様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １７ 日現在

機関番号：１３９０１
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８００４３

研究課題名（和文）高性能計算リソースの抽象化を実現するランタイムシステム

研究課題名（英文）Runtime Systems with Abstraction of High-Performance Computing Resources

研究代表者
加藤　真平（Kato, Shinpei）

名古屋大学・情報科学研究科・准教授

研究者番号：７０６３１８９４

交付決定額（研究期間全体）：（直接経費）　　13,800,000円

研究成果の概要（和文）：本研究課題では、計算科学アプリケーションにおけるプログラミングおよびデータ管理の簡易化を目的とし、（１）メニーコアアクセラレータの仮想化（プロセッサ抽象化）、（２）ソフトウェア定義ネットワークの仮想化（ネットワーク抽象化）、（３）科学データのデータベース化（データ抽象化）の機能を有する高性能計算クラスタ向けランタイムシステムの研究開発を行った。交通流シミュレーションとDNA解析を応用事例として、提案技術の汎用性ならびに仮想化やデータベース化による性能への影響を評価した。

研究成果の概要（英文）：This project aimed at simplifying programming and data management of scientific applications. We developed runtime systems for high-performance computing clusters providing (i) virtualization of many-core accelerators (processor abstraction), (ii) virtualization of software-defined networks (network abstraction), and (iii) database encapsulation of scientific data (data abstraction).

研究分野：並列分散システム

キーワード：高性能計算　データベース　抽象化

１．研究開始当初の背景

計算科学におけるシミュレーションやデータ解析は、科学技術の発展に不可欠な学問である。その基盤となる高性能計算クラスタ技術は、我が国が世界に誇る情報技術の１つであり、今後のイノベーションを促すために極めて重要な研究分野といえる。

スーパーコンピュータの性能はペタスケールに達しており、今後のエクサスケールに向けた取り組みも既に始まっている。一方、計算科学に関する SDHPC の報告によれば、多くの計算科学アプリケーションは小・中規模のジョブを投入する傾向にあるため、大規模なスーパーコンピュータに見られる課題の解決だけでなく、計算機システムと計算科学にかかる根本的な問題にも目を向ける必要がある。特に計算機アーキテクチャのパラダイムシフトに対応することは容易ではなく、最先端の高性能計算クラスタの恩恵を受けられない計算科学の研究者も少なくない。たとえば、アクセラレータと呼ばれるプロセッサ技術が確立され、従来の汎用プロセッサ（CPU）に比べて 10 倍以上のスケールでプログラムの高速化が可能になったが、このアクセラレータと汎用プロセッサが混在する高性能計算クラスタを使いこなせる計算科学の研究者は極めて限定的である。その主な要因はシステムの「見え方」に一貫性がないことである。すなわち、新しい技術の登場により潜在的な性能改善は達成されるが、同時にシステムの構成やプログラミング方法も変わってしまい、結果として過去の資産を流用できない状況を招いている。このような技術革新による使い手との隔たりは、ネットワーク管理やデータ管理においても問題となっており、今後は計算機システムと計算科学の双方に歩み寄りが求められる。

本研究では図１に示すように、高性能計算クラスタに対するプログラミングを簡易的に抽象化できるシステムを想定し、高性能計算リソースの管理手法に関する研究に着手した。

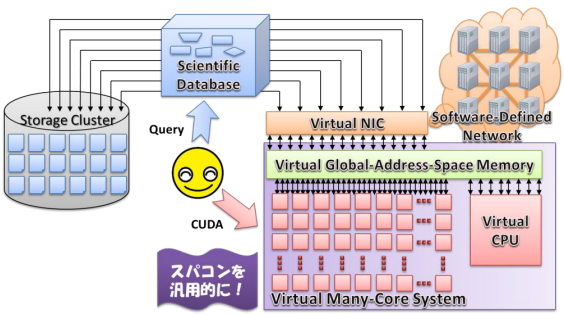


図１　本研究で想定するシステム

２．研究の目的

計算科学アプリケーションにおけるプログラミングおよびデータ管理の簡易化を目指し、以下の３つの機能を有する高性能計算クラスタ向けランタイムシステムの研究開発を行った。

・メニーコアアクセラレータの仮想化（プロセッサ抽象化）
・ソフトウェア定義ネットワークの仮想化（ネットワーク抽象化）
・科学データのデータベース化（データ抽象化）

これらの機能により、小規模から中規模の計算科学アプリケーションの生産性の飛躍的な向上が期待できる。本研究では、交通流シミュレーションと DNA 解析の２つの異なる計算科学の分野を対象とし、提案技術の汎用性ならびに仮想化やデータベース化による性能への影響を評価した。

