DDN Lustre Seminar

東京工業大学様 TSUBAME3用ストレージシステム

DataDirect Networks Japan, Inc.

橋爪信明

2019/10/28

 © 2019 DDN Storage

Agenda

- DDN会社/製品紹介
- Lustreについて
- Lustreの特徴
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- TSUBAME3ファイルシステムの利用
- Q&A

 © 2019 DDN Storage

Agenda

- **DDN会社/製品紹介**
- Lustreについて
- Lustreの特徴
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- TSUBVAME3ファイルシステムの利用
- Q&A

 © 2019 DDN Storage

DDN | DataDirect Networks　会社概要

会社名	株式会社データダイレクト・ネットワークス・ジャパン （米国DataDirect Networks社100%子会社）
設立	日本法人　2008年1月 US本社　1998年
所在 本社製造拠点	日本法人　東京/大阪/名古屋 US　カリフォルニア・チャッツワース
従業員	650名以上 / 20カ国 35名 / 日本法人
事業内容	ストレージプラットフォーム・ファイルシステムソリューションの開発・販売・サポート 高性能ストレージシステムの販売・設計・構築・保守
国内代理店	SCSK・CTCSP・日本コムシス・Thymos HPCソリューションズ　など
販売パートナー	富士通・NEC・日立製作所・Crayジャパン・ HPE・新日鉄ソリューションズ　など

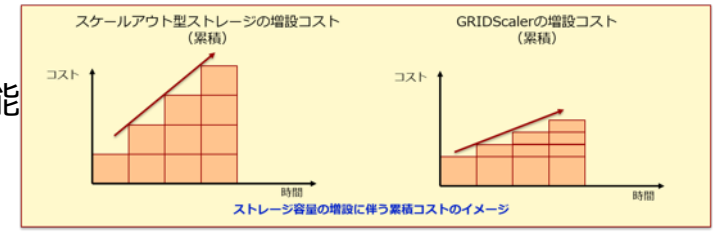
グローバル・マーケットで知られており、受賞実績も多数

DDN STORAGE　IDC　Inc.　Gartner　the 451 group　HPC　STORAGE　Federal Computer Week

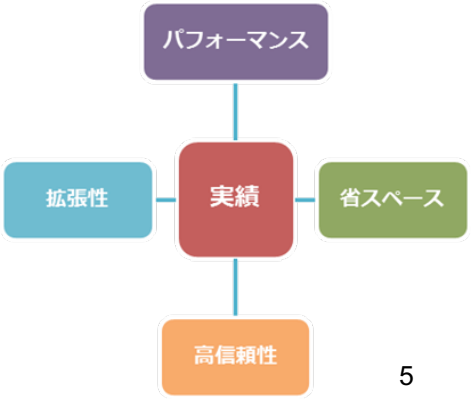
 © 2019 DDN Storage

DDN | DDNストレージの特長

- ハイパフォーマンスストレージ
  - ・最適に設計されたプラットフォームとユニークなOS機能による業界最速の並列ファイルシステムパフォーマンス（最大60GB/s：1ストレージコントローラペアあたり）
  - ・豊富なネットワークインターフェースをサポート（10/25/40/100Gb・16/32FC・Infiniband FDR/EDR）
- 高密度設計
  - ・4Uで90本の高密度設計が可能なディスクエンクロージャ
  - ・1 Rack で、最大物理容量12.6 PBを搭載可能（10ディスクエンクロージャ/14TB HDD使用時）
  - ・40PBを1ファイルシステムで構築した実績
- ご要望に応じた拡張性とシステム可用性
  - ・パフォーマンス（スループット/IOPS)と容量をフレキシブルに拡張可能
  - ・すべてのコンポーネントレベルで冗長化



- 実績
  - ・HPC/CAEの豊富な実績によりお客様が抱える様々なワークロードを解決
  - ・ワールドワイドで1300サイト、日本で350サイトの導入実績
  - ・HPC Top100ランクインの66%のお客様がDDNストレージを利用



- サポート
  - ・日本全国24時間365日サポート対応可能な体制

 © 2019 DDN Storage

DDN | 導入実績① 学術研究機関

- 北海道大学
- 東北大学
- 筑波大学
- 東京大学
- 東京工業大学
- 名古屋大学
- 京都大学
- 大阪大学
- 九州大学
- JCAHPC
- 理研
- 産総研
- NiCT
- JAXA
- JAMSTEC
- KAGRA
- NAOJ
- KEK
- 気象庁
- SACLA
- 統数研
- 慶應大学
- JAIST
- NAIST
- OIST
- 物材研

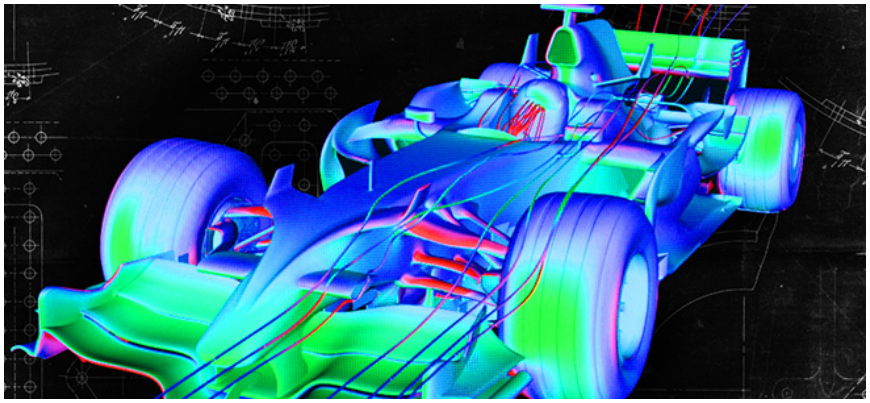
 © 2019 DDN Storage

DDN | 導入実績② ライフサイエンス・メディア・クラウド

- 東北メディカル・メガバンク機構
- 京都大学iPS細胞研究所
- 遺伝研
- かずさDNA研究所
- 東京大学Human Genome Center
- アステラス製薬
- 医薬基盤・健康・栄養研究所
- YAHOO Japan
- イマジカ
- 日本テレビ
- U-NEXT
- 白組

 © 2019 DDN Storage

DDN | 導入実績③　エンジニアリング



- マツダ株式会社　高速CAE用ファイルサーバ　https://ddn.co.jp/media/2016/11/29/24
- スズキ株式会社　高速CAE用ファイルサーバ　https://ddn.co.jp/media/2013/11/05/27
- 某大手自動車メーカー　高速CAE（流体・構造）用ファイルサーバ
- 某大手自動車メーカー　高速CAE・統合ファイルサーバ
- 某大手自動車メーカー　高速CAE・統合ファイルサーバ
- 某大手自動車メーカーサプライヤ　高速CAE用ファイルサーバ
- 某大手自動車メーカーサプライヤ　高速CAE（流体・構造）用ファイルサーバ
- 某大手製薬企業　研究開発用大規模ファイルサーバ
- 某大手光学機器企業　高速CAEファイルサーバ
- 某鉄鋼メーカー　高速大容量ファイルサーバ
- 某半導体企業　高速CAEファイルサーバ
- 某発動機・建機メーカー　高速CAE・統合ファイルサーバ
- 某重工業メーカー　高速CAE・統合ファイルサーバ
- 某総合電機メーカー　高速CAE（流体）ファイルサーバ

TTDC

VMware Horizon Viewを利用したVDI環境を構築

https://ddn.co.jp/media/2018/02/05/140

 © 2019 DDN Storage

DDN | 導入実績④　AI・マシンラーニング

産業技術総合研究所
人工知能研究センター

課題：研究競争力や産業界との連携の強化

解決策：GRIDScalerの採用で4PiBの実効容量、低遅延・高スループット、シングルスレッドI/O 3GB/秒を獲得

成果：Green500 Listで世界第3位を獲得、
AI研究開発用の共用計算プラットフォーム
として国内最大級の性能を獲得

https://ddn.co.jp/issue/machine_learning.html

- 国内　東京工業大学　学術国際情報センター　https://ddn.co.jp/media/2017/02/17/31
- 国内　某自動車メーカーサプライヤ　　自動運転研究開発
- 国内　某精密機械メーカー
- 海外　マイアミ大学　https://ddn.co.jp/media/2017/06/05/69
- 海外　某大手決済サービス会社
- 海外　フォーチュン100 SaaSプロバイダー
- 海外　某大手サービス会社

 © 2019 DDN Storage

DDN | DDN®プロダクトポートフォリオ

Burst Buffer

Infinite Memory Engine (IME)

クライアントと並列ファイルシステムの中間レイヤーに位置づけられるBurst Bufferであり一時的なBurst IOに対応し、多数プロセスからShared File へ小さなランダムIOアクセスなど並列ファイルシステムが不得意としているI/Oの高速化が可能。アプライアンスおよびソフトウエアによる提供。

IME Software IME

IME240

IME140

File System

DDN EXAScaler®

- オープンソースのLustreファイルシステムをベースに構成
- 膨大な数のクライアントに対し高スループットIO環境を提供
- 多数のHPCサイトで多くの実績

DDN GRIDScaler®

- 高いIO性能に加えてマルチプロトコルアクセス環境
- Snapshot/Replication/DR/サイト間連携など多彩な機能を提供
- 低速ストレージ、テープ装置等を用いた階層化機能の実現

Storage Platform + Operation System

SFA18K

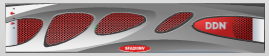
90GB/sec / 2M IOPS
Max 900 Drive

SFA7990

20GB/sec / 1M IOPS
Max 450 Drive

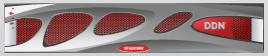
SFA®OS

SFA200NV



20GB/sec / 1M IOPS
Max 24 NVMe

SFA400NV

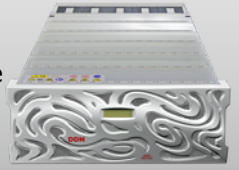


40GB/sec / 2M IOPS
Max 24 NVMe

Drive Enclosure / Drive

SS9012

12Gbps 90Drive
Enclosure



Drive

7.2Krpm NL-SAS
10/15Krpm SAS

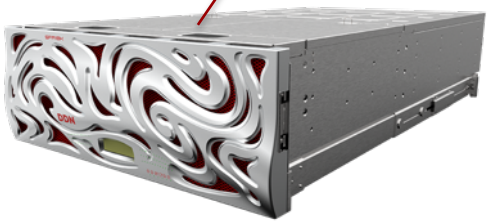
Server

DDN 1U Server

Intel Xeon skylakeベース
ファイルシステム管理用

 © 2019 DDN Storage

DDN | ストレージプラットフォーム コントローラ SFA18K 仕様



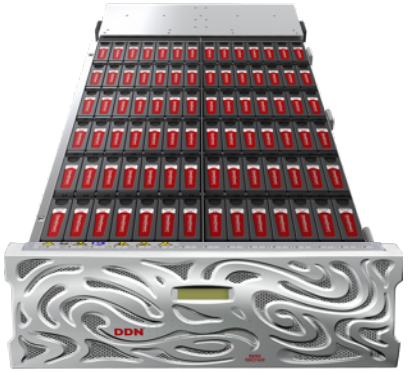
SFA18K+SS9012/10台接続構成

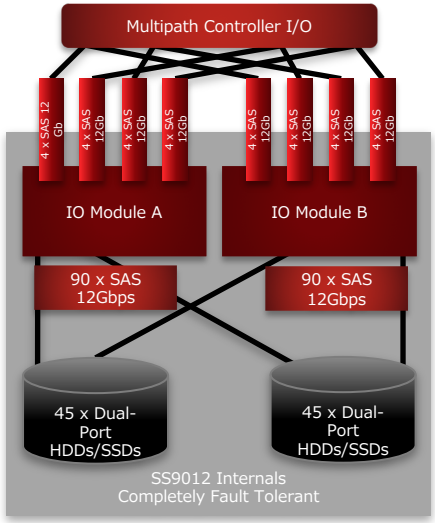
項目	製品仕様
フォームファクター	4U（Dual ControllerおよびBBU、2.5inch Drive slot x72を内蔵）
コントローラ	Active/Active Dual Controller
対応RAIDレベル	RAID1(1+1)・RAID5（8+1 or 4+1)・RAID6(8+2 or 4+2)
最大理論性能	シーケンシャル Read/Write 100GB/s
対応ドライブ数	ベースエンクロージャー 48 x PCIe NVMe SSD ・ 72 x 2.5inch SAS SSD 拡張エンクロージャー 1800 3.5inchドライブ ※20x SS9012構成時
キャッシュ	64GB / Controller（Battery-Backup)
ホストインターフェイス	16x EDR Infiniband 16x 100GbE イーサネット
ソフトウェア機能	LUNマッピングおよびマスキング・Read QoS・ポートゾーニング・Direct Protect DCR(Declustered RAID)
管理機能	Web GUI管理・CUI管理・メールベースの障害発報・SNMP

 © 2019 DDN Storage

DDN | ストレージプラットフォーム ディスク拡張エンクロジャ SS9012

SS9012ディスクエンクロージャは、4U で90本のドライブを収容可能な高密度ストレージエンクロージャです。SSD、SASの各種ドライブを混在可能であり、I/Oパスも含めてコンポーネントが冗長化されております。





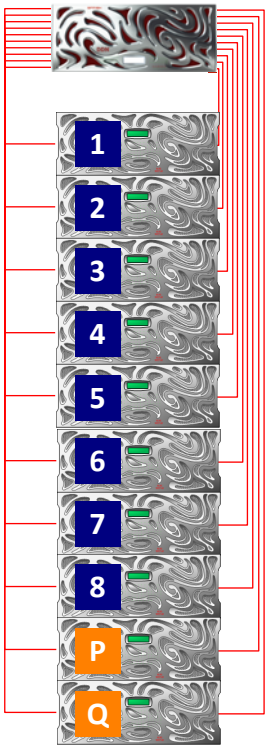
項目	製品仕様
フォームファクター	4U / 90 Drive
スロット数	90 (3.5inch/2.5inch)
対応ドライブ	SSD, SAS, NL-SAS (12Gb/s)
冷却ファン	5x 冷却ファン
電源	4 x 1200W 80PLUS Platinum
ホットスワップ対象部品	ドライブ、電源、冷却ファン、IOモジュール
LCD Display	Status, Power, Environmental Monitoring
LED	Power, Status, Monitoring, Drive Activity

 © 2019 DDN Storage

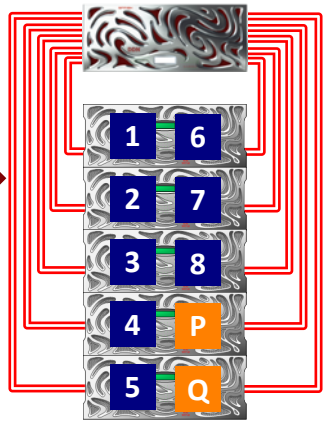
ダイレクト接続モデル

SFA18Kではコントローラ～ストレージエンクロージャ間がSAS 3.0インターフェイスにより直接接続されます。これにより、完全な冗長性と高内部帯域幅を提供します。(20エンクロージャ構成時はDaisy Chain接続)

SFA18KはSFAOS11のみのサポートであり、SFAOS11は従来型のRAIDではなく、Declustred RAID(DCR)を使用します。DCRでエンクロージャリダンダンシーを保証するためには、DCR Poolを構成するDriveの本数がエンクロージャ数の倍数である必要があります。



- 筐体間に跨ってRAID6を構成することにより2台ストレージエンクロージャ（5エンクロージャ構成時は、1ストレージエンクロージャ）に対してアクセス不能になった場合もデータが保護されます。(エンクロージャリダンダンシー)
  ※ 5エンクロージャ未満の構成時は本機能は利用できません。
- コントローラ、ストレージエンクロージャ間はSAS3.0により冗長構成(Active/Active)でダイレクト接続されます。これにより、コントローラは、全てのストレージエンクロージャに対して高帯域でアクセス可能です。また、ストレージエンクロージャ側のIOモジュール も冗長化されております。



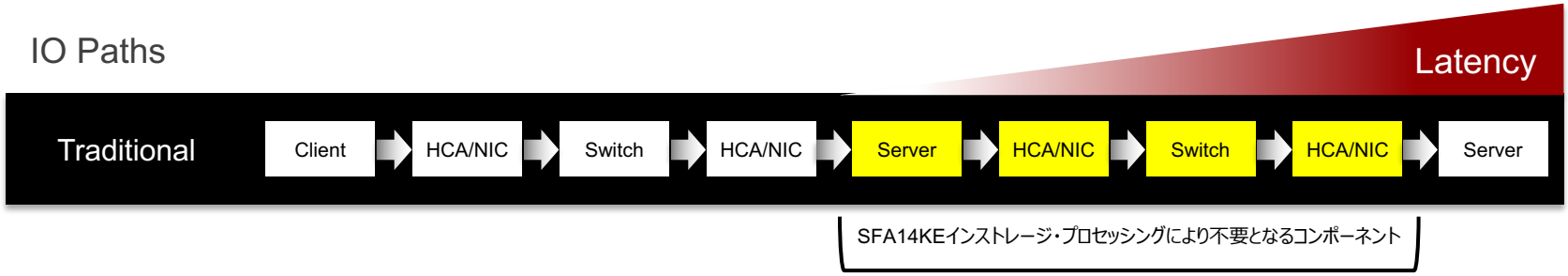
	SFA18K
Enclosures	SS9012
# of Enclosures	5, 6, 8, 10, 16, 20

