多様化する計算要求に柔軟に対応できる計算基盤の実現に向けて

伊達 進 吉川隆士 高橋雅彦+ 菅 真樹+ 渡場康弘++ Chonho Lee 木戸善之 下條真司

大阪大学サイバーメディアセンター
+NEC システムプラットフォーム研究所
++奈良先端科学技術大学院大学
{ date, tyoshikawa, leech, kido, shimojo}@cmc.osaka-u.ac.jp
{m-takahashi@ex, kan@bq}.jp.nec.com
watashiba@is.naist.jp

概要：プロセッサ、アクセラレータを始めとする計算資源の多様化に伴い、ユーザの計算要求もまたますます多様化しつつある。そのような観点から、大阪大学サイバーメディアセンターと日本電気株式会社は、ユーザの様々な計算要求に柔軟に対応できる計算基盤の実現にむけ共同研究開発を推進している。本稿では、当該共同研究開発の背景と目的、現状、およびその課題について記す。

1 まえがき

今日、高性能計算をとりまく計算資源の多様化が進展している。例えば、プロセッサ資源を例にとれば、Intel Xeon、Intel Phi、ARM 等の利用が可能である。さらに、近年のプロセッサ資源においては、その搭載しているコア数、動作周波数もまた多様になっている現状がある。また、アクセラレータを例にとれば、GPU、FPGA 他、各種専用ボードも利用可能である。このような計算資源の多様化は、ユーザによるプログラム開発環境をもまた急速に多様化しつつある。実際、大阪大学サイバーメディアセンターの利用者は、それぞれが各自のプログラム開発環境を有しており、それぞれが異なるプロセッサコア数を搭載したプロセッサをターゲットとしたノード内並列処理、異なるノード数をターゲットとしたノード間分散処理、および、異なるアクセラレータ利用した高性能プログラム開発を行っているのが現状である。このようなユーザのプログラム開発環境の多様化は、ユーザが自身の開発環境で開発したプログラムを本センターの大規模計算機システム上で高い並列度で大規模に実行しようとする際、本センターの大規模計算機資源の計算環境に合わせて改めてプログラムを書き直さなければならないなどのユーザ負担を発生させる。

大阪大学サイバーメディアセンターでは、上述の背景と問題点から、利用者の多様な開発環境で開発されたプログラムをシームレスに収容できる大規模計算環境の必要性に着眼し、ユーザの計算資源要求を柔軟に対応できる計算基盤の実現にむけ、日本電気株式会社との共同研究を推進している。

本稿では，当該研究開発の背景と目的、現状、およびその課題について報告する。本稿の構成は、以下の通りである。2 節では，まず本稿で記す計算基盤が前提とする本センターの大規模計算機システムの特性について紹介する。その後，3 節において，本センターが研究開発をする、ユーザの計算資源要求を柔軟に対応できる計算基盤について概説し、4 節では今後の課題について記す。5 節で本稿をまとめる。

2 大規模可視化対応 PC クラスタ (VCC)

本節では，本稿で報告するユーザの計算資源要求を柔軟に対応できる計算基盤が、その前提とする大規模可視化対応 PC クラスタ(以下、VCC)の再構成可能特性について概説する。

図 1 に VCC の概要を示す。VCC は、総計 65 ノードが InfiniBand FDR で相互接続されたクラスタシステムである。各ノードは Intel Xeon

E5-2670v2 プロセッサ 2 基、64 GB の主記憶を搭載している [1]。この VCC 上では、NEC 独自開発の NEC Scalable Technology File System (ScaTeFS)とよばれる高速・分散並列ファイルシステムを通じて、2 PB のストレージにアクセス可能な構成となっている。

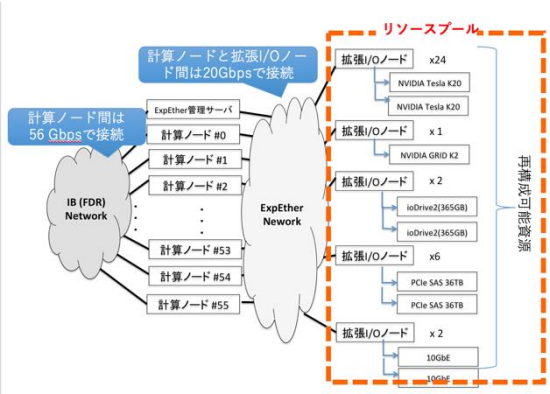

図 1：VCC の概要.

