典型的な中断時間は数十μsから数msであるが，集団通信では多数のノードがデータを待ち合わせるため，処理の遅れが多数のノードに伝播し性能劣化が深刻になる。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る。

む す び

本稿では，10万ノードのスケーラビリティを有するTofuインターコネクトの設計，ネットワーク，通信機能について，概要と特徴を紹介した。Tofuインターコネクトは，ペタスケールを超えてエクサスケール実現に向かうスーパーコンピュータにおいて更に重要となる，スケーラビリティの高さと，高性能，高信頼性，高可用性を兼ね備えたインターコネクトとして，引き続き強化発展させていく。

参考文献

(1) Y. Ajima et al.：Tofu：A 6D Mesh/Torus interconnect for Exascale Computers. *IEEE Computer*，Vol.42，No.11，p.36-40（2009）.

(2) Y. Ajima et al.：The Tofu Interconnect. *The 19th Annual Symposium on High-Performance Interconnects*，p.87-94（2011）.

(3) Y. Ajima et al.：The Tofu Interconnect. *IEEE Micro*，Vol.32，No.1，p.21-31（2012）.

著者紹介

安島雄一郎（あじま　ゆういちろう）
次世代テクニカルコンピューティング開発本部システム開発統括部 所属
現在，スーパーコンピュータの研究開発に従事。

平本新哉（ひらもと　しんや）
次世代テクニカルコンピューティング開発本部システム開発統括部 所属
現在，スーパーコンピュータの開発に従事。

井上智宏（いのうえ　ともひろ）
次世代テクニカルコンピューティング開発本部システム開発統括部 所属
現在，スーパーコンピュータの開発に従事。

清水俊幸（しみず　としゆき）
次世代テクニカルコンピューティング開発本部システム開発統括部 所属
現在，スーパーコンピュータの開発に従事。