HPC ADVISORY COUNCIL

STAR-CCM+ Performance Benchmark and Profiling

July 2012

CD-adapco DELL intel Mellanox TECHNOLOGIES

Note

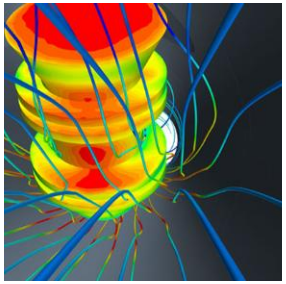
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: CD-adapco, Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

- **The following was done to provide best practices**
  - STAR-CCM+ performance overview
  - Understanding STAR-CCM+ communication patterns
  - Ways to increase STAR-CCM+ productivity
  - MPI libraries comparisons

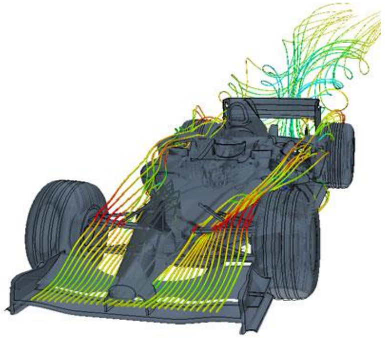
- **For more info please refer to**
  - http://www.cd-adapco.com
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com

STAR-CCM+

- **STAR-CCM+**
  - An engineering process-oriented CFD tool
  - Client-server architecture, object-oriented programming
  - Delivers the entire CFD process in a single integrated software environment



- **Developed by CD-adapco**



Objectives

- **The presented research was done to provide best practices**
  - CD-adapco performance benchmarking
  - Interconnect performance comparisons
  - Ways to increase CD-adapco productivity
  - Power-efficient simulations

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 16-node (256-core) “Jupiter” cluster**
  - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **SwitchX SX6036 InfiniBand switch**
- **MPI: Platform MPI 8.2**
- **Application: STAR-CCM+ version 7.02.008**
- **Benchmarks:**
  - Lemans_Poly_17M (Epsilon Euskadi Le Mans car external aerodynamics)
  - Civil_Trim_20M (Civil Airliner External Aerodynamics)

About Intel® Cluster Ready

- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
  - Focus on your work productivity, spend less management time on the cluster
- **Select Intel Cluster Ready**
  - Where the cluster is delivered ready to run
  - Hardware and software are integrated and configured together
  - Applications are registered, validating execution on the Intel Cluster Ready architecture
  - Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health

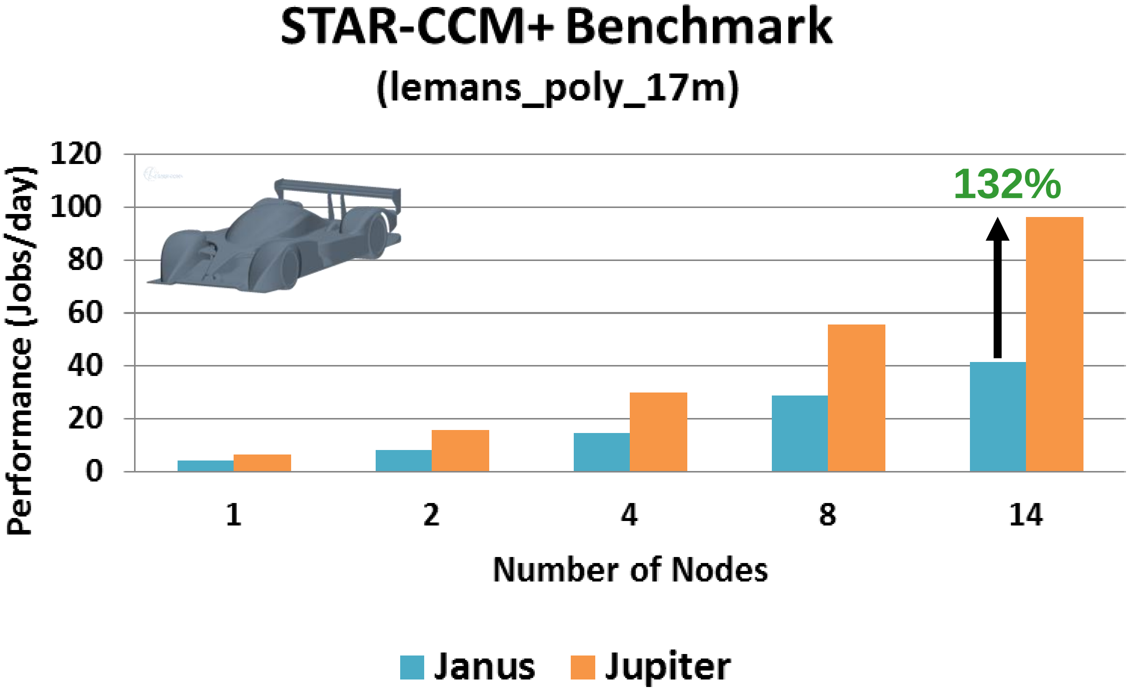
PowerEdge R720xd

Massive flexibility for data intensive operations

- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GigE to IB
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans
- **Benefits**
  - Designed for performance workloads
    - from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    - High performance scale-out compute and low cost dense storage in one package
- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5” - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch

STAR-CCM+ Performance – Processors

- **Intel E5-2600 Series (Sandy Bridge) outperforms prior generations**
  - Up to 132% higher performance than Intel Xeon X5670 (Westmere) at 14-node
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk

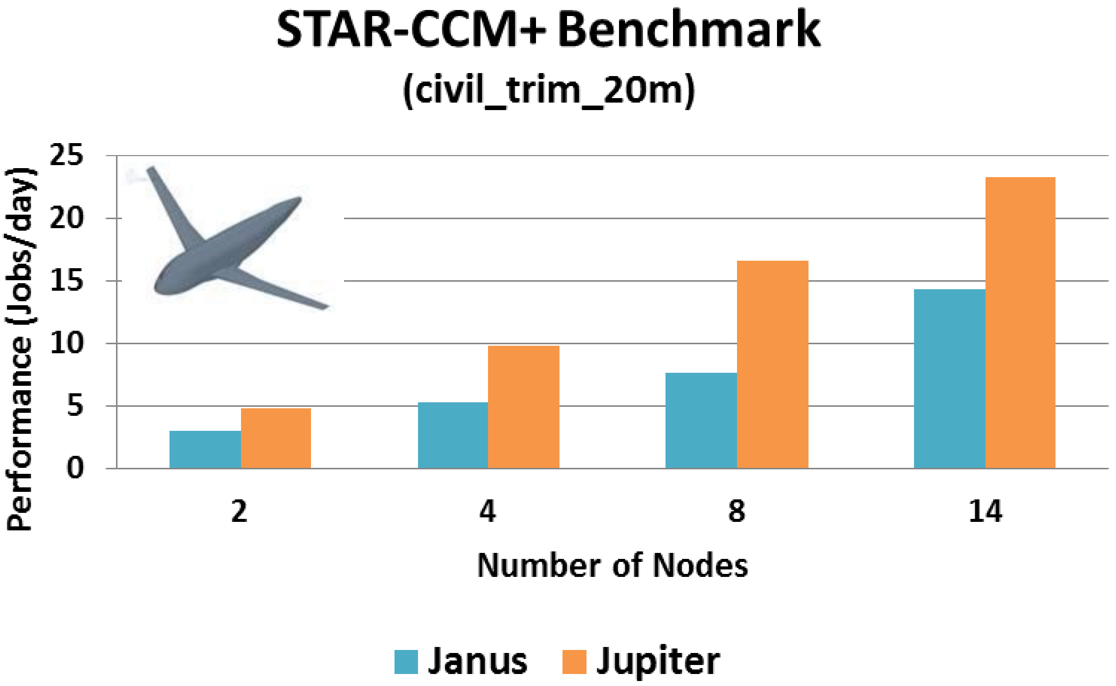


Higher is better

InfiniBand FDR

STAR-CCM+ Performance – Processors

- **Intel E5-2600 Series (Sandy Bridge) outperforms prior generations**
  - Average of 80% gain in performance compared to a X5670 (Westmere) cluster
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk

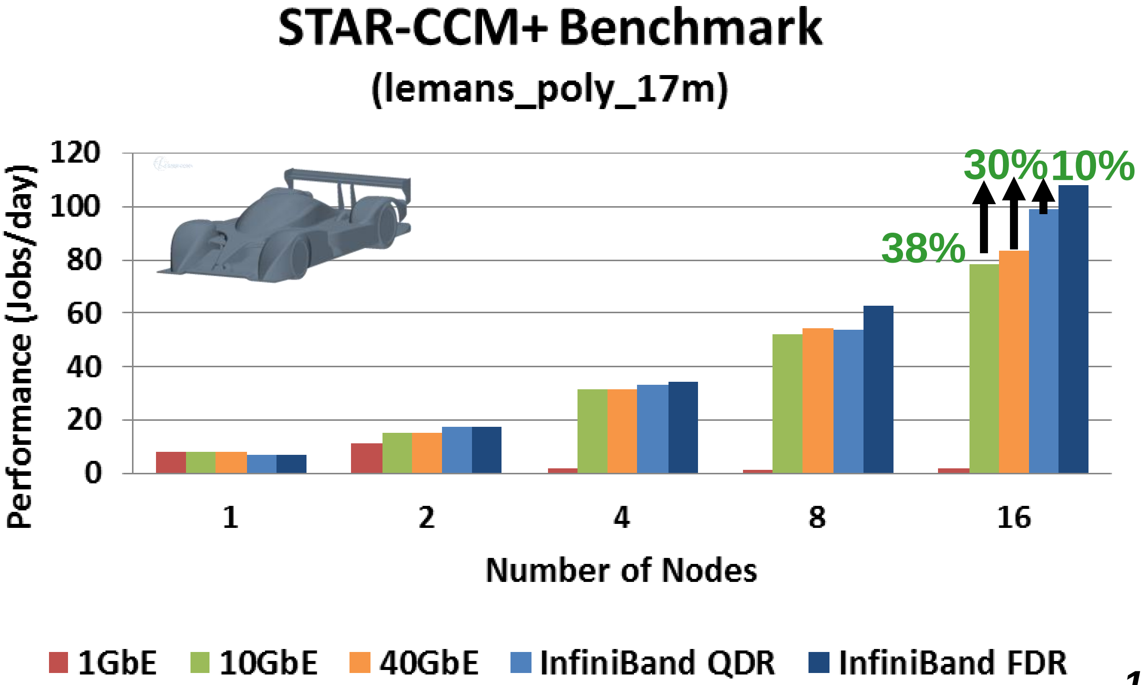


Higher is better

InfiniBand FDR

STAR-CCM+ Performance – Network

- **InfiniBand FDR delivers the best network scalability performance**
  - Provides up to 10% better performance than InfiniBand QDR
  - Provides up to 38% better performance than 40GbE
  - Provides up to 30% better performance than 10GbE
  - 1GbE is seen having scalability issues beyond 2 nodes
- **CPU binding optimization flag used in all cases shown**
  - MPIRUN_OPTIONS="-cpu_bind "

