Performance and analysis of manufacturing applications on an HPC cluster

Performance benchmarking results, tuning and analysis for ANSYS Fluent, CD-adapco STAR-CCM+, LSTC LS-DYNA and NICE DCV and EnginFrame on a high performance system that includes compute, storage, remote visualization and networking.

Garima Kochhar
Joshua Weage
Dell HPC Engineering
December 2015

A Dell Technical White Paper

Revisions

Date	Description
December 2015	Initial release – v1

Table of contents

1 Introduction 5
2 System Architecture 6
2.1 Compute 6
2.2 Accelerators and GPUs for compute 7
2.3 Local storage 8
2.4 Shared NFS storage – Dell NSS-HA solution 8
2.5 Shared parallel file system – Dell Intel Enterprise Edition for Lustre Solution 9
2.6 MPI/computation Network 9
2.7 Administration Network 9
2.8 Remote Visualization 10
2.9 Large memory server 10
2.10 Management 10
2.11 Software 10
2.12 Resource Manager 11
2.13 Racks, Power and Weight considerations 11
3 Test bed details 12
4 Performance results and analysis 17
4.1 STREAM 17
4.2 HPL 17
4.3 ANSYS Fluent 18
4.4 LS-DYNA 24
4.5 STAR-CCM+ 32
5 Remote Visualization 37
6 Power consumption results 41
7 Conclusion 45

Executive summary

This technical white paper describes an example single rack HPC solution for manufacturing applications that combines compute, storage, network and remote visualization. System design based on considerations specific to users in the manufacturing domain, detailed performance results with sample CFD, CAE and remote visualization applications, as well as power characteristics of the system are presented in this document.

1 Introduction

This technical study focuses on the requirements and design of a High Performance Computing (HPC) system for the manufacturing domain. A single rack system comprised of 32 compute nodes, a remote visualization server and 240 TB of storage was built in the Dell HPC Engineering lab to verify the design choices and to benchmark the system performance. The benchmark results from this study are also relevant for smaller and larger sized systems.

The system design and rationale behind the design choices are presented in Section 2. Section 3 lists the hardware, software and application versions, and the benchmark test cases that were used to measure and analyze the performance of the test system. Section 0 quantifies the capabilities of the system and presents performance on three manufacturing applications. The applications included as part of this study include ANSYS Fluent, LSTC LS-DYNA, and CD-adapco STAR-CCM+®. Section 5 includes details on the remote visualization capabilities included with the lab test system and Section 6 contains the power consumption details for the full system across the different workloads.

2 System Architecture

In order to effectively run manufacturing workloads, a system must provide computation capabilities for the simulations, storage that can satisfy the requirements of fast temporary scratch space as well as a repository for project results which allows collaboration, and options for visualizing the simulation results. The system should be easy to use for the engineers while providing quick turnaround time, and be simple to configure and manage for the system administrator and IT staff. Additionally, a well-designed system should be tuned for manufacturing workloads, balancing the unique computation and I/O needs. Keeping these requirements in mind, this section describes the architecture of a single rack HPC system which was built in the Dell HPC Engineering lab and explains the design choices which were made for this system.

At a high level, the configuration included 32 compute nodes, 240 TB of shared storage, a cluster master node and a remote visualization node in one rack. The system was interconnected using InfiniBand and Ethernet networks.

Each of the system components is described in detail below, including the compute, storage, networking, software and management options.

2.1 Compute

The PowerEdge C6320 server was used as the compute component in the system. With 4 server sleds in a 2U chassis, this form factor is very popular in HPC systems for its density, performance and ease of management. The shared infrastructure chassis provides common power and cooling. Each server sled is an individual dual socket system supporting the Intel Xeon E5-2600 v3 family of processors (architecture code named Haswell) and 16 DDR4 memory DIMM slots. The server includes onboard 10GbE network adapters and InfiniBand support via a mezzanine PCI-E slot. Figure 1 shows the chassis and the four server sleds.

Figure 1 Compute node - PowerEdge C6320

For the lab test system, the PowerEdge C6320 servers were configured with dual Intel Xeon E5-2660 v3 processors and 128 GB memory. With 32 servers in the lab test system, this was a total of 640 cores and 4 TB of memory.

A previous white paper included a detailed analysis of processor choices for manufacturing, taking into account CPU frequency, core count, processor cost and license costs. The Intel Xeon E5-2660 v3 processor was used for the lab test cluster; however, any of the E5-2600 v3 family of processors could be used instead based on individual performance, cost, power and cooling considerations.

The previous study also discussed optimal memory configurations. With this generation of servers, all four memory channels for both CPU sockets must be identically populated for best performance. Given the number of compute servers in the lab test system, eight 16GB memory DIMMS per server was the best configuration optimizing for cost per GB as well as performance considerations. This is also a customizable component and the memory capacity can be changed as desired.

2.2 Accelerators and GPUs for compute

Most manufacturing workloads, across all domains, rely on traditional x86 architecture for computation. However, some codes, both commercial and in-house, have been ported to run well on NVIDIA Tesla GPUs and Intel Xeon Phi accelerator cards. If there is a need to support both types of computation, PowerEdge C4130 servers can be included in the HPC manufacturing system. Each C4130 server is capable of supporting up to four GPUs/accelerator cards in addition to two Intel Xeon processors in a 1U form factor as shown in Figure 2. The lab test system did not include GPU servers.

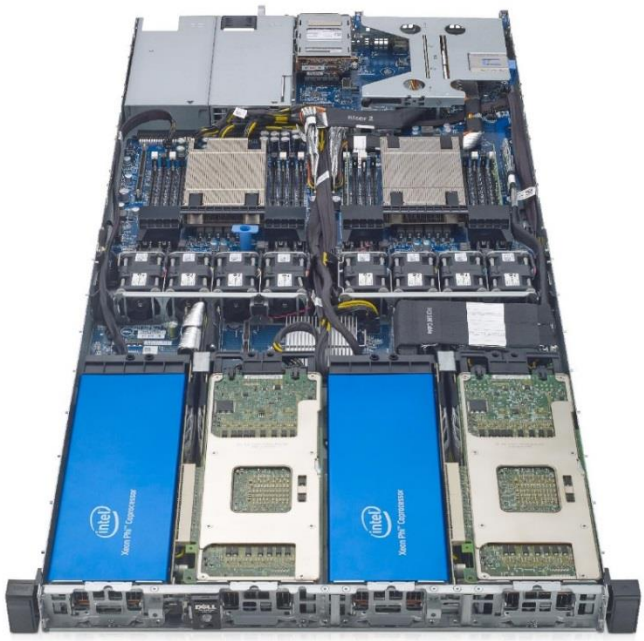

Figure 2 GPU and accelerator node - PowerEdge C4130

2.3 Local storage

Each PowerEdge C6320 server can support up to six 2.5" drives or four 3.5" drives. The 2.5" configuration is shown in Figure 3 with 24 total drives, six drives for each of the four servers in the chassis. The lab test system was configured with six 2.5" 10K 600GB SAS drives per server. The drives were configured in RAID 0 to provide fast local scratch space for structural codes and applications that write a lot of temporary data during analysis. The operating system was also installed on the local disks for each server. This configuration is overkill for most workloads that do not perform significant local I/O but, for simplicity, all servers were configured identically with six drives in the lab test system.

Figure 3 Local drives in the PowerEdge C6320

2.4 Shared NFS storage – Dell NSS-HA solution

The lab test system included 240 TB of shared NFS storage via the Dell NFS Storage Solution (NSS). The version used was the Dell NSS6.0-HA. NSS-HA is a performance tuned NFS solution, providing high availability (HA) features. The file system is based on Red Hat XFS, and is exported to the compute cluster via IPoIB. An example configuration is shown in Figure 4. In the lab test system the NSS-HA storage was used for user home directories and to provide a common repository for application images. It can also be used as the repository for simulation results. Since the NSS storage was used for home directories, the HA version of the solution was selected to provide better data availability and reduce downtime in the event of a hardware or software failure. With the HA version, the NFS service can seamlessly failover between the two NFS servers.

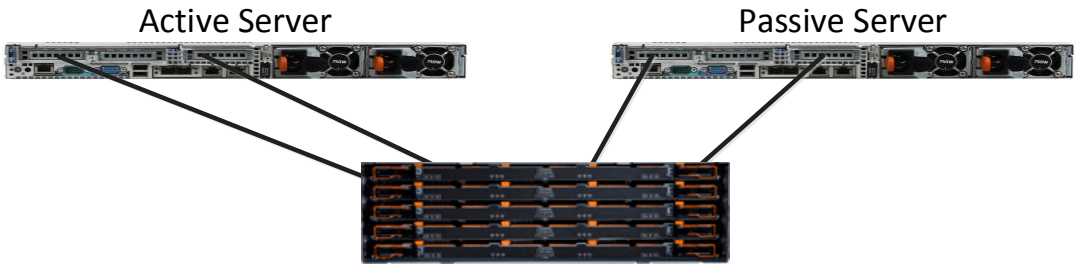

Figure 4 Dell NFS Storage Solution with High Availability

2.5 Shared parallel file system – Dell Intel Enterprise Edition for Lustre Solution

A parallel file system is advisable for HPC systems with more than 100 users, or when the system needs to support large jobs with parallel I/O requirements. When needed, the Dell Intel Enterprise Edition for Lustre Solution can provide a parallel file system for temporary storage of files created by the simulations. Final project results should be stored on the NFS storage. An example Lustre configuration is shown in Figure 5. The lab test system did not use a parallel file system since it did not require the capabilities that a parallel file system provides.

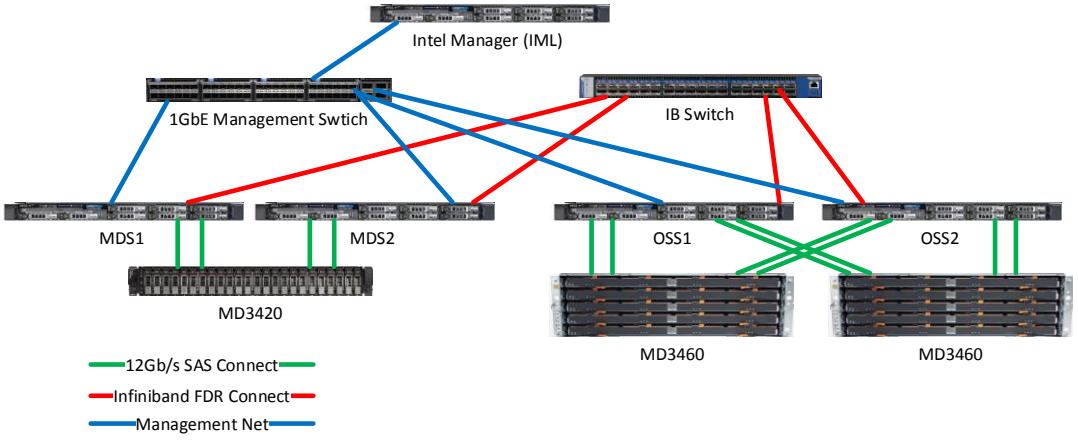

Figure 5 Dell Lustre configuration

2.6 MPI/computation Network

A high-speed fabric is necessary for most manufacturing applications to scale efficiently beyond more than three to four compute nodes. For this system, 56 Gbps FDR InfiniBand was chosen as the fabric, as it provides the right capability for manufacturing workloads for this size of system. The additional bandwidth provided by EDR InfiniBand is unnecessary and 10 Gigabit Ethernet would prove to be a performance bottleneck at this size. QDR InfiniBand or FDR10 are also possible choices for a system of this size.

The fabric for the lab test system was built utilizing Mellanox FDR InfiniBand HCAs and switches. A single 36-port InfiniBand switch was used, with 32 ports used by the compute nodes, two by the NFS servers in the NSS-HA solution, one by the remote visualization node and the last port by the cluster master node. This fully utilized the 36-port switch.

2.7 Administration Network

A Gigabit Ethernet network was used for deploying, administering and managing the system. All components in the system have onboard Ethernet controllers and these were connected to a single Dell Networking S3048-ON switch. The administration network was used for operating system installation, cluster administration, job submission and IPMI traffic.

2.8 Remote Visualization

Visualization of the simulation results is an important consideration for the manufacturing domain and can be accomplished in multiple ways. One option is for each user to view the results outside of the HPC system on individual workstations. These workstations would be located at the user's desk/workspace and each equipped with a high end graphics card as well as enough memory to support the visualization. In this case, the data would be moved from the HPC system to a location accessible by the workstations, or the workstations would need access to the HPC storage sub-system.

Another option is to use remote visualization where the user launches a remote desktop or remote application on the HPC system directly. This use case is appealing since a richly configured workstation per user is no longer required, and the visualization resources can be centrally managed along with the HPC system.

The lab test system included a PowerEdge R730 as a remote visualization system.

2.9 Large memory server

Each compute server in the lab test cluster was configured with 128 GB of memory to balance performance requirements with system cost. Some implicit FEA solvers, such as NASTRAN/AMLS, OptiStruct and ABAQUS, may require more memory per server for optimal performance. PowerEdge R630 or R730 servers providing up to 768 GB of DRAM per server can be included for this purpose. A large memory machine was not included in the lab test cluster, but this would need to be considered based on specific solver workload requirements.

2.10 Management

The lab test system used Bright Cluster Manager (BCM) for cluster deployment and administration. The system used a single master node; BCM does provide an option for two master nodes providing redundancy in an active-passive configuration if needed. The lab test system used the Gigabit Ethernet network to deploy, administer and manage the system.

All server components in the system were equipped with an Intelligent Platform Management Interface (IPMI) compliant Integrated Dell Remote Access Controller (iDRAC). Out-of-band hardware management was accomplished via IPMI, and the Gigabit Ethernet administration network was used for IPMI traffic.

2.11 Software

The cluster software used to deploy, administer and monitor the lab test system was Bright Cluster Manager (BCM). BCM is an easy to use tool that simplifies cluster management with enterprise class support and services.

The operating system used for the lab test system was Red Hat Enterprise Linux (RHEL). Specifically the OS version was RHEL 6.6 with errata kernel 2.6.32-504.16.2.el6.x86_64. This specific version was selected with care as it addresses a user process lock-up issue that has been known to impact some applications.

For the storage component, BCM was integrated with NSS-HA and used the shared NFS file system for user home directories as well as for application installation. For systems that require a parallel file system, the parallel file system can also be mounted and managed using BCM.

2.12 Resource Manager

An HPC system supporting multiple users and projects will require a workload manager such as Grid Engine, PBS Professional, Slurm or Torque. These tools manage the resources of the system, schedule jobs for multiple users, implement system usage policies and provide system usage reporting. They enable multiple simultaneous simulations to be run, scheduling jobs to optimize the utilization of system resources. The lab test system was configured with Torque.

2.13 Racks, Power and Weight considerations

An extra-deep, extra-wide standard height rack was used for the lab test system. These racks can easily accommodate all the components in the system while providing extra space for simplified cable management.

The lab test system was configured to weigh less than 1500 lbs. and consume less than 25kW of power. Today's server and storage component choices can provide very dense solutions and it is easy to pack a single rack with equipment weighing upwards of 2500 lbs., but most data centers only support a much lower per-rack weight limit. Similarly, most data centers can easily supply 25kW of power per rack and provide adequate cooling for that load. The configuration of the lab test system was designed to fit into most data centers by considering typical weight and power limitations.

3 Test bed details

The lab test system was extensively benchmarked using three manufacturing applications: ANSYS Fluent, LS-DYNA from LSTC and CD-adapco STAR-CCM+. This section describes the test environment used for the performance benchmarking.

The hardware configuration used in the lab is shown in Figure 6.

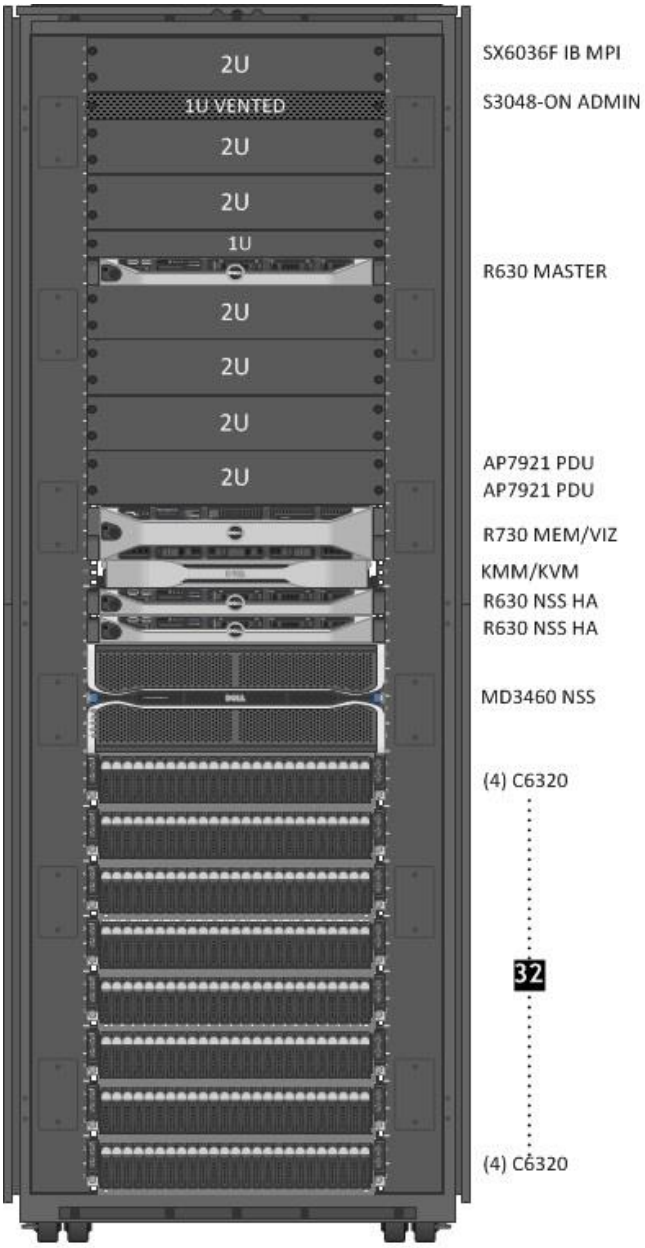

Figure 6 Lab test system

The components of the lab test system are listed in Table 1. Table 2, Table 3 and Table 4 list the detailed hardware configuration for the compute nodes, master node and remote visualization node. Table 5 lists the configuration of the NSS6.0-HA storage component.

