HPC ADVISORY COUNCIL

STAR-CCM+
Performance Benchmark and Profiling

July 2014

CD-adapco DELL intel Mellanox TECHNOLOGIES

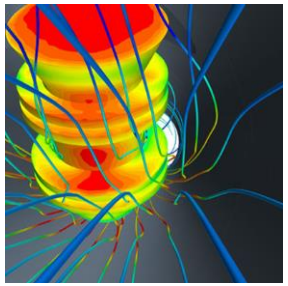
Note

- **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

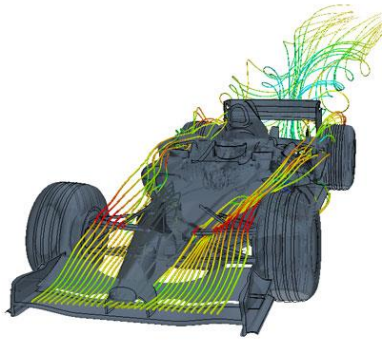
- **The following was done to provide best practices**
  - STAR-CCM+ performance overview
  - Understanding STAR-CCM+ communication patterns
  - Ways to increase STAR-CCM+ productivity
  - MPI libraries comparisons

- **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+**
  - 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 32-node (640-core) "Jupiter" cluster**
  - Dual-Socket Hexa-Core Intel E5-2680 V2 @ 2.80 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 2.1-1.0.0 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox Connect-IB FDR InfiniBand and ConnectX-3 Ethernet adapters**
- **Mellanox SwitchX 6036 VPI InfiniBand and Ethernet switches**
- **MPI: Mellanox HPC-X v1.0.0, Platform MPI 8.3.0.6, Intel MPI 4.1.3**
- **Application: STAR-CCM+ version 9.02.005 (unless specified otherwise)**
- **Benchmarks:**
  - Lemans_Poly_17M (Epsilon Euskadi Le Mans car 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

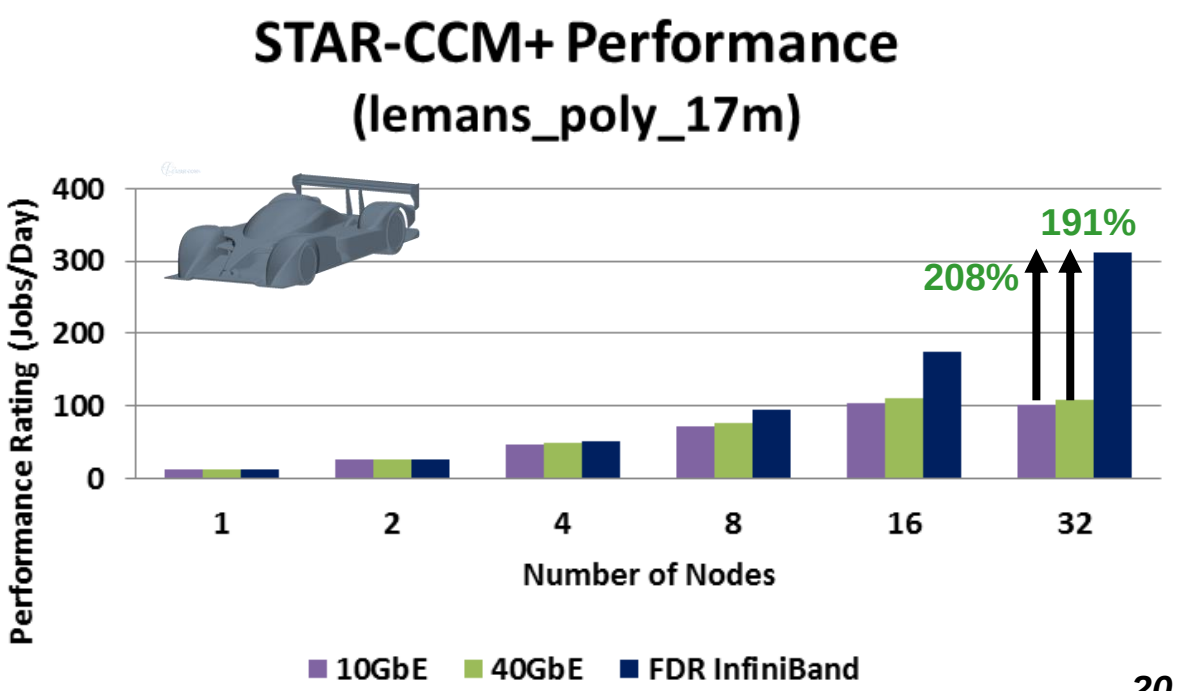
PowerEdge R720xd

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 – Network

- **FDR InfiniBand delivers the best network scalability performance**
  - Provides up to 208% higher performance than 10GbE at 32 nodes
  - Provides up to 191% higher performance than 40GbE at 32 nodes
  - FDR IB scales linearly while 10/40GbE has scalability limitation beyond 16 nodes

