WekaIOとは？

Overview

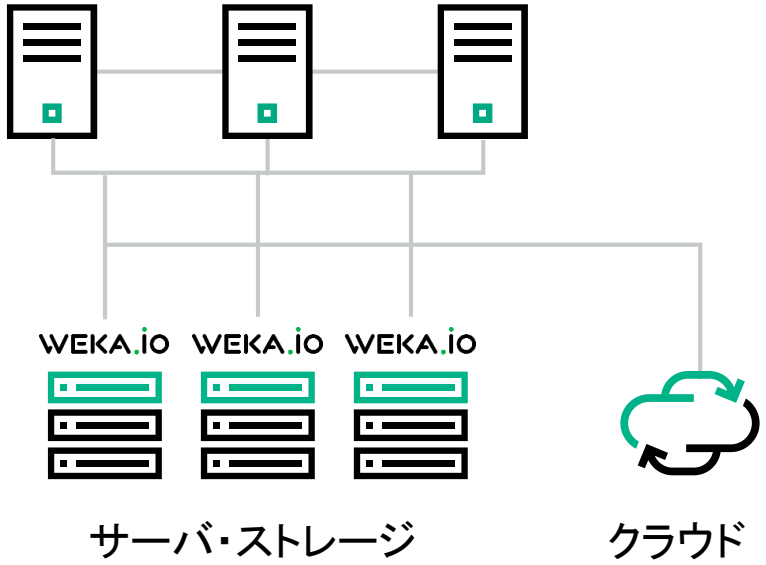
日本ヒューレット・パッカード株式会社
HPC&AI事業統括
テクニカルサポート部
坂詰　唯

2020/2/7

Hewlett Packard
Enterprise

NVMeベースの並列スケールアウトNAS

WekaIO Matrix



高パフォーマンスの共有ストレージを実現

超高性能なパフォーマンス
パフォーマンスとスケーラビリティ

Hybrid IT
バースティング時のクラウド活用

TCOの削減
データの階層化管理が可能

高いパフォーマンスとスケーラビリティ
次世代のストレージ・ディマンドに対応

高帯域幅での効率の良い性能

- 最大NFSの7倍の性能を実現
- Apollo 6500 + Resnet50と組み合わせたソリューションで4万枚/秒の画像処理を可能に。

低レイテンシー

- NVMeを有効活用することで低レイテンシーを実現
- SSDsに最適なドライバーの実装

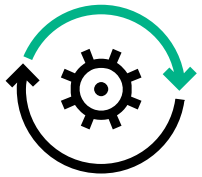
細かいファイルのやり取りにおいても性能を担保

- 細かいファイルのやり取りのボトルネックを排除したメタデータの配布と低レイテンシー

*https://www.weka.io/wp-content/uploads/2018/03/AI-Use-Case-W01UC201803.pdf
**http://dlpg.labs.hpe.com/

Hybrid IT：バースティング時のクラウド活用
場所を選ばない、柔軟性と可用性を提供

まとまったデータの共有



- オンプレ - クライド間の一貫したデータのやり取りを実現
- クライドとオンプレの同期を簡単に実現

災害対策

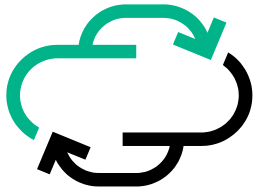
- バックアップデータをS3及びクラウドに保管することが可能
- スナップショットの活用により、柔軟なリカバリーを実現

柔軟な拡張

- 8ノード構成で安価にスモールスタートさせることが可能
- クラウドからオンプレへのマイグレーションも可能

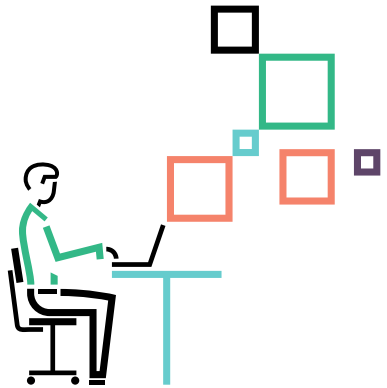
TCOの削減： データの階層化が可能
コールドデータをオフロード

データの階層化



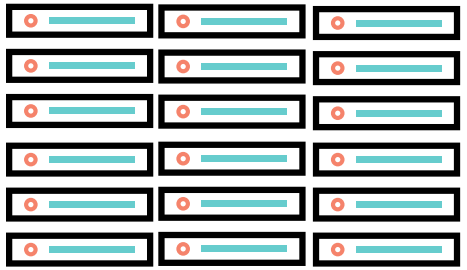
- コールドデータをオフロードさせることでGB単価を安く提供
- ポリシー管理をベースとした自動での階層化管理