Table 1 Lab test system components

System component	Details
Compute nodes	32 PowerEdge C6320 servers
Master node	PowerEdge R630
Remote Visualization node	PowerEdge R730
Shared Storage	Dell NSS6.0-HA 240TB
Ethernet network switch	Dell Networking S3048-ON
InfiniBand network switch	Mellanox FDR SX6036F

Table 2 Compute node configuration

Hardware component	Details
System	PowerEdge C6320 server
Processor	Dual Intel Xeon E5-2660 v3 – 2.6 GHz, 10c, 105W
Memory	128 GB. 8x16 GB 2133 MT/s DDR4 DIMMs
Disks	6x600 GB 10K SAS, RAID0
RAID controller	PERC H330
Network	Onboard Ethernet adapter Mellanox FDR ConnectX-3 mezzanine card

Table 3 Master node configuration

Hardware component	Details
System	PowerEdge R630 server
Processor	Dual Intel Xeon E5-2660 v3 – 2.6 GHz, 10c, 105W
Memory	256 GB. 16x16 GB 2133 MT/s DDR4 DIMMs
Disks	6x600 GB 10K SAS, RAID 5
RAID controller	PERC H730
Network	Onboard Ethernet adapter Mellanox FDR ConnectX-3 mezzanine card

Table 4 Remote visualization node configuration

Hardware component	Details
System	PowerEdge R730 server
Processor	Dual Intel Xeon E5-2695 v3 – 2.3 GHz, 14c, 120W
Memory	256 GB. 16x16 GB 2133 MT/s DDR4 DIMMs
Disks	2x300 GB 15K SAS
RAID controller	PERC H730
Network	QLogic 10GbE network daughter card Mellanox FDR ConnectX-3 mezzanine card
GPU	NVIDIA GRID K2

Table 5 Dell NSS6.0-HA configuration

Hardware component	Details
NFS Servers	Two PowerEdge R630 NFS servers
Processor	Dual Intel Xeon E5-2697 v3 – 2.6 GHz, 14c, 145W
Memory	128 GB. 16x8 GB 2133 MT/s DDR4 DIMMs
Local Disks	5x300 GB 15K SAS, RAID 1 for OS, RAID 0 for swap
RAID controller	PERC H730
Network	Onboard Ethernet adapter Mellanox FDR ConnectX-3 mezzanine card
NFS Storage	One PowerVault MD3460
NFS Disks	240 TB. 60x4 TB NL-SAS drives
PDU	Two AP7921 PDUs for HA functionality
Operating System	Red Hat Enterprise Linux 7.0
Kernel version	3.10.0-210.el7.x86_64
File System	Red Hat Scalable File System, XFS 3.2.0-alpha2

Table 6 lists the BIOS tuning options.

Table 6 BIOS tuning

BIOS option	BIOS setting
Logical Processor	Disabled
Memory Snoop Mode	Early Snoop
Node Interleaving	Disabled
System Profile	DAPC Profile for performance scaling tests Turbo enabled, C-states and C1E enabled. Performance Profile for results in power section Turbo enabled, C-states and C1E disabled

Table 7 lists the software versions on lab test system.

Table 7 Software versions

Components	Software Versions
Operating System	RHEL 6.6
Kernel	2.6.32-504.16.2.el6.x86_64
Bright Cluster Manager	v7.1 with RHEL 6.6 (Dell version)
Intel compilers	2016.0.109
Intel MKL from compilers	2016.0.109
Intel MPI	5.1.1.109
Platform MPI	09.01.00.01
Fluent code	v16.0.0
Fluent benchmarks	v15 and v16 cases
LS-DYNA code	MPP code R7 and R8. AVX2 and SSE2 binaries ls-dyna_mpp_s_r8_0_0_95359_x64_redhat54_ifort131_sse2_intelmpi-413 ls-dyna_mpp_s_r8_0_0_95359_x64_redhat54_ifort131_sse2_platformmpi ls-dyna_mpp_s_r8_0_0_98726_x64_redhat54_ifort131_avx2_intelmpi-413 ls-dyna_mpp_s_r8_0_0_98726_x64_redhat54_ifort131_avx2_platformmpi ls-dyna_mpp_s_r7_1_2_95028_x64_redhat54_ifort131_sse2_intelmpi-413

	ls-dyna_mpp_s_r7_1_2_95028_x64_redhat54_ifort131_sse2_platformmpi ls-dyna_mpp_s_r7_1_2_95028_x64_redhat54_ifort131_avx2_intelmpi-413 ls-dyna_mpp_s_r7_1_2_95028_x64_redhat54_ifort131_avx2_platformmpi
LS-DYNA benchmarks	car2car-ver10, with endtime=0.02 ODB-10M-ver14, with endtime=0.02
STAR-CCM+ code	10.02.012 (linux-x86_64-2.5/gnu4.8), mixed precision
STAR-CCM+ benchmarks	9 benchmark cases as listed. 20 iterations, 40 pre-iterations (-nits 20 -preits 40)
NICE DCV	2014.0 (r16231)
NVIDIA Driver	352.55

4 Performance results and analysis

This section presents the performance results on the lab test system using the configuration described in Section 3. The goals of this exercise were to verify the design, and quantitatively describe the performance characteristics of the system using applications from the manufacturing domain.

The operation of the base system was checked first, prior to any application benchmarking. This verifies that the individual sub-systems work as expected and the system itself is stable, The STREAM memory bandwidth test was used to check the memory configuration and HPL used to check the computational subsystem, power configuration and stress test the individual servers and the full system.

After the system was verified, the performance of the ANSYS Fluent, STAR-CCM+ and LS-DYNA benchmark test cases were measured on the system.

The test system as configured in Section 3 provided 640 cores. Application performance was measured from 1 server (20 cores) up to 32 servers (640 cores). This was done to measure the scalability of the system and the applications. As will be shown in the subsequent sections, some of the benchmark data sets are small in size and the overhead of problem decomposition and inter-process communication at large core counts out-weights the benefit of additional cores. Independent of the size of the data set, a resource manager (Section 2.12) is recommended to submit, schedule and manage jobs on the system. This will allow multiple users and multiple jobs to use the system concurrently, maximizing the utilization of the system while optimizing job turnaround time.

4.1 STREAM

The STREAM benchmark results are presented in Table 8. These results indicate that each server can sustain 113 GB/s memory bandwidth which is as expected for this configuration. All 32 compute servers have similar performance (less than 1% variation) and are working correctly.

Likewise, the memory bandwidth of the master node, remote visualization node and NSS servers was verified prior to application benchmarking.

Table 8 STREAM benchmark results

TRIAD	Min (compute)	Max (compute)	Avg (compute)	Variation
Bandwidth in GB/s	112954	113667	113173	0.63 %

4.2 HPL

High Performance Linpack (HPL) is a popular benchmark that is very computation heavy and stresses the computational sub-system extensively. It is used to rank the TOP500 fastest supercomputers in the world and is an important burn-in test, although not usually representative of actual real-world application performance. As a burn-in test it helps to quickly weed out unstable components and verify the power delivery to the system.

The precompiled Intel HPL binary from Intel MKL was used for this test. The 32 compute servers performed similarly on single-server tests and the results were within expectations, 630-650 GFLOPS per server. Cluster level HPL results are presented in Figure 7, which plots performance as additional cores are added to the test. HPL shows good scalability from one to 32 servers on this system, showing the system has a balanced design. The full compute system is capable of 20.1 TFLOPS!

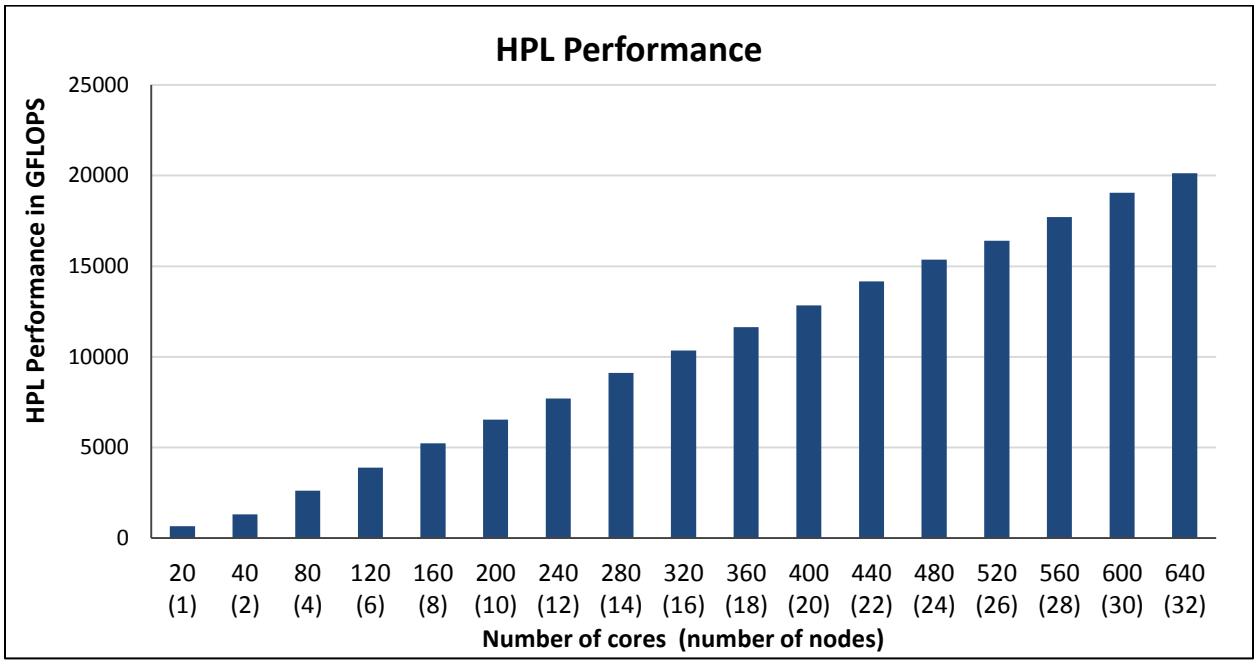

Figure 7 HPL performance

4.3 ANSYS Fluent

Multiple cases from Fluent benchmark suites v15 and v16 were tested on the lab test system. Between the older Fluent v15 benchmark cases and newly released v16 benchmark cases, there are 20 benchmark tests. For simplicity, eight cases are presented in this section. Truck_poly_14m and truck_111m are from the older v15 version; ice_2m, sedan_4m, combustor_12m, aircraft_wing_14m, combustor_71m and exhaust_system_33m are part of the newer v16 benchmark cases.

The graphs in Figure 8, Figure 9 and Figure 10 show the actual measured performance of the lab test system on 1 to 32 nodes using 20 to 640 cores as noted. Each data point on the graphs records the performance of the specific benchmark data set using the number of cores marked on the x-axis in a parallel simulation. The results are presented using the Solver Rating metric which counts the number of jobs that can be run in a day, i.e. <total seconds in a day>/<job runtime in seconds>. A higher value represents better performance. The results are divided into three charts for easy readability—the scale for Solver Rating is large and some models run much faster than others based on size of model, type of solver used, etc.

Combustor_71m and truck_poly_111m are large models and needed a minimum of two servers in this configuration. Recall that each compute node has 128 GB of memory and two nodes with 256 GB total memory are needed to accommodate these model sizes. The results for these cases therefore start at the 40 cores (2 node) mark.

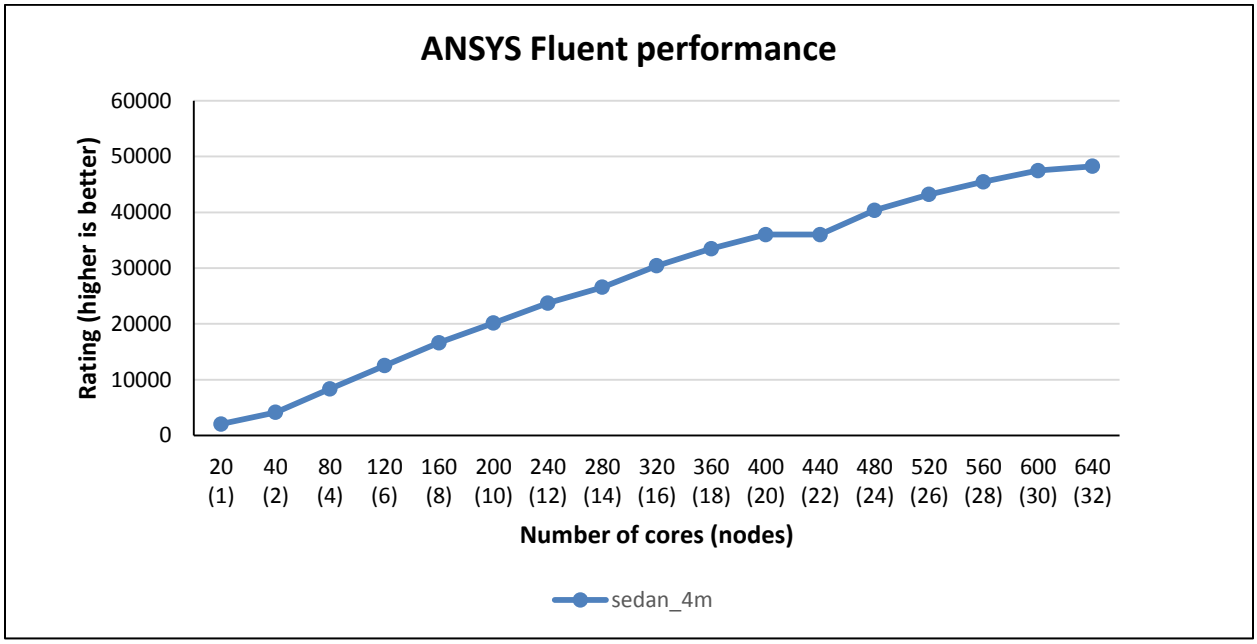

Figure 8 ANSYS Fluent performance (sedan)

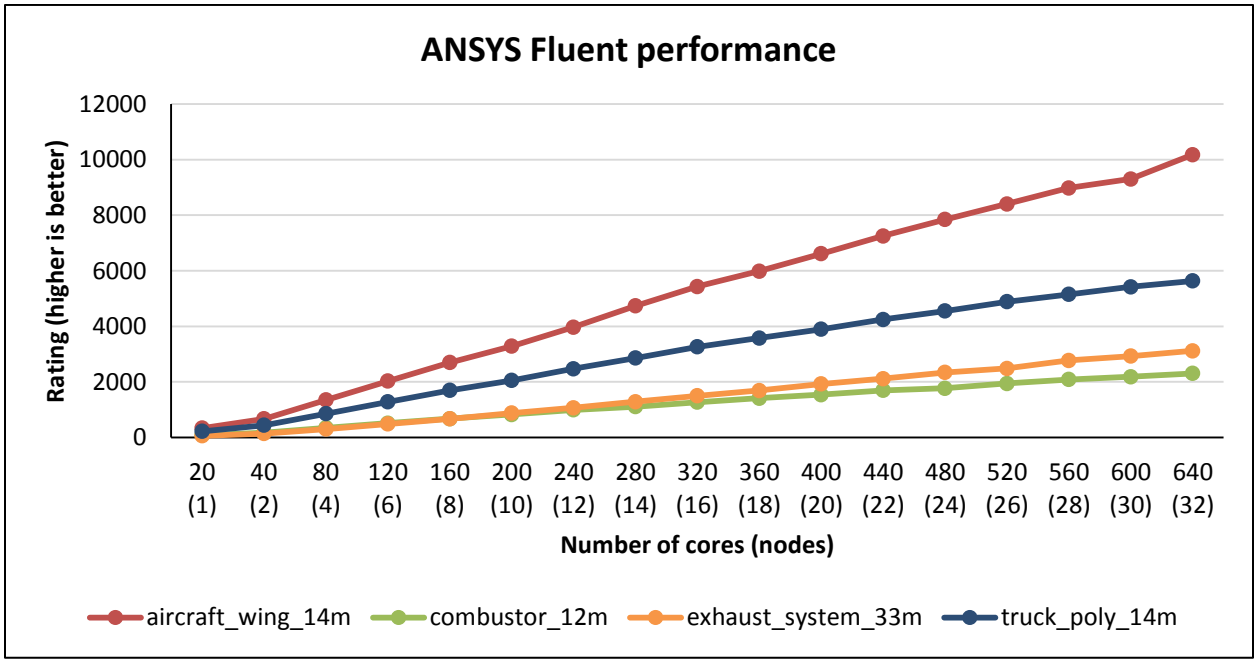

Figure 9 ANSYS Fluent performance (aircraft_wing, combustor_12m, exhaust_system, truck_poly)

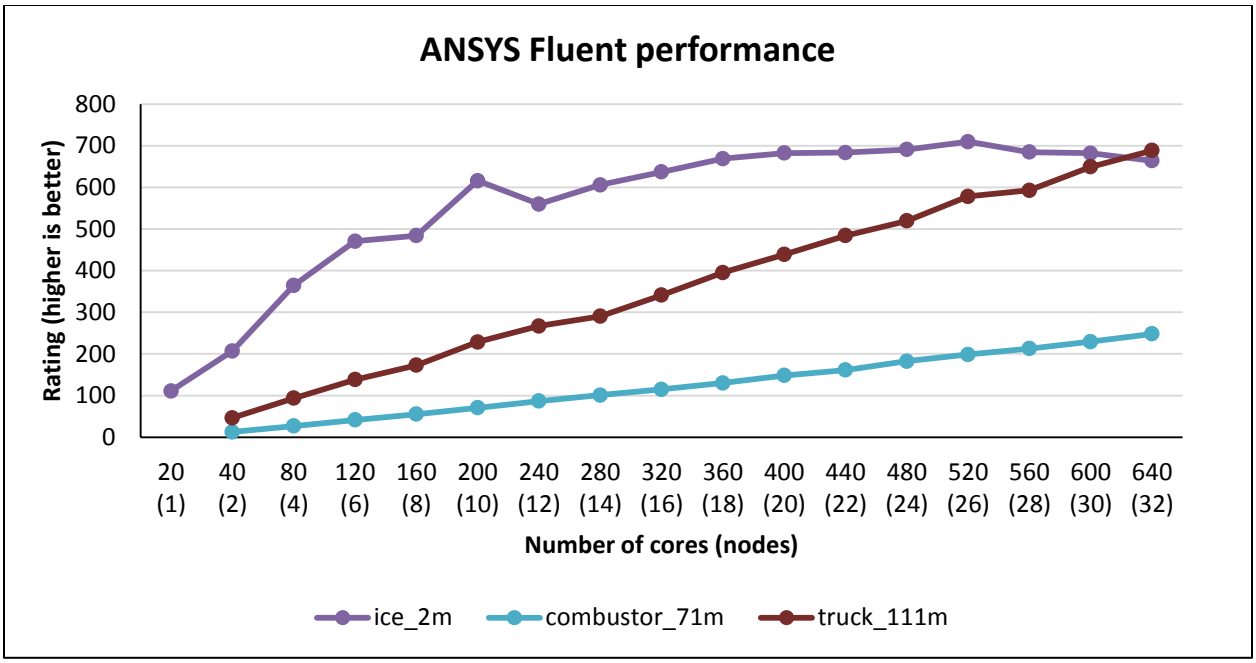

Figure 10 ANSYS Fluent performance (ice, combustor_71m, truck_111m)

