SNIA®

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

NVMe® over Fabrics（NVMe-oF™）の最適化

人工的なワークロードと現実世界のワークロードを使用した、ホスト要因を伴うさまざまなトランスポートの性能の比較

ホワイトペーパー
2021 年 4 月

執筆者：
Eden Kim、Calypso Systems, Inc.
Fred Zhang、Intel Corp.

要約：
NVMe® over Fabrics（NVMe-oF™）の性能を 100Gb イーサネット経由で RDMA（iWARP と RoCEv2）トランスポートと TCP トランスポートを使用して比較する。人工的ワークロードと現実世界ワークロードを標準（1500B）MTU フレームとジャンボ（9000B）MTU フレームを使用して様々なファブリックを介してストレージ・ターゲットに適用する。6-SSD 3D XPoint™ LUN と 6-SSD 3D NAND LUN の性能を比較する。結果は、様々なワークロード（人工的コーナー・ケース・ワークロードと現実世界ワークロード）、RDMA トランスポート・メカニズムと TCP トランスポート・メカニズム（CPU オンロードと CPU オフロード）、および様々なタイプのストレージ LUN（3D XPoint と 3D NAND）の影響を示している。

SNIA CMSI | COMPUTE, MEMORY, AND STORAGE

SNIA NSF | NETWORKING STORAGE

目次

背景

SNIA/Brighttalk Web キャストに対するホワイトペーパー・ガイド兼アップデート

このホワイトペーパーは、2020 年 9 月 15 日に行われた SNIA/Brighttalk Web キャスト「Optimizing NVMe-oF Performance with different Transports: Host Factors」配信に対するガイド兼アップデートである。この Web キャストは、Illuminosi 社のプリンシパルである Tom Friend 氏が司会を務め、導入部分はニューハンプシャー大学の David Woolf 氏が担当した。Web キャスト・プレゼンターは、Intel Corp.の Fred Zhang 氏と Calypso Systems, Inc.の Eden Kim 氏であった。

Web キャストで提示された人工的な Random 4KB & Sequential 128KB Read/Write コーナー・ケース・ワークロードと現実世界の GPS 100% Write Navigation Portal ワークロードの他に、このホワイトペーパーでは、Retail Web 66% Read Portal と VDI 75% Write Storage Server Cluster という 2 つの現実世界ワークロードが追加される。

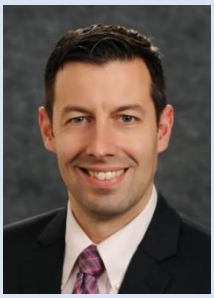
このホワイトペーパーは、SNIA NSF（Networking Storage Forum）、SNIA SSS TWG（Solid State Storage Technical Working Group）、および SNIA CMSI（Compute, Memory & Storage Initiative）の共同制作である。

以下をクリックすると、Web キャストを表示したり、プレゼンテーションをダウンロードしたり、Web キャストに関する質問と回答をダウンロードしたりできる。このホワイトペーパーに関する質問は、Fred.zhang@intel.com または edenkim@calypsotesters.com で受け付けている。

Tom Friend
Illuminosi

Fred Zhang
Intel Corp.

Eden Kim
Calypso Systems, Inc.



David Woolf
Univ. New Hampshire

SNIA リソース

SNIA リソース

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）は、ストレージ技術と情報技術を発展させるための標準と教育プログラムの開発を専門とする世界的非営利組織である。Compute, Memory & Storage Initiative（CMSI）の使命は、メモリやストレージと演算の融合と、次の 10 年間のデータ作成の急増を分析して活用するための新しい計算アーキテクチャおよびソフトウェアの作成を促進する業界をサポートすることである。Networking Storage Forum（NSF）の使命は、ストレージ・ネットワーキング・ソリューションの採用と認識を広めることである。

SNIA、CMSI、および NSF については、snia.org、http://www.snia.org/forums/cmsi、および www.snia.org/forums/nsf/technology を参照されたい。最新のホワイトペーパーは SNIA Educational Library で参照でき、ポッドキャストは snia.org/podcasts で視聴できる。SNIA 関連動画は SNIAVideo YouTube Channel でも視聴できる。

NVMe Web サイトについては、www.nvmexpress.org で参照できる。NVMe 仕様については、www.nvmexpress.org/developers/nvme-specification/で、NVMe-oF 仕様については、https://nvmexpress.org/developers/nvme-of-specification/で参照できる。

性能試験仕様（PTS）を含む SNIA の技術的研究については、https://www.snia.org/tech_activities/work で参照できる。

SNIA、CMSI、または NSF に関する追加情報については、https://www.snia.org/resources で参照するか、askcmsi@snia.org に電子メールで問い合わせいただきたい。

I. 要約

様々なファブリック・トランスポートが利用できるようになるにつれ、NVMe over Fabrics（NVMe-oF）ストレージの人気が高まっている。このホワイトペーパーでは、異なるイーサネット NVMe-oF トランスポート（RDMA（iWARP と RoCEv2）と TCP）の様々な性能特性を調査する。

この比較では、人工的コーナー・ケース・ワークロードと現実世界ワークロードの両方を適用する。試験 IO は、100GbE NVMe-oF ファブリックを介して Host Initiator サーバから Target Storage サーバに適用される。Target Server に直接取り付けられた 2 つのタイプのストレージ（6 ドライブ 3D XPoint および 6 ドライブ 3D NAND 論理ストレージ・ユニット（LUN））を試験する。試験のセットアップと条件は、性能比較に対するホスト要因の影響を分離するように正規化する。

次の 3 つの主要メトリクスを比較するために、性能分析が行われる。

- 最大伝送単位（MTU）フレーム・サイズ（標準フレームとジャンボ・フレーム）
- 3D XPoint 6-SSD LUN と 3D NAND 6-SSD LUN
- iWARP、RoCEv2、および TCP トランスポート経由の人工的ワークロードと現実世界ワークロード

人工的ワークロードは、Random 4K Read/Write（RND 4K R/W）IO と Sequential 128K Read/Write（SEQ 128K R/W）IO で構成される。現実世界ワークロードは、GPS Nav Portal（100% Write）、VDI Storage Cluster（75% Write）、および Retail Web Portal（65% Read）から取得される。

結果は、標準 1500B（バイト）フレーム MTU とジャンボ 9000B フレーム MTU の両方でかなり似通った性能を示す。3D XPoint ストレージ LUN は、3D NAND ストレージ LUN よりかなり高い性能を示す。CPU オフロードありの RDMA トランスポート（iWARP と RoCEv2）は、CPU オフロードなしの TCP より大幅に高い性能を示す。RoCEv2 は、iWARP よりわずかに低い IO 応答時間を示す。これらの試験はすべて指定されたハードウェアを使用して行われたが、別のネットワーク・カードを使用した場合は違う結果が生成される可能性がある。サードパーティ試験は、フル NVMe/TCP ハードウェア・オフロードを使用した場合に、NVMe/TCP で RDMA とほぼ同等の良好な性能と遅延になる可能性があることを示した。

iWARP と RoCEv2 は、65% Read Retail Web Portal ワークロードと 100% Write GPS Nav Portal ワークロードに対してより高い性能を示すが、TCP は、75% Write VDI Storage Cluster ワークロードに対してより高い性能を示す。この差は、それぞれのワークロードの IO ストリーム・コンテンツとブロック・サイズ・コンテンツの違いによるものと思われる（つまり、ワークロード IO ストリーム・コンテンツによって、このスタディで試験した様々なトランスポートおよび/またはストレージ LUN に対する性能上の影響が異なる）。

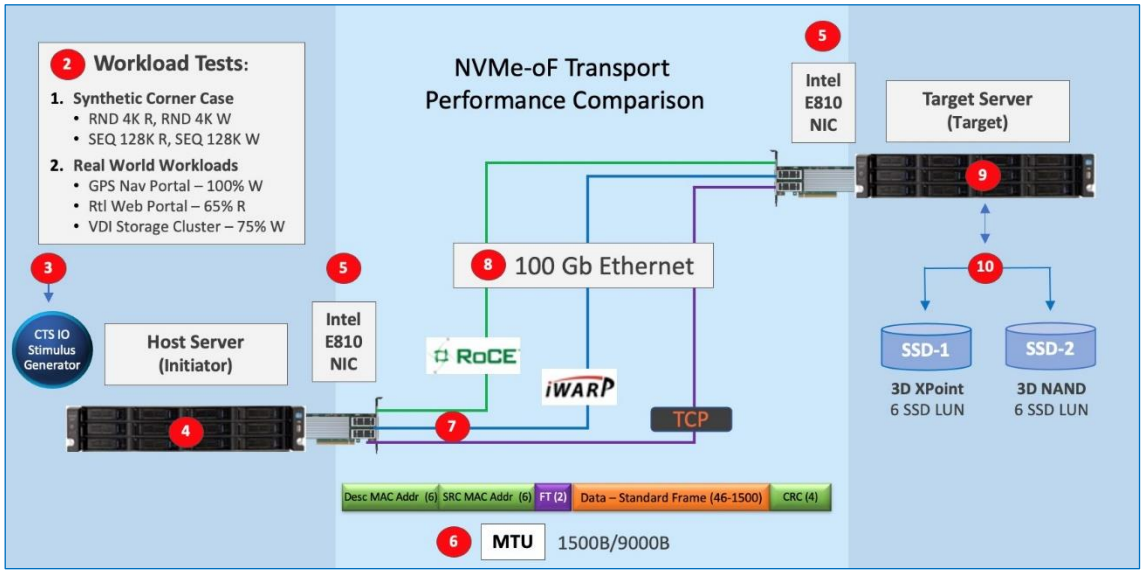


図 1 – セットアップと試験計画

II. はじめに – NVMe over Fabrics（NVMe-oF）

A. NVMe-oF：概要

NVM Express（NVMe）は、PCIe ベースのソリッド・ステート・ドライブ（SSD）用の標準ホスト・コントローラ・インタフェースである。NVMe over Fabrics（NVMe-oF）仕様では、RDMA やファイバ・チャネルなどの相互接続経由で NVMe コマンドを伝送可能にするプロトコル・インタフェースと関連拡張機能が規定されている。また、NVMe-oF は、NVMe ストレージをスケールアウトするために、NVMe の展開をローカル・ホストからリモート・ホストに拡張する。

B. NVMe-oF トランスポート： 相違点

NVMe over Fabrics 用のイーサネットベースのトランスポートには、iWARP RDMA、RoCEv2 RDMA、および TCP の 3 つがある。

リモート・ダイレクト・メモリ・アクセス（RDMA）

リモート・ダイレクト・メモリ・アクセス（RDMA）は、ネットワーク経由でアプリケーション間の低遅延で高スループットのダイレクト・メモリ間データ通信を可能にするホストオフロード・ホストバイパス技術である（RFC 5040 A Remote Direct Memory Access Protocol Specification）。一般的に、RDMA は、ネットワーク・ハードウェア・オフロードを利用して、通常はネットワーク機能専用となるサーバ・リソースを削減する。NVMe-oF 用のイーサネット上の RDMA の実装には、主に、 iWARP と RoCEv2 の 2 つがある。

ベスト・エフォート・ネットワークとロスレス・ネットワークの比較

ベスト・エフォート・ネットワークは、データが到着すること、順番を維持して到着すること、そして整合性を失わずに到着することを保証しないネットワークである。インターネット・プロトコル（IP）ネットワーク層が、ベスト・エフォート・ネットワークの例である。一般的に、IP ネットワークは、信頼できるデータ配信を実現するための追加のメカニズムを提供するために、上位プロトコル（TCP と呼ばれる伝送制御プロトコルなど）に依存する。このようなメカニズムには、フロー制御や輻輳管理を含めることができるが、これらに限定されない。

「ノードロップ」ネットワークとも呼ばれるロスレス・ネットワークは、信頼性が高く、どのパケットもドロップされないことを保証するように設計されているため、そのように呼ばれている。一方、ベスト・エフォート・ネットワークは、定義上、配信を保証できないため、パケットが消失した場合は再送信が必要になる。

ロスレス・ネットワークは、TCP over IP（TCP/IP）などのように、ベスト・エフォート・ネットワーク上に構築できる。UDP over IP（User Datagram Protocol over Internet Protocol、つまり UDP/IP – 下記参照）は、フロー制御と輻輳管理を備えておらず、*保証された*配信も提供しないため、パケットのドロップを回避するためには、追加のイーサネット・ネットワーク構成またはメカニズムが必要である。このような追加の構成には、レイヤ 2 のプライオリティ フロー制御（PFC）（https://www.ieee802.org/1/pages/dcbridges.html）またはレイヤ 3 の差別化サービス・コード・ポイント（DSCP）PFC が含まれる（RFC 2474 IPv4 ヘッダと IPv6 headers ヘッダ内の差別化サービス・フィールドの定義、RFC 2475 差別化サービスのアーキテクチャ）。

iWARP

iWARP は、広く普及している TCP over IP（TCP/IP）プロトコル上に RDMA を実装するコンピュータ・ネットワーキング・プロトコルである （「iWARP」は頭字語でないことに注意されたい）。iWARP は、伝送制御プロトコル（TCP）やストリーム制御伝送プロトコル（SCTP）などの インターネット・エンジニアリング・タスク・フォース（IETF）標準輻輳認識プロトコルの上で動作する。そのため、iWARP RDMA は、標準のネットワーク層とトランスポート層で動作し、TCP/IP をサポートするすべてのイーサネット・ネットワーク・インフラストラクチャで機能する。

TCP は信頼できる配信を提供するため、信頼できない IP ネットワーク上で、上位アプリケーションに信頼できるネットワーク・サービスを提供できる。iWARP は、ネットワーク・アダプタ上の低遅延ハードウェア・オフロード・エンジンとも呼ばれているが、このようなオフロードには iWARP 対応ネットワーク・アダプタが必要である。iWARP では、ネットワークアダプターに搭載された低遅延のハードウェア

オフロードエンジンが知られている。このオフロードには、iWARP に対応したネットワークアダプターが必要である。

RoCE（RDMA over Converged Ethernet）

2009 年に InfiniBand Trade Association（IBTA）によって開発された RoCEv1（RDMA over Converged Ethernet）は、イーサネット・データ・リンク層と物理層を使用して、InfiniBand（IB）トランスポートおよびネットワーク層をサポートする（Annex A16 RoCE、InfiniBand Architecture Specification Volume 1 Release 1.2.1 に対する補足）。RoCEv2 は、さらに、UDP/IP 上で動作するように改良された（Annex A17 RoCEv2、InfiniBand Architecture Specification Volume 1 Release 1.2.1 に対する補足）。RoCEv2 は、低遅延も提供する。下の図 2 を参照されたい。今日、RoCE のほぼすべての実装で、RoCEv2 が使用されており、用語の「RoCE」は一般的に RoCEv2 を意味する。

