High Performance Computing

Roofline

Project 3

Johannes Winklehner 1226104

Armin Friedl 1053597

June 23, 2016

A *roofline model* for a multicore-processor is obtained by calcuating the theoretical peak performance of the processor and benchmarking the peak memory bandwith. Two artificial computational kernels with arithmetic intensities of $\frac{1}{16}$ GFLOPs/Byte and 8 GFLOPs/Byte are devised. The performance of the two kernels is then compared to the theoretical calculations in the roofline model.

Contents

1 Introduction

2 Roofline Model

In this section a roofline model [7] will be created for the Intel® Core™ i5-4210U. In Section 2.1 the theoretical floating-point peak performance of the CPU is calculated. Section 2.2 then shows memory bandwidth measurements gathered with NUMA-STREAM [1]. These ingredients are put together into the roofline model which is constructed in Section 2.3.

2.1 Theoretical Peak Performance

The CPU under test was a Intel® Core™ i5-4210U. Table 1 shows the relevant specifications for this processor according to Intel Ark [5].

Specification	Value
# of Cores	2
# of Threads	4
Microarchitecture	Haswell
Max Turbo Frequency	2.7 GHz
Processor Base Frequency	1.7 GHz
Instruction Set Extension	SSE 4.1/4.2, AVX 2.0

Table 1: Relevant processor specifications

According to Intel [3, 5-2 Vol.1] the 4th generation Intel Core processors provide FMA (Fused Multiply-Add) units and AVX (Advanced Vector Extension). Whereas AVX can be the main driver for floating-point peak performance, the peak in this case is mainly determined by the FMA unit.

In general an FMA unit is capable of multiple floating-point (FP) operations during a single cycle. This is directly backed by the hardware (operations are *"fused"* together). Specifically the FMA unit of a Haswell processor is capable of "[...] 256-bit floating-point instructions to perform computation on 256-bit vectors" [3, 5-28 Vol.1].

Since even a DP (double-precision) FP element has only 64-bit, 256-bit would be obviously overprovisioned. But the FMA instructions do not just take scalars as arguments. Instead up to 4 DP FP elements can be packed together in a vector and operations are conducted pairwise. An example mulitply-add instruction is given in [4].

Unfortunately no definite source could be found but according to Shimpi [6] the Haswell architecture is built with 2 FMA units per core. Taking all together we get:

1. Two operations are conducted at once ("fused") and up to four DP FP elements can be packed into the argument vectors. At optimal untilization the FMA unit therefore provides $2 * 4 = 8$ DP FLOPs each cycle.
2. Two cores each with two FMAs can then calculate $2 * 2 * 8 = 32$ DP FLOPs

At maximum turbo frequency the processor therefore has a theoretical peak performance of $32 * 2.7 = 86.4$ GFLOP/s. At base frequency it is capable of $32 * 1.7 = 54.4$ GFLOP/s.

2.2 Memory Bandwidth

To benchmark the memory bandwidth NUMA-STREAM [1] was used. The binary ran on a Fedora 23 system with kernel `4.5.7-200.fc23.x86_64 x86_64` in `multi-user.target` to turn off as many distractors as possible. Compilation was done with gcc and the following options: `-O3 -std=c99 -fopenmp -lnuma -DN=80000000 -DNTIMES=100`.

Again the details of the processor architecture offer a bit of a challenge. The i5-4210U is hyper threaded meaning it provides 4 hardware threads on 2 physical cores. It is not immediately obvious how many threads NUMA-STREAM should be configured with. For this test both configurations[1] were tested and the best one was chosen. The results for NUMA-STREAM configured with two threads are listed in Listing 1. Prefixes are given in metric scale, i.e. $M = 10^6$ not 2^{20}.The highest achieved rate was 10608 MB/s with the triad function. The triad function is the most demanding kernel of NUMA-STREAM defined at [2] as `a[j] = b[j]+scalar*c[j]`. All other tested configurations had worse results for all 4 kernels although with at most 300 MB/s difference.

```
Function      Rate (MB/s)   Avg time     Min time     Max time
Copy:         9373.3846       0.1368       0.1366       0.1390
Scale:        9414.1304       0.1361       0.1360       0.1381
Add:         10614.6002       0.1812       0.1809       0.1835
Triad:       10607.7910       0.1813       0.1810       0.1834
```

Listing 1: NUMA-STREAM results for two threads

2.3 Graph

The graph of the roofline model is defined by [7]:

```
Attainable GFLOP/s = Min(Peak FLOP,
                         Peak Memory Bandwidth*Operational Intensity)
```

The resulting graph for the values obtained in Section 2.1 and Section 2.2 can be seen in Figure 1.

[1] plus two configurations with 8 and 1 threads respectively for cross checking

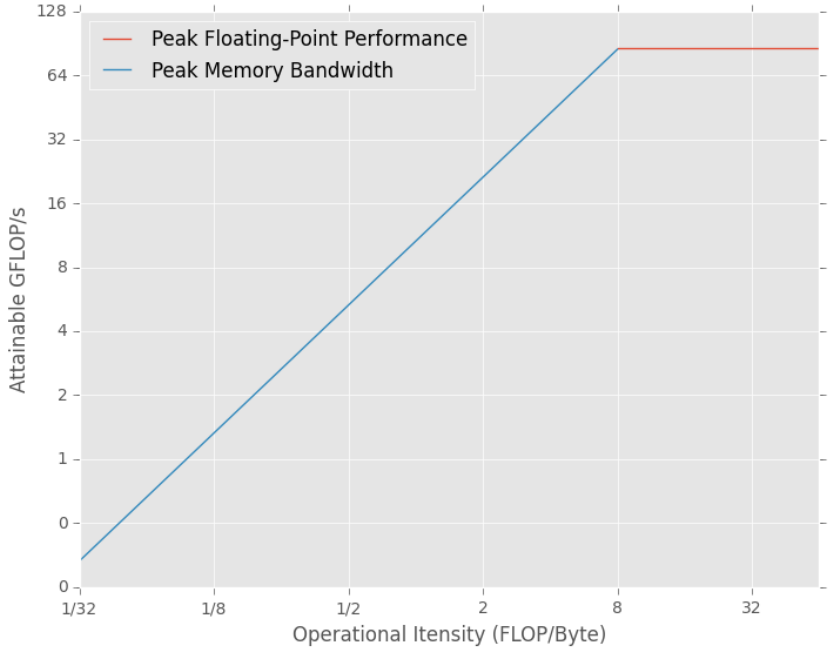

Figure 1: Roofline graph from the values obtained in Section 2.1 and Section 2.2

3 Kernels

References

[1] Lars Bergstrom. *NUMA-STREAM*. URL: https://github.com/larsbergstrom/NUMA-STREAM (visited on 06/20/2016).

[2] Lars Bergstrom. *stream.c*. URL: https://github.com/larsbergstrom/NUMA-STREAM/blob/e5aa9ca4a77623ff6f1c2d5daa7995565b944506/stream.c#L286 (visited on 06/20/2016).

[3] Intel. *Intel® 64 and IA-32 Architectures Software Developer's Manual. Combined Volumes: 1, 2A, 2B, 2C, 3A, 3B, 3C and 3D*. Intel. Apr. 2016. URL: https://www-ssl.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-manual-325462.pdf.

[4] Intel. *Intel Intrinsics Guide*. URL: https://software.intel.com/sites/landingpage/IntrinsicsGuide/#techs=AVX2,FMA&text=madd&expand=2365 (visited on 06/19/2016).

[5] Intel Ark. *Intel® Core™ i5-4210U Processor Specifications*. URL: http://ark.intel.com/products/81016/ (visited on 06/19/2016).

[6] Anand Lal Shimpi. *Haswell's Wide Execution Engine*. Oct. 5, 2012. URL: http://www.anandtech.com/show/6355/intels-haswell-architecture/8 (visited on 06/19/2016).

[7] Samuel Williams, Andrew Waterman, and David Patterson. "Roofline: an insightful visual performance model for multicore architectures". In: *Communications of the ACM* 52.4 (2009), pages 65–76.