T2K オープンスパコン仕様と富士通の T2K 仕様準拠の計算ノード「Fujitsu HX600」

久門耕一

Ⅰ．始めに

本稿では，筑波大学，東京大学，京都大学（T2K）が共同で仕様を制定した T2K オープンスパコンの概要と，T2K オープンスパコン仕様に基づき富士通株式会社が開発した Fujitsu HX600 の特徴についてご紹介します。

Ⅱ．T2K オープンスパコンの技術的な背景

80 年代の後半に生まれた PC クラスタシステムは，ベースとなる PC 用 CPU の性能が急速に向上したことに伴い，高いコストパフォーマンスを武器に急速に普及してきました。またクラスタのノードを構成する PC 間を接続するネットワークも，デファクトスタンダードであるイーサネットが 100Mbps から 1Gbps，将来的には，10Gbps と高速化するにしたがい，コンピュータ間を高速に接続する専用インタコネクトに近づく性能となってきました。しかしながら，高性能科学技術計算分野では，PC 用 CPU の理論的なピーク性能と実際に得られる性能の差が目立つようになっています。

現在主流となっている PC 用 CPU アーキテクチャの一つである Intel 社の Core2 アーキテクチャの場合，科学計算で重視される浮動小数演算能力は，1 クロックあたり最大 4 演算です。この結果，3GHz の CPU では 1 コアあたり理論演算性能は 12GFlops であり，4 コア構成の CPU を 2CPU 搭載するノードの場合，ノード当りの理論演算性能は 96GFlops に達します。

Core2 アーキテクチャにおいては，CPU は，チップセットを介してメモリとデータのやり取りを行いますが，CPU にデータを供給するメモリバンド幅は，8Byte 幅で周波数が 1666MHz だとしても 13.3GB/s しかありません（図 1）。

Core2 アーキテクチャを使った Xeon CPU の場合，CPU ごとにバスを持つので理論上は CPU 数に比例するバンド幅が得られますが，実測値では，5GB/s−10GB/s のバンド幅しか得られません。これは，Intel 系のアーキテクチャの場合 CPU からメモリまでの間にメモリコントロール用のチップセットが存在するため，データの伝送遅延が大きい事と，キャッシュ一貫性を取るコストが高いことが原因です。このような問題で性能低下が起きることがないように，Intel 系の CPU では大き目のサイズのキャッシュを搭載していますが，科学技術計算においてはキャッシュによるデータ再利用には限界があるため，メモリバンド幅が性能を律速する場合が多くなっています。

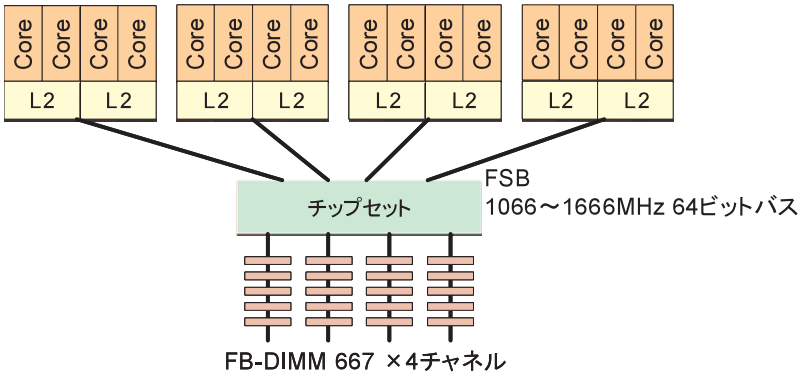

図 1　Intel Core2 アーキテクチャによる 4CPU システムメモリ構成例

かつての高性能科学計算向けコンピュータである，ベクター型スーパーコンピュータでは，1 秒当たりの浮動小数演算能力（Floating Operations = FLOP）と，1 秒あたりのデータ供給能力（Byte）の比を取ると，8B/FLOP または 4B/FLOP でした。

一方，前述の Core2 で同様の計算を行うと 0.05B/Flop−0.08B/Flop と 1/100 程度と大幅に小さくなっています。ただし，キャッシュを搭載するこれらの CPU では，キャッシュバンド幅は十分に高く，Xeon で 4Byte/Flop，Opteron では 7Byte/Flop 程度確保されています。つまり，キャッシュを有効活用すれば，十分高い性能が得られるものの，そうでない場合には，急激に性能が落ちるため，性能を得るためのプログラミングが難しいと言えます。

