



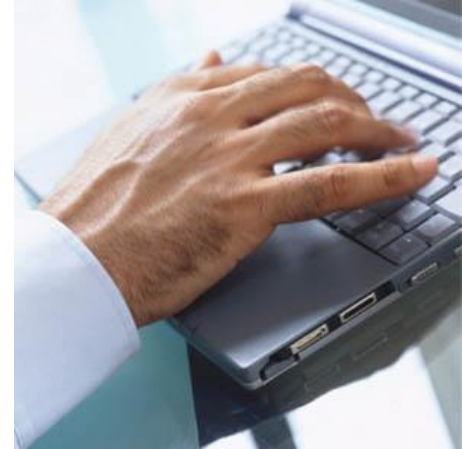
**ORACLE®**

## Обзор Sun Oracle Exadata и Database Machine

Игорь Мельников  
Oracle CIS

# План

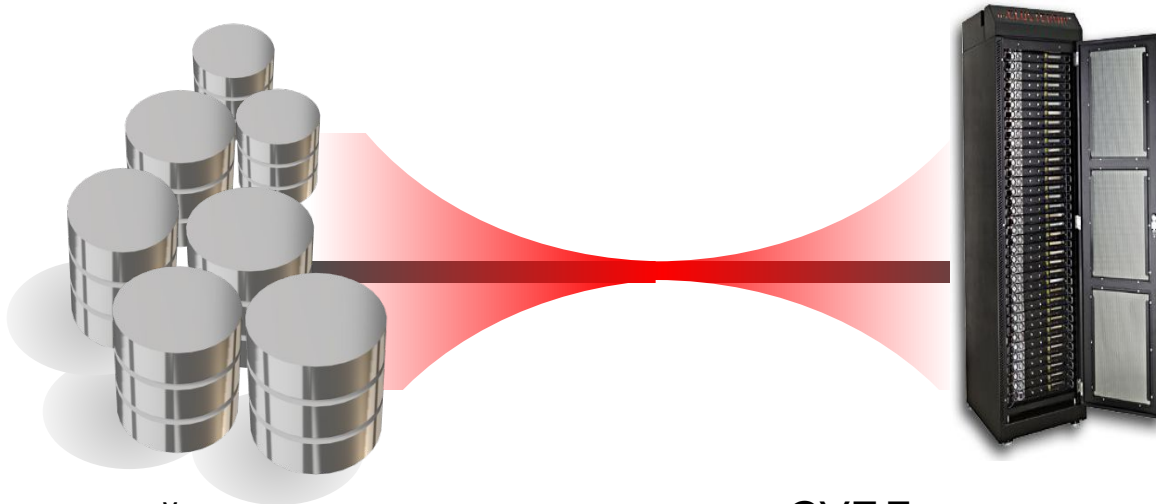
- Введение
- Архитектура и технологии
- Преимущества
- Что говорят заказчики



# Введение

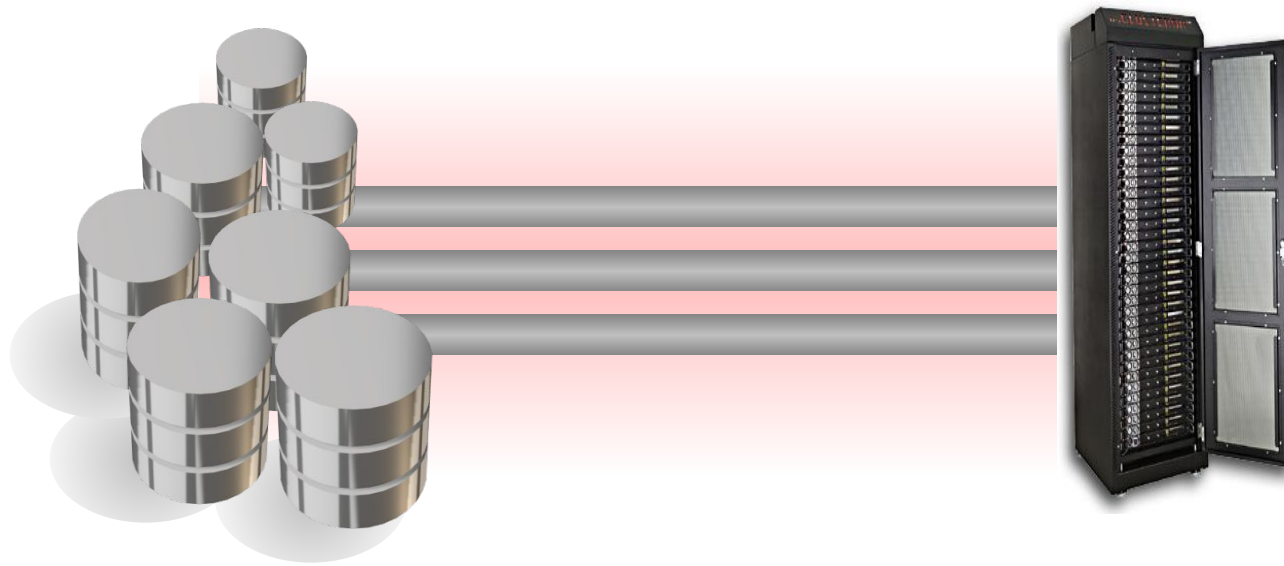


# Узкие места систем хранения



- На сегодняшний день производительность СУБД ограничена возможностями систем хранения
  - Системы хранения ограничены возможностями передачи данных к серверам
  - Внутренние ограничения дисковых массивов
  - Ограничения SAN
  - Скорость дисков ограничивают произвольный ввод/вывод
- Пропускная способность ограничивает производительность хранилищ данных
- Ограничения произвольного ввода/вывода “тормозят” производительность приложений OLTP

