[全国共同利用情報基盤センター研究開発論文集 No.32]より

大規模科学計算システムの紹介と性能評価

山下毅[†]　江川隆輔[‡]　小野敏[†]　伊藤英一[†]　岡部公起[‡]　小林広明[‡]

[†]東北大学情報部情報基盤課
[‡]東北大学サイバーサイエンスセンタースーパーコンピューティング研究部

1　はじめに

東北大学サイバーサイエンスセンター（以下、当センター）では，当センターのユーザの皆様へ，より強力かつ使い易いシステム提供を目指して， 2010年 4 月にベクトル型スーパーコンピュータ SX-7C を SX-9 へ，また，並列型コンピュータシステム TX7/i9610 を，最新鋭のスカラ型サーバ Express5800/A1080a-D へとリプレースした．これにより，当センターの大規模科学計算システムは，ベクトル機能の強化ばかりでなく，汎用アプリケーション，もしくはベクトル型スーパーコンピュータでは効率的に処理できないアプリケーションの実行環境も大幅に強化した．本稿では，Express5800/A1080a-D システムのハードウェア，およびプログラミングを支えるソフトウェアの紹介をはじめ，当センターにおける並列計算機の運用，そして実アプリケーションを用いた性能評価について説明する．

2　大規模科学計算システムの紹介

2.1　システム更新概要

今回のリプレースでは，当センターの大規模科学計算システムのフロントエンドサーバである TX7/i9610 3 ノードを，Express5800/A1080a-D [1]（日本電気（株）製）6 ノードに，計算サーバであるベクトル型スーパーコンピュータ SX-7C の 5 ノードを，同じくベクトル型スーパーコンピュータ SX-9 [2]（日本電気（株）製）の 2 ノードに更新した．図 1 に更新された大規模科学計算システムの構成を示す．ファイルサーバは，100TB の記憶容量を持つ磁気ディスク装置である．Express5800/A1080a-D の各ノードは，高速なノード間接続装置 (Infiniband)で高速に相互接続されている．また SX-9 およびファイルサーバとはギガビットイーサネットスイッチおよびファイバチャネルスイッチにより接続されている．ギガビットイーサネットスイッチは学内 LAN(StarTAINS) と接続され，大規模科学計算システムは学内外のユーザにサービスされている．

旧システムにおけるベクトル型スーパーコンピュータ SX-7C の 5 ノード (ピーク性能 640Gflop/s, 640GB memory) は，2 ノードの SX-9(3.2Tflop/s, 2TB memory) にリプレース

