DELL EMC

Technical White Paper

Dell EMC Ready Solution for HPC Digital Manufacturing—Dassault Systèmes' Simulia Abaqus Performance

Abstract

This Dell EMC technical white paper discusses performance benchmarking results and analysis for Simulia Abaqus on the Dell EMC Ready Solution for HPC Digital Manufacturing.

June 2019

Revisions

Date	Description
January 2018	Initial release with Intel® Xeon® Scalable processors (code name Skylake)
June 2019	Revised with 2nd Generation Intel Xeon Scalable processors (code name Cascade Lake)

Acknowledgements

This paper was produced by the following:

Authors: Joshua Weage
Martin Feyereisen

Table of contents

Revisions....2
Acknowledgements....2
Table of contents....3
1 Introduction....4
2 System Building Blocks....5
2.1 Infrastructure Servers....5
2.2 Compute Building Blocks....6
2.3 Basic Building Blocks....7
2.4 Storage....8
2.5 System Networks....10
2.6 Cluster Management Software....10
2.7 Services and Support....11
3 Reference System....12
4 Abaqus Performance....14
5 Conclusion....20

1 Introduction

This technical white paper discusses the performance of Dassault Systèmes' Simulia Abaqus on the Dell EMC Ready Solution for HPC Digital Manufacturing. This Dell EMC Ready Solution for HPC was designed and configured specifically for Digital Manufacturing workloads, where Computer Aided Engineering (CAE) applications are critical for virtual product development. The Dell EMC Ready Solution for HPC Digital Manufacturing uses a flexible building block approach to HPC system design, where individual building blocks can be combined to build HPC systems which are optimized for customer specific workloads and use cases.

The Dell EMC Ready Solution for HPC Digital Manufacturing is one of many solutions in the Dell EMC HPC solution portfolio. Please visit www.dellemc.com/hpc for a comprehensive overview of the available HPC solutions offered by Dell EMC.

The architecture of the Dell EMC Ready Solution for HPC Digital Manufacturing and a description of the building blocks are presented in Section 2. Section 3 describes the system configuration, software and application versions, and the benchmark test cases that were used to measure and analyze the performance of the Dell EMC HPC Ready Solution for HPC Digital Manufacturing. Section 4 presents benchmark performance for Abaqus.

2 System Building Blocks

The Dell EMC Ready Solution for HPC Digital Manufacturing is designed using preconfigured building blocks. The building block architecture allows an HPC system to be optimally designed for specific end-user requirements, while still making use of standardized, domain-specific system recommendations. The available building blocks are infrastructure servers, storage, networking, and compute building blocks. Configuration recommendations are provided for each of the building blocks which provide good performance for typical applications and workloads within the manufacturing domain. This section describes the available building blocks along with the recommended server configurations.

With this flexible building block approach, appropriately sized HPC clusters can be designed based on individual customer workloads and requirements. Figure 1 shows three example HPC clusters designed using the Dell EMC Ready Solutions for HPC Digital Manufacturing architecture.

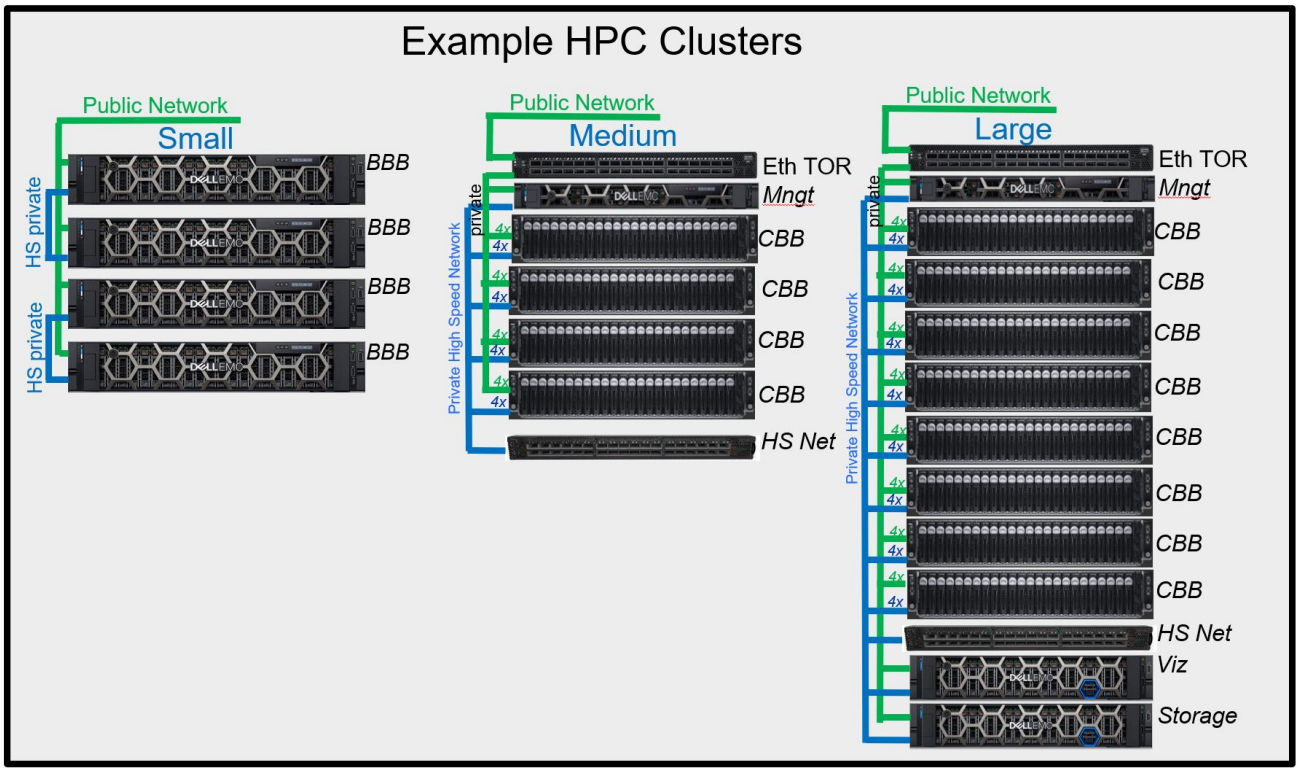


Figure 1 Example Ready Solutions for HPC Digital Manufacturing

2.1 Infrastructure Servers

Infrastructure servers are used to administer the system and provide user access. They are not typically involved in computation, but they provide services that are critical to the overall HPC system. These servers are used as the master nodes and the login nodes. For small sized clusters, a single physical server can provide the necessary system management functions. Infrastructure servers can also be used to provide storage services, by using NFS, in which case they must be configured with additional disk drives or an external storage array. One master node is mandatory for an HPC system to deploy and manage the system. If high-availability (HA) management functionality is required, two master nodes are necessary. Login nodes are optional and one login server per 30-100 users is recommended.

