

CASE STUDY

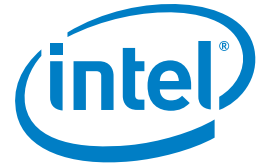
Schlumberger ECLIPSE* 2010 Reservoir Simulation Application

NetEffect™ 10 Gigabit Ethernet Server Cluster Adapters

Internet Wide Area RDMA Protocol (iWARP)

Arista 10 Gigabit Ethernet Switching

Technical and High-Performance Computing



Reservoir Simulation Made Simpler and More Efficient with 10 Gigabit Ethernet



By Owen Brazell, Schlumberger
Steve Messenger, Schlumberger
Tom Stachura, Intel Corporation
Jamie Wilcox, Intel Corporation

CHALLENGE

How can petroleum production be made dramatically more efficient while minimizing the associated capital and operating expenses? Specifically, the objective is to enable very high performance and scalability while leveraging all of the familiar and cost-effective attributes of Ethernet, thus avoiding complex, specialized fabrics.

Schlumberger
ARISTA

SOLUTIONS

Combining ECLIPSE* 2010 reservoir engineering software with 10 Gigabit Ethernet iWARP (Internet Wide Area RDMA Protocol) provides sophisticated modeling that scales, while retaining the advantages of Ethernet. ECLIPSE provides native support for the iWARP protocol to provide the advanced parallelization necessary with multi-core, multi-node installations. Intel Ethernet iWARP delivers the required low latency and high message rate performance, providing a competitive alternative to Infiniband*. When coupled with low-latency, non-blocking, 10 Gigabit Ethernet (10GbE) Arista Networks switches, application performance can be taken to a new level.

CUSTOMER BENEFIT

Using ECLIPSE 2010 with 10GbE iWARP enables performance on par with InfiniBand while realizing the key benefits of using an Ethernet-based fabric:

- **Ease of use.** Ethernet deployment and IP management is familiar and common across data center and cluster facilities by existing personnel.
- **IP routing and storage.** Ethernet enables seamless integration of network-attached storage (NAS) file systems while supporting IP subnets and routing.
- **Fabric consolidation.** A 10GbE switching fabric provides for the consolidation of management and applications to reduce ports, cables, and complexity.

Better Reservoir Engineering Enables Smarter Petroleum Drilling

Oil and gas companies must strive to be more competitive while working on increasingly complex and remote reservoirs. More sophisticated computer modeling and simulation are keys to reducing the uncertainty and risk of working in that environment. ECLIPSE 2010 from Schlumberger helps optimize oil and gas recovery. In addition to increasing operational efficiency, that optimization also reduces the uncertainty of production and its impact on the environment.

Running on high-performance computing clusters, ECLIPSE 2010 quickly and accurately predicts the dynamic behavior of reservoirs with varying complexity, structure, and geology. Customers can tailor simulator capabilities to their specific needs with a wide variety of add-on options for coalbed methane, gas field operations, calorific value-based controls, reservoir coupling, and surface networks. The capabilities of the ECLIPSE Compositional Simulator make this modeling more sophisticated:

- **Multi-component hydrocarbon flow.** Describing reservoir fluid-phase behavior and compositional changes improves the accuracy of modeling complex reservoir dynamics.
- **CO₂ enhanced oil recovery.** Modeling depleted reservoirs and gas mixtures in aquifers helps remove the economic constraints associated with production in these areas.
- **Shale gas and hydraulic fractures.** Characterizing heterogeneous shale gas reservoirs and the associated directional stresses accurately improves modeling efficiency and flexibility.

Because ECLIPSE 2010 is designed to operate on high-performance computing clusters, it benefits dramatically from the software parallelism that Schlumberger introduced in this latest version of the product. Distributed-memory parallel code allows the ECLIPSE 2010 environment to smoothly divide work among the many processor cores available in a

typical cluster topology, a vital aspect of scalability in these systems. By applying large-scale compute resources, oil and gas companies can run larger numbers of simulations in an acceptable time span, ultimately producing more sophisticated results.

Software parallelism is a vital aspect of taking advantage of the increasingly parallel hardware environments in which ECLIPSE tools operate. For example, a typical Intel® Xeon® processor 5600 series used in a modern cluster has six cores, for a total of 12 cores in a two-processor server and 48 cores in a modest four-server cluster. The sophisticated architecture of ECLIPSE 2010 helps take advantage of this hardware parallelism to deliver very high performance and scalability.

Customers Need an Effective Alternative for High-Performance Fabric

Because parallel processing is so vital to large-scale computations such as those associated with reservoir engineering, the environments in which ECLIPSE 2010 runs must be highly optimized to that purpose. High core counts, Intel HT Technology, and advanced support from the memory and I/O subsystems in platforms based on the latest Intel Xeon processors are one key to the robustness of these cluster designs. Selection and implementation of ideal networking technologies is equally important, from the perspectives of both performance and economics.

