HC110

vSAN 6.7 Update 3 で進化した vSAN 性能徹底比較

ヴイエムウェア株式会社
ソリューションビジネス本部
シニア HCI スペシャリスト　知久　貴弘

vmware

免責事項

- このセッションには、現在開発中の製品/サービスの機能が含まれている場合があります。
- 新しいテクノロジーに関するこのセッションおよび概要は、VMware が市販の製品/サービスにこれらの機能を搭載することを約束するものではありません。
- 機能は変更される場合があるため、いかなる種類の契約書、受注書、または販売契約書に記述してはなりません。
- 技術的な問題および市場の需要により、最終的に出荷される製品/サービスでは機能が変わる場合があります。
- ここで検討されているまたは提示されている新しいテクノロジーまたは機能の価格およびパッケージは、決定されたものではありません。

Agenda

vSAN 詳細アーキテクチャ

LSOM 1.5 による性能改善

ハードウェアの進化

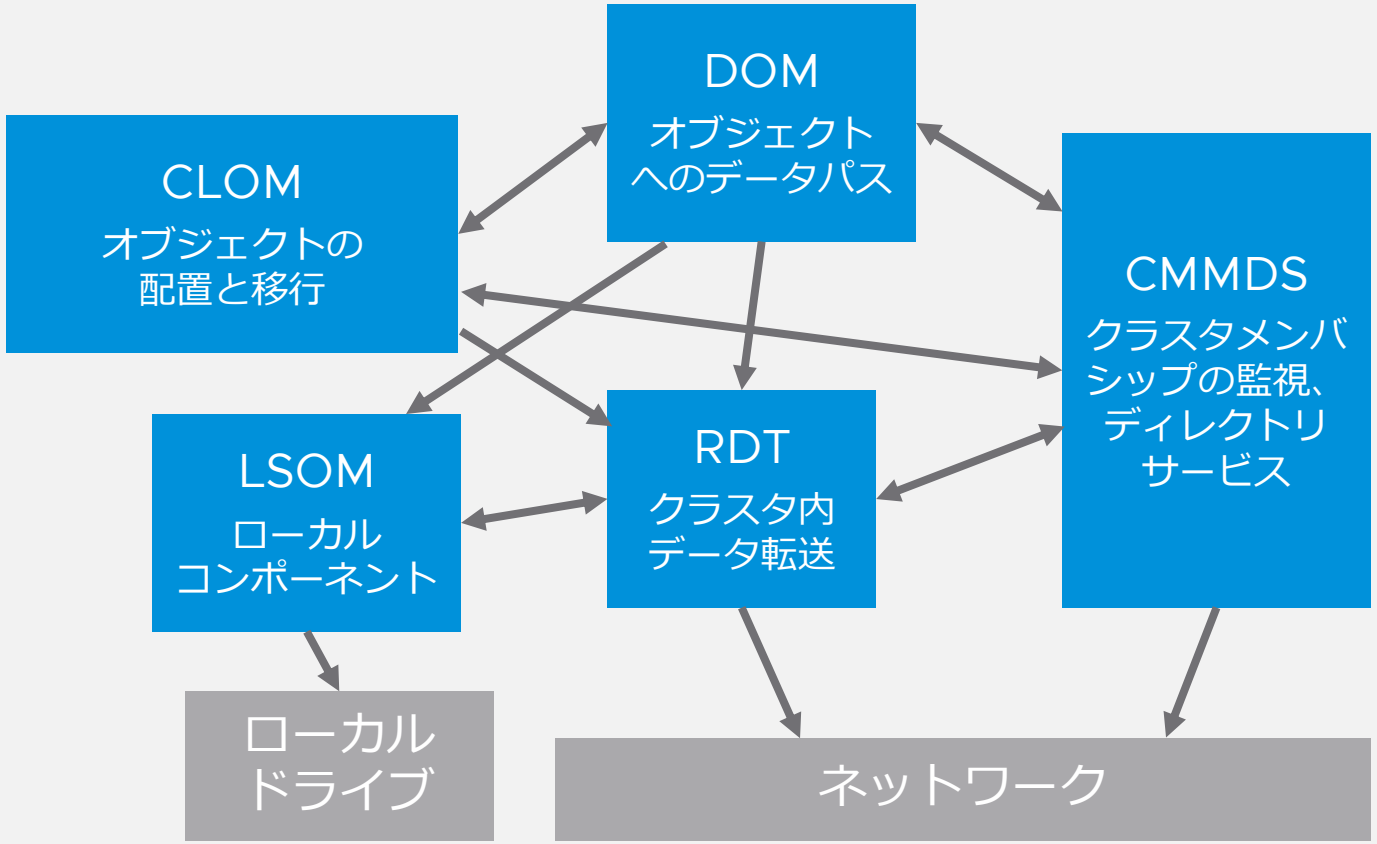
性能検証データ比較

- vSAN 6.7 U1 vs vSAN 6.7 U3

vSAN 詳細アーキテクチャ

vSAN モジュール構造

クラスタ内の全てのノードに実装されている



DOM は VMware vSAN™ のモジュールを接続し、オブジェクトへのデータパスを提供する

CLOM はオブジェクトの配置と移行をハンドリングする

LSOM はコンポーネントの管理を実施

CMMDS はクラスタのメンバシップを維持し、他のモジュールへデータを提供する

RDT は信頼出来るデータ転送を提供する

vSAN Cluster-level Object Manager (CLOM)
オブジェクトの配置を決定

リーフノードの数を計算