このように，1CPU に搭載される CPU コア数が 2 個，4 個と増加しても，処理性能が CPU へのデータ供給能力で制限されます。その結果，高性能科学技術計算ではコモディティーを使った PC クラスタの実効性能が頭打ちとなるケースが多くなっています。

これに対し，図 2 に示す Opteron では，1 クロック当たり最大 4 演算であることは同じですが，メモリと CPU がチップセットを介さず直接接続され，また，CPU ごとにメモリバンクを保有するため，もしプログラムのメモリアクセス局所性をメモリ配置と一致させることが可能なら，シ

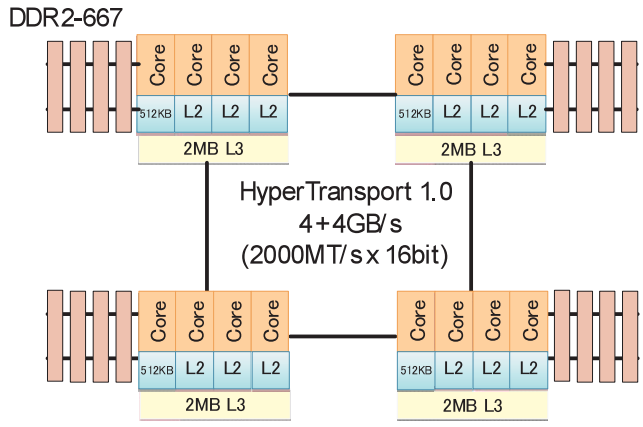

図 2　AMD の Opteron による 4CPU システム構成例

ステムのメモリバンド幅は，メモリのチャネル数に比例して高められる可能性があります。

クラスタのノード間を接続するインタコネクトネットワークにおいてもデータ供給能力がシステム性能の鍵で，ノード内のコア数の増加による演算性能に見合うだけのインタコネクト性能を持たなければ，ノード内の演算性能向上の恩恵を受けることができません。

PC クラスタにおいて，CPU の持つ高い計算能力を生かすために，CPU へのデータ供給速度を如何に高めるかがシステム性能に大きく影響しています。

T2K オープンスパコン仕様では，CPU のピーク性能だけでなく，少しでも多くのプログラムで高性能が得られるように，CPU の最大演算性能と同時に，表 1 のように高いデータ供給能力を要求しています [1]。

表 1　T2K オープンスパコンの要求仕様概要

● 40GB/s の理論メモリバンド幅
● 5GB/s 以上のノード間理論データ転送能力
● 32GB/ ノード以上の搭載メモリ
●ノード内に 16 コア以上の CPU 数

また，T2K オープンスパコンでは，コモディティーである X86CPU を使っており，他の PC クラスタシステムと開発されたソフトウェアを互いに利用できるようにするため，OS としては Linux が指定されています。

Ⅲ．T2K オープンスパコン仕様に基づく計算ノード HX600

T2K の計算ノードには，ノード当たり X86 ベースのコモディティー CPU を 16 コア搭載する必要があるため，富士通では，AMD 社の最新の 4 コア Opteron（開発コード名 Barcelona）を 4CPU 搭載する HPC クラスタ向けの新サーバ HX600 を新たに開発しました（図 3，図 4）。先に述べたように，計算ノードの開発において，データ供給能力の確保が，設計上の最大の留意点となります。

図 3　HX600 概観

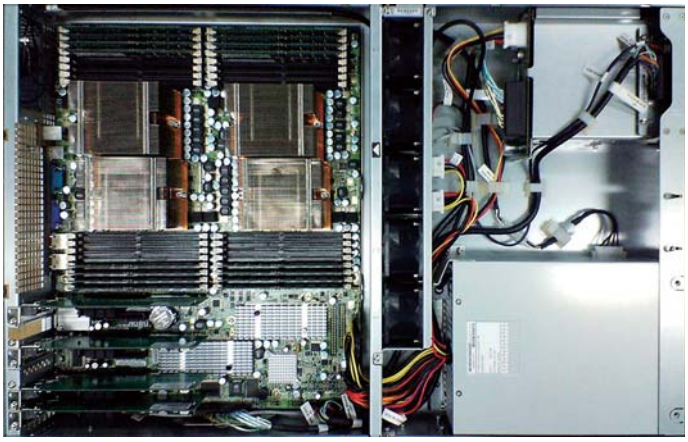

図４　HX600 内部の構造

１．40GB/s 以上のメモリバンド幅の確保

Opteron は CPU とメモリが直接接続され，複数の CPU 間は Hyper Transport と呼ばれる片方向当たり 4GB/s の専用の通信路で接続されたいわゆるダイレクトコネクトアーキテクチャを採用した CPU で，1 つの CPU とメモリの間は 2 チャネルの DDR2 インタフェースで結ばれています。したがって，理論最大バンド幅は，DDR2 インタフェース速度が 667MHz の時 10.3GB/s，4CPU で 41.2GB/s となります。また，実機によるメモリバンド幅計測では，15−20GB/s 程度が得られています。

２．5GB/s 以上のインタコネクトバンド幅の確保

クラスタを構成する上でノード間データを転送するために使われるインタコネクトネットワークは，選択肢としては，2GB/s のデータ転送速度を持つ DDR-InfiniBand（DDR-IB）4 系統，もしくは，1.25GB/s のデータ転送速度を持つ Myrinet−10G4 系統が候補となります。富士通では，従来から高性能 PC クラスタを構築する際に用いてきた経験が豊富にある DDR-InfiniBand（DDR-IB）を採用しました。これにより，ノード間理論最大インタコネクトバンド幅は 8GB/s/片方向となります。1 枚 2GB/s の IB-HCA に十分なデータを供給するには，8 または 16 レーンの PCI-express（PCIe）インタフェースが必要です。

ノードには合計 4 枚の IB-HCA 搭載し，さらに拡張性も考慮し，8 レーンの PCIe インタフェースを 6 スロット設けました。さらに，6 系統の PCIe インタフェースをサポートするため，4 つの Opteron に対し，I/O 接続を行える CPU を 2 つ用意し，それぞれに nVidia 社の MCP55 と IO55 を接続し，6 系統の PCIe の性能に見合う 8GB/s の I/O メモリ間バンド幅を確保しました。

また，CPU のコア電圧と，HyperTransport のためのノースブリッジ電圧を独立に設定できる Split Power Plane を採用し，プロセスが動作していない CPU コアの消費電力を下げつつ，その CPU に接続されたメモリへのアクセス性能を劣化させないようにしました。これにより，例えばノード内でジョブが走っていないときだけではなく，すべてのコアを使っていないとき，例え

表２　HX600 諸元

CPU	プロセッサ	AMD OpteronTM プロセッサ 8300 シリーズ
	クロック	2.3GHz
	コア数	4
	キャッシュ	一次キャッシュ：1 コア当たり，64KB 二次キャッシュ：1 コア当たり，512KB 三次キャッシュ：CPU 当たり，2MB
ノード	CPU 数	4（合計コア数 16）
	演算処理性能	147GFlops
	メモリ容量	標準 32GB，最大 128GB，ECC 付き DDR2
	チップセット	nVidia MCP55/IO55
	ハードディスク	標準 146.8GB×2，最大 300GB×2 RAID1（ハードウェア RAID）
	ノード間インタコネクト	InfiniBand™ DDR（2GB/s）×4
	ネットワークインタフェース	Gigabit Ethernet（1000Base-T）×2
シャーシ	外形寸法（W×D×H）	430（481（突起部含む））×701（774（突起部含む））×87mm ラックマウント 2U
	質量	約 21kg
	電源条件	最大約 780W
ソフトウェア	サポート OS	Red Hat Linux V4.6

