スーパーコンピュータ「京」「富岳」を実現した高次元接続技術

安島　雄一郎

あらまし

世界トップレベルのスーパーコンピュータである「京」およびその後継機「富岳」は，それぞれ88,192ノード, 158,976ノードを接続する超並列計算機である。10万ノードに及ぶ高い拡張性は，富士通が開発した高次元接続技術によって実現された。本技術の区画化機能，仮想トーラス機能によって，複数の並列プログラム間における通信干渉を防止し，各並列プログラム内の通信パターンの最適化を支援すると同時に故障ノードを含む区画を継続使用し，安定した通信性能と高い可用性を備えるシステムを実現した。

本稿では，「京」および「富岳」を実現した高次元接続技術について述べる。

1. まえがき

世界トップレベルのコンピュータ・シミュレーションは，社会が抱える複雑な課題の解決と先端科学技術の発展のために欠かせない道具となっている。大規模なコンピュータ・シミュレーションの実行には，並列計算機と呼ばれる種類のコンピュータが使用される。並列計算機は，プロセッサを搭載する多数のノードと，ノード間を接続するインターコネクトと呼ばれるシステム内ネットワークで構成されている。各プロセッサは大規模なシミュレーションの一部を担当し，インターコネクトを通して計算結果などのデータを交換しながら，シミュレーション処理を進める。

スーパーコンピュータ「京」[1, 2] は，88,192ノードを接続する超並列計算機であり，世界トップレベルのコンピュータ・シミュレーション実行環境を，国内外の研究者・技術者に提供してきた。また，その後継機「富岳」[3] は158,976ノード [4] を接続する。これらは，大型データセンター全体に匹敵する超大規模システムであり，ノード間を接続するインターコネクトには高い通信性能に加え，複数の並列プログラム間における通信の干渉を防ぐ機能と通信パターンの最適化を支援する機能，更にはシステムを停止させず，ノードの稼働率を維持する高い可用性が求められる。

本稿では，「京」および「富岳」のインターコネクトを実現した高次元接続技術について述べる。まず，2章で従来技術と問題，3章で高次元接続技術と実装について述べる。更に，4章で本技術の成果を紹介し，最後に5章で今後の展望を述べてまとめる。

なお，本項に関連して，「超並列計算機のためのプロセッサの高次元接続技術の開発」の業績によって，筆者は令和2年春の褒章において，紫綬褒章を受章した [5]。

2. 従来技術と問題

数百～数千ノード規模のインターコネクトでは，複数段のネットワークスイッチで構成されるFat-treeもしくはFolded-Closと呼ばれるネットワークが現在に至るまで主流である。特に世界トップレベルのスーパーコンピュータでは，Fat-treeを超えるインターコネクトの開発が重要な技術課題の一つとなっている。本技術の開発が始まった2005年度当時，3次元接続のインターコネクトが1万ノードを超える規模を実現していた。例えば，2004年に発表されたIBMのBlue Gene/L [6] では，組込みプロセッサを32,768ノード接続して，初代地球シミュレーターを超える性能を実現した。また，2005年に米国のCray Inc.が発表したRed Storm [7] では，汎用プロセッサを10,880ノード接続している。

3次元接続では，各ノードにネットワークルータを備え，ノードを3次元格子状に相互接続する。各次元を環状に接続するネットワークを3次元トーラス，各次元に端点があるネットワークを3次元メッシュと呼ぶ。4×4×4の3次元トーラスの概念図を図-1に示す。3次元トーラスを採用するシステムと3次元メッシュを採用するシステムには設計思想に相違があり，どちらも固有の問題があった。

Blue Gene/Lに始まる3次元トーラスのシステムでは，システムを複数の区画に分割しても各区画が3次元トーラスになる区画化機能を提供する。各並列プログラムは一つの区画を占有して実行され，他の並列プログラムからの通信干渉を受けない。また，3次元トーラスは3次元空間をシミュレートするアプリケーションに適し，更に各次元は対称なネットワーク，すなわち端，中央のような位置の違いがないネットワークになっているため，効率の

