Stochastic Rounding: Algorithms and Hardware Accelerator

Mantas Mikaitis
University of Manchester, UK

Motivation

- **SpiNNaker2** Spiking Neural Network (SNN) simulation chip (available late 2021) will be based on ARM M4F.
- Many use cases for **rounding and saturating** fixed-point numbers of various lengths in SpiNNaker software.
- **Stochastic rounding (SR)** is increasingly popular in machine learning.
- Limited rounding support in ARM M4F—requires software libraries.
- To **accelerate some parts of arithmetic libraries** we designed a highly customizable rounding accelerator.

Background

SR is useful in various applications such as numerical verification, summation, ODE and PDE solvers. **On SpiNNaker, ODE solvers in fixed-point arithmetic benefit by replacing round-to-nearest (RN) with SR**. Various mixed-precision arithmetic operations are of use in SpiNNaker software. **Next-gen SpiNNaker is based on ARM M4F core** which contains RN support but not SR and custom rounding bit position and saturation. Furthermore, **a binary32 floating-point unit of the M4F can be used to compute on bfloat16 data and round back to reduce memory costs**.

MANCHESTER 1824
The University of Manchester

Rounding accelerator for a digital neuromorphic chip **SpiNNaker2**.

64- to 32/16-bit fixed-point arithmetic (with a preset round bit position) and float to bfloat16 **rounding and saturation**.

Round to neighbours **with probabilities** $\propto$ distances, or **to nearest**.

Accelerator features

- Round and saturate 64-, 32-, or 16-bit to 32- or 16-bit fixed-point numbers.
- SR and RN (ties up) modes.
- Rounding bit position programmable.
- Signed and unsigned format support.
- Rounding and saturation of float (binary32) values to bfloat16 values.
- Uses SpiNNaker2 hardware PRNG (algorithm commonly known as `KISS`).
- Up to four threads with different PRNG seeds are supported.
- Accelerator is general purpose: not limited to SpiNNaker2 hardware or SNN applications.
- $3-4$ cycle latency.

Evaluation

We evaluated three versions of the accelerator with 8-, 16-, and 32-bit SR. Lower bit width in SR means lower precision PRNG and adder needed. In SpiNNaker, SR of as low as 4 bits was shown to work well in ODE solvers. Using a 22 nm library accelerator was synthesized for $f_{clk} \in [50, 400]$ MHz clock frequencies. At $f_{clk} = 150$ MHz 8-bit SR accelerator can provide order of magnitude lower leakage power than 32-bit version, but no substantial savings at higher rates.

Summary

The presented accelerator can provide speedup in arithmetic libraries when transitioning from SpiNNaker to SpiNNaker2, and can be easily extended with other formats. SpiNNaker2 is a collaboration between TUD and Manchester.

TECHNISCHE UNIVERSITÄT DRESDEN
SpiNNaker2 Universal Spiking Neural Network Architecture