Oracle Exadata: первые результаты в демо-центре «Инфосистемы Джет»

Алексей Струченко
Руководитель группы оптимизации СУБД и приложений компании «Инфосистемы Джет»

Демо-центр по Oracle Exadata (далее Exadata) в компании «Инфосистемы Джет» активно функционирует с конца 2010 г. За это время был выполнен значительный объем работ по тестированию, изучению данного комплекса, наращиванию экспертизы. В начале декабря 2010 г. был проинсталлирован первый комплекс Exadata (X2-2 High Performance Quarter Rack) в головном офисе в Москве, в начале марта 2011 г. такой же комплекс был развернут в киевском филиале. В течение прошедших с момента первой инсталляции месяцев были успешно выполнены пять внутренних проектов по изучению возможностей Exadata, а также развернуты и протестированы четыре базы данных наших заказчиков. Параллельно была проведена серьезная работа по организации тренингов и сдаче экзаменов Oracle: Real Application Cluster, Performance and Tuning, Warehousing, Linux, собственно Exadata. В итоге компания «Инфосистемы Джет» смогла получить специализацию по Oracle Exadata в числе первых в Европе. Цель этой статьи — проанализировать первые результаты деятельности демо-центра по тестированию комплекса, причем представить не только цифры и графики, но и качественные оценки совместной работы с линейкой продуктов Oracle.

Exadata позиционируется как программно-аппаратный комплекс (Appliance), обещающий многократное повышение производительности Oracle Database на любых классах задач. Пожалуй, ключевое слово в этом позиционировании — «любых». Наш многолетний опыт оптимизации работы баз данных Oracle позволил сделать вывод, что к большинству систем необходим индивидуальный подход. Задача повышения производительности систем OLTP (системы с преобладанием коротких транзакций), DWH (хранилища с преобладанием тяжелых аналитических запросов) и систем смешанного типа зачастую решается принципиально разными способами. Если на рынок выходит решение, способное решить задачу повышения производительности для любого класса задач, — это, несомненно, революционное предложение, заслуживающее самого пристального внимания.

Архитектура Exadata

С аппаратной точки зрения Oracle Exadata состоит из трех компонентов:

1. Серверы баз данных (Database Server). На сегодняшний день существует две конфигурации Exadata, отличающиеся серверами баз данных. В конфигурации X2-2 каждый Database Server включает два шестиядерных процессора Intel Xeon X5670 и 96 Гб памяти, данная конфигурация продается в виде Quarter Rack (1/4), Half Rack (1/2) и Full Rack. В конфигурации X2-8 каждый Database Server включает восемь восьмиядерных процессоров Intel Xeon X7560 и 1 Тб памяти, данная конфигурация продается только в виде Full Rack. На серверах баз данных функционирует стандартное программное обеспечение Oracle Database.

2. Серверы хранения (Exadata Storage Server). В обеих конфигурациях каждый Exadata Storage Server включает два шестиядерных процессора Intel Xeon L5640, 24 Гб памяти, 384 Гб Exadata Smart Flash Cache и 12 дисков. Диски могут быть двух видов — High Performance (более быстрые, емкостью 600 Гб) или High Capacity

Таблица 1. **Различные конфигурации Exadata**

	X2-2 Quarter	X2-2 Half	X2-2 Full	X2-8
Database Server	2	4	8	2
Exadata Storage Server	3	7	14	14
InfiniBand Switch	2	3	3	3

(более медленные, емкостью 2 Тб). За зеркалирование и «размазывание» данных по дискам отвечает Oracle Automatic Storage Manager (ASM), на серверах хранения функционирует специальное программное обеспечение Exadata Software.

3. Сетевое решение InfiniBand. Серверы баз данных и серверы хранения соединены коммутаторами Oracle Data Center InfiniBand Switch (40 Gb/s, что по крайней мере в пять раз быстрее традиционных сетевых решений на основе Fibre Channel).

Таблица 1 показывает, чем отличаются различные конфигурации Exadata. В Full Rack (независимо от того, X2-2 или X2-8) входят 14 серверов хранения, которые обеспечивают 100 Тб «сырого» дискового пространства High Performance либо 336 Тб «сырого» дискового пространства High Capacity. Следует отметить наличие во всех конфигурациях Exadata (кроме Quarter Rack) третьего коммутатора InfiniBand, с помощью которого осуществляется связь нескольких Exadata между собой. Также с помощью этого коммутатора можно связать Exadata с Exalogic — еще одним программно-аппаратным комплексом от Oracle, являющимся Application Appliance (на основе Weblogic).

