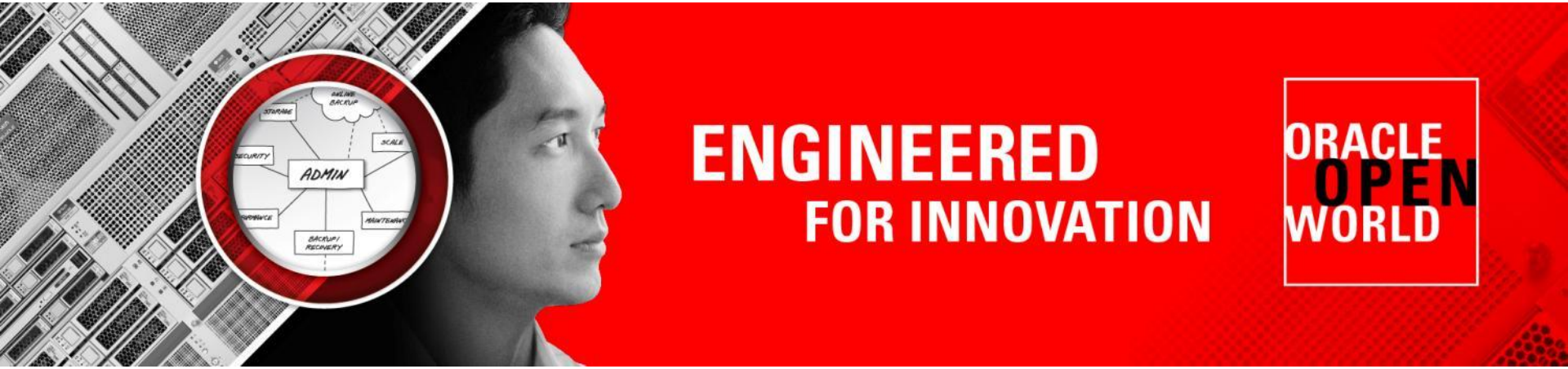


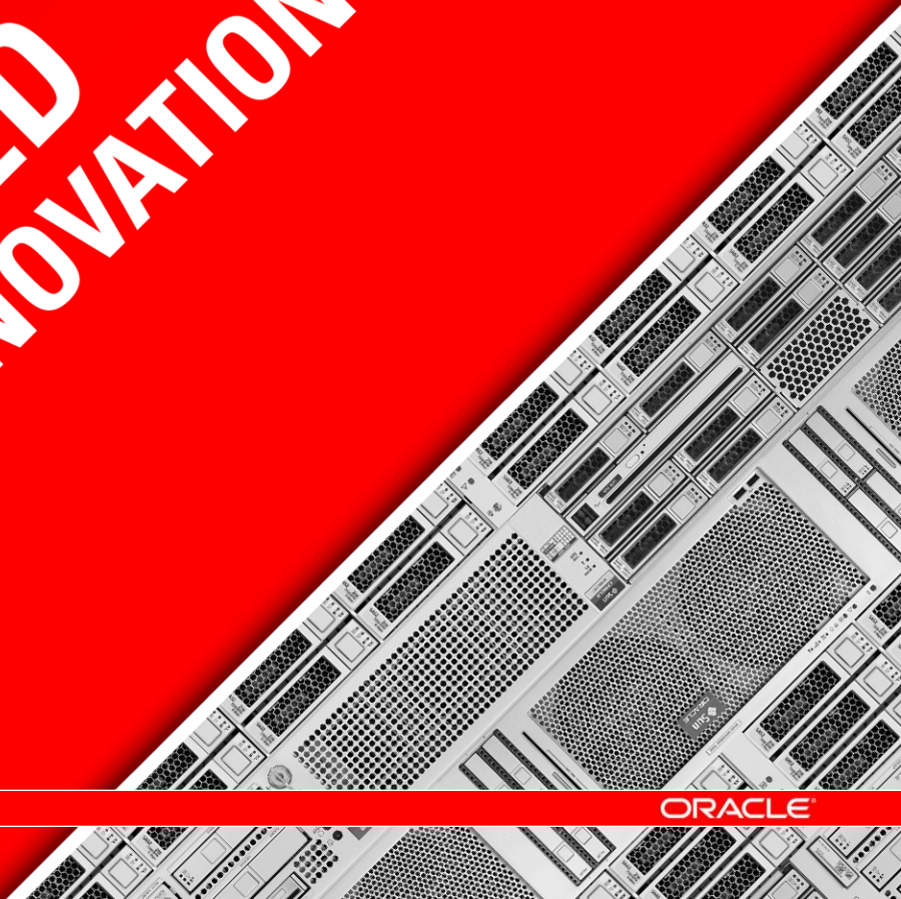
ORACLE®

Oracle Exadataの代表的な機能の特長とその仕組み

テクノロジー製品事業統括本部 技術本部 Exadata技術部 エンジニア
赤木 維磨

ORACLE OPEN WORLD

ENGINEERED FOR INNOVATION



ORACLE

ORACLE DEVELOP

Russia

17–18 April 2012

India

3–4 May 2012

San Francisco

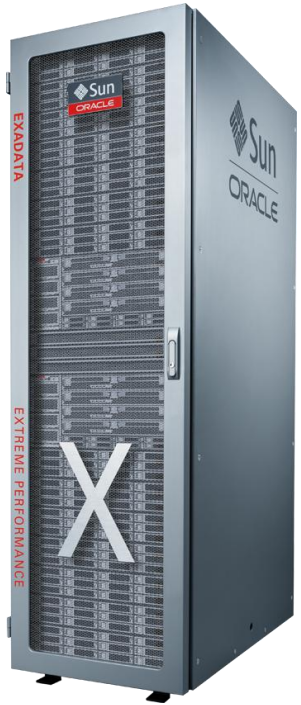
September 30–October 4, 2012

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。

Exadata概要

Oracle Exadata Database Machine

Oracle Databaseに最適化されたEngineered System



Hardware(H/W)

- DBサーバー
  - マルチコア
- InfiniBand
  - 広帯域、低遅延ネットワーク
- Storageサーバー
  - Flashカード
  - 大量ディスク

×

Software(S/W)

- Oracle Database
  - 長年培ってきたデータベース技術
  - Gridアーキテクチャー
- Exadata Storage Server Software
  - I/Oのボトルネックを排除
  - Flash機能
  - リソース制御機能
  - 高圧縮機能

- Best for DWH / OLTP / Consolidation

Hardware概要

製品ラインナップ

Oracle Exadata X2-2

- Quarter, Half, Full and Multi-Racks
  - QuarterからHalf、HalfからFull、Fullから複数Rackへと拡張可能

Oracle Exadata X2-8

- Full and Multi-Racks
  - Fullから複数Rackへと拡張可能

Exadata Hardware アーキテクチャー

Databaseサーバー

- X2-2

  - 2 x 6-core processor
  - 96GB Memory

- X2-8
  - 8 x 10-core processor
  - 2TB Memory

InfiniBandネットワーク

- 36-port 40Gb/s switches
- 一つに統合されたサーバー間ネットワーク

Storageサーバー

- 12 disks

  - Disk Type
    - 600 GB High Performance disk
    - 3 TB High Capacity disk
- 384 GB PCIe Flash