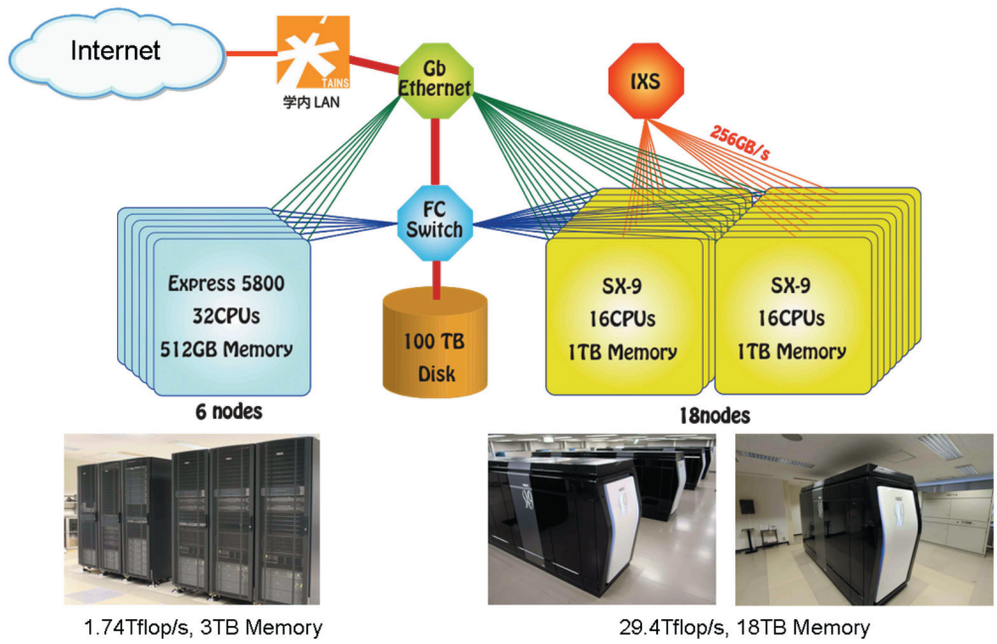

図 1　大規模科学計算システムの構成

され，既に導入済みの 16 ノードの SX-9 と併せてトータル 29.4Tflop/s となり，全国の当センターユーザのベクトル型計算機に対する高いニーズに応えられるシステム構成とした．新システムであるスカラ並列コンピュータ Express5800/A1080a-D は 1.74Tflop/s の演算能力を有し，Gaussian, Marc, Marc Mentat, Mathematica, MATLAB, SAS 等の汎用アプリケーションの高速実行環境を提供するとともに，高いベクトル化率が達成困難なアプリケーション実行をサポートするためにリプレースされた．また，Express5800/A1080a-D は，当センターの主力計算機であるベクトル型システムのフロントエンドサーバとしても利用されており，並列ユーザばかりでなくベクトルユーザにも快適なプリプロセッシング環境の提供を可能にしている．今回のリプレースにより，ベクトル型システムおよびスカラ型システムを合わせた大規模科学計算システムの理論性能は 31.14Tflop/s に達する．

次に今回新たに導入した Express5800/A1080a-D システムの詳細について述べる．

2.2　Express5800/A1080a-D システム

Express5800/A1080a-D は，米国 Intel® 社の最新の 64bit 8 コアプロセッサである Intel Nehalem-EXプロセッサをノードあたり 4 台 (32 コア) 搭載している．当センターでは 6 ノード導入することにより，合計 192 コア，1.74Tflop/s の理論性能を実現している．メモリは各ノードあたり 512GB 搭載し，大規模演算を強力にサポートしている．各ノードは 40GB/sec の通信性能を有する Infinibandで接続され，高速なノード間通信を実現している．旧並列コンピュータシステムと新並列コンピュータシステムの諸元を表 1 に示す．システム全体では

表 1　並列コンピュータ諸元比較

		TX7/i9610	Express5800/A1080a-D	向上比
1 コア	動作周波数	1.6GHz	2.26GHz	1.4 倍
	最大演算能力	6.4Gflop/s	9.06Gflop/s	1.4 倍
1CPU ソケット	コア数	2	8	4 倍
	最大演算能力	12.8Gflop/s	72.48Gflop/s	5.6 倍
	メモリバンド幅	8.3GB/s	34.1GB/s	4.1 倍
	L3 キャッシュ容量	24MB	24MB	1.0 倍
1SMP ノード	最大演算能力	409.6Gflop/s	289.9Gflop/s	0.7 倍
	メモリ容量	512GB	512GB	1.0 倍
システム全体	ノード間通信速度	800MB/s	4GB/s	5.0 倍
	最大演算能力	1.2Tflop/s	1.7Tflop/s	1.4 倍
	メモリ容量	1.5TB	3TB	2.0 倍

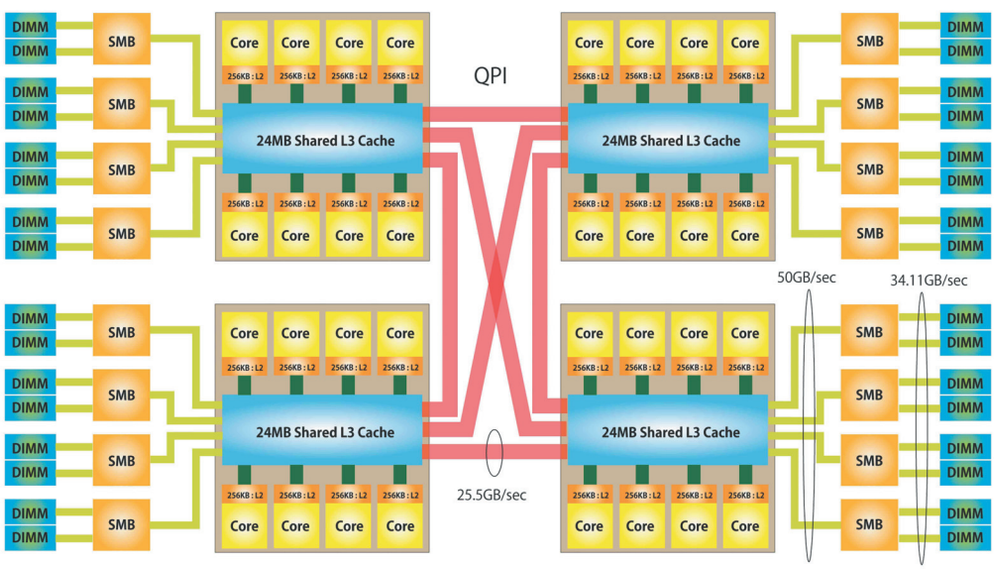

図 2　ノード内構成図

Express5800/A1080a-D は TX7/i9610 と比較して，ノード間通信速度で 5.0 倍，最大演算能力で 1.4 倍, メモリ容量で 2.0 倍の性能を有している．

次に，図 2 に Express5800/A1080a-D のノード内部の構造を示す．Express5800/A1080a-D に搭載されている Intel Xeon プロセッサ 7500 番台は，1 クロックあたり，最大で 4 個の演算を同時に実行可能な Nehalemマイクロアーキテクチャと，単一 CPU ソケット内に 8 つのコアプロセッサを内蔵したマルチコア構造を採用している．これにより，ソケットあたりの最大性能は 72.3Gflop/s に達する．当センターに導入した Express5800/A1080a-D では，各ノードあたり 4 つの CPU を搭載しており，ノードあたりの性能は 289.3Gflop/s となる． これは旧シス

テムの TX7/i9610 と比較して，コアあたりで 1.4 倍，CPU ソケットあたりで 5.6 倍の計算能力を有する．また，各 CPU 間は Quick Path Interconnect (QPI) を介して，高いバンド幅で結合されている．メモリのスループットとレイテンシに関しても，Express5800/A1080a-D では，TX7/i9610 と比較して高い性能を実現している．TX7/i9610 では，2 つの CPU 毎に Front Side Bus (FSB) を共有した論理構造を取っていたのに対して，Express5800/A1080a-D では，プロセッサとメモリコントローラを単一の CPU ソケット内に実装するために，図 2 に示す様に各メモリは SMB(Scalable Memory Buffer) を介して，CPU に直接接続される論理構造を採用している．このため，プロセッサコアから同一ソケット内のメモリへアクセスする際には，DDR3 DIMM のスループットを十分に活用できる．Express5800/A1080a-D では，プロセッサコアあたり TX7/i9610 の約 1.4 倍のメモリスループットを実現している．これらの新たなアーキテクチャの導入により，さらなるユーザプログラムの高速化が期待される．

2.3　運用環境

プログラミング言語および科学技術計算用ライブラリとして，表 2 に示すものが利用出来る．コンパイラは自動並列化機能を有しているので，既存の逐次処理プログラムを修正することなく並列実行が可能である．その他 OpenMP によるノード内並列化および，MPI による複数ノードを使用した並列実行も可能である [4]．また，表 3 に示す汎用アプリケーションの実行環境が提供されている．特に分子軌道計算プログラムの Gaussian は最大 16 並列までの並列処理が可能で，実行時間の大幅な短縮が可能である．バッチ処理としては，スカラ型並列コンピュータおよび，ベクトル型スーパーコンピュータに NQS II(Network Queuing System)[3, 4] を導入することで，ジョブの一元管理が可能となり，利便性の高いジョブ投入環境となっている．スカラ型並列コンピュータで提供しているジョブクラスを表 4 に示す．

表 2　プログラミング言語及びライブラリ

Fortran95	ISO/IEC 1539-1:1997，自動並列化，OpenMP
C/C++	ISO/IEC 14882:1998，自動並列化，OpenMP
MPI	並列処理ライブラリ
ASL	Fortran95用科学技術計算ライブラリ
Math Kernel Library	数値演算ライブラリ

表 3　提供アプリケーション

Gaussian09,03	非経験的分子軌道計算プログラム
MSC.Marc / MSC.Marc Mentat	汎用構造解析プログラム
Mathematica	数式処理プログラム
MATLAB	科学技術計算言語
SAS	データ解析システム

表 4　ジョブクラス

ジョブクラス（キュー名）	利用可能コア数（並列数）	CPU 時間制限	メモリサイズ制限 (GB)
as	1	無制限	16
am（Marc 専用）	1	無制限	16
am2（Marc 専用）	1	無制限	128
a8	8	無制限	128
a16	16	無制限	256
a32	32	無制限	512

3　Express5800/A1080a-D の性能評価

今回導入したシステムの性能を明らかにするために，当センターで実行されているユーザプログラムおよび，汎用アプリケーションである Gaussian03により性能評価を行った．新システムの Express5800/A1080a-D では表 3 の通り Gaussian09の提供も行っているが，旧システムの TX7/i9610 との性能比較を行うために Gaussian03での演算を行った．

3.1　プロセッサ性能

各プロセッサの諸元を表 5 に示す．ここで，B/Flop(Byte per flop) は演算命令あたりのメモリ転送能力を表す．始めに，当センターを利用するユーザが作成した実アプリケーションによる性能評価を行った．評価に用いたベンチマークは，Earthquake (地震解析)，Turbulent Flow (乱流)，Antenna (アンテナ)，Turbine (タービン)，Land Mine (地雷探索)，Plasma (プラズマシミュレーション) の 6 種類である．プロセッサの実行性能を比較した結果を図 3 に示す．上記の 6 つのアプリケーションのうち，4 つ（グラフでは左 4 つ）はプロセッサの演算能力が支配的なプログラムであり，2 つ（グラフでは右 2 つ）はメモリアクセスが支配的なアプリケーションで

表 5　Intel プロセッサ諸元比較

	クロック 周波数 (GHz)	ピーク性能 (Gflop/s)	ピークメモリ バンド幅 (GB/s)	コア数	B/Flop	キャッシュ容量
Express5800/A1080a-D Nehalem-EX	2.26	72.48	34.1	8	0.47	L2:256kB/core L3:24MB
Nehalem-EP	2.93	46.93	25.6	4	0.55	L2:256kB/core L3:8MB
TX7/i9610 Itanium II	1.60	12.80	8.5	2	0.66	L2:256kB/core L3:24MB

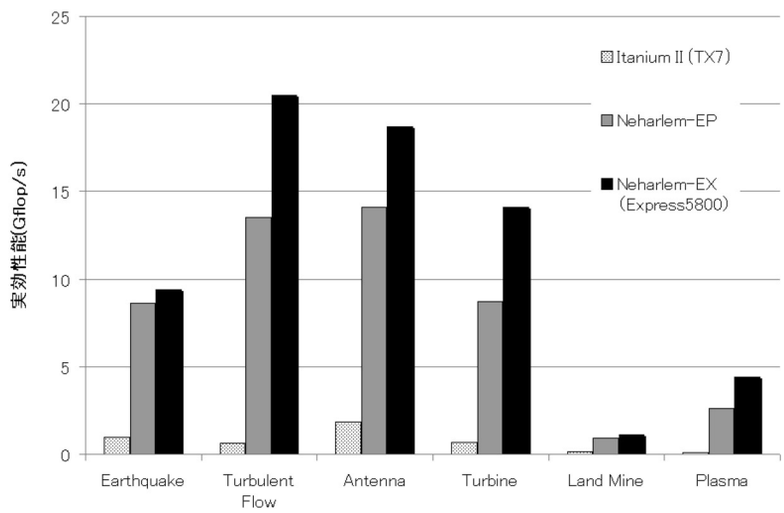

図 3　プロセッサ実効性能

ある．

グラフから分かるように，旧システムの TX7/i9610 と比較して大幅な性能向上を実現している．また参考として，Nehalem-EP (4 コアの Xeon) の値も併記している．ここで，Nehalem-EX の B/Flop の値が Nehalem-EPと比較して低いのは，コアあたりのメモリバンド幅が狭いことが原因として考えられる．高効率なジョブの実行のためにはコアあたりのメモリバンド幅が重要となる．

3.2　ノード性能

次に，次世代 CFD アルゴリズムである BCM(Building Cube Method) アプリケーションをノード内でシングル実行または並列実行し，コア数毎に評価した実効性能を図 4 に示す．新シス

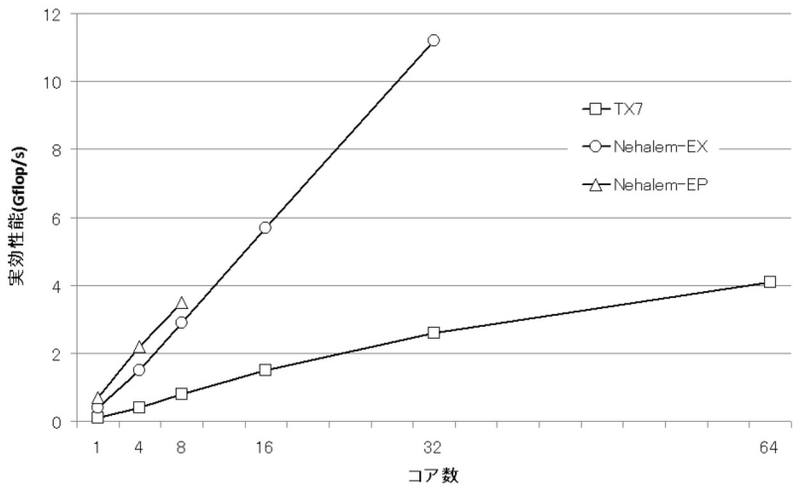

図 4　ノード実効性能

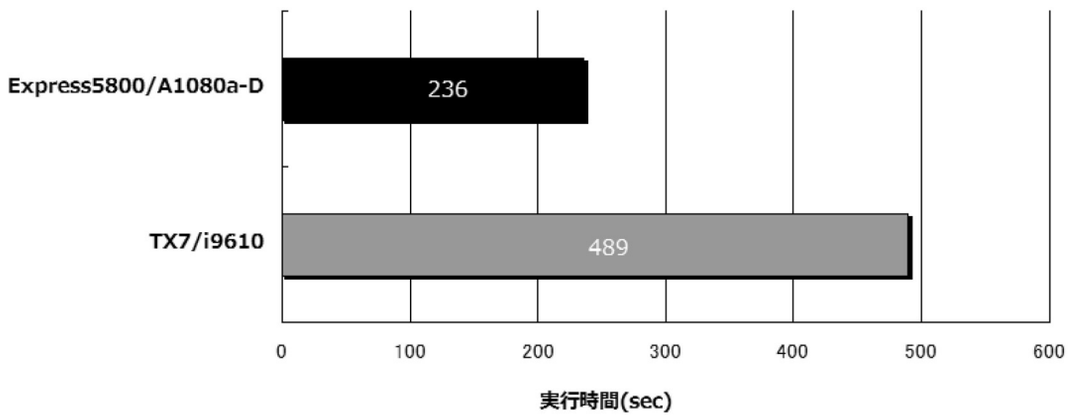

図 5　Gaussian03 32 並列テストプログラムの実行結果比較

テムは旧システムと比較して，同コア数での実効性能で大幅な性能向上を示すとともに，並列コア数に対してリニアに性能を向上させており，並列実行性能の高さも示している．また，旧システムと比較してノードあたりの総コア数は半減しているが，システム全体としても約 2.7 倍の性能向上を示している．

最後に，Gaussian03のテストプログラムを用いて性能評価を行った結果を図 5 に示す．旧システムの TX7/i9610 および新システム Express5800/A1080a-D 上で 32 並列テストプログラムの実行に要した時間を比較すると，新システムは従来システムと比較して約 2 倍の性能向上を実現している．これは各コア間を高速に繋ぐ QPI の性能と，約 1.4 倍高速になったメモリスループットの影響が大きいものと考えられる．

4　おわりに

本稿では，新並列コンピュータシステム Express5800/A1080a-D についてシステム構成，および演算性能評価について紹介した．実効性能，並列実効性能ともに期待通りの性能向上が見られた．当センターにおいては，システム更新によるハードウェアの性能向上を超える勢いで，利用者のジョブは大規模・長時間化してきている．このため，システムの増強と併せて 1997年から高速化推進活動を行い利用者プログラムの高速化および並列化率の向上を図っており，高度化する利用者のニーズに対応できるサービスの提供を目指している．

謝辞

本プログラム評価において快くプログラムをご提供いただいた，東北大学東北アジア研究センター佐藤研究室，東北大学大学院理学研究科地震・噴火予知研究観測センター長谷川研究室，東北大学工学研究科航空宇宙工学専攻升谷研究室，東北大学工学研究科電気・通信工学専攻澤谷研究室，東北大学工学研究科機械システムデザイン工学専攻太田研究室，東北大学工学研究科航空宇宙工学専攻中橋研究室に感謝いたします．

参考文献

[1] 日本電気株式会社　那須康之，鈴木健一，谷岡隆浩 Express5800/A1080a-D のハードウェア 東北大学サイバーサイエンスセンター大規模科学計算システム広報 SENAC Vol.43 No.3 2010-7

[2] 日本電気株式会社　稲坂純，萩原孝 スーパーコンピュータ SX-9 のハードウェア 東北大学サイバーサイエンスセンター大規模科学計算システム広報 SENAC Vol.41 No.3 2008-7

[3] 東北大学情報部情報基盤課 システム管理係，東北大学サイバーサイエンスセンタースーパーコンピューティング研究部 スーパーコンピュータシステム SX-9 利用ガイド 東北大学サイバーサイエンスセンター大規模科学計算システム広報 SENAC Vol.41 No.2 2008-4

[4] 東北大学情報部情報基盤課 共同研究支援係，共同利用支援係，東北大学サイバーサイエンスセンタースーパーコンピューティング研究部 並列コンピュータ Express5800 の利用法 東北大学サイバーサイエンスセンター大規模科学計算システム広報 SENAC Vol.43 No.2 2010-4