Sun Oracle Database Machine 2: интегрированная унифицированная платформа для BI- и OLTP-приложений

В сентябре 2009 г. компания Oracle объявила о доступности в продаже 2-й версии Oracle Database Machine специализированной консолидированной платформы для обработки и хранения данных в составе как для BI- (Business Intelligence), так и для OLTP-приложений (online transaction processing). Решение построено на базе стандартных Sun-серверов с использованием FlashFire технологии и ПО от Oracle: Oracle Database 11g Release 2 и Oracle Exadata Storage Server Software Release 11.2. Это вторая публикация (первая – № 1/38, 2009), посвященная Oracle Database Machine.

Валерий Безруков – заместитель руководителя по продвижению, Центр Вычислительных Комплексов, ЗАО "Инфосистемы Джет".

Введение

За последние несколько лет так называемые серверы стандартной архитектуры стали неотъемлемой частью ИТ-инфраструктуры почти всех предприятий – от небольших до самых крупных. Серверы стандартной архитектуры используются для решения практически всех классов задач – от вспомогательных (размещение служб каталогов и электронной почты) до серверов приложений и даже иногда – баз данных.

Однако, если, например, для БД Microsoft SQL Server в силу архитектурных ограничений СУБД серверы стандартной архитектуры являются практически единственной и как, следствие, самой популярной платформой, то для серверов БД Oracle Server использование серверов стандартной архитектуры все еще является экзотикой. На этом рынке пока лидируют серверы RISC-архитектуры – SPARC, Power и Intanium.

В последние годы Oracle активно развивает направление, связанное с поддержкой своих программных продуктов на серверах стандартной архитектуры. В результате этих усилий на данный момент у Oracle есть полный набор системного программного обеспечения для работы "с уровня железа" на серверах стандартной архитектуры, включая собственный дистрибутив Linux – Oracle Enterprise Linux.

Вероятно, именно благодаря наличию подобного набора системного ПО, а также, принимая во внимание вопросы стандартизации и открытости для первого своего программно-аппаратного комплекса – Oracle Exadata v1, – Oracle решил использовать исключительно серверы стандартной архитектуры. Все аппаратные компоненты Exadata v1 были реализованы на оборудовании производства HP, при этом все серверы были стандартной архитектуры. А все ПО, включая операционные системы, кластерное ПО и непосредственно СУБД, поставлялось от Oracle. Правда, позиционировалась эта система в основном для решения задач класса BI/DWH, то есть для такой нагрузки, где основной операцией является чтение и обработка статических данных. А для решения задач типа OLTP эта платформа, как минимум, не позиционировалась.

Прошел почти год, и появилась Exadata v2. Первое отличие v2 от v1, которое видно сразу, – смена производителя аппаратной части. В контексте решения компании Oracle – купить компанию Sun и идущего сейчас процесса согласования этого решения с уполномоченными государственными органами США и Европы – вполне предсказуемым решением Oracle было перейти на использование оборудования Sun при построении Exadata v2. Но в Exadata по-прежнему используются серверы стандартной архитектуры, просто на этот раз – производства Sun.

Но смена производителя аппаратного обеспечения не обошлась без небольшого сюрприза – в составе сервера, который в Exadata v2 выполняет роль "продвинутого" хранилища информации, – Exadata Storage Cell – появилась достаточно интересная компонента – PCI адаптер 96 GB Sun Flash Accelerator F20 PCIe. Эта компонента, представляющая расположенную на PCI адаптере flash-память в размере 96GB, достаточно интересна сама по себе, а в составе такой системы как Exadata – интересна дважды.

А самое интересное: эволюция Exadata – от v1 к v2 – под воздействием, в том числе и этой компоненты. Если, как помните, ранее Exadata предназначалась в основном для хранения и обработки статической информации, то теперь Oracle не ограничивает типы решаемых Exadata v2 задач только BI/DWH – система позиционируется как универсальная "БД машина" (Sun Oracle Database Machine) и готова принять на себя любую нагрузку, как BI/DHW, так и OLTP.

