

Platform Competence Center

**CERN openlab III
Major Review**

Sverre Jarpe
Alfio Lazzaro
Julien Leduc
Andrzej Nowak

January 25th, 2011



- ❑ Workshops / Teaching / Conferences
- ❑ Performance optimization and monitoring
- ❑ System evaluations
 - Intel Westmere and Sandy Bridge
 - Intel Micro-Server Proof-of-Concept
 - Intel Many Integrated Cores (MIC)
- ❑ Software evaluations and developments
- ❑ Future activities

- ❑ 4 workshops organized at CERN in 2010:
 - Computer Architecture and Performance Tuning: 17/18 February and 22/23 September
 - Performance optimization
 - Computer architecture
 - Compilers
 - Multi-threading and Parallelism: 4/5 May and 10/11 November
 - Parallel concepts descriptions
 - Multi-core architectures
 - Intel Threading Software tools
- ❑ Jeff Arnold from Intel's compiler group continues as a regular teacher
- ❑ Regular number of attendees: more than 20 persons per workshop (with some enthusiasts who attended more than 1 workshop)
 - They use the openlab machines for the hands-on session
- ❑ Same workshop structure planned for the 2011 (at least 4 workshops)

Teaching at International Schools

- CERN School of Computing 2010, 23 August – 3 September, 2010 (Uxbridge, UK)
 - 1 whole day: 3 lectures and 3 hours of exercises
 - Invited to next school in 2011

- INFN International School on “Architectures, tools and methodologies for developing large scale scientific computing applications”, 22 – 27 November, 2010 (Bertinoro, Italy)
 - First whole day of lecture: introduction, performance methodology (lectures + hands-on) + GPUs introduction
 - Last day on parallelization techniques: support to Tim Mattson (Intel) for the OpenMP lectures and exercises
 - Invited to next school in 2011

Recent Intel/Openlab Workshop at CERN

- Organized in the 4th quarter of 2010
 - “Intel Array Building Blocks (ArBB)”, 29 November
 - Hans Pabst
 - “Intel VTune Amplifier XE Performance Profiler”, 30 November
 - Levent Akyil

- Similar “special” workshops planned for 2011

Participation at CHEP 2010

- ❑ 3 presentations (1 plenary and 2 parallels) and 1 poster at the 18th International Conference on Computing in High Energy and Nuclear Physics (CHEP), 18 – 22 October, 2010 (Taipei, Taiwan)
 - Also contribution to the special session on GPUs
- ❑ Interesting discussions (several presentations) on multi-core and optimizations in HEP community
 - Expected more activity in this domain in the coming months/years
- ❑ 4 papers submitted for the review
 - They will be published on Journal of Physics: Conference Series



List of CHEP presentations

- Plenary by S. Jarp: “How to harness the performance potential of current Multi-Core CPUs and GPUs”
 - **Slides:** <http://117.103.105.177/MaKaC/materialDisplay.py?sessionId=2&materialId=4&confId=3>
- Parallel by A. Lazzaro (on behalf of A. Nowak): “Evaluating the Scalability of HEP Software and Multi-core Hardware”
 - **Slides:** <http://117.103.105.177/MaKaC/materialDisplay.py?contribId=77&sessionId=103&materialId=slides&confId=3>
 - **Full details of the evaluation in the published (2010) openlab reports:** https://openlab-mu-internal.web.cern.ch/openlab-mu-internal/03_Documents/3_Technical_Documents/2010/1_Technical_Reports/Doc_list.htm
- Parallel by A. Lazzaro: “Maximum likelihood fits using GPUs” (work done with a summer student, Felice Pantaleo)
 - **Slides:** <http://117.103.105.177/MaKaC/materialDisplay.py?contribId=297&sessionId=79&materialId=slides&confId=3>
- Poster: “The breaking point of modern processor and platform technology”
 - **Paper:** <http://cdsweb.cern.ch/record/1322444/files/CERN-IT-2011-005.pdf>
 - **Poster:** https://openlab-mu-internal.web.cern.ch/openlab-mu-internal/03_Documents/4_Presentations/Slides/2010-list/A-Nowak-CHEP2010-poster.pdf

Performance optimization and monitoring

- Intel Performance Tuning Utility 4 is fully available
 - Made critical stability improvements, especially in the area of CERN code
 - CERN openlab participated closely, also in the area of perfmon compatibility
 - Further work on the Westmere and Nehalem event maps with David Levinthal – improvements, dissemination
 - Additional ideas for a next generation tool
- Intel VTune Amplifier XE 2011 fully available
 - Beta test ended (CERN openlab participated heavily)
 - Tutorial workshop held for CERN experts by Levent Akyil (30 November 2010)
 - “This is genuinely the first piece of performance software that I see that works with our code out of the box” – a senior developer from one of the experiments
- Perfmon2: stable usage with current Scientific Linux 5.5

