ORACLE

Exadata Database Machine : Maximum Availability Architecture (MAA)

技術に関するプレゼンテーション

ExadataおよびMAA製品管理

2025年1月

アジェンダ

1 Maximum Availabilityを重視する理由は?

2 Maximum Availability Architectureとは?

3 ExadataにおけるMaximum Availability Architectureの機能

4 Exadataのライフサイクル操作

5 まとめ

Maximum Availabilityを重視する理由は?

35万ドル

1時間あたりの停止時間の平均コスト

出典：Gartner、Data Center Knowledge、IT Process Institute、Forrester Research

1,000万ドル

データセンターの計画外停止や災害の平均コスト

出典：Gartner、Data Center Knowledge、IT Process Institute、Forrester Research

87時間

年間平均停止時間

出典：Gartner、Data Center Knowledge、IT Process Institute、Forrester Research

91%

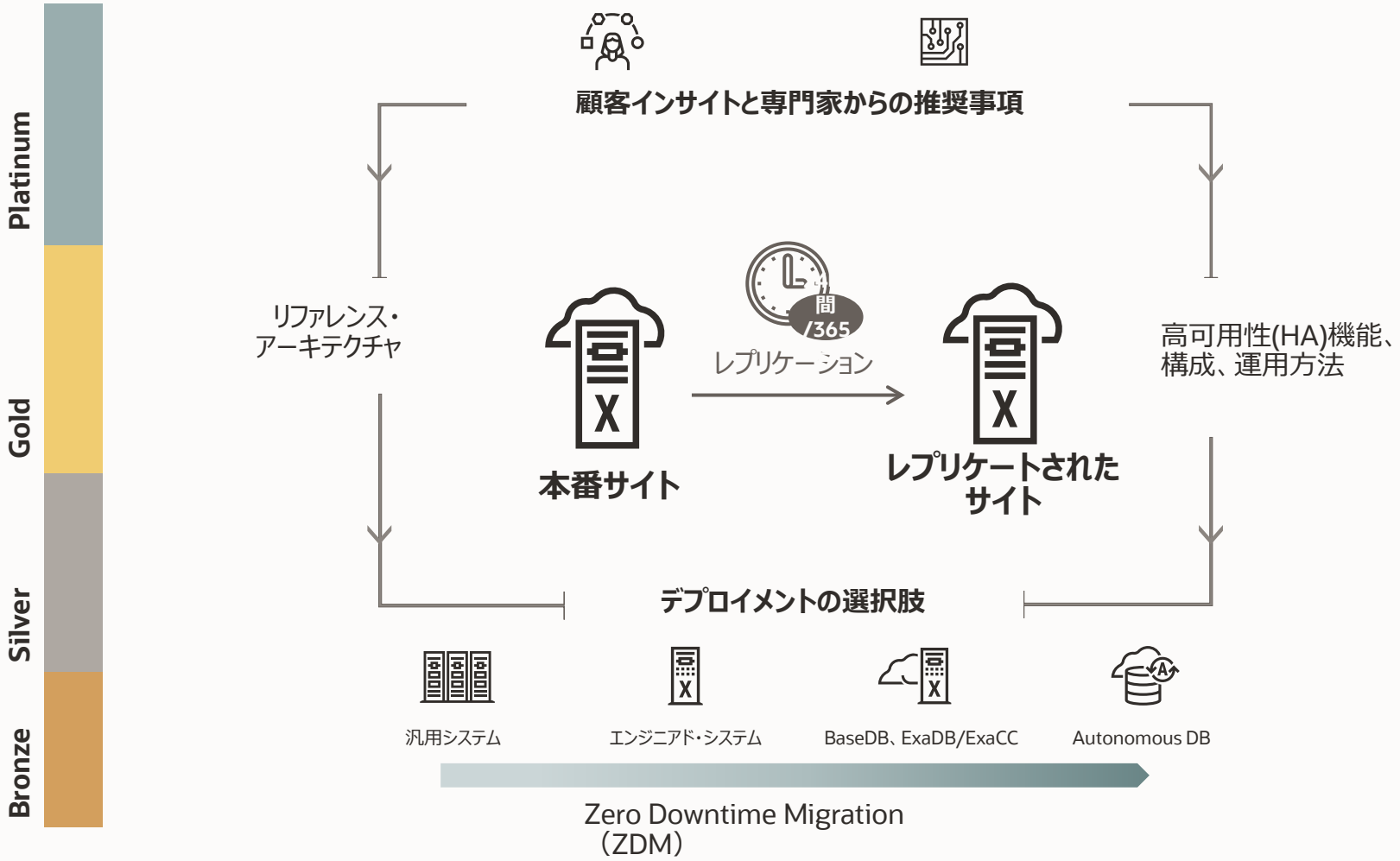
過去24か月以内にデータセンターの計画外停止を経験した企業の割合

出典：Gartner、Data Center Knowledge、IT Process Institute、Forrester Research

Maximum Availability Architectureとは?

Oracle Maximum Availability Architecture（Oracle MAA）

停止が許されないデプロイメント向けに標準化されたリファレンス・アーキテクチャ



継続的な可用性

アプリケーション・コンティニュイティ / オンライン再定義 / エディションベースの再定義

データ保護

フラッシュバック / RMAN / ZDLRA+ ZRCV

アクティブ・レプリケーション

Active Data Guard / Full Stack DR / GoldenGate

スケール・アウトおよびライフサイクル

RAC / FPP / Globally Distributed Database

MAAリファレンス・アーキテクチャ
可用性サービス・レベル

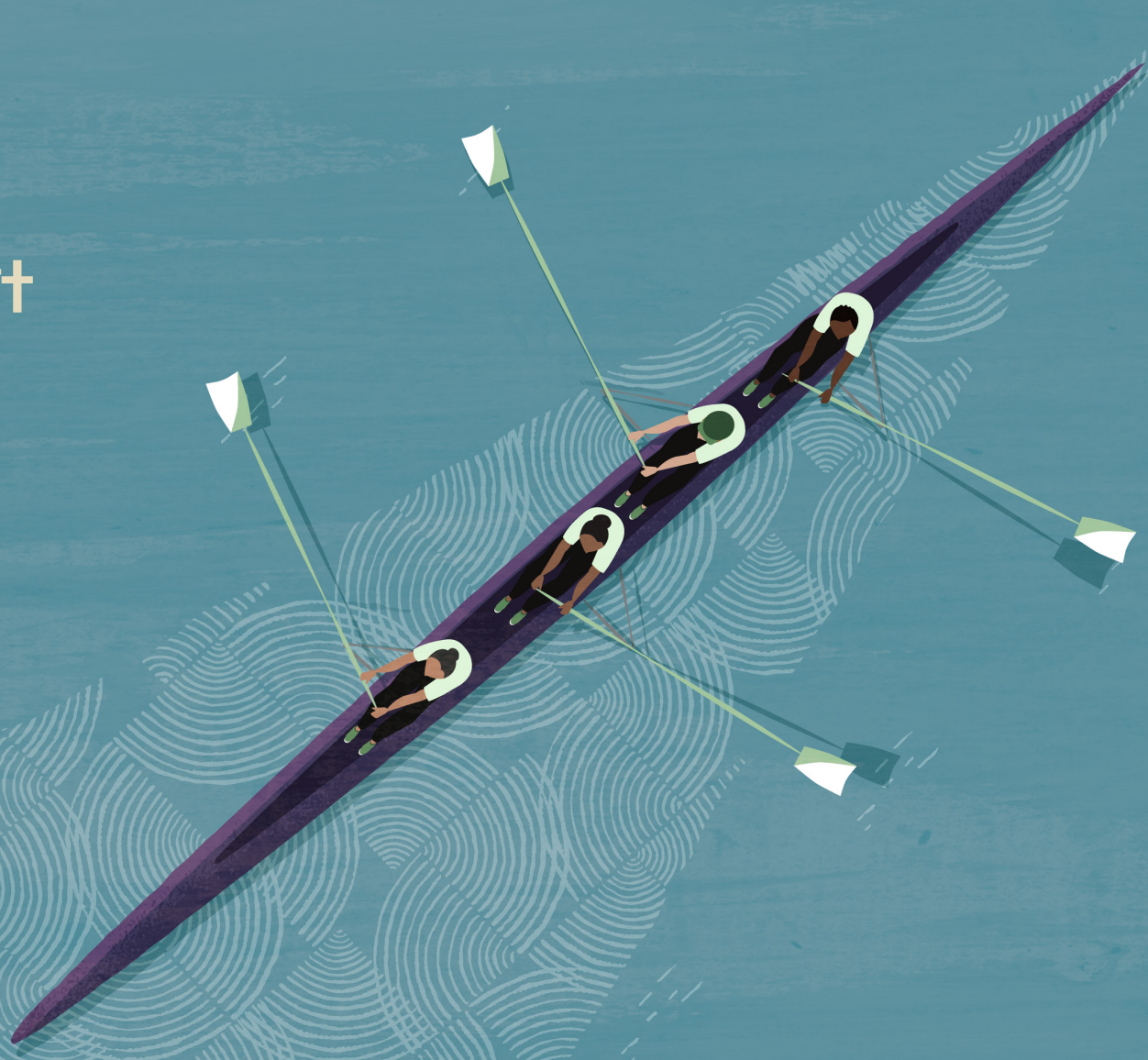
Bronze	Silver	Gold	Platinum
開発、テスト、本番	**本番/部門**	**ビジネス・クリティカル**	**ミッション・クリティカル**
シングル・インスタンスDB 再起動可能 バックアップ/リストア	**Bronze＋** Oracle RACによるデータベースのHA アプリケーション・コンティニュイティ	**Silver＋** Oracle Active Data Guardを使用したDBレプリケーション	**Gold＋** GoldenGate エディションベースの再定義

すべての層をオンプレミスとクラウドで利用可能。

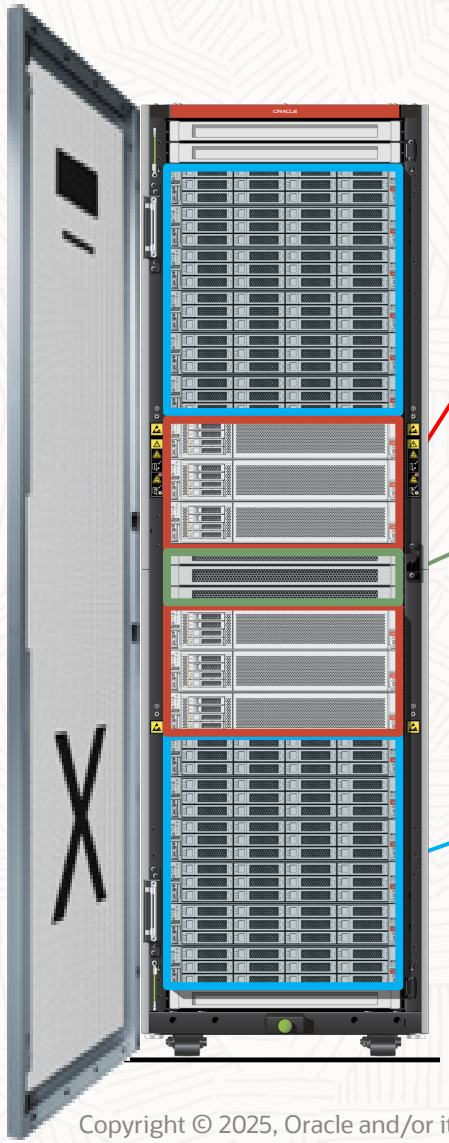
Oracle MAAでのExadataの機能

ハードウェアとソフトウェアをともに設計

- ✓ パフォーマンス
- ✓ 管理性
- ✓ 可用性



Oracle Exadata Database Machine：組み込み済みの高可用性



冗長化されたデータベース・サーバー

クラスタ化された高可用性のアクティブ-アクティブ・サーバー
ホットスワップ対応の電源、ファン、フラッシュ・カード
冗長配電ユニット
統合されたHAソフトウェア/ファームウェア・スタック

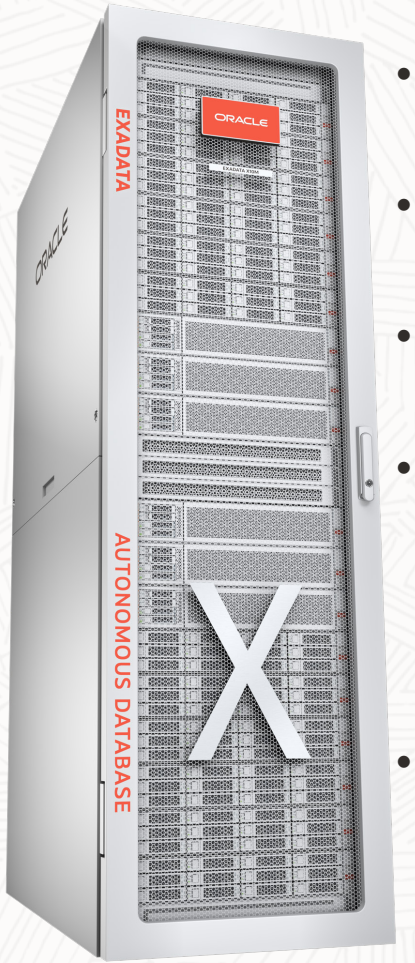
冗長化されたネットワーク

100 Gb/秒の冗長化されたRoCEネットワークとスイッチ
HAボンディング・ネットワークを使用したクライアント・アクセス
統合されたHAソフトウェア/ファームウェア・スタック

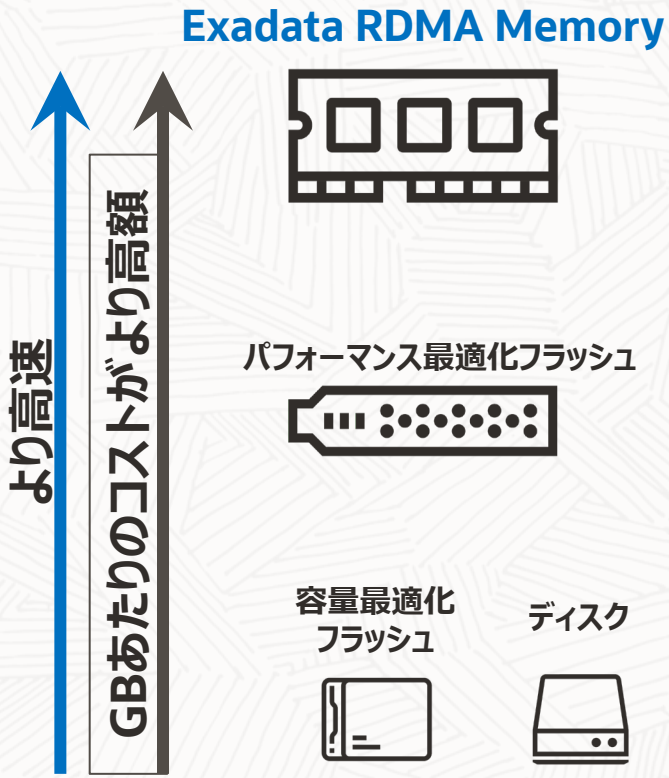
冗長化されたストレージ・グリッド

複数のストレージ・サーバーにまたがってミラー化されたデータ
ホットスワップ対応の電源、ファン、M.2ドライブ、フラッシュ・カード
冗長なノンブロッキングI/Oパス
統合されたHAソフトウェア/ファームウェア・スタック

Exadata：X11M



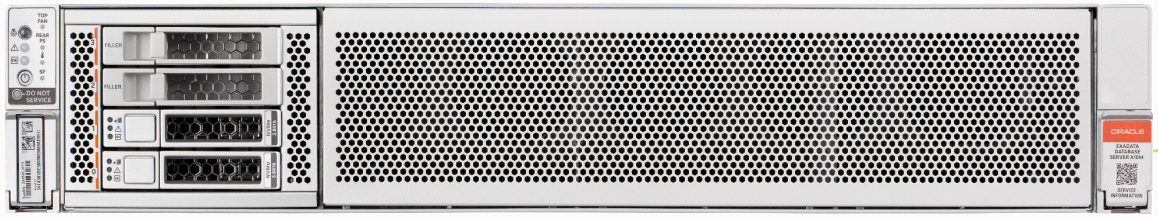
- 100 Gbアクティブ-アクティブRDMA over Converged Ethernet（RoCE）プライベート・ネットワーク
- 待機時間の短い1.25 TBのExadata RDMA Memory（XRMEM）をストレージ・サーバーごとに搭載
- Data Acceleration機能により、読取り待機時間が14 µs未満に短縮
- 3つのストレージ層：
  - Exadata RDMA Memory
  - パフォーマンス最適化フラッシュ
  - 容量最適化フラッシュまたはハード・ディスク
- ベアメタル構成、またはKVMベースの仮想化構成



Exadata：MAAに最適なプラットフォーム

進化：HAに関するもっとも困難な問題からお客様のサービス・レベルを継続して保護

X11 データベース・サーバー



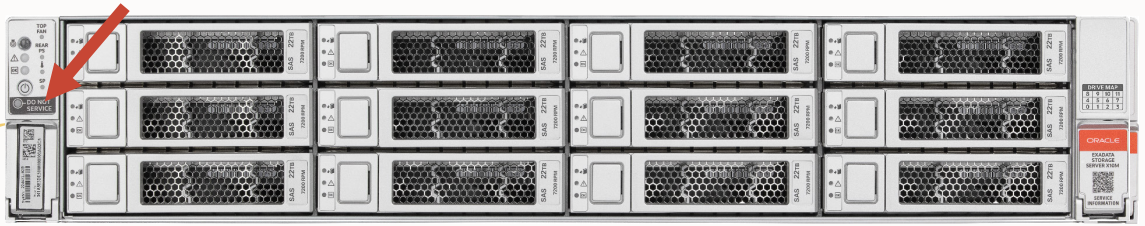
Linuxのメジャー・バージョン・アップグレードへの影響ゼロ
（例: Exadata System Software 23.1 OL8）

セキュリティ・ソフトウェア・アップグレードへの影響ゼロ
（STIGコンプライアンス対応を含む）

データベースとGrid Infrastructureソフトウェアの
インシデントを通知するMS （管理サーバー）

人為的エラーの
防止!

