ORACLE

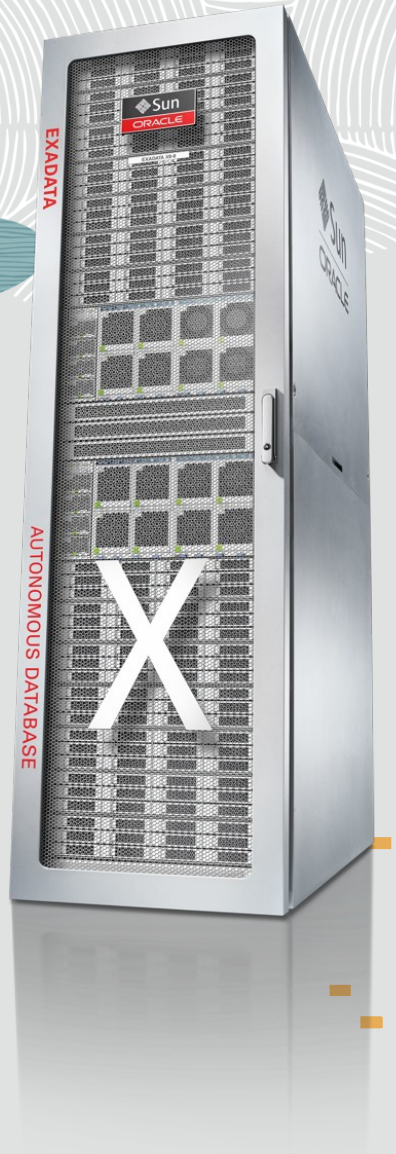
ORACLE

Oracle Exadata Database Machine

Exadata X8M Overview

Николай Диев
Nikolay.diev@oracle.com
Консультант Oracle

декабрь 2019



Safe Harbor

The preceding is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Statements in this presentation relating to Oracle's future plans, expectations, beliefs, intentions and prospects are "forward-looking statements" and are subject to material risks and uncertainties. A detailed discussion of these factors and other risks that affect our business is contained in Oracle's Securities and Exchange Commission (SEC) filings, including our most recent reports on Form 10-K and Form 10-Q under the heading "Risk Factors." These filings are available on the SEC's website or on Oracle's website at http://www.oracle.com/investor. All information in this presentation is current as of September 2019 and Oracle undertakes no duty to update any statement in light of new information or future events.

Концепция Oracle Exadata

Лучшая платформа для всех типов нагрузки Oracle Database
Позволяет получить время отклика БД менее 19 микросекунд

- **Идеальное аппаратное обеспечение для БД** – горизонтальное масштабирование, оптимизированные для СУБД вычислительные мощности, сетевая среда и СХД. Лучшая производительность при оптимальной стоимости.
- **Интеллектуальный системный “софт”** – специализированные алгоритмы существенно улучшают OLTP, Аналитику, Консолидацию
- **Автоматическое управление** – полностью автоматизированные и оптимизированные конфигурирование, производительность, устойчивость к сбоям, обновления

Oracle Exadata – наращивание технологий

Современная версия Exadata позволяет получить время отклика базы данных менее 19 микросекунд

Oracle Database Machine

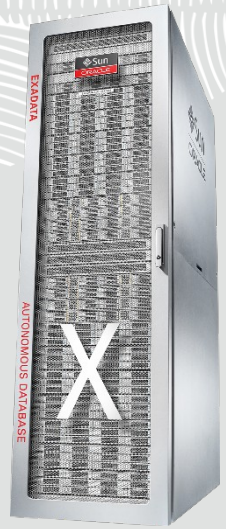
Sun Oracle Database Machine

Ускорение задач OLTP, Аналитики, Консолидации

Ускорение задач In-Memory

Exadata Cloud Service
Exadata Cloud at Customer

Gen 2
Exadata Cloud at Customer



Exadata V1 (2008)
- DDR Infiniband
- Smart Scan

Exadata V2 (2009)
- + QDR Infiniband
- + Flash Cache
- + Hybrid Columnar Compression

Exadata X2 (2010)

Exadata X3 (2011)
- + Flash Cache Write-Back

Exadata X4 (2012)
- + Active/Active Infiniband

Exadata X5 (2013–2014)
- + NVMe
- + Elastic Config
- + Columnar Flash Cache
- + VM Support

Exadata X6 (2016)

Exadata X7 (2017)
- + Storage Tier
- In-Memory Analytics
- + Hot Swap Flash Card

Exadata X8 (2019)
- + XT cells
- + Automatic Indexing
- + 60% Performance Boost

Exadata X8M (2019)
- + PMEM
- + RoCE
- + 160% Performance Boost

ORACLE
Exadata

Exadata Hardware

Exadata: Built-in High Availability

Дублированные серверы баз данных

- Active-Active кластеризованная серверная структура
- Компоненты с горячей заменой
- Дублированная система питания
- Интегрированный стек ПО и firmware

Дублированные сетевые компоненты

- Дублированные сетевые коммутаторы и соединения
- Клиентский доступ через дублированные сетевые линки bonded networks
- Интегрированный стек ПО и firmware

Дублированное хранилище данных

- Зеркалирование данных между серверами хранения
- Дублированные каналы ввода-вывода
- Интегрированный стек ПО и firmware

Отличия Exadata X8 (InfiniBand) и Exadata X8M

Exadata X8 (IB)

- **Database Server Specifications**
  - InfiniBand 4X QDR (40Gb/s)
  - HCA Xen Virtualization
- **Unified Network Technology**
  - InfiniBand (QDR – 40Gbps)
- **Storage Servers**
  - Exadata X8 HC, EF, and XT
    - InfiniBand 4X QDR (40Gb/s) HCA

Exadata X8M

- **Database Server Specifications**
  - QSFP28 100GbE RDMA Ethernet Card
  - KVM Virtualization
- **Unified RDMA Network Fabric**
  - RDMA over Converged Ethernet (RoCE) 100GbE
- **Storage Servers**
  - Exadata X8-2 HC / EF Storage Server
    - QSFP28 100GbE RDMA Ethernet Card
    - 128 GB Intel® Optane™ DC Persistent Memory Modules
  - Exadata X8-2 XT Storage Server
    - QSFP28 100GbE RDMA Ethernet Card

Масштабируемость от Eighth-Rack до Multi-Rack

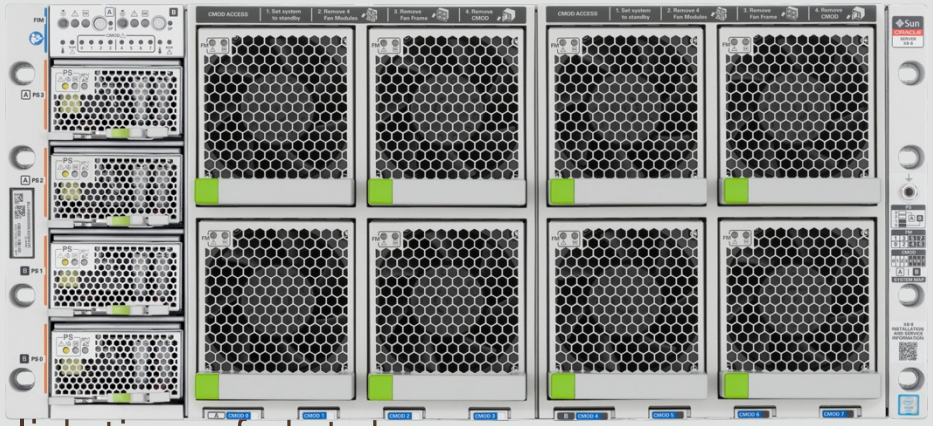
- Начиная с 2 серверов БД и 3 серверов хранения
  - Добавляя сервера БД и хранения по мене необходимости
  - Используя Oracle Exadata Configuration Assistant (OECA) для конфигурирования

Два типа серверов Баз Данных

- X8M-2 использует x86 2-socket серверы БД
  - Наиболее распростаненная Exadata
  - Доступна от Eighth Rack и Quarter Rack до…..