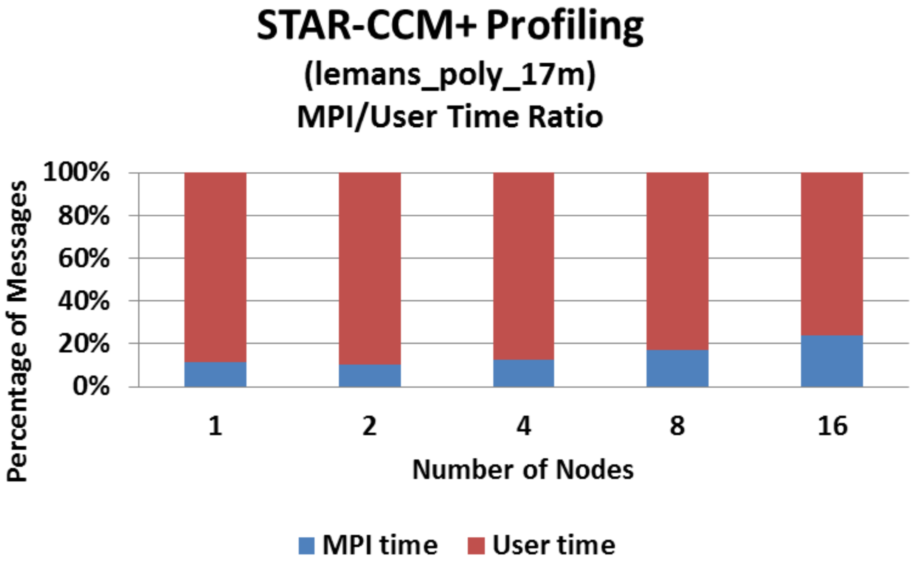


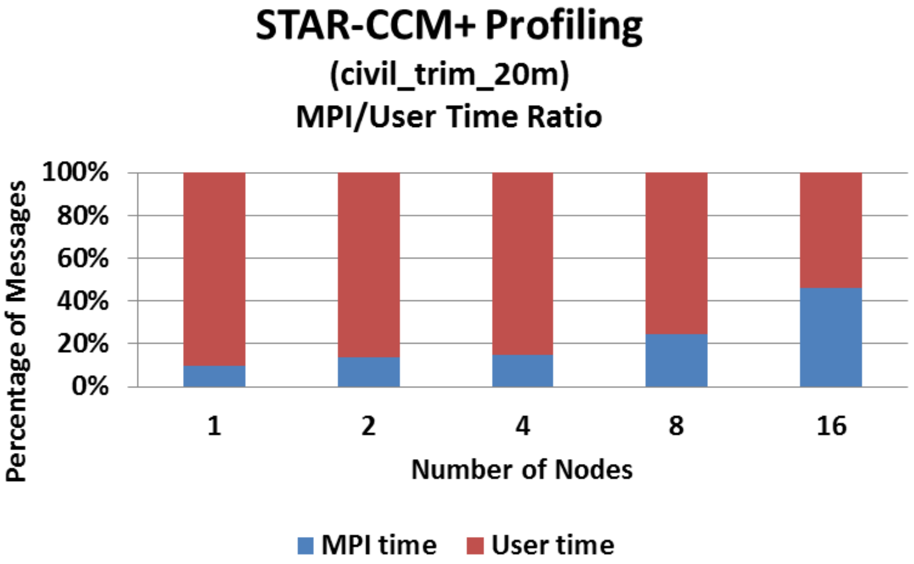
Higher is better

16 Processes/Node

STAR-CCM+ Profiling – # of MPI Calls

- **The overall runtime reduces as more nodes take part of the MPI job**
  - Using more compute nodes to reduce the runtime by spreading out the workload
- **Higher percentage time is spent on CPU than on communications**
  - Civil_trim_20m has communications grow at a faster pace than leman_poly





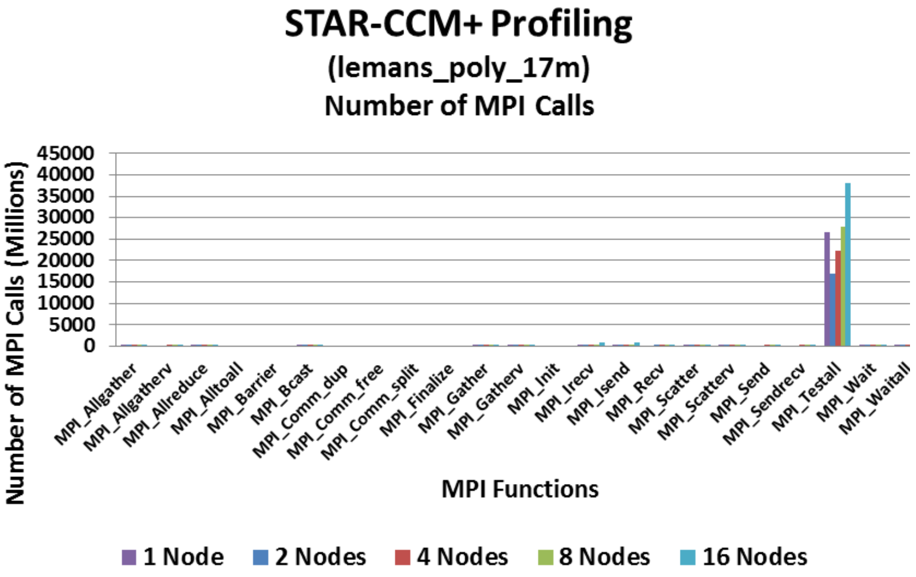
Higher is better

16 Processes/Node

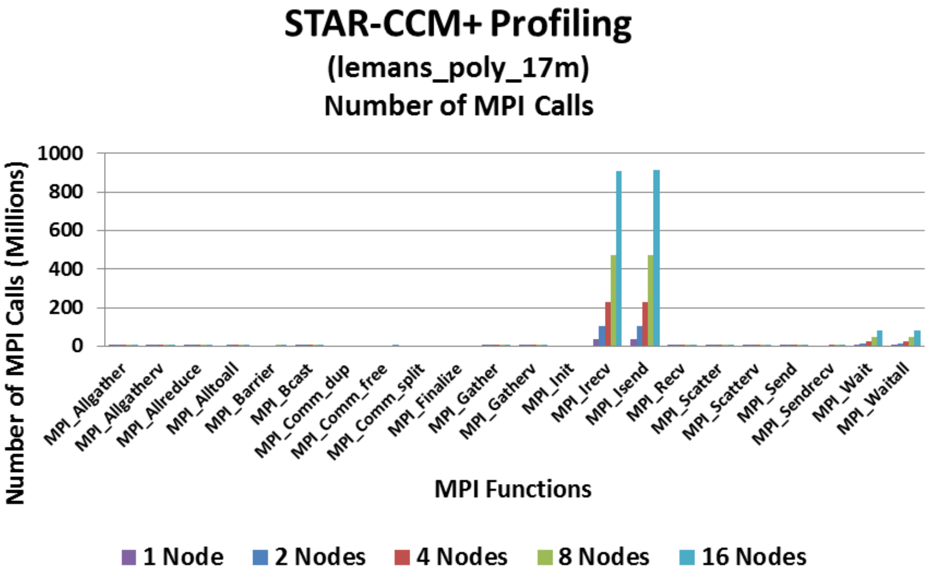
STAR-CCM+ Profiling – # of MPI Calls

- **The most used MPI calls is MPI_Testall**
  – Aside from MPI_Testall, MPI_Isend and Irecv are the next most used calls

