iWARP: From Clusters to Cloud RDMA

RDMA over Ethernet Goes Mainstream

This paper presents an overview of the evolution of iWARP, the standard for RDMA over Ethernet. By going over the history of iWARP, the paper traces the developments that made iWARP a potent high performance replacement for InfiniBand, a true plug-and-play solution that maintains the same APIs while enabling Internet wide scalability. Looking forward, the elimination of the performance gap with IB, coupled with the familiarity and benefits of Ethernet and the scalability and resiliency of TCP/IP, destine Ethernet with iWARP to overtake InfiniBand, as it consistently did with competing technologies.

Introduction

Remote DMA (**RDMA**) is a technology that achieves unprecedented levels of communication efficiency, thanks to direct system or application memory-to-memory transfer, without CPU involvement. With RDMA enabled adapters, all packet and protocol processing required for communication is handled by the network adapter itself, typically in hardware for high performance. In return for the performance and efficiency benefits, RDMA does require application changes from the popular socket paradigm to an asynchronous communication model based on a send and receive "queue pair" concept, using a set of communication "Verbs" or operations. The OpenFabrics Enterprise Distribution (OFED, [1]) is the main open-source RDMA middleware library, with support for the major RDMA providers.

In an era of **Big Data**, massive datacenters, pervasive virtualization and focus on "Green" operation and efficiency, RDMA use is steadily gaining ground. Moreover, RDMA support is integrated into the very core of today's server operating systems (BSD, Linux, and Windows). By providing high level, simplified communication **abstractions**, such integration lowers the barrier to realizing the benefits of RDMA, and is further contributing to the acceleration in RDMA adoption. A leading example of this movement is seen in two key applications that have been identified and targeted in Windows Server 2012, namely high performance file storage (**SMB**) and **Virtual Machine migration** in virtualized systems. In fact, the latter builds upon the native RDMA support introduced into SMB to seamlessly achieve unprecedented levels of performance in Virtual Machine migration.

This paper discusses the benefits of RDMA, and the use cases driving its adoption in today's data centers. It then expands on the evolution of Ethernet and iWARP that allow them to match or outperform esoteric RDMA fabrics, such as InfiniBand, in both micro-benchmark and application level metrics.

An Overview of iWARP

The Internet Wide Area RDMA Protocol (**iWARP**) is the **IETF standard** for RDMA over Ethernet. It was developed by the iWARP Consortium, and standardized by IETF in 2004. The iWARP protocol layers RDMA on top of **TCP/IP**, which is the main transport protocol used in the **Internet**, in data centers and **cloud** installations, and in **Ethernet** networks in general. Network statistics consistently show that TCP carries over 90% of Internet traffic [2]. While IP is clearly needed for routing and packet delivery, TCP was chosen as transport because it provides the following key functionalities:

1. **Reliability** – most packet networks (e.g. Ethernet and IP) are "best-effort" where packets can get dropped or re-ordered. TCP handles re-ordering and data retransmission, providing reliable data transfer in all environments, including **long distance** and next generation high speed **wireless** links, expected to reach 5 to 10Gb speeds in the near-term future [3].
2. **Flow control** – TCP allows the receiver to flow control the sender to avoid over-subscribing its resources. This end-to-end control allows TCP to operate between vastly different endpoints, e.g. servers with 10x or 100x the network connectivity speed of clients.
3. **Congestion control** – TCP implements algorithms to automatically adapt its transmission rate to the network capacity in order to avoid and react to congestion. This prevents collapse when load increases beyond trivial levels, and allows TCP to work at high performance in **large scale or heterogeneous networks** and across **network boundaries**.

Thus, iWARP's native support for reliability and congestion control mechanisms ensures maximum **scalability, routability, reliability and robustness** without requiring a lossless fabric or Ethernet PAUSE to be enabled. It also guarantees ease of deployment and use, and allows leveraging all **existing infrastructure**, including networking, monitoring, security and management with no change required. iWARP is supported in the same OFED distribution as InfiniBand, the incumbent RDMA provider, and requires **no changes to RDMA applications** to run over Ethernet.

A Game of Speeds

When the first iWARP adapters were introduced about a decade ago, they provided 10Gbps speed, while InfiniBand was at DDR speed (called 16Gbps, but lower effective rate). IB distanced itself again with QDR (again called 32Gbps, with lower effective rate). However, both iWARP and IB adapters were limited in practice by the PCI Express attachment speed, at 22Gbps per adapter. This development established the pattern for the following generations, where the effective capacities of iWARP and IB adapters were identical, both being **limited by the PCI bus**, despite a difference in individual port speed. In 2013, iWARP became available at 40Gbps, virtually **eliminating the single port speed gap** with IB.

