BIOS tuning for HPC on 13th Generation Haswell servers

Garima Kochhar, September 2014

This blog discusses the performance and energy efficiency implications of BIOS tuning options available on the new Haswell-based servers for HPC workloads. Specifically we looked at memory snoop modes, performance profiles and Intel's Hyper-Threading technology and their impact on HPC applications. This blog is part two of a three part series. Blog one provided some initial results on HPC applications and performance comparisons on these new servers and previous generations. The third blog in this series will compare performance and energy efficiency across different Haswell processor models.

We're familiar with performance profiles including power management, Turbo Boost and C-states. Hyper-Threading or Logical Processor is a known feature as well. The new servers introduce three different memory snoop modes – Early Snoop, Home Snoop and Cluster On Die. Our interest was in quantifying the performance and power consumed across these different BIOS options.

The "System Profile Settings" category in the BIOS combines several performance and power related options into a "meta" option. Turbo Boost, C-states, C1E, CPU Power Management, Memory Frequency, Memory Patrol Scrub, Memory Refresh Rate, Uncore Frequency are some of the sub-options that are pre-set by this "meta" option. There are four pre-configured profiles, Performance Per Watt (DAPC), Performance Per Watt (OS), Performance and Dense Configuration, that can be used. The DAPC and OS profiles balance performance and energy efficiency options aiming for good performance while controlling the power consumption. With DAPC, the Power Management is handled by the Dell iDRAC and system level components. With the OS profile, the operating system controls the power management. In Linux this would be the cpuspeed service and cpufreq governors. The Performance profile optimizes for only performance – most power management options are turned off here. The Dense Configuration profile is aimed at dense memory configurations, memory patrol scrub is more frequent and the memory refresh rate is higher and Turbo Boost is disabled. Additionally if the four pre-set profiles do not meet the requirement, there is a fifth option "Custom" that allows each of the sub-options to be tuned individually. In this study we focus only on the DAPC and Performance profiles. Past studies have shown us that DAPC and OS perform similarly, and Dense Configuration performs lower for HPC workloads.

The Logical Processor feature is based on Intel® Hyper-Threading (HT) technology. HT enabled systems appear to the operating system as having twice as many processor cores as they actually do by ascribing two "logical" cores to each physical core. HT can improve performance by assigning threads to each logical core; logical cores execute their threads by sharing the physical cores' resources.

Snoop Mode is a new category under Memory Setting. Coherence between sockets is maintained by way of "snooping" the other sockets. There are two mechanisms for maintaining coherence between sockets. Snoop broadcast (Snoopy) modes where the sockets are snooped for every memory transaction and directory support where some information is maintained in memory that gives guidance on whether there is a need to snoop.

The Intel® Xeon® Processor E5-2600 v3 Product Family (Haswell) supports three snoop modes in dual socket systems - Early Snoop, Home Snoop and Cluster On Die. Two of these modes are snoop broadcast modes.

In Early Snoop (ES) mode, the distributed cache ring stops can send a snoop probe or a request to another caching agent directly. Since the snoop is initiated by the distributed cache ring stops itself, this mode has lower latency. It is best for workloads that have shared data sets across threads and can benefit from a cache-to-cache transfer, or for workloads that are not NUMA optimized. This is the default mode on the servers.

With Home Snoop (HS) mode, the snoop is always spawned by the home agent (centralized ring stop) for the memory controller. Since every snoop request has to come to the home agent, this mode has higher local latencies than ES. HS mode supports additional features that provide extra resources for larger number of outstanding transactions. As a result, HS mode has slightly better memory bandwidth than ES - in ES mode there are a fixed number of credits for local and remote caching agents. HS mode is targeted at workloads that are bandwidth sensitive.

Cluster On Die (COD) mode is available only on processor models that have 10 cores or more. These processors are sourced from different dies compared to the 8 core and 6 core parts and have two home agents in a single CPU/socket. COD mode logically splits the socket into two NUMA domains that are exposed to the operating system. Each NUMA domain has half of the total number of cores, half the distributed last level cache and one home agent with equal number of cores cache slices in each numa domain. Each numa domain (cores plus home agent) is called a cluster. In the COD mode, the operating system will see two NUMA nodes per socket. COD has the best local latency. Each home agent sees requests from a fewer number of threads potentially offering higher memory bandwidth. COD mode has in memory directory bit support. This mode is best for highly NUMA optimized workloads.

With Haswell processors, the uncore frequency can now be controlled independent of the core frequency and C-states. This option is available under the System Profile options and is set as part of the pre-configured profiles.

There are several other BIOS options available, we first picked the ones that would be most interesting to HPC. Collaborative CPU Performance Control is an option that allows CPU power management to be controlled by the iDRAC along with hints from the operating system, a kind of hybrid between DAPC and OS. This is a feature we plan to look at in the future. Configurable TDP is an option under the Processor Settings section and allows the processor TDP to be set to a value lower than the maximum rated TDP. This is another feature to examine in our future work.

