Algorithm Engineering Lab Assignment 8

Brian Zahoransky (brian.zahoransky@uni-jena.de)

January 20, 2021

1. Explain the naming conventions for intrinsic functions.

_<vector_size>_<operation>_<suffix>

Commonly, a name for an intrinsic procedure consists of three parts separated by an underscore as shown above. The first part indicates the size of the returned vector. Typically, its 'mm', 'mm256', or 'mm512' which stands for 128-bit, 256-bit, or 512-bit vectors.

The second part specifies the operation of the intrinsic procedure. Operations like add, sub, mul were listed in the lecture as well as more compelex ones like fmadd which represents fused-multiply-add. A fused-multiply-add operation performs a multiplication on the first both arguments and adds the third input vector, all within a single instruction.

The third part ,which can also be declared as suffix, describes the type of the main input arguments. In this context, 'p' stands for packed and is commonly used with 's' or 'd' meaning single-precision or double-precision floating point numbers. The letters 'ep' are the short version for extended packed an is typically used in combination with 'i' or 'u' which is short for either a signed or unsigned integer. Referring to the following link, intel seems to have added the letter 'e' just for distinction. (https://stackoverflow.com/questions /64600563/what-is-packed-and-unpacked-and-extended-packed-data) Furthermore, the integer suffix includes a number, too. It expresses the size of the integer numbers (e.g. 8-bit, 16-bit, 32-bit, ...). However, packed only says that the numbers are stored linear without any padding.

2. What do the metrics latency and throughput tell you about the performance of an intrinsic function?

Latency shows the number of clock cycles which are needed to complete an operation. For example, a processor of the broadwell architecture needs five cycles to complete a fused-multiply-add operation. **Throughput** describes how often it is possible to start an operation of the same kind. For example, the broadwell processor is able to start two independent fused-multiply-add functions each clock cycle. Thus, the throughput is 0.5.

3. How do modern processors realize instruction-level parallelism?

Instruction-level parallelism means that distinct operations can be performed by a single core at once. This is enabled by a scheduler which is able to access multiple ports. Each port provides different instruction units, such as arithmetic logic units (ALU), vector ALUs, floating point units of different kinds, loading / storing units, and so on. The instruction units of one port are independent from the instruction units of all the the other ports. Due to this, the best performance will be achieved if the scheduler is able to assign work to all ports.

4. How may loop unrolling affect the execution time of compiled code?

Loop enrolling may affect the execution time on different stages, thus there is no ever valid answer. The goal of loop enrollment is to give compilers a hint for using vector instructions. As in the lecture presented, hints provided by the use of openMP might be interpreted very different. Some compilers nearly reach the same performance as the loop were unrolled with intrinsic procedures whereas others performing worse than if the for loop were not unrolled at all. Furthermore, the execution time also depends on the processor architecture. While for a processor with avx512 support an enrolment of 16 could be the best, on a processor only supporting avx2 instructions a enrolment of 8 might be faster.

5. What does a high IPC value (instructions per cycle) mean in terms of the performance of an algorithm?

A high IPC value indicates a high instruction-level parallelism. Operating on huge arrays, it also allows to conclude that the algorithm has a few cache losses. However, a high IPC score says nothing about the quality of a algorithm. To substantiate this, bubble sort was mentioned in the lecture. Even if the algorithm is optimized to have a high IPC value, it would have a squared running time.