ソリューション概要

# Yahoo!* がインテル I® CAS 3.0 でスケールアウト・ストレージのパフォーマンスを高速化

## インテルとの導入ソリューションの探索プログラムが画期的なソリューションをもたらす

「 Ceph* への移行により、当社のストレージの費用を削減してスケーラビリティーを高めることができました。PCIe* 対応インテル ® SSD データセンター・ファミリーにインテル ® CAS 3.0 を加えたことでパフォーマンスを飛躍的の向上でき、Ceph* を使って当社のウォームデータとホットデータを処理できるようになりました。」

**– Ruiping Sun、シニア・プリンシパル・アーキテクト、Yahoo!*[1]**

### はじめに

ストレージの費用の高騰と急速なスケーラビリティーの要求により、多くの企業は Ceph* や Swift* などのオープンソースのスケールアウト・ストレージ・ソリューションを取り入れています。これらのソリューションは多数の利点を提供しますが、その反面パフォーマンスの課題がもたらされます。Yahoo!* との導入ソリューションの探索プログラムを通じて、インテルは企業がこれらの課題の多くを克服できるようにする注目せずにいられないソリューションを開発しました。

インテル ® キャッシュ・アクセラレーション・ソフトウェア (CAS) 3.0 と PCIe* 対応インテル ® SSD データセンター・ファミリーの組み合わせが Yahoo!* のユーザー体験を向上させました。スループットが 2 倍になり、レイテンシーが最高 75 % まで減少し、エラー修復時間が 70 % 縮減され、Ceph* 環境で修復中のパフォーマンスの影響が 50 % 削減され、これらはすべてデータの移行またはストレージ・インフラストラクチャーに大きな変更を行う必要なしに実現されました。[2]

高度のパフォーマンスを提供するこのスケールアウト・ストレージ・ソリューションは、Intel® Labs およびインテル ® 不揮発性メモリー・ソリューション・グループ (NSG) による草分け的な作業に由来し、他の企業およびソフトウェア・デファインド・ストレージ・エコシステム全体に幅広く適用できます。

### ストレージの増加により課題が出現

デジタルの世界は 2 年ごとに 2 倍のサイズになり、2020 年までに 44 ゼタバイトに増えることが予測されています。[3] データのこの急激な増加に遅れをとらないように、企業はデータストレージへより多くの IT 予算をつぎ込み、これにより IT への支出が年間 2% から 4% 増加する結果となっています。[4]

Yahoo!* の課題は、このジレンマを例証しています。毎日 10 億人以上が Yahoo!* Mail、Flickr*、Tumblr*、その他の Yahoo!* の無料アプリケーションを使っています。同社は、ストレージの費用を維持したり、ときに削減したりしながら、毎年飛躍的により多くのデータを保管せねばならないため、これらのアプリケーションでデータをすばやく効率的に処理するのは重大な課題となります。

Yahoo!* を含む世界中の多数の企業用のソリューションは、従来の NAS および SAN ソリューションからソフトウェア定義型のスケールアウト・ストレージ・ソリューションに移行しました。2014 年に Yahoo!* は同社のストレージ要件を Ceph* に移行して、より優れたスケーラビリティーと資本コストの 50 % 削減を含む典型的なスケールアウト・ソリューションの多数の利点につながりました。[1]

**インテル ® キャッシュ・アクセラレーション・ソフトウェア (CAS) 3.0**

インテル ® CAS 3.0 は、カーネルをじっくり調べ、詳細分析を実行し、事前決定された分類に基づいて I/O を分類し、キャッシュから除去されるアイテムの順序の優先度を指定するためのユニークでインテリジェントな方法を提供します。インテル ® CAS の他の利点には以下のものが含まれます：

- ユーザーとアプリケーションに透明
- すべてのハードドライブを SSD に置き換えたのに近いほど、高度なレベルのパフォーマンス
- 主要アプリケーションへの重大時における追加のパフォーマンス
- 特定の作業負荷と環境の最適化のための数種類のキャシングモードと機能
- インテル ® CAS ソフトウェアのライセンスは PCIe* 対応インテル ® SSD データセンター・ファミリーに含まれています。

しかし、Ceph* や Swift* などのオープンソースのスケールアウト・ストレージ・ソリューションは、パフォーマンスの課題をもたらすこともあります。特に Yahoo!* は Ceph* では 1 年以上前のメールの添付ファイルなどのコールドデータ ( めったにアクセスしないデータ ) ではパフォーマンスが優れていることを見い出しました。けれども、同日のメールの添付ファイルなど、IOPS 集中のホットデータの処理を試みると、スループットとレイテンシーが重大な問題をもたらしました。

