2L-06

並列計算における FPGA 高速相互接続に関する研究

敖運[†]　　中村大樹[‡]
筑波大学大学院システム情報工学研究科[†‡]

山口佳樹[#§]
筑波大学システム情報系[#]
筑波大学計算科学研究センター[§]

1. はじめに

近年，Big Data や機械学習など，高性能な大規模並列計算システムの需要がより高まっている．並列計算システムの演算性能は，GPU を初めとする演算加速器の飛躍的な性能向上により，高い成長率を維持している．一方，これらに対して PCI Express (PCIe)[1]や InfiniBand[2]等のデータ転送の性能向上は十分と言い難い．

このため，並列計算システムの構築において，PCIe などによるノード内の各コンポーネント間の通信と InfiniBand などのノード間通信をシームレスに接続し，実効通信帯域をどのように高めるか、が重要なテーマとなっている．

筑波大学計算科学研究センターでは，2019 年度に Cygnus というスパコンが稼働予定である．この計算ノードは，CPU，GPU，FPGA を演算装置として持ち，これらは PCIe により接続されている。また，計算ノード間は InfiniBand によるネットワークと FPGA による 2D-Torus ネットワークにより接続されている．そこで本稿では，この実現に先立ち，計算ノード内外の通信効率を高める FPGA 実装について提案し，その性能と拡張性について議論する．

2. 関連研究

本稿の先行研究として，Nakamura ら[3]の研究報告が挙げられる．Nakamura らは，FPGA における PCIe 通信について数多くのプラットフォームで検証を行い，プラットフォームに依らない安定した通信を実現する方法について検証を行っている．単一の計算ノード（ノード内通信）という観点では一定の知見を示したものの，並列計算システム（ノード間通信）という観点では別の知見が必要である．

Takayama ら[4]は，FPGA 上に実装した PCIe 通信と光インタフェースによる高速シリアル通信について性能評価を行っている．PCIe 通信プロトコルから光通信プロトコルの変換において十分に低遅延な FPGA 実装を示し，また，高い拡張性を持つことを報告している．しかし，光通信は peer-to-peer 通信のみであり，ルーティングについては今後の課題とされている．

そこで本研究では，これらの先行研究を参考に，ルーティングを伴った，並列計算システムに応用可能な実装について提案および検証を行った．

3. PCI Express (PCIe)

PCI Express は，PCI-SIG によって策定された全二重シリアル通信インタフェースである．同時に利用する通信線の数をレーンという言葉で表現し，一般には 1，2，4，8，16 と 2 のべき乗のレーン数で通信を行う．

また，PCIe 通信は，ルーティングを実現する Root Complex を中心に構成されている．図 1(a)に PCIe 通信システムの構成を示す．PCIe に接続される各コンポーネントは，Root Complex を経由して，他のデバイスへアクセスできる．

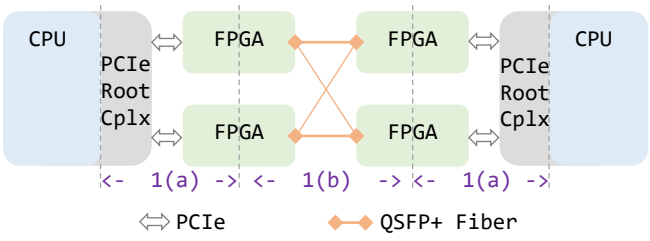

図 1 検証システムの構成
Figure 1 Experimental System Architecture

4. 光通信

本稿では，同一の計算ノードに接続された FPGA 間は PCIe により通信を行い，異なる計算ノードにある FPGA 間は光通信を用いる．この物理接続は QSFP+ポートにより実現される．図 1(b)に本稿で実装する光通信の構成を示す．

5. システム実装と評価

本稿では，PCIe 通信及び光通信について FPGA ボード間の性能を測定し評価する．また FPGA に実装したルーティング制御回路の性能についても検証する．表 1 に実験に使用した環境を示す．

FPGA-based High-speed Interconnect for Parallel and Distributed Computing
Yun AO[†], Hiroki NAKAMURA[‡], and Yoshiki YAMAGUCHI[#§]
Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Ten-ou-dai Tsukuba Ibaraki 305-8573, Japan[†‡]
Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Ten-ou-dai, Tsukuba Ibaraki, 305-8573, Japan[#]
Center for Computational Sciences, University of Tsukuba, 1-1-1 Ten-ou-dai, Tsukuba Ibaraki, 305-8577, Japan[§]

表1　実験環境(以下の構成で2ノード用意した)
Table 1　Experimental System Setup (2 sets of below)