Showing All MPI calls



Excluding MPI_Testall

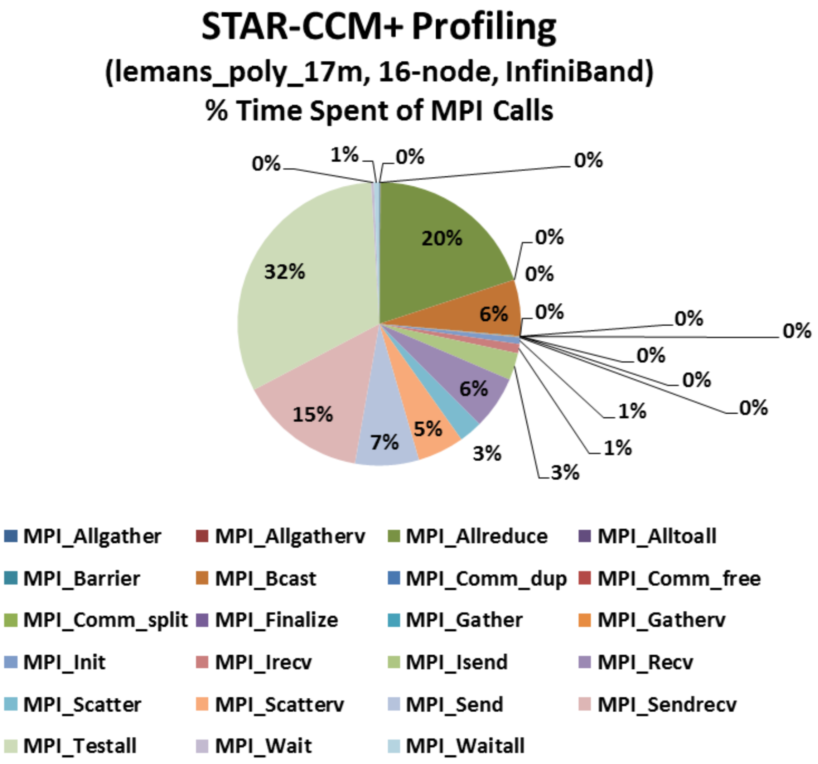


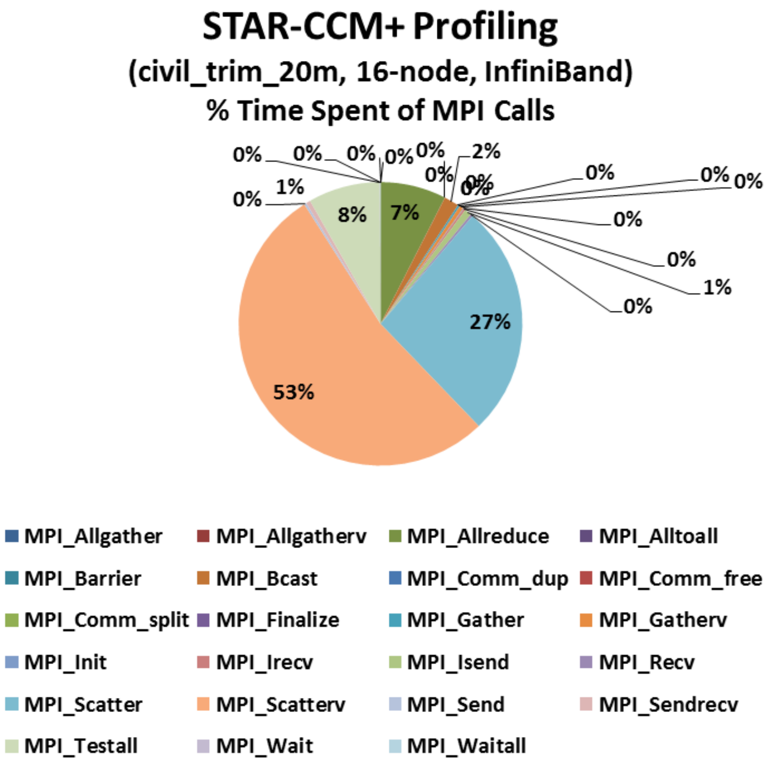
Higher is better

16 Processes/Node

STAR-CCM+ Profiling – MPI Comm. Time

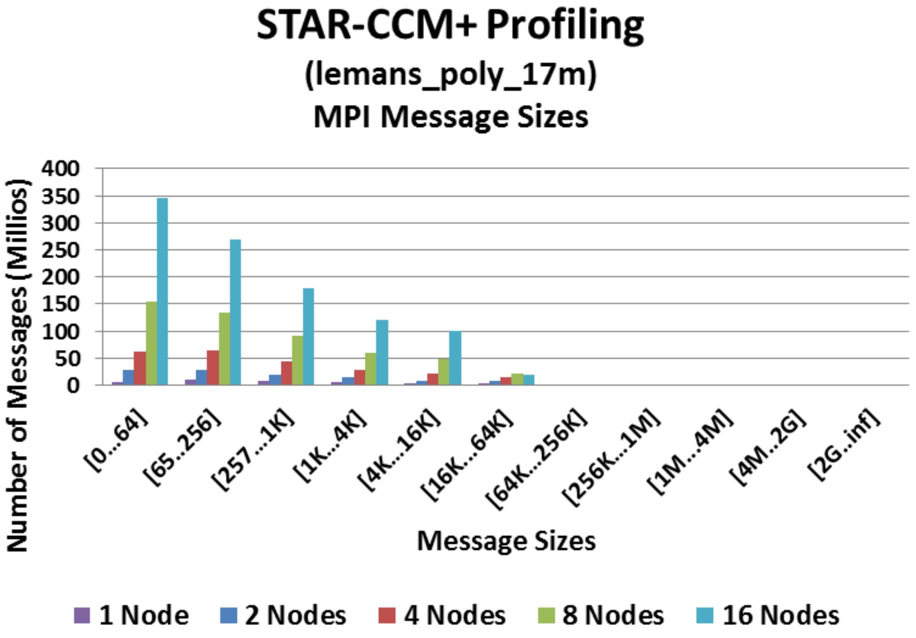
- **Majority of MPI communication time is spent on MPI_Testall**
  - Lemans_poly_17m: MPI_Testall(32%), MPI_Reduce(20%), MPI_Sendrecv(15%)