ストレージの費用を押さえて、業界をリードする同社のアプリケーションで期待される即座のアクセスをカスタマーに提供するために、Yahoo!* はウォームデータとほっとデータをより速く、より効率的に処理する方法を見つける必要がありました。

## 導入ソリューションの探索プログラムが画期的なソリューションをもたらす

インテル・ラボは広範囲にわたる技術革新で可能性の新境地を開くインテルのリサーチ部門です。2008 年以来インテル・ラボは企業のストレージシステムにおけるキャッシングの改善方法について研究を続けており、DSS (Differentiated Storage Services) の発明を含む進歩に至りました。[5] 2014 年後半に Yahoo!* の役員達がインテル ® デベロッパー・フォーラムで DSS を知り、間もなく導入ソリューションの探索プログラムに従事して、同社が Ceph* で経験しているパフォーマンスの課題のいくつかを解決するために、DSS テクノロジーの可能性を探りました。

DSS の潜在的な利点は、Yahoo!* がストレージの費用を削減し、ストレージの使用率を高めるために同社の Ceph* 環境で使用するイレージャー・コーディングを詳しく見てみるとわかります。イレージャー・コーディングは、データを分散型ストレージ環境全体で保管される断片に分割します。全体的なパフォーマンスと顧客体験は、ファイルとディレクトリについての情報を保管する何千もの i ノードを急速に検索して、いくつかのファイルシステムのメタデータのアクセスを行い、データの断片を見つけて、できるだけ速く各オブジェクトを再構成することに依存します。最も遅いディスク I/O によって、パフォーマンスが妨げられます。

DSS には、キャッシュで発生する I/O のタイプにより細かい洞察を提供することによって、イレージャー・コーディングを速めることができる I/O 分類機能が含まれます。たとえば、ブロックにファイルシステムのメタデータ ( 例えば、スーパーブロック、i ノード、ディレクトリのエントリ ) が含まれる場合または通常ファイルの一部である場合、DSS がそれを示すことができます。このように I/O を分類すると、DSS がストレージ・トラフィックを分析して、キャッシュ内の異なるアイテムに優先順位を付けて、必要に応じて降順の優先順でアイテムを破棄できます。

数か月に渡りインテル・ラボは、DSS を最適化して Yahoo!* の特定のストレージの課題に対処するために、Yahoo!* と密接に作業を行いました。やがて、インテル ® NSG の助けを得て DSS テクノロジーはインテル ® CAS 3.0 になったものに組み込まれました。PCIe* 対応インテル ® SSD データセンター・ファミリーと組み合わせたときに、インテル ® CAS 3.0 により Yahoo!* は同社のデータの優先度を指定してデータを選択的にキャッシュして、イレージャー・コーディングと他のストレージの処理に劇的なパフォーマンスの向上をもたらすことができました。

インテル ® CAS 3.0 は PCIe* 対応インテル ® SSD データセンター・ファミリーのドライブで利用可能な NVMe* ストレージ・インターフェイス規格に特に最適化されています。社内テストでは、ハードドライブと SATA SSD の両方に比べて、NVMe* 設計では、はるかに優れたスループットと低いレイテンシーが得られることが見られました。

## Yahoo!* がストレージのパフォーマンスで大規模な改善を達成

インテル ® CAS 3.0 と NVMe* 対応 インテル ® SSD DC P3600 1.6TB を組み合わせることにより、Yahoo!* は Ceph* 環境でスループットを 2 倍に向上させレイテンシーを 75% まで削減しました。[2] Yahoo!* は、各 8TB HDD に対してインテル ® CAS ソフトウェアを搭載したインテル ® SDD に 50GB だけ追加することによって、このパフォーマンスの改善を達成しました。さらに、パフォーマンスの改善は大規模なデータの移行または基礎をなすストレージのインフラストラクチャーへの変更を必要とせずに実現され、ソリューションは必要に応じて規模を変更する余地があり、コストを最適化しています。

さらにインテル® CAS 3.0 は、Yahoo!* が予期されないデータの損傷、ハードドライブまたは他のシステムの障害からのリカバリー時間を 70% 削減するのに役立ちました。リカバリー中のパフォーマンスへの影響も半分に削減され [2]、リカバリー中のカスタマーの体験が大幅に改善される結果となりました。

