

解开exadata的神秘面纱 – flash cache

2020年4月8日上午11:00

Cliff.liu@oracle.com , 甲骨文解决方案构架师

**公益讲座11点准时开始，请大家先浏览云技术微信公众号技术文章
资料会在各群同步发布，已入群客户请勿重复入群！**

扫码加入:

19c新特性讲座群



欢迎关注:

甲骨文云技术公众号





揭开 Exadata 的神秘面纱

关于 **flash** 和 **in-memory**

Cliff.liu@oracle.com

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.

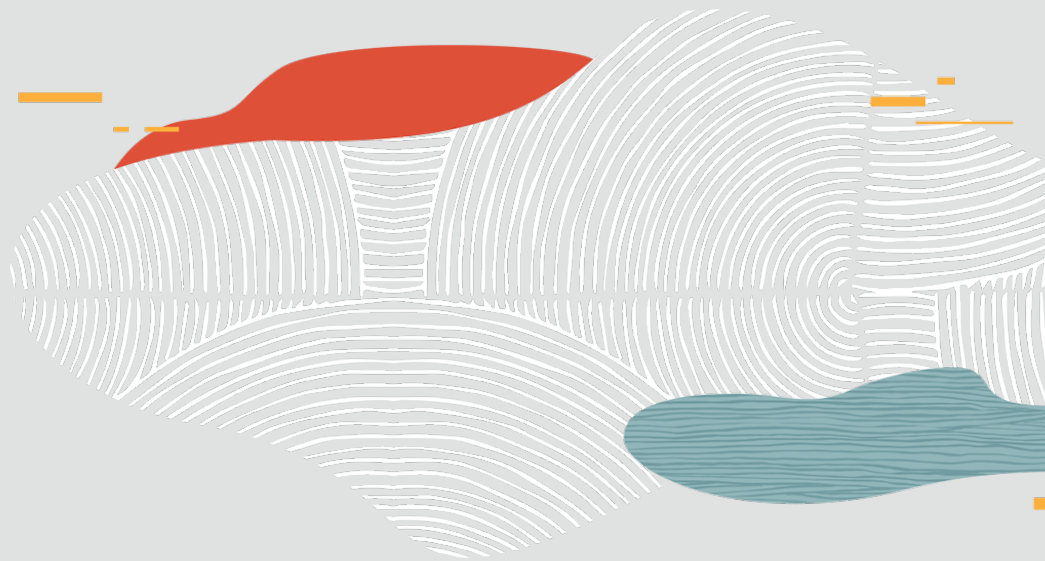
Statements in this presentation relating to Oracle's future plans, expectations, beliefs, intentions and prospects are "forward-looking statements" and are subject to material risks and uncertainties. A detailed discussion of these factors and other risks that affect our business is contained in Oracle's 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.

