DDN Update
@Gfarm Workshop 2021

DataDirect Networks Japan, Inc.
Nobu Hashizume

2021/3/5

アジェンダ

- 2020年実績
- EXAScaler and A3I
- 2021年予定HW新製品
- RED

2020年実績

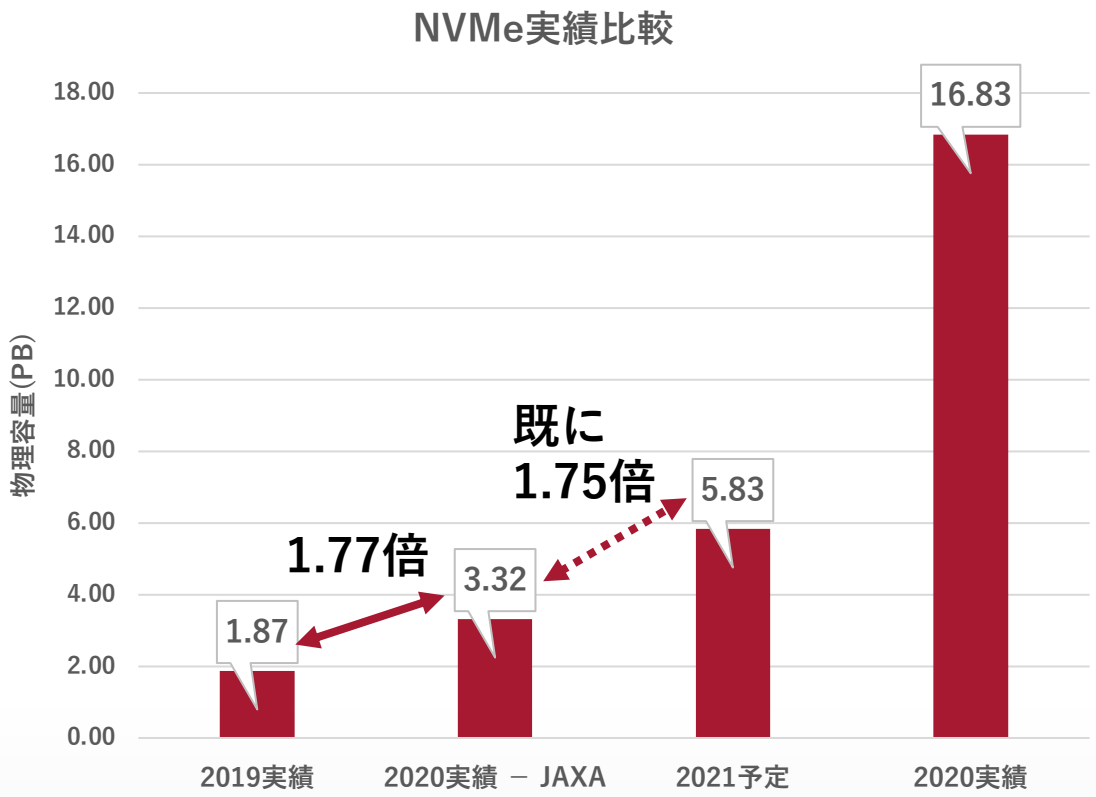
2020年導入実績

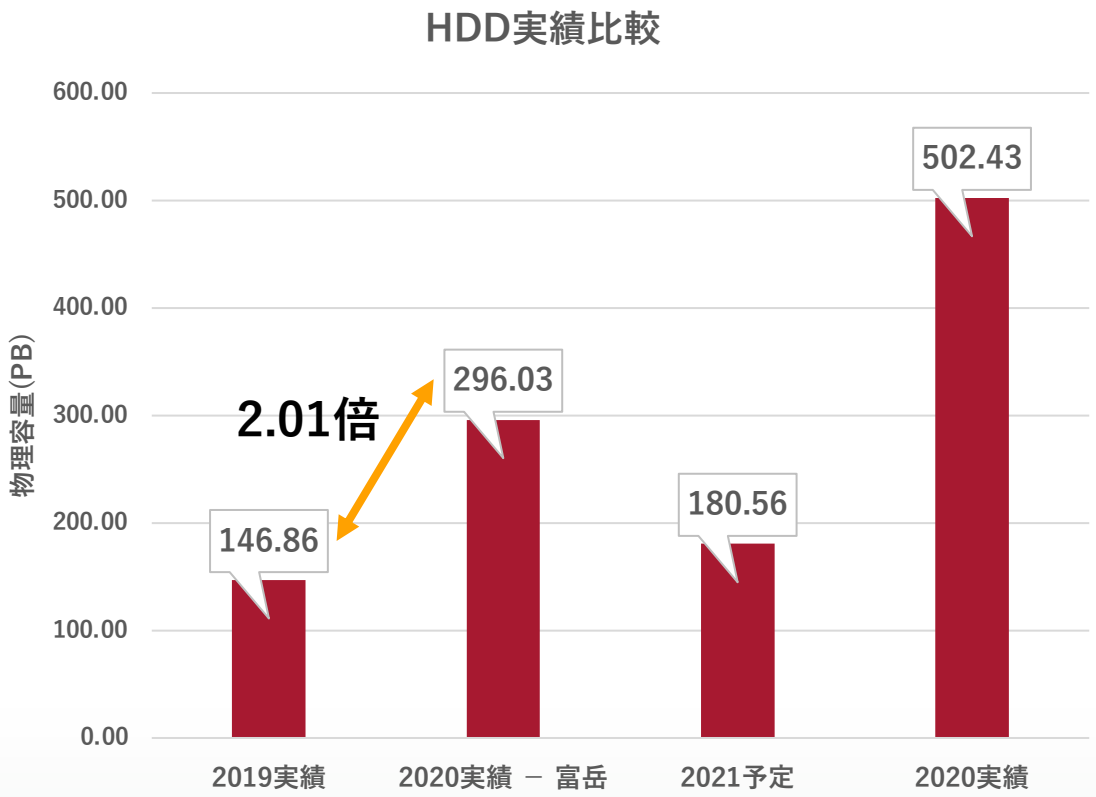
お客様	NVMe物理容量 (TB)	HDD物理容量 (PB)	ファイルシステム
理化学研究所 富岳		206.40	FEFS
海洋研究開発機構(JAMSTEC)	1766.40	83.23	EXAScaler
核融合科学研究所(NIFS) プラズマシミュレータ雷神		44.80	EXAScaler
名古屋大学情報基盤センター 不老		40.32	FEFS
理化学研究所 理研データ科学基盤	322.56	38.22	EXAScaler
宇宙航空研究開発機構(JAXA)	13516.80	16.40	FEFS
日本原子力研究開発機構(JAEA)		23.04	EXAScaler
東北大学メディカル・メガバンク機構(ToMMo)		16.16	EXAScaler
沖縄科学技術大学院大学(OIST)	720.00	8.16	EXAScaler
大阪大学核物理研究センター(RCNP)		8.88	EXAScaler
名古屋大学N研		4.03	EXAScaler
理化学研究所生命医科学研究センター(IMS)		3.20	EXAScaler
東北大学サイバーサイエンスセンター AOBA		2.56	ScaTeFS
某製造業		2.30	EXAScaler
某製造業		2.30	FEFS
京都大学基礎物理学研究所(YITP)		1.56	EXAScaler
某製造業	337.92		EXAScaler
某製造業	84.00		EXAScaler
某製造業	84.48	0.86	EXAScaler
合計	16.83PB	502.43PB	EXAScaler : 14 , FEFS : 4, ScaTeFS : 1

