Hardware Implementation of a Bfloat16 Exponential Function for Softmax Computation

Radosław Feiglewicz, Andrzej Kos
AGH University of Krakow
Faculty of Computer Science, Electronics and Telecommunications
Krakow, Poland
feiglewicz@agh.edu.pl, kos@agh.edu.pl

EXTENDED ABSTRACT

Transformer-based models are increasingly used in robotics and other edge computing applications, where local inference is required due to latency, connectivity, and power constraints. A key component of transformer architectures is the attention mechanism, which relies on the softmax function. The evaluation of the exponential function within softmax is computationally demanding in hardware and can become a bottleneck in efficient implementations of transformer inference.

This work presents a hardware-efficient implementation of the exponential function tailored for softmax computation in transformer models. The proposed approach targets the bfloat16 (BF16) numerical format, commonly used in AI workloads due to its wide dynamic range and reduced precision requirements.

To reduce implementation complexity, the input domain is restricted to non-positive values, consistent with the numerically stable formulation of the softmax function. The exponential operation is implemented using base conversion and a piecewise-linear (PWL) approximation of the function over the interval (−1,0], divided into 128 segments. The coefficients were optimized to minimize the unit-in-the-last-place (ULP) error, achieving a maximum error of 0.5 ULP in BF16 precision.

Figure 1 presents the proposed architecture. The BF16 input is decomposed into its components, special cases are handled, and the exponential value is computed using a fixed-point datapath with PWL approximation. The result is then normalized, rounded, and packed back into the BF16 format.

The proposed method enables an efficient hardware realization of the exponential function suitable for FPGA or ASIC-based transformer accelerators targeting edge computing systems.

This work was supported by the statutory subsidy of AGH University of Krakow (No. 501.00.230000.10000).

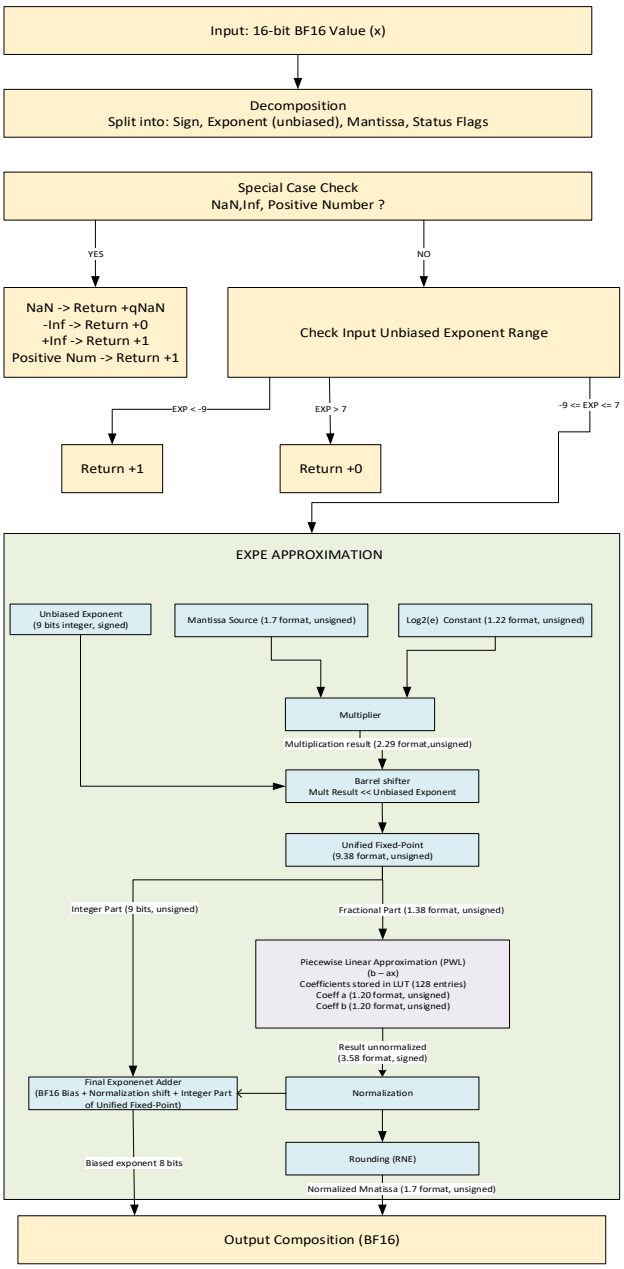

Fig. 1. Block diagram of the proposed BF16 exponential unit.