本システムの特徴は、イーサネット上で PCI Express を仮想化する技術、すなわち、システムハードウェア仮想化技術 ExpEther [2] を導入している点にある。この仮想化技術を活用することで、本システムでは、 GPU などのアクセラレータ資源、SSD、ストレージなどのディスク資源を任意のノードに容易に割り付ける(再構成する)ことができる柔軟性を提供している。例えば、あるユーザから 3 枚の GPU 資源を割り当てた計算ノードを 5 ノード構成してほしいという計算要求があった場合においても、また、別ユーザが 1 枚の GPU 資源を割り当てた計算ノードを 3 ノード、また、そのうち 1 台には SSD を割りつけた計算ノードを構成してほしいといった計算要求があった場合に対しても、本システムの物理的な接続構成を変更するのではなく、 ExpEther 管理ノード上で設定を変更するだけでよい。

本センターでは、リソースプールに登録された任意の再構成可能資源を任意の計算ノードに割りつける（再構成する）ことができる特性を活用して、大阪大学をはじめとし HPCI、JHPCN を中心とした全国の研究者に対し、それぞれの計算要求に柔軟に対応した計算機サービスを提供してきた。しかし、上述したシステムハードウェア仮想化技術 ExpEther を活用した、ユーザの計算要求への柔軟な対応は、本センターの技術職員がユーザからの要求をヒアリングした上で、 VCC の運用支援を担当する日本電気株式会社の担当 SE に連絡し、担当 SE が手動で ExpEther 管理サーバの設定を変更することにより実現されている。そのような問題点と経験から、本センターでは、本システムの運用支援を担当する日本電気株式会社との共同研究を通じて、 VCC の特性を活用しつつ、ユーザの多様な計算要求を収容できる柔軟な計算基盤の実現を目指した取り組みを推進中である。

3 提案計算基盤

3.1 概要

前節で述べた通り、VCC システムは、インターコネクションレイヤ（ExpEther）で、計算機自体のハードウェア構成を管理ソフトウェアで変更する仕組みを備えている。現在、このハードウェアレベルのシステム再構成を、計算機 JOB のスケジューラと連動させ、JOB ごとに必要な計算機構成をダイナミックに構築する検討を行っている。これにより、従来は固定的なサイバーメディアセンターをはじめとする計算機センターや、クラウドデータセンターなどで、アクセラレータや大規模分散コンピューティングなどの様々な高性能な計算機環境を、ユーザの要求に応じて柔軟に提供する事が可能となる。

3.2 背景

今後の計算機センターにおいては、高性能な GPU コンピューティングや MPI を用いた分散並列計算機など多様な環境を提供する必要がある。しかし、それらの高性能計算機の実現のためには、現在、計算機センター内部で大量に用いられている標準的なサーバが使えない。それは、例えば高性能 GPU を搭載するために巨大な拡張機構や大電源や強冷却機構が必要だったり、あるいは高性能

クラスタ構成には InfiniBand などの高速なネットワークアラプターが必要であるからである。これらは高価であり、かつ大量の電力を消費してしまうだけでなく、計算機の構成があらかじめ固定されてしまう問題もある。例えば GPU が 2 台搭載された GPU マシン、あるいは、 Infiniband で密結合された 16 台の PC クラスタ、というように、ある特定の計算機構成を、システム導入時点で始めから作りこんでおく必要がある。一方で、今日のユーザの要求は多様化しているために、このように計算機の構成が固定されてしまうと、ユーザにとっては、自分のやりたい計算に最適な構成が自由に選べない。また、計算機センター運用側にとっては、様々なユーザの要求を満たす計算機構成を個別に用意し、 JOB が走っていないときも維持しておくことになす。結果としてリソースの利用効率が極端に低くなってしまう。

これに対して、 VCC ではシステムの構成要素であるサーバの数、ネットワーク接続、データストレージ、さらにはサーバを構成する GPU や SSD などのハードウェアデバイスまで、管理ソフトウェアで構成を簡単に変更できる計算機システムを開発した。これに加えて、計算 JOB を投入する際に、投入される JOB の性質や必要な機能・性能にあわせてシステムの構成まで変化できる JOB ・リソース連携管理マネージメントシステムである。

これにより、ユーザの投入した JOB で必要な計算機システムをダイナミックに構築して計算リソースを提供できるようになる。さらには単一の JOB だけでなく、クラウドに流れる JOB 全体の待ち時間を減らして効率よく流せるようになると考えている。

3.3 システム構成と動作

現在のサイバーメディアセンターではクラスタ、ベクトル、 VCC への JOB 投入は一元化されており、ユーザがパラメータとしてシステムを指定する形になっている。これを継承する形で JOB 投入からのフローとそれを実現するアーキテクチャを考えた。これは、図２に示す通り、JOB スケジューラを頭に、リソースマネージャ、ネットワークマネージャ、ソフトウェアプラットフォームマネージメントが連携したアーキテクチャである。

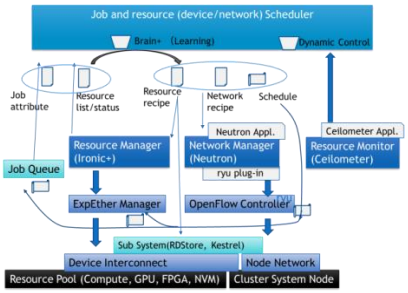

図２　JOB～リソース連携管理のアーキテクチャ.