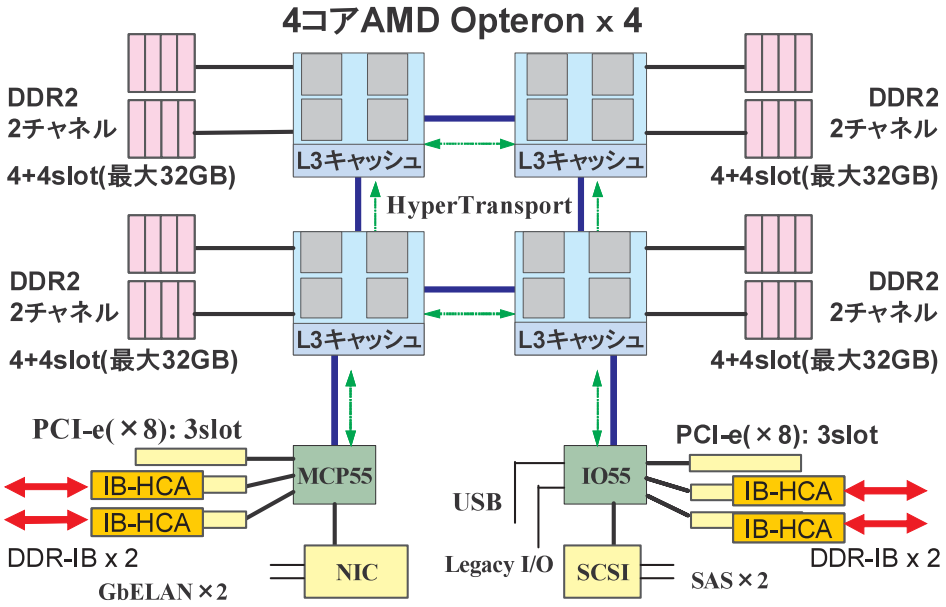

図５　HX600 システムブロック図

ば2ソケット8コアだけを使うようなケースでも，使用されていないCPUのコア周波数を最低に設定し，システムの消費電力を抑えることが可能になりました。このためのコア周波数の制御は，ミドルウェアである富士通Parallelnaviがジョブ投入と連携して行うため，ジョブ実行を行うユーザはこのことを意識する必要はありません。図6では，4つのソケットでそれぞれがマルチスレッドで動作する2ソケットを使用する青，1ソケットを使う黄色と緑で表された3つのプロセスのうち，青のプロセスが終了した結果，2つのソケットのコア電圧，コア周波数を下げて消費電力を抑えている状態を表しています。

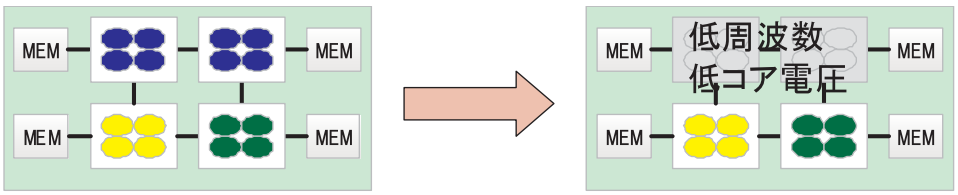

図6　アイドルCPUのソケット単位電圧制御

3．インタコネクトトポロジ

各ノードに4枚搭載されるDDR-IBカードは，全体を一つのネットワークとして構築することも可能ですが，4枚のカードを2つのグループに分け，それぞれを独立なネットワークとして構成することにより，ネットワークコストを低減することも可能です。

富士通のジョブ管理システムであるParallelnavi NQSでは，プロセスの使用メモリとプロセス実行CPU，IB-HCAを一体として管理することにより，ノードに複数のプロセスを割り付ける場合でも高いメモリバンド幅を実現し，ノードに1つあるいは2つのプロセスを割り付ける場合，利用可能なHCAを複数束ねるネットワークトランキングをサポートし，高いノード間データ転送スループットを実現可能としています。

図7において，ノード内に1から4プロセスが配置されたときにプロセス間でハードウェア資源が競合しないように，それぞれのプロセスに対し排他的に最適なInfiniBand HCAが割り付けられている様子を示します。

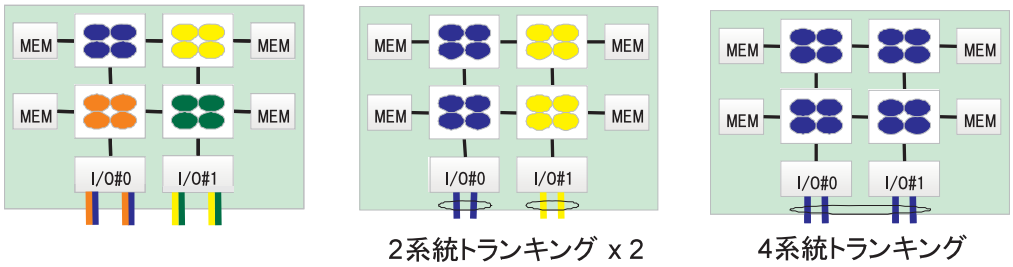

図7　ノードへの複数プロセス割り当てとトランキングの実現

4．ソフトウェア

言語処理系

T2K オープンスパコンでは，1 ノードが 16 コアから構成されており，システム全体で動作させるプログラムの場合，ノード数の 16 倍の並列度を持つプログラムが動作可能です。一方，ノード内には 16 コアあるので，1 ノードを中規模の SMP システムとして考えることもできます。例えば，京都大学のシステムでは全体で 6656 コア（416 ノード *16 コア）が MPI 接続されているとして考えることも，416 台の SMP の MPI 接続クラスタと考えることもできます。