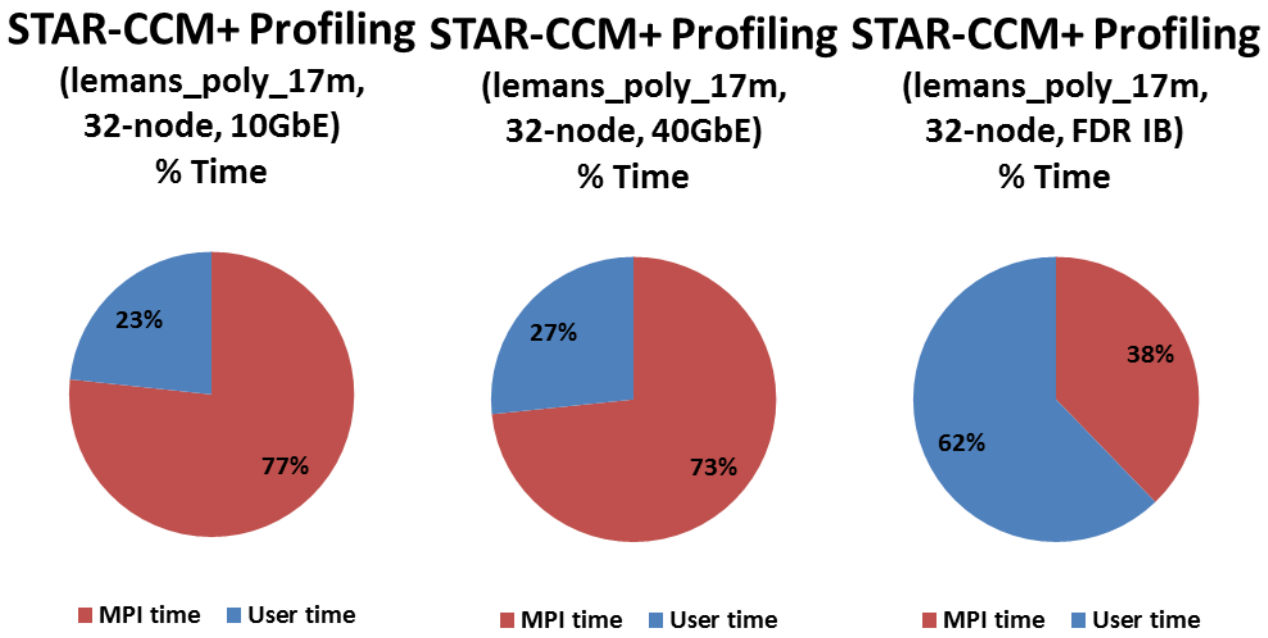


Higher is better

20 Processes/Node

STAR-CCM+ Profiling – Network

- **InfiniBand reduces network overhead; results higher CPU utilization**
  - Reducing MPI communication overhead with efficient network interconnect
  - As less time spent on the network, overall application runtime is improved
- **Ethernet solutions consumes more time in communications**
  - Spent 73%-95% of overall time in network due to congestion in Ethernet
  - While FDR IB spent about 38% of overall runtime