X11 ストレージ・サーバー



計画外停止、計画停止の間の短いI/Oレイテンシーの保持

不具合のあるストレージの自動修復と
緊密に統合されたハードウェアおよびソフトウェア

パフォーマンスと容量が最適化されたフラッシュを搭載した新しい
Exadata X11 Extreme Flashストレージ・サーバー

ストレージサーバー側のXRMEMキャッシュから14マイクロ秒でデータベースI/Oを取得

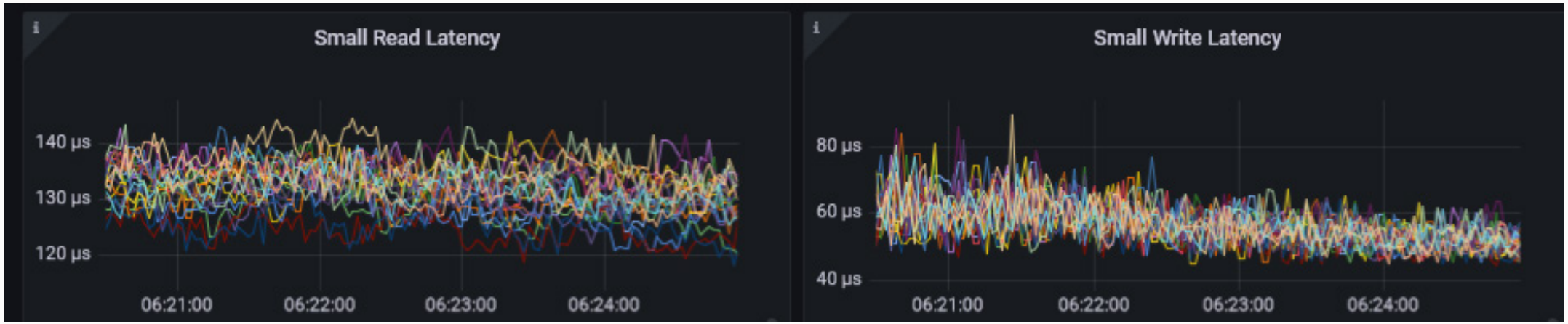
Exachkによる MAAベスト・プラクティスのフル・スタック・コンプライアンス・チェックを実施

Exadata：MAAに最適なプラットフォーム

進化：メトリックを簡略化

*Exadata*の内部で起きていることをどうやって知ることができるでしょうか?

- Exadataの開発以来、Exadataメトリック内にパフォーマンス・データが存在しているが、メトリックを利用して理解することは難しい場合があった
- Exadataリリース22.1で*Real-Time Insight* 機能が実装。 ダッシュボードの1つを拡大するだけで、パフォーマンスの傾向を監視したり、パフォーマンス異常の際に明るく光らせたりすることが可能に



Exadata：組み込み済みの高可用性

ディスク取り外しのための自動LEDサポート

電源停止時の冗長性チェック

読取りおよび書込みにおけるI/O待機時間制限

patchmgrによるデータベース・ノードの更新

コントローラ・キャッシュ障害からの自動修復

ハード・ディスクのドロップと交換

セル・シャットダウン時の冗長性保護

最速のREDO Applyとインスタンス・リカバリ

I/Oハングの検出および修復

EM障害レポート

アクティブ/アクティブRoCEネットワーク

セル間のリバランスによるデータ・アクセラレータ・キャッシュの保持

自動ディスク管理

I/Oエラー破損時における自動ASMミラー読取り

Exadataディスク・スクラビングおよびASM破損修復によるI/Oエラー回避

データベース・サーバーに対する障害モニタリング

ILOMハングの検出および修復

最適化、高速化されたExadataパッチ適用

Exadata HARD

セル間のリバランスによるフラッシュ・キャッシュの保持

ディスク修復時におけるセルからセルへのオフロード

ネットワーク障害の高速な検出

Exadataスマート・ライトバック

I/OおよびネットワークのリソースExachkによるフル・スタック・ヘルス・チェック（重大な問題のアラートあり）

BBUのドロップと交換

VLANのサポートおよび自動化

ディスクの拘束

HARDのサポートによる破損の回避

セルI/Oタイムアウトしきい値

インスタンス・リカバリ時の一時停止の短縮

セル・アラート用カスタム診断パッケージ

取り外しOKを知らせる青色LED通知

自動オンライン

優先順位リバランスのサポート

柔軟性のあるExadata構成

フラッシュとディスクのライフサイクル管理のアラート

フラッシュ障害後の効率的なリシルバ・リバランス

障害と断定されるディスクのヘルス要素

ドライブ障害誤検出の排除

Exadataスマート・フラッシュ・ロギング

cellsrvシャットダウン時の冗長性保護

スマート・ライトバック・フラッシュ・キャッシュの永続性

アプライアンス・モードのサポート

セル・アラートのサマリー

ライフサイクル管理

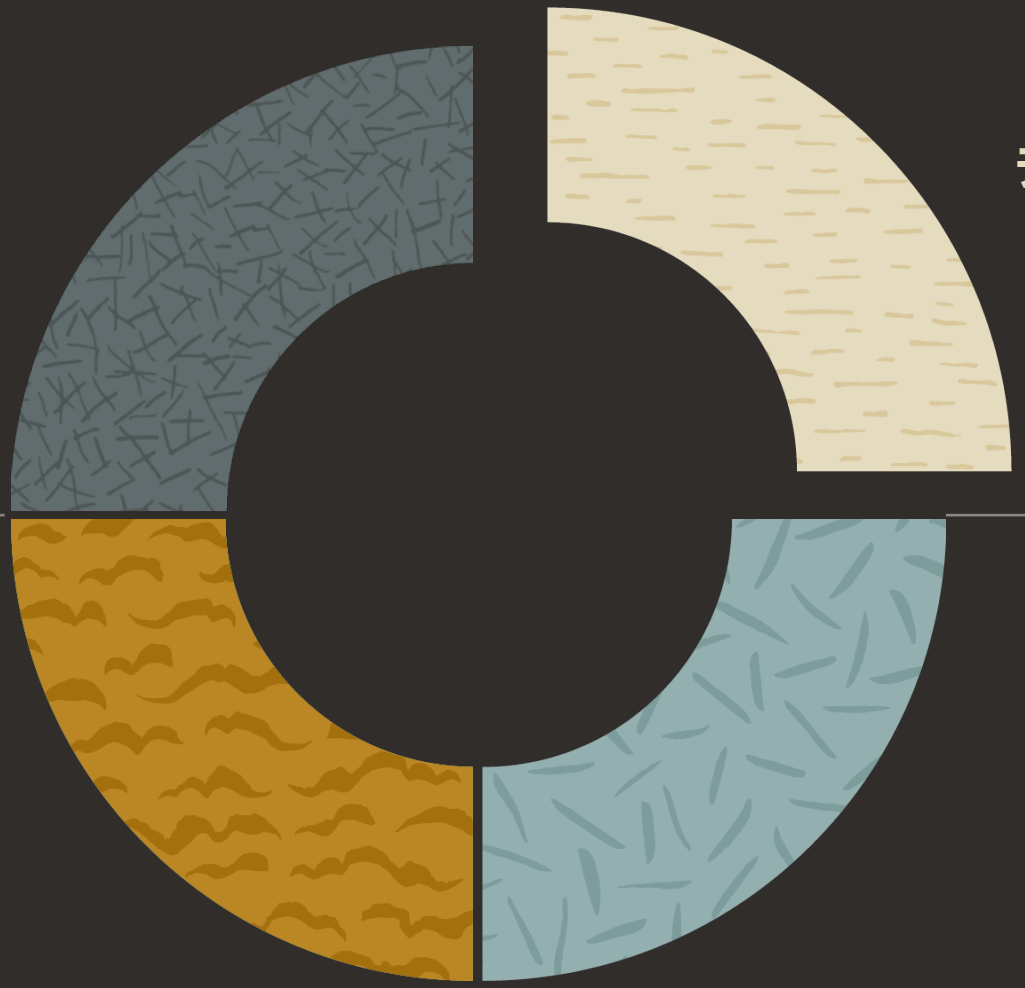
データ保護

一時停止

サービス品質および
パフォーマンス

ライフサイクル管理

データ保護



一時停止

サービス品質および
パフォーマンス

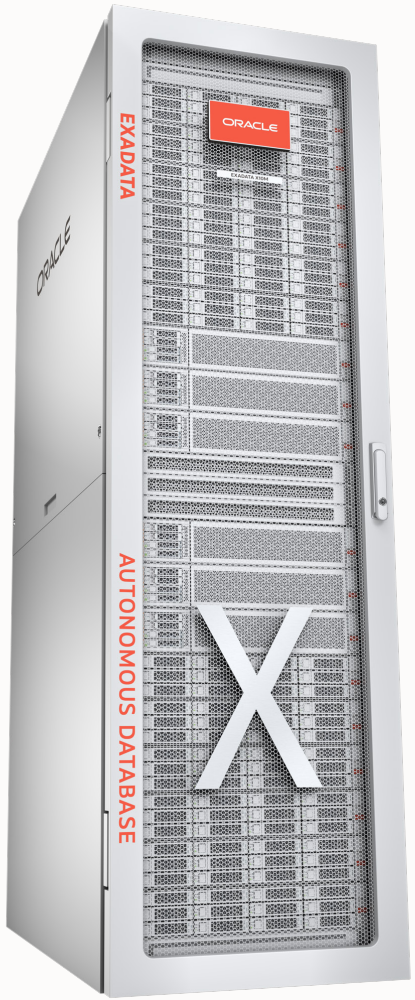
データ破損とは

Data corruption refers to errors in computer data that occur during writing, reading, storage, transmission, or processing, which introduce unintended changes to the original data. Computer, transmission, and storage systems use a number of measures to provide end-to-end data integrity, or lack of errors.

- 物理的破損（メディア破損）
  - データベース・ブロックのチェックサムがコンテンツと一致しない
    - ヘッダーの破損
    - ブロックに0（ゼロ）が含まれる
    - …
- 論理的破損
  - データベース・ブロックのチェックサムは正しいが、論理的に一致していない
    - ヘッダーの下部構造が破損
    - 書込みデータの消失
    - 存在しないトランザクションによって行ロックが発生
    - …
- 通常は気づかないうちに発生

出典：Wikipedia

Exadata：データ保護
破損の検出と回避

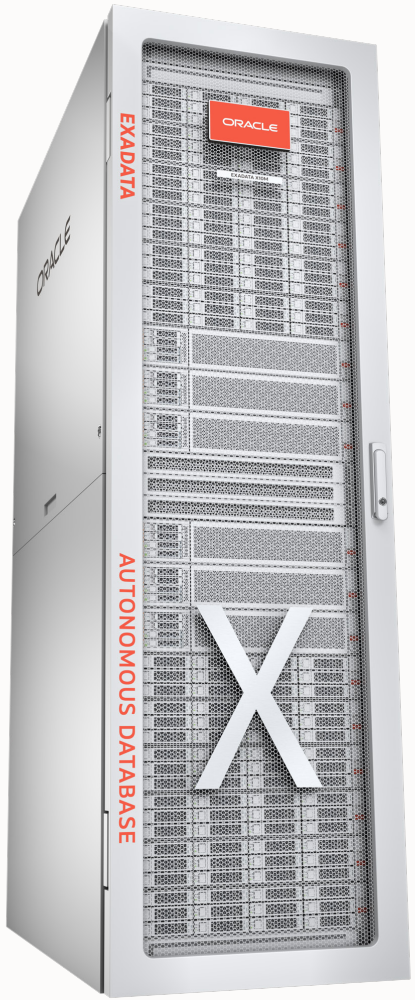


DBサーバーとストレージ・ノード間のI/Oパスの
ネットワーク・パケットが破損した場合

- ストレージ・セルが書込みを防止
- パケットの再送信によるASMの再試行

✓アプリケーションに破損が生じることはない

Exadata：データ保護
破損の修復



アプリケーションのデータベース更新で破損に遭遇した場合

- データベースはASMのミラーからデータブロックを読み取る
- 適切なコピーを使用して破損を修復

✓この修復は、他のデータベース・プロセスやアプリケーションに影響を与えずに行われる

Exadata：データ保護
ストレージ障害

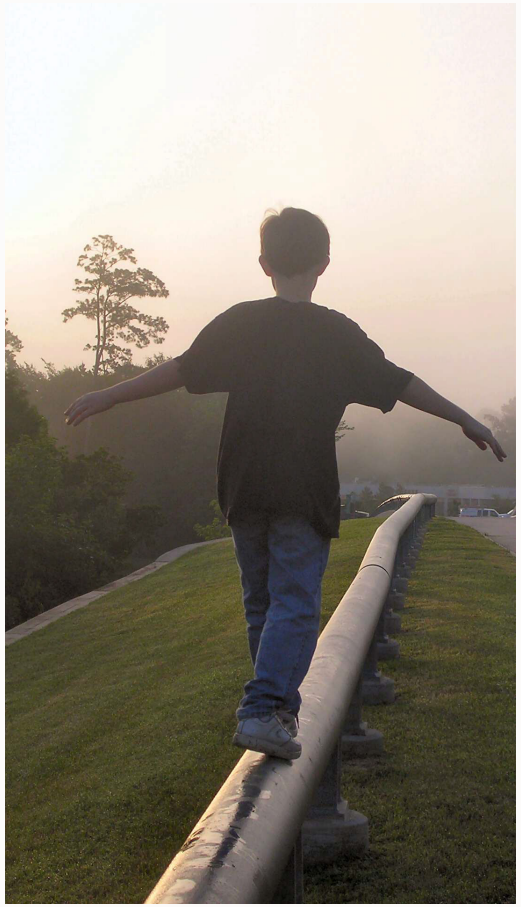
ストレージ・セルで実際に発生していない障害がドライブに報告された場合

- ドライブ障害の誤検出を回避するために、ドライブ/フラッシュの電源の自動再投入を実施

- ストレージ障害が発生すると、冗長性に影響が及ぶ
- データベースを意識した優先順位では、データ保持のために冗長性のリストアが優先される
- 優先順位
  1. 制御ファイル
  2. オンライン・ログ
  3. アーカイブ・ログ
  4. ASM SPFILE
  5. データベースSPFILE
  6. TDEキー・ストア
  7. OCR
  8. スタンバイREDOログ
  9. ウォレット
  10. データファイル

Exadata：データ保護

サービス・レベルを保護しながら効率的なリバランス



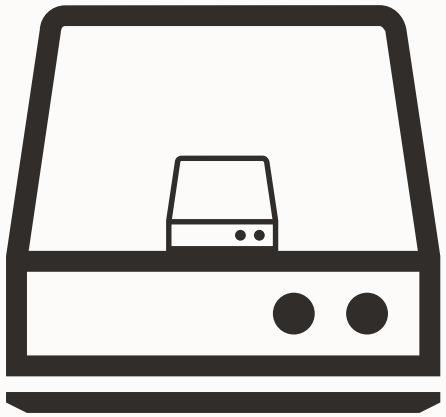
asm_power_limit ：インテリジェントで柔軟な電力設定のリバランス

- 冗長性のリストアとサービス・レベルの確保との間で最良のバランスを探るためにMAAラボにてテスト
- asm_power_limit 値のOracle MAAのベスト・プラクティス
  - デフォルト 4（クラスタ間全体、デプロイ時に設定）
  - **asm_power_limit = 0** の設定は**厳禁**
  - alter diskgroup <diskgroup name> rebalance modify power <value> を使用して動的に変更可能

asm_power_limit の最大推奨値	
Oracle Database 23ai	Oracle Database 21c以前
96	64

出典：LilGoldWmn https://www.freeimages.com/photo/balance-1218685

Exadata：データ保護
Exadata ASM構成のベスト・プラクティス

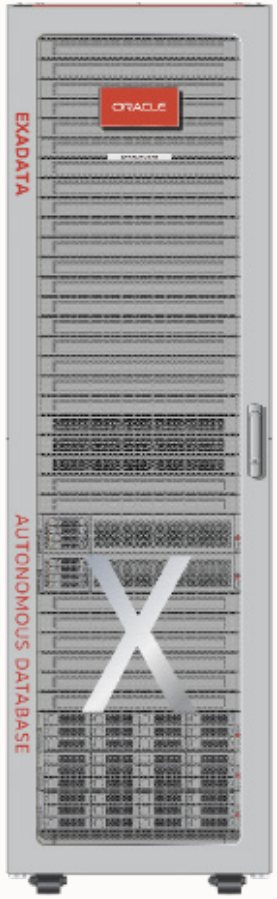


高冗長性ディスクグループ構成を**強く**推奨

- ディスク容量は増大し続ける
  - Exadata System Softwareのローリングでの更新適用中も2つのコピーを保持することが可能
  - 二重のパートナー・ディスク障害はまれだが発生する可能性がある
- **旧式の**ディスクを搭載した古いシステムでは特に重要
- ユーザー・データはプライマリ・エクステントに格納され、2つのコピーがミラー化
- 投票ファイルとASMメタデータ保持のために、5つの障害グループが必要
  - OCRはプライマリに格納され、2つのコピーがミラー化

Exadata：データ保護

Exadata ASM構成のベスト・プラクティス



高冗長性ディスクグループには、5つの障害グループを含むディスク・グループが最低1つ必要

- Eighthラック構成、Quarterラック構成には3つのストレージ・サーバーが存在
- ストレージ・サーバーだけだと障害グループは3つ存在

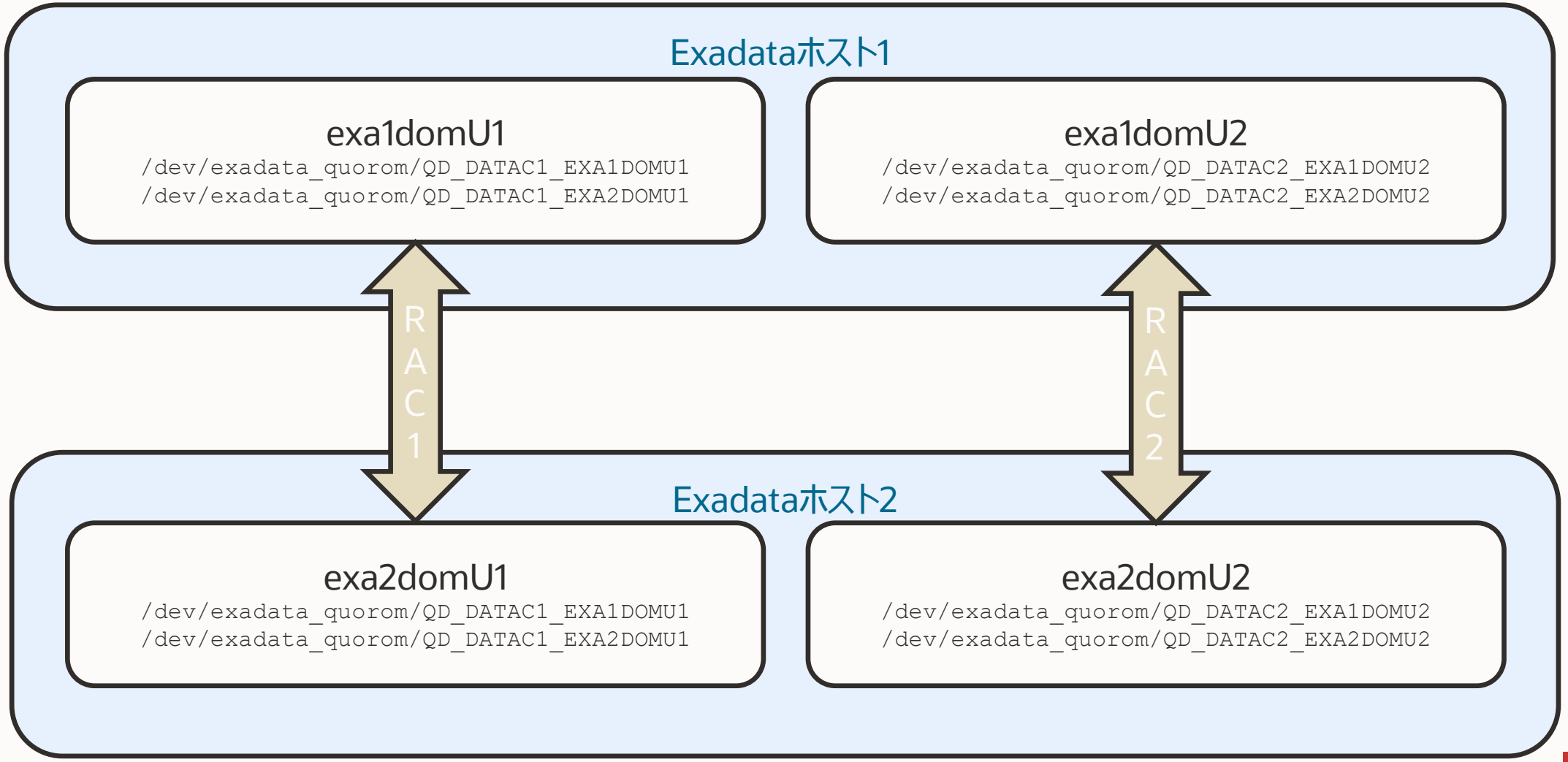
解決策：データベース・サーバー上に構成されたASMクォーラム・ディスク

- OEDAを使用したデプロイ時に自動的に実装
- iSCSIベースの ‘Quorum Failure Group’ を使用
- `quorumdiskmgr` コマンドを使用して管理（必要に応じて）

高冗長性ディスクグループ構成を**強く**推奨

Exadata：データ保護

Exadata ASM構成のベスト・プラクティス：Eighthラック構成、Quarterラック構成の高冗長性



Exadata：データ保護
ストレージ障害

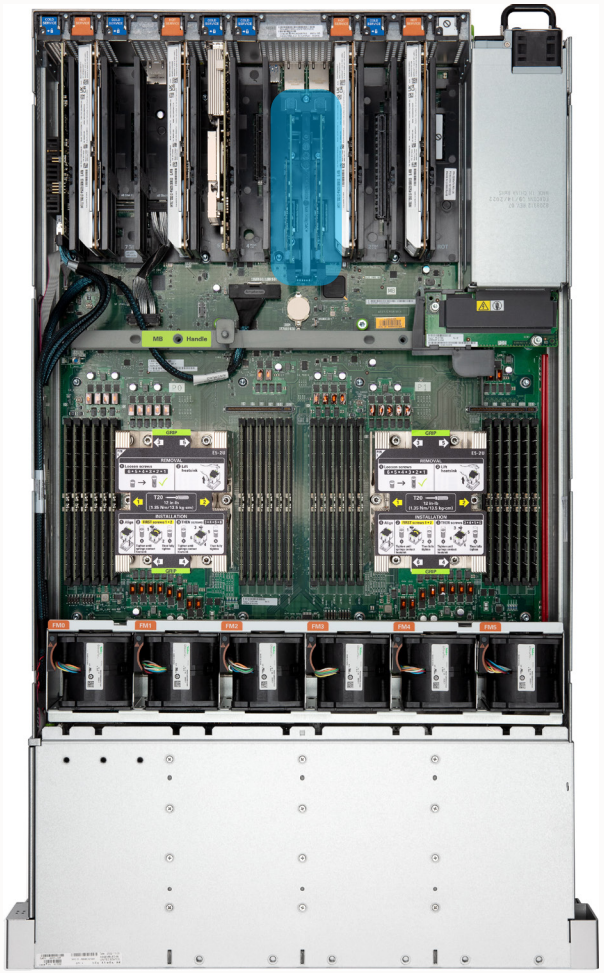
- Exadataには、ディスクに障害が生じたり、事前に問題があるとマークされた場合のディスク・メンテナンス用の自動化された操作が実装済

- ASMはディスク上のデータのリバランシングの前に自動的に冗長性を回復
  - 一部のデータにより冗長性が減少した可能性がある場合は所要時間を短縮

- ディスクを手動でドロップする必要がある場合は、管理者はMAINTAIN REDUNDANCY句を指定することで該当のASMディスクをドロップする前にデータのリバランスが可能
  - DROP FOR REPLACEMENT句を指定することで、実行する通常のチェックに加えて冗長性を維持

Exadata：データ保護

M.2ドライブの障害からの迅速な保護とオンライン交換（X7以降）

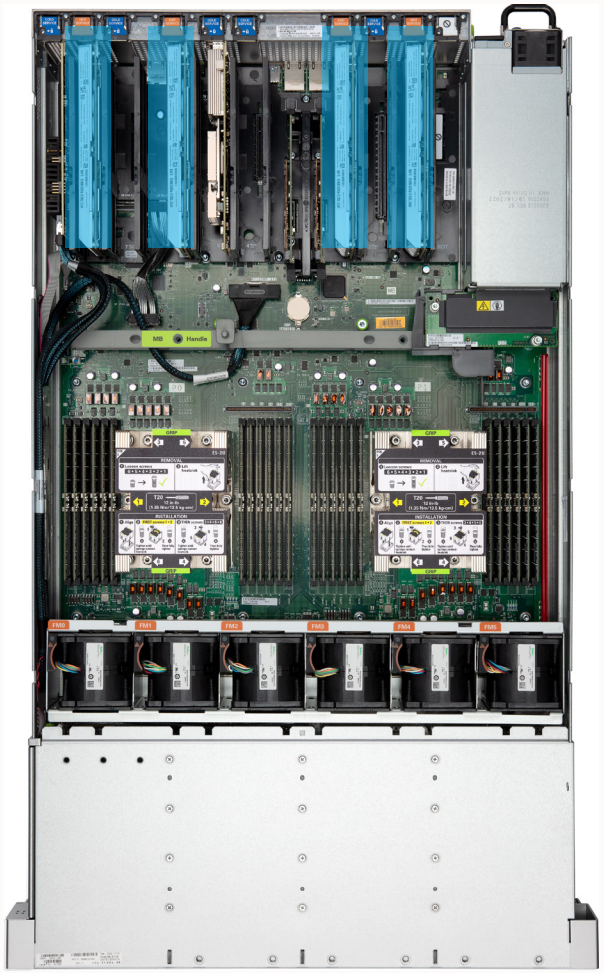


2台のM.2ドライブにOSとセル・ソフトウェアを搭載

M.2ドライブは、Intel RSTe RAIDにより保護

オンラインで交換できるため、
ユーザー・データのオフライン化は不要

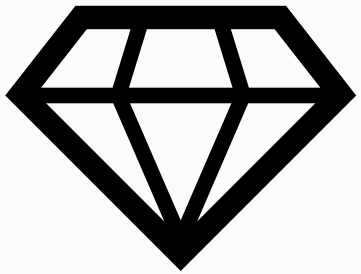
Exadata：データ保護
オンライン・フラッシュ交換（X7以降）



- ストレージ・サーバーを停止せずに、ストレージサーバーの上蓋を開き、オンラインで交換可能
- 障害の発生したドライブは、取り外しの準備が整ってから交換
- オンライン状態のドライブ交換の場合：
  - `CellCLI> alter physicaldisk FLASH_2_2 drop for replacement;`
- 交換後のユーザー操作が不要

Exadata：データ保護
Hardware Assisted Resilient Data

- Exadataには、特定のファイル・タイプの破損を防止するためのHardware Assisted Resilient Data（HARD）によるチェックが含まれる
  - SPFILE
  - 制御ファイル
  - ログ・ファイル
  - データファイル
  - Data Guard Brokerファイル



- HARDのチェックが失敗した場合、破損したデータは書き込まれない

- DB_BLOCK_CHECKSUM パラメーターを有効化した後に透過的に機能
  - ASMのリバランス中またはASM再同期中にアクティブに

Exadata：データ保護
ディスク・スクラビング

- アイドル時間中にハード・ディスクの検査と修復を実行
  - ディスクの不良セクターをチェック
  - Exadata System Softwareによる自動実行
  - 不良セクターが見つかると、Exadata System SoftwareがASMのミラー・コピーをリクエストして、修復を実行

- 自動かつ動的に実行
  - デフォルトで隔週にスケジューリング
  - ディスクがアイドル状態（ディスク・ビジー率が25 %未満）のときに実行
  - アプリケーションのI/Oリソースが必要な場合は自動でスクラブの実行が抑制

データ保護の結論

Exadataは以下を実現：

- 破損の検出、回避、および修復
- H.A.R.D.
- スクラビング
- フラッシュのオンラインでの交換
- M.2ドライブの障害からの迅速な保護とオンライン交換
- 高冗長性 ディスクグループ構成
- Do Not Service LED
- 効率的なリバランス
- （障害の可能性がある）フラッシュ/ドライブの電源自動再投入

出典：David Clode
https://unsplash.com/photos/Yg_sNKOiXvY

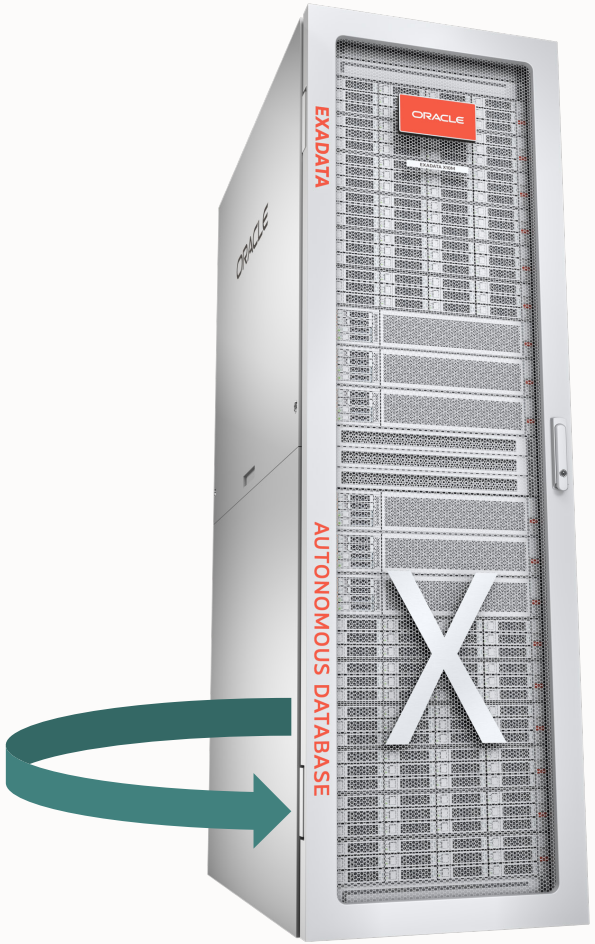
ライフサイクル管理

データ保護

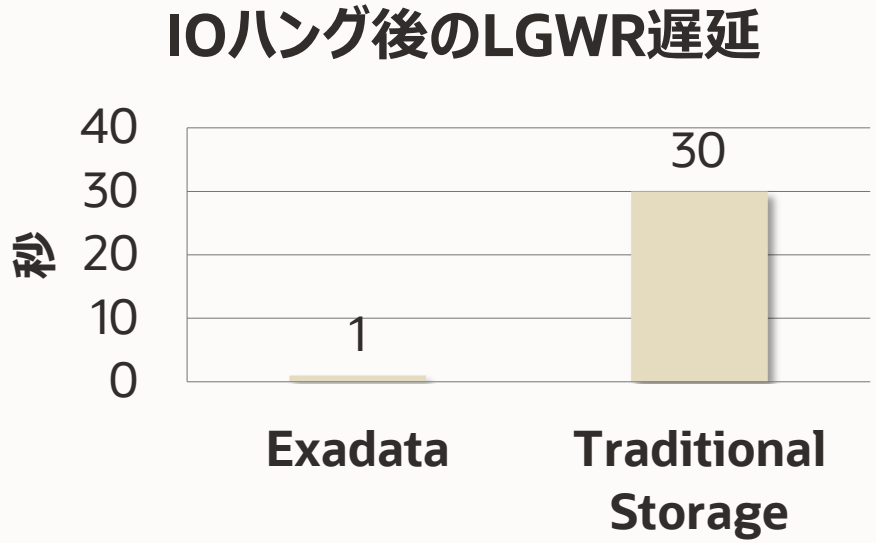
一時停止

サービス品質および
パフォーマンス

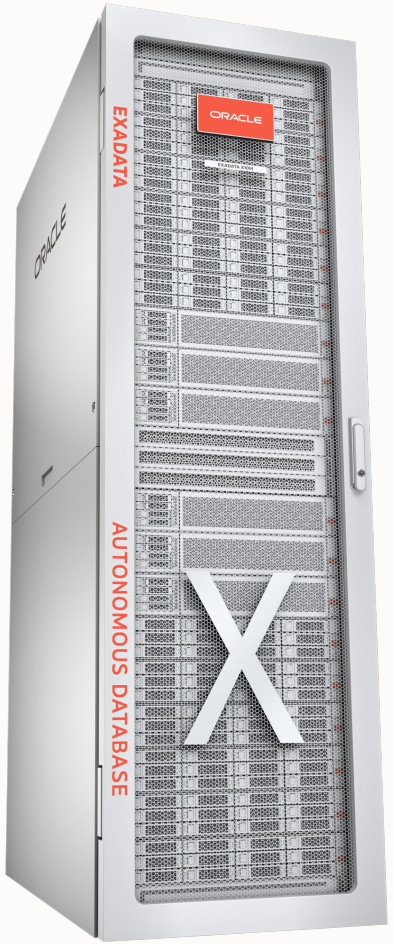
Exadata：サービス品質およびパフォーマンス
I/O Latency Capping



- 待機時間の長いI/Oはパフォーマンスに悪影響を与える可能性がある
- Exadataは待機時間の長いI/Oを検出し、読取りおよび書込みを他のデバイスにリダイレクト
  - 待機時間が長い読取りI/Oをパートナー・ストレージサーバーにリダイレクト
  - 待機時間が長い書込みI/Oはキャンセルされ、同一ストレージサーバー上のフラッシュに一時的に書込み



Exadata：サービス品質およびパフォーマンス
ストレージ・サーバー・ディスクの制限（Storage Server Disk Confinement）



- Exadataはディスクのパフォーマンスと健全性を絶えず監視
- パフォーマンスの低下はディスク障害の前兆であることが多い
- パフォーマンスが低下していると特定されたディスクは、I/Oの制限を加えられ、I/Oは代替ミラーへ送られる
- ストレージ・サーバーはディスクのヘルス・チェックを自動的に実行
- ディスクが健全であると見なされた場合
  - ディスクはサービスに戻り、再同期
- ディスクが健全でないと見なされた場合
  - ディスクはドロップされ、データはリバランスされて冗長性が維持され、サービスLEDが青に点灯
  - この時点でディスクを交換可能に

Exadata：サービス品質およびパフォーマンス
I/Oリソース・マネージャ（IORM）を備えたスマート・ストレージ

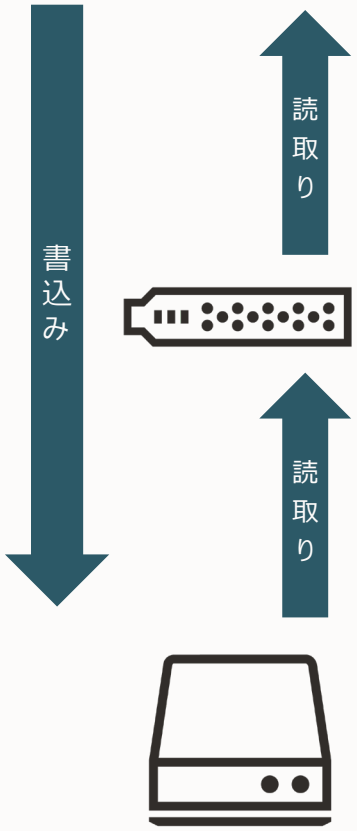
- I/Oの競合が発生した場合、ストレージ・サーバーでIORMを構成して、ストレージ・サーバーI/O関連リソースを構成して管理
- データベース（CDB/PDB/Non-CDB）またはClusterのIORMプランに基づいて、I/Oにタグ付けおよび優先順位付け
- タグに含まれるもの
  - データベース名/PDB/Cluster名
  - Objective
  - 優先順位
- 混合ワークロード環境および統合型ワークロード環境で有用
- データベース・リソース・マネージャと組み合わせて使用可能

```
CellCLI> ALTER IORMPLAN - dbplan=((name=prod, share=16), -
(name=dw, share=4), - (name=prod_test, share=2), -
(name=DEFAULT, share=1))
```

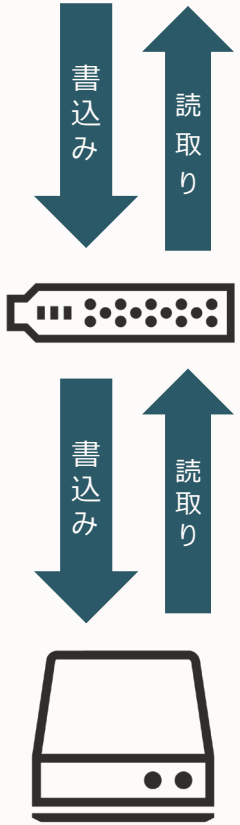
Exadata：サービス品質およびパフォーマンス

フラッシュ・キャッシュ：ライトバックまたはライトスルー

ライトスルー



ライトバック

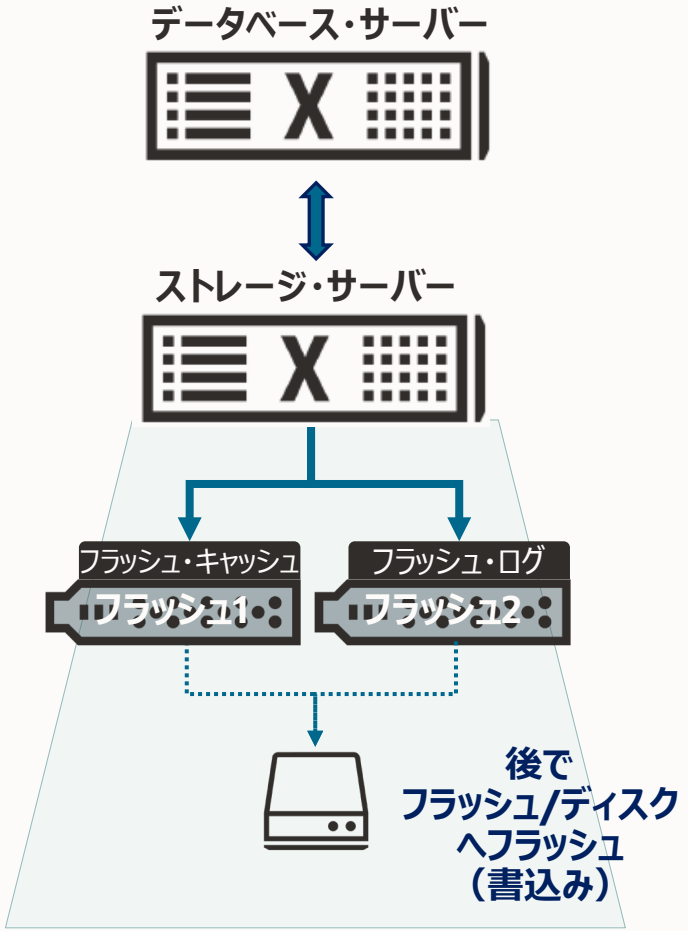


アプリケーションI/Oが、ハード・ディスクや容量最適化フラッシュに到達することはめったにない

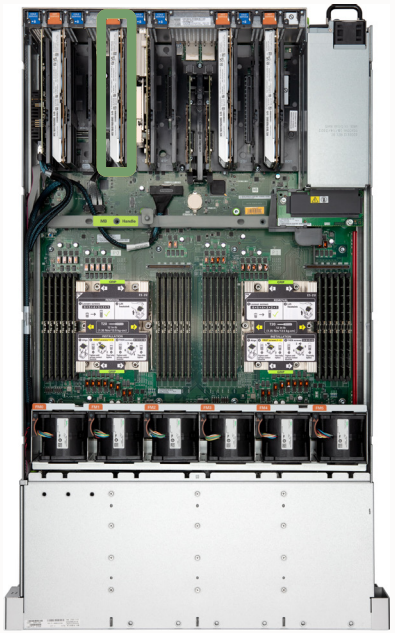
MAAのベスト・プラクティス

Exadata：サービス品質およびパフォーマンス
スマート・フラッシュ・ログ

- 書込み待機時間の外れ値を排除
  - REDOログ書込みの待機時間は、OLTPのデータベース・パフォーマンスにとって非常に重要
  - 異なるフラッシュ・デバイス上のフラッシュ・キャッシュおよびフラッシュ・ログへ同時に書込み
  - 最高速のデバイスが書込みを認識
- ログ書込みのボトルネックとなるストレージを排除
  - オンラインREDOLOG、スタンバイ環境のREDOLOGを、
    自動的かつ透過的に Write-Back Smart Flash Cacheにキャッシュ
  - ディスクではなくフラッシュに書き込むことでログ書込みスループットを向上
  - GoldenGateなどのオンラインREDOLOGを読み取るワークロードに有益
  - ハード・ディスクのI/O帯域幅を必要とする複数の同時ワークロード（バックアップなど）の場合に有益
- 容量最適化フラッシュまたはHDDへの非同期のフラッシュ（書込み）



