Oracleホワイト・ペーパー
2010年7月

Oracle Berkeley Database 11g Release 2 パフォーマンスの概要

ORACLE

概要

パフォーマンスは、データベースを選択する際に最も重要な要素です。このホワイト・ペーパーは、一般的な構成におけるOracle Berkeley Databaseの能力や特性を理解するために、パフォーマンスの測定結果を示したものです。アプリケーションのパフォーマンスは、対象データ、データのアクセス・パターン、キャッシュ・サイズ、他の構成パラメータ、オペレーティング・システム、ハードウェアによっても変わります。ベンチマークは、特定のアプリケーションのパフォーマンスを示すことはほとんどありませんが、ガイドラインを示し、基本的な運用上の想定を設定するために役立ちます。

導入

本書では、ある特定の構成で実施されたテストに基づいた、Oracle Berkeley DB（Oracle BDB）11g Release 2（11.2.5.0.21）のスループットに関する情報を提供します。テストには、さまざまな構成におけるOracle BDBのkey-value APIのスループットの計測と、Wisconsinベンチマークおよびtpc-Bと同等のベンチマークを使用したOracle BDBのSQL実装の検証が含まれています。

テスト環境

すべてのテストは、以下のハードウェア構成で実施されています。

表1 - ハードウェア構成

プロセッサ	オペレーティング・システム	RAM	ストレージ/速度	ファイル・システム
Intel Core 2 Duo E8400、3.00GHz	Red Hat Enterprise Linux Server 5.4	4GB	SATA、7200RPM	EXT3

key-value APIのパフォーマンス概要

Berkeley DBは多様な構成が可能なデータベースで、事前書込みのトランザクション・ロギングや同時実行性制御のロックなど、データベース操作の主要な特性を有効または無効にできます。このテストでは、2つの特定の構成を使用してBerkeley DBを検証します。

まず、Berkeley DBの構成オプションの1つであるData Store（DS）機能セットでテストを実施します。DSは、基本的にシンプルで、シングルスレッド/非トランザクションのストレージ・システムです。次に、フル・トランザクション・セマンティクスを提供するTransactional Data Store（TDS）機能セットでテストを実施します。

表2 - Data StoreおよびTransactional Data Storeの機能

機能	DATA STORE（DS）	TRANSACTIONAL DATA STORE（TDS）
アクセス・メソッド	Btree	Btree
ロック	なし	ページレベル
ロギング	なし	オンディスク（128MBバッファ付き）
トランザクション	なし	同期
共有メモリ	DB_PRIVATE	DB_PRIVATE
バッファ・キャッシュ	512MB	512MB

両テストとも、512MBバッファ・キャッシュ、データベース環境に対するDB_PRIVATEフラグ、10の空間的局所性が使用されます。DB_PRIVATEフラグは、データベース・キャッシュのヒープ・メモリの使用を示します。この構成は、単一プロセス（潜在的にはマルチスレッド）の実行の場合にのみ適しています。

データベースのパフォーマンスの場合、もっとも一般的なのはスループットを計測することです（一定の時間内に読み取り、または書き込まれたレコード数を測定）。これらのテストでは、Berkeley DB 11g Release 2（11.2.5.0.21）を使用し、1秒あたりの処理数をスループットとして表します。

テスト・データベースには、固定サイズのレコード（64バイトのキーおよび64バイトのデータ値）が使用されます。すべてのキーおよびデータ値は整数です。各テストで、10の連続したキー操作を行う、57,600のバッチを実行します。たとえば、挿入テストの場合、まず1つのキーをランダムに生成し、連続した9つのキー値とともに挿入します。このプロセスを続けて57,600回繰り返します。テストはそれぞれ5回実行し、操作タイプごとに平均と標準偏差がレポートされます。

構成ごとに、以下の手順でテストを行います。

- ステップ1 : データベースに576,000レコード（10の連続したキーを57,600セット）を追加します。これで、テスト・データベースが作成され、キャッシュがウォーム状態になりますが、ここでのレポートは行いません。
- ステップ2：10の連続したキーのセットをランダムに57,600セット取得します。
- ステップ3：10の連続したキーのセットをランダムに57,600セット更新します。
- ステップ4：10の連続したキーのセット内のすべてのレコードを削除します。
- ステップ5：ステップ1で行ったように、データベースにレコードを再度追加します。

操作ごとに、最初の操作の前と、最後の操作の後にタイマーを開始します。データベース環境とデータベース・ハンドルを開くまたは閉じるために要する時間は含まれません。

Data Store：シングルスレッド

最初のテストでは、Berkeley DB Data Store（DS）を使用して、シングルスレッドのアプリケーションのスループットを測定します。下記の表は結果を示したものです。

表3 - Oracle BDB DSのシングルスレッド・パフォーマンス

説明	挿入		フェッチ		削除		更新	
	操作/秒	標準偏差	操作/秒	標準偏差	操作/秒	標準偏差	操作/秒	標準偏差
Data Store	208,139	329	264,665	229	158,506	236	250,297	870

Transactional Data Store：シングルスレッド

2番目のテストでは、Berkeley DB TDSを使用して、書込みの永続性、ログの永続性、ストレージ・メディアなどの構成オプションを変えて、シングルスレッドのアプリケーションのスループットを測定します。以下に、これらの構成オプションについて説明します。

- **書込みの永続性** - デフォルトでは、トランザクションのディスクへのコミットは同期的に行われます。DB_TXN_NOSYNCフラグを設定することで、非同期コミットが有効になります。また、DB_TXN_WRITE_NOSYNCフラグを設定すると、ファイル・システムへのデータの書込みが行われますが、ディスク書込みは同期的に行われません。
- **ログの永続性** - デフォルトでは、ログはディスクに格納されます。非永続的なインメモリ・ロギングについても調査します。
- **ストレージ・メディアの違い** - デフォルトの構成では、従来のハード・ディスクを使用します。このパフォーマンスを、2GB RAMディスクを使用し、RAMディスクにデータベースおよびログの両方を格納することで得られるパフォーマンスと比較します。いずれのケースにおいて、トランザクションのコミットは同期的に行われます。

表4 - Oracle BDB TDSのシングルスレッド・パフォーマンス

説明	挿入		フェッチ		削除		更新	
	操作/秒	標準偏差	操作/秒	標準偏差	操作/秒	標準偏差	操作/秒	標準偏差
TDS SYNC	693	10	159,837	895	831	23	1,732	18
TDS RAMDISK[1]	49,673	912	162,848	1,588	44,279	277	60,973	291
TDS WNS[2]	37,664	460	160,307	293	36,110	454	52,922	258
TDS NS[3]	53,199	613	159,980	1,419	49,340	183	86,960	639
TDS INMEM[4]	66,435	102	163,229	815	58,845	101	97,602	396

1. 2GB RAMDISK、データベースとログをRAMDISKに格納
2. トランザクションのコミット・フラグ DB_TXN_WRITE_NOSYNC を設定
3. トランザクションのコミット・フラグ DB_TXN_NOSYNC を設定
4. インメモリ・ロギングを有効化

上述のすべてのテストでは、データベース・キャッシュが十分に効いているため、ディスクI/Oは読取りスループットに影響を与えていません。TDSとTDS RAMDISKの結果を比較すると、I/Oの遅延時間とスループットが、書込みスループットに大きな影響を及ぼすことは明らかです。また、TDS SYNC、TDS INMEM、TDS WNS、TDS NSによる結果の違いは、永続性が書込みスループットに与える影響を示しています。

この結果には、すべてのデータベースに共通した、多数の興味深い特徴が示されています。その中の1つに、パフォーマンスを決めるもっとも重要な要素がI/O処理であるということがあります。ストレージ・ディスクからの読取りでは、わずかなペナルティしか発生しませんが、書込みのほうははるかに重大です。TDSの挿入と更新のテストでは、もっとも大きいI/Oオーバーヘッドが発生しているため、そのパフォーマンスに最大の影響が出ています。

これらのテストは、以下の5つの主な遅延時間のカテゴリを示しています。

- メモリおよびプロセッサ
- ユーザー空間からカーネル空間への転送
- ファイル・システム
- ストレージI/O
- メディア

すべてのログ操作がインメモリで行われるTDS INMEMがもっとも高速という結果が出ています。コミット操作はバッファ・キャッシュで実行されます。そのため、カーネル・ファイル・システムのバッファやディスクのいずれにもデータは転送されません。TDS INMEMのシングルスレッド・テストの挿入操作の計測では、理想的な速度が出ていますが、データは永続的ではないため、あまり有用ではないかもしれません。プロセスが突然終了した場合は、データは消失してしまいます。データがメモリ内にあるため、永続性はありません。

トランザクションのDB_TXN_NOSYNCフラグが設定されている場合、コミット時にデータは、ログ・バッファからファイル・システムへ転送されません。ログ・データはログ・バッファを使い切るまでログ・バッファに保管され、ファイル・システムの書込み時に、ログ・レコードがユーザー・メモリからカーネル/ファイル・システムのメモリに転送されます。その後は、ファイル・システムによって、そのデータがストレージ・メディアに書き込まれるタイミングが決定されます。つまり、TDS INMEMとTDS NSのパフォーマンスの違いは、ログ・バッファが枯渇した際に、ユーザーレベルのログ・バッファからカーネル/ファイル・システムのメモリへログ・レコードが転送されることによって生じるオーバーヘッドということになります。

トランザクションのDB_TXN_WRITE_NOSYNCフラグが設定されている場合、Berkeley DBはログ・レコードをトランザクションごとにユーザーレベルのログ・バッファ・メモリからオペレーティング・システム/ファイル・システムに転送します。これにより、安定したストレージにそのデータが書き込まれる前にオペレーティング・システムに不具合がない限り、永続性が保たれます。アプリケーションに不具合が発生しても、オペレーティング・システムが継続して稼働していれば、データの喪失はありません。しかし、オペレーティング・システムまたはハードウェアに不具合が発生した場合は、オペレーティング・システムに転送されたものの、安定したメディアにはまだ書き込まれていなかったデータは失われます。

TDS WNSとTD SNSのパフォーマンスの違いは、挿入操作中に行われる、このBerkeley DBログ・バッファからファイル・システムへのコミット時のコピーを反映しています。最後に、TDS SYNCとTDS RAMDISKのパフォーマンスを比較すると、メモリへの書込みとディスクへの書込みとの間に違いが見られます。TDS SYNCの場合、SATAバスとハード・ドライブのオーバーヘッドが、パフォーマンスの劇的な低下の原因となっています。

すべてのテストの読取りパフォーマンスを比較すると、その平均値は標準偏差と1%異なるだけで（標準 = 1655.92、平均 = 161240.2、百分率変化 = 1.026989482）、ほぼ同じであることは興味深い結果です。これは、キャッシュがすべてのテストにおいて同等に効率的に使用されたこと、また、キャッシュ内のデータに対するフェッチ（読取り）操作でのトランザクションの保証によるオーバーヘッドがごくわずかであったことを示しています。測定は行っていませんが、キャッシュに存在しないデータへのリクエスト（フェッチ、読取り）では、I/Oのオーバーヘッドが処理時間の大半を占めることが予想できます。

Transactional Data Store：対称型マルチプロセッサ・システム（SMP）での測定

下記の表は、対称型マルチプロセッサ・システム（SMP）を使用してOracle BDB TDSのパフォーマンスを測定したものです。マルチスレッドのアクセスでは、競合を最小限に抑えるために各スレッドにDBハンドルが用意されています。各スレッドは、スループットの計算にそれぞれ1個のタイマーを使用します。総スループットを表に示します。

