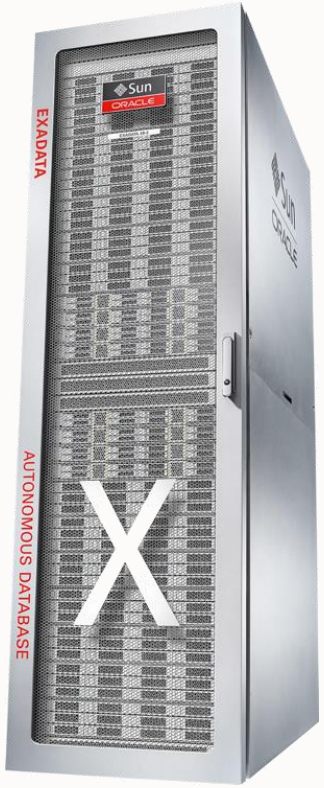
ORACLE

Under the Hood of an Exadata Transaction

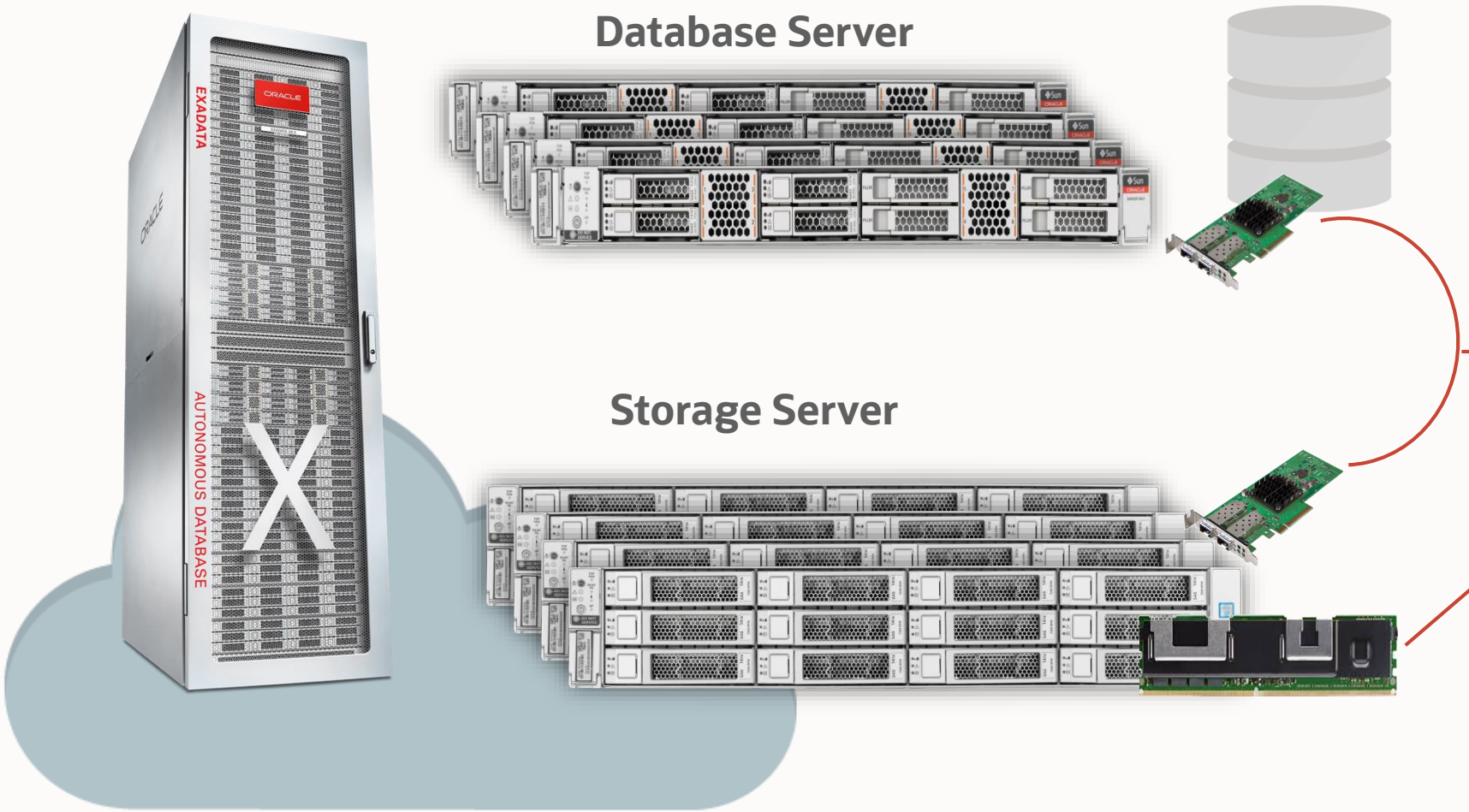
How do we harness the power of Persistent Memory?



Jia Shi

Vice President
Exadata Development, Oracle
jia.shi@oracle.com

Meet an Exadata X8M



X8M-2 socket Xeon
- **48** cores per server
- **384 GB - 1.5 TB** DRAM

X8M-8 socket Xeon
- **192** cores per server
- **3-6 TB** DRAM

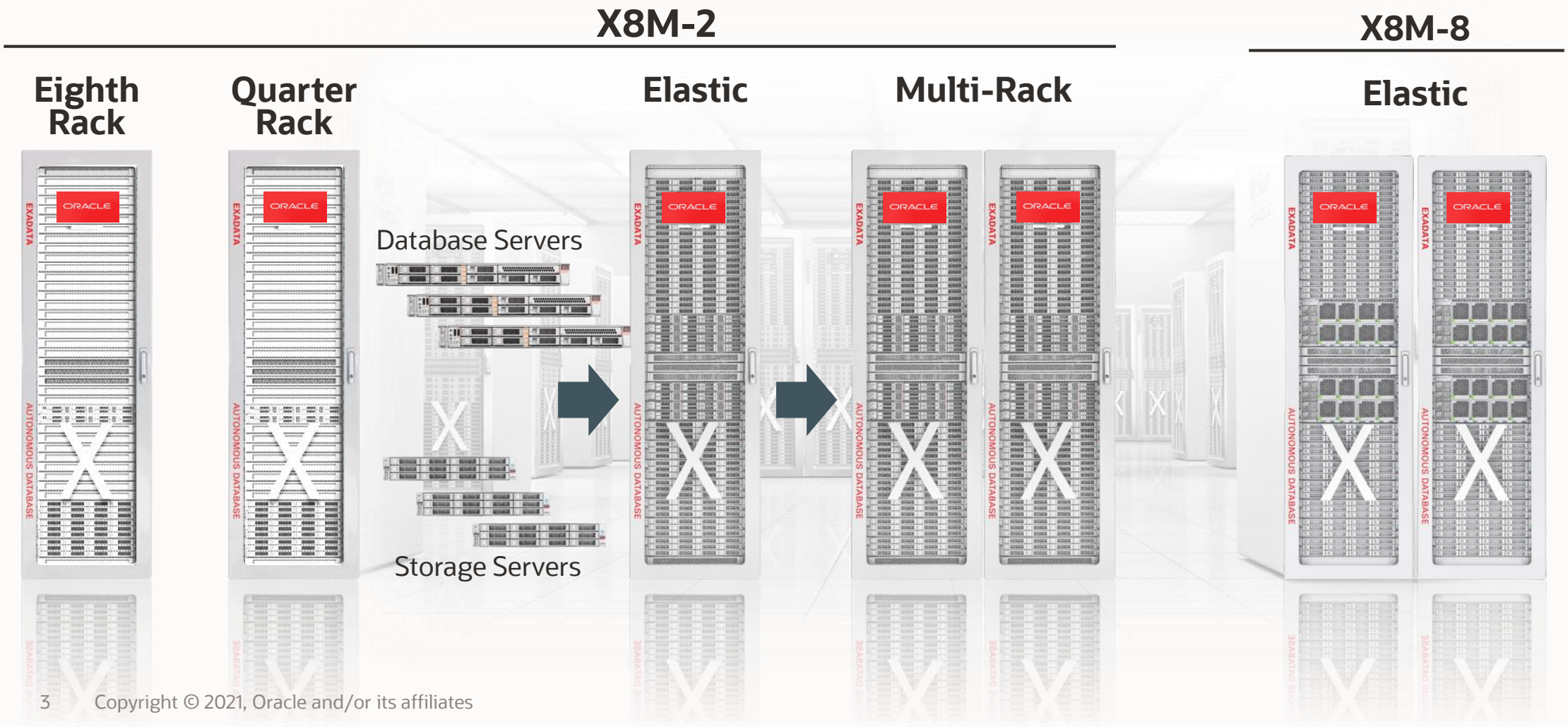
100 Gb/s **RoCE**
RDMA **o**ver **C**onverged **E**thernet

1.5 TB Persistent Memory

High Capacity
- **168** TB HDD
- **25.6** TB PCI NVMe Flash

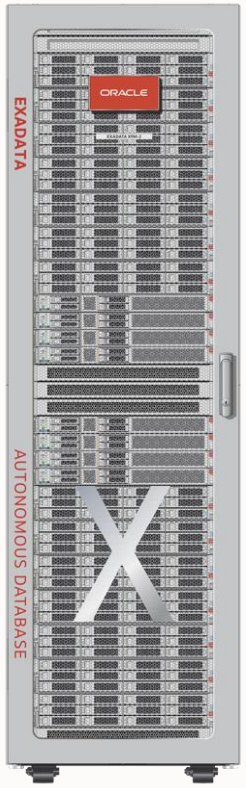
Extreme Flash
- **51.2** TB PCI NVMe Flash

Exadata X8M: Scale-out with Elastic Expansion



Exadata

Extreme Performance and Availability at Lowest Cost



Ideal Database Hardware –
Scale-out, database optimized compute, networking, and storage

Database Aware System Software -
Unique algorithms vastly improve **OLTP, Analytics, Consolidation**

Automated Management -
Fully automated and optimized platform from end to end

Identical Across On-Prem, Cloud, Cloud@ Customer

Thousands of Ultra-Critical Deployments Since 2008

87% of Fortune Global 100 Run Exadata

BEST FOR ALL WORKLOADS

Petabyte Warehouses

Ultra-critical Oltp – Financial Trading, Etc

Complex Business Applications - Sap, Oracle, Siebel

Massive Db Consolidation

target · CaixaBank · Mercedes-Benz · sentry data systems · ANA · Cerner
物美 WUMART · Banco de Chile · GE · METRO · LAND O'LAKES, INC. · KDDI
Rakuten · Exelon · LG · SAMSUNG · 中国移动 China Mobile · Panasonic
swisscom · Medimpact · BT · FICC · EDF · NTT docomo
Canon · GOL · NRI 未来創発 Dream up the future. · ENERGY TRANSFER · P&G · HALLIBURTON 100 YEARS
AmerisourceBergen · HAVI · Swiss Re · xerox · Telefónica · dialog SEMICONDUCTOR

Let’s go under the hood of OLTP - *OnLine Transactional Processing*

Meet a Super Critical OLTP Application

What does such an app do?

Ben wants to deposit $1000 to his bank account.
Alice wants to withdraw $500 from her bank account.
Bob wants to transfer $2000 from checking into savings.
....