データの保護



- 様々な要因からのデータ保護をサーバの観点から実現

拡張性



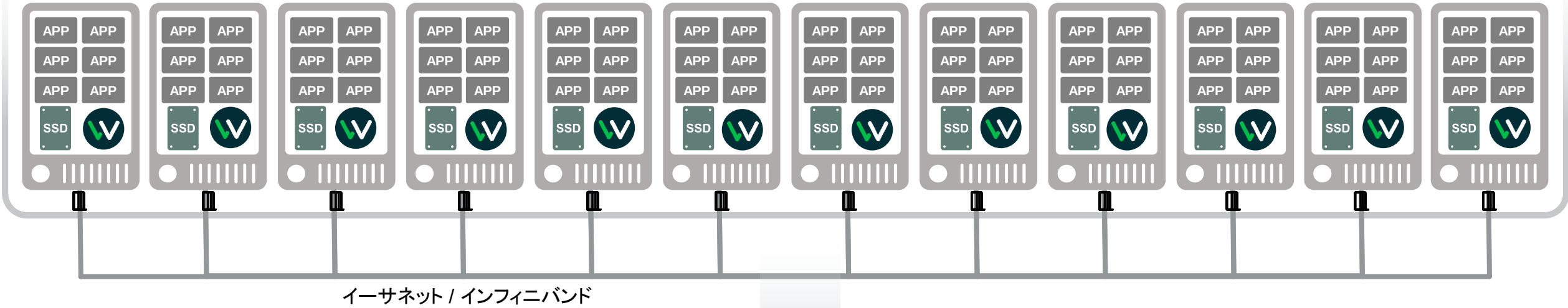
- 容量と性能の拡張を簡単に実現

インフラの最小化

- ストレージ構成の最適化
- NFS, SMB, HDFS, and S3との接続性を担保

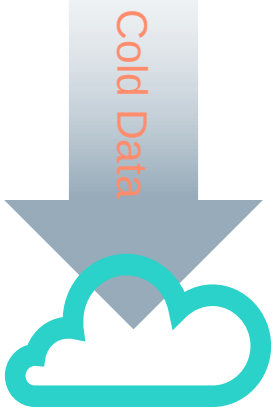
WekaIO 並列ファイルシステムの仕組み

柔軟な構成方法と、クラウドへの拡張性



イーサネット / インフィニバンド

Apollo2000/DL360をストレージサーバーとして構成することも可能



SCALITY ceph DDN STORAGE Hewlett Packard Enterprise S3

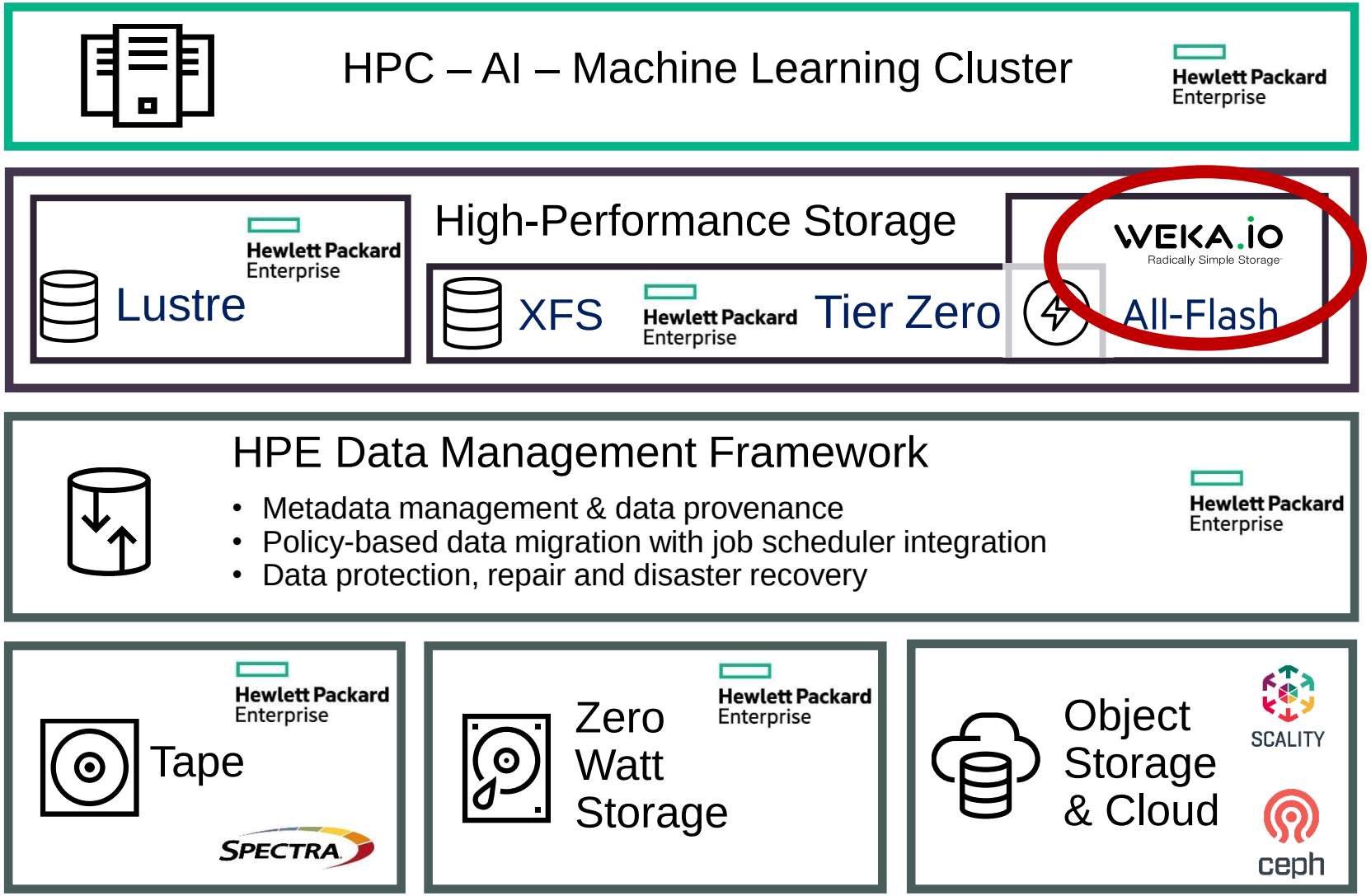
- シングルネームスペースでの管理可能
- ノードの使い方は3通り
  - ✓ FSクライアント
  - ✓ FS サーバー
  - ✓ その両方

