C-039

ストレージシステム適用に向けたプロセッサ PCIe 性能の評価
Evaluation of PCIe performance of processor for storage system

岡田 尚也† 高田 正法† 新井 政弘†
Naoya Okada Masanori Takada Masahiro Arai

1. 背景

年率 40%という急激なデータ量増大[1]に伴いデータ転送処理に時間を要するようになったため、ストレージシステムは高い IO 性能が要求されている。また、近年ストレージシステムに適用されるプロセッサは PCI Express® (PCIe®) コントローラを統合させ、PCIe の世代更新による帯域向上に伴って IO 性能が向上している。ストレージシステムの帯域性能はプロセッサの IO 帯域の合計が左右する事から、高性能化を実現するためにはプロセッサの PCIe スループット性能を出し切る事が重要である。

高い信頼性が求められる IT システムにおいて、ストレージシステムがホストからユーザーデータを受領する時に、プロセッサに接続するホスト I/F 等の PCIe デバイスは T10-Data Integrity Field(DIF)[2]と呼ばれる 8B の保証コード(図 1)をユーザーデータに付加する。T10-DIF は Guard Tag と呼ばれる CRC-16 を含み、ユーザーデータの誤り検出に用いられる。ストレージシステムがユーザーデータをホストに返送する時、ホスト I/F が CRC 演算でユーザーデータの誤りの有無を検証する事で、ストレージシステムはホストに対して格納したデータの信頼性を保証する(図 2)。従い、ストレージシステムでは T10-DIF 分だけサイズが増加したユーザーデータを転送処理する。

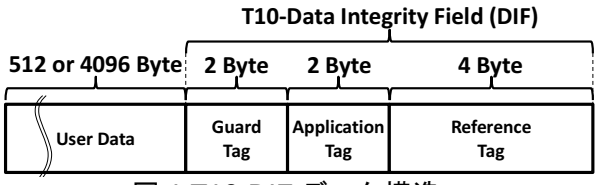

図 1 T10-DIF データ構造

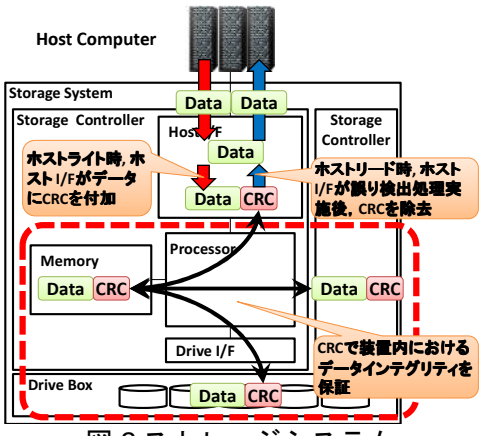

図 2 ストレージシステム

一方、ストレージシステムのプロセッサは主記憶に対し 64B 単位でメモリの読み書きを行う。そのため、PCIe デバイスが T10-DIF を付加したユーザーデータを転送すると、64B より小さなデータのメモリの読み書きや、64B アドレス境界を跨いだメモリの読み書きが生じるため(以降 64B 非アライメントメモリアクセスと呼ぶ)、メモリを一度読み、更新処理をしてから書き戻すリードモディファイライト(RMW)処理が必要になる。 RMW はプロセッサ内部の処理効率の低下を生じさせ、PCIe デバイスのデータ転送性能の低下を引き起こす。

2. 目的

従来の検証[3]では接続 PCIe レーン数が小さい、または 64B 境界に沿った条件下での検証しか行われていかった。しかし、先述の通り、ストレージシステムではプロセッサの PCIe 帯域を使い切るような高負荷条件下で、かつ 64B 非アライメントメモリアクセスが生じる条件での性能特性が重要となる。そこで、同条件におけるプロセッサの PCIe スループット性能への影響を明らかにする事を本研究の目的とした。

3. 評価方法

ハードウェア構成を図 3 に示す。FPGA にホスト I/F を模擬させるため、T10-DIF を付加したユーザーデータ相当のライト転送を行う PCIe パケット生成回路を設計し、Gen3 8 レーンのバスを持つ FPGA に回路を実装してサーバーボード上のプロセッサと接続した。

表 1 に実験環境を示す。ストレージシステムに用いられる典型的なプロセッサである Intel Xeon を実験用プロセッサとして選択した。また、検証に必要十分な帯域を確保するため、PCIe Gen3 8 レーンのバス幅を持つ FPGA を選択した。

表 2 に FPGA が実行する PCIe ライトのデータ転送パターンを示す。ストレージシステムのユーザーデータの典型的なデータ管理単位である 512B と 4096B を想定し、種々の転送長について T10-DIF の有無の違いと、FPGA の接続数で実験条件を変えて性能測定を行った。測定では、表 2 に示した各データ転送長を 1 回のデータ転送単位として、PCIe ライトの転送先アドレスは、64B の整数倍となるアドレス値をインクリメントさせながら、指定した回数だけ FPGA に PCIe ライトパケット発行し、シーケンシャルライト処理を実行させた。FPGA が出力したパケット数と測定時間から PCIe シーケンシャルライトのスループット性能を算出した。

†(株) 日立製作所 研究開発グループ 情報通信イノベーションセンタ, Hitachi Ltd, Research & Development Group, Center for Technology Innovation - Information and Telecommunication.