Exadata：サービス品質およびパフォーマンス
Smart Flashとハード・ディスクの交換



- フラッシュ・ドライブまたはハード・ディスクの交換後、影響を受けたハード・ディスクに“health factor”が設定
- Health factorがオンの間、読取りは**健全なパートナー・ストレージ・セル**から実施され、Exadata System Softwareは障害ドライブを交換したストレージ・サーバー上で**フラッシュ・キャッシュのウォームアップ**を継続
- フラッシュ・キャッシュのウォームアップが完了したら、health factorのステータスはクリアされる
- この機能により、障害ドライブの交換後も**一貫性のある短いI/Oレイテンシー**が実現し、**アプリケーションのサービス・レベルが維持される**

Exadata：サービス品質およびパフォーマンス

計画メンテナンス時または計画外メンテナンス時におけるSLAの維持

- ASMリバランス操作時に
  Exadataフラッシュ・キャッシュ・ステートを保持
  現実的な例としては、
  セル・ソフトウェアのローリング更新時の再同期

- ストレージサーバーに対するI/Oリクエストが
  インテリジェントにルーティングされることにより、
  フラッシュやディスクの障害および交換後も、
  最良のサービスレベルを実現することが可能に

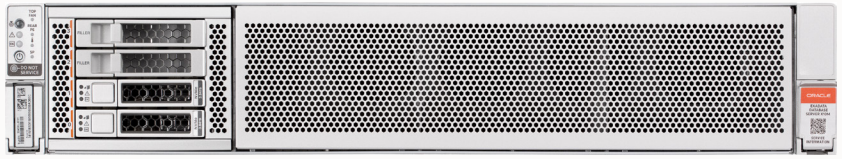
- 計画外停止および計画メンテナンスの両方に適用

パフォーマンスは時間なり
時は金なり

Exadata：サービス品質およびパフォーマンス

データベース層のI/Oキャンセル

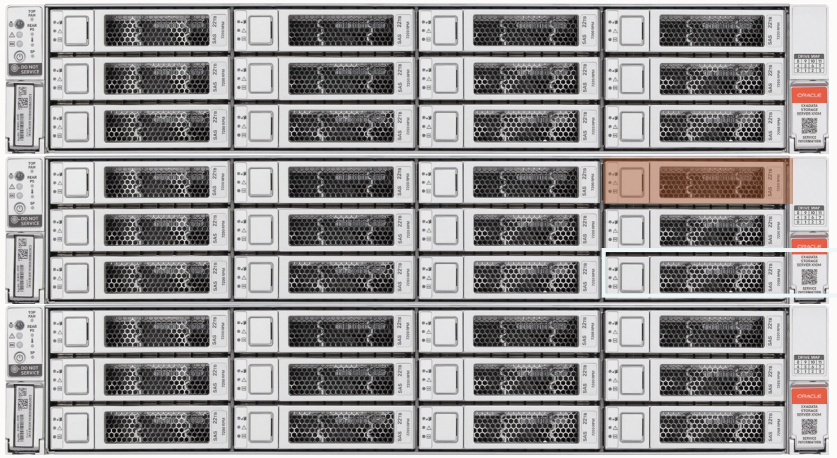
データベース層



データベース層のI/Oレイテンシー制限✓

過剰なI/O

ストレージ層



I/Oが低速? セルのI/Oレイテンシー制限✓

I/Oがハング? I/Oハングの検出/修復✓

ディスクの不良?Disk Confinement✓

検出されないハードウェアまたは
ソフトウェアの問題?

Exadata ASMのリバランス用の予約領域

- ASMには障害発生時にデータのリバランスを可能にするための領域が必要
  - リバランスの成功を確保
  - 冗長性をリストア
  - リバランスの成功を確保するための領域は、予約されていない
  - リバランスを完了するための十分な領域がない場合、ORA-15041が報告される

REQUIRED_MIRROR_FREE_MB
- 障害グループの数およびASMバージョンに依存
  すべてのディスク・グループとすべての冗長性（高または標準）に適用
- すべてのメディア・タイプとハードウェア世代に同じように適用*

Grid Infrastructure のバージョン	障害グループの数	ディスク・グループで必要な空き領域（%）
12.1.0	任意	15
12.2、18.1 以降	5未満	15
12.2、18.1 以降	5以上	9

* Exadata X10M Extreme Flashモデルにはハードウェア固有の要件がある

障害グループの数（8つのASMディスク/障害グループ）	冗長性	単一の物理ディスク障害後にリバランスを正常に行うためにディスク・グループで必要な空き領域（%）
5未満	NORMAL	15 %
5未満	HIGH	29 %
5以上	NORMAL	9 %
5以上	HIGH	11 %

- X10 EFセルには4つの物理フラッシュ・ディスクがあり、物理フラッシュ・ディスクごとに2つのASMディスクがあります。したがって、フラッシュ・カードに障害が発生すると、2つのASMディスクがドロップされます。
- GI/ASM 19c以降（パッチ34281503適用済みの場合）。

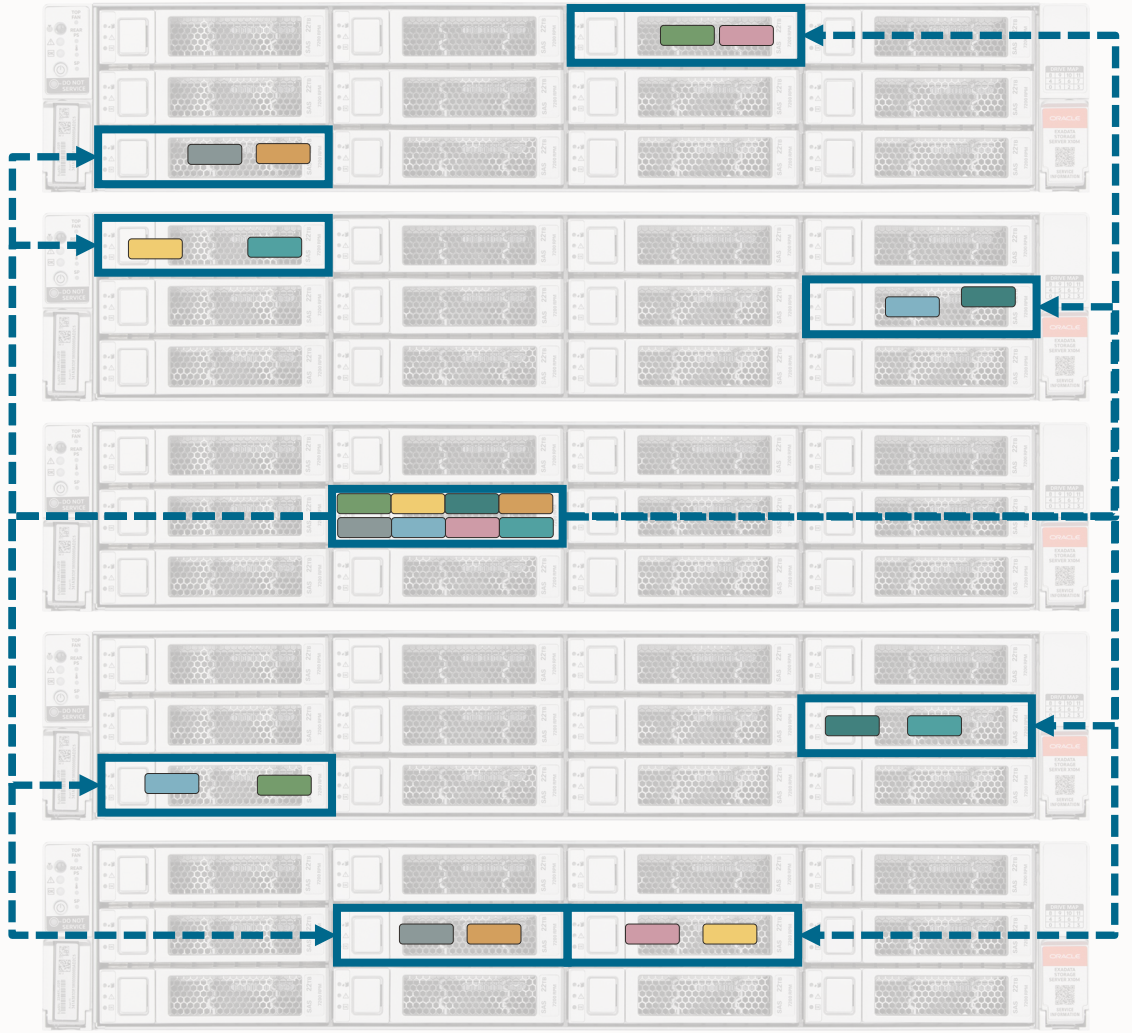
Exadataのスマート・リバランス機能

- スマート・リバランス機能は高冗長性ディスク・グループでの障害発生時に影響を及ぼす
  - ディスク・グループが必要な空き領域を保持していた場合
    - データはリバランスされて冗長性はリストアされる
  - ディスク・グループが必要な空き領域を保持してなかった場合
    - ディスクはオフラインになり、リバランスは延期される
    - ディスクが交換されると、パートナー・ディスクから効率的に再ミラー化される
- データベース・ストレージにさらに容量が必要な場合、障害時の余分なI/Oやデータ移動を削減

スマート・リバランスは安全策
MAAでは十分な空き領域を維持することを強く推奨します。

ASMディスクのパートナー設定
コンセプト

- ASMはディスクのパートナーシップを利用して、エクステントとそのミラー・コピーを配置するディスクを選択
- 各ディスクは他の8つのディスクとパートナーとなる
  - 5セル未満の場合、すべてのディスク・パートナーは2セルから構成
  - 5セル以上の場合、すべてのディスク・パートナーは4セルから構成
- プライマリ・エクステントがミラー化される
  - 高冗長性の場合はこれらのパートナーの内の2つから
  - 標準冗長性の場合はこれらのパートナーの内の1つから
- 読取りI/Oは8つのディスクにより提供される
  - リバランス、再構築、再同期、リシルバ、ディスク/フラッシュのウォームアップ操作で使用される

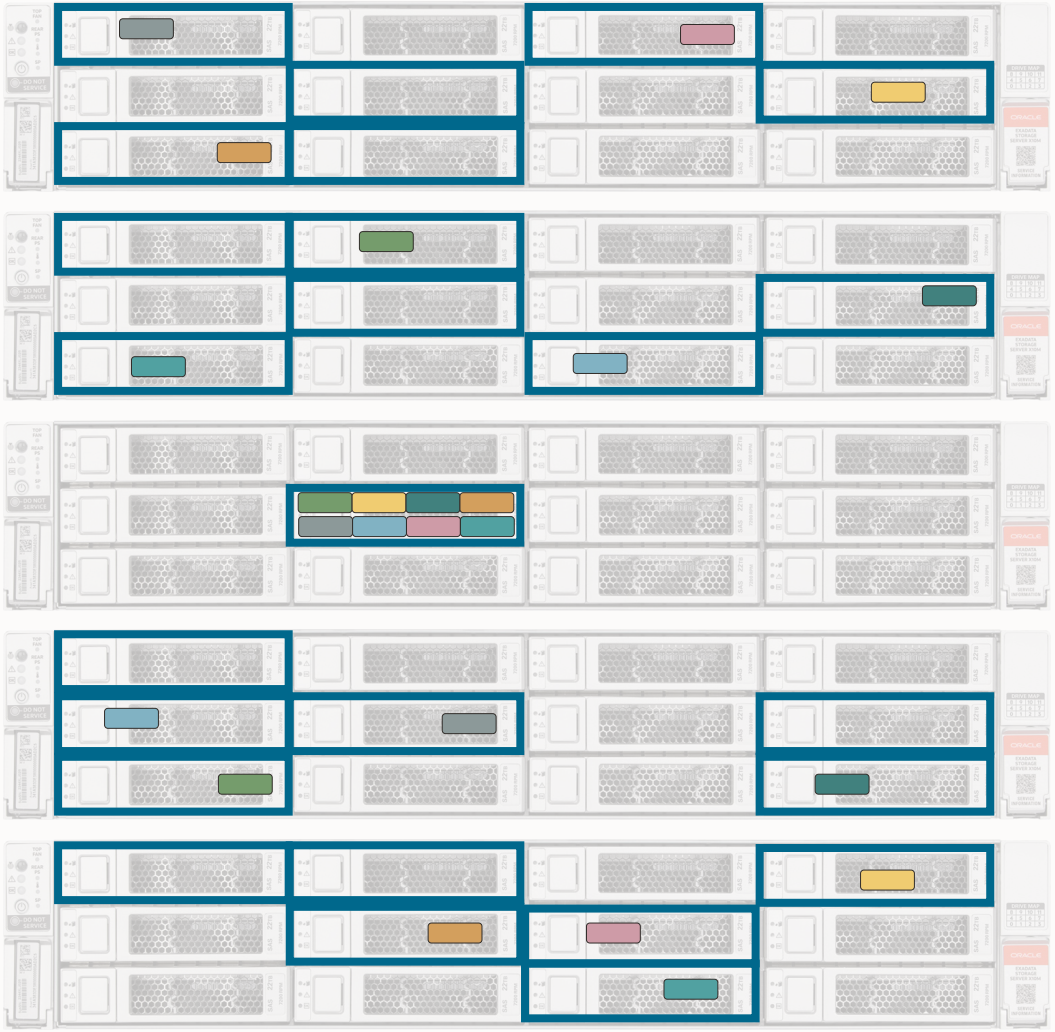


ASM 23aiでのディスク・パートナー数の増加

New In 23ai

- 各ディスクは4個のセル上で他の24個のディスクとパートナーとなる
- 読取りI/Oは24個のディスクにより提供される
  - リバランス、再構築、再同期、リシルバ、ディスク/フラッシュのウォームアップ操作で利用される
- ASMにより自動的に管理
  - 新しいパートナー設定スキームは、アップグレード時には適用されない
  - ディスク・パートナーは次の操作により更新される
    - ディスクの追加
    - 障害グループの追加（セルの追加）
    - リバランス

冗長性のリストアが
最大3倍高速化



その他のストレージ構成の場合は？

New In 23ai

- 異なるストレージ・サーバー構成は異なるパートナーシップ値を利用する

セルの数	ストレージ・サーバーのタイプ	セルあたりのディスクの数	ディスク・パートナーの数（23aiより前）	ディスク・パートナーの数（23ai）
3	Eighth ラック HC	6	8	12
3または4	HC	12	8	12
5以上	HC	12	8	24
3または4	Extreme Flash	8	8	8
5以上	Extreme Flash	8	8	16

- 以下の操作で効果がある
  - 再構築 (REBUILD) – ディスク障害
  - 最同期 (RESYNC) – セルのパッチ適用
  - リシルバ (RESILVER) – フラッシュ・カードの障害
  - ディスク/フラッシュのウォームアップ

注：5番目のセルを構成に追加すると、
適用されるディスク・パートナーの数が増えるため、
リバランスの実行時間が長くなります。

Exadata：サービス品質およびパフォーマンス
容量計画：メモリ構成

メモリのスワップの発生がパフォーマンスと安定性の問題を引き起こす可能性

適切なメモリ構成で以下が回避：

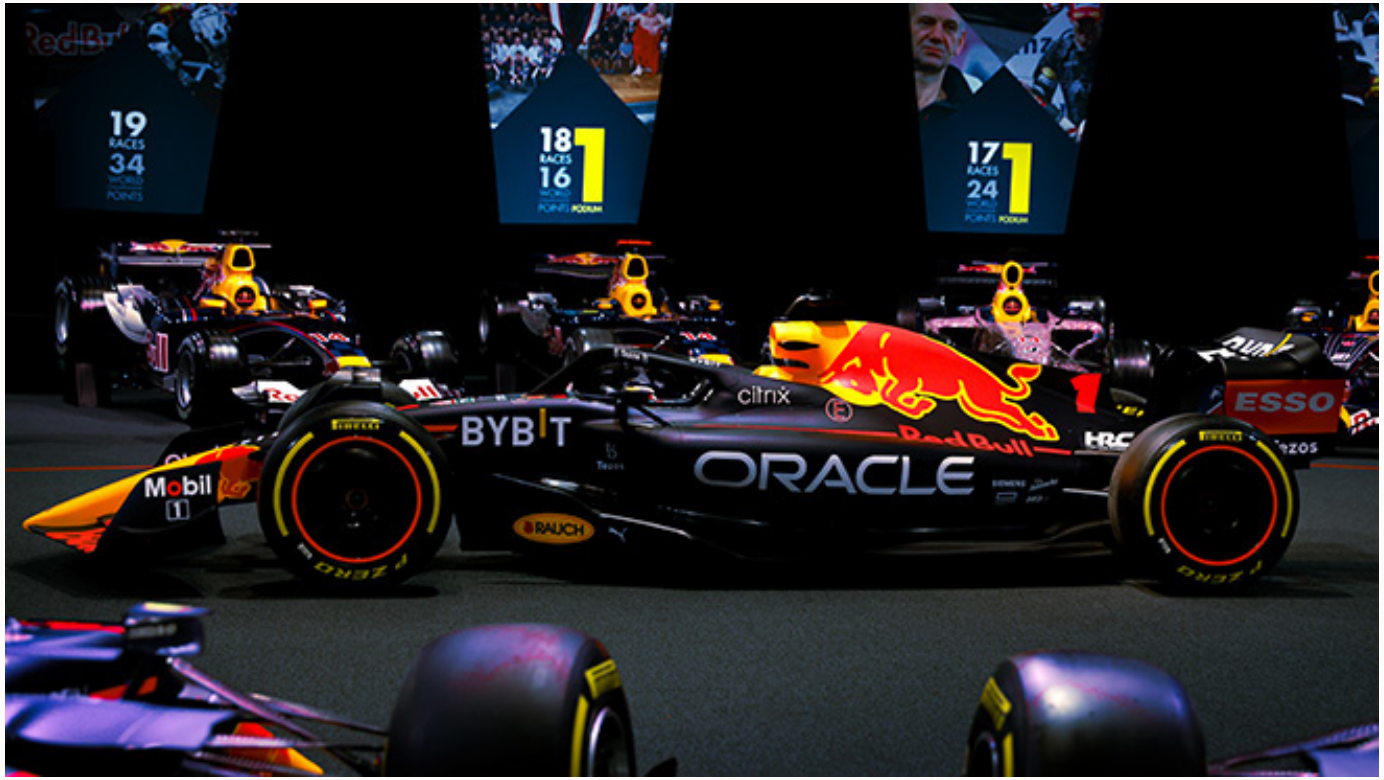
- スワッピング
- 不安定性

出典：Kathy
https://unsplash.com/photos/R7nSPG8edVI

Exadata：サービス品質およびパフォーマンス

Exadataは高いパフォーマンスを実現するために構築されています

- Smart Scan
- スマート・フラッシュ・キャッシュ
- ストレージ索引
  - “最速のI/O処理が必要なわけではない”
- Hybrid Columnar Compression
- インメモリ列形式
- RDMA
- Real-Time Insight



