Cluster Interconnects:
Real Application Performance and Beyond

Gilad Shainer, MSc., Mellanox Technologies

Scientists, engineers and analysts in virtually every field are turning to high performance computing to solve today's vital and complex problems. Simulations are increasingly replacing expensive physical testing, as more complex environments can be modeled and in some cases, fully simulated.

High-performance computing encompasses advanced computation over parallel processing, enabling faster execution of highly compute intensive tasks such as climate research, molecular modeling, physical simulations, cryptanalysis, geophysical modeling, automotive and aerospace design, financial modeling, data mining and more. HPC clusters become the most common building blocks for high-performance computing, not only because they are affordable, but because they provide the needed flexibility and deliver superior price/performance compared to proprietary symmetric multiprocessing (SMP) systems, with the simplicity and value of industry standard computing.

Maui High Performance Computing Center
1280 servers, Mellanox InfiniBand interconnect, 42.3TFlops

Real-world application performance depends on the performance of the various cluster's key elements – the processor, the memory, and the interconnect. The interconnect controls the data transfer between servers, and has a high influence on the CPU efficiency and memory utilization.

Transport offload interconnect architectures, unlike the "on-loading" ones, eliminate the need of dealing with the protocol processing within the CPU and therefore increasing the number of cycles dedicated toward computational tasks. If the CPU is busy moving data and handling network

protocol processing, it is unable to perform computational work, and the overall productivity of the system is severely degraded.

The memory copy overhead includes the resources required to copy data buffers from the network device to the kernel memory and then from the kernel memory to the application memory. This approach requires multiple memory accesses before the data is placed in its final destination. While it is not a major problem for small data transfers, it is a big problem for larger data transfers. This is where the interconnect zero-copy capabilities eliminates the memory bandwidth bottleneck without involving the CPU in the network data transfer.

Sandia National Lab
4500 servers, Mellanox InfiniBand interconnect
53TFlops, 84.66% Linpack efficiency

The interconnect bandwidth and latency have traditionally been used as two metrics for assessing the performance of the system's interconnect fabric. However, these two metrics are typically not sufficient to determine the performance of real world applications. Typical real-world applications send messages ranging from 64 Byte to 4 Megabyte using not only point-to-point communication but a diverse mixture of communication patterns, including collective and reduction patterns in the case of MPI. In some cases, interconnect vendors create artificial benchmarks, such as message rate, and apply bombastic marketing slogans to these benchmarks – such as "Hypermessaging". Message rate is yet another single point in the point-to-point bandwidth graph. If the traditional interconnect bandwidth indicates the maximum available bandwidth (single point), message rate indicates the bandwidth for message size of zero or 2 bytes.

The single points of data, give some indication for the interconnect performance, but are far from describing the real world application performance. The interactive combination of those points, together with others (CPU overhead, zero copy etc.), will determine the overall ability of the connectivity solution.

The difference between theoretical power and what is actually delivered is measured as processor efficiency. The more CPU cycles used to get the data out the door by "filling the wire" due to protocol and data transfer inefficiencies, the less cycles are available for the application. When comparing latencies of different interconnects, one needs to pay attention to the interconnect architecture. 1usec latency "on-loading" interconnect versus 2usec latency "off-load" solution is similar to a case when one needs to decide between two cars that show the same horsepower (i.e. CPU). Both engines are capable of 200 miles per hour, but the first car, due to "on-loading", limits the actual engine power to 75 miles per hour (the engine power must be used for other tasks). The Second car has no limitations on the engine, but its wheels can tolerate only 150 miles per hour. The knowledge on the wheels tolerance (i.e. latency), as a single point of data, is definitely misleading.

There are attempts to provide real world application performance while comparing different interconnects, but in most cases the "comparison" is biased and by using different systems and/or conditions, which makes a true comparison difficult. There have been recent cases comparing 10-Gigabit Ethernet to InfiniBand. While InfiniBand adapters were tested with PCIe x4 (that is limited to ~700MByte/sec bandwidth (due to limitations in the current available systems), the 10 Gigabit Ethernet cards were PCI-X, that is capable to higher bandwidth (~850-900MByte/s). Other cases compare InfiniBand PCIe x4 to other interconnects with PCIe x8 host interface (the only valid conclusion one can make is that PCIe x8 has more lanes than PCIe x4). Another paper compared QLogic InfiniPath on Intel 3GHz CPU based system to Mellanox InfiniBand on 2.2GHz Opteron based system. Any attempt to compare different interconnects in those manners is deceptive.

Real application performance

InfiniBand is a proven interconnect for clustered server solutions, and one of the leading connectivity solution for high-performance computing. InfiniBand was designed as a general I/O and in practice provides low-latency and the highest link speed.

Computational Fluid Dynamics (CFD) is one of the branches of fluid mechanics that uses numerical methods and algorithms to solve and analyze problems that involve fluid flows. ANSYS/FLUENT is a leading commercial software provider for solving fluid flow problems. The broad physical modeling capabilities of FLUENT have been applied to industrial applications ranging from air flow over an aircraft wing to combustion in a furnace, from bubble columns to glass production, from blood flow to semiconductor manufacturing, from clean room design to wastewater treatment plants. The ability of the software to model in-cylinder engines, aero acoustics, turbo machinery, and multiphase systems has served to broaden its reach. At the core of any CFD calculation is a computational grid, used to divide the solution domain into thousands or millions of elements where the problem variables are computed and stored. In FLUENT, unstructured grid technology is used, which means that the grid can consist of elements in a variety of shapes: quadrilaterals and triangles for 2D simulations, and hexahedral, tetrahedral, prisms, and pyramids for 3D simulations. These elements form an interlocking network throughout the volume where the fluid flow analysis takes place.

The performance of a CFD code depends on several factors, including size and topology of the mesh, physical models, numerics and parallelization, compilers and optimization, in addition to performance characteristics of the hardware where the simulation is performed. FLUENT provides a set of benchmark problems which represent typical current usage and covering a wide range of mesh sizes and physical models. The problems selected represent a range of simulations typical of those which might be found in industry. The principal objective of this benchmark suite is to provide comprehensive and fair comparative information of the performance of FLUENT on available hardware platforms.

The following charts compares Mellanox InfiniBand and QLogic InfiniPath interconnects on the same platform – dual core, dual socket, Intel Xeon 3GHz 5100 series (code name Woodcrest) servers, using FLUENT benchmarks. When testing real world applications, the entire architecture makes the difference. The Mellanox architecture is a full transport-offload one, with hardware capabilities of RDMA, while QLogic is a full "on-loading" architecture.

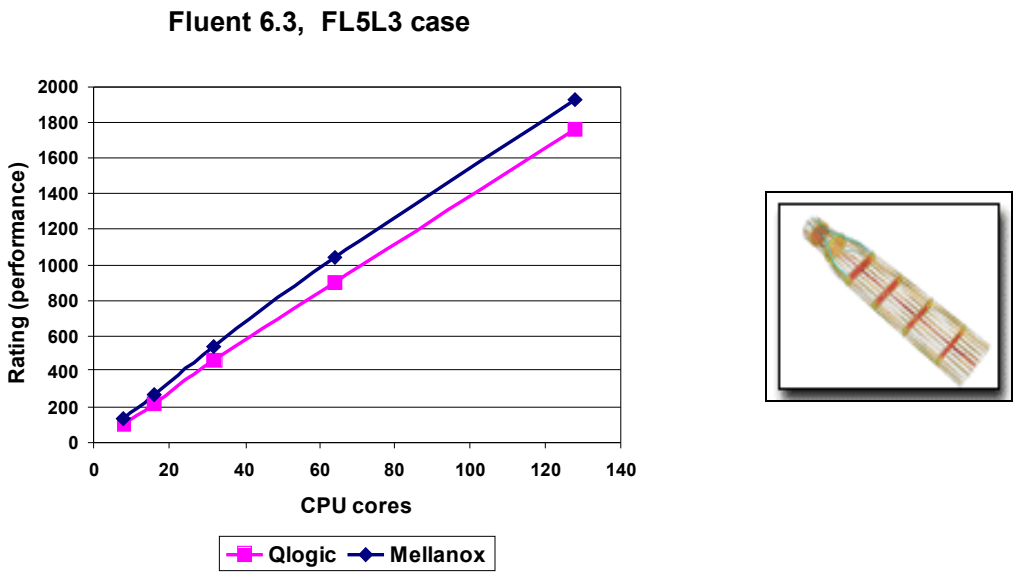

Figure One: Turbulent flow of air through a duct

In Fluent FL5L3 benchmark, a Turbulent flow of air through a duct is computed. The cross-sectional planes of the duct transition from a circle at the inlet to a rectangle at the outflow boundary. The Reynolds-Stress Model is used for computing turbulence (number of cells: 9,792,512, cell type hexahedral, models RSM turbulence, solver segregated implicit).