Exadata: 极致性能简史 flash和in-memory

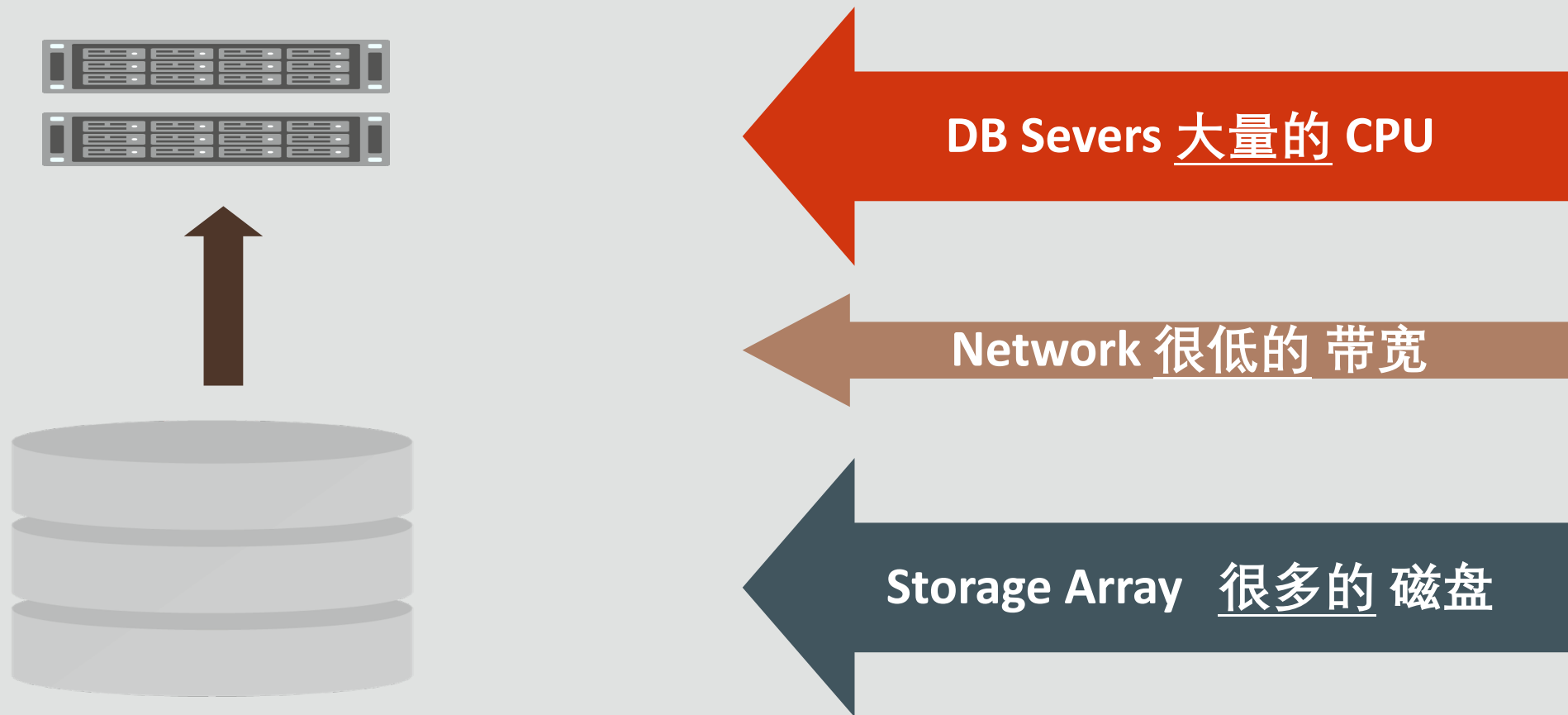




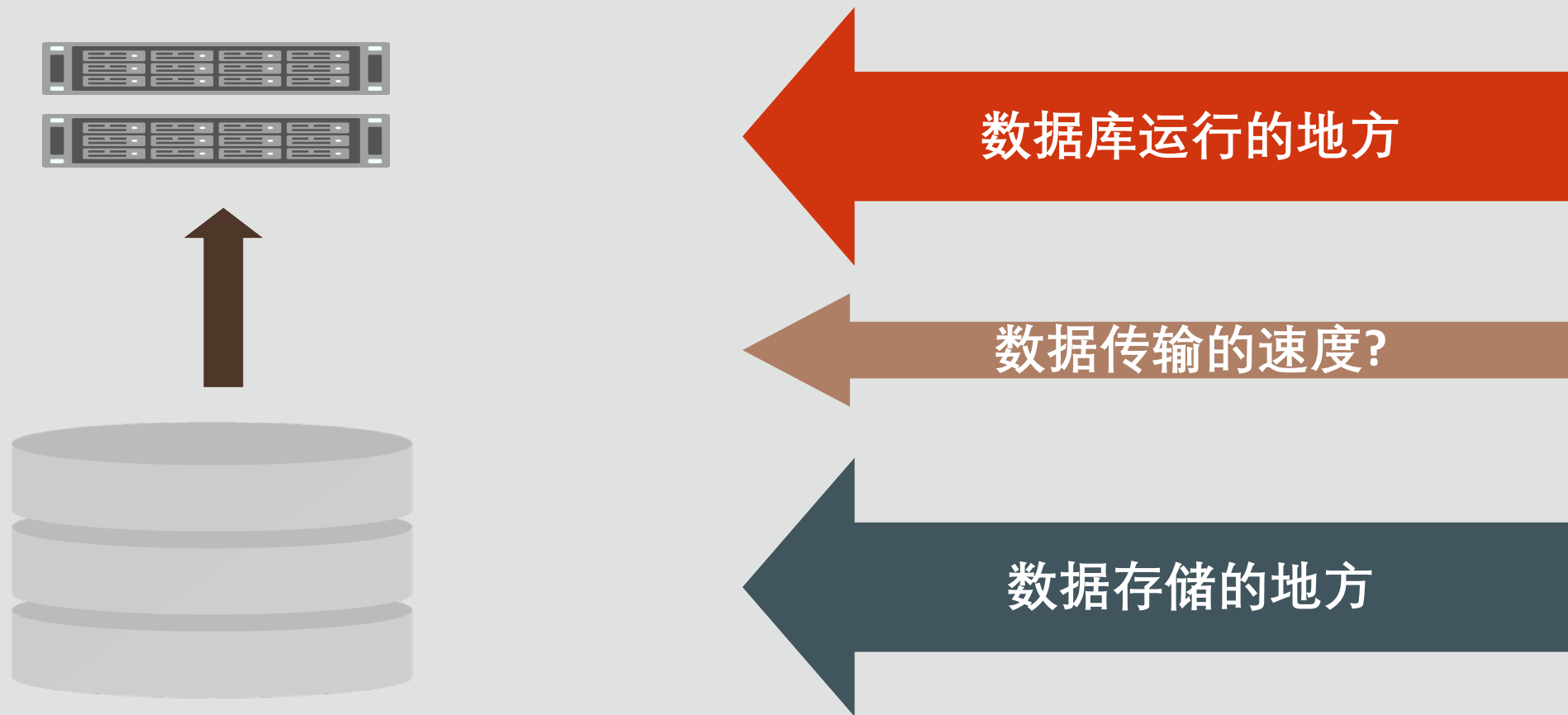
很久很久以前...
那时是在2008年?



