Optane SSD を使用したメモリ拡張技術の性能改善手法

風間 哲1　児玉 宏喜 1　桑村 慎哉 1　山中 英樹2　吉田 英司 1

1. はじめに

インメモリ処理のデータ容量増加の要求に対して，DRAM コストが高止まり傾向にあるため低コストなシステム設計が難しくなっている．そのため，ブロックデバイスである SSD を利用し，アプリケーションを SSD に最適化することにより DRAM 使用量を減らす技術が提案されている[1]．しかし，ブロックデバイスである SSD をメモリとしてアクセスできる仕組みを使えばアプリケーションへの適用はより容易になる．

Linux のスワップはブロックデバイスをメモリとして利用する最も簡単な手法であるが，HDD 時代に設計され性能に課題がある．そこで，近年の高速デバイスへの最適化や，アプリに応じて遅いメモリ領域を効果的に利用する Hybrid memory 技術が提案されている [2].

近年は DIMM に搭載できる大容量不揮発性メモリ製品 Intel Optane DC Persistent Memory Module(DCPMM)が登場し，メモリ拡張用途として利用できるようになってきた．しかし，現時点では物理的なサイズや電力制約の少ない SSD 型(Optane SSD)の方が容量やコストの面ではメリットが大きい．

今回，我々はフラッシュメモリより高速な不揮発メモリを利用した Optane SSD を Hybrid memory に適用し，その性能評価を実施した．その結果，デバイスの高速化により Hybrid memory の制御ソフトウェアのオーバーヘッドが相対的に顕在化し，Optane SSD の性能が引き出せていないことが判明した．本論文では，制御ソフトウェアのオーバーヘッドの原因，対策手法および対策による性能改善効果について述べる．

2. メモリ拡張制御ソフトウェアのオーバーヘッド

Optane SSD を適用した Hybrid memory の性能を把握するため，マイクロベンチマークを使用し Hybrid memory とスワップのリード/ライト性能を測定した．図 1 にその結果を示す．リードスループット性能に関しては Hybrid memory ではスワップより高スループットであり，Optane SSD のハードウェア性能と同等性能が出ていることが確認できた．しかし，ライトスループットに関してはスワップより低く，Optane SSD のハードスループット性能に届かない結果となった．つまりライトに関しては Hybrid memory のソフトウェアオーバーヘッドにより性能が頭打ちになっていると考えられる．このオーバーヘッドの解析の結果，メモリライト時の Read Modify Write とメモリロック競合が原因であることが分かったため，これらのオーバーヘッド削減を行った．

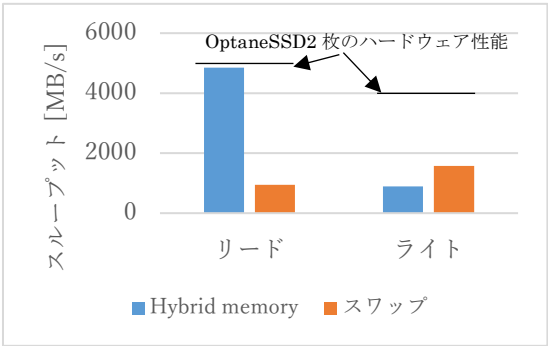

図 1 Optane SSD(750GBx2)を使用したメモリ拡張スループット(DRAM:16GB,データ 40GB)

1 (株)富士通研究所 Fujitsu Laboratories Ltd.
2 (株)富士通コンピュータテクノロジーズ Fujitsu Computer Technologies Ltd.

2.1 Read Modify Write(RMW)

Hybrid memory は，スワップをベースにしているため，ライト時にページキャッシュに対応するページがなければ Optane SSD からページをページキャッシュにリードし，ページ中の対応するデータのみライトする RMW を実行する．しかし，ページキャッシュの管理単位である 4KB のライト時には 4KB のデータすべてが書き換えられるため，RMW によるページのリードは不要である．そこで，4KB ライトの場合には RMW 動作を行なわない write システムコールに自動的に切り替えることにより，リードを行うことなくページキャッシュにライトすることを実現した．これにより，Optane SSD からページキャッシュへのリード回数を削減させた．

2.2 バックグラウンド追い出しスレッドのロック競合

ライト時のページキャッシュから Optane SSD へのデータ追い出しはバックグラウンドスレッドにより実行される．バックグラウンドスレッドはノード毎にページキャッシュのリスト管理を行っている．複数のスレッドがライトアクセスをする際，リスト管理のロック競合が起こり，それが性能低下を引き起こしていた．そこで，ノード毎の管理単位をノードの構成要素であるコア毎の管理に細分化することにより，ロック競合が生じる確率を減少させた．

3. 評価

3.1 評価環境

表 1 に性能測定に使用したハードウェアおよびソフトウェアの構成を示す．性能評価は，40GB のメモリ領域を Hybrid memory または malloc(スワップ評価時)で確保し，全領域に対するマルチスレッドでの 4KB シーケンシャルアクセススループットを計測するマイクロベンチマークを作成して行った．

3.2 評価結果

図 2 に RMW 対策およびロック競合対策後のライト性能改善効果を示す．対策により Optane SSD のハードウェア性能と同等レベルに改善した．

表 1 評価環境

サーバ	Fujitsu PRIMERGY RX2540M4
CPU	Intel Xeon Gold 6148 2.4GHz x2
DRAM	16GB DDR4-2666
Optane SSD	Intel Optane SSD DC P4800X 750GB 2 枚
OS	Linux kernel4.14.84

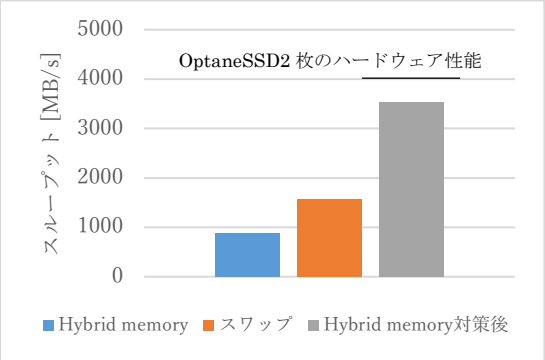

図 2 RMW およびロック競合対策による改善効果 (4KB ライト,DRAM:16GB,データ 40GB)

4. まとめ

メモリ拡張技術 Hybrid-memory のソフトウェアオーバーヘッドによる性能向上を阻止する要因を解析した．その要因として，①RMW による Optane SSD から DRAM へのリード回数増加と②ロック状態の多発であることを明らかにした．さらに，それらの要因に対する対策を講じ，その改善効果を示した．今後，本改善手法をインメモリ処理に適用し，アプリとしての評価を実施する予定である．

参 考 文 献

[1] A.Eisenman, et al., “Reducing DRAM Footprint with NVM in Facebook,” in Proc. of the thirteenth EuroSys Conference Article No.42, Apr. 2018

[2] B. Höppner, et al., “An Approach for Hybrid-Memory Scaling Columnar In-Memory Databases,” in Proc. Int. Workshop Accelerating Data Mana. Syst. Using Modern Processor Storage Archit., pp. 64–73, Sept. 2014.