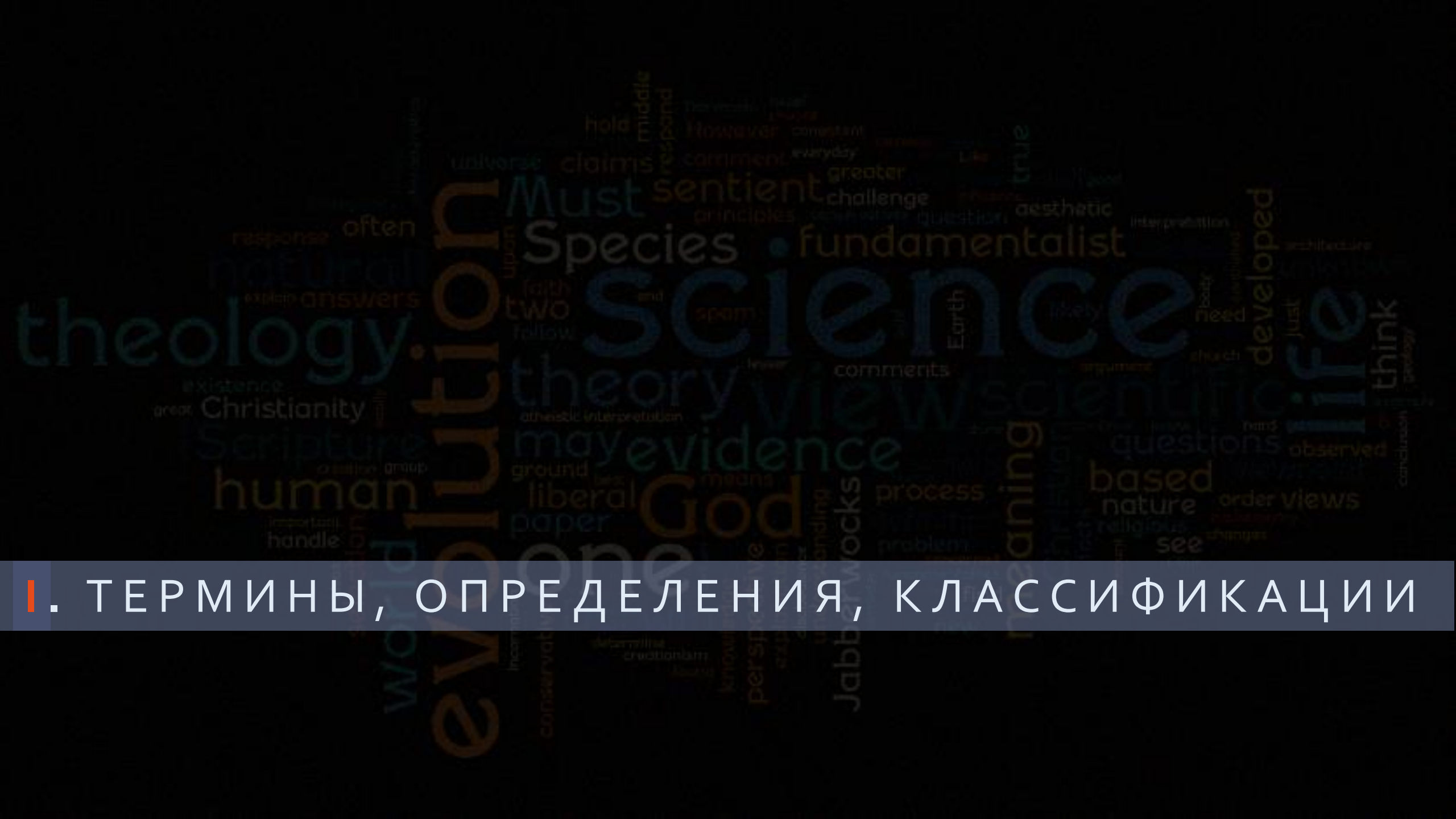


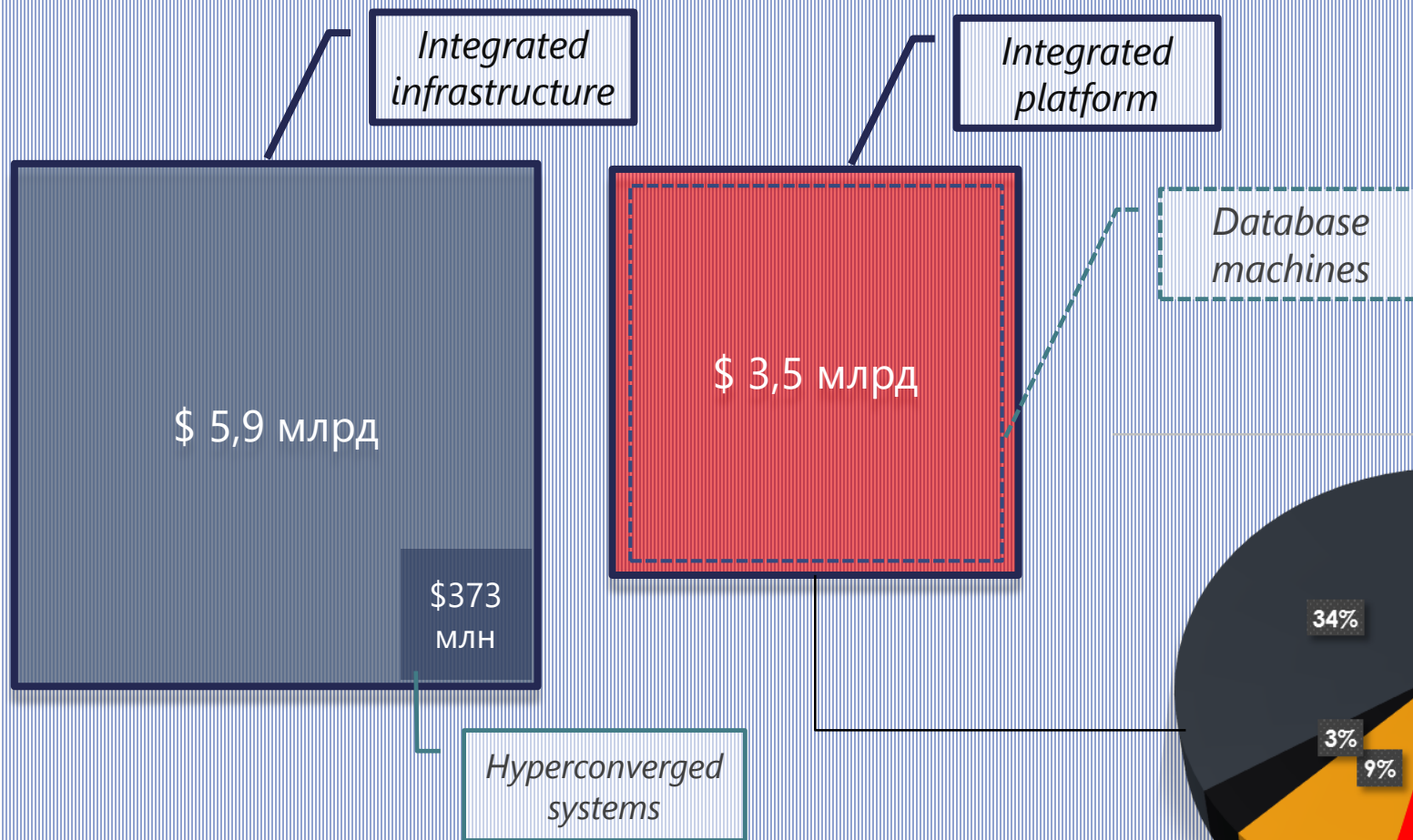
Андрей Николаенко
IBS

Машины баз данных: концентрированное обозрение

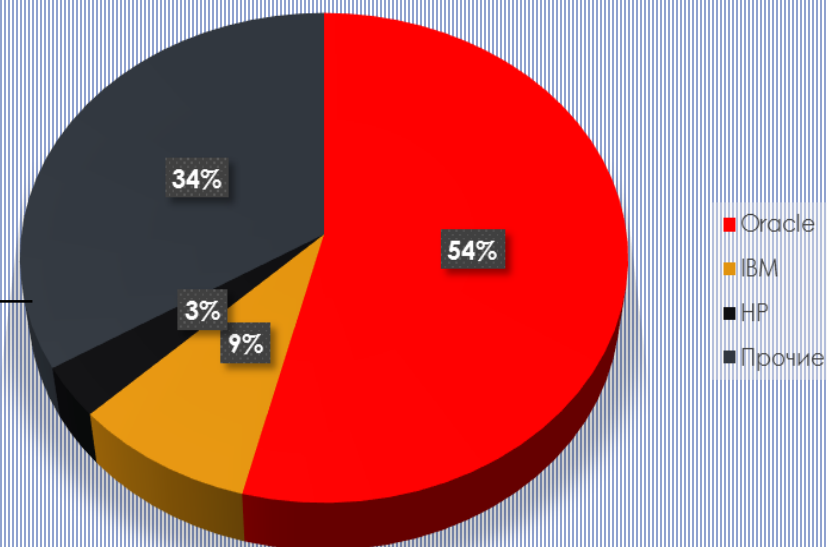


I. ТЕРМИНЫ, ОПРЕДЕЛЕНИЯ, КЛАССИФИКАЦИИ

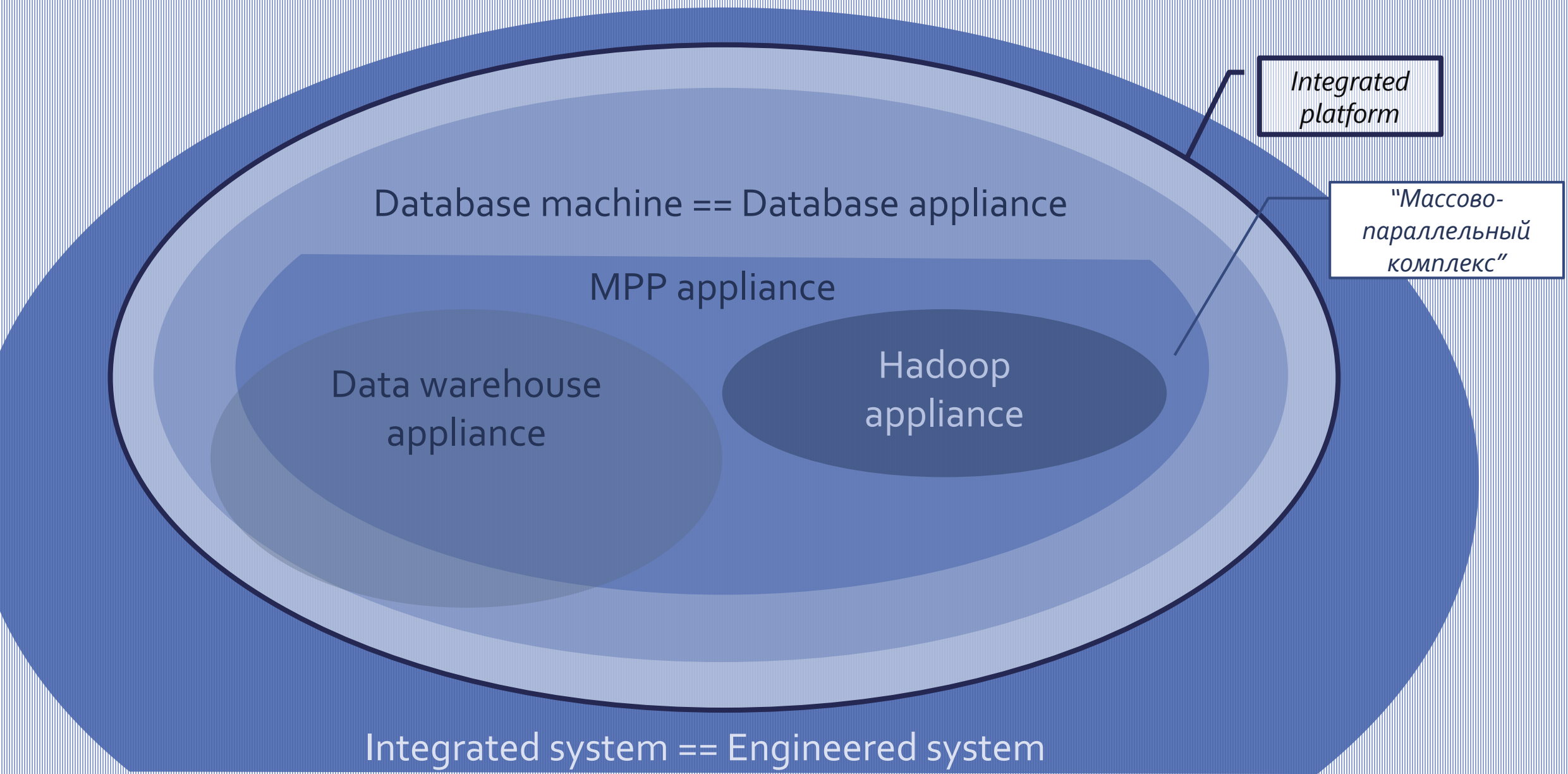
ИНТЕГРИРОВАННЫЕ СИСТЕМЫ

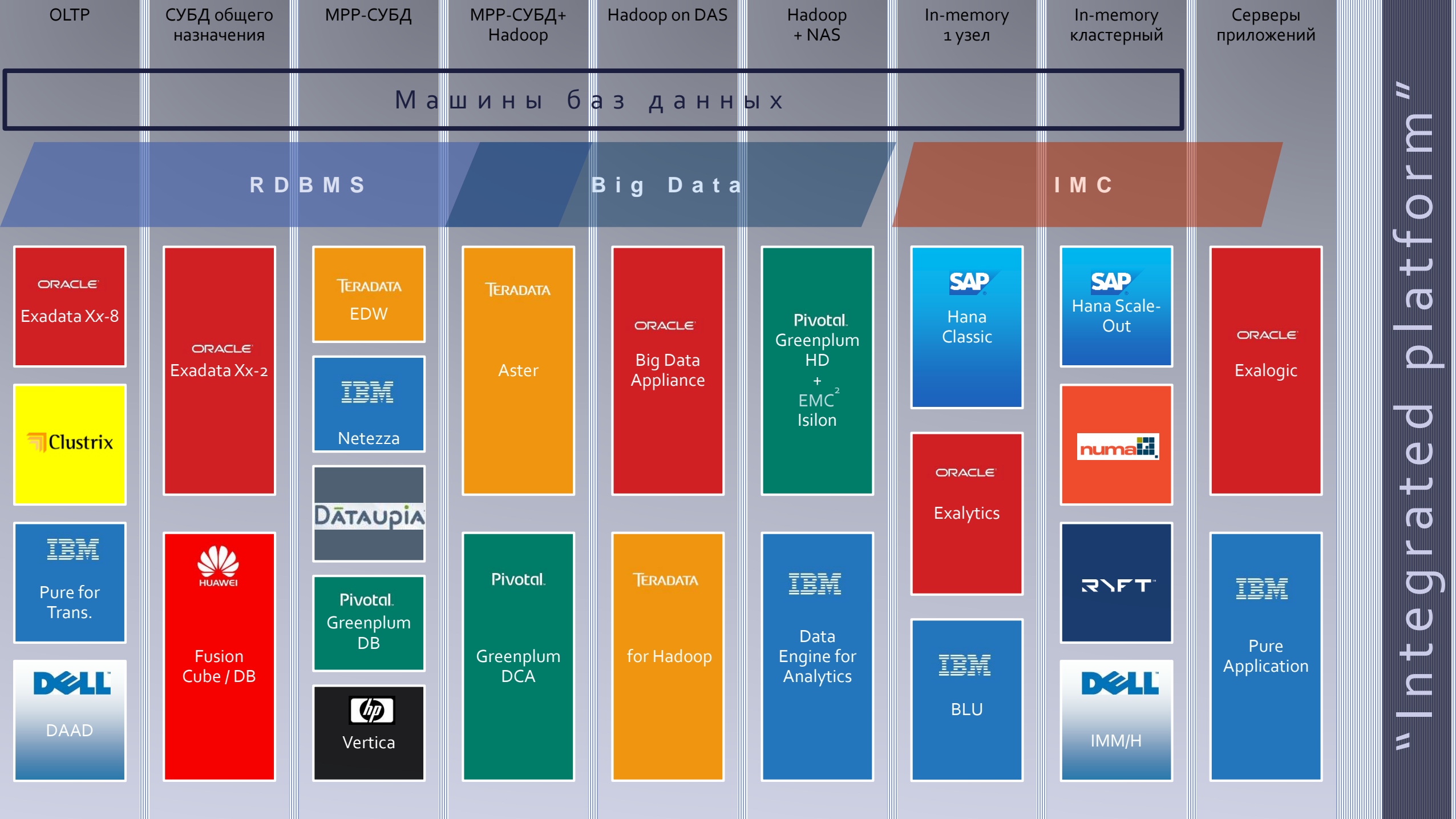


Данные IDC на 2015 год



ТЕРМИНЫ В ВАРИАНТАХ





“Integrated platform”



II. ОТДЕЛЬНЫЕ ПРЕДСТАВИТЕЛИ

TERADATA (классическая)

1 на шкаф

К л и к а
Clique

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

1–4 на клику,
до 512 на кластер

У з е л
Node

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

2 Infiniband-
коммутатора

В Y N e t
Banyan Network