数据库系统构架概况 (2008前后)



数据库系统构架概况 (2008前后)



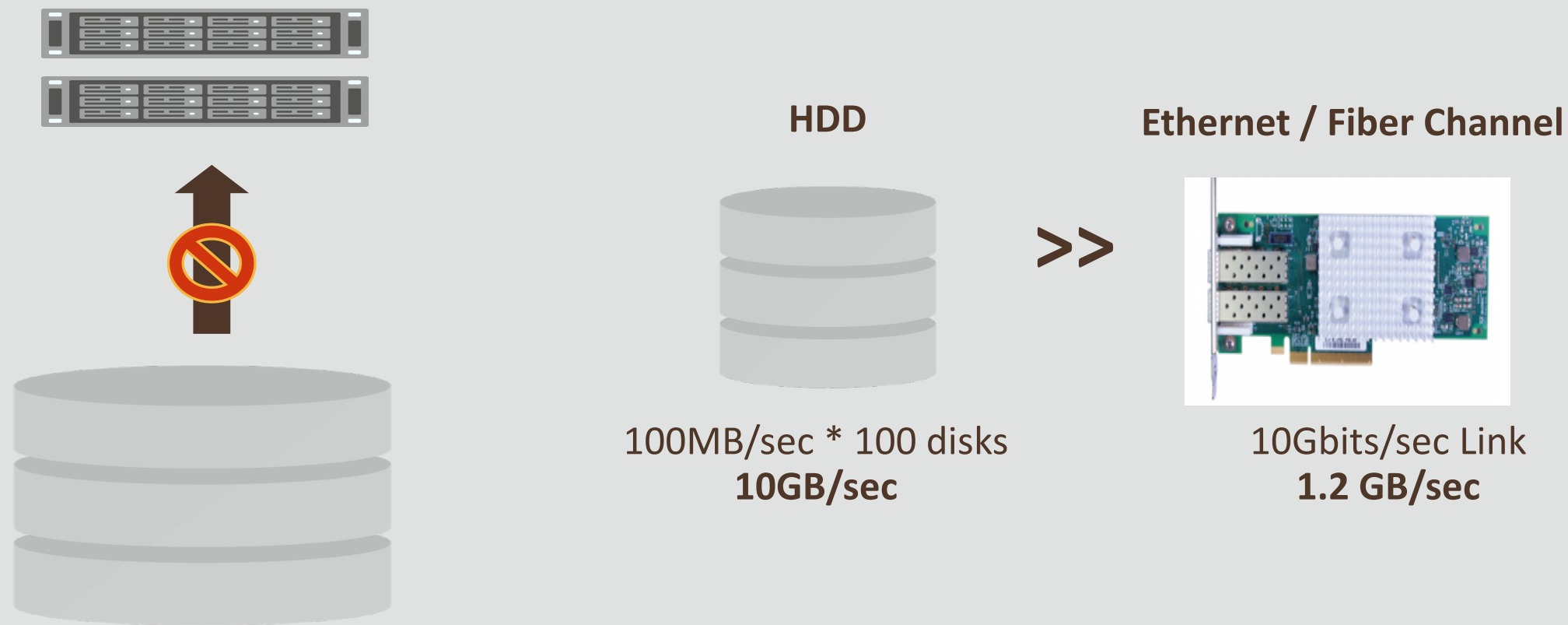
2008年新闻截图

10 Gbits/sec and beyond: High speed in the data center

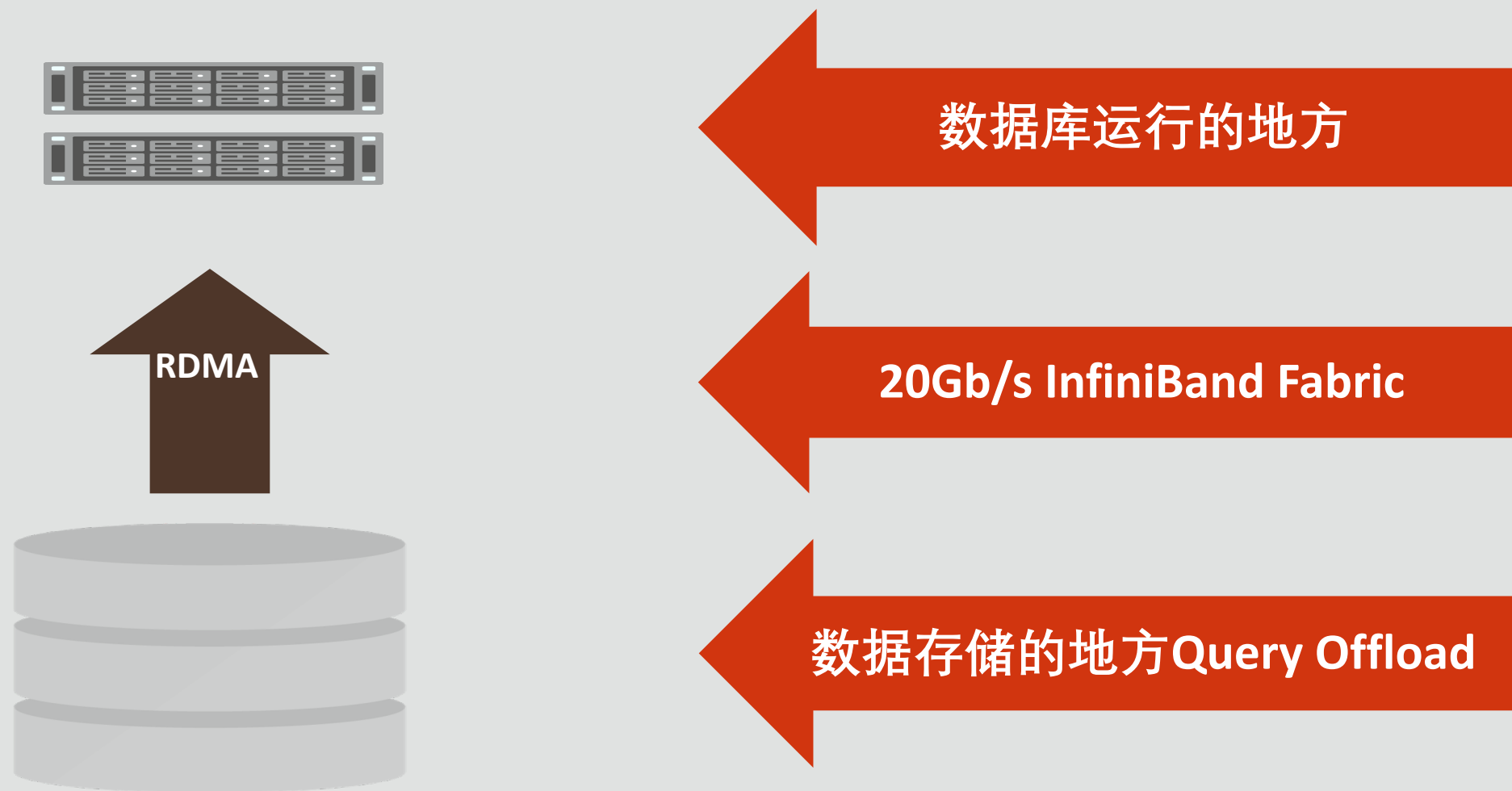
January 1, 2008

Ethernet, InfiniBand, and Fibre Channel all have their own cable and connectivity requirements.

数据库系统瓶颈 (2008年前后)



Exadata Offload 改变了一切 (circa 2008)

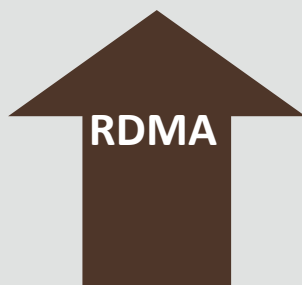


Smart Scan Offload - 智能存储扫描

- 数据库告诉存储系统所需需要的内容
 - 比如：某个省份或城市（filtering）的客户的姓名和地址（projection）
- 存储系统由自己的高性价比 CPUs 对数据块‘过滤和裁剪’
 - 去掉不需要的行{filtering}
 - 去掉不需要的列 {projection}
 - 合成满过滤和裁剪后的数据块并返回给数据库服务器

Database query
processing offloaded
to Storage Servers

Exadata Offload 改变了一切 (2008)



