

PCクラスターワークショップ in 大阪 2023

DDN HPC/AI Solutionのご紹介

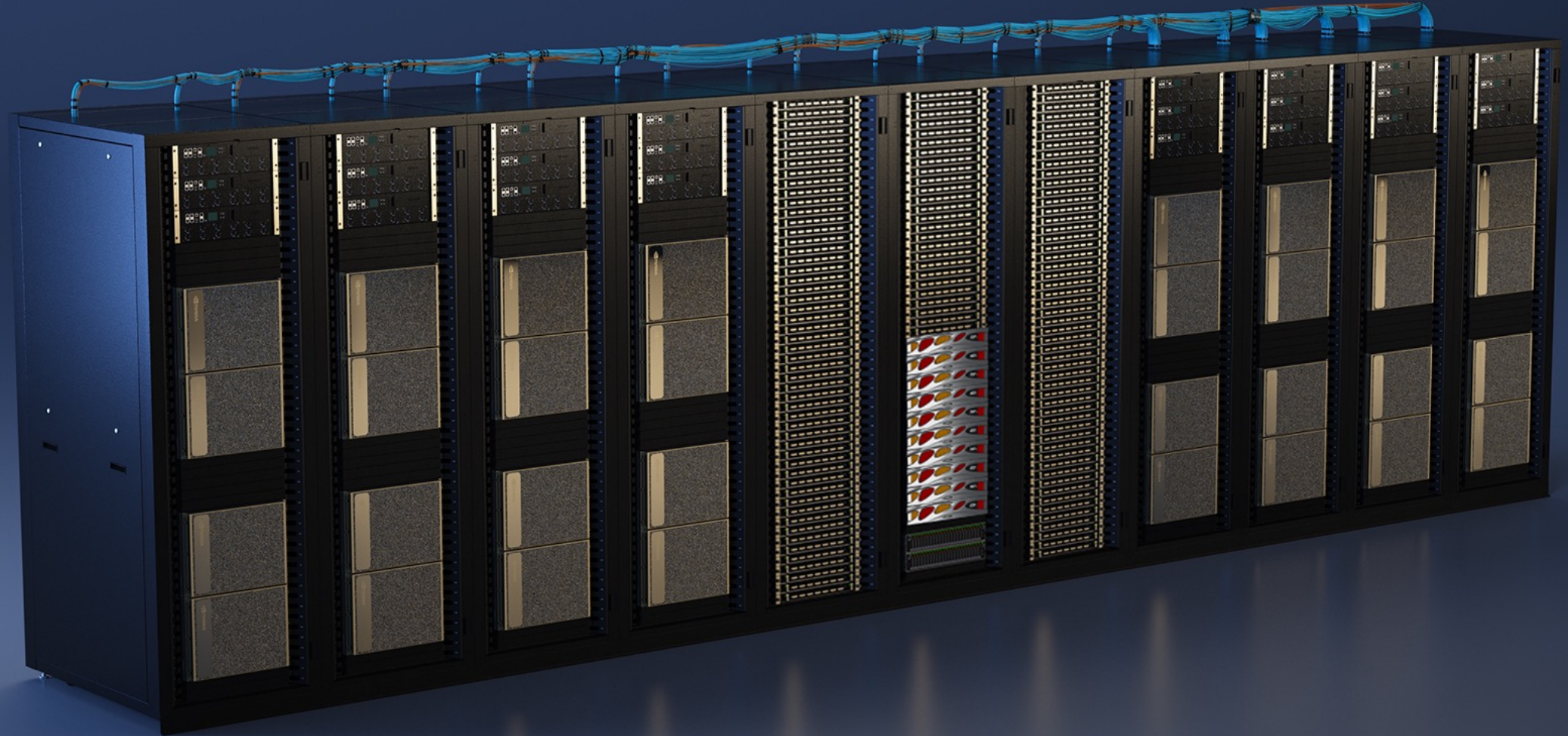
DataDirect Networks Japan, Inc.

橋爪信明

2023/6/22



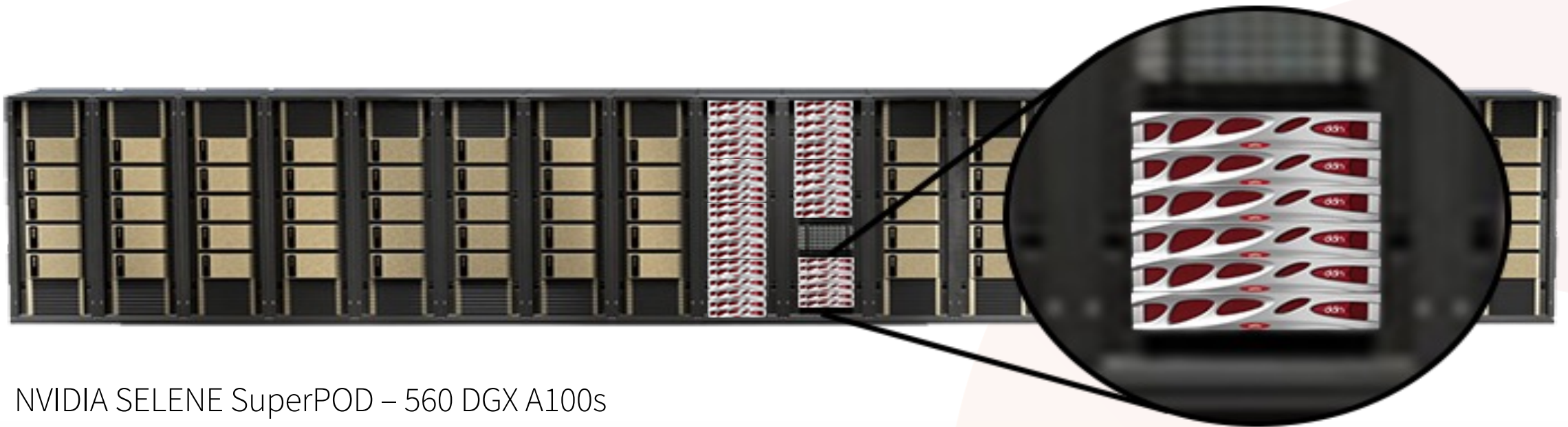
DDN is the Most Efficient Data Solution for AI and HPC
Driving 10x more data in a fraction of the energy of other systems.



