C-018

FPGA による PCIe Gen3 および 100G 超光インターコネクトに関する実験評価：汎用高速演算クラスタシステムの実現に向けて
A study of PCIe Gen3 and 100+G interconnect circuits on an FPGA: toward a commodity high-performance computing cluster system

高山尋考 [1] 山口佳樹 [2, 3] 朴泰祐 [2, 3]
Hirotaka Takayama Yoshiki Yamaguchi Taisuke Boku

1. はじめに

IoT/M2M や BigData の広がりにより、データセンターにおける演算性能の要求は今後ますます増大すると思われる。ここで要求されている高い演算性能を実現するためには、ノード間高速通信技術が欠かせないものとなるはずである。筑波大学計算科学研究センターでは、GPU クラスタにおけるノード間の低レイテンシ通信実現に向け Tightly Coupled Accelerators 機構 [1]を提案し、HA-PACS/TCA として運用を行っている。

本論文では、ノード間通信の高速化に加えて高い拡張性および柔軟性を許容するネットワークをにらみ、FPGA による PCIe Gen3 8 レーンおよび 100G 超光インターコネクトによる通信実験を行った。そして、この実験を通して、HPC システムを支えるネットワークチップとしてみた FPGA の可能性について議論を行う。

2. 技術背景

本章では、本研究の技術的背景である PCI Express と高速シリアル通信についてそれぞれ紹介する。

2.1 PCI Express について

PCI Express (以下、PCIe) は、最新デバイスによって要求される帯域幅を確保するため、計算機関連メーカの業界連合により生まれた高速シリアル転送における業界標準規格である [2]。PCIe は、Gen1、Gen2、Gen3 と新しくなるにつれ、物理層上の実効データ転送速度（片方向 1 レーンあたり）を 0.25[GB/s]、0.5[GB/s]、0.98[GB/s]と向上させてきた。ここで、ヘッダ等のオーバーヘッドを考慮した、トランザクション層における PCIe Gen3 の性能を表 1 に示す。ペイロードサイズが∞の場合、オーバーヘッドを無視した値と等しいため、物理性能と等しい値になっている。

表 1　PCIe Gen3 の性能 [GBytes/s]

ペイロードサイズ	PCIe Gen3 のレーン数					
	1	2	4	8	12	16
128 [bytes]	0.80	1.60	3.19	6.38	9.57	12.76
256 [bytes]	0.88	1.76	3.53	7.05	10.58	14.10
∞ [bytes]	0.98	1.97	3.94	7.88	11.82	15.75

[1]筑波大学大学院システム情報工学研究科
Graduate School of System and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki, 305-8573, Japan
[2]筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Ibaraki, 305-8573, Japan
[3]筑波大学計算科学研究センター
Center for Computational Sciences, University of Tsukuba, Tsukuba, Ibaraki, 305-8577, Japan

2016 年現在、PCIe Gen3 8 レーンに対応した FPGA を市場から入手可能であり、XILINX 社などは PCIe Gen4 など、次世代バス規格の対応も始めている。

2.2 高速シリアル I/O (High-speed serial I/O)について

2000 年頃まではパラレルバスが全盛であったが、個々のレーンにおけるデータおよびクロックスキューが問題となり、2 GHz を超える高速化は非常に困難であった。そこで、1 つの作動信号にデータとクロックの双方を含み伝送するシリアル転送によりスループットの向上を図ってきた。

前節の PCIe もだが、USB、HDMI、SATA、および光通信なども高速シリアル I/O（以下、HSSIO）によって実現されている。そして、光モジュールとの組み合わせによる高速データの長距離伝送にも現在注目が集まっている。

FPGA のトランシーバに目を向けると、この 10 年でピンあたりで 5 倍以上、デバイス単位で 20 倍以上のスループット向上を実現している（表 2）。このため、高速ネットワークをサポートするスイッチチップとして FPGA を活用することは十分に考えられる。

表 2　トランシーバ性能の進化（[3]の表 7・1 を改編）

年	シリーズ	名称	帯域[Gbps]×レーン数[本]
2002	Virtex2 Pro	Rocket IO	3.125×24
2004	Virtex 4	Rocket IO MGT	6.5×24
2006	Virtex 5	Rocket IO GTX	6.5×48
2009	Virtex 6	GTX & GTH	6.6×48 & 11.18×24
2010	Virtex 7	GTH & GTZ	13.1×72 & 28.05×16
2013	UltraScale	GTH & GTY	16.3×60 & 30.5×60
2015	UltraScale+	GTY	32.75×128

3. PCIe Gen3 における先行研究との差分

先行研究 [4]において、50%程度の性能差がシステム構成により生ずる可能性を示唆した。著者らは継続して調査を進めているが、この傾向はあまり変わらない。最新の結果を図 1 に示す.