X8M-8 использует x86 8-socket серверы БД

- Доступна от Half Rack и до …..
- Higher-end configuration
  - Larger SMP nodes
  - More memory
- Well suited for high-end OLTP workloads, large scale consolidation of databases, memory intensive workloads, and multi-rack data warehouses



- Идентичные серверыхранения и **Сетевая структура & RDMA**

Exadata X8M-2 Database Server plus **Network Fabric**

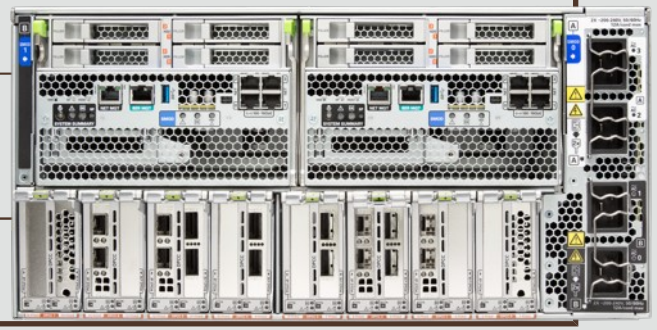
24-core “Cascade Lake” CPUs with up to 1.5TB of Memory

Processors	2 Twenty-four-Core Intel® Xeon® 8260 Processors (2.4 GHz)
DDR4 Memory	384 GB (12 x 32 GB) / 768 GB (12 x 64 GB) – Expandable to 1.5 TB (24 x 64 GB) via memory kit
Local Disks	4 x 1.2 TB 10K RPM SAS Disks (Hot-Swappable)
Disk Controller	Disk Controller HBA with 2 GB Cache – No More batteries
Network	2 x QSFP28 100G RDMA Ethernet Card Ports (PCIe 3.0) – All Ports Active 2 x 10G Base-T Ethernet Ports ***OR*** 2 x 10G/25G Ethernet SFP28 Ports LAN on Motherboard (LOM) based on the Broadcom Limited BCM57417 NetXtreme-E 10G/25G RDMA Ethernet Controller 2 x 10G/25G Ethernet SFP28 Ports Add in Card (AIC) based on the Broadcom Limited BCM57414 NetXtreme-E 10G/25G RDMA Ethernet Controller Click for Port Details
Remote Management	1 Ethernet port (ILOM) 1 Ethernet Port (HOST)
Power Supplies	Redundant Hot-Swappable power supplies and fans

Exadata X8M-8 Database Server plus **Network Fabric**

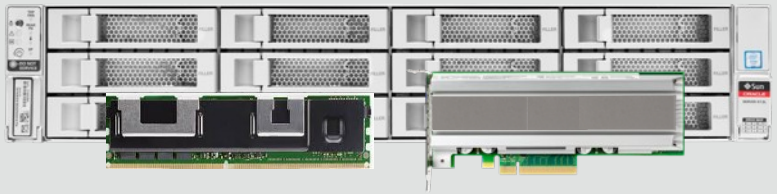
24-core “Cascade Lake” CPUs with up to 6TB of Memory

Processors	8 Twenty-four-Core Intel® Xeon® 8268 Processors (2.9 GHz)
DDR4 Memory	3 TB (48 x 64 GB) – Expandable to 12 TB (96 x 64 GB) via memory kit
Local Disks	2 x 6.4 TB 2.5-inch Flash Accelerator F640 PCIe Cards (Hot-Pluggable)
Disk Controller	Disk Controller HBA with 2 GB Cache – No More batteries
Network	8 x QSFP28 100G RDMA Ethernet Card Ports (PCIe 3.0) – All Ports Active 8 x 10GbE Base-T Ethernet Ports (2 Quad-port Intel Corporation Ethernet Connection X722 for 10GBASE-T)—One used for Host ADMIN 8 x 10G/25G Ethernet SFP28 Ports Add in Cards (AIC) based on the Broadcom Limited BCM57414 NetXtreme-E 10G/25G RDMA Ethernet Controller Click for Port Details
Remote Management	1 Ethernet port (ILOM)
Power Supplies	Redundant Hot-Swappable power supplies and fans

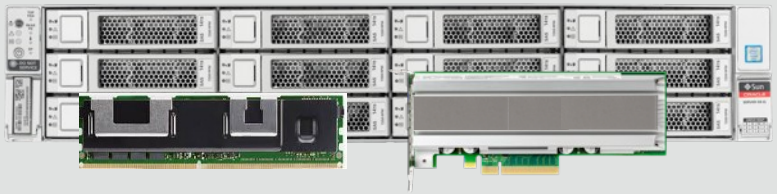


Три типа серверов хранения

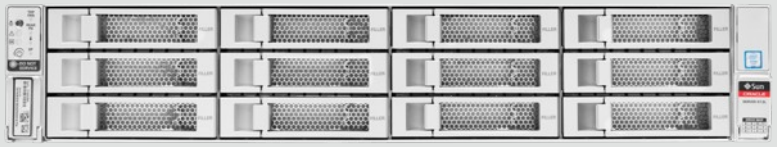
- **X8M-2 Extreme Flash (EF) Storage Server**
- Industry Leading I/O Performance
- All Flash, Scale-out, Highly Available, **RDMA** Connected Smart Storage
- **12 x 128 GB Persistent Memory Modules**
- 8 x 6.4 TB F640 PCIe rear-mounted flash cards per storage server—"Hot Plug" Replaceable
  - State-of-the-art NVMe interface optimized for low-overheard
  - No flash cache misses, so predictably low flash response times



- **X8M-2 High Capacity (HC) Storage Server**
- More Flash, Higher Performance
- Tiered, Scale-out, Highly Available, **RDMA** Connected Smart Storage
- **12 x 128 GB Persistent Memory Modules**
- 12 front mounted 14 TB High Capacity Disks
- 4 x 6.4 TB F640 PCIe rear-mounted flash cards per storage server—"Hot Plug" Replaceable
  - State-of-the-art NVMe interface optimized for low-overheard
  - Smart Flash Cache intelligently manages flash



- **X8M-2 Extended (XT) Storage Server**
- Long Term / Archive Storage Solution
- Tiered, Scale-out, Highly Available, **RDMA** Connected Smart Storage
- 12 front mounted 14 TB High Capacity Disks
  - 3192 TB per Rack



Exadata X8-2 EF Storage Server plus **Network Fabric**

16-core “Cascade Lake” CPUs, PMEM, NVMe Flash Cards, No Spinning Drive

Processors	2 Sixteen-Core Intel® Xeon® 5218 Processors (2.3 GHz)
DDR4 Memory	192 GB (12 x 16 GB)
Persistent Memory	**1.5 TB (12 x 128 GB)**
Drives	8 x 6.4 TB 2.5-inch Flash Accelerator F640 PCIe Drives Non Volatile Memory Express (NVMe) protocol 2 x 240 GB M.2 Drives performing Boot and Rescue Functions—No More USBs
Network	2 x **QSFP28 100G RDMA Ethernet Card** Ports (PCIe 3.0) – All Ports Active Embedded Gigabit Ethernet Ports for management connectivity
Remote Management	1 x Ethernet port (ILOM) 1 x Ethernet Port (HOST)
Power Supplies	Redundant Hot-Swappable power supplies

Exadata X8-2 HC Storage Server plus **Network Fabric**

New 16-core “Cascade Lake” CPUs, PMEM, NVMe Flash Cards

Processors	2 Sixteen-Core Intel® Xeon® 5218 Processors (2.3 Ghz)
DDR4 Memory	192 GB (12 x 16 GB)
Persistent Memory	**1.5 TB (12 x 128 GB)**
Flash	4 x 6.4 TB Flash Accelerator F640 PCIe Card Non Volatile Memory Express (NVMe) protocol
Disk Controller	Disk Controller HBA with 2 GB Cache - No more batteries
Network	2 x **QSFP28 100G RDMA Ethernet Card** Ports (PCIe 3.0) – All Ports Active Embedded Gigabit Ethernet Ports for management connectivity
Remote Management	1 x Ethernet port (ILOM) 1 x Ethernet Port (HOST)
Power Supplies	Redundant Hot-Swappable power supplies a