A recommended base configuration for infrastructure servers is:

- Dell EMC PowerEdge R640 server
- Dual Intel® Xeon® Bronze 3106 processors
- 192 GB of RAM (12 x 16GB 2667 MTps DIMMs)
- PERC H330 RAID controller
- 2 x 480GB Mixed-Use SATA SSD RAID 1
- Dell EMC iDRAC9 Enterprise
- 2 x 750 W power supply units (PSUs)
- Mellanox EDR InfiniBand™ (optional)

The recommended base configuration for the infrastructure server is described as follows. The PowerEdge R640 server is suited for this role. Typical HPC clusters will only use a few infrastructure servers; therefore, density is not a priority, but manageability is important. The Intel Xeon Bronze 3106 processor, with 8 cores per socket, is a basic recommendation for this role. If the infrastructure server will be used for CPU intensive tasks, such as compiling software or processing data, then a more capable processor may be appropriate. 192 GB of memory provided by twelve 16 GB DIMMs provides sufficient memory capacity, with minimal cost per GB, while also providing good memory bandwidth. These servers are not expected to perform much I/O, so a single mixed-use SATA SSD should be sufficient for the operating system. For small systems (four nodes or less), an Ethernet network may provide sufficient application performance. For most other systems, EDR InfiniBand is likely to be the data interconnect of choice, which provides a high-throughput, low-latency fabric for node-to-node communications or access to a Dell EMC Ready Solution for HPC NFS Storage solution or a Dell EMC Ready Solution for HPC Lustre Storage solution.

2.2 Compute Building Blocks

Compute Building Blocks (CBB) provide the computational resources for most HPC systems for Digital Manufacturing. These servers are used to run the Abaqus simulations. The best configuration for these servers depends on the specific mix of applications and types of simulations being performed by each customer. Since the best configuration may be different for each customer, a table of recommended options are provided that are appropriate for these servers. The specific configuration can then be selected based on the specific system and workload requirements of each customer. Relevant criteria to consider when making these selections are discussed in the application performance chapters of this white paper. The recommended configuration options for the Compute Building Block are provided in Table 1.

Table 1 Recommended Configurations for the Compute Building Block

Platforms	Dell EMC PowerEdge R640 Dell EMC PowerEdge C6420
Processors	Dual Intel Xeon Gold 6242 (16 cores per socket) Dual Intel Xeon Gold 6248 (20 cores per socket) Dual Intel Xeon Gold 6252 (24 cores per socket)
Memory Options	192 GB (12 x 16GB 2933 MTps DIMMs) 384 GB (12 x 32GB 2933 MTps DIMMs) 768 GB (24 x 32GB 2933 MTps DIMMs, R640 only)
Storage Options	PERC H330, H730P or H740P RAID controller 2 x 480GB Mixed-Use SATA SSD RAID 0 4 x 480GB Mixed-Use SATA SSD RAID 0
iDRAC	iDRAC9 Enterprise (R640) iDRAC9 Express (C6420)
Power Supplies	2 x 750W PSU (R640) 2 x 2000W PSU (C6400)
Networking	Mellanox® ConnectX®-5 EDR InfiniBand™ adapter

2.3 Basic Building Blocks

Basic Building Block (BBB) servers are selected by customers to create simple but powerful HPC systems. These servers are appropriate for smaller HPC systems where reducing the management complexity of the HPC system is important. The BBB is based on the 4-socket Dell EMC PowerEdge R840 server.

The recommended configuration for BBB servers is:

- Dell EMC PowerEdge R840 server
- Quad Intel Xeon Gold 6242 processors
- 384 GB of RAM (24 x 16GB 2933 MTps DIMMS)
- PERC H740P RAID controller
- 2 x 240GB Read-Intensive SATA SSD RAID 1 (OS)
- 4 x 480GB Mixed-Use SATA SSD RAID 0 (scratch)
- Dell EMC iDRAC9 Enterprise
- 2 x 1600W power supply units (PSUs)
- Mellanox ConnectX-5 EDR InfiniBand (optional)
- Mellanox 25 GbE (optional)

The R840 platform is used to minimize server count and provide good compute power per server. Each server can contain up to four Intel Xeon processors, where each BBB is essentially two CBB's fused into a single server. The Intel Xeon Gold 6142 processor is a sixteen-core CPU with a base frequency of 2.6 GHz and a max all-core turbo frequency of 3.3 GHz. With four processors, a BBB contains 64 cores, a natural number for many CAE simulations. A memory configuration of 24 x 16GB DIMMs is used to provide balanced performance and capacity. While 384GB is typically sufficient for most CAE workloads, customers expecting to handle larger production jobs should consider increasing the memory capacity to 768GB. Various CAE

applications, such as implicit FEA, often have large file system I/O requirements and four Mixed-use SATA SSD's in RAID 0 are used to provide fast local I/O. The compute nodes do not normally require extensive OOB management capabilities; therefore, an iDRAC9 Express is recommended.

Additionally, two BBB's can be directly coupled together via a high-speed network cable, such as InfiniBand or Ethernet, without need of an additional high-speed switch if additional compute capability is required for each simulation run (HPC Couplet). BBB's provide a simple framework for customers to incrementally grow the size and power of the HPC cluster by purchasing individual BBBs, BBB Couplets, or combining the individual and/or Couplets with a high-speed switch into a single monolithic system.

Performance testing for BBB's has been done using both Linux and Windows Server 2016. In general, Linux provides better overall performance, and an easier path to combining BBB's to create larger, more capable HPC clusters. We have tested up to two BBB's with Linux using both 25 Gigabit Ethernet and EDR InfiniBand adapters/cable in the two-node couplet configuration. With Linux, the EDR couplet gave the best overall performance across our tests. While the performance of the 25GbE based couplet was often comparable to the EDR based couplet, we saw little reason not to use EDR to ensure better overall performance at a similar cost and complexity. However, for customers not wishing to deploy InfiniBand in their environment, choosing a 25GbE based couplet is a suitable alternative.

For Windows testing, we tested only a couplet with a 25 GbE network. Support for InfiniBand on Windows is not currently feasible for most customers. Customers wishing for the highest level of performance, and potential cluster expansion would be advised to use Linux as an operating system.