ユーザ・データグラム・プロトコル（UDP）は、特に、インターネット上の動画再生や DNS（ドメイン・ネーム・システム）ルックアップなどの時間的制約のある伝送に使用される通信プロトコルである。データを伝送する前に接続を正式に確立しないことによって通信を高速化する。これにより、データを非常に高速に伝送できる。「このプロトコルはトランザクション指向である。送達も重複の防止も保証されない。 順序が正確で信頼できるデータのストリームの送達を必要とするアプリケーションは、伝送制御プロトコル（TCP）を使用するべきである」（RFC768 ユーザ・データグラム・プロトコル）。

UDP はフロー制御や輻輳管理を提供せず、RoCEv2 は UDP 上で動作するため、一般的に、RoCEv2 は、ロスレス・イーサネットを必要とし、プライオリティ・フロー制御（PFC）または Explicit Congestion Notification（ECN、RFC3168）などの輻輳管理ソリューションの使用に依存して、ネットワーク輻輳時のパケット損失を最小限に抑える。RoCEv2 は、1 つのデータセンター内の展開に最適である。また、RoCEv2 は、ハードウェア・オフロード用に RoCE 対応 RDMA ネットワーク・アダプタ（*RNIC*）も必要とする。PFC や ECN を必要としない高速な RoCE 性能を実現可能な RoCE 対応 RNIC が存在するが、この機能はベンダー固有であり、様々なベンダー製の RoCE 対応 NIC で動作しない可能性がある。

TCP

TCP、つまり、伝送制御プロトコルは、ネットワーク経由でアプリケーション・データを交換する場合にネットワーク通信を確立して維持する方法を定義した、広く受け入れられている標準である。TCP は、各パケットを正しい宛先に届けるためにアドレス指定してルーティングする方法を決定するインターネット・プロトコル（IP）と一体となって動作する。

NVMe over TCP（NVMe/TCP）が NVMe-oF Specification v1.1 に追加された。NVMe/TCP は、標準の TCP を NVMe-oF 用のトランスポートとして使用するため、追加の特定の要求仕様なしで任意のイーサネット・ネットワーク・アダプタを使用でき、ネットワーク構成を変更したり特別な機器を実装したりする必要もない。NVMe-oF での TCP トランスポート・バインディングによって、2 つのホスト間で通常の TCP 接続を使用してデータをカプセル化して配信するための方法が定義される。

ただし、NVMe/TCP には欠点もある。例えば、TCP は、TCP オフロードを実行するためのアダプタがない場合、プロトコル・スタックの処理がホストの CPU と OS に依存するため、余分なホスト・システムの処理能力が必要になる可能性があることから、仕様に応じてシステム・プロセッサの負荷が増える可能性がある。

また、NVMe/TCP は、追加のデータのコピーを TCP スタック内で保持する必要があるため、遅延（応答時間）が大きくなる可能性もある。この遅延の程度は、仕様の実装方法とサポートされているワークロードのタイプに依存し、TCP オフロードをサポートするネットワーク・アダプタを使用して削減できる可能性がある。

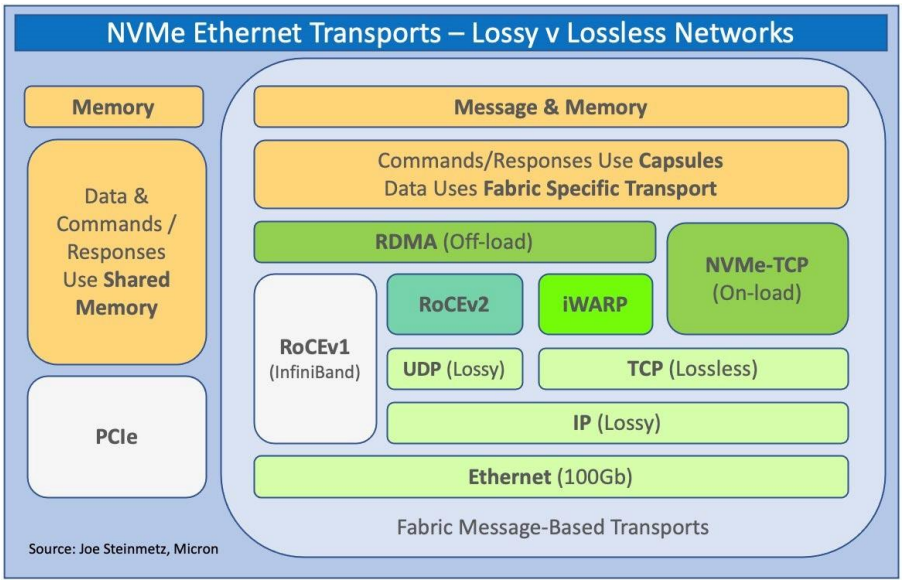


図 2 – NVMe イーサネット・トランスポート：UDP と TCP の比較

C. RoCEv2 と iWARP の比較 – UDP と TCP の比較

iWARP と TCP は、RoCE よりパケット損失に対して耐性がある。iWARP は、フロー制御と輻輳管理を備えた TCP/IP アーキテクチャをベースにしている。TCP のお陰で、パケット損失が発生する場合でも、iWARP は、選択的再伝送と順序から外れたパケットの受信をサポートする。これらの技術は、ベスト・エフォート・ネットワークの性能をさらに高める可能性がある。

RoCEv2 の標準実装にはパケット損失から回復するためのメカニズムが組み込まれているが、伝統的に、「ロスレス」ネットワークが推奨されている。これは、パケット損失が発生すると性能が低下するためである。これを回避するために、RoCE は、レイヤ 2 の IEEE データセンター・ブリッジング拡張機能（特に、プライオリティ・フロー制御）および/またはレイヤ 3 の ECN を使用して、パケット損失を最小限に抑え、順序を維持した送達を保証している。

D. NVMe-oF：成熟度

NVMe-oF v1.0 仕様は、2016 年 6 月にリリースされ、2019 年 10 月に一部の改良と新しいトランスポートとしての TCP の追加を含む v1.1 に改訂された。 今では、NVMe-oF をサポートする多くのイーサネット製品が市販されている。

OS エコシステムに、堅牢なドライバ・サポートが存在する。NVMe-oF の Linux ドライバは、Initiator 用と Target 用の両方に用意されている。VMware は NVMe-oF イニシエータを備えている。Microsoft Windows 用の NVMe-oF イニシエータを提供しているサードパーティもある。

University of New Hampshire Inter-Operability Lab（UNH-IOL）では、複数のイーサネット製品ベンダーの様々なトランスポートの相互運用性試験と適合性試験を体系化している。

III. 様々なイーサネット・トランスポートの性能に影響する要因

A. 考察の範囲

ホスト、スイッチ、ネットワークなど、NVMe-oF の性能に影響する要因は多数存在する。このホワイトペーパーでは、ホスト要因に焦点を当てる。CPU のオフロード技術とオンロード技術（ソフトウェアベース）、様々な NVMe ドライブの属性とその性能に対する影響、および RDMA と TCP の性能分析における最大伝送単位（MTU）フレーム・サイズ（1500B と 9000B）を考慮する。そのため、試験では、ネットワーク（スイッチなど）の構成、設定、トポロジ、ベスト・プラクティスを試験変数として考慮しない。

ホスト

ホスト・サーバでは、CPU とメモリの構成が、NVMe-oF（特にソフトウェアベースのソリューションのホスト OS プロトコル・スタックに依存する NVMe/TCP）の性能に影響する。NVMe ドライブの属性も NVMe-oF の性能に影響する。他に性能ボトルネックがない場合でも、NVMe ドライブの性能は、多くのワークロード・シナリオで見られる IO R/W 構成、伝送サイズ、および遅延属性がボトルネックとなる可能性がある。

スイッチ

スイッチ設定は、NVMe-oF の全体性能に影響する可能性がある。NVMe-oF の性能は、バッファリング、オーバーサブスクリプション、専用トラフィック・クラスのセットアップ、および NVMe-oF 用の輻輳制御メカニズムの影響を大きく受ける可能性がある。このことは、特に、RoCE が高性能のサポートをロスレス・ネットワークに依存していることから、NVMe over RoCE に当てはまる。前述したように、このホワイトペーパーと対応する試験結果の目的は、様々なスイッチの条件やベスト・プラクティスに関するベスト・プラクティスを提案することではない。

ネットワーク

ネットワーク・トポロジは考慮すべきもう 1 つの要因である。性能に関する考慮事項として、ターゲット・ストレージの帯域幅オーバーサブスクリプション、Initiator と Target が必要とするファンイン比(fan-in ratios)、サービス品質の設定、サービス・クラスの構成、およびその他の様々な条件といった要因が含まれる。前述したように、このホワイトペーパーと対応する試験結果の目的は、様々なネットワークの条件やベスト・プラクティスに関するベスト・プラクティスを提案することではない。

B. オンロードとオフロード の比較

RDMA は、結果的に CPU 使用率が下がるホスト・バイパス・オフロード技術である。NVMe over RDMA では、RDMA ネットワーク・インタフェース・カード（RNIC）上の RDMA エンジンが、オペレーティング・システム（OS）のプロトコル・スタックをバイパスして、ダイレクト・リモート・メモリ間データ・アクセスを使用できる。

従来の TCP は、OS カーネルのプロトコル・スタックに依存している。1Gb や 10Gb のイーサネットでは、CPU 使用率はそれほど大きくないかもしれないが、ネットワーク速度が 100Gb になると、CPU 使用率が著しく上昇する。その結果、ソフトウェアベースの NVMe/TCP は、カーネルに対する依存度が高いため、同じワークロードでも RDMA より多くの CPU サイクルを消費することになる。

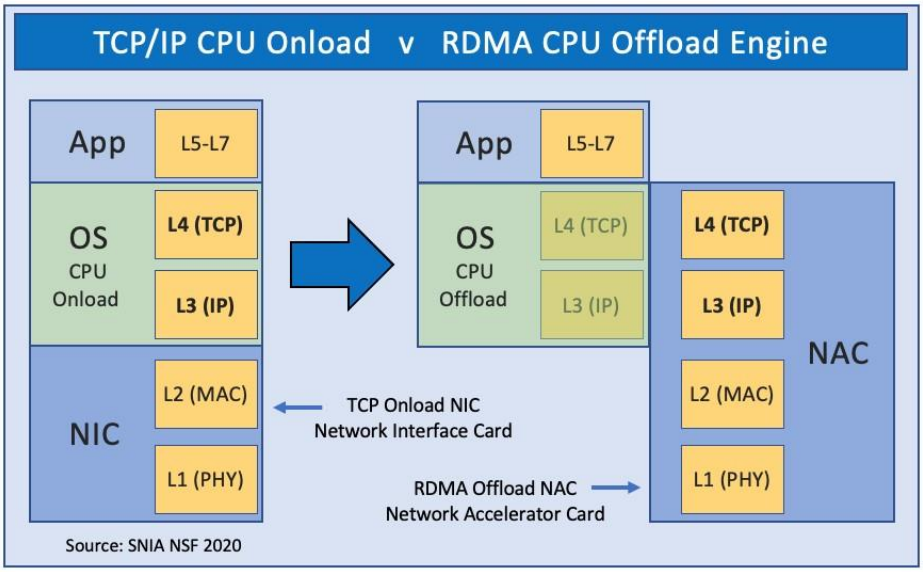


図 3 – オンロード・エンジンとオフロード・エンジンの比較

ネットワーク・アダプタ上に完全な TCP オフロード・エンジンを置けば、低い CPU 使用率で高い性能を実現できる。

注： 他にも、高スループットと低遅延を実現するために、ユーザ空間で機能し、専用の CPU コアを使用してポーリング・モードで動作する技術（Storage Performance Development Kit（SPDK）、（spdk.io）など）がある。

C. MTU：1500B と 9000B の比較

最大伝送単位（MTU）は、インターネット・プロトコル層において特定の媒体上で分割せずに伝送可能なパケットの最大サイズである。イーサネット・フレームには、18 バイトまたは 18 バイトに IEEE 802.1Q タグ用を追加した 22 バイトのフィールドが含まれており、ジャンボ・フレームがサポートされている場合は、イーサネット・フレームを最大 9000 バイト（9KB）にすることができる。MTU サイズが大きいほど、IO ワークロードが高い場合の CPU 使用率と帯域幅が改善されるが、遅延が増大する可能性もある。また、イーサネット・ジャンボ・フレームを使用するには、ネットワーク内のすべてのサーバ、スイッチ、およびオプション・ルータ上のジャンボ・フレーム設定を有効にして適切な動作を保証する必要もある。

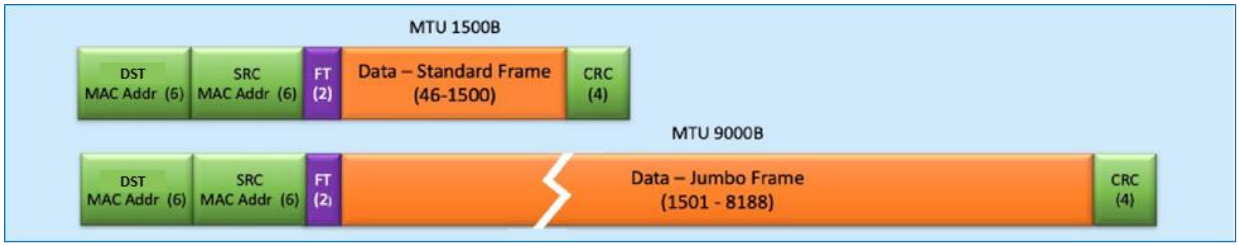


図 4 – MTU フレーム・サイズ： 標準 1500B とジャンボ 9000B の比較

D. 個々のドライブ・レベルの要因

NVMe-oF は、特に、ダイレクト・アタッチド・ストレージ（DAS）の場合に、基礎となる NVMe ドライブの性能に大きく依存している。Target Initiator 上にストレージ・ヘッド・ノードがある他のストレージ・システム（ネットワーク・アタッチド・ストレージ（NAS）デバイスなど）では、性能の議論に新たに抽象化層が追加される。ポリシー設定、サービス品質、様々な消失訂正符号戦略と RAID 戦略などの要因が NVMe ドライブの性能と結び付いて、NVMe-oF の全体性能に影響を及ぼす。

現実世界ワークロード IO パターンも全く異なる。ワークロードは、読み取り中心、書き込み中心、または読み取り（R）と書き込み（W）の何らかの組み合わせにすることができる。 特定のデータ・サイズのランダム（RND）またはシーケンシャル（SEQ）R または W 伝送である IO ストリームは、需要強度（DI）、つまり、未処理 IO（OIO）の影響も受ける。 ここでは、DI、OIO、およびキュー深さ（QD）を同じ意味で使用する。

NVMe ドライブはこれらの要因に応じて異なる動作をするため、基礎となる SSD のタイプも考慮する必要がある。したがって、NVMe ドライブは、最高性能を実現するために想定される IO ストリーム・コンテンツと需要強度に基づいて選択する必要がある。IO 特性や目標性能範囲が異なる様々な NVMe ドライブが設計されている。下の図 5 に、このケース・スタディで使用された RND 4K R/W および SEQ 128K R または W の性能に関する NVMe SSD メーカーの仕様を示す。

