IO-DIMM: A Low-Latency, Power-Efficient Near-Memory I/O interface

Presenter: Igor Sharovar, CTO, Truememorytechnology LLC

Igor.Sharovar@truememorytechnology.com

Credit to Intel Optane and its impact on the industry

Intel Optane: A Pioneer in Persistent Memory

- **What It Achieved:**
- Brought persistent memory closer to DRAM speeds
- Introduced new memory tiering concepts to data centers
- *Proved the value of byte-addressable, non-volatile memory*
- Stimulated industry-wide interest in hybrid memory solutions
- **Key Limitations Observed:**
- Proprietary protocol (DDRT) limited ecosystem adoption
- Ultimately discontinued due to market cost/benefit trade-offs

Influence of NVDIMM-P standard

NVDIMM-P: A Platform for Persistent Memory Standards

- **Positive impact*:***
- Defined interoperable standards for persistent memory on DDR bus
- Opened door for hybrid DRAM/NAND solutions
- Validated use of DDR channels for persistent memory
- *Simulated researches in near-memory computing*
- **Key limitation:**
- The NVDIMM-P standard introduces additional complexity to the memory subsystem, particularly in the memory controller. While it reuses the DDR physical interface for electrical compatibility, it defines a proprietary protocol that is not compatible with existing DDR standards.

The researches to extend the use of Intel Optane and NVDIMM-P standard

Embedded systems

- The research project at Waseda University in Tokyo exploring the use Intel Optane memory in embedded systems[1].

Networking

- “NetDIMM: Low-Latency Near-Memory Network Interface Architecture”, University of Illinois Urbana-Champaign[2].

Near-memory processing

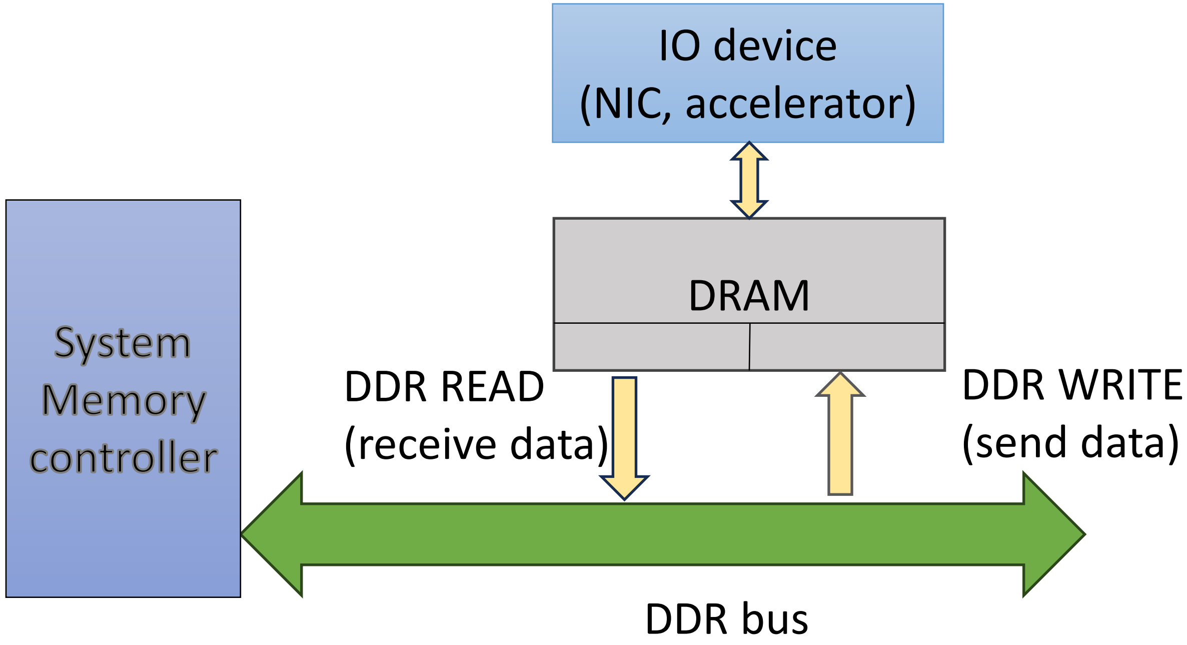
- “AxDIMM: Near-Memory Processing”, Facebook, Washington University in St. Lous, Samsung [3].

IO-DIMM solution

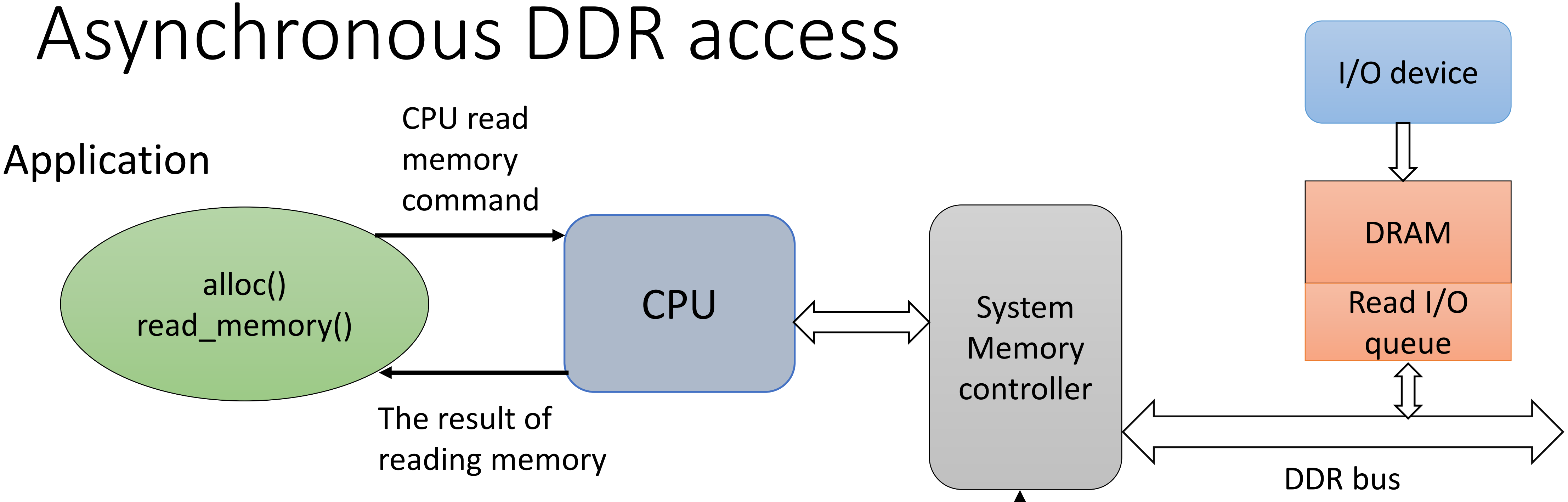
- Provide an interface for connecting high-performance I/O devices to DDR4/5 busses. Examples of such devices include AI accelerators, network cards and sensors.
- Reuse the standard DDR protocols. It minimizes system changes compare to NVDIMM-P.
- Eliminate the need for external I/O buses, reducing system footprint, power consumption, latency, and CPU utilization.

IO-DIMM architecture

- DRAM works a cache for I/O operations
- Flexible I/O queues organization
- Reuse the asynchronous DDR access mechanism employed in the company NVDIMM solution [4]
- If requested data is not available a memory operation is halted.



