Memory Allocation on
[Base / RDMA / GR + NIC / GR + RNIC]

INDEX

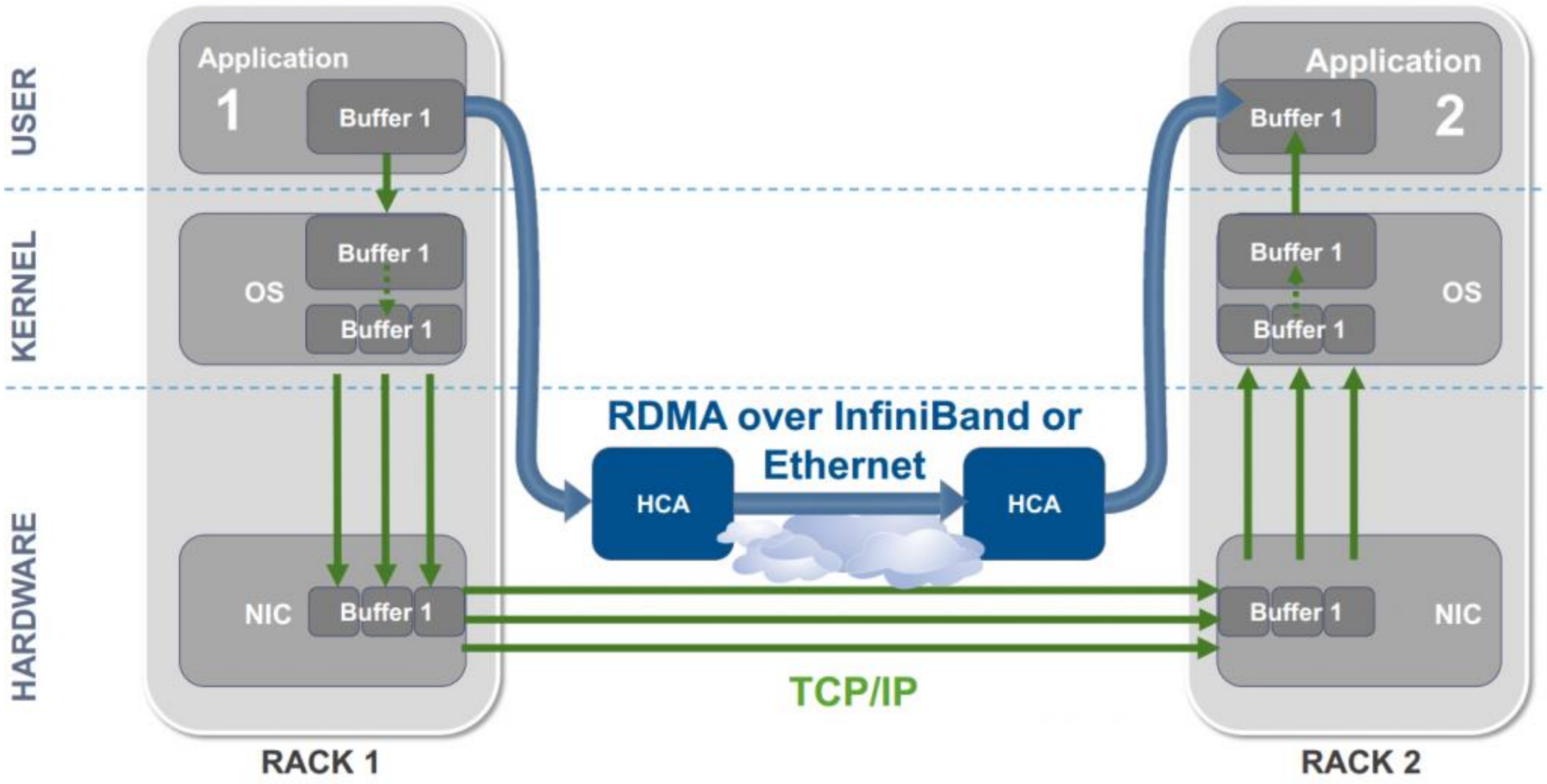
메모리 할당 비교(white paper)

BASE
RDMA
GR + NIC
GR + RNIC

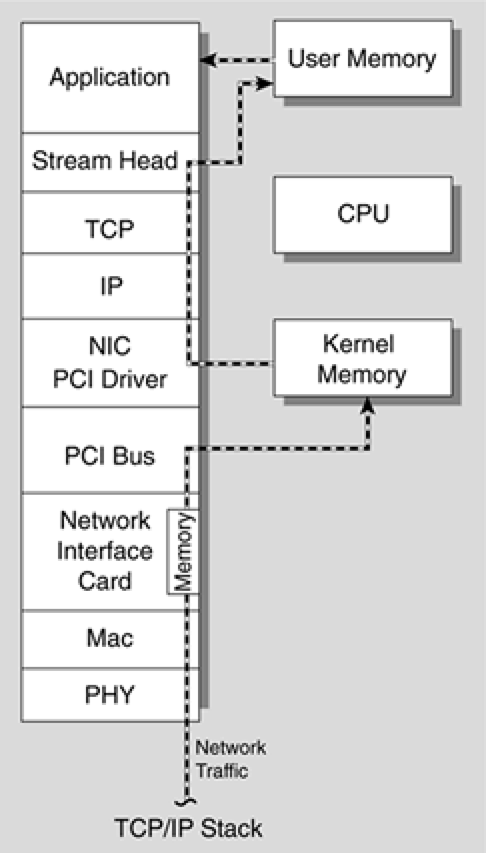
GAM : Efficient Distributed Memory Management with RDMA and Caching

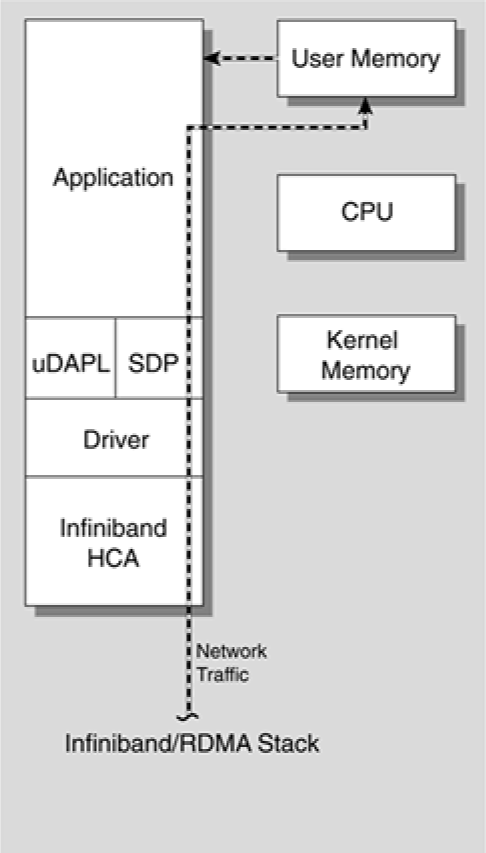
Accelerating TensorFlow with RDMA for high-performance deep learning

BASE vs RDMA

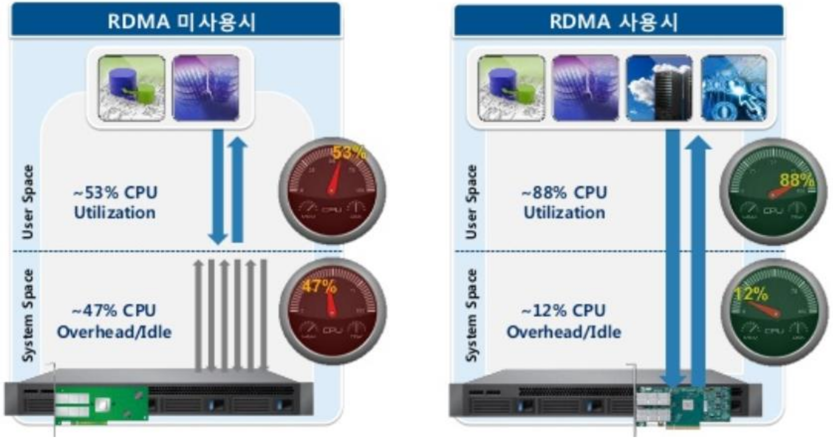

USER
KERNEL
HARDWARE
Application 1
Buffer 1
OS
Buffer 1
Buffer 1
NIC
Buffer 1
RACK 1
RDMA over InfiniBand or Ethernet
HCA
HCA
TCP/IP
Application 2
Buffer 1
Buffer 1
OS
Buffer 1
Buffer 1
NIC
RACK 2

BASE vs RDMA





CPU 효율 차이



GR + NIC

RDMA requires RDMA-enable NIC

RNICs : iWARP, ROCE NICs

Infiniband : already infused in the IB networks

∴ 일반 NIC을 사용한 GPU RDMA 관련된 연구 자료를 찾기 어렵다

Mellanox Requirements

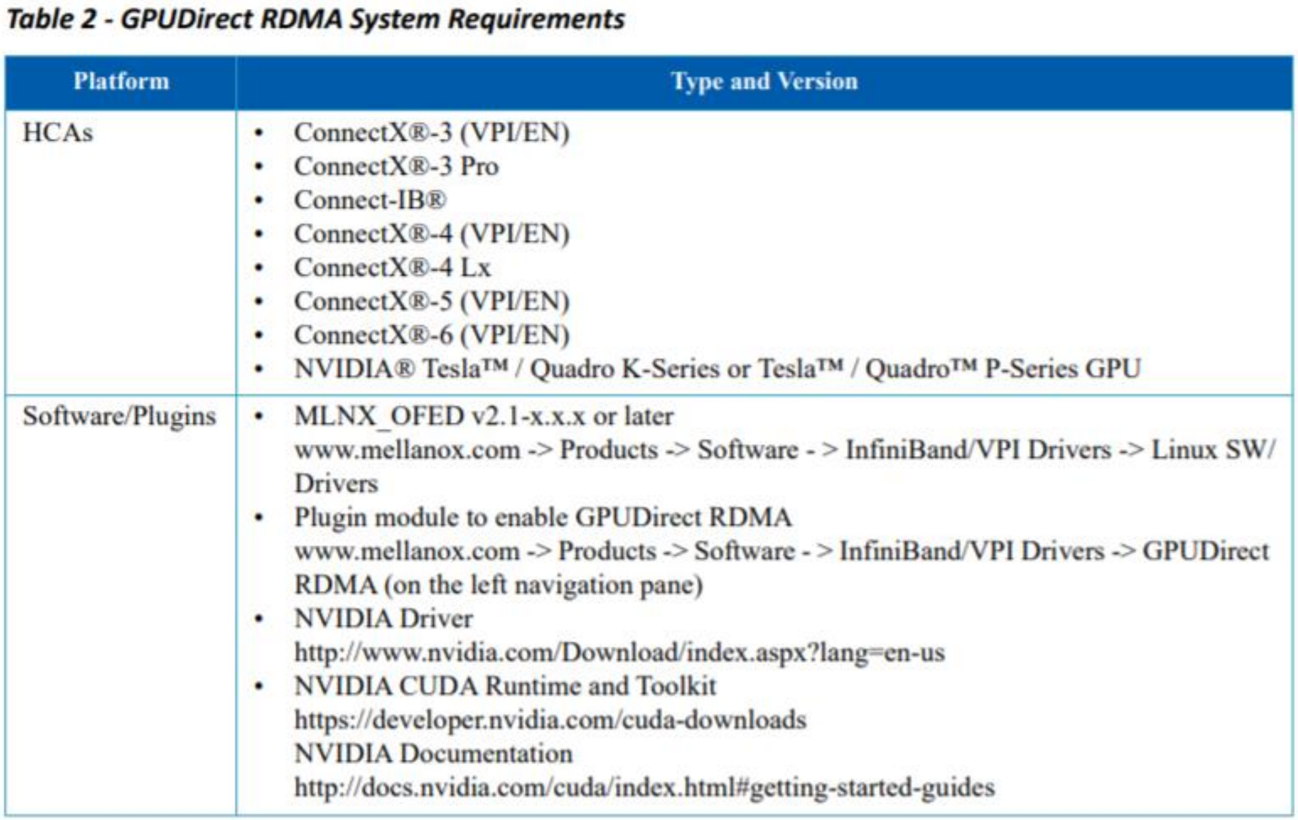
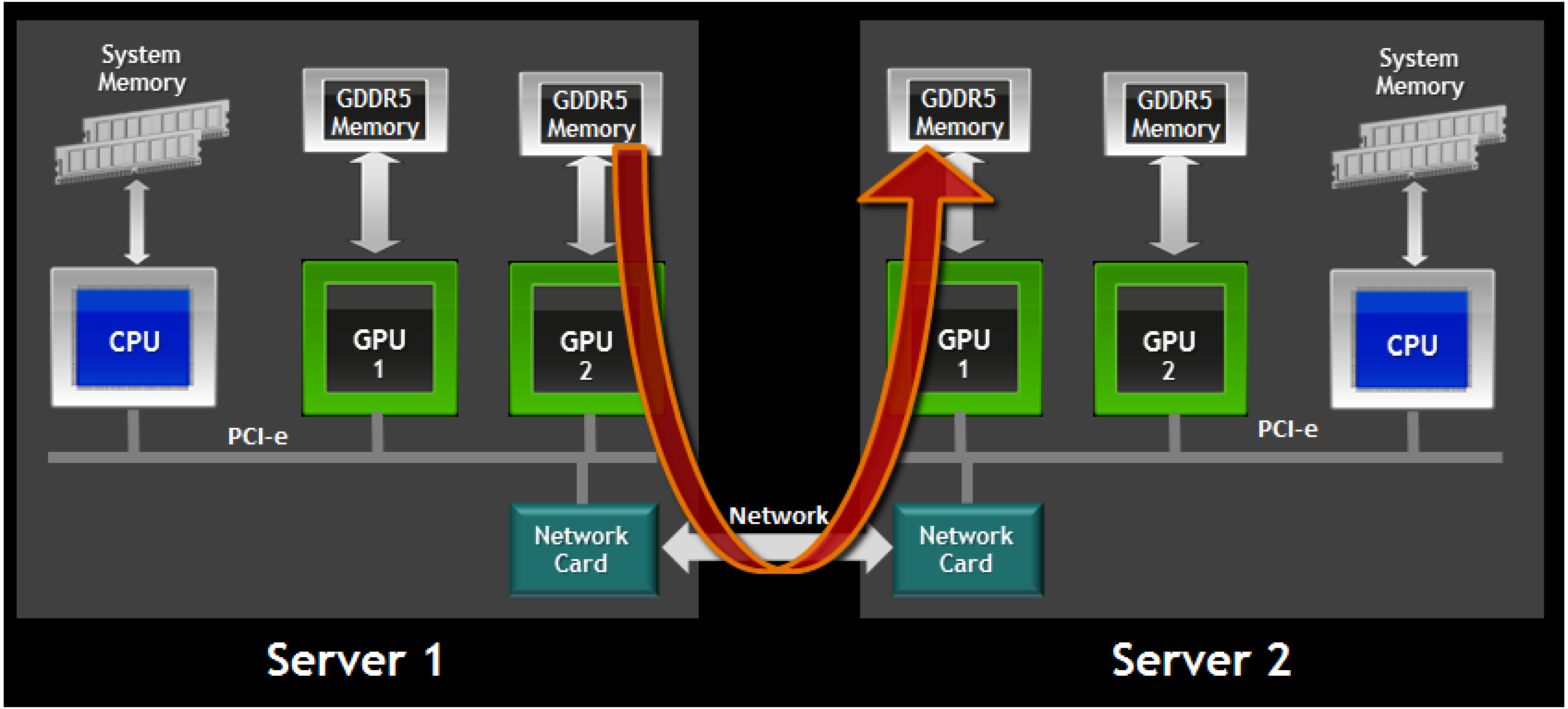


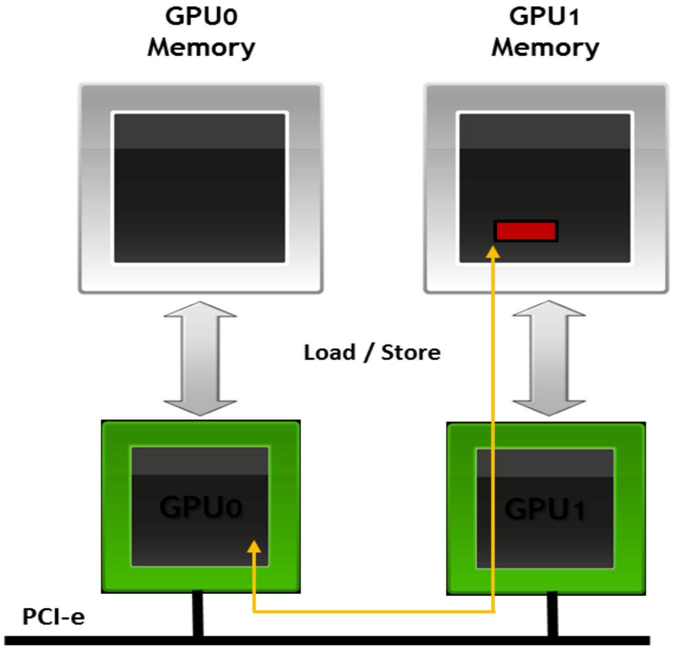
Table 2 - GPUDirect RDMA System Requirements

Platform	Type and Version
HCAs	• ConnectX®-3 (VPI/EN) • ConnectX®-3 Pro • Connect-IB® • ConnectX®-4 (VPI/EN) • ConnectX®-4 Lx • ConnectX®-5 (VPI/EN) • ConnectX®-6 (VPI/EN) • NVIDIA® Tesla™ / Quadro K-Series or Tesla™ / Quadro™ P-Series GPU
Software/Plugins	• MLNX_OFED v2.1-x.x.x or later www.mellanox.com -> Products -> Software - > InfiniBand/VPI Drivers -> Linux SW/ Drivers • Plugin module to enable GPUDirect RDMA www.mellanox.com -> Products -> Software - > InfiniBand/VPI Drivers -> GPUDirect RDMA (on the left navigation pane) • NVIDIA Driver http://www.nvidia.com/Download/index.aspx?lang=en-us • NVIDIA CUDA Runtime and Toolkit https://developer.nvidia.com/cuda-downloads NVIDIA Documentation http://docs.nvidia.com/cuda/index.html#getting-started-guides

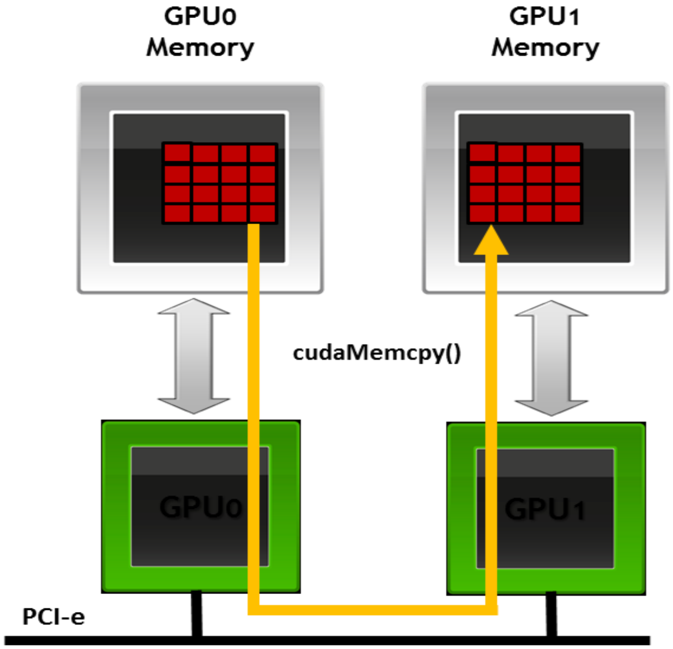
GR + RNIC


System Memory
GDDR5 Memory
GDDR5 Memory
CPU
GPU 1
GPU 2
PCI-e
Network Card
Network
Network Card
GDDR5 Memory
GDDR5 Memory
System Memory
GPU 1
GPU 2
CPU
PCI-e
Server 1
Server 2

GR + RNIC



P2P Direct Access



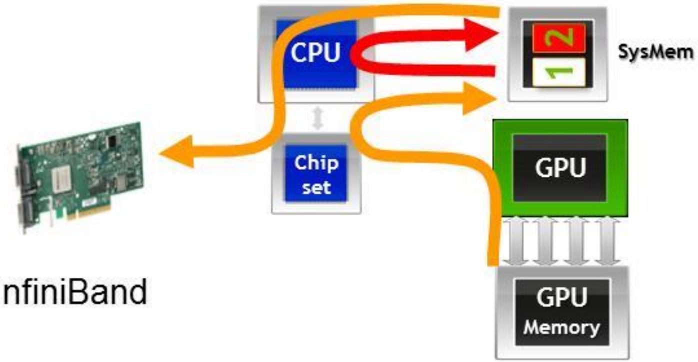
P2P Direct Transfers

GR + RNIC

Without GPUDirect

Same data copied three times:

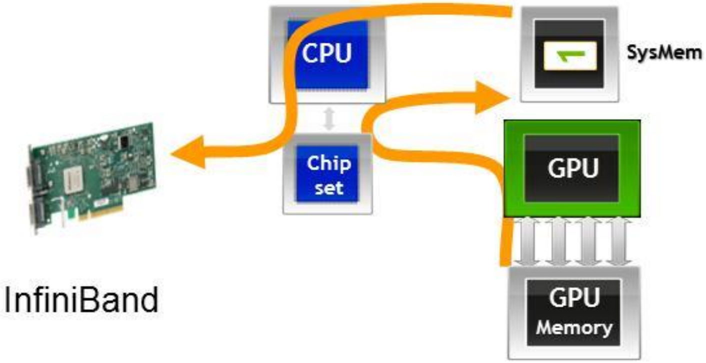
1. GPU writes to pinned sysmem1
2. CPU copies from sysmem1 to sysmem2
3. InfiniBand driver copies from sysmem2



With GPUDirect

Data only copied twice

Sharing pinned system memory makes sysmem-to-sysmem copy unnecessary



GAM : Efficient Distributed Memory Management with RDMA and Caching

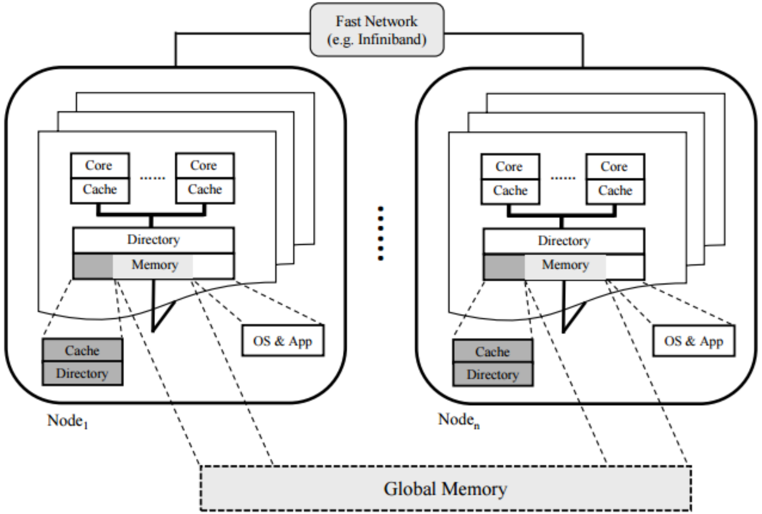


Figure 1: Overview of GAM

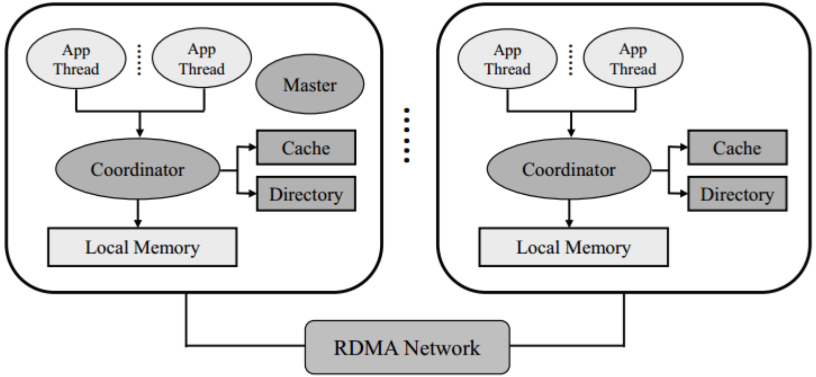
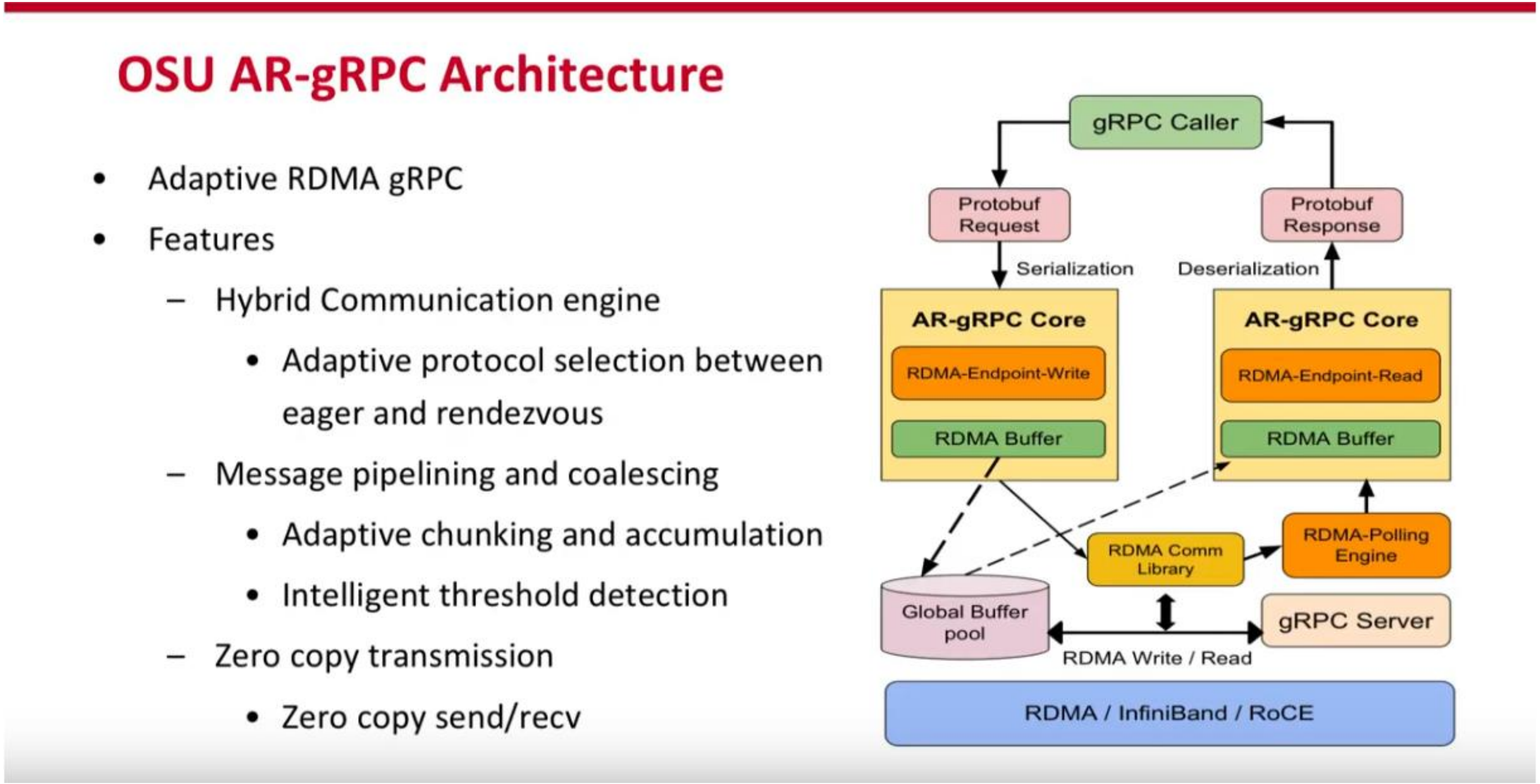


Figure 5: Architecture of GAM

GAM manages the free memory distributed among multiple nodes to provide a unified memory model.

Accelerating TensorFlow with RDMA for high-performance deep learning



OSU AR-gRPC Architecture

- Adaptive RDMA gRPC
- Features
  - Hybrid Communication engine
    - Adaptive protocol selection between eager and rendezvous
  - Message pipelining and coalescing
    - Adaptive chunking and accumulation
    - Intelligent threshold detection
  - Zero copy transmission
    - Zero copy send/recv

그림 출처

Mellanox 블로그랑 white paper

논문들