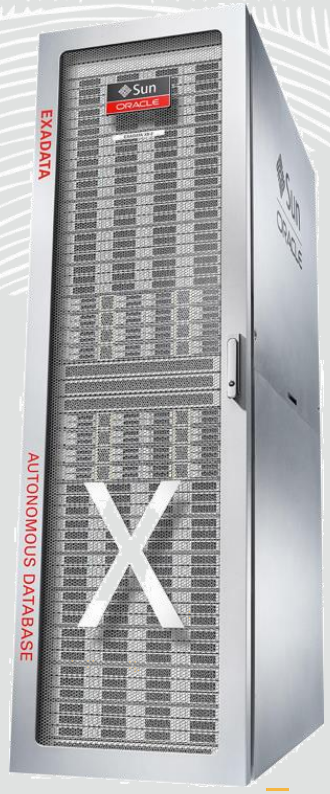
ORACLE

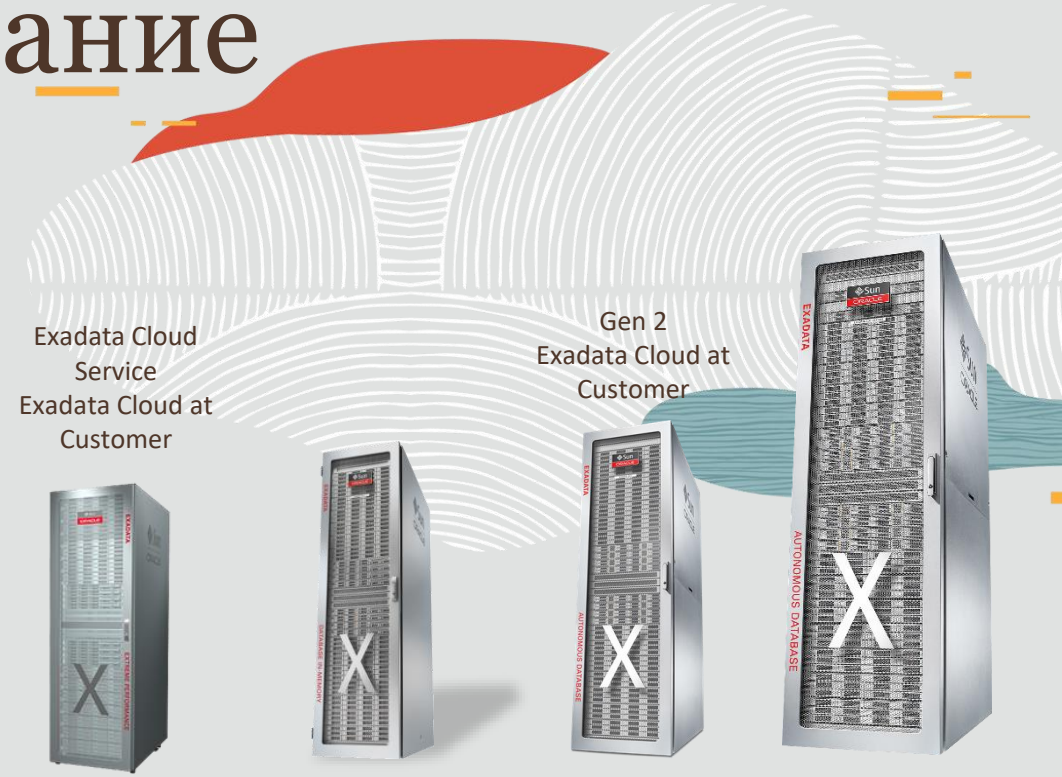
Как научить Machine Learning летать с Oracle Exadata



30 октября 2019

Новая фабрика RDMA
Энергонезависимая память

Oracle Exadata – наращивание технологий



Exadata V1 (2008)
- Oracle Database Machine
- DDR Infiniband
- Smart Scan

Exadata V2 (2009)
- Sun Oracle Database Machine
- + QDR Infiniband
- + Flash Cache
- + Hybrid Columnar Compression

Exadata X2 (2010)

Exadata X3 (2011)
- Ускорение задач OLTP, Аналитики, Консолидации
- + Flash Cache Write-Back

