

Gfarmワークショップ2019



NVMe over Fabricに対応した分散ストレージ 「Excelero」の紹介

2019/2/1

マクニカ・富士エレクトロホールディングス グループ

半導体商社の
Global Leader

Global Top5
(ELECTRONICS DISTRIBUTORS 部門)



2017年度売上
5,041 億円



3132

2500以上の従業員
お客様に根ざしたサポートを
世界 70 拠点以上で展開

200 以上の
最先端製品・技術を提供

mF Macnica Fuji Electronics

GROUP

マクニカ社内分社



Brilliant
Technologies
Company

CLAVIS
Company

TecStar
Company

ELSENA

国内関連会社
(ネットワーク・セキュリティ)

macnica
networks

macnica
solutions

国内関連会社
(コンポーネント・富士エレクトログループ)

FUJI ELECTRONICS CO.,LTD

フレストロニクス



マクニカ海外現地法人
(主に日系顧客対応)

MACNICA

米国 (カリフォルニア、イリノイ)
シンガポール
香港
台湾
中国 (深セン、上海、北京)
韓国
タイ
マレーシア

海外グループ会社
(現地顧客対応)

Cytech
中国・香港

Cytech
ASEAN・インド

Gfec
台湾

MACNICA
AMERICAS
北米

MACNICA
D H W
ブラジル

MACNICA
Gm bH
欧州

取扱い製品



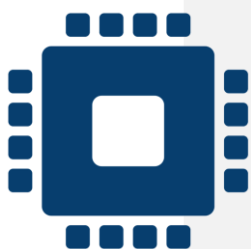
Network



- ストレージネットワークに最適な、**低遅延、高機能バッファ搭載**スイッチ
- ハイパースケールネットワーク(Cumulus)
- Smart NIC, 高機能アダプタカード(GPU-Direct, RDMA, OVS, VXLAN)
- LincX ケーブル、モジュール製品



- インテントベースネットワーキングソフトウェア
デザイン、構築、運用を完全自動化



Compute



- FPGAアクセラレータカード
- Deep Learningアクセラレーション
- DataBaseアクセラレーション



- マルチコア、高密度サーバ(Xeon, AMD)
- GPU搭載サーバ
- All NVMe Flashストレージ



- 各種GPU製品



Storage



- NVMe SSD
- 大容量オブジェクトストレージシステム
- JBOD/JBOF



- 次世代NVMeスケールアウトストレージ
ソフトウェアソリューション



- コンポーザブル・インフラストラクチャ
統合管理ソフトウェア

本日のお話

「NVMe over Fabricに対応した分散ストレージExceleroの紹介」



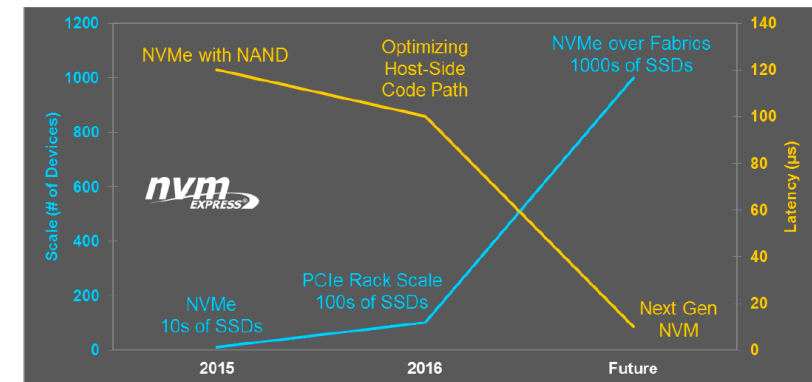
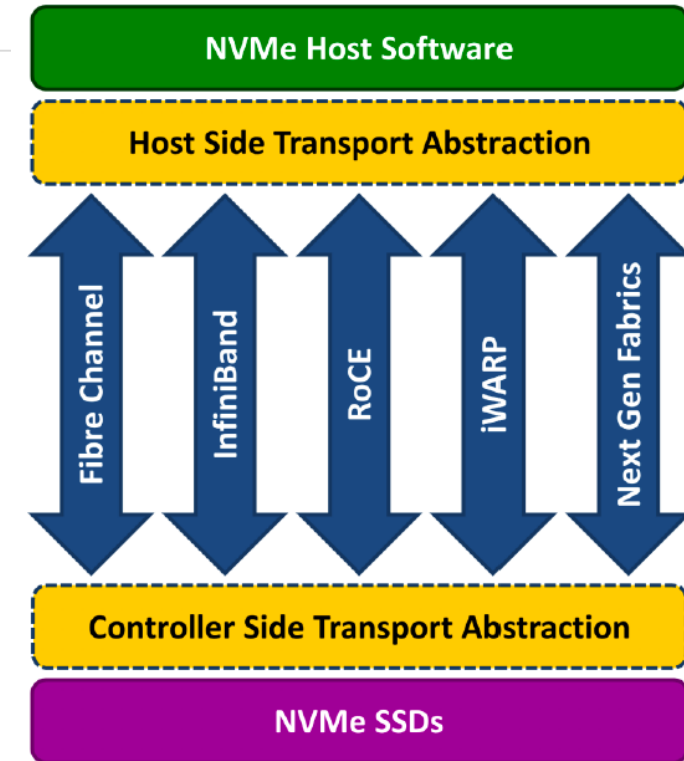
NVMe over Fabricsの現在
(いま)



Excelero NVMeshの紹介 & ベンチマークデータ

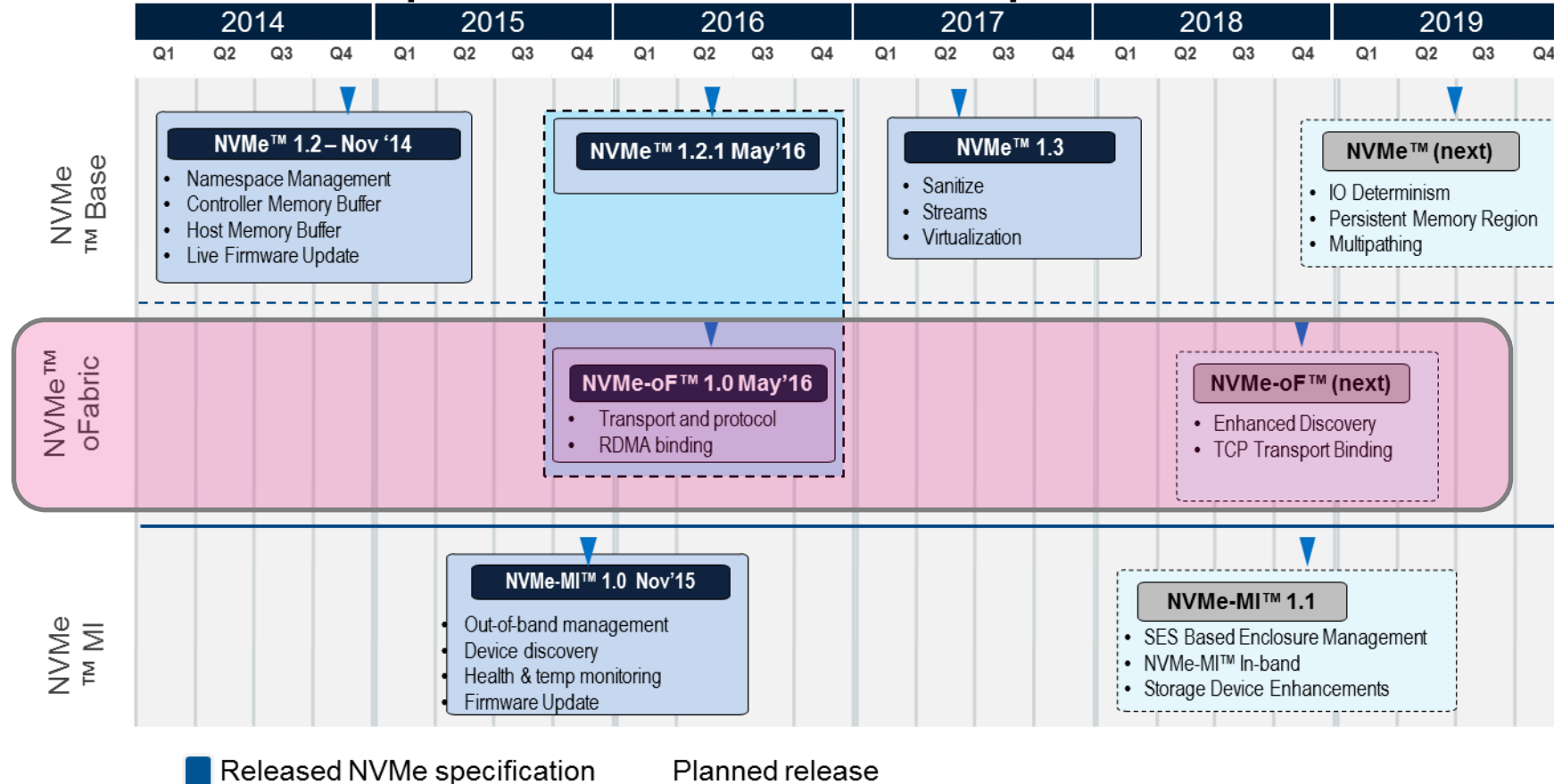
NVMe over Fabrics

- NVMeをローカル利用する課題
 - PCIスロット以上には拡張できない
 - ノード単位の容量制限以上をアプリケーション利用できない
- NVMe over Fabricとは
 - NVMeストレージを複数ノードで共有する仕組み
 - マルチプロトコルサポート
 - データプレーン、マネジメントプレーン機能提供
 - NVMe over Fabrics標準規格化
 - NVMe over Fabrics 1st Version 1.0 (2016/June)
 - NVMe over Fabrics Latest Version 1.0a (2018/July)
- NVMe over Fabricが目指すところ
 - 低レイテンシー性能(over Fabric)
 - ファブリックスケール



