平成 25 年 11 月 1 日

報道関係者　各位

国立大学法人筑波大学
クレイ・ジャパン・インク
エヌビディア ジャパン

スーパーコンピュータ「HA-PACS」を拡張－1 ペタフロップスを超える性能に

概要

筑波大学計算科学研究センターは、スーパーコンピュータ「HA-PACS」に新規開発した「TCA 機構」搭載部を拡張し、これまでの総ピーク演算性能 802 テラフロップス（毎秒 802 兆回）から 1.166 ペタフロップス（毎秒 1166 兆回）に増強させたスーパーコンピュータの運用を開始しました。

「HA-PACS」は平成 24 年 2 月 1 日に運用を始めた、宇宙・素粒子・生命などの研究をけん引する最先端の超並列演算加速器クラスタ型スーパーコンピュータです。268 台の計算ノードからなるベースクラスタシステムに、この度、同センターで開発した密結合並列演算加速機構（TCA 機構）を装備した 64 台の計算ノードを追加しました。その結果、364 テラフロップスの演算性能が増強され、システムの総ピーク演算性能は 1.166 ペタフロップス（毎秒 1166 兆回）となりました。筑波大学として初めて 1 ペタフロップスを超えるシステムです。

TCA 機構は、GPU*1 を搭載した PC クラスタシステムの大きな問題であった、遠隔 GPU 間の通信性能の低さを改善する画期的な機構です。独自開発の通信用チップにより、これまでできなかった異なるノード上の GPU 間の直接通信を実現。通信時間を大幅に短縮させて、GPU クラスタにおける演算性能を大きく改善させることが可能となりました。これにより、計算科学研究センターでは、並列 GPU 計算アプリケーション開発を加速させ、先進的計算科学研究を推進していきます。なお、HA-PACS/TCA 部の構築は、システム実装及び TCA 機構の GPU 向け開発に際し、米エヌビディア社および米クレイ社の技術協力を得て進められました。

1. 背景

10 ペタフロップス級の性能がスーパーコンピュータ「京」によって実現された現在、演算性能をエクサ*2 フロップス級（エクサはペタの 1000 倍）まで高めるための研究がすでに始まっています。しかし、1 台の計算機で使用可能な電力や設置面積の制限から、このような超高性能を実現することはますます難しくなっており、何らかの演算加速装置*3 を持つ

システムが不可欠です。これらのシステムには、演算加速装置と CPU の間の通信や、並列演算加速装置間の通信における様々なボトルネックが存在します。加えて、超並列規模の演算加速装置を用いた大規模プログラムの開発には、アルゴリズムレベルからの改良など大きな人的コストと時間がかかります。

筑波大学計算科学研究センターでは、高密度超並列 GPU クラスタを最先端標準製品技術とわれわれ独自の技術の組み合わせにより実現し、これらの問題に挑戦します。このための研究基盤が「HA-PACS」です。最先端 CPU と GPU の組み合わせによる超並列 GPU クラスタを従来にない規模で定常的に並列利用することにより、エクサスケール時代につながる演算加速型アプリケーションの開発と、われわれが提唱する密結合並列演算加速機構アーキテクチャに基づく次世代 GPU クラスタを実現します。ここで培われたハードウエア及びソフトウエアのシステム開発技術を、エクサスケールシステム実現への基盤技術として熟成させていきます。

2. 詳細

筑波大学計算科学研究センターは、宇宙・素粒子・生命などの研究をけん引する最先端の超並列演算加速器クラスタ型スーパーコンピュータ、密結合並列演算加速機構実証システム「HA-PACS」（Highly Accelerated Parallel Advanced system for Computational Sciences）の導入を平成 23 年度から進め、平成 24 年 2 月 1 日にその基礎となるベースクラスタシステムの稼働を開始。さらに平成 25 年 11 月 1 より、独自開発による密結合演算加速機構 TCA（Tightly Coupled Accelerators）を備えた HA-PACS/TCA システムを追加した拡張システムを稼働しました。追加された HA-PACS/TCA システムの基本部分は米クレイ社により提供され、これに計算科学研究センターで開発された TCA 通信機構を搭載した通信ボードを装着することで、従来のシステムを大幅に上回る GPU 間通信性能を持つシステムが実現されています。

HA-PACS/TCA システムは、米インテル社製の最新 CPU である E5-2680 v2 を 2 基と米エヌビディア社製の最高性能 GPU である Tesla K20X を 4 基搭載した、コンパクトで先進的な計算ノードを 64 台結合した並列システムです。ノード単体のピーク演算性能は 5.688 テラフロップス（毎秒 5 兆 6800 億演算）に達し、これは GPU を搭載した標準的な 2 CPU ソケットタイプのサーバを利用したこの規模の超並列クラスタ型スーパーコンピュータとして世界最高クラスの性能となります。ベースクラスタシステムと一体となった並列処理が可能で、システム全体としての総ピーク演算性能は 1.166 ペタフロップス（毎秒 1166 兆回）となります。

TCA 機構は計算科学研究センターが提唱する「密結合並列演算加速」という概念を実現する新しい技術です。将来のエクサスケール計算システムにおいて、システムの省電力化は最重要課題の一つであり、限られた電力で特定の演算を超高速に実行可能な演算加速装置の重要性が注目されています。しかし、一般的に演算加速装置はその演算性能の高さに

比べ、外部とのデータのやり取りを行う入出力部の性能が弱く、特に大規模並列処理に用いた場合、その潜在的性能が著しく制限されてしまう可能性があります。TCA 機構はこの問題に対し、ハードウエアとソフトウエアの技術により、一つの答えを提供します。

GPU を始めとする演算加速装置は、基本的に PCI Express と呼ばれる標準バス（データ伝送路）によって CPU と結合され、計算の実行や並列処理におけるノード間通信などは CPU のメモリや結合網を用いて行われます。従来の PCI Express バスは、CPU からの制御によってあらゆる通信が実行されていました。TCA 機構は、この PCI Express バスを計算ノード間通信に拡張し、ノードを超えた演算加速装置間の直接通信を実現することにより、演算加速装置が本来持つ性能を最大限に活かした新しい並列処理を実現する技術です。

TCA 機構を GPU に適用するため、われわれは PEACH2（PCI Express Adaptive Communication Hub ver.2）と呼ばれる通信チップを集積回路 FPGA により新規開発。このチップを搭載した通信ボードを HA-PACS/TCA の計算ノードに装着することにより、多数の GPU 間の通信時間を数分の一程にする大幅な短縮を実現しました。

また、TCA 機構が対象とする演算加速装置としては、GPU だけでなくメニーコアプロセッサなどを利用することも可能で、われわれは将来的にいろいろな演算加速装置に適用した実験も視野に入れています。これらの実証実験で培われる新しい形の並列処理や、開発されるアルゴリズム及びアプリケーションは、次世代の超並列演算加速機構の開発につながるものと期待されます。

3. 開発経緯とシステムの特徴

計算科学研究センターは、平成 23 年度から文部科学省から国立大学法人運営費交付金特別経費を受け、3 カ年計画で「エクサスケール計算技術開拓による先端学際計算科学教育研究拠点の充実」事業（責任者 佐藤三久教授）を推進しています。

この事業は、超並列演算加速型クラスタ計算機の「HA-PACS」を開発・製作し、これを用いて宇宙・素粒子・生命の先端的な研究を推進し、さらに次世代の演算加速型並列システムの要素技術となる密結合並列演算加速機構の技術開発を行うものです。HA-PACS の基本部分となる超並列 GPU クラスタは最先端コモディティ技術に基づく CPU と GPU を搭載したシステムとして調達します。密結合並列演算加速機構については、計算科学研究センターにおいてハードウエアからアプリケーションまでの開発を行い、HA-PACS の拡張部分として実装していきます。

システムの特徴

HA-PACS/TCA は 64 台の計算ノードを持ち、クラスタグループと呼ばれる複数の計算ノード上の GPU 間を TCA ネットワークで結合し、さらに全計算ノード間を 2 本の並列 QDR InfiniBand ネットワーク*4 で Fat Tree 結合した並列型の GPU クラスタ計算機です。全体で 364 テラフロップス（毎秒 364 兆回）のピーク計算性能、8 テラバイトのメモリを持っ

