[全国共同利用情報基盤センター研究開発論文集 No.31]より

超大規模ベクトル計算基盤の実現へ向けて

大泉健治†，村田善智††，江川隆輔††，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

超高速コンピュータ網形成プロジェクト(National Research Grid Initiative：NAREGI)は，世界標準に準拠した実運用に耐えうる品質のグリッド基盤ソフトを開発することを目的として2003年に開始された産学官連携による研究開発プロジェクトです[1]. 同プロジェクトにおいて精力的に開発が進められているNAREGIグリッドミドルウェアは，広域に点在する研究開発拠点の大規模計算資源を密に連携することで，各計算資源の効率的な利用だけではなく，これまで不可能であった大規模計算を実現可能な基盤として注目されています．また，ベクトル型スーパーコンピュータは，流体計算，構造解析などに代表される大規模科学技術計算を高い実効効率で処理することが可能であり，最先端の研究開発や製品の設計開発における重要な演算基盤として，広く利用されています．

本稿では，将来のベクトル型計算基盤である“ベクトルコンピューティングクラウド”の実現に向けた，東北大学サイバーサイエンスセンターと大阪大学サイバーメディアセンターの取り組みについて述べます．この取り組みは，NAREGIプロジェクトの基盤ソフトウェアであるNAREGIグリッドミドルウェアと両センターで運用されているベクトル型スーパーコンピュータNEC SXシリーズの仮想化技術に基づき，複数のベクトル型スーパーコンピュータシステムの効率的な利用とこれまで不可能であった超大規模ベクトルコンピューティング基盤の実現を目的としています．

2. ベクトルコンピューティングクラウドの概要

本章では，ベクトルコンピューティングクラウド基盤の概要，本取り組みの基本構成要素であるNAREGIグリッドミドルウェアに基づくシステム構成，ベクトルプロセッサの仮想化計算資源であるGRIDVM for SXについて説明します．

2.1. ベクトルクラウドの概要

現在，国内，または世界に点在するベクトル型コンピュータを利用するには，図1に示すように，ユーザは各ベクトルコンピュータサイトにアクセスし，コンパイル・実行という手順を踏むことになります．しかし，各ベクトルサイトはベクトル型スーパーコンピュータに対する高いニーズにより，常に高い稼働率で運用されています[2]．この様な状況下で，様々な異なる規模のジョブを効率よく実行可能なベクトルコンピューティング環境が強く求められています．また，近年の高精度計算に対する高い要求により，各サイトの計算資源を超えた超大規模計算実行環境への要求が年々高まっています．そこで，我々はベクトルコンピューティングクラウド基盤を構築することで，これらの要求を満たすことができると考えています．

図2にベクトルコンピューティングクラウド基盤の概略図を示します．ベクトルコンピューティングクラウド基盤では，各ベクトルサイトの計算資源を仮想化することで，ユーザが複数のベクトルコンピュータシステムを一つの超大規模ベクトルスーパーコンピュータシステムとして利用可能なシングルサインオン環境を提供します．ユーザ・ジョブスケジューラは仮想化された膨大なベクトル計算資源の中から，これまでよりも柔軟，且つ効率的にジョブの規模

に応じたベクトル計算資源を特定し，ジョブを投入・実行することが可能となります．これにより効率的なベクトルスーパーコンピュータシステムの運用が可能となり，ユーザは長いキューイング状態を回避することが期待できます．また，複数の大規模ベクトルスーパーコンピュータシステムを仮想化し，超大規模ベクトルスーパーコンピューティングシステムを構築することで，これまで不可能であった規模の計算が可能になります．これにより，これまでに無い大規模計算環境における効率的な運用・ジョブ実行を可能にする要素技術の確立が期待できます．

我々は，このベクトルコンピューティングクラウド基盤を確立するために， NAREGI グリッドミドルウェアに着目し，同ミドルウェアに基づくベクトルコンピューティングクラウド基盤の構築を目指します．次に，NAREGI グリッドミドルウェアを用いたベクトルコンピューティングクラウド基盤のシステム構成について述べます．

