バーストバッファの性能評価

尾形 幸亮 [1)], 池田 健二 [1)], 疋田 淳一 [1)]

1) 京都大学 企画・情報部

ogata.kosuke.7c@kyoto-u.ac.jp

A Performance Evaluation of Burst Buffers

Kosuke Ogata [1)], Kenji Ikeda [1)], Junichi Hikita [1)]

1) Planning and Information Management Department, Kyoto Univ.

概要

本学のスーパーコンピュータシステム には、ストレージシステムの新たな機能として、バーストバッファと呼ばれる SSD により構成された高速な一時ストレージを導入している。バーストバッファにより、従来システムでは性能の不足しがちな、転送ブロックサイズの小さい IO や、single-shared ファイルの IO について高速化が期待できるため、本稿ではベンチマークソフトを用いて性能測定を行った結果について報告する。

1. はじめに

京都大学学術情報メディアセンター（以下、当センター）では、スーパーコンピュータシステムとして、Cray XC40（システム A）、Cray CS400 2820XT（システム B）、Cray CS400 4840X（システム C）の 3 種類の演算システム及びストレージを運用している。ストレージは、Lustre による分散ファイルシステムと、バーストバッファと呼ばれる SSD により構成された高速な一時ストレージを導入している。従来のストレージシステムでは読み書きの速度が不足する状況における性能改善を意図している。本稿では、まず当センター保有のスーパーコンピュータシステムの概要を述べたのち、同システムに配置されている 2 種類のバーストバッファについて、仕様を説明する。続いて、バーストバッファの性能測定について条件と結果を述べ、結果について考察する。

2. スーパーコンピュータシステム概要

スーパーコンピュータシステムの機器構成を図 1 に示す[1]。当センターは 3 種の演算システムを保有しており、演算性能を重視したシステム A、研究室とのクラスタとの親和性・連続性を重視したクラスタタイプのシステム B、大規模メモリを必要とするユーザのためのシステム C により構成している。

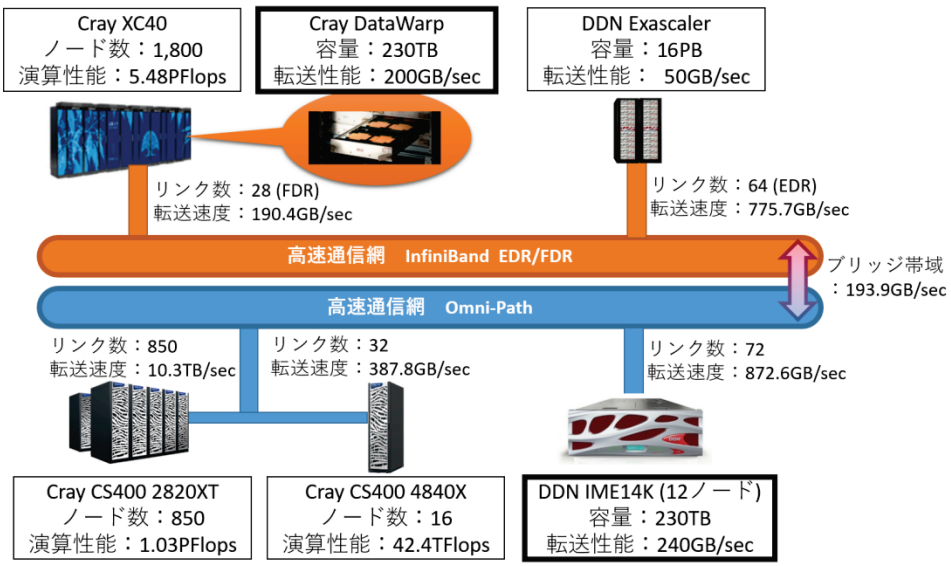

図 1 システム構成

これに加えて、ユーザデータを格納するストレージシステムとバーストバッファを保有している。本稿ではシステムAとBをストレージおよびバーストバッファを用いて性能評価を行っており、以下に構成について説明する。