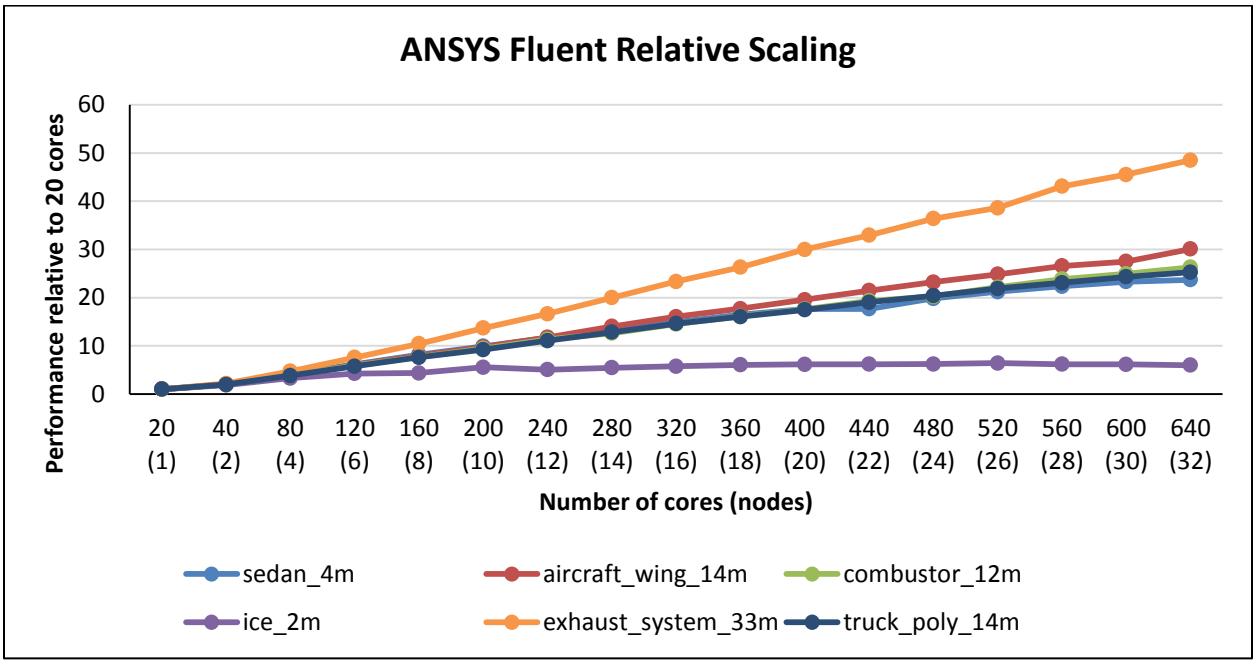

Figure 11 ANSYS Fluent relative scaling

Figure 11 presents the same performance data but plotted relative to the "20 cores (1 node)" result. This makes it easy to see the scaling of the solution, i.e. the performance improvement as more cores are used for the analysis. Most test cases in the graph scale well, almost linearly. Ice_2m plateaus out around 200 cores; this is expected from this small model.

Figure 12 also presents relative data for the two larger cases, combustor_71m and truck_111m. Results are plotted relative to the "40 cores (2 node)" result, the first valid data point for this configuration. These two test cases also scale very well on the lab test system, with close to linear scalability.

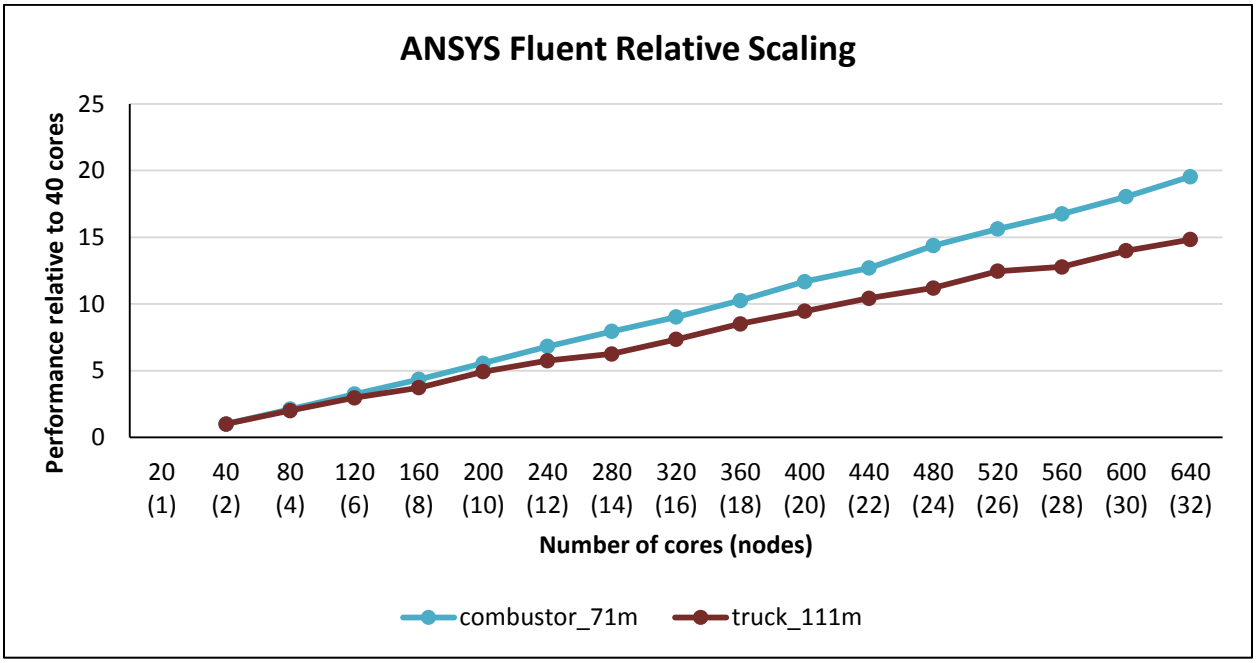

Figure 12 ANSYS Fluent relative scaling (combustor_71m, truck_111m)

Results on a sub-set of these tests cases were presented in a previous study on a 4-node system interconnected using standard 10 Gigabit Ethernet. The results below compare the current results over InfiniBand with the previous 10 Gigabit Ethernet results. The processor and memory configuration of the two configurations is similar, making this a valid comparison.

ANSYS Fluent variation between test runs is ~2-3%, so any performance difference less than 3% is not statistically significant. Looking at the interconnect comparisons in Figure 13 to Figure 17, it is clear that the benefit of a faster interconnect like InfiniBand, as compared with 10 Gigabit Ethernet, is apparent only around three or four nodes. This leads to two conclusions: 10 Gigabit Ethernet on the four node configuration is a good choice for the size of that system, and larger systems like the lab test configurations need a faster, low-latency fabric to make effective use of the greater number of servers and cores. Note that the 10 Gigabit Ethernet network in the previous study was not tuned for latency and tuning of the 10 GbE network could further help ANSYS Fluent performance.

