ユーザー空間におけるオープンソースのストレージスタックのビルド

この記事は、インテル® デベロッパー・ゾーンに公開されている「Build Open Source Storage Stacks in User Space」の日本語参考訳です。

概要

ストレージ要件は年々飛躍的に増加しており、その要件を満たすためにストレージ・ハードウェアも急速に進歩しています。3D XPoint™ テクノロジーのような次世代のハイパフォーマンス・テクノロジーでは、レイテンシーはミリ秒ではなく、ナノ秒で測定されます。次世代ストレージでレイテンシーの桁が変わることから、I/O ボトルネックがストレージメディアからネットワークやソフトウェアへ移行し、ユーザーモードのクラウドストレージの開発が必要になることが予想されます。Containers*、Ceph*、Swift*、Hadoop* のような多くの商用プロジェクトやプロプライエタリー・ソフトウェアはユーザー空間ですでに動作していますが、オープンソースのストレージシステム向けビルディング・ブロックの開発には、まさにうってつけの時期と言えます。

はじめに

ストレージのパフォーマンスで考慮すべき主なポイントは、レイテンシーとスループットの 2 つです。レイテンシーはデータ移動中の遅延時間を指し、スループットは指定時間内に転送されるデータの量を指します。

ストレージ・アーキテクチャーは、メディアのさまざまなプロパティーを最適化する必要があります。例えば、HDD には考慮すべき 3 つの重要な特性があります。

1. ドライブおよびストレージ・サブシステムのバッファーでは、キャッシュレイヤーが必要になるため、読み取りおよび書き込みのレイテンシーは大きくなります。
2. ランダムではなくシーケンシャルな読み取りおよび書き込みには、それほどコストはかかりません。
3. I/O パフォーマンスは、スピンドルの回転速度および数によって制限されます。
4. 機械的なドライブおよび分離されたコントローラーはスループットを低下させます。

これらのハードウェアおよびアーキテクチャー上の制限により、ストレージを最適化するには追加の設計が必要になります。

1. 割り込みベースの I/O モデルはより効率的です。
2. HDD キャッシュに DRAM を使用すると、読み取りとシーケンシャル書き込みのレイテンシーが軽減されます。
3. SSD に DRAM を使用すると、ハードウェア・エラーが減少し、アクセス時間が短くなります。
4. プロセッサーとストレージメディア間のデバイスの数を減らすと、レイテンシーが軽減されます。