2.1. システム A

システム A は、Intel 社の最新の Xeon Phi (Knights Landing)プロセッサを1基搭載したノードを 1,800 台接続した大規模並列コンピュータである。各ノードには68個の CPU コアを搭載し、96GB のメモリを持っている。ストレージシステムとは、InfiniBand FDR x28(190.4GB/s)で接続されている。

2.2. システム B

システム B は、Intel 社の最新の Xeon プロセッサ (Broadwell-EP)を 2 基搭載したノードを Omni-Path ネットワークにより 850 台接続した大規模クラスタシステムである。各ノードには 36 個の CPU コアを搭載し、128GB のメモリを持っている。ストレージシステムとは、Omni-path と InfiniBand のブリッジを介して接続されており、ネットワーク帯域は、193.9GB/s である。

2.3. ストレージ

ストレージシステムとして、DDN 社の EXAScaler による Lustre ファイルシステムを有している。メタデータサーバ（MDS）6 ノード、オブジェクトストレージサーバ（OSS）18 ノードで構成され、物理容量は 16PB、転送性能は 100GB/s である。今回の性能評価で用いたファイルシステムは、MDSx2、OSSx8、転送性能 50GB/s の構成である。システム A との間は InfiniBand EDR および FDR により接続され、システム B および C との間は InfiniBand と Omni-Path のブリッジサーバを経て接続される。

2.4. バーストバッファ

バーストバッファは、Lustre ストレージが苦手とする I/O を補助する目的で導入しており、MPI プロセスが共通のファイル（以下、single-shared ファイル）に行う I/O 性能の改善や、小さい I/O（一度の転送ブロックサイズは小さいが、トータルの転送量は多い状況）における性能改善が期待されるものである。特に single-shared ファイルの性能が高ければ、大規模並列時のファイル I/O において、メタデータアクセスが過大になってしまう点を抑えることができる。当センターでは、バーストバッファとして、システム A 向けに Cray 社の DataWarp、システム B および C 向けに DDN 社の IME を導入している（図 1の太枠で囲われた機器）。

以下、DataWarp、IME の順に、構成、理論性能、利用形態について説明する。

2.4.1. DataWarp

DataWarp は、システム A の計算ノードと同様に Cray 社の Aries インターコネクトに接続されており、総容量は 230TB、転送速度は 200GB/s の構成である。Lustre ファイルシステムの 50GB/s の転送性能と比較し 4 倍高速である。

DataWarp のハードウェアは、2 個の SSD を搭載したブレードを単位としている。SSD の容量は 1 個当たり 3.2TB である[2][3]。このブレード 36 台が IO ノードに装着されており、ここから直接 Aries インターコネクトに接続される。

現段階の DataWarp の利用形態は、ユーザプログラムによるファイルのステージングを伴う高速なスクラッチ領域である。将来的には Lustre との間で透過的に I/O を高速化するストレージキャッシュとして働くようにソフトウェアアップデートが計画されている。

2.4.2. IME

IME は、システム B の計算ノードと同一ネットワークである Omni-Path で接続されている。総容量は 230TB、転送速度は 240GB/s であり、50GB/s の Lustre ファイルシステムよりも転送速度が 4.8 倍高速である。

IME のハードウェアは、IME14K を 6 筐体(12 ノード)導入している。IME14K は NVMe SSD(800GB)を 1 ノードあたり 24 個搭載している[4]。システム B とは Omni Path x72 (872.6GB/s)、Lustre とは変換スイッチを介して InfiniBand EDR x16(193.9GB/s、システム B および C と共用)で接続される。

