DDN Updata @PCCC22

DataDirect Networks Japan, Inc.

橋爪信明

2022/12/5

アジェンダ

- 2022実績(Lustreのみ)
- HW新製品紹介
- EXAScaler Update

2022年実績
(Lustreのみ)

2022年導入実績(1)

お客様	NVMe物理容量 (TB)	HDD物理容量 (PB)	ファイルシステム
某機構		67.00	EXAScaler
某省庁		59.60	FEFS
京都大学		40.00	EXAScaler
東京大学情報基盤センター Ipomoea01		34.40	EXAScaler
某民間企業		24.90	EXAScaler
某機構		19.58	EXAScaler
NEC	1474	18.43	EXAScaler
NICT UCRI	88	12.90	EXAScaler
筑波大学 Pegasus		9.60	EXAScaler
東京大学某研究所		6.10	EXAScaler
理研R-CCS		5.00	EXAScaler
某研究機関		3.33	EXAScaler
某製造業		3.20	EXAScaler
理研 AIP		2.98	EXAScaler
理研 CBS		2.98	EXAScaler
某製造業		2.56	EXAScaler
某民間研究所		2.40	EXAScaler
某機構		2.34	EXAScaler
某民間企業		2.30	FEFS

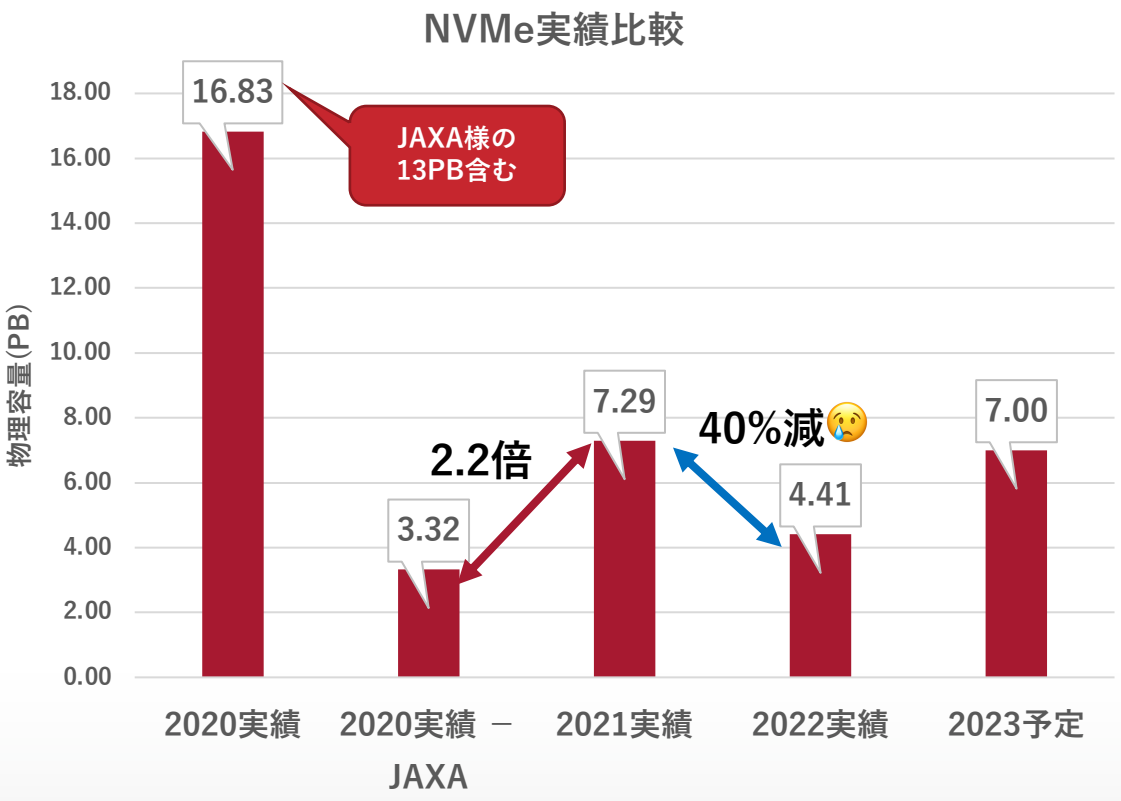
2022年導入実績(2)

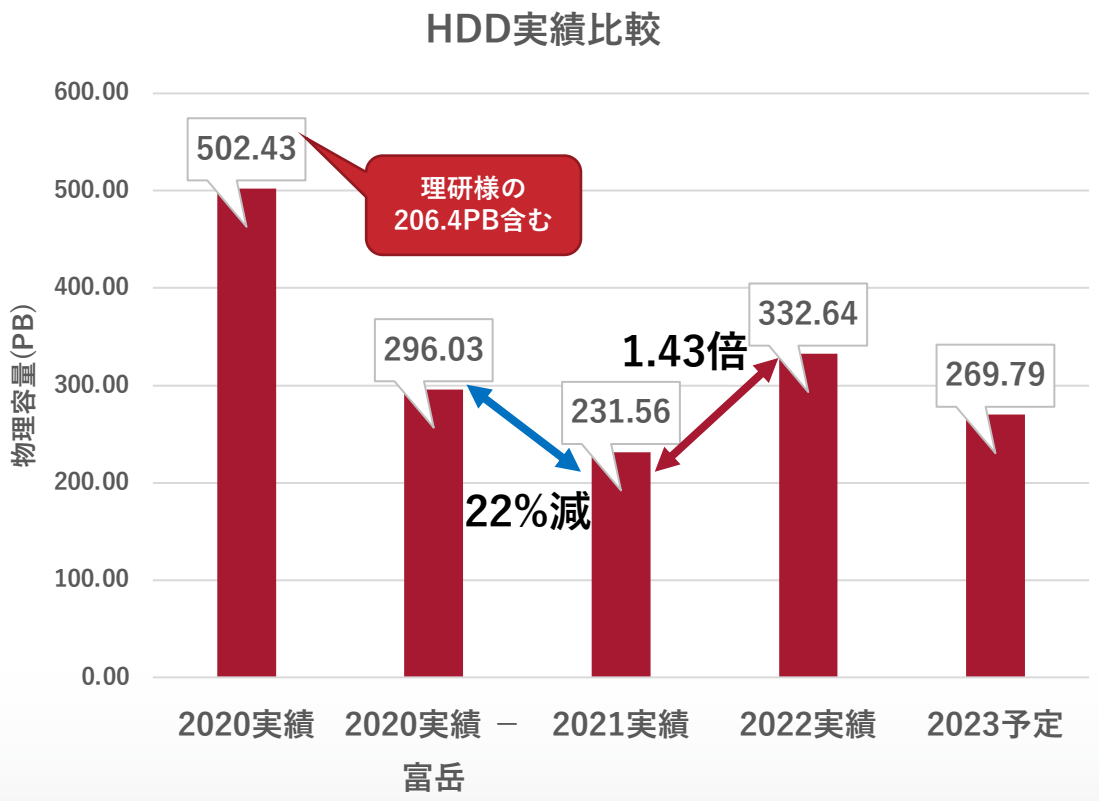
お客様	NVMe物理容量 (TB)	HDD物理容量 (PB)	ファイルシステム
某製造業		1.80	EXAScaler
某民間研究所		1.58	EXAScaler
OIST		1.50	EXAScaler
某製造業		1.28	EXAScaler
某製造業		1.20	EXAScaler
某製造業		1.00	EXAScaler
某製造業		0.86	EXAScaler
理研		0.85	EXAScaler
NAIST		0.84	EXAScaler
某製造業		0.80	EXAScaler
某省庁		0.70	EXAScaler
某製造業		0.63	EXAScaler
某機構	1000		EXAScaler
某製造業	737		EXAScaler
某製造業	368		EXAScaler
JAXA	350		EXAScaler
某機構	177		EXAScaler
某民間企業	135		EXAScaler
某通信事業者	76		EXAScaler
合計	4.41PB	332.64PB	