NVIDIA Partners with DDN to Fast Track HPC and AI in Enterprise Centers Globally



DDN Powers the largest AI Supercomputers in the world



- NVIDIA SELENE SuperPOD – 560 DGX A100s
- **40 DDN AI400X, 2 TB/s read throughput, 15 PB NVME capacity**
- #5 on Top 500 and #2 on Green 500 Lists



NVIDIA Eos

Next-Generation DGX SuperPOD System

“NVIDIA wants the best tools for our research and development teams to use internally. The sheer scale of Eos will make it a boon for next gen AI.”

- Charlie Boyle, VP and GM DGX Systems, NVIDIA

- 576 DGX H100 Systems, 4608 DGX H100 GPUs, **48 DDN AI400X2**
- HDR400 InfiniBand Networks for Compute and Storage
- Predicted to Deliver 18 EXAFlops of AI Performance

NEW QLC Solution

DDNは今後QLCを使用してスパコンの寿命問題を解決していきます

DDN Embedded Technology
ケーブル、スイッチ、サーバ
の削減

大容量QLCドライブ
消費電力あたり2倍の容量

エンドToEnd並列デ
ータパス
消費電力あたり2倍の性能

ワークロードに合わせた性能最適化
アプリケーションの高速化

100%リニアにスケール
するソフトウェア
高コストなサイロ化防止



あらゆるスケールとワークフローに対応



TLC

Up to 30TB
Write 3500MB/s
IOPS 1.4M
Idle/Max 8/22w

最大性能



QLC

Up to 61TB
Write 3000MB/s
IOPS 1.0M
Idle/Max 5/27w

All Flashでありながら
低コスト



HDD

Up to 22TB
Write 250MB/s
IOPS 210
Idle/Max 6.5/10.2w

TB単価最安

DDN EXAScaler QLC Flashソリューション

- Single Port QLC Flashを24本搭載可能なエンクロージャを最大5基接続

CY2023Q4
予定

DDN ES400NVX2/AI400X2

- All Flash EXAScalerファイルシステムを2RUで提供
- 80GB/s Read, 65GB/s Writeスループット
- 3M IOPS
- 700TBを超えるFlash容量(Dual Port TLC)

スイッチレス 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(Single Port QLC)
- 5エンクロージャ構成で3.6PBのFlash容量を実現(30TB使用時)
- 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5エンクロージャ構成
 - 2,4,5エンクロージャ構成をサポート予定

The Fastest Controller to Drive Your QLC

Real-Time, Many-Core RAID Engine, 90GB/s 3M IOPs

Highest Density Flash

Up to 8PB TLC and QLC in 12RU

Unprecedented Resilience

SFAOS Data Protection, Intelligent data placement and SuperFast Rebuilds



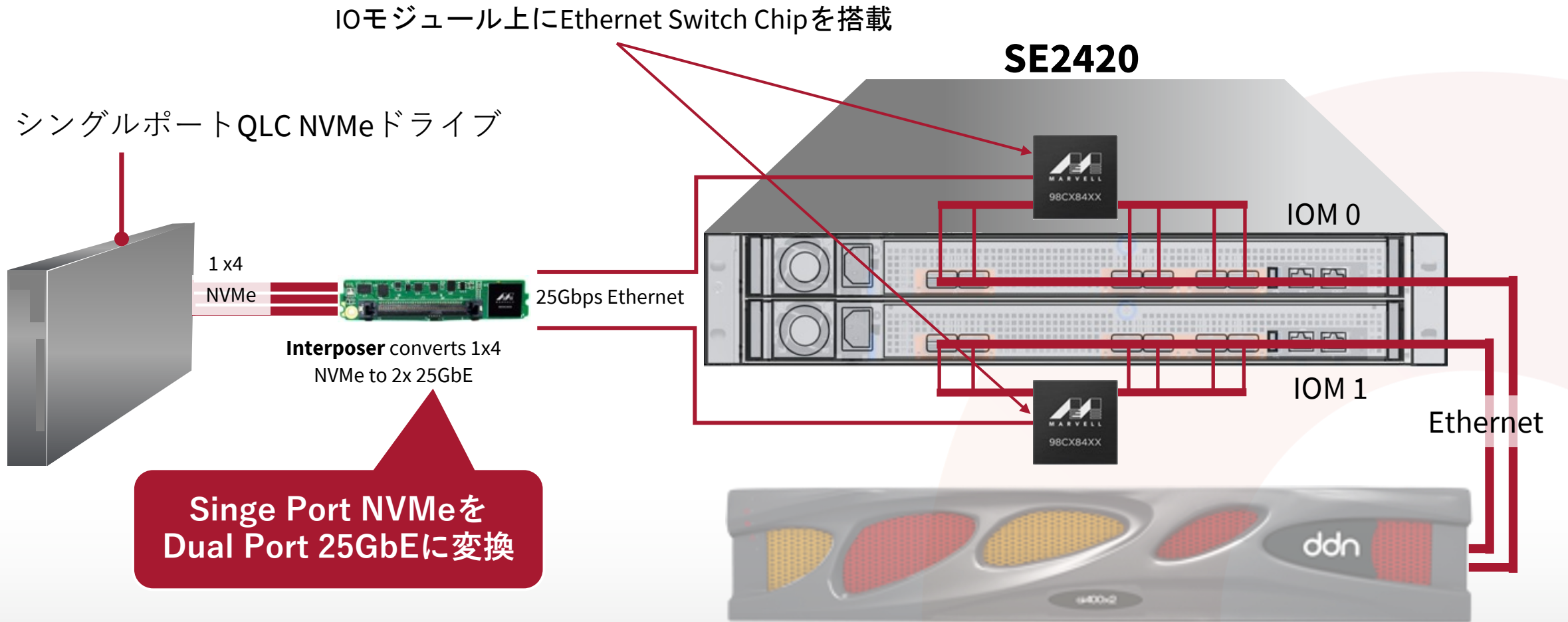
EXAScaler Parallelism

DDN Software extends outstanding performance across the network to your applications

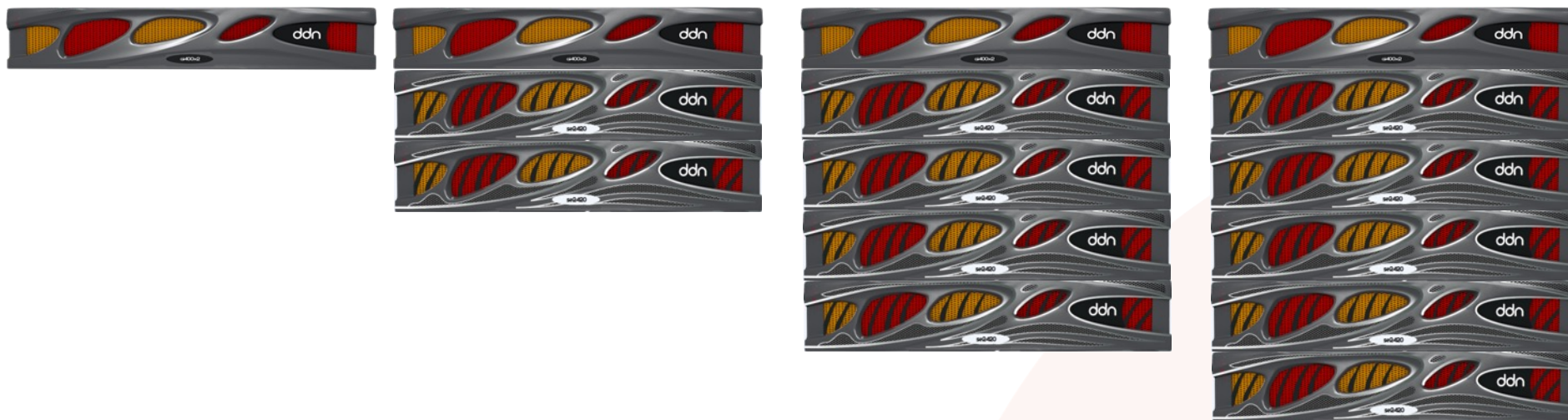
No Single Point of Failure,
Zero Midplane Design

Simple Scale up and Scale out
No External Switching with DDN Storage Fusion NVMe Fabric

SE2420: NVMe Device ➡ Ethernet Fabric



ES400NVX2/AI400X2 QLC Expansion Options



TLC NVMe Slots	24	24	24	24
Max Capacity	732 TB	732 TB	732 TB	732 TB
QLC NVMe Slots	0	48	96	120
Raw Capacity	0	2.9 PB	5.9 PB	7.3 PB
Usable Capacity		2.3 PB	4.7 PB	5.9 PB
Effective Capacity		4.7 PB	9.3 PB	11.7 PB

※ Effective Capacity : EXA6 Client-Side Compressionを利用した場合の容量

QLC構成一覧

項目/構成	1	2	3	4	5	6	7	8	9
ドライブ容量(TB)	15.36			30.72			61.44		
# of SE2420	2	4	5	2	4	5	2	4	5
# of Dataドライブ	46	94	118	46	94	118	46	94	118
# of Spareドライブ	2	2	2	2	2	2	2	2	2
Total # of ドライブ	48	96	120	48	96	120	48	96	120
物理容量(PB)	0.70	1.44	1.81	1.41	2.88	3.62	2.82	5.77	7.24
FS実効容量(PB)	0.54	1.11	1.40	1.09	2.23	2.80	2.19	4.47	5.61
最大消費電力(kW)	4.19	6.74	8.02	4.19	6.74	8.02	4.19	6.74	8.02
Write性能(GB/s)	50～55								
Read性能(GB/s)	60～70								

- 現在、提供する構成をFixにするのかFlexibleにするのかを検討しています
- 性能は現在得られてる性能と最低減の目標値を表しています
- HDDで同等の性能を出すためには、600本以上搭載する必要があります
 - 性能と容量の要件のミスマッチにより、大容量HDDドライブが使用できない場合があります
 - HDD実効容量：22TB x 600 = 10.32PB, 20TB x 600 = 9.38PB, 18TB x 600 = 8.44PB, 14TB x 600 = 6.56PB, 10TB x 600 = 4.69PB
 - HDD600本構成で最大消費電力は、12kW程度

TLCにより最大性能を得る場合

TLC All Flash

全てのワークロードに最高のレイテンシ性能を提供



3.0M

IOPS in 2RU

1.5TB/s

In a single rack

TLC All Flash	
Write (GB/s)	64
Read (GB/s)	80
Raw Capacity (TB)	Up to 720
Random Read IOPS (M)	3.0
Power (Nominal) (KW)	2.1
Power (Idle) (KW)	1.8
Rack Units	2

容量は30TBドライブ利用時の値

QLCコンフィギュレーション例

性能最適化/コスト構成



ES400NVX2 with 2x QLC Expansion	
	QLC
Write (GB/s)	50**
Read (GB/s)	60**
Raw Cap. (TB)	2928
RR IOPS (M)	3.0*
Power* (Nominal) (KW)	4.9
Power* (Idle) (KW)	3.7
Rack Units	6

容量最適化/コスト構成



ES400NVX2 with 5x QLC Expansion	
	QLC
Write (GB/s)	50**
Read (GB/s)	60**
Raw Cap. (TB)	7321
RR IOPS (M)	3.0*
Power* (Nominal) (KW)	9.3
Power* (Idle) (KW)	6.5
Rack Units	12

容量は61.44TBドライブ利用時の値

*推定値

**現在の性能(最適化前)

QLC システムコンフィギュレーション例

性能/コスト最適化構成



READ Performance

420GB/s

WRITE Performance

350GB/s

Capacity

20PB

容量/コスト最適化構成



READ Performance

180GB/s

WRITE Performance

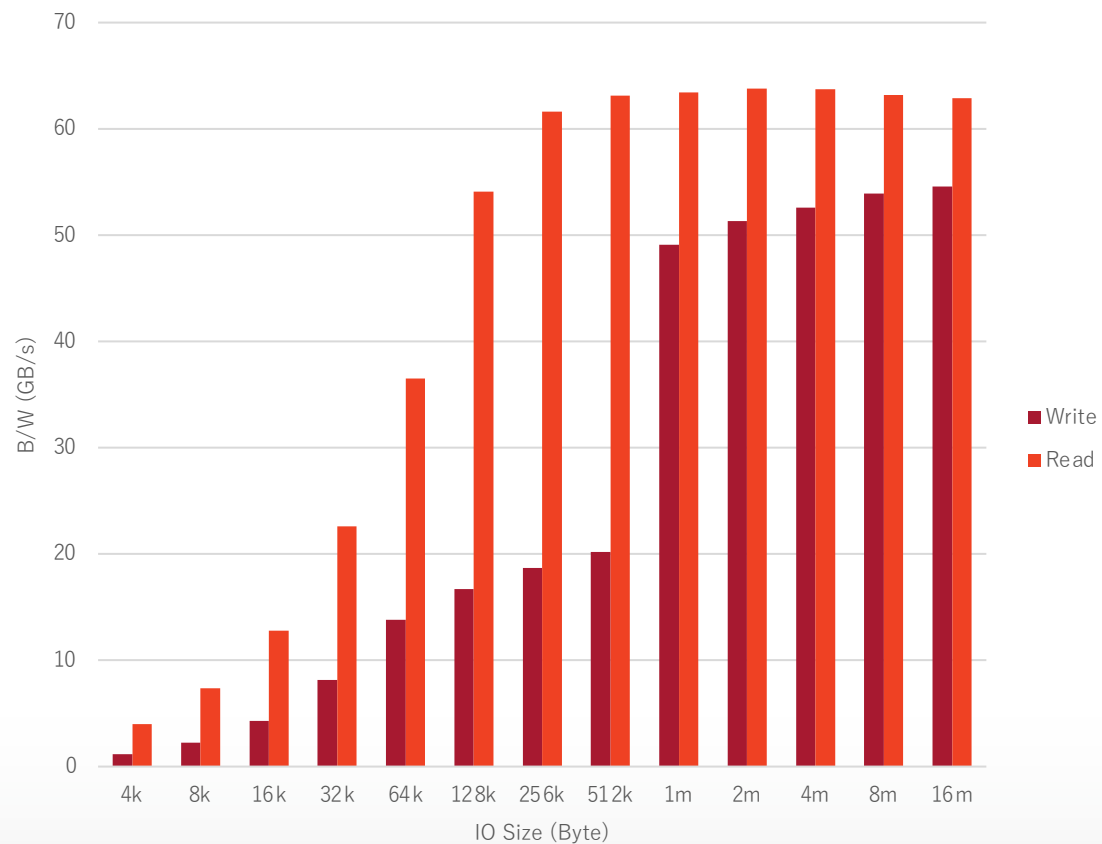
150GB/s

Capacity

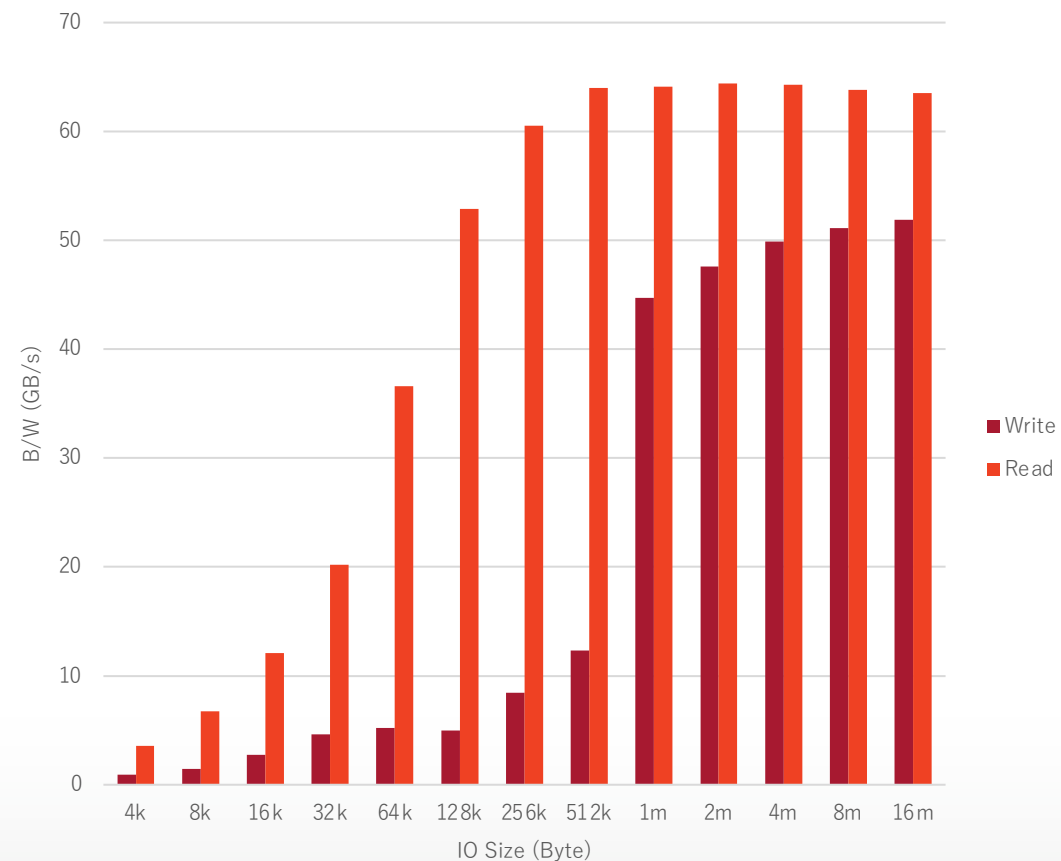
22PB

QLC - 現在の性能値

FIO (Sequential IO, O_DIRECT, QD=1, 4 x VM, 4 x Pool, 80 x QLC NVME)



FIO (Random IO, O_DIRECT, QD=1, 4 x VM, 4 x Pool, 80 x QLC NVME)



Sequential IOでHDDのPeak程度の性能(Read 70GB/s, Write 55GB/s)を得ることが最低目標



QLCを販売するにあ たっての補足

書き込み寿命(Endurance)

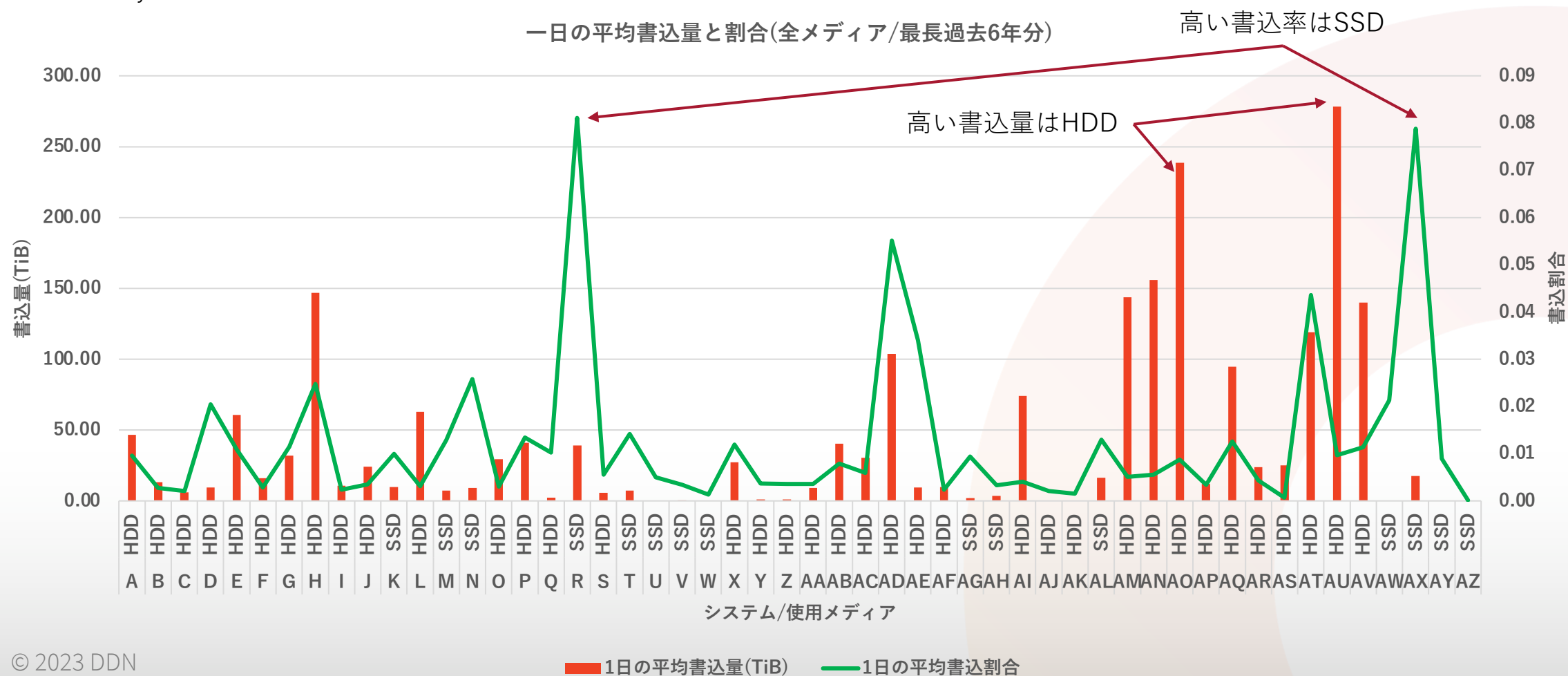
故障率

消費電力

ddn

QLC Endurance(1)

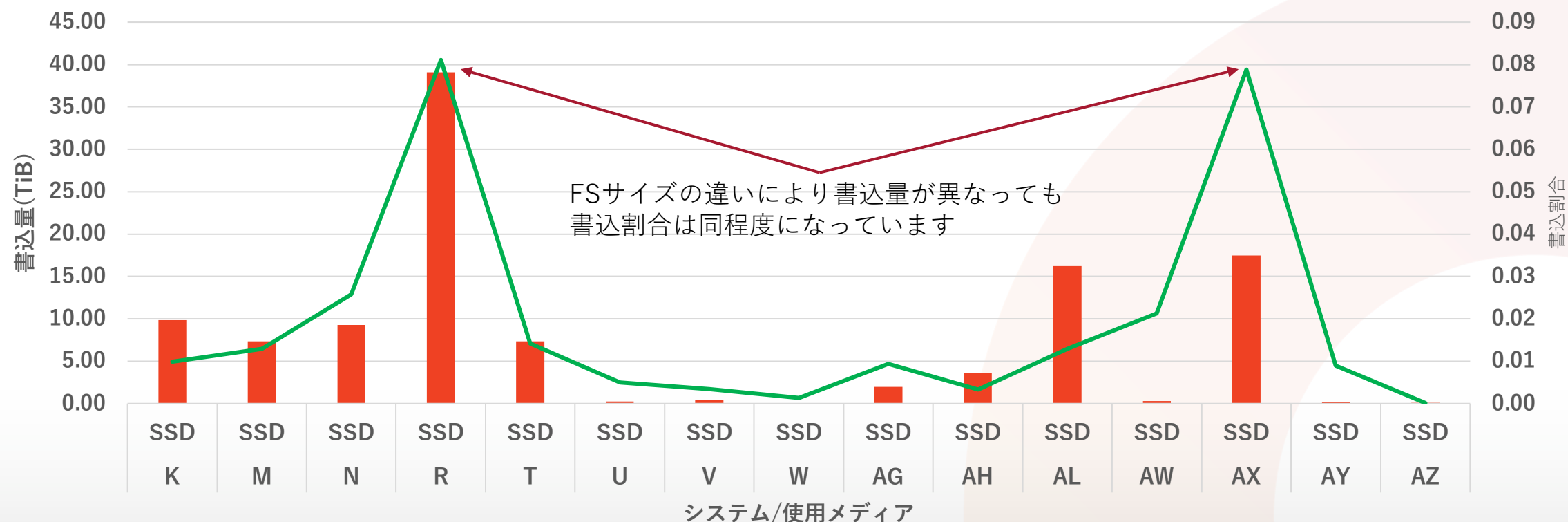
- 国内導入済52システム(最大6年運用)のLustreファイルシステムあたりの一日の書込量とファイルシステムサイズに対する書込量の割合を取得
 - ファイルシステムサイズに対して最大でも一日0.08%程度の書込量
 - Parity分の書き込みを考慮した値



QLC Endurance(2)

- SSDを使用した15システムのための表です
 - SSDはすべて1 DWPDのNVMe TLC SSDです

一日の平均書込量と割合(SSD)



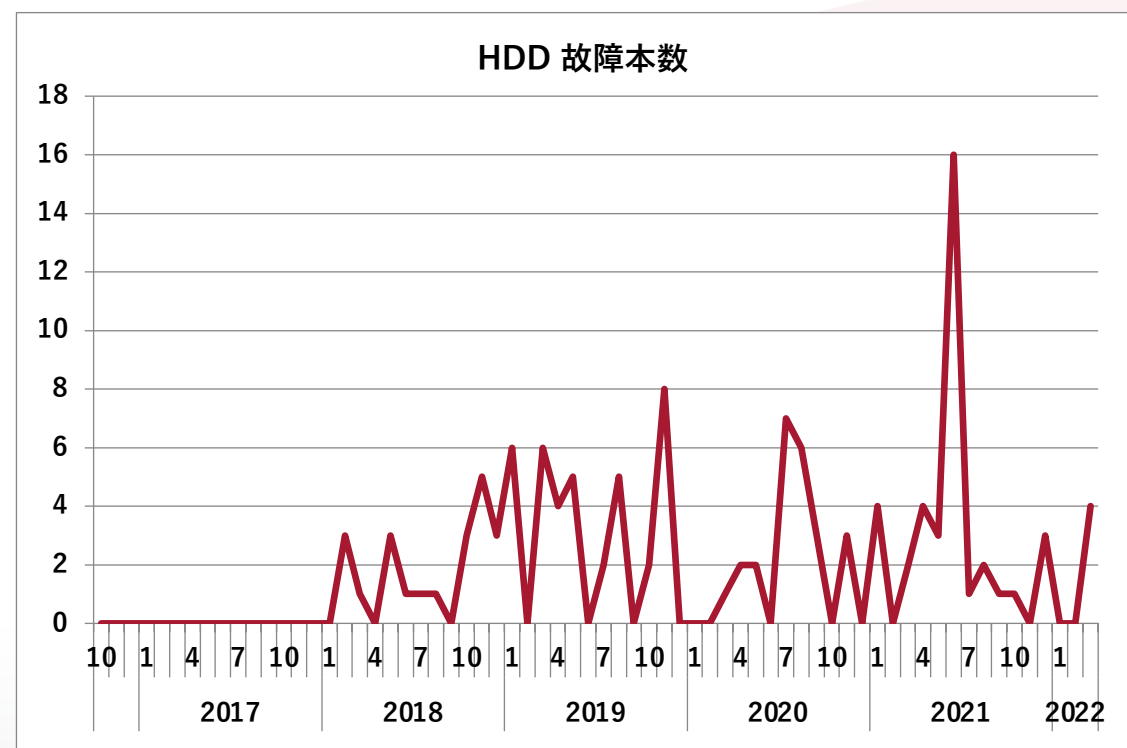
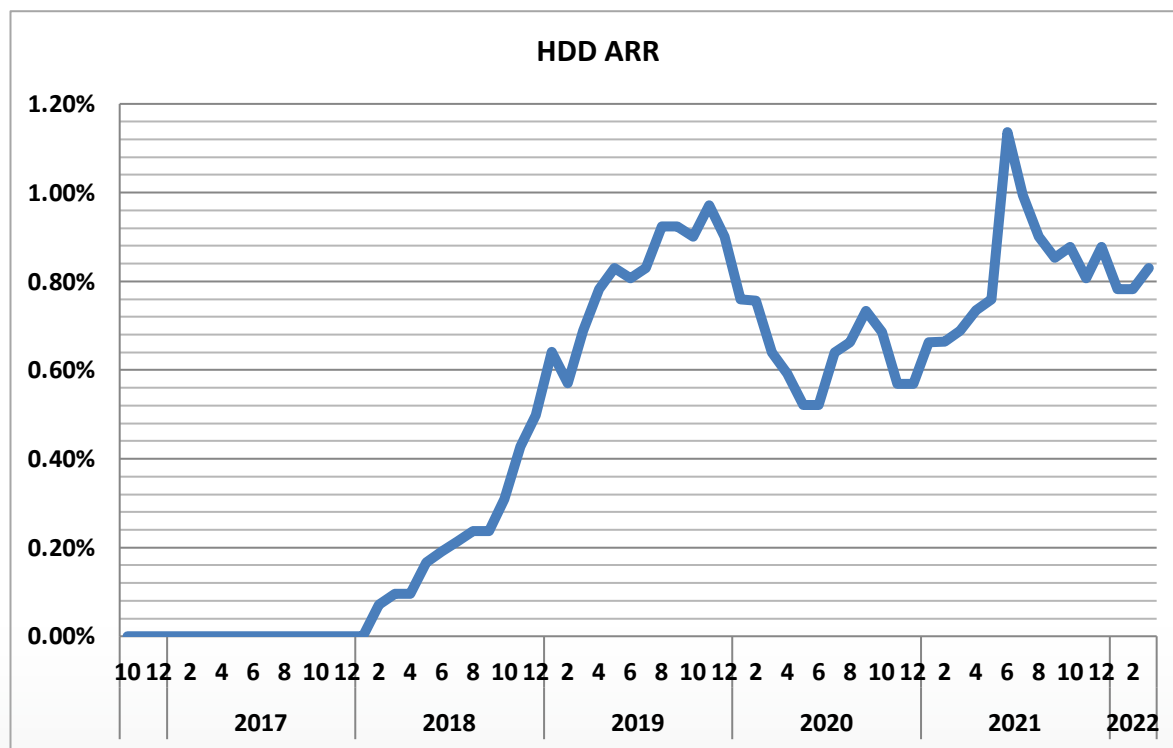
■ 1日の平均書込量(TiB) — 1日の平均書込割合

QLC Endurance(3)

- SSDのシステムはHDDのシステムよりFSの容量とOSTの容量がいずれも小さい場合が多いです
 - 書込量がHDDより少なくても書込割合は高くなります
 - HDDを使用したOSTは数百TBの場合もあるため、書込量がSSDより大きくても書込割合は低くなります
 - 一日100TiB程度の書き込み、書込割合0.04のシステムAUのHDD OSTのサイズは約340TiB
 - 一日40TiB程度の書き込み、書込割合0.08のシステムRのSSD OSTのサイズは約15TiB
- 実システムでの一日の書込割合の平均はHDDで0.01、SSDで0.02程度でした
 - HPCのワークロードでは、QLC SSDのDWPD 0.5でも運用には十分耐えうると考えます

故障率(1)

- 参考までに運用を終了した大規模スパコンサイトで使用していたHDDの故障本数、ARRを示します。
。導入本数は4,200本、総故障本数は124本で導入本数の2.95%です



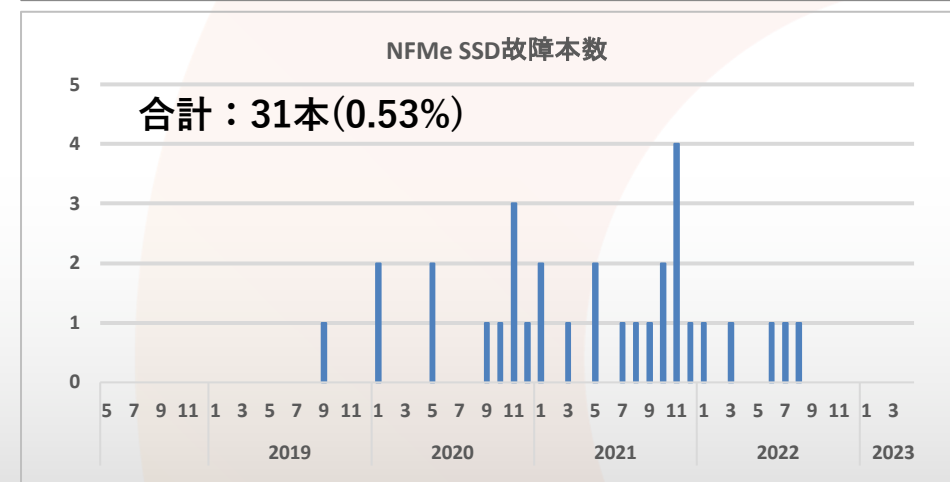
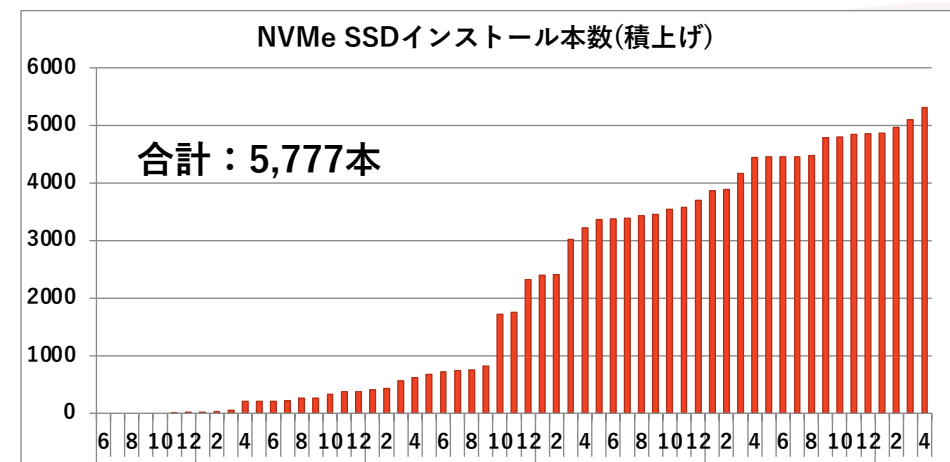
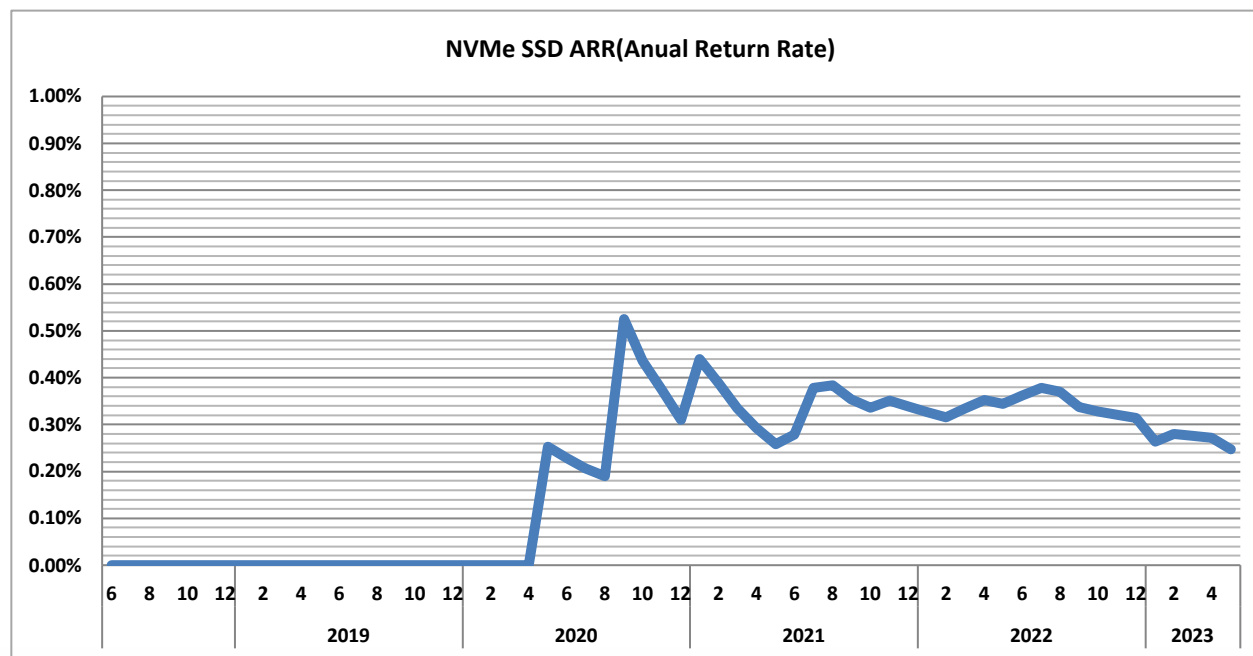
ARR算出式