Exadata X4 (2012)
- Ускорение задач In-Memory
- + Active/Active Infiniband

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

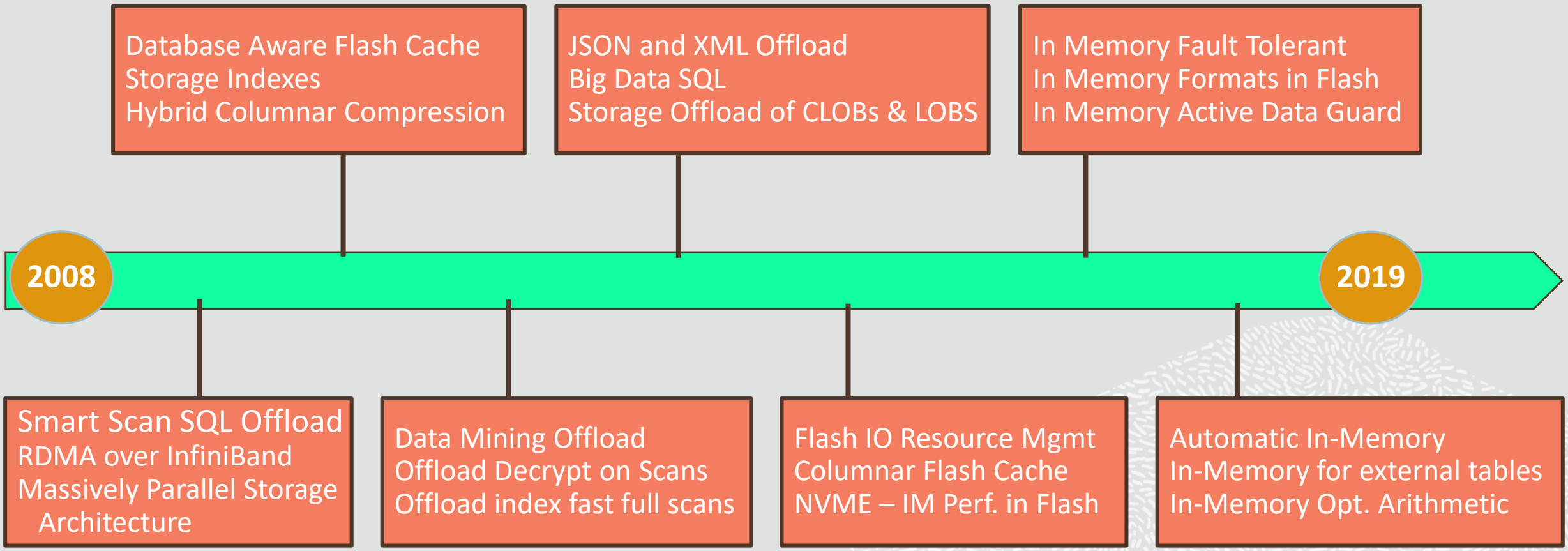
Exadata X6 (2016)
- Exadata Cloud Service
- Exadata Cloud at Customer

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

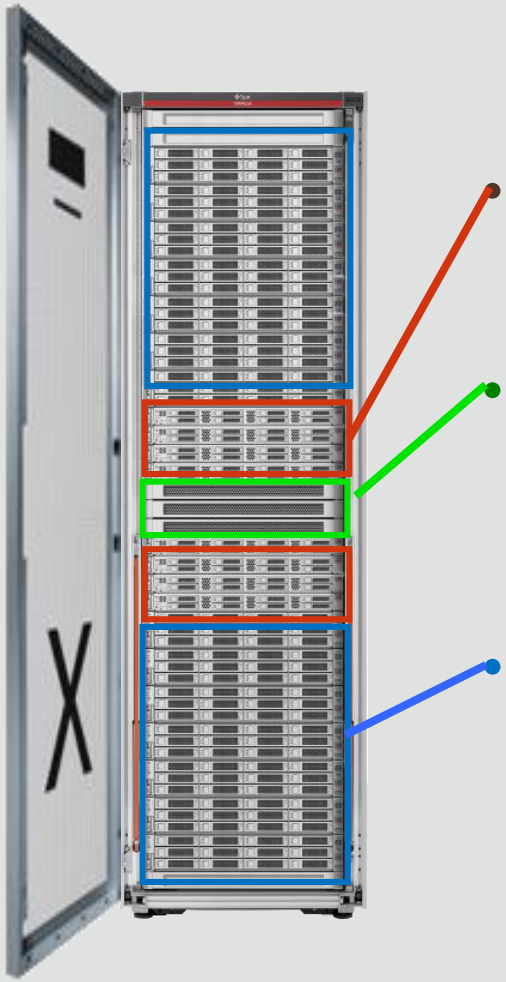
Exadata X8 (2019)
- Gen 2 Exadata Cloud at Customer
- + XT cells
- + Automatic Indexing
- + 60% Performance Boost

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

Exadata— уникальные преимущества для аналитики



Что нового в **Exadata X8M?**



Серверы БД: Гипервизор KVM

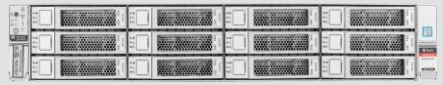
интерконнект 100 Gb/sec RDMA over Converged Ethernet RoCE