2.4 Storage

Dell EMC offers a wide range of HPC storage solutions. For a general overview of the entire HPC solution portfolio please visit www.dellemc.com/hpc. There are typically three tiers of storage for HPC: scratch storage, operational storage, and archival storage, which differ in terms of size, performance, and persistence.

Scratch storage tends to persist for the duration of a single simulation. It may be used to hold temporary data which is unable to reside in the compute system's main memory due to insufficient physical memory capacity. HPC applications may be considered "I/O bound" if access to storage impedes the progress of the simulation. For these HPC workloads, typically the most cost-effective solution is to provide sufficient direct-attached local storage on the compute nodes. For situations where the application may require a shared file system across the compute cluster, a high performance shared file system may be better suited than relying on local direct-attached storage. Typically using direct-attached local storage for most CAE simulations offers the best overall price/performance and is considered best practice. For this reason, local storage is included in the recommended configurations with appropriate performance and capacity for a wide range of production workloads. If anticipated workload requirements exceed the performance and capacity provided by the recommended local storage configurations, care should be taken to size scratch storage appropriately based on the workload.

Operational storage is typically defined as storage used to maintain results over the duration of a project and other data, such as home directories, such that the data may be accessed daily for an extended period of time. Typically, this data consists of simulation input and results files, which may be transferred from the scratch storage, typically in a sequential manner, or from users analyzing the data, often remotely. Since this data may persist for an extended period, some or all of it may be backed up at a regular interval, where the interval chosen is based on the balance of the cost to either archive the data or regenerate it if need be. Archival data is assumed to be persistent for a very long term, and data integrity is considered critical. For many modest HPC systems, use of the existing enterprise archival data storage may make the most sense,

as the performance aspect of archival data tends to not impede HPC activities. Our experience in working with customers indicates that there is no 'one size fits all' operational and archival storage solution. Many customers rely on their corporate enterprise storage for archival purposes and instantiate a high performance operational storage system dedicated for their HPC environment.

Operational storage is typically sized based on the number of expected users. For fewer than 30 users, a single storage server, such as the Dell PowerEdge R740xd is often an appropriate choice. A suitable equipped storage server may be:

- Dell EMC PowerEdge R740xd server
- Dual Intel® Xeon® Bronze 4110 processors
- 96 GB of memory, 12 x 8GB 2667 MT/s DIMMS
- PERC H730P RAID controller
- 2 x 250GB Mixed-use SATA SSD in RAID-1 (For OS)
- 12 x 12TB 3.5: nlSAS HDDs in RAID-6 (for data)
- Dell EMC iDRAC9 Express
- 2 x 750 W power supply units (PSUs)
- Mellanox EDR InfiniBand Adapter
- Site specific high-speed Ethernet Adapter(optional)

This server configuration would provide 144TB of raw storage. For customers expecting between 25-100 users, an operational storage solution, such as the Dell EMC Ready Solution for HPC NFS Storage (NSS), shown in Figure 2, with up 840 TB of raw storage of storage may be appropriate:

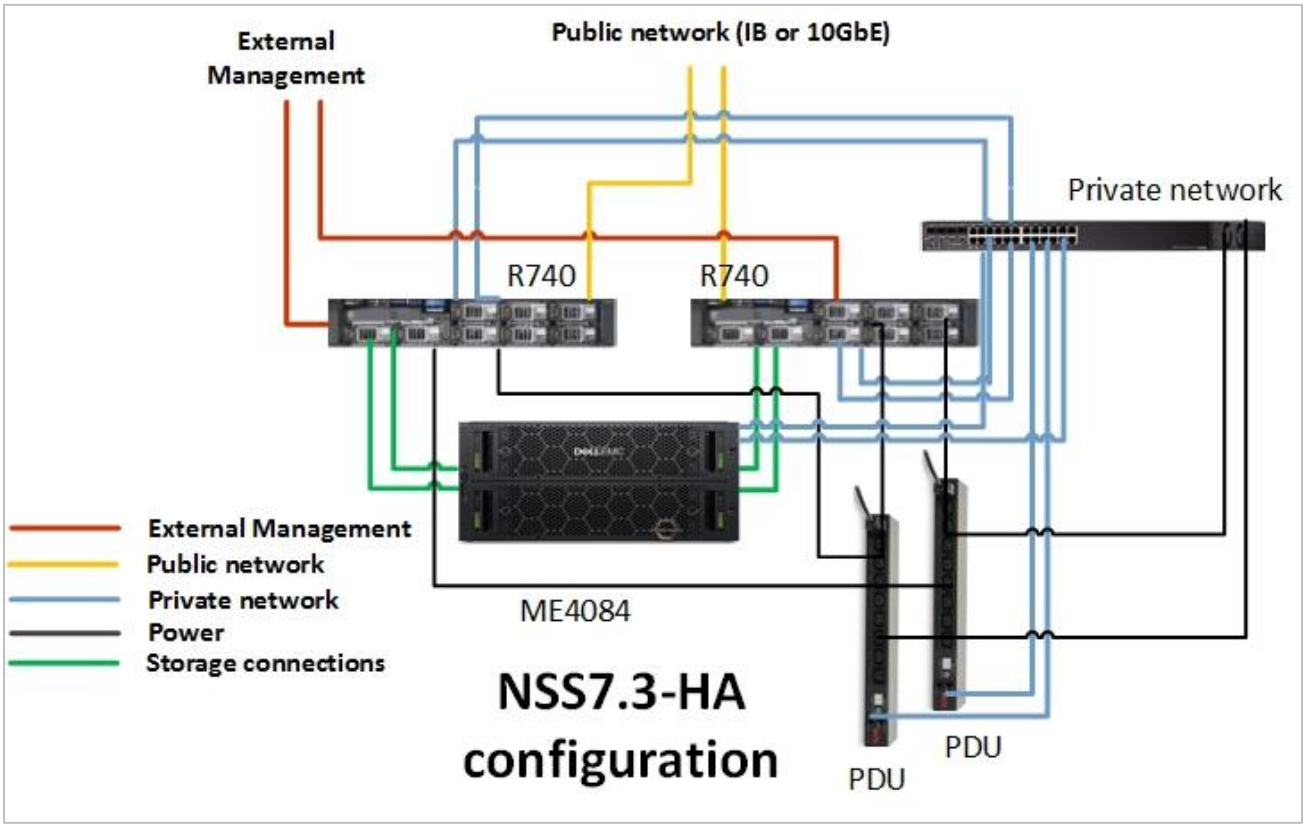


Figure 2 NSS7.3-HA Storage System Architecture

For customers desiring a shared high-performance parallel filesystem, the Dell EMC Ready Solution for HPC Lustre Storage solution shown in Figure 3 is appropriate. This solution can scale up to multiple petabytes of storage.