NVMe over Fabricsのいま

NVMe™ Specification Roadmap: 2014 – 2019

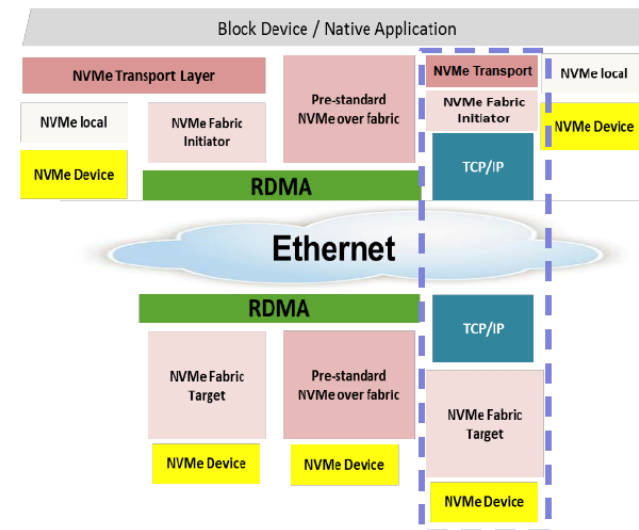
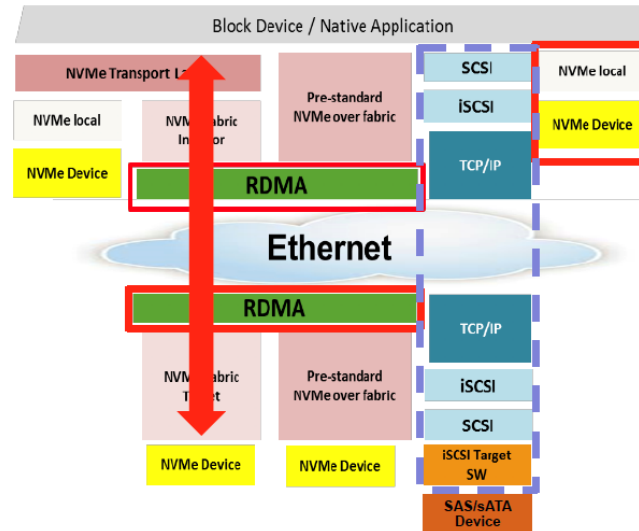
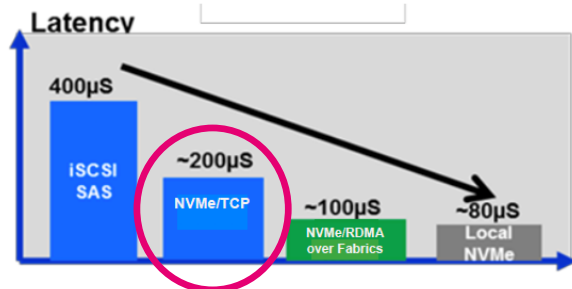
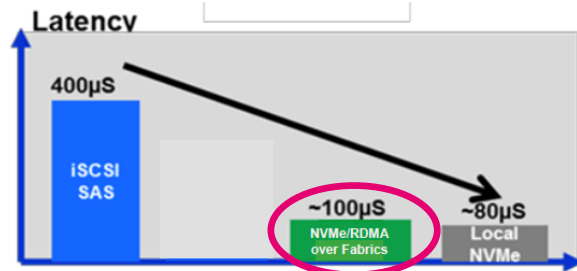


- サービス環境で利用できる機能セットのサポートはこれから
 - SCM, QCLサポート
 - Multipath
 - Namespace Sharing
 - Discovery(Management)
 - Security
 - Etc

https://www.nvmedeveloperdays.com/English/Conference/Proceedings_Chrono.html

NVMe over Fabricsのいま(性能面)

- By extending NVMe efficiency over a fabric
 - NVMe commands and data structures are transferred end to end
- Bypassing legacy stacks for performance
- First products and early demos all used RDMA
- Performance is impressive



- TCP/IP(iSCSI)
 - 性能面でのアドバンテージなし
- RDMA/RoCE
 - 現状、最大限のパフォーマンスを求めらるのであればRDMAが必須
 - Exceleroはさらにレイテンシーのチューニングがされている(+5usec)
- TCP/IP
 - iSCSIに比べ1/2のレイテンシー性能
 - RDMAが勝る

ここまでのまとめ

- NVMe over Fabricsはフラッシュストレージのファブリック技術のスタンダードとして発展していく
- 現状、サービス用途のユースケースでは課題は多い
 - そもそも、バージョン1.0x ...
- 一方で、既にI/Oパフォーマンスが必要なEmergentな市場/アプリケーションはある

Excelero紹介

Exceleroの紹介

Leadership & Sales

Business/Sales leaders with histories of driving revenue growth



Lior Gal
CEO (co-founder)



Michael Chazot
VP Strategic Alliances

Product & Marketing

Defining market-leading products for over a decade



Josh Goldenhar
VP Products



Tom Leyden
VP Corporate Marketing

Technologists with extensive experience with products for the hyper scale storage market



Yaniv Romem
CTO (co-founder)



Ofer Oshri, Ph.D
VP R&D (co-founder)



Omri Mann
Chief Scientist (co-founder)

拠点：イスラエル(HQ)、北米
概要：2014年設立、2017年3月NVMesh GAリリース
その他：投資トータル\$35M(RoundC) as of 2018 Oct



Software-defined Block Storage

Lowest Overhead

Local Flash Latency across the Network

100% Software-defined

Use any Hardware

Block Storage

Use any File System

Exceleroの紹介

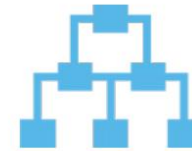
● コアテクノロジー「NVMesh」

- RDDA(標準技術RDMA+NVMEOFがベース)による、ゼロオーバーヘッドのアクセス性能
 - リモート性能 = ローカル性能
 - ゼロCPUアクセス
 - 低レイテンシー
- スケールアウトSAN
 - PBクラスのスケラブルなブロックストレージ提供
 - アプリケーションからの透過性
 - データ冗長(RAID10, Erasure Coding)
- GUI、オープンAPIによる統合管理、監視
 - プロビジョニング
 - モニタリング



