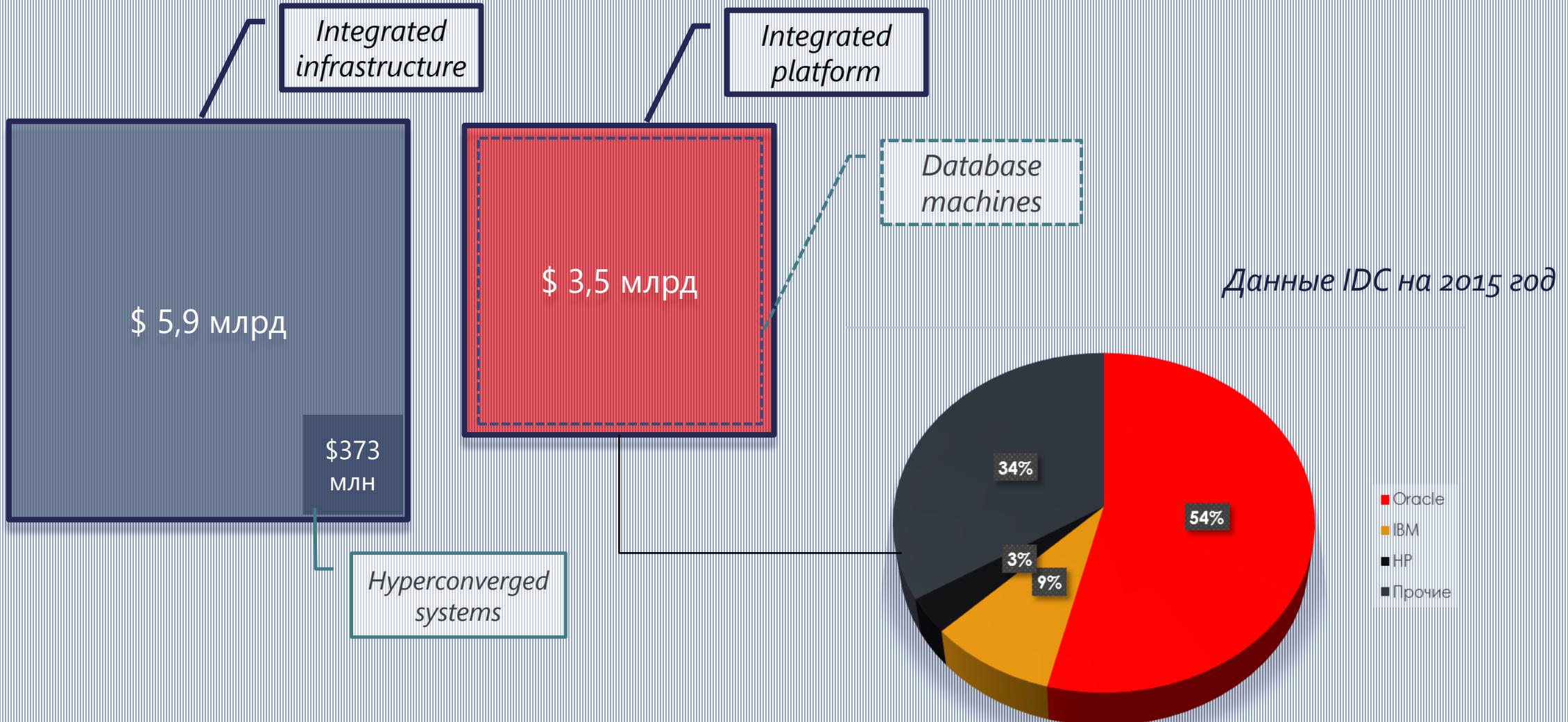


Андрей Николаенко
IBS

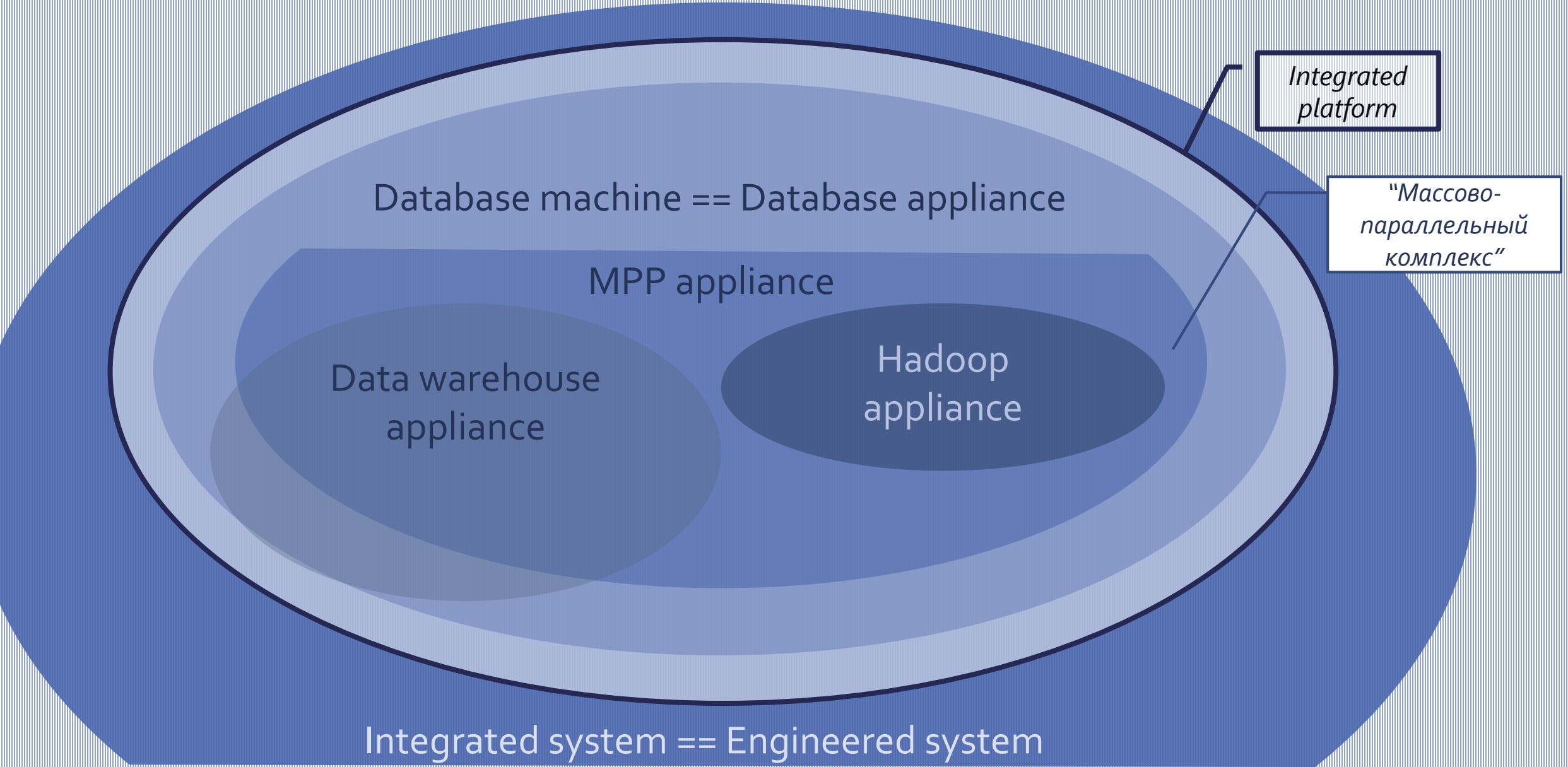
Машины баз данных: концентрированное обозрение

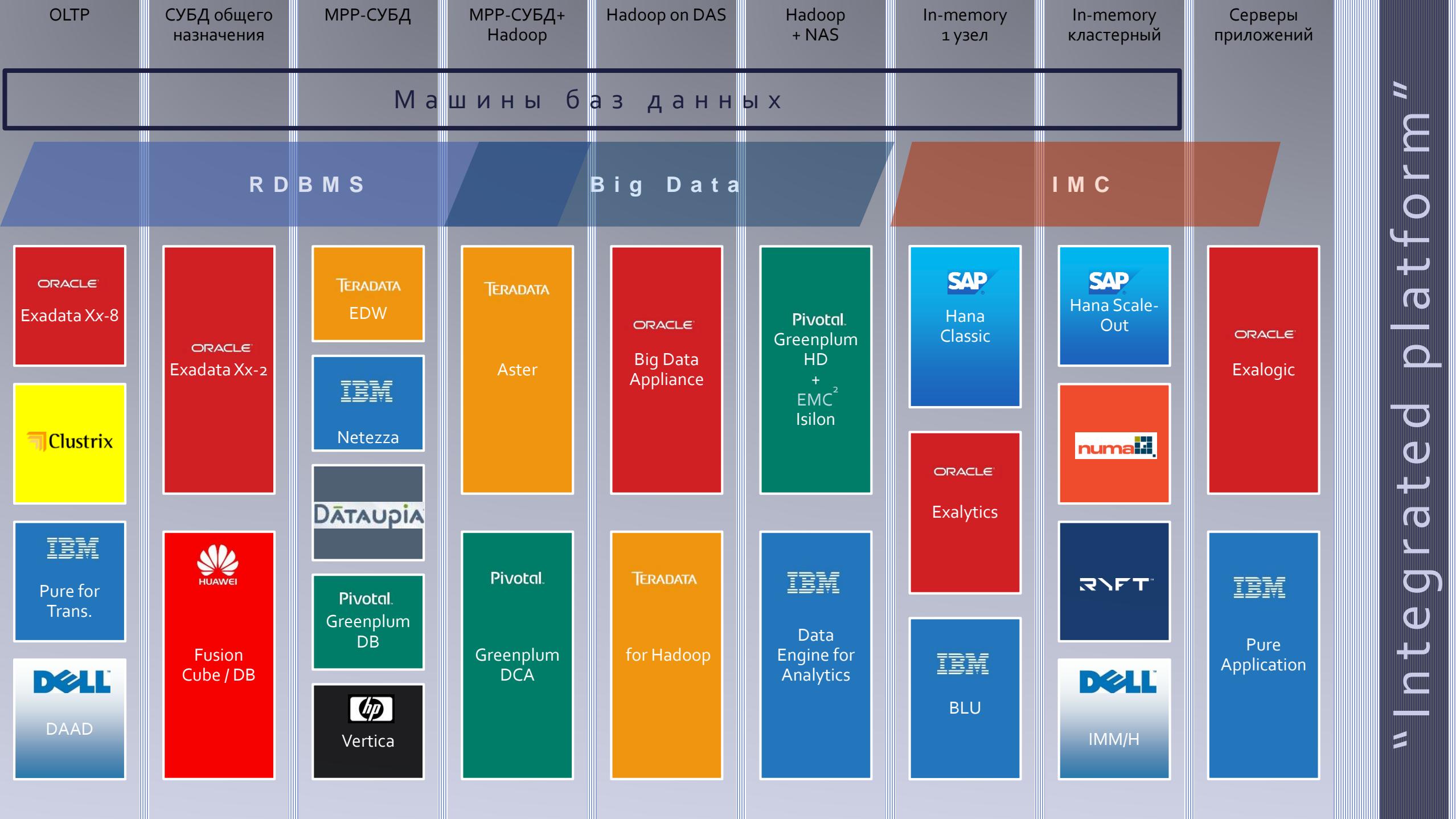
I. ТЕРМИНЫ, ОПРЕДЕЛЕНИЯ, КЛАССИФИКАЦИИ

ИНТЕГРИРОВАННЫЕ СИСТЕМЫ



ТЕРМИНЫ В ВАРИАНТАХ





II. ОТДЕЛЬНЫЕ ПРЕДСТАВИЕЛИ

TERADATA (классическая)

1 на шкаф

1–4 на клику,
до 512 на кластер

2 Infiniband-
коммутатора

2 (max 10) на узел,
120 сессий на РЕ

4–20 на узел
(max 128 на узел)

Клика

Clique

Узел

Node

В Y Net

Banyan Network

РЕ

parsing engine

AMP

access module processor

Дисковая группа

- группа узлов, в рамках которых переподключаемы дисковые группы – область отказоустойчивости

- x86-машина под управлением SuSE
- в каждой клике может быть один узел под горячий резерв

- Межсоединение между узлами

- группа процессов, отвечающих за разбор запросов и раздачу параллельных заданий

- единица параллелизма
- группа процессов, работающая с выделенной дисковой группой

- Подключённая по FC группа накопителей

NETEZZA



Шкаф

• 1...8 на кластер

A[symmetric]MPP

SMP-host

- x86-узел, RHEL
- 1...2 на кластер

Snippet – единица параллелизма

Сниппет-блейд

• x86-узел, RHEL

Сниппет-ПЛИС

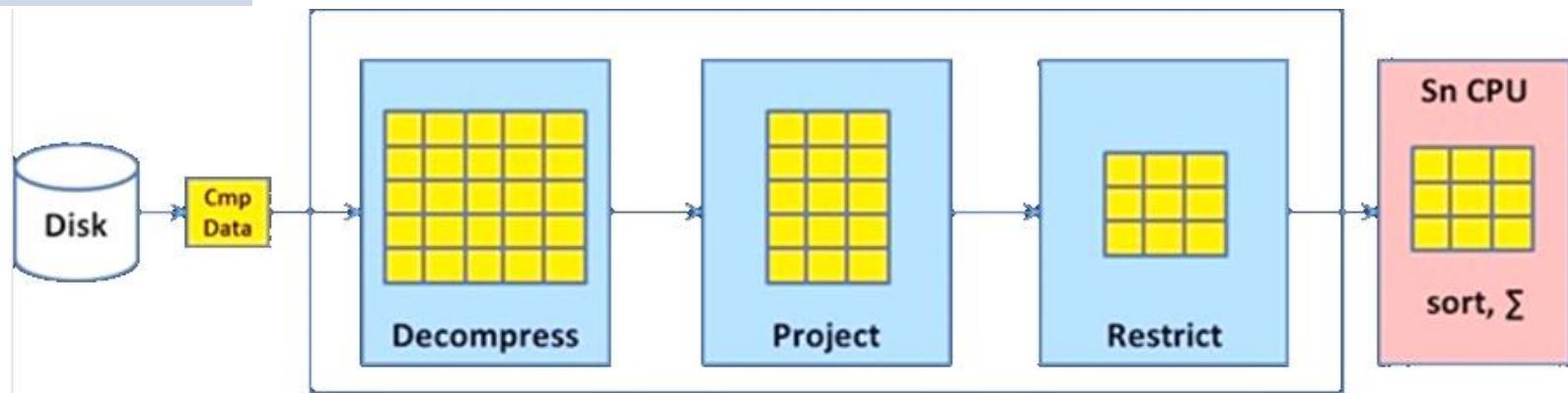
- Xilinx Virtex 6
- 2 на сниппет-блейд
- Обрабатывает весь ввод-вывод к своей дисковой группе