User

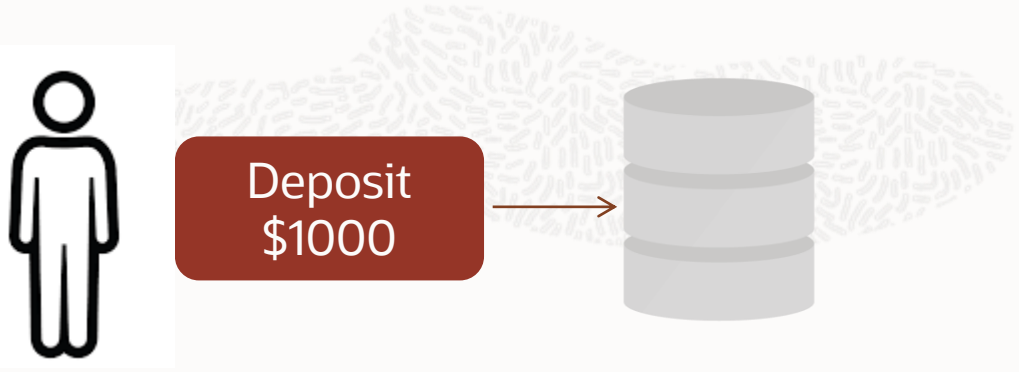
Super Critical OLTP App

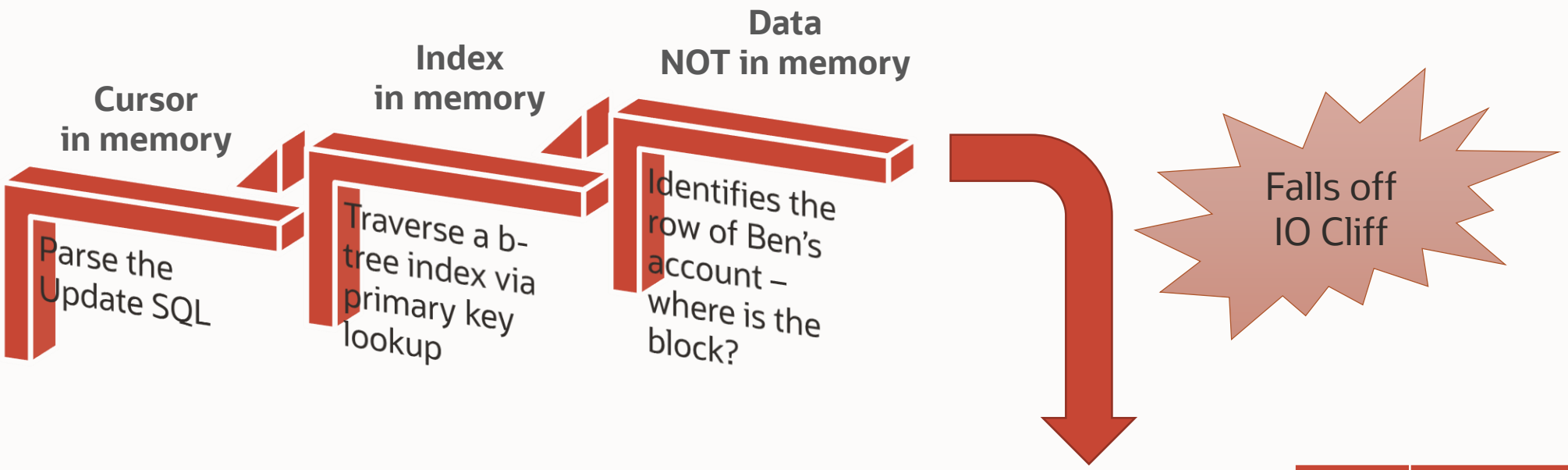
Database Transaction Processing

Meet Ben's Transaction

What constitutes a database transaction?

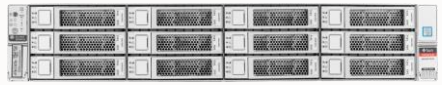
Ben wants to deposit $1000 to his bank account.





Database Server

Storage Server



User	Account Balance
…	…
Ben	$2000

OLTP Challenge #1 -

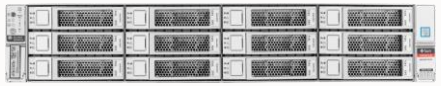
What is the IO cliff for random data reads?

Challenge #1 – Random Data Read

Database Server

1. Identifies the row of Ben's account – where is the block? Miss in the buffer cache!
2. Issues the data read to storage

Storage Server



1. Finds the data block cached in Flash Cache
2. Issues local read to Flash

Flash Cache

...

Flash Cache Line

Flash Cache Line

...

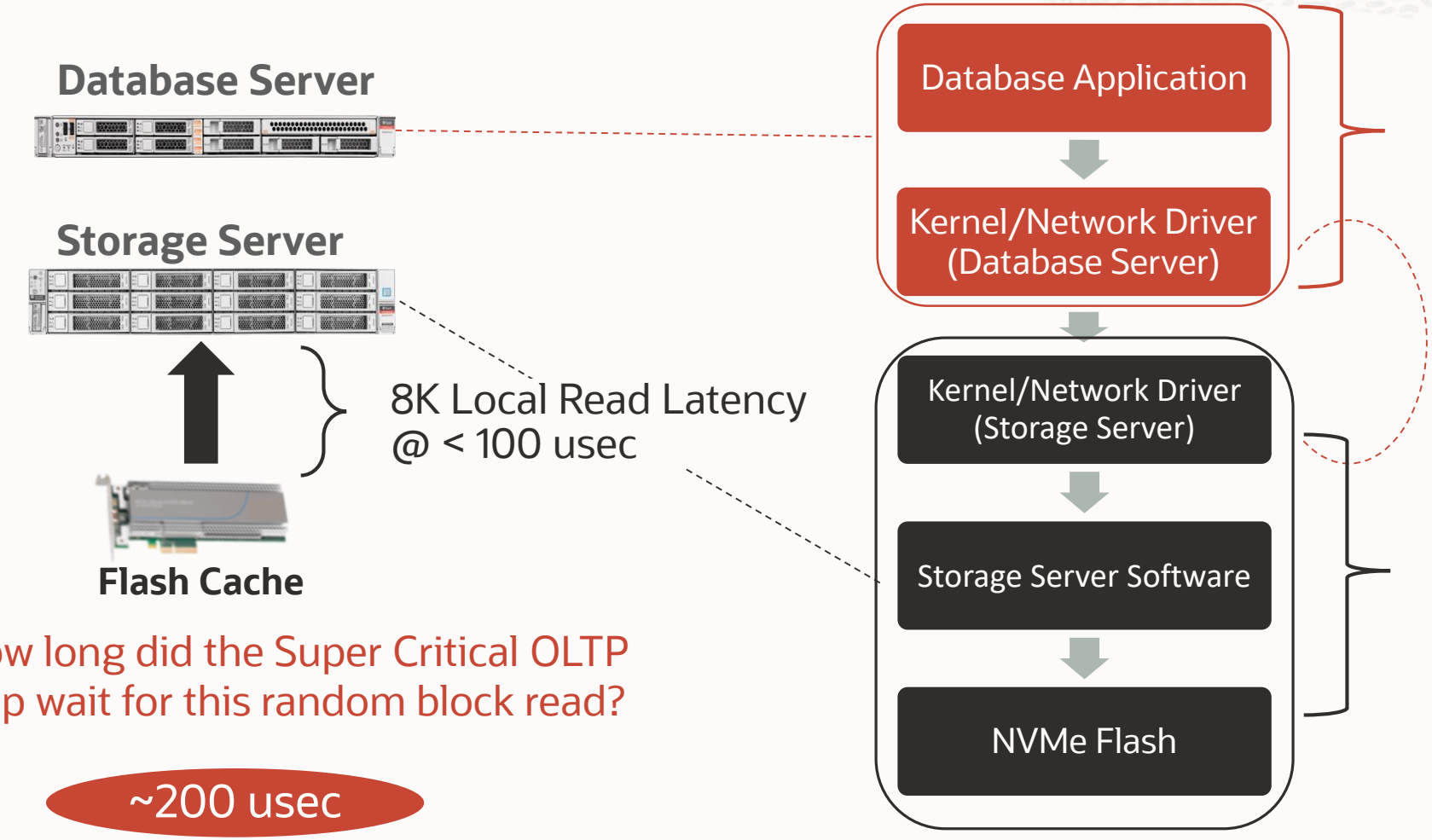
Buffer Cache (DRAM)

User	Account Balance
...	...
Ben	$2000

3.Sends data block to DB

How long did the Super Critical OLTP App wait for this random block read?

How long does 8K random read take?



- User->Kernel & Kernel-> User context switches
- Network IO interrupt processing
- IO software stack

Network Round Trip

- User->Kernel & Kernel-> User context switches
- Network IO interrupt processing
- IO software stack

How long did the Super Critical OLTP App wait for this random block read?

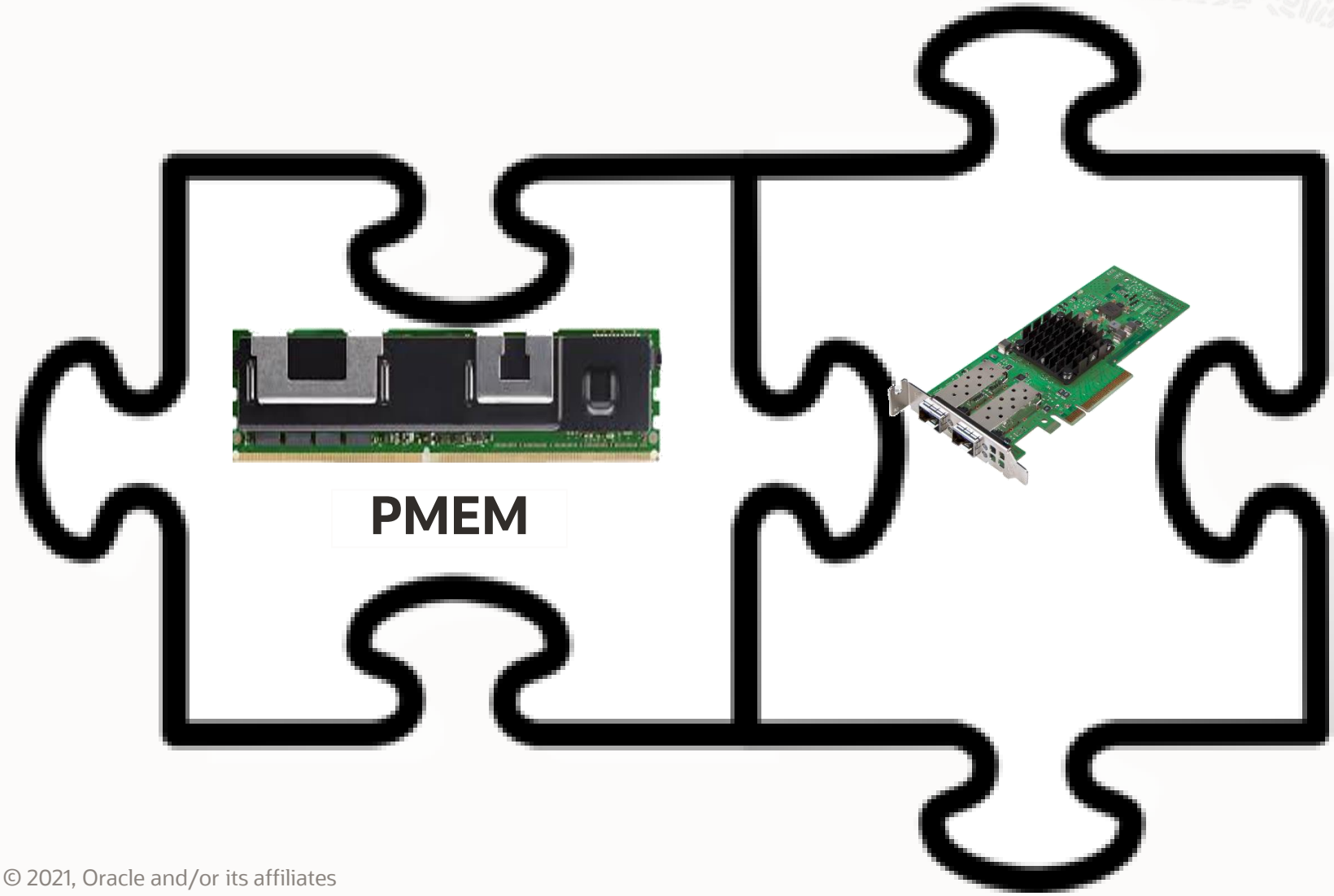
~200 usec

~100 usec

OLTP Challenge #1 -

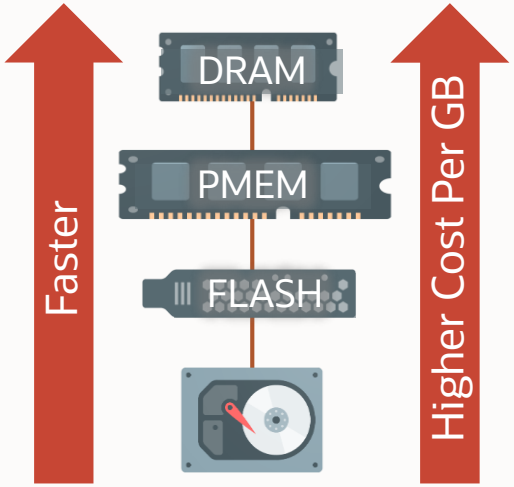
How to conquer the random read IO Cliff?