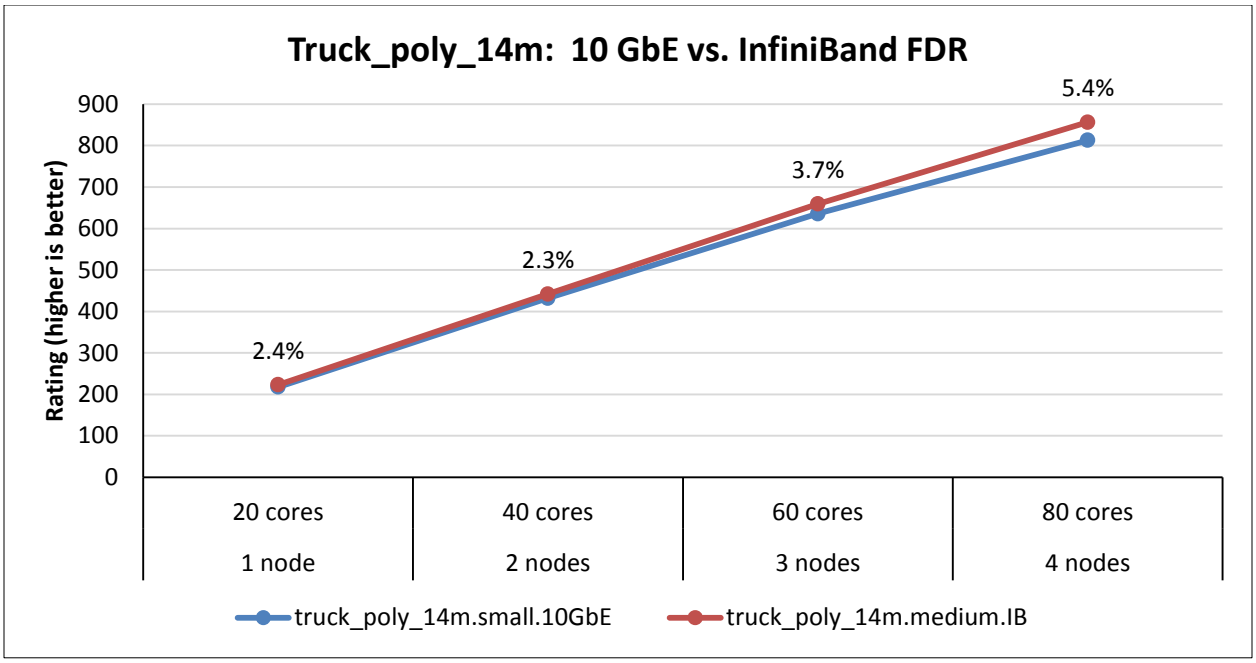

Figure 13 ANSYS Fluent – truck_poly_14m comparison

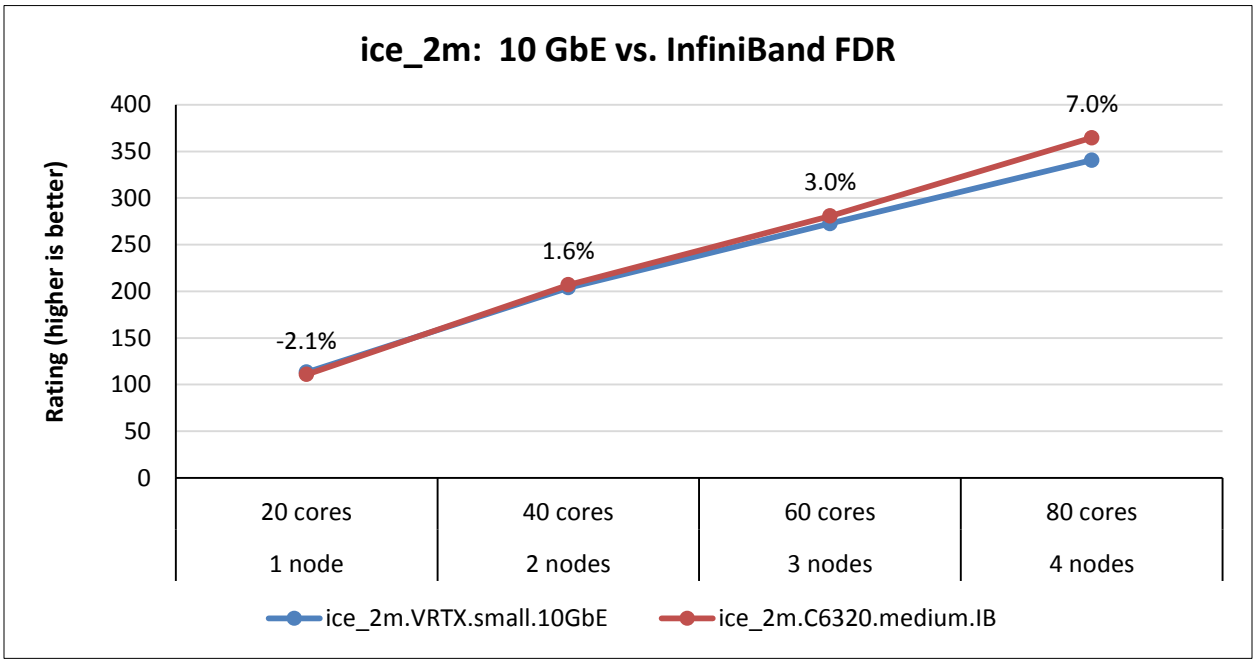

Figure 14 ANSYS Fluent – ice_2m comparison

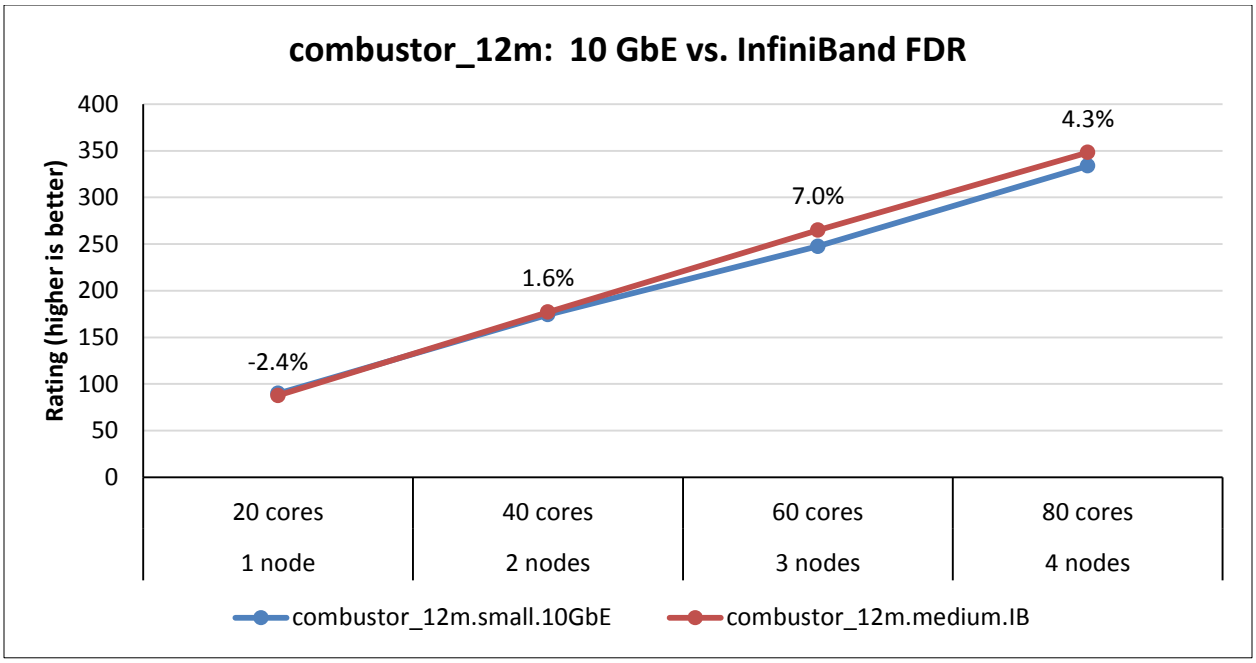

Figure 15 ANSYS Fluent – combustor_12m comparison

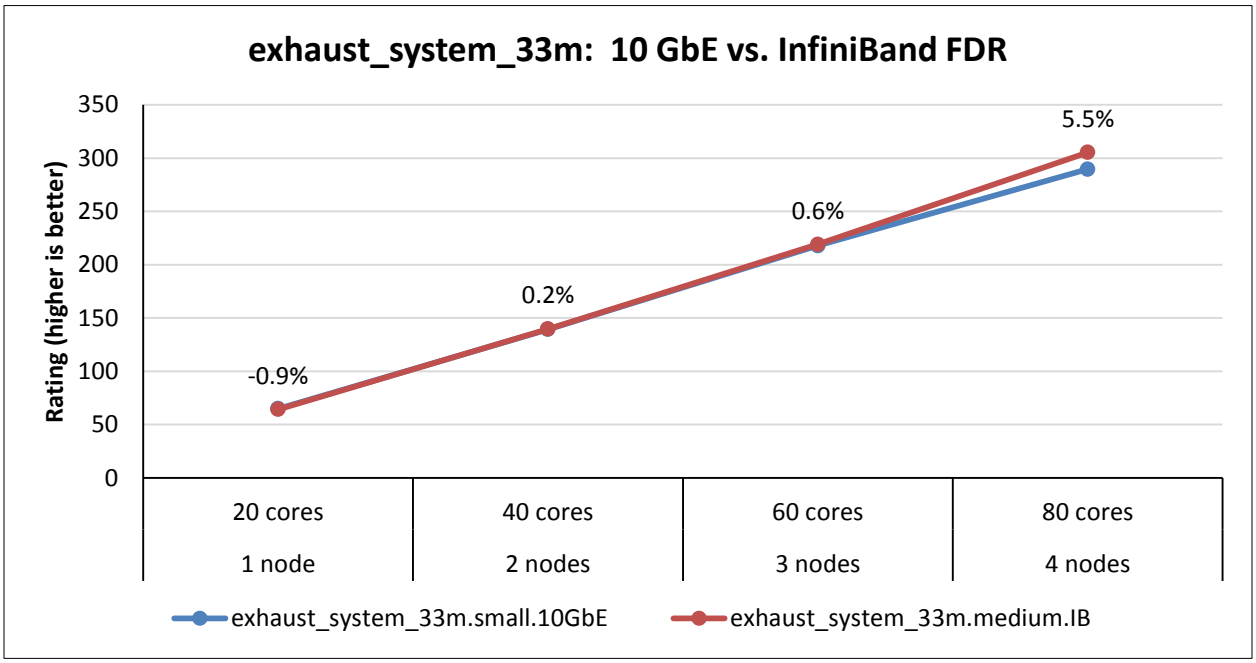

Figure 16 ANSYS Fluent – exhaust_system_33m comparison

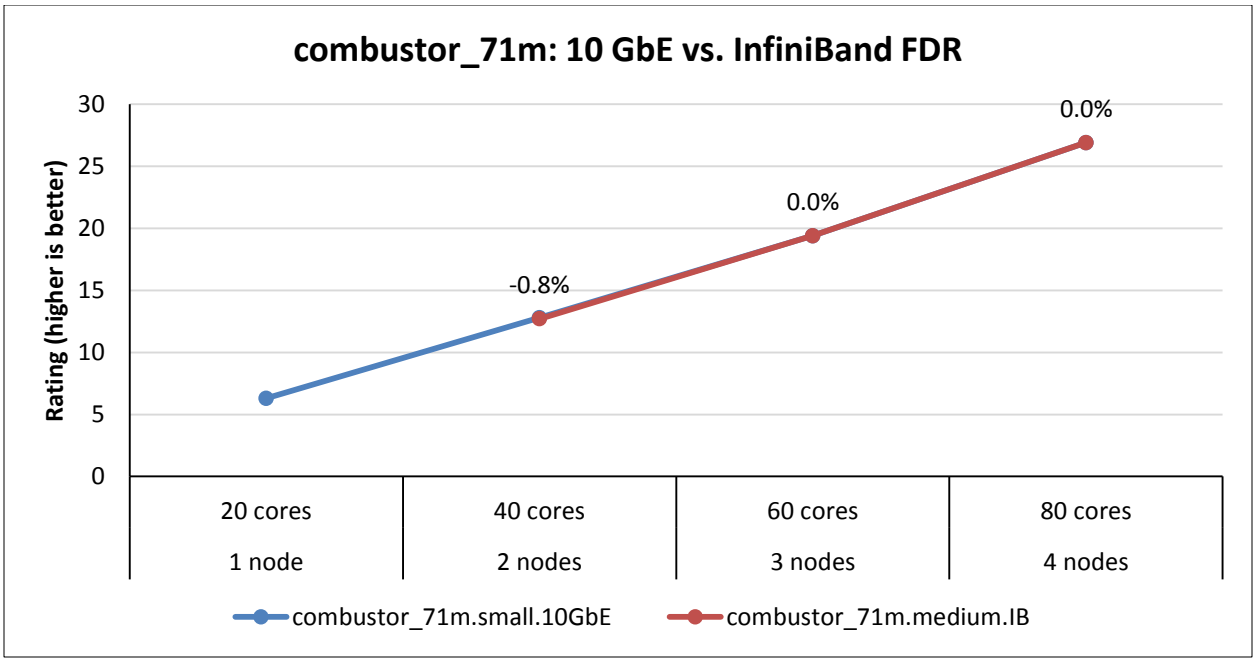

Figure 17 ANSYS Fluent – combustor_71m comparison

4.4 LS-DYNA

Several types of performance analysis tests were run with LS-DYNA on the lab test system. Various factors affecting LS-DYNA performance were evaluated. For example, the location of temporary files written by LS-DYNA is a user configurable option which can affect job elapsed time. Many users of LS-DYNA choose to checkpoint jobs at regular intervals to facilitate easy restart in case of job failure. For these use cases, the file system performance is an important aspect that can affect overall system performance. These factors and others are evaluated in the following analysis.

The LS-DYNA aspects evaluated include:

1. AVX2 vs. SSE2 binary performance
2. Intel MPI vs. Platform MPI comparison
3. R7 vs. R8 comparison
4. local configured on local scratch vs. local configured on shared NFS storage
5. Check-pointing performance
6. Scaling dependency on model size

4.4.1 LS-DYNA car2car

The LS-DYNA car2car-ver10 dataset is a 2.4 million element model. By default, this dataset is configured with a simulation time of 0.120 s. In order to reduce the runtime for benchmark analysis, "endtime=0.02" was used. This allows the evaluation of many more test combinations which are still representative of the performance of the unmodified dataset.

Figure 18 plots the performance of the car2car dataset on the lab test system. The metric for performance is Elapsed Time, with a lower value representing better/faster performance. The graph shows the performance of four R8 LS-DYNA binaries, AVX2 and SSE2 with Intel MPI and Platform MPI. The scaling, i.e. the performance improvement as more cores are used in the simulation, is clear for small number of nodes with the Elapsed Time reducing significantly as more servers are added to the simulation. Performance begins to plateau around 280 cores (14 nodes).

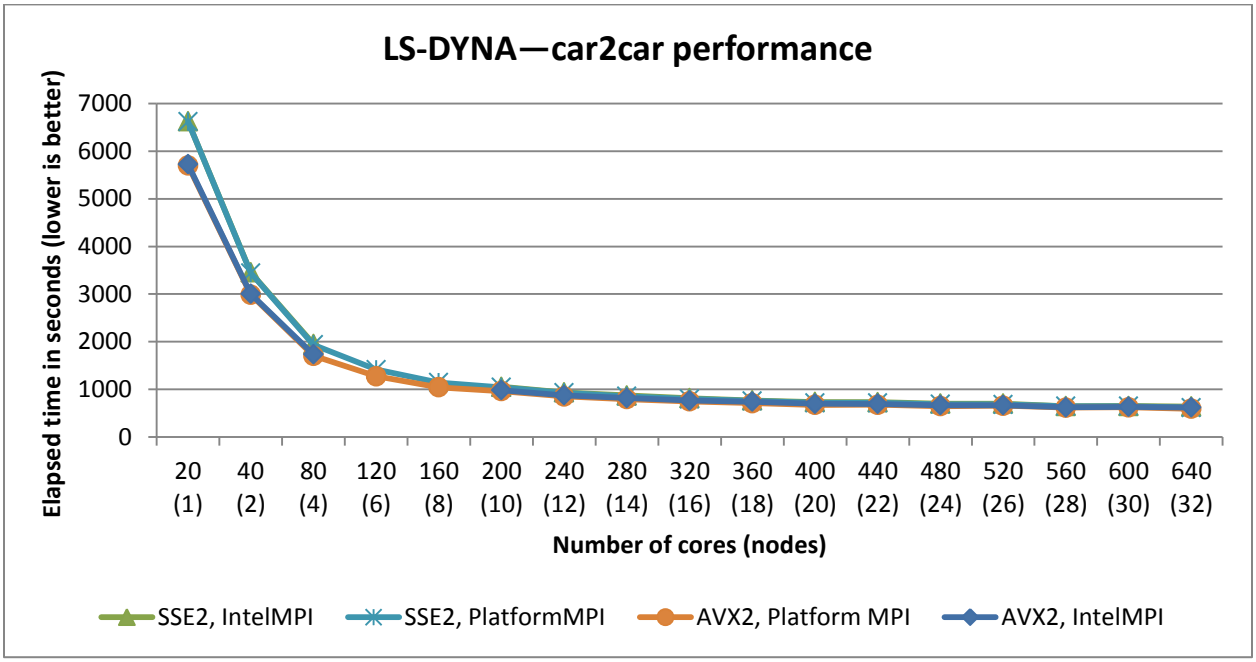

Figure 18 LS-DYNA car2car performance

This is better illustrated in Figure 19 which plots the relative performance of the R8 AVX2, Platform MPI case. For the car2car data set, this type of scaling behavior is expected. The dataset consists of 2.4M elements and is expected to scale to about 240 cores. After that point, adding more cores does not proportionally increase performance.

To take a closer look at the difference in performance between the two MPI implementations and the AVX2 and SSE2 R8 binaries, Figure 20 and Figure 21 plot the same result but at a more granular level. The results show interesting patterns:

1. For each MPI, the AVX2 binaries perform better than the SSE2 versions by 4 to 16%. At 20 cores (1 node), the AVX2 is 16% better than the SSE2 code. At 160 cores (8 nodes), the performance delta is ~10%. This difference decreases at higher core counts, decreasing to 4% at 640 cores (32 nodes).
2. For the versions used in this study, Platform MPI is better than Intel MPI by 0-4%. The difference between the MPIs shows up only after 8 nodes. Up to 8 nodes, the performance of both MPI implementations is similar.

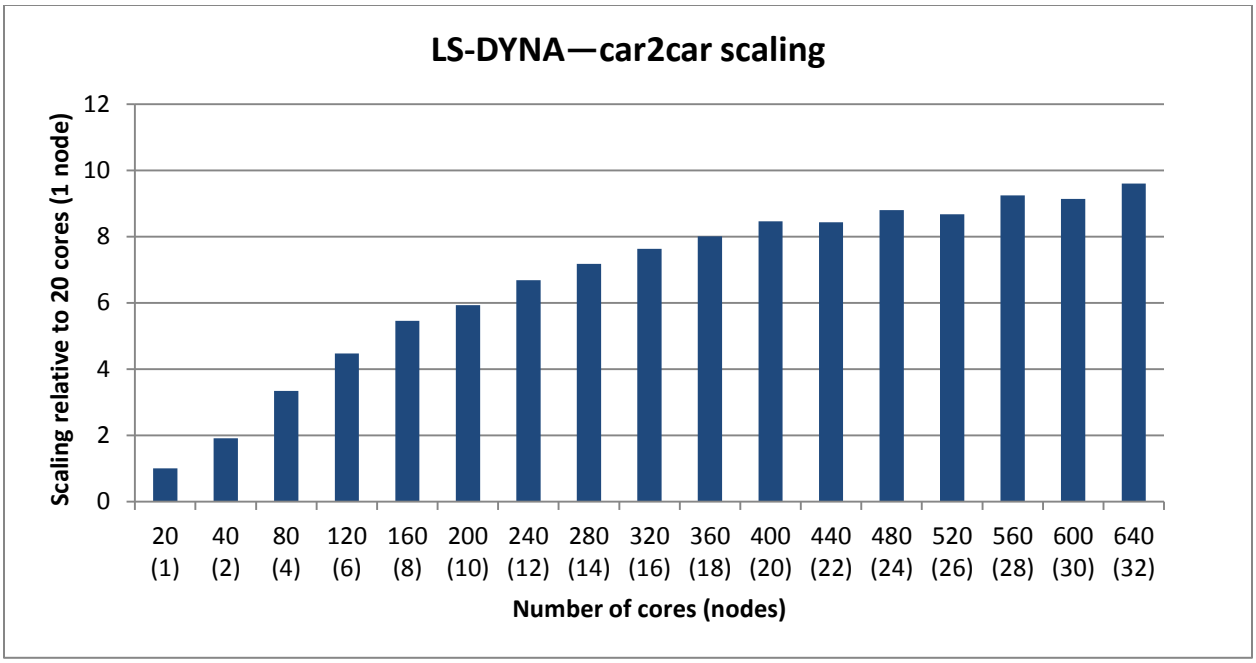

Figure 19 LS-DYNA car2car relative scaling

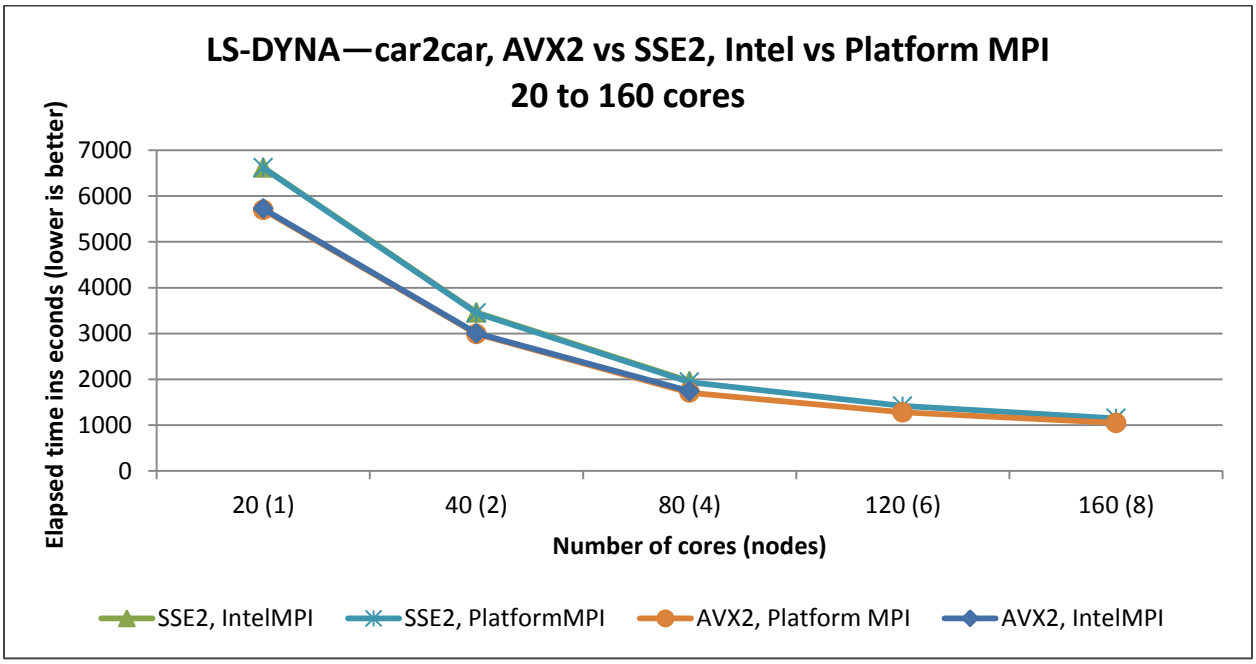

Figure 20 LS-DYNA car2car performance–AVX2, SSE2 and Intel MPI, Platform MPI 20-160 cores

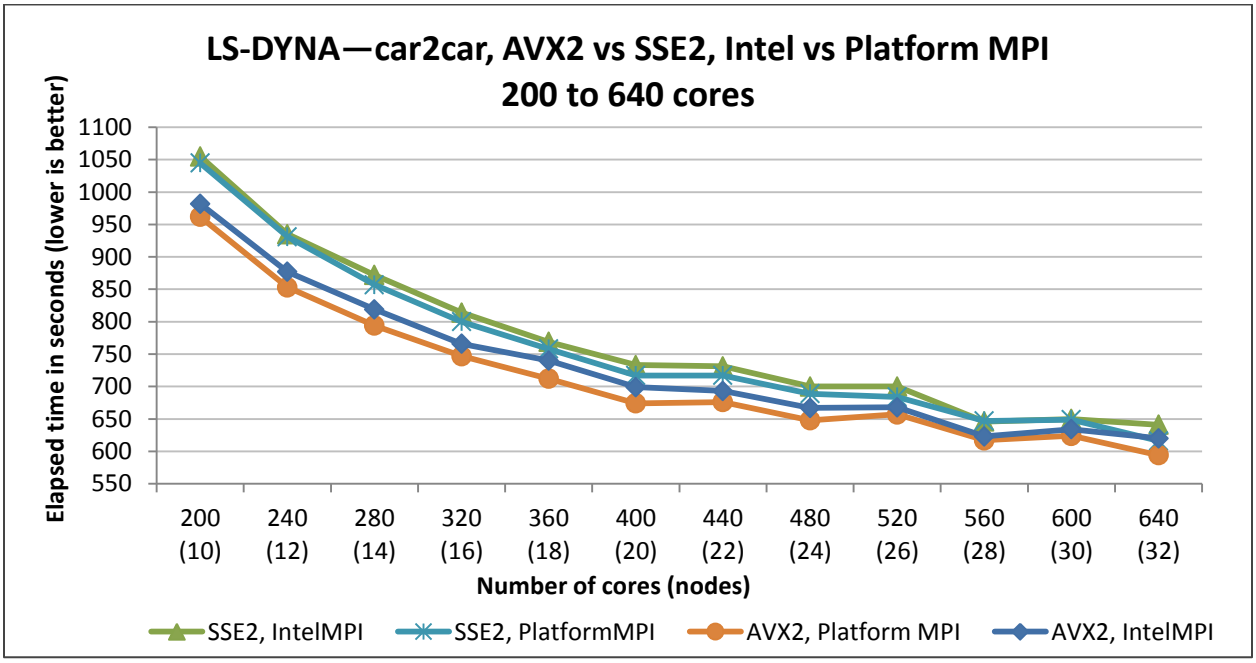

Figure 21 LS-DYNA car2car performance–AVX2, SSE2 and Intel MPI, Platform MPI 200-640 cores

Next, two different LS-DYNA versions, R7.1.2 and R8.0.0, are compared. These results are shown in Figure 22 and plot performance for the car2car benchmark dataset on 640 cores (32 nodes). The R7.1.2 code is known to perform better and that was measured as well. Details of the performance difference relative to R712.SSE2.PMPI are noted in the data labels in the figure. The performance difference between Intel MPI and Platform MPI for R8 of the code as described above (Figure 20 and Figure 21) was observed for R7 as well. The AVX2 versions of the R7.1.2 binaries produced early termination errors with this dataset, so results are not reported for these binaries.

LS-DYNA creates various files during a simulation. It allows the user to specify a global location for shared files, and a local location for files local to each process. On the lab test system, the default choice for local files is on the local disks on each compute node. Another option is to place these files in a NFS exported directory on the NSS storage system. Recall that each compute node has local storage with a six drive RAID 0 virtual disk. The NSS is built on a 60 drive storage enclosure that is configured with RAID 6 and LVM. The next test compares the performance difference of the two local file configuration options and results are presented in Figure 23 and Figure 24. "local on /local" indicates LS-DYNA was configured to use each server's local disks for the local files. "local on NSS" indicates LS-DYNA was configured to use a shared directory on NSS for all local files.

For lower core counts there is no difference between the local drive and the NSS configuration. At higher core counts NSS is 0.5% to 2.8% faster than the local drives.