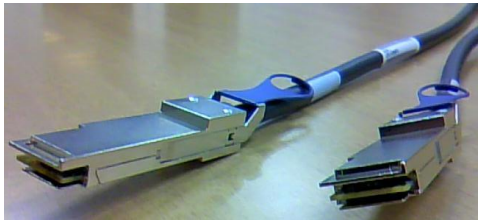
Oracle Exadataの各モデルのH/W構成

	X2-8 Full	X2-2 Full	X2-2 Half	X2-2 Quarter
Database Servers	2	8	4	2
Database CPU Cores	160	96	48	24
Database Memory (GB)	4096	768 (max 1152)	384 (max 576)	192 (max 288)
InfiniBand switches	3	3	3	2
Ethernet switch	1	1	1	1
KVM	No	Yes	Yes	Yes
Exadata Storage Servers	14	14	7	3
Storage CPU Cores	168	168	84	36
Storage Disks	168	168	84	36

InfiniBand & Flashカード

広帯域/低遅延ネットワークとランダムI/Oの高速化

- InfiniBand ネットワーク
  - Sun Datacenter 36ポート Managed QDR(40Gb/s) InfiniBandスイッチ
  - Host Channel Adapter(HCA)
    - InfiniBand用のインターフェース
  - 広帯域/低遅延/CPUオーバーヘッドの小さいネットワーク
    - ネットワーク帯域のボトルネック排除
    - 低遅延通信によりRACのリニアなスケーラビリティを実現

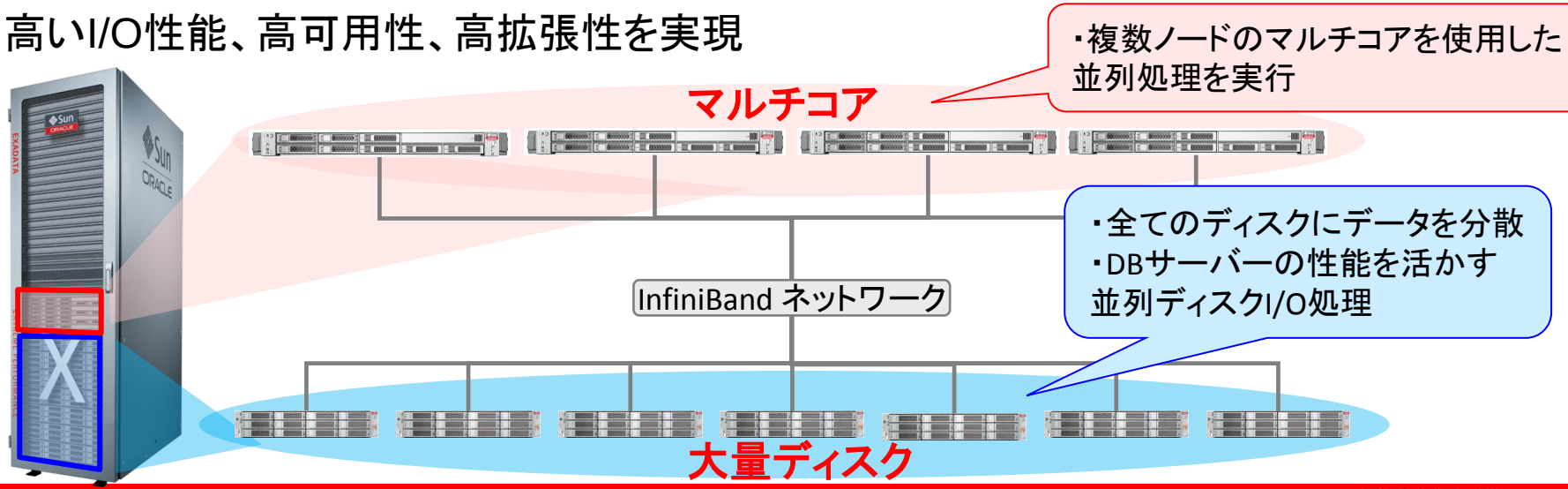


- Flashカード
  - StorageサーバーにFlashカードが搭載
  - Exadataのディスクと比較し約30倍のIOPSを実現
    - ランダム読み込み/Redo書き込みの高速化

Software概要

マルチコア/大量ディスク × Gridアーキテクチャー

- データベース層：Oracle Real Application Clusters(RAC)で仮想化
  - 高可用性、リニアなスケーラビリティ、高拡張性を実現
- ストレージ層：Oracle Automatic Storage Management(ASM)で仮想化
  - 高いI/O性能、高可用性、高拡張性を実現



Exadata Storage Server Software

- for DWH

  システム性能のボトルネックになりやすい「I/O」を効率化

  – Smart Scan、Storage Index
  – Exadata Hybrid Columnar Compression(EHCC)