В ячейках появилась энергонезависимая память на базе Intel DC Optane

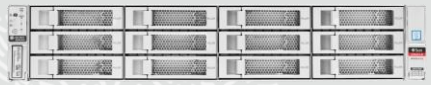
- 1.5 TB PMEM в каждой ячейке

Database Server

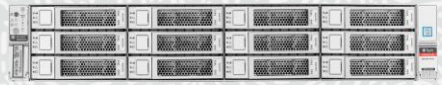
High-Capacity (HC) Storage



Extreme Flash (EF) Storage

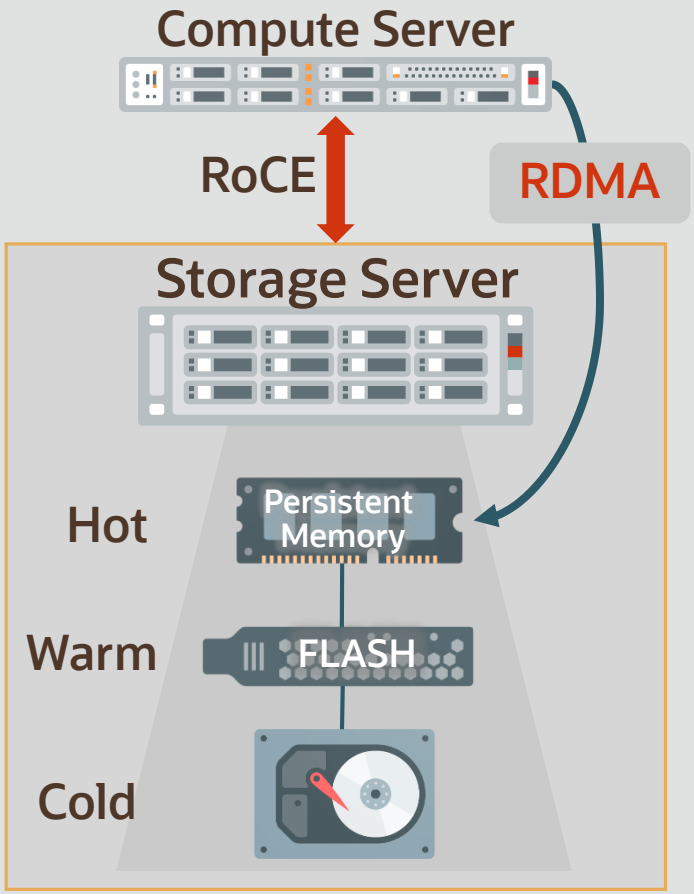


Extended (XT) Storage



Exadata X8M: RoCE + PMEM

Каков результат применения новых технологий?



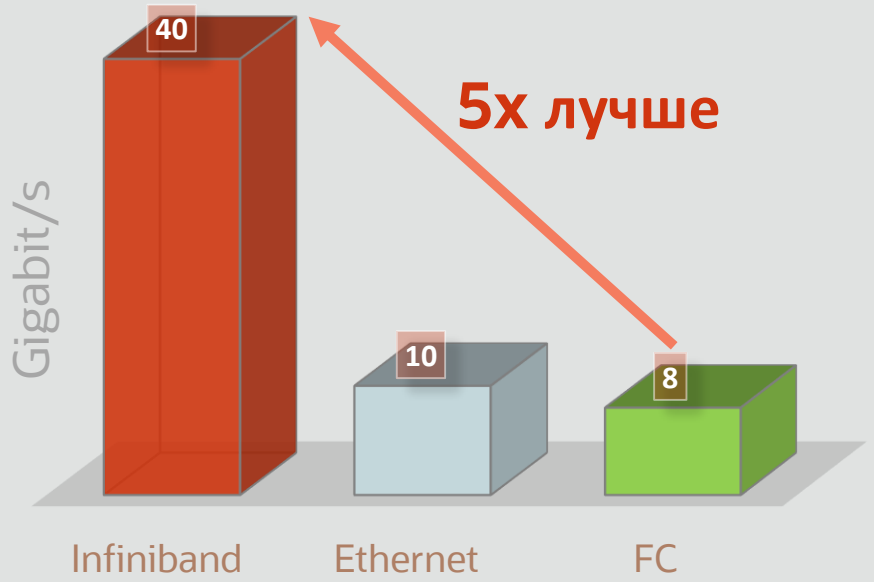
- Новый рекорд по операциям I/O
  - 16 миллионов IOPS – в 2,5 раза лучше показателей предыдущей Exadata X8
- Радикальное улучшение латентности
  - СУБД работает с PMEM по RDMA напрямую минуя системный стек ОС, прерывания
  - <19 µsec латентность на чтении БД блоками 8K в – 10 раз лучше
- Сверхбыстрая запись журнала транзакций
- Новые технологии Exadata:
  - Exadata PMEM cache
  - Exadata PMEM Log

