RDMA OVER CONVERGED ETHERNET: A REVIEW

Gurkirat Kaur[1], Manju Bala[2]
[1]M.Tech Student, CSE Department, CTIEMT Jalandhar, Punjab, India
[2]HOD, CSE Department, Jalandhar, Punjab, India

ABSTRACT
Convergence is a buzzword now a day in the network industry. RDMA has enjoyable consideration in the late 1990 when the Virtual Interface Architecture was introduced. This growth has accelerated with the introduction of Open Fabrics Alliance's (OFA's) Verb Interface. The stability & independence of OFA verb interface facilitated significant growth of software applications that exploit the benefits of RDMA. Until recently, a user wishes to utilize the benefits of RDMA could either use the InfiniBand Architecture or RDDP protocols which are designed to run over an IP network. The recent developments in the Ethernet have expanded the available options for providing RDMA benefits to the applications while using the Ethernet link & physical layers. RoCE is an emerging trend that can be made to work on the Ethernet infrastructures. The goal of this paper is to describe the RDMA over Converged Ethernet (RoCE). The main benefit of RoCE is that it can be implemented in hardware and software also named as Soft RoCE.

KEYWORDS: *RDMA, InfiniBand, Ethernet, Converged Enhanced Ethernet, RoCE, Soft RoCE*

I. INTRODUCTION

Ethernet is most widely used technology in the world. These days Ethernet networks can be found in data-centres, offices, schools and cluster-computing. However, with the growing demand of low latency and high throughput the technologies like InfiniBand and RoCE have evolved with unique features viz. RDMA. InfiniBand is a well-known technology that provides high-bandwidth and low-latency and makes optimal use of in-built features like RDMA (Remote Memory Direct Access). With the rapid evolution of InfiniBand technology and Ethernet lacking the RDMA and zero copy protocol, the Ethernet community has come out with a new enhancements that bridges the gap between InfiniBand and Ethernet. By adding the RDMA and zero copy protocol to the Ethernet a new networking technology is evolved called RDMA over Converged Ethernet (RoCE).

II. REMOTE DIRECT MEMORY ACCESS

RDMA is the interesting network technology that has been dominant in the HPC marketplace [8]. Modern HPC interconnects like InfiniBand, Myrinet, IBM's Federation Technology makes use of RDMA to improve the performance & achieve high throughput considering the hardware bandwidth [4]. 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, e.g. almost any form of cloud computing [8]. Since the mid-1990's, Ethernet has been the dominant LAN technology. Ethernet has replaced many other LAN Technologies like FDDI, token ring etc. 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 supports **zero-copy networking** where "zero-copy" refers to computer operations with no CPU involvement in copying data from one

memory area to another. It also removes CPU from being bottleneck the use of RDMA reduces the cost of data movement by eliminating redundant copies throughout the network path, and reduces overall resource utilization (see Fig.1)

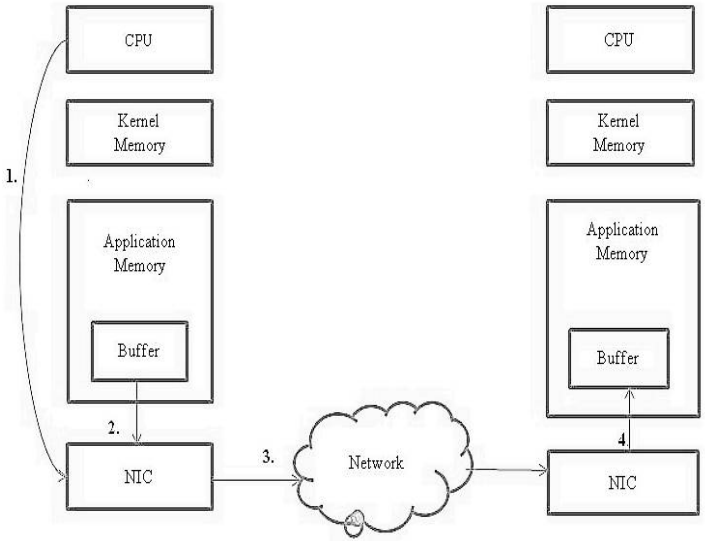

Figure 1: Working of RDMA

III. CONVERGED ENHANCED ETHERNET

Converged Enhanced Ethernet is a single interconnect Ethernet technology developed to converge a variety of data centres. CEE's primary focus is to consolidate the number of cables and adapters connected to servers. 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. [5] Converged Enhanced Ethernet (CEE) is also called Data Center Bridging (DCB). Four specifications from the DCB task force:

3.1 Priority Based Flow Control

It is defined as the *IEEE 802.1Qbb* standard; it focuses on developing a standard mechanism that can control the flow for each traffic class of service independently. The goal of this mechanism is to ensure zero loss under congestion in Data Center Bridging networks.

