Faculty of Physics and Astronomy

University of Heidelberg

Bachelor's Thesis

in Physics,

submitted by

Johannes Weis

born in Heppenheim, Germany

February 2018

Characterization and Calibration of Synaptic Plasticity on Neuromorphic Hardware

This thesis has been carried out by

Johannes Weis

at the

Kirchhoff-Institute for Physics,

Heidelberg University

under the supervision of

Prof. Dr. Karlheinz Meier

Abstract

Neuromophic hardware platforms are developed with different aspects of the human brain's architecture in mind. The HICANN-DLS 3 neuromorphic ASIC, implemented in a $65\,\mathrm{nm}$ process, contains 32 analog neurons and 1024 hardware synapses. This thesis focuses on synaptic plasticity, specifically Short Term Plasticity (STP) and Spike-Timing Dependent Plasticity (STDP), which are both implemented in hardware. Concerning STP, we observe that synapses can be configured to utilize $27\,\%$ to $74\,\%$ of the available neurotransmitters when transferring an action potential. The recovery time constants cover three orders of magnitude, ranging from $2.38\,\mathrm{ms}$ to $2120\,\mathrm{ms}$ of biological time.

Synapse drivers, which process STP, are subject to mismatch. An offset parameter counteracting the mismatch is calibrated for every driver. We implement a new readout method using the neurons' spike events instead of an ADC to acquire the data for calibration. Using this highly scalable readout mechanism, it is possible to calibrate the synapse drivers in 30 seconds.

Lastly, we characterize the STDP correlation sensors. Previous prototypes showed a strong asymmetry of amplitudes between causal and anticausal measurements. The problem is solved for the latest version by soldering a capacitor onto the board next to the chip.

Zusammenfassung

Neuromorphe Hardware wird nach dem Vorbild des menschlichen Gehirns entwickelt. Der neuromorphe HICANN-DLS-3-Prototyp-Chip enthält 32 analoge Neuronen und 1024 Synapsen in Hardware. Diese Arbeit behandelt synaptische Plastizität, genauer gesagt Short Term Plasticity (STP) und Spike-Timing-Dependent Plasticity (STDP). Beide sind in Hardware implementiert. In Bezug auf STP haben wir beobachtet, dass der konfigurierbare Bereich der Ausschüttung von Neurotransmittern bei der Weitergabe eines Aktionspotentials zwischen $27\,\%$ und $74\,\%$ einstellbar ist. Die verfügbaren Zeitkonstanten der Erhohlung decken drei Größenordnungen ab, von $2.38\,\mathrm{ms}$ bis $2120\,\mathrm{ms}$ biologischer Zeit.

Synapsentreiber, in welchen STP modelliert wird, unterliegen Fertigungstoleranzen. Ein Offset-Parameter, welcher genutzt werden kann, den Unterschied auszugleichen, wird für jeden Treiber kalibriert. Wir implementieren eine neue Auslesemethode, welche die Aktionspotentiale der Neuronen anstelle eines ADC benutzt, um die zur Kalibration benötigten Daten aufzunehmen. Mit dieser gut skalierbaren Auslesemethode ist es möglich, die Synapsentreiber in 30 Sekunden zu kalibrieren.

Zuletzt charakterisieren wir die STDP-Korrelationssensoren. Bisherige Prototypen zeigten eine starke Asymmetrie der Amplituden zwischen kausalen und antikausalen Messungen. Das Problem ist für die aktuelle Generation gelöst, indem ein Kondensator auf dem Board neben dem Chip eingelötet wurde.

Contents

1 Introduction

In the Human Brain Project, different subprojects are focused on researching the brain and developing neuromorphic hardware, novel computing systems inspired by the brain's architecture. Strong cooperation between groups working in many different fields of science is necessary to achieve better understanding of this complex matter. When investigating neural networks, simulations are conducted. The available simulation software requires lots of computational power, especially when using complex models of neurons and synapses [Kunkel et al., 2013]. As an alternative to simulations, dedicated hardware is designed to emulate biological systems directly, called neuromorphic hardware. Neuromorphic systems have the potential to overcome the energy efficiency problem while offering further improvements over digital computing [Meier, 2015].