Требуется Exadata System Software 19.3 и Database Software 19c

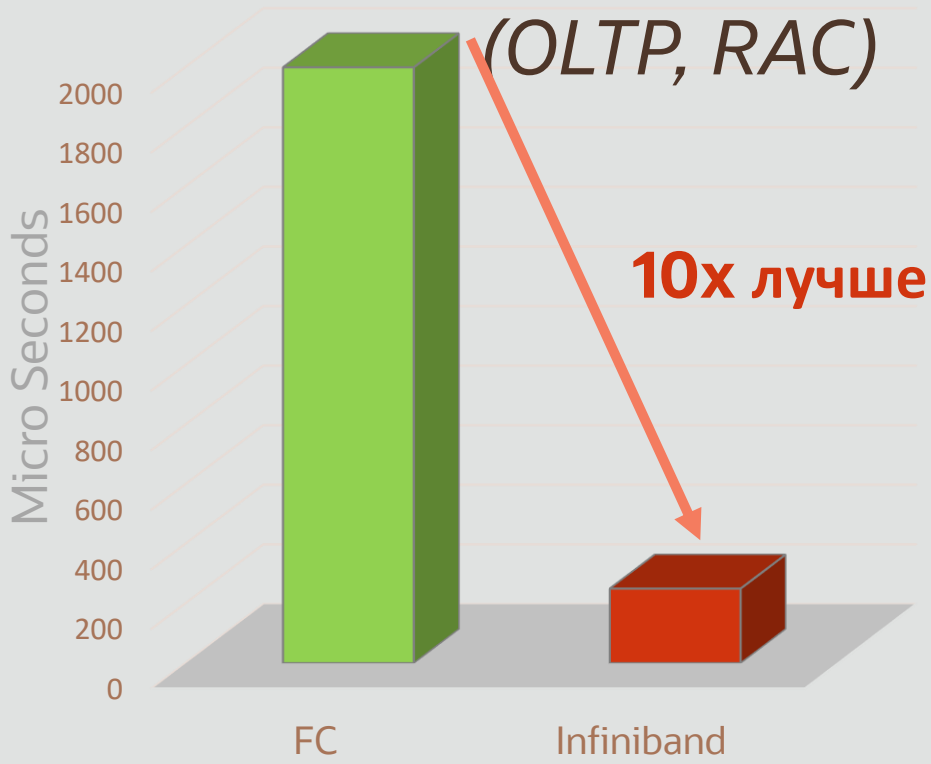
Почему был выбран Infiniband?

вспомним 2008-2009гг

Пропускная способность
(Аналитика)

