ストレージネットワーキング・
インダストリ・アソシエーション

テクニカル・ホワイトペーパー

SNIA 永続メモリ性能試験仕様の概要

2020 年 10 月

要約：
このホワイトペーパーは、SNIA の性能試験仕様（PTS）に精通しているストレージ・プロフェッショナル、ストレージ・アーキテクチャ内の永続メモリと永続メモリ（PM）対応ソフトウェア・アプリケーションの理解、試験、および設計に興味のあるストレージ・アーキテクトとソフトウェア・アーキテクト、および PM ストレージ・ソリューションに興味のあるストレージ・アーキテクトとマーケティング・マネージャを対象とする。

使用にあたって

SNIA では、本書の使用を個人に対しては個人的利用に限定して許可し、法人およびその他の事業主体に対しては社内利用（社内での複製、配布、および掲示を含む）に限定して許可する。ただし、次の要件が満たされていることを前提とする。

1. テキスト、図、チャート、表、または定義を複製する場合は、変更を加えずに全体を複製すること
2. 本書からの資料（または本書の一部）を複製した印刷文書または電子文書は、その資料に対する SNIA の著作権を表示し、SNIA から再利用の許可を得ていることを明示すること

上記で明示的に規定されている場合を除き、本書の商業的利用、本書の一部または全部の販売、または本書の第三者への配布を行ってはならない。明示的に付与されていないすべての権利は、明示的に SNIA に留保されている。

上記以外の目的での本書の使用の許可は、tcmd@snia.org に電子メールを送付して要請する。要請する個人および／または法人の識別情報と、要請する使用の目的、性質、および範囲の簡単な説明を含めること。

この SNIA 文書内のすべてのコード・フラグメント、スクリプト、データ・テーブル、およびサンプル・コードは、次のライセンスに基づいて利用できる。

3 条項 BSD ソフトウェア・ライセンス

Copyright (c) 2020、ストレージネットワーキング・インダストリ・アソシエーション。

ソース形式かバイナリ形式か、変更するかしないかを問わず、以下の条件を満たす場合に限り、再頒布および使用が許可される。

* ソース・コードを再頒布する場合、上記の著作権表示、本条件一覧、および下記免責条項を含めること。

* バイナリ形式で再頒布する場合、頒布物に付属のドキュメントなどの資料に、上記の著作権表示、本条件一覧、および下記免責条項を含めること。

* 書面による特別な事前の許可なしに、本ソフトウェアから派生した製品の宣伝または販売促進に、ストレージネットワーキング・インダストリ・アソシエーション（SNIA）の名前またはコントリビューターの名前を使用してはならない。

本ソフトウェアは、著作権者およびコントリビューターによって「現状のまま」提供されており、明示黙示を問わず、商品性および特定の目的に対する適合性に関する暗黙の保証も含め、またそれに限定されない、いかなる保証も行われない。著作権者もコントリビューターも、事由のいかんを問わず、損害発生の原因のいかんを問わず、かつ責任の根拠が契約であるか厳格責任であるか（過失その他の）不法行為であるかを問わず、仮にそのような損害が発生する可能性を知らされていたとしても、本ソフトウェアの使用によって発生した（代替品または代用サービスの調達、使用の喪失、データの喪失、利益の喪失、業務の中断も含め、またそれらに限定されない）直接損害、間接損害、偶発的な損害、特別損害、懲罰的損害、または結果損害について、一切責任を負わない。

免責事項

この文書に含まれる情報は、事前の通知なく変更される場合がある。SNIA はこの仕様書に関していかなる種類の保証も行わない。これには商品性および特定の目的に対する適合性の暗黙的保証が含まれるが、これらに限定されない。SNIA は、本書に含まれる誤りあるいはこの仕様書の交付、履行、または使用に関連した偶発的または結果的損害に対して責任を負わない。

改訂に関する提案は、http://www.snia.org/feedback/まで。

目次

図表目次

I. 要約

一般的に、永続メモリ（PM）は次のような特徴をもって定義される。

- 超低遅延、メモリ／DIMM 速度を実現
- 電源を入れ直したり、アプリケーションやシステムをリセットしたりしても保持される不揮発性データ
- バイト・アドレス指定可能、バイト・アドレスを使用して直接ストレージ・メディアにアクセスできる
- アプリケーション向けの高い性能

中には、DRAM より低いコストで高い容量をメモリ層に提供するものもある。ほぼすべてのタイプがソリッド・ステート・ドライブより高い性能を示す。

そのため、PM に関する性能試験仕様（PTS）の作成に大きな関心が寄せられている。この仕様には、試験設定、メトリクス、方法論、ベンチマーク、および参照オプションが含まれることになる。これらの試験は、繰り返しの適用にわたって信頼性の高い結果が得られるものとなる。

このホワイトペーパーは、ストレージ・アーキテクチャやメモリ・アーキテクチャを担当している試験開発の専門家を対象とする。最新の PM PTS v1.0.1 は、ブロック IO 読み取り／書き込み試験とバイト・アドレス指定可能なロード／ストア・アーキテクチャの両方に対応している。また、3D XPoint、NVDIMM、MRAM、ReRAM などの様々な PM にも対応している。今後のホワイトペーパーの目標は、より具体的なアーキテクチャに試験を適用し、新しい試験ベンチマークを定義することである。

PM PTS では、人工的ワークロードと現実世界ワークロードの両方が使用される。人工的試験は、SNIA PTS v2.0.1 for NAND Flash SSD の改訂版に基づく。現実世界ワークロード試験は、SNIA Real World Storage Workload (RWSW) PTS for Datacenter Storage v1.0.7 に基づく。

PM PTS v1.0.1 が公開されれば、ユーザは、PM ストレージ用のソフトウェア・スタックを理解して最適化できるようになるはずである。PM PTS の開発には、興味のある業界の専門家、研究者、および教育者の参加を歓迎する。askcmsi@snia.org で SNIA にご連絡いただきたい。

II. はじめに

永続メモリ（PM）は、PCIe バス上に存在する従来のブロック IO ストレージとは異なり、キャッシュ・コヒーレント・リンク上に存在する高性能、低遅延、バイト・アドレス指定可能、不揮発性のストレージとして広く定義されている。将来のコヒーレント PCIe 実装で PM がサポートされることも考えられる。PM は、ストレージ階層内でメイン・メモリ DRAM の下位、NAND フラッシュベースの SSD の上位に位置する 1 つの層を占有するように想定されている。図 1 ー ストレージ階層を参照されたい。