The following diagram shows the evolution of multi-ported adapter capacity over the past few generations, projected into 2015. The diagram shows how the gap between the two technologies has been closing as physical layer technologies (SERDES) converge, resulting in full **parity** at 100Gbps[1].

[1] In fact, most high-speed interface technologies are converging in SERDES design (e.g. FC, SAS, IB, PCI, Ethernet), eliminating what used to be a differentiating aspect of esoteric fabrics. This convergence leverages the increasingly expensive R&D across multiple applications, and will keep Ethernet continually abreast of the fastest link speeds.

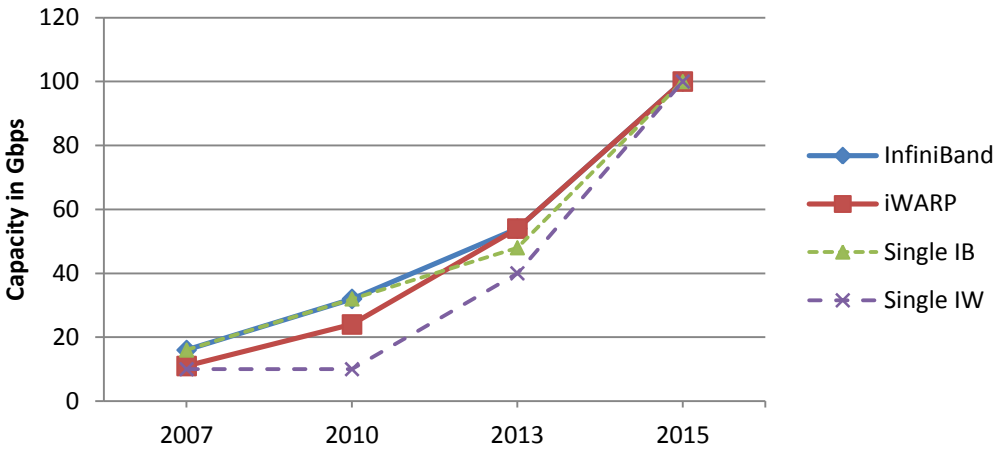

Figure 1 – iWARP vs. IB Single Port and Aggregate Adapter Bandwidth

The following diagram contrasts iWARP and IB **latency**, again showing the two in **near parity** today, with identical, sub-microsecond performance expected in the next iteration. This diagram also dispels the myth that TCP prevents iWARP from achieving the same levels of latency as IB. Much of this myth is based on poor implementations using firmware running on underpowered general-purpose processors. The same approach to implementing IB would yield similarly poor performance. In contrast, **cut-through, specialized-processor** based implementations exist that only need 10nsec per packet to fully process the iWARP protocol stack.

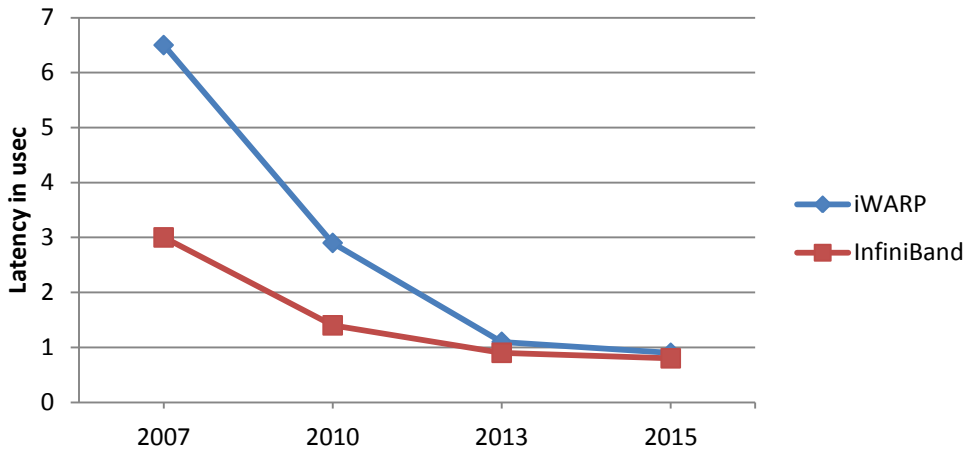

Figure 2 – iWARP vs. IB Latency

Notwithstanding the micro-benchmark latency differences that had existed in the past, iWARP has been shown to match or exceed competing IB gear in **real-life application**-level benchmarks. This is clearly extends to today's more even playing field. The following diagram from an IBM study [4] compares Weather Research and Forecasting (WRF) execution time over 40Gb iWARP and 56G FDR IB. The results show **parity with a slight edge for iWARP**, and are representative of application-level performance across a range of useful cluster applications, which can run **seamlessly** over the two fabrics using identical APIs.

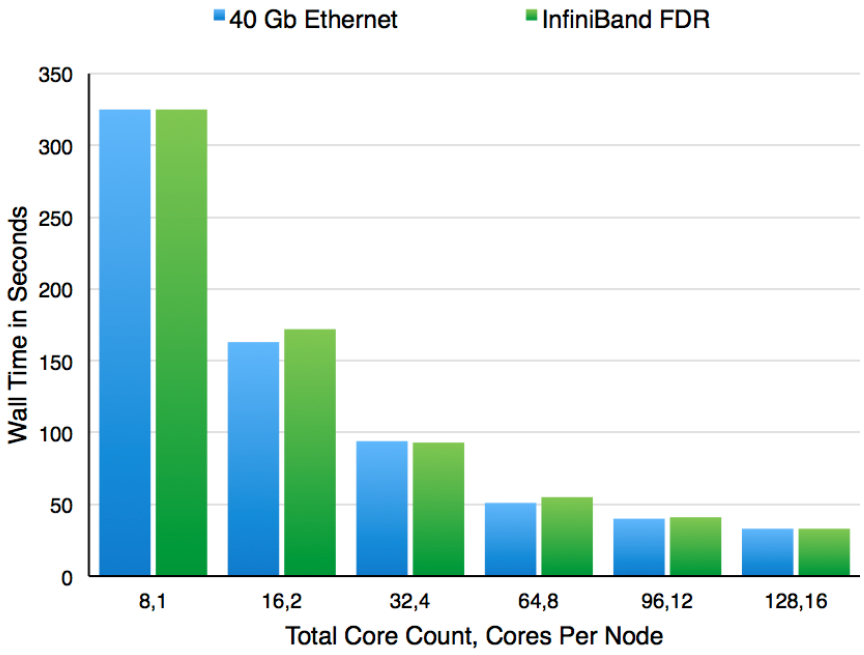

Figure 3 – WRF Execution Time for 40G iWARP vs. 56G IB (Lower is Better)

iWARP Today

With a **performance** profile that matches or exceeds that of the fastest InfiniBand speed, multi-vendor support, **price parity both at the adapter and switch level**, and plug-in ready software **in-boxed** in all major operating systems, iWARP is in a unique position to replace IB in its traditional markets – HPC and specialized clustered applications, such as databases.

However, it is in the **exciting new scale-out applications** of RDMA, that are expected to see the most massive use yet of RDMA – i.e. **Big Data, datacenters and clouds** – where iWARP is particularly well suited for adoption. With its Ethernet and TCP/IP DNA, **iWARP is a native** of these environments, unlike IB, an alien technology, requiring a separate fabric and unjustifiable capital and operating costs.

In this context, Windows Server 2012 embraced RDMA by offering a simplified interface (NDKPI) to kernel applications built on top of RDMA adapters. NDKPI launched with two key use cases: the first is **SMBDirect**, a highly efficient SMB protocol implementation written to take advantage of RDMA, and **Live Migration (VM motion)** in virtualized environments, a killer application that is remarkable in its combination of huge transfer sizes and strict delay requirements. The figure below from [5] shows SMBDirect performance with 40Gbps iWARP, compared to FDR InfiniBand, showing the two again at performance parity.

