Evaluating the scalability of HEP software and multi-core hardware

Sverre Jarp, Alfio Lazzaro, Julien Leduc, Andrzej Nowak

CERN openlab, Geneva, Switzerland

<Sverre.Jarp@cern.ch>, <Alfio.Lazzaro@cern.ch>, <Julien.Leduc@cern.ch>, <Andrzej.Nowak@cern.ch>

Abstract. As researchers have reached the practical limits of processor performance improvements by frequency scaling, it is clear that the future of computing lies in the effective utilization of parallel and multi-core architectures. Since this significant change in computing is well underway, it is vital for HEP programmers to understand the scalability of their software on modern hardware and the opportunities for potential improvements. This work aims to quantify the benefit of new mainstream architectures to the HEP community through practical benchmarking on recent hardware solutions, including the usage of parallelized HEP applications.

CERN-IT-2011-006
15/01/2011

In this paper we report on a set of benchmark results obtained by CERN openlab [1] when comparing two groups of processors. One is the 6-core "Westmere-EP" processor (X5670 at 2.93GHz), which is compared with Intel's previous generation of the same microarchitecture, the 4-core "Nehalem-EP" (X5570 at 2.93GHz). Both platforms are dual-socket servers. The second group compares a 4-socket, 32-core Intel Xeon "Nehalem-EX" server (X7560 at 2.27GHz) with the previous generation 4-socket "Dunnington" server, based on the 6-core Xeon X7460 processor at 2.66GHz. The Xeon X7560 processor represents a major change in many respects, especially the memory sub-system, so it was important to make multiple comparisons. It should be underlined that both servers represent the "top of the line" in terms of frequency.

Multiple benchmarks were used to get a good understanding of the performance of the new processors. We used both industry-standard benchmarks, such as SPEC2006, and specific High Energy Physics (HEP) benchmarks, representing both simulation of physics detectors and data analysis of physics events. In the following text we show some of the results of these benchmarks and additional details can be found in our other research [2][3]. We must stress the fact that benchmarking of modern processors is a very complex affair. One has to control (at least) the following features: processor frequency, overclocking via Turbo mode, the number of physical cores in use, the use of logical cores via Simultaneous Multi-Threading (SMT), the cache sizes available, the memory configuration installed, as well as the power configuration if throughput per watt is to be measured. We have tried to do a good job of comparing like with like.

1. Benchmark overview

We perform power consumption measurements in two different configurations, with and without SMT enabled in the systems. When conducting the tests without SMT, the systems are considered to have *N* cores in total, which corresponds to the total number of physical cores. According to the standard

energy measurement procedure, the Load stress test consists of running *N*/2 instances of CPUBurn along with *N*/2 instances of LAPACK (using 2 GB of memory each) [2]. In a second phase, now with SMT enabled, the systems were considered to have 2*N* cores in total, meaning that the Load stress test should be conducted by running *N* instances of CPUBurn along with *N* instances of LAPACK (using 1 GB of memory each). The standard energy measurement quoted here is a mix of idle power consumption accounting for 20% of the value, and of the power consumption under Load accounting for 80%.

One of the important performance benchmarks in the IT industry is the SPEC CPU2006 benchmark from the SPEC Corporation [4]. This benchmark can be used to measure both individual CPU performance and the throughput rate of servers. We used the HEPSPEC06 variant, which is a C++ subset of SPEC2006 coming from the HEP community [1].

Since HEP has always been blessed with parallelism inherent in the processing model, it is natural to try to utilize modern multi-core systems by converging towards multi-threaded event processing. The multi-threaded Geant4 [5] prototype is one of the key steps in that direction. Based on Geant4, this suite has been updated to support multi-threading by two Northeastern University researchers [6]. The example used in this case is "ParFullCMSmt", a parallelized version of the "ParFullCMS" program. In our tests we focus on the scalability of the application, which is defined as throughput. In such a scenario the amount of work per core is fixed and grows with the number of cores (weak scaling). Another key metric considered in this case is "efficiency", which is defined as the scaling of the software relative to the serial runtime, compared with ideal scaling determined by the core count. The threads were pinned to the cores running them, and the throughput defining factor was the average time it takes to process one 300 GeV pion event in a predefined geometry. The systems in these tests are always SMT-enabled.

The last benchmark we consider is a data analysis application. In this case we are interested in strong scalability and we optimize for latency. The code is based on the RooFit package [7], which is part of the ROOT framework [8], and is parallelized using MPI [9]. It performs an extended unbinned maximum likelihood fit [10] used for a measurement documented in Ref. [11]. The procedure for finding the maximum requires several evaluations of the likelihood function *L*. Given the fact that *L* is the sum of different terms calculated for each event, identified by several variables, for a given data sample, it is possible to distribute the calculation of each event over different processes and then collect all results for the final calculation of *L* for each process. The workload is not entirely balanced amongst the parallel processes since one of the processes does the calculation of the extended term. We look at the efficiency, defined as the scaling of the software relative to the serial runtime (scalability) confronted with ideal scaling determined by the process count. The fraction of time spent for executing code that we can parallelize ranges between 90% and 98% of the sequential execution, depending on the system. Since we are considering strong scaling, there is no benefit from hardware threads. Even though the systems were SMT-enabled, the feature was not used during the tests.

2. Westmere-EP X5670 vs. Nehalem-EP X5570

Results of the power consumption measurements are shown in Table 1. Turning SMT on introduces a minor penalty in power, established to be around 1.5% of our measurement (6-7W). Since CERN usually requires 2GB of memory per core (or process), we extrapolated the measurements from 12GB to 24GB.

Active Power		*Idle*	*Load*	*Standard measurement*
12 GB	SMT-off	215 W	449 W	402 W
	SMT-on	227 W	455 W	409 W
24 GB	*SMT-off*			*426 W*[1]

[1] Extrapolated figures taking 4 W of Standard power consumption for each 2GB memory module

	SMT-on		*433 W*[1]

Table 1: “Westmere-EP” X5670 total power consumption using two power supply units

HEPSPEC06 benchmark results are shown in Figure 1. We compare the performance of the two systems, the “Westmere-EP” X5670 and the “Nehalem-EP” X5570, respectively, running at the same frequency. We note that from 1 to 8 CPUs, the two systems obtain the same performance (within a few percent). Indeed, in this range, both the “Westmere-EP” X5670 and the “Nehalem-EP” X5570 systems have enough physical cores to sustain the benchmark load. Further out along the curve, the “Westmere-EP” X5670 platform extends its performance, thanks to its four additional cores, but an inflexion occurs in the performance increase: from 8 to 12 CPUs, the HEPSPEC06 number increases by 32% while increasing the number of CPUs by 50%, showing a relative overall scalability of 88%. The gain produced by SMT can be computed by comparing the HEPSPEC06 results for 12 and 24 CPUs for the “Westmere-EP” X5670, and for 8 and 16 CPUs for the “Nehalem-EP” X5570: in the former, the SMT gain is 23.7%, while in the latter is 24.4%. This again shows that the two processors have very similar behaviour.

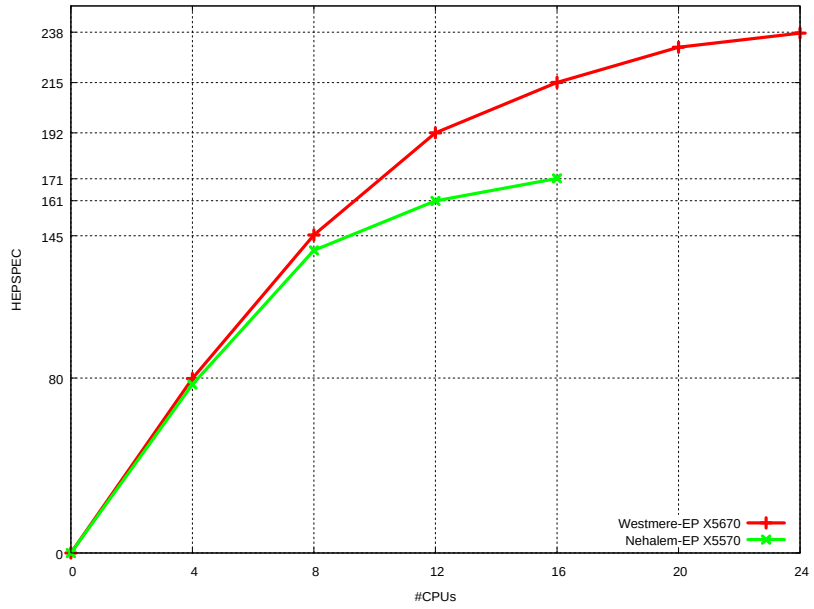

Figure 1: HEPSPEC06 performance comparison SMT-off Turbo-on (higher is better)

Concerning the data analysis benchmark, the application takes about 37 minutes when running in sequential mode. First we run the test with Turbo mode enabled. The total execution time (wall-clock time) and efficiency for 1 to 12 processes are shown in Table 2. We can see that the efficiency is poor when we increase the process count. Looking at the profile, we note that the sequential part for the extended term calculation by a single node takes a significant part of the time, which increases with the number of processes (see Table 2). Therefore it limits the scalability. If we remove the time for the extended term calculation, we obtain the results shown on Figure 2. In this case the efficiency is between 92% and 100%, with a negative slope versus the increase of number of processes. This effect is expected as well, since we know that the application is still not fully parallelized in the remaining part. So the results are in accord with expectations. Repeating the test without Turbo mode enabled, we observe that there is a 10% degradation of performance, with an efficiency surpassing 95%.

Other tests and additional details can be found in Ref.[2].

# Processes	*Wall-time [seconds]*	*Efficiency [%]*	*Sequential part [%]*
1	2215	100	–
2	1239	89	10
4	755	73	26
8	527	52	45
12	448	41	55

Table 2: Wall-clock time and efficiency for the data analysis application requiring different number of processes running on "Westmere-EP" X5670 based system. We show also the percentage of time with respect to the wall-time spent for the extended term calculation.

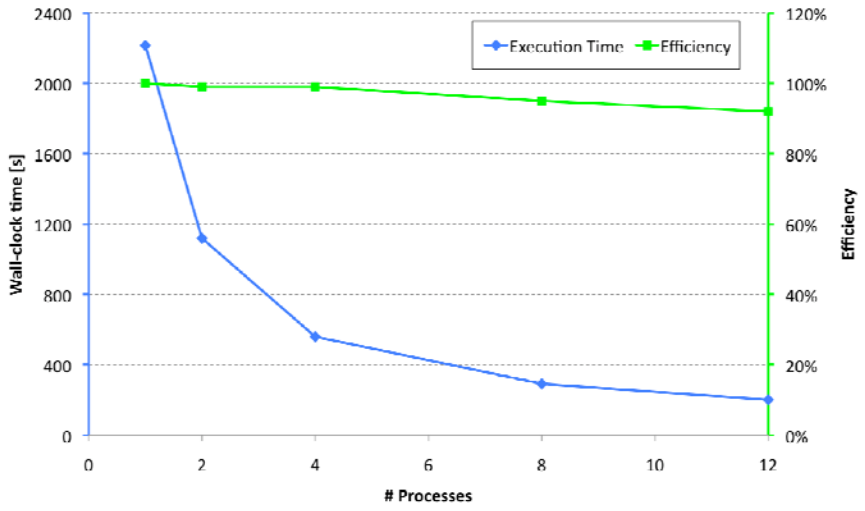

Figure 2: Time spent for the calculation of the data analysis application (blue line) and efficiency (green line) running on "Westmere-EP" X5670 based system. For this plot we subtract the time spent for the extended term calculation.

3. Nehalem-EX X7560 vs. Dunnington X7460

Table 3 contains the results of the test for the power consumption measurements. We reach some sizeable figures, even when the server is idle. The Standard power consumption of all the memory boards and their associated DIMMs is 448W - almost 40% of the total power consumed by the system.

Active Power		*Idle*	*Load*	*Standard measurement*
128 GB	SMT-off	715 W	1209 W	1110 W
	SMT-on	715 W	1243 W	1137 W

Table 3: "Nehalem-EX" X7560 total power consumption using three power supply units

Figure 3 shows HEPSPEC06 results. To compare the two systems, the "Dunnington" X7460 results were frequency scaled, from the initial 2.66GHz clock rate down to 2.27GHz, to match the "Nehalem-EX" X7560 frequency. Both systems are four socket systems, aimed at the "expandable" server market, however scalability testing shows that the behaviour of the two systems running HEPSPEC06 benchmark is inherently different. Where the "Dunnington" X7460 seems to reach quickly a horizontal asymptote, the "Nehalem-EX" X7560 system is able to sustain increasing load. If a direct comparison to the "Dunnington" X7460 server is considered, the tested system allows for 3x more

throughput. Frequency scaled results show that the “Nehalem-EX” X7560 based system yields around 3.5x more throughput.

