

DAOS アップデート

DAOS 2.0の新機能、サポートのご紹介

石橋史康

テクニカル・セールス・スペシャリスト

インテルOptaneグループセールス

インテル株式会社

fumiyasu.ishibashi@intel.com

お知らせと免責事項

インテル® テクノロジーの機能とメリットは、システム構成によって異なり、有効なハードウェア、ソフトウェア、またはサービスのアクティベーションが必要な場合があります。パフォーマンスはシステム構成によって異なります。

絶対に安全な製品や部品はありません。

テストは、特定のシステムにおいて、特定のテストでのコンポーネントの性能を記録するものです。ハードウェア、ソフトウェア、構成の違いにより、実際の性能は異なります。性能やベンチマーク結果についての詳しい情報は、<http://www.intel.com/benchmarks> をご覧ください。

パフォーマンス・テストに使用されたソフトウェアやワークロードは、インテル製マイクロプロセッサ上でのみ動作するように最適化されている場合があります。SYSmark や MobileMark などのパフォーマンス・テストは、特定のコンピューター・システム、コンポーネント、ソフトウェア、操作、機能を用いて測定されます。これらの要素に変更があった場合、結果が異なる可能性があります。他の製品と組み合わせた場合の製品の性能を含め、購入を検討する際には、他の情報や性能テストを参考にしてください。より詳細な情報については、<http://www.intel.com/benchmarks> をご覧ください。

インテル® AVX (Advanced Vector Extensions) は、特定のプロセッサ動作のスループットを向上させます。プロセッサの電力特性が異なるため、AVX 命令を使用すると、a) 一部のパーツが定格周波数以下で動作する、b) インテル® ターボ・ブースト・テクノロジー 2.0 を搭載した一部のパーツがターボ周波数を達成しない、または最大周波数を達成しない、といった現象が発生します。性能はハードウェア、ソフトウェア、システムの構成によって異なります。詳細は<http://www.intel.com/go/turbo> をご覧ください。

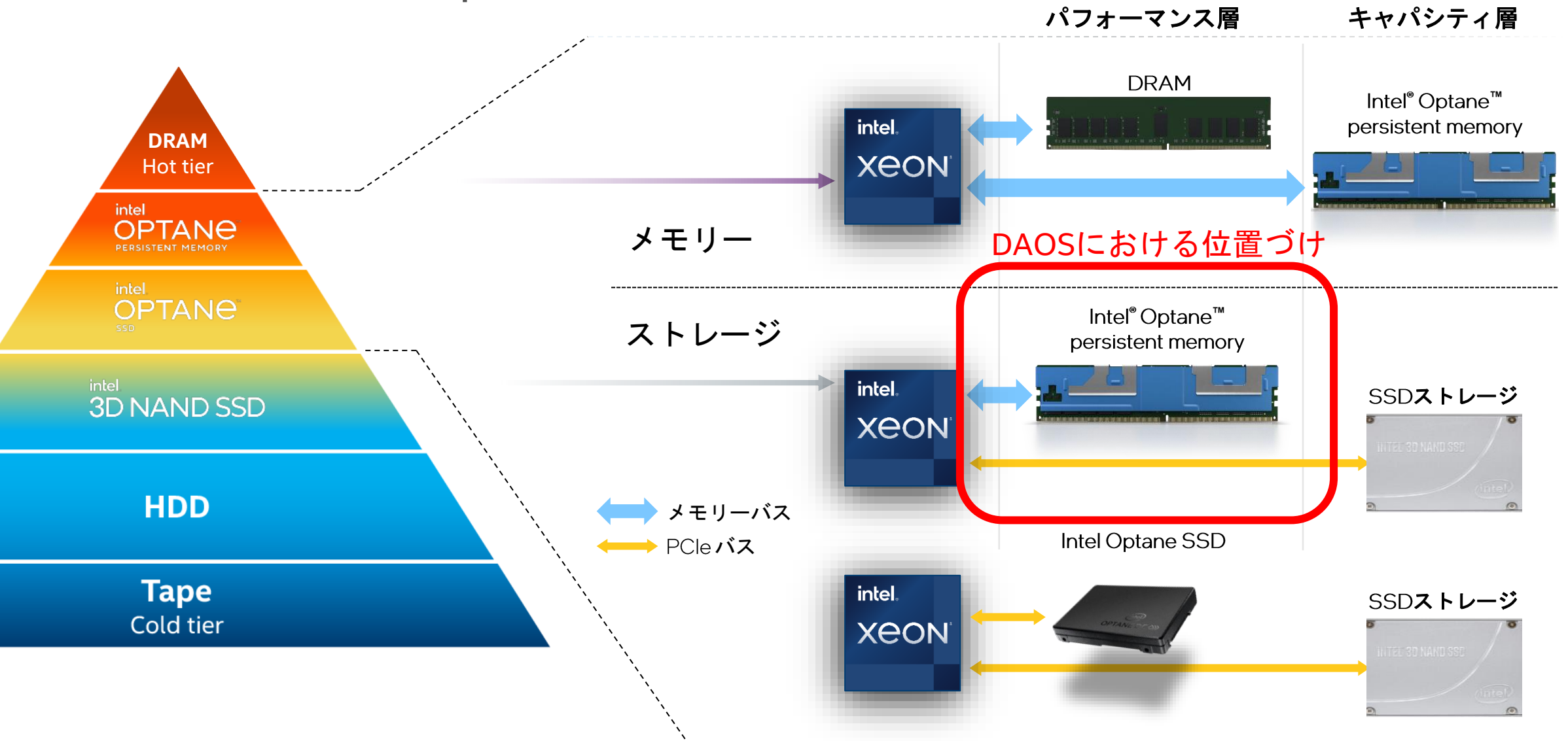
インテルのコンパイラーは、インテル製マイクロプロセッサに固有ではない最適化について、インテル製マイクロプロセッサ以外に対しても同じ程度に最適化する場合もあるかもしれませんが、しない場合もあります。これらの最適化には、SSE2、SSE3、SSSE3 命令セットやその他の最適化が含まれます。インテルは、インテル製以外のマイクロプロセッサにおける最適化の利用可能性、機能性、有効性を保証しません。本製品のマイクロプロセッサ依存の最適化は、インテル製マイクロプロセッサでの使用を意図しています。インテル・マイクロアーキテクチャーに特化していない一部の最適化機能は、インテル・マイクロプロセッサ用に予約されています。本通知の対象となる特定の命令セットに関する詳細は、該当する製品のユーザーガイドおよびリファレンスガイドを参照してください。

