ORACLE

Exadata のパフォーマンスと AWR

AWR による Exadata のパフォーマンス診断

March, 2024, Version 2.0

ORACLE

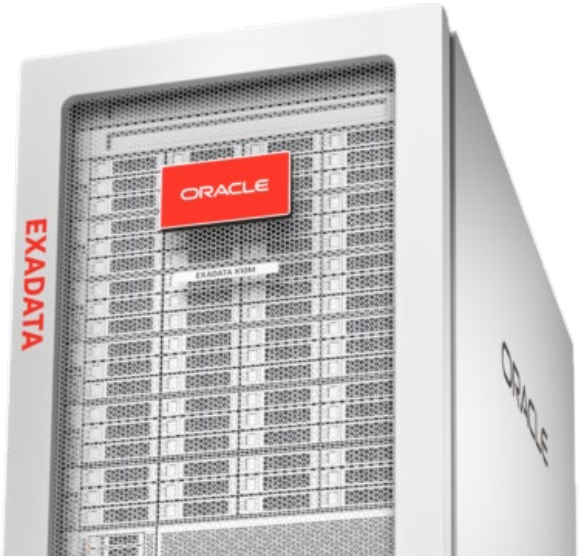
免責事項

本文書には、ソフトウェアや印刷物など、いかなる形式のものも含め、オラクルの独占的な所有物である占有情報が含まれます。この機密文書へのアクセスと使用は、締結および遵守に同意した Oracle Software License and Service Agreement の諸条件に従うものとします。本文書と本文書に含まれる情報は、オラクルの事前の書面による同意なしに、公開、複製、再作成、またはオラクルの外部に配布することはできません。本文書は、ライセンス契約の一部ではありません。また、オラクル、オラクルの子会社または関連会社との契約に組み込むことはできません。

本書は情報提供のみを目的としており、記載した製品機能の実装およびアップグレードの計画を支援することのみを意図しています。マテリアルやコード、機能の提供をコミットメント（確約）するものではなく、購買を決定する際の判断材料になさらないでください。本文書に記載されている機能の開発、リリース、時期および価格については、弊社の裁量により決定されます。製品アーキテクチャの性質上、本書に記述されているすべての機能を安全に組み込むことができず、コードの不安定化という深刻なリスクを伴う場合があります。

ORACLE

内容



はじめに

Oracle Exadata は、Oracle データベースのパフォーマンス、コスト効率、および可用性が劇的に向上するように設計されています。Exadata は、スケールアウト型の高パフォーマンス・データベース・サーバー、最先端のフラッシュ・ドライブを搭載したスケールアウト型のインテリジェント・ストレージ・サーバー、超高速 Remote Direct Memory Access（RDMA）over Converged Ethernet（RoCE）内蔵ファブリックを備えた、最新のクラウドベース・アーキテクチャです。Exadata 独自のソフトウェア・アルゴリズムによって、ストレージ、コンピューティング、RoCE ネットワーキングにおけるデータベース・インテリジェンスを可能にすることで、他のプラットフォームよりも低コストで高パフォーマンスと大容量を実現しています。

Exadata では、オンライン・トランザクション処理（OLTP）、データウェアハウス（DW）、インメモリ分析、複合ワークロードの統合など、あらゆるタイプのデータベース・ワークロードを実行できます。Exadata は、プライベート・データベース・クラウドの基盤としてオンプレミスでデプロイするか、またはオラクルがすべてのインフラストラクチャを管理する Exadata Database Service on Dedicated Infrastructure（ExaDB-D）または Exadata Database Service on Cloud@Customer（ExaDB-C@C）を備えた Oracle Cloud Infrastructure（OCI）でサブスクリプション・モデルを使用して入手できます。

世界中の顧客が Exadata をエンタープライズ・データベース・デプロイメントのプラットフォームに選択し、Exadata システムに統合されるデータベースの数が増え続けているため、データベースのパフォーマンスを Exadata システムの観点から監視することがこれまで以上に重要になっています。この技術概要では、Oracle Database 自動ワークロード・リポジトリ（AWR）機能と Exadata を併用してデータベースのパフォーマンス特性を Exadata の観点から監視および分析する方法について説明します。

この技術概要の内容は、デプロイ先がオンプレミスか、ExaDB-D か、ExaDB-C@C かに関係なく、すべての Exadata デプロイメントに適用されます。特に、OCI の Exadata の場合はデータベースのすべての管理権限をお客様が握っているため、データベースがオンプレミスにデプロイされている場合と同様に Exadata 固有の AWR 機能を使用できます。

AWR の概要

自動ワークロード・リポジトリ（AWR）は Oracle Database 10g で導入された機能で、Oracle Database 向けパフォーマンス診断ツールとしてもっとも広く使用されています。AWR では、問題の検出や自己チューニングを目的として、データベースのパフォーマンス統計データを収集、処理、管理します。このデータ収集プロセスは一定の間隔で繰り返され、結果は AWR スナップショットに取得されます。AWR スナップショットに取得されたデータから計算される差分値は、この期間中の各統計の変化を表すもので、AWR レポートで詳しく分析することができます。デフォルトでは、1 時間間隔で AWR スナップショットが取得され、8 日間保存されます。月次（31 日間）または四半期ごと（90 日間）に比較できるように要件に応じて保存期間を延長することが推奨されます。AWR レポートは期間を指定してオンデマンドで生成することもできます。[1]

パフォーマンスと対象範囲

パフォーマンスの問題を分析するときに重要なのは、パフォーマンスの問題の対象範囲を把握し、分析に使用されるデータとツールが必ず問題の対象範囲と一致するようにすることです。

たとえば、問題がごく一部のユーザーや SQL 文に局所化されている場合、問題の対象範囲に関係するデータは SQL Monitor レポートに含まれます。SQL Monitor レポートでは、SQL 文またはデータベース（DB）操作の 1 回の実行に関する詳細な統計を確認できます。

パフォーマンスの問題がインスタンス全体またはデータベース全体のものである場合、AWR レポートには該当するインスタンスまたはデータベース全体のデータと統計が含まれます。アクティブ・セッションをサンプリングするアクティブ・セッション履歴（ASH）は、インスタンス全体の問題、データベース全体の問題、および局所的な問題に使用できます。ASH では、データのフィルタリングに使用できる複数のディメンションを横断してデータが収集されます。

ベースラインの保守

統計ベースラインは、通常、システムが良好に動作している場合に一定の間隔で取得される統計のコレクションです。ベースラインを使用することで、ベースラインで取得された統計とパフォーマンスの低下中に取得された統計を比較して、パフォーマンス上の問題を診断できます。これにより、問題の原因となり得る、大幅に増加した統計を特定できます。

ベースラインは、月末や年末の処理などの重要な時間枠に加えて、通常の処理期間中にも収集することが推奨されます。ベースラインには、ストレージ・サーバーからの追加の統計（ExaWatcher およびセル・メトリック履歴）とともに、AWR データ [2]と、鍵となるいくつかの SQL 文の SQL Monitor レポートが含まれる必要があります [3]。

AWR での Exadata のサポート

AWR で Exadata がサポートされるようになったのは Oracle Database 12.1.0.2.0 および Exadata System Software 12.1.2.1.0 からです。Exadata の統計が AWR レポートに含まれるようになったため、ストレージ・サーバーからさらにデータを収集しなくても、1 つのレポートでストレージ層まで観察できるようになりました。ストレージ・サーバーにアクセスすることができない ExaDB-D および ExaDB-C@C のお客様はこの点に特に目を引かれるでしょう。

Exadata の統計は、HTML 形式とアクティブ HTML 形式の AWR インスタンス・レポート、および CDB$ROOT からの AWR グローバル・レポートでしか提供されません。Exadata の統計は、テキスト形式のレポートでも、PDB レベルの AWR レポートでも表示することはできません。レポート内の Exadata のセクションは、Exadata ソフトウェアの新しいリリースに新機能が組み込まれるのに合わせて、絶えず強化されています。[4]Exadata の統計は Enterprise Manager の AWR レポートでも表示できます。Enterprise Manager で Exadata を管理する方法について説明しているドキュメントのリストを、この技術概要の参考資料のセクションで示します。

---

[1]AWR について詳しくは、*Oracle Database パフォーマンス・チューニング・ガイド*の"データベース統計の収集"を参照してください。

[2]ベースラインには、AWR レポートではなく、実際の AWR データが含まれる必要があります。

[3] *Oracle Exadata System Software – Exadata の監視*には、AWR、ExaWatcher、およびセル・メトリック履歴に関する広範な情報が含まれます。

[4]Exadata のセクションは絶えず拡張されているため、システム上に表示されるバージョンがこの技術概要のスクリーンショットに一致しない場合があります。

他にも重要な点として、Exadata ストレージ・レベルの統計が AWR レポートに追加されましたが、パフォーマンス・チューニングの手法に変更はありません。まずは DB 時間を調査し、DB 時間を多く消費しているものを分析してパフォーマンス上の問題に対応します。Exadata のセクションの調査を開始するのは、IO に問題がありそうだと判断された場合のみです。Exadata のセクションは既存のツールや手法に置き換わるものではなく、補完するものです。

課題および AWR Exadata ソリューション

Oracle データベース管理者（DBA）が多く直面する課題は、サーバー、ネットワーク、ストレージといった基盤インフラストラクチャに直接関係しているデータベース・パフォーマンス特性をよりよく分析し、理解することです。インフラストラクチャの構成が最適であれば、最適なデータベース・パフォーマンスが得られます。ただし、インフラストラクチャの構成が誤っていたり、コンポーネントに欠陥があったりすると、その結果として引き起こされるデータベース・パフォーマンスの問題を正確に診断し、特定のコンポーネントに関連付けるのは容易ではありません。

Exadata のようなエンジニアド・システムによってもたらされる価値は、Exadata ストレージ・サーバー上で収集および管理される統計情報を Oracle DBA が直接かつ自動的に AWR に統合できるようになったことです。このホワイト・ペーパーで後ほど説明しますが、汎用インフラストラクチャ上のデータベースで費やされたであろう時間やリソースと比較して、この診断プロセスは驚くほど効率的です。また、Exadata のコア・プラットフォームがソフトウェア機能やハードウェア機能によって強化されるのに伴い、継続的に Exadata 固有の AWR コンテンツも継続的に強化され続けているという事実も、Oracle の DBA にとって有益です。

次のセクションからは、Exadata 固有の AWR 機能を活用できると考えられる具体的なシナリオについて概説します。

統合環境

Exadata ストレージにおいて、パフォーマンスの問題を分析する際に新たな視点が追加されます。ストレージ・サブシステムは複数のデータベースで共有されることがあり、そのためストレージ・レイヤーから得られる統計情報はシステム全体を対象としたものとなります。つまり、単一のデータベースや単一のデータベース・インスタンスに限定される統計情報ではないということです。

複数のデータベースを実行している Exadata システムでは、特定のデータベースがシステム上の IO 帯域幅を大量に消費することで他のデータベースに影響を与えている可能性を分析することが重要です。そのような場合は、Exadata に組み込まれている IO リソース管理（IORM）機能を利用して、Exadata Storage Server 内の IO リクエストに優先順位を付け、構成されたリソース・プランに基づいてスケジューリングすることを強くお勧めします。IORM の詳細については、*Oracle Exadata System Software ユーザーズ・ガイドの"IO リソース管理の理解"*を参照してください。

AWR レポートには **Top Databases**（上位データベース）セクション [5]が含まれており、IO リクエスト数と IO スループットが表示されます。このセクションは、各データベースの IO リソースの消費量を比較するのに役立ちます。各ストレージ・サーバー内の上位 N 個のデータベースを特定する内部メトリックに基づいて、一部のデータベースが AWR スナップショットに記録され、Exadata の AWR レポートには、IO リクエスト数と IO スループットが上位のデータベースが表示されます。図 1 に示されているように、データはフラッシュ・デバイス上の IO とハード・ディスク上の IO とに分けて表示されます。

図 2 では、リクエストはさらに IO サイズが小さい IO（Small Requests）と大きい IO（Large Requests）に分類され、平均 IO 待機時間と IO の平均 IORM キュー時間と合わせて表示されます。

図 1 で、レポートには全体に占める割合（%Total）ではなく、記録されたリクエストに占める割合（%Captured）が表示されている点に注意してください。これは、すべてのデータベースのすべての統計が AWR によって記録されるわけではないためです。このデータは、システム全体を集計したものと、ストレージ・セルごとに集計したものを使用できます。この AWR レポートを作成したデータベース DB03 は(*)でマークされていますが、AWR に記録された全 IOPS の 5 パーセントにすぎません。

[5]セキュリティが懸念される場合のために、dbPerfDataSuppress というセル属性があります。これを使用すると、他のデータベースの v$cell_db ビューと、そのデータを取得する AWR ビューにデータベース名が表示されないようにすることができます。別のデータベースからこのビューが問い合わせられた場合、dbPerfDataSuppress にリストされているデータベースの IO は "OTHER（その他）"に含まれることになります。セル属性のリスト表示、変更、記述については、『Oracle Exadata System Software ユーザーズ・ガイド』を参照してください。

図 2 の DB03 データベースでは、フラッシュとディスクの両方の Large Requests をキューイングしていることが示されています。これは、このデータベースからの IO の優先順位を下げる IORM 計画が使用されていることを示している可能性があります。

図 1：Top Databases by IO Requests

Top Databases by IO Requests

- The top 10 databases by IO Requests are displayed
- (*) indicates current database. Current database is always displayed.
- %Captured - % of Captured DB IO requests
- Total - total IO requests or IO throughput (Flash + Disk)
- Ordered by IO requests desc

		IO Requests					IO Throughput (MB)			
DB Name	DBID	%Captured	Total Requests	per Sec	Flash	Disk	Total MB	per Sec	Flash	Disk
******DB04	4217808068	44.59	15,229,908,120	423,229.35	15,030,322,122	199,585,998	1,140,328,136.10	31,688.99	1,030,499,101.46	109,829,034.64
******DB01	3501400968	26.98	9,214,731,876	256,071.47	9,083,420,779	131,311,097	715,829,149.22	19,892.43	633,912,640.09	81,916,509.13
OTHER	0	10.67	3,642,893,165	101,233.66	1,427,978,205	2,214,914,960	47,852,147.47	1,329.78	12,578,541.59	35,273,605.88
******DB02	2997371584	9.63	3,290,534,395	91,441.83	3,217,862,665	72,671,730	278,967,324.25	7,752.32	237,516,435.72	41,450,888.53
******DB03 (*)	3870370343	5.50	1,879,755,241	52,237.19	1,839,253,212	40,502,029	137,229,537.08	3,813.52	125,398,032.13	11,831,504.96
******DB06	3515888175	0.91	311,289,885	8,650.55	296,762,258	14,527,627	22,571,684.45	627.25	20,252,601.63	2,319,082.83
******DB07	2692616685	0.50	170,130,618	4,727.82	161,754,795	8,375,823	11,898,324.80	330.65	10,251,724.97	1,646,599.83
******DB05	1430994058	0.48	165,389,820	4,596.08	153,195,142	12,194,678	11,037,913.32	306.74	8,801,559.60	2,236,353.72
******DB08	3444312789	0.48	165,384,170	4,595.92	156,631,881	8,752,289	11,320,581.74	314.59	10,009,049.81	1,311,531.93
ASM	1	0.18	62,016,763	1,723.41	54,873,516	7,143,247	2,750,631.22	76.44	1,670,361.48	1,080,269.74