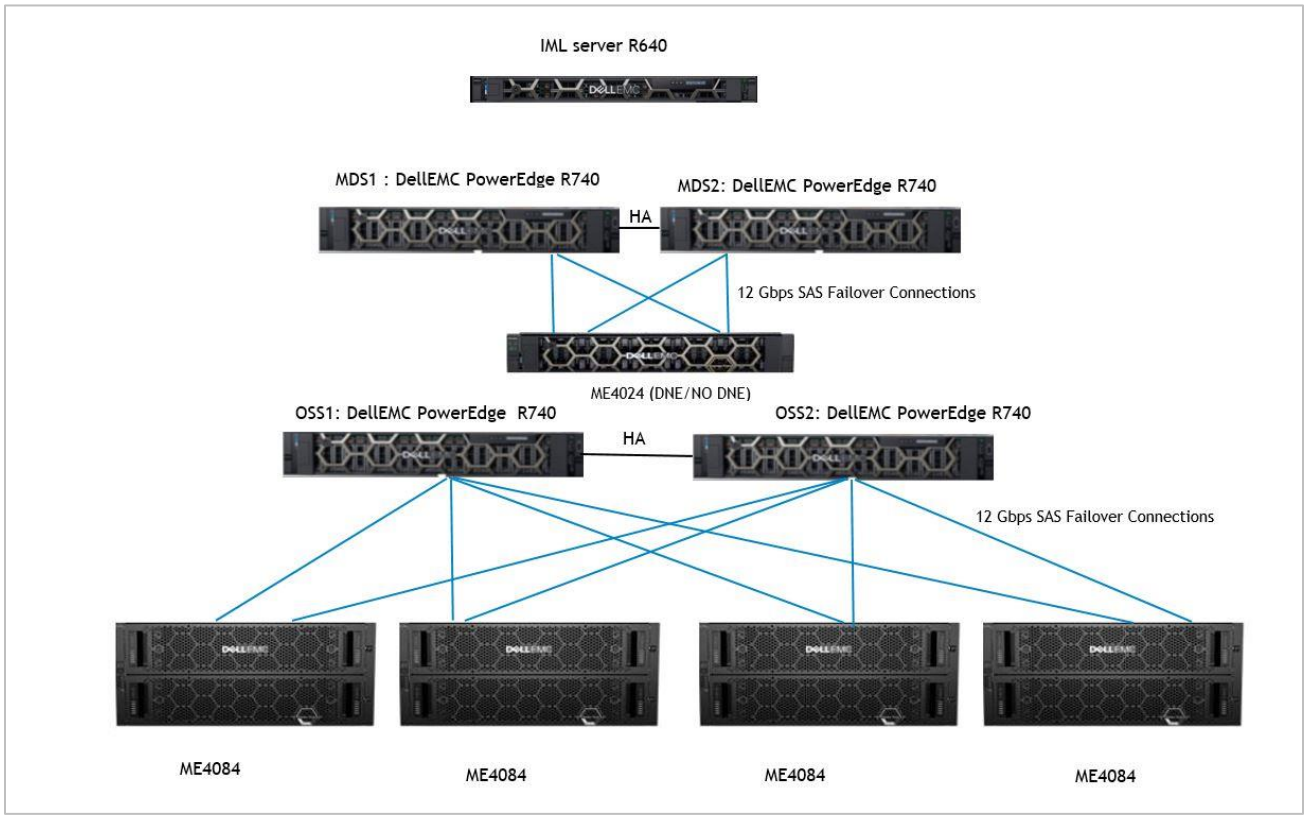


Figure 3 Dell EMC Ready Solution for Lustre Storage Reference Architecture

2.5 System Networks

Most HPC systems are configured with two networks—an administration network and a high-speed/low-latency switched fabric. The administration network is typically Gigabit Ethernet that connects to the onboard LOM/NDC of every server in the cluster. This network is used for provisioning, management and administration. On the CBB servers, this network will also be used for IPMI hardware management. For infrastructure and storage servers, the iDRAC Enterprise ports may be connected to this network for OOB server management. The management network typically uses the Dell Networking S3048-ON Ethernet switch. If there is more than one switch in the system, multiple switches should be stacked with 10 Gigabit Ethernet cables.

A high-speed/low-latency fabric is recommended for clusters with more than four servers. The current recommendation is an EDR InfiniBand fabric. The fabric will typically be assembled using Mellanox SB7890 36-port EDR InfiniBand switches. The number of switches required depends on the size of the cluster and the blocking ratio of the fabric.

2.6 Cluster Management Software

The cluster management software is used to install and monitor the HPC system. Bright Cluster Manager (BCM) is the recommended cluster management software.

2.7 Services and Support

The Dell EMC Ready Solution for HPC Digital Manufacturing is available with full hardware support and deployment services, including additional HPC system support options.

3 Reference System

The reference system was assembled in the Dell EMC HPC and AI Innovation Lab using the building blocks described in section 2. The building blocks used for the reference system are listed in Table 2.

Table 2 Reference System Configuration

Building Block	Quantity
Infrastructure Server	1
Computational Building Block (CBB) PowerEdge C6420 Dual Intel Xeon Gold 6242 192GB RAM 12x16GB 2933 MTps DIMMs Mellanox ConnectX-5 EDR adapter	2
Computational Building Block (CBB) PowerEdge C6420 Dual Intel Xeon Gold 6252 192 GB RAM 12x16GB 2933 MTps DIMMs Mellanox ConnectX-5 EDR adapter	8
Basic Building Block	2
Dell Networking S3048-ON Ethernet Switch	1
Mellanox SB7700 EDR InfiniBand Switch	1

The BIOS configuration options used for the reference system are listed in Table 3.

