Beyond core count: a look at new mainstream computing platforms for HEP workloads

P Szostek, A Nowak, G Bitzes, L Valsan, S Jarp and A Dotti

CERN openlab, Geneva, Switzerland

E-mail: {pawel.szostek, andrzej.nowak, georgios.bitzes, liviu.valsan, sverre.jarp, andrea.dotti}@cern.ch

Abstract. As Moore's Law continues to deliver more and more transistors, the mainstream processor industry is preparing to expand its investments in areas other than simple core count. These new interests include deep integration of on-chip components, advanced vector units, memory, cache and interconnect technologies. We examine these moving trends with parallelized and vectorized High Energy Physics workloads in mind. In particular, we report on practical experience resulting from experiments with scalable HEP benchmarks on the Intel "Ivy Bridge-EP" and "Haswell" processor families. In addition, we examine the benefits of the new "Haswell" microarchitecture and its impact on multiple facets of HEP software. Finally, we report on the power efficiency of new systems.

1. Introduction

1.1. Description of the processors

In this paper we describe four different Intel Xeon processors, belonging to three different generations. In one comparison, we use two dual socket units: E5-2690 "Sandy Bridge-EP" and its successor, the E5-2695 "Ivy Bridge-EP", belonging to server class solutions. In another comparison, we use single socket workstation products: E3-1265L "Ivy Bridge-EN" and E3-1280 "Haswell-EN". The former pair represents the "tick" step in the Intel model, a shrink from 32nm to 22nm silicon process, whereas the latter is the "tock" step, incorporating microarchitectural changes. In this paper we assess the impact of architectural features of new generations of CPUs on the performance obtained on HEP workloads.

The major architectural features in the Haswell family are an expansion of in-core execution ports and support for Fused Multiply-Add. The "Ivy Bridge" server processor differs from its predecessor mainly with a core count – we compare a 12-core part to an 8-core "Sandy Bridge" predecessor.

1.2. Hardware configuration

1.2.1. Dual socket server. The Intel "Sandy Bridge-EP" processor, model E5-2690, is running at 2.9GHz and is equipped with 512KB L1 cache, 2048 L2 cache and 30MB L3 cache. Its Thermal Design Power (TDP) is estimated at 135W. The motherboard installed in the system is a dual socket Intel S2600JP motherboard supporting up to 16 DDR3 memory DIMMs. This system is equipped with 64GB of memory (8x8 GB DIMMs, 1333MHz, low voltage, ECC, registered) and 3 SSDs, each one with a capacity of 240 GB. The motherboard used is designed to fit into a four system chassis with 2 PDU slots. Both server systems were mounted in an Intel H23212JFJR Jefferson Pass 2U chassis with two redundant 1600W Intel 80 PLUS platinum power supplies.

The Intel "Ivy Bridge-EP" processor, model E5-2695 v2, is running at 2.8GHz and has the same cache configuration as the tested Sandy Bridge part. The TDP is estimated at 115W. For this CPU we used exactly the same hardware configuration as for Sandy Bridge to minimize the impact of other system parts on the overall performance and power measurement. The common components include DIMM configuration, PSU, mother board, hard drives and BIOS version.

To simplify the power consumption measurements and to lessen the thermal impact of other systems, the three other slots in the chassis were not installed. We are aware that it might introduce a slight bias in power usage in idle state caused by PSU inefficiencies. Conducting fully unskewed measurements would require further investigations on the PSU efficiency, as it might grow monotonically with load. This kind of analysis is however beyond the scope of our work.

1.2.2. Single socket workstations. The Intel "Ivy Bridge-EN" processor, model E3-1265L v2, is running at 2.5GHz and has 128kB L1 cache, 1024kB L2 cache and 8MB L3 cache and is mounted on a single socket Asus P8Z77-M motherboard. The Intel "Haswell" workstation processor, model E3-1280 v3, is running at 3.6GHz, with the same cache setup as the Ivy Bridge part. The motherboard in the system is a single socket Intel S1200V3RP with 4 DDR3 memory slots.

