インテル Enterprise Edition 2.3 for Lustre 搭載のHPC用 Dellストレージ

Dell テクニカルホワイトペーパー

Quy Ta

Dell HPC エンジニア

2015年10月 | バージョン 1.0

このホワイトペーパーは情報提供のみに使用するものとし、記載内容の誤りや技術精度を保障するものではありません。本コンテンツはDellによって「現状のまま」提供されており、明示黙示を問わずいかなる保証もありません。

© 2015 Dell Inc. All rights reserved. 本サイトに含まれる資料の複製は、Dellからの書面による事前承諾を得ずに使用することは硬く禁じられています。詳細情報はDellへお問い合わせください。

「Dell」、「DELL」のロゴ、「 DELL」の バッジ、 PowerConnect、 PowerVault はDellの商標です。本資料で使用されているその他の商標は他社のものであるか、もしくは他社製品名称に使用されています。Dellは自社商標以外の商標、または商標名における所有の権益を放棄するものとします。

コンテンツ

図 iv
表 v
1. 紹介 1
2. Lustre ファイルシステム 1
3. インテル EE for Lustre搭載のHPC用 Dellストレージとは 3
3.1 マネジメントサーバー 5
3.2 メタデータサーバー 5
3.3 オブジェクトストレージサーバー 7
3.4 スケーラビリティ 8
3.5 ネットワーク 9
3.5.1 管理ネットワーク 9
3.5.2 データネットワーク 9
3.6 インテル EE for Lustre搭載のHPC用 Dellストレージの管理 10
4. パフォーマンス評価と分析 11
4.1 N-to-N 連続読み込み/書き込み 14
4.2 ランダムリードとライト 16
4.3 メタデータテスト 18
5. 結論 21
補足資料 A: ベンチマークコマンドレファレンス 23
参考資料 24

図

図 1 : Lustre を基盤としたストレージソリューションのコンポーネント..... 2
図 2 : インテル EE for Lustre搭載のHPC用 Dellストレージ コンポーネント概要 4
図 3 : Lustre DNE オプション 4
図 4 : Dell PowerEdge R630 5
図 5 : メタデータサーバーペア..... 6
図 6 : Lustre DNEオプション搭載のメタデータサーバーペア 6
図 7 : オブジェクトストレージサーバーペア 7
図 8 : MD3460, MD3060e アレイ上のRAID6レイアウト 8
図 9: OSS スケーラビリティ 9
図 10: インテル Manager for Lustre (IML) インターフェース 10
図 11 : インテル EE for Lustre Solution搭載のHPC用Dellストレージによる連続読み込み／書き込み ... 14
図 12: SAS ドライバー 15
図 13 : ストレージ書き込みキャッシュ..... 16
図 14 : N-to-N ランダムリード／ライト 17
図 15 : ランダムライトにおける書き込みキャッシュの影響..... 17
図 16 : ファイルメタデータオペレーション 19
図 17 : ディレクトリメタデータオペレーション 20
図 18 : MDテスト Lustre DNEによるファイル作成オペレーション 21

表

表 1：テストクライアントクラスターの詳細 11
表 2：インテルEE for Lustre搭載のHPC用Dellストレージ構成 13
表 3：MDテストで使用したパラメータ 18

1. 紹介

ハイパフォーマンス・コンピューティング（HPC）は、コンピュートノードへの効率的なデータ転送がきわめて重要でありながら、その実行は複雑です。研究者によってHPCシステムからのデータ生成や収集を素早く行うことができるようになりましたが、それによりストレージコンポーネントがシステムのボトルネックとなるようになりました。そのような複雑なストレージシステムを管理・監視するのはストレージ担当者や研究者に負荷をかけてしまいます。データに要求されるパフォーマンスと容量に対する需要は急速に増しています。ストレージシステムのスループットやスケーラビリティが増えると、ハイパフォーマンス・コンピューティングシステムに多くのプランニングや構成が必要となります。

本資料内で記載しているインテル® Enterprise Edition (EE) for Lustre® Solution 搭載のHPC用Dellストレージ（以下、インテル EE for Lustre搭載HPC用 Dellストレージ）は学術関連、または業界関連者用に設計されており、フルサポートで、使用が簡単で、情報量が豊富で、スケールアウト可能であり、さらに費用対効果のある並列ファイルシステムストレージソリューションが求められています。インテル EE for Lustre搭載のHPC用 Dellストレージは、スケールアウトストレージソリューションアプライアンスで、高い性能と高可用性を持つストレージシステムを実現します。インテリジェントで幅広く、直感的に扱えるインターフェイスを活用し、インテルManager for Lustre（IML）のソリューションは、すべてのハードウェアとストレージシステムコンポーネントの導入、管理、監視を大幅に簡易化します。容量のスケーリングや性能、またはその両方を備えるのは非常に簡単ですので、将来的な拡張に向けた利便性の高い道筋を提供することができます。ストレージソリューションは業界をリードするHPCオープンソースの並列ファイルシステムであるLustreを使用しています。[1]

ストレージソリューションは第13世代エンタープライズ向け Dell PowerEdge™ サーバーと最新の高密度 PowerVault™ ストレージ製品を使用しています。ハードウェアとソフトウェアのフルサポートをDellならびにインテルが行っているので、インテルEE for Lustre Solution搭載のHPC用Dellストレージは性能、信頼性、密度、簡単な使用、費用対効果など非常に多くの優位性があります。

次の項目ではLustre ファイルシステムとインテルEE for Lustre Solution搭載のHPC用Dellストレージについて説明します。また、引き続いてパフォーマンス分析と結果、補足資料 A ：ベンチマークコマンドリファレンスについて説明していきます。

2. Lustre ファイルシステム

Lustre は並列ファイルシステムで、データへのパラレルアクセスや分散ロッキングなどの高いパフォーマンスを提供します。Lustreのインストールは、メタデータサブシステム、オブジェクトストレージサブシステム（データ）、クライアントによるデータへのアクセスと運用という、3つの鍵となるエレメントで構成されています。

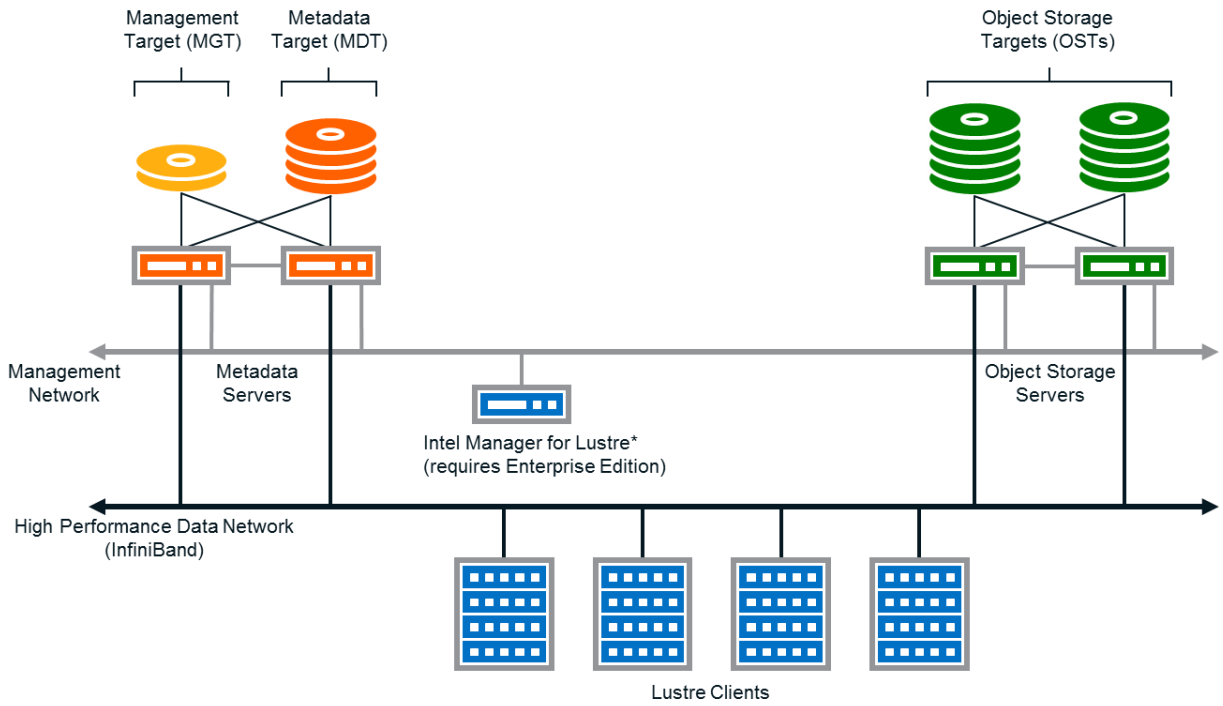
メタデータサブシステムはメタデータターゲット（MDT）、マネジメントターゲット（MGT）、メタデータサーバー（MDS）で構成されています。MDTはファイル名、パーミッション、タイムスタンプ、オブジェクトストレージシステム内のデータオブジェクトのロケーションなどを含む、すべてのメタデータをファイルシステムに蓄積します。MGTは構成情報やレジストリなどの管理データを保存します。MDSはMDTを管理するため専用サーバーです。このソリューションスタックには分散ネームスペース、または Lustre DNE 構成を追加するオプションもあります。

---

[1] 次世代 Lustre Software によってHPCの利用は一般企業へと拡大

オブジェクトストレージサブシステムは1つ以上のオブジェクトストレージターゲット（OST）と、1つ以上のオブジェクトストレージサーバー（OSS）で構成されています。各OSSが1つ以上のOSTの管理を行い、OSTはファイルオブジェクトデータのためのストレージを提供します。通常、複数のアクティブなOSSが同時に存在します。LustreはアクティブなOSS(とそれに連携したOST)を増やすことで、スループットを増大させることができます。これらの増やしたOSSは現行のネットワークスループットを増加させ、増やしたOSTはストレージ容量を拡張させます。図1は、典型的なLustre構成におけるコンポーネントであるMDS、MDT、 MGS、 OSS、OSTの関係を示しています。図中のクライアントはHPCクラスターのコンピュートノードです。

図1：Lustre を基盤としたストレージソリューションのコンポーネント



Lustreのような並列ファイルシステムは、複数のオブジェクトストレージターゲット（OST）にデータを配布する（“ストライピング”する）ことで性能とスケーラビリティを実現し、同時に複数のコンピュートノードがデータアクセスを効率的に行えるようになりました。Lustreの設計においてもっとも配慮が必要とされたのは、システム全体のパフォーマンス向上のため、メタデータアクセスをIOデータアクセスから分離した点です。

LustreのクライアントソフトウェアはLustreファイルシステムに保存されたデータへのアクセスができるよう、コンピュートノードにインストールします。クライアント側は、ファイルシステムに単独のネームスペースが表示され、アクセスのためにマウントされます。この単独のマウントポイントはアプリケーションデータへのアクセスをシンプルにし、ネイティブクライアントオペレーティングシステムツールを介してアクセスができるため、より簡単な運用が可能になります。