2番目のスレッドを導入すると、各操作タイプでスループットが低下します。これはシステムがロック処理を行うため、オーバーヘッドが発生することによるものです。書込みスループットは、スレッドの追加によく対応しています。読取りスループットは前の4つのスレッドで低下します。これは、CPUの使用率が100%の状況でスレッドを追加することで、無用なコンテキスト・スイッチやハードウェア・キャッシュの削除が発生するためです。

表5 - SMPシステムでのマルチスレッドのTDSのパフォーマンス

説明	挿入		フェッチ		削除		更新	
	操作/秒	標準偏差	操作/秒	標準偏差	操作/秒	標準偏差	操作/秒	標準偏差
TDS 1スレッド	693	10	159,837	895	831	23	1,732	18
TDS 2スレッド	647	7	106,196	432	767	20	1,634	27
TDS 3スレッド	690	9	126,762	762	863	20	1,909	32
TDS 4スレッド	721	8	134,581	497	882	18	2,085	54
TDS 8スレッド	859	26	128,928	423	967	27	2,497	32
TDS 12スレッド	930	24	112,735	518	1,013	20	2,658	149
TDS 16スレッド	968	13	99,860	820	1,085	34	2,837	117

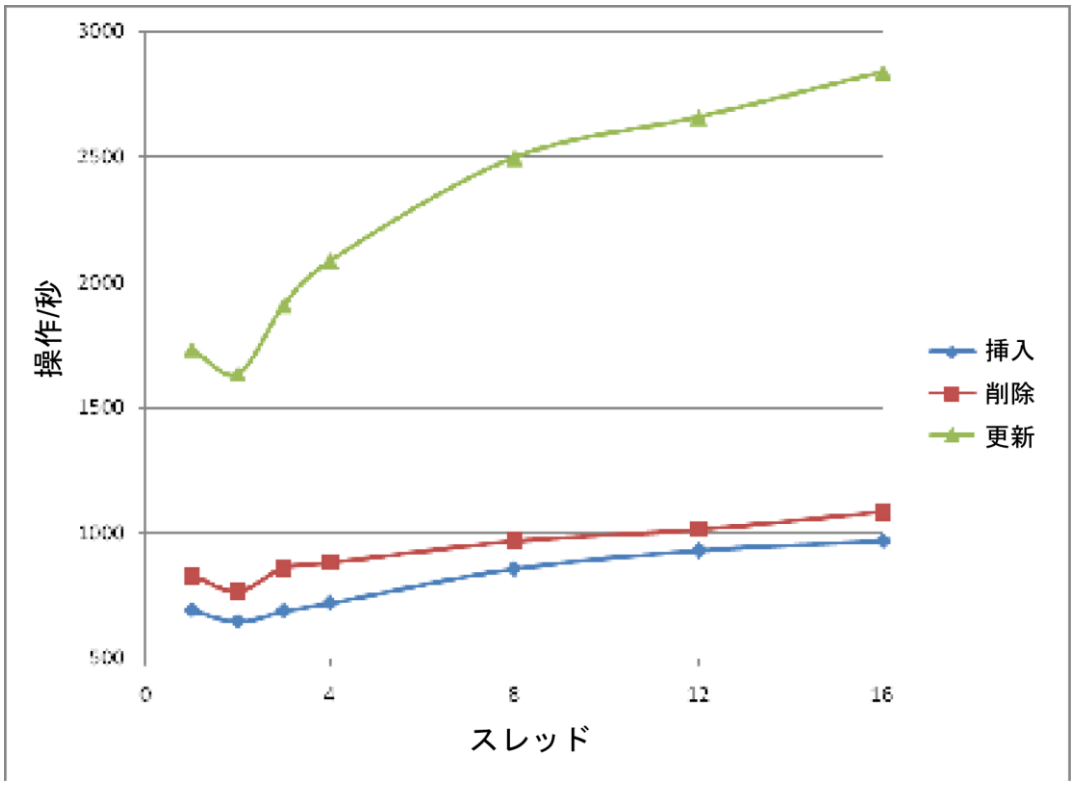

図1：SMPシステムでのマルチスレッドのTDSのパフォーマンス

SQLインタフェースのパフォーマンス概要

SQLベースのリレーショナル・データベース用には、いくつかのよく使われるベンチマークがあります。Berkeley DBのSQLインタフェースは、これらのベンチマークを使用する際に適しています。

以下のテストは、よく使われる2つのベンチマーク、Transaction Processing Performance Councilの TPC Benchmark B[1]とD. J. DeWittのWisconsin Benchmark[2]からアイデアを得ています。

これらの実施およびテストの実行は未認証でかつ第三者による照査も行われていません。これらは、Oracle BDBのパフォーマンスの特性に対する理解を深めることを目的としたもので、他のソフトウェア・システムのテストは参照していません。

TPC Benchmark B

このTPC-B風のテストでは、TPC Benchmark B仕様書[3]で指定されているように、スキーマと標準トランザクションを実施します。実行しやすくするため、口座の数は減らしています。ここでは100,000の口座と10,000の窓口を持つ1000の支店を使用します。

以下の表では、TPC Benchmark Bにおける、トランザクション・スループット率（トランザクション/秒（TPS））を表す数値を示しています。これらは、256MBのデータベース・キャッシュを使用して得られた結果です。

[1] http://www.tpc.org/tpcb/default.asp
[2] http://firebird.sourceforge.net/download/test/wisconsin_benchmark_chapter4.pdf
[3] http://www.tpc.org/tpcb/spec/tpcb_current.pdf

表6 - TPC-Bのトランザクション・スループット率

スレッド	TPS
1	1846.71
2	2310.84
3	2508.26
4	2678.14
5	2808.51
10	2859.15

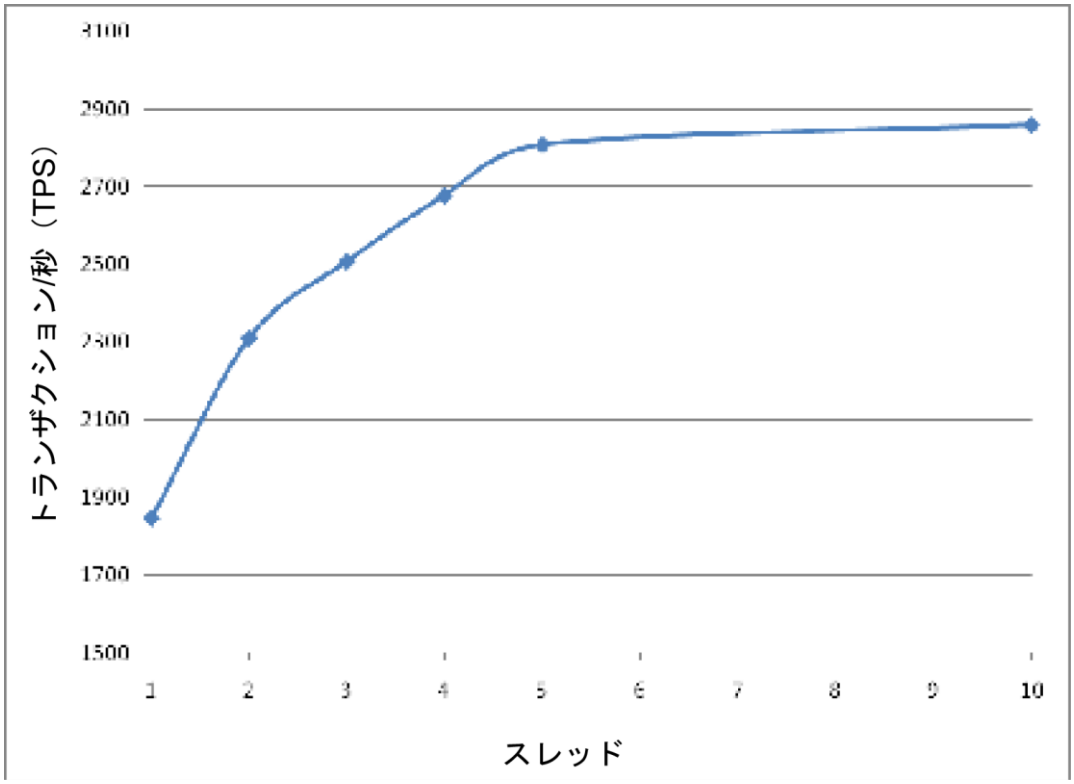

図2 - TPC-Bと同等のベンチマークで計測したトランザクション/秒（TPS） VS ワーカー・スレッド数

Wisconsin

Wisconsin Benchmark仕様書[4]に記載されているスキーマおよび問合せに準じてパフォーマンス・テストを行います。ただし、データベース・システムおよびオペレーティング・システムのキャッシングの影響を緩和するようなことは行っていません。

このテストでは、標準のデータベースで大量かつ多様なテスト・ケースを用いてシングルユーザーのパフォーマンスを計測します。すべてのテスト・ケースの結果を総合すると、データベースの長所と短所を分析することができます。問合せに関する説明は、後述の参考文書の付録Iを参照してください。

これらのテストを実行するために、データベースは、シングルスレッドのアクセスとシングルユーザーのアクセスを最適化するコンパイル時オプションを使用して構築されています。各問合せは10回実行され、その平均経過時間（ミリ秒）が以下の表に掲載されています。

表7 - Wisconsin Benchmarkで計測した平均経過時間

ケース番号	経過時間（ミリ秒）	ケース番号	経過時間（ミリ秒）	ケース番号	経過時間（ミリ秒）	ケース番号	経過時間（ミリ秒）
1	8.175	9	31047.220	17	16.944	25	18.163
2	13.875	10	3160.824	18	58.801	26	0.907
3	0.879	11	5335.219	19	136.595	27	3.844
4	6.749	12	9.920	20	3.944	28	4.380
5	1.368	13	9.942	21	32.566	29	1.081
6	11.017	14	11.585	22	32.774	30	1.273
7	0.244	15	19.407	23	0.102	31	1.252
8	1.373	16	14.759	24	17.742	32	1.081

[4] http://firebird.sourceforge.net/download/test/wisconsin_benchmark_chapter4.pdf

結論

Berkeley DBのkey-value APIは、シングルスレッドのアプリケーションでもマルチスレッドのアプリケーションでも優れたパフォーマンスを示します。また、アプリケーションが完全な永続性を求めていない場合は、パフォーマンスを大幅に高めることもできます。

Berkeley DBの11g Release 2 5.0では、key-value API上にSQL APIが追加されています。

Oracle Berkeley DBは以下のURLからダウンロードできます。

http://www.oracle.com/technology/software/products/berkeley-db/index.html

Oracle Berkeley DBに関するご意見やご質問は、Oracle Technology Network（OTN）フォーラムから投稿できます。

http://forums.oracle.com/forums/forum.jspa?forumID=271

販売やサポートに関する情報については、berkeleydb-info_us@oracle.comまでご連絡ください。新しい製品のリリースについては、bdb-join@oss.oracle.comまでご連絡ください。

Oracle Berkeley Database 11g
Release 2
パフォーマンス
2010年7月
著者：Karl Fu

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

海外からのお問い合わせ窓口：
電話：+1.650.506.7000
ファクシミリ：+1.650.506.7200
www.oracle.com

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

Hardware and Software, Engineered to Work Together