$\text{Annual MTBF} = \text{Total Field Hours(Year)} / \text{RMA counts}$

$\text{ARR} = 1 - \text{EXP}(-8760(\text{Hour}) / \text{Annual MTBF})$

故障率(2)

- SSDは駆動している機械が無い場合、HDDに比べて故障率が低くなります
。直近5年間のNVMe TLC SSDインストール本数、故障本数、ARR



ARR算出式

Annual MTBF = Total Field Hours(Year) / RMA counts

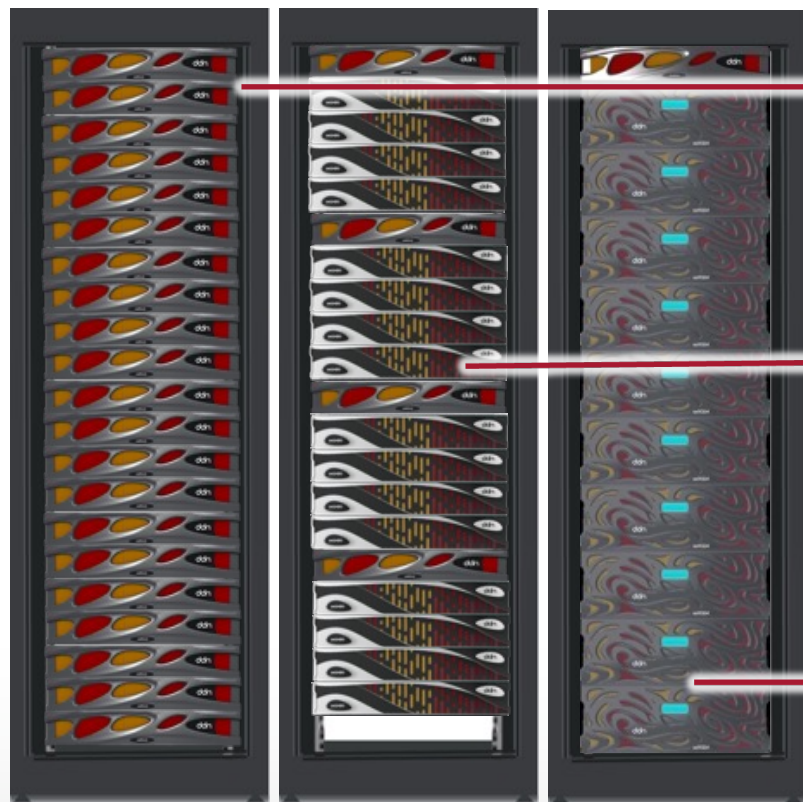
ARR = 1-EXP(-8760(Hour)/Annual MTBF)

消費電力

- HDDを用いた並列ファイルシステムでは、HDDは常に回転しているため、ほぼピークの電力を消費続けます
 - HDDはピーク10W程度、アイドル6.5W程度です
- SSDは操作によって電力が変わります
 - 提案の30.72TB QLC SSDの諸元は確定していませんが、15.36TB QLC SSDと同等の消費電力を見込んでいます
 - 15.36TB QLC SSDの電力は以下のとおりです
 - QD256/4KiB Random Write、500μsサンプル時間 : <27W
 - QD256/128KiB Random Write, 100msサンプル時間 : <22W
 - QD256/128KiB Sequential Read, 100ms サンプル時間 : <21W
 - idle : <5W
- SSDは使用中のみ消費電力が上がるので運用時のトータル消費電力はHDDシステムと同等もしくは低くなると推定していますが、現在、根拠データを持ち合わせておりません
 - 今後、ラボにて消費電力についてのデータ収集、解析を行う予定です

World's Most Performing and Flexible Storage At Any Scale

TLC Flash QLC Flash Hybrid



Best IOPS & Throughput per rack

1ラックあたり (21 x ES400NVX2)

63M IOPS

1.68 TB/s READスループット

1.26 TB/s WRITEスループット

物理容量 15.48PB TLC Flash領域

Best Price per Flash TB

1ラックあたり (4x ES400NVX2+SE2420x5)

12M IOPS

280 GB/s READスループット (目標)

220 GB/s WRITEスループット (目標)

物理容量 29.49PB QLCFlash領域

Best Price per TB

1ラックあたり (1x ES400NVX2+SS9024x10)

物理容量 19.80PB HDD領域, 0.73PB TLC Flash領域

80 GB/s READスループット (Flash)

70 GB/s Read スループット (HDD)

60 GB/s WRITEスループット (NVMe/HDD)



ddn