2020年12月～2021年導入予定

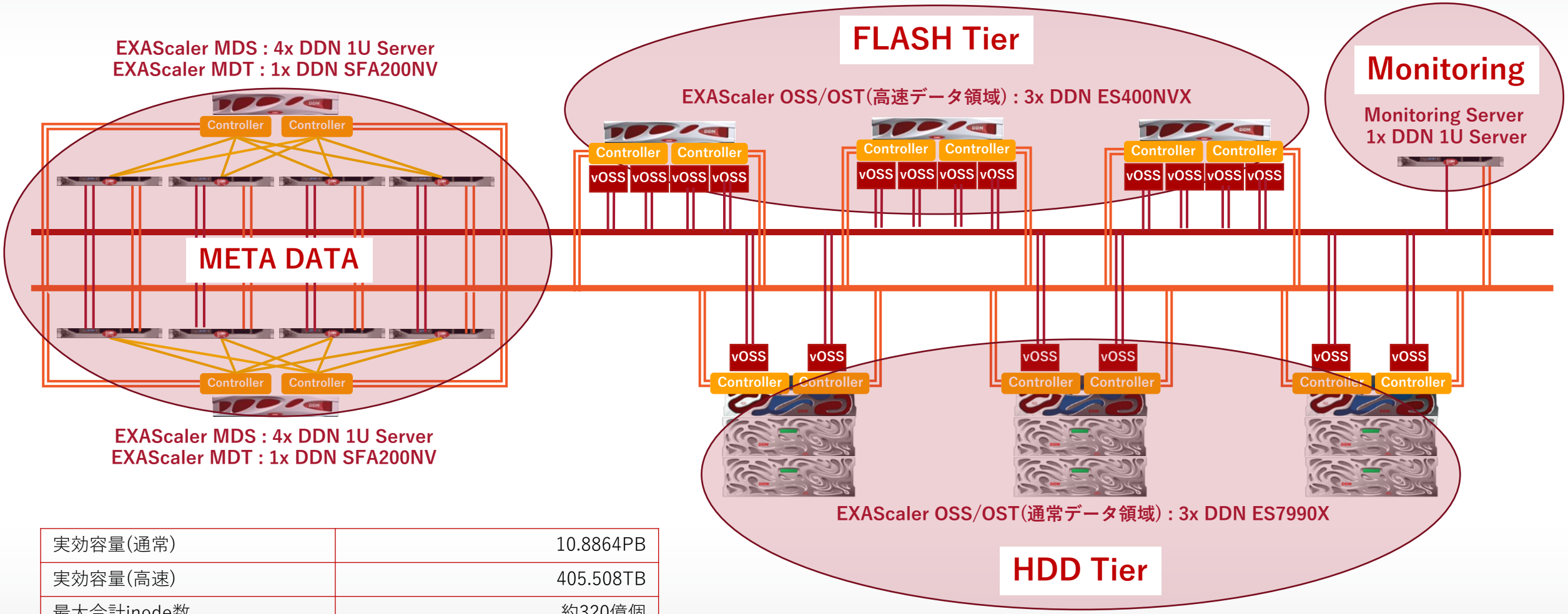
お客様	NVMe物理容量 (TB)	HDD物理容量 (PB)	ファイルシステム
某省庁		43.40	EXAScaler
東京大学情報基盤センター mdx	2027.52	35.78	EXAScaler
東京大学情報基盤センター Wisteria/BDEC-01	1413.12	34.18	FEFS
大阪大学サイバーメディアセンター SQUID	1536.00	26.88	EXAScaler
産業技術総合研究所(AIST) ABCI++	529.90	14.40	EXAScaler
国立環境研究所(NIES) GOSAT/GOSAT-2 プロジェクト		13.87	EXAScaler
国立遺伝学研究所(NIG) DDBJ		4.82	EXAScaler
名古屋大学宇宙地球環境研究所(ISEE)		4.14	EXAScaler
情報通信研究機構(NICT)		3.09	EXAScaler
北陸先端科学技術大学院大学(JAIST)	322.50		EXAScaler
合計	5.83PB	180.56PB	EXAScaler : 9 , FEFS : 1

2019年との比較



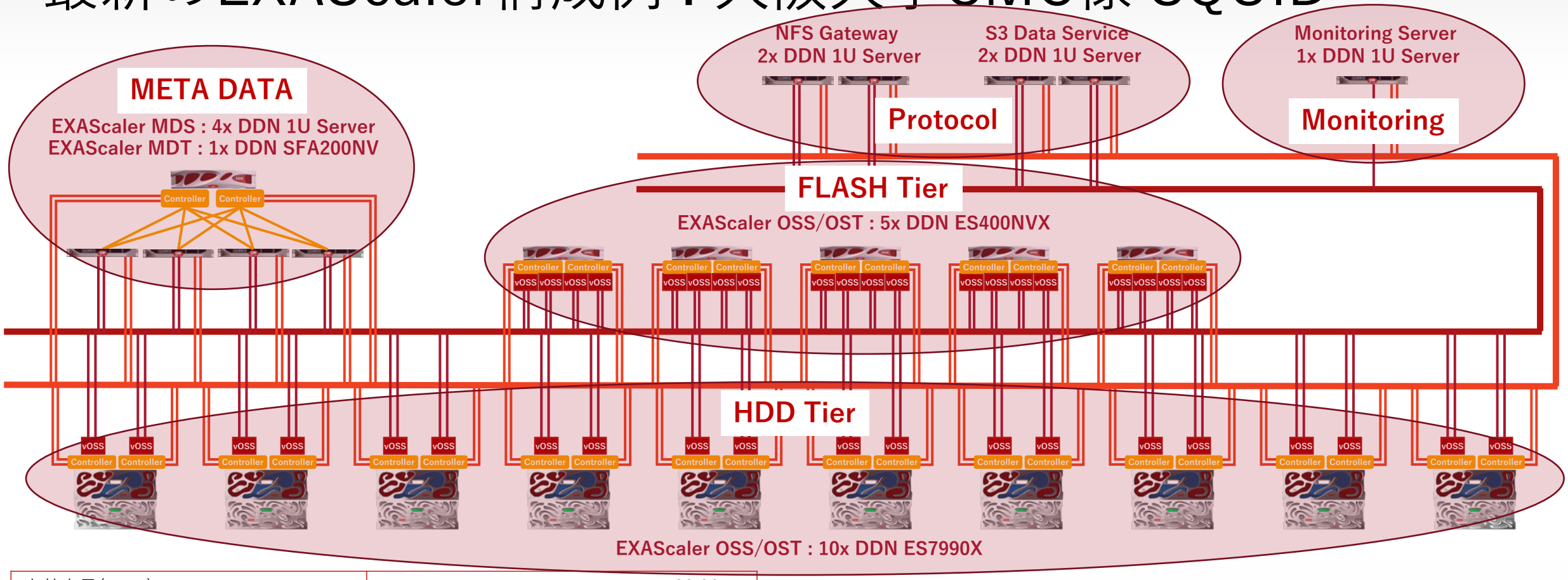


最新のEXAScaler構成例：産総研様 ABCI++



実効容量(通常)	10.8864PB
実効容量(高速)	405.508TB
最大合計inode数	約320億個
最大想定実効スループット(通常)	60GB/s
最大想定実効スループット(高速)	Write : 105GB/s, Read : 120GB/s

最新のEXAScaler構成例：大阪大学CMC様 SQUID



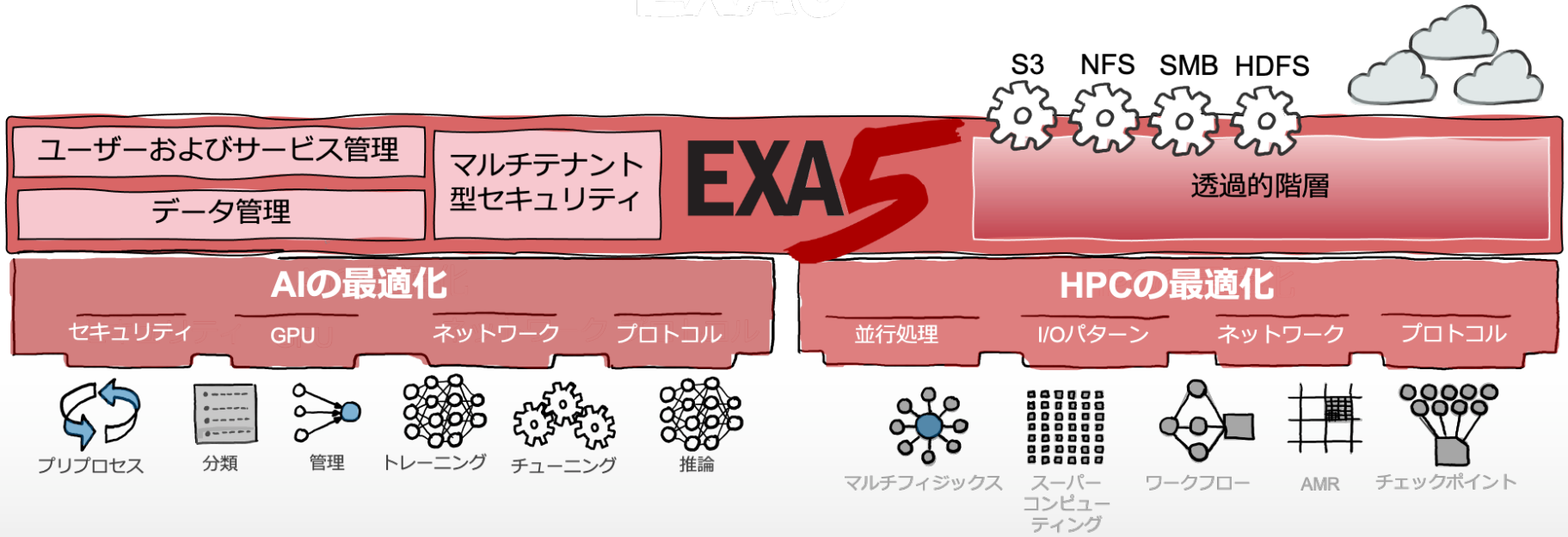
実効容量(HDD)	20.00PB
実効容量(NVMe)	1.20PB
最大合計inode数	約80億個
最大想定実効スループット(HDD)	160GB/s以上
最大想定実効スループット(NVMe)	Write : 160GB/s以上, Read : 180GB/s以上