個々のドライブ・レベルの特性は、ストレージ空間からファブリック空間までとその逆の抽象化層ごとにマスクまたは変更できる。例えば、SSD の前方で大容量キャッシュを使用すると、別のタイプの SSD を使用した場合との観察性能差が縮まる可能性がある。そのため、SSD レベルの要因は、観察対象のホスト・レベルの性能に想定された影響を及ぼさない可能性がある。SSD レベルの要因の例を以下に示す。

読み取り ー 書き込み（RW）構成。少量の書き込み IO は、混在 RW ワークロード性能に偏って影響する可能性がある。また、「読み取り中心または書き込み中心」ワークロード用に設計されたドライブは、実際のアプリケーション生成ワークロードとは全く異なる IO ストリーム・コンテンツに基づいている可能性がある。

ブロック・サイズ／アクセス。小ブロック RND サイズと大ブロック SEQ IO サイズでは、性能が異なる可能性がある。

IO ストリーム。混在 IO ストリームで構成された現実世界ワークロードは、単一の IO ストリームと固定需要強度（DI）で構成された人工的ワークロードとは異なる形で性能に影響する可能性がある。

需要強度（DI）の飽和。低い DI は IOPS を無駄にする一方で応答時間（RT）を減少させる可能性があるが、高い DI は IOPS と応答時間の両方を増加させる可能性がある。

ストレージ容量。少ない SSD 容量は、飽和状態になって、ガベージ・コレクションや RT スパイクを誘発する可能性がある。

IO データ・パスのボトルネック。RT は、IO データ・パス内の各コンポーネントの影響を受ける可能性があり、RT ボトルネックの根本原因の分離は困難である（図 18 ー 応答時間のスパイクを参照）。

個々のドライブ・メーカーの仕様 – 3D XPoint SSD と 3D NAND SSD の比較

下の図 5 に、このスタディで使用された SSD に関するメーカーの仕様を示す。3D XPoint SSD は対称の RND 4K および SEQ 128K R/W の性能を示すが、3D NAND SSD は非対称の RND 4K R/W の性能を示す。3D XPoint SSD は、3D NAND SSD より、RND W 性能が高く、容量が少ない。3D NAND SSD は、3D XPoint SSD より、SEQ R/W 性能が高く、容量が多い。

メーカー仕様の最適なキュー深さ（QD）の範囲は、3D XPoint（QD=16）の場合は 3D NAND（QD=256）の場合より低いことに注意されたい。 これは、ドライブには任意の数の QD ジョブからアクセスできるが、最良の（最適な）QD と関連性能は SSD メーカーの仕様で規定されていることを意味する。

Manufacturer Spec	RND 4K R	RND 4K W	SEQ 128K R	SEQ 128K W
SSD-1: 3D XPoint (6) x 375 GB SSD	550,000 IOPS QD 16	550,000 IOPS QD 16	2,500 MB/s QD 16	2,200 MB/s QD 16
SSD-2: 3D NAND (6) x 4.0 TB SSD	636,500 IOPS QD 256	111,500 IOPS QD 256	3,000 MB/s QD 256	2,900 MB/s QD 256

図 5 ー SSD 特性 – Mfgr SSD の仕様

IV. 試験比較：iWARP、RoCEv2、および TCP の比較

A. 試験計画

目的

iWARP、RoCEv2、および TCP。このスタディの主要試験目標は、100Gb イーサネット経由の iWARP、RoCEv2、および TCP トランスポート・プロトコルの性能を比較することである。CPU オフロードを使用した RDMA トランスポート（iWARP と RoCEv2）の影響を評価し、オフロードを使用しない（つまり、CPU オンロード）従来の TCP トランスポートと比較する。前述したように、ネットワーク・トポロジ、QoS、スイッチ構成、その他のネットワーク・ベスト・プラクティスなどのソリューション性能に関する追加の重要要因は考慮しない。

ワークロード。人工的コーナー・ケース・ワークロード（RND 4K RW および SEQ 128K RW）の性能を、3 つの現実世界ワークロード（GPS Nav Portal、Retail Web Portal、および VDI Storage Cluster）の性能と比較する。小ブロック（4K）RND および大ブロック（128K）SEQ コーナー・ケース・ワークロードに対する人工的ワークロードの RW 比率の影響を観察する。また、異なる RW 比率（100% W、66% R、75% W）の複数の IO ストリームの現実世界ワークロードの影響を評価する。さらに、現実世界ワークロード IO キャプチャでの IO ストリームの組み合わせの様々なシーケンスとキュー深さ（QD）の影響も評価する。

MTU。標準（1500B）MTU フレーム・サイズとジャンボ（9000B）MTU フレーム・サイズの性能を比較する。

ストレージ LUN。6 ドライブ 3D XPoint ストレージ LUN と 6 ドライブ 3D NAND SSD ストレージ LUN（RAID 0）の性能差を測定する。また、IO アクセス・パターン（IO ストリーム・サイズ、RW 比率、RND または SEQ アクセス）、様々なワークロード・タイプを処理するためのストレージ設計の能力、および IO で 100Gb イーサネットを飽和状態にする LUN の能力を考慮する。

Initiator と Target Server にわたるホスト要因

100Gb イーサネット経由のホスト要因の性能への影響を分離するための Host Initiator と Target Storage サーバの構成と設定の正規化を試みる（後述する試験セットアップを参照）。これらのホスト要因には、以下の項目などが含まれる。

- RDMA トランスポート（iWARP と RoCEv2）の CPU オフロードと CPU オンロード・トランスポート（TCP）
- MTU フレーム・サイズ（標準とジャンボ）
- 試験ワークロードの組成と設定（後述する試験ワークロードと試験方法を参照）
- IO ストリーム・コンテンツ（IO 伝送サイズ、RW 比率、RND または SEQ アクセス、および QD）
- 未処理 IO（OIO）／需要強度（DI） ― スレッド・カウント（TC） x QD で測定
- IO サイズ（伝送サイズまたはブロック・サイズ）
- 多様な需要強度における性能

試験トポロジ

「バックツーバック試験トポロジ」では、試験 IO を NVMe-oF トランスポート経由でホストからスイッチを使用せずにターゲットに適用する（下の図 6 を参照）。試験ワークロード（2）は、Host Initiator サーバ（4）上にマウントされた Calypso Test Suite（CTS）IO Stimulus ジェネレータ（3）から生成される。CTS IO Stimulus ジェネレータは、論理ストレージの直接、リモート、またはファブリック試験を可能にする Calypso IO エンジンである。CTS IO Stimulus ジェネレータ（3）は、48 ビット乱数発生器を利用して、非反復ランダム・バイナリ・データを試験スレッドとキューにロードする Linux ベースの libaio（非同期 IO）である。

試験スクリプトは、CTS Control Server データベース（1、2、3）から生成される。人工的ワークロードと現実世界ワークロードの IO ストリーム・コマンド、試験設定、および試験手順（2）は、CTS データベース（1）から収集され、Host Initiator サーバ（4）上に存在する CTS IO Stimulus ジェネレータ（3）に送信される。

その後で、CTS IO Stimulus ジェネレータ（3）が、論理ストレージ（10）に対する試験スクリプト IO コマンドを Intel E810-CQDA2 NIC（5）を介して 100Gb イーサネット・ファブリック（7、8）経由でターゲット・サーバ（9）と試験ストレージ LUN（10）に送信する。試験結果データのパケットは、CTS Control Server データベース（1）に送信され、アーカイブされ、後で、表示、レポート作成、およびデータ分析に使用される。

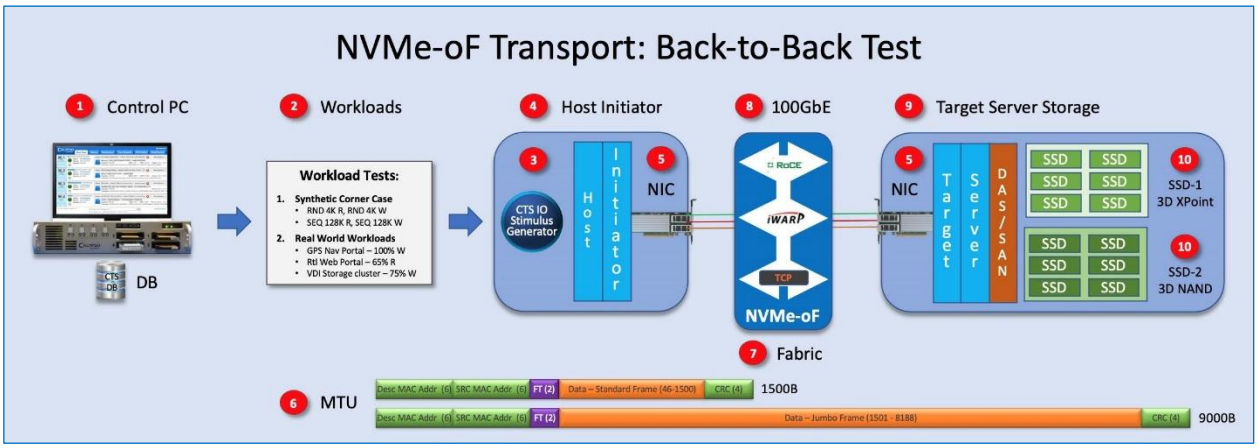


図 6— バックツーバック試験トポロジ

B. 試験ワークロード

NVMe-oF 試験は、人工的コーナー・ケース・ベンチマークと現実世界アプリケーション・ワークロードの両方を 3D XPoint SSD LUN と 3D NAND SSD LUN に適用する。

人工的コーナー・ケース・ワークロード

人工的コーナー・ケース・ベンチマーク試験は、固定された一定のワークロードをターゲット・ストレージに適用する。コーナー・ケース試験は、ターゲット・ストレージを飽和状態にして、通常動作の範囲外の性能を特定し、IO 性能をメーカー仕様に照らして比較するように設計されている。

一般的に、「全域」ベンチマーク試験は、小ブロック RND RW（RND 4K）と大ブロック SEQ RW（SEQ 128K）である。通常は、各 IO ストリームが、OIO = 1 (T1/Q1)や OIO = 128 (T4/Q32)などの固定未処理 IO（OIO）で 1 つずつ別々に定常状態に適用される。

現実世界ワークロード

現実世界アプリケーション・ワークロードは、現実世界ワークロード IO キャプチャから観察されたように、IO ストリームとキュー深さ（QD）の様々な組み合わせおよびシーケンスをストレージに適用する。各現実世界ワークロードには、それぞれ異なる IO ストリームと QD の組成とシーケンスが割り当てられる。これらの IO ストリームと QD は、特異なブロック・サイズと様々な未処理 IO（OIO）の絶えず変化する組み合わせである。現実世界アプリケーション・ワークロード試験の目的は、現実世界の展開におけるアプリケーションとストレージの使用中に観察されるものと同様の IO ワークロードに対するストレージの性能を評価することである。後に示す図 7：現実世界ワークロード比較表を参照されたい。

IO ストリーム・マップを使用した現実世界ワークロードの可視化

IO ストリーム・マップは、現実世界ワークロードの IO ストリーム、キュー深さ、および IO メトリクスを表現するために使用される。IO ストリーム・マップは、現実世界のアプリケーションが実行されたときにソフトウェア・スタックの特定のレベル（ファイル・システムやブロック IO など）で発生する IO ストリームの IO キャプチャから導出される。IO ストリーム統計は、特定の時間ステップ、つまり、観察の時間間隔で平均化される。その後で、これらの IO ステップが、IO キャプチャ中に発生した IO ストリーム、メトリクス、およびイベント（プロセス ID）を示す IO ストリーム・マップを作成するために使用される。IO キャプチャ時間ステップ統計を使用すれば、巨大なデータ・セットが関連付けられることなく、長時間のキャプチャの表示が可能になる。現実世界アプリケーション・ワークロード試験は、IO キャプチャから導出された IO ストリーム統計を利用する。

現実世界ワークロード比較表

図 7 は、このスタディで使用された 3 つの現実世界ワークロードを示している。すべての現実世界ワークロードが一意であるが、一般的に、ワークロードは、その全体的な RW 比率、選択されたまたは発生中の IO ストリーム、合計 IO、キャプチャ中に発生した合計 IO ストリーム、および QD の範囲（最小、最大、および中央）で特徴付けられる。

各ワークロードは、IO キャプチャの全体的な RW 比率、IO キャプチャ・レベル（ファイル・システムやブロック IO）、観察された合計 IO、観察された合計 IO ストリーム、IO の割合の最も発生頻度の高い 9 つの IO ストリーム、およびワークロードの最小、最大、および中央 QD を示す。

複数のドライブの IO キャプチャ（2 ドライブ 24 時間や 6 ドライブ 12 時間など）は、各ストレージ・デバイスの IO ストリームとメトリクスが単一の合成 IO ストリーム・マップと関連するメトリクスおよび統計に統合されることを意味する。複数ドライブ合成 IO ストリーム・マップの統合は、CTS IOProfiler ツールセットの機能である。

Real-World Workload	RW Mix Normalized	IO Capture Level	Total IOs	Total IO Streams	9 Most Frequent IO Streams by % of IOs						Min QD	Max QD	Median QD
Retail Web Portal: 2-Drive, 24-hour	65% R	Block IO	4.5 M	5,086	18.5% 10.0% 4.0% 3.4% 2.7%	RND RND RND SEQ RND	64K R 8K R 4K W 64K R 8K W	17.0% 8.4% 3.7% 2.9%	SEQ SEQ SEQ RND	0.5K W 8K R 64K W 4K R	5	306	19
GPS Nav Portal: 1-Drive, 24-hour	100% W	Block IO	3.5 M	1,033	21.6% 11.7% 9.6% 3.4% 2.1%	SEQ SEQ RND RND SEQ	4K W 0.5K W 4K W 8K W 1.5K W	12.0% 10.7% 4.9% 2.4%	RND SEQ RND RND	16K W 16K W 8K W 2K W	6	368	8
VDI Storage Cluster: 6-Drive, 12-hour	75% W	Block IO	167 M	1,223	19.3% 9.1% 4.2% 3.3% 2.3%	RND SEQ SEQ SEQ SEQ	4K R 4K R 128K R 4K W 8K R	11.3% 8.2% 3.6% 3.3%	RND SEQ RND RND	4K W 32K R 32K R 8K R	64	1024	128
SNIA CMSI Reference Workloads – Retail Web Portal, GPS Nav Portal and VDI Storage Cluster workloads can be viewed at www.testmyworkload.com													

図 7 ー 現実世界ワークロード： 比較表

Retail Web Portal（図 8 を参照）は、様々な小売 SQL Server イベント（朝のブート・ストーム、開店、日常活動、閉店活動、午前 2 時のバックアップなど）、様々な IO ストリームの混在、0.5K～64K のブロック・サイズの範囲、および 65% R IO の正規化 RW 比率で構成された 2 ドライブ 24 時間ワークロードとして特徴付けられる。

GPS Nav Portal（図 10 を参照）は、GPS ナビゲーションに関連する同質な IO 活動、小ブロック・サイズ IO ストリーム、定期的な SEQ 0.5K IO スパイクの発生、および 100% W IO の正規化 RW 比率で構成された 1 ドライブ 24 時間ワークロードとして特徴付けられる。

VDI Storage Cluster（図 12 を参照）は、従来のストレージ・ブロック・サイズ（断片化されていないブロック・サイズ）、QD が高い IO ストリーム（最大 1,024）、および 75% W IO の正規化 RW 比率で構成された 6 ドライブ RAID 0 LUN 12 時間ワークロードとして特徴付けられる。

Retail Web Portal

下の図 8 は、1 日の様々な活動と IO ストリームで構成された 2 ドライブ 24 時間 66% Read Retail Web Portal の IO ストリーム・マップを示している。x 軸は時間を表し、24 時間 IO キャプチャにおける時間と主要イベントの両方が示されている。各データ点は、IO 統計が平均化される 5 分の時間ステップを表す。y 軸は、IO メトリクスが関連付けられた *IO*、*IOPS*、および *IO ストリーム*を示す。

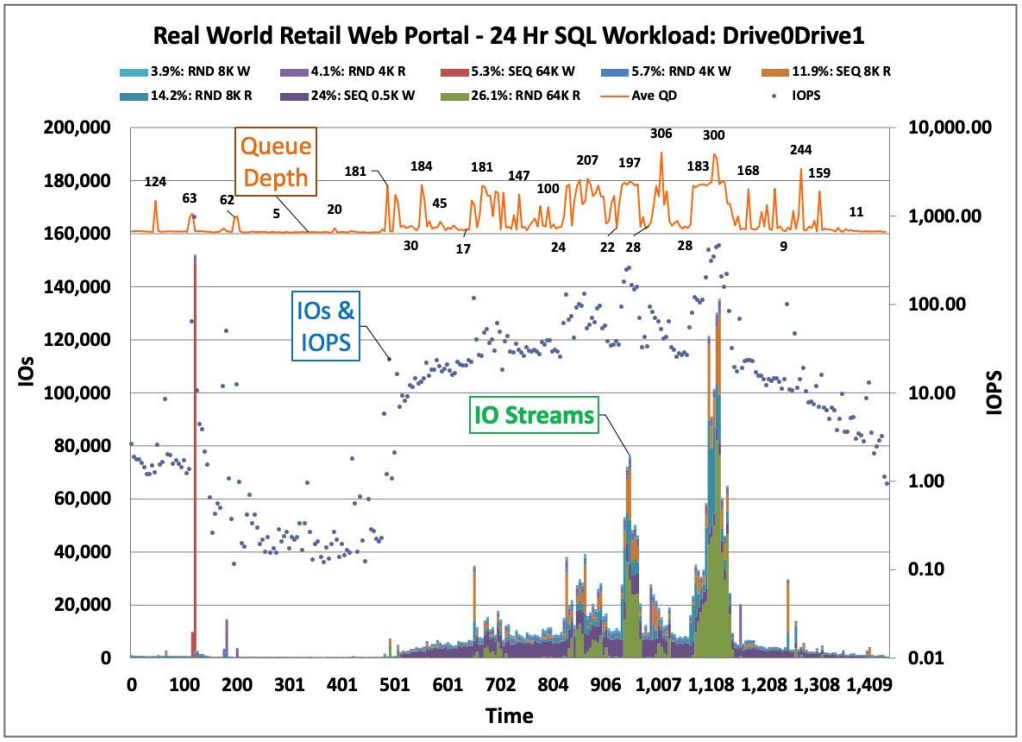


図 8 – IO ストリーム・マップ： Retail Web Portal

積み上げ縦棒の各色は、RND/SEQ アクセス、ブロック・サイズ、および RW 比率の個別の IO ストリームを表している。オレンジ色の線は、IO キャプチャの各ステップの平均 QD を示しており、QD=7～QD=306 の範囲で、中央値が QD=19 である。青色のドットは、24 時間キャプチャ中に発生した IO と IOPS である。

図 9 - Rtl Web 9 IO ストリーム

注： 午前 2 時のバックアップ用の赤色の SEQ 64K IO ストリーム・スパイク、使用が限定的な早朝の時間帯の低レベルの青色の IOP と IO、朝のブート・ストーム中の紫色の SEQ 0.5K W 上位 IO ストリーム、および日常の取引中や活動中に発生した混在 IO ストリーム、ピーク IO、およびピーク IOPS。

図 9 は、IO ストリーム・マップに表示するために選択された Retail Web Portal の 9 つの IO ストリームの累積ワークロードを示している。全体的に 66:34 の RW 比率で 24 時間 IO キャプチャ中に合計 5,086 個の IO ストリームが観察された。最も発生頻度の高い 9 つの IO ストリームが、発生した合計 450 万個の IO の 71%を占める。

9 つの IO ストリームのワークロードの正規化 RW 比率は 65:35 RW である。つまり、全 5,086 個の IO ストリームによる全体ワークロードの RW 比率が 66:34 であることに対して、選択された 9 つの IO ストリームのみに基づくと読み取りが 65%である。

IO 発生割合で上位 4 つの IO ストリームは、RND 64K R（18.5%）、SEQ 0.5K W（17%）、RND 8K R（10%）、および SEQ 8K R（8.4%）である。現実世界ワークロードごとの主な特徴を図 7：現実世界ワークロード比較表にまとめる。

GPS Navigation Portal

下の図 10 と図 11 は、24 時間 GPS Navigation Portal の 1 ドライブ IO ストリーム・マップと 9 つの IO ストリームのワークロードを示している。この IO ストリーム・マップは、上の図 8 の複数イベントベースの Retail Web Portal より同質的な IO ストリームの組成と発生を示している。

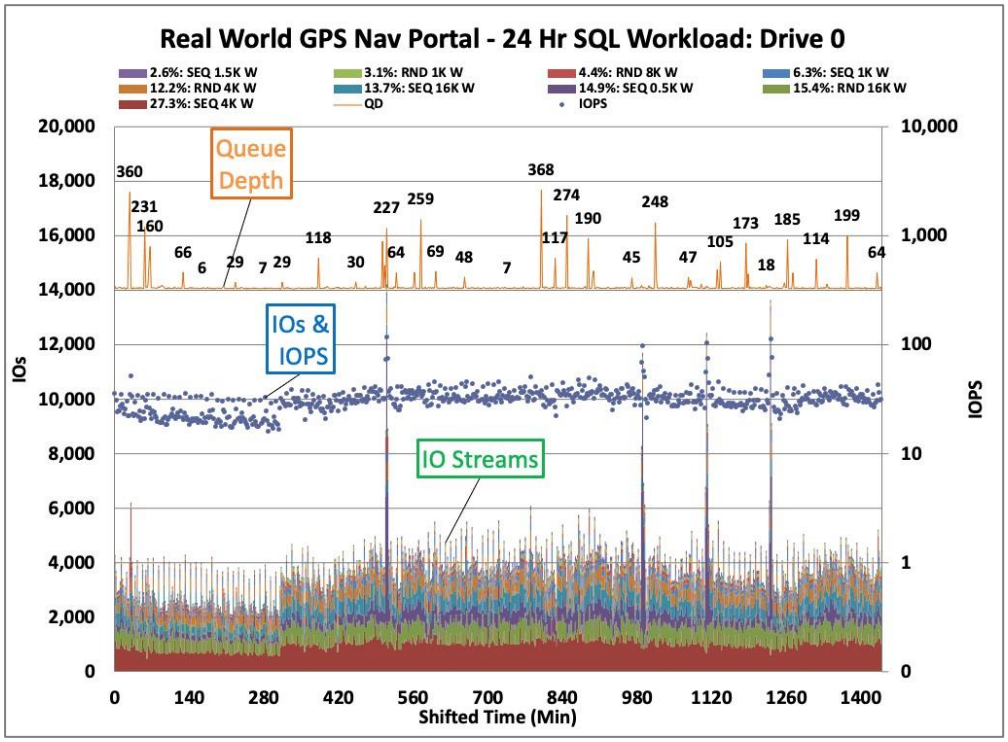


図 10 ー IO ストリーム・マップ：24 時間 GPS Nav Portal

注： IO ストリーム・マップには、1 つ以上のドライブ IO キャプチャを表示できる。 CTS IO ストリーム・マップ機能は、複数の同時ドライブ IO キャプチャからの IO ストリームと統計を単一の合成 IO ストリーム・マップに結合できる。

図 11 - GPS Nav 9 IO ストリーム

図 10 は、12,000 個の IO の 4 つの SEQ 0.5K W スパイクと、QD=6～QD=368 の QD 範囲と中央値 QD=8 を示している。Retail Web Portal の IO の範囲が 40,000～160,000 個であったことに対し、IO が 10,000 個を中心としたバンドに集中していることに注意されたい。

最も発生頻度の高い 9 つの IO ストリームが、発生した合計 350 万個の IO の 78%を占める。9 つの IO ストリームのワークロードの正規化 RW 比率は 100% W である。つまり、全 1,033 個の IO ストリームの RW 比率が 94% W であることに対して、選択された 9 つの IO ストリームのみに基づくと 100% W である。

9 つの IO ストリームのブロック・サイズは、Retail Web Portal IO ストリームより小さく、範囲は最大で 16K（Retail Web Portal ワークロードでは 64 K）である。

IO 発生割合で上位 4 つの IO ストリームは、SEQ 4K W（21.6%）、RND 16K W（12%）、SEQ 0.5K W（11.7%）、および SEQ 16K W（10.7%）である。図 7：現実世界ワークロード比較表を参照されたい。

VDI Storage Server Cluster

下の図 12 と図 13 は、12 時間 75%書き込み VDI 6 ドライブ・ストレージ・クラスタの 6 ドライブ IO ストリーム・マップと 9 つの IO ストリームのワークロードを示している。この IO ストリーム・マップは、IO ピークが異なる IO ストリームと QD の同質的な組成を示している。

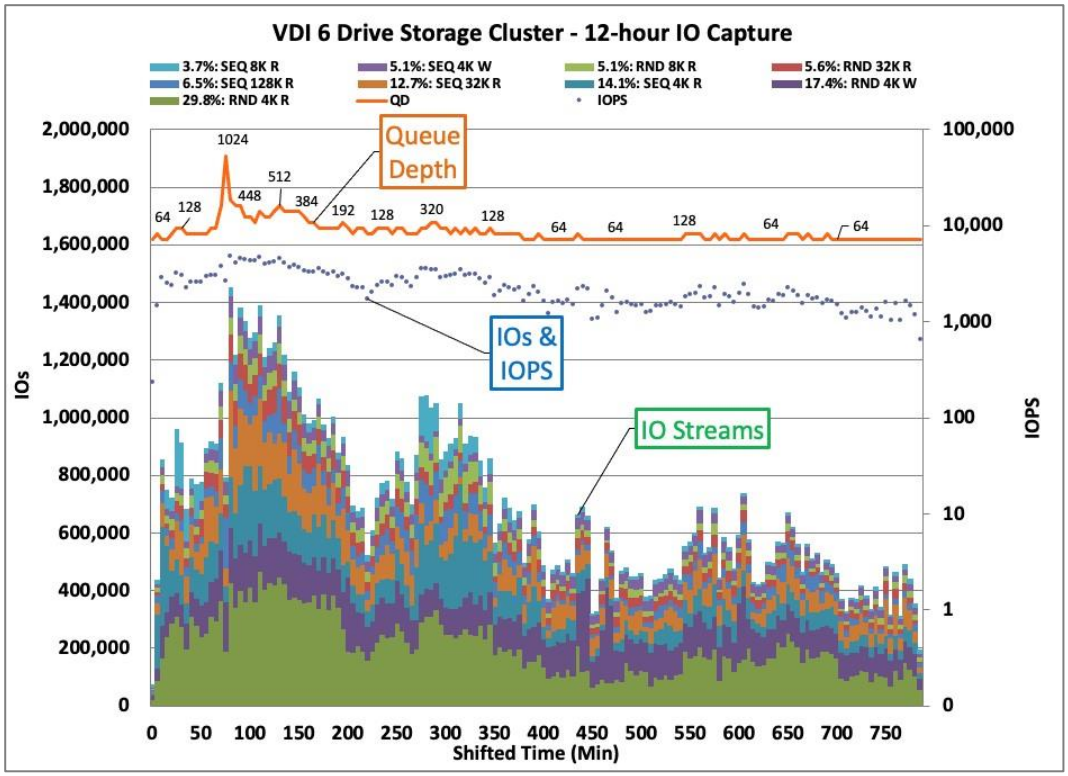


図 12 ー IO ストリーム・マップ：12 時間 VDI Storage Cluster

上の図 12 では、IOPS の変化と QD の変化が連動しており、QD が 1,024 でピークのときに 1.4M IO のピークが来ている。QD 範囲は、QD=7～QD=1,024 で、中央値は QD=128 である。Retail Web Portal の IO が 40,000～160,000 個の範囲であったことや GPS Nav Portal バンドの IO が 10,000 個であったことに対し、今回の IO は 1.4 M 個を中心としたバンドに集中していることに注意されたい。

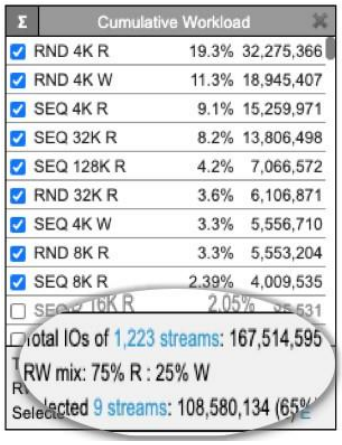


図 13 - VDI Cluster 9 IO ストリーム

図 13 の VDI Storage Cluster 累積ワークロードでは、9 つの IO ストリームのブロック・サイズは、RND/SEQ 4K RW と RND/SEQ 8K RW が大部分を占め、一部が SEQ 32K RW と SEQ 128K R である。これは、IO ブロック・サイズがブロック IO ストレージにより多く関連付けられることを反映している。

最も発生頻度の高い 9 つの IO ストリームが、発生した合計 IO の 65%を占める。ここでは、Retail Web Portal ワークロードと GPS Nav Portal ワークロードでの 4.5 M IO と 3.5 M IO に比べて、167 M IO になっている。9 つの IO ストリームのワークロードの正規化 RW 比率は 75% W である。つまり、全 1,233 個の IO ストリームではなく 9 つの選択された IO ストリームのみに基づくと、RW 比率が 75% W になる。

IO 発生割合で上位 4 つの IO ストリームは、RND 4K R（19.3%）、RND 4K W（11.3%）、SEQ 4K R（9.1%）、および SEQ 32K R（8.2%）である。図 7：現実世界ワークロード比較表を参照されたい。

C. 試験セットアップ

NVMe-oF ホスト要因の正規化 IO ストリームのワークロードと組成はソフトウェアの各層と抽象化の影響を受けるため、性能測定に対するハードウェア／ソフトウェア・スタックの影響を正規化するようにあらゆる努力が払われた。これにより、NVMe-oF ファブリック・トランスポートの*ホスト要因*の影響を評価することができる。図 6：バックツーバック試験トポロジを参照されたい。

現実世界ワークロードが、IO ストリームと QD の絶えず変化する組み合わせで構成されていることを確認した。そのため、現実世界ワークロードをキャプチャしてキュレートするには、IO キャプチャ手順と IO メトリクスの試験ワークロードとスクリプトの忠実かつ正確な編纂、再現、およびキュレーションが必要である。

Control PC、データベース、および CTS スクリプト作成 Calypso Control Server は、Calypso Test Software（CTS）Database 4.0.1、CTS 6.5 programmatic test scripting、および IOProfiler Real-world application workload IO Capture Module 6.5 を使用して、現実世界ワークロードの試験をサポートする。CTS Control Server は、IO キャプチャを作成し、それを編纂して現実世界ワークロード試験スクリプトを作る。

CTS Control Server は、SuperMicro X11SSH-F-O LGA 1151 Micro ATX Server、Intel Xeon QuadCore E3-1225 v5 3.3GHz CPU、80W 8MB L3、32 GB 2133 Mhz DDR4 ECC RAM、64 ビット Windows 7 Pro OS、Calypso CTS 6.5、CTS DB 4.0.1、および TCP 経由で Host Initiator サーバにリモートで接続されている 10/100Mb/s イーサネットで構成される。

CTS IO Stimulus ジェネレータ CTS IO Stimulus ジェネレータ 2.0.2 は、Host Initiator サーバ上にマウントされている。編纂された試験スクリプトは、Control PC から CTS IO Stimulus ジェネレータに送信され、そこで試験 IO がターゲット論理ストレージに適用される。試験測定データのパケットは、Control Server データベースに送信され、アーカイブされ、後で、再現、レポート作成、およびデータ分析に使用される。

Host Initiator Intel サーバ Host Initiator は、Intel Server Board S2600WF、Intel Xeon Platinum 8280 2.7 GHz 28 コア CPU、198 GB 2166 Mhz DDR4 ECC RAM、RHEL OS 8.1 カーネル 5.7.8、および Intel Ethernet Network Adapter E810-CQDA2 で構成される。

Intel Ethernet Network Adapter E810-CQDA2 試験 IO は、ホスト・サーバ NIC カードを介してターゲット・サーバ NIC への 100Gb イーサネット経由で論理ストレージに適用される。Intel Ethernet Network Adapter E810-CQDA2 Network Interface Card（NIC）は、NVMe over Fabrics 用のすべてのイーサネットベースのトランスポートをサポートする。これにより、試験セットアップ内での単一 NIC の使用が可能になり、イーサネット・トランスポートが変更されたときに NIC を変更する必要がなくなる。

1 台のアダプタで、RDMA iWARP、RDMA RoCEv2、標準 TCP を含むすべてのイーサネットベースのトラフィックを処理できる。Intel E810-CQDA2 は、100Gb/50Gb/25Gb/10Gb を含む複数ポート構成をサポート可能な 8 SerDes と MAC も有する。Intel E810 は、サーバあたり最大 4x 25Gb ポートまたは 8x 10 Gb ポートもサポートする。

100Gb イーサネット・ケーブル QSFP28 Direct Attach 100Gb イーサネット・ケーブルは、短距離直接相互接続用に設計された高密度、低電力、パッシブ、直接接続 100Gb イーサネット・ケーブルである。ここでは、1m のケーブルが「バックツーバック」（スイッチを使用しない）構成のホスト・サーバとターゲット・サーバの Intel E810 NIC を接続している。

ターゲット・サーバ ターゲット・サーバは、Intel Server Board S2600WF、Intel Xeon Platinum 8280 2.7 GHz 28 コア CPU、198 GB 2166 Mhz DDR4 ECC RAM、RHEL OS 8.1 カーネル 5.7.8、および Intel Ethernet Network Adapter E810-CQDA2 で構成される。

ターゲット・ストレージ LUN ターゲット・ストレージは、2 つの別々の 6 ドライブ RAID 0 LUN で構成される。SSD-1 は、6 台の 375 GB 3D XPoint SSD で構成され、LUN 容量は 2.25 TB である。SSD-2 は、6 台の 4TB 3D NAND SSD で構成され、LUN 容量は 24 TB である。SSD-1 は 3D XPoint SSD に基づく低容量（2.25 TB）LUN であるのに対して、SSD-2 は 3D NAND SSD に基づく高容量（24 TB）LUN である。

注： メーカー指定の QD は、表に記載された性能値を満たす最適な QD 設定であって、通常のドライブ操作に必要な最小 QD または最大 QD を示しているわけではない。

Test Set-Up: NVMe-oF Transport Test		
Item	Description	Note
Control Server	Calypso CTS Control Server; CTS test Software 6.5; CTS Database 4.0.1; IOProfiler IPF 6.5	CTS Software, Database & Test Scripting IO Capture; Curation & Creation of Real-world workload scripts; Archival, Analytics & Reporting of Test Results
Real-World Workload IO Capture	Calypso IOProfiler (IPF) Real-World Application Workload IO Capture	Time-step IO Capture of Real-World Application Workloads: Block IO level IO Captures
Test Workloads	Synthetic Corner Case Real-World Application	RND 4K RW; SEQ 128K RW – Single Stream T4Q32 Rtl Web; GPS Nav; VDI Storage Cluster – 9 IO Stream
IO Stimulus Workload Generator	CTS IO Stimulus Generator 2.0.2; Libaio 48-bit RND number generator	Host based AIO Stimulus Generator; Application of Test IOs across NVMe-oF Fabrics to Target Storage; Transmits test results data to CTS Control Server
Host Initiator Server	Intel Server Board S2600WF; Intel Xeon Platinum 8280, single 28 core CPU, 198 GB DDR4 ECC RAM	Intel Xeon 8280 2.7Ghz 28 core CPU, 198GB 2166 Mhz DDR4 ECC RAM, RHEL 8.1 kernel 5.7.8
Network Interface Card (NIC)	Intel Ethernet Network Adapter E810CQDA2Multi-transport NIC	Intel Ethernet Network Adaptor E810-CQDA2Host NIC & Target NIC; ice-1.1.3, rdma-1.1.21, NVM 2.1 0x8000433E; iWARP, RoCEv2, TCP; Link Flow Control On
100Gb Ethernet Cable	QSFP28 Direct Attach 100Gb Ethernet cable	High density, low power, passive, short distance (1.0m) direct attach 100Gb cable
Maximum Transmission Unit (MTU)	1500B 9000B	Standard frame Jumbo frame
Ethernet Transport	RDMA (iWARP, RoCEv2) TCP	RDMA offset (iWARP, RoCEv2) No RDMA non stateful offload (TCP)
Target Storage Server	Intel Server S2600WF; single 28 core CPU, 198 GB DDR4 ECC RAM	XEON 8280 2.7Ghz 28 core CPU, 198 2166 Mhz DDR4 ECC RAM, RHEL 8.1 kernel 5.7.8
Target Storage LUN – SSD-1	3D XPoint – LUN capacity 2.25TB RAID 0 SSD LUN - 6 x 375 GB SSD	Mfgr Spec: RND 4K IOPS: 550K IOPS R; 550K IOPS W – QD16 SEQ 128K MB/s: 2500 MB/s R; 2200 MB/s W – QD16
Target Storage LUN - SSD-2	3D NAND - LUN capacity 24.0TB RAID 0 SSD LUN - 6 x 4TB SSD	Mfgr Spec: RND 4K IOPS: 636K IOPS R; 111K IOPS W – QD256 SEQ 128K MB/s: 3000 MB/s R; 2900 MB/s W – QD256

図 14 ― 試験セットアップ： NVMe-oF トランスポート試験

D. 試験方法

この試験セットアップでは、NVMe-oF の性能に対するホスト要因の影響を分離するためのハードウェア／ソフトウェア環境の正規化を試みる。2 つの RDMA（iWARP と RoCEv2）トランスポートと TCP を介して様々な人工的アプリケーション・ワークロードと現実世界アプリケーション・ワークロードを適用する。2 つのタイプのストレージ LUN（SSD-1 3D XPoint と SSD-2 3D NAND）と MTU フレーム・サイズ（標準 1500B とジャンボ 9000B）を使用して性能差を評価する。

単一 IO ストリームの人工的コーナー・ケース試験は、試験ソフトウェアによって生成される。複数 IO ストリームの現実世界アプリケーション・ワークロードは、このケースではブロック IO レベルの IO ステップ・キャプチャ・ツールによって観察される IO ストリームに基づく。IO キャプチャは、Control Server データベースにアーカイブされ、9 つの IO ストリームの試験ワークロード 3 つを編纂し、現実世界ワークロード試験スクリプトを作成するために使用される。

試験スクリプトは、Control Server データベースから Host Initiator 上の IO Stimulus ジェネレータに伝送される。IO Stimulus ジェネレータは、試験 IO を、Host Initiator NIC - 100Gb イーサネット・ケーブル – NIC を介して、ターゲット・ストレージ LUN（SSD-1 または SSD-2）に適用する。試験測定データのパケットは、NVMe-oF を介して Control Server データベースに戻され、表示、データ分析、後処理、およびレポート作成に使用される。

メトリクス

IOPS、MB/s、および応答時間（RT）は、SNIA PTS 定義ごとに参照され、使用される。高い IOPS および MB/s と低い RT は高性能を意味する。平均応答時間（ART）はすべての RT の平均であるのに対して、Maximum RT（MRT）は観察された単一の最も高い RT である。「99.999」（5 つの 9）RT サービス品質（QoS）は、100,000 IO RT ごとの 99.999（5 つの 9）（または 100,000 IO RT に 1 回のドロップ）を評価する。99.999 QoS は、「ロング・テール」RT と呼ばれることもあるが、RT 分布をより正確に反映しているため、IO「サービス品質」として知られている。

未処理 IO（OIO）、つまり、需要強度（DI）は、特定の時間にストレージに適用される試験 IO の合計 TC x QD である。コーナー・ケース試験 Max OIO=128、現実世界ワークロード再現試験 Max OIO の範囲は 306～1,024 であるのに対して、現実世界ワークロード TC/QD スイープ試験は Max OIO=576 である。

現実世界ワークロード IO キャプチャ方法

ここで使用される現実世界アプリケーション・ワークロードは、IO キャプチャ・ツールを使用してブロック IO レベルで取得された現実世界ワークロード IO キャプチャから導出される。現実世界ワークロードは、現実世界アプリケーションとストレージの使用中に観察された IO ストリームと QD に対するストレージとアプリケーションの性能を評価するように設計されている。

現実世界ワークロード・ソース・キャプチャは IO ストリーム活動のキャプチャであり、その統計は、一連の事前定義のステップ、つまり、時間ステップにわたって平均化される。個人データや実データは記録されない。IO ストリーム・メトリクスは、時間ステップごとに平均化され、一定期間（秒、時間、日、または週）の IO ストリーム活動を特徴付け、データベースから IO ストリーム・マップを再作成するために使用される。SNIA CMSI 基準ソース・ワークロードと IO キャプチャ・ツールは、www.TestMyWorkload.com から無料で表示してダウンロードできる。

事前調整と定常状態

現実世界ワークロード・スレッド・カウント／キュー深さスイープ（TC/QD スイープ）試験は、すべての試験の事前調整と定常状態決定因子として使用される。この試験は、最初に、他のすべての試験がデバイス・パージや事前調整なしでいきなり実行された後の定常状態に対して実行される。以降の試験ワークロードのアプリケーションは、初期 TC/QD スイープ試験で実現された定常状態に基づく。

TC/QD スイープ試験の定常状態は、LUN ユーザ容量の 2 倍分の SEQ 128K W 事前調整を適用してから、定常状態に達するまで SNIA Real World Storage Workload（RWSW）Performance Test Specification（PTS）定常状態方法を実行することによって、決定される。ここで、SNIA RWSW PTS v1.0.7 を参照されたい。

SEQ 128K W 事前準備後は、Applied Test Workload（または試験ワークロードとして選択された IO ストリーム – RWSW PTS Definitions 2.1.2 を参照）が、TC/QD 追跡変数の 5 回連続測定が最大 20%のデータ偏位／すべてのデータ測定が報告される 10%の傾斜定常状態ウィンドウを満たすまで実行される。この場合は、Applied Test Workload の最も高い OIO（または TC/QD）の組み合わせが定常状態追跡変数として使用される。

人工的コーナー・ケース・ベンチマーク試験

人工的コーナー・ケース試験は、4 つの単一 IO ストリーム評価パターン試験（RND 4K R、RND 4K W、SEQ 128 K R、および SEQ 128K W）で構成される。各ケースでは、ストレージが、最初に現実世界ワークロード TC/QD スイープ試験を適用することによって、定常状態に事前調整される（上記参照）。

TC/QD スイープ事前調整／定常状態が実現されたら、各コーナー・ケース・ワークロードが、スレッド・カウントを 4 に設定し、キュー深さを 32（T4Q32）に設定することによって、128 の未処理 IO（OIO）で 10 分間実行される。

人工的コーナー・ケース・ベンチマーク試験は、MTU フレーム・サイズ、ストレージ LUN、および NVMe-oF トランスポートの IOPS と RT QoS の性能を比較するために使用される。試験結果は、NVMe-oF LUN の性能を基礎となる SSD のメーカー性能仕様と比較するためにも使用できる（上の図 5）。

現実世界ワークロード再現試験

このスタディでは、2 つのタイプの現実世界ワークロード試験（再現試験と TC/QD スイープ試験）を実行する。再現試験では、IO キャプチャで観察された IO ストリームと QD のシーケンスと組み合わせを再

作成して実行する。TC/QD スイープ試験では、OIO（未処理 IO または需要強度）の範囲を変えながら TC/QD スイープ試験の各ステップに IO ストリームの固定の組み合わせを適用し、アプリケーションとストレージの IO と応答時間の飽和状態を評価する。

再現試験は、現実世界で使用中に観察された実際のワークロードに対するストレージの性能を測定するように設計されている。現実世界ワークロードが IO キャプチャ（IO キャプチャ方法を参照）としてキャプチャされたら、時間ステップ統計を使用して再現試験スクリプトが編纂される。試験スクリプトの時間ステップ期間は、非常に長い現実世界のキャプチャ（分、時間、日）がそれより短い試験期間で再現されるように、または、短いが粒度の細かい分解能のキャプチャ（マイクロ秒、秒）がそれより長い試験期間に再現されるようにユーザが設定できる。

再現試験スクリプトによって生成され実行される IO 操作は様々な時間ステップ統計を反映している。 再現試験スクリプトは、IO キャプチャ内で観察されたようにすべての IO 操作の正確な順序とタイミング・シーケンスを再現しない。

再現試験結果は、比較（つまり、再現試験全体で平均化した IOPS、MB/s、および応答時間の比較）がしやすいように「単一値」（つまり、一定期間の平均値）として表示できる。ただし、さらに重要なことに、再現試験を使用すれば、試験操作者は、様々な時間のサブセット（つまり、IO キャプチャ中に発生した特定の期間、イベント、またはプロセス ID（PID））の測定値を観察することができる。これにより、バックアップ、ブート・ストーム、日常活動など、注目する離散的イベントの分析が可能になる。

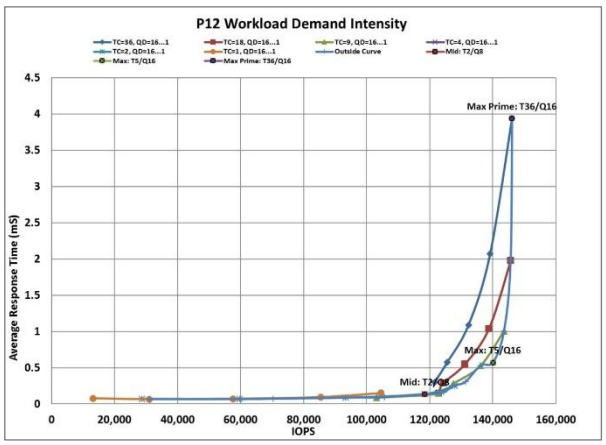
現実世界ワークロード TC/QD スイープ試験

TC/QD スイープ試験は、需要強度（OIO）の増加に伴う現実世界ワークロードの性能の飽和状態を評価するように設計されている。この試験では、注目する数個の IO ストリーム、通常は、IO キャプチャ中の IO 発生割合上位の IO ストリームを使用して、TC/QD スイープ試験の各ステップの IO ストリームの固定の組み合わせが構築される。ここでは、合成 Applied Test Workload として、現実世界ワークロードごとに最も発生頻度の高い 9 つの IO ストリームを選択した。図 7、9、11 ― 9 つの IO ストリーム・ワークロードと図 13 ― 現実世界ワークロード比較表を参照されたい。

この 9 つの IO ストリームの組み合わせが試験スクリプトの各ステップの試験ストレージに適用される。LUN 容量の 2 倍分の SEQ 128K W を適用することによってストレージを事前調整したら、需要強度を 1 分単位で低 OIO から高 OIO に（例：OIO=1 から OIO=576 に）変化させながら、9 つの IO ストリーム・ワークロードが実行される。定常状態追跡変数（このケースでは OIO=576）は、定常状態ウィンドウ（5 連続 OIO ラウンドの最適な線形近似曲線の 20%のデータ偏位と 10%の傾斜を超えないと定義）内に性能が入るまで観察される。

TC/QD スイープ試験結果は、需要強度（DI）外（OS）曲線として表示され、簡単に比較できる。DI OS 曲線は、増加する OIO の関数としての IOPS を表す。IOPS が x 軸上に表示され、平均応答時間（ART）が y 軸上に表示される。IOPS と ART は、最小 OIO から最大 OIO までの最適な線形近似曲線でプロットされる。結果の DI OS 曲線は、OIO の増加に伴う IOPS と ART の増加を示す。

DI OS 曲線内の性能指数は、IOPS が高く、ART がまだ急激に増加していない DI OS 曲線の「屈曲点」の手前の最適な OIO ポイントである。DI 曲線の屈曲点はアルゴリズム（60%傾斜増加式）で決定されるが、DI OS 曲線は、通常、ART の急増とそれに伴う IOPS の横ばいまたは減少（「反転」）を示す。下の図 15 ― 需要強度曲線と図 16 ― 需要強度外曲線を参照されたい。



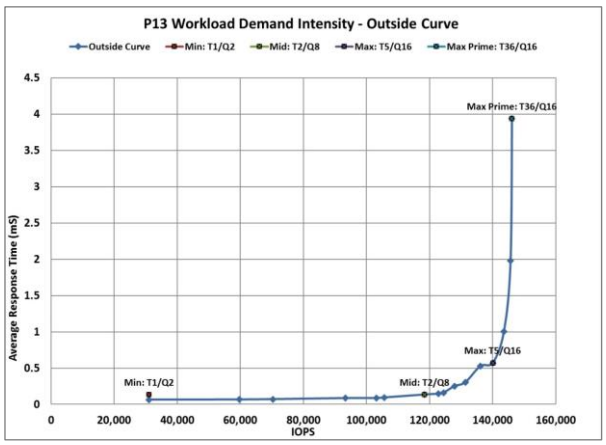


図 15 ― 需要強度曲線

図 16 ― 需要強度外曲線

上の図 15 の DI 曲線は、各スレッド・カウント（TC）を QD が QD=1 から QD=16 まで増加するデータ系列として表示している。黄色の線は TC=1 で、青色の線は TC=36 である。この DI 曲線では、Min IOPS OIO=T1Q1、Mid IOPS OIO=T2Q8、および Max IOPS OIO=T5Q16 である。Max Prime IOPS OIO（T36Q16）は、ART とは無関係の最大 IOPS、つまり、最大 TC/QD での IOPS として定義される。性能指標は、Max IOPS OIO ポイント（T5Q16）、つまり、IOPS が高く、ART がまだ急増していない OIO ポイントである。

上の図 16 は、Min、Mid、Max、および Max Prime IOPS OIO と ART がアルゴリズムで決定され、最適な線形近似外曲線が形成される DI 外曲線を示している。ここでは、Min IOPS OIO=T1Q2、Mid IOPS OIO=T2Q8、Max IOPS OIO=T5Q16、および Max Prime IOPS OIO=T36Q16 である。簡略化された DI OS 曲線を使用すれば、複数の DI OS 曲線、試験、またはストレージ・ユニットを簡単に比較することができる（下の試験結果 ― TC/QD スイープを参照）。

DI OS 曲線の解釈では、この特定の 9 つの IO ストリーム・ワークロード（Retail Web Portal）を使用すると、OIO=80（T5Q16）の飽和点に達して IOPS が 140,000 に ART が 0.5 mS になるまで、IOPS が OIO と一緒に増加を続けることになる。この後、IOPS は 145,000 で横ばいになるが、ART は OIO=576（T36Q16）になるまで OIO と共に増加し、4 mS に達する。別の言い方をすれば、最適な OIO=80（140,000 IOPS と 0.5 mS ART）に達すると、OIO は OIO=576 まで増加して IOPS を名目上押し上げるが、ART が 800%（0.5 mS から 4.0 mS まで）増加するコストがかかっている。つまり、追加の 5,000 IOPS のコストは 3.5 mS ART である。

試験フロー

このスタディでは、（2）MTU フレーム・サイズ、（3）NVMe-oF トランスポート、および（2）ストレージ LUN を変化させながら、（7）人工的ワークロード試験と現実世界ワークロード試験を適用する。その結果、84 試験の 7 x 2 x 3 x 2 の試験マトリックスができる。

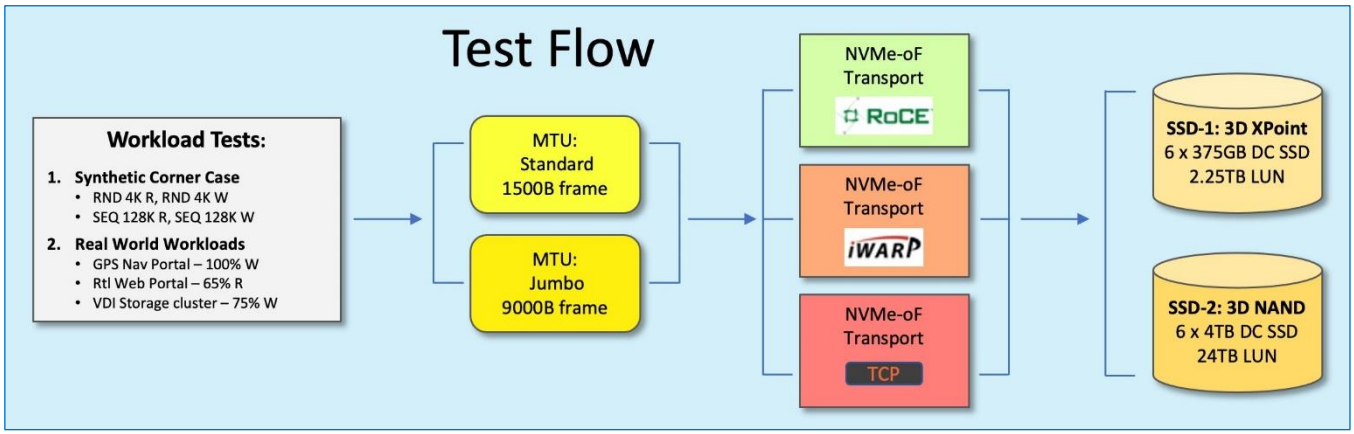


図 17 ― 試験フロー：84 試験 ― 試験マトリックス

E. 試験結果

人工的コーナー・ケース： RND 4K RW と SEQ 128K RW

iWARP、RoCEv2、および TCP 100Gb イーサネット・トランスポートにわたって、1500B 標準（青）フレーム・サイズと 9000B ジャンボ（赤）フレーム・サイズを比較する。単一の IO ストリームの人工的な RND 4K RW および SEQ 128K RW ワークロードを 3D XPoint および 3D NAND ストレージ LUN に適用する。OIO は OIO=128（T4/Q32）に設定し、IOPS と 99.999 の応答時間サービス品質（RT QoS）を測定する。注 ― SEQ 128K RW IOPS は 8 で割って MB/s に変換できる。

サマリー。標準フレームとジャンボ・フレームは、後述する場合を除いて、実質的に同等の IOPS を示す。iWARP 読み取りワークロードは、非常に大きな RT QoS スパイク（147mS から 429mS）を示す。CPU オフロードを使用した RDMA（iWARP、RoCEv2）の性能は、実質的に同等である。RDMA は、ソフトウェアベース（オフロードなし）の TCP より大幅に高い性能である。3D XPoint の性能は、iWARP 経由の RND 4K および SEQ 128K R RT と TCP 経由の RND 4K R IOPS を除いて、3D NAND より高い。

3D XPoint LUN

- **iWARP** – RND 4K R ジャンボ IOPS は標準フレームより高い。読み取りワークロードは非常に高い RT QoS スパイクを示すが、これは、SSD ストレージ・レベルを超えるホスト要因が原因だと考えられる（図 18 を参照）。
- **RoCEv2** – RND 4K RW および SEQ 128 K RW は、標準フレーム・サイズとジャンボ・フレーム・サイズでかなり似通った IOPS と RT QoS を示す。RoCEv2 は、読み取りワークロードで RT QoS スパイクを示さない（図 19 を参照）。
- **TCP** – RND 4K W IOPS はジャンボ・フレームで高いが、RND 4K R IOPS と SEQ 128K RW IOPS はかなり似通っている。TCP SEQ 128K RW は高い RT QoS を示す（図 20 を参照）。

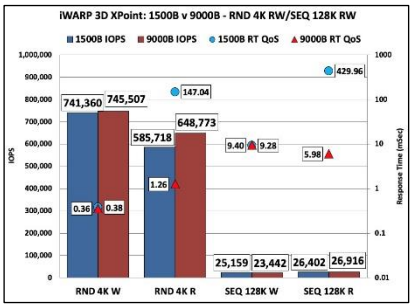


図 18 – 人工的コーナー・ケース iWARP 3D XPoint： IOPS と QoS – 1500B と 9000B の比較

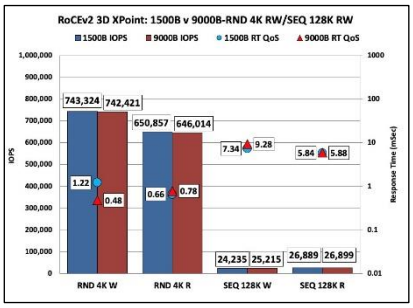


図 19 – 人工的コーナー・ケース RoCEv2 3D XPoint： IOPS と QoS – 1500B と 9000B の比較

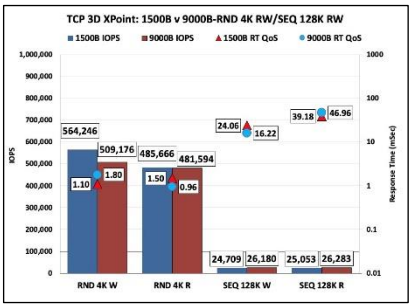


図 20 – 人工的コーナー・ケース TCP 3D XPoint： IOPS と QoS – 1500B と 9000B の比較

3D NAND LUN

- **iWARP** – RND 4K R および SEQ 128K W 標準フレーム IOPS はジャンボ・フレームより高い。RND 4K W および SEQ 128K R IOPS は似通っている。SEQ 128K W は高い RT QoS を示す（図 21 を参照）。
- **RoCEv2** – RND 4K RW および SEQ 128 K RW は、標準フレーム・サイズとジャンボ・フレーム・サイズでかなり似通った IOPS と RT QoS を示す。RoCEv2 SEQ 128K W は iWARP と同様の高い RT QoS を示す（図 22 を参照）。
- **TCP** – RND 4K R および SEQ 128K W ジャンボ・フレーム IOPS は標準フレームより高い。SEQ 128K RW は高い RT QoS を示す（図 23 を参照）。

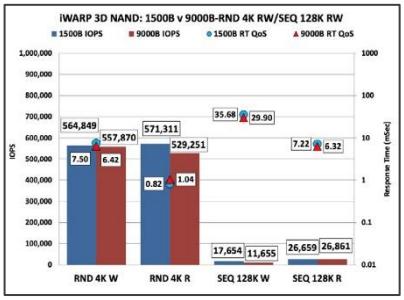


図 21 – 人工的コーナー・ケース iWARP 3D NAND： IOPS と QoS – 1500B と 9000B の比較

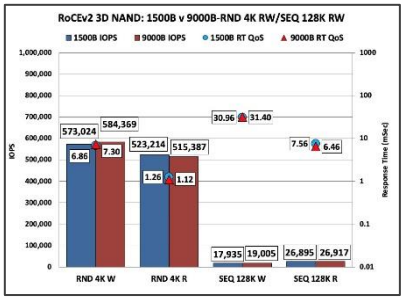


図 22 – 人工的コーナー・ケース RoCEv2 3D NAND： IOPS と QoS – 1500B と 9000B の比較

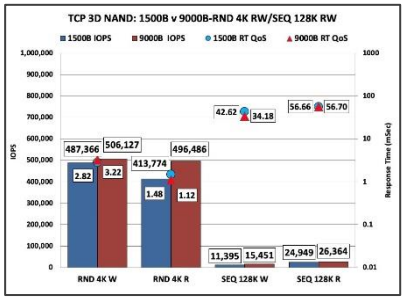


図 213 – 人工的コーナー・ケース TCP 3D NAND： IOPS と QoS – 1500B と 9000B の比較

現実世界ワークロード： 再現試験

3 つの現実世界ワークロード（Retail Web（65% R）、GPS Nav（100% W）、および VDI Storage Cluster（75% W））を比較する。IO キャプチャ中に観察された 9 つの IO ストリーム・ワークロードを実行する。再現試験は、標準（青）MTU フレーム・サイズとジャンボ（赤）MTU フレーム・サイズを使用して、3D XPoint ストレージと 3D NAND ストレージに対して実行される。結果は、再現試験全体で平均化された IOPS と RT QoS を示す。最大 OIO は、306（Rtl）、368（GPS）、および 1,024（VDI）である。