LustreはLustre ネットワーク（以下LNet）と呼ばれる高度に洗練、また拡張されたストレージネットワークプロトコルを持っています。LNetはあるタイプのネットワークに強みがあります。例えば、インテルEE for Lustre Solution搭載のHPC用Dellストレージが InfiniBand® をクライアント、MDS、OSSと繋ぐネットワークとして使用する際、LNetは通常のネットワークプロトコルに比べ、より高速なI/O転送と低いレイテンシーを提供する、InfiniBand ファブリックにおけるRDMAの利点をLustreで活用することができます。

Lustreファイルシステムの特徴を要約すると：

- メタデータターゲット(MDT) - データの“ストライピング”のロケーション、ファイル名、タイムスタンプなどの保存
- マネジメントターゲット(MGT) - 構成やレジストリのような管理データを保存
- メタデータストレージサーバー(MDS) - MDTを管理し、Lustreのクライアントにファイルへのアクセスを提供
- オブジェクトストレージターゲット(OST) - ファイルシステムでのデータストライプの保存、またはファイル拡張を行う
- オブジェクトストレージサーバー(OSS) - OSTを管理し、Lustreクライアントにデータへのアクセスを提供
- Lustreクライアント - ファイルがどこに位置するかを決定するMDSへのアクセス、またOSSへアクセスを行いデータ読み込みや書き込みを行う

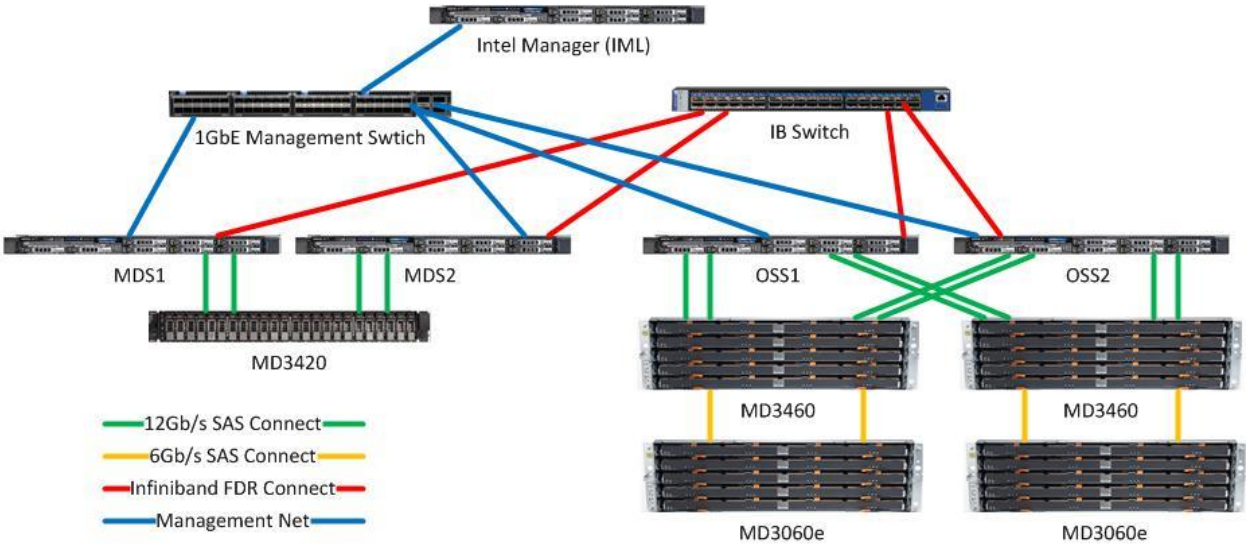
通常、Lustreの構成と導入は非常に複雑で時間のかかるタスクです。Lustreのインストールと管理は通常、コマンドラインインターフェイス（CLI）を介して行っています。この作業にはファイルシステム運用に関する高度な知識と、LNetやロッキングメカニズムのような予備ツールが必要です。 また、Lustreストレージシステムを導入した後のシステムメンテナンスとパフォーマンスの最適化は手間のかかる作業ともいえます。このような要件や学ばなければならない事項が多くあるということで、Lustreに馴染みの薄いシステム管理者は並列ファイルシステムの利点を自社に活かしきれないこともあります。Lustreの経験があるシステム管理者でも、Lustreファイルシステムをメンテナンスするのに大幅な時間を費やしているかもしれません。

3.インテル EE for Lustre搭載のHPC用 Dellストレージとは

インテル EE for Lustre Solution搭載のHPC用 Dellストレージは、管理サーバー、Lustreメタデータサーバー、Lustreオブジェクトストレージサーバーと、それらに付随するバックエンドストレージで構成されるストレージソリューションを提供します。ソリューションはクラスターのコンピュートノードで簡単にアクセスできる、シングルネームスペースを使用するストレージを提供します。また、ウェブベースインターフェースであるインテル Manager for Lustre（IML）で管理します。図2は、240個のドライブで構成されたインテル EE for Lustreシステム搭載のHPC用 Dellストレージの概要です。後程具体的に説明する異なったコンポーネントについても基本情報については触れていきます。主に3つのコンポーネントがあります。

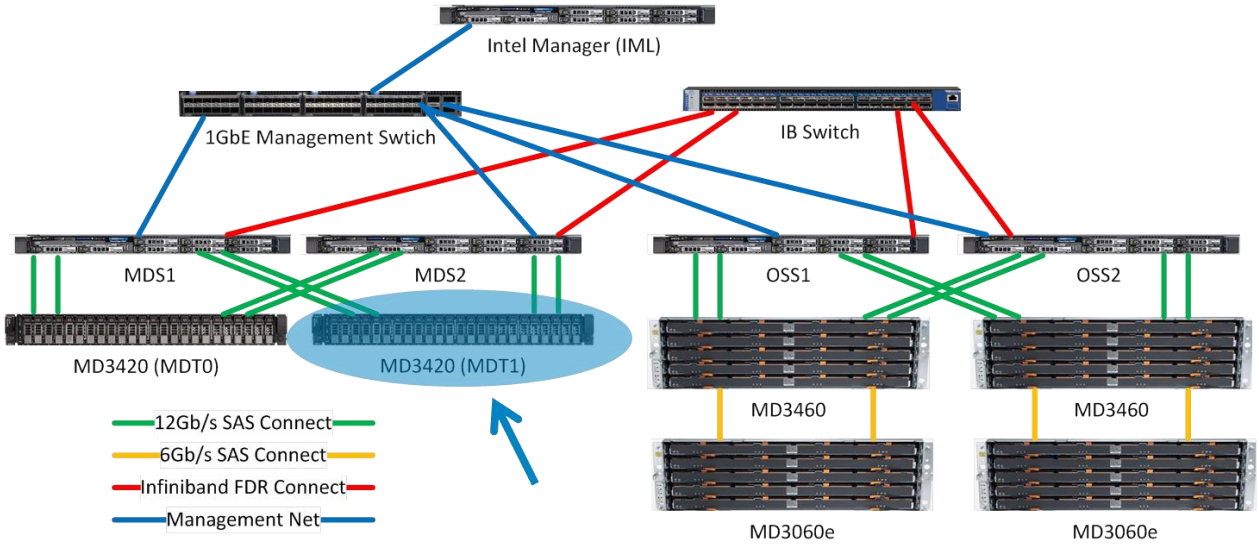
- マネジメントサーバー (IML)
- メタデータサーバーペア (MDS)
- オブジェクトストレージサーバーペア (OSS)

図2：インテル EE for Lustre搭載のHPC用 Dellストレージ コンポーネント概要



このリリースにおける新しい機能は、Lustre分散メタデータを活用するオプションを含んでいます。この機能はDNEと呼ばれ、ソリューション能力をいかんなく発揮するため容量とメタデータパフォーマンスをスケールさせる能力を提供します。また、マルチプルメタデータターゲット（MDT）を介し、メタデータ（ファイルとディレクトリ両方）を分散させることができます。加えて、メタデータスループットと同様にネームスペースのサイズをスケールさせることも可能です。Lustre DNEの機能は、開発とリリースの2つにフェーズが分かれています。フェーズ1は、遠隔ディレクトリで定義済みのMDTにLustreサブディレクトリを設定、分散させることができます。フェーズ2は、ストライプディレクトリで複数のMDTを介して得たディレクトリのコンテンツを分散させることができます。この資料作成時にはすでにフェーズ1がリリースされ、インテルEnterprise Edition for Lustre v.2.3 Softwareによってサポートされています。フェーズ2は開発段階でリリースに向けて準備中のため、この資料はLustre DNEのフェーズ1について網羅しています。図3はLustre DNEオプションのソリューションアーキテクチャを可視化したものです。

図 3：Lustre DNE オプション



インテル EE for Lustre Solution搭載のHPC用 DellストレージはDell PowerEdge R630 サーバープラットフォームを管理用サーバー、オブジェクトストレージサーバーやメタデータサーバーとして構成を行っています。ソリューションはDellの第13世代サーバーによってサポートされるPCIe 3.0の利点を活かしたMellanox ConnectX-3 InfiniBand FDR (56 Gb/s)アダプターをサポートしています。インテル EE for Lustre Solution搭載のHPC用 Dellストレージの最大能力を明らかにするため、FDRスピードのInfiniBandを基盤としたパフォーマンスに特化して調査をしていきます。図4のPowerEdge R630は、1Uフォームファクターにおいてソリューションサーバーコンポーネントに要求されるパフォーマンスや運用性、サーバー密度を備えています。

図 4: Dell PowerEdge R630

3.1 マネジメントサーバー

インテルマネジメントサーバーは、内部1GbEネットワークを介してメタデータサーバーとオブジェクトストレージサーバーに繋がる単独のサーバーです。

マネジメントサーバーは、インタラクティブWeb GUIコンソールであるインテル Manager for Lustreを介してシステムの正常性管理や基本的なモニタリングデータ収集を行うのと同様に、ユーザーインタラクションを提供します。また、全ての管理とインテル EE for Lustre solution搭載のHPC用 Dellストレージへの管理者アクセスはこのサーバーを通じて行います。ただし、Lustreファイルシステムに関連するデータ収集やソリューションに対するマネジメントは提供しますが、Lustreファイルシステムまたは、データ経路についてアクティブな運用を行うことはありません。

インテル Manager for Lustre (IML) GUIはインストールプロセスの複雑さを取り除き、Lustreシステムの導入や構成にかかる時間を短縮します。また、異なるコンポーネントの正常性やパフォーマンスのモニタリングも自動的に行います。導入と構成にかける時間と労力を削減することで、業務に使用するための準備のスピードアップができます。モニタリングの自動化は、システム管理者に負担をかけることなくエンドユーザーにより良いサービスを提供することができます。また、このソリューションはファイルシステムのパフォーマンスに関連したトラブルシューティングを手助けするツールの提供も行います。さらに、モニタリングツールの履歴保存能力はストレージアプライアンスの拡張やメンテナンス、更新のプランニングに役立ちます。