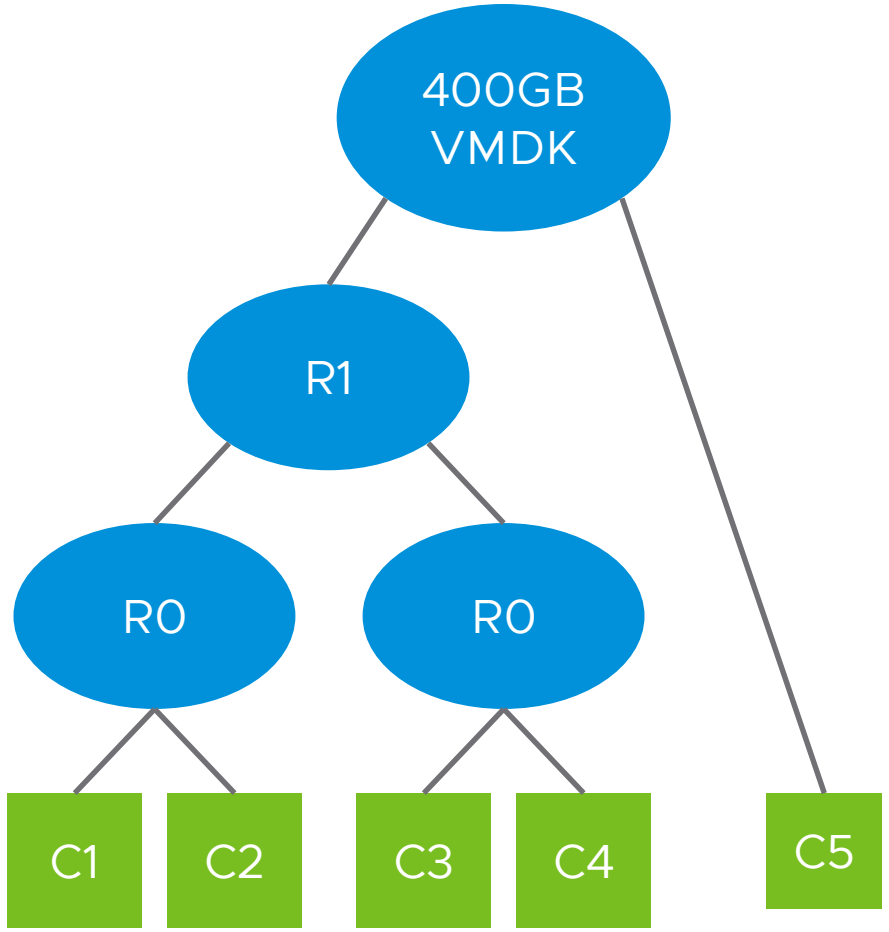
ツリートポロジーの候補を生成

コンプライアンスとコストの観点で各トポロジーを評価
- よりシンプルなツリーはコストが低いと評価され優先される

監視コンポーネントと投票を決定

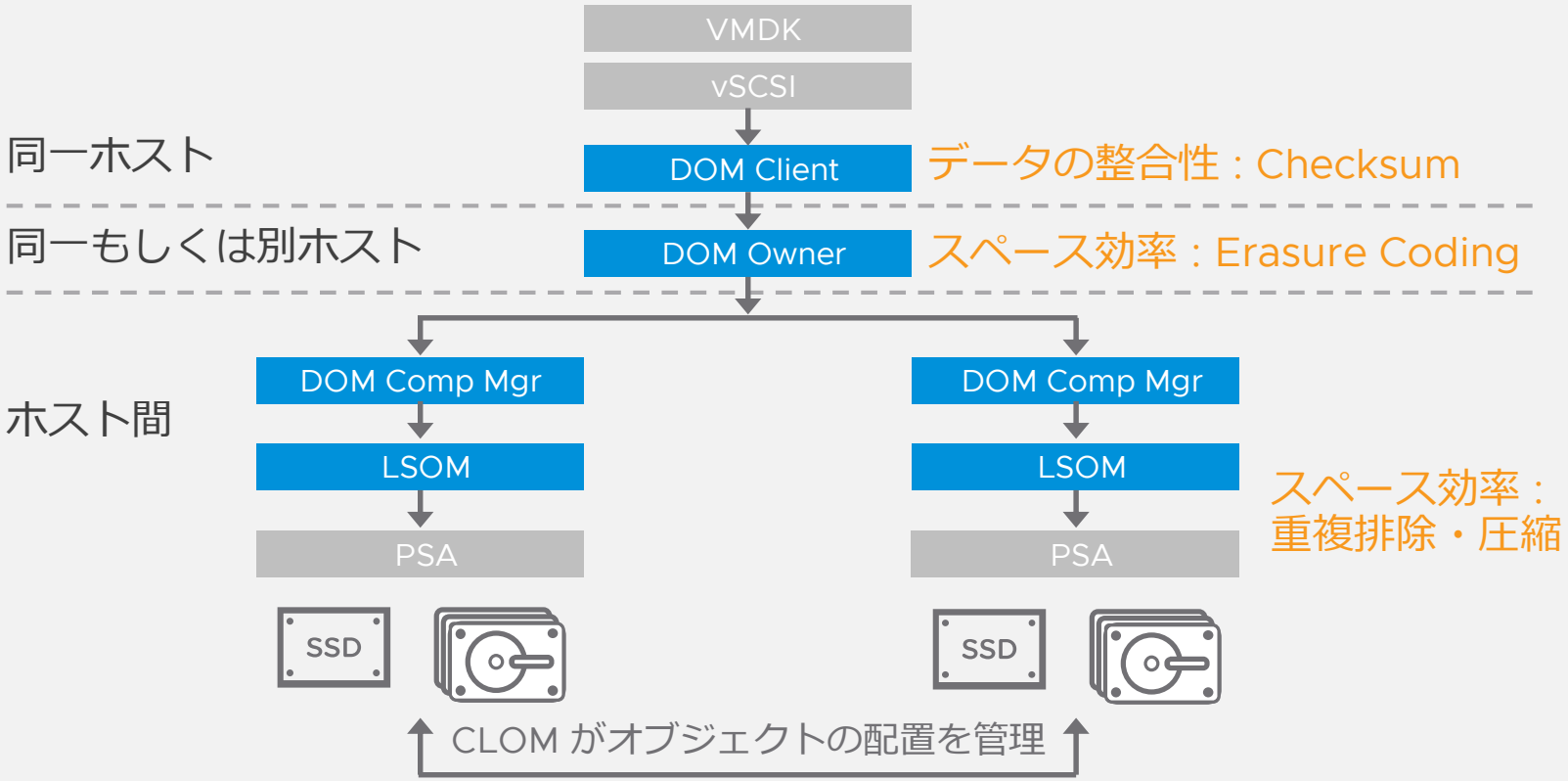
コンポーネントが配置されるディスクをアサイン

以下の処理も実施 :
- 再構成
- スワップオブジェクトの生成
- ロードバランス (データリバランス)



vSAN Distributed Object Manager (DOM)

オブジェクトへの分散アクセスを実装し提供



DOM は CLOM が決定したオブジェクトの配置構成を実行する

DOM の役割:

- Owner
- Client
- Component

DOM Owner はオブジェクトへのアクセスを仲介する

RDT はピア DOM 間の設立された接続を利用する

vSAN Cluster Monitoring, Membership, and Directory Services

通称 CMMDS

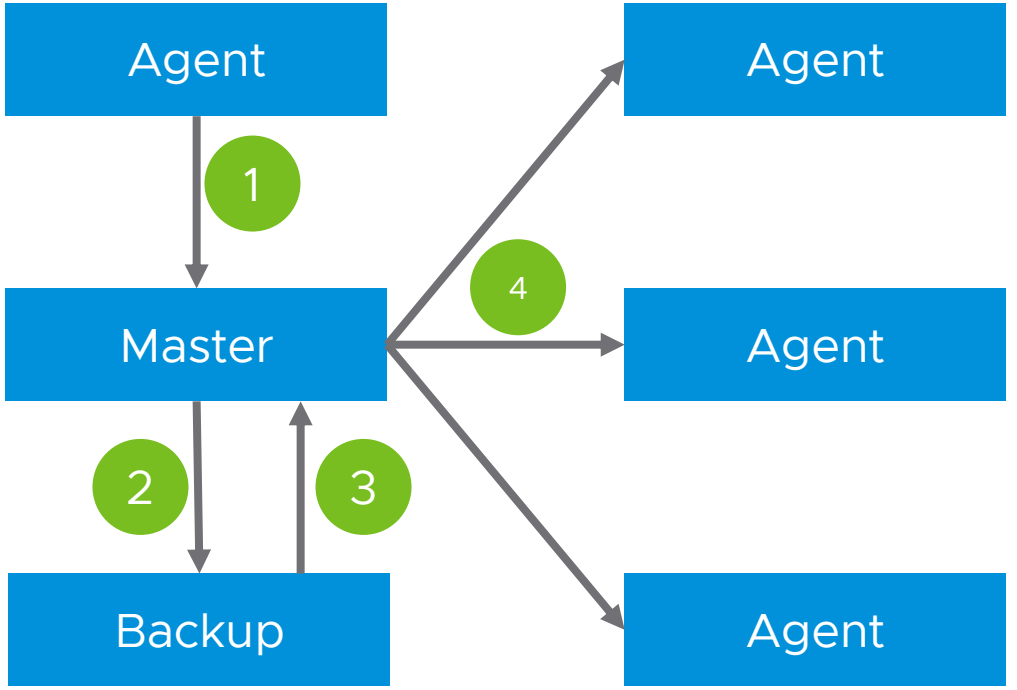
クラスタへのリンクとクラスタメタデータの優先分散ファブリックを監視

各エージェントはローカルの情報を更新し、マスタにアップデート情報を送付する

DOM は CMMDS を利用して、どのホストにオブジェクトのコンポーネントが配置されているかを確認する

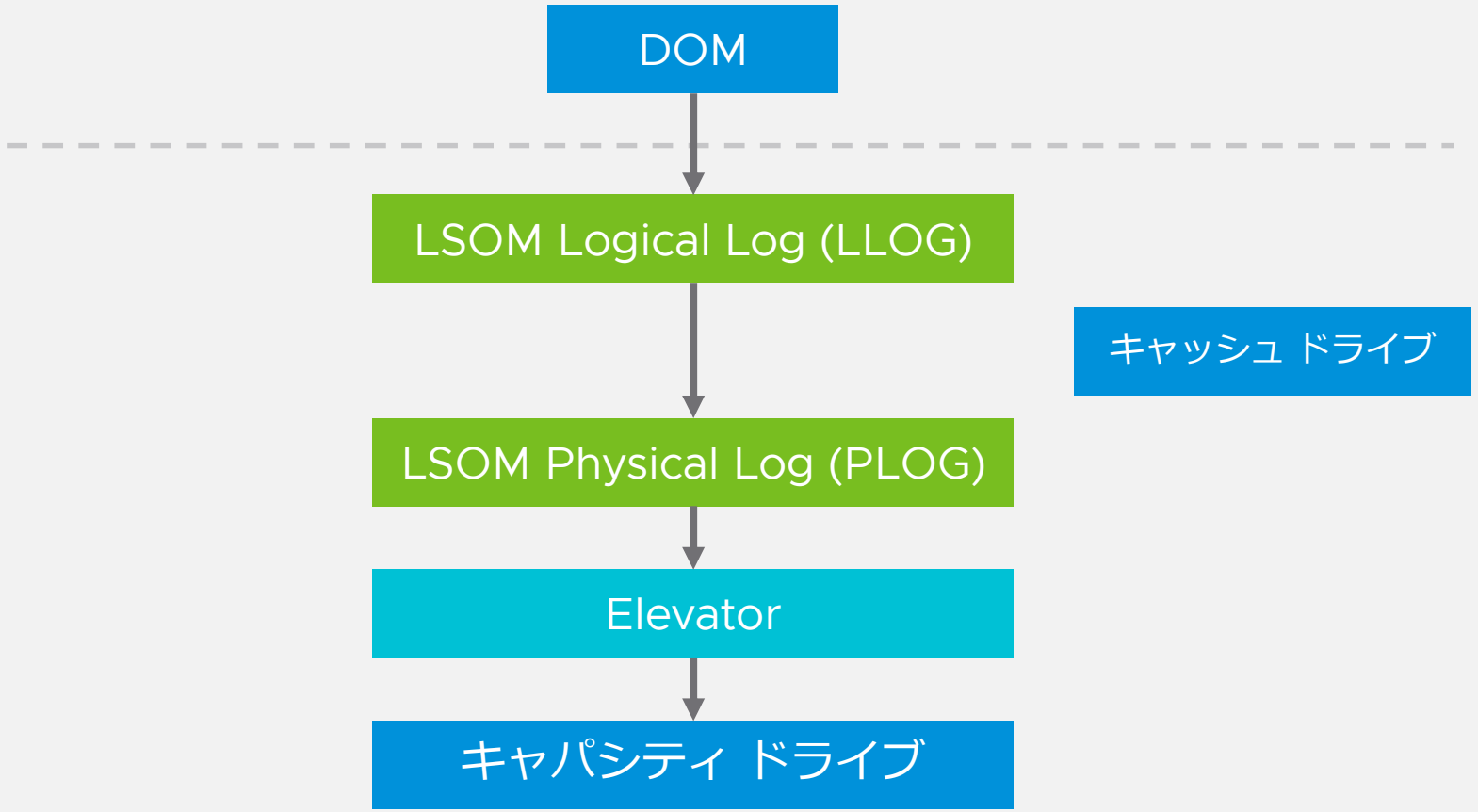
CMMDS は DOM のオブジェクトオーナー選出を手助けする

CMMDS はクラスタとネットワークの健全性維持も担当する



vSAN Local Log-structured Object Manager (LSOM)