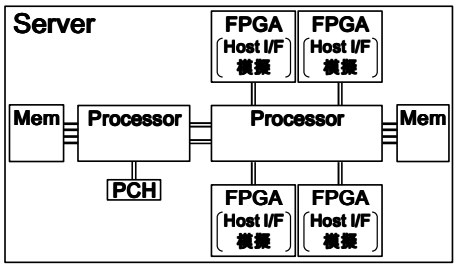

図 3 ハードウェア構成

表 1 実験環境

項目	条件
プロセッサ種別	Intel Xeon® E5-2698 v4
プロセッサ搭載数	2
メモリ I/F	DDR4-2133MHz、4ch
メモリアクセスモード	Uniform Memory Access
QPI	9.6GT/s 2 link
FPGA 種別	Altera Stratix® V GX
FPGA I/F 種別	PCIe® Gen3 8 レーン
FPGA 接続数	1、2、3、4
PCIe® Max Payload Size	256 Byte
OS	Cent OS 7.1

表 2 データ転送条件

ケース	転送長	データ転送パターン	64B 非アライメントメモリアクセス頻度
(A)	512 B	256 256	0 回/kB
(B)	520 B (512B+DIF)	256 256 8	8.86 回/kB
(C)	1040 B (520B*2)	256 256 256 256 16	6.89 回/kB
(D)	2080 B (520B*4)	256 ・・・ 256 32	4.43 回/kB
(E)	4104 B (4kB+DIF)	256 ・・・ 256 8	7.24 回/kB
(F)	4160 B (520B*8)	256 ・・・ 256 64	0 回/kB

4. 結果

測定結果を図 4 に示す。64B 非アライメントメモリアクセスが生じないデータ転送パターンである表 2 のケース(A)及び(F)では、PCIe スループット性能は FPGA 接続 PCIe レーン数に概ね比例する事が分かった。一方、64B 非アライメントメモリアクセスが生じるデータ転送パターンでは、合計接続 PCIe レーン数が 8 レーンでは性能低下率は高々7%に留まったものの、FPGA 接続 PCIe レーン数が増加し PCIe ライトデータ流量が大きくなるに従って性能低下率が大きくなり、高負荷を想定した 32 レーン接続時には最大 38%の性能低下が生じる事がわかった。

また、測定結果から 64B 非アライメントメモリアクセス発生頻度の大きさに応じて性能低下の度合いが大きくなる事がわかった。特に 64B 非アライメントメモリアクセス発生頻度が 7.24 回/kB から 8.86 回/kB にかけて性能低下量が大きかった。

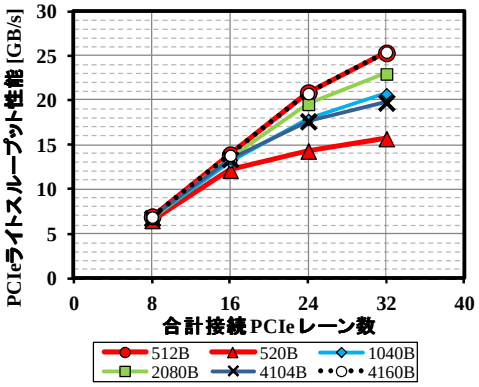

図 4 性能測定結果

5. 考察

64B 非アライメントメモリアクセス発生頻度が一定であれば接続 PCIe レーン数が増加しても性能低下量は一定となるはずである。しかし実際は、64B 非アライメントメモリアクセス発生頻度の増加と FPGA のライト流量増加に比例して、PCIe スループット性能の低下量は大きくなった。これは、プロセッサ内部の RMW 処理回路が各 FPGA の接続 PCIe レーン間で共有されている事を示していると考えられる。64B 非アライメントメモリアクセス発生頻度と FPGA 接続 PCIe レーン数に比例して RMW 処理回路の競合度合が大きくなった結果、プロセッサの I/O 処理効率が低下し、PCIe スループットの性能低下率が大きくなったと考えられる。

上記性能低下の対策として、ソフトウェアを修正してメモリレイアウトやデータ管理方法を変える代わりに、表 2 のケース(B)のデータ転送を 8 回繰り返さずに、ユーザーデータと T10-DIF を分離し T10-DIF を最後にまとめて転送する表 2 のケース(F)で転送する事で 64B 境界のずれを解消する方法が考えられる。もしくは、PCIe デバイスのハードウェアに待ち合わせバッファを設けて回路規模拡大を許容する代わりに、待ち合わせバッファ内で 64B 境界に沿うように PCIe ライトパケットを分割処理してからメモリライトを実施する事で 64B 境界のずれを解消する方法も考えられる。これらの対策を実施する事でプロセッサの PCIe 帯域を使い切る事が可能になると考えられる。

6. 結論

汎用プロセッサの PCIe データ転送効率は PCIe デバイスのデータ転送パターンにおける 64B 非アライメントメモリアクセス発生頻度によってデータ転送効率が変化し、更に接続レーン数を増やすに従って、その低下率は拡大する事がわかった。

参考文献

[1] John Gantz, et al, “THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East”, IDC – iView, DEC 2012
[2] Keith Holt, “T10/03-224”, T10 Technical Committee, JULY 2003
[3] L. Rota, et al, “A PCIe DMA Architecture for Multi-Gigabyte Per Second Data Transmission”, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 62, NO. 3, JUNE 2015