Exadata X8-2 XT Storage Server plus **Network Fabric**

16-core “Cascade Lake” CPU

Processors	1 Sixteen-Core Intel® Xeon® 5218 Processor (2.3 GHz)
DDR4 Memory	96 GB (6 x 16 GB)
Disk Controller	Disk Controller HBA with 2 GB Cache
Network	2 **QSFP28 100G RDMA Ethernet Card** Ports (PCIe 3.0) – All Ports Active Embedded Gigabit Ethernet Ports for management connectivity
Remote Management	1 x Ethernet port (ILOM) 1 x Ethernet Port (HOST)
Power Supplies	Redundant Hot-Swappable power supplies ar

Exadata X8-2 Eighth Rack Storage Server HC plus **Network Fabric**

Processors	2 Sixteen-Core Intel® Xeon® **5218** Processors (**2.3** GHz), **16** cores enabled
DDR4 Memory	192 GB (12 x 16 GB)
Persistent Memory	**1.5 TB (12 x 128 GB)**
Flash	2 x 6.4 TB Flash Accelerator F640 PCIe Card Non Volatile Memory Express (NVMe) protocol
Disk Controller	Disk Controller HBA with 2 GB Cache - No more batteries
Network	2 x **QSFP28 100G RDMA Ethernet Card** Ports (PCIe 3.0) – All Ports Active Embedded Gigabit Ethernet Ports for management connectivity
Remote Management	1 x Ethernet port (ILOM) 1 x Ethernet Port (HOST)
Power Supplies	Redundant Hot-Swappable power supplies and

Exadata использует сбалансированные, но стандартные аппаратные компоненты. Что же позволяет демонстрировать выдающиеся результаты?

Exadata X8M ключевые отличия :

- энергонезависимая память (PMEM)
- RDMA over Converged Ethernet (RoCE)

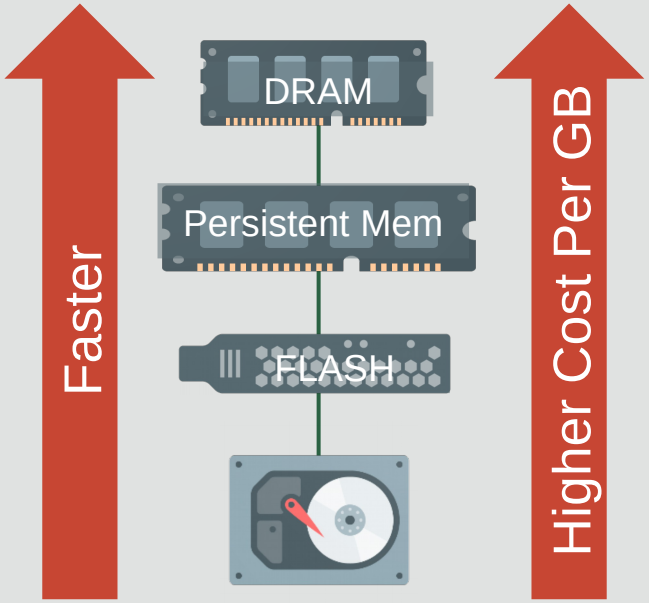
Энергонезависимая Память (PMEM)

Энергонезависимая память

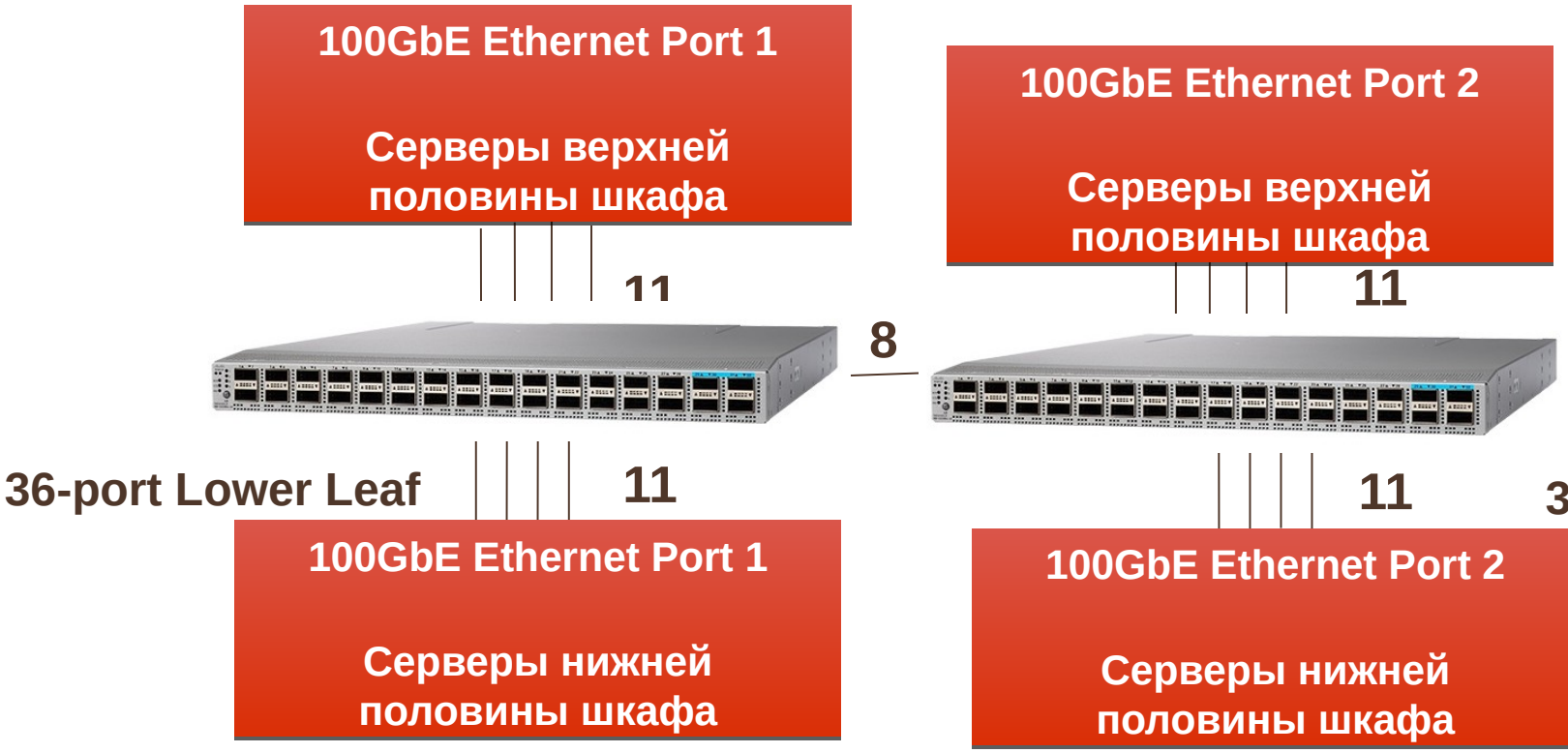
- Объём, производительность и стоимость между DRAM и Флэш

Intel® Optane™ DC Persistent Memory:

- Скорость обмена сравнима с DRAM
- Энергонезависимость в отличии от DRAM
- Использование PMEM Application Mode
- Размещение в PMEM наиболее горячих данных и их перемещение на Flash/HDD по специальному алгоритму