Focusing on HPC applications, we ran two benchmarks and four applications on our server. The server in question is part of Dell's PowerEdge 13th generation (13G) server line-up. These servers support DDR4 memory at up to 2133 MT/s and Intel's latest Xeon® E5-2600 v3 series processors (architecture code-named Haswell). Haswell is a net new micro-architecture when compared to the previous generation Sandy Bridge/Ivy Bridge. Haswell processors use a 22nm process technology, so there's no process-shrink this time around. Note the "v3" in the Intel product name – that is what distinguishes a processor as one based on Haswell micro-architecture. You'll recall that "E5-2600 v2" processors are based on the Ivy Bridge micro-architecture and plain E5-2600 series with no explicit version are Sandy Bridge based processors. Haswell processors require a new server/new motherboard and DDR4 memory. The platform we used is a standard dual-socket rack server with two Haswell-EP based processors. Each socket has four memory channels and can support up to 3 DIMMs per channel (DPC).

Configuration

Table 1 below details the applications we used and Table 2 describes the test configuration on the new 13G server.

Table 1 - Applications and benchmarks

Application	Domain	Version	Benchmark
Stream	Memory bandwidth	v5.9	Triad
HPL	Computation - solve a dense system of linear equations	From Intel MKL	Problem size 90% of total memory
Ansys Fluent	Computational fluid dynamics	v15.0	truck_poly_14m
LS-DYNA	Finite element analysis	v7_0_0_79069	car2car with endtime=0.02
WRF	Weather Research and Forecasting	v3.5.1	Conus 2.5km
MILC	Quantum chromo dynamics	v7.7.3, v7.7.11	Input data file from Intel

Table 2 - Server configuration

Components	Details
Server	PowerEdge R730xd prototype
Processor	2 x Intel® Xeon® E5-2693 v3 – 2.6/2.2 GHz, 14c, 145W 2 x Intel® Xeon® E5-2660 v3 – 2.6/2.2 GHz, 10c, 105W * Frequency noted as "Rated base/AVX base GHz"
Memory	128GB - 8 x 16GB 2133 MHz DDR4 RDIMMs
Hard drive	1 x 300GB SAS 6Gbps 10K rpm
RAID controller	PERC H330 mini
Operating System	Red Hat Enterprise Linux 6.5 x86_64
Kernel	2.6.32-431.el6.x86_64
BIOS settings	As noted per test
MPI	Intel® MPI 4.1.3.049
Math Library	Intel® MKL 11.1.3.174
Compilers	Intel® 2013_sp1.3.174 - v14.0.3.174 Build 20140422

All the results shown here are based on single-server performance. The following metrics were used to compare performance.

- Stream – Triad score as reported by the stream benchmark.
- HPL – GFLOP/second.
- Fluent - Solver rating as reported by Fluent.
- LS DYNA – Elapsed Time as reported by the application.

- WRF – Average time step computed over the last 719 intervals for Conus 2.5km
- MILC – Time as reported by the application.

Power was measured by using a power meter attached to the server and recording the power draw during the tests. The average steady state power is used as the power metric for each benchmark.

Energy efficiency (EE) computed as Performance per Watt (performance/power).

Snoop modes

As described above, with Cluster On Die as the Memory Snoop mode the operating systems sees two NUMA nodes per socket for a total of four NUMA nodes in the system. Each NUMA node has three remote nodes, one on the same socket and two on the other socket. When using a 14core E5-2697 v3 processor, each NUMA node has 7 cores and one fourth of the total memory.

Figure 1 plots the Stream Triad memory bandwidth score in such a configuration. The full system memory bandwidth is ~116 GB/s. When 14 cores on a local socket access local memory, the memory bandwidth is ~ 58GB/s - half of the full system bandwidth. Half of this, ~29 GB/s, is the memory bandwidth of 7 threads on the same NUMA node accessing their local memory.

When 7 threads on one NUMA node access memory belonging to the other NUMA node on the same socket there is a 47% drop in memory bandwidth to ~15GB/s. This bandwidth drops a further 11% to ~14GB/s when the threads access remote memory across the QPI link on the remote socket. This tells us there is significant bandwidth penalty in COD mode when data is not local.

