［大学 ICT 推進協議会 2015 年度年次大会論文集より転載］

スーパーコンピュータ SX-ACE の運用について

小野　敏 1), 齋藤　敦子 1), 森谷　友映 1), 佐々木　大輔 1), 山下　毅 1),
大泉　健治 1), 岡部　公起 2), 江川　隆輔 2), 小林　広明 2)

1) 東北大学　情報部　情報基盤課
2) 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
ono@cc.tohoku.ac.jp

概要：東北大学サイバーサイエンスセンターは、全国共同利用機関として学内外の研究者に対して大規模科学計算システムを提供している。本稿では、2015 年 2 月より運用している新ベクトル型スーパーコンピュータ SX-ACE の運用状況等や、本システムを用いた共同研究成果について報告する。

1　はじめに

東北大学サイバーサイエンスセンター（以下、本センターと略す）では、常に最高・最新鋭の大型計算機を導入し、大規模科学技術計算の要求に応えてきた。本センターの大規模科学計算システムは二種類の計算機を運用し、主に研究室のワークステーションでは行えないような大規模シミュレーションに用いられている。このうちベクトル演算に適したものは、ベクトル型スーパーコンピュータ SX-ACE[1]で、それ以外のものはスカラ型並列コンピュータ LX 406Re-2 で実行されている。これまで本センターでは大規模・長時間シミュレーションの実行を運用の中心としており、2015 年 2 月より運用している新スーパーコンピュータシステム SX-ACE の運用においても同様の方針で運用環境を構築した。また大規模・長時間シミュレーションの実行には、利用者アプリケーションの高度化・高速化、MPI による並列化が必須であり、高速化支援活動が重要であると考えている。

本稿では、新スーパーコンピュータシステム SX-ACE を含む大規模科学計算システムの概要、付帯設備、運用環境、運用状況および高速化支援活動について紹介する。

2　大規模科学計算システムの概要

本センター大規模科学計算システムの構成を図 1 に示す。本センターの大規模科学計算システムはベクトル型スーパーコンピュータ SX-ACE、スカラ型並列コンピュータ LX 406Re-2、ストレージシステム、および三次元可視化システムから構成される。

ベクトル型スーパーコンピュータ SX-ACE は 2,560 のノードから構成される。1 つのノードは 276GFLOPS のベクトルプロセッサ 1 基と 64GB の主記憶容量を有している。各ノードは、ノード間接続装置(IXS)で高速に相互接続されており、1 ジョブで最大 1,024 ノードを占有する並列処理が可能である。なお OS は UNIX System V に準拠した SUPER-UX である。スカラ型並列コンピュータ LX 406Re-2 は 68 のノードから構成される。1 つのノードは 460GFLOPS の Intel Xeon プロセッサ 2 基と 128GB の主記憶容量を有している。なお OS は Linux である。ストレージシステムは lustre で構成される 1 次ストレージ 1PB と ScaTeFS(NEC Scalable Technology File System)で構成される 2 次ストレージ 3PB の合計 4PB の容量を有している。1 次ストレージは並列コンピュータとの高速な I/O を、2 次ストレージは利用者のホーム領域として並列コンピュータおよびスーパーコンピュータとの高速な I/O