記載されているコスト削減シナリオは、特定の状況および構成において、特定のインテルベースの製品が将来のコストにどのように影響し、コスト削減をもたらすかの例として意図されています。状況は様々です。インテルはいかなるコストやコスト削減をも保証するものではありません。

インテルは、サードパーティのベンチマークデータや本文書で参照されている Web サイトの管理や監査を行っていません。参照先の Web サイトにアクセスして、参照データが正確であるかどうかを確認する必要があります。

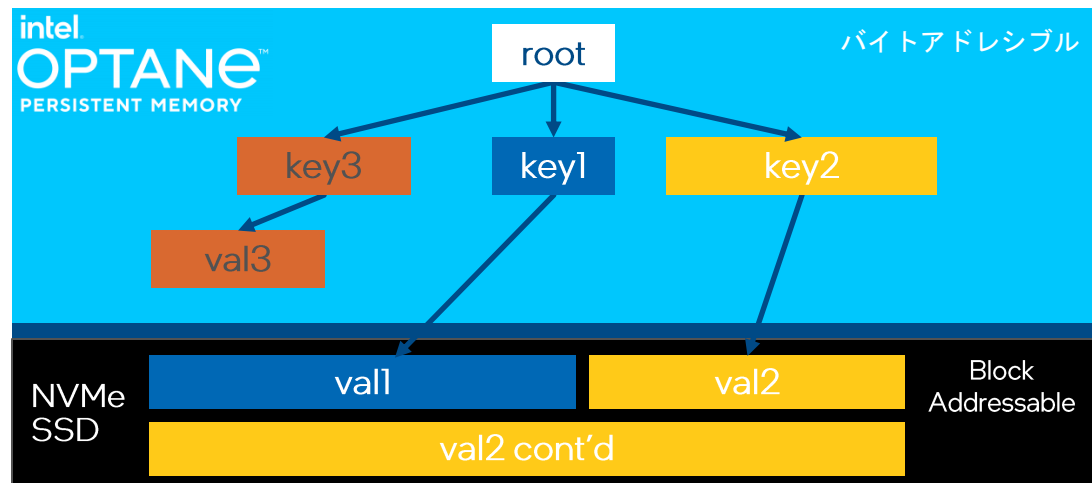
© Intel Corporation. Intel、Intel ロゴ、およびその他の Intel マークは、Intel Corporation またはその子会社の商標です。その他の名称やブランドは他者の所有物として主張されることがあります。