Table 3 BIOS Configuration

BIOS Option	Setting
Logical Processor	Disabled
Virtualization Technology	Disabled
System Profile	Performance Profile
Sub NUMA Cluster	Enabled (CBB) Disabled (BBB)

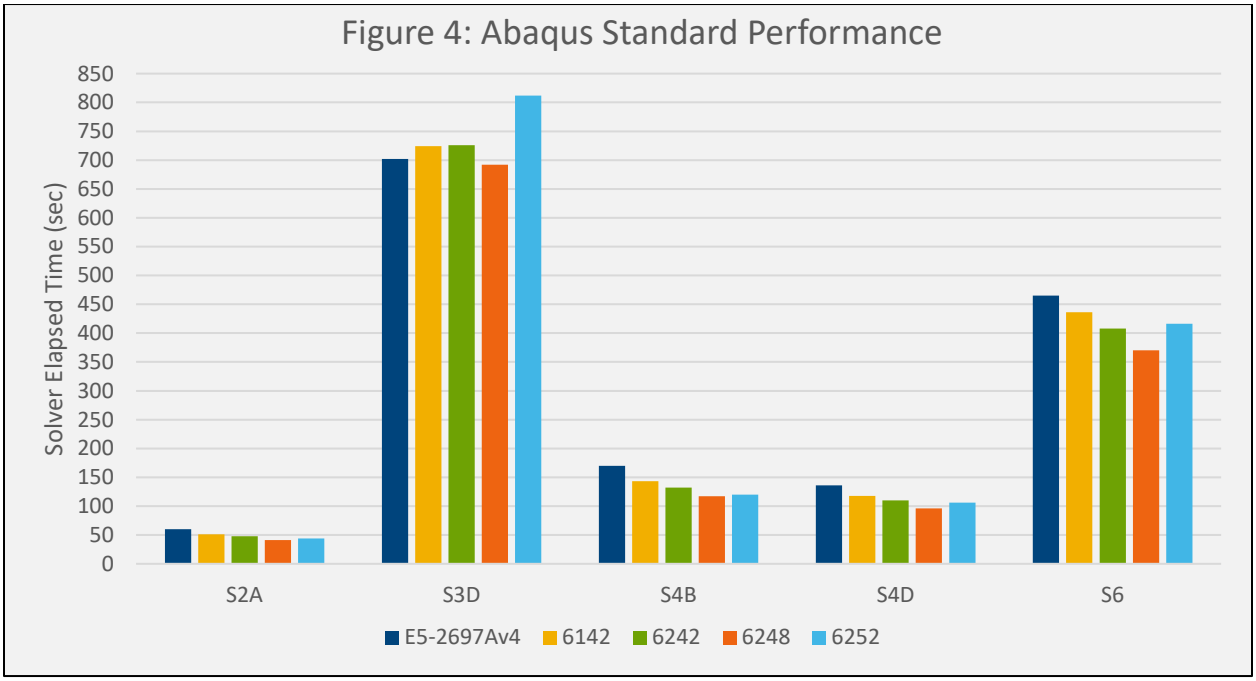
The software versions used for the reference system are listed in Table 4.

Table 4 Software Versions

Component	Version
Operating System	RHEL 7.6 Windows Server 2016 (BBB)
Kernel	3.10.0-957.el7.x86_64
OFED	Mellanox 4.5-1.0.1.0
Bright Cluster Manager	8.2
Simulia Abaqus	2019

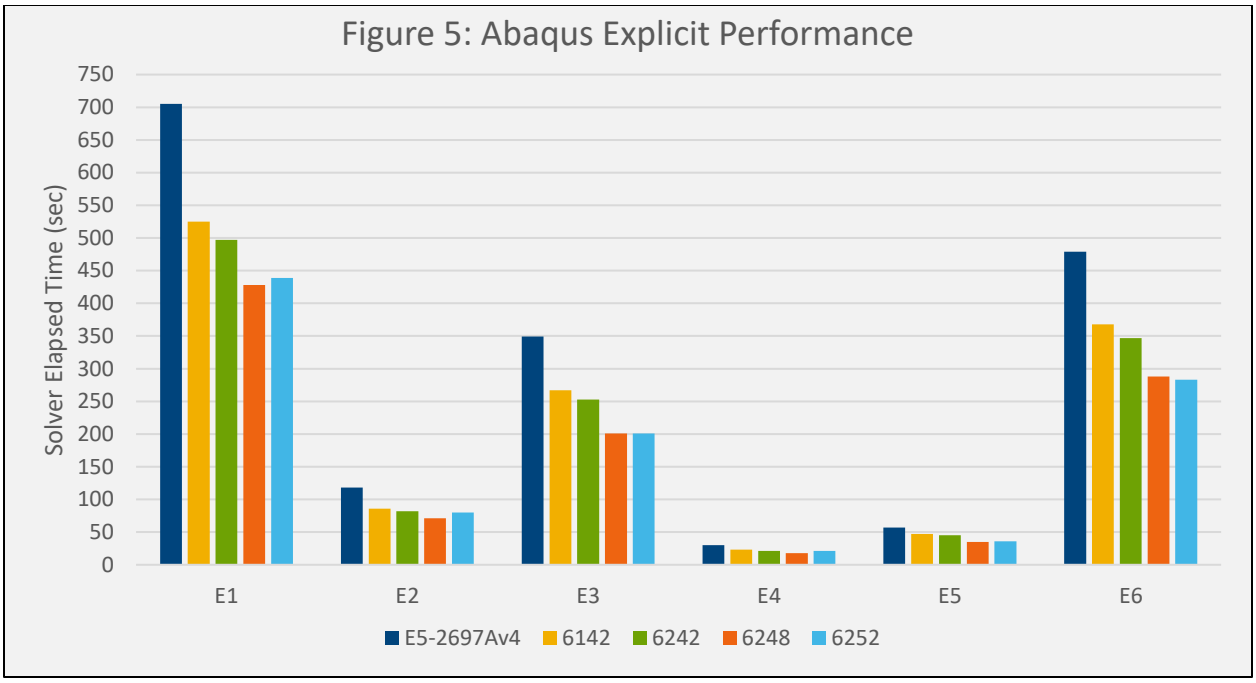
4 Abaqus Performance

Abaqus is a multi-physics Finite Element Analysis (FEA) software commonly used in multiple engineering disciplines. Depending on the specific problem types, FEA codes may or may not scale well across multiple processor cores and servers. Implicit FEA problems often place large demands on the memory and disk I/O sub-systems. Abaqus contains several solver options, both implicit and explicit. As such it is difficult to summarize the overall performance potential of Abaqus with a few benchmarks. Each Abaqus release distribution does contain some standard benchmarks, both for the implicit solver (Sxx or standard) and the explicit solver (Ex for explicit). These benchmarks are useful to get an indication of the relative performance potential for different systems, so should be viewed more qualitatively than quantitatively. Figure 4 shows a single server performance comparison for four standard Abaqus benchmarks when using all processors cores on the server, where the value for each benchmark is the solver wall time based on the output at the bottom of the .msg file.