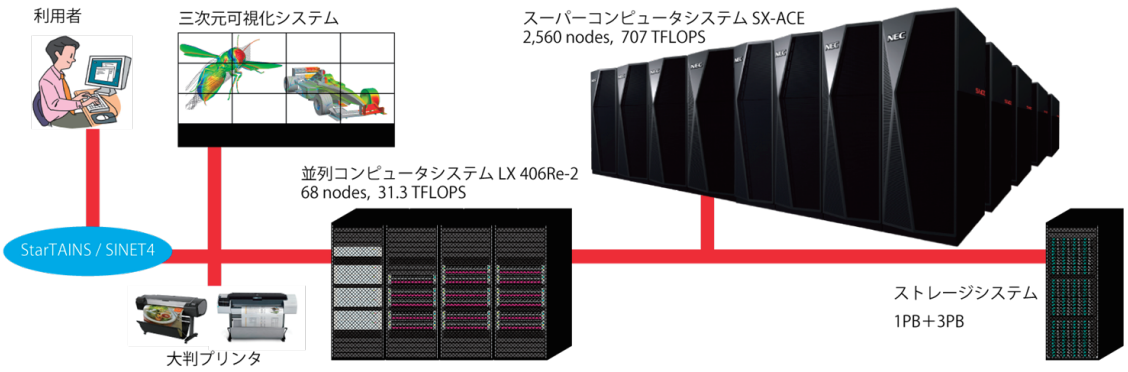

図 1　大規模科学計算システム構成図

を実現している。

本センターの大規模科学計算システムは、ベクトル型スーパーコンピュータを主力計算機としつつ二種類の計算機から構成され、また三次元可視化システム、ストレージシステムとも連携することで、利用者の幅広いニーズに応える計算環境を学内外の研究者へ提供している。

3　付帯設備について

SX-ACE を設置している本センター2 号館の内部構成を図 2 に示す。2 号館は SX-ACE の導入時に必要とされる電源設備、空調機等を含めた計算機の設置スペース、水冷方式に対応した冷却システムを既存の計算機室に確保することが困難であったため、2014 年 11 月に竣工した 2 階建の計算機棟である。スーパーコンピュータ SX-ACE は 2 号館の 1 階に設置しており、2015 年 9 月からは 2 号館の 2 階へ並列コンピュータ LX 406Re-2、ストレージシステムを移設した。計算機を集約することで、冷却設備の一元的な運用が可能となり、省エネルギー運用を実現している。

SX-ACE は水冷・空冷の両冷却方式で運用されている。水冷設備の冷却水と空調に必要な冷水は密閉型フリークーリング方式の冷却塔(クーリングタワー)で生成される。2 号館屋上に設置した冷却塔一体型チリングユニットを図 3 に示す。冷却の仕組みは、冷却塔が外気取り入れ送風ファンまたは散布水により、冷媒の通るコイルを冷やし冷水を作り出している。夏場等の暑い期間は、外気や散布水だけでは冷やしきれないため、チラーを併用し電気を使って冷水を作り出している。

SX-ACE ではさらに冷却の効率を高めるため

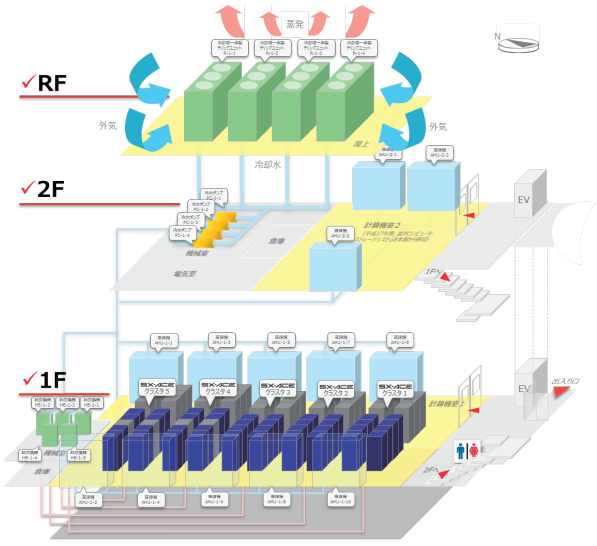

図 2　2 号館内部構成

図 3　冷却塔一体型チリングユニット

図 4　SX-ACE のコールドアイルキャッピング

の設備としてアイルキャッピングを導入している。その様子を図 4 に示す。SX-ACE および付帯装置が収納されるラックは、冷気の吸入側である前面側を向かい合わせてレイアウトされている。フリーアクセスフロアの底面から供給される冷気をラック前面から効率的に吸入するために、向かい合うラックの側面と上面をビニールカーテンにより仕切るコールドアイルキャッピングを採用している。また、計算機ラック背面からの排気は天井面の吸入口から天井裏を介し、空調機へ戻す方式である。

4　運用環境

4.1 プログラミング言語とライブラリ

SX-ACE で利用可能なプログラミング言語および科学技術計算用ライブラリを表 1 に示す。SX-ACE ではこれまで提供していたコンパイラに加え、新たに Fortran 2003 に対応したコンパイラも提供している[2]。並列処理については、MPI による並列処理を前提としているが、ノード内については自動並列処理、OpenMP による並列処理が利用可能であり、MPI と組み合わせたハイブリット並列、またはフラット MPI による並列処理によりシミュレーションの大規模並列実行が可能となっている。

