［大学 ICT 推進協議会 2014 年度 年次大会論文集より転載］

スーパーコンピュータシステム SX-ACE の紹介

†山下毅 †森谷友映 †佐々木大輔 †齋藤敦子
†小野敏 †大泉健治 ‡岡部公起 ‡江川隆輔 ‡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
yamacta@cc.tohoku.ac.jp

概要：東北大学サイバーサイエンスセンターは、全国共同利用設備として大規模科学計算システムの整備と、HPCI の資源提供機関としての役割を担っている。本稿では、2015 年初頭に運用を開始する本センターの主力計算機である新ベクトル型スーパーコンピュータ SX-ACE と、その運用方針およびユーザの利用環境について紹介する。

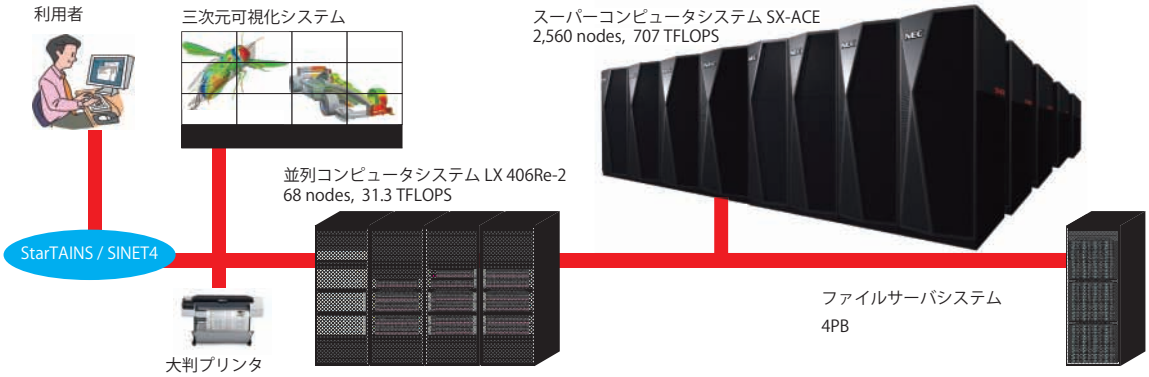

図 1　大規模科学計算システムの構成（2015 年以降）

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）では、2015 年初頭に新スーパーコンピュータシステム SX-ACE（日本電気株式会社製）の運用開始を目指し、本年 11 月に竣工の HPC 新棟（仮称）内において、現在ハードウェアおよびソフトウェアの環境構築作業を行っている。今回導入した SX-ACE のシステムは全 2,560 ノードで構成され、1 ノードあたり理論最大演算性能 276GFLOPS の世界初のマルチコア（4 コア）ベクトルプロセッサを 1 基搭載し、システム全体では約 707TFLOPS となる。主記憶は 1 ノードあたり 64GB を搭載し、256GB/s という高いメモリバンド幅でプロセッサと接続されることで、高い演算性能とメモリ性能の最適化を実現している。

本稿では、新スーパーコンピュータシステムのハードウェアおよびソフトウェアの特徴と、新システムの導入に伴い今回新設された HPC 新棟の概要、および利用者の利便性とセキュリティの向上を考慮して構築を行った、大規模科学計算システムの運用方針について紹介する。

2　大規模科学計算システム

2.1　システムの概要

本センターの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を主力計算機とし、汎用アプリケーションの実行環境とし

