BIOS characterization for HPC with Intel Skylake processor

Ashish Kumar Singh. Dell EMC HPC Innovation Lab. Aug 2017

This blog discusses the impact of the different BIOS tuning options available on Dell EMC 14th generation PowerEdge servers with the Intel Xeon® Processor Scalable Family (architecture codenamed "Skylake") for some HPC benchmarks and applications. A brief description of the Skylake processor, BIOS options and HPC applications is provided below.

Skylake is a new 14nm "tock" processor in the Intel "**tick-tock**" series, which has the same process technology as the previous generation but with a new microarchitecture. Skylake requires a new CPU socket that is available with the Dell EMC 14th Generation PowerEdge servers. Skylake processors are available in two different configurations, with an integrated Omni-Path fabric and without fabric. The Omni-Path fabric supports network bandwidth up to 100Gb/s. The Skylake processor supports up to 28 cores, six DDR4 memory channels with speed up to 2666MT/s, and additional vectorization power with the AVX512 instruction set. Intel also introduces a new cache coherent interconnect named "Ultra Path Interconnect" (UPI), replacing Intel® QPI, that connects multiple CPU sockets.

Skylake offers a new, more powerful **AVX512 vectorization technology** that provides 512-bit vectors. The Skylake CPUs include models that support two 512-bit Fuse-Multiply-Add (FMA) units to deliver 32 Double Precision (DP) FLOPS/cycle and models with a single 512-bit FMA unit that is capable of 16 DP FLOPS/cycle. More details on AVX512 are described in the Intel programming reference. With 32 FLOPS/cycle, Skylake doubles the compute capability of the previous generation, Intel Xeon E5-2600 v4 processors ("Broadwell").

Skylake processors are supported in the Dell EMC PowerEdge 14th Generation servers. The new processor architecture allows different tuning knobs, which are exposed in the server BIOS menu. In addition to **existing options for performance and power management**, the new servers also introduce a clustering mode called Sub NUMA clustering (SNC). On CPU models that support SNC, enabling SNC is akin to splitting the single socket into two NUMA domains, each with half the physical cores and half the memory of the socket. If this sounds familiar, it is similar in utility to the Cluster-on-Die option that was available in E5-2600 v3 and v4 processors as described **here**. SNC is implemented differently from COD, and these changes improve remote socket access in Skylake when compared to the previous generation. At the Operating System level, a dual socket server with SNC enabled will display four NUMA domains. Two of the domains will be closer to each other (on the same socket), and the other two will be a larger distance away, across the UPI to the remote socket. This can be seen using OS tools like numactl –H.

In this study, we have used the Performance and PerformancePerWattDAPC system profiles based on our earlier experiences with other system profiles for HPC workloads. The Performance Profile aims to optimize for pure performance. The DAPC profile aims to balance performance with energy efficiency concerns. Both of these system profiles are meta options that, in turn, set multiple performance and power management focused BIOS options like Turbo mode, Cstates, C1E, Pstate management, Uncore frequency, etc.

We have used two HPC benchmarks and two HPC applications to understand the behavior of SNC and System Profile BIOS options with Dell EMC PowerEdge 14th generation servers. This study was performed with a single server only; cluster level performance deltas will be bounded by these single server results. The server configuration used for this study is described below.

Testbed configuration:

Table 1: Test configuration of new 14G server

Components	**Details**
Server	PowerEdge C6420
Processor	2 x Intel Xeon Gold 6150 – 2.7GHz, 18c, 165W
Memory	192GB (12 x 16GB) DDR4 @2666MT/s
Hard drive	1 x 1TB SATA HDD, 7.2k rpm
Operating System	Red Hat Enterprise Linux-7.3 (kernel - 3.10.0-514.el7.x86_64)
MPI	Intel® MPI 2017 update4
MKL	Intel® MKL 2017.0.3
Compiler	Intel® compiler 17.0.4

Table 2: HPC benchmarks and applications

Application	**Version**	**Benchmark**
HPL	From Intel® MKL	Problem size - 92% of total memory
STREAM	v5.04	Triad
WRF	3.8.1	conus2.5km
ANSYS Fluent	v17.2	truck_poly_14m, Ice_2m combustor_12m

Sub-NUMA cluster

As described above, a system with SNC enabled will expose four NUMA nodes to the OS on a two socket PowerEdge server. Each NUMA node can communicate with three remote NUMA nodes, two in another socket and one within same socket. NUMA domains on different sockets communicate over the UPI interconnect. With the Intel® Xeon Gold 6150 18 cores processor, each NUMA node will have nine cores. Since both sockets are equally populated in terms of memory, each NUMA domain will have one fourth of the total system memory.

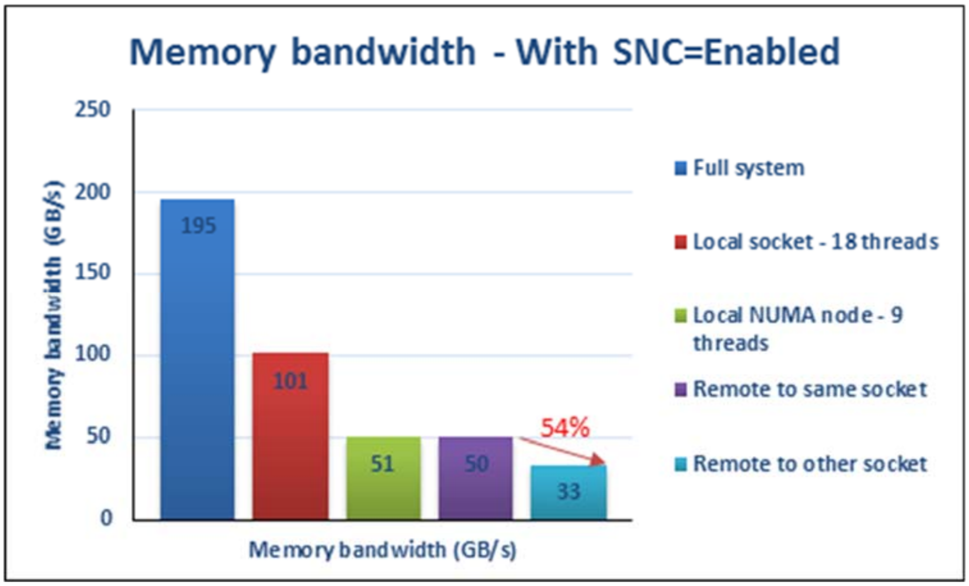

Figure 1: Memory bandwidth with SNC enabled

Figure 1 plots the memory bandwidth with SNC enabled. Except SNC and logical processors, all other options are set to BIOS defaults. Full system memory bandwidth is ~195 GB/s on the two socket server. This test uses all available 36 cores for memory access and calculates aggregate memory bandwidth. The “Local socket – 18 threads” data point measures the memory bandwidth of single socket with 18 threads. As per the graph, local socket memory bandwidth is ~101 GB/s, which is about half of the full system bandwidth. By enabling SNC, a single socket is divided into two NUMA nodes. The memory bandwidth of a single SNC enabled NUMA node is noted by “Local NUMA node – 9 threads”. In this test, the nine local cores access their local memory attached to their NUMA domain. The memory bandwidth here is ~50 GB/s, which is half of the total local socket bandwidth.

The data point “Remote to same socket” measures the memory bandwidth between two NUMA nodes, which are on the same socket with cores on one NUMA domain accessing the memory of the other NUMA domain. As per the graph, the server measures ~ 50GB/s memory bandwidth for this case; the same as the “local NUMA node – 9 threads” case. That is, with SNC enabled, memory access within the socket is similar in terms of bandwidth even across NUMA domains. This is a big difference from the previous generation where there was a penalty when accessing memory on the same socket with COD enabled. See Figure 1 in the previous **blog** where a 47% drop in bandwidth was observed and compare that to the 0% performance drop here. The “Remote to other socket” test involves cores on one NUMA domain accessing the memory of a remote NUMA node on the other socket. This bandwidth is 54% lower due to non-local memory access over UPI interconnect.

These memory bandwidth tests are interesting, but what do they mean? Like in previous generations, SNC is a good option for codes that have high NUMA locality. Reducing the size of the NUMA domain can help some codes run faster due to less snoops and cache coherence checks within the domain. Additionally, the penalty for remote accesses on Skylake is not as bad as it was for Broadwell.

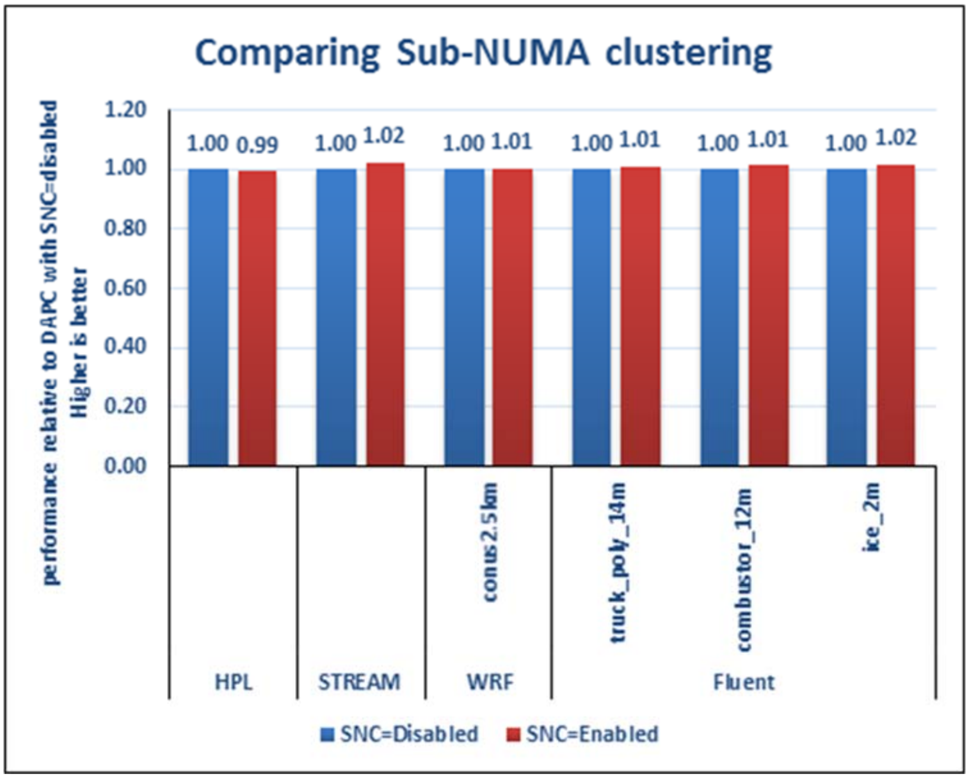

Figure 2: Comparing Sub-NUMA clustering with DAPC

Figure 2 shows the effect of SNC on multiple HPC workloads; note that all of these have good memory locality. All options except SNC and Hyper Threading are set to BIOS default. SNC disabled is considered as the baseline for each workload. As per Figure 2, all tests measure no more than 2% higher performance with SNC enabled. Although this is well within the run-to-run variation for these applications, SNC enabled consistently shows marginally higher performance for STREAM, WRF and Fluent for these datasets. The performance delta will vary for larger and different datasets. For many HPC clusters, this level of tuning for a few percentage points might not be worth it, especially if applications with sub-optimal memory locality will be penalized.

The Dell EMC default setting for this option is “disabled”, i.e. two sockets show up as just two NUMA domains. The HPC recommendation is to leave this at disabled to accommodate multiple types of codes, including those with inefficient memory locality, and to test this on a case-by-case basis for the applications running on your cluster.

System Profiles

Figure 3 plots the impact of different system profiles on the tests in this study. For these studies, all BIOS options are default except system profiles and logical processors. The DAPC profile with SNC disabled is used as the baseline. Most of these workloads show similar performance on both Performance and DAPC system profile. Only HPL performance is higher by a few percent. As per our **earlier studies**, DAPC profile always consumes less power than performance profile, which makes it suitable for HPC workloads without compromising too much on performance.

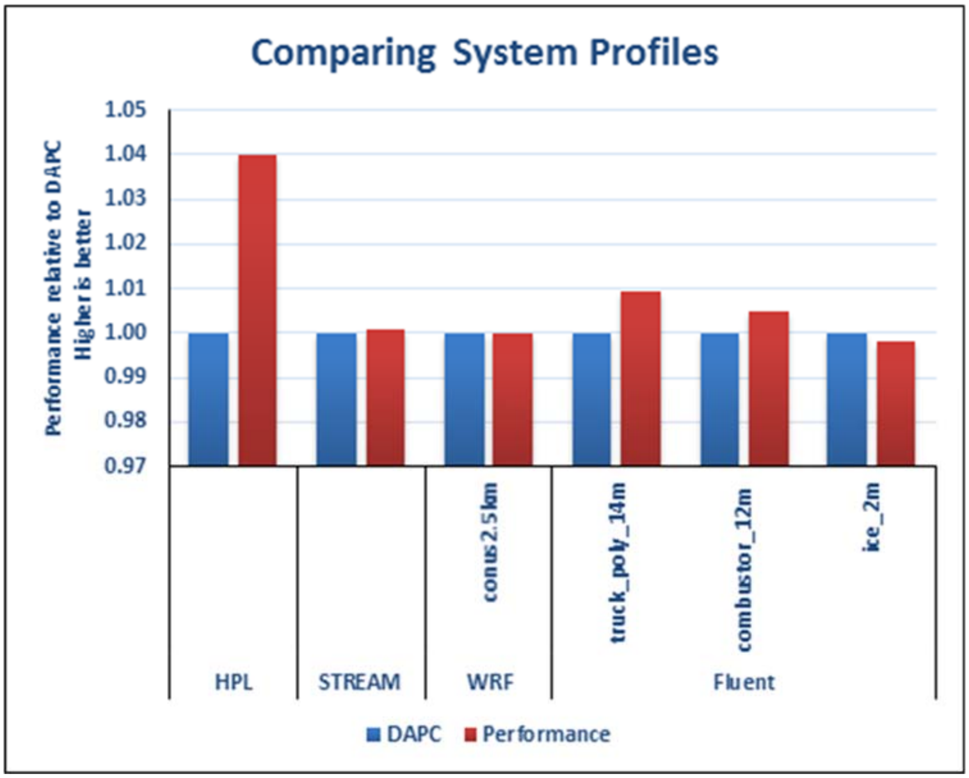

Figure 3: Comparing System Profiles

Power Consumption

Figure 4 shows the power consumption of different system profiles with SNC enabled and disabled. The HPL benchmark is suited to put stress on the system and utilize the maximum compute power of the system. We have measured idle power and peak power consumption with logical processor set to disabled.

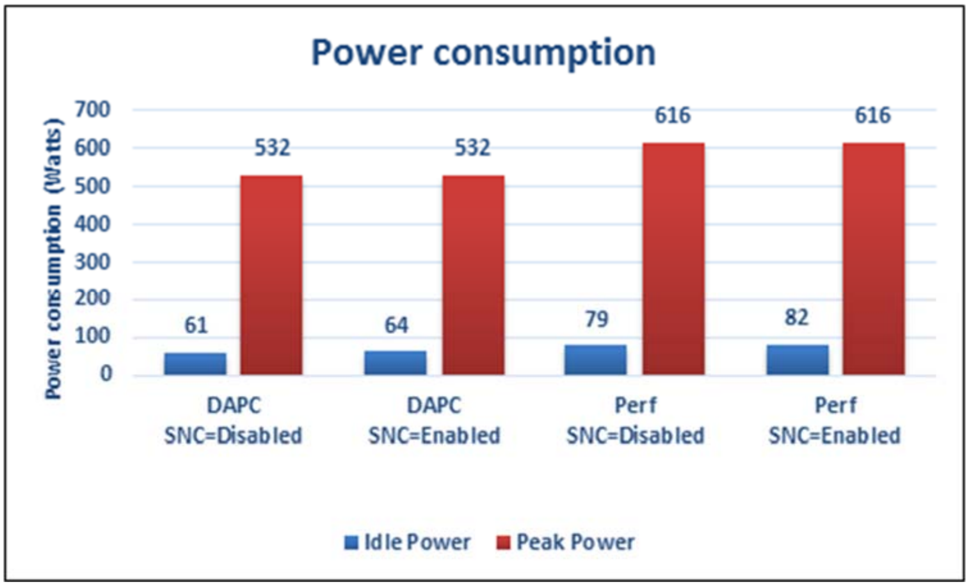

Figure 4: Idle and peak power consumption

As per Figure 4, DAPC Profile with SNC disabled shows the lowest idle power consumption relative to other profiles. Both Performance and DAPC system profiles consume up to ~5% lower power in idle status with SNC disabled. In idle state, Performance Profile consumes ~28% more power than DAPC.

The peak power consumption is similar with SNC enabled and with SNC disabled. Peak power consumption in DAPC Profile is ~16% less than in Performance Profile.

Conclusion

Performance system profile is still the best profile to achieve maximum performance for HPC workloads. However, DAPC consumes less power than performance increase with performance profile, which makes DAPC the best suitable system profile.

Reference:

http://en.community.dell.com/techcenter/extras/m/white_papers/20444326