科学技術計算用ライブラリは、SX-ACE 用に最適化された ASL、ASLSTAT が利用可能であり、SX-9 を利用していたプログラムについては、再コンパイルすることにより実行可能である。BLAS、LAPACK、ScaLAPACK を利用しているプログラムについては、MathKeisan ライブラリを利用することで、プログラムを変更することなく実行可能である。

表 1　SX-ACE で利用可能なプログラミング言語と科学技術計算用ライブラリ

プログラミング言語・ライブラリ	規格・機能
Fortran 90/95	ISO/IEC 1539-1:1997 自動並列、OpenMP
Fortran 2003	ISO/IEC 1539-1:2004 自動並列、OpenMP
C, C++	ISO/IEC 9899-1999 ISO/IEC 14882:2003 自動並列、OpenMP
MPI/SX	MPI-3.0
ASL	数値計算ライブラリ
ASLSTAT	統計計算ライブラリ
MathKeisan	数学ライブラリ集 (BLAS、LAPACK、 ScaLAPACK を含む)

4.2 ジョブ管理および利用形態

ジョブスケジューリング機能は NQSⅡを導入し並列コンピュータとスーパーコンピュータの一元管理が可能となっており、利便性の高いジョブ投入環境となっている。また、SX-ACE へのバッチジョブ投入、コンパイルなどの作業は SX-ACE ではなくフロントエンドサーバである並列コンピュータにて行う[3]。

SX-ACE の利用形態は占有利用と共有利用を用意している。占有利用は一定数のノードを利用者、あるいは研究グループが期間を定めて占有し利用する方式である。契約されたノード数内での並列処理が可能であり、ジョブの実行待ちについては、研究グループ内のジョブにのみ影響され、他の利用者のジョブによる実行待ち

は発生しない。共有利用は他ユーザと利用するノードを共有する利用方式であり、利用する時期や利用ノード数によってジョブの実行待ちが発生するが、最大で 1,024 ノードでの大規模並列実行が可能である。

また共有利用には通常利用、無料利用、デバッグ利用の 3 つの利用形態があり、利用形態はジョブ投入時のオプションによって指定することができる。これまでの運用方針と同様に、大規模・長時間ジョブをサポートするため、実行時間制限の最大値は 1 ヶ月としている。各利用形態の詳細を表 2 に示す。

通常利用はジョブ投入用コマンド qsub のオプションとして、-q で sx を -b でノード数を指定することで利用可能であり、利用ノード数によって実行時間制限の規定値が異なっている。デバッグ利用は、-q で debug を -b でノード数を指定することによって利用可能であり、利用ノード数によって実行時間制限が異なっている。

SX-ACE では研究室の PC、ワークステーション等で実行されているプログラムの SX-ACE への移行を支援する目的で、最大 1 ノード、1 時間の実行時間制限付きでジョブの実行を無料利用できる利用形態を用意している。これは-q で sx を-b で f を指定することによって利用可能である。

表 2　共有利用での利用形態

利用形態	利用ノード数	実行時間制限
通常	1～256	規定値:2 週間 最大値:1 ヶ月
	257～1,024	規定値:1 ヶ月 最大値:1 ヶ月
デバッグ	1～16	2 時間
	17～32	24 時間
無料	1	1 時間

5　運用状況

SX-ACE の今年度 4 月から 9 月までのノード時間とジョブ件数の推移を図 5 に示す。4 月、5 月と比較して 6 月から 8 月にかけてはジョブ件数が大きく増加しているが、7 月、8 月についてのノード時間は 4 月と同程度にとどまっている。6 月から 8 月にかけてジョブ件数が大きく増加しているのは、システム更新後の新システムに向けた最適化のための短時間ジョブが、多数投入されたためと考えられる。9 月に投入件数、ノード時間が大きく減っているが、これは停電・計算機保守作業、並列コンピュータシステムの移設作業により約 3 週間計算機サービスを休止していた影響によるものである。また例年 4 月から 9 月の期間については、利用者の入れ替わりなどもあり、下半期に比べ利用時間が少なくなる傾向があり、今後の利用に期待したい。

