

White paper

FUJITSU Supercomputer PRIMEHPC FX1000

AI・エクサスケール時代を切り拓く HPC システム

富士通株式会社

目次

FUJITSU Supercomputer PRIMEHPC FX1000 の概要	2
HPC、AI 領域向け Arm プロセッサ A64FX	3
Arm 命令セットのベクトル拡張 Scalable Vector Extension 2.5 次元パッケージ技術による	4
CPU と 3 次元積層メモリ HBM2 の異種統合	6
高密度システム向け Tofu インターコネクト D	7



FUJITSU Supercomputer PRIMEHPC FX1000 の概要

はじめに

富士通は 1977 年に日本初のスーパーコンピュータを開発して以来、40 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。FUJITSU Supercomputer PRIMEHPC FX1000（以降、PRIMEHPC FX1000 と表記）は、AI・エクサスケール時代を切り拓き、スーパーコンピュータ「富岳」の世界をより身近にする、最新鋭のスーパーコンピュータです。

HPC、AI 領域向けの高性能設計

PRIMEHPC FX1000 は、富士通が HPC、AI 領域向けに設計した A64FX プロセッサを搭載する超並列計算機です。A64FX はスマートフォン等で普及している Arm アーキテクチャを採用しています。また Armv8-A 命令セットの HPC、AI 領域向けベクトル拡張 Scalable Vector Extension（以降 SVE と表記）を世界で初めて実装しました。A64FX の CPU チップは 48 個の計算コアと 4 個のアシスタントコアを有し、主記憶に 3 次元積層メモリ High Bandwidth Memory2（以降 HBM2 と表記）4 スタックを同一パッケージ内に搭載します。A64FX に実装された Tofu インターコネクト D（以降 TofuD と表記）は 20 レーンの高速信号でノード間を接続し、拡張性の高い 6 次元メッシュ／トーラス構成でシステムを構築します。

高信頼直接水冷

A64FX を搭載する CPU メモリユニット（以降 CMU と表記）ではコールドプレートに冷水を循環して A64FX、光トランシーバ、直流電圧変換素子を冷却し、半導体温度の上昇を防ぎます。低温を保つことにより素子の故障率を低く抑えます。

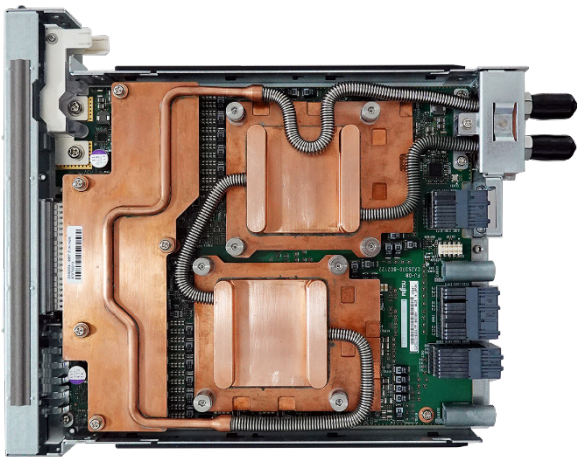


図 1 PRIMEHPC FX1000 CPU メモリユニット

本体装置およびラック構成

PRIMEHPC FX1000 の本体装置は CMU 24 枚、起動用ディスク 3 台、システム監視用サービスプロセッサ 3 台、Low Profile PCI Express 拡張スロット 6 本、電源ユニット 12 台を搭載可能です。ラックは前面と背面に本体装置を最大 4 台ずつ搭載します。ラックあたりの最大ノード数は 384 台です。

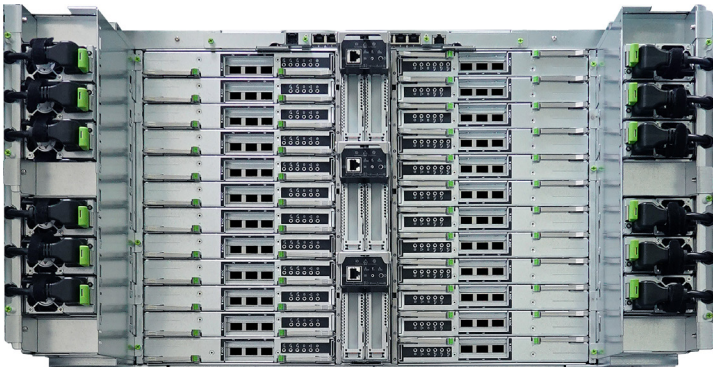


図 2 PRIMEHPC FX1000 本体装置

システム構成

PRIMEHPC FX1000 は 1 ラックあたり 1.297 ペタフロップスのピーク性能を有します。最大構成は 1,024 ラックでピーク性能は 1.328 エクサフロップスになります。