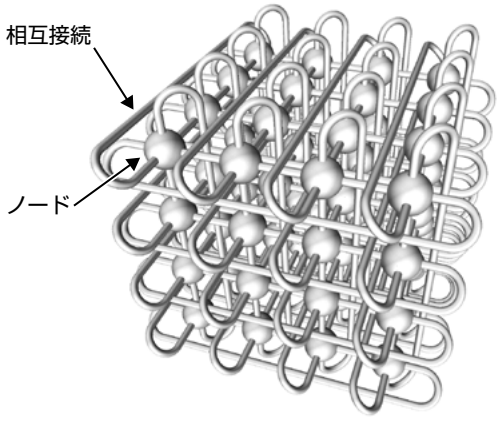


図-1　3次元トーラスの概念図

良いデータ転送をプログラムしやすい。3次元トーラスのシステムは，区画化機能を備えるために8×8×8のような一定の大きさの3次元メッシュを単位とし，各単位は各次元の両端を区画化スイッチに接続する。しかし，この方式ではノードが故障すると，故障していないノードを多数含む単位全体を停止し，システム中の利用可能なノード数が大幅に減少することが問題であった。

Red Stormを商用化したCray XTシリーズは，区画のない3次元メッシュを採用し，各並列プログラムはシステム中で離散的な位置にある空きノードを使用する。この場合，故障による可用性低下の問題は起こらないが，3次元のネットワークを想定した通信性能の最適化は困難である。それだけでなく，複数の並列プログラムが3次元ネットワーク上で重なり合って実行されるため，通信の干渉によって通信性能が低下する問題があった。加えて，3次元メッシュのシステムでは，ノードが故障しても周囲のノードを継続使用するため，故障ノード周辺におけるデータ通信経路の迂回（うかい）が更に通信の干渉を引き起こすという問題もあった。

3. 高次元接続技術と実装

高次元接続技術は，6次元メッシュ/トーラスと呼ぶネットワークによって，従来の3次元接続より1桁大きい規模の拡張性を実現する。具体的には，12ノードのグループを単位として，そのグループを更に3次元接続する。図-2に，6次元メッシュ/トーラスの概念図を示す。X，Y，Zはグループ間3次元接続における座標，A，B，Cはグループ内3次元接続における座標を表す。グループ内の12ノードは，A軸方向に2ノード，B軸方向に3ノード，C軸方向に2ノードの大きさの3次元格子状に接続されている。6次元メッシュ/トーラスを採用するシステムは3次元トーラスのシステムと同様に，区画化機能を提供して通信性能の低下を解決し，更に3次元メッシュのシステムと同様に，故障ノードを含む領域を継続使用して可用性低下の問題を解決する。

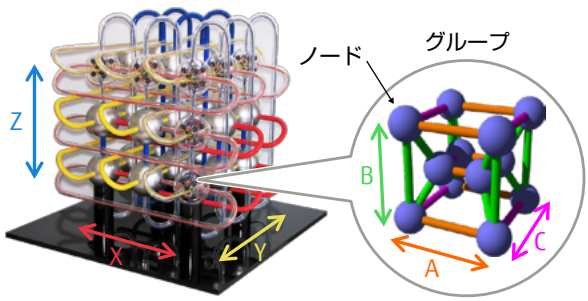


図-2　6次元メッシュ/トーラスの概念図

6次元メッシュ/トーラスは，グループ単位の任意の位置で区画化できる。3次元トーラスに比べて区画化スイッチが不要であると共に，区画の単位が従来技術より細かいため，様々な並列度の並列プログラムを同時に，並列プログラム内の通信が他の並列プログラムと干渉せずに実行することが可能である。高次元化によって増えたリンクは全て通信経路として使用されるため，ノード間の通信経路が短くなり，通信の混雑が緩和されて通信性能が向上する。

各並列プログラムは，一つの6次元区画を占有して実行されるが，仮想3次元トーラス機能によってノードの座標は3次元で提供される。その際に，仮想3次元トーラス座標で隣接するノードは，実際の6次元ネットワークでも隣接関係にあることが保証される。この機能によって，並列プログラムは3次元トーラスを想定した通信性能の最適化が可能になる。仮想3次元トーラス機能は，物理6次元ネットワークを3組の2次元空間に分割し，それぞれの2次元空間の全ノードを一筆書きで一度ずつ通って元の位置に戻る経路（ハミルトン閉路）を仮想1次元と見なして，座標を与えることで実現される。区間内に故障ノードが含まれていても，いずれか1組の2次元空間上で故障ノードを除外したハミルトン閉路を構成できれば，その区間は使用可能である。また，グループ内に大きさが3となる次元があるため，一つの故障ノードを含む区画は使用可能であることが保証される。ただし，複数の故障ノードを含む区画が使用可能かどうかは，区間内の故障ノードの位置に依存する。