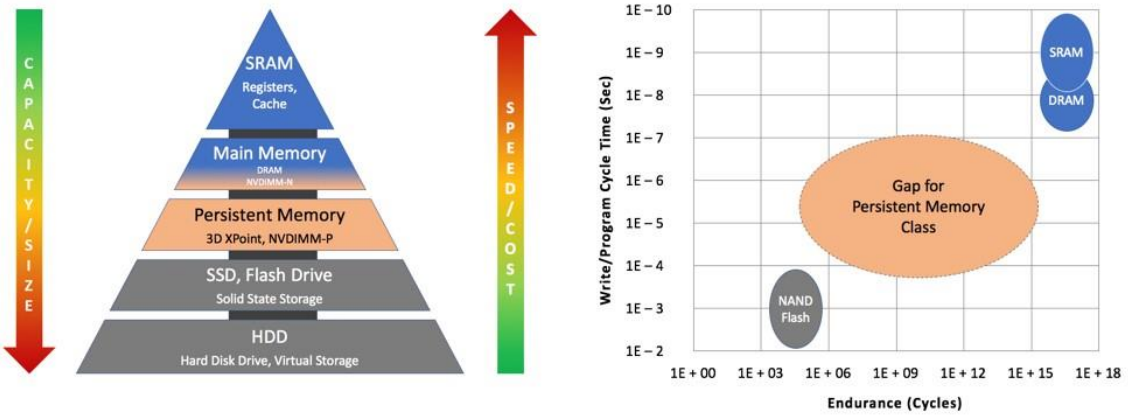

図 1 ー ストレージ階層

PM は、メイン・メモリ DRAM より安価で高容量だが、SSD より高価で高速である。図 2 ー ストレージ階層：容量と速度を参照されたい。

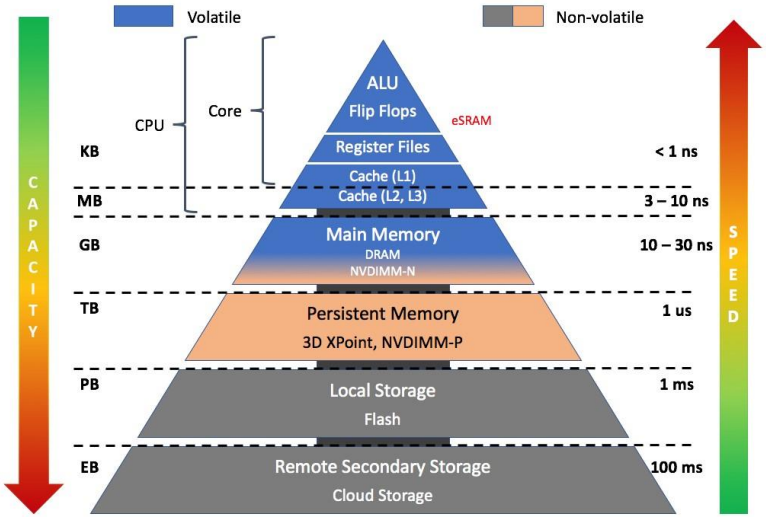

図 2 ー ストレージ階層： 容量と速度

この WP では、PM ストレージ・アーキテクチャの特性、最適化、および試験に焦点を当てる。他の SNIA の技術的研究（https://www.snia.org/sites/default/files/technical_work/final/NVMProgrammingModel_v1.2.pdf で参照可能な NVM プログラミング・モデルなど）で取り上げられている PM プログラミング・モデルやソフトウェアの最適化とは異なる。PM ストレージ技術の例として、3D XPoint、NVDIMM DRAM、相変化メモリ、MRAM、ReRAM、STRAM などが挙げられる。

III. 背景 ― 性能試験標準の必要性

ハード・ディスク・ドライブ

ハード・ディスク・ドライブ（HDD）に関する初期の性能仕様では、購入者が容易に HDD 性能を比較できるようにするための特定の人工的なアクセス試験パターンを定義することに重点が置かれていた。最初のベンチマーク「コーナー・ケース・ストレス」試験（ランダム 4K 書き込み飽和試験など）は、通常のオペレーションの範囲外の HDD 性能をモニターするように設計されていた。そのため、「コーナー・ケース・ストレス」試験と呼ばれている。一般的なベンチマーク試験は、様々なブロック・サイズ（BS）の読み取りまたは書き込み（R/W）IO のランダム（RND）またはシーケンシャル（SEQ）アクセスを測定し、IO 速度（IOPS）、帯域幅（MB/s）、および応答時間（または遅延）の観点から性能を報告していた。

HDD の「速度、供給量、および容量」は進歩し続けていたが、HDD の性能の向上は、CPU、メイン・メモリ、メモリ・キャッシュ、およびソフトウェア／ハードウェア（SW/HW）スタック全体の性能の向上に比べて遅れていた。この性能面のギャップが、低い HDD の性能をカバーするための SW/HW スタックの最適化につながった。そこで、速い SW/HW スタックが遅い HDD ストレージ層に対する IO 要求をキューに入れることで、高い性能（および高いコスト）の CPU リソースとメイン・メモリ・リソースを解放し、キューに入れられた HDD IO 要求が履行されるまで他のタスクに専念できるようにした。この補正機能は伝統的に、DRAM ベースのキャッシュを介して実装されたが、これは非常に高価であることが実証された。

NAND フラッシュ

この低い HDD 性能と速い SW/HW スタック間の性能分離によって、速いストレージ層のニーズが生まれ、NAND フラッシュベースのソリッド・ステート・ドライブ（SSD）の導入への道筋が付いた。SSD は HDD より 1 桁以上速い。ただし、NAND フラッシュベースの SSD では、SSD 性能の一貫して反復可能な測定に影響する複数の変数が新たに関係する。

SSD の速度の向上と NAND フラッシュ・ストレージの特殊性（R/W 非対称、FOB（Fresh-Out-of-Box）ピーク動作、書き込みサイクル寿命の制限、ブロック書き込み／ページ消去オペレーション、および性能を定常状態で測定するための事前調整の必要性）のために、様々な SSD 製品の性能を一貫して比較するために使用される試験ハードウェア・プラットフォーム、データ・パス・コンポーネントおよびドライバ、オペレーティング・システム、試験ソフトウェア、試験方法、試験設定、および特定の試験手順の標準化が必要になった。