表 1 PRIMEHPC FX1000 システム諸元

		1 ラック構成	最大構成
ラック数		1	1,024
本体装置数		8	8,192
ノード数		384	393,216
ピーク性能	倍精度	1.297 Pflops	1.328 Eflops
	単精度	2.595 Pflops	2.657 Eflops
	半精度	5.190 Pflops	5.315 Eflops
メモリ容量		12 TiB	12 PiB
メモリ帯域		393 TB/s	402 PB/s
インターコネクト帯域		31 TB/s	32 PB/s
PCIe 拡張スロット数		48	49,152
接続トポロジ		2x4x4x2x3x2	32x32x32x2x3x2
		2x2x8x2x3x2	

HPC、AI 領域向け Arm プロセッサ A64FX

A64FX の概要

A64FX の CPU チップは 7nm プロセステクノロジーで製造され、約 90 億個のトランジスタを実装しています。図 3 に CPU チップの写真を示します。

CPU チップあたり計算コア 48 個、アシスタントコア 4 個の合計 52 コアを搭載し、計算コア 48 個の合計ピーク性能は倍精度浮動小数点演算で 3.3792 Tflops です。A64FX の CPU チップは入出力に HBM2 インタフェース 4 組、TofuD インタフェース、PCIe インタフェースを備えます。

Core Memory Group

A64FX のようなメニーコアプロセッサではコア、キャッシュ、メモリ間の接続方法は非常に重要な課題です。A64FX では Core Memory Group (以降 CMG と表記) という 4 つのグループに分割する方式を採用しています。1 つの CMG は、12 個の計算コア、1 個のアシスタントコア、2 次キャッシュ、メモリコントローラで構成されます。4 つの CMG 間はキャッシュ一貫性が維持されており、システムソフトウェアは CMG を NUMA ノードとして扱うことができます。

CPU メモリ異種統合

A64FX は 2.5 次元パッケージ技術により CPU チップと 3 次元積層メモリ HBM2 を単一パッケージに異種統合し、1,024 GB/s の高い理論帯域を確保しています。A64FX の CPU メモリ異種統合の詳細は 6 ページで説明します。

Tofu インターコネクト D 内蔵

A64FX は超並列システムを実現する富士通独自のインターコネクト TofuD を CPU に内蔵しています。TofuD の詳細は 7-8 ページで説明します。

I/O 接続

A64FX は I/O 接続として、標準インタフェース PCIe Gen3 を 16 レーン備えています。

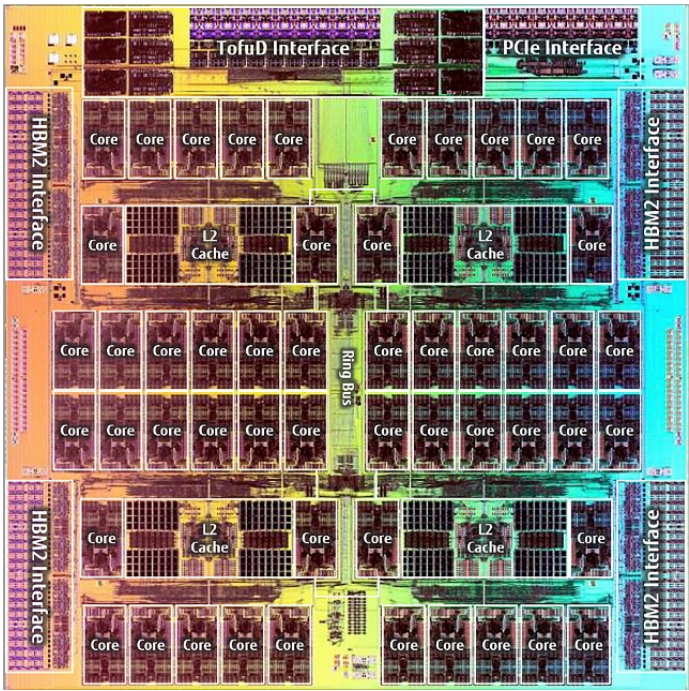


図 3 A64FX CPU チップの写真

Scalable Vector Extension

A64FX は Armv8-A 命令セットのベクトル拡張 SVE を実装しています。SIMD ビット幅は SPARC64™XIfx の 2 倍、512 ビットに拡張し、さらに AI で使用される半精度浮動小数点数 (FP16) に対応しました。SVE の詳細は 4-5 ページで説明します。

アシスタントコア

アシスタントコアは主に OS、I/O 処理等を行い、計算コアにおける OS ジッタを軽減します。集団通信の遅延原因である OS ジッタを軽減することで、並列処理の効率が向上します。