The Electronic Vision(s) group at the Kirchhoff-Institute for Physics in Heidelberg develops a neuromorphic Application-Specific Integrated Circuit (ASIC) containing both digital and analog signals. We are working on a new chip generation called High Input Count Analog Neural Network (HICANN) with Digital Learning System (DLS). Being based on current theories of how neurons and synapses work, the hardware we build enables brain researchers to investigate their models in accelerated time while using large systems [Schemmel et al., 2010]. There are also purely digital approaches: the SpiNNaker platform, also part of the Human Brain project, was created to provide highly parallel computing using ARM9 cores [Furber et al., 2013].

Machine learning allows systems to learn solving problems without being provided a specific algorithm. Using conventional computers, machine learning is increasingly applied to physics research, including particle physics, nuclear physics and condensed matter physics [Pang et al., 2018]. Custom hardware is even developed in the commercial sector: one of many examples is Google, where a specialized tensor processing unit yields advantages towards traditional processors [Jouppi et al., 2017].

For the process of learning, synapses are important. They form the connections between neurons, which receive, process and transmit information. Synaptic plasticity, the change of connection strength over time, is characterized in this thesis. We focus on two effects, happening on different timescales: Firstly, Spike Timing Dependent Plasticity (STDP) [Bi and Poo, 2001] models long-lasting changes in connection strength between neurons. If a synapse transfers input towards a neuron shortly before the neuron sends out an action potential itself, the strength of this synaptic connection, its weight, is typically increased. Secondly, the weights of synaptic connections can vary significantly over short times as well. We call this effect Short Term Plasticity (STP) [Zucker and Regehr, 2002]. The release of neurotransmitters in chemical synapses induces various processes that lead to either depression or amplification of Post-Synaptic Potentials (PSPs) over time. Depending on the brain region the synapses are located in, one of the effects is stronger, but generally both effects can take place at the same time.

During this thesis, we will present basics of the current implementation of STP and STDP on the HICANN-DLS 3 prototype chip. We will conduct experiments characterizing the function of the models and the available configuration range. In particular, the STP mechanism will be calibrated in order to counteract variances in the circuits' behaviour caused by production variances.

2 Principles

2.1 Biological Principles

Neurons in a brain can be coarsly split into different segments: dendrites, a soma and an axon. Synaptic inputs travel along dendrites towards the soma, which can be understood as the essence of the neuron. Once inputs have excited the neuron sufficiently, an action potential is created and sent along the axon. This action potential reaches the dendrites of averagely 7000 other neurons. Synapses normally connect axons and dendrites of different neurons. In total, the human brain is estimated to consist of 20 billion neurons and 100 to 500 trillion synapses [Drachman, 2005].

Most synapses transfer action potentials by using neurotransmitters [Lytton, 2007]. Stored in vesicles in the presynaptic neuron's axon, neurotransmitters get emitted when the axon is activated during an action potential. They travel through the synaptic cleft and reach receptors on a dendrite of the postsynaptic neuron.

2.1.1 Short Term Plasticity

The path of neurotransmitters along the synaptic cleft allows for synaptic plasticity. Lasting at most a couple of minutes, this is referred to as Short Term Plasticity [Zucker and Regehr, 2002]. The received input at the postsynaptic neuron decreases over the course of repeated stimulation due to the limited amount of disposable neurotransmitters. This vesicle depletion leads to Short Term Depression (STD) [Hennig, 2013]. There is also the opposite phenomenon: an increase in the synaptic efficiency leads to higher received inputs, called Short Term Facilitation (STF) [Stevens and Wesseling, 1999]. The term *Short Term Plasticity* includes both depression and facilitation, they can happen at the same time [Hennig, 2013].

