様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ５ 月 ２８ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００１０４

研究課題名（和文）NVDIMM を用いたレジリエントなストレージの実現

研究課題名（英文）Implementing NVDIMM-based resilient storage

研究代表者
河野　健二（Kono, Kenji）

慶應義塾大学・理工学部（矢上）・教授

研究者番号：９０３０１１１８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：ミッションクリティカルなサービスに限らず，あらゆるサービスに対し高い信頼性と可用性が求められるようになっている．本研究の目的は，不揮発性 DIMM (NVDIMM) という比較的新しいメモリ・デバイスを活用したレジリエント (resilient) なストレージ基盤を実現することである．ジャーナリング領域を NVDIMM 上で管理することによって，旧来は性能上の観点から実用的ではなかったデータジャーナリングのオーバーヘッドを低減し，さらにファイルの一貫性保持のために頻繁に同期的書き込みを行なっても高い性能を得ることが可能となった．特に，高いデータ保全性が求められるワークロードで効果が大きい．

研究成果の学術的意義や社会的意義
いわゆる情報システムはすでに社会的インフラストラクチャーとなっており，信頼性の高いサービスを安価に提供することが求められている．本研究の社会的意義は，安価にデータの保全性を高める手法を確立した点にある．将来，クラウド・ストレージなどの信頼性向上およびより安価な実現手法として実用化されることを期待している．また，NVDIMM という比較的新しいデバイスを利用した点に特徴がある．NVDIMM はその利用法が活発に研究されており，本研究の内容は学術的にみても重要なものとなっている．

研究成果の概要（英文）：High reliability and availability are required for all services, not just mission critical services. The purpose of this research is to realize a resilient storage infrastructure that utilizes relatively new memory devices such as non-volatile DIMMs (NVDIMMs). Managing the journaling area on the NVDIMM reduces the data journaling overhead that was traditionally not practical from a performance point of view, and it is also expensive to do frequent synchronous writes to maintain file consistency. It became possible to obtain performance. In particular, it is effective for workloads that require high data integrity.

研究分野：システムソフトウェア

キーワード：ストレージ　NVDIMM

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

ミッションクリティカルなサービスに限らず，あらゆるサービスに対し高い信頼性と可用性が求められるようになっている．たとえば，Google Drive 等のクラウド・ストレージや Facebook 等に保存されるデータについても，多くのユーザは，その内容が損なわれることはないと想定している．しかし，さまざまな文献等（参考文献 [1] など）で報告されているように，電源異常等の障害発生時にデータ損失が発生することは稀ではなく，一般ユーザが期待しているような高いレジリエンスの提供は容易ではない．クラウド・ストレージなどの一般向けサービスは，安価に利用可能であるという点に特色があり，レジリエンス向上のために多額のコストをかけることが難しい．また，保持しているデータ量が膨大であり（Facebook では 65 ペタバイトを超える [3]），ストレージの多重化といった旧来の方式をそのまま適用することは難しい．そのため，ストレージ基盤のレジリエンス向上は重要な研究課題となっている[2][4].

２．研究の目的

本研究の目的は，不揮発性 DIMM (NVDIMM) という比較的新しいメモリ・デバイスを活用したレジリエント (resilient) なストレージ基盤を実現することである．NVDIMM とは DRAM にフラッシュメモリによるバックアップを設けたデバイスであり，メモリと同等な高速かつバイト単位でのアクセスが可能でありながら，データの不揮発性を保証する．NVDIMM をストレージの高速化に用いるのではなく，レジリエンスの向上に用いる点に特徴がある．

３．研究の方法

本研究の狙いは，比較的新しいメモリ・デバイスである不揮発性 DIMM (NVDIMM) を活用し，レジリエントなストレージ基盤を実現することである．特に，NVDIMM-N という規格の NVDIMM を用いる．NVDIMM-N は，

1) DRAM と同等のアクセス遅延である
2) バイト単位でのアクセスが可能である
3) 不揮発性を持ち，システムの異常停止や再起動後も記憶内容が失われない，

という特徴を持つ．こうした NVDIMM-N の特徴を活用し，高レジリエンスかつ低アクセス遅延のストレージ基盤を実現する．従来のファイルシステムやキーバリューストアが高いレジリエンスを達成できない理由は，高レジリエンスと低遅延アクセスのトレードオフが存在し，性能上の要求からレジリエンスを犠牲にせざるを得ないからである．

本研究の狙いは，高レジリエンスと低アクセス遅延のトレードオフをぶち破り，高レジリエンスと低遅延アクセスの両立を達成することである．具体的には，次の視点からアプローチを行う．