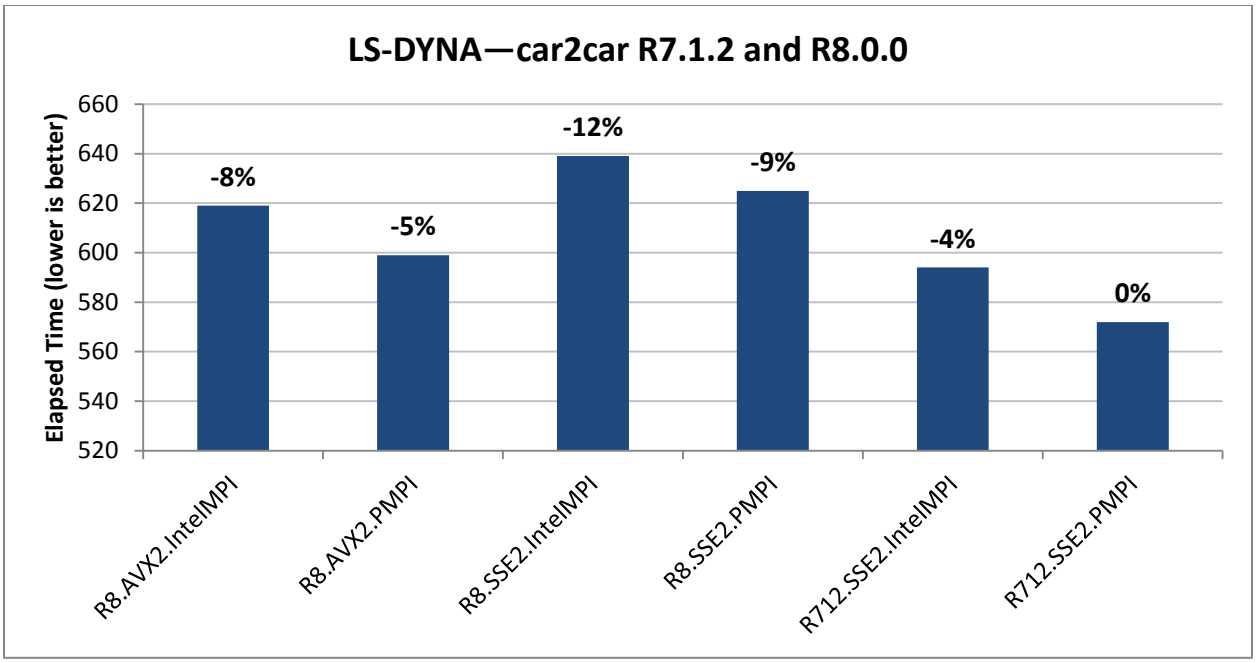

Figure 22 LS-DYNA car2car performance across different versions

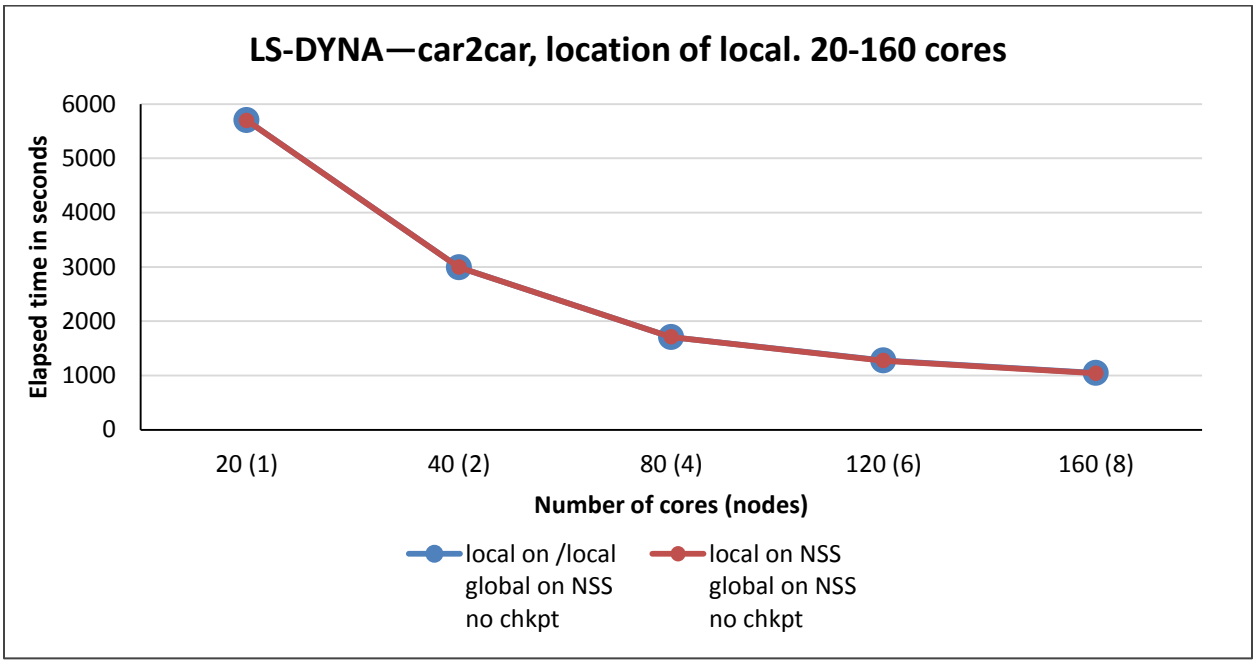

Figure 23 LS-DYNA car2car performance–location of local, 20-160 cores

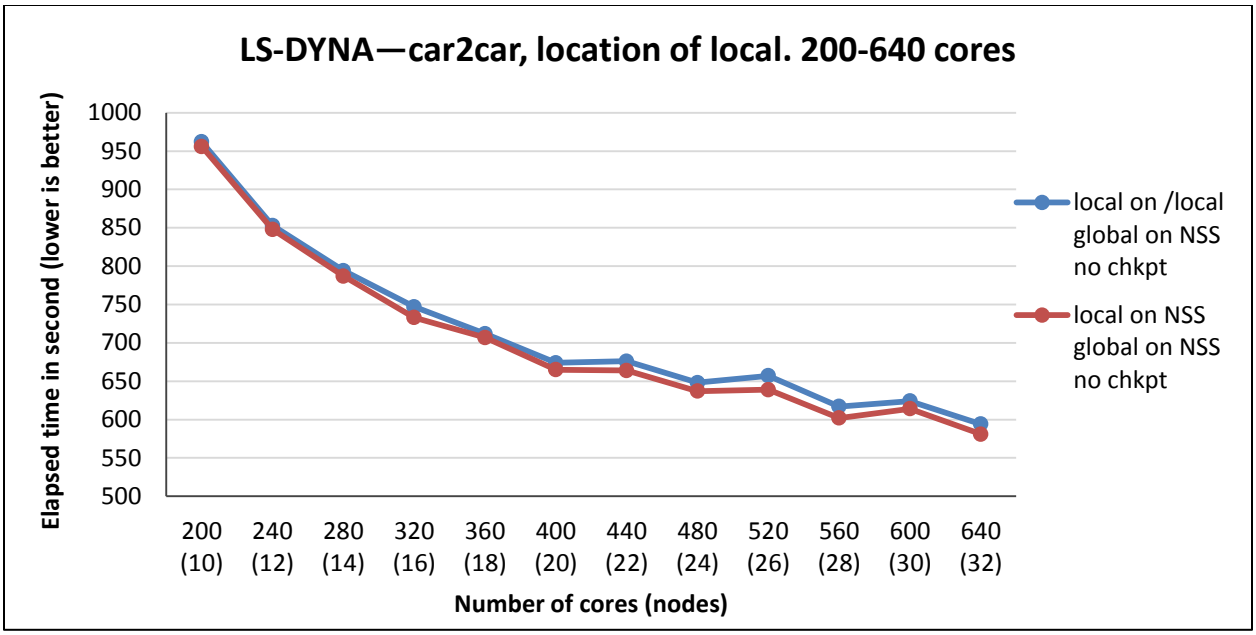

Figure 24 LS-DYNA car2car performance–location of local, 200-640 cores

The last set of LS-DYNA tests looked into check-pointing. LS-DYNA was configured to create restart files every 500 cycles which translated to 80 check-points for the car2car endtime=0.02 test case. This is a high frequency of check-pointing and this test was used to stress the storage systems. Check-pointing adds an additional step to the simulation and the test ran slower for the check-point case as expected. Absolute performance is presented in Figure 25. The graph also compares two scenarios – where the restart files are written to the local disks on each server and where they're written out to the NSS. The results show that check-pointing to the NSS is faster by up to 70% at lower core counts. Recall that without check-pointing NSS was only 0.5% to 2.8% faster than the local drives (Figure 24); the significant performance benefit with NSS under check-pointing I/O load speaks to the value and robustness of the shared storage component in the lab test system.

Similar to the results for ANSYS Fluent, the LS-DYNA results on the current lab test system presented in this white paper are compared to the corresponding results on a 4-node system interconnected using 10 Gigabit Ethernet. The processor and memory configuration of the two configurations is similar, making this a valid comparison. This comparison is shown in Figure 26. As with ANSYS Fluent, the benefit of a faster interconnect like InfiniBand over 10 Gigabit Ethernet is apparent at three or four nodes making InfiniBand necessary on larger systems to make effective use of the greater number of servers and cores.

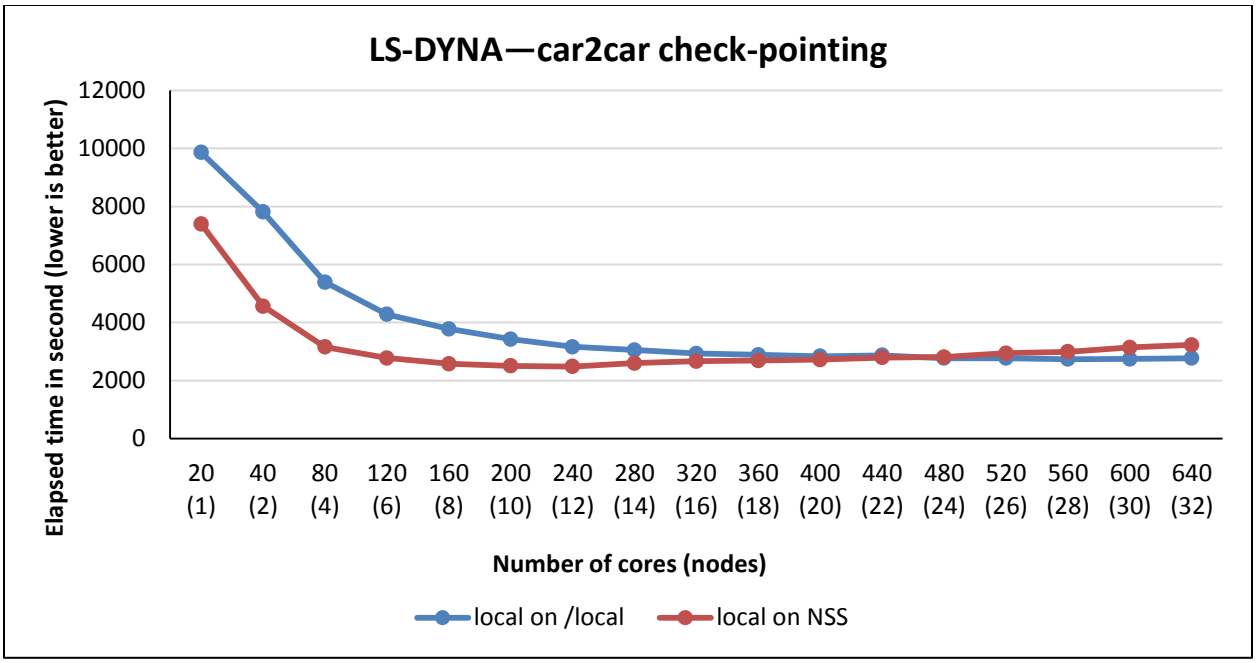

Figure 25 LS-DYNA car2car check-point performance

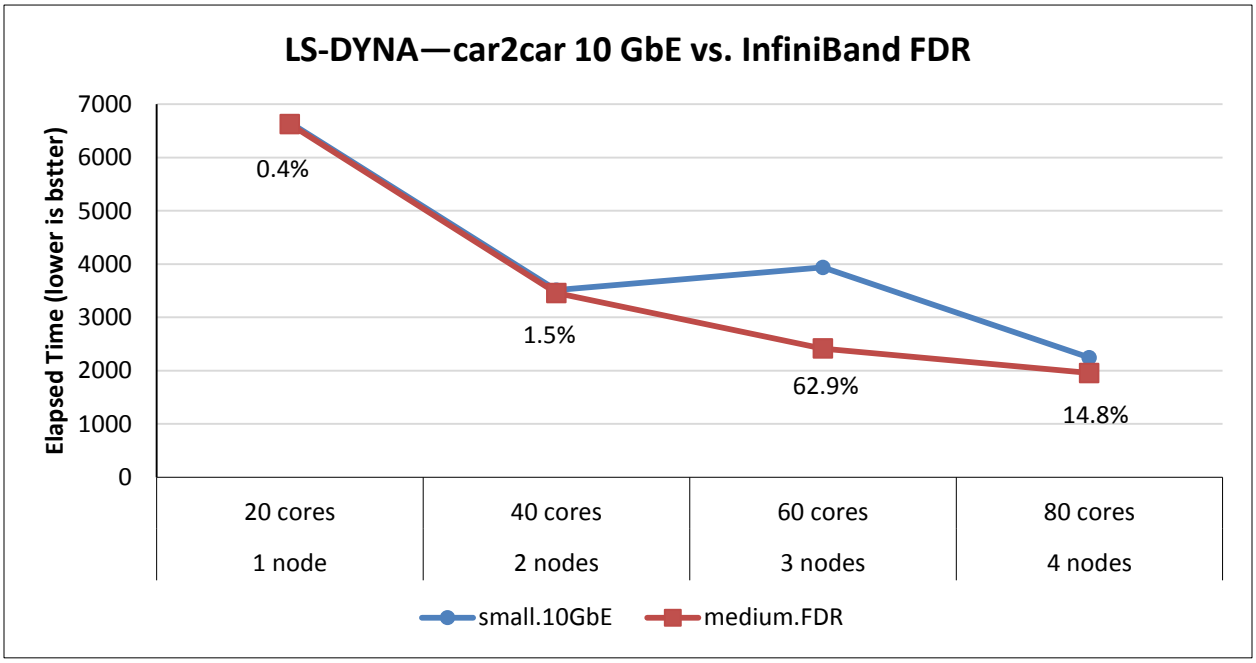

Figure 26 LS-DYNA car2car performance comparison— 10 GbE vs. InfiniBand FDR

4.4.2 LS-DYNA ODB-10M

At 2.4 million elements, the car2car-ver10 dataset is relatively small by current automotive industry standards, and as shown in the previous section, this dataset doesn't scale well beyond about 280 cores. Because of this, the ODB-10M-ver14 dataset is also used for additional performance tests.

The ODB-10M-ver14 dataset is a 10 million element LS-DYNA model, configured with a simulation time of 0.120 s. In order to reduce the runtime for benchmark analysis, "endtime=0.02" was used. This allows the evaluation of many more test combinations which are still representative of the performance of the unmodified dataset.

Figure 27 plots the performance of the ODB-10M dataset on the lab test system. The metric for performance is Elapsed Time, with a lower value representing better/faster performance. The graph shows the performance of two R8 LS-DYNA binaries, AVX2 and SSE2, with Platform MPI. Similar trends regarding performance of the SSE2 vs AVX2 binaries are seen with this dataset as with the car2car dataset.

Figure 28 plots the scaling of the ODB-10M and car2car datasets relative to 20 cores or 1 node. The data shows that the larger ODB-10M model continues to scale up to 640 cores or 32 nodes.