Do you recognize the dynamic duo?

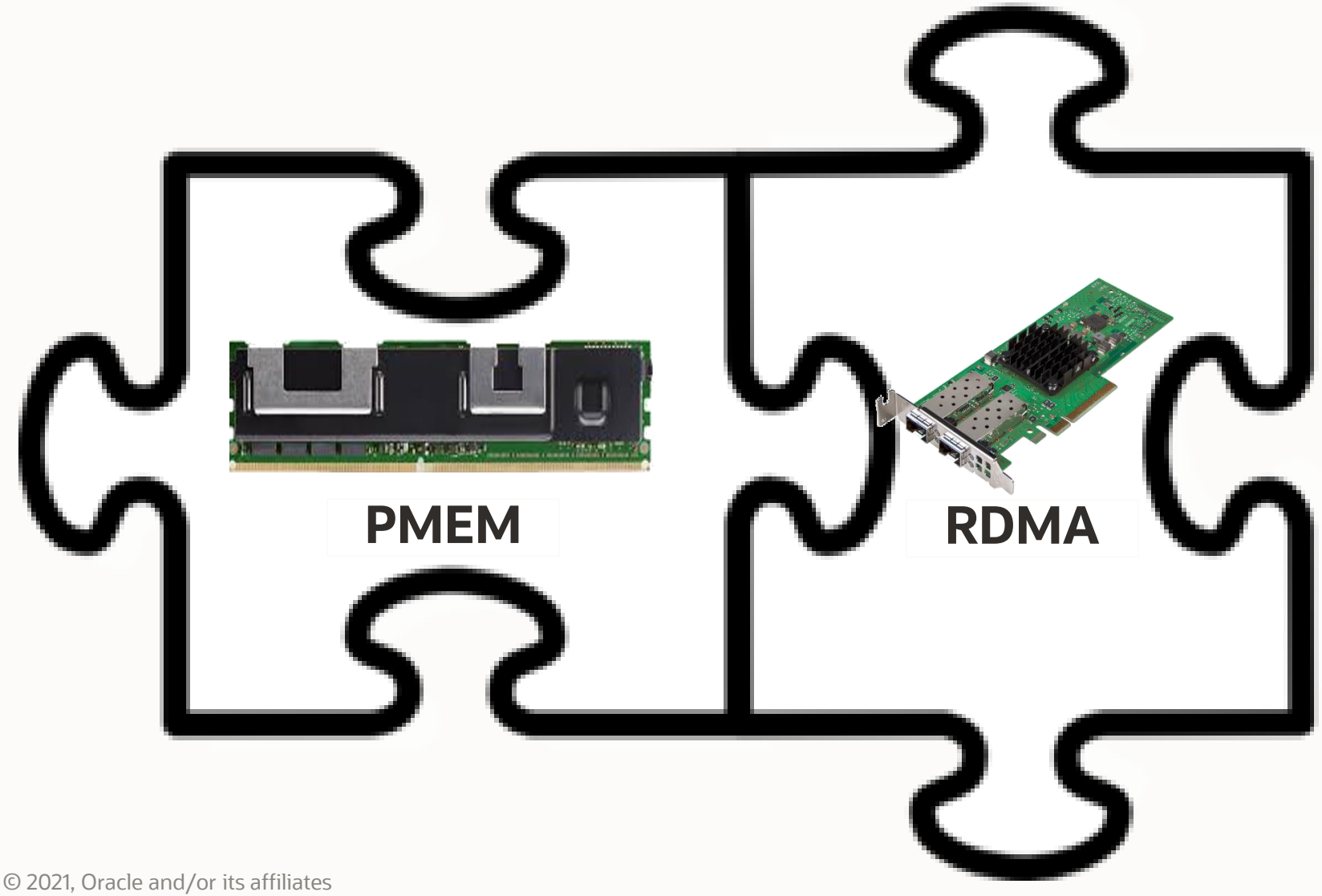


Persistent Memory (PMEM)

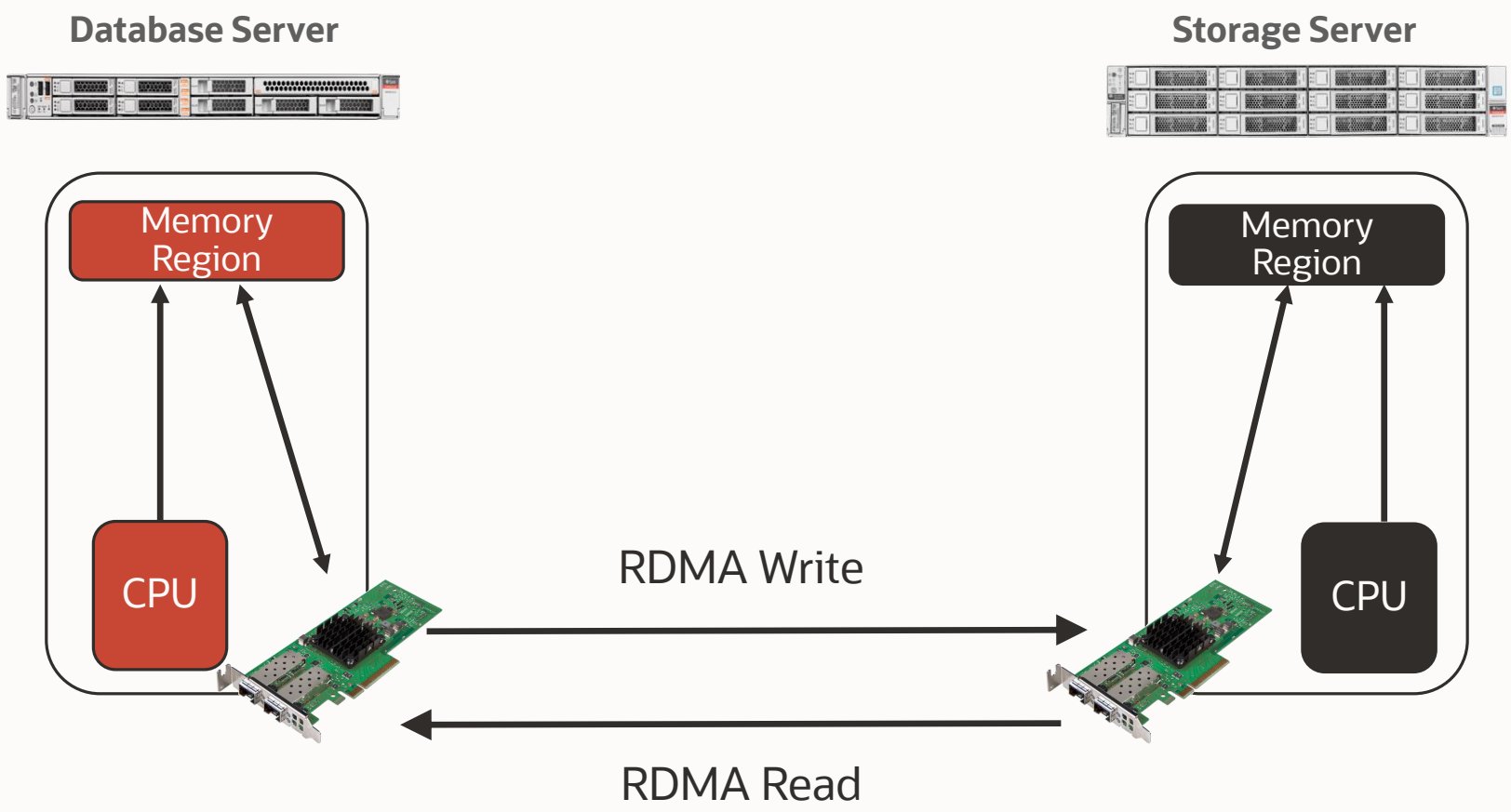
- Persistent memory is a new silicon technology
  - Capacity, performance, and price are between DRAM and flash
- Intel® Optane™ Persistent Memory:
  - Reads at memory speed – much faster than flash
  - Writes survive power failure unlike DRAM
- Requires sophisticated algorithms to maintain integrity of data on PMEM during failures
  - Call special instructions to flush data from CPU cache to PMEM
  - Complete or backout sequence of writes interrupted by a crash



Do you recognize the dynamic duo?



Remote Direct Memory Access (RDMA)



Introducing RDMA over Converged Ethernet

RDMA over Converged Ethernet is a protocol that runs InfiniBand RDMA software on top of **Ethernet**

- Same software at upper levels of network protocol stack
- InfiniBand packets sent as ethernet UDP packets at low level

Exadata RoCE provides **RDMA speed** and reliability on **Ethernet** fabric

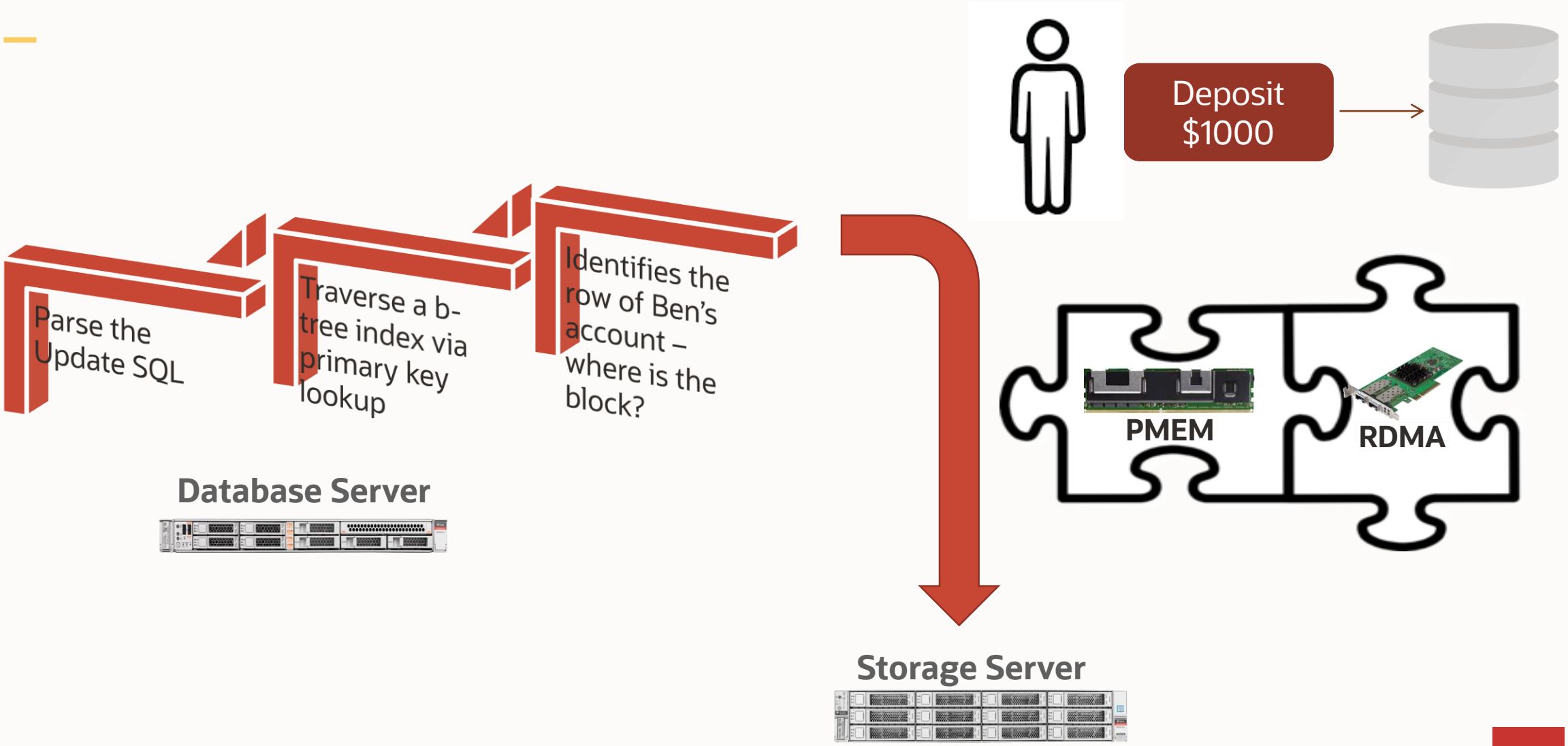
- 100Gb/s link throughput
- Zero packet loss messaging
- Prioritization of critical database messages

Defined by an open consortium IBTA, developed in open-source Linux

- Supported by major network card and switch vendors

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

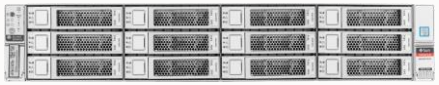
How do they conquer the IO Cliff?



Drop in solution: Flash -> PMEM?

Database Server

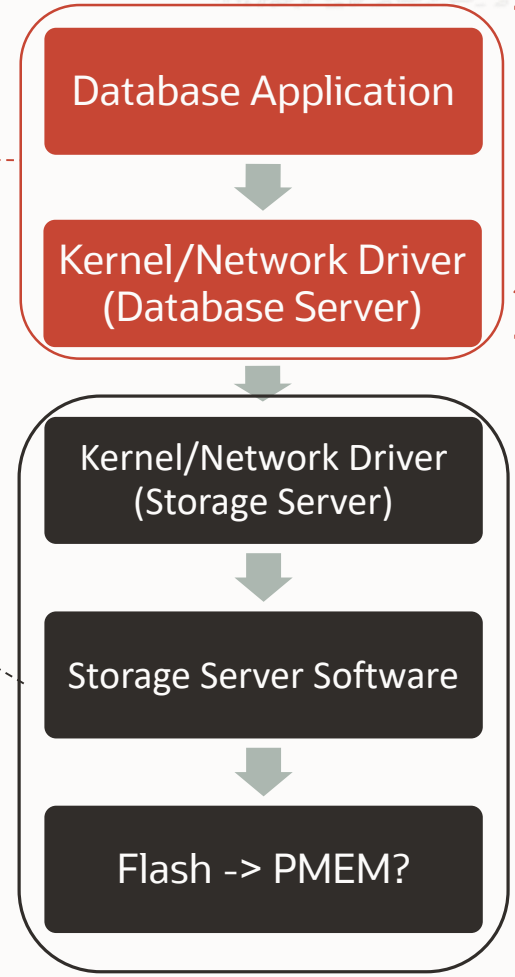
Storage Server



PMEM Cache

What happens to the IO Cliff?

~100 usec



- User->Kernel & Kernel-> User context switches
- Network IO interrupt processing
- IO software stack

Network Round Trip

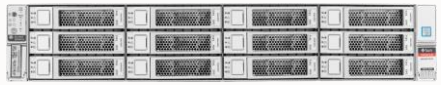
- User->Kernel & Kernel-> User context switches
- Network IO interrupt processing
- IO software stack

~100 usec

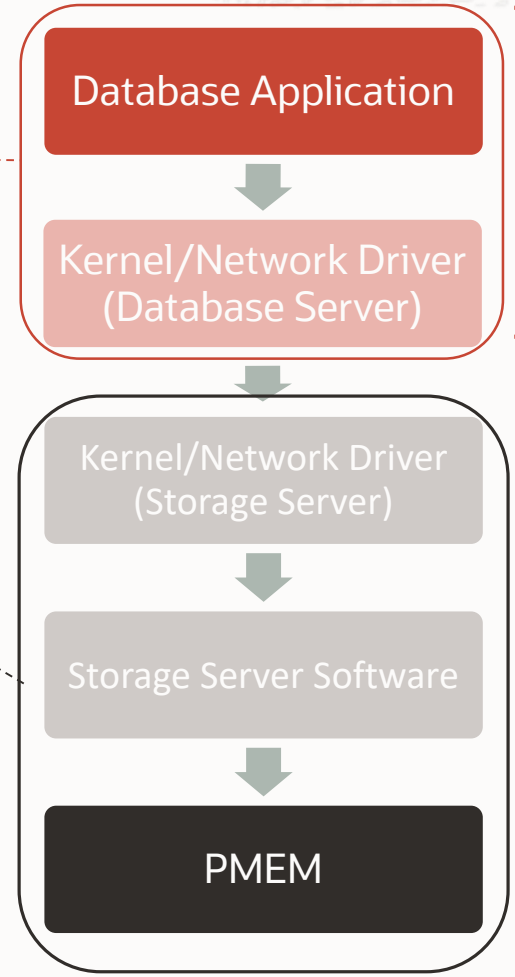
How about a radical approach: RDMA to PMEM?

Database Server

Storage Server



PMEM Cache



- User->Kernel & Kernel-> User context switches
- Network IO interrupt processing
- IO software stack

RDMA over 100Gb/s RoCE

- User->Kernel & Kernel-> User context switches
- Network IO interrupt processing
- IO software stack

What happens to the IO Cliff?

<19 usec

~100 usec

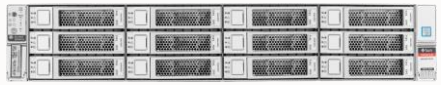
10x Faster Random Read!

RDMA Read from PMEM Cache in Storage Server

Miss in Buffer Cache:
Need to fetch the data from storage

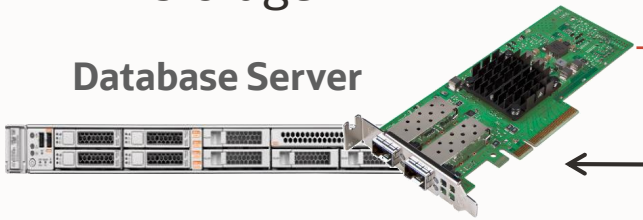
1. DB issues RDMA Read from PMEM Cache

Storage Server



PMEM Cache

Database Server



8K Data Block returned to DB from PMEM
- No software involved

User	Account Balance
…	…
Ben	$2000

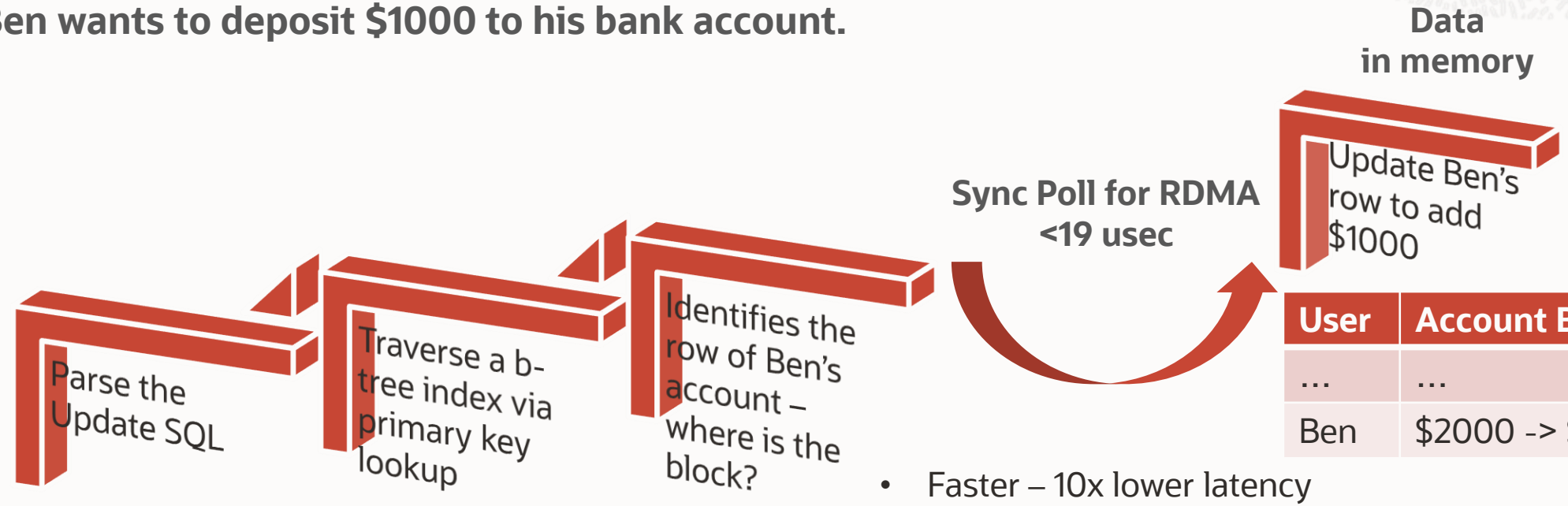
…
PMEM Cache Line
PMEM Cache Line
PMEM Cache Line

User	Account Balance
…	…
Ben	$2000

PMEM + RDMA

Transforming IO bound application to near memory performance

Ben wants to deposit $1000 to his bank account.