- Database 的功能智能地延伸到 可扩展的 存储服务器
- 存储运行在最佳磁盘速度
- 不再有 网络瓶颈
- RDMA 提供低延迟和低 CPU 消耗

V1



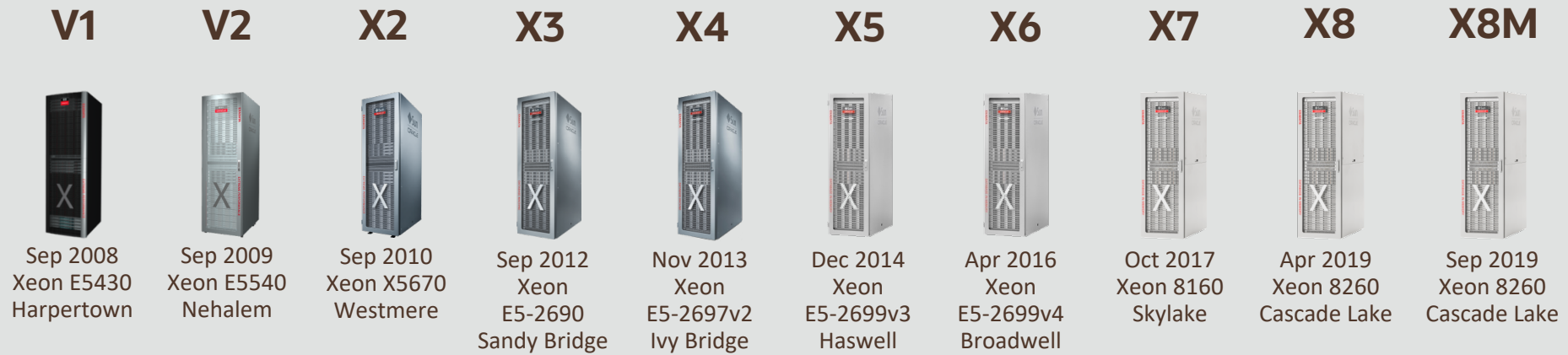
Scan Rate
14GB/sec

Sep 2008

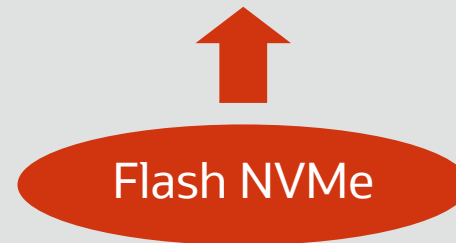


后来的10年？

flash的变革(2009 – 2019)

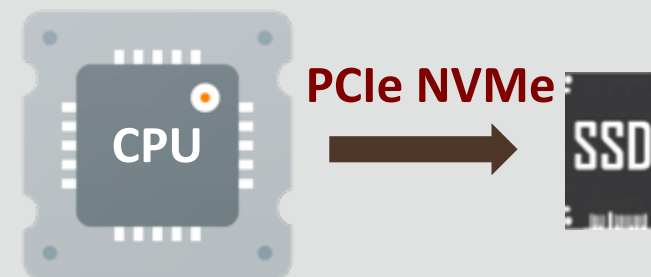


Flash Cache (TB)	0	5.3	5.3	22.4	44.8	89.6	179.2	358	358	358
Scan Rate (GB/s)	14	50	75	100	100	263	301	350	560	560
Read IOPS (M)	.05	1	1.5	1.5	2.66	4.14	5.6	5.97	6.57	16



PCI Express Flash 使用 NVMe 接口

- Non Volatile Memory Express 非易失性内存主机控制器接口是全新设计的flash存储的接口
- NVMe 支持大规模并行访问。
- NVMe 原生支持数据库块大小IO。
- NVMe IO 协议栈大幅缩短，减小系统CPU消耗和延迟



**NVMe is 2.2x
Faster than SCSI^x**

PCI Express Flash with NVMe Interface is the right choice for your Database

Exadata 引领NVMe的使用潮流

从2014开始，成千上万的Exadata装备NVMe Flash



Exadata X5-2
Industry's First
Enterprise System
with NVMe



Exadata Cloud
Service uses NVMe
in Public Cloud

Exadata X6-2
2nd Generation
NVMe with
3D VNAND

Exadata X7
3rd Generation
NVMe

2014

1st NVMe Drive by
Samsung

2015

1st NVMe Drive by
Intel



2016

Facebook launches
Lightning based on
NVMe



EMC Announces
DSSD D5 with
NVMe

2017

EMC Closes DSSD
division

数据库系统瓶颈 (2019年前后)



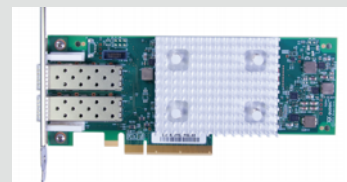
- 一个 NVMe Flash 设备就使网络吞吐饱和
- 其他的Flash设备的吞吐能力白白浪费掉