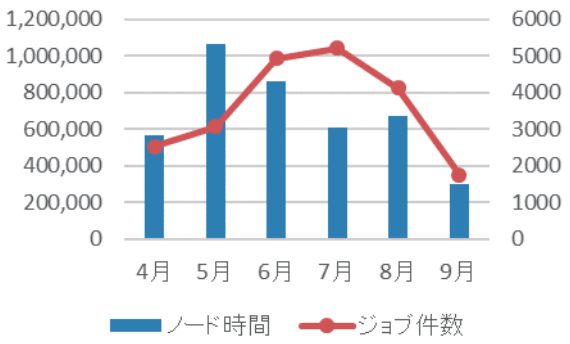

図 5　ノード時間とジョブ件数の推移

SX-ACE の今年度 4 月から 9 月までの利用ノード数に対する経過時間の分布を図 6 に示す。利用ノード数が 32 ノード以下のジョブ件数が一番多くなっており約 80%程度を占めている。これはデバッグ利用や無料利用についての利用ノード数は 32 ノード以下であり、短時間で実行が終了するものが多いためこのような状況であると考えられる。実行時間制限の最大値である 1 ヶ月間実行するジョブは存在しなかったが、規定値である 2 週間に達するジョブについては 10 件程度存在した。また、利用ノード数にかかわらず、ジョブの実行が 48 時間以内で終了するジョブが大部分を占めていた。4 月から 8 月は並

列数の上限が 512 ノードであったが、すでに 512 ノードを利用したジョブを実行している利用者もいた。9 月以降は 1,024 ノードでの並列実行も可能となっており、より大規模なジョブが実行できる環境を整えている。

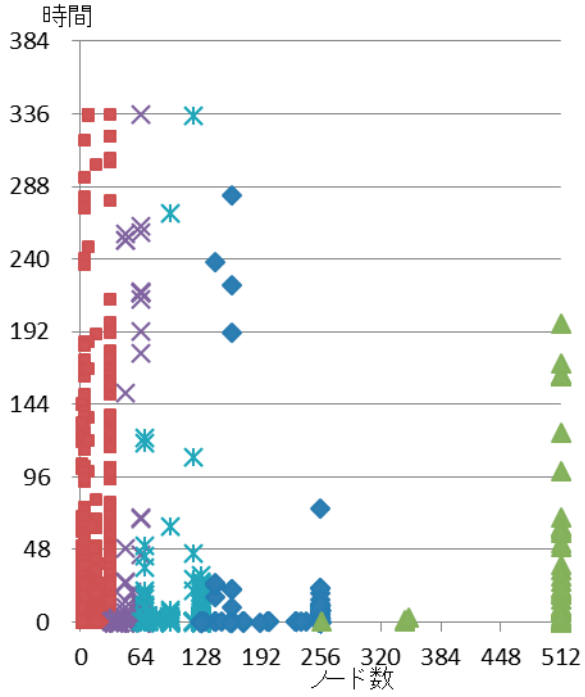

図 6　ノード数に対する経過時間分布図

6　高速化支援活動

本センターでは 1999 年よりユーザアプリケーションの高精度化、大規模化の支援を目的とした共同研究制度を実施している。また社会貢献の一環として、サイバーサイエンスセンター共同研究制度の他に、産学連携共同研究に基づく民間企業利用も実施しており、学術分野のみならず産業のイノベーション創出にも貢献している。さらには全国の情報基盤センター等と連携して行う JHPCN(学際大規模情報基盤共同利用・共同研究拠点)、HPCI(革新的ハイパフォーマンス・コンピューティング・インフラ)システム利用研究課題を通じて、計算科学分野の利用者と共同研究を行い、様々な分野におけるアプリケーションの最適化・高度化、並列化に取組み高速化支援活動を行ってきた。