Management

- Provisioning
- Monitoring



Storage Services

- Data Protection
- Multipathing
- Logical Volumes

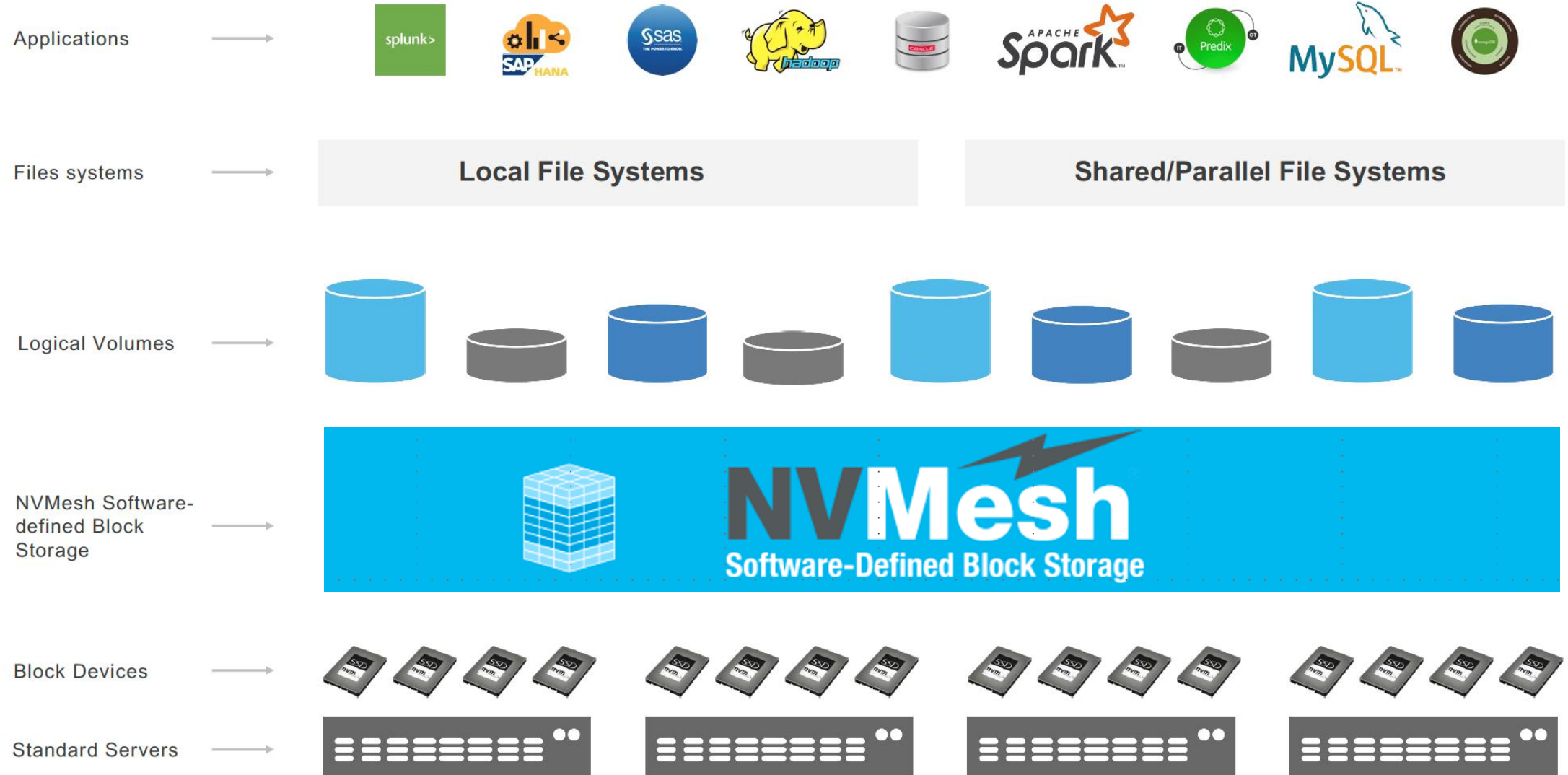


Hardware Abstraction

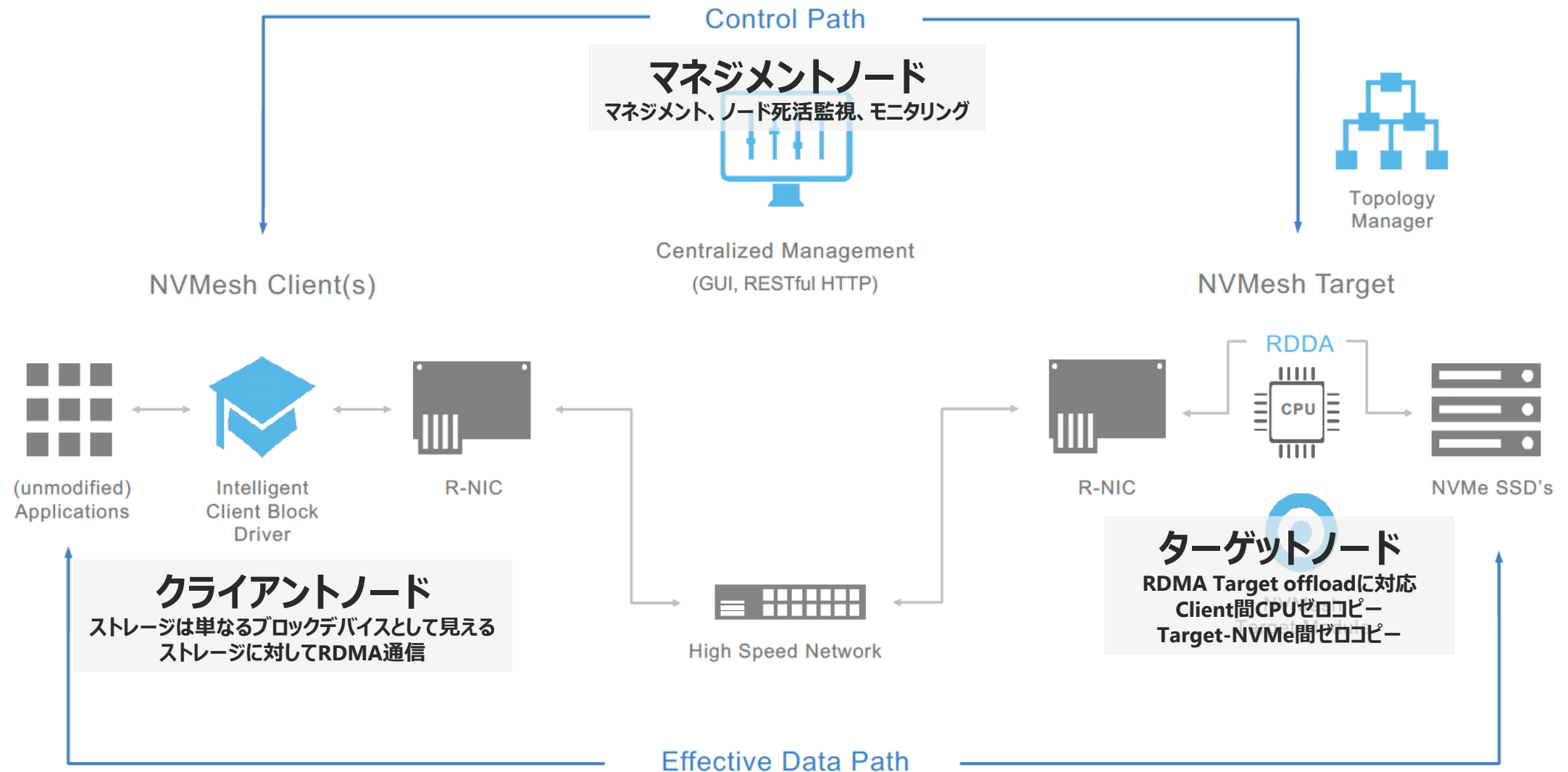
- Create a pool of block storage out of Non-Volatile devices and networking
- Supports RDDA & NVMf for transport



NVMeshアーキテクチャ

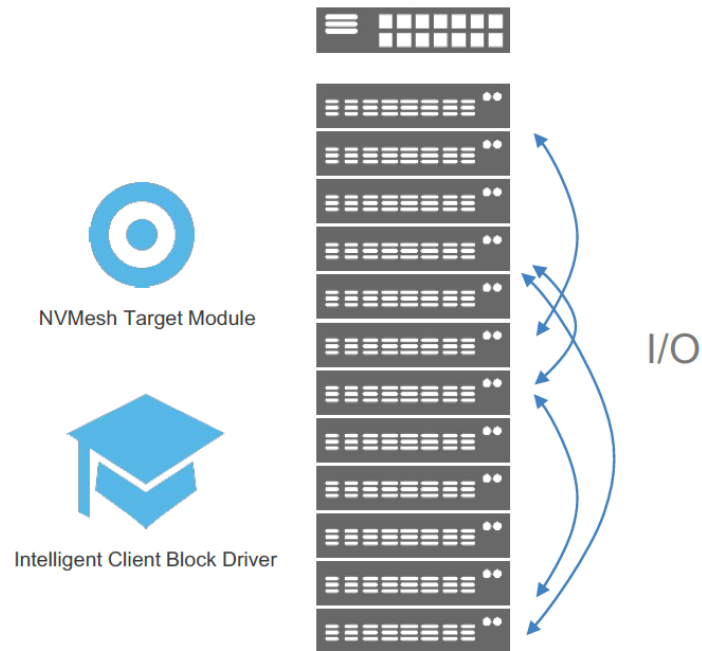


NVMeShアーキテクチャ(コンポーネント)



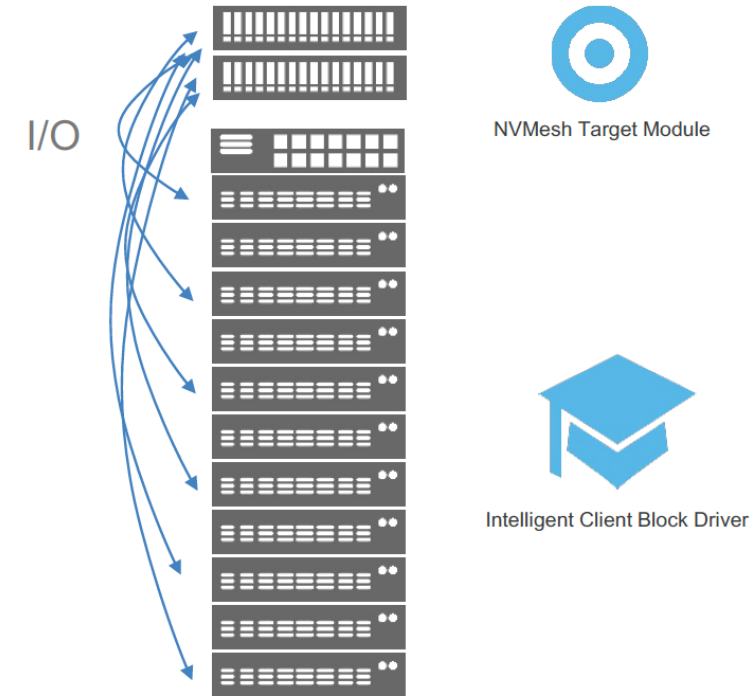
柔軟な構成が可能

Local Storage in Application Server



- Storage is unified into one pool
- NVMesh Target Module & Intelligent Client Block Driver run on all nodes
- NVMesh bypasses server CPU
- Linearly scalable

Storage is Centralized



- Storage is unified into one pool
- NVMesh Target Module runs on storage nodes
- Intelligent Client Block Driver runs on server nodes
- Applications get performance of local storage

NVMeshサポート機能

- NVMeshサポート機能(v1.2.x – 2019/1月)

Data Management & Protection

Multiple transports	NVMf-ready, Patented RDDA,
Logical Volumes	Concatenated, RAID 0, RAID 1, RAID 10 Erasure Coding 8+2 (NVMesh2 coming soon)
Multiple Drive Types	NVMe, NVMf, SATA, SAS, 3D XPoint
Failure Domains	Host, rack & row aware

Management & Monitoring

Interactive Interfaces	Web GUI & CLI commands
Automated Provisioning	RESTful API, Docker Persistent Volumes

Next-gen Data Center

Flexible Topologies	Physically converged, disaggregated or mixed
High Performance	Limitlessly scalable at near 100% linear performance
Scale-out Architecture	Intelligent clients utilizing multiple hosts, drives, network-paths and racks
Supports multiple fabrics	Ethernet (RoCE v2), InfiniBand TCP/IP, FC (NVMesh2 coming soon)

