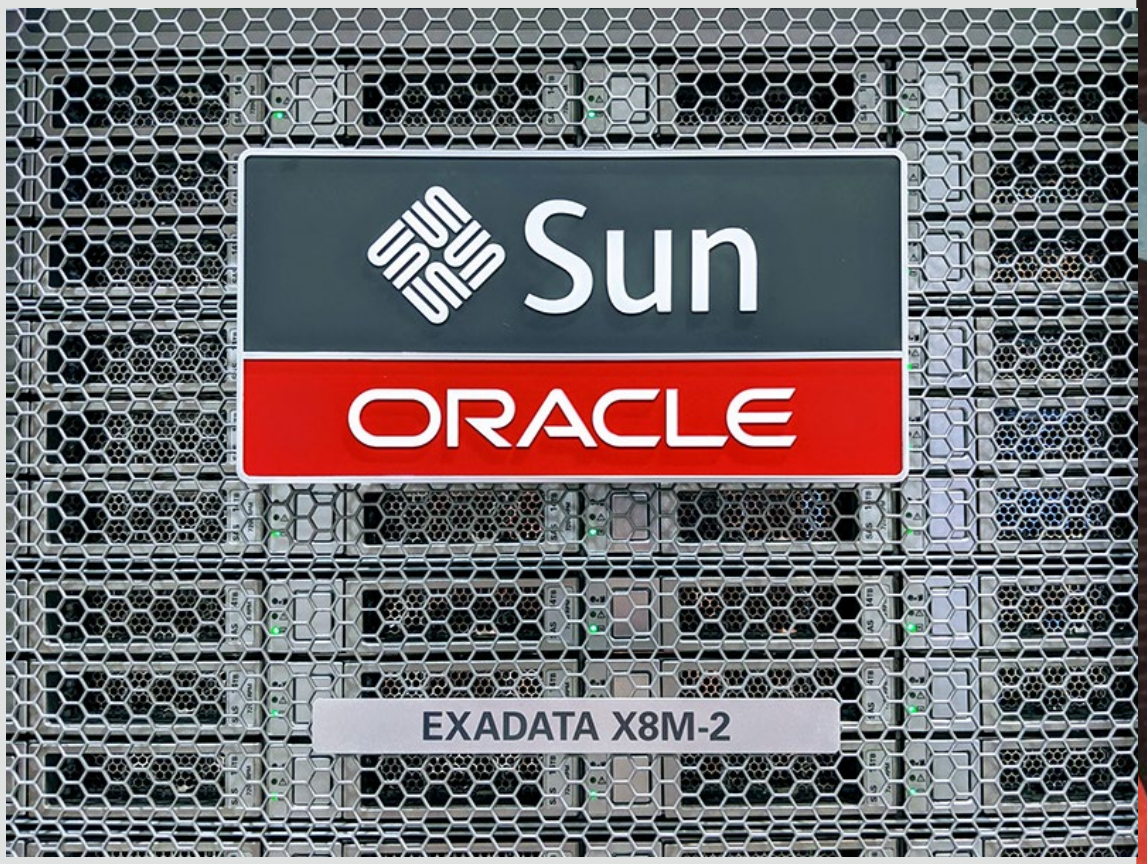
ORACLE

Oracle Database Technology Night

Exadata X8M



2020年 1月31日
日本オラクル株式会社
田口裕也

Safe Harbor Statement

以下の事項は、弊社の一般的な製品の方向性に関する概要を説明するものです。また、情報提供を唯一の目的とするものであり、いかなる契約にも組み込むことはできません。以下の事項は、マテリアルやコード、機能を提供することをコミットメント（確約）するものではないため、購買決定を行う際の判断材料になさらないで下さい。

オラクル製品に関して記載されている機能の開発、リリースおよび時期については、弊社の裁量により決定されます。

OracleとJavaは、Oracle Corporation 及びその子会社、関連会社の米国及びその他の国における登録商標です。

文中の社名、商品名等は各社の商標または登録商標である場合があります。

Exadata V2のスライドよりInfiniBandについて

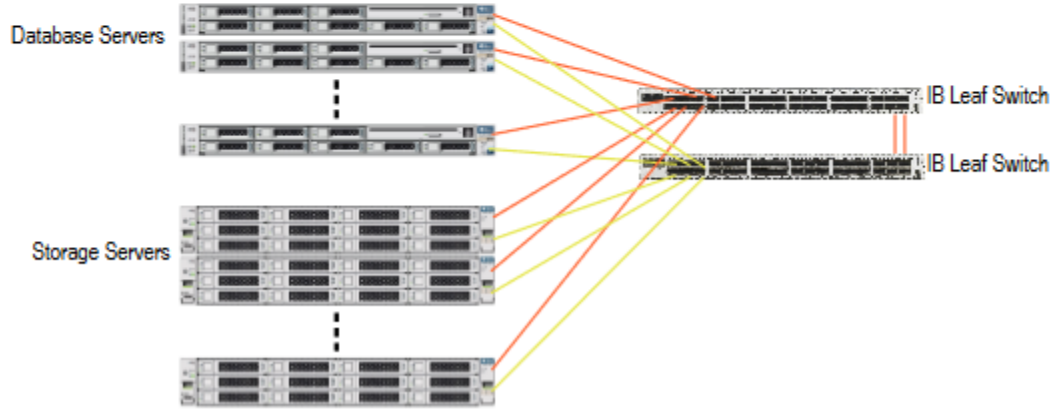
Agenda

- 既存データベースの課題
- Sun Oracle Database Machine
- 製品紹介～ Database Server, Storage Server
  - > 製品概要
  - > アーキテクチャ
  - > フラッシュテクノロジー
  - > ハードウェア構成
- 製品紹介～ InfiniBand
  - > 製品概要
- 製品紹介～ Sun Rack II
  - > 製品概要