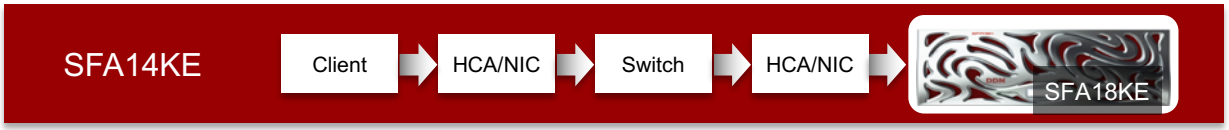
Enclosures Support/Configuration

SFA18KE
インストレージ・プロセッシング

SFA18KEはストレージコントローラ内でファイルサービスを実装することにより、外部サーバ、ネットワークアダプタ、スイッチなどを必要としません。SFA18KEの主なメリットは以下の通りです。

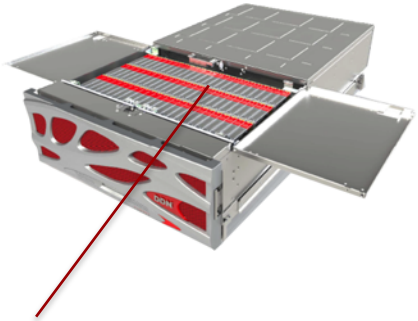
- 外部サーバ、ネットワークアダプタ、スイッチ購入コスト削減
- 構成の簡素化による管理コストの削減
- 低遅延でのストレージアクセスを実現（下記図参照）





データ/パケットの経路が簡素化することによりIOPSを向上

TSUBAME3使用ストレージ
DDN | ストレージプラットフォーム コントローラ SFA14KX 仕様



2.5インチSSD 72本搭載可能
もしくは48 個のNVMeを搭載可能

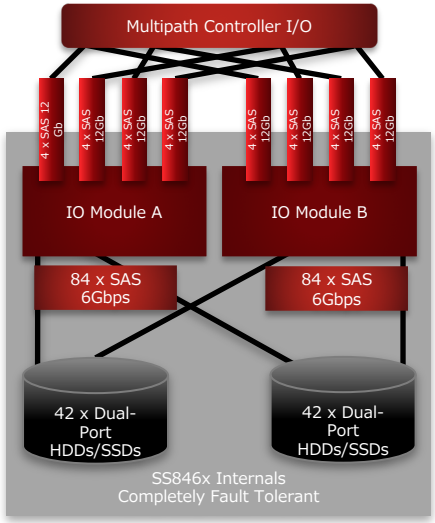
SFA14K+SS8462/10台接続構成

項目	製品仕様
フォームファクター	4U（Dual ControllerおよびBBU、2.5inch Drive slot x72を内蔵）
コントローラ	Active/Active Dual Controller
対応RAIDレベル	RAID1(1+1)・RAID5（8+1 or 4+1)・RAID6(8+2 or 4+2)
最大理論性能	シーケンシャル Read/Write 45GB/s
対応ドライブ数	ベースエンクロージャー 48 x PCIe NVMe SSD ・ 72 x 2.5inch SAS SSD 拡張エンクロージャー 1680 3.5inchドライブ ※20x SS8462構成時 1440 3.5inchドライブ ※16x SS9012構成時
キャッシュ	64GB / Controller（Battery-Backup)
ホストインターフェイス	6x EDR/FDR Infiniband 4 x Omni-path 4x 100GbE イーサネット 24 x16GbpsFC
ソフトウェア機能	LUNマッピングおよびマスキング・Read QoS・ポートゾーニング・Direct Protect
管理機能	Web GUI管理・CUI管理・メールベースの障害発報・SNMP

 © 2019 DDN Storage

TSUBAME3使用ストレージ
DDN | ストレージプラットフォーム　ディスク拡張エンクロジャ　SS8462

SS8462ディスクエンクロージャは、4U で84本のドライブを収容可能な高密度ストレージエンクロージャです。SSD、SAS、SATAの各種ドライブを混在可能であり、I/Oパスも含めてコンポーネントが冗長化されております。



項目	製品仕様
フォームファクター	4U / 84 Drive
スロット数	84 (3.5inch/2.5inch)
対応ドライブ	SSD, SAS, NL-SAS (6Gb/s)
冷却ファン	4x 冷却ファン
電源	2 x 1200W 80PLUS Platinum
ホットスワップ対象部品	ドライブ、電源、冷却ファン、IOモジュール
LCD Display	Status, Power, Environmental Monitoring
LED	Power, Status, Monitoring, Drive Activity

 © 2019 DDN Storage

Agenda

- DDN会社/製品紹介
- **Lustreについて**
- Lustreの特徴
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- TSUBAME3ファイルシステムの利用
- Q&A

 © 2019 DDN Storage

Lustreについて

- コンピュートクラスタ(HPC)向け並列分散共有ファイルシステム
  - 多くのシステムから一つのファイルシステムを利用可能
  - 高IOスループットと大容量を実現
  - 10万台規模のクライアントを構成可能
- HPCおよびその他の分野で利用
- オープンソース、コミュニティ主導の開発
- 種々の高速ネットワークインタフェースをサポート
  - Infiniband, OmniPath等
- 高可用性運用構成も可能

 © 2019 DDN Storage

Why Lustre?

- ネットワークファイルシステム：CIFS, NFS,,,
- 通常はボトルネックがあり、低パフォーマンス
- 理由は少ないサーバ台数、低いネットワーク帯域等々、、、
- HPCおよびIOインテンシブアプリケーションにとって大きな問題
- Lustreは、多くのクライアントからなる大規模システム向けに高いパフォーマンスを提供可能

 © 2019 DDN Storage

Lustreと言えば、、、

- 古典的ファイルシステムでは、、、
  - RAIDストレージ、非RAIDストレージ
  - ノードから直接ファイルシステムにアクセス可能
  - スケーラビリティ・ボトルネック

- 並列分散ファイルシステム
  - ネットワークベースの分散データシステム
  - サーバ数に比例してパフォーマンスが向上
  - データの一貫性確保のための機能実装

- **Lustreは代表的な並列ファイルシステム**

 © 2019 DDN Storage

Lustreの歴史

- 2000年代半ば、米国エネルギー省(DOE)科学部および国家核安全保障局(NNSA)のファンドによって開発開始
- ClusterFS(CFS)がファンドを受ける
- 2007年、サンマイクロシステムズがCFSを買収
- 2010年、オラクルがサンを買収したが、オラクルはHPCを知らなかった。。。
- Lustreエンジニアはオラクルを離れWhamCloud/Xyratex/DDNへ
- 2011年、インテルがWhamCloudを買収
- XyratexがLustre商標、ロゴ、関連資産(サポート)をオラクルから引継ぎ（その後、Communityに返還)
- 2014年、SeagateがXyratexを買収
- 2017年、CrayがSeagateからClusterStor製品ライン(Xyratex)を買収

- **2018年、DDNがインテルのLustre Business Unitを買収**

 © 2019 DDN Storage

DDNによるIntel Lustre Business Unit買収

- 買収後も100% Lustre CommunityにCommit
  - 独自Branchの作成はしません
- Intel Lustreの顧客に対するサポートを継続
  - 他社製HW上のLustreもサポートします
- DDN ExaScaler
  - 従来通り、Lustre Community EditionをCoreとし付加価値を追加しパッケージング
- Lustre開発者の殆どがDDNに移籍
  - 元々CFS/Whamcloudの人間が殆どで顔見知りであり、Communityで協業していた

 © 2019 DDN Storage

DDN ExaScalerとLustre

- DDN Lustre(EXAScaler)は、Lustre Community Editionをベースに性能および機能強化したDDNが提供するLustreディストリビューションです。全世界で300を超えるシステムで動作しています

- DDNのハードウェアに最適化され、お客様に提供するシステムと同等構成にてテストおよび検証されています。また、DDNのレベル1, 2, 3サポートにてストレージハードウェアとLustreのシームレスなサポートをDDNよりワンストップにて提供します

DDNのハードウェアに最適化、追加機能強化 追加テストおよびbugfix	DDN Lustre (EXAScaler)
コミュニティバージョンのLustre-2.xコードベース	Lustre Community Edition

© 2019 DDN Storage

Agenda

- DDN会社/製品紹介
- Lustreについて
- **Lustreの特徴**
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- TSUBAME3ファイルシステムの利用
- Q&A

 © 2019 DDN Storage

Lustreの特徴

パフォーマンス	1TB/secを越えるスループットを実現可能
アーキテクチャ	オブジェクトベースファイルシステム
適応性	様々なネットワークインタフェースとストレージ製品をサポート
スケーラブル	分散ファイルオブジェクト処理により10万クライアント規模をサポート
安定性	製品自体の安定性とフェイルオーバ機能
モジュラー型	機能拡張が容易なインタフェース
高可用性	HA構成によりSPOF無し
拡張性	オンライン容量拡張
OSサポート	ほとんどのLinuxディストリビューションで動作
互換性	POSIX互換
アカウンティング	ユーザ、グループおよびプロジェクトクォータ
オープン	オープンソース
コミュニティ	LAD(Lustre Admins Developers) & LUG(Lustre User Group)
Get Involved!	OpenSFS and/or EOFS

 © 2019 DDN Storage

Lustre Numbers

- TOP100システム中の75以上のシステムで使用
- 導入システムでの最大値

Rawバンド幅と比較したFile IO%	~90%
実効シングルOSS IO性能	> 8 GB/s
実効シングルクライアントIO性能	> 4 GB/s
実効総IO性能	> 1 TB/s
平均メタデータ操作回数	50,000+ ops/s ???
最大ファイルサイズ	64PB

 © 2019 DDN Storage

Lustre Numbers

- 実装上の制限値

最大ファイルシステムサイズ	1 EiB
最大OST数	8,150
最大MDT数	256 / ファイルシステム
最大ファイル数 / ディレクトリ	100万(ldiskfs)、2^48(ZFS)
最大ストライプ数	2,000
最大ストライプサイズ	< 4 GiB
最小ストライプサイズ	64 KiB
最大オブジェクトサイズ	16 TiB(ldiskfs)、256 TiB(ZFS)
最大ファイルサイズ	31.25 PiB(ldiskfs)、8 EiB(ZFS)

 © 2019 DDN Storage

Lustre Limitations

- データとメタデータが厳密に分離
  - 2.11よりDoM(Data on MDT)が実装されたため、2.11以降は異なる
- シングルMDTのメタデータ性能
  - DNE(Distributed Namespace Environment)の利用により改善
- RPC Size : 1MB, 最大16MB(Large Bulk IO)
- 最大OSTサイズはLustreのバージョンによる
  - 実システムでは256TiB程度が上限
- ext4(ldsiskfs) バックエンドファイルシステムによる制限
  - ZFSも利用可能
- ~~ディレクトリクォータ~~
  - 以前はユーザ、グループクォータのみ
  - プロジェクトクォータにてディレクトリクォータをサポート

 © 2019 DDN Storage

Lustreの利用

- Lustreに適するのは、、、
  - ラージファイル
  - シーケンシャルアクセス
  - プロセス毎に異なるファイルに書き込む並列アプリケーション

- Lustreが苦手とするのは、、、
  - 大量の小さいファイル
  - 莫大なメタデータ操作、ただしバージョンアップの度にに改善
  - OST容量の平準化

 © 2019 DDN Storage

今後リリース予定の機能

- 障害復旧処理の迅速化
- ChangeLogsを利用したHSMインテグレーション
- ZFS、BtrFSをバックエンドファイルシステムとして利用
- ただし、Roadmapはかなり流動的です、、、
- OpenSFSとEOFSが新機能の実装について決定します

 © 2019 DDN Storage

Agenda

- DDN会社/製品紹介
- Lustreについて
- Lustreの特徴
- **Lustreのアーキテクチャ**
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- TSUBAME3ファイルシステムの利用
- Q&A

 © 2019 DDN Storage

Lustre略語集

DNE	**D**istributed **N**amespace **E**nvironment	単一ネームスペースを複数MDTで管理する機能
FID	**F**ile **ID**entifier	1ファイルシステム内のオブジェクトを示す128bitの一意な識別子
LMV	**L**ogical **M**etadata **V**olume	複数MDTを扱う為のクライアントソフトウェアレイヤ
LNET	**L**ustre **NET**twork	Lustre Networkの総称
LOD	**L**ogical **O**bject **D**evice	複数のMDT, OSTへのアクセスを管理するMDSソフトウェアレイヤ
LOV	**L**ogical **O**bject **V**olume	複数OSTを扱う為のクライアントソフトウェアレイヤ
MDC	**M**etadata **C**lient	MDSへのインタフェースとなるクライアントソフトウェアレイヤ
MDD	**M**etadata **D**evice **D**river	POSIXファイルアクセスを管理するMDSソフトウェアレイヤ
MDS	**M**eta**D**ata **S**erver	クライアントからファイルシステムネームスペース(inode, path, permission等)へのアクセスを管理するソフトウェアサービス
MDT	**M**eta**D**ata **T**arget	ファイルシステムメタデータ(属性, inode, ディレクトリ, xattr等)を保持するストレージデバイス
MGS	**M**ana**G**ement **S**erver	クライアント・サーバの構成情報を管理するサービス
MGT	**M**ana**G**ement **T**arget	コンフィグレーション情報を保持するストレージデバイス
NID	**N**etwork **ID**entifier	ノードIDとネットワークタイプによってLustre Networkのエンドポイントを一意に識別するために使用

 © 2019 DDN Storage

Lustre略語集

OFD	**O**bject **F**ilter **D**evice	ファイルIOを扱うOSSソフトウェアレイヤ
OSC	**O**bject **S**torage **C**lient	OSSへのインタフェースとなるクライアントソフトウェアレイヤ
OSD	**O**bject **S**torage **D**evice	ldiskfs(ext4)やZFSなどのバックエンドディスクファイルシステムへアクセスするMDDとOFDを抽象化するサーバソフトウェアレイヤ
OSP	**O**bject **S**torage **P**roxy	あるMDSから別のMDSもしくはOSS上のOSDにアクセスするインタフェースとなるサーバソフトウェアレイヤ
OSS	**O**bject **S**torage **S**erver	ファイルシステムデータへのアクセス（read, write, truncate等）を管理するソフトウェアサービス
OST	**O**bject **S**torage **T**arget	ファイルシステムデータを保持するストレージデバイス

 © 2019 DDN Storage

Lustreを構成するノードタイプ

- ManaGement Servers (MGS)
  - システムに一つ、MDSと共存可能
- Meta-Data Servers (MDS)
  - ファイルシステム毎に最低一つ
  - 1 MDSに付き一つ以上のMeta Data Target(MDT)
- Object Storage Servers(OSS)
  - 通常ファイルシステム毎に複数存在
  - 各OSSが実ファイルデータを保持する一つ以上のObject Storage Target(OST)を提供
- クライアント
  - LustreクライアントソフトウェアがLustreファイルシステムへのアクセスを提供

 © 2019 DDN Storage

Lustre MGS

- ManaGement Server
- Lustreシステムの構成DBを管理するサーバ
- クライアント、OSSおよびMDSはMGSにアクセスできなければならない
- MDSと共存可能
  - DDNが国内で提供するLustreはほぼ100%共存構成

 © 2019 DDN Storage

Lustre MDS

- Meta Data Server
- ネームスペースを管理し、ファイルパーミッションやOST上のデータオブジェクトへのポインタなどをMDT上に保持する
- 通常、MDSはActive/Standby HA構成とする
- 一つのMDTは、一つのMDSで管理する必要がある
- メタデータのサイズは2KiB～16KiB

 © 2019 DDN Storage

Lustre OSS

- Object Storage Server
- 実際のファイルIOを提供し、配下のOSTへのリクエストを処理する
- 通常、OSSはActive/Active HA構成とする
- OSSが複数のOSTを管理する構成が一般的

 © 2019 DDN Storage

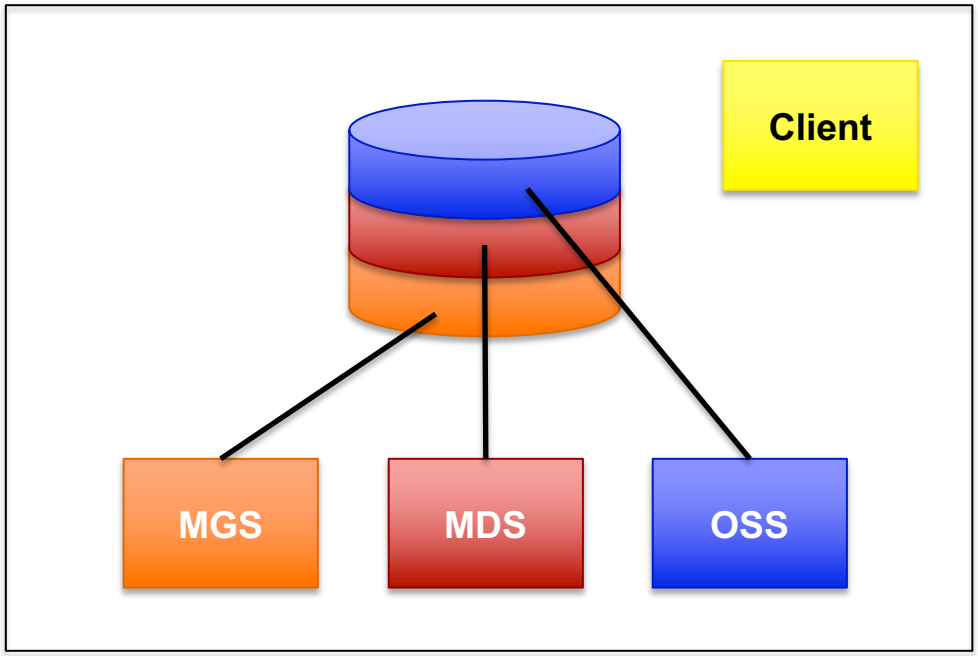
Lustreクライアント

- POSIX互換ファイルアクセスレイヤ
- DNE環境下での複数MDTへのアクセスを管理するLogical Metadata Volume(LMV)レイヤを持つ
- 複数のOSTへのファイルストライピングを管理するLogical Object Volume(LOV)レイヤを持つ
- 各サーバに対するクライアントレイヤを持つ
  - Object Storage Client(OSC) : OSSと協調
  - Metadata Client(MDC) : MDSと協調

 © 2019 DDN Storage

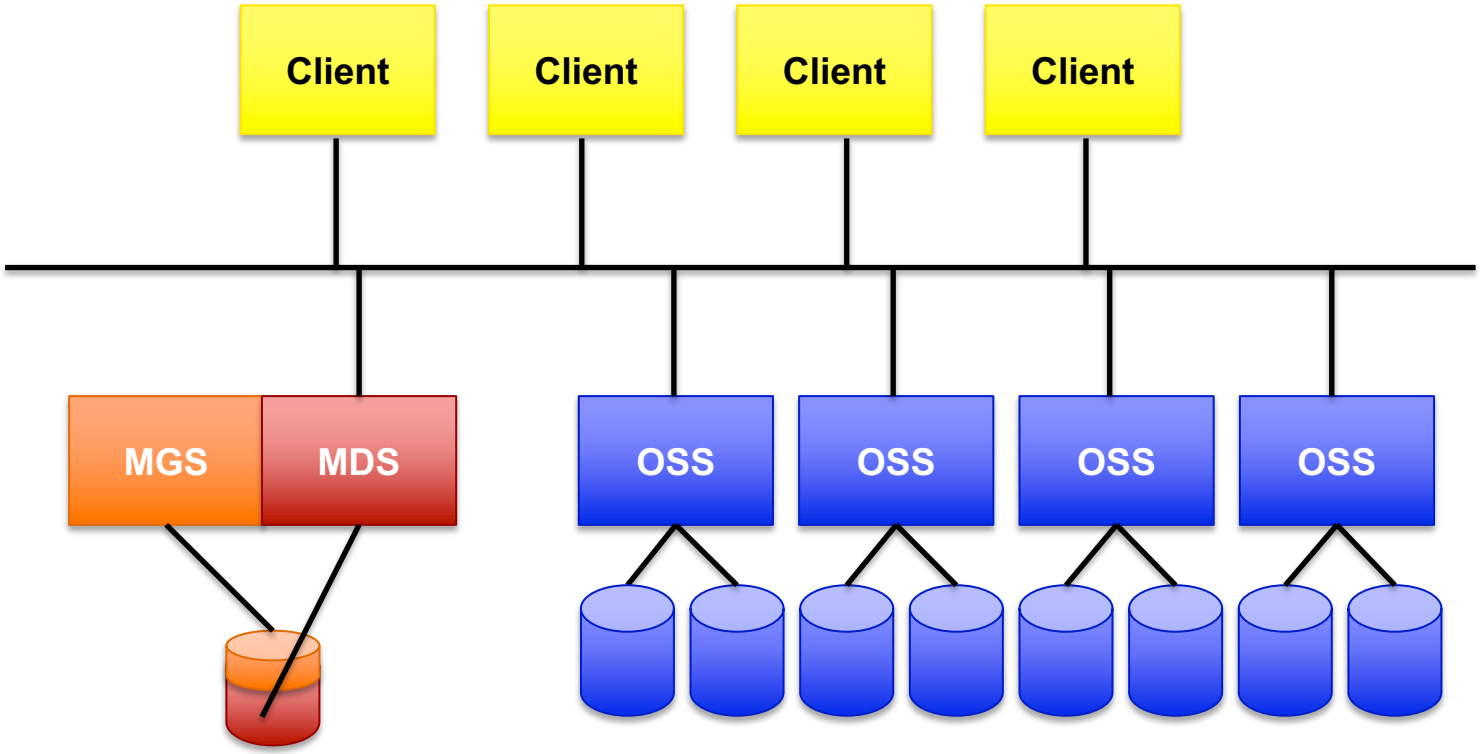
Lustre最小構成例

- 単一ノード構成
- 全機能を1サーバノードに実装

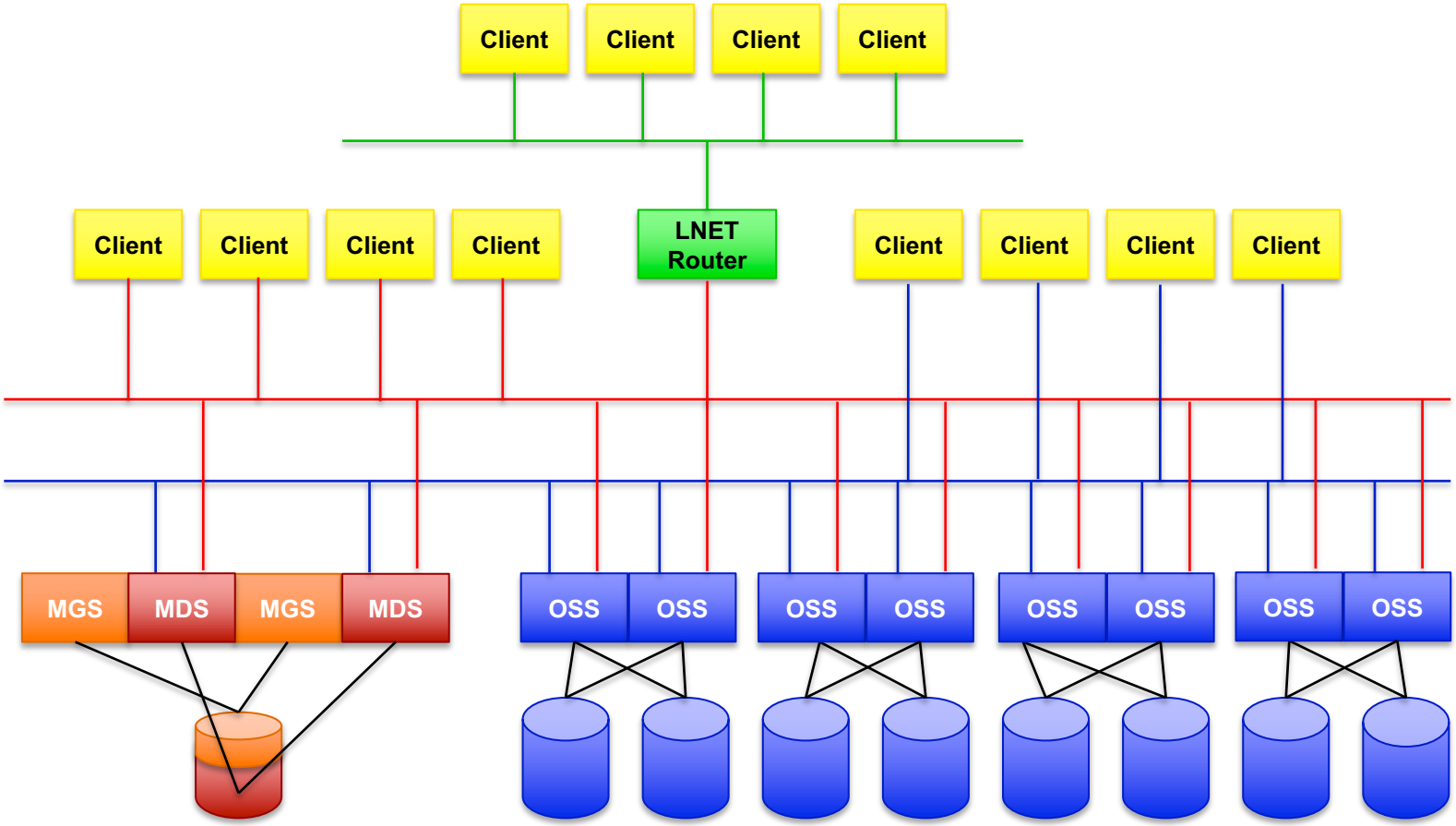


 © 2019 DDN Storage

一般的なLustre基本構成



複雑な構成

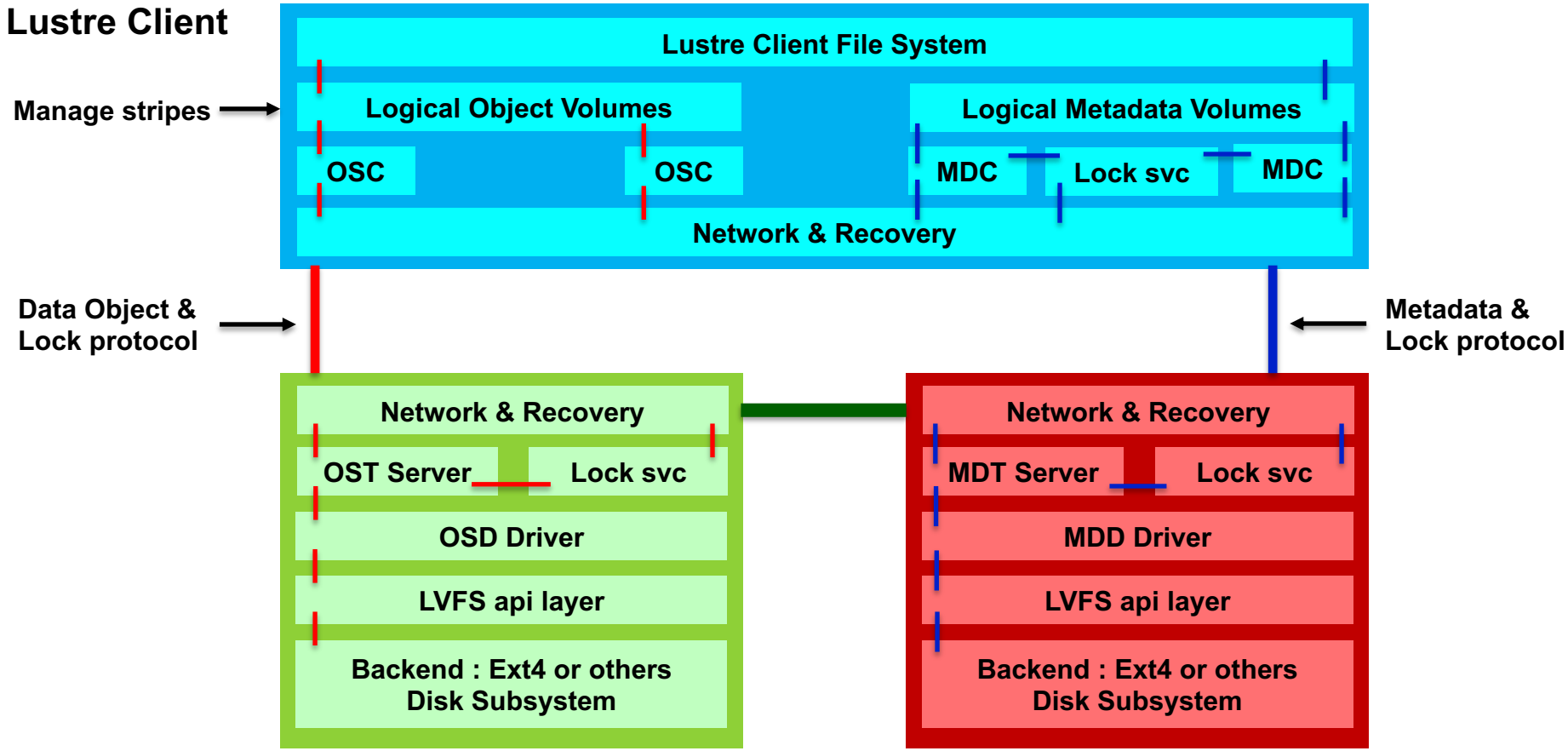

Client
Client
Client
Client
Client
Client
Client
Client
LNET Router
Client
Client
Client
Client
MGS
MDS
MGS
MDS
OSS
OSS
OSS
OSS
OSS
OSS
OSS
OSS

Agenda

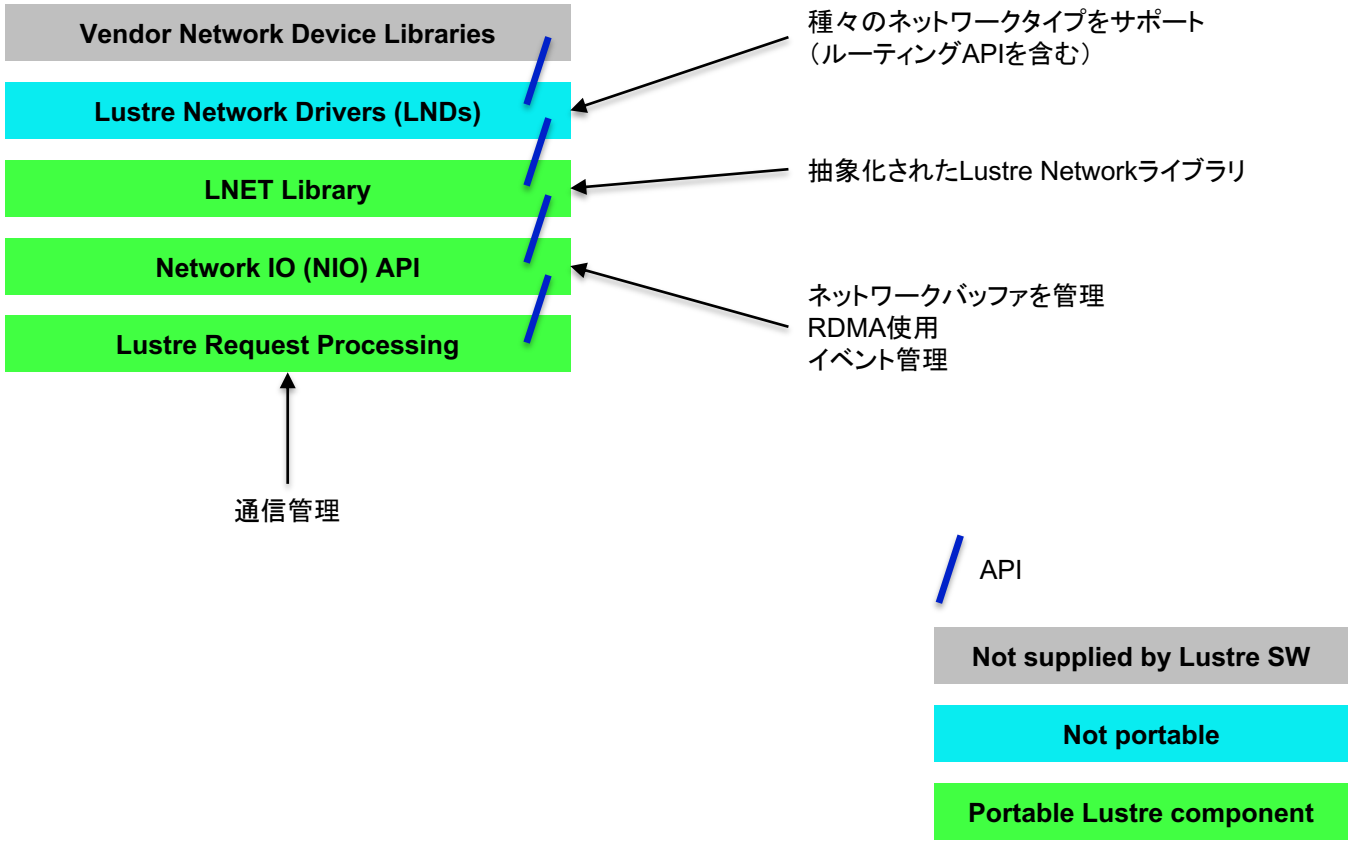
- DDN会社/製品紹介
- Lustreについて
- Lustreの特徴
- Lustreのアーキテクチャ
- **Lustreの各コンポーネント**
- TSUBAME3ストレージ概要
- TSUBAME3ファイルシステムの利用
- Q&A

