DDNのHPCストレージへの貢献と新しい課題への取り組み

DDN Storage
井原 修一

アジェンダ

1. DDNについて
2. HPCストレージのトレンド
3. ExascalserにおけるHPCストレージの課題に対する取り組み
4. 次世代のストレージシステム”DDN RED”について

DDNは世界最大手の非上場ストレージベンダー

世界規模のマーケットリーダー

- 20年以上に渡る業界リーダーシップ
- 10,000以上の顧客
- 世界10拠点にテクノロジーセンターを設置

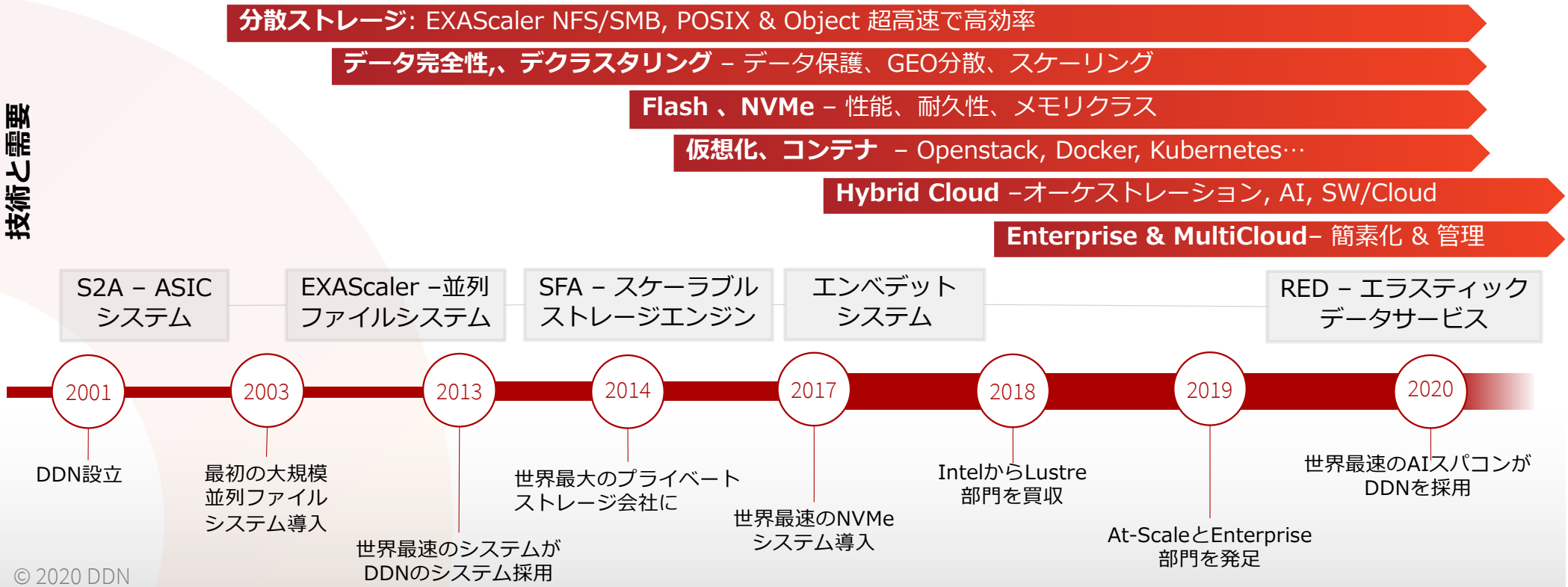
At Scale	Enterprise
AIストレージシステム	仮想環境
大規模HPCストレージ	ソフトウェアデファインド統合ストレージ
スケーラブルデータ管理	高性能統合ストレージ