- Межсоединение между узлами

2 (*max* 10) на узел,
120 сессий на PE

PE
parsing engine

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

4–20 на узел
(*max* 128 на узел)

AMP
access module processor

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

Дисковая группа

- Подключённая по FC группа накопителей

NETEZZA



Шкаф

- 1...8 на кластер

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
- Обработывает весь ввод-вывод к своей дисковой группе

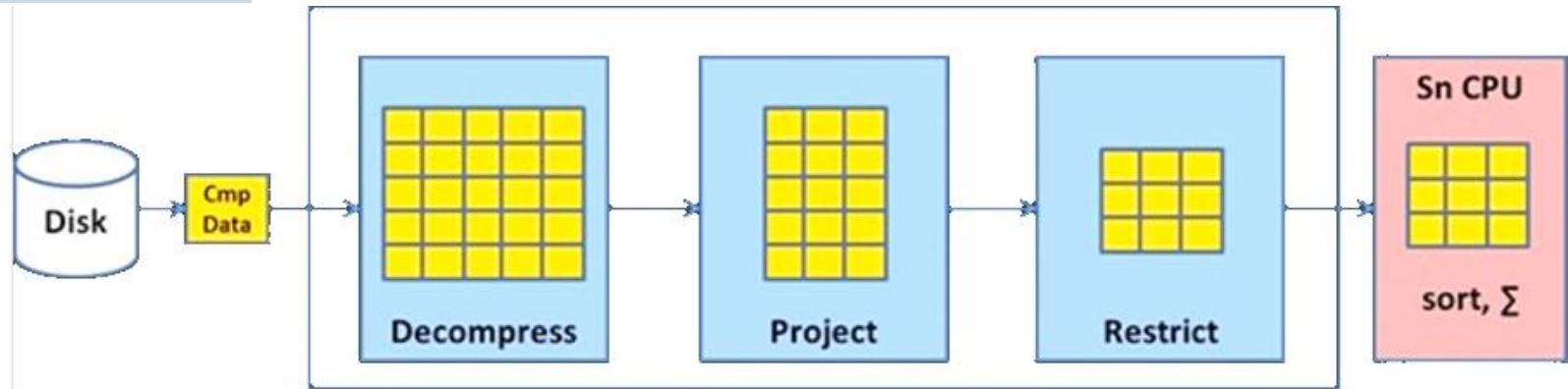
Диски

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

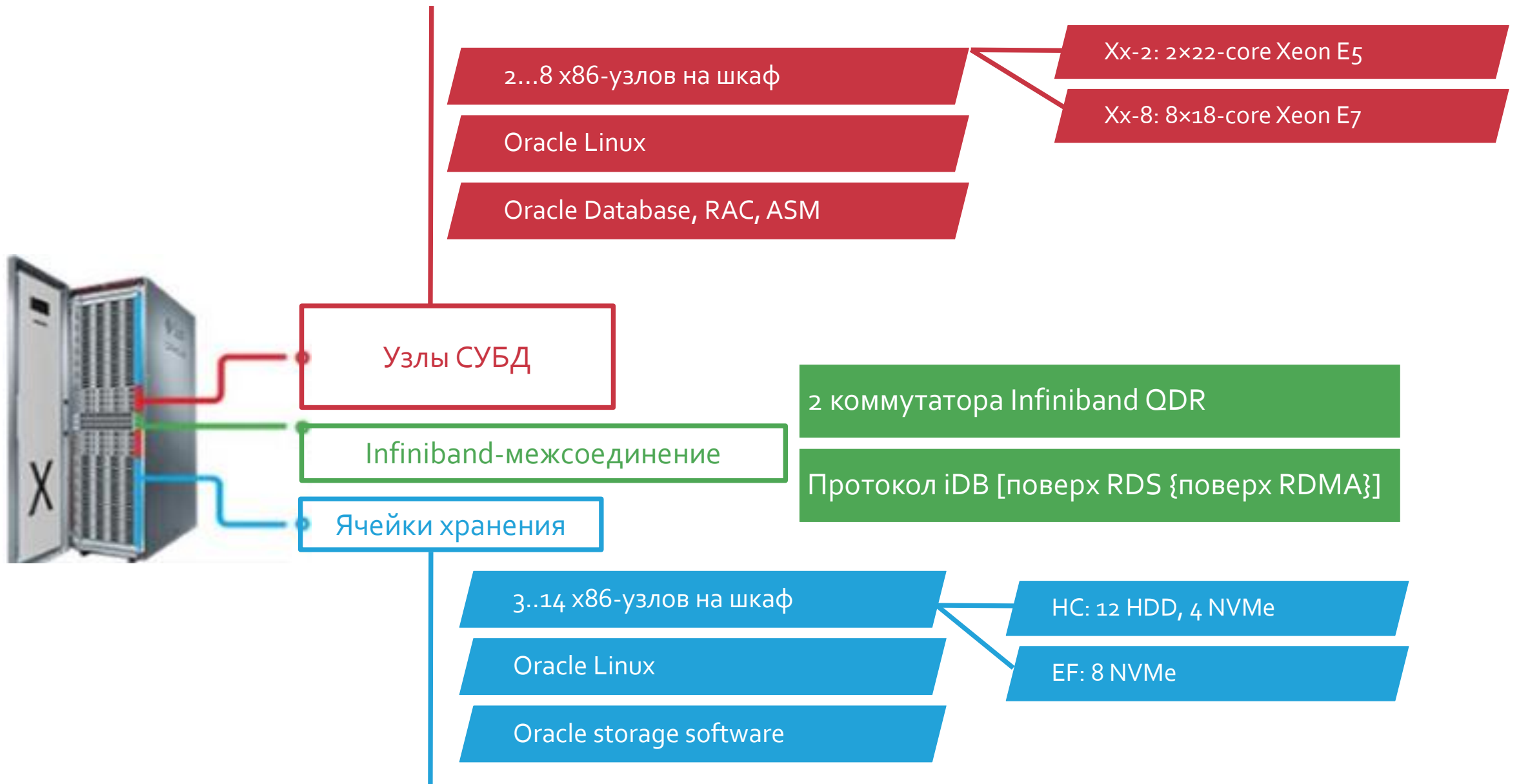
$A_{\text{[symmetric]}}$ MPP

Операции на ПЛИС:

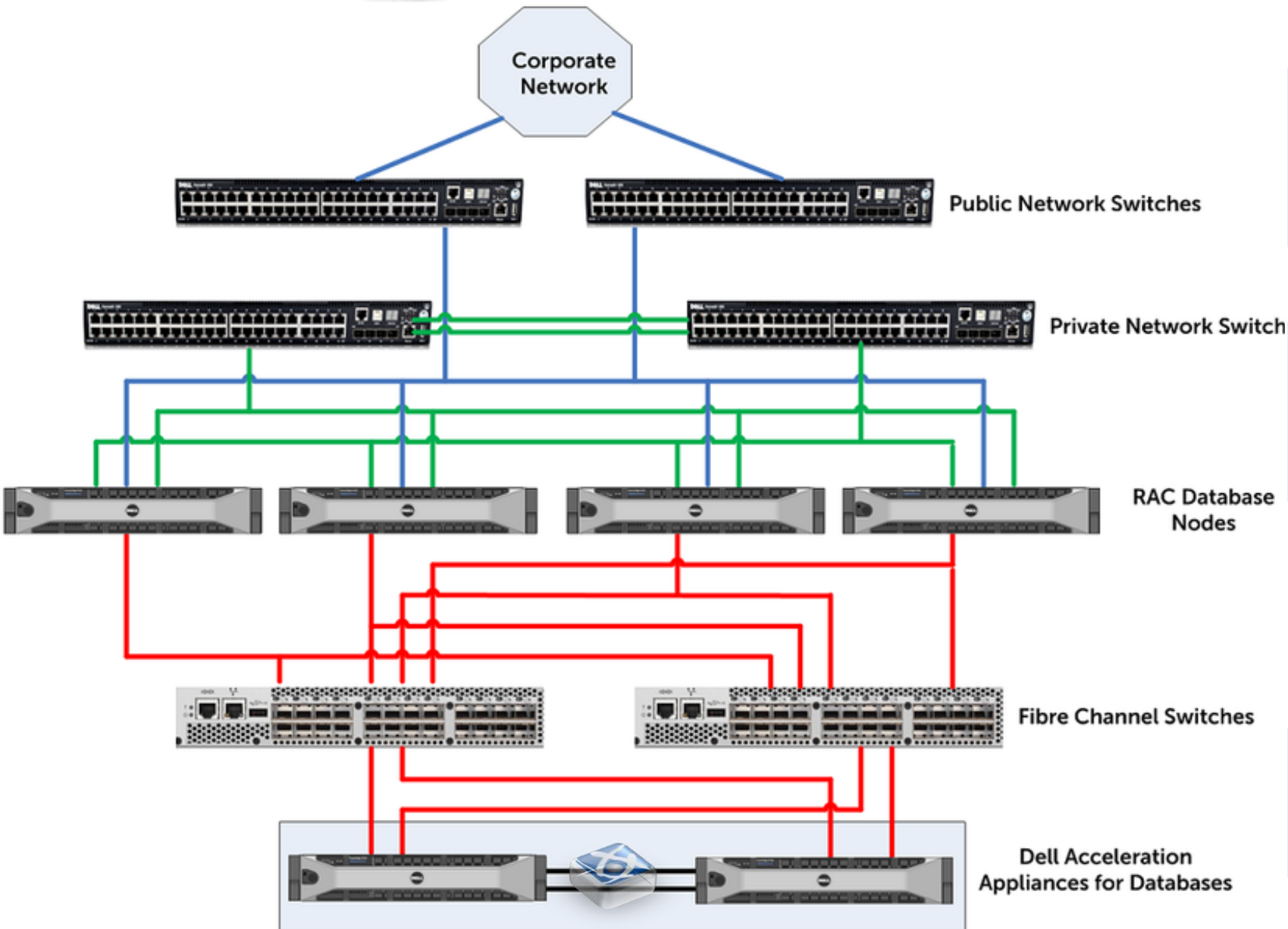
...на ЦПУ:



EXADATA



DELL DAAD



х86-стройблок для узлов СУБД и узлов хранения

- Dell PowerEdge, 2 × Xeon E5

Три вида сетевых технологий

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

Fusion-io ION на узлах хранения

ORACLE BIG DATA APPLIANCE & TERADATA APPLIANCE FOR HADOOP

18x2xCPU x86

- 12x8TB JBOD в каждом

Infiniband/QDR

- Межсоединение IPoIB
- Соединяемость с Exadata

Предустановленный софт

- Cloudera Data Hub Edition
- Oracle Table Access for Hadoop
- Oracle NoSQL CE
- Oracle R Enterprise
-



18x2xCPU x86

- 12x8TB JBOD в каждом

BYNet [Infiniband/QDR]

- Межсоединение IPoIB
- Соединяемость с Teradata

Предустановленный софт

- Cloudera CDH или Hortonworks



DELL-EMC DCA & ISILON

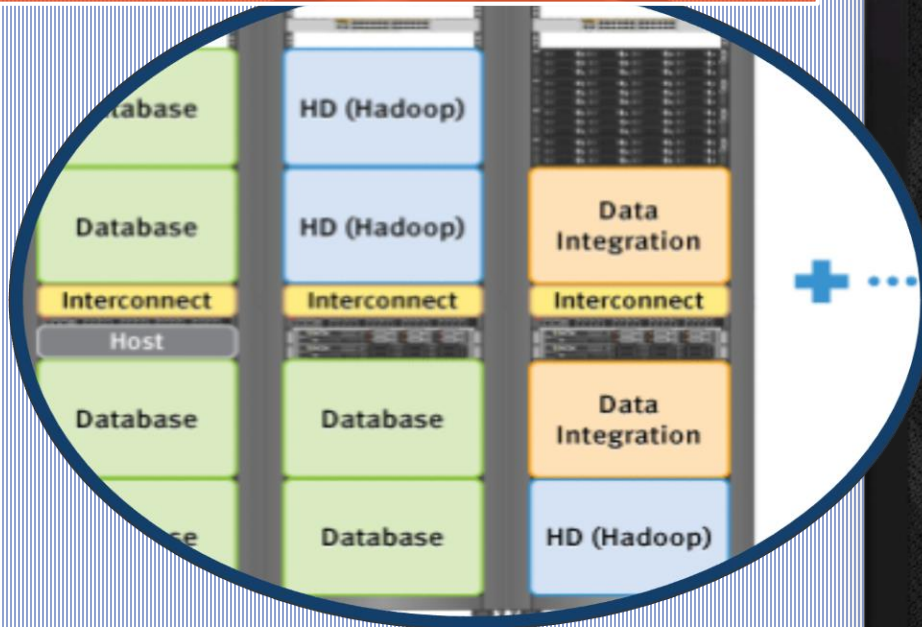


Greenplum appliance (DB only)

EMC Greenplum Data Computing
Appliance (+Hadoop, 2013)

EMC Data Computing Appliance (2015)

NAS,
доступный по HDFS



SAP HANA

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

Narrow your Search

- ▼ Certification Status
- ▼ Vendor
- ▼ CPU Architecture
- ▼ Memory
- ▼ Appliance Type
- ▼ Certification Date
- ▼ File System
- ▼ Operating Systems
- ▼ Certification Scenarios

Search Results as of 2017-05-24

1159 appliance models found

[certified](#)
[clear selection](#)

[Export as Pdf](#) [Compare](#)

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

Scale-out

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

РЕЗИДЕНТНЫЕ ВЫЧИСЛЕНИЯ «В ЖЕЛЕЗЕ»

	OLTP-SQL	MOLAP	BI	Data Discovery	Stats	Apps	OLTP-NoSQL	Hadoop
SAP Hana	Hana in-memory database			KXen	PAL R	SAP S/4		Vora (Spark)
ORACLE Exalytics	DB 12c (Timesten)	Essbase	OBIEE	Endeca	R (DB option)	Hyperion		
IBM BLU	DB2 BLU				R			
numa	MonetDB				RevolutionR			Spark
RYFT								Spark
IBS Скала-СР / Аналитика				Полиматика				
DELL IMM/H								Spark

ШАБЛОНЫ ПОСТРОЕНИЯ



Гиперконвергенция

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



Дезагрегация

- Узлы обработки
- Узлы хранения
- Специальный протокол общения

ХАРАКТЕРИСТИЧЕСКИЕ ПРИЗНАКИ



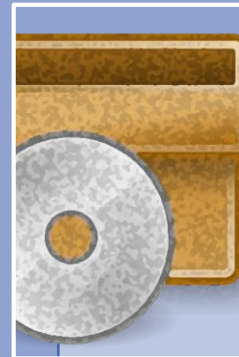
$n \times$ [Commodity x86]
как строительный
блок



Infiniband/RDMA
[часто – просто 10G]



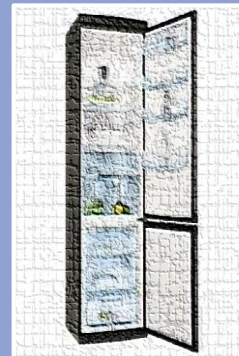
FPGA
[редко]



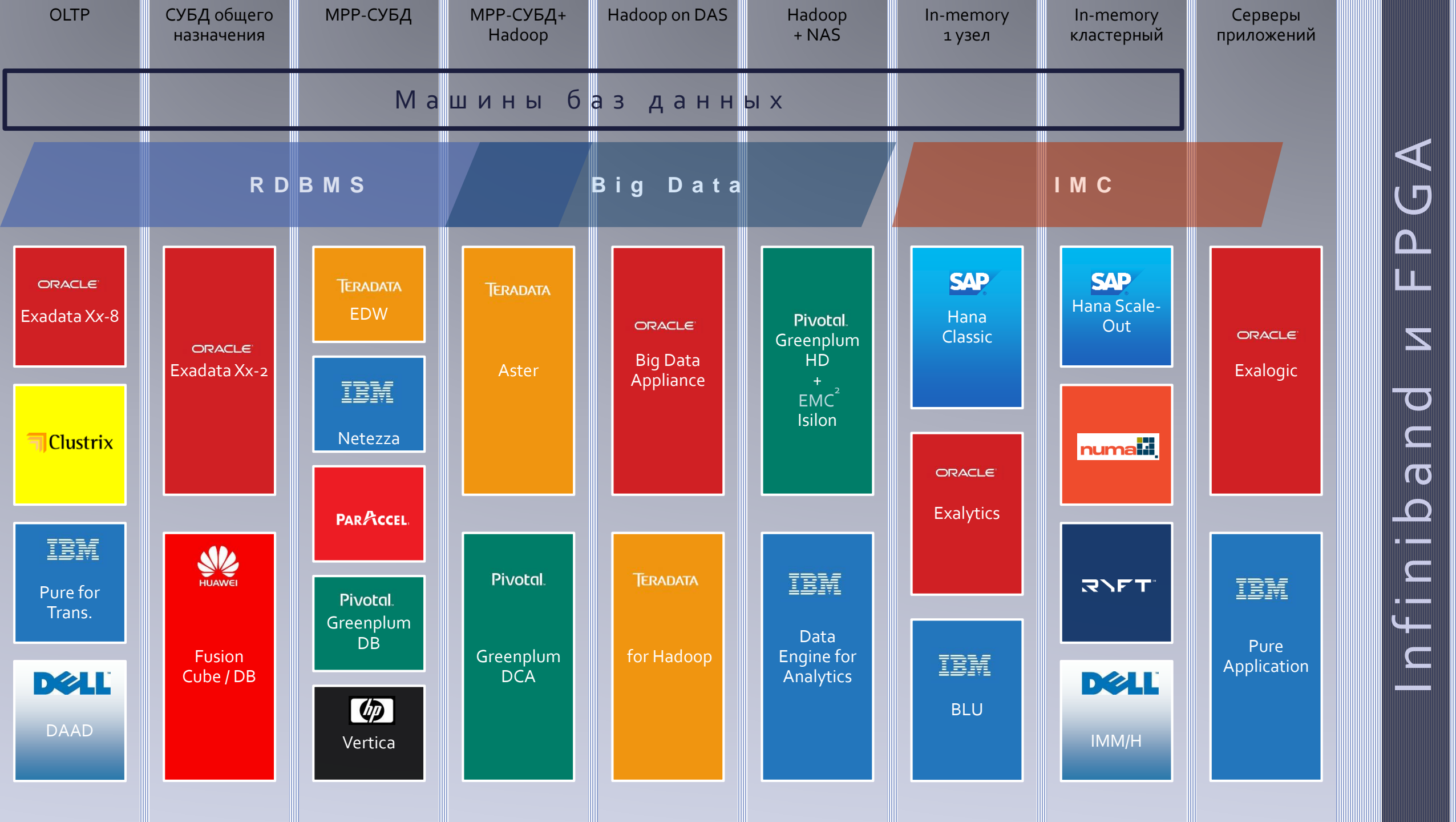
Программные
добавки, программная
определяемость

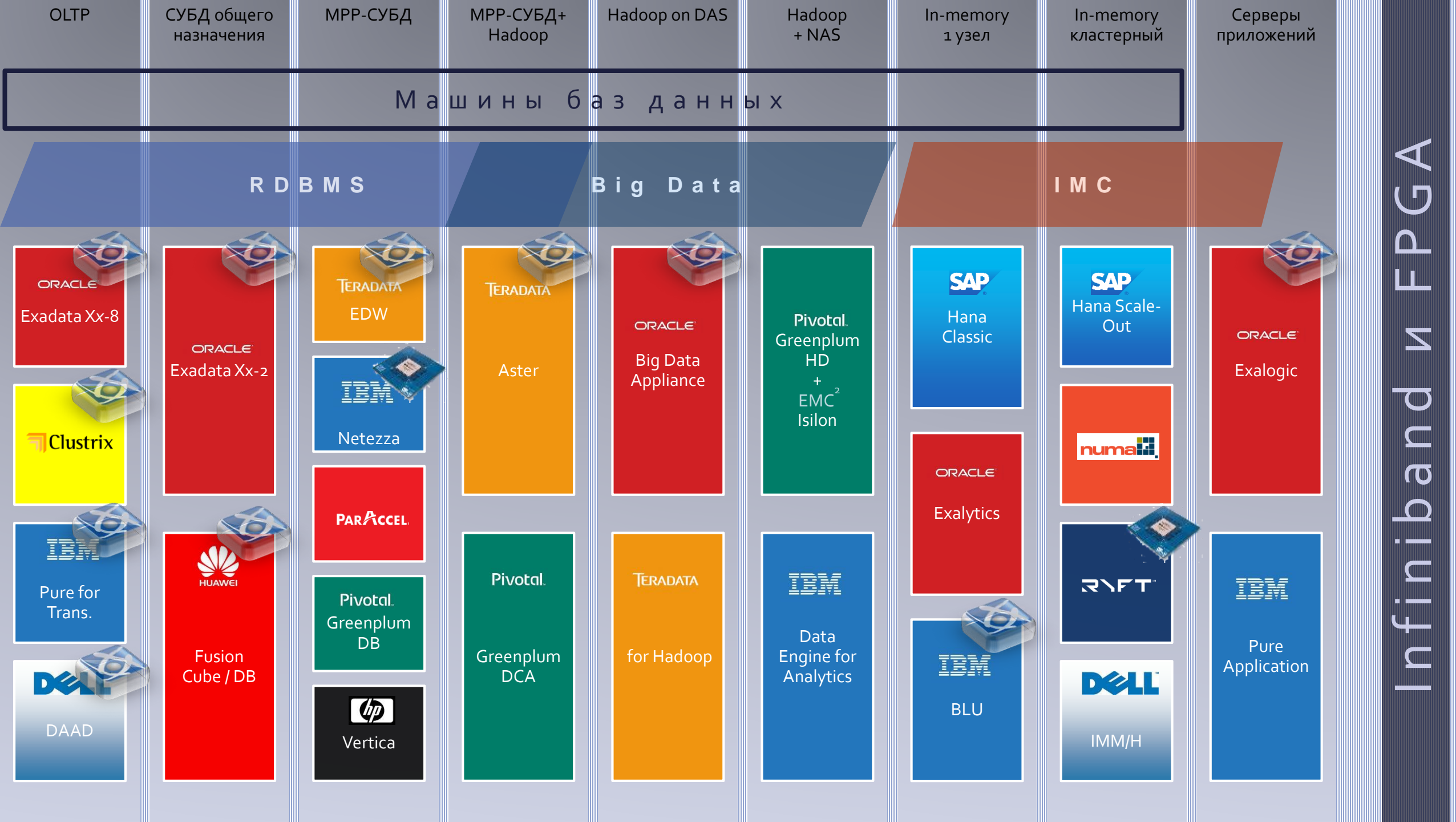


Без единых точек отказа,
дублирование компонентов,
самолечение
[не всегда и не во всём]