CPU	Intel Xeon Gold 5122 (Skylake-SP)
Motherboard	Supermicro X11DRG-QT
FPGA Board	Xilinx KCU1500(Chip: XCKU115)
FPGA IDE/EDA	Xilinx Vivado 2018.3

5.1 PCIe 通信性能および光通信性能の検証

PCIe 通信性能の検証は，PCIe 3.0 x8 モードで実行した．光通信の検証は，40Gbps の QSFP+ モジュールと長さ 2m の光ファイバーが使用され，通信プロトコルには Xilinx 社の Aurora プロトコルを用いて，1 パケットあたりのデータサイズは 1,536bytes とされ，実験結果を表 2 に示す．

表 2 PCIe 通信と光通信の実験結果（FPGA 同士）
Table 2 Test Result of PCIe & Fiber Comm. between FPGAs

Transfer Method	Throughput[GB/s]	Latency[ns]
PCIe in One CPU	5.8185	~500
PCIe across Sockets	5.8172	~500
Fiber Connection	4.8899	259

5.2 ルーティング制御回路の検証

FPGA 上には，プロトコルのラッピングを含む，ルーティング制御回路が実装されている．本実験では，FPGA1（ノード 1）=（PCIe）⇒ FPGA2（ノード 1）=（光通信）⇒ FPGA3（ノード 2）の経路で測定を行った．FPGA2 がルーティングを担当しており，FPGA1 から FPGA3 までのレイテンシは 783ns であった．

5.3 FPGA のリソース使用量

本実験における FPGA に実装された通信制御回路（PCIe IP，Aurora IP とルーティング機能を含む）のリソース使用量を表 3 にまとめる．

表 3 FPGA リソース使用量
Table 3 Resource Utilization of FPGA Chip

Type	Board	Utilized	Utilization
Flip-Flop	1,326,720	16507	1%
LUT	663,360	7,442	1%
BRAM	2,160	117	5%
GTH	64	16	25%

6. 考察

PCIe 通信にはパケットヘッダが必要であり，本実装の実スループットは理論値（7.88GB/s）の 74% にとどまっている．これは，ヘッダとペイロードの割合（16bytes : 48bytes）から考えると妥当な数値であり，高効率な実装を実現できたと考えている．また，通信レイテンシは約 500ns であった．

光通信の性能はオーバーヘッドを考えた理論値（4.90GB/s）の 99.8%に到達し，通信レイテンシも 259ns と十分な値であると考えている．

ルーティングについては，転送におけるレイテンシの値が PCIe レイテンシ + 光通信レイテンシ + ルーティングであることを考えると，約 50ns と言える．これは，回路のパイプライン段数から推測される値（52ns）ともほぼ一致しており，適切な値であると考えている．

最後に，リソース使用量は転送用トランシーバを除いて全体の 5%以下であり，FPGA に実装する他の機能への影響は少ないことが確認できた．

7. おわりに

本稿は FPGA で PCIe と光通信によるノード間通信システムの構築にむけて，PCIe 通信と光通信の性能評価を行った．

提案する通信制御回路の規模は，FPGA 全体に対し非常に小さく，FPGA を演算加速装置として利用する際にも十分利用できると考えている．

PCIe 通信および光通信の実効帯域は，ヘッダ部分も含めると理論値の 99.8%に達しており，またルーティングにおいても 50ns 程度のレイテンシで実装されていることから十分な性能を実現できたと考えている．

今後は，本稿で提案したルーティング手法を基づき，FPGA による 2D-Torus 通信網およびそれによる科学技術計算の実装について取り組む．

謝辞

本研究の一部は，JSPS 科研費「JP17H01707」「JP18H03246」，文科省「次世代計算技術の開拓による学際計算科学連携拠点の構築」，文科省「次世代領域研究開発（高性能汎用計算機高度利用事業）」における「次世代演算通信融合型スーパーコンピュータの開発」による．また，Xilinx 社より「Xilinx University Program」を通じて開発ソフトウェアの支援を受けており，ここに謝意を表する．

参考文献

[1] *Specifications PCI-SIG*, https://pcisig.com/specifications/.
[2] *InfiniBand Trade Association*, https://www.infinibandta.org/.
[3] H. Nakamura, et al.: *Thorough analysis of PCIe Gen3 Communication*, Int'l Conf. on Reconfigurable Computing and FPGAs, pp.1-6, 2017.
[4] H. Takayama, et al.: *Performance Assessment of PCIe Gen3 and 100+G High Speed Serial Link Communication on FPGAs*, Int'l Symp. on Highly Efficient Accelerators and Reconfigurable Technologies, pp.1-2, 2016.