NVMe Flash



5.8 GB/sec

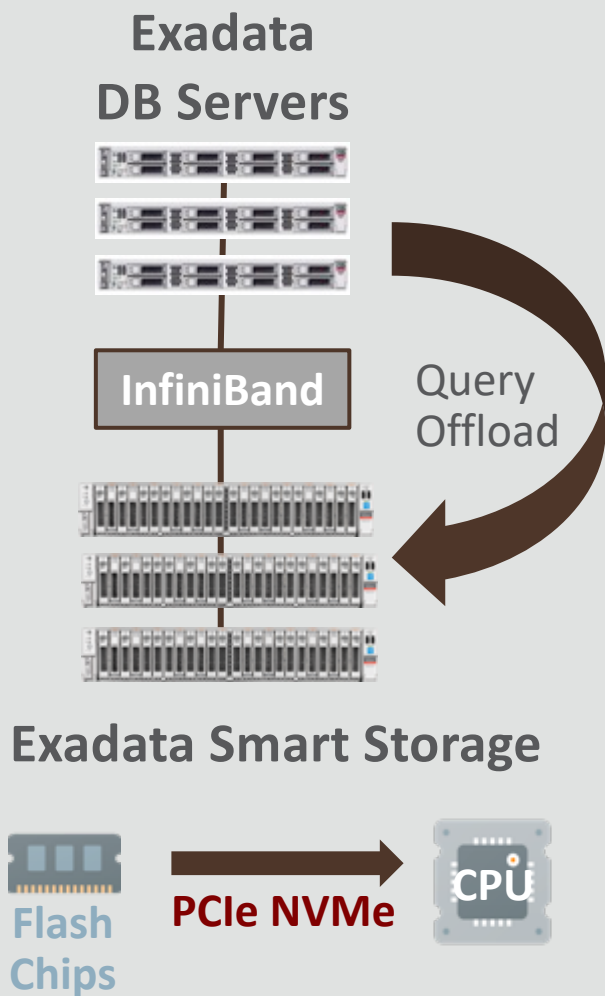
40Gb/s SAN Link



5 GB/sec

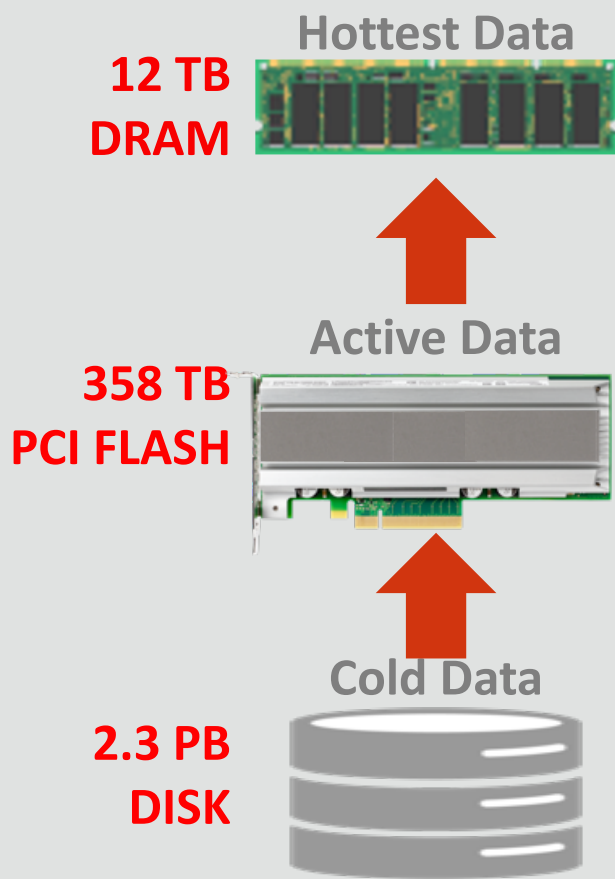
>

Exadata 使用Flash达到内存级的性能



- Exadata X8 拥有高达 **560GB/每秒的flash带宽**到任意数据库服务器
 - 接近 800GB/每秒 数据库服务器 **DRAM 内存整体带宽**
- **必须把计算下移到存储才能发挥全部的Flash IO能力**
 - 需要从数据库到硬件全栈支持，不可能单靠存储解决
- **本质上,传统存储系统只能共享Flash的容量，并不能完全发挥flash的性能**
 - 无论使用下一代的可扩展PCIe 网络或NVMe 光纤接口
- **拥有内存级别带宽性能的共享存储是业界的典范**
 - 拥有DRAM的吞吐能力和Flash的容量

