

不揮発性メモリ 3D XPoint の AI/ビッグデータ処理への適用に向けた初期評価

佐藤 仁^{1,a)} 溝手 竜¹ 小川 宏高¹

概要: 不揮発性メモリ 3D XPoint の AI/ビッグデータ処理への適用に向けた初期評価として、現在市販されている 3D XPoint を不揮発性メモリとして用いた SSD である Intel Optane SSD DC P4800X に対して、ストレージ I/O(fio), ストレージ I/O の遅延隠蔽 (libaio), メモリバンド幅 (STREAM), 演算性能 (GEMM), ビッグデータ処理性能 (Graph500) など AI/ビッグデータ処理を模したワークロードを実行し、性能評価を行った。その結果、3D XPoint SSD をメモリの拡張として用いた場合に、ベアメタル環境での実行と比較して、計算インテンシブなワークロードである GEMM で 95.0%(単精度), 95.8%(倍精度), メモリインテンシブなワークロードである Graph500 で 78.3%(Scale30) 程度の性能で実行でき、DRAM メモリを超える規模のデータセットに対しても性能低下を抑えて透過的なメモリアクセスを提供できることを確認した。

1. はじめに

近年、AI やビッグデータ処理においても高性能計算の必要性が著しく高まっている。とりわけ、深層学習は、自動車の自動運転、製造業、ロボット、医療、創薬、金融など様々な分野への応用が期待されており、アルゴリズム (Algorithm Theory) の進展だけでなく、大量のビッグデータ (Big Data) を蓄えるストレージ技術や、それらのデータに対して高速に処理する計算能力 (Computation) など三位一体となった解決が求められている。

AI/ビッグデータ処理を高速に行うためには処理の対象となるデータセットをメモリ上に置くことが必要になるが、現状でも数万～数千万以上規模の映像・画像・音声・テキストデータであり、今後も爆発的に増大する傾向であるため、将来的にはメモリ上にデータを置くことができなくなり、ストレージへのデータをオフロードする必要があると予想される。特に、大規模グラフ処理や分散深層学習などでは、その処理の性質から、メモリやストレージへの大量のランダム I/O が発生する傾向があり、問題となると考えられる。

一方、不揮発性メモリなどのような DRAM メモリと比較するとレイテンシやスループットなど性能面で劣るものの、容量あたりの価格や消費電力の点で優れたデバイスが登場し始めている。このような不揮発性メモリを DRAM

メモリに対して補助的に利用することで、メモリを多階層化し、1 台の計算ノード上の DRAM メモリには収まらない規模のデータセットに対して AI/ビッグデータ処理を行うことができる可能性がある。例えば、Intel と Micron により開発が進められ、2015 年に一般発表された不揮発性メモリ技術である 3D XPoint [1] は、Intel の主張によると、従来型の NAND フラッシュメモリの 1000 倍高速であり、1000 倍書き換え寿命が長く、DRAM メモリの 10 倍の記憶密度を有すると主張しており、現状の NAND フラッシュメモリに代わり将来普及が見込める不揮発性メモリとして期待されている。しかし、3D XPoint メモリを AI/ビッグデータ処理に適用した際のその具体的手法やどの程度性能低下が起きるのかなどの定量的な指標は明らかではない。

我々は、不揮発性メモリ 3D XPoint の AI/ビッグデータ処理への適用に向けた初期評価として、現在市販されている 3D XPoint を不揮発性メモリとして用いた SSD である Intel Optane SSD DC P4800X に対して、ストレージ I/O(fio), ストレージ I/O の遅延隠蔽 (libaio), メモリバンド幅 (STREAM), 演算性能 (GEMM), ビッグデータ処理性能 (Graph500) など AI/ビッグデータ処理を模したワークロードを実行し、性能評価を行った。その結果、3D XPoint SSD をメモリの拡張として用いた場合に、ベアメタル環境での実行と比較して、計算インテンシブなワークロードである GEMM で 95.0%(単精度), 95.8%(倍精度), メモリインテンシブなワークロードである Graph500 で 78.3%(Scale30) 程度の性能で実行でき、DRAM メモリを

¹ 国立研究開発法人産業技術総合研究所

^{a)} hitoshi.sato@aist.go.jp

表 1 SSD のスペック

	Intel DC P4800X	Intel DC P3700
レイテンシ (Write)	10usec	20usec
IOPS (Write)	500000	175000
バンド幅 (Write)	2000MB/s	1900MB/s
レイテンシ (Read)	10usec	20usec
IOPS (Read)	550000	450000
バンド幅 (Read)	2400MB/s	2800MB/s
容量	375GB	2TB

超える規模のデータセットに対しても性能低下を抑えて透過的なメモリアクセスを提供できることを確認した。

2. 3D XPoint による不揮発性メモリ

2.1 3D XPoint

3D XPoint [1] は, Intel と Micron により開発が進められ, 2015 年に一般発表された不揮発性メモリの技術である。Intel によると, 従来型の NAND フラッシュメモリの 1000 倍高速であり, 1000 倍書き換え寿命が長く, DRAM メモリの 10 倍の記憶密度を有すると主張しており, 将来の AI/ビッグデータ処理の計算基盤を支える重要な技術の一つである。