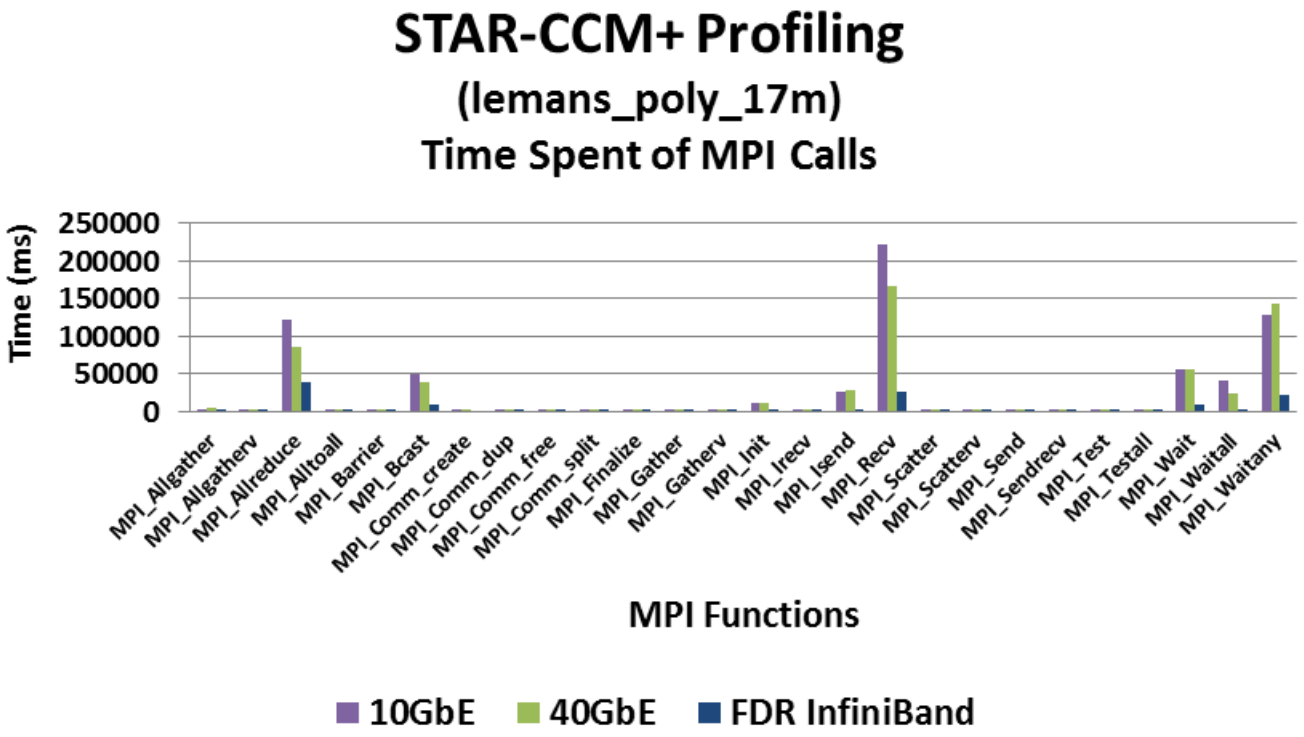


Higher is better

20 Processes/Node

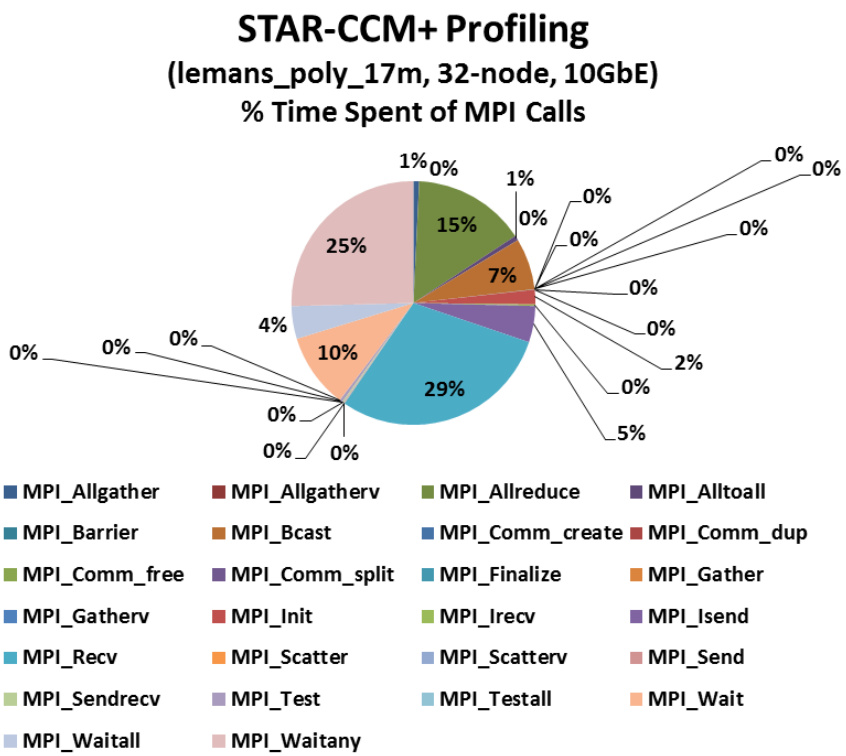
STAR-CCM+ Profiling – MPI Comm. Time

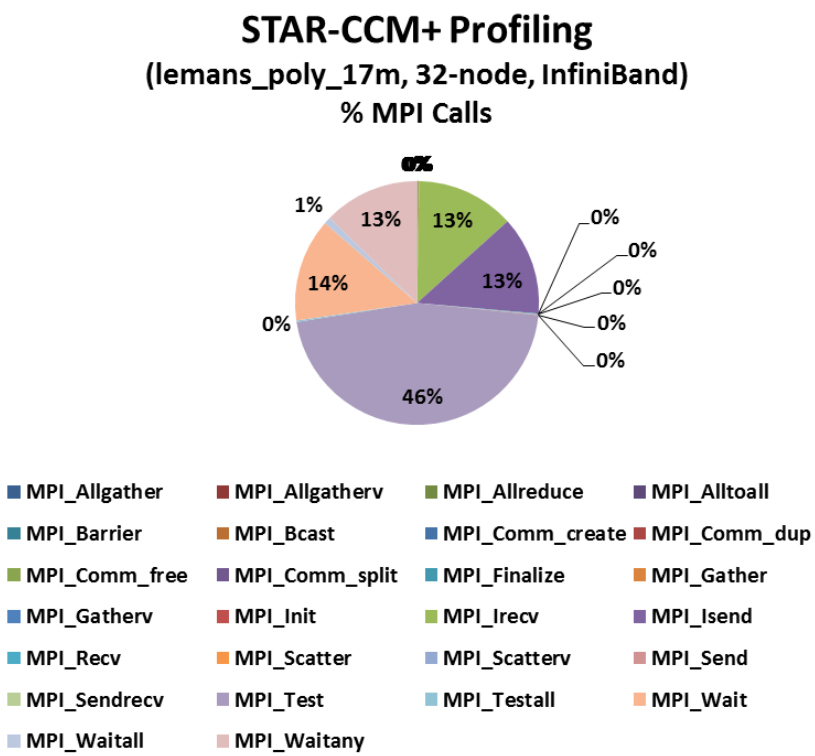
- **Identified MPI overheads by profiling communication time**
  - Dealt with communications in collective, point-to-point and non-blocking operations
  - 10/40GbE vs FDR IB: Spent longer time in Allreduce, Bcast, Recv, Waitany



STAR-CCM+ Profiling – MPI Comm. Time

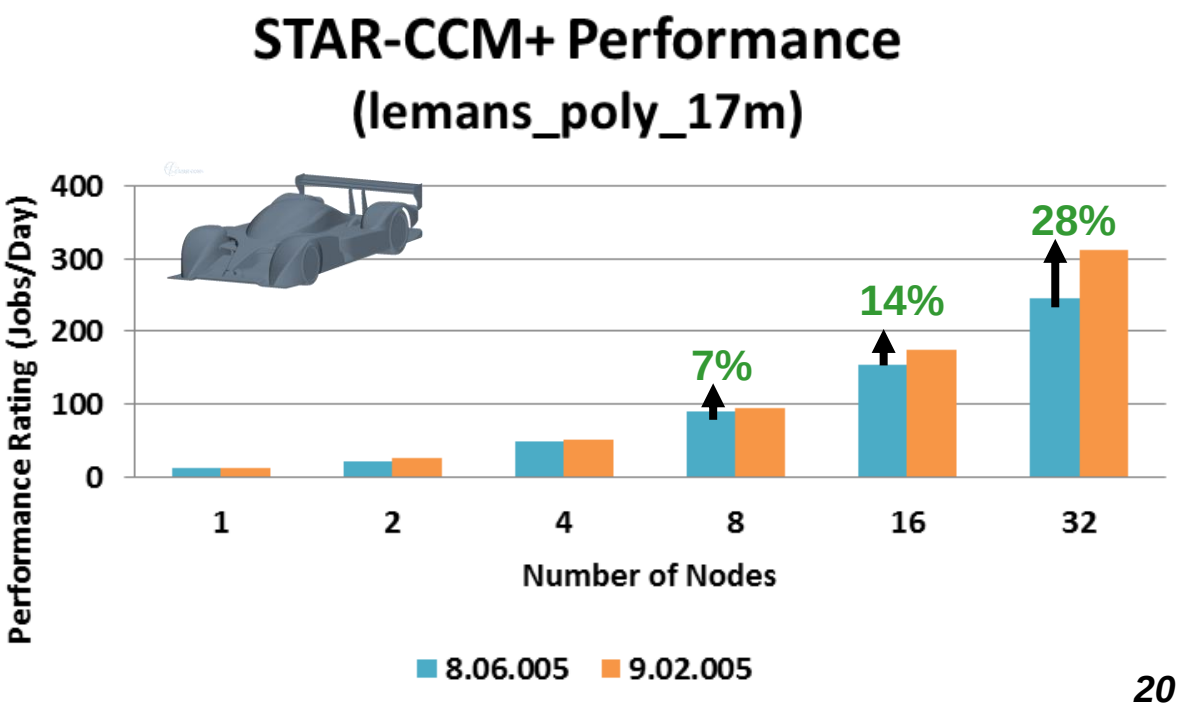
- **Observed MPI time spent by different network hardware**
  - FDR IB: MPI_Allreduce(32%), MPI_Recv(22%), MPI_Waitany(15%), MPI_Bcast(8%)
  - 10GbE: MPI_Recv(29%), MPI_Waitany(25%), MPI_Allreduce(15%), MPI_Wait(10%)





STAR-CCM+ Performance – Software Versions

- **Improvement in latest STAR-CCM+ results in higher performance at scale**
  - v9.02.005 demonstrated a 28% gain compared to the v8.06.005 on 32-node run
  - Slight Change in communication pattern helps to improve the scalability
  - Improvement gap expects to widen at scale
  - See subsequence slides in the MPI profiling to show the differences