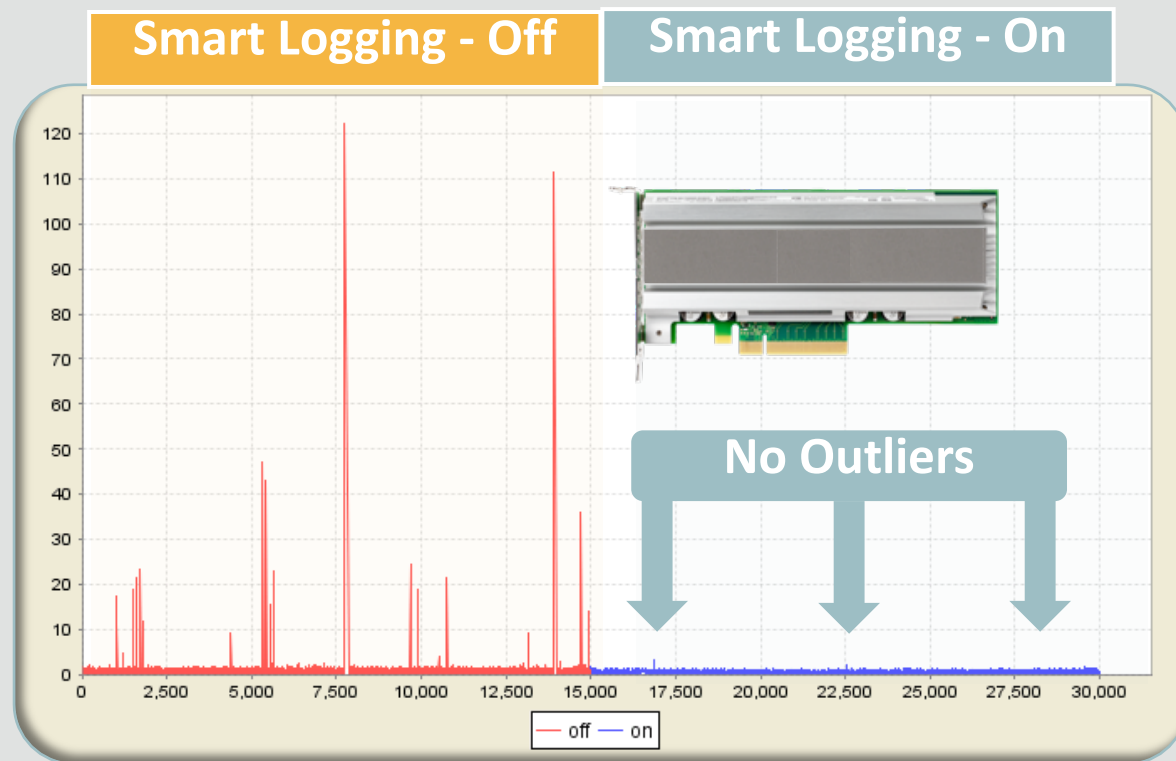
Exadata 智能 Flash Cache



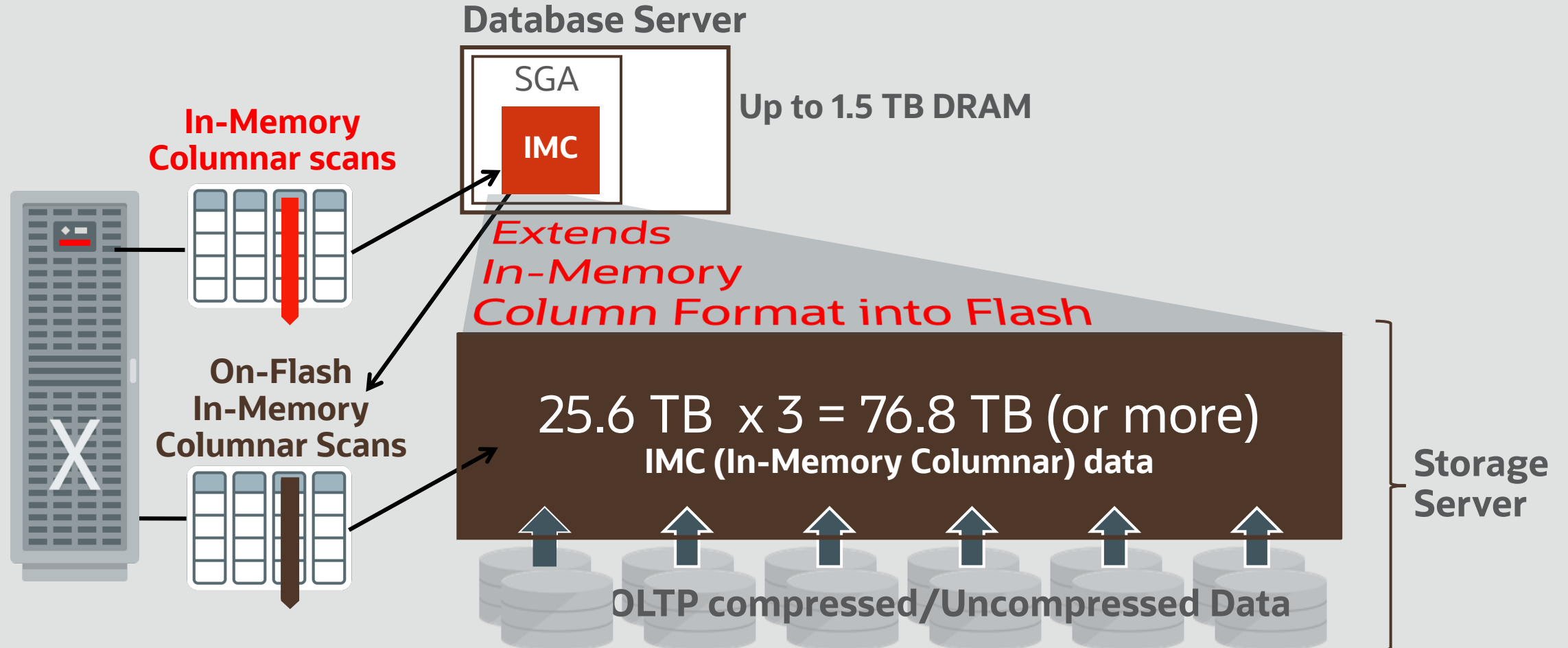
- 懂得不通种类的数据库 I/Os
 - 不会缓存backup I/Os, data pump I/O, archive logs, tablespace formatting
 - 会缓存控制文件读写, Control File Reads and Writes, file headers, data and index blocks
 - 保留更多空间给用户数据
- 立即响应负载类型变化
- Write-back flash cache
 - 不仅仅对读操作提供缓存, 还支持写操作
- 对读操作不需要在flash生成镜像或冗余数据
 - 全闪存阵列的数据冗余镜像存储在Flash中, 增加成本
- 通过智能扫描可以达到以所有Flash Driver的带宽传输数据
 - 对于SQL查询, 虽然全闪存阵列拥有大量的服务器和处理器, 但是还是不能达到智能扫描的带宽。
 - 以普通磁盘的成本提供高于flash的性能。

Exadata Smart Flash Log

- Smart Flash Log并行使用flash缓存和磁盘控制器写入
- 无论flash还是磁盘，任意一个写成功就算日志写入完成。
- 减少响应时间和异常峰值毛刺