3D XPoint の技術詳細は公式には明らかにはなっていないが, Flash Memory Summit 2017 で発表された Techinsights 社の解析 [2] によると, 3D XPoint は相変化メモリ (PCM: Phase Change Memory) の一種で, メモリの記憶素子とセルの選択 (セレクト) スイッチにカルコゲナイド材料を用い, 動作原理には, カルコゲナイド材料が電気的性質と光学的性質の異なる状態を可逆的に変化することができるという性質を利用したオブシンスキー効果に基づいている。また, ワード線とビット線が交差した微小な領域にメモリセル全体を収めるクロスポイント構造を構成し積層することで, 高密度なメモリセルアレイを構成する。3D XPoint のメモリセルアレイは 2 層構造となっている。

2.2 Intel Optane SSD

Intel Optane SSD [3] は, 3D XPoint を不揮発性メモリとして用いた SSD の一実装である。2018 年 1 月現在で市場に出ており, 今回のベンチマークの対象とする 3D XPoint SSD である Intel DC P4800X のスペックを表 1 に示す。比較対象として, 従来型の NAND SSD である Intel DC P3700 のスペックも記載する。NAND SSD と比較して, 3D XPoint SSD は遅延が 10usec と著しく小さいことが特徴となっている。

2.3 IMDT (Intel Memory Drive Technology)

IMDT (Intel Memory Drive Technology) [4] は, ソフトウェアにより 3D XPoint SSD (Intel Optane SSD) を DRAM メモリのサブシステムとして透過的に拡張するこ

表 2 計算機のスペック

CPU	Intel Xeon E5-2699 v4 (2.2GHz, 22core) × 2 sockets
Mem	256 GiB (DDR4-2400)
SSD	Intel DC P4800X 375GB × 2 Intel DC P3700 2TB × 2

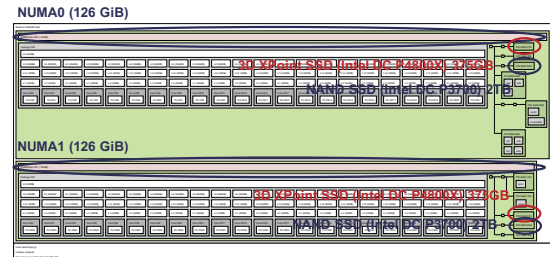


図 1 計算機の内部構成 (ベアメタル環境)

とを可能にする Software-Defined Memory である。IMDT のソフトウェアとしての実体は, SMP システムを構築するための仮想化ソフトウェアである ScaleMP の vSMP Foundation [5] を Intel Optane SSD 向けに最適化したものとなっている [6]。メモリの構成方法としては, 3D XPoint SSD を 1 つの独立した NUMA ノードとしてメモリを構成する Unify モードと, 3D XPoint SSD を DRAM メモリの拡張として透過的に構成する Expand モードの 2 通りがある。

3. IMDT を用いた計算機の構成

IMDT により 3D XPoint SSD をメインメモリの拡張として用いた際の計算機の構成を, ベンチマークの計測環境を対象として説明する。

3.1 ベアメタル環境

表 2 に計算機のスペック, 図 1 に hwloc [7] の lstopo コマンドで取得した計算機の内部構成を示す。1 つの NUMA ノードにつき CPU が 1 ソケット, 128GiB のメインメモリ (実際には 126GiB に見えている), SSD が属している。SSD には, 3D XPoint SSD である Intel Optane SSD DC P4800X 375GB 1 基と, 比較対象として従来型の NAND SSD である Intel DC P3700 2TB 1 基が属している。

3.2 IMDT により構成された環境

3.2.1 Unify モード

図 2 に hwloc の lstopo コマンドで取得した IMDT の Unify モードで構成した際の計算機の内部構成を示す。IMDT の Unify モードでは, 既存の DRAM メモリからなる NUMA ノードの他に, 3D XPoint SSD からなる NUMA ノードを構成する。図 2 では, ベアメタル環境でみえる 2 つの NUMA ノード (NUMA0, NUMA1) に加えて, 447GiB の 3D XPoint SSD からなる独立した NUMA ノー

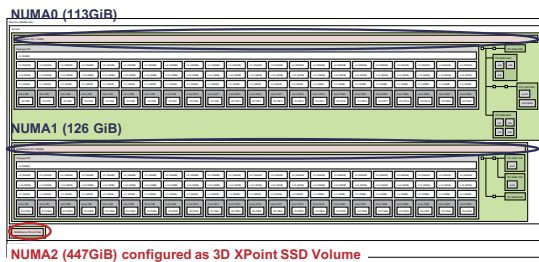


図2 計算機の内部構成 (IMDT Unify モード)

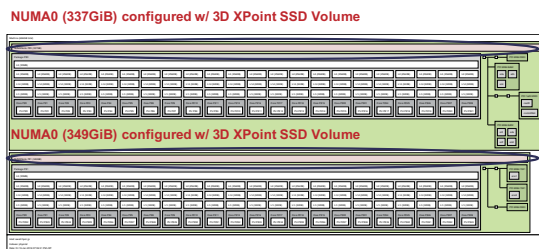


図3 計算機の内部構成 (IMDT Expand モード)

ド (NUMA2) としてみえる。IMDT のソフトウェア動作やキャッシュのために、DRAM メモリからなる NUMA0 の NUMA ノードは 113GiB とベアメタル環境と比較して少ないメモリ容量として構成される。また、3D XPoint SSD からなる NUMA2 の NUMA ノードも、同様に、物理容量 (375GB $\times$ 2 = 750GB) と比較して少ない容量として構成される。