  - Civil_trim_20m: MPI_Scatterv(53%), MPI_Scatter(27%), MPI_Testall(8%)

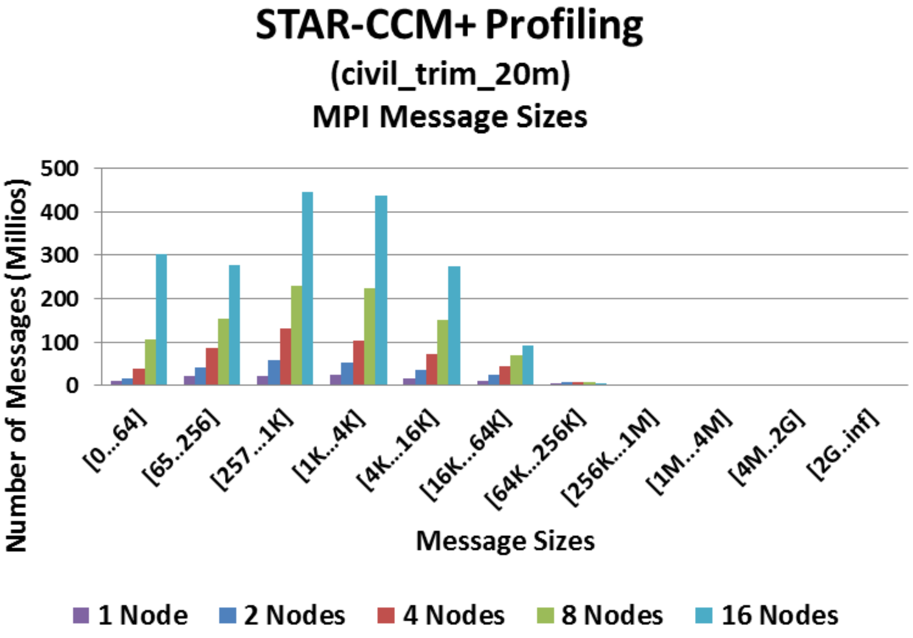




STAR-CCM+ Profiling – Message Sizes

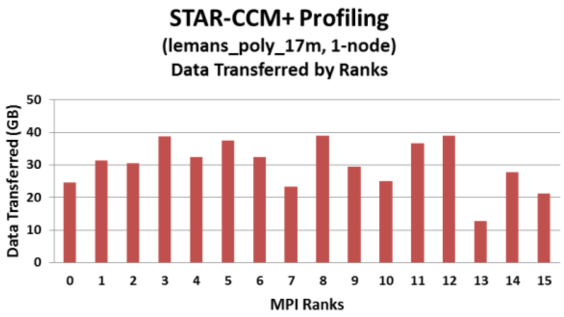
- **Majority of messages are small messages**
  - Lemans_poly_17m: Messages below 64B are mostly used
  - Civil_trim_20m: messages between 256B to 4KB are mostly used.
- **Number of messages increases with the number of nodes**

