Distributed Shared Memory Architecture

Di Xu, Jun Song, Fu Wang
Alibaba Group
{di.x, justin.song, yunfan.wf}@alibaba-inc.com

- We propose distributed shared memory, an architecture that provides a shared and tiered memory space using a pool of servers with expansion memory modules attached to the high bandwidth, low latency, cache coherent interface such as Compute Express Link (CXL) [1] on each server.
- The distributed shared memory architecture is motivated by three datacenter trends: data-intensive applications that require data sharing; emerging high bandwidth low latency cache coherent interfaces that enable memory expansion within a compute node or cross nodes; storage class memory technologies [2] that bridges the gap between DDR and flash.
- With CXL, applications can perform native memory load and store instructions to access both local and far memory in this global memory space
- Application data can be persistent if incorporate storage class memory technologies.
- In memory database, PageRank, graph processing etc. make up an important class of data-intensive applications. People usually deploy a cluster of server nodes connected via a high-bandwidth commodity network, which requires careful handling of the data placement in order to reduce the overhead on data movement. It involves replications of data on different server nodes that cost expensive DDR capacity.
- Our goal is to build a shared memory layer that is transparent to the applications and ease programmer's work, lower system memory TCO by reducing data replications and improving memory utilizations.

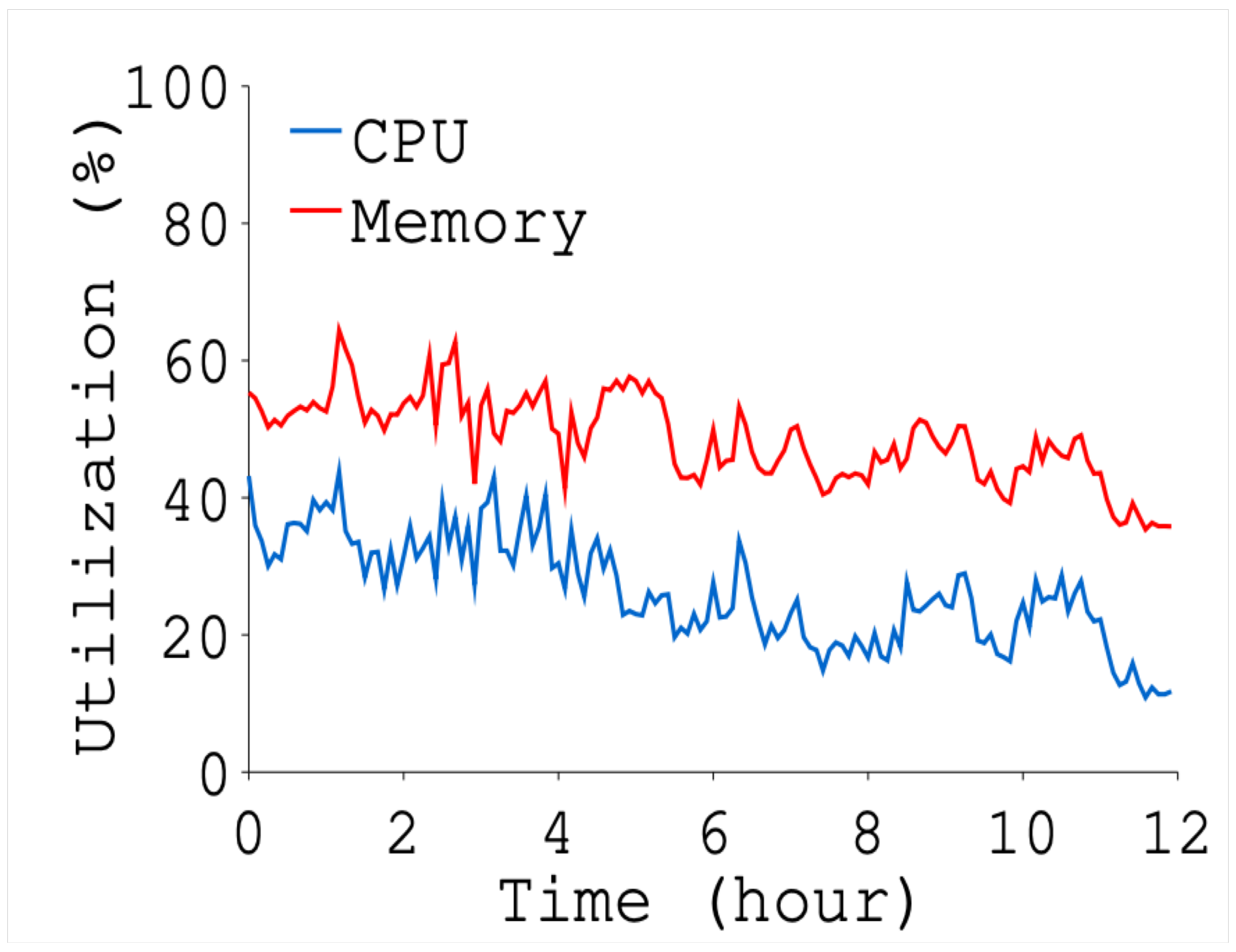

Fig. 1. Alibaba production cluster traces

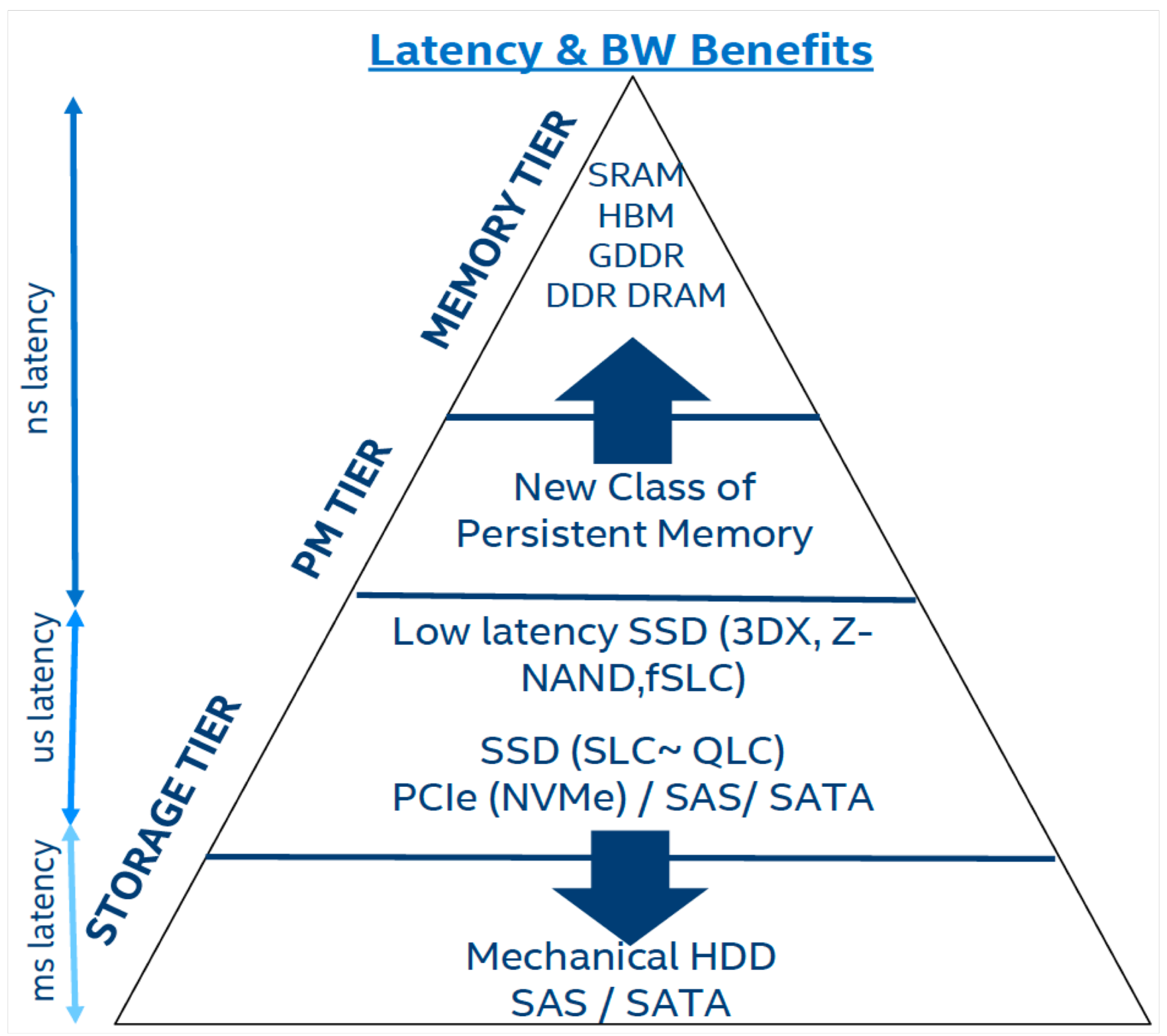

Fig. 2. Memory and Storage Media Hierarchy

Figure 1 shows Alibaba's production cluster traces [3]. Without memory sharing, a physical server becomes the boundary of resource allocation and it is difficult to achieve high memory utilization. Figure 2, Storage Class Memory is an emerging new class of persistent memory. Compared with DDR, SCM offers advantages such as capacity expansion with lower cost and increased data reliability due to its persistency.

- The distributed shared memory architecture that we are proposing is illustrated in Figure 3.
- Tier 1 memory is composed of local system DDR which has the best performance and is intended to serve high SLA workloads.
- Tier 2 memory is a pool of SCM that has large capacity to supplement Tier 1. We plan to use FPGAs with CXL interface to control SCMs. The FPGA will support RDMA or GenZ [4] protocols for interconnecting with each other to form a high-speed data plane.
- The FPGA also serves as a near memory computing engine, that accelerates typical operations such as compression and encryption etc.

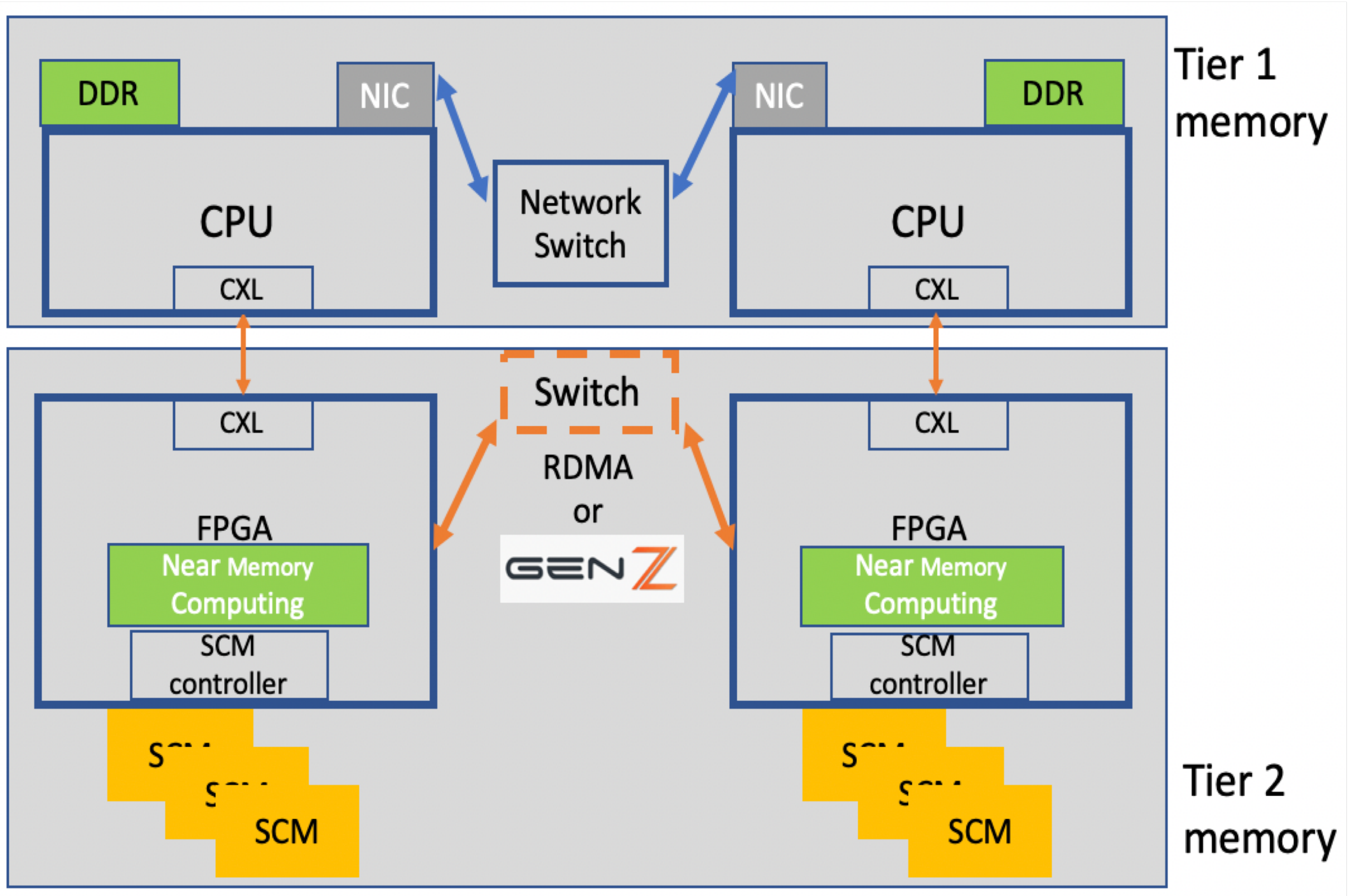

Fig. 3. Distributed Shared and Tiered Memory Architecture with inline acceleration

[1] https://www.computeexpresslink.org/about-cxl
[2] Intel Corporation. Intel Non-Volatile Memory 3D XPoint. http: //www.intel.com/content/www/us/en/architecture-and-technology/non-volatile-memory.html?wapkw=3d+xpoint.
[3] Yizhou Shan, Yutong Huang, Yilun Chen, Yiying Zhang LegoOS: A Disseminated, Distributed OS for Hardware Resource Disaggregation. In Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation
[4] https://genzconsortium.org/white-papers/

OCP FUTURE TECHNOLOGIES SYMPOSIUM

2020 OCP Global Summit