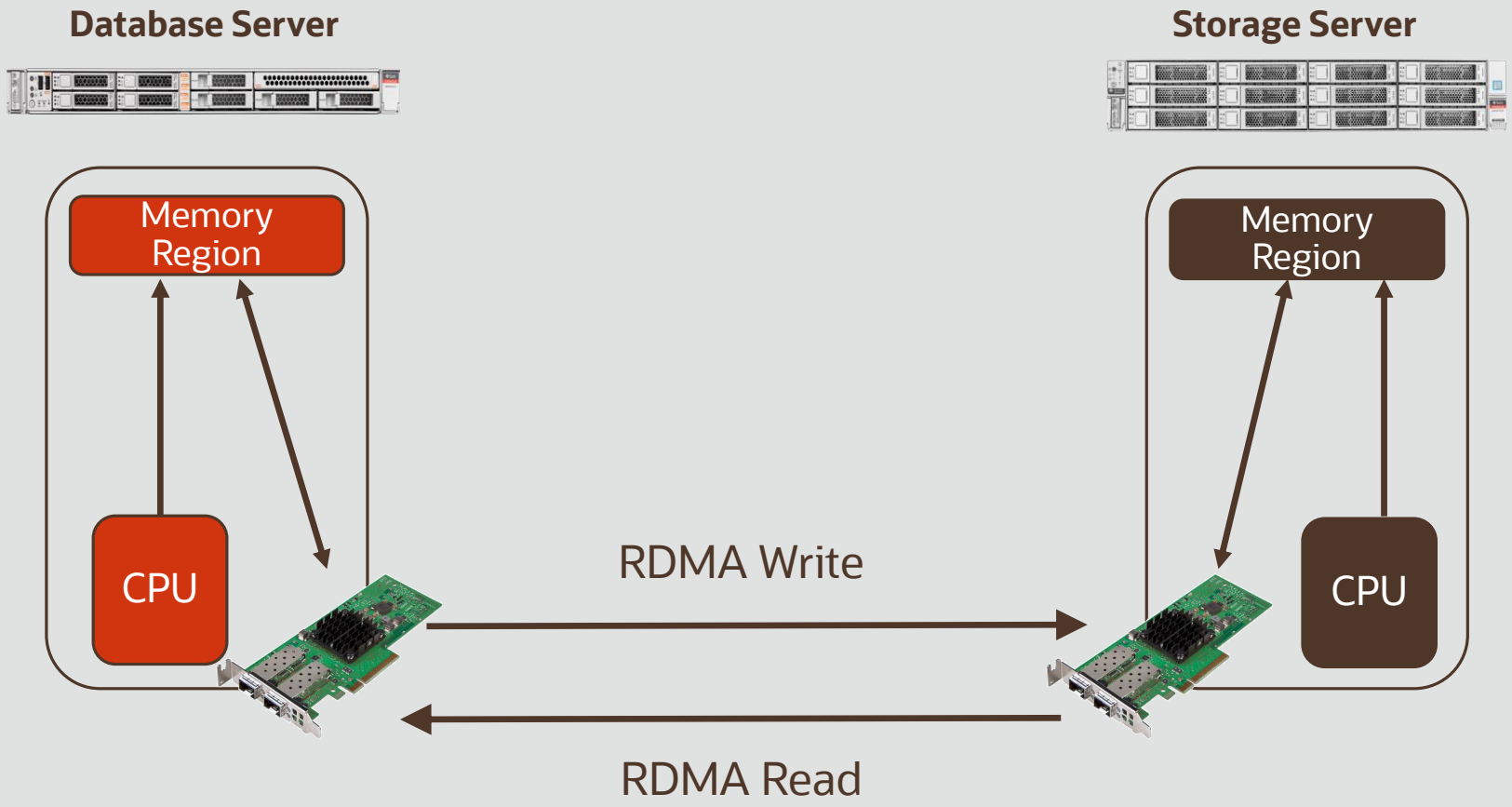


Латентность
(OLTP, RAC)



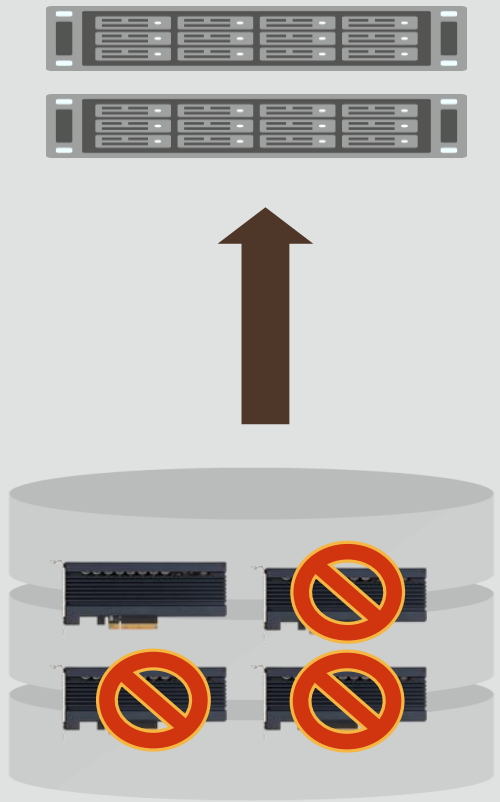
Почему был выбран Infiniband?

сверхнизкая латентность – заслуга RDMA (Remote Direct Memory Access)




Если Infiniband
так хорош,
зачем же его
заменять?

Сегодня флэш создаёт ГИГАНТСКОЕ узкое место



- Один NVMe-носитель насыщает канал 40 Gbit
- Пропускная способность остальных носителей теряется

NVMe Flash

5.8 GB/sec

>

40Gb/s Link

5 GB/sec

Как удалось добиться прорыва в Exadata X8M?

V1	V2	X2	X3	X4	X5	X6	X7	X8	X8M
									
Sep 2008 Xeon E5430 Harpertown	Sep 2009 Xeon E5540 Nehalem	Sep 2010 Xeon X5670 Westmere	Sep 2012 Xeon E5-2690 Sandy Bridge	Nov 2013 Xeon E5-2697v2 Ivy Bridge	Dec 2014 Xeon E5-2699v3 Haswell	Apr 2016 Xeon E5-2699v4 Broadwell	Oct 2017 Xeon 8160 Skylake	Apr 2019 Xeon 8260 Cascade Lake	Sep 2019 Xeon 8260 Cascade Lake
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**

Flash SCSI (SSD)

