DDN Update @Gfarm WorkShop2022

DataDirect Networks Japan
橋爪信明
2022/2/10

アジェンダ

- 2021実績(Lustreのみ)
- HW新製品紹介
- EXA6新機能紹介

2021年実績
(Lustreのみ)

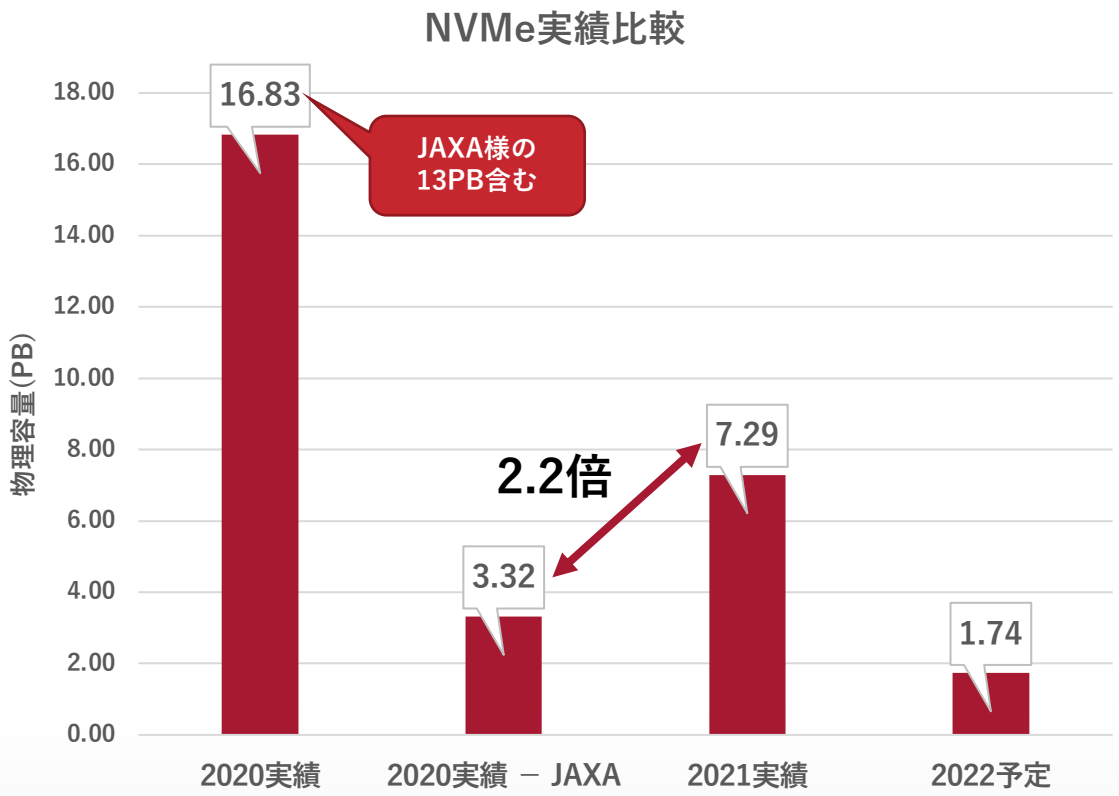
2021年導入実績

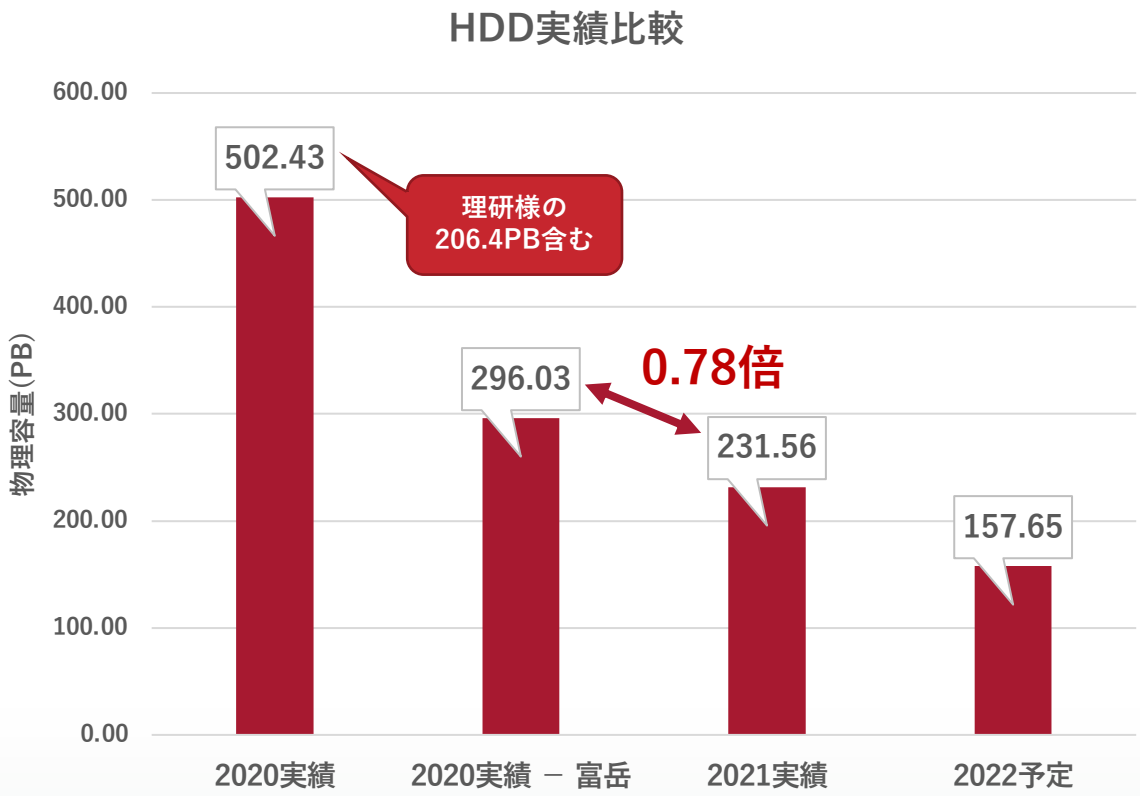
お客様	NVMe物理容量 (TB)	HDD物理容量 (PB)	ファイルシステム
某省庁		43.4	EXAScaler
東京大学HGC	307.2	36.1	EXAScaler
東京大学情報基盤センター mdx	2027.52	35.784	EXAScaler
東京大学情報基盤センター Westeria/BDEC-01	1413.12	34.176	FEFS
大阪大学サイバーメディアセンター SQUID	1536	26.88	EXAScaler
産業技術総合研究所(AIST) ABCI2.0	529.9	14.4	EXAScaler
国立環境研究所(NIES) GOSAT/GOSAT-2 プロジェクト		13.87	EXAScaler
理化学研究所Spring-8		5.6	EXAScaler
国立遺伝学研究所(NIG) DDBJ		4.816	EXAScaler
名古屋大学宇宙地球環境研究所(ISEE)		4.14	EXAScaler
某機構		3.1	EXAScaler
情報通信研究機構(NICT)		3.09	EXAScaler
マクロジェン・ジャパン		3.0	EXAScaler
某民間企業	161.28	1.62	EXAScaler
某研究所		1.58	EXAScaler
北陸先端科学技術大学院大学(JAIST)	322.5		EXAScaler
理化学研究所R-CCS	307.2		EXAScaler
某民間企業	276.48		EXAScaler
某民間企業	161.28		EXAScaler
某民間企業	161.28		EXAScaler
某民間企業	88.32		EXAScaler
合計	**7.29PB**	**231.6PB**	

2022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
某機構		67.0	EXAScaler
東京大学情報基盤センターIpomoea-01		34.4	EXAScaler
某機構		19.58	EXAScaler
情報通信研究機構(NICT)	88	12.9	EXAScaler
東京大学HGC		9.0	EXAScaler
某大学		6.00	EXAScaler
某民間企業		2.56	EXAScaler
某民間企業		1.8	EXAScaler
某機構		1.58	EXAScaler
沖縄科学技術大学院大学(OIST)		1.5	EXAScaler
某民間企業		1.2	EXAScaler
某研究機関	1000	0.13	EXAScaler
某民間企業	737.28		EXAScaler
合計	1.73PB	157.7PB	

2020年との比較





HW新製品紹介

New “NVX2” 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	~46GB/s, 1.5M IOP/s	~90GB/s, 3M IOP/s
HDD	No Support	2021 Q3以降：Max 360 Drive（4x SAS3 90Slot Enc） 2022 Q3以降：Max 900 Drive（10x SAS4 90Slot Enc）
HDD Performance	No Support	2021 Q3以降：~40GB/s 2022 Q3以降：~90GB/s
Connectivity	HDR IB（200Gb/100Gb）（4） Or 100/200 GbE（4）	HDR IB（200Gb/100Gb）（8） Or 100/200 GbE（8）

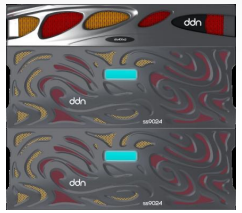
ES400NVX2 SAS-3 Expansion Option (2021～2022 1H)