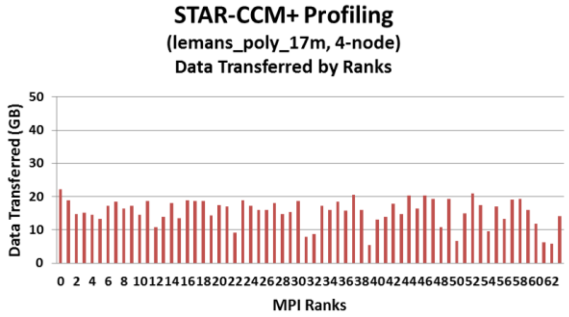


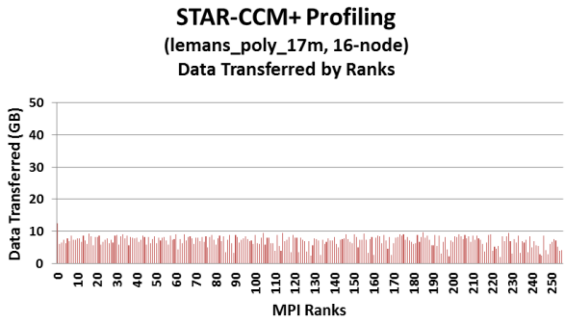


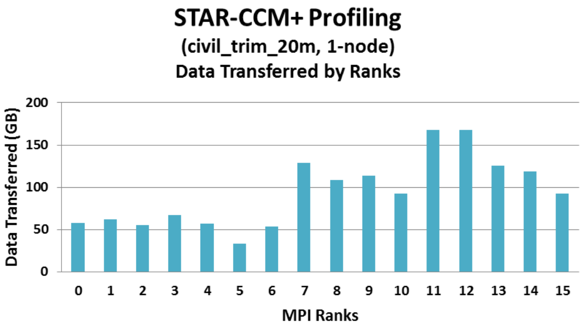
STAR-CCM+ Profiling – MPI Data Transfer

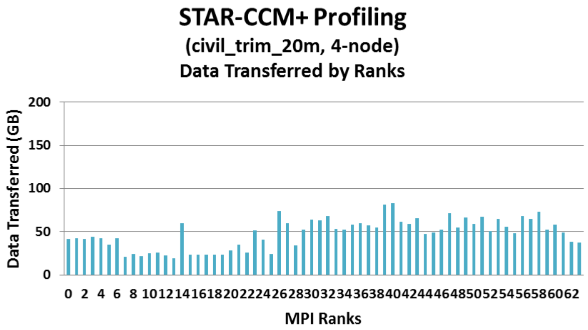
- **As the cluster grows, less data transfers between MPI processes**
  - Lemans_poly_17m: Drops from ~30GB per rank at 1-node vs ~9GB at 16-node
  - Civil_trim_20m: Drops from ~150GB per rank at 1-node to ~20GB at 16-node