以下のフローで JOB が実行される。

・最初にユーザが JOB を投入する。その際に利用するシステム、すなわち、アクセラレータやストレージなどのマシン構成やネットワーク構成、さらには Spark などの分散システムやクラスタ MPI システムを用いる場合はそのノード構成の要求などを指定する。

・スケジューラ内の JOB ～リソース連携制御部（Brain）で、現時点以降の JOB スケジュール、及びリソース利用状況を考慮した上で、実際にその JOB を流す場合のリソース・ネットワーク構成をレシピとして JOB に関連づけしたうえでスケジューラのキューに投入する。

・ JOB が投入される際に、 OpenStack ベースのリソース管理ソフトウェアが、アクセラレータなどのハードウェアをリソースプールから選択してインターコネクションで結合し、ハードウェア再構成を行う。それと同時に必要なネットワークを設定し、 JOB を処理するためのリソース群（パーティション）を設定する。

・ OpenStack ベースで、 OS や分散システムミドルウェアなどのソフトウェアプラットフォームの入ったシステムイメージを導入する。

・スケジューラの指示に沿ってパーティション上で JOB を実行する。

・ JOB 実行後、不要になった GPU アクセラレータなどのデバイスはシステムから切り離し、他の

JOB で使われるまでリソースプールで待機させる。

3.4　現状

図３に示す通り、コンピュートノードが 4 台の小規模な構成で上記のフローを実現するプロトタイプを実装した。このシステムでは JOB を投入することで、その処理に必要なハードウェアが再構成され、さらに分散ミドルウェアやストレージソフトの入ったシステムイメージがインストールされる。

図３プロトタイプ.

図４は動作の例を示す。ここで投入される Job は spark 上で走る Logistic Regression 処理である。JOB 投入前のハードウェア構成（図４左側）では CPU ・メモリからなるコンピュートノードに対して、 ExpEther で結合された外部共有 NVMe ストレージデバイスが共有されている。この状態に対して GPU アクセラレータを用いる指定をした JOB が、スケジューラのキューに登録され実行される際に図４右側に示す通り両方のノードに GPU が組み込まれた GPU クラスタシステムが構築されている。

図４　JOB～リソース管理連携動作.：JOB の投入に連動して、GPU が 2 個、追加された状況を示している。

以下に主な技術要素を示す。本システムには UCLA(University of California Los Angeles)、奈良先端大、NEC、Falcon Computing Solutions の研究開発成果が含まれている。

（１） JOB・ネットワークリソース連携管理ソフトウェア（大阪大学サイバーメディアセンター、奈良先端大学）ユーザからの JOB を一元的に受け付けるスケジューラ。指定されたリソース条件に対して最適な JOB 実行環境をリソース管理と連携して整えたうえで JOB を実行する

（２） Kestrel ： YARN をベースとした GPU ・FPGA アクセラレータを用いたヘテロコンピューティング用の JOB・リソース連携管理ソフトウェア（UCLA、Falcon Computing System）

（３）RDStore

リソース分離型計算機（Disaggregated Computing Architecture）の利点を利用した高速分散 kvs ストレージソフトウェア。安価で大容量な SSD を用いながら DRAM ベース分散ストレージより高性能、かつ、スケールアップが容易に行える特徴を有する [2]。

（４）EEStack：OpenStack の Ironic をベースに、管理対象をベアメタルだけでなく、サーバを構成する PCIe デバイスレベルまで拡張したリソース管理ソフトウェア。ハードウェアレベルのダイナミックなサーバ再構成技術と OpenStack ベースのリソース管理技術（NEC、ExpEhter コンソーシアム）の連携で実現 [3]。

4 今後の課題

現在までのところ、JOB の特性、要求にあわせて柔軟に構成を変更することで、多様な高性能コンピューティイングの要求に対して最適なシステム構成を効率よく提供するための、アーキテクチャと小規模なコンセプト試作ができた。今後は、実際のサイバーメデイァセンターでの運用実績を基に、本システムの有用性を評価していく必要がある。また、実際のアプリケーションに基づいて、その高速処理を実現するソフトウェアプラットフォームや手法も含めて、利用者に提供できるような仕組みを実現していく

5 まとめ

本稿では、多様化する計算要求に着眼し、大阪大

学サイバーメディアセンターが日本電気株式会社の共同研究を通じ研究開発を行う、ユーザの計算資源要求を柔軟に対応できる計算基盤について報告した。今後は現行の VCC システムの運用位実績をもとに本システムの有用性を示すと同時に、実際の具体的なアプリケーションに対しての、適用の仕方の検討と課題抽出を行って、本技術のシステム実用化へと発展させていく。

参考文献

[1] 大規模可視化対応 PC クラスタ(VCC), http://www.hpc.cmc.osaka-u.ac.jp/vcc-sys/.

[2] 菅　真樹, 鈴木　順, 林　佑樹, 吉川隆士, 宮川伸也: リソース分離アーキテクチャにおける分散キーバリューストア, 信学技報, vol. 115, no. 174, CPSY2015-31, pp. 167-172, 2015 年 8 月.

[3] 吉川 隆士, 菅 真樹, 高橋 雅彦, 宮川 伸也, 飛鷹 洋一, 阿部 晋樹: スケールアップにより多種多様なコンピューティングを実現する Resource Disaggregated Platform, NEC 技報 Vol67. No.2(2015/3)