KIOXIA

Performance Brief

KumoScale™ソフトウェアは、Cephソフトウェアと比較して、データセンタースケールでフラッシュパフォーマンスを向上させます

NVMe-oF™プロトコルとの組み合わせにより、ローカル接続ストレージに匹敵するパフォーマンスで分散型ネットワークストレージを実現

NVMe-oF規格は、ローカル接続された[1]NVMe™ SSDに匹敵するパフォーマンスでネットワークストレージをプロビジョニングすることを可能にし、ステートフルアプリケーションをどこでも動作させる柔軟なスケジューリングを可能にします。またネットワーク全体での低レイテンシアクセスと複数のアベイラビリティゾーン間でのデータレプリケーションを実現することができます。これらは通常、最大規模のパブリッククラウドプロバイダーからのみ利用可能な機能です。ハードドライブ用に設計されたプロトコルに比べて、NVMe-oFプロトコルが提供するパフォーマンス上のメリットにより、フラッシュベースのストレージの利点を生かすことが出来ます。

多くのデータセンターでは、Cephストレージを導入し続けています。Cephストレージはオープンソースのソフトウェアストレージプラットフォームで、汎用ハードウェアを使用して分散コンピュータクラスタにオブジェクトストレージを実装します。Cephストレージはデータをオブジェクトに変換し、複数のサーバー間で分散して格納するRADOS（Reliable Autonomic Distributed Object Store）に基づいています。RADOSは分散型のアーキテクチャであり、データの位置や複数のネットワークホップを計算するために、すべての入出力に複数回システムCPUが関与する必要があります。また複数サーバー間でデータの正確性を維持する為に、複数のメッセージのやり取りが必要になります。以上よりCephストレージは、通常のフラッシュストレージ環境において遅くなりがちです。またCephの復旧はターゲット側で行われるため、復旧を完了させるためには他のターゲットそれぞれについて同期した書き込みオペレーションが必要になり、このような場合にも速度が低下することがあります。

NVMe-oFプロトコルは、最新のフラッシュベースストレージ用に設計されているため、以前のハードドライブ ストレージよりも、これらのCephアーキテクチャの限界を改善します。高速で並列化され、高速100～200ギガビットイーサネット（GbE）ネットワーク パスを利用します。NVMe-oFプロトコルの主なメリットの1つは、Linux® オペレーティング システムのI/Oスタックを全体的にバイパスするため、I/O アクセスにはCPU処理がまったく必要ないことです。クライアントがターゲット上のボリュームに接続すると、データパスはNVMe-oFプロトコルのフルスピードで実行され、KumoScaleソフトウェアなどのネイティブNVMe-oFソフトウェア定義ストレージ（SDS）ソリューションが次世代のデータセンターストレージパフォーマンスを提供できるようになります。

キオクシア株式会社（旧東芝メモリ株式会社）が開発したKumoScaleソフトウェアは、高性能ストレージプラットフォームです。ローカル接続された[1]NVMeドライブに匹敵するパフォーマンスを提供する分離型ストレージの仮想化レイヤーと、データセンタースケール向けに設計された仮想化リソース管理の柔軟性を備えています。このソフトウェアは、ストレージリソースのプロビジョニングと管理を自動化して、適切な量の共有され復元力に富んだストレージを個々のワークロードに動的に接続し、分析を行うことでデータセンター全体のストレージ使用率を最適化します。

KumoScaleソフトウェアはCephソフトウェアの代替であり、NVMe-oFプロトコル上にネイティブに構築されたSDSプラットフォームに移行することで、パフォーマンスが向上します。テスト比較[2]では、サーバクラスタとSSDで同一のベンチマークプロセスを使用しました。

テスト環境

KumoScaleソフトウェア（バージョン3.13）とCephソフトウェア（バージョン14.2.11）の両方が、NVMe-oF規格（バージョン1.0a）とTCP/IPトランスポートを含むネットワーク環境でテストされました。テスト環境は3つのストレージノードで構成され、それぞれに5つの容量3.84テラバイト[3]（TB）のKIOXIAエンタープライズCM5シリーズPCIe® Gen3 SSDが含まれており（図1）、ノードあたり最大20 TBのストレージを備えています。