表 1　SX-9 と SX-ACE の性能比較

	性能	SX-9	SX-ACE	向上比
CPU あたり	コア数	1 個	4 個	4 倍
	理論最大演算性能	118.4GFLOPS	276GFLOPS	2.3 倍
	最大ベクトル演算性能	102.4GFLOPS	256GFLOPS	2.5 倍
	メモリバンド幅	256GB/sec	256GB/sec	1 倍
	ADB	256KB	1,024KB/コア× 4	16 倍
ノードあたり	CPU 数	16 個	1 個	0.06 倍
	理論最大演算性能	1,894GFLOPS	276GFLOPS	0.15 倍
	最大ベクトル演算性能	1,638GFLOPS	256GFLOPS	0.16 倍
	メモリ容量	1TB	64GB	0.06 倍
	メモリバンド幅	4TB/sec	256GB/sec	0.06 倍
	ノード間通信速度	256GB/sec	8GB/sec	0.03 倍
システムあたり	CPU 数	288 個	2,560 個	8.9 倍
	理論最大演算性能	34.1TFLOPS	706.6TFLOPS	20.7 倍
	最大ベクトル演算性能	29.5TFLOPS	655.4TFLOPS	22.8 倍
	メモリ容量	18TB	160TB	8.9 倍
	最大消費電力	590kVA	1,080kVA	1.8 倍
	計算機室床面積	293 平米	430 平米	1.5 倍

てスカラ型の並列コンピュータの運用も行っている。この二種類の計算機の運用により、利用者の幅広いニーズに応えるサービスを提供している。

SX-9 システムは、2008 年 3 月から運用を開始し本年で 7 年目を迎えているが、6 年半の平均利用率は 80% を超え、また今年度上半期の平均利用率は過去最高の 90% と高い利用率となっており、科学技術計算においてベクトル型スーパーコンピュータのニーズの高さを伺うことが出来る。(利用率：ユーザプログラムの CPU 時間合計÷システムの稼働時間合計× 100【%】)

並列コンピュータシステムの LX 406Re-2 と、合計 4PB の容量を有するファイルサーバシステム、および三次元可視化システムは、2014 年 4 月に導入され運用を行っている。これらのシステムは、スーパーコンピュータシステムとの連携、および分散ファイルシステムを活用した計測データの高速な I/O により、高速かつ高精度な防災・減災シミュレーションを行い、シミュレーション結果を三次元可視化するシステムとして活用される。これらのシステムを利用した、ものづくり分野における萌芽的研究、産業利用の促進も期待される。

以下では、2015 年初頭に運用を開始するスーパーコンピュータシステム SX-ACE について説明する。

2.2　SX-ACE システムの紹介

2.2.1　ハードウェアの特徴

■システム構成　SX-9 システムと SX-ACE システムの性能比較を表 1 に示す。SX-9 システム全体 18 ノードでの理論最大演算性能が 34.1TFLOPS であるのに対し、SX-ACE システム全体の 2,560 ノードでは 706.6TFLOPS となり、約 21 倍の性能向上となる。本センターの SX-ACE システムでは最大 512 ノード並列の実行環境に加え、来年度からは最大 1,024 ノード並列の大規模な実行環境が利用者に提供される。

■消費電力・設置面積　SX-ACE は SX-9 と比較して、同一性能時の LSI 数を約 1/100 へと削減したため、SX-9 と同等の性能を 1/10 の消費電力と 1/5 の設置面積で実現している。