これらの新しい課題を解決すべく、SNIA Solid State Storage（SSS）Technical Working Group（TWG）が、クライアント・クラス・ストレージとエンタープライズ・クラス・ストレージに関する一連の SSS 性能試験仕様（PTS）を策定して発行した。これらの PTS の目的は、SSD 性能結果に対する OS、ハードウェア、試験設定、試験方法、および試験ソフトウェアの影響を正規化するための標準の人工的ベンチマーク試験と参照試験プラットフォーム（RTP）を明示することである。SSS PTS v2.0.1 では、定常状態で NAND フラッシュ SSD を準備して試験するための標準化された試験方法が定義されている。

現実世界ワークロード

ストレージ・ソリューション・アーキテクチャおよびアプリケーションの進歩（ストレージ仮想化、ストレージ階層化、リモート・ストレージとファブリック・ストレージ、データ圧縮、データ重複排除、暗号化、オープン SSD、計算ストレージ、AI と機械学習、およびその他の最適化を含む）によって、SSD 性能のベンチマーキングにおける試験ワークロードの重要性が明確になった。SSD は、本質的に、IO ストリームのコンテンツと強度、つまり、ワークロードのキュー深さ（QD）の影響を受けやすい。様々なタイプの IO ストリームとワークロードの需要強度（または QD）が、IO、帯域幅と応答時間の飽和、および全体性能に大きく影響する。アプリケーション・ワークロード・コンテンツも SW スタックや抽象化の各階層で変化するため、性能ベンチマーキングにとって、現実世界のワークロードを使用したキャプチャ、分析、および試験がますます重要になっている。

この現実世界アプリケーション・ワークロードのコンテンツの重視と、ワークロードの組成やSSD 性能に対する SW/HW スタックの影響が、SNIA Real World Storage Workload (RWSW) PTS for Datacenter Storage v1.0.7 のリリースにつながった。この RWSW PTS では、現実世界のワークロードのキャプチャ、分析、および試験に関する標準が明示されている。

不揮発性メモリ・プログラミング・モデル

SW/HW スタック性能の継続的な進歩は、メイン・メモリにより近い新しいストレージ層のニーズも生み出した。SNIA は、新しい不揮発性メモリ（NVM）機能と新しい NVM 技術の継続的増加に対応し、様々なユーザ空間と NVM をサポートするオペレーティング・システム（OS）カーネル・コンポーネント間の推奨動作を定義するために、NVM プログラミング・モデルをリリースした。この NVM プログラミング・モデルは、PM にアクセスするために使用される動作ユーザ空間ソフトウェアを定義しており、永続メモリと関連 PM ドライバ用の標準プログラミング・モデルの作成を促した。これが、さらにもう 1 つの新しいストレージ層が求められる状況を作り上げた。

永続メモリ・ストレージ

永続メモリ（PM）ストレージは、開発対象の最新のストレージ層であり、メイン・メモリ DRAM と NVMe SSD ストレージの間に位置する。この新しい PM ストレージ・クラスが、ブロック IO 読み取り／書き込みとバイト・アドレス指定可能なロード／ストアの両方の性能の最適化とベンチマーキングに関する業界標準方法を PM PTS で明示する必要性を高めている。

PM ストレージと PM アプリケーションは、従来の IO スタック経由かユーザ空間からの直接アクセスにより、バイト・アドレス指定可能な IO とブロック・アドレス指定可能な IO の両方を使用してストレージにアクセスできる。従来の IO スタック・アクセスは、同期的に（pread または pwrite）または非同期的に（libaio）に行うことができる。永続メモリ・ユーザ空間アクセスのもう 1 つの興味深い属性は、ストアを永続化するためにキャッシュ・フラッシュ・ステップが必要となる可能性であり、これは性能に関係する。また、PM アプリケーションは、従来のブロック IO 転送サイズと異なり、小さいデータ転送サイズ（64 バイト～8KB）を使用する傾向がある。これらの要因が PM ストレージ試験設定、ハードウェア構成、および試験方法を定義して標準化する必要性につながった。

SNIA のベンダー中立性と製品不問性

PM PTS v1.0.1 の目的は、ベンダー中立性とストレージ・アーキテクチャ不問性の SNIA 方針を維持しながら、ブロックとバイトの両方のアドレス指定可能な性能ベンチマーク方法および試験を定義する必要性に対処することである。そのため、現時点で新しい PM の具現化はまだほとんど確認されていないが、ベンダー中立性とストレージ不問性を維持するためのあらゆる取り組みが行われている。

例えば、PM モジュールは、市販名ではなく、一般的な技術タイプで呼ばれている（3D XPoint PM モジュールやデータ・センターPM モジュールなど）。そのため、PM PTS で取り上げられた最初の PM ストレージ技術は 3D XPoint 技術と NVMDIMM-N/P であるが、他の PM ストレージ・アーキテクチャおよび技術についてはそれらが市販可能になった時点で今後の改訂版に追加される予定である。

IV. 適用範囲

この PM PTS ホワイトペーパー（WP）は、SNIA PTS 仕様に精通しているストレージ・プロフェッショナル、ストレージ・アーキテクチャ内の永続メモリと PM 対応ソフトウェア・アプリケーションの理解、試験、および設計に興味のあるストレージ・アーキテクトとソフトウェア・アーキテクト、および PM ストレージ・ソリューションに興味のあるストレージ・アーキテクトとマーケティング・マネージャを対象とする。

V. 背景：ブロック IO アクセスと PM バイト・アクセスの比較

従来のブロック IO アクセスでは、データのアクセスは、論理ブロック・アドレス（LBA）として表現される一定数のセクタ単位でアクセスされる。一般的に、セクタ・サイズは、最小 LBA サイズの 512 バイト（0.5 KB）または 4096 バイト（4 KB）である。このセクタ・サイズに、データ整合性フィールドに対応するためのバイトを追加することができるため、結果的に、520 バイトや 528 バイトなどのサイズになる。

論理ブロックは、NAND フラッシュ SSD やその他のストレージ（ハード・ディスク、光ディスク、テープなど）につながる PCIe などのストレージ・プロトコルを介してアクセスされる。この場合、CPU IO 要求はブロック IO ストレージに直接アクセスできないため、ドライブ・プロトコルに頼って LBA にアクセスする必要がある。

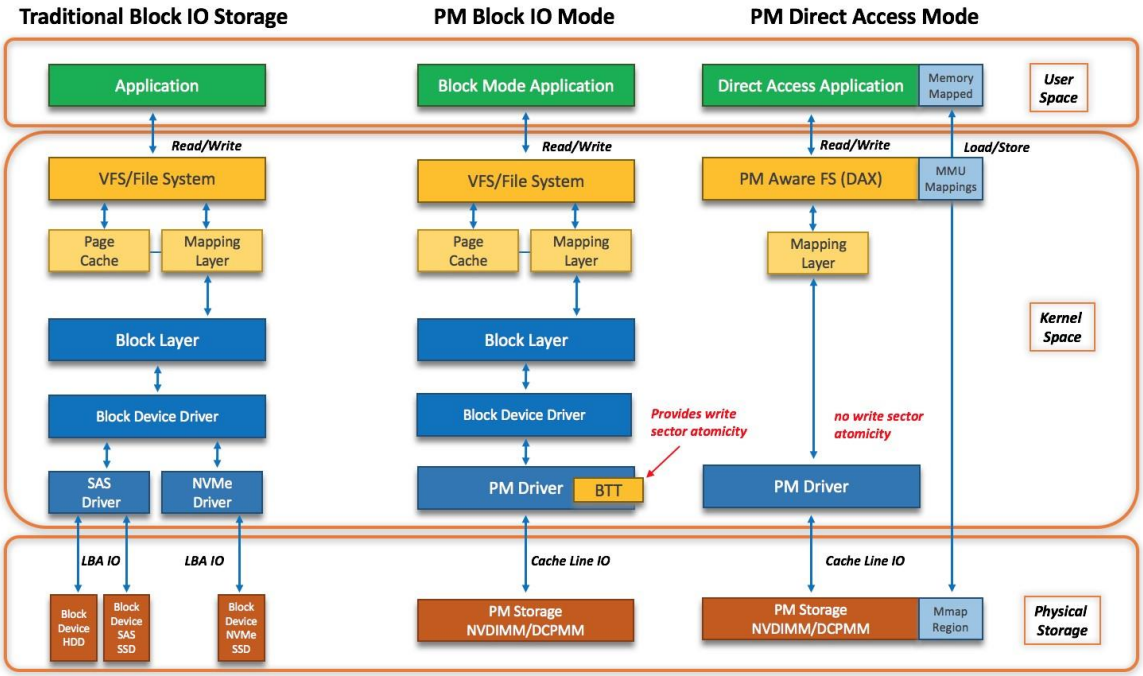

図 3 – 従来のブロック IO アクセス・モードと PM ダイレクト・アクセス・モードの比較

バイト・アクセス・ロード／ストアでは、CPU（および IO 要求）がキャッシュ・ライン（64 バイト・キャッシュ・ラインなど）へのバイト・アドレス・アクセスを使用して直接ストレージ・メディアにアクセスできる。PM デバイスをブロック IO モードで使用している場合は、PM 対応ドライバがブロック IO 要求をバイト・アドレス指定可能な（キャッシュ・ライン）メモリ・コピーに変換する。この変換は、ユーザにとって透過的であり、PM ドライバを介して PM に適用された場合に「従来の」ブロック IO の性能が向上する。

PM 対応アプリケーションの場合は、CPU から直接キャッシュ・ラインにアクセスできるため、PM ドライバを介したブロック IO や PCIe を介した従来のブロック IO より大幅に性能が向上する。

VI. バイト IO とブロック IO の比較

ブロック IO アクセス。CPU とストレージが分離されているという事実にはいくつかの意味合いが含まれているが、その中で最も重要なものは、CPU が*直接*ストレージ・サブシステムと対話できないという事実である。直接とは、CPU によって実行される命令で、ストレージ・デバイスに直接読み取り／書き込み（つまり、IO）オペレーションを発行できないことを意味する。代わりに、CPU は、PCIe などのプロトコルを使用して IO バス経由でデバイスに*要求*を送信する必要がある。このやり取りはドライバによって処理されるため、オペレーティング・システム（OS）に対する呼び出しが避けられない（多くの場合、OS への必須のコンテキスト・スイッチを含め、同じ CPU コア上で他のプロセスを実行するためのコンテキスト・スイッチも発生する）。発生する一部の遅延を隠すために、IO 要求は 4 キロバイト（KiB）（4096 バイト）などのあらかじめ決められた大きいサイズのブロック単位で行われる。さらに、データをユーザ空間内の中間ダイナミック・ランダム・アクセス・メモリ（DRAM）バッファにコピーしなければならない場合がある。これは、アプリケーションが正式なロード命令とストア命令を使用してアクセスするページ・キャッシュとは異なる。

バイト・アクセス。前の項ではブロック・アクセスについて説明した。ブロック IO アクセスとは対照的に、バイト・アクセスでは、CPU がメディアと直接対話できる。一部では、この種のメディアが*バイト・アドレス指定可能*と呼ばれている。バイト・アドレス指定可能なメディアの例にダブル・データ・レート（DDR）メモリがある。メモリ・バス上に存在する永続メモリ技術もバイト・アドレス指定可能である。図 4 に、この 2 つのアクセス・モードのハードウェア・レベルの違いを示し、図 5 に、そのソフトウェア・レベルの違いを示す。

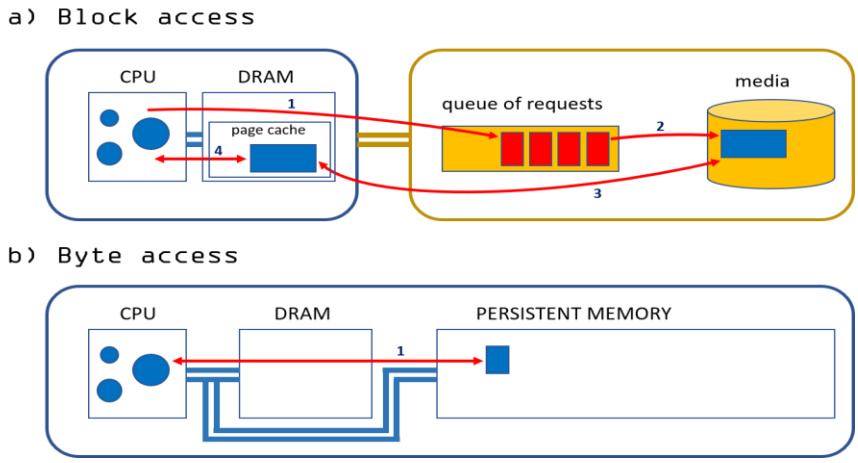

図 4 – ハードウェア・ビュー：ブロック・アドレス指定可能アクセスと
バイト・アドレス指定可能アクセスの比較

ハードウェア・ステップ – ブロック IO アクセス・パス。図 4 内の数字は IO パスの様々なステップを示している。図 4a のブロック・アクセスのケースでは、すべての IO オペレーションで CPU が以下を行う必要がある。

(1) デバイス内に要求をキュー。
(2) デバイス自体による要求の実行。これが追加のステップをトリガする可能性がある – 次を参照。
(3) DRAM からブロックを読み取り、それをストレージ・メディアに書き込むためのダイレクト・メモリ・アクセス（DMA）オペレーション、または、その逆（ページ・キャッシュはバイパス可能）。簡単にするために、図 4 ではページ・キャッシュを使用したケースのみを考慮する。また、重要な留意点として、ブロック IO では長い時間がかかることが普通であるため、OS はデバイスが IO を完了するまで待っている間に他の実行可能なプロセスに切り替え、IO が完了した時に元に切り替える。
(4) アプリケーションは単純にデータを読み取る（オペレーションが読み取りの場合）かデータを書き込む（書き込みオペレーションの発行前）。

ハードウェア・ステップ ― バイト IO アクセス・パス。図 5b のバイト・アクセスでは、CPU がキャッシュ・ラインの粒度でシングル・ステップでデバイス内のデータに直接アクセスできる。

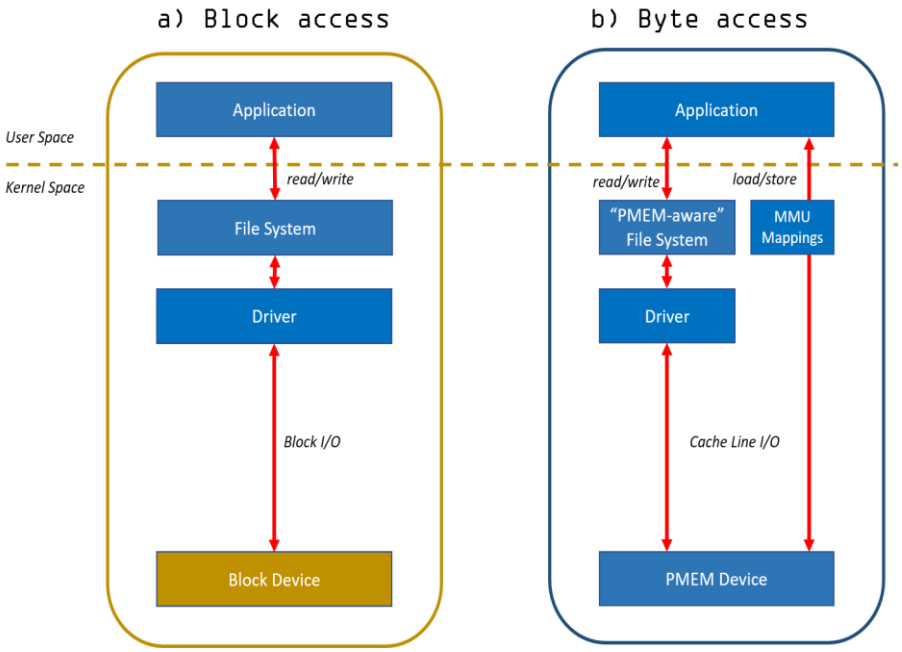

図 5 ― ソフトウェア・ビュー： ブロック・アドレス指定可能アクセスと
バイト・アドレス指定可能アクセスの比較

ソフトウェア・ステップ ― ブロック・アクセスとバイト・アクセスの違い。図 5 では、ソフトウェア・レベルでのブロック・アクセスとバイト・アクセスの 2 つの主な違いを確認できる。

1) 1 つ目の違いは、ブロックがファイル・システム経由で読み書きされた場合でも、必ず、図 4b で行われるすべての IO がキャッシュ・ラインの粒度で実行されることである。

2) 2 つ目の違いは、ファイルをメモリ・マップすることによって OS をバイパスできることである。このケースでは、アプリケーションが、ユーザ空間から直接、ロードとストアを介してファイルにアクセスしたり、データを CPU キャッシュからフラッシュしたりできる。

図 4 と図 5 は、アプリケーションで IO を実行可能なすべての方法を網羅しているわけではないが、この考察の目的には十分に役立つ。例えば、図 5a のブロック・アクセス・ケースでファイルをメモリ・マップしてから、ロード命令とストア命令を使用してデータにアクセスするアプリケーションのケースを考えてみる。このケースでも、ファイル・システムはまだデバイスからブロックを取り込んで DRAM に保存する必要がある。実際には、キャッシュされていない（がマップされている）ブロックへのバイト・アクセスは、デバイスに対してブロック全体の読み取り要求を発行することと本質的に同じである。また、アプリケーションが保留中の書き込みを DRAM バッファからフラッシュして持続性を確保したい場合には、ファイル・システムが関与する。

先に進む前に、このセクションの重要なポイントをまとめる。

ブロック・アクセスでは：

1. メディア内のデータへのアクセスがブロック単位で行われる。
2. CPU は直接データにアクセスできない。
3. IO パス内で常に、OS が必要とされる。

バイト・アクセスでは：

1. 読み取り／書き込みを要求するファイル・システム経由で IO が行われる場合でも、必ず、メディア内のデータへのアクセスがキャッシュ・ラインの粒度で行われる。キャッシュ・ラインより大きいオペレーションは分割される。
2. CPU は直接データにアクセスできる。
3. OS に対するシステム・コールは、ファイルをメモリ・マップすることによって避けられる。これにより、ユーザは、ユーザ空間からロードとストアを発行し、CPU キャッシュからデータを完全にフラッシュできる。

VII. PM アクセス用の様々な IO パス

前述したように、PM は、高速（低遅延）かつ永続的（不揮発）であり、市場で最速の SSD と DRAM メモリの間のストレージ階層の隙間に位置する（図 1 を参照）。遅延の観点では、PM の応答時間は SSD の遅延より DRAM の遅延にはるかに近い。この強力な低遅延と永続性の組み合わせにより、ストレージ・アーキテクトやソフトウェア開発者はさらなる IO アクセスのメリットを享受できる。

アプリケーション開発者は、使用する IO アクセスのタイプで PM をフル活用できる。PM を使用しない場合、開発者は、非 PM の小さいブロックの RND アクセスと SEQ アクセスに伴う大きな（遅い）アクセス遅延を隠すために大きい SEQ ブロック・サイズを使用するような次善策に頼らざるを得ない。しかし、PM を使用すれば、開発者は、より速くより小さいブロックの RND アクセスと SEQ アクセスだけでなく、OS システム・コールを呼び出さないユーザ空間からの直接的メモリ・コピー（ロード命令とストア命令）に頼ることができる。

しかも、PM は従来の IO スタックで使用することもでき、既存の PM 非対応アプリケーションでの使用も可能である。以降のサブセクションでは、アプリケーションで使用可能な PM の様々なモードを紹介する。

