ExadataでのAWRレポートの使用

AWRによるExadataのパフォーマンス診断

ホワイト・ペーパー / 2018年9月18日

ORACLE

免責事項

下記事項は、弊社の一般的な製品の方向性に関する概要を説明するものです。また、情報提供を唯一の目的とするものであり、いかなる契約にも組み込むことはできません。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。オラクルの製品に関して記載されている機能の開発、リリース、および時期については、弊社の裁量により決定されます。

目次

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 PCI フラッシュ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、超高速の InfiniBand 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、InfiniBand ネットワーキングにデータベース・インテリジェント機能を実装することで、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、またはサブスクリプション・モデルを使用して入手し、オラクルがすべてのインフラストラクチャを管理する Oracle Public Cloud または Oracle Cloud at Customer でデプロイすることができます。

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。このホワイト・ペーパーでは、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の概要を説明します。

このホワイト・ペーパーの内容は、デプロイ先がオンプレミスか、Oracle Public Cloud か、Oracle Cloud at Customer かに関係なく、すべての Exadata に該当します。特に、Exadata Cloud 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データベース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。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、管理するのがAWRの機能です。このデータ収集プロセスは一定の時間隔で繰り返され、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値から計算されるデルタ値は、この時間隔中の各統計の変化を表すもので、AWRレポートに表示して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、このスナップショットは8日間にわたって管理されます。AWRレポートは時間隔を指定してオンデマンドで生成することもできます。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。

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、必ず問題の対象範囲と一致するデータとツールを使用することです。

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レポートまたは SQL Detailsレポート[1]から取得できます。SQL Monitorレポートでは、SQL文またはDB操作の1回の実行に関する詳細な統計を確認でき、SQL Detailsレポートでは、指定した時間枠内に実行された1つのSQL文の複数回分の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とデータベース全体の問題の両方に使用できるうえ、局所的な問題にも使用できます。というのも、ASHでは複数ディメンションを横断してデータが収集され、ディメンションでデータをフィルタリングできるためです。

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。Oracle Public Cloud環境およびOracle Cloud at Customer環境ではお客様がストレージ・サーバーにアクセスすることはできないため、これらの環境でExadataを使用する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AWRグローバル・レポートでしか提供されません。テキスト形式のレポートで統計を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。また、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、このホワイト・ペーパーの参考資料の項で示します。

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、パフォーマンスの問題に対処するために、DB時間を多く消費しているものを見つけます。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。

---

[1] AWRを使用するにはOracle Diagnostics Packが必要です。SQL MonitorレポートおよびSQL DetailsレポートにはOracle Diagnostics and Tuning Packが必要です。

課題およびAWR EXADATAソリューション

Oracle DBAが極めて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の分析と理解を改善することです。この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が、このインフラストラクチャの構成が誤っていたり、一部のコンポーネントで障害が発生したりした場合は、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、それを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などのエンジニアド・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AWRに自動的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にデータベースがデプロイされていた場合に別のやり方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。ソフトウェア機能やハードウェア機能が追加されてExadataのコア・プラットフォームが強化されるのに合わせてExadata固有のAWRコンテンツが強化され続けているという事実も、OracleのDBAにはメリットです。

次の項からは、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。

ノイジー・ネイバー（うるさい隣人）

Exadataストレージでは、パフォーマンスの問題を分析するときの対象範囲が増えます。ストレージ・サブシステムは複数のデータベースに共有されている可能性があり、そのためストレージ・レイヤーに由来する統計はシステム全体のものとなります。つまり、1つのデータベースまたは1つのデータベース・インスタンスの統計ではないということです。

複数のデータベースが統合されているExadataシステムでは、システムのIO帯域幅を大量に消費しているがゆえにシステム上の他のデータベースに影響を与えていると考えられるデータベースを特定することが重要です。これは一般にクラウド・デプロイメント内のノイジー・ネイバーと呼ばれる問題です。ただし、Exadataに組み込まれているIOリソース管理（IORM）機能を利用すると、Exadata Storage Server内のIOリクエストに優先順位を付け、構成済みのリソース・プランに基づいてスケジューリングができるため、IORMの活用を強く推奨します。IORMについて詳しくは、『Oracle Exadata System Softwareユーザーズ・ガイド』の"I/Oリソース管理の理解"を参照してください。

このノイジー・ネイバー問題に対処できるように、AWRレポートにはTop Databases by IO Requests（IOリクエスト件数上位データベース）セクション[2]が含まれており、IO Requests（IOリクエスト件数）とIO Throughput (MB)（IOスループット MB/s）の両方が表示されます。AWRのスナップショット時に、データベースのサブセットがAWRに取得されます。取得されるデータベースは、各ストレージ・サーバー内の上位N個を特定する内部メトリックに基づいて決まります。図1に示すように、レポート時にデータが集計され、IOリクエストおよびIOスループットからみた上位データベースがレポートされます。また、データはFlash（フラッシュ・デバイス上）のIOとDisk（ハード・ディスク上のIO）とに分けて表示されます。

Top Databases by IO Requests

- The top databases by IO Requests are displayed
- At most 10 databases are 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
***074IR	2784208988	64	130,710,242	18,139.08	101,962,540	28,747,702	3,448,493.83	478.56	2,742,990.47	705,503.36
***074PR	1423592318	19	39,086,267	5,424.13	28,100,373	10,985,894	1,145,975.19	159.03	369,402.64	776,572.54
***087PR	4249955985	7	14,788,183	2,052.20	10,812,944	3,975,239	238,509.77	33.10	130,012.12	108,497.64
OTHER	0	4	8,952,348	1,242.35	8,867,303	85,045	163,349.91	22.67	86,751.92	76,597.98
***004PR	3271184140	3	6,252,671	867.70	4,458,922	1,793,748	85,240.87	11.83	60,779.42	24,461.45
***089PR	3833642602	1	3,036,429	421.38	2,553,247	483,182	322,062.77	44.69	90,101.66	231,961.12
***085PR	1146478208	1	2,152,209	298.67	1,566,173	586,036	202,850.86	28.15	54,280.02	148,570.84
***059PR	16998699	0	0	0.00	0	0	0.00	0.00	0.00	0.00
***027PR	498900796	0	0	0.00	0	0	0.00	0.00	0.00	0.00
***052PR	3053934856	0	0	0.00	0	0	0.00	0.00	0.00	0.00

