

PERNIX: Floating-Point Number De/Compression on CPUs

Authors:

Felix Sternberg, Robert Schade, Christian Plessl

Background: The memory wall limits many HPC applications in which large floating-point datasets must be transferred between accelerators, CPUs, and memory. In hybrid density-functional theory, electron repulsion integrals (ERIs) in codes such as CP2K are a representative case: their transfer and storage costs can dominate execution, motivating controlled lossy compression.

Aims: We present PERNIX, an open-source C++ library for SIMD-optimized quantization and bit-packing of Float32 data, available at <https://github.com/pc2/pernix>. The goal is to provide a maintainable, application-integrable implementation that improves de/compression throughput sufficiently to reduce CPU involvement and expose bandwidth limits in heterogeneous HPC pipelines.

Method: PERNIX follows the block-wise quantization scheme of Guidon et al. For each block, values are scaled by the maximum absolute value, quantized to signed 8–16 bit integers, and densely packed into 512-bit blocks containing $\text{floor}(512/n)$ values. Decompression performs unpacking, sign extension, and rescaling to Float32. The library provides C/C++ and Fortran APIs and aligns its block format with FPGA compression schemes by Wu et al. We implement three x86 variants: AVX2, AVX2+BMI2 using PEXT/PDEP, and AVX-512-VBMI using wide byte permutations and vector bit manipulation. Byte-aligned widths receive specialized fast paths. Benchmarks were run on one core of an AMD EPYC 9655 on the Otus cluster at Paderborn University with GCC 14.3.0. Core-throughput experiments exclude memory transfers, while end-to-end experiments include load/store effects across cache and DRAM. MAQAO was used to analyze port pressure and guide instruction-level optimization.

Results: The AVX-512-VBMI implementation reaches up to 1500 MBlocks/s for 8-bit compression, about 2–3× faster than the AVX2 variants, which reach up to 700 MBlocks/s. For N=8 bit, AVX-512-VBMI achieves up to 300 GB/s in L1 and becomes transfer-bound beyond L2, following measured memcpy bandwidth limits. Non-byte-aligned widths such as N=13 bit remain compute-bound longer due to permutation-heavy packing. Decompression reaches memory-bandwidth limits earlier because unpacking is simpler than packing.

Conclusion: PERNIX substantially improves throughput over the current CP2K-style implementation and can saturate memory bandwidth with fewer CPU cores. This makes compressed ERI reuse and accelerator-to-CPU data exchange more efficient while preserving a controlled quantization error model.