A. セクタ原子性を使用したブロック・アクセス

システム内に適切な PM ドライバとツールがインストールされていれば、アプリケーションやファイル・システムを変更せずに、永続メモリをストレージとして使用することができる。

前述したように、永続メモリに対して行われるすべての IO がキャッシュ・ラインの粒度（64 バイトが一般的）で行われる。これは、すべての IO オペレーションがドライバによってメモリ・コピーに変換されることを意味する。PM は、ブロック／セクタ・レベルでの書き込み原子性に依存しているアプリケーションやファイル・システムに対して電源障害書き込み原子性を提供しない。これは、クラッシュや電源障害が発生した場合にセクタがちぎれることによってアプリケーションのデータが破損する可能性があることを意味する（x86 アーキテクチャはクラッシュや電源障害によって 8 バイトがちぎれないことしか保証しないため）。

このようなシナリオを回避するために、PM ドライバは、*セクタ・モード*と呼ばれる原子セクタ書き込みを可能にするモードをサポートしている。このモードでは、ドライバが、Block Translation Table（BTT）と呼ばれるデータ構造を維持することで、ちぎれたセクタが発生しないことを保証する（図 3 を参照）。

B. セクタ原子性を使用しないブロック・アクセス

Linux の ext4 や xfs などの最新のファイル・システムは、Linux カーネル・バージョン 4.2 から永続メモリ対応になっているため、電源障害書き込み原子性を提供しない永続メモリ・メディアとうまく連動する。電源障害書き込み原子性を使用せずに PM を使用するためのモードは fsdax と呼ばれている。

ただし、永続メモリ対応ファイル・システムが提供している保護は、アプリケーションのデータではなく、ファイル・システムのメタデータのみを対象とすることに留意されたい。アプリケーションがデータのセクタに対する書き込み原子性に依存している場合は、アプリケーションを再設計しなければならない可能性がある。同様に、ファイル・システムなしでデバイスを使用している（一部のデータベースがそうなっている）場合も、この点を考慮したアプリケーション設計にする必要がある。

C. ダイレクト・アクセス

ダイレクト・アクセス（DAX）は、アプリケーションがファイルをメモリ・マップし、それをユーザ空間からメモリ・コピーを通して直接読み書きできる fsdax の「特殊なケース」である。システム内での DAX の構成方法は OS によって異なる。例えば、Linux では、オプション「DAX」を指定して永続メモリ対応ファイル・システムをマウントすればよい。

本書では、DAX はアプリケーションの観点からのダイレクト・アクセスを意味する（CPU は永続メモリ・デバイスに対して常にダイレクト・アクセス可能であることを思い出してもらいたい）。

fsdax の DAX オプションは、ストレージネットワーキング・インダストリ・アソシエーション（SNIA）が策定した NVM プログラミング・モデル（NVMPM）標準に準拠した永続メモリ・デバイスに対するプログラミング用に設計されていることを指摘しておく。つまり、アプリケーションがローカル・バッファへのコピーを行うことなくメディアへのダイレクト・ポインタを通してバイト・データにアクセスできるように特に設計されている。実際には、ローカル・バッファにコピーした場合は、同じバス上の 2 つのメモリ・デバイスの間（DRAM から PM へ）のデータ・コピーが 1 つ追加的に（恐らく不必要に）行われるだけである。

DAX オプションは、ファイルがメモリ・マップされている場合にアプリケーションがユーザ空間から直接永続メディアにアクセスできるようにすることに加えて、ページ・キャッシュを自動的にバイパスする（そうする必要がある）。ページ・キャッシュのバイパスは従来のファイル・システムでもすでに使用できたことに言及しておく必要がある。例えば、POSIX の場合は、O_DIRECT を指定してファイルを開くことによってページ・キャッシュをバイパスすることができる。相違点として、従来のケースでは以下のようになっていた。

1. すべての IO がユーザ空間バッファに対して、またはユーザ空間バッファから実行される（メディアへのダイレクト・ポインタなし）。
2. IO パス内に OS が含まれる。

一般的に、PM デバイスを使用している場合は、DAX の方がページ・キャッシュをバイパスするためのより良い方法と考えることができるが、DAX と O_DIRECT のどちらを使用すべきかはアプリケーションによって異なる。

VIII. PM の試験方法と試験

PM PTS では、様々なタイプの永続メモリを人工的ワークロードと現実世界ワークロードの両方を使用してベンチマークするように設計された試験方法と試験が明示されている。これらの試験は、主に、SNIA SSS PTS v2.0.1 for SSDs と RWSW PTS for Datacenter Storage v1.0.7 をベースにしている。どちらのケースでも、NAND フラッシュ SSD の動作とは異なる PM ストレージの動作に適応するための改訂がなされている。例えば、PM は NAND フラッシュの書き込みヒステリシスを示さないため、試験フロー内のデバイス・パージ、定常状態への事前調整、およびブロック・サイズ／需要強度の順序付けに関連する試験プロセス・ステップの大部分が不必要になる。

ドラフト PM PTS v0.3 内の人工的試験は次のとおりである。

1. DIRTH 単一ストリーム試験 ― 単一ブロック・サイズ／RW 比率 IO ストリームが様々な需要強度で適用される
2. DIRTH 複数ストリーム試験 ― 複数ブロック・サイズ／RW 比率 IO ストリームが様々な需要強度で適用される

ドラフト PM PTS v0.3 内の現実世界ワークロード試験は次のとおりである。

1. 再現試験 ― 現実世界のワークロード・キャプチャで観察された IO ストリームと QD のシーケンスと組み合わせを適用する
2. 個別ストリーム試験 ― 現実世界のワークロードで観察され、選択された IO ストリームのそれぞれを単一ブロック・サイズ／RW 比率飽和試験として試験する

A. DIRTH 試験

DIRTH 試験は、Demand Intensity Response Time Histogram（需要強度応答時間ヒストグラム）試験の頭字語である。これは、様々な需要強度のストレージ性能（IO、帯域幅、応答時間、CPU 飽和）を評価する際に、どのような所定の刺激（人工的ベンチマークや現実世界ワークロードの IO ストリーム）を適用するためにも使用できることから、SNIA PTS for SSDs と Real World Storage Workload PTS for Datacenter Storage の基礎試験になっている。