図 1 従来の利用環境

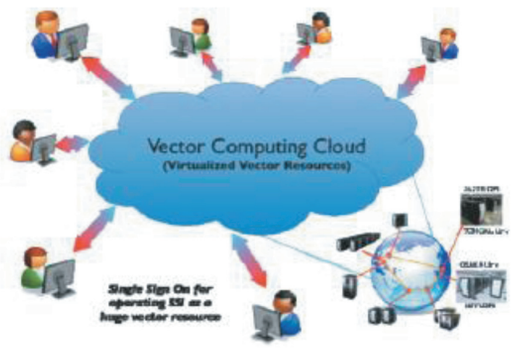

図 2 ベクトルコンピューティングクラウド環境

2.2. システム構成

本稿で提案する NAREGI グリッドミドルウェア ver.1 によるベクトルコンピューティングクラウド基盤のシステム構成を図 3 に示します．NAREGI グリッドミドルウェア(2009 年 10 月現在 ver1.1)，および関連情報は国立情報学研究所の web サイトで公開されていますので，詳細は(http://www.naregi.org/)をご覧ください．このシステムでは各ベクトルコンピュータサイトがそれぞれ，ポータル(portal)，ユーザ管理サーバ(User Management Server:UMS)，仮想化組織管理サービス(VO Management Service:VOMS)，スーパースケジューラ(Super Scheduler :SS)，情報サービス（Information Service: IS)，NAREGI 用仮想 SX 計算資源管理ミドルウェア(GRIDVM for SX)で構成される事を想定しています．各コンポーネントの主な機能は以下の通りです．

Portal	: 仮想化されたシステムのインターフェース群の提供
UMS/ VOMS	: ユーザ・サーバの認証・管理
SS	: 利用者ジョブの要求に応じた資源を探索し，スケジューリング
IS	: グリッドを構成する計算資源の管理．各計算資源稼働状況の収集蓄積
GRIDVM for SX	: 計算資源を仮想化による計算資源の同期制御およびメタコンピューティング環境の提供

各サイトのベクトルコンピュータ資源の連携は，それぞれの SS がリザベーション・キャッシュ・サービス(Reservation Cache Service : RCS)を介して連携することで実現します．RCS は複数のサイトの計算資源の情報を常に監視し，各 SS の要求を調停・管理します．これにより，ユーザはポータルサイトにログイン後，個人・サーバの認証を受け，ポータルサイトに用

意されているワークフローツール，グリッド可視化システム等を用いて，複数の計算資源を一つの計算システムとして扱い，ジョブを投入することができます．投入されたジョブは SS, RCS に渡され，RCS によって仮想化された GRIDVM for SX にスケジューリングされた後，実行されます．この間，各サイト・各サイト間の SS と IS は定期的に同期をとることで，常に最新の資源情報を集積・蓄積していきます．

しかし，現状公開されている NAREGI ver.1.1 が提供している GridVM はベクトル型スーパーコンピュータである NEC SX は対応していません． ベクトルコンピューティングクラウド基盤を実現するためには，NAREGI ver.1.1 で提供されている GridVM を NEC SX 用に移植する必要があります．

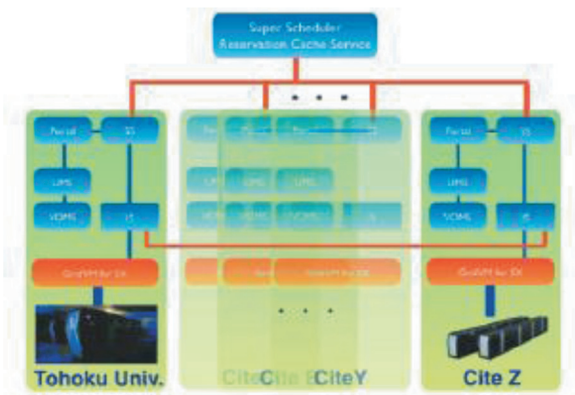

図 3　ベクトルコンピューティングクラウドシステム概念図

2.3. GRIDVM for SX

GRIDVM for SX は，NAREGI ver. 1.1 の GridVM をベクトル型スーパーコンピュータシステム（NEC SX）上で利用可能にし，また本センターのスーパーコンピュータシステム固有の運用性向上と機能強化を目的に開発されました．開発にあたっては，NAREGI ver1.1 が提供している機能のうち，ジョブ管理機能，情報プロバイダ機能，資源利用量制限機能を SX 用に移植するとともに，当センターの大規模科学計算システム固有の機能強化として，ローカルジョブと GRID ジョブの共存の為の強化，NEC SX 固有の MPI のサポートを実現しました．次に,各機能の特徴を説明します．

ジョブ管理機能では，当センターの大規模科学計算システムのローカルスケジューラである NEC 製 NQSII とその拡張モジュールである JobManipulator を用いて，資源予約を行う予約ジョブと資源予約を行わない非予約ジョブをサポートし，その混在を可能にしています．情報プロバイダ機能では，NEC SX の各種ハード・ソフトウェア情報，グリッド環境で利用可なローカルスケジューラのキュー情報，ジョブのステータス情報を NAREGI グリッドミドルウェアに登録することができます．また，グリッド環境にある各ベクトルサイトにおいて，システム管理者が指定したポリシーに基づき，ジョブが利用する資源量を監視し，且つ必要なジョブ制御を行う機能を実装しています．これらの機能を NAREGI ver.1.1 のインターフェースに基づいて実現できるように移植を行いました． 次に，当センター運営を考慮した固有の強化機能として，NAREGI ver.1.1 ではサポートされていなかったローカルジョブと GRID 予約ジョブの共存を，ジョブ毎に資源を分割すること無く可能にしています．また，NAREGI ver.1.1 における MPI 実行は，GridMPI を使用したものに限定されています．これを NAREGI ver.1.1 の JSDL 仕様を変更することなく，MPI/SX による MPI 非予約ジョブの実行を可能にしています．

これらの移植・新機能開発により，GRIDVM for SX では，通常の運用と NAREGI に基づく

グリッド運用の共存を可能にし，効率的なシステム運用を実現します．次章では実際のアプリケーションを用いた NAREGI 環境と GRIDVM for SX の動作確認を目的とした，評価実験について説明します．

3.　実験評価

我々が提案するベクトルコンピューティングクラウド基盤の実現可能性の確認と今回開発した GRIDVM for SX の動作検証を目的として，当センターの SX-9 システムと大阪サイバーメディアセンターの SX-9 システムを用いて実験評価を行いました．今回評価に用いたシステム構成を図 4 に示します．今回の評価では各センターの SX-9 システム 1 ノード（16CPU）に，NAREGI ver.1.1 と GRIDVM for SX を用いた環境を構築しました．600Km 離れた東北大学・大阪大学の両センター間は SINET3 によって結合されています．RCS は大阪大学に配備し，東北大学のポータルからログインし，ジョブ投入を行うことで動作検証を実施致しました．以下に操作の流れに沿って，動作を確認していきます．

はじめに当センターで立ち上げた，図 5 に示すポータルサイトにログインします．シングルサインオン環境のため，このポータルサイトから両サイトのシステムを利用することができます．続けて Submit New Job タグを選択すると，図 6 に示すジョブ投入画面が表示されます．ジョブの実行に際しては，あらかじめ用意したジョブサブミット用コマンドを入力し Submit ボタンを押します．本稿における評価では実際にサイバーサイエンスセンターで実行されている SMP16 並列の電磁界分布シミュレーションプログラムを用います．演算を実行するサイトは，基本的に RCS が自動的に選択し割り当てます．利用できるサイト，各サイトのシステム基本情報，稼働状況をもとに，RCS は適切なサイトにジョブを割り当てます．各サイトの情報や稼動状況は，Server List 画面（図 7）で確認することができますので，利用者が希望のサイトを選択してジョブを実行することも可能です．今回，ジョブの実行サイトは，RCS が自動で選択して評価を行いました．

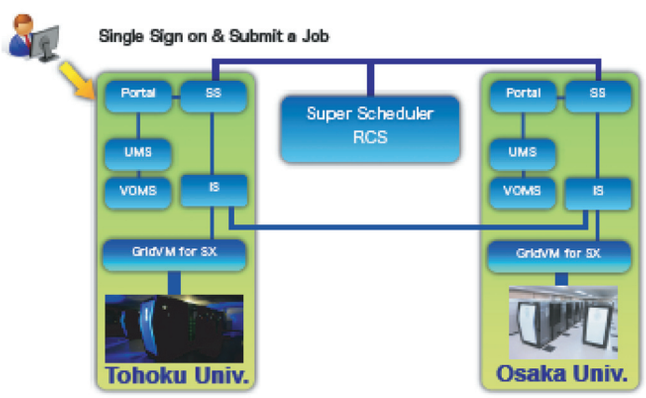

図 4　評価システム

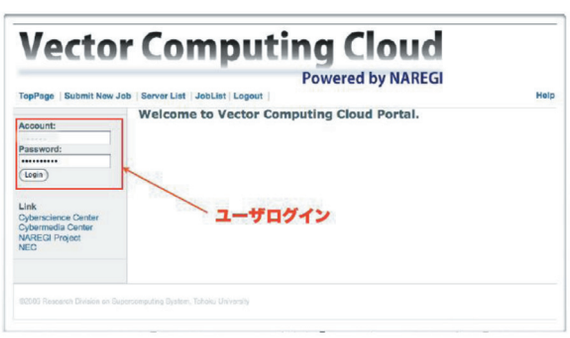

図 5　portal ログイン画面

図 6　ジョブ投入画面

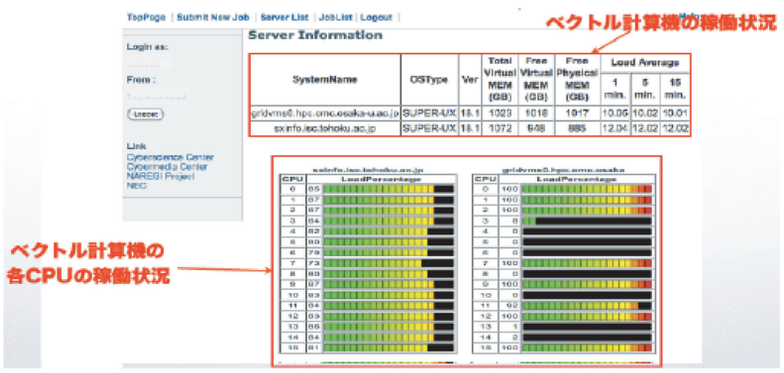

図 7　ベクトル計算機の稼動状況

先に述べたように RCS が適切な演算サイトを自動的に選択しますので，ユーザは演算サイトを選択する必要はありません. ユーザは, 待ち時間表を参考にジョブの投入を行うだけです. Queue Status 表（図 6）から必要とする CPU 数ごとに大まかな待ち時間を知ることができます. 評価に用いたプログラムは 16CPU を必要とする SMP16 並列のプログラムです. Queue Status 表（図 6）中 Job Size 16 の項目が赤信号であることから，しばらく待つことがわかります．さらに詳細な稼動状況を知りたい場合は，Server List 画面（図 7）のように CPU 稼動状況を表示させることも可能です. 上段の表に大阪大学と東北大学の各サイトのシステム情報を，下段にシステムごとの CPU 稼動状況をそれぞれ確認することができます．稼動状況の左側が東北大学(sxinfo.isc.tohoku.ac.jp)，右側が大阪大学(gridvms0.hpc.cmc.osaka-u.ac.jp)のシステムを示しており，両システムとも 16CPU の負荷状況をバーグラフメータ(%)で表示します．このとき，東北大学側の CPU は全て 70%から 90%の負荷状況で空きがありません．一方の大阪大学側も負荷が 0 状態で空いている CPU は，5 個のみであることがわかります．

ジョブの投入状況は，図 8 に示す Job List 画面でステータスを確認することができます．上段から最近のジョブが順に表示され，Status 項目が Queued で待ち状態，Submitted で実行中，Done で終了状態をそれぞれ示しています．Site 項目はジョブが割り当てられたサイトを示します. 図 8 の場合, 最上段のジョブは大阪大学(gridvms0.hpc.cmc.osaka-u.ac.jp)の SX-9 にスケジュールが割り当てられ，実行中(Submitted)であることがわかります．プログラムが終了すると，Status が Done 状態に変わりジョブは完了です.

本評価では，東北大学のポータルにログインし，先に述べた SMP16 並列のプログラムを実行し，今回構築した環境で正常に動作することを確認しました．今回の評価ではノード内の評

価だけではありますが，シングルサインオンで，複数サイトの SX-9 ベクトルスーパーコンピュータを利用，ノード内の SMP プログラムを自動的に演算サイトに割り当て実行が可能であることを示しました．また，東北大学・大阪大学の両サイトはそれぞれ独自の運用ポリシーに沿ってジョブを運用していますが，従来のジョブ運用と今回のベクトルコンピューティングクラウドによるジョブ運用を共存できることも確認しました．

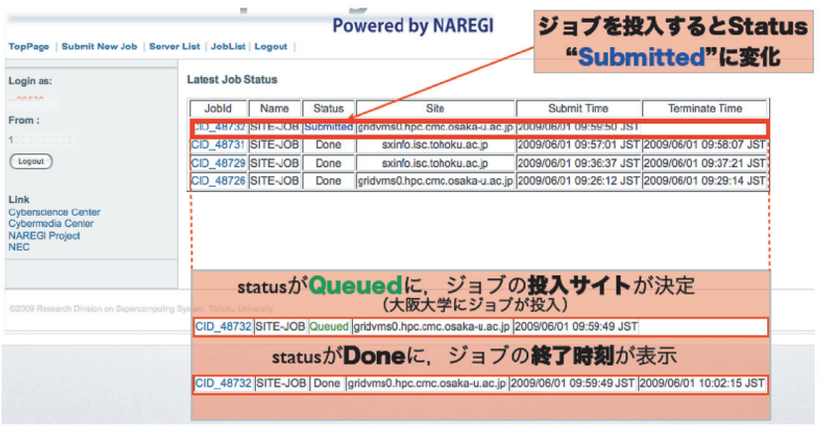

図 8　ジョブのステータス

4.　まとめ

本稿では，ベクトルコンピューティングクラウド基盤構築に向けた，東北大学，大阪大学の広域ベクトル型スーパーコンピュータ連携について述べました．NAREGI グリッドミドルウェアによるグリッド環境の構築と GRIDVM for SX の開発により，約 600Km 離れた両センターの計算資源を仮想化し，シングルサインオンで双方の計算資源の利用が可能になり，演算サイトの割り当てを自動的に選択できることを示しました．また，両サイトにおいて従来のジョブ運用と，ベクトルコンピューティングクラウドによるジョブ運用を共存できることも示しました．今後は，詳細な性能評価，両サイトを跨いだ大規模並列処理が可能なシステムの構築，および効率的な計算資源の利用を目指した RCS スケジューラの改良に取り組み，ベクトルコンピューティングクラウドのさらなる高度化を目指します．併せて，ベクトルコンピュータとの連携のみならずベクトル-スカラ連携も視野に入れたメタコンピューティング環境の構築にも取り組んでいく予定です．

謝辞

本稿を執筆するにあたり多くの方々にご協力ご支援を賜りました．大阪大学サイバーメディアセンター東田学助教をはじめ運用スタッフの皆様，東北大学サイバーサイエンスセンターおよび大阪大学サイバーメディアセンターご担当の日本電気㈱スタッフの皆様には NAREGI 環境構築に際して，多大なるご協力を頂きました．ここに深く感謝申し上げます．なお，本研究開発は，国立情報学研究所が推進する最先端学術情報基盤(CSI:Cyber Science Infrastructure)整備の一環として行っているものです．

参考文献

[1]　超高速コンピュータ網形成プロジェクト，http://www.naregi.org.

[2]　小林広明，”東北大学サイバーサイエンスセンター.新大規模科学計算システム SX-9 全国共同利用施設としての役割.,”　SX-9 導入披露&SENAC50 周年記念式典・講演会資料集，pp.21-35, 2008 年 11 月．