Топология подключений внутренней сети Exadata



- Каждый сервер имеет дублированные подключения

RDMA Network Fabric

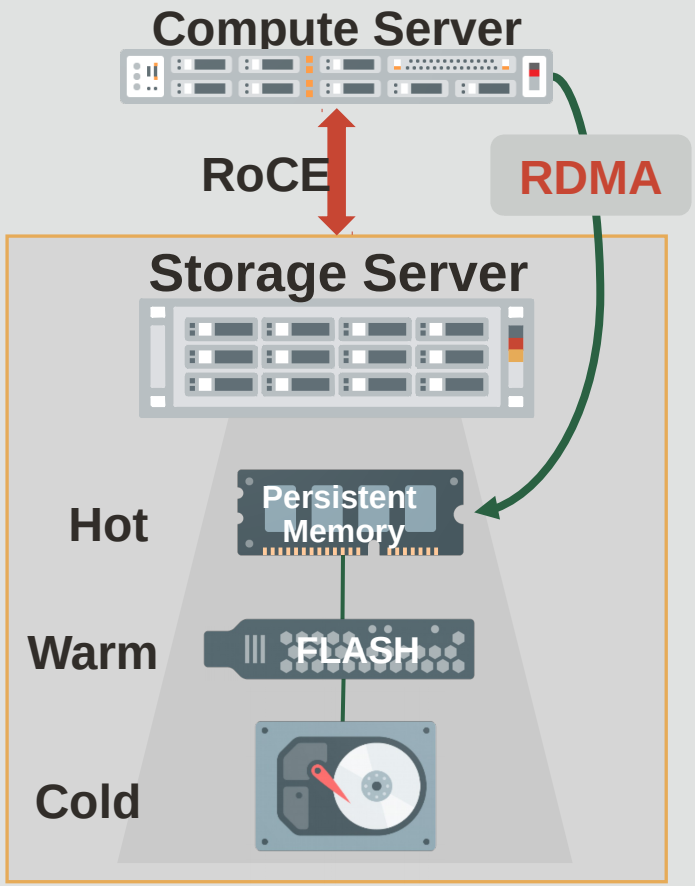
- Unified RoCE Network Fabric
  - Storage Network
  - RAC Cluster Interconnect
  - External Connectivity (optional)
- High Performance, Low Latency Network
  - 200 GbE bandwidth per link (100 GbE each direction)
  - SAN-like Efficiency (Zero copy, buffer reservation)
  - Simple manageability like IP network
- Protocols
  - Zero-copy Zero-loss Database Communication Protocol (ZDP RDSv3) Used by Database
    - Low CPU overhead (Transfer multiple GB/s with 2% CPU usage)
  - Internet Protocol over InfiniBand (IPoIB)
    - Looks like normal Ethernet to host software (tcp/ip, udp, http, ssh,…)

Использование Exadata RDMA для повышения производительности

- Remote Direct Memory Access (RDMA) предоставляет компьютеру доступ к памяти другого компьютера минуя слои операционной системы
  - Сетевой контроллер имеет прямой доступ чтения/записи данных в памяти с очень низкими задержками
- RDMA – встроенная часть высокопроизводительной архитектуры Exadata:
  - Высокая пропускная способность с низкой нагрузкой на CPU для передачи больших объемов данных
  - Уникальный протокол Direct-to-Wire для межузлового оповещения OLTP (3-х кратное ускорение)
  - Уникальный протокол RDMA для координации межузловых транзакций
  - Прямой доступ к данным в PMEM серверов хрвнения
  - Сверхнизкие задержки записи database logs в энергонезависисую память СХД

Exadata X8M : RoCE + PMEM

Оптимизированное использование для СУБД энергонезависимой памяти



СУБД использует RDMA для обмена данными в PMEM на серверах СХД

В 10x ниже задержки: менее 19 µsec для 8K чтений БД

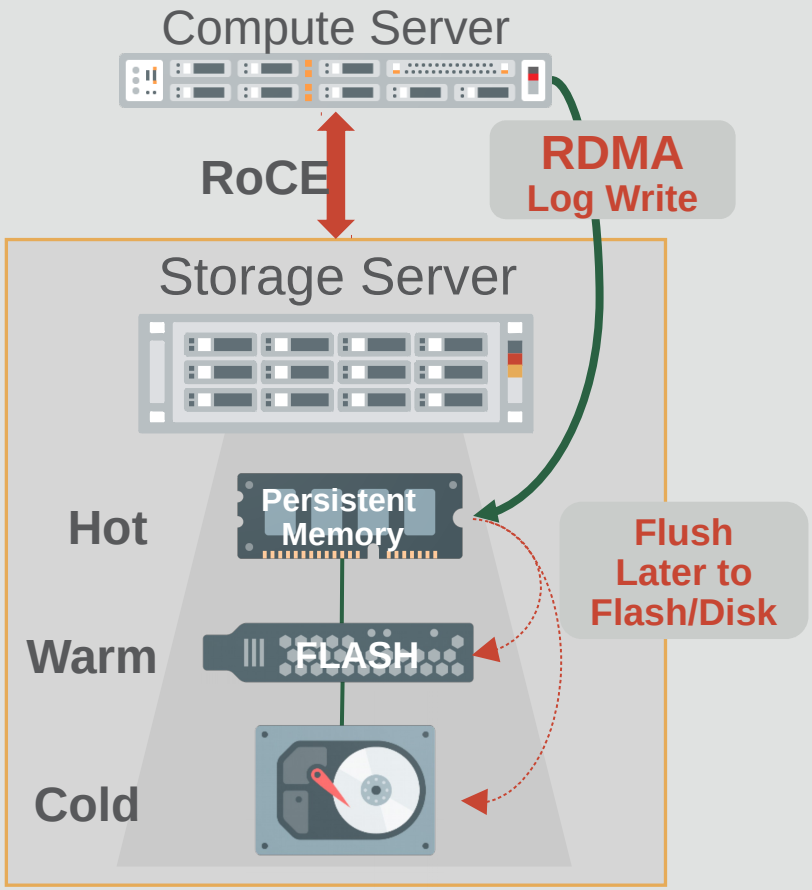
PMEM автоматически одновременно используется для множества БД на Exadata

Используется как КЭШ для самых “горячих” данных увеличивая скорость обработки в десятки раз

Exadata СХД Серверы прозрачно добавляют PMEM Ускоритель Данных перед флэш памятью

В 2.5x раза больше IOPS чем на предыдущих схемах – 16 Миллионов IOPS

Exadata X8M : RoCE + PMEM



Сокращение времени выполнения записи в журнальный файл критично для производительности OLTP