- for OLTP

  Flashテクノロジーによる、安定した高IOPSの実現

  – Smart Flash Cache、Smart Flash Log

- for Consolidation

  ワークロードごとのリソースを動的かつ容易に制御可能

  – I/O Resource Manager

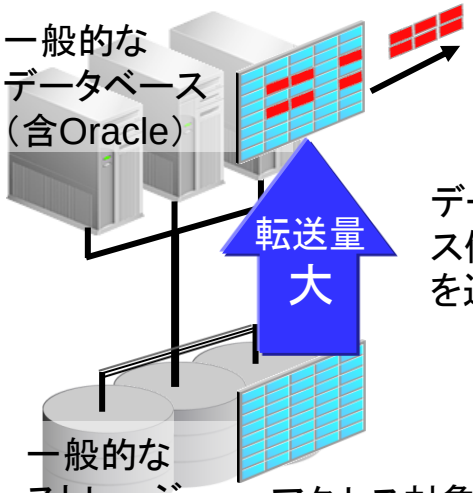
Smart Scan

for DWH

Storageサーバーへの処理のオフロード

ストレージが問い合わせを解釈し、必要なデータだけをDBサーバーへ転送する
サーバーとストレージ間のI/O量を最小限に留め、安定した性能を実現

従来のデータベースシステム



③計算・集計

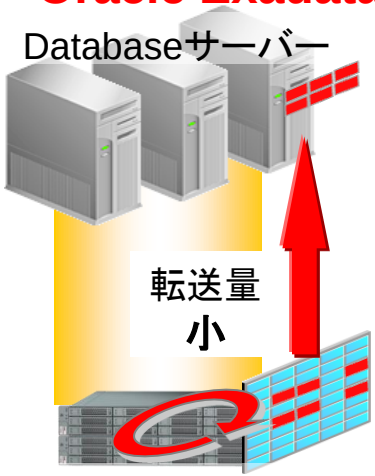
②データ選別

データの選別をデータベース側で行う為、ネットワークを通るデータ量は多い

①データ抽出

アクセス対象のデータ全てを抽出

Oracle Exadata



③計算・集計

計算・集計に使用されるデータのみを選別して返却する為ネットワークを通るデータ量は常に最小限

②データ選別

①データ抽出

Storageサーバーが必要な行・列データのみを選別・抽出

Exadata Hybrid Columnar Compression

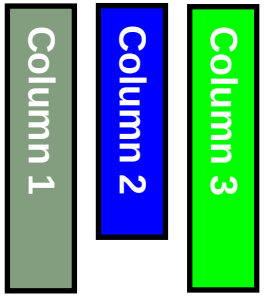
for DWH

- 列単位でデータを圧縮し格納することで高い圧縮率を実現
- 物理I/O量が減少することによる読み取り高速化

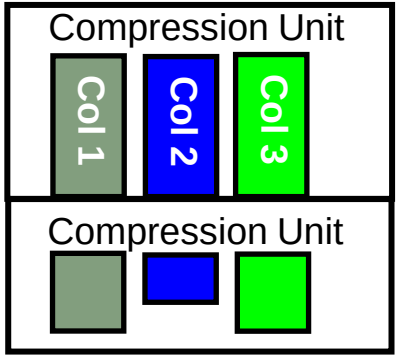
行指向の格納方式

※BASIC/OLTP圧縮時は、
重複データが圧縮される

列指向の格納方式



行指向 + 列指向の格納方式



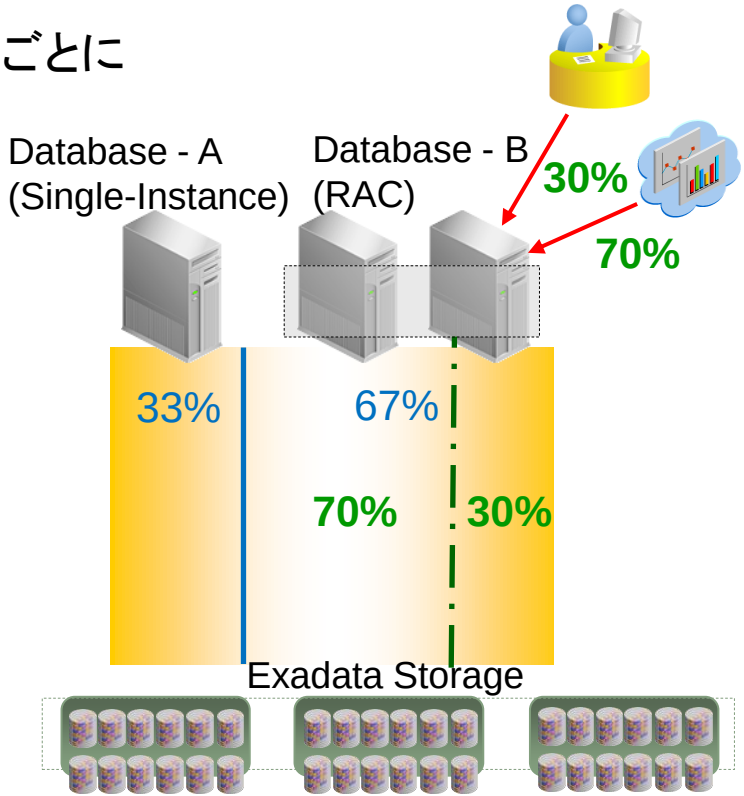
ランダムアクセス	◎		×	○
全表検索	○		◎	◎
Oracleの実装	非圧縮時	BASIC・OLTP圧縮時	なし	EHCC時
圧縮率	なし	中（3 ～ 5倍）	高	高（10 ～50倍）

I/O Resource Manager(IORM)

for Consolidation

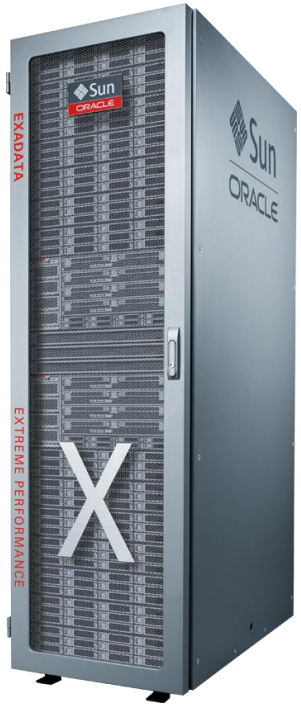
I/Oリソースを論理的に制御

- ストレージサーバーで、ユーザーごとやデータベースごとに使用するディスクのI/Oリソースを論理的に制御
- IORMの使用例
  - DBごとのリソース配分
    - Database A : 33%のI/Oリソース
    - Database B : 67%のI/Oリソース
  - DB内のリソース配分
    - インタラクティブ処理：30%のI/Oリソース
    - バッチ処理：70%のI/Oリソース
- 使用するI/Oリソースの上限値も設定可能
- スループット重視かレイテンシー重視かを選択可能
  - OLTP処理のI/OをDWH処理のI/Oから保護



Oracle Exadata Database Machine

Oracle Databaseに最適化されたEngineered System



Hardware(H/W)

- **DBサーバー**
  - マルチコア
- **InfiniBand**
  - 広帯域、低遅延ネットワーク
- **Storageサーバー**
  - Flashカード
  - 大量ディスク

×

Software(S/W)

- Oracle Database
  - 長年培ってきたデータベース技術
  - Gridアーキテクチャー
- Exadata Storage Server Software
  - I/Oのボトルネックを排除
  - Flash機能
  - リソース制御機能
  - 高圧縮機能

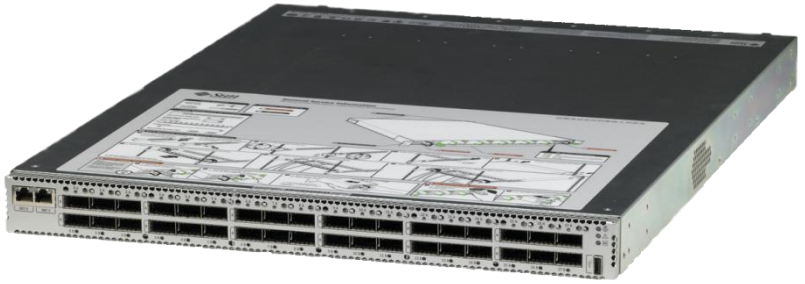
- Best for **DWH** / **OLTP** / **Consolidation**

Best for OLTP (InfiniBand & Flash)

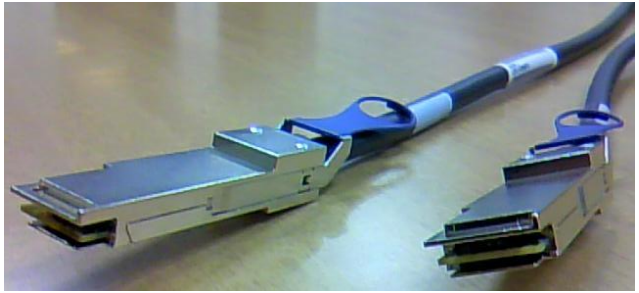
InfiniBandテクノロジー

InfiniBandネットワーク

広帯域、低遅延のネットワーク



- Multilane
  - 複数レーンを束ねて、**広帯域を実現**
  - Exadataでは、4 × QDR(40 Gbit/s) を採用
- ZDP RDS v3 プロトコル
  - **低遅延を実現 、CPUオーバーヘッドも大幅に軽減**
    - トランスポート層/ネットワーク層の処理を HCAのオフロード（Transport Offload）
    - RDMAを使用し、余計なデータ・コピーは作らず、 Storageサーバー上のデータをデータベース・バッファに 直接転送
  - 3 GB/secの転送に2%程度のCPU使用率
- カットスルー方式の採用（スイッチ）