ユースケース

Excelero Customer Success



62% reduction in Oracle
licensing costs doing the
same amount of processing



Reducing NVMe TCO
for Predix Cloud with
70%



128 x86 servers with NVMe
SSD's
140GB/s of throughput
30M random 4k IOPS



80 pooled NVMe devices
give 250GB/s of throughput
and 20M random 4k IOPS



4K streaming @ 60fps to 50+
workstations concurrently



10x lower IO latency than
Ceph through all layers in
the virtual machine

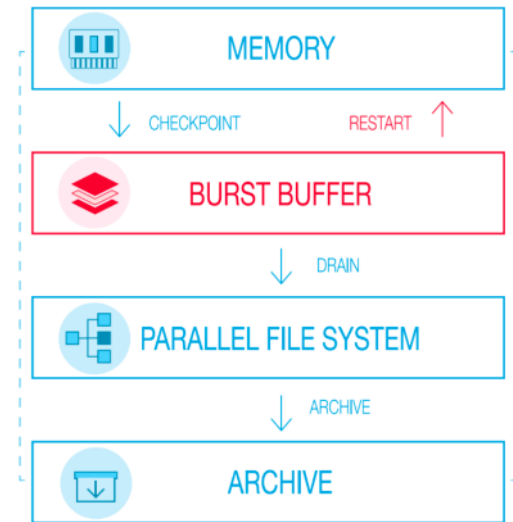
- 採用ポイント

- 250GB/sec バーストバッファ
- チェックポイント・リスタート

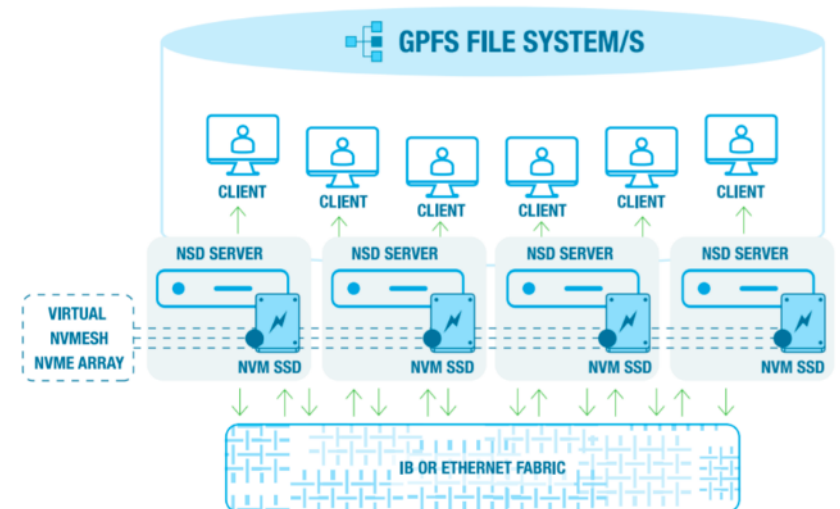
- ハイライト

- 圧倒的パフォーマンス！
 - 80 pooled NVMe devices
 - 148 GB/s of write burst (device limited)
 - 250GB/s read throughput (network limited)
 - Well over 20M random 4k IOPS
- ファブリック化により大容量化、階層化などを実現化
 - ブロックストレージ : NVMe
 - ファイルシステム : IBM Spectrum Scale

HPC STORAGE ARCHITECTURE WITH BURST BUFFER

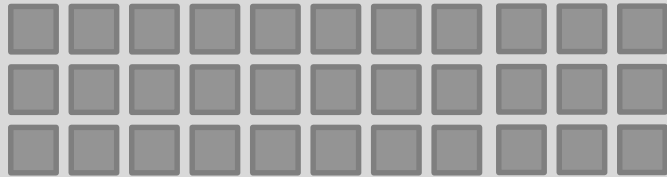


CONVERGED NSD SERVER TOPOLOGY

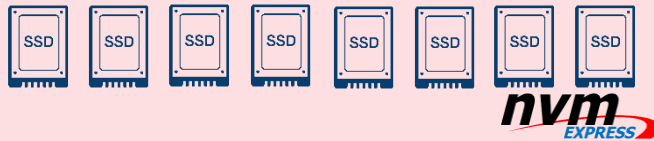


NVMeを活用した並列ファイルシステムの高速化ソリューション

Compute Cluster



NVMe based distributed Hot-Tier



NVMesh



IBM Spectrum Scale

並列ファイルストレージ



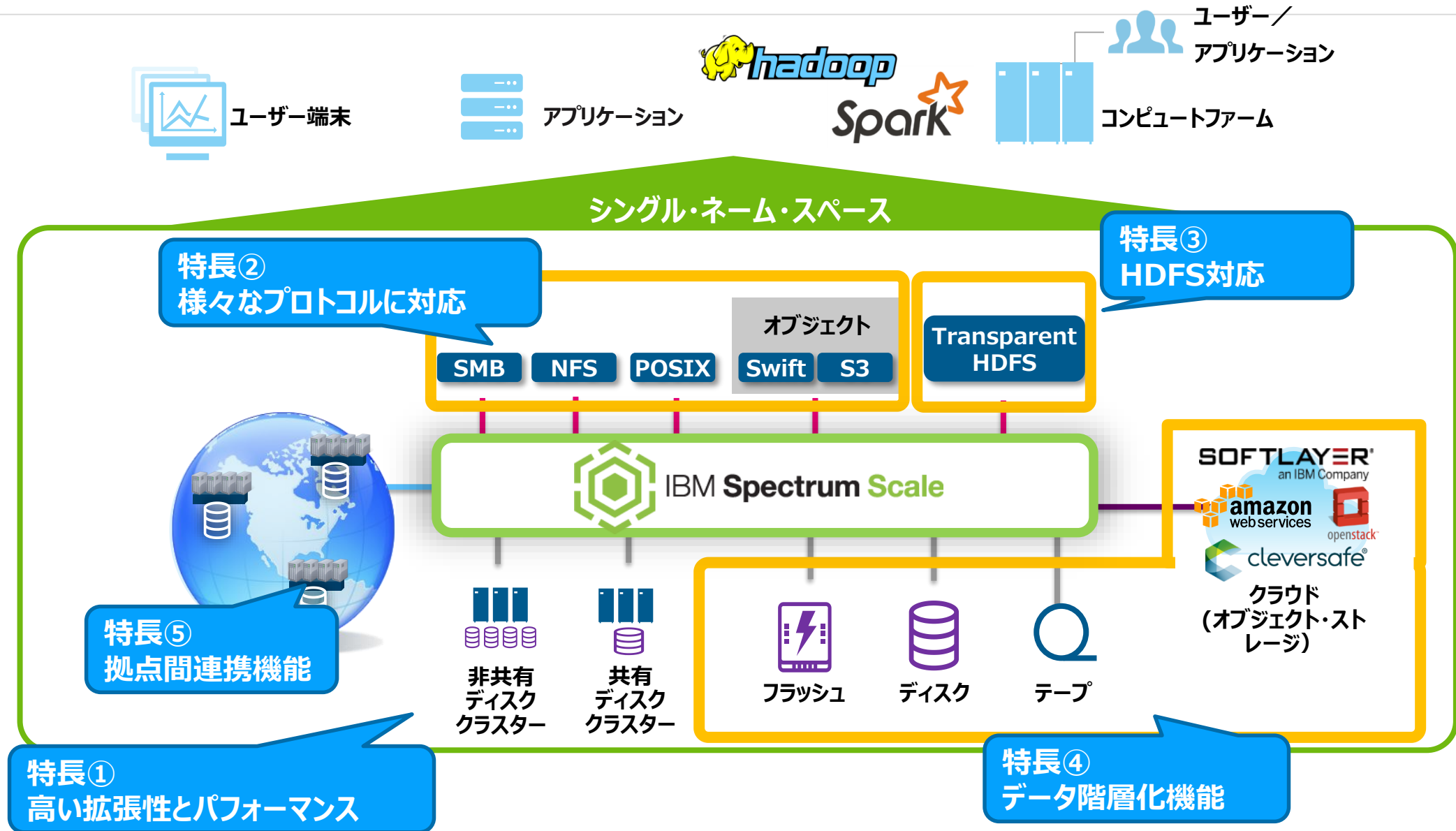
● 特徴

- 高速read/writeバーストバッファ
- PFSへのIOアクセス高速化
- スケールアウト
 - 性能 = NVMeローカルパフォーマンス
 - 容量 = ～PB

● 機能概要

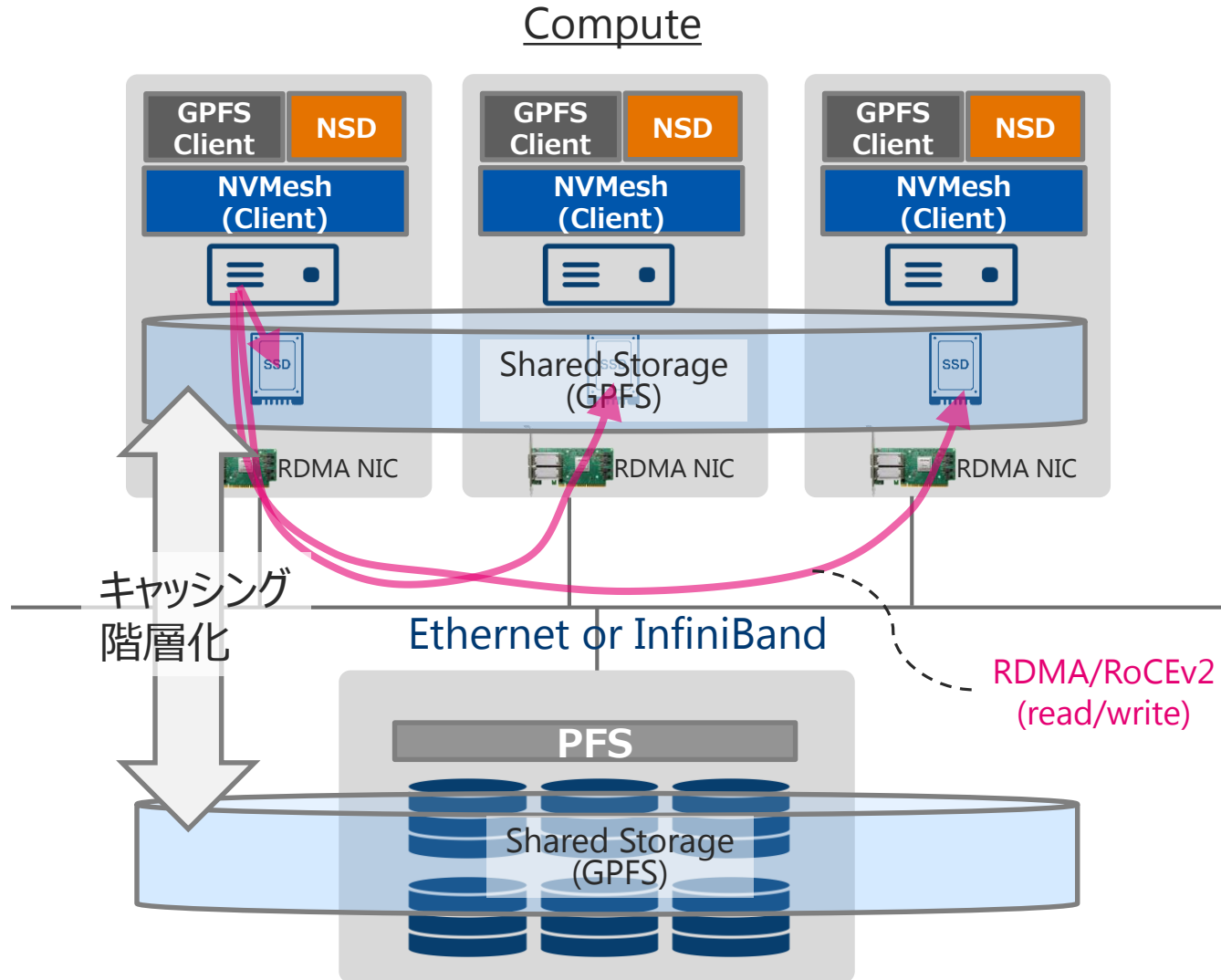
- Excelero NVMesh
 - NVMe over Fabric(RDMA[Eth/IB])による超高速IO
 - スケールアウトブロックストレージ
 - データ冗長(RAID10, Erasure-Coding)
- IBM Spectrum Scale
 - 並列ファイルシステムの実現
 - 高速バッファ/大容量ストレージのキャッシング/階層化