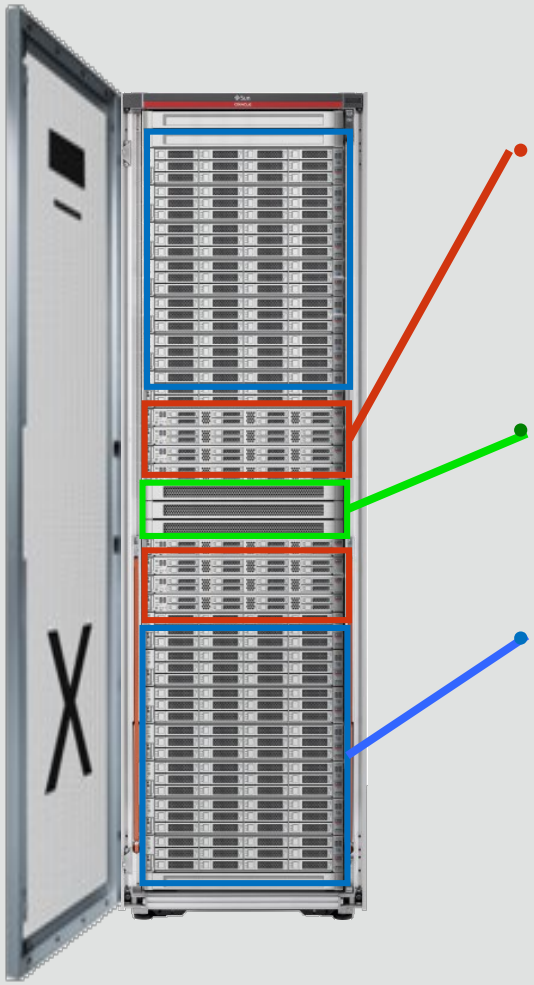
IB Network Topology

Quarter and Half Rack IB Network

- 2 台の InfiniBand 36 Leaf Switch
- 各サーバに 1 枚の Dual-port QDR InfiniBand HCA を実装
- 各サーバは 2 台の IB Leaf Switch に接続（冗長性確保）
- 各サーバは 2 ホップで接続（低レイテンシ確保）



Exadata X8Mの変更点

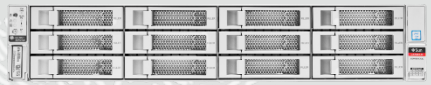


- **2ソケットのスケールアウト可能なデータベース・サーバー**
  - 最新の24コア Intel Cascade Lake
- **100Gb RDMA over Converged Ethernet (RoCE) の内部ネットワーク**
- **2ソケットのスケールアウト可能なインテリジェント・ストレージ・サーバー**
  - ストレージ・サーバーあたり 1.5 TB の**Persistent Memory**を搭載
  - 3階層のストレージ: PMEM, NVMeFlash, HDD
- **仮想化技術 KVMを新たに採用**

Database Server

High-Capacity (HC) Storage

Extreme Flash (EF) Storage



Extended (XT) Storage

Exadata X8Mの変更点

H/W

① 内部ネットワークに
100 Gbps RoCEを採用
（InfinibandからEthernetに変更）

100 Gbps RoCE （RDMA over Converged Ethernet）

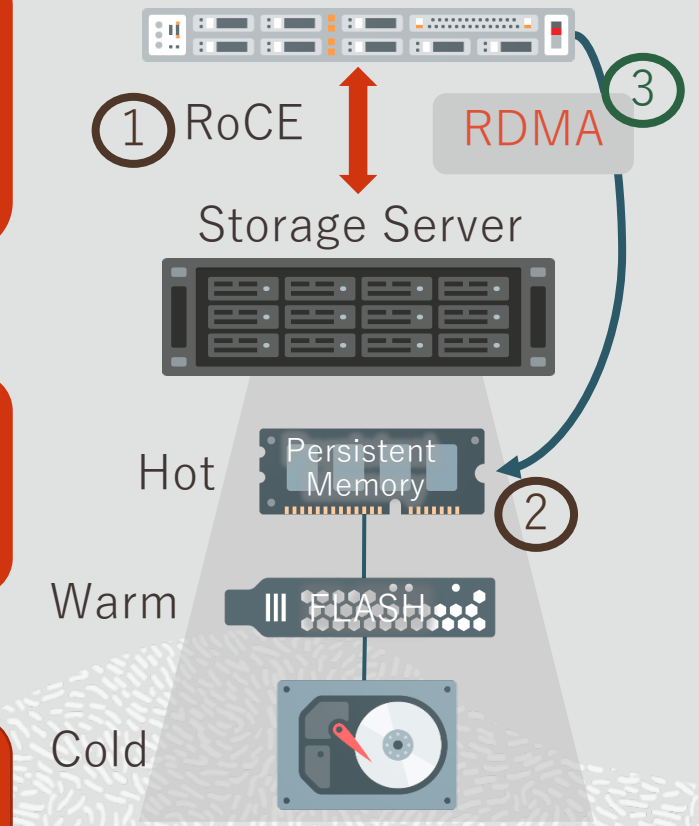
H/W

② Persistent Memory（PMEM）を
ストレージサーバーに標準搭載

Intel® Optane™ DC Persistent Memory（ストレージサーバーあたり 1.5TB）

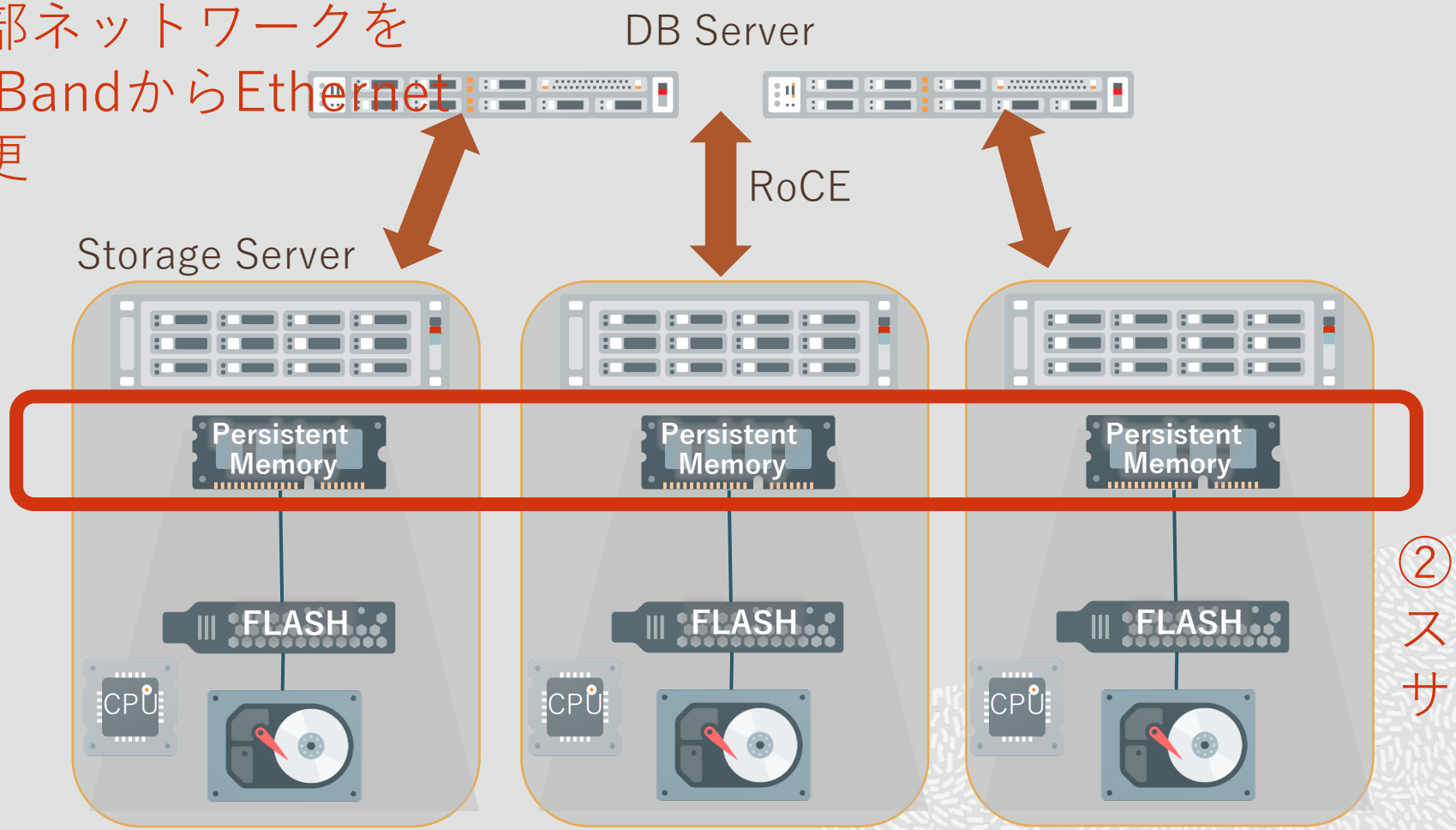
S/W

③ データベース からPMEMへ
I/OではなくRDMAでアクセス



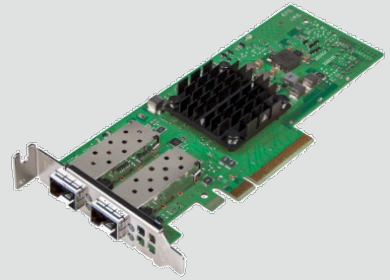
Exadata X8Mの構成イメージ

①内部ネットワークを
InfiniBandからEthernet
に変更



②PMEMは
ストレージ
サーバ側に搭載

① 新しいネットワーク構成（RoCE）



Exadataが最初にリリースされた時

10Gb Ethernet vs 20Gb InfiniBand
(and then 40Gb InfiniBand)

No RDMA on Ethernet vs RDMA on InfiniBand

現在

100Gb Ethernet vs 100Gb InfiniBand

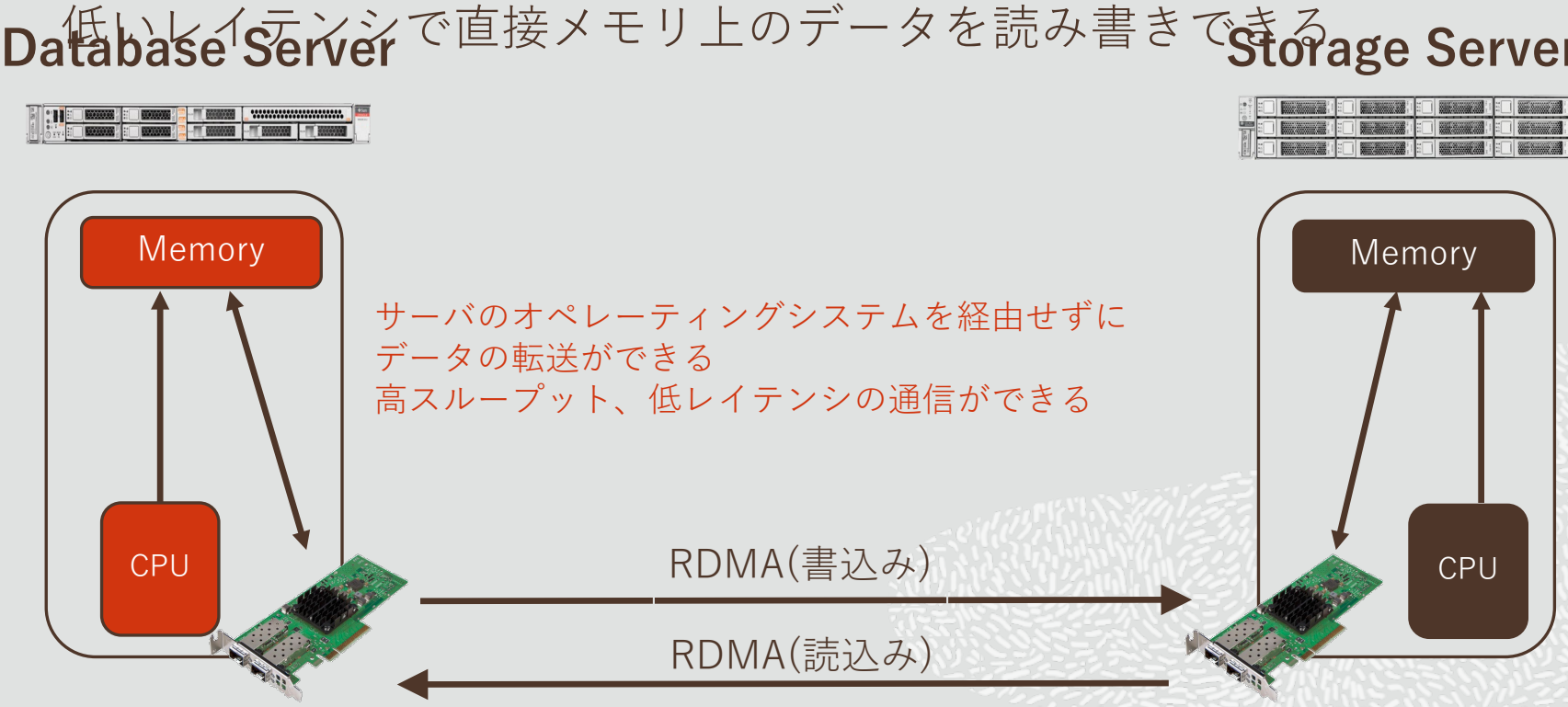
RDMA available now on Ethernet and InfiniBand