Для того чтобы понять суть произошедших в Exadata изменений, целесообразно сначала рассмотреть структуру системы, а уже только после этого провести анализ нововведений и сформулировать возможные ожидания от обновленной системы.

Что такое Exadata v2?

Если сказать кратко, то это Oracle Real Application Clusters 11g R2 на серверах стандартной архитектуры + "активная" система хранения. Причем, эта система хранения построена достаточно необычным, по крайней мере, для Oracle Server способом – также с использованием серверов стандартной архитектуры, а именно Sun Fire X4275.

В составе Oracle Exadata v2 такой сервер называется Exadata Storage Server, в его состав входит (рис. 1):

- 2x Quad-Core Intel® Xeon® E5540 Processors (2.53 GHz), 24GB RAM;

Рис. 1. Внешний вид Exadata Storage Server.

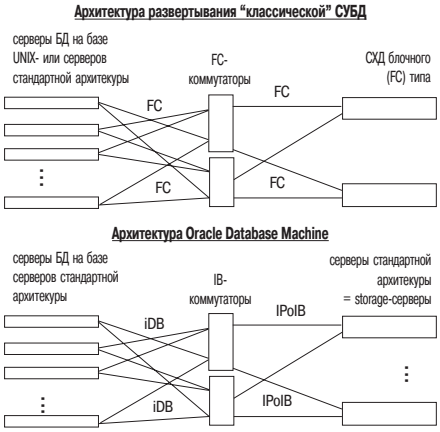

Рис. 2. В Oracle Database Machine FC-СХД заменены storage-серверами (на базе стандартных серверов), каждый из которых содержит часть функционала серверов БД и часть всех таблиц БД.

- 12 x 600 GB 15K RPM SAS или 12 x 2 Тбайт 7.2K RPM SATA;
- 4 x 96 GB Sun Flash Accelerator F20 PCIe Cards;
- Disk Controller HBA with 512MB Battery Backed Cache;
- 2 InfiniBand 4X QDR (40Gb/s) Ports (Dual-port HCA).

Именно в состав этого сервера и входит PCI адаптер 96 GB Sun Flash Accelerator F20 PCIe Card.

Итак, несколько серверов SF X4275 – Storage Server'ов – подключаются к IB коммутаторам по соответствующему протоколу. "Поверх" IB серверы Storage Server'ы и другие компоненты Exadata v2 "общаются" между собой по разработанному и реализованному Oracle протоколу ZDP (Zero-loss Zero-copy Datagram Protocol). Протокол ZDP, в свою очередь, основан на другом разработанном Oracle протоколе – Reliable Datagram Sockets (RDS), версия 3, который в терминологии OSI является транспортным протоколом 4-го уровня семиуровневой модели и работает "поверх" протокола IPv4, а если быть совсем точным – IPoIB.

"Выше" протокола ZDP находится еще один разработанный Oracle сетевой протокол – iDB. Этот протокол во многом похож на широко распространенный открытый протокол iSCSI. Однако, помимо поддержки обмена данными между сервером и системой хранения на классическом для систем хранения "блочном" уровне, в iDB реализован еще один механизм. Именно этот механизм позволяет превратить систему хранения Exadata v2 в "активную" систему хранения за счет передачи на ее сторону не просто команд на возврат той или иной последовательности блоков данных. iDB позволяет серверу посылать в сторону Storage Server некий упрощенный аналог SQL-запроса, который, в свою очередь, будет обработан Storage Server'ом самостоятельно, и в ответ серверу вернется уже конкретный ответ на запрос, а не просто набор "сырых" блоков данных (рис. 2).

Такой подход в обработке упрощенных запросов силами Storage Server'а Oracle называется "Smart Scans" или "Scan Offload". Использование этого механизма позволяет Storage Server'ам обрабатывать следующие запросы:

- фильтрация строк на основе "where" предиката;
- фильтрация колонок;
- фильтрация соединений (join);
- фильтрация зашифрованных данных;
- работа с функциями Data Mining.