3.2.2 Expand モード

図3に hwloc の lstopo コマンドで取得した IMDT の Expand モードで構成した際の計算機の内部構成を示す。IMDT の Expand モードでは、既存の DRAM メモリからなる NUMA ノードを拡張し、3D XPoint SSD のボリュームも利用可能なように構成する。図3では、DRAM メモリと 3D XPoint SSD から構成される 337GiB と 349GiB の NUMA ノードが2つみえる。IMDT のソフトウェア動作やキャッシュのために、ベアメタル環境と比較して少ない物理容量として構成される。

4. 実験

4.1 実験内容

不揮発性メモリ 3D XPoint の AI/ビッグデータ処理への適用に向けた初期評価として、現在市販されている 3D XPoint を不揮発性メモリとして用いた SSD である Intel DC P4800X に対して、AI/ビッグデータ処理を模したワークロードを実行し、予備的な性能評価を行った。具体的には、まず、Intel DC P4800X をストレージとして用いた際の性能を理解するために、ストレージ I/O (fio)、ストレージ I/O の遅延隠蔽 (libaio) のベンチマークを実行した。次に、Intel DC P4800X をメモリとして用いた際の性能を理解するために、メモリバンド幅 (STREAM) や演算

性能 (GEMM)、ビッグデータ処理性能 (Graph500) のベンチマークを実行した。

4.2 実験環境

実験は全て3節に記述した1台の計算機上で行った。計算機の設定として、C-State, CPU Power Saving Mode, IRQ Balance Service, Udev Service を無効とし、CPU Governor を Performance とした。計算ノードの OS は CentOS 7.3 (1611) で構成され、Linux のカーネルはセルフビルドした 4.14.12 である。ただし、PTI (Page Table Isolation) を無効にして起動している。各種ベンチマークプログラムのビルドには gcc 4.8.5 を用いた。また、IMDT は 8.5.1955.1 を用いた。

4.3 ストレージとしての評価

4.3.1 ストレージ I/O (fio)

3D XPoint SSD のストレージ I/O の性能を計測するために Flexible I/O Tester (fio) ベンチマーク [8] を実行した。実験の設定として、ファイルシステムを構成しないもの (nofs) と xfs ファイルシステムを構成したもの (xfs) とを 3D XPoint SSD である Intel DC P4800X と従来型の NAND SSD である Intel DC P3700 とで比較した。fio は v3.3 を用いた。

まず、4KiB 単位でランダムに読み込み・書き込み I/O (rand-r, rand-w) を行った際の平均レイテンシを計測した結果を図4に示す。ここで、Queue Depth は 1 と 16 として実行している。結果より、Intel DC P4800X の方が Intel DC P3700 と比較して概ね低い平均レイテンシを示していることが伺える。とりわけ、Queue Depth を 1 と小さく設定した際の読み込み I/O の平均レイテンシを 10usec 程度にできることを確認した。このことから、3D XPoint SSD は深層学習の I/O で多くみられるような画像やテキストなど小さいサイズのファイルへの大量のアクセスなどのシナリオにおいて非常に有効であることが伺える。一方で、書き込み I/O の平均レイテンシは Intel DC P4800X と Intel DC P3700 とでどちらも同程度を示すことを確認した。これは、書き込みの際のキャッシュなどの影響があったためであると考えている。

次に、4KiB 単位でランダムに読み込み・書き込み I/O を行った際の IOPS を計測した結果を図5に示す。ここでは、全て 4KiB 単位でのランダムな読み込み・書き込み I/O (rand-r, rand-w) を行うものの他、4KiB 単位でのランダムな読み込み I/O と書き込み I/O を 70% と 30% の比率で混合させた I/O (rand-rw) を行った。また、Queue Depth は 1 と 16 として実行している。結果より、Intel DC P4800X のほうが Intel DC P3700 と比較して概ね高い IOPS を示していることが伺える。とりわけ、Queue Depth を大きく

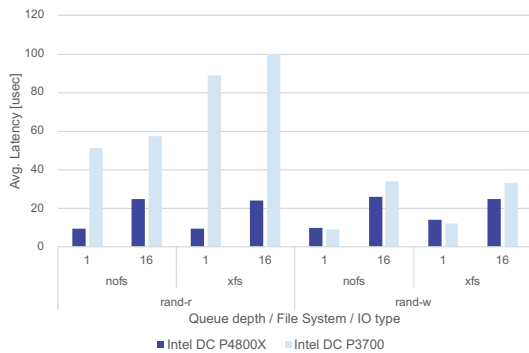


図 4 4KiB 単位でのランダム読み込み・書き込み I/O の平均レイテンシ

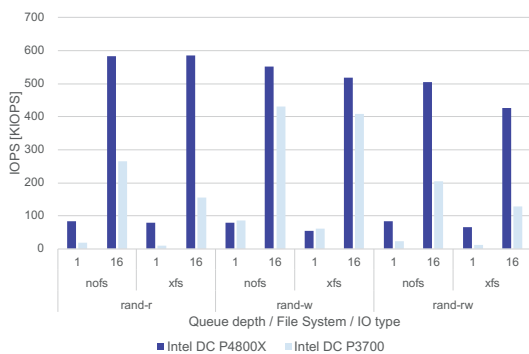


図 5 4KiB 単位でのランダム読み込み・書き込み I/O の IOPS

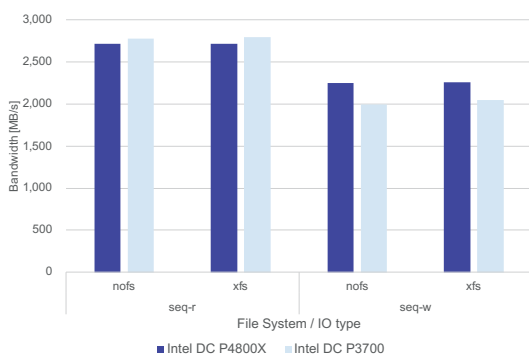


図 6 128KiB 単位でのシーケンシャルな読み込み・書き込み I/O のバンド幅

設定したときにおいて顕著な性能差を示した。

最後に、128KiB 単位でシーケンシャルに読み込み・書き込み I/O(seq-r, seq-w) を行った際のバンド幅を計測した結果を図 6 に示す。この場合は、Intel DC P4800X と Intel DC P3700 とでほぼ同等の性能を示した。

上記の結果から、現状の 3D XPoint SSD は従来の NAND SSD と比較して、ランダムな読み込み・書き込み I/O において非常に良好な性能を示すことを確認した。とりわけ、読み込み I/O の平均レイテンシが 10µsec 程度であり、非常に小さく抑えられる。一方で、バンド幅は従来の NAND SSD とほぼ同等の性能を示すため、シーケンシャル I/O を多く行うアプリケーションでは 3D XPoint SSD の性能特性を活かしきれないことが伺える。

4.3.2 ストレージ I/O の遅延隠蔽 (libaio)

非同期 I/O を用いることによりストレージに対する I/O の遅延隠蔽ができ、アプリケーションの演算性能を向上させることができる場合がある [9]。ここでは、3D XPoint SSD に対して Linux Asynchronous I/O(libaio) で非同期 I/O を行った際の性能を計測し、比較対象として同期 I/O(psync) を行った場合の性能を計測した。計測には我々が開発しているベンチマークソフトウェアである aiortest [10] を用い、ファイルシステムを構成しないもの (nfs) と xfs ファイルシステムを構成したもの (xfs) を Intel DC P4800X と Intel DC P3700 とで比較した。ワークロードとしては、16GiB の単一ファイルに対して 1GiB 単位で 4 events, 4 ファイルディスクリプタで読み込み・書き込み I/O をするシナリオと、240GiB の単一ファイルに対して 1GiB, 60 events, 4 ファイルディスクリプタで読み込み・書き込み I/O をするシナリオを設定し、各々、Buffered I/O (buffered) と Direct I/O (direct) を行った。

図 7 に libaio による非同期 I/O の性能の結果を示す。図 7(a), 図 7(c) より、読み込みの非同期 I/O、とくに Buffered I/O において、Intel DC P4800X の方が Intel DC P3700 と比較して良好な性能を示している。一方で、書き込みの非同期 I/O の場合、図 7(b), 図 7(d) より、Intel DC P4800X と Intel DC P3700 でほぼ同等の性能を示している。また、非同期 I/O と同期 I/O の比較では、図 7 より、Intel DC P4800X では概ね同期 I/O でも十分な性能が達成されていることが伺える。これは、3D XPoint SSD がランダム I/O に非常に特化しているため、一般的な同期 I/O を用いた場合でも十分に高速な I/O が達成されるためであると考えられる。

4.4 メモリとしての評価

4.4.1 メモリバンド幅 (STREAM)

IMDT により 3D XPoint SSD を透過的なメモリとして構成した際のメモリバンド幅を計測するために STREAM ベンチマークを実行した。実行にはバージニア大学のサイトより入手した v5.10 のコード [11] を mmap, mbind を利用するように改変して用いた。実験としては、IMDT によりメモリを Unify モードで構成した場合と Expand モードで構成した場合の 2 通りを行った。

図 8 に IMDT によりメモリを Unify モードで構成した際の実験の概要を示す。実験のシナリオとしては、NUMA0 の NUMA ノードに属する CPU コアから同じ NUMA ノード (NUMA0) に属する DRAM メモリへアクセスする場合 (Local NUMA), NUMA0 の NUMA ノードに属する CPU コアから遠隔の NUMA ノード (NUMA1) の DRAM メモリへアクセスする場合 (Remote NUMA), NUMA0 の NUMA ノードに属する CPU コアから遠隔の NUMA

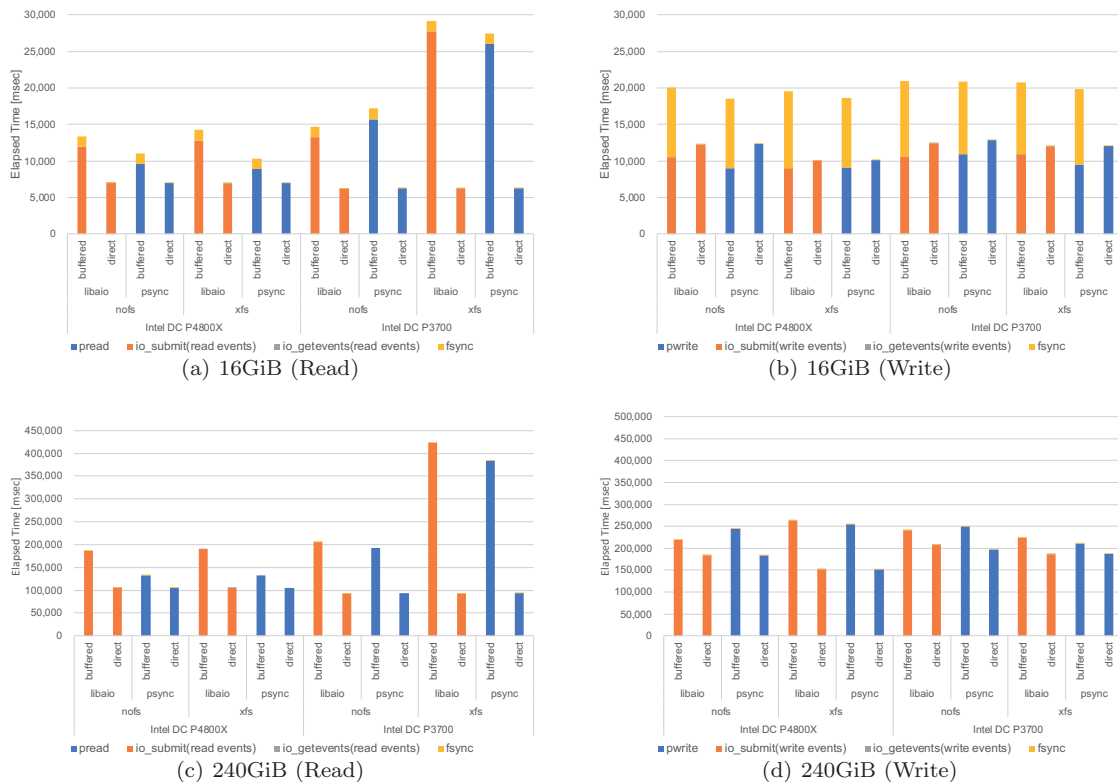


図 7 libaio による非同期 I/O の性能

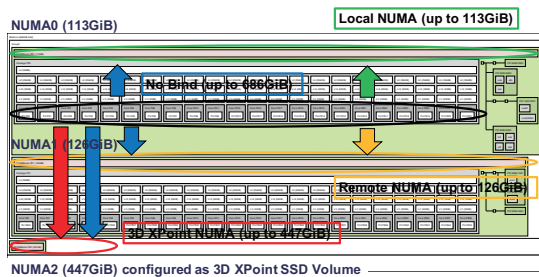


図 8 IMDT Unify モードでの STREAM の概要

ノード (NUMA2) の 3D XPoint メモリへアクセスする場合 (3D XPoint NUMA), NUMA ノードを考慮しない場合 (No Bind) の 4 通りのメモリアクセスを設定し、比較対象として IMDT を無効にしベアメタル環境で実行した (Baremetal). Local NUMA では 113GiB (ベアメタル環境の場合は 126GiB), Remote NUMA では 126GiB, 3D XPoint NUMA では 447GiB, No Bind では 686GiB (ベアメタル環境の場合は 252GiB) のメモリ容量が利用できる。

IMDT Unify モードでの STREAM ベンチマークの結果を図 9 に示す。IMDT の有無に関わらず、Local NUMA に対しては 47GiB/s 程度、Remote NUMA に対しては 28GiB/s 程度の性能を達成した。また、3D XPoint NUMA や No Bind などデータサイズを 384GiB や 576GiB など大きくし 3D XPoint SSD のボリュームへのアクセスが発生した場合、デバイスのバンド幅に律速され、4GiB/s 程度の性能であることを確認した。一方で、興味深いことに、

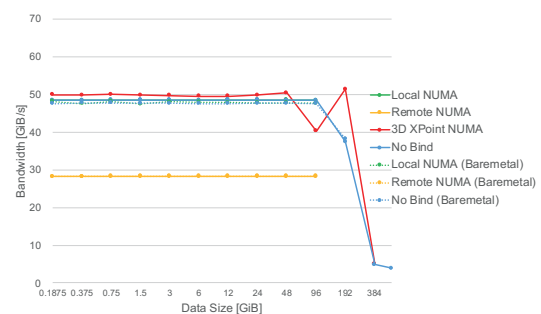


図 9 IMDT Unify モードでの STREAM の性能

3D XPoint NUMA へのアクセスでデータサイズが小さい場合は、3D XPoint SSD のバンド幅に律速されず、Local NUMA 相当の性能を達成することを確認した。これは、IMDT のソフトウェアにより何らかの 3D XPoint SSD から DRAM メモリへのキャッシュやプリフェッチなどが行われているためであると考えている。

次に、IMDT によりメモリを Expand モードで構成した際の実験の概要を図 10 に示す。実験のシナリオとして、NUMA0 の NUMA ノードに属する CPU コアから同じ NUMA ノード (NUMA0) に属するメモリへアクセスする場合 (Local NUMA), NUMA0 の NUMA ノードに属する CPU コアから遠隔の NUMA ノード (NUMA1) のメモリへアクセスする場合 (Remote NUMA), NUMA ノードを考慮しない場合 (No Bind) の 3 通りのメモリアクセスを

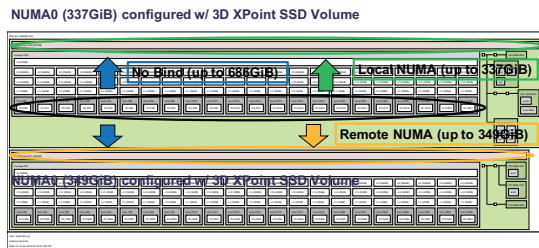


図 10 IMDT Expand モードでの STREAM の概要

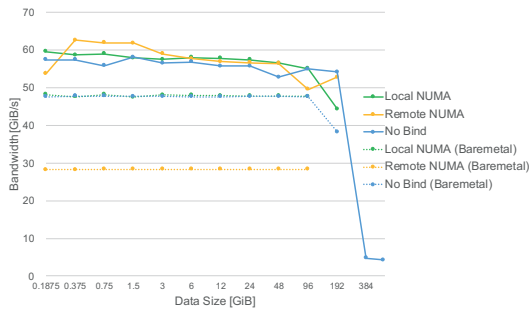


図 11 IMDT Expand モードでの STREAM の性能

設定し、比較対象として IMDT を無効にしベアメタル環境で実行した (Baremetal). IMDT Expand モードの場合は、Local NUMA では 337GiB, Remote NUMA では 349GiB, No Bind では 686GiB のメモリ容量が利用でき、ベアメタル環境の場合は、Local NUMA, Remote NUMA ともに 126GiB 程度、No Bind では 252GiB のメモリ容量が利用できる。