株式データダイレクトネットワークスジャパン

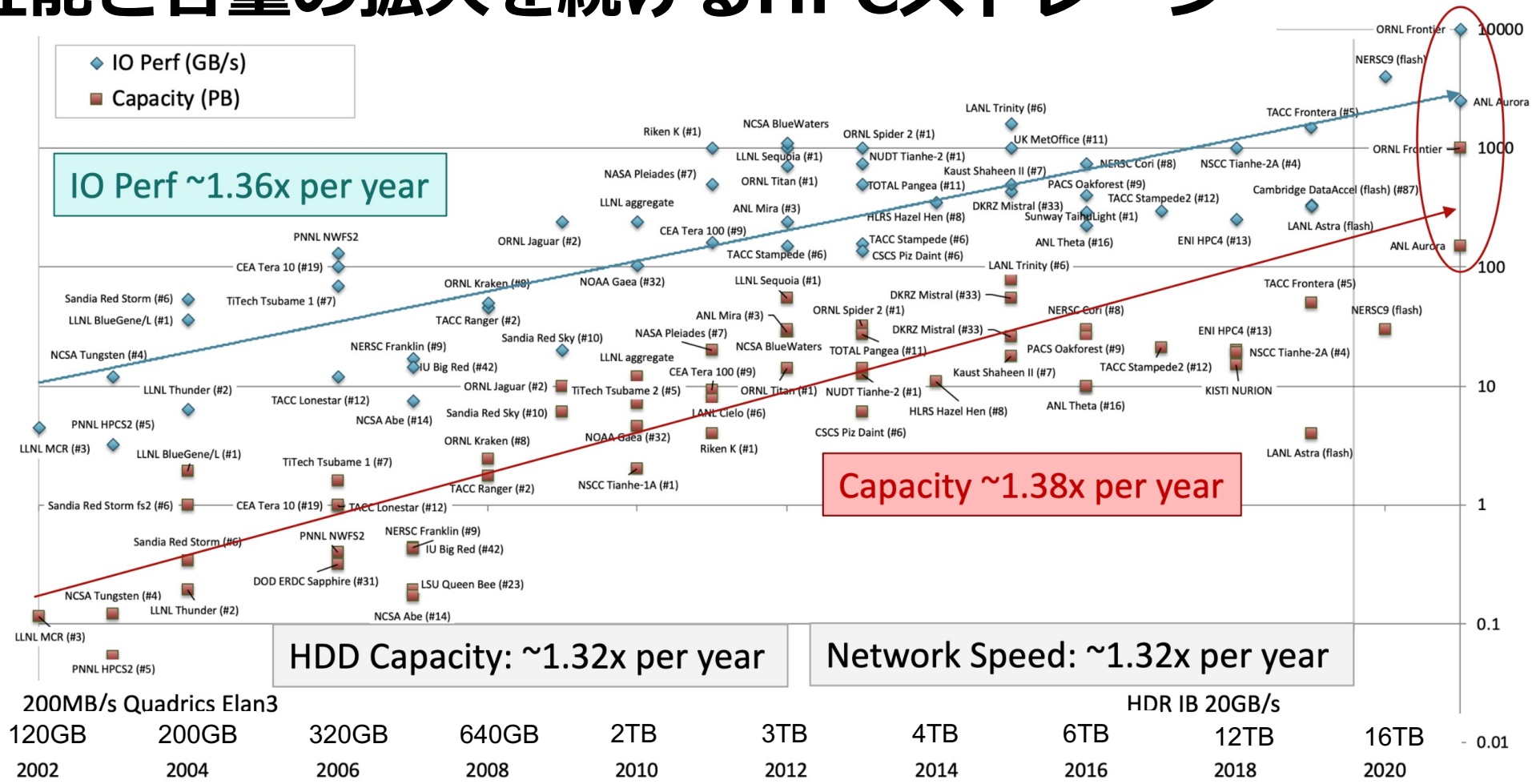
(米国DDN社100%子会社)、2008年１月設立

- ストレージのフルスタックソリューションを提供する営業、エンジニアリング集団
- 高性能ストレージシステムの販売、設計、構築、保守
- 販売パートナー: 富士通、日本電気、日立製作所、SCSK、日本コムシスなど

DDNにおける20年間の技術革新



性能と容量の拡大を続けるHPCストレージ



Source: *Rock Hard Lustre*, Nathan Rutman, Cray (with updates for recent years); *Disk Drive Prices (1955-2019)*, John C. McCallum

参照：「Lustre:The Next 20 Years(次の20年)」、ISC19、Andreas Dilger

DDN AI400Xが支える NVIDIA SuperPODとDGX A100

DDNストレージがスケーラブルなソリューションであることが最大のSuperPODとDGX A100のプロダクション環境で証明

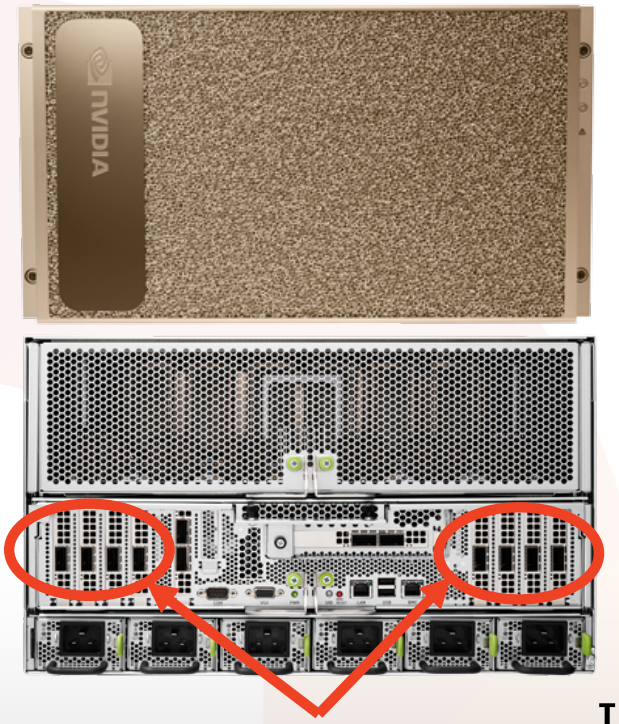
Top500のNo7にランキングされるSelene

- 280 NVIDIA DGX A100 システム
- 170 Mellanox 200G HDR スイッチ
- 40 x DDN AI400X アプライアンス
  - 最大2TB/secのスループットと1.2億 IOPS
  - 最適かつ完全にDGX A100向けにインテグレーション
  - 最初の10システムを4時間で導入し、その後シームレースな拡張を提供

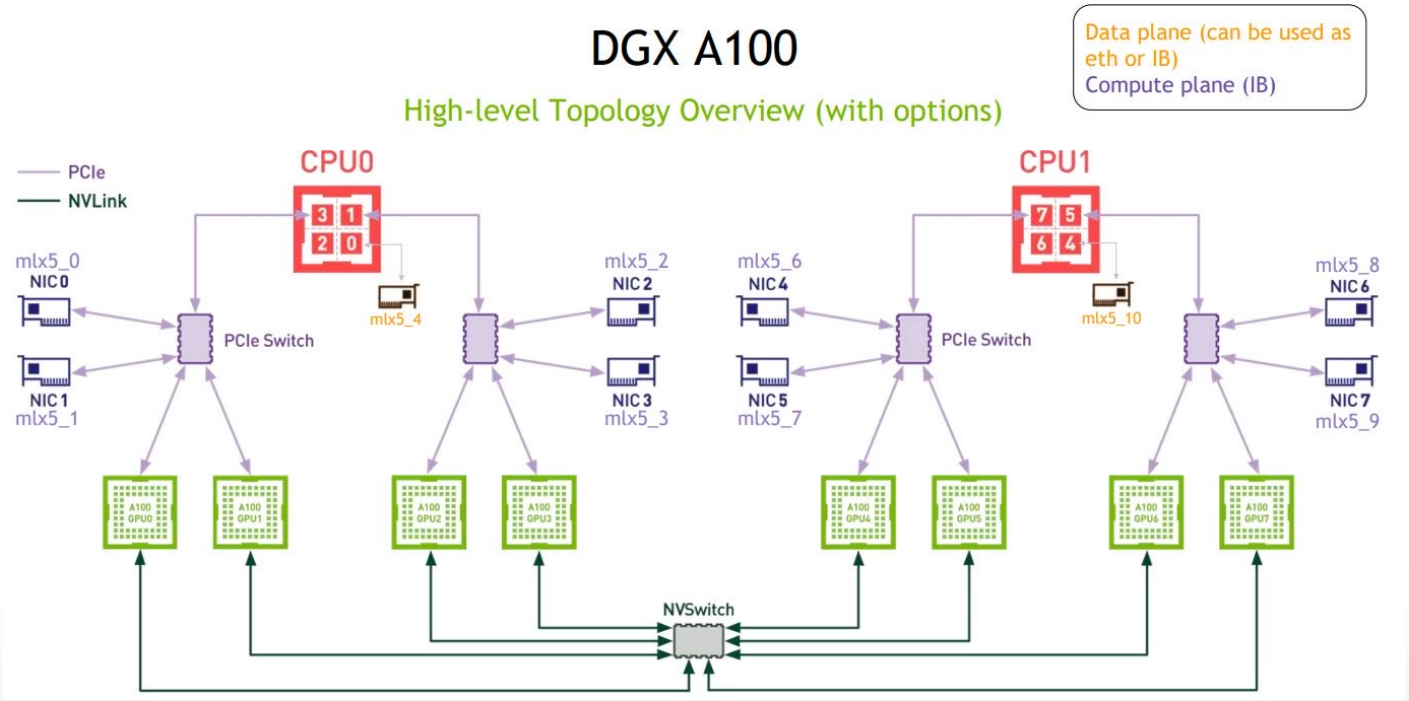
NVIDIA社によってテスト、検証されたAI400Xは 非常に簡素化されDGX A100, DGX PODまたはDGX SuperPODのいずれにおいても利用可能

ddn

Nから1へダウンスケール、性能はN倍



8 x IB-HDR200+
2 x IB-HDR200



I/Oネットワーク帯域は一般的なクライアントの20倍以上*

*IB-EDRを比較した場合

自動運転の研究開発基盤にDDNを採用

DDN EXAScalerを用いてハイパースケールR&Dを実用化(米国,2016)

- 全米規模のハイパースケール自動運転ソリューション
- 全体ストレージ容量は350PB以上、性能は1.6TB/s以上
- 各リージョンごとに非常に大きな容量を持つ単一ネームスペースに対する性能を、コスト効率の高いアプローチで実現
- 完全な大規模ストレージ管理環境を提供

大手自動車メーカー向けにモジュール型でスケーラブルなAIストレージ(欧州, 2019-2020)

- １つのモジュールは27PB、200GB/sec
- 複数サイトの複数モジュールにて構成
- システムの性能、状態だけでなく、ユーザとワークロードのIOアクティビティを詳細に把握可能

指数関数的に増加する大規模データ処理の需要

AI & Analytics

自立型IT のマーケット規模は2025年に$25B (USD)

75%のAIシステムを採用しているところは今後もAI環境のための予算が増加

Web/Cloud

2025年には世界の約半分のデータがクラウドに

2025年にはインターネットユーザデータは300%の増加

Government/Academia

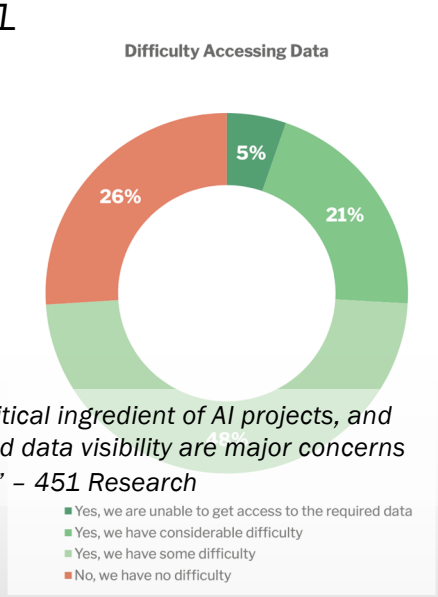
$150B (USD)規模がCovid19関連の研究のために利用

データインテンシブのHPCでは 7% 平均成長率

データ解析17%の成長率

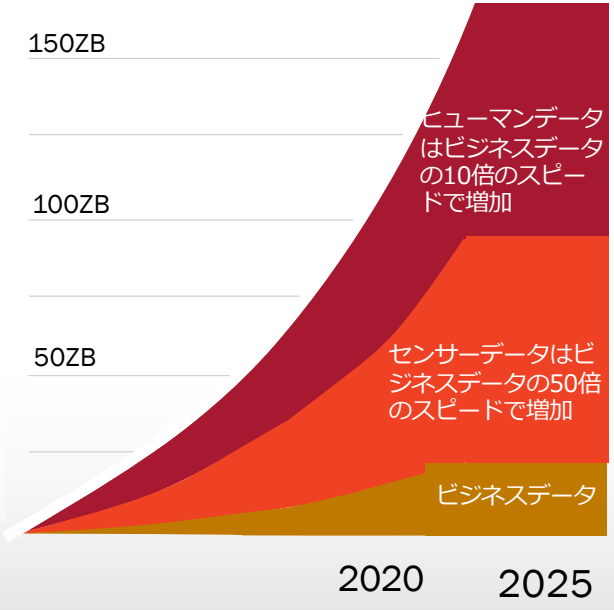
Enterprise at Scale

データ戦略における迅速な採用。ただ、74%がデータアクセスの困難さに直面。



"Data is the critical ingredient of AI projects, and data access and data visibility are major concerns for AI adopters" – 451 Research

爆発的に増加するヒューマンデータとセンサーデータ



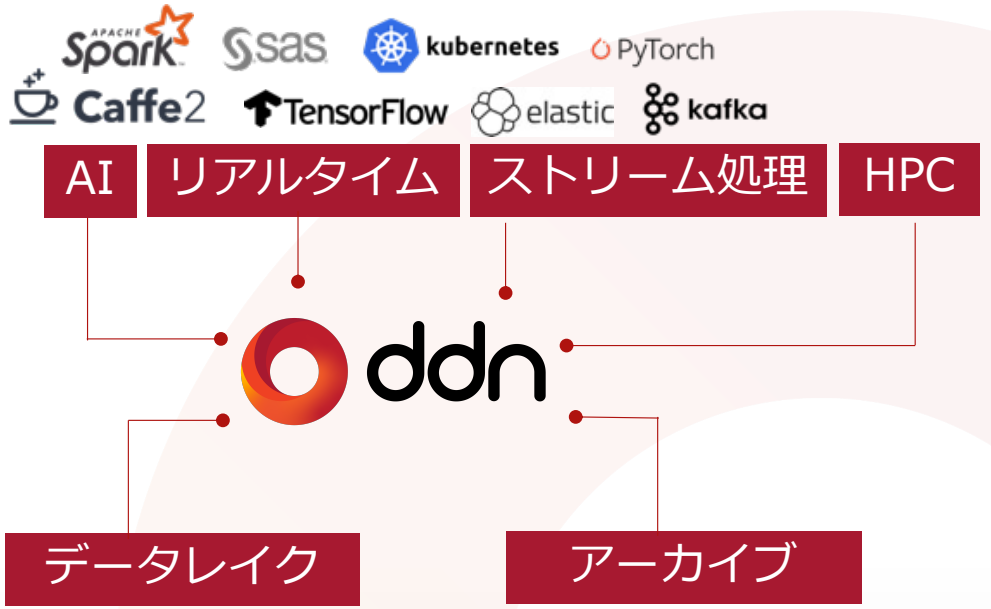
*451 Research – AI & Machine Learning Infrastructure 2020 *inside Big Data

SSD Expands, but HDD Does Not Die - *Gartner*

- DDNは大規模HPCに多くのNVMeのストレージを提供
- ガートナーの予測では:
  - 2030年、エンタープライズにおけるPBの70%はHDD
    - 技術とサプライヤーの経済活動
    - HDDとSSDのギャップは供給とコスト/GB
  - これまでは~25%で推移していた 2020年のNANDベースSSDの生産コストの削減は10-15%
  - 2030年までにすべてのエンタープライズストレージがNANDベースになるのは懐疑的
- 今後もHDDとNVMe/SSDとのハイブリッドは非常に重要

全てのデータを１つの場所へ

- 従来の”ビックデータ” HDFSやデータレイクは限界
  - HDFSはバッチ処理、データレイクはコールドデータを想定
  - これまで全体の1%が、今後は30%以上のデータ解析が必要
  - データと計算リソースの分離は加速。データ領域は共有され、多くの他種多様の計算解析エンジンで処理。
  - データ保護、コンプライアンスは計算リソースと独立
  - 計算リソースの導入、拡張サイクルはストレージと異なる
- 新しいパラダイムは、単にデータをインジェストして共有するだけではなくどのようにデータから価値を生み出すか。リアルタイム性も要求 (Kafka, Spark等)
- ファイルやオブジェクトに対する高いIOスループットとIOPS
- スケールアウトデザイン、多次元的な性能、大規模な並列アーキテクチャ



DDN EXAScaler

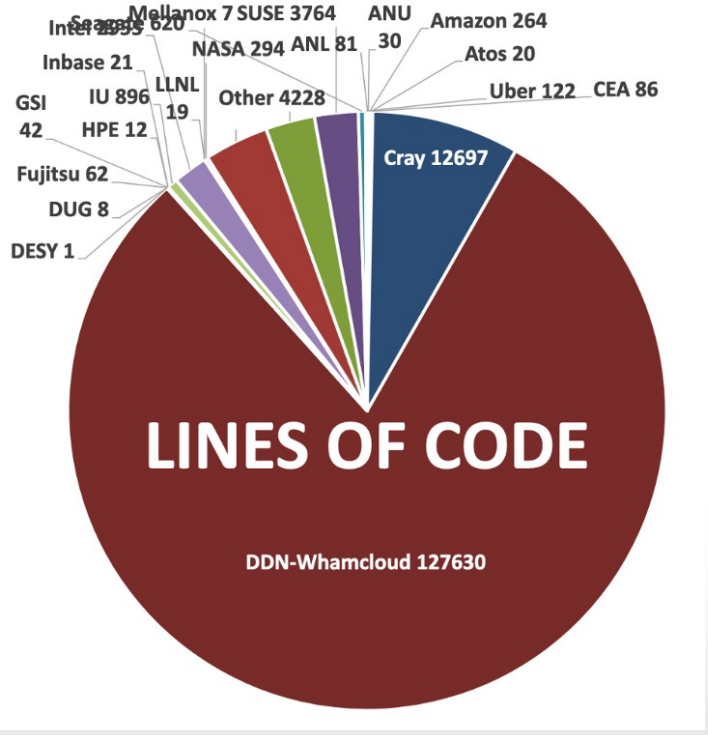
HPC/HPDAストレージへの課題への取り組み

ddn

EXA5

HPC/AIの大規模なストレージ環境においても
データ管理、共有が要求される時代

- 高速なデータ解析のためのデータアクセラレータ
- データ管理とデータ相互交換
- 様々なプロトコルでのデータサービス
- クラウドシステムとの連携
- マルチテナントとセキュリティ
- データアクセスにおける性能の見える化

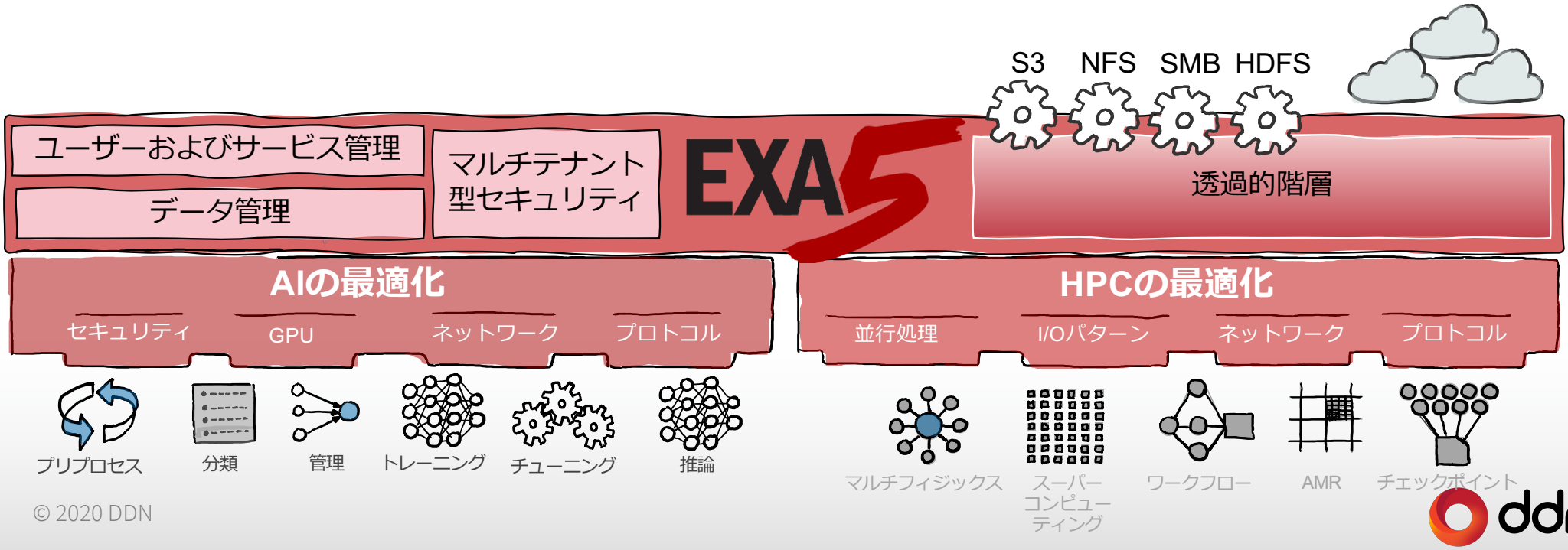


Lustre-2.13への貢献

EXA5 (ExaScaler5)

AIおよびHPC向けに最適化されたストレージソフトウェア環境

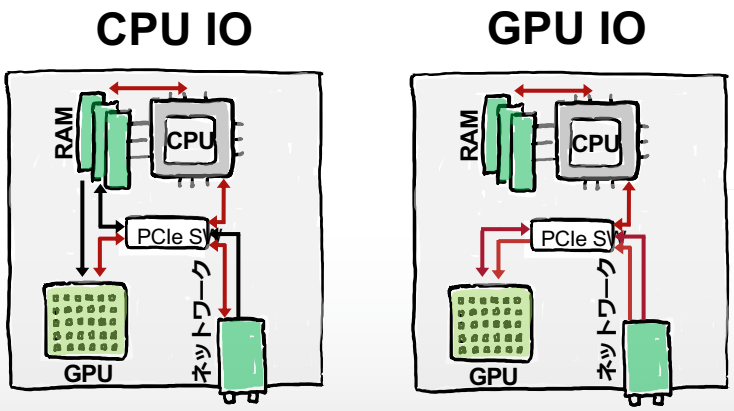
- AIとHPCの双方に対する詳細な最適化により最高の効率性と適正な機能を実現
- データを必要な場所に必要なタイミングで提供

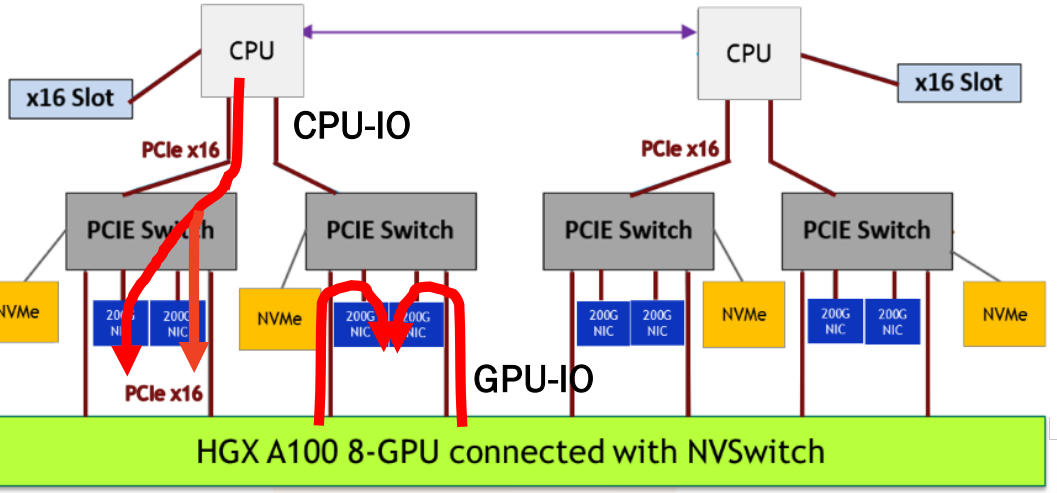


GPU DIRECT STORAGE

GPUから直接ストレージアクセスを可能に

- 不必要なメモリコピーを排除し低レイテンシーを実現
- ホストCPUやメモリサブシステムの消費量が減少
- ストレージとGPU間のBWデータ転送が高速化
- AI、深層学習、HPCアプリケーションの性能向上





Source: Introducing NVIDIA HGX A100: The Most Powerful Accelerated Server Platform for AI and High Performance Computing https://developer.nvidia.com/blog/introducing-hgx-a100-most-powerful-accelerated-server-platform-for-ai-hpc