3.2 メタデータサーバー

図5のメタデータサーバーペアは、アクティブ／パッシブの高可用性クラスターで構成される2つのDell PowerEdge R630 サーバーによって構成されています。それぞれのサーバーはLustre MDT と MGTが入っている単独のDell PowerVault MD3420ストレージアレイに直接接続されています。Dell PowerVault MD3420は24 – 300GB、15K回転、2.5” ニアラインSASドライブが搭載された、22個のディスクからなるRAID10と2つのホットスペアと構成されています。このメタデータターゲット（MDT）によるソリューションは、ファイルシステムメタデータ用に約3TB分のスペースを提供することです。MDSは適正なオブジェクトストレージターゲットのファイル、ディレクトリ要求、ルーティングタスクのハンドリングを担います。このサイズの単独MDTだけで扱うことができる最大ファイル数は16億を上回ります。このソリューションで、ストレージ要求はシングル56 Gb/s FDR InfiniBand接続によるLNetを介してハンドリングされます。

図 5: メタデータサーバーペア

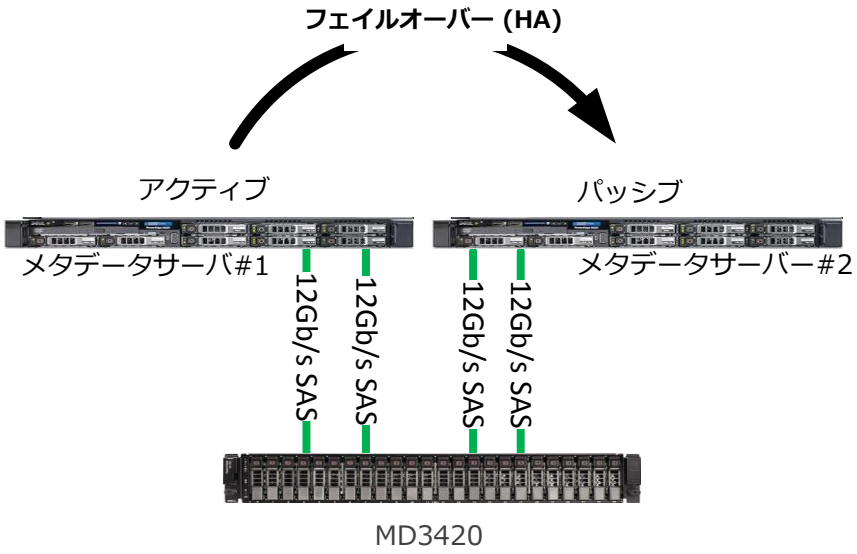
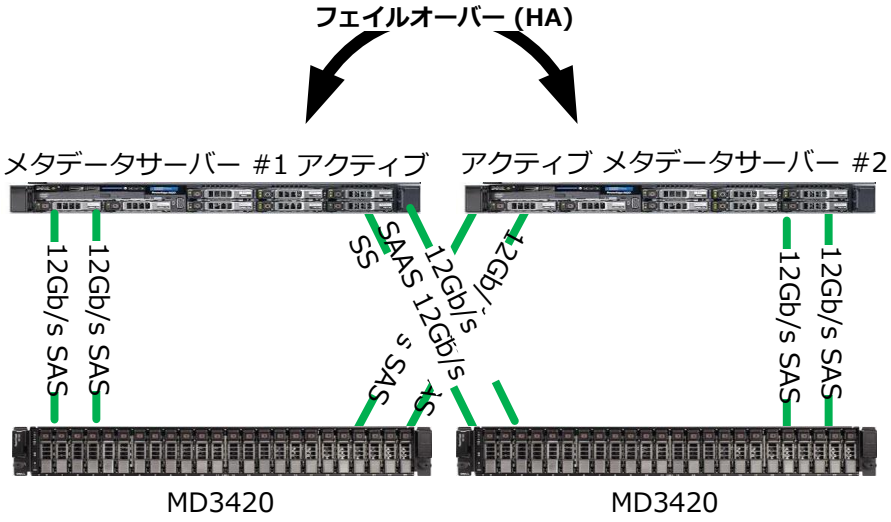


図6はLustre DNE 構成オプションを表しており、アクティブ／アクティブ構成の2つのDell PowerEdge R630サーバーで高可用クラスターを構成しています。各サーバーは2つのDell PowerVault MD3420ストレージアレイに直接繋がっています。1つのMD3420がMDTと同じように1つのLustre MGTを格納し、セカンダリのMD3240が別のMDTを格納しています。サーバーは、各サーバーが１つのMDTに対してプライマリとなり、セカンダリは別のMDTに対応するように構成されています。それぞれのMD3420は24 – 300GB、15K 回転、2.5” SASドライブが搭載された、22個のディスクからなるRAID10と2つのホットスペアと構成されています。Lustre DNEオプションは、ファイルシステムメタデータとファイル数を2倍使用することができます。

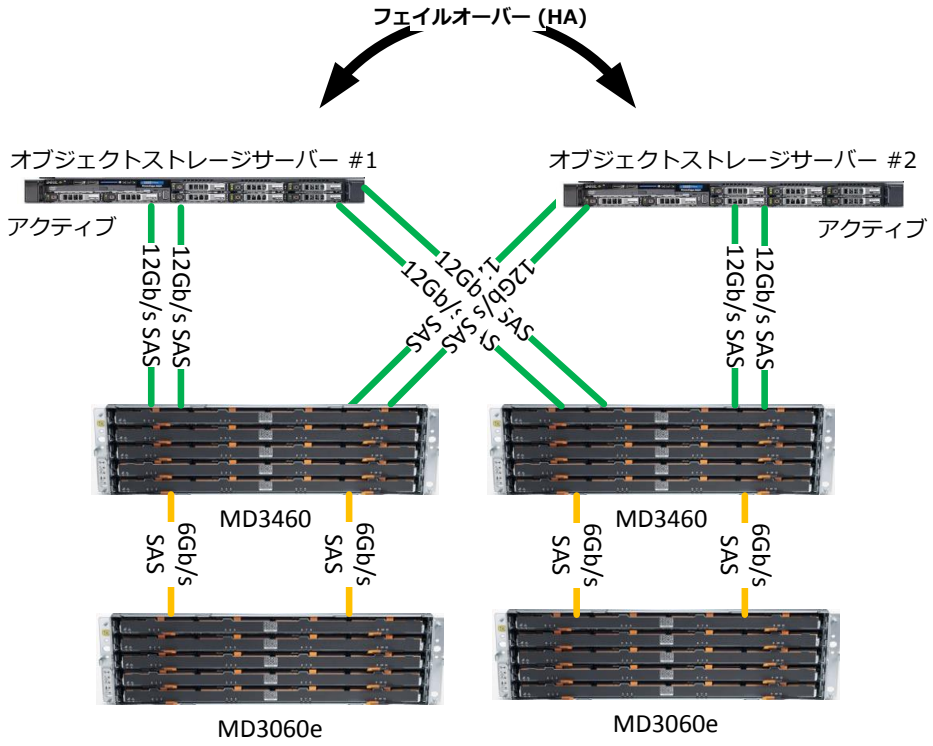
図 6 : Lustre DNEオプション搭載のメタデータサーバーペア



3.3 オブジェクトストレージサーバー

図7はオブジェクトストレージサーバーを示しています。2ノードで構成された高可用性（HA）クラスターにより、MD3060e の拡張機能を備えた2つのDell PowerVault MD3460高密度ストレージアレイに対してアクティブ／アクティブなアクセスを可能にしています。それぞれのPowerVault MD3460アレイは、60個の4TB 3.5″ ニアライン SAS ドライブで構成されています。それぞれのMD3460は、PowerVault MD3060e高密度拡張アレイにより拡張されています。この構成は各OSSペアに960TBのローストレージ容量を提供することができます。

図 7：オブジェクトストレージサーバーペア

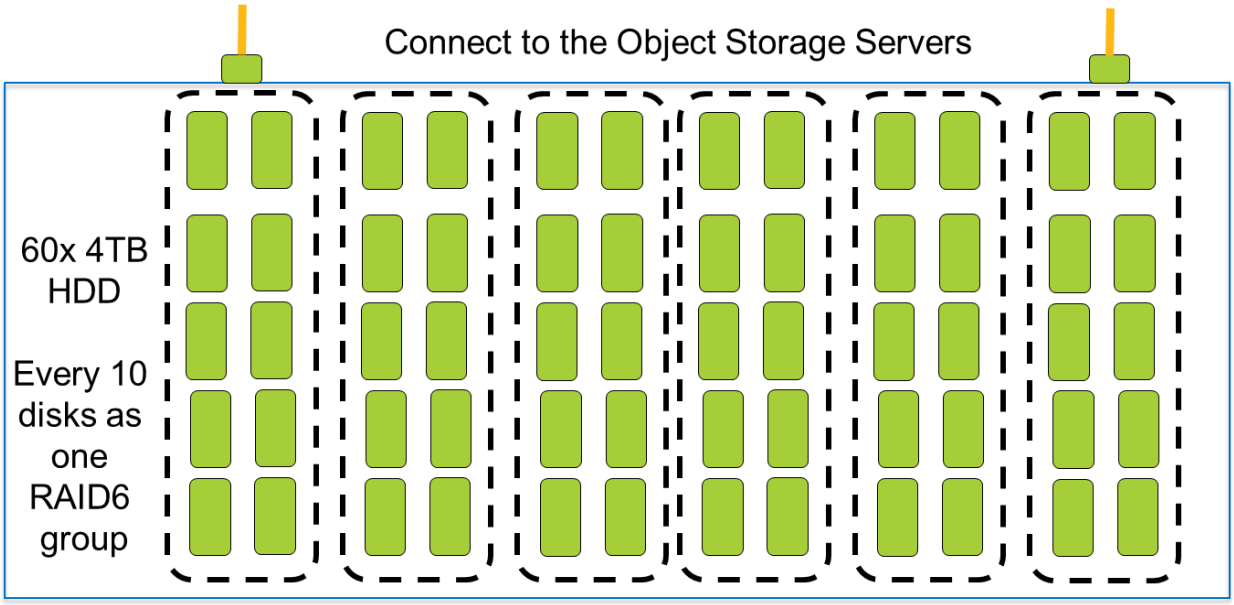


オブジェクトストレージサーバーはソリューションの構成要素です。各PowerEdge R630の2つのデュアルポート12Gb/s SASコントローラーで、2つのサーバーが各PowerVault MD3460高密度ストレージアレイに冗長的に接続されます。

図8は、各ストレージアレイが6つのRAID6バーチャルディスクにどのように分かれているかを示しています。各バーチャルデスクには8つのデータと2つのパリティデスクが入っており、アレイのトレーごとに2つのデスクを使用しています。エンクロージャ毎に6つのオブジェクトストレージターゲットを備えています。RAID6を使用することで、書き込みの性能（各RAID6が必要とする余剰のパリティデータにより）の高い信頼性を提供します。各OSTは約29TBのフォーマットされたオブジェクトストレージスペースを提供します。インテル EE for Lustre solution搭載のHPC用 Dellストレージにより、シングルOSSペアはPowerVault MD3060e拡張アレイをMD3460アレイに加えることで、24 のOSTを持ちます。 OSTは56Gb/s Infiniband FDR接続により、クライアント側からもLNet経由で見ることができます。

Lustre　クライアントを備えたどのコンピュートノードからも見ることができるので、Lustre管理の下でその他のファイルシステムと同じように、すべてのネームスペースの閲覧と管理ができます。