*Roadmap to HPE DMF Integration

Hewlett Packard Enterprise

HPC & AI | **Data Management**

Software Defined Storage



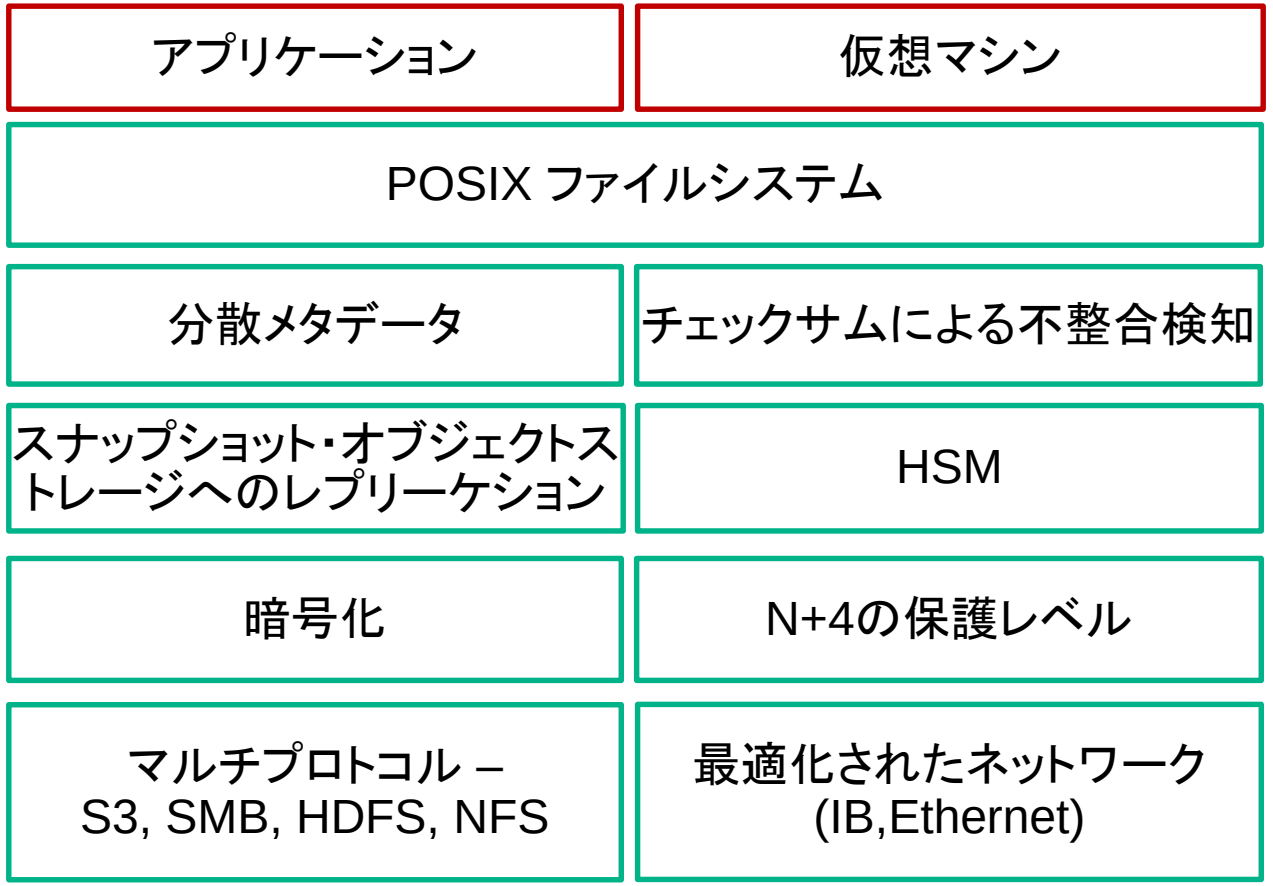
WekaIO のアーキテクチャ

Technical Overview

Hewlett Packard Enterprise

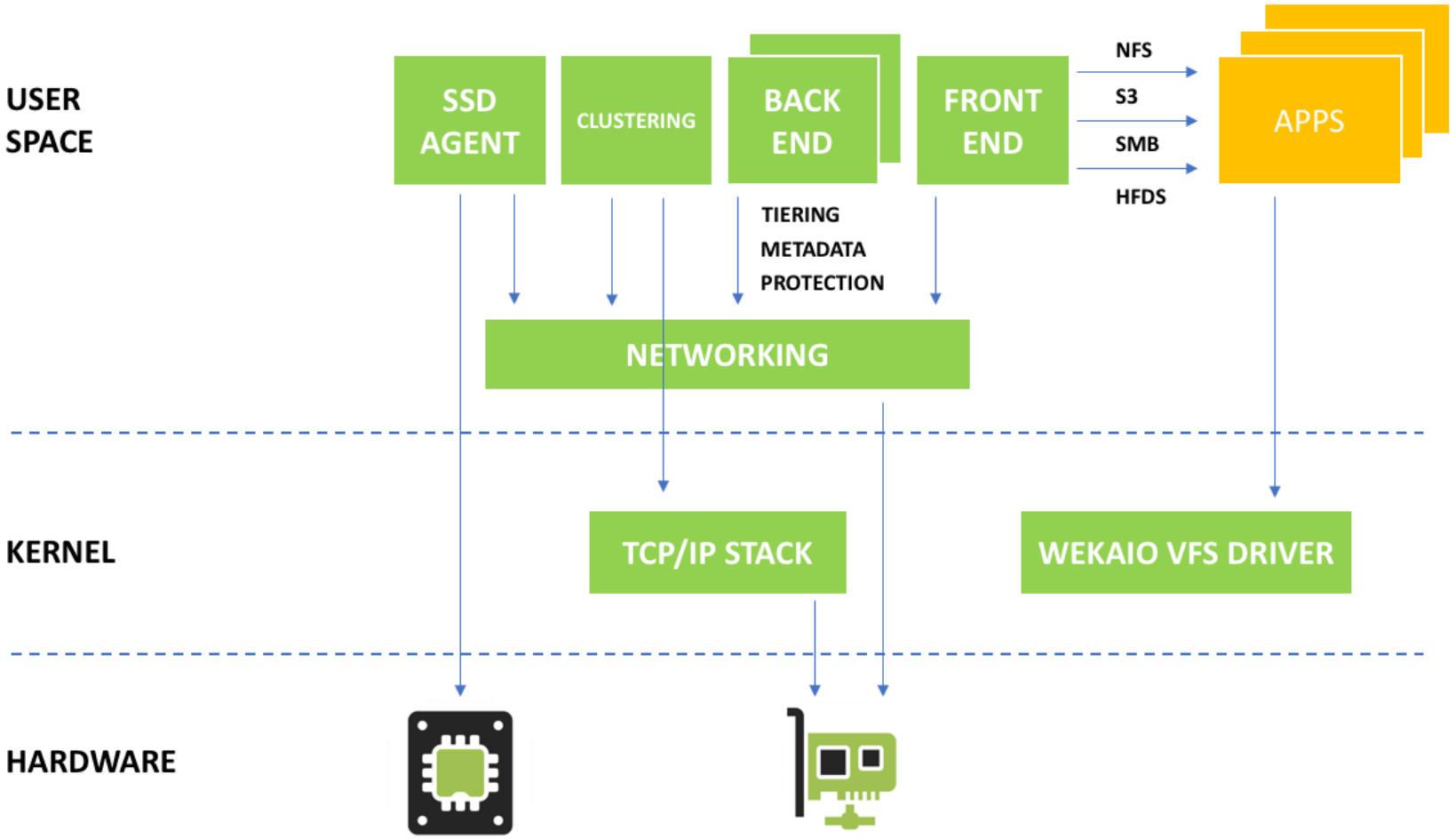
WekaIO Matrix FSの構成

- フラッシュに最適化されたファイルシステムで、高帯域及び低レイテンシーを実現
- 億単位のファイルとEBの容量を管理することが可能な分散メタデータのアーキテクチャ
- 特許取得のデータ冗長化と迅速な復旧、高い容量効率
- スナップショット/レプリケーションでのデータ保護
- S3と連携した階層化管理を実現させることが可能



WekaIOのアーキテクチャ

- 標準的なプロトコルとして、SMB、NFS、HDFS REST(S3)をサポート。これらのプロトコルのサポートはWekaIOのkernel moduleとして実装されている。
- クライアントにWeka VFSドライバを用いることでPOSIXファイルシステム準拠のAPIと高いパフォーマンスを提供。(MatrixFS)



WekaIO Matrix Distributed Data Protection (MatrixDDP)

イレージャーコーディングの冗長性と柔軟性をそのままに、計算負荷を低減を実現

- N+2またはN+4の保護レベルを選択可能
- ストライプサイズは4~16に任意に選択可能(ただし、ノード障害の許容の範囲)
- スループットへの影響がほとんど無い
- 障害時に数分程度で一つ下の保護レベルに復帰する
- リビルドに割り当てるバンド幅を任意に変更可能

Stripe Size(data+parity)	16+2	16+2	16+4
Number of Nodes	50	100	100
Capacity per Node	10TB	10TB	30TB
Rebuild Bandwidth (GB/Second)	0.5	0.5	1.0
Time to Full Protection (h:m:s)	2:54:34	1:26:22	2:25:50
Time to 1st failure resiliency (h:m:s)	1:00:37	0:14:52	0:28:00
Time to 2nd failure resiliency (h:m:s)			0:05:11
Time to 3rd failure resiliency (h:m:s)			0:00:55

Table 1 - Sample Rebuild Time to Return to Protected Level

Hewlett Packard Enterprise

End-to-End Data Protection(EEDP)

- データ書き込み時にチェックサムを計算し、書き込み後にverifyを行う。不整合が検出された場合、そのデータはコミットされない。
- チェックサムを保存し、読み出し時に整合性を再度チェックする。
- ジャーナリングファイルシステムを採用しており、停電時にfsckが不要。

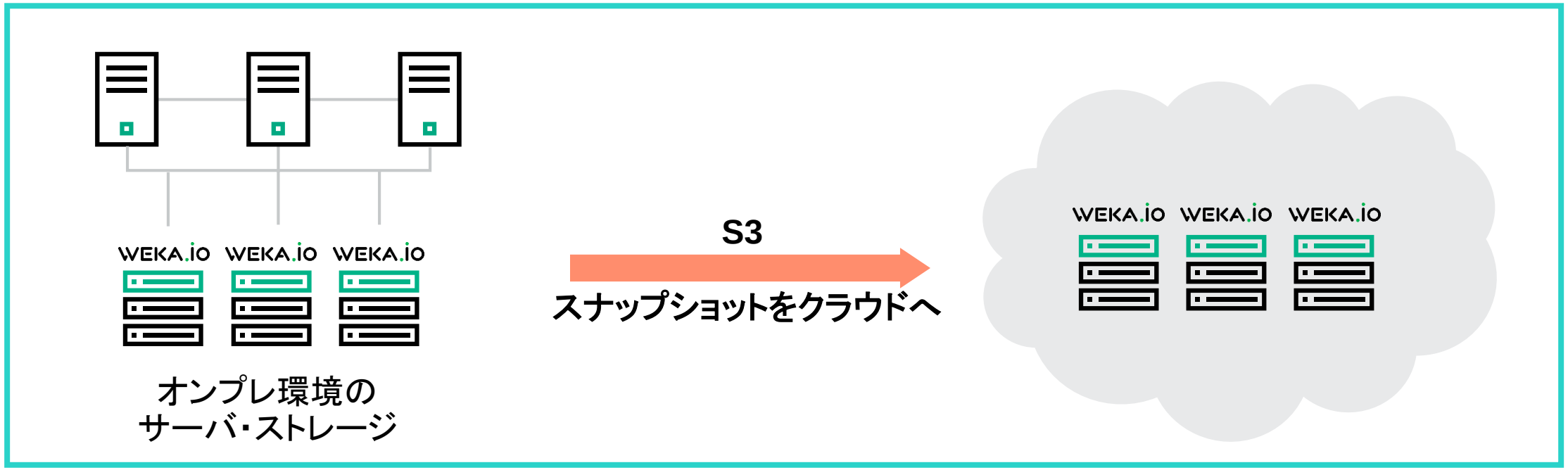
Full end-to-end encryption

- 高い暗号強度 (XTS-AES 512bit keyをサポート)
- KMIPS準拠のキーマネージメントシステムをサポート
- ファイルシステム単位で暗号化ポリシーを設定可能
- クライアントからオブジェクトストレージに至るまで、end-to-endの暗号化

Automated Data Rebalancing

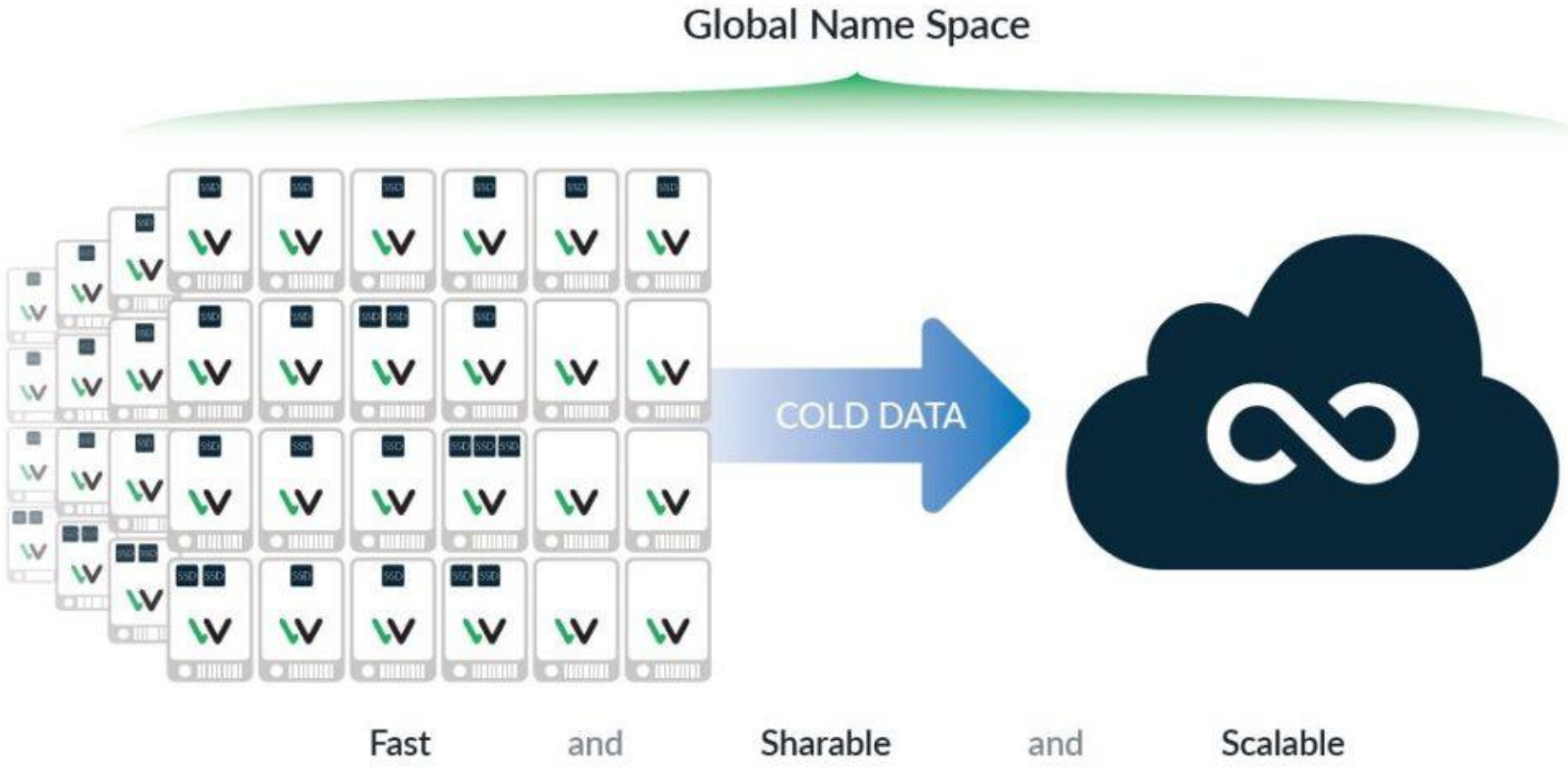
クラスタノードの平均容量を大きく超えているノードは、データの再分配により自動的にリバランスされる。

スナップショット ＆ クラウドへの連携



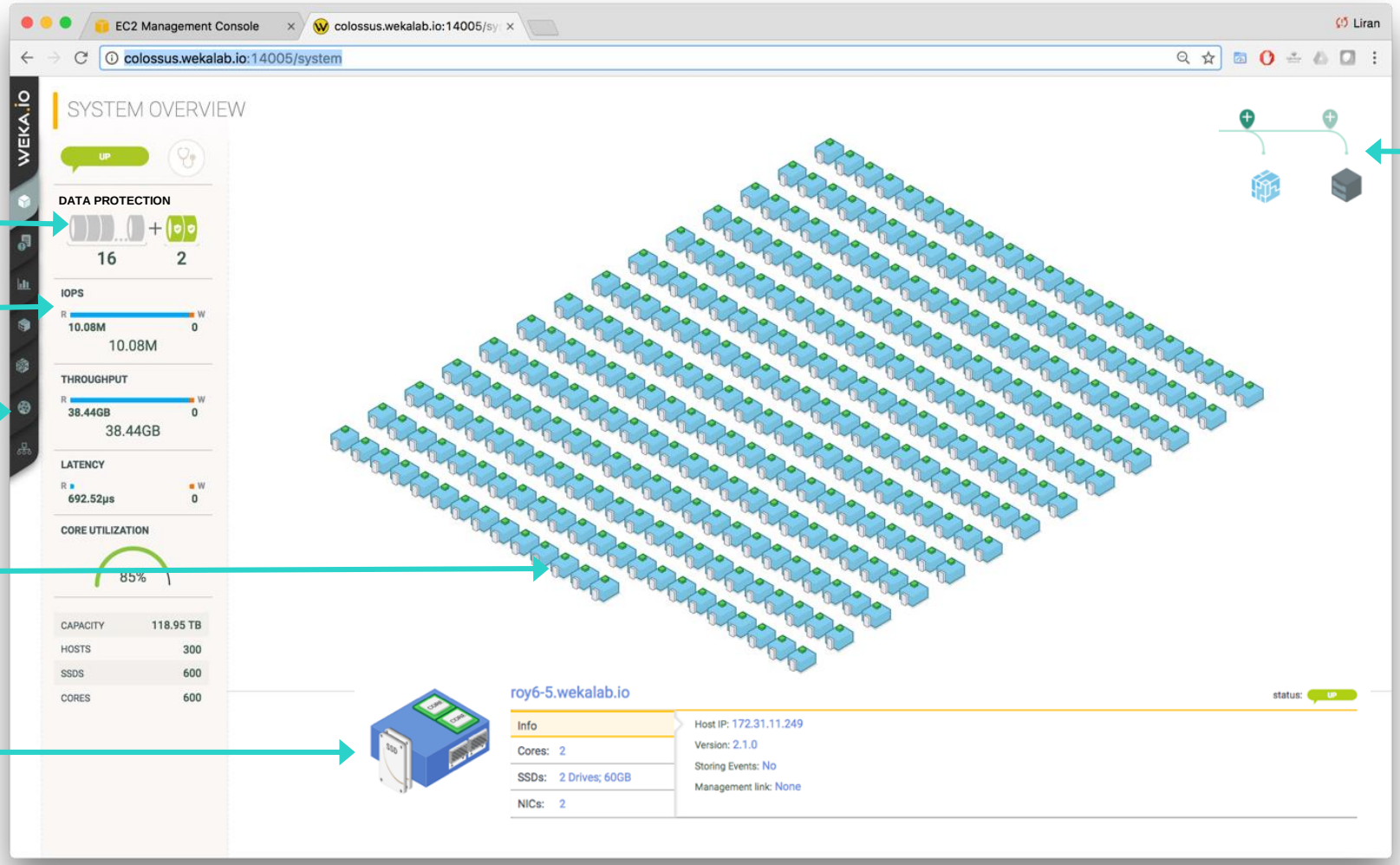
定期的にスナップショットを取得し、オブジェクトストレージへレプリケーション可能
オンプレとクラウドで共通したネームスペースを使用

階層化管理



階層化のポリシー
- ファイルシステム単位
- ファイル更新からの経過時間
- 暗号化の設定が可能

管理を容易にするGUIとCLIの管理ツール(Trinity)を提供



データ保護

パフォーマンス
(IOPS)
スループット管理

サーバ・
クラスター

サーバの詳細

オブジェクト
ストレージ

WekaIO Limitation

- 最大1024個のファイルシステム
- 最大6.4兆個 (T) ファイル/ディレクトリをサポート
- 最大8EBの容量保管を実現（SSD上は512PBまで）
- 1ディレクトリ当たり、6.4億個のファイルを格納可能
- ラージファイルは4PBまでをサポート
- Tiering先は8個

Licensing

Hyperconverged Licensing

Base license

ノードをクラスターに参加させるには、各ノードにノードロックライセンスが必要。標準で2CPUコアまで使用可能なライセンス、計算ノードや、Matrix server licenseと組み合わせてobject storage S/Wと併用する場合に使用される。

