Exadata with Persistent Memory: An Epic Journey

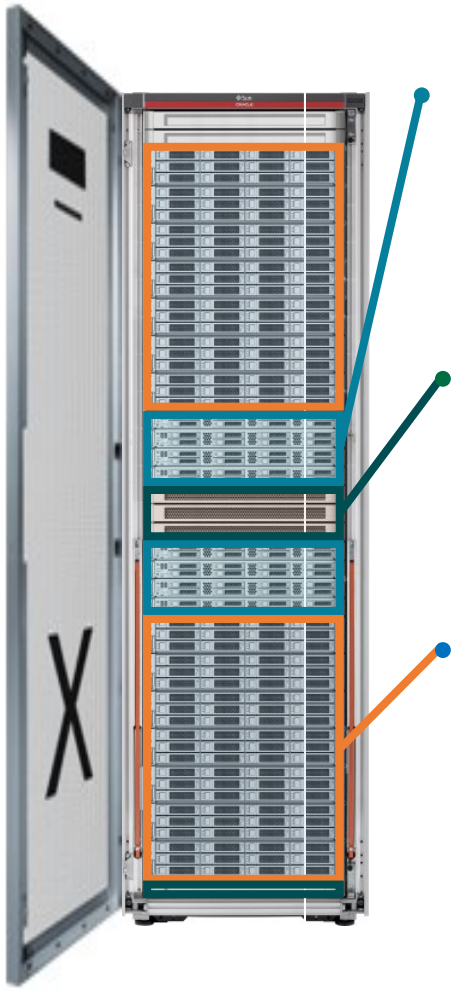
Jia Shi, Senior Director of Exadata Development, Oracle

What is an Exadata?

Exadata Vision

- **Ideal Database Hardware** – Scale-out, database optimized compute, networking, and storage for fastest performance and lowest cost
- **Smart System Software** – Specialized algorithms vastly improve all aspects of database processing: OLTP, Analytics, Consolidation
- **Automated Management** – Automation and optimization of configuration, updates, performance, and management culminating in Fully Autonomous Infrastructure and Database

Exadata X8M-2



Scale-Out 2-Socket Database Servers

2 socket Xeon processors
48 cores per server
384 GB - 1.5 TB DRAM

100Gb **RDMA** over Converged Ethernet (RoCE) Network Fabric

100 Gb/s RoCE network fabric
25/10 GigE external network

Scale-Out ***Intelligent*** 2-Socket Storage Servers

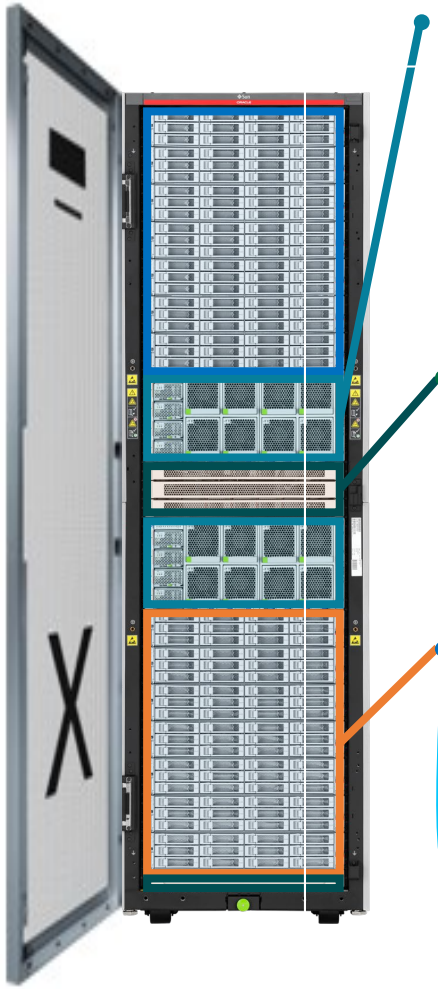
High Capacity

168 TB disk capacity
25.6 TB PCI NVMe Flash
1.5 TB Persistent Memory
32 cores for SQL offload

Extreme Flash

51.2 TB PCI NVMe Flash
1.5 TB Persistent Memory
32 cores for SQL offload

Exadata X8M-8



Scale-Out 8-Socket Database Servers

- **8** socket Xeon processors
- **192** cores
- **3-6 TB** DRAM

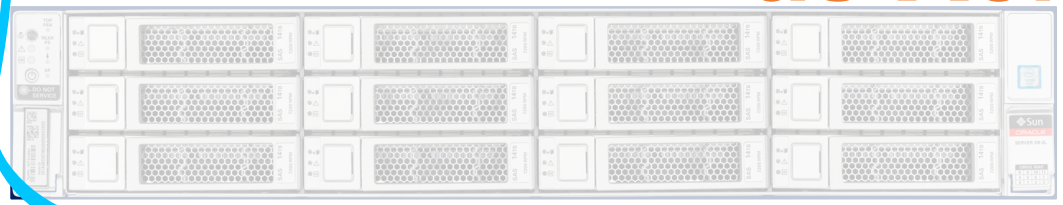
Large SMP Processor Model
Big data warehouses
Massive database consolidation
In-Memory databases

