Selected References on RoundOdd

When double rounding is Odd

Many .. processors .. may not always produce the correctly rounded result of a floating-point operation due to double rounding. Instead of rounding the value to the working precision, the value is first rounded in an intermediate extended precision and then rounded in the working precision; this often means a loss of accuracy. We suggest the use of rounding to odd as the first rounding in order to regain this accuracy: we prove that the double rounding then gives the correct rounding to the nearest value. To increase the trust on this result, as this rounding is unusual and this property is surprising, we formally proved this property using the Coq automatic proof checker.

> S. Boldo and G. Melquiond, "When double rounding is odd," Dec. 2021, [Online]. Available: https://www.lri.fr/~melquion/doc/05-imacs17_1-article.pdf

Emulation of a FMA and Correctly Rounded Sums

Rounding to odd is a non-standard rounding on floating-point numbers. By using it for some intermediate values instead of rounding to nearest, correctly rounded results can be obtained at the end of computations.

> S. Boldo and G. Melquiond, "Emulation of a FMA and Correctly Rounded Sums: Proved Algorithms Using Rounding to Odd," *IEEE Transactions on Computers*, vol. 57, no. 4, pp. 462–471, Apr. 2008, doi: 10.1109/TC.2007.70819. Available: odd rounding.pdf.

Bfloat16 processing for Neural Networks

This paper proposes a possible implementation of a [BFloat16] multiply accumulation operation that relaxes several IEEE Floating-Point Standard features to afford low-cost hardware implementations. Specifically, subnorms are flushed to zero; only one nonstandard rounding mode (Round-Odd) is supported; NaNs are not propagated; and IEEE exception flags are not provided. The paper shows that this approach achieves the same network-level accuracy as using IEEE single-precision arithmetic ("FP32") for less than half the datapath area .. with greater throughput.

> N. Burgess, J. Milanovic, et. Al., "Bfloat16 Processing for Neural Networks," in *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, Jun. 2019, pp. 88–91. doi: 10.1109/ARITH.2019.00022 url: https://ieeexplore.ieee.org.

RLibm-Prog: Fast Correctly Rounded Math Libraries

The key idea behind RLibm-All is to generate a polynomial approximation that produces correctly rounded results for a floating-point (FP) representation with two additional bits of precision (i.e., n +2-bits) using the round-to-odd mode.

> M. Aanjaneya, J. P. Lim, and S. Nagarakatte, "RLIBM-PROG: Progressive Polynomial Approximations for Fast Correctly Rounded Math Libraries." arXiv, Mar. 17, 2022. doi: 10.48550/arXiv.2111.12852.