The investigated workstation platforms have 16GB of memory (2x8 GB DIMMs, 1333MHz, ECC, unbuffered) and one Intel S3500 SSD. They were installed in a pedestal mount desktop chassis with standard 500W desktop PSU.

1.3. Software configuration of the machines

All systems were running 64-bit Scientific Linux CERN (SLC) 6.3, based on Red Hat Enterprise Linux 6 (Server). The default kernel for the distribution, that is 2.6.32-358.18.1.el6, was used for all the measurements.

Both server machines used the same BIOS versions and have enabled the following performance-related features: MLC Streamer, MLC Spatial Prefetcher, DCU Data Prefetcher and DCU Instruction prefetcher.

2. Standard energy measurement

2.1. Methodology

For our energy measurements we adopted a well-established procedure from the IT department at CERN. This procedure was already thoroughly described in previous openlab papers [7][8]. A description is attached as an appendix to this paper. During all measurements Intel Speed Step Technology was turned on. For the measurements on server machines we used two PSUs at a time (as opposed to one) in order to prevent the CPU from downscaling the clock. The final score is a sum of measurements from both PSUs.

2.2. Results

The dual socket machines were tested with both two power supply units (PSU) enabled. To follow the CERN standard energy consumption procedure, all machines were loaded with the same number of LAPACK and CPU Burn instances, each being equal to the half of all available logical cores. For instance, on the "Sandy Bridge-EP" with SMT enabled there were 48 cores available, out of which 24 were populated with LAPACK instances and the remaining 24 with CPU Burn instances. The tables below compare power consumption of all tested systems.

Table 1. Power consumption in idle mode.

Active power	Turbo enabled		Turbo disabled	
	SMT on	SMT off	SMT on	SMT off
E3-1265L V2 **“Ivy Bridge”**	31	31	31	31
E3-1285L V3 **“Haswell”**	24	24	24	24
E5-2690 **“Sandy Bridge-EP”**	126	126	126	126
E3-2695 V2 **“Ivy Bridge-EP”**	99	99	97	97

Table 2. Power consumption when loaded.

Active power	Turbo enabled		Turbo disabled	
	SMT on	SMT off	SMT on	SMT off
E3-1265L V2 **“Ivy Bridge”**	62	60	60	60
E3-1285L V3 **“Haswell”**	63	63	63	60
E5-2690 **“Sandy Bridge-EP”**	433	421	421	375
E3-2695 V2 **“Ivy Bridge-EP”**	403	406	338	321

Of particular notice are PSUs characteristics of tested servers. When the tested server system had “Ivy Bridge-EP” processors installed, the load was distributed unevenly between the power supplies, that is one of them was running idle with output power of 4W, whereas the other one produced between 90W and 430W of output power. When the chassis was populated with “Sandy Bridge-EP” processors, the load was equally distributed between two PSUs. BIOS versions and settings were equal in both cases.

3. Benchmarks

3.1. HEP-SPEC06

SPEC CPU2006 is one of the most widely used benchmarks in the IT industry. The workloads are designed to stress the CPU, usually requiring neither high memory bandwidth nor capacity, and there is no significant I/O either that could influence the results. The suite consists of real-life applications available on a commercial basis. Several years ago a working group at HEPiX has shown a significant correlation between a C++ subset of the tests and HEP applications. In consequence the HEP community decided to adopt this subset, denominated “HEP-SPEC06”, as a reference benchmark.

In our tests the HEP-SPEC06 suite was compiled with GCC 4.4.7 in 64-bit mode with –O2 and −pthread used as switches. The tests were carried out with Intel Hyper-Threading Technology (thereafter also called SMT) enabled and Turbo Boost disabled. The processes were assigned to the cores by the operating system kernel. Since HEP-SPEC06 consists of several single-threaded workloads, scalability testing required running several instances in a parallel fashion. The overall result is an addition of results from separate runs.

These benchmarks serve us as a measure of scalability of different architectures. To make the comparison between different CPUs fair, all the results were scaled to a reference frequency of 2.7GHz.

3.1.1. HEP-SPEC06 results. Both tested server systems are dual socket machines. Figure 1 shows the scalability of HESPEC06 results with an increasing number of processes. In the leftmost part of the plot, when the two systems have enough physical cores to accommodate the workload, they scale linearly, with 12% advantage of “Ivy Bridge-EP” HEP-SPEC06 performance per core. When we compare results for the maximum number of threads, “Sandy Bridge-EP” has a light advantage of 2% per core. Nevertheless, when comparing maximal performance, “Ivy Bridge-EP” has an advantage of 47%.

Figure 2 shows that the scalability of the “Ivy Bridge” and “Haswell” workstation processors is comparable, with a small advantage of the “Ivy Bridge” part under the full load. It must be noted here that the both workstation CPUs were mounted on motherboards from different manufacturers. The results for the workstations cannot be extrapolated to the server configuration. With this benchmark we were not able to exploit the advantage of the new “Haswell” microarchitecture as the workloads’ code in not parallelism-oriented.

3.1.2. Hyper-Threading gain. The gain obtained with SMT is calculated as a ratio of the results produced with maximum number of threads to the result when all physical cores were used, this is 32 and 16 cores for the "Sandy Bridge-EP" and 48 and 24 cores in case of "Ivy Bridge-EP". In the case of "Sandy Bridge" we observed a gain of 24%, whereas "Ivy Bridge" offered only 19% of improvement. When comparing workstation processors we take performance on 8 and 4 cores into account. On "Haswell", Hyper-Threading benefit reaches 23% and 20% for "Ivy Bridge". Results show that SMT is a well-established technology delivering around 20% of extra performance. The results show that "Ivy Bridge-EP" offers slightly worse SMT gain than "Sandy Bridge-EP". The matter will be investigated further.

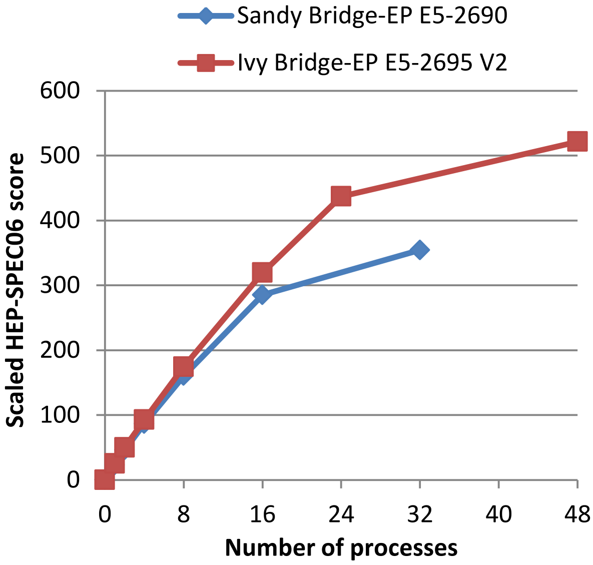

Figure 1. HEP-SPEC06 frequency scaled performance comparison for server machines (higher is better).

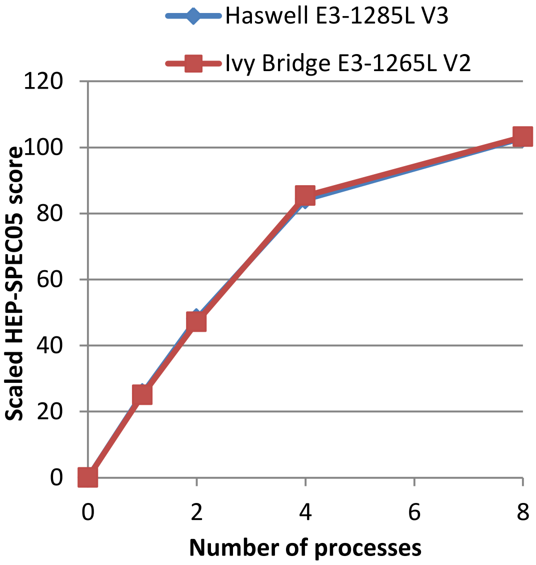

Figure 2. HEP-SPEC06 frequency scaled performance comparison for workstation machines (higher is better).

3.2. Power efficiency

The power efficiency is a key factor for computing centers. In this paper we implement CERN IT standard power consumption measurement, which is described in the appendix. Platform power efficiency is calculated as an unscaled HEP-SPEC06 score per watt of power consumption and incorporates all elements of the system, including hard drives, power supplies, motherboards and others. Therefore it cannot be translated directly to efficiency of the CPU itself. Our goal in this case was to assess the power efficiency of the whole platform, not only the CPU. Therefore we kept the amount of RAM memory and other configuration parameters the same in both machines.

In Table 3 one can see calculated platform efficiency. It shows that the "Haswell" platform offers 19% more performance per watt compared to "Ivy Bridge". Comparing the results for server platforms, we observe that "Ivy Bridge-EP" shows a significant improvement over its predecessor – nearly 54%, which was achieved in 18 months It is worthwhile to notice that in openlab's previous paper on platform comparison [7], which focused on two consecutive platform families, namely "Sandy Bridge-EP" and "Westmere-EP", we observed improvement of 70%.

Table 3. Standard energy measurement and HEP-SPEC06 per watt results for the tested machines.

	"Ivy Bridge" E3-1265L V2	**"Haswell" E3-1285L V3**	**"Sandy Bridge-EP" E5-2690**	**"Ivy Bridge-EP" E5-2695 V2**
Standard measurement [W]	54	56	362	290
HEP-SPEC06 score	94	115	381	463
HEP-SPEC06 score per watt	1.74	2.05	1.05	1.60

3.3. Multi-threaded Geant4 prototype

Geant4 is one the main toolkits used in the Large Hadron Collider (LHC) simulations. It is used to simulate the passage of particles through matter. It is a very good representative of real life workloads being run in Worlwide LHC Computing Grid (WLCG) and constitutes a significant part of its CPU time. The multi-threaded Geant4 prototype is one of the steps towards exploiting thread-level parallelism in event processing.

In this test we ran Geant4 [1][2], version 9.6-ref09a (released end of September 2013), which is an internal development version of Geant4 in preparation for v 10.0 (to be released in December 2013). The next major release will include event-level parallelism via multi-threading [3]: after the geometry and physics processes have been initialized, threads are spawned and events are simulated in parallel. To reduce memory footprint the read-only memory parts of the simulation area shared among threads. Memory increase for each additional thread has been measured to be in the order of 40-70 MB depending on the application [4].

We used the "ParFullCMS" application as a benchmark, which has been developed by the Geant4 collaboration and earlier optimized by openlab. The application uses realistic CMS experiment [5] geometry expressed in GDML format [6], where highly energetic (50 GeV) negatively charged pions are shot in random directions inside the detector. The recommended HEP physics list (FTFP_BERT) is used for simulating the interaction of particles with matter. Tracking in a solenoidal magnetic field is also included. No digitization of energy deposits is performed since this strongly depends on the experimental framework software and we are interested in evaluating the performance of Geant4 itself.

