The following research was performed under the HPC Advisory Council activities
- Participating vendors: 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

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
  - STAR-CCM+ performance benchmarking
    - MPI Library performance comparison
    - Interconnect performance comparison
    - CPUs comparison
    - Optimization tuning
  - The presented results will demonstrate
    - The scalability of the compute environment/application
    - Considerations for higher productivity and efficiency
Test Cluster Configuration

- **Dell PowerEdge R730 32-node (896-core) “Thor” cluster**
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs (BIOS: Maximum Performance, Home Snoop)
  - Memory: 64GB memory, DDR4 2133 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 6.5, MLNX_OFED_LINUX-3.0-1.0.1 InfiniBand SW stack
  - Hard Drives: 2x 1TB 7.2 RPM SATA 2.5” on RAID 1
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7700 36-port EDR 100Gb/s InfiniBand Switch**
- **Mellanox ConnectX-3 FDR VPI InfiniBand and 40Gb/s Ethernet Adapters**
- **Mellanox SwitchX-2 SX6036 36-port 56Gb/s FDR InfiniBand / VPI Ethernet Switch**
- **Dell InfiniBand-Based Lustre Storage based on Dell PowerVault MD3460 and Dell PowerVault MD3420**
- **MPI: Platform MPI 9.1.2**
- **Application: STAR-CCM+ 10.02.012**
- **Benchmarks: lemans_poly_17m, civil_trim_20m, reactor_9m, LeMans_100M.amg**
PowerEdge R730
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
EDR InfiniBand delivers superior scalability in application performance:
- IB delivers 66% higher performance than 40GbE, 88% higher than 10GbE at 32 nodes.
- Scalability stops beyond 4 nodes for 1GbE; scalability is limited for 10/40GbE.
- Input data: Lemans_poly_17m: A race car model with 17 million cells.

**Higher is better**
**EDR InfiniBand delivers superior scalability in application performance**
- EDR IB provides 177 higher performance than 40GbE, 194% than 40GbE at 32 nodes
- InfiniBand demonstrates continuous performance gain at scale
- Input data: reactor_9m: A reactor model with 9 million cells

**STAR-CCM+ Performance**
*(reactor_9m)*

![Performance Chart](chart.png)

*Higher is better*

28 MPI Processes / Node
• **EDR InfiniBand demonstrates linear scalability in STAR-CCM+ performance**
  – The benchmark option in STAR-CCM+ calculates the speed up for the input data tested
  – STAR-CCM+ is able to achieve linear scaling with EDR InfiniBand
  – Other interconnects only provided limited scalability, as demonstrated in previous slides

*Higher is better*
STAR-CCM+ Performance – Input Data

- Larger cell sizes would consume more time to compute
  - 100 million cells takes about 219% longer to simulate than the 17 million cell case
  - About ~5 times increase in cell size would result in 219% increase in runtime
Current system generations of HW & SW configuration outperform prior generations

- Current Haswell systems outperformed Ivy Bridge by 38%, Sandy Bridge by 149%, Westmere by 409%
- Dramatic performance benefit due to better system architecture in compute and network scalability

System components used:

- Haswell: 2-socket 14-core E5-2697v3@2.6GHz, DDR4 2133MHz DIMMs, ConnectX-4 EDR InfiniBand, v10.02.012
- Ivy Bridge: 2-socket 10-core E5-2680v2@2.8GHz, DDR3 1600MHz DIMMs, Connect-IB FDR InfiniBand, v9.02.005
- Sandy Bridge: 2-socket 8-core E5-2680@2.7GHz, DDR3 1600MHz DIMMs, ConnectX-3 FDR InfiniBand, v7.02.008
- Westmere: 2-socket 6-core x5670@2.93GHz, DDR3 1333MHz DIMMs, ConnectX-2 QDR InfiniBand, v5.04.006

Higher is better
STAR-CCM+ Profiling – % of MPI Calls

- **For the most time consuming MPI calls:**
  - Lemans_17m: 55% MPI_Allreduce, 23% MPI_Waitany, 7% MPI_Bcast, 7% MPI_Recv
  - Reactor_9m: 59% MPI_Allreduce, 21% MPI_Waitany, 7% MPI_Recv, 4% MPI_Bcast

- **MPI as a percentage in wall clock times:**
  - Lemans_17m: 12% MPI_Allreduce, 5% MPI_Waitany, 2% MPI_Bcast, 2% MPI_Recv
  - Reactor_9m: 15% MPI_Allreduce, 5% MPI_Waitany, 2% MPI_Recv, 1% MPI_Bcast

**lemans_17m – 32 nodes / 896 Processes**

- MPI_Allreduce
- MPI_Waitany
- MPI_Bcast
- MPI_Recv
- MPI_Test
- MPI_Isend
- MPI_Waitall
- MPI_Alltoall
- MPI_Irecv
- MPI_Allgather
- MPI_Allgatherv
- MPI_Barrier
- MPI_Wait
- MPI_Send

**reactor_9m – 32 Nodes / 896 Processes**

- MPI_Allreduce
- MPI_Waitany
- MPI_Bcast
- MPI_Recv
- MPI_Test
- MPI_Isend
- MPI_Waitall
- MPI_Alltoall
- MPI_Irecv
- MPI_Allgather
- MPI_Allgatherv
- MPI_Wait
- MPI_Send
- MPI_Gatherv
For the most time consuming MPI calls
- Lemans_17m: MPI_Allreduce 4B (30%), 16B (19%), 8B (6%), MPI_Bcast 4B (4%)
- Reactor_9m: MPI_Allreduce 16B (35%), 4B (15%), 8B (8%), MPI_Bcast 1B (4%)
Majority of the MPI time is spent on MPI collective Operations and non-blocking communications
- Heavy use of MPI collective operations (MPI_Allreduce, MPI_Bcast) and MPI_Waitany

Some node imbalances characteristics shown on both input dataset
- Some processes appeared to take more time in communications, in MPI_Allreduce

lemans_17m – 32 nodes / 896 Processes

reactor_9m – 32 Nodes / 896 Processes
STAR-CCM+ Summary

**Performance**

- Compute: cluster of the current generation outperforms system architecture of previous generations
  - Outperformed Ivy Bridge by 38%, Sandy Bridge by 149%, Westmere by 409%
  - Dramatic performance benefit due to better system architecture in compute and network scalability
  - The input case of 100 million cells takes about 219% longer to simulate than the 17 million cell case
- Network: EDR InfiniBand demonstrates superior scalability in STAR-CCM+ performance
  - EDR IB provides higher performance by over 4-5 times vs 1GbE, 10GbE and 40GbE, 15% vs FDR IB at 32 nodes
  - Lemans_17m: Scalability stops beyond 4 nodes for 1GbE; scalability is limited for 10/40 GbE
  - Reactor_9m: EDR IB provides 177 higher performance than 40GbE, 194% than 40GbE at 32 nodes
  - EDR InfiniBand demonstrates linear scalability in STAR-CCM+ performance on the test cases
- MPI Profiles
  - Most MPI times are spent on MPI collective operations and non-blocking communications
  - Heavy use of MPI collective operations (MPI_Allreduce, MPI_Bcast) and MPI_Waitany
  - Lemans_17m: 55% MPI_Allreduce, 23% MPI_Waitany, 7% MPI_Bcast, 7% MPI_Recv
  - Reactor_9m: 59% MPI_Allreduce, 21% MPI_Waitany, 7% MPI_Recv, 4% MPI_Bcast
Thank You

HPC Advisory Council