QLogic's InfiniBand Technology Selected by Top-Tier HPC Vendors

Roy Krischer

QLogic Inc. announced on November 16th, 2009 a number of new partnerships involving its InfiniBand products. HP will resell a range of QLogic's QDR InfiniBand offerings including host channel adapters (HCAs), directors, and switches as part of the HP Unified Cluster Portfolio. SGI also selected QLogic's 12000 series QDR directors and switches with its CloudRack high-performance computing (HPC) offerings. Similarly, Dell will offer QLogic QDR switches and directors in addition to the QLogic HCAs it already sells. Finally, IBM will integrate QLogic's 7300 series QDR InfiniBand host channel adapters into its System x servers as part of the IBM System Cluster 1350. This agreement expands on a previous one in which IBM similarly selected QLogic's TrueScale-based 12000 Series QDR InfiniBand switch. These developments increase QLogic's footprint in the QDR InfiniBand business as well as its access to the HPC market and pose a challenge to InfiniBand vendors Mellanox and Voltaire.

The Technology

High-Performance Computing

High-Performance computing (HPC) is a discipline in computing that addresses extreme computational problems. Much scientific, governmental, or industrial research involves large-scale numerical computation such as computational fluid-dynamics. Examples include weather forecasting, climate modeling, nuclear detonation simulation, crash test simulation, drug testing, financial modeling, and electronic product engineering. These computational requirements far exceed the CPU power and memory bandwidth of a single host. Hence, HPC systems are usually clusters, i.e., a large number of hosts (nodes) that are connected among each other through a local area network (LAN). In order for a computational problem to be efficiently solvable in such a system, it needs to be divided into sub-problems that can then be solved in independent threads on the individual hosts. Unfortunately, most interesting problems do not allow for these threads to operate in perfect independence from one another. Instead, they need to communicate, i.e., exchange data. Since network I/O is generally orders of magnitude slower than data processing in the CPU, scalability of computation is inversely related to the amount of communication between threads/hosts required. The choice of networking technology to connect the nodes in such a cluster is therefore essential: depending on the problem to be computed, bandwidth and, often more importantly, latency limitations can dramatically alter the computation time. Given the large scale of such an HPC installation, practical scalability of the network and its energy consumption are further considerations.

As an example, the long-time top-performing cluster (recently dropped to second place) is IBM's Roadrunner in Los Alamos, New Mexico, USA, with 6,480 Opteron processors containing 2 cores each, 12,960 PowerXCell 8i processors containing 9 cores each, and

a total of 103.6 TB of RAM (computational units only). These are distributed among 270 racks of 12 TriBlade nodes each, for a total of 3240 TriBlades. These nodes employ a two-tiered InfiniBand topology as their network.

InfiniBand

InfiniBand (IB) is a high-performance local networking technology based on open standards. It employs a switched-fabric topology in which hosts (or rather: their host channel adapters, HCAs) connect to each other via switches. Peripherals such as storage hardware can also be attached using target channel adapters (TCAs).

Data transmission is performed over serial bi-directional links of 2.5 Gb/s, 5 Gb/s (double data rate, DDR), or 10Gb/s (quad data rate, QDR). These links can be aggregated, most commonly in bundles of 4x yielding 50 Gb/s of maximum raw bandwidth (with QDR); for extreme bandwidth requirements,12x bundling can yield a maximum bandwidth of 120 Gb/s. Data packets are called *messages* and can be up to 4 kB in length. The latency of an application message (e.g., MPI) between network end points is in the range of 1-2 µs.

InfiniBand offers quality-of-service, failover, and special features such as remote direct memory access (RDMA), which allows a host to access the memory of another remote host directly without incurring the extra overhead of involving the remote host's operating system or CPU. These properties and the outstanding performance in terms of bandwidth and latency have made InfiniBand the de-facto standard in top high-performance computing.

According to the TOP500[1] list from November 2009, InfiniBand is employed as the interconnect technology in 36% (up from 30%) of the listed systems. While Gigabit Ethernet has a share of 52% (down from 56%), this apparent majority mainly represents smaller installations. When it comes to high-end systems, InfiniBand rules supreme, with 9 out of the 20 fastest systems (including the aforementioned Roadrunner), whereas Gigabit Ethernet has no entry in the top-20. Thus, the 181 InfiniBand systems have a combined maximum performance that is 70% higher than the 259 systems employing Ethernet. More tellingly, when comparing actual with theoretical performance given sheer CPU power, Ethernet-equipped systems achieve only 50% of their theoretical combined maximum, whereas InfiniBand systems manage to yield 77%. Since interconnect performance is a major contributor to discrepancies between actual vs. theoretical performance, this result underlines the superiority of InfiniBand compared to Gigabit Ethernet.