インテル® Optane™ テクノロジー



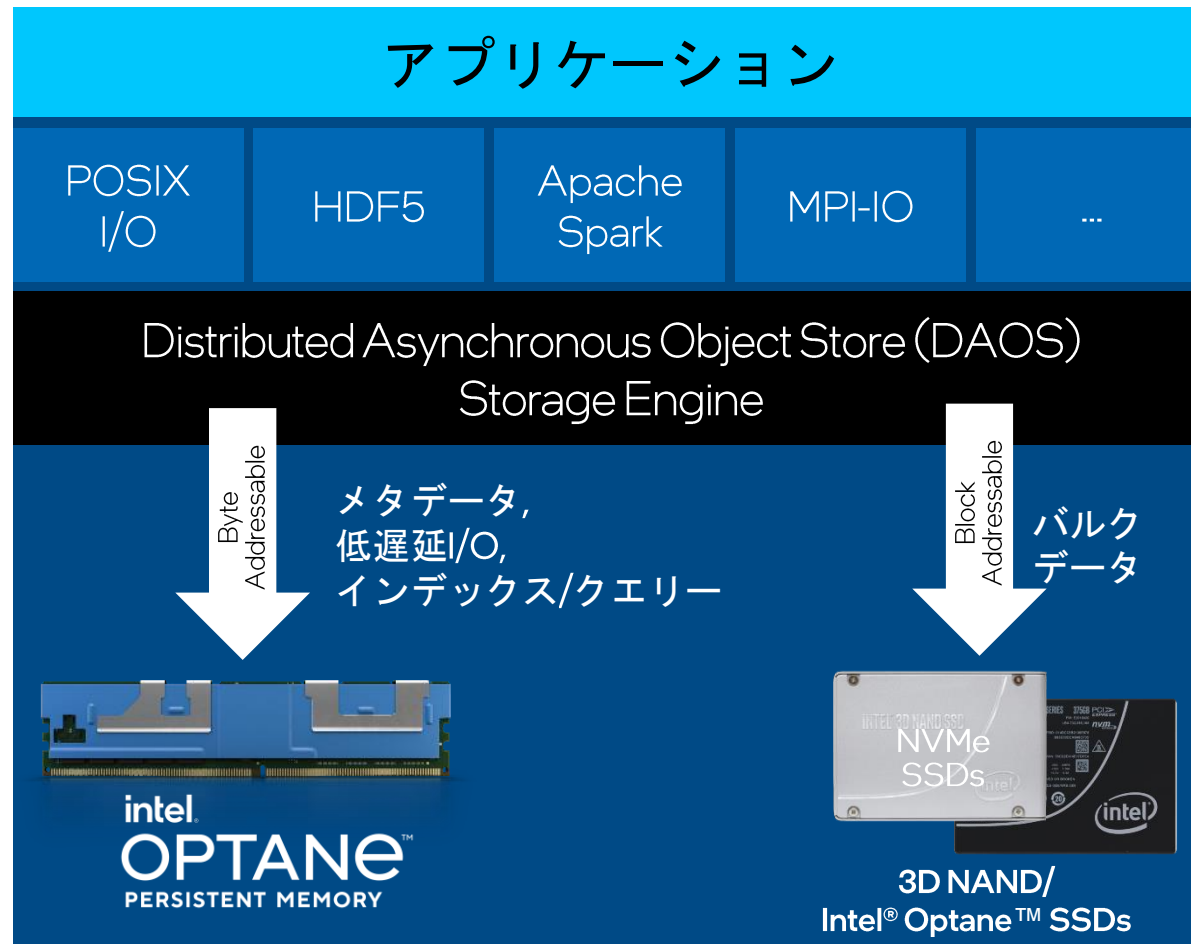
DAOS とは

革新的分散ストレージ



DAOS: POSIX を超えて

- 構造化データと非構造化データの両方に適した新しいスケーラブルなストレージモデル
- キーバリューストア、多次元配列、カラムナーデータベースなど
- POSIXシリアライゼーションの制約を受けない
- データ分析/AIフレームワークの高速化



DAOSストレージモデル

■ DAOS Pool

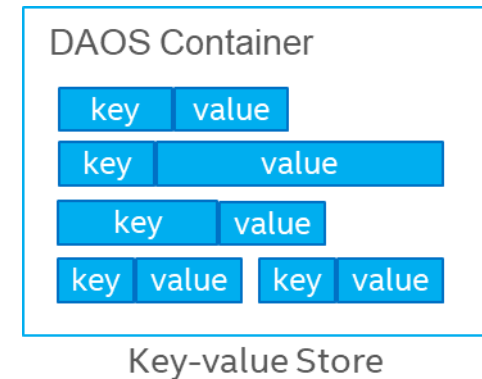
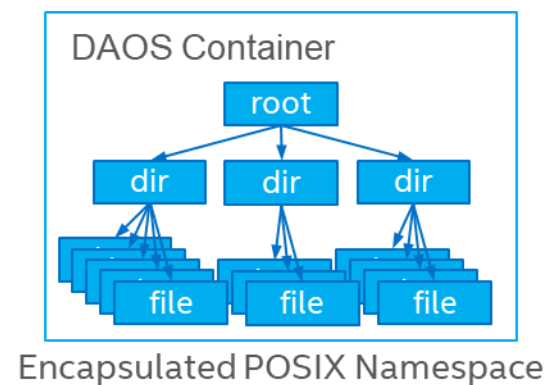
- SCMやNVMe SSDなどのハードウェアリソースと仮想的なストレージとして定義したもの
- リサイズ可能

■ DAOS Container

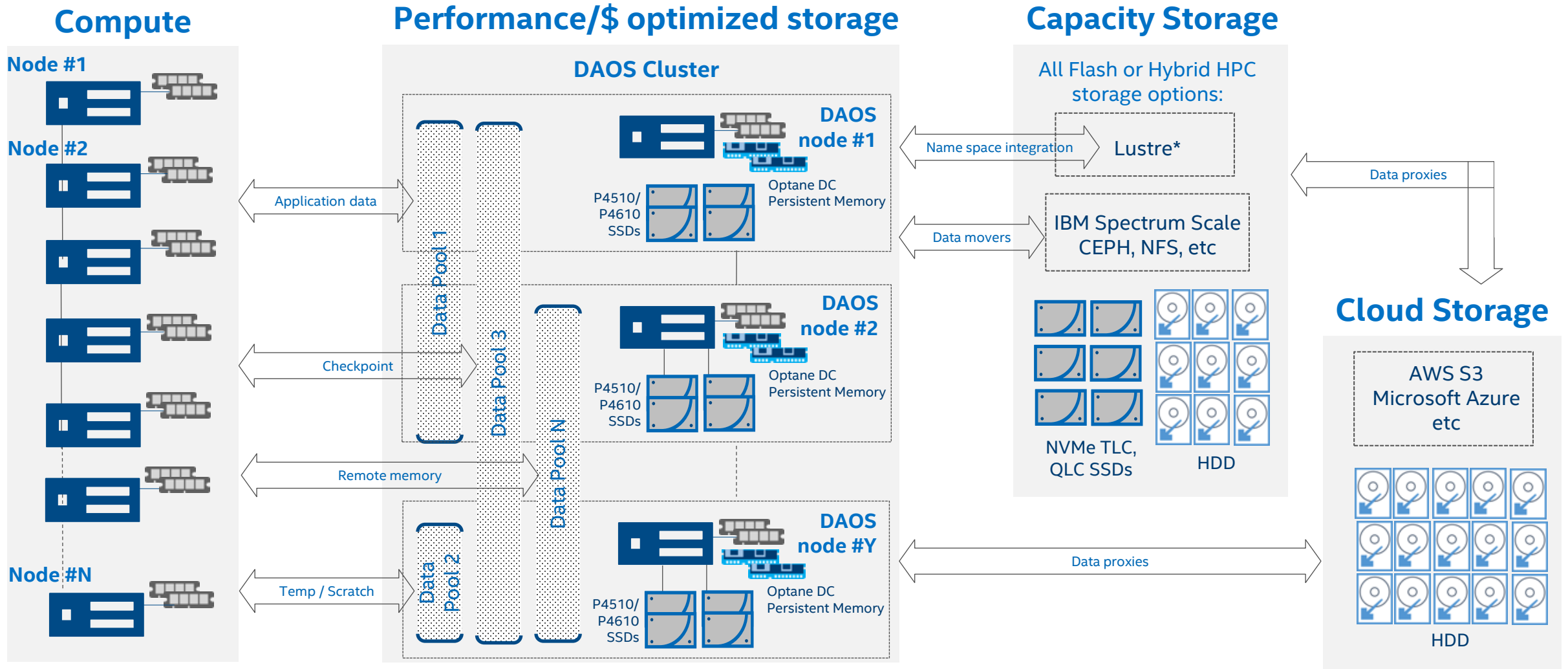
- DAOS Poolに保管
- ユーザーにより管理されるデータセット
- POSIX I/O, HDF5, Key-Value Storeなどのデータモデルを管理
- スナップショット、ロールバックをサポート

■ DAOS Object

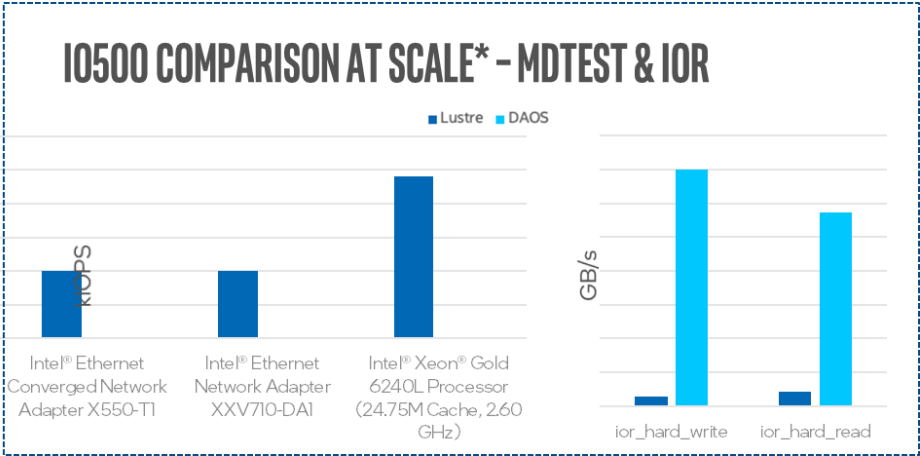
- コンテナは、トランザクションとバージョニングの基本単位
- 全てのオブジェクト操作には、エポックと呼ばれるタイムスタンプが付与



HPC クラスターにおける DAOS



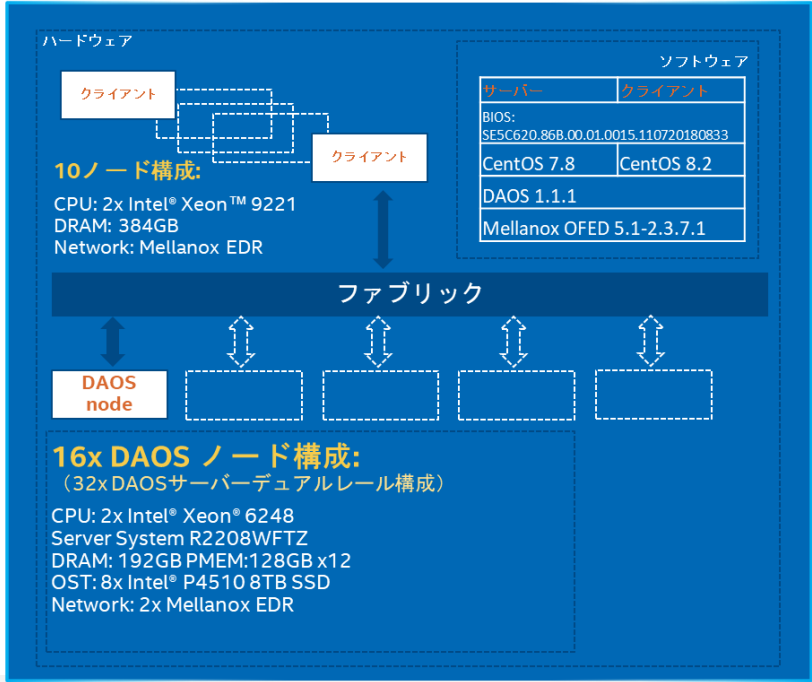
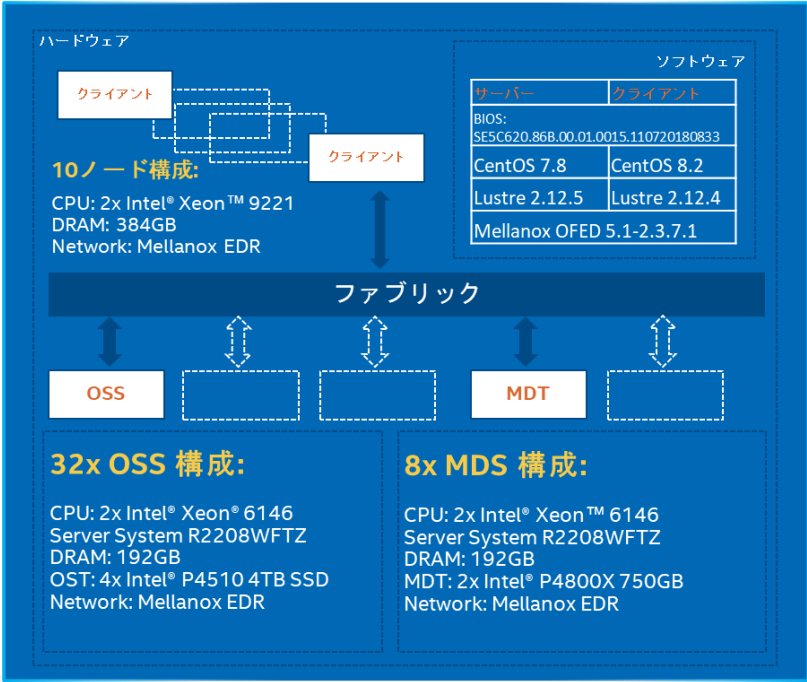
DAOS vs Lustre パフォーマンス比較



LUSTRE

DAOS

- mdtestのhard writeで**70X**の性能向上
- mdtestのhard readで**5X**の性能向上
- mdtestのhard statで**6X**の性能向上
- iorのhard writeで**24X**の性能向上
- iorのhard readで**16X**の性能向上



DAOS 2.0の新機能

- イレイジャーコード
- ストレージターゲットの再統合
- テレメトリーとモニタリング機能
- 安定性、操作性、ドキュメントの改善
- L3 テクニカルサポートの提供を発表
- 第3世代インテル® Xeon® スケーラブル・プロセッサー（Ice Lake）のサポート

DAOS L3 テクニカルサポート

- Intel DAOSチームがパートナー向けに商用L3サポートを提供
- 最低サポート契約期間は3年で、延長可能
- フルフィルメント:
 - MMID SKU 標準デリバリープロセス
 - ハードウェアサイズによる課金
- エスカレーションパス
 - PoCプログラム - 初期のエンゲージメント/PoCは、SLAを保証することなく、オープンソースのコミュニティサポートのメーリングリストとSlackチャンネルにアクセス可能
 - L2からL3へのエスカレーションを追跡するためのコミュニティJIRAチケットを運用中

L3サポートの前提HW・SW (DAOS 2.0)

■ DAOSサーバー

- PMDKをサポートするSCM : Intel Optane Persistent Memory
- SPDKをサポートするNVMe SSD
- OS : CentOS Linux 8(CentOS 8.5), RHEL 8(currently RHEL 8.5), openSUSE Leap 15.3, SLES 15 SP3

■ DAOSクライアント

- OS : DAOSサーバーと同じ (将来的には拡張予定)

■ Fabric

- Infiniband(OPA, RoCEはサポート予定)
- libfabric Version 1.14以降
- MLNX_OFED Version 5.3、 Version 5.4

https://docs.daos.io/v2.0/release/support_matrix/

DAOS Ecosystem

Hardware Partners	       
Reseller Partners	     
Software Development and 3 rd party support	     
End Customers	    

DAOSスケーリング

DAOSは、エクストリームスケールに対応したスケールアウト型のストレージソリューションです。DAOSのアーキテクチャ上の限界と、DAOS Ver.2.0の現在の検証値を示します。

DAOSコンポーネント	理論上のオーダー	DAOS 2.0で検証済みの数値
DAOSシステムあたりの最大DAOSクライアント数	数十万	256
DAOSシステムあたりの最大DAOSサーバー数	数千	128
DAOSサーバーあたりの最大DAOSエンジン数	10未満	1か2か4
DAOSエンジンあたりの最大DAOSターゲット数	数十	4-16
DAOSエンジンあたりの最大SCM数	数十	6 (Optane PMem 100), 8 (Optane PMem 200)
DAOSエンジンあたりの最大NVMe SSD数	数十	4-12
DAOSシステムあたりの最大DAOS Pool数	数百	100
DAOS Poolあたりの最大DAOSコンテナ数	数百	100
DAOSコンテナあたりの最大DAOSオブジェクト数	数百億	60億 (mdtestによる)
DAOSコンテナあたりにアクセスできる 最大アプリケーションタスク数	数百万	3072 (verbsによる)

