Department of Physics and Astronomy
University of Heidelberg

Bachelor Thesis in Physics
submitted by

Timo Wunderlich

born in Freising (Germany)

August 2016

Synapse Calibration on the HICANN-DLS Neuromorphic Chip

This Bachelor Thesis has been carried out by Timo Wunderlich at the
Kirchhoff Institute for Physics
Heidelberg University
under the supervision of
Prof. Karlheinz Meier

Abstract

HICANN-DLS is a novel neuromorphic Application-Specific Integrated Circuit (ASIC) developed in Heidelberg within the scope of the Human Brain Project. It consists of 32 neurons and a total of 1024 synapses that mimic their biological counterpart by means of full-custom analog electronic circuits. The synapses contain individual circuitry designed to implement Spike-Timing-Dependent Plasticity (STDP), a synaptic mechanism that is believed to play a vital role in learning and memory. Because of inherent transistor mismatch, the parameters describing STDP (time constants and amplitudes) vary across synapses. In order to counteract this mismatch, each synapse contains four calibration bits which are intended to compensate parameter variation. Within this thesis, the synaptic circuits tailored for STDP are described and their functionality quantified. In particular, it has been tested to which degree transistor mismatch can be counteracted. The combinatorial problem of assigning calibration settings to synapses in order to reduce the Mean Absolute Deviation (MAD) of the parameter values is described using Mixed Integer Linear Programming (MILP). Using this approach on the tested chip, it is possible to reduce the spread of parameters by a factor of 1.4 to 2.1.

Zusammenfassung

HICANN-DLS ist ein neuartiger neuromorpher Application-Specific Integrated Circuit (ASIC), der im Rahmen des Human Brain Projects in Heidelberg entwickelt wird. Er besteht aus 32 Neuronen und insgesamt 1024 Synapsen, die ihr biologisches Gegenstück in analogen Full-custom Schaltkreisen emulieren. Jede Synapse enthält elektrische Komponenten die Spike-Timing-Dependent Plasticity (STDP) implementieren, ein synaptischer Mechanismus, dem eine bedeutende Rolle in Lernprozessen und Erinnerungsvermögen zugesprochen wird. Aufgrund von unausweichlichen Fertigungsunterschieden der Transistoren variieren die STDP beschreibenden Parameter von Synapse zu Synapse. Um diesem Effekt entgegenzuwirken, beinhaltet jede Synapse vier Kalibrationsbits, die Unterschiede ausgleichen sollen. Im Rahmen dieser Arbeit werden die STDP-relevanten Schaltkreise beschrieben und ihre Funktionalität bestätigt. Es wird insbesondere untersucht, inwiefern die Varianz der Transistoren ausgeglichen werden kann. Das kombinatorische Problem, den einzelnen Synapsen Kalibrationseinstellungen zuzuweisen, die die Mean Absolute Deviation (MAD) minimieren, wird mittels Mixed Integer Linear Programming (MILP) behandelt. Dieser Ansatz erlaubt es, den Versatz auf dem getesteten Chip um einen Faktor 1.4 bis 2.1 zu reduzieren.

Contents

1 Introduction

Computers have developed at a rapid pace and are arguably the most important technological feat in recent human history, spurring scientific progress by providing means for data evaluation and numeric calculations. Since their conception, they vastly outperform humans in tasks such as arithmetic and with the most refined algorithms available today, slightly outperform humans in domains such as traffic sign recognition (Stallkamp *et al.*, 2012) or multi-talker speech recognition (Hershey *et al.*, 2010). However, these results were obtained using traditional computer hardware consuming power in the order of kilowatts while a human brain operates with around 20 Watts and is not performing only a single task, thereby being a much more efficient computational device.

Traditional computers are designed to perform calculations in a largely sequential manner with the code not changing in time once compiled, while mammalian brains operate massively in parallel and have the ability to learn, thus change their internal structure and interconnection. This makes it possible to adapt to changing circumstances and new kinds of sensory input. The brain is also robust to probabilistic behaviour of the biological transmission mechanisms between its constituents (Otmakhov *et al.*, 1993) and may recover from damage caused by a stroke or even outright loss of brain matter by structural re-organization (Feuillet *et al.*, 2007; Weiller *et al.*, 1992). This kind of robustness and ability to adapt is not present in traditional computer hardware.

Artificial neural networks are inspired by biological neural networks and are used to model the functional and learning behaviour of the brain. They are used in neuroscientific experiments with the goal of gaining a better understanding of brain function as well as in solving tasks such as the recognition tasks mentioned before. Such networks can be simulated on traditional computer hardware albeit very inefficiently when compared to their biological counterpart. Some neural networks cannot be simulated faster by increasing parallelism in general-purpose com-

puters which means that their run time exhibits a hard boundary (Zenke & Gerstner, 2014). It is therefore of interest to design new, non-Von-Neumann hardware architectures that mimic brain function, are able to provide efficiency comparable to biological neural networks and even emulate them faster than biological real-time by orders of magnitude.

This is the goal of the neuromorphic computing platform within the Human Brain Project which succeeds the BrainScaleS project. This project has led to the development of the *High Input Count Analog Neural Network* (HICANN) neuromorphic chip that emulates neural networks with a speed-up factor of about 10^4 compared to biology (Millner *et al.*, 2012). This chip represents a physical model of neural networks that is realized using Complementary Metal-Oxide-Semiconductor (CMOS) integrated circuit technology. The functional units of the brain, neurons and synapses, are modeled using electronic circuits that emulate their properties as dictated by a certain model.

The latest prototype design of this chip, code-named HICANN Digital Learning System (HICANN-DLS), includes functionality to emulate models of learning and synaptic plasticity and specifically is able to emulate the biological learning process called *Spike-timing-dependent synaptic plasticity* (STDP). This process takes place in each individual synapse independently and can be characterized by parameters which in the case of HICANN-DLS are determined by electronic circuits made up of transistors. Inevitably, these transistors do not behave identically due to the nature of the chip manufacturing process and the parameters of STDP will therefore vary across different synapses. For this reason, the prototype design includes a calibration mechanism in each synapse that is supposed to make up for these differences and provide a set of parameters across all synapses that is as homogeneous as possible. It is the goal of this thesis to investigate this and provide software to calibrate a chip for given target parameters.

2 Methods

This chapter will illustrate the methods used for investigating the synaptic correlation circuits on the HICANN-DLS. First, the learning mechanism for which the circuits are tailored, Spike-Timing-Dependent Plasticity, is introduced.

2.1 Spike-Timing-Dependent Plasticity

Spike-Timing-Dependent synaptic Plasticity (STDP) is a biological process that is widely believed to play a vital role in learning and memory. Historically, it was developed on the basis of *Hebb's postulate*. In 1949, Hebb published his book *Organization of Behavior* in which he formulated the greatly influential Hebb's postulate that describes a proposed mechanism of learning in mammalian brains. The postulate states that the interconnection between two neurons is strengthened if one of them is repeatedly responsible for firing the other:

> *"When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."*
> Hebb (1949)

This has been famously but imprecisely sloganized as *"Cells that fire together, wire together"* (Shatz, 1992). The imprecision stems mostly from the fact that the postulate speaks of a directional relationship (cell A excites cell B). Variants of learning rules based on this postulate have been implicated in the emergence of mirror neurons (Keysers & Perrett, 2004) and classical conditioning (Bi & Poo, 2001) which underscores the fundamental importance of Hebbian learning for social and environmental functioning in mammals.

The postulate in its original form does not include depression of the synaptic efficacy and will cause an increase without bounds - this problem is mitigated by introducing a synaptic weakening mechanism through competition between synapses. The competition can naturally be introduced by a global mechanism that incorporates the state of many synapses, e.g. by establishing a constant value for the sum of synaptic values (Rochester *et al.*, 1956), yet it is bio-physically more realistic to introduce a competition mechanism that is local to each synapse and does not know about the state of any other synapses.

In biological experiments, it was observed that the relative timing of excitatory pre- and post-synaptic spikes determined the modification in efficacy of the connecting synapse, where a smaller temporal delay between the spikes yielded a larger change and the sign of the change (potentiation or depression) depended on the order of the spikes (Bi & Poo, 1998; Markram *et al.*, 1997). This is consistent with Hebb's postulate as the time difference between pre- and post-synaptic spike can be seen as a measure for the correlation of the pre-synaptic neuron firing the post-synaptic neuron. The change in synaptic efficacy mediated by the temporal correlation of pre- and post-synaptic spikes can last for months (Douglas & Goddard, 1975) and this effect has therefore been dubbed long-term potentiation (LTP) or long-term depression (LTD) for an increase or decrease in efficacy, respectively (Ito, 1989; Teyler & DiScenna, 1987).

This form of synaptic plasticity, where the temporal correlation of spikes is key, is called spike-timing-dependent plasticity (Song *et al.*, 2000). It provides a plasticity mechanism that leads to competing synapses while being local for each of them. Competition arises because synapses compete to control the timing of post-synaptic spikes and the efficacy of synapses that do not do so is reduced. The amount of synaptic modification is determined by a function $F(\Delta t)$ of the time difference between pre- and post-synaptic spike Δt:

$$F(\Delta t) = \begin{cases} A_+ \exp\left(-\frac{\Delta t}{\tau_+}\right) & \Delta t > 0 \\ -A_- \exp\left(\frac{\Delta t}{\tau_-}\right) & \Delta t < 0 \end{cases} \tag{2.1}$$

where τ_+ and τ_- control the time window over which the synaptic modification occurs and A_+ and A_- control the maximum amount of modification. For $\Delta t > 0$, the pre-synaptic spike occurred before the post-synaptic spike and vice-versa. A_+ and τ_+ are therefore referred to as the *causal* amplitude and time constant, A_- and τ_- are the *anti-causal* (shorthand: acausal) parameters. The integral over Δt has to be negative if an uncorrelated spike train is to produce a weakened synapse (Song *et al.*, 2000).

This model makes the assumption that the synaptic modification caused by a spike pair does not depend on previous spike pairs, i.e. that the modification sums linearly and each pair makes an independent contribution. It has been shown that this is not the case in biology, where other

factors like firing rate and cooperativity play a role (Clopath *et al.*, 2010; Froemke & Dan, 2002; Sjöström *et al.*, 2001), but this simplification is used in the following. However, the presented neuromorphic chip possesses circuitry to measure the firing rate and may use it in custom plasticity models (Hartel, 2016).

As examples of applications, STDP has been used to extract car trajectories from a spiking silicon retina (Bichler *et al.*, 2012) as well as to perform character recognition using the Mixed National Institute of Standards and Technology (MNIST) benchmark (Diehl & Cook, 2015), both with a high degree of accuracy and robustness.

2.2 HICANN-DLS

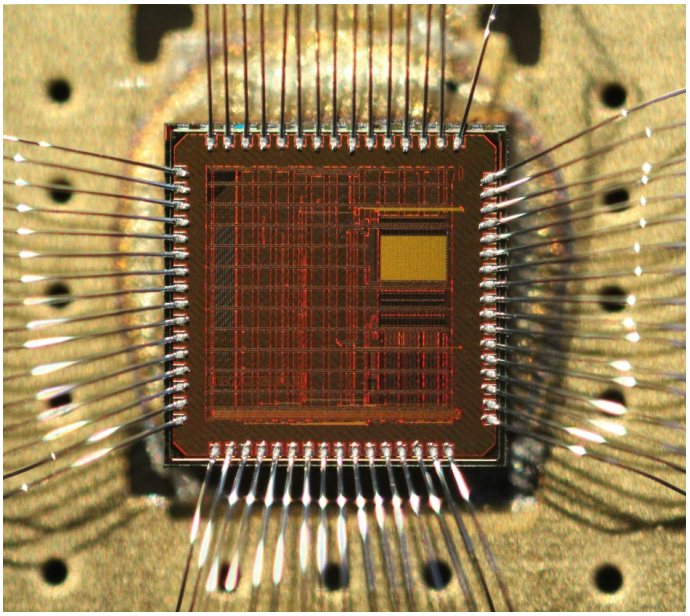

Figure 2.1: The HICANN-DLS neuromorphic chip. Image provided by Hartel (2016).

The HICANN-DLS (Fig. 2.1) is the successor to the HICANN neuromorphic chip and was available in the form of several individual prototype chips to be used with evaluation boards during the authorship of this thesis. All measurements in this thesis were conducted exclusively with chip No. 23. During that time, the HICANN chip was in operation on a platform made up of several wafer modules, each including 348 HICANN chips for a total of around 200k neurons and 60M synapses enabling the emulation of large neural networks (Hock, 2014).

The main difference of the HICANN-DLS compared to HICANN is that it is manufactured using a 65 nm process instead of a 180 nm process, allowing for a considerably higher integration density and the incorporation of a digital extension dubbed the Plasticity Processing

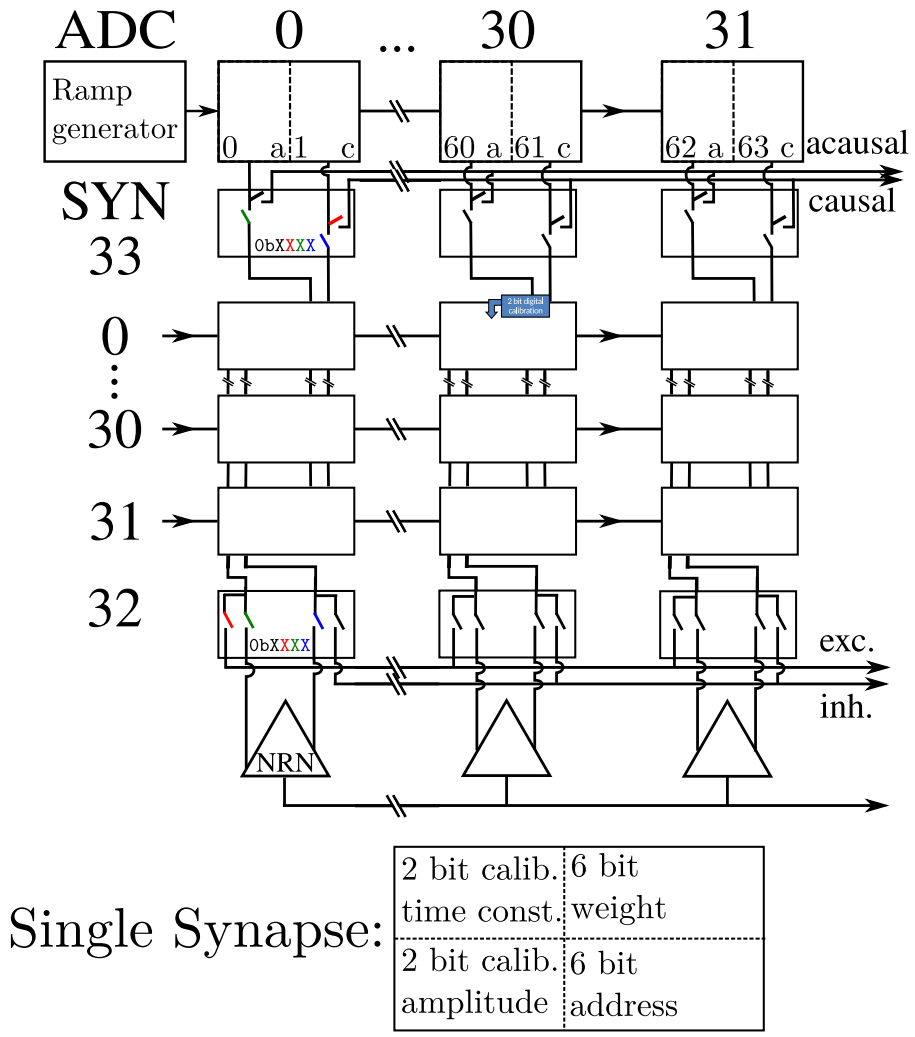

Figure 2.2: Schematic overview of parts of the HICANN-DLS that are relevant for this thesis. A detailed description is given in 2.2.1. The content of each synapse is visualized in Fig. 2.3.