IMDT Expand モードでの STREAM ベンチマークの結果を図 11 に示す。IMDT が無効な場合、Local NUMA に対しては 47GiB/s 程度、Remote NUMA に対して 28GiB/s 程度の性能を達成するが、IMDT を有効にした場合、これらの性能よりも良好な性能を示すことを確認した。これは、IMDT のソフトウェアにより何らかの 3D XPoint SSD から DRAM メモリへのキャッシュやプリフェッチなどが行われているためであると考えている。一方で、DRAM メモリの容量を超えるサイズを実行した場合、3D XPoint SSD のボリュームへのアクセスが発生し、IMDT Unify モードの場合と同様にデバイスのバンド幅に律速され、4GiB/s 程度の性能であることを確認した。

4.4.2 演算性能 (GEMM)

DRAM メモリを超える規模のデータサイズに対する演算性能を計測するために BLAS(Basic Linear Algebra Subprograms) ライブラリの GEMM(General Matrix Multiplication) カーネルを CPU 上で実行した。メモリを IMDT で Expand モードとして構成し、Intel Math Kernel Library 2018 Update 1(2018.1.163) の CBLAS [12] を用いて、単精度 (FP32) と倍精度 (FP64) を DRAM メモリを超える規模のサイズとなるように計算機上で実行した。実験のシナ

リオとしては、データサイズを変えながら 1CPU(1CPU) 及び 2CPU(2CPUs) で GEMM を実行し、比較対象として IMDT を無効にしベアメタル環境で実行した (Baremetal). 図 10 と同様に、単体の NUMA ノードへは 126GiB 程度、DRAM 上のメモリへは 239GiB, 3D XPoint SSD により拡張したメモリ全体へは 686GiB のメモリ容量が利用できる。

IMDT Expand モードでの GEMM ベンチマークの実行結果を図 12 に示す。図 12(a) に単精度の GEMM を実行した結果、図 12(b) に倍精度の GEMM を実行した結果を示す。図中、演算の対象となるデータサイズを Size とし、その際のメモリ消費量を Memory Consumption として示している。IMDT Expand モードで実行した場合、遠隔の NUMA ノードへのアクセスや 3D XPoint SSD のボリュームへのアクセスが発生した場合においても良好な性能を維持することを確認した。例えば、IMDT Expand モードで実行可能な最大サイズである $n = 244736$ (単精度), 173056(倍精度) のときで 2733.51GFLOPS(単精度), 1274.82GFLOPS(倍精度) であり、DRAM メモリ内で実行可能な最大サイズである $n = 122368$ (単精度), 86528(倍精度) と比較しても 95.0%(単精度), 95.8%(倍精度) 程度と遜色のない性能を示した。

これらの結果から、計算インテンシブなワークロードの場合、3D XPoint SSD を IMDT Expand モードにより透過的なメモリとして構成しても DRAM メモリと遜色のない性能が達成でき、3D XPoint SSD のランダム I/O に特化した低レイテンシで高 IOPS な性能特性や IMDT ソフトウェアのキャッシュやプリフェッチが良好に機能することが伺える。

4.4.3 ビッグデータ処理性能 (Graph500)

ビッグデータ性能を計測するベンチマークである Graph500 [13] を対象に、NUMA に最適化したアルゴリズム [14] を採用した実装 (In-core) と、NUMA に最適化し Out-of-core 処理に最適化したアルゴリズム [15] を採用した実装 (Out-of-core) を用いて性能評価を行った。実装としては我々が開発している Graph500 実装である NETALX [16] を用いた。In-core 実装では 3D XPoint SSD である Intel DC P4800X を IMDT Expand モードでメモリとして構成してベンチマークを実行し、Out-of-core 実装では Intel DC P4800X と従来型の NAND SSD である Intel DC P3700 をストレージとして xfs ファイルシステムを構成してベンチマークを実行した。また、比較対象として、IMDT を無効にしベアメタル環境でも NUMA に最適化された In-core 実装を実行した (Baremetal). 実行の対象とするグラフは、Graph500 ベンチマークの仕様と同様に、頂点が 2^{SCALE} , 平均次数が 16 である Kronecker Graph とし、各 Graph500 実装でのパラメータはデフォルトのものを用了。具体的には、Hybrid BFS の切り替えのパラメータを $\alpha = 64, \beta = 4$ とし、Out-of-core 実装での DRAM メモリへの辺のキャッ

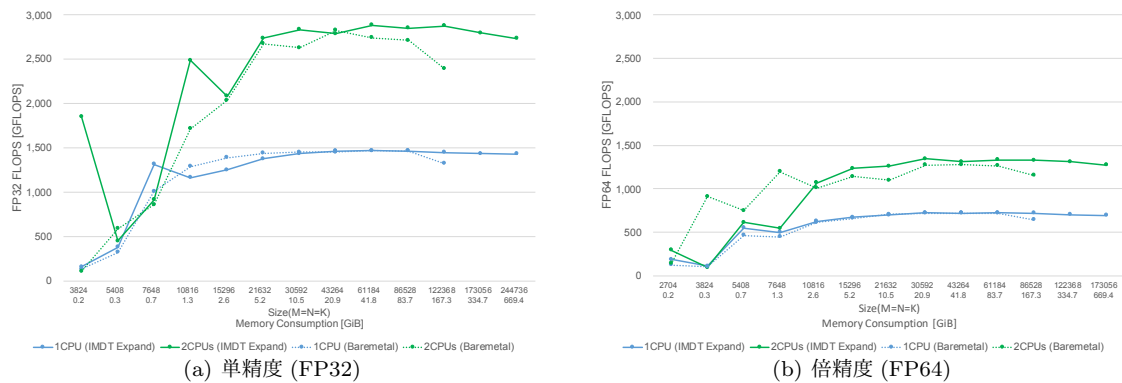


図 12 GEMM の性能

シュサイズを on-mem-edges = 10 とした。

図 13(a) に 3D XPoint SSD をメモリとして構成し In-core 実装を実行した結果、図 13(b) に 3D XPoint SSD をストレージとして構成し Out-of-core 実装を実行した結果を示す。Scale 29 が DRAM メモリのみを用いた最大の実行サイズであり、Scale 30, Scale 31 が SSD のボリュームを必要とする実行サイズとなる。3D XPoint SSD をメモリとして構成し In-core 実装を実行した場合、Scale 30 の実行で 7.52GTEPS と良好な性能を示す。一方で、ストレージとして構成し Out-of-core 実装を実行した場合、Scale30 の実行で 4.41GTEPS となった。また、ベアメタル環境での Scale 29 での実行と比較すると、メモリとして構成した場合は 78.3%、ストレージとして構成した場合は 45.8%程度の性能だった。ただし、3D XPoint SSD をストレージとして構成した場合は、メモリとして構成した場合と比較して、さらに 2 倍程度大きくした Scale 31 のグラフも実行可能であることを確認した。これは、Out-of-core 実装において DRAM メモリへの辺のキャッシュサイズを制御できるためであり、このパラメータを変化させることにより更なる性能向上を示す可能性はある。

これらの結果から、メモリインテンシブなワークロードの場合においても、IMDT Expand モードにより比較的性能低下を抑えて 3D XPoint SSD を DRAM メモリの拡張として透過的に利用できることが伺える。

4.5 議論

4.3 節の実験により、3D XPoint SSD をストレージとして用いた場合、ランダム I/O が頻出するワークロードに対して非常に有効に働くことが伺える。従来の HPC ではランダム I/O はシーケンシャル I/O へ変換しバーストアクセスを行うことで性能向上を達成する場合が多い。しかし、このような利用用途では 3D XPoint SSD の特性は活かしきれず、バンド幅性能は従来型の NAND SSD とあまり変わらない。大規模グラフ処理や分散深層学習など AI/ビッグデータ処理では、その処理の性質から、メモリやス

トレージへの大量のランダム I/O が発生する傾向があり、3D XPoint SSD は特別な最適化を行わなくても性能が達成できることが期待できる。

また、4.4 節の実験により、3D XPoint SSD をメモリとして用いた場合においても、ランダム I/O に特化した高 IOPS で低レイテンシな性能特性や、IMDT ソフトウェアのキャッシュやプリフェッチ機能により、計算インテンシブやメモリインテンシブなワークロードにおいても、DRAM メモリを超える規模のデータセットに対しても性能低下を抑えて透過的なメモリアクセスを提供できることが伺える。

5. 関連研究

不揮発性メモリの利用に関してはこれまで様々な議論が行われてきた [17], [18]。とりわけ、不揮発性メモリに対して、Memory-mapped I/O を行う方法やファイルシステムを構成する方法は、既存のアプリケーションに対する修正を最小限にすることができるため有力であると考えられている。しかし、従来の NAND フラッシュメモリでは、DRAM と比較して性能差が非常に大きかったため、現実的に DRAM メモリの代替として利用することは難しく、非同期 I/O による遅延隠蔽などのソフトウェア実装による最適化が必要であった [9]。SDSC の Dash [19] や Gordon [20] などのビッグデータ処理に特化したスーパーコンピュータでは、ソフトウェアにより NAND フラッシュメモリを集約して共有メモリマシンのように利用する試みが行われてきた。IMDT による 3D XPoint SSD の利用はこのようなアプローチに類似する。

6. おわりに

我々は、不揮発性メモリ 3D XPoint の AI/ビッグデータ処理への適用に向けた初期評価として、Intel Optane SSD DC P4800X に対して AI/ビッグデータ処理を模したワークロードを実行し性能評価を行った。その結果、3D XPoint SSD のランダム I/O に特化した低レイテンシで高 IOPS である性能特性や、IMDT ソフトウェアのキャッシュやプリ