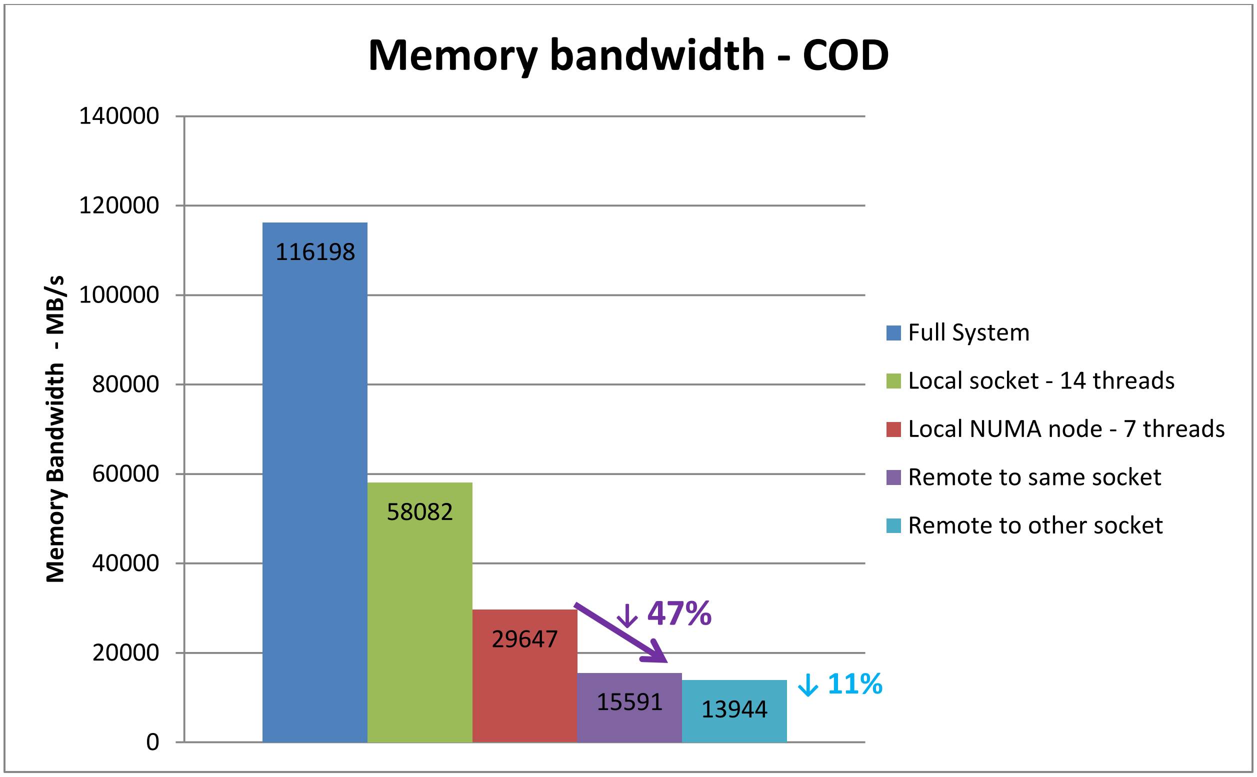

Figure 1 - Memory bandwidth with COD mode

Figure 2 and Figure 3 compare the three different snoop modes on two processor models across the different applications. The system profile was set to DAPC, HT disabled. All other options were at BIOS defaults.

The graphs plot relative performance of the three modes in the height of the bar. This is plotted on the y-axis on the left. The relative power consumed is plotted on the secondary y-axis on right and is noted by a marker. The text value noted in each bar is the energy efficiency, higher is better. The baseline used for comparison is HS mode.

For both the processor models, the performance difference between ES and HS is slight – within a couple of percentage points for most applications for these single-server tests. This difference is expected to be even smaller at the cluster-level. COD performs better than ES/HS for all the applications, up to 4% better in the best case.

In terms of power consumption, COD consumes less power than ES and HS in most cases. This combined with better performance gives COD the best energy efficiency of the three modes, again by a few percentage points. It will be interesting to see how this scales at the cluster level (more future work!).

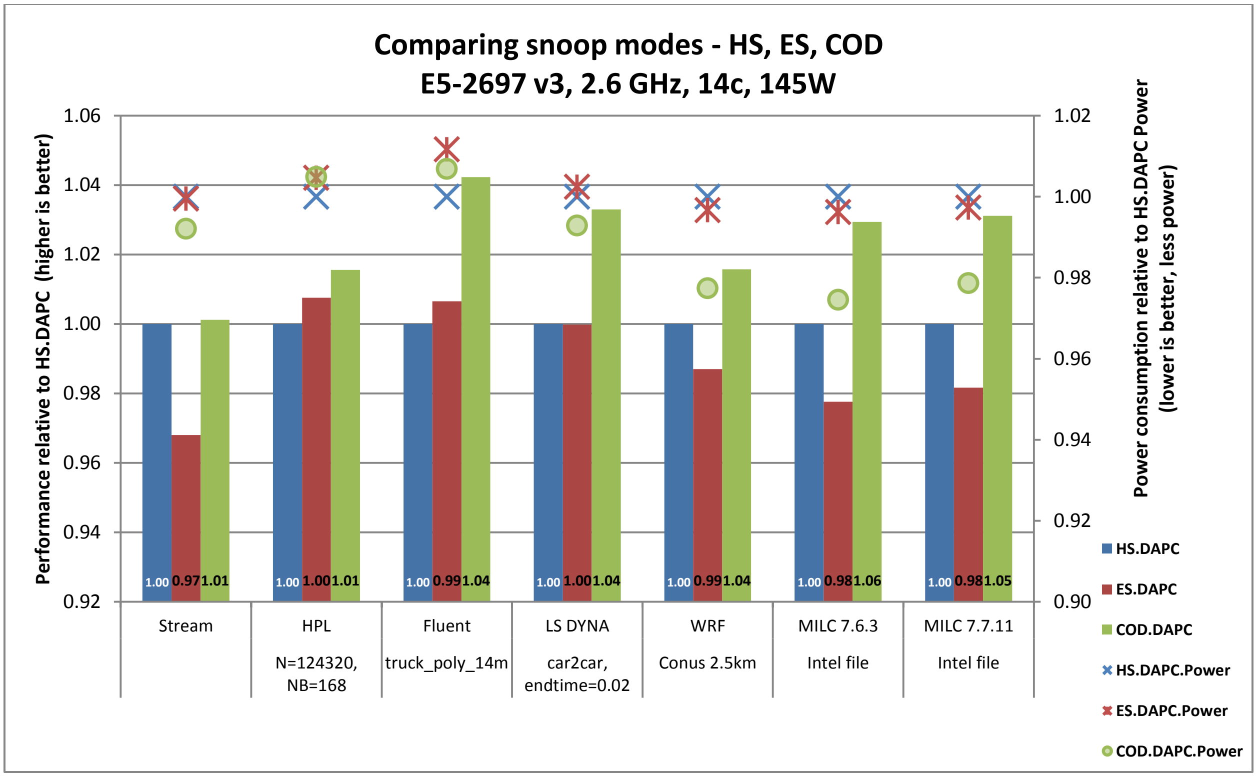

Figure 2 - Snoop modes - E5-2697 v3

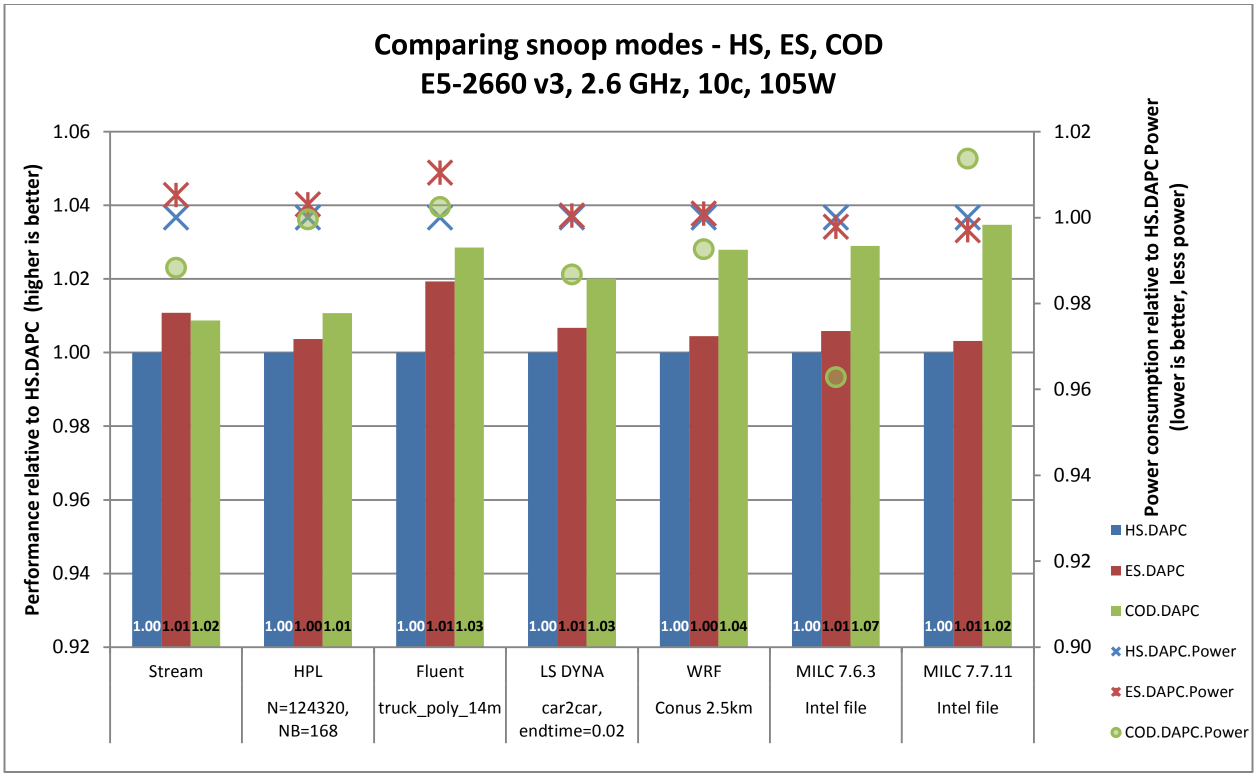

Figure 3 - Snoop modes - E5-2660 v3

System Profile

Figures 4 and 5 compare the System Profile across the two processor models for the HS and COD modes. HT is disabled. All other options were at BIOS defaults.

The graphs plot relative performance in the height of the bar. This is plotted on the y-axis on the left. The relative power consumed is plotted on the secondary y-axis on right and is noted by a marker. The text value noted in each bar is the energy efficiency, higher is better. The baseline used for comparison is HS mode with DAPC profile.

For both the processor models the two profiles DAPC and Performance (Perf) show similar performance, within a couple of percentage points. From the graphs, HS.DAPC is similar to HS.Perf, COD.DAPC is similar to COD.Perf. The bigger differentiator in performance is HS vs. COD, going from DAPC to Perf improves performance by a smaller factor. WRF is the only application that shows better performance with DAPC when compared to Perf.

The Performance profile consumes more power than DAPC by design since many power management features are turned off. This is shown in markers plotted on the secondary-y-axis. As expected, energy efficiency is better with the DAPC profile since the performance improvement with the Perf profile is less than the additional power consumed.

