富士通 PRIMEHPC FX10 チューニング連載講座
1. ハードウェア概要

大 島 聡 史
東京大学情報基盤センター 助教

1　はじめに

本稿では富士通 PRIMEHPC FX10（以下FX10）のハードウェア概要について解説します。プログラムを最適化するためには対象ハードウェアについて理解することが非常に重要です。またFX10は既設のHITACHI SR16000/M1(以下SR16000)とは大きく構成や特徴が異なります。システム間の違いも含めて正しく理解したうえで最適化プログラミングに挑んでください。

2　全体構成

図1にFX10の全体構成図を示します。また表1にはFX10の性能諸元を示します。FX10は4800の計算ノード、3.2PBのストレージ(ローカルファイルシステム1.1PB+共有ファイルシステム2.1PB)、高速な計算ノード間ネットワーク（Tofu)、そしてログインノードや各種管理用ノードから構成される計算機システムです。

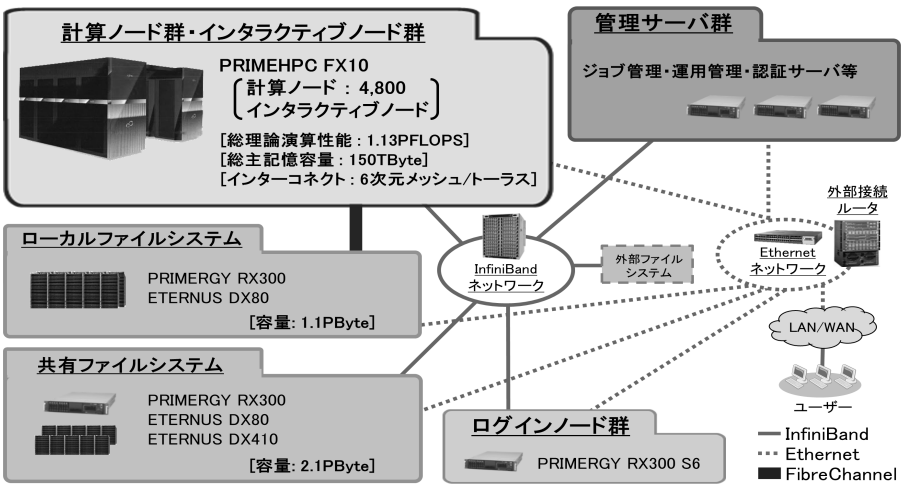

図 1　FX10 の全体構成図

3　CPU構成

FX10は計算ノードのCPUとして富士通が開発したSPARC64IXfxを搭載しています。CPUアーキテクチャは本センターの既存のシステムとは異なるSPARC64系(SPARC64V9+HPC-ACE)です。プログラムの作成方法・最適化方法については本連載中に別の記事で解説しますが、既存のx86系やPower系とは異なる最適化手法やパラメタが重要となることが考えられます。

表 1　FX10 の性能諸元

	FX10	SR16000（参考）
CPU	SPARC64IXfx 1.848 GHz	Power7 3.83 GHz
ノード数	4800	56
コア数/計算ノード	16	32
理論演算性能/1 コア	14.784 GFLOPS	30.64 GFLOPS
理論演算性能/1 計算ノード	236.5 GFLOPS	980.48 GFLOPS
理論演算性能/全計算ノード	1.13 PFLOPS	54906.88 GFLOPS
主記憶容量/1 計算ノード	32 GByte	200 GByte
主記憶容量/全計算ノード	150 TByte	11.2 TByte
物理転送性能/1 計算ノード	85 GByte/sec	512 GByte/sec
Byte/FLOPS 値	0.36	0.52
SMT 機能	-	最大 4 スレッド/1 コア
計算ノード間 ネットワーク構成	6 次元メッシュ/トーラス	階層型完全結合
計算ノード間転送性能	20GB/s（単方向）× 双方向 （4 方向同時通信可能）	96GB/s（単方向）× 双方向
ストレージ容量	1.1 PByte + 2.1 PByte	556 TByte

SPARC64IXfxと計算ノードの構成を図2に示します。SPARC64IXfxは16つのコアによって構成されているマルチコアプロセッサであり、1ノードには本CPUが1基搭載されています。ラックへの搭載については、4CPUおよびメインメモリとICC(インターコネクトコントローラ)からなる「システムボード」を1単位として行われています。