2023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	944	91.30	EXAScaler
某研究機関		53.20	EXAScaler
NICT Cinet	706	30.60	EXAScaler
理研Spring-8	154	28.50	EXAScaler
AIST北陸デジタルものづくりセンター(仮称)		17.00	EXAScaler
東京大学HGC		16.00	EXAScaler
某研究機関		8.92	EXAScaler
AIST		5.97	EXAScaler
東北大学		5.90	EXAScaler
東京大学 マテリアル先端リサーチセンター(ARIM)		4.00	EXAScaler
理研		2.98	EXAScaler
某製造業	460	1.62	EXAScaler
豊橋技術科学大学		1.60	EXAScaler
某製造業		1.20	EXAScaler
国立成育医療研究センター(NCCHD)		1.00	EXAScaler
某製造業	737		EXAScaler
京都大学	4000		EXAScaler
合計	7.00PB	269.79PB	

2022/21との比較





HW新製品紹介

New “X2” Platform

	ES200NVX2	ES400NVX2
		
Class / Controller	2U All NVMe Platform, Active/Active Dual Controller	
CPU	2x Ice Lake CPUs	4x Ice Lake CPUs
NVMe	24 Drive（PCI Gen 4）	
NVMe Performance	~45GB/s以上, 1.5M IOP/s	~90GB/s, 3M IOP/s
HDD	N/A	Max 900 Drive (10x SAS4 90Slot Enc)
HDD Performance	N/A	~80GB/s
Connectivity	HDR IB (200Gb/100Gb) (4ポート) Or 100/200 GbE (4ポート)	HDR IB (200Gb/100Gb) (8ポート) Or 100/200 GbE (8ポート) 提供予定：NDR200 IB (8ポート)

400NVX2E SAS-4 Expansion Options

NVMe Slots	24
Capacity	732 TB
SS9024 qty	2
Slots	180
Capacity	3.2 PB

NVMe Slots	24
Capacity	732 TB
SS9024 qty	4
Slots	360
Capacity	6.4 PB

NVMe Slots	24
Capacity	732 TB
SS9024 qty	5
Slots	450
Capacity	8.0 PB

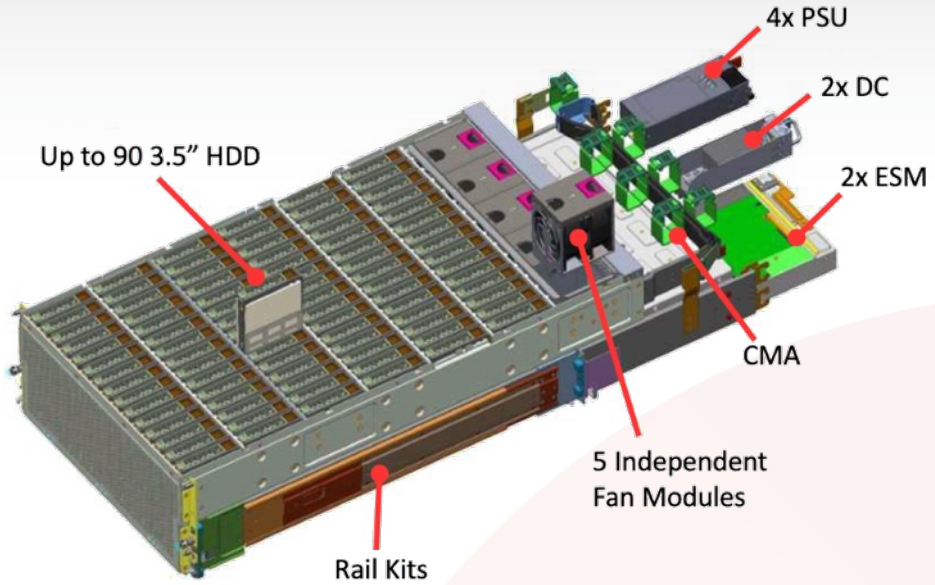
NVMe Slots	24
Capacity	732 TB
SS9024 qty	8
Slots	720
Capacity	12.8 PB

NVMe Slots	24
Capacity	732 TB
SS9024 qty	10
Slots	900
Capacity	16 PB

100% NVME without use of SAS expansion is also supported.
Capacity is maximum raw device capacity based on 30TB NVMe flash and 18TB SAS HDD. SAS SSD (up to 30TB) can also be used in SS9024.

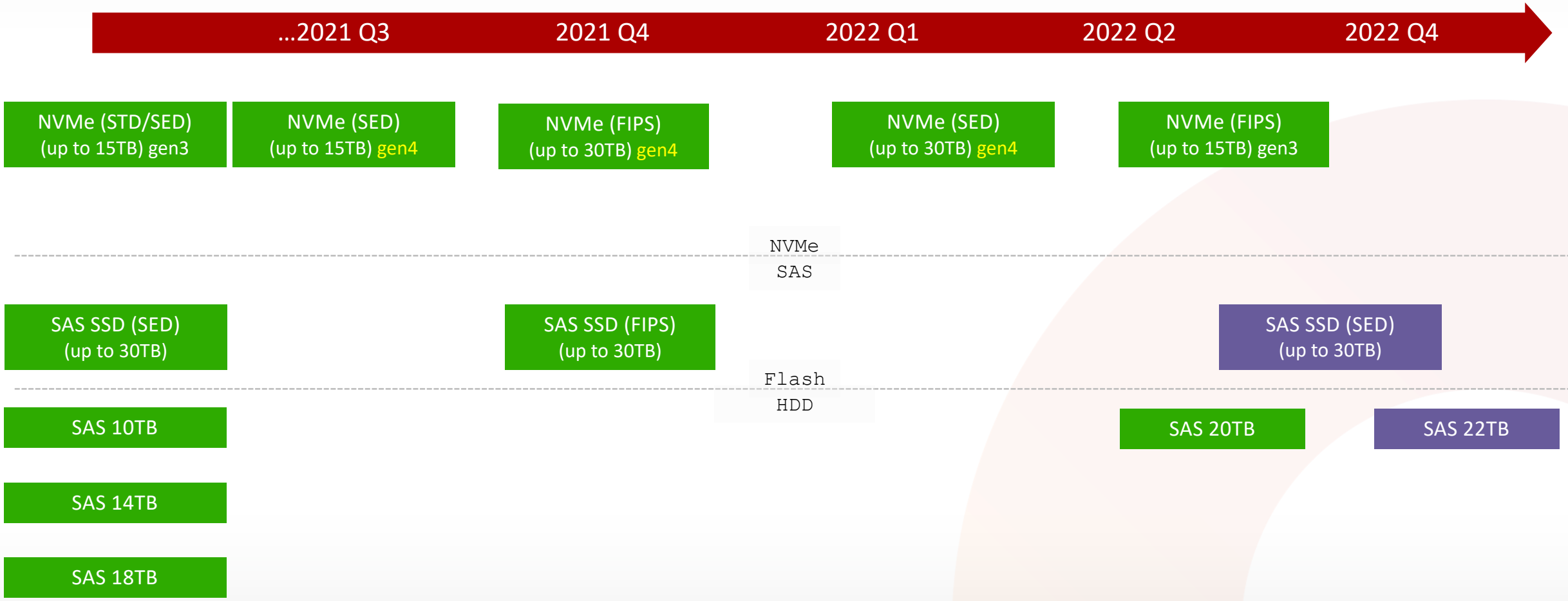
DDN SS9024

SS9024ディスクエンクロージャは、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
シャーシ	4U / 90 Drive
スロット数	90　(3.5inch/2.5inch)
IOモジュール	2x IOモジュール, SAS4(24Gb) 4x 4lane SAS 24Gb Mini SASポート/IOモジュール
対応ドライブ	SSD, SAS, NL-SAS
冷却ファン	5x 冷却ファン
電源	4 x 1300W 冗長電源 (2+2)
ホットスワップ対象部品	ドライブ、電源、冷却ファン
LCD Display	Status, Power, Environmental Monitoring
LED	Power, Status, Monitoring, Drive Activity