In order to express depression and facilitation in equations, we divide synaptic neurotransmitters into three partitions using the *Tsodyks-Markram model* [Tsodyks and Markram, 1997]. It uses a recovered partition R, an effective partition E, and the current I flowing onto a neuron's membrane. The values of R and E lay between 0 and 1, the time of the action potential is given by t_{AP}. Denoting states before an action potential with upper indices x^- and afterwards with x^+, the transfer of neurotransmitters over time t is described by a set of three differential equations [Tsodyks and Wu, 2013]:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{E}{\tau_{\mathrm{facilitation}}} + U_{\mathrm{SE}} \cdot (1 - E^-) \cdot \delta(t - t_{\mathrm{AP}}) \tag{2.1}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{1 - R}{\tau_{\mathrm{depression}}} - E^+ \cdot R^- \cdot \delta(t - t_{\mathrm{AP}}) \tag{2.2}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{I}{\tau_{\mathrm{syn}}} + A \cdot E^+ \cdot R^- \cdot \delta(t - t_{\mathrm{AP}}) \tag{2.3}$$

When the synapse is idle, the effective partition E decays to 0 with a time constant $\tau_{\text{facilitation}}$ and the recovered partition R decays to 1 with a time constant $\tau_{\text{depression}}$. Typical values range from hundreds of milliseconds to seconds of biological time [Regehr, 2012]. If an action potential is transmitted, a fraction U_{SE} is added to the effective partition E. We call the factor U_{SE} the utilization. The enlarged effective partition E^+ is then used to shrink the recovered partition since depression occurs. R gets smaller by an amount proportional to E^+. The actual synaptic output is given by the current I. Its amplitude at an action potential is given by a maximum amplitude A and the product of E and R, thus handling depression as well as facilitation. The current then decays back to 0 with a time constant τ_{syn}.

2.1.2 Spike Timing Dependent Plasticity

Another form of synaptic plasticity is the Spike Timing Dependent Plasticity (STDP). The weight of a synapse is typically increased when it stimulates a neuron shortly before it spikes, which we call causal correlation. It is typically decreased when the synapse sends input to a neuron shortly after it has spiked, called anticausal correlation [Sjöström and Gerstner, 2010]. This allows changing synaptic weights according to Hebbian theory [Hebb et al., 1949]:

> When an axon of cell A is near enough to excite cell B or repeatedly or consistently takes part in firing it, some growth or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

For implementing learning algorithms, STDP is a key feature: *Spike timing-dependent modifications, together with selective spread of synaptic changes, provide a set of cellular mechanisms that are likely to be important for the development and functioning of neural networks* [Bi and Poo, 2001]. Typical timescales for changes in synaptic weights based on correlation are in the order of 10 µs biological time [Bi and Poo, 1998].

2.2 The HICANN-DLS 3 ANNCORE

In order to include analog neurons on a silicon chip, the membrane capacity is represented by a capacitor. Synaptic inputs are integrated on this capacitor, while a leakage current pulls the voltage back to a resting potential. On HICANN-DLS, neurons also include an exponential term and adaption based on the Adaptive Exponential Integrate-and-Fire model (AdEx) [Brette and Gerstner, 2005].

The chip runs network emulations sped up by a factor of 10^3 due to intrinsic time constants. This means the typical timescales of milliseconds in biological domain translate to microseconds of actual time. In this thesis, all times are given in actual chiptime, if not specified otherwise.

On the HICANN-DLS 3 prototype ASIC that is used in this thesis, there are 32 neurons and 1024 synapses available. The synapses are arranged in a 32×32 array, a neuron is located at the bottom of each column. The neuron integrates the input of all synapses in its column. Each synapse can have an individual weight and address, which are both 6-bit values.

On the left side of the array, there are 16 synapse drivers, driving two rows of synapses each. A sketch of the setup is shown in figure 2.1. The synapse driver passes the target address along the line and processes STP. Each of the 64 addresses, which are available to configure synapse connections, can have an individual STP state, which is encoded in the length of the `dacen` pulse enabling the synapses. Drivers receive their data on 4 PArallel Debug Interface (PADI) buses.

