Performance Analysis of HPC Applications on Several Dell PowerEdge 12th Generation Servers

This Dell technical white paper evaluates and provides recommendations for the performance of several HPC applications across three server architectures, the Dell PowerEdge M620, M420, and R820.

Nishanth Dandapanthula and Garima Kochhar

High Performance Computing Engineering

September 2012| Rev 1.0

Contents

Tables

Figures

Executive summary

In the last six months there have been a variety of new servers available in the market. These servers have several architectural differences as well as support for different amounts of memory, PCI-E slots, hard disks, and so on. All these models are good candidates for High Performance Computing clusters, but certain questions remain unanswered: Which server model is best suited for a specific application? What features of the architecture make it the ideal choice?

In this technical white paper, different server designs are compared using several high performance computing workloads. At a cluster-level, a quantitative study is undertaken analyzing both performance and energy efficiency across the different server models. The paper analyses the measured results and concludes with recommended configurations for each application.

1. Introduction

The Dell PowerEdge 12th generation server line up armed with the latest processors from Intel has been well received by the High Performance Computing (HPC) community. The new servers provide more choice than before; however, with these choices there is a need for quantitative recommendations and guidelines to match an application's requirements to the ideal cluster configuration. The latest Dell servers can support the Intel® Xeon® processor E5-2400 product family, the Intel® Xeon® processor E5-2600 product family, or the Intel® Xeon® processor E5-4600 product family, giving HPC users numerous choices to configure a server for specific CPU, memory and I/O requirements. It is a daunting, although necessary task for HPC users to understand the performance characteristics of each of these server models to be able to make well-informed decisions regarding which server platform is best for their purposes. This white paper analyses the performance and power consumption characteristics of these server platforms at an application level to help HPC users make this choice with confidence.

The latest Dell PowerEdge 12th generation servers include support for the new processors from Intel. The PowerEdge M420 servers armed with the Intel Xeon processor E5-2400 product family cater to users who need a dense compute intensive platform by accommodating 32 servers in 10 U rack space. This allows 512 cores in 10 U, doubling the typical rack density. The 4 socket PowerEdge R820 servers tap into the processing power of the Intel Xeon processor E5-4600 product family and provide massive processing power and memory density. These characteristics are attractive to users who need *fat nodes* in their clusters. Finally, the PowerEdge M620 server strikes a balance between performance, energy efficiency, scalability, and density with the Intel Xeon processor E5-2600 product family.

This white paper describes the behavior of select HPC workloads on these three Intel Xeon processor families with focus on performance and energy efficiency. The focus is on a cluster-level analysis as opposed to a single-server study. The paper first introduces each of the three Intel architectures and compares the three different processor families. It provides cluster-level results for different HPC workloads. Subsequent sections analyze the results in order to provide better understanding and recommendations regarding which type of server platform best fit a particular workload. The study characterizes each server platform based not only on its performance but also on its energy efficiency.

The behavior and guidelines presented here apply to HPC workloads similar to those tested as part of this study. The recommendations in this document may not be appropriate for general enterprise workloads.

2. Dell PowerEdge 12th generation server platforms and Intel processor architecture

A detailed comparison of the latest Intel processor architectural variants (Xeon E5-2400/E5-2600/4600, architecture codenamed Sandy Bridge) is provided in Table 1. It also provides a comparison to the previous generation Intel Xeon processor 5600 series (architecture codenamed Westmere). At a glimpse, the major improvements on new Sandy Bridge based servers when compared to the previous generation Westmere servers are the 33 percent increase in core count, increase in memory channels, support for higher memory speeds and higher QPI speeds. A previous study by the authors [1] describes the Dell PowerEdge 12th generation server models and the architecture of Sandy Bridge EP (Intel Xeon processor E5-2600 product family) in great detail. It also explains the differences between Westmere-EP and Sandy Bridge-EP at a more granular level. A block diagram of the Sandy Bridge EP processor architecture is included in this document in Figure 1 as reference.

Table 1. Intel architecture comparison

	Intel Xeon Processor 5600 Series	Intel Xeon Processor E5-2400 Product Family	Intel Xeon Processor E5-2600 Product Family	Intel Xeon Processor E5-4600 Product Family
Architecture Codename	Westmere-EP	Sandy Bridge-EN	Sandy Bridge-EP	Sandy Bridge-EP 4S
Max Sockets / Cores per socket	2/6	2/8	2/8	4/8
Memory channels	3 per socket	3 per socket	4 per socket	4 per socket
Max Memory speed	1333 MHz	1600 MHz	1600 MHz	1600 MHz
QPI links per CPU	2	1	2	2
Max QPI Speed	6.4 GT/s	8 GT/s	8 GT/s	8 GT/s
Max Processor TDP	130 W	95W	135W	130W
Max DIMMs Per Channel (DPC)	3DPC	2 DPC	3 DPC	3 DPC
Dell PowerEdge servers models	R610, R710, M610	R420, R520, M420	R620, R720, M620	R820, M820

Figure 2 outlines the block diagram of the Sandy Bridge-EN platform architecture. Compared to the Sandy Bridge-EP platform, the differences lie in the number of QPI links, the number of memory channels and the number of DIMMs per channel. The EN based processors operate at a lower maximum wattage compared to the EP based processors. InfiniBand FDR is not supported on the EN based processers. In its place, InfiniBand FDR10 [2] is used. The EN processor is a balanced configuration in terms of bandwidth. The processors can support a theoretical maximum of 32 GB/s through the QPI link between the sockets and the theoretical maximum memory bandwidth from a socket to its memory is 38.4 GB/s.

Figure 1. Platform architecture for Sandy Bridge EP (Intel Xeon E5-2600)

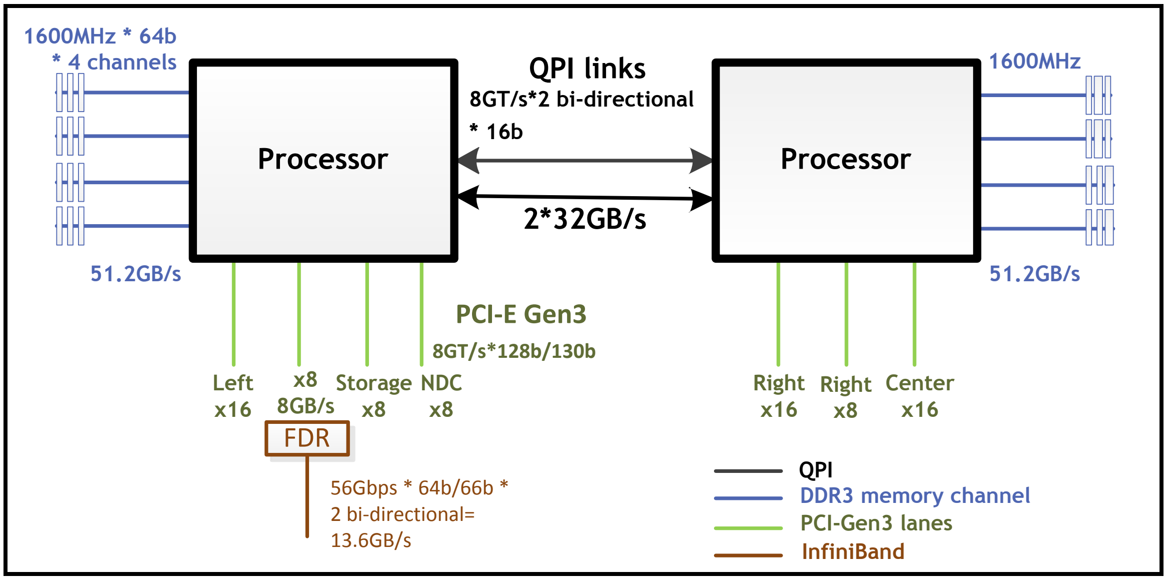


Figure 2. Platform architecture for Sandy Bridge EN (Intel Xeon E5-2400)