図1.Top Databases by IO requests（IOリクエスト件数順上位データベース）の例。レポートには%Totalではなく%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によって取得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（上の図）と、セルごとに集計したものを使用できます

[2] 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$cell_dbビューと、v$cell_dbのデータを取得する後続のAWRビューにデータベース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。

セルまたはディスク上のワークロードの不均衡

Exadataは並列システムであり、想定では、すべてのストレージ・サーバーとディスクにワークロードが均等に分散されることになっています。他のストレージ・サーバーやディスクよりも多くの作業を実行しているストレージ・サーバーやディスクがある場合は、パフォーマンスの問題がそこで発生する可能性があります。並列システムでよくあるように、システムの処理速度はもっとも遅いコンポーネントと同程度にしかなりません。

Exadataレポートでは、多数のメトリックを使用してデバイス同士を比較する簡単な外れ値分析を実行します。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(Outlier)分析に使用される統計にはOSの統計が含まれます。これはiostatと同様で、IOPs、スループット、使用率、サービス時間、キュー時間などが含まれています。セル・サーバーの統計もここに含まれ、IOPs、IO MB/s（スループット）、待機時間が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。

外れ値(Outlier)分析の他に、Exadata AWRレポートでは、システムが最大容量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セルから問い合わせられるもので、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変化するため、レポートで使用されるこれらの最大値は、厳格なルールではなくガイドラインとして使用する必要があります。

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

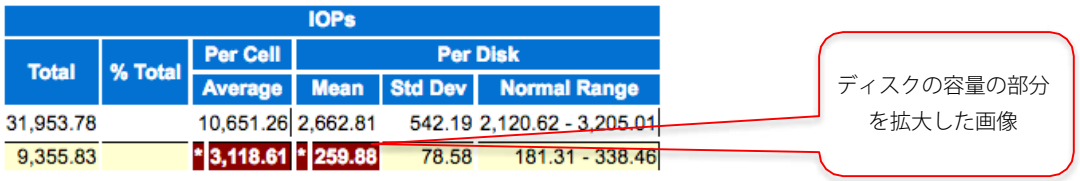


図2a.簡単な外れ値分析の例。この例に外れ値はありませんが、ハード・ディスクのIOPS容量が最大に達している可能性があることが特定され、（*）と濃い赤色の背景で示されています。ハード・ディスクに対するシステムの最大値は6,408 IOPS、レポートの現在の表示は9,355.83 IOPSとなっています

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		

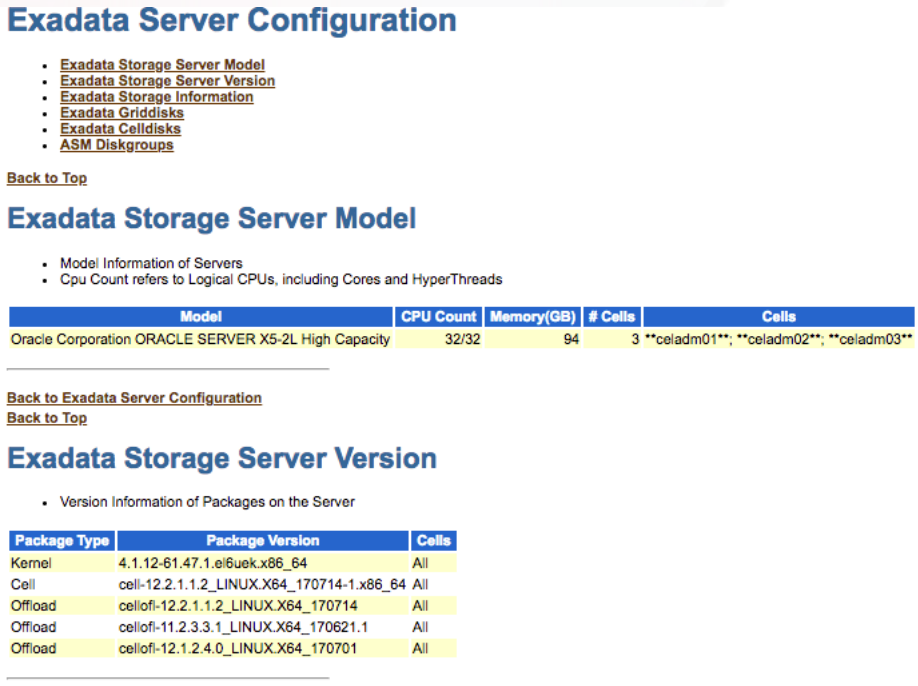
拡大した図

図2b.簡単な外れ値分析の例。この例では、ハード・ディスクの容量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

構成の相違

ストレージ・サーバーまたはディスク上のワークロードが不均衡な場合と同様、ストレージ・サーバー同士で構成が異なっている場合も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、フラッシュ・キャッシュやフラッシュ・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。

図3および図3aに示すように、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。



Exadata Server Configuration

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

Back to Top

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 X5-2L High Capacity	32/32	94	3	**celadm01**; **celadm02**; **celadm03**

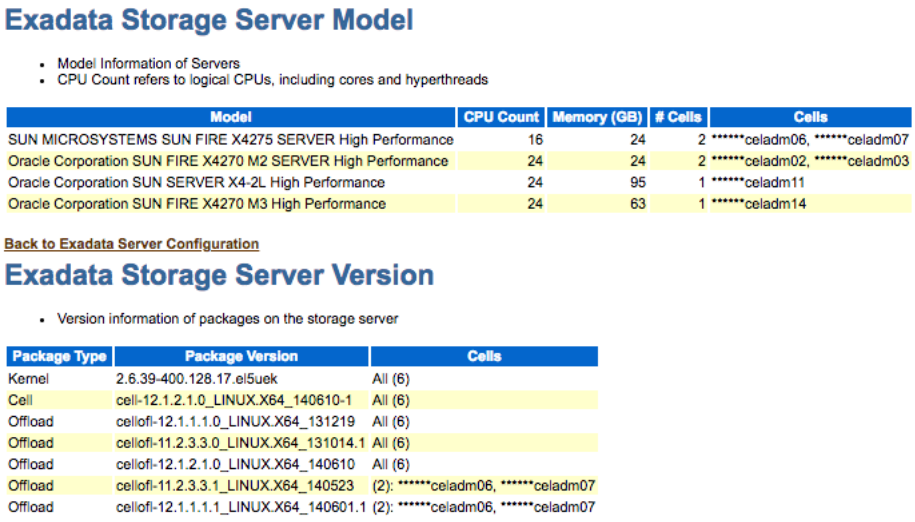
Back to Exadata Server Configuration
Back to Top

Exadata Storage Server Version

- Version Information of Packages on the Server

Package Type	Package Version	Cells
Kernel	4.1.12-61.47.1.el6uek.x86_64	All
Cell	cell-12.2.1.1.2_LINUX.X64_170714-1.x86_64	All
Offload	cellofl-12.2.1.1.2_LINUX.X64_170714	All
Offload	cellofl-11.2.3.3.1_LINUX.X64_170621.1	All
Offload	cellofl-12.1.2.4.0_LINUX.X64_170701	All

図3.取得された構成情報の例。Storage Server Modelではセルの名前が表示されます。'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ます



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
SUN MICROSYSTEMS SUN FIRE X4275 SERVER High Performance	16	24	2	******celadm06, ******celadm07
Oracle Corporation SUN FIRE X4270 M2 SERVER High Performance	24	24	2	******celadm02, ******celadm03
Oracle Corporation SUN SERVER X4-2L High Performance	24	95	1	******celadm11
Oracle Corporation SUN FIRE X4270 M3 High Performance	24	63	1	******celadm14

Back to Exadata Server Configuration

Exadata Storage Server Version

- Version information of packages on the storage server

Package Type	Package Version	Cells
Kernel	2.6.39-400.128.17.el5uek	All (6)
Cell	cell-12.1.2.1.0_LINUX.X64_140610-1	All (6)
Offload	cellofl-12.1.1.1.0_LINUX.X64_131219	All (6)
Offload	cellofl-11.2.3.3.0_LINUX.X64_131014.1	All (6)
Offload	cellofl-12.1.2.1.0_LINUX.X64_140610	All (6)
Offload	cellofl-11.2.3.3.1_LINUX.X64_140523	(2): ******celadm06, ******celadm07
Offload	cellofl-12.1.1.1.1_LINUX.X64_140601.1	(2): ******celadm06, ******celadm07

図3a.種類の異なるシステムから取得された構成情報の例。Storage Server Modelではモデルごとにセルの名前が表示されます。'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ます。構成に相違があれば、セルの名前が表示されます。Exadata Storage Server Versionのセクションを確認してください

高負荷

システムにかかる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、IO負荷の増加またはストレージ・サーバー上のCPUの増加があります。IOの負荷が増加する原因としては、バックアップなどの保守行為、ユーザー・ワークロードの増加や実行計画の変更などによるユーザーIOの変化などが考えられます。

ExadataシステムではIOごとに追加情報が送信されますが、データベースでIOが実行されている理由もそこに含まれます。IOの理由が含まれているため、IOの負荷が増えた原因が保守行為なのか、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、スマート・スキャン、スマート・フラッシュ・ログ、スマート・フラッシュ・キャッシュといったExadataのスマート機能に関する情報も表示されます。

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

図4.各ストレージ・セルについてIOの理由をIOリクエスト件数順に表示した例。これを見ると、通常のデータベース・ワークロード（REDOログの書込みやバッファ・キャッシュの読取り）からIOリクエストが発生していることがわかります

Exadata固有のAWRデータの分析

AWRのさまざまなセクションのことがよくわかるように、実際のお客様のユースケースを例にして説明します。

このお客様は、4つのコンピュート・ノードを搭載したフル・ラック構成のExadata X5-2にデータベースをデプロイしたばかりですが、その直後からパフォーマンスの問題が発生し始めました。次の項からは、レポートに含まれるExadata固有のセクションを分析することで簡単に実行できる診断について概説します。

データベースの統計情報の確認

最初に行ったチェックは、システムのIO特性の検証です。図5は、1つのインスタンスで時間を要したイベントの上位のものを示していますが、このケースでは4つのインスタンスはどれもかなり似通っていました。待機イベントを見ると、DB時間のほぼ75％がcell single block physical read（セルのシングル・ブロック物理読取り）に費やされ、平均待機時間が8.32ミリ秒だったことがわかります。この読取り待機時間は、フラッシュ・キャッシュではなくハード・ディスクからデータが読み取られていることを示している可能性があります。

Top 10 Foreground Events by Total Wait Time

Event	Waits	Total Wait Time (sec)	Wait Avg(ms)	% DB time	Wait Class
cell single block physical read	238,950	1989.2	8.32	74.4	User I/O
DB CPU		560.9		21.0	
control file sequential read	73,397	24.5	0.33	.9	System I/O
enq: WF - contention	174	21.7	124.67	.8	Other
SQL*Net more data to client	58,582	19.7	0.34	.7	Network
reliable message	109,910	18.8	0.17	.7	Other
PX Deq: Slave Session Stats	57,477	11	0.19	.4	Other
gc current block busy	17,420	10.3	0.59	.4	Cluster
enq: PS - contention	56,927	8.4	0.15	.3	Other
REPL Capture/Apply: miscellaneous	4	5.8	1459.37	.2	Other

図5.AWRレポートのTop 10 Foreground Events by Total Wait Time（合計待機時間順上位10フォアグラウンド・イベント）

データベースの統計情報を詳しく見ていくと、このインスタンスから発行されたIOが154.6 IOPSと、比較的少量であったことがわかりました（図6）。これでは4つのインスタンス全体でも合計IOPSが600程度にしかなりません。

ところが、図6には、Optimized Requests（最適化されたリクエスト）がほぼ0であることも示されています。Exadataに対するOptimized IO（最適化されたIO）には以下が含まれます。

- スマート・フラッシュ・キャッシュに対するIO
- スマート・スキャンからのIO（ストレージ索引によって保存されたIO、列指向キャッシュによって保存されたIOを含む）

このデータベースはスマート・スキャン・ワークロードを実行していなかったようです。そのため、Optimized IO（最適化されたIO）が欠如しているという点は、「IOでフラッシュ・キャッシュが使用されていない」という仮説の裏付けを強化するものといえます。

IO Profile

	Read+Write Per Second	Read per Second	Write Per Second
Total Requests:	154.6	108.2	46.4
Database Requests:	82.5	61.8	20.8
Optimized Requests:	0.8	0.8	0.0
Redo Requests:	32.1	8.6	23.5
Total (MB):	18.9	18.5	0.4
Database (MB):	0.7	0.5	0.2
Optimized Total (MB):	0.0	0.0	0.0
Redo (MB):	16.3	16.2	0.1
Database (blocks):	88.8	62.9	25.9
Via Buffer Cache (blocks):	87.3	61.7	25.6
Direct (blocks):	1.5	1.2	0.3

図6.このインスタンスの最小IOが表示されたAWRレポートのIOプロファイル

図7は、データベース・インスタンスが発行しているIOのタイプを示しています。このリストを見る限り、IOのタイプに関してデータベースによるフラッシュ・キャッシュの使用を妨げそうなものは特に見当たりません。読取りの大半はBuffer Cache Reads（バッファ・キャッシュの読取り）で、これはバッファ・キャッシュに移入するためにデータベースによって実行されるディスク読取りです。このような読取りは通常、フラッシュ・キャッシュを使って行われるべきです。

IOStat by Function summary

- 'Data' columns suffixed with M,G,T,P are in multiples of 1024 other columns suffixed with K,M,G,T,P are in multiples of 1000
- ordered by (Data Read + Write) desc

Function Name	Reads: Data	Reqs per sec	Data per sec	Writes: Data	Reqs per sec	Data per sec	Waits: Count	Avg Tm(ms)
Streams AQ	66.1G	15.57	17.245M	0M	0.00	0M	44.9K	3.78
Others	3.1G	30.92	.805M	498M	2.24	.127M	102K	0.54
Buffer Cache Reads	1.8G	60.93	.481M	0M	0.00	0M	227.3K	8.71
DBWR	0M	0.00	0M	785M	20.52	.2M	6	1.67
LGWR	0M	0.01	0M	363M	23.48	.092M	92.2K	0.21
Direct Reads	32M	0.74	.008M	4M	0.14	.001M	0	
Smart Scan	4M	0.00	.001M	0M	0.00	0M	0	
Direct Writes	0M	0.00	0M	2M	0.04	.001M	22	0.05
TOTAL:	71.1G	108.17	18.541M	1.6G	46.42	.421M	466.3K	4.77

図7.フラッシュ・キャッシュにキャッシュされるべき通常のデータベースIOが表示された、AWRレポートのIOStat by Function summary（IOStatのファンクション別概要）

IOパフォーマンスに関する問題を確認したら、次はExadataの構成と統計の分析に移ります。

Oracle Exadataの構成

Exadataの構成のセクションを見ると、ストレージ・サーバーを14台搭載したフル・ラック構成のX5-2であることがわかります（図8）。残りの構成のセクションと動作状態のセクションからは何も異常は見つかりませんでした。すべてのストレージ・サーバーが、想定どおりの構成で同じように構成されていました。アラートはなく、オフラインになっているディスクもありませんでした。簡略化するため、これらのセクションはこのホワイト・ペーパーに掲載していません。

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 X5-2L High Capacity	32/32	95	14	*****celadm01; *****celadm02; *****celadm03; *****celadm04; *****celadm05; *****celadm06; *****celadm07; *****celadm08; *****celadm09; *****celadm10; *****celadm11; *****celadm12; *****celadm13; *****celadm14

図8.フル・ラック構成のX5-2であることが表示された、AWRレポートのExadata Storage Server Model

IOの分布

ストレージ・サーバーに対するIOを確認すると、ストレージ・サーバーに対するIOはさほど発生していないことがわかります（図9）。ところが、ハード・ディスクのIOはセルあたり564.28IOPsと、フラッシュ・デバイスでのセルあたり119.80IOPsより多くなっています。Exadata環境では通常ほとんどのIOがフラッシュで発生するため、このような分布は一般的ではありません。

デバイスのタイプが異なるとパフォーマンス特性も当然異なると考えられるため、Outlier（外れ値）のセクションでは、デバイス・タイプごとにIOがレポートされます。デバイス・タイプの特定に使用される形式は<FまたはH>/<サイズ>です（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）。

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 280.56 (1% of maximum capacity of 28,056)
- Outliers for flash disks are displayed when Flash Disk IOPs exceeds 5000.24 (1% of maximum capacity of 500,024)
- Outliers for flash disks will not be displayed. There are only 1,690 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
- % 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/1.5T	All	14	56	1,677.26		119.80	29.95	12.16	17.79 - 42.11	115.97	28.99	11.50	17.49 - 40.49	1.55	0.39	0.66	0.00 - 1.04	2.29	0.57	2.23	0.00 - 2.80	0.00	0.00	0.00	0.00 - 0.00
H/7.2T	All	14	168	7,899.87		564.28	47.02	10.75	36.28 - 57.77	133.66	11.15	6.67	4.49 - 17.82	429.31	35.78	7.62	28.16 - 43.39	0.15	0.01	0.11	0.00 - 0.13	0.96	0.08	0.27	0.00 - 0.35

				IOPs					
				Total	% Total	Per Cell	Per Disk		
Disk Type	Cell Name	# Cells	# Disks			Average	Mean	Std Dev	Normal Range
F/1.5T	All	14	56	1,677.26		119.80	29.95	12.16	17.79 - 42.11
H/7.2T	All	14	168	7,899.87		564.28	47.02	10.75	36.28 - 57.77

拡大した図

Small Reads/s				Small Writes/s				Large Reads/s			
Per Cell	Per Disk			Per Cell	Per Disk			Per Cell	Per Disk		
Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range	Average	Mean	Std Dev	Normal Range
115.97	28.99	11.50	17.49 - 40.49	1.55	0.39	0.66	0.00 - 1.04	2.29	0.57	2.23	0.00 - 2.80
133.86	11.15	6.67	4.49 - 17.82	429.31	35.78	7.62	28.16 - 43.39	0.15	0.01	0.11	0.00 - 0.13

図9.AWRレポートに示されたセル・サーバーのIOPS

スマート・スキャン

次に行ったのは、スマート・スキャンのセクション（図10）の確認です。ディスク読取りがスマート・スキャンによるものかどうかを判定することが目的ですが、その可能性は低そうです。というのも、図9を見ると、ほとんどのIOがSmall Reads/s（1秒あたりの小規模読取り回数）（図9には、Small Reads/s（1秒あたりの小規模読取り）が133.88と表示されています）ですが、スマート・スキャンはLarge Reads/s（大規模読取り）として観察されるのが一般的だからです（図9には1秒あたりのLarge Reads/s（大規模読取り）が0.15と表示されています）。スマート・スキャンの対象となったデータはセルあたりわずか1.5 MB/秒であるため、図10のスマート・スキャンのセクションを簡単に見ただけで、IOの発生源がスマート・スキャンではなかったことを確認できます。

Smart IOのセクションにはシステム上のスマートIOアクティビティの全体像も表示されるため、スマート・スキャンの実行状況が良好かどうかや、ストレージ・サーバーがCPUバウンドになっていないかどうかの判断ができます。ただし、ごく一部のSQL文に限ったスマート・スキャンの問題を調査する場合は、SQL*Monitorレポートのほうが、使用するパフォーマンス診断ツールとして適しています。

Smart IO

- These statistics are collected by the cells and are not restricted to this database or instance
- The top cells by Total MB Requested are displayed
- 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 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

	MB Requested			Storage Index		Flash Cache		Offload		Passthru		Reverse Offload	
Cell Name	% Total	Total	per Sec	MB	% Optimized	MB	% Optimized	MB	% Efficiency	MB	% Passthru	MB	% ReverseOffload
All		82,802.22	21.10	0.00	0.00	0.00	0.00	82,050.46	99.09	0.00	0.00	0.00	0.00
*****celadm03	7.17	5,939.99	1.51	0.00	0.00	0.00	0.00	5,889.19	99.14	0.00	0.00	0.00	0.00
*****celadm08	7.16	5,932.31	1.51	0.00	0.00	0.00	0.00	5,871.12	98.97	0.00	0.00	0.00	0.00
*****celadm09	7.16	5,930.88	1.51	0.00	0.00	0.00	0.00	5,882.45	99.18	0.00	0.00	0.00	0.00
*****celadm01	7.16	5,926.45	1.51	0.00	0.00	0.00	0.00	5,878.62	99.19	0.00	0.00	0.00	0.00
*****celadm04	7.15	5,922.05	1.51	0.00	0.00	0.00	0.00	5,863.57	99.01	0.00	0.00	0.00	0.00
*****celadm06	7.14	5,914.13	1.51	0.00	0.00	0.00	0.00	5,857.17	99.04	0.00	0.00	0.00	0.00
*****celadm02	7.14	5,912.31	1.51	0.00	0.00	0.00	0.00	5,864.23	99.19	0.00	0.00	0.00	0.00
*****celadm13	7.14	5,912.06	1.51	0.00	0.00	0.00	0.00	5,857.04	99.07	0.00	0.00	0.00	0.00
*****celadm14	7.14	5,911.75	1.51	0.00	0.00	0.00	0.00	5,860.00	99.12	0.00	0.00	0.00	0.00
*****celadm07	7.14	5,911.69	1.51	0.00	0.00	0.00	0.00	5,854.32	99.03	0.00	0.00	0.00	0.00
*****celadm11	7.13	5,903.44	1.50	0.00	0.00	0.00	0.00	5,853.50	99.15	0.00	0.00	0.00	0.00
*****celadm10	7.13	5,902.24	1.50	0.00	0.00	0.00	0.00	5,842.91	98.99	0.00	0.00	0.00	0.00
*****celadm12	7.12	5,895.19	1.50	0.00	0.00	0.00	0.00	5,843.38	99.12	0.00	0.00	0.00	0.00
*****celadm05	7.11	5,887.73	1.50	0.00	0.00	0.00	0.00	5,832.96	99.07	0.00	0.00	0.00	0.00

図10.スマート・スキャンの情報が表示された、AWRレポートのスマートIO

Flash Cache

データベースの待機イベントを確認して最初に疑ったのは、フラッシュ・キャッシュではなくハード・ディスクからデータが読み取られていたのではないか、ということでした。その場合は、フラッシュ・キャッシュのヒット率が低くなっていると予測されます。フラッシュ・キャッシュのヒット率の定義と計算方法は、表1を参照してください。

図11を見ると、Cell OLTP Hit%（セルOLTPのヒット率）（およびCell Scan Hit%（セル・スキャンのヒット率））が非常に高く、ほぼ100%となっています。ところが、Database Flash Cache Hit%（データベースのフラッシュ・キャッシュのヒット率）はほぼ0です。このような違いが生じているのはなぜでしょうか。

Flash Cache Savings

- Disk write savings (overwrites) - writes absorbed by flash cache that would have otherwise gone to disk
- Database Flash Cache Hit% - for the database, not restricted to an instance
- Cell OLTP and Cell Scan Flash Cache Hit% - for the cells, not restricted to this database or instance

Database Flash Cache Hit %	.08
Cell OLTP Hit %	99.7
Cell Scan Hit %	99.9
Disk Write savings/s	2.49

図11.AWRレポートのFlash Cache Savings（フラッシュ・キャッシュによる節約）

統計情報	説明
Database Flash Cache Hit%（データベースのフラッシュ・キャッシュのヒット率）	データベースからの読取りリクエストのうち、フラッシュ・キャッシュで処理されたものの割合
Cell OLTP Hit%（セルOLTPのヒット率）	ストレージ・サーバーに対するOLTP読取りリクエストのうち、フラッシュ・キャッシュで処理されたものの割合。これは次の計算式で求められます。 $100 * \frac{\text{フラッシュ・キャッシュによる読取りリクエストのヒット数}}{\text{フラッシュ・キャッシュによる読取りリクエストのヒット数} + \text{フラッシュ・キャッシュのキャッシュ・ミス}}$
Cell Scan Hit%（セル・スキャンのヒット率）	ストレージ・サーバーに対するスキャン・リクエストのうち、フラッシュ・キャッシュで処理されたものの割合。これは次の計算式で求められます。 $100 * \frac{\text{フラッシュ・キャッシュによるスキャンの読取りバイト数}}{\text{フラッシュ・キャッシュが読取りを試行したバイト数}}$

表1.フラッシュ・キャッシュのヒット率の定義

セルのヒット率は、フラッシュ・キャッシュへの***キャッシング対象***である読取り件数に基づいて計算されます。Exadataには、データベースから発行されるIOリクエストのタイプを区別でき、データをフラッシュ・キャッシュにキャッシングすることにメリットがあるかどうかを判断できる***Smart***Flash Cacheがあります。セルのヒット率には、フラッシュ・キャッシュにキャッシングするメリットがないデータを取得した読取りの件数は含まれません。

ただし、データベースのフラッシュ・キャッシュのヒット率は、データベースから発行されるすべてのIOリクエストを対象としています。

この2つの間の差は、データベースからの読取りがフラッシュ・キャッシュの***キャッシング対象ではない***ことを示すものと言えそうです。

セルからのSmall Reads Distribution（小規模読取りの分布）（図12）を確認すると、Small Reads Distribution（小規模読取り）はフラッシュとハード・ディスクにほぼ均等に分割され、53.58 %の小規模読取りがハード・ディスク上で行われたことがわかります。