Intel VTune Amplifier XE 2011

Hotspots - View hotspots colored by CPU usage

Analysis Type | Collection Log | Summary | Bottom-up | Top-down Tree

Function	CPU Time	Module
grid_intersect	13.725s	tachyon_find_hotspots
initialize_2D_buffer	52.020s	tachyon_find_hotspots
setup_2D_buffer	52.020s	tachyon_find_hotspots
draw_trace		tachyon_find_hotspots
thread		tachyon_find_hotspots
intersect_objects	0.070s	tachyon_find_hotspots
light_intersect	0.130s	tachyon_find_hotspots
Selected (1 row(s)):		52.020s

Thread (0x7ffffff)

CPU Usage

Thread: Running, CPU Time

CPU Usage: CPU Time

No filters are applied. Module: [All] Call Stack Mode: Only user functions

□ No significant change since the last meeting

- 60 Viglen L5520 based systems
 - Standard CERN systems (CDB, IPMI management...)
 - 3 * 1TB disks in RAID0
 - Used for all the workshops (up to 40 systems used already)
- New Westmere systems expected in February
 - Around 60 dedicated new systems

□ System to be withdrawn

- Itaniums: we will keep 5-10 of them with an updated OS (RHEL 5.5) for some specific users at CERN (AMS experiment)
- Blade systems: encol101 (dual density blades: 32 Harpertown based systems) and encol301 (16 Harpertown based systems), just keeping one enclosure of 16 blades for openlab

□ Configuration of a shared file system for the user files

- 6TB (5x1.5TB drives in RAID5) exported over NFS (UNIX) and Samba (Windows) on Gb Ethernet



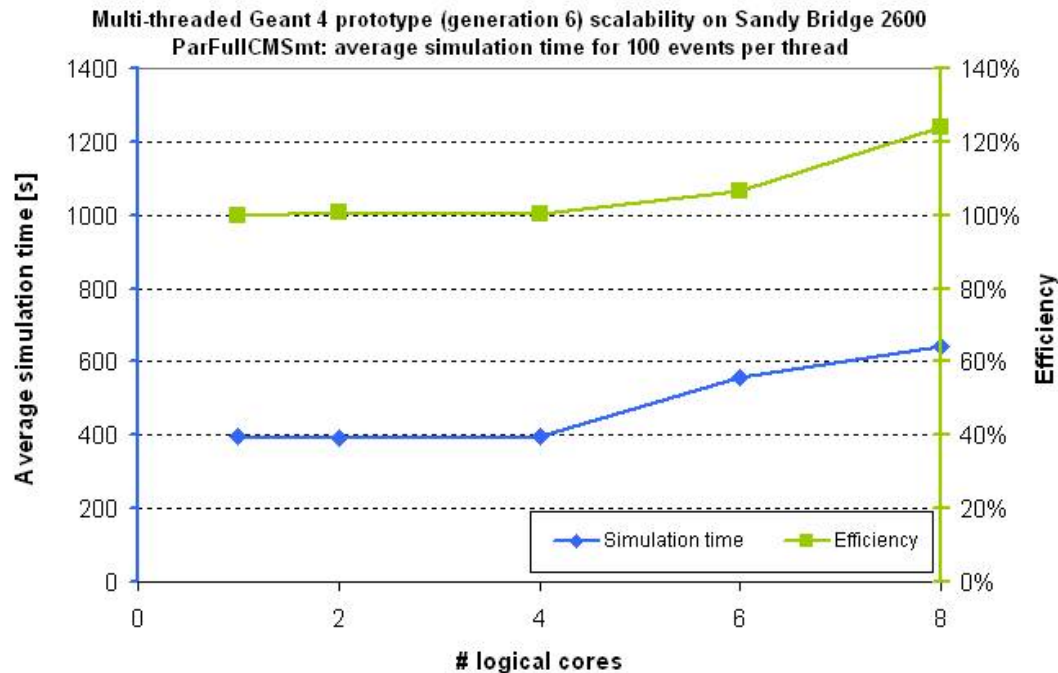
Hardware evaluation: Benchmarks

- “Bouquet” of applications to check different HEP workloads on different hardware
 1. HEPSPEC06 performance
 - “Brute” performance with the standard HEP benchmark
 2. Multi-threaded Geant4 prototype (Offline Simulation)
 - Throughput performance scalability (ptheaded workload)
 3. Parallel Maximum Likelihood fit with ROOT/RooFit (Data analysis)
 - Strong scaling (latency)
 - Prototype developed by us (Vectorized, OpenMP, MPI)
 4. ALICE Trackfitter/Trackfinder (Online)
 - Throughput performance scalability (Vectorized, pthread, OpenMP, ArBB)
- Other benchmarks
 - Power consumption vs performance: HEPSPEC06 per Swiss Franc per Watt
 - Non Uniform Memory Access aspects
 - Solid State Disk performance

- Intel Sandy Bridge (“tock” at 32nm)
 - New design respect to previous CPU, i.e. Westmere
 - Many new features, such as introduction of AVX instructions for vector operation at 256bit
 - Desktop version (single socket)
 - Core i7-2600 CPU @ 3.4GHz, 4 cores, 4GB memory
- AMD Magny-Cours (as reference to the Intel systems)
 - Last purchase at CERN, used for batch farm processing
 - Server version (4 sockets)
 - Opteron Processor 6164 HE @ 1.7GHz, 12 cores per processor, 48 cores in total, 96GB memory
 - Comparison with respect to Westmere-EP, Nehalem-EX systems
- Intel Micro-Server Proof-of-Concept
- Intel MIC “Knights Ferry Software Development Platform”

Multi-threaded Geant4 on Sandy Bridge

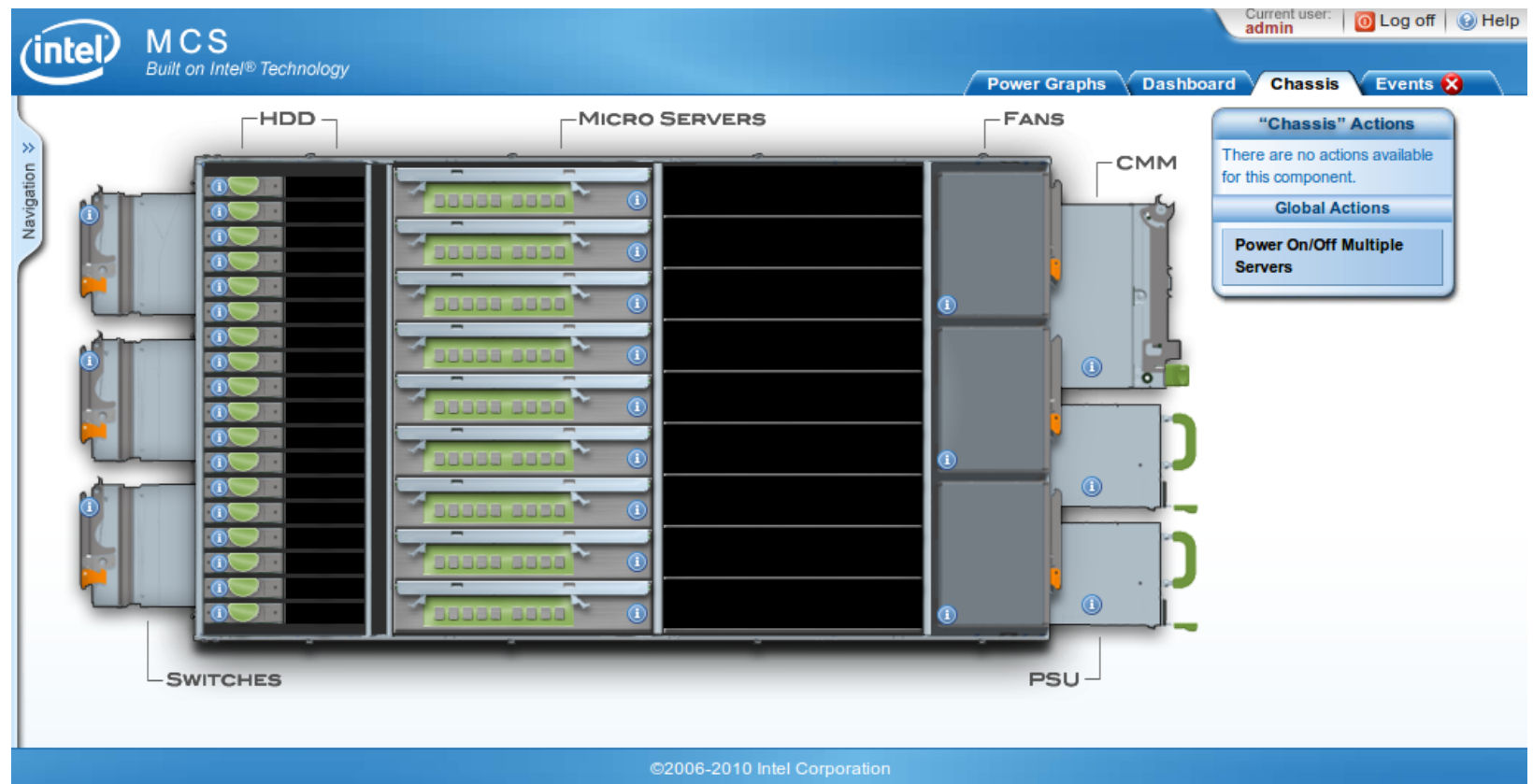
- ❑ Good scalability up to 4 cores
- ❑ Hardware multi-threading benefit is 25% (4 to 8 cores)
- ❑ Comparing to Westmere-EP, Core i7-2600 (Sandy Bridge based desktop) has ~10% better performance (frequency scaled)
 - Mainly due to the new chip design



