

次世代スパコン『京』¹について

追永 勇次
富士通株式会社

[アブストラクト]

次世代スパコン「京」について、ハードウェア、ソフトウェアの概要、および性能について紹介する。「京」では、100 万コアクラスの超並列計算を効率良く実現するため、並列計算モデルからノード間同期、データ転送、OS、並列ファイルシステム、ソフトウェア開発ツールに至るまで一新した。本発表では、「京」システムで採用した超並列技術について紹介し、基礎的な性能評価結果について議論する。

[キーワード]

次世代スパコン、スーパーコンピュータ『京』、SPARC64 VIIIIfx、HPC-ACE、Tofu インターコネク、性能評価

1. はじめに

理化学研究所様と共同開発を進めている、スーパーコンピュータ「京」は、最先端の 100 万コアクラスの超並列スーパーコンピュータである。このような大規模なシステムを、効率良く安定して動作させるためには、従来からの延長の技術だけでは対応が難しい。並列化されたプログラムを効率よく動作させる実行モデルと、それをサポートするハードウェア技術、消費電力を抑え、さらに高い信頼度を実現する実装技術、ハードウェア性能を引き出すチューニングやアルゴリズムの開発が欠かせない。

2. 「京」の概要

「京」は、富士通が独自に開発・製造した LSI、SPARC CPU とインターコネクトコントローラ、それぞれ 1 チップと、DDR3 メモリ 16G バイトの主記憶から成るノードを、8 万以上のノード接続した超並列スパコンである⁽⁵⁾。SPARC CPU (SPARC64 VIIIIfx⁽²⁾) は、HPC-ACE⁽¹⁾と呼ぶ SPARC V9 アーキテクチャ拡張を採用した 8 コアの 64 ビットプロセッサである。インターコネクトコントローラ (ICC: Interconnect controller)⁽⁴⁾は、Tofu インターコネクトアーキテクチャ⁽³⁾を実装した ASIC である。

「京」では、100 万コアクラスの規模の超並列計算の効率的な実行のため、VISIMPACT と呼ぶ並列計算モデルの導入、Tofu インターコネクトアーキテクチャの開発、OS の最適化、高並列・高信頼ファイルシステム (FEFS) の開発、運用管理、ソフトウェア開発ツールの整備など、数々の技術開発を行った。さらに、CPU を低消費電力化するとともに、水冷技術を用いることで、システムの消費電力の低減と、高い信頼性を実現した。

「京」は 2011 年 6 月に発表された Top500 で、8.162PetaFlops の性能で 1 位となった。2 位に対して 3 倍以上の性能を実現しながら、電力は 9.98MW であった。

SPARC64 VIIIIfx は、汎用性の高い SIMD 拡張を行ない、1CPU あたり、128GFLOPS の性能を実現した。SIMD 拡張によりコアあたり 9%の面積増加で 2 倍の性能を実現している。

1 2010 年 7 月に理化学研究所様が発表した「次世代スーパーコンピュータ」の愛称です

Tofu インターコネクは、6次元メッシュ/トラスにより、保守時での継続運用、I/Oの分離、故障ノードの回避などを実現している。また、システムを分割して複数ジョブを走行させる場合にも、各々のジョブに対して論理的な3次元空間を割り当てることが可能である。これらの機能により、システムの高い運用性を実現している。

ソフトウェアとして、ファイルシステム、OS、MPI、などについては、一部OSS（オープンソースソフトウェア）も活用し、汎用性と相互運用性を担保するとともに、コンパイラや運用ソフトなどについては、独自に開発を進めることで、信頼性と、性能を確保し差別化を実現している。

汎用Linuxを計算ノードのOSとして利用可能とするために、OS内のノイズ源を把握し、ノイズ間隔と継続時間を規定内(200ms周期で50us以内の処理)に抑えるなどの取り組みも進めている。集団通信のハード化などもノイズの拡大を防いでいる。

システムの信頼性を確保するために、代替制御パスの確保や徹底した2重化を行い、単一故障によるシステムの停止を防いでいる。

3. 性能

HPC-ACEによって拡張されたレジスタ（256本へ拡張）の効果とSIMD拡張の効果、コンパイラ開発に用いている89本の実アプリケーションによって評価した。結果、拡張レジスタの使用により、平均1.43倍、最大3.20倍の性能向上が確認された。SIMD拡張と拡張レジスタ両方を活用することで、平均1.84倍、最大4.50倍の性能向上となった。

Tofu インターコネクで用意した4本のTNI（Tofu Network Interface, DMAエンジン）を活用することで、1TNI時は、4.76GB/sの転送スループットが、4TNI同時利用時には、15.03GB/sのスループットとなることが確認された。

Tofu高機能バリアにより、9,216プロセスで、バリア成立までの時間が、ソフトウェアのみの場合に87.9usであったものが、12.2usと7.2分の1に短縮された。Tofu高機能バリアの集団通信(allreduce)とVISIMPACTによるハイブリッド並列により、姫野ベンチマークが65,536コアにおいても高いスケラビリティを示すことが確認された。

Tofuの複数同時通信機能により、1リンクのバンド幅を超えた高い通信性能が実現されることも確認できた。単一スループット性能だけではなく、Alltoall通信の様なシステムワイドな通信において、1,536プロセスでも、同時に開発した専用アルゴリズムの適用により3次元トラスのインターコネク利用効率を94%にまで向上させることができた。複数同時通信機能を利用したアルゴリズムにより、InfiniBandなどのFat treeでのバイセクションバンド幅の利用効率を上回る、高いインターコネク利用効率を実現した。

4. まとめ

超並列スパコンで高い実効性能と安定した運用を実現するために導入・開発したハードウェア、ソフトウェアについて紹介した。

性能評価を行った結果、目論見通りの性能が実現されていることを確認した。「京」の開発により、超並列システムの技術基盤を確立した。

今後「京」が、種々のアプリケーションに適用され、運用性含め多くのフィードバックが得られることが期待される。これらのフィードバックと、CPUとインターコネク両方のLSI設計技術を生かして、エクサシステムへの展開も見据え、高性能スパコンシステムの開発を継続する。

[参考文献]

- (1) SPARC64 VIIIfx Extensions,
<http://img.jp.fujitsu.com/downloads/jp/jhpc/sparc64viiifx-extensionsj.pdf>
- (2) SPARC64 VIIIfx: Fujitsu's new generation octo core processor for PETA scale computing Takumi Maruyama, August 25, 2009, Hot chips21,
http://www.hotchips.org/archives/hc21/3_tues/HC21.25.500.ComputingAccelerators-Epub/HC21.25.51A.Maruyama-Fujitsu-Octo-Core-VIIIfx.pdf
- (3) Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers, Yuichiro Ajima, Shinji Sumimoto, Toshiyuki Shimizu, Computer, pp. 36-40, November, 2009
- (4) ICC: An interconnect controller for the Tofu interconnect Architecture, Takashi Toyoshima, August 24, 2010, Hot chips22,
<http://www.hotchips.org/uploads/archive22/HC22.24.510-1-Toyoishima-Tofu-icc.pdf>
- (5) 次世代スパコン「京(けい)」のコアテクノロジー, 追永勇次, 応用物理、第80巻、第7号、p.0590-0593 (2011)