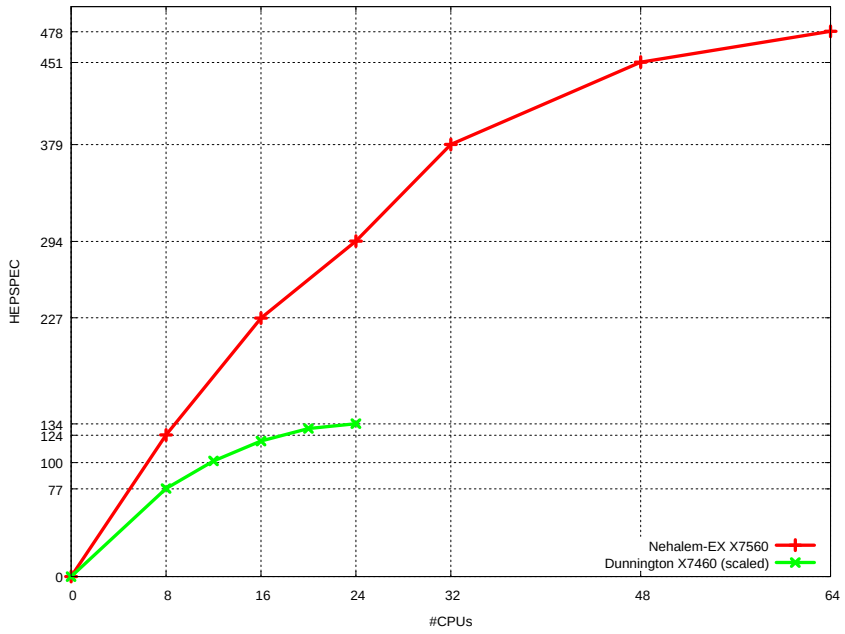

Figure 3: HEPSPEC06 performance comparison: SMT-off Turbo-on, “Dunnington” X7460 frequency scaled. The value for 1 core in the case of “Nehalem-EX” X7560 is 15.5.

Given the high number of cores available, these systems are particularly suitable for throughput computing. The multi-threaded Geant 4 prototype application scales quite well up to 32 physical cores.

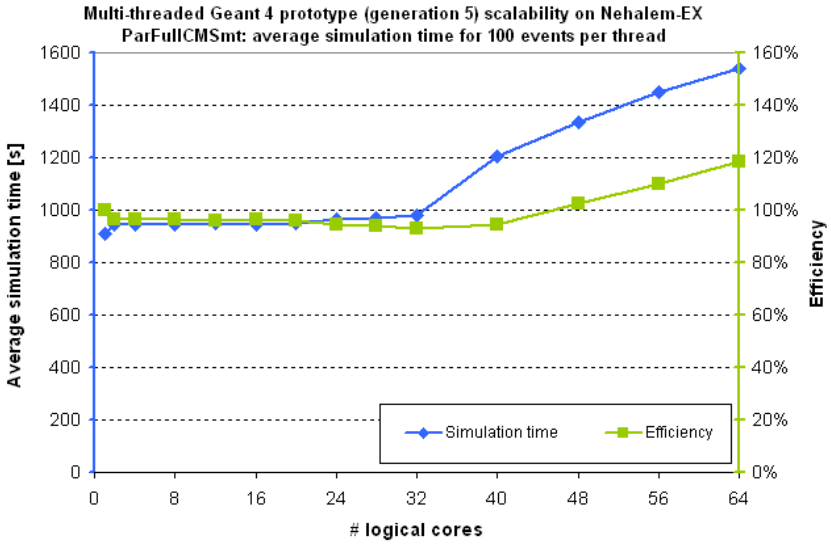

Figure 4: ParFullCMSmt scalability on a “Nehalem-EX” X7560 based system (with SMT)

The efficiency under full physical core load was 93%, which corresponds to a scaling factor of 29.7x. Relative scaling up to 8 or 12 cores is comparable to the Nehalem-EP X5570 and “Westmere-

EP" X5670 systems. A slight drop in efficiency is observed beyond 24 cores, the origin of which remains to be discovered. In essence, the efficiency curve as seen on Figure 4 is nearly flat, which means that one can expect excellent, predictable scalability with this kind of applications. Figure 4 also shows the data for points between 1 and 64 threads. The efficiency curve recovers beyond 32 cores and ultimately surpasses 100%, since for thread counts higher than 32, expected scalability is fixed at 32x. Thus a final value of 118% indicates that the system loaded with 64 threads of the benchmark yields 18% more throughput than a perfectly scaled serial version on 32 physical cores.

When compared to the "Dunnington" X7460 platform, the "Nehalem-EX" X7560 performs 17% better with one thread on one core, and 12% better with 24 threads on 24 cores. The tested "Nehalem-EX" X7560 platform also delivers a 33% increase in core count, and adds Hyper Threading functionality in comparison to the previous solution. A "Nehalem-EX" machine with 32 cores fully loaded provides 47% more throughput than the equivalent "Dunnington" solution, and 87% more when loaded with 64 threads on 32 cores. In essence, a "Nehalem-EX" core can be up to 17% faster than an equivalent Core 2 based one, while the overall system provides up to 87% more throughput when using SMT. However, if we consider that both the "Nehalem-EX" X7560 and its Core 2 counterpart are "top of the line" processors with the highest frequency bins available, we can also make a realistic, direct comparison of the two. In this case, the "Nehalem-EX" performs slightly (1-5%) worse in terms of absolute performance, but the increased core count allows for a 25% advantage with 32 cores loaded, and the addition of SMT increases this advantage to 60% with 64 threads.

Also for these systems additional details can be found in Ref. [3].

4. Conclusion

The Xeon 5600 "Westmere" platforms provide a consistent and "across the board" improvement in performance with respect to the previous Xeon 5500 generation ("Nehalem"). Depending on the workload and measurement plane, the specific benefits vary from case to case. The additional two cores yield an extra 32% to nearly 50% in performance. The overall advantage over Nehalem based servers is established to be at 39% for the HEPSPEC06 benchmark, and between 46% and 61% for in-house applications. Also, a 10-23% performance per Watt improvement was observed. Although the gain is smaller than the 35% improvement measured during the transition from Harpertown to Nehalem, this development is quite important for CERN's computer centre [12]. Platform power consumption as an important factor in the tendering process for its power constrained computing centre building.

The Xeon 7500 "Nehalem-EX" platform provides a significant jump in performance and efficiency compared to the previous Xeon 7400 "Dunnington" generation. Part of the improvements can be credited to the architectural jump and related developments – "Nehalem" represents a significant improvement over "Core 2". The new processors represent a 33% core increase over the previous generation, and the cache increase was of 8MB (24MB cache in total). The performance figures thoroughly represent these additions. If we consider clock for clock performance, it is roughly 15% to 20% better than the tested "Dunnington" system. If we consider the overall frequency-scaled throughput of the whole system, a significant increase can be seen in comparison to "Dunnington" based servers. The measured throughput slightly exceeded 3.5x that of the "Dunnington" for the HEPSPEC06 benchmark (SMT included), and has increased between 47% and 87% for in-house applications (SMT off and on respectively). The noteworthy result for HEPSPEC06 could be credited in a large part to weak performance on the "Dunnington", possibly stemming from an FSB and cache coherency bottleneck. The power-performance of the "Dunnington" system is not directly comparable and was not evaluated in this context.

References

[1] http://cern.ch/openlab

[2] Jarp S, Lazzaro A, Leduc J and Nowak A 2010 *Evaluation of the Intel Westmere-EP server processor* CERN openlab report

[3] Jarp S, Lazzaro A, Leduc J and Nowak A 2010 *Evaluation of the Intel Nehalem-EX server processor* CERN openlab report
[4] http://www.spec.org/
[5] http://cern.ch/geant4
[6] Apostolakis J, Cooperman G and Dong X 2010 *Multithreaded Geant4: Semi-Automatic Transformation into Scalable Thread-Parallel Software* Europar 2010
[7] Verkerke W, Kirkby D 2006 *The RooFit Toolkit for data mod- eling* proceedings of PHYSTAT05, Imperial College Press
[8] Brun R, Rademakers F 1997 *ROOT An object oriented data analysis framework* Nuclear Instruments and Methods in Physics Research Section A **389** 81
[9] Lazzaro A and Moneta L 2010 *MINUIT package parallelization and applications using the RooFit package* J. Phys.: Conf. Ser. **219** 042044
[10] Cowan G 1998 Statistical Data Analysis (Oxford: Clarendon Press)
[11] Aubert B *et. al.* (*BABAR* Collaboration) 2007 *Observation of CP Violation in* B^0 *to* $\eta' K^0$ *Decays*, Physics Review Letters **98** 031801
[12] Busch A and Leduc J 2009 *Evaluation of energy consumption and performance of Intel's Nehalem architecture* CERN openlab report