図8 : MD3460, MD3060e アレイ上のRAID6レイアウト



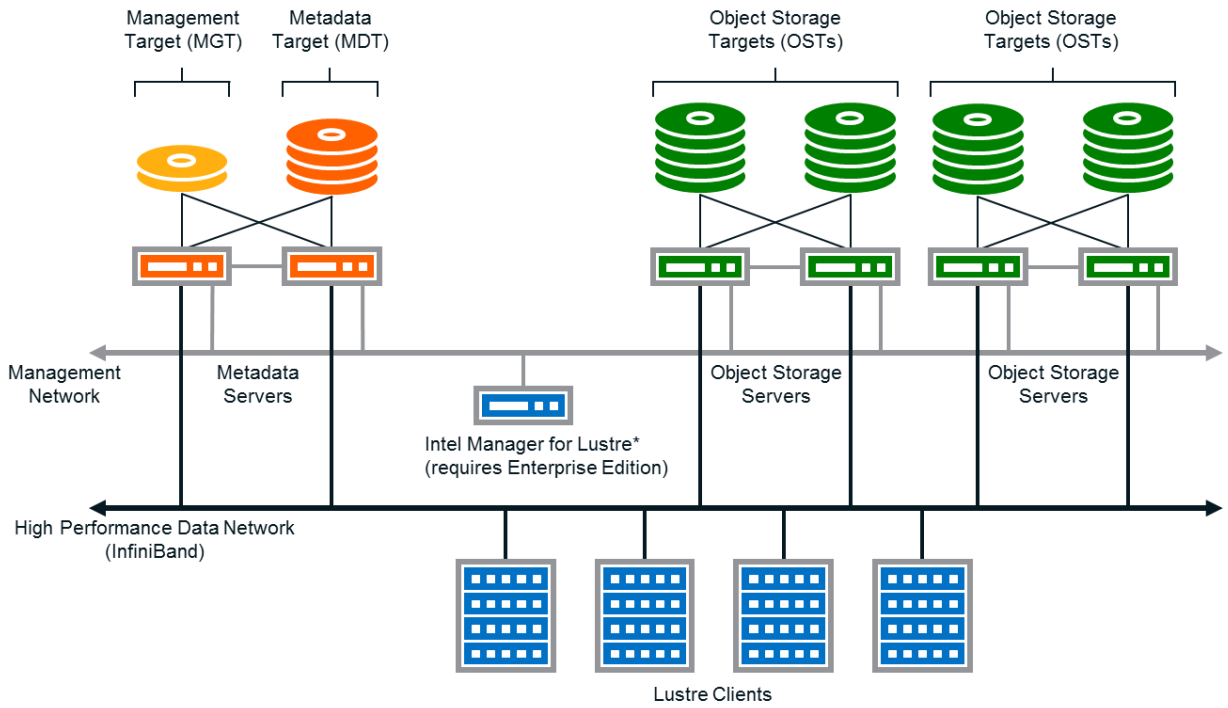
3.4 スケーラビリティ

アクティブ/アクティブクラスター構成のオブジェクトストレージサーバーにより、膨大なスループットとプロダクトの信頼性を生みだしています。この構成は高可用性を提供し、メンテナンスの必要性を削減するので、結果としてダウンタイムの削減にもつながります。

PowerEdge R630サーバーは、性能と高いサーバー密度を提供します。このソリューションは各OSSペアに960TBのローストレージを提供します。また、FDR InfiniBandによるローレーテンシストレージ転送を使用した相互接続を超高速で行うことができるという利点もあります。PowerEdge R630はOSSでより高いネットワークスループットを実現するために、FDR InfiniBandのPCIe Gen3 インターフェースを利用することができます。

RHEL6.6 kernel with Mellanox OFED v.2.4のLustreクライアントバージョン2.5.37は、インテル EE for Lustre搭載のHPC用 Dellストレージへのアクセスが可能です。（詳細はインテル EE for Lustre Solution搭載のHPC用 Dellストレージ構成ガイドを確認ください）

図 9：OSS スケーラビリティ



インテル EE for Lustre搭載のHPC用 Dellストレージは、ストレージバックエンドを備えた追加のOSSペアを加えることでスケーリングを行うことができます。（図9参照） これは、ストレージ容量と全体のネットワークスループットを同時に増やすことが可能だということを意味しています。つまり、最大ネットワークスループットを維持しながらもストレージ容量を増やすことができるということです。

3.5 ネットワーク

3.5.1 管理ネットワーク

プライベート管理ネットワークはストレージ構成やモニタリング、メンテナンスと同様、LustreとLustre HA機能のためのコミュニケーションインフラストラクチャーを提供します。このネットワークは日々の運用を担うセグメントと、トラブルシューティングやメンテナンスのための制限されたスコープのための分離されたセグメントを作り出します。管理サーバーはこのネットワークを使用し、管理者によるシステム変更や正常性の確認、情報収集を行いながら異なるソリューションコンポーネント間で相互に作用します。OSSとMDSはどちらも管理サーバーと連携し、正常性に関する情報やパフォーマンスデータを提供しています。また、管理作業中であっても動作することが可能です。PowerVault MD3420 と MD3460コントローラーはアウトオブバンド（イーサーネットポート）を介し、ストレージアレイ正常性のモニタリングとストレージバックエンドの管理作業を行っています。

このインテグレーションレベルは、経験の浅い運用者でもソリューションの監視と管理を容易に、かつ効果的行うことを可能にします。デフォルトの情報はクイックインスペクションにおいて簡略化されていますが、ユーザー自身でサーバーまたはストレージコンポーネントからある一定のレベルの詳細な情報を確認することがきます。

3.5.2 データネットワーク

LustreファイルシステムはInfiniBand FDR、イーサーネット、もしくはその両方のLustreネットワーク（LNet）を介して提供されています。これはクライアントがデータアクセスをするために使用するものです。

インテル Manager for Lustre（IML）のGUIインターフェイスは、MDSサーバーとOSSサーバーのシングルかマルチいずれかがLustreネットワークに加わるためのLustreネットワーク識別子（NID）を構成するオプションを提供しています。例えば、ご自分のOSSサーバーにある10Gbイーサーネットインターフェイス（例：eth0）をLustreネットワークに加えることができると同じように、Infinibandインターフェイス（例：ib0）を構成することができます。

InfiniBandネットワークでは、ローレイテンシ―での高速な転送スピードを達成しています。LNetの利点はRDMAによってクライアントとMDT、OST間の高速なデータおよびメタデータ転送を提供できるということです。OSSとMDSサーバーはシングルポートのMellanox ConnectX-3 56 Gbアダプターを使用したFDR InfiniBandファブリックを利用できるという利点もあります。FDR InfiniBand HBAは現行のQDRやDDRネットワークに統合することも可能です。

3.6 インテル EE for Lustre搭載のHPC用 Dellストレージの管理

インテル Manager for Lustre（IML）は管理のためのセントラルウェブGUIを備えることで、Lustreファイルシステムの管理を行いやすくしました。例えば、IMLはノードからノードへの（OSSもしくはMDS）ファイルシステムのフェイルオーバーの開始、ファイルシステムのフォーマット、ターゲットのマウントとアンマウントやLustreファイルシステムのパフォーマンスとコンポーネント状態の監視など、これらのアクションを標準化させるツールとして使用されています。図10はいくつかのIMLモニタリングインターフェースを示しています。

IMLはウェブベースの管理コンソールで、ソリューション（すべてのセキュリティ上の要件が満たされた場合）を管理しています。これによりモニタリングや管理を行っているハードウェアやソフトウェア、ファイルシステムコンポーネントを確認することが可能です。

IMLの使用により、複雑なCLI操作を必要としていた多くのタスクが、わずか数クリックで行うことができるようになりました。また、IMLはMDSから他へのフェイルオーバーや監視、ファイルシステムのシャットダウンなどにも使用されます。

図 10：インテル Manager for Lustre (IML) インターフェース

4. パフォーマンス評価と分析

本資料のパフォーマンス調査結果は、240個のドライブで構成されたインテル EE for Lustre solution搭載のHPC用 Dellストレージを基に作成されています。構成は240個の4TBディスクドライブ（960TBロースペース）です。目標はソリューションの性能とピークパフォーマンス、スケーリングに最も適した方法を明らかにすることです。ソリューションテストのためのI/Oワークロードを提供するクライアントのテストベッドは、PowerEdge M610ブレードサーバー上のDell HPCコンピュートクラスターです。表1に構成を表しています。

パフォーマンスの限界やその持続可能性を測るため、異なる種類のワークロードによる構成への負荷に焦点を置いた多くのパフォーマンス調査が実施されてきました。InfiniBandはネットワークのボトルネックをなくし、そのハイスピードとローレイテンシーがインテル EE for Lustre solution搭載のHPC用 Dellストレージのパフォーマンスを最大限引き出すことができるため、しばしばこのような調査に使われてきました。

私たちは通常「標準的で一貫性のある」テスト環境と方法を維持するよう心がけています。ただし、キャシュの影響限度を計るのと同じく、サーバーやストレージ構成を最適化することはあります。目標はパフォーマンスへのインパクトをより明らかにすることです。本資料はそのような構成の仕様に関する詳細も含んでいます。

表 1：テストクライアントクラスターの詳細

コンポーネント	解説
コンピュートノード:	Dell PowerEdge M610, 64 ノード
ノード BIOS:	6.3.0
プロセッサー:	2 x Intel Xeon™ X5650 @ 2.67GHz 6コアプロセッサー
メモリ:	24GB DDR3 1333MHz
インターコネクト:	InfiniBand - Mellanox Technologies M3601Q (QDR)
Lustre:	Lustre 2.5.37 + Mellanox OFED クライアント
OS:	Red Hat Enterprise Linux 6.6 (2.6.32-504.16.2.el6.x86_64)
IB ソフトウェア:	Mellanox OFED 2.4

パフォーマンス分析は、鍵となるパフォーマンスの指標3つに焦点を絞りました：

- 秒ごとのスループットとデータ転送速度
- 秒ごとのI/Oオペレーション（IOPS）
- 秒ごとのメタデータオペレーション（OP/s）

目標はインテル EE for Lustre搭載のHPC用 Dellストレージの広範かつ正確なレビューをすることです。2つのベンチマーク評価を行うために IOzone と MDtest を採用し、この目標達成にむけた評価を行いました。

ベンチマークが使用するファイルへのアクセス方法は2種類あります。1つ目はN-to-Nと呼ばれる方法で、ストレージシステム内の異なるファイル（Nファイル）にベンチマークの全スレッドを書き込む（Nクライアント）方法です。IOzoneとIORはどちらもこの評価方法が可能です。この調査ではIOzoneのN-to-Nアクセス方法によるワークロードを使用します。2つ目のファイルアクセス方法はN-to-1と呼ばれ、全スレッドを同じファイルに書き込みます（Nクライアント、1ファイル）。この調査では、IORのN-to-1アクセス方法によるワークロードを使用します。IORはN-to-1ファイルアクセステストをするため、MPI-IO、 HDF5、 POSIXが使用可能です。分析の目的上、最適なPOSIXを使用します。

N-to-1テストは、ファイルシステムが複数スレッドを同じファイルに書き込みと読み込みを行う際の同時要求のオーバヘッドにどのように対応するかを調べます。オーバーヘッドはスレッドを扱うLutresファイルのロッキングとデータ書き込みから発生します。補足資料Aはこの2つのベンチマークで使用したコマンド例を記載しています。