For comparison, the figure also includes performance data for prior generations of the Ready Solution for HPC Digital Manufacturing the 13th generation system using 16-core Intel E5-2697Av4 processors, and the 14th generation system using Intel Xeon Gold 16-core 6142 processors. With the exception of the “s3d” model, all other models were run with “mp_mode=MPI” and “mp_host_split=8”, which we have found is typically optimal for systems with more than 16 cores. The “s3d” benchmark is a modal analysis which operates only in “mp_mode=threads” mode. These results demonstrate that the latest Intel Xeon Scalable (code-name Cascade Lake) processors 62XX deliver noticeably improved performance other its predecessors. Typically, the processors with more cores, outperformed processors with fewer cores, while there were a few cases where the 20-core 6248 outperformed the 24-core 6252. However, these standard benchmarks are much smaller than typical production size cases, where the additional processors likely will improve overall performance.

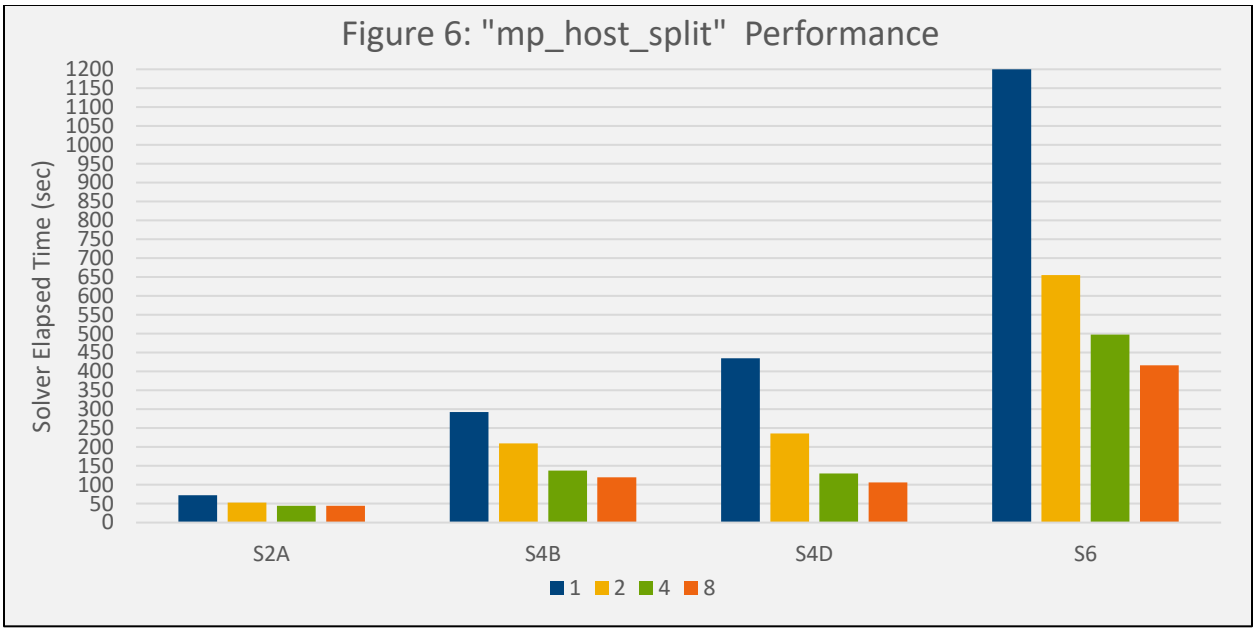
Figure 5 shows the singer server performance for the Abaqus Explicit benchmark models E1-E6, on the same servers noted above. The benchmark timings were based on the wall clock timing listed at the bottom of the .sta file.



These results are consistent with the Standard results in Figure 4, where the newer Cascade Lake processors with the most cores performed the best.

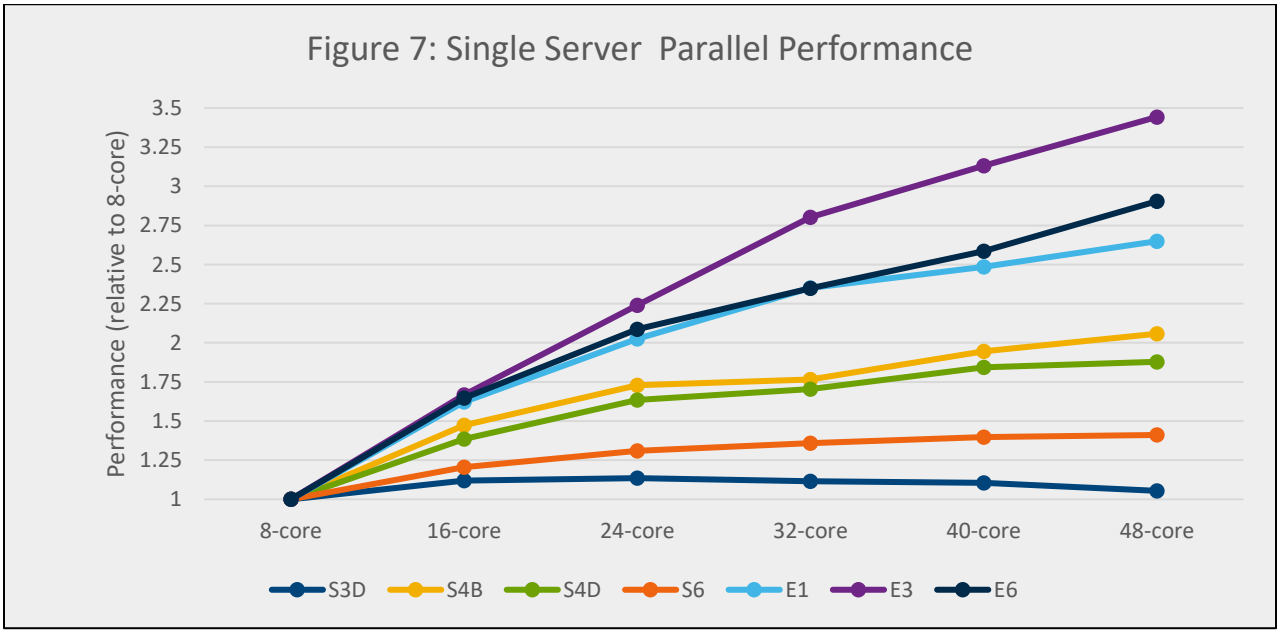
While the explicit solver in Abaqus is a straight forward MPI parallel implementation, the typical standard solver employs a hybrid parallel algorithm using both shared memory parallel threads and MPI domain parallelism. The default run mode for the standard solver is to use a simple MPI domain per server, with parallel threads for each available core on the server. However, the parallel efficiency of the thread parallelism tends to drop off depending on the model size and features after 5-10 threads. Abaqus enables the user to carry out simulations by placing more than a single MPI domain on a server to reduce the number of shared memory parallel threads per domain to increase overall program efficiency. This can be easily activated with the command line argument “mp_host_split=xx” argument. There is no absolute method to *a priori* determine the optimal number of MPI domains per server.