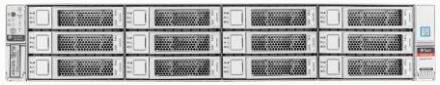


User	Account Balance
…	…
Ben	$2000 -> $3000

- Faster – 10x lower latency
- Cheaper – less CPU than context switch

Database Server

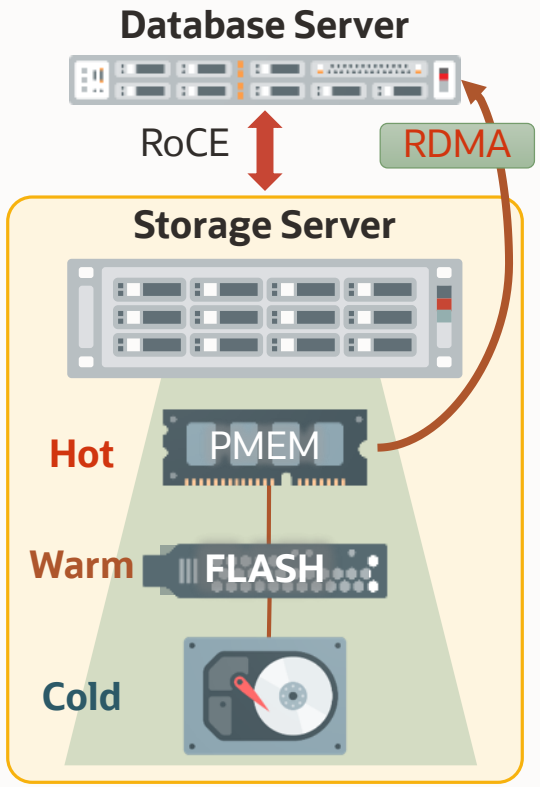
Storage Server



User	Account Balance
…	…
Ben	$2000

Exadata X8M With Persistent Memory

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



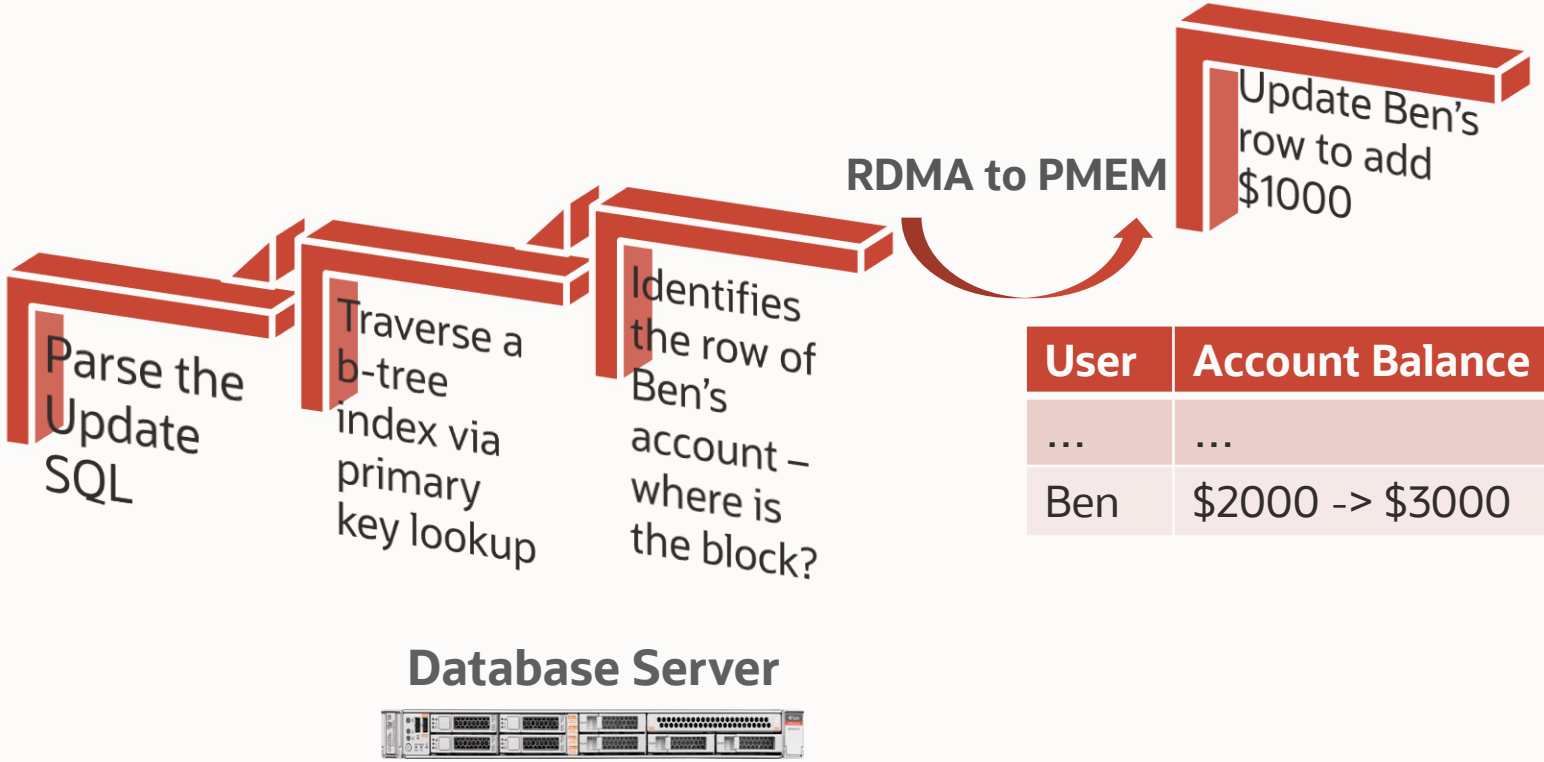
- Shared capacity across Databases
- Smart Capacity management for maximum performance
  - Primaries on persistent memory, secondaries on flash
  - Secondaries moved into persistent memory when they become active primaries
- Management server on storage server automatically manages persistent memory
  - Creates regions, namespaces, DAX devices etc as needed
  - Persistent memory replacement follows standard hardware replacement process
- End to End Security
  - Persistent memory is not visible to the OS on the database server
  - Protected end to end while serving RDMA

Meet Ben's Transaction – time to commit?

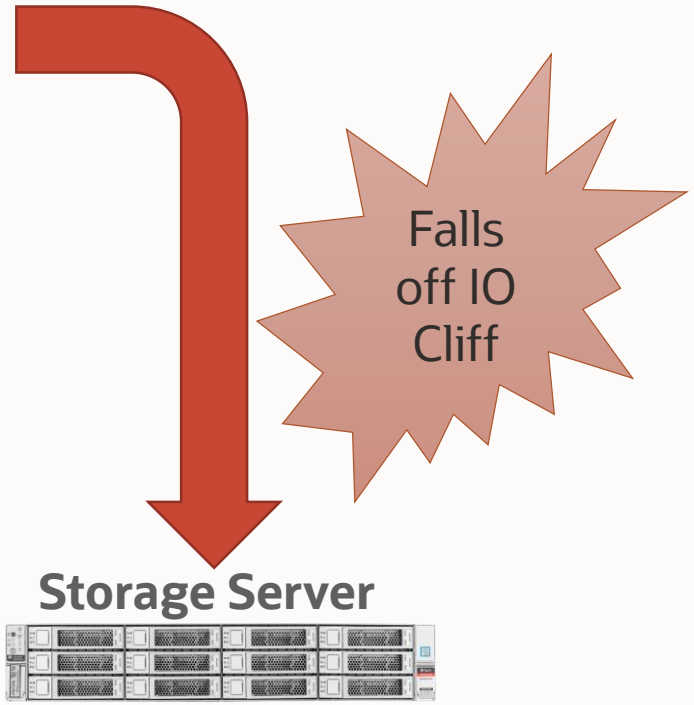
Putting the **D(urability)** into Database Transaction ACID Properties

Ben wants to deposit $1000 to his bank account.

How do we commit a transaction?



User	Account Balance
…	…
Ben	$2000 -> $3000

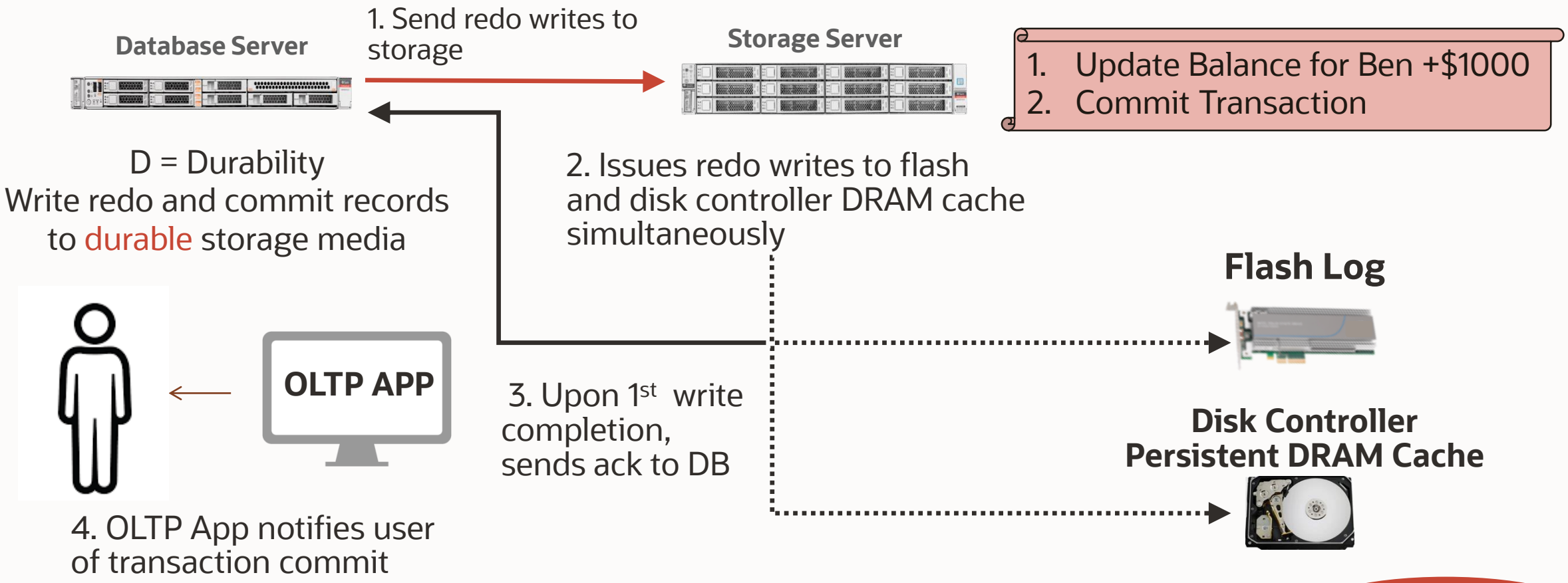


1. Update Balance for Ben +$1000
2. Commit Transaction

OLTP Challenge #2 -

Can lightning strike the same place twice?
What is the IO cliff for redo log writes?

Challenge #2 – Redo Log Writes



How long did the Super Critical OLTP App wait for the log write?

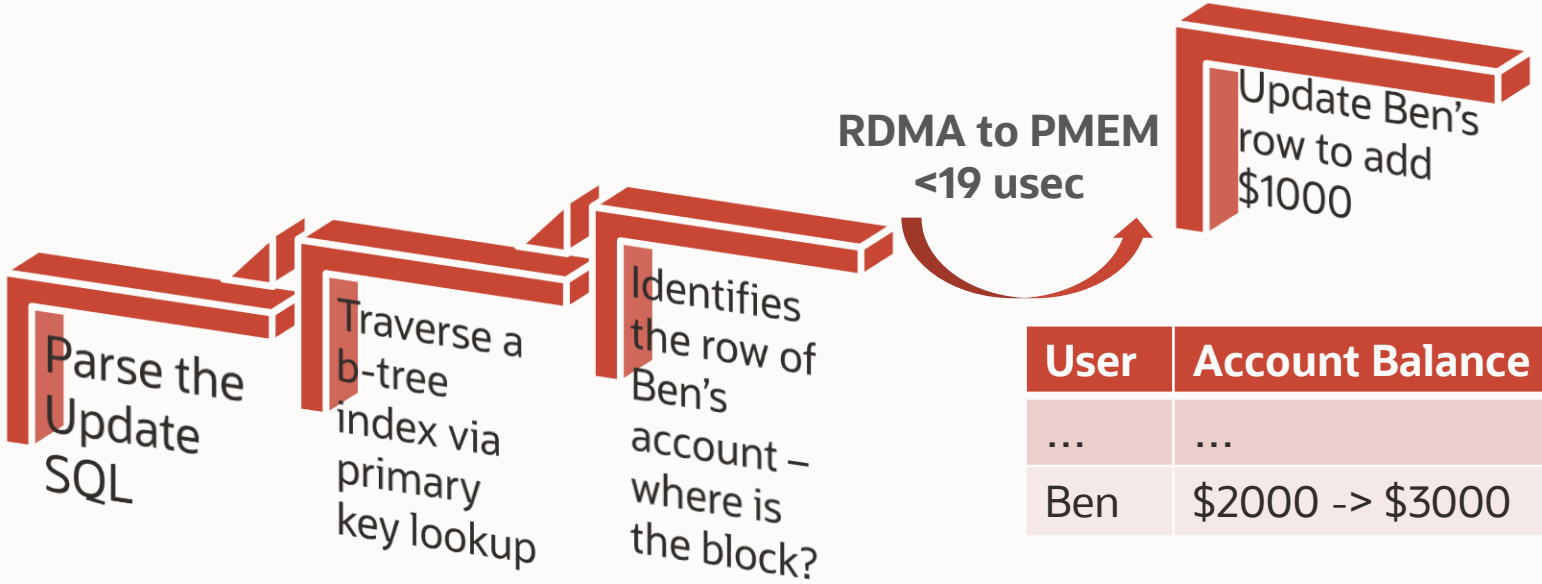
~200 usec

Can PMEM + RDMA come to the rescue again?

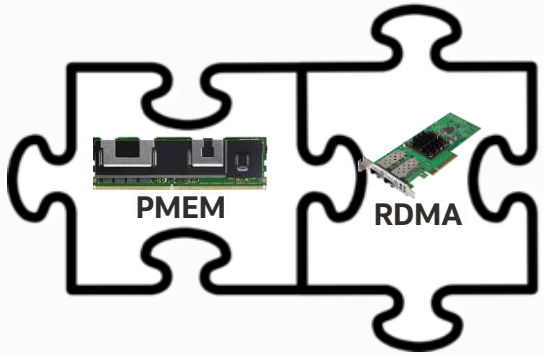
Lightning strikes the same place twice!

Ben wants to deposit $1000 to his bank account.

How do we not fall off the IO cliff?

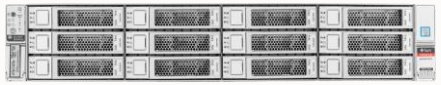


User	Account Balance
…	…
Ben	$2000 -> $3000



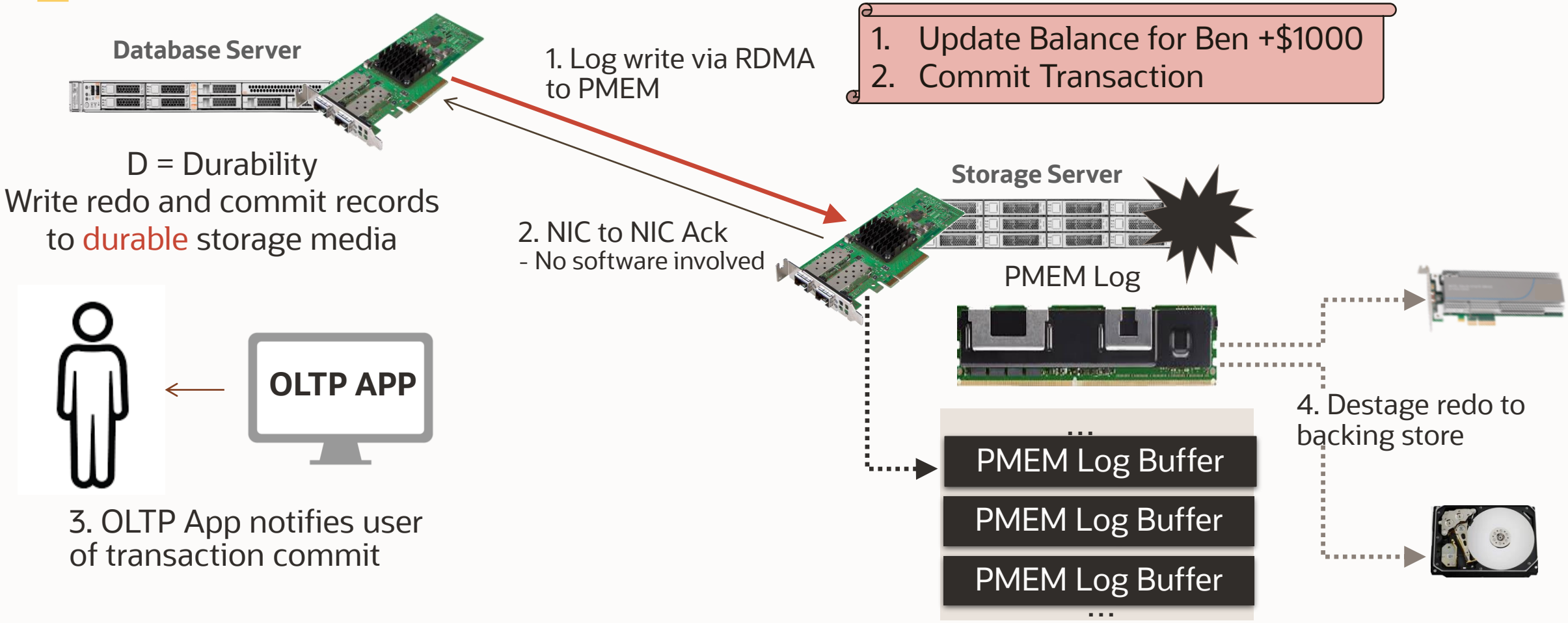
Database Server

Storage Server

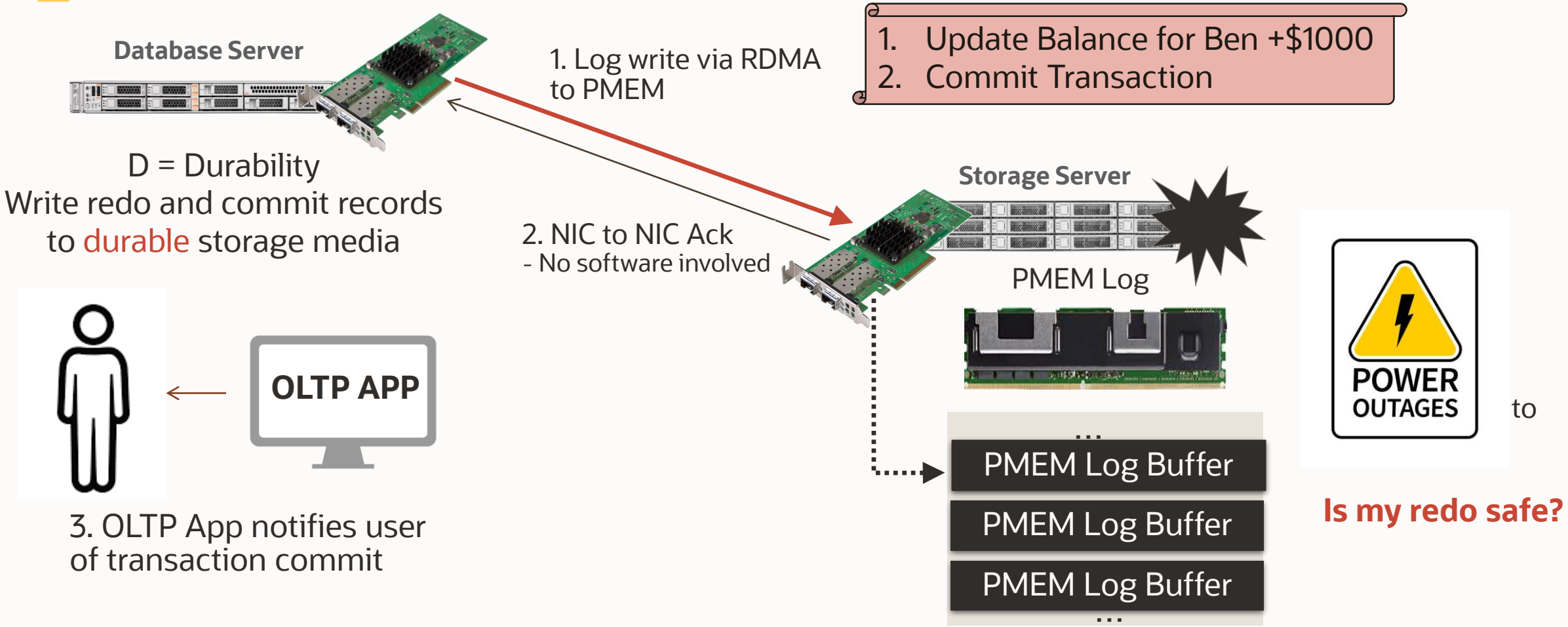


1. Update Balance for Ben +$1000
2. Commit Transaction

Redo Log RDMA write to PMEM in Storage Server

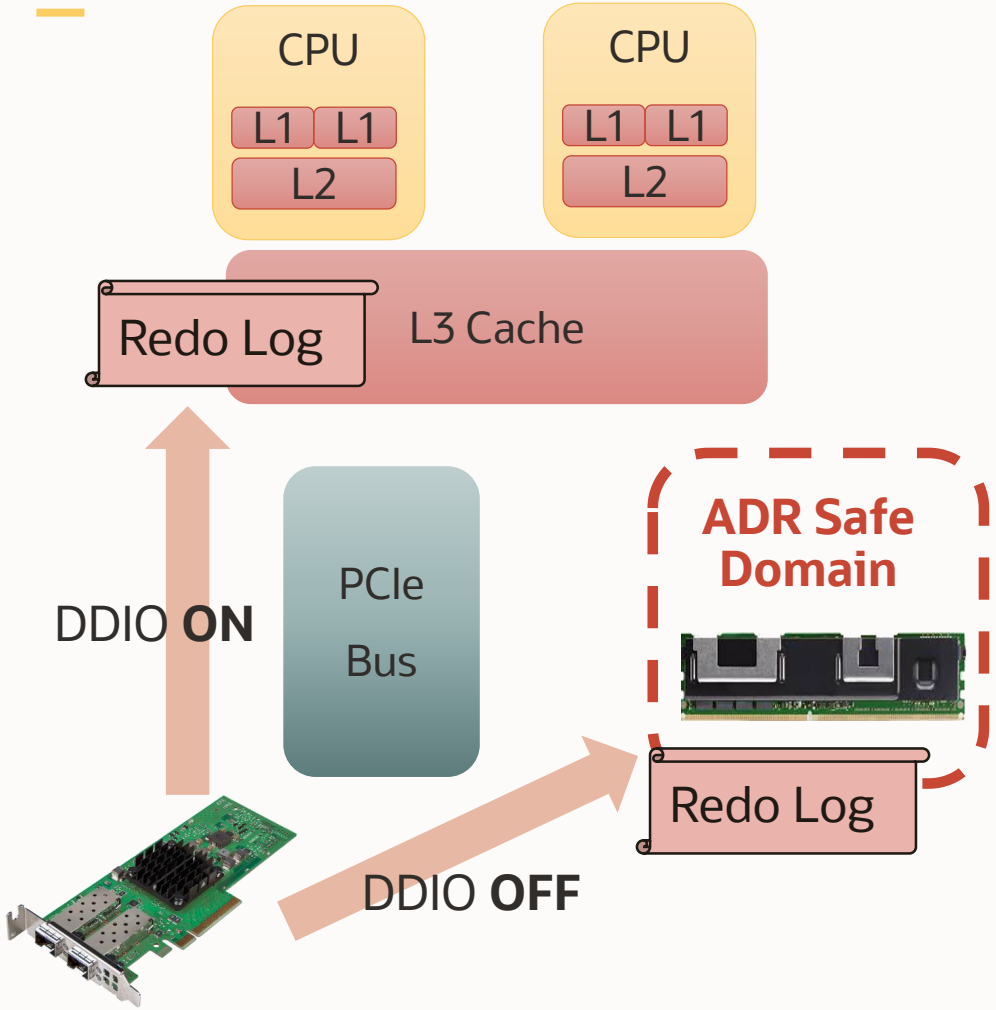


Redo Log RDMA write to PMEM in Storage Server



Which Begs the Question:

Are all writes to PMEM automatically persistent across power failures?



Intel Data Direct IO (DDIO) enabled by default

- Performs Write Allocate/Update into L3 Cache
- Places data outside Asynchronous DRAM Refresh (ADR) Safe Domain
  - Not Durable upon Power Fail

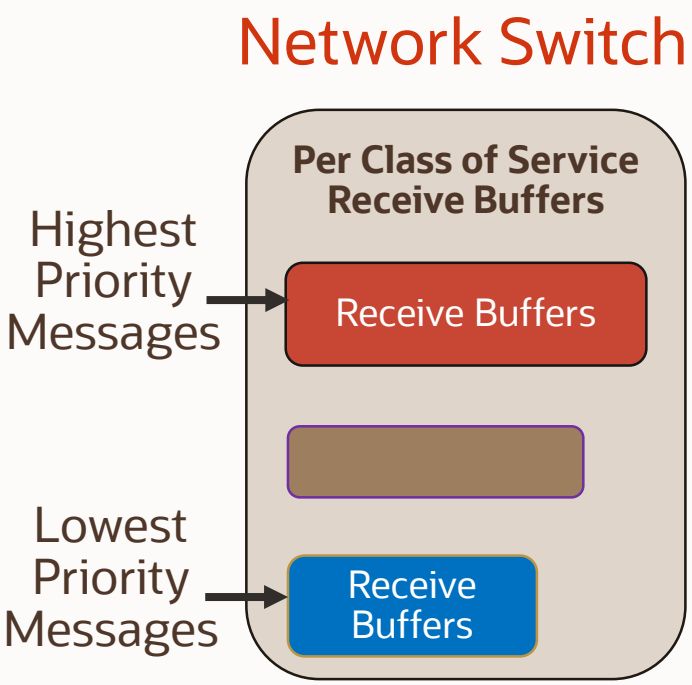
How to ensure redo log writes are persistent in PMEM?

Turn off DDIO for Network Card PCI Slot in BIOS

- Global Knob per PCI Device:
  - Affects DMA to both DRAM and PMEM
- Be aware of potential impacts to other network receive performance due to LLC misses

What happens if the network is congested?

Smart network prioritization of redo log writes!



Network prioritization for latency-sensitive DB algorithms

- Ensures that messages requiring low latency are not slowed by high throughput messages
  - Low latency: transaction commit, redo log writes
  - Medium priority: OLTP data reads
  - High throughput: backup, reporting, batch workloads

RoCE Class of Service (CoS)

- Allows packets to be sent on multiple classes of service, each with separate network buffers for independence

Exadata uniquely chooses the best Class of Service for each database message

Remember Ben's Transaction?

What happens to the IO Cliff?

8x Faster Log Writes!

Ben wants to deposit $1000 to his bank account.

Database Server

Parse the Update SQL

Traverse a b-tree index via primary key lookup

Identifies the row of Ben's account – where is the block?

RDMA Read from PMEM Cache < 19 usec

Update Ben's row to add $1000

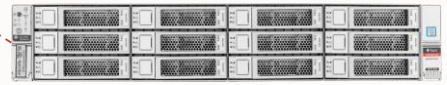
User	Account Balance
…	…
Ben	$2000 -> $3000

RDMA Write to PMEM Log 10's usec

Write the redo log and commit the transaction

1. Update Balance for Ben +$1000
2. Commit Transaction

Storage Server



Are we done?

Yes and No

Ben wants to deposit $1000 to his bank account – YES!

Database Server

Parse the Update SQL

Traverse a b-tree index via primary key lookup

Identifies the row of Ben's account – where is the block?

RDMA Read from PMEM Cache < 19 usec

Update Ben's row to add $1000

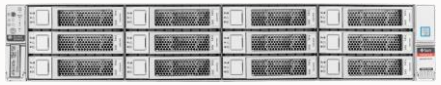
User	Account Balance
…	…
Ben	$2000 -> $3000

RDMA Write to PMEM Log 10's usec

Write the redo log and commit the transaction

1. Update Balance for Ben +$1000
2. Commit Transaction

Storage Server



PMEM Cache

User	Account Balance
…	…
Ben	$2000 -> $3000

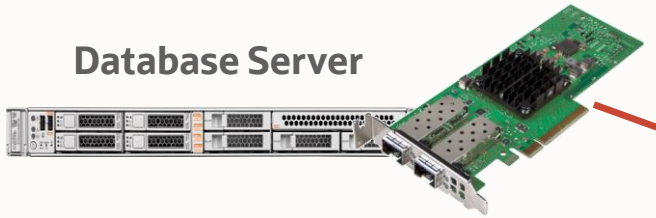
Not yet – Database needs to checkpoint updated data to storage!

OLTP Challenge #3 -

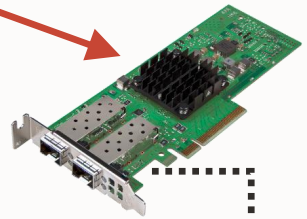
What is so difficult about writing data back to PMEM Cache in Storage Server?

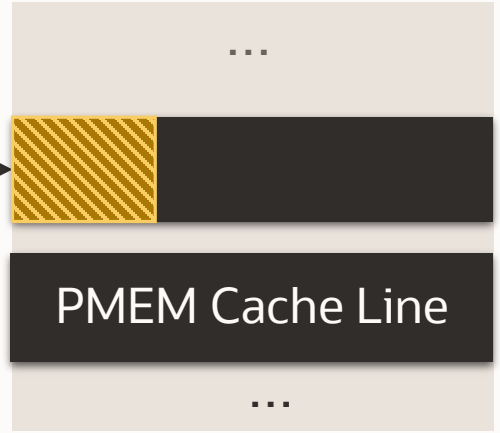
What is splinch?

No torn blocks!



User	Account Balance
…	…
Ben	$2000 -> $3000





PMEM has 8-byte Atomic Write Guarantee

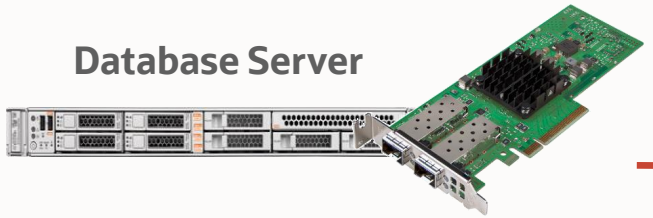
- Database blocks are much larger – 8K being the typical size
- No power failure should result in torn blocks!

Is RDMA to PMEM a good choice here?

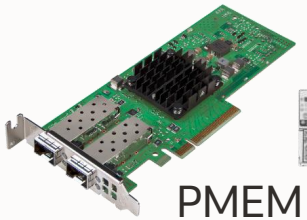
NO!!!

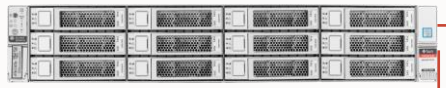
How do we prevent splinch?

3 D's - Destination, Determination, and Deliberation



User	Account Balance
…	…
Ben	$2000 -> $3000





PMEM

Destination:
DB sends the 8K data block to storage via regular messaging

Determination:
Storage software recognizes that the data needs to be updated on PMEM Cache

Deliberation:
Storage software guarantees atomic database block writes to PMEM via staging writes

What happens during a power fail?

- Staging buffer recovery!

User	Account Balance
…	…
Ben	$2000 -> $3000

1. Write to 8K Staging Buffer on PMEM

…

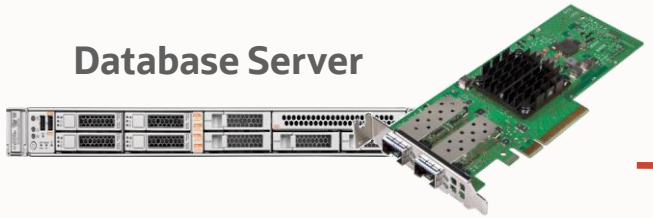
PMEM Cache Line

User	Account Balance
…	…
Ben	$2000 -> $3000

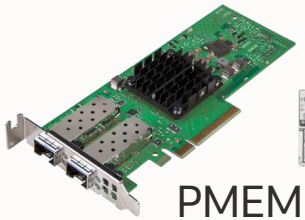
2. Write to 8K block in PMEM Cache

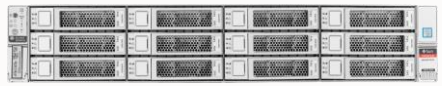
How do we prevent splinch?

3 D's - Destination, Determination, and Deliberation

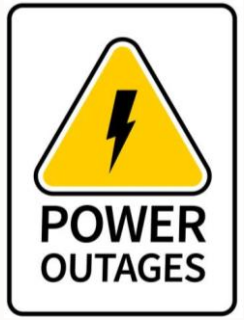


User	Account Balance
…	…
Ben	$2000 -> $3000





PMEM



Destination:
DB sends the 8K data block to storage via regular messaging

Determination:
Storage software recognizes that the data needs to be updated on PMEM Cache

Deliberation:
Storage software guarantees atomic database block writes to PMEM via staging writes

What happens during a power fail?

- Staging buffer recovery!

User	Account Balance
…	…
Ben	$2000 -> $3000

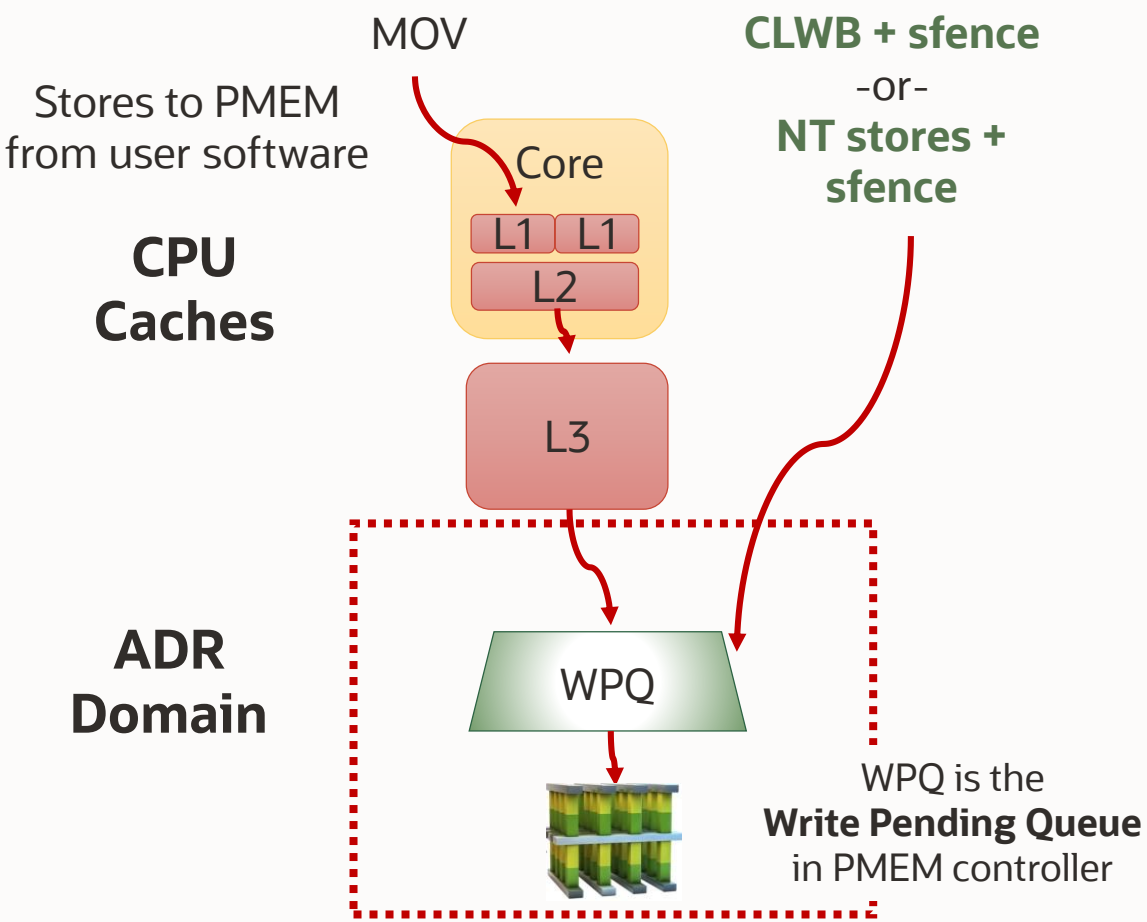
…

PMEM Cache Line

User	A e
…	…
Ben	$2000 -> $3000

Re-apply the 8K staging buffer to PMEM Cache

How do we ensure persistence across power failure?



Stores in CPU Cache are NOT Persistent

- Use Non-Temporal Stores to Bypass CPU Cache

  _mm512_stream_si512 or
  _mm_stream_si64
  + sfence

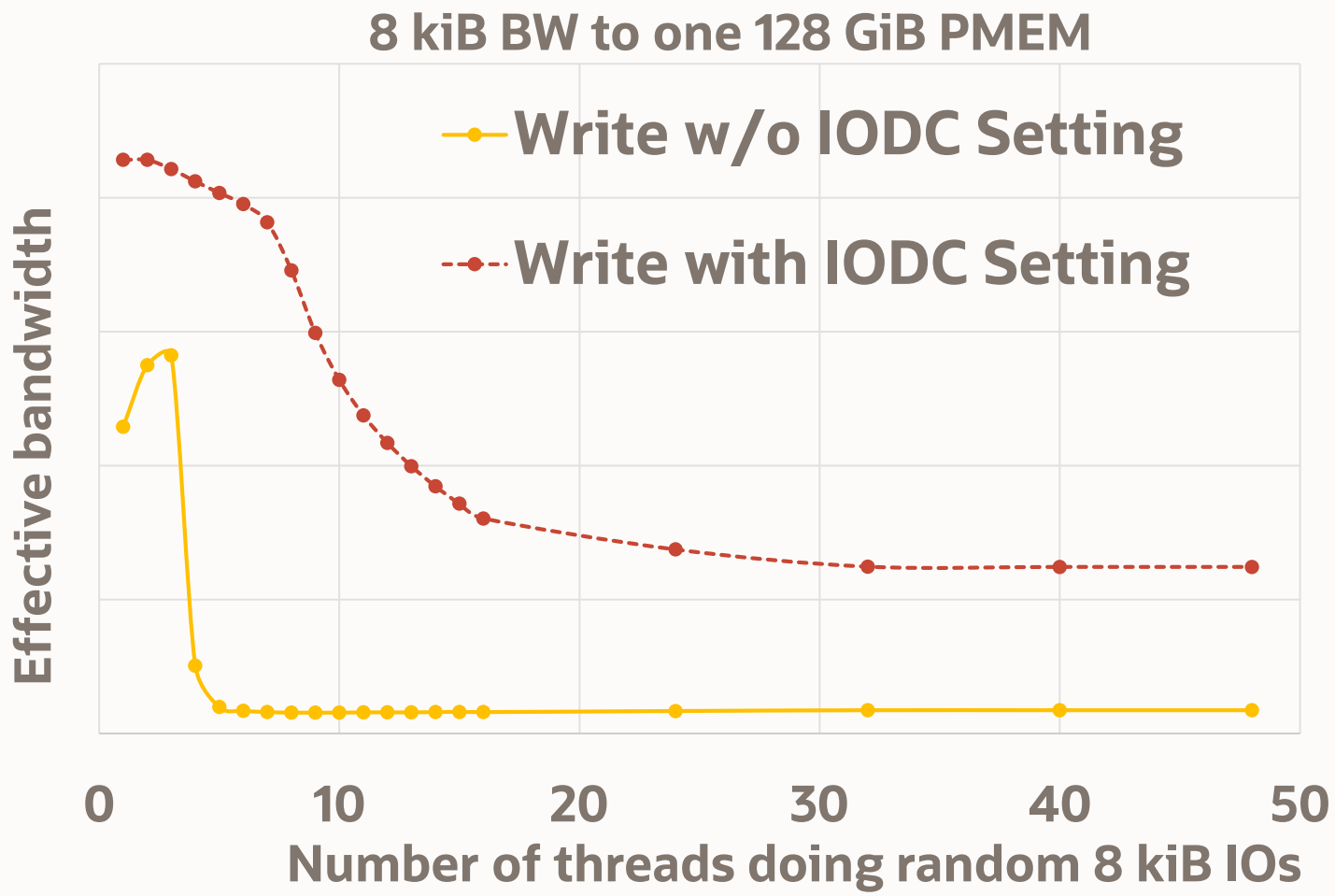
- Use CLWB to Flush into ADR Domain

  _mm_clwb_ + sfence

We use **NT stores** for PMEM Cache:

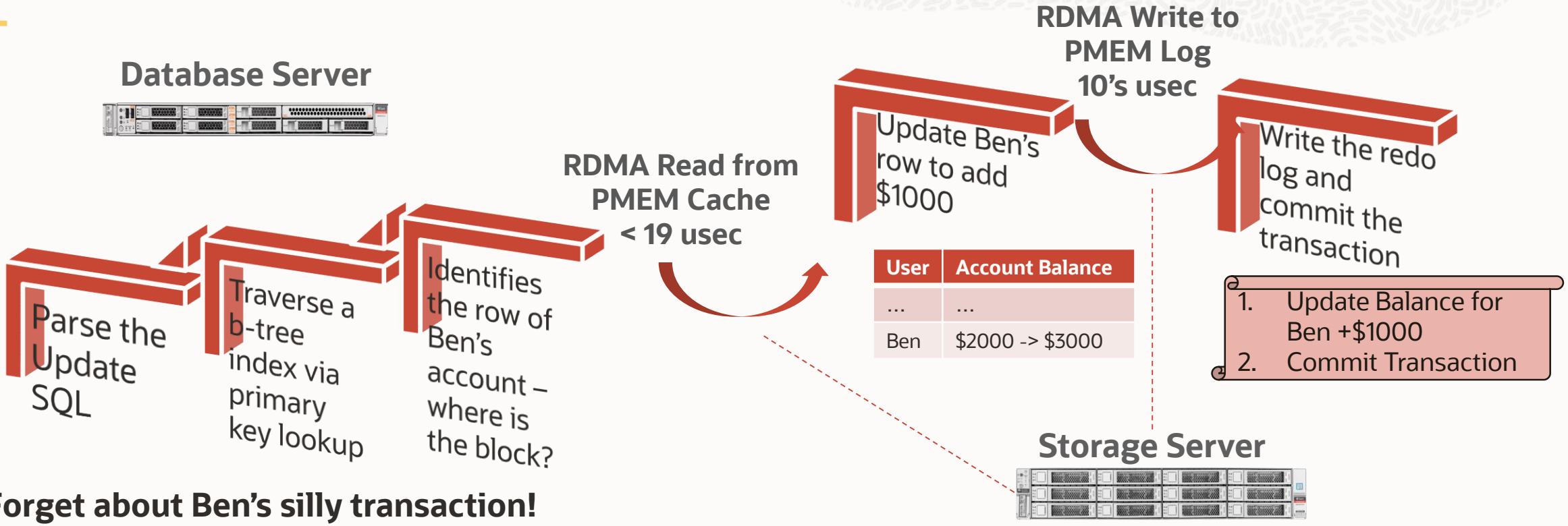
- Faster for sequential 8K data writes
- Avoid polluting CPU Cache

How do we achieve better PMEM write scaling?



- Previously, write scales poorly with concurrent writes
  - Especially with writes from remote core
- IO Directory Cache contention
  - Specific IODC setting reduces contention
  - Still room for future improvements
- Each thread slows down writes from another thread
- Best performance with fewer threads doing IOs

Take Away – Do not fall off the IO Cliff!

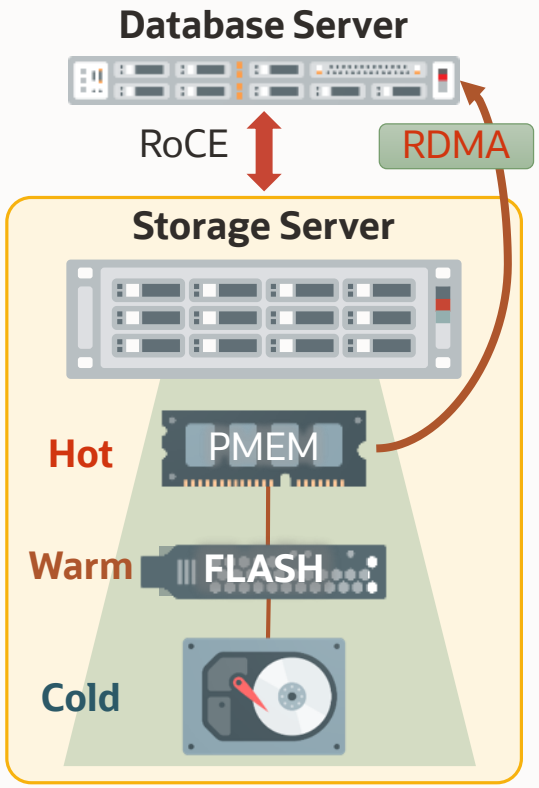


Forget about Ben's silly transaction!
What is a truly Super Critical OLTP Application in the Real World?

- **Fraud tracking**
- **Real-time analytics upon a click on the screen**
- **...**

OLTP on Exadata

How do we harness the power of PMEM?



How to have a cake and eat it too?

- >99% of PMEM used for PMEM Cache – 1.5TB per server
- <1% of PMEM used for PMEM Log – 10G per server

New Leading-Edge RDMA fabric

- Based on 100 Gb/sec Converged Ethernet

RDMA to PMEM in Storage

- 10X better transaction processing IO latency @ <19 usec
- 8X faster log writes for faster commit processing
- 16 Million read IOPS on a full rack of ultimate OLTP database machine

PMEM automatically tiered and shared across Databases

Our mission is to help people see data in new ways, discover insights, unlock endless possibilities.