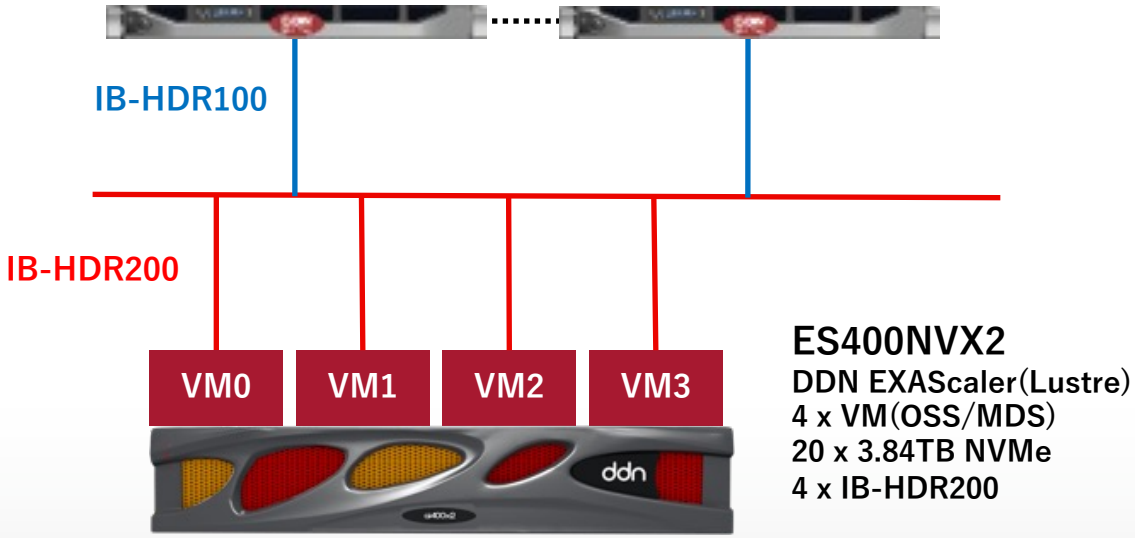
SFA Drive Lineup



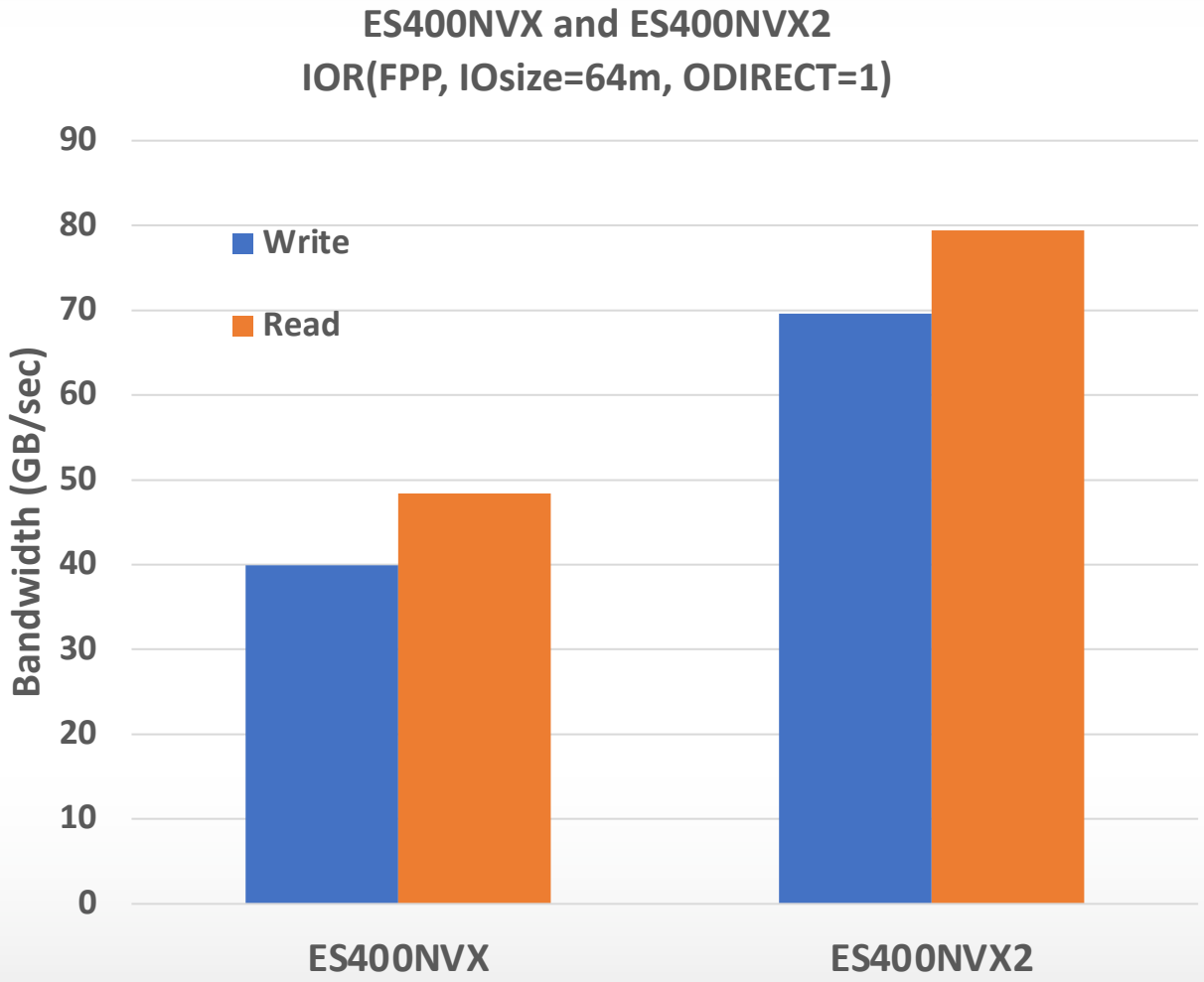
KEY: Released | Scheduled | Planning | Investigation

ES400NVX2 Flash Performance

40 x EXAScaler(Lustre) Client
1 x Gold 5218
96GB DDR4(2666MHz)
1 x IB-HDR100
CentOS8.1 (4.18.0-147.el8.x86_64)
Mellanox OFED-5.1-2.5.8.0



ES400NVX2
DDN EXAScaler(Lustre)
4 x VM(OSS/MDS)
20 x 3.84TB NVMe
4 x IB-HDR200



ES400NVX2 HDD Performance

高スループット

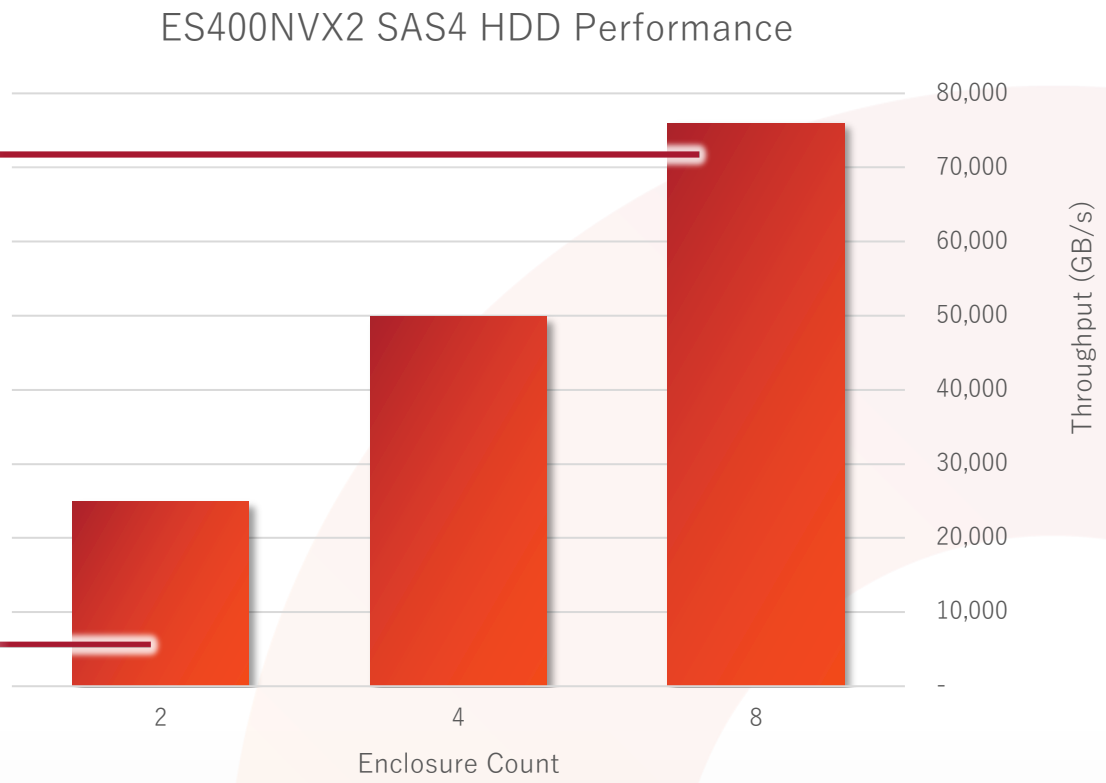
70GB/sを超えるスループット性能

高密度実装

1台のES400NVX2で20PBに迫る物理容量を提供

高い性能効率

アプリケーションに対して、ドライブあたり170MB/sを超えるファイルシステムレベルの性能効率を実現



DDN EXAScaler QLC Flashソリューション

- QLC Flashを24本搭載可能なエンクロージャを最大4基接続(2023年中)
  - 2024年に5エンクロージャ構成をサポート予定

2023Q4

DDN ES400NVX2/AI400X2

- All Flash EXAScalerファイルシステムを2RUで提供
- 90GB/s Read, 65GB/s Writeスループット
- 3M IOPS
- 700TBを超える**TLC** Flash容量

スイッチレスな Ethernet Fabricを構成

- 高性能で低遅延なNVMeoFを提供
- 単一障害点のない完全な冗長構成

DDN SE2420

- NVMeoF Flash拡張エンクロージャ
- 2RUで720TB以上搭載可能な高密度構成
- 4エンクロージャ構成で2.88PBの**QLC** Flash容量を実現
- IOM上にSwitch Chipを搭載するEmbedded Switching Fabric

SE2420：ES400NVX2 NVMeoF拡張エンクロージャ

- DDNが提供する高密度QLCを使用した新しいNVMeoF拡張エンクロージャ
  - 最大4エンクロージャ構成(2023)
  - 最大5エンクロージャ構成(2024)
  - 2,4,5エンクロージャ構成をサポート予定

Real-Time, Many-Core RAID Engine
Full Implementation of Real-Time Storage OS, Interrupt-Free; Massively Parallel I/O Delivery System

High Density Flash
24 QLC SSDs per enclosure
30TB initially, moving to 60TB

Parallel multi-channel rebuilds
Super fast rebuilds for large SSDs

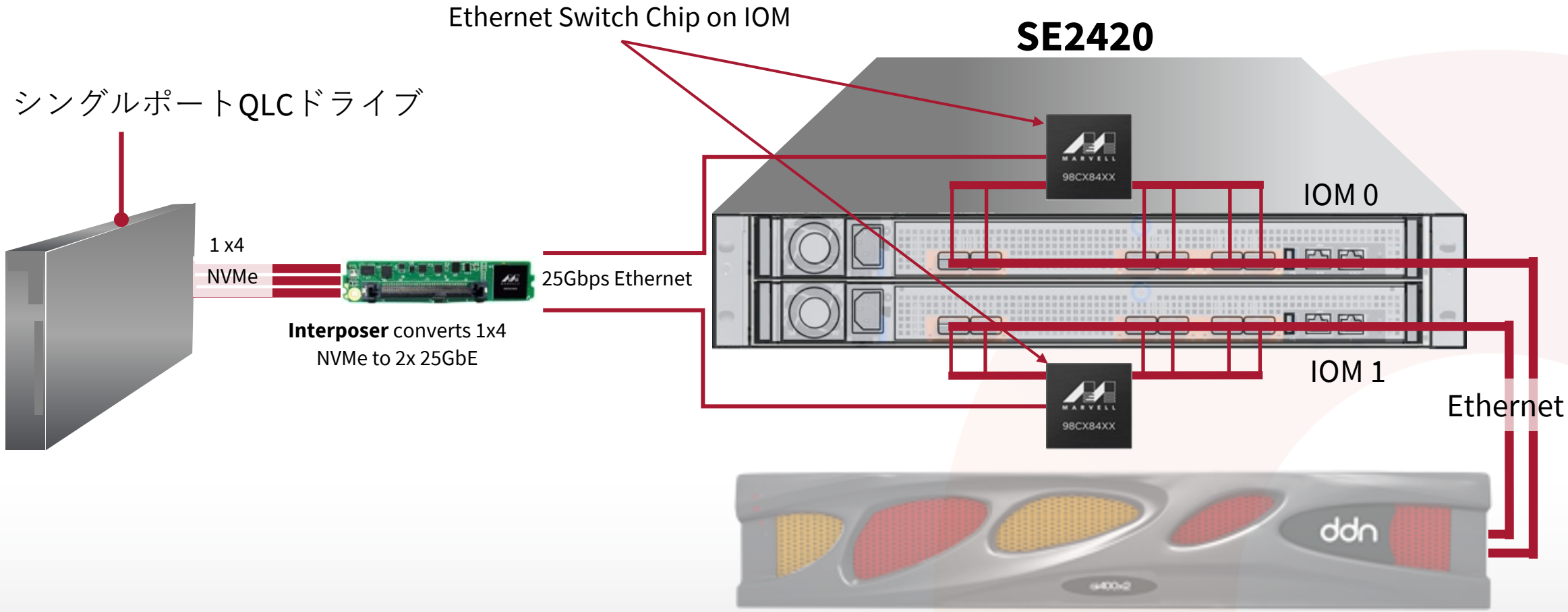
ReACT Cache Management
Real-Time I/O Data-Aware Cache Optimizer, Small IOs use Mirrored Battery Backed Cache (Write-Back), Large & Streaming I/Os Dynamically use Write-Through