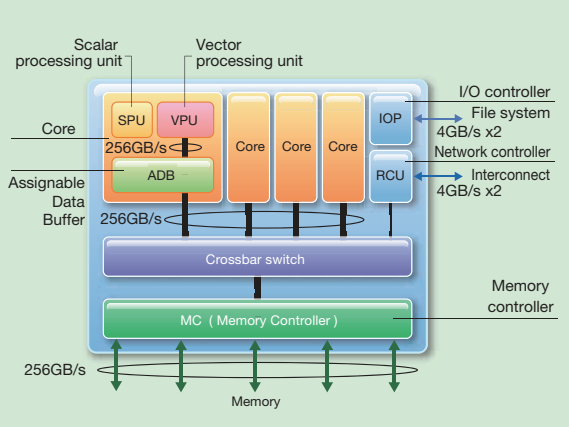

図 2　SX-ACE マルチコアプロセッサ

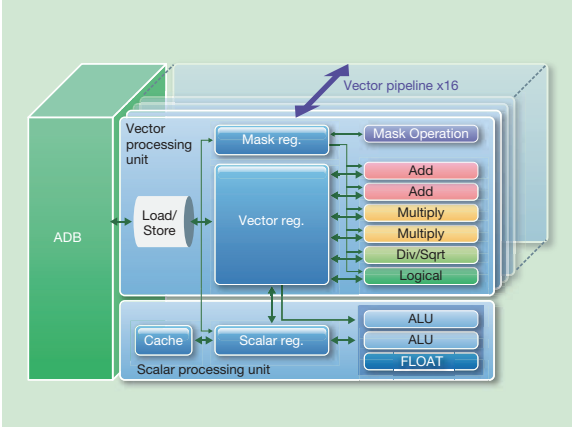

図 3　ベクトルプロセッシングユニット

SX-ACE システムは SX-9 システムと比較して約 21 倍の性能向上であるが、最大消費電力および、計算機室の床面積（付帯装置および冷却装置を含む）はそれぞれ、1.8 倍および 1.5 倍として設計した。SX-ACE システムの導入によって最大理論演算性能の飛躍的向上と、省エネおよび省スペース化の両立を実現することが可能となった。

■ベクトルプロセッサ　SX-ACE のマルチコアプロセッサの模式図を図 2 に、ベクトルプロセッシングユニットの模式図を図 3 に示す。

SX-ACE の CPU はこれまでのベクトルプロセッサと同様の高いベクトル性能と、高いメモリバンド幅を継承し、またシリーズ初のマルチコア化を行った。1CPU は、それぞれ 64GFLOPS のベクトル演算性能を持つ計 4 コアで構成され、256GFLOPS のベクトル演算性能を持つ。

CPU 内のコア間は 256GB/sec のクロスバスイッチにより高速接続され、クロスバスイッチと主記憶間はメモリコントローラを介し 256GB/sec で高速接続されており、CPU あたり 1Byte/FLOP を達成している。

各コアには、容量が 1,024KB に拡張された HPC 専用設計キャッシュである ADB（Assignable Data Buffer）が搭載され、ADB とコア間のメモリバンド幅は 256GB/sec を有する。データが ADB 経由でアクセスされる場合は、4Bytes/FLOP という高いバンド幅によるデータ供給性能により、メモリ負荷の高いアプリケーションも高い実行効率での実行が可能となる。また、ノード内並列数を 1 として、単一のコアのみを演算に使用する場合も 256GB/sec のメモリバンド幅が利用でき、このときも 4Bytes/FLOP での実行が可能である。

■ノード間通信　ノード間の通信は最大 4GB/sec × 2（双方向）で接続された 2 段ファットツリーネットワークで構成される。SX-ACE は 1 ノードあたり 256GFLOPS という高いベクトル演算性能により、プログラムの実行に必要な並列度を低く抑えることが可能であり、多並列での実行時に並列性能の高いスケーラビリティが期待出来る。

以上のような SX-ACE のハードウェアの特徴により、これまでベクトル向けに開発されてきたアプリケーションは勿論のこと、学術研究者の幅広い分野のアプリケーションを高い実行効率で実行されることが期待される。

2.2.2　ソフトウェアの特徴

■オペレーティングシステム　オペレーティングシステムは前システムから引き続き、POSIX 準拠の SUPER-UX を採用している。OS レベルでマルチノードをサポートし、大規模マルチノードシステムにおいても安定した利用環境を提供して

表 2　SX-ACE で利用可能なプログラミング言語と数値演算ライブラリ

言語・ライブラリ	コンパイラ名・ライブラリ名	準拠規格・機能
Fortran 90/95	FORTRAN90/SX	ISO/IEC 1539-1:1997 準拠 自動ベクトル化、自動並列化、OpenMP 対応
Fortran 2003	NEC Fortran 2003 コンパイラ	ISO/IEC 1539-1:2004 準拠 自動ベクトル化、自動並列化、OpenMP 対応
C,C++	C++/SX	ISO/IEC 9899:1999 C 準拠 ISO/IEC 14882:2003 C++ 準拠 自動ベクトル化、自動並列化、OpenMP 対応
MPI ライブラリ	MPI/SX	MPI-3.0 準拠
科学技術計算ライブラリ	ASL ASLSTAT MathKeisan	数値計算ライブラリ 統計計算ライブラリ 数学ライブラリ集 （BLAS, LAPACK, ScaLAPACK を含む）