Oracle заявляет, что smart scan прозрачен для приложений – то есть Oracle Server 11g R2 каким-то образом решает, что ему делать в каждой отдельной ситуации: работать со Storage Server на уровне блоков или smart scan. Вероятно, это делается в момент составления плана запроса, и, вероятно, что Oracle Server 11g R2 для этого должен понимать, где он работает, – на Exadata или на "простом" сервере. Это подтверждается появлением в "обычном" Oracle Server 11g R2 инициализационных параметров, которые отвечают за передачу управления запросами на сторону Exadata, например, cell_offload_processing.

Какая бывает Exadata?

Exadata по своей структуре построена по принципу MPP-систем – ядром системы является высокопроизводительный коммутатор к которому подключаются все компоненты, при этом все эти компоненты активно задействованы в процессе обработки информации. Одним из преимуществ такого подхода является хорошая теоретическая масштабируемость системы и возможность расширять ее аппаратные ресурсы по мере необходимости. Однако у этого расширения есть организационные рамки – Exadata расширяется этапами, и этапы эти определяются тем, какой процент от стойки (rack) занимает оборудование той или иной конфигурации.

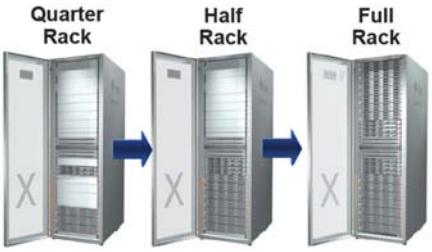

Рис. 3. Варианы исполнения Sun Oracle Database Machine: 1/4 стойки, 1/2 стоки, полная стойка.

На данный момент времени есть четыре основные конфигурации Exadata v2 (рис. 3):

- Sun Oracle Database Machine Full Rack Hardware: 8x Database server SF X4170, 14 Exadata storage server, 3 Sun Datacenter InfiniBand Switch 36;
- Sun Oracle Database Machine Half Rack Hardware: 4x Database server SF X4170, 7 Exadata storage server, 2 Sun Datacenter InfiniBand Switch 36;
- Sun Oracle Database Machine Quarter Rack Hardware: 2x Database server SF X4170, 3 Exadata storage server, 2 Sun Datacenter InfiniBand Switch 36;
- Sun Oracle Database Machine Basic System Hardware: 1x Database server SF X4170, 1 Exadata storage server, 2 Sun Datacenter InfiniBand Switch 36.

Как видно из названия, первые три являются вариантами заполнения стойки – от 1/4 до полной, самый же нижний – Basic – из-за отсутствия дублирования компонентов представляет интерес разве что для разработчиков или организации тестовых систем.

На серверы Exadata storage server – SF X4275 – ставится ПО Oracle Enterprise Linux и Oracle Exadata Storage Server Software – именно оно делает из сервера стандартной архитектуры под управлением OEL полноценного "участника" системы Exadata v2. На Exadata Database Server – SF X4170 – ставится все тот же Oracle Enterprise Linux и Oracle Server 11gR2. Набор ПО одинаков для всех конфигураций Exadata.

Как видно, в максимальной конфигурации, да и не только в ней, Exadata v2 представляет собой достаточно мощную систему. Оснащенные двумя четырехъядерными процессорами Intel Xeon E5540 серверы Sun Fire X4170 сами по себе могут посоревноваться по производительности со многими RISC серверами, а под управлением Oracle Real Application Clusters да еще и в количестве восьми штук, могут представлять достаточно серьезную конкуренцию многим промышленным RISC серверам.

Если, например, принять потери на межкластерное взаимодействие на каждый сервер в размере 25% от совокупной вычислительной мощности каждого сервера и посчитать "чистую" вычислительную мощность Exadata Full Rack, то получим следующее значение:

8 (северов) x 2 (процессора) x 4 (ядра) x 0,75 (потери на RAC) = 48 ядра или 24 процессора Intel E5540

Если это значение теперь попытаться сравнить с процессорами и серверами другой архитектуры, например HP Itanium, для чего использовать, например, публичный и хорошо зарекомендовавший себя тест SAPS, то получим следующие данные:

24 (процессоров) x 8282 (SAPS SAP ERP 6.0 EHP4 / NW7.0 / 7.01, 64 Bit, Unicode для процессора Xeon Quad Nehalem 2.53GHz) / 2925 (SAPS cern no 2007040) = 68 процессоров Dual-Core Intel Itanium Processor 9140M, 1.66 GHz, 32 KB L1 cache

Таким образом, можно ожидать, что Exadata Full Rack вполне сможет посостязаться, например, с HP Superdome в максимальной комплектации.