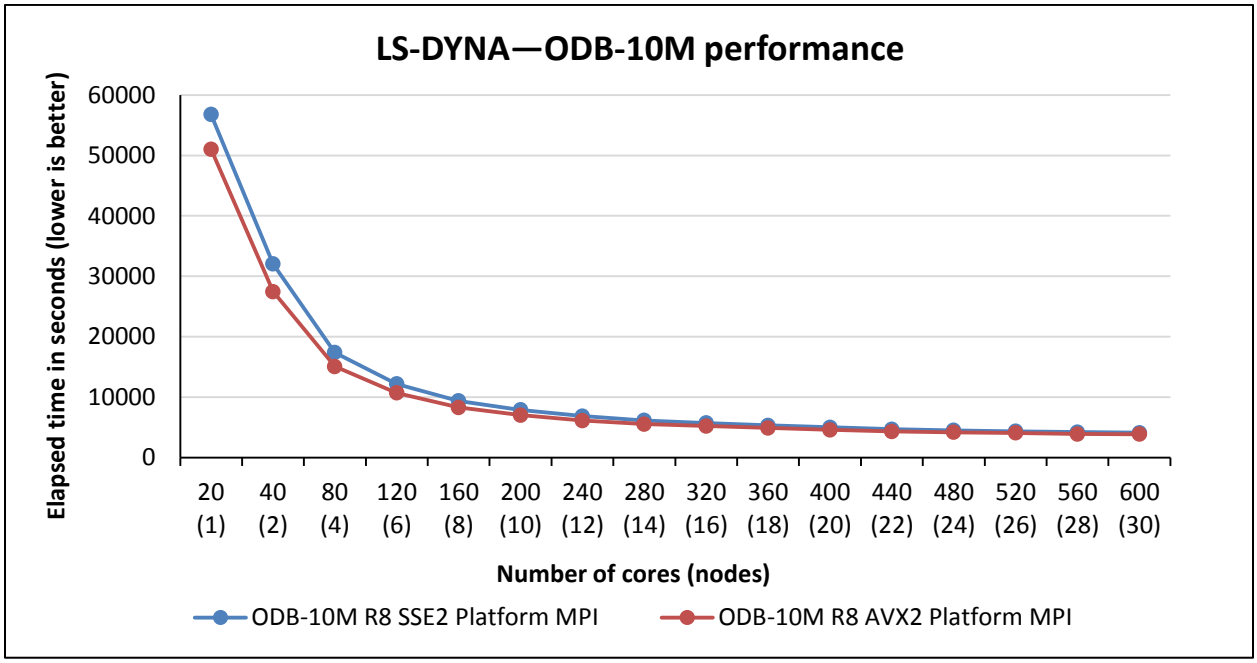

Figure 27 LS-DYNA ODB-10M performance

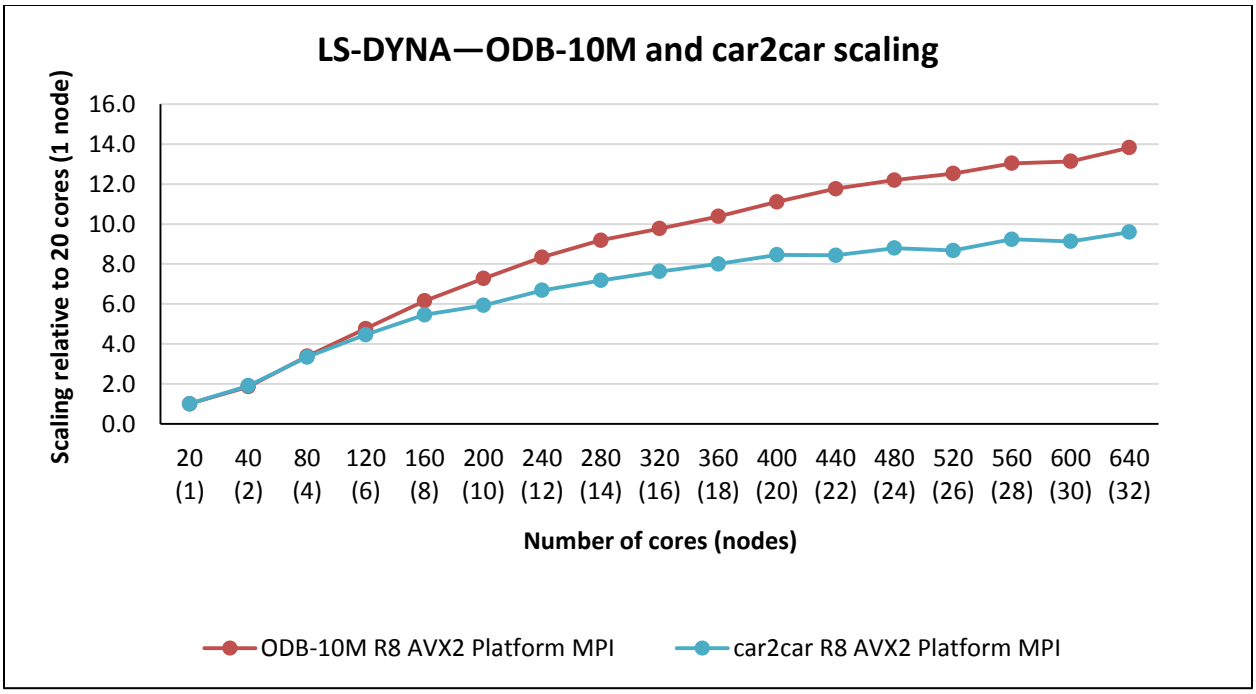

Figure 28 LS-DYNA ODB-10M and car2car relative scaling

4.5 STAR-CCM+

Nine STAR-CCM+ benchmark test cases are presented in this study: civil_trim_20m, EglinStoreSeparation, HlMach10Sou, KcsWithPhysics, LeMans_100M, lemans_poly_17m, reactor_9_million, TurboCharger and VtmUhoodFanHeatx68m.

Results for these STAR-CCM+ test cases on the lab test system are plotted in Figure 29 and Figure 30. The performance metric used here is AverageElapsedTime as reported by the STAR-CCM+ benchmark output, and lower is better since this is a time based metric. From the graphs it is easy to see that the benchmark data sets scale well for lower core counts, i.e. performance improves as more cores are provided for the test. Due to the scale of the graph it is hard to analyze the performance at larger number of cores. Those patterns are easier to see in Figure 31 and Figure 32 which plot the same data but relative to the "20 cores (1 node)" data point.

The datasets presented in Figure 31 scale almost linearly—as more cores are added, performance improves. The three datasets presented in Figure 32 show different scaling patterns—Turbo is known not to scale due to the type of simulation which is a conjugate heat transfer analysis, the number of continua in the model, and other aspects that make the case more complex and less scalable. Eglin and Kcs are small models with 4.6M cells and 2.9M cells respectively and are only expected to scale to smaller core counts due to the small model size. It is noteworthy that the performance of these three cases doesn't decrease at high core counts; the performance either plateaus or continues to increase up to 640 cores.

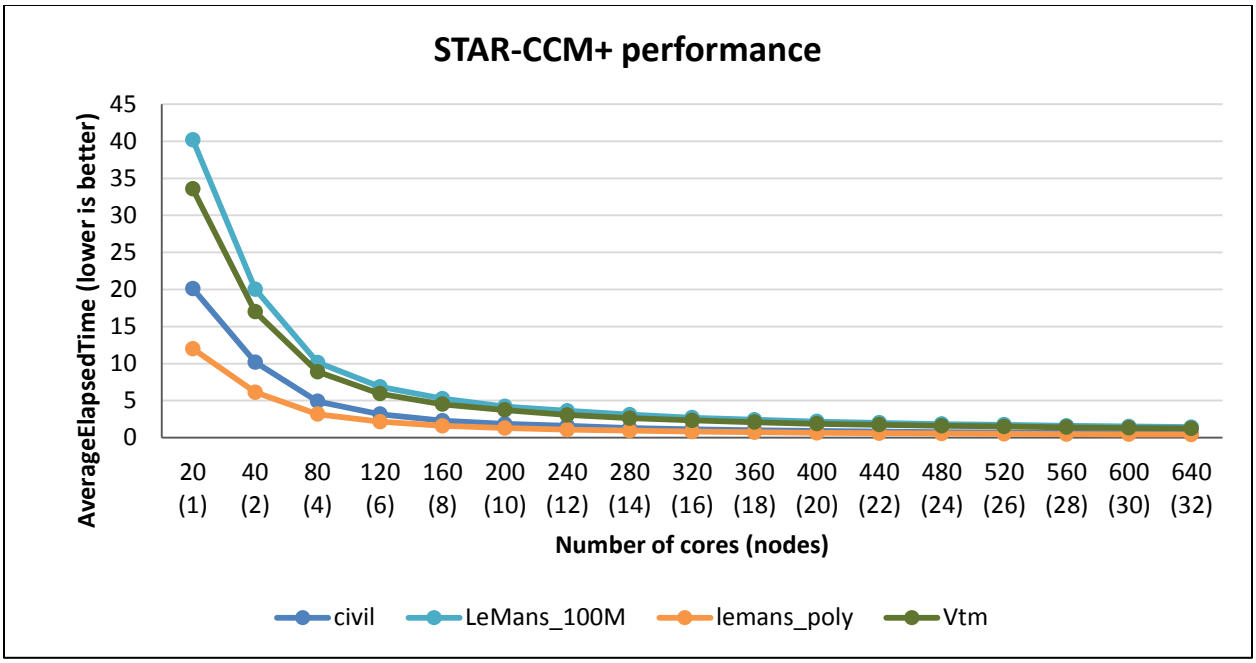

Figure 29 STAR-CCM+ performance (civil, LeMans_100M, lemans_poly, Vtm)

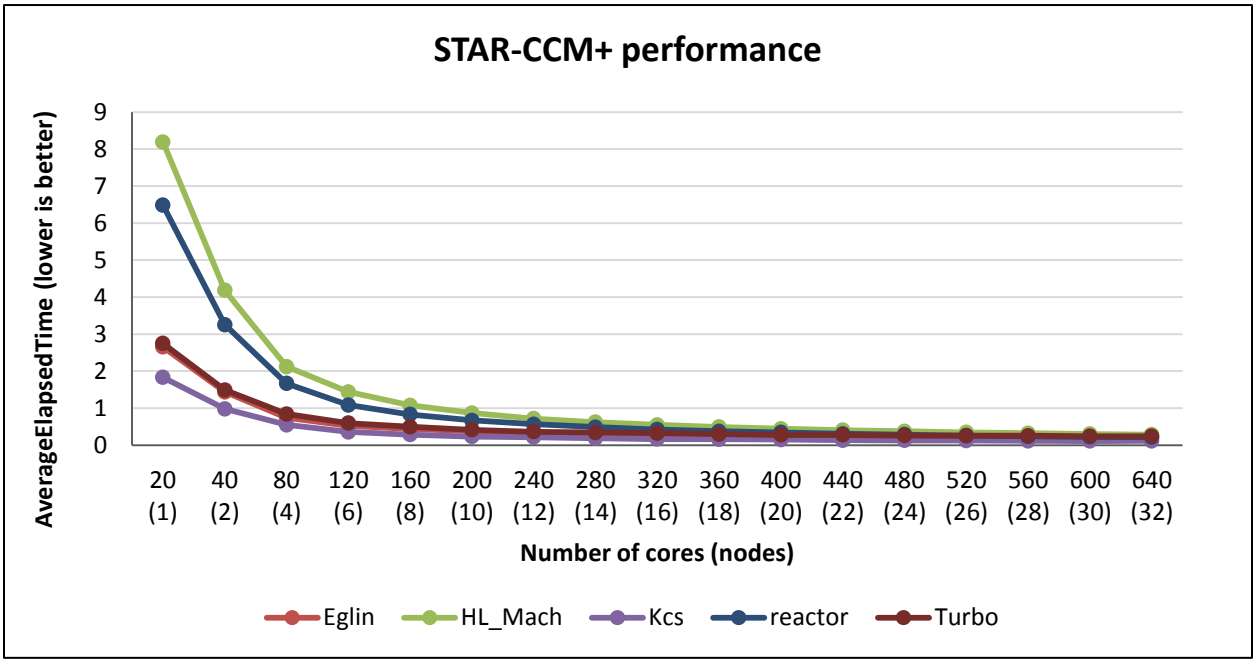

Figure 30 STAR-CCM+ performance (Eglin, HL_Mach, Kcs, reactor, Turbo)

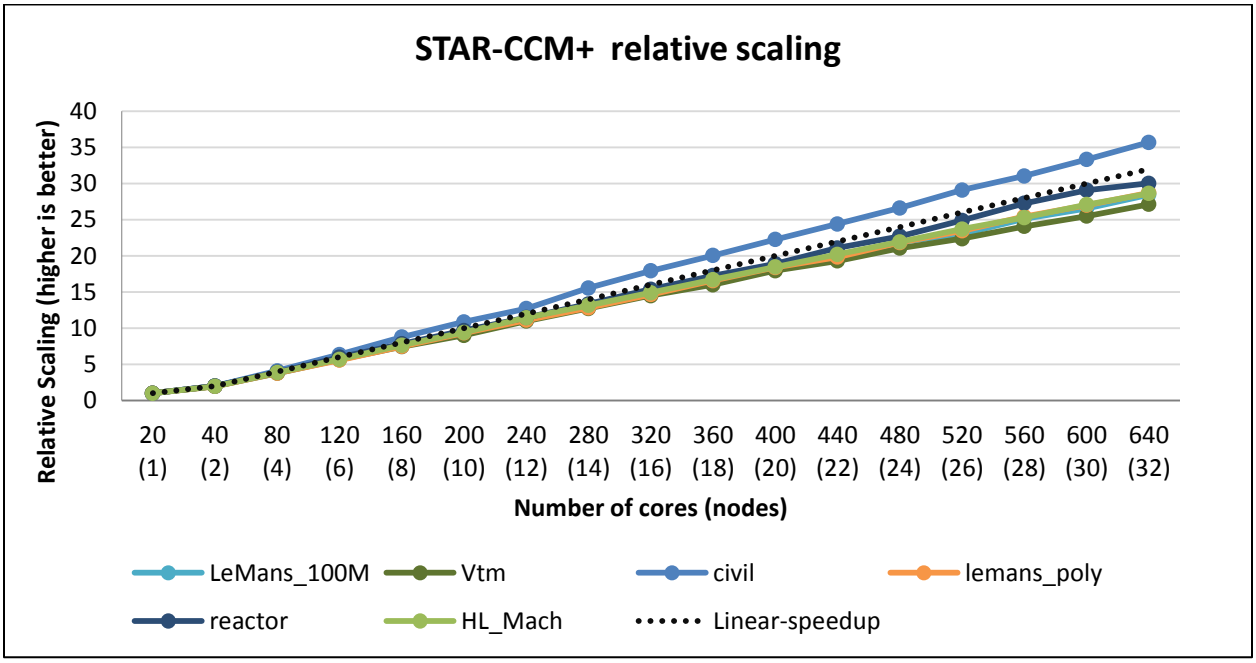

Figure 31 STAR-CCM+ relative scaling (LeMans_100M, Vtm, civil, lemans_poly, reactor, HL_Mach)

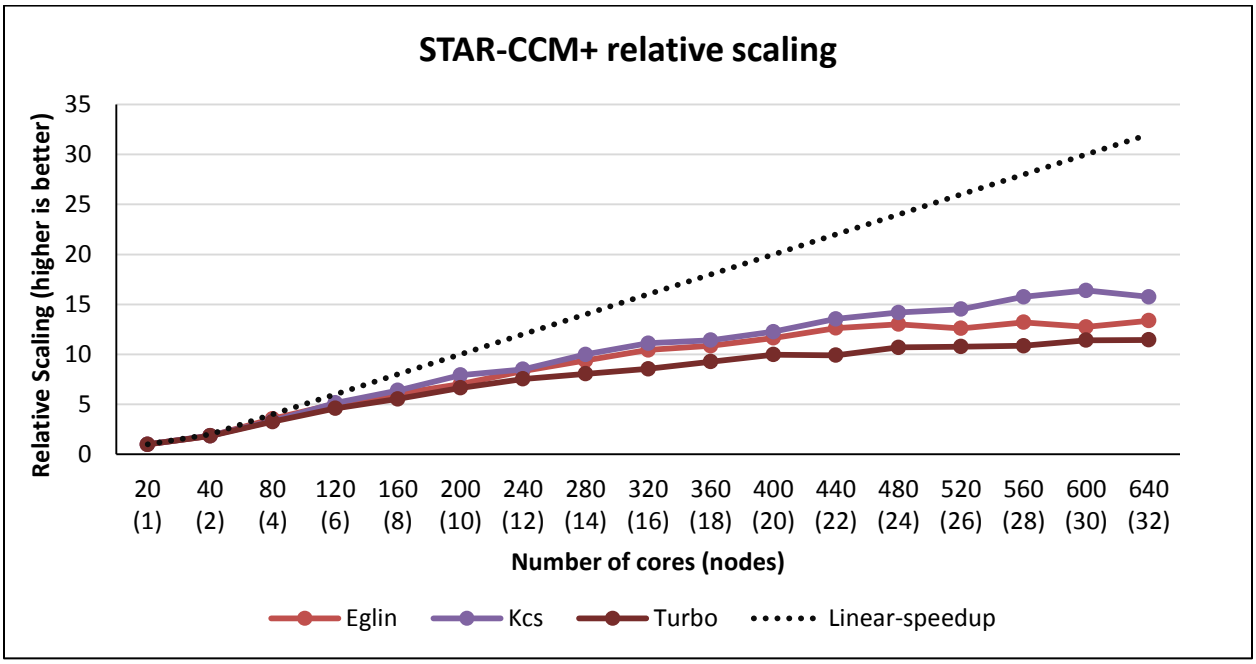

Figure 32 STAR-CCM+ relative scaling (Eglin, Kcs, Turbo)

Similar to the other applications, the STAR-CCM+ results on the lab test system presented here are compared to the corresponding 4-node results on 10 Gigabit Ethernet. The processor and memory configuration of the two configurations is similar, making this a valid comparison. This comparison is shown in Figure 26. Similar to ANSYS Fluent and LS-DYNA, the benefit of a faster interconnect like InfiniBand over 10 Gigabit Ethernet is apparent at four nodes making InfiniBand necessary on the larger system to make effective use of the greater number of servers and cores, and demonstrating that 10 Gigabit Ethernet is sufficient for the smaller 4-node system.