ています。既に稼働しているベースクラスタシステムと合わせ、総演算ピーク性能 1.166 ペタフロップスが実現されます。計算科学の大規模計算を実現可能とする特徴は次のとおりです。

1) 独自開発の PEACH2 チップ及びこれを搭載した通信ボードを 64 台の全ての計算ノードに装着することにより、併設する InfiniBand ネットワークよりもはるかに短い時間での高速通信を実現します。また、単に通信が速いだけでなく、計算ノード上の GPU と他のノードの GPU 間の直接通信が可能となり、これに基づく新たな GPU アプリケーションやアルゴリズムの開発を通じて、大幅な計算性能の向上が見込まれます。

2) 豊富な PCI Express チャネル数を持つ米インテル社の最新 CPU である E5 v2（IvyBridge-EP）プロセッサを 2 基搭載することにより、4 基の最新型 GPU（米エヌビディア社製 Tesla K20X）をストレスなく CPU と結合させることを可能にしました。これにより、GPU への通信性能を損なうことなく、5.688 テラフロップスという世界最高クラスのノード単体性能を 3U 相当のコンパクトな構成で実現しました。

3) TCA 機構を持つ拡張部を既設のベースクラスタと InfiniBand ネットワークによってシームレスに結合し、全システムで 1 ペタフロップスを超える超並列 GPU 計算を実行可能としました。

4. 今後の見通し

今回の HA-PACS の拡張により、科学技術の基礎となる大規模行列演算の並列処理の加速、宇宙物理分野における大規模並列処理の加速など、従来の GPU による並列処理の効率を上げ、先端的計算科学の諸分野に貢献することが可能となります。

5. 用語解説

＊1　GPU

Graphics Processing Unit の略。本来 PC サーバにおけるグラフィックス処理を目的として作られた専用プロセッサだが、近年はその高い演算性能とメモリバンド幅を利用した高性能計算への転用が活発化している。

＊2　エクサ

10 の 18 乗。ペタ（10 の 15 乗）の 1000 倍。エクサフロップスとは、現在、スーパーコンピュータ「京」が持つ 10 ペタフロップスの性能の 100 倍、すなわち毎秒 100 京回の演算性能に相当する。

＊3　演算加速装置

汎用計算を行う CPU に対する拡張機構として、PCI Express などの汎用バスを介して接続される高性能演算装置。計算を自律的に行うことは不可能で、CPU から起動されることに

より、アプリケーションの一部または全部を高速に実行する。ただし、演算装置やアーキテクチャが高性能浮動小数点演算向けに特化され、必ずしも全てのアプリケーションプログラムが高速化されるとは限らない。一般的に利用可能な演算加速装置の例としては、GPU やメニーコアプロセッサなどがある。

＊4　QDR InfiniBand ネットワーク
高性能クラスタ型計算機で多用される高性能ネットワーク。Ethernet などに比べて数倍～数十倍の通信性能を持ち、さらに数百～数千ノード規模のシステムを Fat Tree と呼ばれるネットワーク構成で結合可能である。

6. 関連情報

筑波大学計算科学研究センターホームページ
　http://www.ccs.tsukuba.ac.jp/CCS
「HA-PACS」プロジェクト特設ページ
　http://www.ccs.tsukuba.ac.jp/CCS/research/project/ha-pacs

<問い合わせ先>
梅村雅之（研究代表者）
　筑波大学計算科学研究センター長／数理物質系教授
　TEL 029-853-6485　E-mail：umemura@ccs.tsukuba.ac.jp
朴 泰祐（「HA-PACS」開発担当主査）
　筑波大学計算科学研究センター／システム情報系教授
　TEL 029-853-5518　E-mail：taisuke@cs.tsukuba.ac.jp

報道担当：
筑波大学計算科学研究センター広報室
　TEL：029-853-6260（直通）、6487（代表）　E-mail：pr@ccs.tsukuba.ac.jp

クレイ・ジャパン・インク 製品企画本部
　TEL：03-3503-0901（代表）　E-mail：jpsales_online@cray.com

エヌビディア ジャパン マーケティング本部 広報/マーケティングコミュニケーションズ
中村かおり
　TEL: 03-6743-8712（直通）　E-mail：knakamura@nvidia.com