ローカル物理ドライブ上のストレージコンポーネントを提供



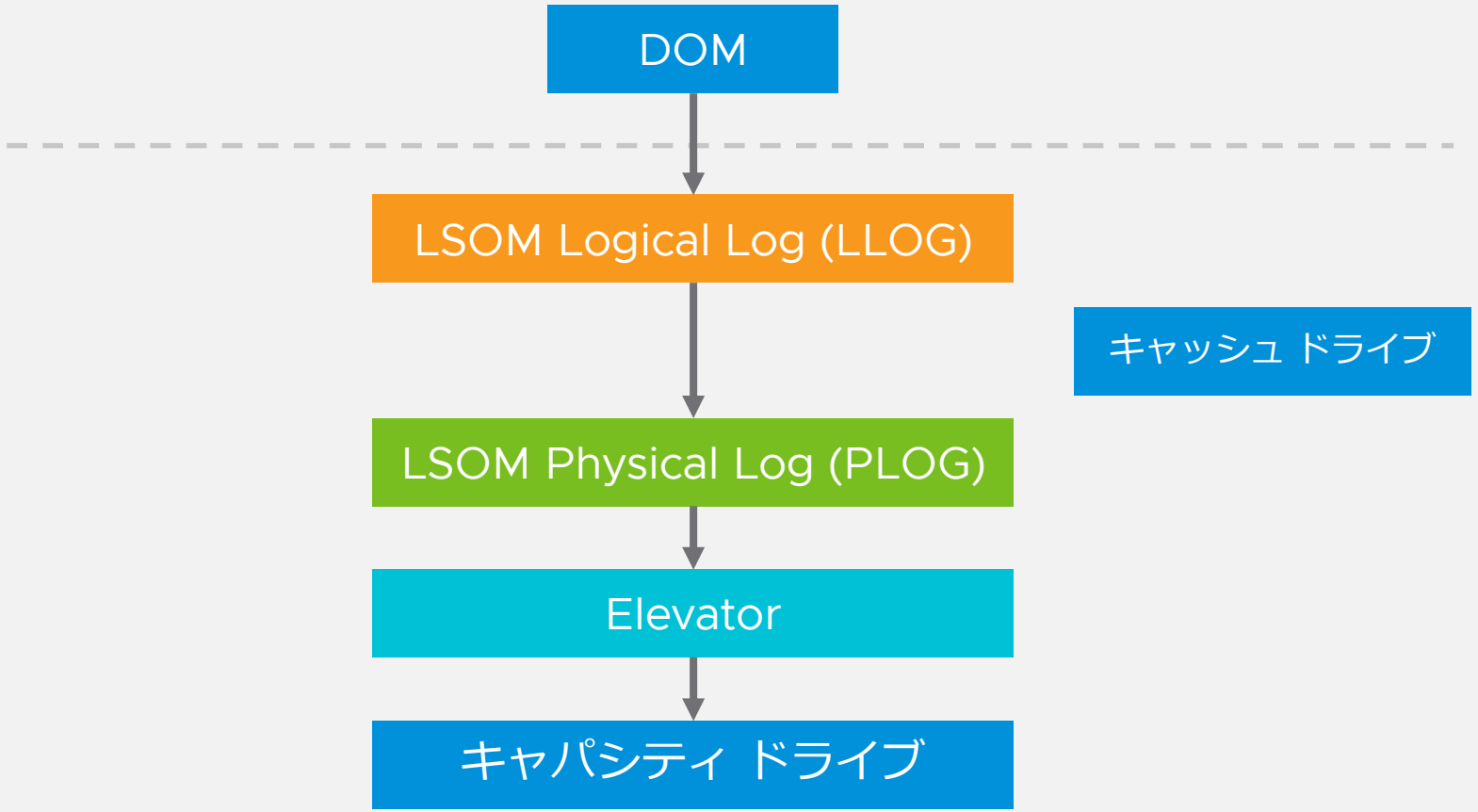
DOM の要求に応じてコンポーネントを作成・変更・削除する

ディスクボリュームと全てのコンポーネントを CMMDS へ公開する

hostd や CMMDS に報告される容量や使用量な情報を CLOM は配置決定に利用する

vSAN Local Log-structured Object Manager (LSOM)

ローカル物理ドライブ上のストレージコンポーネントを提供



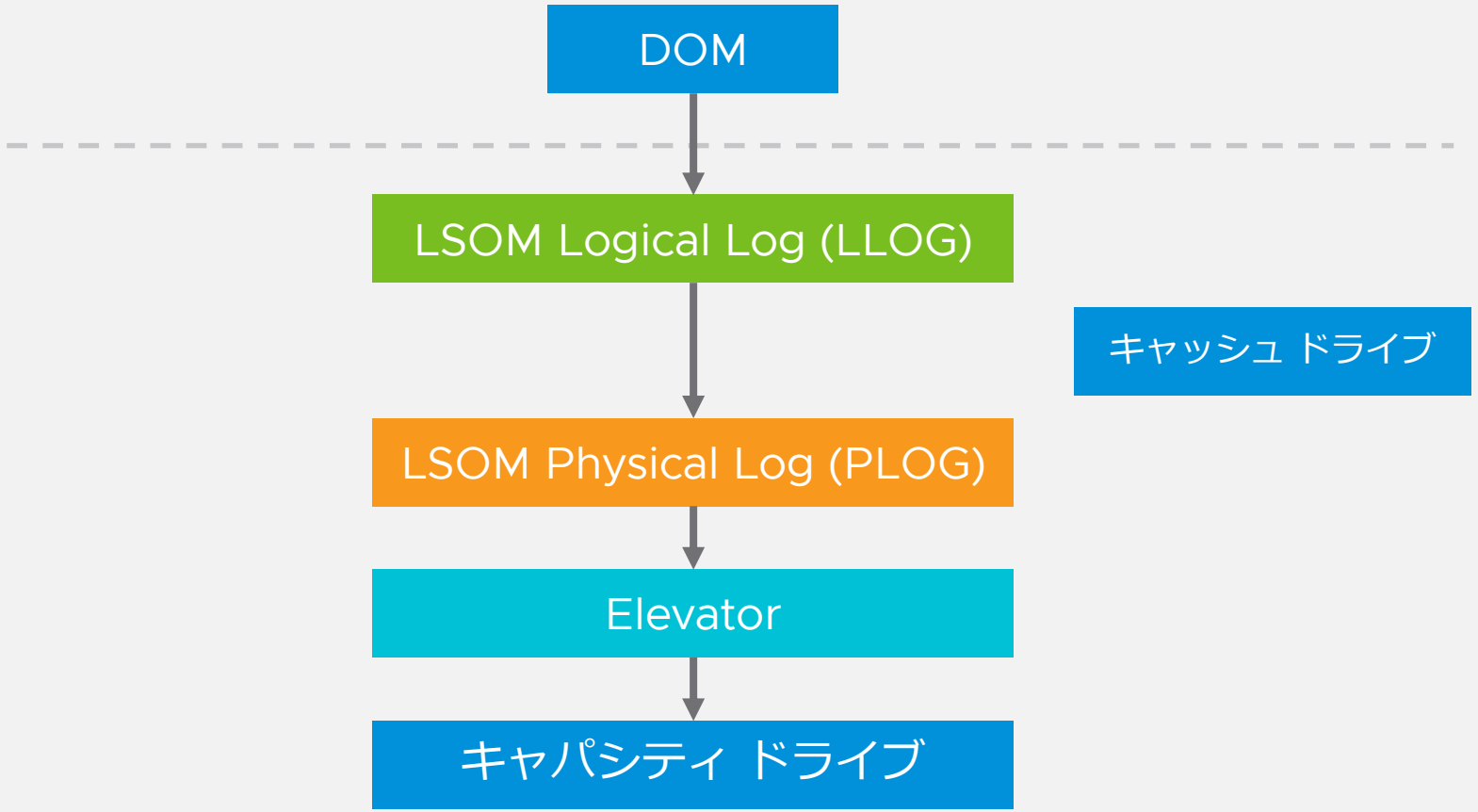
Logical Log (LLOG)