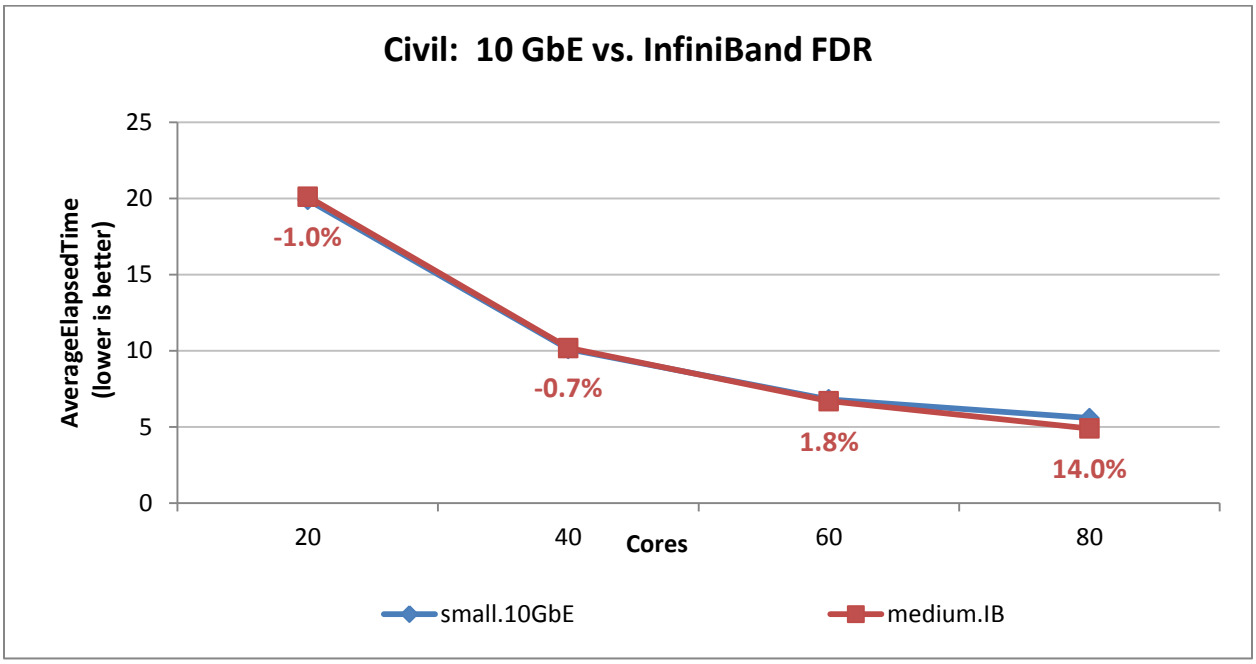

Figure 33 STAR-CCM+ performance comparison (civil)

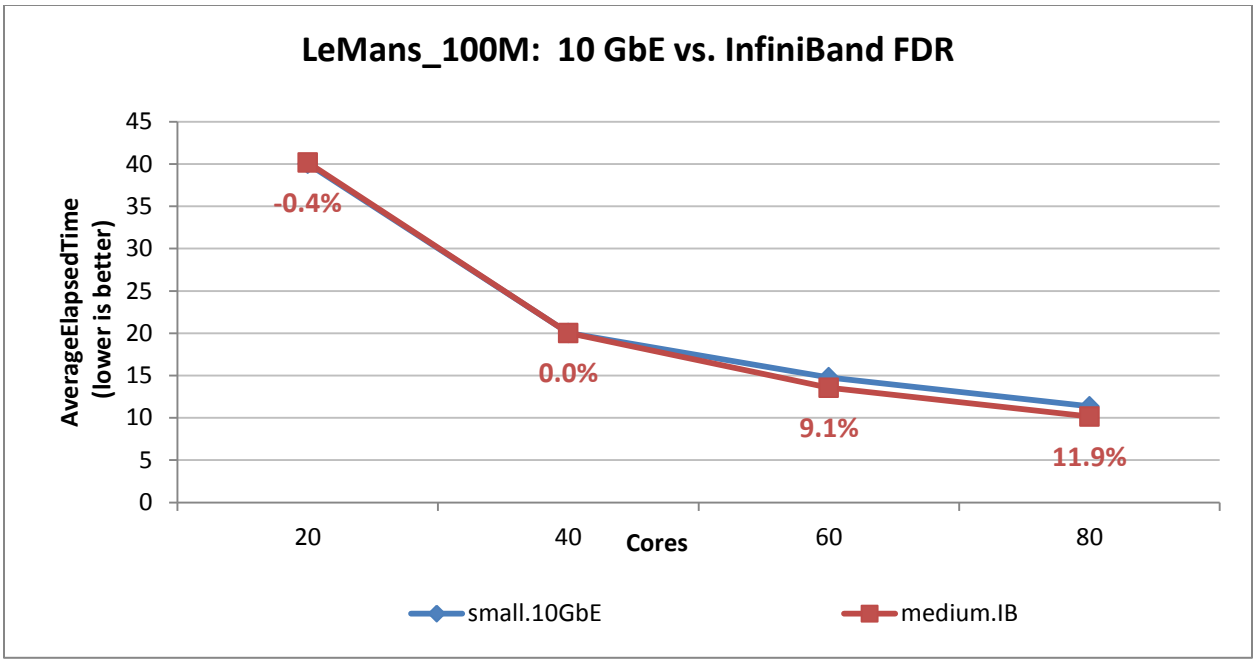

Figure 34 STAR-CCM+ performance comparison (LeMans_100M)

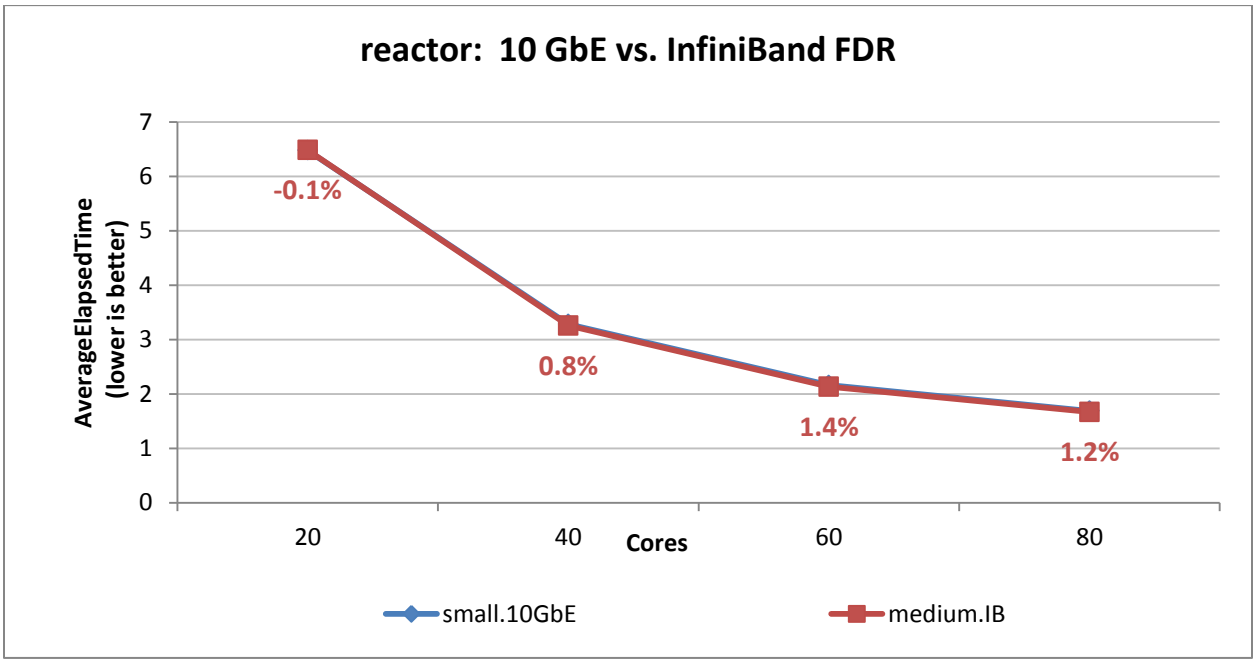

Figure 35 STAR-CCM+ performance comparison (reactor)

5 Remote Visualization

A PowerEdge R730 remote visualization node was included in the lab test system and configured as previously described in Table 4. In order to evaluate the remote visualization system, NICE EnginFrame and Desktop Cloud Visualization (DCV) were installed on the lab test system.

NICE is a software company that provides remote visualization software and a Grid Portal for managing remote visualization sessions, HPC job submission, job control and monitoring. NICE EnginFrame is the Grid Portal component. For this evaluation, EnginFrame 2015.0 r37468 was installed on the cluster master node. NICE DCV enables remote access to 2D and 3D applications over a standard network, providing remote GPU acceleration for 3D applications. DCV 2014.0 r16231 was installed on the remote visualization node.

With the NICE remote visualization solution, EnginFrame primarily provides management of remote visualization sessions and has no impact on the performance of the DCV component. For this evaluation, EnginFrame was tested to verify correct operation and successful integration with the overall system. It was also used to manage the remote desktop sessions on the DCV/remote visualization server. A screen capture of the EnginFrame VIEWS portal, showing an active Linux Desktop session, is shown in Figure 36.

Various applications and datasets were used to verify the operation of DCV, as listed in Table 9. This evaluation primarily focused on stability and correct operation of the NICE solution and a qualitative evaluation of the interactive application performance in both LAN and WAN environments. Screen captures showing several of the applications and datasets used for the evaluation are shown in Figure 37 through Figure 39

Table 9 DCV Evaluation Software

Software	Version	Datasets
LSTC LS-PrePost	4.2_centos6	car2car-ver10 ODB10M-ver14
ANSYS Fluent	v16.0.0	JetDemo small-indy
Blender	2.75a	BMW27 BMW27GE "For You"
glmark2	9/29/15	OpenGL ES 2.0
BETA CAE mETA Post	v16.0.0	OpenFOAM 2.4.0 motorBike

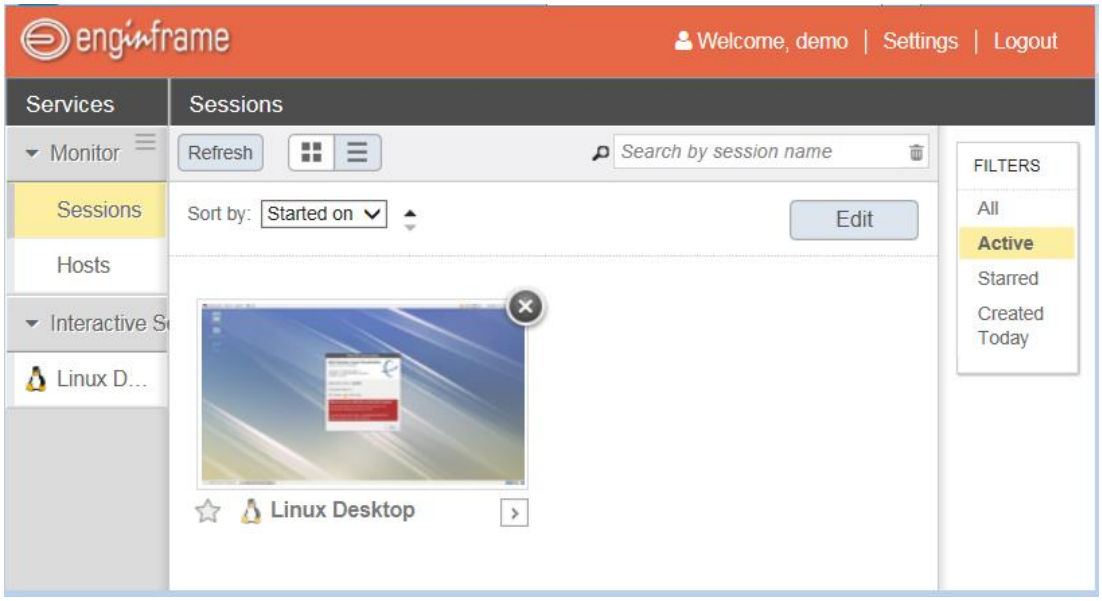

Figure 36 NICE EnginFrame VIEWS Portal

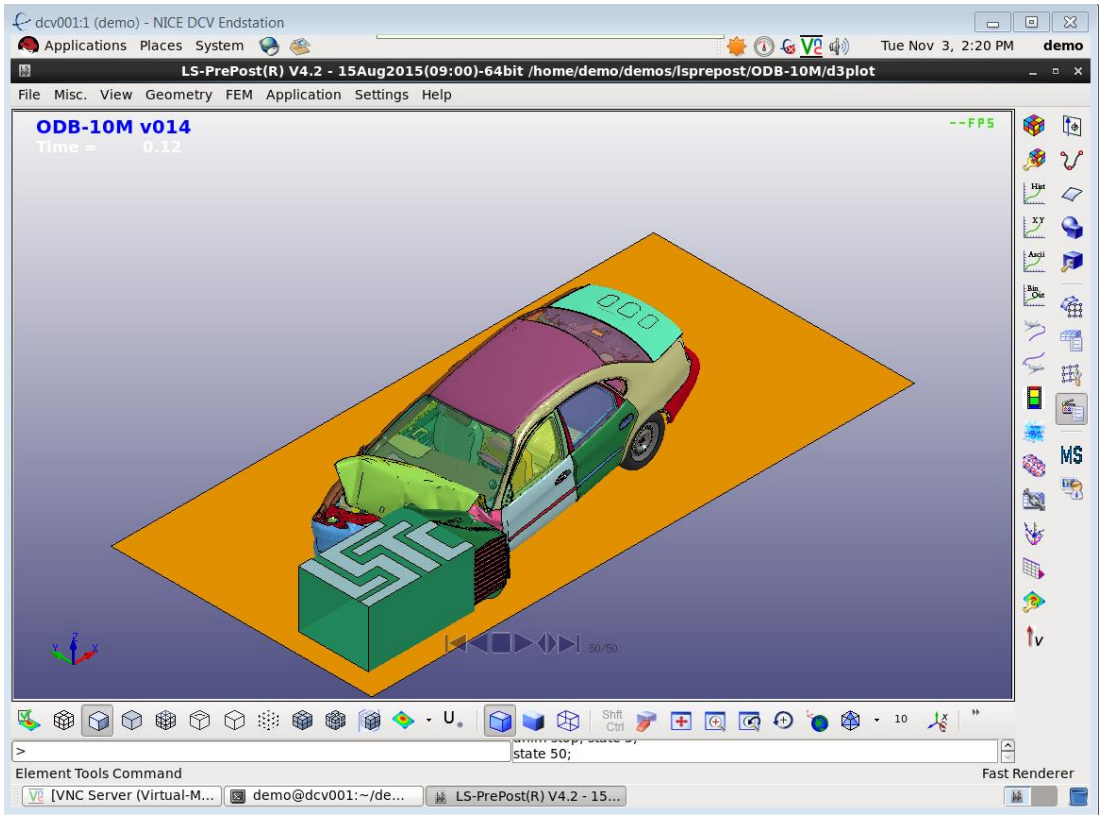

Figure 37 LS-PrePost 4.2 with ODB-10M

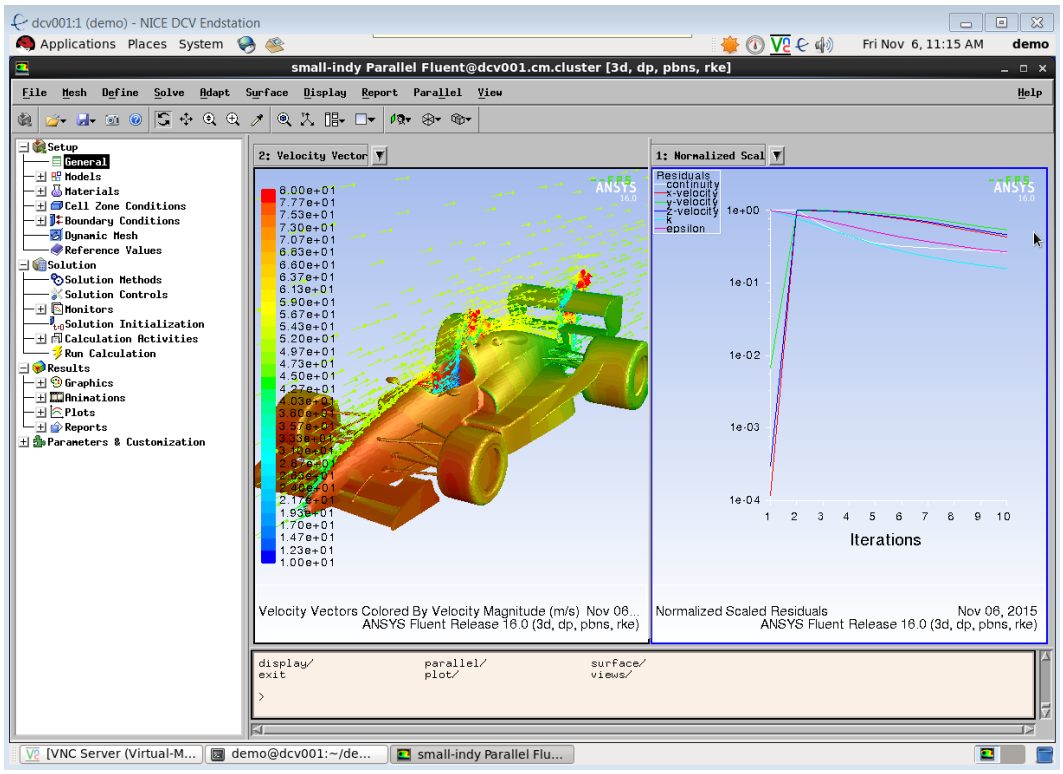

Figure 38 Fluent v16.0 with small-indy

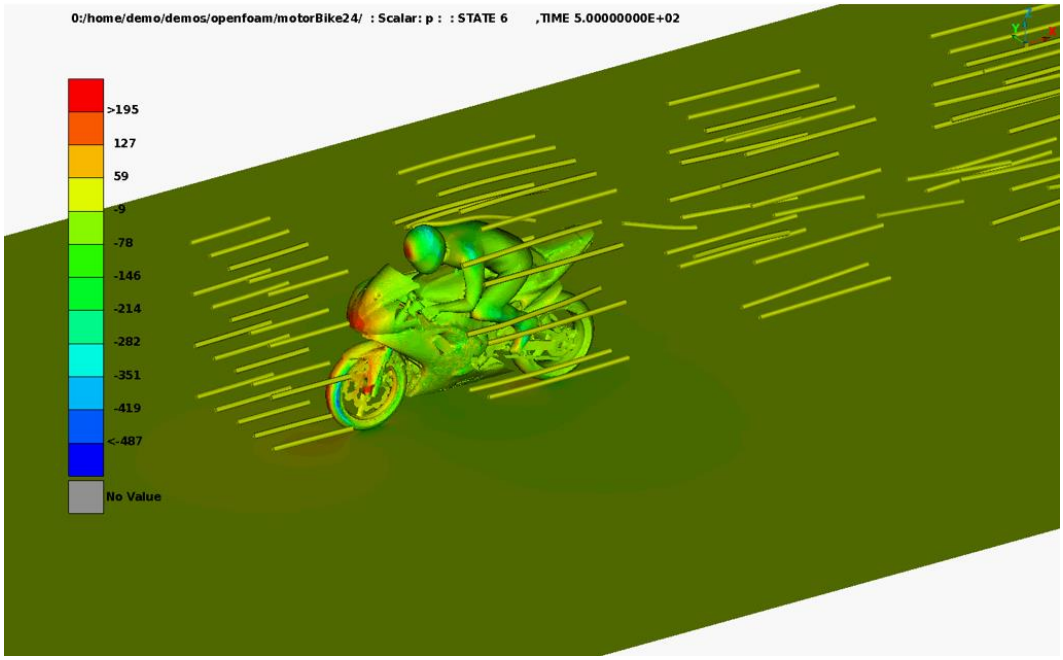

Figure 39 mETA Post with motorBike

One of the features of the NICE DCV Endstation client is the Endstation console, shown in Figure 40. The console allows the end user to dynamically adjust quality vs network bandwidth utilization using a slider bar and to monitor the bandwidth being used by the client. For most uses, the 50% setting provides a good balance between bandwidth usage and image quality. An aspect to note about the NICE DCV solution is that the final image delivered to the client after display updates have stopped is always lossless, regardless of the quality level setting. This ensures that static images are always shown with full quality on the client.