3.2 Enhanced Transmission Selection (ETS)

It is defined as the *IEEE 802.1Qaz* standard; it provides a common management framework for assignment of bandwidth to frame priorities.

3.3 Congestion Notification

It is defined as the IEEE 802.1Qau standard; it provides end to end congestion management for protocols that are capable of transmission rate limiting to avoid frame loss. It is expected to benefit protocols such as TCP that do have native congestion management as it reacts to congestion in a timelier manner.

3.4 Data Center Bridging Exchange (DCBX) Protocol

It is a discovery & capability protocol and allows the automatic exchange of Ethernet parameters and discovery functions between switches & endpoints to ensure consistent configuration across the network.

IV. RDMA OVER CONVERGED ETHERNET

Remote Direct Memory Access (RDMA) is an effective technology to reduce the system load & to improve the throughput. Recently, Ethernet has exploited the RDMA technology that can provide a high performance fabric for MPI communications at lower cost than other competing technologies.

The emerging RDMA over Converged Ethernet (RoCE) standards enables the InfiniBand transport for use over the existing and widely deployed network infrastructure [2]. The RoCE standards allow the users to take the advantage of low latency, high efficiency; high performance. RoCE is basically an InfiniBand (IB) protocol that can be used over the Ethernet infrastructures. RoCE provide all of the InfiniBand transport benefits and well established RDMA ecosystem combined with 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. [6]

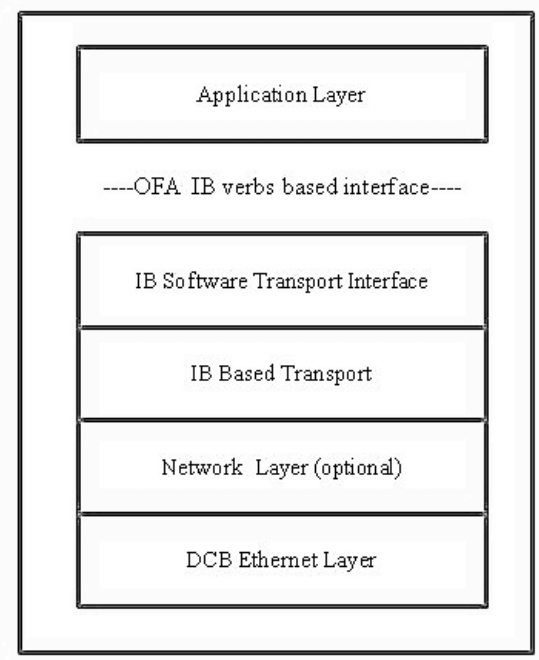

Figure 2: RDMA over Converged Ethernet

With the initiation of CEE, a new option to use a non-IP-based transport option is available, which we shall call RDMA over Converged Ethernet (RoCE, pronounced as "rock-ie"). RoCE uses the upper layers of the InfiniBand Architecture, including the transport layer and above, directly on top of CEE or optionally IP over Converged Enhanced Ethernet. RoCE provides all of the InfiniBand transport benefits and well-established RDMA ecosystem combined with all the
CEE advantages [3]. InfiniBand was designed with these architectural goals:

- Communication architecture optimized for Message passing.
- Ease of implementation on either hardware or software.

The realization of these goals produced an architecture which maximizes the delivered performance like bandwidth, latency and latency variation (jitter), and minimizes resource utilization (CPU, memory bandwidth and wire bandwidth). The goal was to deliver performance capacity as much possible to application processing by reducing the resources required to support communications [3].

Software Interface & Transport Layer: RoCE is complaint with OFA verbs definition & is interoperable with OFA software stack which is similar to InfiniBand & iWARP (Internet Wide Area Network Protocol). InfiniBand transport layer check Ethernet layer 2 addresses instead of the InfiniBand layer 2 addresses. The IB transport layer provide services like data link layer, especially related to lossless delivery of packets, and these services are delivered by a CEE based data link 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 verbs-defined operations including kernel bypass, Send/Receive, RDMA Read/Write, and Atomic operations. UDP and multicast operations are also fully supported [6].