準備とコミット(2フェーズコミット)オペレーションを DOM の指示で実施

クラッシュ時のリカバリに備えて準備はロギングされる

vSAN Local Log-structured Object Manager (LSOM)

Provides storage of components on local physical drives



Physical Log (PLOG)

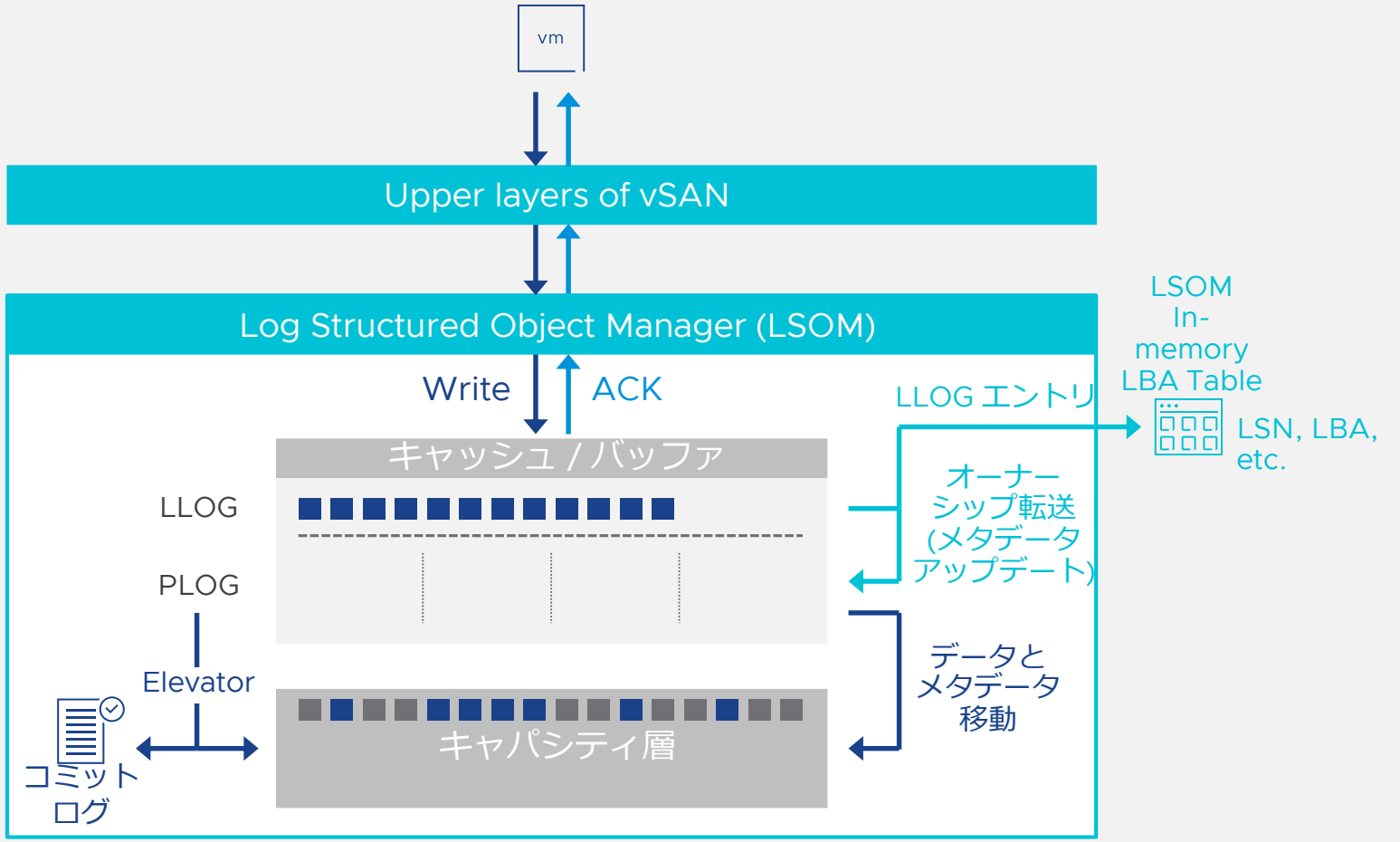
デステージのために全データとメタデータが準備される

vSAN デバイスとデバイスイベントを報告する

キャッシュからキャパシティへデータを排出するエレベータアルゴリズムを実装する

デステージ

カスタムエレベータアルゴリズムが最も効果的な方法でデステージする



複数のデステージトリガ

- バッファフル
- LSOM のメモリ負荷
- SSDLog エントリ

必要に応じてバッチ処理で
デステージ

順序と近接性を理解

頻繁に上書きされるデータは
バッファに保持される

LSOM 1.5による性能改善

従来の LSOM の課題

- シーケンシャル書き込みの性能
- 重複排除・圧縮有効時の書き込みバッファの効率性
- より一貫した遅延の提供
- 再同期時間の短縮

vSAN 6.7 Update 3 での改善項目

ワークロード

より迅速な再同期を可能にするデステージの最適化

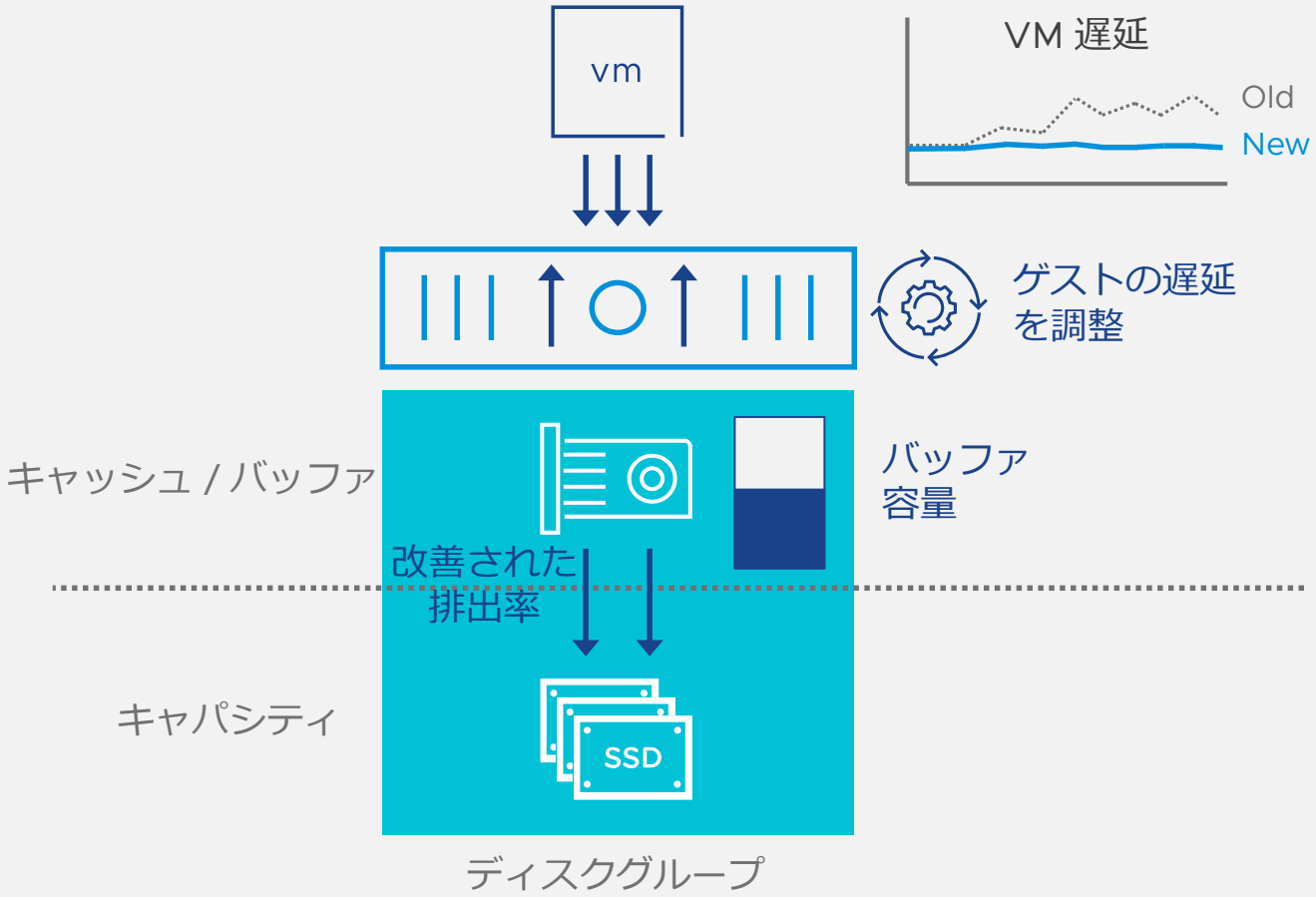
シーケンシャル IO の改善

センシティブなアプリのために予測可能な遅延を提供

項目	改善目標	目標への達成手法
一貫した性能	書き込み I/O 遅延の標準偏差を削減する	• 一貫した書き込み遅延を提供するために I/O フロー制御の改善
より高いスループット	オールフラッシュでの重複排除有効時のシーケンシャル書き込み I/O 性能を向上させる	• キャッシュでステージング中のプロアクティブな I/O バッファリング • 重複排除のタスク並行化
より迅速な再同期	オールフラッシュでの重複排除有効時の再同期時間を短縮する	• 異なる 書き込み I/O ステージ間のタスク並行化