Figure 6 shows the effect of modifying the number of MPI domains for the standard benchmarks on a single server with the 24-core Intel Gold 6252 processor.



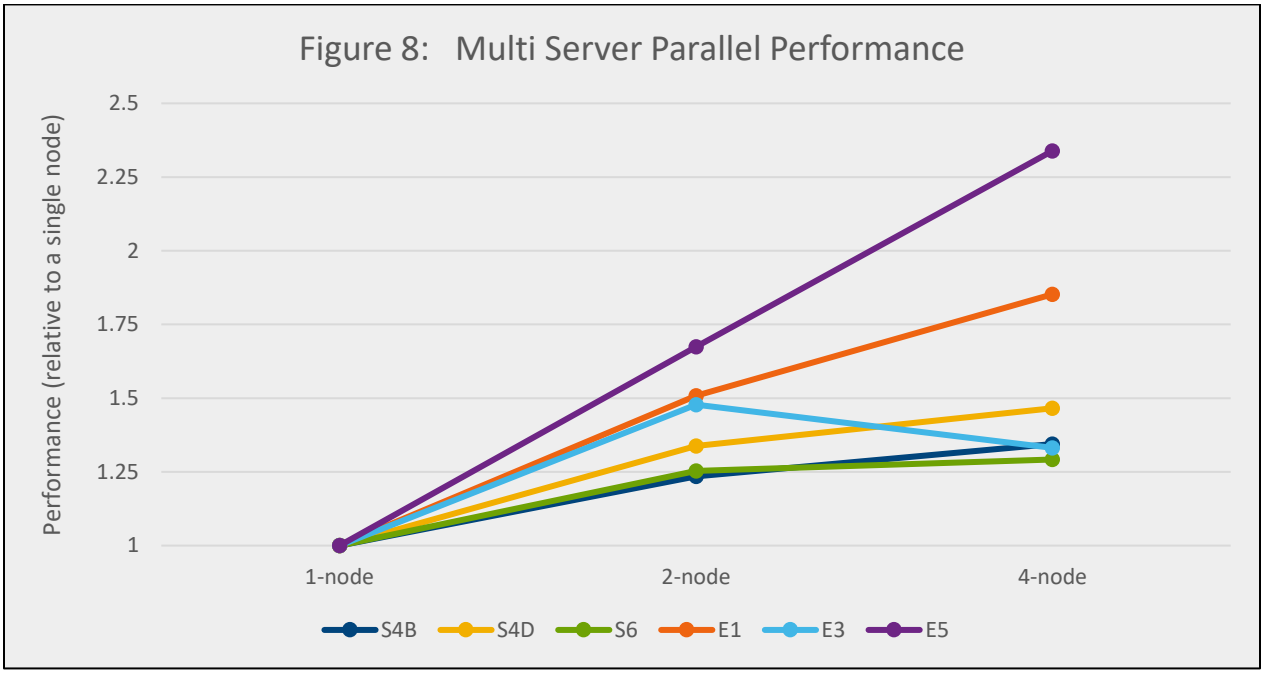
For all of the models test, substantial performance gains can be made using multiple domains per node, where using 8 domains (6 threads per domain) delivers the optimal performance. Users are encouraged to examine this option with their models to determine the optimal value. An even number is preferred, since it would allow MPI processor binding to be enabled to further improve performance. There may be an increase in the amount of memory required to minimize I/O when more than a single domain is placed on a node and one needs to be careful to avoid “out-of-core” solutions, causing potentially significant I/O activity, decreasing the overall performance. The user can examine the domain memory requirements to minimize I/O in the .dat file to make sure this does not occur.

Figure 7 shows the performance on the larger standard and explicit benchmarks mentioned above on a 6252 based system, where the number of cores used was varied from 8 to 48 cores.



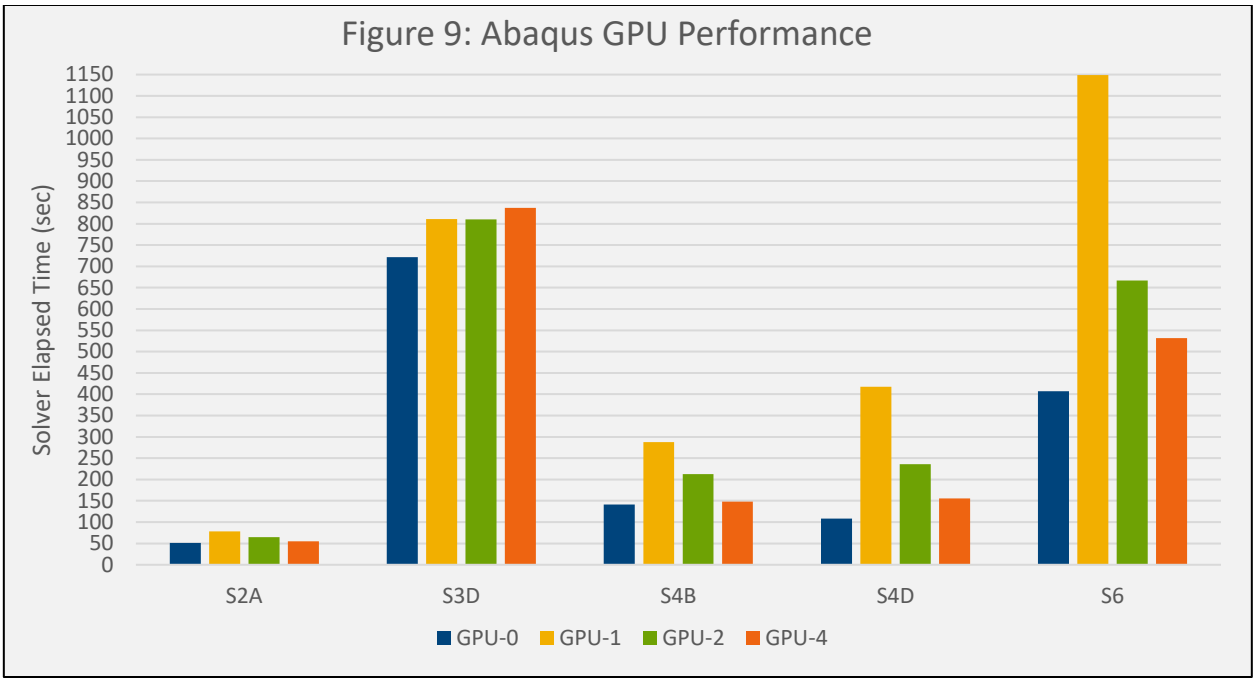
These results demonstrate that while the performance increase is not linear with respect to the number of cores per node, there can be a substantial benefit in using system with large numbers of cores, and typically the best performance is obtained when using all of the cores available. The exception for these cases is the modal analysis model S3D, which is only thread parallel. The MPI mode is required to take full advantage of several cores per server.