いる。

■言語とライブラリ　SX-ACE で利用可能な言語およびライブラリを表 2 に示す。SX-ACE では新たに Fortran 2003 に対応したコンパイラを導入し、幅広い Fortran コードの実行が可能である。また、MPI-3.0 に準拠した MPI ライブラリは Fortran、 C/C++ から利用可能であり、同一ノード内では共有メモリの特徴を活かした自動並列／ OpenMP 並列の利用、または Flat MPI 実行による大規模並列実行が可能である。

また、SX シリーズに最適化された科学技術計算ライブラリとして、表 2 に示すライブラリが引き続き利用出来る。ライブラリを利用して SX-9 向けに作成したプログラムも、SX-ACE 用にコンパイルし直すだけで実行が可能である。BLAS、LAPACK、ScaLAPACK ライブラリを利用して記述したプログラムは、MathKeisan ライブラリをリンクすることで、ソースコードの変更なく実行が可能である。

■高速化支援ツール　SX-9 で高速化支援ツールとして活用されてきた、プログラム実行解析情報（PROGINF）と簡易性能解析情報（FTRACE）は引き続き利用可能であり、加えて GUI でプログラムの性能を解析可能な NEC Ftrace Viewer を導入する。これは FTRACE 機能で採取された性能解析情報をグラフィカルに表示し、ベクトル性能や OpenMP、MPI を利用した並列プログラムのスレッド・MPI プロセス毎の実行時間、MPI プロセス間の通信時間をグラフ表示することにより、性能のボトルネックやロードインバランスを把握するためのツールである。これらのツールを用いることで、実行コストの高いサブルーチンの特定や、並列実行時の演算量の均一化、ノード間通信の最適化等の高速化作業が容易になる。

■分散・並列ファイルシステム　HPC システムの大規模化やデータの大容量化に対応するために、分散・並列ファイルシステムである、NEC Scalable Technology File System（ScaTeFS）を採用した。SX-ACE および並列コンピュータの各ノードとストレージシステムを ScaTeFS で接続し、データおよびメタデータを複数の I/O サーバに分散配置することで負荷分散とスケールアウトを実現し、システム全体のスループットの向上が見込まれる。

■ジョブ管理システム　ジョブスケジューリング機能は前システムと同様 NQSⅡ を採用し、計算リソースを管理して効率的なユーザのジョブ管理を行う。

NQSⅡ によるバッチジョブ投入方法の概略を図

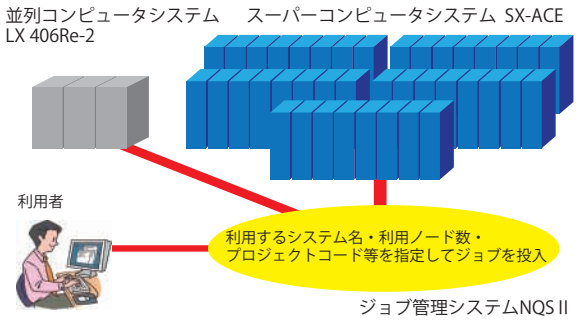

図 4　NQSII によるバッチジョブ投入方法

図 5　HPC 新棟外観（2014.10.31 撮影）

4 に示す。利用者はジョブ投入の際に、利用するコンピュータシステム、利用ノード数、プロジェクトコード名、ジョブスクリプトファイル名、ノードあたりの実行プロセス数、ノード内並列数等を指定することにより、同一のフロントエンドサーバから SX-ACE、LX 406Re-2 の両システムに対してジョブの投入が可能である。