IBM Spectrum Scale (GPFS) 概要



NVMesh x Spectrum Scale 構成例

例：Local NVMeを活用したHyper Converged構成



- ✓ ローカルNVMeをストレージHCI化
- ✓ リモートノード間のファイル共有をLocaliz (ローカル=リモート性能) ※レイテンシー +5u
- ✓ GPFSによる高速ローカルNVMe領域と大容量ストレージ間のキャッシング、階層管理

構成概要

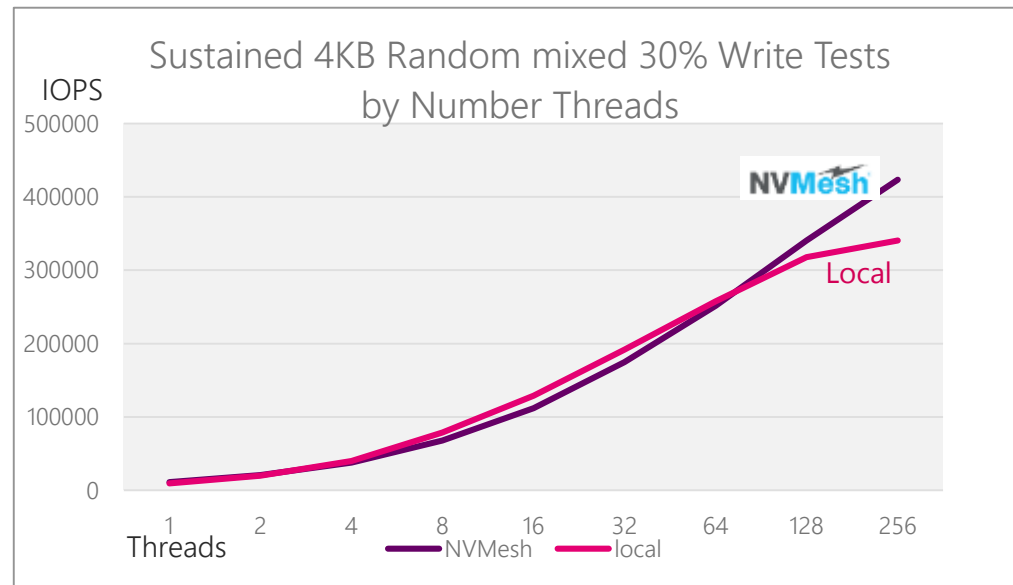
- **ハード**
 - コモディティサーバ & 汎用NVMe
 - インターコネクト Eth or IBサポート
 - RDMA対応NIC(Mellanox)
- **ソフト**
 - Excelero NVMesh Client/Target
 - IBM Spectrum Scale
- **IO**
 - プロトコル : RDMA/RoCEv2
 - データ冗長 : RAID10/Erasure Coding(NVMesh)
 - キャッシング/階層化 : Spectrum Scale

ベンチマーク

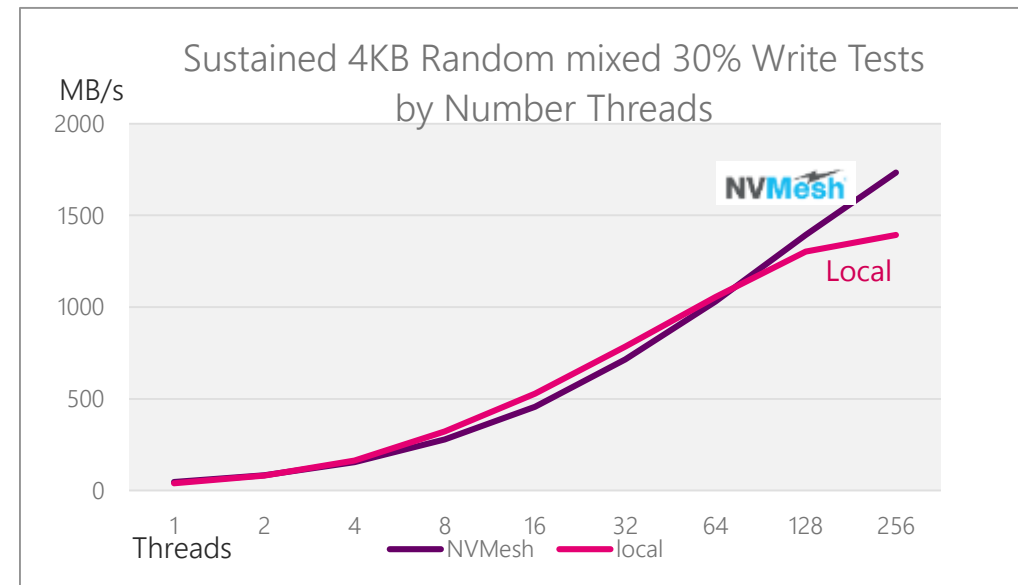
NVMeのローカル性能とほぼ同等のパフォーマンス

※1クライアント vs 1ターゲット(x1NVMe)の性能

※HGST SN150 PCIe SSD使用



IOPS

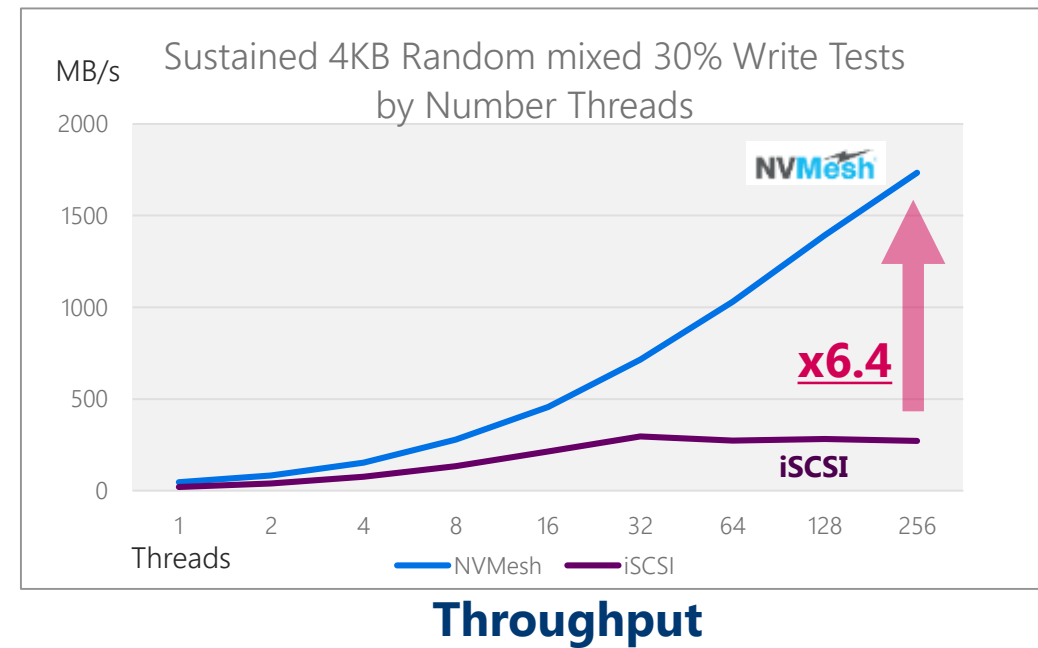
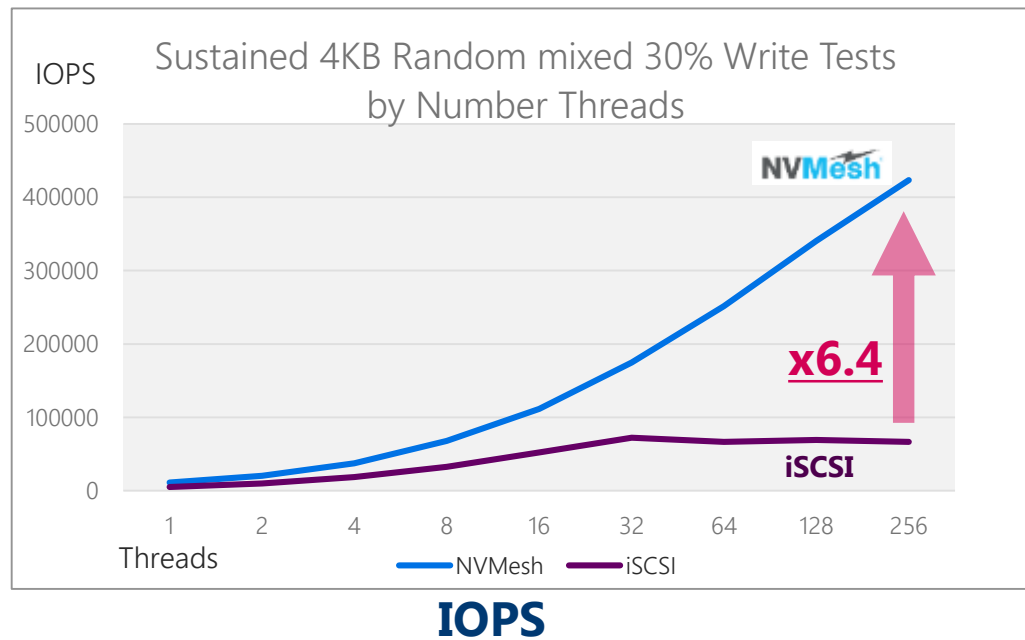