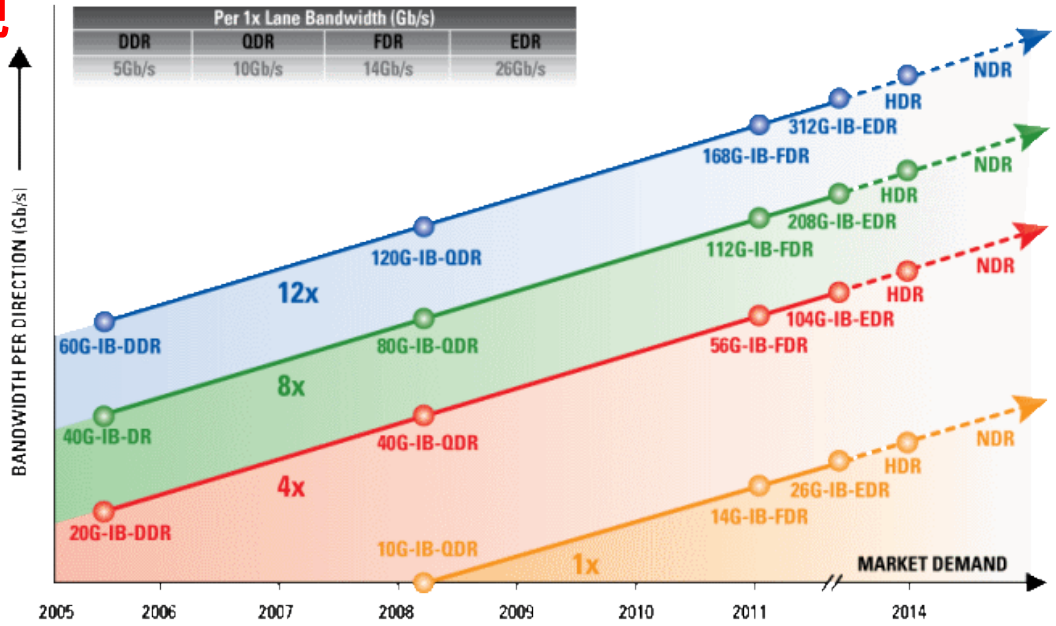
Exadataでの使用
- ストレージネットワーク
- RACのインターコネクト
- 外部通信(optional)

広帯域

何故広い!? InfiniBand

Multilaneによる広帯域の実現

- 1レーン
  - 一対の信号線のこと
- 転送レート
  - SDR 2.5Gbps / 1レーン
  - DDR 5Gbps / 1レーン
  - QDR 10Gbs / 1レーン
  - FDR 14Gbps / 1レーン
- Multilane
  - 複数レーンを束ねる通信規格
    - 1レーン(1×)、4レーン(4×)、8レーン(8×)、12レーン(12×)
  - 束ねたことによる転送速度の低下なし



InfiniBand Trade AssociatoinのWEBページより
http://www.infinibandta.org/content/pages.php?pg=technology_overview

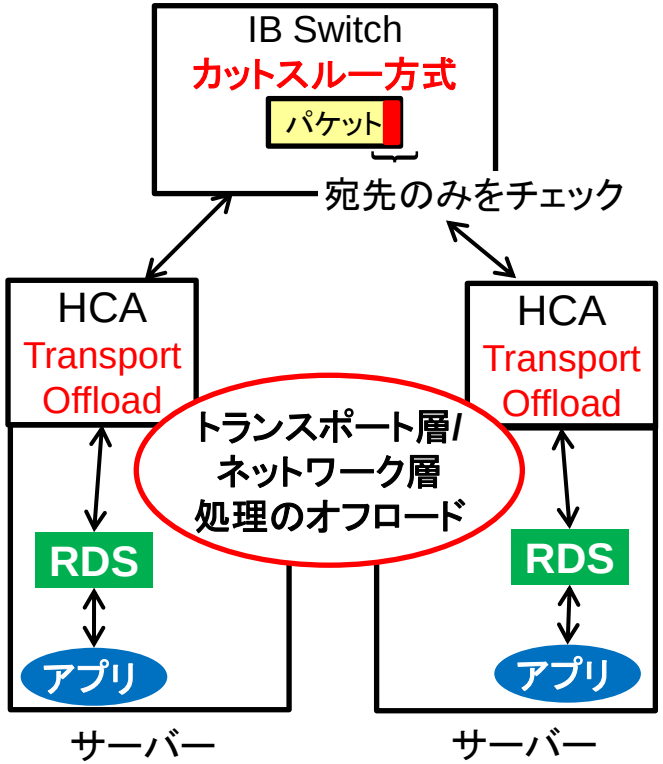
低遅延

何故速い!? InfiniBand①

カットスルー方式

- スイッチでの遅延を軽減
  - InfiniBand Switchは、カットスルー方式を採用

一般的なEthernet Switch	InfiniBand Switch
ストア&フォワード方式	**カットスルー方式**
• パケットを蓄積し、すべてのパケットをチェック後転送	• パケットの宛先のみをチェックし転送
• エラーを見つけたらパケットを破棄	• エラーチェックはエンドノードのみで実施
• 異なる速度のLANに対応	• 低遅延での転送が可能