Flash NVMe

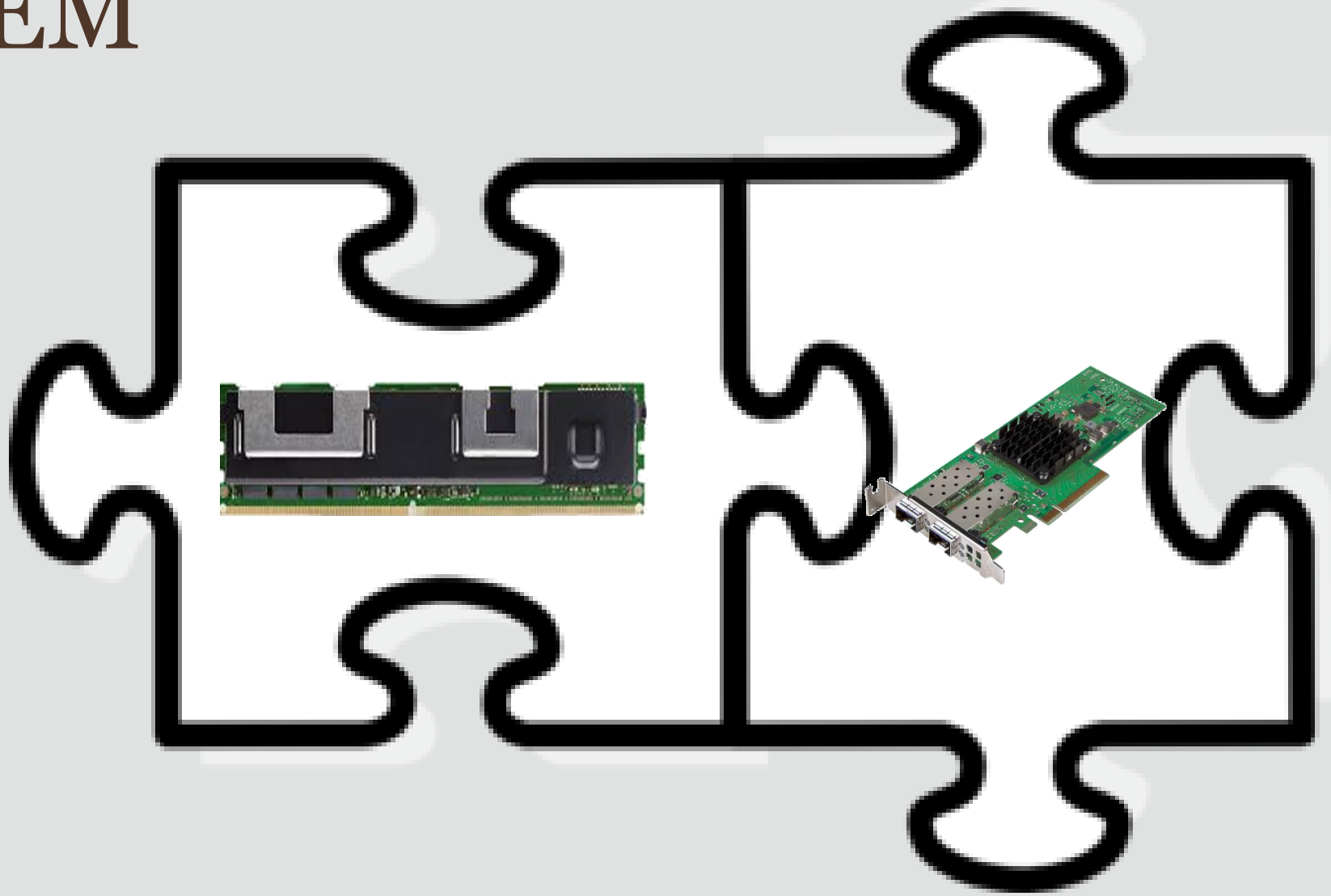
КАК???

100 GbE Ethernet с RoCE – замена Infiniband

- RoCE = RDMA over Converged Ethernet – протокол с InfiniBand RDMA API на основе Ethernet
  - то же ПО на верхнем уровне сетевого стека
  - сочетает скорость Infiniband и гибкость IP
- Сохраняются все оптимизации Exadata на основе RDMA
- Стандарт разработан InfiniBand Trade Association (IBTA)
  - Разрабатывается как Open Source
  - Широко поддержан производителями сетевого оборудования

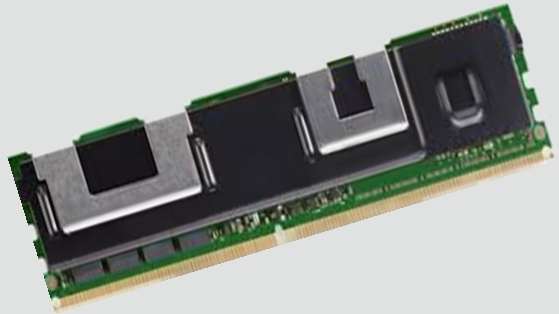
Layer	RoCE	InfiniBand
Application	User Application	
	Transport (InfiniBand)	
Network	IP Network	InfiniBand Network
Hardware	Ethernet	InfiniBand

