C-042

PEACH3 の通信性能の測定

金田　隆大 †　鶴田　千晴 †　塙　敏博 ‡　天野 英晴 †
† 慶應義塾大学　‡ 東京大学

1　はじめに

近年,Graphic Processing Unit(GPU) を搭載したハイパフォーマンスコンピューティングが一般的になった. しかし, ノードをまたいだ GPU 同士の通信は遅延が大きく, 改善する必要がある. この問題を解決するために開発された PEACH3 の性能評価を行った.

2　TCA(Tightly Coupled Accelerators)

TCA アーキテクチャは，GPU クラスタにおける GPU 間のデータ転送ボトルネックを解消するために開発されたシステムである [1].

2.1　PEACH2

PEACH2 は TCA アーキテクチャが利用可能なスイッチである.PEACH2 によりノードをまたいだ GPU 間で直接通信が出来る． PEACH2 は，Altera 社製 FPGA Stratix IV を用いて実装され，4 本の PCI-Express(PCIe) Gen2 x8 コネクタ (N, E, W, S) と DMA コントローラ，NIOS ソフトコア・プロセッサ，メモリ領域 (DDR3-SDRAM など) を搭載する．各ノードは, PEACH2 上のコネクタ間を PCIe ケーブルで接続することで通信が出来る.

PEACH2 と GPU 間で直接通信するために,GPUDirect RDMA を使用している.

2.2　PEACH3

PEACH2 が利用する PCIe Gen 2 x8 は, バンド幅の理論最大値が 4.0GB/sec であり, ノード間通信に一般的に利用される Infiniband にバンド幅で劣る. そのため通信データサイズが大きいプログラムでは PEACH2 ではパフォーマンスが不十分であった． そのため, バンド幅を強化した PEACH3 を開発した.PEACH3 では PCIe Gen 3 x8 を利用し, 理論バンド幅の最大値が 7.9GB/sec まで上昇している.

基本的な仕様は PEACH2 と同等で,FPGA を PCIe Gen 3 のハード IP が利用出来る Stratix V に変更した.

現状の PEACH3 では S ポートを省略している.

3　通信試験と性能評価

本報告では PEACH3 を使った GPU 間通信の性能を測定し, 結果を示す. 実験環境スペックを表 1 に、概略図を図 1 に示す.PEACH3 間の接続には 50cm の PCIe x8 ケーブルを用い,W ポートと E ポート同士を接続する.

3.1　ノード内通信

まず初めに同一ノード内の CPU-GPU 間通信の性能測定を行う．TCA の API と cudaMemcpy() それぞれ

表 1: 実験環境のシステム仕様

CPU	Intel Xeon E5-2680 v2(Ivy Bridge-EP)
メモリ	32 GB, DDR3 1600MHz,
GPU	NVIDIA Tesla K40m

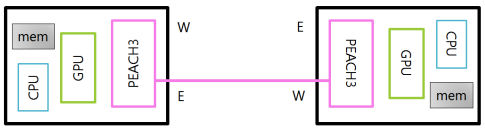

図 1: PEACH3 の実験環境

で転送サイズを変えながら,100 回転送したときの平均の値を計測している．この時,CUDA ではホストメモリは Page-locked メモリとして確保している.

図 2 にレイテンシを, 図 3 にバンド幅を示す.

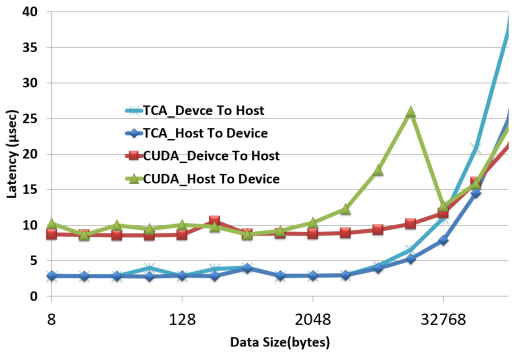

図 2: レイテンシ (同一ノード内のホスト-デバイス間)

図 2 より,TCA では最小レイテンシは 2.8 μ sec で,CUDA より約 3 倍高速だった． これは PEACH3 ではアドレス変換をしないので CUDA より高速にルーティングが可能なためである．また,HostToDevice に比べ DeviceToHost の通信性能が低いのは GPU の write より read の性能が低いことによる．PEACH3 の場合, 転送データは必ず PEACH3 の N ポートを経由する. よってデータサイズが大きいと N ポートのバンド幅がボトルネックとなり,cudaMemcpy() の方が性能が良い.