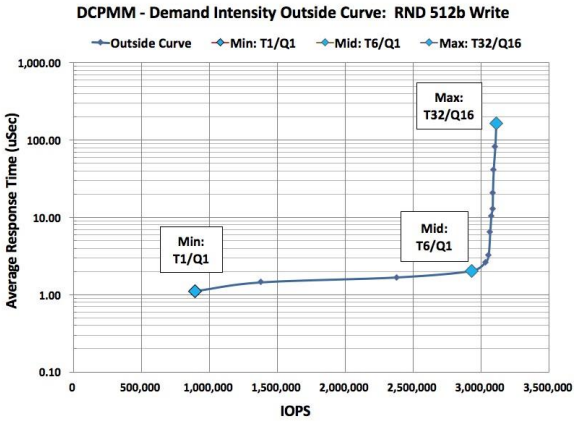

図 6 ― 需要強度外曲線

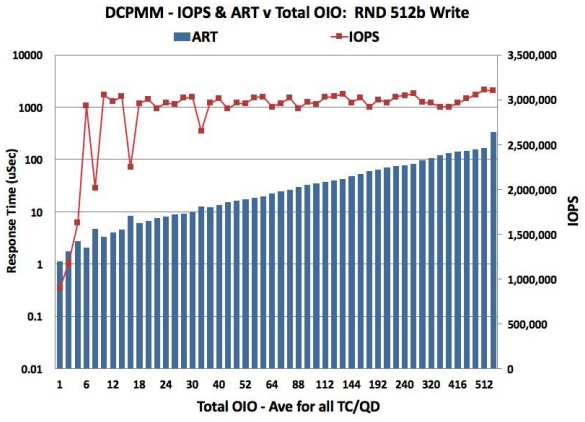

図 7 ― IOPS および ART と合計 OIO の比較

ストレージ性能の測定値は、主に、ワークロードの需要強度に依存する。この場合、需要強度（DI）はストレージに適用される未処理 IO（簡単に言えば、ジョブ、スレッド、または要求）の数を表す。DI が不十分な場合は、ストレージの IO または帯域幅の最高性能に達するために十分な要求が存在していない。DI が過剰な場合は、ストレージへのアクセスを同時に試みる IO が多すぎることによって引き起こされるボトルネックが応答速度の増大と IO 速度の低下を招く。

DIRTH 試験のメリットは、ユーザが同時に複数の性能の側面（IO、帯域幅（BW）、応答時間（RT）、CPU 使用率、および需要強度）を評価できることである。一般的に、DI が低いと、IOPS は低く、応答時間は短い。逆に、DI が高いと、IOPS と帯域幅は高く、関連する応答時間は長く、CPU システムの使用率は高い。図 6 と図 7 を参照されたい。

この試験の重要な価値は、ストレージがピーク IO/BW/RT 性能に達する DI のレベルと、RT が飽和して速度がどんどん低下したり、CPU システムの使用率がどんどん高くなったりするポイントを確認することである。

応答時間に対する IO、BW、および RT をプロットすることによって、試験操作者は、a) DI の関数として IO/BW を示す需要強度外曲線と、b) 合計未処理 IO（つまり DI）に対する IO/BW/RT/CPU 使用率の両方をプロットすることができ、そうすることで、様々な需要強度のワークロード（またはユーザ／スレッド・カウント）の下でのストレージ・デバイスの予想性能を示すことができる。図 7 を参照されたい。

B. 再現試験

再現試験は、オリジナルの現実世界ワークロード IO キャプチャで観察された IO ストリームと QD のシーケンスを試験ストレージに適用する。RWSW PTS 1.0.7 for DC Storage で定義された再現試験では、IO キャプチャ・ワークロードからの変化する IO ストリームと QD のシーケンスが再現される。再現試験の各ステップが、試験操作者の要求どおりに、一定時間（つまり、ステップ時間）にわたって試験ストレージに適用される。

再現試験を使用すれば、利用者は、現実世界のアプリケーション・ストレージ・サーバからキャプチャされたワークロード IO ストリームと QD の下でのターゲット試験ストレージの性能を確認することができる。IOPS、帯域幅、応答時間、キュー深さ、CPU 使用率、電力消費、温度、LBA 範囲アドレス、TRIM、書き込み増幅率などの主要な性能指標を表示できる。下の図 8 を参照されたい。

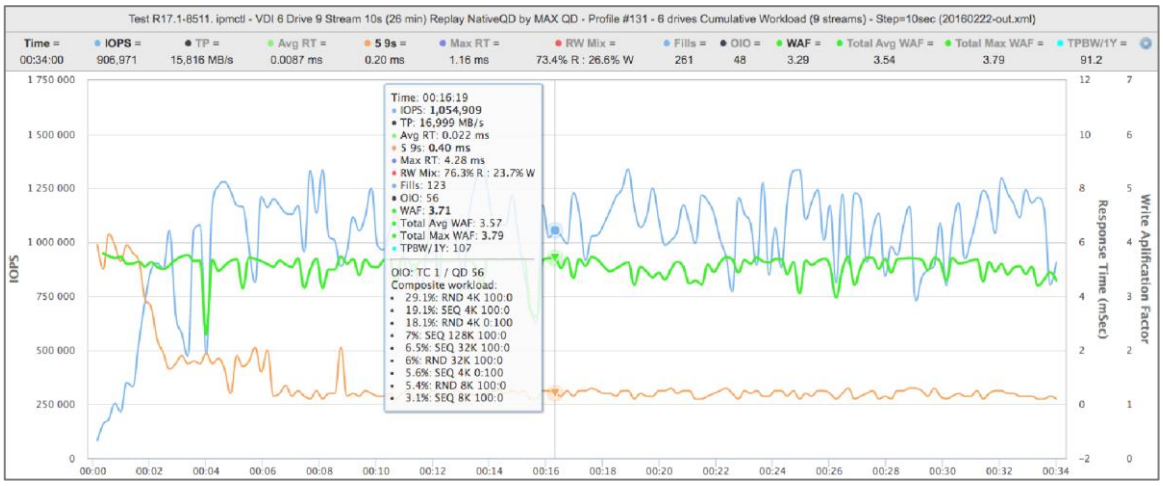

図 8 ― リアル・タイム・プロット： IOPS と応答時間サービス品質

再現試験の性能は、一定期間の性能（再現試験の各時間ステップ）を示すリアル・タイム・プロットで報告することも、単一の集計値（再現試験の時間ステップの一部または全部の平均 IOPS など）として報告することもできる。

再現試験の重要な価値は、単一の IO ストリームまたは複数 IO ストリームの固定の組み合わせを測定する人工的コーナー・ケース試験とは異なり、現実世界のワークロード・キャプチャで観察された IO ストリームと QD の変化する組み合わせに対して試験ストレージがどのように応答するかを示すことである。したがって、再現試験は、意図された現実世界のアプリケーション・ワークロードに対する試験ストレージの応答性能を示すことができる。