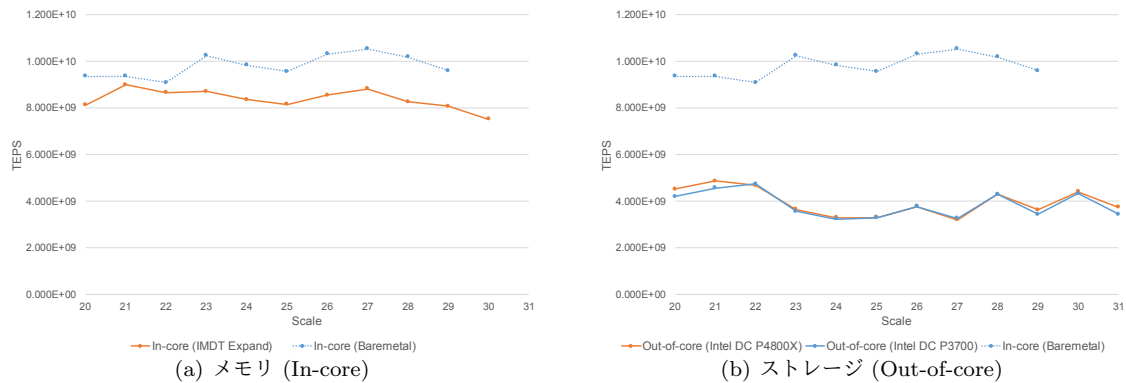


図 13 Graph500 の性能

フェッチ機能により、DRAM メモリを超える規模のデータセットに対しても性能低下を抑えて透過的なメモリアクセスを提供できることを確認した。

今後の課題として、大規模グラフ処理や分散深層学習など、メモリやストレージへの大量のランダム I/O が発生する実アプリケーションを用いた 3D XPoint SSD の検証を目指したい。

謝辞 この研究の一部は、JSPS 科研費 26540050、及び、NEDO 次世代人工知能・ロボット中核技術開発の一環で実施した。

参考文献

- [1] Intel: Intel and Micron Produce Breakthrough Memory Technology, <https://newsroom.intel.com/news-releases/intel-and-micron-produce-breakthrough-memory-technology/> (2015).
- [2] Choe, J.: XPoint Memory Comparison Process & Architecture, Flash Memory Summit Persistent Memory Forum R-12 3D XPoint: Current Implementations and Future Trends (2017).
- [3] Intel: Intel Optane Memory, <https://www.intel.com/content/www/us/en/architecture-and-technology/optane-memory.html> (2017).
- [4] Intel: Intel Memory Drive Technology (IMDT), <https://www.intel.co.jp/content/www/jp/ja/software/intel-memory-drive-technology.html> (2017).
- [5] ScaleMP: vSMP Foundation, <http://www.scalemp.com/products/vsmp-foundation/> (2017).
- [6] ScaleMP: ScaleMP's Software-Defined Memory for NVMe and Intel Optane Technology to be Presented at Intel Developer Forum, <http://www.scalemp.com/media-hub-item/scalemp-software-defined-memory-for-nvme-and-intel-optane-technology-presented-at-intel-developer-forum-2016/> (2016).
- [7] OpenMPI: Portable Hardware Locality (hwloc), <https://www.open-mpi.org/projects/hwloc/> (2018).
- [8] Axboe, J.: Flexible I/O Tester (FIO), [git://git.kernel.dk/fio.git](https://github.com/axboe/fio).
- [9] Sato, H., Mizote, R., Matsuoka, S. and Ogawa, H.: I/O chunking and latency hiding approach for out-of-core sorting acceleration using GPU and flash NVM, *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 398–403 (online), DOI: 10.1109/BigData.2016.7840629 (2016).
- [10] Sato, H.: aiotest, <https://github.com/htsst/aiotest> (2017).
- [11] McCalpin, J. D.: STREAM: Sustainable Memory Bandwidth in High Performance Computers, <https://www.cs.virginia.edu/stream/>.
- [12] Intel: Intel Math Kernel Library, <https://software.intel.com/en-us/mkl>.
- [13] Graph500: The Graph500 List, <http://www.graph500.org>.
- [14] Yasui, Y., Fujisawa, K. and Goto, K.: NUMA-optimized Parallel Breadth-first Search on Multicore Single-node System, *2013 IEEE International Conference on Big Data (IEEE BigData 2013)*, IEEE, pp. 394–402 (online), DOI: 10.1109/BigData.2013.6691600 (2013).
- [15] Iwabuchi, K., Sato, H., Yasui, Y., Fujisawa, K. and Matsuoka, S.: NVM-based Hybrid BFS with Memory Efficient Data Structure, *2014 IEEE International Conference on Big Data (IEEE BigData 2014)*, IEEE, pp. 529–538 (online), DOI: 10.1109/BigData.2014.7004270 (2014).
- [16] Sato, H.: NETALX, <https://github.com/htsst/netalx>.
- [17] Mittal, S. and Vetter, J. S.: A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 9219, No. c, pp. 1–14 (online), DOI: 10.1109/TPDS.2015.2442980 (2015).
- [18] Rudoff, A.: Programming Models for Emerging Non-Volatile Memory Technologies, *login:*, No. June 2013, Volume 38, Number 3, pp. 40–45 (2013).
- [19] He, J., Jagatheesan, A., Gupta, S., Bennett, J. and Snively, A.: DASH: a Recipe for a Flash-based Data Intensive Supercomputer, *2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, pp. 1–11 (online), DOI: 10.1109/SC.2010.16 (2010).
- [20] Strande, S. M., Cicotti, P., Sinkovits, R. S., Young, W. S., Wagner, R., Tatineni, M., Hocks, E., Snively, A. and Norman, M.: Gordon: Design, Performance, and Experiences Deploying and Supporting a Data Intensive Supercomputer, *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment on Bridging from the eXtreme to the campus and beyond - XSEDE '12*, New York, New York, USA, ACM Press, p. 1 (online), DOI: 10.1145/2335755.2335789 (2012).