ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

OPERATING SYSTEMS & PARALLEL COMPUTING

Final Exam Preparation

Final Exam

- Parallel Computing — 20 Questions
  - MPI / OpenMP / PThreads
  - Shared / Distributed / Hybrid Memory
- Big Data — 10 Questions
- GPU / NVIDIA — 10 Questions
- Quantum — 10 Questions
- HPC — 10 Questions
- Cloud Solutions — 10 Questions
  - AWS, Azure, Google, Oracle, IBM

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Questions Examples

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

What is Parallel Computing?

Answer: Parallel Computing resembles the study of designing algorithms such that the time complexity is minimum. Thus the speed up factor is taken into consideration.

Differentiate between SIMD and MIMD?

Answer: In architectures referred to as single instruction stream, multiple data stream (SIMD), a single control unit dispatches instructions to each processing unit. In an SIMD parallel computer, the same instruction is executed synchronously by all processing units.

Computers in which each processing element is capable of executing a different program independent of the other processing elements are called multiple instruction stream, multiple data stream (MIMD) computers.

SIMD computers require less hardware than MIMD computers because they have only one global control unit. SIMD computers require less memory because only one copy of the program needs to be stored.

In contrast, MIMD computers store the program and operating system at each processor.

Differentiate between UMA and NUMA.

Answer: If the time taken by a processor to access any memory word in the system (global or local) is identical, the platform is classified as a uniform memory access (UMA) multicomputer.

On the other hand, if the time taken to access certain memory words is longer than others, the platform is called a non-uniform memory access (NUMA) multicomputer.

What is Cache Coherence?

Answer: In shared address space platform ensuring that concurrent operations on multiple copies of the same memory word have well-defined semantics is called cache coherence.

What is PRAM Model?

Answer: A Model of computation (the Random Access Machine, or RAM) consists of p processors and a global memory of unbounded size that is uniformly accessible to all processors. All processors access the same address space. Processors share a common clock but may execute different instructions in each cycle. This ideal model is also referred to as a parallel random access machine (PRAM).

What is Snoopy cache system?

Answer: Snoopy caches are typically associated with multiprocessor systems based on broadcast interconnection networks such as a bus or a ring. In such systems, all processors snoop on (monitor) the bus for transactions. This allows the processor to make state transitions for its cache-blocks.

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

What is MPI?

Answer: Message Passing Interface (**MPI**) is a standardized and portable message-passing system designed by a group of researchers from academia and industry to function on a wide variety of parallel computers. The standard defines the syntax and semantics of a core of library routines useful to a wide range of users writing portable message-passing programs in the C programming language.

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Networking

RDMA for Extreme Performance

- Remote Direct Memory Access (RDMA) is the ability for one computer to access data from a remote computer without any OS or CPU involvement
  - Network card directly reads/writes memory with no extra copying or buffering and very low latency

- RDMA is an integral part of the Exadata high-performance architecture enabling:
  - High throughput and low-CPU usage for large data transfers
  - Unique Direct-to-Wire Protocol to deliver 3x faster inter-node OLTP cluster messaging
  - Unique Smart Fusion Block Transfer that eliminates log write on inter-node block move
  - Unique RDMA protocol to coordinate transactions between nodes

RoCE – RDMA Over Converged Ethernet

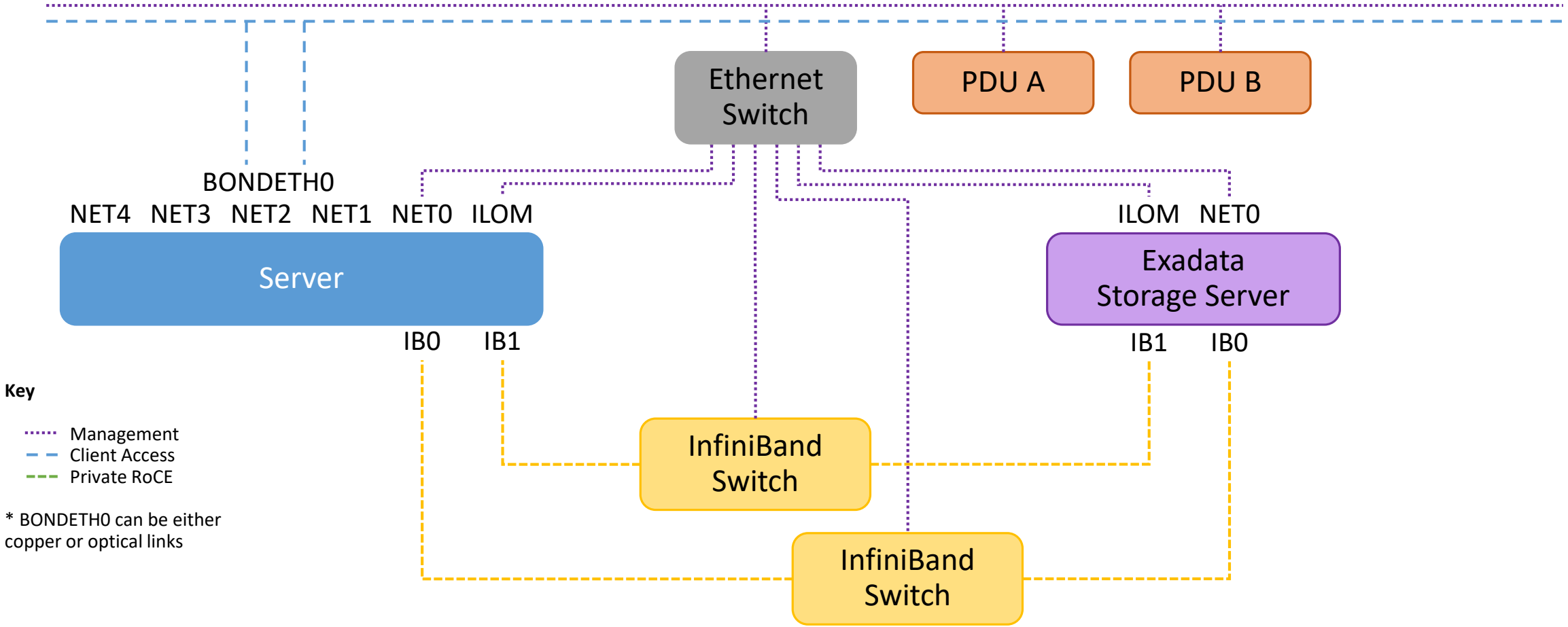
- RDMA over Converged Ethernet is a protocol that runs InfiniBand RDMA software on top of Ethernet
  - Same software at upper levels of network protocol stack
  - InfiniBand packets sent as ethernet UDP packets at low level
- RoCE on Exadata supports all Exadata RDMA optimizations
- RoCE enables scalability and volume of Ethernet with speed of RDMA
- Defined by an Open Consortium
  - InfiniBand Trade Association (IBTA)
  - Developed in open-source and maintained in upstream 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

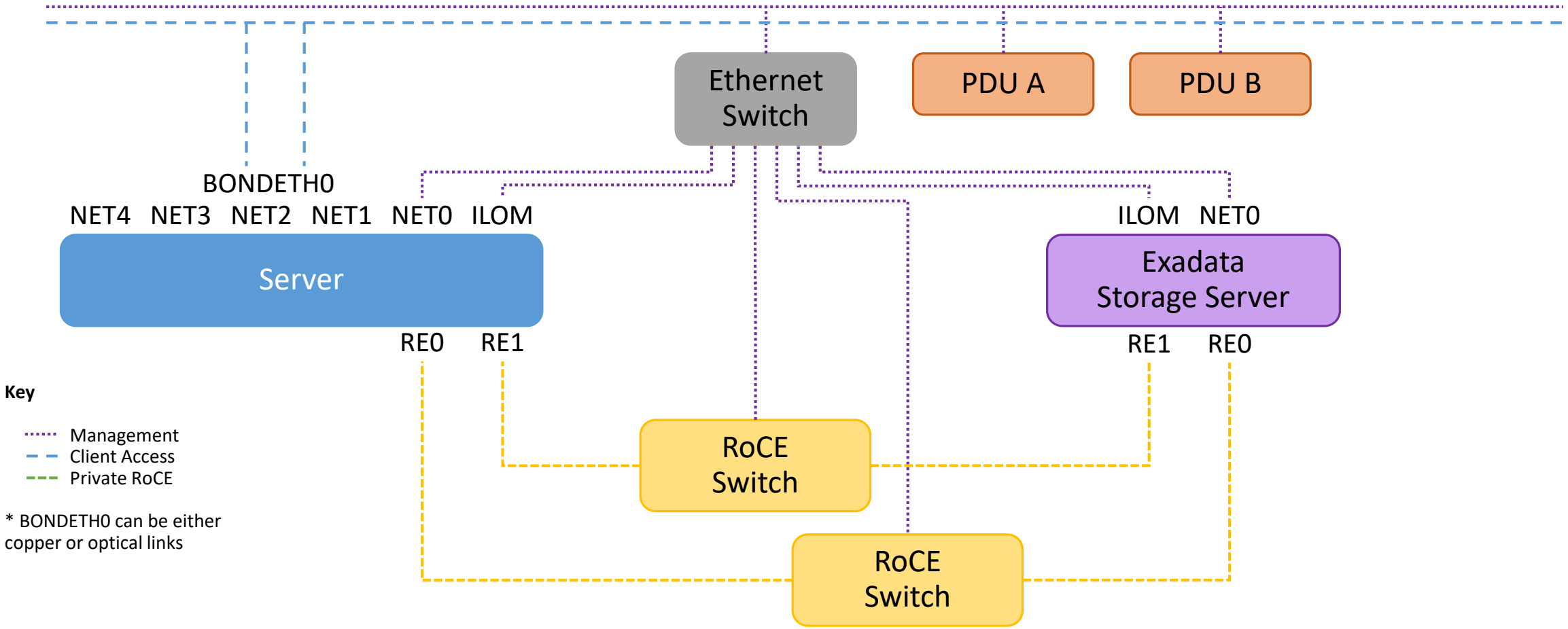
New RoCE Internal Network Fabric

- InfiniBand was the only viable RDMA capable network at the inception of Exadata, but now Ethernet has caught up
- Exadata RoCE provides RDMA speed and reliability on Ethernet fabric
  - 100Gb throughout
  - Zero packet loss messaging
  - Prioritization of critical database messages
  - Latest KVM based virtualization