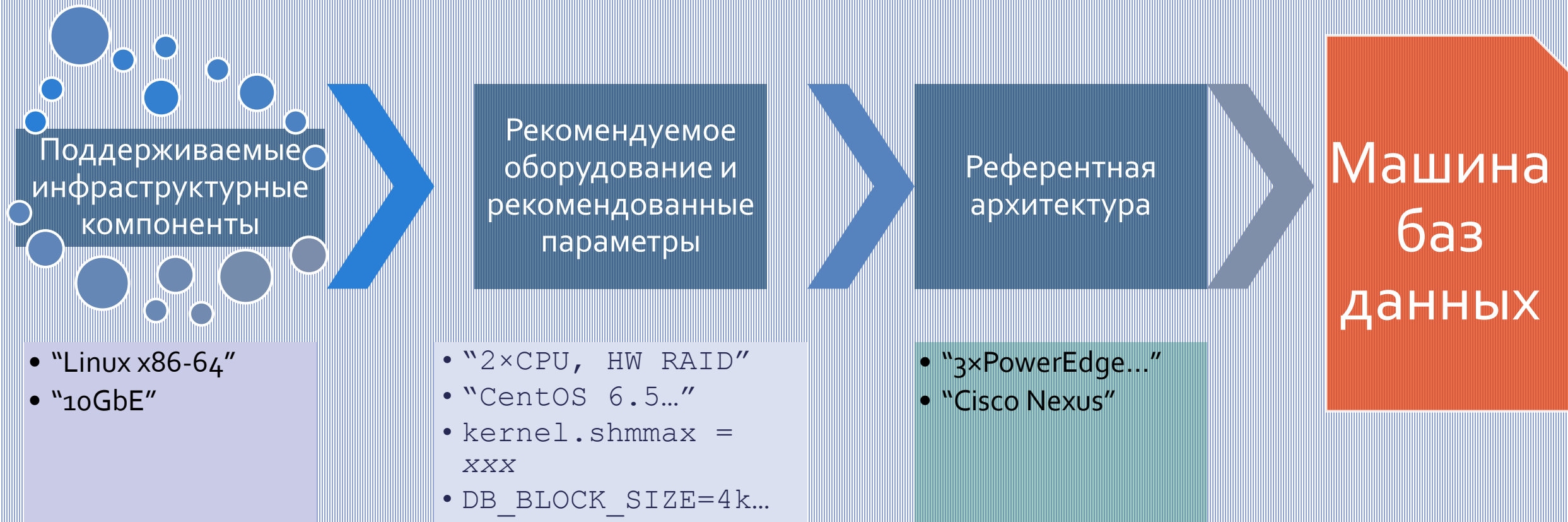
Pre-configured,
self-managed,
DBaaS off-the-shelf





Infiniband и FPGA

АРХИТЕКТУРНЫЙ КОНТИНУУМ





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

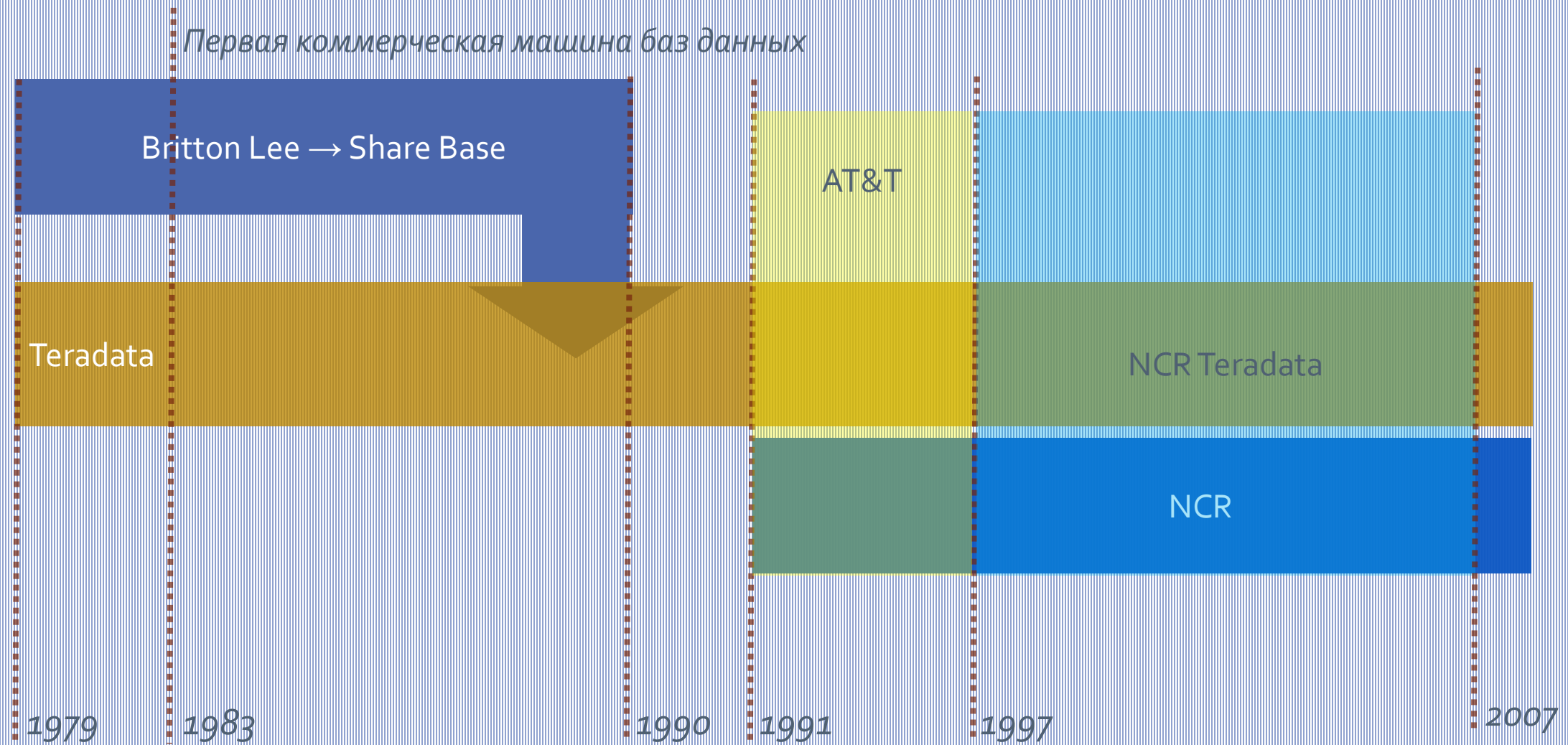
by

Haran Boral
David J. DeWitt

Computer Sciences Technical Report #504
July 1983

*Прогресс в дисковых массивов
и подсистем 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

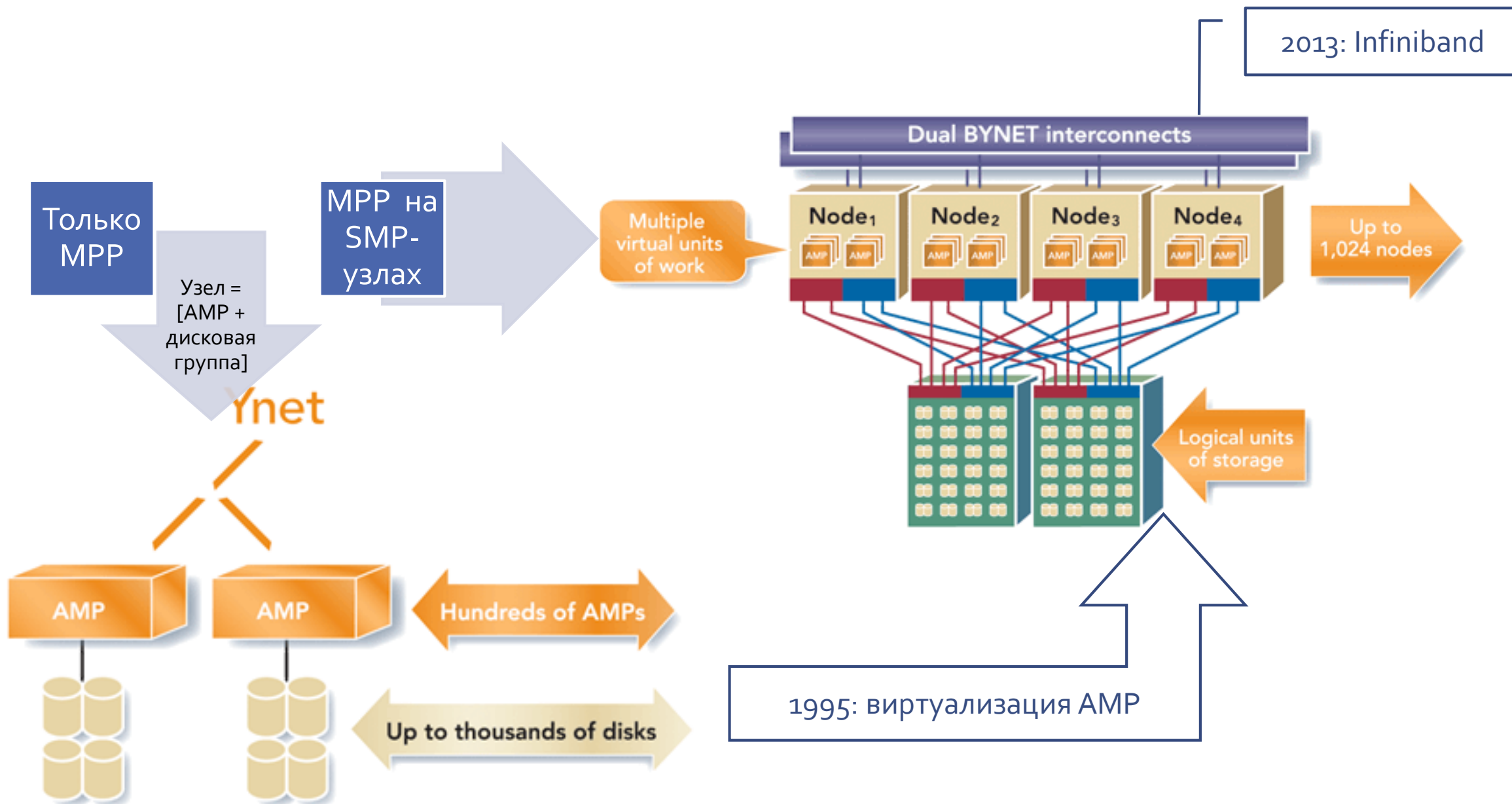
SD

- shared disk

SN

- shared nothing

TERADATA: I ПЕРЕДЕЛ (1995)



2002:

момент идентификации
рынка машин хранилищ
данных



Фостер Хиншоу
сооснователь Netezza,
основатель Dataupia

«ВОЗЬМЁМ-КА POSTGRESQL!»

Netezza

- IBM

Greenplum

- Pivotal / Dell-EMC

ParAccel

- Actian

Aster Data

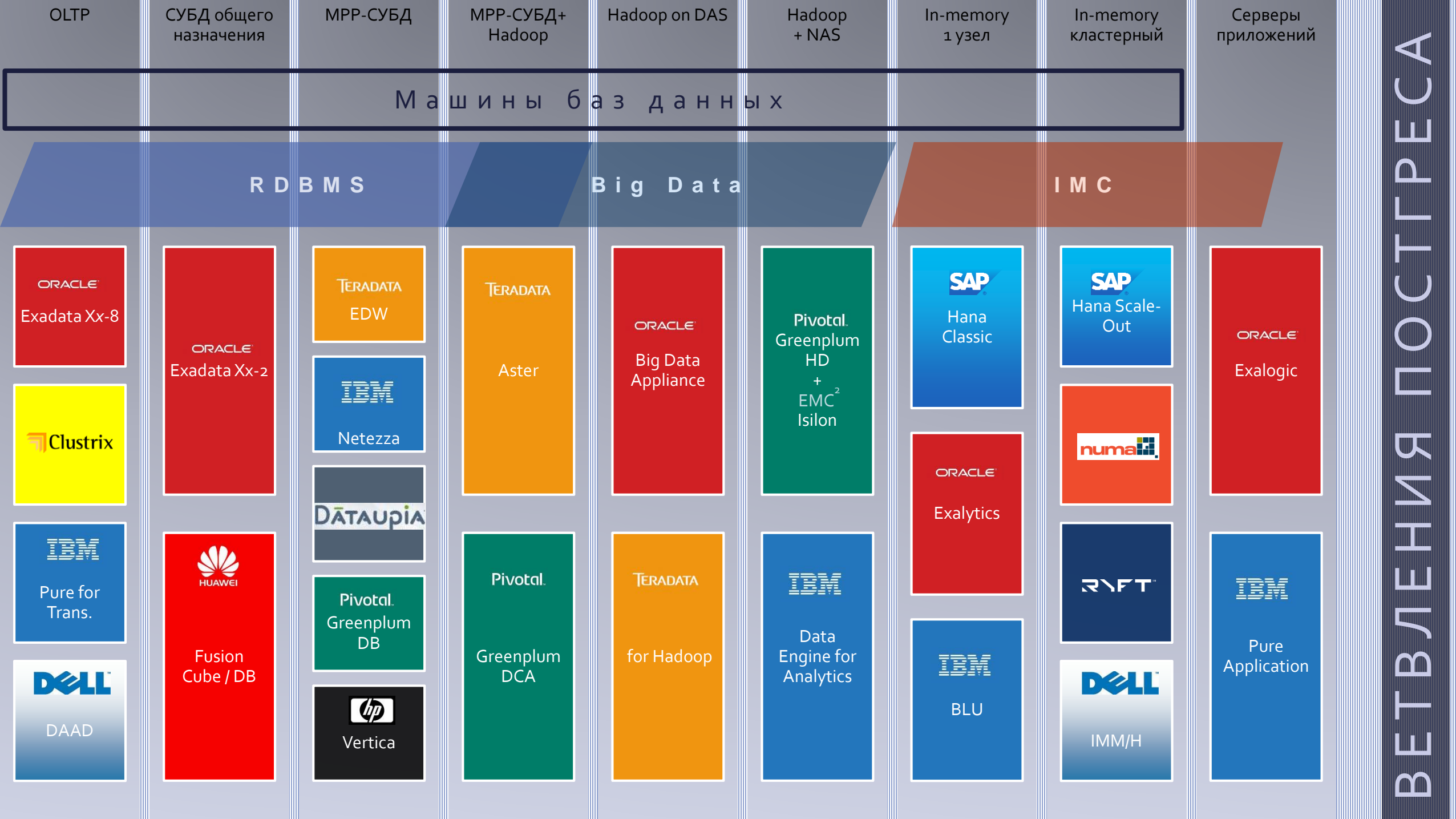
- Teradata

Dataupia

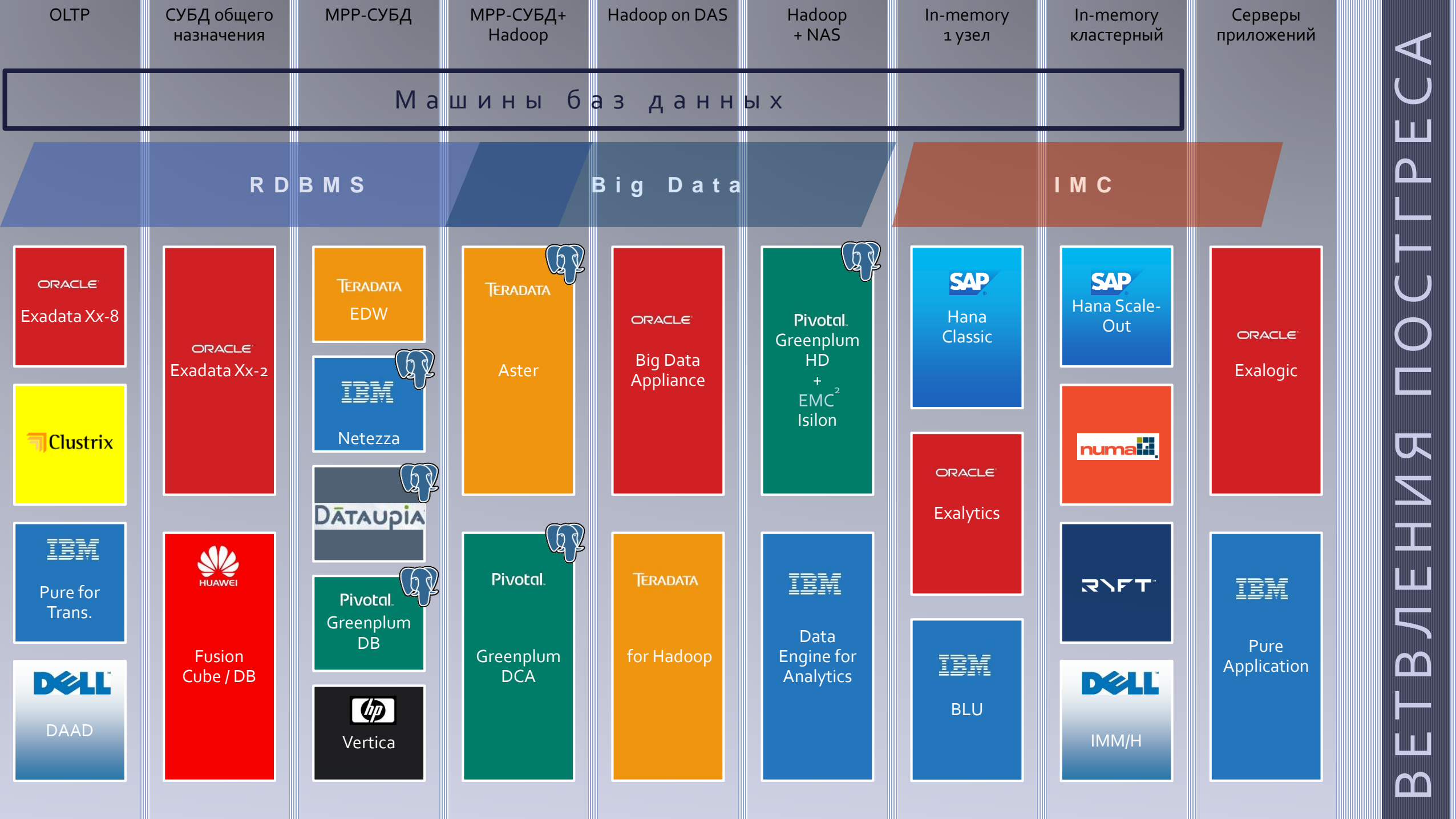
- *Seeking for acquisition...*

Hadapt

- Teradata



ВЕТВЛЕНИЯ ПОСТГРЕСА

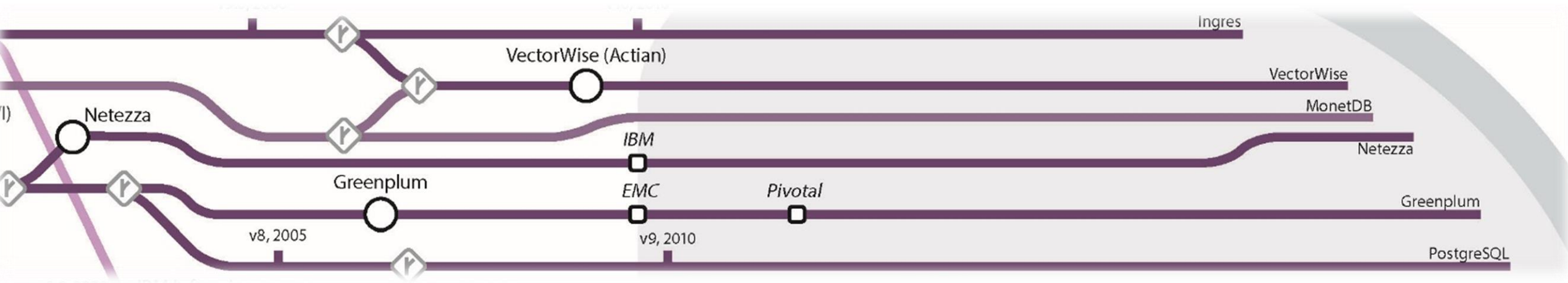


ВЕТВЛЕНИЯ ПОСТГРЕСА

МРР-машины 2000–2005 годов (на пятидесятимиллионном рынке)

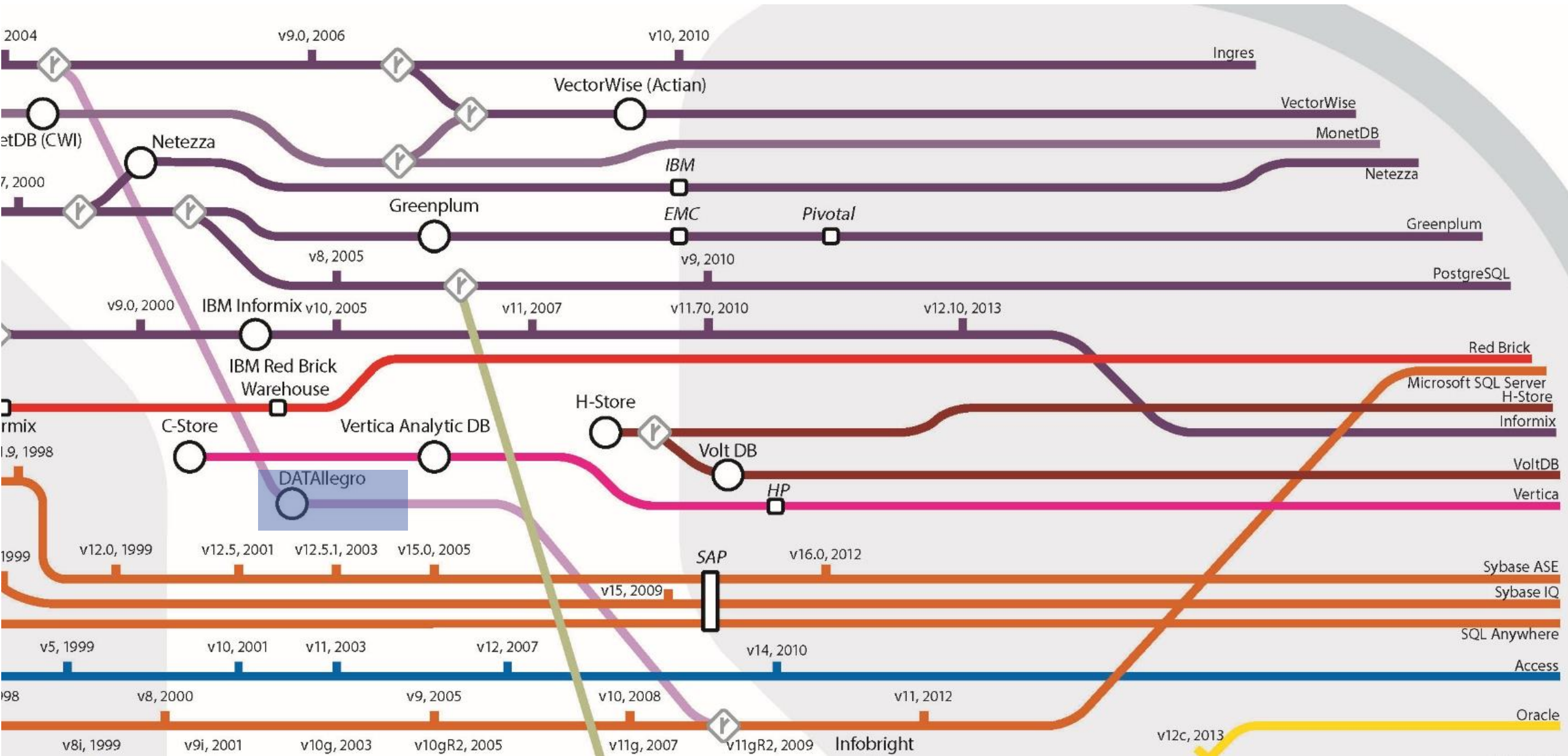
	Netezza	Datallegro	Greenplum	ParAccel
Возникновение	2002	2003	2003	2005
Строительный блок	«Сниппет» [PowerPC + FPGA]	Dell PowerEdge [x86]	Sun Fire на AMD [x86]	“Blade cluster” [x86]
Кодовая база	PostgreSQL	Ingres	PostgreSQL	PostgreSQL
Поглощение	2013 IBM	2008 Microsoft	2010 EMC	2011 Actian
Современность	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*

DATALEGRO НА СТЫКЕ СЕМЕЙСТВ



NON-APPLIANCE MPP DBMS

Exasol

Резидентная

Колоночная

Чемпион в
кластерном TPC-H

InfiniDB

На основе MySQL

Колоночная

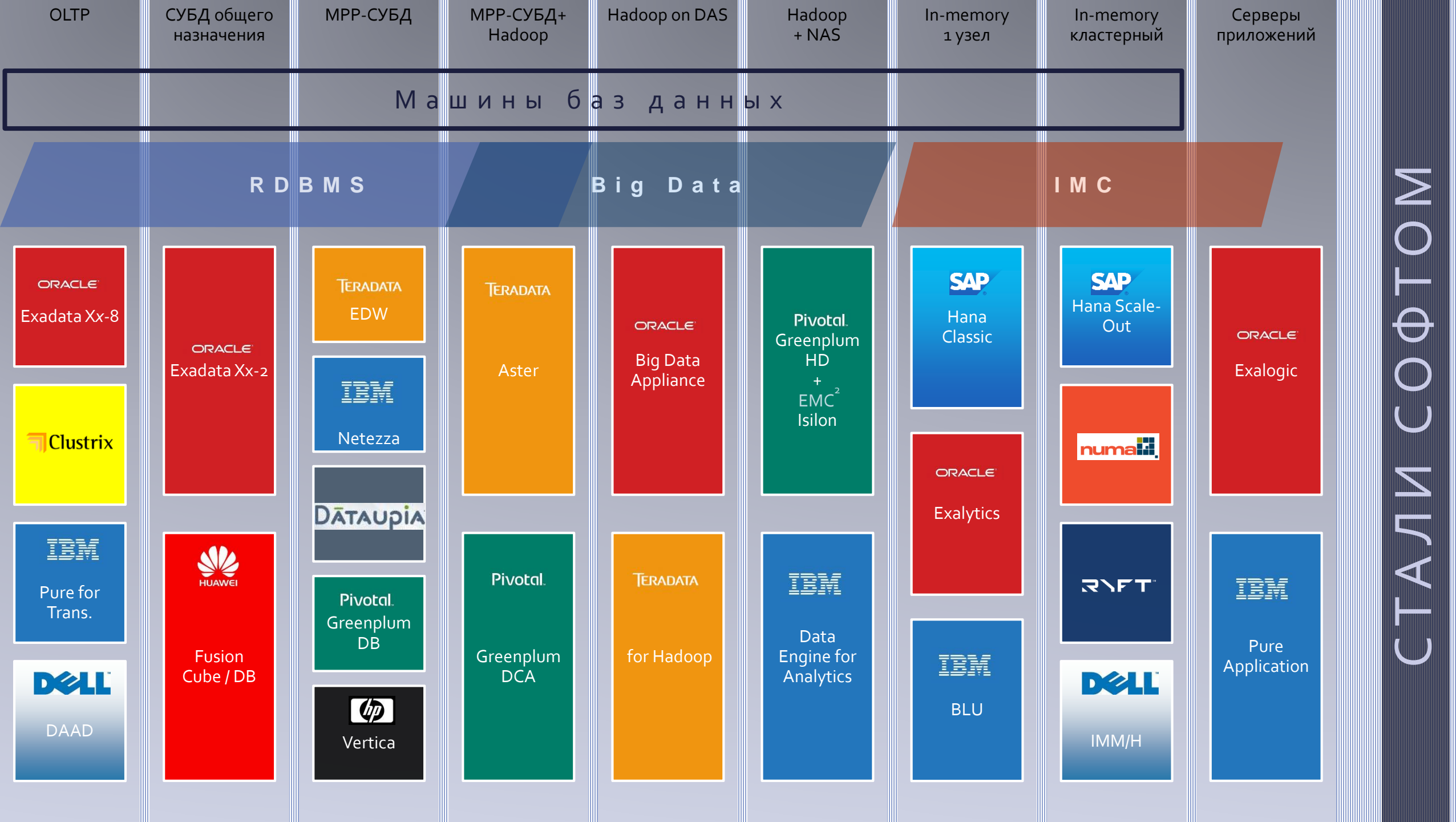
Компания Calpont
обанкрочена в 2014
году

Ряд решений над PostgreSQL

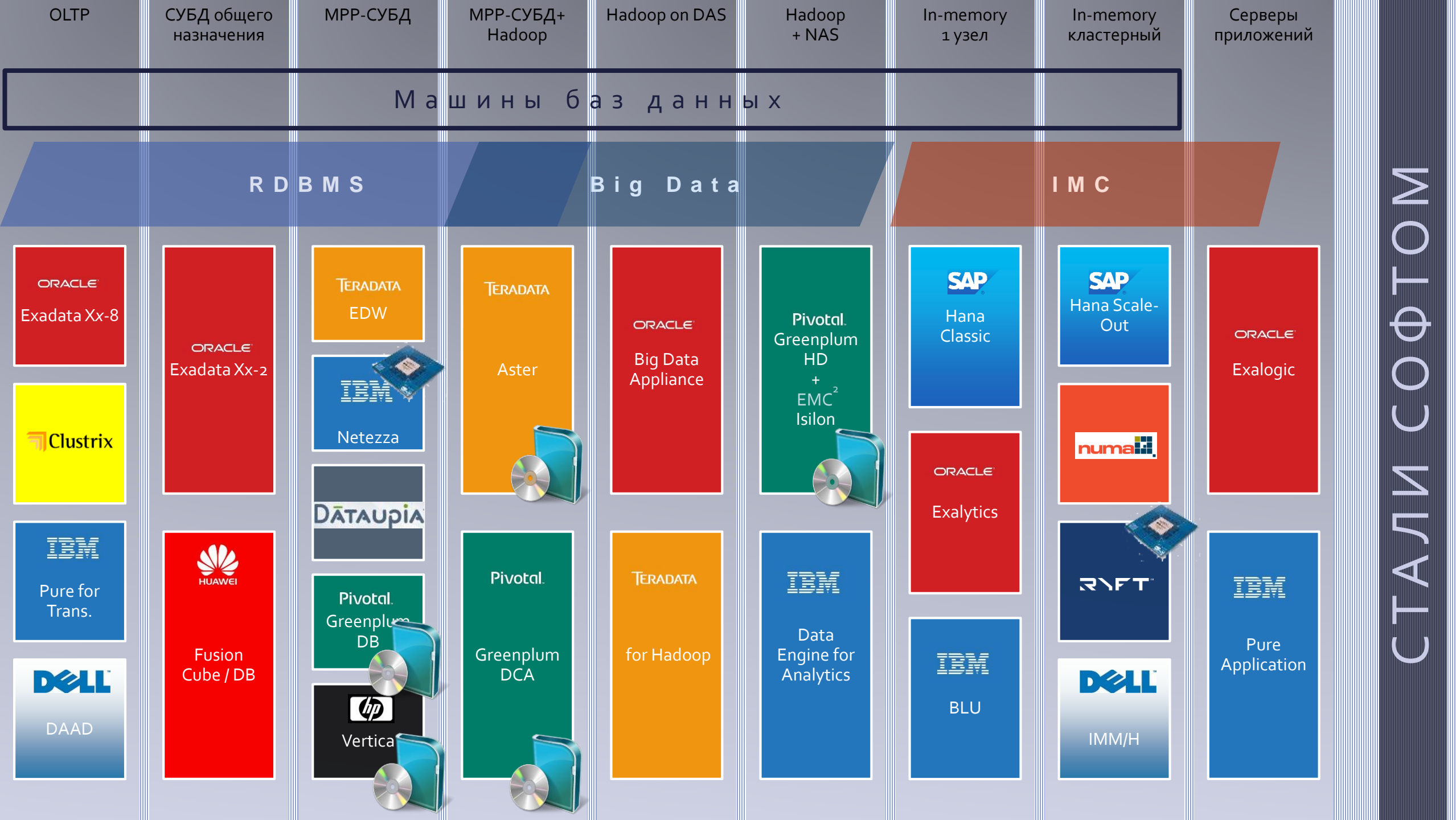
Postgres-XL

Hadapt

Citus DB

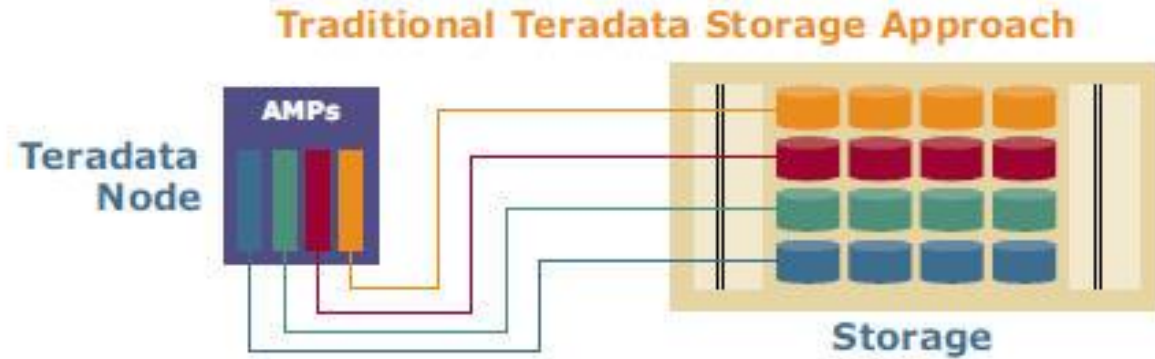


СТАЛИ СОФТОМ

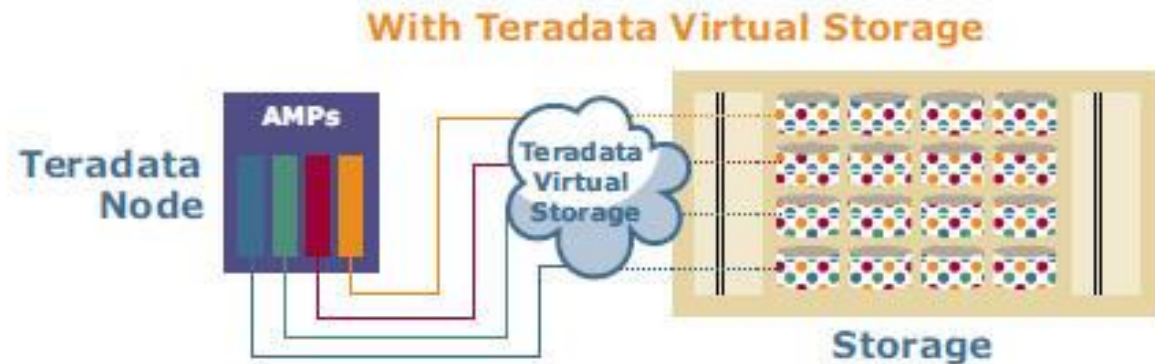


СТАЛИ СОФТОМ

TERADATA: II ПЕРЕДЕЛ (2009)