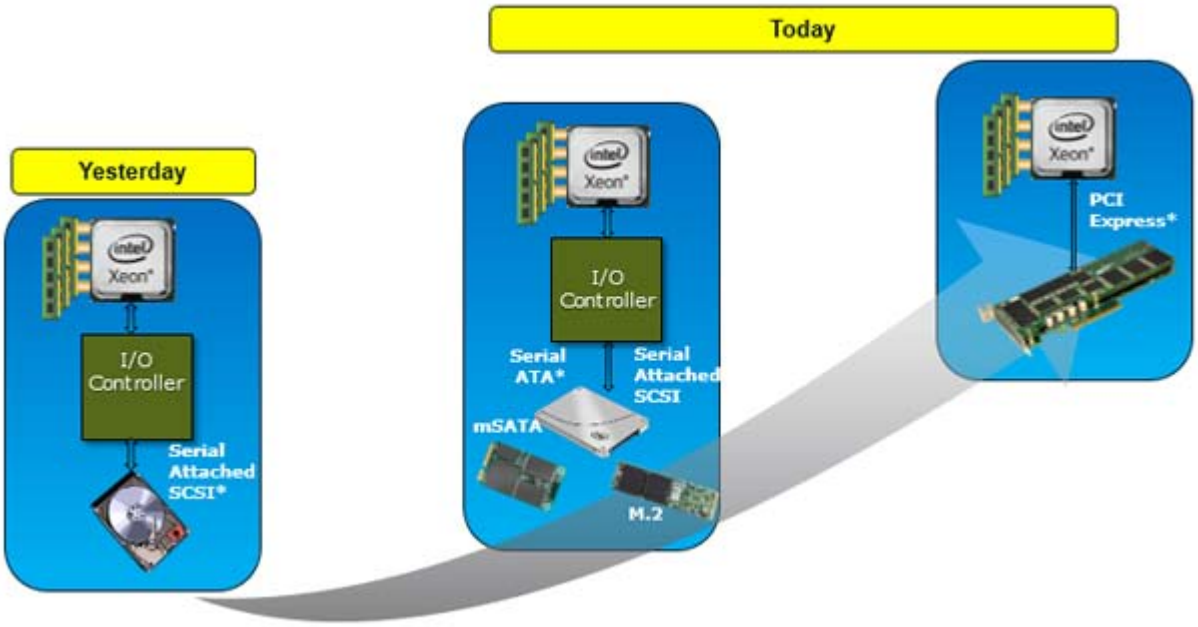

図 1 – パフォーマンスの向上とともにデータはプロセッサーへ接近

ハードウェアのブレークスルー

どうすればより高いパフォーマンスとより低いレイテンシーを両立できるのでしょうか。まず、ハードウェア (特に密度と電力要件および CPU へのホップ数) の改良から見ていきましょう。

ハードウェア側のさまざまなブレークスルーにより、より少ないレイテンシー、より高いスループット、より低い電力要件、より高い密度でパフォーマンスは向上しています。

- プロセッサーに統合型 NVMe (Non-volatile Memory Express) インターフェイスを使用することにより、外部コントローラーを排除 (1 ホップから 0 ホップへ)。
- 不揮発性メモリーへの移行により、データの保持に必要な電源を排除。
- 3D XPoint™ メモリー (最大で密度が NAND の 10 倍、速度が NAND の 1,000 倍) を使用できる、インテル® Optane™ テクノロジーを備えた新しいタイプの SSD の登場。

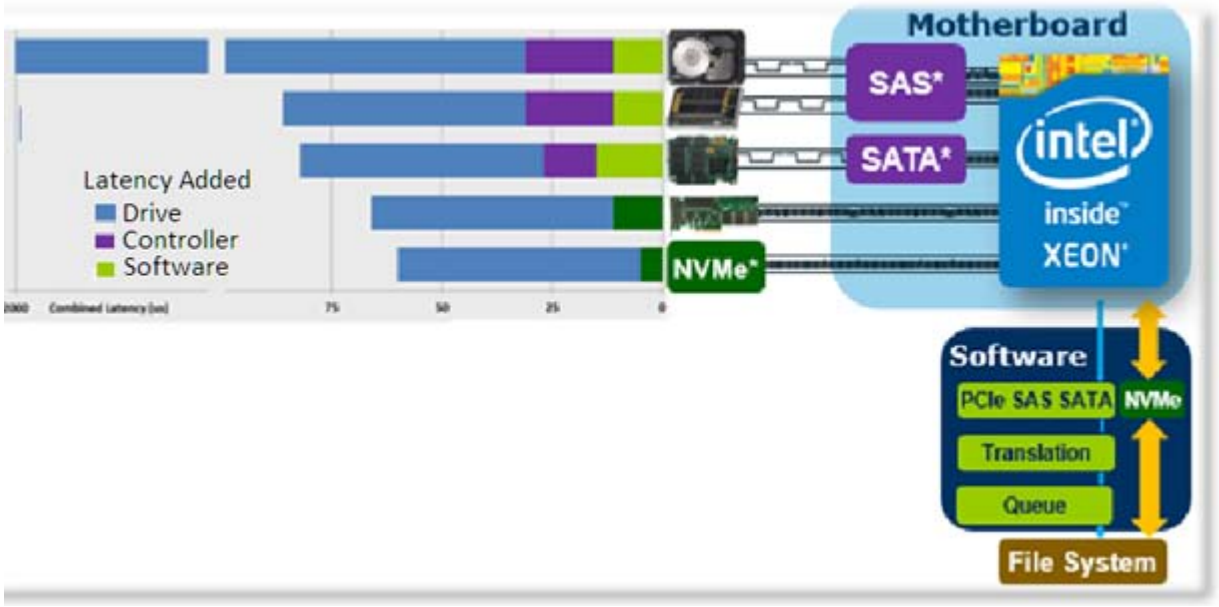

図 2 – ストレージのパフォーマンスとレイテンシーの向上

図 2 のように、コントローラーやソフトウェアの遅延により、15,000 rpm の HDD では約 2,000 µs のレイテンシーがあります。SSD では、機械的な動作による遅延がなくなり、コントローラーを CPU に移動することにより余分な手順が省略され、継続的なソフトウェアの改良によりレイテンシーが大幅に軽減されました。