３．研究の方法

本研究では、オペレーティングシステムとユーザプログラムの双方にランタイムシステムを導入し、プロセッサ、ネットワーク、データの管理を抽象化するレイヤを作った。リソース管理に関する各機能のダイアグラムを図２に示す。

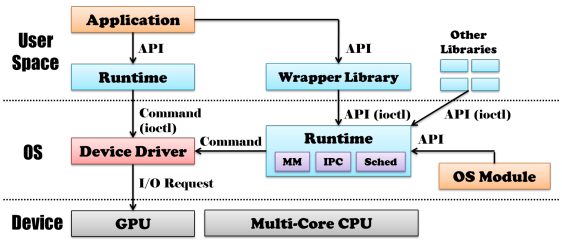


図２　本研究のリソース管理モデル

【プロセッサ抽象化】

本研究では、高性能計算クラスタの単一ノードには複数の GPU が搭載され、各 GPU は数百～数千コアを集積しているシステムを対象とした。従来は CUDA や MPI、OpenMP といった種々のプログラミング言語とライブラリを用いて CPU と GPU を協調させる労力が生産性を損なう要因となっていた。本研究では、複数の GPU を１つの巨大な GPU を搭載するシステムに仮想化する技術を目指した。特に GPU プログラミング言語である CUDA に焦点を絞り、ユーザが要求する仮想 GPU のコア数やメモリ量と物理 GPU のコア数とメモリ量をマッピングする OS、仮想化技術の研究開発を行った。主に NVIDIA 社製の GPU を利用し、オープンソースの GPU ドライバである Gdev やハイパバイザである Xen をベースとしたシステムソフトウェアの設計と実装を行った。

【ネットワーク抽象化】

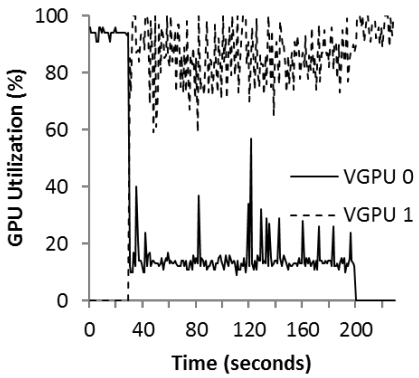
高次元トポロジの仮想化においては、計算ノード間の数 KB 未満のデータ通信が性能の鍵を握っており、通信速度には物理的な上限

があるため、限られたネットワーク資源のもとでデータ通信の遅延を最小化することが重要課題となった。本研究では、ネットワークに内包される高次元トポロジそのものを仮想化し、アプリケーションの特性に応じたツリー型やトーラス型のトポロジを構築し、さらにデータ通信の遅延を最小化する仕組みを研究開発する。これにより、高次元トポロジを抽象化しつつ、ネットワーク性能を最大限に引き出す技術の研究開発を行った。

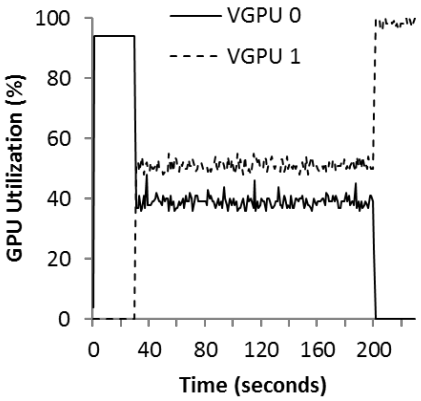
【データ抽象化】
統計解析やマイニングにおける科学データ管理を容易化するためのデータ管理技術に焦点を充てた。科学データ管理に関するこれまでの研究成果を基に、科学者にとって不評極まりないリレーショナルデータモデルではなく、多くの科学データの適切な表現である多次元配列を基礎とするアレイデータモデルによりデータの抽象化を実現した。特にアレイデータモデルの中のウィンドウ集約処理を事例として、差分計算の概念などを導入しつつ、処理の高速化を実現した。