Unit (PPU), which allows for implementation of arbitrary plasticity rules and specifically STDP by means of a general-purpose processor right on the chip (Friedmann *et al.*, 2016). A single HICANN-DLS chip contains 1024 synapses and 32 neurons that emulate their biological counterpart in fully analog, continuous-time circuits. Communication with a host computer takes place via a Field-Programmable Gate Array (FPGA) within a test controller. In the following, the parts of the HICANN-DLS relevant for this thesis (the synaptic circuits, correlation mechanisms and the PPU) are described.

2.2.1 Synapses and Neurons

A single HICANN-DLS chip contains an array of 32 by 32 synapses, where each of the 32 synaptic columns is connected to one neuron at the bottom (Fig. 2.2). The used neuron model is

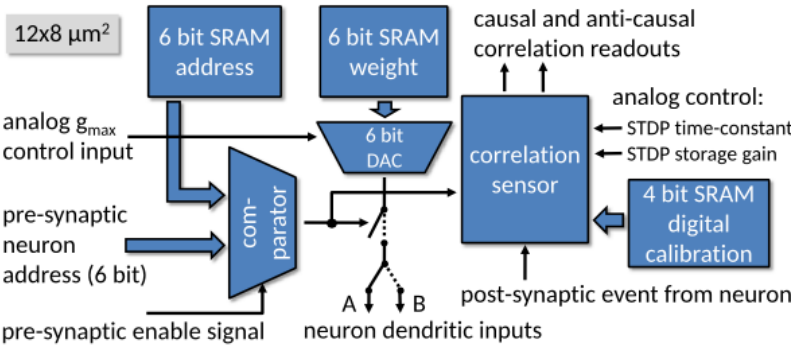

Figure 2.3: Block diagram of synapse circuit, taken from Friedmann *et al.* (2016).

Leaky Integrate & Fire (LIF). A block diagram of each synaptic circuit is shown in Fig. 2.3. Pre-synaptic spikes are sent into the array row-wise and addressed to specific synapses by using the 6-bit address stored in each synapse using Static Random-Access Memory (SRAM). In addition to this address, each synapse has SRAM to store a 6-bit weight and 4 calibrations bits. When a synapse receives a spike addressed to itself, it generates a current pulse with an amplitude proportional to the stored weight via a 6-bit Digital-to-Analog Converter (DAC) that travels through the synaptic column toward the neuronal input via either the excitatory or inhibitory line. There exist four output lines for debugging that can be switched by configuration bits within two special synaptic rows (number 32 and 33, see Fig. 2.2) and lead to the outside of the chip. When the neuron fires a post-synaptic pulse, this pulse is fed back to the synapse so that it may record the correlation of pre- and post-synaptic firing. The correlation sensor used in this process is described in the following.

2.2.2 Synaptic Correlation Sensor

Every synapse circuit also contains a correlation sensor that saves timing information of received pairs of pre- and post-synaptic spikes. The circuit diagram for the correlation sensor is given in Fig. 2.4. Describing the operation of the circuit in detail is out of the scope of this thesis. An elaborate description can be found in Friedmann *et al.* (2016). It shall suffice to elucidate the general operating principle.

The arrival of a post-synaptic spike instantly discharges C_{acausal} and then triggers charging it with a constant current generated by M4 and controlled by V_{ramp}. The equivalent is done for a pre-synaptic spike and C_{causal}. The charges on these capacitors are then a linear measure for the time that has passed since the latest arrival of a pre or post pulse.

A post-synaptic pulse will also trigger transferring the time difference on C_{causal} to one of the two C_{storage} capacitors designated for causal storage. Before storing it on the storage capacitor, the

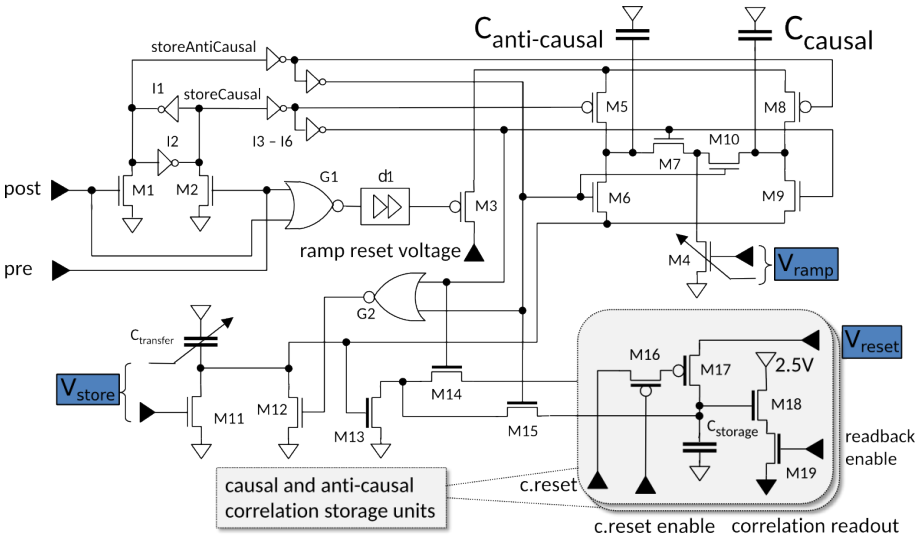

Figure 2.4: Schematic of the correlation sensor, adapted from Friedmann *et al.* (2016). The voltages mentioned in the text are marked using blue boxes.

time difference (equivalent to the charge on C_{causal}) is weighted exponentially. The amplitude of the exponential weighting is controlled by transistor M11 and V_{store}. This means that the charge on the storage capacitor is an exponential measure for the time difference of the latest pre-before-post spike pair. Storing the time difference is done by discharging the respective C_{storage} capacitor from the reset voltage V_{reset}.

The voltage on the storage capacitors is then ideally given as follows (a_+ for causal, a_- for anti-causal)

$$a_+ = V_{\text{reset}} - \sum_{\text{pre-post pairs}} \eta_+ \exp\left(-\frac{\delta_{\text{pair}}}{\tau_+}\right) \tag{2.2}$$

$$a_- = V_{\text{reset}} - \sum_{\text{post-pre pairs}} \eta_- \exp\left(\frac{\delta_{\text{pair}}}{\tau_-}\right) \tag{2.3}$$

where V_{reset} is the reset voltage, δ_{pair} is the time difference for each pair, $\eta_\pm$ are the analog accumulation amplitudes per pair (controlled by V_{store}) and $\tau_\pm$ are time constants (controlled by V_{ramp}). V_{reset} is a global parameter that is the same for all synapses while $\eta_\pm$ and $\tau_\pm$ are subject to variations due to transistor mismatch. The time measurement circuit is shared for pre-before-post pairs and post-before-pre pairs and it is therefore expected that $\tau_\pm$ and $\eta_\pm$ exhibit a high degree of equality.

By using an enable signal, the causal or anti-causal storage voltage can be connected to the corresponding line in that column which leads to a source follower biased using the voltage $V_{\text{coroutbias}}$

that is connected to the single-slope Analog-to-Digital Converter (ADC, two per column for causal and anti-causal readout). The enable signal allows for reading out the correlation of each synapse in that column individually. The ADC transforms the applied voltage to a 8-bit output code which can be read out using the PPU or directly using external control logic. It works by using the voltage ramp generated commonly for all ADCs: a digital counter (255 to 0) is started upon commencement of the ramp and stopped as soon as the voltage on the ADC matches the ramp voltage. The output code is then the code at which the counter was stopped. This signaling chain is constructed in a way so that proper configuration leads to an output code offset that corresponds to zero correlation, while an output code of 255 implies maximum measurable correlation. The offset will vary across synapses.

The correlation data obtained by the PPU in this way can be used to execute arbitrary plasticity and learning rules and is specifically suited for STDP, as a comparison of equations 2.2, 2.3 and 2.1 instantly suggests. The PPU is able to change synaptic weights and reset the correlation voltage. It will generally determine the offset correlation for all synapses at the beginning of an experiment and correct for this later.

2.2.3 Available Parameters

All parameters available to influence the plasticity mechanism on HICANN-DLS are summarized in Tab. 2.1 and further described in the following.

The derived parameters $\eta_\pm$ and $\tau_\pm$ of a correlation sensor are in first order determined by the global parameters V_{store} and V_{ramp}, respectively (there are plans to make these row-wise parameters in future prototypes (Hartel, 2016)).

Additional control local and specific to each synapse can be exerted using the 4 calibration bits. Here two bits control $\eta_\pm$ and the other two $\tau_\pm$ by varying the length of the current sink M4 and the capacity of C_{transfer}, respectively. These have been implemented because transistor mismatch causes substantial varia-

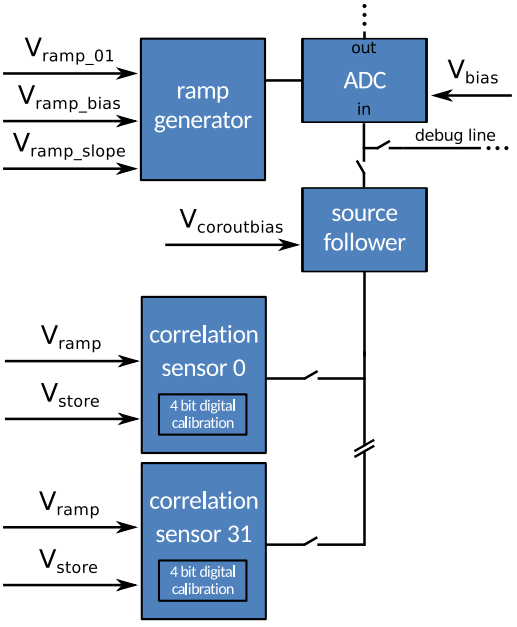

Figure 2.5: Correlation signal chain of one synapse column.

Parameter	Affected quantity	Type
V_{ramp}	Time constants $\tau_{\pm}$	global (planned: row-wise)
V_{store}	Amplitudes $\eta_{\pm}$	global (planned: row-wise)
V_{reset}	Correlation reset voltage, thereby correlation offsets	global
$V_{\text{coroutbias}}$	Source follower bias voltage	global
$V_{\text{ramp\_slope}}$	Slope of ADC ramp	global
$V_{\text{ramp\_bias}}$	Ramp bias	global
$V_{\text{ramp\_01}}$	Offset of ADC ramp	global
V_{bias}	ADC comparator bias	global
First two calibration bits	Time constants $\tau_{\pm}$	local
Last two calibration bits	Amplitudes $\eta_{\pm}$	local

Table 2.1: Parameters available to influence plasticity. Global: set for all synapses. Local: set per synapse.

tion between synapses as mentioned before.
The next chapter will describe how these calibration bits can be used to provide a set of parameters that is as homogeneous as possible.

V_{reset} is the voltage to which the storage capacitors are pulled after a reset. It should be chosen such that transistor mismatch between synapses or ADC trace variations do not in any synapse cause a baseline output code of zero, i.e. a voltage beyond the range of the ADC as this would entail loss of correlation information.

$V_{\text{coroutbias}}$ is the bias of the source followers right in front of the ADCs and should be chosen in a way that provides the most linear source follower while using the whole range of the ADC.

$V_{\text{ramp\_slope}}$, $V_{\text{ramp\_bias}}$ and $V_{\text{ramp\_01}}$ are controlling the characteristics of the ramp generator used with the ADCs such as slope and offset of the ramp. These parameters therefore control the measurement range of the ADCs. V_{bias} is the comparator bias voltage within each ADC and impacts the linearity of the ADC characteristic curves.

2.2.4 Plasticity Processing Unit

The PPU is a general-purpose processor with some specialization that receives correlation information from the synaptic circuits processing spikes. A thorough introduction is found in Friedmann *et al.* (2016). It constitutes a digital extension to the analog circuits that emulate

synapses and neurons and is able to execute arbitrary learning rules by being programmed using the C programming language. The processor is clocked at $96\,\mathrm{MHz}$ which allows for weight update rates consistent with a speed-up factor of 10^3 compared to biology. The implemented instruction set is Power ISA 2.06 with added instructions for using the vector unit within the PPU for row-wise parallel processing of synapses (16 at a time, if no synapses were combined to provide a 12-bit weight). The PPU reads the 8-bit output codes of the correlation ADCs which correspond to the causal or anti-causal correlation values of one synapse row and can use these to calculate weight updates for the synapses, for example using an STDP rule.

2.3 Synaptic Calibration

The time constants $\tau_\pm$ and amplitudes $\eta_\pm$ of the correlation sensor in each synaptic circuit are subject to transistor mismatch and will therefore vary across synapses. It is the first task to empirically determine the values and variation of both parameters for which a calibration can then take place. In order to be able to determine $\tau_\pm$ and $\eta_\pm$ for a synapse, it is a prerequisite to ensure that the correlation signal chain operates as desired. This entails ensuring proper operation of the ADC ramp, measuring the characteristic curves of the ADCs and the correlation source follower. A measurement protocol to determine $\tau_\pm$ and $\eta_\pm$ is then devised.

The calibration bit settings of the time constant and amplitude (two bits each) are given as a four bit number, where the two most significant bits are the amplitude bits and the two least significant bits are the time constant bits. The ADC output code (0 to 255) is reported using arbitrary units, abbreviated $\mathrm{a.u.}$.

2.3.1 ADC Ramp

The ADC ramp may be configured through $V_{\text{ramp\_slope}}$, $V_{\text{ramp\_bias}}$ and $V_{\text{ramp\_01}}$. An oscilloscope (LeCroy WaveRunner HRO 64Zi) was used to investigate the ramp generated using these parameters by connecting it to a debug pin on the chip evaluation board. The maximum voltage of the ramp is limited to $1.2\,\mathrm{V}$ because of the kind of transistors used in the generating circuit. The ideal waveform for the ramp is therefore a saw-tooth signal with an amplitude of $1.2\,\mathrm{V}$, thus maximizing the dynamic range of the ADCs.

2.3.2 ADC Characteristic Curves

The ADCs digitize the voltage coming from the correlation sensor of each synapse but can also convert an external voltage by using the debug lines visible in Fig. 2.2 and Fig. 2.5. This allows for measuring the characteristic curves of the ADCs and this information can be used in measurements to discern if effects are caused by the ADC or the correlation circuit. The characteristic curves can be measured by setting the proper switches, providing the external voltage and then reading out the ADCs using the evaluation board.

2.3.3 Correlation Source Follower

The voltage on the causal or anti-causal correlation storage capacitor on each synapse is not directly connected to the ADC but via a column-wise source follower in front of the ADC and only when a "readback enable" signal is activated (see Fig. 2.5). The source follower is biased using $V_{\text{coroutbias}}$. This voltage has an impact on the linearity and offset of each source follower.

The characteristic curves of the 64 source followers can be be taken by setting V_{reset}, reading out the ADC and correcting the result using the ADC characteristic curves. This allows to translate any given V_{reset} (equivalent to a non-reset correlation voltage that is the result of arriving spike pairs) into an ADC input voltage.

2.3.4 Measurement Protocol

The time constants $\tau_{\pm}$ and amplitudes $\eta_{\pm}$ of a correlation sensor can be empirically determined by recording the causal or anti-causal correlation when spike pairs with defined spike intervals Δt are sent into the synaptic circuit. $\Delta t > 0$ implies that the pre-spike arrives Δt units of time before the post-spike (thereby causing causal correlation) and vice-versa. The correlation has to be reset before probing. The data of both branches can be fitted with an exponential decay as given in 2.2 and 2.3.

The correlation voltage can be read out either directly from the ADCs using the FPGA on the evaluation board or indirectly by using the PPU. In the latter case, the PPU is programmed so that it reads out the ADCs and saves the values to memory that can be accessed using the FPGA. Both methods yield equal results.

As described in 2.2, pre-synaptic spikes can be sent to an entire row of synapses at a time. Post-synaptic spikes can also be triggered simultaneously for all 32 neurons, so that all columns of

synapses receive post-synaptic spikes. All 64 ADCs operate in parallel and the PPU is designed to process 16 ADC values in parallel. It is therefore the natural approach to measure the $A_{\pm}(\Delta t)$ curve (causal or anti-causal correlation as a function of Δt) for an entire row of synapses at a time by subjecting all synapses of the row to spike pairs simultaneously and then reading out the correlation.

At small amplitudes $\eta_{\pm}$, sending single spike pairs into the circuit will not produce data suitable for a fit and it is warranted to send multiple pairs. After sending each pair, an idle time is required so that the pairs are temporally spaced far from each other. This idle time was set to 10 ms. The resulting fit parameter for the amplitude must then be normalized using the number of sent pairs. As the time constant can vary from approximately 10 µs to 300 µs (Friedmann *et al.*, 2016), it is also advisable to adapt the measurement range, that is the range of used values for Δt.

The measurement protocol can be summarized as follows:

1. Configure the parameters given in table 2.1.
2. Reset the correlation.
3. Repeat for a range of Δts:
   (a) Send spike pairs (row-wise).
   (b) Read out correlation (row-wise).
4. Fit exponential decay to data.

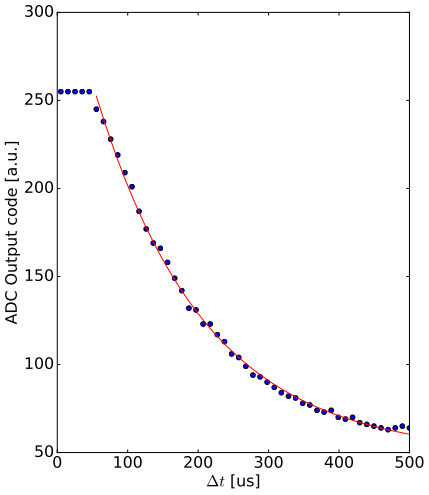

Figure 2.6: Example of the fit extrapolating beyond saturation.

If the correlation saturates for small absolute values of Δt, the saturated part is disregarded for the fit. Any fit parameter for the amplitude that is larger than 255 (maximum correlation output code of the ADC) is the result of extrapolation beyond the saturation.

For some outliers, the fits might fail, for example when the Δt range used in the measurement is much smaller than the outlying time constant. The fit is defined to have failed when the error of one fit parameter is above 20%. The value is then disregarded.

3 Results

3.1 ADC Ramp

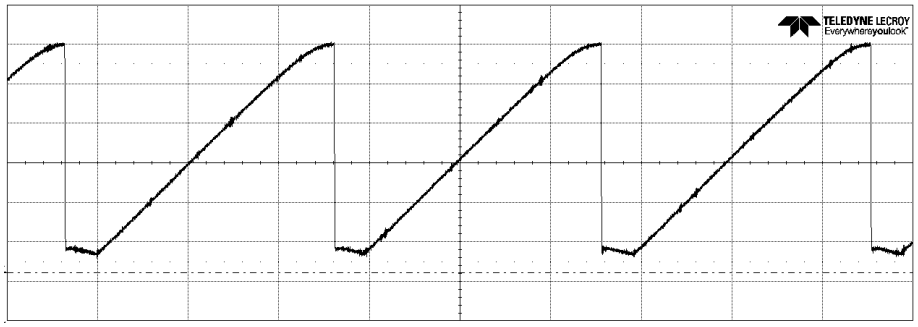

Figure 3.1: Adjusted ADC ramp taken with oscilloscope for $V_{\text{ramp\_slope}} = 0.33\,\text{V}$, $V_{\text{ramp\_bias}} = 0.88\,\text{V}$, $V_{\text{ramp\_01}} = 0.4\,\text{V}$. The scaling of the vertical axis is $200\,\text{mV}$ per division, the scaling of the horizontal axis is $1\,\mu\text{s}$ per division.

A ramp that closely resembles the ideal saw-tooth waveform is visible in Fig. 3.1. The slight delay between adjacent ramps, causing an offset of $(114 \pm 5)\,\text{mV}$, is a measurement artifact caused by the ramp output buffer used for the measurement. The characteristic curves of the ADC presented in the following were created using the pictured ramp and show that the ADC is able to digitize voltages below $114\,\text{mV}$. This ramp was used in all following measurements.

3.2 ADC Characteristic Curves

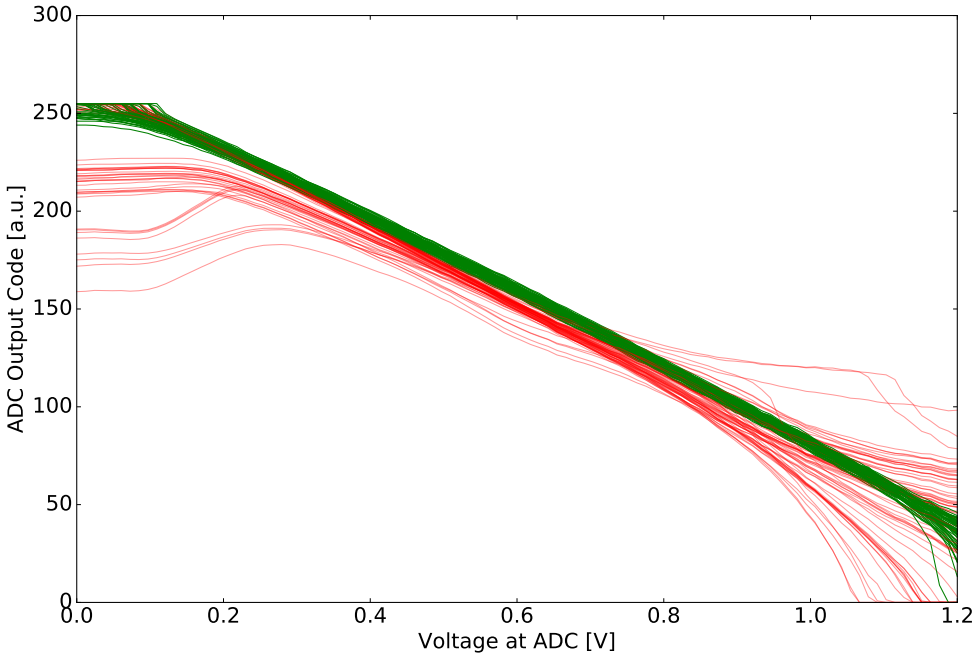

Figure 3.2: Characteristic curves for all 64 ADCs when using $V_{\text{ramp\_slope}} = 0.33\,\text{V}$, $V_{\text{ramp\_bias}} = 0.88\,\text{V}$, $V_{\text{ramp\_01}} = 0.4\,\text{V}$. For the green curves, $V_{\text{bias}} = 0.7\,\text{V}$ and for the red curves $V_{\text{bias}} = 0.3\,\text{V}$. External voltage input was provided to the chip using a debug line.

The characteristic curves of all 64 ADCs were taken and are presented in Fig. 3.2. The chosen value for V_{bias} is the result of taking the curves when varying V_{bias} and using the ramp parameters found previously. This parameter was optimized for curve linearity and homogeneity by manual optical judgment, so as to provide linearity in the greatest possible range. The red curves demonstrate that bad values of V_{bias} lead to erratic curves.

3.3 Correlation Source Follower

The characteristic curves of the 64 source followers are presented in Fig. 3.3. These were taken by varying V_{reset} of the first synapse row, measuring the ADC output code and correcting this data using the ADC characteristic curves. In this way, it was possible to determine good values for $V_{\text{coroutbias}}$ and V_{reset}.While different combinations of $V_{\text{coroutbias}}$ and V_{reset} can yield the same behaviour, $V_{\text{coroutbias}} = 0.43V$ and $V_{\text{reset}} = 2.5V$ were chosen for the following measurements.

Fig. 3.3 shows that these values provide a baseline correlation larger than zero for all synapses and the source follower characteristic curves are largely linear in the relevant interval ranging from an ADC voltage of $0.1\mathrm{V}$ to $1.2\mathrm{V}$. It was experimentally confirmed that these curves are identical with respect to the synapse row that is being used to provide V_{reset} to the ADCs.

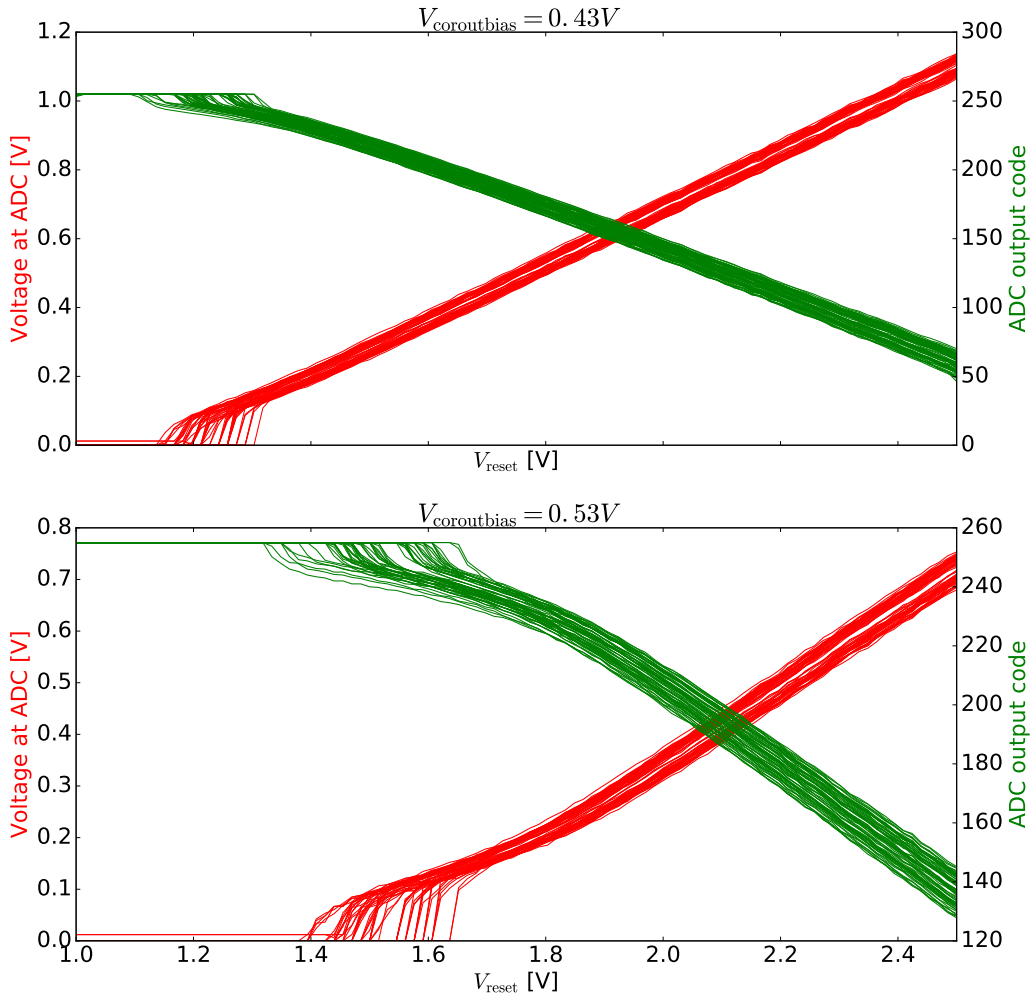

Figure 3.3: ADC output code and ADC input voltage (obtained by correcting the data using the curves from 2.3.2) for different correlation voltages set using V_{reset} when using two different voltages for $V_{\text{coroutbias}}$. The value for $V_{\text{coroutbias}}$ from the upper plot together with $V_{\text{reset}} = 2.5V$ provides a non-zero correlation offset for all ADCs and a relatively linear curve. The lower plot illustrates the consequences of a bad choice for $V_{\text{coroutbias}}$.

3.4 Asymmetric Amplitudes

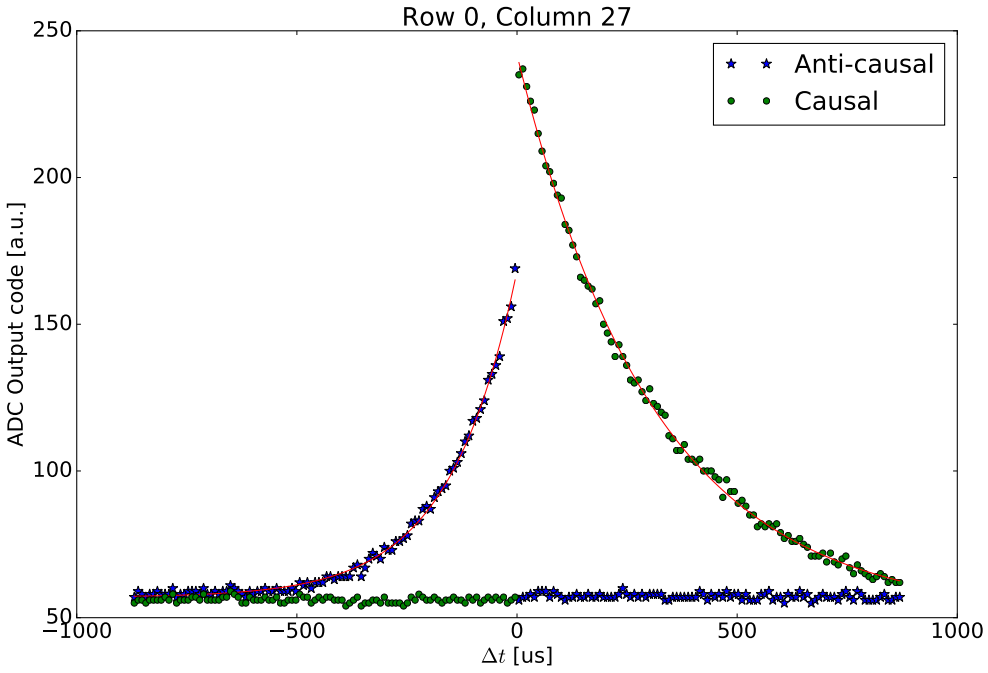

Figure 3.4: Exemplary measurement of the causal and anti-causal channel of a specific synapse on the chip using $V_{\text{ramp}} = 0.18\text{V}$ and $V_{\text{store}} = 0.32\text{V}$ after sending one spike pair. The amplitude of the anti-causal channel is clearly smaller.

Using the measurement protocol described earlier, curves such as the exemplary one presented in 3.4 were recorded and fitted. Each set of parameters (V_{ramp}, V_{store}, calibration bits) defines 2048 of such curves (1024 synapses, causal and anti-causal branch).

During measurements, it was observed that the anti-causal amplitude was systematically smaller than the causal amplitude. This was confirmed by comparing the average over all synapses for both branches. It was further observed that the magnitude of the asymmetry depended on the number of simultaneously fired post-synaptic pulses: the anti-causal amplitude average increases for a smaller number of fired neurons and in that sense approaches the causal amplitude. This was done by measuring the correlation curves for all synapses individually, where in each measurement pre-synaptic spikes were sent to all synapses of the corresponding row and a certain number of post-synaptic spikes from random neurons was added to the post-synaptic spike from the neuron corresponding to the individual synapse. The results are shown in Fig. 3.5. Output codes larger than 255 are the result of the fit extrapolating beyond the saturation.

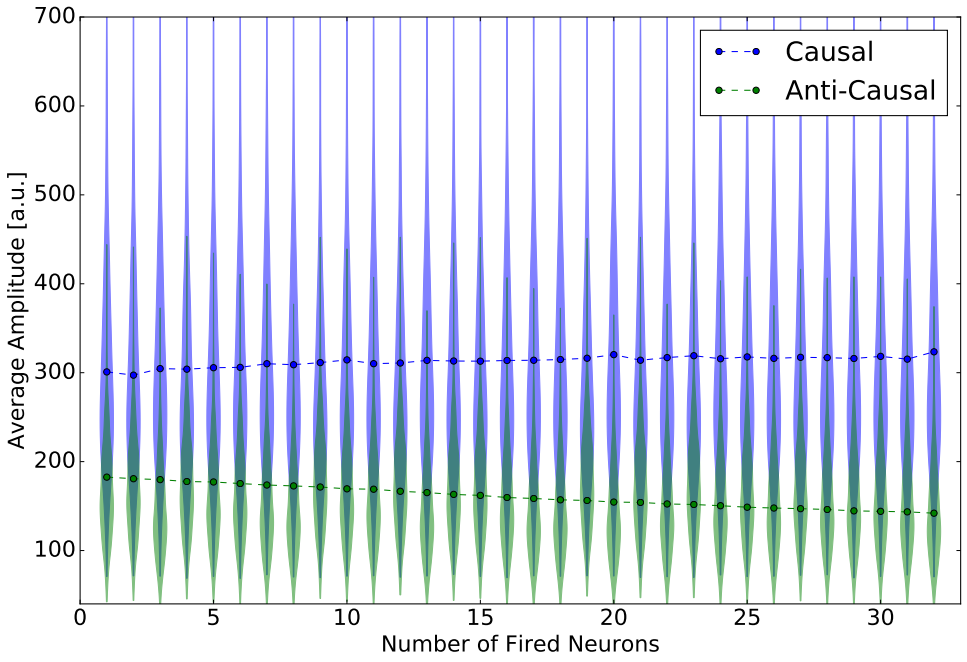

Figure 3.5: The amplitude of the causal and anti-causal correlation curves, when averaged over all 1024 synapses, exhibits a clear asymmetry: the anti-causal amplitude is systematically smaller. The asymmetry decreases for a smaller number of neurons that are simultaneously fired per measurement. The distribution of values is visualized as a violin plot. The used parameters were $V_{\text{store}} = 0.3\text{V}, V_{\text{ramp}} = 0.24\text{V}$.

It was found that this asymmetry is present regardless of the values chosen for V_{store} and V_{ramp} and independent of pre-synaptic spike addressing. The asymmetry could not be mitigated during the authorship of this thesis and because firing only one neuron per measurement (implying lowest asymmetry) increases the measurement time for all synapses 32-fold, all following measurements were taken firing all 32 neurons at once. The time constants $\tau_{\pm}$ exhibited no systematic asymmetry and did not change when firing different numbers of neurons.

Although the specific cause for the asymmetry could not be determined, it is expected that this issue will be resolved in the next prototype. The following endeavors for calibration were conducted using only the causal branch but can be applied to both branches, if the asymmetry is absent.

3.5 Distribution of Values

The synapse-to-synapse distribution of the amplitude and time constant at fixed analog parameters and calibration bits is presented in Fig. 3.6 and 3.7. In both cases, the data is clearly skewed towards higher values. In consequence, the data may be better described using a log-normal distribution rather than a normal distribution. This is supported by fitting both data sets with a log-normal distribution and performing the Kolmogorov–Smirnov (KS) test : $\mathrm{p} = 0.46$ for the time constants, $\mathrm{p} = 0.99$ for the amplitudes. The KS test measures the distance between the actual and the fitted distribution and p is the probability of the KS test yielding a distance equal or greater than the distance found, assuming the data is log-normally distributed. Small values therefore call the hypothesis of the data being log-normally distributed into question but this is not the case for the presented data.

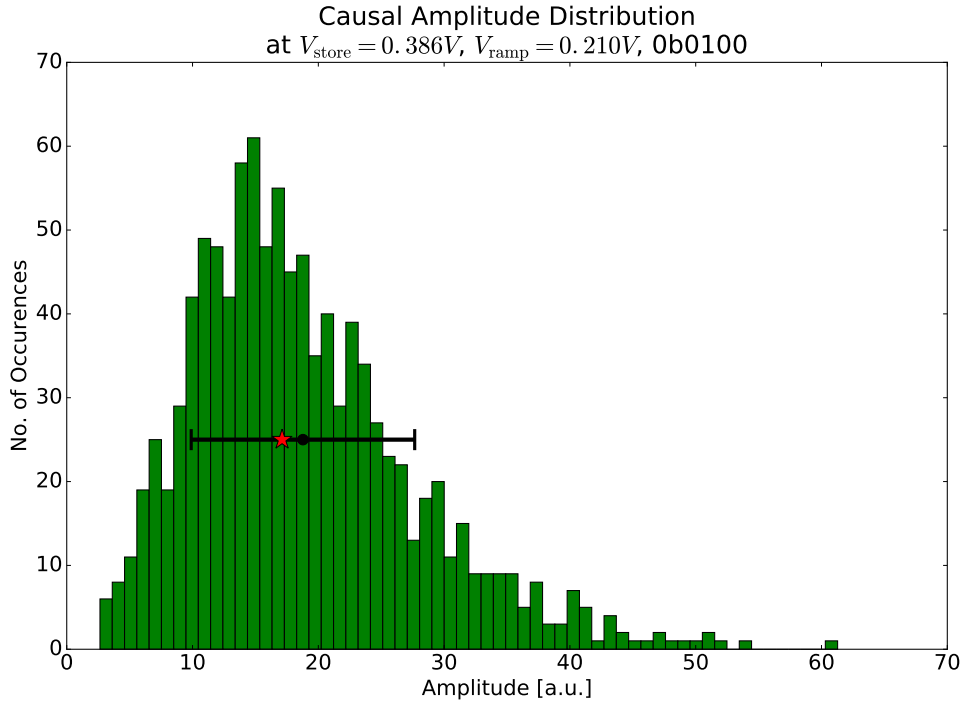

Figure 3.6: The distribution of the amplitudes over the synapses for a fixed set of randomly chosen parameters with mean and standard deviation. Nine (lower plot) or seven (upper plot) spike pairs and 100 linearly spaced Δt in the range $-500\,\mu\mathrm{s} < \Delta t < 500\,\mu\mathrm{s}$ were used in the measurement. Mean and SD are given in black, the median is given as a red star.

The presented time constant distribution has a mean and standard deviation (SD) of $(62 \pm 16)\,\mu\mathrm{s}$ and median of $60\,\mu\mathrm{s}$, the mean and SD for the amplitude distribution are $(19 \pm 9)\,\mathrm{a.u.}$ with a

median of 17 a.u.. Already, this data suggests a substantial degree of variation over the synapses when the calibration bits are all at the same configuration. Because of the aforementioned asymmetry issue, only the causal values were included in both histograms as the systematically decreased amplitude in all anti-causal channels decreases fit quality.

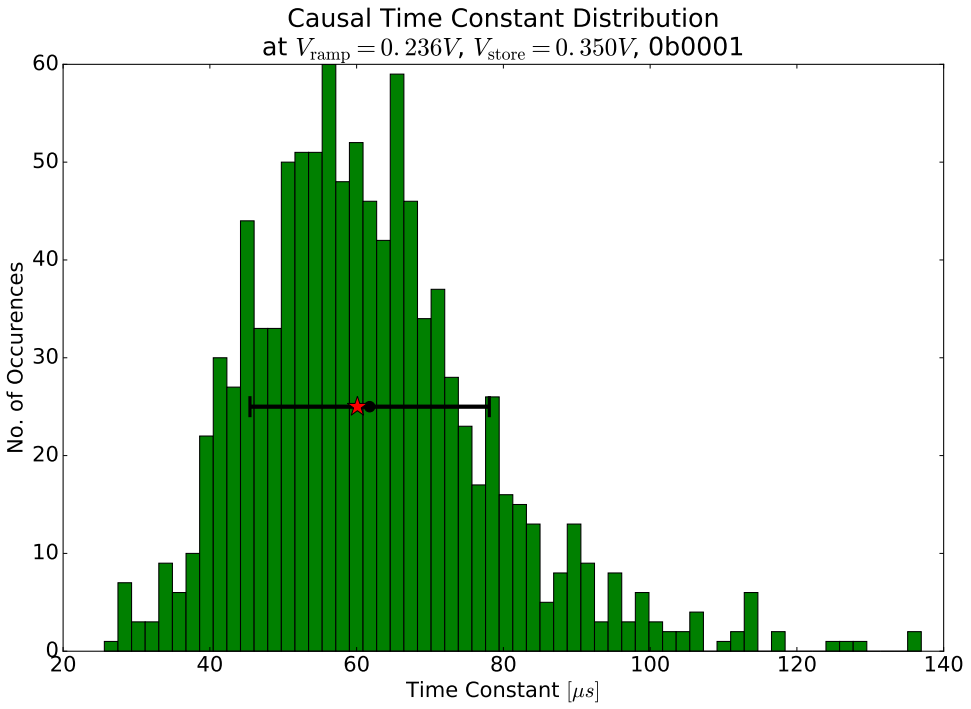

Figure 3.7: The distribution of the time constants with parameters as in Fig. 3.6.

3.6 Independence of Parameters

Ideally, V_{store} should not influence the time constants and V_{ramp} should not influence the amplitudes. The two pairs of calibration bits should also operate independently. Both idealizations needed to be verified. To this end, a range of values for V_{store} and V_{ramp} was measured when the calibration bits were set to $0\mathrm{b}0101$. The two LSBs represent the time constant calibration bits, the two MSBs represent the amplitude calibration bits. A setting of zero implies the greatest values, while setting both bits implies the greatest possible reduction in both cases. The results are given in the two-dimensional colormap plot in Fig. 3.8. The plot suggests that in the measured range, the independence of V_{store} and V_{ramp} does hold. The coefficient of variation of the amplitude increases for smaller amplitudes while the Coefficient of Variation (CV) of the time constant increases for larger time constants. From the plots, it can be estimated that the varia-

tion of the amplitude ranges from 40% to 60% and that of the time constant from 20% to 40%. The quality of the fits is influenced by the available data, which can be sub-optimal for certain synapses at small amplitudes, in the sense that the exponential curve is barely distinguishable from the offset. This explains the unevenness of the CV of the time constant at large values for V_{store} (small amplitudes).

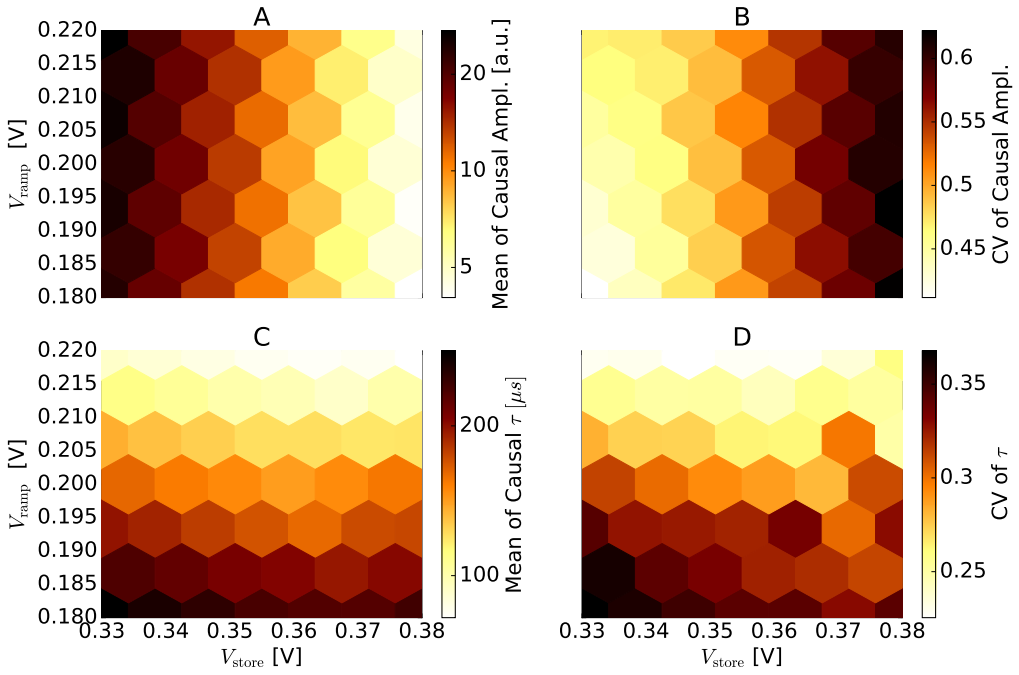

Figure 3.8: A: The mean amplitude of the causal branch of all 1024 synapses for different settings of V_{store} and V_{ramp} using logarithmic color scaling. B: The coefficient of variation of the causal amplitudes using a linear color scale. C: The mean time constant of the causal branch using logarithmic color scaling. D: The coefficient of variation of the time constant using a linear color scale. All measurements were taken with seven spike pairs and a calibration bit setting of 0b0101. 100 linearly spaced data points in the range $-500\,\mu\text{s} < \Delta t < 500\,\mu\text{s}$ were taken.

3.7 Effect of Calibration Bits

The effect of the sixteen possible calibration bit settings on the amplitude and time constant, which is demonstrated in Fig. 3.9 for a specific synapse, was evaluated by measuring the amplitude and time constant of each synapse for each bit setting and across a range of values for V_{ramp} (time constant calibration bits) or V_{store} (amplitude calibration bits). The results at fixed analog parameters are presented as violin plots in Fig. 3.10. It is noticeable that the amplitude calibra-

tion bits have a very strong effect: relative to the 0b01XX setting (X suggests arbitrary setting), the amplitudes differ by around a factor of 0.2 (01/00), 2 (01/10) and 5 (01/11). Judging by the previously demonstrated variation of at most 60%, this seems disproportionate. It is therefore expected that only a subset of the amplitude calibration settings are useful in a calibration. However, it is evident that the amplitude is not influenced by the time constant calibration bits. The same plot for the time constant shows that the bit-to-bit ratio is approximately constant at around 1.2. As the variation of the time constant was shown to be at most 40%, this seems reasonable. Again, the time constant is not systematically influenced by the amplitude calibration bits.

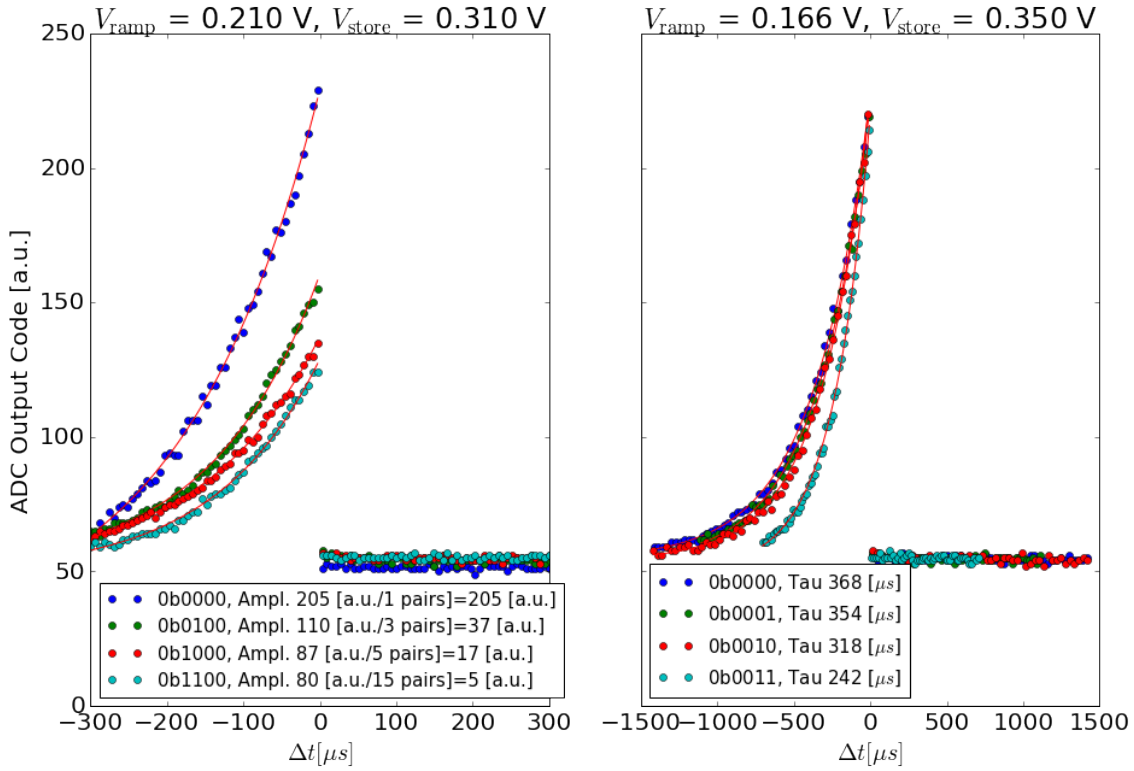

Figure 3.9: Example of the effect of the calibration bits on one specific synapse. Note that for the amplitude scaling, several spike pairs were sent into the synapses. The amplitude normalized to one spike pair at the setting 0b1100 would not be visible in this plot.

Regarding the amplitude calibration bits, the lowest setting (0b00XX) was by design not intended to be used (Schemmel, 2016). Yet when disregarding this setting, the scaling factors of the remaining settings do not seem appropriate for the amount of spread that was determined. It is therefore expected that calibrating the amplitude has little effect on the spread of the amplitudes.

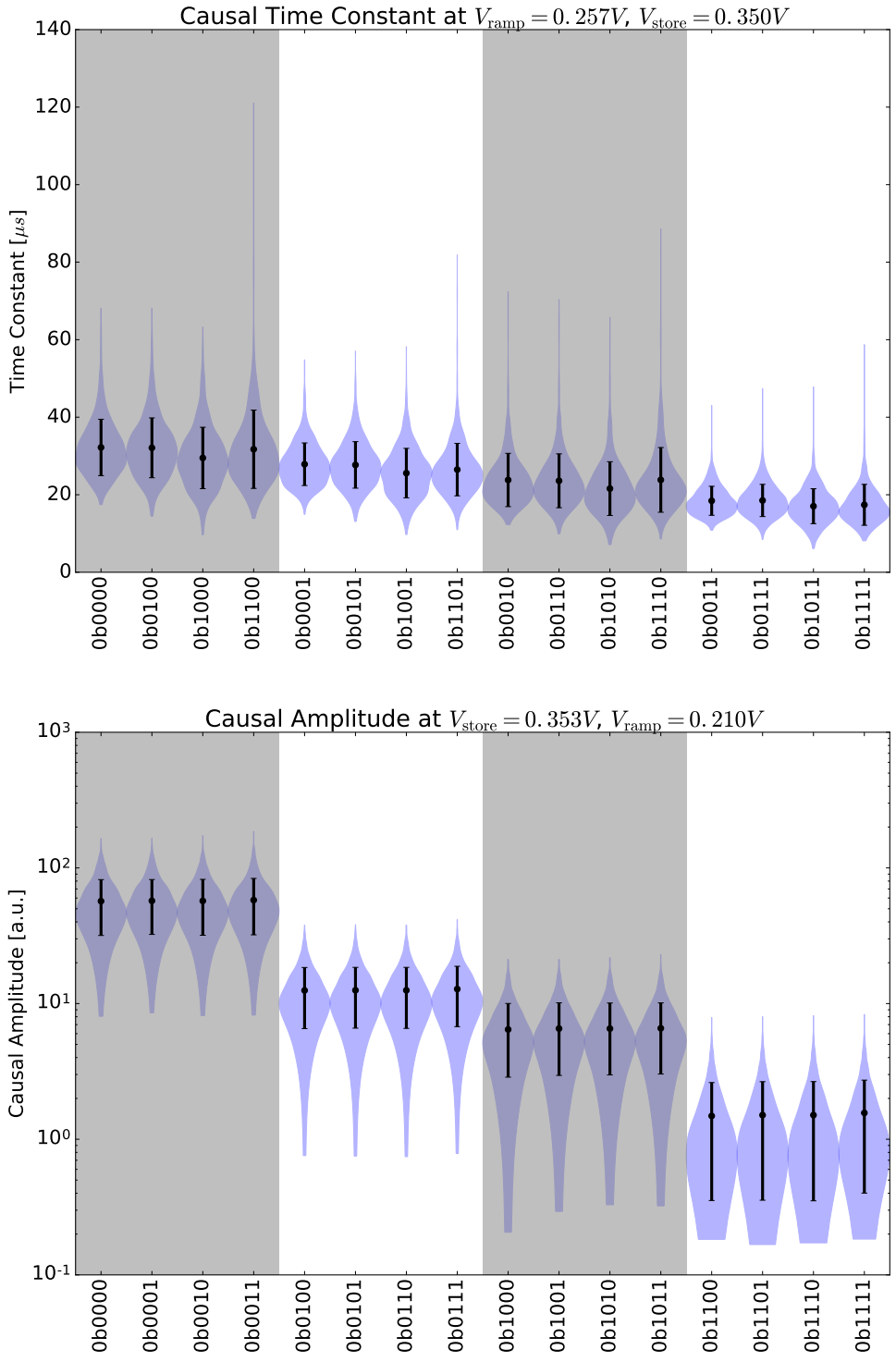

Figure 3.10: The effect of the calibration bits on the causal amplitude and time constant. For each calibration setting, the distribution of the synaptic values was violin-plotted together with arithmetic mean and standard deviation. The shading delimits the ranges of calibration settings for which the amplitude or time constant should stay constant (the two LSBs should not have an effect on the amplitude and vice-versa for the time constant). Note the different ordering of bit settings in both plots.

3.8 Trial-to-Trial Variation

In order to quantify the trial-to-trial variation, 100 measurements of the causal time constants and amplitudes at a fixed set of parameters were conducted. The used parameters were $V_{\text{store}} = 0.31\,\text{V}$, $V_{\text{ramp}} = 0.21\,\text{V}$ and a calibration setting of 0b0101. Each measurement took 100 data points linearly spaced in $-500\,\mu\text{s} < \Delta t < 500\,\mu\text{s}$ and sent in three spike pairs for each data point. The results are presented in Fig. 3.11. For both the time constants and amplitudes, the CVs over the 100 trials were calculated and plotted as a histogram over the synapses (plot A and D). Both distributions of CVs are markedly skewed. For the time constants, the CVs have an arithmetic mean and SD of $(1.6 \pm 0.8)\,\%$ and median of $1.4\,\%$. For the amplitudes, the arithmetic mean and SD of the CVs are $(1.0 \pm 0.4)\,\%$ with a median of $0.9\,\%$. This is the degree of variation to expect as trial-to-trial variation for a single synapse with the used parameters.

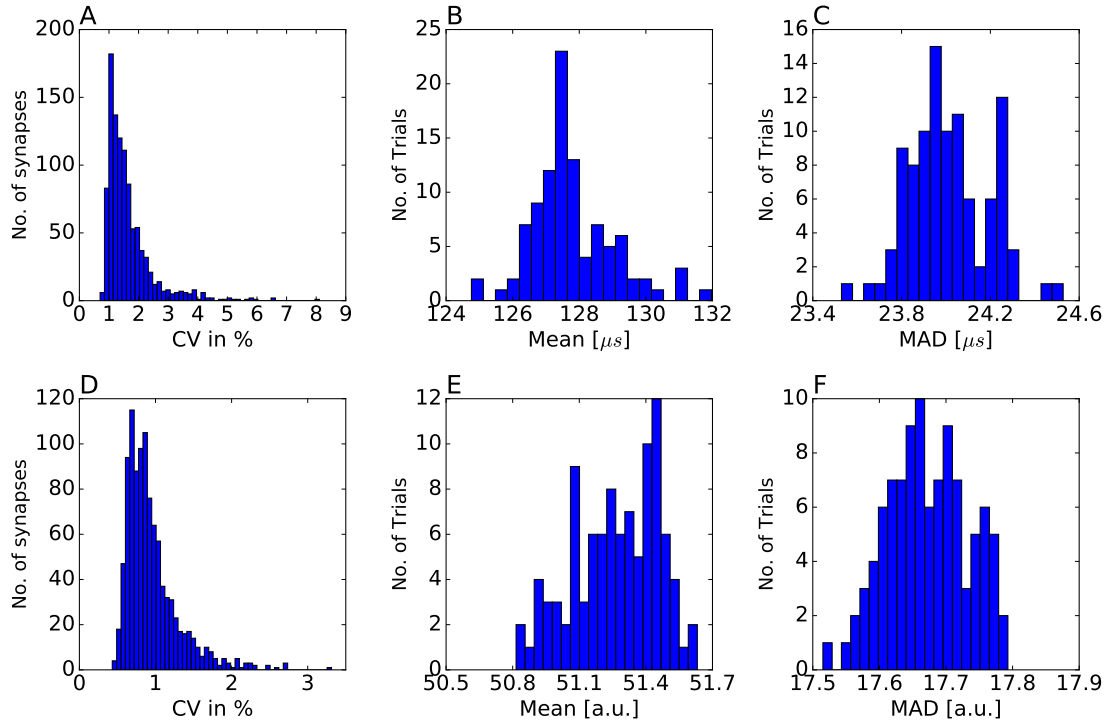

Figure 3.11: A: Histogram of the CVs of the time constants of all synapses based on 100 trials. B: Histogram of the mean of the time constants (averaged over all synapses per trial) for 100 trials. C: Histogram of the MAD of the time constants for 100 trials. D-F: The equivalent of A-C for the amplitude. Used parameters: $V_{\text{store}} = 0.31\,\text{V}$, $V_{\text{ramp}} = 0.21\,\text{V}$, 0b0101.

Plots B, C, E and F show the mean and Mean Absolute Deviation (MAD) calculated over the synapse array for all 100 trials for the time constant or amplitude. The MAD is a linear measure

for the spread of data and defined as $\frac{1}{N}\sum_i(\bar{X} - X_i)$, where the X_i are the values for the time constants or amplitudes and $\bar{X}$ is the arithmetic mean. It is calculated here as it will be used in later calibrations. The mean and SD of the mean over the time constants is $(127.8 \pm 1.2)\,\mu\text{s}$, that of the MAD $(24.0 \pm 0.2)\,\mu\text{s}$. The mean and SD of the mean of amplitudes over the 100 trials is (51.3 ± 0.2) a.u., for the MAD the mean and SD are (17.67 ± 0.06) a.u..

3.9 Calibration

The calibration bits are intended to be used to minimize the spread of the synaptic parameters $\eta_\pm$ and $\tau_\pm$ while the analog parameters (V_{ramp} and V_{store}) can be used to continuously shift the mean of the distribution. Therefore, the natural approach is to minimize the standard deviation (SD) at fixed analog parameters using the calibration bits. This will result in a mean value that is incidental, as it was not the subject of optimization. Doing this for a range of values of V_{ramp} or V_{store} will then enable one to find the best possible configuration, i.e. the set of one analog parameter and calibration bits that minimizes the spread for a given target mean, while providing a mean that is equal to the target mean.

Let X_i be the configured value of X (either time constant or amplitude) in synapse i, N the number of synapses and $\bar{X}$ the arithmetic mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. The standard deviation is then given as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X} - X_i\right)^2} \tag{3.1}$$

which can be conveniently reformulated to

$$\sigma = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)^2\right)}. \tag{3.2}$$

The standard deviation is therefore minimal precisely when

$$\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)^2 \tag{3.3}$$

is minimal. At a fixed analog parameter, the X_i may in each synapse be chosen from the discrete set of four values corresponding to the four possible calibration bit settings. Minimizing the above equation under this constraint is highly non-trivial because of the interaction terms

resulting from squaring the sum over X_i. An exhaustive search is not feasible, as $N = 1024$ but the problem can be tackled using Mixed Integer Quadratic Programming (MIQP). However, even using the fastest available MIQP solver (GUROBI 6.5, Mittelmann (2016)) the full problem converges very slowly with the available computing resources. For this reason, minimizing the standard deviation was deemed impractical.

The problem can be relaxed by minimizing the Mean Absolute Deviation (MAD)

$$\frac{1}{N}\sum_{i=1}^{N}\left|\bar{X} - X_i\right| \tag{3.4}$$

instead of minimizing the standard deviation. This presents a problem of Mixed Integer Linear Programming (MILP) which can be solved more efficiently, in this case using GUROBI 6.5 as solver. In case of a normal distribution, the SD and MAD are related by $\mathrm{SD} = \sqrt{\frac{\pi}{2}}\mathrm{MAD}$. This implies that in that case, minimizing the MAD is equivalent to minimizing the SD. As mentioned earlier, the synaptic values seem to be better described using a log-normal distribution for which the SD and MAD are monotone functions of the SD of the normally distributed logarithmized data (Weisstein, 2016). This means that also in the case of a log-normal distribution, the SD is minimal when the MAD is minimal. Disregarding the constant factor $\frac{1}{N}$, defining x_{ij} as the empirical value of X (time constant or amplitude) in synapse i (1, ..., 1024) at calibration setting j (1, 2, 3, 4), it is formalized as

$$\text{Minimize} \quad \sum_{i=1}^{N}(A_i + B_i) \tag{3.5}$$

$$\text{subject to} \quad \forall i: \quad A_i \in \mathbb{R}_{>0} \tag{3.6}$$

$$\forall i: \quad B_i \in \mathbb{R}_{>0} \tag{3.7}$$

$$\forall i: \quad \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{4} a_{jk}x_{jk} - \sum_{k=1}^{4} a_{ik}x_{ik} = A_i - B_i \tag{3.8}$$

$$\forall i \forall j: \quad a_{ij} \in \{0, 1\} \tag{3.9}$$

$$\forall i: \quad \sum_{j=1}^{4} a_{ij} = 1 \tag{3.10}$$

where the first term on the left hand side of 3.8 is simply the mean and the second term is the value of X in synapse i. The binary values a_{ij} are the subject of optimization and signal whether calibration bit setting j is used in synapse i. A_i and B_i are auxiliary variables in order to introduce the absolute value of the deviation into the linear problem. Because both are constrained

to positive values and they are subtracted in 3.8, either A_i or B_i contains the absolute value of $\bar{X} - X_i$, depending on the sign. Eq. 3.10 constrains the synapses to one active configuration setting.

3.9.1 Time Constant

Regarding the time constants $\tau_{\pm}$, it was previously shown that the scaling factors of the four calibration settings are approximately equal and in proportion to the spread of the empirical values of the time constants. After measuring the values of the time constants in all synapses at all calibration bit settings at different V_{ramp}, the calibration was conducted using all four calibration settings and the strategy outlined before. Solving the minimization problem to within 5% of optimality on a machine with 8 cores and 16GB RAM for $N = 1024$ takes on the order of hours, where the degree of optimality is determined by the solver by finding upper bounds on the optimal minimized solution.

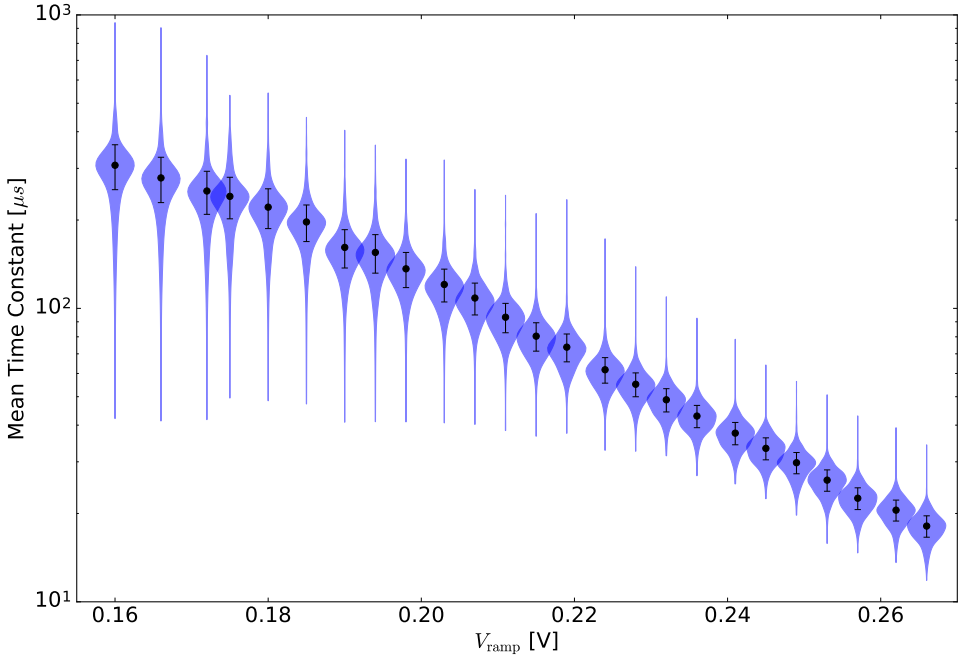

Figure 3.12: Results of the calibration of the time constants at different V_{ramp}. The black dots and error bars show the mean (incidental) and MAD (minimized) of the calibrated distribution. The distributions are visualized using blue violin plots. The goal of the calibration was to center the violin around the mean as much as possible. Each calibration minimized the MAD using the four possible calibration settings for 8000 seconds, generally within 5% of optimality.

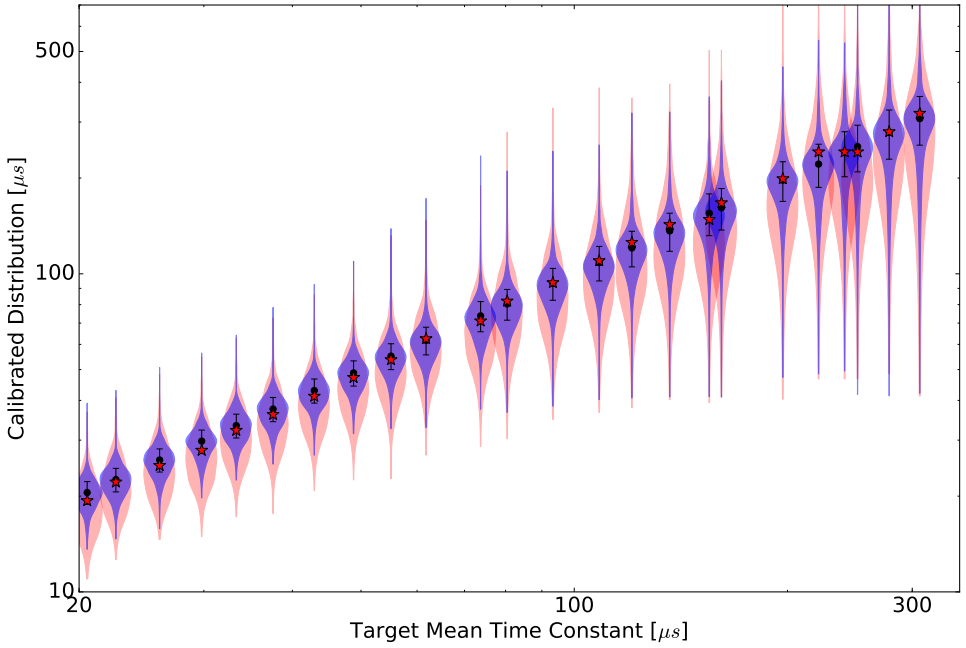

Figure 3.13: Calibration results for the time constants as function of the target mean (i.e. the mean resulting from the calibration) in order to compare the calibrated distribution to the baseline distribution. The distributions of the calibrated values are given as blue violin plots together with arithmetic mean (equal to the target mean) and MAD as error bars. The red violin plots show the distribution at closest measured means if all synapses are set to $0\mathrm{bXX01}$ with a red star marking the mean of this distribution. That is, if only the $0\mathrm{bXX01}$ setting was available, this mean together with the red distribution would be provided for the given target mean as it is the closest measured mean. The red star generally only approximates the target mean because a finite set of values for V_{ramp} was measured.

Fig. 3.12 and 3.13 show the results of the minimization of the time constants. The solver was run for 8000 seconds at a range of values for V_{ramp}. Fig. 3.12 shows the calibration result as a function of V_{ramp}, while Fig. 3.13 shows the calibration result as a function of the mean of the calibrated distribution. The latter figure serves to compare the calibrated distribution to the uncalibrated distribution, i.e. for a specified target mean, compare the calibrated distribution to the closest uncalibrated distribution with regards to the target value for the mean.

A value for the mean (or any other measure of central tendency) specified by the user leads to a look-up on the curve of minimized MADs in Fig. 3.12. The user can then be either presented with a set of calibration bits that stems from a point on the curve that is as close as possible to the desired value or interpolation can take place. In the latter case, additional measurements to ensure the optimality of the calibration bit settings may be warranted.

The values of V_{ramp} were chosen in accordance with the previously used range and the usable range specified in Friedmann *et al.* (2016) (another chip was used which can lead to a shift in the usable range, but the reported values were still usable for the chip used in this thesis). It is evident that the calibration was able to significantly reduce the spread of the values compared to a baseline setting of $0\text{bXX}01$, as the after-calibration distributions (blue) are clearly more centered around the mean and less spread out when compared to the baseline distributions (red).

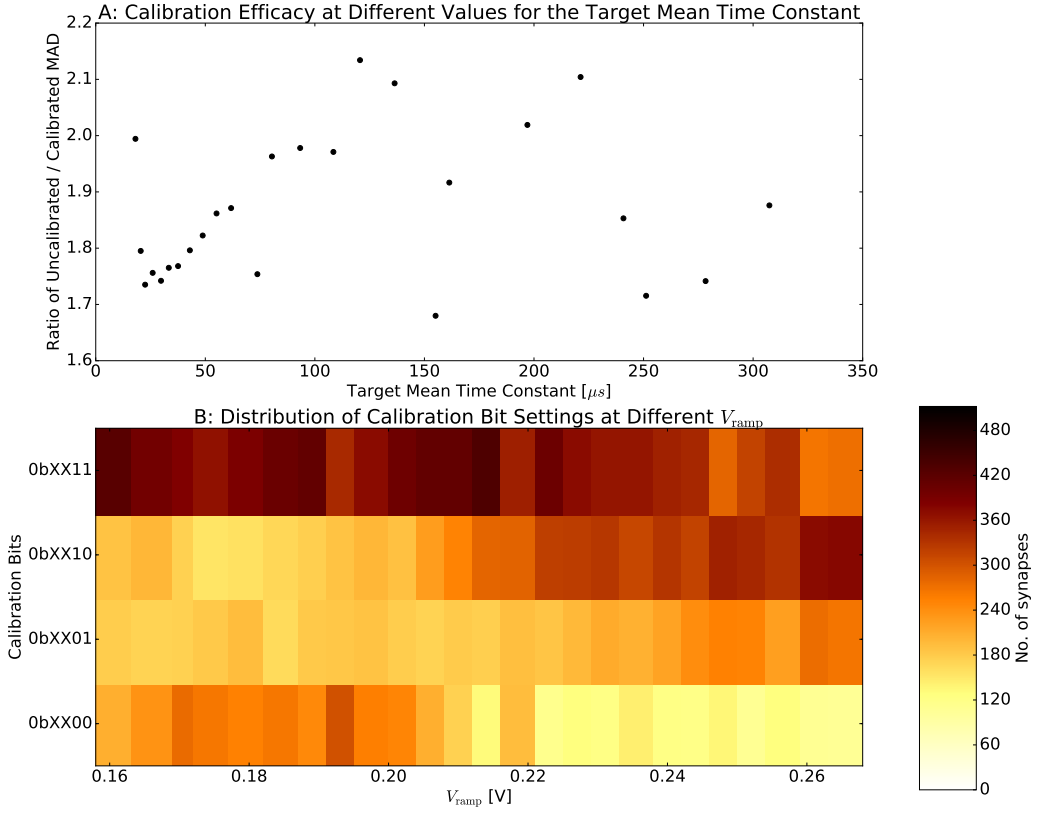

Figure 3.14: A: The ratio of the calibrated and uncalibrated MADs at different target mean time constants (equal to the calibrated means), where $0\text{bXX}01$ is used as the reference setting. The calibration improves (decreases) the MAD by a factor of 1.7 to 2.1. There are more points at smaller time constants because linearly spaced values of V_{ramp} were measured, but the mean time constant is not a linear function of V_{ramp}. B: Visualization of the distribution of the calibration bit settings over the 1024 synapses after the calibration.

The decrease in spread can be quantified by again choosing $0\text{bXX}01$ as a baseline setting and then calculating the ratio of the MADs at different target means, which are chosen to be the means resulting from the calibration. The baseline MADs are determined by choosing the value

for V_{ramp} that provides the closest mean to the target mean (see Fig. 3.13). The results are presented in plot A of Fig. 3.14. Calibrating the synapses decreases the MAD by a factor of 1.7 to 2.1.

It is of interest how the calibration distributes the calibration bit settings over the synapses. For each calibration at a value for V_{ramp}, the distribution is visualized in plot B of Fig. 3.14. Note that there is a total of 1024 synapses and the color scale was set to a maximum of 512 synapses. Already, this shows that no calibration bit setting is used in more than 50% of the synapses after any calibration. At smaller values for V_{ramp} (larger time constants), the strongest calibration setting (0bXX11) tends to get used more often. This is probably related to the fact that the relative spread (coefficient of variation) increases for larger time constants which was demonstrated previously, as this implies a larger number of outliers and the distribution is skewed towards higher values.

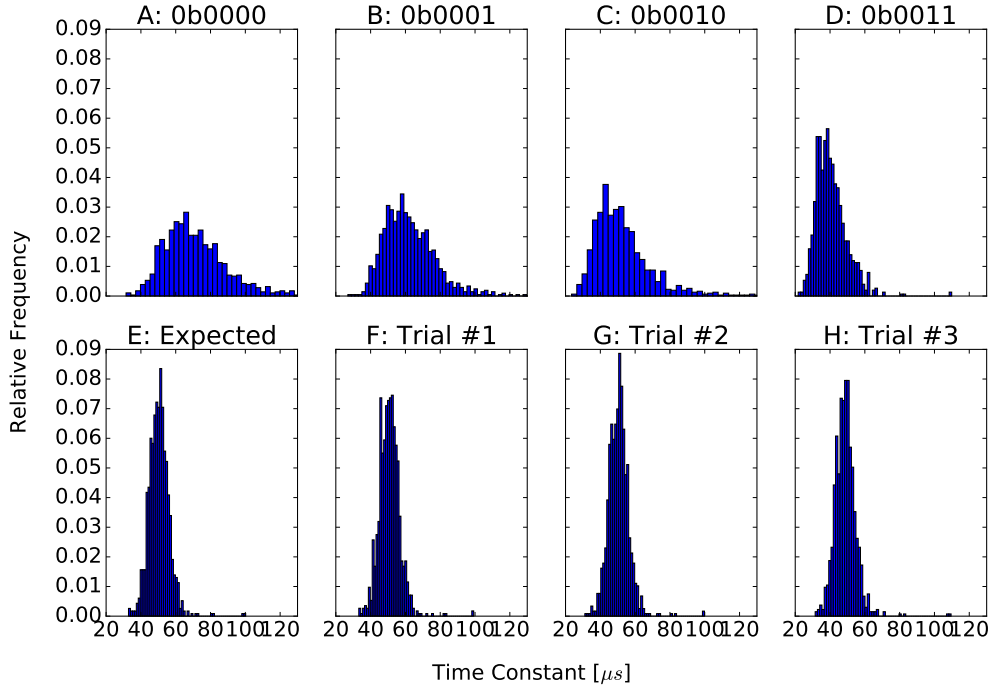

Figure 3.15: A-D: Time constant distribution when all synapses are set to the given bit pattern. These are the values that have been used as empirical values for the calibration. E: Expected distribution after calibration, resulting from combining the values in A-D as dictated by the calibration. F-H: Actual value distribution in three trials when setting the synapse array with the bit pattern resulting from the calibration. Used parameters: $V_{\text{ramp}} = 0.232\,\text{V}$, $V_{\text{store}} = 0.35\,\text{V}$.

The calibration bit pattern resulting from one calibration at $V_{\text{ramp}} = 0.232\,\text{V}$ was used to per-

form a new measurement where the synapses were set using this pattern, in order to verify that the desired calibration effect takes place. The results are presented in Fig. 3.15 and suggest that the distribution resulting from setting all synapses to their calibrated configuration approximates the expected distribution that results from combining the empirical synaptic values which were taken when all synapses were set to the same configuration. The expected minimized value for the MAD was $4.44\,\mu s$ and the MADs of the three trials were $4.38\,\mu s$, $4.45\,\mu s$ and $4.37\,\mu s$. The expected mean was $48.8\,\mu s$, the actual means were $50.4\,\mu s$, $50.8\,\mu s$ and $50.5\,\mu s$. Comparing this with the trial-to-trial variation reported previously, it is found that the expected MAD is consistent with the measured MADs but the mean differs by at least 3%, which is above the degree of trial-to-trial variation that was found before. The cause for this is not clear.

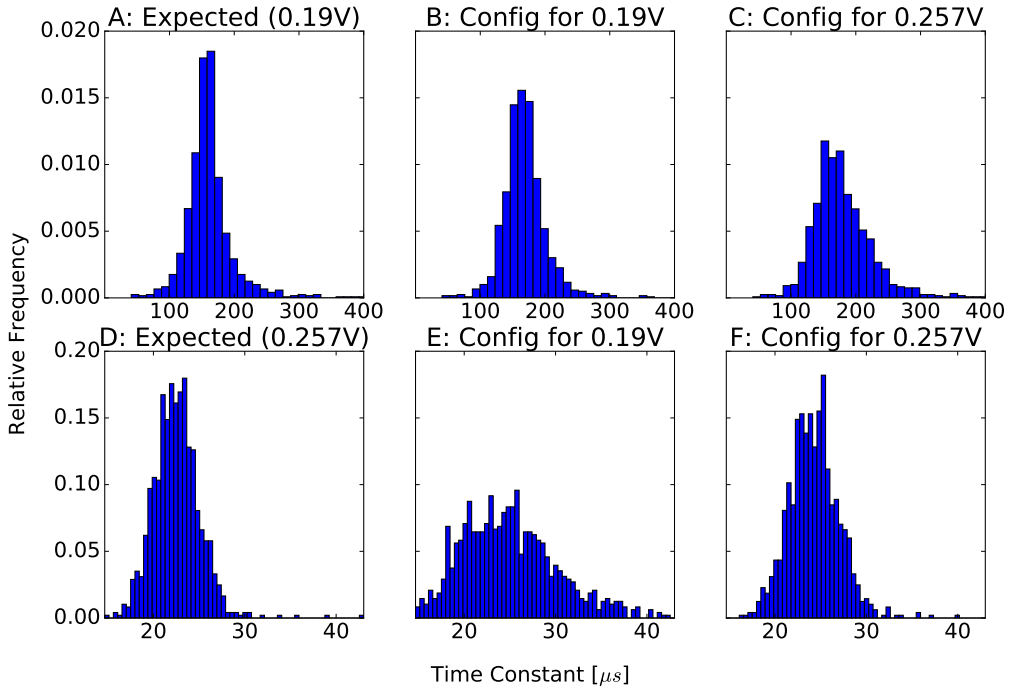

Figure 3.16: The calibration bit pattern found for $V_{\text{ramp}} = 0.19\,\text{V}$ was tested at $V_{\text{ramp}} = 0.257\,\text{V}$ and vice-versa. A, D: The expected distribution at $V_{\text{ramp}} = 0.19\,\text{V}$ or $V_{\text{ramp}} = 0.257\,\text{V}$ after calibration, based on the empirical values used in the calibration. B: The measured distribution using the "right" bit pattern at $0.19\,\text{V}$. C: The measured distribution using the "wrong" bit pattern at $0.19\,\text{V}$. E: The measured distribution using the "wrong" bit pattern at $0.257\,\text{V}$. F: The measured distribution using the "right" bit pattern at $0.257\,\text{V}$.

In order to test the hypothesis whether the calibration bit pattern of a specific value of V_{ramp} also provides good performance at other values, the calibration bit pattern for $0.19\,\text{V}$ (lower end of the usable range) was measured at $0.257\,\text{V}$ (upper end of the usable range) and vice-versa. The

results are given in Fig. 3.16. It is clearly visible that the distributions when using the "wrong" bit pattern are less narrow and more spread out when compared to the distribution resulting from the original bit pattern, suggesting that there is no single bit pattern that is optimal for all values of V_{ramp}. For $V_{\mathrm{ramp}} = 0.19\,\mathrm{V}$, the expected mean and MAD were $161\,\mu\mathrm{s}$ and $24\,\mu\mathrm{s}$ (plot A). Using the corresponding calibration, the mean and MAD were $169\,\mu\mathrm{s}$ and $24\,\mu\mathrm{s}$ (plot B) while the values when using the other calibration pattern were $181\,\mu\mathrm{s}$ and $37\,\mu\mathrm{s}$ (plot C). The expected values for $V_{\mathrm{ramp}} = 0.257\,\mathrm{V}$ were $22.5\,\mu\mathrm{s}$ and $1.9\,\mu\mathrm{s}$ (plot D), the correct bit pattern gave $24.3\,\mu\mathrm{s}$ and $2.1\,\mu\mathrm{s}$ (plot E), the other bit pattern gave $25.1\,\mu\mathrm{s}$ and $4.2\,\mu\mathrm{s}$ (plot F). These results show that using the wrong bit pattern yields a significantly higher MAD. Again, the measured values for the mean deviate from the expected mean, the cause for which is unclear.

3.9.2 Amplitude

It was previously shown that the scaling factors of the amplitude calibration bits are very large (at least 2) and overall in disproportion to the spread of the amplitudes. The calibration was therefore limited to the middle settings ($0\mathrm{b}01\mathrm{XX}$ and $0\mathrm{b}10\mathrm{XX}$), which differ by around a factor of two, the smallest possible combination. This severely decreases the time required for minimization of the MAD compared to the time constant: a solution within 5% of optimality is reached on the order of a few minutes instead of hours.

As might be expected because two instead of four degrees of freedom per synapse are available, the amplitude calibration performs worse than the time constant distribution when comparing the ratios of uncalibrated to calibrated MADs. Still, it is able to reduce the MAD by a factor of 1.4 to 1.8 compared to a baseline setting of $0\mathrm{b}01\mathrm{XX}$ as is visible in Fig. 3.18. Among the two calibration bit settings, the stronger one ($0\mathrm{b}10\mathrm{XX}$) is clearly favored. This might be because the calibration decreases the absolute, not relative spread and the scaling factor is around 2, which means that the distribution of $0\mathrm{b}10\mathrm{XX}$ provides smaller absolute values.

Fig. 3.19 was created to verify the result of the calibration by using the bit pattern resulting from a calibration at $V_{\mathrm{store}} = 0.319\,\mathrm{V}$ in a new measurement. The calibration gave an expected mean of $24.4\,\mathrm{a.u.}$ and expected MAD of $5.6\,\mathrm{a.u.}$, the measured results in the three trials were equal up to the first decimal point: a mean of $25.6\,\mathrm{a.u.}$ and a MAD of $6.3\,\mathrm{a.u.}$. While the discrepancy between expected and measured values that is not explained by trial-to-trial variation remains an unsolved issue, the measured distributions seem to resemble the expected distribution.

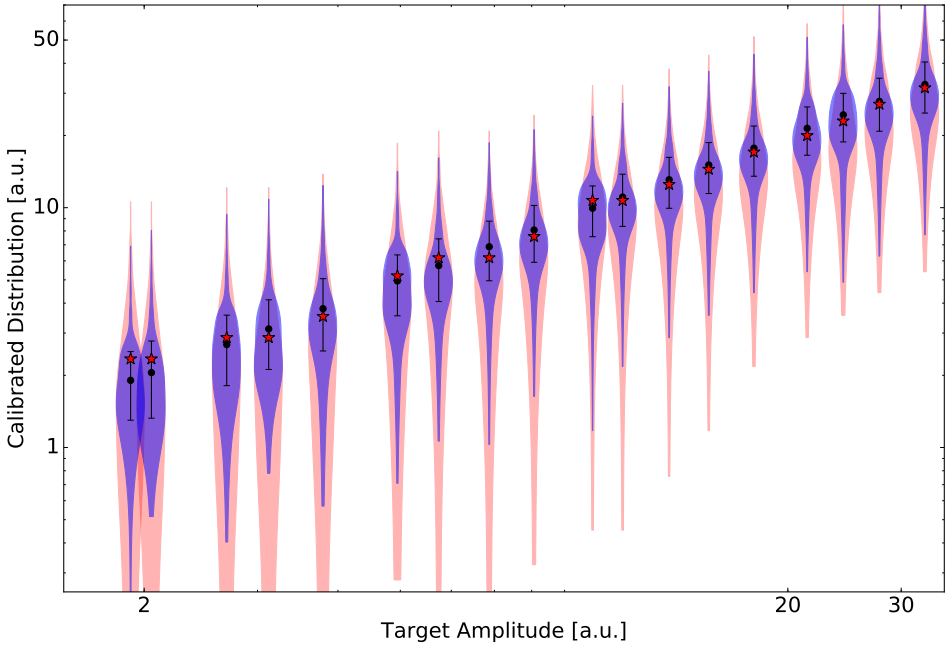

Figure 3.17: Results of the amplitude calibration as a function of the mean after calibration, i.e. the mean for which a user might request the minimized distribution. The blue violin plots and black error bars show the distribution and MAD of the calibrated values. The uncalibrated distribution is given in red and uses only 0b01XX as bit setting. The red star marks the mean of the distribution, which is chosen to be as close as possible to the calibrated mean.

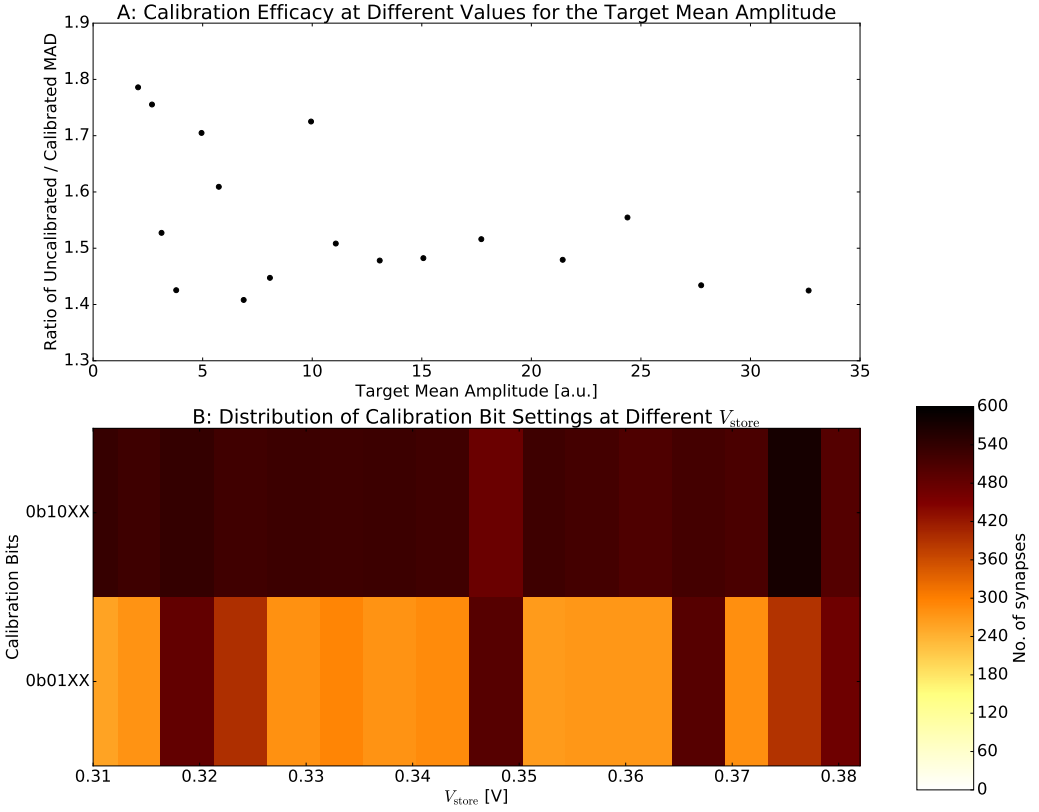

Figure 3.18: A: The ratio of uncalibrated to calibrated MAD at different values for the target mean amplitude (equal to the calibrated means). For the baseline reference, only the 0b01XX was allowed. The calibration reduces the MAD by a factor of 1.4 to 1.8. B: The distribution of the calibration bit settings among the two possible settings at different V_{store}.

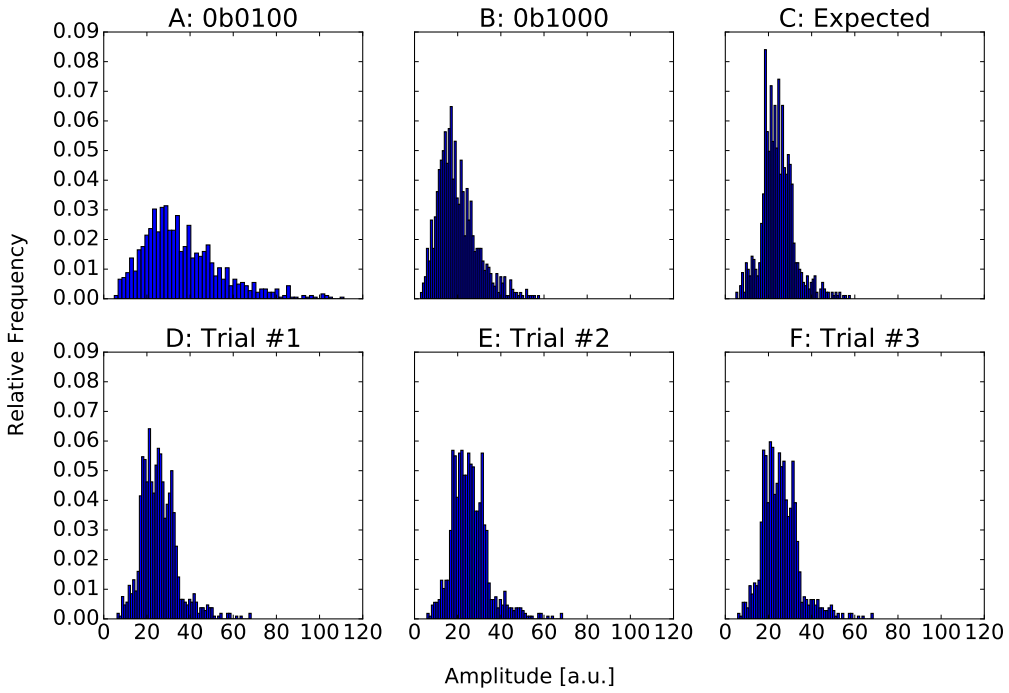

Figure 3.19: A, B: The distributions of $0\mathrm{b}01\mathrm{XX}$ and $0\mathrm{b}10\mathrm{XX}$ at $V_{\mathrm{store}} = 0.319\,\mathrm{V}$ that were used as empirical values for the calibration. C: Distribution after calibration resulting from combining the values in A, B. D-F: Measured distributions using the calibration bit pattern in three trials. Used parameters: $V_{\mathrm{ramp}} = 0.21\,\mathrm{V}$, $V_{\mathrm{store}} = 0.319\,\mathrm{V}$.

4 Discussion

During this thesis, the parts of the HICANN-DLS responsible for storage and readout of synaptic correlation were characterized and their functionality verified. Software was developed that enables the user to probe the correlation mechanism of the chip and repeat all measurements mentioned in this thesis. Two unexpected issues emerged, most prominently the asymmetry of amplitudes in the causal and anti-causal channel. The cause for this issue was narrowed down to the signal representing a post-synaptic spike by finding the dependency on the number of fired neuronal columns. This issue was under scrutiny at the time of publication of this thesis and is expected to be resolved. Another cause for concern was the effect of the calibration bits for the amplitude, as no scaling factor is below 2 while the coefficient of variation is well below 1, implying that no matter how the baseline is defined, even the lowest scaling factor is not proportionate to the spread of the values.

In the scope of the calibration, the MAD was chosen as a measure for the spread of the data and therefore as the subject of minimization. As was shown, this is equivalent to minimizing the standard deviation of the synaptic values. Using a MILP solver for the combinatorial problem of assigning calibration settings to the synapses guarantees that an optimal solution will be found or at least, that the found solution is within a certain margin of optimality. Other solutions may use heuristic methods like genetic algorithms or simulated annealing. However, it was demonstrated that it is not necessary to resort to purely heuristic methods.

While the calibration result is in principle arbitrarily close to optimal, it is naturally tainted by the quality of the empirical data used to judge a certain calibration bit pattern. This data consists of fit results produced for each synapse at a specific set of chip parameters (voltages and bit settings) and measurements parameters (Δt range, number of spike pairs, number of samples). The presented data suggests that outliers either in amplitude or time constant are not uncommon in the synapse array, which can make it difficult to use one set of measurement parameters

for all synapses as the outlier fits might fail. If this was the case, the value for that synapse was disregarded for the calibration. The goodness of the fit parameters plays a critical role in the calibration and improved measurement methods could account for this by using synapse-specific measurement parameters, instead of using one set of parameters for the entire synapse array as is the case in this thesis.

Measuring a single data point for the entire synapse array using one spike pair with specific Δt currently takes around one second. This is vastly disproportionate to the duration of the spike train, as Δt is generally well below $1\,\mathrm{ms}$. The overhead is caused by controlling the measurement using Python and communicating data to the host computer row-wise. This could be improved by extending the role of the PPU in measurements, as one set of correlation values corresponds to $2 \cdot 1024 = 2048$ bytes and the PPU has 16 KiB memory (Friedmann *et al.*, 2016). The PPU could therefore be used to reduce the amount of communication required to transfer data for the entire array.

The presented calibration method is able to reduce the chosen measure of spread (MAD) of the time constants by a factor of 1.7 to 2.1 and that of the amplitude by a factor of 1.4 to 1.7. The variation in the factor of improvement may in part stem from the fact that in both cases the convergence to the optimal solution was not complete, as the minimization process of the time constants was time-limited to $8000\,\mathrm{s}$ and the minimization of the amplitude MAD was aborted after the result was within 5% of optimality. The problem model and solver parameters were not specifically optimized and doing this in the future could make the optimization process more efficient. Still, the marked improvement of in both cases shows that the calibration bits can be used to homogenize the synapse array.

The amplitude calibration has shown that only two calibration settings can suffice to reduce the MAD by a significant factor. The limitation to two calibration settings drastically reduces the complexity of the chosen method to tackle the minimization problem and solving the minimization problem is faster by around two orders of magnitude. Therefore, it might be of interest to investigate how the number and scaling of calibration settings should be chosen in order to provide a good trade-off between calibration speed and efficiency. In the case of the time constant, it was demonstrated that the four different calibration settings are distributed relatively evenly across the synapse array, suggesting that the chosen scaling factors are reasonable.

For both the amplitude and time constant, the after-calibration measurement that was intended to reproduce the calibration result using the corresponding bit pattern yielded results which showed a similar distribution of values and MAD, yet the mean was significantly different from the expected results. The cause for this could not be conclusively determined due to the temporal limitation on the authorship of the thesis but it is assumed to be a measurement artifact

rather than a malfunction of the chip. The latter would arise from setting each synapse configuration to individual values instead of setting the whole array to the same value, as the empirical values for the calibration were determined by setting the entire synapse array with a constant bit pattern. It might rather be the case that a difference in measurement parameters (Δt range, number of spike pairs, number of samples) led to a systematic shift of the values.

In summary, it was demonstrated that the synaptic correlation mechanism on the HICANN-DLS is generally functioning, may therefore be used to implement STDP-like learning mechanisms and that calibrating the synaptic parameters via calibration bits in each synapse is feasible and effective.

Bibliography

Bi, G Q, & Poo, M M. 1998. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, 18(24), 10464–72.

Bi, Guo-qiang, & Poo, Mu-ming. 2001. Synaptic Modification by Correlated Activity: Hebb's Postulate Revisited. *Annual Review of Neuroscience*, 24(1), 139–166.

Bichler, Olivier, Querlioz, Damien, Thorpe, Simon J, Bourgoin, Jean-Philippe, & Gamrat, Christian. 2012. Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity. *Neural Networks*, 32(8), 339–48.

Clopath, Claudia, Büsing, Lars, Vasilaki, Eleni, & Gerstner, Wulfram. 2010. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neuroscience*, 13(3), 344–352.

Diehl, Peter U, & Cook, Matthew. 2015. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience*, 9(1), 99.

Douglas, R M, & Goddard, G V. 1975. Long-term potentiation of the perforant path-granule cell synapse in the rat hippocampus. *Brain research*, 86(2), 205–15.

Feuillet, Lionel, Dufour, Henry, & Pelletier, Jean. 2007. Brain of a white-collar worker. *Lancet (London, England)*, 370(9583), 262.

Friedmann, Simon, Schemmel, Johannes, Gruebl, Andreas, Hartel, Andreas, Hock, Matthias, & Meier, Karlheinz. 2016. Demonstrating Hybrid Learning in a Flexible Neuromorphic Hardware System. 4.

Froemke, Robert C., & Dan, Yang. 2002. Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, 416(6879), 433–438.

Hartel, A. 2016. *Personal communication*.

Hebb, Donald. 1949. *The Organization of Behavior*. New York: Wiley & Sons.

Hershey, John R., Rennie, Steven J., Olsen, Peder A., & Kristjansson, Trausti T. 2010. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech & Language*, 24(1), 45–66.

Hock, Matthias. 2014 (7). *Modern Semiconductor Technologies for Neuromorphic Hardware*.

Ito, M. 1989. Long-term depression. *Annual review of neuroscience*, 12, 85–102.

Keysers, Christian, & Perrett, David I. 2004. Demystifying social cognition: a Hebbian perspective. *Trends in cognitive sciences*, 8(11), 501–7.

Markram, H, Lübke, J, Frotscher, M, Sakmann, B, Hebb, D. O., Bliss, T. V. P., Collingridge, G. L., Friedlander, M. J., Sayer, R. J., Redman, S. J., Stuart, G., Sakmann, B., Regehr, W. G., Connor, J. A., Tank, D. W., Markram, H., Lübke, J., Frotscher, M., Roth, A., Sakmann, B., Markram, H., Tsodyks, M., Gustafsson, B., Wigstrom, H., Abraham, W. C., Huang, Y.-Y., Stanton, P. K., Sejnowski, T. J., & Singer, W. 1997. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science (New York, N.Y.)*, 275(5297), 213–5.

Millner, S., Hartel, A., Schemmel, J., & Meier, K. 2012. Towards biologically realistic multi-compartment neuron model emulation in analog VLSI. *In: ESANN 2012 proceedings*.

Mittelmann, H. 2016. *Mixed Integer QP Benchmark*.

Otmakhov, Nikolai, Shirke, Aneil M., & Malinow, Roberto. 1993. Measuring the impact of probabilistic transmission on neuronal output. *Neuron*, 10(6), 1101–1111.

Rochester, N., Holland, J., Haibt, L., & Duda, W. 1956. Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IEEE Transactions on Information Theory*, 2(3), 80–93.

Schemmel, J. 2016. *Personal Communication*.

Shatz, C J. 1992. The developing brain. *Scientific American*, 267(3), 60–7.

Sjöström, Per Jesper, Turrigiano, Gina G, & Nelson, Sacha B. 2001. Rate, Timing, and Cooperativity Jointly Determine Cortical Synaptic Plasticity. *Neuron*, 32(6), 1149–1164.

Song, S, Miller, K D, & Abbott, L F. 2000. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, 3(9), 919–926.

Stallkamp, J, Schlipsing, M, Salmen, J, & Igel, C. 2012. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32(8), 323–32.

Teyler, T J, & DiScenna, P. 1987. Long-term potentiation. *Annual review of neuroscience*, 10, 131–61.

Weiller, C, Chollet, F, Friston, K J, Wise, R J, & Frackowiak, R S. 1992. Functional reorganization of the brain in recovery from striatocapsular infarction in man. *Annals of neurology*, 31(5), 463–72.

Weisstein, Eric W. 2016. Mean Deviation, Log Normal Distribution. *MathWorld–A Wolfram Web Resource*.

Zenke, Friedemann, & Gerstner, Wulfram. 2014. Limits to high-speed simulations of spiking neural networks using general-purpose computers. *Frontiers in neuroinformatics*, 8(1), 76.

Erklärung

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 10.08.2016,