Диски

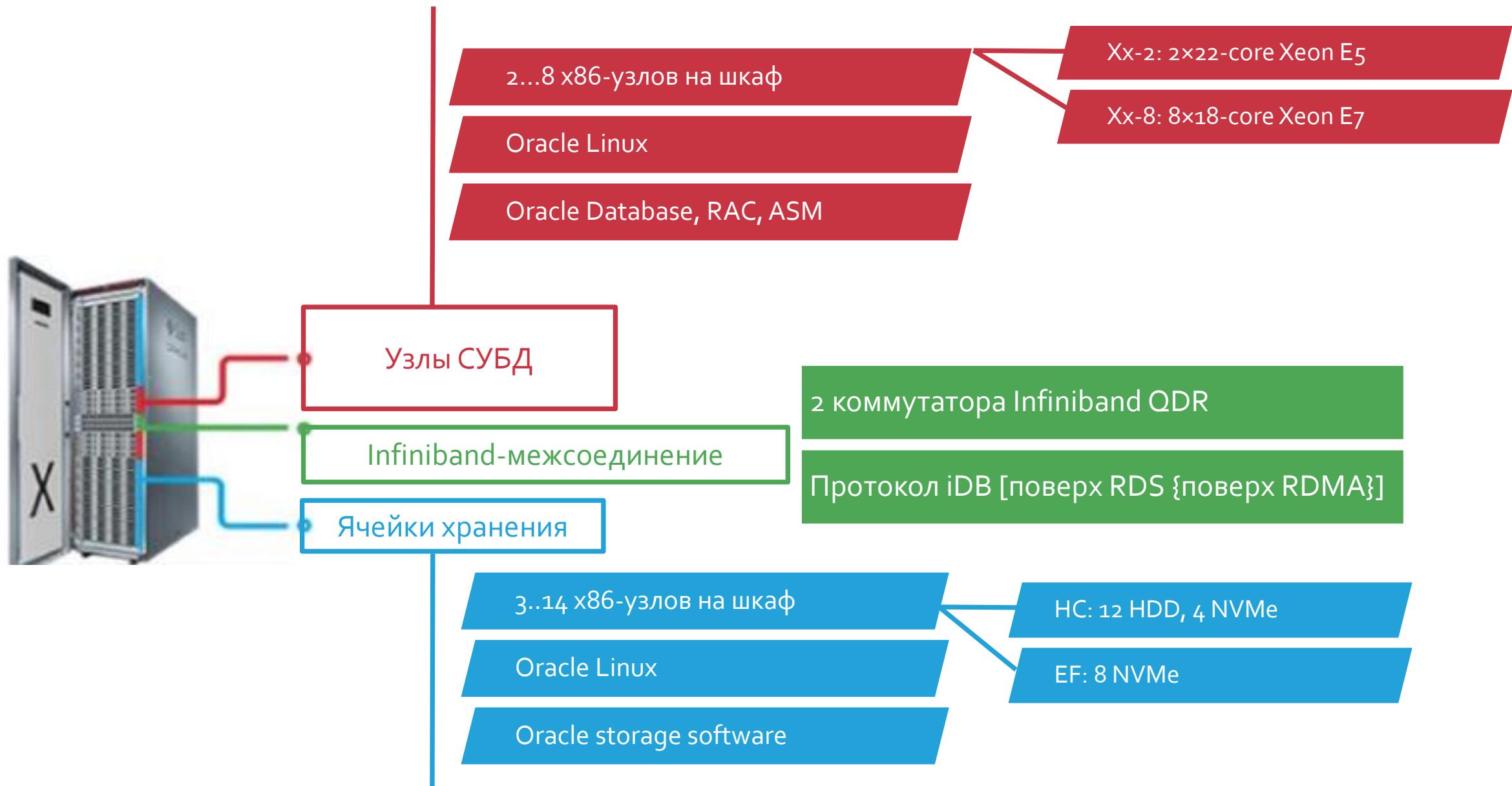
- Сгруппированы в RAID1-группы
- Напрямую подключены через ПЛИС в сниппеты

Операции на ПЛИС:

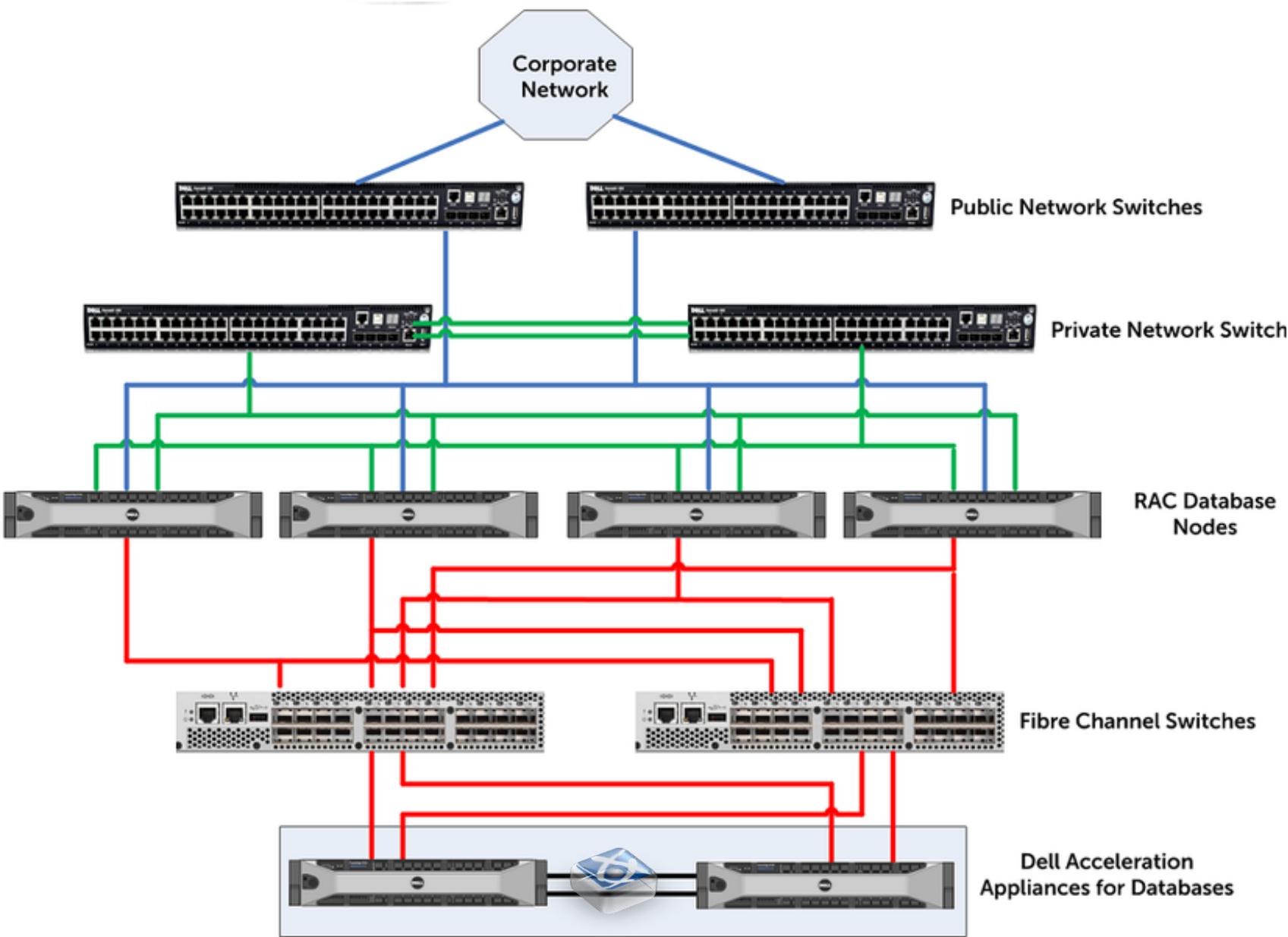
...на ЦПУ:



EXADATA



DELL DAAD



х86-страница для узлов СУБД и узлов хранения

- Dell PowerEdge, 2 × Xeon E5

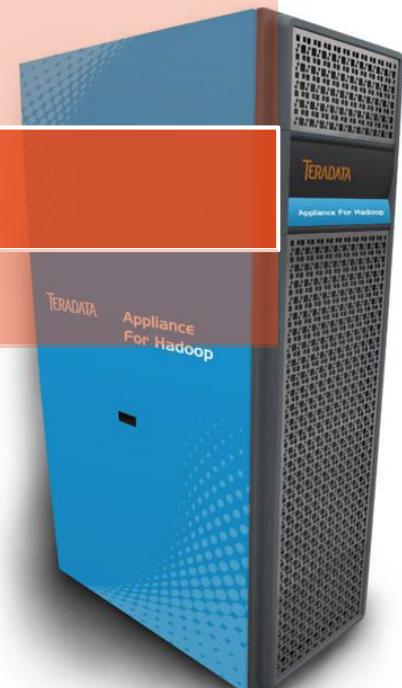
Три вида сетевых технологий

- 10G
- FC в межсети для ASM
- Infiniband для соединения узлов хранения

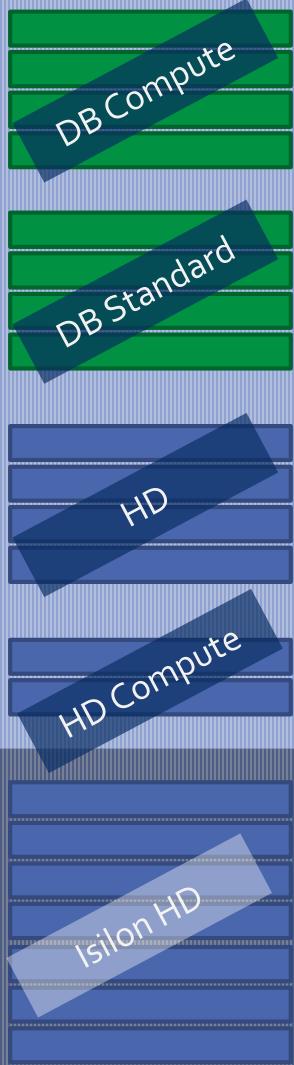
Fusion-io ION на узлах хранения

ORACLE BIG DATA APPLIANCE & TERADATA APPLIANCE FOR HADOOP

18x2xCPU x86	18x2xCPU x86
<ul style="list-style-type: none">12x8TB JBOD в каждом	<ul style="list-style-type: none">12x8TB JBOD в каждом
Infiniband/QDR	BYNet [Infiniband/QDR]
<ul style="list-style-type: none">Межсоединение IPoIBСоединяемость с Exadata	<ul style="list-style-type: none">Межсоединение IPoIBСоединяемость с Teradata
Предустановленный софт	Предустановленный софт
<ul style="list-style-type: none">Cloudera Data Hub EditionOracle Table Access for HadoopOracle NoSQL CEOracle R Enterprise....	<ul style="list-style-type: none">Cloudera CDH или Hortonworks



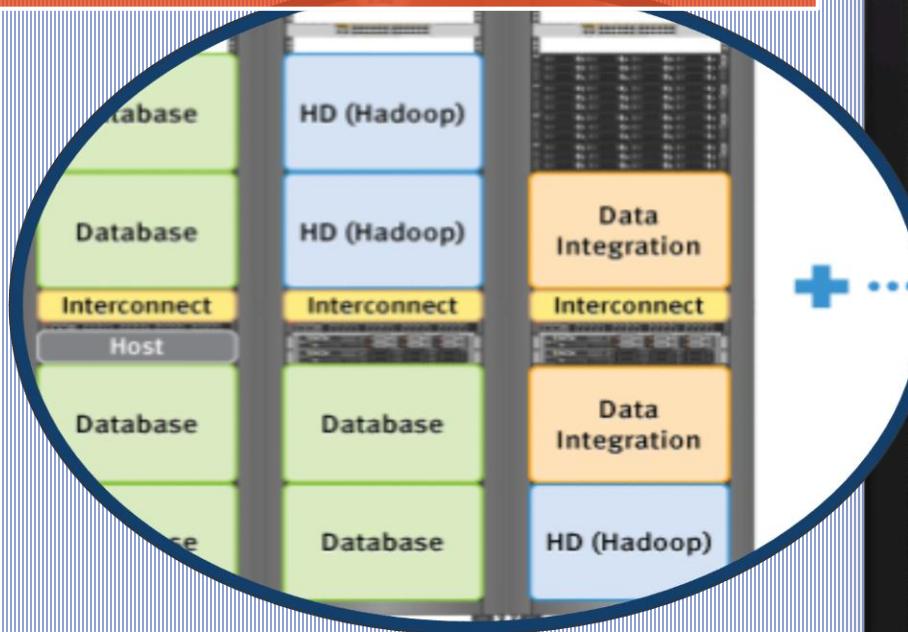
DELL-EMC DCA & ISILON



Greenplum appliance (DB only)

EMC Greenplum Data Computing Appliance (+Hadoop, 2013)

EMC Data Computing Appliance (2015)



Фактически: сертифицированные на стороне SAP некоторые аппаратные платформы

Некоторые базовые условия

Xeon E7

Есть системы на IBM Power

Scale-up

- Один узел
- Для ERP
- Много ОЗУ
- Много ЦПУ

Scale-out

- Много узлов
- Для BW, аналитики...
- С марта 2017 – и для S4/Hana

as of 2017-05-24

Export as Pdf Compare

Narrow your Search

Search Results

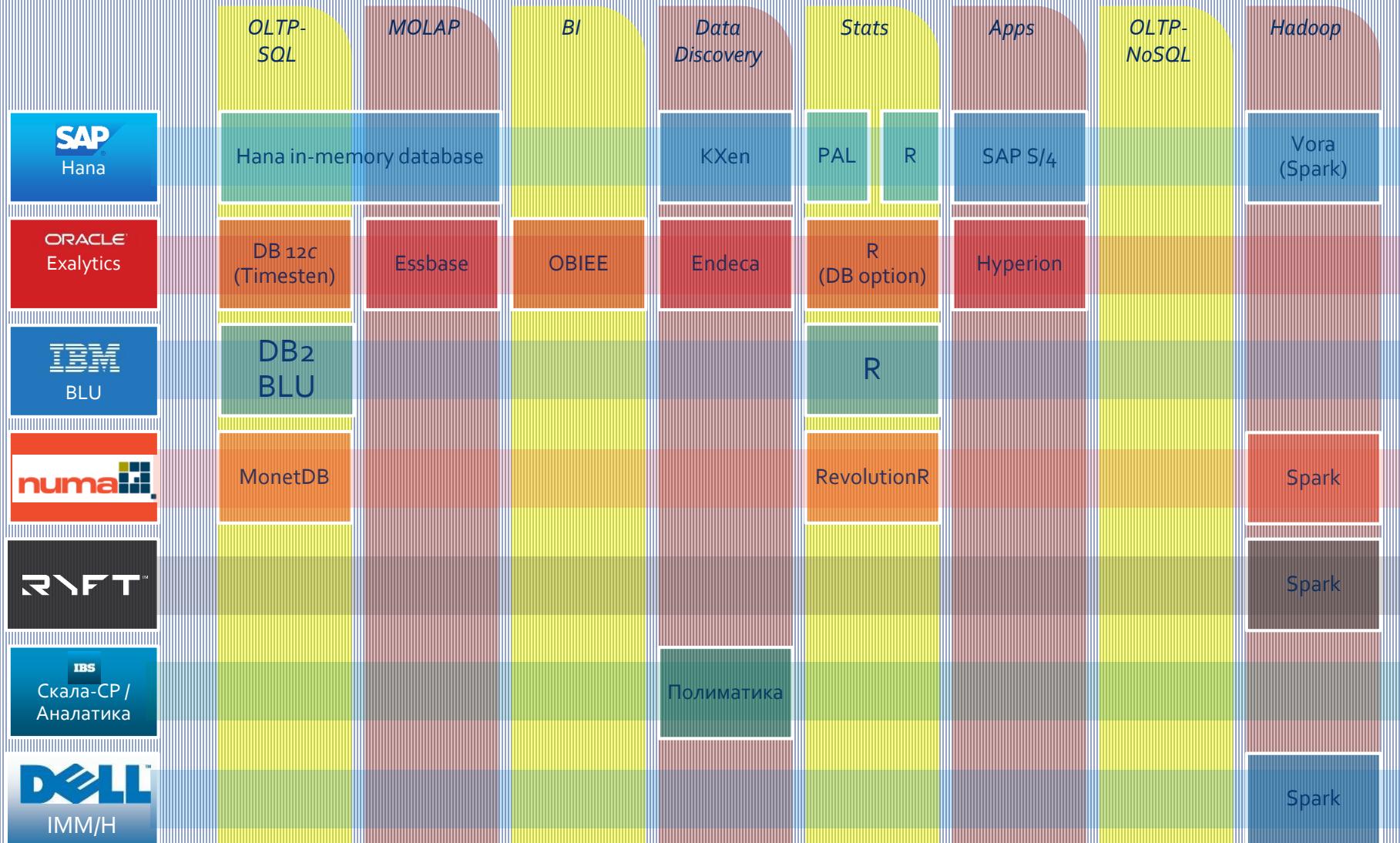
1159 appliance models found

certified

clear selection

Vendor	Model	CPU min.	CPU Architecture	Appliance Type	Memory
Bull SAS	bullion S16	10	Intel Broadwell EX E7	Scale-up: SoH/S4H	7.5 TB ✓
		12			10 TB ✓
		14			9 TB ✓
					12 TB ✓
					10.5 TB ✓
					14 TB ✓