ES400NVX2 SAS3

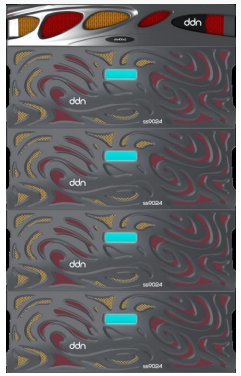
Platform	ES400NVX2 SAS-3
NVMe Slots	24
Capacity max (raw)	368 TB (15.36TB NVMe)
SAS Chassis	4
Slots	360
Capacity max (raw)	6.4 PB (18TB HDD)

100% NVME without use of SAS expansion is also supported

Planned ES400NVX2 SAS-4 Expansion Options (2022 Q3以降)

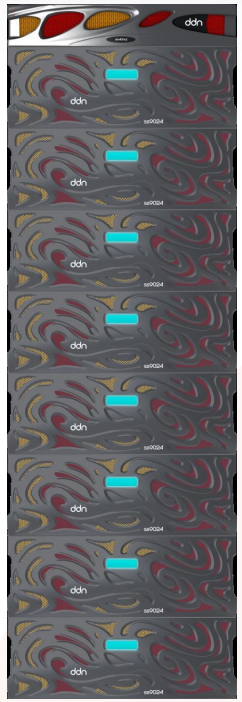


NVMe Slots	24
Capacity	732 TB
SS9024 qty	2
Slots	180
Capacity	3.2 PB



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

NVMe Slots	24
Capacity	732 TB
SS9024 qty	6
Slots	540
Capacity	9.6 PB

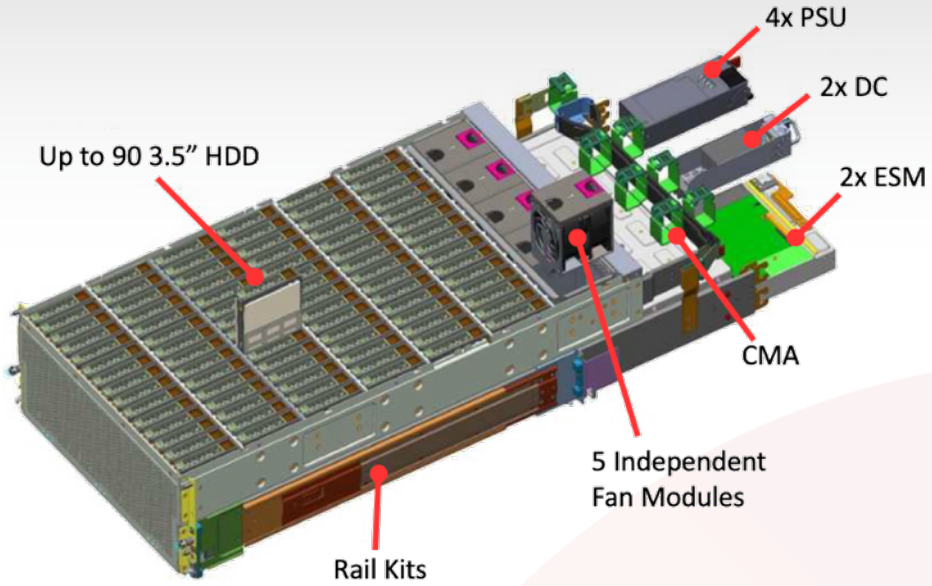


NVMe Slots	24
Capacity	732 TB
SS9024 qty	8
Slots	720
Capacity	12.8 PB

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.

SS9024 ENCLOSURE



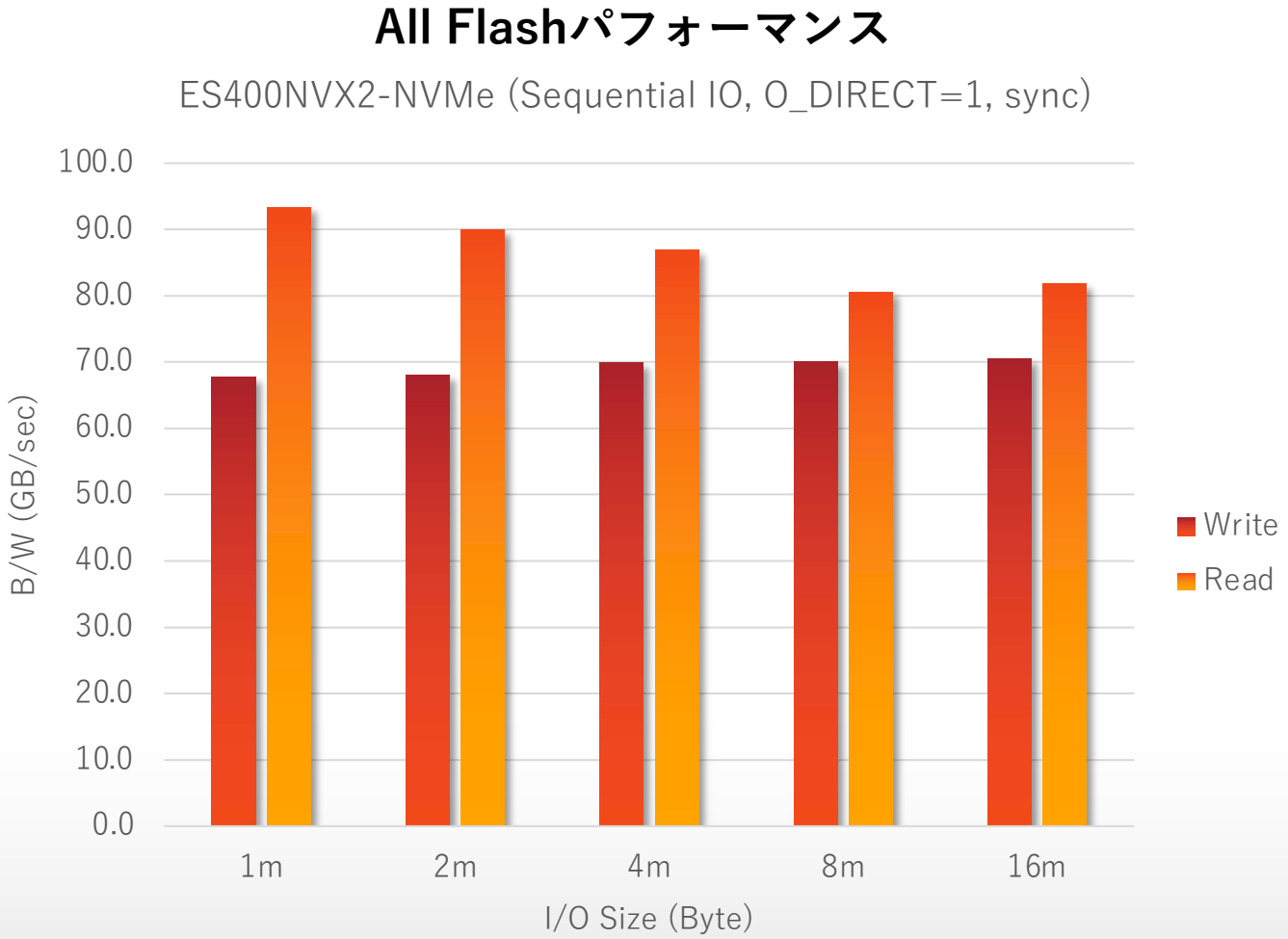
SAS4サポートエンクロージャ

SS9024 ENCLOSURE SPECIFICATIONS	
Chassis	Redundant 4U
Disk Slots	90 top accessible 3.5” drive slots, SAS-3 (12G)
PSU/Cooling	4 PSUs (2+2 redundant), 5 Independent Fans
Monitoring	LCD displays for providing system status Per drive activity LEDs
IO Modules	2x IO Modules. SAS-4 (24G) 4x 4 lane SAS 24Gb Mini SAS HD ports on each IO Module

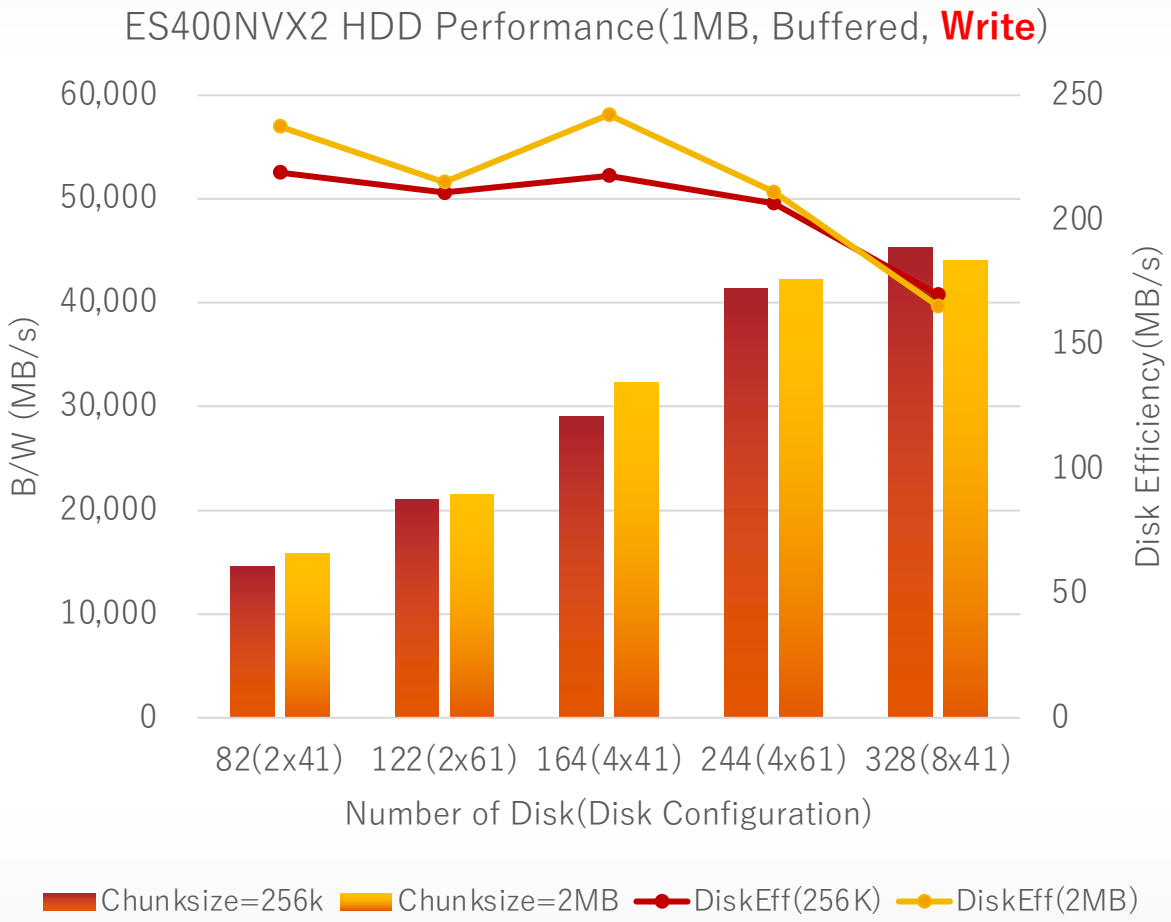
Hyper Fast AI & HPC Data Storage

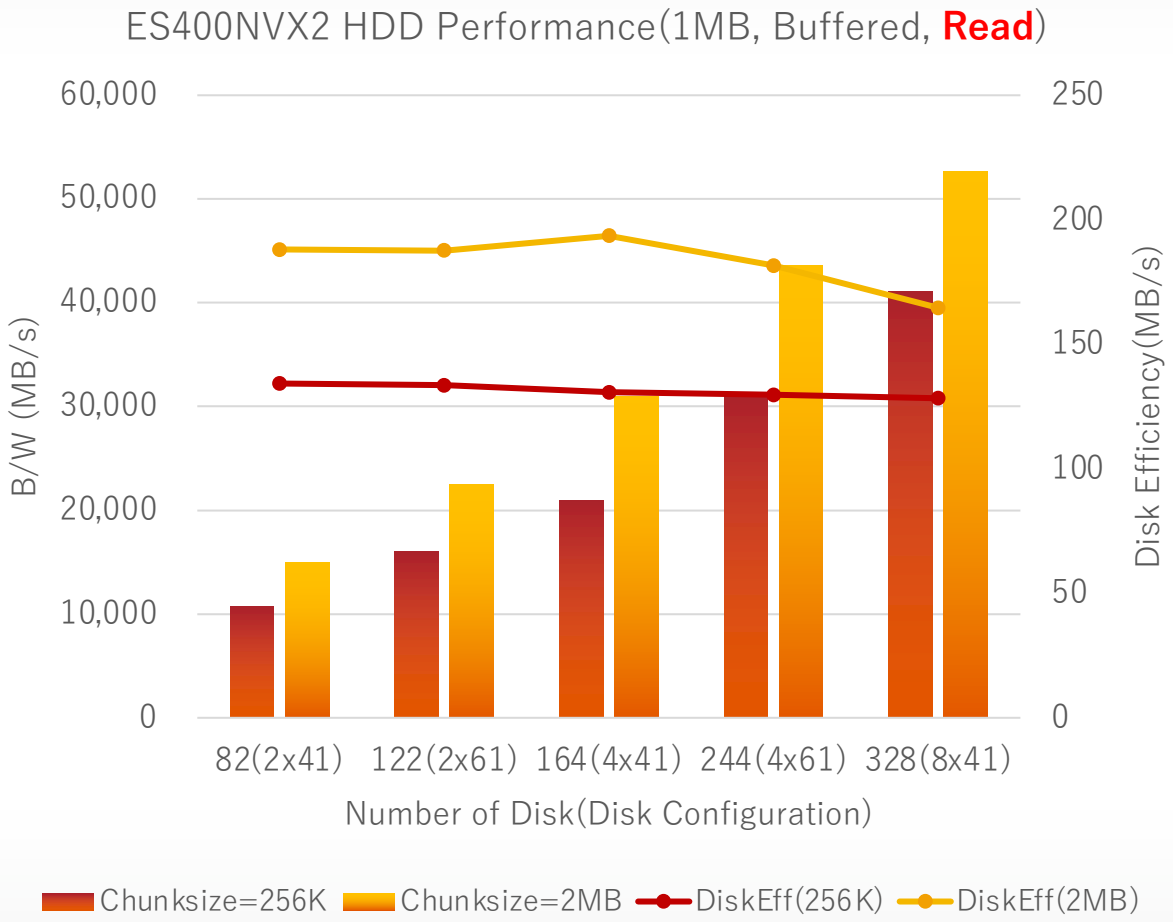
世界最速ストレージ

- Over 90GB/s for reads
- Over 65GB/s for writes



ES400NVX2 HDD Performance (HDD/w SAS3)





Chunksize : RAIDのChunksize
DiskEff : Disk 1本あたりの性能
WriteはParity 2本分のWrite性能を付与するため、性能の1.2倍をDisk本数で除した値
Readは性能をDisk本数で除した値

EXA6新機能

EXA6の主な新機能

Lustre2.14をベースとした新EXAScalerバージョンEXA6をリリース

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

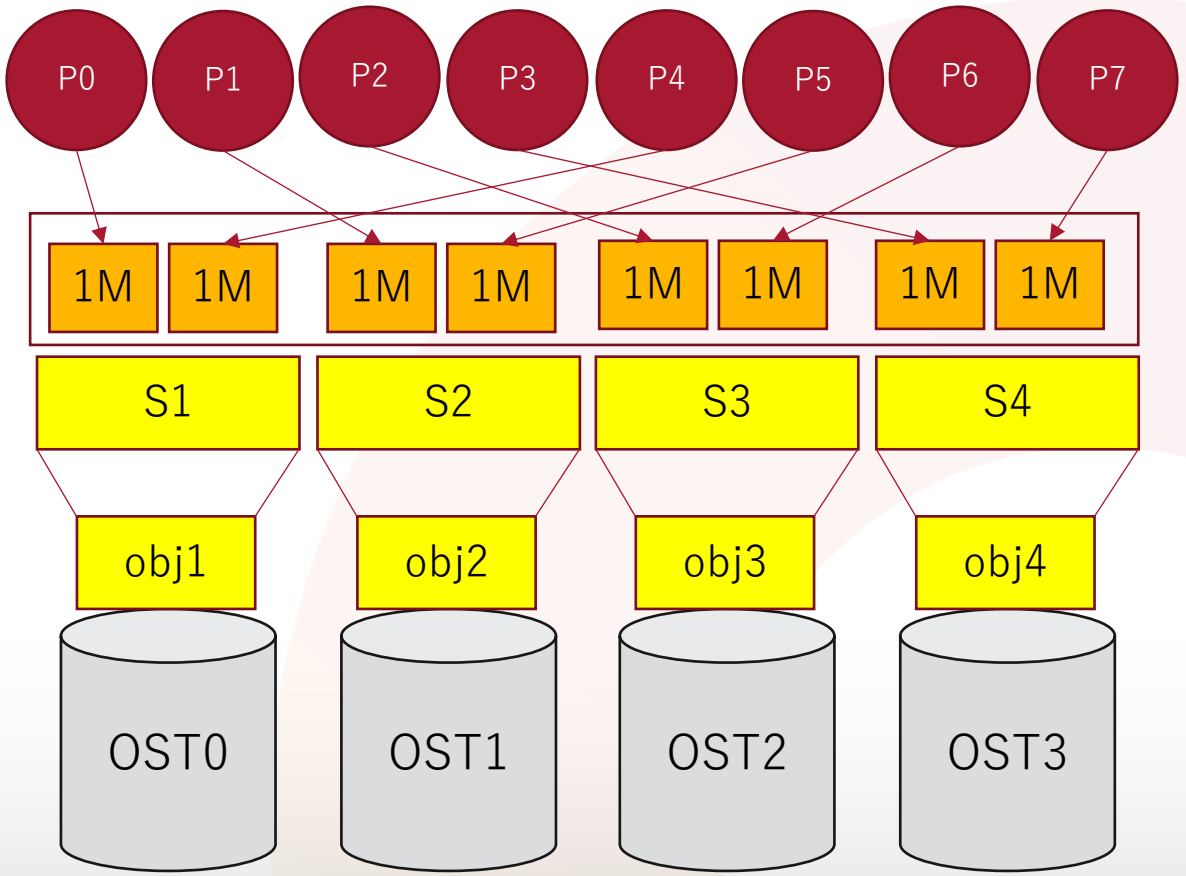
- **OST Pool Quota**
  １つのファイルシステムに混在する異なるデバイス(HDD, NVMe)毎にそれぞれOST Poolを作成して、異なるQuota設定可能
- **MDT Auto Balancing**
  同一ディレクトリ内でinode数が設定値を超えた時点から複数のMDTを自動的に利用

Lustre Striping

```
root@ubuntu1804-1:~# lfs setstripe -c -1 /ai200x1/shared-file
root@ubuntu1804-1:~# lfs getstripe /ai200x1/shared-file
/ai200x1/sharedfile
lmm_stripe_count:  4
lmm_stripe_size:   1048576
lmm_pattern:       raid0
lmm_layout_gen:    0
lmm_stripe_offset: 1
obdidx objid objid group
     1             2           0x2                0
     3             2           0x2                0
     0             2           0x2                0
     2             2           0x2                0
```

("-c -1"は全てのOST)

shared-file



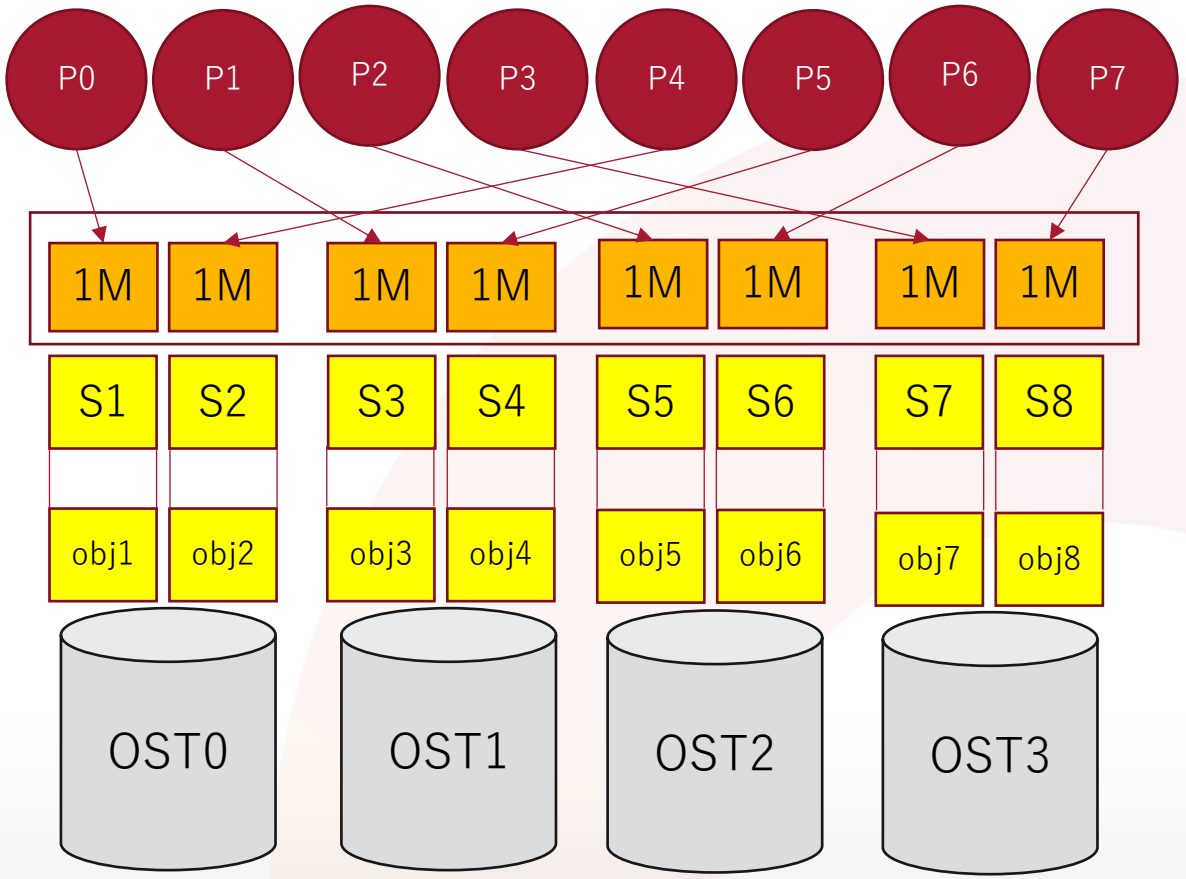
プロセス数 > ストライプ数の場合 OST オブジェクトに対する競合が発生

※ OST : Object Storage Target

Lustre OverStriping

```
root@ubuntu1804-1:~# lfs setstripe -C 8 /ai200x1/shared-file-OS
root@ubuntu1804-1:~# lfs getstripe /ai200x1/shared-file-OS
/ai200x1/shared-file-OS
lmm_stripe_count:  8
lmm_stripe_size:   1048576
lmm_pattern:       raid0,overstriped
lmm_layout_gen:    0
lmm_stripe_offset: 0
obdidx objid objid group
     0             4           0x4                0
     2             4           0x4                0
     1             4           0x4                0
     3             4           0x4                0
     0             5           0x5                0
     2             5           0x5                0
     1             5           0x5                0
     3             5           0x5                0
```

shared-file-OS

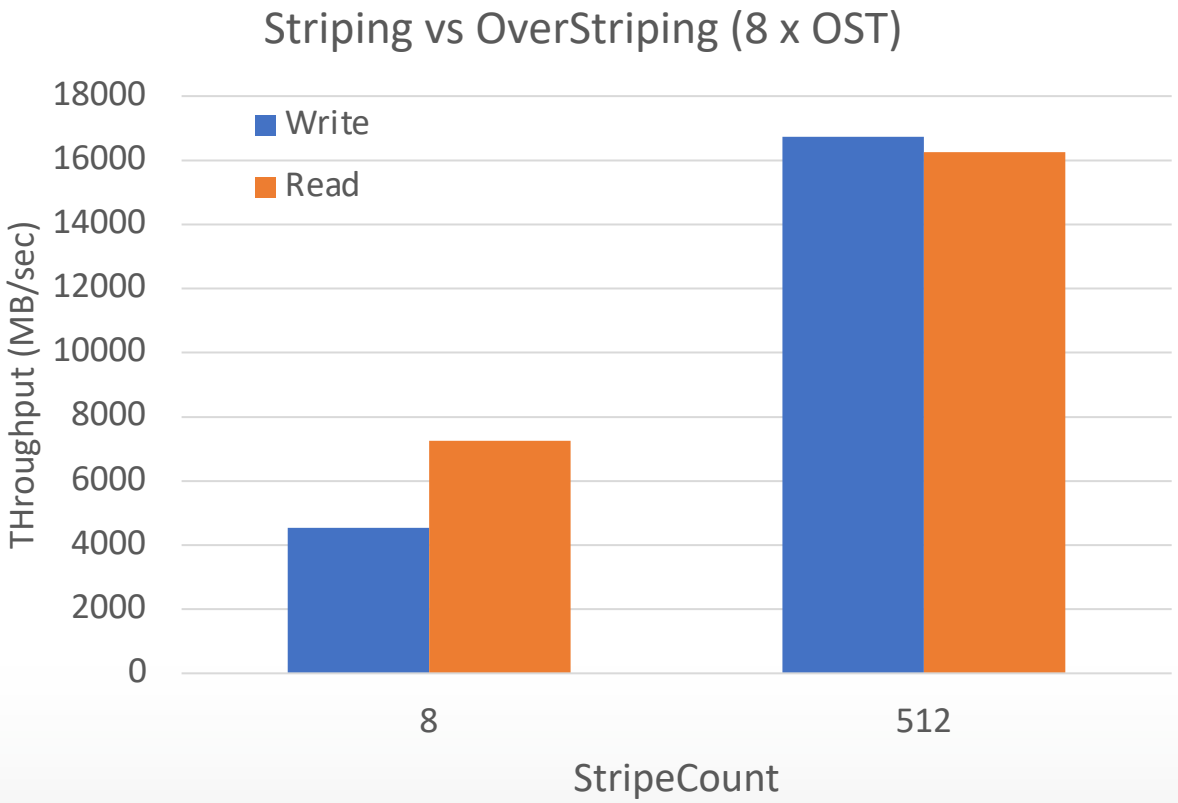


OverStriping機能によってOST数以上のストライプ数を設定可能になりオブジェクトアクセスにおける競合を排除

※ OST : Object Storage Target

Striping vs Over Striping性能比較

１つのファイルに複数プロセスでアクセスした場合の性能比較

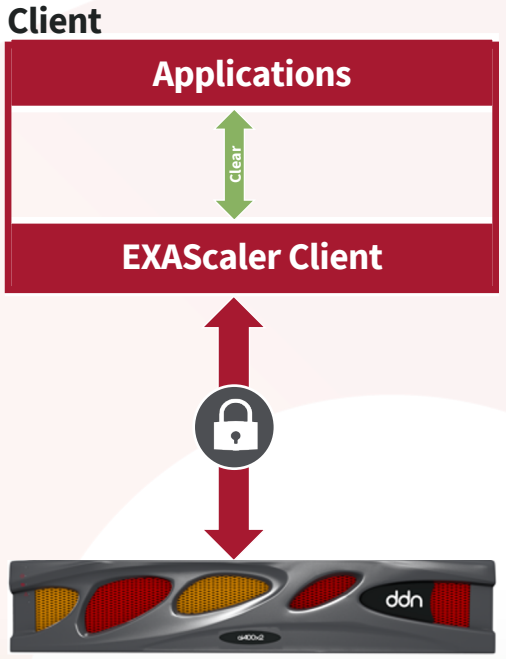


- ES7990(160 x HDD, 2 x OSS, 8 x OST)
- 32クライアント, 512プロセス
- 1MB, Single Shared File

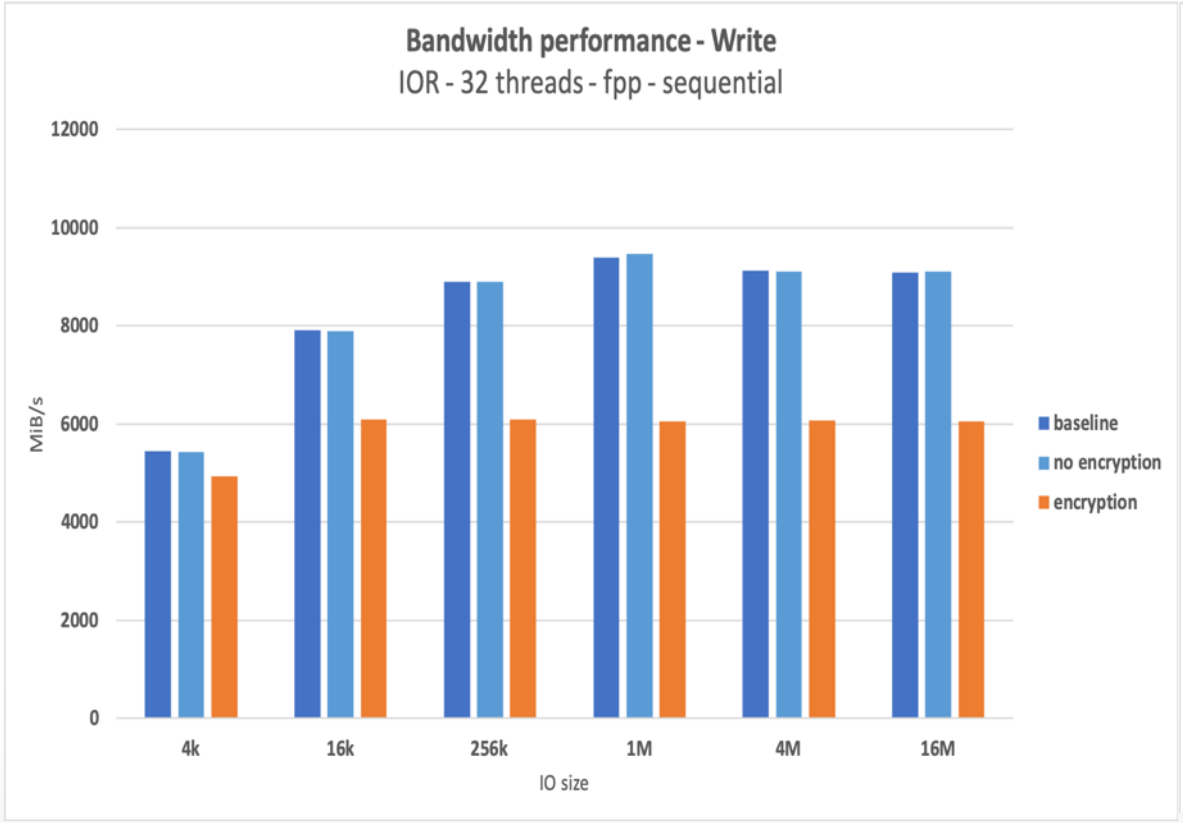
  # ior/src/ior -w -r -C -g -i 3 -vv -s 13000 -b 1m -t 1m -a POSIX -e
- ストライプカウント8と512で比較

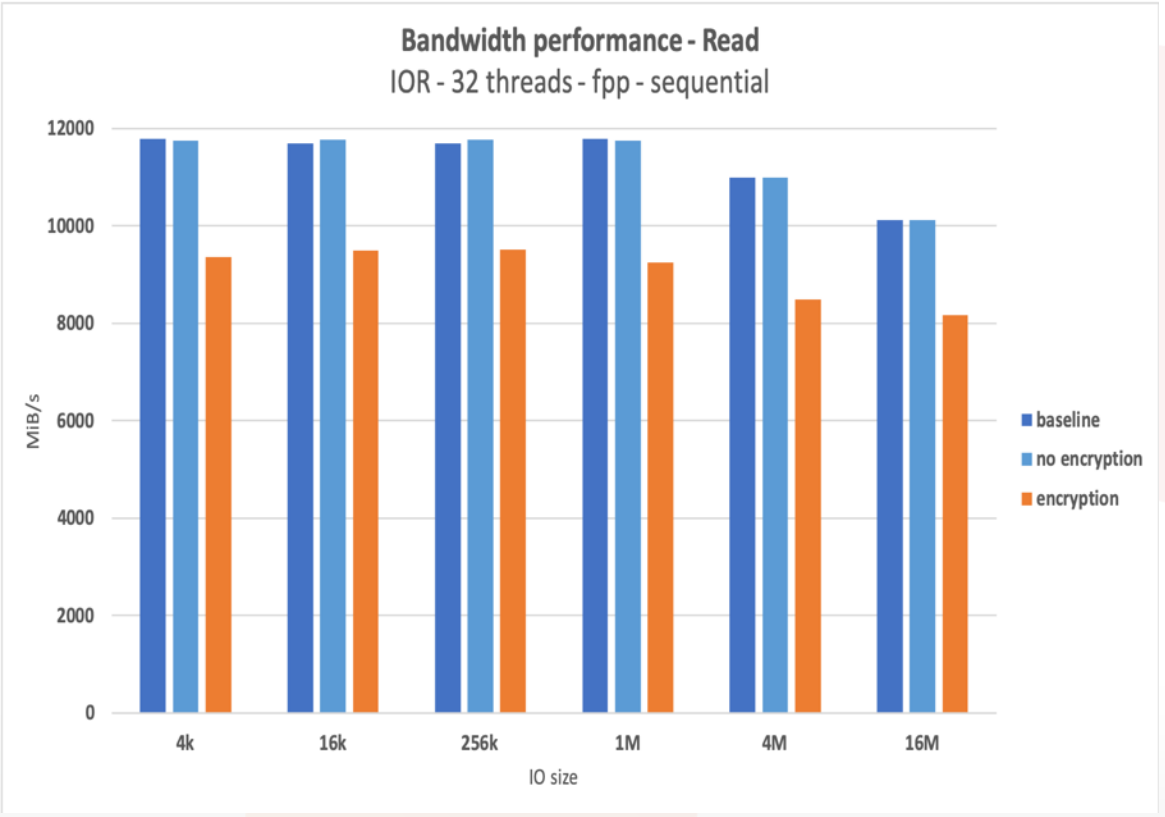
Lustreにおける暗号化

- ユースケース:
  - 各ユーザの特定のディレクトリに含まれるファイルに対する機密性を提供
- ゴール:
  - クライアントおよびサーバ間でデータ保護
  - 保存データの保護
- ソリューション
  - fscrypt kernel APIに準拠
    - ext4, F2FS, and UBIFSにて使用されているAPI
    - 基本原則: Page Cacheに含まれるPageはクリアテキストデータを含む
  - fscrypt ユーザスペースツールを活用
- Lustreにおける実装
  - 暗号化の方法
    - Lustre Clientにて透過的にWrite時に暗号化、Read時に復号化を実施
  - ディレクトリにおけるポリシーの適応方法
    - fscryptユーザスペースを使った新しいIOCTLのサポート
    - アトミックな暗号化コンテキストの処理



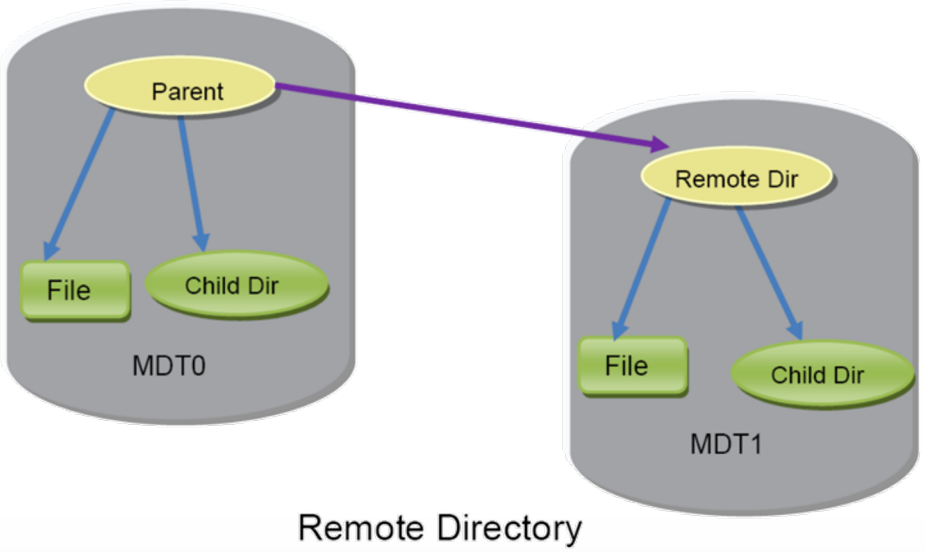
Lustre Client Encryption – bandwidth performance





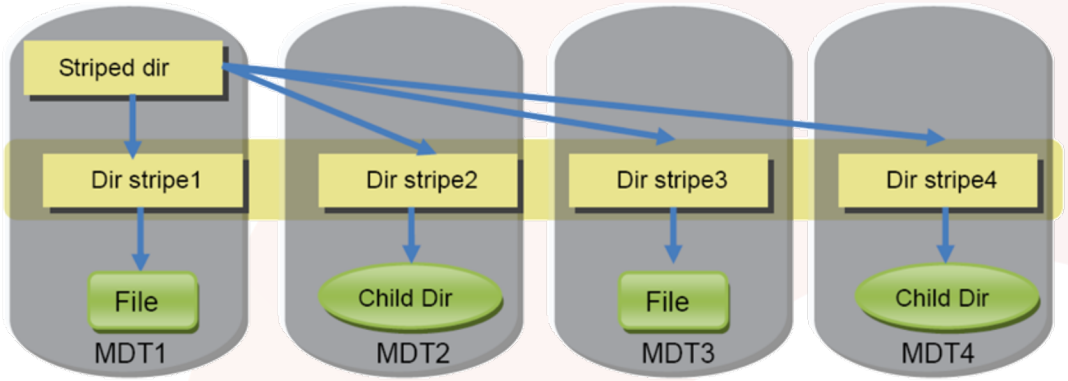
従来のDNE(Distributed Namespace Environment)

- 2.14以前のLustreは２つのタイプのDNEをサポート
  - Remote Directory(DNE1)とStriped Directory(DNE2)
  - いずれも動的に設定できDNE1とDNE2の混在も可能



Remote Directory

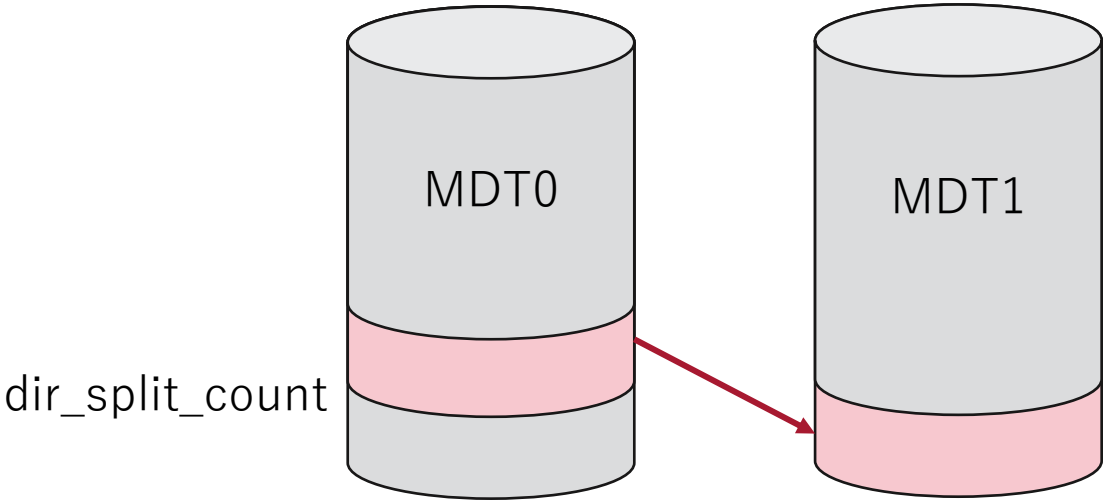
 # lfs mkdir –i 0 /ai200x1/mdt0
 # lfs mkdir –i 1 /ai200x1/mdt1



Striped Directory

 # lfs mkdir –c 4 /ai200x1/striped-dir
 # lfs mkdir –c 4 -D /ai200x1/ striped-dir

DNE Auto Rebalancing (2.14で追加)




```
[root@vexa01 ~]# lctl get_param mdt.*.enable_dir_auto_split
mdt.*.dir_split_count mdt.*.dir_split_delta
mdt.ai200x1-MDT0000.enable_dir_auto_split=1
mdt.ai200x1-MDT0000.dir_split_count=50000
mdt.ai200x1-MDT0000.dir_split_delta=2
```

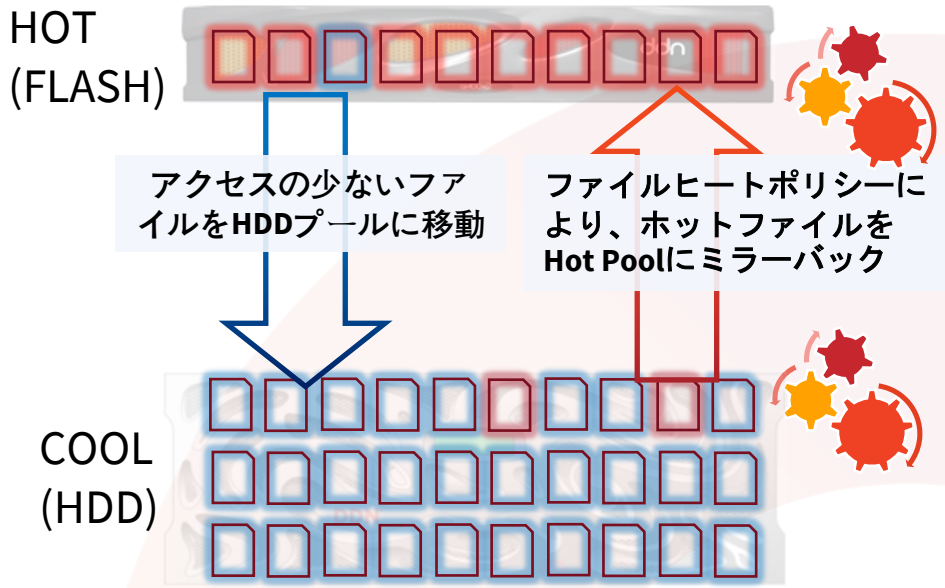
enable_dir_auto_split=1(Auto rebalancing)が有効の場合、
同一ディレクトリ内でinode数がdir_split_countを超えた時点から
dir_split_deltaに基づき複数のMDTに自動的に分散される

閾値を設定することでディレクトリのサイズが大きくなる前に複数のMDTに自動的に分散し性能劣化を防ぐ

Hot Pools

ストレージプール間のTieringを実現

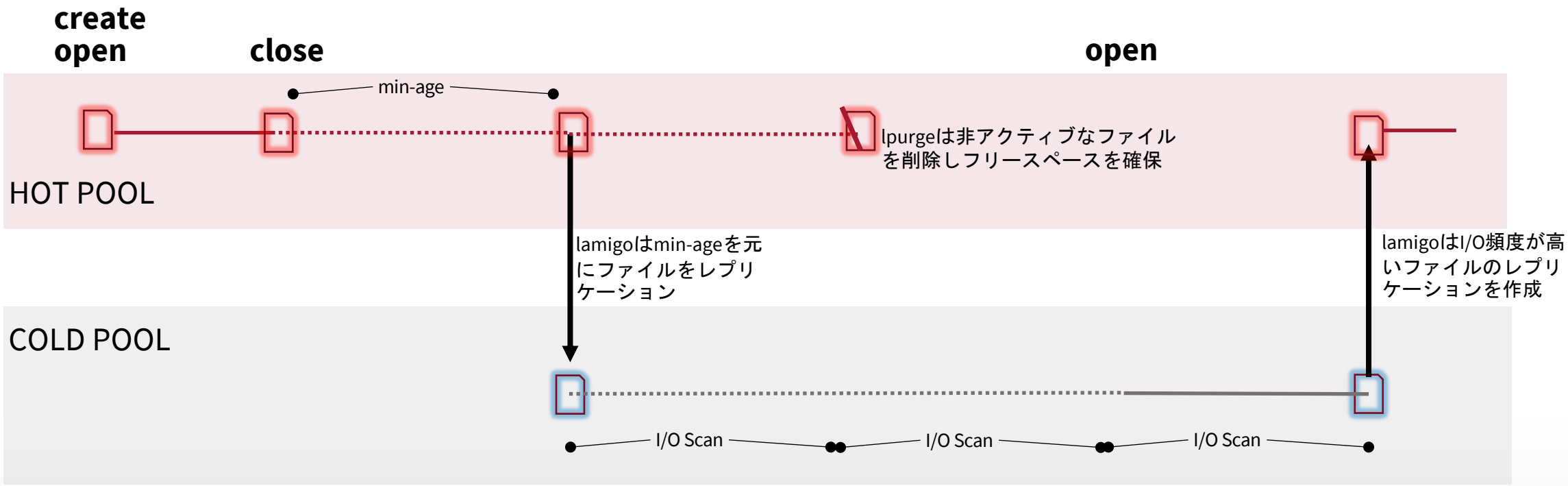
- ユーザ透過的なFlash層(Hot Pool)からHDD層(Cold Pool)へのアクセスが少ないファイルの自動マイグレーション
- File Level Replication(FLR)機能がベース
- 使用量の増加に合わせて、HDD容量、Flash性能のいずれかもしくは両方の層を個別にスケールできます
- ユーザはHot Poolsの機能を意識することなくメリットを享受できます
- 新たにスパースファイルをHot Pools管理下でサポート