 © 2019 DDN Storage

Lustreコンポーネント配置図



階層型ソフトウェアモジュール設計



Lustre FS Storage & IO

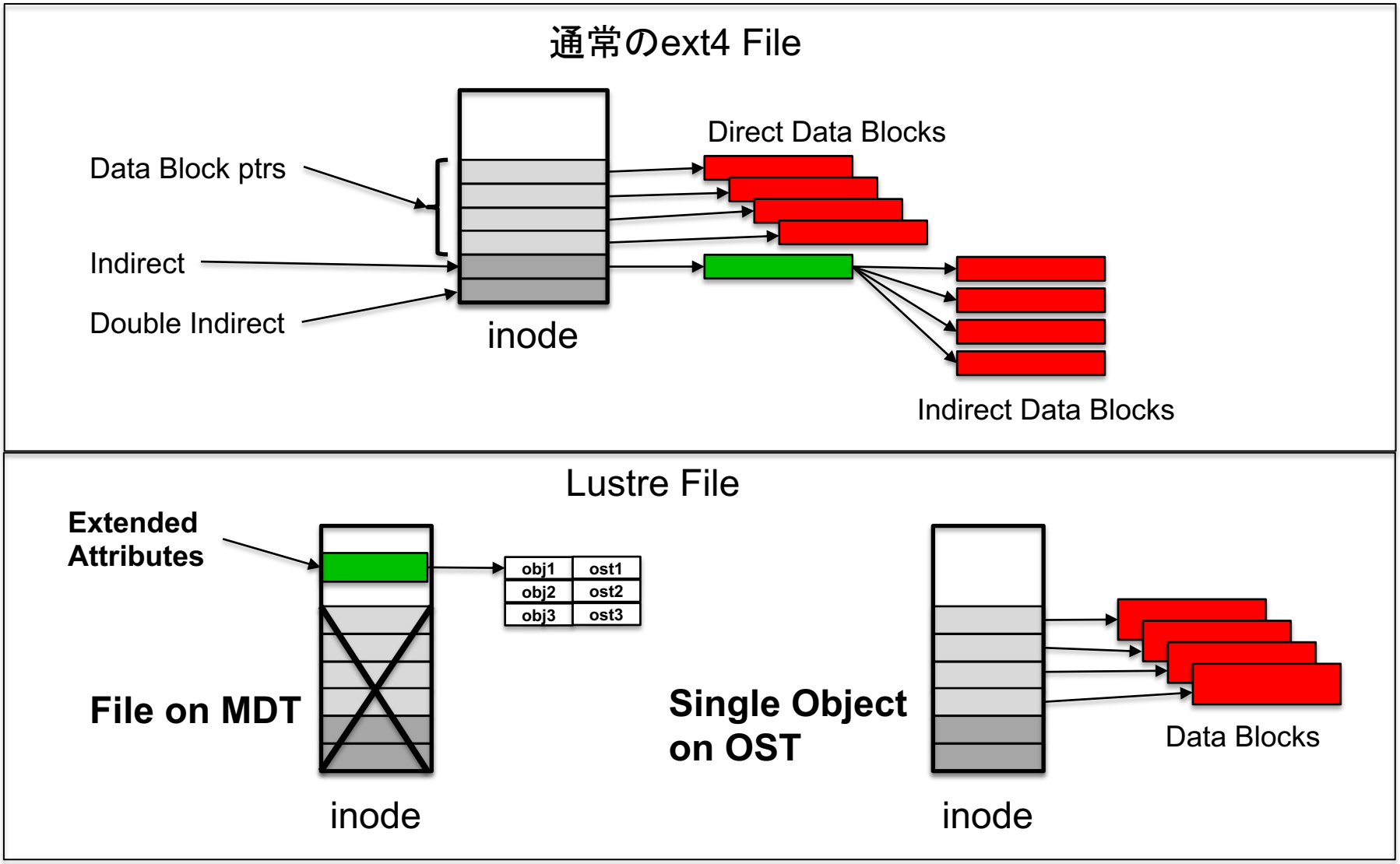
- FID (Lustre File IDentifiers)
  - 128bit
    - 64bit sequence number (1 FS内でユニーク)
    - 32bit object id
    - 32bit version number

- OST上のファイルデータの位置
  - MDTオブジェクトのExtend Attributeに保持

- クライアントからのIO(read/write)操作要求
  - MDS/MDTからファイルデータの位置情報を取得
  - 位置情報を元にOSS/OSTにファイルIOを要求

 © 2019 DDN Storage

Disk Layout

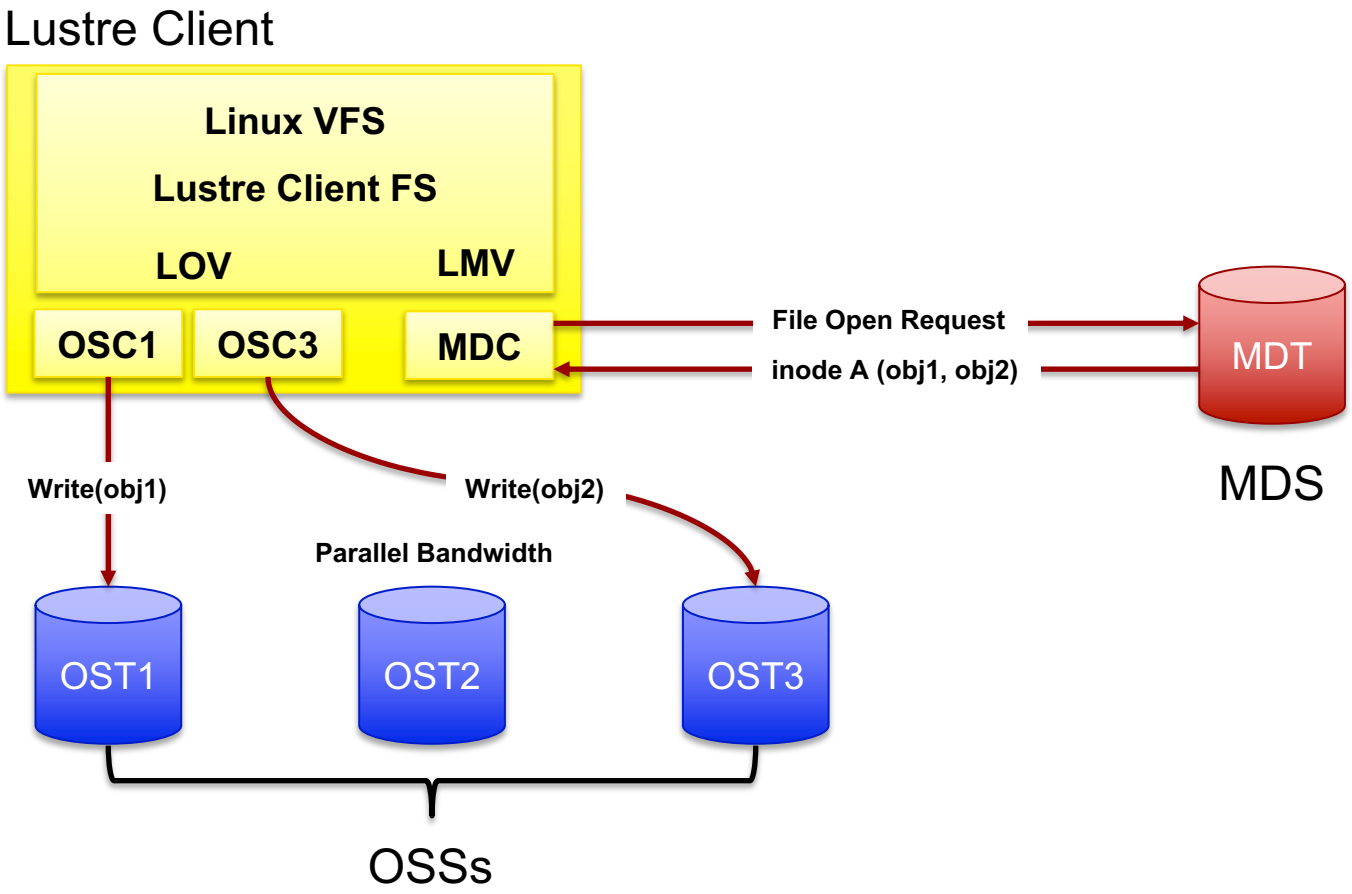

通常のext4 File
Data Block ptrs
Indirect
Double Indirect
inode
Direct Data Blocks
Indirect Data Blocks
Lustre File
Extended Attributes
obj1 ost1
obj2 ost2
obj3 ost3
File on MDT
inode
Single Object on OST
Data Blocks
inode

Lustre File IO

- クライアントがファイルOpenを要求
  - LMVがMDSへリクエストを送信
  - MDSはdirectory treeをロックし、ファイルをCreate/Openした後、directory treeのロックを解除
  - MDSはOpenしたファイルのinodeをLMVに返信
- LMVがクライアントアプリケーションにファイルディスクリプタを返す
  - inodeはOST位置情報を含むextended attributesを持つ
  - LOVはストライプを含むOSTの情報を元にファイルIOを行う
- Open後のファイルIOは主にクライアントとOSS間で行われる

 © 2019 DDN Storage

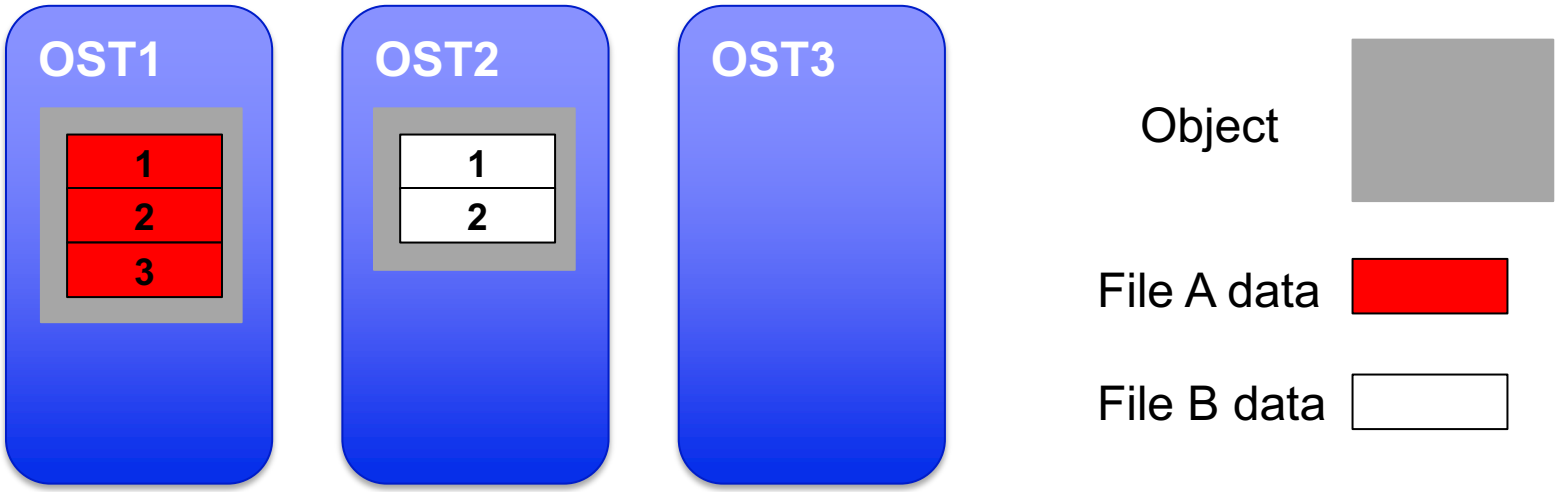
Lustre File IO



Lustre File Striping

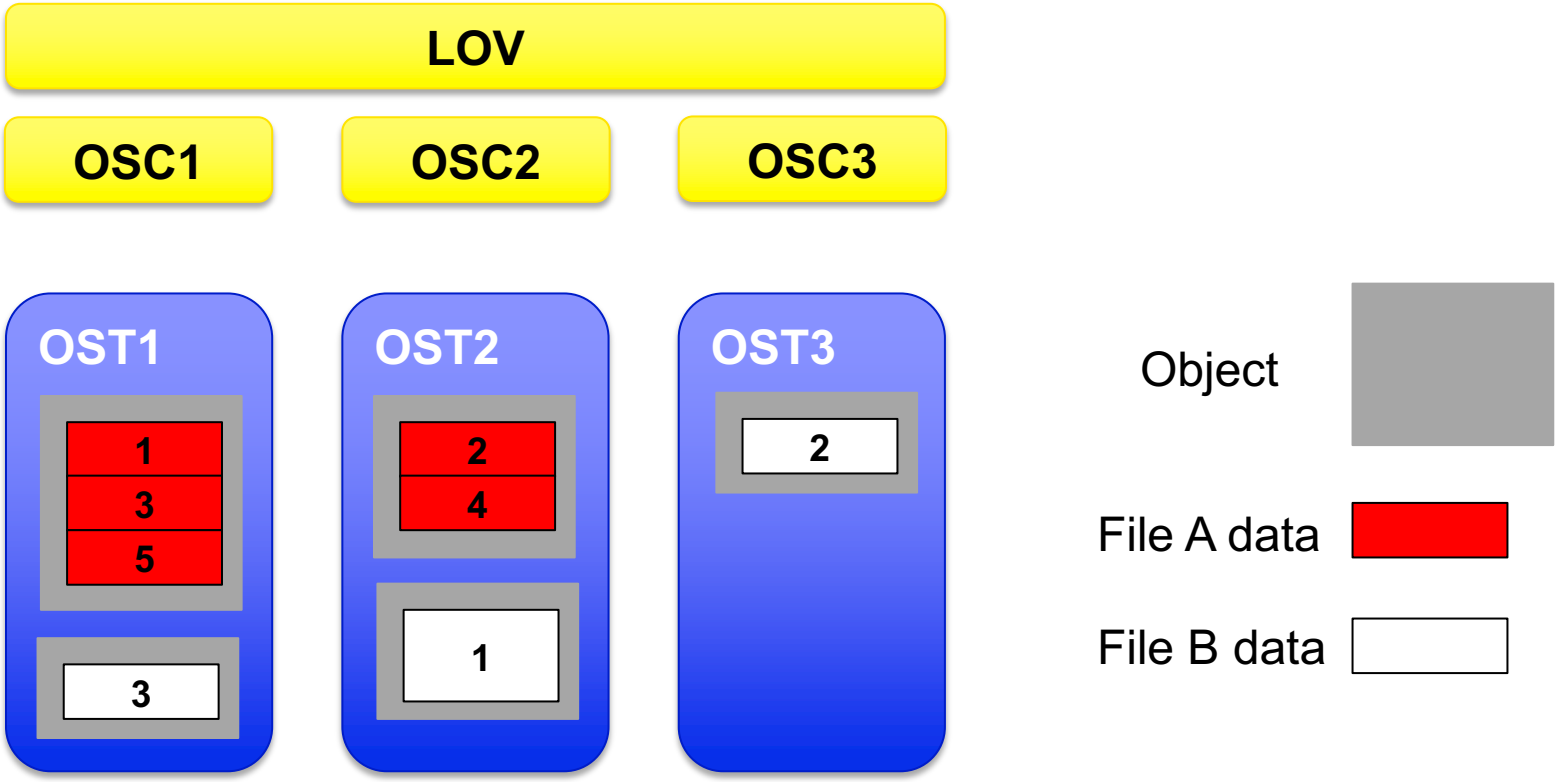
- File Stripingを利用することによりファイルを複数のOSS/OSTに分散して書き込むことが可能
  - Stripeをしない場合、Objectの最大サイズ(16TiB) == 最大ファイルサイズ

- Single Stripe(Stripe無し)例



Lustre File Striping

- Multiple Stripes例



Agenda

- DDN会社/製品紹介
- Lustreについて
- Lustreの特徴
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- **TSUBAME3ストレージ概要**
- TSUBAME3ファイルシステムの利用
- Q&A

