



Production HPC Clusters: State-of-the-Art Performance via a Best-of-Breed Solution Stack

A Technical White Paper by CD-adapco, SilverStorm and Scali, Inc.,

Introduction

From Top 500 presence to press releases, it's clear that there is an increasing use of clusters in High Performance Computing (HPC) deployments. With a fairly mature product and services ecosystem in place, this uptake is understandable. However as cluster adoption continues to enter the mainstream, it's important to identify and leverage the state-of-the-art software and hardware innovations that will result in the highest levels of performance in production HPC cluster deployments. By benchmarking the CD-adapco STAR-CD Computational Fluid Dynamics (CFD) application, state-of-the-art performance characteristics are demonstrated for best-of-breed components – i.e., AMD Opteron processors, SilverStorm Infiniband interconnects and Scali MPI Connect. In the most-compelling example, this component combination yields relative performance, speedups and efficiencies more than double those achieved by common historically used cluster technologies of MPICH MPI with Gigabit Ethernet interconnect. After briefly describing the benchmarking environment, results comparing these two environments are provided for two, independent CFD datasets; this is followed by a summary and compilation of resources.

Benchmarking Environment

The AMD Developer Center's Niobe cluster provides the reference configuration for benchmarking. Niobe consists of 256 AMD Opteron Model 246 processors (single-core, 2.0 GHz) in a two-processor-per-server configuration. Each server runs Rocks 3.3.0, a derivative of Red Hat Enterprise Linux 3.0 for AMD64 (Update 3). Niobe obtains 922.8 GFLOPS (measured) for High-Performance Linpack (HPL), and has placed on the Top 500. The only variations to this base platform are in the MPI implementation and interconnect.

Baseline results derive from MPICH – an Open Source implementation of the Message Passing Interface (MPI) running on Gigabit Ethernet. Against this baseline, results are obtained for the Scali MPI Connect implementation of MPI running on SilverStorm Technologies InfiniBand.

Scali MPI Connect makes use of an operating-system bypass mechanism. This bypass mechanism makes use of a standards-based (Remote Direct Memory Access, RDMA) implementation (Direct Access Provider Library, DAPL). Scali MPI Connect has a number of additional features and functionalities that both enhance and extend the implementation of the MPI standard – e.g., high availability through fault tolerance; see Resources for more.

SilverStorm 9120 is the InfiniBand switch in the benchmarks. Based on the InfiniBand standard, this switch offers a bandwidth of 840 Mbits/sec and latency of 5.2 microsec via a PCI-X bus. The SilverStorm interconnect exclusively supports the passing of messages relating to the CFD computations.

Using the CD-adapco STAR-CD CFD model with two input datasets, several combinations of benchmarks are provided. CD-adapco STAR-CD has been compiled for use with MPICH via Gigabit Ethernet and for use with Scali MPI Connect.

Benchmarking Results

Two datasets provide inputs for CD-adapco STAR-CD CFD models and establish the performance characteristics of the ‘configuration’ described above.

ACCLASS Benchmark

The ACCLASS dataset is the first input for the CD-adapco STAR-CD CFD model used to demonstrate the performance characteristics of the benchmarking configuration. This test case simulates the turbulent flow around an ACCLASS car and consists of 5,914,426 cells in a hybrid mesh. Figure 1 provides results for performance versus processor count in a bar-graph format. Because the Scali MPI Connect results with SilverStorm Infiniband have been scaled by the MPICH results, the vertical axis provides a percentage-based measure of relative performance. The combined effect of Scali MPI Connect and SilverStorm Infiniband High Speed Interconnect is illustrated by the dark-red bars of the figure. Systematically increasing performance relative to MPICH is evident throughout the range of processor counts, with a 135% achievement for 64 CPUs. Figure 1 quantitatively demonstrates the performance gains achieved by the combination of Scali MPI Connect and SilverStorm Infiniband.

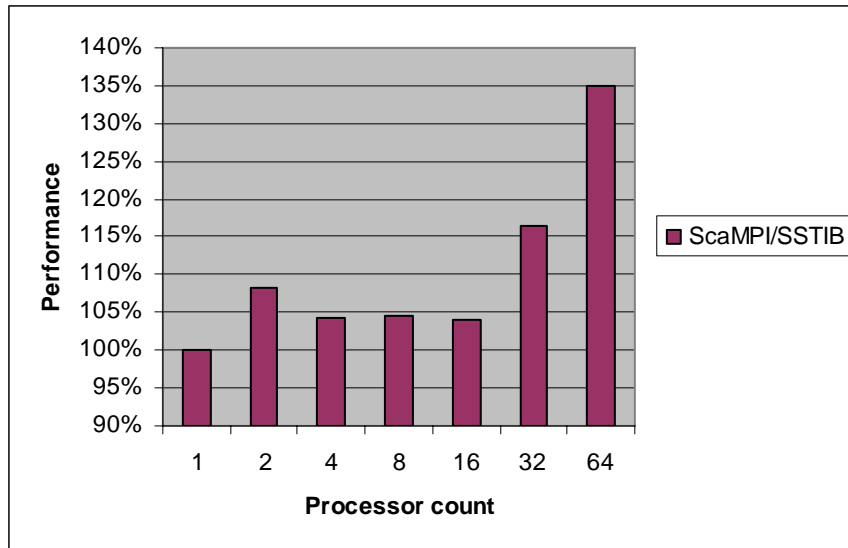


Figure 1 Performance versus processor count for the ACLASS dataset as input to the CD-adapco STAR-CD CFD model. Percentage-based performance values are normalized relative to the results for MPICH via Gigabit Ethernet.

Speedup¹ and efficiency calculations are routinely used to quantify the impact of parallel computing. In the ideal case where serial computations can be eliminated completely, speedup varies linearly with processor count. This linear speedup is shown by the teal-blue line that connects the “x” in Figure 2. (Note that the near-exponential appearance of the linear speedup data in this figure is an artifact of the graphing – specifically the axes.) Also illustrated in Figure 2 are speedup results for the CD-adapco STAR-CD Advanced CFD Solver² used in the ACLASS model. Because the ideal of linear speedup is not attained in this case, all results plot below the linear-speedup curve – i.e., the parallel solver experiences a sublinear speedup with processor count. Results for the two environments are comparable up to 64 CPUs. However, beyond 64 CPUs, MPICH via Gigabit Ethernet degrades dramatically, due to Gigabit Ethernet’s high latency and inherent limitations found in the Ethernet protocol. Scali MPI Connect results with SilverStorm Infiniband more closely mimic the ideal of linear speedup. The combination of Scali MPI Connect and SilverStorm Infiniband offers the most aggressive ability to mimic the idealized linear speedup.

1 Speedup is related to Amdahl’s Law.

2 The solver in the CD-adapco STAR-CD software is a parallel one.

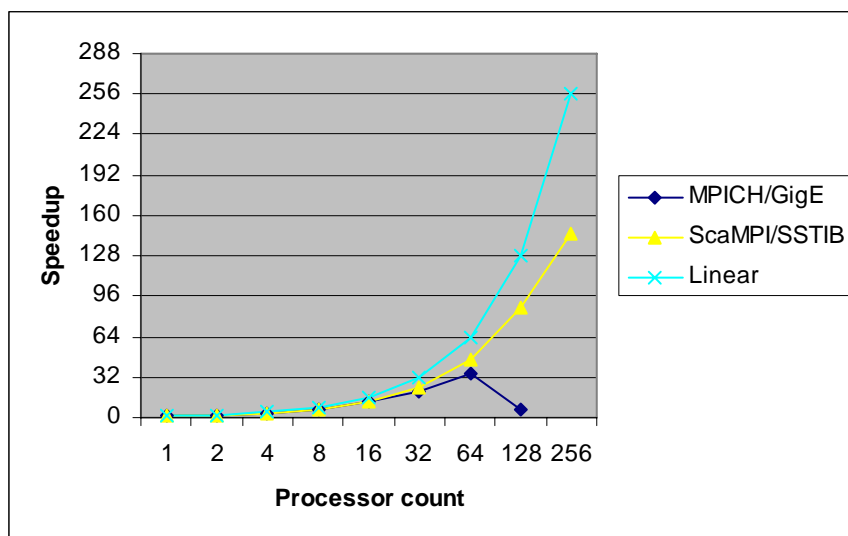


Figure 2 Speedup of the Advanced CFD Solver for the ACLASS dataset as input to the CD-adapco Star-CD CFD model as a function of processor count. Results are presented relative to the ideal of linear speedup.

As the name implies, efficiency calculations quantify usage of the processors employed in a parallel calculation. Specifically, the speedup divided by the processor count quantifies the efficiency. Figure 3 illustrates the efficiencies obtained by the Advanced CFD Solver in the case of the ACLASS dataset as input to the CD-adapco STAR-CD CFD model. In the ideal case, a 100% efficiency is obtained regardless of processor count. Results for the two environments tested are comparable up to processor counts of 16. The small differences evident by 32 CPUs are magnified considerably at larger processor counts – with MPICH via Gigabit Ethernet degrading dramatically by 128 CPUs. As communication increases, the need for a low latency interconnect, provided by SilverStorm Infiniband interconnect, becomes apparent. At 64 processors, the Gigabit Ethernet network becoming saturated with packets and race conditions, causing processors to wait for their required communication before proceeding. At almost 60% efficiency, the combination of Scali MPI Connect and SilverStorm Infiniband clearly offers the best results.

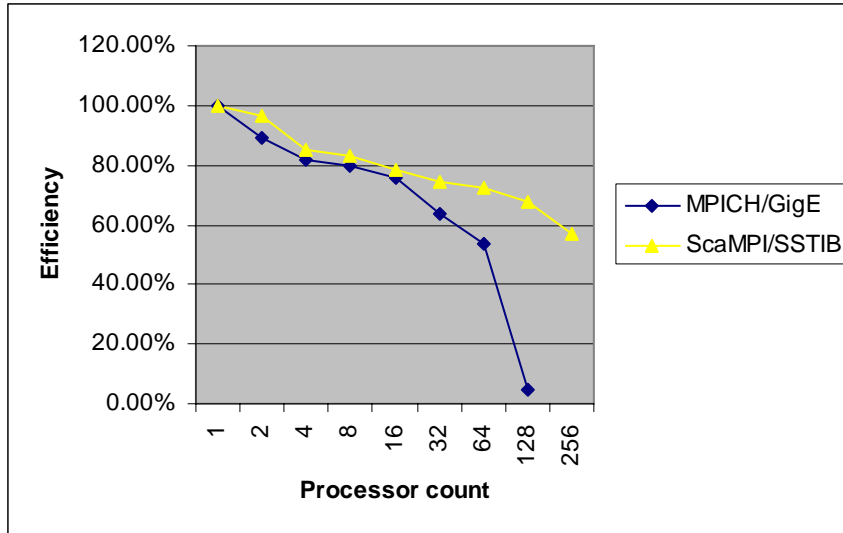


Figure 3 Efficiency of the Advanced CFD Solver for the ACLASS dataset as input to the CD-adapco STAR-CD CFD model as a function of processor count. Results are presented relative to the ideal of 100% efficiency regardless of processor count.

Engine Dataset

The Engine dataset is the second input for the CD-adapco STAR-CD CFD model used to demonstrate the performance characteristics of the benchmarking configuration. This dataset simulates the engine cooling in an automobile engine block and consists of 156,739 cells in a hexahedral mesh. Figure 4 illustrates performance as a function of processor count. Scali MPI Connect results with SilverStorm Infiniband are again normalized relative to MPICH via Gigabit Ethernet. When Scali MPI Connect makes use of the SilverStorm Infiniband interconnect, the performance relative to MPICH exceeds 340% by 32 CPUs. In the case of the Engine dataset, the performance gains for the Scali MPI Connect / SilverStorm Infiniband combination are more than two-and-a-half times those achieved with the ACLASS dataset.

