Vol.2015-ARC-215 No.8
Vol.2015-OS-133 No.8
2015/5/27

ストレージシステムにおける大容量データ転送時の経路制御方式

鈴木貴敦[†1]　吉原朋宏[†1]

概要：　複数のインタフェースを用いる大規模なストレージシステムにおいて、高いシーケンシャルリード性能実現に必要な、キャッシュメモリへのデータ転送経路制御方式を検討した。本研究では、複数 DMA のデータ転送経路の競合により性能低下が発生することを明らかにし、転送先キャッシュメモリに合わせて転送経路を選択し競合を回避する方式を提案した。プロトタイプにて経路制御方式の実測評価を行い、提案手法の有効性を確認した。

1. はじめに

1.1 ストレージシステムとは

ストレージシステムとは，多数のホストコンピュータを接続可能な，大容量・大規模なデータ記憶装置である．ストレージシステムは，一般に，多数の HDD や SSD 等の記憶媒体と，それらの記憶媒体とコンピュータの間を取り持つコントローラから構成される(図 1)．記憶媒体は，信頼性向上のため，Redundant Array of Independent Disk（RAID）[1]構成を取ることが多い．

コントローラは，以下の 4 つの要素から構成される．それぞれの要素は，信頼性と可用性の向上のため，冗長化されている．

- ホストコンピュータとの接続インタフェース(以後，ホストインタフェース)
- 記憶媒体に格納するデータを一時的に保持するキャッシュメモリ(以後，キャッシュメモリ)
- 記憶媒体との接続インタフェース(以後，記憶媒体インタフェース)
- ホストインタフェース，キャッシュメモリ，記憶媒体インタフェースを制御する制御部(以後，制御部)

ストレージシステムは，企業の基幹系システムに使われることが多く，その場合，高い信頼性と可用性はもちろんのこと，多数のホストコンピュータを接続することから，柔軟に性能設計変更が可能なものが求められる．

このニーズに対し，ホストインタフェース，キャッシュメモリ，記憶媒体インタフェース，制御部の各部を分離し，独立して増減設可能なストレージコントローラアーキテクチャがある[2][3]．このアーキテクチャでは，例えば記憶媒体を増やしたい場合は，記憶媒体インタフェースを増設すると言った，必要に応じたシステム構成を取ることが可能である．

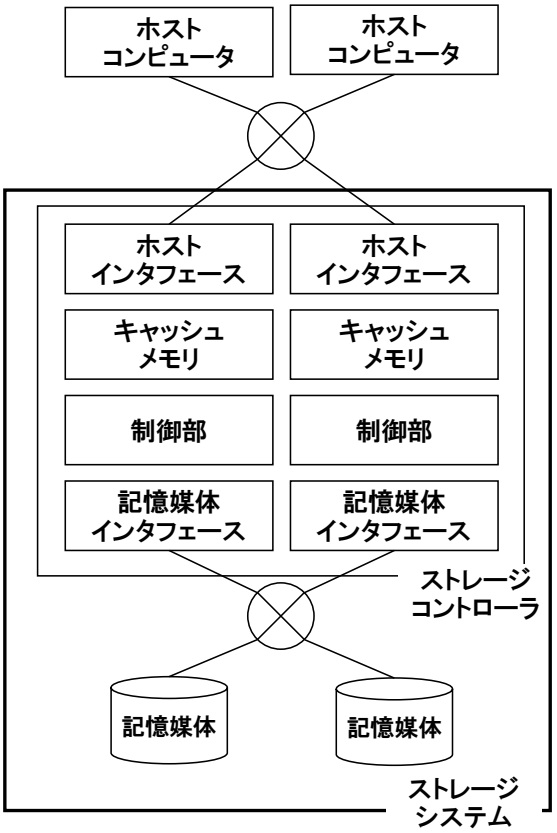

図 1　ストレージシステム概要

1.2 ストレージシステムへの I/O 種別と本研究の目的

ストレージシステムへの I/O は，ランダムアクセスとシーケンシャルアクセスの 2 つに分けることができる．

- ランダムアクセス：データの転送長が小さく，ホストコンピュータがアクセスするアドレスが非連続であるアクセスパターンのこと
- シーケンシャルアクセス：データの転送長が大きく，ホストコンピュータがアクセスするアドレスが連続であるアクセスパターンのこと