各評価はソリューションのスケーラビリティをテストするため、さまざまなクライアントを使用し実施しています。また、シングルクライアント～64クライアントまで多様な物理クライアントを同時に使用したテスト評価を行います。物理的なサーバー数に対応するスレッド数は最大64個までになります。64以上の全スレッド数は、全クライアントに対する1クライアント当たりのスレッド増加数によってシミュレーションを行っています。例えば、128スレッドは、64クライアントそれぞれが2つのスレッドを実行します。

ソリューションのテスト環境は全体で960TBのローディスクスペースを持つシングルMDSペアとシングルOSSペアです。OSSペアは256GB メモリーと2つの12Gbps SASコントローラー、さらにシングルMellanox ConnectX-3 FDR HCAを持った2つのPowerEdge R630で構成されています。詳細はインテル EE for Lustre搭載のHPC用 Dellストレージの構成ガイドのケーブル配線や拡張カードに関する項目をご確認ください。MDSは256GBメモリーやMellanox ConnectX-3 FDR HCA、デュアル12Gbps SASコントローラーによって構成されています。

InfiniBandファブリックは、インテル EE for Lustre Solution搭載のHPC用 Dellストレージ用サーバーのための、クライアントクラスター用の32ポートMellanox M3601Q QDR InfiniBand スイッチと36ポートMellanox SX6025 FDR InfiniBandスイッチで構成されています。M3601Q スイッチの3つのポートがSX6025スイッチに接続されます。

表2は異なるソフトウェアとハードウェアコンポーネントの特徴についての詳細を示しています。

表 2：インテル EE for Lustre 搭載のHPC用 Dellストレージ構成

構成サイズ	960TB RAW
Lustre サーバーバージョン	2.5.37.7
インテル EE for Lustre バージョン	v2.3
OSS ノード	2 x PowerEdge R630 サーバー
OSS メモリ	256GB DDR4 2133MT/s
OSS プロセッサー	2 x インテル Xeon™ E5-2660V3 @ 2.60GHz 10コア
OSS サーバー BIOS	0.3.28
OSS ストレージアレイ	2 x PowerVault MD3460, 2 x PowerVault MD3060e
OSS ストレージアレイドライブ	240 3.5" 4 TB 7.2K 回転 ニアライン SAS
OSS SAS コントローラー	2 x SAS 12Gbps HBA LSI 9300-8e
MDS ノード	2 x PowerEdge R630 サーバー
MDS メモリ	256GB DDR4 2133MT/s
MDS プロセッサー	2 x インテル Xeon™ E5-2660V3 @ 2.60GHz 10コア
MDS サーバー BIOS	0.3.28
MDS ストレージアレイ	1 x PowerVault MD3420
MDS ストレージアレイドライブ	24 - 2.5" 300GB ニアライン SAS
MDS SAS コントローラー	2 x SAS 12Gbps HBA LSI 9300-8e
データネットワーク - InfiniBand	
OSS、MDS サーバー	Mellanox FDR HCA MT27500
コンピュートノード	Mellanox QDR HCA MT26428
クライアントQDR IBスイッチ	Mellanox M3601Q
FDR IBスイッチ	Mellanox 36ポート SX6025
IBスイッチ コネクティビティ	クライアント: QDR ケーブル; サーバー: FDR ケーブル QDRスイッチからFDRスイッチへのアップリンク x 3

キャッシュ効果による誇張された結果を避けるため、テストは下記の技術を活用したコールドキャッシュで行いました。テスト開始前にLustreファイルシステムの再マウントを行いました。syncが実行され、カーネルは下記のコマンドによって全クライアントのキャッシュをクリアするようにしました：

- sync
- echo 3 > /proc/sys/vm/drop_caches

さらに、サーバーのコールドキャッシュをシミュレーションするため、 ”sync”で全アクティブサーバー（OSSとMDS）のすべてのテストとカーネルにおいて、クライアントに使用するものと同じコマンドでキャッシュクリアを実行しました。

インテル EE for Lustre Solution搭載のHPC用 Dellストレージのパフォーマンスを計るため、すべてのテストは同じような初期状態で開始しています。ファイルシステムは全機能を活かせるように構成し、ターゲットはテスト前にファイルとディレクトリを空にしています。

ソリューションのパフォーマンス特徴の1つとして、テスト時に比較が可能であった最新のLSI SASドライバー（バージョンP8）とRHEL/CentOS6.6.のネイティブSASドライバーのパフォーマンスインパクトを比較しました。また、異なるドライバーやOSレベルチューニングについても実験を行いました。さらに、ストレージアレイのキャッシュ状態も実験を行い、ソリューションのパフォーマンス全体におけるインパクトも確認しました。次の項目ではその結果について触れていきます。

4.1 N-to-N 連続読み込み/書き込み

連続動作による評価はIOzoneテストツールバージョン3.429を使用して実施しました。図11のスループット結果は毎秒のMBに換算されています。評価用のファイルサイズは全スレッドからサンプルとして採取した、合計2TBからなるサイズです。つまり、連続読み込みと書き込みは総サンプルサイズである2TBを、テストで使用する総スレッド数で均等に割ったものになるということです。IOzoneのブロックサイズはLustreの要求サイズである１MBに設定しました。

各ファイルへの書き込みはOSSとクライアントのキャッシュ効果を最小にするため、十分に大きなものにしました。さらに、キャッシュ効果を避けるため他の技術も活用しました。各ファイルへの書き込みはOST（ラウンドロビン）によって均等に配分されています。これらにより、シングルSAS接続やOSTでの不均衡なI/O負荷を避け、同じ意味でユーザーのワークロードのバランシングも可能になります。

図 11：インテル EE for Lustre Solution搭載のHPC用 Dellストレージによる連続読み込み／書き込み

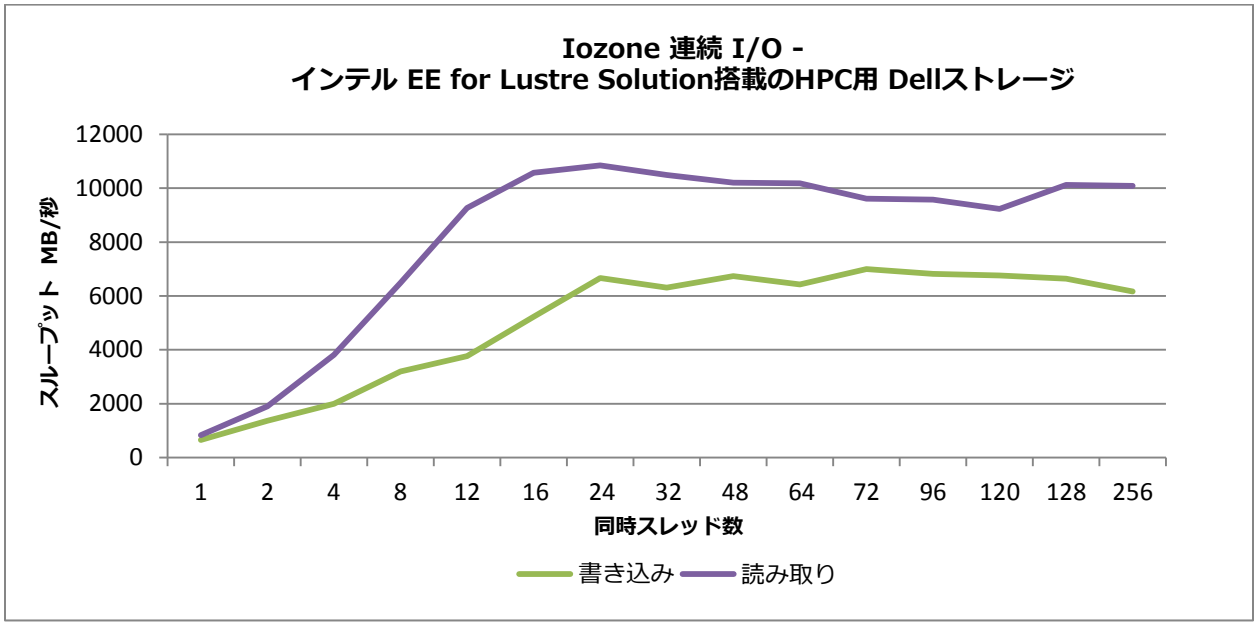
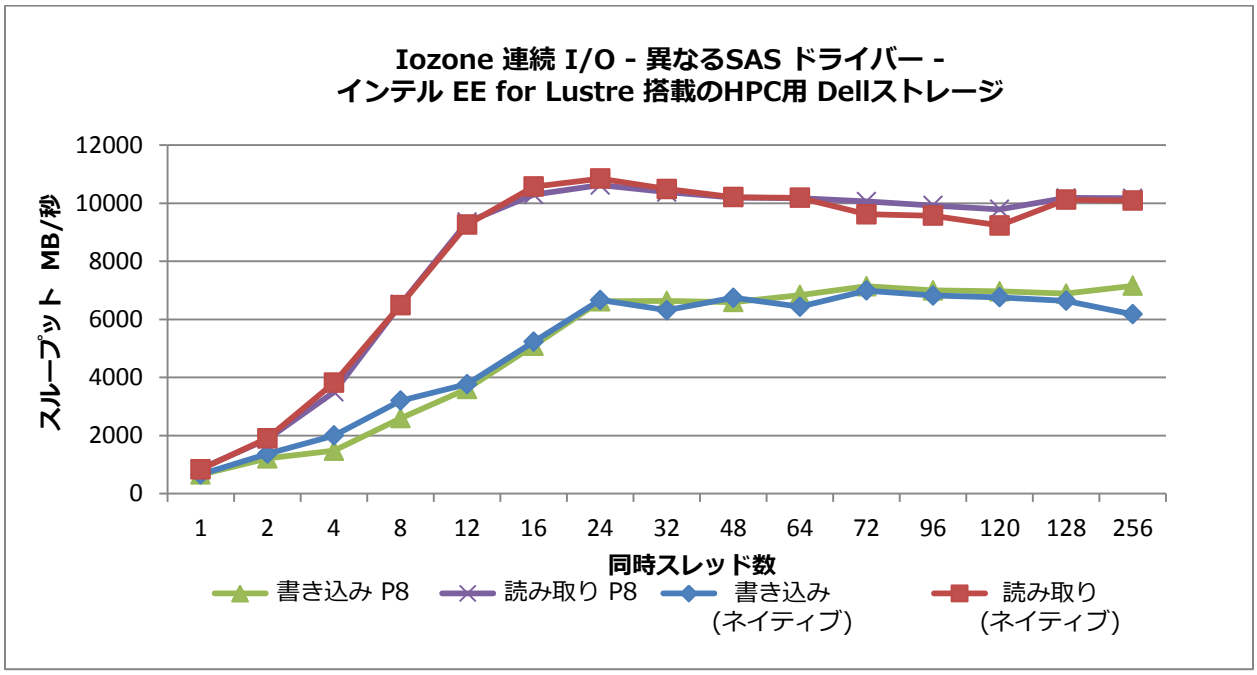