Network Layer: The network layer can be used for routing even though, routing for RoCE packets are undesirable, where latency, jitter & throughput are the main considerations. When necessary, 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. 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.

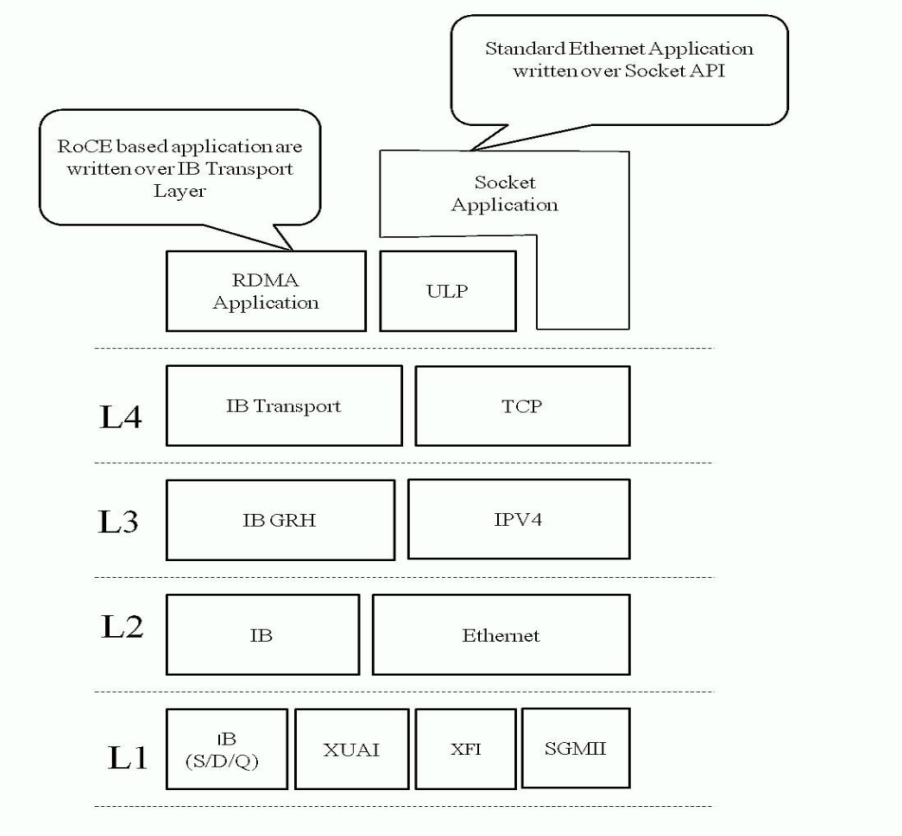

Figure 3: Protocol Stack of RoCE

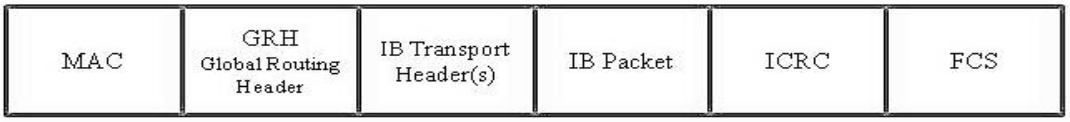

Figure 4: RoCE Packet Format

Data Link Layer: At Data link Layer, standard link layer services are needed, as well as IEEE802.1Qbb is required at minimum. IEEE802.1Qau is desirable but not mandatory unless there is congestion while transferring the data between server to server or server to storage. Addressing of nodes is based on MAC address.MAC address is an Ethertype field which uniquely identifies an RDMA network. Many Linux Distributors included OFED (Open Fabrics Enterprise Distributors), support a wide and rich range of middle wares and application solutions like IPC, sockets, messaging, virtualization etc. RoCE is implemented & available at the OFED stack. RoCE can be implemented in Hardware as well as software. April 22, 2010 – System Fabric Works (SFW) is a systems integration company delivering a high quality integration, development & deployment of high performance software solutions to the global clients. SFW is delivering powerful, open-source fabric and I/O solutions in high performance software engineering, announced support for RDMA over Converged Ethernet (RoCE) implemented in software as an addition to the OpenFabrics Enterprise Distribution (OFED) release 1.5.1 for Linux. RoCE is a new standard announced earlier by the InfiniBand Trade Association (IBTA) and supported by the OpenFabrics Alliance. SFW is announcing the availability of a software implementation of the RoCE standard – compatible with standard Ethernet networks – called "**Soft RoCE**." With **Soft RoCE**, SFW offers the opportunity for data center technologists to implement RDMA for their business solutions to improve computing efficiency, simplify infrastructure, and future proof their networks for scaling from 1 to 10 gigabits per second. [7]

