InfiniBand’s Fifteen Minutes

InfiniBand (IB) has languished away from the limelight since its inception in 1999, remaining largely unknown outside of a close circle of early adopters and niche applications. Indeed, even as it continued to enjoy a significant speed advantage over the Ethernet of the day, IB has seen limited deployment and market traction for the best part of a decade. Ironically, while it seems today to experience some long elusive measure of success, IB has never been as close to being completely obviated by Ethernet.

This paper starts by giving a brief overview of InfiniBand and its development, discussing the reasons for its recent rise in adoption. It then presents Ethernet’s answers to InfiniBand’s strengths, and argues that every one of them is either already matched or exceeded, or would be so within a year. Left with no intrinsic value to justify the adoption pain, all indications are that InfiniBand’s fifteen minutes of fame are soon to be over.

Background

Conceived as a high speed system interconnect, InfiniBand was designed for relatively short distance communication, with static routing and strict link-by-link flow control. Nevertheless, it soon started to be promoted as an all encompassing fabric, poised to replace both local busses for intra-system component connectivity, and inter-system networking and storage communication.

Despite these lofty aspirations, IB failed to make inroads in both, as PCI-Express took over within the system and the esoteric fabric failed to displace incumbent networking technologies. Unable to penetrate at either intra-system level or wide area uses, it remained limited to short span high speed compute cluster communication. It has recently seen some deployment in similarly constrained storage interconnects. In such environment, IB’s strengths offered an edge over the available Ethernet technologies.

InfiniBand’s Strengths

InfiniBand has enjoyed a number of advantages over Ethernet, which have enabled it to gain traction in some specialized environments.

High Speed

For the larger part of its existence, IB enjoyed a link speed and general performance advantage over Ethernet, mainly thanks to the use of serial communication. Before proceeding further, it would be useful to note that IB’s speed ratings are doubly confusing at best, if not intentionally misleading. The naming convention is in units of “Data Rate”, where Single Data Rate is a claimed data rate of 2.5Gbps but a raw data rate of 2Gbps (i.e. removing signaling overhead but not protocol overhead). Most adapters have multiple such lanes, further confusing the users.

At the time of IB's introduction then, IB links could operate at several multiples of the bandwidth of Gigabit Ethernet, although normally limited by the PCI bus speed available. It took until 2003 when the first 10Gbps Ethernet products started shipping for a comparable speed to be available (both IB and Ethernet would then be limited by PCI bus speed – first at 7Gbps then 14Gbps half-duplex). However, 10Gbps Ethernet port prices started out significantly higher for both switches and adapter cards. As Ethernet's economies of scale kicked into gear, Ethernet prices dropped until parity was recently reached.

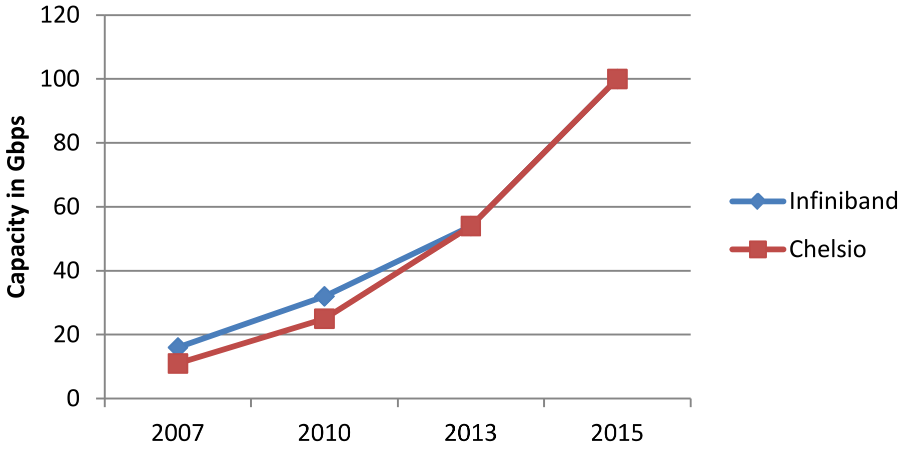

Figure 1 – InfiniBand vs. Ethernet Available Capacity

The advent of PCI-Express brought system bus speeds to new levels, allowing full duplex 10Gbps operation with Gen1 (2006), doubling that with Gen2 (2008), and again with Gen3 (2012). In the meantime, IB single port speeds scaled accordingly, while Ethernet remained at 10Gbps, making use of link aggregation to increase effective link capacity. However, Ethernet 40G and 100G speeds have recently become available, closing the link speed gap with IB for the first time in almost a decade. Going forward, having adopted similar physical layer (SERDES) specifications, Ethernet is no longer at a disadvantage.

Low Latency

IB has maintained an edge in latency over Ethernet for most of the past decade. At a time where Ethernet switches imparted latencies in the 10s of microseconds, IB virtual circuit switches were in the microsecond levels. However, low latency cut-through Ethernet switches appeared in 2008 and have since leveled the playing field, with switch latencies as low as 200 nanoseconds. On the adapter side, latencies have progressively dropped as illustrated in the following graph, which shows the evolution of back-to-back latency for the two technologies. A key observation is apparent: both IB and Ethernet have basically approached speed of light limits, and within a matter of months any IB latency advantage will have evaporated.

