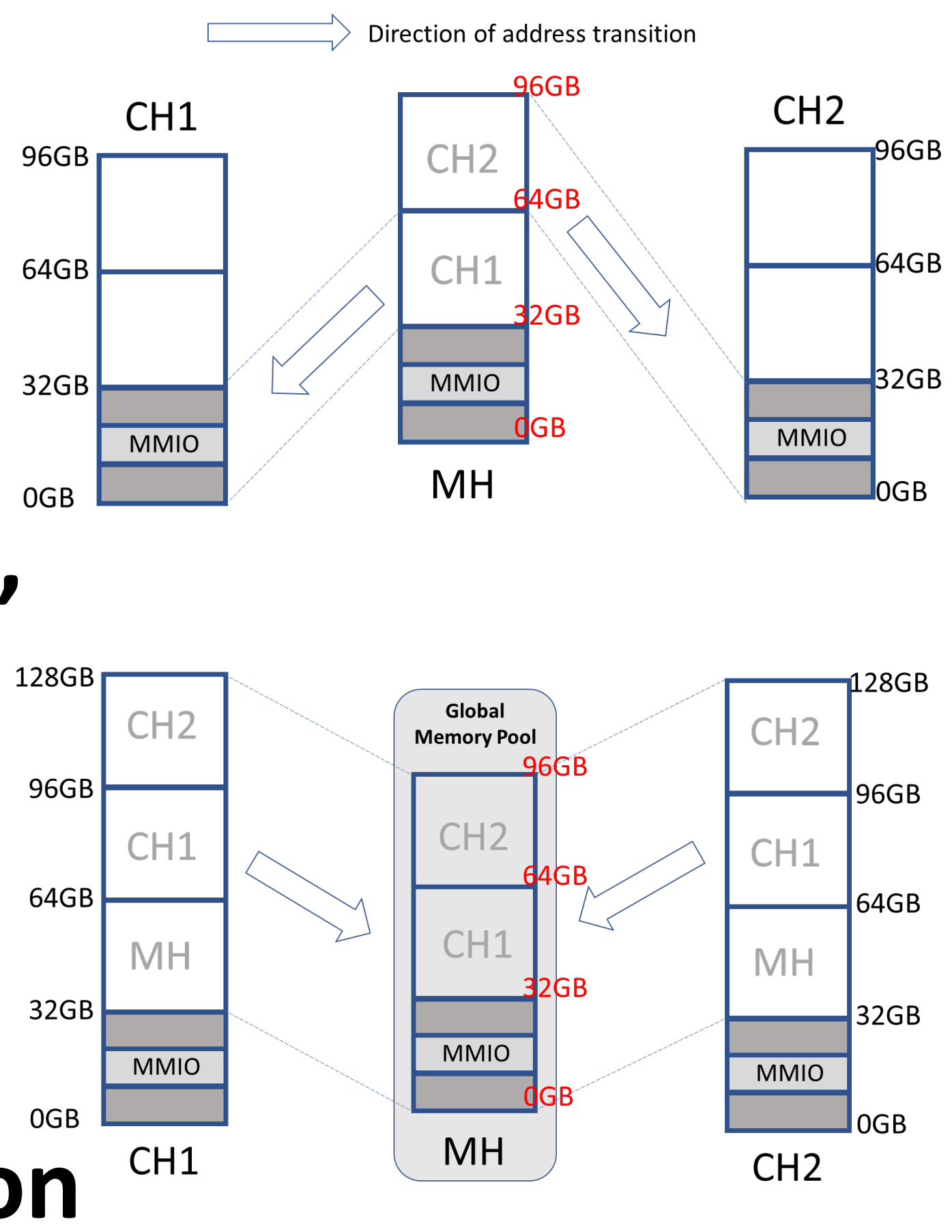
PCIe Networking as Open Accelerator Interconnect

Tzi-cker Chiueh, Chao-Tang Lee
Industrial Technology Research Institute
Hsinchu, Taiwan, R.O.C

Introduction

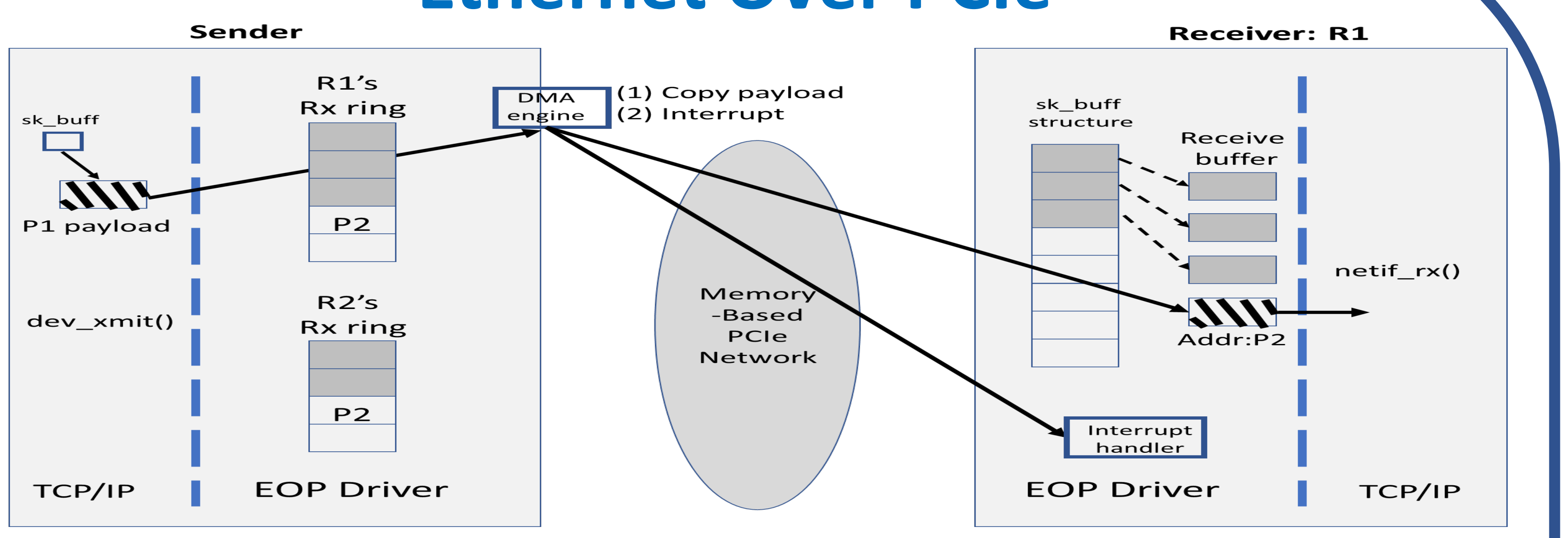
- Open Accelerator Infrastructure (OAI) is a new initiative for scalable DNN training workloads. OAI consists of a standard specification for the OAI accelerator module (OAM) and the interconnect among these modules.
- This poster describes the design and implementation of a potential OAI inter-accelerator interconnect architecture called Ladon, which is built entirely from commercially available PCIe Express (PCIe) components.
- Ladon supports direct access to remote PCIe devices, hardware-based cross-machine DMA (HRDMA) and Ethernet-over-PCIe (EOP) communications

PCIe Connected As A Network



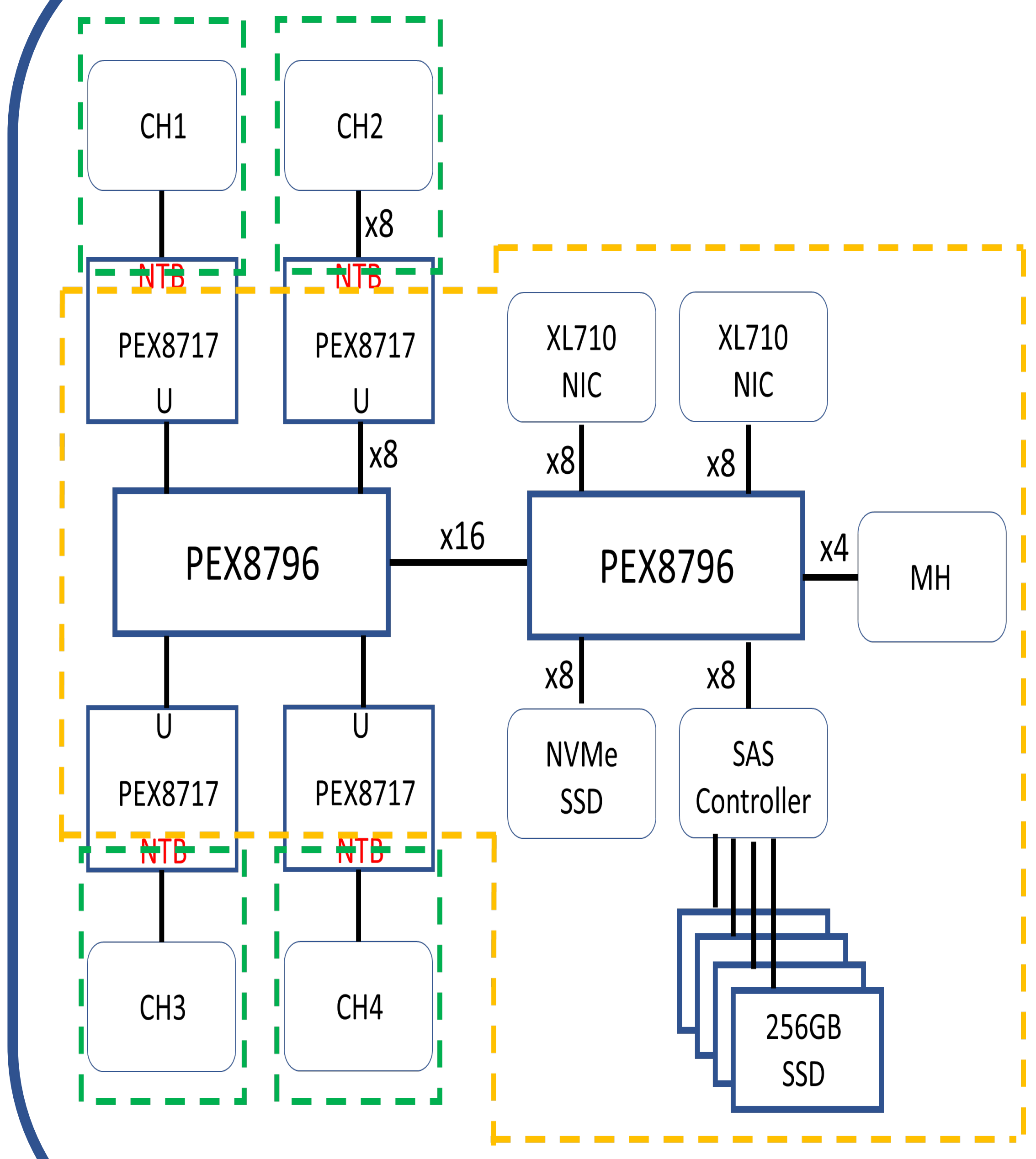
- Physically connect servers (CH1, CH2, MH) by Non-Transparent bridge
- Put the resources of three computers, CH1, CH2 and MH, into a shared address space
  - ✓ Maps CH1's and CH2's resources to MH's physical address space
  - ✓ Then maps the active portion of MH's physical address space back into the physical address space of CH1 and CH2
- CH1, CH2 and MH now could each directly access any resource residing in CH1, CH2 and MH, with proper address translation

Ethernet Over PCIe



- A virtual Ethernet driver that turns Ethernet packets into PCIe frames and enables all existing TCP/IP applications to run on PCIe-based communications networks without modification.
- The EOP driver on a sending server sets up a receive buffer pointer ring for every potential receiver server, and uses HRDMA for packet transfer, and generates MSI-based inter-server interrupts to notify the receiver of packet arrival events.

NexTCA System



- 4 compute hosts (CH) and 2 management hosts (MH) for high availability, with hot plug and play support
- Ethernet over PCIe inter-server communication
- Sharing of all devices mapped on the cluster-wide address space
  - ✓ 40Gbps Ethernet adapter (XL710)
  - ✓ SAS/SATA hard disk and SSD
  - ✓ NVMe SSD
  - ✓ DRAM
- High availability support for
  - ✓ EOP intra-cluster communication
  - ✓ 40G NIC inter-cluster communication
  - ✓ SAS/SATA hard disk and SSD access
  - ✓ NVMe SSD access

2020 OCP Global Summit