[1] The TOP500 list (http://www.top500.org) is compiled biannually and ranks computer systems according to benchmark results submitted. It is generally regarded as the global supercomputer ranking list. Note, not all organizations submit benchmark results for their systems, either because they cannot or choose not to, so the list is not exhaustive. Furthermore, the Linpack benchmark employed is based on solving a massive set of dense linear equations, which generally offers good scalability, but may or may not be an adequate approximation of a system's performance depending on what the system is actually used for.

Products

QLogic's InfiniBand offerings are based on its TrueScale architecture, which employs ASIC technology to provide highly-integrated, low-power, high-performance InfiniBand solutions. QLogic's expertise in ASIC design is apparent throughout its product portfolio, e.g., its Fiber Channel over Ethernet (FCoE) converged network adapters (see also [1]).

The QLogic 7300 series x8 PCI-Express 2.0 HCAs offer a unidirectional bandwidth of 3400 MB/s, with a message rate of up to 30 million per second. According to QLogic, this message rate is five times that of its competitors and is achieved by taking better advantage of modern multi-core architectures. The 7300 series sports a very low MPI latency of 1 µs that, due to its connection-less design, remains low as the number of nodes scales. Its typical power consumption of around 6W is the lowest in the industry.

QLogic's 12000 series of QDR InfiniBand director switches are available in a variety of flexible configurations, from 18 up to a maximum of 864 ports, which is the largest in the industry (competing products support up to 648 ports, though this configuration requires two interconnected units). Installation, configuration, and fabric monitoring are eased through QLogic's comprehensive InfiniBand Fabric Software Suite.

The Players

QLogic

QLogic is a leading supplier of high-performance networking solutions for FCoE, Ethernet, Fiber Channel (FC), iSCSI, and InfiniBand. Its portfolio includes devices such as network adapters and switches, as well as technology supplied to various tier-1 OEM vendors. According to market research firm Dell'Oro Group, QLogic is the leading provider of FC technology, with a market share by revenue of over 50%. It has extensive know-how in using ASIC technology to create highly-integrated products that provide high performance with low energy requirements.

The Partners

According to the November 2009 TOP500 list, 42% of all listed systems were sold by HP, making it the largest contributor by number of systems. It is followed by IBM, which sold 37% of the systems, though this number can be misleading as IBM's combined performance of 9.7 PFlops is higher than HP's of 6.3 PFlops. SGI follows with 3.8% and Dell is in fifth place with 3.2% (Dell could be awarded another 0.8% for systems built collaboratively with other companies or by subsidiaries, putting it in third place). Overall, these QLogic partners make up an overwhelming 86% of all TOP500 systems.

The Bottom Line

HPC is a growing field. While the computational problems for traditional customers such as governments and academic institutions do not go away, additional HPC applications in commercial interests keep popping up. In tough economic times such as those experienced recently, a corporation has to make tough strategic choices. Cost cutting can help in the short term, but in the long term, innovation is the better bet to ensure a

prosperous future. Without innovative new products to entice them, customers who are already used to cutting costs may otherwise be content to make do with what they already own. In order to drive innovation, companies need to invest in research and development, which nowadays increasingly involves large-scale computation. Hence, HPC installations are bound to grow in number, and since InfiniBand is *the* interconnect technology standard for HPC, the future for InfiniBand should be quite rosy.

It is not always easy to compare performance numbers between competitors as these depend heavily on the network topology, the benchmark (parameters) chosen, and a myriad of other details. However, some numbers can be compared: QLogic's 12000 series director offers the largest number of ports in the market place, which is important considering the continuously growing scale of clusters. Additionally, tests under the industry-standard SPEC-MP2007 benchmark reveal that QLogic's new QDR adapters perform faster as more cores are added to a cluster: at 256 cores, the 7300 Series HCAs perform over 20% faster than comparative ConnectX adapters from Mellanox, all while exhibiting the lowest power consumption. Again, with the growing scale of installations and energy prices that are more likely to rise than to fall (apart from environmental concerns), low power consumption along with peak performance is top of mind for many HPC sites.

In the end, the sheer breadth of vendors that now rely on QLogic's QDR InfiniBand technology should be a good indicator of its merits. If four out of the top-5 HPC vendors that together sold 86% of the TOP500 systems trust QLogic technology, this can be regarded as a strong endorsement from those in the know. Especially noteworthy is IBM, the largest HPC vendor by performance: After recently choosing to integrate QLogic's FCoE technology into its products, it now selects QLogic's QDR InfiniBand technology as well. QLogic's business is clearly gaining momentum across all of the networking markets it currently serves.

References

1. **Krischer, Josh.** *QLogic Continues to Innovate in Connectivity and Network Infrastructures.* 2009. http://joshkrischer.com/files/QLogic_CNA.pdf.