Секретное оружие Exadata v2, или Exadata Smart Flash Cache

Как уже говорилось выше, в составе Exadata v2 появился новый компонент – Sun Flash Accelerator F20 PCIe Card (рис. 4). На каждом из Exadata Storage Server устанавливается по четыре Sun F20 PCIe, таким образом, в каждом Exadata Storage Server содержится по 384GB flash-памяти типа Single Level Cell (SLC). Память эта доступна со стороны Exadata Storage Server по протоколу SAS, то есть с точки зрения операционной системы представляет собой обычный SSD диск, вернее, четыре диска. Sun F20 PCIe состоит из четырех частей – доме-

Рис. 4. Конструктивное исполнение Sun Flash Accelerator F20 PCIe Card.

нов, или, как их еще называют, Solid State Disks-on-Modules (DOMs), каждый из них содержит по 24GB flash-памяти. Каждый из доменов виден со стороны операционной системы по двум "широким" (wide) портам SAS. Sun F20 PCIe поддерживается несколькими операционными системами, в том числе Linux и Solaris. Однако купить этот адаптер можно только в составе Exadata и только у Oracle – после объявления Exadata v2 Sun решил не продавать его отдельно.

Согласно паспортным данным, Sun F20 PCIe обеспечивает до 100K IOPS на операциях чтения при размере блока данных 4K и 84K IOPS – на операциях записи при том же размере блока. При этом задержка на обработку операций чтений и записи составляет 0.32 и 0.22 миллисекунды соответственно. Эти данные вполне сопоставимы с параметрами SSD дисков других производителей с той поправкой, что таких SSD дисков в Sun F20 PCIe четыре.

Вся эта быстрая память может быть использована в Exadata v2 в двух режимах:

- для кэширования данных на уровне Exadata Stotage Server;
- в новом появившемся в Oracle Server 11gR2 режиме Flash Cache.

В первом случае Exadata Storage Server использует Sun F20 PCIe для хранения наиболее востребованной информации, при этом используется достаточно сложный механизм определения уровня этой самой востребованности, а не ставший уже привычным алгоритм LRU. Этот режим использования является прозрачным для пользователя и приложения, то есть присутствие в системе flash-памяти "ощущается" только через ускорение операций ввода-вывода.

Второй режим – Flash Cache – это поддержка появившегося в Oracle Server 11gR2 одноименного механизма. Механизм этот позволяет организовать "расширение" SGA buffer cache на SSD дисках. Стоит отметить, что режим Flash Cache поддерживается Oracle Server'ом только для двух операционных систем – Linux и Solaris. Flash Cache может управляться как в прозрачном для пользователя и администратора БД автоматическом режиме, так и в ручном. В ручном режиме пользователь или администратор могут закреплять в кэше отдельные объекты – таблицы и индексы – благодаря появившемуся в 11gR2 расширению синтаксиса, так называемого "storage clause" – части синтаксической конструкции SQL DDL оператора, ответственной за задание параметров хранения объекта.

