EE542 - Reading Assignment – 09

RDMA over Converged Ethernet: A Review

Presenter: Boyang Xiao

USC id: 3326-7302-74

Email: boyangxi@usc.edu

Index

- RDMA review
- Converged enhanced ethernet review
- RDMA over converged ethernet
- Q&A

RDMA review

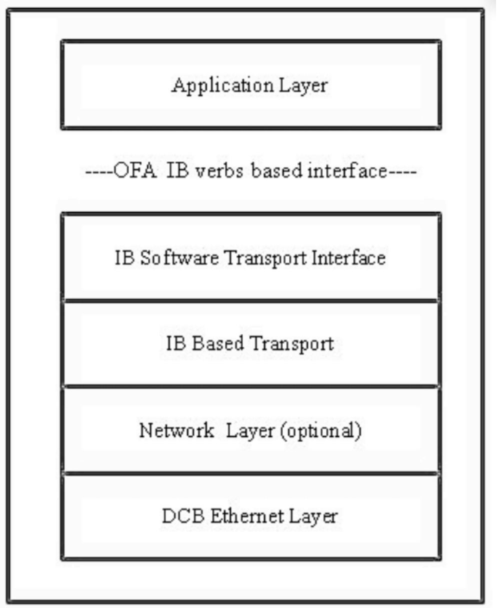
- RDMA is a Direct Memory Access from the memory of one computer into that of another without involving either one's operating system RDMA implements a reliable transport protocol in hardware on the NIC that enables the NIC itself to transfer data directly to or from application memory, without having to execute a kernel call
- RDMA is quickly becoming a necessity in performance-critical networking. It is now finding the increase in applications in modern commercial data centers, especially in performance sensitive environments.
- RDMA supports zero-copy networking where "zero-copy" refers to computer operations with no CPU involvement in copying data from one memory area to another.

Converged enhanced ethernet review

- Converged Enhanced Ethernet is a single interconnect Ethernet technology developed to converge a variety of data centers. Converged Enhanced Ethernet is a term used to refer to the IEEE 802.1 standard version, and is considered to be the next generation Ethernet, providing a standardized packet lossless technology. a.k.a Data Center Bridging (DCB)
- Four specifications from the DCB task force:
    - Priority Based Flow Control
    - Enhanced Transmission Selection (ETS)
    - Congestion Notification
    - Data Center Bridging Exchange (DCBX) Protocol

RDMA over converged ethernet

- RDMA over converged ethernet (RoCE) is network protocol which allows RDMA access over the Ethernet. It is also called link layer protocol which allows the communication between the two hosts on the same Ethernet broadcast domain.
- RoCE architectures:



Application Layer

----OFA IB verbs based interface----

IB Software Transport Interface

IB Based Transport

Network Layer (optional)

DCB Ethernet Layer

RDMA over converged ethernet

- Transport Layer
  - ROCE inherits a rich set of transport services beyond those required to support OFA verbs including connected and unconnected modes and reliable and unreliable services. It also has a full set of verbsdefined operations including kernel bypass, Send/Receive, RDMA Read/Write, and Atomic operations. UDP and multicast operations are also fully supported
- Network Layer
  - ROCE requires InfiniBand Global Routing Header (GRH) based network layer functions. In GRH, routing is based on GID (Global Identifier) which is equivalent to IPv6 addressing and can be adapted to IPv4 addressing.
- Data Link Layer
  - RoCE is implemented & available at the OFED (Open Fabrics Enterprise Distributors) stack. RoCE can be implemented in Hardware as well as software.
  - With Soft RoCE, SFW offers the opportunity for data center technologists to implement RDMA for their business solutions to improve computing efficiency, simplify infrastructure

Q&A

- Question: How does RoCE handle ip addressing in the network layer?

- Answer:
  - ROCE requires InfiniBand Global Routing Header (GRH) for network layer functionalities, where ip addressing and routing is based on a Global Identifier(GID) which is equivalent to IPV6 address and can be adapted to IPV4 addressing. Layered addressing is based on GID. GID resolves to a Queue Pair number plus MAC Address to an End node is referred to by their IP addresses, where the GID is derived from the IP address.

Thanks for watching!

Presenter: Boyang Xiao

USC id: 3326-7302-74

Email: boyangxi@usc.edu