サマリー。標準フレームとジャンボ・フレームは、同等の IOPS と RT QoS を示す。RDMA（iWARP、RoCEv2）IOPS および RT QoS の性能は、ソフトウェアベースの TCP より大幅に高い。3D XPoint ストレージの性能は、特に、RDMA Retail Web Portal ワークロードと GPS Nav Portal ワークロードの場合に、3D NAND ストレージよりかなり高い。ただし、VDI Storage Cluster ワークロードの場合は、3D XPoint と 3D NAND の性能が似通っている。

この再現試験の IOPS は、人工的コーナー・ケース試験の IOPS よりかなり低いが、これは、一部には、現実世界の再現ワークロードで使用された OIO と、IO ストリームのタイプとシーケンスが異なることが原因であることに注意されたい。

3D XPoint LUN

- iWARP の標準フレームとジャンボ・フレームは、かなり似通った性能を示す（図 24 を参照）。
- RoCEv2 標準フレームは、GPS Nav と VDI Cluster Storage の場合に高い IOPS を示す（図 25 を参照）。
- TCP ジャンボ・フレームは、Retail Web と VDI Cluster Storage の場合に高い IOPS を示す（図 26 を参照）。
- RT QoS は、標準フレームとジャンボ・フレームで似通っている。RT QoS は、iWARP と RoCEv2 の場合に TCP より低い。
- 75% W VDI ワークロードは、100% W GPS Nav ワークロードと 65% R Retail Web ワークロードより高い IOPS を示す。

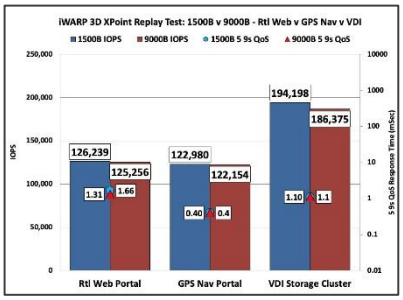


図 22 ― 再現試験： Rtl と GPS と VDI の iWARP 3D XPoint の比較： IOPS と QoS – 1500B と 9000B の比較

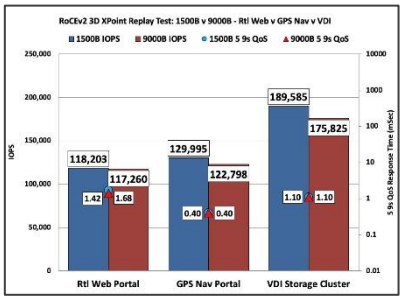


図 23 ― 再現試験： Rtl と GPS と VDI の RoCEv2 3D XPoint の比較： IOPS と QoS – 1500B と 9000B の比較

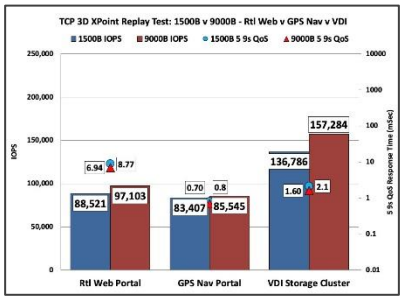


図 24 ― 再現試験： RtlとGPSとVDIのTCP 3D XPointの比較： IOPSとQoS – 1500Bと9000Bの比較

3D NAND LUN

- iWARP の標準フレームとジャンボ・フレームは、かなり似通った性能を示す（図 27 を参照）。
- RoCEv2 の標準フレームとジャンボ・フレームは、かなり似通った性能を示す（図 28 を参照）。
- TCP ジャンボ・フレームは、VDI Cluster Storage の場合に高い IOPS を示す（図 29 を参照）。
- RT QoS は、iWARP、RoCEv2、および TCP の場合に標準フレームとジャンボ・フレームで似通っている。
- Rtl Web および GPS Nav 3D NAND RDMA の IOPS は、Rtl Web および GPS Nav 3D XPoint RDMA より低い。
- 75% W VDI ワークロードは、100% W GPS Nav ワークロードと 65% R Retail Web ワークロードより高い IOPS を示す。

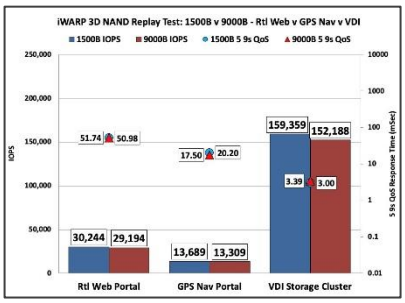


図 25 ― 再現試験： Rtl と GPS と VDI の iWARP 3D NAND の比較： IOPS と QoS – 1500B と 9000B の比較

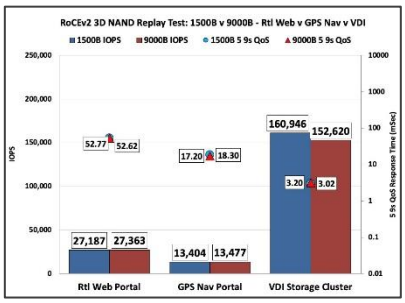


図 26 ― 再現試験： RtlとGPSとVDIのRoCEv2 3D NANDの比較： IOPSとQoS – 1500Bと9000Bの比較

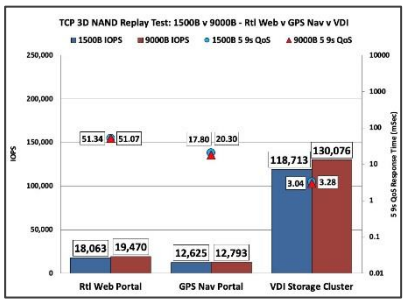


図 27 ― 再現試験： RtlとGPSとVDIのTCP 3D NANDの比較： IOPSとQoS – 1500Bと9000Bの比較

現実世界ワークロード： TC/QD 深さスイープ試験

スレッド・カウント／キュー深さ（TC/QD）スイープ試験は、Retail Web Portal（65% R）、GPS Nav Portal（100% W）、および VDI Storage Cluster（75% W）の現実世界ワークロードを固定 IO ストリーム・ワークロードとして比較する。9 つの IO ストリームの Applied Test Workload は、OIO が OIO=1（T1Q1）～OIO=576（T36Q16）の範囲で変化しながら、試験の各ステップに固定の組み合わせとして適用される。標準 1500B フレーム・サイズの結果のみを示す。

TC/QD スイープ試験の目的は、需要強度外曲線（DI OS）を使用して、OIO の増加に伴う IOPS と ART の飽和状態を観察することである。拡大されたデータ点で示される主要性能指標は、IOPS は最も高いが、ART はまだ急増していない Max IOPS OIO ポイントである。

サマリー。RDMA は、max ART が低く、IOPS が高い滑らかな DI OS 曲線を示す。ただし、TCP は、3 つのすべてのワークロードの Max IOPS OIO ポイントで高い IOPS を示す。RDMA は、OIO が Max IOPS OIO ポイントを超えて増加したときに TCP より高い IOPS を示す。つまり、RDMA IOPS は OIO とともに増加するが、このケースでは、ART の受け入れ難い増加が見られる。すべてのトランスポートが似通っているが、ワークロードによって違いが見られる。

3D XPoint LUN

- **Retail Web 65% R** ― TCP は、優れた Max IOPS OIO ポイント ART を示すが、T36Q16 のときに高い max ART を示す。RDMA は、曲線の屈曲点以降で TCP より高い IOPS を示す（図 30 を参照）。
- **GPS Nav 100% W** ― TCP は、優れた Max IOPS OIO ポイント IOPS および ART を示す（図 31 を参照）。
- **VDI Cluster 75% W** ― TCP は、優れた Max IOPS OIO ポイント IOPS ART を示すが、T36Q16 のときに高い max ART QoS を示す。RDMA は、曲線の屈曲点以降で TCP より高い IOPS を示す（図 32 を参照）。

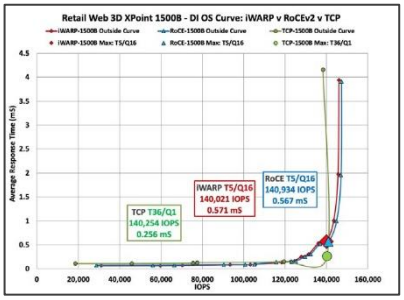


図 28 ― DI OS 曲線： Rtl Web 3D XPoint： iWARP と RoCEv2 と TCP の比較

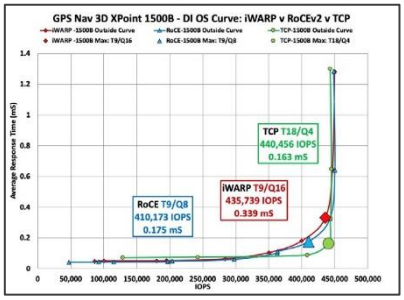


図 29 ― DI OS 曲線： GPS Nav 3D XPoint： iWARP と RoCEv2 と TCP の比較

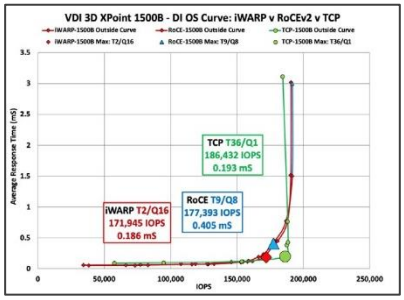


図 30 ― DI OS 曲線： VDI Cluster 3D XPoint： iWARP と RoCEv2 と TCP の比較

3D NAND LUN

- **Retail Web 65% R** – RDMA と TCP は、かなり似通った DI OS 曲線を示す（図 33 を参照）。
- **GPS Nav 100% W** – RoCEv2 は RT QoS が最も低く IOPS が低い。iWARP は IOPS と ART が高い。TCP は曲線の屈曲点で IOPS と RT が最も高い（図 34 を参照）。
- **VDI Cluster 75% W** ― TCP は RT が最も低く IOPS が最も高い。iWARP と RoCEv2 は低い IOPS と似通った RT を示す（図 35 を参照）。

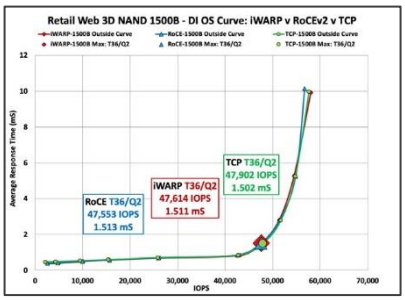


図 31 ― DI OS 曲線： Rtl Web 3D NAND： iWARP と RoCEv2 と TCP の比較

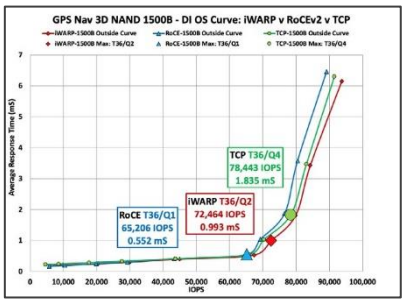


図 32 ― DI OS 曲線： GPS Nav 3D NAND： iWARP と RoCEv2 と TCP の比較

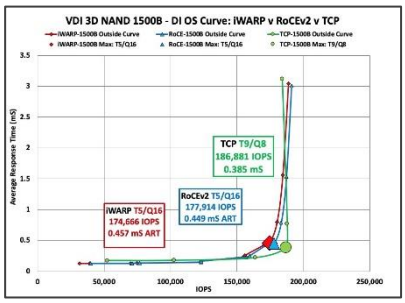


図 33 ― DI OS 曲線： VDI Cluster 3D NAND： iWARP、RoCEv2、および TCP の比較

3D XPoint ストレージ LUN と 3D NAND ストレージ LUN の比較

人工的コーナー・ケース RND 4K および SEQ 128K RW ワークロードと、TC/QD スイープ現実世界の Retail Web Portal（65% R）、GPS Nav Portal（100% W）、および VDI Storage Cluster（75% W）ワークロードに関して、3D XPoint（青）ストレージ LUN と 3D NAND（赤）ストレージ LUN を比較する。MTU 標準 1500B フレーム・サイズの結果のみを示す。

人工的コーナー・ケース・ワークロードの Max OIO（需要強度）は OIO=128（T4Q32）である。現実世界ワークロードの Max OIO は OIO=576（T36Q16）である。コーナー・ケース・ワークロードは IOPS と RT QoS の比較を示すが、現実世界ワークロード需要強度外曲線（DI OS 曲線）は IOPS と平均応答時間（ART）の比較を示す。

サマリー ー 人工的コーナー・ケース。3D XPoint の IOPS は、特に、書き込みワークロード（RND 4K/SEQ 128K）の場合に 3D NAND より高い。iWARP 3D XPoint は、読み取り（RND 4K/SEQ 128K）で高い RT QoS スパイクが見られる。

- **iWARP** – 3D XPoint 書き込みワークロードは IOPS がかなり高い。読み取りワークロードの IOPS は実質的に同等である（図 36 を参照）。3D XPoint 読み取りワークロードは、ホスト・レベル、スイッチ、またはネットワーク・トポロジの要因が原因で RT QoS が非常に高い（セクション III.D.「個々のドライブ・レベルの要因」を参照）。
- **RoCEv2** – 3D XPoint は、SEQ 128K R IOPS が似通っていることを除いて、IOPS がかなり高い。3D NAND は、RND 4K W と SEQ 128K W の場合に RT QoS がかなり高い（図 37 を参照）。
- **TCP** ー 3D XPoint は、書き込みワークロードの場合に IOPS がかなり高い（図 38 を参照）。

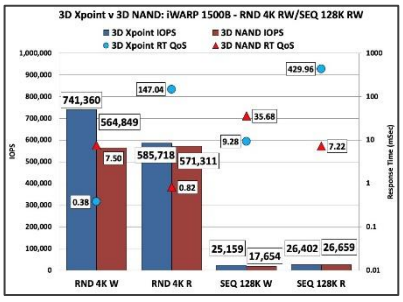


図 34 ー 人工的コーナー・ケース iWARP： 3D XPoint と 3D NAND の比較 ー IOPS と QoS

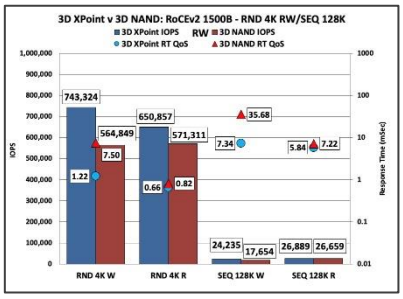


図 35 ー 人工的コーナー・ケース RoCEv2： 3D XPoint と 3D NAND の比較 ー IOPS と QoS

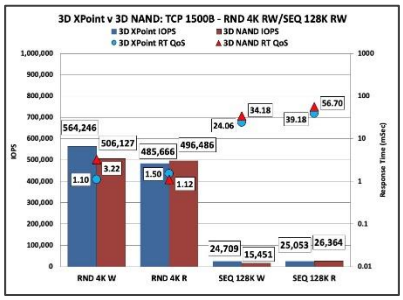


図 36 ー 人工的コーナー・ケース TCP： 3D XPoint と 3D NAND の比較 ー IOPS と QoS

サマリー – 現実世界ワークロード。DI OS 曲線は、3D XPoint と 3D NAND の性能差と様々な RW-IO ストリーム ー 高 DI ワークロードに対する OIO 飽和の影響を明確に示している。3D XPoint RDMA（iWARP と RoCEv2）は、IOPS がかなり高く、ART がかなり低い。VDI 75% W ワークロードでは、3D XPoint と 3D NAND の IOPS は実質的に同等だが、ART は 3D NAND の方が高い。すべてのケースで、完全飽和 OIO（T36Q16）時の最大応答時間は、3D XPoint より 3D NAND の方がはるかに高い。

- **Retail Web 65% R** – 3D XPoint IOPS はかなり高く、ART は低い（図 39 を参照）。
- **GPS Nav 100% W** – 3D XPoint IOPS はかなり高く、ART は低い（図 40 を参照）。
- **VDI 75% W** – IOPS は実質的に同等だが、3D NAND の方が ART が高い（図 41 を参照）。

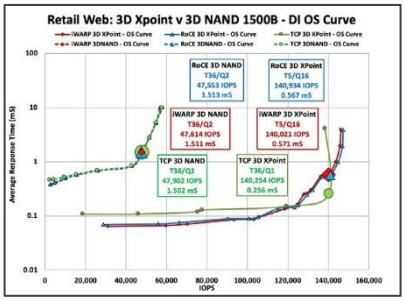


図 37 ー DI OS 曲線： Rtl Web Retail Web： 3D XPoint と 3D NAND の比較 ー IOPS と OIO

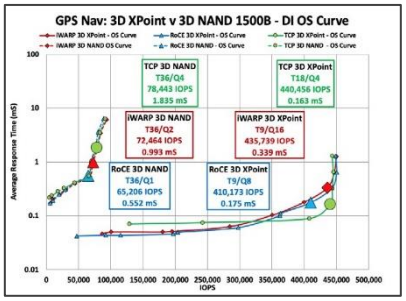


図 38 ー DI OS 曲線： GPS Nav GPS Nav： 3D XPoint と 3D NAND の比較 ー IOPS と OIO

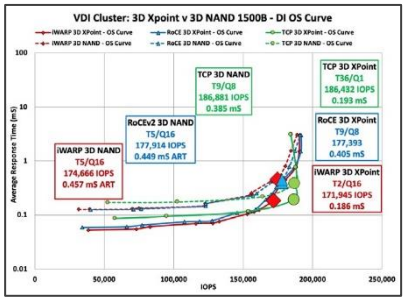


図 39 ー DI OS 曲線： VDI Cluster VDI Cluster： 3D XPoint と 3D NAND の比較 ー IOPS と OIO

V. 結論

この試験の主な目的は、ワークロード、MTU、およびターゲット・ストレージを変化させた場合の RDMA（iWARP と RoCEv2）と TCP 100Gb イーサネット・トランスポートの性能差を評価することである。ホスト要因（MTU フレーム・サイズ、ワークロード、DI、IO ストリーム・コンテンツ、RW 比率など）の影響を評価するために HW/SW 環境の正規化を試みた。このスタディでは、スイッチやネットワーク・トポロジの影響は評価していない。

試験 IO を NVMe ファブリックを介して論理ストレージに適用するため、試験 IO トラフィックを Host Initiator レベルでは観察できるが、100Gb イーサネット・ケーブル経由や、その下では直接観察することはできない。そのため、試験性能結果がホスト・レベルでの論理ストレージに限定され、スイッチ、

ネットワーク・トポロジ、ストレージ LUN、または個々の SSD ストレージの性能による影響や貢献度は区別されない。

基礎となる SSD の性能差は、介在する抽象化層によって曖昧になっている可能性がある。例えば、NVMe-oF で、ストレージ IO がプールされるターゲット・ストレージ・サーバに RAM キャッシュが搭載されているとする。これは性能に影響する可能性があり、書き込み IO が高速な RAM キャッシュに保存される一方で特定の読み取り IO が基礎となるストレージに直接アクセスする必要がある場合があり、すると、相対的な応答時間が遅くなる。加えて、3D NAND SSD は、3D XPoint SSD より、読み取り IOPS（635K 対 550K）スペックが高く、書き込み IOPS（111.5K 対 550K）スペックが低いが、様々な RW 比率と IO ワークロードに対する実際の性能はこの想定された性能差を反映していない可能性がある。

にもかかわらず、一般的に、RDMA トランスポートは、コーナー・ケース読み取り IO（iWARP 読み取りワークロード）が高い RT スパイクを示す特定のケースを除いて、非オフロード TCP と同等の優れた性能を示すことが分かっている。人工的コーナー・ケース・ワークロードの場合は、単一 IO ストリームの性能が 3D NAND ストレージより 3D XPoint ストレージの方が優れている。ただし、1500B 標準フレーム・サイズと 9000B ジャンボ・フレーム・サイズの両方が、人工的ワークロードと現実世界ワークロードの両方で実質的に同等の性能（IOPS、ART、および 99.999 QoS で）を示すことから、MTU フレーム・サイズによる性能差はあまりない。

99.9 IO ストリームの現実世界ワークロードの試験を使用すれば、複数の IO ストリーム、様々な需要強度、異なる RW 比率の性能差と、ストレージと NVMe-oF トランスポートの性能に対する絶えず変化する IO ストリームとキュー深さの組み合わせの影響を評価することができる。

RDMA の性能は、100% Write GPS Nav ワークロードと 65% Read Retail Web ワークロードの場合に、非オフロード TCP より優れている。非オフロード TCP の性能は、75% Write VDI Storage Cluster ワークロードの場合に、RDMA を超えるまではいかないとしても、同等である。この差は、様々な現実世界ワークロードの IO ストリーム・コンテンツが原因だと考えられる。

現実世界ワークロード需要強度外曲線（DI OS 曲線）の評価は、性能に対する DI の増加の影響を示す。未処理 IO を増やしてストレージを飽和状態にさせると、IOPS の横ばいまたは減少が見られ、RT が急増し始める。現実世界ワークロードには複数の IO ストリーム・コンテンツ（および IO ストリームと QD の変化する組み合わせ）が含まれているため、DI OS 曲線の解釈時に平均応答時間（ART）の差を評価する。これにより、ストレージとファブリックの性能に対する現実世界ワークロードの全体的影響を把握することができる。

RDMA DI OS 曲線は、非オフロード TCP より決定的動作を示しているように見える。これは、IOPS と ART が変化する OIO に対してより直接的に一貫して反応しているように見える（前述した基礎となる SSD の性能要因にかかわらず）ためである。加えて、RDMA は、非オフロード TCP に比べて、DI OS 曲線上で低い ART と、最大 OIO 飽和ポイントで低い最大応答時間を示す。

最後に、RDMA トランスポートを比較した場合、iWARP はコーナー・ケース・ワークロードと現実世界ワークロードで名目上高い IOPS を示すのに対して、RoCEv2 は低いがより一貫した RT を示す。ただし、このスタディの iWARP と RoCEv2 間で観察された性能は実質的に同等であることに注意する必要がある。

いずれかの NVMe-oF トランスポートの実装に有利に働く可能性のある他の要因は、このスタディの試験変数として含まれていない。特定のファブリック構成、NVMe-oF ハードウェア全体、サーバと CPU ノードの構成、基礎となるストレージ・デバイス、CPU リソース割り当て、およびネットワーク・アーキテクチャの影響はこのスタディの範囲を超えている。ただし、追加のスタディで、スイッチとネットワーク・トポロジを試験変数として実行する予定である。このホワイトペーパーに関する質問は、冒頭に掲載されている電子メール・アドレスで執筆者に問い合わせることができる。

執筆者について

Fred Zhang, Intel Corp.

Fred Zhang は、Intel Corporation のプロダクト・マーケティング・マネージャで、Intel のイーサネット・コントローラ製品を担当している。シリコン製品やソフトウェア製品を含め、20 年以上の製品管理の経験がある。

SNIA Network Storage Forum（NSF）のメンバーであり、イーサネットベースのストレージ・ソリューションの業界への普及に積極的に取り組んでいる。

Eden Kim、CEO Calypso Systems, Inc.

Calypso Systems の CEO である Eden Kim は、SNIA SSS Technical Working Group の議長であり、SSD ストレージとデータセンター・ストレージに関する SNIA PTS 仕様（https://www.snia.org/tech_activities/work）の主要な執筆者であり、多数の SNIA ホワイトペーパー（https://www.snia.org/forums/cmsi/knowledge/whitepapers）の執筆者である。

また、米国と中国の世界的な業界団体イベントで発表や出版を行ったり、SNIA PM Summit、SNIA SDC、Santa Clara Flash Memory Summit（FMS）、および北京、上海、深川、武漢、およびその他の場所の DOIT がスポンサーを務める中国展示会で定期的に講演を行ったりしている。

Calypso Systems, Inc.は、高度なワークロード分析、試験ソフトウェアと測定ソフトウェア、ハードウェア、および試験サービスのサプライヤおよびメーカーであり、www.testmyworkload.com サイトをホスティングしている。SSD とデータセンターの試験用の SNIA Solid State Storage Performance Test Specification（PTS）Reference Test Platforms（RTP）のサプライヤでもある。Calypso SSD RTP は、SSD ODM および OEM における標準ツールであり、Calypso IPF サーバは、クラウド、データセンター、およびエンタープライズの顧客向けの RWW 用の機能をフル装備している。

付録 A： トランスポートの比較 ― 人工的ワークロード

Appendix A: NVMe-oF Transport Comparison [1] – Synthetic Workloads [2,3,4]

	IO Rate IOPS			Bandwidth MB/sec			Average Response Time (ART) mSec			5 9s Quality of Service (QoS) mSec		
Synthetic Workloads [5,6]	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP
RND 4K W – QD128	564,849	573,024	487,366	2,206	2,238	1,904	0.227	0.224	0.263	7.560	6.860	2.820
RND 4K R – QD128	571,311	523,214	413,774	2,232	2,044	1,616	0.224	0.245	0.309	0.820	1.260	1.480
SEQ 128K W – QD128	17,654	17,935	11,395	2,207	2,242	1,424	7.364	7.196	11.256	35.680	30.969	46.620
SEQ 128K R – QD128	26,659	26,895	24,949	3,332	3,362	3,119	4.801	4.759	5.130	7.220	7.560	56.660

Notes

1 RDMA iWARP & RDMA RoCEv2 NAC with Offload; TCP NIC - no Offload
2 Back-to-Back NVMe-oF Transport Topology – 100GbE, Intel E810-CQDA2, No Network Switch; MTU Frame Size 1500B
3 Intel Server S2600WF; XEON 8280 2.7 Ghz 28 core single CPU, 198 GB 2166 Mhz DDR 4 ECC RAM, RHEL 8.1, kernel 5.7.8
4 SSD-2 Storage LUN – 3D NAND NVMe SSD x 6; CTS IO Stimulus Generator, CTS Test Software
5 Synthetic Workloads Pre-conditioned to Steady State per SNIA Solid State Storage Performance Test Specification (PTS) v2.0.2
6 Calypso Test Software (CTS), CTS IO Synthetic Workload Generator and IOProfiler Real World Workload IO Capture toolset

付録 B： トランスポートの比較 ― 現実世界ワークロード

Appendix B: NVMe-oF Transport Comparison [1] - Real World Workloads [2,3,4]

	IO Rate IOPS			Bandwidth MB/sec			Average Response Time (ART) mSec			5 9s Quality of Service (QoS) mSec		
Thread Count/Queue Depth Sweep Test [5,6]	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP
GPS Nav Portal [7] QD=72/144	72,465	72,465	72,465	477	423	485	0.990	0.550	1.840	13.500	7.400	21.900
Rtl Web Portal [8] QD=72/144	47,614	47,614	47,614	1,205	1,204	1,191	1.510	1.510	1.500	14.050	13.200	14.150
Storage Cluster [9] QD=96/144	174,666	174,666	174,666	2,962	3,031	3,200	0.460	0.450	0.390	2.600	2.200	2.6540
Replay Test [5,6]	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP
GPS Nav Portal [7] QD=72/144	13,689	13,404	12,625	96	94	94	0.991	1.015	1.036	20.200	18.300	20.300
Rtl Web Portal [8] QD=72/144	30,234	27,187	18,063	320	298	298	9.778	9.970	9.841	50.980	52.615	51.073
Storage Cluster [9] QD=96/144	159,359	160,946	118,713	2,623	2,648	2,648	0.401	0.398	0.553	3.000	3.020	3.279

Notes

1 RDMA iWARP & RDMA RoCEv2 NAC with Offload; TCP NIC - no Offload
2 Back-to-Back NVMe-oF Transport Topology – 100GbE, Intel E810-CQDA2, No Network Switch; MTU Frame Size 1500B
3 Intel Server S2600WF; XEON 8280 2.7 Ghz 28 core single CPU, 198 GB 2166 Mhz DDR 4 ECC RAM, RHEL 8.1, kernel 5.7.8
4 SSD-2 Storage LUN – 3D NAND NVMe SSD x 6; CTS IO Stimulus Generator, CTS Test Software
5 TC/QD Sweep & Replay Tests per SNIA Real World Storage Workload Performance Test Specification (RWSW PTS) v1.0.7
6 Calypso Test Software (CTS), CTS IO Synthetic Workload Generator and IOProfiler Real World Workload IO Capture toolset
7 GPS Navigation Portal – 24-hour IO Capture; 9 IO Stream; 2 min resolution; 720 steps; Drive0; Block IO level; QD Range 6-368
8 Retail Web Portal - 24-hour IO Capture; 9 IO Stream; 5 min resolution; 290 steps; Drive0 & Drive1; Block IO level; QD Range 5-306
9 Storage Cluster - 13-hour IO Capture; 9 IO Stream; 5 min resolution; 158 steps; Drive0,1,2,3,4,5; Block IO level; QD Range 64-1,024