V. CONCLUSION & FUTURE SCOPE

The importance of RDMA is growing in the industry, driven by increased use of clustered computing. RDMA enables low latency, which is a foundation stone for delivering efficient computing and linear scaling of clusters, resulting in higher ROI [1]. The use of RDMA reduces the cost of data movement by eliminating redundant copies throughout the network path, and reduces overall resource utilization. InfiniBand is an emerging Technology & RoCE is the InfiniBand protocol which can be used over the Ethernet infrastructure. Because of the RDMA RoCE provides the efficient data transfer with very low latencies at lossless Ethernet. RoCE can be implemented in software too. In future, the comparison of Soft RoCE with other existing technologies like Ethernet can be evaluated using different benchmarks by choosing different parameters. By doing this type of comparison, it is possible that RoCE will become a low cost solution for that who want to stick to the Ethernet always & also avail the benefits of InfiniBand on the Ethernet infrastructure.

REFERENCES

[1]. Motti Beck& Michael Kagan, "Performance Evaluation of the RDMA over Ethernet (RoCE) Standard in Enterprise Data Centers Infrastructure", in Proceedings of the 2011 3rd Workshop on Data Center Converged and Virtual Ethernet Switching, 978-0-9836283-2-3 © 2011 ITC

[2]. Ezra Kissel &Martin Swany, "Evaluating High Performance Data Transfer with RDMA-based Protocols in Wide-Area Networks", in 2012 IEEE 14th International Conference on High Performance Computing and Communications, 978-0-7695-4749© -7/12 2012 IEEE

[3]. David Cohen, Goldman Sachs; Thomas Talpey, Consultant; Arkady Kanevsky, Consultant; Uri Cummings, Fulcrum Microsystems, Michael Krause, HP; Renato Recio, IBM; Diego Crupnicoff, QLogic, Mellanox Technologies; Lloyd Dickman,; Paul Grun,& System Fabric Works, "Remote Direct Memory Access over the Converged Enhanced Ethernet Fabric: Evaluating the Options", 1550-4794/09 © 2009 IEEE

[4]. Michael Oberg, Henry M. Tufo, Theron Voran, and Matthew Woitaszek, "Evaluation of RDMA over Ethernet Technology for Building Cost Effective Linux Clusters", University of Colorado, Boulder , National Center for Atmospheric Research.

[5]. (2013), The techopedia website.[Online].Available at: http://www.techopedia.com/definition/1295/converged-enhanced-ethernet-cee

[6]. (2013), The Wikipedia Website.[Online] available at: http://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

[7]. (2013), The System Fabrics Works Website, [Online] Available at http://www.systemfabricworks.com/news/system-fabric-works-ofed-distribution-supports-rdma-over-converged-ethernet-roce

[8]. (2013), The IBTA Blog Website,[Online] Available at http://blog.infinibandta.org/?s=roce+vs+infiniband

AUTHORS

Gurkirat Kaur received her B.Tech Degree from Uttar Pradesh Technical University Lucknow and is pursuing her M.Tech Degree from the Department of Computer Science & Engineering in CT Institute of Engg. & Technology, Jalandhar, Punjab. Her interest area is Parallel Computing, Computer Networks.

Manju Bala received her B.Tech. from U P Technical University Lucknow, India and Master of Technology in Computer Science and Engineering from Punjab Technical University Jalandhar in the year of 2007 and completed her PhD from NIT, Hamirpur (HP) in 2013 . She has worked eight years as a Lecturer in Information Technology at DAV Institute of Engineering and Technology, Jalandhar (Punjab) .Currently she is working as Associate Professor and head of the Department in computer science and Engineering Department at CT Institute of Engineering Management and Technology, Jalandhar (Punjab). Currently she is working in the area of data communication, computer network and wireless sensor networks. She has published 40 research papers in the International/National/Conferences. She is member of Punjab Science Congress, Patiala, India.