Быстрая запись redo = быстрый отклик БД

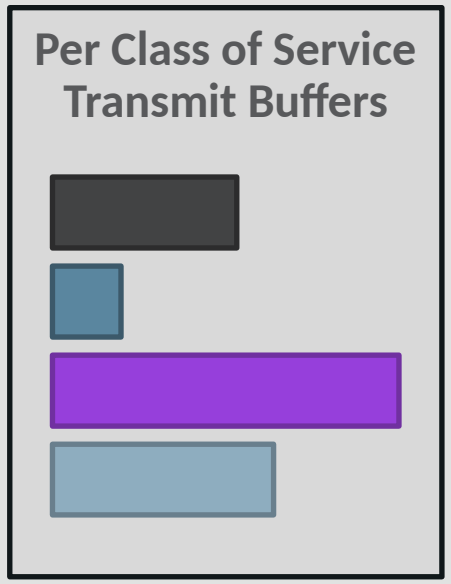
Запись в журналы ускоряется в разы

Любые задержки записи могут вызывать каскадные ожидания множества сессий

Зеркалирование данных PMEM

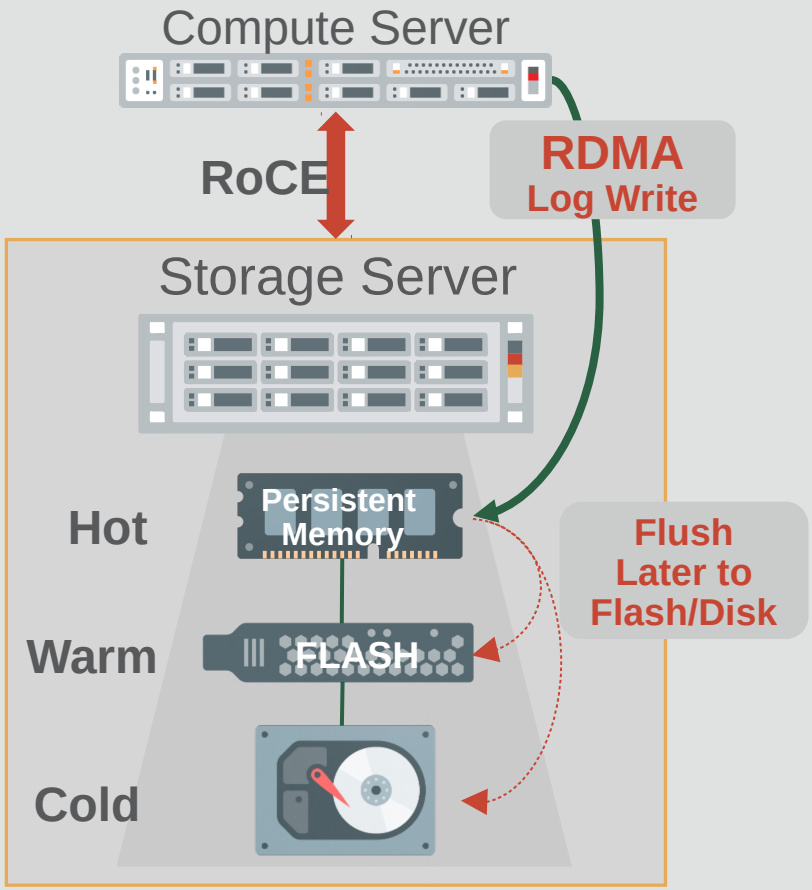
Данные в PMEM зеркалируются СУБД между серверами хранения (СХД), что исключает потерю целостности данных при отключении одного из серверов СХД

Exadata X8M RoCE – Приоритезация обмена



- Алгоритмы СУБД для приоритезации обмена
  - Исключение задержек для данных с высоким приоритетом со стороны другого типа обмена
  - Примеры Low latency данных: cluster heartbeat, transaction commit, cache fusion
  - Низкоприоритетные операции: backup, reporting, batch
- RoCE Class of Service (CoS)
  - Обеспеыение независимости пересылки пакетов по нескольким каналам, с присвоением каждому своего буфера
- Exadata автоматически оптимизирует обмен СУБД, используя политики CoS

Зависимость производительности Exadata от версий ПО



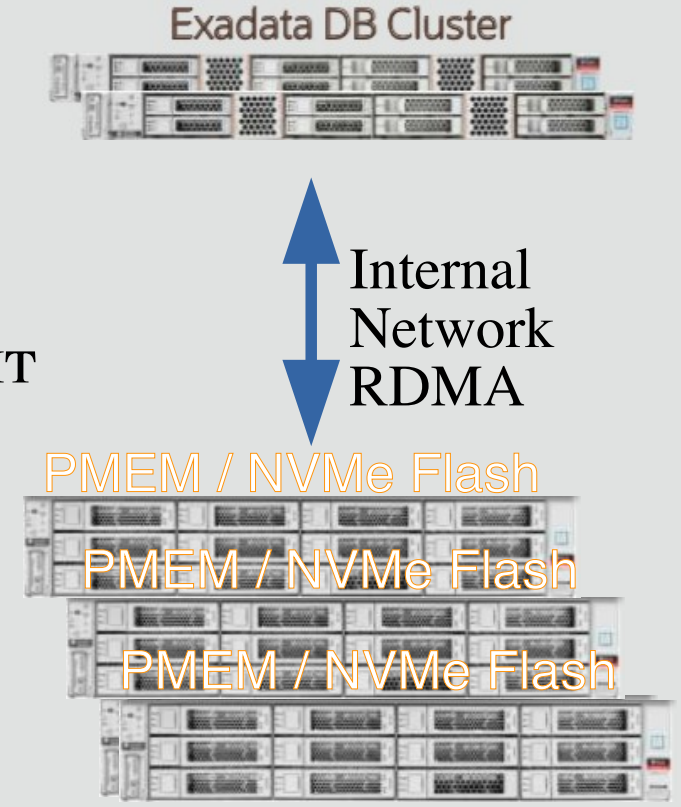
- DB 19 (включая GI 19) + Exadata System SW 19.3 с включенным PMEMCache
  - Латентность ~ 19usec
- DB 18, 12, 11.2 + Exa ПО 19.3 Cell с PMEMCache
  - ~100 usec
- Без использования PMEMCache
  - ~200 usec

Exadata уникально оптимизирует OLTP

Передовые технологии, настроенные для массированного произвольного I/O с предсказуемой низкой задержкой
- горизонтально масштабируемая СХД
- PMEM & PCIe NVMe Flash
- Сеть 100GbE
- Smart Flash Logging
- OLTP Cache на СХД

Непрерывные мониторинг и обнаружение сбоев компонент
- Автоматическое обнаружение сбоев без таймаута
- Перенаправление I/O со сбоящего устройства в доли секунды

Исключение узких мест в координажии кластера БД
- Direct-To-Wire протокол → обмен OLTP между серверами в 3 раза быстрей
- Smart Fusion Block Transfer исключает ожидания записи redo при передаче между узлами
- Протокол RDMA для координации транзакций между узлами



Exadata Storage Server

Exadata уникально оптимизирует Аналитику

Smart Scan (SQL Offload)

интенсивная обработка данных* выполняется на СХД, снижая нагрузку на каналы и ЦРУ серверов СУБД

Smart Scan (SQL Offload)

активные данные сохраняются на быстрых PCI Flash, неактивные – на недорогих HDD

Storage Indexes

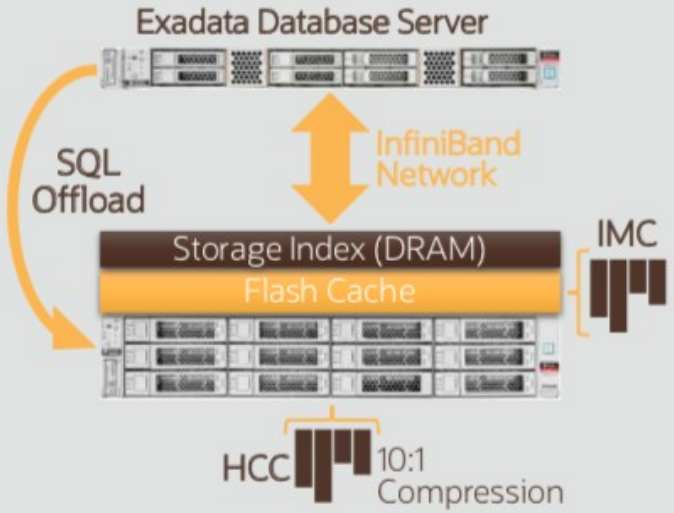
исключают нерелевантный для SQL запроса обмен

Hybrid Columnar Compression (HCC)

поколоночное сжатие выполняется на СХД, экономя пространство, обмен и ускоряя аналитические SQL

In-Memory Columnar (IMC)

использует для опции In-Memory СУБД Flash СХД



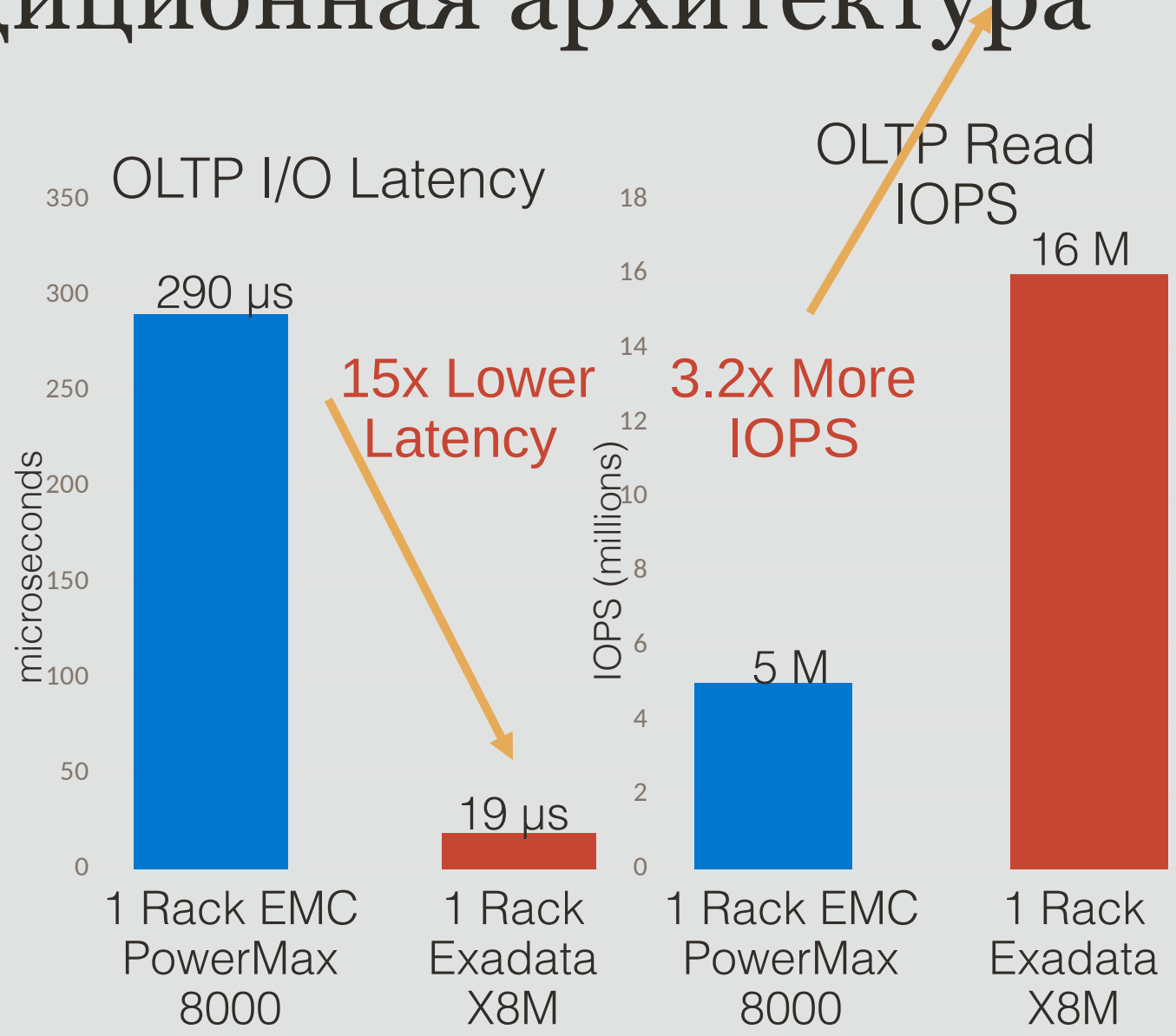
* Включая длительные SQL запросы, резервные копии, дешифрование, фгрегацию, data mining

Exadata vs традиционная архитектура

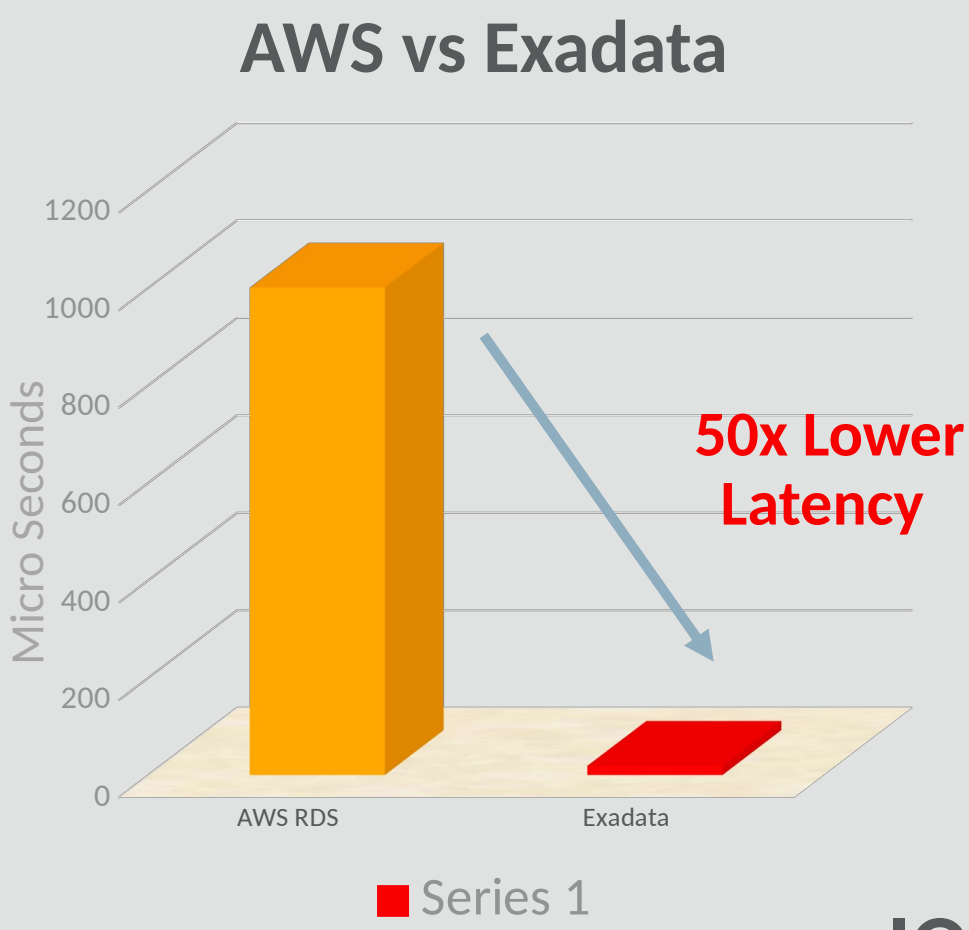
Одна стойка X8M vs СХД EMC PowerMax **all-flash**

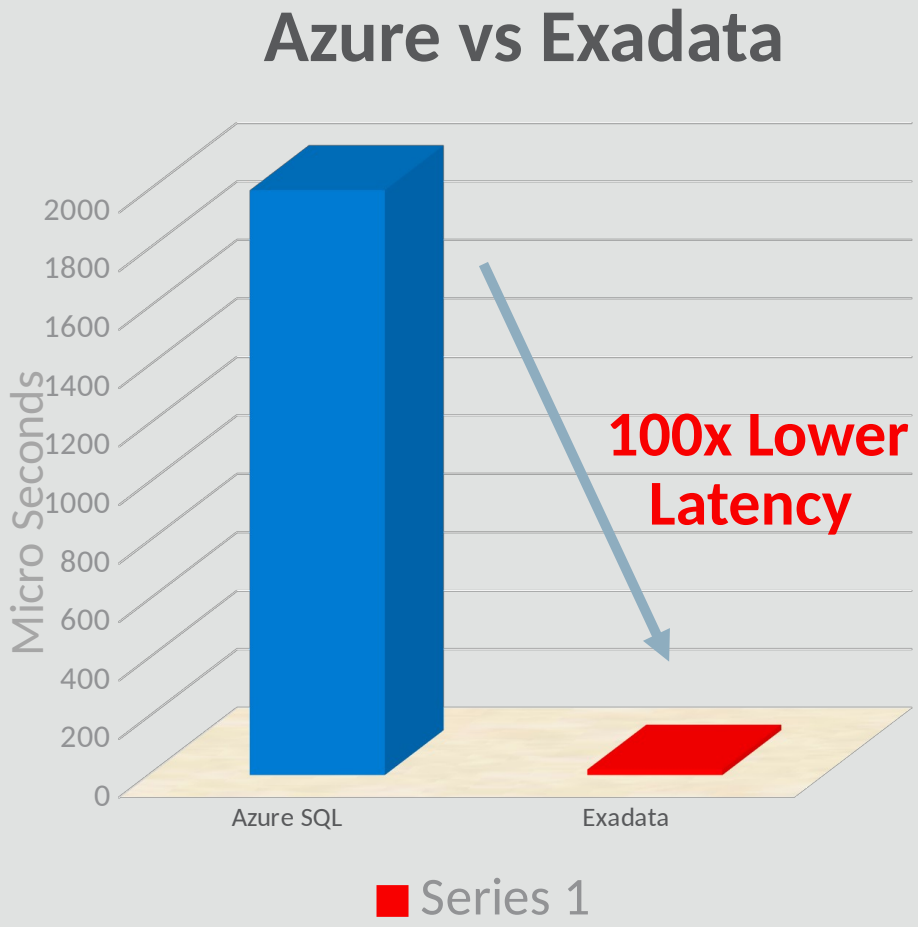
- 7.4x выше сканирование
- 3.2x больше IOPS
- 15x ниже задержки

Производительность Exadata масштабируется при добавлении стоек



Exadata vs AWS & Azure





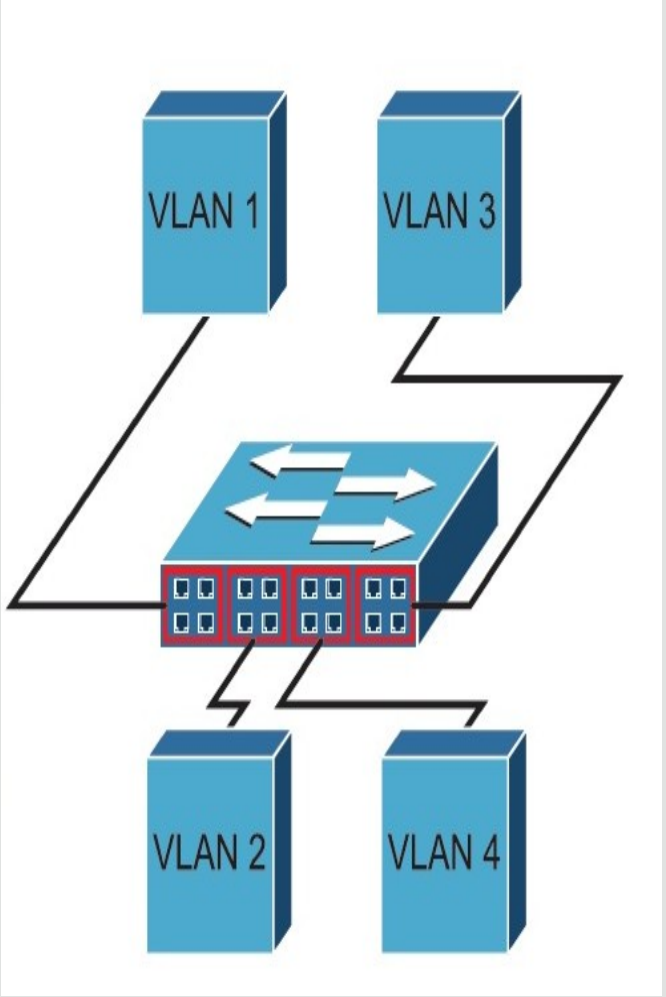
IO Latency

Виртуальные машины Exadata X8M

- Новый гипервизор - KVM
  - Type 2 гипервизор в ядре Linux с улучшенной производительностью
- ВМ изолируют CPU/память и администрирование для различных нагрузок
  - Хостинг, облако, консолидация между отделами, test/dev, не-БД и приложения сторонних производителей
  - Поддерживается множество кластеров в серверах (напр., для SAP)
- ВМ на Exadata обеспечивают производительность близкую “голому” железу
  - СУБД I/O идёт напрямую к высокоскоростному RDMA Network Fabric минуя гипервизор
- Комбинируется с сетью Exadata и приоритезацией I/O для достижения уникальной изоляции во всём стеке
- Trusted Partitions позволяют ограничивать лицензирование

Exadata: Virtual Local Area Network

- Снижение количества занимаемых портов коммутаторов
- OEDA поддерживает создание виртуальных сетей в сети управления, ILOM, клиентской сети и сети резервного копирования
- VLAN tagging требуется при подключении к нескольким коммутаторам
- Поддержка QoS между сетью disaster recovery network и клиентскими сетями
- VLAN tagging разделяет клиентские сети виртуальных машин

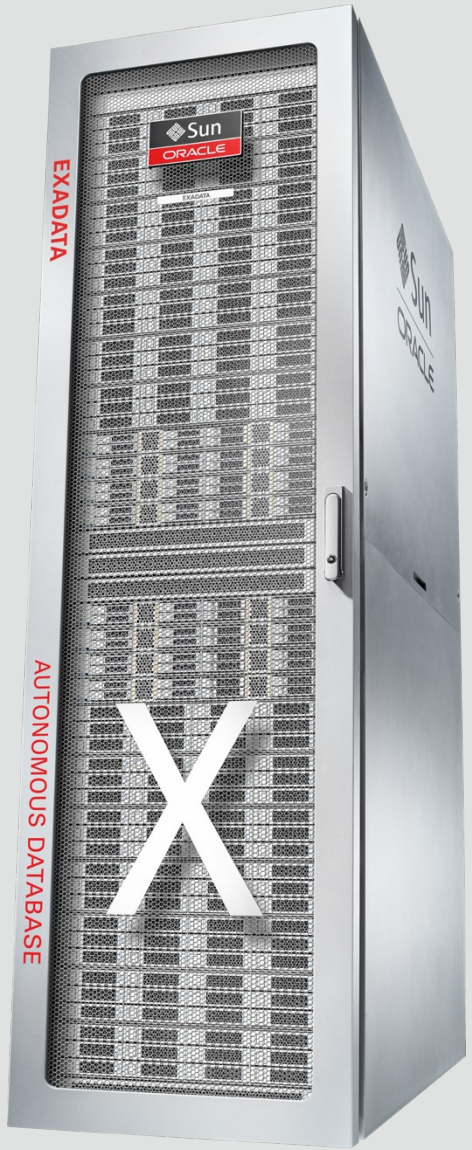


Client VLAN Support

- Обязательное дублирование VLAN (bond)
- Сети резервного копирование и клиентского доступа должны иметь различные tagged VLAN, но могут использовать одни и те же порты
- IPv6 VLAN Exadata Storage Server поддерживается только для сети управления
- Виртуальная конфигурация не поддерживает IPv6 VLANs

Exadata :

Эксплуатация и доступность данных



Резервное копирование Exadata

InfiniBand Network

Storage Expansion Rack and X8-2 Extended (XT)

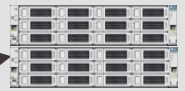
- Fastest Backup and Restore
- ILM Historical Archive
- Second Disk Group
- Expansion of DATA

Public Network

Database Backup Cloud Service

- Offsite Storage
- Low Cost

Media Server



Fiber Channel SAN

IB,10GigE, or 25GigE

10gigE or 25GigE

Fiber Channel SAN

Recovery Appliance

- Delta Push & Backup Validation
- Incremental Forever
- Zero Data Loss Recoverability

Tape library

- Offsite Backups
- Vaulting

Automatic Service Request (ASR)

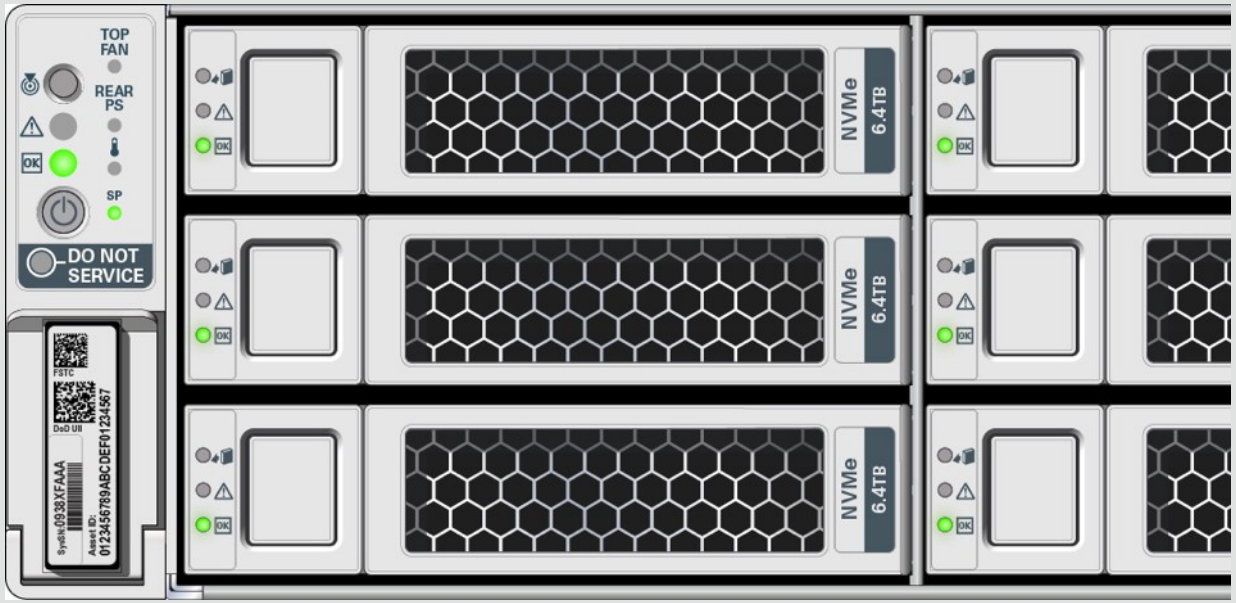
Автоматическое заведение заявок (SR) в службе поддержки Oracle при возникновении аппараных сбоев, таких как ошибки на накопителях (flash & hdd), памяти, процессорах, блоках питания, вентиляторах и т.д.

Механизм ASR может работать совместно с другими системами оповещения о сбоях (таких как SMTP, SNMP и т.д.) и не является их заменой

Exadata X8-2 “Do-Not-Service” LED on Storage Servers

“Application Driven” indicator

Close Up View



- Начиная с версии Exadata X7 : для предотвращения ошибки выключения сервера СХД разработан дополнительный индикатор **показывающий возможность безопасного выключения**
  - Автоматически включается при снижении избыточности, чтобы проинформировать администратора и сервисного инженера о невозможности выключения сервера хранения
  - Поддерживается начиная с Exadata SW 18.1.0.0.0 (см. “What’s New”)

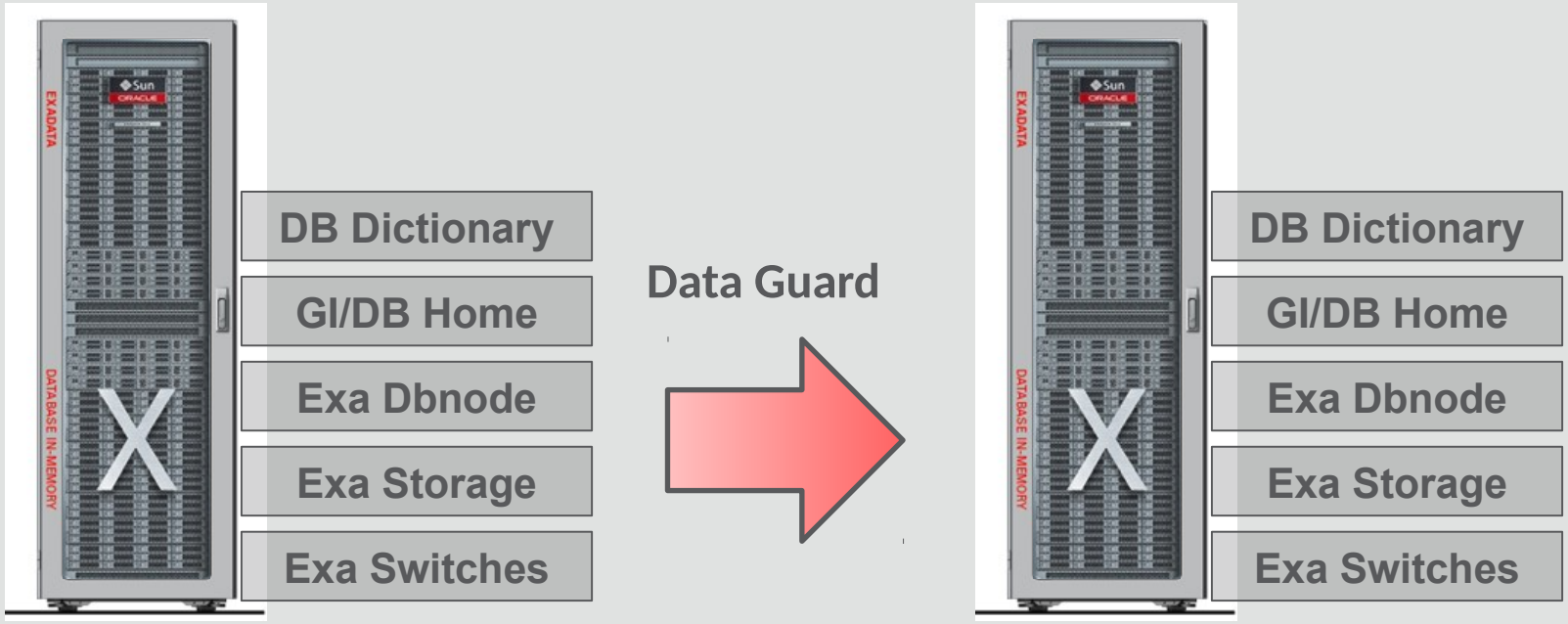
Zero Downtime Software Maintenance

Rolling Software Update Support

Component to Update	How to Mitigate Impact and Risk
Database / Grid Infrastructure	Rolling GI / DB updates with Fleet Patching and Provisioning Application Continuous Availability Data Guard Standby First
Exadata Database Server	Rolling Database Server updates Application Continuous Availability and RHPhelper Data Guard Standby First
Exadata Storage Server	Rolling Storage Server updates ASM HIGH redundancy Data Guard Standby First
Exadata switch	Rolling switch updates Data Guard Standby First

Reduce Risk and Downtime with Data Guard

Data Guard Standby First Patching (MOS 1265700.1)



Site A

Site B

Standby First Patching Steps

1. Update software on Site B (Standby)
2. Test new software
3. Switchover (optional)
4. Update software on Site A
5. Run SQL portion of RU on Primary databases

Стоимость Exadata X8M

Специальная цена для Exadata X8M :

Exadata X8M будет поставляться с сетью RoCE RDMA и с полным комплектом PMEM без увеличения стоимости

1.5 ТБ PMEM на каждый сервер СХД Exadata

21 ТБ PMEM на полную стандартную стойку

Exadata X8 (InfiniBand) будет продолжать поставляться

Для заказчиков расширяющих существующие Exadata

Для заказчиков, которые нуждаются в сертификации X8M до её использования

ORACLE