これまで，ストレージシステムには主に Online Transaction Processing (OLTP) 処理の高速化のために，高いランダムアクセス性能が求められていた[4]．近年では，ラ

†1 (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 Telecommunications

Vol.2015-ARC-215 No.8
Vol.2015-OS-133 No.8
2015/5/27

ンダム性能に加えて, Hadoop などのビッグデータ解析処理の効率向上のため，大量データを読み込むシーケンシャルリード性能が求められている[5].

ストレージシステムにおいて，高いシーケンシャルリード性能を出すためには，データを転送するハードウェアの帯域と，ハードウェアの帯域を使い切るためのデータの転送経路制御方式が必要である.

本研究の目的は，ストレージシステムを構成するハードウェアの持つシーケンシャルリード性能を出し切るためのデータ転送経路制御の方式を検討することである.

2. 研究対象

2.1 ハードウェア構成

以下の図 2 に，本研究で前提とするストレージシステムのハードウェア構成を示す．ストレージシステムは，ストレージコントローラと記憶媒体からなる．ストレージコントローラは，ホストインタフェース部，記憶媒体インタフェース部，制御部，キャッシュメモリ部，及びそれらを互いに接続するネットワーク部からなり，それぞれ信頼性・可用性の向上のため，冗長化されている.

記憶媒体インタフェースは以下の 3 つの要素から構成される．内部ネットワークを介して，それぞれが相互に接続されている.

- 記憶媒体と通信するためのプロトコルチップ(以後，プロトコルチップ)
- 記憶媒体へ転送するデータ，記憶媒体から転送されたデータを一時的に保持するためのバッファメモリ
- キャッシュメモリとバッファメモリ間のデータ転送を行う Direct Memory Access (DMA) エンジン

データは，ネットワーク部，および記憶媒体インタフェース内部ネットワーク上をあるサイズのパケットに分割されて，シリアルに転送される．また，ネットワーク部の帯域は，記憶媒体インタフェース内部ネットワークの帯域よりも大きいものとする.

また，記憶媒体には，デュアルポートの SAS ドライブを利用する[6].

2.2 シーケンシャルリード処理概要

シーケンシャルリードでは，以下の流れにそってデータをホストまで転送する.

(1) 制御部が、ホストインタフェースを経由してサーバからリード要求を受信する.
(2) 制御部が，より使用量の少ないキャッシュメモリを選択する.
(3) 制御部が，より低負荷な記憶媒体インタフェースを選択する.

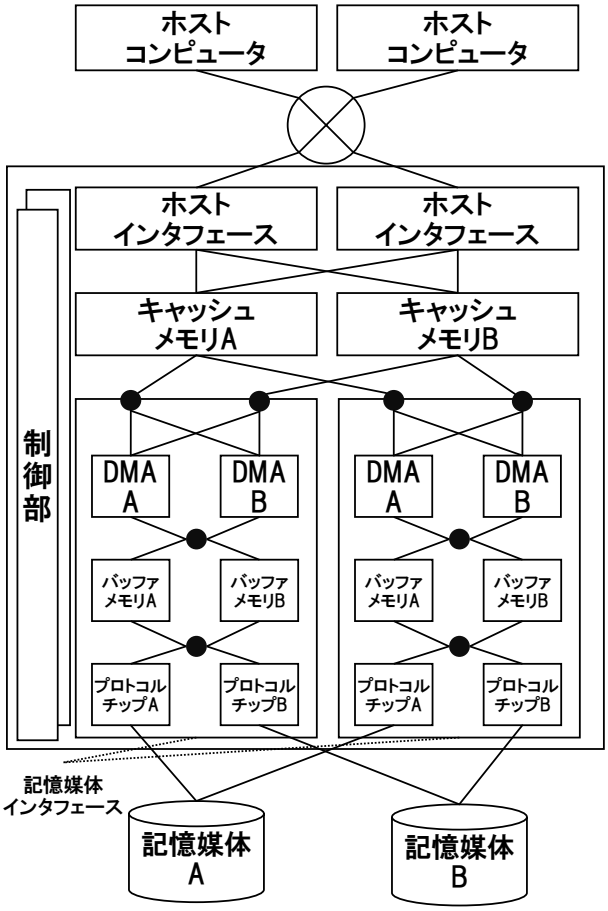

図 2 ハードウェア構成概要

(4) プロトコルチップが記憶媒体からバッファメモリへデータを転送する.
(5) DMA がバッファメモリからキャッシュメモリへデータを転送する.
(6) リード要求を受信したホストインタフェースがキャッシュメモリからホストコンピュータへデータを転送し，処理が完了する

本研究では，特に記憶媒体インタフェース内の DMA エンジンが性能ボトルネックとなる場合に，シーケンシャルリード性能を最大限引き出すための, 記憶媒体-キャッシュメモリ間のデータ転送経路制御方式を明らかにする.

3. 課題

DMA エンジンのデータ転送能力の低下要因となるのは，データ転送経路の競合である(図 3)．データ転送経路の競合は，同一記憶媒体インタフェース内の複数 DMA エンジンが, 同じ経路を使用してデータ転送する場合に発生する．データ転送経路の競合が発生すると，DMA エンジンが稼働時間あたりに転送可能なデータ量が減少するため，シーケンシャルリード性能が低下する．そのため，DMA エンジンが使用するデータ転送経路での競合を少なくすること

Vol.2015-ARC-215 No.8
Vol.2015-OS-133 No.8
2015/5/27

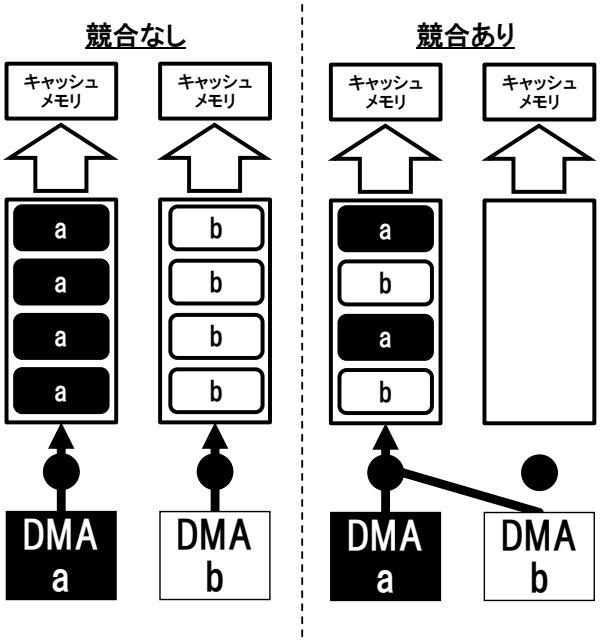

図 3　DMA によるデータ転送(競合ありの場合と競合なしの場合との比較)

が課題である.

4. データ転送経路制御方式

4.1 転送経路制御方式の制約

記憶媒体-キャッシュメモリ間のデータ転送経路制御方式は，キャッシュメモリの負荷分散や交換・増設に対応するため，記憶媒体から任意のキャッシュメモリへ転送可能なものにしなければならない．以後，この制約を守った経路制御方式を 4 つ挙げる.

4.2 経路制御方式 1：キャッシュメモリ-DMA エンジン間経路切替方式

図 4 に，キャッシュメモリ-DMA エンジン間経路切替方式の概要を示す．本方式では，転送先のキャッシュメモリに合わせて，DMA エンジンが使用するデータ転送経路を選択する方式である．具体的には，記憶媒体-キャッシュメモリ間を以下の流れに沿ってデータを転送する.

(1) プロトコルチップが，記憶媒体からバッファメモリへデータを転送する．このとき，プロトコルチップは同一記憶媒体インタフェース内で，同一の ID が割り振られているバッファメモリへ転送する

(2) DMA エンジンが，バッファメモリからキャッシュメモリへデータを転送する．このとき，同一記憶媒体インタフェース内に関して，データ転送元のバッファメモリと同一の ID が割り振られている DMA エンジンを利用する.

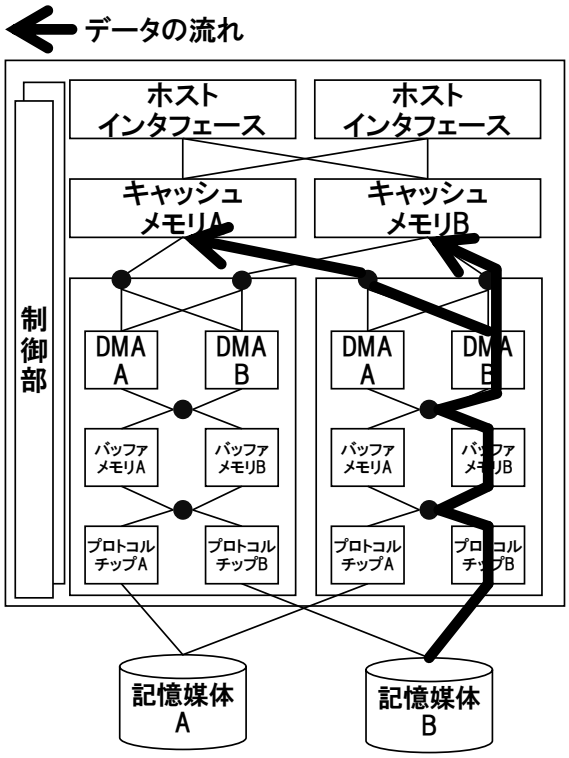

図 4　キャッシュメモリ-DMA エンジン間経路切替方式
(片側の記憶媒体インタフェース内経路線省略)

4.3 経路制御方式 2：DMA エンジン-バッファメモリ間経路切替方式

図 5 に， DMA エンジン-バッファメモリ間経路切替方式の概要を示す．本方式では，転送先のキャッシュメモリに合わせて，DMA エンジン-バッファメモリ間のデータ転送経路を選択する方式である．具体的には，記憶媒体-キャッシュメモリ間を以下の流れに沿ってデータを転送する.

(1) プロトコルチップが，記憶媒体からバッファメモリへデータを転送する．このとき，プロトコルチップは同一記憶媒体インタフェース内で，同一の ID が割り振られているバッファメモリへ転送する

(2) DMA エンジンが，バッファメモリからキャッシュメモリへデータを転送する．このとき，同一記憶媒体インタフェース内に関して，データ転送先のキャッシュメモリと同一の ID が割り振られている DMA エンジンを利用する.

4.4 経路制御方式 3：バッファメモリ-プロトコルチップ間経路切替方式

図 6 に， バッファメモリ-プロトコルチップ間経路切替方式の概要を示す．本方式では，転送先のキャッシュメモリに合わせて，バッファメモリ-プロトコルチップ間のデータ転送経路を選択する方式である．具体的には，記憶媒体-キャッシュメモリ間を以下の流れに沿ってデータを転送す

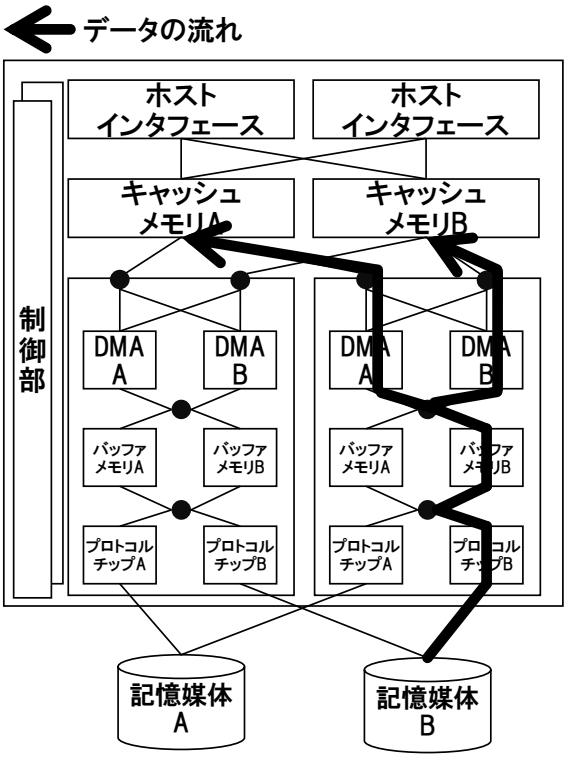

図 5　DMA エンジン-バッファメモリ間経路切替方式
(片側の記憶媒体インタフェース内経路線省略)

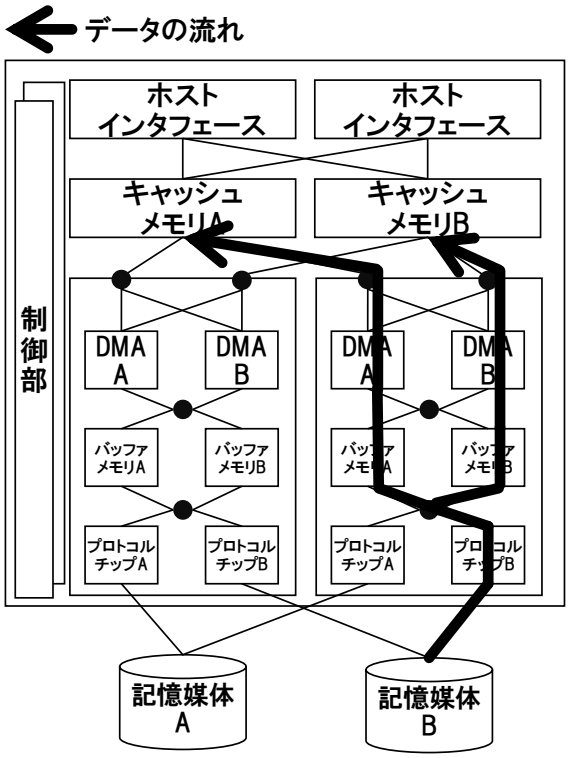

図 6　バッファメモリ-プロトコルチップ間経路切替方式
(片側の記憶媒体インタフェース内経路線省略)

る．

(1) プロトコルチップが，記憶媒体からバッファメモリへデータを転送する．このとき，プロトコルチップは同一記憶媒体インタフェース内で，転送先のキャッシュメモリと同一の ID が割り振られているバッファメモリへ転送する

(2) DMA エンジンが，バッファメモリからキャッシュメモリへデータを転送する．このとき，同一記憶媒体インタフェース内に関して，データ転送先のキャッシュメモリと同一の ID が割り振られている DMA エンジンを利用する．

4.5 経路制御方式 4：記憶媒体インタフェース切替方式

図 7 に， 記憶媒体インタフェース切替方式の概要を示す．本方式では，信頼性・可用性向上のために冗長化されている記憶媒体インタフェースに着目し，それぞれを別のキャッシュメモリへのデータ転送経路とみなす．具体的には，記憶媒体-キャッシュメモリ間を以下の流れに沿ってデータを転送する．

(1) プロトコルチップが，記憶媒体からバッファメモリへデータを転送する．このとき，一方の記憶媒体インタフェースのプロトコルチップは，同一記憶媒体インタフェース内で，同一の ID が割り振られているバッファメモリへデータを転送する．もう一方の記憶媒体インタフェースのプロトコルチップは，異なる ID が割り当てられているバッファメモリへデータを転送する．

(2) DMA エンジンが，バッファメモリからキャッシュメモリへデータを転送する．このとき，同一記憶媒体インタフェース内に関して，データ転送先のキャッシュメモリと同一の ID が割り振られている DMA エンジンを利用する．