より予測可能なアプリケーション性能を提供

LSOM 1.5 重複排除・圧縮有効時での一貫した性能の改善



シケーンシャル書き込みのスループットを向上

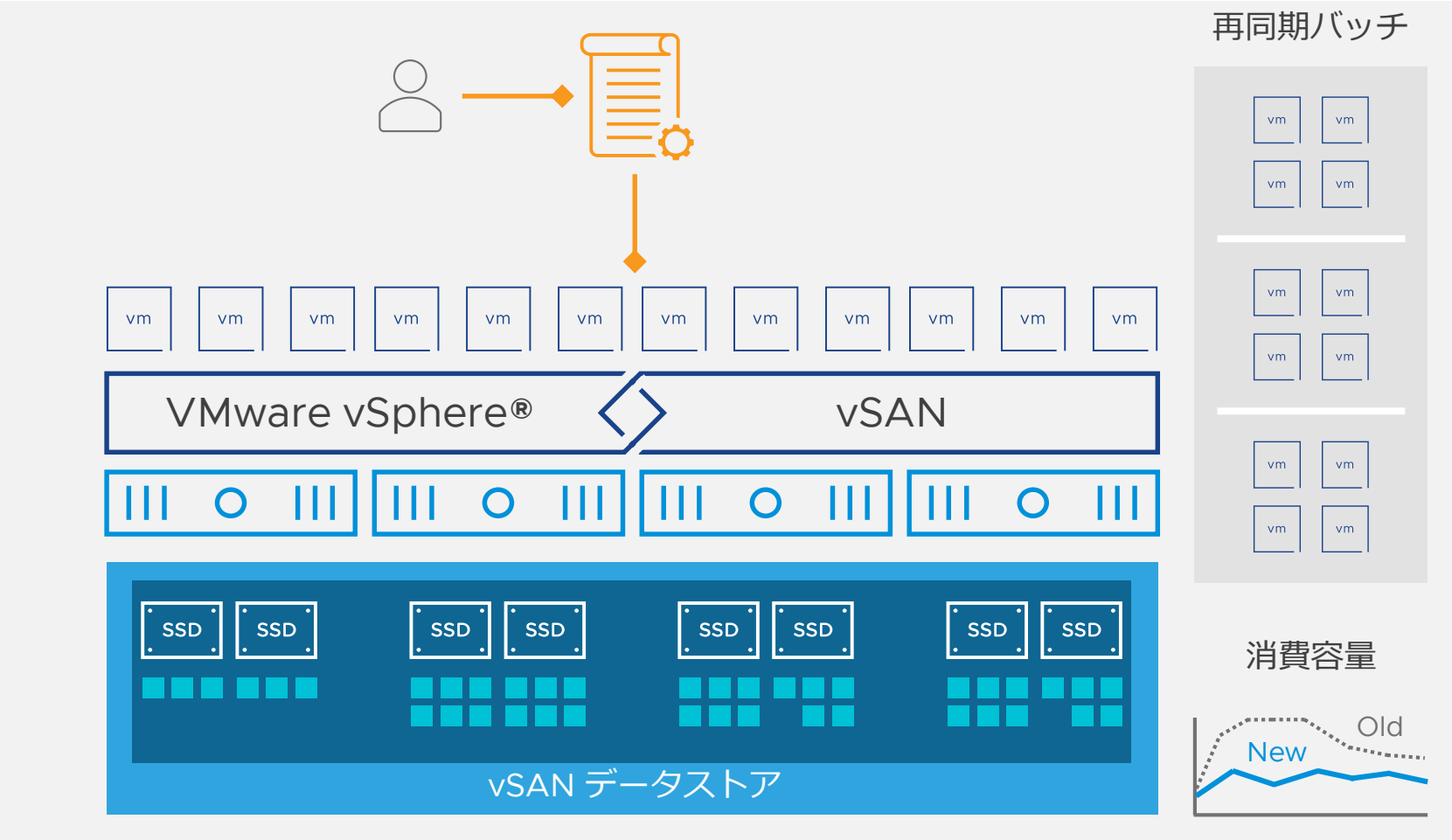
- 仮想マシンのスループット改善
- 再同期時間の削減

遅延にセンシティブなアプリケーションのための一貫性の向上

- 持続的な大量書き込み下での遅延をスムーズにする
- 高遅延と低遅延感の偏差をより小さく

ポリシー変更に伴う再同期時の容量管理の改善

ポリシー変更に伴う再同期のための一時領域使用量の自動制限



ポリシー変更の場合、vSAN は**バッチ処理**で再同期を実施

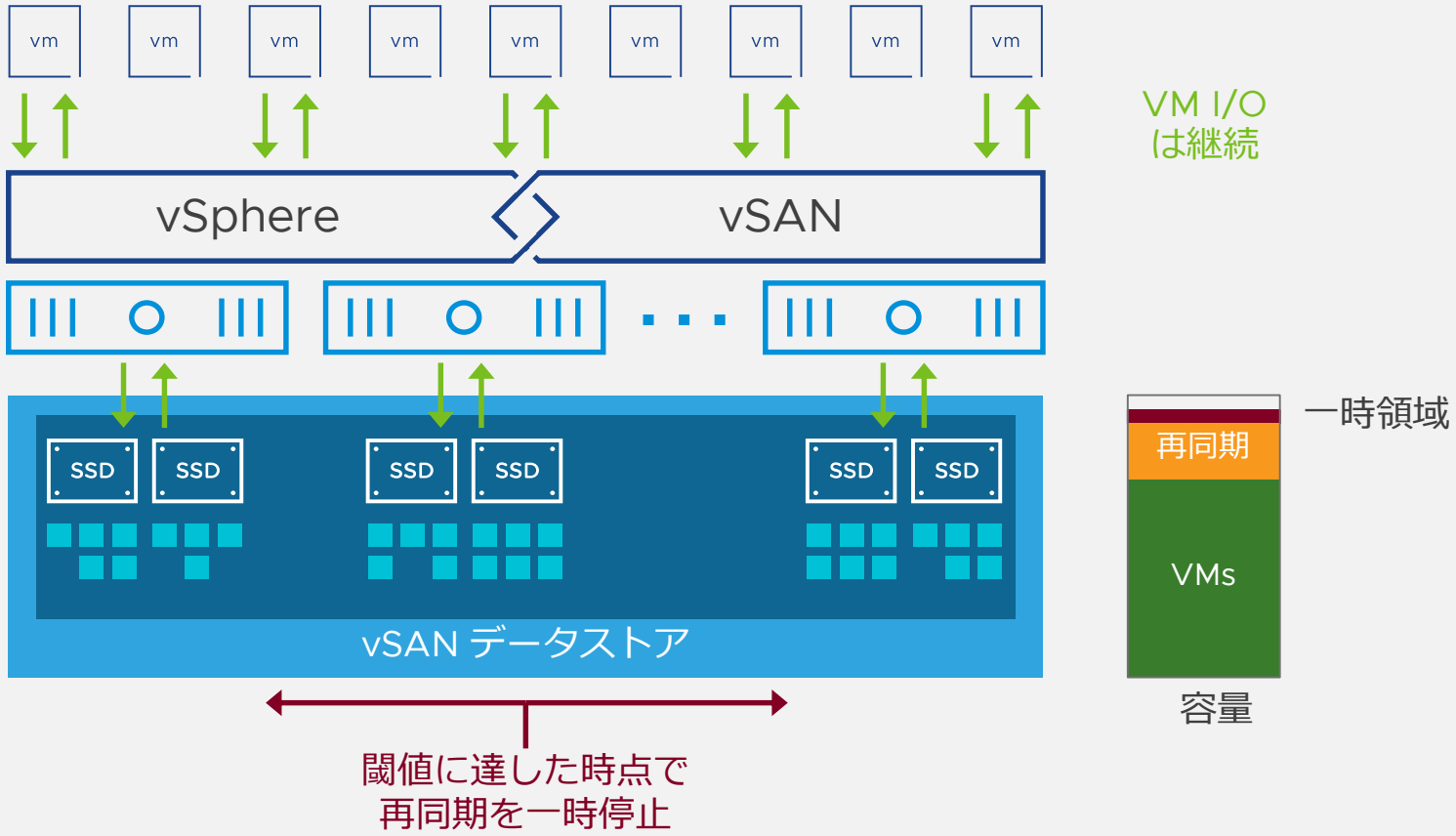
クラスタ全体で消費される一時**再同期領域の量**を**削減する**

複数の VM に影響するポリシー変更のユーザー操作を**簡素化**