Exadata：サービス品質およびパフォーマンス

RDMAネットワーク・ファブリック

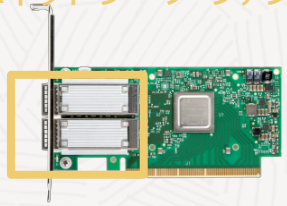
2 つのアクティブ/アクティブ・ポート
（すべてのRDMAネットワーク・ファブリック・アダプタ用）

2 つのRDMAネットワーク・ファブリック・スイッチ
（すべてのExadataシングルラック用）

22 のポート（スイッチあたり）を内部クラスタ・ネットワークに使用し、
シングル・ポイント障害が存在しないように配線

- Exadataが専用で使用
- スイッチ・レベルの設定は変更されない
- ZFSシステムを構築する場合は、スケーラビリティと柔軟性のために、
  ZFSシステムはTop of Rack（ToR）スイッチ経由で接続することが推奨

RDMAネットワーク・ファブリック・アダプタ



RDMAネットワーク・ファブリック・スイッチ

Exadata：サービス品質およびパフォーマンス
ワークロードの自動優先順位付け

RDMAファブリックは自動サービス品質（QoS）を実装

特定のトラフィック用のQoSレーンを分離
- 重要なI/O–LGWR
- ディスク読取り
- ディスク書込み

RoCEネットワーク・レジリエンス

Exadata RoCE IPは高可用性である必要がある

- 各サーバーにはデュアル・ポートRoCE NICがあり、各ポートは異なるリーフ・スイッチに接続
- スイッチ・ポートが"ダウン"している場合は自動的にフェイルオーバー

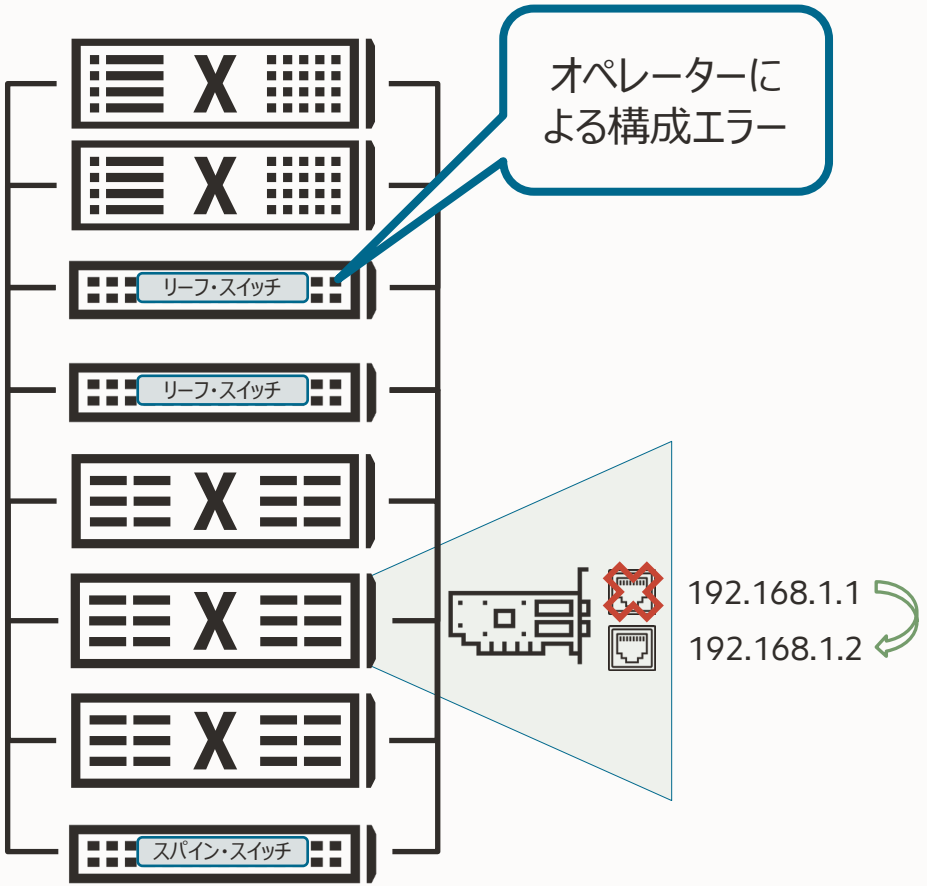
スイッチまたはネットワークに異常があると、ポートは"稼働"状態のままになる場合があるが、ネットワーク・トラフィックは停止して流れなくなる

- スイッチの不適切な構成
- 過剰な一時停止（PAUSE）フレーム

ネットワーク・トラフィックの停止により、
データベースが不安定になったり停止したりする可能性がある

ExaPortMonプロセスはホスト上で実行され、
両方のRoCEポートのライブ・トラフィックを監視

- 停止が検出された場合は、IPを動作しているポートに移行
- アップストリームの問題が解決されたらIPを元のポートに戻す

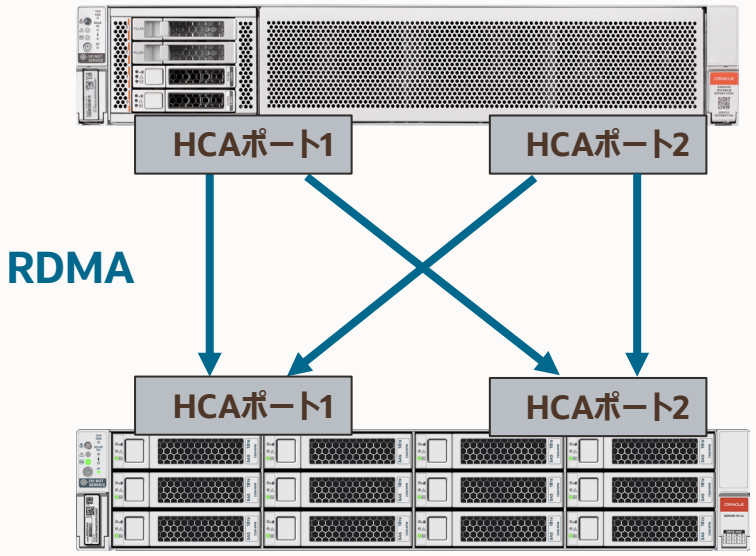


Exadata：サービスとパフォーマンスの品質
Instant Failure Detection（IFD）

- 従来のシステムはソフトウェアを使用して可用性を確認
  - 高負荷ではパフォーマンス上の問題を引き起こす場合がある
  - TCPのタイムアウトを待つ必要がある

- Exadataではサーバーの可用性の確認にRDMAを使用
  - Instant Failure Detection
  - 冗長性のために4つのRDMAパスを利用
    - データベース・サーバー↔ストレージ・サーバー
    - データベース・サーバー↔データベース・サーバー

- 短期間の後に4つのパスの通信がいずれも使用できない場合、該当サーバーは排除（eviction）される

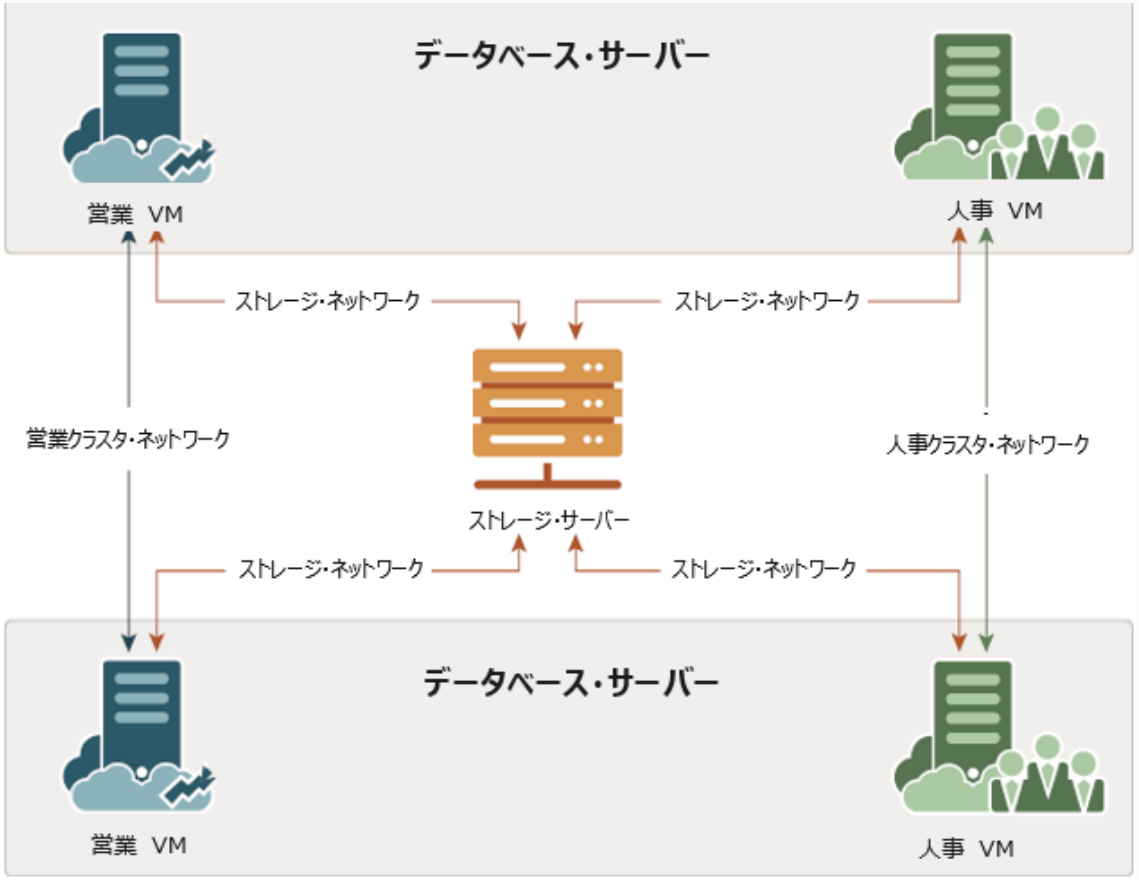


Exadataでは1秒未満で不具合が通知されるのに対し、非Exadataプラットフォーム上ではタイムアウトに最大1分間必要

Exadata Secure RDMA Fabric Isolation for RoCE

Exadata Secure Fabric for RoCE システムでは、共有のExadata Storage Serverへのアクセスを許可しながら、仮想マシンのネットワーク分離を実装

- 各VMクラスタにはプライベート・ネットワークが割り当てられる
- VMクラスタ間は相互に通信不可能
- すべての仮想マシンが共有ストレージ・インフラストラクチャと通信可能
- セキュリティのバイパスは不可
  - ネットワーク・カードにより、すべてのパケットに対しての強制的にセキュリティが実装
  - ルールはハイパーバイザによって自動的にプログラムされる



サービス品質を実現するための機能

出典：Towfiqu barbhuiya
https://unsplash.com/photos/0ZUoBtLw3y4

- セル側のI/Oレイテンシー制限
- ストレージサーバーのDisk Confinement
- I/Oリソース・マネージャ（IORM）を備えたスマート・ストレージ
- スマート・フラッシュ・ログ
- ライトバック・スマート・フラッシュ・ログ
- スマートなフラッシュ交換
- Exadata RDMA Memory Data Accelerator
- RDMAネットワーク・ファブリック
- RDMAでのQoS
- Instant Failure Detection
- Exadata Secure RDMA Fabric Isolation

ライフサイクル管理

データ保護

一時停止

サービス品質および
パフォーマンス

Exadata：一時停止

停止と一時停止

停止

- サービス・レベルの**完全な**中断

一時停止

- サービス・レベルの**大幅な**低下

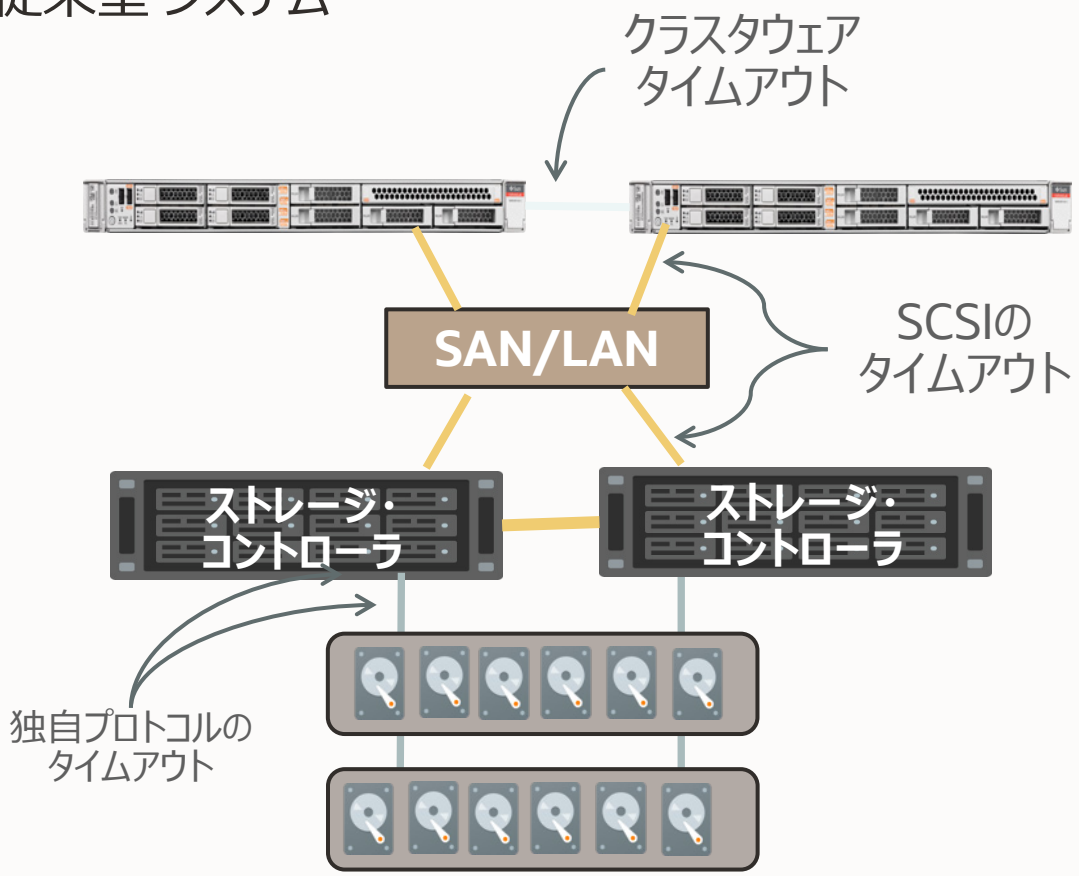
生産性の低下と収益の損失

システムは複雑で、1つのレイヤーの問題が他のレイヤーに連鎖する可能性がある

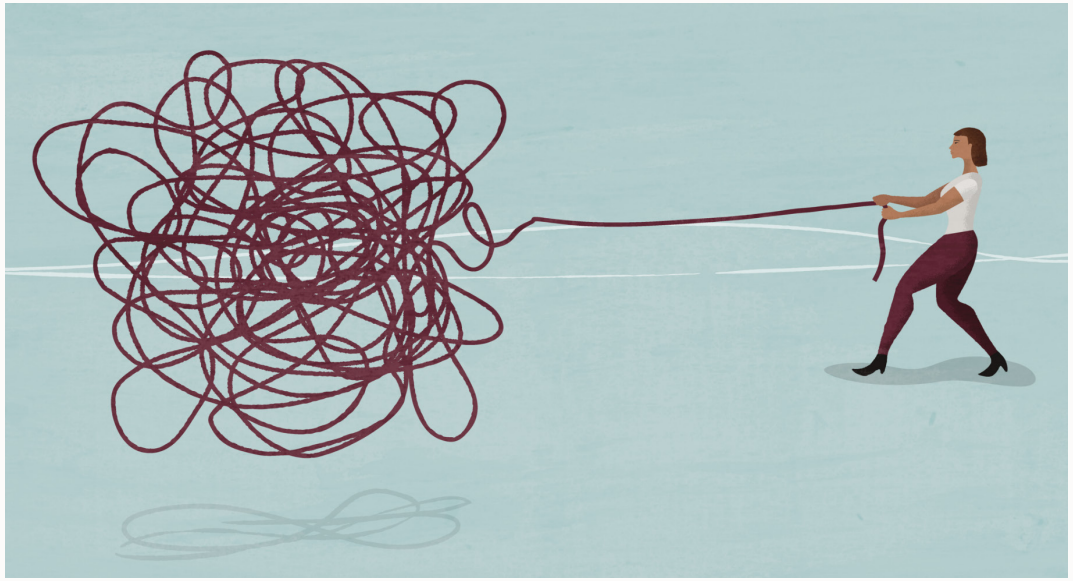
オラクルのエンジニアド・システムとOracle MAAのベスト・プラクティスは、
この問題に対処できるように設計およびチューニングされている

Exadata：一時停止

従来型 システム



- レイヤーごとに個別に障害が検出され、タイムアウトが発生
- 通常は障害検出の時間が付加される
  例：ストレージ・コントローラが故障すると、DBサーバーがこの障害を検出するために、2回のSCSIタイムアウト発生が必要



ストレージサーバー・コントローラ・キャッシュ障害への対処

データ損失の自動回避

コントローラのキャッシュ障害は、カスタム構築されたシステムや、
初期のExadataシステムでは、複雑になる可能性があった

Exadata System Software 21.2 以前は、ユーザーは以下の手順を手動で実行して、コントローラ・キャッシュの障害からリカバリする必要があった

- コンソールから、続行する方法について、コンソールで不可解な質問に回答
- コントローラーを交換する前に、
  グリッド・ディスクを必ず強制的にドロップしてからコントローラを交換

Credit Liam Riby
https://unsplash.com/photos/j8h2_9UDqrM

ストレージサーバー・コントローラ・キャッシュ障害への対処

データ損失の自動回避

Exadata System Software 21.2 以降を使用している場合、以下を実行することで、ストレージサーバー・コントローラ・キャッシュ障害からの修復を自動で処理

- 障害発生後にセル・サービス （cellsrv）を起動する前に問題を検出
- グリッド・ディスクへのアクセスを無効化
- 障害が発生したディスクをリカバリ

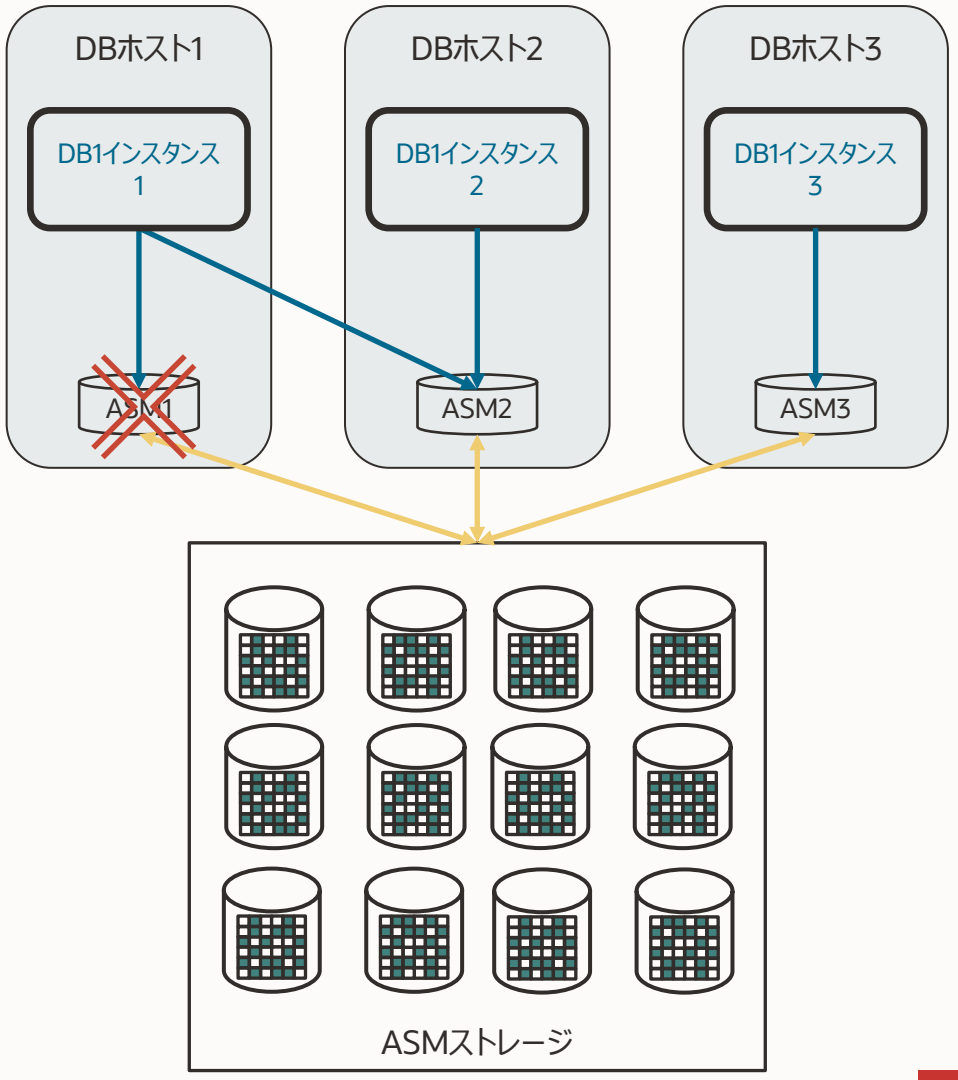
出典：Connor McSheffrey
https://unsplash.com/photos/MIspM6Hlit8

Exadata：一時停止

一時停止と停止：Oracle Flex ASM

Oracle Flex ASM機能により、特定のデータベース・サーバーからと別の物理サーバー上でOracle ASMインスタンスを実行可能

- RDBMSとASM間の継続的な通信が可能
- ASMインスタンスの障害時、データベース・サービスのフェイルオーバーは不要
- アプリケーションに影響はなく、サービス・レベルにも影響なし
- ExadataではカーディナリティⅠ（ASMインスタンス数）のパラメーターをALLに設定

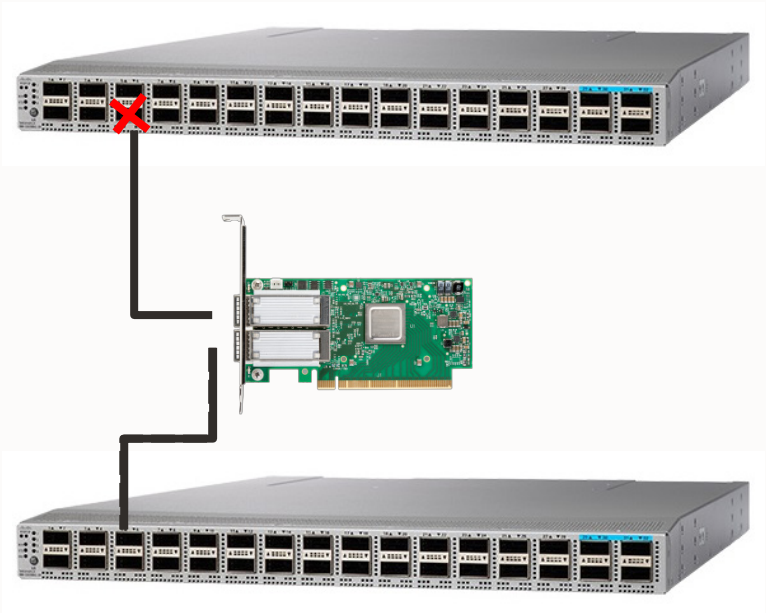


Exadata：一時停止

クライアント・ネットワーク・ポート障害における一時停止の減少

アクティブ/パッシブ設定のクライアント・アクセス・ネットワーク・ポートは、障害に関連する一時停止が極めて少ない

LACP “アクティブ/アクティブ”も構成可能。
この場合、ネットワーク・インフラストラクチャの設定変更が必要



アクティブ接続

パッシブ接続

Exadata：一時停止

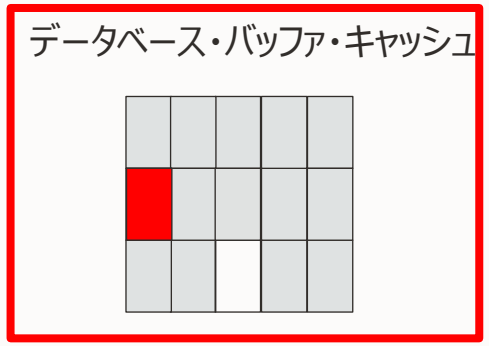
ストレージ・サーバーのシャットダウン時におけるスマート・ハンドシェイク

- ストレージ・サーバーがシャットダウンされると、データベース・サーバー上のGrid Infrastructureのdiskmonプロセスに通知が送信

- メンテナンスのためにストレージ層がシャットダウンされる場合は停止は発生しない

Exadata：一時停止
スマートOLTPキャッシング

Exadataのデータ・アクセス層の概要



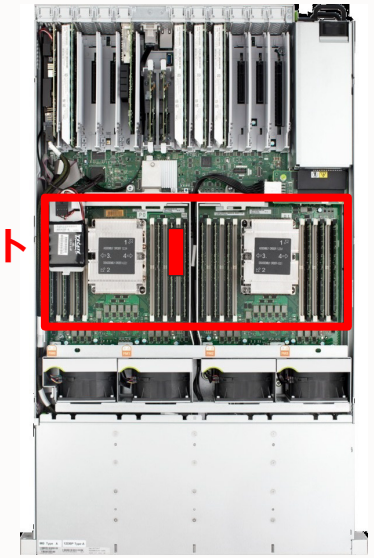
1.バッファ・キャッシュへのデータ読取り

2.バッファ・キャッシュの領域を解放するために
DBWRがバッファを排除

頻繁なアクセス

1.プライマリ・ミラーのセルを
待機時間が極めて短いデータ・アクセラレータに移入

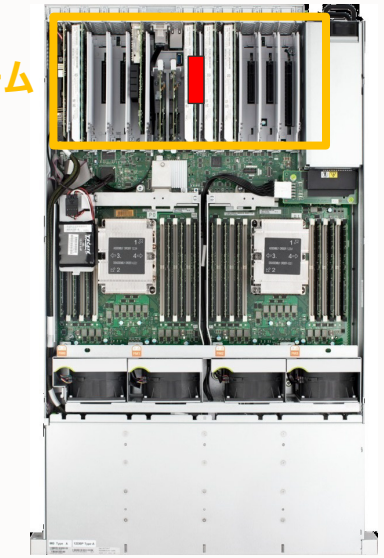
2.プライマリのセルは
待機時間が極めて短いデータ・アクセラレータへの移入を維持



ホット

2.セカンダリ・ミラーのセルを
待機時間が短いフラッシュ・キャッシュに移入

ウォーム



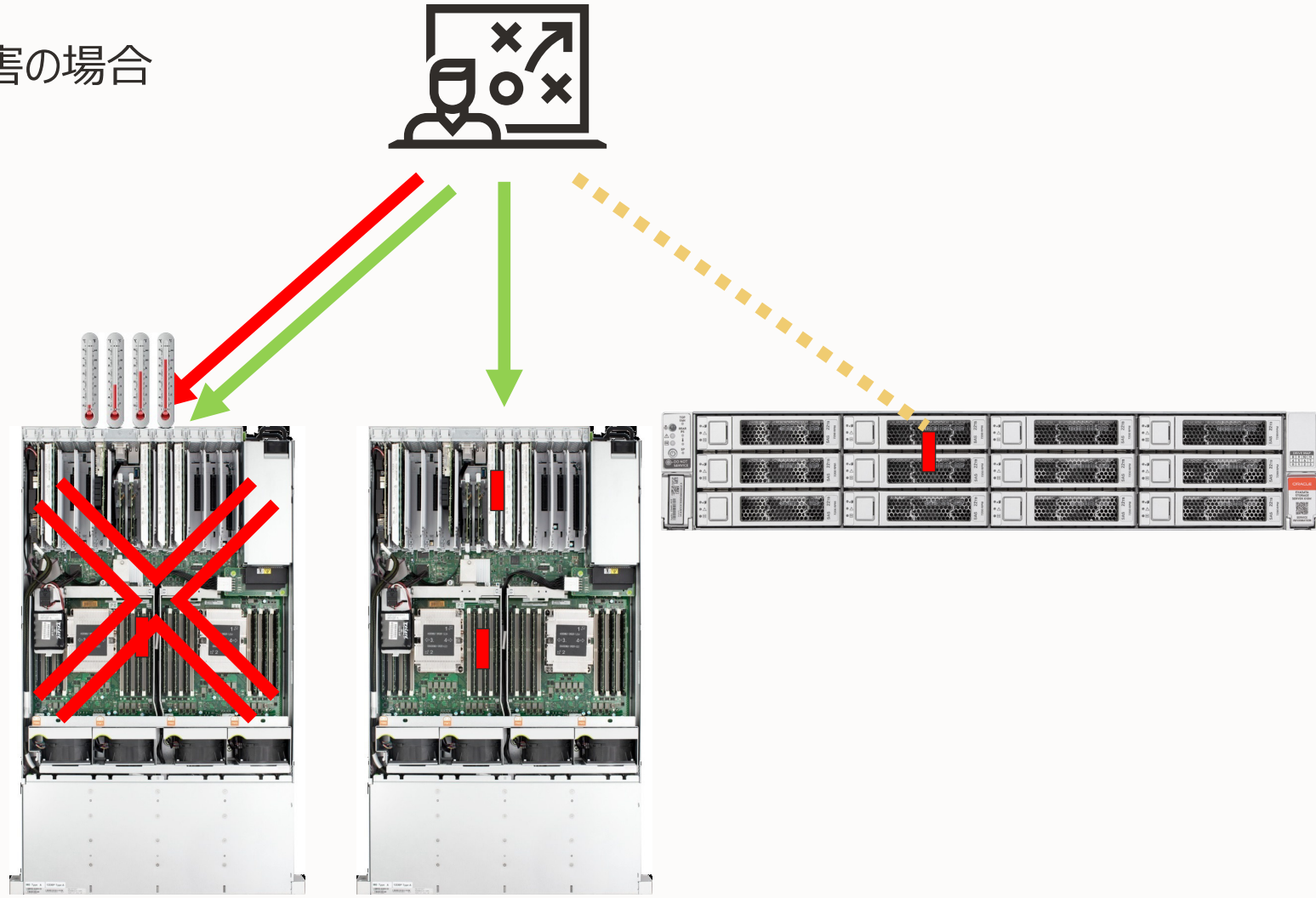
3次ミラーのセルを
待機時間が長いハード・ディスクに配置

コールド

Exadata：一時停止

スマートOLTPキャッシング– ストレージ障害の場合

- アプリケーションは
  プライマリ・ミラーからデータを読取り
- プライマリ・ミラーを保持するセルで
  ストレージ障害が発生
- レイテンシーの短いフラッシュ上のセカンダリ・ミラーからデータを取得し、
  待機時間が極めて短いデータ・アクセラレータに移入
- マーフィーの法則が正しい場合に備えて、
  3次ミラーが引き続き保護を提供
- ストレージ障害を修復し、フラッシュ・キャッシュにデータをウォームアップしたら、
  アプリケーションの読み込み先がプライマリ・コピーに戻る

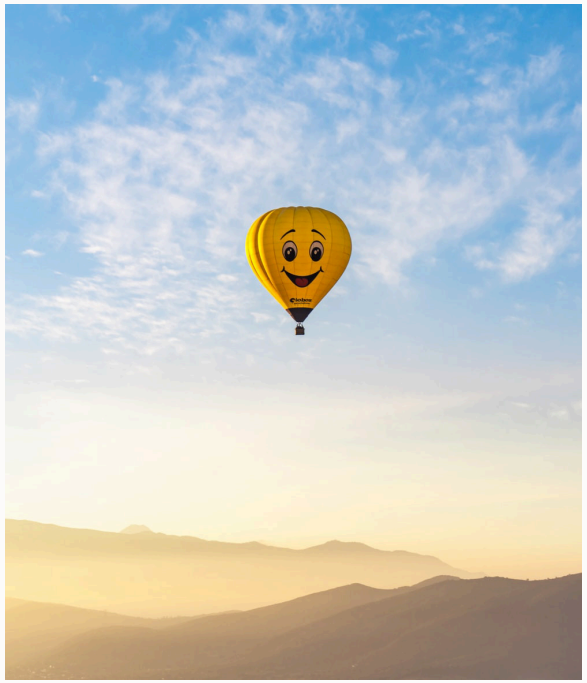


エクステントを移動する際は、*ASM*によるリバランス、再同期、リシルバによって、
常にフラッシュ・キャッシュ・ステートが維持される

Exadata：一時停止

セル間のリバランス時のデータ・アクセラレータ移入の保持

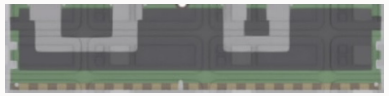
- ディスク障害によりリバランスが発生
- プライマリ・ミラーがデータ・アクセラレータにキャッシュされている
- プライマリ・ミラーが他のセルに移動
- データ・アクセラレータ上にキャッシュされたデータも同様に移動
- レイテンシーが保持され、エンドユーザーの満足度も維持される



出典：Jacob Vizek
https://unsplash.com/photos/ibvHQnpk4LE

データ・
アクセラレータ

データ・
アクセラレータ



Exadata：一時停止

Cisco RoCEスパイン・スイッチ・ソフトウェアの更新

- Oracle MAAチームはマルチラック構成でスパイン・スイッチのリブートをテスト

- Exadata System Software 21.2.* 以降
  - 停止なしでファームウェア更新が可能に
  - 一時停止が大幅に減少

ライフサイクル管理

データ保護

一時停止

サービス品質および
パフォーマンス

Exadata：ライフサイクル管理

Oracle Exachk

- EXAchkの推奨事項は、現場で発生するお客様の障害について、開発エンジニアが週次ミーティングで検討して実装

- 常に最新バージョンのExachkを使用することが重要
  - 重大な問題の追跡
  - 特定のリリースにおける問題
  - 180日以上前のバージョンのExachkの実行は許可されない

- EXAchkを以下の頻度で実行することを強く推奨
  - 1か月に1度
  - 構成の大規模な変更前後
    （パッチの適用、ストレージの追加など）

- ベスト・プラクティスのヘルス・チェック

Database Server

Status	Type	Message	Status On	Details
CRITICAL	Database Check	Database parameter CLUSTER_INTERCONNECTS is not set to the recommended value	random01client02:rac1,random01client02:cdbm18c	View
CRITICAL	Database Check	Database parameters log_archive_dest_n with Location attribute are not all set to recommended value	All Databases	View
CRITICAL	OS Check	Hardware and firmware profile check is not successful. [Database Server]	All Database Servers	View
CRITICAL	OS Check	The InfiniBand Address Resolution Protocol (ARP) Configuration on Database Servers should be as recommended	All Database Servers	View
FAIL	SQL Check	Some data or temp files are not autoextensible	cdbm122	View
FAIL	OS Check	Memlock settings do not meet the Oracle best practice recommendations	All Database Servers	View
FAIL	ASM Check	Fast recovery area allocation totals are greater than the total space of the DB_RECOVERY_FILE_DEST disk group	All ASM Instances	View
FAIL	OS Check	Active kernel version should match expected version for installed Exadata Image	All Database Servers	View
FAIL	OS Check	One or more database server has non-test stateless alerts with null "examinedby" fields	All Database Servers	View
FAIL	OS Check	One or more database servers have stateful alerts that have not been cleared	All Database Servers	View
FAIL	Database Check	Hidden database Initialization Parameter usage is not correct	All Databases	View
WARNING	Database Check	Local listener init parameter is not set to local node VIP	random01client02:cdbm18c	View
WARNING	Database Check	Database parameter DB_BLOCK_CHECKING on PRIMARY is NOT set to the recommended value.	All Databases	View
INFO	OS Check	Exadata Critical Issues (Doc ID 1270094.1):- DB1-DB4,DB6,DB9-DB41, EX1-EX54,EX56 and IB1-IB3,IB5-IB8	All Database Servers	View
INFO	Database Check	One or more non-default AWR baselines should be created	All Databases	View