# Способы увеличения пропускной способности



- Передавать меньше данных
- Добавить ещё каналов
- Сделать каналы шире

# Exadata – интеллектуальная система хранения

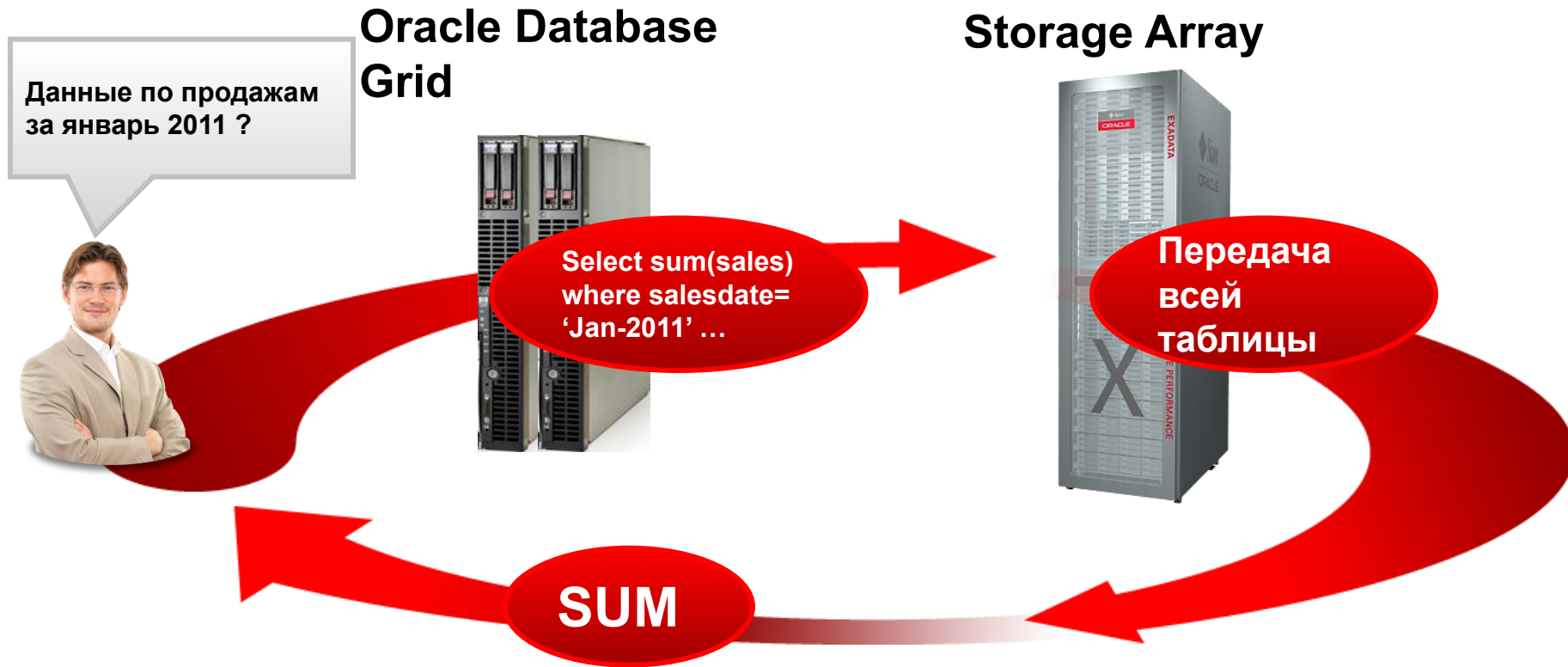
- Oracle решает проблему узких мест в потоках данных 3-мя способами
  - Storage grid с массовым параллелизмом серверов хранения Exadata (ячеек).
    - Пропускная способность растет с увеличением объема данных
  - Интенсивная обработка данных осуществляется в ячейках Exadata.
    - Прimitives операции запросов выполняются над потоком данных с диска, разгружая ЦПУ серверов СУБД
  - Компрессия по столбцам сокращает объем данных в десятки раз
    - Exadata Hybrid Columnar Compression обеспечивает в десятки раз уменьшает стоимость и увеличивает производительность
- Oracle решает проблему ограничения количества произвольных операций I/O с использованием Exadata Smart Flash Cache
  - Увеличение произвольных операций I/O в порядке 20X

## Exadata Storage Cells

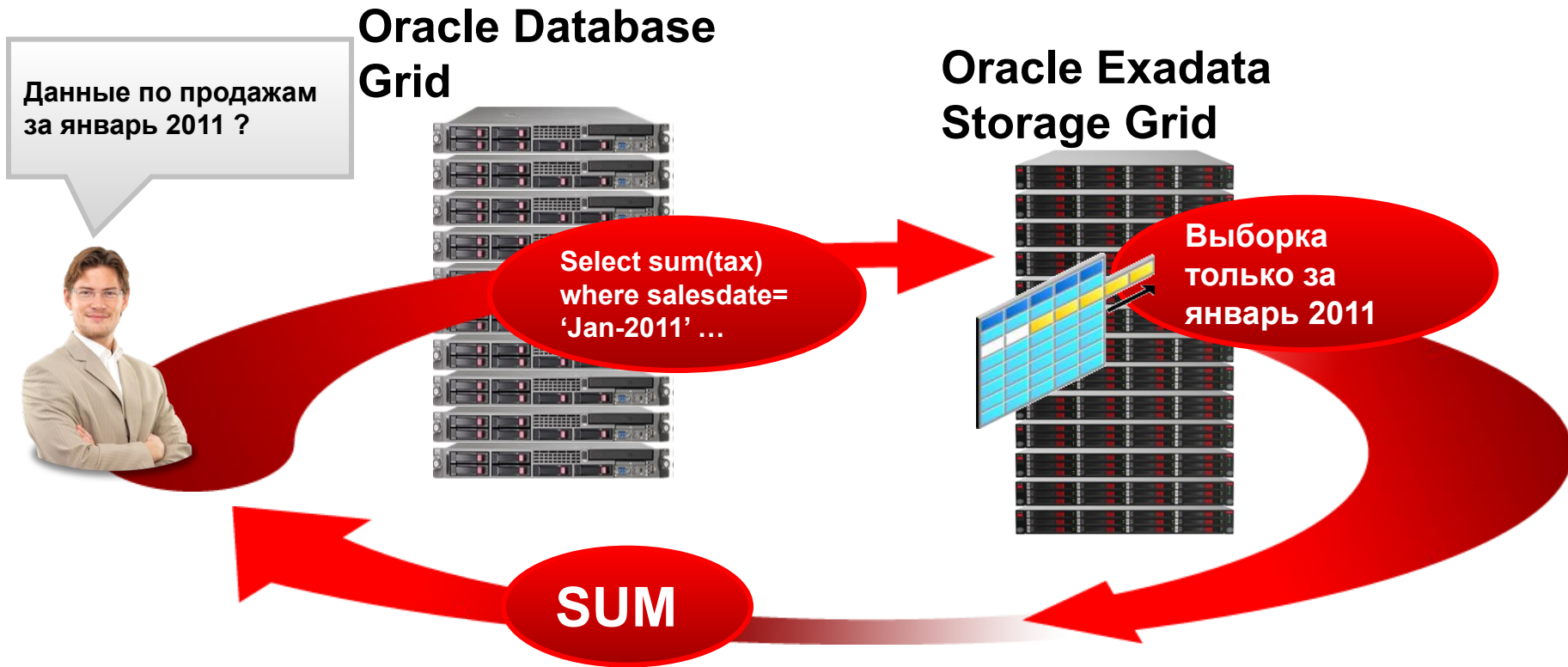


# Обработка запросов:

*Используя обычный дисковый массив*



# Обработка запросов: *Используя Exadata Storage Server*





# Архитектура



# Sun Oracle Database Machine

- Grid – архитектура будущего
  - Высочайшая производительность, низкая стоимость, резервирование, инкрементальная масштабируемость
- Sun Oracle Database Machine первая и единственная полная архитектура для управления данными

## Database Grid

- 8 compute servers (1U) или 2 сервера (4U)
- 96 / 128 Intel cores



## InfiniBand Network

- Скорость 40Gb/s с резервированием



## Storage Grid

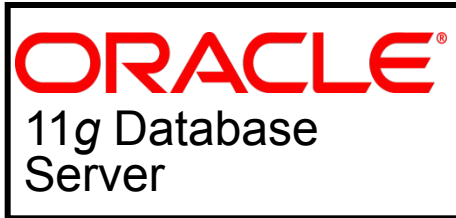
- 14 серверов хранения (2U)



- 168 ядер в системе хранения
- 100 TB SAS, или 336 TB SATA диски
- 5 TB PCI Flash
- Данные зеркалируются между серверами хранения

# Database Machine Software

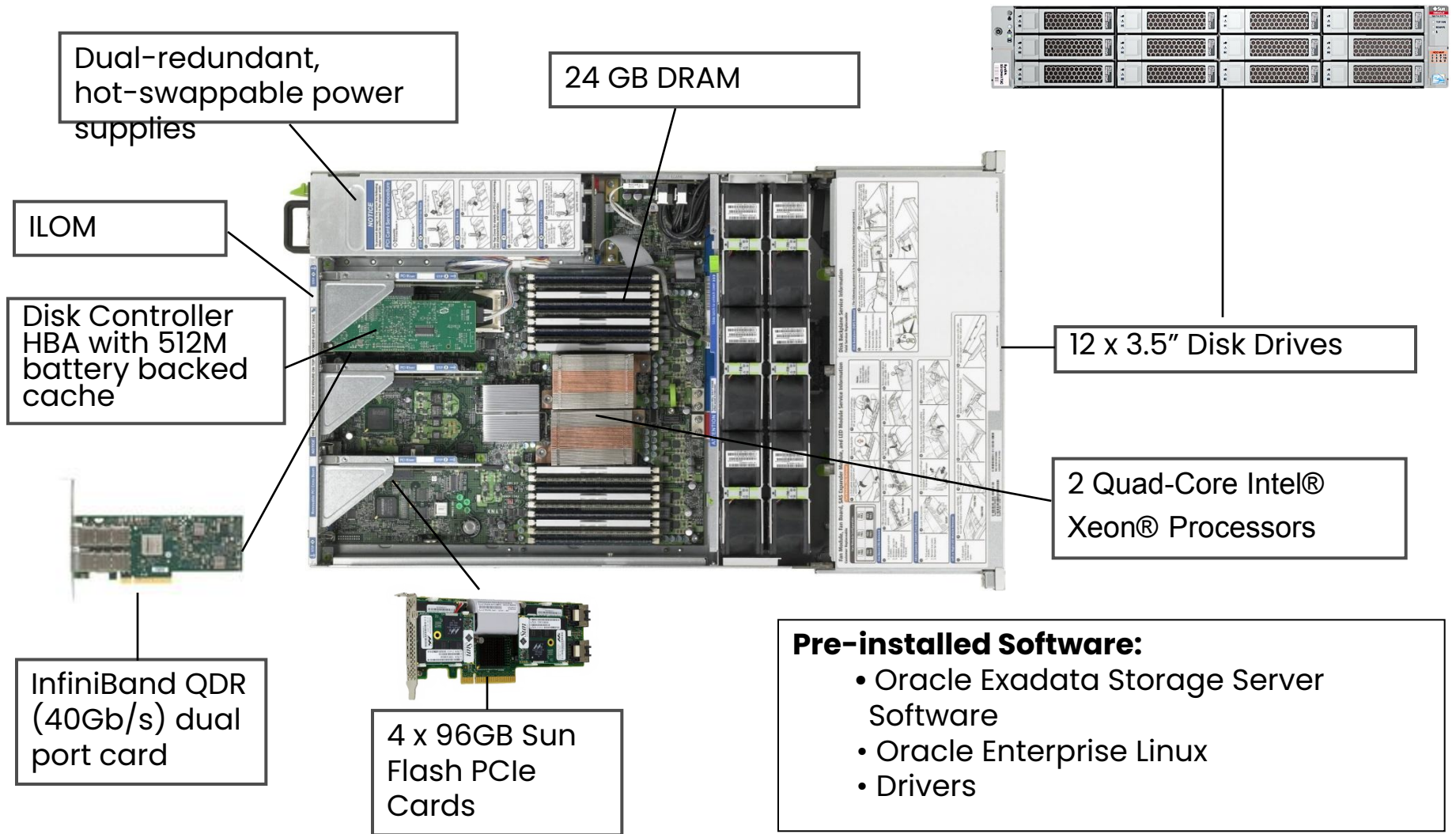
## Архитектура



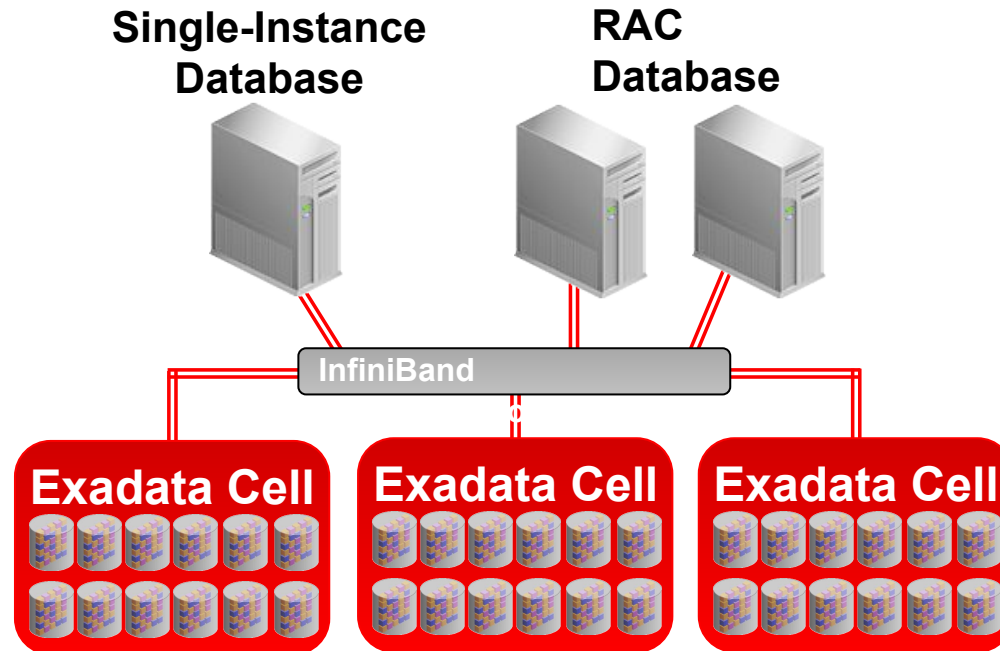
- Oracle Database 11g на 8 или 2 узлах RAC
- ASM обеспечивает зеркалирование, чередование и балансировку
- ПО Exadata обеспечивает smart scan с помощью протокола iDB



# “Железо” Sun Exadata Storage Server



# Конфигурация Exadata



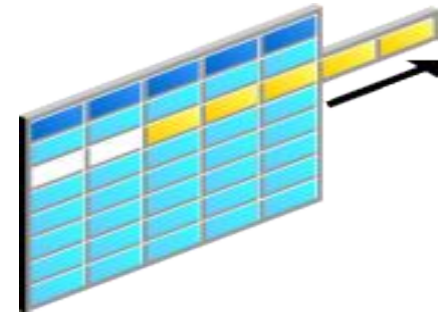
- Каждая ячейка Exadata это самостоятельный сервер, который обеспечивает дисковое пространство и работу ПО Exadata
- БД Oracle располагается на множестве ячеек Exadata
- Производительность СУБД Oracle увеличивается за счет кооперации Exadata Storage Server
- Нет практического лимита на количество ячеек, которые могут быть в GRID

# Возможности ПО Exadata

- **Exadata Smart Scans**
  - Многократное сокращение объема данных в сторону серверов БД
- **Exadata Storage Indexes**
  - Исключает ненужные дисковые операции В/В
- **Hybrid Columnar Compression (HCC)**
  - Компрессия увеличивает эффективность использование дискового пространства и на порядок увеличивает скорость сканирования данных
- **Exadata Smart Flash Cache**
  - Ломает ограничения произвольных операций В/В, увеличивая их количество в 20 раз
  - Удваивает производительность сканирования данных
- **I/O Resource Manager (IORM)**
  - Обеспечивает приоритет операций В/В для обеспечения предсказуемой производительности

# Exadata Smart Scan

- Ячейки Exadata реализуют механизм передачи запросов на сторону хранилища (scan offload) с тем, чтобы значительно уменьшить объем данных возвращаемых на сторону серверов БД
  - Фильтрация строк на основе “where” предиката
  - Фильтрация колонок
  - Фильтрация соединений (join)
  - Фильтрация инкрементального backup
  - **Фильтрация зашифрованных данных**
  - **Работа с функциями Data Mining**
- **10x уменьшение данных является обычным (на тестах заказчиков)**
- Полностью прозрачно для приложения
  - **Даже если происходит сбой ячейки или диска во время запроса**



# Exadata: изменение плана запроса

- TABLE ACCESS STORAGE FULL
- INDEX ACCESS STORAGE FULL
- storage(SYS\_OP\_BLOOM\_FILTER(...))

```
-----  
| Id  | Operation                               | Name          | E-Rows |  
-----  
|  0  | SELECT STATEMENT                       |               |        |  
|  1  |   SORT AGGREGATE                       |               |        |  
|  2  |     PX COORDINATOR                      |               |        |  
|  3  |       PX SEND QC (RANDOM)                | :TQ10000     |        |  
|  4  |          SORT AGGREGATE                  |               |        |  
|  5  |            PX BLOCK ITERATOR             |               |        |  
|*  6  |              TABLE ACCESS STORAGE FULL| SALES        |        |  
-----
```

Predicate Information (identified by operation id):

```
-----
```

```
6 - storage(:Z>=:Z AND :Z<=:Z)  
   filter("PRICE"<25000)
```



# Прозрачность технологии Smart Scan для приложений



- Smart scans прозрачен для приложения
  - Не требуется изменения приложения или SQL кода
  - Возвращаемые данные полностью консистентны
  - В случае выхода из строя ячейки во время smart scan незавершенная часть запроса прозрачно перенаправляется на ячейку, содержащую копию данных
- Smart Scans корректно обрабатывает следующие случаи:
  - неподтвержденные записи (uncommitted) и заблокированные записи
  - Цепочки строк (chained rows)
  - Сжатые таблицы
  - Обработку национальных языков
  - Работа с датами
  - Регулярные выражения
  - Партиционированные таблицы

# Передача функций data mining на сторону Exadata



- Data mining запрос на Exadata:

```
select cust_id
from customers
where region = 'US'
and prediction_probability(churnmod, 'Y' using *) >
0.8;
```

← Функции Data Mining выполняется на Exadata

- Функции Data Mining scoring перегружаются на Exadata
- Выигрыш производительности до 10x раз
- Уменьшает утилизацию ЦПУ на стороне сервера БД

# Exadata Storage Index



*Прозрачно исключает ненужные чтения*

## Таблица Индексы

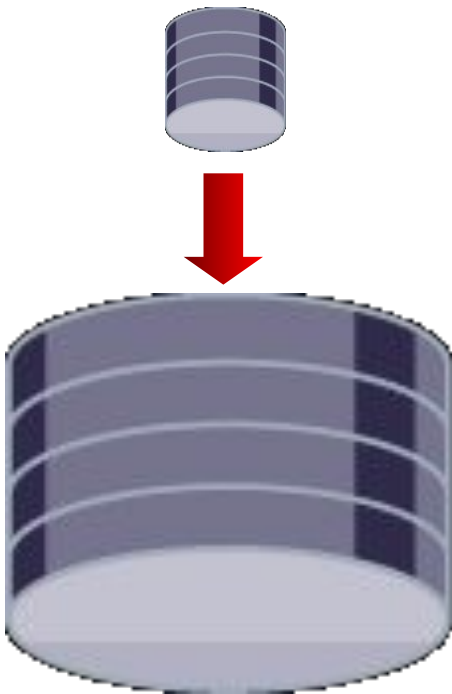
A	B	C	D
	1		
	3		
	5		
	5		
	8		
	3		

Min B = 1  
Max B = 5

Min B = 3  
Max B = 8

- Exadata Storage Index – структуры в памяти, которые хранят информацию о распределении данных между блоками данных.
  - Хранит МИН и МАКС значения для каждой колонки
  - Обычно одна запись в индексе для каждого Мб диска
- Исключает ввод-вывод для тех дисков, где МИН и МАКС не соответствуют условию “where”
- Полностью автоматически и прозрачно

# Проблема роста данных

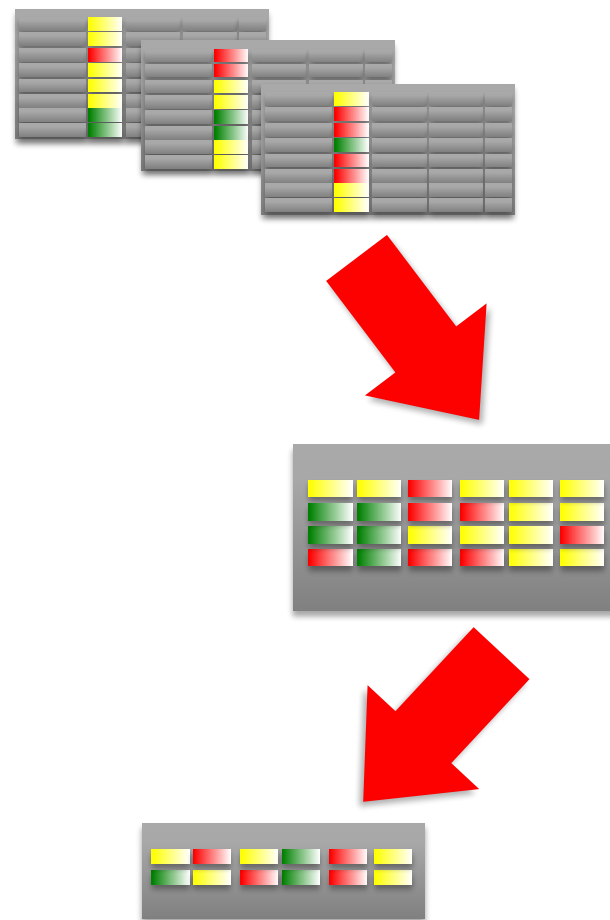


- ИТ-структура должна поддерживать экспоненциальный рост объема данных
  - Без воздействия на производительность
  - Без увеличения стоимости
- Мощное и эффективное сжатие – ключ решения

# Гибридное колоночное сжатие

## Hybrid Columnar Compression

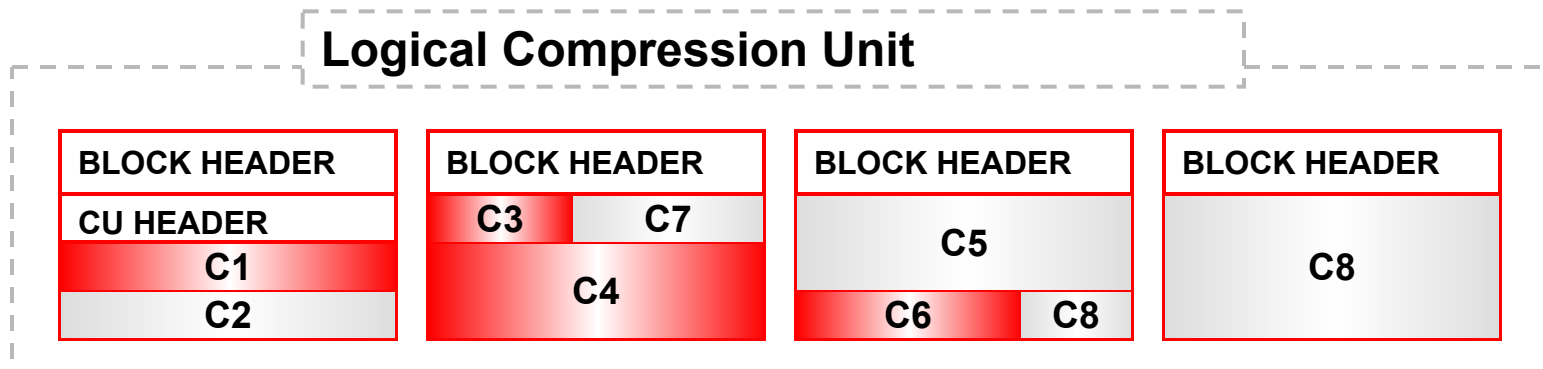
- Данные группируются по колонкам и затем сжимаются
- **Query Mode** для хранилищ данных
  - Оптимизированы для быстрого доступа
  - **10X сжатие**
  - Время сканирования уменьшается соответственно
- **Archival Mode** для редко используемых данных
  - Оптимизировано для уменьшения занимаемого места
  - 15X сжатие
  - До 50X раз для некоторых данных
- **Помощник по сжатию**  
DBMS\_COMPRESSION PL/SQL пакет



# Гибридное колоночное сжатие

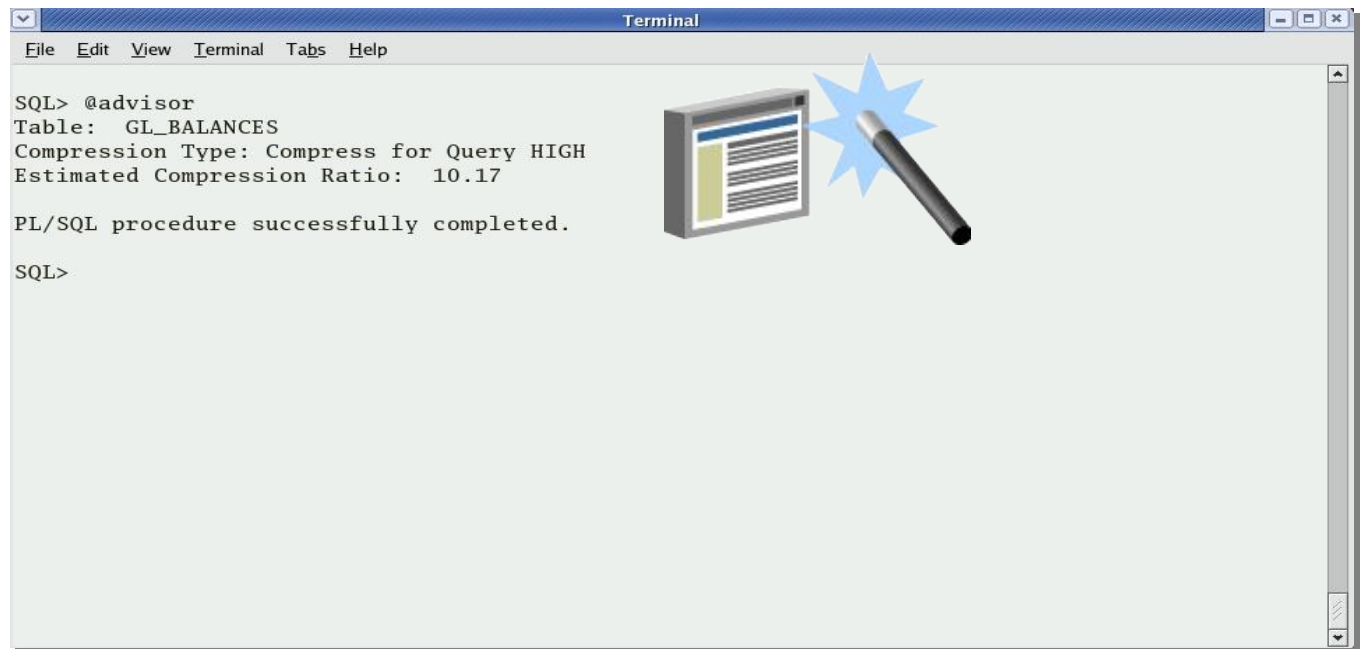
## *Как это работает?*

- Единица сжатия
  - Логическая структура, содержащая несколько блоков данных
  - Данные организуются по колонкам во время загрузки
  - Каждая колонка сжимается отдельно
  - Все данные для этой колонки для всех записей, хранимых в compression unit
  - Редкие изменения



# Помощник по сжатию

- Новый помощник (advisor) в 11g Release 2
  - DBMS\_COMPRESSION PL/SQL пакет
  - Оценивает процент сжатия с помощью гибридной колоночной компрессии на не-Exadata “железе”



```
Terminal
File Edit View Terminal Tabs Help

SQL> @advisor
Table:  GL_BALANCES
Compression Type: Compress for Query HIGH
Estimated Compression Ratio:  10.17

PL/SQL procedure successfully completed.

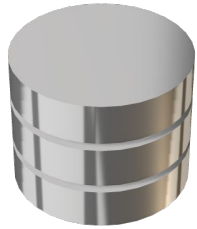
SQL>
```

The image shows a terminal window titled "Terminal" with a menu bar (File, Edit, View, Terminal, Tabs, Help). The terminal displays the output of the SQL command @advisor. The output shows the table name GL\_BALANCES, the compression type as "Compress for Query HIGH", and an estimated compression ratio of 10.17. The procedure completed successfully. A blue starburst graphic is overlaid on the terminal window, pointing to the output text.

# Exadata Smart Flash Cache

## Расширяет ограничения произвольного в/в дисков

300 I/O в секунду



- Компромисс между традиционными дисками и Флэш памятью
  - Диски дешевы, имеют большую ёмкость, но ограничены низким в/в (300 IOPS на диск)
  - Флэш память дорогая, имеет малую ёмкость, но может поддержать тысячи операций в/в в секунду

Десятки тысяч операций в секунду



- Идеальное решение - Exadata Smart Flash Cache
  - Хранение данных на диске из-за стоимости
  - **Прозрачно перемещает “горячие” данные на флэш кэш**
  - Используются **флэш карты вместо флэш дисков**, что исключает ограничения дисковых контроллеров
  - Флэш карты в Exadata
    - Высокая пропускная способность, низкая латентность
    - 4 x 96GB PCI Express Flash Cards на Exadata Server





# Алгоритмы кэширования Exadata



- Эффективное интеллектуальное кэширование часто читаемых данных;
- Автоматически пропускает кэширование для редко читаемых объектов
  - Резервное копирование не кэшируется
  - Вторичные копии пользовательских данных не кэшируются
  - Операции перераспределения данных (rebalancing) ASM не кэшируются
- Пользователи могут определить политики кэширования для конкретных объектов БД.

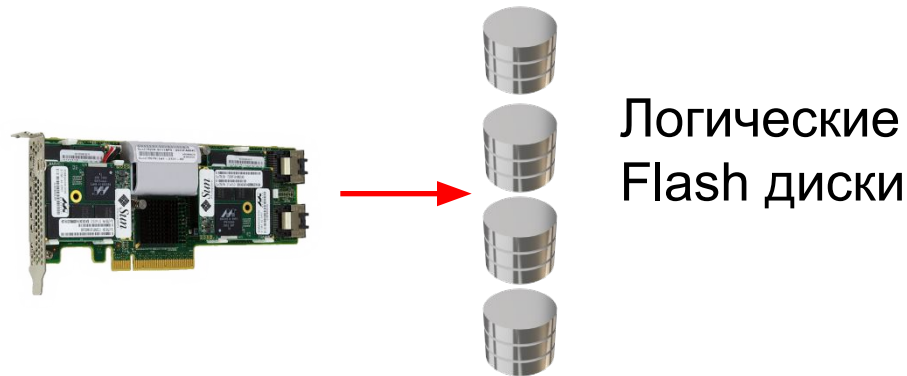
# Smart Flash Cache

Указать, что таблица должна быть закэширована:

```
ALTER TABLE accounts  
  STORAGE (CELL_FLASH_CACHE KEEP)
```



# Exadata логические Flash диски



- Настроить часть или всю Flash память как логический Flash диск
- Дальше он работает как обычный диск
  - Можно создать дисковую группу ASM состоящую из нескольких Flash дисков
  - Данные автоматически зеркалируются ASM между Flash дисками других Exadata Storage серверов
- Высокая производительность для приложений с большим количеством операций записи

# Преимущества



# Сбалансированно и оптимизированно



**Полностью  
оптимизированно**

- Спроектированные системы редко работают с максимальной производительностью:
  - Несбалансированные компоненты, ошибки в конфигурации, “узкие” места
- Exadata спроектирована и оптимизированна от начала до конца:
  - Двухзначные скорости Гбит/сек передачи данных с диска в БД
  - Библиотеки доступа к БД оптимизированны с BIOS, драйверами, ОС, сетевыми протоколами
  - Годы тестирования и отладки
- Удовлетворение требованиям бизнеса за меньшую стоимость

**Прозрачно для существующих приложений – не нужно никаких изменений!**

# Машина БД Exadata

## Консолидация всех существующих приложений



- На Exadata могут совместно выполняться приложения любого типа. Это гарантируется:
  - Широкими каналами и масштабируемой системой ввода/вывода;
  - Instance Caging – ограничение на ресурсы ЦПУ между БД на одном узле;
  - Менеджер ресурсов ввода/вывода;
- Большой объем памяти и процессорные мощности для онлайн задач;
- Оффлоадинг операций (smart scans, storage indexes) для пакетных задач, отчетности, хранилищ;
- **Встроенная компрессия** – существенная экономия дискового пространства для любых приложений.
  - Архивы и данные для отчетности

# Модели Exadata Database



**X2-2 Full Rack**



**X2-2 Half Rack**



**X2-2 Quarter Rack**



**X2-8 Full Rack**

# Масштабируем производительность и объем



## • Масштабируемость

- До 8 стоек в одну систему простым подключением кабелей
  - Больше с использованием внешних InfiniBand коммутаторов
- Масштабируется до сотен серверов хранения
  - Многопетабайтные БД

## • Избыточность и иммунитет к сбоям

- Иммунитет к сбою любого компонента
- Данные зеркалируются между серверами хранения



# Sun Oracle Database Machine

## Экстремальная Производительность для всего



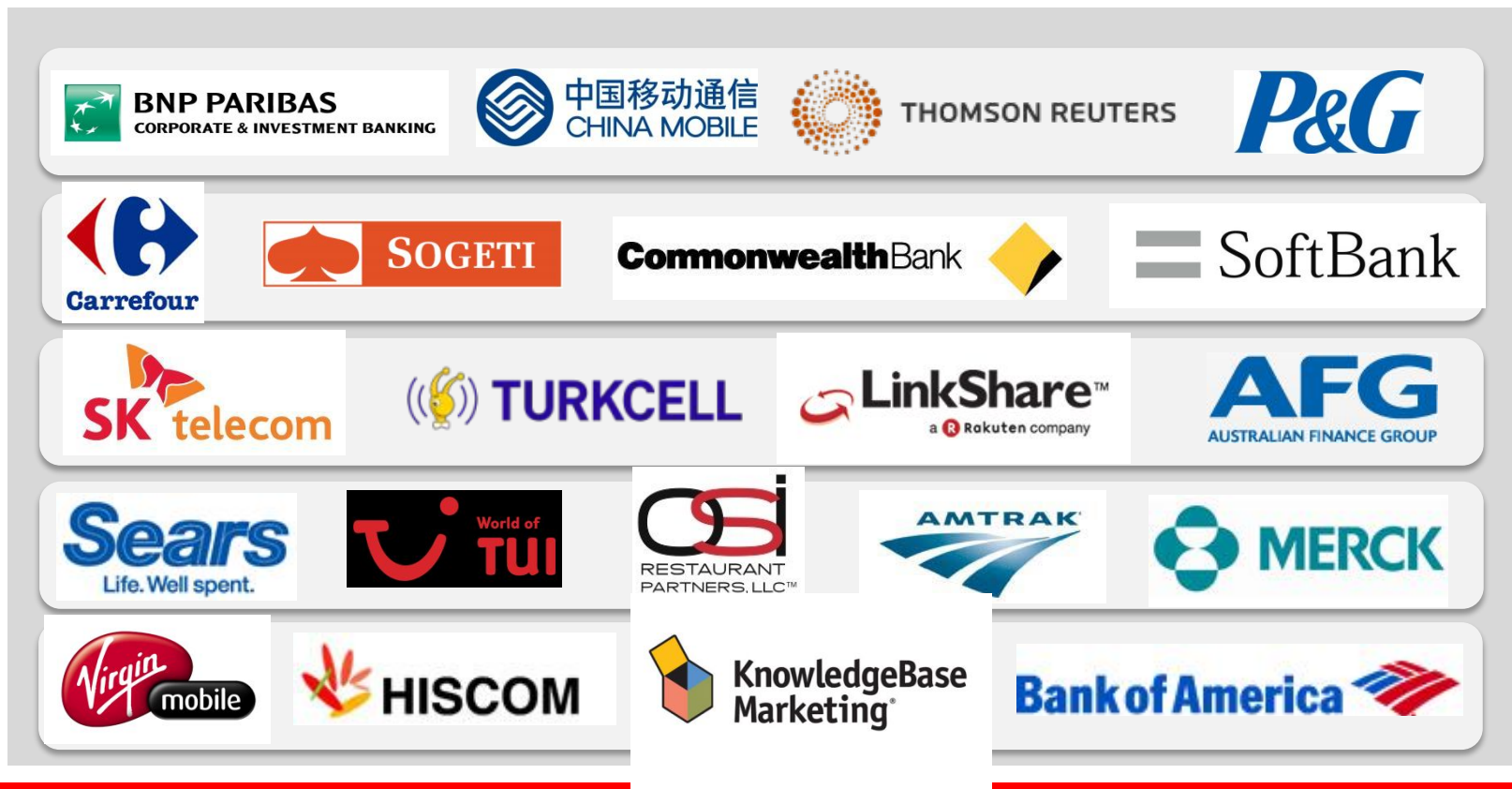
- Для хранилищ данных
  - Параллельные запросы в памяти или в Flash
    - Сжатые 4TB данных в памяти, 50 TB на flash
  - В среднем в 10X-20X быстрее традиционных хранилищ
- Для OLTP-систем
  - Масштабирование реальных приложений в grid - среде
  - Smart flash кэш обеспечивает 1 млн операций ввода/вывода в секунду
    - Сжатые 1.2 TB данных в памяти, 15 TB в Flash
  - Сжатие в 50x для архивных данных
  - Защищенность и отказоустойчивость
- Для консолидации баз данных
  - Поддерживает масштабирование любых типов нагрузки
  - Предсказуемое время отклика в многопользовательском окружении

# Что говорят заказчики



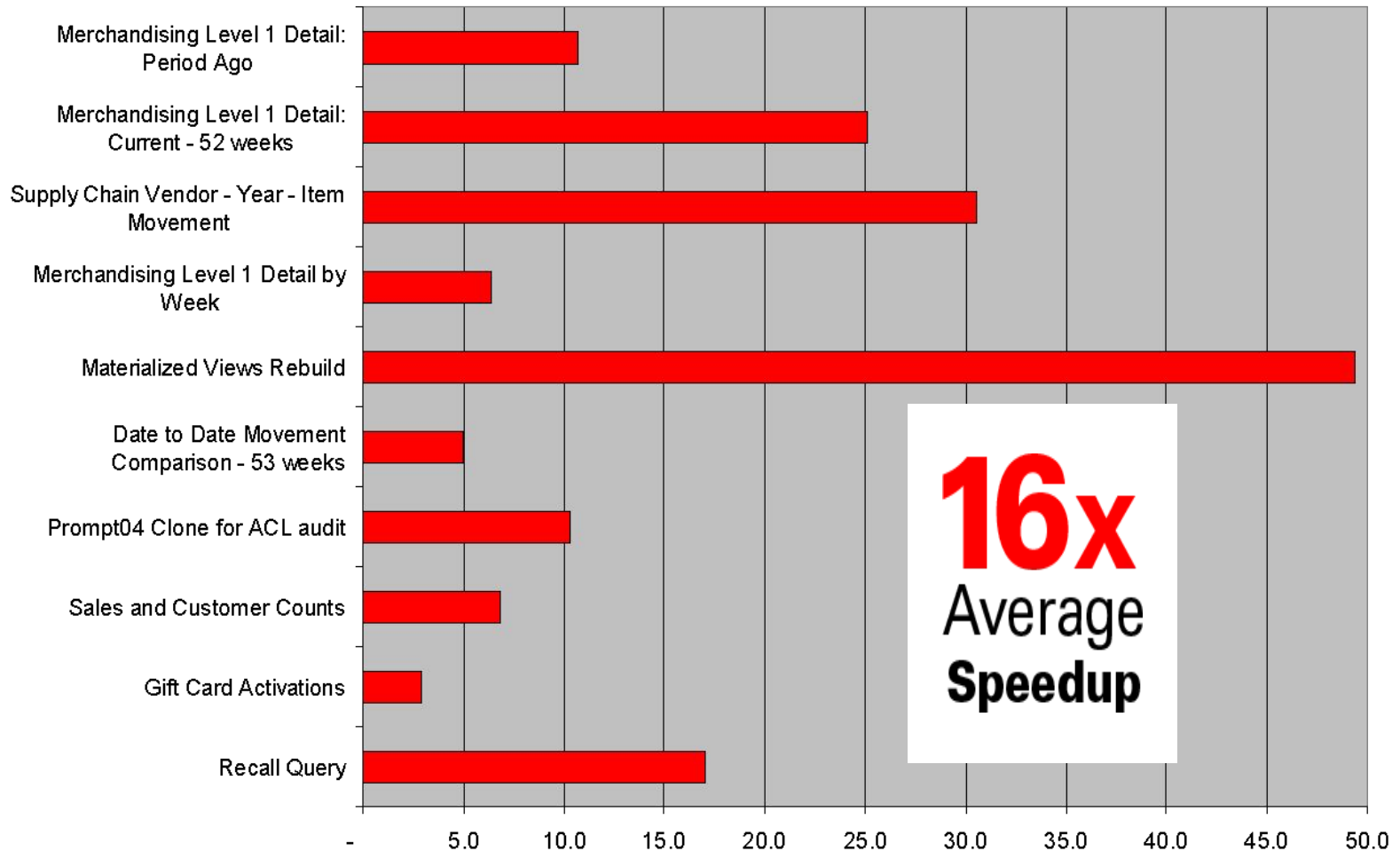
# Exadata на рынке

- Была выпущена в 2008
- Применяется в всех регионах и индустриях



# Giant Eagle ускорение – от 3X до 50X

Ретейлер, хранилище данных по продажам



# Giant Eagle



## Существующая система



13 IBM  
P570 CPUs



EMC CLARiiON and  
DMX Storage Array

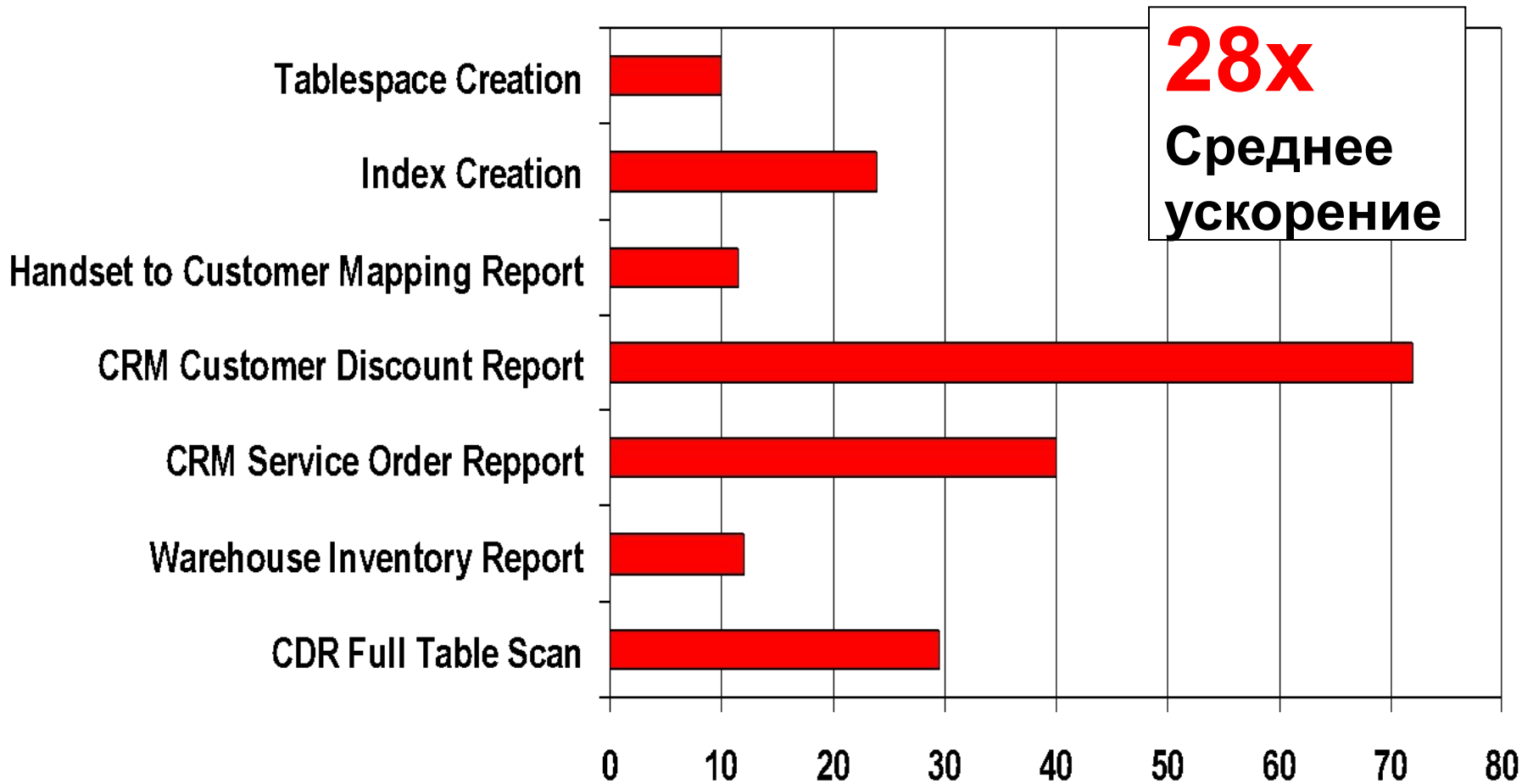
## 1/2 Database Machine



Рост производительности в  
среднем в **16** раз

# М-Тел ускорения операций – 10X to 72X

Мобильный оператор, Болгария



# Оборудование M-Tel



## Существующая система



2 IBM  
P570s



EMC CX3-40  
Storage

## 1/2 Database Machine



# Успех Database Machine



“Запрос, который раньше выполнялся 24 часа, сейчас выдает результат за 30 минут. Oracle Database Machine превзошла конкурентов в **пропускной способности, скорости загрузки, объеме системы хранения и прозрачности.**”

Christian Maar, CIO



“Oracle Database Machine **идеальная эффективная по стоимости платформа для удовлетворения наших потребностей в скорости и масштабируемости.**”

Ketan Parekh, Manager Database Systems



“После тщательного тестирования нескольких платформ для хранилищ данных, мы выбрали Oracle Database Machine. Oracle Exadata способна ускорить наши критичные процессы **с дней до минут..**”

Brian Camp, Sr. VP of Infrastructure Services



# Exadata в 6 раз дешевле

*Самый мощный в IBM - Power 795*

2 Exadata X2-8

**\$3,000,000**



- Больше ядер **CPU**
- Более производительный I/O
- Одинаковый объем дисков
  - Но еще не считая сжатия !
- Отказоустойчивая конфигурация

---

**IBM P795 + 4 DS8700s with Flash**

**\$18,860,000**



**Сравнение цен только на железо**

ORACLE

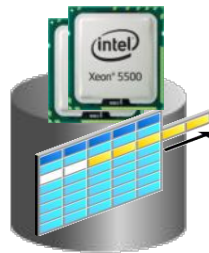
# Oracle Sun Exadata

*Идеальная платформа для баз данных*

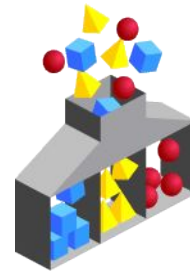
- Для хранилищ данных
- Для оперативных систем (OLTP)
- Для консолидации баз данных



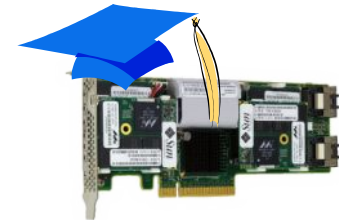
**Быстрее, с меньшей стоимостью**



**Интеллектуальная  
система  
хранения**



**Сжатие по  
столбцам**



**Быстрый  
Flash  
Cache**

# Ресурсы

- Oracle.com:  
<http://www.oracle.com/exadata>
- Портал о технологиях Oracle Exadata:  
<http://www.oracle.com/technology/products/bi/db/exadata>
- Документы об Oracle Exadata:  
<http://www.oracle.com/technology/products/bi/db/exadata/pdf/exadata-technical-whitepaper.pdf>

<http://www.oracle.com/technology/products/bi/db/exadata/pdf/migration-to-exadata-whitepaper.pdf>

# Database Machine Hardware Price

- Price includes only base hardware and 1 year basic warranty
- See Exadata Price List for definitive prices

	List Price
X2-8 Full Rack	\$1,500,000
X2-2 Full Rack	\$1,000,000
X2-2 Half Rack	\$550,000
X2-2 Quarter Rack	\$300,000
X2-2 Half Rack to Full Rack Upgrade	\$525,000
X2-2 Quarter Rack to Half Rack Upgrade	\$300,000
X2-2 Storage Server	\$55,000
X2-2 Expansion Switch Kit	\$23,000

# Exadata Storage Software Licenses

- Customer must purchase (or transfer existing) Exadata Storage software licenses
  - Note that the licenses can be transferred to another machine in the future. For example, 5 years later if the customer replaces their hardware with a new version, they can transfer their existing licenses from the old hardware.
- Licensing metrics based on number of disk drives in storage servers in the Database Machine.
- List Price is \$10,000 / disk drive

	Required Exadata Licenses
X2-8 Full Rack	168
X2-2 Full Rack	168
X2-2 Half Rack	84
X2-2 Quarter Rack	36
X2-2 Half Rack to Full Rack Upgrade	84
X2-2 Quarter Rack to Half Rack Upgrade	48

**ORACLE®**