SMP サーバ内では，コンパイラによる自動並列化により，マルチスレッド実行が可能であるため，ユーザはアプリケーションの持つ並列度，実効効率などさまざまな要求に応じ，MPI 並列とスレッド並列を組み合わせることが可能です（図 8）。

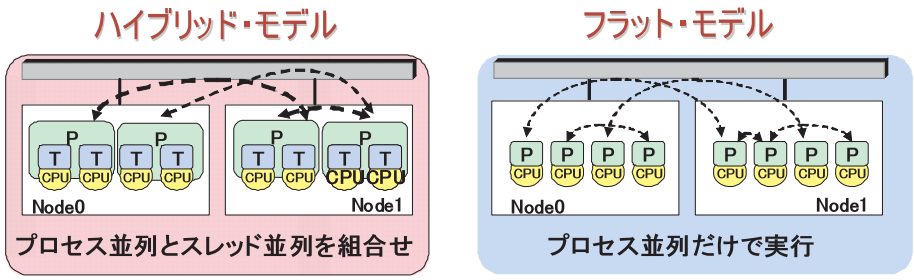

図８　プロセスとスレッドを用いたプログラミングモデル

このようなプログラムを開発する環境として，HX600 ベース T2K オープンスパコン向けに，富士通が従来から提供してきた XPFortran をはじめとする下記の言語処理環境を提供しています。

➢国際規格及び業界標準言語仕様のサポート

- Fortran95 準拠，Fortran2003 規格（一部），Fortran90 規格，Fortran77 規格
- C99，C89，K&R，C ＋＋
- OpenMP API V2.5

➢ VPP/PRIMEPOWER シリーズ等従来システムのお客様の資産の継承と動作互換保証

➢コンパイラによるノード内並列のサポート

- OpenMP による明示的並列化
- コンパイラによる自動並列化

➢性能解析機能

- キャッシュミス等の動特性計測（Opteron CPU 内の性能評価カウンタ利用）

➢高性能ライブラリ

- 4 コア Opteron（Barcelona）に最適化した DGEMM（行列積ルーチン）

特に，科学演算では必須の高性能行列演算に関しては，富士通独自の高度なチューニングを行っ

た行列積ルーチン（DGEMM）を X86 システム用に開発しています。

スーパーコンピュータのランキングである Top500 サイトの 2007 年秋のリストで，富士通の Xeon 用 DGEMM を使った，九州大学の Core2 ベース Xeon（Woodcrest）クラスタシステムが Woodcrest 利用のクラスタ内で最高の Linpack 実効効率（81.9%）を得ています。

また，Opteron 用に最適化された DGEMM により，京都大学学術情報メディアセンタ様に納入し，2008 年 6 月から運用開始の予定である，416 ノードの T2K 仕様 Opteron クラスタシステムでも，Linpack において 82.49% と高い実効効率が得られており，2008 年 6 月に発表された最新の Top500 リスト上で世界 34 位の性能となりました。[4][5]。

2008 年 6 月の Top500 リストにおいて，4 コア Opteron 搭載システムは 13 システム登録されていますが，Linpack 実行効率が 80% を超えるシステムは，京都大学システムを含め 2 システムしかありません。

V．終わりに

本稿では，T2K オープンスパコン仕様に基づく計算ノード Fujitsu HX600 の概要についてご説明しました。今後，CPU がマルチコア時代となり，処理性能が向上するに伴い，処理すべきデータを CPU とメモリあるいは他のノードの間で移動する速度をますます向上させていく必要があり，コモディティベースサーバに対するデータ供給能力の追及は必要不可欠のものになっていくため，スパコン用サーバと汎用サーバの間の隙間を埋める努力が必要になるものと考えます。

参考文献

[1] 中島浩，T2K オープンスパコン設計思想とアーキテクチャ
http://www.pccluster.org/event/symp/2007/Nakashima_pccc.pdf

[2] 佐藤三久，石川裕，他，T2K シンポジウムつくば 2008
http://www.ccs.tsukuba.ac.jp/workshop/t2k-sympo2008/

[3] 富士通株式会社，HPC サーバ「HX600」の概要・仕様
http://pr.fujitsu.com/jp/news/2008/02/15−2a.pdf

[4] Top500 サイト
http://top500.org/list/2007/11/100

[5] Jack, Dongarra, Performance of Various Computers Using Standard Linear Equations Software, http://netlib.org/benchmark/performance.pdf

（くもん　こういち：（株）富士通研究所 IT システム研究所主席研究員）