Figure 8 shows the parallel performance improvement when running a single job in parallel across multiple nodes. These benchmarks were carried out on a cluster with four 6252 based servers using all available cores in each server.



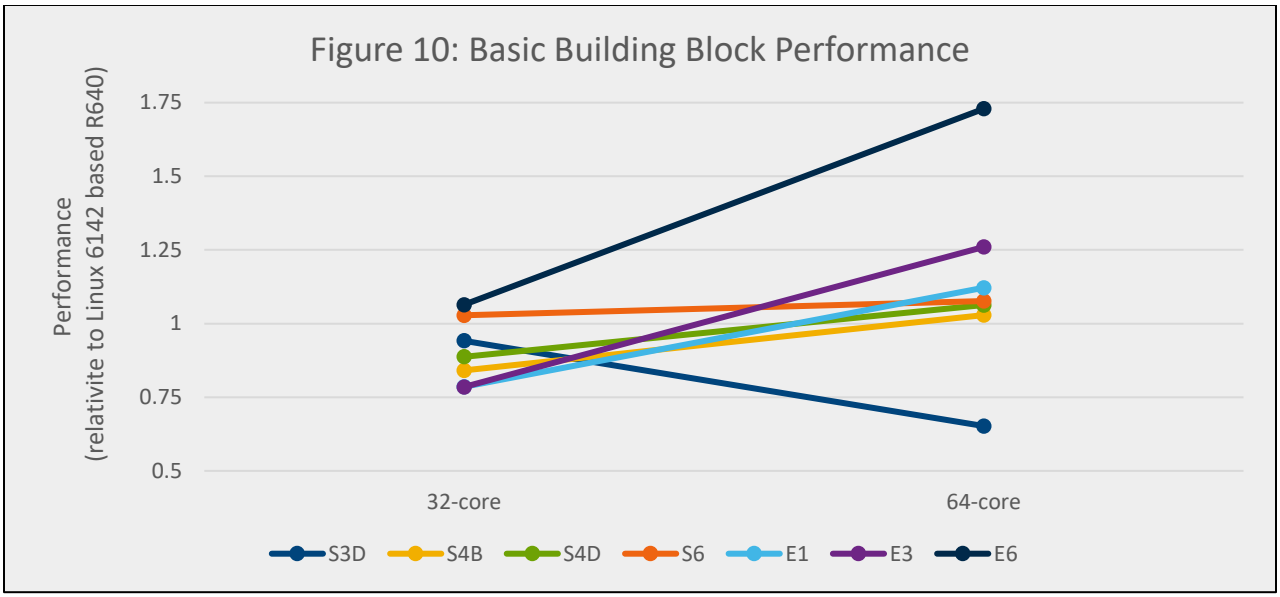
The parallel speedup when running jobs across more than a single node is somewhat mixed. These datasets are rather small by current production standards, and do not represent typical production sized simulations. However, these models do show significant speedup across two nodes, and modest speedups up to four nodes.

Many FEA applications such as Abaqus, have been modified to allow for GPU acceleration with appropriate GPU enabled servers. Figure 9 contains performance information for the Abaqus standard benchmarks models described above run on a Dell EMC PowerEdge C4140 server, equipped with dual Intel Xeon Gold 20-core 6148 processors and four NVIDIA V100 GPUs.



For each benchmark, the wall clock time (in sec) is shown. For the GPU enabled runs, all 40 Xeon cores were used. Benchmarks were carried out using the base system with no GPU acceleration, and using 1,2,4 GPUs. With Abaqus, the number of GPUs for each of the MPI domain must be the same, so one a single server, it is possible to using the GPUs in a variety of ways. As an example, with four GPUs, the benchmarks can be made with a single MPI domain, using 0,1,2,4 GPUs. With two MPI domains, each domain could use 0,1,2 GPUs, and with four MPI domains, each domain could use 0,1 GPUs. Only certain code sections have been enabled to take advantage of GPUs, and as shown in figure 5 above, Abaqus typically runs better having more domains per node. As a result, when MPI mode is possible (S2a, S4b, S4d, S6), the best results are obtained when running with each GPU in its own MPI domain, allowing for more domains per node. With the modal analysis D3d case, all GPUs were used in the single domain. For these benchmarks, GPUs did not improve performance over the existing Xeon processors. It may be the case that for certain simulations, GPUs may offer significant performance advantages, so care should be taken when choosing systems with GPUs. to determine whether utilizing GPUs would be appropriate for their simulations.

Figure 10 demonstrates the performance of a Window's based R840 Basic Building Block on the larger benchmark models.



Overall, these benchmarks display the performance of the Window based Basic Building Block is comparable to or greater than the typical dual-socket based Linux server. Except for the modal analysis S3d benchmark models, there was a modest to noticeable performance gain from using 32 cores (two sockets) to 64 cores (four sockets) with the server. Since these test models were small in size, testing above 64 cores would not have been advantageous. However, we anticipate the potential for good performance gains for larger customer datasets using a cluster of two Basic Building Blocks described above. Customers could incrementally increase their solution capabilities by adding more BBBs, aggregating them into more powerful couplets, re-provision them from Windows to Linux, and eventually aggregate them with an InfiniBand switch into a single HPC cluster based on BBBs, all while preserving their hardware investment, and minimizing production downtime.

5 Conclusion

This technical white paper presents the Dell EMC Ready Solution for HPC Digital Manufacturing. The detailed analysis of the building block configurations demonstrate that the system is architected for a specific purpose—to provide a comprehensive HPC solution for the manufacturing domain. Use of this building block approach allows customers to easily deploy an HPC system optimized for their specific workload requirements. The design addresses computation, storage, networking 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 solution design, demonstrating system performance with Simulia Abaqus software.