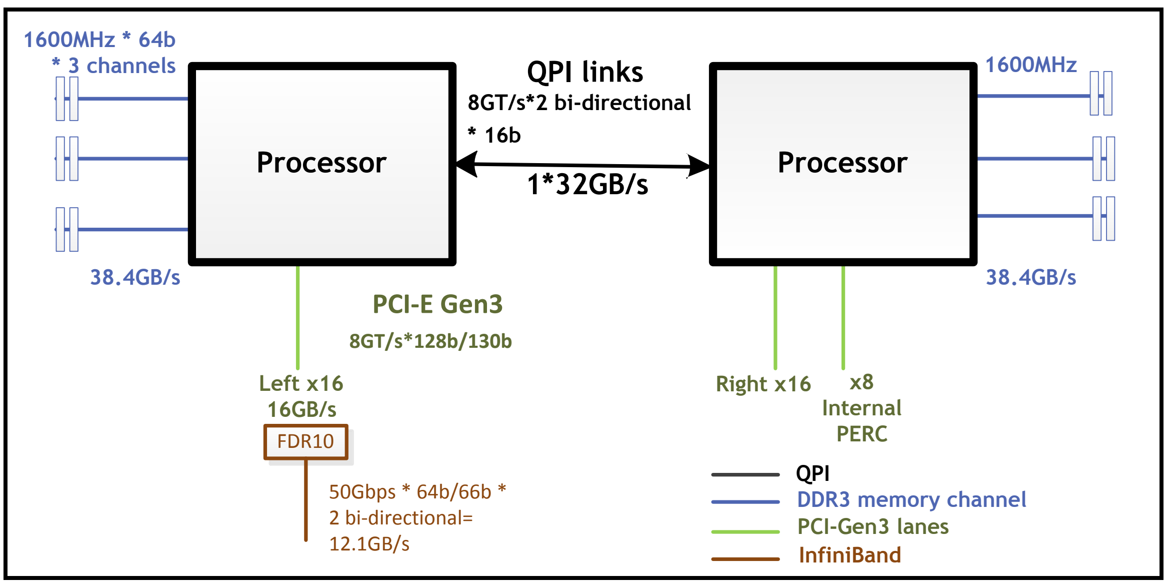
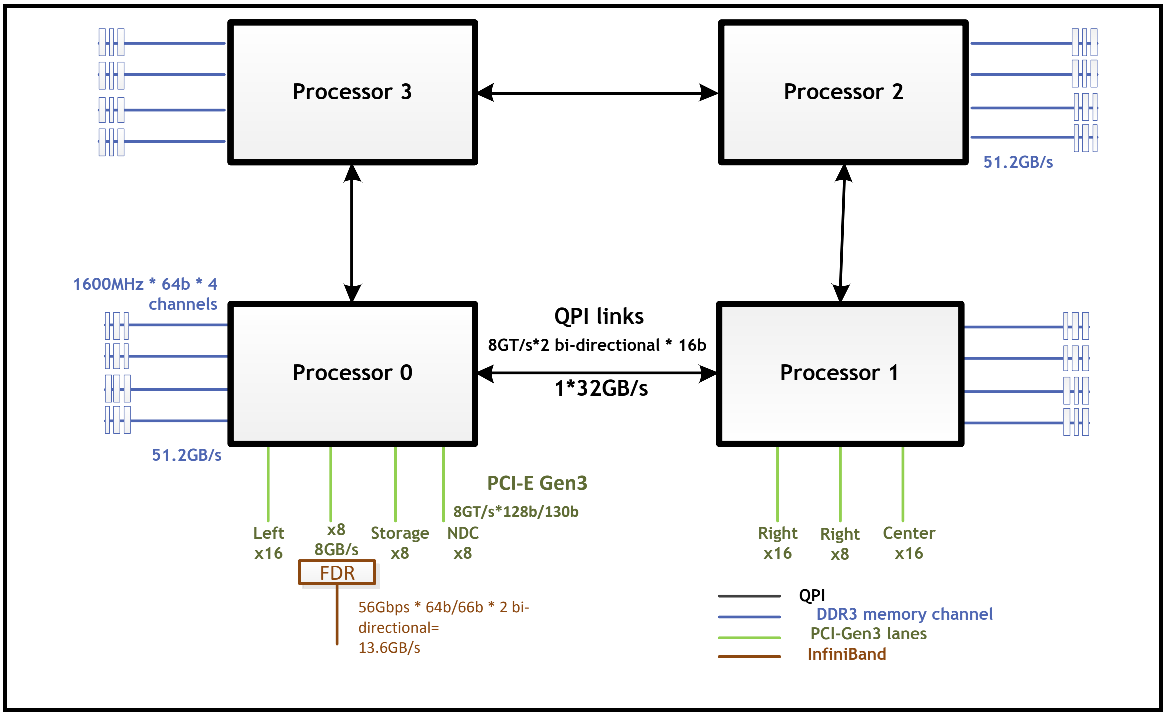


Figure 3 describes the platform architecture of Sandy Bridge-EP 4 socket platform. Each socket has two QPI links, but any two adjacent sockets are connected by just one QPI link in a ring-structure. There is no cross link between processors one and three, and between processors zero and two. Thus, any communication between these two socket pairs needs to traverse two QPI links. Only two of the sockets have PCI-lanes and therefore can be *local* to PCI cards installed in the system. Other than the differences in number of QPI links, the 4 socket platform architecture is very close to the 2 socket EP platform architecture.

Figure 3. Platform architecture for Sandy Bridge EP - 4 Socket (Intel Xeon E5-4600)



Information regarding any of the Intel Sandy Bridge processors can be obtained from [3].

3. Test bed and applications

The previous section presented the differences between the three Sandy Bridge based architectures. This section details the test bed used in the study, explains the choices selected in configuring the test bed, and describes the HPC applications used in this study. Subsequent sections evaluate the performance of these HPC workloads on the different architectures.

Three types of HPC clusters were configured for this purpose. The details of this test bed are provided in Table 2. A 16 server Dell PowerEdge M620 cluster was deployed to represent Sandy Bridge-EP while a 32 server PowerEdge M420 cluster represented the Sandy Bridge-EN. A four server PowerEdge R820 cluster was used for Sandy Bridge-EP 4S. A PowerEdge R620 rack server was used as the master node of the cluster.

Table 2. Test bed details

Component	PowerEdge M420 Cluster	PowerEdge M620 Cluster	PowerEdge R820 Cluster
Server Configuration	PowerEdge M420 blade server (32) in a PowerEdge M1000e chassis	PowerEdge M620 blade server (16) in a PowerEdge M1000e chassis	PowerEdge R820 rack server (4)
Architecture	Sandy Bridge EN	Sandy Bridge EP	Sandy Bridge EP - 4S
Processor	Dual Intel Xeon E5-2470 @ 2.3GHz	Dual Intel Xeon E5-2680 @ 2.7GHz	Quad Intel Xeon E5-4650 @ 2.7GHz
Memory	6 * 8GB @ 1600MT/s	8 * 8GB @ 1600 MT/s	16 * 8GB @ 1600 MT/s
Memory Configuration	1 DIMM Per Channel at 1600 MHz		
InfiniBand	Mellanox ConnectX-3 FDR10 Two Mellanox M4001T FDR10 IO modules for the PowerEdge M1000e blade chassis	Mellanox ConnectX-3 FDR Mellanox M4001F FDR IO module for the PowerEdge M1000e blade chassis	Mellanox ConnectX-3 FDR Mellanox FDR rack switch SX6036
Cluster Size	32 Servers, 512 Cores	16 Servers, 256 Cores	4 Servers, 128 Cores
Turbo bins [4] (100 MHz)	+5 (when 8 cores active)	+4 (when 8 cores active)	+2 (when 8 cores active)
Disk	1*50GB SSD	1*146GB 15K SAS	1*146GB 15K SAS
Disk Controller	PERC H310		
BIOS	1.2.4	1.1.2	1.1.5
iDRAC	1.20.20 (Build 24)	1.06.06 (Build 15)	1.20.20 (Build 24)
OFED	Mellanox OFED 1.5.3-3.0.0		
OS	RHEL 6.2 - 2.6.32-220.el6.x86_64		

Even though the absolute memory configurations appear differently, all servers contain a balanced one DIMM per channel configuration running at a memory speed of 1600 MT/s. For the PowerEdge R820 and PowerEdge M620 cluster, the amount of memory per core is also identical.

The Intel Xeon processor E5-2680 on the PowerEdge M620 is the highest bin 130W part available in that product family. The Intel Xeon processor E5-4650 is the highest bin processor available in the EP 4 socket product family and the Intel Xeon processor E5-2470 is the highest bin processor available in the EN product family which supports 8 GT/s QPI speed and 1600 MT/s memory.

The PowerEdge M620 and PowerEdge M420 are blade based servers. The PowerEdge M620 is a half-height blade; the PowerEdge M1000e chassis can house up to 16 such blades. The PowerEdge M420 is a denser, quarter-height blade and the same PowerEdge M1000e chassis can house up to 32 such blades. A full chassis of servers was used in each case to allow meaningful power measurements and to properly amortize the shared infrastructure cost of power supplies, fans, and so on. The other differences in the size of the clusters are due to resource limitations; however the results sections compares performance based on the number of cores to eliminate total cluster size as a factor.

The PowerEdge M420 supports only SSD drives. The operating system for the server was installed on this drive. None of the applications were configured to write local files on each compute node; therefore, the choice of SSD versus SAS is not relevant to the results in this study.

The BIOS on all the servers are set to Dell HPC defaults, which include the Performance per Watt Optimized DAPC System Profile, Node Interleaving disabled and Logical Processor disabled. This System Profile balances power saving and performance options by enabling Turbo Boost, C states and C1-E. The Power Profile is set to DAPC (Dell Advanced Power Controller) and the Memory Frequency is set to max performance.

StackIQ Rocks+ 6.0.1 Dell edition [5] was used to deploy and manage the cluster.

Table 3 illustrates the applications that were studied, the benchmarks used, and their characteristics. The applications chosen are a mix of open source and commercial applications.

Table 3. Application and benchmark details

Application	Domain	Version	Benchmark data set
High Performance Linpack	Floating point CPU intensive system benchmark	Intel MKL v10.3.9.293	Problem size set to 90 percent of total memory.
Stream	Memory Bandwidth micro-benchmark	v5.9	Array size 160000000
ANSYS Fluent	Computational Fluid Dynamics application	v14.0.0	truck_poly_14m and truck_111m
WRF	Weather modeling application	v3.1	Conus 12k
NAMD	Molecular Dynamics application	v2.9	STMV
MILC	Quantum Chromo-dynamics application	v7.6.3	fnl-2009-intel.in Based on Medium-NSFt3
LU	Lower-upper decomposition, physical systems	NPB v3.3.1	Class D

For HPL, the performance metric used for comparison is ***GFLOPS*** and for WRF, the performance metric used is the *average time step*. For NAMD, the performance metric used is *days per nanosecond*. For all other applications the metric used is *rating*. Rating is defined as the number of times an application can be executed in a single day. In addition to quantifying the performance on the above mentioned server platforms, the power consumed is also measured by using a rack power distribution unit (PDU). Because an *apples to apples* comparison is not possible with the test bed configuration, a cluster level comparison of power consumption is provided in Power Consumption and Energy Efficiency.

A previous study [1] characterized the performance and energy impact of different BIOS tuning options on the PowerEdge M620 servers. The PowerEdge M620 cluster test bed and applications used in that study were identical to this one, and therefore data from that analysis is leveraged for the Sandy Bridge EP portion of this work.

4. Results and analysis

This section compares the performance characteristics of each of the above mentioned applications on the three different server platforms.

Because the Dell PowerEdge R820 has double the number of cores per server over the PowerEdge M620 and the PowerEdge M420, the performance comparison is made on the basis of core count rather than the number of servers. This comparison is also helpful when studying applications that have per-core licensing costs. For example, the PowerEdge R820 needs double the number of ANSYS Fluent licenses for each server (32) when compared to the 16 needed for a PowerEdge M620 or M420. For all tests, the cores in the server were fully subscribed. For example, a 32-core result indicates that the test used two PowerEdge M620 (2*16 cores/server), one PowerEdge R820, and two PowerEdge M420s. All application results in this section are plotted relative to the performance on the PowerEdge M620 cluster.

Before jumping into application performance, the obvious differences in the memory subsystem of the three server platforms are studied first. The impact each server's architecture has on system memory bandwidth is demonstrated at a micro benchmark level using the Stream benchmark [6]. Subsequent sections analyze and explain the application level performance.

4.1. Memory bandwidth

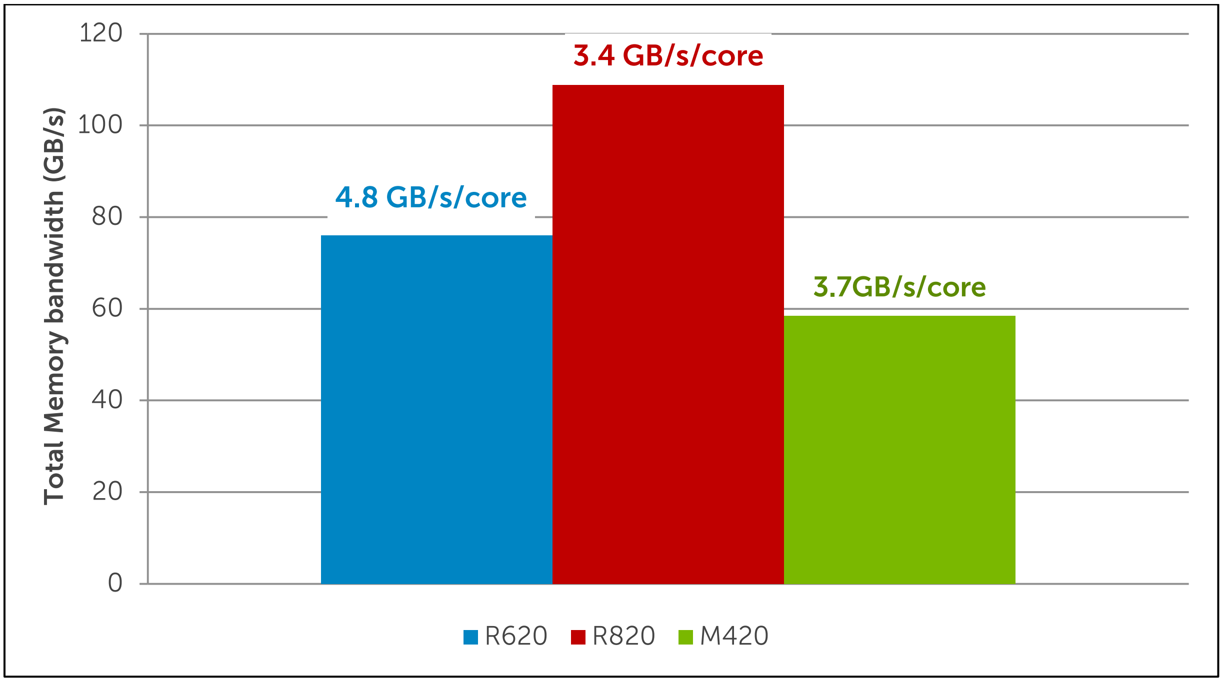
The memory bandwidth and memory bandwidth per core for the three platforms measured using the Stream benchmark is plotted in Figure 4. The height of the bar indicates the total memory bandwidth of the system. The value above each bar marks the memory bandwidth per core. The Dell PowerEdge R620 is a rack based server with a similar architecture and expected performance as the PowerEdge M620 blade server.

As expected, the PowerEdge R820 has the maximum total memory bandwidth measured at ~110GB/s. The corresponding bandwidth for the 2 socket PowerEdge R620 is 78GB/s. The Stream Triad benchmark performs two reads, and one write to memory. If additional data is transferred to/from memory during this benchmark measurement period, it is not counted towards the total memory bandwidth capability. Therefore, the memory bandwidth available to certain applications may be higher than reported by Stream. On the Intel Xeon processor E5-4600 product family, an issued non-cacheable write instruction still triggers a read for ownership due to the cache coherency protocol. This extra read is not counted when running the benchmark but takes memory bandwidth to accomplish. This is explained in more detail in [7]. If this extra read was counted by the benchmark, the effective memory bandwidth of the PowerEdge R820 would be approximately two times that of the PowerEdge R620.

This study uses the actual measured memory bandwidth as reported by Stream. An application may have the same behavior and incur the same RFO penalty. This measured value provides a baseline for the analysis.

At 4.8GB/s per core, the PowerEdge R620 has the highest memory bandwidth per core whereas the memory bandwidth per core on the PowerEdge R820 is measured to be ~30 percent lower. Because the PowerEdge M420 has three memory channels when compared to PowerEdge R620 or PowerEdge R820 that have four memory channels, the total memory bandwidth is ~22 percent lower and the memory bandwidth per core is ~25 percent lower than the PowerEdge R620.

Figure 4. Memory bandwidth



* BIOS options: System Profile is set to Max Performance and C states and C1E are disabled.

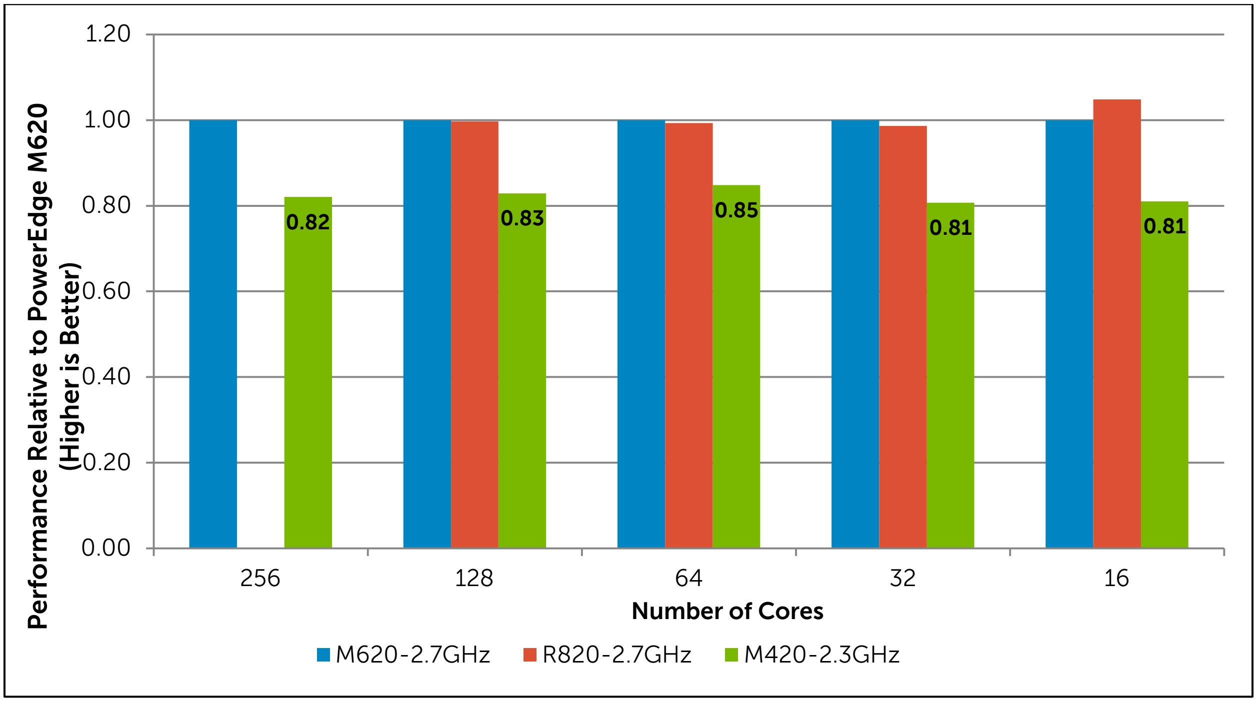
4.2. HPL

Moving on to the application level performance, High performance Linpack (HPL), a popular computationally intensive application is analyzed first. The problem size used for all the runs is maintained at 90 percent of the entire memory as described in Table 3. Note that this translates to different problem sizes (values of “N”) for each configuration.

The results are plotted in Figure 5. From the figure it is clear that the Dell PowerEdge M620 and the PowerEdge R820 perform similarly when comparing the same number of cores. This performance is attributed to the similar core frequency and memory frequency of these two configurations. Clearly HPL is not affected by the difference in memory bandwidth at these core counts. HPL also scales well and the interconnect is not a bottleneck at these core counts. This is apparent from the graph because the number of PowerEdge R820 servers needed to achieve a certain core count is half that of the PowerEdge M620 servers, but the performance of both clusters is similar.

The PowerEdge M420s perform consistently lower than the M620s by ~15 to 19 percent irrespective of core count. The difference in core frequency between the PowerEdge M420 (2.3 GHz) and the PowerEdge M620 (2.7 GHz) is 15 percent. The PowerEdge M420 also has a lower total memory configuration, and uses InfiniBand FDR10, which is slower than the InfiniBand FDR used in the PowerEdge M620s. This explains the consistent lower performance of the PowerEdge M420s.

Figure 5. HPL performance

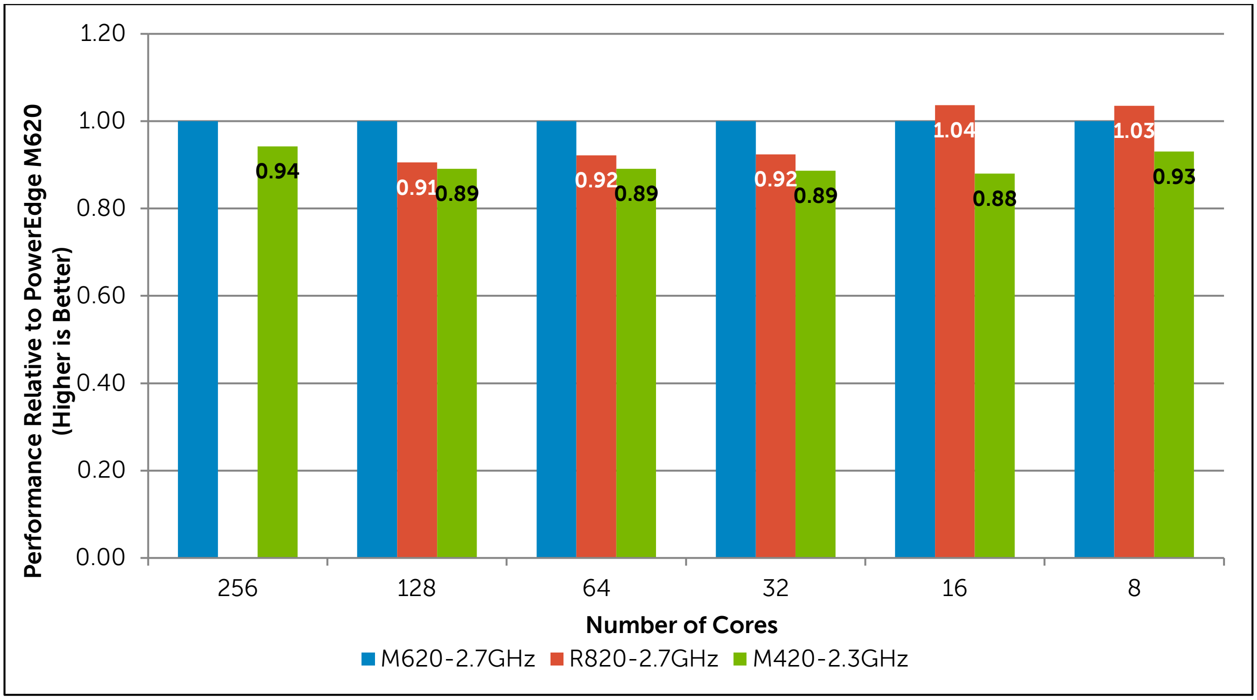


4.3. LU

Figure 6 presents the performance of the LU benchmark from the NAS Parallel Benchmarks (NPB) suite on the three clusters. When the servers are fully subscribed, the Dell PowerEdge M620 performs ~8 percent better than the PowerEdge R820 and ~6 to ~12 percent better when compared to the PowerEdge M420. From a previous study analyzing the various memory configurations on Dell PowerEdge 11th generation servers [8], a 16 percent drop in measured memory bandwidth led to a 2 percent drop in LU performance. This indicates that LU is not a memory intensive application. The PowerEdge R820 has a single QPI link connecting the sockets whereas the PowerEdge M620 has two QPI links. The extent of intra-node communication is higher on the PowerEdge R820 because of the higher core count. Recall that there are no crosslinks between sockets zero and two on the PowerEdge R820 and thus the messages need to traverse two QPI links for any communication as described in Figure 3. The difference in this QPI bandwidth can be associated with the lower performance on the PowerEdge R820. However, the value of the PowerEdge R820 is that a fewer number of servers are needed to achieve a certain performance or core count because this is a quad-socket system.

The performance drop on the PowerEdge M420 when compared to the PowerEdge M620 can be attributed to the 15 percent lower clock speed, single QPI link, and lower memory configuration.

Figure 6. LU performance



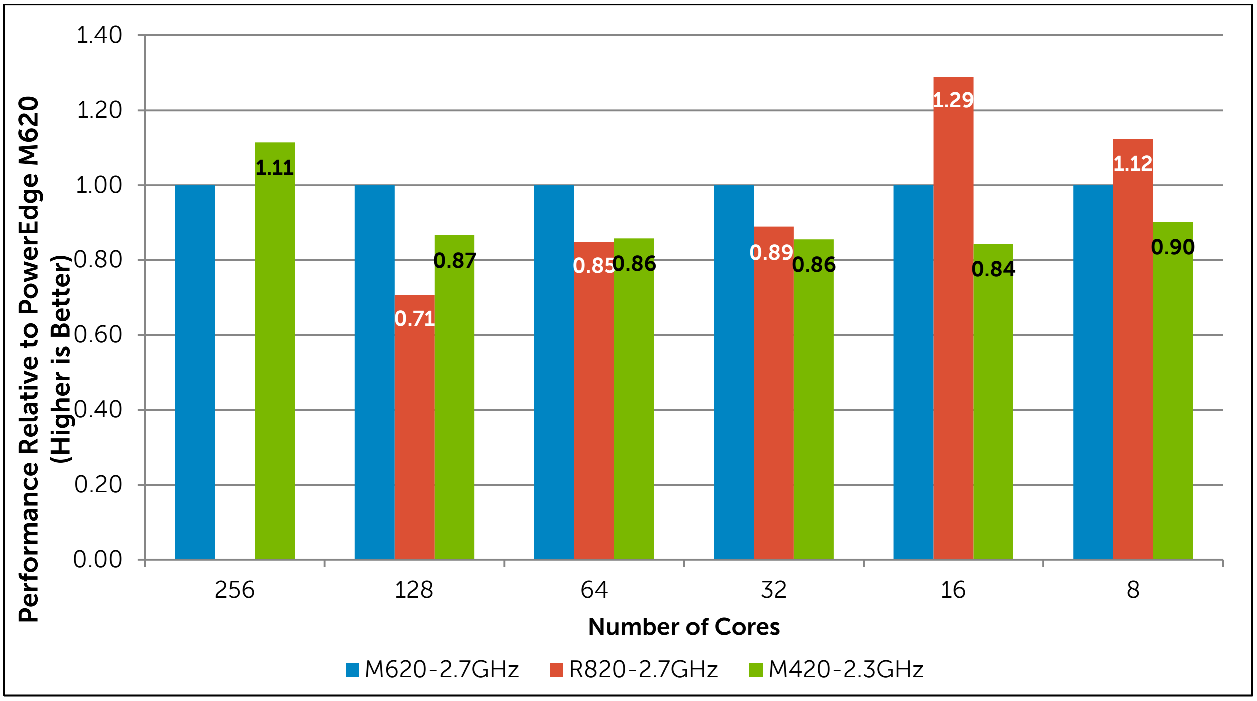
4.4. WRF

The performance of WRF on the three clusters is shown in Figure 7. The Dell PowerEdge M620 performs increasingly better than the PowerEdge R820 as the cluster size increases. From a previous study of the impact of memory bandwidth [8], it is observed that a 16 percent drop in memory bandwidth translates to a 4 percent drop in WRF performance. In this case, the drop in memory bandwidth per core of the PowerEdge R820 when compared to the PowerEdge M620 is ~30 percent. Thus, a portion of this performance drop on the PowerEdge R820s can be attributed to the lower memory bandwidth per core. However, for a fixed cluster size, fewer PowerEdge R820 servers will be needed to achieve this level of performance since it is a four socket system.

WRF is impacted by the difference in processor frequency on the PowerEdge M420. The PowerEdge M420 cluster performs consistently lower than the PowerEdge M620s by a factor of ~15 percent until 128 cores. There is a 23 percent drop in memory bandwidth per core on the PowerEdge M420 when compared to the PowerEdge M620, which also contributes to this drop in performance.

At 256 cores, the PowerEdge M420s perform 11 percent better than the PowerEdge M620s. This data point is repeatable and is not explained by the difference in the InfiniBand network architecture between these two clusters. At the time of writing, this aspect was still under study.

Figure 7. WRF performance



4.5. ANSYS Fluent

Two benchmark datasets, truck_poly_14m and truck_111m, are used for performance evaluation with ANSYS Fluent. ANSYS Fluent is not sensitive to memory bandwidth but is sensitive to core speed. From Figure 8, the Dell PowerEdge M420 performs 9 to 12 percent less than the PowerEdge M620. This can be attributed to the 15 percent drop in CPU frequency when compared to the PowerEdge M620.

For the same dataset, the PowerEdge R820 performs 7 to 11 percent less than the PowerEdge M620 when the server is fully-subscribed. The difference in the number of QPI links, the number of turbo bins and the turbo headroom available are possible factors that contribute to the drop in performance on the PowerEdge R820. The 30 percent drop in memory bandwidth per core on the PowerEdge R820 when compared to the PowerEdge M620 is not considered a significant factor since ANSYS Fluent is not memory bandwidth sensitive.

Figure 9 compares the performance among the three clusters when using the truck_111m benchmark dataset. The minimum amount of memory required for the benchmark is 128GB. The PowerEdge M620 and PowerEdge M420 were configured with 64GB each. Allowing some space for the operating system, a minimum of three servers is needed to run this benchmark data set. Similarly, a minimum of two PowerEdge R820 servers (configured with 128GB memory each) is required to run truck_111m. Thus the first data point plotted is with 64 cores and not 8, 16, or 32 cores. The PowerEdge R820 performs ~8 percent lower than the PowerEdge M620 and the PowerEdge M420 perform 8 to 11 percent lower than the PowerEdge M620s. The trends and analysis are similar to the ones observed with the truck_poly_14m data set.

Figure 8. ANSYS Fluent performance – truck_poly_14m

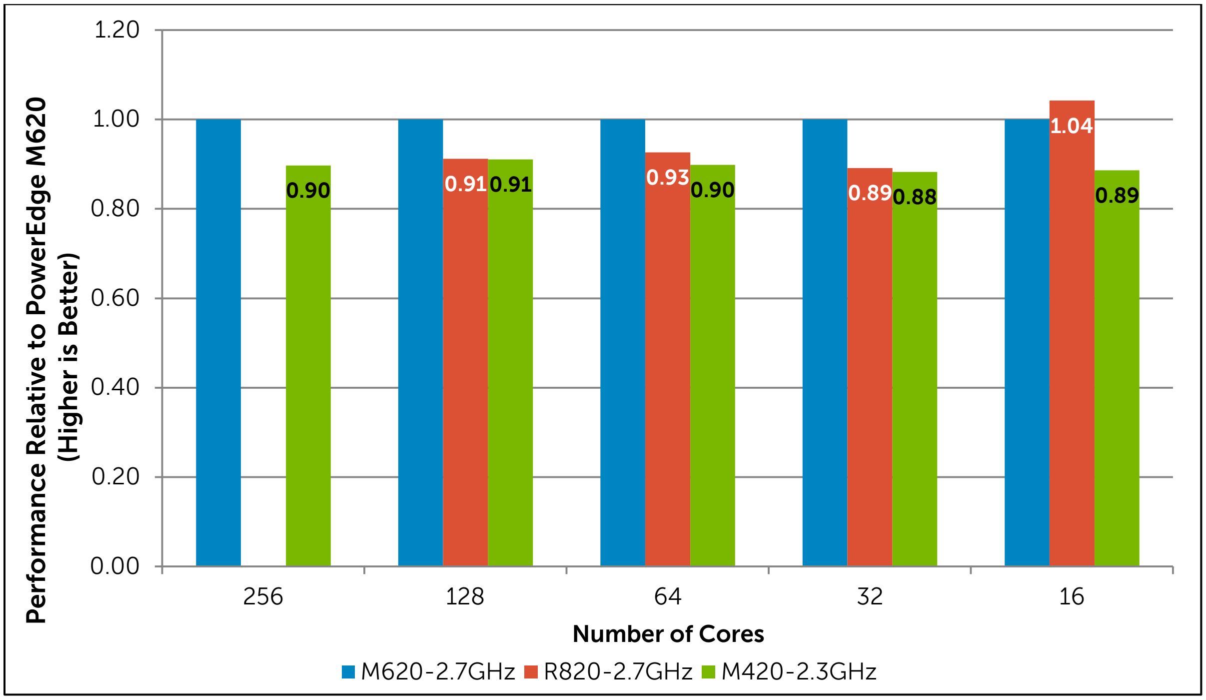
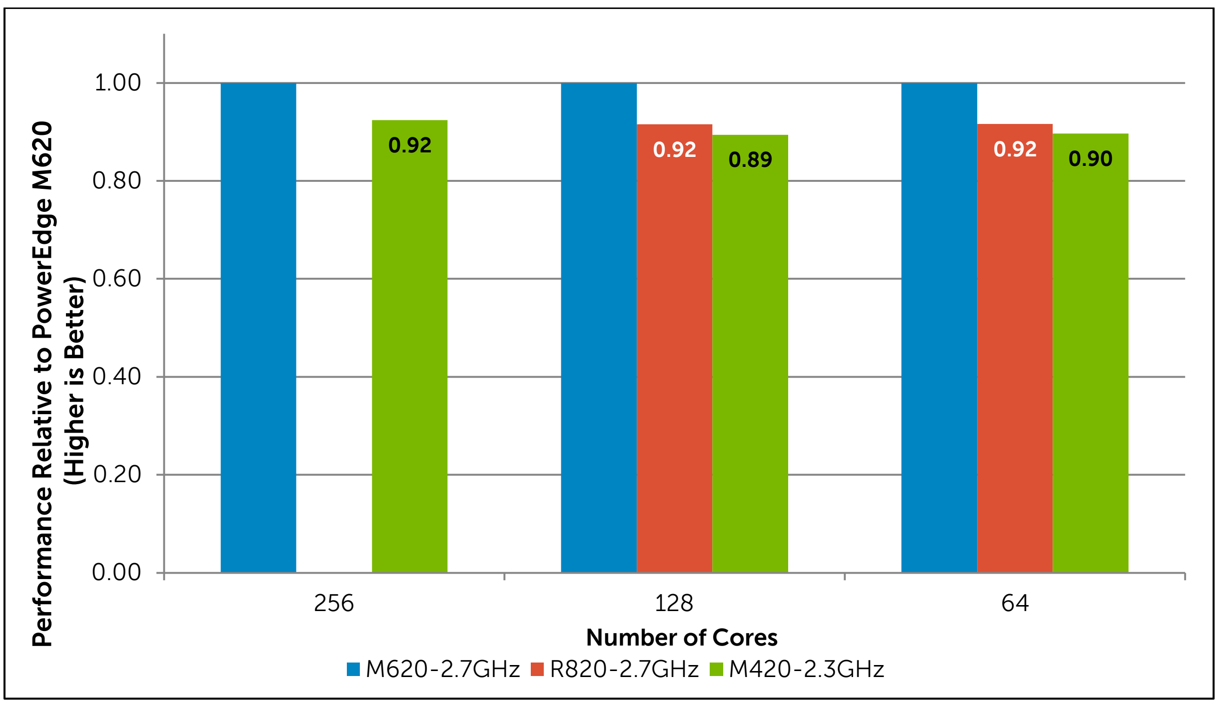


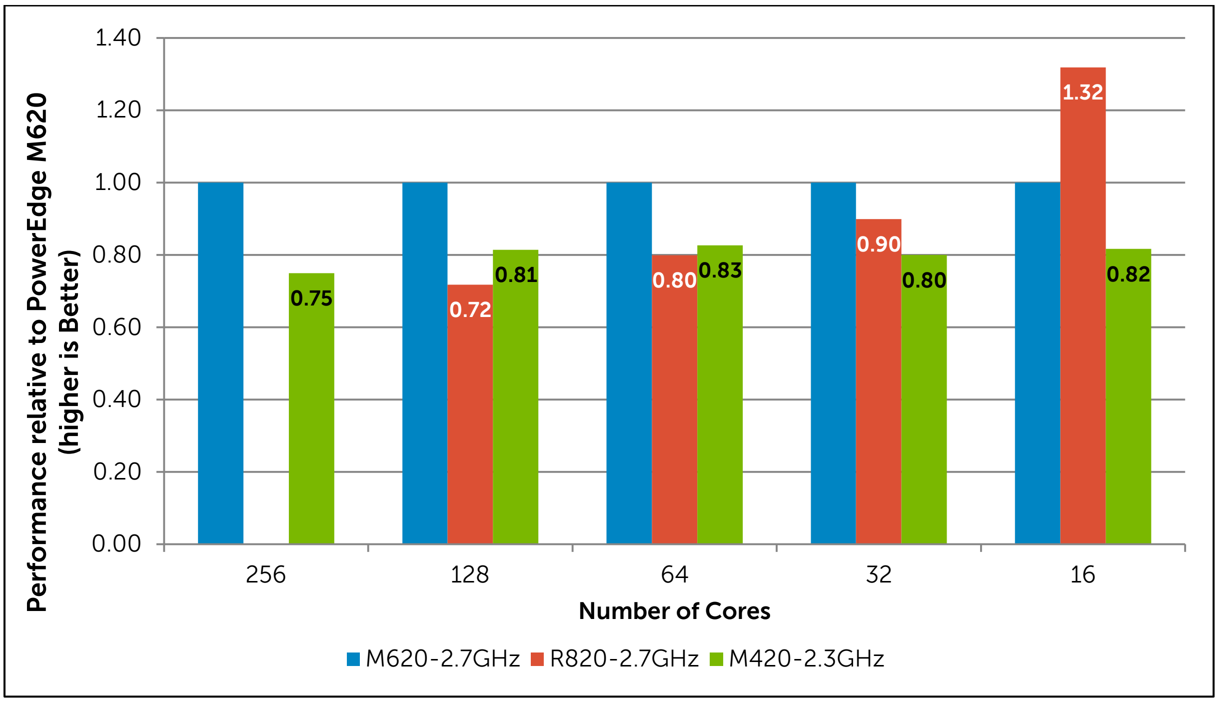
Figure 9. ANSYS Fluent performance – truck_111m



4.6. MILC

Figure 10 illustrates that as the core count increases the Dell PowerEdge M620 outperforms the PowerEdge M420 and the PowerEdge R820 when running the MILC application. MILC is sensitive to memory bandwidth while core speed and Turbo Boost do not contribute to the difference in performance in this scenario. The memory bandwidth per core is ~30 percent lower on the PowerEdge R820 when compared to the PowerEdge M620 and is ~25 percent lower on the PowerEdge M420 when compared to the PowerEdge M620. When fully subscribed, the PowerEdge M420 performs 21 to 33 percent less than the PowerEdge M620 and the PowerEdge R820 performs 11 to 39 percent less than the PowerEdge M620s. However, for a fixed cluster size, fewer PowerEdge R820 servers will be needed because it is a four socket system.

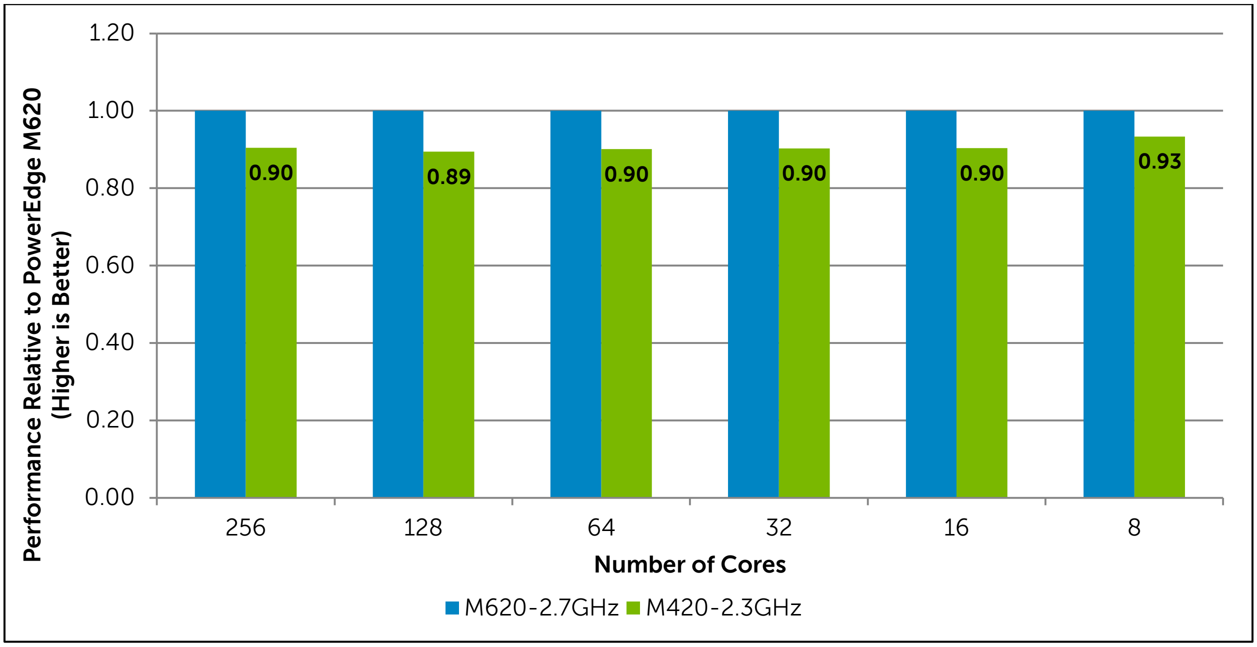
Figure 10. MILC performance



4.7. NAMD

Figure 11 plots NAMD performance on the three clusters. From the figure it is observed that the Dell PowerEdge M420 performs consistently lower than the PowerEdge M620 by 10 percent. NAMD is not memory bandwidth sensitive and the delta in performance can be attributed to the 15 percent drop in core frequency. There were issues running NAMD fully subscribed on more than one server on the PowerEdge R820 platform. The authors are working towards resolution with the NAMD developers.

Figure 11. NAMD performance



4.8. Power consumption and energy efficiency

This section discusses the power consumption and the energy efficiency of the three clusters described in Test bed and applications. Energy efficiency is used as the metric for power comparisons and is computed as performance obtained for each watt of power consumed.

The Dell PowerEdge M1000e chassis supports multiple server blades and includes shared infrastructure such as fans, power supplies, management modules and switches. The power consumption of the chassis includes the power consumed by these shared resources. Power measurements for a smaller number of servers that do not fully make use of the chassis would not be a fair measure of the actual power consumed. Therefore, only a fully populated PowerEdge M1000e blade chassis with 16 PowerEdge M620 servers or 32 PowerEdge M420 servers is used for this portion of the study. For the PowerEdge R820 cluster, the power consumed by four PowerEdge R820s (128 cores) and the InfiniBand switch are taken into consideration. The power consumed by the Gigabit Ethernet management switch is not taken into account.

Note that the energy efficiency data presented here is **not** an apples-to-apples comparison. First, each type of cluster has a different total core count; for example, different number of servers in the cluster. Additionally, the results compare rack servers to blade servers. These results are an attempt to extract trends and report measured results. An ideal comparison would compare similar types of servers and identical number of cores in each of the three clusters.

Figure 12 presents the energy efficiency of HPL. The metric used here is GFLOPS/Watt. The PowerEdge R820 cluster with 128 cores has 4 percent better efficiency than the 256 core PowerEdge M620 cluster. This can be attributed to the lower core count of the PowerEdge R820 cluster; lower total performance, but lower power consumption as well. The 512 core PowerEdge M420 cluster is 15 percent more energy efficiency than the 256 core PowerEdge M620 cluster. This is due to two factors - double the core count on the PowerEdge M420 cluster boosts the performance. Additionally, the lower wattage processors and the lower DIMM configuration on the PowerEdge M420 contribute to the lower power consumption.

Figure 12. HPL performance/Watt

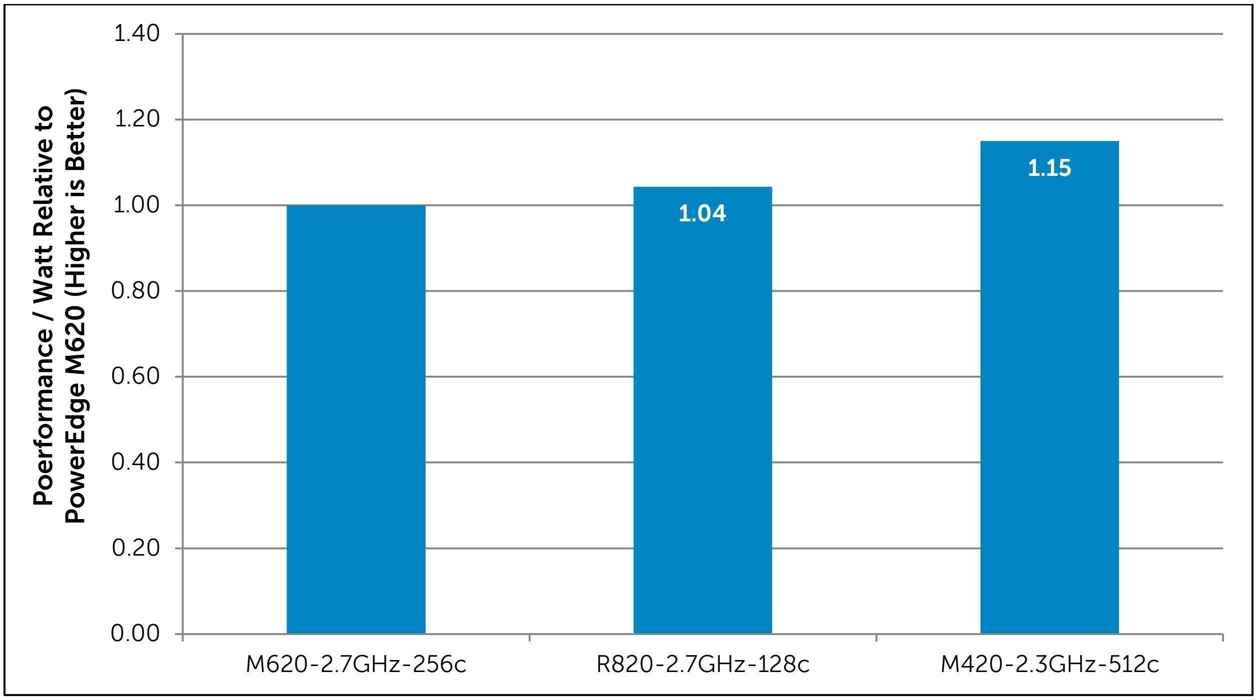


Figure 13. LU performance/Watt

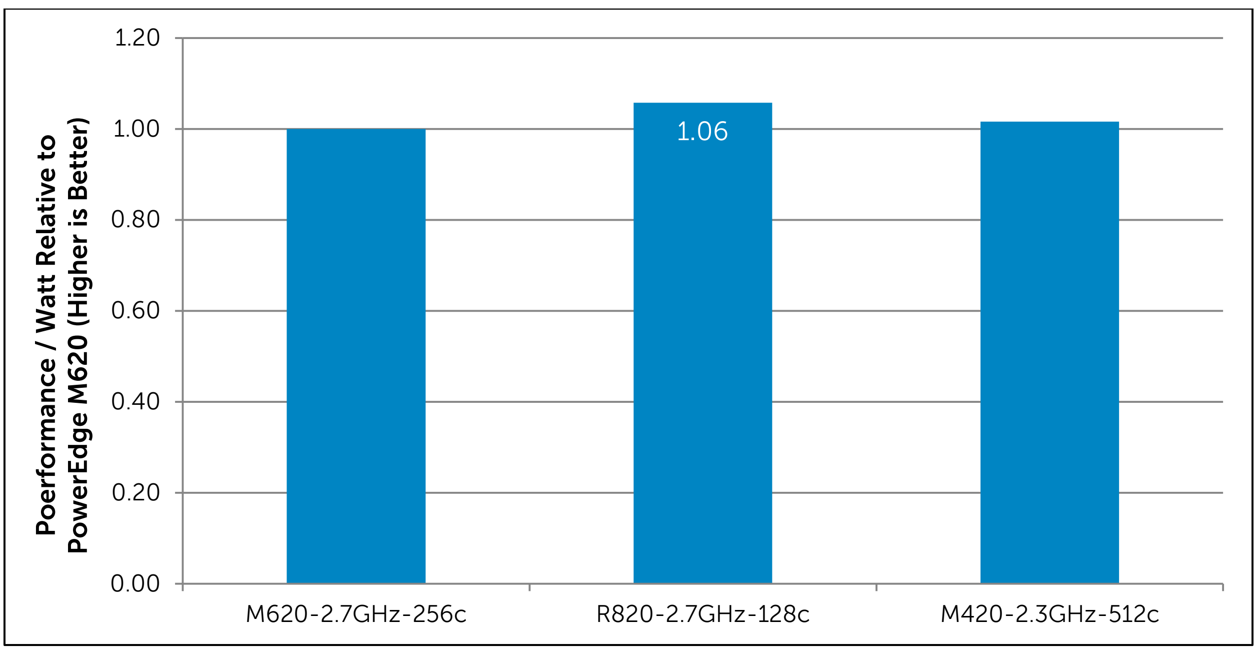


Figure 13 plots the energy efficiency of the three clusters when executing the LU benchmark. The observation is that the PowerEdge M620 cluster and the PowerEdge M420 cluster have similar energy efficiencies. The differences in the values are within the statistical variation. The PowerEdge R820 cluster shows slightly better energy efficiency than the other two clusters because of the lower core count. The metric used here is rating/Watt, which translates to (number of jobs of LU which can be run during a period of one day)/power consumed in Watts.

Figure 14 shows the energy efficiency of WRF. The PowerEdge M420 cluster has 13 percent better energy efficiency when compared to the PowerEdge M620 cluster. The performance metric used here is average time step/Watt. The lower wattage processors and the lower memory configuration attribute to this improvement in energy efficiency on the PowerEdge M420 cluster.

The PowerEdge R820 is 12 percent less energy efficient when compared to the PowerEdge M620. The performance of the 128 core PowerEdge R820 cluster is lower than half of the 256 core PowerEdge M620 cluster's performance. But the power consumed by the PowerEdge R820 cluster is higher than half the power taken by the PowerEdge M620 blades. This contributes to the 12 percent drop in energy efficiency.

Figure 14. WRF performance/Watt

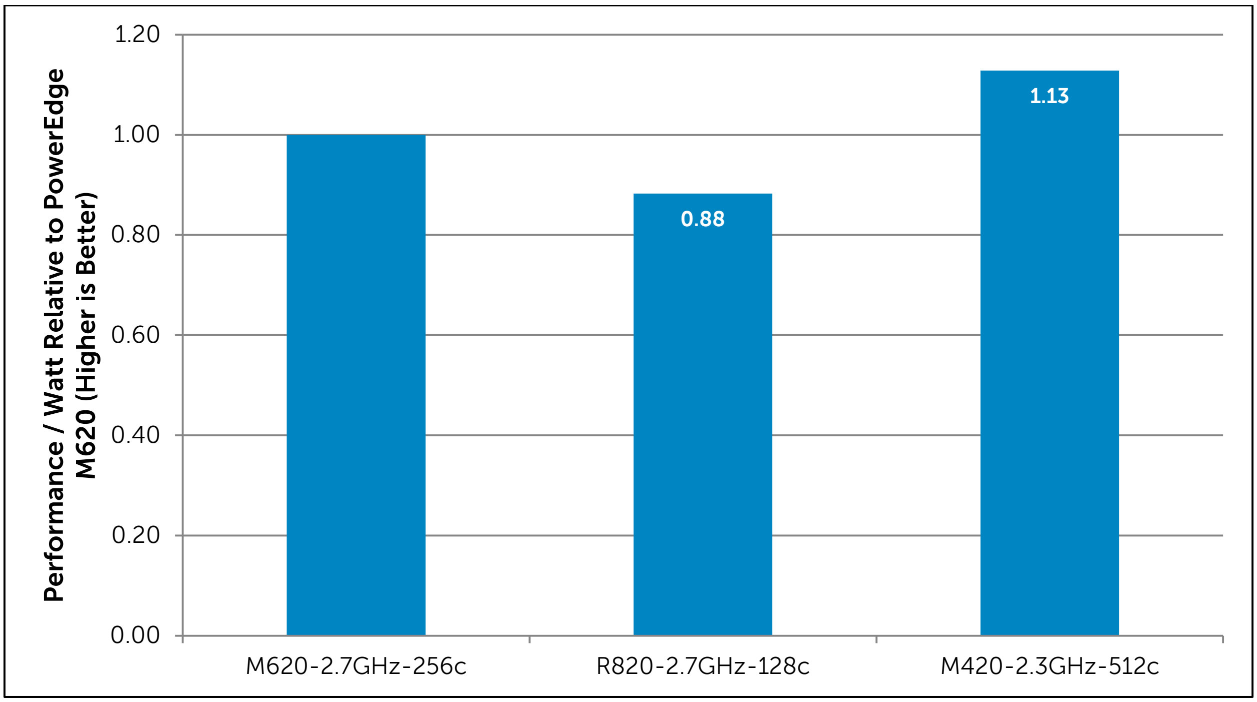


Figure 15 shows the energy efficiency with ANSYS Fluent. The metric used here is rating per watt. The PowerEdge R820 cluster has ~14 to 16 percent better energy efficiency than the PowerEdge M620 cluster. The performance is approximately half of the PowerEdge M620 cluster but the power consumption is accordingly lower too because of the lower total core count. The 512 core PowerEdge M420 cluster has similar energy efficiency compared to the 256 core PowerEdge M620 cluster for the truck_poly_14m dataset. Interestingly, the PowerEdge M420 cluster is 23 percent more energy efficient than the PowerEdge M620 cluster for the truck_111m data set.

Figure 15. ANSYS Fluent performance/Watt

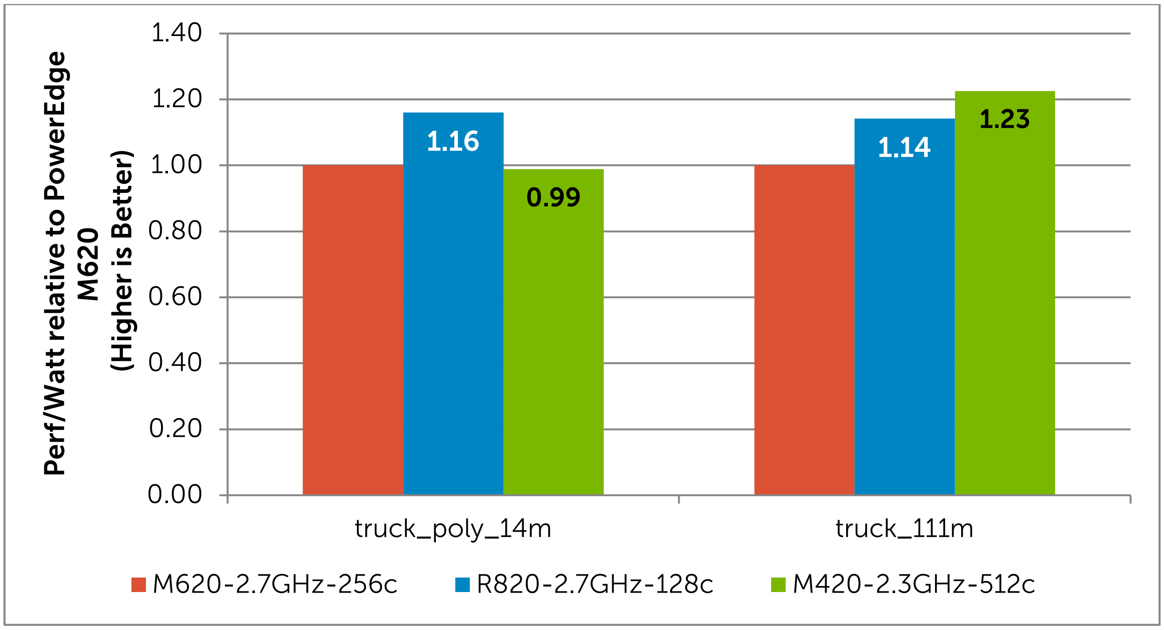


Figure 16 illustrates the energy efficiency of MILC. The metric used here is rating/Watt. The PowerEdge M620 cluster provides approximately double the performance for less than double the power consumed by the PowerEdge R820 cluster. Thus, the PowerEdge R820 cluster measures 16 percent lower energy efficiency than the PowerEdge M620 cluster. The 512 core PowerEdge M420 cluster provides double the performance for less than double the power consumption of the 256 core PowerEdge M620 cluster. This results in the 37 percent better energy efficiency of the PowerEdge M420 cluster over the PowerEdge M620 cluster. The shared infrastructure of the 32 node PowerEdge M420 chassis clearly provides a benefit.

Figure 16. MILC performance/ Watt

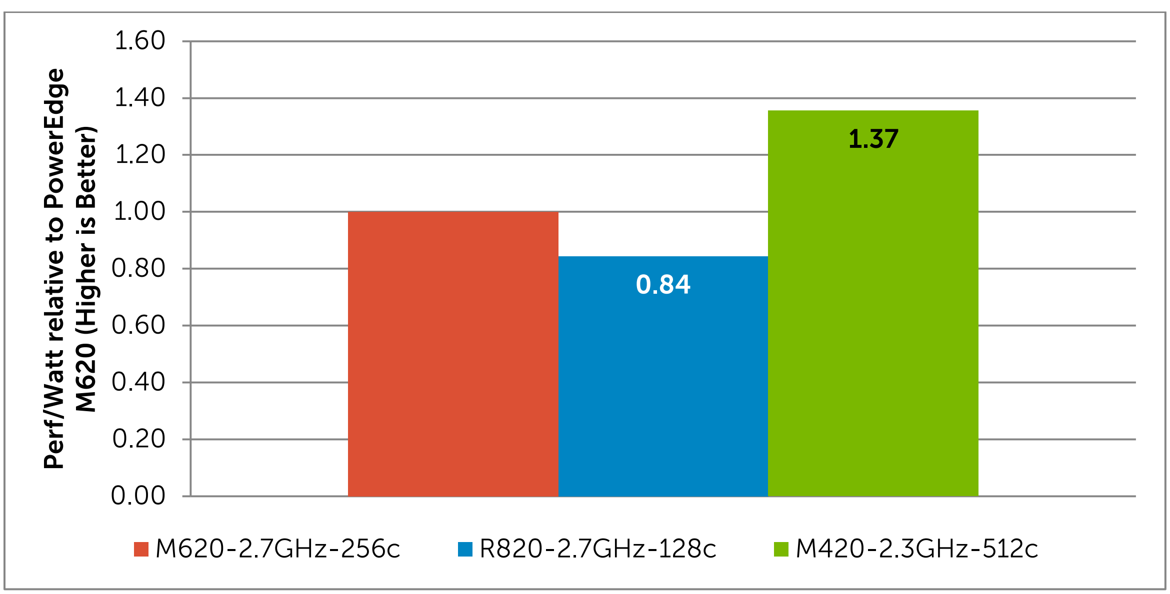
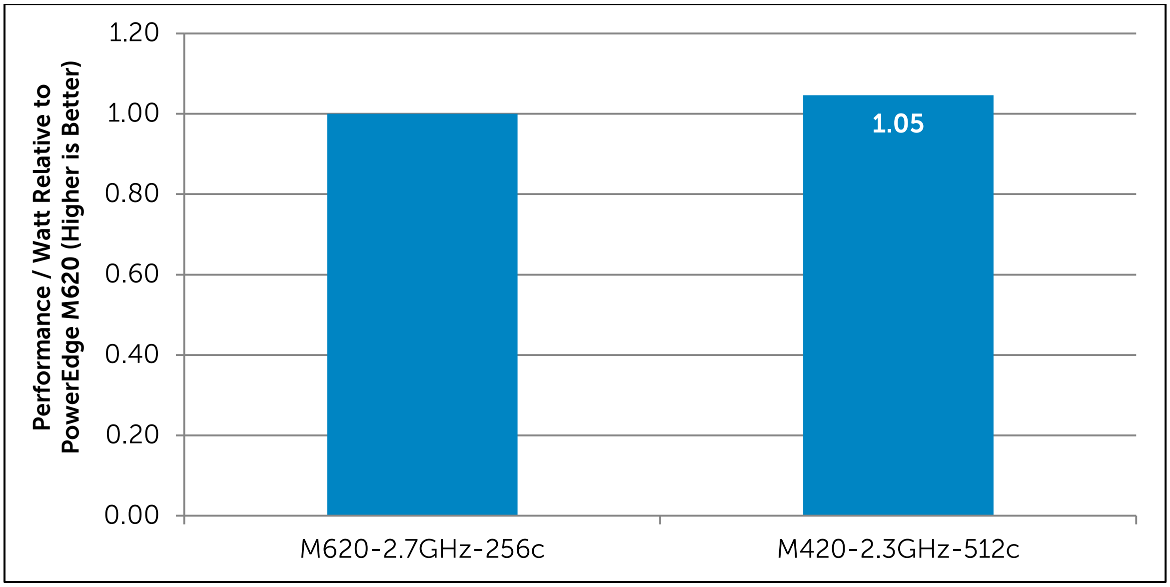


Figure 17 plots the energy efficiency of NAMD. The PowerEdge M420 cluster performs better in this scenario providing 5 percent better energy efficiency than the PowerEdge M620 cluster. As mentioned in NAMD, there were issues running NAMD on the PowerEdge R820 cluster; therefore, that data point is missing in this graph.

Figure 17. NAMD performance/ Watt



5. Conclusion

This white paper provides a quantitative analysis of the performance and energy efficiency of three different Intel Xeon E5 Family (Sandy Bridge) based clusters using a sample of HPC workloads. Table 4 and Table 5 summarize the performance and energy efficiency characteristics of each HPC application on the three types of clusters. They present the results of this study in an easy to read format. The Dell PowerEdge M620 cluster is used as the baseline for performance and energy efficiency comparisons. Clusters of the same size, for example, with the same number of cores are used for performance comparisons. As described in Power consumption and energy efficiency, power and energy efficiency measurements were conducted on clusters of different sizes. Table 5 explicitly notes the cluster size used for power measurements for easy reading. Details of the test bed were provided in Test bed and applications.

Table 4. Results summary - performance

Application	PowerEdge M620 @ 2.7 GHz (baseline)	PowerEdge R820 @ 2.7 GHz	PowerEdge M420 @ 2.3 GHz
HPL	Similar performance to PowerEdge R820	Similar performance to PowerEdge M620	Performance: 17% lower
LU	Best Performance	Performance: ~8% lower	Performance: ~11% lower
WRF	Best Performance	Performance: up to ~20% lower	Performance: up to 9% lower
ANSYS Fluent	Best performance	Performance: ~8% lower	Performance: ~10% lower
MILC	Best performance	Performance: ~39% lower	Performance: ~33% lower
NAMD	Best performance	(Issues when executing program)	Performance: ~9% lower

* The PowerEdge M620 cluster is used as the baseline for performance comparisons. Higher is better.

Table 5. Results summary – energy efficiency (EE)

Application	PowerEdge R820 @ 2.7 GHz 4 servers, 128 cores	PowerEdge M420 @ 2.3 GHz 32 servers, 512 cores
HPL	EE: 4% higher	Best EE: 15% higher
LU	Best EE: ~6% higher	Similar EE to PowerEdge M620
WRF	EE: 12% lower	Best EE: ~13% higher
ANSYS Fluent	Best EE for truck_poly_14m (+16%)	Best EE for truck_111m (+23%)
MILC	EE: ~18% lower	Best EE: ~37% higher
NAMD	(Issues when executing program)	Better EE: ~5 % higher

* EE is shorthand for energy efficiency. The PowerEdge M620 cluster is used as the baseline for comparison. Higher is better for EE.

From an engineering and design perspective, performance and energy efficiency considerations are important to best-fit a cluster to an application's requirements. However other factors do influence the final decision. These include total cost of ownership aspects that differ from data center to data center like the total number of servers, number of switches, power and cooling availability, ease of administration and cost.

6. References

1. Optimal BIOS settings for HPC with Dell PowerEdge 12th generation servers
   http://www.dellhpcsolutions.com/assets/pdfs/Optimal_BIOS_HPC_Dell_12G.v1.0.pdf

2. Mellanox FDR10 Product line description
   http://www.mellanox.com/pdf/products/oem/RG_IBM.pdf

3. Intel® Xeon® Processor E5 Family (Servers)
   http://ark.intel.com/products/family/59138/Intel-Xeon-Processor-E5-Family/server

4. Intel® Xeon® Processor E5 Family Specifications
   https://www-ssl.intel.com/content/www/us/en/processors/xeon/xeon-e5-family-spec-update.html

5. StackIQ Rocks+
   http://www.stackiq.com/

6. Stream benchmark for memory bandwidth
   https://www.cs.virginia.edu/stream/

7. Measuring Memory Bandwidth on Intel® Xeon® Processor 7500 Platform (RFO penalty)
   https://www-ssl.intel.com/content/www/za/en/benchmarks/resources-xeon-7500-measuring-memory-bandwidth-paper.html

8. Memory Selection Guidelines for High Performance Computing with Dell™ PowerEdge™ 11G Servers
   http://i.dell.com/sites/content/business/solutions/whitepapers/en/Documents/11g-memory-selection-guidelines.pdf