４．研究成果

プロセッサ抽象化については、提案当初は複数ノードに跨る GPU 仮想化を目標としていたが、GPU の構造とデバイスドライバの実装の制約上、ページフォールトのハンドリングが困難であることがわかり、ノード内 GPU の仮想化に焦点を絞って研究開発を進めた結果、これまで１つのアプリケーションが GPU を占有するプログラミングモデルが主流であったのに対し、仮想化を通じて、複数のアプリケーションが GPU を効率的に利用できるメカニズムを創出することができた。本成果により、GPU を CPU のような計算リソースとしても利用することができ、より多くの科学者に GPU の恩恵がもたらせると期待する。
ネットワーク抽象化については、GPU の存在を意識しない研究を進め、ネットワークスイッチで相互接続した典型的なネットワーク構成に対して、スイッチのルーティングテーブルのエントリをリネーミングするアプローチにより、マイクロ秒単位の切替を可能にすることで、ユーザ定義のネットワークを超高速に実現できる機構を構築した。
データ抽象化については、アレイデータベースのウィンドウ集約処理を高速化するための再帰的インクリメンタル法を提案し、既存の SciDB に対して 60 倍以上の高速化を達成した。
本研究成果の一例を図 3 に示す。高性能計算リソースを抽象化し、複数のアプリケーションにプロセッサやネットワークといったリソースの使用を許可すると、これまでのランタイムシステムでは図３（１）に示すように、各々のアプリケーションのリソース使用が競合してしまい、サービスの質を確保できなくなってしまった。一方で、本研究成果を適用することで、図３（２）に示すように、一定のリソース使用量を維持することが可能になり、結果としてシステムのサービスの質を確保することができるようになった。今後、GPU やメニーコア技術が発展し、ネットワークトポロジも多次元化されていくことを考慮すると、システムのリソース管理は更に複雑化していくことが考えられるが、本研究成果はそれら今後の課題を解決するための基盤となる技術として期待できる。



（１） 従来の仮想マシンによる管理



（２） 本研究の抽象化による管理

図３　本研究成果の効果の一例

これらの研究成果により、GPU のようなメニーコアアクセラレータを簡易化し、ネットワークやデータ管理も抽象化しつつ、既存手法よりも高速に処理できる新規技術を創出した。これまでは、抽象化を行うと性能が損なわれるというトレードオフがあったが、本研究成果によりこのトレードオフが緩和され、複雑な高性能計算システムの抽象化アプローチが実用に一歩近づいたといえる。
本研究成果はオープンソースとして公開し、第 3 者にも利用可能な環境を提供している。オープンソース化することにより、第 3 者でも本研究成果を再現できることが可能になり、今後の本研究分野の技術発展の促進につながることが期待できる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

1. Y. Suzuki, S. Kato, H. Yamada, and K. Kono, GPUvm: GPU Virtualization at the Hypervisor, IEEE Transactions on Computers, pp. 1-14, 2016.**（査読有）**

〔学会発表〕（計 19 件）

2. J. Li, H. Kawashima, and O. Tatebe, Recursive Incremental Computation on Efficient Window Aggregate over Array Database, **情報処理学会研究報告システムソフトウェアとオペレーティングシステム，理化学研究所計算科学研究機構**, 2016 **年** 2 **月** 29 **日**－3 **月** 1 **日**. **（査読有）**

3. M. Koibuchi, Singularity of Future Computer-System Networks, ACM International Symposium on Information and Communication Technology, Hue, Vietnam, December 3-4, 2015. **（招待講演）**

4. S. Kato, J. Aumiller, and S. Brandt, Zero-Copy I/O Processing for Low-Latency GPU Computing, ACM/IEEE International Conference on Cyber-Physical Systems, Philadelphia, U.S., April 8-11, 2013. **（査読有）**

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

Gdev, https://github.com/cpfl/gdev
GPUvm, https://github.com/cpfl/pvdrm

６．研究組織
(1)研究代表者

加藤真平（KATO, Shinpei）
名古屋大学大学院情報科学研究科・准教授
研究者番号：７０６３１８９４

(2)研究分担者

鯉渕道紘（KOIBUCHI, Michihiro）
国立情報学研究所アーキテクチャ科学研究系・准教授
研究者番号：４０４１３９２６

川島英之（KAWASHIMA, Hideyuki）
筑波大学大学院システム情報工学研究科・准教授
研究者番号：９０４０７１４８

(3)連携研究者

藤原一毅（FUJIWARA, Ikki）
国立情報学研究所アーキテクチャ科学研究系・特任研究員
研究者番号：９０６４８０２３

油井誠（YUI, Makoto）
独立行政法人産業技術総合研究所情報技術部門・研究員
研究者番号：１０５８６７１２

大野欽司（OHNO, Kinji）
名古屋大学大学院医学系研究科・教授
研究者番号：８０３９７４５５

杉山雄規（SUGIYAMA, Yuki）
名古屋大学大学院情報科学研究科・教授
研究者番号：２０１９６７７８

石井克哉（ISHII, Katsuya）
名古屋大学情報連携基盤センター・教授
研究者番号：６０１３４４４１