［全国共同利用情報基盤センター研究開発論文集 No.33より］

ベクトル型スーパーコンピュータ広域連携基盤の性能評価

†山下　毅　‡村田善智　‡江川隆輔
†小野　敏　†大泉健治　‡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

平成 23 年 3 月 11 日 14 時 46 分 18 秒，宮城県石巻市牡鹿半島の東南約 130km 地点（三陸沖）の深さ約 24km を震源とする東北地方太平洋沖地震が発生した．本報告では，震災後の厳しい電力制限の下で，如何に現状の大規模科学計算システムを通常運用に復旧させたかについての経緯を述べる．また，震災等により限られた計算資源下においても安定した大規模計算サービスを提供するために，計算資源の運用形態の一つとして期待されるハイパフォーマンスコンピューティング (HPC) クラウドの開発状況と性能評価について報告する．

2　東北地方太平洋沖地震発生からスーパーコンピュータシステム復旧まで

3 月 11 日に発生した M9.0 の本震により，東北電力管内では青森県・岩手県・秋田県の全域，山形県・宮城県のほぼ全域と，福島県の一部で停電した．これに伴いセンターの全スーパーコンピュータシステムが停止した．居室の物品が床に散乱するなどはあったが，幸いなことにセンター関係者の人的被害は無く，建物の損傷も軽微なものであった．またスーパーコンピュータシステムも，今後発生するとされていた宮城県沖地震の対策として行っていた耐震施工が功を奏し，筐体の転倒や空調機の落下などは無かった．その後の点検で CPU・メモリ交換で対応可能な故障はあったが，運用に支障がない軽微なものであった．表 1 に 3 月 11 日の地震発生から 5 月 9 日の全スーパーコンピュータシステム復旧までの経緯を示す．

サイバーサイエンスセンターのある宮城県仙台市青葉区では 13 日の 14 時頃までに停電が解消し，運用再開に向けての作業を開始した．15 日には全スーパーコンピュータシステムの起動が確認されたが，震災直後の利用者の需要低下と節電対応のため部分運転を行っていた．その後，利用者のジョブの増加に対応するために 4 月 4 日までにはスーパーコンピュータの 55%，並列コンピュータの 33% が運用再開していたが，7 日夜に発生した余震による瞬停で，再度スーパーコンピュータの停止を余儀なくされた．

4 月 19 日にはスーパーコンピュータの稼働率が未だ 77% ではあるものの，センターの社会貢献の一環として，震災にかかわる復興研究を支援するために，災害・防災，安全技術，環境エネルギーに関する研究分野を対象として，演算負担額の全額または一部を免除する課題の募集を行った．5 月 9 日までに全スーパーコンピュータシステムが通常運用となり，現在に至っている．

表 1　災害・復旧状況の経緯

日時	災害・復旧状況など
3 月 11 日 14 時 46 分	地震発生 M9.0 仙台市青葉区で震度 6 弱を観測
14 時 48 分	停電により全スーパーコンピュータシステム停止
13 日 14 時頃	センターの電気復旧
15 日	全スーパーコンピュータシステムの起動を確認 節電対応のためログインサーバおよびファイルサーバのみの運用開始
22 日	飲料用給水タンクが空になる 空調用の給水を井戸水に切り替える
24 日	スーパーコンピュータ 2 ノード，並列コンピュータ 2 ノードの運用開始
28 日	センターの水道復旧 空調用の給水を水道水に切り替える
29 日	スーパーコンピュータ 4 ノードを追加運用開始
4 月　4 日	スーパーコンピュータ 4 ノードを追加運用開始
7 日 23 時 30 分	余震（M 7.4 震度 5 弱）により運用中のスーパーコンピュータ 10 ノードが停止
8 日	スーパーコンピュータ 10 ノード運用開始
15 日	並列コンピュータ 2 ノードを追加運用開始
18 日	スーパーコンピュータ 4 ノードを追加運用開始
19 日	特別復興研究支援課題の募集開始
5 月　9 日	スーパーコンピュータ，並列コンピュータ全ノードの運用開始

災害や電力供給力の問題により本センターのシステムが縮小運転される状況下において，北海道大学，名古屋大学，京都大学，大阪大学，九州大学の各情報基盤センターから可能な範囲での計算資源提供の申し出を頂いた．この支援計画は「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（HPCI)」構築の準備段階の活動として実施されたものである．この活動の一つとして，本センターの利用者が本センターの利用者 ID をそのまま使用して，大阪大学サイバーメディアセンターの大規模計算システムでのジョブの実行を可能とする，アカウント連携の導入を行った．