本技術は，最初に「京」のTofuインターコネクト［8－10］に適用された。Tofuとは，Torus fusionを短縮した名称である。Tofuインターコネクトは，専用のICC（インターコネクト・コントローラ）に

実装され，最大投入帯域20 GB/s，最大スイッチング容量140 GB/sの広帯域を実現した。「富岳」のインターコネクトは，最新のTofuインターコネクトD [11] に強化された。記号「D」は，高密度（High-density）を意味する。TofuインターコネクトDは，A64FX™ CPUに内蔵され，最大投入帯域は40.8 GB/s，最大スイッチング容量は217.6 GB/sに増強された。図-3，図-4にICCおよびA64FX CPUのダイ写真を示す。

TofuインターコネクトのX，Y，Z軸の接続を環状に閉じるかどうかは，構築・運用の利便性を考慮してシステムごとに構成の選択が可能である。例えば「京」のシステム構成では，X軸はシステムを部分停止して保守する際にシステムが分断されないように環状に接続され，Y軸はシステムの拡張が容易となるように両端が開いており，Z軸は座標0に配置した入出力ノードへの通信経路を短くするために環状に接続された。

4. 本技術の成果

「京」は2010年に出荷が開始されて構築が始まり，2012年から正式運用を開始した。本技術によって実現されるTofuインターコネクトの高い拡張性，柔軟な区画化機能，高い可用性によって，「京」は88,192ノードを接続することが可能となった。創薬，地震・津波，気象，宇宙，ものづくり，材料の開発など幅広い分野で研究機関，大学，企業の研究者，技術者に利用され，計画保守期間を除いて97%を超える高いシステム稼働率を記録した。「京」はGraph500 [12] およびHPC Challenge Global FFT [13] といった，インターコネクト性能が重要となるベンチマークで世界1位の性能を記録しており，Tofuインターコネクトの高い性能が貢献している。「京」は2019年に稼働を停止した。本技術は，2014年度の恩賜発明賞を受賞した [14]。

後継機「富岳」は，2019年に出荷が開始され，TofuインターコネクトDによって158,976ノードを接続する。本技術は，エクサスケール世代まで世界トップレベルのスーパーコンピュータを支えることを目標として，2009年に発表された [8]。「京」に引き続いて本技術を採用した「富岳」が無事稼働することよって，本技術の目標が完遂される。

近年，微細電子工学の分野では半導体微細化技術の減速，高密度パッケージング技術の台頭など，新しい技術トレンドが生まれている。また，計算機科学・工学の分野では人工知能（AI）向けなどの領域特化型アーキテクチャが注目されている。この