EXAScalerは**ファイルヒート**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

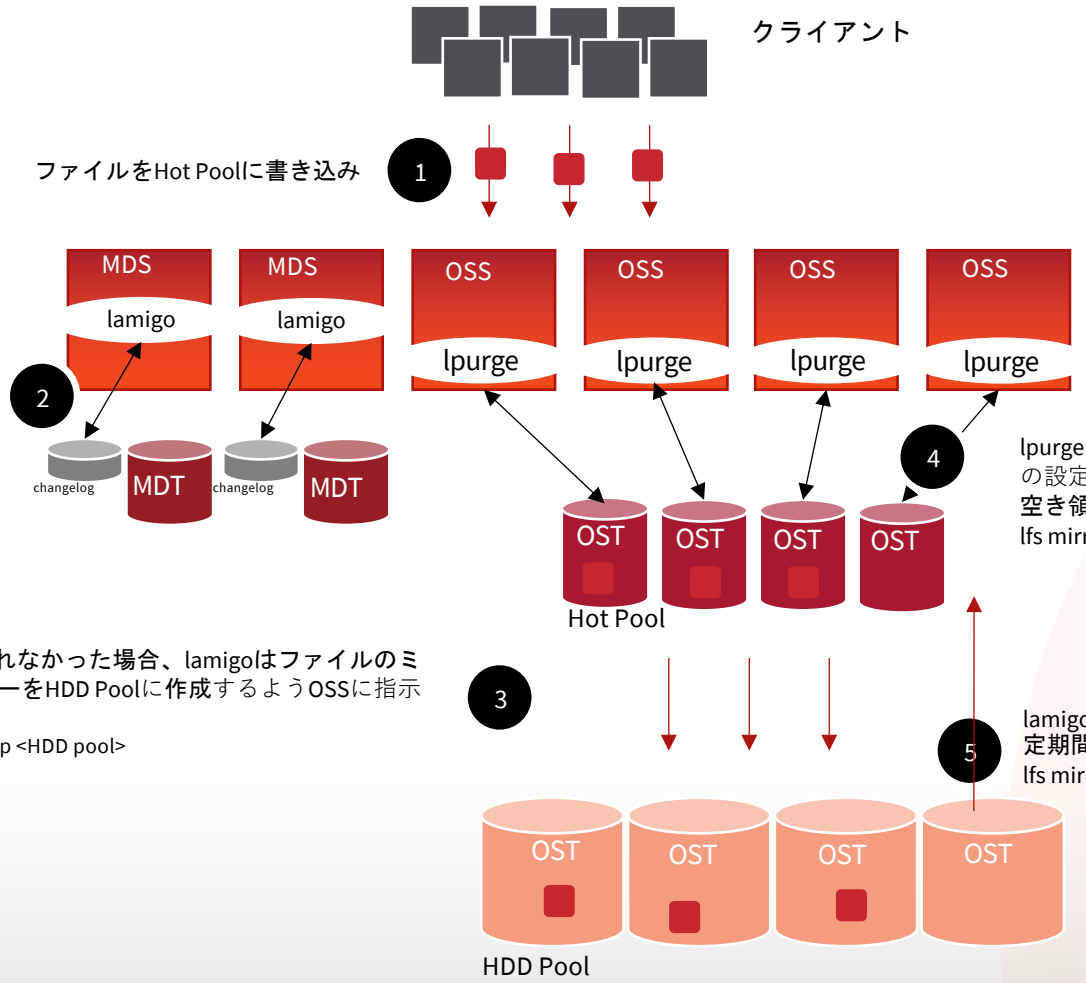
EXA6 Hot Poolsの実装

Hot Poolsによるファイル管理のタイムライン



※lamigo, lpurgeはHot Poolsのコマンド名

EXA6 Hot Poolsの実装

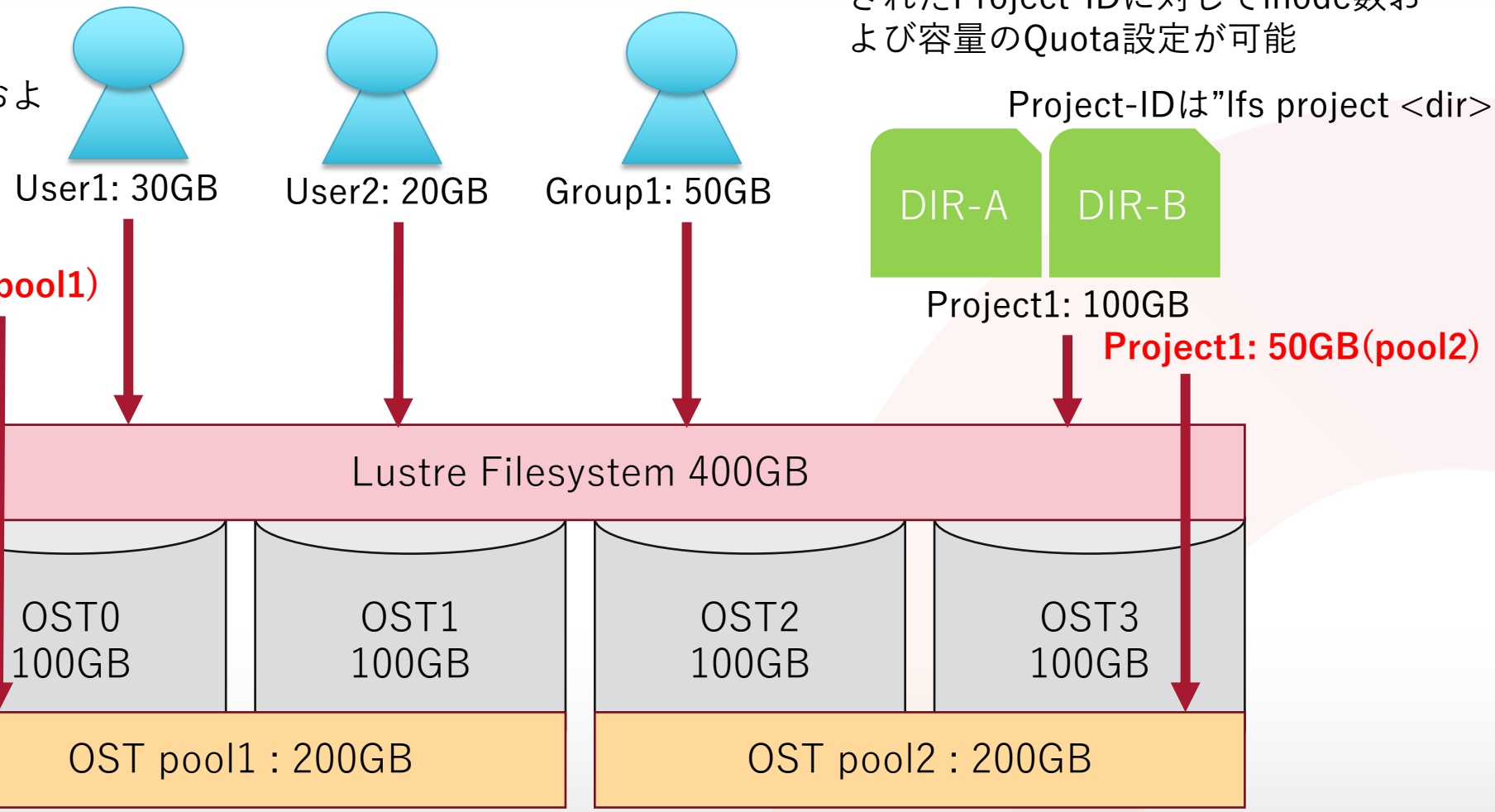


OST Pool Quota

特定のディレクトリファイルに付与されたProject-IDに対してinode数および容量のQuota設定が可能

UID, GIDに対してinode数および容量の Quota設定が可能

Project-IDは”lfs project <dir>”で設定



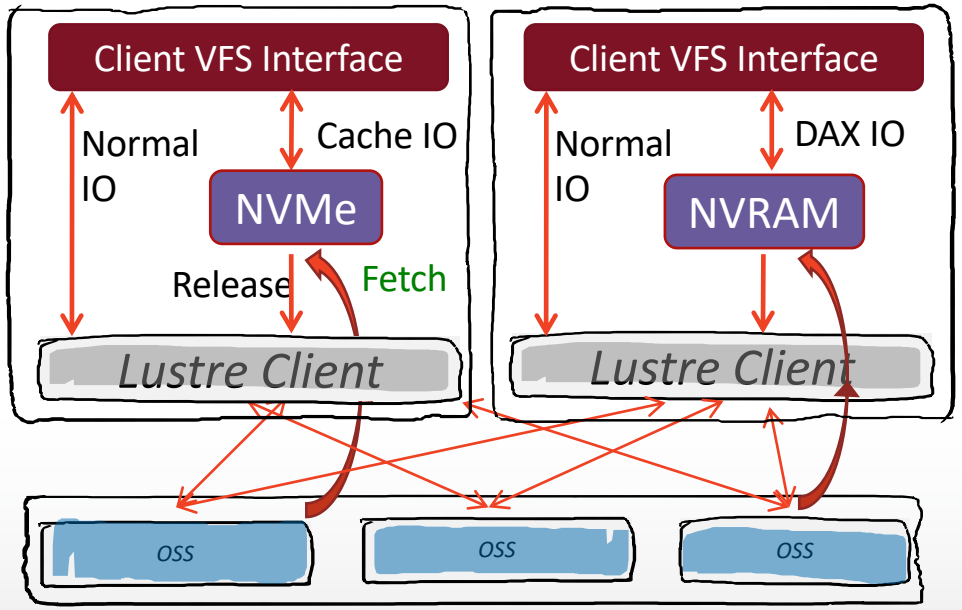
事前に作成されたOST Poolに対してUID, GIDに対する容量のQuota設定が可能

OST Poolは”lctl pool_create/pool_add“で動的に作成

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

Hot Nodes

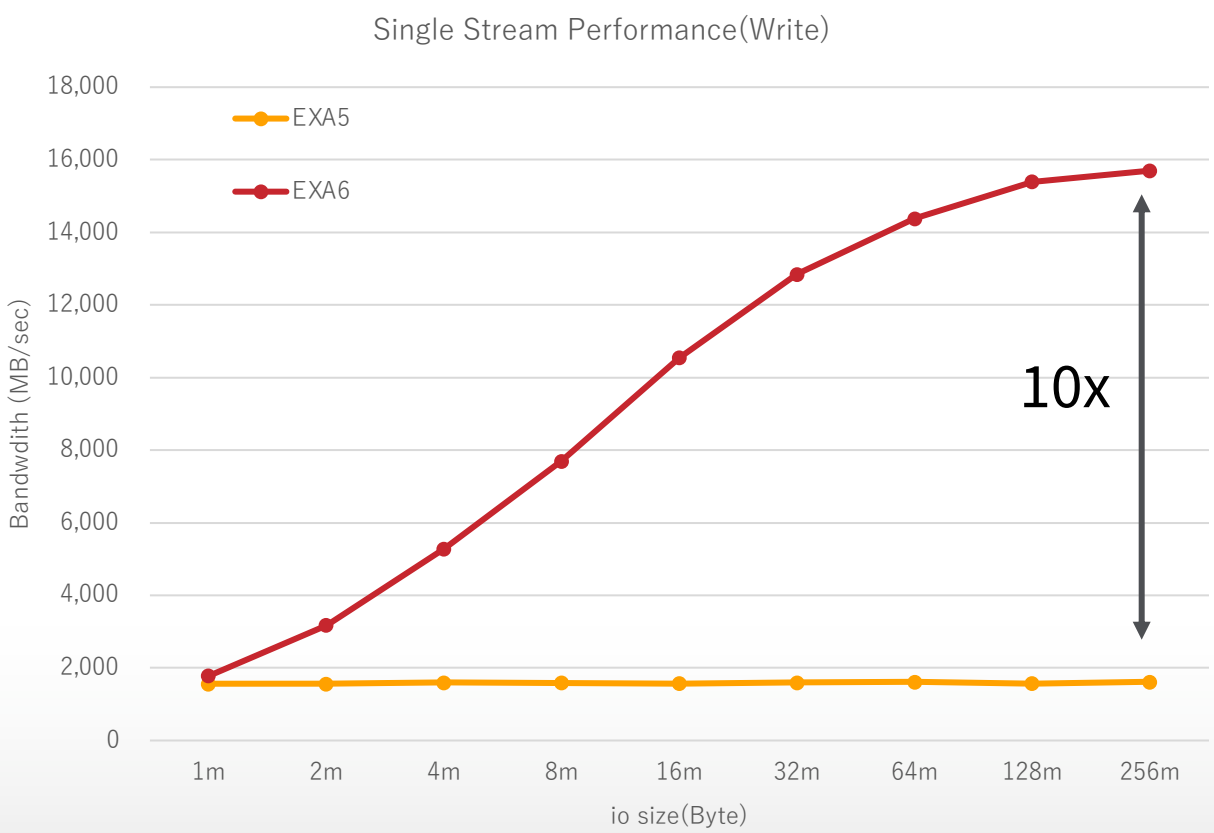
- クライアントのローカルストレージとのインテグレーション
  - レイテンシの削減, ネットワークトラフィックの削減
  - DAX(Direct Access)可能なNVDIMMデバイスとインテグレーション
- ReadおよびWriteの透過的なキャッシュを提供
  - Read Onlyキャッシュもサポート
- ワークフロー全体の最適化に活用
  - 同一クライアントでのデータの再利用
  - データの先読みにて、IOと計算利用のネットワークリソースの分散化

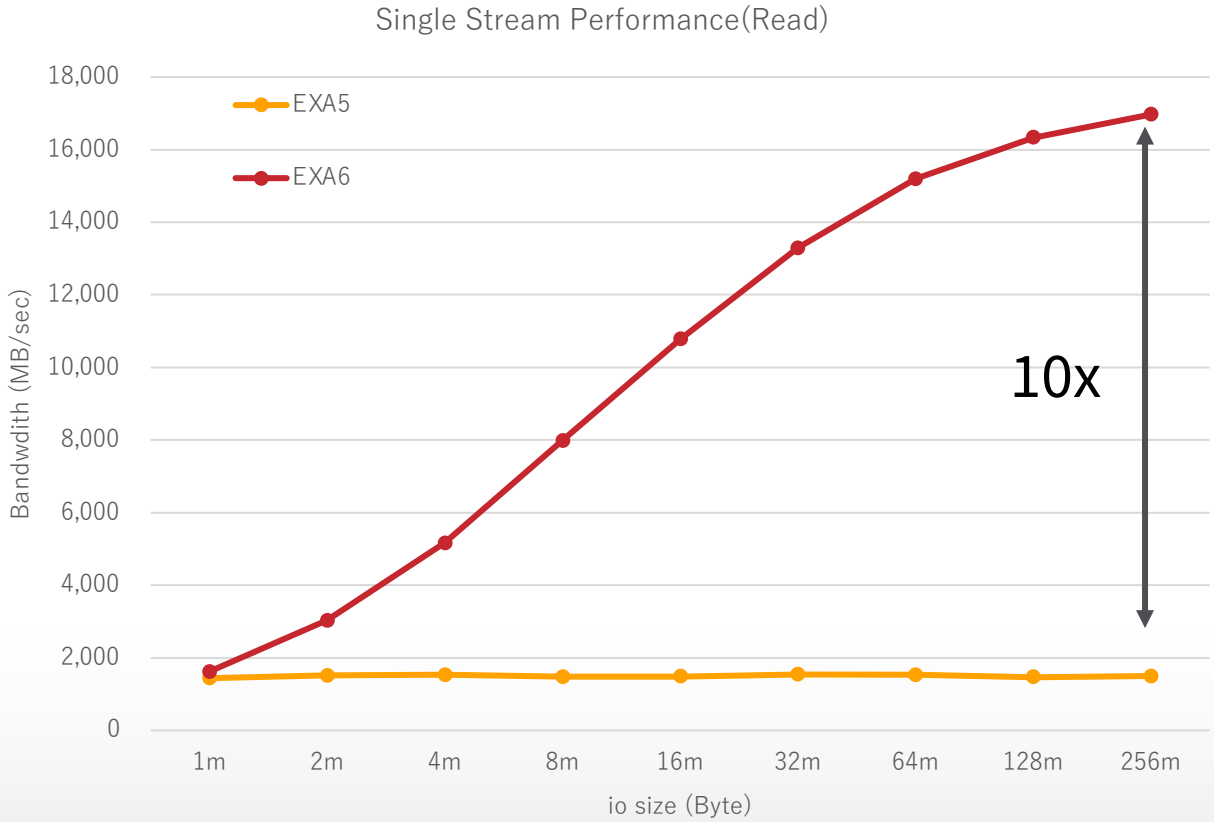


Single Stream Performance(IOR)

IOR(O_DIRECT, 256GB File), Lustre StripeCount=8, StripeSize=1MB

mpirun -np 1 ior -w -r -t $t -b 256g -e -o $DIR/file --posix.odirect





単一プロセス/DirectIO性能