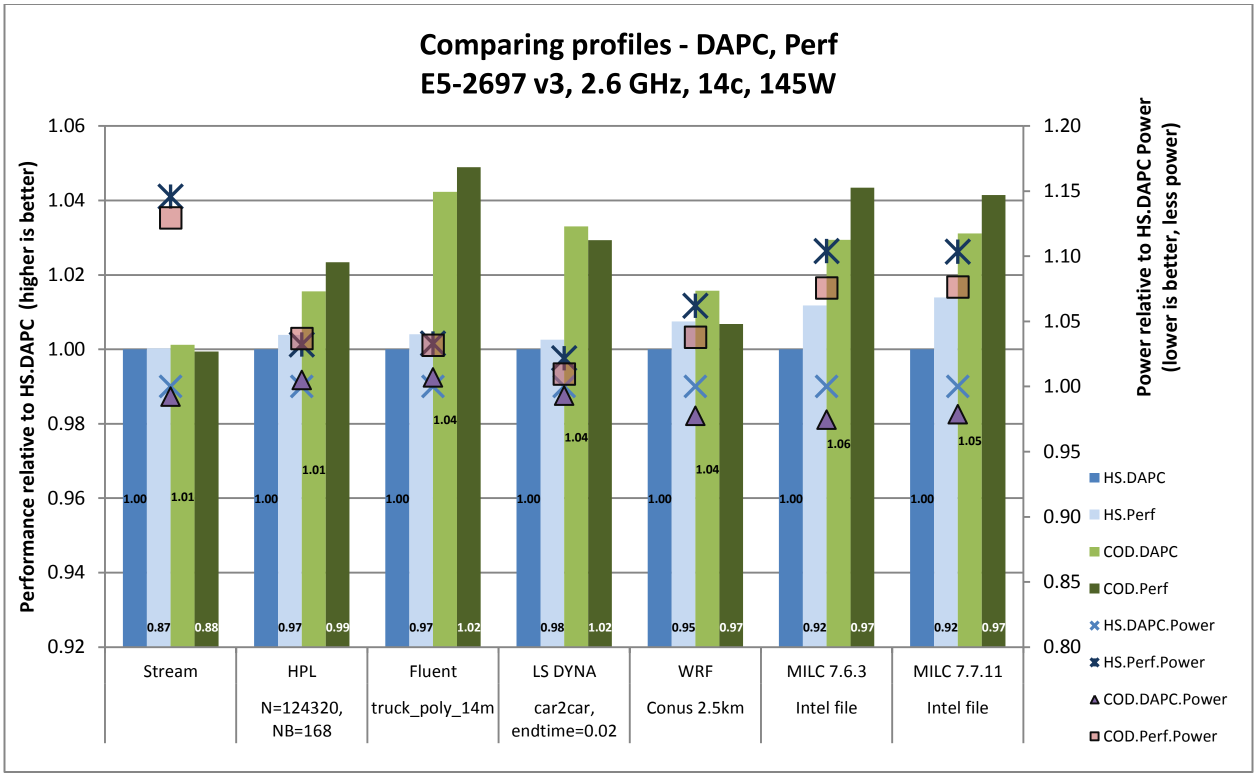

Figure 4 - System profiles - E5-2697 v3

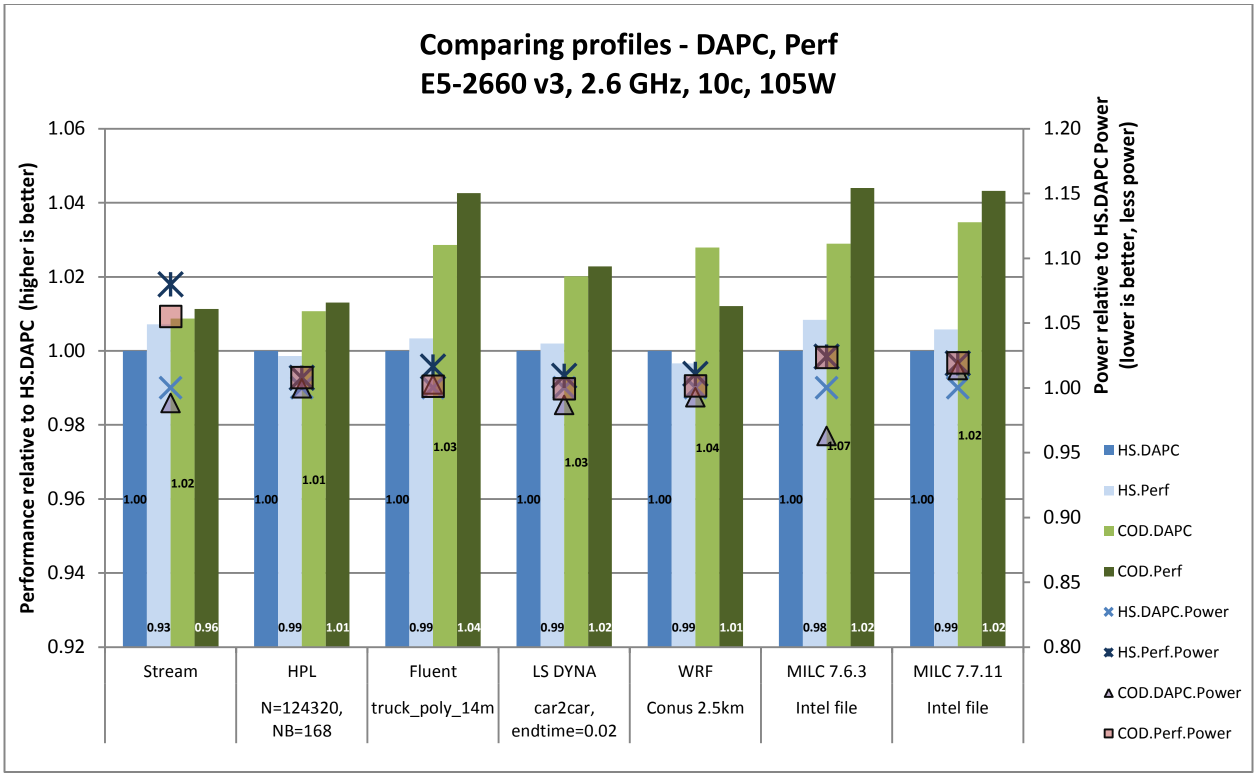

Figure 5 - System Profiles – E5-2660 v3

Hyper-Threading or Logical Processor

Figures 6 and 7 evaluate the impact of Hyper-Threading. These tests were conducted with the HS mode and DAPC System Profile. All other options were at BIOS defaults.

The graphs plot relative performance in the height of the bar. The relative power consumed is plotted on the secondary y-axis on right and is noted by a marker. The text value noted in each bar is the energy efficiency, higher is better. The baseline used for comparison is HS mode, DAPC profile and HT off. Where used in the graph, “HT” implies Hyper-Threading is enabled.

For all applications except HPL, the HT enabled tests used all the available cores during the benchmark. For HPL, only the physical number of cores was used irrespective of HT enabled or disabled. This is because HPL is used as a system benchmark for stress tests and is known to have significantly lower performance when using all HT cores. HT enabled is not a typical use-case for HPL.

HT enabled benefits MILC and Fluent. Fluent is licensed per core, and the ~12% improvement in performance with hyper-threading enabled probably does not justify doubling the license cost for the logical cores.

We measured a 3% improvement in LS-DYNA with HT enabled on the 10c E5-2660 v3. Again, the extra cost for the HT cores probably does not justify this small improvement. The 14c E5-2697 v3 does not show any performance improvement for LS-DYNA with HT. The total memory bandwidth for both these processor models is similar; both support 2133 MT/s memory. With the 14c model, we've added 40% more cores when compared to the 10c model. It's likely that the memory bandwidth per core with HT enabled on the 14c processor is smaller than LS-DYNA's requirement and that is why there is no performance improvement with HT enabled on the 14c processor.

The power consumption with HT enabled was higher for all cases when compared to HT disabled. The EE therefore depended on whether the additional performance with HT enabled was on par with the higher power consumption and is noted as text values in the bars in the graph.

These application trends for HT are similar to what we have measured in the past on Dell's 12th generation Sandy-Bridge based servers.

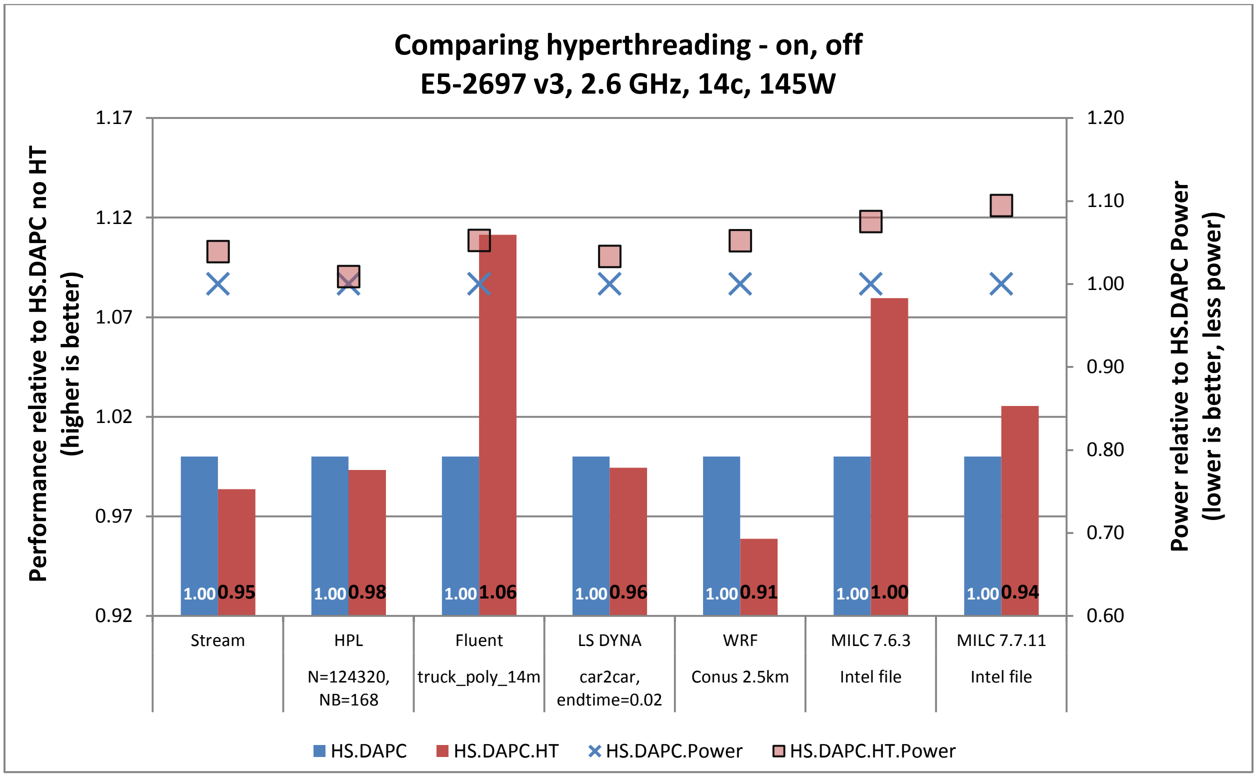

Figure 6 - Hyper-Threading - E5-2697 v3

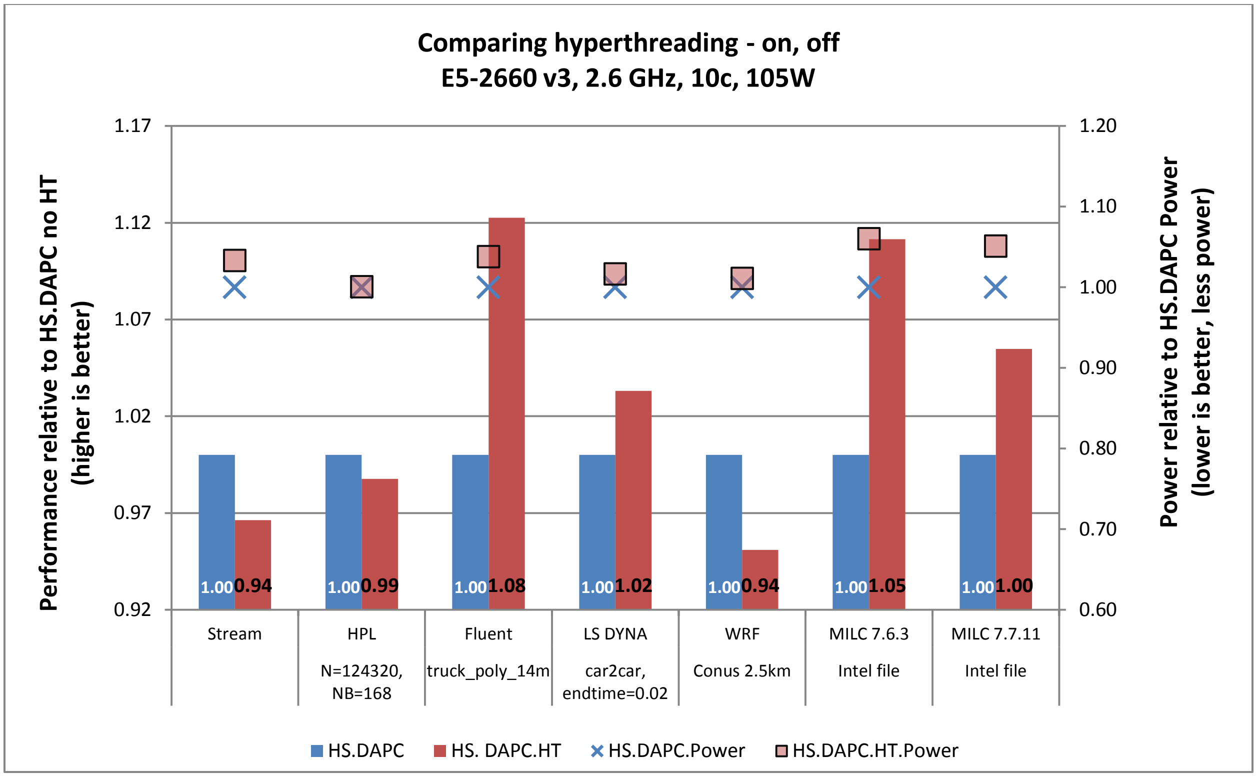

Figure 7- Hyper-Threading - E5-2660 v3

Power consumption – idle and peak

Figures 8 and 9 plot the idle and peak power consumption across the different BIOS settings. Note this data was gathered on an early prototype unit running beta firmware. The power measurements shown here are for ***comparative purposes*** across profiles and not an absolute indicator of the server's power requirements. Where used in the graph, "HT" implies Hyper-Threading is enabled.

The idle power consumption across different snoop modes is similar. The Performance profile adds 60-70 Watts over the DAPC profile for the configuration used in these tests.

The peak power consumption (during HPL initialization) is similar for the 14c E5-2697 v3 across the different BIOS configurations. On the 10c E5-2660v3 the Performance profile consumes ~5% more power than DAPC.