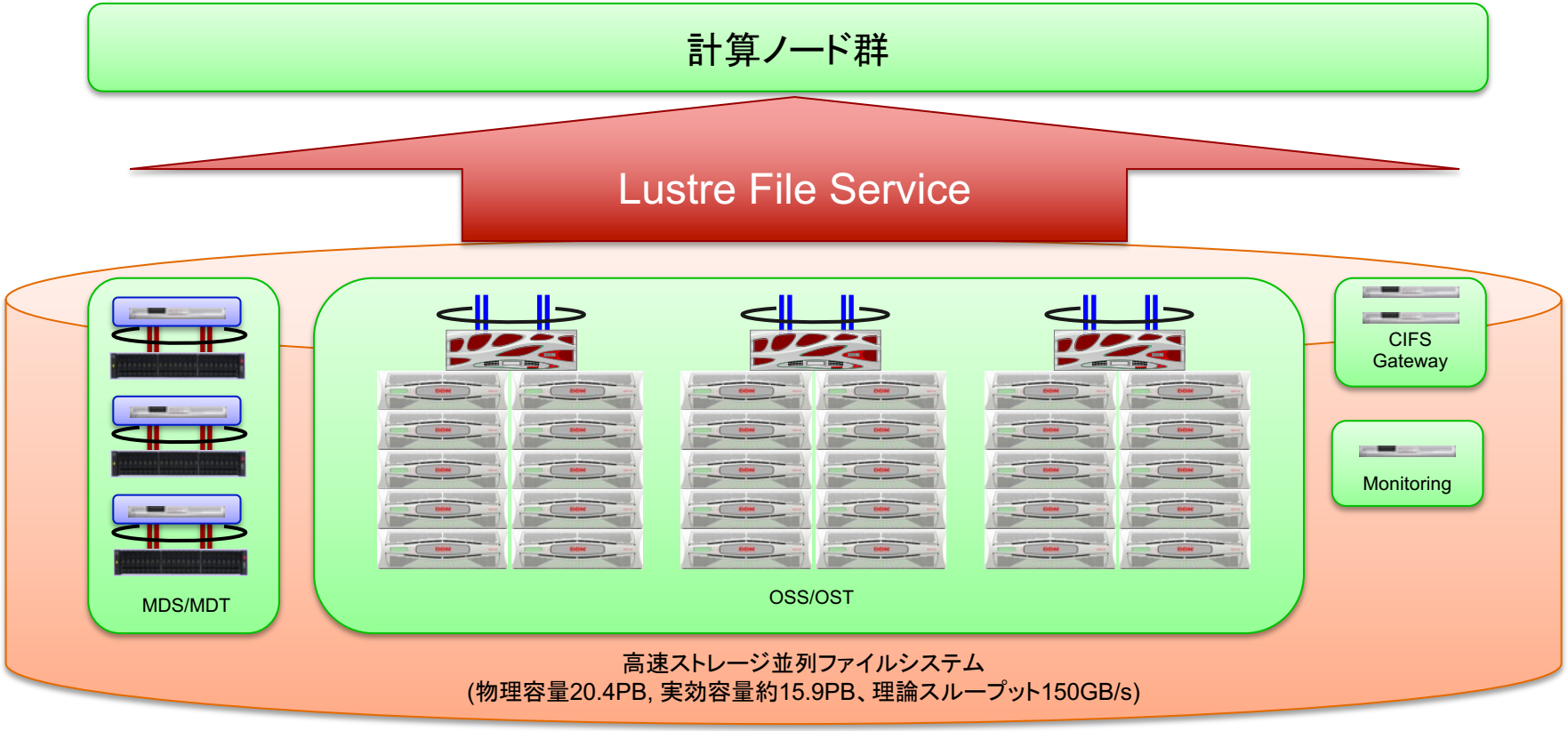
 © 2019 DDN Storage

TSUBAMEとLustre

- TSUBAMEは1.0からLustreを使用しています
- TSBUAME1.x
  - Sun Thumper + CFS Lustre
- TSUBAME2.x
  - DDN SFA10K + DDN Lustre(ExaScaler)
- TSUBAME3.0
  - DDN ES14KX + DDN Lustre(ExaScaler)

 © 2019 DDN Storage

TSUBAME3ストレージ構成図



TSUBAME3 Lustre FS概要

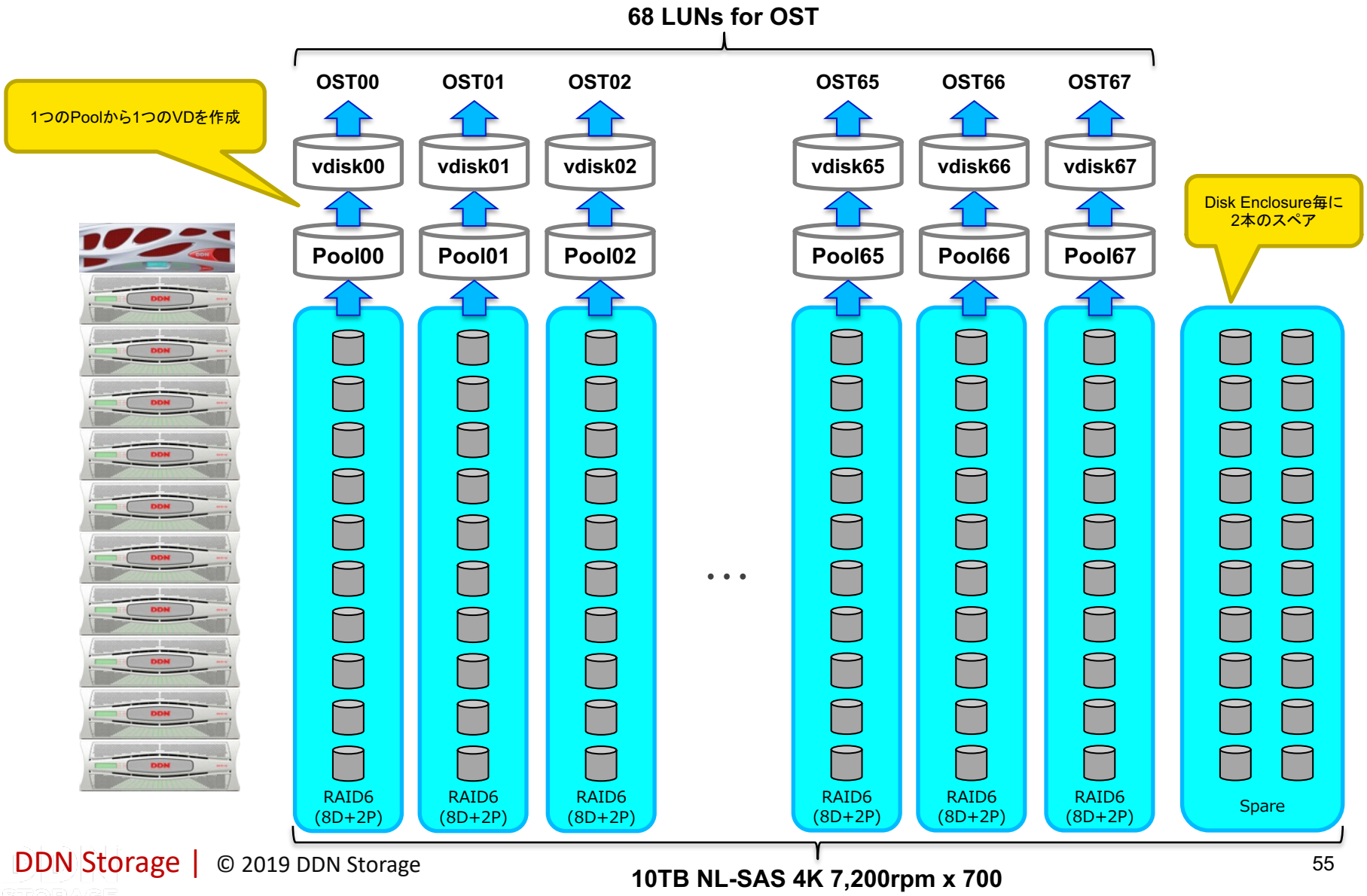
- 以下の3つのLustre FSを構成

FS Name	Mount Point	inode数	Size
T3_HS0	/gs/hs0	4,294,443,008	4,860.49 TiB(約5.344PB)
T3_HS1	/gs/hs1	4,294,443,008	4,860.49 TiB(約5.344PB)
T3_HS2	/gs/hs2	4,294,443,008	4,860.49 TiB(約5.344PB)

- MDSは DDN 1U Server 2台でActive/Standby構成
- MDTはDDN EF4024(600GB 15Krpm SAS x 24) 1基で構成
  - 10本のDiskを用いたRAID6(8D+2P) x 2で構成し、OS上でLVMを用いて１つのMDTとして構成
- OSSはDDN SFA14KXE上のVM 4台で2台ずつペアのActive/Active構成
- OSTはDDN SFA14KXE(10TB 4Kn 7.2Krpm NL-SAS x 700) 1基で構成
  - 10本のDiskを用いたRAID6(8D+2P)で構成し、１つのRAID6から１つのVolume(LUN)を切り出し構成
- ファイルシステムにはDNEを使用しない
- Lustre ClientはIntel OmniPath HFIを持つLinux Client