Перейдём к второй части пазла - PMEM

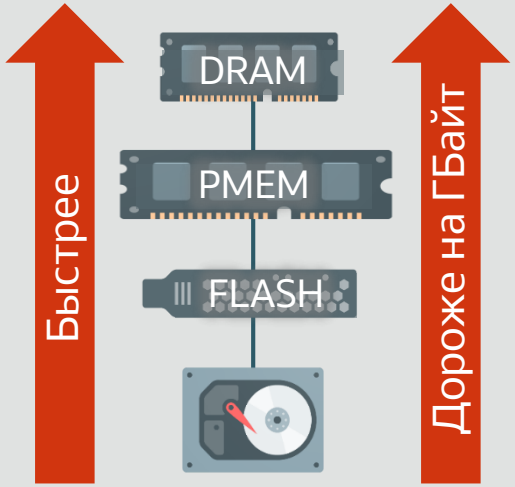


Энергонезависимая память

PMEM = Persistent MEMory



- Энергонезависимая память
  - По ёмкости и цене между динамической памятью и флэш
  - По производительности близка к динамической памяти

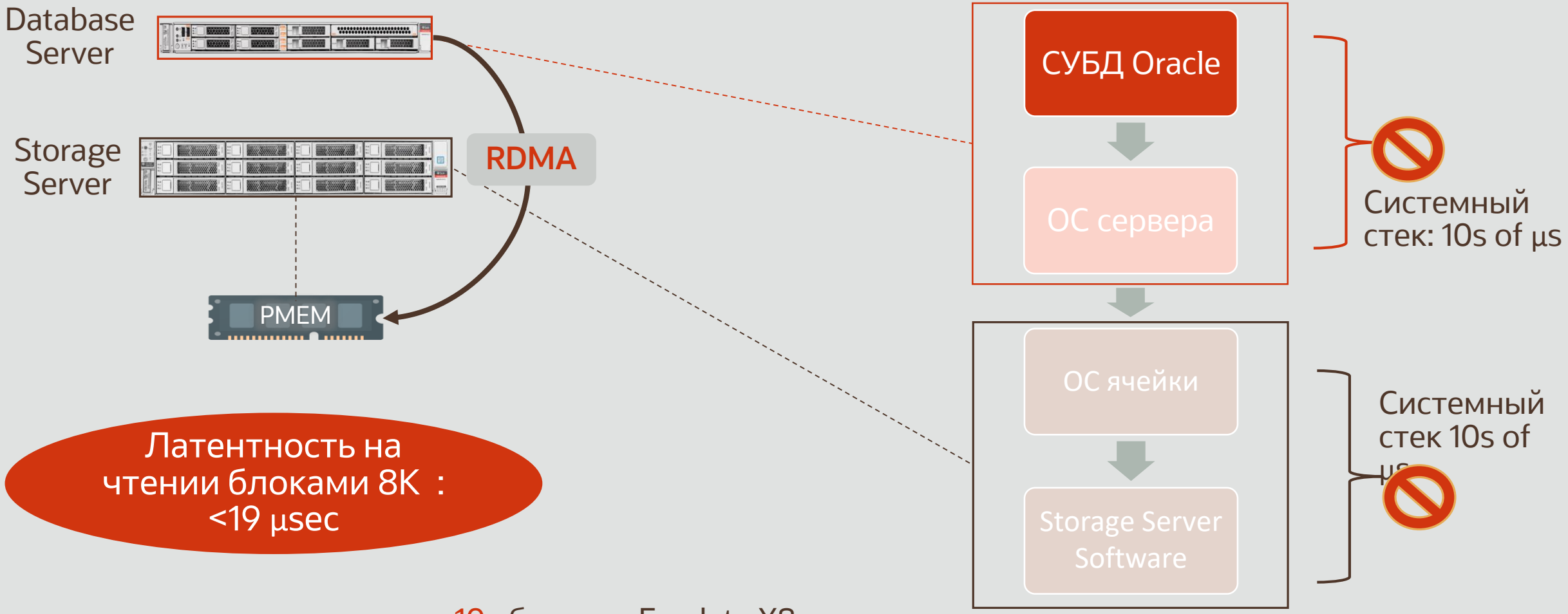


- Intel® Optane™ DC Persistent Memory:
  - Операции чтения/записи на скорости, почти как у памяти: ~50-100 быстрее флэш
  - Отключение питания переносит как флэш