3.2　ノード間通信

次に, ノードをまたぐ GPU-GPU 通信の試験を行った. 図 4 にレイテンシ, 図 5 にバンド幅を示す. MPI/IB

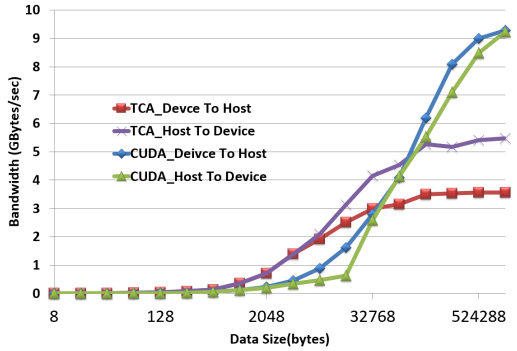

図 3: バンド幅 (同一ノード内のホスト-デバイス間)

と PEACH2 を利用した通信の結果も比較対象として示した [2]. MPI/IB では 512KB 以上のデータサイズで GPUDirect RDMA を使用しない設定になっている. 図 4 より, レイテンシは MPI/IB に比べるとかなり低く,PEACH2 と同等の約 2 μ sec である. 図 5 より, バンド幅は 2KB 以上の時,PEACH2 の約 1.2-1.3 倍であり,MPI/IB と比べると 1KB 4KB の転送時におよそ 4 倍のバンド幅を達成している. データサイズが 1MB 以上で PEACH3 は MPI/IB の結果を下回るが, これは PEACH3 で利用している GPUDirect RDMA が, データサイズが大きいときに性能低下を引き起こすためである. MPI/IB では,GPUDirect を使用するデータサイズを設定できるが,PEACH3 では同等の機能がない. PEACH3 でも GPUDirect を使用せず通信を行えるよう改良する必要がある.

また,PCIe Gen3 x8 の最大バンド幅に比べて PEACH3 のバンド幅が小さい原因を探るため, 図 6 に示すようにノード間通信として 4 種類の組み合わせでバンド幅の計測を行った. 図 6 より, ホストメモリからデータを読み出す通信はピークバンド幅が約 7GB/sec を達成している事がわかる. それに対し, デバイスから読み出す通信はピークバンド幅が約 3.5GB/sec とかなり小さい. このことから GPU の read 性能がボトルネックになっている事が分かる. そのため GPU 間通信の最大バンド幅が約 3.5GB/sec で止まっているのは PEACH3 ではなく GPU 側の原因であると考えられる.

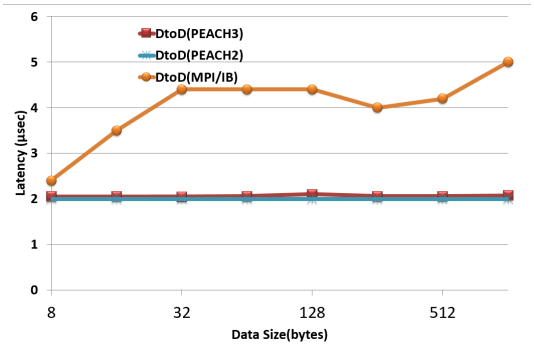

図 4: レイテンシ (Device to Device)

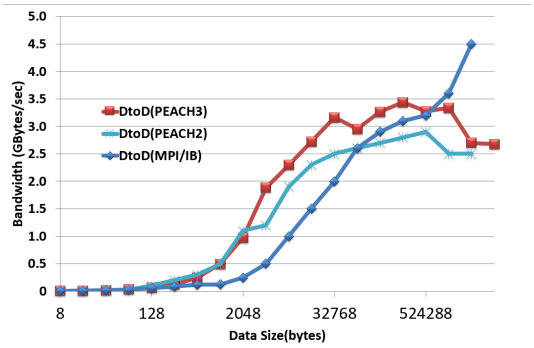

図 5: バンド幅 (Device to Device)

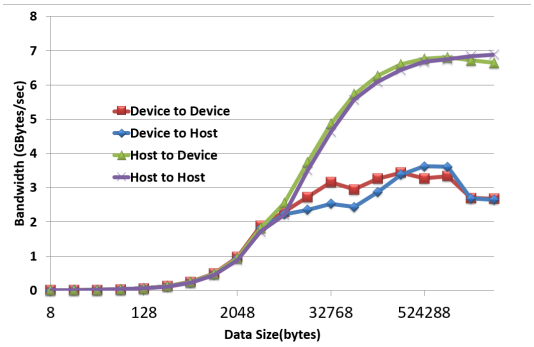

図 6: バンド幅 (ノード間の全組み合わせ)

4 結論

本報告では,GPU 間直接通信が可能な PEACH3 の通信性能の測定を行った.

同一ノード内の CPU-GPU 間通信ではレイテンシ, バンド幅ともデータサイズが 32KB 以下の時.TCA API を利用した方が高性能であった. ノード間通信では PEACH2 より PEACH3 の方が約 1.3 倍のバンド幅を記録し, 最小レイテンシも PEACH2 と同等だった. 特に GPU からデータを読み出す通信では,GPU の read 性能がボトルネックとなる.

謝辞

本件研究は、JST-CREST 研究領域「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」、研究課題「ポストペタスケール時代に向けた演算加速機構・通信機構統合環境の研究開発」による。

参考文献

[1] T. Hanawa, Y. Kodama, T. Boku, and M. Sato, “Interconnect for tightly coupled accelerators architecture,” ”IEEE 21st Annual Sympsium on High-Performance Interconnects (HOT Interconnects 21)”, 2013.

[2] 松本和也, 塙 敏博, 児玉祐悦, 藤井久史, 朴 泰祐, “密結合並列演算加速機構 TCA による GPU 間直接通信における Collective 通信の実装と性能評価,” 情報処理学会論文誌コンピューティングシステム (ACS), pp.36–49, 11 月 2015 年.