・NVDIMM-N の低遅延アクセスを利用する．NVDIMM-N 上にジャーナル領域を設け，ジャーナル領域へのアクセス遅延を隠蔽する．その結果，データジャーナリングのように多量のデータをジャーナル領域に書き出しても，性能上のペナルティは小さく，レジリエンスの向上が期待できる．さらに synch 処理による遅延も隠蔽可能であると期待できる．ただし，これまでの予備調査により，ジャーナル領域を NVDIMM 上に配置しただけではアクセス遅延の隠蔽は達成できないことがわかっている．これは，既存の I/O 処理スタックが，高遅延でブロックアクセスのみが可能なハードディスクやフラッシュメモリ等に最適化されているためであり，バッファリングや同期制御など複雑な処理が行われているためである．

・ NVDIMM-N がバイトアクセス可能であるという性質を利用する．ジャーナル領域を NVDIMM 上に配置しただけではアクセス遅延の隠蔽は達成できない．その理由のひとつは，ジャーナルのスナップショット処理がコミット処理に追いつかず，結局，ジャーナリングにおける遅延が生じるためである．この問題については，ジャーナル領域を仮想的に大きく見せることによって隠蔽できるのではないかと考えている．具体的には，NVDIMM-N はバイト単位でのアクセスが可能であるという特徴を利用し，ジャーナル領域への書き込みに対して，デルタ圧縮などの技法を用いることでコミットによるジャーナル領域の圧迫を遅らせ，スナップショット処理がコミット処理に追いつかないようにする．

参考文献

[1] M. Zheng, et al., “Torturing Databases for Fun and Profit”, USENIX Symposium on Operating Systems Design and Implementation, pp. 449-464, Oct. 2014.
[2] H. Chen, et al., “Using Crash Hoare Logic for Certifying the FSCQ File System”, ACM Symposium on Operating Systems Principles, pp. 18-37, Oct. 2015.
[3] S. Muralidhar, et al., “f4: Facebook ’s Warm BLOB Storage System”, USENIX Symposium on Operating Systems Design and Implementation, pp. 383-398, Oct. 2014.
[4] T. Ridge, et al., “SibylFS: formal specification and oracle-based testing for POSIX and real-world file systems”, ACM Symposium on Operating Systems Principles, pp. 38-53, Oct. 2015.

４．研究成果

① データジャーナリングの必要性

本研究では，データジャーナリングを頻繁に行ってもそのオーバーヘッドを低く抑えること

ができれば，ファイルシステムそのものの堅牢性を大きく改善できることを示す．すでに述べたように既存のファイルシステムの多くは，性能上の懸念からデータジャーナリングを利用することはなく，メタデータのジャーナリングに留まっている．また，ファイルシステムの一貫性保証のためのセマンティクスは，ファイルシステムやその動作モードによって異なるため，ソフトウェア開発者が全てのファイルシステム，それらのすべての動作モードに対して期待したように動作するように実装するのは難しい． 表 1 にさまざまなファイルシステムおよびその設定によって一貫性の保証がどのように異なってくるのかをまとめた．

表 1 一貫性保証のセマンティクス

	ext2	ext2-sync	ext3-writeback	ext3-ordered	ext3-datajournal	ext4-writeback	ext4-ordered	ext4-nodealloc	ext4-datajournal	btrfs	xfs	xfs-wsync	reiserfs-nolog	reiserfs-writeback	reiserfs-ordered	reiserfs-datajournal
Atomicity																
Single sector overwrite																
Single sector append	×		×		×									×		
Single block overwrite	×	×	×	×		×	×	×			×	×	×	×	×	×
Single block append	×	×	×			×							×	×		
Multi-block append/writes	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Multi-block prefix append	×	×	×			×							×	×		
Directory op	×	×											×			
Ordering																
Overwrite->Any op	×		×	×		×	×	×			×	×	×	×	×	×
[Append, rename]->Any op	×		×			×							×	×		
O_Trunc Append->Any op	×		×			×							×	×		
Append->Append (same file)	×		×			×							×	×		
Append->Any op	×		×			×	×			×	×		×	×		
Dir op->Any op	×									×			×			

この問題を回避するためには，ファイルシステムへの書き込みごとにバッファキャッシュの同期を行えばよい．現在のファイルシステムではジャーナリングに伴うオーバーヘッドのために，このような実装を行うことは現実的ではない．図 1 から分かるように同期を頻繁に行った場合（w/ fsync）大きく性能が劣化していることが分かる．

さらにジャーナリングを行った場合のジャーナリングのオーバーヘッドの測定を行った．図 2 に示したようにジャーナリングを一切行わない場合と比較し，データジャーナリングを行った場合のスループットは最大 30 分の 1 程度にまで劣化してしまう．メタデータのみのジャーナリングを行った場合，ジャーナリングを行わない場合と比較し，若干の性能低下は見られるものの，この程度の劣化であれば許容範囲内であると言える．実際，メタデータのジャーナリングさえも行わないようにすることは稀であり，信頼性向上のためであればこの程度のオーバーヘッドは実運用上も許容されることが分かる．

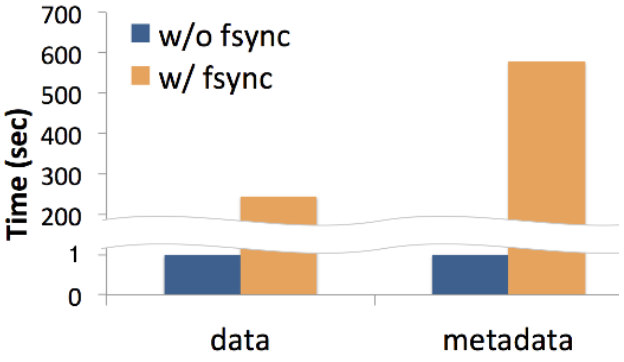

図 1 同期を頻繁に行なった場合のオーバーヘッド

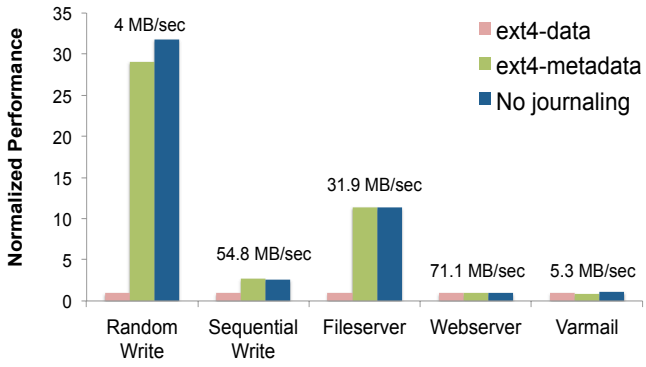

図 2 ジャーナリングのオーバーヘッド

② NVDIMM 上へのジャーナリングの実現

Linux 上のファイルシステムである ext4 を対象としてジャーナル領域を NVDIMM 上で管理する方式の実現を行なった．そのためには，ジャーナリング領域を NVDIMM 上に配置するだけで十分であるように見えるものの，実装上はジャーナリングの堅牢性を保証するためにさまざまな工夫が必要とされる．

通常の磁気ディスク等ではセクタ単位での書き込みのアトミシティ (atomicity) がハードウェアによって保証されている．そのため，ジャーナリングを実現する既存のコードはそのアトミシティを前提として設計・

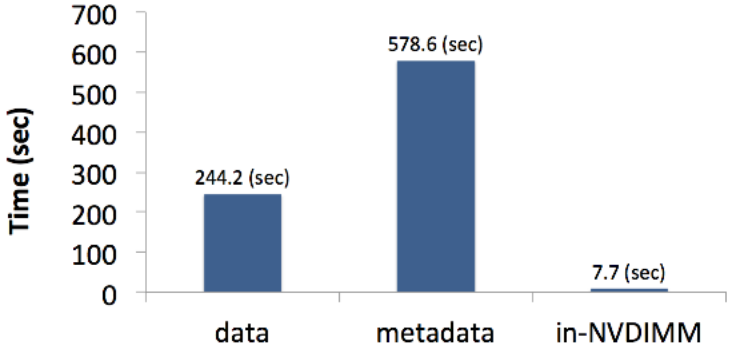

図 3 ロギングワークロードにおける実行時間

実装されている．通常時は NVDIMM は単なる DRAM として動作しているため，そのようなアトミシティは提供されていない．実装の詳細等は割愛するものの，NVDIMM 上でのジャーナリングを実現するためにはさまざまな技術的課題があったことを明記しておく．

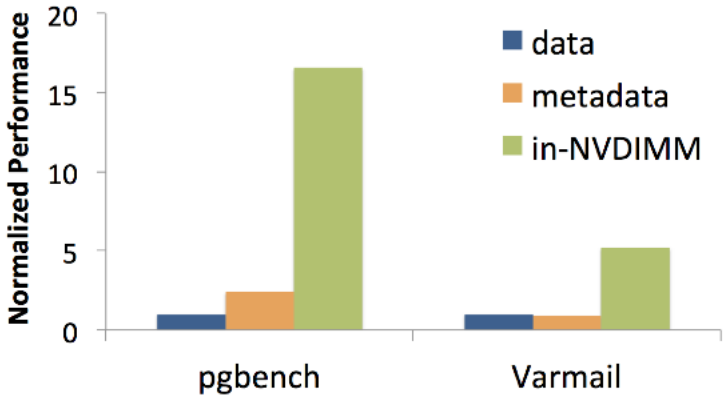

図 4 同期的書き込みの多いワークロード

NVDIMM 上でジャーナリングを行なった場合，ワークロードによって得られる性能向上が大きく異なることが判明した．Fsync を多く実行するような同期的なワークロードではその性能向上が著しい．頻繁に同期処理を行うロギングワークロードを対象に性能測定を行なった結果を図 3 に示す．この図からわかるように NVDIMM 上でジャーナリングを行うとその性能向上は著しい．図 4 に他のふたつのベンチマークに対する結果を示す．これらのベンチマークはどちらも比較的ファイルシステムに対する同期的な書き込みが頻繁に行われるため，大きな性能向上が達成されている．

しかしながら，同期的な書き込みが頻繁に行われることのないワークロードでは，従来のデータ・ジャーナリングに対しては大きな性能向上を得ることはできるものの，そのオーバーヘッドはメタデータ・ジャーナリングと比べて優位に大きい．その結果を図 5 に示す．なお，図 5 には比較のために Varmail の測定結果も掲載している．この図からわかるようにディスクに対するデータ・ジャーナリング（グラフ中の青いバー）に対しては大きな性能向上が得られているもの，メタデータ・ジャーナリングと比較した場合，ある程度のオーバーヘッドが発生している．すでに述べたとおり，本研究で提案する方式ではすべてのデータに対してジャーナリングを行なっているため，データ・ジャーナリングと同等の信頼性を保証しており，メタデータのみのジャーナリングよりもその堅牢性は高い．従って，このオーバーヘッドは許容範囲であると考えている．また，データの高い保全性が求められるアプリケーションは同期的な書き込みを頻繁に行なっており（同期的な書き込みを頻繁に行わない限り，データ損失が起きる可能性が高くなる），本提案方式と相性がよい．高いデータ保全性を必要としないアプリケーションでは旧来どおりにメタデータのジャーナリングのみを行えばよく，管理者の裁量によって判断すればよいと考えている．

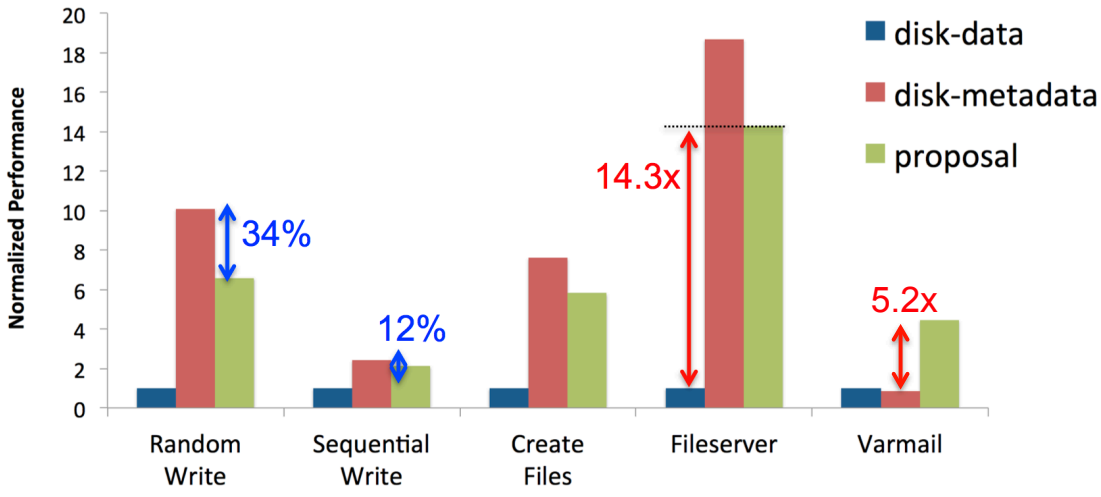

図 5 同期的書き込みの少ないワークロード

③ ファイルシステムの信頼性に関する調査

本研究課題の発展的内容として，そもそも信頼性の高いファイルシステムを実現する上での阻害要因について調査を行なった．この調査を行うことで，ジャーナリング領域以外の領域（例

えば inode 領域など）を NVDIMM 上で管理することで，ファイルシステム全体の信頼性向上に繋げることができないかどうかといった知見を得ることを狙って行なった調査である．ディスク上のブロック領域の使用状況を表すビットマップの信頼性を高めるだけでファイルシステム全体の信頼性を高めることができるといったようなブレークスルーにつながるような研究を狙ったものである．具体的には xfstest というファイルシステム向けのテストスーツを利用し，テストケースでカバーされていない領域に信頼性阻害要因が潜んでいるはずだという仮定で調査を行なった．図 6 に調査結果を示す．

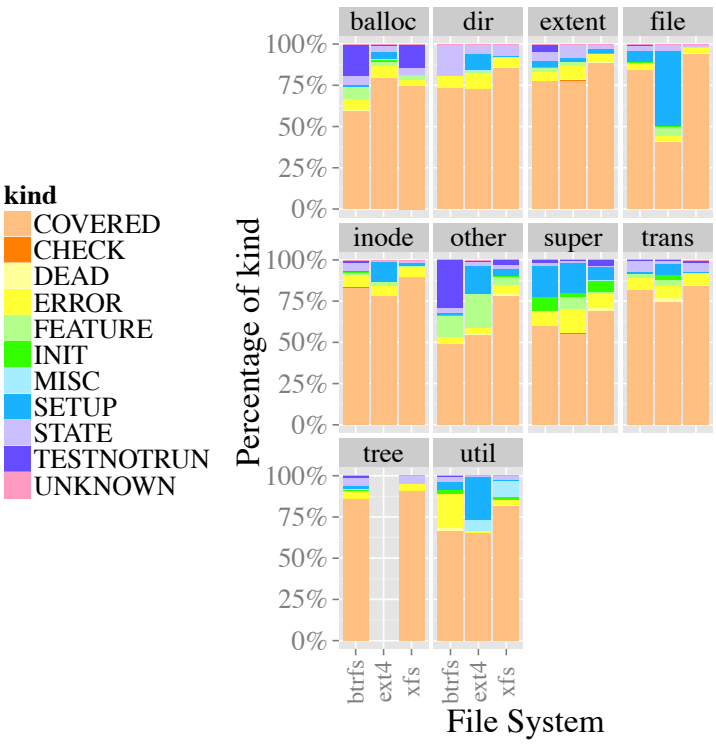

図 6 ファイルシステムの信頼性阻害要因

ファイルシステムの信頼性阻害要因としてもっとも著しいのは，そもそも多様なファイルシステムの機能が網羅的に検査されていないためであった．しかし，ファイルシステムの管理情報のコーナーケースは十分に検査されておらず，信頼性阻害の大きな要因の一つであることが判明した．ディスク上のデータ構造は，性能向上のために複雑な構造となることが多い．そのため，検査の難しいコーナーケースが多い．管理情報の一部を NVDIMM 上に委譲し，NVDIMM の低レイテンシでのアクセスを前提とし，より単純な構造をとれば信頼性の向上に寄与するのではないかという知見を得た．これは将来の研究課題としたい．

５．主な発表論文等

〔雑誌論文〕（計　2　件）

① Naohiro Aota, Kenji Kono, “File Systems are Hard to Test - Learning from Xfstests”, IEICE Transactions on Information and Systems, 査読有, Vol. E102-D, No. 2, pp. 269-279, 2019.
DOI: 10.1587/transinf.2018EDP7006

② Asraa Abdulrazak Ali Mardan, Kenji Kono, “Containers or Hypervisors: Which Is Better for Database Consolidation?”, International Workshop on Quality of Service Assurance in the Cloud in conjunction, 査読有, pp. 564-571, 2016.
https://ieeexplore.ieee.org/document/7830739

〔学会発表〕（計　3　件）

① 飛松 秀三郎，青田 直大，Asraa Abdulrazak Ali，河野 健二，“コンテナ環境におけるジャーナリング I/O の制御”，情報処理学会研究会報告（SIGOS），2018 年.

② 青田 直大，河野 健二，“ファイルシステムにおける tail latency の定量的分析”，情報処理学会研究会報告（SIGOS），2018 年.

③ 迫田 賀章，青田 直大，河野 健二，“不揮発性 DIMM を用いた LSM-tree によるキーバリューストアの性能向上”，情報処理学会研究会報告（SIGOS），2016 年.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：

権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：

ローマ字氏名：

所属研究機関名：

部局名：

職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：