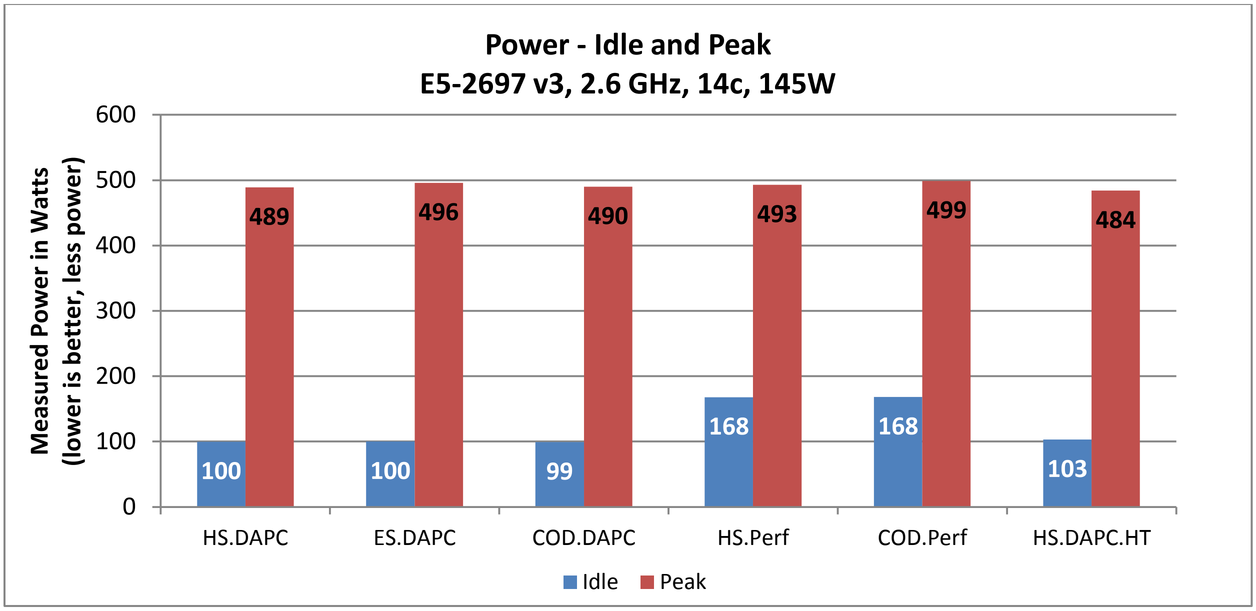

Figure 8 - Idle and Peak power - E5-2697 v3

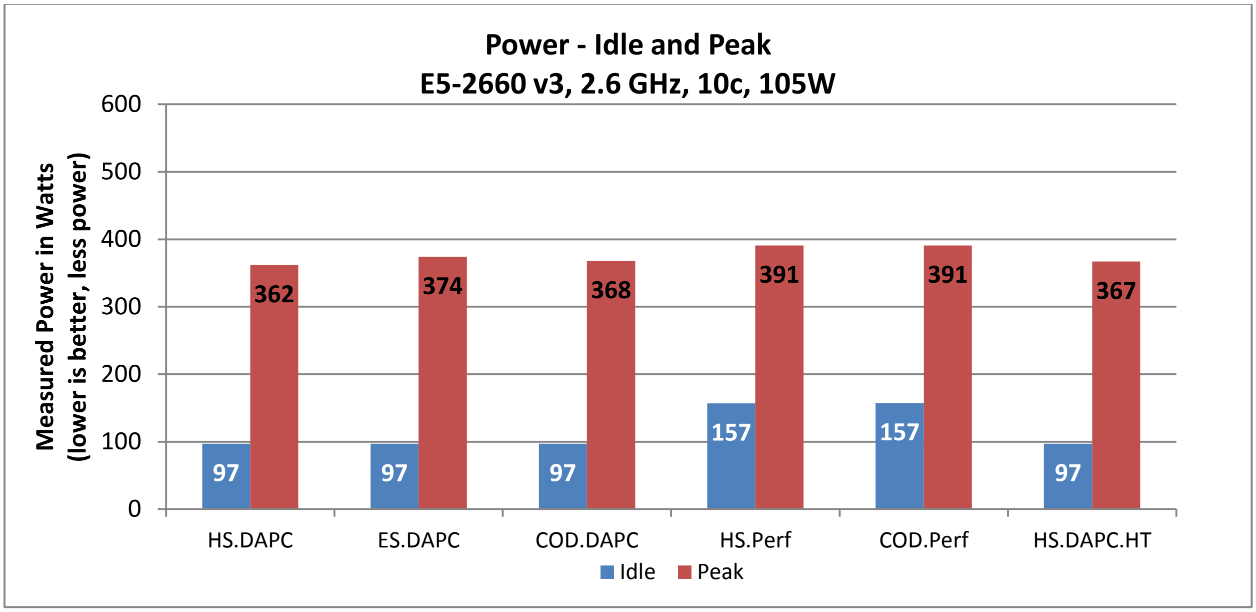

Figure 9 - Idle and Peak power - E5-2660 v3

Conclusion

We expect the ES and HS snoop modes to perform similarly for most HPC applications. Note we have not studied latency sensitive applications here and the benefits of ES mode might be more applicable in that domain. The data sets used in this study show the advantage of COD mode, but the benefit of COD depends greatly on data locality (as shown in Figure 1). We look forward to hearing about COD with real-world use cases.

In terms of System Profile, DAPC appears to be a good choice providing performance similar to Performance profile but with some energy efficiency benefits. Note that the profile DAPC enables C-states and C1E, and will not be a good fit for latency sensitive workloads.

It is recommended that Hyper-Threading be turned off for general-purpose HPC clusters. Depending on the applications used, the benefit of this feature should be tested and enabled as appropriate.

We've evaluated a couple of Haswell processor models in this blog; look out for the third blog in this series that will compare performance and energy efficiency across four different Haswell processor models.