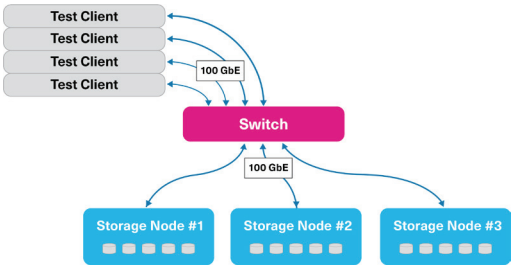

図1：ソフトウェアテスト環境

4つのテストクライアントを用いてテストを実施し性能を測定しました。すべてのストレージノードは、1台の100 GbEネットワークスイッチを介し、100 ギガビット/秒（Gb/s）のイーサネットスピードでテストを行いました。両方のソフトウェアテストに同一のハードウェア構成が使用され、4つの論理 200 GiBボリュームを作成しました。それらは各テストクライアントに 1 つずつ割り当てました。

テストプロセスの一環として、それぞれのデータレプリケーションを次のように設定しました。

KumoScaleソフトウェアの場合、ボリュームは3重に複製されました（ひとつの論理ボリュームに対し一つのレプリカを設定。それを3台のストレージノードに割り当てる）。

Cephソフトウェアの場合、ボリュームは「CRUSH」ハッシュ方式を使用してシャーディング（水平分割）され、レプリケーション計数は3に設定されました。

テストの説明

KumoScaleとCephソフトウェアの両方で使用されるテストプロセスには、次のワークロードの概要と予想される動作が含まれていました。

ワークロード指標:

- *100%ランダムの4kB読み取り、および100 %ランダムの4kB書き込み*
- *未処理の読み取りコマンドの総数（キュー深度）を制御することにより、負荷を変化させる*
- *キュー深度に対するレイテンシ（平均および99パーセンタイル）と1秒あたりの入出力数（IOPS）を記録*

想定される動作:

- *最初にキュー深度が増加すると、IOPS数は直線的に増加し、レイテンシは一定のままであるか、ゆっくりと増加する。その結果、負荷はストレージシステムに固有の並列処理によって処理される。*
- *あるキュー深度で、IOPS数の上昇が止まる一方、レイテンシが増加し始める。これは、ストレージシステムのパフォーマンスが最大に達したことを示しており、より多くの読み取りコマンドをキューに入れると単に待機時間が長くなる。*

免責条項：実施されたパフォーマンステストとは異なり、データセンターの負荷は通常、未処理のキュー深度によって制限されず、任意の数値でコマンドを発行します。この数値がストレージシステムの最大IOPS能力を超えると、発行されたコマンドは急速に失敗し始めます。データセンターのオペレータは、ストレージシステムが最大IOPSを下回って運用されるようにしなければなりませんが、最大の資本効率を達成するためには、その最大能力に近い負荷をかける必要があります。

テスト実施結果: IOPS

KumoScaleおよびCephソフトウェアのIOPSパフォーマンステストの結果（図2）は次のとおりです。

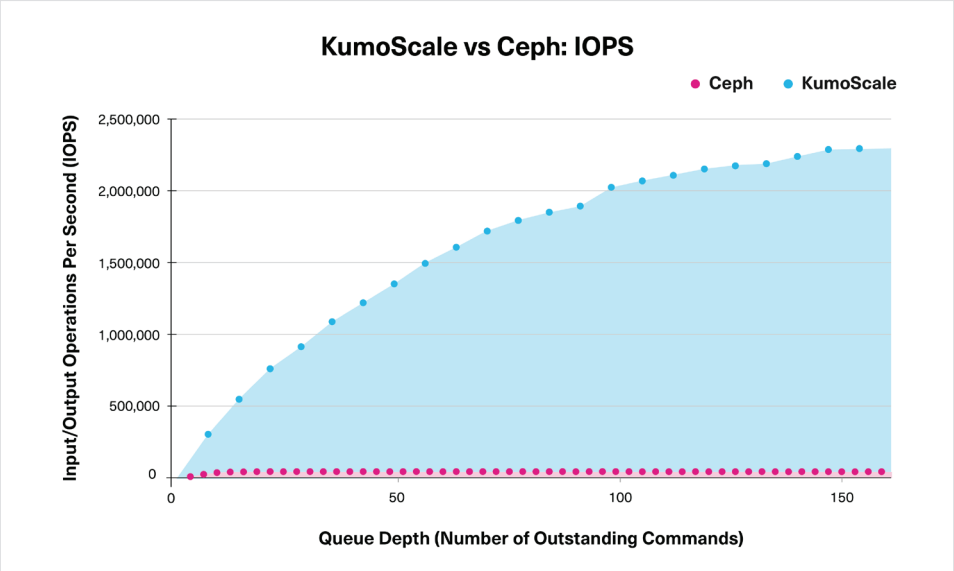

図2：KumoScaleとCephのIOPSテスト結果

Copyright © 2020 KIOXIA Corporation. All rights reserved.

KumoScaleソリューションは、同一のハードウェア上でCephソフトウェアの52倍を超えるIOPSパフォーマンスを実現しました（表1）。このパフォーマンス指標の場合は、高いほど優れています。

テスト指標	KumoScale	Ceph	KumoScale Advantage
最大負荷でのIOPS	2,293,860	43,852	>52x

表1：KumoScaleソフトウェアとCephソフトウェアを比較したIOPSパフォーマンステスト結果

テスト実施結果: 読み取りレイテンシ

KumoScaleおよびCephソフトウェアの読み取りレイテンシテストの結果（図3）は次のとおりです。

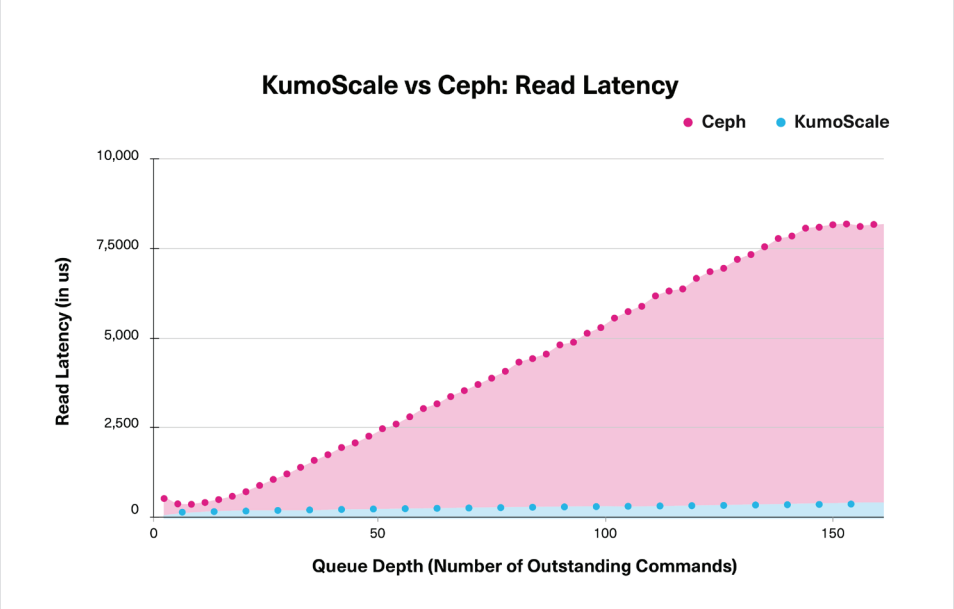

図3：KumoScaleとCephの読み取りレイテンシテスト結果

キュー深度が150の時に、KumoScaleソリューションの読み取りレイテンシはCephソフトウェアより22倍以上高速でした（表2）。このレイテンシ指標では、低いほど優れています。

テスト指標	KumoScale	Ceph	KumoScale Advantage
最大負荷での読み取りレイテンシ	369µs	8,169µs	>22x

表2：KumoScaleソフトウェアとCephソフトウェアを比較した読み取りレイテンシテスト結果

KumoScaleソリューション:

- *キュー深度が150で、最大229万IOPSに到達*
- *キュー深度が150で、読み取りレイテンシは369usで安定状態を保持*

Cephソフトウェア:

- *キュー深度が150で、最大43.8kIOPSに到達*
- *キュー深度が150で、読み取りレイテンシは8,169 us前後で推移*

Copyright © 2020 KIOXIA Corporation. All rights reserved.

テスト実施結果: 書き込みレイテンシ

KumoScaleおよびCephソフトウェアの書き込みレイテンシテストの結果(図4)は次のとおりです。

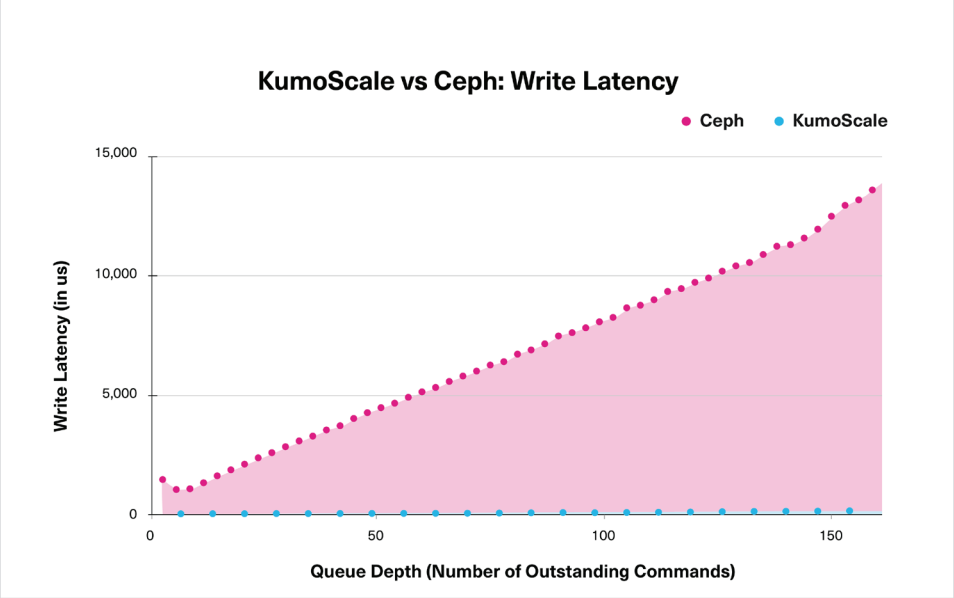

図4: KumoScaleとCephの書き込みレイテンシテスト結果

KumoScaleソフトウェアの書き込みレイテンシは、Cephソフトウェアよりも81倍高速でした(表3)。このレイテンシ指標では、低いほど優れています。

テスト指標	KumoScale	Ceph	KumoScale Advantage
最大負荷での書き込みレイテンシ	167µs	13,610µs	>81x

表3:KumoScaleソフトウェアとCephソフトウェアを比較した書き込みレイテンシテスト結果

KumoScaleソリューション:

- *キュー深度が150で、最大229万IOPSに到達*
- *キュー深度が150で、書き込みレイテンシは167usで安定状態を保持*
- *書き込みパフォーマンスは、SSDを5台のみ搭載しているストレージノードでテストしました。これらのKIOXIA CM5 SSDドライブは、それぞれ160kの書き込みIOPSをサポートし、5つのドライブで合計800kを実現します。今回のテストでは最大負荷で2.293kが達成されましたが、全24台のドライブでは500万から600万の書き込みIOPSに達すると予想されます。*

Cephソフトウェア:

- *キュー深度が最大150で、最大43.8kIOPSに到達*
- *キュー深度が最大150で、書き込みレイテンシは12,509usであり、増大*
- *書き込み確認に非常に時間が掛かりレイテンシに悪影響を及ぼす結果として、最大書き込みスループットが非常に低い*

経済的影響

読み取りおよび書き込みパフォーマンス テストの結果から、1つのストレージノードがサポートできるユーザー数と言う観点で、データセンターにおける経済的影響を評価できます(表1を参照)。ストレージノードあたりのユーザー数を増やすことは、ユーザーあたりの設備投資の削減に直接つながります。上記のIOPSの結果に基づくと、KumoScaleソフトウェアは、Cephソフトウェアよりもストレージノードあたり約52倍多くのクライアントをはるかに低いレイテンシでサポートでき、購入するストレージノードの数を大幅に削減できるため、データセンターへの投資収益率を向上できます。

Copyright © 2020 KIOXIA Corporation. All rights reserved.

まとめ

この比較では、KumoScaleソフトウェアのパフォーマンス能力を明確にし、この高いパフォーマンスによってデータセンターユーザにもたらされる経済的なメリットを示しています。このソリューションは、NVMe-oF規格のパフォーマンスとスループットにおけるアーキテクチャ上のメリットを最大限に活用し、データセンタースケールの仮想化、自動化、プロビジョニング、管理、およびストレージリソースの最適化をすることで、Cephソフトウエアに対してほぼ52倍の入出力処理が可能です。これにより運用コストを削減し、顧客満足度を向上させることができます。

KumoScaleソフトウェアの詳細については、https://kumoscale.kioxia.com/jp にアクセスしてください。

KUMOSCALE™

Notes:

[1] 2020年9月時点のNVMe SSDベンダーによる公開済みパフォーマンス仕様に基づいています。

[2] パフォーマンス比較テストは、2019年10月にKIOXIA Corporationによって実施され、同一のテストプロセスと同一のサーバークラスタ、5つの同一のKIOXIAエンタープライズCM5シリーズPCIe Gen3 SSD 4TB容量を使用しました。

[3] 記憶容量 - 1MB(1メガバイト)=1,000,000(10の6乗)バイト、1GB(1ギガバイト)=1,000,000,000(10の9乗)バイト、1TB(1テラバイト)=1,000,000,000,000(10の12乗)バイトによる算出値です。
しかし、1Gbit = 1,073,741,824(2の30乗)ビット、1GB=1,073,741,824(2の30乗)バイト、1TB = 1,099,511,627,776(2の40乗)バイトによる算出値をドライブ容量として用いるコンピューターオペレーティングシステムでは、記載よりも少ない容量がドライブ容量として表示されます。使用可能なストレージ容量(各種メディアファイルを含む)は、ファイルサイズ、フォーマット、セッティング、ソフトウェアやオペレーティングシステムおよび/またはプリインストールされたソフトウェアアプリケーション、またはメディアコンテンツおよびその他の要因で変わります。実際のフォーマット済み容量は異なる場合があります。

Linuxは、米国およびその他の国におけるLinus Torvaldsの商標または登録商標です。NVMeおよびNVMe-oFは、NVM Express, Inc.の米国またはその他の国における登録商標または商標です。PCIeは、PCI-SIGの登録商標です。その他記載されている社名・商品名・サービス名などは、それぞれ各社が商標として使用している場合があります。

© 2020 KIOXIA America, Inc. All rights reserved. 無断複写・転載を禁じます。このパフォーマンスブリーフに記載されている製品仕様、テスト内容、評価などの情報は、本書の発行日現在のものであり、ドキュメントが公開された日付時点で正確であると考えられていますが、予告なく変更される場合があります。ここに記載されている技術情報およびアプリケーション情報は、最新の該当するKIOXIA製品仕様に従います。