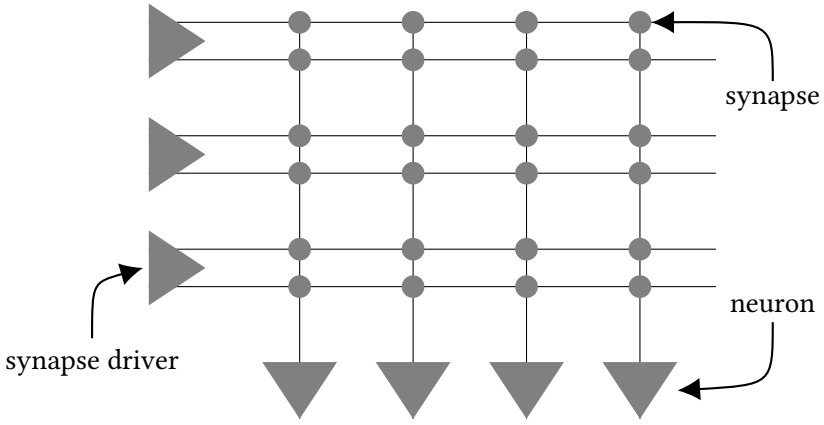


Figure 2.1: Sketch of the ANNCORE on HICANN-DLS 3. On the left, synapse drivers are located that can drive two rows of synapses each. At the bottom of synapse columns, neurons receive input of the synapses above them.

The synapse weights characterize the connections between neurons. When spiking, neurons send out a digital signal that can be fed back into one of the synapse drivers and thus travel through the synapse array, reaching other neurons. External input can be sent to the drivers as well, which will be the focus in this thesis.

Since both the short term (STP) and long term (STDP) plasticity mechanisms have to be able to change the synaptic input amplitudes received at a neuron, synapses multiply two analog parameters. The electric current flowing through the synapse is set by the weight, which can be modified by STDP. The time this current is flowing for is the length of the `dacen` pulse, modified by STP. The transferred charge, which is proportional to the received synaptic input, is then subject to both STP and STDP. The typical shape of a synaptic input over time is generated in the neuron.

2.3 STP implementation

Processing STP is done entirely in the synapse driver. Its output consists of the spike address and the `dacen` pulse. In case that STP is disabled, the `dacen` pulse reaches its maximum duration: with the intended chip clock of $250\,\mathrm{MHz}$ this means $4\,\mathrm{ns}$. If STP is enabled, the `dacen` pulse width changes with the level of depression and facilitation, respectively.

In order to store STP states for all addresses, individual capacitors are used. The voltage on the latter, V_{STP}, represents the state of neurotransmitters. With one capacitor each, we can store only one parameter. Since equations 2.1 and 2.2 require two parameters to process depression and facilitation, we have to decide which one to use. Equation 2.3 is processed at the input stage of the neuron.

Therefore, on the HICANN-DLS 3 chip, we can configure synapse drivers to use either depression or facilitation. The desired mode is set by inverting the output signal. For depression, an initially long pulse gets shorter and for facilitation, an initially short pulse gets longer.

The voltages V_{STP}, that are stored on the capacitors, recover exponentially towards a global voltage V_{recover} with a selectable time constant τ_{rec}. When an action potential is forwarded, the capacitor for the associated address gets connected to a net on voltage V_{charge}, thus the charge is

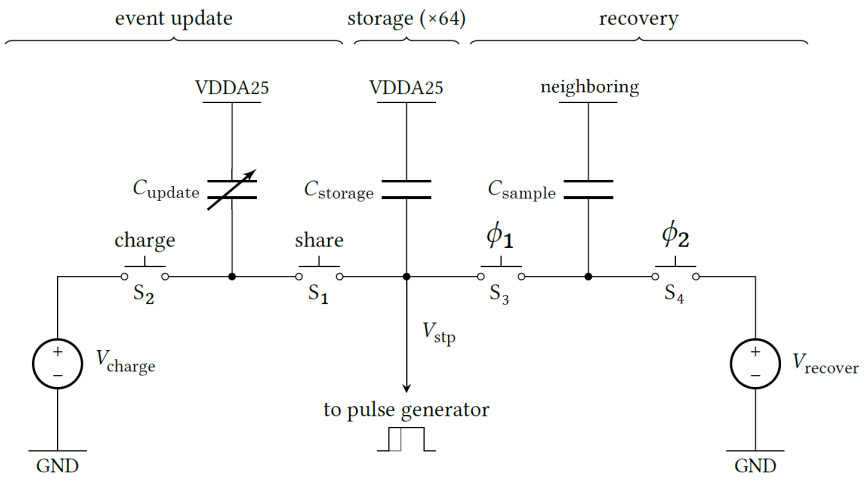


Figure 2.2: Schematic of the STP circuit implemented on HICANN-DLS 3. Charge stored on capacitor C_{storage} represents the current state of neurotransmitters. The left side of the circuit is used to update the state after handling an action potential, the right side is responsible for recovery and runs continuously. Figure adapted from [Billaudelle, 2017, figure 11].

shared with an update capacity C_{update}. The ratio of STP storage capacity C_{storage} and C_{update} is the utilization U_{SE}, since the remaining voltage $V_{\text{STP},i+1}$ will drop from $V_{\text{STP},i}$ according to

$$V_{\text{STP},i+1} = V_{\text{charge}} + (V_{\text{STP},i} - V_{\text{charge}}) \cdot \frac{C_{\text{storage}}}{C_{\text{storage}} + C_{\text{update}}}. \tag{2.4}$$

Since the recovery of neurotransmitters is an exponential process, it could be implemented using a resistor between V_{recover} and V_{charge}. However, considering the used capacities, a resistance in the range of gigaohms would be necessary to achieve the desired time constants ranging from 1 µs to 1000 µs. This would take up a lot of space on the chip. Instead, a small sample capacity C_{sample} is switched between V_{STP} and V_{recover}, transferring a minimal amount of charge each time it is switched. This switched-capacitor circuit forms a pseudo-resistor with a resistance $R = (C_{\text{sample}} \cdot f)^{-1}$, where f denotes the switching frequency. This allows an exponential convergence of V_{STP} towards V_{recover} with a time constant configurable by the frequency of switching. This frequency can be set using a global clock divider, referred to as `prescaler` and a local `recovery` setting, which is a counter-based switch: the sample capacity is switched after `recovery` global clock cycles have passed.

In figure 2.2, a schematic of the used circuits is shown. V_{STP} is stored on the storage capacity C_{storage}. The update capacity is C_{update}, which is configurable in 4-bit resolution via the `utilization` parameter. The recovery sample capacity is C_{sample}. Switch S_2 is normally closed and gets opened after an action potential was transmitted to update the STP state. Switch S_1 is normally open and closes during an update. Switches S_3 and S_4 are switched continuously to enable STP recovery. When an action potential is processed, the voltage on C_{storage} is used to create the `dacen` pulse.

To translate the voltages V_{STP} to durations of the dacen pulse, a comparator is used. It compares a linear voltage ramp with the given voltage V_{STP}. Depending on the STP mode, the pulse

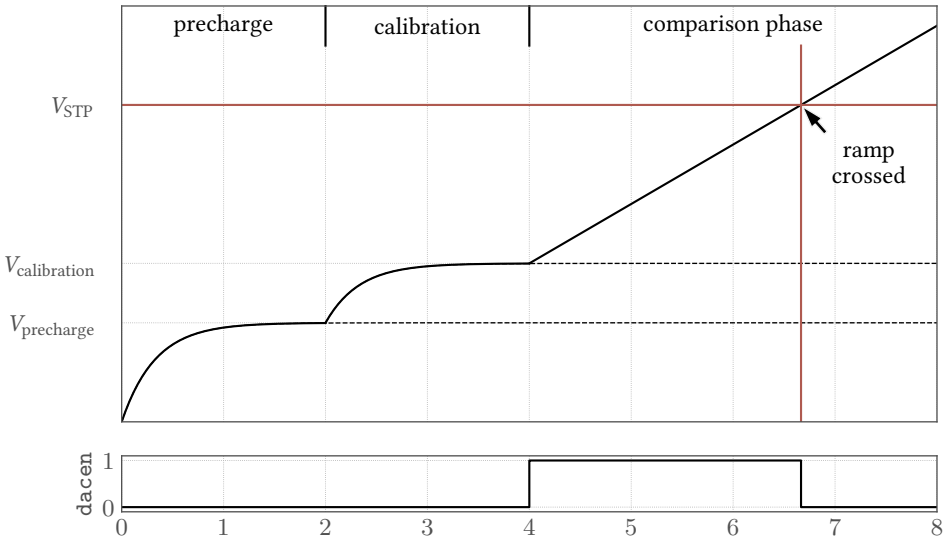


Figure 2.3: Sketch of the two voltages the comparator in the synapse driver is connected to. The capacitor that the ramp is produced on is plotted in black, the STP storage capacitor in red. The ramp capacitor is precharged during the first 2 ns, shares charge with the offset capacitors to correct individual offsets in the following 2 ns. Afterwards, the ramp is generated and the comparator toggles the pulse when it crosses V_{STP}.

is initially high or low and gets toggled once the ramp crosses V_{STP}. In order to map the full range of V_{STP} to the full range of `dacen` times, the voltage ramp should start near V_{charge} and end near V_{recover}.