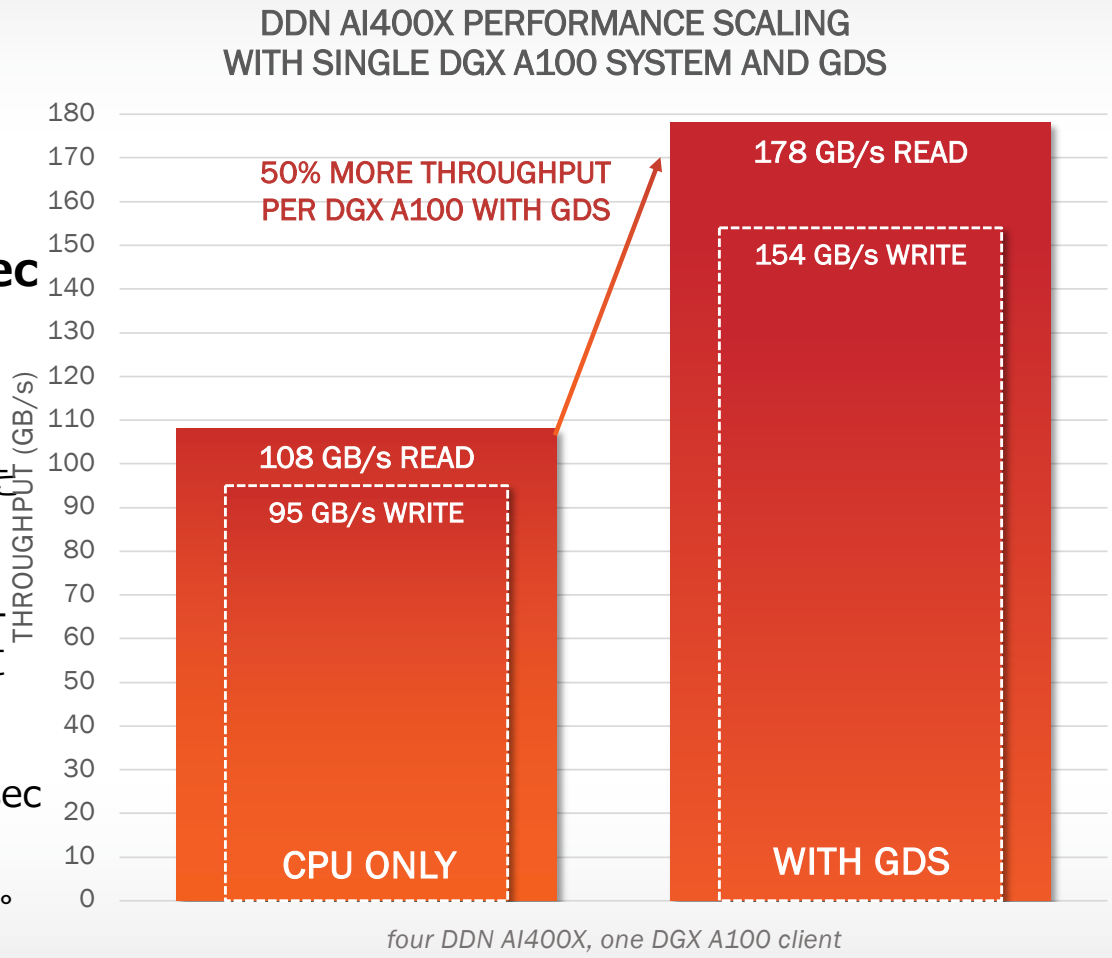
GPU ダイレクトストレージ

ハードウェアの制限をバイパスし178GB/sec性能を達成

DDNはNVIDIA GPUDirect Storageをシステムに統合した最初のストレージベンダー。

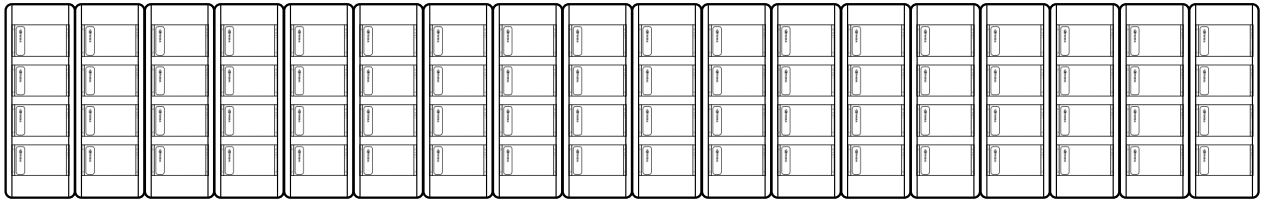
GPUメモリへの直接データ転送を可能とし、システムアーキテクチャにある制限をバイパス、レイテンシを最小限に抑えることによりGPUがもつ最大IO性能を達成。

GPU Direct Storageにおいて、DDN AI400Xは178GB/sec Readと154GB/sec Write性能を達成。
同システムにおけるCPUだけのIOと比較して1.7倍の性能。



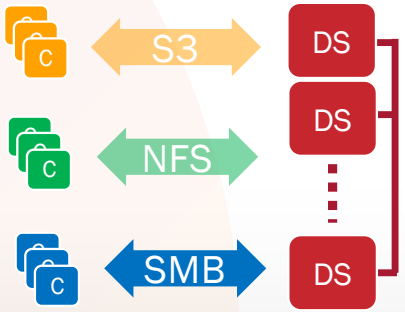
four DDN AI400X, one DGX A100 client

Exascaler データサービスとデータ相互交換



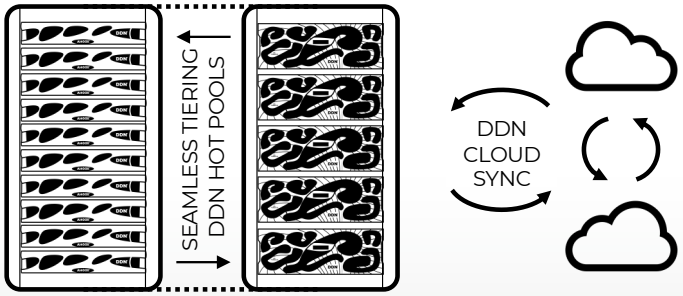
データサービス

様々なプロトコルで
同じ名前空間を共有



高速データアクセス

高速並列アクセス
高帯域、低レイテンシネットワーク
Infiniband, Ethernet(RoCE)