Что ждать от Exadata: вместо заключения

Основной вопрос, который вызывает появление на свет второй версии Exadata – сможет ли эта система конкурировать с существующим подходом к построению инфраструктуры для Oracle Server, то есть с современными многопроцессорными RISC-северами и системами хранения. В Exadata v2 заложено достаточно много конкурентных преимуществ, которые вполне вероятно смогут сделать ее серьезным игроком на рынке инфраструктуры для Oracle Server, вот некоторые из них:

- *соотношение цена–производительность: применение серверов стандартной архитектуры позволяет, с одной стороны, получить достаточно серьезную производительность, с другой – минимизировать затраты на аппаратную часть. При этом снижаются как затраты на серверы БД, так и затраты на систему хранения;*
- *масштабируемость: построение системы по принципу grid-технологий позволяет надеяться на достаточно хорошую ее масштабируемость, а наличие в прайс-листе Oracle четырех вариантов Exadata v2 позволит заказчикам сохранять инвестиции в инфраструктуру путем приобретения начальной конфигурации с последующим ее расширением до максимальной по мере необходимости;*
- *интегрированность: Exadata v2 позволяет получить готовую инфраструктуру для работы Oracle Server 11g "под ключ" – в одном решении присутствуют тесно интегрированные серверы БД, система хранения и необходимая для их взаимодействия сетевая инфраструктура. Все, что остается сделать потенциальному заказчику, – установить стойку с Exadata v2 в ЦОД, подключить к питанию и Ethernet-сети. Такой подход выглядит весьма заманчивым как с точки зрения сокращения сроков внедрения, так и с точки зрения расходов на внедрение инфраструктуры. Понятно, что в реальной жизни процесс инсталляции и интеграции Exadata v2 все равно будет занимать некоторое время, однако благодаря стандартизации, с точки зрения используемых компонентов, по сути, есть только один вариант комплектации Exadata. С течением времени этот процесс будет сокращаться и выполняться "с закрытыми глазами";*
- *унифицированность: возможность развертывания на базе одного решения BI- и OLTP-приложений с высоким уровнем производительности, избегая при этом дополнительных затрат и сложности администрирования, требующихся для поддержания дополнительной инфраструктуры.*

Валерий Безруков,
ЗАО "Инфосистемы Джет"

10GbE решения Dell для ЦОД

Декабрь 2009 г. – Dell представила новые инструменты управления инфраструктурой, решения для сетей хранения данных и услуги, которые позволяют упрощать управление, стандартизировать и автоматизировать ЦОД.

В число новых решений для управления инфраструктурой входят:

Advanced Infrastructure Manager

Infrastructure Manager позволяет динамически распределять рабочую нагрузку за минуты, внося изменения в конфигурацию серверов, сетей и устройств хранения данных – без необходимости повторной прокладки кабелей, перенастройки или перезагрузки программного обеспечения. Новое решение гладко интегрируется в рамках существующих, неоднородных сред и поддерживает как физические, так и виртуальные рабочие нагрузки. Это позволяет ИТ-администраторам развертывать образы физических серверов так же быстро, как виртуальных, уменьшая количество отказоустойчивых серверов, и быстро реагировать на изменение требований со стороны приложений путем перевода задач на новые системы по мере необходимости.

Готовые бизнес-конфигурации Dell

Готовые бизнес-конфигурации Dell (Dell Business Ready Configurations – BRC) – это решения, которые объединяют новейшие технологии в области серверов, систем хранения данных и сетевых систем простым для понимания, интеграции и приобретения способом. Каждая конфигурация сопровождается техническими характеристиками, практическими рекомендациями и указаниями по организации сети.

Dell представила первую такую готовую конфигурацию, цель которой – радикальное сокращение времени развертывания инфраструктуры заказчиками. Эта BRC основана на двух блейд-серверах PowerEdge M610 с программным обеспечением Infrastructure Manager, решении сети хранения данных (SAN) EqualLogic PS6000 ISCSI Storage Area Network, двух блейд-коммутаторах PowerConnect M6220 и 24-портовом сетевом коммутаторе Brocade Foundry 424 или Dell PowerConnect 6224. Готовые конфигурации поставляются в стандартной редакции или в редакции с резервированием.

Dell Lifecycle Controller версии 1.3

Dell Lifecycle Controller – это встроенное решение управления, предназначенное для облегчения развертывания и технического обслуживания серверов PowerEdge 11-го поколения. Новая версия 1.3 позволяет администраторам серверов дистанционно обнаруживать серверы в сети, обновлять необходимые драйверы и развертывать операционные системы практически в любое время и находясь в любом месте. Замененные пе-