GTC Japan 2016 参加報告

大島聡史, 塙敏博
東京大学情報基盤センター

1　はじめに

本記事では2016年10月15に開催されたGTC Japan 2016 (以下GTCJ2016) について報告する。GTC JapanはNVIDIA社と東京工業大学 学術国際情報センターが共催して年に1回、夏から秋にかけて開催しているGPUコンピューティング (GPGPU) のイベントである。本スパコンニュースでもVol.14 No.5, Vol.15 No.5, Vol.16 No.5, Vol.17 No.6にて参加報告 (GTC Japan 201x参加報告) を掲載しているため参考にしていただきたい。

GTC Japanは前身であるNVIDIA GTC Workshop Japan 2011から毎回、六本木や虎ノ門など都心にて開催されてきたが、今回は出展数や参加人数が増えてきたためか、お台場のヒルトン東京お台場[*1]での開催となった。しかし、参加人数が3,500人を数え、約50もの展示が行われたため、大盛況どころか会場が狭すぎるという印象を受けた。

GTCJ2016では初めて文部科学省が「OFFICIAL SUPPORT」としてスポンサーに名を連ねた。これは文部科学省が近年AIやディープラーニングに注目していることと、その研究開発にGPUが盛んに用いられていることによるものである。実際、今回のGTCJ2016は前回にもましてAIやディープラーニングにフォーカスした内容となっており、Jen-Hsun Huang氏 (NVIDIA共同創設者, CEO)による基調講演、発表セッション、展示やデモまで、その多くがAIやディープラーニングに関係するものであった。

なお、一部の講演については`https://www.gputechconf.jp/sessions.html`から資料をダウンロードすることができるため、参照されたい。

図 1　Jen-Hsun Huang 氏による基調講演の様子とポスターセッション会場の様子

2　PascalアーキテクチャとCUDA8

NVIDIA 社の成瀬氏から、最新のPascalアーキテクチャに基づくGPUであるTesla P100を中心としたいくつかの製品、およびそれらに対応した開発環境CUDA8の紹介があった。

[*1] さらに前身である GPU コンピューティング 2010 Winter が開催されたホテル日航東京と同じ場所である。

PascalアーキテクチャのTesla GPUシリーズは大きく分けて、数値計算向けのP100と、ディープラーニング向けのP40(およびP4)が発表されている。P40は、これまでのGPUに搭載されてきたメモリの後継版であるGDDR5Xメモリを搭載し、単精度(FP32)演算性能の高速化と新たにINT8演算に対応したGPUである。一方のP100は、従来のGPU向けメモリよりも高速な新しいメモリであるHBM2 (High Bandwidth Memory 2) メモリを搭載してメモリ転送性能を高めるとともに、SM (Stream Multiprocessor)の構成を見直し倍精度(FP64)、単精度(FP32)、半精度(FP16)の演算性能を高速化したGPUである。さらにP100にはNVlinkという新しい高速インタフェースを搭載したバージョンも存在するが、その実際の転送性能が初めて明らかにされた。

CUDA8については、Pascalアーキテクチャへの対応とあわせて、Unified memoryサポートが強化され、ホストメモリとGPUメモリの間でのページ転送が高速化された。これまでも同種の機能は用意されていたが、GPUメモリとホストメモリを意識したプログラミングをしなければまともな性能が得られなかったものが、あまり意識せずに記述しても良い性能が得られやすくなり、また性能チューニングのためのヒント情報なども指定できるようになった。そのため、これまでGPUによる高速化が困難だと考えられていたアプリケーションについても性能向上が期待できる。

3　企業展示

企業による展示においては、AIやディープラーニングに関する応用事例が多数展示され注目を集めていたが、最新のGPUであるTesla P100を搭載したサーバー製品の展示も多く見られた。現在HPCにて用いられているGPUは2012年に開発されたKeplerアーキテクチャのGPUであり、久々の大きな更新となることから注目も大きい。特にNVLinkによってGPU同士の接続やCPU-GPU間の接続が従来より大幅に高速化されたため、CPUに加えて4枚のGPUを1シャーシに搭載した製品が多く見られた。なお東京大学情報基盤センターにて現在設置が進められているReedbush-H（データ解析・シミュレーション融合スーパーコンピュータシステムのGPU部）の計算ノード(120ノード)には、1ノードあたり2ソケットのCPUに加えて2枚のP100が搭載される予定である。

4　OpenACCに関する研究発表

例年のGTC Japanと同様、GPUコンピューティング研究会によるテクニカルセッションが並行して開催された。その中でも特にOpenACC関連の発表が目を引いた。

JAXAの宮島氏からは、膨大なレガシーコードをいかにGPUに移植するかについて、OpenACCを使った移行プロセスの提案があった。OpenACCを用いれば、既存のプログラムを単純にGPU上で動かせるようにする程度ならばそれほど手間はかからないが、性能を最大限引き出すためには、データ構造を変更するなど、大掛かりな修正が必要になってしまう。それでもOpenACCによって段階を踏んだ最適化ができるようになったことは非常に有効そうであった。

筑波大の廣川氏からは、第一原理電子動力学コードの最適化について、いずれも最新のプロセッサである、Intel Xeon PhiとNVIDIA Tesla P100とを用いた性能比較の結果が示された。実効性能と実効効率の面で興味深い議論がなされた。しかし実際には、元となるコードがすでに京コンピュータやXeonクラスタなどに最適化されているなど、最適化の度合いに依存するのでは、という質問があり、特にXeon Phi向けではそのような傾向が見られそうである。