5. 評価

5.1 机上評価

4 章で挙げた 4 つの転送経路制御方式に関して，それぞれの性質について机上で比較評価を行う．経路競合観点での評価まとめを表 1 に示す．

キャッシュメモリ-DMA エンジン間経路切替方式，DMA エンジン-バッファメモリ間経路切替方式，バッファメモリ-プロトコルチップ間経路切替方式では，複数の記憶媒体から同時にデータを転送する場合，それぞれ経路を切り替える箇所で，経路競合が発生する． 図 8 に，キャッシュメモリ-DMA エンジン間経路切替方式での例を示す．

一方，記憶媒体インタフェース切替方式では，複数の記憶媒体から同時にデータを転送しても，経路競合が発生しない．そのため，シーケンシャルリード性能が最も高くなるのは，記憶媒体インタフェース切替方式であると考えられる．

Vol.2015-ARC-215 No.8
Vol.2015-OS-133 No.8
2015/5/27

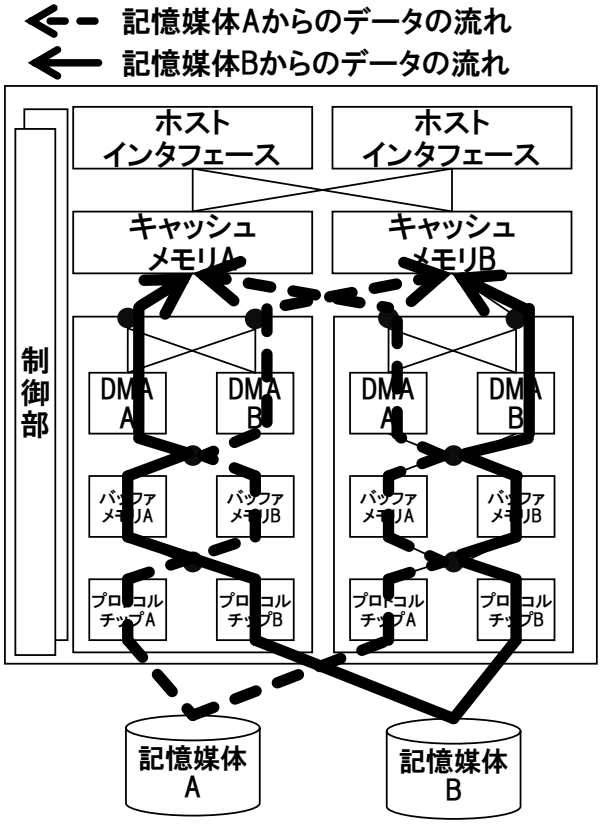

図 7 記憶媒体インタフェース切替方式

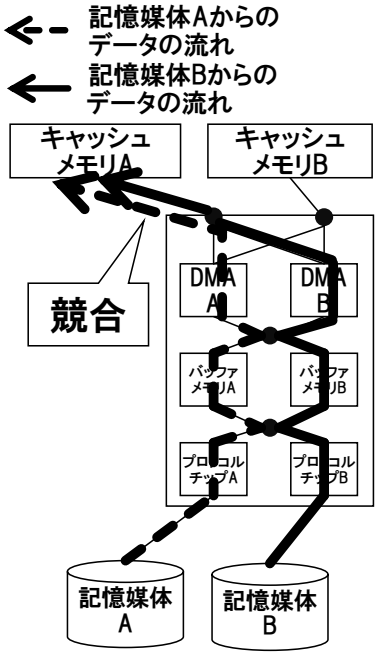

図 8 キャッシュメモリ-DMA エンジン間経路切替方式における経路競合

表 1 机上評価まとめ(競合なし:○，競合あり:×)

	方式 1	方式 2	方式 3	方式 4
DMA エンジン-キャッシュメモリ間競合	×	○	○	○
DMA エンジン-バッファメモリ間競合	○	×	○	○
バッファメモリ-プロトコルチップ間競合	○	○	×	○