Программное обеспечение Exadata

Итак, с аппаратной точки зрения Exadata представляет собой сбалансированное под задачи Oracle Database стандартное Oracle/Sun Hardware на основе платформы Intel x86. Но «волшебством» Exadata является специальное ПО, функционирующее на серверах хранения — Exadata Software. На рис. 1 схематично представлена архитектура Exadata, программное обеспечение Exadata Software выделено красным цветом.

Первой ключевой особенностью Exadata Software является Offloading или Smart Scan — возможность переноса части SQL-логики на уровень серверов хранения.

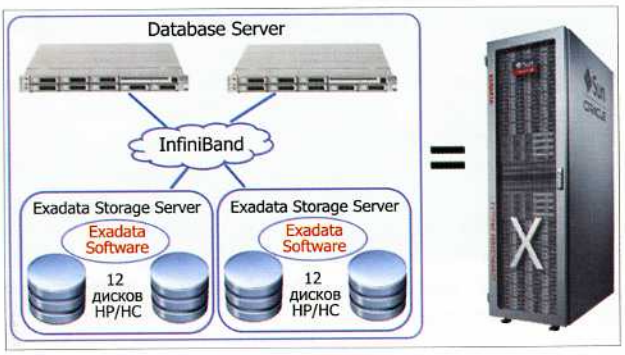

Рис. 1. Архитектура Exadata

Можно сказать, что Exadata включает в себя систему хранения данных, «понимающую SQL». Для объяснения процесса Offloading рассмотрим идеальную модель: пусть перед базой данных Oracle стоит задача прочитать миллиард блоков, обработать их и в результате обработки вернуть несколько строк. Традиционная архитектура справляется с этой задачей следующим образом: дисковые массивы читают миллиард блоков, передают их по сети на сервер баз данных, который обрабатывает эти блоки (как правило, в памяти) и выбирает нужные строки. Архитектура Exadata подойдет к решению этой задачи иначе (конечно, если оптимизатор Oracle сочтет способ Offloading более эффективным): серверы хранения будут параллельно обрабатывать миллиард блоков и посылать по сети отобранные строки на серверы баз данных. Если в системе есть задачи, похожие на эту идеальную модель (как правило, это тяжелые аналитические запросы в хранилищах класса DWH), то Exadata обещает значительный выигрыш в производительности.

Вторая важная особенность Exadata Software — специальный механизм сжатия Exadata Hybrid Column Compression. Здесь ключевое слово — Column: данные сжимаются поколоночно, за счет чего достигается более высокая степень компрессии, чем при традиционном механизме сжатия Oracle (так называемый Basic). Этот механизм реализован в виде двух различных алгоритмов сжатия — Query Low/High (для часто используемых данных) и Archive Low/High (для редко используемых, архивных данных). На Oracle Open World 2010 были представлены результаты по компрессии баз данных турецкого телеком-оператора Turkcell — согласно презентации, при миграции данных в Exadata различными способами удалось получить общее сжатие в десять раз за счет Exadata Hybrid Column Compression. Нам практически удалось повторить этот впечатляющий результат для данных DWH-хранилища одного из наших заказчиков (телеком). Большинство таблиц этого хранилища изначально были сжаты традиционным механизмом Basic в 2—2,5 раза, и поверх этого данные были сжаты еще в 3,35 раза Exadata алгоритмом Query High, что дало общую степень компрессии в 6,7—8,3 раза!

И, наконец, рассмотрим задачи, хорошо известные специалистам по оптимизации СУБД Oracle — задачи, в которых существуют критичные объекты по вводу-выводу (подобные задачи встречаются как среди систем OLTP, так и среди хранилищ). Серверы хранения в Exadata обладают собственным флеш-кешем Exadata Smart Flash Cache размером 384 Гб. Комбинируя различные способы использования этого кеша, можно добиться ускорения ввода-вывода для конкретных объектов базы данных и тем самым значительно повысить производительность системы в целом.

Масштабирование в Oracle Real Application Clusters

Лицензии Oracle Database являются минимальным набором при лицензировании серверов баз данных. Если рассматривать Exadata под задачи консолидации не-

скольких небольших баз, то можно обойтись без опции Oracle Real Application Clusters (далее RAC), однако работа одной базы данных на нескольких Database Server в Exadata возможна только с помощью RAC. Поэтому первые эксперименты на нашей Exadata были связаны с вопросами масштабирования в RAC.

На рис. 2 приведен характерный результат масштабирования в RAC аналитической задачи одного из наших заказчиков. Для тестирования была обезличена часть крупного DWH-хранилища размером 1,3 Тб, при этом исследовалось среднее время исполнения запросов.

При достаточном количестве сессий (50 и более одновременно выполняемых запросов) два узла RAC отработали практически в два раза быстрее одного, причем без какой-либо дополнительной оптимизации.