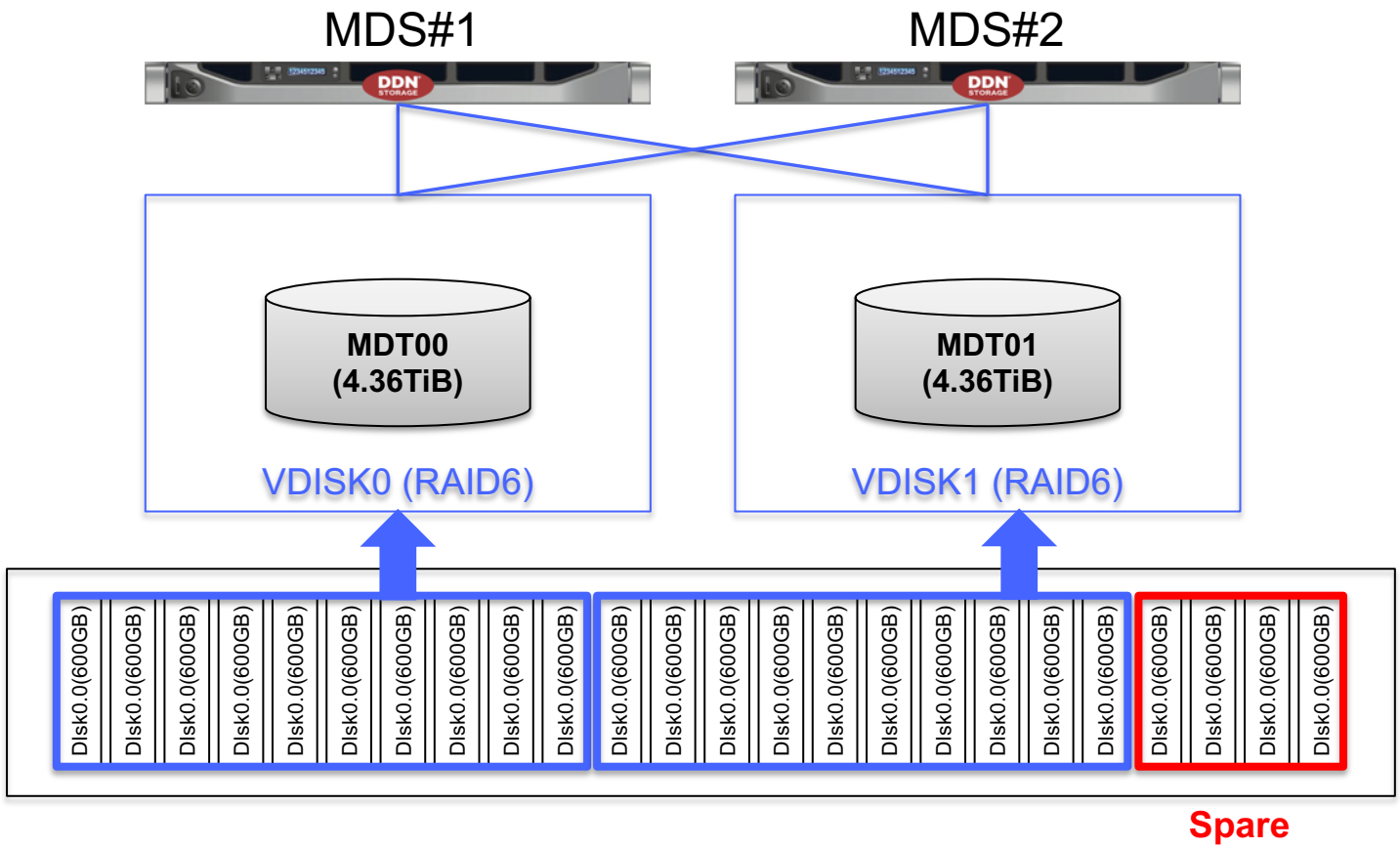
 © 2019 DDN Storage

TSUBAME3 OST構成



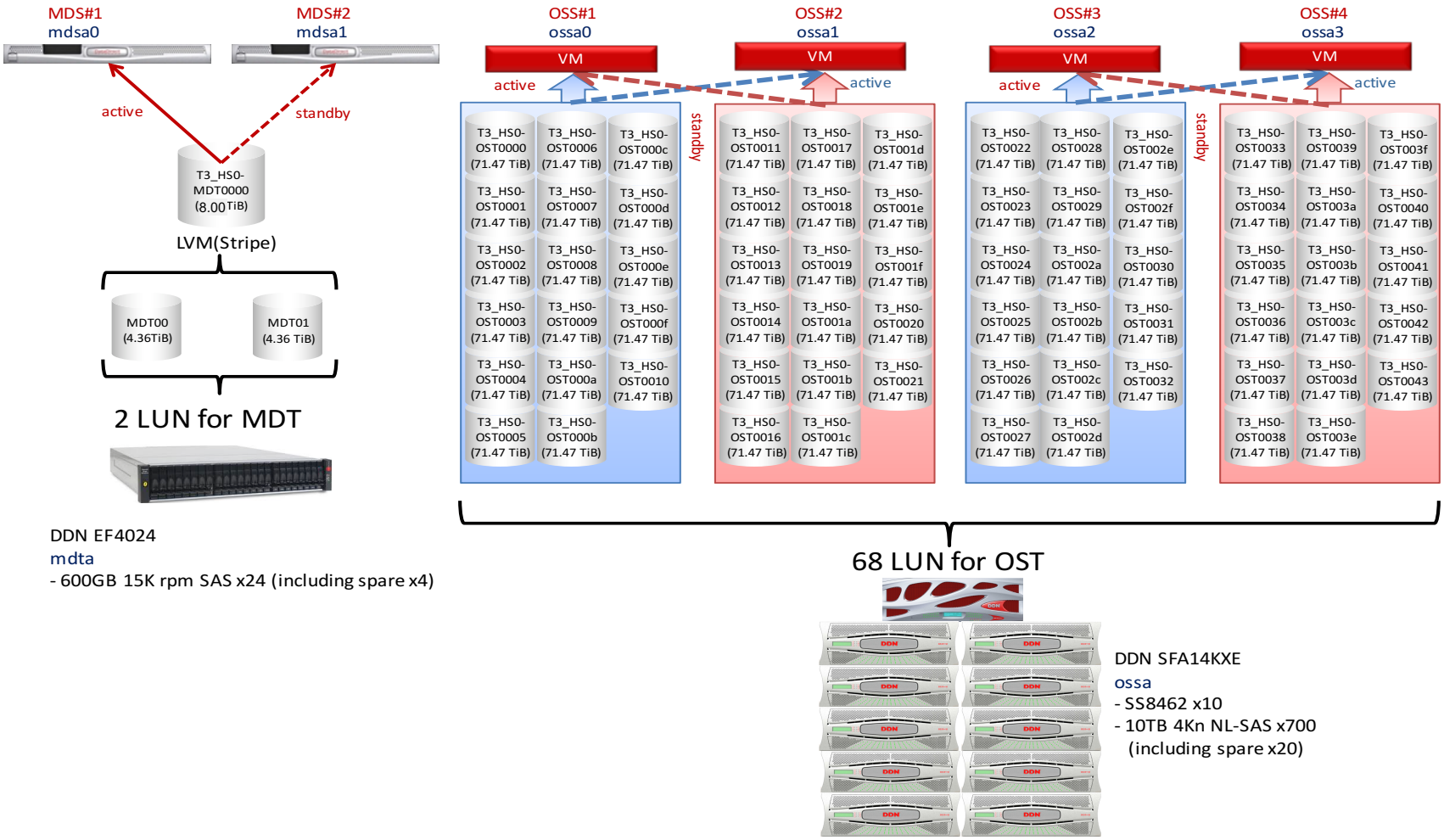
TSUBAME3 MDT構成

本構成を3セット使用



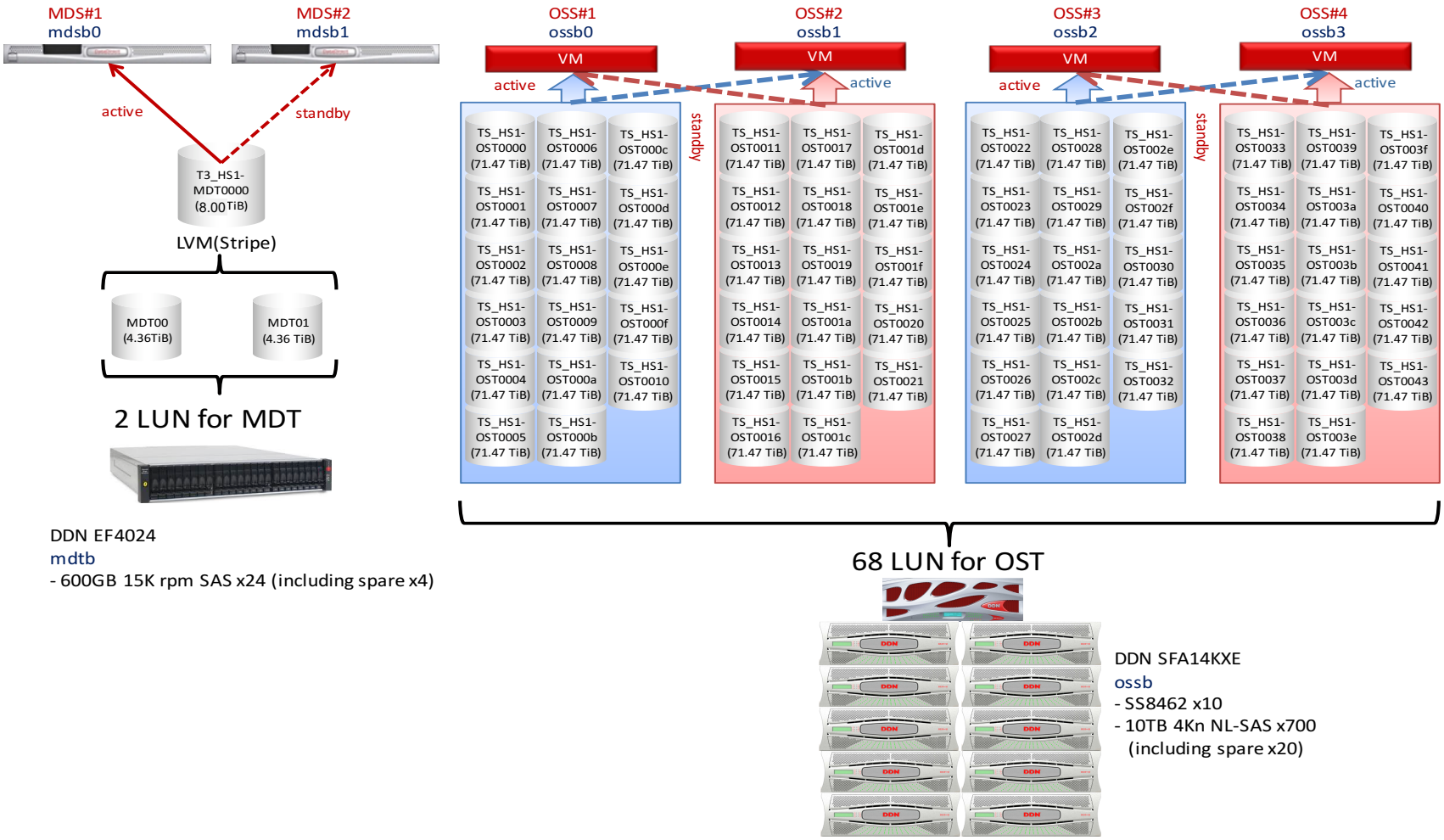
TSUBAME3 Lustre FS構成：T3_HS0

T3_HS0: /gs/hs0
実効容量: 4,860.499 TiB (FS format後)
inode数: 4,294,443,008



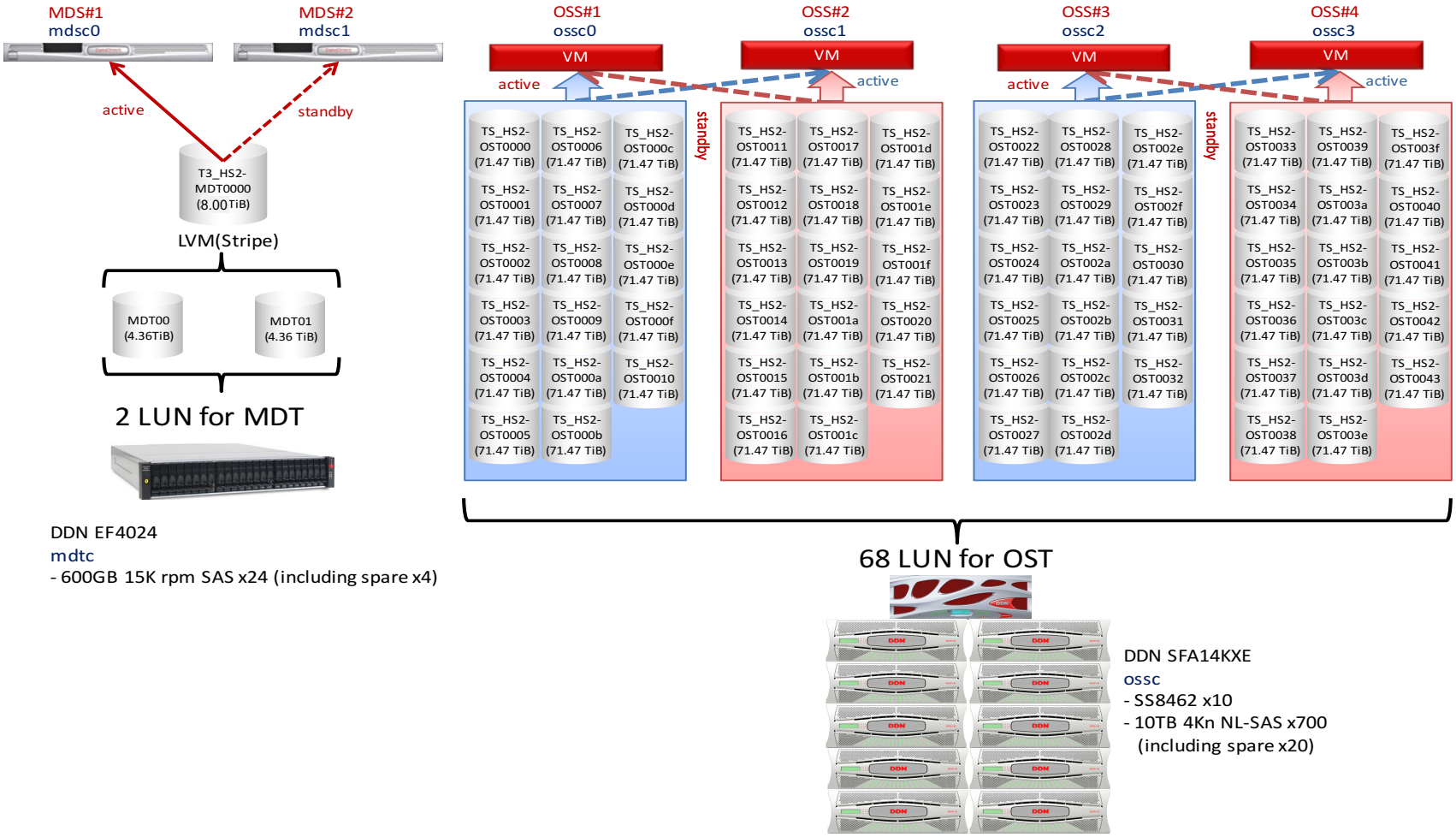
TSUBAME3 Lustre FS構成：T3_HS1

T3_HS1: /gs/hs1
実効容量: 4,860.499 TiB (FS format後)
inode数: 4,294,443,008

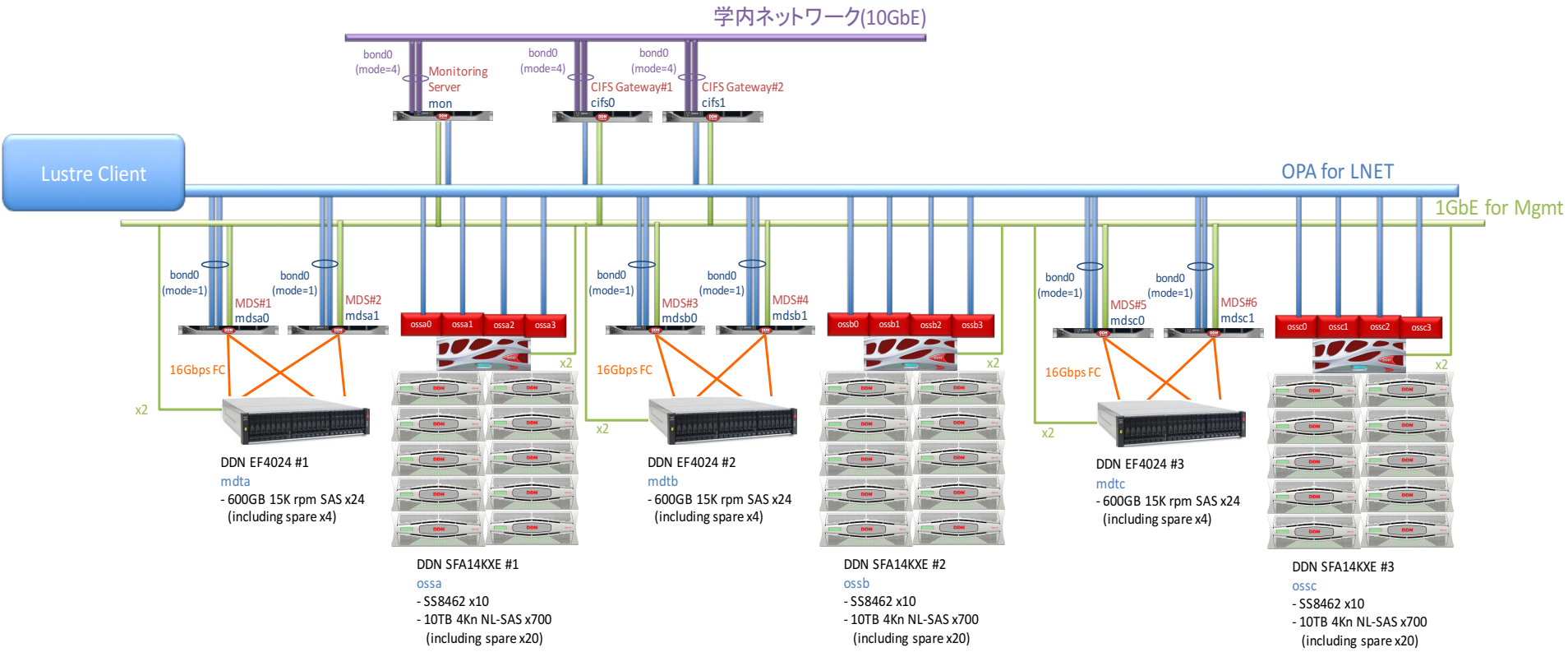


TSUBAME3 Lustre FS構成：T3_HS2

T3_HS2: /gs/hs2
実効容量: 4,860.499 TiB (FS format後)
inode数: 4,294,443,008



TSUBAME3 LNET構成



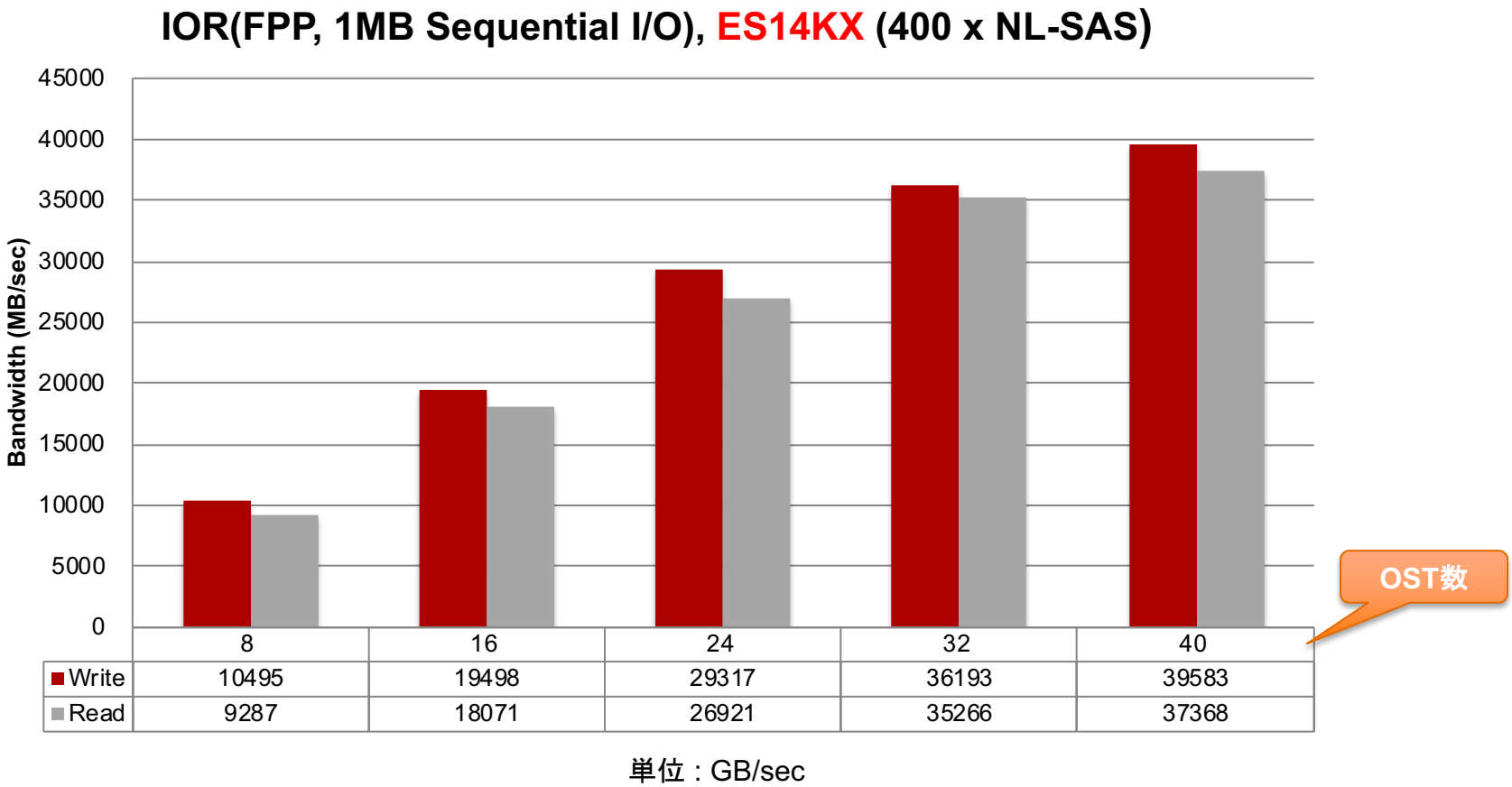
TSUBAME3 LNET概要

- LNETとして使用するOPA Networkは１つだが、LNETは4つ構成
- LNETとしては、OPA Network(@o2ib10, @o2ib11, @o2ib12, @o2ib13)を使用
- Lustre FSを構成するMDSは、2本のOPAをLinux Bondingで1つのInterface(bond0)として、LNET(@o2ib10, @o2ib11, @o2ib12, @o2ib13)を構成
- Lustre FSを構成するOSSは、1本のOPAでLNET(@o2ib10, @o2ib11, @o2ib12, @o2ib13)を構成
- OPA HFIを持つLinux Clientは、o2ib10～13を通して直接MDS/OSSへアクセス
- Windows ClientはCIFS Gateway経由でアクセス

 © 2019 DDN Storage

SFA14KXE(ES14KX)の性能

- TSUBAME3ストレージと同構成のピーク実効スループット
  - ベンチマーク参考値



	8	16	24	32	40
Write	10495	19498	29317	36193	39583
Read	9287	18071	26921	35266	37368

 © 2019 DDN Storage

Agenda

- DDN会社/製品紹介
- Lustreについて
- Lustreの特徴
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- **TSUBAME3ファイルシステムの利用**
- Q&A

 © 2019 DDN Storage

TSUBAME3ファイルシステム

- Home領域 : /home
  - ユーザあたり25GiBを割当
  - DDN GridScaler NAS (GPFSをNFS Export)

- 高速ストレージ領域 : /gs/hs0, /gs/hs1, /gs/hs2
  - グループディスクとして購入
  - DDN ExaScaler (Lustre FS)
  - Lustre Project Quotaを利用
  - 学内のWindows/Mac端末からCIFSアクセスが可能

今回の講習はLustreを利用する高速ストレージのみ説明しています

- ローカルスクラッチ領域 : /scr
  - 計算ノード上のSSD領域
  - XFS

- 共有スクラッチ領域 : /beeond
  - 複数計算ノードのSSDを纏めてスクラッチ領域として共有
  - 資源タイプFのf_nodeを利用したバッチスクリプトでのみ利用可能
  - BeeGFS On Demand(BeeOND)

 © 2019 DDN Storage

TSUBAME3 Lustreファイルシステム

```
ddn_user00@login0:~> df -H
Filesystem                                  Size  Used Avail Use% Mounted on
devtmpfs                                     34G     0   34G   0% /dev
tmpfs                                       34G  100M   34G   1% /dev/shm
tmpfs                                       34G  3.4G   31G  10% /run
tmpfs                                       34G     0   34G   0% /sys/fs/cgroup
/dev/sda3                                   251G   50G  202G  20% /
/dev/sda2                                   489M  188M  271M  41% /boot
/dev/sda5                                   739G  137G  602G  19% /data1
10.0.254.88:/gs/apps                              46T   16T   31T  34% /apps
10.0.254.89:/gs/home                              46T   16T   31T  34% /home
10.0.254.41@o2ib10:10.0.254.42@o2ib10:/T3_HS0  5.4P  1.6P  3.8P  29% /gs/hs0
10.0.254.43@o2ib10:10.0.254.44@o2ib10:/T3_HS1  5.4P  747T  4.6P  15% /gs/hs1
10.0.254.45@o2ib10:10.0.254.46@o2ib10:/T3_HS2  5.4P  5.4T  5.3P   1% /gs/hs2
```

 © 2019 DDN Storage

Lustreコマンド

- lctl
  - 管理者向けコマンド
  - Lustre FS Configuration Utility
  - Lustre Parameterの設定・変更などを行うことが可能

- lfs
  - 一部一般ユーザ権限で利用可能なコマンド
  - Lustre Utility
  - Stripe管理
  - Quota管理
  - find : 特定のOST上のファイル、stripeされたファイルなどを指定可能

 © 2019 DDN Storage

Lustreの状態確認

```
lfs > mdts
MDTS:
0: T3_HS0-MDT0000_UUID ACTIVE
MDTS:
0: T3_HS1-MDT0000_UUID ACTIVE
MDTS:
0: T3_HS2-MDT0000_UUID ACTIVE
lfs > osts
OBDS::
0: T3_HS0-OST0000_UUID ACTIVE
1: T3_HS0-OST0001_UUID ACTIVE
2: T3_HS0-OST0002_UUID ACTIVE
3: T3_HS0-OST0003_UUID ACTIVE
4: T3_HS0-OST0004_UUID ACTIVE
5: T3_HS0-OST0005_UUID ACTIVE
6: T3_HS0-OST0006_UUID ACTIVE
7: T3_HS0-OST0007_UUID ACTIVE
8: T3_HS0-OST0008_UUID ACTIVE
9: T3_HS0-OST0009_UUID ACTIVE
10: T3_HS0-OST000a_UUID ACTIVE
11: T3_HS0-OST000b_UUID ACTIVE
12: T3_HS0-OST000c_UUID ACTIVE
13: T3_HS0-OST000d_UUID ACTIVE
14: T3_HS0-OST000e_UUID ACTIVE
15: T3_HS0-OST000f_UUID ACTIVE
16: T3_HS0-OST0010_UUID ACTIVE
17: T3_HS0-OST0011_UUID ACTIVE
18: T3_HS0-OST0012_UUID ACTIVE
19: T3_HS0-OST0013_UUID ACTIVE
20: T3_HS0-OST0014_UUID ACTIVE
21: T3_HS0-OST0015_UUID ACTIVE
22: T3_HS0-OST0016_UUID ACTIVE
23: T3_HS0-OST0017_UUID ACTIVE
24: T3_HS0-OST0018_UUID ACTIVE
25: T3_HS0-OST0019_UUID ACTIVE
26: T3_HS0-OST001a_UUID ACTIVE
27: T3_HS0-OST001b_UUID ACTIVE
28: T3_HS0-OST001c_UUID ACTIVE
29: T3_HS0-OST001d_UUID ACTIVE
30: T3_HS0-OST001e_UUID ACTIVE
31: T3_HS0-OST001f_UUID ACTIVE
32: T3_HS0-OST0020_UUID ACTIVE
33: T3_HS0-OST0021_UUID ACTIVE
34: T3_HS0-OST0022_UUID ACTIVE
35: T3_HS0-OST0023_UUID ACTIVE
```

MDTの状態

OSTの状態

Lustreの使用状況確認 – 容量(block)

```
lfs > df -h
UUID                    bytes      Used   Available Use% Mounted on
T3_HS0-MDT0000_UUID      5.9T     173.4G      5.7T   3% /gs/hs0[MDT:0]
T3_HS0-OST0000_UUID     71.5T      20.2T     50.6T  28% /gs/hs0[OST:0]
T3_HS0-OST0001_UUID     71.5T      22.1T     48.7T  31% /gs/hs0[OST:1]
T3_HS0-OST0002_UUID     71.5T      19.6T     51.2T  28% /gs/hs0[OST:2]
T3_HS0-OST0003_UUID     71.5T      19.4T     51.3T  27% /gs/hs0[OST:3]
T3_HS0-OST0004_UUID     71.5T      19.9T     50.9T  28% /gs/hs0[OST:4]
T3_HS0-OST0005_UUID     71.5T      20.4T     50.3T  29% /gs/hs0[OST:5]
T3_HS0-OST0006_UUID     71.5T      20.4T     50.3T  29% /gs/hs0[OST:6]
T3_HS0-OST0007_UUID     71.5T      20.7T     50.1T  29% /gs/hs0[OST:7]
T3_HS0-OST0008_UUID     71.5T      19.6T     51.2T  28% /gs/hs0[OST:8]
T3_HS0-OST0009_UUID     71.5T      20.1T     50.7T  28% /gs/hs0[OST:9]
T3_HS0-OST000a_UUID     71.5T      20.2T     50.6T  29% /gs/hs0[OST:10]
T3_HS0-OST000b_UUID     71.5T      19.7T     51.1T  28% /gs/hs0[OST:11]
T3_HS0-OST000c_UUID     71.5T      20.4T     50.3T  29% /gs/hs0[OST:12]
  ・
  ・
[省略]
  ・
  ・
T3_HS0-OST0033_UUID     71.5T      20.1T     50.7T  28% /gs/hs0[OST:51]
T3_HS0-OST0034_UUID     71.5T      19.8T     51.0T  28% /gs/hs0[OST:52]
T3_HS0-OST0035_UUID     71.5T      19.9T     50.8T  28% /gs/hs0[OST:53]
T3_HS0-OST0036_UUID     71.5T      20.2T     50.5T  29% /gs/hs0[OST:54]
T3_HS0-OST0037_UUID     71.5T      20.5T     50.3T  29% /gs/hs0[OST:55]
T3_HS0-OST0038_UUID     71.5T      20.7T     50.0T  29% /gs/hs0[OST:56]
T3_HS0-OST0039_UUID     71.5T      20.4T     50.4T  29% /gs/hs0[OST:57]
T3_HS0-OST003a_UUID     71.5T      19.5T     51.3T  28% /gs/hs0[OST:58]
T3_HS0-OST003b_UUID     71.5T      20.9T     49.9T  30% /gs/hs0[OST:59]
T3_HS0-OST003c_UUID     71.5T      20.4T     50.3T  29% /gs/hs0[OST:60]
T3_HS0-OST003d_UUID     71.5T      23.9T     46.9T  34% /gs/hs0[OST:61]
T3_HS0-OST003e_UUID     71.5T      19.0T     51.8T  27% /gs/hs0[OST:62]
T3_HS0-OST003f_UUID     71.5T      20.2T     50.5T  29% /gs/hs0[OST:63]
T3_HS0-OST0040_UUID     71.5T      19.2T     51.5T  27% /gs/hs0[OST:64]
T3_HS0-OST0041_UUID     71.5T      20.4T     50.4T  29% /gs/hs0[OST:65]
T3_HS0-OST0042_UUID     71.5T      21.2T     49.5T  30% /gs/hs0[OST:66]
T3_HS0-OST0043_UUID     71.5T      19.8T     51.0T  28% /gs/hs0[OST:67]


filesystem summary:      4.7P       1.3P       3.4P  29% /gs/hs0
```

Lustreの使用状況確認 – inode

```
lfs > df -i -h
UUID                    Inodes      IUsed      IFree IUse% Mounted on
T3_HS0-MDT0000_UUID       4.0G     747.6M       3.3G  18% /gs/hs0[MDT:0]
T3_HS0-OST0000_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:0]
T3_HS0-OST0001_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:1]
T3_HS0-OST0002_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:2]
T3_HS0-OST0003_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:3]
T3_HS0-OST0004_UUID     577.5M      10.5M     567.0M   2% /gs/hs0[OST:4]
T3_HS0-OST0005_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:5]
T3_HS0-OST0006_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:6]
T3_HS0-OST0007_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:7]
T3_HS0-OST0008_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:8]
T3_HS0-OST0009_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:9]
T3_HS0-OST000a_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:10]
T3_HS0-OST000b_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:11]
T3_HS0-OST000c_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:12]
  :
  :
[省略]
  :
  :
T3_HS0-OST0033_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:51]
T3_HS0-OST0034_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:52]
T3_HS0-OST0035_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:53]
T3_HS0-OST0036_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:54]
T3_HS0-OST0037_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:55]
T3_HS0-OST0038_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:56]
T3_HS0-OST0039_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:57]
T3_HS0-OST003a_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:58]
T3_HS0-OST003b_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:59]
T3_HS0-OST003c_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:60]
T3_HS0-OST003d_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:61]
T3_HS0-OST003e_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:62]
T3_HS0-OST003f_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:63]
T3_HS0-OST0040_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:64]
T3_HS0-OST0041_UUID     577.5M      10.6M     566.9M   2% /gs/hs0[OST:65]
T3_HS0-OST0042_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:66]
T3_HS0-OST0043_UUID     577.5M      10.7M     566.8M   2% /gs/hs0[OST:67]

filesystem summary:       4.0G     747.6M       3.3G  18% /gs/hs0
```

 © 2019 DDN Storage

Lustre Quota

- Lustre 2.4.1以降、block/inodeに対してそれぞれUser/Group Quotaを設定可能

- Lustre 2.7.21以降、Project Quotaをサポート
  - 同一Project IDを持つディレクトリ/ファイルに対するQuota
  - 複数のディレクトリを纏めてQuota設定可能

- TSUBAME3高速ストレージ領域は、Project Quotaのみ利用
  - Quotaの設定、変更、解除は管理者のみ可能
  - UserおよびGroup Quotaも有効であるが、使用していない
  - Project IDの付与、確認などの管理は管理者のみ可能

 © 2019 DDN Storage

Lustre Quota設定(1)

Quota	対象	単位	説明
User block soft limit	User	Kbytes	対象ファイルシステムへ、あるユーザが書き込めるサイズの制限値 この値を超えている場合でも、grace timeの間は書き込み可能
User block soft limit grace time	User	時間	User block soft limitを超えていられる時間
User inode soft limit	User	Inode数	対象ファイルシステムへ、あるユーザが書き込めるファイル数の制限値 この値を超えている場合でも、grace timeの間は書き込み可能
User inode soft limit grace time	User	時間	User inode soft limitを超えていられる時間
User block hard limit	User	Kbytes	対象ファイルシステムへ、あるユーザが書き込めるサイズの制限値 この値を超えて書き込みを行うことは不可能
User inode hard limit	User	Inode数	対象ファイルシステムへ、あるユーザが書き込めるファイル数の制限値 この値を超えて書き込みを行うことは不可能
Group block soft limit	Group	Kbytes	対象ファイルシステムへ、あるグループが書き込めるサイズの制限値 この値を超えている場合でも、grace timeの間は書き込み可能
Group block soft limit grace time	Group	時間	Group block soft limitを超えていられる時間
Group inode soft limit	Group	Inode数	対象ファイルシステムへ、あるグループが書き込めるファイル数の制限値 この値を超えている場合でも、grace timeの間は書き込み可能
Group inode soft limit grace time	Group	時間	Group inode soft limitを超えていられる時間
Group block hard limit	Group	Kbytes	対象ファイルシステムへ、あるグループが書き込めるサイズの制限値 この値を超えて書き込みを行うことは不可能
Group inode hard limit	Group	Inode数	対象ファイルシステムへ、あるグループが書き込めるファイル数の制限値 この値を超えて書き込みを行うことは不可能

© 2019 DDN Storage

Lustre Quota設定(2)

Quota	対象	単位	説明
Project block soft limit	Project	Kbytes	対象ファイルシステムへ、あるプロジェクトが書き込めるサイズの制限値 この値を超えている場合でも、grace timeの間は書き込み可能
Project block soft limit grace time	Project	時間	Project block soft limitを超えていられる時間
Project inode soft limit	Project	Inode数	対象ファイルシステムへ、あるプロジェクトが書き込めるファイル数の制限値 この値を超えている場合でも、grace timeの間は書き込み可能
Project inode soft limit grace time	Project	時間	Project inode soft limitを超えていられる時間
Project block hard limit	Project	Kbytes	対象ファイルシステムへ、あるプロジェクトが書き込めるサイズの制限値 この値を超えて書き込みを行うことは不可能
Project inode hard limit	Project	Inode数	対象ファイルシステムへ、あるプロジェクトが書き込めるファイル数の制限値 この値を超えて過去込みを行うことは不可能

Quotaの確認方法

- TSUBAME3ではt3-user-infoコマンドによってHomeおよび高速ストレージ領域の利用状況を確認ができますが、ここではlfsコマンドによる確認方法を示します

- lfs quota [-u <user_name | uid> | -g <group_name | gid> | -p project_id] mount_point

- lfs quotaその他のオプション
  - -t [-u | -g | -p] mount_point
    - User, Group, Projectそれぞれに対するblock およびinodeのgrace timeを表示
  - -v
    - MDT, OSTごとの使用状況を確認可能
  - -h
    - human readable format

 © 2019 DDN Storage

Quotaの確認例

```
lfs > quota -u ddn_user00 /gs/hs0
Disk quotas for usr ddn_user00 (uid 2181):
     Filesystem  kbytes   quota   limit   grace   files   quota   limit   grace
        /gs/hs0       0       0       0       -       0       0       0       -
```

見出し	意味
Filesystem	ファイルシステム名(mount_point)
kbytes(-hを使用した場合はused)	現在の使用量、超過している場合数字の後ろに * が付与される
quota	block soft quota
limit	block hard quota
grace	block grace time
files	現在のファイル数、超過している場合数字の後ろに * が付与される
quota	inode soft quota
limit	inode hard quota
grace	inode grace time

 © 2019 DDN Storage

File Stripingの利用

- Lustreでは、File Stripingをサポートしています
  - 1つのファイルを複数OSTに書き込む機能

- 各ファイルを格納するStripe Count(使用するOSTの数)およびStripe Sizeをディレクトリやファイルごとに設定可能
  - TSUBAME3のデフォルトはStripe設定無し
  - 各ファイルは1つのOSTへ格納
  - ディレクトリにStripeを設定した場合、以降そのディレクトリ以下に作られるファイルは全てStripeされる

 © 2019 DDN Storage

Stripeの設定方法

```
lfs setstripe [--stripe-size | -S <stripe_size>] [--stripe-count | -c <stripe_count] [--stripe-index | -I <start_ost_idx>] [--pool | -p <pool_name>] [--block | -b] <directory | filename> [--ost-list|-o <ost_indices>]
```

Option項目	意味
stripe_size	各OSTをStripeする際のStripe Sizeを指定 k, m, gでそれぞれKbytes, Mbytes, GBytes指定が可能 デフォルトは0でファイルシステムデフォルト(TSUBAME3では1MiB)を意味する
start_ost_idx	Stripeを開始するOSTを指定 デフォルトは-1でランダム
stripe_count	使用するOSTの数を指定 -1で全OST デフォルトは0でファイルシステムデフォルト(TSUBAME3では1)を意味する
pool_name	使用するプール名を指定 ※ TSUBAME3では OST Poolを使用していません
block	data migration時にファイルへのアクセスをblockする
ost_indices	使用するOSTのインデックス番号を指定

 © 2019 DDN Storage

Stripeの確認方法

```
lfs getstripe [--ost | -O <uuid>] [--quiet | -q] [--verbose | -v] [--recursive | -r] <directory | filename>
```

Option項目	意味
--ost|-O <uuid>	uuidで指定したOSTに属するファイルを表示
--quiet|-q	出力項目の制限
--verbose|-v	Verbose Mode
--recursive|-r	Recursive Mode

 © 2019 DDN Storage

Stripe使用例(1)

- 作成したディレクトリに対し、2x OST, Stripe Size=1MBytesでStripeを設定

```
ddn_user00@login0:~> mkdir /gs/hs2/ddn_user00/s2
ddn_user00@login0:~> lfs setstripe -c 2 -S 1m /gs/hs2/ddn_user00/s2
ddn_user00@login0:~> lfs getstripe /gs/hs2/ddn_user00/s2
/gs/hs2/ddn_user00/s2
stripe_count:   2 stripe_size:    1048576 stripe_offset:  -1
```

- 作成したディレクトリに対し、全OST, Stripe Size=4MBytesでStripeを設定

```
ddn_user00@login0:~> mkdir /gs/hs2/ddn_user00/st_all
ddn_user00@login0:~> lfs setstripe -c -1 -S 4m /gs/hs2/ddn_user00/st_all
ddn_user00@login0:~> lfs getstripe /gs/hs2/ddn_user00/st_all
/gs/hs2/ddn_user00/st_all
stripe_count:   -1 stripe_size:    4194304 stripe_offset:  -1
```

 © 2019 DDN Storage

Stripe使用例(2)

- 作成したs2ディレクトリ以下にファイルを作成

```
ddn_user00@login0:/gs/hs2/ddn_user00/s2> pwd
/gs/hs2/ddn_user00/s2
ddn_user00@login0:/gs/hs2/ddn_user00/s2> dd if=/dev/zero of=testfile1 bs=1M count=1000
1000+0 records in
1000+0 records out
1048576000 bytes (1.0 GB, 1000 MiB) copied, 0.721639 s, 1.5 GB/s
ddn_user00@login0:/gs/hs2/ddn_user00/s2> lfs getstripe testfile1
testfile1
lmm_stripe_count:  2
lmm_stripe_size:   1048576
lmm_pattern:       1
lmm_layout_gen:    0
lmm_stripe_offset:  33
obdidx objid objid group
    33      13340797     0xcb907d          0
    50      13344430     0xcb9eae          0
```

- testfile1はobdidxで示される2つのOSTにStripeされており、obdidxはlfs ostsで確認可能です

 © 2019 DDN Storage

File Stripe使用上の注意

- 必ずしも、大きいStripe Count == 高性能ではありません
  - ファイルシステム全体のスループットはStripe無しが最も高い
- Stripe Sizeはデフォルトの1MiBが最適です
  - ストライプする場合
- 充分大きなファイルはStripeした方が良い
  - 1ファイルが1OSTに書ける最大サイズは16TiB
  - 各OSTの使用量がアンバランスになると性能が劣化する可能性あり
- シングルシェアードファイル（後述）を使用する場合はStripe必須
- シングルスレッドアプリケーションでは、Stripeの性能に対する効果はほぼ無い
- OSTは手動で指定せずランダムが最適

 © 2019 DDN Storage

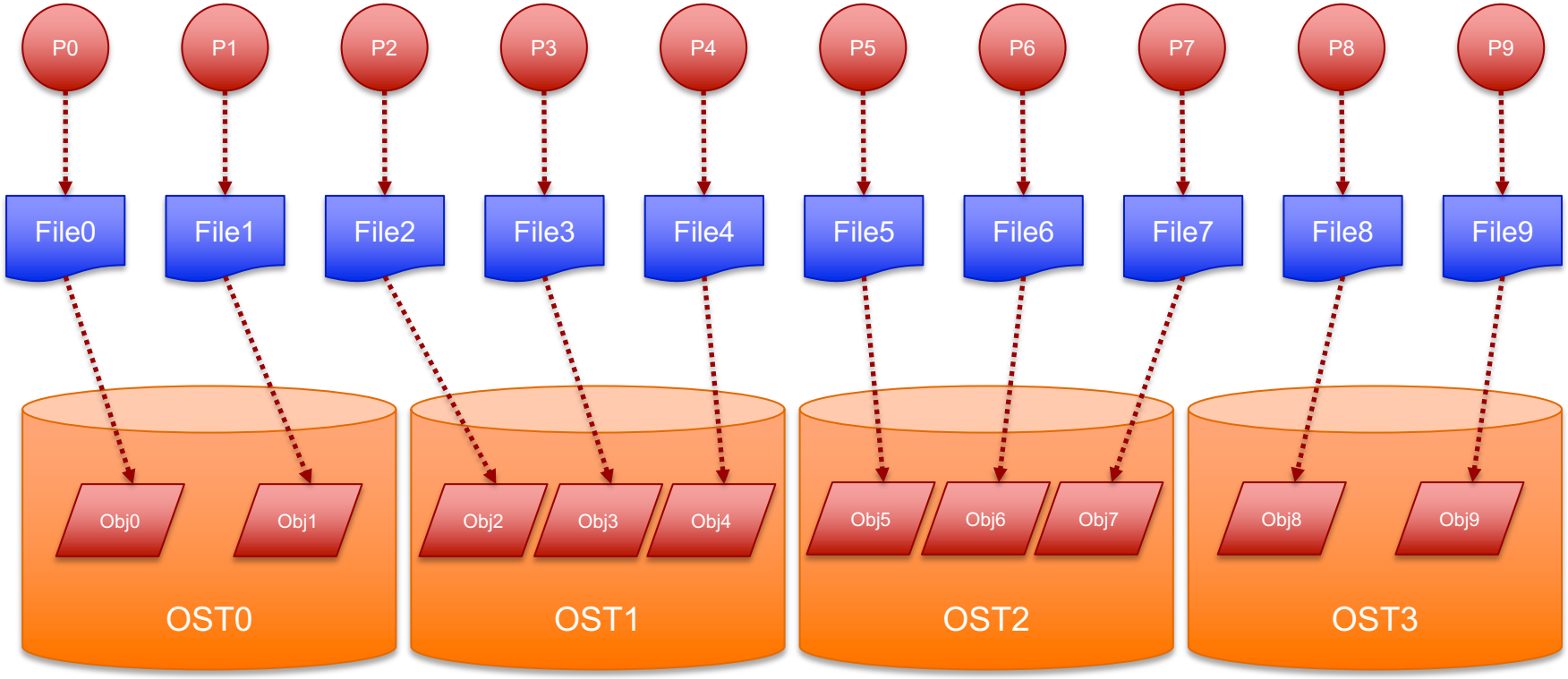
FPP(File Per Process)とSSF(Single Shared FIle)

- FPP : プロセスごとに異なるファイルを作成
  - 高速なファイル入出力
  - ファイル数が膨大になった場合の処理が複雑、且つメタデータ性能が劣化する

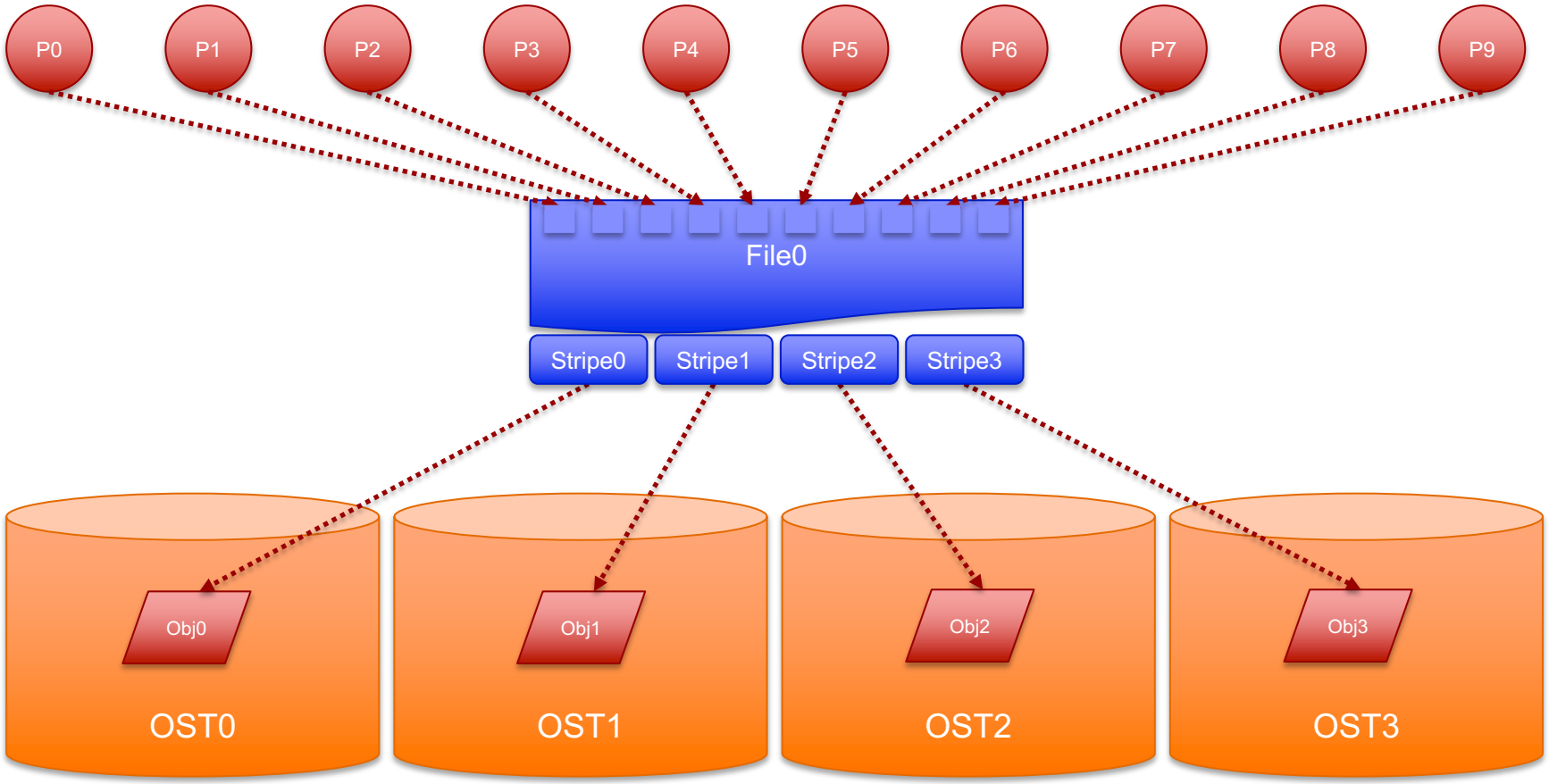
- SSF : 全プロセスで単一ファイルにアクセス
  - ファイルの処理が容易
  - FPPに比べて低速
  - 性能向上の為にはStripe必須
    - $ lfs setstrip –c -1 ファイル名
    - ストライプカウント(-c) -1 : 全OSTに跨がってStripe
    - ストライプサイズのデフォルトは1MiB (デフォルトのままで良い)
  - MPIIO, NetCDF, HDF5などを利用

 © 2019 DDN Storage

FPP(File Per Process)概念図



SSF(Single Shared File)概念図



Lustre Tips

- 苦手なこと
  - 大量の小さいファイル
  - 小さいIO Size
  - ランダムアクセス

- メタデータ性能
  - 複数プロセスによる同一ディレクトリへの操作は遅くなる
  - プロセス毎に異なるディレクトリを操作する方が良い

- スループット性能
  - プロセス毎に異なるファイルを操作
  - できる限り各プロセスにIOを行わせる
  - 1 Clientのスループット性能は最高でも1リンクあたり10GB/s
    - 単一プロセスでは2-3GB/sでれば良い方、Xeon Phiならもっと出ない
    - 1クライアント/複数IOプロセス + 複数ノード = スループット向上

 © 2019 DDN Storage

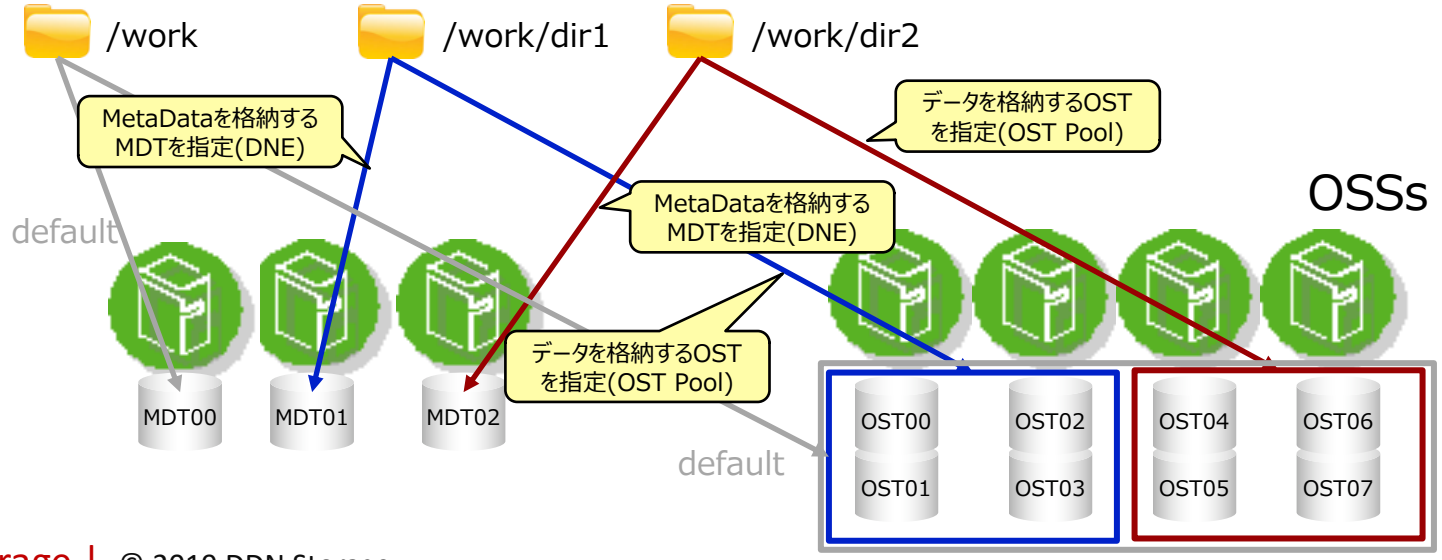
DNEとOST Poolのご紹介(1)

- TSUBAME3では利用されていませんが、Lustreが提供しているDNEとOST Poolを紹介します
  - 次回調達のご参考までに、、、

- DNE(Distributed Namespace Environment)
  - Directoryごとに使用するMDS/MDTを指定可能
  - メタデータオペレーションのスケーリング
  - 他ジョブによる性能への影響を回避

- OST Pool
  - Directoryごとに使用するOSS/OSTを指定可能
  - Disk性能が異なる際に有用
  - 他ジョブによる性能への影響を回避

One Namespaceで
複数FSのような利用が可能



DNEとOST Poolのご紹介(2)

- DNEのメリット・デメリット

メリット	デメリット
複数台のMDS/MDTを使用することによる全体でのメタデータオペレーション性能の向上	2つのRAIDを1つにまとめて使用する場合に比べて、DNEとして構成するMDTを意図的に使用する必要がある
複数台のMDS/MDTを使用することでジョブ間での影響を抑える	DNE間を跨る場合、Directory間におけるfile/directoryのmv時にはdataのcopy処理も実施される
	MDT用RAIDの数に依存するが、後でやり直すことがほぼ不可能。但し、動的な追加は可能

⇒ 構築前に構成FIXが必要

- Stripeのメリット・デメリット

メリット	デメリット
1 FileへのIO性能向上	FS全体でのスループットではStripe無しが高い
1 Fileあたりの最大サイズ向上	大きなファイルを作成した場合、特定OSTの使用量が大幅に増加し、そのOSTが使われにくくなる

⇒ 使いながら調整可能であり、アプリやデータごとに変更することがBEST

© 2019 DDN Storage

DNEとOST Poolのご紹介(3)

- 例えばTSUBAME3であれば、、、
  - 高速ストレージ領域の1FSを複数MDTで構成し、専用MDSを持つ領域を作る
  - ただし、MDTの追加が必要です
  - 明示的に3FSに分かれているので、利用ユーザを分散することによってDNEの代わりになります

- OST Pool
  - Sub Directory Mountと組み合わせた専有OST領域
    - セキュリティを考慮したコンテナからの利用など
  - ただし、スループットは限定される

© 2019 DDN Storage

おまけ
高速データ転送ツール

- 並列ファイルシステムと高速にデータ転送を行うための機能として、DDNがコントリビュートしているオープンソースツールである、mpiFileUtilがあります
  - https://github.com/hpc/mpifileutils

- mpiFileUtilに含まれるツール群はMPIを利用して並列にファイルのコピー、削除、比較などを行えるツールです
  - 本ツールを計算ノード上で利用し、MPIによる並列ファイルコピーによる高速なデータ転送を実現可能です

- 本ツールはLustre(DDN ExaScaler)の様な、並列ファイルシステムでの利用を想定し開発されたツールです
  - rsync, scpなどの転送ツールより高速な転送を行うことが可能です

- **BeeONDを使用する共有スクラッチ領域と高速ストレージ領域間のデータ転送に使用すると、利用手引きにあるcpコマンドを使用するより高速にデータのStage-In/Outが可能です**
  - 使用して良いかはGSIC殿にご相談ください

 © 2019 DDN Storage

Agenda

- DDN会社/製品紹介
- Lustreについて
- Lustreの特徴
- Lustreのアーキテクチャ
- Lustreの各コンポーネント
- TSUBAME3ストレージ概要
- TSUBAME3ファイルシステムの利用
- **Q&A**

 © 2019 DDN Storage

Thank You!

Keep in touch with us

Team-JPSales@ddn.com

Tokyu Bancho Bldg. 8F
6-2 Yonbancho Chiyoda-ku,
Tokyo 102-0081

@ddn_limitless

+81-3-3261-9101
+81-3-3261-9140

company/datadirect-networks

 © 2019 DDN Storage