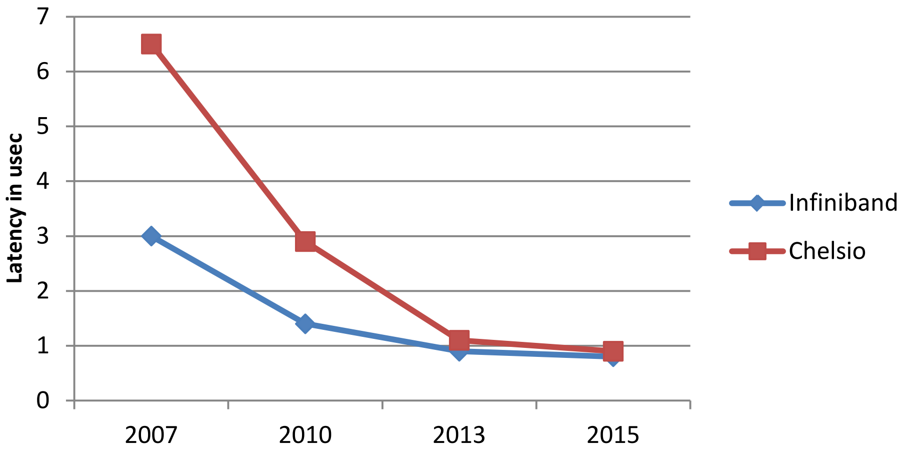

Figure 2 – InfiniBand vs. Ethernet Latency

RDMA Efficiency

Finally, a key differentiator for IB and the main reason for its recent resurgence is the Remote DMA (RDMA) communication interface it provides. RDMA allows very efficient communication, where most of the data transfer is handled silently by the adapter, without the involvement of the main CPU. Thus, RDMA frees up the cycles for the host system to process useful application workloads. In the datacenter age, at a time where system efficiency and power savings are critical metrics, increased efficiency is directly translated into dollars – both in terms of CAPEX and OPEX. Although making use of RDMA requires rewriting of applications, the gained efficiencies offer sufficient return on investment in areas such as High Performance Compute (HPC), storage system back-end and some datacenter and cloud applications.

Since 2007, however, Ethernet has in iWARP a standard RDMA specification (IETF standards [RFC 5040] and [RFC 5041]), with products from Broadcom, Chelsio and Intel. Second generation iWARP products are now available, with application level performance matching that of current IB technology. Third generation iWARP products will close the last gap by offering competitive performance at the micro benchmark level. The following section discusses iWARP in more detail.

iWARP

The standard Internet Wide Area RDMA Protocol (iWARP) offers a native RDMA verb interface implemented over TCP/IP/Ethernet, and can be easily interchanged with IB without any application modifications. In fact, Chelsio's implementation has been supported in a common software distribution with InfiniBand since 2008 (OpenFabrics Enterprise Distribution [OFED]).

In contrast, IB over Ethernet (marketed as RoCE) is commonly regarded as an attempt by an InfiniBand vendor to alleviate the concerns of running an esoteric infrastructure, with a solution of limited applicability and scalability. Perhaps good enough for back-to-back micro-benchmarks, RoCE's lack of network protection mechanisms and single subnet limitations raise serious concerns once a more realistic data center or cloud environment is considered.

An Ethernet native, iWARP easily coexists with all other traffic and implements network critical congestion and flow control mechanisms. Unlike IB – and RoCE – iWARP is routable, robust against packet loss, goes over long distances and runs over legacy equipment rather than expensive switches with oversized packet buffers, and is therefore significantly more scalable.

Performance

InfiniBand today enjoys an advantage over Ethernet in micro-benchmarks. However, multiple studies have shown that this perceived superiority dissipates when looking at actual useful application level performance. Chelsio's hardware iWARP implementation has been shown to offer identical application level performance to the competing IB speeds in several studies. Even in terms of micro-benchmarks, Ethernet is poised to close the gap in short order. As latency reaches the sub-microsecond levels, it is well into the zone of diminishing returns, nearing speed of light limits. iWARP will attain this level of performance by 2014, and will eliminate a critical pillar of the IB marketing structure.

Note that unlike previous speed transitions, 40Gbps and 100Gbps Ethernet are coming in close succession and will not offer the usual window of opportunity for other technologies to gain traction.

Deployment

Despite its recent successes, InfiniBand remains a largely unknown and unfamiliar technology, with confusing terminology and bandwidth/goodput characterization. Moving to IB requires completely new network monitoring and management tools and procedures. In addition, running lifeblood IP traffic over IB is complex and unnatural, typically resulting in a two network configuration. Therefore, the needs and costs in terms of personnel and training are significant. Furthermore, IB is effectively sole-sourced and suffers from the double risk of price gouging and/or vendor acquisition by a single OEM. In contrast, the Ethernet ecosystem is diverse and competition abounds. Further, iWARP requires no changes to existing Ethernet infrastructure. iWARP traffic can be observed, monitored, debugged and managed identically to all other TCP/IP and Ethernet traffic, using familiar tools such as Wireshark and tcpdump. Therefore, iWARP is significantly more cost effective overall than IB.

Conclusion

The matchup of InfiniBand vs. Ethernet is not the first where Ethernet eventually absorbs the strengths and builds up the feature set and performance levels needed to obviate the need for an esoteric technology, and take over its market. With a very high performance, robust, native Ethernet RDMA implementation, Chelsio is uniquely positioned to deliver InfiniBand's Ethernet nemesis.

Related Links

IBM Research Report on IB and 10GbE Performance for HPC Applications
IBM/Blade Networks Presentation
Cisco 10G for ECLIPSE Reservoir Simulation
Purdue University 10GbE Coates Cluster Whitepaper
Open Fabrics Enterprise Alliance

References

[RFC 5040] Recio et al., "A Remote Direct Memory Access Protocol Specification", RFC 5040, October 2007.
[RFC 5041] Shah et al., "Direct Data Placement over Reliable Transports", RFC 5041, October 2007.