クラウド連携

AWS、Azure, GogoleCloud
上でのExaScaler
オンプレミスなデータとの
データコピー

ALL-NVME

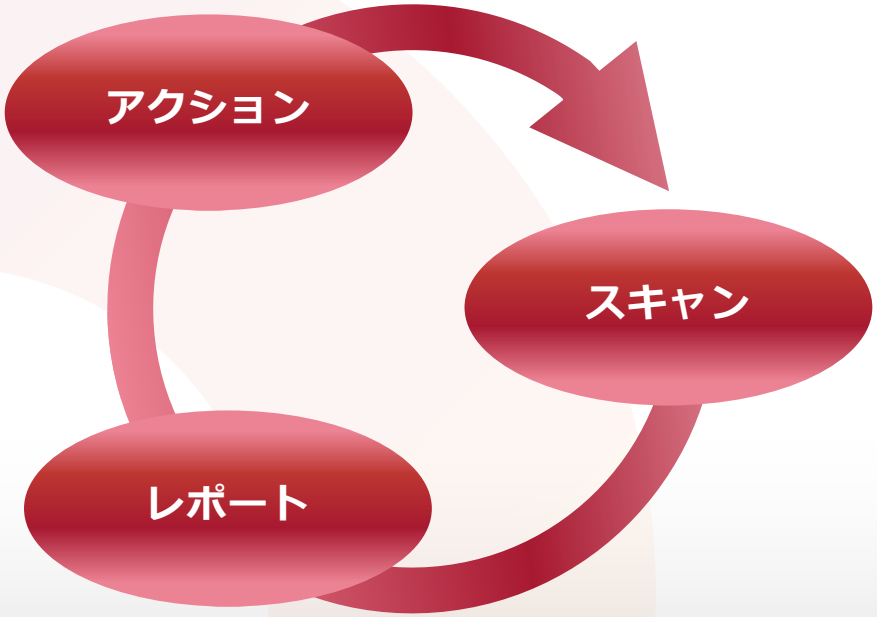
低レイテンシ高いIOPS性能

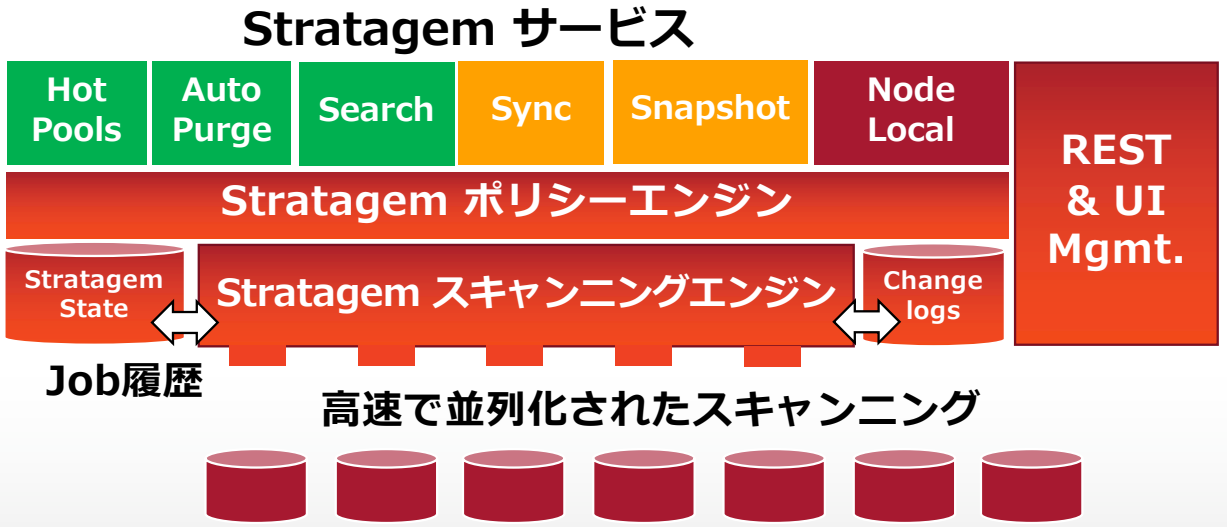
NVME/HDD混在

大容量、高いIO帯域性能

Exascaler データ管理フレームワーク - Stratagem

ファイルシステムを高速にスキャンし、レポートし
ポリシーに基づきアクションを実施

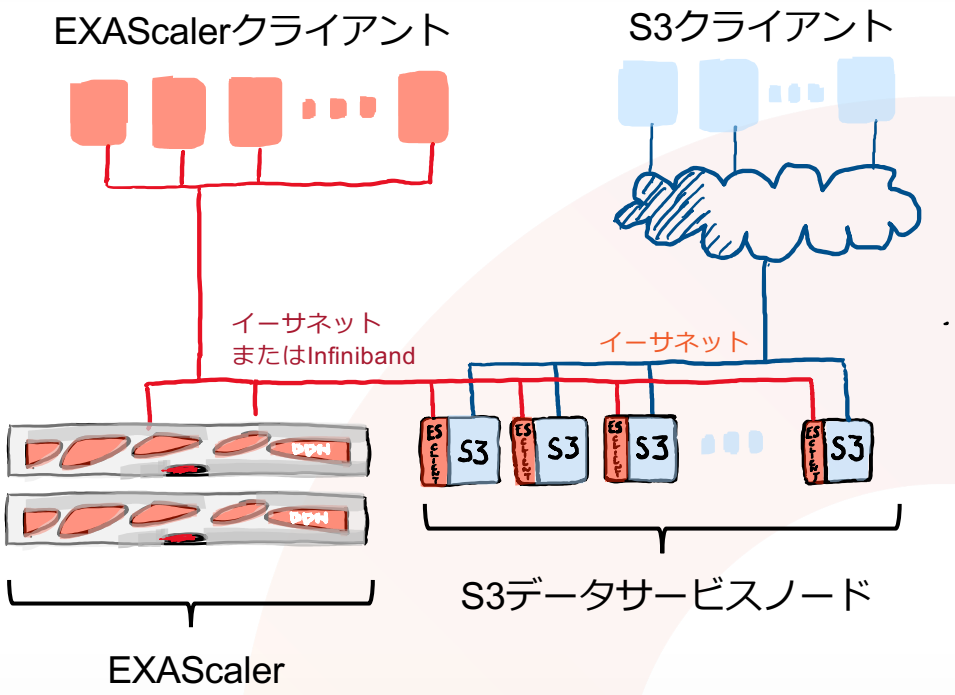




EXAScaler S3データサービス

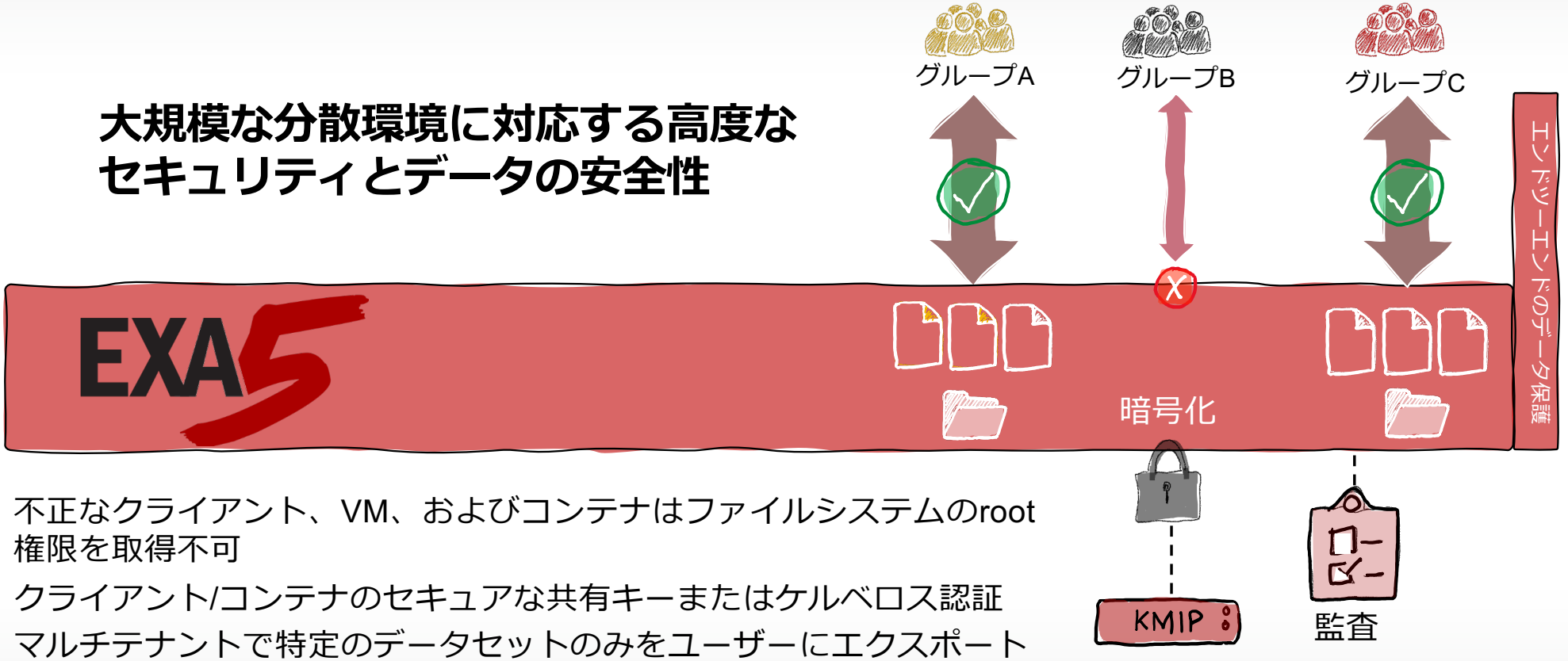
統合されたスケールアウト型S3データサービス

- ExaScaler ストレージ上にてS3 APIを提供
- S3とPOSIXのデータアクセスの統合
- S3経由でデータの書き込み、S3もしくはExaScalerからデータの読み込み
  - 格納される:
    “/s3_content/<バケット>/<ファイル名>…"
- ExaScalerのデータをS3にて読み出し
- S3経由にてデータをインジェストし、追加のデータコピーなしに深層学習の入力データとして直接利用可能



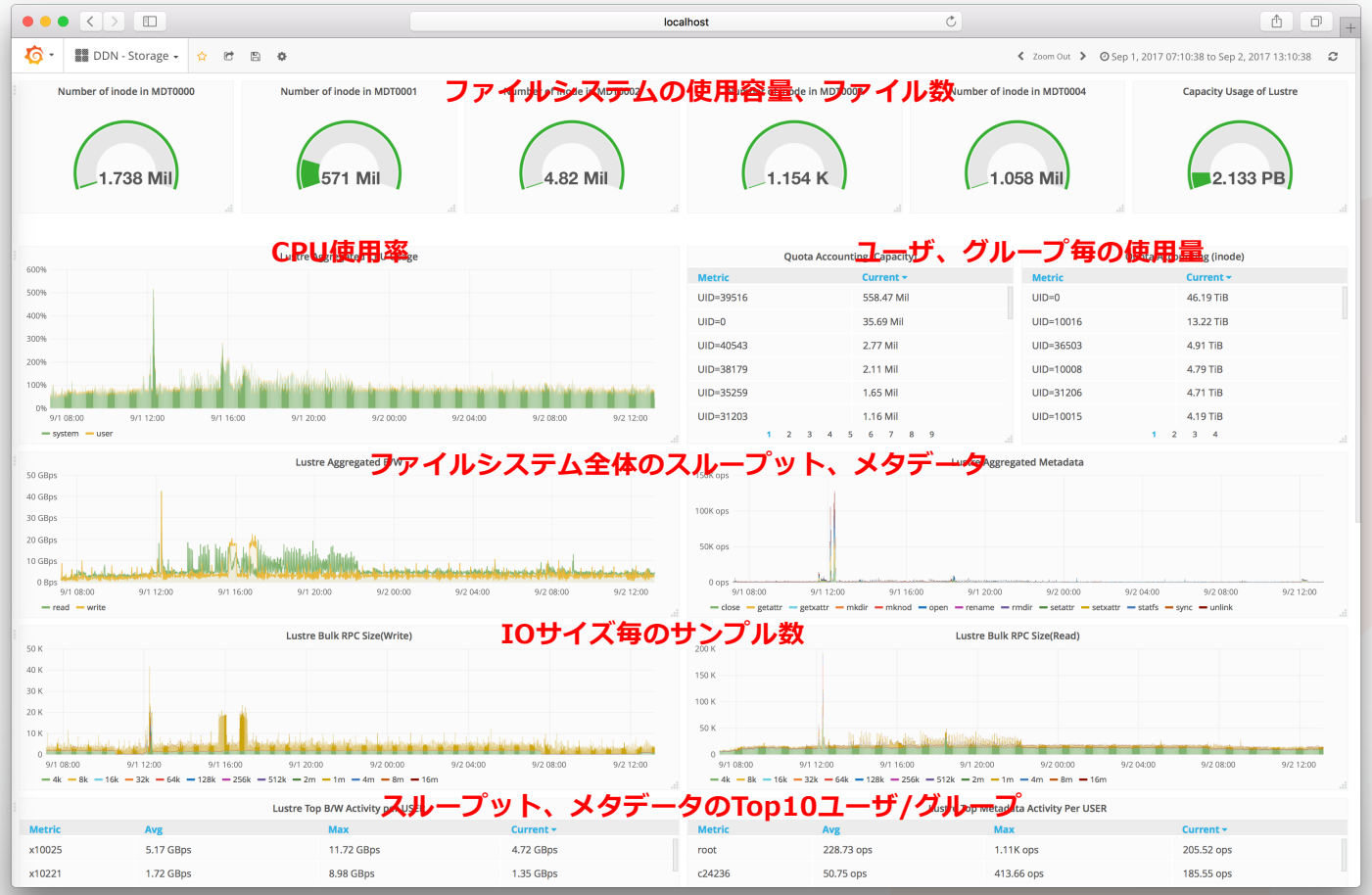
EXAScaler マルチテナント型セキュリティ

大規模な分散環境に対応する高度な
セキュリティとデータの安全性



- 不正なクライアント、VM、およびコンテナはファイルシステムのroot権限を取得不可
- クライアント/コンテナのセキュアな共有キーまたはケルベロス認証
- マルチテナントで特定のデータセットのみをユーザーにエクスポート
- セキュアな監査ログ
- ドライブおよびNVMe*における保存データ(Data at Rest)の暗号化

EXAScaler I/O モニタリング



詳細なIO性能をリアルタイムにモニタリングし、ユーザ毎のIO統計情報の取得やアプリケーションのI/Oパターンまたは特徴などの解析に有用


ddn