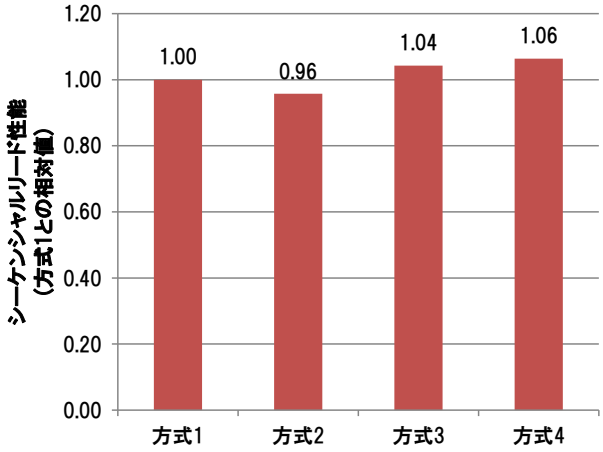

図 9 各方式の実測結果(※方式 1 からの相対値)

5.2 実測評価

図 2 で示されるハードウェアアーキテクチャを持つストレージシステム上に，4 章の 4 つの方式を実装し，シーケンシャルリード性能の測定を行った．結果を図 9 にまとめる．机上評価の通り，データ転送経路上で競合が発生しない記憶媒体インタフェース切替方式を適用した場合に，最もシーケンシャルリード性能が高くなる結果が得られた．

5.3 考察

今回測定を行ったストレージシステムでは，記憶媒体インタフェース内の DMA エンジンがボトルネックであるため，キャッシュメモリ-DMA エンジン間経路切替方式，DMA エンジン-バッファメモリ間経路切替方式，バッファメモリ-プロトコルチップ間経路切替方式の中では，DMA エンジンに関係するデータ転送経路上で競合が発生しないバッファメモリ-プロトコルチップ間経路切替方式を適用した場合に，シーケンシャルリード性能が高くなった．バッファメモリ-プロトコルチップ間経路切替方式は，性能ボトルネックである DMA エンジンが直接使用するデータ転送経路上では競合が発生しないが，バッファメモリ-プロト

コルチップ間のデータ転送経路上で競合が発生するため，記憶媒体インタフェース切替方式と比較して性能低下が見られた．

キャッシュメモリ-DMA エンジン間経路切替方式とDMA エンジン-バッファメモリ間経路切替方式は，いずれも DMA エンジンに関係するデータ転送経路上で競合が発生しているが，記憶媒体インタフェース内部のネットワーク帯域のほうが，キャッシュメモリ-DMA エンジン間のネットワーク帯域よりも小さいこと，また，バッファメモリの帯域が，キャッシュメモリの帯域よりも小さいことが影響したためであると考えられる．

6. おわりに

本稿では，ストレージシステムにおいて，ハードウェアのシーケンシャルリード性能を出し切るためのデータ転送経路制御方式について述べた．4 つの制御方式について，データ転送経路競合の観点から評価を行った．その結果，記憶媒体インタフェースを切り替える方式が最もシーケンシャルリード性能が高かった．

今後，今回の結果を元に，ストレージシステムのシーケンシャルリード性能見積もり手法を検討する．

参考文献

1) David A. Patterson, Garth Gibson, Randy H. Katz: A case for redundant arrays of inexpensive disks(RAID), Proceedings of the ACM SIGMOD international conference on Management of data, pp.109-116 (1988).
2) Josh Krischer: The Virtual Storage Platform (VSP) from Hitachi Data Systems-Setting New Levels of Excellence, Josh Krischer & Associates GmbH (2010).
3) 高田正法，下薗紀夫，藤本和久，坂下悠貴，藤林昭，細谷睦: スケーラブルストレージシステムにおけるアクセス要求振り分け方式，電子情報通信学会技術研究報告. CPSY，コンピュータシステム ，pp.25-30 (2013)
4) 早水悠登，合田和生，中野美由紀，喜連川優：オンライントランザクション処理における高速フラッシュストレージの性能活用に関する実験的考察，情報処理学会第 74 回全国大会，1N-5 (2012)
5) Apache Hadoop, https://hadoop.apache.org/
6) エンタープライズ SSD インターフェイスの比較, http://www.seagate.com/files/www-content/product-content/_cross-product/ja/docs/enterprise-interface-comparisons-tp625-1-1203jp.pdf