図 2：Top Databases by Requests - Details

Top Databases By Requests - Details

- Request details for the top databases by IO requests

			Small Requests							Large Requests						
			Reqs/s			Latency		Queue Time		Reqs/s			Latency		Queue Time	
DB Name	DBID	IOs/s	Total	Flash	Disk	Flash	Disk	Flash	Disk	Total	Flash	Disk	Flash	Disk	Flash	Disk
******DB04	4217808068	423,229.35	39,784.17	37,511.95	2,272.22	164.90us	909.87us		41.06us	383,445.18	380,171.03	3,274.15	1.10ms	8.59ms	1.72ms	1.57ms
******DB01	3501400968	256,071.47	21,523.61	20,501.67	1,021.94	158.87us	1.92ms		118.64us	234,547.86	231,920.75	2,627.11	931.35us	6.36ms	0.95ms	558.97us
OTHER	0	101,233.66	100,583.85	39,036.83	61,547.02	109.58us	144.75us		1.46ms	649.80	645.76	4.04	332.24us	3.17ms	125.70us	65.36us
******DB02	2997371584	91,441.83	5,530.82	4,780.39	750.43	182.92us	1.60ms		45.52us	85,911.01	84,641.94	1,269.07	0.96ms	5.87ms	569.59us	375.39us
******DB03(*)	3870370343	52,237.19	4,567.22	3,785.53	781.69	172.45us	2.34ms		1.65ms	47,669.97	47,326.13	343.84	1.04ms	10.32ms	1.27ms	2.33ms
******DB06	3515888175	8,650.55	837.99	505.88	332.11	157.96us	583.88us	71.00ns	151.25us	7,812.56	7,740.95	71.61	920.63us	2.86ms	386.81us	218.26us
******DB07	2692616685	4,727.82	1,347.88	1,160.43	187.46	133.92us	1.77ms		123.15us	3,379.94	3,334.64	45.30	0.95ms	4.04ms	1.72ms	355.54us
******DB05	1430994058	4,596.08	1,395.30	1,129.98	265.33	193.08us	4.46ms		3.31ms	3,200.77	3,127.22	73.56	704.66us	7.55ms	90.97us	3.10ms
******DB08	3444312789	4,595.92	821.24	614.14	207.10	125.58us	538.22us		27.62us	3,774.68	3,738.56	36.12	648.42us	1.75ms	189.29us	67.62us
ASM	1	1,723.41	1,071.05	906.42	164.62	171.28us	2.11ms		17.14us	652.36	618.48	33.88	482.33us	303.78us	54.19us	4.46us

セルまたはディスクへの負荷の偏り

Exadata は、すべてのストレージ・サーバーとディスクに負荷を均等に分散するように設計されています。ストレージ・サーバーやディスクが他と比較して過剰な負荷を処理している場合、パフォーマンスの問題を引き起こす可能性があります。

Exadata AWR レポートでは、いくつかのメトリックを使用してデバイス同士を比較する外れ値(Outlier)分析を実行します。デバイスはタイプ別とサイズ別にグループ化して比較されます。デバイス・タイプが異なればパフォーマンス特性も異なるためです。たとえば、フラッシュ・デバイスのパフォーマンスはハード・ディスクとは大きく異なるはずです。同様に、1.6 TB のフラッシュ・デバイスでは 6.4 TB のフラッシュ・デバイスと同じ量の IO を維持できないでしょう。

外れ値分析に使用される統計には IOPS、スループット、使用率、サービス時間、キュー時間などを含む iostat のような OS 統計が含まれます [6]。ストレージ・サーバーの統計も外れ値分析に含まれ、IOPS、スループット、レイテンシーが IO のタイプ（読取りまたは書込み）と IO のサイズ（小または大）別に分けられています。

Exadata AWR レポートでは、システムが処理能力の限界に達したかどうかが特定されます。レポートで使用される最大値は Exadata のデータ・シートで公開されているものと同じです。お客様のワークロードは多種多様であるため、ここでの最大値は厳格なルールというよりもガイドラインとして使用されることを意図しています。

[6]OS 統計の一部として報告されるキュー時間は、デバイスのキュー時間であり、IORM のキュー時間ではありません。

自動ハード・ディスク・スクラブと修復（スクラブ）は、ディスク上のセクターを事前に検査する Exadata の機能であり、内蔵ハード・ドライブ機能では検出できない問題を検出することができます。スクラブは Exadata の自動化プロセスで、データベース・パフォーマンスに影響しないようにディスクがアイドル状態（25 %未満のビジー状態）の場合に開始され、デフォルトでは隔週にスケジューリングされています。[7]

通常、Exadata のスクラブが実行されている間は、読取り数は最大 IOPS を超過します。スクラブは、16 KB のシーケンシャル・リードを実行します。Exadata ソフトウェアは、スクラブ IO よりもクライアント IO を優先するように設計されています。クライアント IO が発行されると、スクラブ IO はクライアント IO の処理を優先するために抑制されるので、スクラブはクライアント IO に影響しないと見込まれます。

[7] https://blogs.oracle.com/oracle4engineer/post/exadata-disk-scrubbing-jp

図 3 は、ストレージ・セルの外れ値分析の例を示したものです。この例に外れ値（他と大きく異なる値）はありませんが、ハード・ディスクの IOPS 性能が最大に達している可能性があることが特定され、*と濃い赤色の背景で示されています。このシステムのハード・ディスクの最大性能は 6,408 IOPS ですが、レポートの現在の表示は 9,355.83 IOPS となっています。

図 4 は、別システムの AWR レポートのディスクの外れ値分析の例を示したものです。この例では、ハード・ディスク性能が最大に達していることが特定されています。また、他のディスクよりも多くの IOPS を実行している 2 つのディスクも特定されています。

図 3：Exadata OS IO Statistics - Outlier Cells

Exadata OS IO Statistics - Outlier Cells

- These statistics are collected by the OS on the cells and are not restricted to this database or instance
- Outliers are cells whose average performance is outside the normal range, where normal range is -/+ 1 standard deviation
- Outlier cells must have a mininum of 10 IOPs. Idle cells are not considered for outlier analysis.
- Outliers for hard disks are displayed when Hard Disk IOPs exceeds 64.08 (1% of maximum capacity of 6,408)
- Outliers for flash disks are displayed when Flash Disk IOPs exceeds 27599.88 (1% of maximum capacity of 2,759,988)
- 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
- %Total - Avg [IOPs |IO MB/s] of the cell as a percentage of total [IOPs | IO MB/s] for the disk type

				IOPs						IO MB/s						% Disk Utilization		
				Total	% Total	Per Cell	Per Disk			Total	% Total	Per Cell	Per Disk					
Disk Type	Cell Name	# Cells	# Disks			Average	Mean	Std Dev	Normal Range			Average	Mean	Std Dev	Normal Range	Mean	Std Dev	Normal Range
F/1.5T	All	3	12	31,953.78		10,651.26	2,662.81	542.19	2,120.62 - 3,205.01	630.98		210.33	52.58	12.69	39.89 - 65.27	16.36	4.24	12.12 - 20.60
H/3.6T	All	3	36	9,355.83		* 3,118.61	* 259.88	78.58	181.31 - 338.46	471.03		157.01	13.08	9.76	3.32 - 22.84	13.24	12.23	1.01 - 25.47

IOPs					
Total	% Total	Per Cell	Per Disk		
		Average	Mean	Std Dev	Normal Range
31,953.78		10,651.26	2,662.81	542.19	2,120.62 - 3,205.01
9,355.83		* 3,118.61	* 259.88	78.58	181.31 - 338.46

図 4：Exadata OS IO Statistics - Outlier Disks

Exadata OS IO Statistics - Outlier Disks

- These statistics are collected by the OS on the cells and are not restricted to this database or instance
- 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 231.6 (1% of maximum capacity of 23,160)
- Outliers for flash disks are displayed when Flash Disk IOPs exceeds 37500 (1% of maximum capacity of 3,750,000)
- 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
- % Total - Avg [IOPs |IO MB/s] of the disk as a percentage of total [IOPs | IO MB/s] for the disk type

				IOPs				IO MB/s				% Disk Utilization		
Disk Type	Cell Name	Disk Name	# Disks	% Total	Mean	Std Dev	Normal Range	% Total	Mean	Std Dev	Normal Range	Mean	Std Dev	Normal Range
F/2.9T	All	All	40		1,682.23	1,600.22	0.00 - 6,482.90		39.15	36.14	0.00 - 147.59	6.39	6.24	0.00 - 25.10
H/7.2T	All	All	120		* 213.08	38.57	97.38 - 328.79		120.15	48.70	0.00 - 266.26	70.32	24.21	0.00 - 142.95
Outlier	***celadm04	CD_06_***celadm04		1.39	* 354.58			0.74	107.41			58.68		
Outlier	***celadm06	CD_07_***celadm06		1.33	* 340.73			0.72	104.35			57.25		

			IOPs			
Cell Name	Disk Name	# Disks	% Total	Mean	Std Dev	Normal Range
All	All	40		1,682.23	1,600.22	0.00 - 6,482.90
All	All	120		* 213.08	38.57	97.38 - 328.79
***celadm04	CD_06_***celadm04		1.39	* 354.58		
***celadm06	CD_07_***celadm06		1.33	* 340.73		

構成の相違

ストレージ・サーバー同士で構成が異なっている場合は、パフォーマンスの問題が発生する可能性があります。構成の問題としては、Smart Flash Cache（フラッシュ・キャッシュ）や Smart Flash Log（フラッシュ・ログ）のサイズの相違、または使用されているセル・ディスク数またはグリッド・ディスク数の相違が考えられます。

AWR レポートには Exadata の構成情報が含まれており、構成が異なるストレージ・サーバーが特定されます。図 5 は、ストレージ・サーバー構成が同一であるシステムの例を示したものです。Exadata Server Configuration セクションで、'All'はすべてのストレージ・サーバーの構成が同一であることを示しています。図 6 の **Exadata Storage Server Model** セクションに示すように、セル構成に相違があればセルの名前が表示されます。

図 5：ストレージ・サーバー構成が同一のシステム

Exadata Storage Server Model

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

Model	CPU Count	Memory (GB)	# Cells	Cells
Oracle Corporation ORACLE SERVER X9-2L High Capacity	96/96	252	12	***celadm01, ***celadm02, ***celadm03, ***celadm04, ***celadm05, ***celadm06, ***celadm07, ***celadm08, ***celadm09, ***celadm10, ***celadm11, ***celadm12

Back to Exadata Server Configuration

Exadata Storage Server Version

- Version information of packages on the storage server

Package Type	Package Version	Cells
Kernel	4.14.35-2047.518.4.3.el7uek.x86_64	All (12)
Cell	cell-22.1.13.0.0_LINUX.X64_230818-1.x86_64	All (12)
Offload	cellofl-11.2.3.3.1_LINUX.X64_220513	All (12)
Offload	cellofl-12.1.2.4.0_LINUX.X64_230109	All (12)
Offload	cellofl-22.1.13.0.0_LINUX.X64_230818	All (12)

図 6：ストレージ・サーバー構成が異なるシステム

Exadata Storage Server Model

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

Model	CPU Count	Memory (GB)	# Cells	Cells
Oracle Corporation ORACLE SERVER X7-2L High Capacity	40/40	188	30	***r1celadm01, ***r1celadm02, ***r1celadm03, ***r1celadm04, ***r1celadm05, ***r1celadm06, ***r1celadm07, ***r2celadm01, ***r2celadm02, ***r2celadm03, ***r2celadm04, ***r2celadm06, ***r2celadm07, ***r2celadm08, ***r2celadm09, ***r3celadm01, ***r3celadm02, ***r3celadm03, ***r3celadm04, ***r3celadm06, ***r3celadm07, ***r3celadm08, ***r3celadm09, ***r4celadm01, ***r4celadm02, ***r4celadm03, ***r4celadm04, ***r4celadm06, ***r4celadm07, ***r4celadm08
Oracle Corporation ORACLE SERVER X8-2L High Capacity	64/64	188	3	***r2celadm05, ***r3celadm05, ***r4celadm05

Back to Exadata Server Configuration

Exadata Storage Server Version

- Version information of packages on the storage server

Package Type	Package Version	Cells
Kernel	4.14.35-2047.516.2.4.el7uek.x86_64	All (33)
Cell	cell-21.2.18.0.0_LINUX.X64_221111.1-1.x86_64	All (33)
Offload	cellofl-11.2.3.3.1_LINUX.X64_220513	All (33)
Offload	cellofl-12.1.2.4.0_LINUX.X64_220712	All (33)
Offload	cellofl-21.2.18.0.0_LINUX.X64_221111.1	All (33)

高負荷

システム負荷の増加が原因でパフォーマンスが変化することがあります。考えられる原因としては、ストレージ・サーバー上の IO または CPU 負荷の増加があります。IO 負荷の増加は、バックアップなどのメンテナンス作業、またはユーザーのワークロードの増加や SQL 実行計画の変更などによるユーザーIO の変化によって引き起こされることがあります。

Exadata システムでは、データベースが IO を実行する理由を示す追加情報が、各 IO と共にストレージ・サーバーへ送信されます。IO Reasons を確認することで、IO 負荷の増加の原因がメンテナンス作業なのか、それともデータベース・ワークロードの増加なのかを簡単に判断できます。

レポートには、Smart Scan、Smart Flash Log、Smart Flash Cache といった Exadata *Smart* 機能に関する情報も表示されます。

図 7 は、典型的なのデータベース・ワークロードによって上位の IO リクエスト（redo log write や buffer cache reads）が発生している例を示しています。

図 7：Top IO Reasons by Requests

Top IO Reasons by Requests

- The top IO reasons by requests per cell are displayed
- Only reasons with over 1% of IO requests for each cell are displayed
- At most 5 reasons are displayed per cell
- %Cell - the percentage of IO requests on the cell due to the IO reason
- Ordered by Cell Name, Requests Value desc

		Requests			MB	
Cell Name	IO Reason	%Cell	Total Requests	per Sec	Total MB	per Sec
celadm01	redo log write	34.04	33,008,655	4,580.72	320,986.18	44.54
	buffer cache reads	17.02	16,499,555	2,289.70	505,945.44	70.21
	database control file read	13.08	12,679,567	1,759.58	212,912.14	29.55
	dbwr media recovery writes	9.48	9,194,222	1,275.91	116,868.18	16.22
	aged writes by dbwr	6.17	5,985,270	830.60	87,165.43	12.10
celadm02	redo log write	31.40	32,973,741	4,575.87	320,897.67	44.53
	database control file read	18.95	19,901,980	2,761.86	328,248.14	45.55
	buffer cache reads	15.99	16,798,150	2,331.13	494,659.73	68.65
	dbwr media recovery writes	8.56	8,994,422	1,248.19	113,998.26	15.82
	aged writes by dbwr	5.68	5,960,292	827.13	86,947.17	12.07
celadm03	redo log write	35.02	33,067,661	4,588.91	319,872.01	44.39
	buffer cache reads	16.98	16,028,690	2,224.35	497,268.40	69.01
	database control file read	10.43	9,848,423	1,366.70	168,741.53	23.42
	dbwr media recovery writes	9.76	9,214,635	1,278.74	116,003.46	16.10
	aged writes by dbwr	6.28	5,929,270	822.82	86,725.01	12.04

DB Time と待機イベント (Wait Event)

多くの場合、ストレージに関連するパフォーマンス上の問題によって、IO 関連の待機イベントでの DB Time が増加します。データベースには、実行中の IO の種類を示すさまざまな待機イベントがあります。ストレージに関連する問題が発生すると、これらの待機イベントの平均待機時間が長くなり、それに伴い、DB Time の内でこれらの待機イベントに費やされた割合も AWR レポートで確認できます。

前節で説明した AWR Exadata セクションと組み合わせることで、データベース待機イベントと Exadata セクションの特定の統計を関連付けて、ストレージ・サーバー上で IO がどのように処理されているかを判断することができます。

多くの場合、IO レイテンシーの遅延は、フラッシュ・キャッシュや XRMEM Cache を利用せずに、ハード・ディスクで処理する IO 数が増加した結果です。そのため、IO が Exadata キャッシュで処理されていない理由を特定することが重要となります。

Exadata Statistics - Performance Summary と範囲

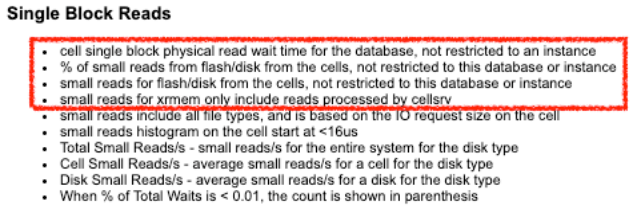
AWR レポートの Exadata セクションを確認する場合、表示されている統計の範囲を示す説明に留意してください。Performance Summary には、データベース統計とストレージ統計の両方が含まれます。これにより、両者の関連付けが容易になります。ただし、データベース統計には AWR レポートの生成元である単一のデータベースのデータが表示される一方、ストレージ統計にはストレージ・サーバー上で収集されたデータが表示され、それらのサーバー上で実行されているすべてのデータベースが含まれます。

シングル・ブロック読取り (Single Block Reads)

図 8 に示すように、シングル・ブロック読取りに関連する待機イベントは、ストレージの IO パフォーマンスを的確に表しています。データベース上のシングル・ブロック読取りは、ストレージ・サーバー上の small read に相当します。待機イベント名から読取りが発生したメディアを判別することも可能です。

多くの場合、シングル・ブロック読取りは、OLTP システムの主要な IO 待機イベントです。cell single block physical reads 待機イベントの待機時間が長い場合、ストレージ関連の問題が潜んでいる可能性があります。

図 8：統計の範囲



- データベース全体の cell single block physical read の待機時間であり、個別のインスタンスに限定されません。
- セル側のフラッシュ / ディスクから読んだ Small Reads の割合(%)は、個別のデータベースやインスタンスに限定されません。
- セル側のフラッシュ / ディスクから読んだ Small Reads の回数も、個別のデータベースやインスタンスに限定されません。
- XRMEM の Small Reads の回数には、ストレージ・サーバー内の cellsrv プロセスが処理した読取りのみが含まれます。

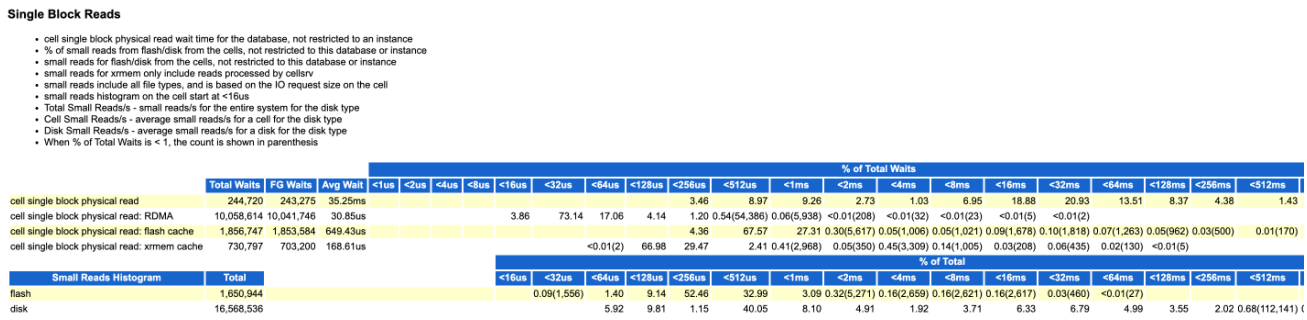
表 1 は、cell single block physical read 待機イベントの種類を示したものです。各シングル・ブロック読取りの平均応答時間は、読取り元のメディアによって大きく異なります。

図 9 は、シングル・ブロック読取りの待機数がおもに RDMA から生じているシステムを示しています。ただし、RDMA からの応答時間は非常に短いため、通常は待機数が多くなると全体の DB 時間は短縮されます。

表 1：cell single block physical read 待機イベントの種類

待機イベント	説明
cell single block physical read: RDMA	RDMA を使用して XRMEM Cache からの読取りを行うシングル・ブロック読取りの待機イベント
cell single block physical read: xrmem cache	XRMEM Cache からのシングル・ブロック読取りの待機イベント
cell single block physical read: flash cache	フラッシュ・キャッシュからのシングル・ブロック読取りの待機イベント
cell single block physical read	ディスク、または容量に最適化されたフラッシュからのシングル・ブロック読取りの待機イベント

図 9：Single Block Reads



Single Block Reads

- cell single block physical read wait time for the database, not restricted to an instance
- % of small reads from flash/disk from the cells, not restricted to this database or instance
- small reads for flash/disk from the cells, not restricted to this database or instance
- small reads for xrmem only include reads processed by cellsrv
- small reads include all file types, and is based on the IO request size on the cell
- small reads histogram on the cell start at <16us
- Total Small Reads/s - small reads/s for the entire system for the disk type
- Cell Small Reads/s - average small reads/s for a cell for the disk type
- Disk Small Reads/s - average small reads/s for a disk for the disk type
- When % of Total Waits is < 1, the count is shown in parenthesis

	Total Waits	FG Waits	Avg Wait	% of Total Waits <1us	<2us	<4us	<8us	<16us	<32us	<64us	<128us	<256us	<512us	<1ms	<2ms	<4ms	<8ms	<16ms	<32ms	<64ms	<128ms	<256ms	<512ms
cell single block physical read	244,720	243,275	35.25ms									3.46	8.97	9.26	2.73	1.03	6.95	18.88	20.93	13.51	8.37	4.38	1.43
cell single block physical read: RDMA	10,058,614	10,041,746	30.85us					3.86	73.14	17.06	4.14	1.20	0.54(54,386)	0.06(5,938)	<0.01(208)	<0.01(32)	<0.01(23)	<0.01(5)	<0.01(2)				
cell single block physical read: flash cache	1,856,747	1,853,584	649.43us									4.36	67.57	27.31	0.30(5,617)	0.05(1,006)	0.05(1,021)	0.09(1,678)	0.10(1,818)	0.07(1,263)	0.05(962)	0.03(500)	0.01(170)
cell single block physical read: xrmem cache	730,797	703,200	168.61us							<0.01(2)	66.98	29.47	2.41	0.41(2,968)	0.05(350)	0.45(3,309)	0.14(1,005)	0.03(208)	0.06(435)	0.02(130)	<0.01(5)		

Small Reads Histogram	Total	% of Total <16us	<32us	<64us	<128us	<256us	<512us	<1ms	<2ms	<4ms	<8ms	<16ms	<32ms	<64ms	<128ms	<256ms	<512ms
flash	1,650,944		0.09(1,556)	1.40	9.14	52.46	32.99	3.09	0.32(5,271)	0.16(2,659)	0.16(2,621)	0.16(2,617)	0.03(460)	<0.01(27)			
disk	16,568,536			5.92	9.81	1.15	40.05	8.10	4.91	1.92	3.71	6.33	6.79	4.99	3.55	2.02	0.68(112,141)

Single Block Reads

- cell single block physical read wait time for the database, not restricted to an instance
- % of small reads from flash/disk from the cells, not restricted to this database or instance
- small reads for flash/disk from the cells, not restricted to this database or instance
- small reads for xrmem only include reads processed by cellsrv
- small reads include all file types, and is based on the IO request size on the cell
- small reads histogram on the cell start at <16us
- Total Small Reads/s - small reads/s for the entire system for the disk type
- Cell Small Reads/s - average small reads/s for a cell for the disk type
- Disk Small Reads/s - average small reads/s for a disk for the disk type
- When % of Total Waits is < 1, the count is shown in parenthesis

	Total Waits	FG Waits	Avg Wait	<1us	<2us
cell single block physical read	244,720	243,275	35.25ms		
cell single block physical read: RDMA	10,058,614	10,041,746	30.85us		
cell single block physical read: flash cache	1,856,747	1,853,584	649.43us		
cell single block physical read: xrmem cache	730,797	703,200	168.61us		

Small Reads Histogram	Total
flash	1,650,944
disk	16,568,536

表 2 は、ディスク読取りのコストを示したものです。cell single block physical read 関連の待機イベントの総回数は 12,890,878 で、これらの待機イベントに費やされた合計待機時間は 9,544.72 秒です。ディスクに対して発生している cell single block physical read 待機回数は 2 パーセントにすぎませんが、合計待機時間の 82 パーセント超を占めています。このため、ハード・ディスクで多数の読取りを処理しなければならない場合、明らかなパフォーマンスの問題が発生することがありえます。

表 2：ディスク読取りのコスト

待機イベント	Total DB time （待機回数 * 平均待機時間）	待機数の割合（%）	待機時間の割合（%）
cell single block physical read	7,892.22 秒（244,720 * 32.25 ミリ秒）	1.9	82.7
cell single block physical read: RDMA	310.30 秒（10,058,614 * 30.85 マイクロ秒）	78.0	3.2
cell single block physical read: flash cache	1,205.83 秒（1,856,747 * 649.43 マイクロ秒）	14.4	12.6
cell single block physical read: xrmem cache	136.37 秒（730,797 * 168.61 マイクロ秒）	5.7	1.4

Exadata Statistics - Performance Summary セクションには、small reads がどのように処理されたかを示すセクションが含まれます。図 10 は、データベース IO の 30.73 パーセントがフラッシュ・キャッシュで処理され、71.97 パーセントが XRMEM cache で処理されていることを示しています（53.59 パーセントは RDMA 読取り経由で実行）。これは、ほとんどの読取りが Exadata キャッシュで処理され、パフォーマンスの良いデータベースです [8]。

データベースが XRMEM cache を使用して RDMA 読取りを実行する場合、読取りリクエストはストレージ・サーバーに送信されません。代わりに、データベースはストレージ・サーバー上の XRMEM キャッシュから読取りを直接行うことで、非常に高速な読取りレイテンシーを実現します。この場合、ストレージ・サーバーはデータベースによって実行された RDMA 読取りをカウントできないので、Exadata セクションでもこれらの RDMA 読取りは計上されません。代わりに、RDMA 読取りはデータベース統計から取得されます。

Exadata XRMEM cache セクションは、RDMA 経由で送信されなかったデータベース読取りリクエストを反映します。この場合、読取りリクエストはストレージ・サーバーに送信され、ストレージ・サーバーが読取りリクエストを処理します。これにより、XRMEM cache のヒットまたはミスのいずれかが生じます。ほとんどの読取りリクエストが XRMEM cache で処理される場合、図 10 に示すように、Flash Cache Hit%が低下する可能性があります。これは単純にフラッシュ・キャッシュに対する読取りリクエスト数が少ないことが原因である可能性があるため、必ずしも懸念する必要はありません。このパターンは通常、キャッシュへの読取りが必要な新しいデータがアクセスされていることを示しています。

図 10：Performance Summary – Cache Savings

Cache Savings

- Disk write savings (overwrites) - writes absorbed by flash cache that would have otherwise gone to disk
- Database Flash Cache Hit% - percentage of database reads from all instances satisfied from Flash Cache
- Database XRMEM Cache Hit% - percentage of database reads from all instances satisfied from XRMEM Cache
- Database XRMEM Cache RDMA Hit% - percentage of database reads from all instances satisfied from XRMEM Cache, including RDMA reads
- Cell Flash Cache Hit% - percentage of cell reads satisfied from Flash Cache
- Cell XRMEM Cache Hit% - percentage of cell reads satisfied from XRMEM Cache

Database Flash Cache Hit%	30.73
Database XRMEM Cache Hit%	71.97
Database XRMEM Cache RDMA Hit%	53.59
Cell Flash Cache OLTP Hit%	37.86
Cell Flash Cache Scan Hit%	75.70
Cell XRMEM Cache Hit%	75.33
Disk Write savings/s	94,307.31
Large Writes/s	1,099.87

[8] キャッシュ・ヒットの総パーセンテージが 100 パーセントを超える場合がありますが、これはキャッシュ・ヒット（分子）としてカウントされても、物理読取り IO リクエスト（分母）としてカウントされないタイプの読取りがあるためです。制御ファイルの読取りはカウントされない一つの例です。

パフォーマンスの問題は、Exadata キャッシュで処理されない過剰なディスク IO が原因となっている場合があるため、図 11 にあるように、**Exadata Statistics - Performance Summary – Disk Activity** のセクションでもディスク IO の潜在的な原因が示されています。

図 11：Performance Summary – Disk Activity

Disk Activity

- The following are possible causes of disk IO
- Smart Scan (estd) are estimated as 1MB per IO request
- Redo log writes to disk are calculated using redo write requests and redo writes absorbed by flash cache. Total redo write requests are in parenthesis.

I/O per second	Total	per Cell
Redo log writes	19,636.61 (19,636.61)	1,402.62 (1,402.62)
Smart Scans (estd)	197.82	14.13
Flash Cache misses (OLTP)	4,603.85	328.85
Flash Cache read skips	2,120.78	151.48
Flash Cache write skips	20,644.36	1,474.60
Flash Cache LW rejections (total)	5,696.43	406.89
Disk writer writes	1,443.92	103.14

図 12 に示すように、セル・ストレージ・サーバーごとのフラッシュ・キャッシュに対する OLTP 読取りリクエストは、1 秒あたりわずか 200 ほどです。これには、Exadata システム上で実行されているすべてのデータベースが含まれます。フラッシュ・キャッシュからの読取りリクエスト率がかなり低く、XRMEM cache からのヒット率が高いことを考慮すると、フラッシュ・キャッシュのヒット率が低いことがこの特定のデータベースのパフォーマンスに与える影響は最小限であると思われます。しかし、データベースのキャッシュ・ヒット率が低いことと、フラッシュ・キャッシュからのミス率が高いことが同時に確認できる場合、読取りがディスクで処理されている可能性があるため、詳細な調査が必要となります。

図 12：Flash Cache User Reads Per Second

Flash Cache User Reads Per Second

- These statistics are collected by the cells and are not restricted to this database or instance
- Total - total number of reads per second from Flash Cache
- OLTP/Scan/Columnar reads include reads on keep objects
- Ordered by Total Hit Read Requests per Second desc

	Read Requests per Second						Read MB per Second				
Cell Name	Total Hits	OLTP	Scan	Columnar	Keep	Misses	Total Hits	OLTP	Scan	Columnar	Keep
Total (14)	6,853.29	2,805.51	4,047.71		0.06	4,603.85	4,183.83	184.23	3,999.60		0.00
***celadm01	624.38	259.33	365.05		0.00	289.53	377.49	16.88	360.61		0.00
***celadm06	556.75	228.27	328.47		0.00	303.60	338.80	14.85	323.95		0.00
***celadm04	499.28	202.46	296.82		0.01	331.17	306.06	13.26	292.80		0.00
***celadm13	493.95	195.48	298.46		0.01	330.42	307.92	12.89	295.02		0.00
***celadm05	492.40	198.86	293.54		0.01	322.51	303.07	13.10	289.98		0.00
***celadm10	490.06	197.61	292.45		0.00	328.03	302.26	12.90	289.36		0.00
***celadm03	485.90	203.02	282.87		0.01	345.95	292.91	13.35	279.56		0.00
***celadm02	483.84	199.00	284.84		0.00	343.28	294.97	13.09	281.89		0.00
***celadm14	476.50	194.72	281.78		0.00	318.70	291.57	12.89	278.68		0.00
***celadm07	473.98	196.83	277.14		0.01	334.39	286.00	12.81	273.19		0.00
***celadm12	472.67	192.20	280.47		0.01	330.68	290.18	12.78	277.40		0.00
***celadm11	467.64	194.42	273.22		0.00	319.17	283.22	12.71	270.51		0.00
***celadm09	434.94	176.11	258.82		0.00	344.58	267.41	11.60	255.80		0.00
***celadm08	400.99	167.21	233.77		0.00	361.85	241.96	11.11	230.85		0.00