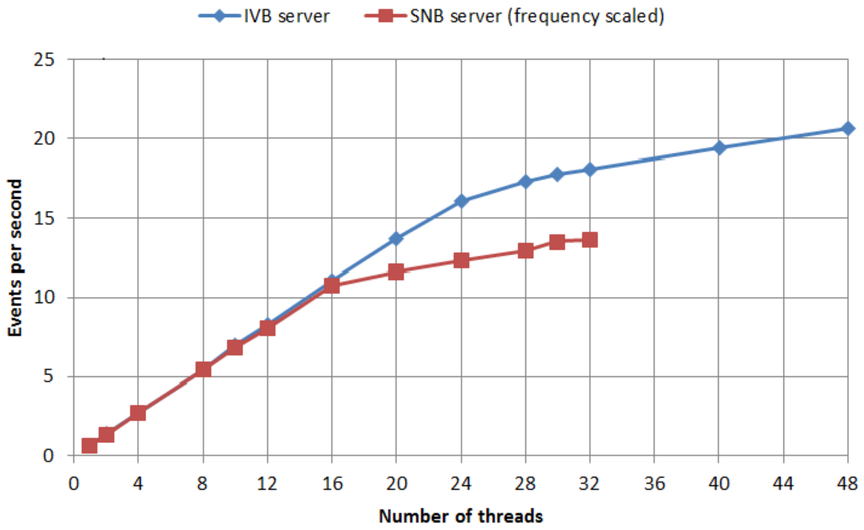

Figure 3. New MTG4 prototype – pi@50GeV events/s ("weak scaling" – preliminary results).

Performance is measured by keeping the workload of each thread constant (weak scaling) and measuring the time they spend in the event loop. The benchmark was compiled using icc with following switches: -O2, –fp-model-precise, –ansi, –fPIC. Results are preliminary.

3.3.1. Results.

Scalability results for the new prototype are shown on Figure 3. Linear scalability with nearly full efficiency (99% @ 24 cores) is observed until all cores are populated, and there is a 28% benefit from SMT. The platform provides good, predictable scaling for this C++ oriented benchmark.

3.4. Maximum likelihood fit prototype

The sheer amount of data produced by physics detectors requires that some of it be discarded, as storing all would be infeasible. Therefore, there’s a need to filter detector events and store only the most interesting ones that might lead to the discovery of new physics phenomena. One of the techniques for this is the maximum likelihood fit that is used in our benchmark. It is therefore representative a HEP analysis workload.

The benchmark was compiled with the latest icc 14.0, using various vectorization options: -no-vec, -xAVX, -xsse3 and in the case of Haswell, -xcore-avx2 as well.

3.4.1. Results. An important factor affecting this benchmark has been Hardware Prefetching – we have observed speedups of 2x simply by enabling BIOS hardware prefetching options.This application is memory bandwidth intensive and as a result prefetching helps reduce last level cache misses. The acquired results are presented in the Table 4. In particular, we enabled “DCU Data Prefetcher”, “MLC Streamer”, “MLC Spatial Prefetcher” and “DCU Instruction Prefetcher” in the conducted tests. All results that follow were obtained with these options enabled.

Table 4. MLFit execution times with respect to BIOS options.

	With prefetching	Without prefetching
Runtime (seconds)	**143.8**	289.7
LLC references	2,145,369,131	4,982,710,972
LLC misses	588,563,757	4,335,777,151
LLC miss ratio	27%	87%

Overall MLFit scalability, as measured on “Ivy Bridge” server, reaches 15.5x at 24 threads and 17x at 48 threads. This cannot be directly compared with previous results, since in this case we opted for a single process model. Multi-process results were reported earlier, and the compilers and runtimes have changed significantly in the past couple of years. Frequency scaled results show that this workload is 40% slower on the previous generation platform equipped with “Sandy Bridge” CPUs. The leftmost “dip” visible on Figure 4 comes from the penalty that is associated with a switch to a second socket. The second “dip” occurs as the workload is distributed onto SMT threads. In this case, there is little benefit from Hyper-Threading.

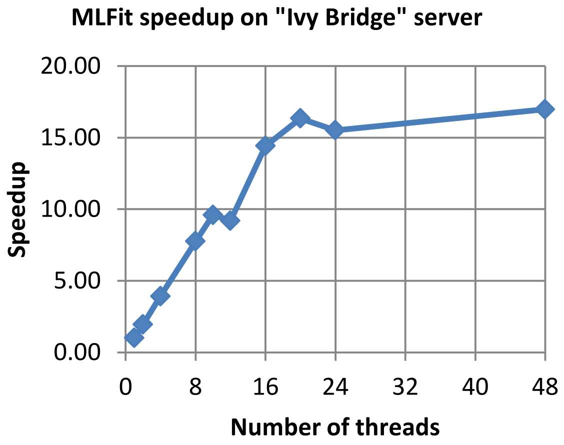

Figure 4. MLFit speedup on “Ivy Bridge-EP” E5-2695 v2.

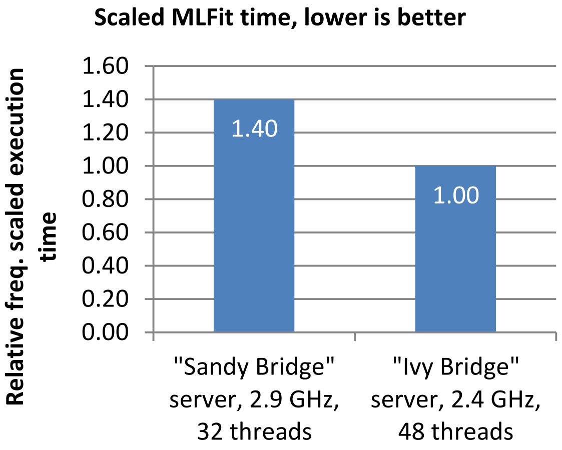

Figure 5. MLFit on ”Sandy Bridge-EP” E5-2690 and “Ivy Bridge-EP” E5-2695 v2 CPUs compared.

Although the “Ivy Bridge” and “Haswell” architectures are not prime candidates for this workload because of their low core count, an interesting comparison can still be made (Figure 6). Here we note that “Haswell” provides an improved AVX implementation that yields a small additional advantage when explicitly enabling the AVX2 flag in icc 14. The small AVX-AVX2 difference comes from the lack of major opportunities for FMA usage in the code.

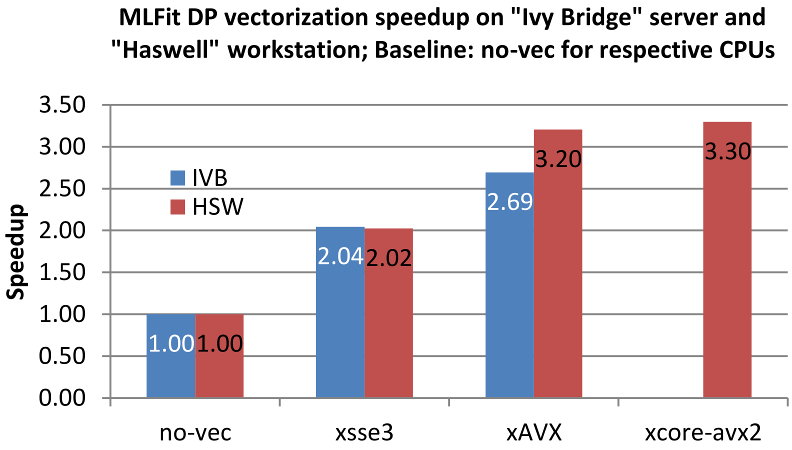

Figure 6. DP vectorization speedup on IVB server and HSW workstation.

3.5. The evolved Trackfitter

Since 2008, openlab has been collaborating with the GSI institute on a novel track fitting benchmark for the ALICE experiment. The workload has been since modernized by Ivan Kisel, Igor Kulakov and Maksym Zyzak of GSI [9][10] and ported to the Vc library developed by Matthias Kretz [11]. The workload is a vectorized Kalman filter, used in a track fitting algorithm. The results for a modernized version of the benchmark using OpenMP are reported below.

Intel C++ Compiler (icc) 14.0 was used to compile the benchmarks, with minor changes to Vc code for C++11 compatibility as understood by the icc compiler – the performance of some of Vc’s routines might have been slightly affected. The following flags were activated: –O3, -force-inline and one of –xSSE4.2, –xAVX or –xcore-avx2.

3.5.1. Results. In the case of fully loaded servers, the newer "Ivy Bridge" platform delivers 57% more throughput than its predecessor (frequency scaled). That can be attributed in a big part to the increased core count and in a smaller part to slightly improved vector handling in the "Ivy Bridge" core. When comparing workstations, "Haswell" delivers 45% more throughput thanks to FMA and extra execution ports. On Haswell, using a single thread, the correspondence in performance between the SSE version, the AVX version and the AVX2 variant is 1.0 : 1.64 : 1.94. This result also highlights the gap between reality and the theoretical SSE-AVX scaling factor of 2 – as suggested by the AVX vector size.

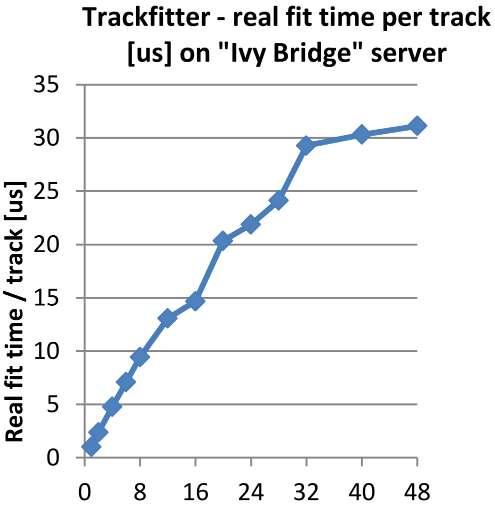

Figure 7. Trackfitter time scaling.

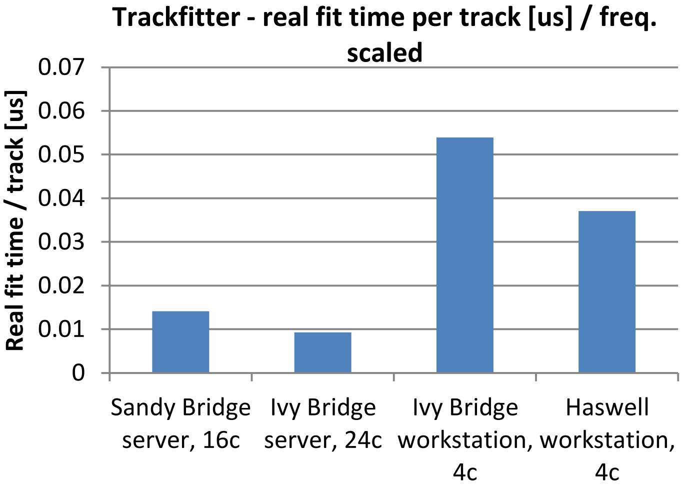

Figure 8. Trackfitter scaled time comparison.

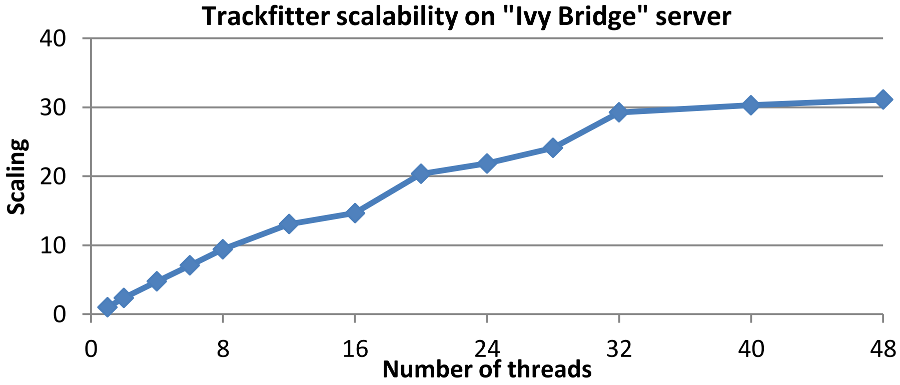

Figure 9. Trackfitter throughput scaling.