100Gb **RDMA** over Converged Ethernet (RoCE) Network Fabric

100 Gb/s RoCE network fabric
25/10 GigE external network

Same Storage, Network and Software as X8M-2

Scale-Out ***Intelligent*** 2-Socket Storage Servers



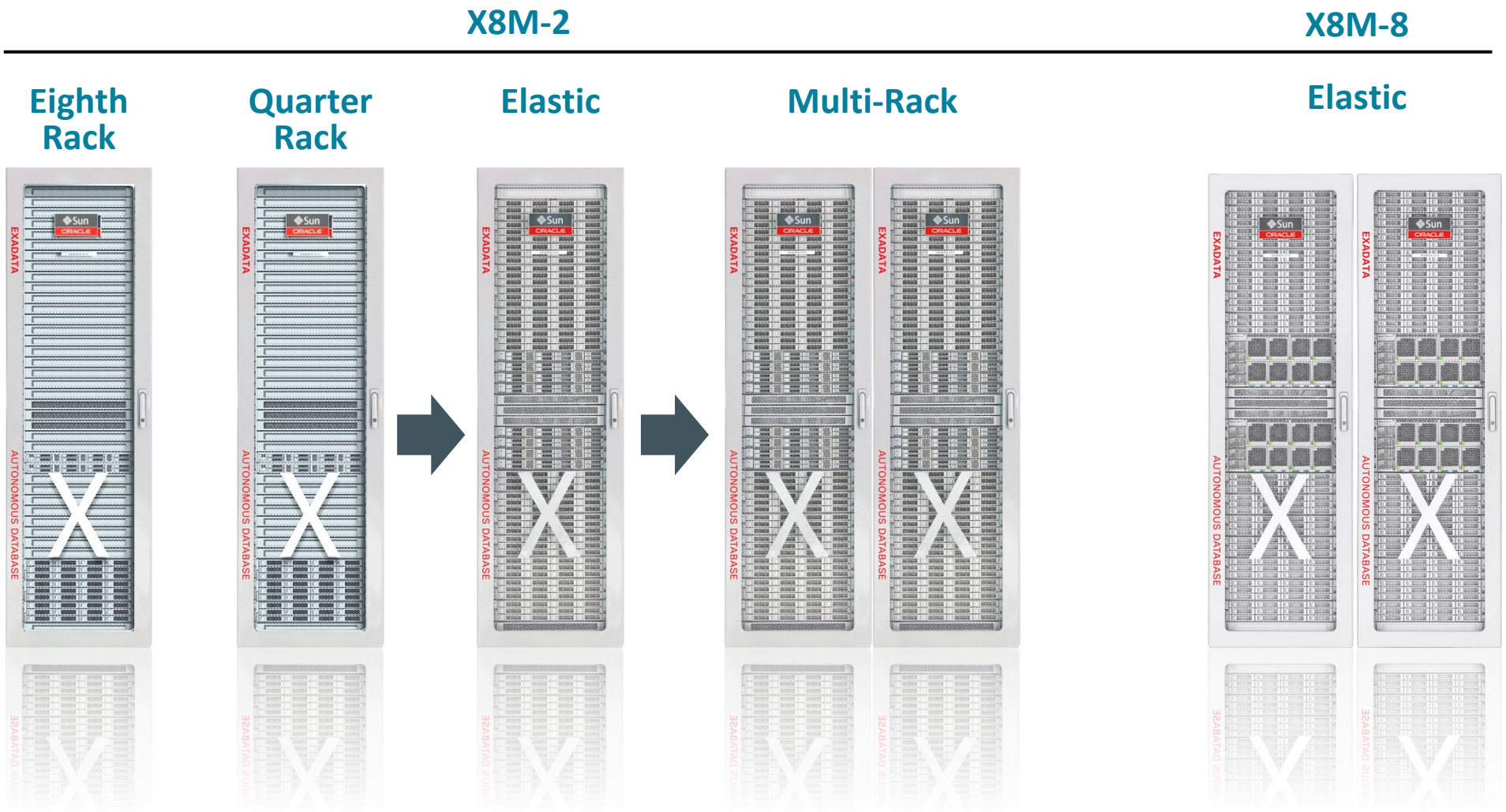
High Capacity

168 TB disk capacity
25.6 TB PCI NVMe Flash
1.5 TB Persistent Memory
32 cores for SQL offload

Extreme Flash

51.2 TB PCI NVMe Flash
1.5 TB Persistent Memory
32 cores for SQL offload

Exadata X8M: Scalability



What Happened In The Last Decade?

The Flash Revolution (2009-2019)

	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

Flash NVMe

Is the Flash Revolution Plateauing?

	V1	V2	X2	X3	X4	X5	X6	X7	X8	X8M
										
	Sep 2008 Xeon E5430 Harperto wn	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

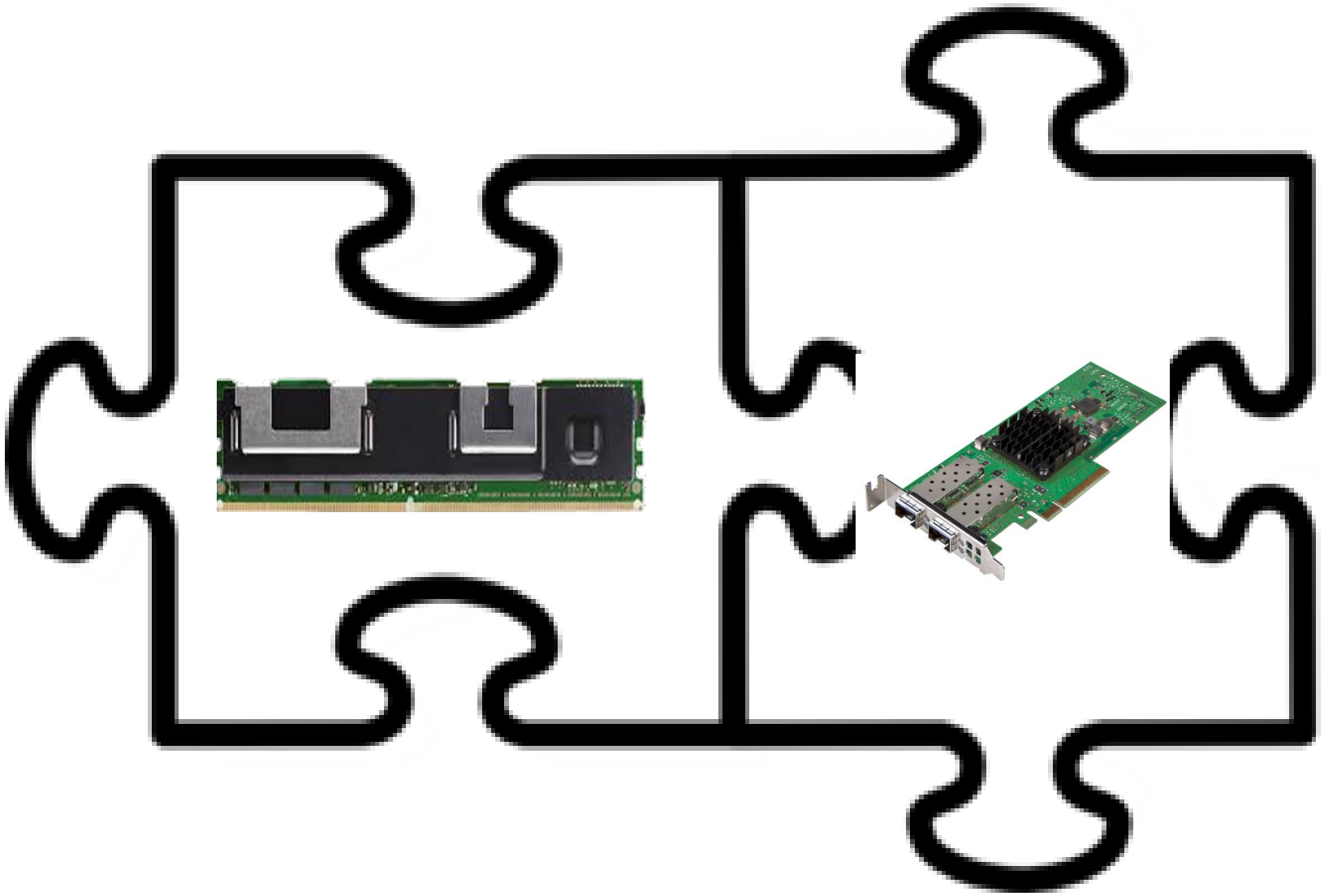
Flash SCSI

Flash NVMe

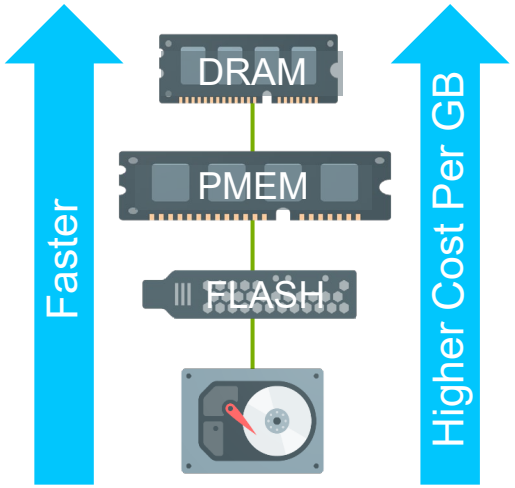
16M IOPS
19 usec latency

What Is The Secret Sauce In X8M?

Do You Recognize The Dynamic Duo?

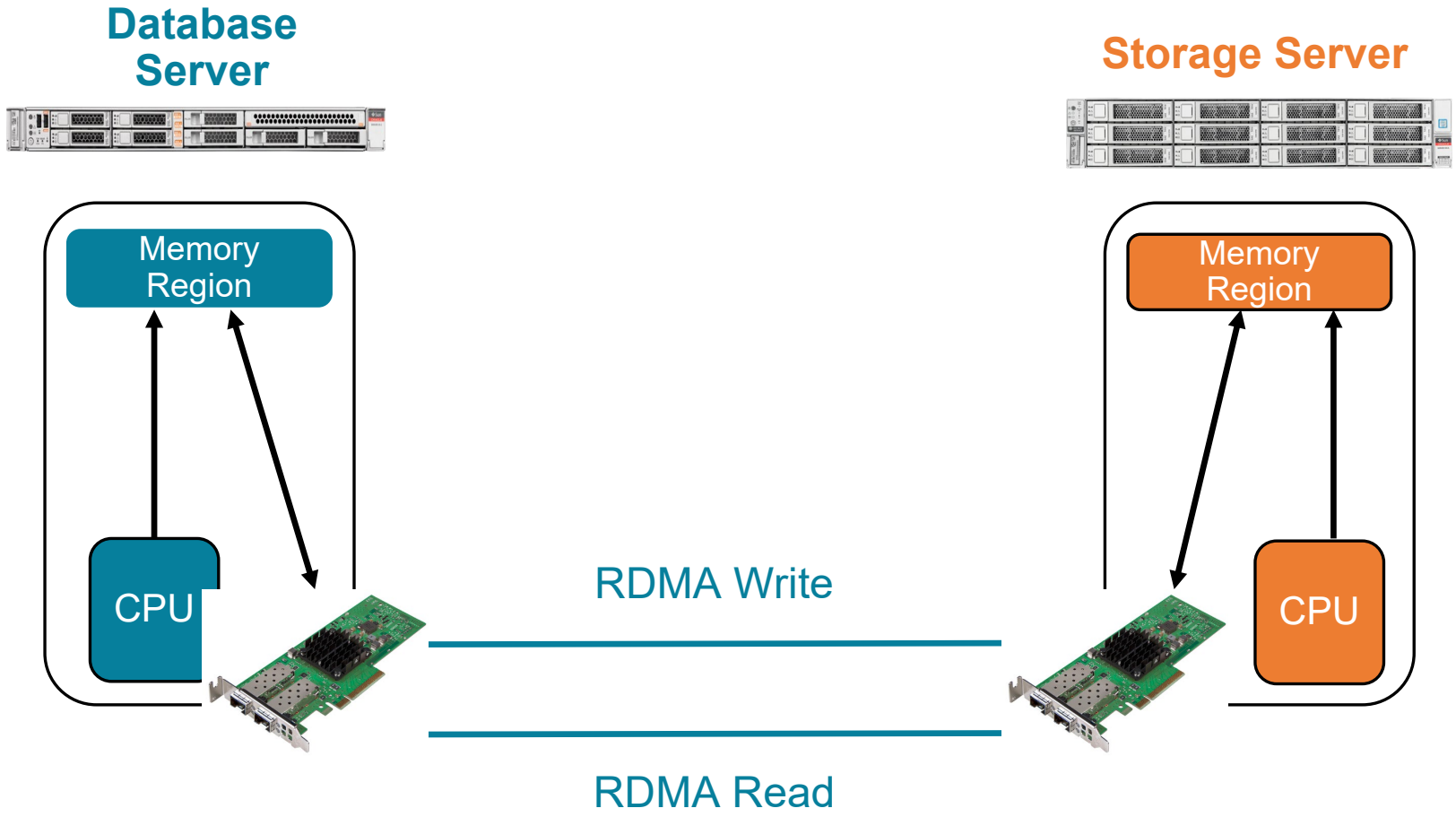


Intel® Optane™ DC Persistent Memory

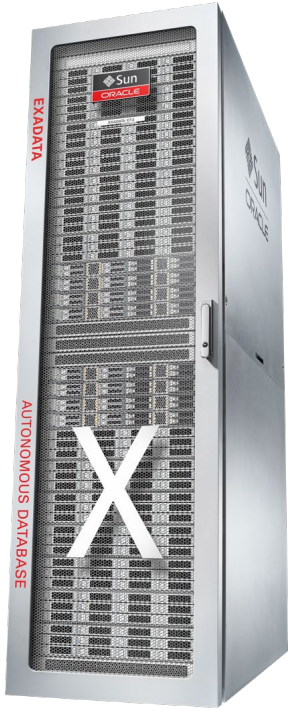


- Capacity, performance, price between DRAM and flash
- Reads at memory speed – much faster than flash
- Writes survive power failure unlike DRAM
- Requires sophisticated algorithms for PMEM data integrity upon failure
  - Call special instructions to flush data from CPU cache to PMEM
  - Complete or backout sequence of writes interrupted by a crash

Remote Memory Direct Access (RDMA)