No Single Point of Failure,
Zero Midplane Design

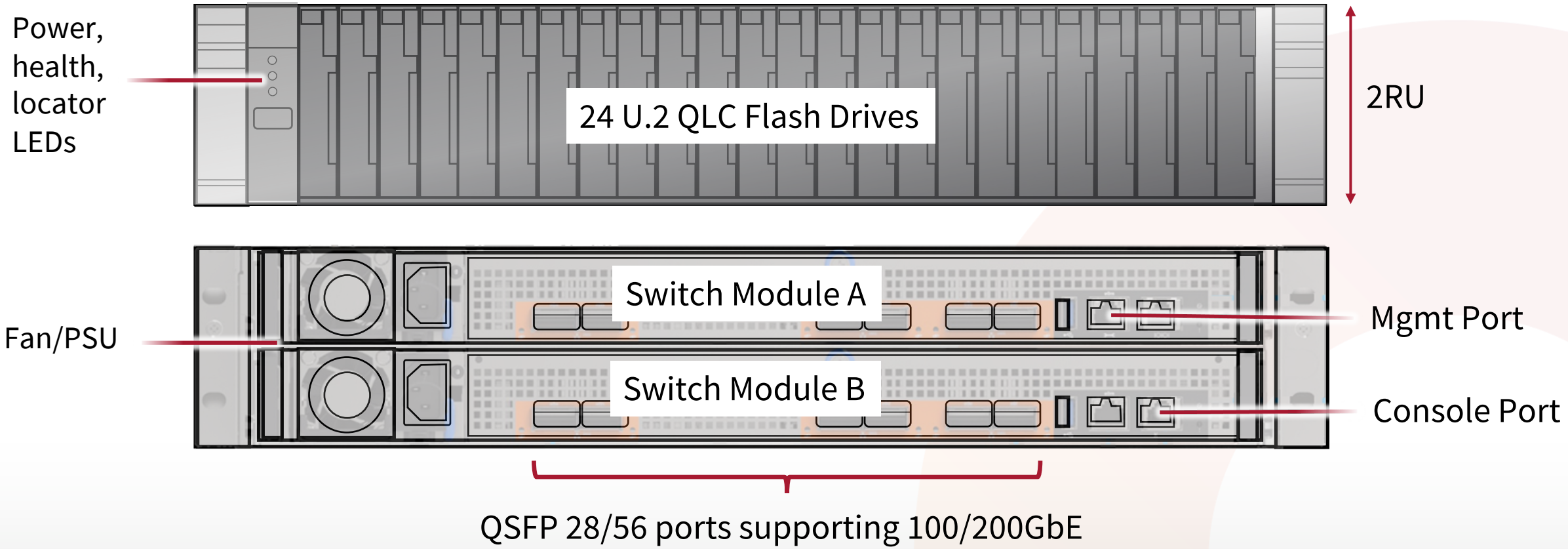
No External Switching - Storage Fusion NMVe Fabric
Highly-Over-Provisioned Backend RAIDed Fabric Withstands Failures of Drives, Enclosures, Cables, etc

Data Protection on disk
T10 PI Type 2

SE2420: NVMe Device ➡ Ethernet Fabric

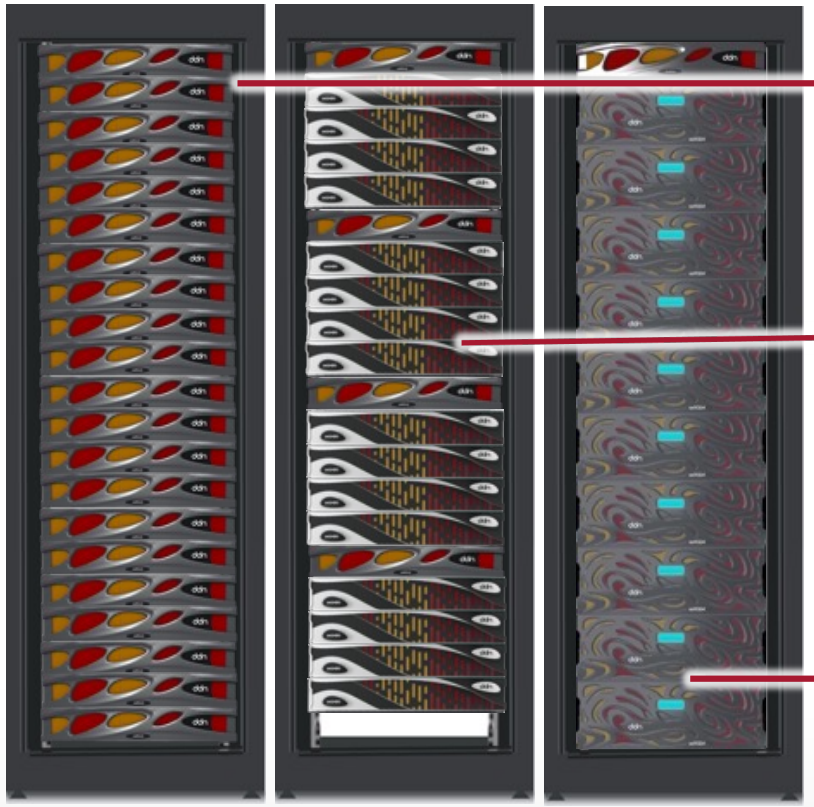


SE2420 - Front/Rear View



性能・容量要件に応じた柔軟な構成を提供

TLC Flash　QLC Flash　Hybrid



Best IOPS & Throughput per rack

63M IOPS by 21x ES400NVX2
1.89 TB/s READスループット
1.26 TB/s WRITEスループット
物理容量 **15PB** Flash領域

Best Price per Flash TB

12M IOPS by 4x ES400NVX2
360 GB/s READスループット
240 GB/s WRITEスループット
物理容量 **26PB** Flash領域

Best Price per TB

物理容量 **19.8PB** HDD領域, **0.7PB** Flash領域by 1x ES400NVX2
80 GB/s READスループット
65 GB/s WRITEスループット

EXAScaler Update

EXA6の主な新機能

Lustre2.14をベースとした新EXAScalerバージョンEXA6を昨年リリース済

Security・Compliance

- **Client-side file Encryption**
  fscrypt APIによるファイル暗号化に対応。ディレクトリ単位で暗号化を適用可能

Performance

- シングルスレッド性能の向上 "15GB/sec”
- **ロックレスIO**
  Direct IO時、Server, Client間でファイルlockを行わないことでオーバヘッドを削除し低Latencyアクセスを実現
- **Lustre Over Striping**
  OST数以上のストライプ数を設定可能→Single Shared Fileの性能向上

Cache Management

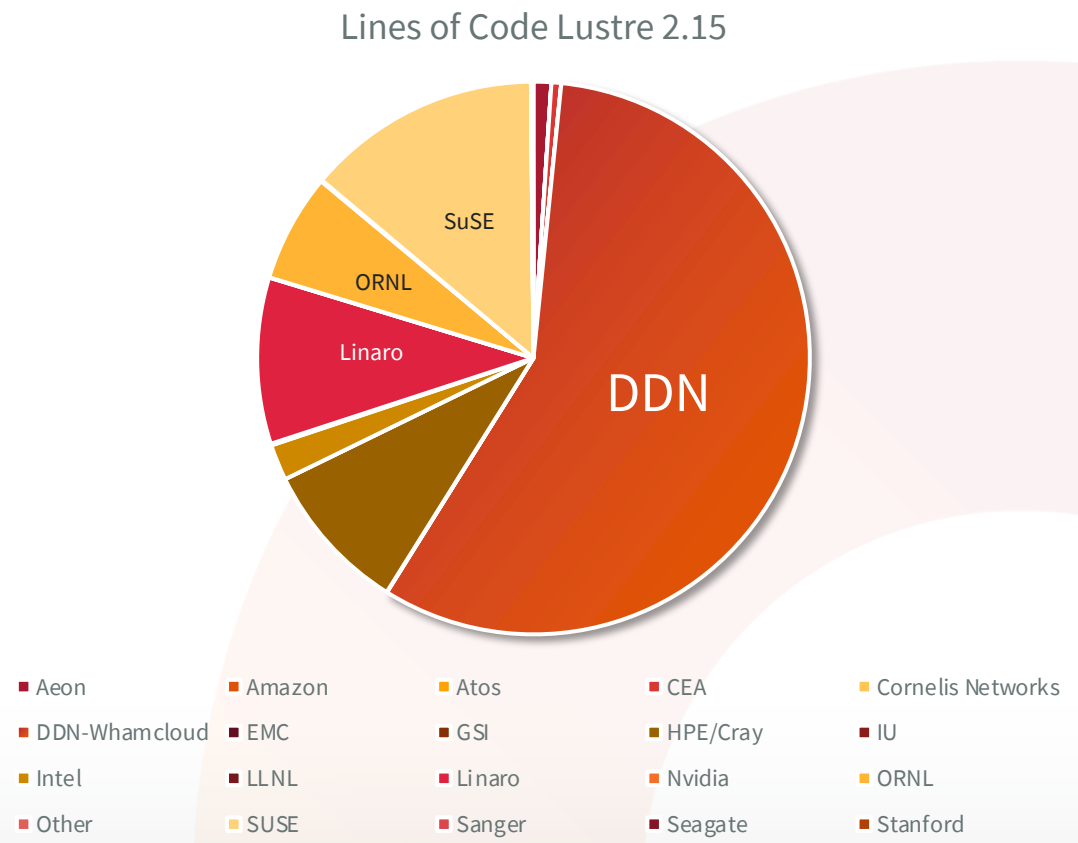
- **Hot Pools**
  NVMe OST、HDD OST間のTieringを実現
- **Hot Nodes**
  クライアントのローカルストレージを Cacheとして利用
  IOPSが必要なアプリケーションの性能向上

Efficiency

- **OST Pool Quota**
  １つのファイルシステムに混在する異なるデバイス(HDD, NVMe)毎にそれぞれOST Poolを作成して、異なるQuota設定可能
- **Auto Directory Split**
  同一ディレクトリ内でinode数が設定値を超えた時点から複数のMDTを自動的に利用

Lustre 2.15 – Communityへのコントリビューション

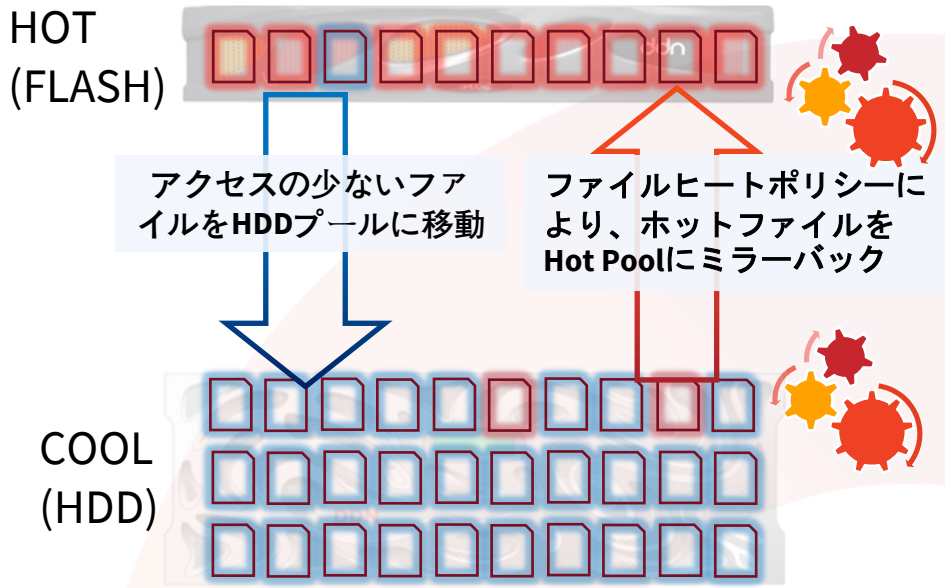
- DDN/Whamcloudがオープンソースプロセスとツリーすべてのメンテナンス管理を行っています
- DDNは引き続きオープンソース（コミュニティ）Lustreのメイン・コントリビューターです
- OpenSFSは全体的な方向性とユーザ間の議論のためのフォーラムを提供します
- EOFSが主催するヨーロッパでのユーザ会議
- DDN/Whamcloud主催のアジアでのユーザ会議



Hot Pools - 記憶メディアの階層構造に対応

EXAScalerファイルシステム上のストレージプール間の自動階層化を実現

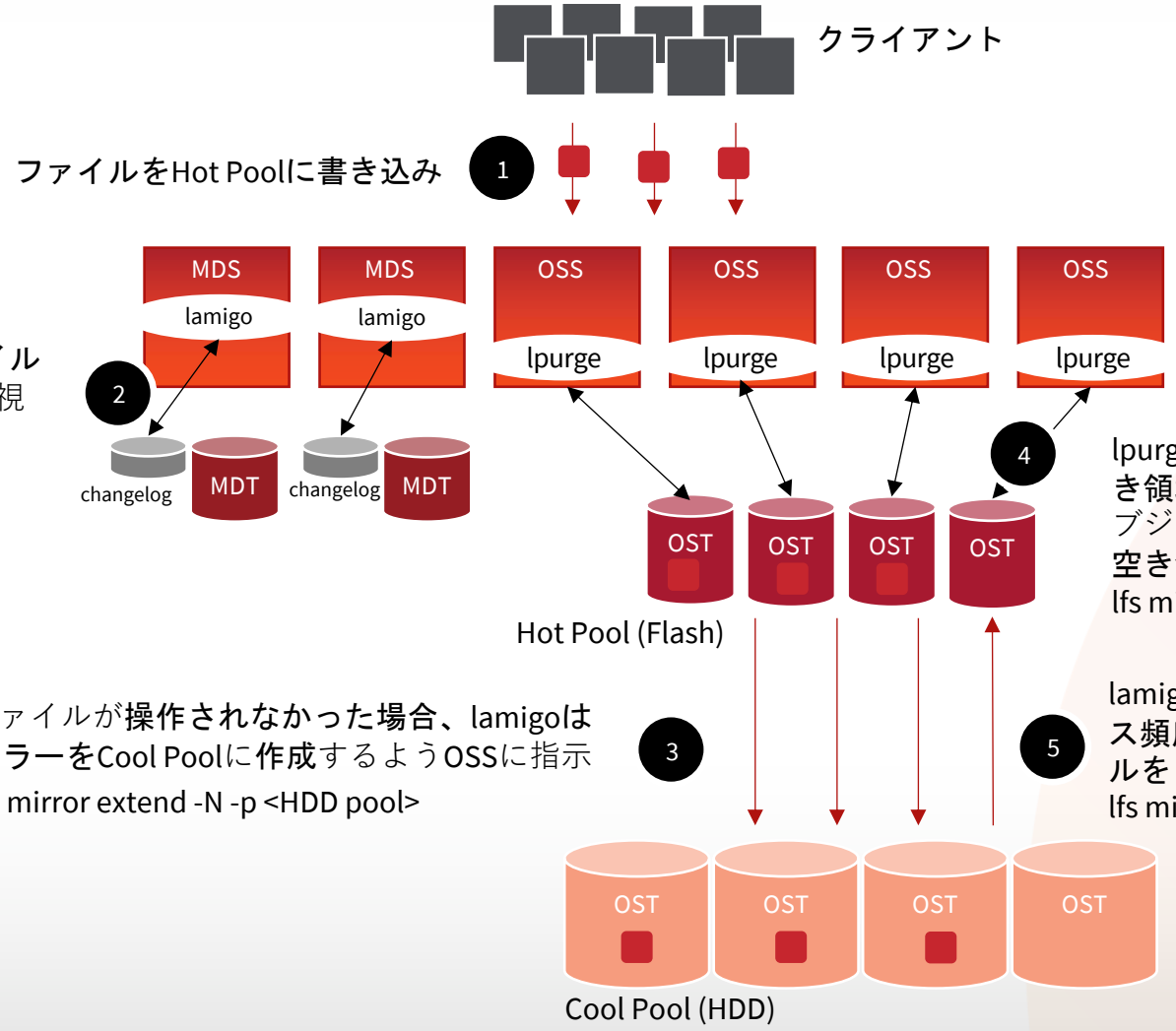
- EXAScalerはOSTの集まりをPoolとして管理可能
- Flash層(Hot Pool)からHDD層(Cool Pool)へのアクセスが少ないファイルの自動マイグレーションを提供
- Cool Pool上のアクセス頻度が高いファイルをHot Poolに自動マイグレーションを提供
  - ファイルヒートポリシー
- LustreのFile Level Replication(FLR)機能がベース
- 使用量の増減に合わせて、HDD容量とFlash性能を効率的に使用
- ユーザはHot Poolsの構成を意識することなく利用可能



EXAScalerは**ファイルヒートポリシー**に基づき、継続的にデータのコピーをFlash層に作成/削除し、最もアクセスされるファイルをFlash層に維持します

※ OST(Object Storage Target)：ファイルデータを格納

EXA6 Hot Poolsの実装



lamigoはHot Poolのファイルのクローズイベントを監視
<min-age>時間の間待機

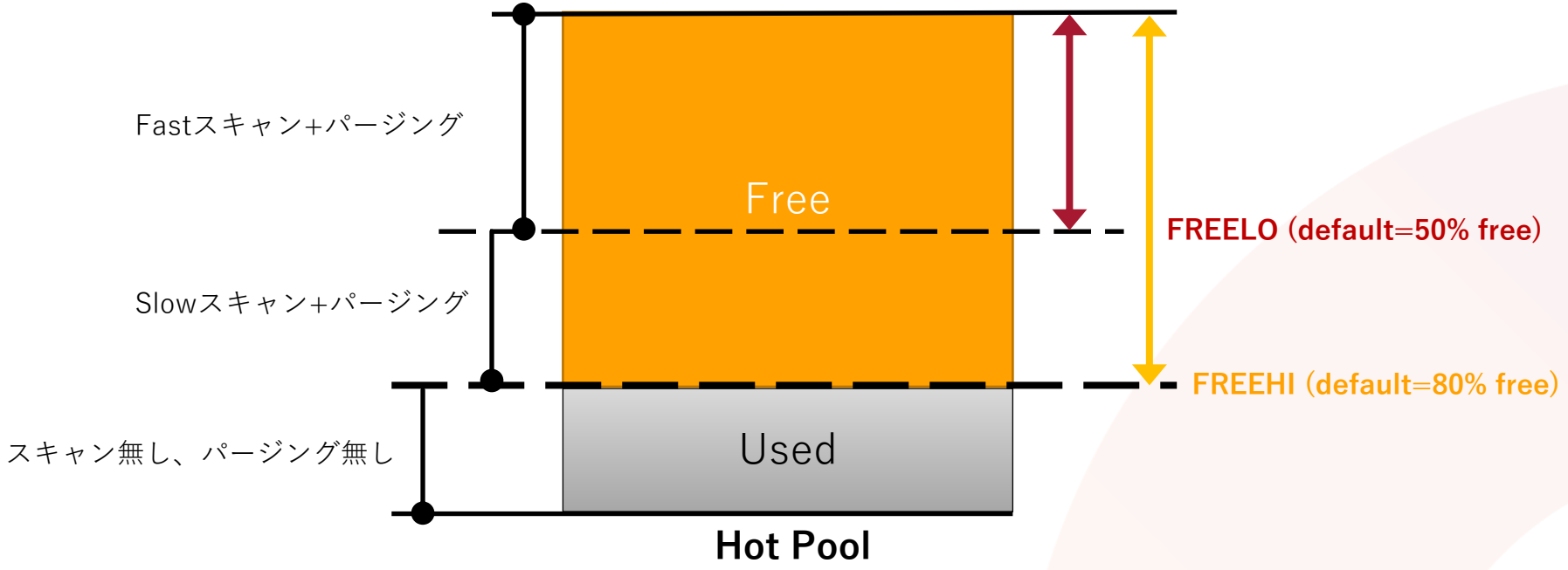
lpurgeはパージするオブジェクトをスキャンHot Poolの空き領域がfreeloの設定値 (%age) を下回ったらHot Poolのオブジェクトを削除
空き領域がfreehiの設定値(%age)を確保するまで継続
lfs mirror split -d <fast pool>

<min-age>の間ファイルが操作されなかった場合、lamigoはファイルのミラーをCool Poolに作成するようOSSに指示
lfs mirror extend -N -p <HDD pool>

lamigoはアクセス頻度が高いread-onlyファイルとアクセス頻度は高いが一定期間書き込みが無いread-writeファイルをミラーバックします
lfs mirror extend -N -d <fast pool>

Hot Poolsのスペース管理

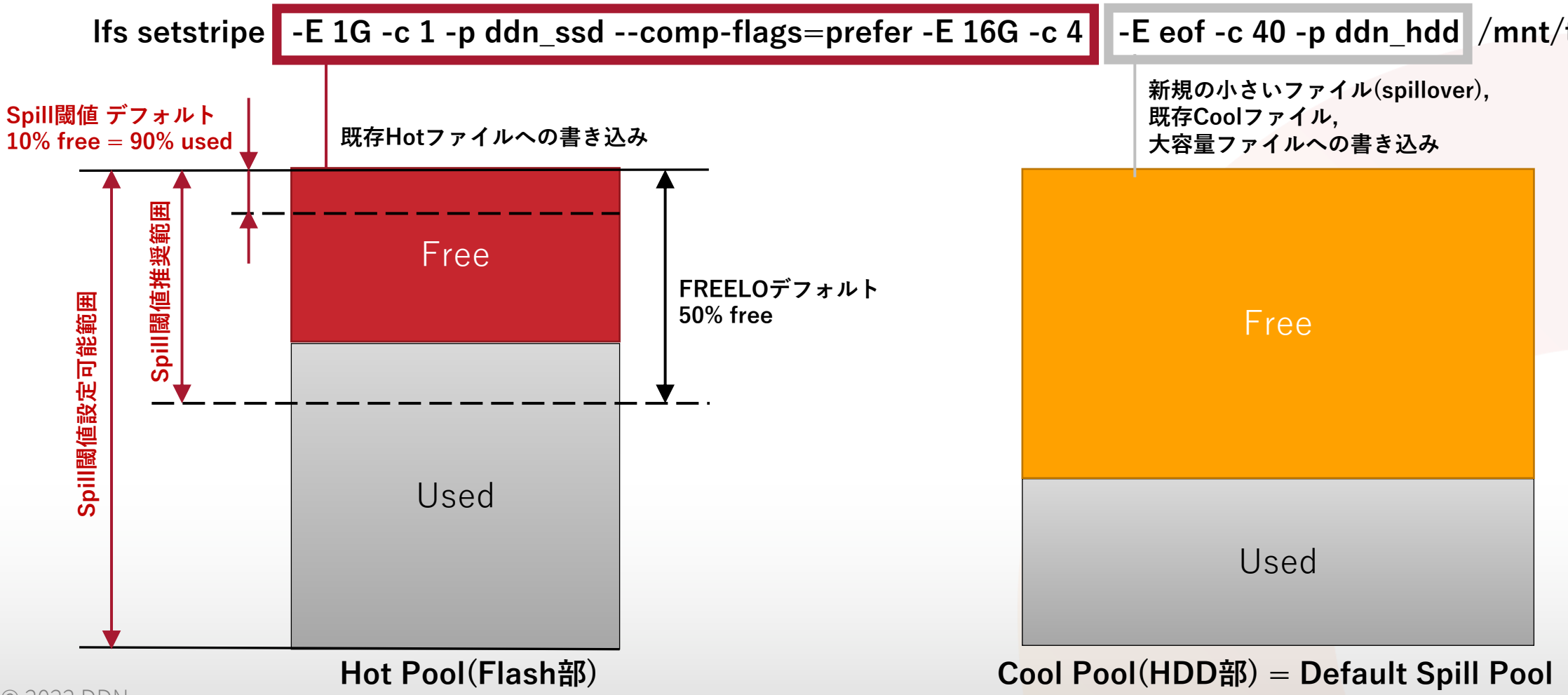
- 閾値FREEHIとFREELOによるHot Pool OST(Flash部)からのファイルパージング



  - FREEHIとFREELOは、消費されたスペースの量では無く、Hot Pool(Flash部)のフリースペースの残量を示します
    - FREEHIはシステムがHot Pool(Flash部)からCool Pool(HDD部)へのファイルミラーリングを開始（同時にHot Poolからファイルをパージ)するフリースペースのレベルを示します
    - FREELOはスキャンプロセスの負荷が高くなるフリースペースのレベルを示します

Hot Pools容量逼迫時の仕組み

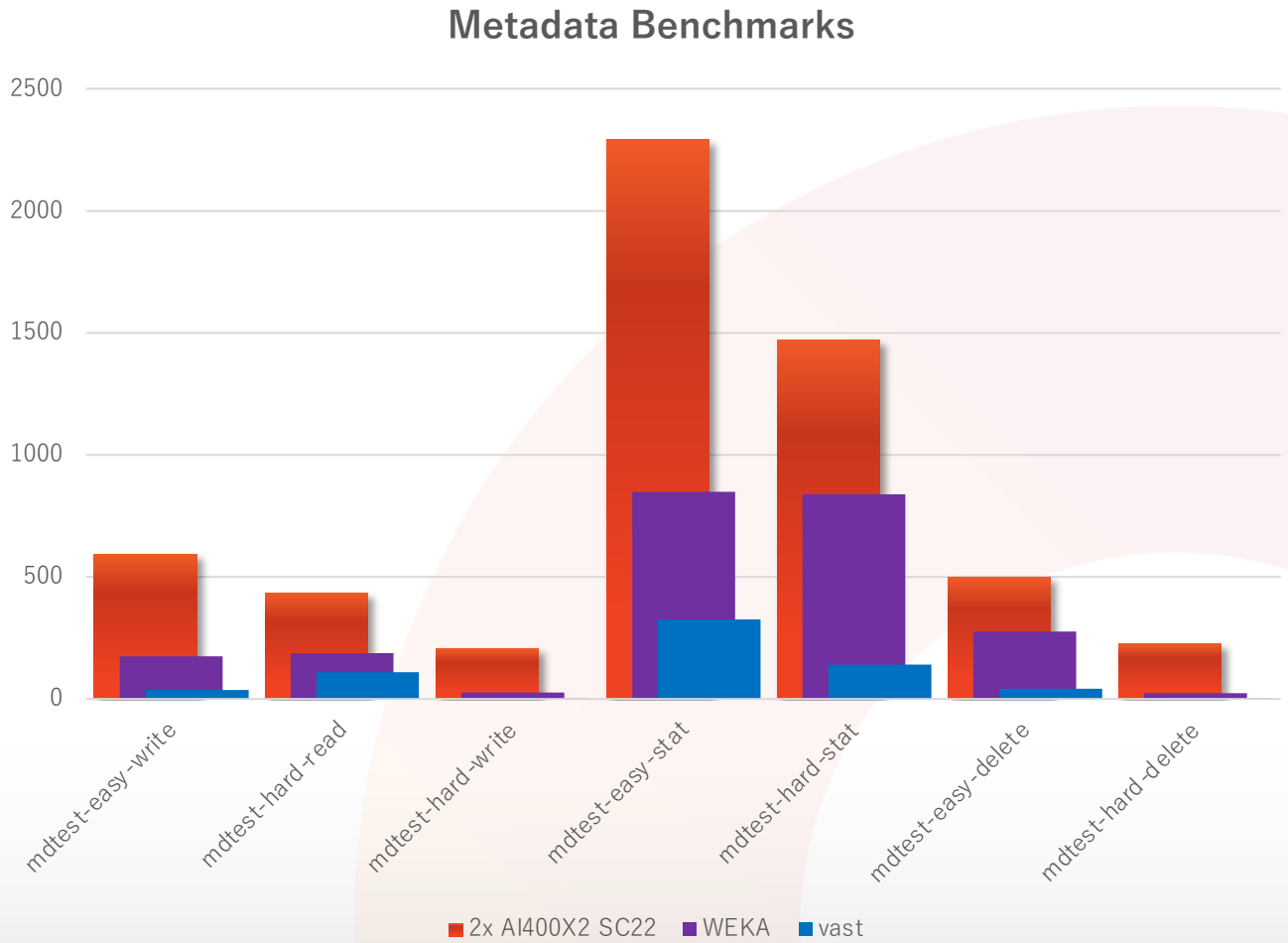
- Hot Pool(Flash部)のSpill閾値とSpill Poolによる容量枯渇回避



EXAScalerの高いメタデータ性能

File statオペレーションの改善

EXAScalerはFile statオペレーションを改善し、他社に比べて2倍以上の性能を達成しました。
stat ahead機能により、連続したstatオペレーションをEXAScalerクライアントが検知し、複数のstatオペレーションを纏めて発行して性能向上を測っています



https://io500.org/submissions/view/30
https://io500.org/submissions/view/555

JLUG 2022 – 2022年12月9日(金)

- 今年はハイブリッド開催
- 詳細/お申し込みは今すぐ！
  https://www.jlug.info/


ddn