VM のディスク領域不足の発生を回避

使用容量状況の堅牢な処理

再同期中の容量制限シナリオの強化



一時的な再同期処理の
自動容量管理

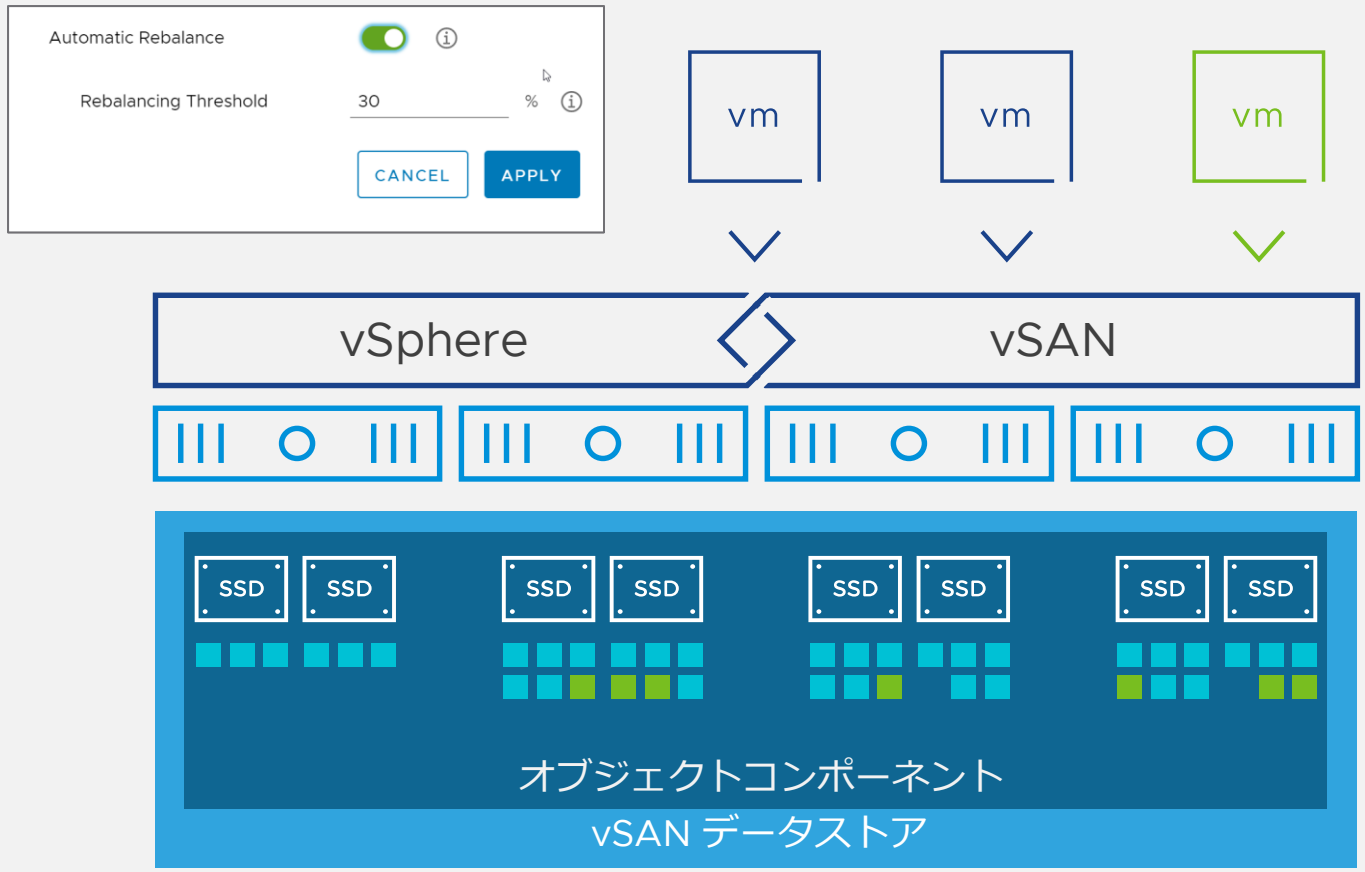
- 再同期処理によって制限値に達すると**一時停止**
- 空き領域が発生した後に再同期を**再開**
- VM の I/O は中断されない

ディスクグループ**完全な容量フル**シナリオの改善

- 容量の順次復旧のためのワークフロー
- vSAN データストアからの VMDK のアップロード / ダウンロードのサポート

自動インテリジェントデータリバランス

プロアクティブなリバランスの強化



完全自動化

- 手動による介入なし
- 必要に応じてリバランス

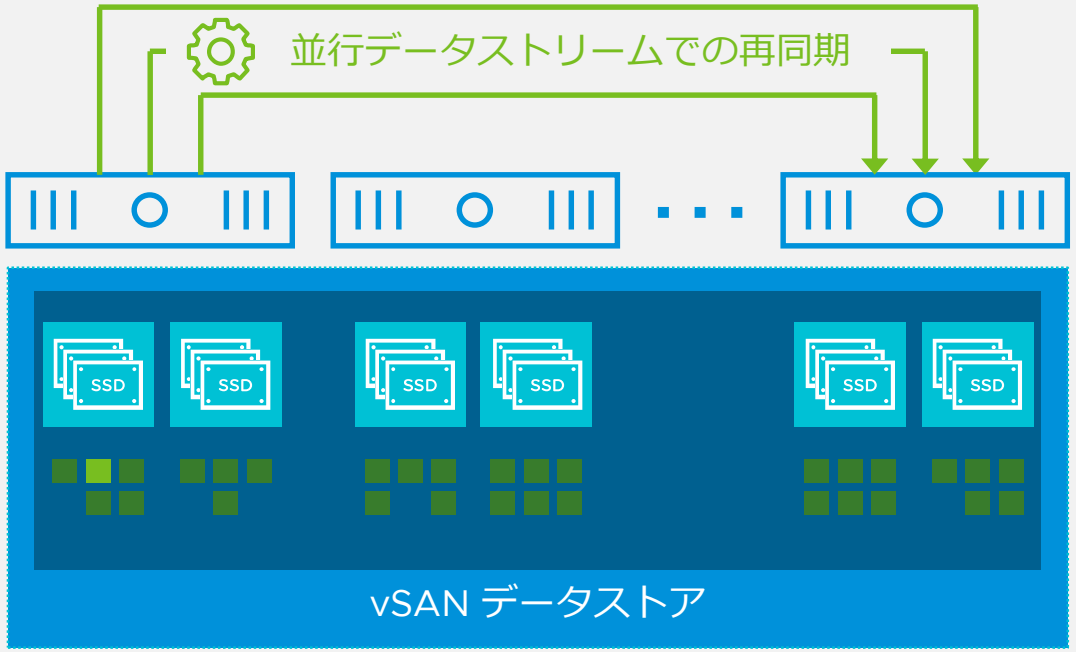
ユーザーインターフェイスから
クラスタレベルで**有効 / 無効**を
設定可能