Traditionally, InfiniBand has been the fabric of choice for high-performance computing clusters such as those employed by ECLIPSE 2010. Unfortunately, the use of InfiniBand continues to have significant limitations that reduce its suitability:

- **Multiple fabrics.** The addition of an InfiniBand network alongside existing Ethernet for management and storage makes the cluster environment more complex and layered, adding to requirements in terms of cost, power, and density.

- **Required expertise.** InfiniBand uses a different paradigm for configuration, operation, and management than Ethernet. This typically requires trained personnel and specialized tools.

- **Non-IP management.** InfiniBand doesn't use IP management and therefore cannot leverage standard data center management consoles, limiting integration and compatibility with information technology policies and procedures.

Customers have long searched for a more efficient, easier-to-use alternative to InfiniBand that can meet their objectives for performance and flexibility. The maturation of iWARP provides an effective means of addressing these challenges of building high-performance computing clusters. Implementations such as reservoir engineering with ECLIPSE 2010 reap the benefits.

Single-Fabric Implementations with 10GbE iWARP Reduce Cost and Complexity

High-performance clusters conventionally use message passing interface (MPI) to allow processes to communicate with one another using messages, reducing the need for nodes in the cluster to access non-local memory. While this approach enables the very high degree of parallelization required by technical and scientific computing solutions such as ECLIPSE 2010, it also requires extensive communication over the network, which must be accomplished with low levels of latency. Remote Direct Memory Access (RDMA) arose as a means of delivering efficient network communication to support MPI within clusters.

RDMA allows direct, zero-copy data transfer between RDMA-capable server adapters and application memory, removing the need in Ethernet networks for data to be copied multiple times to OS data buffers. The mechanism is highly efficient and eliminates the associated processor-intensive context switching between kernel space and user space. ECLIPSE 2010 can therefore reduce

latency and perform message transfer very rapidly by directly delivering data from application memory to the network.

InfiniBand and iWARP both use RDMA and a common API for applications such as ECLIPSE 2010, but iWARP enables use of RDMA over the familiar Ethernet fabric. Because iWARP runs over Ethernet, it allows for both application and management traffic operating over a single wire. Unlike InfiniBand, iWARP is an extension of conventional IP, so standard IT management tools and processes can also be used to manage the traffic and resources associated with iWARP.

ECLIPSE 2010 implementations can use iWARP technology with NetEffect™ Ethernet Server Cluster Adapters from Intel and low-latency Arista 10GbE switches to provide a high-performance, low-latency Ethernet-based solution. By making Ethernet networks suitable for these high-performance clustering implementations, iWARP provides a number of benefits:

- **Fabric consolidation.** With iWARP technology, LAN and RDMA traffic can pass over a single wire. Moreover, application and management traffic can be converged, reducing requirements for cables, ports, and switches.
- **IP-based management.** Network administrators can use standard IP tools to manage traffic in an iWARP network, taking advantage of existing skill sets and processes to reduce the overall cost and complexity of operations.
- **Native routing capabilities.** Because iWARP uses Ethernet and the standard IP stack, it can use standard equipment and be routed across IP subnets using existing network infrastructure.
- **Existing switches, appliances, and cabling.** The flexibility of using standard TCP/IP Ethernet to carry iWARP traffic means that no changes are required to Ethernet-based network equipment.

Using 10GbE iWARP, end-customers have the option of meeting their performance needs for cluster implementations such as ECLIPSE 2010 with a superior cost and complexity profile compared to what they could expect with InfiniBand.

10GbE iWARP Delivers Equivalent Performance to 40Gb InfiniBand

To compare the performance of GbE, 10GbE iWARP, and InfiniBand, Intel engineers created test beds based on a one-million cell ECLIPSE model. The system and network configurations were identical for all three test beds, except for the connectivity fabric, as shown in Table 1. The 10GbE switch used was the Arista 7124S 10GbE switch, which provides ultra-low port-to-port latency and wirespeed forwarding. Combining iWARP support with the low-latency Arista switching enables Ethernet to deliver a cost-effective, high-performance computing solution as an alternative to InfiniBand.

Testing based on the ECLIPSE workload showed approximately equivalent performance between the GbE and 10GbE iWARP test beds in the single-node (12 cores) case, but poor scaling with GbE as the number of cores applied to the problem was increased, as shown in Figure 1. By contrast, the 10GbE iWARP test bed scales exceedingly well to the upper limit of the testing scenario, where the fabric can support 10 nodes (120 cores) without causing a significant increase in execution time. Significantly, migrating a GbE fabric to 10GbE iWARP solves bandwidth limitations while retaining the familiar Ethernet topology, making use of existing skill sets.