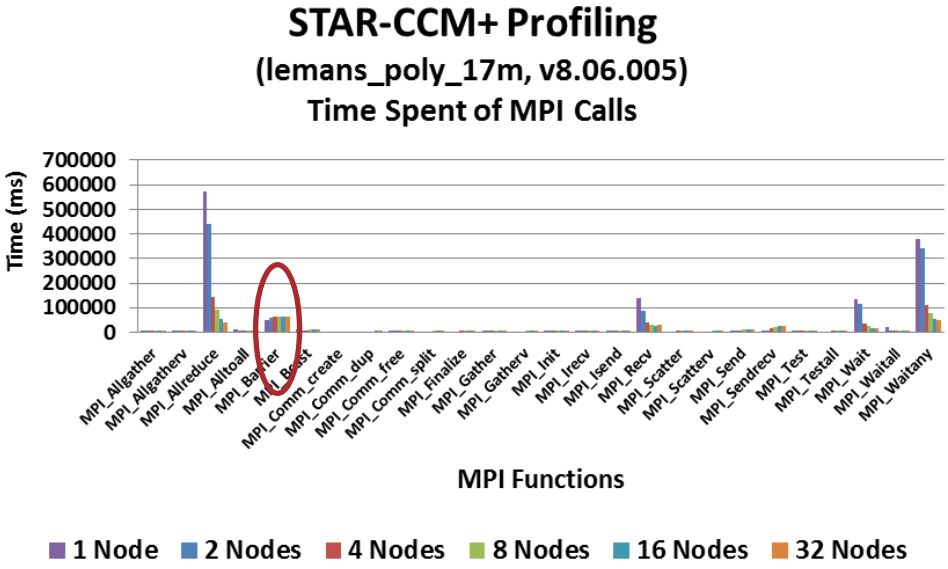


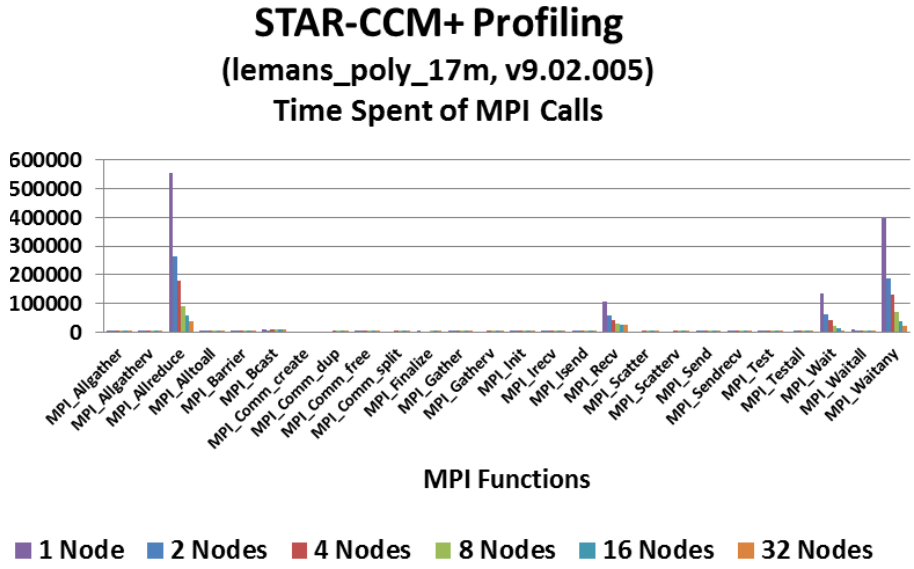
Higher is better

20 Processes/Node

STAR-CCM+ Profiling – MPI Time Spent

- **Communication time has dropped with the latest STAR-CCM+ version**
  - Observed less time spent in MPI, although communication pattern is roughly the same
  - MPI Barrier time is reduced significantly between the 2 releases



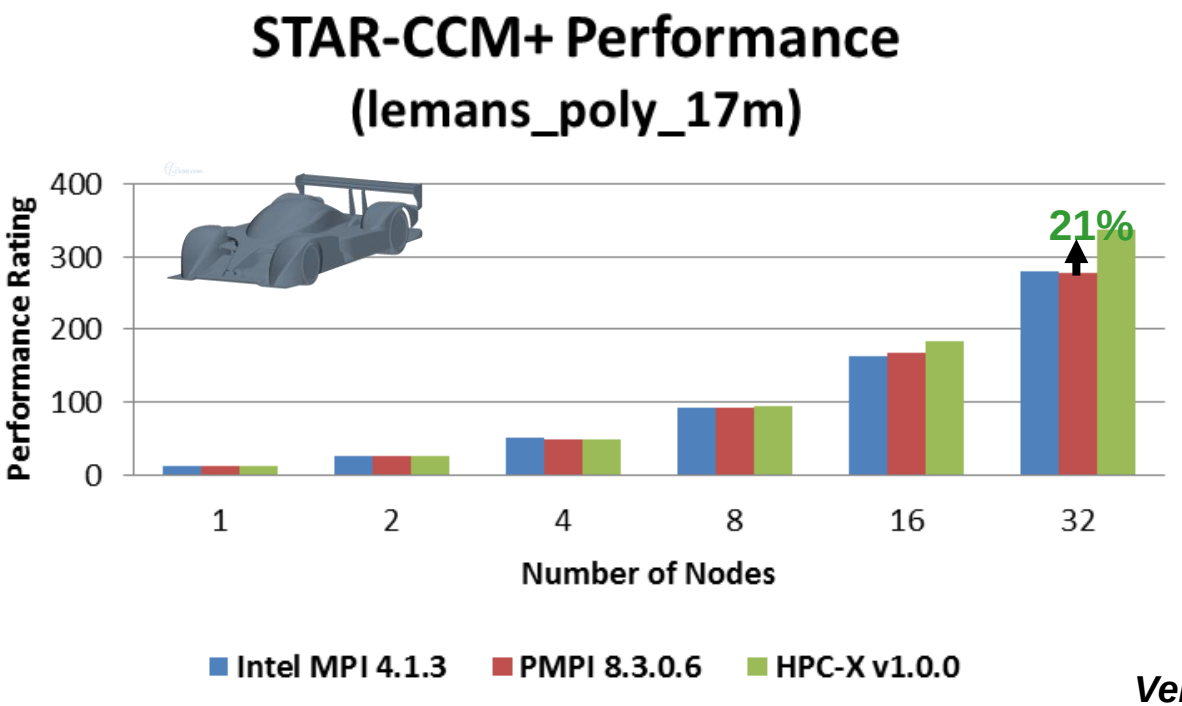


Higher is better

20 Processes/Node

STAR-CCM+ Performance-MPI Implementations

- **STAR-CCM+ has made various MPI implementations available to run**
  - Default MPI implementation used in STAR-CCM+ is Platform MPI
  - MPI implementations started to differentiate beyond 8 nodes
  - Optimization flags have been set already in vendor's startup scripts
  - Support for HPC-X is based on the existing Open MPI support in STAR-CCM+
  - HPC-X provides 21% of higher scalability than the alternatives

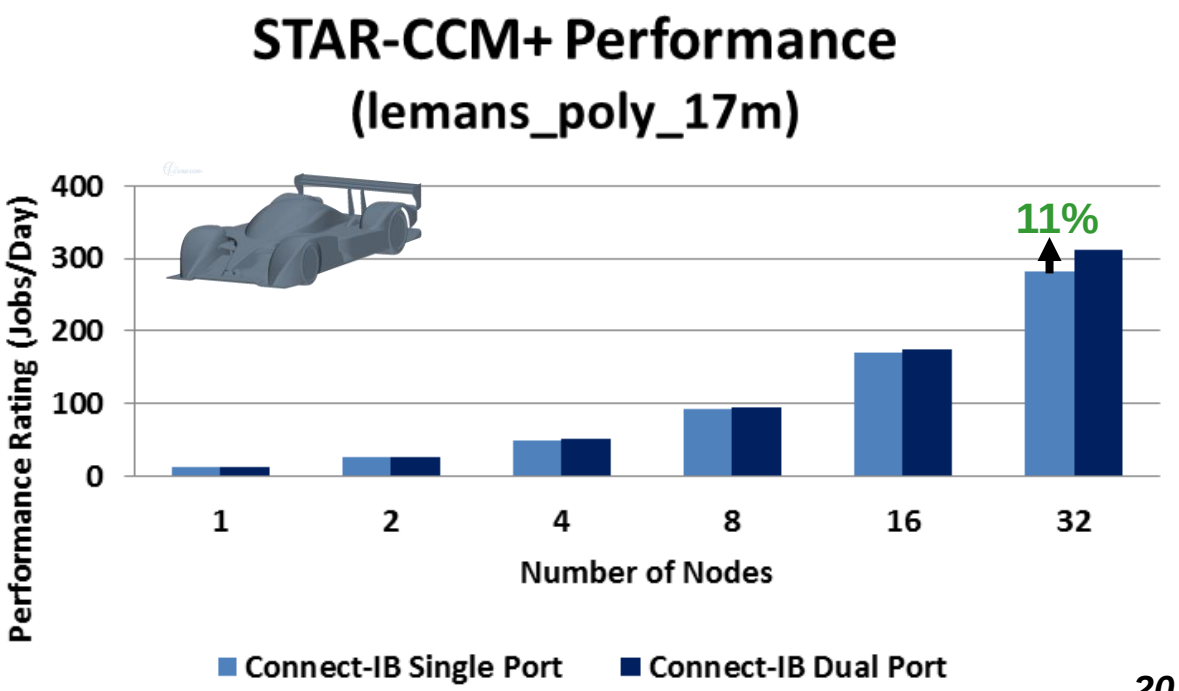


Higher is better

Version 9.02.005

STAR-CCM+ Performance – Single/Dual Port

- **Benefit of deploying dual-port InfiniBand is demonstrated at scale**
  - Running with dual port provides up to 11% higher performance at 32 nodes
  - Connect-IB on PCIe Gen3 x16 slot which can provide additional throughput with 2 links

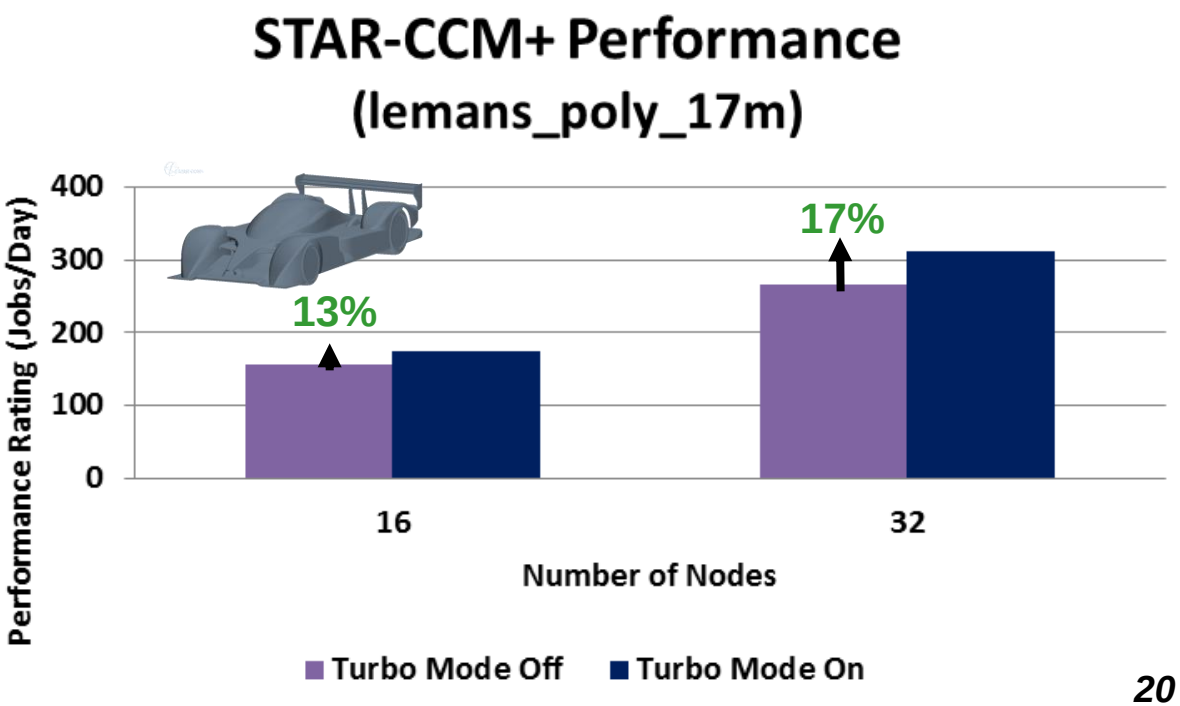


Higher is better

20 Processes/Node

STAR-CCM+ Performance – Turbo Mode

- **Enabling Turbo mode results in higher application performance**
  - Up to 17% of the improvement seen by enabling Turbo mode
  - Higher performance gain seen with higher node count
  - Boosting base frequency; consequently resulted in higher power consumption
- **Using kernel tools called "msr-tools" to adjust Turbo Mode dynamically**
  - Allows dynamically turn off/on Turbo mode in the OS level

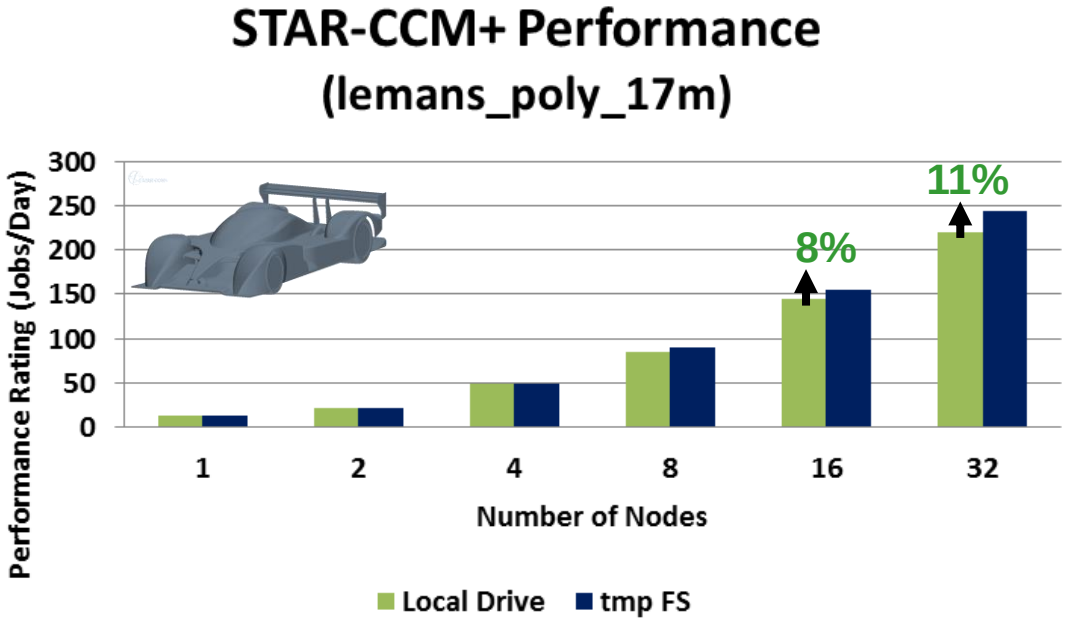


Higher is better

20 Processes/Node

STAR-CCM+ Performance – File IO

- **Advantages for staging data to temporary file system in memory**
  - Data write of ~8GB occurs at the end of the run for the benchmark
  - By staging on local FS, which avoid accessing by all processes (vs NFS)
  - By staging on local tmpfs, even higher performance gain is seen (up to 11% gain)
- **Using temporary storage is not recommended for production environment**
  - While tmpfs reduces outperforms localfs, it is not recommended for production
  - If available, parallel file system is more preferred solution versus local or tmpfs

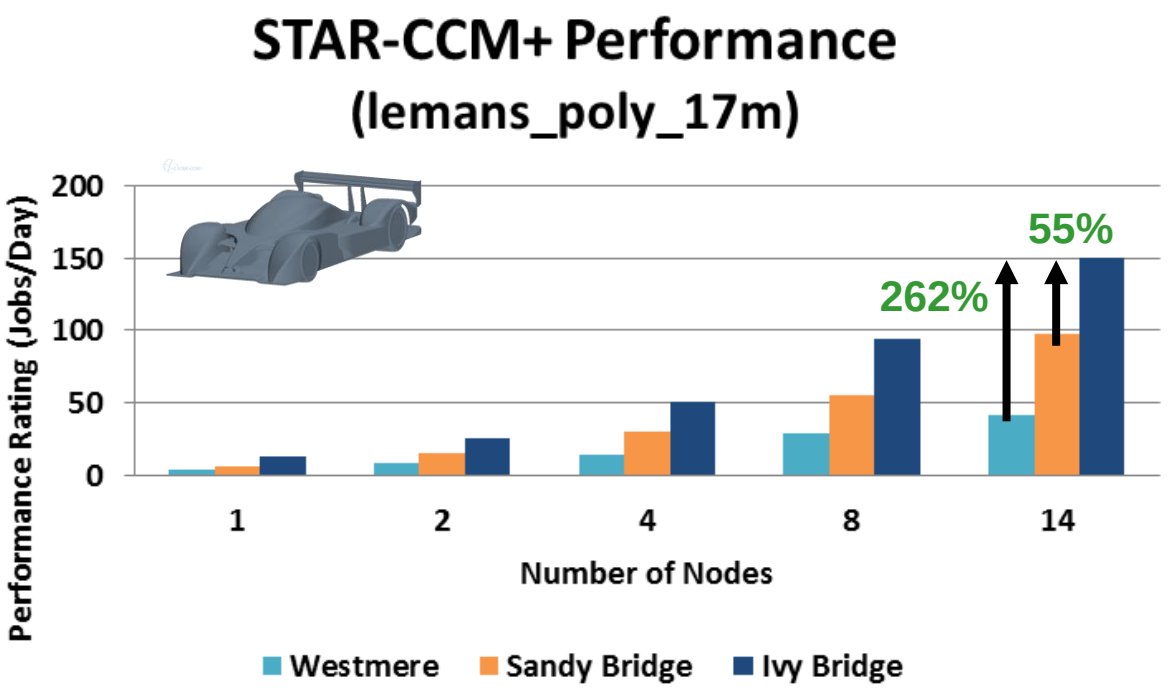


Higher is better

Version 8.06.005

STAR-CCM+ Performance - System Generations

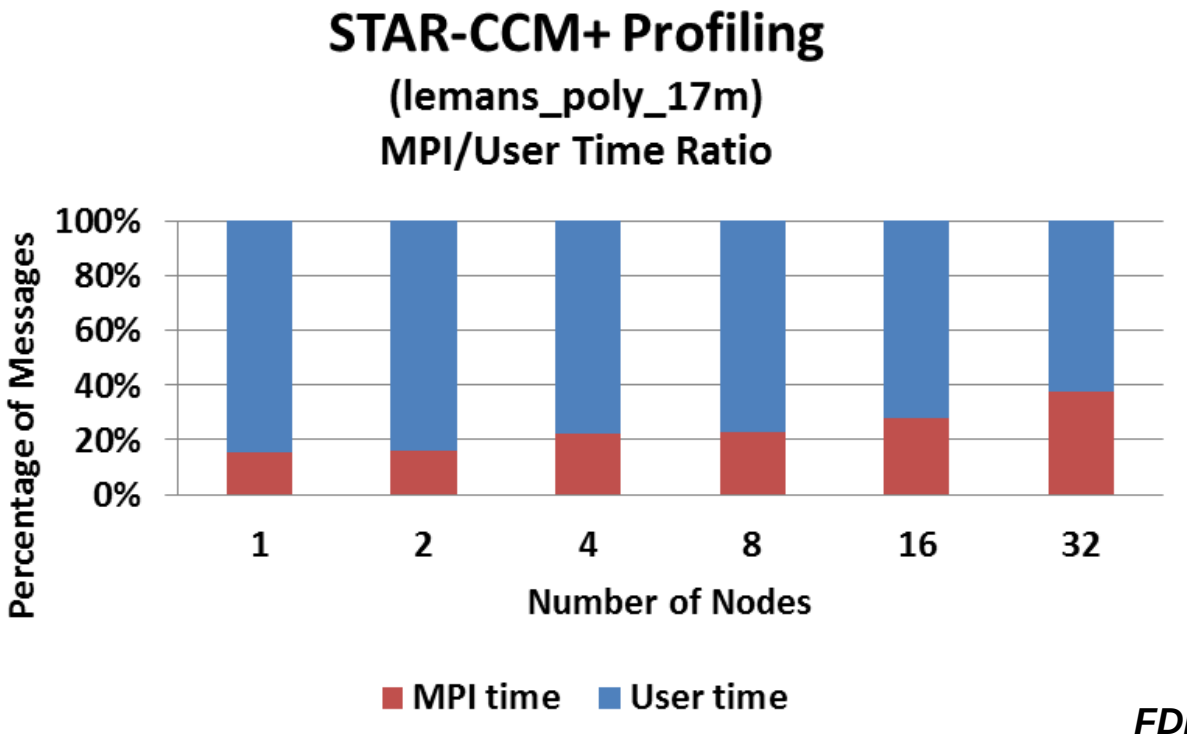
- **New generations of software and hardware provide performance gain**
  - Performance gain demonstrated through variables in HW and SW
  - Latest stack provides ~55% higher performance versus 1 generation behind
  - Latest stack provides ~2.6x higher performance versus 2 generations behind
- **System components used:**
  - WSM: X5670@ 2.93GHz, DDR3-10666, ConnectX-2 QDR IB, 1 disk, v5.04.006
  - SNB: E5-2680@2.7GHz, DDR3-12800, ConnectX-3 FDR IB, 24 disks, v7.02.008
  - IVB: E5-2680v2@2.8GHz, DDR3-12800, Connect-IB FDR IB, 24 disks, v9.02.005



Higher is better

STAR-CCM+ Profiling – User/MPI Time Ratio

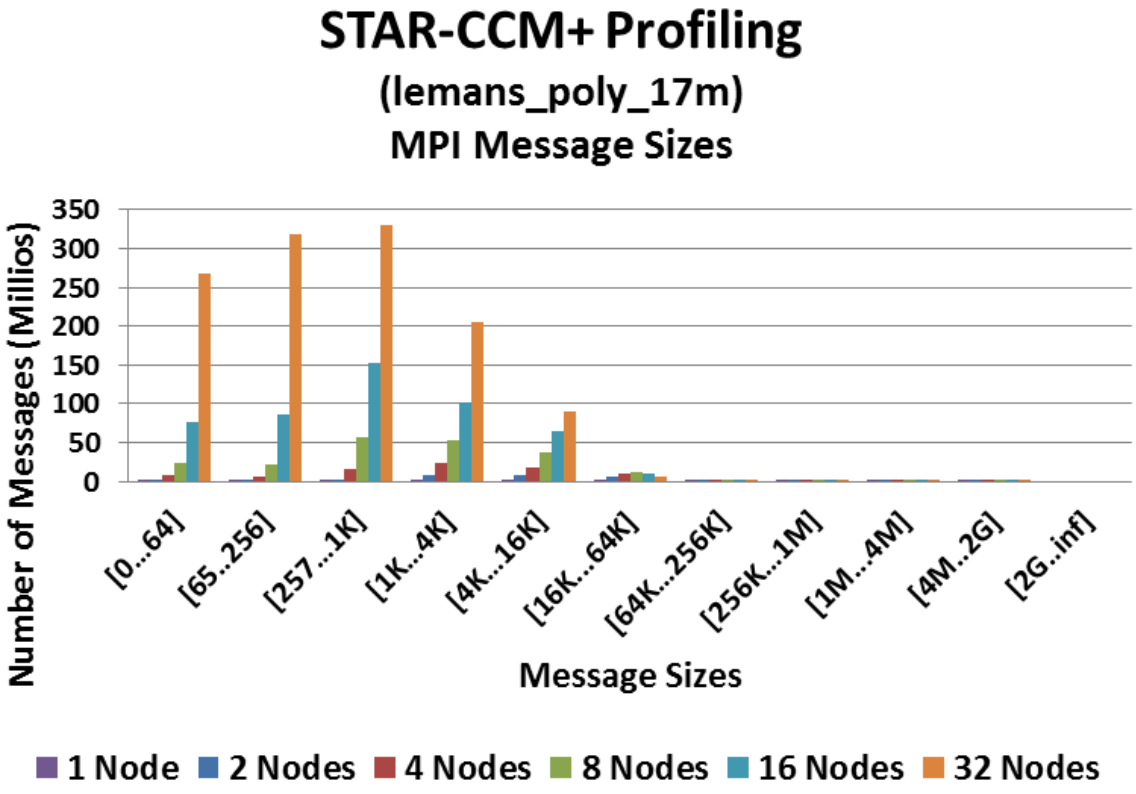
- **STAR-CCM+ spent more time in computation than communication**
  - The time ratio for network gradually increases with more nodes in the job
  - Improvement on network efficiency would reflect in improvement on overall runtime