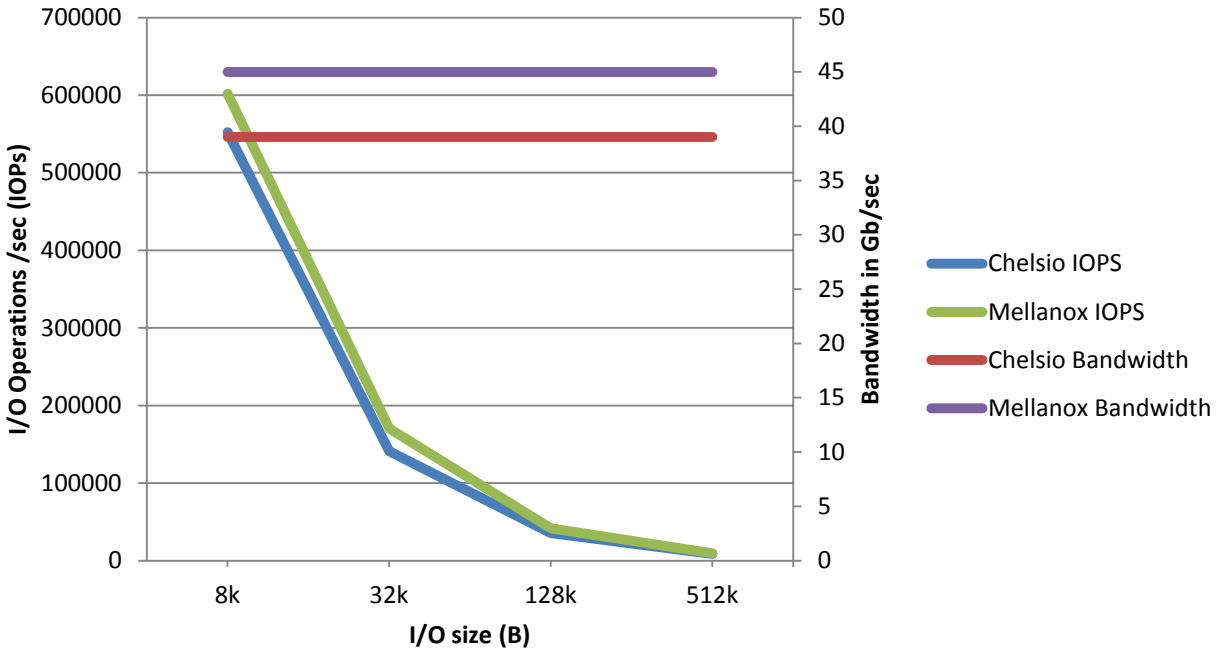

Figure 4 – iWARP vs. IB SMBDirect Throughput and IOPS performance

iWARP support is in-boxed in Windows Server 2012 R2, and is automatically enabled when an iWARP adapter is detected. The **transparent operation, ease of deployment and use** and notable performance benefits are expected to fuel iWARP RDMA's adoption in large Windows installs. The NDKPI is further expected to be leveraged by other applications, which can readily be deployed thanks to its simplification of the Verbs interface. These advantages are likely to trigger the adoption of RDMA and iWARP in other hypervisors, **virtualized environments and cloud stacks**.

Another driving force behind RDMA's mainstream adoption is the rise of **non-volatile RAM (flash)**-based storage, which breaks performance barriers that have held back storage latency and I/O capacity to the slow mechanical spinning disk spindles. RDMA allows the network fabric to match the performance of **flash-based arrays**, thus fully realizing their advantages. iWARP is again exceptionally well suited for this application, as a high-performance transport that can **natively share the same Ethernet** infrastructure with other storage protocols.

The Future of iWARP

iWARP is available today at 40Gb speeds, with 1.5usec end-to-end latency and less than 1usec hardware latency, making the case for using IB harder than it ever was. The next step in speed is slated for 2015, bringing iWARP to 100Gb and sub-microsecond end-to-end latency, going to **400Gb** with the following Ethernet iteration (IEEE 802.3bs [6]). In parallel, multiple vendors are working on extending the iWARP standards with features that cover the remaining IB capabilities, such as Atomics [7]. This practically obviates the need for this particular "Ethernot" technology, that is likely to join Token Ring, ATM, FibreChannel and other contenders to Ethernet's preferred position in everyone's heart and mind.

In fact, this realization – the inevitable decay of IB's advantages compared to Ethernet with iWARP – has resulted in the IB vendor-spawned standard for InfiniBand over Ethernet, called RoCE.

A Note on RoCE

RDMA over Converged Ethernet, or RoCE, is indeed InfiniBand over Ethernet, where the transport and network layers of IB are replaced by raw Ethernet encapsulation. Long claimed to be routable, RoCE is currently undergoing a **major overhaul** to include UDP and IP layers to actually provide that capability, clearly a **non-backward compatible** revision. Even with this change, RoCE Version 2 still **requires lossless** networks with **complex and restrictive** configuration to get it working, and places the burden of correct operation on the users and IT staff, while presenting them with a **hard to debug**, unfriendly stack [8]. This has hindered its deployment in general, and in the large scale applications in particular, and definitely **precludes** it from **high-speed wireless** applications. It is now clear that RoCE is on a costly journey to rediscover the foundation of the Ethernet world, in spite of a tried and ready standard solution being available, in what may merely be an attempt to maintain InfiniBand's market presence.

Summary

This paper discussed the past, present and future of iWARP. The past decade of experience has enabled **iWARP to mature** and increase in robustness, to get included in all major software distributions, all the while gaining in **performance and capabilities**. A true **plug-and-play native of the Cloud and datacenter** era, iWARP is the safe RDMA over Ethernet solution that is available today at **40Gbps from multiple vendors**.

References

[1] OpenFabrics Enterprise Distribution, OFED Overview
[2] CAIDA, Analyzing UDP Usage in Internet Traffic
[3] WiFi Alliance, WiGig Certified
[4] IBM, 40GbE A Competitive Alternative to InfiniBand
[5] Chelsio Communications, SMB Direct RDMA Performance
[6] IEEE P802.3bs, 400Gb/s Ethernet Task Force
[7] Internet Draft, RDMA Protocol Extensions
[8] Chelsio Communications, *RoCE Plug and Debug*

Further Reading

Chelsio Communications, *RoCE the Fine Print*
Chelsio Communications, *RoCE FAQ*
Chelsio Communications, *RoCE at a Crossroads*
Chelsio Communications, *RoCE is Dead, Long Live RoIP?*