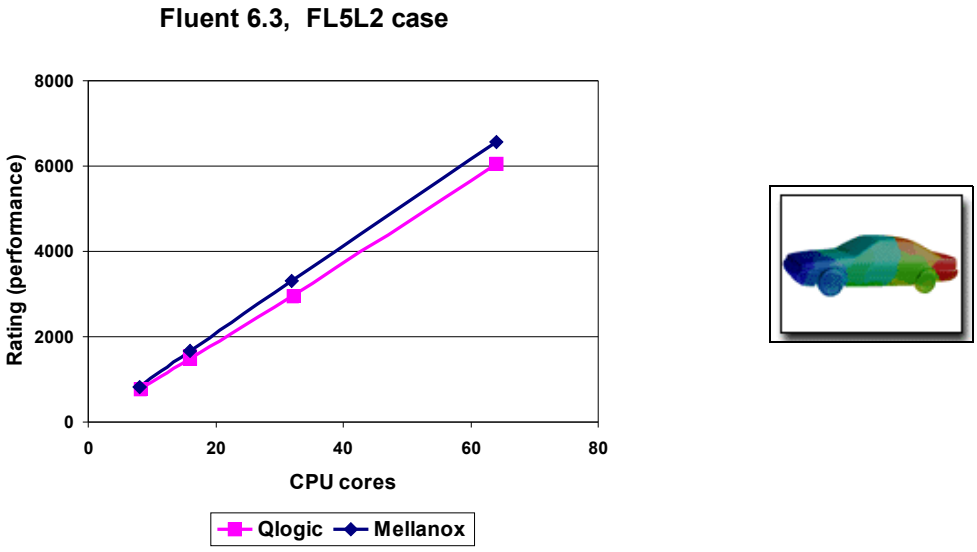

Figure Two: Air flow around sedan

FLUENT FL5L2 benchmark represents the computation of the exterior flow field around a simplified model of a passenger sedan. The simulation geometry was used for the Japan External Aerodynamics competition. A viscous-hybrid grid with prismatic cells is used to adequately model the boundary layer regions (number of cells 3,618,080, cell type hybrid, models k-epsilon turbulence, solver segregated implicit).

Choosing the right interconnect

In both cases of FLUENT benchmarks, Mellanox InfiniBand shows higher performance and better super-linear scaling comparing to QLogic InfiniPath.

FLUENT's CFD application is a latency-sensitive application, and the results shown here are good examples on how pure latency benchmarks can be misleading when choosing the right interconnect. In order to determine the system's performance, one should take into consideration the entire interconnect architecture (such as off-loading versus on-loading) and the ability of scaling, rather than just single points of data.

In order to provide better applications sight, Mellanox has created the Mellanox Cluster Center. The Mellanox Cluster Center offers an environment for developing, testing, benchmarking and optimizing products based on InfiniBand technology. The center, located in Santa Clara, California, provides on-site technical support and enables secure sessions onsite or remotely. More details can be achieved through Mellanox web site (*http://www.mellanox.com*).

Gilad Shainer is a senior technical marketing manager at Mellanox technologies focusing on high performance computing. He joined Mellanox Technologies in 2001 to develop Mellanox's InfiniHost PCI-X Host Channel Adapter (HCA) device and later led the development of Mellanox's InfiniHost III Ex PCI Express HCA device. Gilad Shainer holds MSc. degree (2001, Cum Laude) and a BSc. degree (1998, Cum Laude) in Electrical Engineering from the Technion Institute of Technology in Israel. He is also a member of the PCISIG PCI-X and PCI Express Working Groups and has contributed to the definition of the PCI-X 2.0 specifications.