Table 1. System and Network Configurations Used in Performance Testing

Configuration Details Common to All Three Fabrics >	<ul style="list-style-type: none"> • Intel® Xeon® processors 5670 at 3.3 GHz (two-way servers, six cores per processor) • 16 GB RAM • Red Hat Enterprise Linux* 5.5 • ECLIPSE* 2010.2 Reservoir Engineering Software • OpenFabrics Enterprise Distribution (OFED) 1.5.2 • Intel® Message Passing Interface (MPI) 4 		
Configuration Details Specific to Individual Fabrics >	GbE Test Bed Configuration: <ul style="list-style-type: none"> • HP ProCurve* 2824 Ethernet switch • Intel® 82575EB Gigabit Ethernet Controller 	10GbE iWARP Test Bed Configuration: <ul style="list-style-type: none"> • Arista 7124S 10GbE switch • NetEffect™ 10GbE Serve Cluster Adapters 	InfiniBand* Test Bed Configuration: <ul style="list-style-type: none"> • Mellanox MTS3600 InfiniBand switch • Mellanox ConnectX* QDR (40Gb) InfiniBand Host Channel Adapters

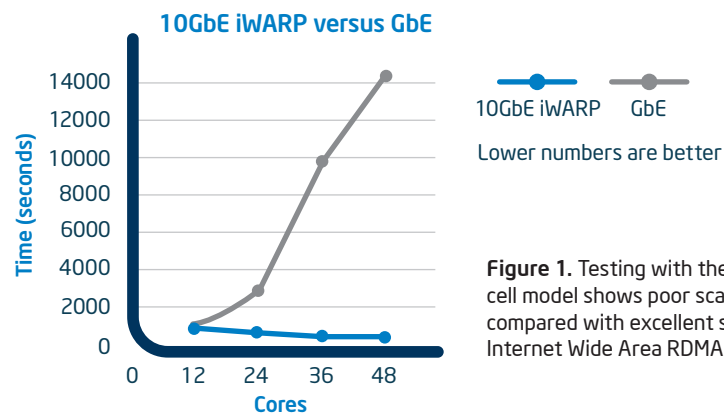


Figure 1. Testing with the ECLIPSE* one million cell model shows poor scaling with GbE, compared with excellent scaling using 10GbE Internet Wide Area RDMA Protocol (iWARP).

After establishing that 10GbE iWARP is an effective means of addressing the GbE scalability limitations, the testing team compared 10GbE iWARP performance directly to that of QDR (40Gb) InfiniBand. Despite the low-latency, high-bandwidth (40Gb) attributes of InfiniBand, 10GbE delivers comparable application results, as shown in Figure 2.

This testing showed that pairing the Arista 7124S 10GbE switch with NetEffect 10GbE Server Cluster Adapters provides an excellent means of delivering performance equivalent to that of InfiniBand for ECLIPSE implementations. The Intel 10GbE iWARP-based solution helps meet the demands of high-performance computing while retaining all the benefits of Ethernet, enhancing information technology skill sets, and providing both superior capital expense and operating expense profiles, compared to solutions that use InfiniBand.

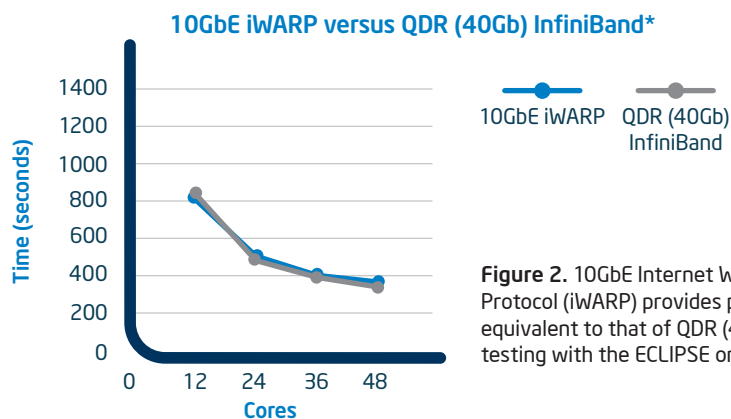
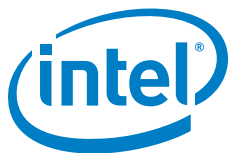


Figure 2. 10GbE Internet Wide Area RDMA Protocol (iWARP) provides performance equivalent to that of QDR (40Gb) InfiniBand* in testing with the ECLIPSE one-million cell model.

Conclusion

With the emergence of 10GbE iWARP from Intel and cost-effective, high-performance network solutions from Arista Networks, it is no longer necessary to implement complex and specialized fabric solutions for high-performance cluster implementations, including ECLIPSE 2010. Dramatic value advantages are available for the petroleum industry and other scientific and technical computing customers. Software providers such as Schlumberger have recognized this opportunity, and by validating and recommending the use of iWARP to their customers, they can now offer more favorable cost-benefit profiles, for a significant overall competitive advantage.

SOLUTION PROVIDED BY:



Learn more about NetEffect™ Ethernet Server Cluster Adapters:
www.intel.com/Products/Server/Adapters/Server-Cluster/Server-Cluster-overview.htm

Learn more about Arista 7100 Series Switches:
<http://www.aristanetworks.com/en/products/7100series>

Learn more about ECLIPSE 2010 Reservoir Engineering Software:
www.slb.com/services/software/reseng/eclipse2010.aspx

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance>.

Intel, the Intel logo, and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.