- Обладает интерфейсом для работы из приложения (СУБД Oracle)

PMEM Cache

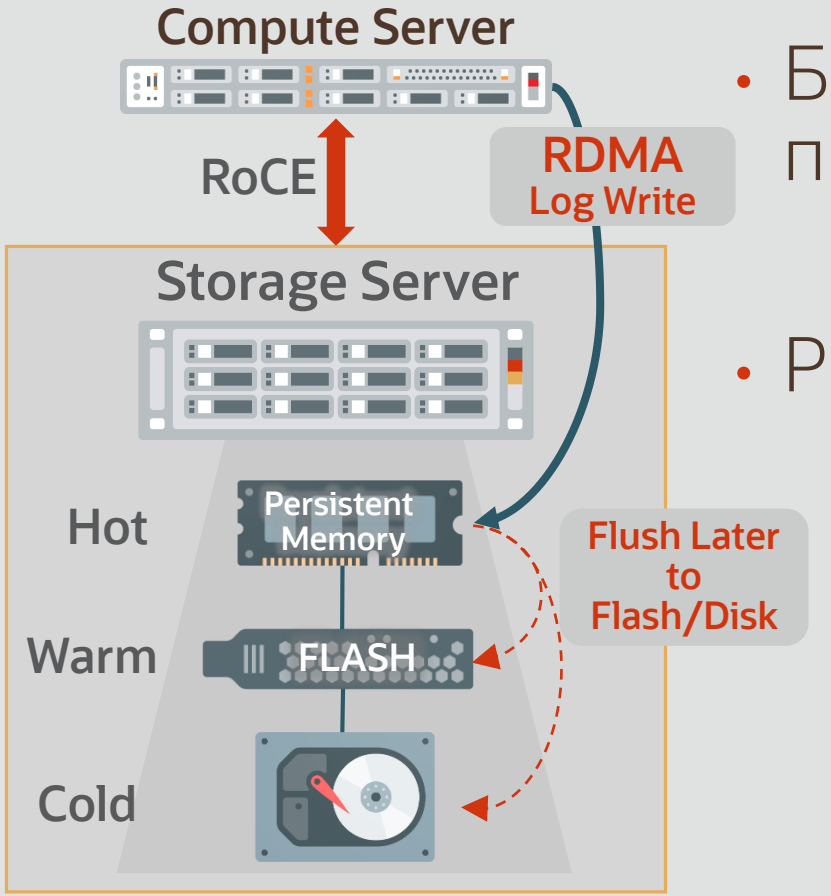
Как исключение лишнего системного стека ускоряет I/O



10x быстрее Exadata X8

Exadata PMEM Log

ускорение на аналогичных принципах – исключение системного стека



- Быстрая запись REDO Log критична для производительности OLTP приложений
- PMEM + RDMA ускоряют запись REDO Log
  - СУБД напрямую работает с PMEM, не обращаясь даже к системному ПО ячейки
  - Ускорение ~8 раз

Необходимо Exadata System Software 19.3 и Database Software 19c

Exadata: портрет семейства

	V1  Sep 2008 Xeon E5430 Harpertown	V2  Sep 2009 Xeon E5540 Nehalem	X2  Sep 2010 Xeon X5670 Westmere	X3  Sep 2012 Xeon E5-2690 Sandy Bridge	X4  Nov 2013 Xeon E5-2697v2 Ivy Bridge	X5  Dec 2014 Xeon E5-2699v3 Haswell	X6  Apr 2016 Xeon E5-2699v4 Broadwell	X7  Oct 2017 Xeon 8160 Skylake	X8  Apr 2019 Xeon 8260 Cascade Lake	X8M  Sep 2019 Xeon 8260 Cascade Lake	V1 – X8M
Storage (TB)	168	336	504	504	672	1344	1344	1.68B	2.3	2.3 PB	14 X
Flash Cache (TB)	0	5.3	5.3	22.4	44.8	89.6	179.2	358	358	358 TB	64 X
CPU (ядер)	64	64	96	128	192	288	352	384	384	384	6 X
Max Mem (GB)	256	576	1152	2048	4096	6144	12288	12288	12288	12 TB	48 X
Internal Network Fabric (GB/s)	20	40	40	40	80	80	80	80	80	200 Gb/s	**10x**
Ethernet (Gb/s)	8	24	184	400	400	400	400	800	800	800 Gb/s	100 X
Scan Rate (GB/s)	14	50	75	100	100	263	301	350	560	560 GB/s	40 X
Read IOPS (M)	.05	1	1.5	1.5	2.66	4.14	5.6	5.97	6.57	16 M	**320 X**

Задавайте вопросы, пожалуйста!

Больше информации:

- oracle.com/exadata
- blogs.oracle.com/exadata

ORACLE