この機能により、以下のように利用者のアプリケーションの特性に合わせて、柔軟にジョブの投入を行うことが可能となる。

・ロードストアを多用する、メモリ負荷の高いアプリケーションは SX-ACE を利用し、ノード内は単一コアのみを使用して高いメモリ転送速度を活用。
・並列実行性能が高いアプリケーションは SX-ACE を利用し、コンパイラの自動並列化機能を用いてノード内は 4 スレッド並列で実行。
・高いスカラ演算性能が要求されるアプリケーションは LX 406-Re2 を利用し、単一コアで実行することでターボブースト機能を活用。

また、スケジューラマップの導入により、ジョブの実行予定がない実行ノードを検出した場合に自動的に CPU のコア縮退運転、またはノードが省電力運転へ移行することが可能であり、高い QoS を保ったまま消費電力を削減することが出来る。この機能により、計画停電などの運用予定に合わせて指定された期日に運用ノード数を調整することも可能である。

実行時に稼働していたノードに障害が発生した場合は、他に空きノードがある場合はそれらを割り当てることにより、利用者に対して迅速に計算機環境を提供することが可能である。

3　HPC 新棟について

3.1　新棟の概要

SX-ACE システムの導入に先立ち、本センターの正面に隣接して HPC 新棟が竣工した。HPC 新棟の外観写真を図 5 に示す。

SX-ACE システムで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、および SX-ACE の水冷方式に対応した冷却システム（水冷装置設備本体と配管設備一式）を、現在 SX-9 を運用している本センターの計算機室に確保することは困難であったため、新棟の建設が行われた。

今後の HPC システムの規模拡大も見据え、継続的に情報処理基盤拠点としての役割を担うことが出来る計算機棟としての理念の下、新棟の設計が行われた。

3.2　空調・水冷設備

SX-ACE は水冷・空冷の両冷却方式で運用されるため、水冷方式に対応するための配管設備と、空冷方式に対応するための空調設備の設置が必要となる。

冷却水と空調に必要な冷水は、屋上に設置した密閉形フリークーリング方式の冷却塔（チラー）で生成される。周囲温度の低下時（中間期・冬期）

に、冷却塔のみで冷水を生成させ直接負荷側に送水することで、エネルギーの使用量を低減させることが出来る。

3.3　アイルキャッピング

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ルカーテンにより仕切る、コールドアイルキャッピングの方式を採用したことで省エネ効果が期待される。

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、空調機に還気する方式を採用している。

4　システムの運用方針

4.1　利用者環境について

4.1.1　ログイン認証方式

■パスワード認証方式の廃止　本センターの SX-9 システムではフロントエンドサーバへのログインの際に、パスワード認証方式および公開鍵暗号方式の両方が利用可能であり、どちらの方式でログインを行うかは利用者の判断に任されている。SX-ACE システムでは、昨今の漏洩パスワードによる不正アクセスのセキュリティインシデントへの対策として、フロントエンドサーバへログインする際のパスワード認証方式を廃止し、公開暗号鍵による認証方式のみ利用可能とした。

■鍵ペアの生成　ログインに必要な公開鍵・秘密鍵ペアの生成においては、パスフレーズを設定しない、あるいは強度の低いパスフレーズにより秘密鍵が作成されることを防ぐために、フロントエンドサーバへの初回接続時には、新たに設置した SSH アクセス認証鍵生成サーバ（以下、鍵サーバ）で鍵ペアを生成する方式を採用した。

利用者は発行された利用者番号と初期パスワードにより鍵サーバへログインし、専用の鍵ペア生成プログラムにより、一定強度のパスフレーズを持つ秘密鍵を作成する。画面に表示された秘密鍵のテキストをローカル PC にコピー&ペーストにより保存し、このファイルを秘密鍵として利用する。この秘密鍵を利用して、フロントエンドサーバに公開暗号鍵方式でのログインが可能となる。なお、鍵サーバ上で鍵ペアを作成すると、鍵サーバへのログインはロックされる。

■HPCI 課題利用者のログイン方法　HPCI 課題利用者は、HPCI が提供する GSI 認証による電子証明書を用いたシングルサインオンでのログインのみが利用可能である。

4.1.2　プロジェクトコード

NQSⅡのジョブアカウント機能によって、ユーザが異なるプロジェクトで計算機資源を利用する際に、ジョブ単位の課金と予算管理を行うことが出来る。SX-ACE システムではこの機能を利用し、1 つの利用者番号で複数の請求先の使い分けを可能とするために、利用者管理の 1 つとしてプロジェクトコードを導入した。プロジェクトコードの導入前と導入後の利用者と請求先の関係を図 6 に示す。プロジェクトコードの導入により、利用者には以下の様な利便性と、セキュリティ性の向上が期待出来る。

■複数の請求先の利用　従来の研究室予算での利用に加え、課題採択形式で利用されるケースが近年増加している。SX-9 システムで複数の請求先を利用する場合の模式図を図 6 左に、SX-ACE システムでの場合を図 6 右に示す。

SX-9 システムまでは、複数の請求先を利用する場合、請求先毎に支払責任者番号を発行する必要があった。そのため、利用者 A が複数の請求先を使い分けるには、請求先ごとの支払責任者番号（u2000, u2100）のそれぞれに利用者番号（c2200, d2300）を取得し、ログインする利用者番号を使い分ける必要があった。

SX-ACE システムからはプロジェクトコードの導入により、利用者 A は 1 つの利用者番号

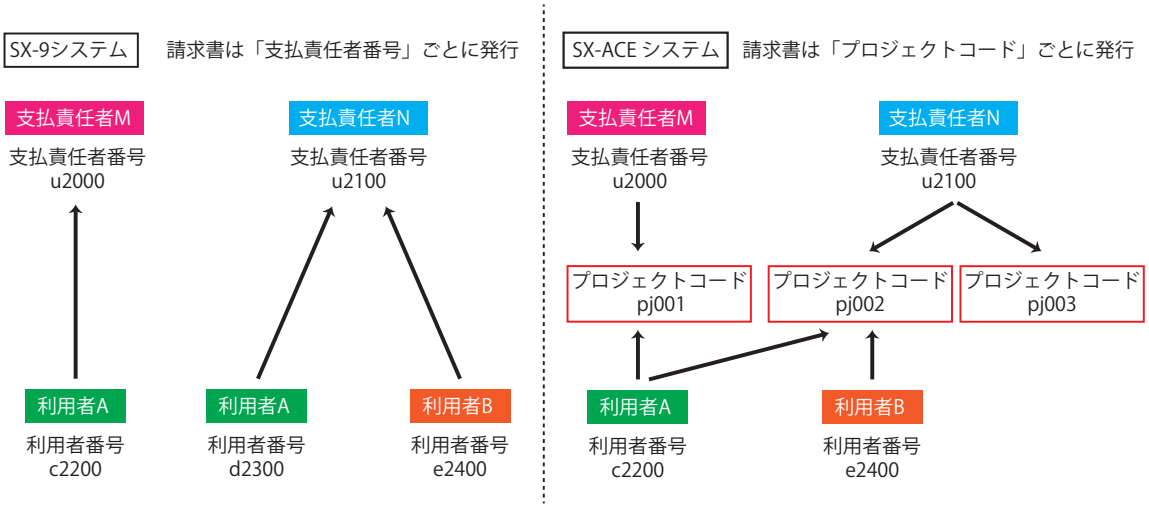

図 6　プロジェクトコードの導入による利用者と請求先の関係

(c2200) から請求先の異なる複数のプロジェクトコード (pj001, pj002) を使い分けることが可能になる。バッチジョブ投入の際に NQSII のジョブアカウント機能を用い、請求先としたいプロジェクトコードを指定することで、利用者が複数の請求先を使い分けることが可能となる。また、支払責任者が複数のプロジェクトコード (pj002, pj003) を保有することも出来る。

■利用者環境　SX-9 システムでは請求先毎に利用者番号を使い分ける必要があったため、利用者がそれぞれの利用者番号に割り当てられたユーザのデータ領域内で環境の構築が必要であった。また、利用者番号・パスワードの管理も利用者番号毎に必要であった。SX-ACE システムでは、利用者は 1 つの利用者番号で、複数の請求先を利用することが可能であるので環境の構築は 1 箇所で済み、また鍵ペアの管理も簡便になるためセキュリティ性も向上する。