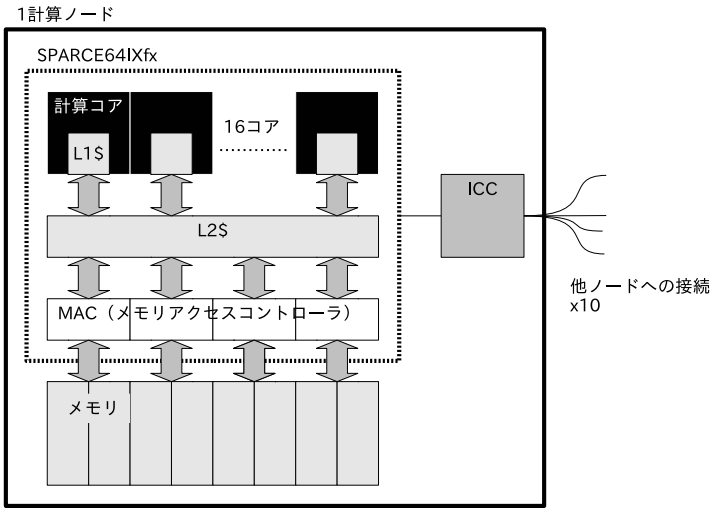

図 2　SPARC64IXfx と計算ノードの構成

FX10に搭載されているSPARC64IXfxの動作周波数は1.848GHzであり、理論倍精度浮動小数

点演算性能は

- 1コアあたり 14.784GFLOPS (1.848GHz×8回=14.784FLOPS)
- 1CPUあたり 14.784GFLOPS×16コア=236.5GFLOPS
- 計算ノード群全体 236.5GFLOPS×4800ノード=1.13PFLOPS

です。本センターの計算機システムとして初めて1PFLOPSを超えるシステムとなります。なおノードあたりのFLOPS値を見てみると、HA8000(T2K東大版)と比べると2倍近いですが、SR16000と比べると4分の1程度しかありません。そのため、FX10の高い演算性能を有効に活用するには多数のノードを適切に活用せねばなりません。

キャッシュについては、L1キャッシュをデータと命令それぞれにコアごとに32KB、L2キャッシュを1CPUごとに12MB搭載しています。既に運用しているHA8000やSR16000と異なりL3キャッシュは搭載されていませんが、L2キャッシュのサイズはHA8000より大きく、また再利用性のあるデータを選択的にキャッシュに残すセクターキャッシュ機能があります。

FX10も既存の他のシステムと同様に、一般ユーザはログインノードを介して計算ノードを利用します。本システムのログインノードはCPUとしてIntel社のXeonプロセッサを搭載しており、ログインノードがx86系アーキテクチャ(x86_64アーキテクチャ)なのに対して計算ノード（およびインタラクティブノード）ではSPARC64系アーキテクチャとなります。そのため、プログラム開発の際にはクロスコンパイラの使用が必須となります。

CPUの演算性能に関するベンチマークの結果などは今後のスパコンニュースの記事にて紹介予定です。

4　メモリ構成

SPARC64IXfx CPUとメインメモリ（主記憶）の接続については、図2中に示すように計算ノード上に搭載されているメモリアクセスコントローラ(MAC)を介して行います。FX10の計算ノードには1ノードあたり32GByte、計算ノード群全体では150TBのメインメモリが搭載されています。メインメモリの種別についてはDDR3 SDRAMメモリ(DDR3-1333)が搭載されています。メモリ帯域幅は85GByte/sec(8Byte×1333MHz×8Channel)、B/F値については0.36となります。B/F値を他のシステムと比較してみると、SR11000/J2(1.39)やSR16000(0.52)には及びませんが、HA8000(0.29, 0.17)と比べると向上しており、大規模計算における良好なスケーラビリティが期待できます。

なお、FX10にはHA8000のように大容量のメモリを搭載したノードが用意されていません。そのため、大規模なプリ・ポスト処理など大容量のメモリを必要とする計算を行いたい場合にはHA8000の128GBノードを利用していただくことになります。（本件の詳細については後日改めてご案内する予定です。）

