High-Performance Computing Group
Prof. Dr. Christian Plessl
Computer Science Department and
Paderborn Center for Parallel Computing

Floating-Point Number De/Compression on CPUs

Bachelor/Master Thesis

At a glance

- The compression and decompression of floating-point numbers to/from n-bit quantized numbers will be implemented and optimized for different CPU-architectures.

In high-performance computing for scientific simulations it is often possible to approximate 64-bit floating-point numbers by quantization. A 64-bit floating-point number x is then approximated via y=ANINT(x*eps), where eps is the quantization value and ANINT returns the nearest integer number. y is then a n-bit signed integer. Since the bit-length n of y can be less than the 64 bit of x, less memory is required to store and transfer y, which results in faster transmission or lower storage size requirements.

The compression and decompression on CPUs will be implemented/optimized in this project making use of AVX2 and AVX512 instructions for AMD and Intel CPUs with the help of compiler intrinsics (function wrappers for specific CPU instructions).

Depending on the interest of the student also a GPU implementation can be attempted.

Further reading:
https://arxiv.org/abs/2303.13632 see section II.D on page 3
https://www.intel.com/content/www/us/en/docs/intrinsics-guide/index.html

Contact:
Robert Schade, E-Mail: robert.schade@uni-paderborn.de