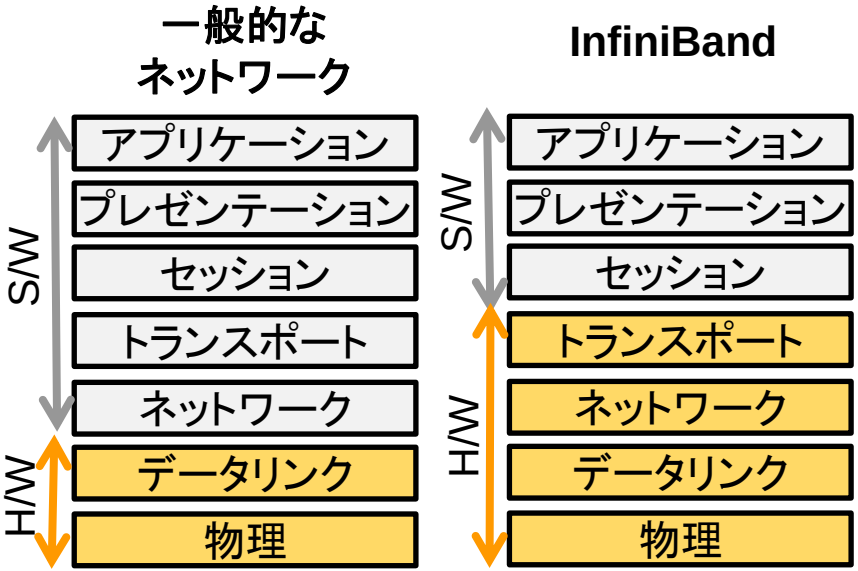


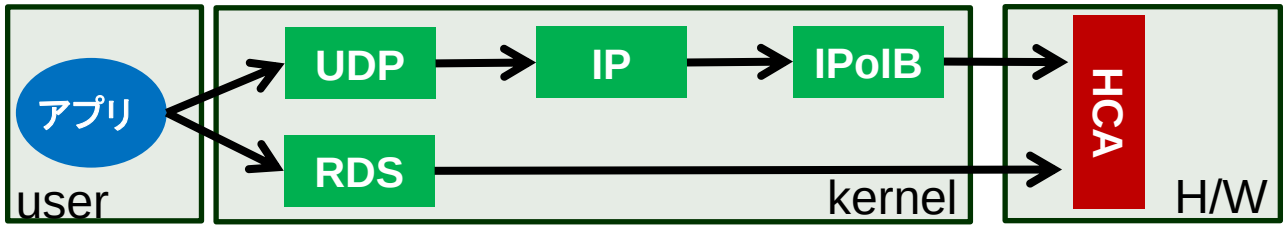
何故速い!? InfiniBand②

低遅延

Transport Offload/RDSプロトコル

- Transport Offload
  - ネットワーク層/トランスポート層を HCAのチップがハードウェアで処理
    - エラーチェックもHCAにて実施
- Reliable Datagram Socketsプロトコル
  - Transport Offload機能を活かすプロトコル
    - RDSプロトコルではエラーチェックを行わない
  - アプリケーションからは、UDPとして使用可能



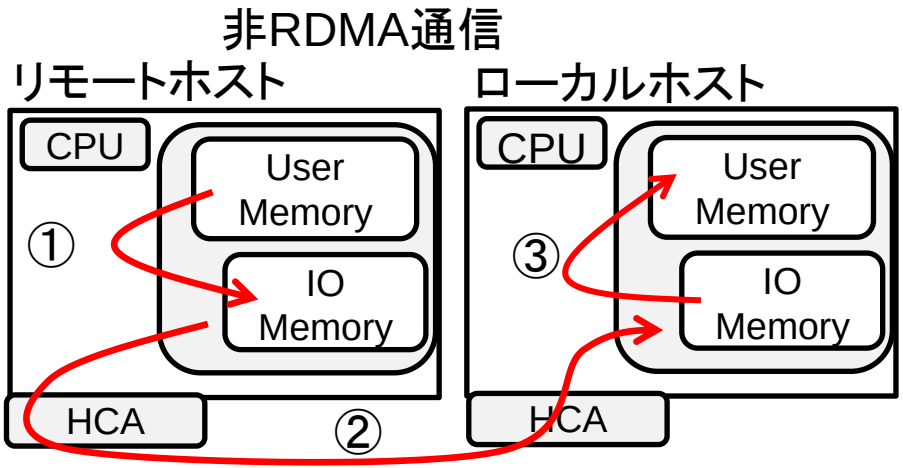


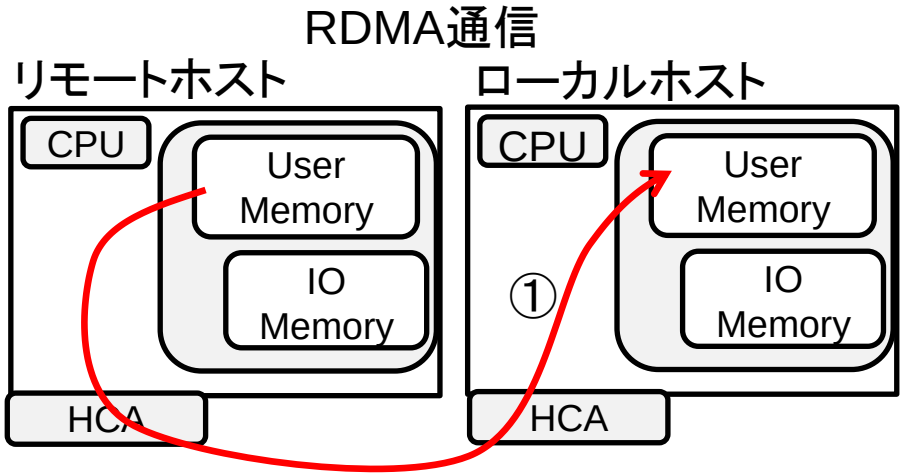
低遅延

何故速い!? InfiniBand③

RDMA通信

- Remote Direct Memory Access
  - リモートホストのメモリ間で直接データを転送できる技術
- CPUオーバーヘッドの削減 / メモリ使用量の削減 / レイテンシーの削減



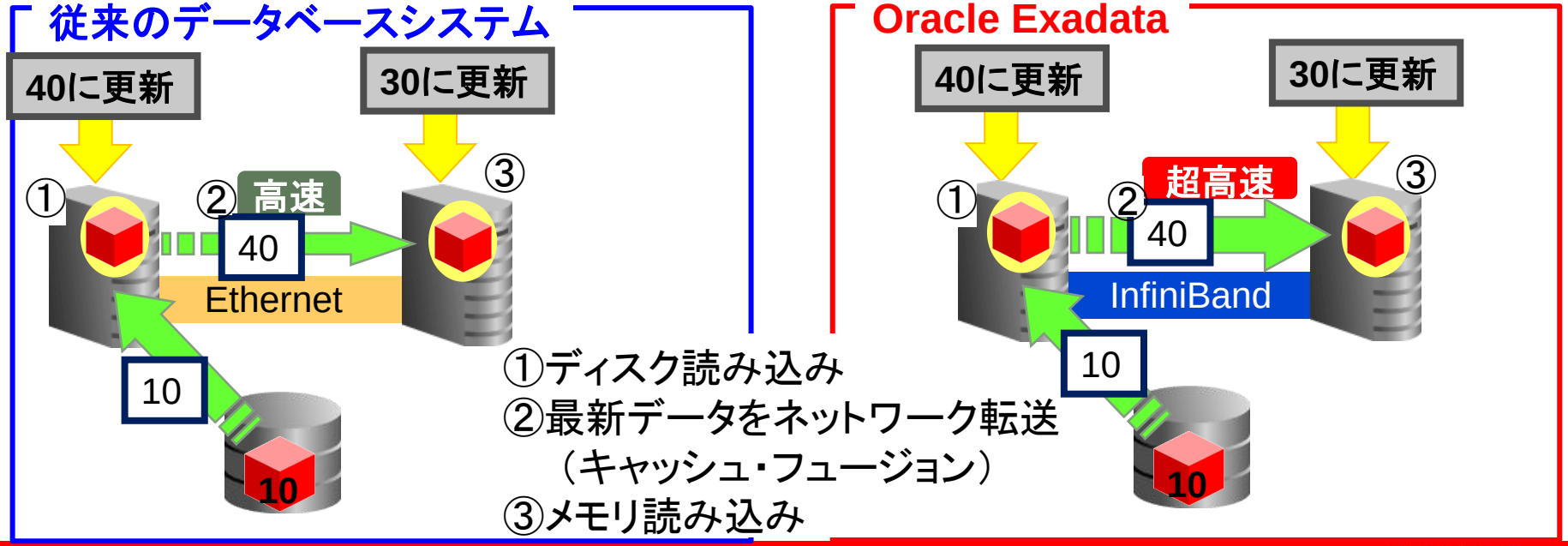


InfiniBand × RAC

低遅延

低遅延通信によるRACのリニアなスケーラビリティの実現

ノード間の通信が超高速になるため、
キャッシュフュージョン多発時もスムーズなブロック転送



スクリーンオンリー

InfiniBand × RAC

1GbE(UDP) vs InfiniBand(RDS) 検証結果

Flashテクノロジー

ExadataのFlashテクノロジー

各StorageサーバーにFlashカードが内蔵

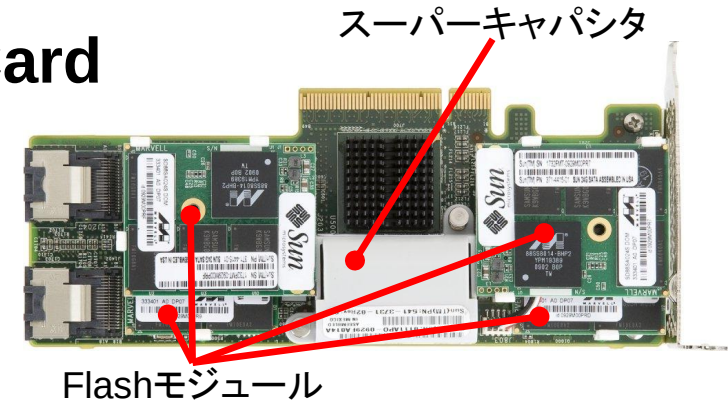
- Sun Flash Accelerator F20 PCIe Card × 4枚
  - ストレージ容量: 384GB(1台)
  - IOPS: 125,000 IOPS(1台)
  - スループット: 5.4 GB/s(1台 ※非圧縮時)
- Flash PCIeカードを採用
  - ディスクコントローラーのボトルネックを解消

Flash × Exadata Storage Server Software

- Smart Flash Cache(ランダム読み取り高速化)
- Smart Flash Log(Redo書き込み高速化)

Sun Flash Accelerator F20 PCIe Card

Enterprise Flash



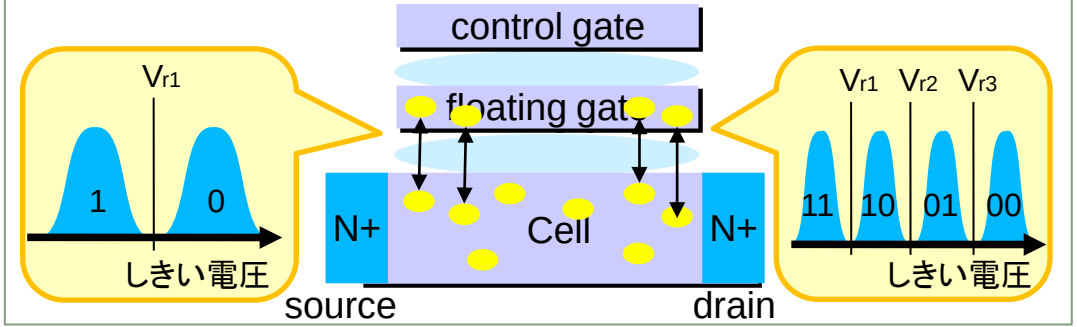
- バックアップ電源としてスーパーキャパシタを搭載

4つのFlashモジュールを搭載

- Flashモジュールに実装されるNANDにはSLCを採用
  - 4KBの読込/書込単位（ページサイズ）
  - 256KBの消去単位（ブロックサイズ）
- 32GBに対して8GBの予備領域
  - ウェアレベリング
  - 無効ブロック管理

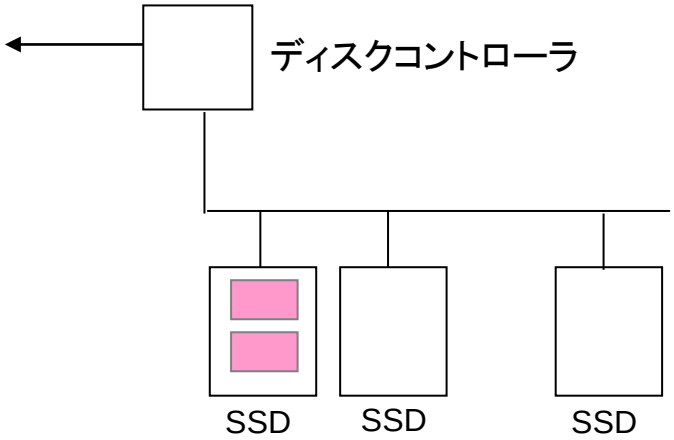
SLC(Single Level Cell)
- 1ビット/cell
- 書込回数：約100,000

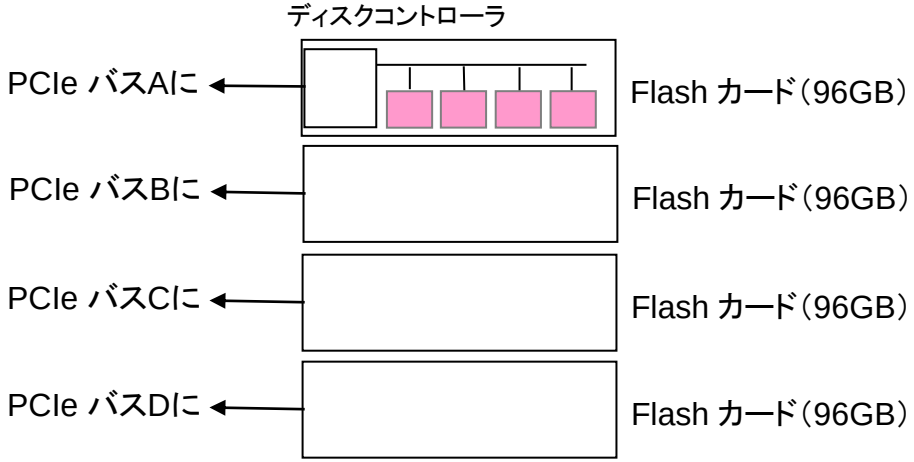
MLC(Multi Level Cell)
- 複数ビット/cell
- 書込回数：約10,000



FlashカードとSSDの違い

ExadataではSSDではなくFlashカードを搭載





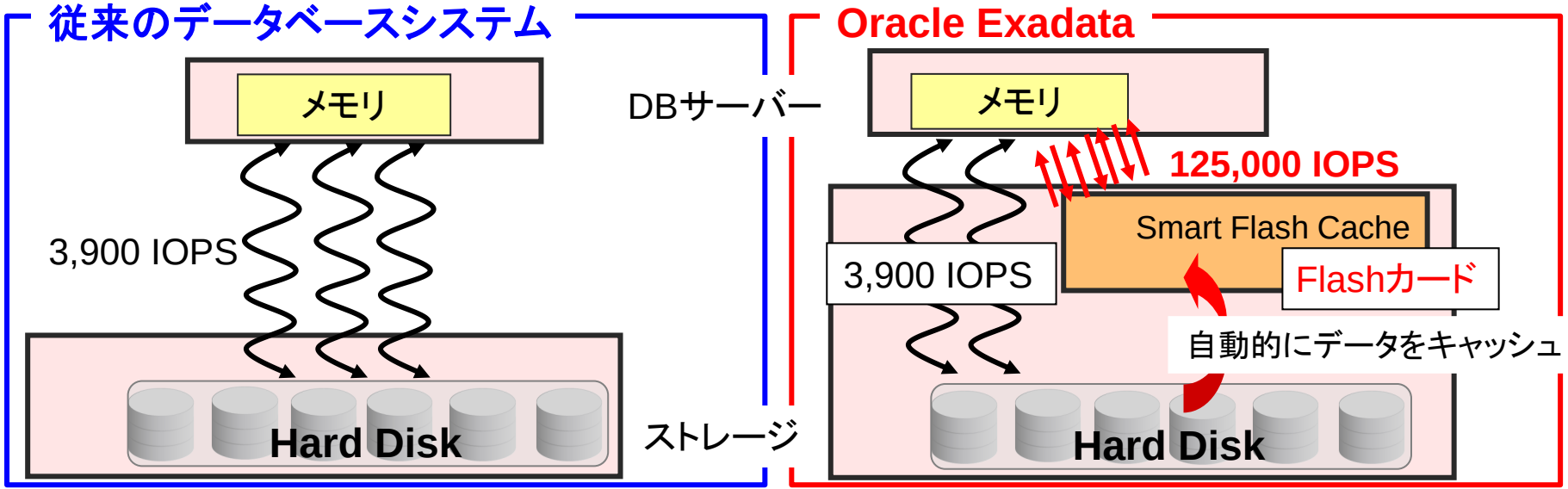
- NAND型フラッシュメモリの構造は同じ、サーバに搭載される際のインターフェースに違いがある
- FlashではPCIeスロットに搭載することにより、I/Oスループットもスケールし、広帯域が確保できる
- SSDで搭載する場合には、ディスク用のスロットが必要になってしまうが、カード型の場合ディスク容量を追求しながら集積性も追及できる

Exadata Smart Flash Cache

ランダムRead I/Oのボトルネックを解消

	メモリ	フラッシュ	ハードディスク
アクセス	◎	○	×
価格	×	○	◎

- 利用頻度の高いデータをFlashに自動でキャッシュ
- Flashとディスクの併用も可能（Small I/O => Flash、Large I/O => ディスク）



Exadata Smart Flash Cache

- Oracle Databaseに最適化されたキャッシング
  - DatabaseのI/Oの種類を自動で理解してキャッシュ

キャッシュしないオブジェクト

- ASMのミラーコピー
- バックアップ、DataPump
- 表領域のフォーマット
- テーブルフルスキャン

キャッシュするオブジェクト

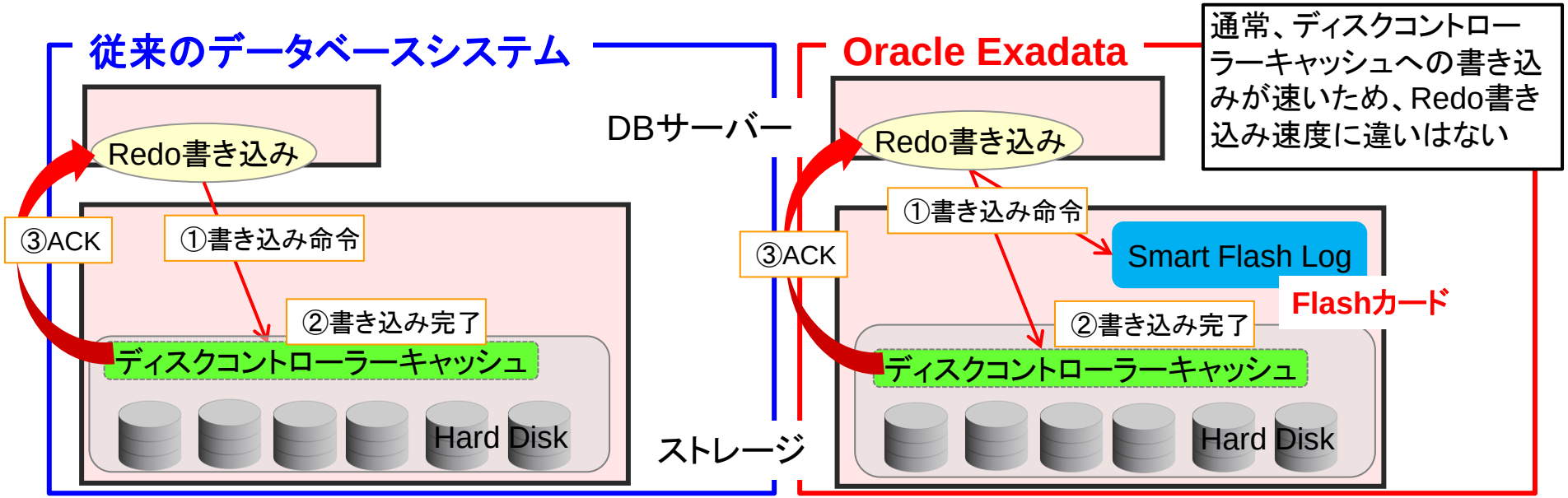
- 制御ファイルのRead/Write
- ファイルヘッダーのRead/Write
- Data BlockとIndex Block

- 管理者は特定のオブジェクトをFlash Cacheにキープさせるように指定も可能
- 完全に自動化され、透過的に実行される

Exadata Smart Flash Log

Redo書き込みを高速化（通常時）

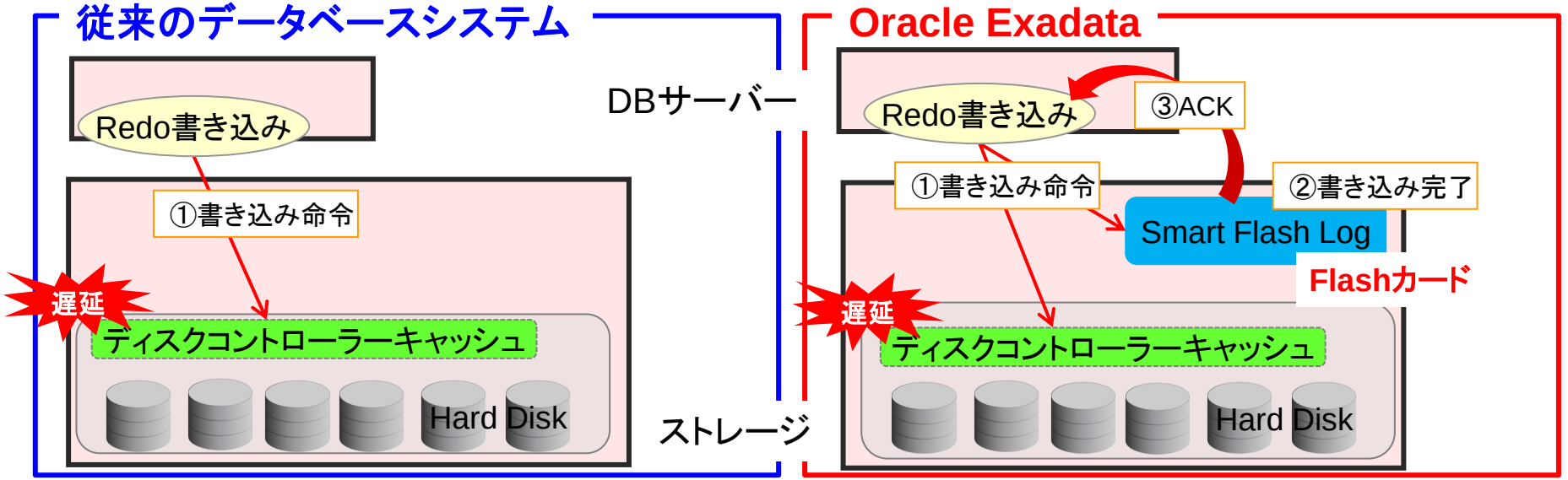
- ディスクコントローラーキャッシュとFlashメモリーの両方同時にRedo書き込みを行い、どちらかが完了したら、書き込みを完了する



Exadata Smart Flash Log

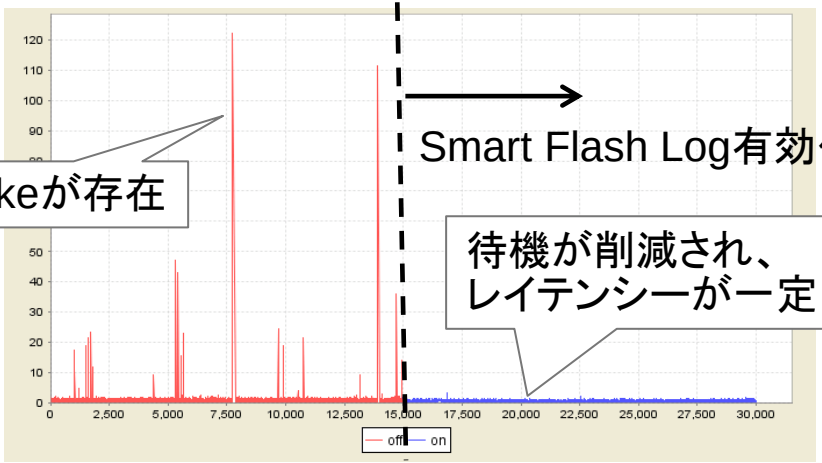
Redo書き込みを高速化（遅延発生時）

- ディスクコントローラーキャッシュへの書き込みが遅い場合、Flashメモリーへの書き込みが終了後、Redo書き込みが終了する



Exadata Smart Flash Log

- レスポンスタイムの向上と、待機の異常値を削減し、データベース全体のスループットの向上を実現
  - OLTP処理のOracle Databaseの同期書き込みは、Redo書き込みのみ
  - “log file parallel write”や”log file sync“の改善
- 完全に自動化され、透過的に実行される



スクリーンオンリー

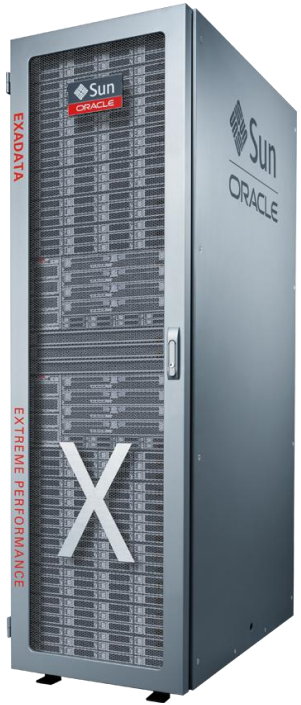
Exadata Smart Flash Log

社内検証結果

まとめ

Oracle Exadata Database Machine

Oracle Databaseに最適化されたEngineered System



- **Best for DWH**
  - マルチコア/大量ディスクを活かすGridアーキテクチャー
  - システムのボトルネックになりやすい「I/O」を効率化
- **Best for OLTP**
  - InfiniBand × RAC
    - 低遅延通信によりRACのリニアなスケーラビリティを実現
  - Flashデバイス × Exadata Storage Server Software
    - Smart Flash Cache（ランダム読み取りを高速化）
    - Smart Flash Log（Redo書き込みを高速化）
- **Best for Consolidation**
  - ワークロードのリソースを動的かつ容易に制御可能

Q&A

ご質問・ご相談はOpenWorld終了後もお受けしております

あなたにいちばん近いオラクル

Oracle Direct

0120-155-096

（平日9:00-12:00 / 13:00-18:00）

http://www.oracle.com/jp/direct/index.html

各種無償支援サービスもございます。

Hardware and Software

ORACLE

Engineered to Work Together

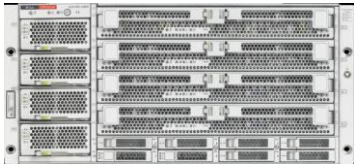

ORACLE®

Appendix

X2-2 Database Server (Sun Fire X4170M2)

プロセッサ	2つの 6Core Intel® Xeon® X5675 プロセッサ(3.06 GHz)
メモリ	96GB (12 x 8GB) (144GB（18X8GB）まで拡張可能)
内蔵ディスク	ホットスワップ可能 4 x 300GB 10K RPM SAS ディスク
ディスク・コントローラ	Disk Controller HBA with 512MB Battery Backed Cache
ネットワーク	2 つの InfiniBand 4X QDR (40Gb/s) ポート (1 枚のDual-port PCIe 2.0 HCA) 4 つの 1GbE Ethernet ポート 2 つの 10GbE Ethernet SFP+ ポート
リモート管理	1 Ethernet port (ILOM用)
電源	冗長化された、ホットスペア対応電源（2基）およびファンモジュール（4基）

X2-8 Database Server (Sun Fire X4800M2)



プロセッサ	8 つの 10Core Intel® Xeon® E7-8870 プロセッサ (2.40 GHz)
メモリ	2 TB (128 x16 GB)
内蔵ディスク	ホットスワップ可能 8 x 300GB 10K RPM SAS ディスク
ディスク・コントローラ	Disk Controller HBA with 512MB Battery Backed Cache
ネットワーク	8 つの InfiniBand 4X QDR (40Gb/s) ポート (4 枚のDual-port PCIe エクスプレス・モジュールが搭載) 2つの Network Express Modules (NEM)に下記が搭載 • 8 つの 1GbE Ethernet ポート • 8 つの10 GbE Ethernet SFP+ ポート
リモート管理	1 Ethernet port (ILOM用)
電源	冗長化された、ホットスペア対応電源（4基）およびファンモジュール(4基)

Exadata Storage Server X2-2 (Sun Fire X4270 M2)

プロセッサ	2 つの6Core Intel® Xeon® L5640 プロセッサ (2.26 GHz)
メモリ	24 GB (6 x 4GB)
ディスク	12 x 600 GB 15K RPM High Performance SAS ディスク もしくは 12 x 3 TB 7.2K RPM High Capacity SAS ディスク
Flash	4 枚の 96 GB Sun Flash Accelerator F20 PCIe カード
ディスク・コントローラ	Disk Controller HBA with 512MB Battery Backed Cache
ネットワーク	2つの InfiniBand 4X QDR (40Gb/s) ポート (1 枚のDual-port PCIe 2.0 HCA) 4つのオンボード Gigabit Ethernet ポート
リモート・マネジメント	1 Ethernet port (ILOM用)
電源	冗長化された、ホットスペア対応電源(2基)およびファンモジュール（6基）


ORACLE®