## 読み取り要求のレイテンシー

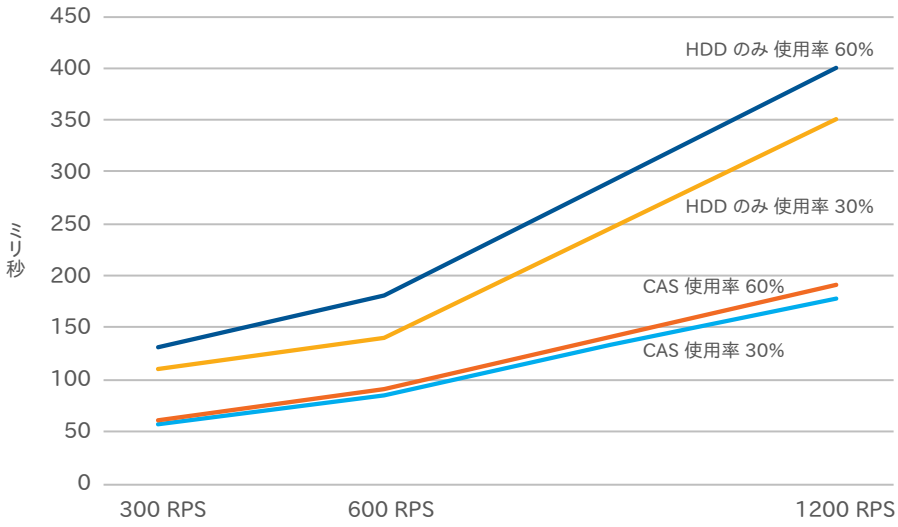


図 1. インテル® CAS により強化されたシステムでの読み取りレイテンシーは、HDD のみのソリューションのレイテンシーの約半分です。[2]

## 書き込み要求のレイテンシー

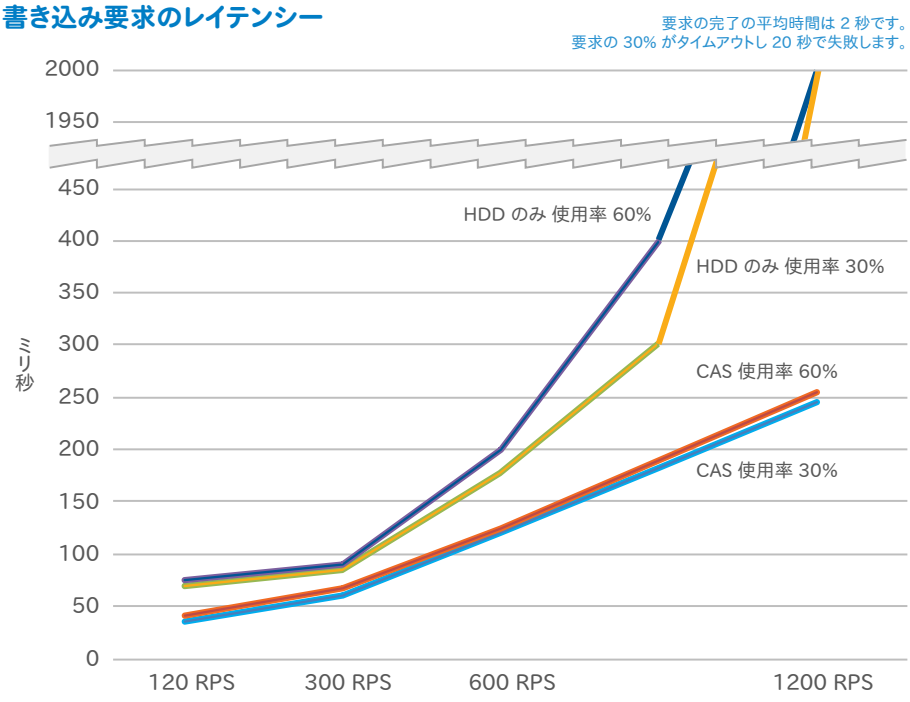


図 2. インテル® CAS によって強化されたシステムで、すべての書き込み要求が完了し、システムのストレージ容量がいっぱいになったときでさえ、レイテンシーが増加しません。[2,6]