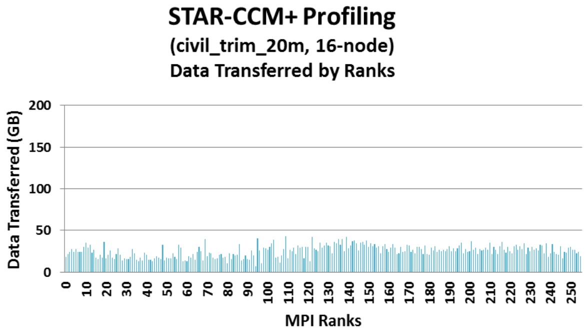

STAR-CCM+ Profiling
(lemans_poly_17m, 1-node)
Data Transferred by Ranks


STAR-CCM+ Profiling
(lemans_poly_17m, 4-node)
Data Transferred by Ranks


STAR-CCM+ Profiling
(lemans_poly_17m, 16-node)
Data Transferred by Ranks

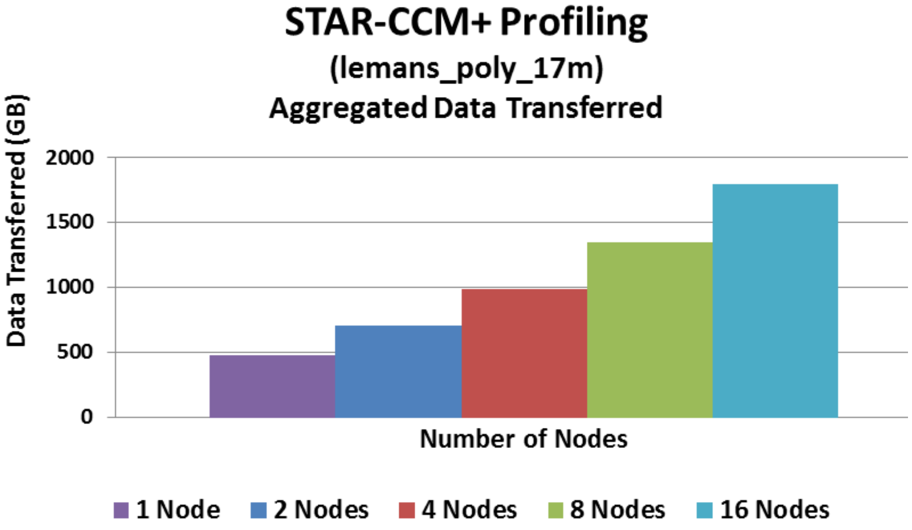

STAR-CCM+ Profiling
(civil_trim_20m, 1-node)
Data Transferred by Ranks

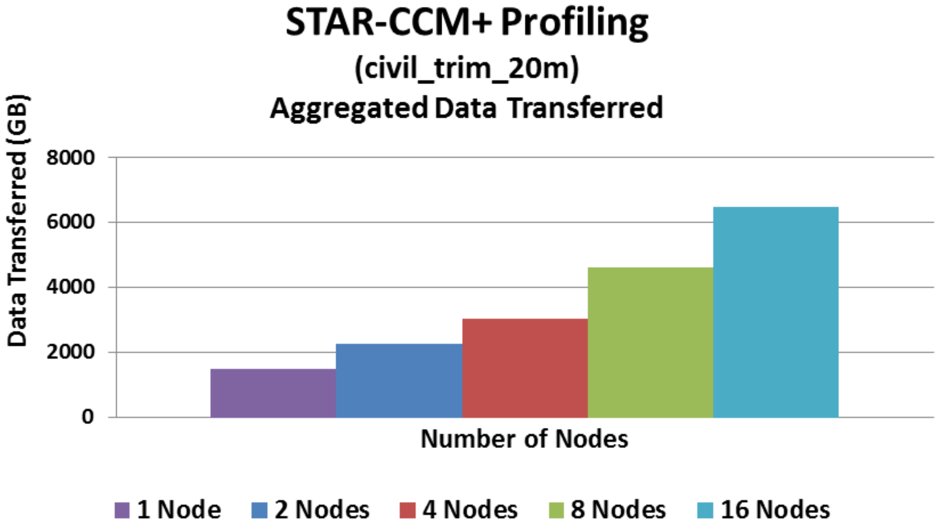

STAR-CCM+ Profiling
(civil_trim_20m, 4-node)
Data Transferred by Ranks


STAR-CCM+ Profiling
(civil_trim_20m, 16-node)
Data Transferred by Ranks

STAR-CCM+ Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Very large data transfer takes place in STAR-CCM+**
  - High network throughput is required for delivering the network bandwidth
  - Lemans_poly_17m: 1.8TB of data transfer takes place between the MPI processes
  - Civil_trim_20m: 6.1TB of data transfer takes place between the MPI processes





InfiniBand FDR

STAR-CCM+ – Summary

- **Performance**
  - Intel Xeon E5-2600 series and InfiniBand FDR enable STAR-CCM+ to scale
  - The E5-2680 cluster outperforms X5670 cluster by 132% for leman_poly_17m at 16-node
  - The E5-2680 cluster outperforms X5670 cluster by 80% on average for civil_trim_20m
- **Network**
  - InfiniBand FDR allows STAR-CCM+ to run at the highest network throughput at 56Gbps
  - InfiniBand FDR provides up to 38% of performance gain over 10GbE
- **Profiling**
  - High network throughput is required for delivering the network bandwidth
  - Majority of MPI time is spent on MPI_Wait for pending non-blocking sends and receives

Thank You

HPC Advisory Council

All trademarks are property of their respective owners. All information is provided “As-Is” without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein