
Contribution ID: **20** Type: **Standard Talk**

# Optimizing for Imperfections in Analog Neural Computations on BrainScaleS-2

*Wednesday 27 September 2023 16:15 (15 minutes)*

Machine Learning has gone through major revolutionary phases over the past decade and neural networks have become state-of-the-art approaches in many applications, from computer vision to natural language processing. However, these advances come at ever-growing computational costs, in contrast, CMOS scaling is hitting fundamental limitations such as power consumption and quantum mechanical effects, thus lags substantially behind these growing demands. To approach this discrepancy, novel computing technologies have come into focus in recent years. In this setting, the field of analog computing is gaining research interest, as it represents a possible solution to the scaling problem.

While analog computing can dramatically increase the energy efficiency of computations and thereby contribute to a continued performance scaling of CMOS, it comes with the emblematic caveats of analog computations like noise, temporal drift, non-linearities, and saturation effects. Even though calibration routines and intelligent circuit design try to compensate for these imperfections, in practice they cannot be fully avoided. Therefore, applications running on these analog accelerators must find solutions to limit their impact.

One implementation of an analog accelerator is the BrainScaleS-2 system from the Kirchhoff Institute at Heidelberg University. It's a neuromorphic mixed-signal system with an analog chip at its core, that serves as a research platform for spiking neural networks and as an analog matrix multiplication accelerator at the same time. The primary design goals of the system are energy-efficient computing and a scalable chip design, which allows the system to be extended up to wafer-scale processor sizes.

This work is concerned with such analog computations and the implications of calibration in terms of result quality and runtime cost. We conduct several experiments with artificial neural network models on the BrainScaleS-2 system as an accelerator for matrix multiplications and familiarize ourselves with these difficulties. A central approach to overcoming analog imperfections is hardware-in-the-loop training, in which the model is trained on the actual inference hardware. This compensates for remaining calibration offsets and allows the model to tolerate a certain level of noise. Further, we improve the performance of these models by adjusting calibration parameters as well as the mapping strategy of the linear layers to the analog hardware.

Our major contributions to the FastML workshop are a short introduction to the circuit which executes the analog matrix multiplication, related tooling, and calibration parameters. We then show how these parameters affect the multiply-accumulate operation, with the respective impact on hardware imperfections and how they impact the training results of the model itself. We show, that optimizing these parameters involves tradeoffs with respect to remaining imperfections that cannot be further improved, and plan on discussing these to foster an overall understanding of the challenges of analog hardware in the community.

**Authors:** Mr KLEIN, Bernhard (Heidelberg University); KERN, Eric (Heidelberg University); BORRAS, Hendrik (Heidelberg University); FRÖNING, Holger

**Presenters:** KERN, Eric (Heidelberg University); BORRAS, Hendrik (Heidelberg University)

**Session Classification:** Contributed Talks

**Track Classification:** Contributed Talks