A. 個別ストリーム試験

個別ストリーム試験は、再現試験で観察され、試験された個々の IO ストリームの性能を測定する。WSAT（書き込み飽和）試験を適用することによって、試験操作者は、個々の IO ストリームの定常状態性能を測定し、その値をメーカー仕様値またはコーナー・ケース試験の試験仕様値と比較できる。

従来のストレージ・メーカー性能試験仕様は、RND 4K 読み取り／書き込み（IOPS 用）、SEQ 128K 読み取り／書き込み（帯域幅用）、RND/SEQ 4K 読み取り／書き込み遅延（1 つの未処理 IO での応答時間）などの広く使用されているいくつかの測定値を強調する傾向がある。

ただし、すぐに分かることだが、現実世界のワークロード IO ストリーム・コンテンツが単一の IO ストリームで構成されることはほとんどない。また、現実世界のワークロードの主要な IO ストリームには、通常、1K、2K、0.5K、396K などの BS フラグメントの非伝統的なベンチマーク BS/RW が含まれている。現実世界のワークロードに伝統的な BS ストリーム（RND 4K R/W や SEQ 128K R/W など）が*含まれている*場合でも、これらの IO ストリームが現実世界のワークロードの合計 IO ストリームに占める割合は小さい。

下の図 9 は、すべての IO ストリームが試験期間中にわたって平均化されるデータ・センター・ストレージ再現試験の 3 つのモード（Mmap、Msync、および Non Temporal Writes）での IOPS、平均応答時間、および 99.999 サービス品質（QoS）を示している。下の図 10 は、各 IO ストリームが定常状態に対して個別に試験される再現試験で観察された個々の IO ストリームの性能を示している。

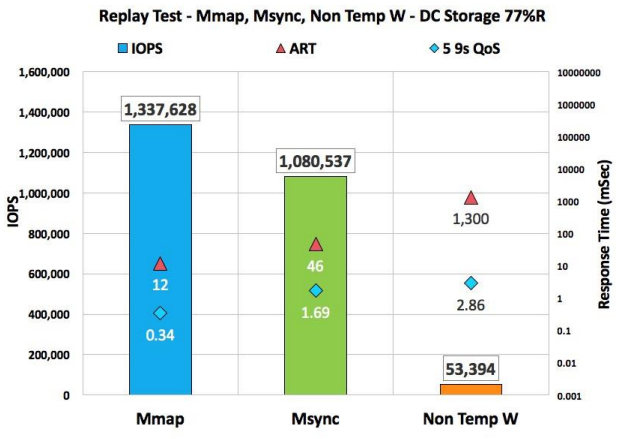

図 9 ― 再現試験：Mmap、Msync、Non Temp W

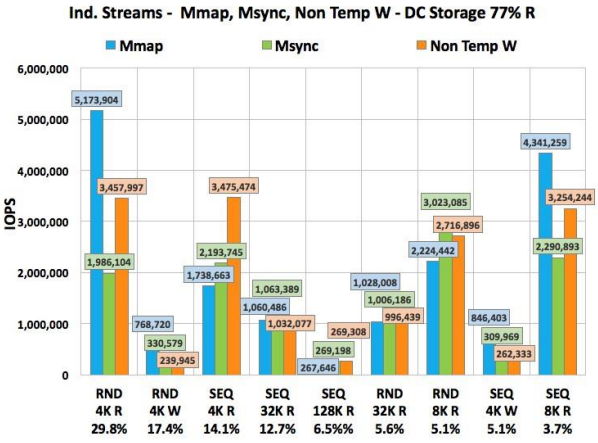

図 10 ― 個別ストリーム：Mmap、Msync、Non Temp W

個別ストリーム試験の重要な価値は 3 つあり、1) 現実世界のワークロード内に存在する BS/RW 比率の IO ストリームを示すこと、2) 定常状態での個々の IO ストリームの性能を示すこと、3) 現実世界ワークロード IO ストリームの性能と同じ単一 BS/RW 比率 IO ストリームのメーカー仕様ベンチマーク値を比較することである。

IX. 試験設定と参照試験プラットフォーム

PM PTS では、仕様でカバーされる PM のタイプごとの PM 参照試験プラットフォーム（PM RTP）が定義される。この RTP は、単なる推奨試験プラットフォームであって、PM PTS の要件ではない。また、一貫した反復可能で比較可能な性能ベンチマーク・データを得られるようにするためにハードウェアとソフトウェアの環境と試験設定を正規化する取り組みの中に含まれる。

PM RTP は、SNIA の Web サイトに掲載され、PM PTS に対する変更とソフトウェア上のより高度な商用プラットフォームのリリースを反映するようにときどき更新される。

X. 結論

永続メモリ・ストレージ製品は、メモリを CPU に近付ける高速で低遅延の永続ストレージを提供することによって、ストレージ階層内の隙間を埋めている。PM PTS の目的は、永続メモリと PM ストレージの概念に加えて、PM ストレージと PM 対応アプリケーションの使用、開発、および統合に関するソフトウェアとハードウェアの要件を紹介することである。

3D XPOINT と NVDIMM-N の早期導入は、高速で計算量の多いアプリケーションのための大きなインメモリ・データ・セット（AI 基幹システム向けのインメモリ・データベース、メタデータ、およびロギングなど）、高速な不揮発性ストレージ層、および簡単にアクセス可能なストレージ空間が必要なアプリケーションで見られる。

このホワイトペーパーの目的は、ストレージ・アーキテクト、ソフトウェア開発者、およびマーケティング・プロフェッショナルが、PM がストレージ階層のどこに位置付けられるかを理解し、新しい PM 対応アプリケーション用のテクニカル・ロードマップとマーケティング・ロードマップを定義できるように支援することである。

この面白い分野への関心のある当事者の参加を歓迎する。質問やコメントは、askcmsi@snia.org にお送りいただきたい。SNIA に入会した個人および企業は、SNIA の技術作業部会やイニシアチブに参加して、永続メモリ標準の作成と採用促進の支援、PM ストレージ・アーキテクチャの伝道、および PM 教材への貢献をすることができる。

XI. 執筆者と寄稿者について

本書を策定して審査した SNIA Solid State Storage Technical Working Group（SNIA ソリッド・ステート・ストレージ技術作業部会）は、以下の個人と企業から多大な貢献をいただいた。

- Eduardo Berrocal、Intel
- Jim Fister、The Decision Place
- Eden Kim、Calypso Systems, Inc.
- Chuck Paridon、dxc
- Andy Rudoff、Intel