3　広域連携高性能基盤の概要

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はかねてより，計算資源を仮想化することで，ユーザが使用する計算資源数等を意識することなくジョブの投入が可能なサービスの提供，かつユーザジョブのターンアラウンドタイムの短縮と両センターが有する資源の高効率利用を目的に，ベクトルコンピューティングクラウドと呼ばれる広域連携高性能計算基盤の研究開発に取り組んできた．今回の被災を受けて，当センターではノード数を減らし

た部分運転を強いられたため，大規模なジョブクラスのサービスを制限せざるを得なかった．このような視点からみると，これまで取り組んできた広域連携による計算資源のシームレスな利用と，物理的に離れた計算資源の有機的連携による大規模計算サービスの提供は，サービスの向上や，資源の高効率利用，ジョブのターンアラウンドタイムの短縮ばかりでなく，災害時にも利用可能なディペンダブルな計算環境として活用する事が期待できる．

これまでのクラウドコンピューティングでは，ユーザジョブの実行に際して，計算資源のみを仮想化し，ユーザがクラウドの中にある計算資源を意識する事無く利用できる環境を提供するものであった．これは，クラウドを構成する計算資源は大規模データセンター等に集約されている場合が多く，計算資源問を結ぶネットワークの遅延をそれほど考慮する必要がないためである．しかし，高性能計算のアプリケーションを対象とするハイパフォーマンスコンピューティングクラウド (以下：HPC クラウド) では，現存するスーパーコンピュータの計算資源が限られていること，分散する各計算資源の規模を超えた計算の実行が期待されていることから，物理的に分散した計算資源を高速なネットワーク網で結合し有機的に連携させることで，クラウドコンピューティングサービスを提供する必要がある．

このため，HPC クラウドには，ユーザが複数の計算資源をシングルサインオンで，かつ物理的に計算資源を意識することなく利用可能な環境が求められている．このような状況下で，我々は大阪大学サイバーメディアセンターと共同で”ベクトルコンピューティングクラウド”と呼ばれる計算基盤の構築に取り組み，プロトタイプシステムを構築している．このプロトタイプシステムでは，ユーザは Web インターフェースからシングルサインオンでシステムを利用可能で遠隔にあるベクトルコンピュータ双方の計算資源を利用する MPI プログラムの実行を可能にしている．クラウド基盤として，高性能計算をサービスとして提供できることは確認できたものの，東北大学，大阪大学問を結ぶ 800 ｋｍにもおよぶネットワーク性能・帯域が性能に与える影響は大きい．

現在の IP ネットワークでは，ネットワークに点在するルータが経路を決定し，かつその経路は１つに定まっている．このため，ネットワークが有する通信性能 (帯域) を，ネットワークを利用している全てのサービスで共有することになる．また，ネットワーク内部の接続構成は実質，動的に制御不能で信頼性・安全性を保証することができず，ベストエフォートで粗い経路制御しかできない．加えて，複雑な制御機能を有するルータが数多く点在し， IP ネットワークにルータの対故障性や，消費電力の増加が問題視され始めているにも関わらず，これらの管理維持コスト，電力コスト等が非常に高いものとなっている．

これらの既存の数々の問題を解決する新世代ネットワーク基盤技術として，近年 Open Flow Network が注目を集めている．Open Flow Network では既存の IP ネットワークにおいて，ルータが持つ複雑な制御機能をネットワークの外に分離して，制御サーバとして集約することで，ネットワーク全体の制御を集中的に管理することが可能になる．これにより，制御サーバがネットワークを瞬時に構築し，素早くかつ動的に最適化することが可能になるばかりでなく，ネットワークの管理維持コスト，消費電力等を大幅に削減することが期待できる．このような状況下で，現在我々は，計算資源だけではなく，計算資源問を結ぶネットワークも同時に仮想化し，ユーザのアプリケーションが求める計算資源・ネットワークをテーラーメードで提供できる技術として Open Flow Network 技術に着目し，図１に示す HPC クラウドの開発に取り組んでいる．

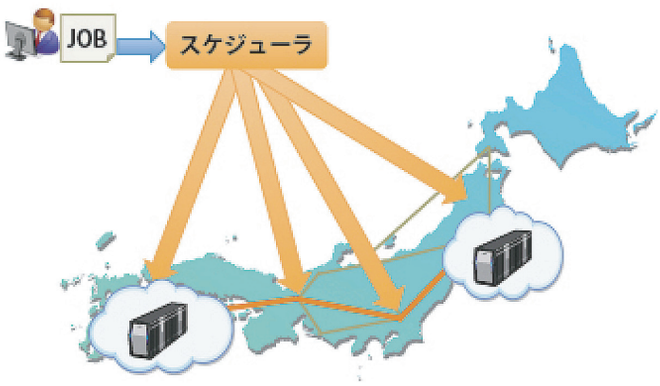

図 1　最適経路選択による HPC クラウドの実現

我々が目指す HPC クラウドでは，実行されるアプリケーションが求めるネットワーク性能を適宜判断し，必要なネットワークをアプリケーションに応じて割り当てる事が可能になる．これらの取り組みと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が開発している実行履歴に基づくジョブスケジューリング機構と連動させることで，高効率かつ，短いターンアラウンドタイムを実現する広域連携計算基盤の構築が可能であると考えている．

次節では，これらの前述の高度化の足がかりとして，HPL（High Performance Linpack）による広域連携計算基盤のネットワーク性能評価と，実アプリケーションを用いた広域連携基盤の基本性能評価について述べる．

4　性能評価と今後の課題

4.1　HPL ベンチマーク

本節では，高性能計算資源の計算性能を評価するために HPL を用いた，広域連携環境の性能評価を示す．評価では，単一拠点の計算資源を用いた環境 (inhouse) と，複数拠点の計算資源を広域連携させた環境 (wide area) において評価を行った．表 2 に示す 6 通りの環境において，inhouse 環境は IXS および 1GbE を用いたネットワーク環境，wide area 環境は SINET3 および JGN2Plus を用いたネットワーク環境である．また MPI 実装は評価に用いた MPI の環境であり，MPI/SX はベンダー提供の MPI 環境，GridMPI はベンダー提供の MPI 環境をラッピング関数を用いて仮想化した環境，GridMPI/IMPI は GridMPI と遠隔 2 拠点間での MPI 通信を実現する IMPI の 2 つを組み合わせた環境である．

図 2 に，それぞれの環境における HPL の性能評価の結果を示す．図 2 より，高速なネットワーク性能を持つ MPI/SX(IXS) と GridMPI(IXS) の 2 つの環境で高い性能を示している．一方，ネットワーク性能が限られる，GridMPI(SINET3) と GridMPI(JGN2Plus) の広域連携環境でも，問題サイズに比例した性能の向上が確認できる．特に，10Gb の転送性能を持つ JGN2Plus を用いた GridMPI (JGN2Plus) の環境では，実行効率が 51.3% を達成しており，ネットワーク性能が限定される広域連携環境においても高い計算性能が得られることが示された．

4.2　実アプリケーションを用いた基本性能評価

実アプリケーションとして理化学研究所・計算科学研究機構 松岡浩教授が開発した 2 次元格子ガス法コードを用い，本計算基盤の評価を行った．本計算基盤を用いて 2 次元格子ガス法コードを実行した時の計算結果を可視化したものを図 3 に示す．図 3 から確認できるように，本計算基盤を用いた場合でも，流体の挙動を正確に計算することができている．本評価では，本計算基盤においてもローカル環境と同様に大規模 MPI プログラムを実行可能であることを確認すると共に，遠隔地にある計算機間のデータ通信量が実行性能に与える影響を明らかにし，通信量を削減する最適化を施すことによって実行効率を向上できることを確認した．

表 2　HPL 評価環境

名称	MPI 実装	ネットワーク
MPI/SX (IXS)	MPI/SX	IXS(64GB/s)
GridMPI (IXS)	GridMPI	IXS(64GB/s)
GridMPI (JF)	GridMPI	1GbE(Jumbo Frame)
GridMPI (NF)	GridMPI	1GbE(Normal Frame)
GridMPI (SINET3)	GridMPI/IMPI	SINET3(1GbE)
GridMPI (JGN2Plus)	GridMPI/IMPI	JGN2Plus(10GbE)