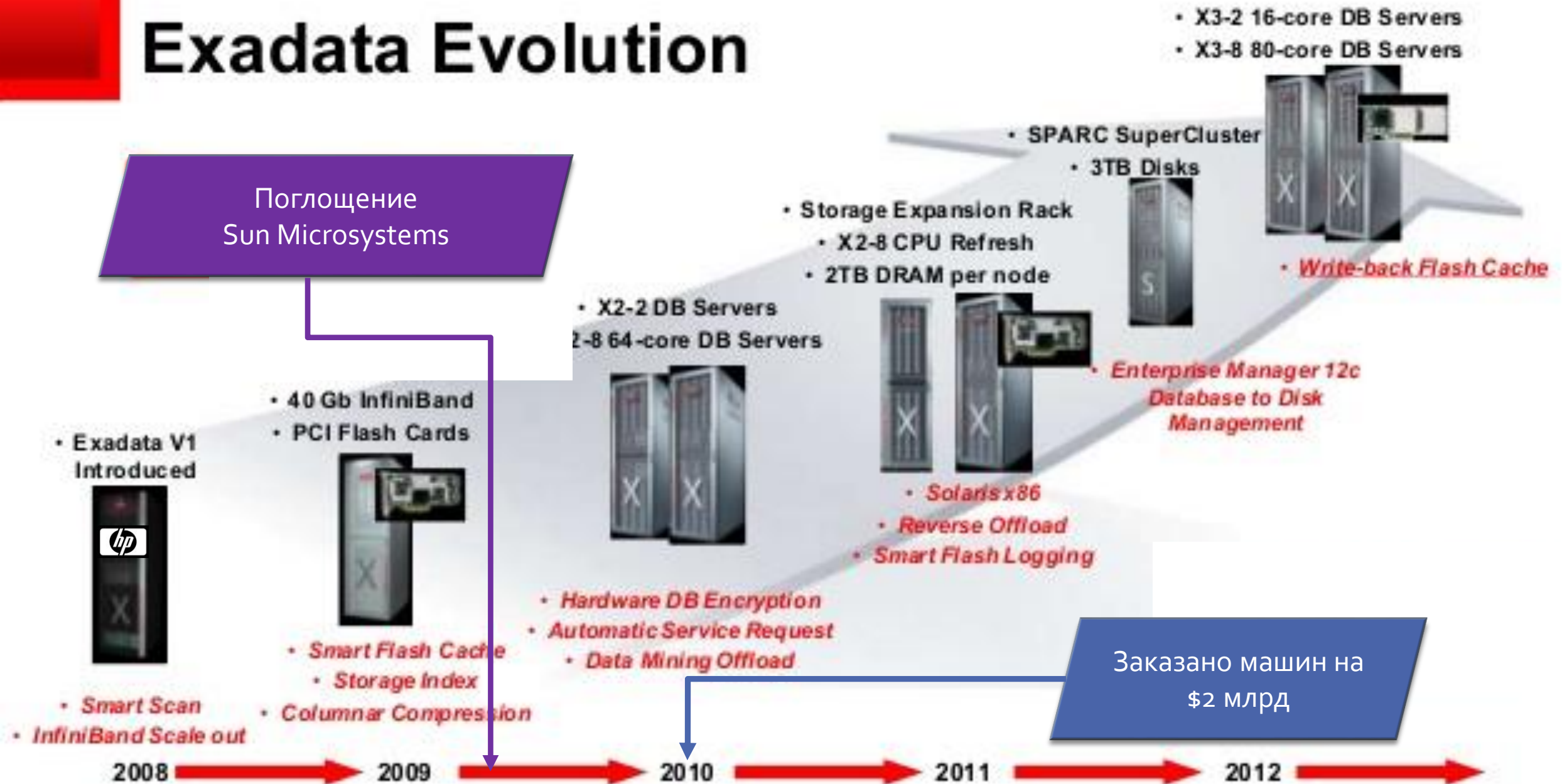


Каждый AMP подключался напрямую к своей дисковой группе [(Drive#,...)];
одинаковое число одинаковых накопителей на AMP



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

Exadata Evolution



ЦЕНЫ И МАСШТАБ В DWH

1990-е

\$1млн / ТБ

- [Teradata]

1 ТБ

- 1995, Winter

2000-е

\$200 тыс.

- [Netezza]

110 ТБ

- 2005, Winter

2010-е

Десятки тыс.

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

12 ПБ

- 2014, SAP

РЕКЛАМНЫЙ ШУМ КАК ЗЕРКАЛО ЭВОЛЮЦИИ

Softbank Runs 2x–8x Faster on Exadata
36 Teradata Racks Replaced by 3 Exadata Racks

Teradata
36 Racks



Exadata
3 Racks



2017?
[300 ТБ]

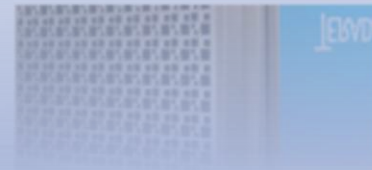
2009

2004

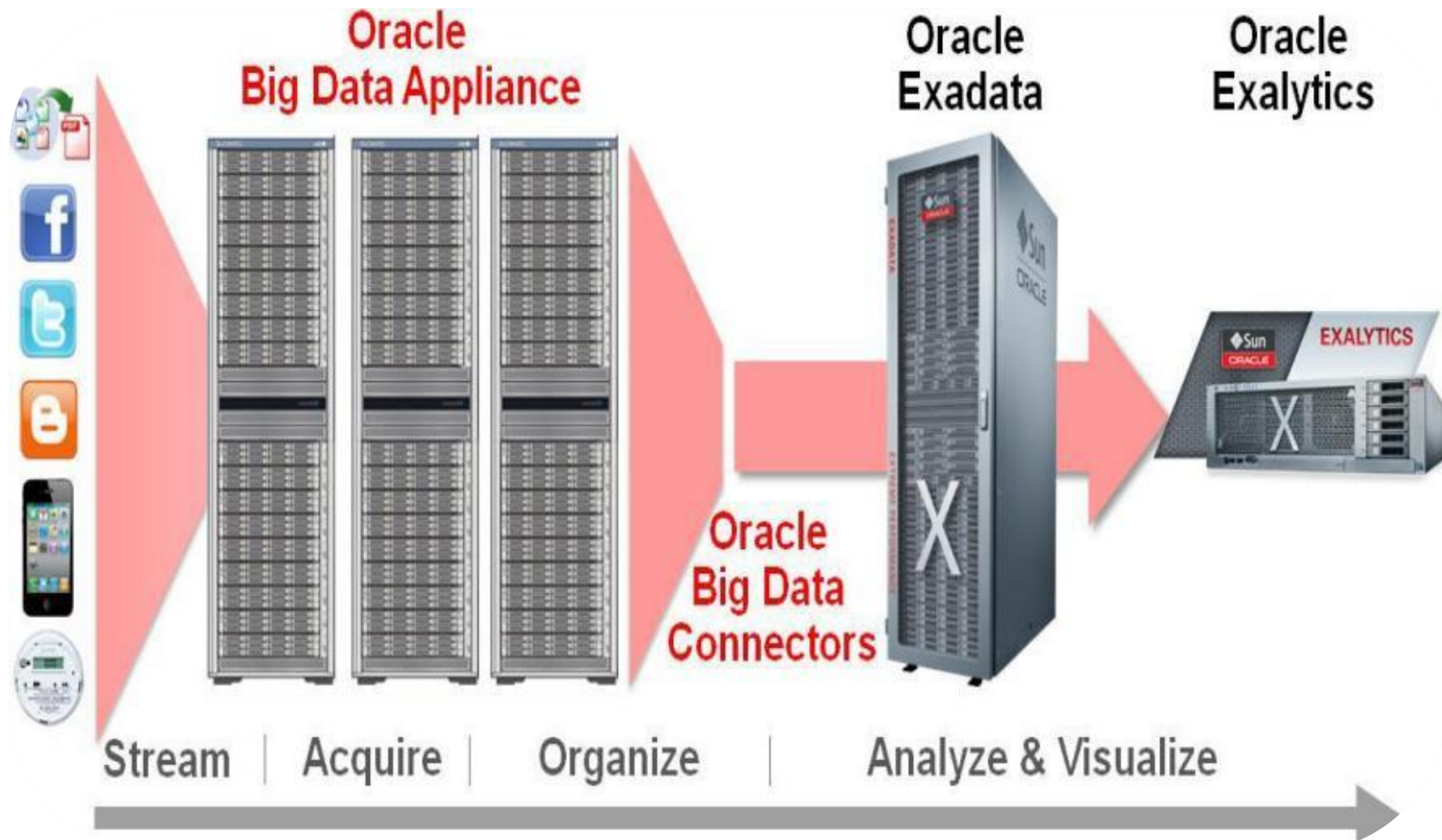
ORACLE

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

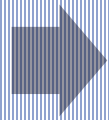


УКЛАДКА В ЛИНЕЙКУ ОТ ORACLE



МУЛЬТИАРЕНДНОСТЬ В СУБД

Сценарии
консолидации



Изоляция
приложений

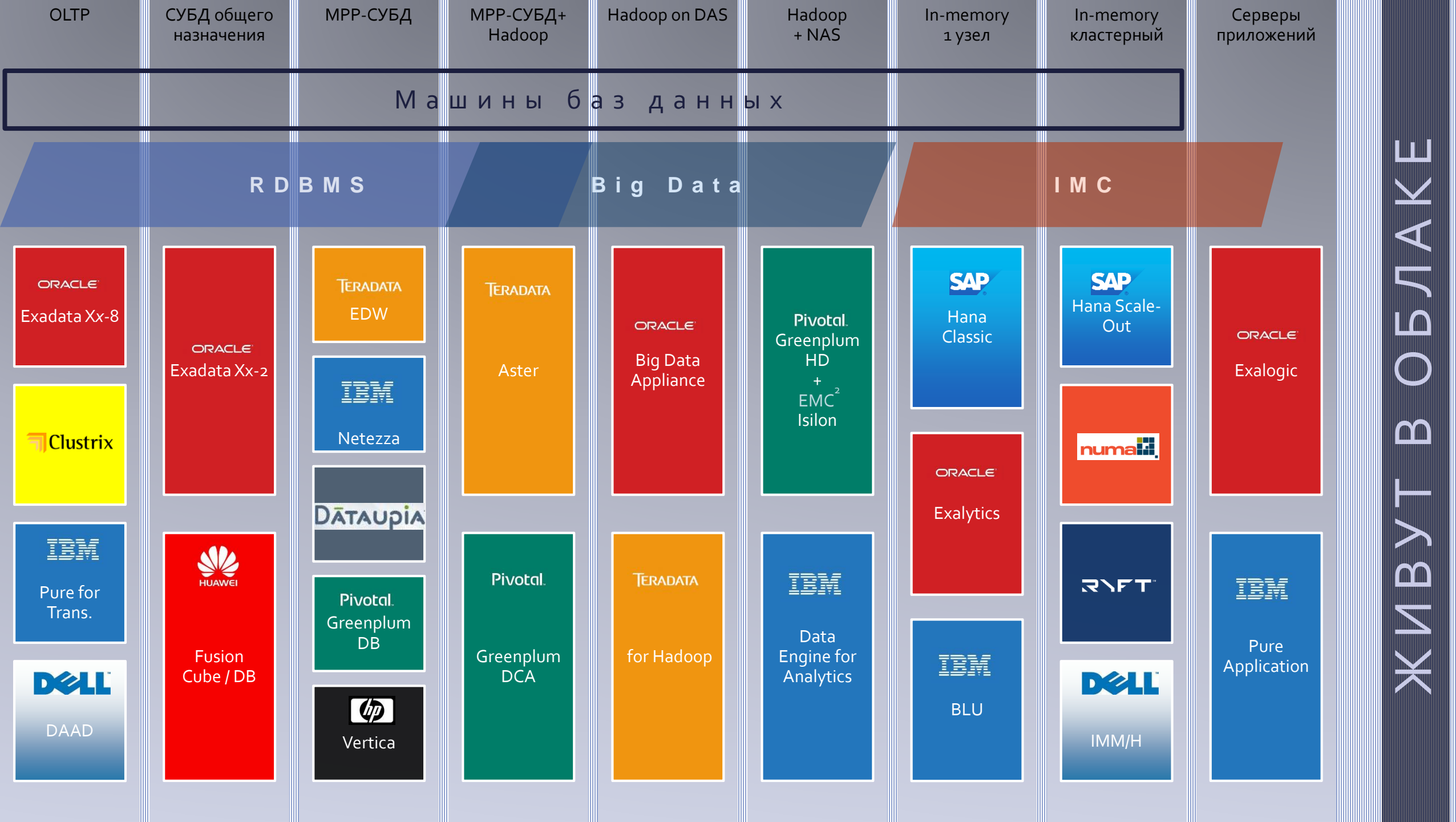


Множество сред
в одном
комплексе

- PROD, TEST, DEV...



DBaaS



ЖИВУТ В ОБЛАКЕ



ЖИВУТ В ОБЛАКЕ

ORACLE CLOUD MACHINE



То же оборудование, но в собственности Oracle

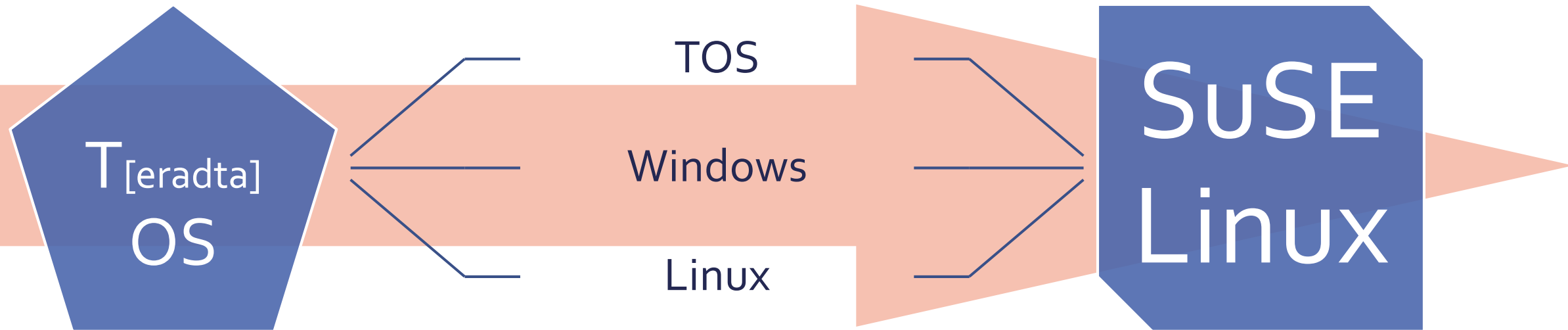
То же ПО, но в собственности Oracle

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 TERA DATA

Банки

Большие
розничные сети

Большие
телеком-
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- фигурированное

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 СОБСТВЕННОРУЧНО...