■課題利用期間とプロジェクトコード　採択課題の利用期間が終了したものについては、該当するプロジェクトコードを無効にすることで、利用者はジョブを投入することが不可能となる。また、利用者が利用可能な課題が追加された場合は、利用者番号に対してプロジェクトコードを追加設定することでジョブの投入が可能となるため、それまで利用していた環境を引き続き利用することが可能である。

4.2　利用ノード数と利用負担金について

大学・学術利用における、SX-ACE システムの利用ノード数と利用負担金を表 3 に示す。なお、民間企業利用に関しては、大学・学術利用単価の 3 倍の単価設定としている。

4.2.1　共有利用

共有利用は、他のユーザと利用するノードを共有する方式である。本センターの運用方針である大規模ジョブの長時間実行環境を提供する目的で、ジョブの実行時間はスケジューラマップ時間以内で無制限としている。待ち行列は FIFO を基本とするが、利用者がジョブの実行時間をジョブスクリプトファイルに明示することで、リソースに空きがある場合はジョブのエスカレーションが自動的に行われる。

研究室の PC、あるいは共有サーバ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、経過時間制限を設定した 1 ノード利用を無料としている。また、ジョブの並列化を促進させる目的で、1 ノードから 32 ノードの利用は利用負担金単価は一定とし、33 ノード以上の利用では利用するノード数が多くなると、利用負担金

表 3　SX-ACE システムの利用ノード数と利用負担金（大学・学術利用）

【共有利用】

利用ノード数	経過時間制限※	最大メモリサイズ	利用負担金単価【円/秒】
1	あり	64GB	無料
1～32	なし	2TB	0.06
33～256	なし	16TB	(利用ノード数-32) × 0.002+0.06
257～1,024	なし	64TB	(利用ノード数-256) × 0.0016+0.508

※ジョブの実行時間は、スケジューラマップ時間以内とする。

【占有利用】

利用ノード数	最大メモリサイズ	利用期間	利用負担金【円】
32	2TB	3 ヶ月間	400,000
		6 ヶ月間	720,000
64	4TB	3 ヶ月間	720,000
		6 ヶ月間	1,300,000
128	8TB	3 ヶ月間	1,300,000
		6 ヶ月間	2,340,000

単価の増加率が減少する 2 段階の単価設定としている。

また、SX-ACE では NQSⅡの機能として会話リクエスト機能が追加され、クライアント環境から SX-ACE のノードに直接ログインすることなく対話型操作が可能である。会話リクエストの場合、1 ノードで実行された際の課金体系(0.06 円/秒)が適用される。

4.2.2　占有利用

占有利用は一定数のノードを利用者、あるいは利用者グループが占有して利用する方式である。この場合、利用者は占有利用として設定されたプロジェクトコードを指定してジョブの投入を行う。共有利用で使用されるスケジューラマップにはアサインされないため、契約されたノード数まではジョブが優先的に実行される。ジョブが利用可能なノード数を超えた場合は、実行中のジョブが終了するまでジョブは実行されないが、研究グループ内でジョブ実行までの待ち時間や、利用額を考慮した計画的な実行が可能である。

5　おわりに

本稿では 2015 年初頭に運用を開始する、サイバーサイエンスセンターのスーパーコンピュータシステム SX-ACE について紹介した。また、SX-ACE システムの導入に伴うシステムの運用方針の変更点について説明した。最新のベクトル型スーパーコンピュータ SX-ACE を、皆様の研究にご活用いただけたら幸いである。

謝辞

可視化画像をご提供いただきました、宇宙航空研究開発機構の中橋和博先生、スーパーコンピュータシステム SX-ACE の導入および環境構築にあたり、日本電気株式会社、NEC ソリューションイノベータ株式会社、NEC フィールディング株式会社の皆様には多大なるご協力をいただきました。皆様に深く感謝の意を表します。