表 2 A64FX 諸元

コア数	計算コア アシスタントコア	48 4
ピーク性能	倍精度 単精度 半精度	3.3792 Tflops 6.7584 Tflops 13.5168 Tflops
L2 キャッシュ容量		32 MiB
メモリ容量		32 GiB
メモリ理論帯域		1,024 GB/s
インターコネクト理論帯域		68 GB/s x2 (in/out)
I/O 理論帯域		15.75 GB/s x2 (in/out)
プロセステクノロジー		7nm CMOS FinFET
トランジスタ数		約 90 億個

Arm 命令セットのベクトル拡張 Scalable Vector Extension

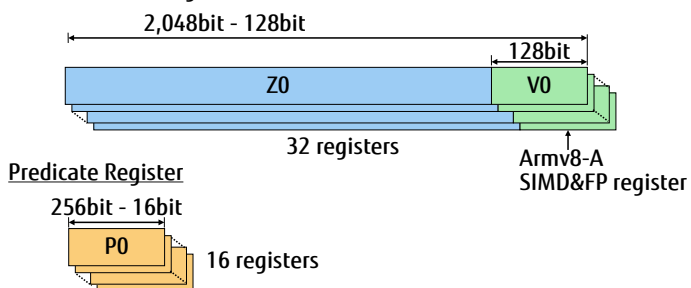
SVE の概要

SVE は Armv8-A 命令セットにスケーラブルなベクトル演算を拡張します。富士通は Arm 社のリードパートナーとして SVE の仕様策定に取り組みました。従来の Armv8-A がサポートするベクトル演算の SIMD ビット幅は 128 と定められています。これに対し SVE の SIMD ビット幅は 128 から 2,048 の範囲で、ハードウェア実装に依存します。A64FX はビット幅 512、256、128 の動作モードをサポートしています。SVE のデータ型は一般的な倍精度浮動小数点数、単精度浮動小数点数に加え、深層学習を加速する半精度浮動小数点数 (FP16) をサポートします。また 16、8 ビット整数のベクトル演算にも対応し、内積演算命令により深層学習の推論を高速化します。SVE の命令種は SPARC64 と同様に 4 オペランド浮動小数点積和演算、Gather/Scatter 命令、数学関数補助命令、Predicate 操作の命令種をサポートします。さらに新規導入の First Fault Load 命令により、事前にループ回数がわからないループの SIMD 化が可能です。

SVE のレジスタ構成

SVE では 32 本の Scalable Vector Register が使用できます。Scalable Vector Register の下位 128bit は Armv8-A SIMD & FP Register と共有しています。また、Scalable Vector Register とは別に 16 本の Predicate Register を使用できます。Predicate Register のビット幅は Scalable Vector Register のビット幅の 8 分の 1 です。レジスタイメージを図 4 に示します。

Scalable Vector Register

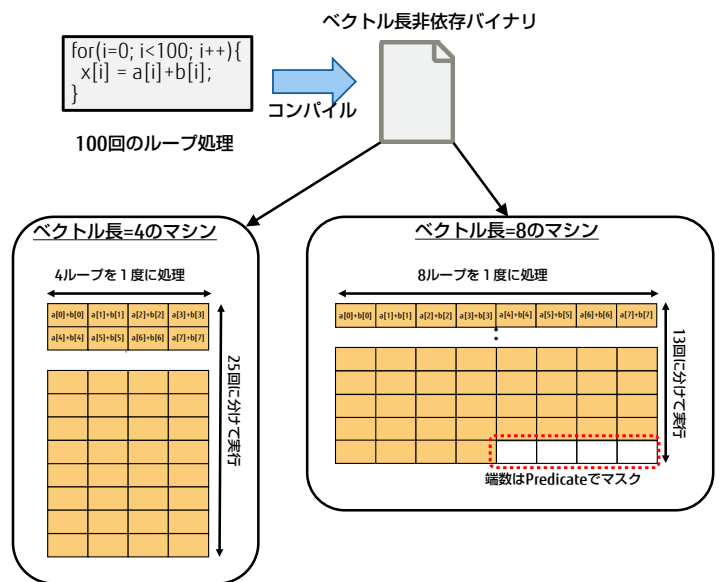


ベクトル長非依存バイナリ (Vector Length Agnostic)

SVE では、ハードウェアが実装する SIMD ビット幅に依存せずに動作可能なバイナリを作成できます。このようなバイナリをベクトル長非依存バイナリと呼びます。ベクトル長非依存バイナリは実装

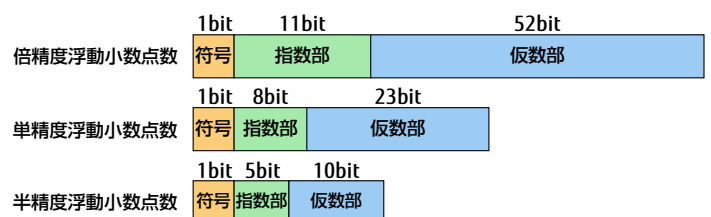
された SIMD ビット幅の異なる SVE マシンにおいても、再コンパイルなしに実行可能です。