オープンソースのユーザー・モード・ストレージ・スタックを利用する

改良されているのはハードウェアだけではありません。カーネルモードを使用する代わりにユーザー空間で次世代のストレージ・アプライアンスとアーキテクチャーを開発する理由はいくつかあります。

理由 1: 信頼性、可用性、保守性 (RAS) の向上

- カーネル・モード・ドライバーはシステムをリブートしないと再起動できませんが、ユーザー・モード・ドライバーはリブートすることなく再起動できます。
- ユーザー・モード・ドライバーを利用することにより、データ収集とロギングをより柔軟に行うことができます。

理由 2: ユーザー空間コードのライセンスの容易さ

- Storage Performance Development Kit (SPDK) は、NVMe および CBDMA ドライバーについては Berkeley Software Distribution License、ユーザー空間ネットワーク・スタックおよび iSCSI ターゲットについてはインテルのプロプライエタリー・ライセンスの下でそれぞれライセンスされているため、プロプライエタリー製品およびオープンソース製品の両方で自由に使用することができます。

理由 3: 効率

- 技術的な理由: カーネルの割り込み制御やコンテキスト・スイッチのオーバーヘッドはプロセッサーのリソースを浪費しますが、SPDK NVMe ポーリングモード・ドライバー・アーキテクチャーは効率的なパフォーマンスを提供し、シングル・プロセッサー・コアで大量の IOPS を処理できます。

- 大きなページが使用されると、TLB (トランスレーション・ルックアサイド・バッファー) ミスが発生し、ページウォークが行われます。

理由 4: 設計の簡素化

- 低レイテンシーの NVM メディアには、バッテリーバックアップ DIMM は必要ありません。
- ユーザー空間ドライバーは、サーバーベースのストレージやソフトウェア定義のアーキテクチャーに直接適用することができます。
- 物理メディアのレイテンシーが低くなると、要件の最も高いワークロードを除いて大きな DRAM キャッシュを使用する必要性はなくなります。

まとめ

インテル® Optane™ テクノロジーのような新しい NVM テクノロジーの潜在的なパフォーマンスを引き出せるように、開発者がユーザー空間で作業できるようにすると、クラウドストレージ開発者のパフォーマンスは向上し、開発は単純化されます。インテルは、SPDK を備えたストレージ環境にオープンソース Data Plane Development Kit (DPDK) のハイパフォーマンス・パケット処理フレームワークを適用します。SPDK は、カーネルへのコンテキスト・スイッチが必要な関数の Linux* ユーザー空間を含む、1 セットのドライバーと完全なリファレンス・ストレージ・アーキテクチャーです。 DPDK と SPDK はどちらもオープンソースとして利用可能です。

参考文献 (英語)

- The New Breakthrough Memory 3D XPoint™ Technology Delivers
- A Storage Framework For Cloud Storage Developers
- Introducing Intel Optane™ Technology – Bringing 3D XPoint™ Memory to Storage and Memory Products
- インテル® SSD と HDD のビデオ
- Data Plane Development Kit
- 年表: ハードドライブの 50 年

この記事の作成に寄与いただいた Jonathan Stern、Debra Graham、Mike Pearce、Colleen Culberson、Victoria Davis の各氏に感謝します。

著者紹介

Thai Le は、ソフトウェア・エンジニアとして、インテル コーポレーションで 20 年以上にわたって、ソフトウェアの自動化、サーバーの電力およびパフォーマンス解析、クラウド・コンピューティングに取り組んでいます。