図 7 に 1999 年から取り組んでいる共同研究の推移を示す。この図からわかるように、サイバーサイエンスセンター共同研究は恒常的に年 10 件程度実施されていることに加え近年、JHPCN、HPCI を介した共同研究数が増加していることが確認できる。これは、サイバーサイエンスセンター共同研究を通してユーザアプリケーションが高度化、大規模化し JHPCN、HPCI 採択課題へとステップアップしているためだと考えられ、継続的な高速化支援活動が一定の成果をあげていることがわかる。また表 3 にこれまで本センターが行ってきた高速化支援の成果を示す。1997 年から 2014 年にかけて 200 件の共同研究に基づくプログラムの最適化に取り組んでおり、単体性能、並列性能ともに大幅な向上を実現している。

本センターで導入したスーパーコンピュータ SX-ACE での大規模ジョブの実行には MPI ライブラリによる並列化が必須である。これまでコンパイラによる自動並列処理、OpenMP による並列処理のみ利用していた利用者の MPI 化支援も積極的に行ってきた。この他各種利用者講習会の開催、大阪大学との高速化に関するハンズオンセミナーの共同開催、高速化推進研究活動報告書の発行、メール、電話での利用・高速化相談に加え、窓口での対面による利用・高速化相談などを行ってきた。この取り組みを今後も継続し、新たに大規模科学計算システムを利用する利用者、高速化・MPI 化が必要な利用者に支援を行い、研究を進める上での強力なツールとして役立てて頂きたいと考えている。

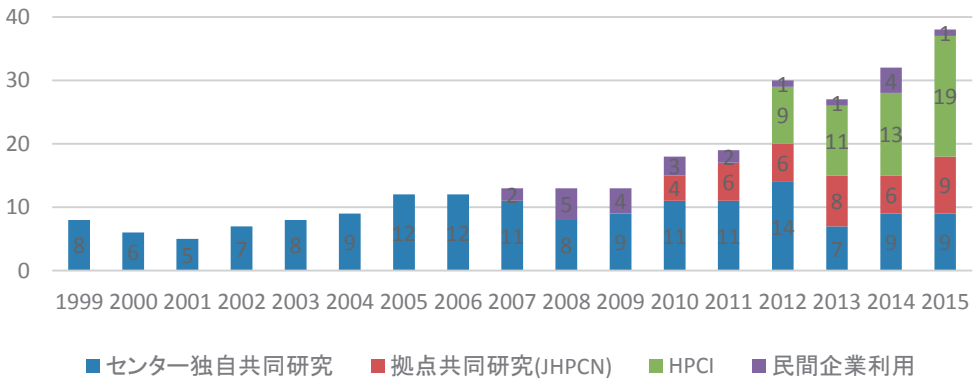

図 7　共同研究数の推移

表 3　高速化支援活動実績

年度	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
件数	2	9	8	9	10	7	18	20	8	29	10
単体性能向上比	1.9	46.7	4.5	2.5	1.6	2.2	6.7	2.9	1.5	3.1	33.0
並列性能向上比	11.1	18.4	31.7	8.6	4.9	2.8	18.6	4.5	4.1	8.0	1.9

年度	2008	2009	2010	2011	2012	2013	2014
件数	15	8	8	13	6	11	9
単体性能向上比	9.3	47.0	47.2	16.2	19.7	16.7	10.3
並列性能向上比	5.1	3.6	48.5	17.2	15.3	12.9	8.0

7　おわりに

本稿では、2015 年 2 月より運用を開始した新スーパーコンピュータシステム SX-ACE を含む大規模科学計算システムの概要、付帯設備、運用環境、運用状況および高速化支援活動について紹介した。SX-ACE での大規模シミュレーションの実行には MPI による並列化が必須であり、今後も利用者アプリケーションの高速化・高度化、並列化による高速化支援活動は必要である。それとともに、高速化に対する認識を利用者に高めてもらうよう努力していきたいと考えている。

8　参考文献

[1] 日本電気株式会社、萩原 孝、浜口 博幸、山信田 恒、スーパーコンピュータ SX-ACE 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 Vol.48、No.1、pp.5-14、2015.

[2] 日本電気株式会社、工藤 淑裕、横谷 雄司、SX-ACE でのプログラミング(ベクトル編)、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 Vol.48、No.1、pp.15-32、2015.

[3] 東北大学情報部情報基盤課共同利用支援係、共同研究支援係、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、スーパーコンピュータ SX-ACE 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 Vol.48、No.2、pp.1-61、2015.