フラッシュ・キャッシュ・ミスのほかに、フラッシュ・キャッシュのスキップがディスクからのシングル・ブロック読取りが行われる原因となる場合があります。フラッシュ・キャッシュのスキップは、読取りがフラッシュ・キャッシュでのキャッシングの対象外であるとマークされているためにフラッシュ・キャッシュを使用せずにバイパスする場合に生じます。図 13 のレポートでは、ストレージ句の設定に依存してフラッシュ・キャッシュをバイパスしている（フラッシュ・キャッシュのスキップが発生している）ことを示しており、これはセグメントの cell_flash_cache 属性を NONE を設定するストレージ句が指定されていることを意味します。

図 13：Flash Cache User Reads - Skips

Flash Cache User Reads - Skips

- These statistics are collected by the cells and are not restricted to this database or instance
- Flash Cache User Read Skips are reads that bypass the flash cache
- Total Skipped includes all reads that have bypassed flash cache
- Only the following possible reasons for bypassing the flash cache are displayed:
- Storage Clause - flash cache skipped due to storage clause
- IOReason - flash cache skipped due to IO reason sent by the database
- GridDisk Policy - flash cache skipped due to griddisk caching policy
- Large IO - flash cache skipped due to size of IO
- Throttle IO - flash cache skipped due to throttling
- Throttle Large IO - flash cache skipped due to exceeding limit for outstanding large IOs

	Requests Skipped		Read Requests Skipped per Second					
Cell Name	Total	per Second	Storage Clause	IOReason	GridDisk Policy	Large IO	Throttle IO	Throttle Large IO
Total (3)	12,411,084	3,447.52	3,383.42	48.49				
***celadm02	4,169,925	1,158.31	1,136.84	16.29				
***celadm01	4,147,818	1,152.17	1,130.89	16.08				
***celadm03	4,093,341	1,137.04	1,115.69	16.12				

フラッシュ・キャッシュ関連セクションの確認に加えて、以下の観点で IO リクエストの状況をチェックする必要もあります。

- セル/ディスク間の偏りの有無
- Top Databases の傾向
- Small Read Histogram – セルのヒストグラムに問題が示されていなくても、データベース側で記録された待機イベントのヒストグラムに表れる待機時間が長い場合、これは IO が原因で待機時間が増加しているのではなく、ネットワークまたは IORM キューイングに何か問題がある可能性があります。

スマート・スキャン (Smart Scan)

スマート・スキャン関連の待機イベントは、システムによって、さらにはクエリーによって異なる可能性があります。待機イベントには、IO 時間に加えて、ストレージ側にオフロードされたすべての処理時間が含まれます。処理コストはオフロードされる操作の種類によって異なり、一部の操作では CPU 負荷が高くなることがあります。

ほとんどの場合、*cell smart table scan* 待機イベントを確認することになります。パススルーが発生している場合（オフロードできない場合）、待機イベントその理由が表示されます。

表 3：スマート・スキャン関連の待機イベント

待機イベント	説明
cell smart table scan	セッションがスマート・スキャンの完了を待機している場合の待機イベント
cell smart table scan: db timezone upgrade	データベースのタイムゾーンがアップグレード・モードであるためにオフロードができない場合の待機イベント
cell smart table scan: disabled by user	ユーザー設定によりオフロードができない場合の待機イベント
cell smart table scan: passthru	スマート・スキャンのオフロードができない場合の待機イベント

cell smart table scan の待機時間が増加する理由はいくつかあります。一般的な原因としては、パススルー、ディスク IO の増加、ストレージ索引の欠如、列キャッシュの不足が挙げられます。場合によっては、データベースがブロック IO モードでの実行に戻る可能性があります。[9]

スマート・スキャンでは、パフォーマンスが低下している特定のクエリーを確認することが役立つことがあります。SQL Monitor は、スマート・スキャンに関する問題を診断するのに非常に有用なツールです。また、スマート・スキャンのパフォーマンスを把握し、それをストレージ・サーバーの統計と関連付けるために使用できるデータベース統計もあります。

Performance Summary のセクションには、**Smart Scan Summary** が含まれます。ここでは、データベース側の統計とストレージ側の統計を関連付けることができます。図 14 では、データベースが約 5.2GB/s の速度でスマートスキャンの対象となるデータを処理していますが（*cell physical IO bytes eligible for smart IOs*）、そのほとんどが進行中のオンライン暗号化のため、ブロック I/O モードに戻っています（cell num bytes in block IO during predicate offload）。

スマート・スキャンがオフロードされない理由を示すデータベース側の統計がいくつかあります。統計については、*Oracle Exadata Storage Software ユーザーズ・ガイド – Exadata の監視*で説明されています。

図 14：Performance Summary: Smart Scan Summary

Smart Scan Summary

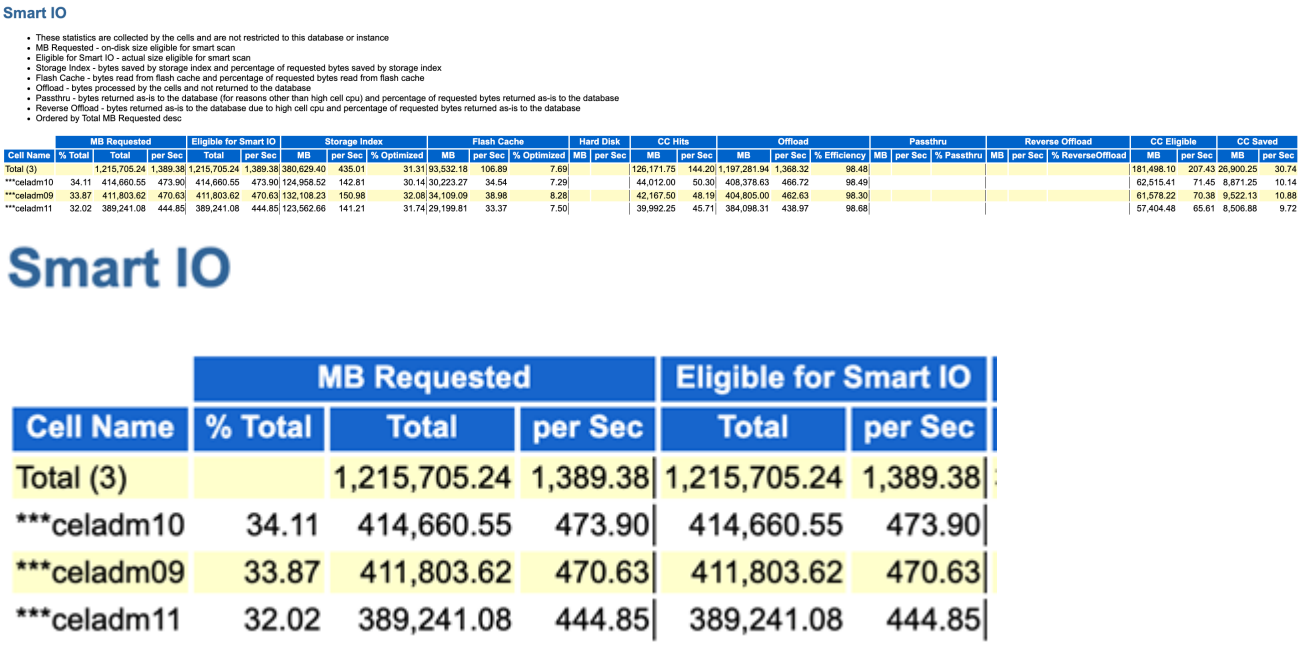
- Database activity and reasons are for this database, not restricted to an instance

Device Type	%MB	MB/s	
Flash	99.99	599.26	
Disk	0.01	0.07	
Database Smart Scan Savings	**MB**	**per Sec**	**% Saved**
cell physical IO bytes saved by columnar cache	5,352	6.12	0.12
cell physical IO bytes saved by storage index	40,919	46.77	0.89
Cell Smart IO Activity	**MB**	**per Sec**	
eligible	1,215,705	1,389.38	
eligible for smart IO	1,215,705	1,389.38	
Database Smart Scan Activity	**MB**	**per Sec**	
cell physical IO bytes eligible for predicate offload	4,614,445	5,273.65	
cell physical IO bytes eligible for smart IOs	4,552,081	5,202.38	
Database Passthru or Block IO	**Total**	**per Sec**	
cell num bytes in block IO during predicate offload (MB)	4,454,347	5,090.68	
cell num smart IO sessions in rdbms block IO due to online encr	903		

[9] パススルーまたはブロック IO モードが発生したことを示すデータベース統計があります。これらについては、『Oracle Exadata Storage Software ユーザーズ・ガイド』の「Exadata の監視」で説明されています。

図 15 の **Exadata Smart IO** セクションでは、ストレージ・サーバーが処理したスマート・スキャン対象の IO 量が表示されています。また、フラッシュから読み取られたストレージ索引によって削減されたバイト数、ディスクから読み取られたバイト数、および列キャッシュ使用量も表示されています。さらに、パススルーやリバース・オフロードが発生しているか否かも表示されます。

図 15：Exadata Smart IO



Smart IO

- These statistics are collected by the cells and are not restricted to this database or instance
- MB Requested - on-disk size eligible for smart scan
- Eligible for Smart IO - actual size eligible for smart scan
- Storage Index - bytes saved by storage index and percentage of requested bytes saved by storage index
- Flash Cache - bytes read from flash cache and percentage of requested bytes read from flash cache
- Offload - bytes processed by the cells and not returned to the database
- Passthru - bytes returned as-is to the database (for reasons other than high cell cpu) and percentage of requested bytes returned as-is to the database
- Reverse Offload - bytes returned as-is to the database due to high cell cpu and percentage of requested bytes returned as-is to the database
- Ordered by Total MB Requested desc

	MB Requested			Eligible for Smart IO		Storage Index			Flash Cache			Hard Disk		CC Hits		Offload			Passthru			Reverse Offload			CC Eligible		CC Saved	
Cell Name	% Total	Total	per Sec	Total	per Sec	MB	per Sec	% Optimized	MB	per Sec	% Optimized	MB	per Sec	MB	per Sec	MB	per Sec	% Efficiency	MB	per Sec	% Passthru	MB	per Sec	% ReverseOffload	MB	per Sec	MB	per Sec
Total (3)		1,215,705.24	1,389.38	1,215,705.24	1,389.38	380,629.40	435.01	31.31	93,532.18	106.89	7.69			126,171.75	144.20	1,197,281.94	1,368.32	98.48							181,498.10	207.43	26,900.25	30.74
***celadm10	34.11	414,660.55	473.90	414,660.55	473.90	124,958.52	142.81	30.14	30,223.27	34.54	7.29			44,012.00	50.30	408,378.63	466.72	98.49							62,515.41	71.45	8,871.25	10.14
***celadm09	33.87	411,803.62	470.63	411,803.62	470.63	132,108.23	150.98	32.08	34,109.09	38.98	8.28			42,167.50	48.19	404,805.00	462.63	98.30							61,578.22	70.38	9,522.13	10.88
***celadm11	32.02	389,241.08	444.85	389,241.08	444.85	123,562.66	141.21	31.74	29,199.81	33.37	7.50			39,992.25	45.71	384,098.31	438.97	98.68							57,404.48	65.61	8,506.88	9.72

Smart IO

	MB Requested			Eligible for Smart IO	
Cell Name	% Total	Total	per Sec	Total	per Sec
Total (3)		1,215,705.24	1,389.38	1,215,705.24	1,389.38
***celadm10	34.11	414,660.55	473.90	414,660.55	473.90
***celadm09	33.87	411,803.62	470.63	411,803.62	470.63
***celadm11	32.02	389,241.08	444.85	389,241.08	444.85

Exadata の **Smart IO** セクションに加えて、**Flash Cache User Reads** セクションでも、図 12 で前述したように、スキャンおよび列キャッシュに対して実行されている IO の量が表示されます。

Smart IO やフラッシュ・キャッシュおよび列キャッシュ関連のセクションの確認に加えて、以下についてもチェックする必要があります。

- **セル/ディスク間の偏りの有無：**スマート・スキャンはすべてのセル/ディスクに均等にヒットすることが期待されます。1 つのセル/ディスクのパフォーマンスが低下している場合、スキャン性能に影響します。スキャンはストレージ・サーバー上での大きなサイズの読取りが発生することが多いです。
- **Top Databases：**Large IO に対する IORM キュー時間も、スマート・スキャンの性能に影響を与える可能性があります。

一時領域への書き出し（Temp Spills）

データベースが TempIO（例えば、一時表領域への IO）を実行する場合、その IO はフラッシュ・キャッシュにキャッシュされることが期待されます。他の Large Writes もフラッシュ・キャッシュに書き出される可能性がありますが、さまざまな理由によって拒否されることもあります。データベースの Temp IO に関連する待機イベントの待機時間が増加する場合、フラッシュ・キャッシュにキャッシュされていないことが原因であることが多いです。

表 4：TEMP 関連の待機イベント

待機イベント	説明
direct path write temp	セッションによる temp の書込み時の待機イベント
direct path read temp	セッションによる temp の読取り時の待機イベント

図 16 で示すように、**Flash Cache User Writes - Large Writes** セクションには、ストレージ・サーバーによって処理されている Large Writes の書込みの量と種類が表示されています。

図 16：Flash Cache User Writes - Large Writes

Flash Cache User Writes - Large Writes

- These statistics are collected by the cells and are not restricted to this database or instance
- Large Writes consist of Temp Spills, Writes to Data and Temp Tables, and Write Only Operations
- Ordered by Total Write Requests desc

	Write Requests									
	Total					per Sec				
Cell Name	Total	Large Writes	Temp Spill	Data/Temp Tables	Write Only	Total	Large Writes	Temp Spill	Data/Temp Tables	Write Only
Total (16)	78,570,670	51,177,646	14,969,257	8,368,898	27,839,491	32,642.57	21,262.00	6,219.05	3,476.90	11,566.05
***celadm04	5,324,968	4,314,251	1,304,924	694,886	2,314,441	2,212.28	1,792.38	542.14	288.69	961.55
***celadm03	5,170,925	4,482,013	1,344,731	725,871	2,411,411	2,148.29	1,862.08	558.68	301.57	1,001.83
***celadm15	5,167,105	4,651,084	1,421,612	753,077	2,476,395	2,146.70	1,932.32	590.62	312.87	1,028.83
***celadm16	5,157,014	4,499,185	1,355,282	735,761	2,408,142	2,142.51	1,869.21	563.06	305.68	1,000.47
***celadm06	5,129,718	2,955,845	871,879	471,708	1,612,258	2,131.17	1,228.02	362.23	195.97	669.82
***celadm05	5,116,376	4,295,316	1,287,423	703,820	2,304,073	2,125.62	1,784.52	534.87	292.41	957.24
***celadm01	5,067,334	4,138,155	1,232,259	676,916	2,228,980	2,105.25	1,719.22	511.95	281.23	926.04
***celadm14	4,993,901	3,374,687	1,008,500	549,288	1,816,899	2,074.74	1,402.03	418.99	228.20	754.84
***celadm02	4,931,043	3,637,116	1,078,614	591,491	1,967,011	2,048.63	1,511.06	448.12	245.74	817.20
***celadm13	4,926,135	2,781,534	797,003	445,803	1,538,728	2,046.59	1,155.60	331.12	185.21	639.27
***celadm11	4,888,895	1,844,922	496,253	294,324	1,054,345	2,031.12	766.48	206.17	122.28	438.03
***celadm12	4,854,915	1,895,348	519,348	306,377	1,069,623	2,017.00	787.44	215.77	127.29	444.38
***celadm07	4,733,392	2,768,619	797,538	448,980	1,522,101	1,966.51	1,150.23	331.34	186.53	632.36
***celadm08	4,619,417	2,273,596	646,629	360,154	1,266,813	1,919.16	944.58	268.65	149.63	526.30
***celadm09	4,541,146	1,838,674	497,531	294,842	1,046,301	1,886.64	763.88	206.70	122.49	434.69
***celadm10	3,948,386	1,427,301	309,731	315,600	801,970	1,640.38	592.98	128.68	131.12	333.18

Flash Cache User Writes – Large Writes Rejections には、Large Writes（または Temp Spills）がフラッシュ・キャッシュに書き込まれない理由が表示されています。図 17 では、Large Writes の大部分が Global Limit のために拒否されています。これは、Large Writes が、フラッシュ・キャッシュ領域内で Large Writes 用に割り当てられた最大領域サイズを超えていることを意味します。

図 17：Flash Cache User Writes - Large Write Rejections

Flash Cache User Writes - Large Write Rejections

- These statistics are collected by the cells and are not restricted to this database or instance
- Eligible - does not include Global Criteria rejections
- Large writes may be rejected for the following reasons:
- Disk Not Busy - IORM determined the hard disk is not busy
- ASM - tagged as not cacheable by ASM
- CG Thrashing - large writes are causing flash cache group thrashing
- LW Thrashing - large writes are causing thrashing
- Max Limit - flash cache size for large writes is at flash cache group limit
- Global Limit - flash cache size for large writes is over global limit
- Flash Wear - large writes are causing excessive flash wear
- Flash Busy - flash is busy
- Keep - keep needs the cache lines
- Misc - large write rejections after passing global criteria

	Eligible		Rejections per Second			Global Criteria Rejections per Second						
Cell Name	Total	per Second	Disk Not Busy	ASM	Misc	CG Thrashing	LW Thrashing	Max Limit	Global Limit	Flash Wear	Flash Busy	Keep
Total (16)	53,289,117	22,139.23	5,642.79				14.98		36,708.18			
***celadm15	5,001,472	2,077.89	592.51						1,567.66			
***celadm16	4,839,143	2,010.45	568.40				8.42		1,632.24			
***celadm03	4,788,010	1,989.20	556.72						1,659.17			
***celadm04	4,593,342	1,908.33	529.06						1,748.64			
***celadm05	4,579,101	1,902.41	526.89						1,751.57			
***celadm01	4,389,978	1,823.84	501.43						1,833.99			
***celadm02	3,814,743	1,584.85	420.37						2,062.48			
***celadm14	3,529,072	1,466.17	383.05						2,215.82			
***celadm06	3,053,996	1,268.80	315.83						2,395.72			
***celadm13	2,851,013	1,184.47	287.25						2,497.34			
***celadm07	2,821,283	1,172.12	279.29						2,503.72			
***celadm08	2,255,863	937.21	194.20						2,793.79			
***celadm09	1,769,397	735.10	128.97						3,012.72			
***celadm12	1,816,520	754.68	134.11						2,964.19			
***celadm11	1,756,393	729.70	121.22						2,997.00			
***celadm10	1,429,791	594.01	103.52				6.56		3,072.11			

図 18 の **Flash Cache Space Usage** は、フラッシュ・キャッシュの約 20％が大きなサイズの書込み(Large Writes)に使用されていることを示しています。これは、フラッシュ・キャッシュの大きなサイズの書込みに使用できる最大量です。多くの場合、これはクエリーをチューニングして一時領域を必要とする量を減らすことで対処できます。場合によっては、一時的な需要の増加の原因は、一連の新しいクエリーをもたらすアプリケーション・アップグレード、またはオプティマイザの実行計画の変更をもたらすデータベース・アップグレードであると考えられます。*Oracle Exadata Database Machine System Software ユーザーズ・ガイド – 分析ワークロードのための Exadata スマート・フラッシュ・キャッシュの最適化*で説明されているように、ワークロードが一時 IO のためにより多くのフラッシュ・キャッシュ領域を必要とする場合、データベース・パラメータ main_workload_type の使用を検討することもできます。

図 18：Flash Cache Space Usage

Flash Cache Space Usage

- These statistics are collected by the cells and are not restricted to this database or instance
- Space is at the time of the end snapshot
- Ordered by Space (GB) desc

		Default								Keep			
		OLTP			Large Writes			%Scan	%Columnar	OLTP		%Scan	%Columnar
Cell Name	Space (GB)	%Clean	%Synced	%Unflushed	%Temp Spill	%Data/Temp	%Write Only			%Clean	%Unflushed		
Total (16)	380,539.40	26.49	22.46	23.48	0.21	0.63	19.16	5.59	1.99	0.00	0.00		
***celadm01	23,783.71	25.94	23.50	22.51	0.26	0.88	18.85	5.95	2.09	0.00	0.00		
***celadm03	23,783.71	25.94	23.40	22.71	0.28	0.95	18.77	5.87	2.07	0.00	0.00		
***celadm05	23,783.71	26.08	23.46	22.37	0.27	0.93	18.80	5.97	2.11	0.00	0.00		
***celadm08	23,783.71	26.34	20.38	24.83	0.16	0.29	19.55	6.23	2.22	0.00	0.00		
***celadm12	23,783.71	27.30	20.71	26.02	0.13	0.23	19.64	4.34	1.63		0.00		
***celadm13	23,783.71	25.71	24.23	21.11	0.19	0.63	19.17	6.67	2.29	0.00	0.00		
***celadm14	23,783.71	25.73	23.83	21.65	0.24	0.78	18.99	6.53	2.26	0.00	0.00		
***celadm15	23,783.71	25.65	23.56	22.81	0.29	0.99	18.71	5.88	2.11	0.00	0.00		
***celadm06	23,783.71	25.90	23.85	21.52	0.21	0.72	19.07	6.48	2.25	0.00	0.00		
***celadm10	23,783.71	30.49	16.26	28.54	0.08	0.16	19.76	3.46	1.26	0.00	0.00		
***celadm16	23,783.71	25.66	23.60	22.81	0.28	0.96	18.76	5.83	2.09	0.00	0.00		
***celadm04	23,783.71	26.08	23.50	22.43	0.27	0.93	18.80	5.87	2.13	0.00	0.00		
***celadm09	23,783.71	27.43	20.64	26.57	0.12	0.16	19.71	3.91	1.44		0.00		
***celadm11	23,783.71	27.68	20.66	26.44	0.12	0.13	19.75	3.79	1.42		0.00		
***celadm07	23,783.71	25.84	24.12	21.15	0.19	0.55	19.26	6.62	2.27		0.00		
***celadm02	23,783.71	26.03	23.60	22.17	0.25	0.82	18.93	6.06	2.14	0.00	0.00		

フラッシュ・キャッシュ・セクションの確認に加えて、以下をチェックする必要もあります。

- **IO latency：**一時領域への書き出しはストレージ・サーバー上で Large Writes になり易いです。
- **Top Databases：**Large IO に対する IORM によるキュー時間も、スマート・スキャンの性能に影響を与える可能性があります。ハード・ディスクがビジー状態の場合、ハード・ディスクの IORM キュー時間が増加することが予想されます。

シナリオ例：Exadata 固有の AWR データの分析

AWR の Exadata 関連セクションに慣れていただくために、実際のお客様のユースケースを反映した例を見ていきましょう。この例では、新しい Exadata システムに移行したお客様でパフォーマンスの低下が発生しています。

データベース統計の確認

まずは、**時間を要した上位の待機イベント**を確認して、パフォーマンス問題がストレージに関連している可能性を確認します。図 19 と図 20 は、単一インスタンスの**時間を要した上位の待機イベント**が示されています。なお、このデータベースは Real Application Clusters（RAC）データベースで複数のインスタンスから構成されていますが、すべてのインスタンスが同じような傾向をしていていたので、単一インスタンスの AWR レポートだけを確認しています。合計待機時間が最も長い待機イベントは、*cell smart table scan* であり、DB time のおよそ 62 %の時間を費やし、平均待機時間は 1.6 秒強でした。

また、この例のデータベース・リリース（バージョン）には、表 3 で前述した *cell single block physical read* 待機イベントの読取りがどこで発生したかを示す機能が実装されていません。ただし、図 19 で *cell single block physical read* の平均待機時間が 71.06 マイクロ秒であるということは、ほとんどのシングル・ブロック読取りが RDMA 経由で実行されていることを示唆しています。

図 19：Top 10 Foreground Events by Total Wait Time

Top 10 Foreground Events by Total Wait Time

Event	Waits	Total Wait Time (sec)	Avg Wait	% DB time	Wait Class
cell smart table scan	64,214	108.2K	1684.62ms	61.6	User I/O
DB CPU		17.2K		9.8	
gc buffer busy acquire	154,755	11.7K	75.81ms	6.7	Cluster
log file sync	219,631	8622.8	39.26ms	4.9	Commit
gc current block busy	847,915	8375.9	9.88ms	4.8	Cluster
gc cr disk read	60,262	7564.9	125.53ms	4.3	Cluster
gc buffer busy release	89,875	7254.8	80.72ms	4.1	Cluster
cell single block physical read	82,516,794	5863.9	71.06us	3.3	User I/O
gc cr block busy	136,084	4355	32.00ms	2.5	Cluster
cell multiblock physical read	61,819	2456.2	39.73ms	1.4	User I/O

図 20 に示す **Exadata Statistics - Performance Summary – Single Block Reads** のセクションから、全体の IO リクエスト数を示す *physical read total IO requests* に対して、RDMA 読取りを行った回数を示す *cell RDMA reads* が割合の大部分を占めていることから、非常に多くの IO が RDMA 読取りによって処理されていることが分かります。ただし、RDMA 読取りは通常、OLTP や単一ブロック IO で使用され、スマート・スキャンのような一度の IO リクエストで複数の連続ブロックを読み取る処理には効果がありません。

図 20：Exadata Statistics - Performance Summary – Single Block Reads

Database IOs	Value	per Sec
physical read total IO requests	504,142,681	140,743.35
physical read IO requests	503,650,620	140,605.98
cell flash cache read hits	6,972,977	1,946.67
cell ram cache read hits		
cell pmem cache read hits	7,099,606	1,982.02
cell RDMA reads	483,532,248	134,989.46

図 21 の Disk Activity をさらに詳しく見てみると、*Flash Cache read skips* と *Flash Cache write skips* の値が大きくなっているのが分かります。スキップは、フラッシュ・キャッシュの対象外の IO を示しています。また、スクラブ処理も実行中であることが確認できますが（*Scrub IO*）、前述の通り、Exadata システムはユーザー処理に起因するクライアント IO をスクラブ IO よりも優先するように設計されています。

図 21：Exadata Statistics - Performance Summary –Disk Activity

Disk Activity

- The following are possible causes of disk IO
- Smart Scan (estd) are estimated as 1MB per IO request

I/O per second	
Redo log writes	22,953.59
Smart Scans (estd)	38.20
Flash Cache misses (OLTP)	75.56
Flash Cache read skips	209.91
Flash Cache write skips	1,261.69
Flash Cache LW rejections (total)	2,121.87
Disk writer writes	2,392.09
Scrub IO	454,234.42

Exadata の構成

図 22 の Exadata Server Configuration のセクションから、このシステムは 12 台のストレージ・サーバーを搭載した X9M-2 であることが分かります。

図 22：Exadata Server Configuration: Exadata Storage Server Model

Exadata Storage Server Model

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

Model	CPU Count	Memory (GB)	# Cells	Cells
Oracle Corporation ORACLE SERVER X9-2L High Capacity	96/96	252	12	***celadm01, ***celadm02, ***celadm03, ***celadm04, ***celadm05, ***celadm06, ***celadm07, ***celadm08, ***celadm09, ***celadm10, ***celadm11, ***celadm12

Back to Exadata Server Configuration

Exadata Storage Server Version

- Version information of packages on the storage server

Package Type	Package Version	Cells
Kernel	4.14.35-2047.518.4.3.el7uek.x86_64	All (12)
Cell	cell-22.1.13.0.0_LINUX.X64_230818-1.x86_64	All (12)
Offload	cellofl-11.2.3.3.1_LINUX.X64_220513	All (12)
Offload	cellofl-12.1.2.4.0_LINUX.X64_230109	All (12)
Offload	cellofl-22.1.13.0.0_LINUX.X64_230818	All (12)

図 23 を見て最初に気付く潜在的な問題は、celadm11 と celadm12 が他のストレージ・セルの構成と異なっているということです。フラッシュ・ログが存在せず、フラッシュ・キャッシュがわずかに大きく状態が確認できます。

図 23：Exadata Server Configuration: Exadata Storage Information

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
10	23,845.81	1,500.56	0.50	12	4	12	72	(10): ***celadm01, ***celadm02, ***celadm03, ***celadm04, ***celadm05, ***celadm06, ***celadm07, ***celadm08, ***celadm09, ***celadm10
2	23,846.31	1,500.56	0.00	12	4	12	72	(2): ***celadm11, ***celadm12
Total (12)	286,150.75	18,006.75	5.00	144	48	144	864	All (12)

IO の分布

ディスクタイプごとにパフォーマンス特性は異なるため、Outlier セクションにはディスク・タイプごとの IO 量がレポートされます。ディスク・タイプを識別するための形式は*<F または H>/<サイズ>*で、F はフラッシュ・デバイス、H はハード・ディスクを表します。Extreme Flash ストレージ・サーバーの場合、容量最適化フラッシュ・デバイスとパフォーマンス最適化フラッシュ・デバイスの両方が *F/<サイズ>* としてレポートされ、<サイズ>にフラッシュ・デバイスの種類を示します（例：X10M EF サーバーでは、容量最適化フラッシュ・デバイスは *F/14.0T*、パフォーマンス最適化フラッシュ・デバイスは *F/5.8T* としてレポートされます）。

図 24 でストレージ・サーバーの IO 量を確認すると、celadm11 および celadm12 で外れ値が検出されています。他のセルでは（スクラブ処理による）大量の IO が発生していますが、celadm11 と celadm12 ではそれらと IO の傾向がかなり異なります。これら 2 つのセルは、479IOPS 以下にもかかわらず、ほぼ 100%の稼働率（% Disk Utilization）に達しています。

図 24：Exadata OS Statistics Outlies – Exadata OS IO Statistics - Outlier Cells

Exadata OS IO Statistics - Outlier Cells

- These statistics are collected by the OS on the cells and are not restricted to this database or instance
- Outliers are cells whose average performance is outside the normal range, where normal range is -/+ 1 standard deviation
- Outlier cells must have a mininum of 10 IOPs. Idle cells 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)
- 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
- %Total - Avg [IOPs |IO MB/s] of the cell as a percentage of total [IOPs | IO MB/s] for the disk type
- **There are no cell outliers on flash**

				IOPs						IO MB/s						% Disk Utilization		
				Total	% Total	Per Cell	Per Disk			Total	% Total	Per Cell	Per Disk					
Disk Type	Cell Name	# Cells	# Disks			Average	Mean	Std Dev	Normal Range			Average	Mean	Std Dev	Normal Range	Mean	Std Dev	Normal Range
F/5.8T	All	12	48	155,876.49		12,989.71	3,247.43	1,550.33	1,697.09 - 4,797.76	8,691.12		724.26	181.06	167.93	13.13 - 349.00	4.48	5.35	0.00 - 9.83
H/16.0T	All	12	144	466,731.39		38,894.28	3,241.19	1,498.45	1,742.74 - 4,739.64	8,467.33		705.61	58.80	17.98	40.82 - 76.78	40.96	26.08	14.88 - 67.04
Outlier	***celadm11		12		1.23	* 5,751.37	* 479.28				4.87	v 412.28	v 34.36			^ 97.99		
Outlier	***celadm12		12		1.23	* 5,748.10	* 479.01				4.74	v 401.34	v 33.44			^ 98.19		

				I/O		
Disk Type	Disk Name	Cell Name	Statistic	per Disk	Std Dev	Range
H/16.0T	All	All	Cell Server IOPs	* 3,252.91	1,476.46	0.00 - 7,682.29
			Cell Server IO MB/s	59.80	17.97	5.88 - 113.72
			Cell Server IO Latency	30.89ms	90.81ms	0.00ns - 303.33ms
H/16.0T	CD_03_***celadm11	***celadm11	Cell Server IOPs	* 586.93		
			Cell Server IO MB/s	44.43		
			Cell Server IO Latency	223.35ms		

図 25 の Cell Server Statistics を詳しく見ると、celadm11 と celadm12 の IO タイプが他のセルと異なっていることが改めて分かります。IO のほとんどが Small Writes で、Large Writes は一部にとどまっています。他のストレージ・セルにおける Small Reads はスクラブ関連のものです。合計待機時間が最も長い *cell smart table scan* 待機イベントは、ストレージ・サーバーで Large Reads を発生させますが、それとは異なる IO 特性、特に celadm11 と celadm12 における大量の書込みと高いディスク稼働率が懸念されます。

図 25：Cell Server Statistics - Outlier Cells

Exadata Cell Server IOPS Statistics - Outlier Cells

- These statistics are collected by the cells and are not restricted to this database or instance
- Outliers are cells whose average performance is outside the normal range, where normal range is -/+ 1 standard deviation
- Outlier cells must have a mininum of 10 IOPs. Idle cells are not considered for outlier analysis.
- Outliers for small reads, small writes, large read, large writes, must have a minimum of 10 requests for the corresponding small read, small write, large read, large write statistic.
- 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)
- 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
- Disk Type <F|H|M>/<size>: F-Flash, H-Hard Disk, M-pMEM; PMEM I/O only include remote I/Os processed by cellsrv
- % Total - Avg IOPs of the cell as a percentage of total IOPs for the disk type

				IOPs						Small Reads/s				Small Writes/s				Large Reads/s				Large Writes/s			
				Total	% Total	Per Cell	Per Disk			Per Cell	Per Disk			Per Cell	Per Disk			Per Cell	Per Disk			Per Cell	Per Disk		
Disk Type	Cell Name	# Cells	# Disks			Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range
F/5.8T	All	12	48	151,863.64		12,655.30	3,163.83	1,525.14	1,638.68 - 4,688.97	139.65	34.91	19.21	15.70 - 54.12	7,724.76	1,931.19	991.09	940.10 - 2,922.28	4,573.56	1,143.39	1,318.26	0.00 - 2,461.65	217.34	54.33	33.66	20.68 - 87.99
Outlier	***celadm11		4		6.45	9,789.33	2,447.33			^ 263.62	^ 65.90			4,282.57	1,070.64			5,243.15	1,310.79			0.00	0.00		
Outlier	***celadm12		4		6.42	9,745.03	2,436.26			^ 265.22	^ 66.31			4,293.59	1,073.40			5,186.22	1,296.56			0.00	0.00		
H/16.0T	All	12	144	468,419.16		39,034.93	3,252.91	1,476.46	1,776.45 - 4,729.37	37,819.52	3,151.63	1,638.80	1,512.83 - 4,790.42	1,093.43	91.12	204.94	0.00 - 296.06	13.85	1.15	2.49	0.00 - 3.65	108.13	9.01	11.31	0.00 - 20.32
Outlier	***celadm11		12		1.42	* 6,680.35	* 555.03			v 230.88	v 19.24			^ 6,016.97	^ 501.41			68.98	5.75			^ 343.52	^ 28.63		
Outlier	***celadm12		12		1.42	* 6,657.00	* 554.75			v 239.00	v 19.92			^ 6,017.22	^ 501.44			58.58	4.88			^ 342.20	^ 28.52		

Small Reads/s				Small Writes/s			
Per Cell	Per Disk			Per Cell	Per Disk		
Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range
139.65	34.91	19.21	15.70 - 54.12	7,724.76	1,931.19	991.09	940.10 - 2,922.28
^ 263.62	^ 65.90			4,282.57	1,070.64		
^ 265.22	^ 66.31			4,293.59	1,073.40		
37,819.52	3,151.63	1,638.80	1,512.83 - 4,790.42	1,093.43	91.12	204.94	0.00 - 296.06
v 230.88	v 19.24			^ 6,016.97	^ 501.41		
v 239.00	v 19.92			^ 6,017.22	^ 501.44		

スマート・スキャン(Smart Scan)

Smart IO のセクションでは、システム全体におけるスマート IO の実行状況が示され、スマート・スキャンがどの程度効率的に行われているかを判断できます [10]。図 26 では、celadm11 と celadm12 の IO 傾向が他のセルと異なっていることが分かります。ハード・ディスクの IO の絶対量はそれほど多くないが他のセルよりも明らかに高く、スマート IO のほとんどがフラッシュ・キャッシュで実行されています。

図 26：スマート・スキャン情報が表示された AWR レポートの Smart IO セクション

Smart IO

- These statistics are collected by the cells and are not restricted to this database or instance
- Storage Index - bytes saved by storage index and percentage of requested bytes saved by storage index
- Flash Cache - bytes read from flash cache and percentage of requested bytes read from flash cache
- Offload - bytes processed by the cells and not returned to the database
- Passthru - bytes returned as-is to the database (for reasons other than high cell cpu) and percentage of requested bytes returned as-is to the database
- Reverse Offload - bytes returned as-is to the database due to high cell cpu and percentage of requested bytes returned as-is to the database
- Ordered by Total MB Requested desc

	MB Requested			Storage Index			Flash Cache			Hard Disk		Offload			Passthru			Reverse Offload		
Cell Name	% Total	Total	per Sec	MB	per Sec	% Optimized	MB	per Sec	% Optimized	MB	per Sec	MB	per Sec	% Efficiency	MB	per Sec	% Passthru	MB	per Sec	% ReverseOffload
Total (12)		49,987,686.30	13,955.24	11,167,783.30	3,117.75	22.34	2,390,468.05	667.36	4.78	136,817.91	38.20	49,781,224.74	13,897.61	99.59						
***celadm04	8.85	4,425,114.65	1,235.38	972,641.27	271.54	21.98	175,545.73	49.01	3.97	1,234.32	0.34	4,404,539.76	1,229.63	99.54						
***celadm03	8.85	4,424,744.76	1,235.27	1,030,869.01	287.79	23.30	163,250.29	45.58	3.69	1,304.19	0.36	4,407,445.84	1,230.44	99.61						
***celadm02	8.84	4,419,207.44	1,233.73	935,552.32	261.18	21.17	173,433.65	48.42	3.92	1,199.35	0.33	4,401,926.87	1,228.90	99.61						
***celadm06	8.76	4,376,865.38	1,221.91	923,810.27	257.90	21.11	168,191.65	46.95	3.84	1,272.13	0.36	4,360,726.65	1,217.40	99.63						
***celadm07	8.29	4,144,706.42	1,157.09	986,400.34	275.38	23.80	165,419.30	46.18	3.99	1,373.77	0.38	4,128,838.01	1,152.66	99.62						
***celadm09	8.25	4,125,603.35	1,151.76	994,047.81	277.51	24.09	160,083.99	44.69	3.88	1,106.09	0.31	4,109,619.69	1,147.30	99.61						
***celadm10	8.17	4,081,556.81	1,139.46	922,042.13	257.41	22.59	159,786.55	44.61	3.91	1,084.40	0.30	4,064,979.35	1,134.84	99.59						
***celadm01	8.12	4,057,502.46	1,132.75	951,237.98	265.56	23.44	161,481.88	45.08	3.98	1,530.80	0.43	4,041,807.26	1,128.37	99.61						
***celadm11	8.08	4,041,011.18	1,128.14	889,041.72	248.20	22.00	379,993.20	106.08	9.40	77,307.77	21.58	4,019,395.12	1,122.11	99.47						
***celadm12	8.08	4,037,783.49	1,127.24	737,339.64	205.85	18.26	342,545.17	95.63	8.48	45,986.15	12.84	4,019,724.91	1,122.20	99.55						
***celadm08	8.03	4,014,117.75	1,120.64	948,053.77	264.67	23.62	167,107.78	46.65	4.16	1,425.48	0.40	3,997,760.53	1,116.07	99.59						
***celadm05	7.68	3,839,472.61	1,071.88	876,747.03	244.76	22.84	173,628.86	48.47	4.52	1,993.47	0.56	3,824,460.75	1,067.69	99.61						

	MB Requested			Storage Index			Flash Cache			Hard Disk	
Cell Name	% Total	Total	per Sec	MB	per Sec	% Optimized	MB	per Sec	% Optimized	MB	per Sec
Total (12)		49,987,686.30	13,955.24	11,167,783.30	3,117.75	22.34	2,390,468.05	667.36	4.78	136,817.91	38.20
***celadm04	8.85	4,425,114.65	1,235.38	972,641.27	271.54	21.98	175,545.73	49.01	3.97	1,234.32	0.34
***celadm03	8.85	4,424,744.76	1,235.27	1,030,869.01	287.79	23.30	163,250.29	45.58	3.69	1,304.19	0.36
***celadm02	8.84	4,419,207.44	1,233.73	935,552.32	261.18	21.17	173,433.65	48.42	3.92	1,199.35	0.33
***celadm06	8.76	4,376,865.38	1,221.91	923,810.27	257.90	21.11	168,191.65	46.95	3.84	1,272.13	0.36
***celadm07	8.29	4,144,706.42	1,157.09	986,400.34	275.38	23.80	165,419.30	46.18	3.99	1,373.77	0.38
***celadm09	8.25	4,125,603.35	1,151.76	994,047.81	277.51	24.09	160,083.99	44.69	3.88	1,106.09	0.31
***celadm10	8.17	4,081,556.81	1,139.46	922,042.13	257.41	22.59	159,786.55	44.61	3.91	1,084.40	0.30
***celadm01	8.12	4,057,502.46	1,132.75	951,237.98	265.56	23.44	161,481.88	45.08	3.98	1,530.80	0.43
***celadm11	8.08	4,041,011.18	1,128.14	889,041.72	248.20	22.00	379,993.20	106.08	9.40	77,307.77	21.58
***celadm12	8.08	4,037,783.49	1,127.24	737,339.64	205.85	18.26	342,545.17	95.63	8.48	45,986.15	12.84
***celadm08	8.03	4,014,117.75	1,120.64	948,053.77	264.67	23.62	167,107.78	46.65	4.16	1,425.48	0.40
***celadm05	7.68	3,839,472.61	1,071.88	876,747.03	244.76	22.84	173,628.86	48.47	4.52	1,993.47	0.56

Smart Flash Log

Exadata Server Configuration のセクションで、celadm11 および celadm12 ではフラッシュ・キャッシュとフラッシュ・ログが他のストレージ・セルと異なることが分かりました。Top 10 Foreground Events by Total Wait Time セクションを確認する限り、REDO ログ書込みに関連した問題は顕在化していませんでしたが、図 27 では、これら 2 つのセルではフラッシュ・ログの動作が異なっていることが確認できます。フラッシュ・ログが構成されていないため、これらの 2 つのセルでのスキップが発生していることが示されています。これは **Exadata Server Configuration - Storage Information** セクションで示された情報と一致しています。

図 27：celadm11 および celadm12 にフラッシュ・ログが含まれないことを示す Flash Log Skip Details

Flash Log Skip Details

- These statistics are collected by the cells and are not restricted to this database or instance
- The top cells by Total Skip Count are displayed
- Outliers - # of outliers when redo log write skips use of Flash Log
- The Flash Log write may be skipped due to the following reasons:
- Busy - data pending to be written to disk
- Large Data - size of data larger than available space
- No Buffer - Flash Log buffer allocation failure
- On Flash - redo log resides on flash disk
- No FL Disk - no active Flash Log disks
- Disabled Grid Disk - flash log disabled for underlying grid disk (due to recent write errors)
- IORM Plan - disabled by IORM plan
- IORM Limit - IORM limit reached for disk containing redo log

	Skip Count									
Cell Name	% Total	Total	Outliers	Busy	Large Data	No Buffer	On Flash	No FL Disks	Disabled Grid Disk	IORM Plan	IORM Limit
Total (12)		13,868,818	1,763,587					13,868,818			
***celadm11	50.01	6,935,543	866,969					6,935,543			
***celadm12	49.99	6,933,275	896,618					6,933,275			

[10]特定の SQL 文にのみ影響しているスマート・スキャンの問題を調査する場合は、AWR レポートよりも SQL Monitor が診断ツールとして適しています。

Smart Flash Cache

Exadata Server Configuration - Storage Information セクションで、調査対象の 2 つのセルではフラッシュ・キャッシュのサイズがわずかに大きいことを確認していますが、図 28 の Flash Cache Configuration セクションからはさらに詳しい情報が得られ、このシナリオの最も可能性の高い問題の原因が強調されています。**Flash Cache Configuration セクション**では、対象の 2 つのセルのステータスが *normal - flushing* の状態であることが分かります。ストレージ・セルが *normal - flushing* ステータスにあるということは、フラッシュ・キャッシュ内のデータをハード・ディスクへフラッシュ（書出し）している最中であることを示しています。この間、クライアント IO はフラッシュ・キャッシュを使用できず、ハード・ディスクにリダイレクトされることになります。

図 28：セル間の相違を表示する Flash Cache Configuration

Flash Cache Configuration

- These statistics are collected by the cells and are not restricted to this database or instance
- Size (GB) - configured size for Flash Cache

Mode	Compression	Status	Size (GB)	Cells
WriteBack		normal	23845.81	***celadm01, ***celadm02, ***celadm03, ***celadm04, ***celadm05, ***celadm06, ***celadm07, ***celadm08, ***celadm09, ***celadm10
WriteBack		"normal - flushing"	23846.31	***celadm11, ***celadm12

これが問題の原因であるように見えますが、、念のため、別のセクションをレビューして、他のセルで他の問題がないことを確認します。

図 29 の **Flash Cache User Reads Per Second** セクションからは、問題の 2 つのセルは他のセルよりもフラッシュ・キャッシュ上のデータにヒットしなかった回数（Misses）が高く、IO リクエスト数も相対的に低いことが分かります。

図 29：Flash Cache User Reads Per Second

Flash Cache User Reads Per Second

- These statistics are collected by the cells and are not restricted to this database or instance
- Total - total number of reads per second from Flash Cache
- OLTP/Scan/Columnar reads include reads on keep objects
- Ordered by Total Hit Read Requests per Second desc

	Read Requests per Second						Read MB per Second				
Cell Name	Total Hits	OLTP	Scan	Columnar	Keep	Misses	Total Hits	OLTP	Scan	Columnar	Keep
Total (12)	11,594.01	1,049.25	518.75	10,026.01		75.56	5,861.62	129.48	501.12	5,231.03	
***celadm02	1,079.07	96.54	49.70	932.83		3.29	526.96	11.99	48.23	466.75	
***celadm04	1,057.50	81.90	49.74	925.86		3.39	529.01	11.38	48.06	469.56	
***celadm06	1,052.37	86.37	51.92	914.08		3.24	519.85	11.66	50.20	457.98	
***celadm03	1,040.34	86.63	46.94	906.77		3.13	521.26	11.70	45.25	464.31	
***celadm10	992.36	100.98	47.78	843.61		3.34	498.03	12.58	46.10	439.36	
***celadm01	983.60	102.56	50.30	830.73		3.25	495.66	12.58	48.63	434.45	
***celadm07	979.97	88.84	47.77	843.37		3.08	508.87	11.60	46.15	451.12	
***celadm09	979.88	94.21	49.75	835.92		3.46	488.41	11.90	48.26	428.24	
***celadm08	950.58	85.28	48.78	816.53		3.09	487.76	11.25	47.15	429.36	
***celadm05	917.97	86.82	54.12	777.03		3.37	484.20	11.49	52.40	420.32	
***celadm11	796.71	66.01	11.85	718.84		20.46	405.39	5.50	11.18	388.71	
***celadm12	763.67	73.12	10.10	680.45		22.47	396.24	5.85	9.52	380.88	

また、図 30 の **Flash Cache User Reads Efficiency** では、問題の 2 つのセルのヒット率（%Hit）が OLTP とスキャンの両方で著しく低いことが分かります。

図 30：Flash Cache User Reads Efficiency

Flash Cache User Reads Efficiency

- These statistics are collected by the cells and are not restricted to this database or instance
- Ordered by Total Hit Requests desc

	Total Hits		OLTP			Scan		
Cell Name	Requests	MB	Read Hits	Misses	%Hit	Read MB	Attempted MB	%Hit
Total (12)	41,529,752	20,996,337.23	3,758,428	270,645	93.28	1,794,995.98	2,197,159.57	81.70
***celadm02	3,865,226	1,887,569.34	345,810	11,789	96.70	172,746.67	175,365.11	98.51
***celadm04	3,787,962	1,894,896.40	293,380	12,129	96.03	172,155.65	174,275.78	98.78
***celadm06	3,769,580	1,862,103.42	309,365	11,595	96.39	179,825.88	184,415.85	97.51
***celadm03	3,726,486	1,867,160.86	310,299	11,216	96.51	162,086.95	165,356.18	98.02
***celadm10	3,554,633	1,783,944.87	361,698	11,976	96.80	165,114.74	166,989.78	98.88
***celadm01	3,523,252	1,775,436.43	367,387	11,626	96.93	174,178.87	176,720.34	98.56
***celadm07	3,510,267	1,822,764.91	318,217	11,023	96.65	165,300.50	167,560.34	98.65
***celadm09	3,509,938	1,749,476.91	337,467	12,385	96.46	172,873.25	174,874.56	98.86
***celadm08	3,404,981	1,747,157.86	305,456	11,069	96.50	168,898.65	171,190.92	98.66
***celadm05	3,288,153	1,734,405.35	310,988	12,058	96.27	187,684.50	190,720.71	98.41
***celadm11	2,853,803	1,452,091.16	236,455	73,281	76.34	40,044.82	246,543.56	16.24
***celadm12	2,735,471	1,419,329.73	261,906	80,498	76.49	34,085.52	203,146.44	16.76

同様に、図 31 の **Flash Cache User Writes** セクションを確認すると、celadm11 および celadm12 には、この AWR レポートの期間中に 1,900 万回を超える Partial Writes が発生していたことがわかります。これは、1 秒あたりに換算すると 5,000 回を超えています。Partial Writes は一般的ではなく、Write IO がフラッシュ・キャッシュとハード・ディスクの両方に対して行われる場合に発生します。これら 2 つのセルでの大量の Partial Writes も、やはりフラッシュ・キャッシュの *normal - flushing* ステータスが原因です。

図 31：Flash Cache User Writes

Flash Cache User Writes

- These statistics are collected by the cells and are not restricted to this database or instance
- Total - total number of write requests or write megabytes to Flash Cache
- First Writes/Overwrites also include Keep Writes and Large Writes
- Ordered by Total Write Requests desc

	Write Requests											
	Total						per Sec					
Cell Name	Total	First Writes	Overwrites	Partial Writes	Keep	Large Writes	Total	First Writes	Overwrites	Partial Writes	Keep	Large Writes
Total (12)	256,788,205	2,932,508	214,697,488	39,158,209	456	4,652,964	71,688.50	818.68	59,937.88	10,931.94	0.13	1,298.98
***celadm07	21,748,043	286,311	21,458,417	3,315	96	389,083	6,071.48	79.93	5,990.62	0.93	0.03	108.63
***celadm01	21,691,411	309,183	21,378,193	4,035	30	380,944	6,055.67	86.32	5,968.23	1.13	0.01	106.35
***celadm06	21,587,882	293,244	21,290,912	3,726	42	374,080	6,026.77	81.87	5,943.86	1.04	0.01	104.43
***celadm09	21,553,320	288,407	21,261,174	3,739	62	377,659	6,017.12	80.52	5,935.56	1.04	0.02	105.44
***celadm10	21,550,908	291,059	21,256,414	3,435	16	367,204	6,016.45	81.26	5,934.23	0.96	0.00	102.51
***celadm02	21,541,988	296,061	21,242,393	3,534	21	386,547	6,013.96	82.65	5,930.32	0.99	0.01	107.92
***celadm08	21,539,550	284,177	21,252,062	3,311	47	393,403	6,013.27	79.33	5,933.02	0.92	0.01	109.82
***celadm04	21,514,265	294,699	21,216,224	3,342	26	391,584	6,006.22	82.27	5,923.01	0.93	0.01	109.32
***celadm03	21,507,652	295,984	21,207,577	4,091	19	403,091	6,004.37	82.63	5,920.60	1.14	0.01	112.53
***celadm05	21,461,587	293,383	21,164,568	3,636	47	381,920	5,991.51	81.90	5,908.59	1.02	0.01	106.62
***celadm11	20,674,017		1,040,249	19,633,768	32	445,183	5,771.64		290.41	5,481.23	0.01	124.29
***celadm12	20,417,582		929,305	19,488,277	18	362,266	5,700.05		259.44	5,440.61	0.01	101.13

さらに、図 32 の **Flash Cache User Writes - Skips** では、celadm11 および celadm12 でフラッシュ・キャッシュをバイパスする書込みが確認されます。（このシナリオの AWR レポートでは、書き込みがフラッシュ・キャッシュをバイパスしていることは確認できますが、その理由までは示されていません。Oracle Database 19.19 以降では、その理由をより明確に理解するための追加情報も表示されるようになっています。）このケースでは、これらの書込みはおそらくフラッシュ・キャッシュの *normal - flushing* ステータスによるものと考えられます。

図 32：Flash Cache User Writes - Skips

Flash Cache User Writes - Skips

- These statistics are collected by the cells and are not restricted to this database or instance
- Flash Cache User Writes Skips are writes that bypass the flash cache
- Total Skipped includes all writes that have bypassed flash cache
- Only the following possible reasons for bypassing the flash cache are displayed:
- Storage Clause - flash cache skipped due to storage clause
- GridDisk Policy - flash cache skipped due to griddisk caching policy
- Large IO - flash cache skipped due to size of IO
- Throttle IO - flash cache skipped due to throttling

	Requests Skipped		Read Requests Skipped per Second			
Cell Name	Total	per Second	Storage Clause	GridDisk Policy	Large IO	Throttle IO
Total (12)	4,519,376	1,261.69	11.70		267.34	
***celadm12	1,409,655	393.54	1.09		0.53	
***celadm11	1,323,763	369.56	1.26		0.64	
***celadm05	197,654	55.18	0.88		26.54	
***celadm10	190,646	53.22	0.84		26.57	
***celadm03	189,131	52.80	1.04		26.72	
***celadm07	177,513	49.56	1.00		26.72	
***celadm01	177,300	49.50	0.84		26.61	
***celadm09	177,124	49.45	0.96		26.55	
***celadm08	173,334	48.39	1.15		26.55	
***celadm02	170,419	47.58	1.04		26.64	
***celadm04	168,017	46.91	0.95		26.64	
***celadm06	164,820	46.01	0.65		26.62	

図 33 の **Flash Cache Internal Reads** セクションには、Disk Writer のアクティビティが表示されています。Disk Writer は、フラッシュ・キャッシュ上のダーティ・データをハード・ディスクへ同期させる役割を担います。Oracle Database 19.19 以降、AWR レポートには Disk Writer による書込みのタイプも含まれており、その書込みがフラッシュ処理（Flush）に関連するかどうかが表示されます。このケースでは、celadm11 および celadm12 でフラッシュ・キャッシュからの読み込みとハード・ディスクへの書込みが多くなっており、これはフラッシュ処理（Flush）によるものと推測されます。

図 33：Flash Cache Internal Reads

Flash Cache Internal Reads

- These statistics are collected by the cells and are not restricted to this database or instance
- Read to Disk Write - reads from flash cache to write to hard disk
- Disk Writer IO Detail - actual number of IOs
- Ordered by Total Read Reqs desc

					Disk Writer IO Detail			
	Read to Disk Write Reqs		Read to Disk Write MB		Reads from Flash		Writes to Hard Disk	
Cell Name	Total	per Sec	Total	per Sec	Requests/s	MB/s	Requests/s	MB/s
Total (12)	4,208,711	1,174.96	2,588,053.79	722.52	6,364.80	795.61	2,392.09	722.51
***celadm12	1,052,863	293.93	743,314.99	207.51	1,828.07	228.51	575.90	207.51
***celadm11	1,029,360	287.37	753,638.78	210.40	1,820.96	227.62	544.93	210.39
***celadm02	216,889	60.55	110,835.30	30.94	278.28	34.79	121.33	30.94
***celadm03	214,490	59.88	113,184.04	31.60	280.03	35.01	113.90	31.60
***celadm05	214,428	59.86	108,711.97	30.35	273.28	34.16	134.01	30.35
***celadm01	214,382	59.85	111,911.20	31.24	279.20	34.90	122.61	31.24
***celadm06	213,901	59.72	106,094.49	29.62	265.81	33.23	137.77	29.62
***celadm10	212,302	59.27	107,022.62	29.88	266.42	33.30	129.87	29.88
***celadm08	211,852	59.14	107,374.48	29.98	267.01	33.38	131.28	29.98
***celadm04	210,528	58.77	114,243.83	31.89	276.09	34.51	113.00	31.89
***celadm09	210,516	58.77	107,038.97	29.88	269.84	33.73	138.46	29.88
***celadm07	207,200	57.84	104,683.12	29.22	259.81	32.48	129.02	29.23

最後に、図 34 の **Flash Cache Internal Writes** セクションでは、celadm11 および celadm12 のフラッシュ・キャッシュに対して書込みが行われていないことを示しています。通常、フラッシュ・キャッシュへのデータをキャッシュする処理（Population）は、フラッシュ・キャッシュのミスを契機に発生します。このケースでは、ミスの数が多いにもかかわらず、Population が行われていません。つまり、これはフラッシュ・キャッシュ内のデータをハード・ディスクへフラッシュ処理（書出し）している最中であることとも一致します。

図 34：Flash Cache Internal Writes

Flash Cache Internal Writes

- These statistics are collected by the cells and are not restricted to this database or instance
- The top cells by Total Write Requests are displayed
- Population - population writes due to read misses
- Metadata - Write-Back Flash Cache metadata persistence writes
- ordered by Total Write requests desc

	Write Requests						
			Population				Metadata
Cell Name	Total	per Sec	Total	per Sec	Columnar per Sec	Keep per Sec	per Sec
Total (12)	32,636,749	9,111.32	681,124	190.15	48,852		8,921.17
***celadm12	12,483,139	3,484.96					3,484.96
***celadm11	12,435,571	3,471.68					3,471.68
***celadm01	881,979	246.23	77,739	21.70	5,125		224.52
***celadm03	849,571	237.18	66,726	18.63	4,289		218.55
***celadm02	844,777	235.84	67,159	18.75	4,777		217.09
***celadm04	807,968	225.56	63,558	17.74	4,967		207.82
***celadm05	780,215	217.82	62,572	17.47	5,948		200.35
***celadm10	737,358	205.85	74,045	20.67	4,288		185.18
***celadm09	724,837	202.36	69,138	19.30	4,280		183.05
***celadm06	702,366	196.08	65,620	18.32	4,150		177.76
***celadm08	695,602	194.19	66,441	18.55	5,675		175.65
***celadm07	693,366	193.57	68,126	19.02	5,353		174.55

Flash Cache のすべてのセクションで、celadm11 および celadm12 でのフラッシュ・キャッシュのアクティビティ（場合によっては IO の欠如）が示されており、どちらにも *normal - flushing* のステータスの状況が確認できます。

この時点で、これら 2 つのセルのフラッシュキャッシュの状態が、観察された問題の主な原因であると言うことができます。しかし、仮説を検証するために残りのセクションも確認していきます。

Exadata IO Reasons

Exadata IO Reasons のセクションを見ると、ストレージ・サーバーに対して IO が発行された理由がわかります。Exadata IO Reasons に表示される IO は、読取りと書込みの両方を含み、またハード・ディスクとフラッシュ・デバイスを含みます。

図 35 と図 36 からは、celadm11 および celadm12 の IO Reasons が他のストレージ・セルと異なっていることが分かります。

他のストレージ・セルで示されている内容は以下のとおりです。

- scrub IO – スクラブ処理の IO で、ハード・ディスクからの小規模読取りにつながり、通常はクライアントに影響しません。
- redo log write – REDO ログへの書込み。Smart Flash Log および Smart Flash Log Write-Back により、[11]REDO ログ書込みはフラッシュ・ログおよびフラッシュ・キャッシュに対して実行されます。
- smart scan – スマート・スキャン関連のアクティビティ。
- limit dirty buffer writes、 aged writes by dbwr、medium-priority checkpoint writes – データベース・ライター（DBWR）によるバッファ・キャッシュからストレージ・セルへのデータブロックの書込み。

ストレージ・セル celadm11 および celadm12 では、IO の 36～37 パーセントが Internal IO によるもので、これは他のセルよりもはるかに多い値です。また、問題があるとみられる 2 つのセル（celadm11 および celadm12）では、スクラブ処理が実行されていないことに注意してください。

図 35：IO Reasons by Requests

Top IO Reasons by Requests

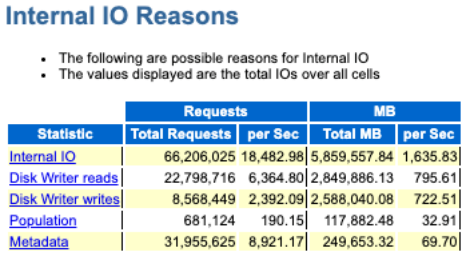
- The top IO reasons by requests per cell are displayed
- Only reasons with over 1% of IO requests for each cell are displayed
- At most 10 reasons are displayed per cell
- %Cell - the percentage of IO requests on the cell due to the IO reason
- Ordered by Cell Name, Requests Value desc

		Requests			MB	
Cell Name	IO Reason	%Cell	Total Requests	per Sec	Total MB	per Sec
Total (12)	scrub IO	73.25	1,627,067,705	454,234.42	25,422,932.89	7,097.41
	redo log write	7.58	168,474,205	47,033.56	4,243,834.80	1,184.77
	smart scan	7.16	158,944,470	44,373.11	20,026,441.63	5,590.85
	limit dirty buffer writes	3.71	82,315,998	22,980.46	1,038,852.29	290.02
	Internal IO	2.98	66,206,025	18,482.98	5,859,557.84	1,635.83
	REQ list writes	1.78	39,635,344	11,065.14	330,712.77	92.33
	medium-priority checkpoint writes	1.03	22,773,974	6,357.89	256,306.84	71.55
	aged writes by dbwr	1.02	22,650,295	6,323.37	235,935.35	65.87
***celadm01	scrub IO	76.25	153,474,696	42,846.09	2,398,042.13	669.47
	redo log write	7.55	15,194,120	4,241.80	387,254.92	108.11
	smart scan	6.55	13,182,622	3,680.24	1,651,006.60	460.92
	limit dirty buffer writes	3.45	6,944,945	1,938.85	86,823.82	24.24
	REQ list writes	1.63	3,276,019	914.58	27,304.96	7.62
	Internal IO	1.20	2,415,435	674.33	261,543.13	73.02
***celadm10	scrub IO	77.22	161,035,167	44,956.77	2,516,174.48	702.45
	redo log write	7.23	15,073,012	4,207.99	388,150.37	108.36
	smart scan	6.36	13,256,373	3,700.83	1,660,229.47	463.49
	limit dirty buffer writes	3.29	6,859,747	1,915.06	86,198.26	24.06
	REQ list writes	1.59	3,317,680	926.21	27,701.16	7.73
	Internal IO	1.08	2,255,257	629.61	248,914.79	69.49
***celadm11	Internal IO	36.89	21,552,847	6,016.99	1,671,865.83	466.74
	smart scan	20.47	11,960,257	3,338.99	1,575,214.93	439.76
	redo log write	15.01	8,768,125	2,447.83	185,314.59	51.73
	limit dirty buffer writes	11.73	6,851,995	1,912.90	87,662.70	24.47
	REQ list writes	5.75	3,357,205	937.24	28,120.29	7.85
	medium-priority checkpoint writes	3.26	1,902,422	531.11	21,629.33	6.04
	aged writes by dbwr	3.10	1,813,853	506.38	19,267.80	5.38
***celadm12	Internal IO	37.36	21,748,398	6,071.58	1,664,979.73	464.82
	smart scan	20.02	11,655,198	3,253.82	1,508,940.39	421.26
	redo log write	15.09	8,785,232	2,452.61	185,801.26	51.87
	limit dirty buffer writes	11.59	6,747,185	1,883.64	86,797.12	24.23
	REQ list writes	5.71	3,322,863	927.66	27,790.08	7.76
	medium-priority checkpoint writes	3.20	1,863,310	520.19	21,287.06	5.94
	aged writes by dbwr	3.12	1,817,180	507.31	19,253.05	5.37

[11]Smart Flash Log Write-Back は、Oracle Exadata System Software 20.1.0 で導入されました。

図 36 は、Internal IO の潜在的な原因を示しています。これは、すべてのセルを集計したものです。リンクは、AWR レポートの該当セクションを直接表示し、各セルの統計情報を確認できます。

図 36：Internal IO Reasons



Internal IO Reasons

- The following are possible reasons for Internal IO
- The values displayed are the total IOs over all cells

	Requests		MB	
Statistic	Total Requests	per Sec	Total MB	per Sec
Internal IO	66,206,025	18,482.98	5,859,557.84	1,635.83
Disk Writer reads	22,798,716	6,364.80	2,849,886.13	795.61
Disk Writer writes	8,568,449	2,392.09	2,588,040.08	722.51
Population	681,124	190.15	117,882.48	32.91
Metadata	31,955,625	8,921.17	249,653.32	69.70

Exadata Top Databases Consumers

図 37 の Top Databases By Requests - Details セクションでは、すべてのデータベースで Disk に対する Large IO の応答時間（Latency）が増加しています。これは Large IO の IORM キューイングの増加にもつながり、データベース側の *cell smart table scan* 待機イベントに直接影響します。

図 37：Top Databases By Requests - Details

Top Databases By Requests - Details

- Request details for the top databases by IO requests

			Small Requests							Large Requests						
			Reqs/s			Latency		Queue Time		Reqs/s			Latency		Queue Time	
DB Name	DBID	IOs/s	Total	Flash	Disk	Flash	Disk	Flash	Disk	Total	Flash	Disk	Flash	Disk	Flash	Disk
OTHER		471,429.20	464,754.32	9,008.30	455,746.02	51.80us	282.23us		204.17us	6,674.88	5,871.83	803.04	533.66us	19.94ms	919.00ns	
DB01		92,796.91	60,105.88	51,924.86	8,181.02	40.21us	272.21ms	1.00ns		32,691.03	32,458.66	232.37	686.36us	91.96ms	200.72us	490.39ms
DB02		18,860.88	12,690.71	11,575.37	1,115.35	52.76us	209.51ms	75.00ns	5.00ns	6,170.17	6,062.98	107.19	625.21us	51.54ms	273.14us	2,566.40ms
DB03(*)		12,878.98	7,513.95	6,694.89	819.05	67.67us	165.69ms	274.00ns	0.00ns	5,365.03	5,226.35	138.68	1.02ms	104.20ms	629.08us	2,000.53ms
DB04		6,359.14	6,092.04	5,323.34	768.69	42.14us	208.86ms	25.00ns		267.11	235.60	31.51	392.63us	99.62ms	1.87us	304.89ms
DB05		4,952.38	2,850.33	2,494.68	355.66	74.29us	126.80ms	880.00ns	224.39ns	2,102.05	1,953.13	148.92	329.24us	72.63ms	35.16us	657.51ms
DB07		1,782.76	1,407.48	1,288.09	119.39	94.70us	188.69ms	689.00ns		375.27	368.41	6.86	705.33us	147.00ms	361.06us	2,519.26ms
DB08		1,485.30	380.64	353.75	26.90	66.30us	136.40ms	6.00ns		1,104.65	1,094.84	9.81	573.96us	67.28ms	80.78us	650.33ms
DB09		1,129.34	977.70	917.39	60.31	54.12us	64.34ms			151.64	145.26	6.38	406.82us	45.33ms	64.84us	113.09ms
DB10		952.33	905.59	860.82	44.78	45.98us	103.34ms	20.00ns		46.73	41.32	5.42	415.04us	9.98ms	40.00ns	114.05ms

図 38 の Top Databases by IO Requests per Cell – Details セクションで各ストレージ・セルを詳しく見てみると、長い応答時間（Latency）と長い IORM キュー時間（Queue Time）が発生しているのは、celadm11 および celadm12 の 2 つのセルのみであることが分かります。これまで継続して確認してきた 2 つのセルです。

図 38：Top Databases by IO Requests per Cell - Details

Top Databases by IO Requests per Cell - Details

- Request details for the top databases per cell

				Small Requests							Large Requests						
				Reqs/s			Latency		Queue Time		Reqs/s			Latency		Queue Time	
Cell Name	DB Name	DBID	IOs/s	Total	Flash	Disk	Flash	Disk	Flash	Disk	Total	Flash	Disk	Flash	Disk	Flash	Disk
***celadm01	OTHER		0	43,438.80	43,136.51	232.56	42,903.95	42.85us	90.64us	179.51us	302.29	266.37	35.92	420.13us	216.49us	2.27us	
	DB01	3370969835	7,948.39	5,217.00	5,213.99	3.01	40.06us	2.89ms	2.00ns		2,731.39	2,717.88	13.51	650.31us	3.52ms	241.06us	60.19us
	DB02	3422047742	1,566.25	1,082.57	1,081.61	0.96	50.77us	6.68ms	77.00ns		483.68	479.91	3.77	652.21us	2.43ms	347.78us	42.84us
	DB03(*)	3517124528	1,124.03	645.27	634.11	11.16	64.53us	3.63ms	273.00ns		478.75	473.91	4.84	1.07ms	5.09ms	578.96us	81.87us
	DB04	4180093614	526.74	501.59	500.56	1.03	42.14us	913.19us	28.00ns		25.15	23.70	1.44	397.10us	249.55us	517.00ns	3.70us
***celadm10	OTHER		0	45,506.19	45,209.29	194.22	45,015.07	41.17us	90.51us	166.53us	296.90	260.73	36.17	423.13us	213.16us	1.37us	
	DB01	3370969835	7,971.13	5,206.28	5,203.27	3.02	39.78us	2.85ms	2.00ns		2,764.84	2,751.39	13.45	635.86us	3.37ms	229.60us	38.95us
	DB02	3422047742	1,574.99	1,093.53	1,092.47	1.06	50.91us	6.52ms	77.00ns		481.46	477.71	3.75	617.74us	2.35ms	263.74us	39.86us
	DB03(*)	3517124528	1,111.16	641.98	629.40	12.58	65.02us	3.11ms	266.00ns		469.18	464.43	4.75	0.97ms	4.40ms	751.57us	51.61us
	DB04	4180093614	526.18	501.52	500.61	0.92	42.09us	719.67us	37.00ns		24.66	23.22	1.44	386.71us	253.09us	3.76us	3.73us
***celadm11	DB01	3370969835	6,591.98	4,143.92	43.10	4,100.83	73.93us	266.62ms	20.00ns		2,448.05	2,401.70	46.35	866.38us	217.31ms	5.58us	1,115.89ms
	OTHER		0	5,814.81	3,934.23	3,473.59	460.64	54.48us	94.34ms	20.34ms	1,880.58	1,653.69	226.89	620.31us	37.26ms	34.00ns	
	DB02	3422047742	1,407.22	977.06	424.42	552.64	86.03us	203.31ms	161.00ns	0.00ns	430.16	395.67	34.49	654.73us	78.98ms	15.30us	4,531.93ms
	DB03(*)	3517124528	802.35	506.09	157.96	348.13	168.41us	188.54ms	797.00ns		296.26	248.43	47.83	0.98ms	155.24ms	46.76us	2,391.32ms
	DB04	4180093614	476.37	465.32	86.13	379.19	62.84us	205.82ms	76.00ns		11.05	2.40	8.65	690.81us	178.21ms	1.86us	480.95ms
***celadm12	DB01	3370969835	6,570.31	4,099.89	52.88	4,047.01	70.06us	280.08ms	17.00ns		2,470.41	2,420.99	49.42	895.83us	219.21ms	2.02us	1,259.04ms
	OTHER		0	5,755.89	3,984.28	3,487.51	496.77	54.54us	88.93ms	18.78ms	1,771.61	1,563.06	208.56	623.28us	35.84ms		
	DB02	3422047742	1,388.04	980.84	432.39	548.46	85.44us	221.09ms	144.00ns		407.19	375.18	32.01	694.95us	84.51ms	17.95us	3,709.89ms
	DB03(*)	3517124528	750.08	526.47	168.58	357.88	180.29us	194.72ms	919.00ns		223.61	181.22	42.39	1.19ms	160.34ms	60.84us	3,846.45ms
	DB04	4180093614	471.94	462.23	86.03	376.20	65.96us	219.28ms	72.00ns		9.71	1.26	8.45	637.07us	188.56ms	1.12us	644.36ms

分析のまとめ

今回の例の分析から得られた結果のまとめです。

- データベースでは IO パフォーマンスが低下しており、おもに *cell smart table scan* 待機イベントが顕著に発生しています。
- Flash Cache Configuration から、2 つのセルが *normal - flushing* のステータスにあることが確認されました。これは、フラッシュ・キャッシュがすべてのデータをハード・ディスクにフラッシュ処理（書出し）していることを意味し、これらのセル上の IO はフラッシ

ュ・キャッシュを利用するできない可能性があります。これら 2 つのセルの IO パターンに違いは、他の Flash Cache 関連のセクションでも明らかです。

- IO Outliers では、これらの 2 つのセルで IO の異常値が見られ、IO パターンは、書込み処理の増加を表しています。他のセルでは、スクラブ処理による Small Read 回数の増加が見られます。
- Smart IO でも、これら 2 つのセルにおける IO 処理方法の違いが再び示されています。
- IO Reasons セクションでは、これら 2 つのセルで異なる IO パターンが見られ、他のフラッシュ・キャッシュ関連のセクションで観察された内容と一致しています。
- Top Databases セクションでは、Large IO の応答時間の増加が確認されていて、その結果、これら 2 つのセルで IORM のキュー待機時間が増加しています。これらの待機時間が、cell smart table scan の待機イベントに直接影響を与えています。

データを確認したところ、主な問題はこれら 2 つのセル・ストレージ・サーバーのフラッシュ・キャッシュからのフラッシュ処理（書き出し）が実行されていたことにあると判断されました。フラッシュ処理（書き出し）は、アクティブなデータベース・ワークロードが流れている最中のシステムでは実行されるべきではありません。データがフラッシュ・キャッシュにキャッシュされる動作が停止してしまうためです。このケースでは、2 つのセルのフラッシュ・ログが構成されていないことを改善するためにメンテナンス作業が行われていましたが、誤ってワークロード負荷がピークの期間中に実行されたため、今回のパフォーマンスの問題が発生しました。この問題を解消するためには、ALTER FLASHCACHE CANCEL FLUSH コマンドを実行してフラッシュ処理（書き出し）をキャンセルする必要がありました。

Exadata のパフォーマンス・データ

AWR レポートの他にも、Exadata ではセル・メトリックや ExaWatcher など、多数のパフォーマンス・データを利用することができます。

表 5 は、それぞれのパフォーマンス・データの特徴と確認可能な情報をまとめたものです。

パフォーマンス・データ

表 5：Exadata 上で入手できるパフォーマンス・データ

AWR	セル・メトリック	ExaWatcher
特性		
• 広範囲に入手可能 • 通常はこれで十分 • 既存のデータベース・ツールと統合 • システム・レベルのビュー（すべてのセル）、セルごとのビューを提供 • レポート期間（デフォルトは 1 時間）での平均	• セルごとに収集 • 累積値と秒あたりの割合（1 分ごとに計算）を含む • 保存：7 日間 • より詳細なデータと保存期間の延長には、Real Time Insight を使用可能 [12]	• セルごとに収集 • 5 秒ごと • 保存：7 日間 • GetExaWatcherResults.sh でグラフの作成が可能
確認可能な情報		
• 構成情報 • OS の統計情報（iostat など） • セル・サーバーの統計 • Exadata のスマート機能 • IO Reasons • Top Databases	• Exadata のスマート機能（Smart Flash Cache、Smart Flash Log、IORM、Smart Scan など）	• OS の統計情報 • Cellsrvstat（Exadata のスマート機能）

まとめ

自動ワークロード・リポジトリ（AWR）は、Oracle Database で最も広く使用されているパフォーマンス診断ツールです。Exadata 上で稼働する Oracle データベースの AWR データには、Exadata の統計情報が追加されています。Exadata の統計情報を AWR に統合することで、データベースのパフォーマンスに問題が発生しても、一般的なインフラストラクチャにデータベースが配置されていた場合よりも格段に優れた分析を容易に実行できるようになっています。

詳しくは、*Oracle Exadata System Software ユーザーズ・ガイド – Exadata の監視*を参照してください。

[12] *Oracle Exadata System Software – Using Real-Time Insight* を参照

参照

1. Oracle Exadata System Software ユーザーズ・ガイド – Exadata の監視
2. Exadata の動作状態およびリソース使用率の監視
3. Exadata Health and Resource Utilization Monitoring - Exadata Database Machine KPIs
4. Exadata Health and Resource Utilization Monitoring - Adaptive Thresholds
5. Exadata Health and Resource Utilization Monitoring - System Baselining for Faster Problem Resolution
6. Oracle Exadata Cloud のための Oracle Enterprise Manager - 実装、管理、および監視のベスト・プラクティス
7. Enterprise Manager Oracle Exadata Database Machine スタート・ガイド

ORACLE

Connect with us

+1.800.ORACLE1までご連絡いただくか、**oracle.com**をご覧ください。北米以外の地域では、**oracle.com/contact**で最寄りの営業所をご確認いただけます。

blogs.oracle.com facebook.com/oracle twitter.com/oracle