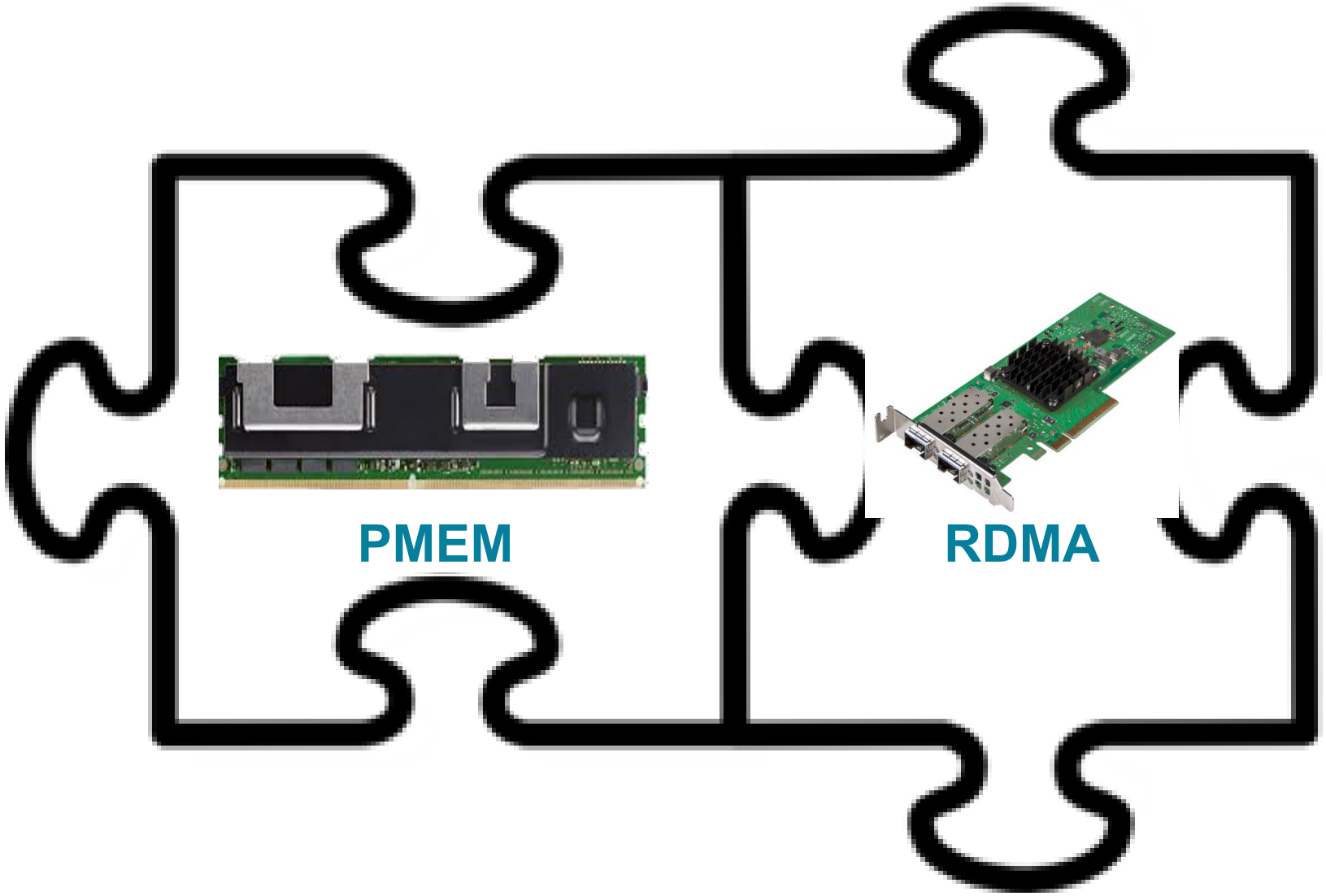
New RoCE Internal Network Fabric



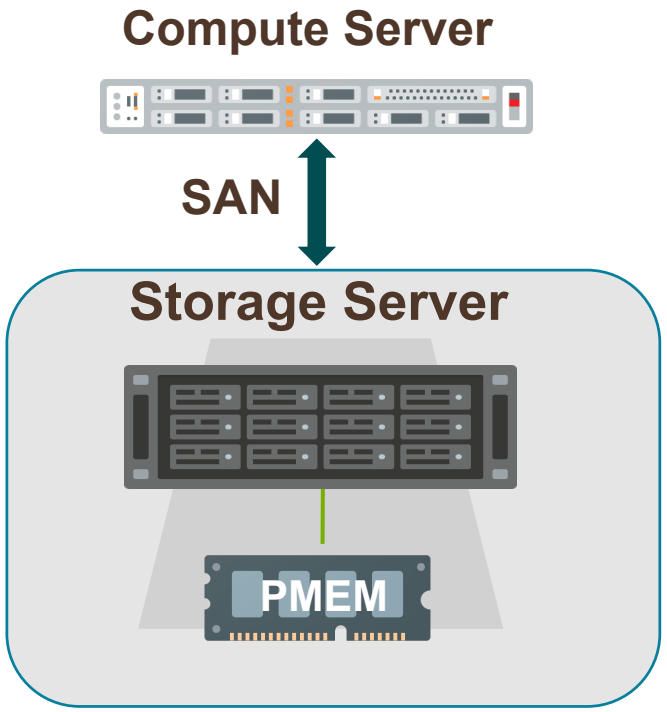
World's First and Only
RoCE-based Database Machine

- Exadata 100Gb/s RoCE enables RDMA on Ethernet fabric
- Exadata runs RDMA on InfiniBand since 2008 for high throughput and low latency
- What is the new trick?

What can they do together?



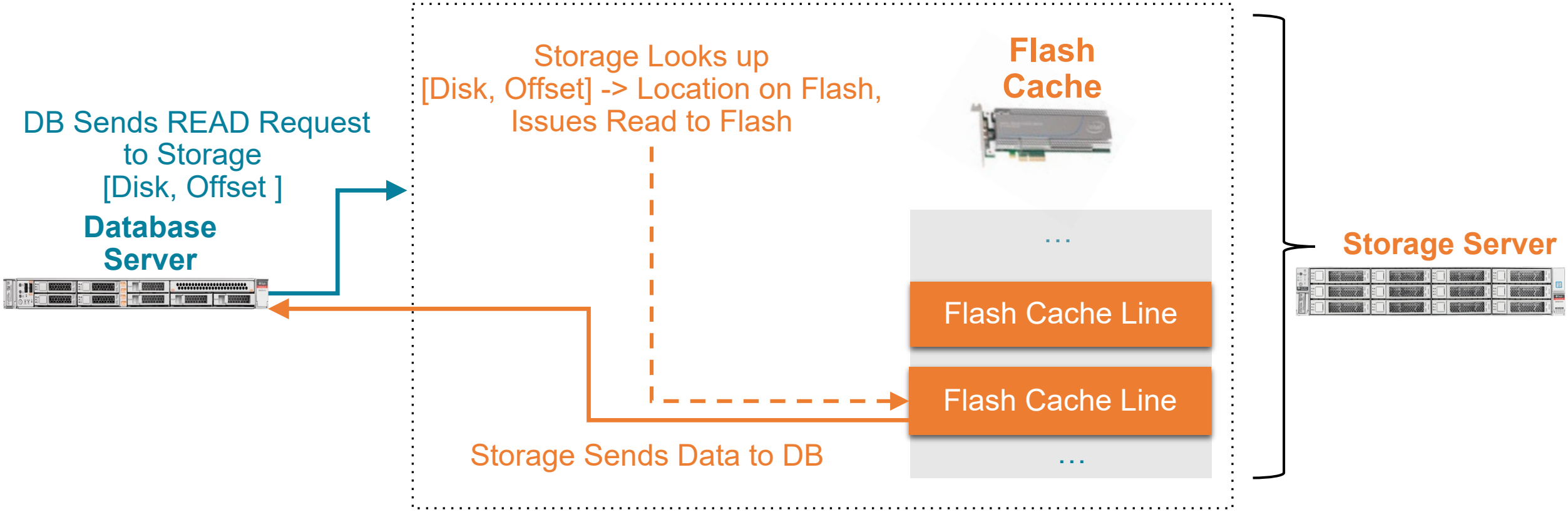
PMEM in Conventional Storage



- Persistent Memory usage with conventional storage:
  - Database issues read I/O call to OS
  - OS sends message to storage
  - Storage CPU issues read to Persistent Memory
  - Storage CPU sends reply to Server OS
  - Server OS wakes up Database
- Overwhelmed by high cost network and I/O software, interrupts, and context switches
- Performance benefit from PMEM is **erased**

Conventional Read

Conventional Storage Performs Server-side Cache Read



Dissect the Exadata Flash IO Latency

Database Server

Storage Server

Flash Read Raw Latency: <100 us

Database 8K Read End-to-end Latency: ~200 usec

Database Software

Kernel/OS (Database Server)

Context Switch: 10s of us

Kernel/OS (Storage Server)

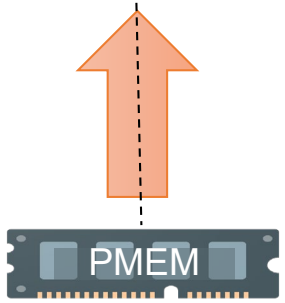
Storage Server Software

Context Switch: 10s of us

What if we drop in PMEM as is?

Database Server

Storage Server



PMEM Read Raw Latency: **~1 us**

Database Software

Kernel/OS (Database Server)

Context Switch: 10s of us

Kernel/OS (Storage Server)

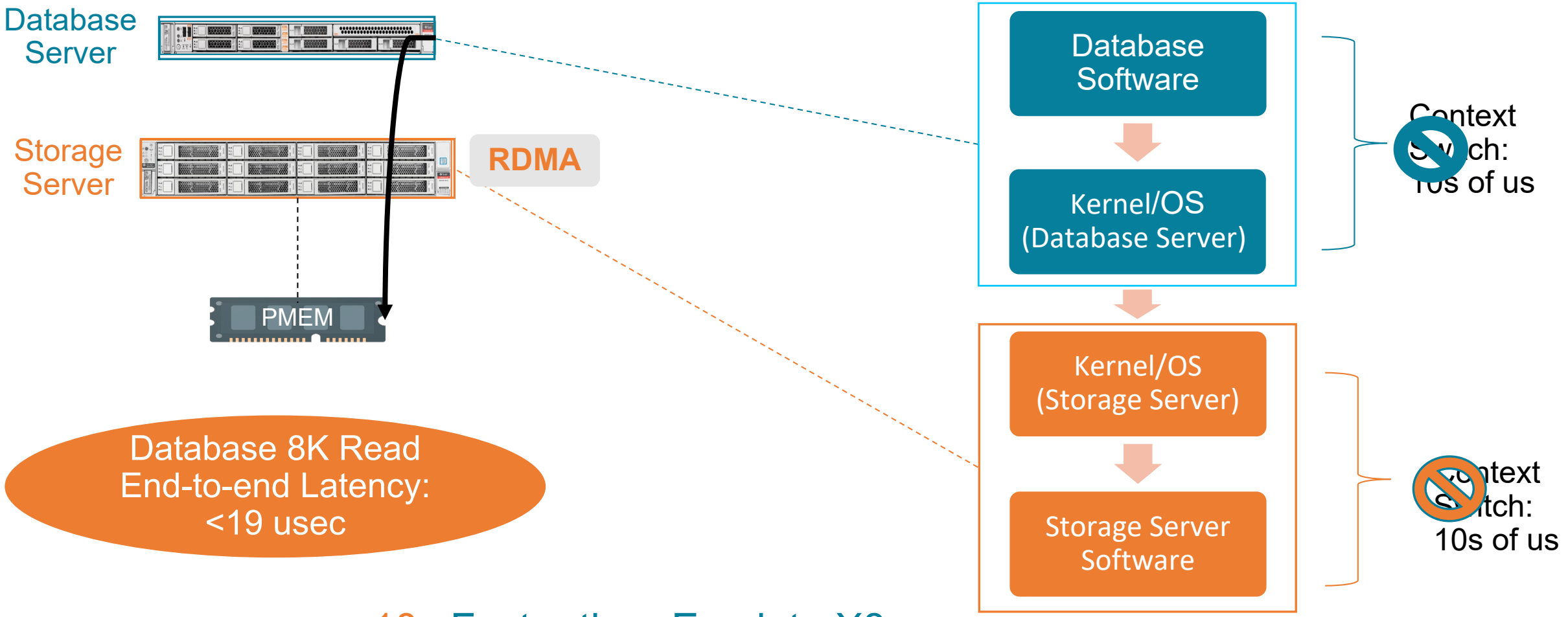
Storage Server Software

Context Switch: 10s of us

Database 8K Read End-to-end Latency: ~100 usec

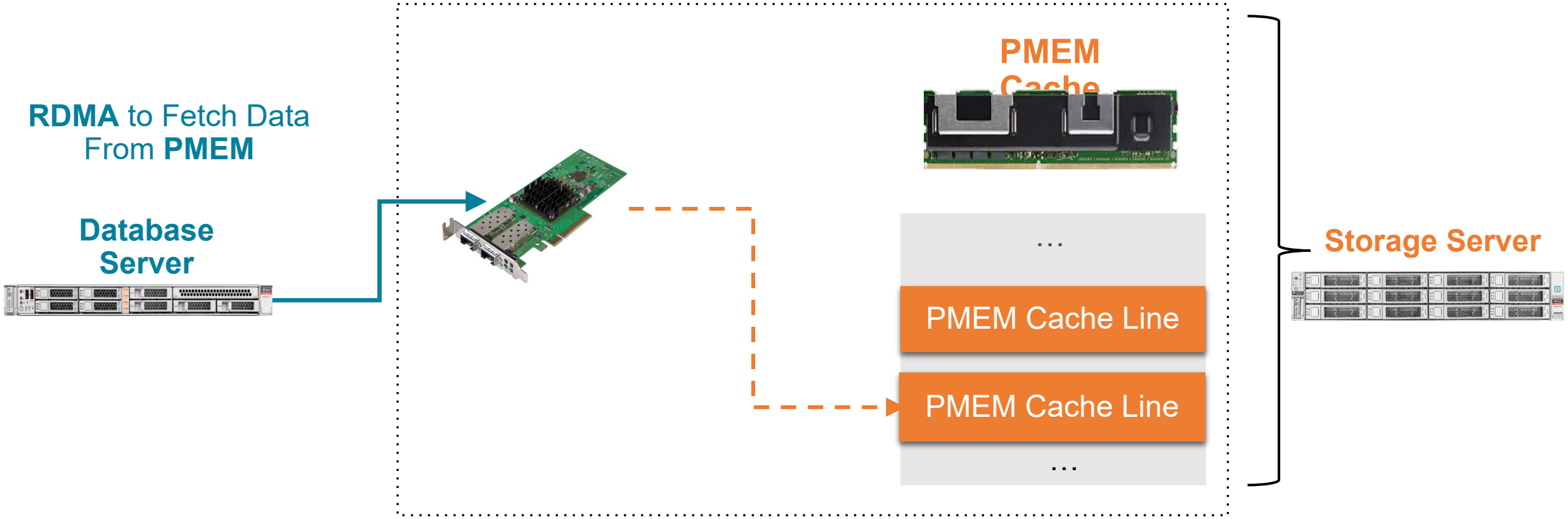
Over 90% of the Time *Wasted*.

A Radical Approach – RDMA to PMEM



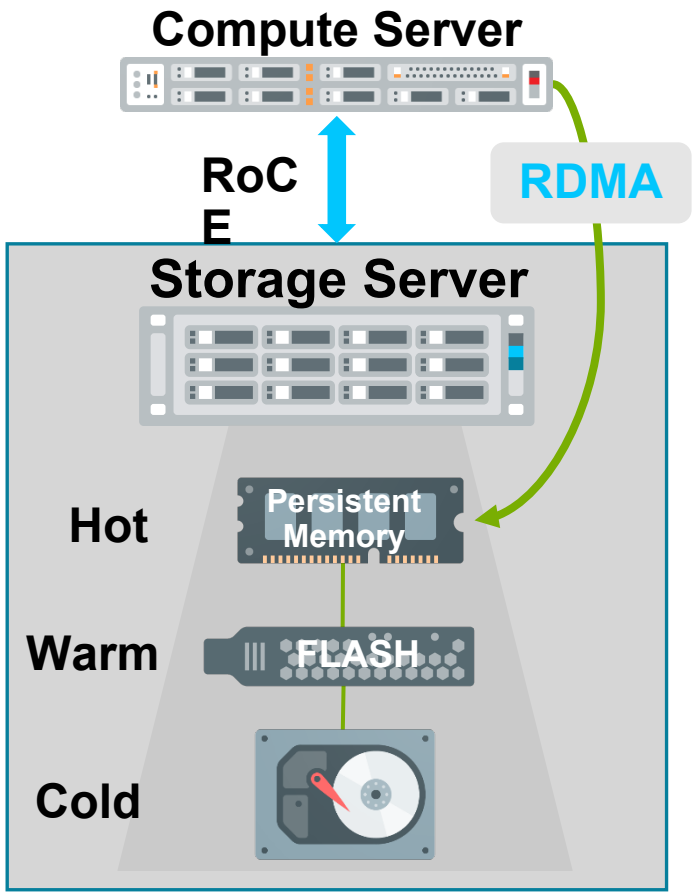
Ultra Fast RDMA Read

New Disruptive Technology Enables PMEM Cache Read via RDMA



Exadata X8M With Persistent Memory Data Accelerator

World's First and Only Shared Persistent Memory Optimized for Database

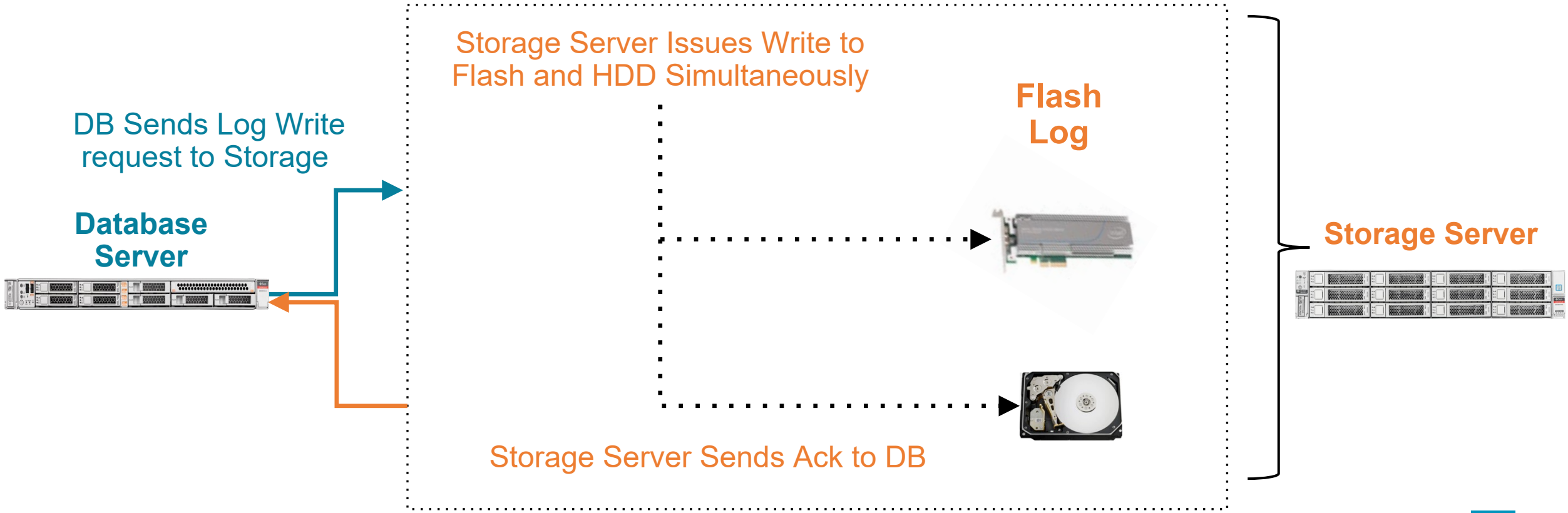


- Exadata Storage Servers transparently add Persistent Memory Accelerator in front of Flash memory
- Database uses RDMA instead of I/O to read remote PMEM
  - Bypasses network and IO software, interrupts, context switches
- PMEM Automatically tiered and shared across DBs
  - Using as a cache for hottest data increases effective capacity 10x
- Persistent Memory mirrored automatically across storage servers for fault-tolerance
- 16Million IOPS, <19us latency for 8K I/Os from the database

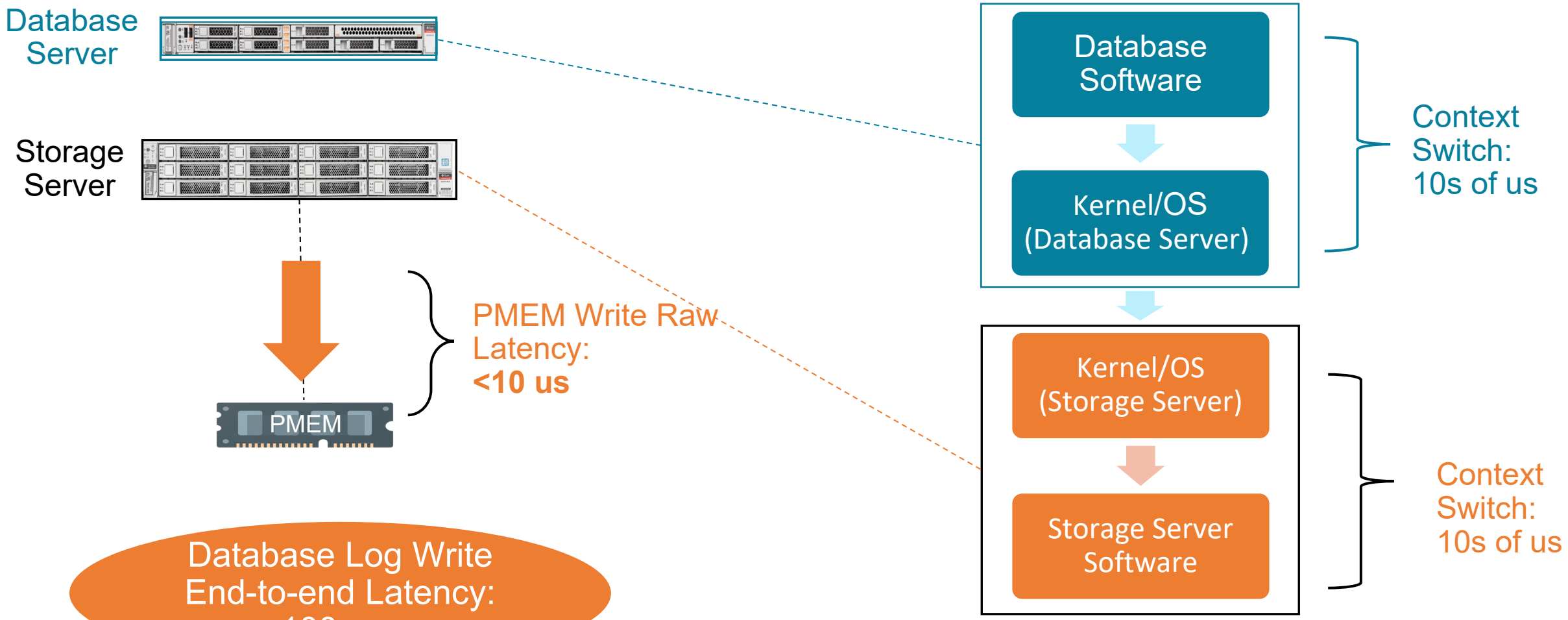
Enabled with Exadata System Software 19.3 and Database Software 19c

Conventional Log Write

- DB Sends Requests to Storage
- Storage Writes to Flash Log and Sends Ack



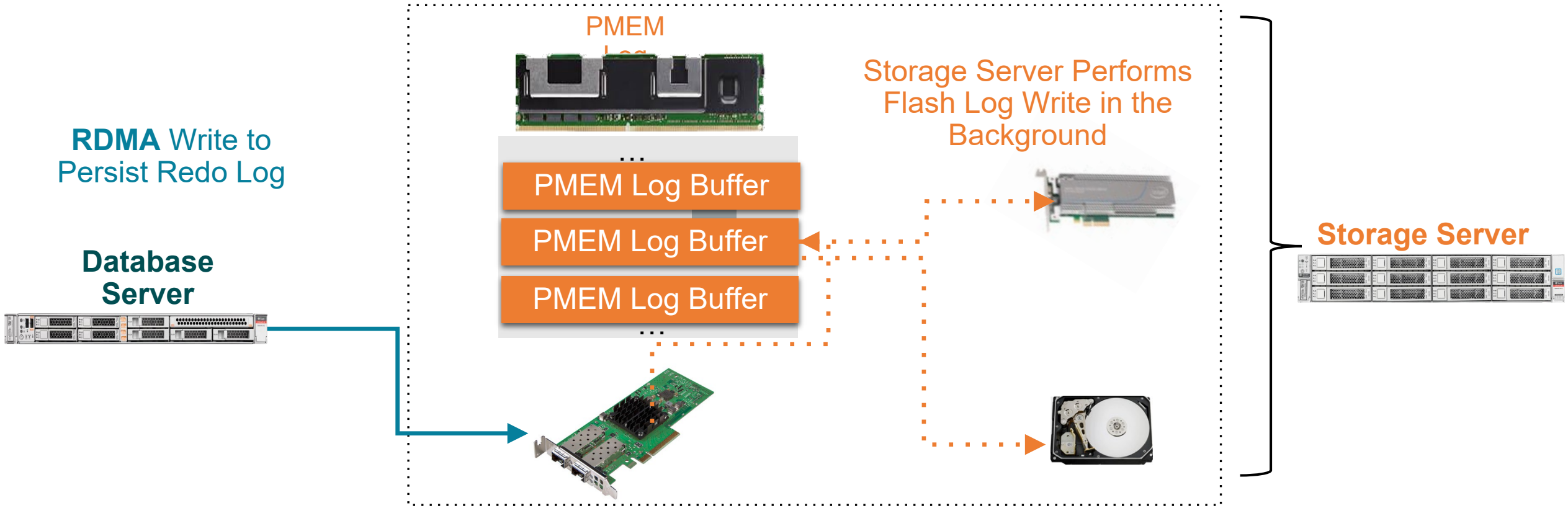
What about Redo Log Writes to PMEM?



Over 90% of the Time *Wasted*.

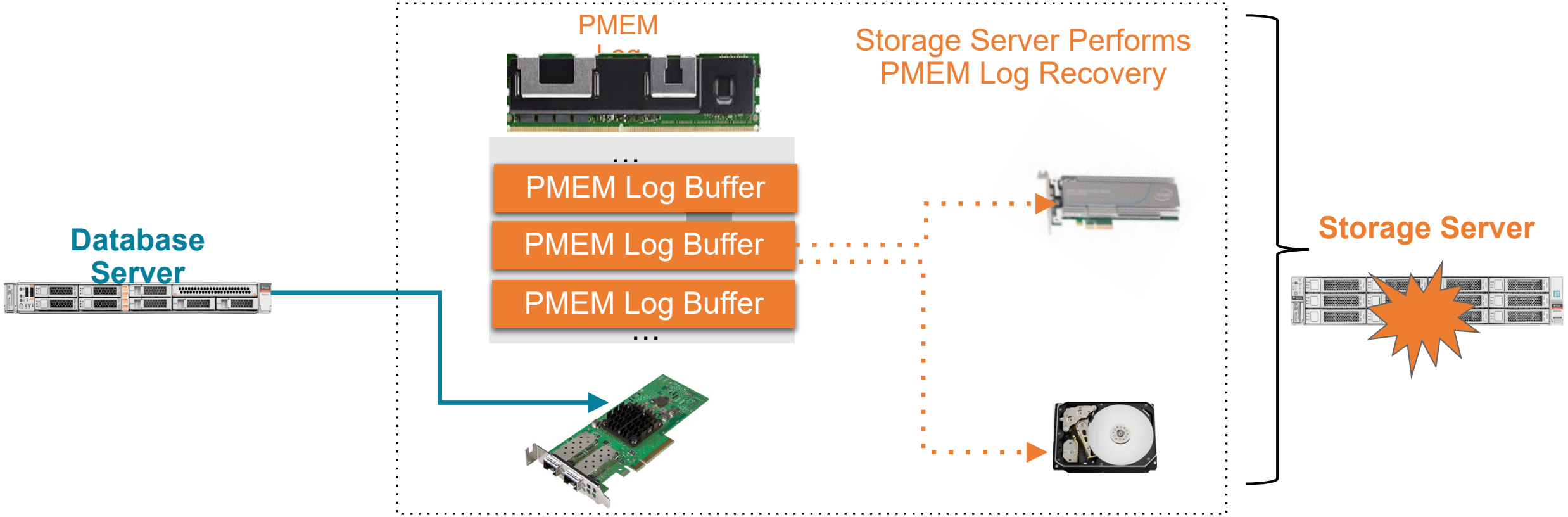
Ultra Fast RDMA Log Write

New Disruptive Technology Enables Durable PMEM Log Write via RDMA

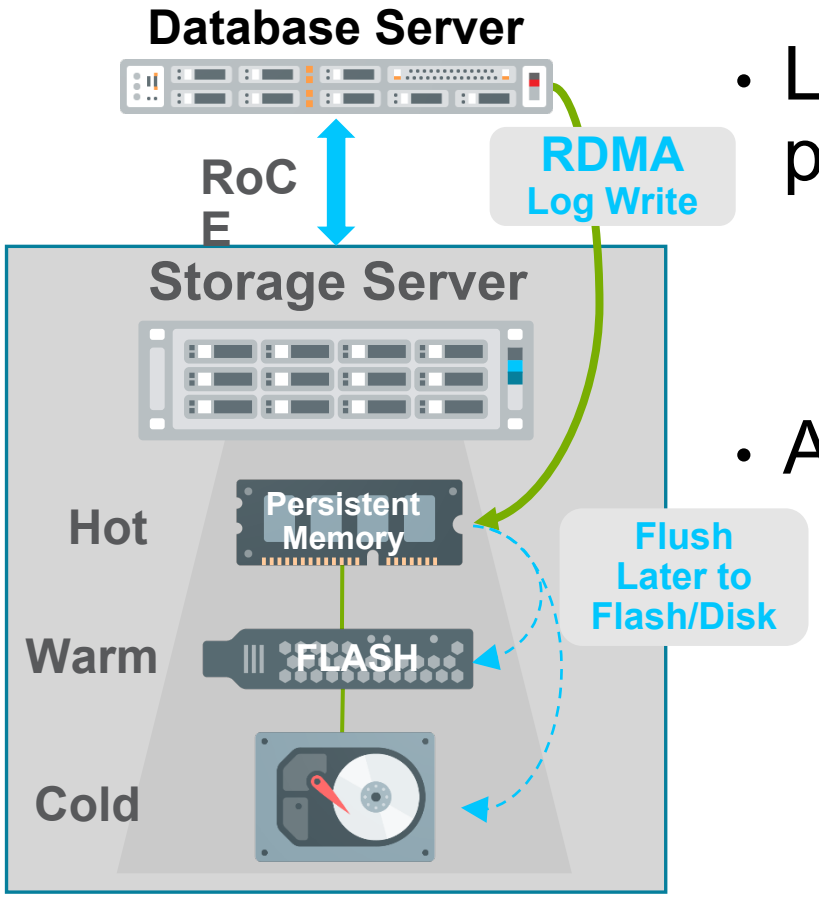


Ultra Fast RDMA Log Write

Storage Server Crash Safe!



Exadata X8M Persistent Memory Commit Accelerator



- Log Write latency is critical for OLTP performance
  - Faster log writes means faster commit times
  - Any log write slowdown stalls the whole database
- Automatic Commit Accelerator
  - Database issues one-way RDMA writes to PMEM on multiple Storage Servers
  - Bypasses network and I/O software, interrupts, context switches, etc.
  - Up to 8x faster log writes

Enabled with Exadata System Software 19.3 and Database Software 19c

An Epic Journey

	V1	V2	X2	X3	X4	X5	X6	X7	X8	X8M	V1 – X8M Growth
											
	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	
Storage (TB)	168	336	504	504	672	1344	1344	1.68	2.35	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 (cores)	64	64	96	128	192	288	352	384	384	384	6 X
Max Mem (GB)	256	576	1152	2048	4096	6144	12288	12288	12	12 TB	48 X
RDMA Network (Gb/s)	20	40	40	40	80	80	80	80	80	200 Gb/s	**10x**
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**

Thank you

Please visit www.snia.org/pmsummit for presentations