図 5 のように 100 回のループを実行するプログラムをベクトル長非依存バイナリにコンパイルすると、マシンの SIMD ビット幅から 1 命令で演算する要素数 (ベクトル長) を計算し、ループ回数を調整するコードが生成されます。このコードはベクトル長 = 4 のマシンでは 25 回、ベクトル長 = 8 のマシンでは 13 回ループします。元のループ回数がベクトル長の倍数でない場合、端数の要素は Predicate 操作でマスクされます。



データ形式と命令種

SVE では、ほとんどの浮動小数点演算命令において倍精度、単精度、半精度浮動小数点数を用いた演算が可能です。必要な精度に合わせてデータ形式を変更しても同じ命令種で演算できるため、演算スループットを予測しやすくなっています。図 6 にサポートする浮動小数点数の形式を示します。



さらに、SIMD 整数演算も可能です。64 ビット、32 ビットに加え 16 ビット、8 ビットの演算も可能で、特に 16 ビットと 8 ビットは推論処理に効果的な内積演算をサポートしています。

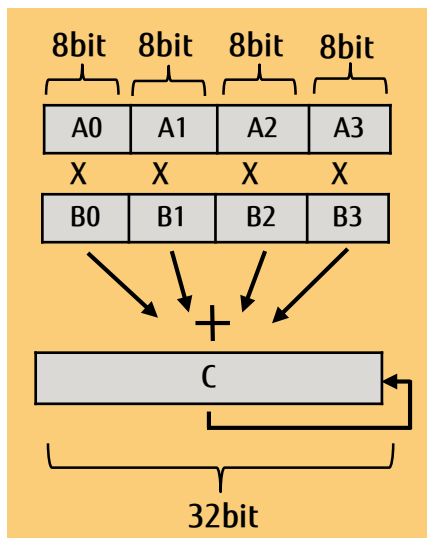


図 7 8 ビット整数内積演算

SVE の数学関数補助命令は SPARC64 と同様に三角関数、指数関数をサポートします。

do-while ループ/break 構文の SIMD 化

SVE では新たに First Fault Load 命令と FFR レジスタを導入し、ループ内処理でループを抜ける do-while ループ/break 構文の SIMD 化が容易になりました。

通常、ループ構造を SIMD 化する場合、プログラムで書かれた複数回分のループ内処理をまとめて SIMD 演算で処理します。しかしループ内処理に依存してループ回数が決まる場合、無理に SIMD 化すると図 8 に示すようにプログラムが使用するデータ領域を超えてメモリにアクセスし、Fault が発生してアプリケーションが強制終了する場合があります。

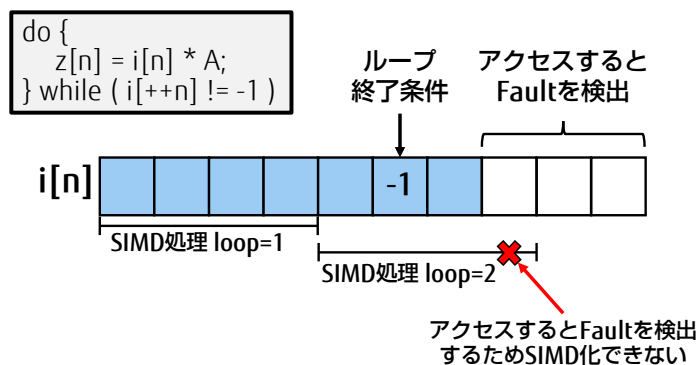


図 8 do-while ループの SIMD 化の難しさ

SVE で導入された First Fault Load 命令を用いると、この問題に対処可能です。First Fault Load 命令は、メモリアクセスで Fault を検出した際に、検出した要素により異なる動作をします。もし SIMD 演算の先頭の要素で Fault を検出した場合は実際に Fault を発生させます。それ以外の要素で Fault を検出したケースでは、Fault の検出情報を FFR レジスタに記録し、Fault が検出した要素以降のメモリアクセスを抑制します。FFR レジスタに記録された情報は Predicate 操作のマスクに使用することが可能です。

図 9 に First Fault Load 命令を用いた SIMD 化例を示します。First Fault Load 命令が記録した FFR レジスタの情報を読み出して Predicate 操作に使うことで、ループを正しく処理できます。

```
ptrue p0.d
ld1rd z1.d, p0/z, [x2]

.loop:
setffr
ldff1d z0.d, p0/z, [x1, x3, lsl #3]
rdffr p1.b, p0/z
cmpeq p2.d, p1/z, z0.d, #-1
brkbs p2.b, p1/z, p2.b
mul z0.d, p2/M, z0.d, z1.d
st1d z0.d, p2, [x0, x3, lsl #3]
incp x3, p2.d
b.last .loop
```