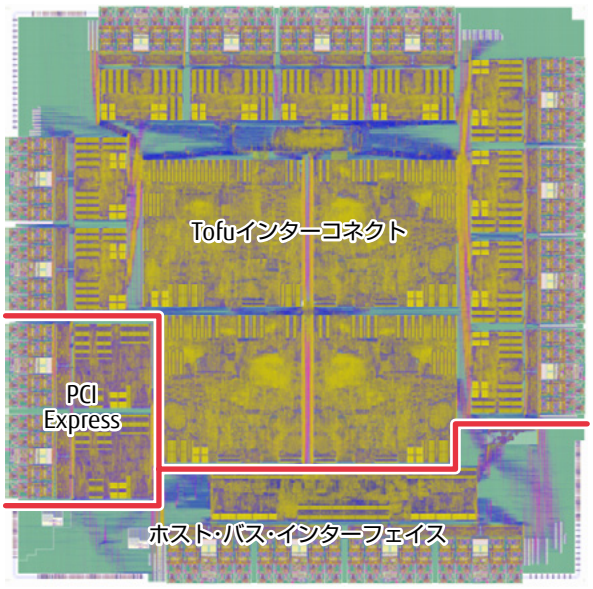


図-3　ICCのダイ写真

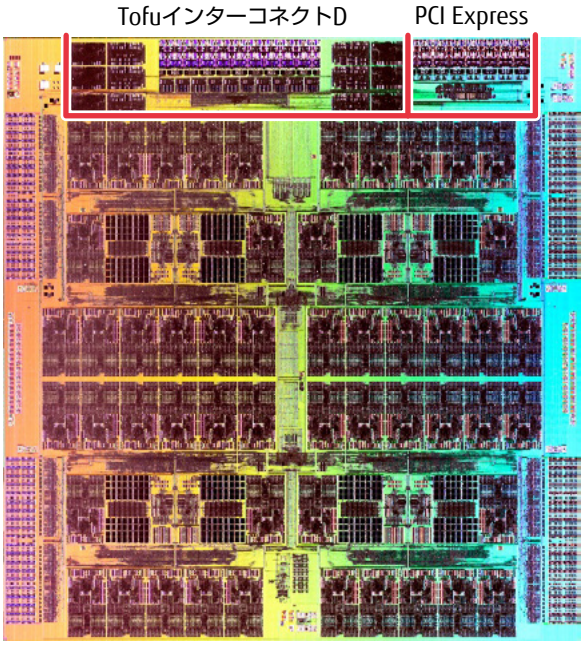


図-4　A64FX CPUのダイ写真

ようなことから，「富岳」の先にあるスーパーコンピュータにも，技術トレンドに対応した新しいシステムアーキテクチャとインターコネクトが求められる。

5. むすび

本稿では，高い拡張性，柔軟な区画化機能，高い可用性を備えるインターコネクトを実現する高次元接続技術について述べた。高次元接続技術によって，「京」および「富岳」は高い拡張性に加えて，性能が安定し，通信を最適化しやすく，信頼性の高いシステムを実現した。関係各位の長年に渡る多大なご尽力に深く感謝する。

富士通はスーパーコンピューティング分野で長年重要な役割を担い，社会の課題解決と科学技術の発展に貢献している。「京」の成功と「富岳」の実現によって，引き続き重要な役割を担っていくことが期待されている。

参考文献・注記

[1] 理化学研究所 計算科学研究センター:「京」について. https://www.r-ccs.riken.jp/jp/k/

[2] 宮崎博行 他：スーパーコンピュータ「京」の概要. FUJITSU，Vol. 63，No. 3，p. 237-246（2012）. http://img.jp.fujitsu.com/downloads/jp/jmag/vol63-3/paper02.pdf

[3] 理化学研究所 計算科学研究センター：スーパーコンピュータ「富岳」について. https://www.r-ccs.riken.jp/jp/post-k

[4] 理化学研究所：スーパーコンピュータ「富岳」. https://www.r-ccs.riken.jp/wp-content/uploads/2020/05/20200515_matsuoka.pdf

[5] 富士通：令和2年春の褒章において紫綬褒章を受章. https://pr.fujitsu.com/jp/news/2020/04/28.html

[6] N. R. Adiga et al.：An Overview of the BlueGene/L Supercomputer. Proceedings of the SC 2002 Conference on High Performance Networking and Computing (2002).

[7] R. Alverson：Red storm. Hot Chips 15 (2003).

[8] Y. Ajima et al.：“Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers.” IEEE Computer，Vol. 42，No. 11，p. 36-40 (2009).

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

[10] 安島雄一郎 他：スーパーコンピュータ「京」のインターコネクトTofu. FUJITSU，Vol. 63，No. 3，p. 260-264（2012）. http://img.jp.fujitsu.com/downloads/jp/jmag/vol63-3/paper05.pdf

[11] Y. Ajima et al.：The Tofu Interconnect D. IEEE International Conference on Cluster Computing，p. 646-654 (2018).

[12] Graph500. https://graph500.org/

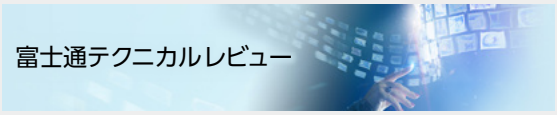
[13] HPC Challenge. https://www.hpcchallenge.org/

[14] 富士通：スーパーコンピュータの高次元接続技術が「恩賜発明賞」を受賞. https://pr.fujitsu.com/jp/news/2014/05/29.html

著者紹介

安島　雄一郎（あじま　ゆういちろう）
富士通株式会社
プラットフォーム開発本部
スーパーコンピュータのアーキテクチャ開発に従事。

この記事は，富士通の技術情報メディア「富士通テクニカルレビュー」に掲載されたものです。
他の記事も是非ご覧ください。



https://www.fujitsu.com/jp/technicalreview/