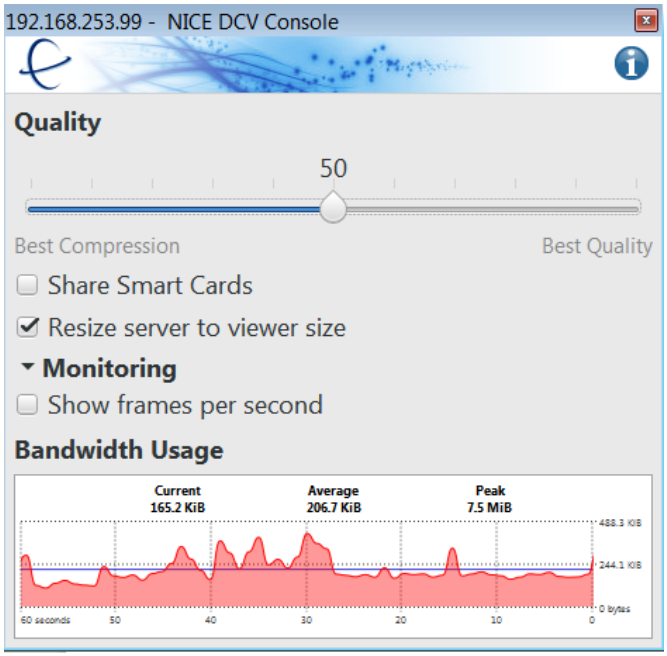

Figure 40 NICE DCV Endstation Console

For testing, the 50% quality setting was used for the client. In a LAN setting, with significant bandwidth and low latency, the remote application responsiveness and rendering performance was very good. In a WAN environment, application responsiveness and rendering performance was also very good as long as network latency remained less than about 150 ms and sufficient network bandwidth was available. When network latency exceeded about 150 ms, lags in the application response became noticeable. This is expected behavior and NICE recommends changing some of the DCV configuration parameters for use in high latency network environments; however, since these changes increase response time for low latency networks they are not recommended for most usage scenarios.

For typical applications at the 50% quality level, average network bandwidth utilization ranged from 150 to 600 KiB/s during display updates and dropped to 0 kb/s for static images. Peak network bandwidth was approximately 1.0 MiB/s for all of the tested applications at the 50% quality setting. At higher quality settings, average network bandwidth gradually increased, with a significant increase in bandwidth utilization from the 90% to the Lossless setting. At 90%, network bandwidth averages about 1.0 MiB/s compared with 5 to 8 MiB/s at the lossless setting.

Overall, the NICE DCV solution performed well and offers a good solution for remote visualization users.

6 Power consumption results

Power requirements and power budgeting is an important consideration when installing any new equipment. This section reports the power consumed by the lab test system for the different applications described in Section 0. This data was obtained by using metered rack power distribution units (PDU) and recording the actual power consumption of the full rack system during test.

For each application and benchmark data set, the test was run using all the cores and servers in the system, i.e. 640 cores (32 nodes). This was to provide an accurate assessment of power requirements when the whole system is in use.

For each application, two BIOS profiles were evaluated and power consumption and application performance measured for each profile. The first is the DAPC Profile. In this setting power management is controlled by the server hardware, Turbo mode is enabled, and C-states and C1E is enabled. This is a performance per watt optimized profile balancing performance needs with power consumption for an energy efficient configuration. The second test is with the Performance Profile. In this test, power management is set to max performance, Turbo mode is enabled, and C-states and C1E are disabled. This profile maximizes performance. Average power consumption data and peak power consumption was measured for both profiles. Average power is the average steady state power consumption during the steady state portion of the test. Peak power is the instantaneous maximum power recorded during the test at any time from start to finish. These values are marked as "Average-DAPC", "Peak-DAPC", "Average-Perf" and "Peak-Perf" in the results below.

Since the Performance Profile attempts to favor system performance while the DAPC Profile attempts to balance performance with energy efficiency, it is expected that the system will consume more power when in Performance Profile mode while possibly providing better performance. This is quantified on the graphs below—the "perf-advant." value recorded in the bubble on the graph calculates the application performance advantage of using the Performance Profile over the DAPC Profile; and the data label noted in "Average-Perf" shows the average additional power consumed by the system when in Performance Profile as compared to the DAPC Profile.

Figure 41 plots the idle power consumption of the system. This is the power draw when there are no jobs running and no activity on the system. This was recorded as 4430W in DAPC Profile. The system consumes 67% more power at 7384 W when in Performance Profile. This measurement includes the power draw of all 32 compute nodes, the remote visualization node, the master node, the NSS-HA, the KMM and all switches. As expected, there is no difference between average and peak power when the system is idle.

Figure 41 also plots the power consumption of the system when running HPL. The peak power draw during HPL is likely to be the maximum power draw of the system under load. Most applications will not stress the system as much as HPL and will not consume as much power as HPL. This is also evident from the subsequent graphs in this section. With HPL, the Performance Profile consumes 5% more power on average and provides 1% better performance. Note the peak power consumption of the system is almost identical irrespective of the profile and that is ~14.1kW for the rack.

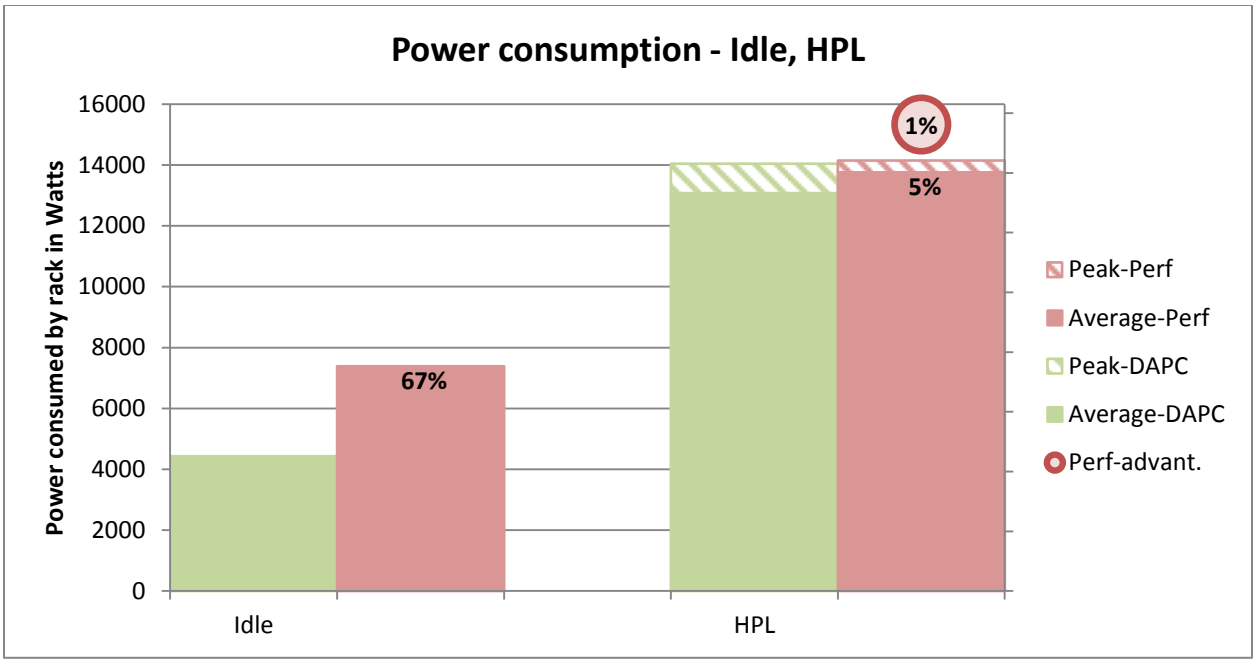

Figure 41 Power consumption – idle, HPL

Figure 42 and Figure 43 plot the power consumption of the system when running the ANSYS Fluent and STAR-CCM+ benchmark datasets respectively. The maximum power consumed for these cases is under 13kW.

In most ANSYS Fluent cases, the power consumed with the Performance Profile is more than the associated performance improvement. For example, sedan_4m consumed 10% more power for 8% better performance. With aircraft_wing_14m, combustor_71m and exhaust_system_33m, the additional power consumed in Performance Profile is commensurate with the additional performance.

When running STAR-CCM+, the Performance Profile consumes significantly more power without providing a significant performance advantage.

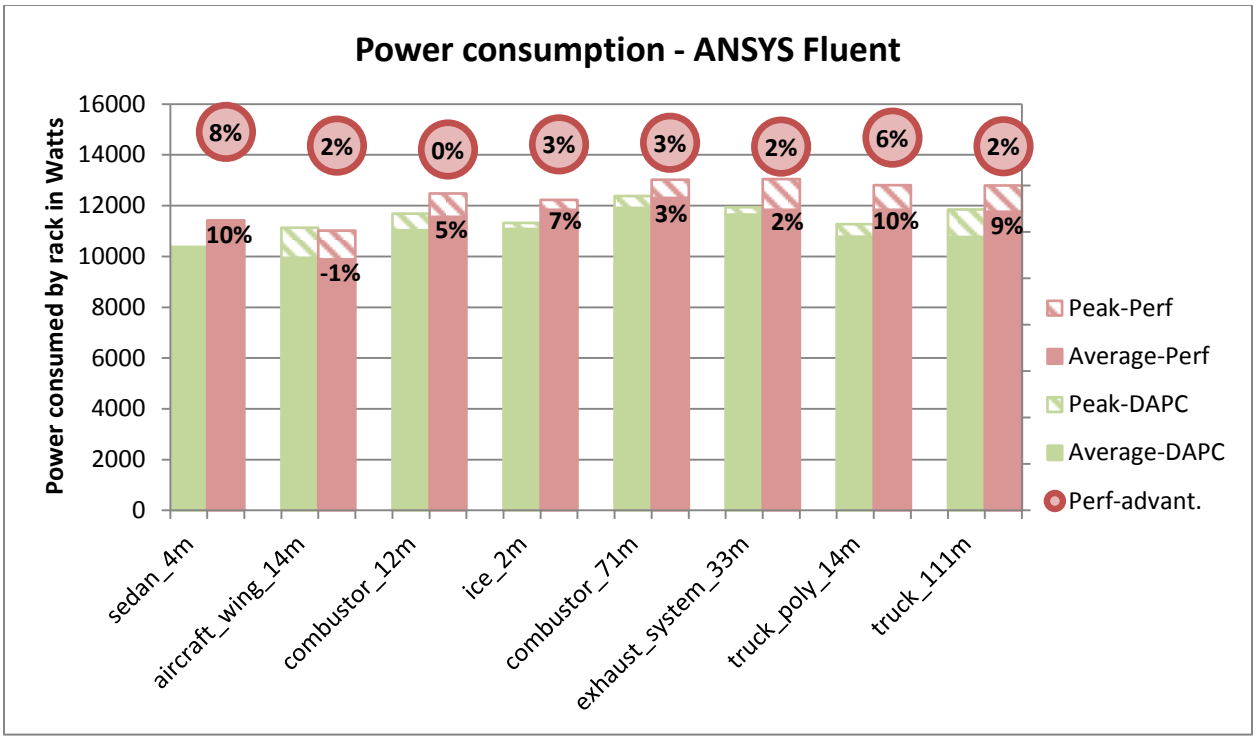

Figure 42 Power consumption – ANSYS Fluent

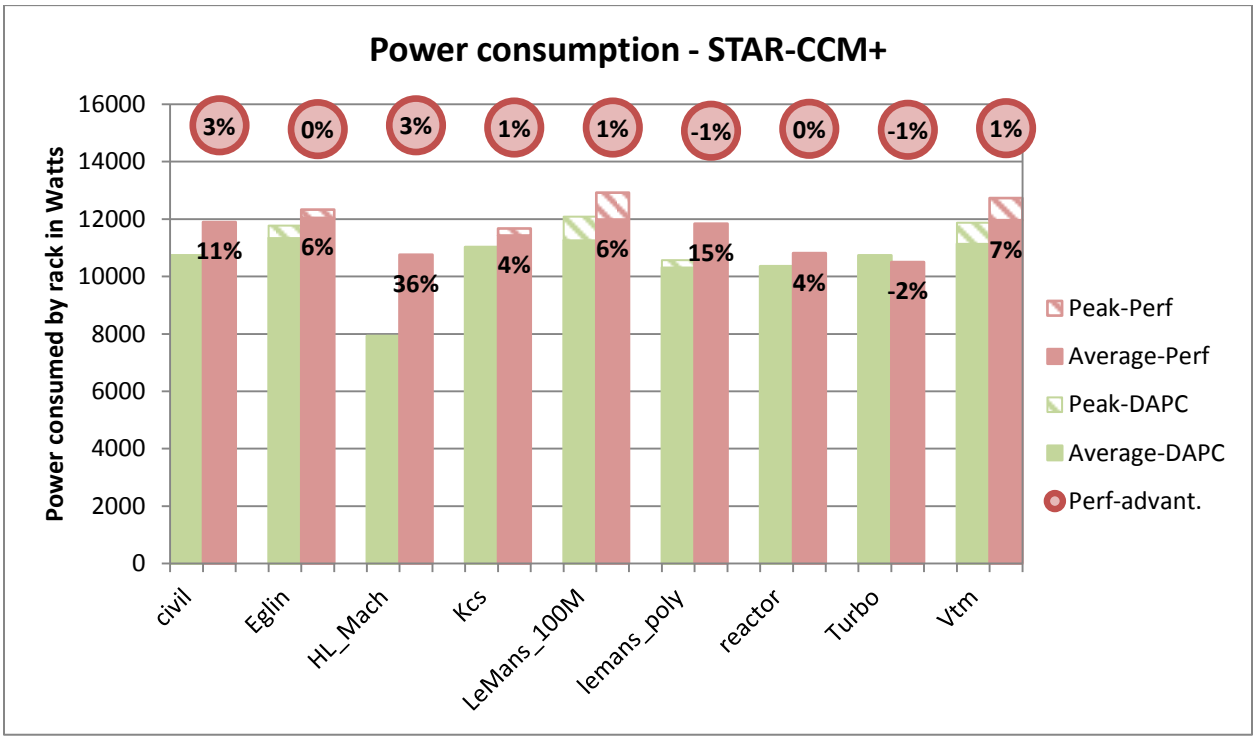

Figure 43 Power consumption – STAR-CCM+

Figure 44 plots the power consumption of the system when running different LS-DYNA tests. For all cases the binary used is the R8 AVX Platform MPI version. The "nochkpt" tests involve no check-pointing and the "chkpt" tests include the creation of restart files every 500 cycles. The "localonlocal" tests write all local files to the compute nodes' local disks, the "localonNSS" tests write all local files to the shared NSS directory.

All four tests show similar performance-power results as seen before—Performance Profile consumes more power for a smaller performance advantage.

For the no check-point tests, the peak power consumed is close to the steady state average consumption for these tests.

The check-point tests show an interesting pattern. There is a lot of I/O during the check-point portion of the simulation and during this period the power consumed by the compute component of the system drops noticeably. This results in lower average power consumption for the duration of the tests. The peak power is similar for all cases, as it is influenced primarily by the computational workload which is the same for all tests.

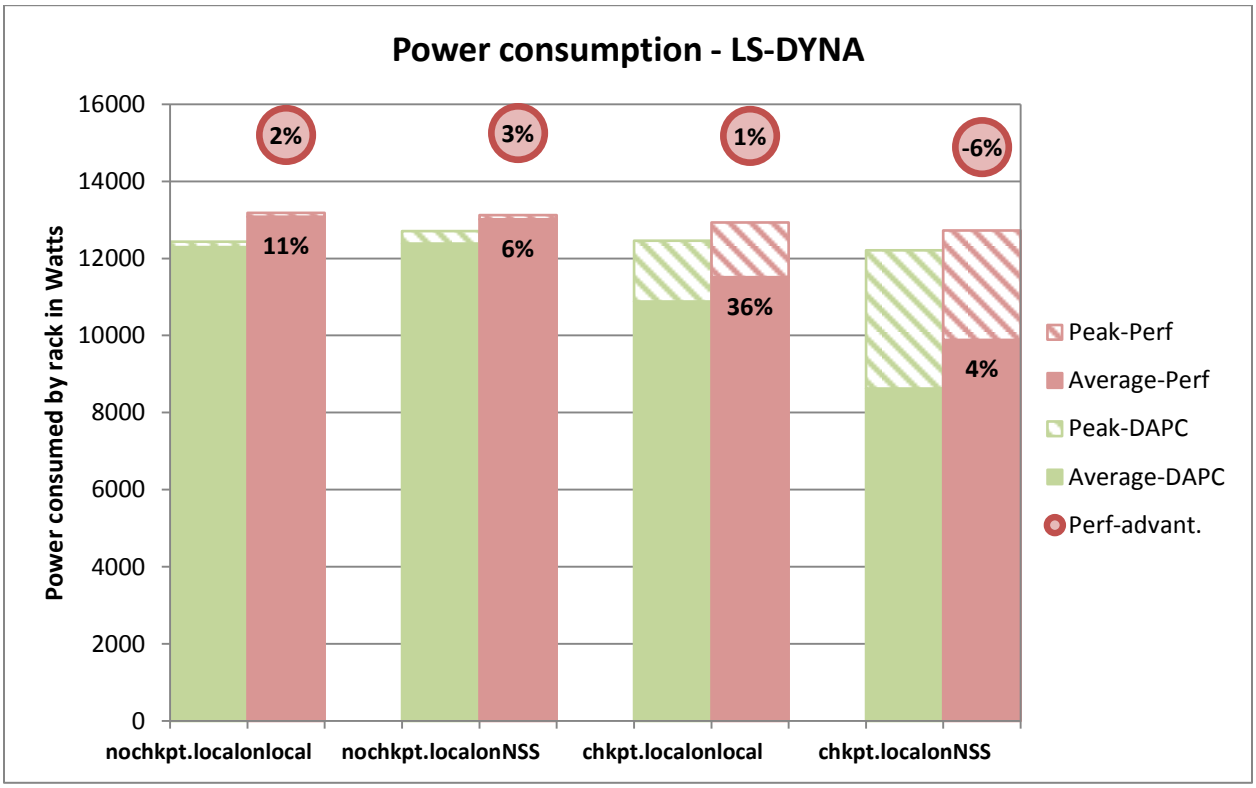

Figure 44 Power consumption – LS-DYNA

7 Conclusion

This technical white paper presents a validated reference design for a single rack HPC system for manufacturing. The detailed analysis of the design options demonstrate that the system is architected for a specific purpose—to provide a comprehensive HPC solution for the manufacturing domain. The design takes into account computation, storage, networking, visualization and software requirements and provides a solution that is easy to install, configure and manage, with installation services and support readily available.

The performance benchmarking bears out the system design, providing actual measured results on the system for specific manufacturing applications. Additionally, system power data is presented to allow for upfront power budgeting for this solution.