ユーザーがカスタマイズ可能な
差異設定

ヘルスチェック調整の有効に
よる不必要なアラート抑制

インテリジェント I/O 管理による性能の改善

アダプティブな並行同期により、修復と再構築の時間を削減



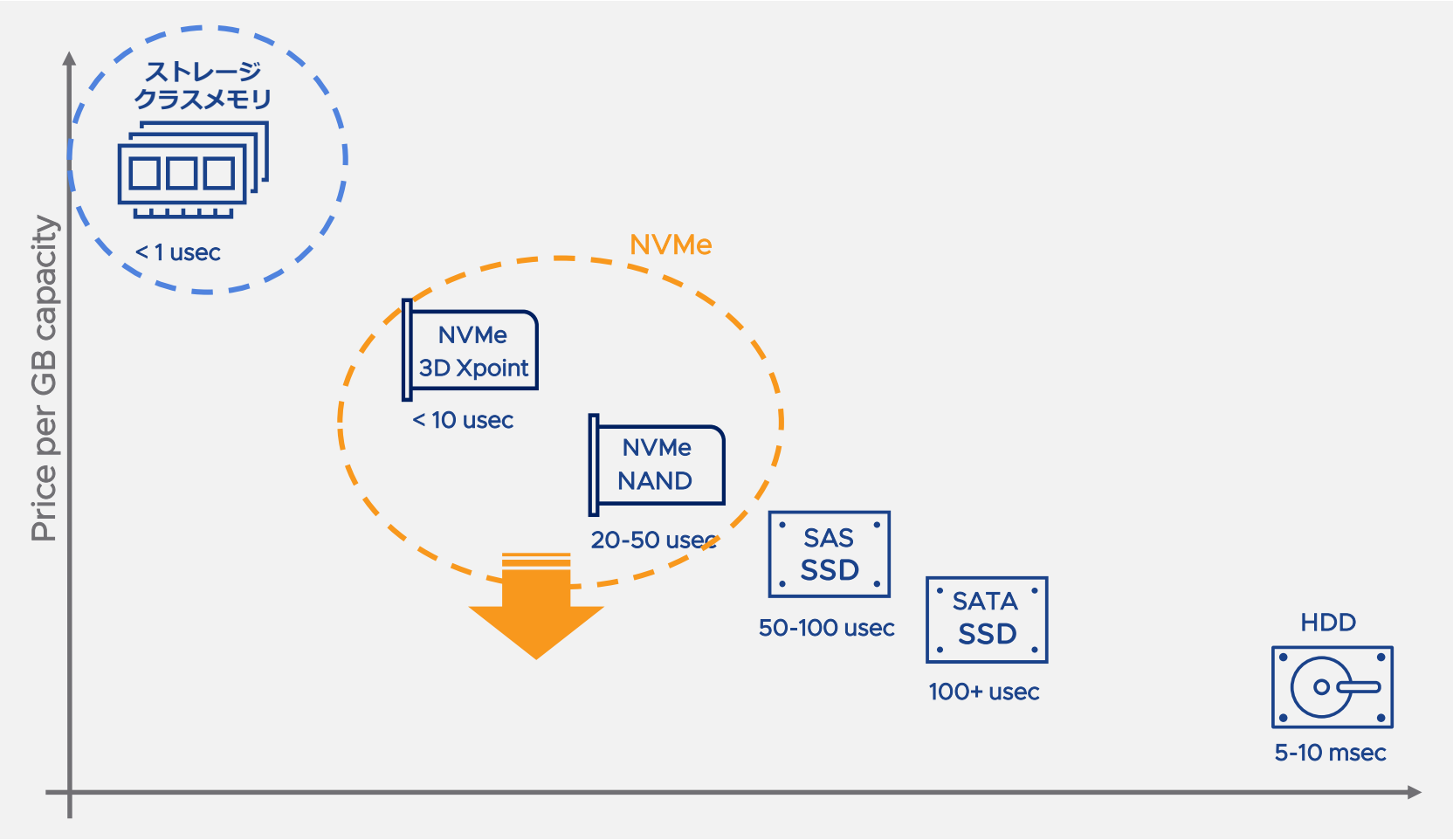
再同期タスクが環境の要求に素早く**適応**

再同期あたり数十での並列ストリームで実行可能

アダプティブ再同期によって管理される**リソース帯域幅**の範囲で利用

ハードウェアの進化

ストレージデバイス技術の進化が新しい可能性を提供する



3D NAND デバイスはより大きく、より安価になっている

耐久性は問題ではない

NVMe がデファクトのプロトコルになる

高性能を実現する新しい技術 (3D Xpoint)

PMEM (ストレージクラスメモリ) の市場はまだ初期段階

インテル® Optane™ SSD DC P4800X

ビルディング・ブロック

＋

インテル® インターコネクト IP

インテル® Software

インテル® メモリー / ストレージ・コントローラー

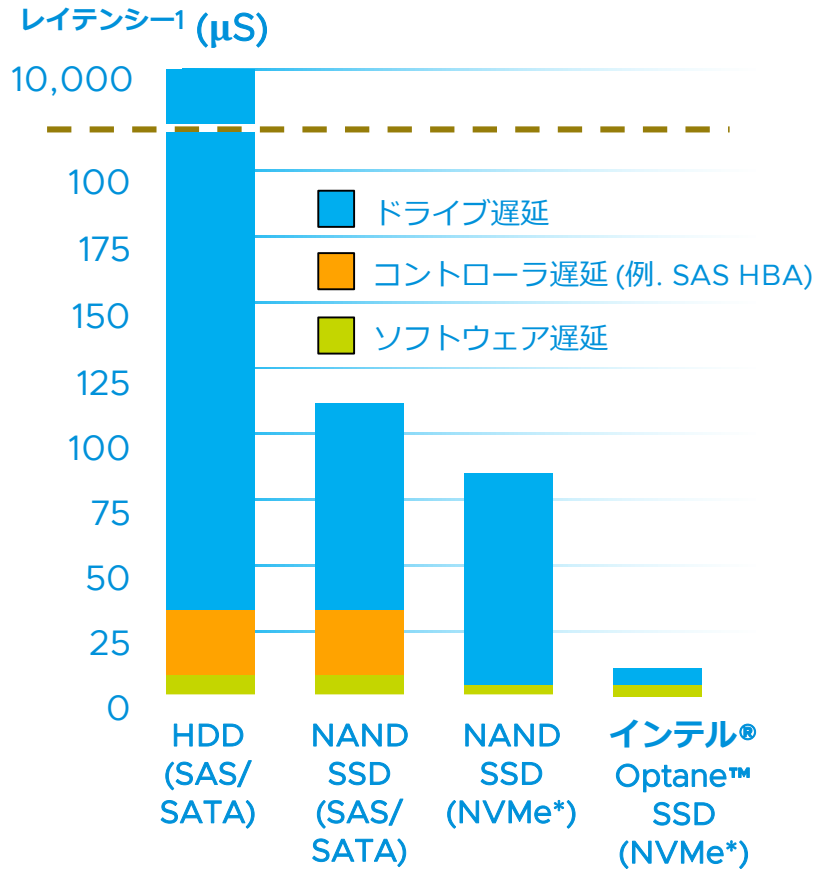
エンドユーザーへの価値[1]

高い
耐久

負荷への応答性
低遅延

予想以上に早いサービス
QoS

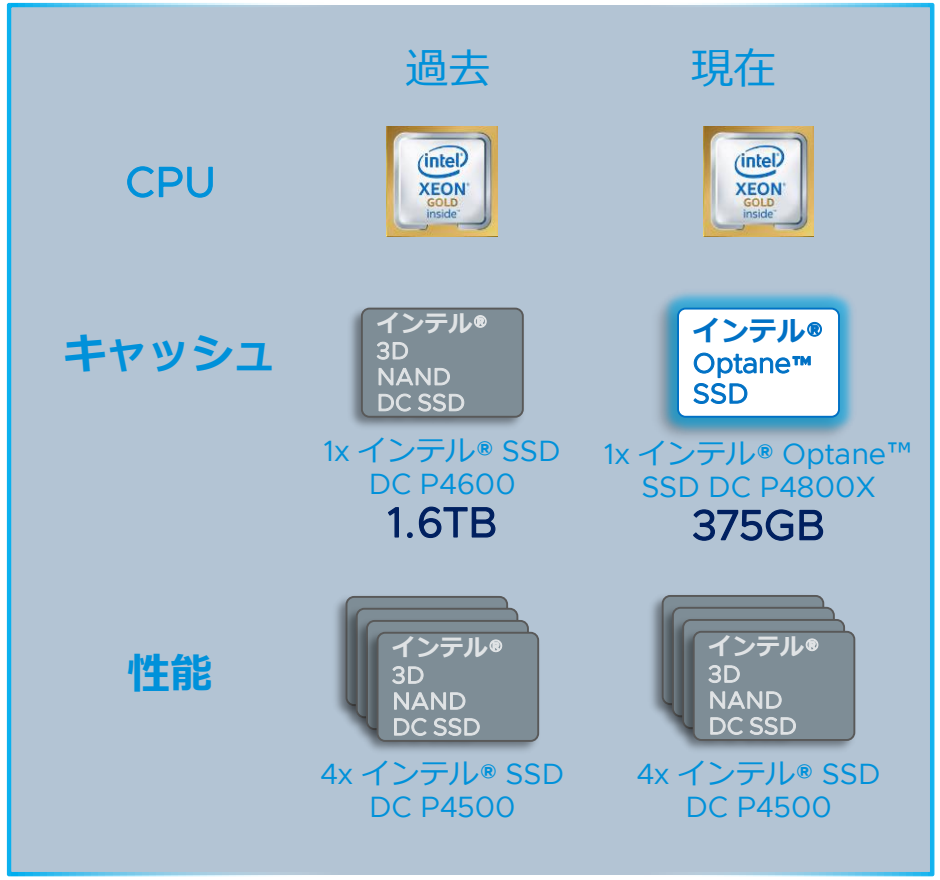
画期的な性能
IOPS



[1]End User Value Source – Intel-tested: Average read latency measured at queue depth 1 during 4k random write workload. Measured using FIO 3.1. Common Configuration - Intel 2U Server System, OS CentOS 7.5, kernel 4.17.6-1.el7.x86_64, CPU 2 x Intel® Xeon® 6154 Gold @ 3.0GHz (18 cores), RAM 256GB DDR4 @ 2666MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P4600 1.6TB. Latency – Average read latency measured at QD1 during 4K Random Write operations using FIO 3.1. Intel Microcode: 0x2000043; System BIOS: 00.01.0013; ME Firmware: 04.00.04.294; BMC Firmware: 1.43.91f76955; FRUSDR: 1.43. SSDs tested were commercially available at time of test Performance results are based on testing as of July 24, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks. * Other names and brands may be claimed as the property of others

VMware vSAN* とのソリューションによる価値

インテル® Optane™ SSD vs. 現在の TLC NAND



up to **60%** VM 密度の増加[1]

up to **$25%** 潜在的な 3 年間のコスト削減[1]