- Exadataの初期リリース時は
  InfiniBandだけがRDMAを利用できた
  現在はEthernetもRDMAを利用できる

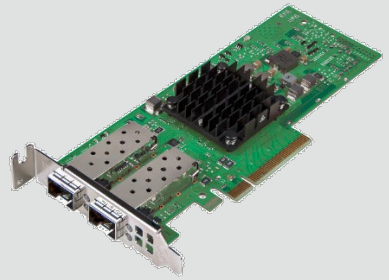
レイヤー	RoCE	InfiniBand
アプリ	ユーザー・アプリケーション	
	トランスポート (InfiniBand)	
ネットワーク	IP ネットワーク	InfiniBand ネットワーク
ハードウェア	Ethernet	InfiniBand

RDMA（Remote Direct Memory Access）について

- **RDMA(Remote Direct Memory Access)は、OSやCPUの関与なしにリモートサーバにあるデータにアクセスできる**
  - ネットワークカードが、追加のコピーやバッファリングせずに、低いレイテンシで直接メモリ上のデータを読み書きできる



RoCEは業界標準



1.オープン・コンソーシアムで定義された技術
   InfiniBand Trade Association (IBTA)
   オープンソースで開発され、Linuxアップストリームでメンテナンス

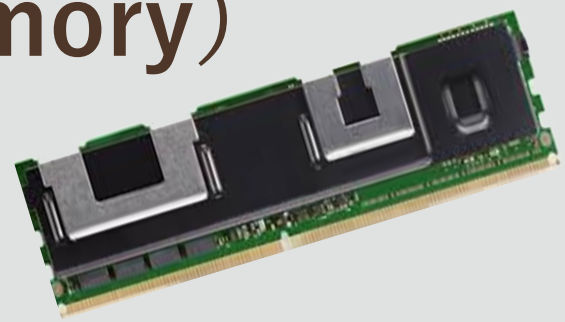
2.主要ネットワーク・カード・ベンダーがサポート:
Broadcom, Intel, Mellanox

3.主要スイッチ・ベンダーがサポート:
Arista, Cisco, Juniper, Mellanox

4.Exadata X8MはMellanoxのカードとCiscoのスイッチを利用

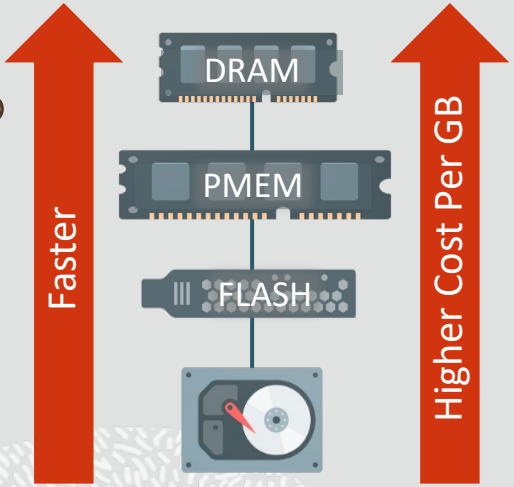
5.RoCEはネットワーク経由でリモートサーバのデータにアクセスができる

② 新しい メモリを搭載（Persistent Memory）



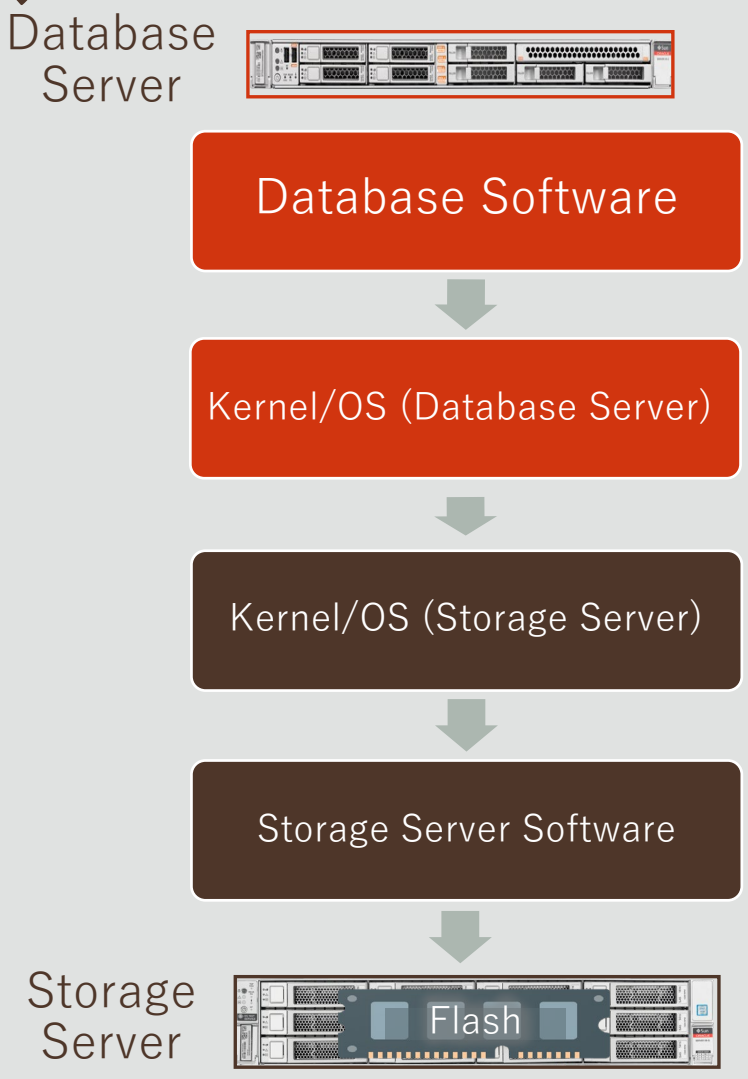
Persistent Memoryの特徴

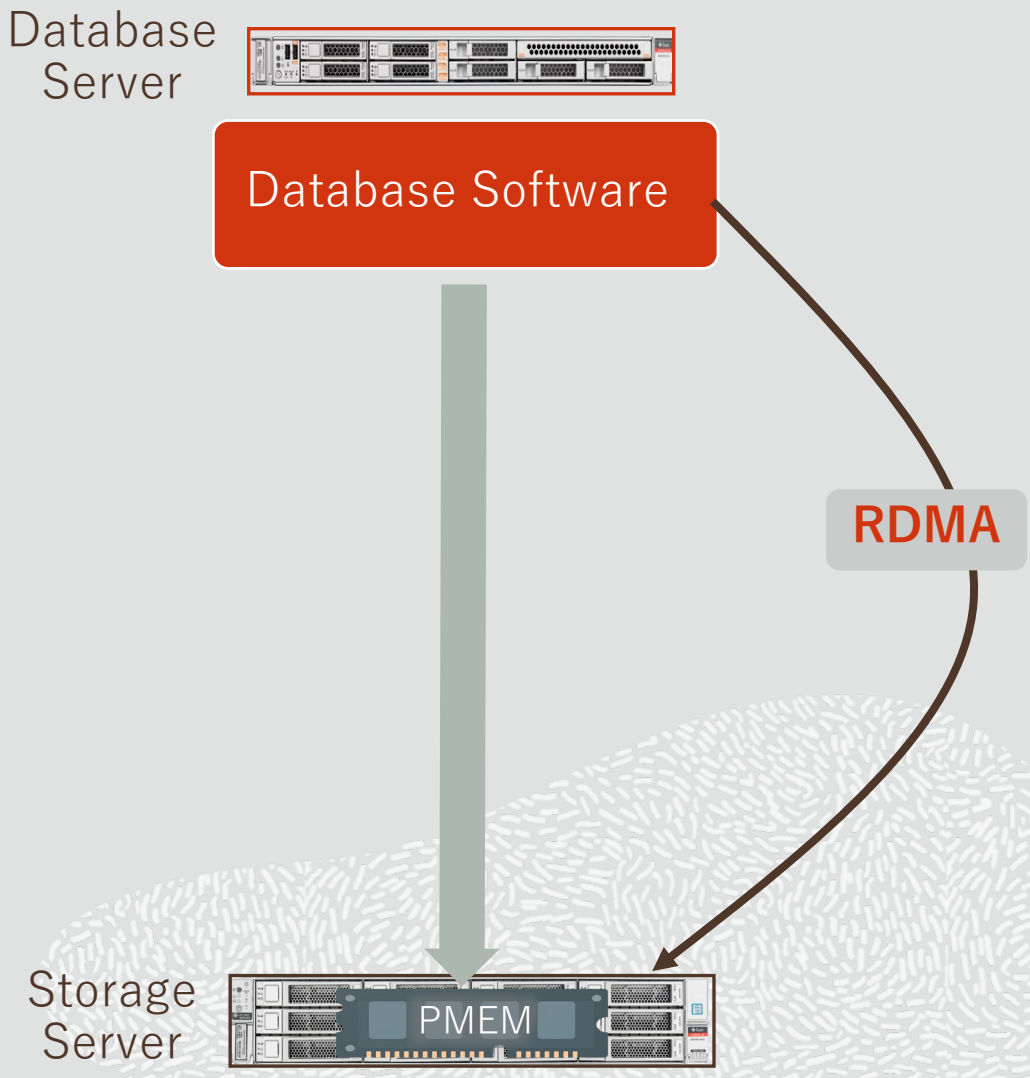
- DRAMの性能に近い読込み処理 – Flashよりかなり高速
- DRAMと違い、電源障害発生しても書込みしたデータが残る
- 容量、性能、価格はDRAMとFlashの間



Exadata X8Mはストレージサーバー上にPMEMを搭載
ストレージサーバーあたり 1.5TBを標準構成として搭載

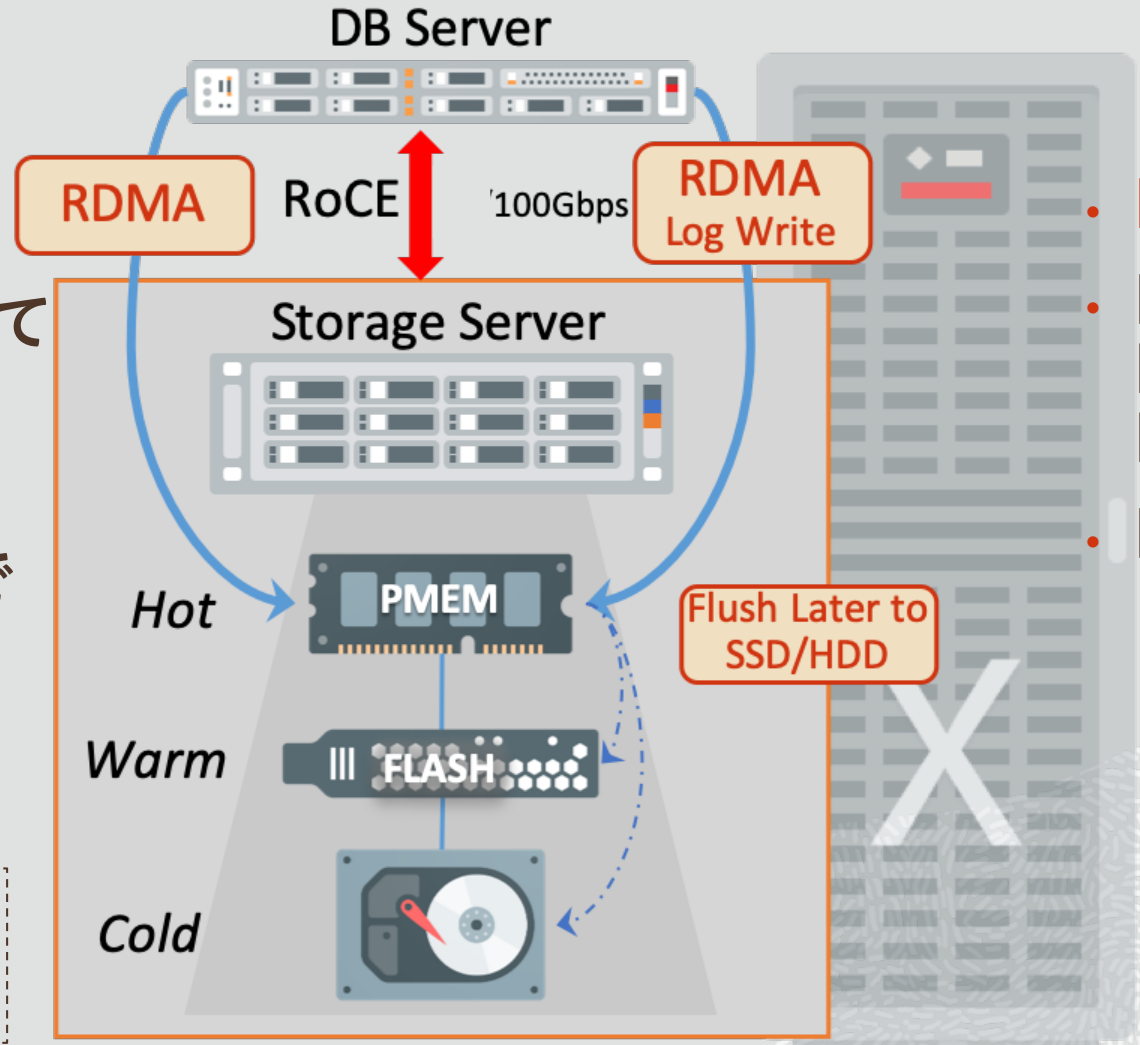
③I/Oではなく、RDMAを利用してPMEMにアクセス





X8MはPMEMをCacheとREDOログの書き込みに利用

- **PMEMCache**
- DBはRDMAを利用して Cacheを読み取る
- PMEMCacheは ストレージサーバ間で ミラー化して格納

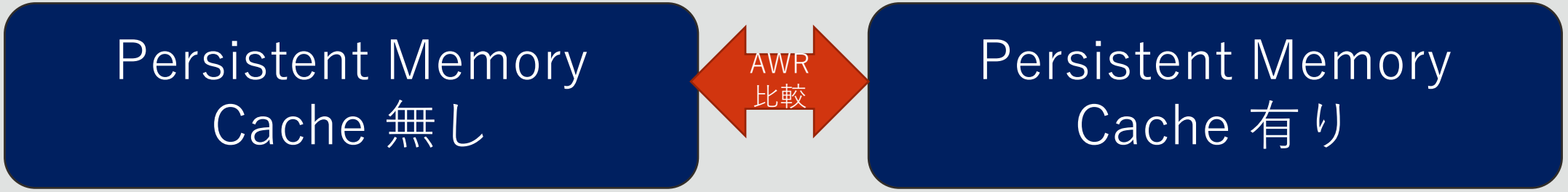


- **PMEMLog**
- DBのログバッファを RDMAを利用して PMEMLog上に書き込み
- PMEMLogにREDOログを書き込み （一方通行、ACKなしで完了）

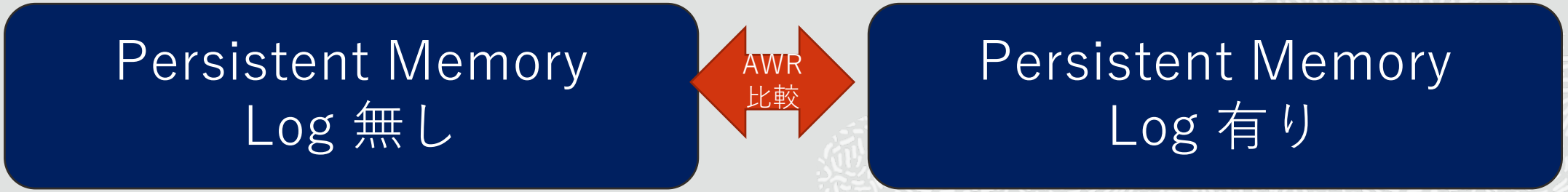
この動作は以下の要件が必要
・Exadata System Software 19.3.0以降
・Oracle Database 19c以降
・Exadata X8M

Exadata X8Mの動作

動作①

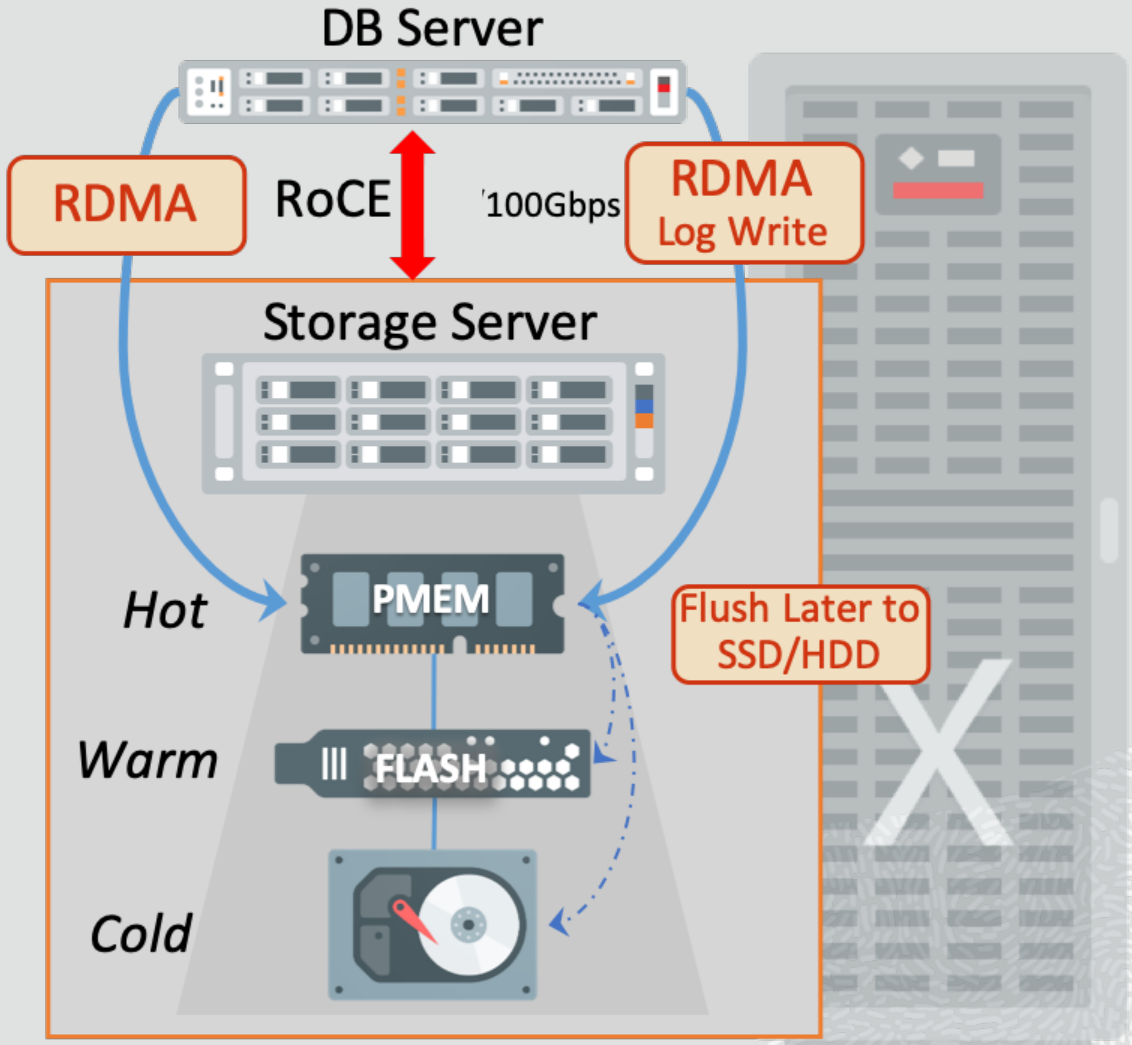


動作②



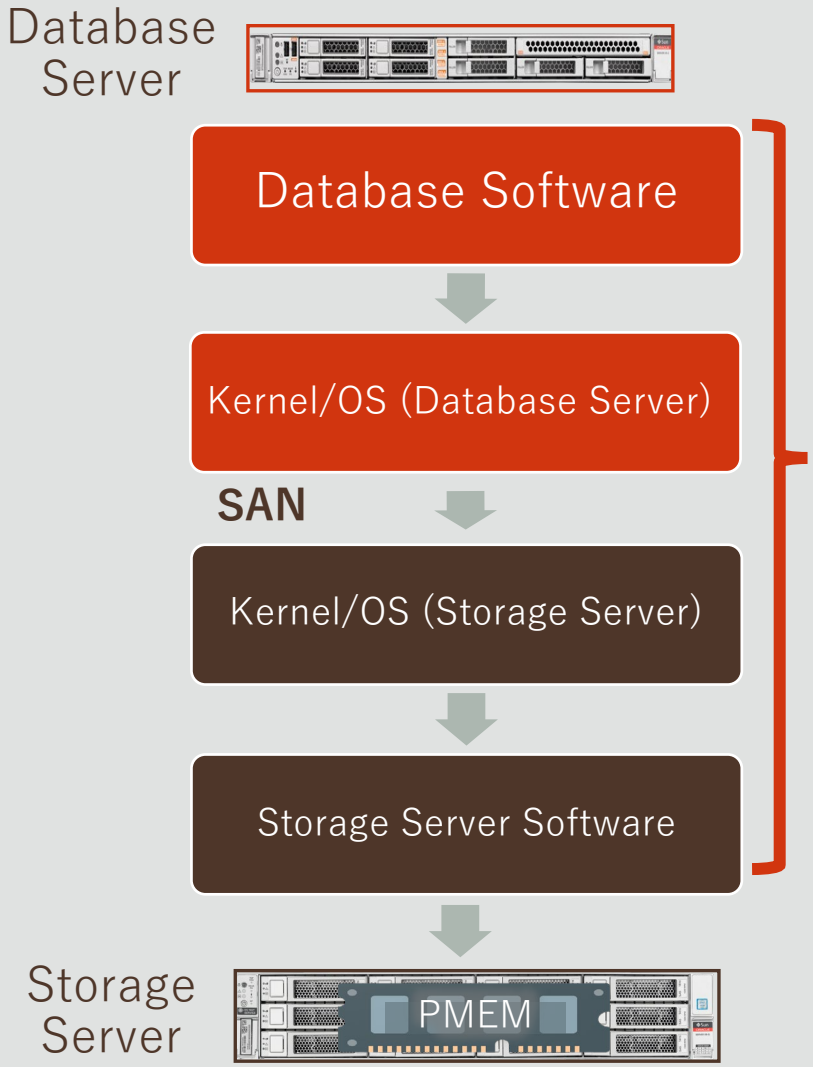
データベースは、I/Oではなく RDMAを利用してPMEMのデータにアクセス

- OLTP処理の高速化
- ネットワーク、I/Oソフトウェア、割り込み、コンテキストスイッチをバイパスするので、**レイテンシ10倍改善、IOPS 2.5倍高速**



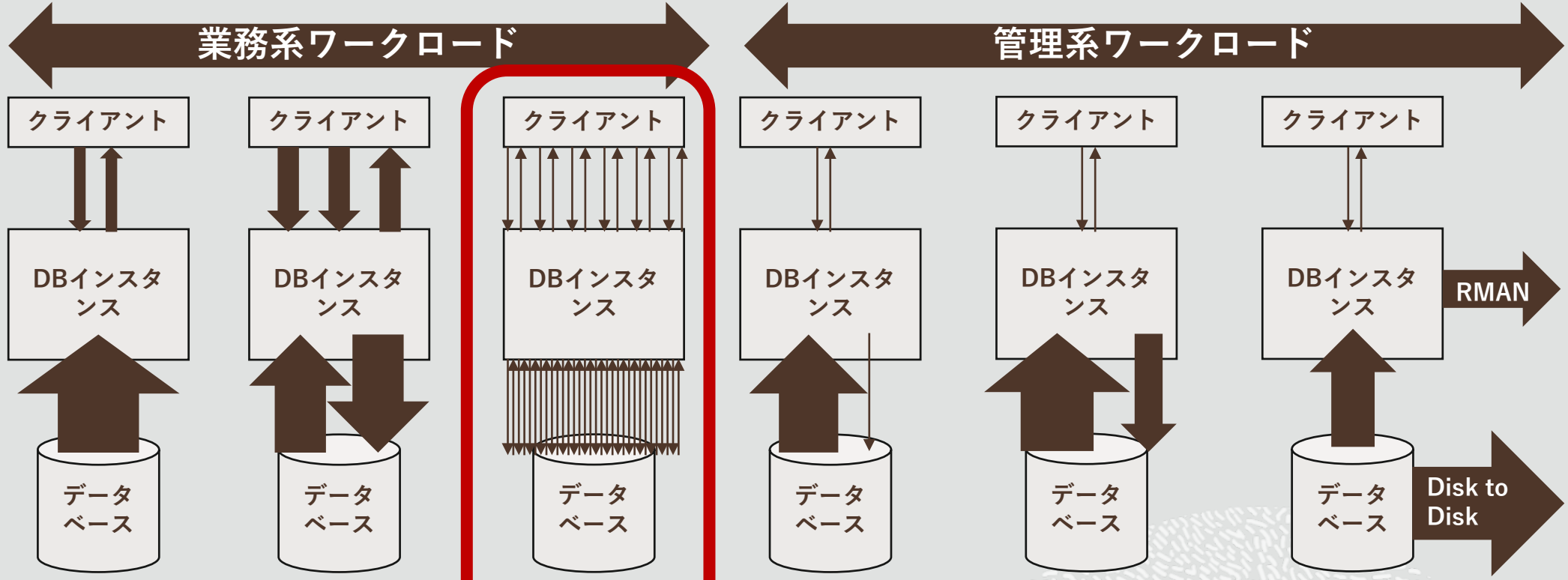
- REDOログ書込みの高速化
- ネットワーク、I/Oソフトウェア、割り込み、コンテキストスイッチをバイパスするので、**最大8倍のログ書き込みを高速化**

PMEMだけではストレージのI/O高速化はできない



- PMEMの読込みの性能は、ネットワークとI/Oソフトウェア、割り込み、コンテキスト・スイッチなどの処理が負荷となり活かしきれない

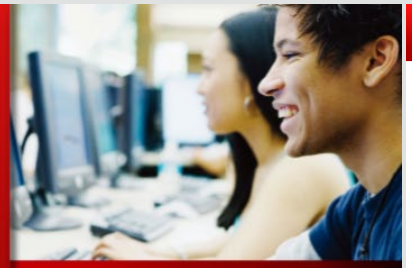
システムのワークロードとExadataX8Mの効果



	データウェアハウス 検索・分析	バッチ、大量帳票	オンライン アプリケーション	オプティマイザ 統計情報集計	索引再作成	バックアップ取得
1.	Smart Scan：○	：○	：△	：△	：○	：NA
2.	Smart Flash Cache：○	：○	：○	：○	：○	：○
3.	Smart Flash Logging ：NA	：○	：○	：○	：○	：NA
4.	Hybrid Columnar Compression：○	：○	：△	：○	：○	：○
5.	Persistent Memory Data Accelerator：○	：○	：○	：○	：○	：○

Exadataのチューニング方針のスライドの例

ORACLE



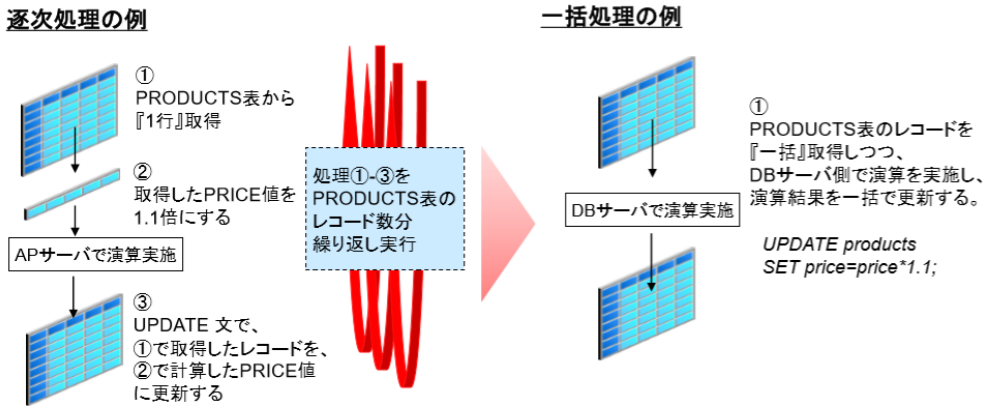
オラクル・コンサルが語る！
Oracle Exadataの性能を引き出す
アプリケーション設計/開発の勘所

日本オラクル株式会社

#odddtky

逐次処理と一括処理の例

『PRODUCTS表のPRICE列を全て1.1倍にする』



逐次処理（ループ）はExadataの性能を引き出せない

Exadataの高いI/O性能を如何に引き出せるかが重要

【Question】
なぜ、AとBのケースで高速化の程度に違いがあったのでしょうか？

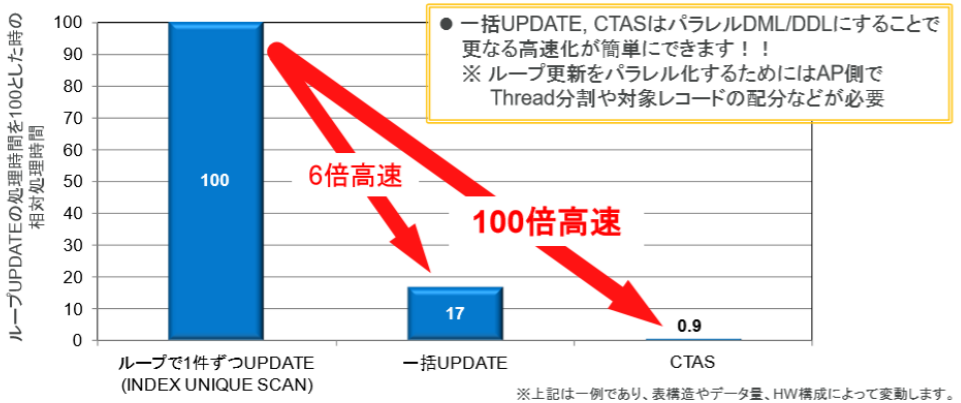
【Answer】
アプリケーションロジックがループ処理中心の逐次処理だったから

Exadataの最大の特徴は、「**広いI/Oバンド幅**」と
「**Smart ScanやStorage IndexによるI/O削減機能**」なので、
1処理あたりのI/O量が小さいループによる逐次処理ではなく、
1処理あたりのI/O量が大きい一括処理で性能を最大限に発揮できます。

書き換えパターンごとの処理時間比較

Tips-1

1,000万件のテーブルの全件更新処理 on Exadata



※上記は一例であり、表構造やデータ量、HW構成によって変動します。

まとめ
Exadata X8Mについて各アナリストからコメント
Analyst Reaction to Exadata X8M

IDC ANALYZE THE FUTURE

“この新しいExadataは、何の変換も構築も適用のためのプログラミングも不要で、その効果は本当に並外れている”

“this new system, which requires no conversion, configuration, or programming effort to adopt, offers benefits that are truly extraordinary.”

“Exadata X8Mは全てのデータベースシステムやDIYデータベースシステムに圧勝する”

“Exadata X8M annihilates every database system or DIY database system”

ESG Enterprise Strategy Group

“(19μ秒の)レイテンシという主要な競合の5倍以上から始まり、このマーケットセグメントでまっとうな競争相手としての選択肢を探すのが本当に難しい”

“With latencies (of 19 microseconds) that start at 5 times better than the most popular competitor, it is genuinely hard to view most alternatives as legitimate competition in this market segment”

Ovum

OracleがPersistent Memoryへ直接アクセスするためにRDMAを実装した初のシステムだ…そして大きなOLTP性能が得られる

Oracle has become the first to implement RDMA for direct access to persistent memory…And that’s where the big OLTP performance gains come

ORACLE