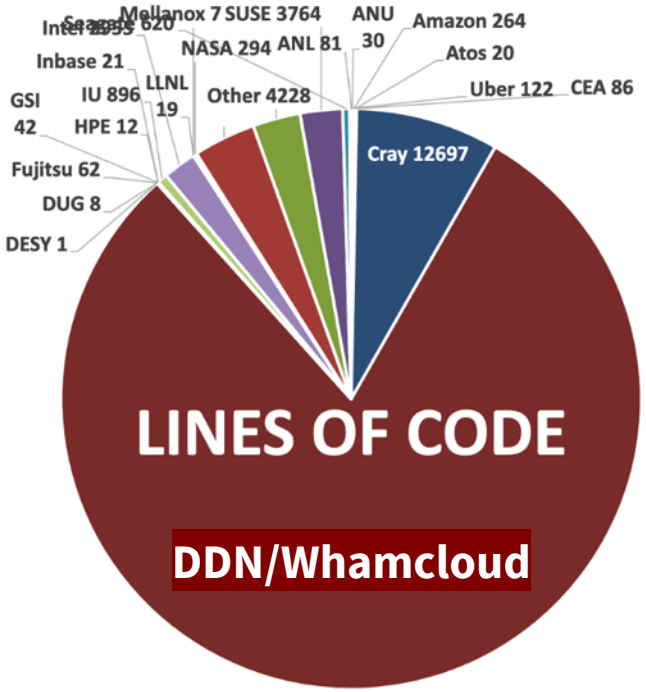
EXAScaler and A3I (Accelerated, Any-Scale AI)

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

- コアのファイルシステムはLustreを採用
- 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

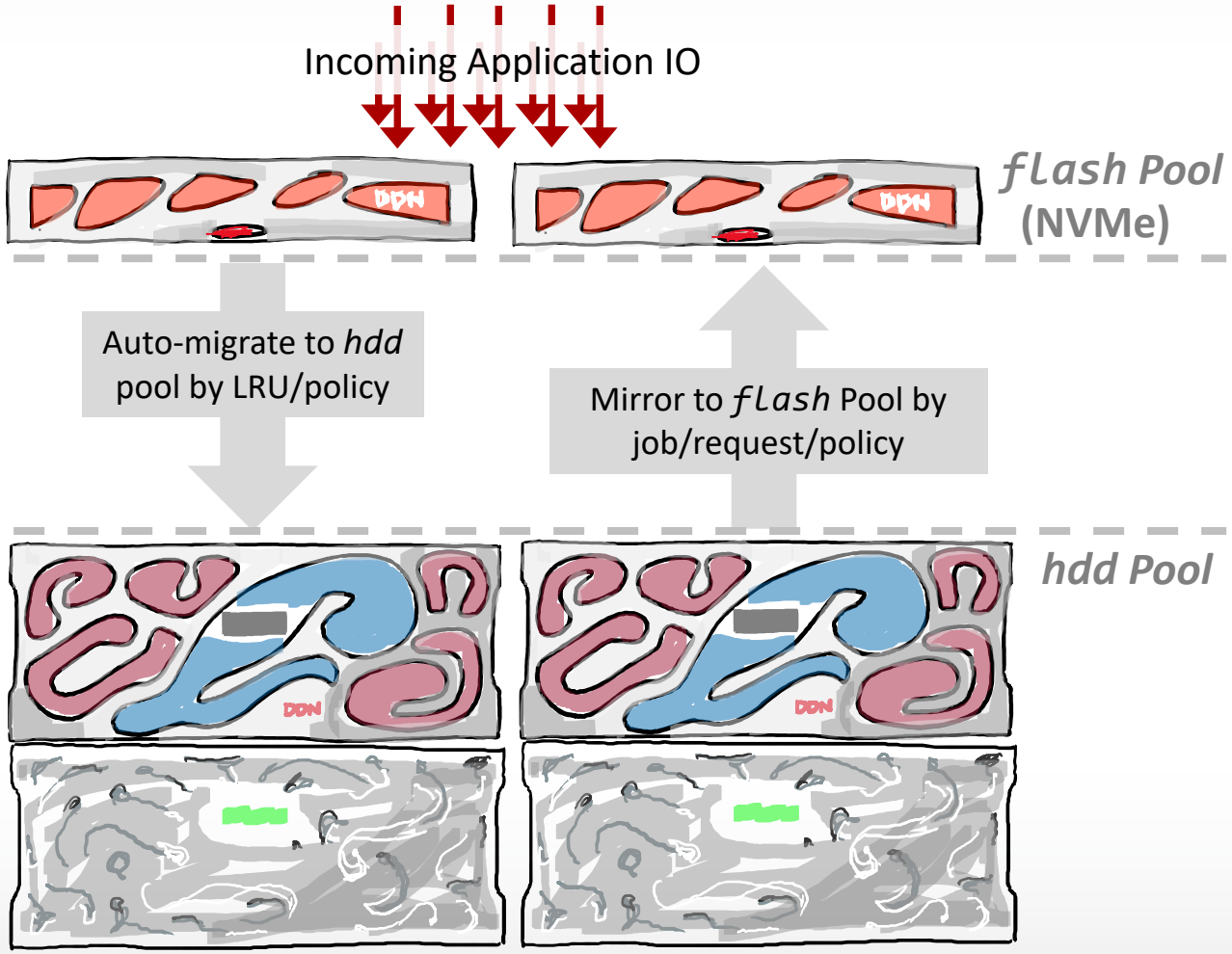




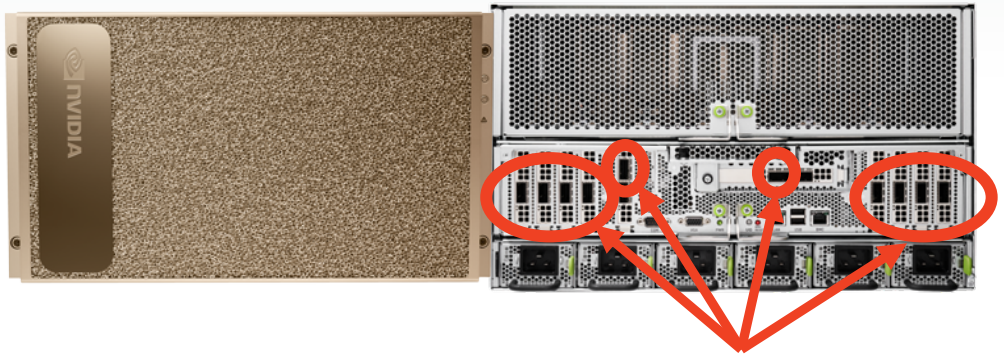
Lustre-2.13への貢献

EXA5 Hot Pools: データ自動キャッシュ/階層管理

- 新しいファイルは優先的にFlashへ
- 自動的にHDD poolへレプリカを作成
- 最大Flashの容量、使用率を考慮
- もしファイルが更新された場合、バックグラウンドにて再同期
- LRUポリシーにてFlashのレプリカはリリースされる
- 拡張のためCLI/APIを提供
- スケジューリング, ユーザリクエス等
- レプリカ HDD pool -> Flash pool
- 使用頻度の多いファイルをトラッキング