Generating the ramp starts with precharging the ramp capacity to a global voltage V_{offset} during the first 2 ns of synapse driver activity. Next, an individual offset voltage is added to the ramp: a configurable capacitor with a resolution of 4 bits is charged to an correction potential V_{zero} and connected to the ramp capacitor. Sharing their charge, the voltage on the ramp capacitor rises depending on the selected capacity. The used parameter is called `offset`, it provides a constant offset to the ramp voltage before the ramp starts rising. This offset is added during another 2 ns. Once the correct initial voltage is reached, a constant current proportional to the parameter I_{ramp} starts flowing onto the ramp capacitor, increasing the voltage linearly. This happens within 4 ns, the maximum time the `dacen` signal can be active. Comparing V_{STP} with the voltage on the ramp capacitor toggles the pulse. The whole process is visualized in figure 2.3. Detailed information on the whole synapse driver implementation can be found in [Billaudelle, 2017].

For the STP model implemented on HICANN-DLS 3, this means each driver can use either the facilitating or depressing mode, but not both at the same time. Using short term depression with an inactive partition I and a recovered partition R, the amplitudes of synaptic input received at neurons are proportional to R, as shown in these equations [Schemmel et al., 2006]:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{I}{\tau_{\mathrm{rec}}} + U_{\mathrm{SE}} \cdot R \cdot \delta(t - t_{\mathrm{AP}}) \tag{2.5}$$

$$R + I = 1. \tag{2.6}$$

$$w \propto R \tag{2.7}$$

During depression, the weight w of the connection is proportional to R. For synapses configured in facilitating mode, the `dacen` pulse is inverted. This means equation 2.7 is replaced in favor of $w \propto I$. The weight gets proportional to I. Therefore, facilitation is based on the same differential equations with inverted roles.

Due to variations in the manufacturing process of the chip, the whole circuitry, especially the voltage comparator, are subject to mismatch. The STP circuit has to be calibrated. The `offset` parameters are used to shift the ramps of all drivers in a way that yields similar amplitudes for all drivers at similar STP states. This calibration of the offset parameter is the focus of this thesis, comparing two different readout mechanisms.

2.4 STDP implementation

The STDP feature relies on correlation measurements of every individual synapse. Synapses measure the time between the presynaptic spike and the spiking of their neuron, the postsynaptic spike. The measured time can also be negative if the neuron spikes before the synapse sends an input. This way, classification of causal and anticausal correlation is possible for all synapses. The time differences get weighted exponentially and can be read out as amplitudes of correlation. In detail, voltages that indicate causal and anticausal correlation are stored on two capacitors that can be read out using the Correlation Analog to Digital Converter (CADC). Complex algorithms can be used to tune synaptic weights in a large network based on correlation timings, since the Plasticity Processing Unit (PPU), an on-chip microprocessor, is able to access CADC results and change synaptic weights, using various operations.

In order to achieve symmetry between anticausal and causal results, most parts of the STDP circuit are shared between causal and anticausal measurements. Thus, using the same hardware, there can not be any mismatch. This means that every time synapses forward an action potential they start a causal measurement, and every time a neuron spikes, the synapses in its column start an anticausal measurement. Every end of a measurement means the start of the opposite one. A basic schematic of the used circuit is shown in figure 2.4. However, the actual circuit is much more complex, including the difference between anticausal and causal measurements.