Exadata：ライフサイクル管理

Exachk

Cluster Summary

Cluster Name	Cluster-c1
OS/Kernel Version	LINUX X86-64 OELRHEL 7 4.14.35-2047.505.4.4.el7uek.x86_64
CRS Home - Version	/u01/app/21.0.0.0/grid - 21.3.0.0.0
DB Home - Version - Names	/u01/app/oracle/product/21.0.0.0/dbhome_1 - 21.3.0.0.0 - cdbm213 database /u01/app/oracle/product/19.0.0.0/dbhome_1 - 19.12.0.0.0 - cdbm19c database /u01/app/oracle/product/18.0.0.0/dbhome_1 - 18.14.0.0.0 - cdbm18c database /u01/app/oracle/product/12.2.0.1/dbhome_1 - 12.2.0.1.210720 - cdbm122 database /u01/app/oracle/product/12.1.0.2/dbhome_1 - 12.1.0.2.210720 - 3 databases
Exadata Version	21.2.4.0.0
Number of nodes	8
Database Servers	2
Storage Servers	3
IB Switches	3
EXAchk Version	21.3.0_20211029
Collection	exachk_random01client01_rac12c_031[illegible]257
Duration	32 mins, 6 seconds
Executed by	root
Arguments	-hardwaretype X4-2
Collection Date	10-Mar-2022 00:57:[illegible]4

- There are 6 flagged critical checks, 1[illegible] flagged failed checks , 7 flagged warning checks, 17 flagged info checks. By defa
- This version of EXAchk is considered[illegible]alid for 48 days from today or until a new version is available

Exadata Critical Issues

The following Exadata Critical Issues (MOS Note 1270094.1) have been checked in this report:

- Exadata Storage Server : EX1-EX65,EX67,EX69,EX70
- Database Server : DB1-DB4, DB6, DB9-DB49
- InfiniBand switch : IB1-IB3,IB5-IB9

古いEXAchk

Cluster Summary

Cluster Name	Cluster-c1
OS/Kernel Version	LINUX X86-64 OELRHEL 7 4.14.35-2047.505.4.4.el7uek.x86_64
CRS Home - Version	/u01/app/21.0.0.0/grid - 21.3.0.0.0
DB Home - Version - Names	/u01/app/oracle/product/21.0.0.0/dbhome_1 - 21.3.0.0.0 - cdbm213 database /u01/app/oracle/product/19.0.0.0/dbhome_1 - 19.12.0.0.0 - cdbm19c database /u01/app/oracle/product/18.0.0.0/dbhome_1 - 18.14.0.0.0 - cdbm18c database /u01/app/oracle/product/12.2.0.1/dbhome_1 - 12.2.0.1.210720 - cdbm122 database /u01/app/oracle/product/12.1.0.2/dbhome_1 - 12.1.0.2.210720 - 3 databases
Exadata Version	21.2.4.0.0
Number of nodes	8
Database Servers	2
Storage Servers	3
IB Switches	3
EXAchk Version	21.4.2_20220211
Collection	exachk_random01client01_rac12c_030922_151642
Duration	30 mins, 48 seconds
Executed by	root
Arguments	-hardwaretype X4-2
Collection Date	09-Mar-2022 15:22:38

- There are 5 flagged critical checks, 19 flagged failed checks , 6 flagged warning checks, 18 flagged info checks. By defa
- This version of EXAchk is considered valid for 154 days from today or until a new version is available

Exadata Critical Issues

The following Exadata Critical Issues (MOS Note 1270094.1) have been checked in this report:

- Exadata Storage Server : EX1-EX65,EX67,EX69,EX70,EX71,EX72
- Database Server : DB1-DB4, DB6, DB9-DB49
- InfiniBand switch : IB1-IB3,IB5-IB9

Exadata：ライフサイクル管理
EXAchk：よく見受けられる問題点

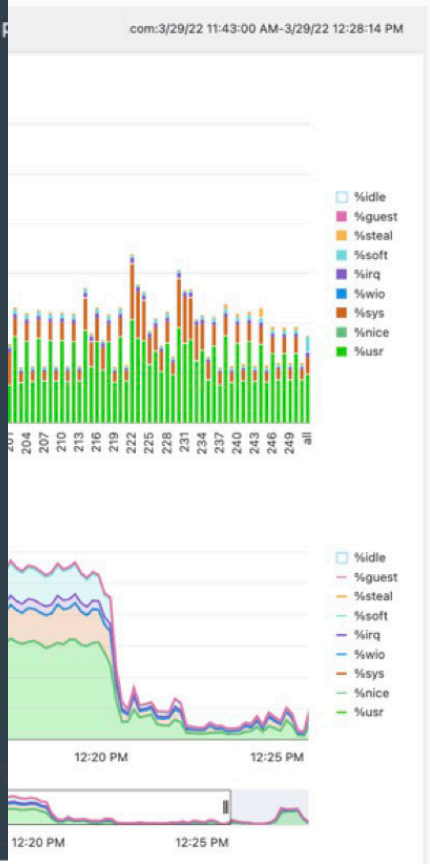
- HugePagesが正しく設定されていないケース
  - 最近のDBリリースでは、SGA > 32 GB以上の構成がHugePages構成なしで使用されているかどうかを確認
  - その場合はインスタンスは起動しない
- 冗長性の推奨事項に従っていないケース
- すでに修正済みの重大な問題（Critical issue）への対応が実施されていないケース

Exadata：ライフサイクル管理

Oracle ExaWatcher：グラフ化

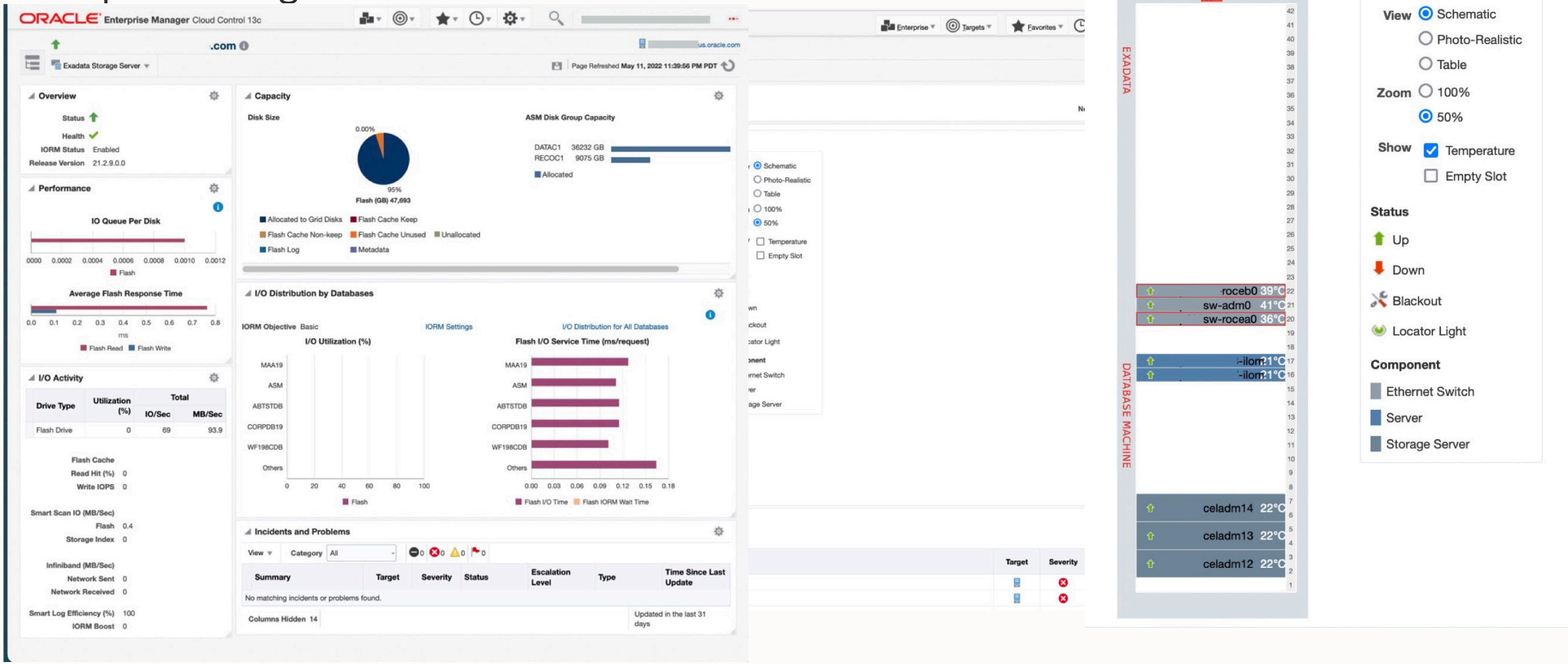
```
   zzz <08/26/2021 01:07:37> subcount: 1
top - 01:07:39 up 23 days, 11:34,  0 users,  load average: 7.01, 7.94, 8.74
Threads: 14011 total,   5 running, 11133 sleeping,   0 stopped,   0 zombie
%Cpu(s):  4.8 us,  2.1 sy,  0.0 ni, 92.6 id,  0.0 wa,  0.2 hi,  0.3 si,  0.0 st
KiB Mem : 15834616+total, 22833467+free, 10630640+used, 29206297+buff/cache
KiB Swap: 25165820 total, 25158640 free,     7180 used. 45395289+avail Mem

    PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+ COMMAND
  31661 splunk    20   0  185704  12144   7236 R 99.9  0.0   0:07.23 python /opt/splunkforwarder/etc/apps/bk_elp_update_read_perm/bin/generate_monitored_list.p
  35530 exawatch  20   0  177024  19592   4188 R 80.0  0.0   0:00.72 /usr/bin/top -b -c -H -n 1 -w512
  35561 exawatch  20   0  169680  12436   4160 R 40.0  0.0   0:00.32 /usr/bin/top -b -c -n 1 -w512
      1 root      20   0  217384  11300   6244 R 20.0  0.0   7669:00 /usr/lib/systemd/systemd --switched-root --system --deserialize 22
 257794 oracle    20   0   19.0g 156488 139184 S  7.5  0.0   6:13.84 oracle          (LOCAL=NO)
 137634 grid      20   0 3658556 104120  98748 S  5.0  0.0 156:28.75 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 229236 grid      20   0 3658556 110608 105236 S  5.0  0.0 156:45.62 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
  35565 oracle    20   0 7050676  69780  65252 S  3.8  0.0   0:00.03 ora_c
 110869 grid      20   0 3654816  80352  76448 S  3.8  0.0  33:29.37 asm_pmon_+ASM4
 137609 grid      20   0 3658560 113088 107716 S  3.8  0.0 157:26.15 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 137647 grid      20   0 3659408 112392 106304 S  3.8  0.0 158:49.41 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 172236 grid      20   0 3658556 107148 101720 S  3.8  0.0  10:25.29 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 172284 grid      20   0 3658560 103324  97956 S  3.8  0.0  10:20.94 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 178226 grid      20   0 3658564 104116  98748 S  3.8  0.0  10:20.00 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 192223 grid      20   0 3668324 113380 107376 S  3.8  0.0  14:48.76 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 225979 grid      20   0 3658560 118500 113124 S  3.8  0.0 163:44.29 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 226001 grid      20   0 3658556 105780 100404 S  3.8  0.0 163:30.96 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 226633 grid      20   0 3658560 111036 105680 S  3.8  0.0 164:25.67 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 227118 grid      20   0 3659412 113076 106964 S  3.8  0.0 161:17.46 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 227320 grid      20   0 3658564 114312 108940 S  3.8  0.0 162:27.96 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 227628 grid      20   0 3658560 118404 113028 S  3.8  0.0 162:31.65 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 228001 grid      20   0 3658560 112884 107512 S  3.8  0.0 164:01.06 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 228157 grid      20   0 3658564 106080 100704 S  3.8  0.0 163:24.08 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 229276 grid      20   0 3658564 103220  97840 S  3.8  0.0 158:57.30 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 230200 grid      20   0 3658564 117016 111640 S  3.8  0.0 165:29.22 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 230209 grid      20   0 3658564 116896 111520 S  3.8  0.0 165:27.66 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 249654 grid      20   0 3668324 110712 104708 S  3.8  0.0  57:58.89 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 253216 grid      20   0 3669152 109952 102936 S  3.8  0.0  10:11.55 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 254595 grid      20   0 3669160 114024 107272 S  3.8  0.0 159:07.87 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 255128 grid      20   0 3658556 109576 104200 S  3.8  0.0 159:17.33 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 257510 grid      20   0 3658556 104528  99160 S  3.8  0.0  10:06.37 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 257528 grid      20   0 3659388 105640  99552 S  3.8  0.0  10:01.25 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 257568 grid      20   0 3659408 104836  98740 S  3.8  0.0   9:55.44 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 264924 grid      20   0 3669156 120184 113128 S  3.8  0.0 156:47.16 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
 307188 grid      20   0 3668328 109736 103732 S  3.8  0.0  12:38.59 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
   2185 grid      20   0 3668332 108372 102356 S  2.5  0.0  12:04.51 oracle+ASM4_a             (DESCRIPTION=(LOCAL=YES)(ADDRESS=(PROTOCOL=beq)))
```



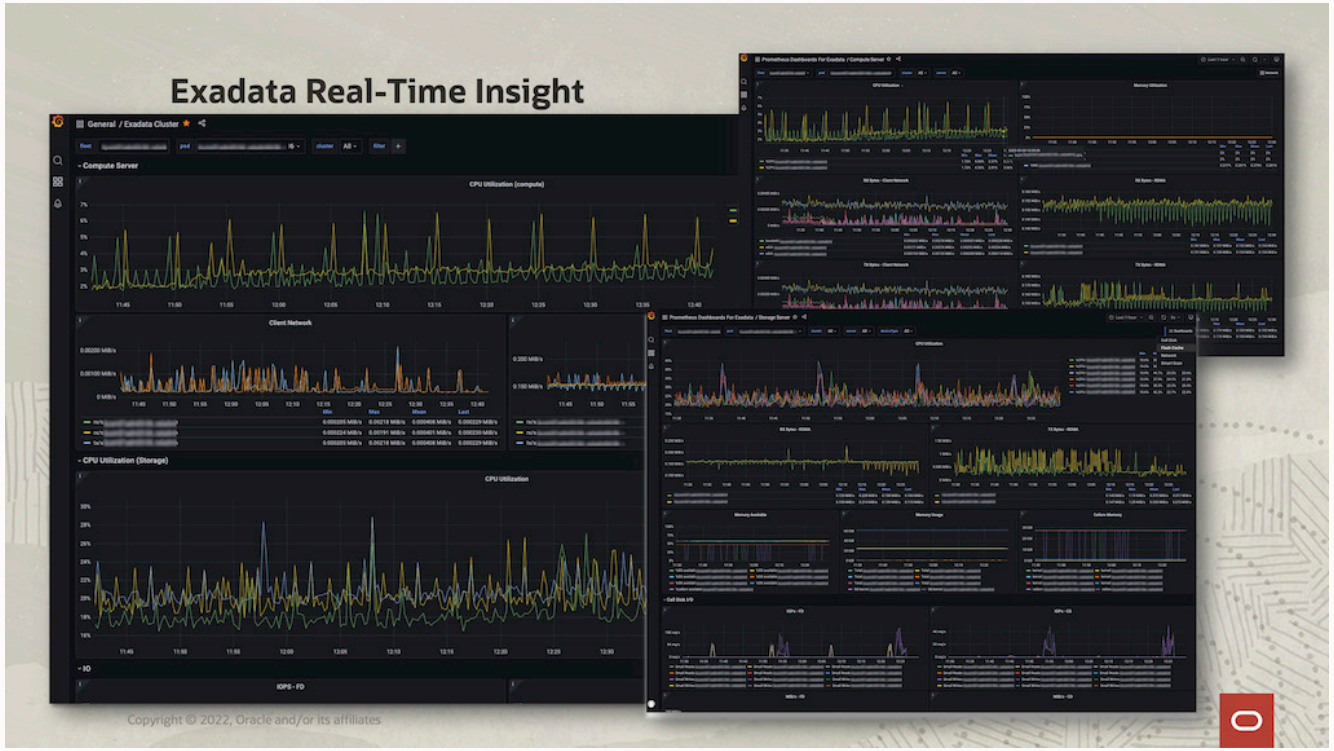
Exadata：ライフサイクル管理

Enterprise Manager 13cで監視



Exadata：ライフサイクル管理
Exadata Real-Time Insight

- Exadataフリートのすべてのサーバーから最新メトリックの所見を自動的にストリーム
- カスタマイズ可能な監視ダッシュボードに、リアルタイム分析と問題解決策のデータをフィード
- 包括的：Exadataソフトウェアおよびハードウェアの200を超えるメトリック
- プロアクティブな問題検出とリアルタイムの意思決定が可能に



Exadata：ライフサイクル管理
Exadata Real-Time Insight

- Exadataフリートのすべてのサーバーから最新メトリックを自動的にストリーム
- カスタマイズ可能な監視ダッシュボードに、リアルタイム分析と問題解決策のデータをフィード

- **包括的**
  - 200を超えるExadataソフトウェアおよびハードウェアのメトリック
  - 1秒ごとの頻度での、きめ細かいメトリックの収集が可能

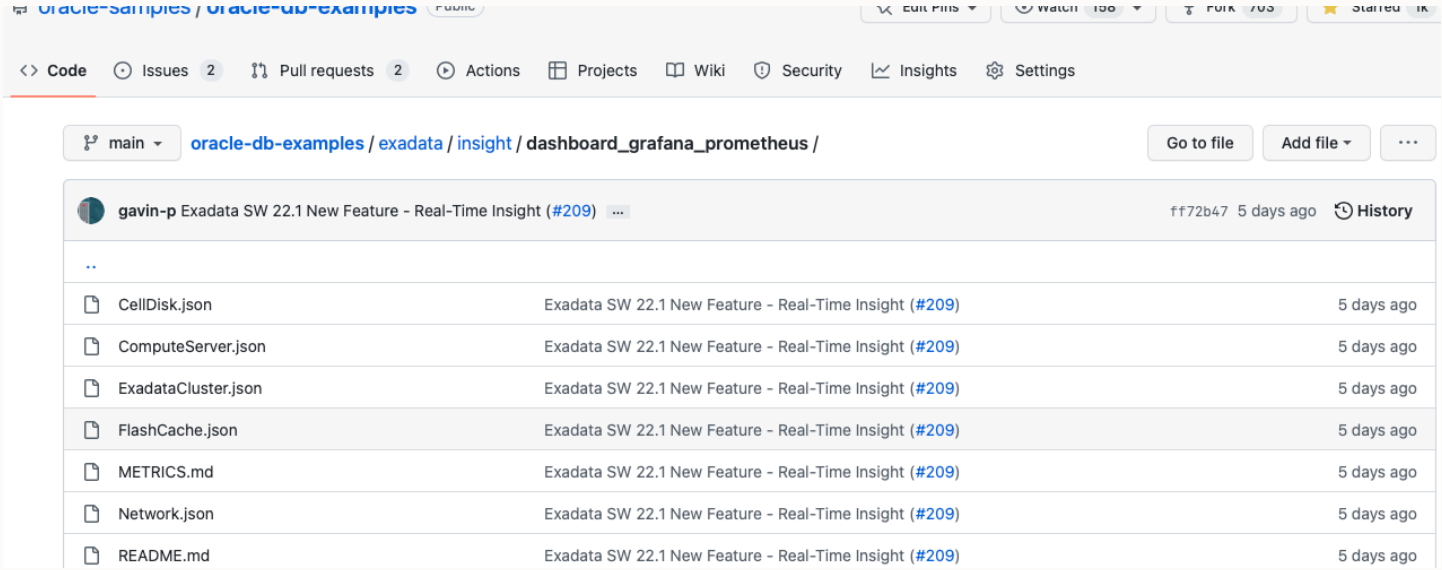
- **統合的**
  - 一般的な時系列/可観測性のプラットフォームとの統合
  - ユーザー定義のエンドポイントにきめ細かいメトリックをリアルタイムでストリーム