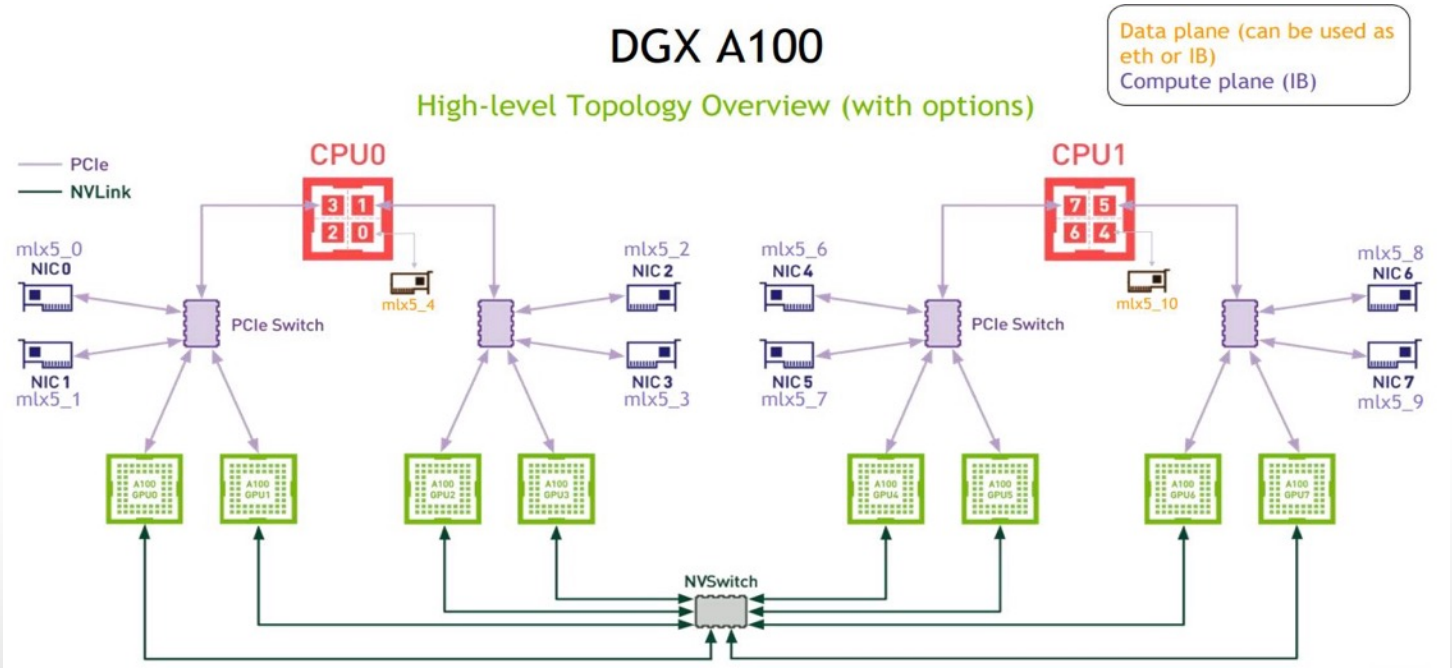


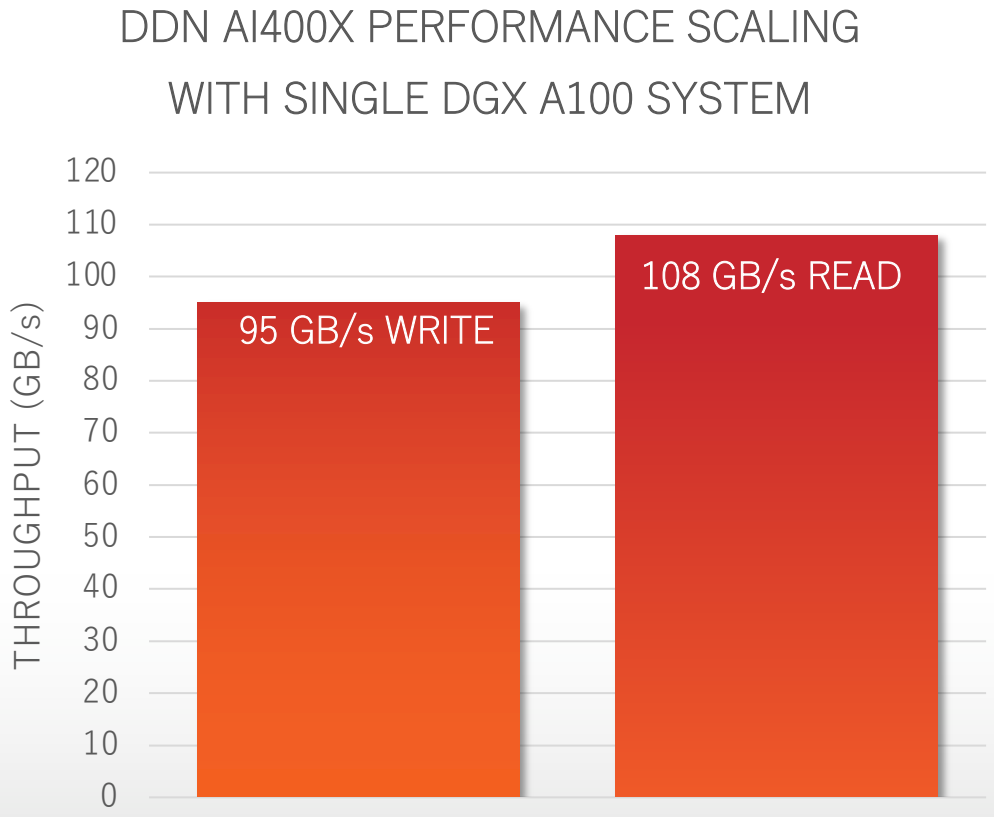
EXA5: シングルクライアント性能



8 x IB-HDR200+2 x IB-HDR200

- NUMAに最適化されたLustre Multi-rail
- 高いシングルクライアント性能

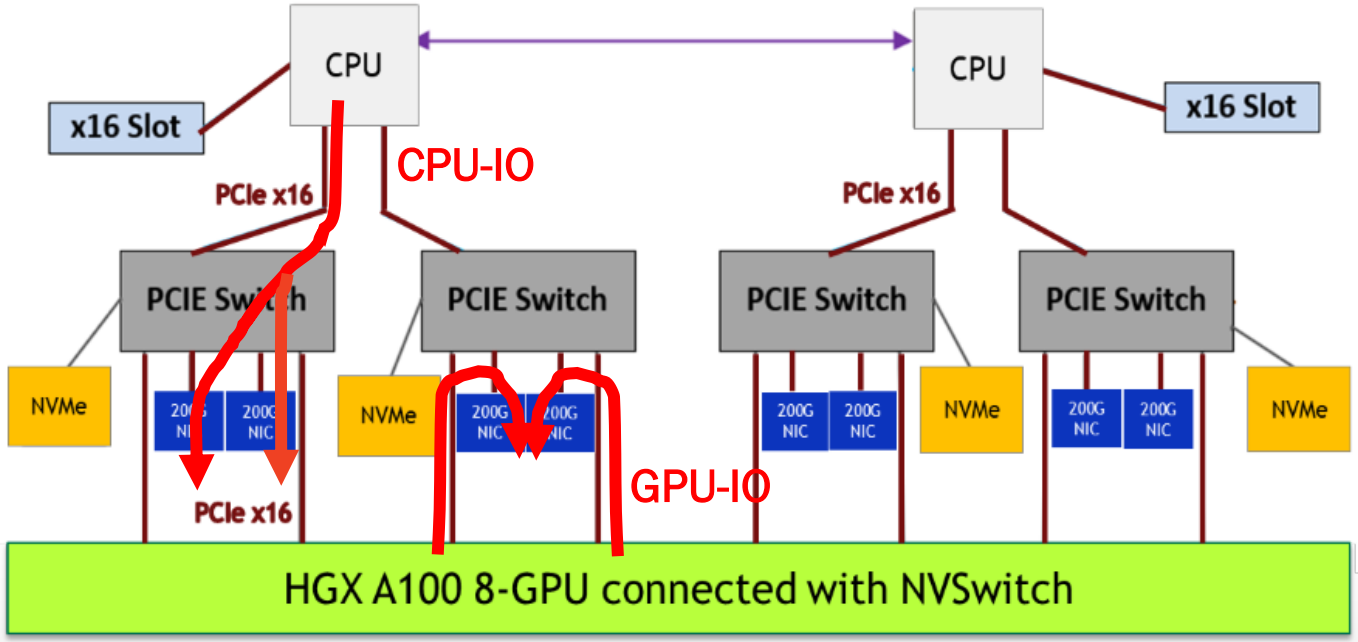


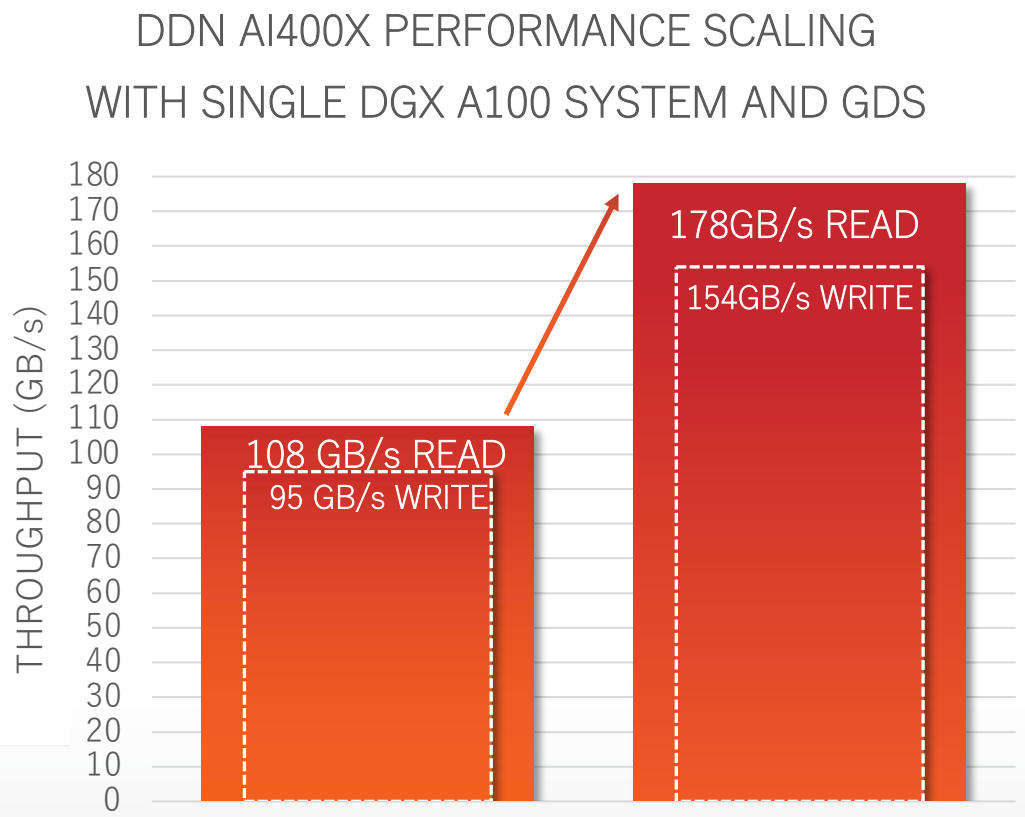


CPU-IO とGPU DIRECTストレージを使ったGPU-IO

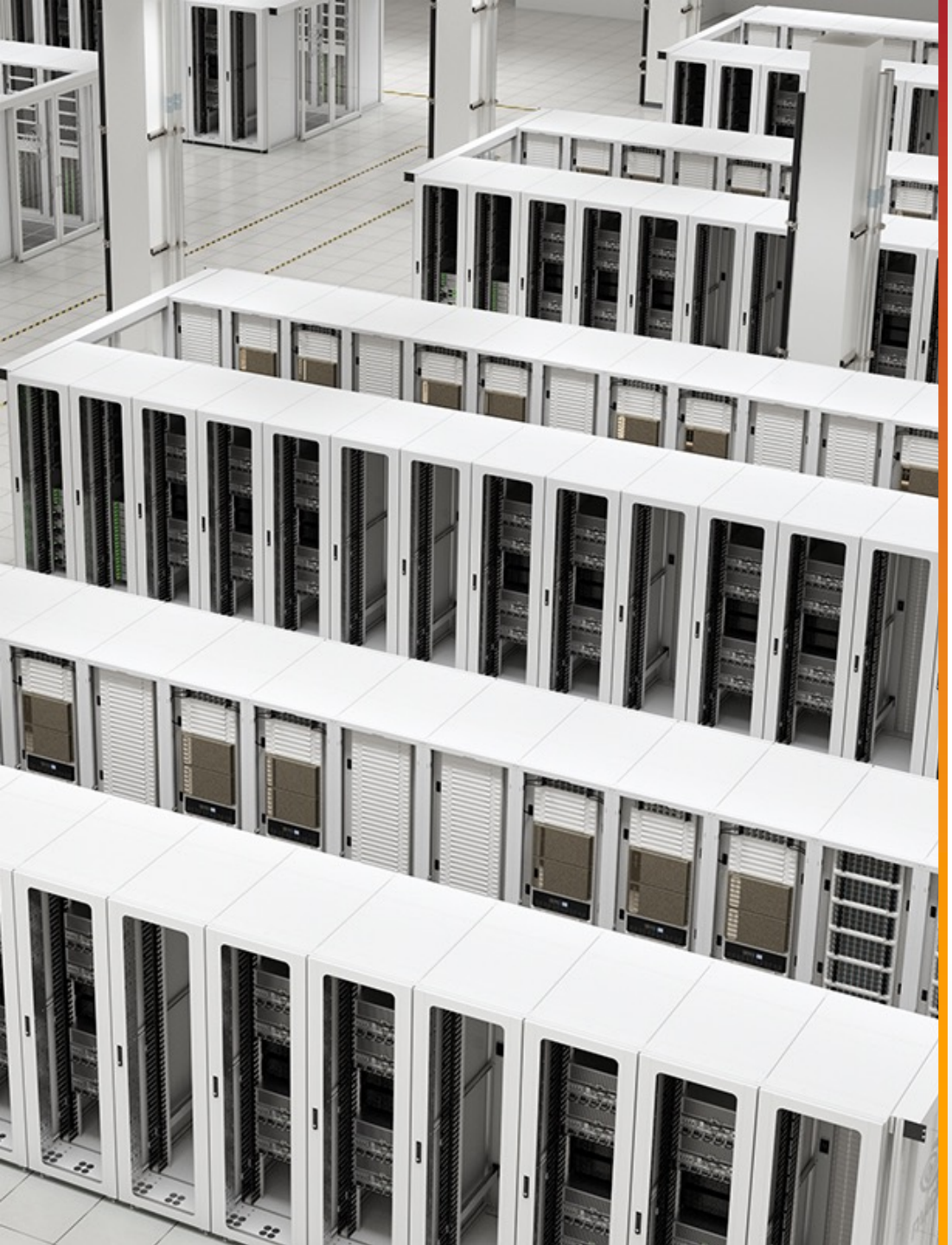
EXA5: GPU DIRECT STORAGE

- 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





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



