Intel neue Prozessorgeneration
Sandy Bridge:
Neuerungen und erste Erfahrungen

J. Treibig
RRZE
Zeuthen 19.5.2011
How to achieve Performance (for data intensive FP codes)

Performance is enabled through:

- **ILP (Pipelining 3-4X, Superscalar 2X)**
- **SIMD (DP 2-4X, SP 4-8X)**
- **Thread level parallelism (4-32X)**
- **NUMA (2-4X)**
- **Caches (4-6X)**
- **Distributed memory parallelism (unlimited 😊, 1000X)**

Node Performance: 1TFlops/s (200-300 GFlops/s)

50-100 GB/s memory bandwidth
Differences to Micro Architecture Intel Nehalem

- **Successor of P6 architecture**
- **1.5 k uop Cache** in L1 level
- **CISC instructions are converted in uops**
- Very mature implementation of out-of-order superscalar design
- **New OOO logic**
- **Throughput 4 uops/cycle**
- **Two 16 byte transfers per cycle**
- **Two 16-byte wide SIMD instructions/cycle (DP 2 multadd)**
- **AVX:** 256bit SIMD registers, 3 operand instructions

hpc@rrze.uni-erlangen.de
New Segmented L3 cache with Ringbus. Similar to Nehalem EX solution.
Bottomline: What does that mean for me?

- Identical Instruction code runs more efficient: CPI for SSE code on Nehalem 0.7-0.9, on Sandy Bridge 0.5-0.6
- It is less sensitive against optimizations: all codes run fast 😊
- Load throughput limited SSE codes (Stencil codes) benefit of the doubled load performance
- AVX results in double FP performance (Compilers get better in vectorization)
- And this is achieved with an even higher clock

Sandy Bridge is the strongest x86 single core ever!
Medical volume reconstruction: RabbitCt

- 3D Volume reconstruction used in CT applications
- Open competition benchmark for comparable results
- Uses CT images from a reference rabbit
- Runtime limit: 15s for complete backprojection
Results: Harpertown vs. Sandy Bridge

![Graph showing performance comparison between Harpertown and Sandy Bridge]

- 54 GFlop/s
- 56 GFlop/s
- # threads
- Blocking
- Load balancing
- Clipping
- SSE
- Plain C

hpc@rrze.uni-erlangen.de
Results: Westmere vs. Nehalem EX

101 GFlop/s

20 sec reconstr.

159 GFlop/s

20 sec reconstr.
Likwid measurement on SMT and blocking influence

Westmere single socket timelines
Problems with current AVX extension

- Only a subset of instructions support the full 256 bit register width
- There is no hardware gather/scatter instruction (Larabee)
- The instruction set is complex (over 350 instructions)
- The penalty for non vectorized code gets larger
CPU vs. GPGPU

![Graph comparing relative performance of different CPUs and GPGPUs.](image)

- 512^3
- 1024^3

Relative performance

- GTX260 (OpenCL)
- FX5800 (CUDA)
- Westmere
- Nehalem EX
- Sandy Bridge

20 sec (512^3)
Conclusion

- Sandy Bridge is a big step forward in performance
- Also non vectorized and SSE codes perform much better
- CPUs are improving. The next Xeon generation will be highly competitive to GPUs even for GPU suited Algorithms
- AVX needs to be improved to allow more algorithms to fully benefit from it
University spin-off offering
- Consulting
- Training
- Services
to industrial partners.

Website: http://likwid-software.com