### Non-Volatile Memory Express* (NVMe*) の利点

インテルは、PCI Express* (PCIe*) SSD で利用可能な標準化された高性能ソフトウェア・インターフェイス NVMe* の作成で業界をリードしました。NVMe* は SAS および SATA SSD のパフォーマンスの限界を克服し、SSD をより効率的でスケーラブルにして管理しやすくするために新たに設計されました。マルチコア CPU のスケールと I/O ごとの低クロックサイクルを提供するための合理化されたプロトコルとより効率的なキューイング・メカニズムが特徴の 1 つです。インテル® CAS は PCIe* 対応インテル® SSD データセンター・ファミリーのハードウェア機能を最大限に活用するために設計され最適化されました。

Yahoo が展開したソリューションは Ceph* 向けに設計されていますが、ベアメタル・ストレージ・システムとソフトウェア・デファインド・ストレージ・システム ( 例：Swift* および Lustre*) に幅広く適用できます。ソリューション全体をインテルから 1 つの部品番号を使って購入し、ライセンスの配布を簡易化することができます。

## まとめ

企業はクラウドストレージの要求における飛躍的な増加を経験し、多数の企業が費用を削減しスケーラビリティーを向上するのに役立つオープンソースのスケールアウト・ストレージ・ソリューションを取り入れています。Yahoo!* との導入ソリューションの探索プログラムで示すように、NVMe* 対応インテル ® SSD データセンター・ファミリーに搭載されたインテル ® CAS 3.0 は Ceph* のようなスケールアウト・ストレージ・ソリューションの利点を拡張して、高価なインフラストラクチャーの全面的な見直しを必要とせずにパフォーマンスと復元の可能性に劇的な改善を提供します。

**詳細情報：**
www.intel.com/cas
www.intel.com/nvme
www.intel.com/ssd

[1] www.youtube.com/watch?v=vtIIbxO4Zlk

[2] Yahoo!* 社内計測に基づく、2015 年。1 つの 600 OSD Ceph* クラスター、3PB ストレージ、イレージャー・コーディング 8+3、(10) 8T SATA ディスク、1MB オブジェクトサイズ、インテル ® CAS 3.0、インテル ® P3600 1.6 TB NVMe* SSD、単一の読み取りと書き込みを使用。テストはインテル ® CAS/SSD コンポーネントを使用して、あるいは使用せずに実行されました。
クラスターの各 OSD ノードには、サーバー：HP ProLiant DL180 G6 ySPEC 39.5、CPU：2x Xeon X5650 2.67GHz (HT 対応、合計 12 コア、24 スレッド )、チップセット：インテル ® 5520 IOH-36D B3 (Tylersburg)、RAM：48GB 1333MHz DDR3 (12 x 4GB PC3-10600 Samsung DDR3-1333 ECC 登録 CL9 2Rx4)。HDD：(10) 8TB 7200 RPM SATA HDD、ネットワーク：(2) HP NC362i/ Intel 82576 Gigabit、(2) Intel 82599EB 10Gbe、OS：RHEL 6.5、カーネル 3.10.0-123.4.4.el7 が含まれます。システムの各ノードに 1 台の SSD を追加したインテル ® CAS 3.0 の構成：SSD：(1) 1.6TB インテル ® P3600 SSD (OSD, 1.5TB キャッシュごとに 10GB ジャーナル )。

[3] www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[4] Gartner (2014 年 6 月 )。 www.gartner.com/newsroom/id/2783517

[5] www.sigops.org/sosp/sosp11/current/2011-Cascais/printable/05-mesnier.pdf

[6] 書き込みレイテンシーは負荷の軽いシステムで 40% 低く、負荷の重いシステムでは 90% 低く、平均すると約 75% 低くなっています。

テストでは、特定のシステムでの個々のテストにおけるコンポーネントの性能を文書化しています。ハードウェア、ソフトウェア、システム構成などの違いにより、実際の性能は掲載された性能テストや評価とは異なる場合があります。
購入を検討される場合は、ほかの情報も参考にして、パフォーマンスを総合的に評価することをお勧めします。

インテル ® テクノロジーの機能と利点はシステム構成によって異なり、対応するハードウェアやソフトウェア、またはサービスの有効化が必要となる場合があります。実際の性能はシステム構成によって異なります。絶対的なセキュリティーを提供できるコンピューター・システムはありません。詳細については、各システムメーカーまたは販売店にお問い合わせいただくか、www.intel.co.jp を参照してください。



0316/tc/rt 334106-001JA