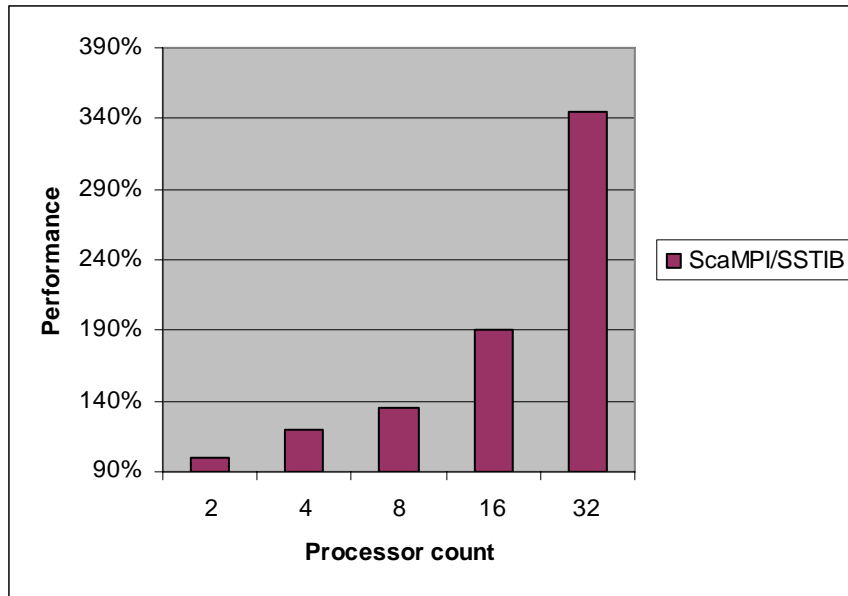


Figure 4 Performance versus processor count for the Engine dataset as input to the CD-adapco STAR-CD CFD model. Percentage-based performance values are normalized relative to the results for MPICH via Gigabit Ethernet.

Speedup and efficiency results for the Engine dataset as input to the Advanced CFD Solver of the CD-adapco STAR-CD CFD model are presented in Figure 5 and Figure 6, respectively. After 4 CPUs, the superiority of the Scali MPI Connect / SilverStorm Infiniband based solution starts to appear in speedup, while the MPICH / Gigabit Ethernet combination rapidly degrades (Figure 5). The Scali MPI Connect / SilverStorm Infiniband results are nothing short of striking – near linear speedup irrespective of processor count. Results for efficiency (Figure 6) corroborate the findings for speedup. The MPICH / Gigabit Ethernet combination degrades rapidly with processor count while the Scali MPI Connect / SilverStorm Infiniband solution exhibits efficiencies consistently in excess of 100% for processor counts of 2 to 16 CPUs, and in excess of 95% for 32 CPUs. Stated differently, super-linear speedups and efficiencies in excess of 100%, together suggest that MPICH / Gigabit Ethernet underestimates what is achievable in cluster configurations. Engine-dataset speedups and efficiencies significantly outperform analogous results obtained for the ACLASS dataset.

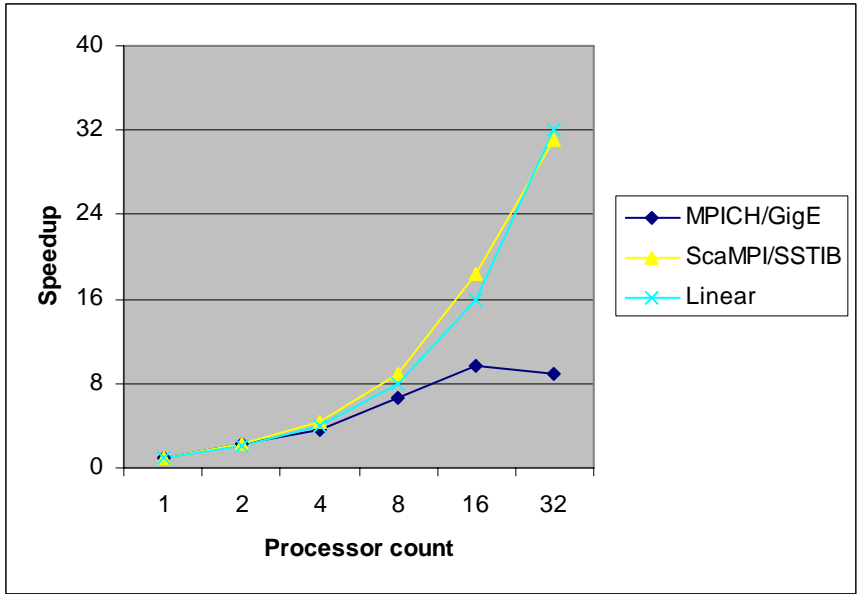


Figure 5 Speedup of the Advanced CFD Solver for the Engine dataset as input to the CD-adapco STAR-CD CFD model as a function of processor count. Results are presented relative to the ideal of linear speedup.

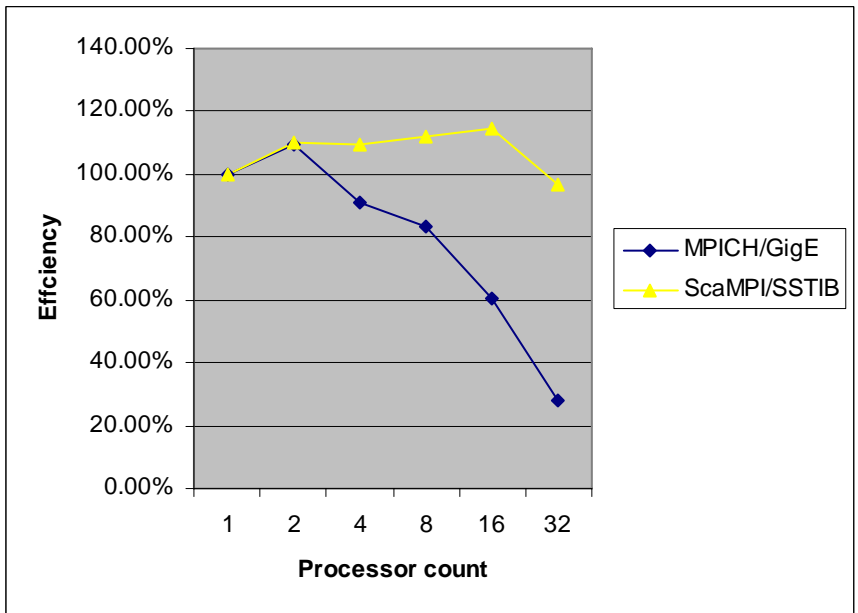


Figure 6 Efficiency of the Advanced CFD Solver for the Engine dataset as input to the CD-adapco SAR-CD CFD model as a function of processor count. Results are presented relative to the ideal of 100% efficiency regardless of processor count.

Summary

It is firmly established that clustering provides a proven approach for production HPC. Given that the concept is proven, attention shifts to innovating within the framework to achieve state-of-the-art performance. An AMD Opteron based platform running Red Hat Enterprise Linux 3.0 provided the foundation for the benchmarking configuration. MPICH via Gigabit Ethernet provided the MPI implementation / interconnect platform against which Scali MPI Connect and SilverStorm Infiniband were measured via two, independent datasets used as input for the CD-adapco STAR-CD CFD model. Scali MPI Connect in tandem with SilverStorm Infiniband provided the best-overall performance – in some cases, more than doubling outcomes achieved with the MPICH / Gigabit Ethernet combination. Speedup and efficiency can be so compelling that super-linear and greater-than-100% results, respectively, are possible. It is also clear that the Scali MPI Connect / SilverStorm InfiniBand combination is even more attractive as processor count increases – as its performance gain over MPICH / Gigabit Ethernet accelerates with processor count.

Resources

- CD-adapco STAR-CD – <http://www.cd-adapco.com>
- DAPL – <http://sourceforge.net/projects/dapl>
- High-Performance Linpack – <http://www.netlib.org/benchmark/hpl>
- HP Linux Solutions – <http://www.hp.com/linux>
- SilverStorm Infiniband – <http://www.SilverStorm.com>
- MPI – <http://www-unix.mcs.anl.gov/mpi>
- MPICH – <http://www-unix.mcs.anl.gov/mpi/mpich>
- OpenIB – <http://www.openib.org>
- RDMA – <http://www.rdmaconsortium.org>
- Scali MPI Connect – <http://www.scali.com>
- The AMD Developer Center - <http://devcenter.amd.com>
- The Top 500 Supercomputer Sites - <http://www.top500.org>