MatrixEP(Extreme performance) license

2コア以上を使用したい場合、MatrixEP licenseを導入する事でCPUコア数の制限がなくなります。

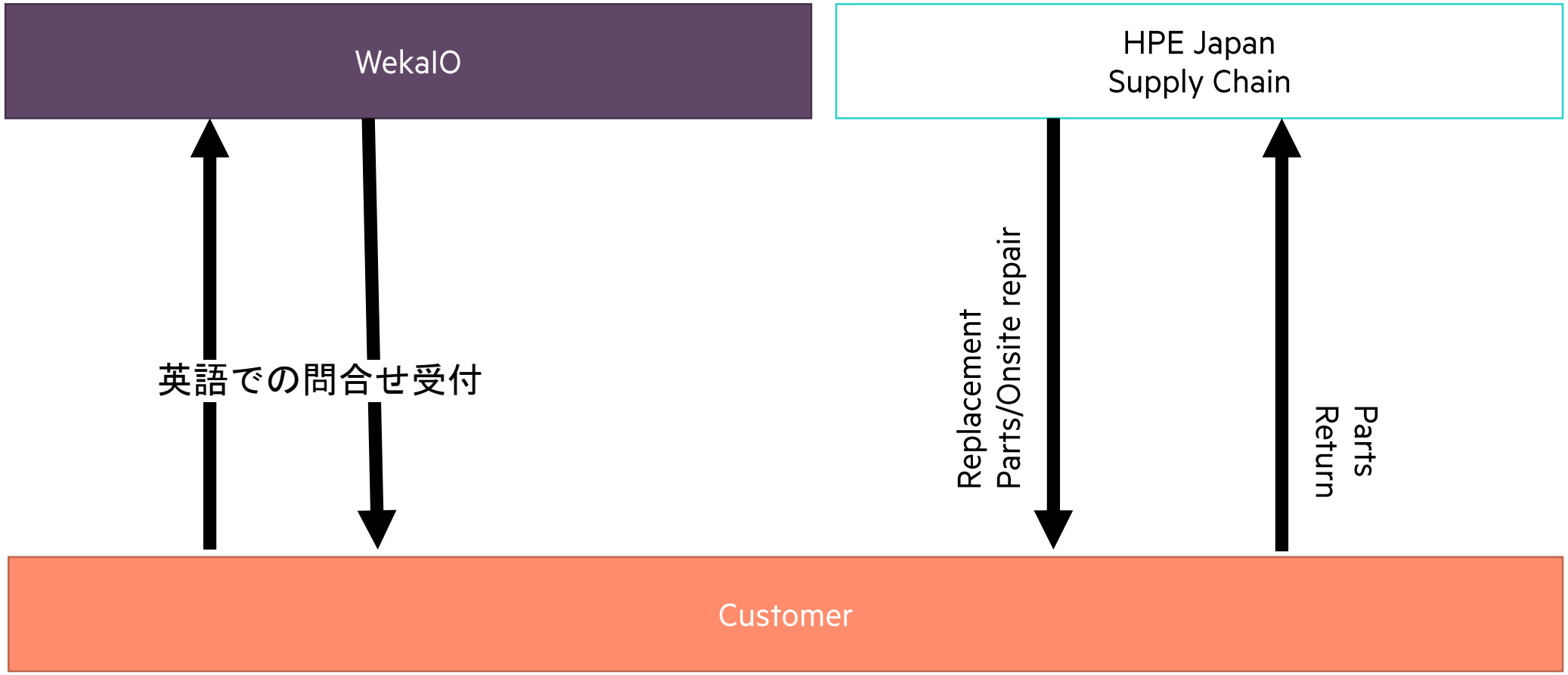
Matrix server license

階層化を行う場合に必要となるライセンスで、object storageに保存されたデータ量[TB](storageの容量ではない)で課金されます。

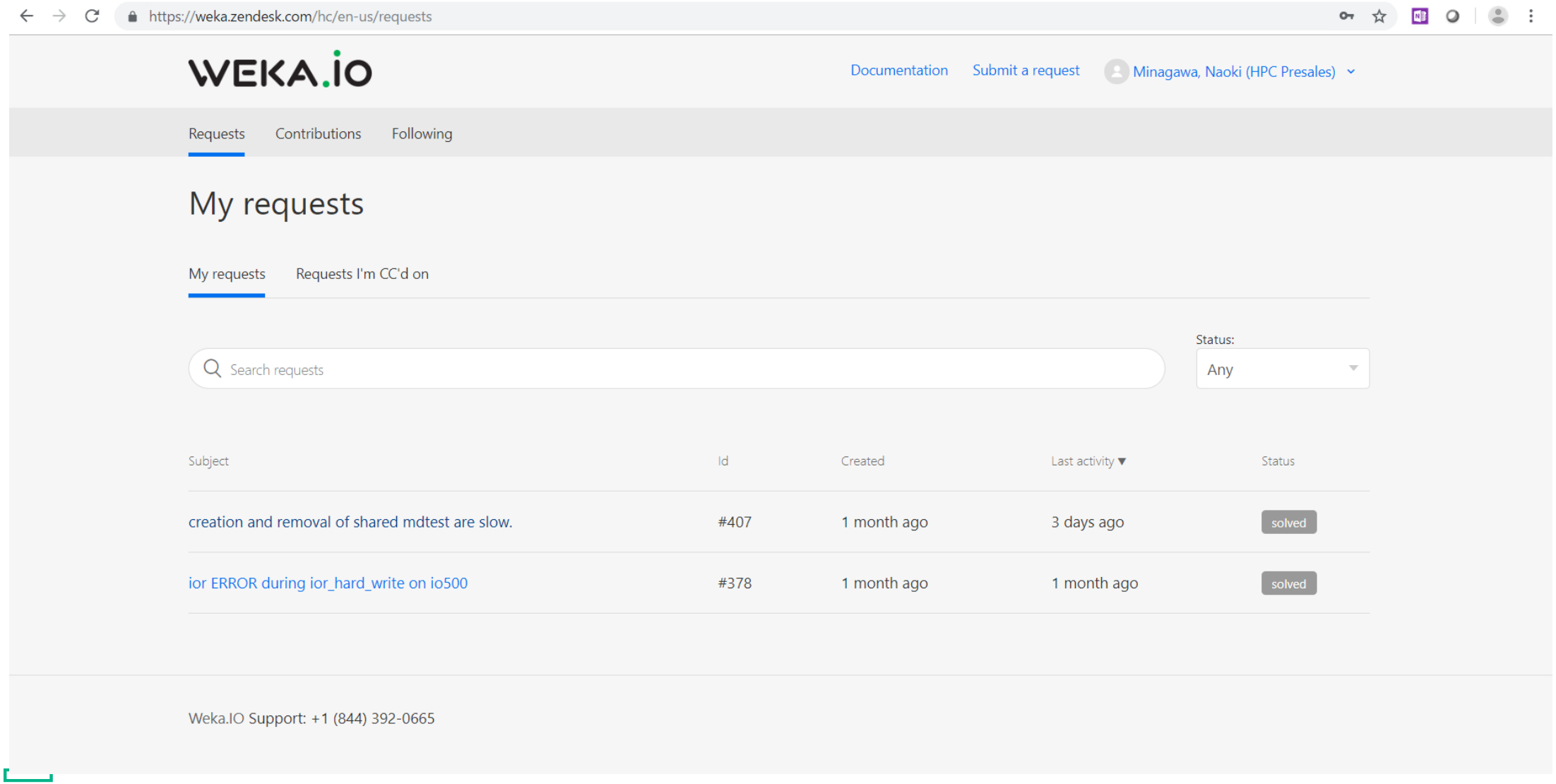
サポート体制

Hewlett Packard
Enterprise

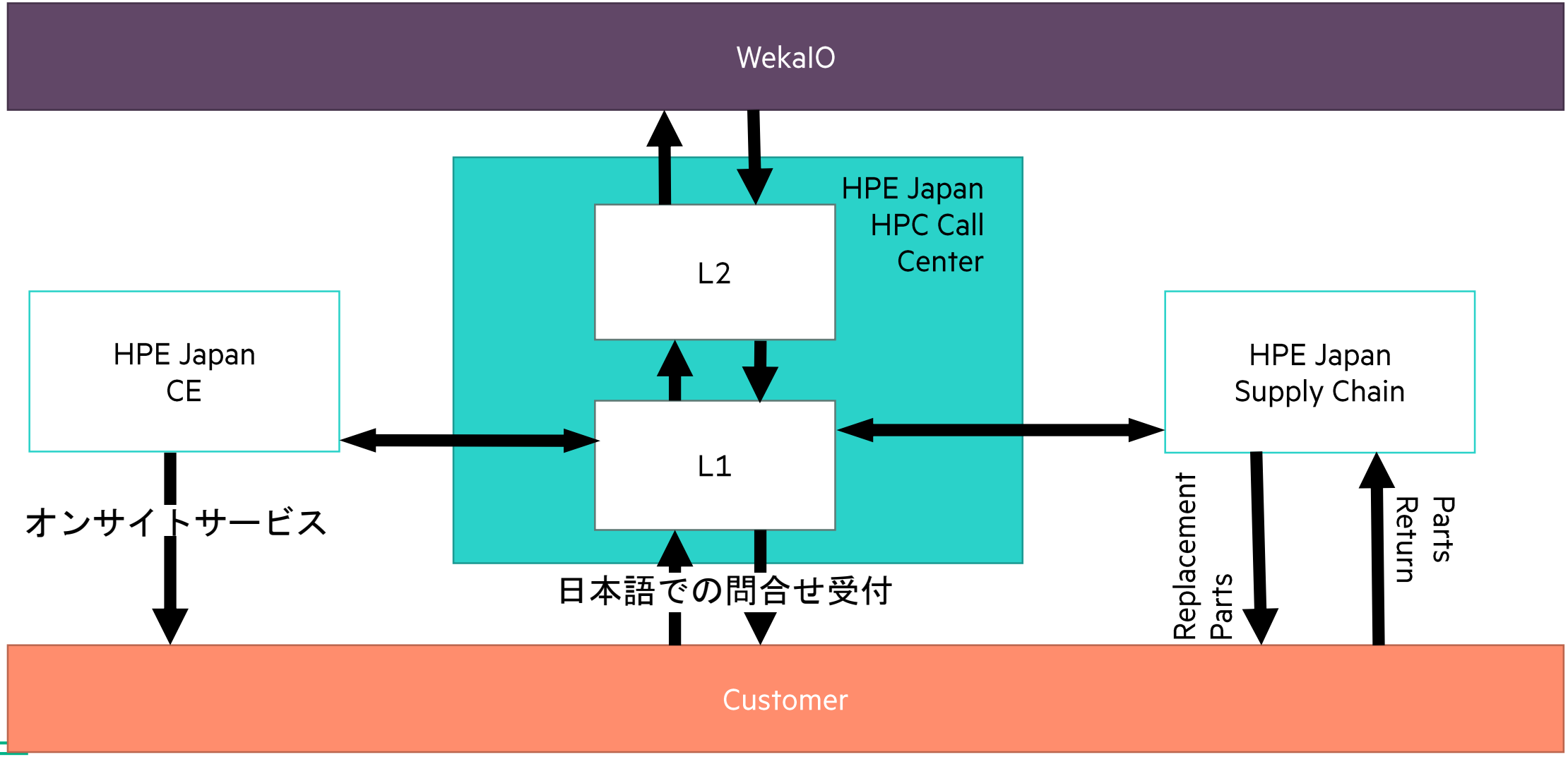
WWのSupport Model



WekaIOのサポートページ



WekaIO 日本語 Support Model

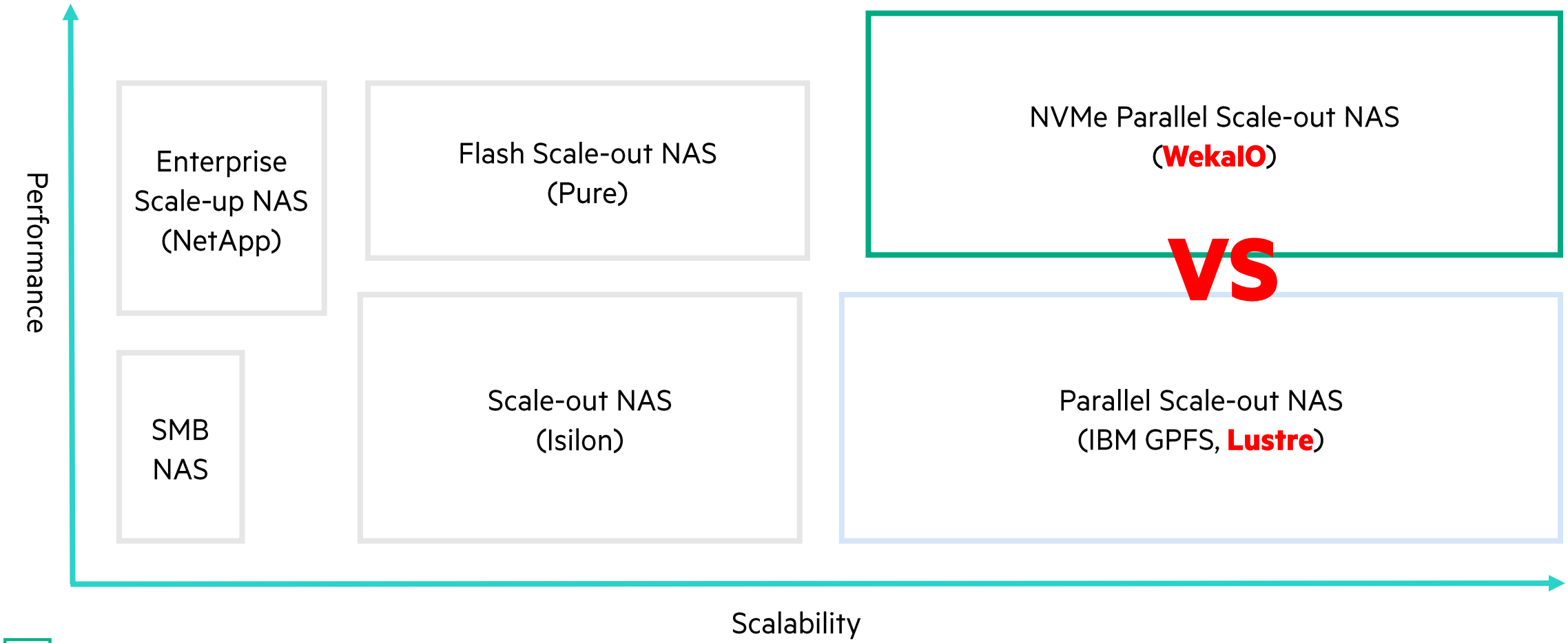


Performance Report