Asynchronous DDR access



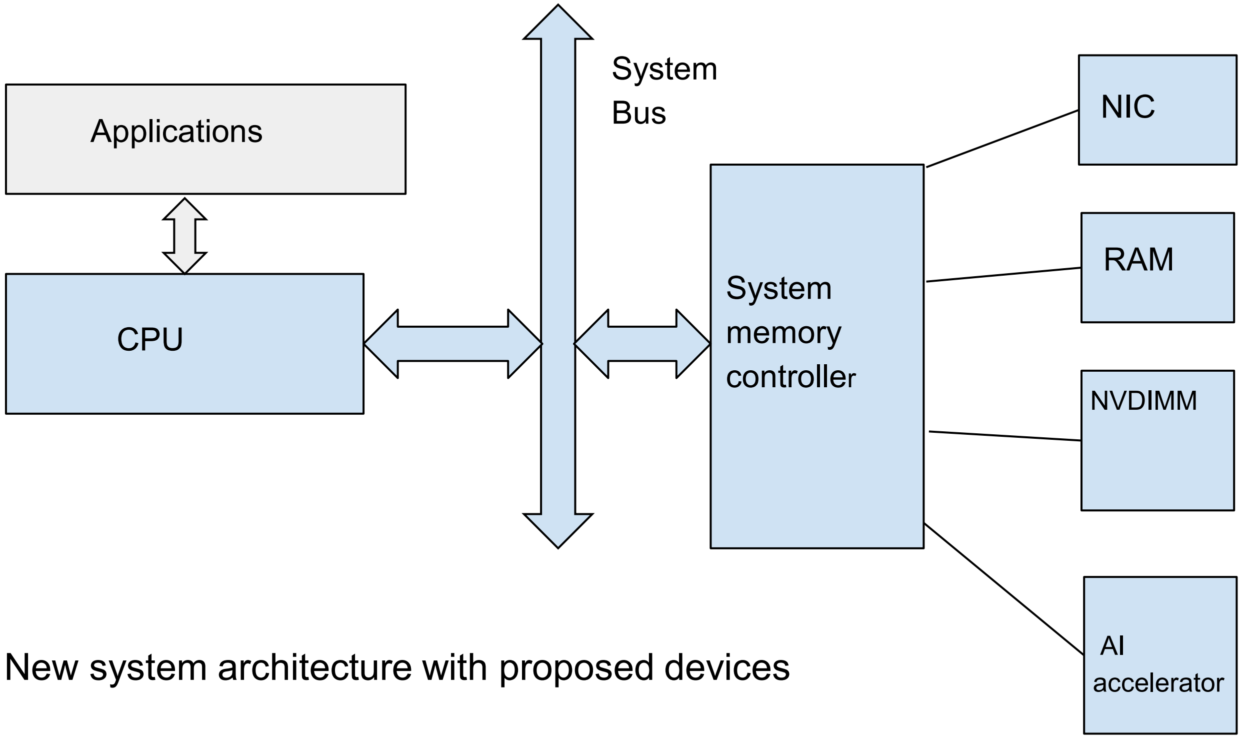
A read command suspended until data in the Read I/O queue is available

There is neither pooling (CPU utilization) nor an interrupt-based method (context switching)

Example of reading data from IO-DIMM

Simplified computer architecture

- **Unmatched Bandwidth:** Achieve up to 20 GB/s with DDR4 or 40 GB/s with DDR5.
- **Optimized CPU Utilization:** Eliminate the need for software management components, allowing applications to access data directly from I/O devices.
- **Compact System Footprint:** Integrate system and non-volatile memory through a unified interface.
- **Reduce power consumption:** Eliminating external I/O buses and reducing memory copy overheads.



New system architecture with proposed devices

References

[1] https://www.jstage.jst.go.jp/article/transinf/E104.D/5/E104.D_2020EDP7092/_pdf/-char/en

[2] NetDIMM: Low-Latency Near-Memory Network Interface Architecture. Authors: Mohammad Alian, Nam Sung Kim, MICRO '52: Proceeding of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture

[3] Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM, Facebook, Washington University in St. Louis, Samsung

[4] www.truememorytechnology.com

Q&A