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を最大のSuperPODとDGX A100のプロダクション環境で証明

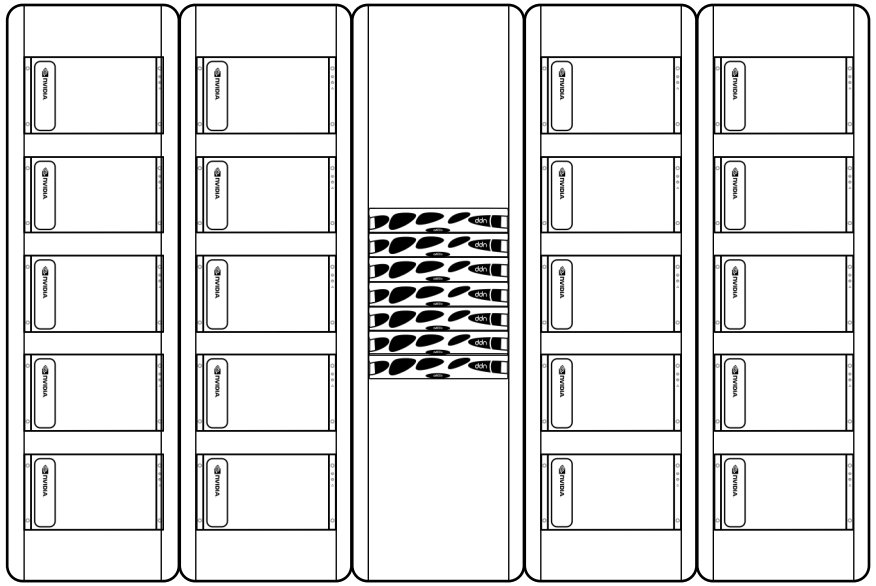
Top500のTop5にランキングされたSelene

- 560 NVIDIA DGX A100 システム
- 850 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に拡張

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 AI400X WITH SUPERPOD
ALL-NVME CONFIGURATION

大規模システムでリニアにスケールする 20x DGX A100用コンフィグレーション

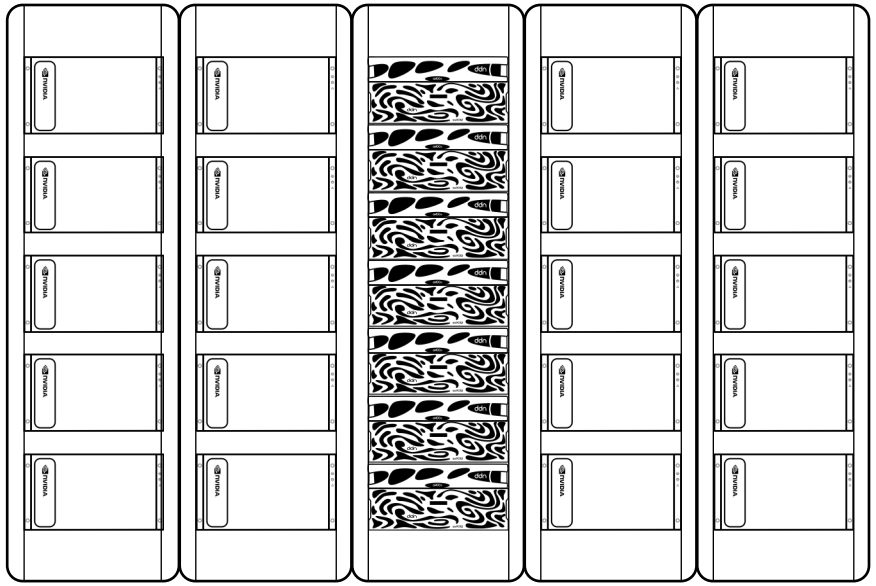


- 7 x DDN AI400Xアプライアンス
- 2 PB (NVME)
- スループット336 GB/s, 21M IOPS
- 56 x HDR100 IB or 100GbE
- 14 RU, 10.5 kW Nominal

- **あらゆるワークロードとデータタイプに最高のパフォーマンスを提供するAll-NVMe構成。どのスケールに於いても運用の柔軟性を最大化する究極のアーキテクチャ。**
- 統合された単一ネームスペースによってデータ管理のオーバーヘッドを排除し、GPUやコンテナ化されたアプリケーションがNVMeの利点を最大限に利用できます。
- **シンプルなAll-NVMeアプライアンスをビルディングブロックとし、リニアにスケールする環境を容易に、確実に構築できます。**
- **インジェストとアーカイブのために、ファイル、オブジェクト、クラウドベースのデータリポジトリと容易に連携可能です。**
- **NVIDIA DGX-2およびDGX A100 SUPERPODを利用するお客様サイトで導入され使用されています。**

DDN AI400X WITH SUPERPOD
HYBRID CONFIGURATION

大規模システムでリニアにスケールする 20x DGX A100用コンフィグレーション



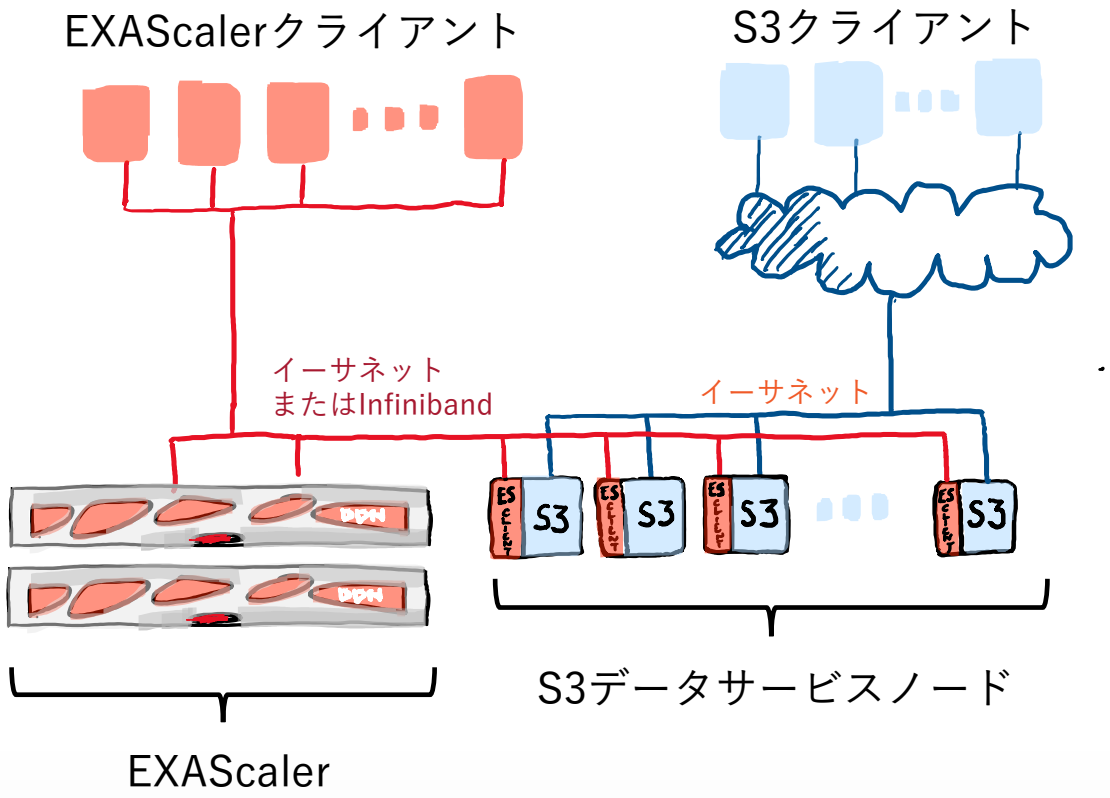
- 7 x DDN AI400Xハイブリッドアプライアンス
- 2 PB (NVME), 7PB/14PB (HDD)
- 336 GB/s スループット, 21M IOPS (NVME)
- 56 x HDR100 IB or 100GbE
- 42 RU, 18.5 kW Nominal

- ハイブリッド構成は、大容量ストレージリポジトリによりデータプラットフォームを増強でき、SuperPODでより大きなデータセットを扱いたいお客様に最適です。
- 変化するニーズに合わせて、性能と容量をそれぞれ個別に拡張できます。
- **可能な限り最大の密度と性能を最小限のフットプリントで構成できます。**
- 統合された安全なデータインジェスト、管理および保持機能により、エンドツーエンドのディープラーニングワークフローを簡素化します。
- **NVIDIA DGX-2 SUPERPODを利用するお客様サイトで導入され使用されています。**

EXA5: 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

- Lustreストレージ上にてS3 APIを提供
- S3とPOSIXの名前空間を共有
  - S3とPOSIXのデータアクセスの統合
  - 双方向からのデータアクセス
  - S3もしくはLustreからのデータの書き込み、読み込み
    “/s3_content/<バケット>/<ファイル名>…"
- シンプルなワークフローを実現
  - S3経由にてデータをインジェスト
  - データコピーなしに入力データとして活用
  - 結果をS3経由でデータ共有
- 複数サイトで採用
  - 東京大学mdx
  - 大阪大学SQUID
  - 理研データ科学基盤



2021年予定新HW製品

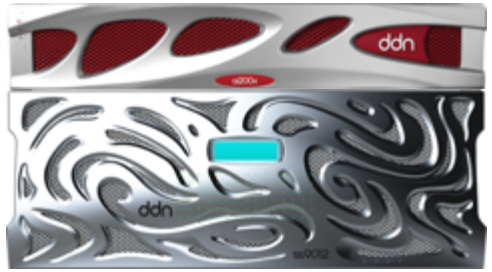
SFA Platform@2021

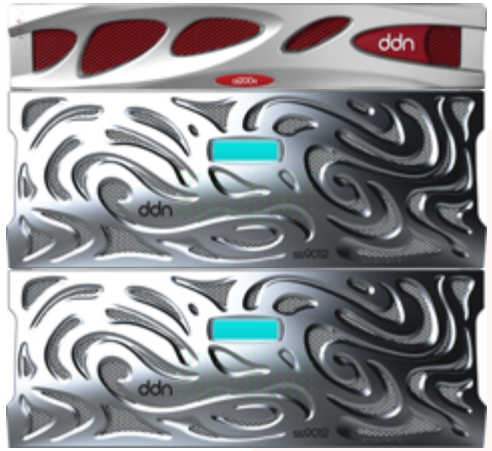
- 2021年Q3以降、PCIe gen4対応SFA Platformを提供予定
  - SAS-3 Hybridを提供するかは未定
  - SAS-4は2022年予定

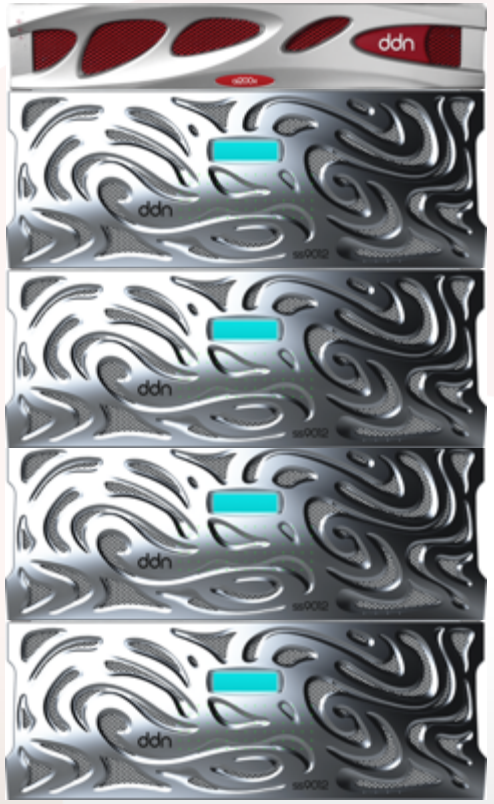
	Entry	High End
		
Class	All NVMe or Hybrid	All NVMe or Hybrid
Peak Performance	**40+GB/s**, 1.5M IOP/s	**90+GB/s**, 3M IOP/s
Media	24 NVME Slots Up to 360 SAS Slots	24 NVME Slots Up to 900 SAS Slots
Connectivity	HDR IB (200Gb) (4) Or 100/200 GbE (4)	HDR IB (200Gb) (8) Or 100/200 GbE (8)

Entry Planned SAS-4 Expansion Options@2022

24 NVMe Slots	90 SAS Slots	180 SAS Slots	360 SAS Slots
Max 366TB (15T NVMe)	Max 1.6PB (18T HDD)	Max 3.2PB (18T HDD)	Max 6.4PB (18T HDD)







- ALL SYSTEMS AVAILABLE EMBEDDED:
- ES AND A3I BLOCK IB UNDER REVIEW. NO FC PLANNED

High End Planned SAS-4 Expansion Options @2022

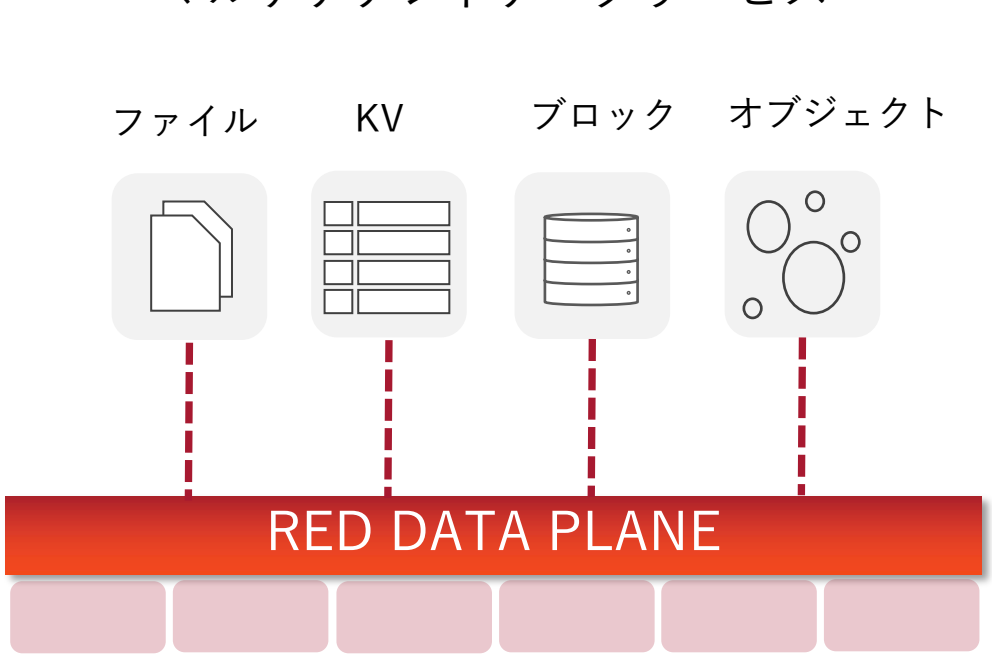
24 NVMe Slots	540 SAS Slots	720 SAS Slots	900 SAS Slots
Max 366TB (15T NVMe)	Max 9.6PB (18T HDD)	Max 12.8PB (18T HDD)	Max 16PB (18T HDD)

- ALL SYSTEMS AVAILABLE EMBEDDED:
- ES AND A3I BLOCK IB UNDER REVIEW. NO FC PLANNED

RED
(Reliable Elastic Data Services)

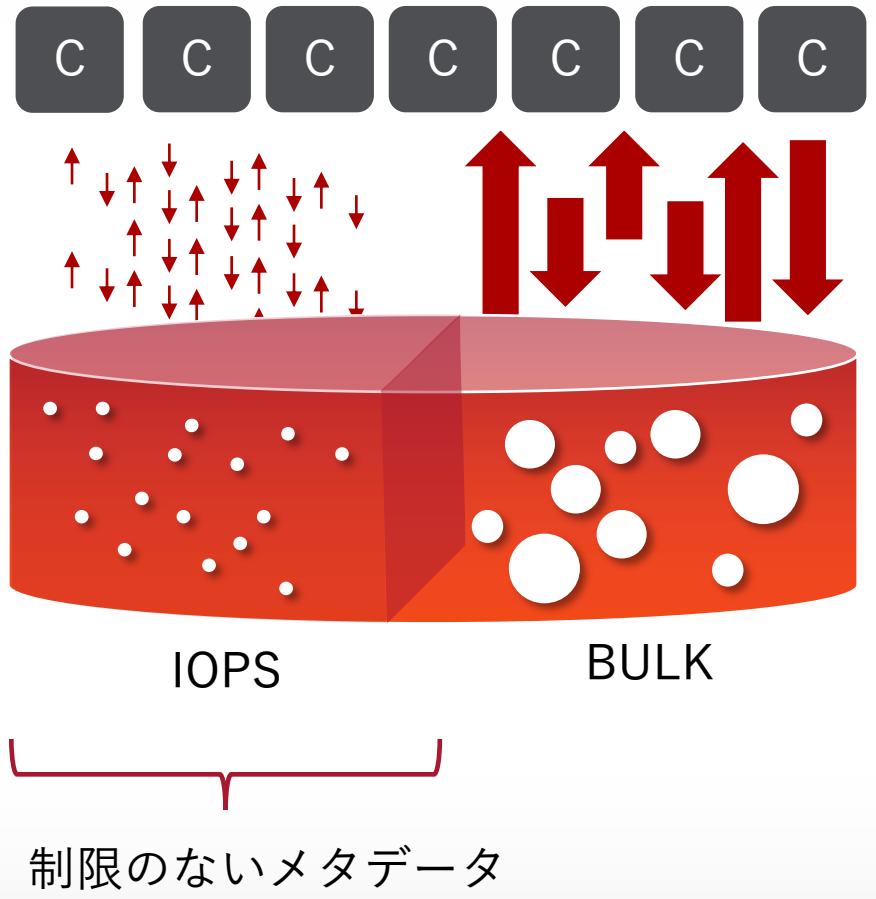
REDとは?

- NVMeにフォーカスした100%ソフトウェア・ ディファインド ・ストレージ
- スケーラブル・エラスティック・データサービス
- クラウドネイティブ、マルチテナント、仮想的なパーティション設計
- 様々なプロトコルでのデータサービス
- エンタープライズ機能を完備
- 制限のないメタデータ



RED 性能

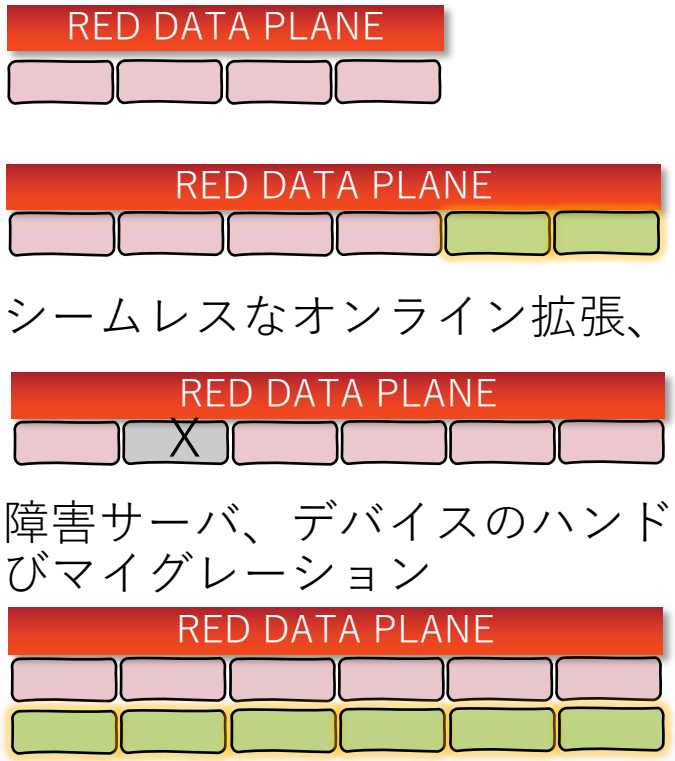
- ハイブリッドエラスティックエンジンを実装した唯一のストレージシステム
- 入出力データを自動的に正しいエンジンを判断
- それぞれのエンジンのデータ配置はIOパターンとサイズによって最適化される
- バイトアドレス、分散、ログ構造
- フレキシブルで効率的なイレイシャーコーディングでデータを保護



RED 柔軟性とシンプルさ

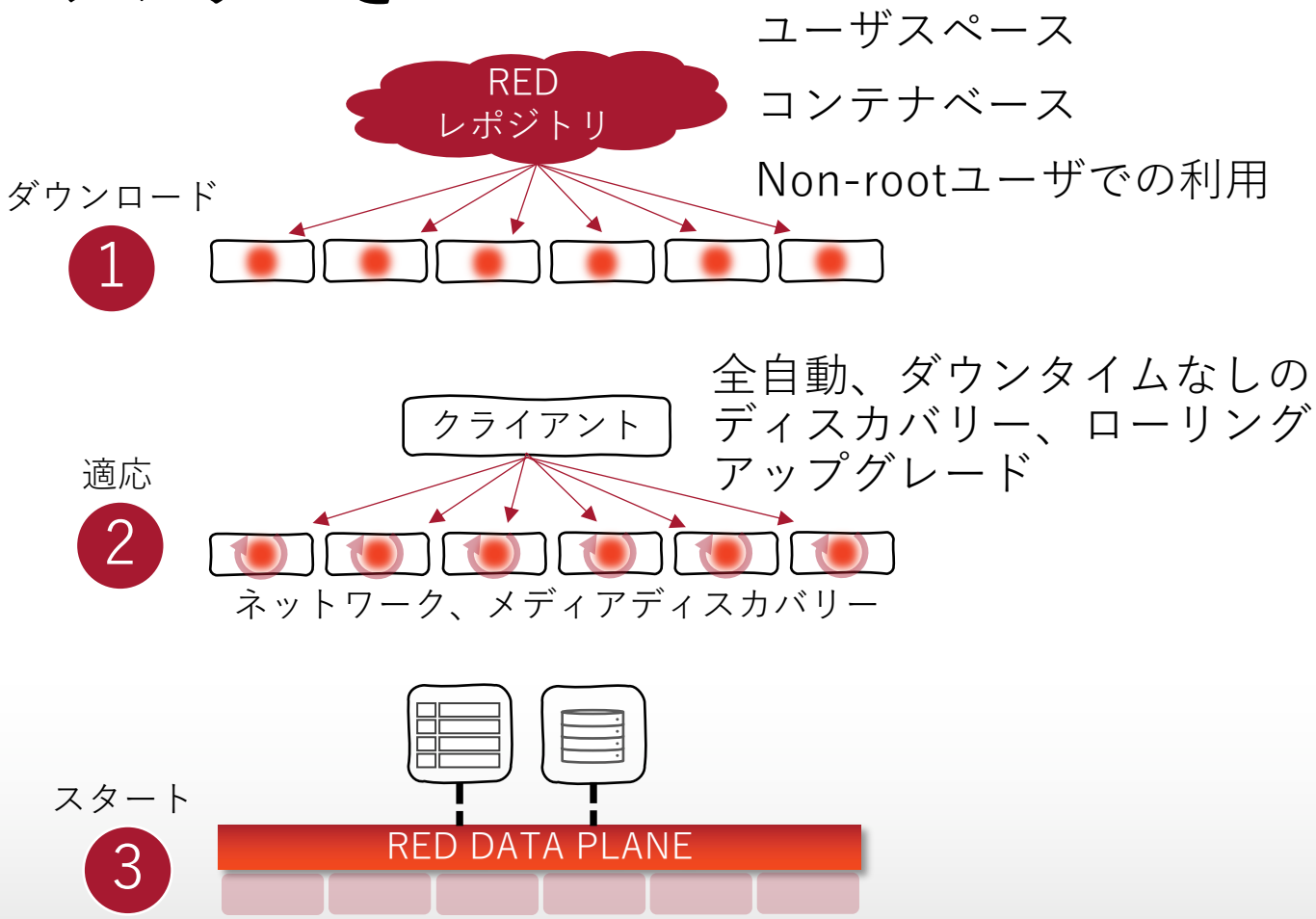
柔軟性

継続的なサービス



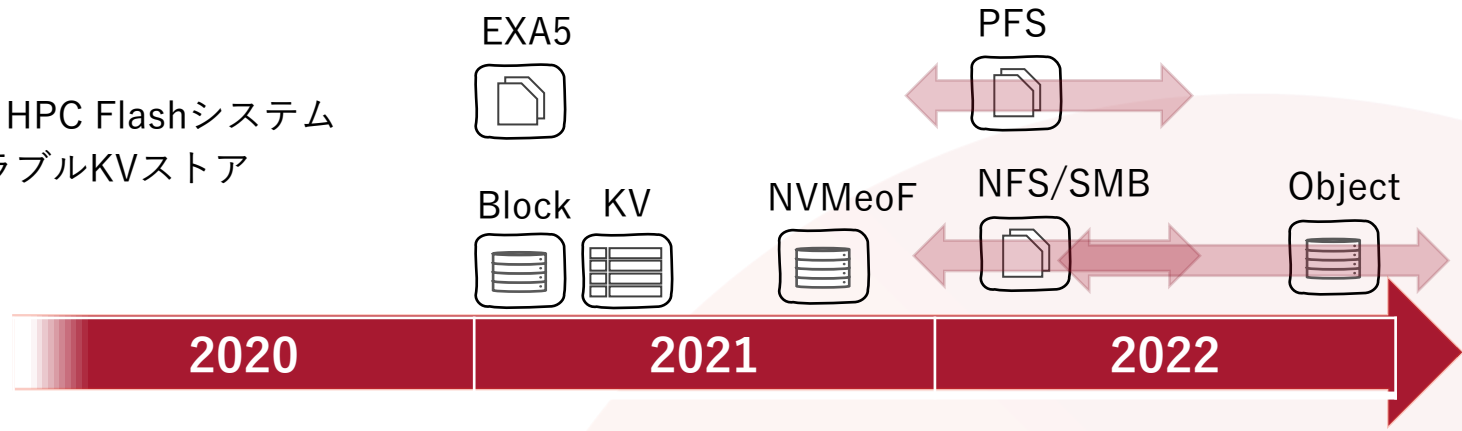
新しい世代のデバイスにおいても継続的なストレージサービスと性能を提供

シンプルさ

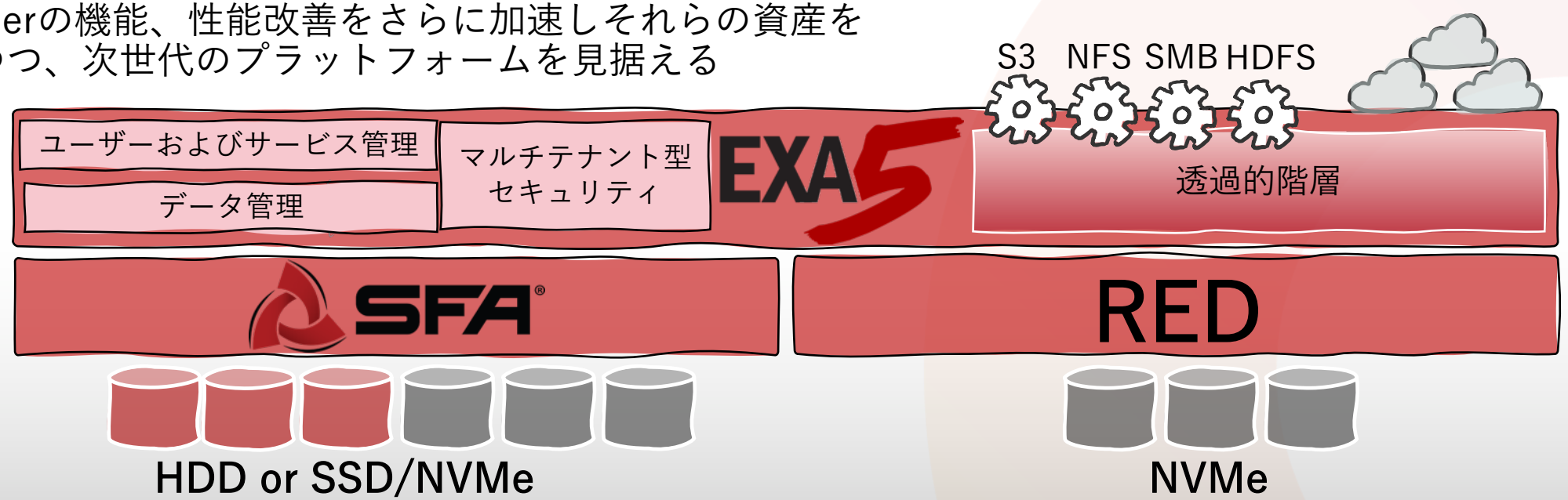


RED 包括的なデータサービス

- 2021 Q1
  - RED Block サポート EXAScaler: スケーラブル HPC Flashシステム
  - スケーラブルKey-Valueストア: 大規模スケーラブルKVストア
- 2021/2022以降
  - ネイティブファイルとオブジェクト



EXAScalerの機能、性能改善をさらに加速しそれらの資産を維持しつつ、次世代のプラットフォームを見据える



ddn