図 9 First Fault Load を用いた SIMD 化コード例

Gather/Scatter 命令

HPC アプリケーションでは、整数配列に格納したインデックスを使用してほかの配列を間接参照することで、非連続なデータの読み書きをすることがあります。SVE ではこのアクセスを実現するための Gather/Scatter 命令をサポートしています。以下の Gather 命令の例では、整数レジスタ X0 に参照する配列の先頭アドレス、Scalable Vector Register の Z1 に参照する配列のインデックスが入っており、Predicate Register の P0 にロードのマスクが入っています。この時に、P0 が 1 の要素のみ間接参照するデータをレジスタに集めて格納することが可能です。

Gather命令 : LD1D Z0.D, P0/Z, [X0, Z1.D, LSL #3]

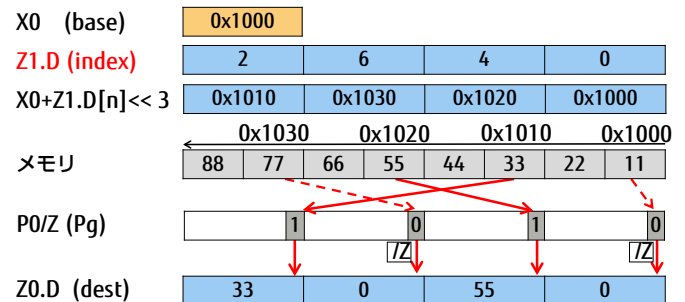


図 10 Gather 命令動作

2.5 次元パッケージ技術による CPU と 3 次元積層メモリ HBM2 の異種統合

CPU メモリ異種統合

PRIMEHPC FX100 では CPU チップを単独で IC パッケージに実装し、3 次元積層メモリは別の IC パッケージに実装されていました。CPU パッケージと 8 つの 3 次元積層メモリパッケージはプリント基板の配線で接続されていました。

これに対して PRIMEHPC FX1000 の A64FX では 2.5 次元パッケージ技術を使用し、CPU チップと 3 次元積層メモリを単一のパッケージに異種統合しました。図 11 に A64FX の 2.5 次元パッケージの写真を示します。CPU チップと 4 スタックの HBM2 は近接配置され、微細配線で高密度に接続されます。

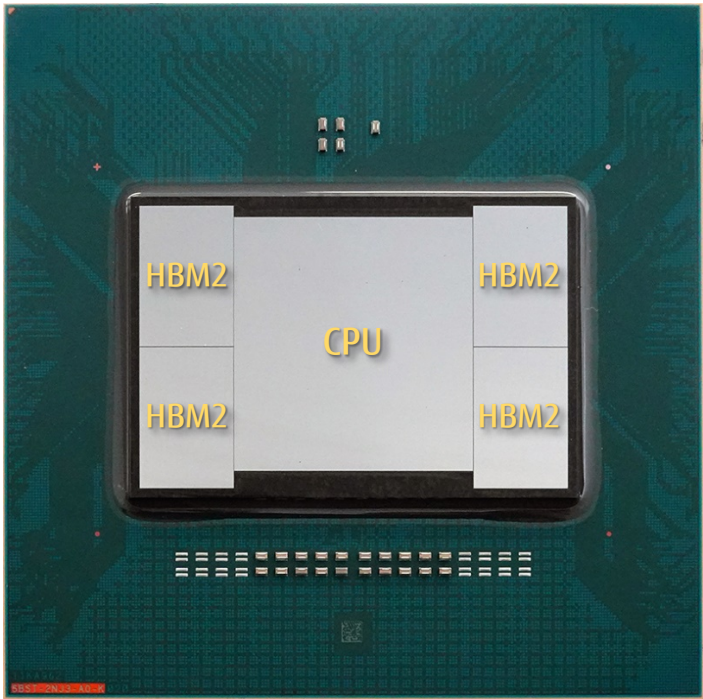


図 11 A64FX 2.5 次元パッケージの写真

2.5 次元パッケージ技術

A64FX の CPU チップおよび 4 スタックの HBM2 はシリコンインターポーザ（以降 Si-IP と表記）上に搭載されています。

図 12 に 2.5 次元パッケージの構成を断面模式図で示します。Si-IP は配線層とシリコン貫通ビア（以降 TSV と表記）が形成されたシリコン基板で、トランジスタは形成されていません。Si-IP は Cu ピラーや C4 バンプなどの接続技術でパッケージ基板に接合されます。CPU チップおよび HBM2 は 40～55μm ピッチのマイクロバンプで Si-IP の配線層に接続します。パッケージ基板は Ball

Grid Array（以降 BGA と表記）と呼ばれる 1mm ピッチの半田ボール端子でプリント基板に接続します。マイクロバンプと BGA のピッチを比較すると、Si-IP 配線はプリント基板配線より約 20 倍高密度と言えます。

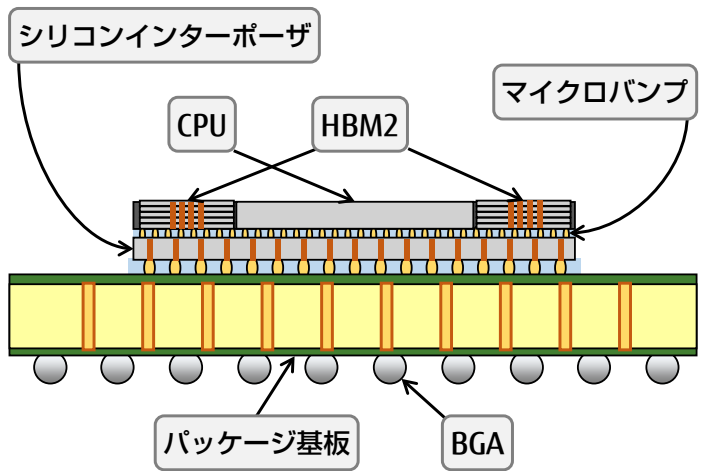


図 12 2.5 次元パッケージの構成（断面模式図）

3 次元積層メモリ HBM2

A64FX が使用する HBM2 は 2.5 次元実装向けの 3 次元積層メモリです。HBM2 は TSV が形成されたメモリダイを最大 8 枚積層し、独立した 128 ビット幅のチャネルを 8 つ搭載します。

表 3 に HBM2 を 4 スタック使用する A64FX 主記憶の諸元を示します。HBM2 のデータ信号伝送速度は 2.0 Gbps で、128 ビット幅のチャネルあたりメモリ帯域は 32 GB/s です。1 スタックのメモリ帯域は 8 チャネル合計で 256 GB/s、メモリ容量は 8 GiB です。HBM2 を 4 スタック使用する A64FX のメモリ帯域は 1,024 GB/s、メモリ容量は 32 GiB となります。

表 3 A64FX 主記憶諸元

メモリ帯域		1,024 GB/s
メモリ容量		32 GiB
パッケージあたり HBM2 数		4
HBM2	データ信号伝送速度	2.0 Gbps
	データ幅	1,024 bit
	メモリ帯域	256 GB/s
	メモリ容量	8 GiB

高密度システム向け Tofu インターコネクト D

TofuD の概要

TofuD は A64FX に内蔵された、10 万ノードを超える超並列システムを構築するインターコネクトです。PRIMEHPC FX100 の Tofu インターコネクト 2（以降 Tofu2 と表記）をベースに高密度システム向け機能と耐故障性を向上させました。

6 次元メッシュ／トーラス・ネットワーク

TofuD は 6 次元メッシュ／トーラス・ネットワークで 10 万ノードを超える超並列システムを構成します。このネットワークはスーパーコンピュータ「京」向けに開発され、Tofu2、TofuD でも引き続き使用されています。図 13 は 6 次元メッシュ・トーラスのトポロジーモデルです。6 次元のうち X, Y, Z 軸はシステム構成により長さが可変です。残りの A, B, C 軸は長さが 2, 3, 2 に固定されています。6 次元での相互接続のため、各ノードは 10 個の接続ポートを備えます。

ユーザービュアのネットワークトポロジーは、1 次元／2 次元／3 次元の仮想的なトーラスです。ユーザーが指定した次元数、大きさの仮想トーラス空間は 6 次元メッシュ／トーラス・ネットワーク上にマップされ、ランク番号に反映されます。この仮想トーラス方式により故障ノードを含む領域をトーラスとして利用できるように、システムの耐故障性、可用性が向上します。

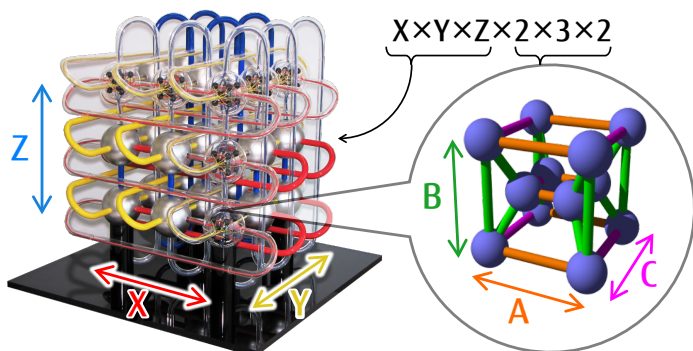


図 13 6 次元メッシュ／トーラスのトポロジーモデル

高密度システム構成

TofuD では各ノードが 2 レーンのリンク 10 本で相互接続し、同じ装置内のノード間は電気伝送、異なる装置のノード間は光伝送のリンクを使用します。ここで 1 つの装置に収容されるノード数が増えるほど、コストの低い電気伝送の比率が高くなります。PRIMEHPC FX1000 は本体装置 4 台に搭載した最大 192 ノード間

を電気伝送で接続する高密度システムとなっています。CMU 上の 2 ノードは C 軸、本体装置内の 48 ノードは Z, A, B 軸、4 台の本体装置間は X, Y 軸で相互接続します。CMU の 6 次元構成は $(X, Y, Z, A, B, C) = (1, 1, 1, 1, 1, 2)$ 、本体装置は $(1, 1, 4, 2, 3, 2)$ 、本体装置 4 台は $(2, 2, 4, 2, 3, 2)$ となります。光伝送を使用する相互接続は、全ノードの半数では X, Y 軸、残り半数では X, Y, Z 軸です。図 14 に CMU にアクティブ光ケーブル（以降 AOC と表記）を接続する様子を示します。TofuD の 6 次元ネットワーク構成に必要な 4 レーン AOC の本数は、1 ノードあたり 0.625 本です。

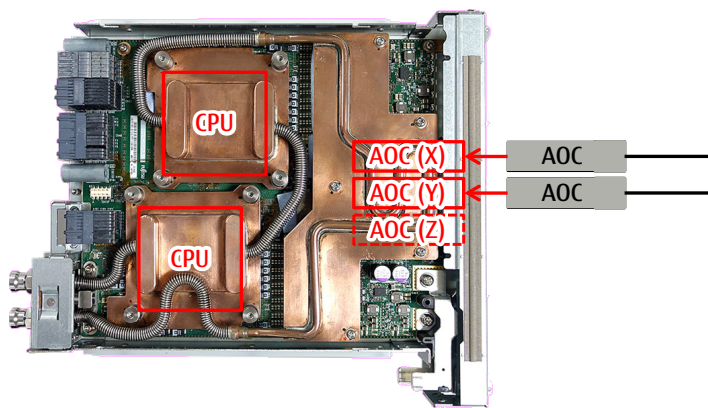


図 14 CMU への AOC 接続

6 基の RDMA エンジン

トーラスは隣接ノード間の通信同士が干渉しないという強い局所性を有するネットワークです。トーラス・ネットワークの局所性は、複数 RDMA エンジンの同時通信により利用できます。

「京」および Tofu2 は 4 基の RDMA エンジンを搭載し、専用通信ライブラリはネットワークの局所性を利用して集団通信を高速に実行しました。TofuD では CPU メモリ異種統合によるメモリ帯域の大幅向上に合わせて RDMA エンジン を 6 基に増強し、10 万ノードを超える「富岳」でも集団通信を高速に実行します。

Tofu バリア増強

Tofu バリアは低遅延にバリア同期および AllReduce 集団通信を実行するハードウェアで、RDMA エンジンに同居して実装されています。「京」および Tofu2 では Tofu バリアを使うプロセスは 1 ノードあたり 1 つという前提で、RDMA エンジン 1 基にだけ Tofu バリアを実装しました。しかし A64FX では CMG が 4 つに増えてノード内のプロセス数も 4 以上となり、ノード内の複数プロセス

が異なるコミュニケータで集団通信する機会が増加します。そこでTofuDではTofuバリア通信資源を増強するために6基のRDMAエンジンすべてTofuバリアを実装し、さらにRDMAエンジン1基あたりのTofuバリアチャンネルをコミュニケータ8個分から16個分に増やしました。

また、バリア同期と同時に実行可能なAllReduce演算も強化しました。一度に縮約できる要素数を浮動小数点数は1から3、整数は1から8に増やしました。さらに、4要素のMAXLOC演算を新たにサポートしました。

動的パケット分割

超並列計算機は膨大な数の部品で構成されるため、障害、故障があってもシステム運用を継続する耐故障性が重要です。Tofu2はリンクの特定のレーンに故障を検出すると、リンクのレーン数を縮退して通信を継続する耐故障性機能を有していました。

TofuDではさらに進んだ障害回復機能として、動的パケット分割技術を開発しました。図15、図16に動的パケットの動作を示します。通常動作である分割モードでは、送出側はパケットをスライスに分割し、2つのレーンで同時転送します。ここで受信側は各レーン独立で誤り検出を行い、誤り頻度を送出側に通知します。送出側は誤り頻度が高い場合は障害が発生しているとみなし、パケットを分割せずに2つのレーンに送出する複製モードに移行します。動的パケット分割の複製モードはTofu2のレーン縮退と同様に有効な帯域が半減しますが、各レーンの誤り検出が継続する点が異なります。通知される誤り頻度が低下した場合、送出側は動的パケット分割のモードを複製から分割に戻します。

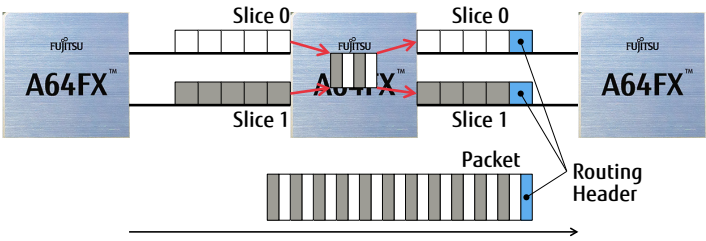


図 15 動的パケット分割：分割モード

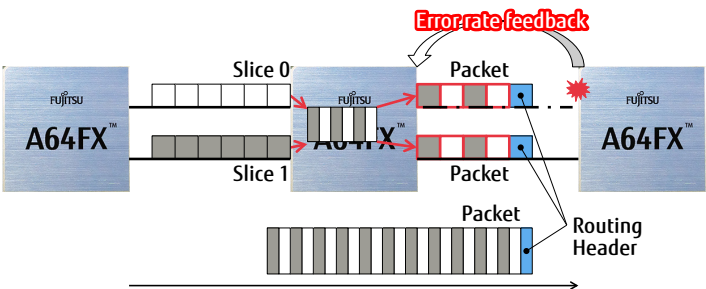


図 16 動的パケット分割：複製モード

低遅延通信

TofuDのRDMA Put通信遅延は0.49μs（最短）です。これはTofu2と比べると0.22μs、「京」と比べると0.42μs短くなっています。遅延の内訳を図17に示します。遅延が短縮した理由は、Tofu2ではキャッシュインジェクション導入とCPU内蔵によるバス削除で、TofuDでは物理コーディング層でのレーン間位相差調整が不要になったことです。1ホップ遅延は約80nsです。

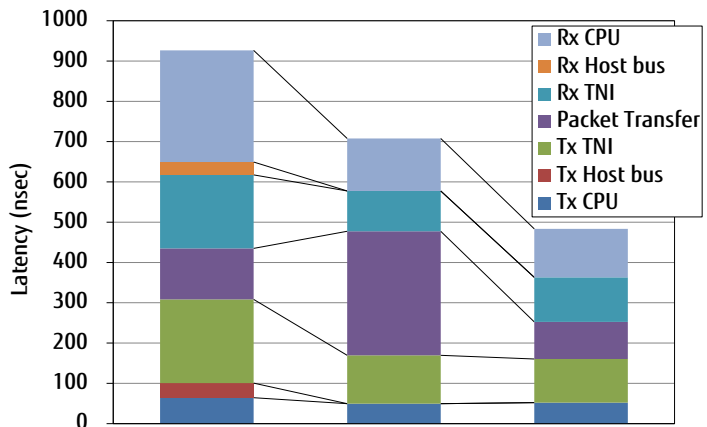


図 17 通信遅延の内訳

表 4 Tofu インターコネクト D 諸元	
データ転送レート	28.05 Gbps
エンコーディング	64b/66b
リンクあたりレーン数	2
リンク帯域	6.8 GB/s
インジェクション帯域	40.8 GB/s
ノードあたり接続ポート数	10
ネットワークポロジ	6次元メッシュ／トーラス
ルーティング方式	拡張次元オーダー
仮想チャンネル数	4
最大パケット長	1,984 バイト
パケット転送方式	バーチャル・カッスルー
フローコントロール方式	クレジットベース
送達保障方式	リンクレベル再送信
RDMA 通信機能	Put/Get/Atomic RMW
RDMA エンジン数	6 (同時通信可能)
RDMA エンジンあたり CQ 数	12 組
アドレス変換方式	Memory Region + Page Table
Tofu バリアチャンネル数	96
通信保護方式	グローバルプロセス ID
動作周波数	425 MHz

参考情報

PRIMEHPC FX1000 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。

<https://www.fujitsu.com/jp/products/computing/servers/supercomputer/index.html>

FUJITSU Supercomputer PRIMEHPC FX1000 AI・エクサスケール時代を切り拓く HPC システム
富士通株式会社
2019 年 11 月 12 日 第 1 版
2019-11-12-JP

- ・ ARM, ARM ロゴは ARM Ltd またはその関連会社の商標または登録商標です。
- ・ 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、®) 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。