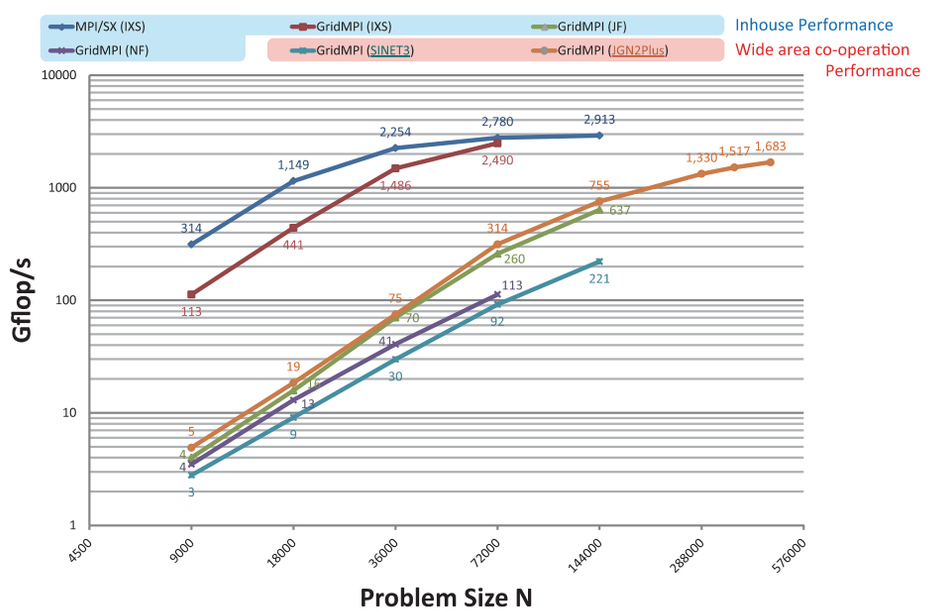

図 2　HPL 評価結果

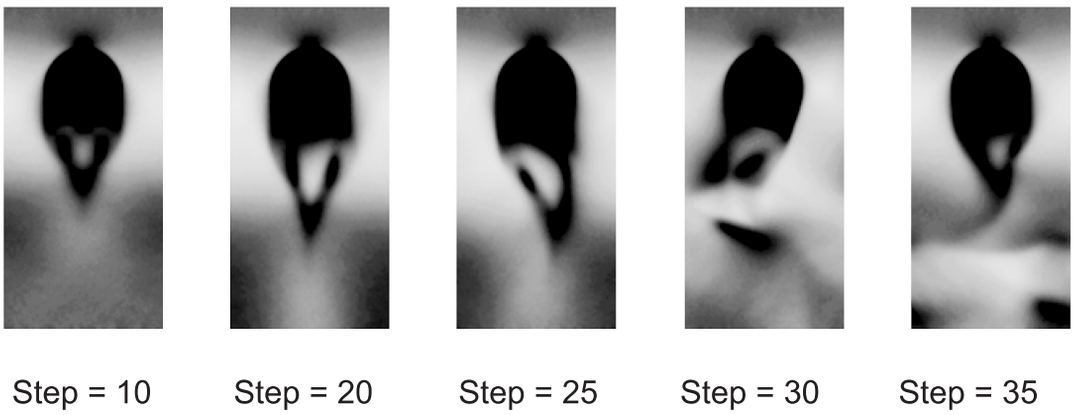

図 3　本計算基盤を用いて得られた円柱周りの流体の挙動

5　おわりに

本稿では，震災後から現在に至るまでの当センターの大規模科学計算システムの運用状況と，HPC クラウド構築に向けたスーパーコンピュータ広域連携の性能評価について述べた．評価の結果，800km 離れたスーパーコンピュータを連携させることで，51.3% と高い実行効率を実現できることが明らかになった．今後の課題としては，アプリケーションが必要とする計算性能，ネットワーク性能の明確化と，アプリケーションの要求する計算・ネットワーク資源を仮想化してユーザに提供するシステムの構築が挙げられる．

謝辞

震災直後いち早く支援のお申し出をいただいた各情報基盤センターの関係者各位に深く感謝する．大阪大学との連携実験において支援頂いた大阪大学サイバーメディアセンター 朿田学助教，技術的支援を頂いた日本電気株式会社の皆様，2 次元格子ガス法コードを提供して頂いた理化学研究所・計算科学研究機構 松岡浩教授，余震の続く中スーパーコンピュータシステムの復旧に尽力して頂いた NEC フィールディング株式会社の皆様，日本電気株式会社の皆様に深く謝意を表する．また広域連携高性能計算基盤の研究は，国立情報学研究所受託事業「最先端学術情報基盤の構築を推進する事業」の成果である．