IME は Lustre のファイルシステムとメタデータを共有し、透過的にファイルアクセス可能な機構を備えている。ファイル読み込みは、直接 Lustre を参照するが、事前に IME 上にプリフェッチすることも可能である。ファイル書込みは IME 上に書き込んだ後、IME のコントロールコマンドにより、明示的に Lustre 側にシンクさせる必要がある[5]。当センターの運用ではジョブスケジューラのジョブ終了処理にシンクコマンドが自動で実行されるよう設定することで、ストレージのキャッシュ領

域としての透過的な利用を実現している。

3. 性能測定

今回の性能測定では、ベンチマークソフトとして IO スループットベンチマークである IOR[6]と、より実際の応用を想定し、並列計算ベンチマークスイートである NAS Parallel Benchmark[7]（以下、NPB）を用いた。NPB については、サブセットである BT-IO (full) を用いている。BT-IO は、ディスク IO が比較的多いブロック三重対角行列の計算を行うベンチマークであり、single-shared のファイルにランダムに近いアクセスを行うことが特徴である。両者のバージョン、コンパイル環境および動作環境は表 1 のとおりである。

表 1 ソフトウェア環境

項目	値
ソフトウェア・バージョン	IOR 2.10.3 NPB 3.3.1
コンパイル環境	Intel 17.0.2
最適化オプション	-O2
MPI ライブラリ	Cray MPICH 7.5.3 (システム A) Intel MPI 2017.2 (システム B)

3.1. 測定条件

IOR は、Lustre とバーストバッファの転送性能を評価するために、表 3 に示す 4 つの条件を設定した。条件 1 は、単一ノードの転送性能を測るためにプロセス数を変化させて測定した。MPI プロセスごとに別ファイルに対して I/O を行う方式（以下、file-per-process）を選択している。また、プロセスあたりの処理するファイルサイズと転送時のブロックサイズを同一としている。条件 2 は、ノード当たりのプロセス数を 10 で固定し、複数ノードにおける転送性能を測定した。他の条件は、条件 1 と同じである。条件 3 は、single-shared ファイルに対する性能を測定であること以外は、条件 2 と同じである。条件 4 は、プロセス数及びノード数を 10 で固定し、I/O における 1 回あたりの転送サイズを変化させて測定した。特に小さい転送サイズにおける I/O 性能を確認する意図である。

アプリケーションベンチマークである NPB BT-IO は、クラス D の問題サイズ を用いている。プロセス数は N の二乗である制約があるが、各プロセスを最小のノードに詰め込む形で測定した。測定条件を表 2 に示す。

いずれの測定においても、3 回以上測定した最大性能の値を採用した。

表 2 NAS Parallel Benchmark の測定条件

パラメータ	値
種類	BT-IO (full)
行列の規模	408x408x408 の行列式 （クラス D）
作成されるファイルのサイズ	126GB
総プロセス数 (カッコ内はノード数)	システム A： 16(1),36(1),49(1),64(1), 81(2),100(2),121(2),144(3), 169(3),196(3),225(4),256(4) システム B： 16(1),36(1),49(2),64(2), 81(3),100(3),121(4),144(4)

表 3 IOR の測定条件

項目	条件 1	条件 2	条件 3	条件 4
I/O API	POSIX			
ファイルアクセス	File-per-process	File-per-process	Single-shared	File-per-process
アクセスオーダ	Sequential			
計算ノード数	1	1,2,4,6,8,10	1,2,4,6,8,10	10
ノードあたりプロセス数	1,4,8,16,32	10	10	10
プロセスあたりのファイルサイズ	10GB			
1 回あたり転送サイズ	10GB	10GB	10GB	16KB,32KB,64KB, 128KB, 256KB, 512KB, 1MB,2MB

4. 測定結果と考察

4.1. 単一ノードでの測定結果(条件１)

まず、IOR を用いた条件 1 での測定結果について、リード速度を図 2、ライト速度を図 3 に示す。

図 2、図 3 の通り、Lustre はプロセス数の増加とともに転送速度が少しずつ改善するのに対し、DataWarp 及び IME では 4~8 プロセスで高い性能を示し、それ以降は最大値の約半分の性能となっており、現状の DataWarp 及び IME のクライアントの実装では、1 ノード当たりのプロセス数が増えると、性能が低下することが分かる。

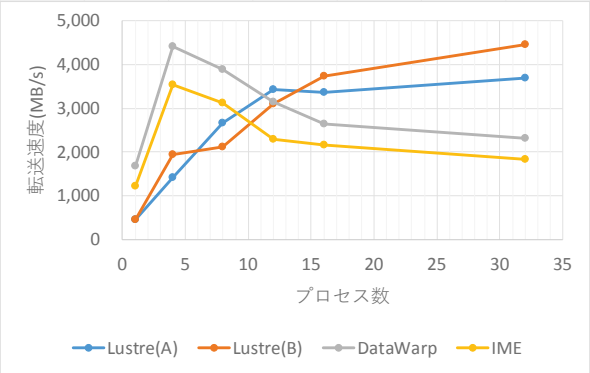

図 2 単一ノード（条件１）でのリード速度

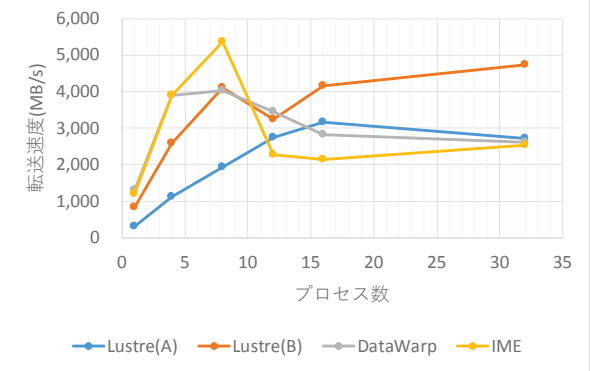

図 3 単一ノード（条件１）でのライト速度

4.2. 複数ノードでの測定結果(条件 2)

次に、IOR を用いた条件 2 での測定結果について、リード速度を図 4、ライト速度を図 5 に示す。

DataWarp、IME ともに、計算ノードを増やしてクライアントの処理能力およびインターコネクトの帯域幅が増えるのと比例し、転送速度が向上している。10 ノード時の転送速度では、DataWarp が Lustre との比でリード 3.5 倍、ライト 3.8 倍であり、IME が Lustre との比でリード 2.0 倍、ライト 2.4 倍である。多ノードを使用した File-per-process においてバーストバッファが Lustre に対してアドバンテージを得られた結果となっている。

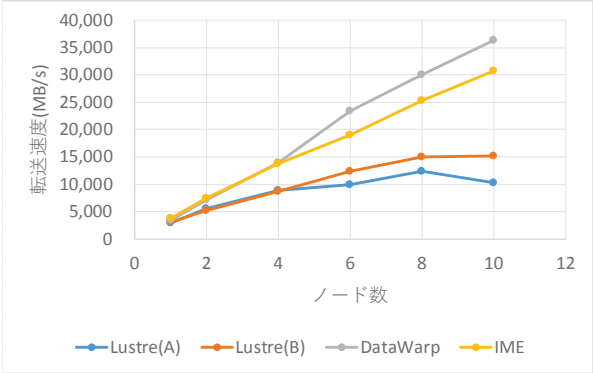

図 4 file-per-process でのリード速度(条件 2)

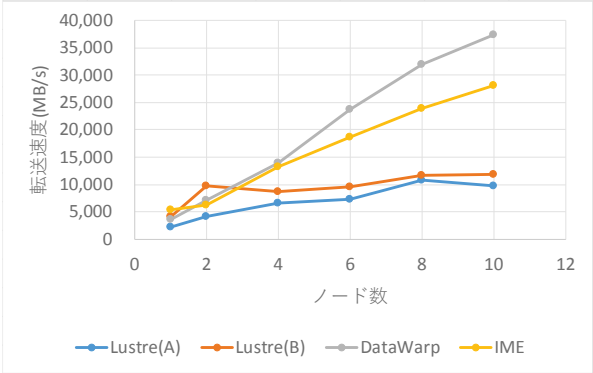

図 5 file-per-process でのライト速度(条件 2)

4.3. single-shared での測定結果(条件 3)

次に、IOR を用いた条件 3 での測定結果について、リード速度を図 6、ライト速度を図 7 に示す。

DataWarp では計算ノードを増やしてクライアントの処理能力およびインターコネクトの帯域幅が増えるのと比例し、転送速度が向上しており、10 ノードではリード 36104MB/s、ライト 38184MB/s である。一方、Lustre では計算ノードの数によらず転送速度が低迷しており、10 ノード時にリード 1002MB/s、ライト 1128MB/s となっている。DataWarp と Lustre の 10 ノード時のライト速度の比は、33.9 倍となっている。

IME についても、DataWarp と同様により転送速度が向上している。10 ノード時、IME の転送速度はリード 29390MB/s、ライト 14404 MB/s である。なお、本来ライト性能もリードと同程度の性能がでる想定のため、この点は調査中である。一方、Lustre はリード 951MB/s、ライト 1085MB/s であり、IME との比はリードで 30.9 倍、ライトで 13.3 倍と

なっている。

バーストバッファで single-shared ファイルが高速に扱えることで、Lustre における file-per-process の I/O では超大規模並列を行う際のストレージのメタデータへの負荷が大きくなる問題点を改善することができるため、大規模計算時の速度改善が期待できる。

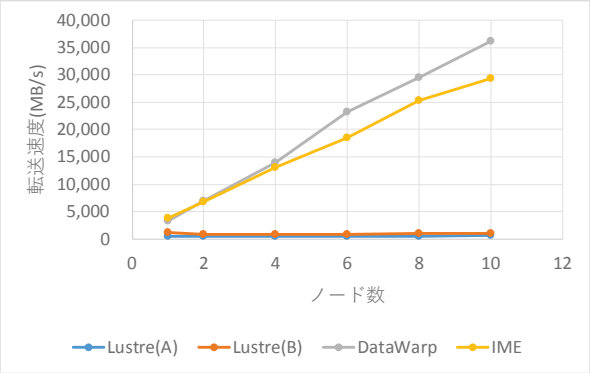

図 6　single-shared（条件 3）でのリード速度

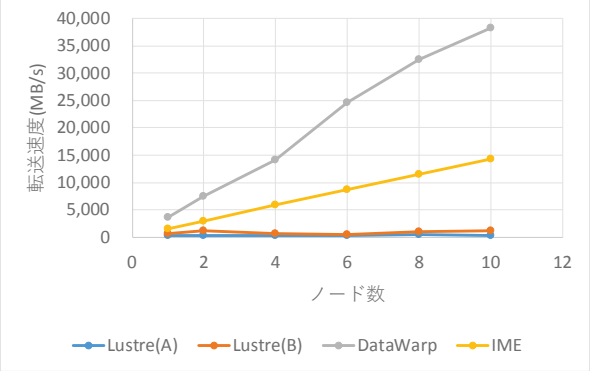

図 7　single-shared（条件 3）でのライト速度

4.4.　小さい IO の測定結果(条件 4)

次に、IOR を用いた条件 4 での測定結果について、リード速度を図 8、ライト速度を図 9 に示す。

DataWarp は小さすぎる I/O は不得手であるが、128KB 以降で Lustre を上回る性能を示している。IME については、32KB までは Lustre が優位であり、それ以降に Lustre を上回る性能を示している。

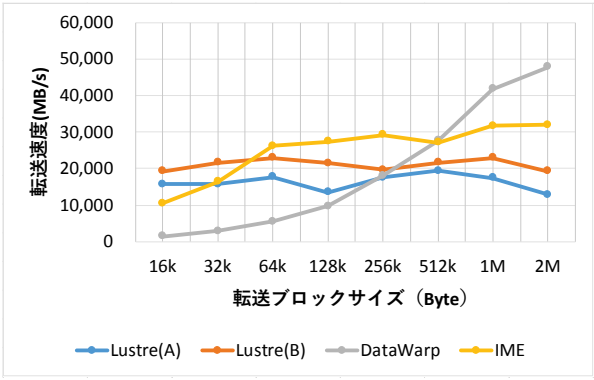

図 8　小さい IO（条件 4）でのリード速度

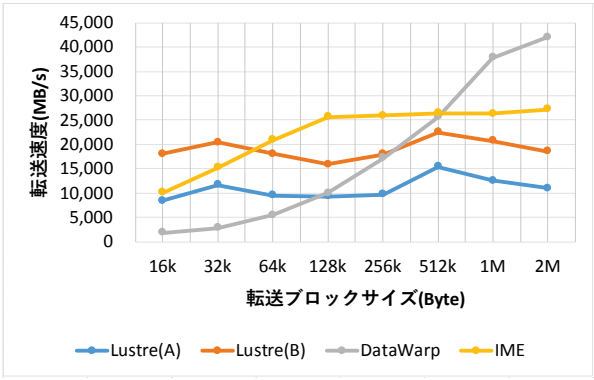

図 9　小さい IO（条件 4）でのライト速度

4.5.　NAS Parallel Benchmark の測定結果

NPB BT-IO を用いて測定した結果を示す。まず、システム A での DataWarp と Lustre の比較を図 10 に示す。

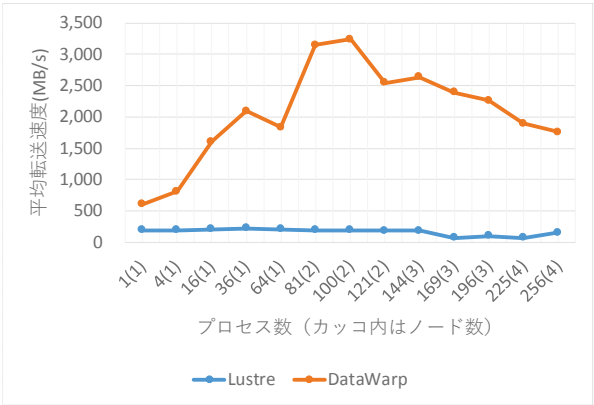

図 10　転送速度（DataWarp 対 Lustre）

DataWarp は、Lustre に比べ最大 16.8 倍の転送速度向上が見られたが、100 プロセス以降は転送速度が下がっている。転送速度が下がる原因として、126GB のデータ量では、100 プロセス超の並列処理には不足していることが原因と考えられるが、それまでは順調に並列化効果を得られていると言える。より大規模な並列性能を確認するために、

データサイズが大きいクラス E の追加測定を行った。なおクラス E では、総データサイズは 2072GB である。測定結果を図 11 に示す。400（20x20）～784（28x28）プロセスでは、並列化効果が得られており、データ量が十分であれば、100 プロセス以上でも転送性能が伸びていることが確認できる。400 プロセス以下の実行は、ノードのメモリが不足するため測定していない。1024（32x32）プロセス以降は、転送性能が低下し横ばいになっているが、1024 並列ではデータサイズが約 2GB (2072GB / 1024) となり、クラス E でも 1024 並列以上ではデータ量が不足するためと考えられる。