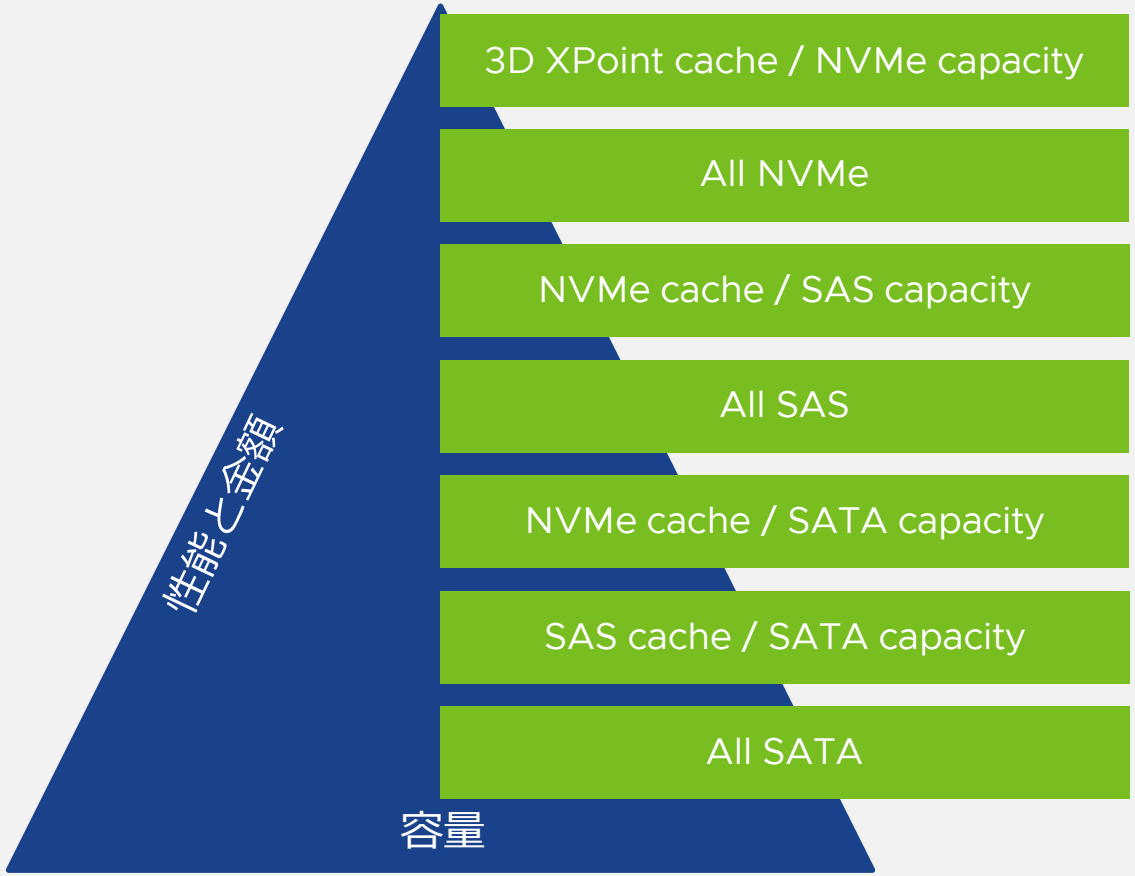
Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks. Performance results are based on testing as of October, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

[1] Tests by The Evaluator Group. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Configuration details available from The Evaluator Group at https://www.evaluatorgroup.com/document/lab-insight-latest-intel-technologies-power-new-performance-levels-vmware-vsan-2018-update/. See Appendix A for server cost estimate details and assumptions.

* Other names and brands may be claimed as the property of others

ハードウェアによる vSAN の性能

ストレージデバイス毎の性能 / コストピラミッド



オールフラッシュ：ハイブリッドよりも予測可能でレスポンスが高い

SATA プロトコルは 1 対 1でバスをロックするため、**出来るだけ避ける**

ストレージコントローラがボトルネックになる可能性がある

SAS エキスパンダーは限定的なサポート(性能のため)

NVMe は最速かつシンプルで**CPU のオーバーヘッドが低い**

現在と将来のハードウェアを活用する

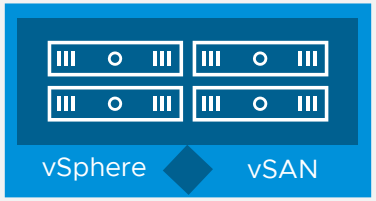
どの様に現在及び将来の進化に対応するか

CPU
より早く
より多くのコア数

NVMe
より高密度
よりシンプル
バイトアドレスアクセスの可能性

パーシステントメモリ
NAND の欠点を回避
DRAM よりはるかに安価

Networking
25 / 100Gb
RDMA



進化するアーキテクチャ
前例のない性能
より効率的なデータサービス
可用性の向上

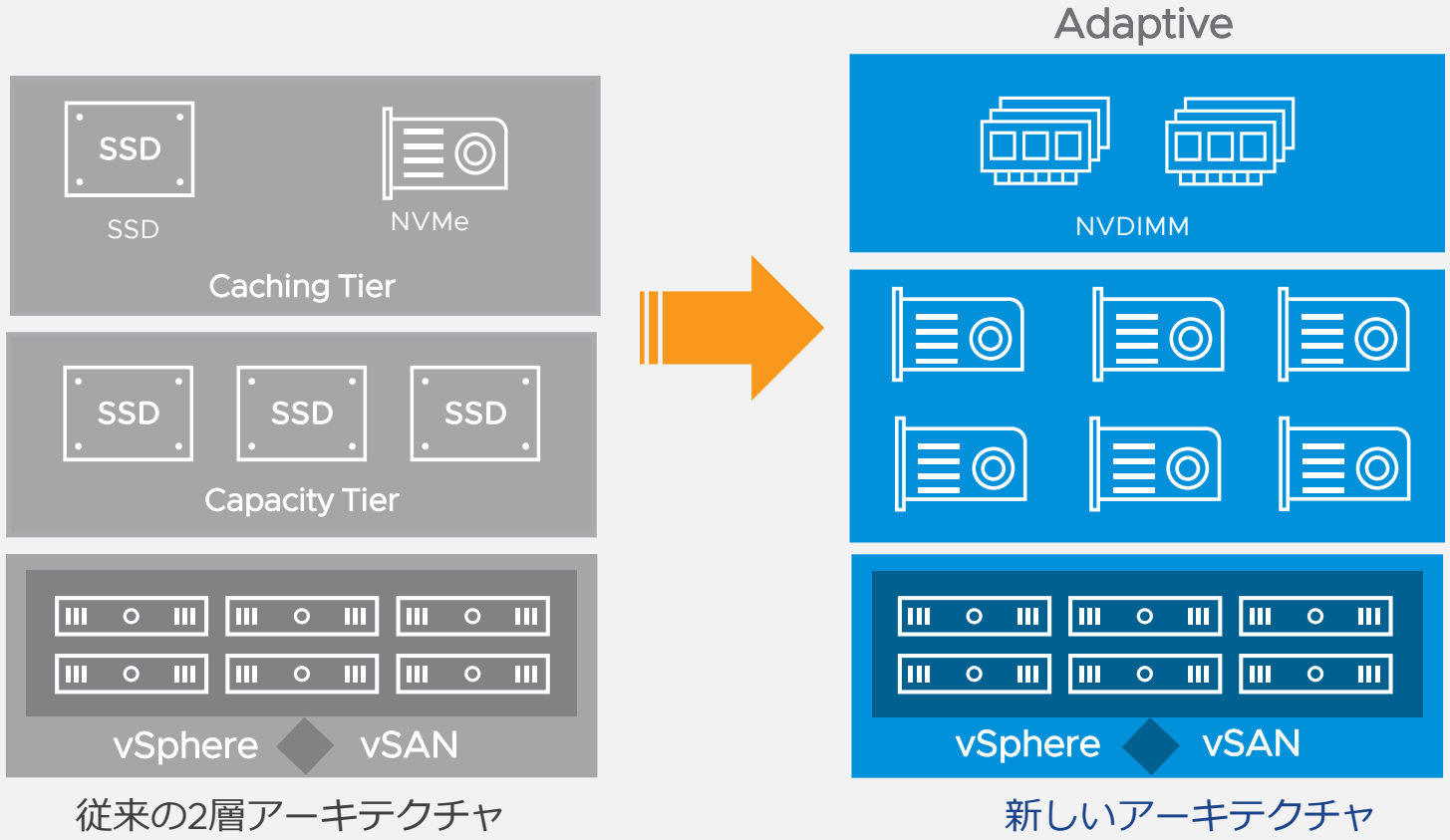
個別のハードウェアの進化が
ボトルネックを変える

最大限の性能を得るには
ソフトウェアの最適化が必要

性能が向上すると CPU の
効率性がより重要になる

次世代ストレージデバイスのための vSAN

ビジョンと方向性



従来の2層アーキテクチャ

新しいアーキテクチャ

極めて高性能が必要な
ワークロードの要件を満たす

All NVMe プラットフォーム上の
アプリケーションに最高の性能を
提供

優れた容量効率性を提供

迅速な再構築および再構成時間

IO ボトルネックの解消

性能検証データ比較

vSAN 6.7 U1 vs vSAN 6.7 U3

まとめ

- vSAN は機能だけでなく、**時代の流れやトレンド**に基づいて**常に最適化**を行なっている
- **デバイス**の進化は著しく、今後数年で大きな**ターニングポイント**をむかえる
- 従来の SAS ベースでは 10GbE で十分だが、**NVMe 以降**はネットワークに関しても
  より上位(**25 GbE,100 GbE**)を検討する必要がある
- **高性能**な**ハードウェアデバイス**を利用する場合、**ソフトウェア**も**最適化**されていなければ最
  大の性能を引き出すことは出来ない

Thank You

vmware®