The function of this simplified circuit is explained using a causal measurement that starts with the synapse transferring an action potential and ends when the neuron spikes. The storage capacitor C_{storage} is reset manually and then accumulates correlation amplitudes. Its voltage is read out by the CADC. Before starting a measurement, the voltage on C_{measure} is reset to V_{resmeas} by closing a switch S_{resmeas}. All other switches are initially open. A measurement starts when the synapse transfers an action potential. The measurement capacitor C_{measure} is charged to a voltage of 1.2 V. Closing S_{ramp}, it is then connected to ground via a transistor M_2 that limits the current flowing depending on its gate potential V_{ramp}. The characteristic time of exponentially weighted measurements can be set using this potential.

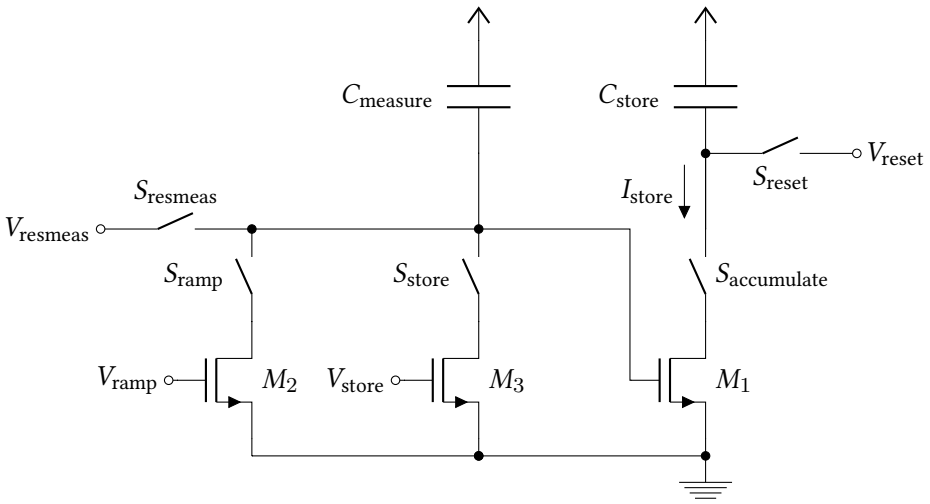


Figure 2.4: Schematic representing core elements of the STDP circuit. The actual implementation is shown in [Friedmann et al., 2017].

Once the neuron spikes, the measurement ends. Depending on the elapsed time, the charge on the measurement capacitor is now smaller. Opening S_{ramp} but closing S_{store} and $S_{\text{accumulate}}$, the discharge current from the measurement capacitor now flows through a different transistor M_3. On its gate, the voltage V_{store} allows setting the amplitude of the correlation signal. The potential on the capacitor C_{measure} is also connected to the gate of a transistor M_1 connected to the STDP storage capacitor C_{store}, where the final correlation signals are stored. Being a subthreshold voltage, the current I_{store} flowing from C_{store} is an exponential function of the voltage on the measurement capacitor C_{measure}. The linearly dropping voltage on the measurement capacitor limits the charge that flows off the storage capacitor. Using the fact that the integral of an exponential function is still an exponential function, this means that the charge flowing off the storage capacitor is proportional to the exponentially weighted time difference between start and end of the measurement.

Despite the circuit sharing the transistors for anticausal and causal measurements, previous experiments conducted on HICANN-DLS 2 have shown a strong asymmetry in amplitudes between causal and anticausal measurements. Those asymmetric results are presented in [Stöckel, 2017, figure 3.3], showing statistics of all synapses, and in [Wunderlich, 2016, figure 3.4], showing correlation measurements for a single synapse. In this thesis we will investigate whether the observed asymmetry in correlation amplitudes is still present using HICANN-DLS 3.