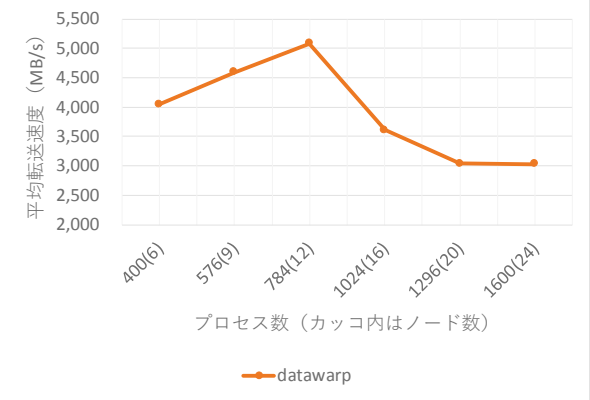

図 11　転送速度（DataWarp、クラス E）

次に、システム B で IME と Lustre の比較を行ったものを図 12 に示す。IME では、Lustre に比べ最大 8.0 倍の転送速度向上が見られた。IME についてもプロセス当たりのデータ量が 2GB 以下となる 64 プロセス以降に DataWarp 同様転送性能が落ちる結果となった。なお、大規模な計算ノードの確保の確保ができなかったため、クラス E については未測定である。

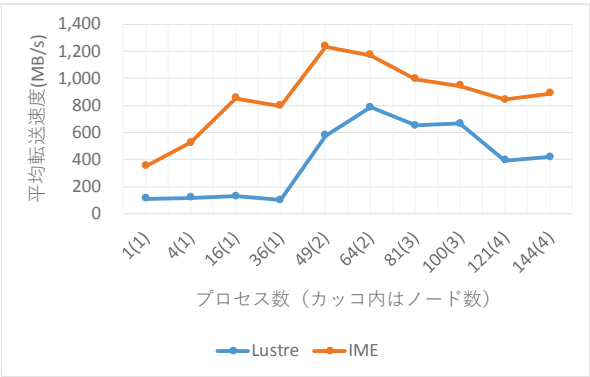

図 12　転送速度（IME 対 Lustre）

5.　おわりに

本稿では、当センターのスーパーコンピュータシステムに配置されている 2 種類のバーストバッファおよび Lustre ストレージシステムについて、仕様と性能測定の結果と考察を述べた。

今回はプログラムの改変を伴わずに扱える POSIX インターフェイスによる性能を評価したが、IME にはより高速にアクセス可能な MPI-IO（ROMIO）インターフェイスがあるため、今後の追加調査についても検討している。DataWarp 及び IME は、初期バージョンがリリースされたばかりの発展途上中の状態であるため、今後のソフトウェアアップデートによる機能追加及び性能の改善が期待されるため、2018 年度に予定している本格運用開始に向けて継続的に性能評価を行いたいと考えている。

参考文献

[1] 京学術情報メディアセンター、“スーパーコンピュータシステム”、京都大学、http://www.iimc.kyoto-u.ac.jp/ja/services/comp/supercomputer (参照 2017-09-27)

[2] “Cray XC40 DataWarp Applications I/O Accelerator”、Cray Inc.、http://www.cray.com/sites/default/files/resources/CrayXC40-DataWarp.pdf (参照 2017-09-27)

[3] “Accelerate your IO with the NERSC Burst Buffer”、NERSC、https://www.nersc.gov/assets/Uploads/DataDay-BurstBuffer.pdf (参照 2017-09-27)

[4] “IME - アプリケーション認識型 I/O アクセラレーションおよびバーストバッファ”、DDN Inc.、https://ddn.co.jp/products/ssdstorage/ime.html (参照 2017-09-27)

[5] 橋爪 信明、“バーストバッファ型 IO アクセラレーション製品 DDN IME のご紹介” DDN Inc.、https://ddn.co.jp/dcms_media/other/IME_14_vol21no4_shinseihin.pdf (参照 2017-09-27)

[6] “IOR HPC Benchmark”、Lawrence Livermore National Laboratory、https://sourceforge.net/projects/ior-sio/files/ (参照 2017-09-27)

[7] “NAS Parallel Benchmarks I/O Version 2.4”、NASA、https://www.nas.nasa.gov/assets/pdf/techreports/2003/nas-03-002.pdf (参照 2017-09-27)