メモリの性能を測定するSTREAMベンチマーク[*1]の結果としては、計算ノード1ノードあたり60GByte/秒の値が得られています。

[*1] http://www.cs.virginia.edu/stream/

5　ノード間ネットワーク構成

FX10のノード間ネットワーク構成の概要を図3に示します。FX10のネットワーク構成は6次元メッシュ/トーラスネットワーク（Tofuネットワーク）という独自のネットワーク構成となっています。

実際にネットワークを制御しているのは各計算ノードに搭載されているICCであり、各ICCから同一計算ノード上のCPUや他計算ノード上のICCへ接続しています。各ノードは6次元座標を持ち(X,Y,Z,a,b,c)、x,y,z全てが一致する12計算ノードをTofu単位と呼びます[*2]。1Tofu単位内では各ノードが4本ずつのリンク(a,b+,b-,c)によって接続されており、Tofu単位間は3次元(X+,X-,Y+,Y-,Z+,Z-)に接続されています。これらの構成により、低レイテンシ、広帯域、高信頼性を達成しています。

理論性能については、各ICCと同一ノード上のCPUとの接続については20GByte/秒（×双方向）、ICC間の接続については5GByte/秒（×双方向）です。ノード間の通信性能については、隣接ノード間の物理転送遅延が1マイクロ秒、ノード間の理論通信速度が20GByte/秒（×双方向、4方向同時通信可能）、バイセクションバンド幅が6TByte/秒となっています。

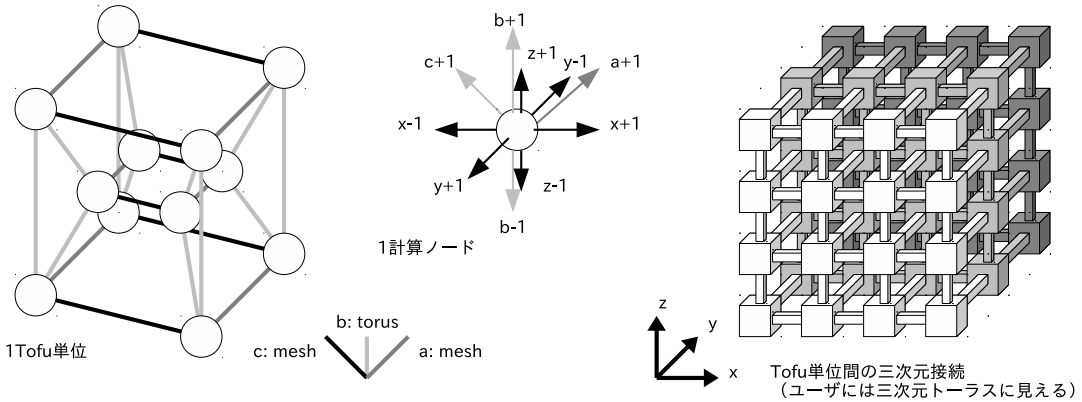

図 3　計算ノード間ネットワーク構成（Tofu ネットワーク）

6　ストレージ構成

FX10は容量1.1PBのローカルストレージと容量2.1PBの共有ストレージの2種類のストレージを備えています。ローカルストレージはETERNUS DX80 S2とPRIMERGY RX300 S6を用いて構築されており、共有ストレージはETERNUS DX410 S2、ETERNUS DX80 S2そしてPRIMERGY RX300 S6を用いて構成されています。いずれも計算ノード群とFibreChannelやInfiniBandによって高速に接続されており、RAIDや機器の多重化による冗長構成をとっています。ファイルシステムについてはLustreをベースとした富士通製のファイルシステムFEFS(Fujitsu Exabyte File System)を採用しています。想定されている用途としては、共有ストレージについてはいわゆるストレージとしてユーザ毎のデータを保持するために、一方のローカル

[*2] 本システムの利用において 2 のべき乗ではなく 12（もしくは 12 の倍数）を単位とした設定や制限があるのは Tofu ネットワークが 12 計算ノードを単位として構成されていることによるものが多いです。

ストレージについてはステージング用として使用されます。

以上、本稿ではFX10のハードウェア概要について、一部ベンチマークの結果等を交えて紹介しました。FX10に関するさらに詳しい情報やベンチマークの結果等については、本連載における他の記事や、当センターのwebサイトに掲載されている資料（http://www.cc.u-tokyo.ac.jp/system/fx10/、今後も必要に応じて情報を追加していく予定です）もご覧ください。