Нелегально, но ...

Доступно для
скачивания

Oracle
Database
Exadata
Storage
Software

Сообщения
о запуске
на AWS

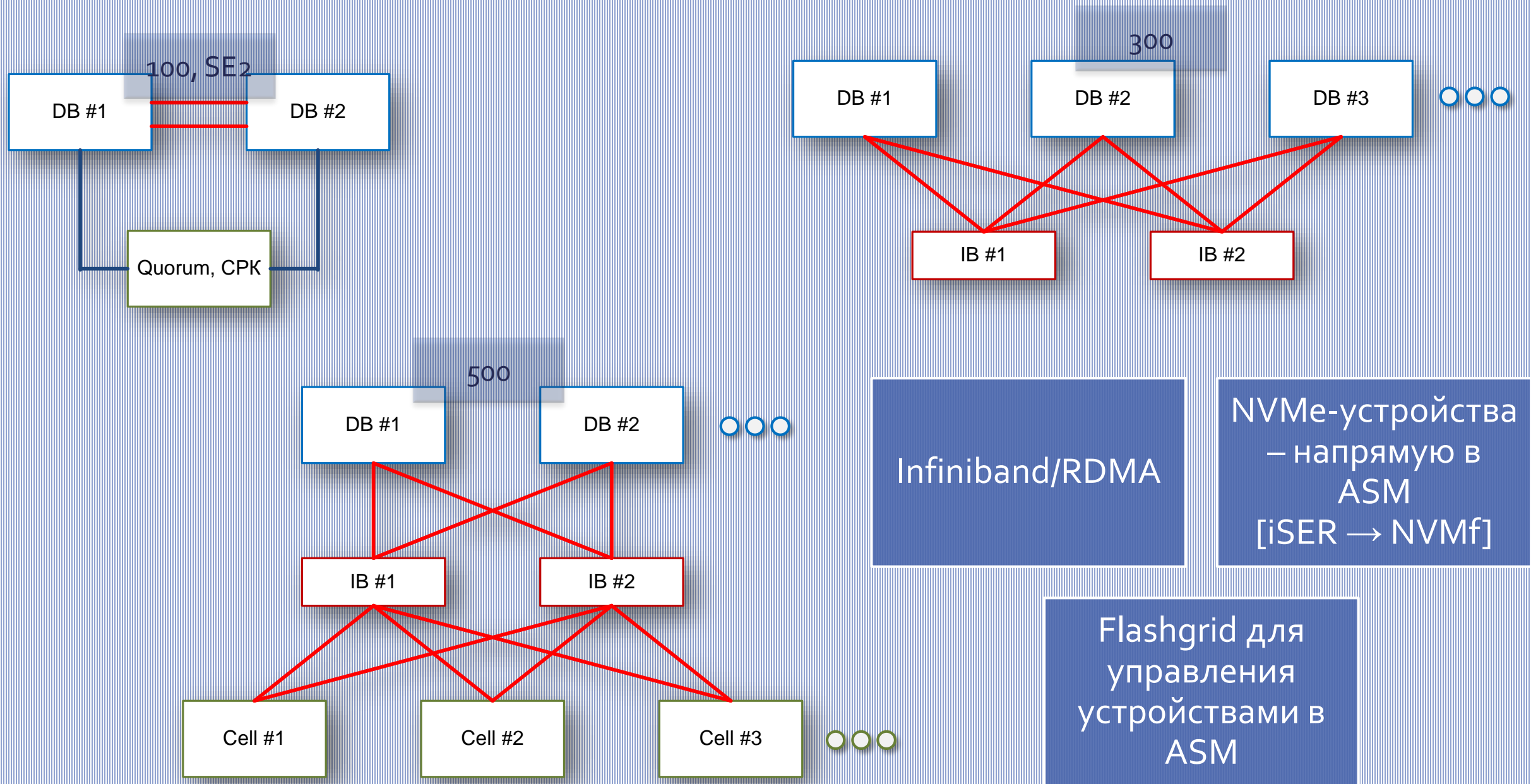
Сообщения
о запуске с
Infiniband-
сетью

В целом –
успешные

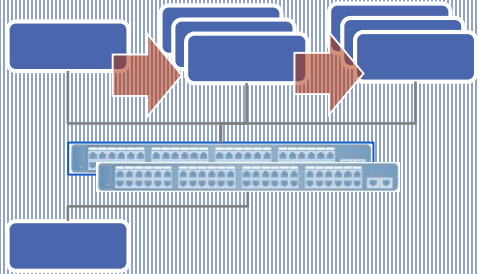
...с
особенностями

...и уже не чёрный ящик

СКАЛА-СР / ORACLE DB



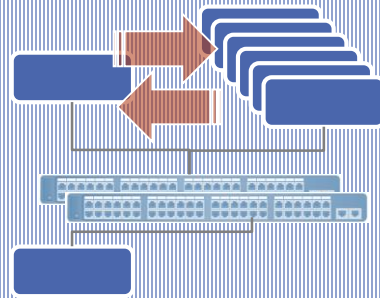
СКАЛА-СР / POSTGRES PRO: ПОСТГРЕС С RDMA



300

Кластер общего назначения

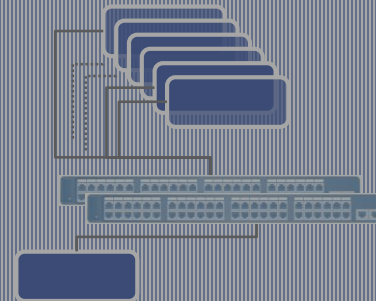
доставка WAL по RDMA
мастер
 n синхронных реплик
 m асинхронных реплик



500

Мультимастер

координация по RDMA



// TODO...

700

ROLAP без разделяемых ресурсов

RDMA
over
Converged
Ethernet



СКАЛА-СР: С МЫСЛЮ 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

Гиперконвергентное ядро



☒ Hyperconverged core

☐ Storage-optimized node

☐ NVMe-storage node

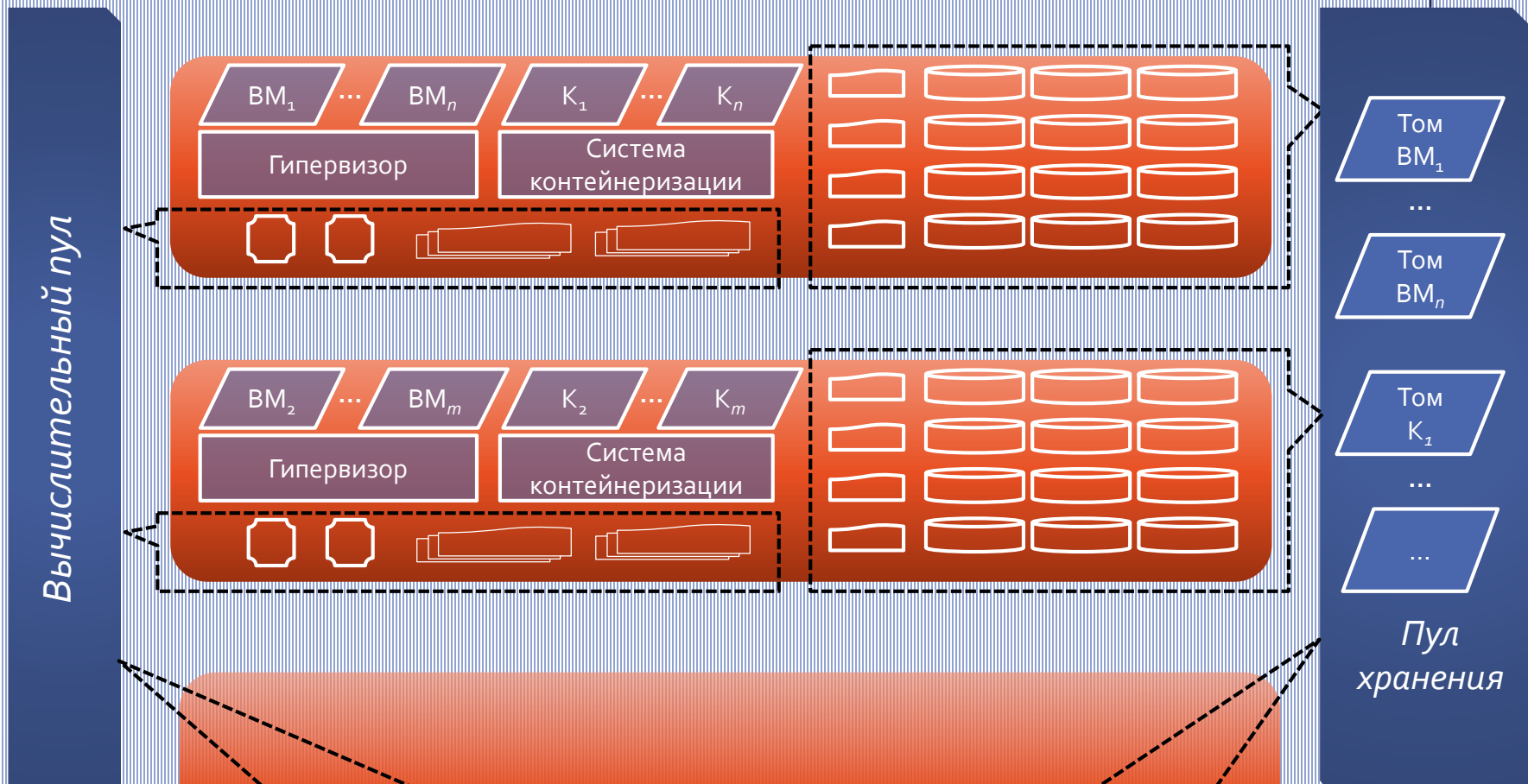
☐ Database-storage node

☐ Compute node

☐ ...

ГИПЕРКОНВЕРГЕНТНАЯ ИНФРАСТРУКТУРА

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



«Серебряная пуля»

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

НСІ: ПЛАТА ЗА УНИВЕРСАЛЬНОСТЬ

Фиксированный (большой)
размер блока в SDS

- Выгодно в условиях для множества виртуальных машин со «средним» вводом-выводом
- Пределы: ~50 kiops с хоста виртуализации

Oracle Database

- Собственный менеджер том, работающий с блочными устройствами
- На практике: пробрасывают блочные устройства в виртуальные машины

PostgreSQL

- POSIX-ортодоксальная система (отработка `fsync()` и т. п.)
- Требуются оптимизации кода SDS для удовлетворительной работы даже для средних нагрузок

Hadoop, ElasticSearch,
Cassandra, ScyllaDB

- JBOD-aware-дизайн с межузловым параллелизмом: гиперконвергентная подоснова избыточна

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

Граничные вычисления

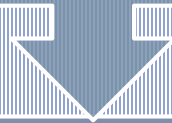
Потоковая обработка

Брокеры сообщений

Мониторы транзакций

Тестовые и разработческие среды

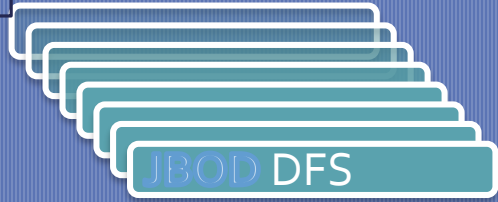
Не требовательны ко вводу-выводу, но многочисленны и изменчивы



Кандидаты к виртуализации в гиперконвергентной инфраструктуре

ВОКРУГ ГИПЕРКОНВЕРГЕНТНОГО ЯДРА

Виртуальные машины
управления, мониторинга,
развёртывания...



Дисковые блоки массового
параллелизма



Флэш-блоки массового
параллелизма



Блоки резидентных
вычислений

Аппаратные блоки с
NVMe-накопителями
и большой ёмкостью ОЗУ



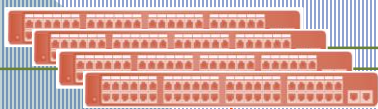
Блоки резидентных
вычислений
с акселерацией

Аппаратные блоки с
SSD-накопителями



Блоки баз данных

хосты виртуализации



НСИ-ядро

коммутаторы
межсоединя (RoCE)

Спасибо за внимание!

anikolaenko@ibs.ru
anikolaenko@acm.org