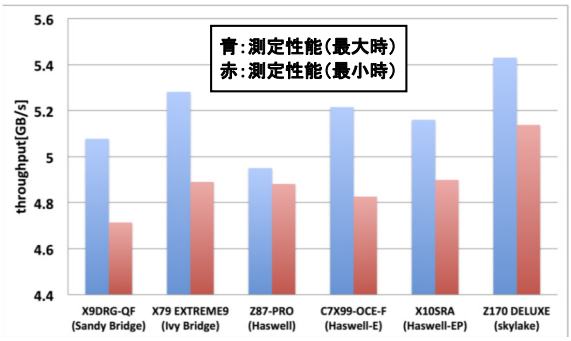

図 1 FPGA➔(PCIe Gen3 x8)➔FPGA 通信性能の比較

なお、システム構成による性能低下の原因は、一つはマザーボードの BIOS 設定に絡むものと推定されたが、もう一つはエラー発生時のデータ再送におけるものであった。

4. 100+G インターコネクト実験

本インターコネクト実験では、PCIe Gen3 の組み合わせを考慮し、図 2 に示す 3 つの構成について評価を行った。

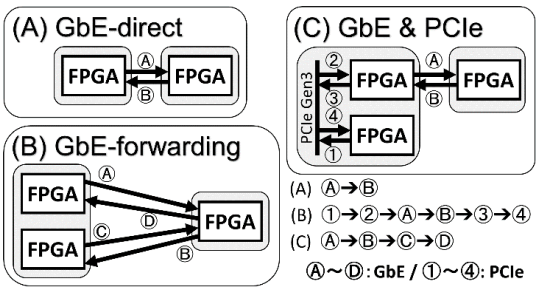

図 2　100+G インターコネクト実験における 3 つの分類

4.1.1 (A) ダイレクト接続実験

FPGA 間に他のノードを挟まないダイレクト通信の結果を示す。本計測には、東京エレクトロンデバイス社製 TB-7VX-1140T-PCIEXP を使用した。4 枚のメザニンカードの利用により最大 400Gbps までの光通信が可能である。

図 3 に、このときのスループット性能を示す。データが十分にそろっていないため掲載していないが、300Gbps（37.5GB/s）まではほぼ理論性能に近く、かつ高いスケーラビリティが実現可能なことを確認している。

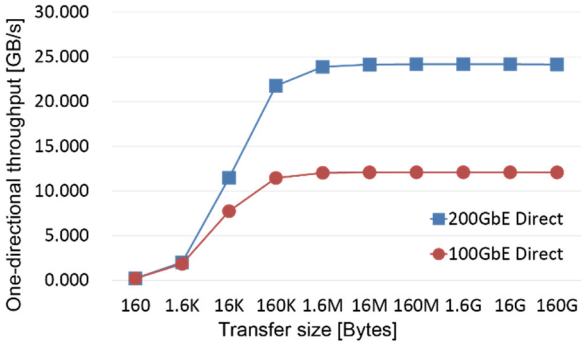

図 3　ダイレクト接続実験におけるスループット性能

4.1.2 (B) 経由を伴う接続実験

図 4 に、FPGA を仲介(ワンホップ)させた、接続実験の結果を示す。本実験において、ホップ挿入による性能低下は確認できなかった。通信レイテンシ計測も行ない、FPGA チップによるレイテンシの増加（FPGA そのものによる性能低下）は 100 ナノ秒より十分に小さいことが確認された。

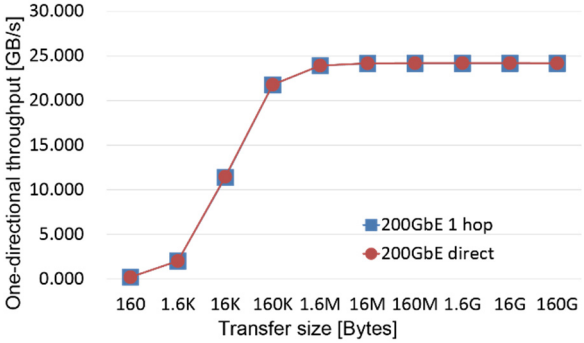

図 4　経由を伴う接続実験におけるスループット性能

以上から、FPGA をネットワークチップとして汎用高速演算クラスタシステムを構築する際、FPGA チップによるレイテンシの増加は小さく、完全に無視はできないが、この効果を過分に考慮する必要はないと著者らは考えている。

4.1.3 (C) PCIe 経由実験

最後に、PCIe を経由した性能について評価を行った。実験結果については図 5 に示す。スループットは、PCIe Gen3 x8 の性能に制約を受けているが、先行研究と合わせて考えるとほぼ想定した性能が得られることを確認した。

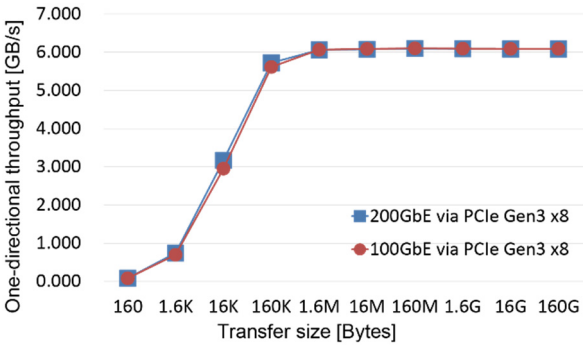

図 5　ノード内 PCIe バスを経由したスループット性能

5. おわりに

100+G の光通信においても、理論性能にほぼ近い形で実システムを構成できることが確認できた。今後は、これらの結果を基にしたインターコネクトを実システムに組込み、アプリケーションレベルでの評価を行っていく予定である。

謝辞

本研究の一部は 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．

参照文献

[1] 塙, 児玉, 朴 , 佐藤, “Tightly Coupled Accelerators アーキテクチャに基づく GPU クラスタの構築と性能予備評価,” *情報処理学会論文誌. コンピューティングシステム,* Vol.6, No.3, pp. 14-25, 2013.

[2] PCI-SIG, “PCI Express Base Specification,” 10 November 2010. [Online]. Available: http://composter.com.ua/documents/PCI_Express_Base_Specification_Revision_3.0.pdf. [Accessed 30 June 2016].

[3] 天野(編), 山口 , 長名(著), “FPGA の原理と構成（7 章：PLD/FPGA の応用事例）,” オーム社, 2016, pp. 209-245.

[4] 高山 , 山口, “FPGA 間通信で発生する諸問題 - PCIe Gen3 通信における注意点-,” *信学技報,* Vol.115, No.343, pp. 33-38, 2015.