4. *Conclusions*

4.1. Comments on scalability and power

In the server comparison, the Ivy Bridge server performed close to theoretical predictions, giving a 42% frequency scaled improvement over its predecessor and a considerable 53% improvement in performance per watt. As far as workstations are concerned, the new "Haswell" microarchitecture provides highly visible benefits only for vectorized workloads. There, the power/performance ratio improved by 18.5% over the predecessor. Since this is only a workstation processor, wider reaching conclusions cannot be drawn at this point.

The advantage of SMT was well noted, at 19% for HEP-SPEC06 and at 30-40% for the revised Trackfitter. It is expected that this gain will translate onto HEP workloads, which still scale with the SPEC-derived benchmarks. However, in dense, bandwidth-oriented workloads, such as MLFit, SMT benefits were lower, around 10%.

4.2. Vector efficiency

We take note that only prototype HEP software and benchmarks attempt to make full use of vector units. This particular performance dimension is still growing in mainstream Xeon servers, but is not actively exploited by production HEP programs. Like HEP workloads, the HEP-SPEC06 benchmark is not suited for vectorization, even though it still exhibits good scalability across cores.

Given the above, the Haswell workstation processor, with its vector unit updates, does not provide significant improvements on production-like code. In order to take advantage of the benefits, more work must be done on the HEP software side. As demonstrated in the revised Trackfitter case, gains can be substantial. There, the difference between the non-vectorized x87 version and single precision AVX2 on Haswell reaches a factor of 7x!

4.3. Opportunities for low power

The platform we built based on “Ivy Bridge” processors exhibited highly improved power characteristics with respect to its predecessor, as observed in the past with Intel hardware. In particular, power constrained data centers can benefit from efficiency improvements.

Another interesting development comes from the low power consumer space – the Intel Atom processor is making its way into the data centers, and initial openlab tests (the results of which are still under NDA as of writing) demonstrate excellent SPEC/watt characteristics. Broadly speaking, low power architectures such as Atom or ARM have a data center niche to fill in, so we are looking forward to new computing products based on those architectures.

References

[1] Agostinelli S et al. 2003 Geant4 - a simulation toolkit *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated* Equipment **506** 250 - 303

[2] Allison J et al. 2006 Geant4 developments and applications *IEEE Transactions on Nuclear Science* **53** 270-278

[3] Dong X, Cooperman G and Apostolakis J 2010 Multithreaded Geant4: Semi-automatic Transformation into Scalable Thread-Parallel Software *Lecture Notes in Computer Science* **6272** 287–303

[4] Apostolakis J et al. 2014 The path toward HEP High Performance Computing *In this proceeding*

[5] CMS Collaboration 2008 The CMS experiment at the CERN LHC *Jinst* **3** 187

[6] Chytracek R 2001 The geometry description markup language *Proc. Computing in High Energy and Nuclear Physics* 473–476

[7] Jarp S, Lazzaro A, Nowak A, Leduc J. Evaluation of the Intel Sandy Bridge-EP server processor [Internet]. 2012.

[8] Busch A, Leduc J. Evaluation of energy consumption and performance of Intel’s Nehalem architecture [Internet]

[9] Kisel I, Kulakov I and Zyzak M Parallel Algorithms for Track Reconstruction in the CBM Experiment *Computing in High Energy and Nuclear Physics 2012*

[10] Kisel I, Kulakov I and Zyzak M Parallel Implementation of the KFParticle Vertexing Package for the CBM and ALICE Experiments *Computing in High Energy and Nuclear Physics 2012*

[11] Kretz M and Lindenstruth V 2012 Vc: A C++ library for explicit vectorization *Software: Practice and Experience* **42** 1409–30