 - “log file parallel write” 性能提升
 - 大幅改善“log file sync”
- 占用非常少的flash(< 0.1%)
- 网络资源管理提供高优先级给 redo log I/Os
- OLTP负载的系统性能极大加速并且响应时间得到保证



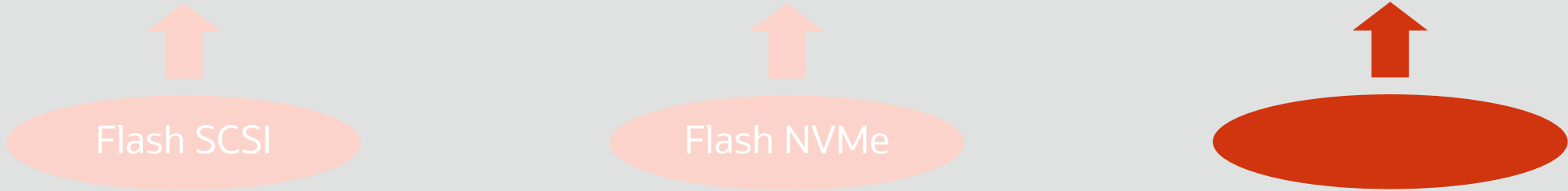
Exadata In-Memory Columnar Flash Cache



Flash带来的革命



Flash Cache (TB)	0	5.3	5.3	22.4	44.8	89.6	179.2	358	358	358
Scan Rate (GB/s)	14	50	75	100	100	263	301	350	560	560
Read IOPS (M)	.05	1	1.5	1.5	2.66	4.14	<u>5.6</u>	<u>5.97</u>	<u>6.57</u>	16



解开exadata的神秘面纱 – flash cache

2020年4月8日上午11:00

Cliff.liu@oracle.com , 甲骨文解决方案构架师

**公益讲座11点准时开始，请大家先浏览云技术微信公众号技术文章
资料会在各群同步发布，已入群客户请勿重复入群！**

扫码加入:

19c新特性讲座群



欢迎关注:

甲骨文云技术公众号





ORACLE