Hewlett Packard
Enterprise

WekaIO VS Lustre

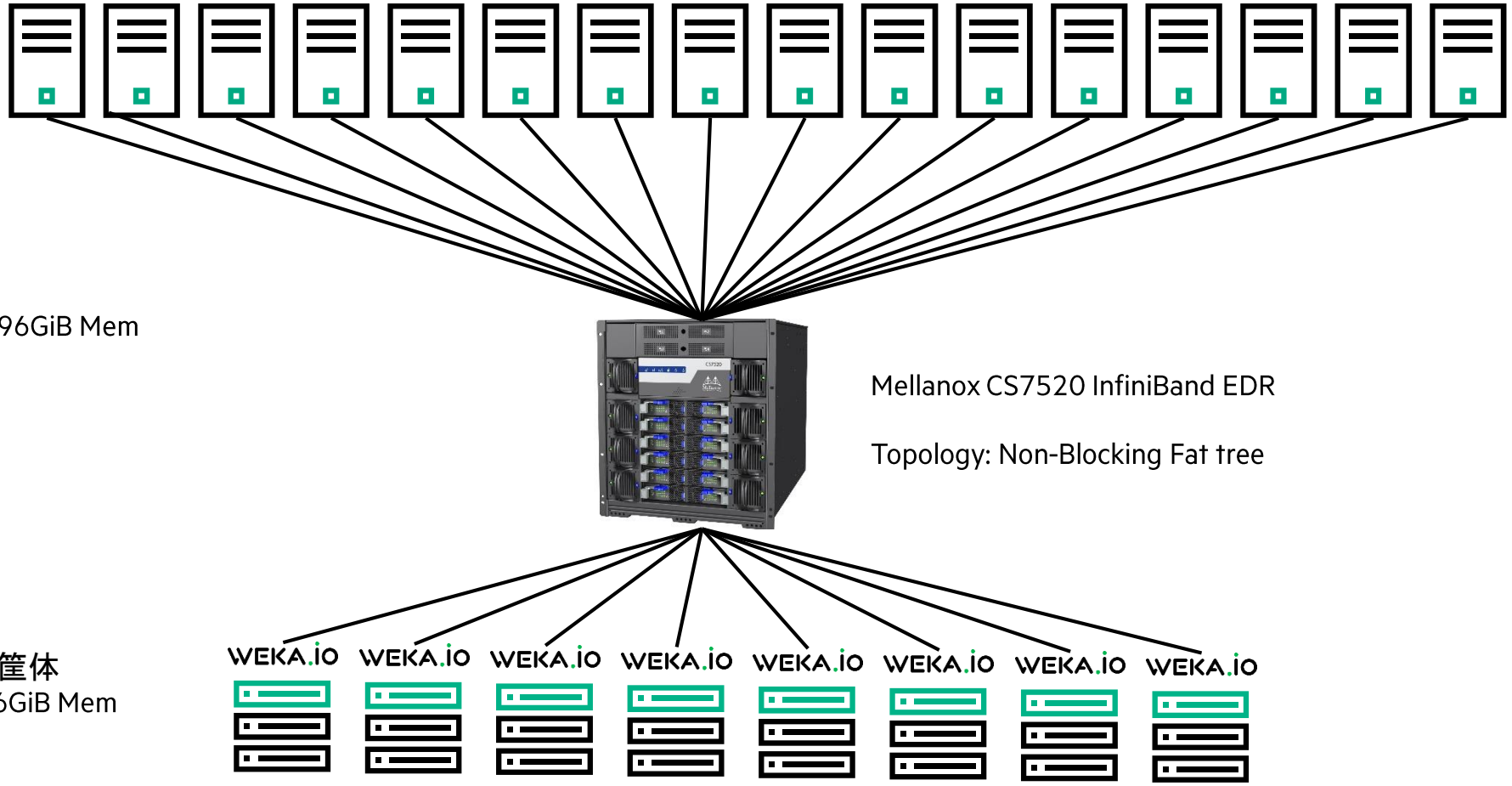
WekaIOとHPC分野でよく使われるLustreを比較した。Lustreは基本はHDDを使うが、NVMeを使うものも増えてきている。特にAI分野においてはNVMeを用いたLustreが多い。



WekaIOテスト環境

計算サーバ(Apollo2000 Gen10) x4筐体
XL170r(Intel Gold 6134(3.2GHz 8core)x2 96GiB Mem
HyperThreading off) x16サーバ

ストレージサーバ(Apollo2000 Gen10) x2筐体
XL170r(Intel Gold 6134(3.2GHz 8core)x2 96GiB Mem
HyperThreading off) x8サーバ
Intel P4600 1.6TB x4/サーバ
protection: 4+2
drive storage: 24.83 TiB total



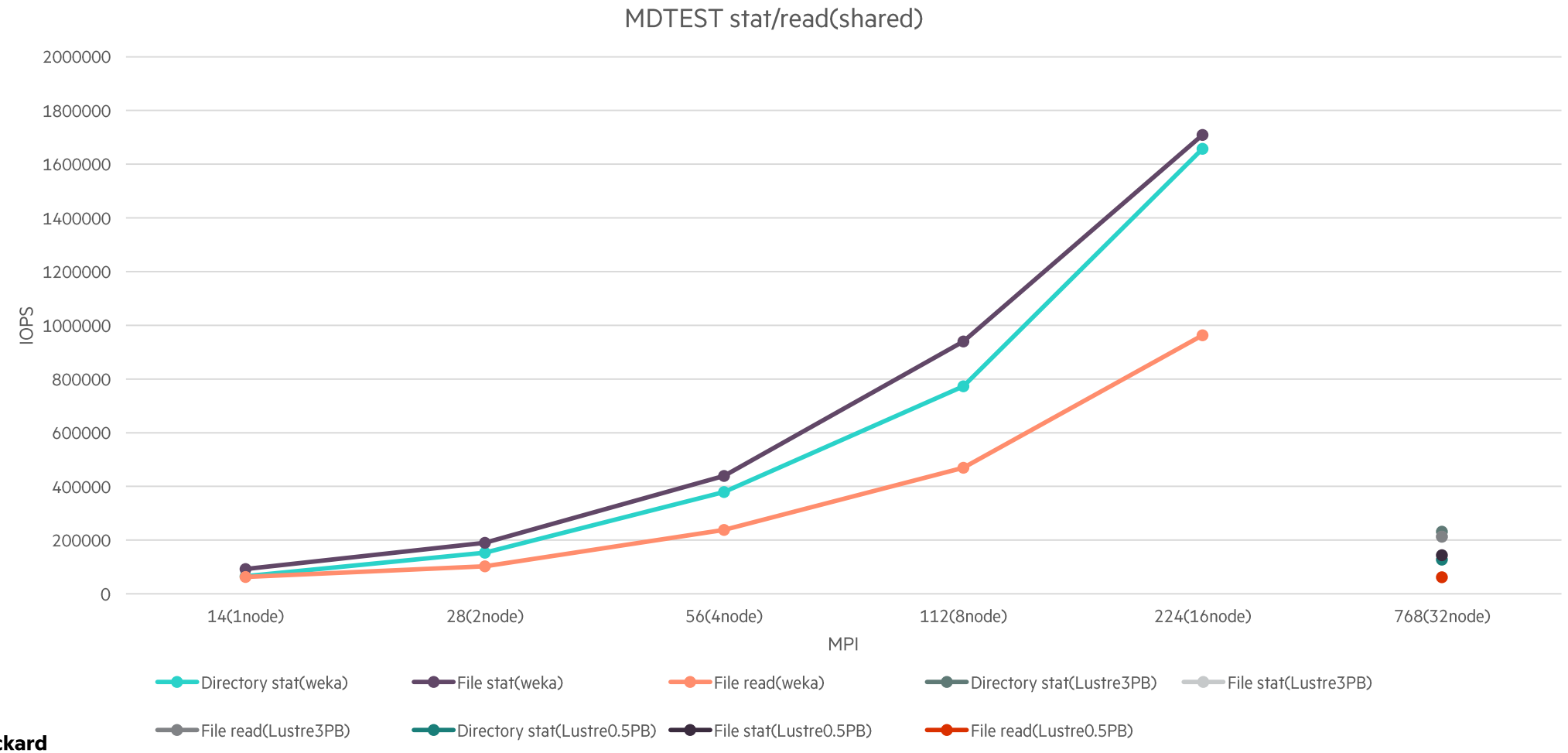
Mellanox CS7520 InfiniBand EDR

Topology: Non-Blocking Fat tree

Hewlett Packard Enterprise

IOPS(MDTEST)

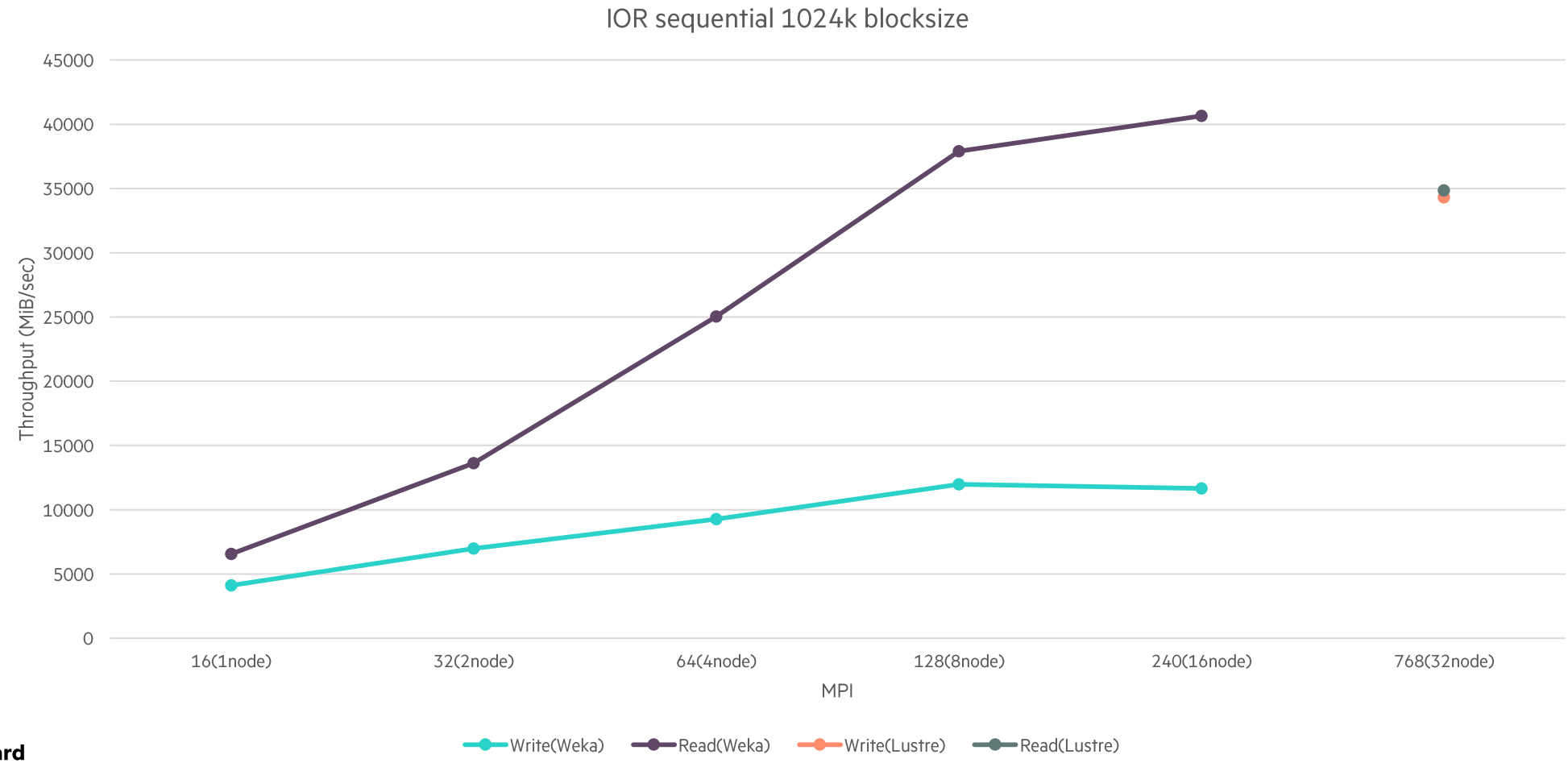
stat/readはLustreと比較し非常に速い。File readで約4.5倍、statは約7倍速い。create/removalに差は無かった。
Lustreはクライアント32ノード768プロセスでの実行に対し、WekaIOではクライアント16ノード224プロセスで実行している。Lustreは3PB(MDT: HDDx20 OST: HDDx400)、0.5PB(MDT: SSDx4 OST: HDDx120)の結果を示している。



Hewlett Packard Enterprise

IOR(sequential throuthput)

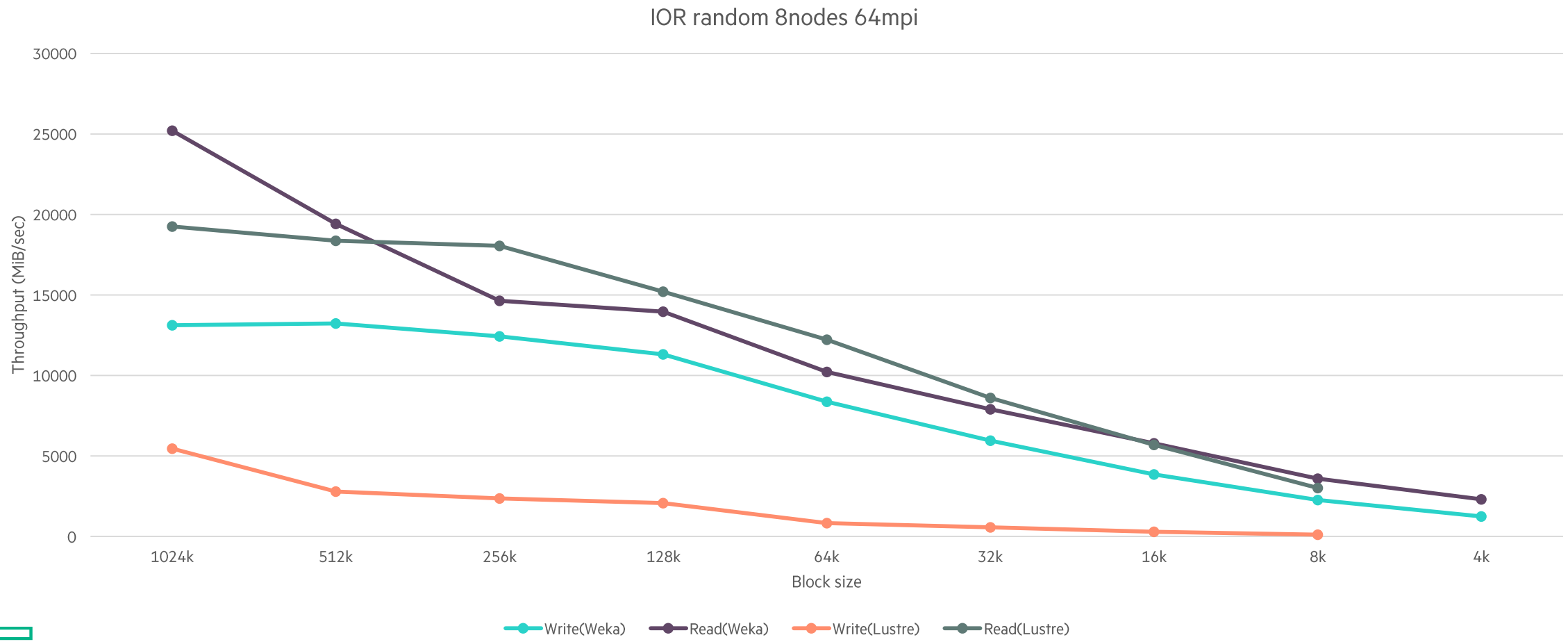
Lustreと比較しReadはWekaの方が1.16倍速い。Writeは2.9倍Lustreが速い。
Lustreはクライアント32ノード768プロセスでの実行に対し、WekaIOではクライアント16ノード240プロセスで実行している。Lustreは3PB(MDT: HDDx20 OST: HDDx400)のファイルシステムに対し、Wekaは25TB(NVMex32)のファイルシステムであることを考慮すると、十分な性能である。



Hewlett Packard Enterprise

IOR(random throuthput) WekaIO vs Lustre

8ノード64mpiにおけるblock size別random write/readの結果から、細かいファイルサイズにおいてもWekaIOは性能が出せることがわかる。(4kにおけるLustreはwriteが100MiB/sec以下で遅すぎたため、測定を中断した)
WekaIOはread intensiveなファイルシステムではあるが、random writeにおいては優位性が認められる。



Performanceまとめ

- Read Intensiveなファイルシステム
  - IOPSにおいてもThroughputにおいても、Read Intensiveなファイルシステム
  - ランダムアクセスであればWrite Intensiveと言って差し支えない

- 少ない本数(NVMe)でも性能が出せる
  - Readでは大規模Lustreファイルシステム(3PB)に対して、WekaIO(25TB)がIOPS/Throughputで性能を上回った

- 使い勝手がよい
  - 特にチューニング無しで良い性能がでる
  - 今回のベンチマークではチューニグしていない
  - Lustreはstripeの数などのチューニングが必須

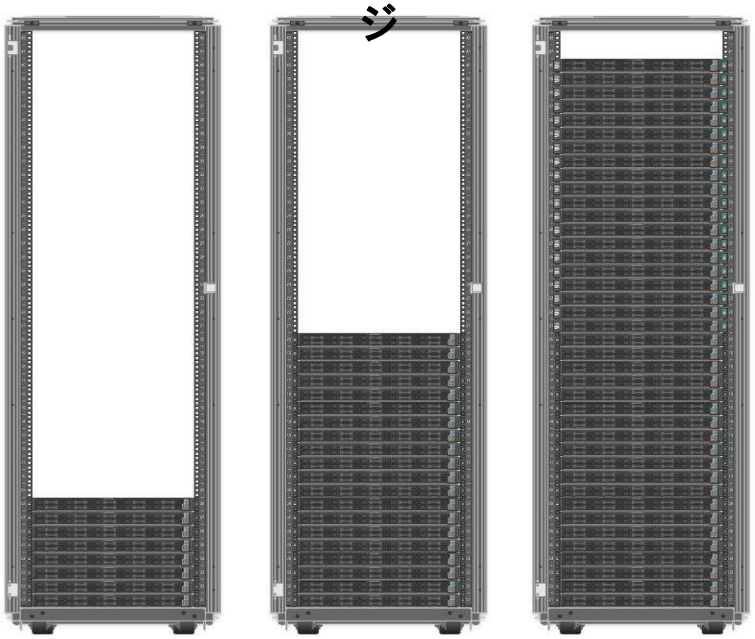
HPE製品情報

Hewlett Packard
Enterprise

DL360 Gen10 における WekaIO の構成

容量に最適化された安価なGB単価を提供

エントリー　ミッドレンジ　ハイエンド

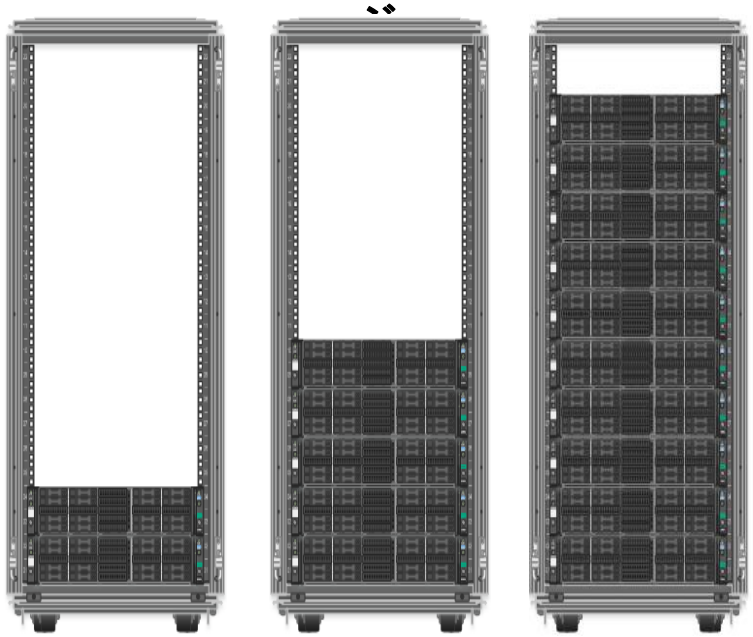


タイプ	ノード数	Raw 容量 (TB)	ユーザ容量 (TB)	Read 4K IOPS (Millions)	Read 1M Bandwidth (GB/s)
エントリー	8	102.4	54.6	2.5	30
ミッドレンジ	20	512	364.1	8.7	160
ハイエンド	40	1280	910.2	17.4	320

Apollo 2000 Gen10 における WekaIO の構成

パフォーマンスに最適なモデル

エントリー　ミッドレンジ　ハイエンド



タイプ	ノード数	Raw 容量 (TB)	ユーザ容量 (TB)	Read 4K IOPS (Millions)	Read 1M Bandwidth (GB/s)
エントリー	8	102.4	54.6	2.5	30
ミッドレンジ	20	256	182.0	8.7	125
ハイエンド	40	512	364.1	17.4	250

Hewlett Packard Enterprise

WekaIO SKU型番

- **WekaIO Matrix SKUs for NVMe Storage**
  - WekaIO Matrix 1yr Subscription/Support per TB E-LTU for HPE Servers Q9Q94AAE
  - WekaIO Matrix 3yr Subscription/Support per TB E-LTU for HPE Servers Q9Q95AAE
  - WekaIO Matrix 5yr Subscription/Support per TB E-LTU for HPE Servers Q9Q96AAE
  - WekaIO Matrix Education/Government 1yr Subscription/Support per TB E-LTU for HPE Servers Q9Q98AAE
  - WekaIO Matrix Education/Government 3yr Subscription/Support per TB E-LTU for HPE Servers Q9Q99AAE
  - WekaIO Matrix Education/Government 5yr Subscription/Support per TB E-LTU for HPE Servers Q9R00AAE
- **WekaIO Matrix SKUs for Tiering to Object Storage**
  - WekaIO Matrix 1yr Tiering per TB E-LTU for HPE Servers Q9R02AAE
  - WekaIO Matrix 3yr Tiering per TB E-LTU for HPE Servers Q9R03AAE
  - WekaIO Matrix 5yr Tiering per TB E-LTU for HPE Servers Q9Q97AAE

- **QuickSpecs WekaIO Matrix for HPE Servers**
  - https://h20195.www2.hpe.com/v2/GetDocument.aspx?docname=a00042248enw

Where to get additional information

Website	Product page on HPE.com for WekaIO: https://www.hpe.com/us/en/product-catalog/detail/pip.weakaio-matrix.1010676546.html HPE partner page on WekaIO site: https://www.weka.io/solutions/hpe/ WekaIO Archtecture https://www.weka.io/wp-content/uploads/2019/07/Architectural_WhitePaper-W02r3WP201812.pdf Security Feature https://www.weka.io/press-releases/wekaio-updates-file-system-with-advanced-security-features-for-multi-user-enterprise-hpc/
Tools	WekaIO Briefcase: https://psnow.ext.hpe.com/#/tiles/search/search?from=dashboard&t=wekaio WekaIO Matrix datasheet: https://www.weka.io/wp-content/uploads/2018/03/Matrix-DS-HPE-W01r7DS201803.pdf

Thank You