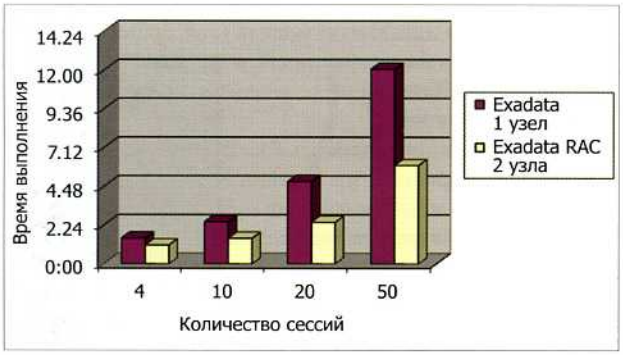

Рис. 2. Масштабирование в RAC отчета DWH

Exadata в сравнении с продуктивными системами

После исследования вопросов масштабирования в RAC была протестирована база данных размером 1,9 Тб (задача класса OLTP от крупного телеком-оператора). Исследовалось время работы процедуры перерасчета абонентской задолженности биллинговой системы. Процедура представляет собой многопоточный PL/SQL-код, обрабатывающий значительную часть центральной таблицы (аналогично банковской процедуре капитализации либо закрытия периода). На момент тестирования данная процедура в продуктиве длилась не менее четырех часов (240 минут) на 32-ядерном сервере, при этом из-за ограниченных возможностей СХД более четырех потоков запустить не удавалось.

В таблице 2 приведены результаты сравнения работы процедуры перерасчета на Exadata и в продуктиве.

На Exadata удалось запустить процедуру перерасчета в 40 потоков, после небольшой оптимизации под RAC два узла Exadata (24 ядра) показали результат в 12 минут, что в 20 раз быстрее продуктива. Заметим, что полученный результат не изменился после сжатия основной таблицы алгоритмом Query High в четыре раза, при этом несколько выросла утилизация процессоров серверов баз данных.

Таблица 2. **Многопоточная процедура перерасчета на Exadata**

Потоки	1 узел RAC		2 узла RAC	
	Мин	CPU %	Мин	CPU %
4 (продуктив)	Более 240	Менее 20		
40	16	90	12	40
40 (с компрессией)			12	50

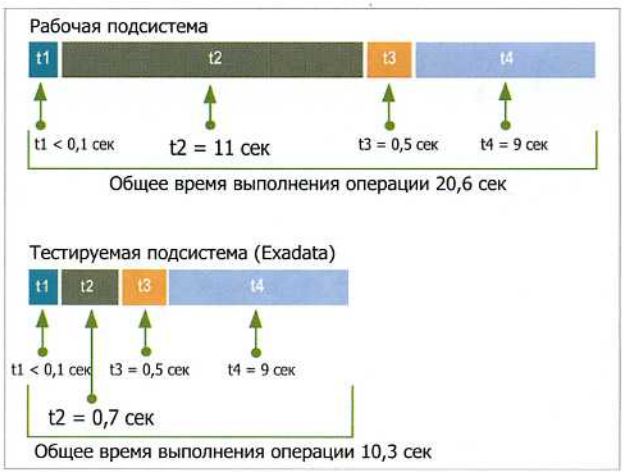

Рис. 3. Общее время исполнения в продуктиве и на Exadata

На рис. 3 показаны результаты сравнения продуктива и Exadata, полученные в рамках тестирования клиент-серверного приложения, обслуживающего хранилище класса DWH размером 400 Гб. Общее время отклика складывалось из времени передачи команды от клиента серверу по сети (t1), времени работы базы данных (t2), времени передачи результатов от сервера клиенту по сети (t3) и времени обработки полученных данных на клиенте (t4). Общее время отклика на Exadata оказалось в среднем в два раза меньше, чем на продуктиве. Если сравнивать по параметру t2 (собственно время работы базы данных), то Exadata показала себя в среднем в 15—16 раз производительнее текущей системы!

Exadata: первые итоги

Объем данной статьи не позволяет охватить все результаты, полученные на сегодня в нашем демо-центре. За рамками материала остались исследование работы на Exadata систем на основе Oracle Siebel CRM и Oracle E-Business Suite, проигрывание на Exadata нагрузки с помощью опции Real Application Testing, настройка процедур резервного копирования и мониторинга и т. д. Тем не менее на основе приведенных в статье первых результатов уже можно сделать вывод, что на большинстве задач, типичных для наших заказчиков, Exadata дает существенное повышение производительности Oracle Database. Кроме того, реализованный в Exadata специальный механизм компрессии Exadata Hybrid Column Compression на ряде задач показал сжатие данных в 8—10 раз при той же производительности. По итогам полугодового знакомства с Exadata мы сделали однозначный вывод: продукт заслуживает самого пристального внимания. Мы будем рады любым новым задачам по тестированию в нашем демо-центре по Oracle Exadata!