FDR InfiniBand

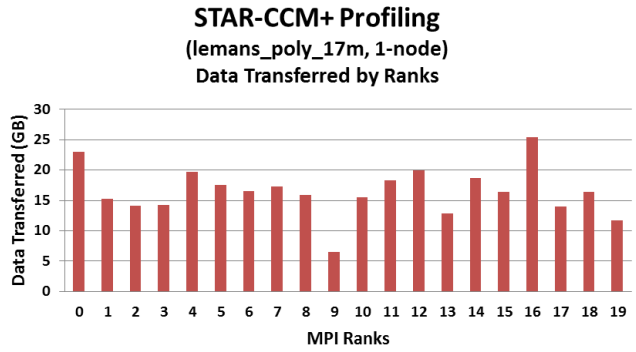
STAR-CCM+ Profiling – Message Sizes

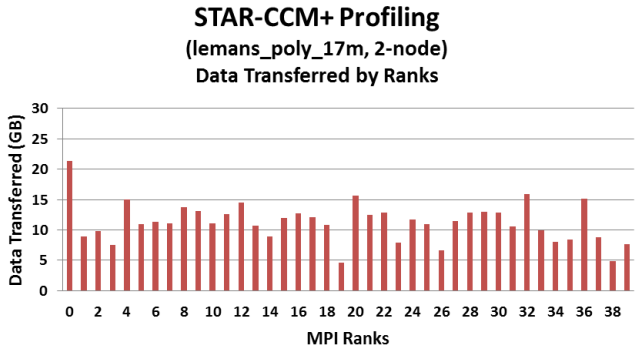
- **Majority of messages are small messages**
  - Messages are concentrated below 64KB
- **Number of messages increases with the number of nodes**

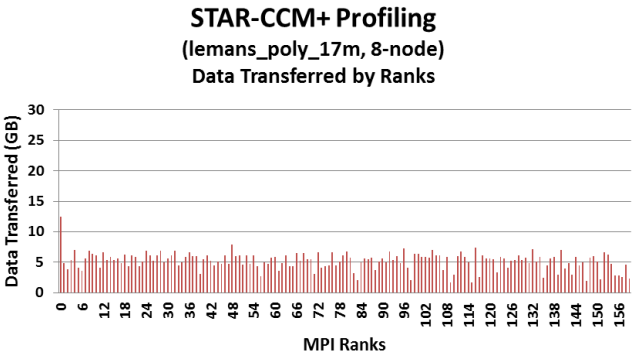


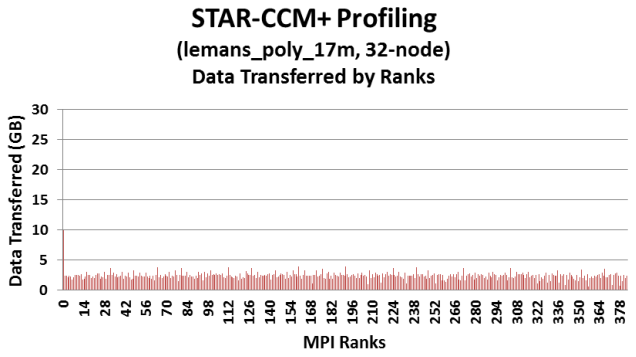
STAR-CCM+ Profiling – MPI Data Transfer

- **As the cluster grows, less data transfers between MPI processes**
  - Drops from ~20GB per rank at 1 node vs ~3GB at 32 nodes
  - Some node imbalances are seen through the amount of data transfers
  - Rank 0 shows significantly higher network activities than other ranks



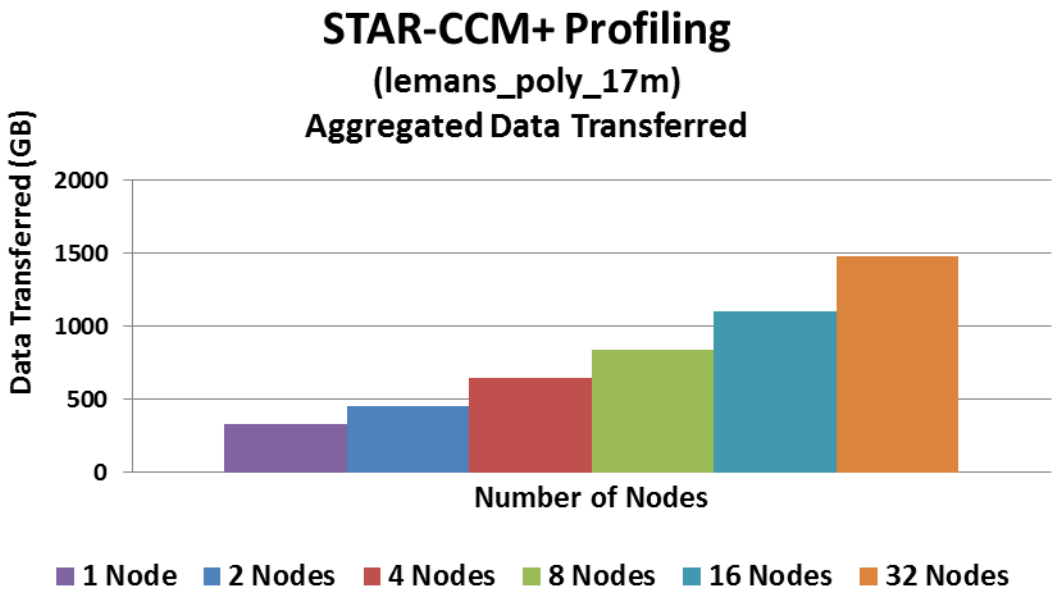






STAR-CCM+ Profiling – Aggregated Transfer

- **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
  - 1.5TB of data transfer takes place between the MPI processes at 32 nodes



Version 9.02.005

STAR-CCM+ – Summary

- **Performance**
  - STAR-CCM+ v9.02.005 improved on scalability over v8.06.005 by 28% at 32 nodes
    - Performance gap expect to widen at higher node count
  - FDR InfiniBand delivers the highest network performance for STAR-CCM+ to scale
  - FDR IB provides higher performance against other networks
    - FDR IB delivers ~191% higher compared to 40GbE, ~208% vs 10GbE on a 32 node run
  - Deploying dual-port Connect-IB HCA provides 11% performance at 32 nodes
  - Performance improvement seen compared to older hardware/software generations
    - Approximately 55% higher performance for 1 generation and 2.6x for 2 generations
  - Enabling Turbo mode results in higher application performance
    - Up to 17% of the improvement seen by enabling Turbo mode
  - Mellanox HPC-X provides better performance than the alternatives
- **MPI Profiling**
  - Communication time reduction with v9.02.005 which improves overall performance
  - Ethernet solutions consumes more time in communications
    - Spent 73%-95% of overall time in network due to congestion in Ethernet, while IB spent ~38%

Thank You

HPC Advisory Council