РЕЗИДЕНТНЫЕ ВЫЧИСЛЕНИЯ «В ЖЕЛЕЗЕ»



ШАБЛОНЫ ПОСТРОЕНИЯ



Гиперконвергенция

- Универсальный строительный блок
- Каждый узел выполняет работу по хранению и обработке



Дезагрегация

- Узлы обработки
- Узлы хранения
- Специальный протокол общения

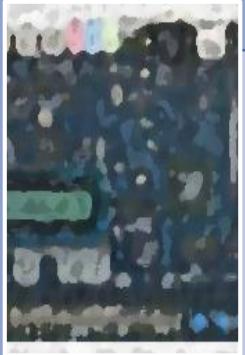
ХАРАКТЕРИСТИЧЕСКИЕ ПРИЗНАКИ



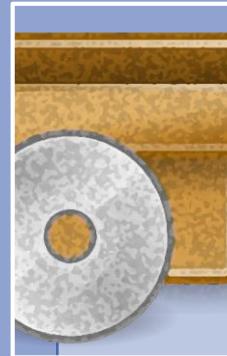
$n \times$ [Commodity x86]
как строительный
блок



Infiniband/RDMA
[часто – просто 10G]



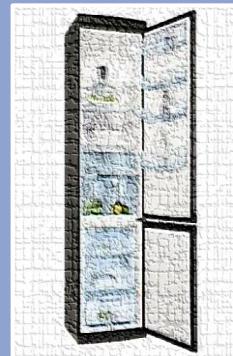
FPGA
[редко]



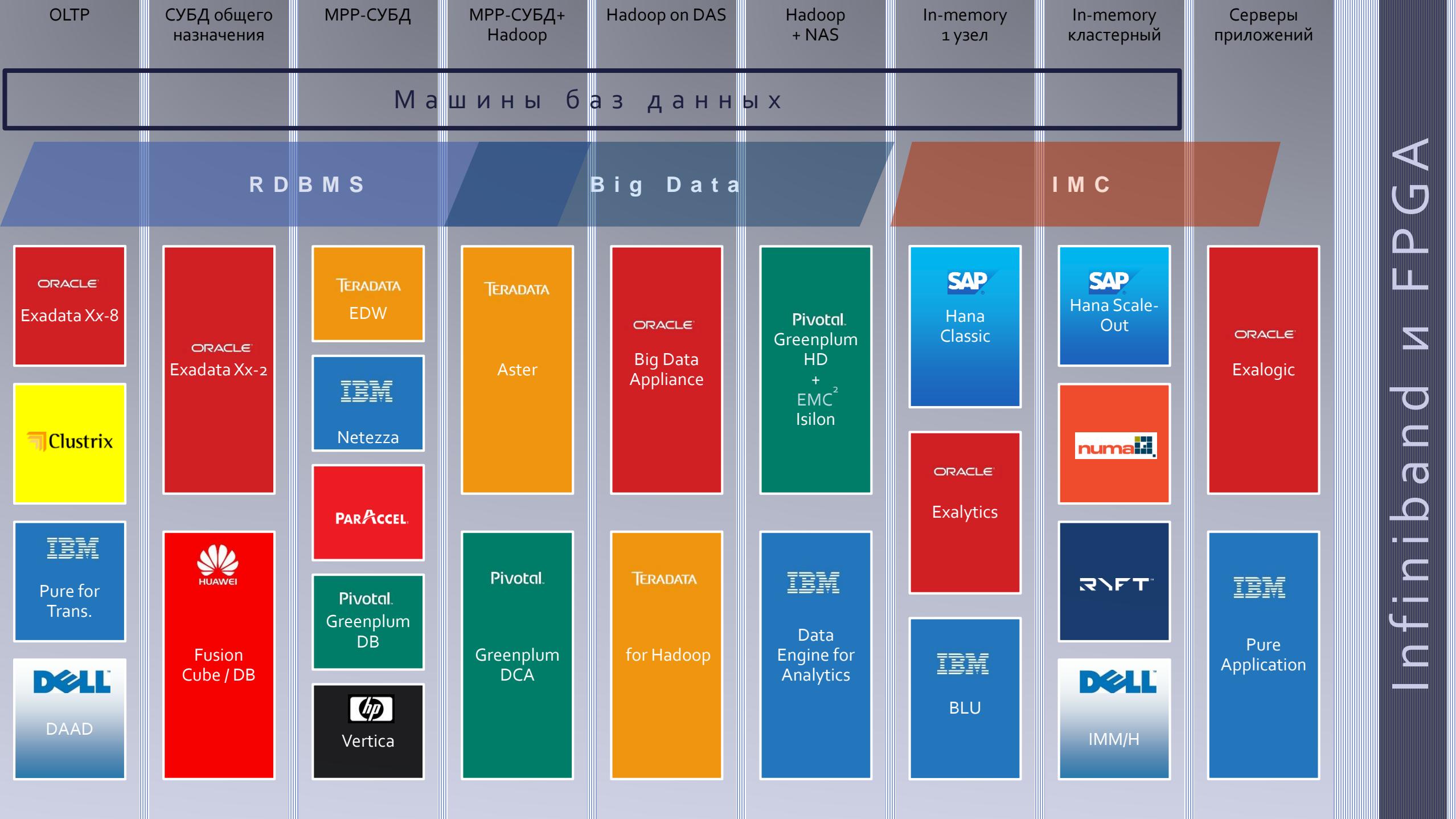
Программные
добавки, програмная
определяемость

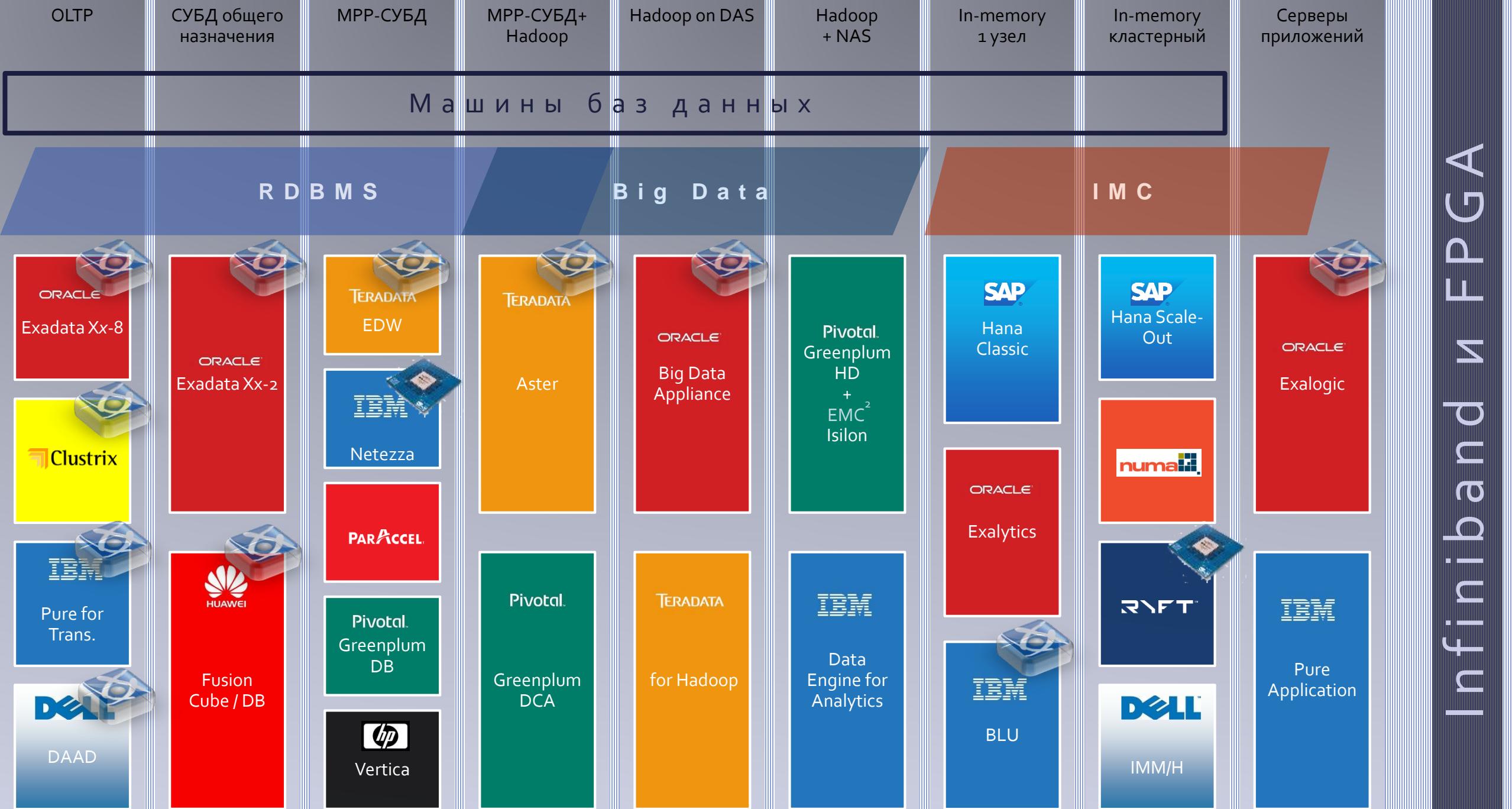


Без единых точек отказа,
дублирование компонентов,
самолечение
[не всегда и не во всём]



Pre-configured,
self-managed,
DBaaS off-the-shelf





АРХИТЕКТУРНЫЙ КОНТИНУУМ



- “Linux x86-64”
- “10GbE”

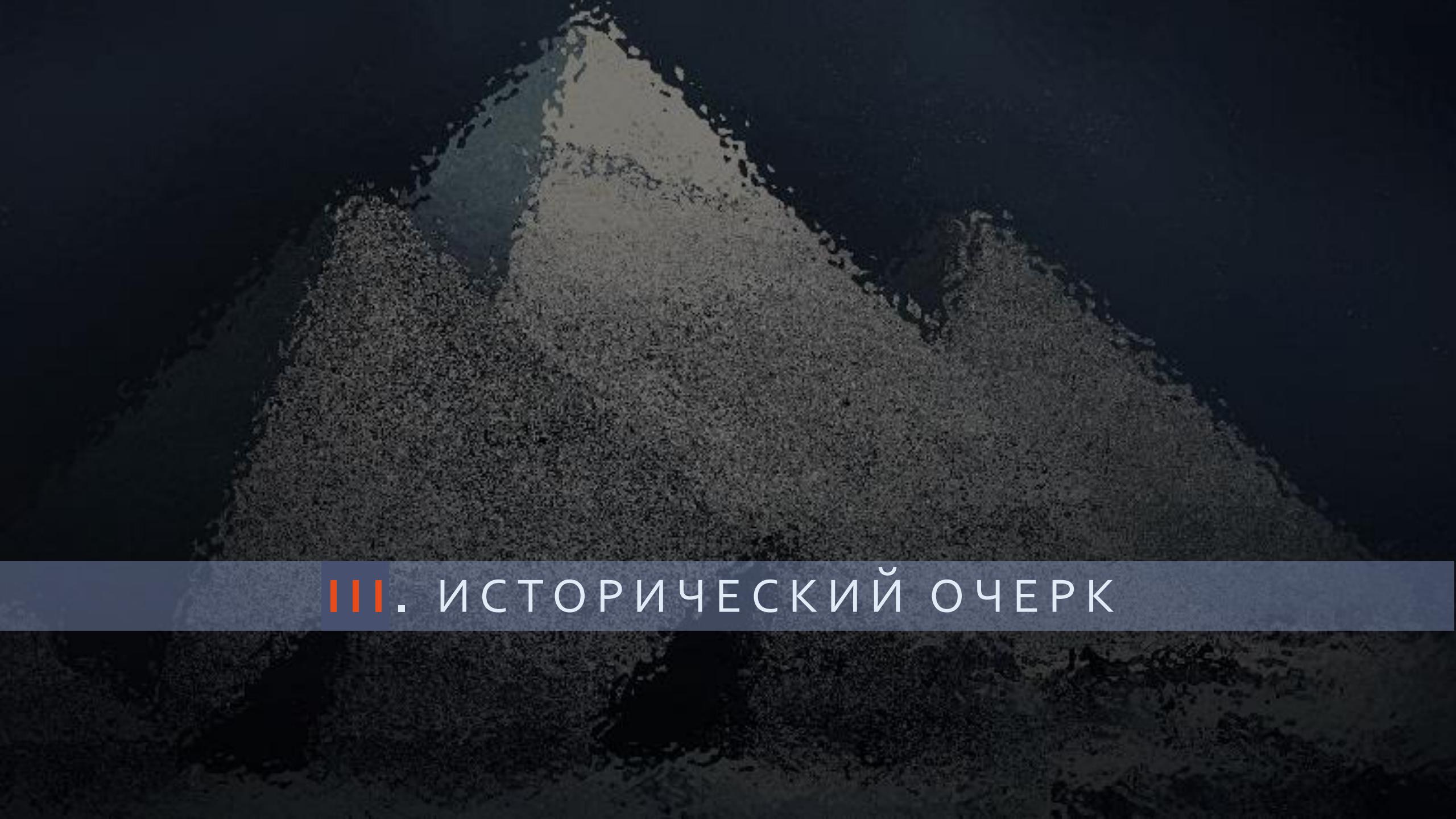
Рекомендуемое оборудование и рекомендованные параметры

- “2×CPU, HW RAID”
- “CentOS 6.5...”
- `kernel.shmmax = xxx`
- `DB_BLOCK_SIZE=4k...`

Референтная архитектура

- “3×PowerEdge...”
- “Cisco Nexus”

Машина баз данных



III. ИСТОРИЧЕСКИЙ ОЧЕРК

ДОИСТОРИЧЕСКАЯ ДИСКУССИЯ

D. Slotnik. Logic per Track Devices // Adv. In Computing, vol. 10 (1970)

David Hsiao. Data Base Machines Are Coming, Data Base Machines Are Coming!
// IEEE Computer, vol. 12 (1979), Issue 3 (March)

Недостатки фон-неймановской архитектуры для СУБД

Альтернативные
процессорные
архитектуры

Ассоциативная память

Ассоциативная
обработка

Разделение
центральных
процессоров и
процессоров ввода-
вывода

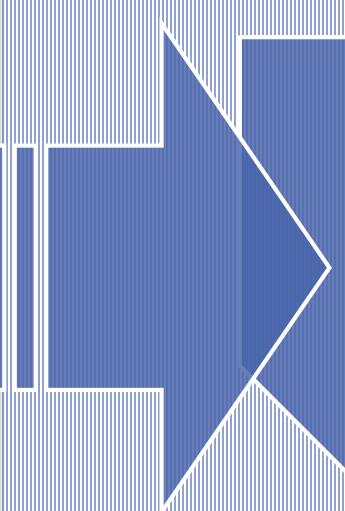
...в одной машине

DATABASE MACHINES: AN IDEA WHOSE TIME HAS PASSED?
A CRITIQUE OF THE FUTURE OF DATABASE MACHINES

by

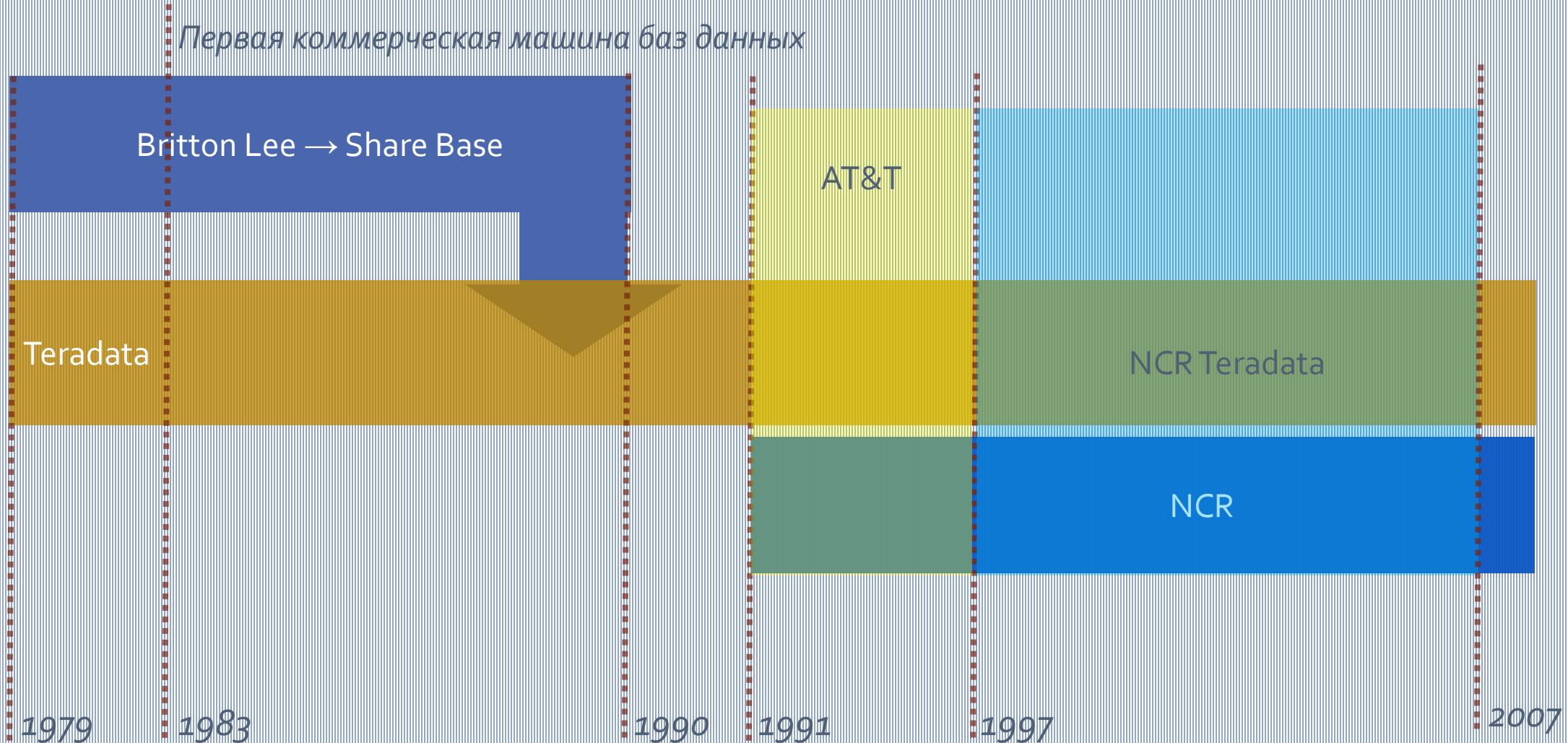
Haran Boral
David J. DeWitt

Computer Sciences Technical Report #504
July 1983



Прогресс в дисковых массивах
и подсистемах ввода-вывода
миникомпьютеров сделали
разработку машины баз данных
экономически нецелесообразной

BRITTON LEE И ЕЁ ПОСЛЕДСТВИЯ



MPP DATABASE, SHARED-NOTHING ARCHITECTURE

Michael Stonebraker. The Case for Shared Nothing // Database Engineering, vol. 9 (1986), No. 1



SM

- shared memory

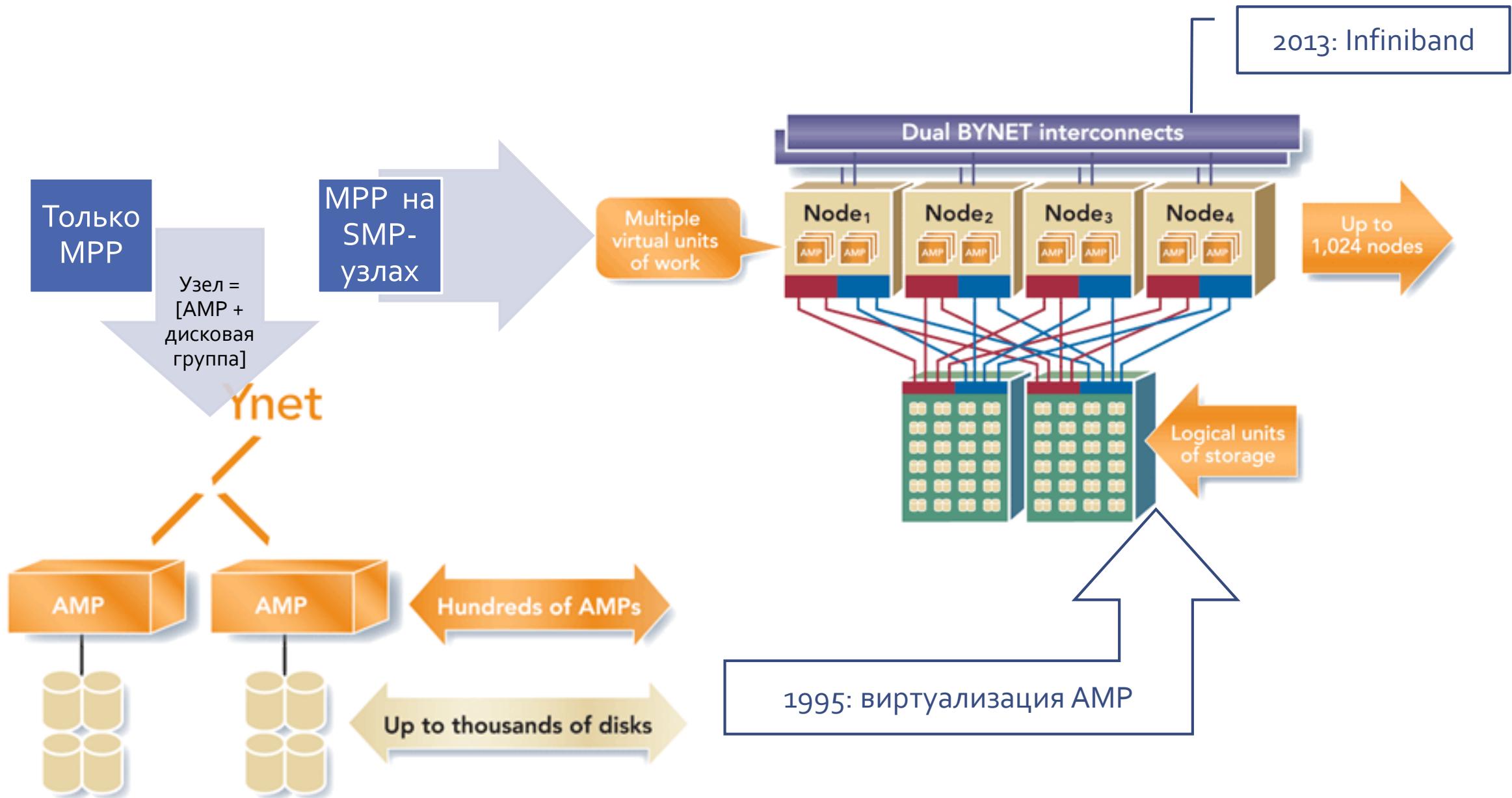
SD

- shared disk

SN

- shared nothing

TERADATA: И ПЕРЕДЕЛ (1995)



2002:

момент идентификации
рынка машин хранилищ
данных



Фостер Хиншоу
сооснователь Netezza,
основатель Dataupia

«ВОЗЬМЁМ-KA POSTGRESQL!»

Netezza

- IBM

Greenplum

- Pivotal / Dell-EMC

Aster Data

- Teradata

Dataupia

- *Seeking for acquisition...*

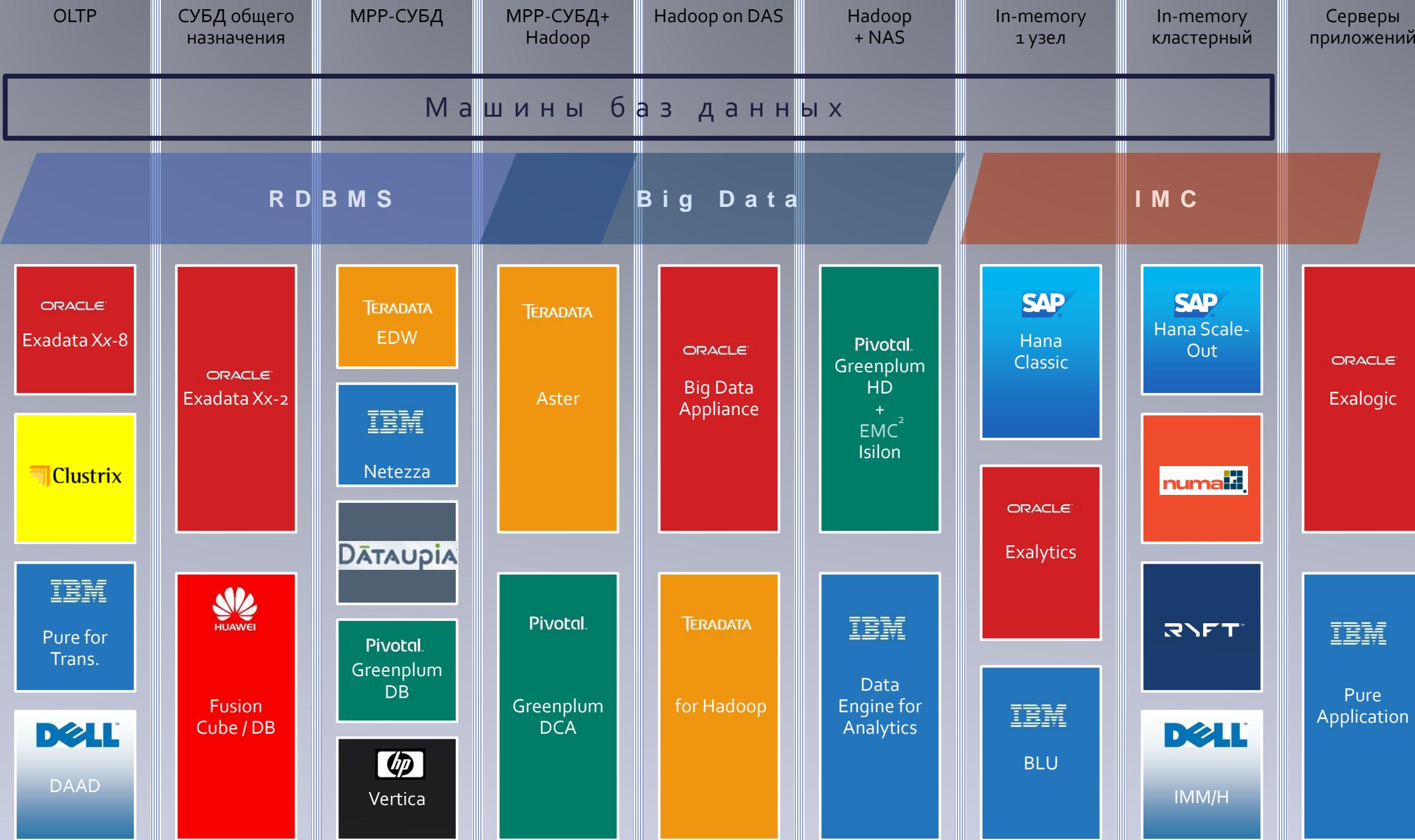
ParAccel

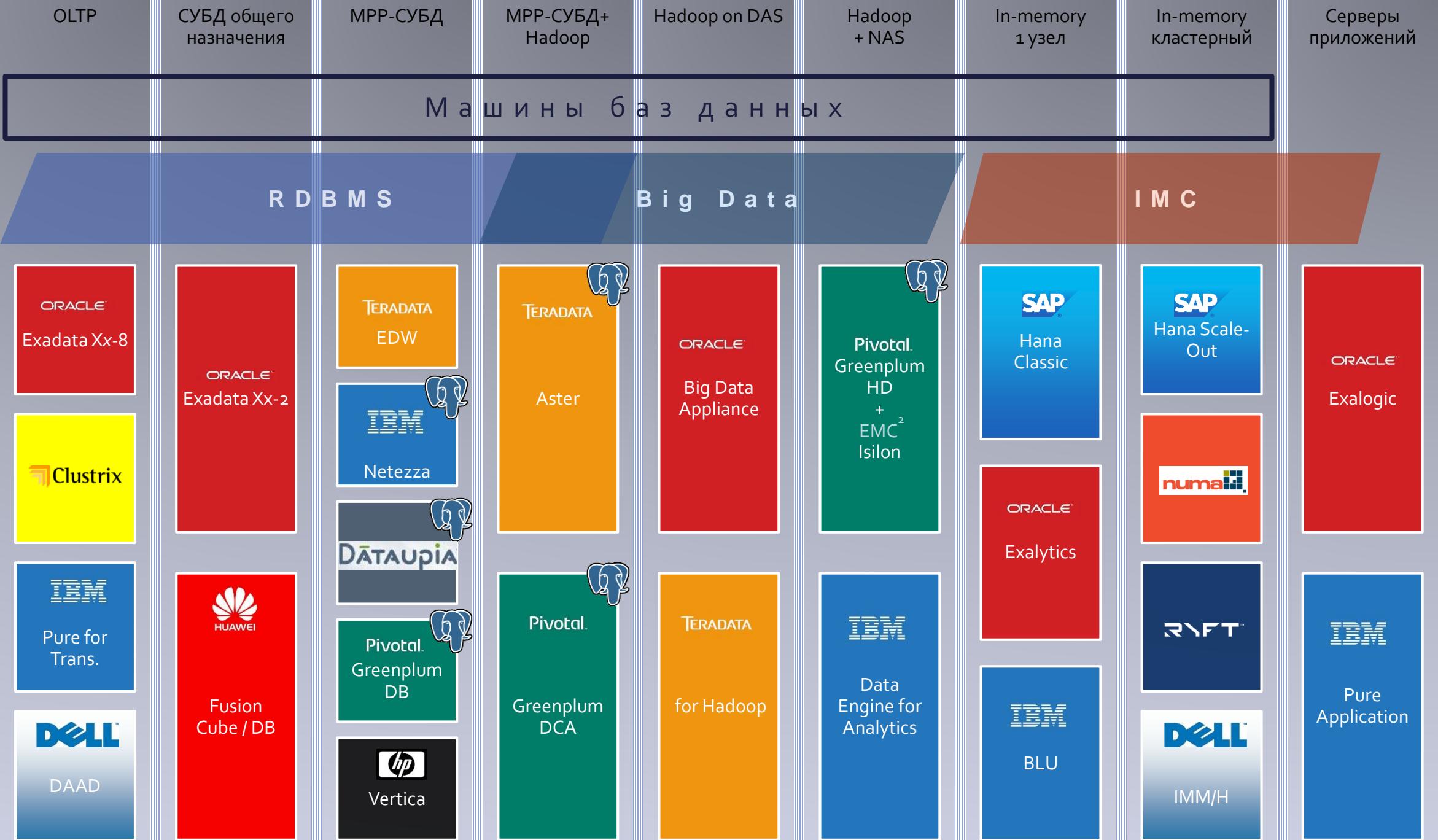
- Actian

Hadapt

- Teradata

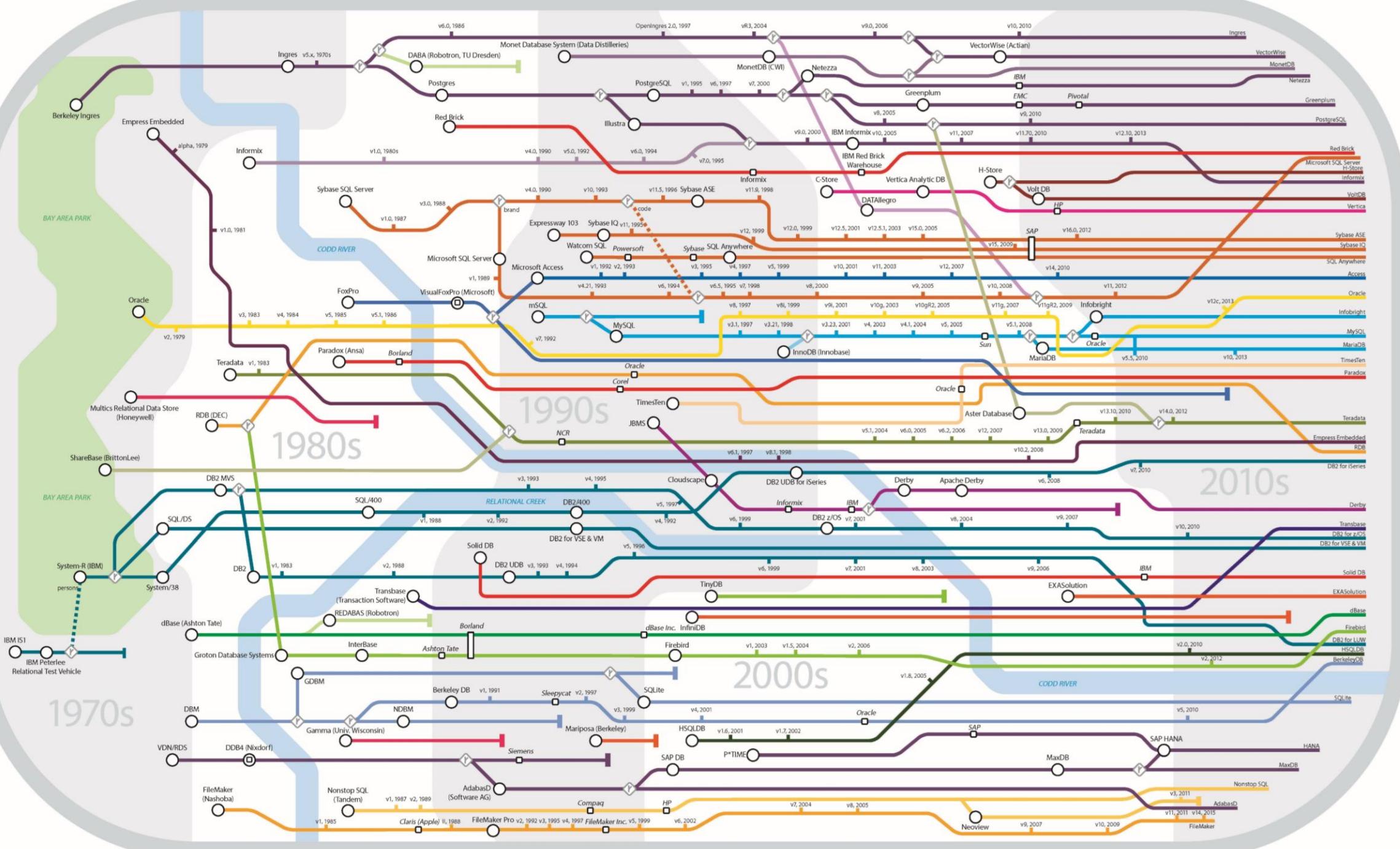
ВЕТВЛЕНИЯ ПОСТГРЕСА



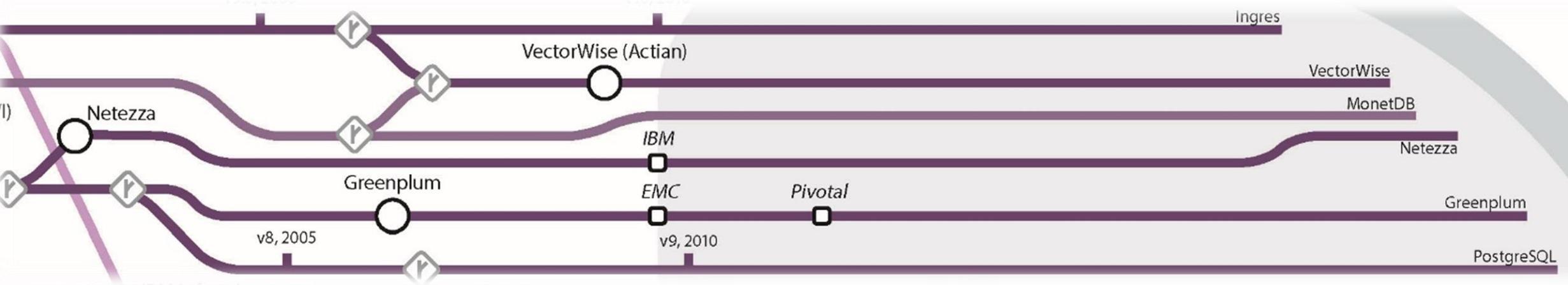


МРР-машины 2000–2005 годов (на пятидесяти миллионном рынке)

	Netezza	Datallegro	Greenplum	ParAccel
Возникновение	2002	2003	2003	2005
Строительный блок	«Сниппет» [PowerPC + FPGA]	Dell PowerEdge [x86]	Sun Fire на AMD [x86]	“Blade cluster” [x86]
Кодовая база	PostgreSQL	Ingres	PostgreSQL	PostgreSQL
Поглощение	2013 IBM	2008 Microsoft	2010 EMC	2011 Actian
Современность	IBM Pure Data for Analytics	Microsoft SQL Server PDW	Dell-EMC-Pivotal Greenplum DB	Amazon RedShift

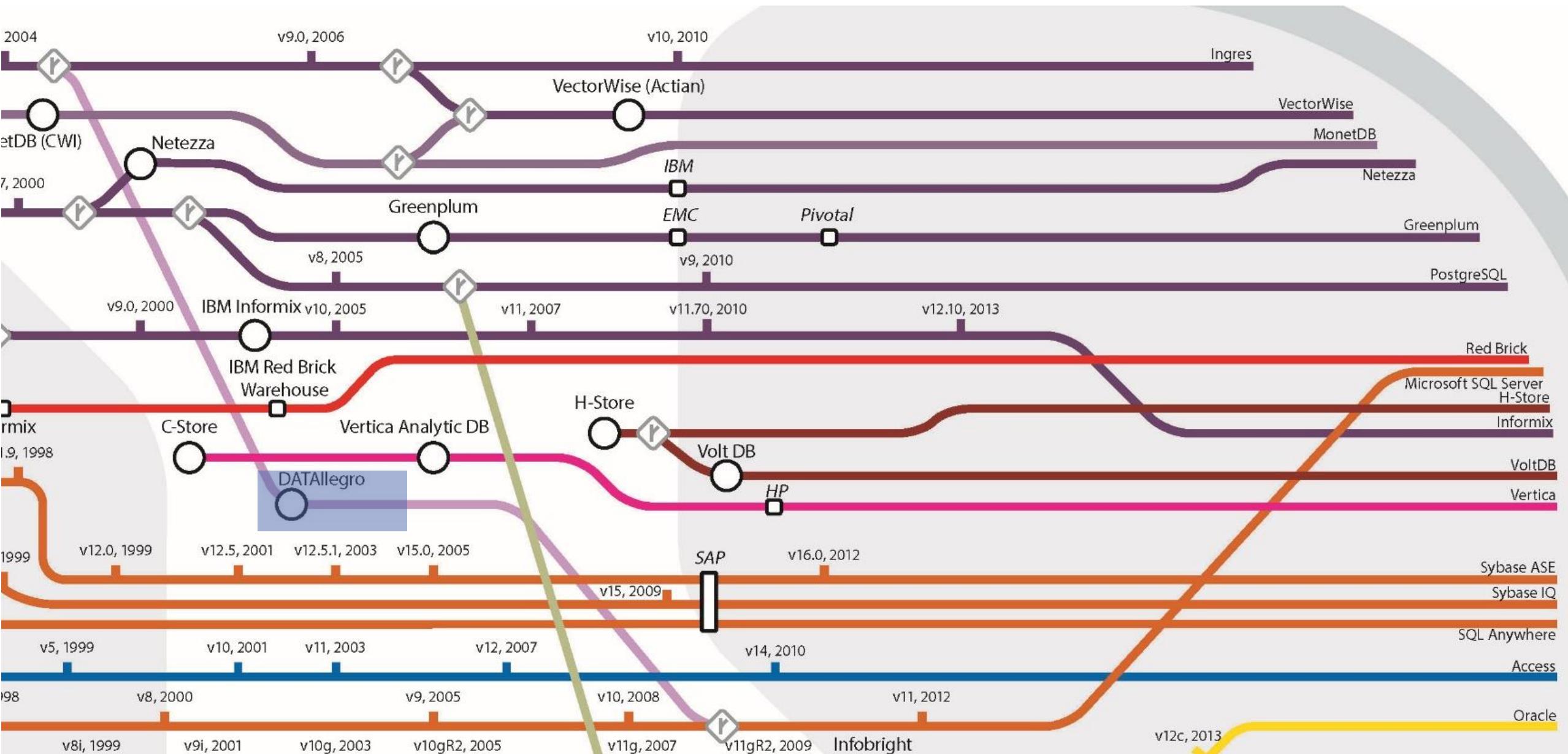


УГОЛ ПОСТГРЕСА (И ИНГРЕСА)



*История машин баз данных
как часть мировой истории СУБД*

DATALLEGRO НА СТЫКЕ СЕМЕЙСТВ



Изображение: Hasso Plattner Institute, CC-BY-SA, 2015

NON-APPLIANCE MPP DBMS

Exasol

Резидентная

Колоночная

Чемпион в
кластерном ТРС-Н

InfiniDB

На основе MySQL

Колоночная

Компания Calpont
обанкрочена в 2014
году

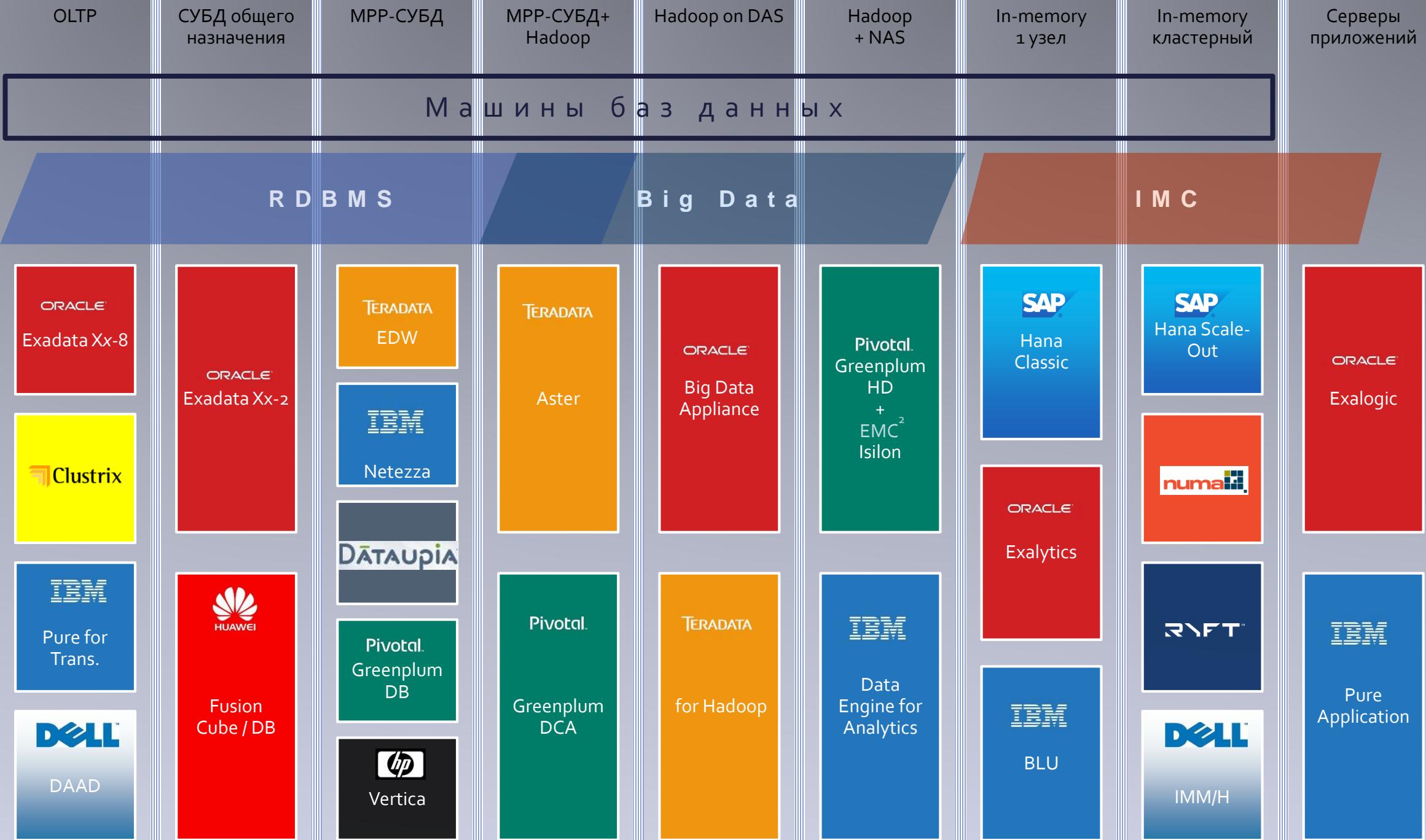
Ряд решений над PostgreSQL

Postgres-XL

Hadapt

Citus DB

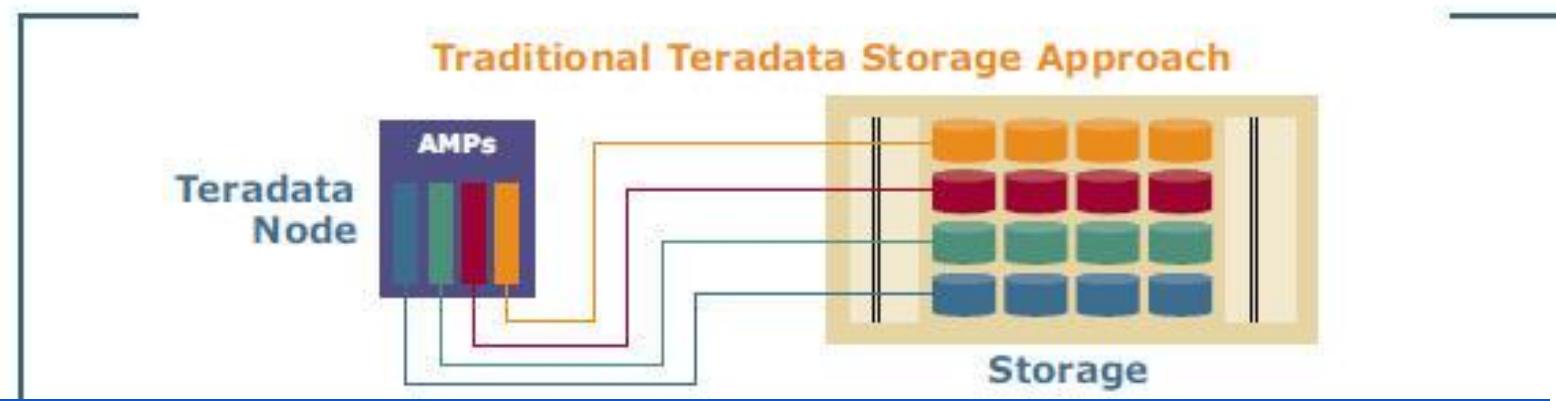
СТАЛИ СОФТОМ



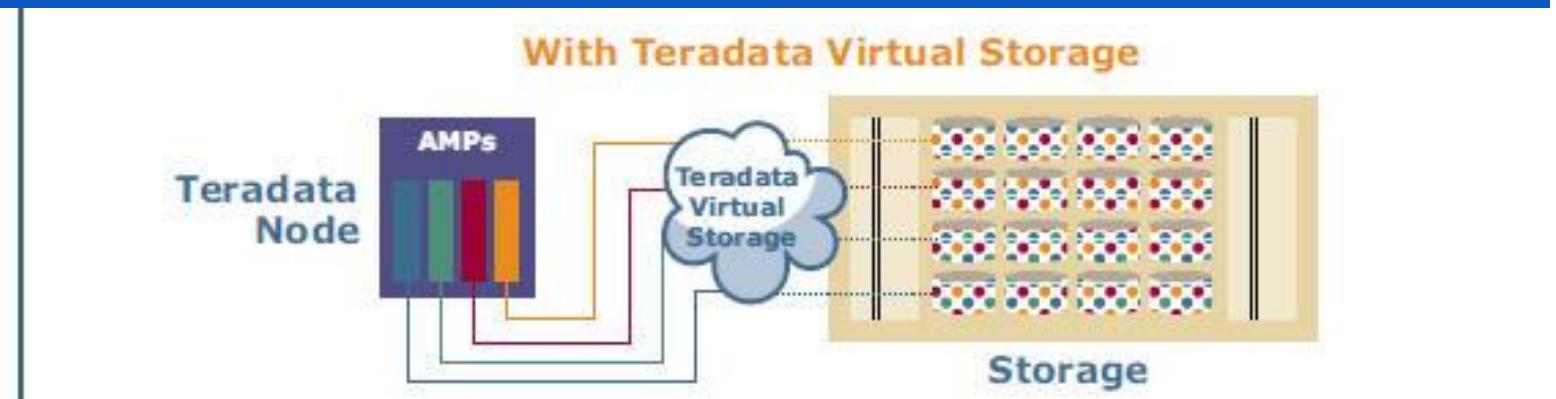
СТАЛИ СОФТОМ



TERADATA: II ПЕРЕДЕЛ (2009)



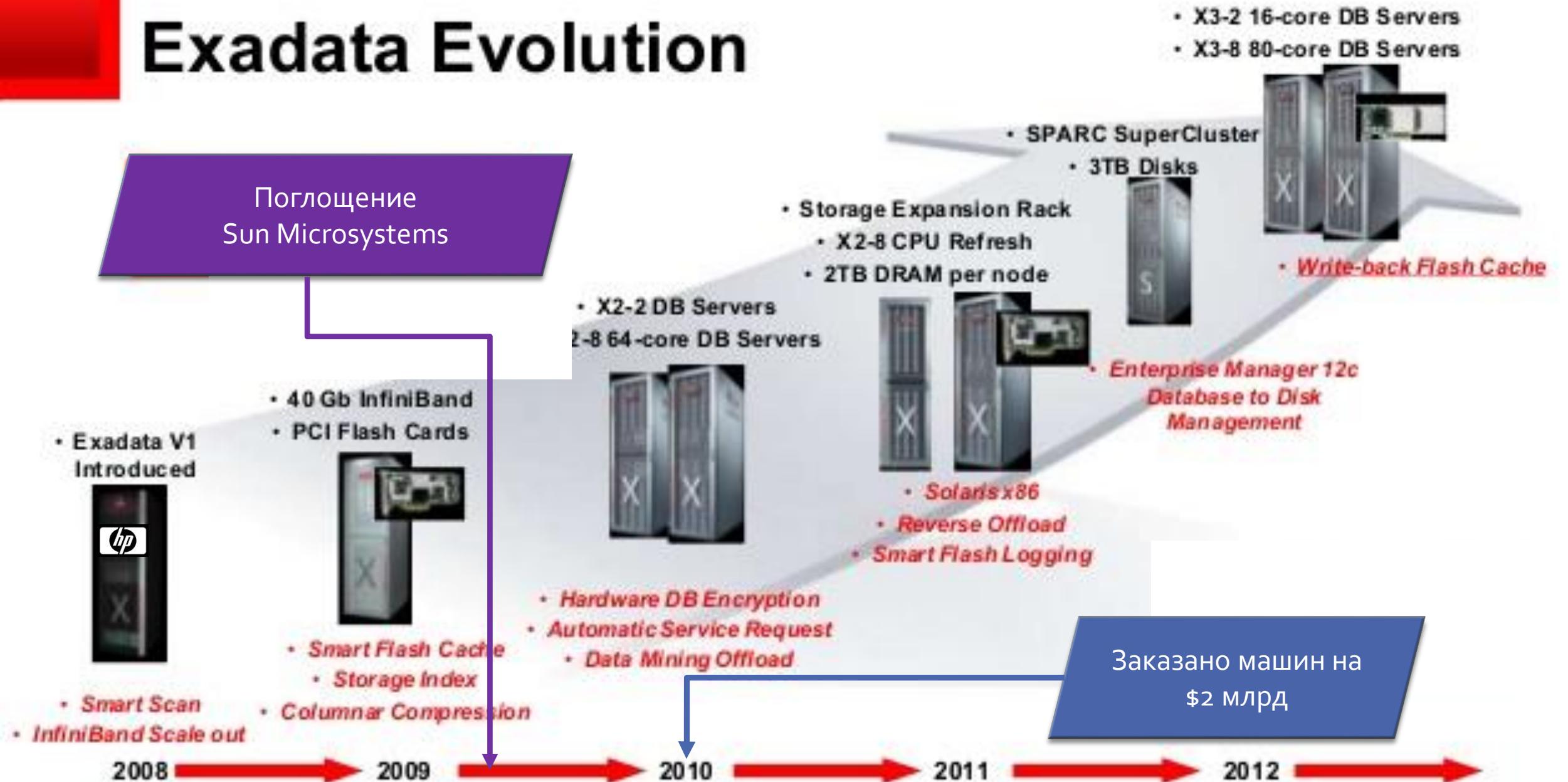
Каждый AMP подключался напрямую к своей дисковой группе [(Drive#,...)];
одинаковое число одинаковых накопителей на AMP



Подключённая к клике подсистема хранения пулирована,
каждому AMP раздаётся в равных долях,
поддерживается многоуровневое хранение и теплокарты

ШАГ В МИЛЛИАРД: EXADATA

Exadata Evolution



ЦЕНЫ И МАСШТАБ В DWH

1990-е

2000-е

2010-е

\$1 млн / ТБ

- [Teradata]

1 ТБ

- 1995, Winter

\$200 тыс.

- [Netezza]

110 ТБ

- 2005, Winter

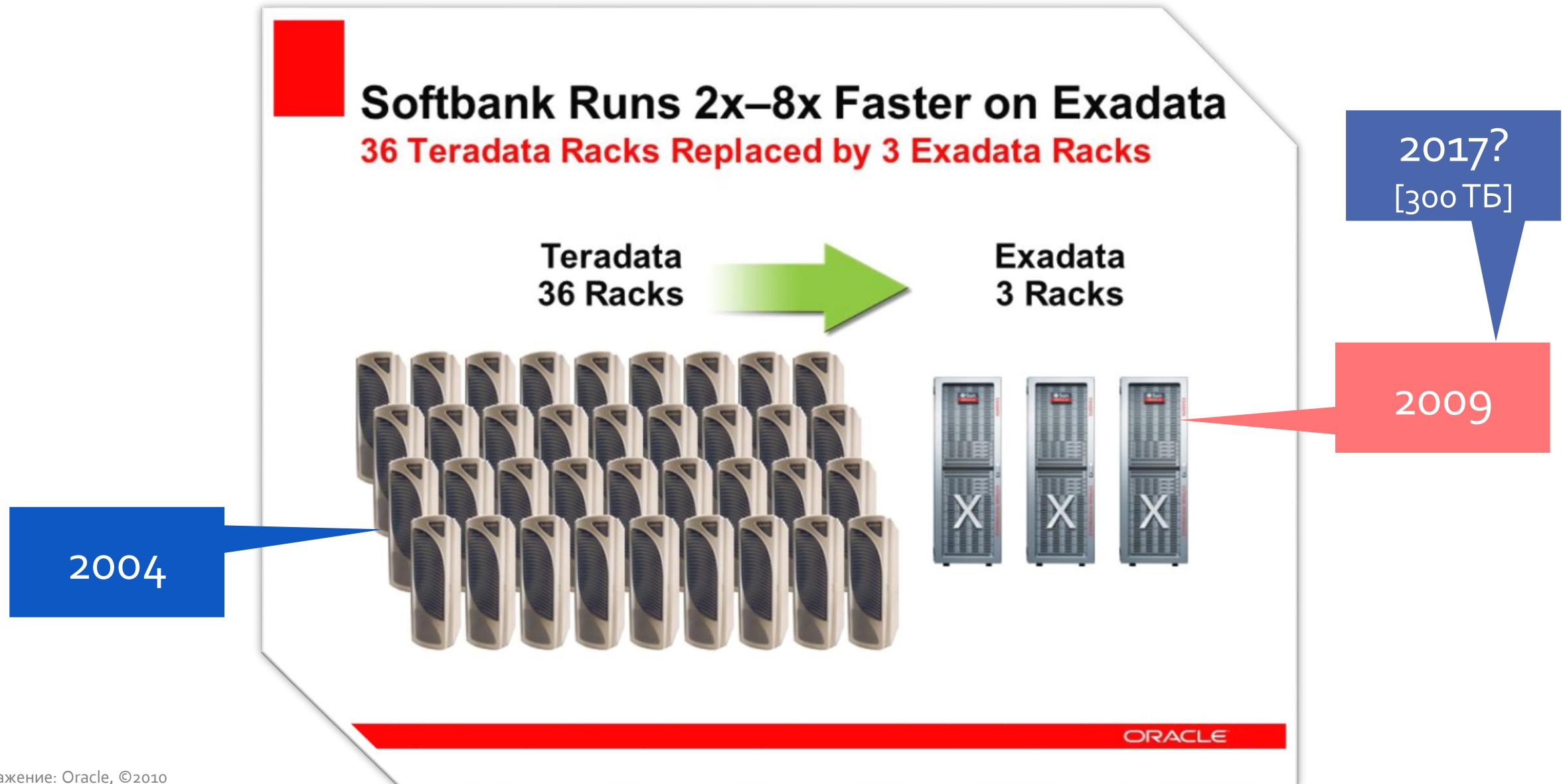
Десятки тыс.

- [при стоимости накопителей в \$200 за ТБ]

12 ПБ

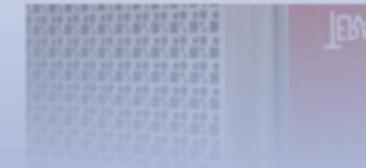
- 2014, SAP

РЕКЛАМНЫЙ ШУМ КАК ЗЕРКАЛО ЭВОЛЮЦИИ

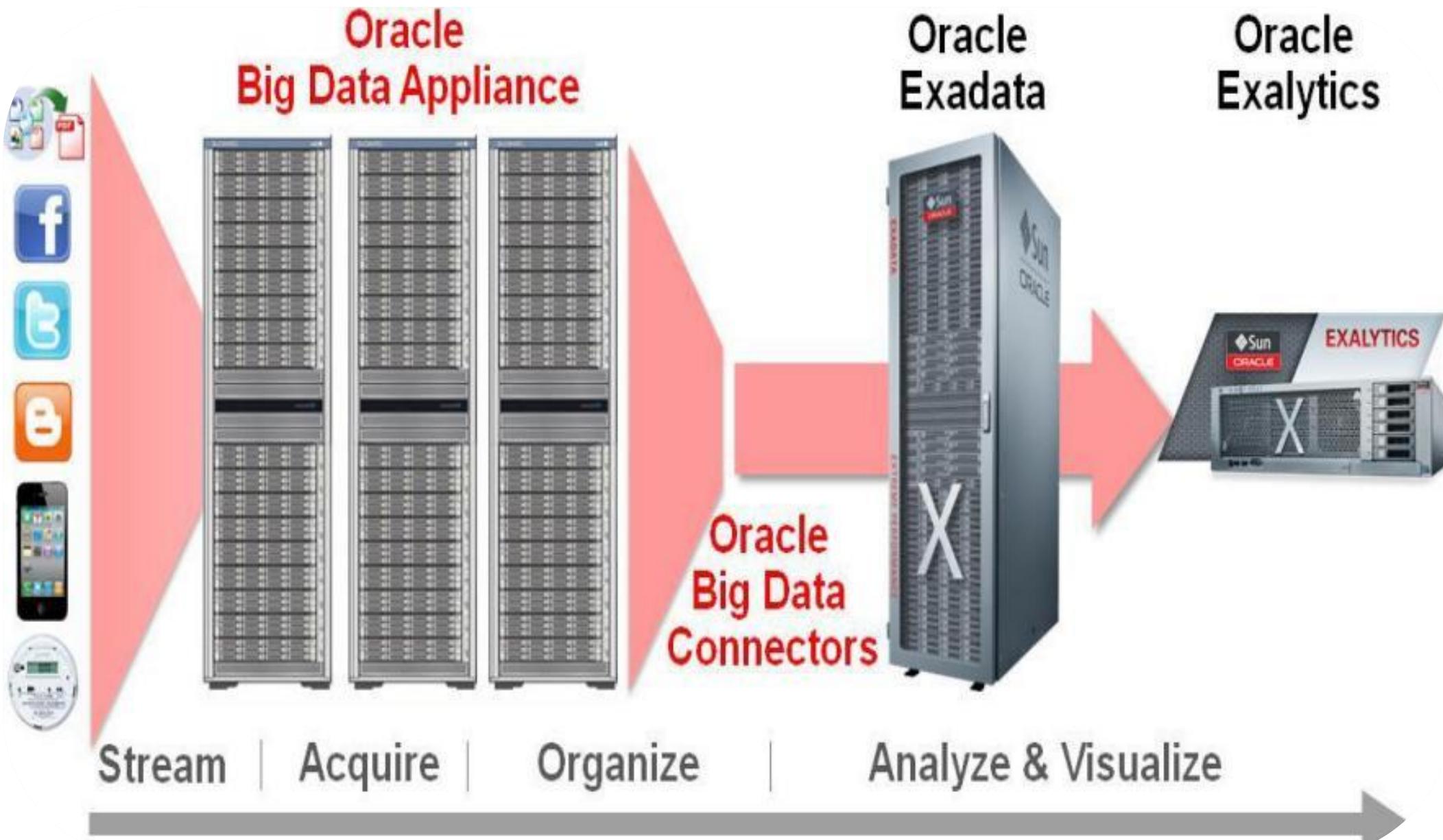


ОТ СУБД К HADOOP: АСПЕКТ ЦЕНЫ

Integrated 1. Big Data Platform	Data Warehouse Appliance	Active Enterprise Data Warehouse	Aster Big Analytics Appliance	Appliance for Hadoop
\$3.8K/TB	\$34K/TB	\$34K-\$69K/TB	\$8K/TB	\$2K/TB



УКЛАДКА В ЛИНЕЙКУ ОТ ORACLE



МУЛЬТИАРЕНДНОСТЬ В СУБД

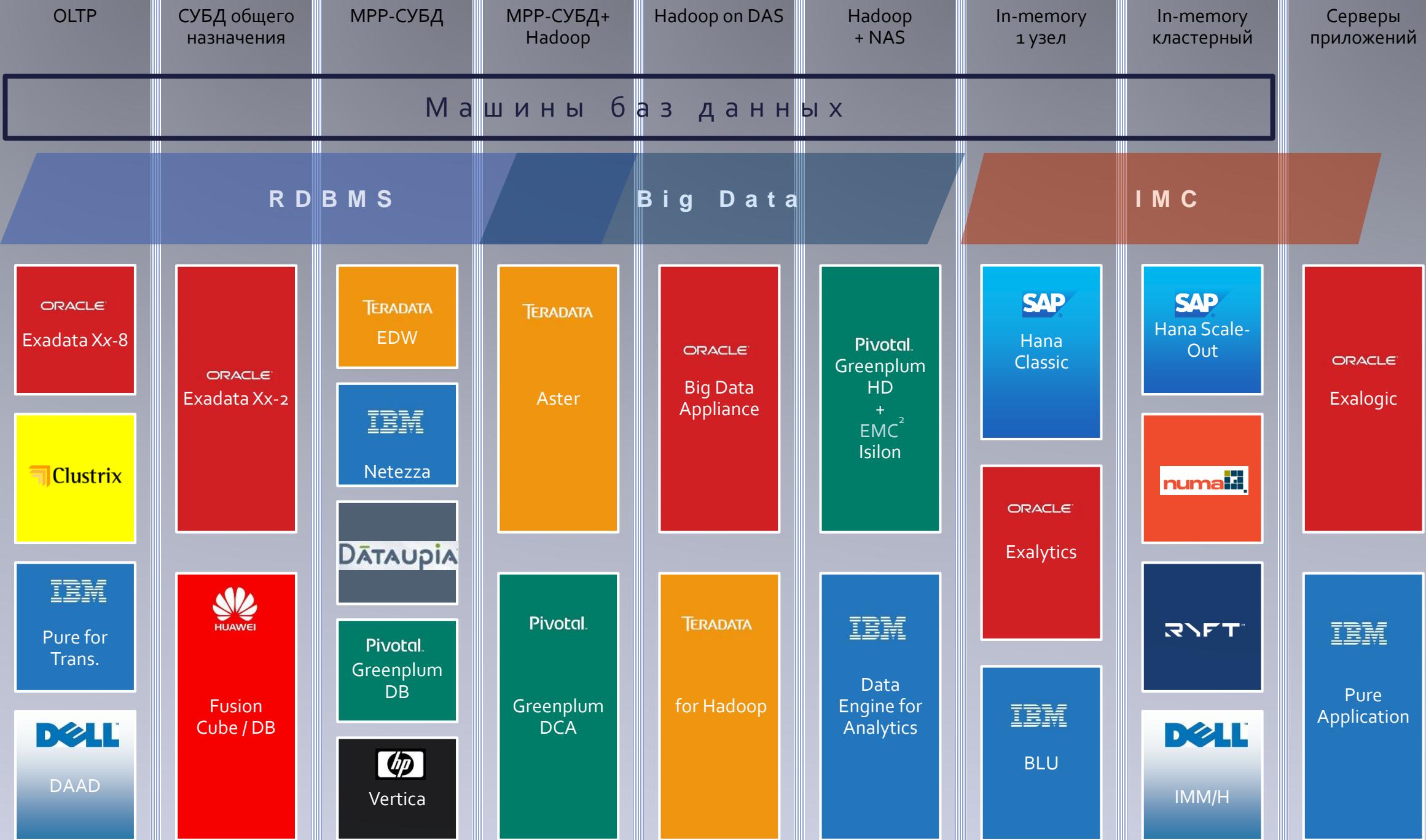
Сценарии
консолидации

Изоляция
приложений

Множество сред
в одном
комплексе
• PROD, TEST, DEV...

DBaaS

ЖИВУТ В ОБЛАКЕ



ЖИВУТ В ОБЛАКЕ



ORACLE CLOUD MACHINE



То же оборудование, но в собственности Oracle

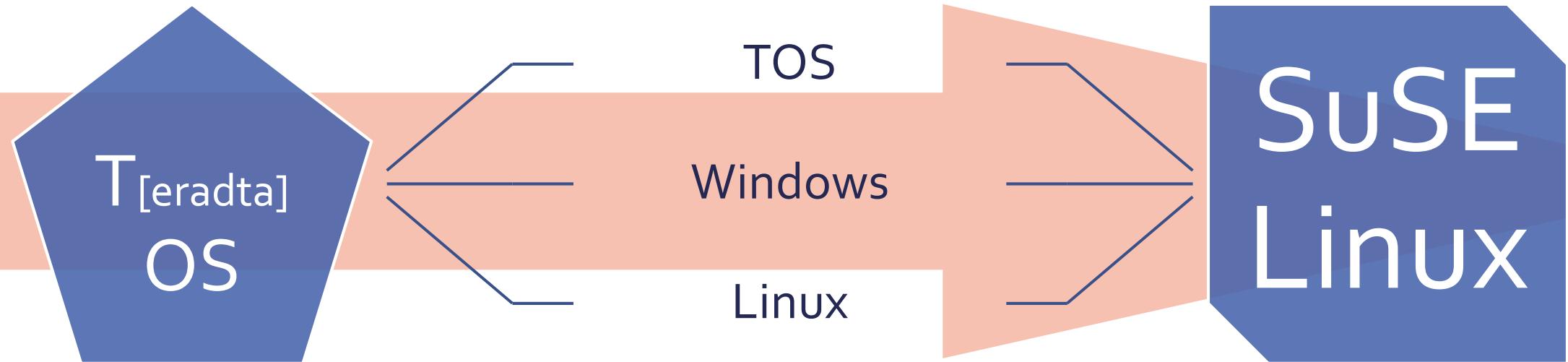
То же ПО, но в собственности Oracle

Устанавливается в ЦОД заказчика

Выделенное подключение к Oracle Corp.

Цены – как за публично-облачную
подписку

TERADATA: СТАНДАРТНЕЕ, ГИБЧЕ, ВИРТУАЛЬНЕЕ



Teradata on VMWare

Aster Data on VMWare

IntelliFlex

плотнее, стройблоки атомарнее

ЭВОЛЮЦИОННЫЕ ТЕНДЕНЦИИ

«Коммодитизация»

К x86

К упрощению топологий

«Демашинизация»

Greenplum,
Vertica...

Виртуальный
Aster

Специализация по типам нагрузок и цене терабайта

Линейки
Teradata

Линейки IBM

Линейки
Oracle

К облачной модели потребления

Oracle cloud machine

Amazon
RedShift



IV. ПОТРЕБИТЕЛИ

ПОКУПАТЕЛИ TERADATA

Банки

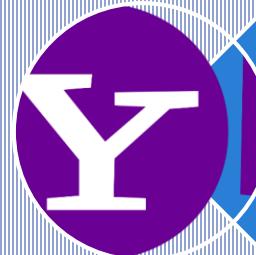
Большие
розничные сети

Большие
теле-
ком-
операторы

но
также....

2 ПБ

6,5 ПБ в
Greenplum –
у них же



Yahoo!



Cisco



EBay

2011: Джобс представляет ЦОД в Северной Каролине





Южнопортовый ЦОД
Сбербанка

ПОКУПАТЕЛИ EXADATA

Промышленность и энергетика

Госсектор

Банки

но также....

И снова Apple
для AppStore

Allegro Group

LinkShare
(онлайн-
маркетинг)

TargetBase
(онлайн-
реклама)

PRO – CONTRA для клиентов (общего характера)

Предкон- фиgурированное

Понятная
производительность
(проверить на готовом)

...не нужно
проектировать

“привезли – включили
– работает”

Одно окно

Проверка на стороне
одного производителя

Уже не скажут:
«а это не мы, а ОС
(гипервизор, СХД...)»

...

Vendor Lock-in

Запчасти –
от единственного
производителя

Ограниченный
манёвр по
поставщику

Сложности миграции

Цены...



V. ВОССОЗДАНИЕ

EXADATA СОБСТВЕННО РУЧНО...



Нелегально, но ...

Oracle
Database
Exadata
Storage
Software

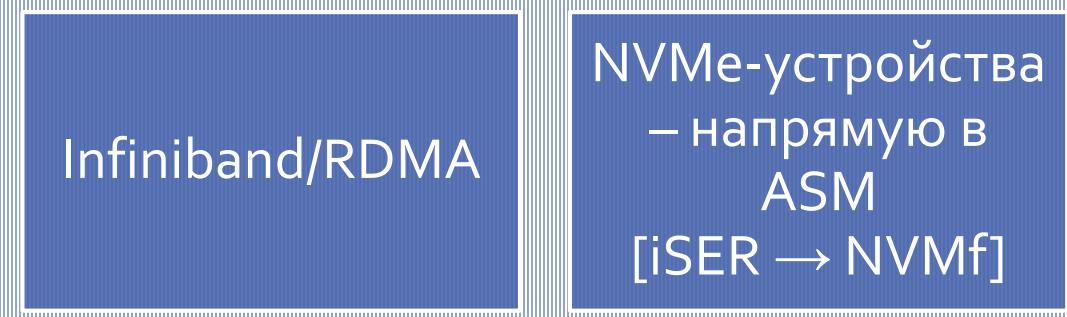
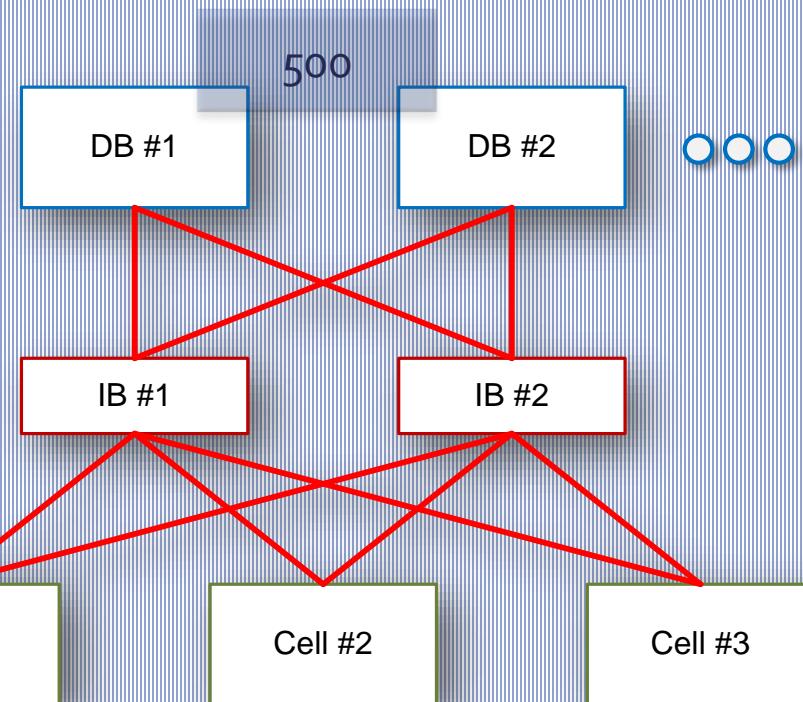
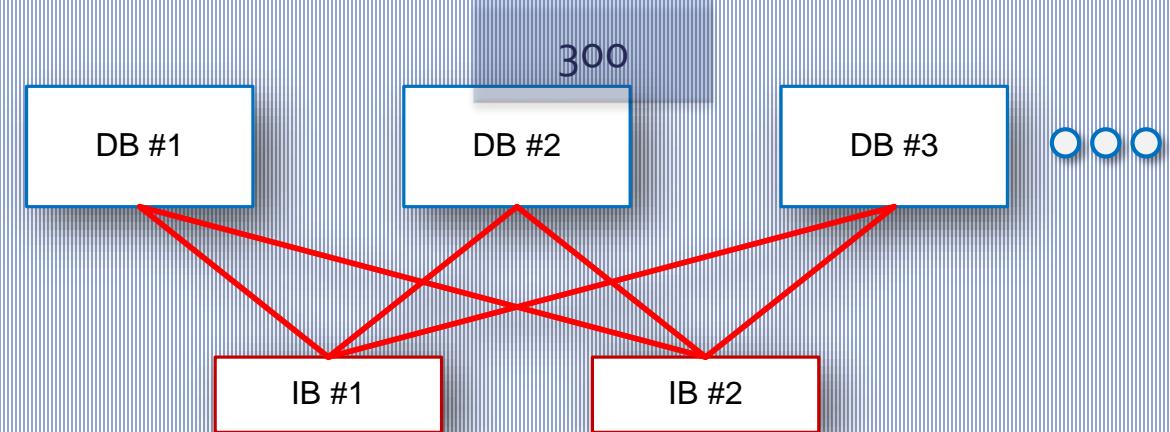
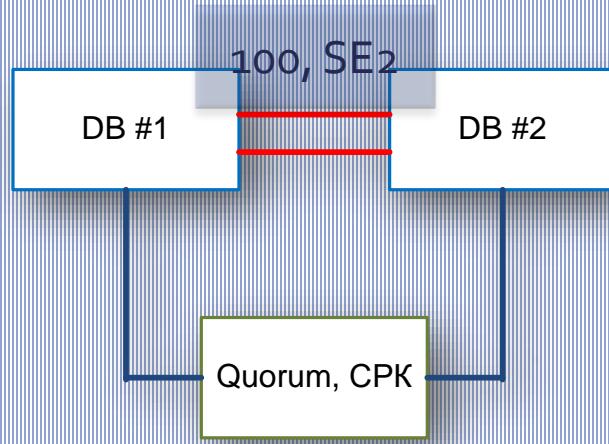
Сообщения
о запуске
на AWS

Сообщения
о запуске с
Infiniband-
сетью

В целом –
успешные
... с
особенностями

... и уже не чёрный ящик

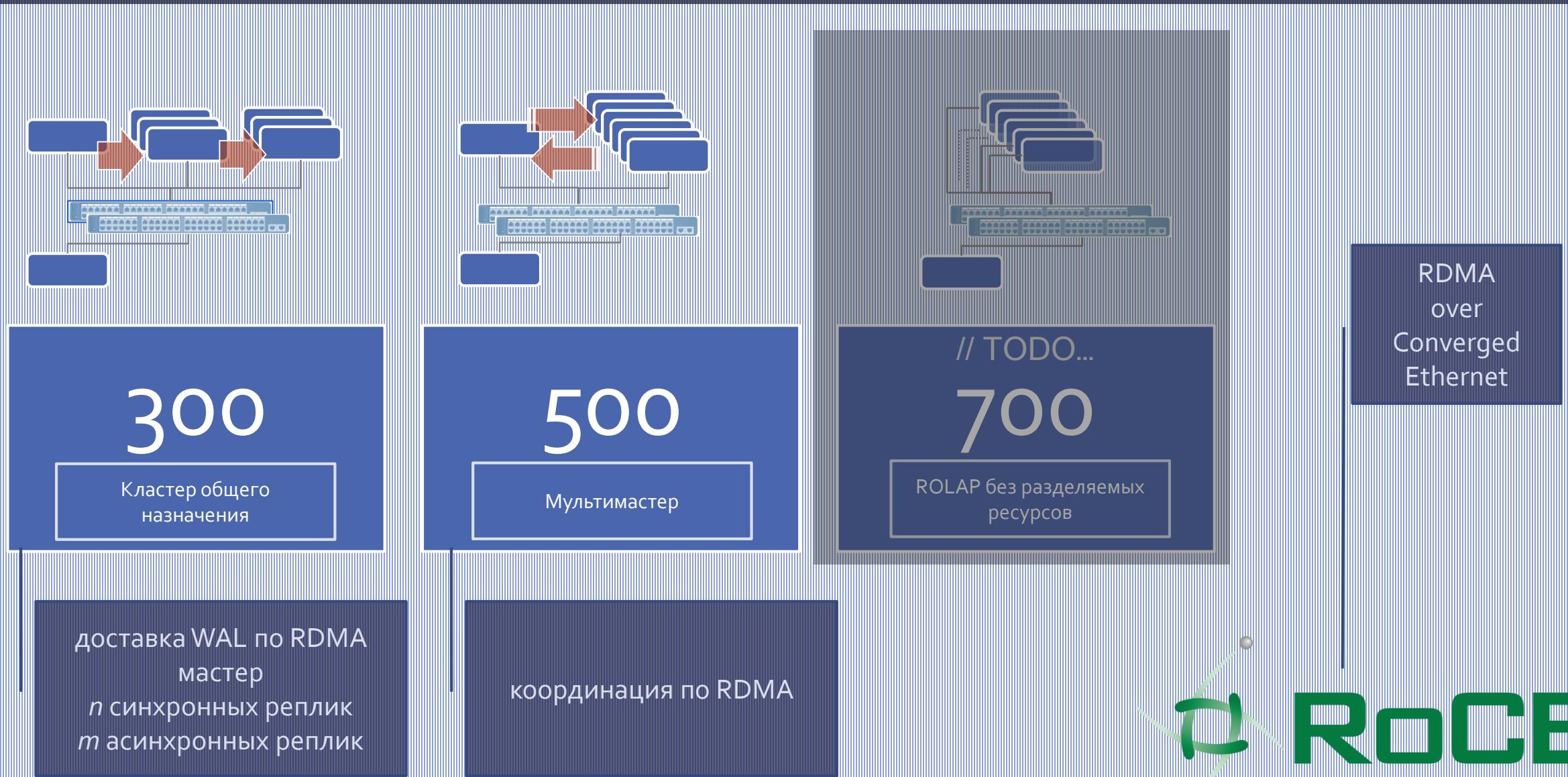
СКАЛА-СР / ORACLE DB



Flashgrid для
управления
устройствами в
ASM

NVMe-устройства
– напрямую в
ASM
[iSER → NVMe]

СКАЛА-СР / POSTGRES PRO: ПОСТГРЕС С RDMA



СКАЛА-СР: С МЫСЛЬЮ О БУДУЩЕМ

Комплекс для Hadoop и современных нагрузок



Пути оптимизации стоимости терабайта
(Data Lake, HDFS-enabled NAS)

Поддержка резидентных вычислений
(Spark, in-memory-NoSQL)

Федерация SQL-on-[MPP, Hadoop]

Кластер с графическими ускорителями

Адаптация новых аппаратных технологий



NVMe over Fabrics

GPGPU

FPGA-сопроцессоры (Intel+Altera)

Вычисления на сетевых картах

3D Xpoint (и SCM как класс)

Gen-Z

Гиерконвергентное ядро



Hyperconverged core

Storage-optimized node

NVMf-storage node

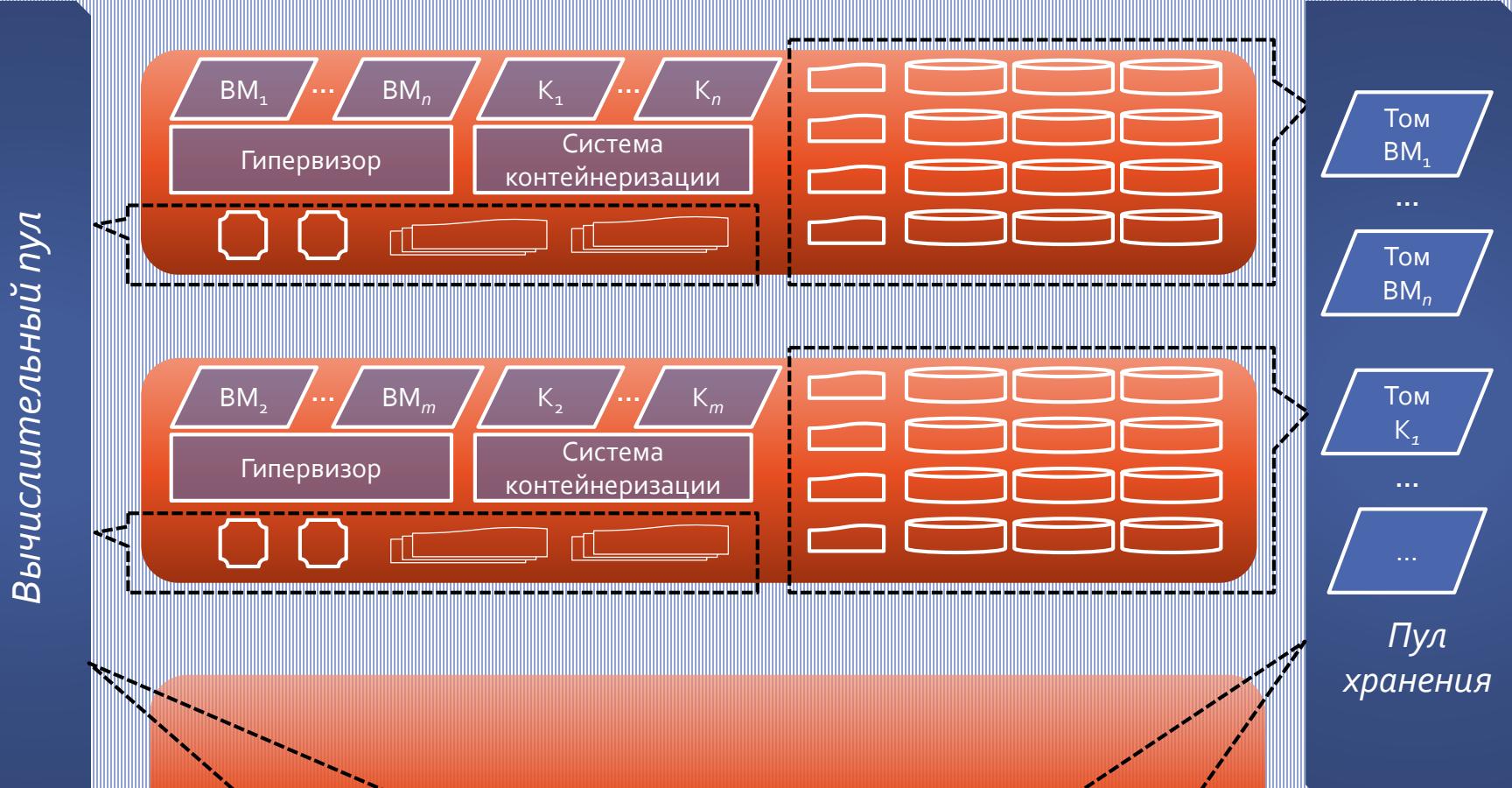
Database-storage node

Compute node

...

ГИПЕРКОНВЕРГЕНТНАЯ ИНФРАСТРУКТУРА

Кластерные системы виртуализации, использующие вместо выделенного массива программно-определенную сеть хранения, собранную из внутренних накопителей

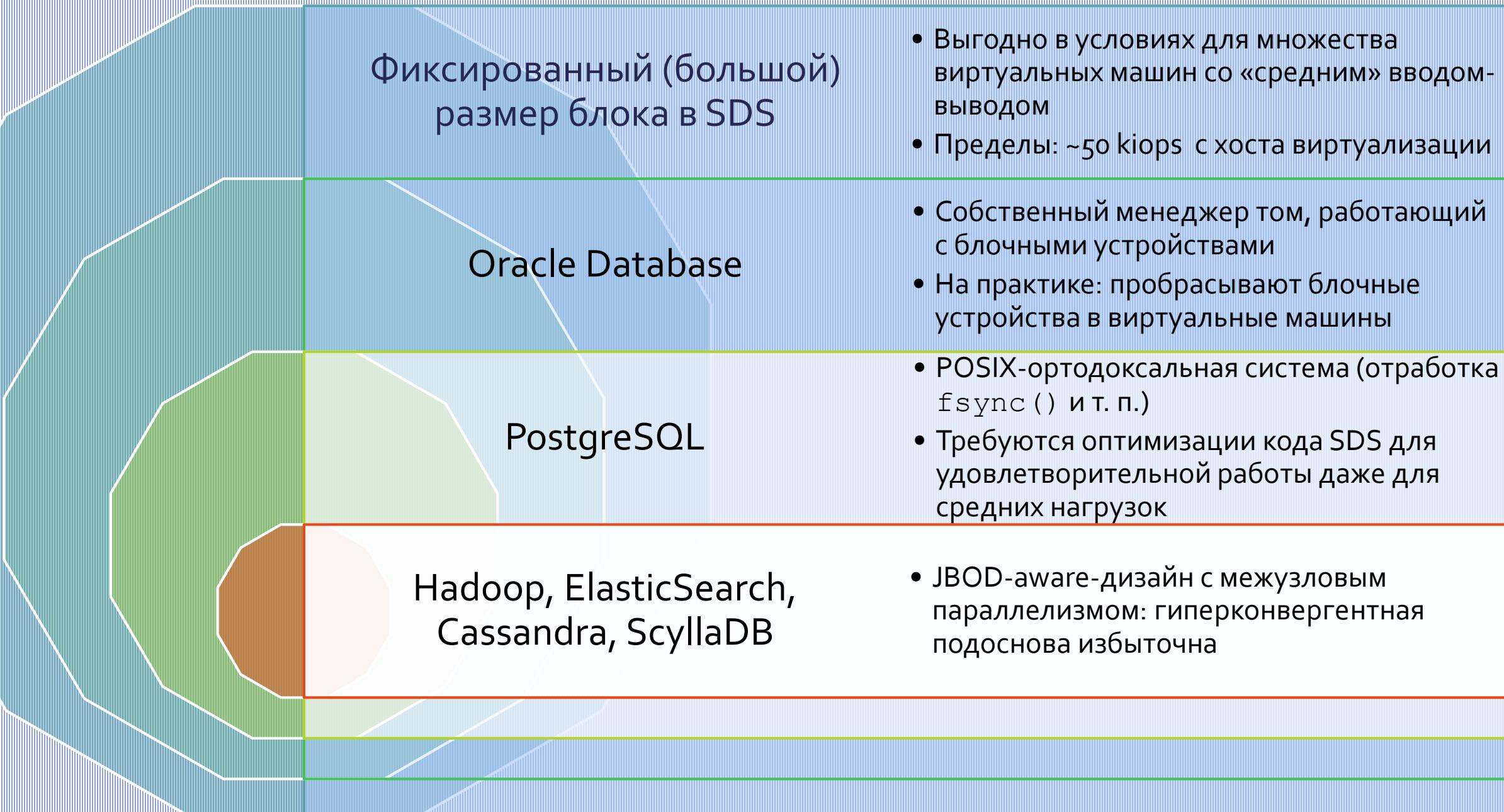


Вычислительный пул

«Серебряная пуля»

- Для нагрузок «общего назначения»
- Эластичная, экономичная, масштабируемая, отказоустойчивая...
- Основной движитель рынка ИТ-инфраструктуры

HCI: ПЛАТА ЗА УНИВЕРСАЛЬНОСТЬ



ПРОЕКТНЫЙ ОПЫТ: СТОРОННИЕ НАГРУЗКИ

Система развёртывания

Разливка ПО: включение нового узла, обновления, ...

Мониторинг

Комплекс в целом: платформа виртуализации, базы данных, сеть...

Управление

Узлы имён, ведущие, координаторы

Резервное копирование

Консоли администрирования

Переключатели при сбоях

Границевые вычисления

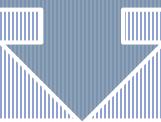
Потоковая обработка

Брокеры сообщений

Мониторы транзакций

Тестовые и разработческие среды

Не требовательны ко вводу-выводу, но многочисленны и изменчивы



Кандидаты к виртуализации в гиперконвергентной инфраструктуре

ВОКРУГ ГИПЕРКОНВЕРГЕНТНОГО ЯДРА

Виртуальные машины
управления, мониторинга,
развёртывания...

JBOD DFS

Дисковые блоки массового
параллелизма

хосты виртуализации

VM



HCI-ядро

коммутаторы
межсоединения (RoCE)

JBOD RS

Флэш-блоки массового
параллелизма

IM

Блоки резидентных
вычислений

Аппаратные блоки с
SSD-накопителями
и большой ёмкостью ОЗУ

RAID DB

Блоки баз данных

GPGPU

Блоки резидентных
вычислений
с акселерацией

Спасибо за внимание!

anikolaenko@ibs.ru

anikolaenko@acm.org