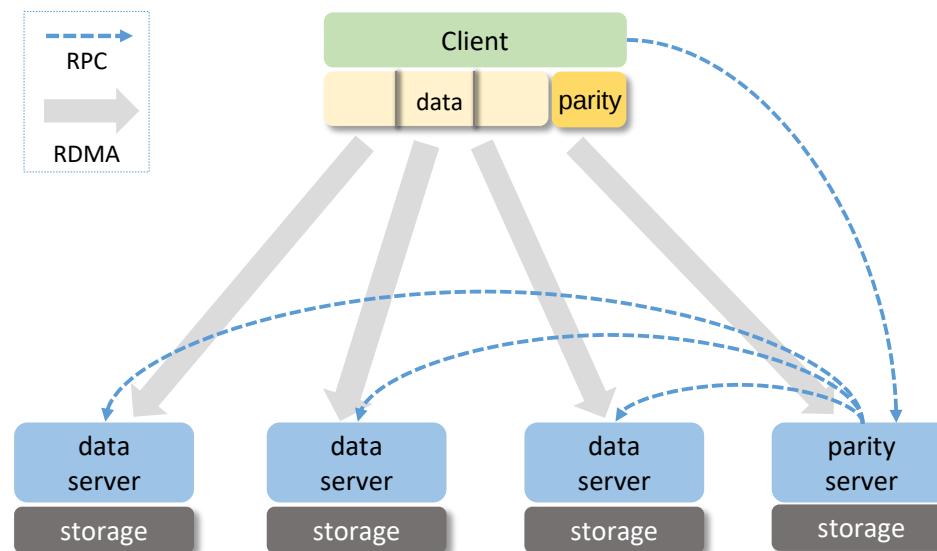
イレージャーコード

DAOS 2.0よりイレージャーコード（Erasure Code）をサポート

- リード・ソロモン方式によるイレージャーコードのサポート
 - ・ レプリケーションよりも容量を節約し、帯域幅の効率化を実現
- パフォーマンス
 - ・ 正常時、Read性能への影響なし。Write性能へ影響あり。
 - ・ レプリケーションよりも
- クライアントサイド・エンコーディング
- デグレード・モード
 - ・ クライアント側でデータの再構築が可能
 - ・ サーバー側では自己修復サービスがフォールバックノードに再格納する

ことでユーザーデータを元の冗長レベルまで回復が可能
- リビルド

- ユーザーはオブジェクト単位でイレージャーコードを設定可能
- ユーザーはセルサイズを設定可能な小さすぎるとメタデータ処理が増えPMEMパフォーマンスに影響

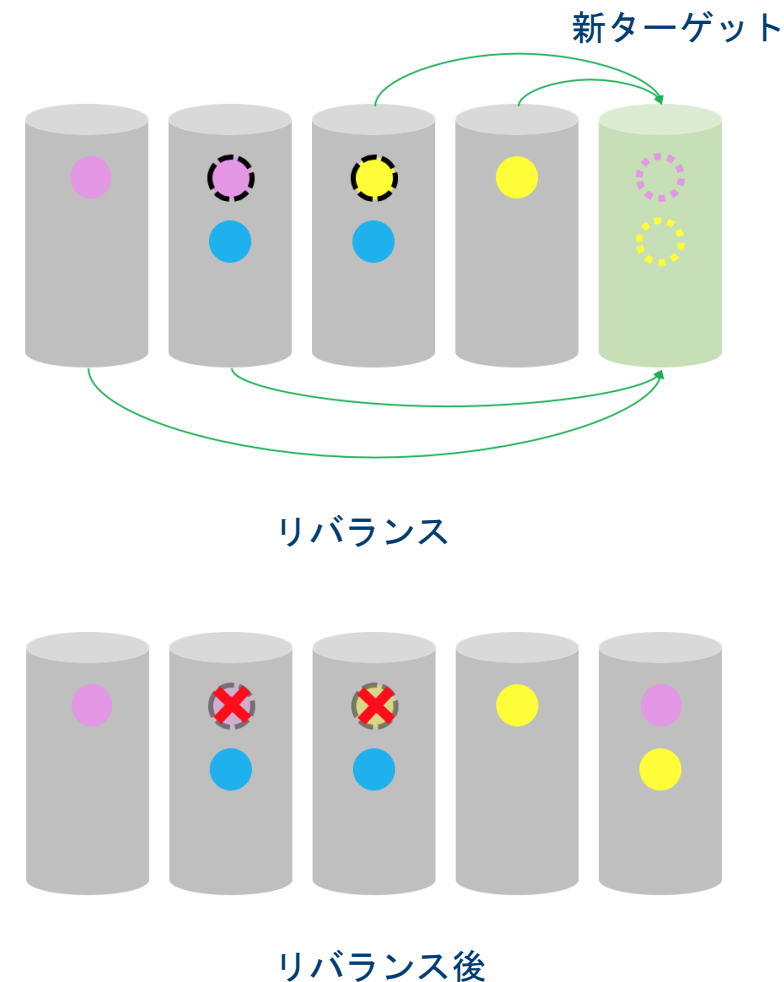


テレメトリとモニタリング

- DAOSは、各ストレージエンジンのメトリクスと統計情報をエンジンの稼働中に維持し、DAOSの健全性とパフォーマンス、トラブルシューティングに関する情報を提供
- 全体的な帯域幅やその他の統計情報の取得が可能
- 提供される情報には、エンジンのストレージに対する読み書きのバイト数、I/Oレイテンシー、I/O操作、エラーイベント、内部状態など
- Prometheus（オープンソースのシステム監視ツール）との連携
- PrometheusのデータをGrafana（可視化ツール）による可視化

ストレージターゲットの再統合

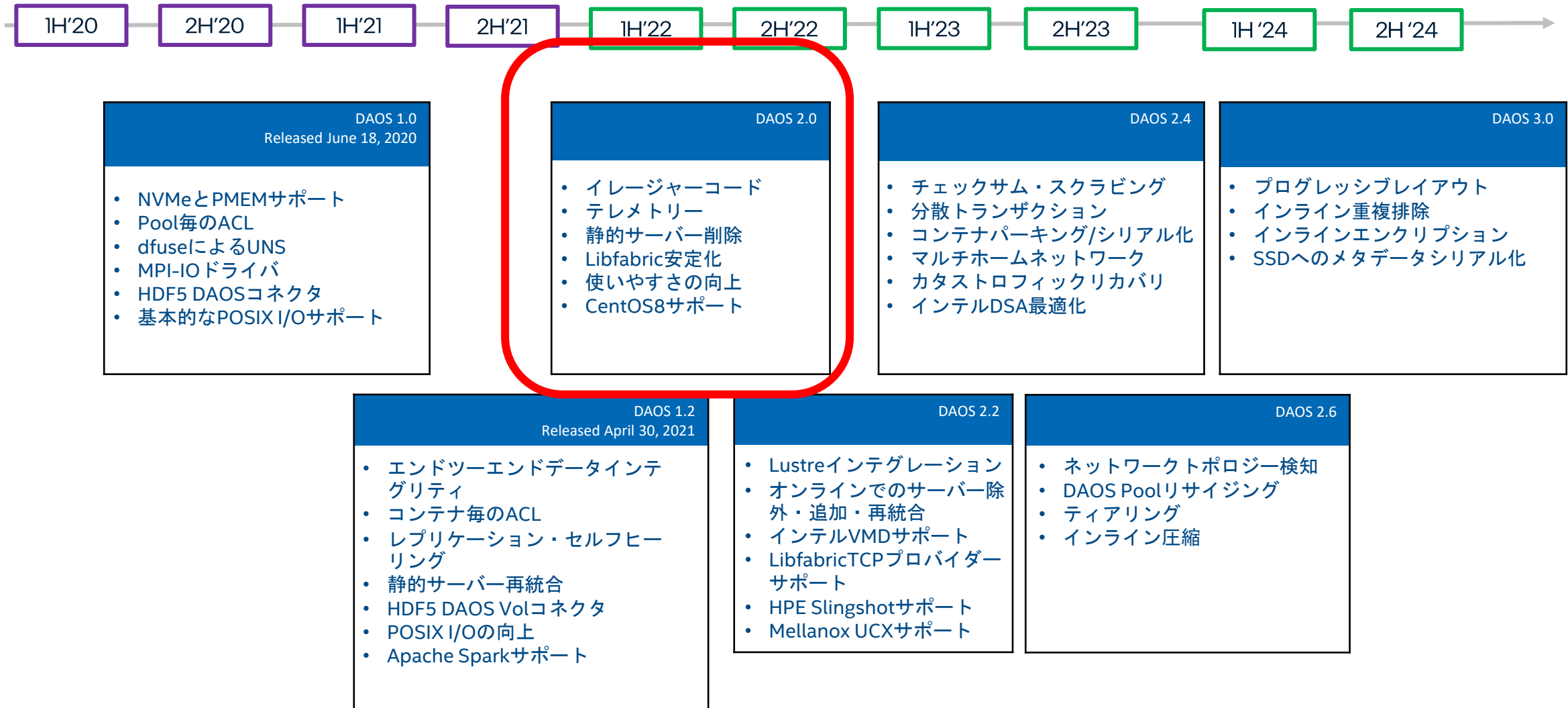
- 復旧したターゲットのPoolへの再統合
 - 障害などにより一時的に除外したストレージターゲットをPoolに戻すことが可能
 - 再統合後リビルド処理が発生
- Poolサイズの拡張
 - 物理ノードやストレージの追加により可能
 - Pool内でリバランス処理が発生



DAOS2.0アップグレードの注意点

- DAOS Version 1.0または1.2からDAOS Version 2.0へのアップグレードは、基本的に新規インストールとなります。
- DAOSの既存のユーザーデータは、DAOSストレージを再フォーマットする必要があるため、アップグレードの前にDAOS以外のストレージにバックアップする必要があります。

DAOS ロードマップ



NOTE: All information provided in this roadmap is subject to change without notice.

リソース

コミュニティウェブサイトとドキュメント

- <https://daos.io>

GitHub上のソースコード

- <https://github.com/daos-stack/daos>

コミュニティメーリングリスト

- daos@daos.groups.io

コミュニティ Slack チャンネル

- <https://bit.ly/38SOLEy>

サポート

- <https://jira.hpdd.intel.com>

Intel DAOS ウェブサイト

- <https://www.intel.com/content/www/us/en/high-performance-computing/daos.html>

DAOS ソリューションブリーフ

- <https://www.intel.com/content/www/us/en/high-performance-computing/>

DAOS SCA'20からのテクニカルペーパー

- <https://rdcu.be/caYtU>

DAOS YouTube チャンネル

- <https://bit.ly/3pHHxcl>