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
- 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

				% of Total Waits														
	Total Waits	FG Waits	Average(ms)	<64us	<128	<256	<512	<1ms	<2ms	<4ms	<8ms	<16ms	<32ms	<64ms	<128ms	<256ms	<512ms	<1s	>=1s
cell single block physical read	5,664,532	5,619,326	8.46					7.21	0.52	9.89	38.82	37.81	4.58	0.90	0.16	0.06	0.04		

Small Reads Distribution	%Small Reads	Small Reads/s
Flash	46.42	1,623.57
Disk	53.58	1,874.00

Small Reads - Flash	# Cells	Total Small Reads/s	Cell Small Reads/s	Disk Small Reads/s	Latency (ms)
F/1.5T	14	1,623.57	115.97	28.99	0.10

Small Reads - Disk	# Cells	Total Small Reads/s	Cell Small Reads/s	Disk Small Reads/s	Latency (ms)
H/7.2T	14	1,874.00	133.86	11.15	6.93

Small Reads Distribution	%Small Reads
Flash	46.42
Disk	53.58

Small Reads - Flash	# Cells
F/1.5T	14

Small Reads - Disk	# Cells
H/7.2T	14

拡大した図

図12.OLTPワークロードを示すAWRレポートのcell single block physical read（シングル・ブロック読取り）

図13のFlash Cache User Reads Per Second（1秒あたりのフラッシュ・キャッシュ・ユーザー読取り）を見ると、フラッシュ・キャッシュからの読取り件数が非常に少ないことがわかります。これにより、「IOリクエストがフラッシュ・キャッシュの対象になっていないためフラッシュ・キャッシュで読取りが発生していない」という仮説の裏付けがさらに強化されます。

Flash Cache User Reads Per Second

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

	Read Requests per Second						Read MB per Second				
Cell Name	Total	OLTP	Scan	Columnar	Keep	Misses	Total	OLTP	Scan	Columnar	Keep
All	4.91	4.77	0.15			0.01	0.14	0.07	0.07		
*****celadm10	2.60	2.58	0.01			0.00	0.05	0.04	0.01		
*****celadm02	1.01	1.00	0.01			0.00	0.02	0.02	0.00		
*****celadm14	0.15	0.14	0.01			0.00	0.01	0.00	0.00		
*****celadm12	0.14	0.13	0.01			0.00	0.01	0.00	0.01		
*****celadm06	0.14	0.13	0.01			0.00	0.01	0.00	0.00		
*****celadm03	0.12	0.10	0.01			0.00	0.01	0.00	0.01		
*****celadm04	0.11	0.11	0.01			0.01	0.00	0.00	0.00		
*****celadm08	0.11	0.10	0.01			0.00	0.01	0.00	0.01		
*****celadm07	0.11	0.09	0.02			0.00	0.01	0.00	0.01		
*****celadm13	0.11	0.10	0.01			0.00	0.01	0.00	0.00		
*****celadm05	0.10	0.09	0.01			0.00	0.01	0.00	0.01		
*****celadm01	0.09	0.08	0.01			0.00	0.00	0.00	0.00		
*****celadm09	0.08	0.07	0.01			0.00	0.01	0.00	0.00		
*****celadm11	0.06	0.05	0.01			0.00	0.01	0.00	0.00		

図13.AWRレポートのFlash Cache User Reads（フラッシュ・キャッシュ・ユーザー読取り）

図14は個々のセルの%Hit（ヒット率）を示しています。これを見ると、フラッシュ・キャッシュの動作はどのセルでも似通っていることがわかります。ストレージ・セルではフラッシュ・キャッシュからの読取りがほとんどありませんが、%Hitは高くなっています。つまり、ディスクIOが発生する原因はフラッシュ・キャッシュのキャッシュ・ミスではなく、意図的にフラッシュ・キャッシュをバイパスしているIOにあるということです。

Flash Cache User Reads Efficiency

- These statistics are collected by the cells and are not restricted to this database or instance
- The top cells by Total Read Requests are displayed

	Total		OLTP			Scan			Columnar				Keep		
Cell Name	Requests	MB	Read Requests	Misses	%Hit	Read MB	Attempted MB	%Hit	Read MB	Eligible MB	Saved MB	% Efficiency	Read Requests	Misses	%Hit
All	19,290	544.03	18,716	57	99.70	273.68	273.95	99.90							
*****celadm10	10,192	182.73	10,141	1	99.99	24.50	24.50	100.00							
*****celadm02	3,952	77.78	3,912	0	100.00	18.51	18.51	100.00							
*****celadm14	600	24.09	564	0	100.00	17.56	17.56	100.00							
*****celadm12	560	27.37	513	4	99.23	21.71	21.71	100.00							
*****celadm06	531	23.50	491	7	98.59	18.50	18.50	100.00							
*****celadm03	453	25.28	409	12	97.15	21.13	21.13	100.00							
*****celadm04	440	16.16	416	21	95.19	11.41	11.41	100.00							
*****celadm08	433	27.66	384	0	100.00	23.30	23.57	98.87							
*****celadm07	432	35.18	372	7	98.15	29.73	29.73	100.00							
*****celadm13	416	21.84	378	0	100.00	17.92	17.92	100.00							
*****celadm05	402	24.53	358	0	100.00	21.02	21.02	100.00							
*****celadm01	334	17.18	305	1	99.67	14.23	14.23	100.00							
*****celadm09	303	20.98	267	0	100.00	17.35	17.35	100.00							
*****celadm11	242	19.77	206	4	98.10	16.80	16.80	100.00							

図14.AWRレポートのFlash Cache User Reads Efficiency（フラッシュ・キャッシュからのユーザー読取りの効率）

図15はFlash Cache User Writes（フラッシュ・キャッシュへのユーザー書込み）を表示したものです。これを見ても、ストレージ・セルでの書込み量が最小限であることがわかります。これは、データベースから発行される読取りと書込みの両方がなんらかの理由でフラッシュ・キャッシュの対象になっていないことを示していると考えられます。

Flash Cache User 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
- Total - total number of write requests or write megabytes to Flash Cache

	Write Requests								Write Megabytes							
	Total				per Sec				Total				per Sec			
Cell Name	Total	First Writes	Overwrites	Keep	Total	First Writes	Overwrites	Keep	Total	First Writes	Overwrites	Keep	Total	First Writes	Overwrites	Keep
All	12,395	2,608	9,787		3.16	0.66	2.49		206.90	54.40	152.51		0.05	0.01	0.04	
*****celadm04	2,090	295	1,795		0.53	0.08	0.46		32.75	5.01	27.74		0.01	0.00	0.01	
*****celadm02	2,036	105	1,931		0.52	0.03	0.49		32.59	2.50	30.09		0.01	0.00	0.01	
*****celadm03	1,969	162	1,807		0.50	0.04	0.46		32.33	3.63	28.70		0.01	0.00	0.01	
*****celadm08	940	197	743		0.24	0.05	0.19		14.65	4.14	10.52		0.00	0.00	0.00	
*****celadm10	771	120	651		0.20	0.03	0.17		12.46	2.38	10.08		0.00	0.00	0.00	
*****celadm09	763	115	648		0.19	0.03	0.17		13.80	2.84	10.97		0.00	0.00	0.00	
*****celadm06	722	384	338		0.18	0.10	0.09		10.90	6.51	4.39		0.00	0.00	0.00	
*****celadm14	634	260	374		0.16	0.07	0.10		11.09	5.56	5.53		0.00	0.00	0.00	
*****celadm12	558	317	241		0.14	0.08	0.06		9.80	5.97	3.84		0.00	0.00	0.00	
*****celadm01	543	86	457		0.14	0.02	0.12		10.72	3.05	7.67		0.00	0.00	0.00	
*****celadm11	443	231	212		0.11	0.06	0.05		8.62	4.48	4.14		0.00	0.00	0.00	
*****celadm05	323	155	168		0.08	0.04	0.04		6.02	3.67	2.35		0.00	0.00	0.00	
*****celadm13	306	88	218		0.08	0.02	0.06		4.98	2.19	2.80		0.00	0.00	0.00	
*****celadm07	297	93	204		0.08	0.02	0.05		6.17	2.48	3.70		0.00	0.00	0.00	

図15.AWRレポートのFlash Cache User Writes（フラッシュ・キャッシュへのユーザー書込み）

Flash Cache のセクションで行う最後のチェックとして、Flash Cache Internal Writes（フラッシュ・キャッシュ内部書込み）のセクションも確認します。ここに表示されるのはPopulation Write Requests（フラッシュへの移入書込みリクエスト）です。フラッシュ・キャッシュでキャッシュ・ミスが発生すると、通常はPopulation Write（移入書込み）が行われるため、同じデータに対する後続のリクエストではフラッシュ・キャッシュがヒットします。この事例では、フラッシュ・キャッシュへのPopulation Write （移入書込み）も非常に少量です。

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 Write - population writes due to read misses

	Population Write Requests				Population Write Megabytes			
Cell Name	Total	per Sec	Columnar	Keep	Total	per Sec	Columnar	Keep
All	55	0.01			2.58	0.00		
*****celadm04	20	0.01			1.02	0.00		
*****celadm02	0	0.00			0.00	0.00		
*****celadm03	11	0.00			0.33	0.00		
*****celadm08	1	0.00			0.03	0.00		
*****celadm10	1	0.00			0.03	0.00		
*****celadm09	0	0.00			0.00	0.00		
*****celadm06	6	0.00			0.29	0.00		
*****celadm14	0	0.00			0.00	0.00		
*****celadm12	4	0.00			0.19	0.00		
*****celadm01	1	0.00			0.06	0.00		
*****celadm11	4	0.00			0.25	0.00		
*****celadm05	0	0.00			0.00	0.00		
*****celadm13	0	0.00			0.00	0.00		
*****celadm07	7	0.00			0.38	0.00		

図16.AWRレポートのFlash Cache Internal Writes（フラッシュ・キャッシュ内部書込み）

Flash Cacheのセクションのデータはどれも、フラッシュ・キャッシュのアクティビティが非常に少ないこと、またIOがキャッシングの対象とみなされていないことを示しています。

IO Reasons

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。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を含みます。

図17に表示されているIO Reasonsのほとんどは、通常、フラッシュ・キャッシュにキャッシュされるものです。

- limit dirty buffer writes – バッファ・キャッシュ内の使用済みバッファの数を制限するためにDBWRから発行された書込み。
- data file reads to private memory（プライベート・メモリへのデータファイル読取り） – これらは大規模読取りであるため、いつもフラッシュ・キャッシュにキャッシュされるとは限りません。ただし、これはリクエストのわずか17%です。[3]
- buffer cache reads – データベース・バッファ・キャッシュへの読取り。これは通常のユーザー読取りで、通常はフラッシュ・キャッシュにキャッシュされるはずです。
- redo log writes – REDOログへの書込み。スマート・フラッシュ・ログを使用すると、この書込みはスマート・フラッシュ・ログとREDOログの両方に対して行われます。

残りのリクエストは最大でもわずか4%程度で、データベース制御ファイルの読取りかInternal IO（内部IO）のいずれかです。Internal IO（内部IO）はストレージ・サーバーによって行われるIOです。

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

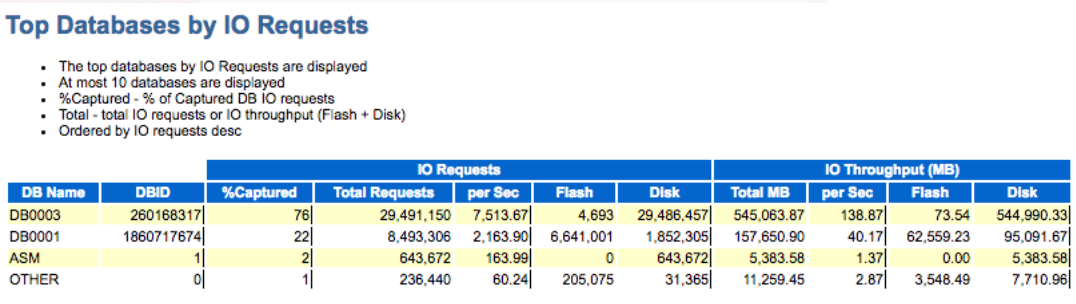
		Requests			MB	
Cell Name	IO Reason	%Cell	Total Requests	per Sec	Total MB	per Sec
*****celadm01	limit dirty buffer writes	40.32	1,177,513	300.00	12,176.33	3.10
	data file reads to private memory	18.32	535,035	136.31	7,373.28	1.88
	buffer cache reads	15.02	438,557	111.73	3,440.62	0.88
	redo log write	14.33	418,541	106.63	6,561.88	1.67
	Internal IO	3.95	115,348	29.39	651.29	0.17
*****celadm02	limit dirty buffer writes	41.28	1,175,349	299.45	12,220.91	3.11
	data file reads to private memory	16.99	483,811	123.26	5,207.55	1.33
	redo log write	14.22	404,816	103.14	6,117.16	1.56
	buffer cache reads	14.16	403,156	102.71	3,170.30	0.81
	database control file read	4.40	125,396	31.95	2,924.70	0.75
*****celadm03	limit dirty buffer writes	42.07	1,173,090	298.88	12,253.26	3.12
	data file reads to private memory	17.82	496,908	126.60	6,837.38	1.74
	redo log write	15.25	425,271	108.35	6,471.48	1.65
	buffer cache reads	14.52	404,997	103.18	3,183.83	0.81
	Internal IO	4.11	114,567	29.19	540.01	0.14
*****celadm04	limit dirty buffer writes	42.19	1,182,429	301.26	12,278.58	3.13
	data file reads to private memory	18.40	515,717	131.39	6,649.20	1.69
	redo log write	15.54	435,426	110.94	6,158.98	1.57
	buffer cache reads	14.42	404,178	102.98	3,172.83	0.81
	Internal IO	4.06	113,865	29.01	644.66	0.16

図17.AWRレポートの Top IO Reasons by Requests（リクエスト件数順IO理由）

3 『Oracle Exadata Database Machineシステム概要』の付録A「Oracle Exadata Database Machineの新機能」のA.2.6「大量分析問合せおよび大量ロードのパフォーマンス高速化」を参照してください。

Top Databases

図18のTop Databases by IO Requests（IOリクエストの件数順上位データベース）を見ると、データベースDB0003に対するIOの大半がディスク上で発生していることがわかります。



Top Databases by IO Requests

- The top databases by IO Requests are displayed
- At most 10 databases are 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
DB0003	260168317	76	29,491,150	7,513.67	4,693	29,486,457	545,063.87	138.87	73.54	544,990.33
DB0001	1860717674	22	8,493,306	2,163.90	6,641,001	1,852,305	157,650.90	40.17	62,559.23	95,091.67
ASM	1	2	643,672	163.99	0	643,672	5,383.58	1.37	0.00	5,383.58
OTHER	0	1	236,440	60.24	205,075	31,365	11,259.45	2.87	3,548.49	7,710.96

図18.AWRレポートのTop Databases by IO requests（IOリクエスト件数順上位データベース）

分析のまとめ

ここまでの分析でわかったことは次のとおりです。

- データベースのIOパフォーマンスが低下しています。DB時間のほぼ75％が'cell single block physical read（セルのシングル・ブロック物理読取り）'に費やされ、平均待機時間が8ミリ秒を超えています。
- Flash Cacheの各セクションは、フラッシュ・キャッシュのアクティビティが非常に少ないこと、また**IOがキャッシングの対象とみなされていない可能性がきわめて高いこと**を示しています。
- IO Reasonsとして示されているのは、通常であればフラッシュ・キャッシュの対象になるはずのごく典型的なIOです（ただし、PGAに対する読取りがある場合は例外です）。
- Top Databasesを分析したところ、このデータベースのIOリクエストのほとんどがフラッシュ上ではなくハード・ディスク上で実行されていることが確認されました。

以上のデータから、おそらく構成上の問題があり、それが原因でIOがフラッシュ・キャッシュをバイパスしている可能性がきわめて高いことがわかりました。構成データをお客様と確認したところ、データベースに対するフラッシュ・キャッシュの使用を誤って無効にしていたIORMプランが見つかりました。IORMプランを修正すると、パフォーマンスの問題は自動的に解決しました。

Exadataのパフォーマンス・データ

AWRレポートのほかにも、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Exadata上で入手できます。

表2は、入手できるデータと、その相対的な特性をまとめたものです。

パフォーマンス・データ

AWR	セル・メトリック[4]	EXAWATCHER
• 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 • レポート時間隔（デフォルトは1時間）での平均	• セルごとに収集 • 累積値と1秒あたりの割合（1分ごとに計算）を含む • 7日間保存	• セルごとに収集 • 5秒ごと • 7日間保存 • GetExaWatcherResults.shでグラフの作成が可能
入手可能なデータ • 構成情報 • OSの統計情報（iostatなど） • セル・サーバーの統計情報 • Exadataのスマート機能 • IOの理由 • 上位のデータベース	入手可能なデータ • Exadataの“スマート”機能（Flash Cache、Flash Log、IORM、Smart Scansなど）	入手可能なデータ • OSの統計情報 • Exadataの“スマート”機能（Flash Cache、Flash Log、IORM、Smart Scansなど）

表2.Exadata上で入手できるパフォーマンス・データ

結論

もっとも広く使用されているOracleデータベース向けパフォーマンス診断ツールであるAWRに、現在はExadataの統計情報が含まれています。Exadataの統計情報がAWRに統合されたことで、データベースのパフォーマンスに問題が発生しても、汎用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。

[4] セルのメトリックについて詳しくは、『Oracle Exadata System Softwareユーザーズ・ガイド』の「Oracle Exadata System Softwareの監視およびチューニング」を参照してください。

参考資料

1. Exadataの動作状態およびリソース使用率の監視
2. Exadata Health and Resource Utilization Monitoring - Exadata Database Machine KPIs
3. Exadata Health and Resource Utilization Monitoring - Adaptive Thresholds
4. Exadata Health and Resource Utilization Monitoring - System Baselining for Faster Problem Resolution
5. 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
6. Enterprise Manager Oracle Exadata Database Machineスタート・ガイド

ORACLE CORPORATION

Worldwide Headquarters
500 Oracle Parkway, Redwood Shores, CA 94065 USA

海外からのお問い合わせ窓口

電話　+ 1.650.506.7000　+ 1.800.ORACLE1

FAX　+ 1.650.506.7200
oracle.com

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

Integrated Cloud Applications & Platform Services

ExadataでのAWRレポート
の使用 2018年9月
著者：Cecilia Grant、Ashish Ray共著者：Curtis Dinkel

Oracle is committed to developing practices and products that help protect the environment