［研究成果］

NAREGIによる広域ベクトル型スーパーコンピュータ連携

―ベクトルコンピューティングクラウドの実現に向けて―

江川隆輔[†]，大泉健治[††]，伊藤英一[††]，山形正明[†††]，神山典[†††]，　金野浩伸[†††]，
東田学[††††]，大西健太郎[†††††]，遠藤直弥[†††††]，山崎潤一[†††††]，関充男[†††††]，
岡部公起[†]，小林広明[†]
[†]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
[††]東北大学情報部情報基盤課
[†††]日本電気株式会社　文教・科学ソリューション事業部
[††††]大阪大学サイバーメディアセンター
[†††††]ＮＥＣシステムテクノロジー株式会社　　第一公共システム事業部

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の一つの在り方として“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

本節では，ベクトルコンピューティングクラウド基盤の構想，本取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 ベクトルコンピューティングクラウドの概要

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．本技術の拡張により，2012 年に稼働が期待されている次世代スーパーコンピュータも仮想化されれば，ユーザがシングルサインオンであらゆる規模のベクトル型スーパーコンピューティングサービスを享受できるようになることが期待できます．

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

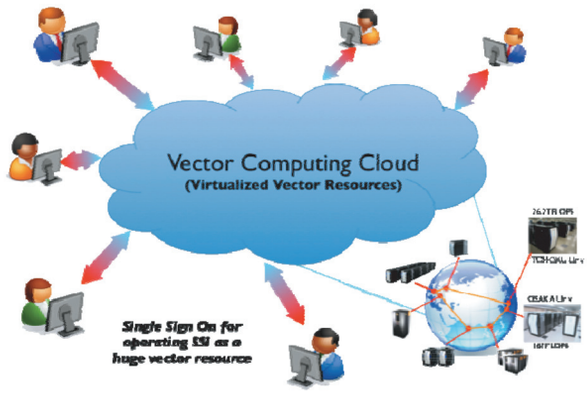

図 2 ：ベクトルコンピューティングクラウド環境

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.1.1 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 4 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)から構成されます．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
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
SS	：利用者ジョブに要求に応じた資源を探索し，スケジューリング
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．

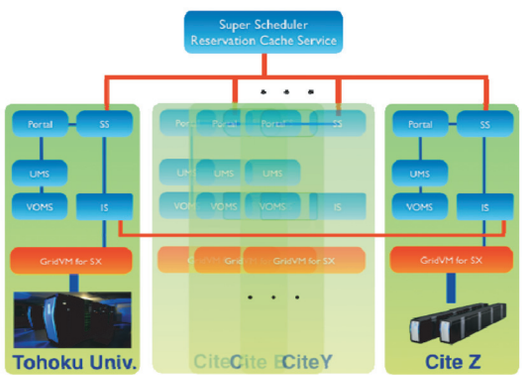

図 3 ： ベクトルコンピューティングクラウドシステム概念図

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 の動作確認を行います．

3. 評価実験

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 の動作検証と評価を目的として，当センターの SX-9 システムと大阪大学サイバーメディアセンターの SX-9 システムを用いて実証実験を行いました．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．東北大学・大阪大学の両センター間は CSI グリッド網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，東北大学，大阪大学，双方へジョブの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．WFT(Work Flow Tool)を選択すると，図 6 に示す WFT ウィンドウのユーザディレクトリに，今回の評価で用いる評価プログラム用の Work Flow アイコンを確認することができます．本稿における評価では実際に東北大学サイバーサイエンスセンターで実行されている SMP16 並列のプログラム (xxx-01, xxx-06, xxx-01-s)と MPI16 並列のプログラム（xxx-10）を用います．”xxx”は”SX9”が東北大学，”CMC”が大阪大学での実行のために用意した Work Flow アイコンを示しています．

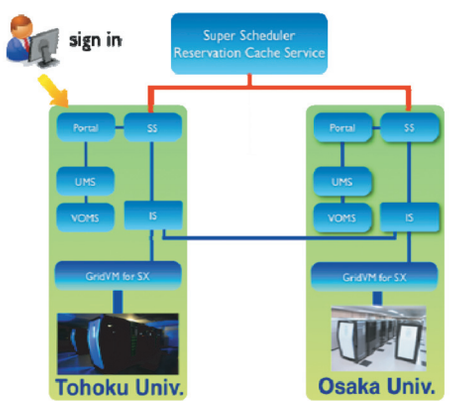

図 4 ：評価システム

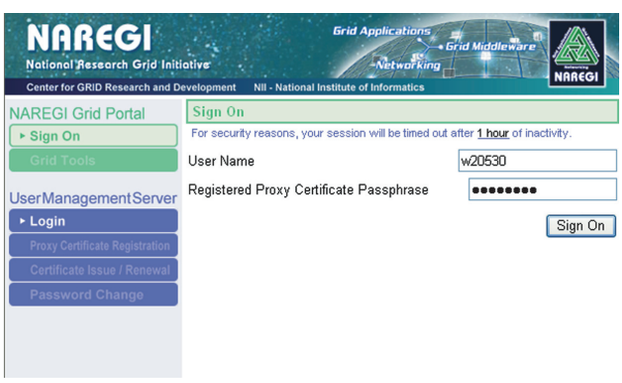

図 5　portal ログイン画面

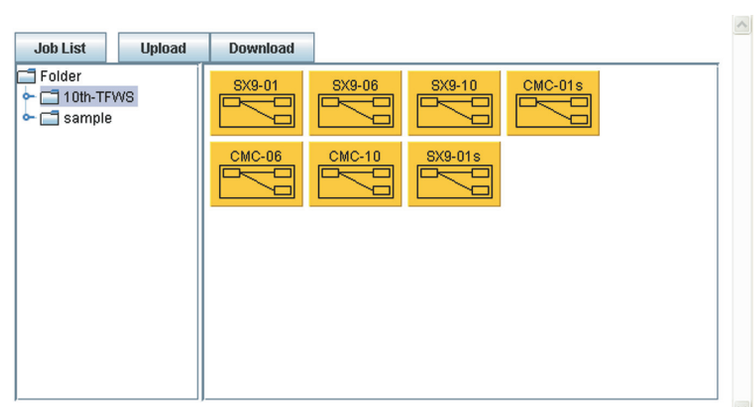

図 6　Work Flow Tool ウィンドウ

図 7：サイト情報の確認

各 Work Flow アイコンを開き，そのプロパティを確認することで図 7 に示す様に，利用できるサイト，各サイトのシステム基本情報，稼働状況を確認することが可能であり，ユーザ・スケジューラは適切なサイトでジョブを実行することが可能になります．図 7 の赤い箇所に着目すると東北大学と大阪大学の計算資源が利用できることが確認できます．また，青い箇所では大阪大学サイバーメディアセンターの SX-9 システム 1 ノード分の基本情報，緑の箇所では 1 ノードを構成する 16 個の CPU の負荷状況をそれぞれ確認することができます．このとき，大阪大学の SX-9 ノードの全 CPU が 1.6GHz(スカラ部)，ピーク性能に対して 97%の負荷で稼働しています．次に WFT から，実行する Work Flow アイコンを選択し，ジョブを実行すると図 8, 9 に示す様に各ジョブのステータスを確認することができます．このとき，大阪大学の計算資源が高負荷状態であるため，東北大学の SX-9 へジョブの投入を行いました．図 8 はジョブを実行中の状態を示し，図 9 はジョブが完了した状態を示しています．

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 3 つの SMP, MPI 16 並列全てのプログラムを当センター，大阪大学サイバーメディアセンターで実行し，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, MPI 並列プログラムの実行が可能であることを確認しました．

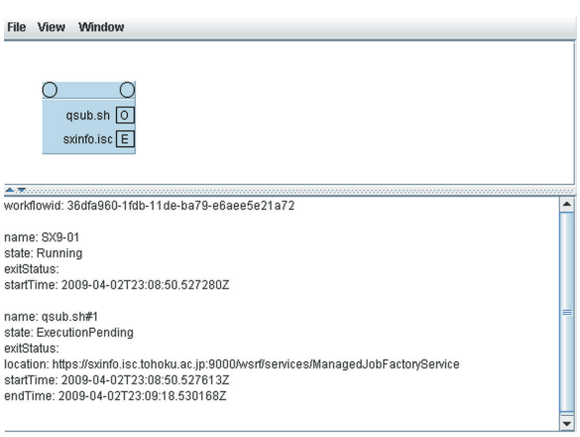

図 8 ジョブステータス(実行状態)

図 9 ジョブステータス（終了状態）

4. おわりに

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の第一歩として，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，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であることを示しました．今後は，詳細な性能評価，ノード間を跨ぐ超大規模並列処理の評価，および多数のベクトルコンピュータを有する組織との連携により，ベクトルコンピューティングクラウド基盤の構築に取り組んで行く予定です．併せて，portal に実行可能なジョブクラスのみを表示することで，ユーザが物理的に各サイトの計算資源を意識することなく，複数の計算資源を効率的に利用できる計算環境の開発を行いたいと考えております.

また，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では，2012 年度に稼働が予定されている理化学研究所の次世代スーパーコンピュータにおける効率的な大規模ベクトル演算実行，スカラ・ベクトル混在の大規模演算実行の要素技術の確立と言う視点に於いても，全国共同利用情報基盤センター群（7 センター，国立情報学研究所，筑波大学，東京工業大学）における本格的なグリッド環境構築[3]と併せて本研究を推進して行く予定です.

謝辞

本プロジェクトは，平成 20 年度国立情報学研究所 CSI 委託事業として行われました．また，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. 国立情報学研究所 リサーチグリッド研究開発センター合田憲人教授，　同 GOC グループ川井優様には 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.

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.

[3] 小林泰三, ”基盤センターにおけるグリッド連携,” 大阪大学スーパーコンピュータシンポジウム，2008 年 10 月.