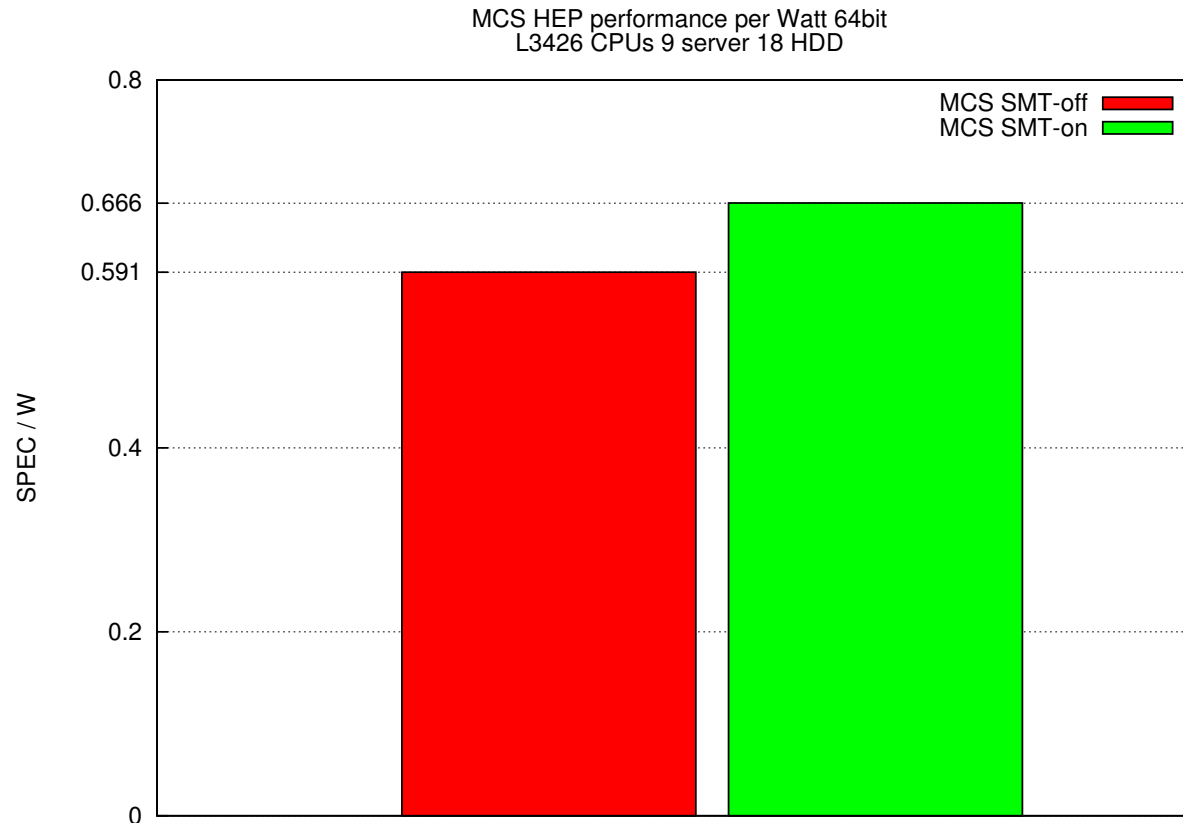
- ❑ More details will be available in a report which will be published in the upcoming months

Micro-Server Proof-of-Concept

- ❑ Small system that can be densely packed in a larger chassis
- ❑ Openlab system embeds
 - 9 microserver boards (18 fit in the chassis)
 - each microserver board counting 1x Intel Xeon Processor L3426 (Nehalem, 8M Cache, 1.86 Ghz, 4 core) + 4x2GB of memory and 2x120GB 2.5" SATA HDD



□ Results for HEPSPEC06 per Swiss Franc per Watt

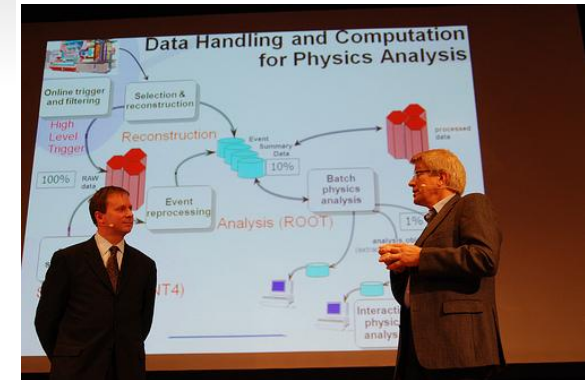


For reference Westmere-EP has:

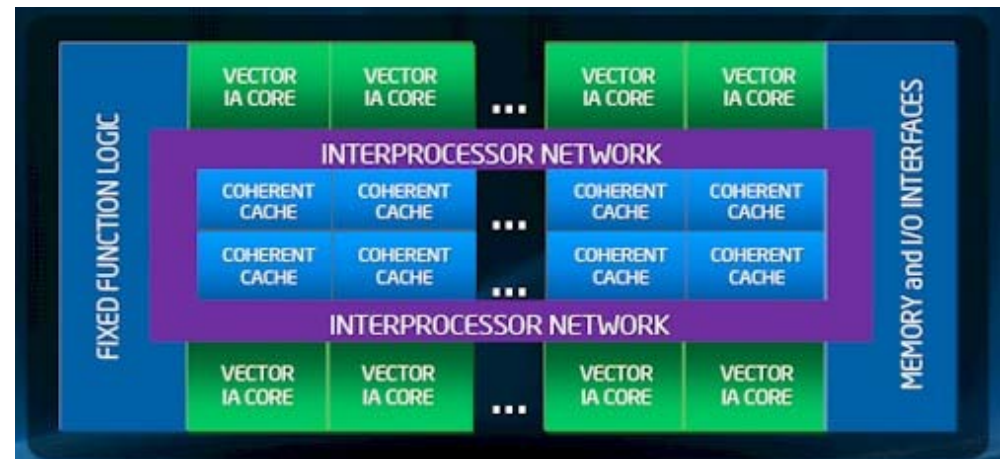
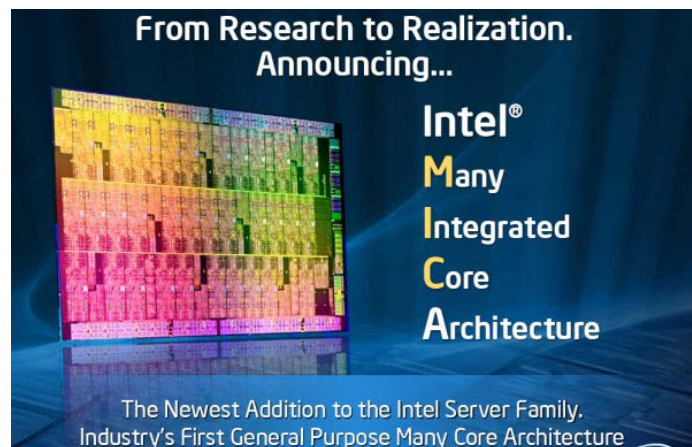
- SMT-off: 0.506 (-16%)
- SMT-on: 0.611 (-9%)

Intel “Many Integrated Cores” architecture

- ❑ Announced at ISC10 (June 2010)
 - S. Jarp participated at the presentation
- ❑ Current version (codename “Knights Ferry SDP”)
 - Enhanced x86 instruction set + vector extensions
 - 32 cores + 4-way hardware multithreaded + 512-bit vector/SIMD units
- ❑ Successful (easy) porting of our benchmark applications
 - ALICE Trackfitter/Trackfinder
 - Multithreaded Geant4 prototype
 - Maximum Likelihood data analysis prototype



Graphics: INTEL



□ Intel Parallel Composer 2011

- Compiler ICC v12.0, performance libraries, and supports Intel Parallel Building Blocks
- It includes vector support for AVX (supported by Sandy Bridge)
- We have participated during the beta testing, reporting bugs and regressions
- Officially released, but feedback is still being provided and features are being tested
 - Inside the Intel compiler project
 - Also working closely with Intel expert Jeff Arnold
- Comparison with latest GNU compiler versions
- Available at CERN for Linux, Windows and (first time) MacOSX

- Intel VTune Amplifier XE 2011 & Intel Inspector XE 2011
 - New tools for monitoring performance and checking correctness, respectively
 - Support parallel execution of the applications
 - We have participated during the beta testing, reporting bugs and regressions
 - Two workshops at CERN last year by Intel experts describing the tools
 - Officially released, but feedback is still being provided and features are being tested
 - Some adaptations needed to streamline work with large or very buggy applications
 - Available at CERN for Linux, Windows (not for MacOSX)
 - GUI available also in Linux (previous applications supported only Windows)
 - Same functionalities and interface
 - In the context of the openlab workshops, we have already moved to new tools (before their official publication), so we will be teaching the very latest technology

- ❑ **Development of “Oplabench” framework**
 - A framework for benchmarks, developed in collaboration with India (Imon Banarjee Master project)
 - Goal of the project is to ensure reproducibility of benchmarks across different systems, recording the parameters in a database
- ❑ **Evaluation of a nVidia GPU using CUDA**
 - We use the Maximum Likelihood data analysis application
 - Activity started by the summer student Felice Pantaleo last Summer
 - Submitted a paper to 12th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing, May 16-20, 2011, Anchorage (USA)
- ❑ **Evaluation of Intel Atom processor based system**
 - Recent test: Atom Pineview D510 @ 1.66GHz
- ❑ **Evaluation of Intel Solid State Disks**
 - Particularly interesting for applications with high I/O, such as database applications and physics data selection

- Visitors from Intel in the 4th quarter of 2010 (close collaboration with Intel teams in the US and in Europe)
 - HPC: Tim Mattson
 - Performance: David Levinthal (also collaborating with the experiments on improvements in their code)
 - Now in GOOGLE
 - Compilers and techniques: Jeff Arnold
 - Roadmap: Mark Myers
 - Tech: Herbert Cornelius, Klaus-Dieter Örtel, Levent Akyil, Hans Pabst
 - Mgmt/support: Claudio Bellini, Jean- Faust Mukumbi (replaced by Andrea Toigo)

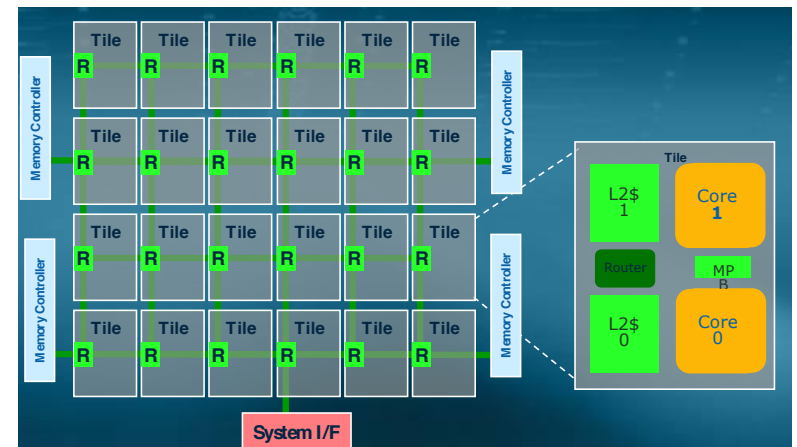
 - Visits/trips by openlab in 2010
 - ACAT (Jaipur, India)
 - ISC (Hamburg, Germany)
 - IDF (San Francisco, USA)
 - ERIC (Braunschweig, Germany)
 - CHEP (Taipei, Taiwan)
- Similar level of activity expected in 2011

Near term future activities (1)

- ❑ Preparation of the report for the evaluation of the Intel Sandy Bridge architecture
- ❑ Continuing server evaluations
 - Expecting Sandy Bridge EP and Westmere-EX
- ❑ Continuing “Knights Ferry SDP” activities
 - Benchmarking, performance analysis, compiler tests
- ❑ Evaluation of the vector capabilities of AVX on Sandy Bridge
- ❑ Continuing testing of the Intel tools and compiler versions (close collaboration with Intel experts)
- ❑ Testing node management technology
- ❑ Courses:
 - Openlab workshops (at least four in 2011)
 - Intel special workshops with physics community
- ❑ Evaluation of the Intel Single-chip Cloud Computer chip

Single-chip Cloud Computer

- ❑ 48 Core Research Microprocessor
 - Experimental Research Processor – Not A Product
 - “Cluster-on-die” architecture (new concept): 48 independent Pentium cores
 - Parallel programmability using MPI
 - Interesting possibilities: a lot of parameters can be configured via software, such as operational voltage and frequency
- ❑ Our research proposal was accepted and we are waiting for the system to be delivered
 - Participated in MARC forum in Braunschweig (9 November 2010)
 - Close relation with Tim Mattson, who described to us the chip during his visit at openlab in September



Near term future activities (2)

- Planning to move the Maximum Likelihood application code to use OpenCL parallelization
 - Particularly interesting for the Sandy Bridge, where GPU is integrated on the same die of the CPU
 - A new technical student, Yngve Sneen Lindal, will work on code optimization and accelerators (starting in February)
 - Close collaboration with Tim Mattson
- Improve our benchmark applications
 - Evaluate new technology for parallelization, such as Intel Concurrent Collections C++ (CnC)
- Update of the OS to the Scientific Linux CERN 6
- Several broad workshops planned

Backup slides

“How to harness the performance potential of current Multi-Core CPUs and GPUs”

Today:
Seven dimensions of multiplicative performance

First three dimensions:

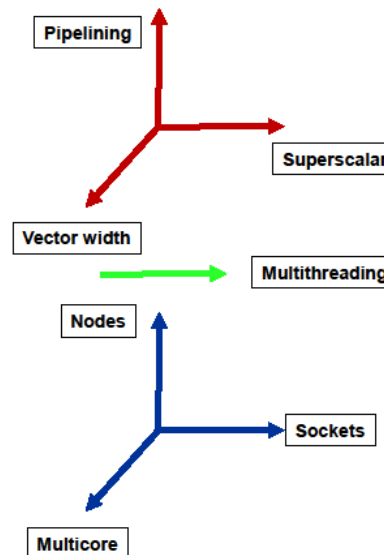
- Pipelined execution units
- Large superscalar design
- Wide vector width (SIMD)

Next dimension is a “pseudo” dimension:

- Hardware multithreading

Last three dimensions:

- Multiple cores
- Multiple sockets
- Multiple compute nodes



SIMD = Single Instruction Multiple Data

Sverre Jarp - CERN

What are the multi-core options?