Throughput

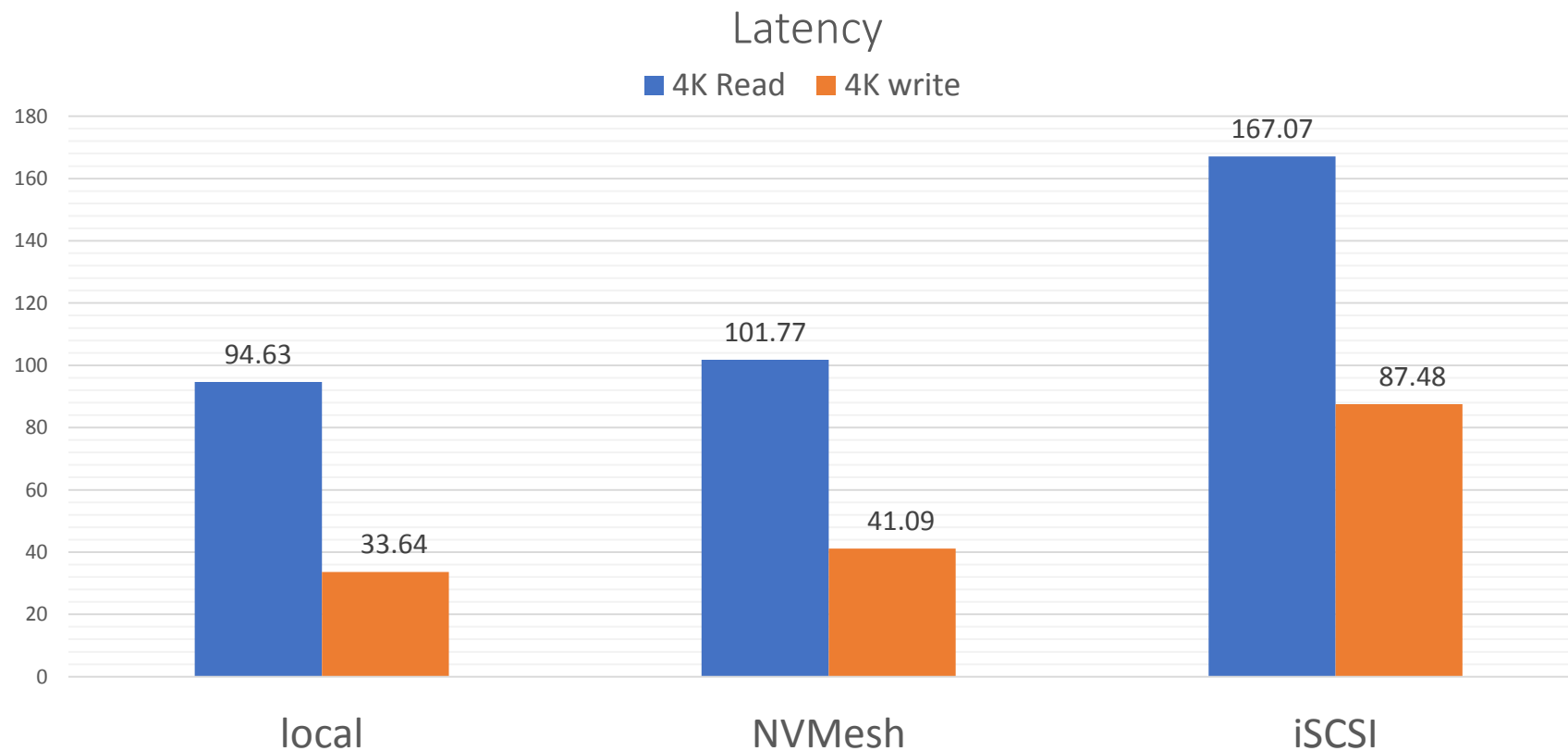
iSCSIに比べ6.4倍の性能差

※1クライアント vs 1ターゲット(x1NVMe)の性能

※HGST SN150 PCIe SSD使用



ローカル≒リモート(+7usec)の低レイテンシー性能

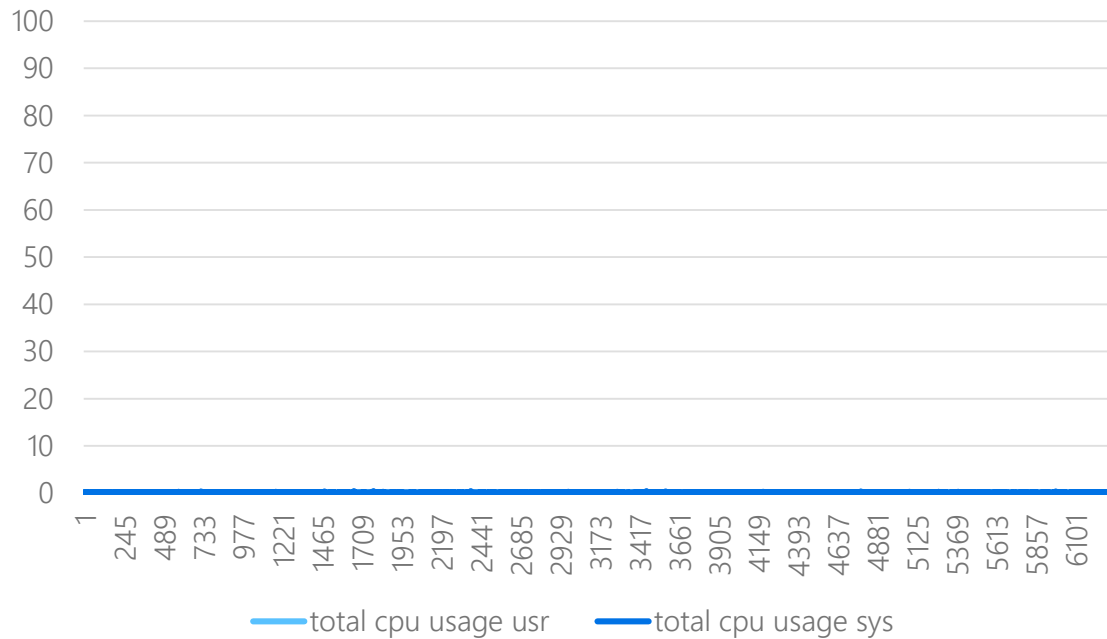


fioツールのアウトプット値

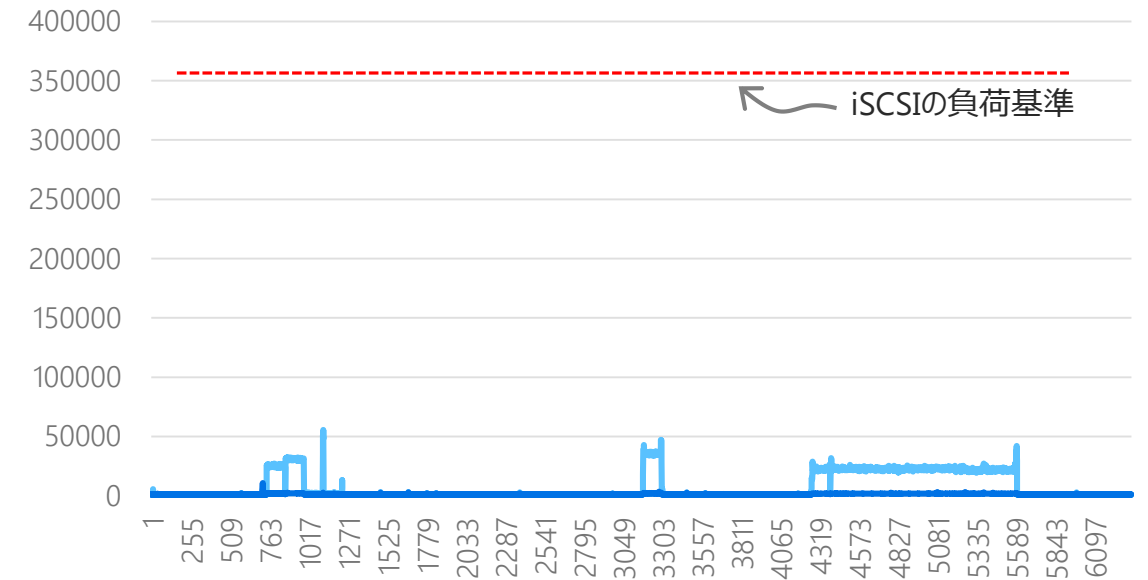
CPU Utilization

CPU使用率ほぼゼロ%、割り込み負荷極少

CPU Utils



割り込み、コンテキストスイッチ



スケーラビリティテスト

検証機器



- ✓ 25/100GE対応 低レイテンシー/ノンブロッキングスイッチ
- ✓ 6.4Tbpsのフルワイヤーレート
- ✓ スマートバッファ搭載(動的バッファリング)
- ✓ RDMA(RoCE)対応アダプタ

ネットワーク



- ✓ HGST社製 Ultrastar Serv24
- ✓ 2U / x24 NVMe SSD搭載
- ✓ 最大容量 184TB

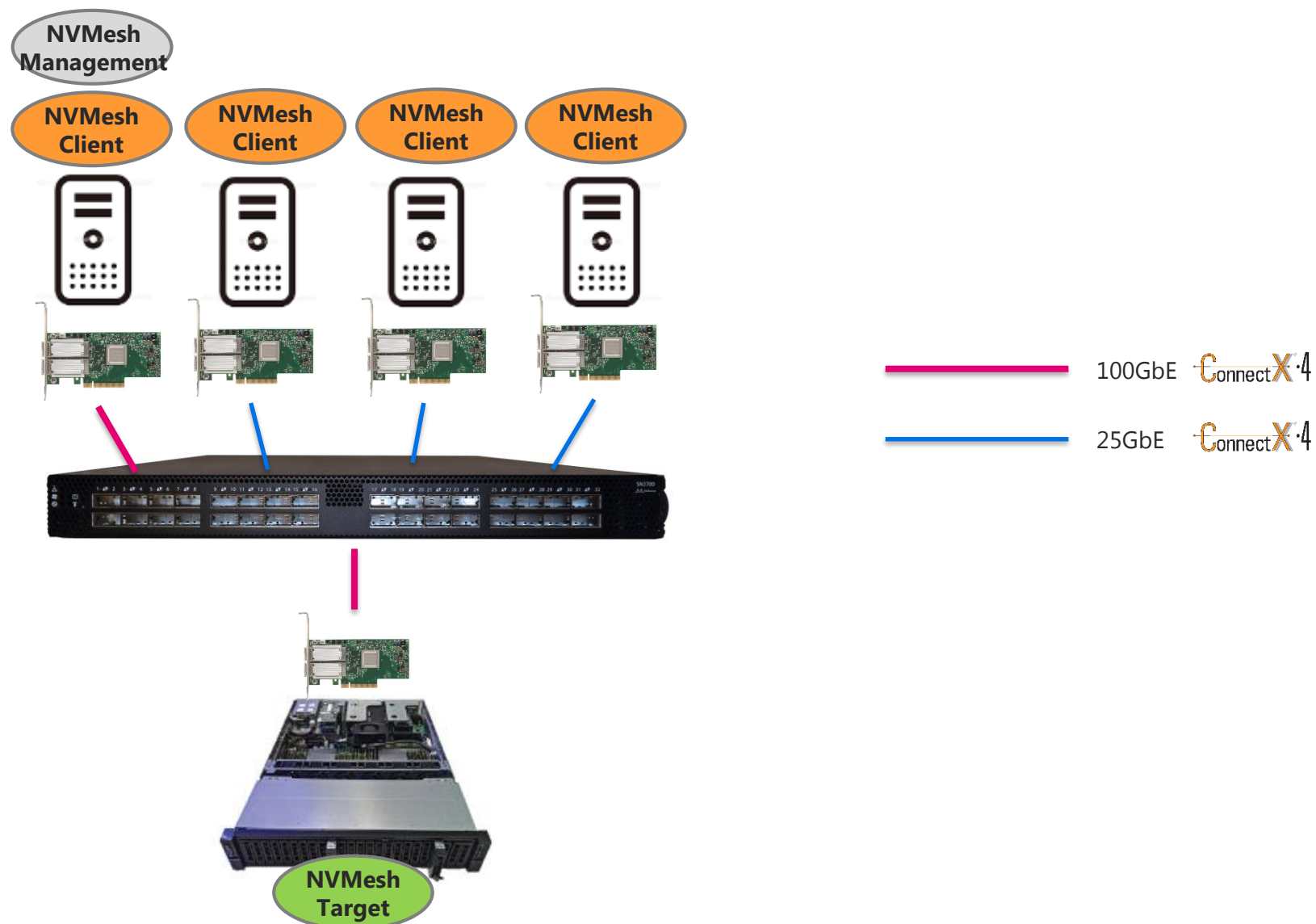
NVMeストレージ



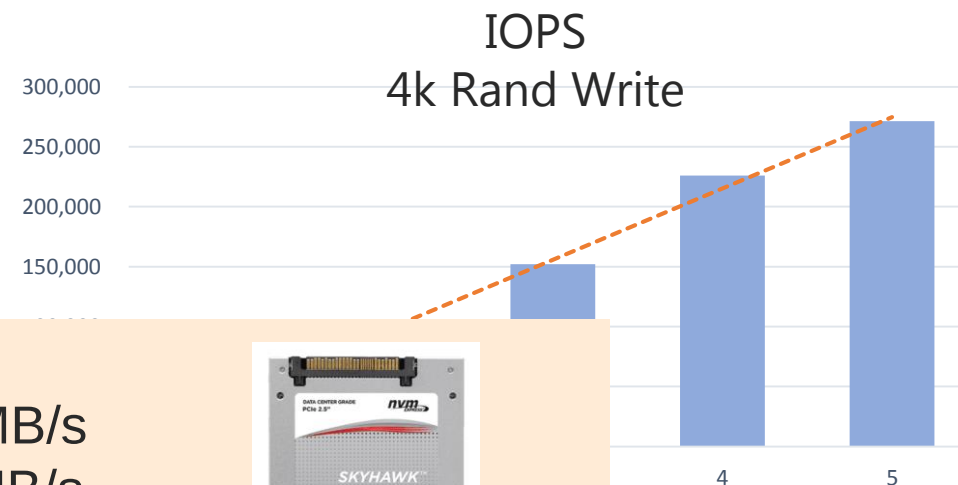
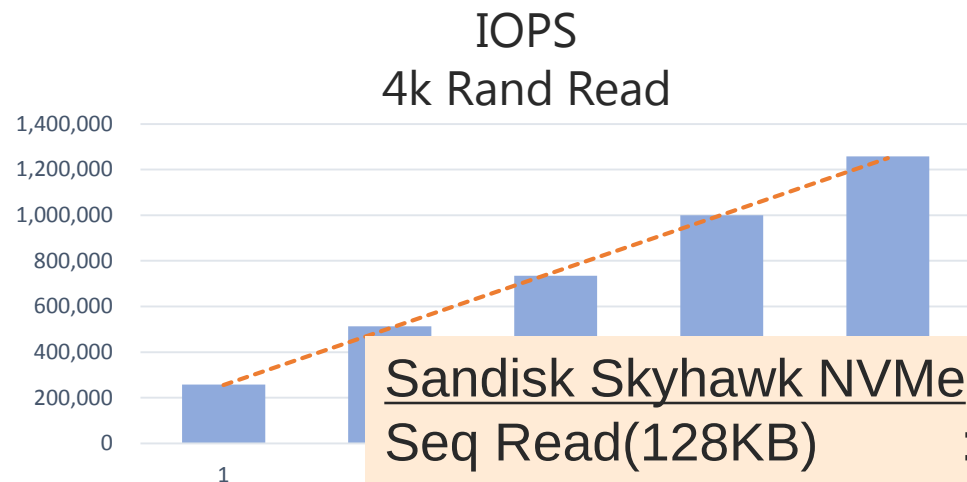
- ✓ NVMe over Fabric対応スケールアウトブロックストレージソフトウェア
- ✓ 高可用性、データ冗長対応
- ✓ RDDA機能による超低遅延性

NVMeofソフトウェア

検証環境

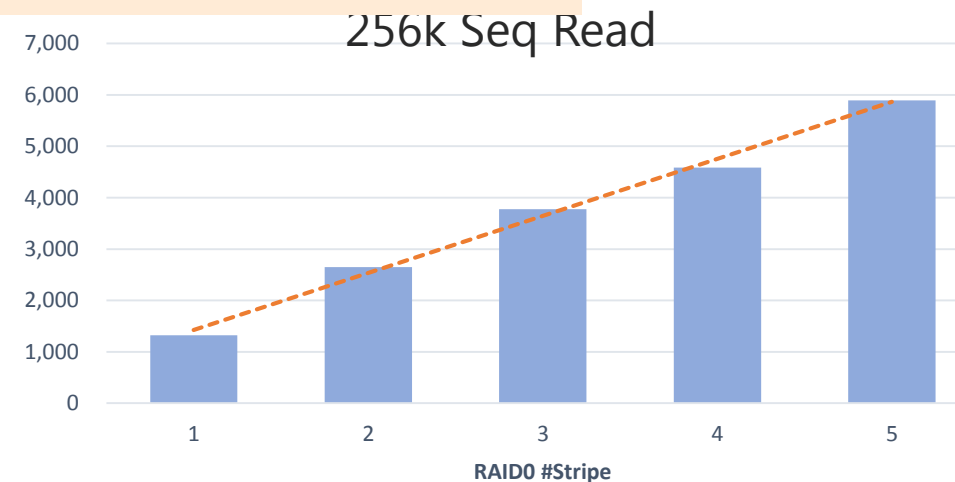
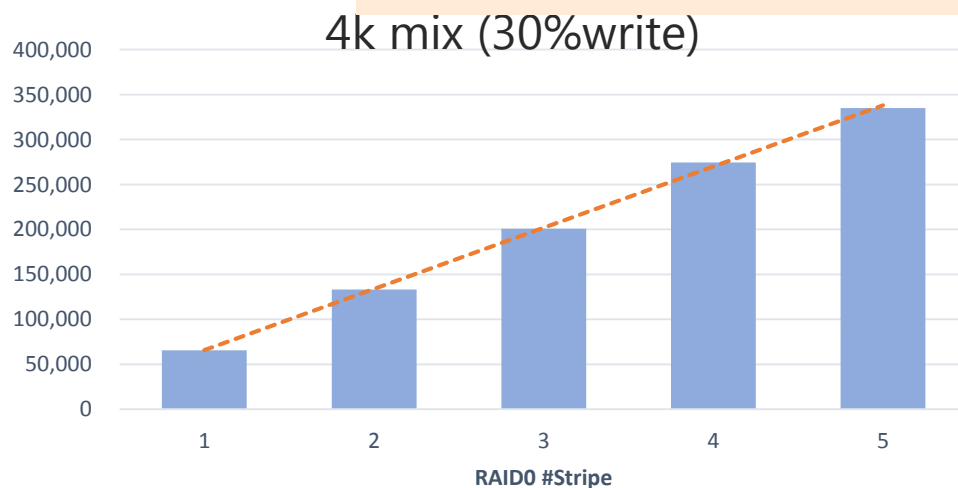


オーバヘッドなしでリニアに性能UP



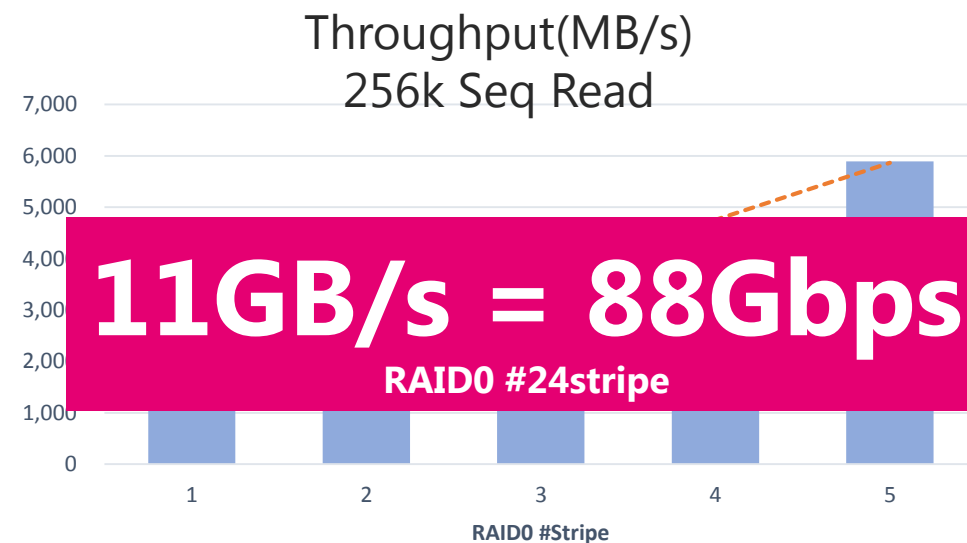
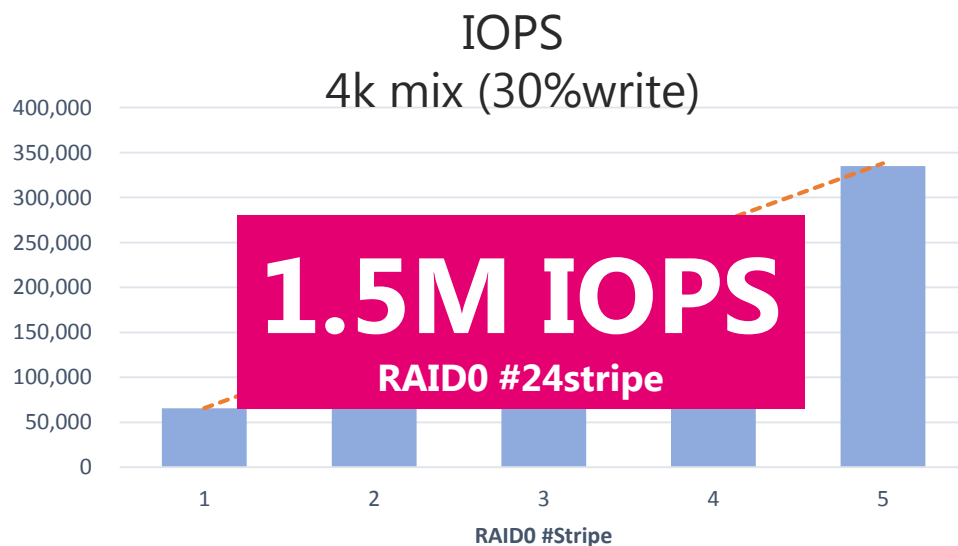
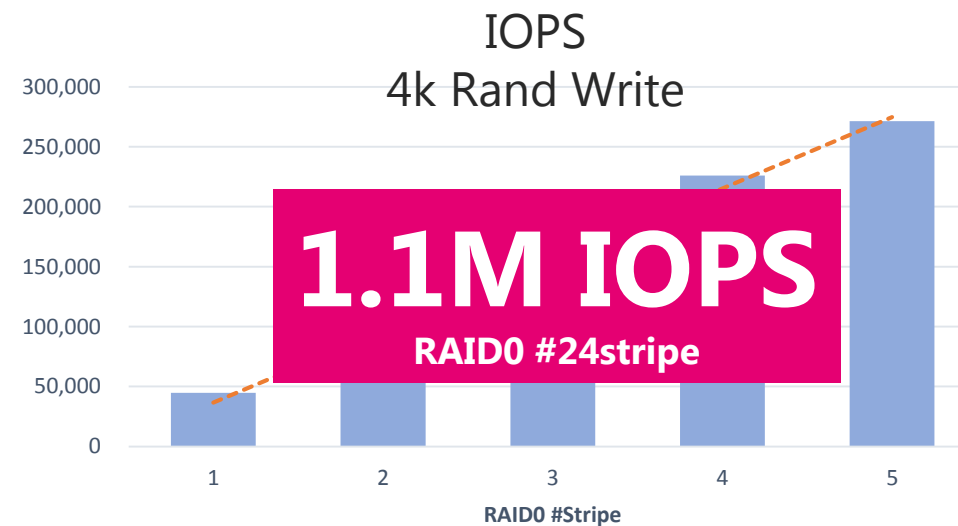
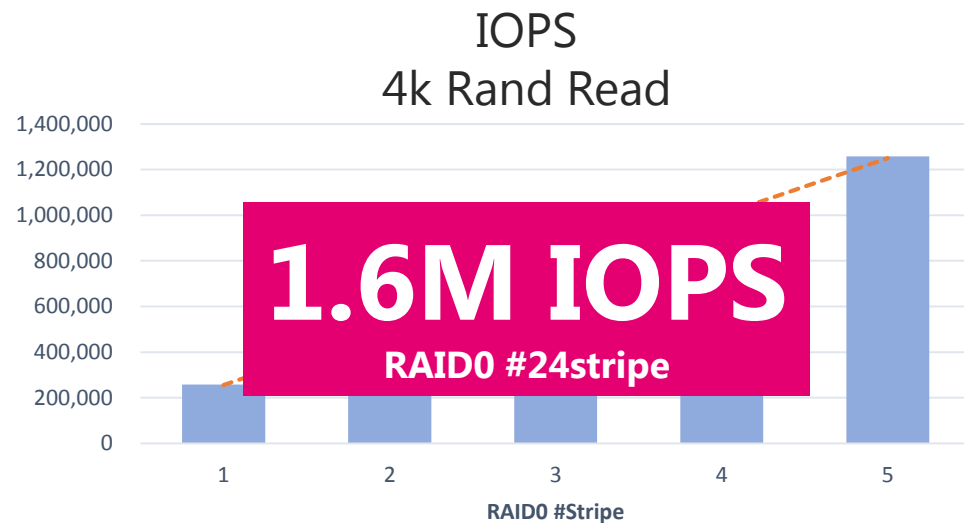
Sandisk Skyhawk NVMe

Seq Read(128KB) : 1500MB/s
Seq Write(128KB) : 1170MB/s
Rand Read(4KB) : 250k IOPS
Rand Write(4KB) : 47k IOPS



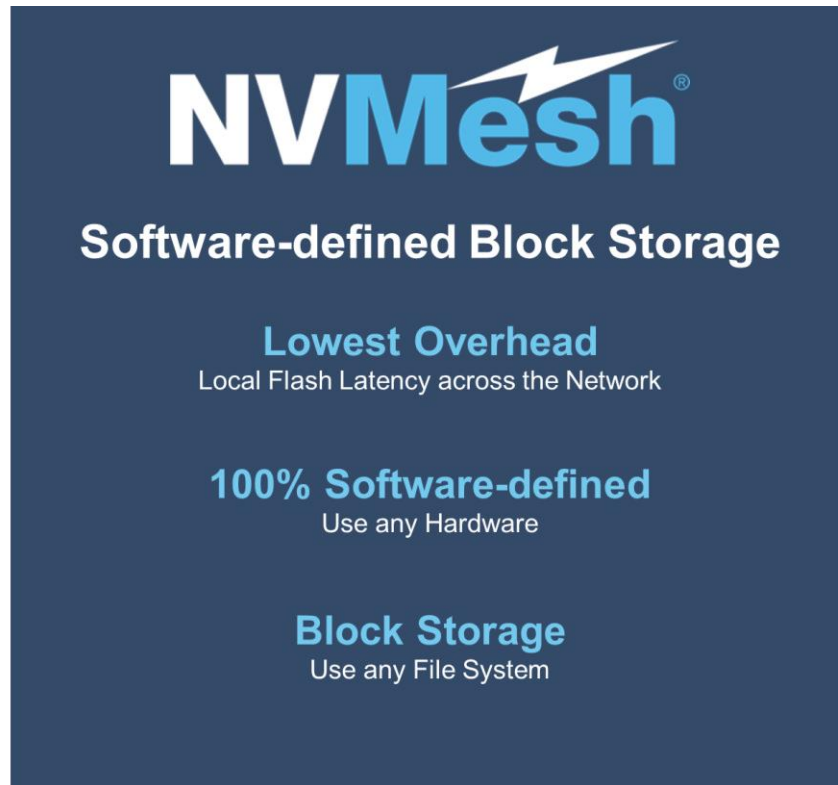
最大パフォーマンス性能

※今回の構成における、RAID0 24ストライプ時



まとめ

性能・スケーラビリティ・可用性を兼ね揃えた
NVMeファブリックストレージソフトウェアは既にReady!!



NVMe Mesh
Software-defined Block Storage

Lowest Overhead
Local Flash Latency across the Network

100% Software-defined
Use any Hardware

Block Storage
Use any File System

HPCアプリケーションを高速化

- ✓ スケールアウト高速NVMeファブリックを実現
- ✓ 高速な並列ファイルシステム