図11は960TBのテスト構成においての連続パフォーマンス評価を示しています。テストベッドでは、書き込みパフォーマンスピークが毎秒7GB弱、読み込みパフォーマンスピークは毎秒11GBでした。テストで特徴を明らかにするため、シングルクライアントワークロードとOSTの割合についての最適なペアリングも行いました。テストはシングルスレッドを使用したシングルクライアントで、いくつかのOST（2、4、6、8、24のOST）から成るストレージを対象に実施しました。OSTの割合に対する異なった各クライアントのパフォーマンスが比較的近い値を示すのに対し、8つのOSTに対するシングルクライアントの割合は毎秒1005MBでの読み込みと毎秒1090MBの書き込みと、若干高くなっていることが分かりました。書き込み読み込みのパフォーマンスはスレッドプロセス数を24まで上げるたびに一定で伸び、256で横ばいになりました。これはスレッドの数を増す（私たちのシステム内のOSTを24まで上げた）たび、OSTの数が増えた結果です。

莫大な量のファイルに対する高いスループットを維持するには、OST数を増やすのが有効と考えられます。Dell PowerVault モジュールディスクストレージマネジャーとパフォーマンスモニターのツールを使用したストレージアレイパフォーマンスの評価は、ベンチマーク測定ツールから独立した形でスループットの有効性について実施しました。

また、ストレージサーバーで現在使用可能なLSI SASドライバーを使用して調査を行いました。これはRHEL/CentOS 6.6.で利用可能なネイティブバージョンと比較しています。下記の図12は若干種類の異なるワークロードを使用した、非常に類似した動きをする2つのドライバーバージョンを示しています。

図 12 : SAS ドライバー

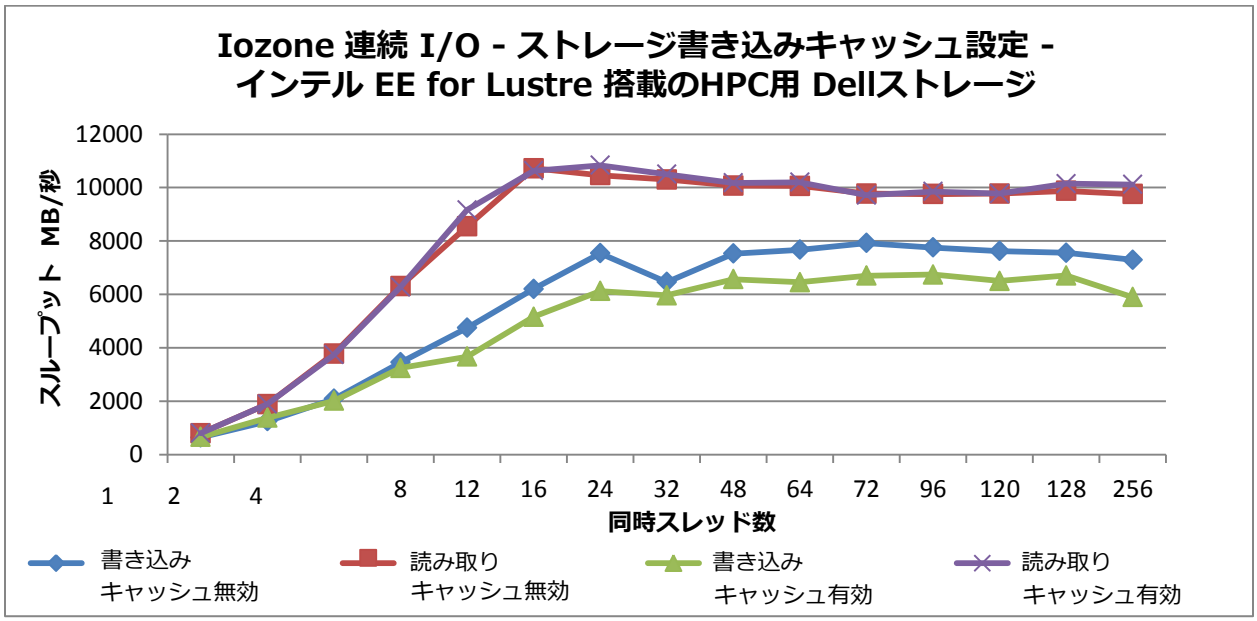


さまざまなOSとSASドライバーのレベルチューニングパラメーターがあり、特定のワークロード用途に合わせてLustreストレージサーバーを最適化するために使用されます。例えば、Lustreのデフォルトは１MBの転送が理想です。そのため、一般的なベストプラクティスはストレージシステムをI/O要求サイズを1MBに合わせることです。例えば、ドライバーパラメーターがデフォルトでの上限であるI/O要求サイズで構成されるストレージサーバーの場合、1MB Lustre RPCによってフラグメンテーションを起こします。このフラグメンテーションを回避するため、max_sgl_entriesをデフォルト値の128から256へと設定変更し、Lustre brw statsを監視し I/O 使用率が100%の状態が1MBブロックの原因であることを確認しました。テストワークロードの構成で、このチューニングがパフォーマンスに大きな差を及ぼさないことが分かりました。

連続書き込みパフォーマンスに大きく利点がでるチューニングの一つとして、ストレージシステムの書き込みキャッシュ機能を無効にすることが非常に有効であると分かりました。デフォルトでDell PowerVault MD3460ストレージアレイのバーチャルディスクを作成し、構成する際に“書き込みキャッシュ”が有効になります。各バーチャルディスク（OST）がこの書き込みキャッシュパラメーターを無効にするようにDell PowerVaultモジュールディスクストレージマネージャーを使用しました。

図13は書き込みキャッシュを無効にした際に連続書き込みパフォーマンスが大幅に向上したことを示しています。

図13：ストレージ書き込みキャッシュ



しかし、このパフォーマンスの向上は、書き込みキャッシュ無効時のランダムライトパフォーマンスによって大きく低下します。このことから、ランダムライトパフォーマンスが影響しない環境下で書き込みキャッシュを無効にすることが望ましいということができます。

4.2 ランダムリードとライト

ランダムリードとライトの測定結果を収集するためIOzoneベンチマークを使用しました。ファイルサイズは全スレッドの合計が1TBとなるように設定しました。つまり、ランダムリードとライトは合計サイズである1TBを、テストにおける総スレッド数で均等に割ったということです。IOzoneホストファイルはワークロードをコンピュートノード間で均等に分散させるようにしました。ストレージはストライプカウント1と4MBのストライプサイズのシングルボリュームとして処理しました。Lustreの4KBファイルシステムブロックサイズと連携していて、しかもランダムワークロードのスモールブロックアクセスの主なサイズであることからテストでは4KBを使用しました。パフォーマンスは毎秒のI/O運用（UOPS）で測定しました。

図14はランダムライトピークが12.6K IOPSを若干超えていることと144スレッドにおけるレベルを示しています。ランダムリードはスレッドが増えるにつれて一定間隔で下がりました。ランダムリードのIOPは48から96スレッドへと急速に増加し、120から144スレッドにかけて再度急増しました。読み込みがOSTアクセス毎のファイルロックを必要とするために飽和状態となります。読み込みはLustreの能力を活かすことで部分的、もしくはすべてのファイルにおけるオーバーラッピングによるファイルロックを拡大することができます。

図14 : N-to-N ランダムリード／ライト

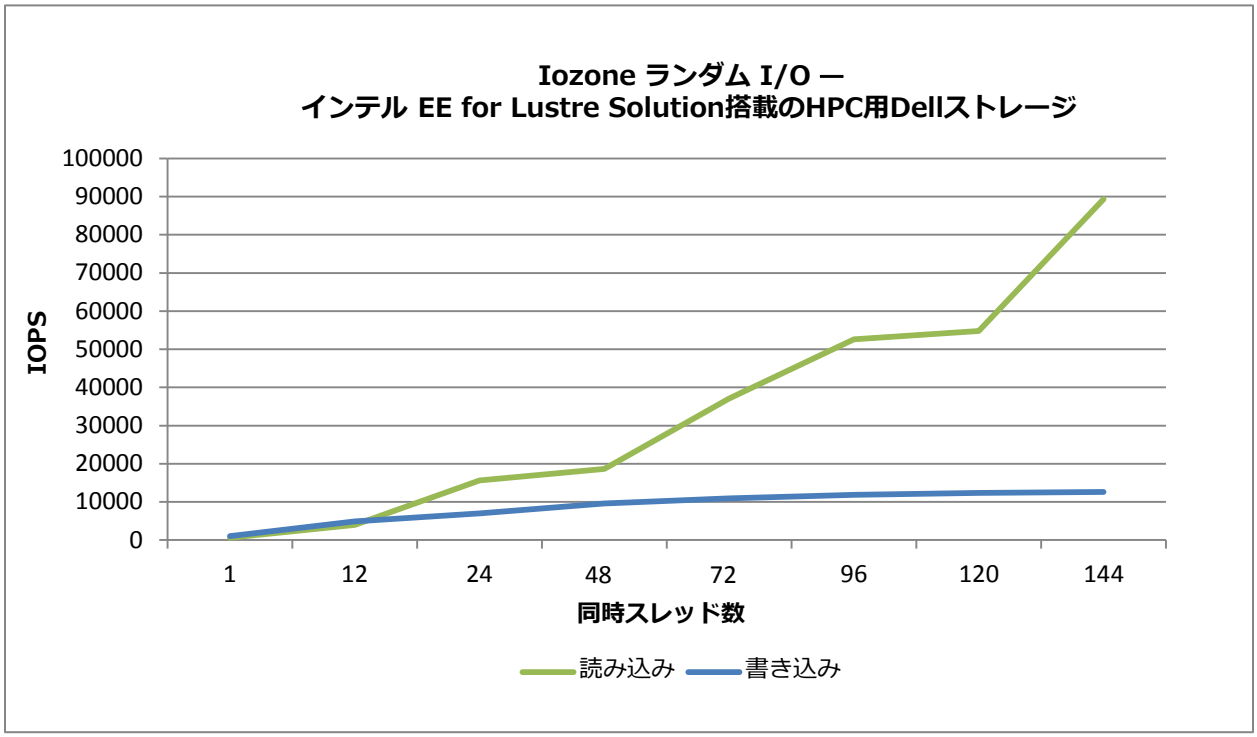
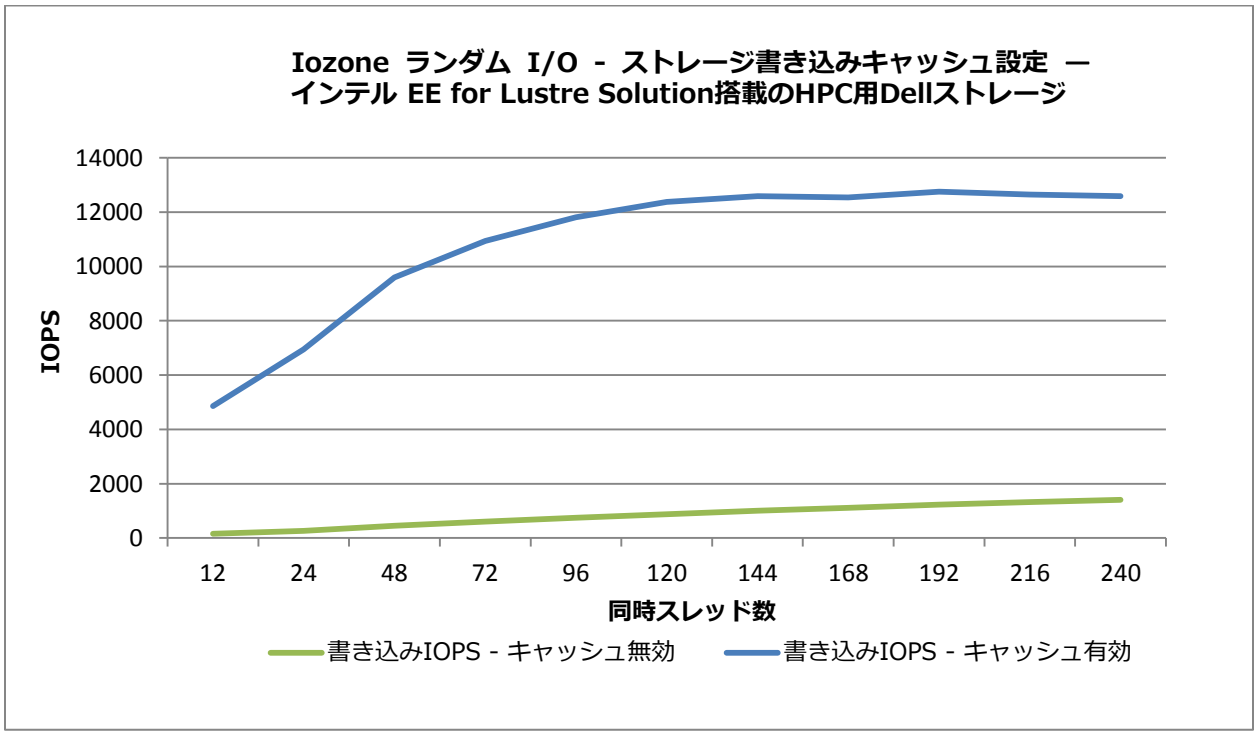