- There is a discussion in the community about the best way(s) forward:

- 1) Stay with event-level parallelism (and entirely independent processes)
 - Assume that the necessary memory remains affordable
 - Or rely on tools, such as KSM, to help share pages
- 2) Rely on forking:
 - Start the first process; Run through the first “event”
 - Fork N other processes
 - Rely on the OS to do “copy on write”, in case pages are modified
- 3) Move to a fully multi-threaded paradigm
 - Still using coarse-grained (event-level) parallelism
 - But, watch out for increased complexity

17

Sverre Jarp - CERN

Presentation at: <http://117.103.105.177/MaKaC/materialDisplay.py?sessionId=2&materialId=4&confId=3>

“Evaluating the Scalability of HEP Software and Multi-core Hardware”

1. HEPSPEC06 performance
 - a standard HEP benchmark
2. Multi-threaded Geant4 prototype scalability (J. Apostolakis et al. *Multithreaded Geant4: Semi-automatic transformation into scalable thread-parallel software*, Europar 2010)
 - parallel implementation of the test40 example from Geant4
 - 200 random events per thread
 - ParFullCMSmt, a full CMS simulation ported to a parallel model
 - 100 pi- events per thread @ 300 GeV
3. MPI Parallel Maximum Likelihood (ML) fit with ROOT/RooFit (A. Lazzaro and L. Moneta, *MINUIT package parallelization and applications using the RooFit package*, *J. Phys.: Conf. Ser.* **219** 04204)
4. Power consumption vs performance
5. NUMA aspects (Nehalem-EX)

- Westemere-EP VS Nehalem-EP
 - 50% core increase, but HEPSPEC06 numbers only 32% better
 - Overall improvements between 39% and 61% (mostly due to core increase)
 - SMT benefit: 15% - 24% (unchanged)
 - 10% - 23% performance per Watt improvement
 - The previous transition (Core 2 -> Nehalem) was ~35%
- Nehalem-EX VS Dunnington (frequency scaled)
 - 33% core increase reflected in performance
 - Total TP increase: 3.5x on HEPSPEC06!
 - Credited to weak Dunnington performance
 - 47% - 87% more TP on in-house applications
 - SMT benefit: 19% - 28% (no SMT on Dunnington)
 - Significant power consumption

Presentation at: <http://117.103.105.177/MaKaC/materialDisplay.py?contribId=77&sessionId=103&materialId=slides&confId=3>

Full details of the evaluation in the published (2010) openlab reports:
https://openlab-mu-internal.web.cern.ch/openlab-mu-internal/03_Documents/3_Technical_Documents/2010/1_Technical_Reports/Doc_list.htm

“Maximum likelihood fits using GPUs”

Test environment

Work done with a summer student, Felice Pantaleo

- PCs
 - CPU: Nehalem @ 3.2GHz: 4 cores – 8 hw-threads
 - OS: SLC5 64bit - GCC 4.3.4
 - ROOT trunk (October 11th, 2010)

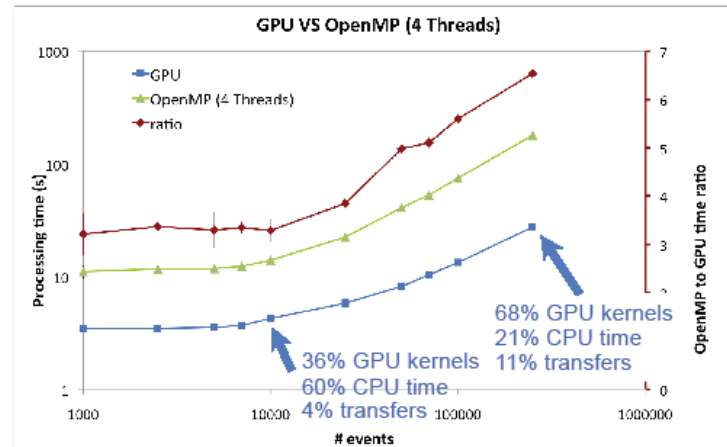
- GPU: ASUS nVidia GTX470 PCI-e 2.0
 - Commodity card (for gamers)
 - Architecture: GF100 (Fermi)
 - Memory: 1280MB DDR5
 - Core/Memory Clock: 607MHz/837MHz
 - Maximum # of Threads per Block: 1024
 - Number of SMs: 14
 - CUDA Toolkit 3.1 06/2010
 - Developer Driver 256.40
 - Power Consumption 200W
 - Price ~\$340



Alfio Lazzaro (alfio.lazzaro@cern.ch)

PDF-event-base: GPU VS OpenMP

- Fair comparison
 - Same algorithm
 - Algorithm on CPU optimized and parallelized (4 threads)
 - CPU does the final sum of the *NLL* and normalization integral calculations
- Check that the results are compatible: asymmetry less than 10^{-12}



- Speed-up increases with the dimension of the sample, taking benefit from the data streaming on GPU and the integral calculation only on the CPU
- ~3x for small samples, up to ~7x for large samples

Alfio Lazzaro (alfio.lazzaro@cern.ch)

15

Presentation at: <http://117.103.105.177/MaKaC/materialDisplay.py?contribId=297&sessionId=79&materialId=slides&confId=3>

“The breaking point of modern processor and platform technology”

The breaking point of modern processor and platform technology

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

The quest for performance

Increased accuracy and detail of processing is expected
Constantly increasing demand for computing resources
Future upgrades of the LHC: data rates higher up to 2 orders of magnitude

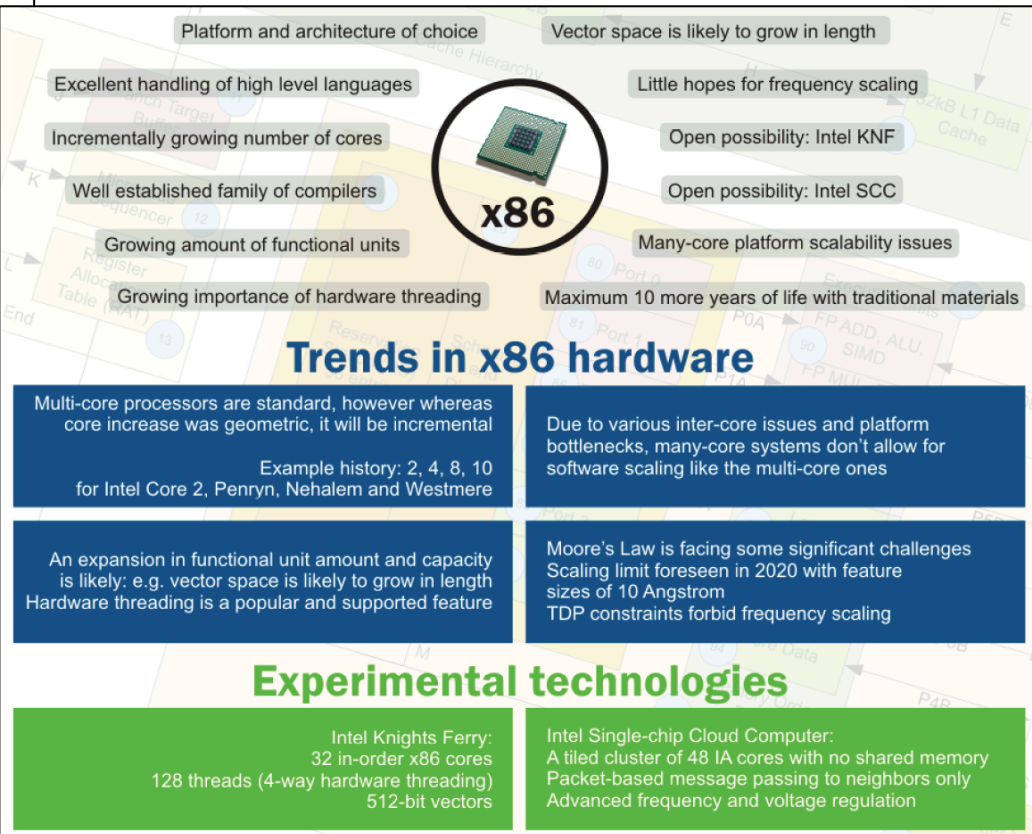
Trends in HEP software

Significant (at least an order of magnitude) speedups possible through careful algorithmic and precision related considerations

In spite of being blessed with trivial parallelism, HEP code is very rarely making use of multi-threaded technologies to save memory
Post-initialization forking is a good start

Advanced, highly beneficial processor features remain unused, wasting capacity: 30% wasted for hardware threading, over 75% for vectors
64-bit not used widely

Complex HEP C++ code produces binaries which are not processor-friendly: processors are under-utilized, to the point where 75%-90% of theoretical execution capacity in a core is wasted

Trends in x86 hardware

- Platform and architecture of choice
- Excellent handling of high level languages
- Incrementally growing number of cores
- Well established family of compilers
- Growing amount of functional units
- Growing importance of hardware threading
- Vector space is likely to grow in length
- Little hopes for frequency scaling
- Open possibility: Intel KNF
- Open possibility: Intel SCC
- Many-core platform scalability issues
- Maximum 10 more years of life with traditional materials

Multi-core processors are standard, however whereas core increase was geometric, it will be incremental Example history: 2, 4, 8, 10 for Intel Core 2, Penryn, Nehalem and Westmere	Due to various inter-core issues and platform bottlenecks, many-core systems don't allow for software scaling like the multi-core ones
An expansion in functional unit amount and capacity is likely: e.g. vector space is likely to grow in length Hardware threading is a popular and supported feature	Moore's Law is facing some significant challenges Scaling limit foreseen in 2020 with feature sizes of 10 Angstrom TDP constraints forbid frequency scaling

Experimental technologies

Intel Knights Ferry: 32 in-order x86 cores 128 threads (4-way hardware threading) 512-bit vectors	Intel Single-chip Cloud Computer: A tiled cluster of 48 IA cores with no shared memory Packet-based message passing to neighbors only Advanced frequency and voltage regulation
--	--

Paper at: <http://cdsweb.cern.ch/record/1322444/files/CERN-IT-2011-005.pdf>

Poster at: https://openlab-mu-internal.web.cern.ch/openlab-mu-internal/03_Documents/4_Presentations/Slides/2010-list/A-Nowak-CHEP2010-poster.pdf