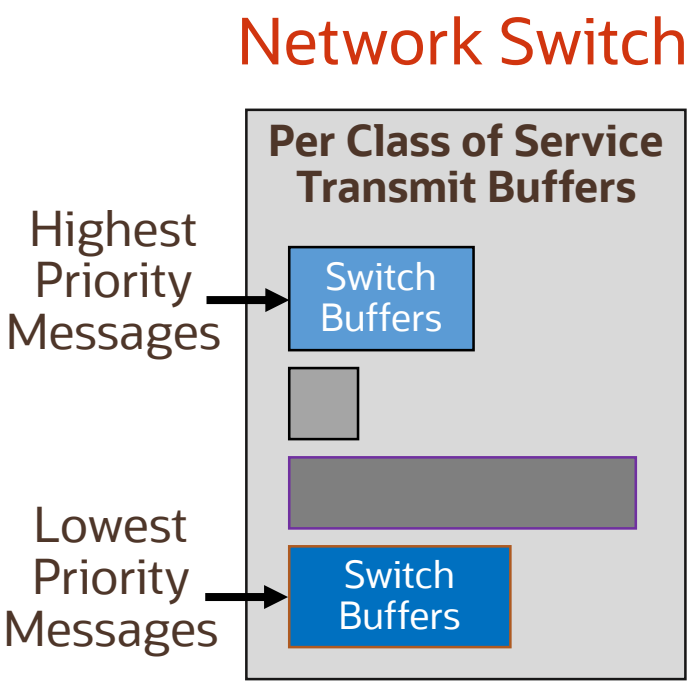
Network Architecture - InfiniBand



Network Architecture - RoCE

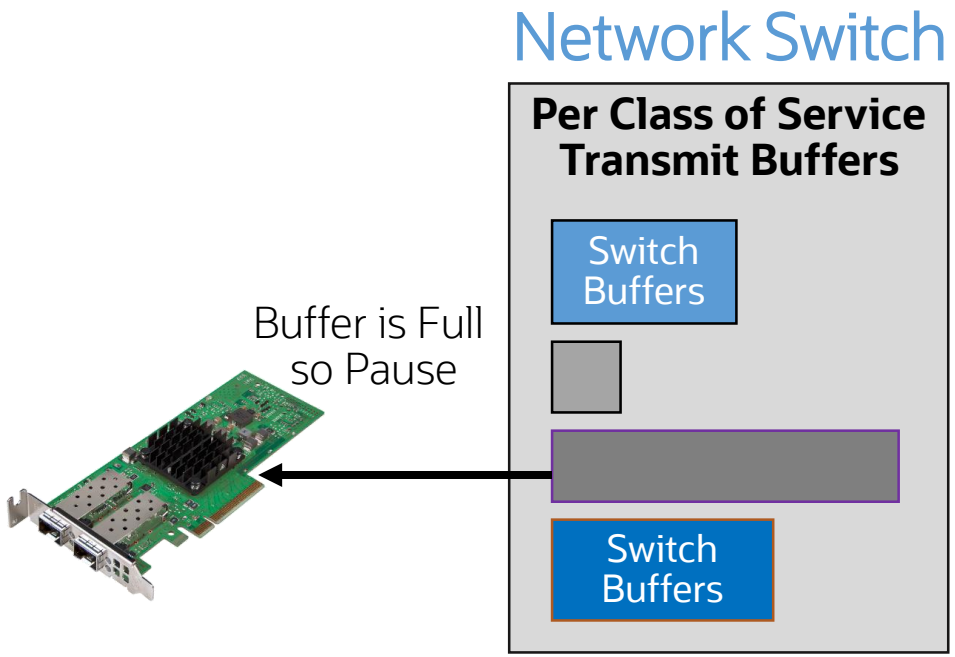


RoCE – High Priority Networking



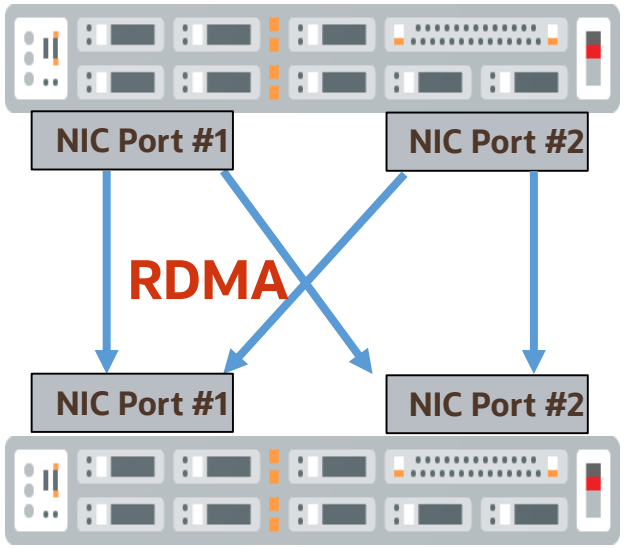
- Network prioritization for latency sensitive DB algorithms
  - Ensures that messages requiring low latency are not slowed by high throughput messages
    - Examples: cluster heartbeat, transaction commit, cache fusion
  - High throughput examples: backup, reporting, batch
- RoCE Class of Service (CoS)
  - Allows packets to be sent on multiple classes of service, each with separate network buffers for independence
- Server uniquely chooses the most optimal Class of Service for each message

RoCE – Avoiding Packet Loss



- Packet loss is usually caused by congestion
  - Packets sent faster than receiver or switches can process
  - Less common sources – switch failure, link failure
- Conventional Ethernet silently drops packets and expects retransmit if sends are too fast
  - Packet Drop causes huge hit to latency and throughput
- RoCE avoids packet drops using:
- RoCE Priority-based Flow Control (PFC)
  - RoCE switch tells sender to pause if switch buffer is full
- RoCE Explicit Congestion Notification (ECN)
  - RoCE switch marks packet flow as too fast, telling source to slow down packet sends

RoCE Instant Failure Detection



- Server uses frequent heartbeat messages between nodes to detect possible server failure
- Server failure detection normally requires long timeout to avoid false server evictions from cluster
  - Hard to quickly distinguish between slow response to heartbeat due to very high CPU load, and server failure
- Server uses RDMA to quickly confirm server failure
  - RDMA uses hardware, so remote ports respond even if software is slow
  - 4 RDMA Reads are sent to suspect server
  - Across all combinations of source and target ports
  - If all 4 RDMAs fail, server is evicted from cluster

InfiniBand vs RoCE

Feature	InfiniBand	RoCE
Fabric Management	Centralized using Subnet Manager	Decentralized Autonomous Fabric Management
Speed	40Gb/s	100Gb/s
Lossless Network	✓	✓
Multi-rack*	✓	✓
All Servers Performance Features	✓	✓
Server Virtualization	Xen o other	KVM
Instant Failure Detection	Via Subnet Manager Query	Via RDMA Queries

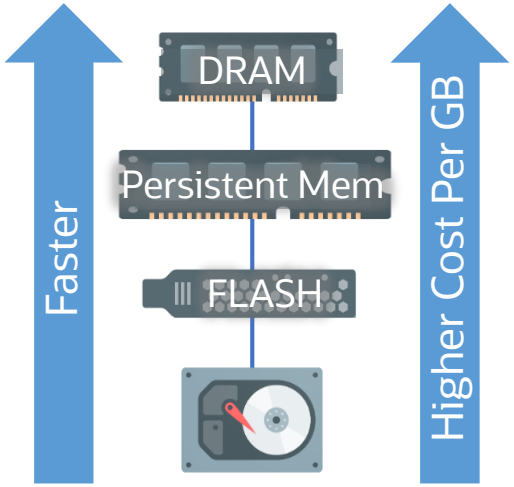
* Multi-racking between InfiniBand and RoCE is not possible

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Memory

New Persistent Memory

- Persistent memory is a new silicon technology
  - Capacity, performance, and price are between DRAM and flash
- Intel® Optane™ DC Persistent Memory:
  - Reads at memory speed – much faster than flash
  - Writes survive power failure unlike DRAM
- Database Server implements 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



SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY