Performance study of four Socket PowerEdge R940 Server with Intel Skylake processors

Author: Somanath Moharana, Dell EMC HPC Innovation Lab, August 2017

This blog explores the performance of the four socket Dell EMC PowerEdge R940 server with Intel Skylake processors. The latest Dell EMC 14th generation servers support the new Intel® Xeon® Processor Scalable Family (processor architecture codenamed "Skylake"), and the increased number of cores and higher memory speed benefit a wide variety of HPC applications.

The PowerEdge R940 is Dell EMC's latest 4-socket, 3U rack server designed to run complex workloads, which supports up to 6TB of DDR4 memory and up to 122 TB of storage. The system features the Intel® Xeon® Scalable Processor Family, 48 DDR4 DIMMs, up to 13 PCI Express® (PCIe) 3.0 enabled expansion slots and a choice of embedded NIC technologies. It is a general-purpose platform capable of handling demanding workloads and applications, such as data warehouses, ecommerce, databases, and high-performance computing (HPC). With the increase in storage capacity the PowerEdge R940 makes it well-suited for data intensive applications that require greater storage.

This blog also describes the impact of BIOS tuning options on HPL, STREAM and scientific applications ANSYS Fluent and WRF and compares performance of the new PowerEdge R940 to the previous generation PowerEdge R930 platform. It also analyses the performance with Sub NUMA Cluster (SNC) modes (SNC=Enabled and SNC=Disabled). SNC enabled will expose eight NUMA nodes to the OS on a four socket PowerEdge R940. Each NUMA node can communicate with seven other remote NUMA nodes, six in other three sockets and one within same socket. NUMA domains on different sockets communicate over the UPI interconnect. Please visit BIOS characteristics of Skylake processor-blog for more details on BIOS options. Table 1 lists the server configuration and the application details used for this study.

Table 1: Details of Server and HPC Applications used for R940 analysis

Platform	**PowerEdge R930**	**PowerEdge R930**	**PowerEdge R940**
Processor	4 x Intel Xeon E7-8890 v3@2.5GHz (18 cores) 45MB L3 cache 165W Codename=Haswell-EX	4 x Intel Xeon E7-8890 v4@2.2GHz (24 cores) 60MB L3 cache 165W Codename=Broadwell-EX	4 x Intel Xeon Platinum 8180@2.5GHz, 10.4GT/s (Cross-bar connection) Codename=Skylake
Memory	1024 GB = 64 x 16GB DDR4 @1866MHz	1024 GB = 32 x 32GB DDR4 @1866 MHz	384GB = (24 x 16GB) DDR4@2666MT/s
CPU Interconnect	Intel QuickPath Interconnect (QPI) 8GT/s	Intel QuickPath Interconnect (QPI) 8GT/s	Intel Ultra Path Interconnect (UPI) 10.4GT/s
BIOS Settings			

BIOS	Version 1.0.9	Version 2.0.1	Version 1.0.7
Processor Settings > Logical Processors	Disabled	Disabled	Disabled
Processor Settings > UPI Speed	Maximum Data Rate	Maximum Data Rate	Maximum Data Rate
Processor Settings > Sub NUMA cluster	N/A	N/A	Enabled, Disabled
System Profiles	PerfOptimized (Performance), PerfPerWattOptimizedDapc (DAPC),	PerfOptimized (Performance), PerfPerWattOptimizedDapc (DAPC),	PerfOptimized (Performance), PerfPerWattOptimizedDapc (DAPC),)
Software and Firmware			
Operating System	Red Hat Enterprise Linux Server release 6.6	Red Hat Enterprise Linux Server release 7.2	Red Hat Enterprise Linux Server release 7.3 (3.10.0-514.el7.x86_64)
Intel Compiler	Version 15.0.2	Version 16.0.3	version 17.0.4
Intel MKL	Version 11.2	Version 11.3	2017 Update 3
Benchmark and Applications			
LINPACK	V2.1 from MKL 11.3	V2.1 from MKL 11.3	V2.1 from MKL update 3
STREAM	v5.10, Array Size 1800000000, Iterations 100	v5.10, Array Size 1800000000, Iterations 100	v5.4, Array Size 1800000000, Iterations 100
WRF	V3.5.1 Input Data Conus12KM, Netcdf-4.3.1	V3.8 Input Data Conus12KM, Netcdf-4.4.0	v3.8.1, Input Data Conus12KM, Conus2.5KM, Netcdf-4.4.2
ANSYS Fluent	v15, Input Data: truck_poly_14m	v16, Input Data: truck_poly_14m	v17.2, Input Data: truck_poly_14m, aircraft_wing_14m, ice_2m, combustor_12m, exhaust_system_33m

Note: The software versions were different for the older generation processors and results are compared against what was best configuration at that time. Due to big architectural changes in servers and processors generation over generation, the changes in software versions is not a significant factor.

The High Performance Linpack (HPL) Benchmark is a measure of a system's floating-point computing power. It measures how fast a computer solves a dense n by n system of linear equations Ax = b, which is a common task in engineering. HPL performed with block size of NB=384 and problem size of N=217754. Since HPL is an AVX-512-enabled workload, we would calculate HPL theoretical maximum performance as (rated base frequency of processor * number of cores * 32 FLOP/second).

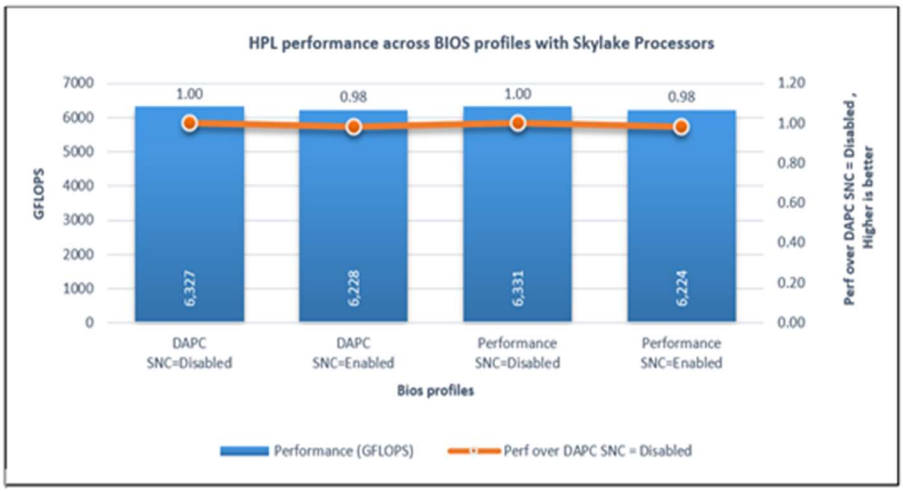

Figure 1: Comparing HPL Performance across BIOS profiles

Figure 1 depicts the performance of the PowerEdge R940 server described in Table 1 with different BIOS options. Here the “Performance SNC = Disabled” gives better performance compared to other bios profiles. With “SNC=Disabled” we can observe 1-2% better performance as compared to “SNC = Enabled” for all the BIOS profiles.

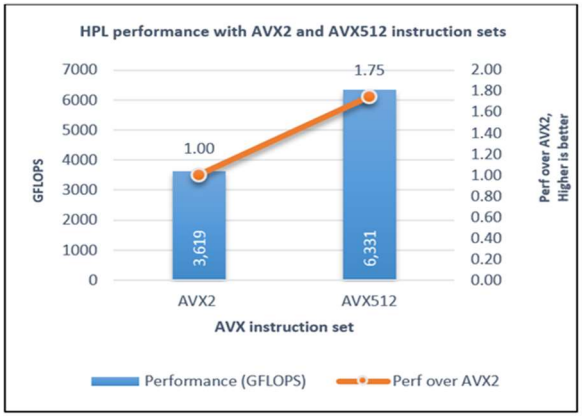

Figure 2: HPL performance with AVX2 and AVX512 instructions sets

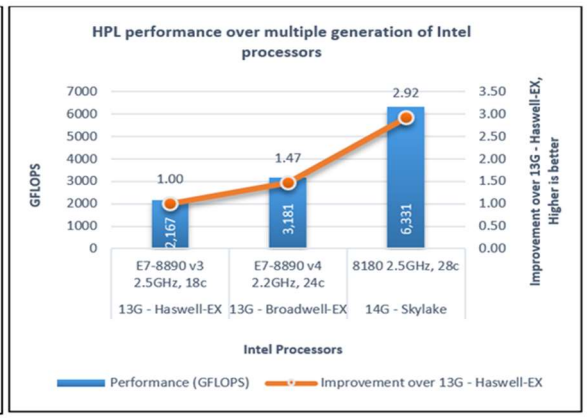

Figure 3: HPL Performance over multiple generations of processors

Figure 2 compares the performance of HPL ran with AVX2 instructions sets and AVX512 instructions sets on PowerEdge R940 (where AVX=Advanced Vector Instructions). AVX-512 are 512-bit extensions to the 256-bit AVX SIMD instructions for x86 instruction set architecture. By Setting "MKL_ENABLE_INSTRUCTIONS=AVX2/AVX512" environment variable we can set the AVX instructions. Here we can observe by running HPL with AVX512 instruction set gives around 75% improvement in performance compared to AVX2 instruction set.

Figure 3: Compares the results of four socket R930 powered by Haswell-EX processors and Broadwell-EX processors with R940 powered by Skylake processors. For HPL, R940 server performed ~192% better in comparison to R930 server with four Haswell-EX processors and ~99% better with Broadwell-EX processors. The performance improvement we observed in Skylake over Broadwell-EX is due to a 27% increase in the number of cores and 75% increase in performance for AVX 512 vector instructions.

The **Stream** benchmark is a synthetic benchmark program that measures sustainable memory bandwidth and the corresponding computation rate for simple vector kernels. The STREAM benchmark calculates the memory bandwidth by counting only the bytes that the user program requested to be loaded or stored. This study uses the results reported by the TRIAD function of the stream bandwidth test.

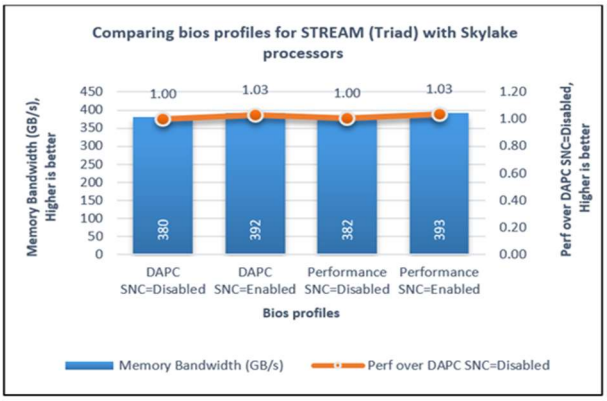

Figure 4: STREAM Performance across BIOS profiles

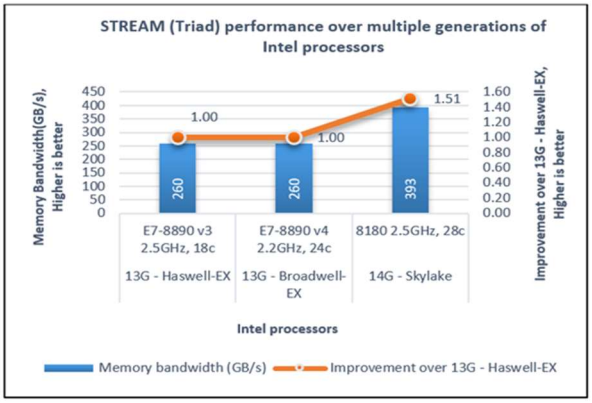

Figure 5: STREAM Performance over multiple generations of processors

As per Figure 4, With "SNC = Enabled" we are getting up to 3% better bandwidth in comparison to "SNC = Disabled" across all bios profiles. Figure 5, shows the comparison of memory bandwidth of PowerEdge R930 server with Haswell-EX, Broadwell-EX processors and PowerEdge R940 server with Skylake processors. Both Haswell-EX and Broadwell-EX support DDR3 and DDR4 memories respectively, while the platform with this configuration supports 1600MT/s of memory frequency for both generation of processors. Due to use of DIMMs of same memory frequency for both generation of processors on PowerEdge R930, both Broadwell-EX and Haswell-EX processors have same memory bandwidth but there is ~51% increase in memory bandwidth with Skylake compared to Broadwell- EX processors. This is due to use of 2666MHz RDIMMS, which gives around ~66% increase in maximum memory bandwidth compared to Broadwell-EX and the second factor is ~50% increase in the number of memory channels per socket which is 6 channels per socket for Skylake Processors and 4 channels per socket in Broadwell-EX processors.

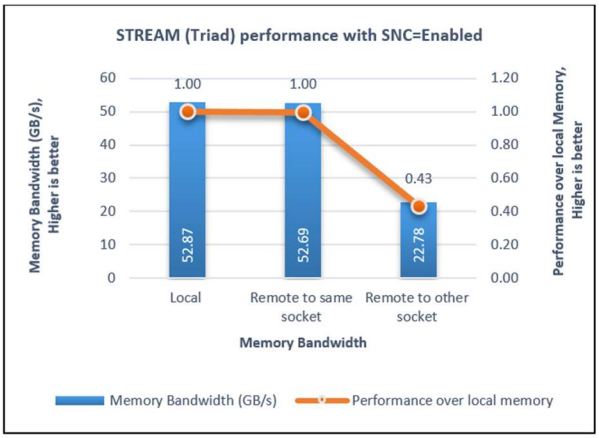

Figure 6: Comparing STREAM Performance with "SNC = Enabled"

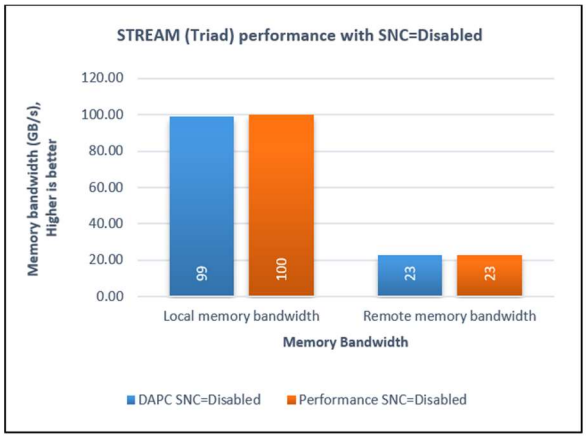

Figure 7: Comparing STREAM Performance with "SNC = Disabled

Figure 6 and Figure 7 describe the impact of traversing the UPI link to go across sockets on memory bandwidth for the PowerEdge R940 servers. With "SNC = Enabled" the local memory bandwidth and remote to same socket memory bandwidth is nearly same (0-1% variations) but in

case of remote to other socket, the memory bandwidth shows ~57% decrease in comparison to local memory bandwidth. With "SNC = Disabled" remote memory bandwidth is 77% lower compared to local memory bandwidth.

The **Weather Research and Forecasting (WRF)** Model is a mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting needs. It features two dynamical cores, a data assimilation system, and a software architecture facilitating parallel computation and system extensibility. The model serves a wide range of meteorological applications across scales from tens of meters to thousands of kilometers. WRF can generate atmospheric simulations using real data or idealized conditions. We used the CONUS12km and CONUS2.5km benchmark datasets for this study.

CONUS12km is a single domain and small size (48hours, 12km resolution case over the Continental U.S. (CONUS) domain from October 24, 2001) benchmark with 72 seconds of time step. CONUS2.5km is a single domain and large size (Latter 3hours of a 9hours, 2.5km resolution case over the Continental U.S. (CONUS) domain from June 4, 2005) benchmark with 15 seconds of time step.

WRF decomposes the domain into tasks or patches. Each patch can be further decomposed into tiles that are processed separately, but by default there is only one tile for every run. If the single tile is too large to fit into the cache of the CPU and/or core, it slows down computation due to WRF's memory bandwidth sensitivity. In order to reduce the size of the tile, it is possible to increase the number of tiles by defining "numtile = x" in input file or defining environment variable "WRF_NUM_TILES = x". For both CONUS 12km and CONUS 2.5km the number of tiles are chosen based on best performance.

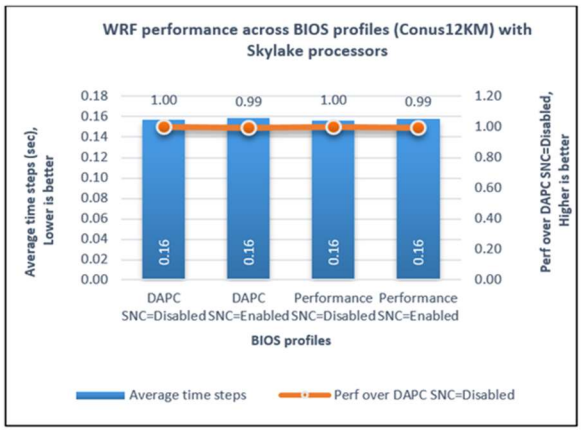

Figure 8: WRF Performance across BIOS profiles (Conus12KM)

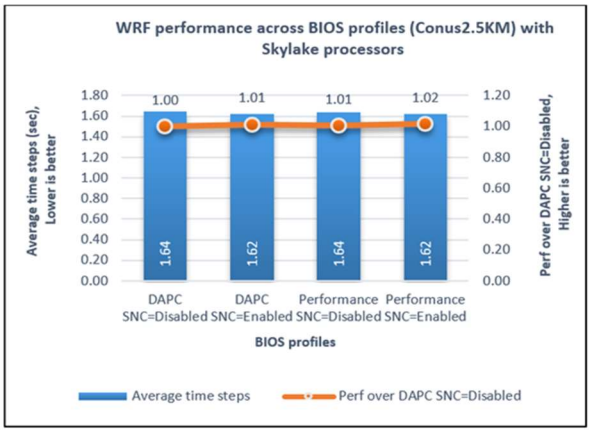

Figure 9: WRF Performance across BIOS profiles (Conus2.5KM)

Figure 8 and Figure 9 demonstrates the comparison of WRF datasets on different BIOS profiles. With Conus 12KM data and CONUS 2.5KM "Performance SNC = Enabled" gives best performance. For both conus12km and conus2.5km, the "SNC = Enabled" performs ~1%-2% better than "SNC = Disabled". The performance difference across different bios profile is nearly equal for Conus12Km as it uses a smaller dataset size, while in CONUS2.5km which is having

larger dataset we can observe 1-2% performance variations across the system profiles as it utilizes larger number of processors more efficiently.

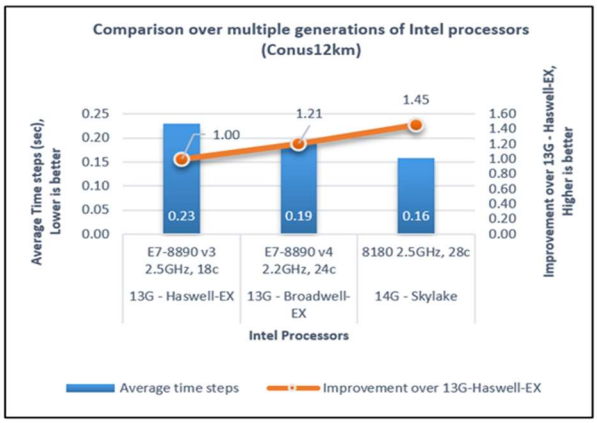

Figure 10: Comparison over multiple generations of processors

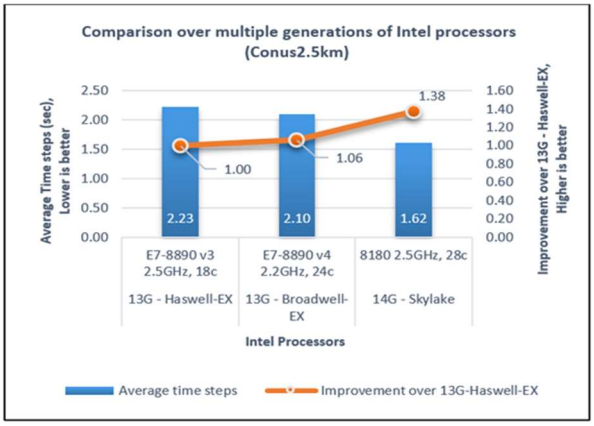

Figure 11: Comparison over multiple generations of processors

Figure 10 and Figure 11 shows the performance comparison between PowerEdge R940 powered by Skylake processors and PowerEdge R930 powered by Broadwell-EX processors and Haswell-EX processors. From the Figure 11 we can observe that for Conus12KM performance of PowerEdge R940 with Skylake is ~18% better as compared to PowerEdge R930 with Broadwell EX and ~45% better compared to Haswell-EX processors. In case of for Conus2.5 Skylake performs ~29% better than Broadwell-EX and ~38% better than Haswell-EX processors.

ANSYS Fluent is a computational fluid dynamics (CFD) software tool. Fluent includes well-validated physical modeling capabilities to deliver fast and accurate results across the widest range of CFD and multi physics applications.

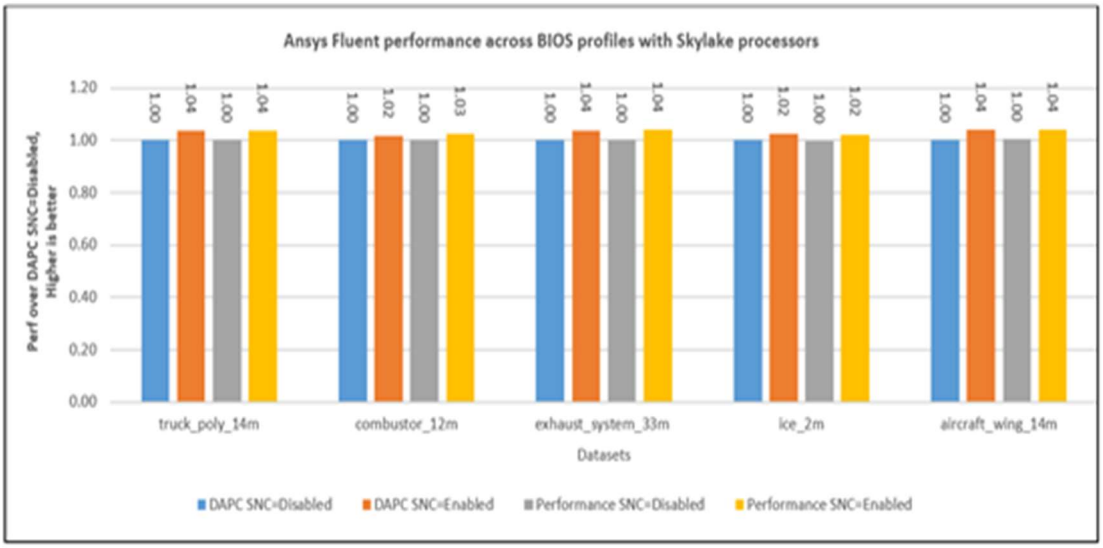

Figure 12: Ansys Fluent Performance across BIOS profiles

We used five different datasets for our analysis which are truck_poly_14m, combuster_12m, exhaust_system_33m, ice_2m and aircraft_wing_14m. Fluent with 'Solver Rating' (Higher is better) as the performance metric. Figure 12 shows that all datasets performed better with "Performance SNC = Enabled" BIOS option than others. For all datasets, the "SNC = Enabled" performs 2% to 4% better than "SNC = Disabled".

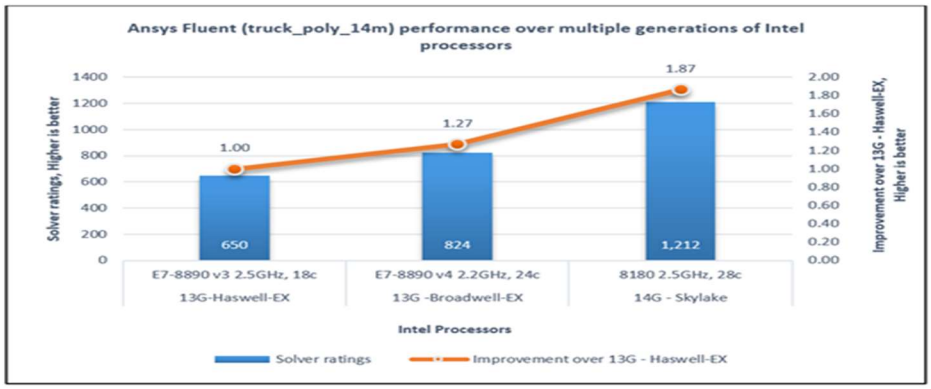

Figure 13: Ansys Fluent (truck_poly_14m) performance over multiple generations of Intel processors

Figure 13, shows the performance comparison of truck poly on PowerEdge R940 with Skylake processors and PowerEdge R930 with Broadwell-EX and Haswell-EX processors. For PowerEdge R940 fluent showed 46% better performance in-comparison to PowerEdge R930 with Broadwell-EX and 87% better performance compared to Haswell-EX processors.

Conclusion:

The PowerEdge R940 is a highly efficient 4 socket next generation platform which provides up to 122TB of storage capacity with 6.3TF of computing power options, making it well-suited for data intensive applications, while not sacrificing performance. The Skylake processors gives PowerEdge R940 a performance boost in comparison to its previous generation of server (PowerEdge R930), we can observe more than 45% performance improvement across all the applications.

Considering our above analysis, we can observe that if we compare system profiles "Performance" gives better performance with respect other system profiles.

In conclusion, PowerEdge R940 with Skylake processors is good platform for all variety of applications and may fulfill the demands of more compute power for HPC applications.