図15は書き込みキャッシュ無効時のランダムライトパフォーマンスの影響について示しています。ランダムライトパフォーマンスに大きな影響があることが分かります。

図15 : ランダムライトにおける書き込みキャッシュの影響



4.3 メタデータテスト

メタデータテストは、特定のファイルもしくディレクトリ運用においてアトリビュートの値を返すアクションが完了するまで時間について計測を行いました。MDテストはMPIが設定したベンチマークで、ファイルやディレクトリの作成、Stat、削除などのオペレーションを行います。この調査ではMDテストバージョン1.9.3.を使用しました。また、使用したMPI stackはインテルMPIバージョン5.0アップデート1です。MDテストの測定結果は、秒毎のオペレーション完了割合（OP/秒）になります。MDテストはディレクトリとファイルにおけるメタデータパフォーマンスを比較するためのものであるといえるでしょう。このテストでは、ファイルオペレーションを行ったのち、ディレクトリオペレーションを個別に行いました。

Lustreファイルシステムでは、OSTはアロケーションするためのオブジェクト識別子を要求したり、メタデータオペレーションに関する情報を紐づけたりします。この相互運用は、OSTにおけるほとんどのメタデータ運用において必要となります。ラボのテストベッドでの検証では、ほとんどすべての環境において1つのOSTを使用するほうがより効率的にメタデータ運用ができることが分かり、特にスレッド数が多いほど効率的なことが分かりました。検証では、次のLustreトポロジーの64クライアントまでで行い、このバージョンのソリューションにとってもっとも効率的なメタデータパフォーマンステスト構成を選びました。（結果については記載しておりません）

- 1 OST、1MB ストライプ
- 1 OST、4MB ストライプ
- 24 OST、1MB ストライプ
- 24 OST、4MB ストライプ

最も効率的な構成は1 OST 1MB ストライプで、この項目で示されている結果についてはLustreトポロジーで取得することができました。

予備的なメタデータテスト中にも、ディレクトリ毎のファイル数が結果に大きく影響することがみられ、作成されたファイルの総数も一定となりました。その例として、1つのディレクトリで64スレッドが3125ファイルを作成するテスト（1つのスレッドに5つのディレクトリ）と、1つのディレクトリで624ファイルを作成するテスト（1つのスレッドに25のディレクトリ）をテストしたところ、どちらも100万個のファイル作成をしましたが、IOPSのパフォーマンス測定結果は異なりました。これはディレクトリを変える際のOSTオーバーヘッドシークによるものと考えられます。一貫した結果を示すため、1ディレクトリ毎のファイル数を3125に固定し、スレッド毎のディレクトリ数が作成する合計ファイル数は100万以下になるように変更しました。また、スレッド数の増加に合わせ、ディレクトリ数も増やすようにしました。表3はそれぞれのテストで使用された値を示しています。どちらのファイルオペレーションテストもディレクトリオペレーションテストと同じく8回繰り返しテストし、その平均値を結果として記録しました。

表3：MDテストで使用したパラメータ

スレッド数 (N)	ディレクトリ毎の ファイル数	スレッド毎の ディレクトリ数	合計ファイル数
1	3125	320	1000000
2	3125	160	1000000
4	3125	80	1000000
8	3125	40	1000000
12	3125	27	1012500

16	3125	20	1000000
24	3125	14	1050000
32	3125	10	1000000
48	3125	7	1050000
64	3125	5	1000000
72	3125	5	1125000
96	3125	4	1200000
120	3125	3	1125000
128	3125	3	1200000
144	3125	3	1350000
168	3125	2	1050000
192	3125	2	1200000
216	3125	2	1350000
240	3125	2	1500000

図16：ファイルメタデータオペレーション

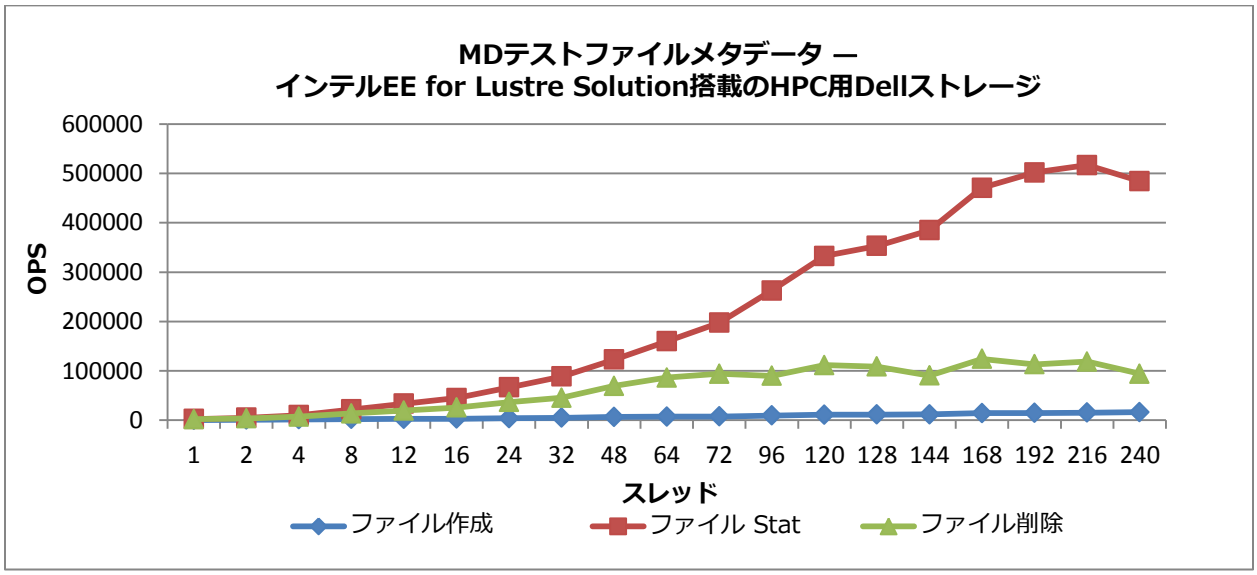


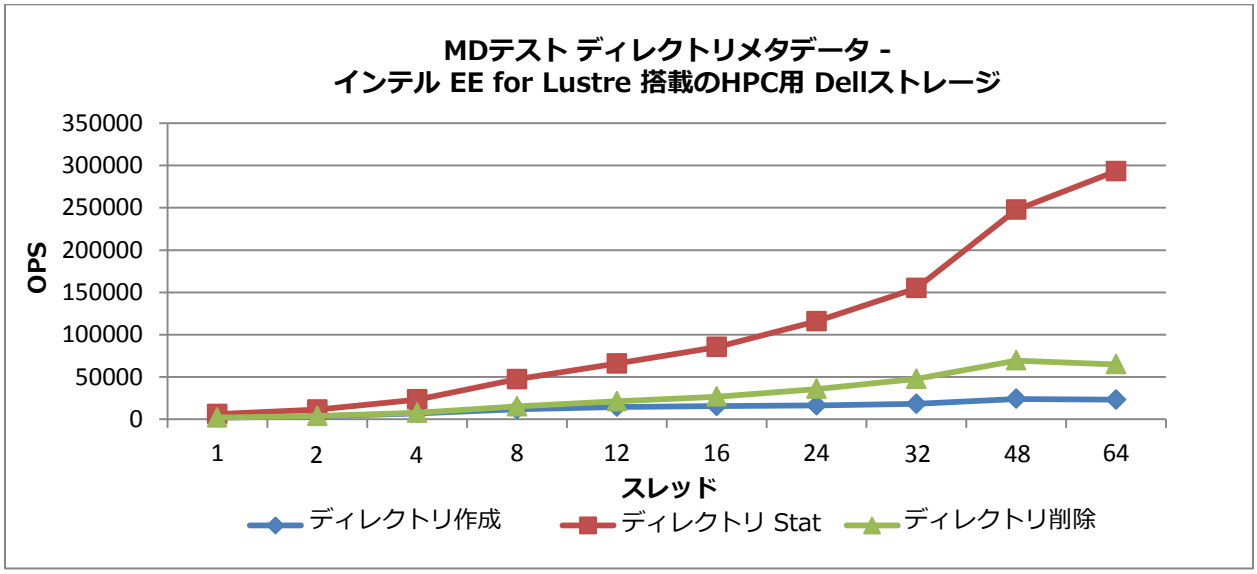
図16はMDテストを使用したファイルメタデータ結果を示しています。このグラフで示すように、メタデータオペレーションにおけるファイル作成は1スレッドで547 OSP以下で開始し、同時スレッド240による16K OPS以上にまでスケールさせました。24のストライプカウントを使用することでパフォーマンスが極端に落ちたのは、MDT上だけでなくOSTにおいてもLustreロックが必要だったためと考えられます。240スレッドで、2つのディレクトリ（-b2）と220万ファイルが作成されました。

ファイルstatメタデータオペレーションは3つのオペレーションの中で圧倒的に軽いものでした。シングルスレッドテストで2K OPS以上となり、同時スレッド216の500K OPS以上にまでスケールしました。パフォーマンスの向上は、Lustreバージョン2.5メタデータオペレーションの改善によるものと考えられます。また、MDTの15K RPMドライブの使用も影響していると考えられます。

ファイル削除もまたOSTへのアクセスが限られており、オペレーション作成に似ています。しかし、削除オペレーションは総スレッド数を増やす際（1スレッド1.8K OPS以上で開始し、同時スレッド増加とともに100K OPS以上にまでスケールしたとき）に、作成オペレーション以上の利点が生じます。