- **優れたインサイト**
  - プロアクティブな問題検出とリアルタイムの意思決定が可能

https://blogs.oracle.com/exadata/post/real-time-insight-quick-start

Exadata：ライフサイクル管理
Exadata Real-Time Insight - サンプル・ダッシュボード・コード

- GitHub.comのOracle Samplesリポジトリには、Real-Time Insightのダッシュボードのサンプルがあります。
  - 次のダッシュボードのコードが含まれています（Grafana/Prometheus）。
    - Exadataクラスタ
    - コンピューティング
    - ストレージ・サーバー
    - セル・ディスク
    - フラッシュ・キャッシュ
    - スマート・スキャン
    - ネットワーク



- https://github.com/oracle-samples/oracle-db-examples/tree/main/exadata/insight

Exadata：ライフサイクル管理

Exadata Real-Time Insight - サンプル・ダッシュボード・コード

- **Exadata Cluster：**コンピューティング・ノードとストレージ・サーバーのメトリックを表示する、クラスタ全体のビューを提供
- **Compute：**コンピューティング・ノードのCPUおよびネットワーク使用率を表示するコンピューティング・ノード・ビューを提供
- **Storage Server：**ストレージ・サーバーのCPUとI/Oのメトリック、およびSmart Flash Cache、Smart Flash Log、Smart I/OのExadataメトリックに焦点を当てた、ストレージ・サーバーを中心としたビューを提供
  - **Cell Disk：**ストレージ・サーバーのセル・ディスクI/Oのメトリックを表示
  - **Flash Cache：**ストレージ・サーバーのフラッシュ・キャッシュのメトリックを表示
  - **Smart Scan：**ストレージ・サーバーのSmart Scanメトリックを表示

Exadata：ライフサイクル管理

Exadata AWRサポート

Exadata Configuration and Statistics

- Exadata Report Summary
- Exadata Server Configuration
- Exadata Server Health Report
- Exadata Statistics

Back to Top

Exadata Server Configuration

- Exadata Storage Server Model
- Exadata Storage Server Version
- Exadata Storage Information
- Exadata Griddisks
- Exadata Celldisks
- ASM Diskgroups
- IORM Objective

Exadata Storage Server Model

- Model Information of Servers
- CPU Count refers to logical CPUs, including cores and hyperthreads

Model	CPU Count	Memory (GB)	# Cells
Oracle Corporation ORACLE SERVER X9-2L High Capacity	96/96	252	12

Back to Exadata Server Configuration

Exadata Storage Server Version

- Version information of packages on the storage server

Package Type	Package Version	Cells
Kernel	4.14.35-2047.511.5.5.el7uek.x86_64	All (12)
Cell	cell-21.2.10.0.0_LINUX.X64_220317-1.x86_64	All (12)
Offload	cellofl-11.2.3.3.1_LINUX.X64_200526	All (12)
Offload	cellofl-21.2.10.0.0_LINUX.X64_220317	All (12)
Offload	cellofl-12.1.2.4.0_LINUX.X64_210708.1	All (12)

Back to Exadata Server Configuration

Exadata Storage Information

- Storage information per cell
- 'Total' is the sum for all cells

	Size (GB)			# Celldisks				
# Cells	Flash Cache	PMEM Cache	Flash Log	Hard Disk	Flash	PMEM	# Griddisks	Cell Name
11	23,845.81	1,500.56	0.50	12	4	12	24	(11):
1	23,845.81	1,500.56	0.38	12	4	12	24	(1):
Total (12)	286,149.75	18,006.75	5.88	144	48	144	288	All (12)

Back to Exadata Server Configuration

Exadata Griddisks

- Griddisks on the storage servervs
- Disk Type <F|H|M>/<size>: F-Flash, H-Harddisk, M-persistent Memory
- Size (GB) - Griddisk: indicates size of individual Griddisks in the cells
- Size (GB) - Cell Total: indicates total size per cell
- Size (GB) - System Total: indicates total size over all cells

Outlier Summary - Disk Level

- Outliers are disks whose average performance is outside the normal range, where normal range is -/+ 3 standard deviation
- Outlier disks must have a mininum of 10 IOPs. Idle disks are not considered for outlier analysis.
- Outliers for hard disks are displayed when Hard Disk IOPs exceeds 306.72 (1% of maximum capacity of 30,672)
- Outliers for flash disks are displayed when Flash Disk IOPs exceeds 95260.8 (1% of maximum capacity of 9,526,080)
- Outliers for flash disks will not be displayed. There are only 86,566 flash IOPs
- A 'v' and a dark yellow background indicates an outlier value below the low range
- A '^' and a light red background indicates an outlier value above the high range
- A '*' and a dark red background indicates over maximum capacity
- Disk Type <F|H|M>/<size>: F-Flash, H-Hard Disk, M-pMEM; PMEM I/O only include remote I/Os processed by cellsrv
- Maximum hard disk capacity: IOPS: H/16.0T: 213 | IO MB/s: H/16.0T: 148
- Maximum flash disk capacity: IOPS: F/5.8T: 198,460 | IO MB/s: F/5.8T: 11,250

Disk Name	Cell Name	Disk Type	Statistic Name	Value	Mean	Std Dev	Normal Range
CD_01_		H/16.0T	% Disk Utilization	^ 2.83	0.34	0.61	0.00 - 2.16
CD_02_		H/16.0T	Small Reads/s	^ 39.95	3.37	7.43	0.00 - 25.67
CD_06_		H/16.0T	Small Reads/s	^ 39.44	3.37	7.43	0.00 - 25.67
CD_08_		H/16.0T	Small Reads/s	^ 39.24	3.37	7.43	0.00 - 25.67
CD_08_		H/16.0T	Small Reads/s	^ 39.39	3.37	7.43	0.00 - 25.67
CD_10_		H/16.0T	Small Reads/s	^ 39.83	3.37	7.43	0.00 - 25.67
CD_10_		H/16.0T	% Disk Utilization	^ 2.45	0.34	0.61	0.00 - 2.16

Back to Exadata Outlier Summary
Back to Exadata Resource Statistics

Exadata：ライフサイクル管理

ストレージ・サーバー・アラート用のカスタム診断パッケージ

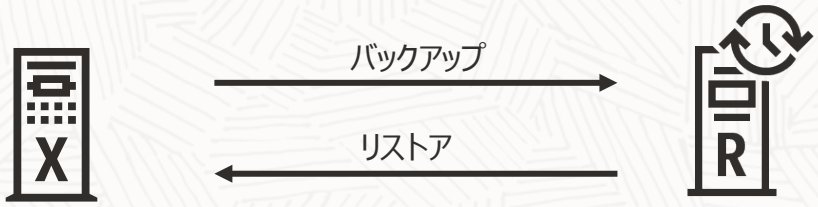
Exadata：ライフサイクル管理

バックアップ

- データベースのバックアップ☺
  - Oracle RMANを搭載したZDLRAまたはZFSアプライアンスを推奨

- KVM HOSTとKVM GUESTをバックアップ
  - 詳細については、プレゼンテーションの最後を参照

- バックアップをテスト



シュレーディンガーのバックアップ：

バックアップの状態はリストアが試行されるまで分からない

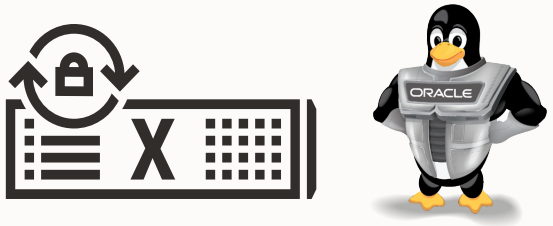
出典：Fabienne Fierens

Exadata Live Update

セキュリティを向上させてデータベース・サーバーとVMの再起動を最低限に抑制

Exadata System Softwareは、Exadataデータベース・サーバーと、
Oracle Databaseの最適でセキュアな運用にとって重要な、
オペレーティング・システム、ファームウェア、およびExadataソフトウェアの更新を提供

更新は、データベース・サーバー全体にローリング形式で適用



Exadataデータベース・サーバー

Exadata Live Update は、更新をオンラインで適用し、
残りの作業を遅延させてスケジュールされた時刻に実施することが可能に

Exadata Live Update は、RPMやkspliceなどの馴染みのあるLinuxテクノロジーを
使用し、更新をオンラインでデータベース・サーバー/VMに適用するため、
再起動は不要

Exadata Live Update のオプション

Exadata Live Update には、共通脆弱性評価システム（CVSS）に基づく複数のオプションがある
Exadata Live Update を使用する場合は、以下のオプションから選択

highcvss	セキュリティ更新のみを適用し、CVSSスコアが7以上の脆弱性に対処
allcvss	セキュリティ更新のみを適用し、あらゆるCVSSスコアの脆弱性に対処
full	セキュリティ関連のあらゆる更新とセキュリティ以外の他のあらゆる更新を含む、完全更新を実行 サーバー/VMの再起動によって適用される定期更新と同等

```
$ patchmgr   --dbnodes kvm_guests.lst --upgrade  --repo <repo.zip location> --rolling ¥
             --target_version 24.1.0.0.0.240517 --live-update-target highcvss|allcvss|full
```

未処理の作業の表示

すべての更新内容がオンラインで適用できるとは限らず、再起動せずにアクティブ化できるとも限らない

- 例：ファームウェア、最新のカーネルによる起動、JDK

これらの更新は‘outstanding work（未処理の作業）’と呼ばれ、次の正常停止時の適用のためにステージングされる

未処理の項目を表示するには、patchmgr --live-update-list-outstanding-workを使用

```
$ patchmgr --dbnodes kvm_guests.lst --live-update-list-outstanding-work
***
Summary of outstanding work for Exadata Live Update:
exdpm1adm01vm01.example.com: (*) 2024-08-15 00:17:08: Exadata Live Update outstanding work is
scheduled for completion at the next reboot
                                - The Linux kernel will be updated from version 5.4.17-
                                  2136.330.7.5.el8uek to 5.4.17-2136.333.5.1.el8uek.
                                  Current Uptrack kernel version: 5.4.17-2136.333.5.1.el8uek.x86_64
                                - New package uptrack-updates-5.4.17-2136.333.5.1.el8uek.x86_64
                                  (version 20240725-0) will be installed.
```

未処理の作業の適用

デフォルトでは、未処理の作業は次回の正常停止時に適用される

管理者は、patchmgr --live-update-schedule-outstanding-workを使用して以下を実行可能

- 再起動ウィンドウを指定する - "YYYY-MM-DD HH24:MM:SS TZ"

```
$ patchmgr --dbnodes kvm_guests.lst --live-update-schedule-outstanding-work ¥
                                    ”2024-11-04 22:00:00 AEDT”
```

- 未処理の作業の適用を遅延させる – ‘never’

```
$ patchmgr --dbnodes kvm_guests.lst --live-update-schedule-outstanding-work never
```

- 以前に設定したスケジュールをデフォルトの動作にリセット

```
$ patchmgr --dbnodes kvm_guests.lst --live-update-schedule-outstanding-work reset
```

オラクルは、少なくとも3か月に一度は未処理の作業を適用することを推奨

Exadata Live Update のベスト・プラクティス

データベース・サーバー/VMのバックアップ

- Patchmgrにより、すべての更新時にシステム・バックアップが自動的に作成されるため、必要に応じて迅速なロールバックが可能
- 管理者が管理するバックアップを追加することを推奨

正常な再起動

- vm_maker --stop_domain/--start_domain操作、ホストの再起動（shutdown –r）、サーバーの電源ボタンの短押しなどを含む
- 物理データベース・サーバーを再起動することにより、VMも再起動
  - VMと物理サーバーの再起動を整合させるのに便利（必須ではない）
- 未処理の作業（outstanding work）を適用している間は、VMと物理サーバーのリセットを回避

アプリケーションやユーザーに計画再起動を認識されないようにするには、
透過的アプリケーション・コンティニュイティなどのDatabase MAAの機能を使用

Exadata Live Update

月次メンテナンス・リリースの適用 - 例

四半期ごとの更新ウィンドウ（推奨）

8月	9月	10月	11月
• 24.1.3 • 完全更新 • サーバー/VMの再起動	• 24.1.4 • Exadata Live Update • 再起動なし	• 24.1.5 • Exadata Live Update • 再起動なし	• 24.1.6 • 完全更新 • サーバー/VMの再起動
12月	**1月**	**2月**	**3月**
• 24.1.7 • Exadata Live Update • 再起動なし	• 24.1.8 • Exadata Live Update • 再起動なし	• 24.1.9 • 完全更新 • サーバー/VMの再起動	• 24.1.10 • Exadata Live Update • 再起動なし

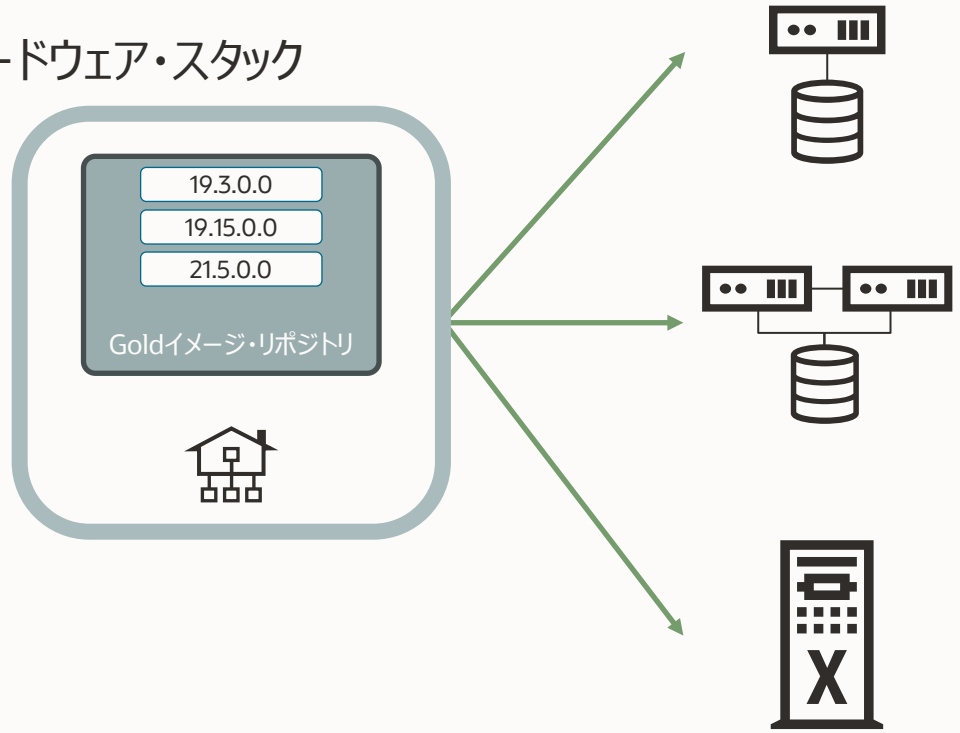
半年ごとの更新ウィンドウ

8月	9月	10月	11月
• 24.1.3 • 完全更新 • サーバー/VMの再起動	• 24.1.4 • Exadata Live Update • 再起動なし	• 24.1.5 • Exadata Live Update • 再起動なし	• 24.1.6 • Exadata Live Update • 再起動なし
12月	**1月**	**2月**	**3月**
• 24.1.7 • Exadata Live Update • 再起動なし	• 24.1.8 • Exadata Live Update • 再起動なし	• 24.1.9 • 完全更新 • サーバー/VMの再起動	• 24.1.10 • Exadata Live Update • 再起動なし

Exadata：ライフサイクル管理
計画メンテナンス

Exadata `patchmgr`ユーティリティを使用することで、以下のハードウェア・スタック全体にパッチを適用可能：

- ストレージ・セル
- RoCEスイッチ
- 管理スイッチ
- ベアメタルとKVM HOST
- KVM GUEST



Oracle Fleet Patching and Provisioning：
アウトオブプレース・パッチ適用のためのツール

- データベース・ホーム
- Grid Infrastructure、およびGIとDBの組合せへのパッチ適用
- Exadataへもパッチ適用
- https://www.oracle.com/jp/database/technologies/rac/fpp.html
- 1つのツールでOracle DBスタック全体のパッチ適用とアップグレードに対処

Exadata：参考資料

バックアップ

- https://www.oracle.com/jp/a/tech/docs/recovery-appliance-maint-practices-ja.pdf

KVMの仮想化

- https://www.oracle.com/jp/a/tech/docs/exadata-kvm-overview-ja.pdf

ライフサイクル管理

- https://www.oracle.com/jp/a/tech/docs/exadata-software-maintenance-2022-ja.pdf

セキュリティ

- https://www.oracle.com/jp/a/tech/docs/exadata-maximum-security-architecture-ja.pdf

Exadataのリアルタイムのインサイト

- https://blogs.oracle.com/oracle4engineer/post/exadata-real-time-insight-jp

参考資料
有用なリソース

Exadata製品管理ブログ - https://blogs.oracle.com/exadata/
MOS Note Referenceブログ - https://blogs.oracle.com/oracle4engineer/post/exadata-mos-notes-jp

Exadata Database Machine and Exadata Storage Server Supported Versions（Doc ID 888828.1）
Oracle Exadata Database MachineEXAchk（Doc ID 1070954.1）
『Oracle Exadata Best Practices』（Doc ID 757552.1）
『 Exadata Critical Issues 』(Doc ID 1270094.1)
『Exadata Patching Overview and Patch Testing Guidelines』（Doc ID 1262380.1）
『The ASM Priority Rebalance feature - An Example』（Doc ID 1968607.1）
『Physical and Logical Block Corruptions. All you wanted to know about it.』（Doc ID 840978.1）
『Best Practices for Corruption Detection, Prevention, and Automatic Repair - in a Data Guard Configuration』（Doc ID 1302539.1）
『Understanding ASM Capacity and Reservation of Free Space in Exadata』（Doc ID 1551288.1）

Exadata MAA：結論

岩のように強固

出典：Zoltan Tasi https://unsplash.com/photos/QxjEi8Fs9Hg

この世のものとは思えないパフォーマンス

出典：Space X https://unsplash.com/photos/OHOU-5UVIYQ

ありがとうございました

Our mission is to help people see data in new ways, discover insights, unlock endless possibilities.


ORACLE