図17はMDテストを使用したディレクトリメタデータ結果を示しています。このグラフから、メタデータオペレーションを作成しているディレクトリはほとんどの場合、1スレッド2K OPS以上で開始し、スレッド増加に合わせて32K OPSにまで増えるという、非常に効率の高いオペレーションであることが分かります。ディレクトリオペレーションは（-b）を使用したトップディレクトリの数にも影響されますが、ファイルオペレーションにはあまり影響がありません。オペレーションの削除はディレクトリ作成とほぼ同じく高効率で、1スレッド2K OPSで開始し、スレッド増加に伴い70K OPS以上に達します。ディレクトリstatオペレーションは再び、3つの中で圧倒的に軽いものとなりました。シングルスレッドテストは1スレッドで6K OPS以上を作成し、64スレッドで300K以上にまでスケールしました。

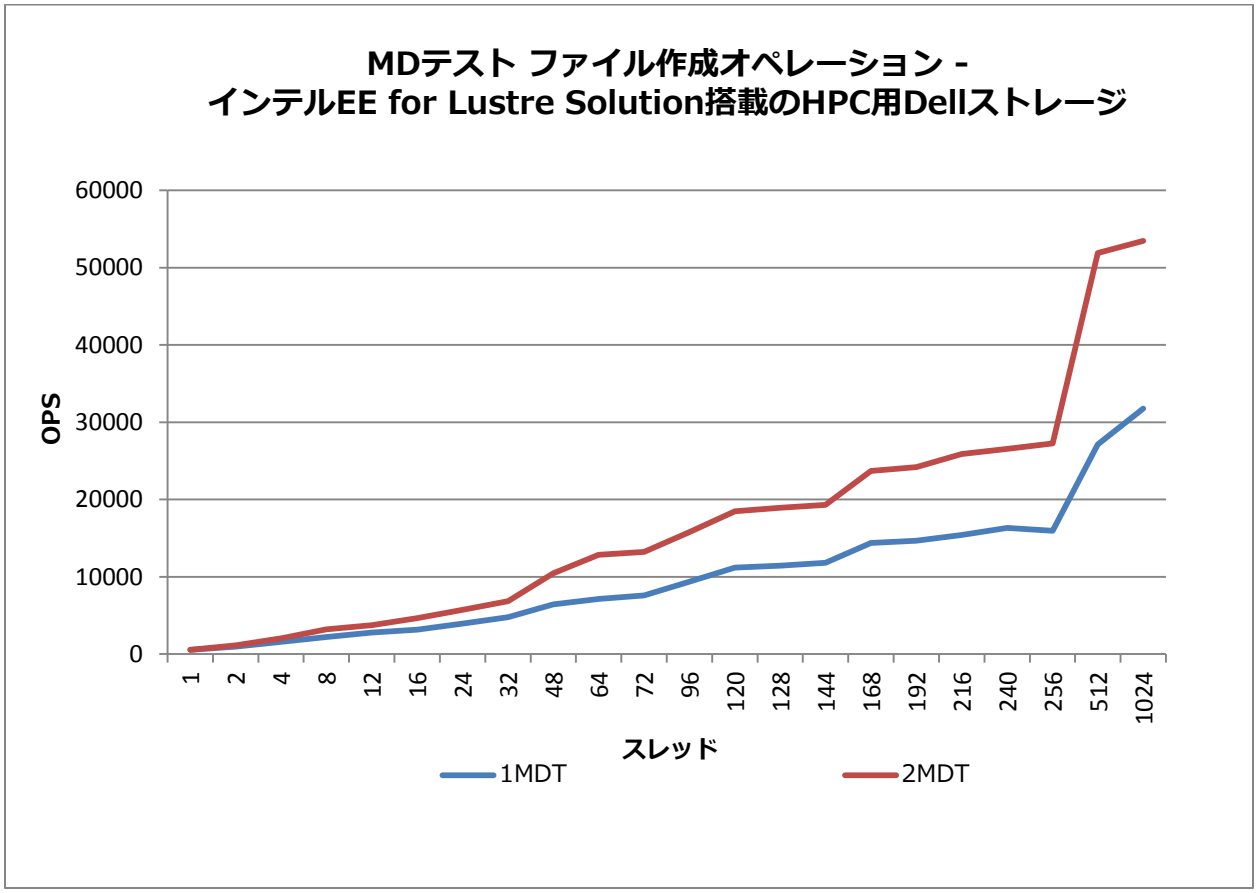
図17：ディレクトリメタデータオペレーション



このソリューションリリースはまだ新しいため、Luster Distributed Namespace featureもしくは略してLustre DNEとよんでいます+。Lustre DNEは新しい技術で、複数のLustreメタデータターゲット（MDT）に対してメタデータ（ファイルとディレクトリ）を配布することが可能です。インテル Enterprise Edition 2.3 for Lustre software 搭載のHPC用 DellストレージはLustre DNEフェーズ1をサポートしています。Lustre フェーズ1はLustre DNEリモートディレクトリを指定先です。これはLustreサブディレクトリが複数のMDTに配分されている場合となります。これはメタデータ容量を増やすだけでなく、メタデータパフォーマンスも伸ばしてくれます。図17はLustre DNEを使用したときのメタデータパフォーマンスにおける特徴を示しています。

このソリューションは、2つ目のメタデータターゲット（MDT）を構成するためのPowerVault MD3420アレイを追加することによってLustre DNEをサポートしています。

図18：MDテスト Lustre DNEによるファイル作成オペレーション



5.結論

スケーラブルでハイパフォーマンスなクラスターファイルシステムソリューションが必要とする要件があることはよく知られています。インテル EE for Lustre Solution搭載のHPC用 Dellストレージは、この要件を満たすために簡単な管理でフルサポートできるソリューションを提供します。このソリューションにはDell PowerEdge™第13世代サーバープラットフォーム、PowerVault™ストレージ製品、Lustre®テクノロジー、業界をリードする並列ファイルシステムのためのオープンソースソリューションなどの利点も含んでいます。インテル Manager for Lustre (IML)は、Lustreファイルシステム管理とソリューションコンポーネントにおけるコントロールと監視を1つのパネルで簡単に利用できるよう統合されています。

ローストレージのキャパシティは1つのオブジェクトストレージサーバーペアに対し960TBを持ち、パッケージコンポーネントの中で最大毎秒11GBの読み込みと毎秒7GBの書き込みスループットを実現することで、ハイパフォーマンスコンピューターに必要とされる環境に対応しています。また、インテル EE for Lustre Solution搭載のHPC用 Dellストレージは、キャパシティのスケールと同じくらい簡単にスループットをスケールすることもできます。

パフォーマンス調査ではN-to-NとN-to-1ファイルタイプアクセスにおける、読み込みと書き込みの高いスループットを示しました。MDテスト結果では、メタデータファイルオペレーションのキャパシティが向上しました。PCI-e 3.0インターフェイスを使用すると、IB FDR HCAは高帯域幅を必要とするアプリケーションのための優れたスループットを実現します。

IOzoneやMDテストのような入手可能かつ業界スタンダードであるベンチマークツールは、簡単な方法で現在と将来の予測されるパフォーマンスについての概略を提供してくれます。これらのベンチマークツールによるテスト結果は、多くのアプリケーションやアプリケーショングループに必要とされるインテル EE for Lustre Solution搭載のHPC用 Dellストレージの構成要件を整理してくれる、極めて重要な情報となります。

インテル EE for Lustre Solution搭載のHPC用 Dellストレージはハイパフォーマンスコンピューティングに求められ、また並列ファイルシステムベースのストレージから得られるすべての恩恵を提供することができます。

補足資料 A：ベンチマークコマンドレファレンス

この項目では、インテル EE for Lustre Solution搭載のHPC用 Dellストレージのベンチマークテストで使用したコマンドについて説明していきます。

IOzone

IOzone 連続書き込み –

```
iozone -i 0 -c -e -w -r 1024K -I -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone 連続読み込み -

```
iozone -i 1 -c -e -w -r 1024K -I -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone IOPS ランダムリード / ライト –

```
iozone -i 2 -w -c -O -I -r 4K -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone コマンドライン	解説
-i 0	書き込みテスト
-i 1	読み込みテスト
-i 2	ランダムIOPSテスト
-+n	再テストなし
-c	算出中のクローズを含む
-e	算出中のフラッシュを含む
-r	記録サイズ
-s	ファイルサイズ
-+m	クラスターモードでIOzoneを実行するクライアントロケーション
-I	O ダイレクト使用
-w	テンポラリファイルのリンク解除（削除）不可
-+n	再テスト選択不可
-O	OPSの結果へ戻る

O_ダイレクトコマンドラインパラメータ (“-I”) により、IOzoneスレッドが実行されているコンピュートノードのキャッシュを回避することができます。

MDテスト - メタデータ

ファイルオペレーション -

```
mpirun -np $Threads -rr --hostfile /share/mdt_clients/mdtlist.$Threads /share/mdtest/mdtest.intel
-v -d /mnt/lustre/perf_test24-1M -i $Reps -b $Dirs -z 1 -L -I $Files -y -u -t -F
```

ディレクトリオペレーション -

```
mpirun -np $Threads -rr --hostfile /share/mdt_clients/mdtlist.$Threads /share/mdtest/mdtest.intel
-v -d /mnt/lustre/perf_test24-1M -i $Reps -b $Dirs -z 1 -L -I $Files -y -u -t -D
```

MDテスト コマンドライン引数	解説
-d	テスト実施用のディレクトリ
-v	verboseモード (オプションのインスタンスが 1 つずつ増えていく)
-i	繰り返しのテスト実施
-b	階層型ディレクトリ構造の分枝因子
-z	階層型ディレクトリ構造の深さ
-L	ツリー構造におけるリーフレベルのみのファイル
-I	ツリーのディレクトリ毎のアイテム数
-y	書き込み後のファイル同期
-u	各タスクによる独自のワーキングディレクトリ
-t	独自のワーキングディレクトリオーバーヘッドの時間
-F	ファイルのみのテスト実施（ディレクトリなし）
-D	ディレクトリのみのテスト実施（ファイルなし）

参考資料

インテル EE for Lustre Solution 搭載のHPC用 Dellストレージ概要
http://salesedge.dell.com/doc?id=0901bc82808e334f&ll=d&pm=160376162

インテル EE for Lustre Solution 搭載のHPC用 Dellストレージ構成ガイド
* 本ドキュメントについてはDellセールス担当者にお尋ねください

Dell PowerVault MD3420
http://www.dell.com/support/home/us/en/04/product-support/product/powervault-md3420/research

Dell PowerVault MD3460 & MD3060e
http://www.dell.com/support/home/us/en/04/product-support/product/powervault-md3460/research

Lustre ホームページ
http://www.whamcloud.com/lustre/
http://wiki.lustre.org/index.php/Main_Page

Dell HPC Solutions ホームページ
http://www.dell.com/hpc

Dell HPC Wiki
http://www.HPCatDell.com

インテルホームページ
http://www.intel.com

インテル HPDD Wiki
https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page

Mellanox Technologies ホームページ
http://www.mellanox.com

LSI 12Gb/s SAS HBA
http://www.lsi.com/downloads/Public/Host%20Bus%20Adapters/
LSI_PB_SAS9300_HBA_Family.pdf