3 Characterization of Short Term Plasticity

3.1 Input and Output

3.1.1 Experiment setup

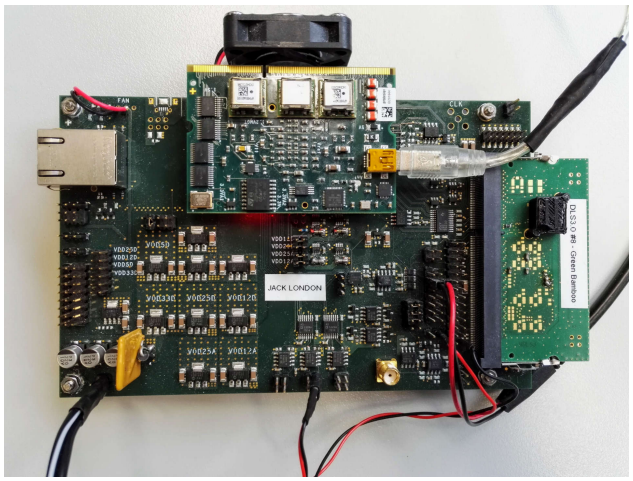
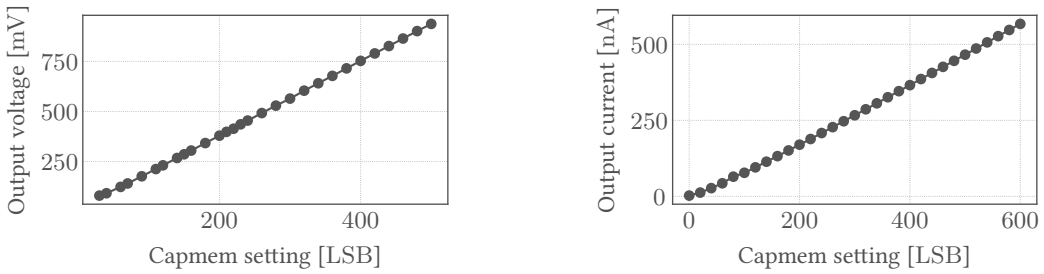


Figure 3.1: Photo of the HICANN-DLS 3 setup that was used during this thesis. It includes the baseboard, the Flyspi board (the board with the USB connection) and the HICANN-DLS 3 chip, which is located below the black cover at the right.

All experiments have been done on a HICANN-DLS 3 setup, as shown in figure 3.1. The host computer runs frickel-dls software, which contains C++ code, and the pydls bindings, which wrap the code in Python. This allows describing experiments using Python. It connects via USB to the Flyspi-board that contains an FPGA and memory. Experiments are stored in memory and controlled by the FPGA in real-time. There is a Analog-Digital-Converter (ADC) available, too. The DLS 3 chip itself is bonded onto a SODIMM module that is inserted into a socket on the baseboard.

The baseboard mainly provides power supply and generates voltages used inside the chip, using Digital-Analog-Converters (DACs). There are pin headers for analog readout of several signals, e.g. the synaptic input of neurons or their membrane potential. Since most signals are driven off the chip, an active probe is required to maintain high amplitudes during oscilloscope readout. Here, the *LeCroy WaveRunner HRO 64Zi* or the *LeCroy WaveSurfer 44Xs* oscilloscope together with *LeCroy ZS1000* active probes were used. To process data directly in Python, the oscilloscopes can be accessed via an Ethernet connection [Stradmann, personal communication, August 2017]. For most experiments however (unless stated otherwise), the Flyspi ADC was used, since acquiring data is a lot faster.

Unless stated otherwise, chip number 8 (DLS 3a) has been used for our measurements, together with baseboard “Jack London”.



(a) Capmem cell output voltage over digital setting.

(b) Capmem cell output current over digital setting.

Figure 3.2: Measuring dependence between capacitive memory configuration and output.

3.1.2 Capacitive memory

Most voltages and currents used on the chip are generated in the capacitive memory (capmem) [Hock et al., 2013]. There are voltage and current cells available, arranged in an array of 34 columns and 24 rows, the latter are split into 8 voltage and 16 current rows. All cells can be configured using digital values between 0 and 1023. This typically yields currents of $15\,\mathrm{nA}$ to $1000\,\mathrm{nA}$ and voltages of $0.2\,\mathrm{V}$ to $1.8\,\mathrm{V}$ [Aamir et al., 2018]. Using the capmem output pins and a *Keithley 2100 multimeter*, the actual voltages are measured for all used chips using cell (32, 0), which is a buffered cell, suited for readout, holding the STP V_{charge} voltage. In figure 3.2a, the data for chip 8 shows a linear function, as expected. This dataset is later used to map the digital settings to voltages, e.g. when plotting STP comparator ramps. This is why we are particularly interested in voltages below $1\,\mathrm{V}$.

The same measurement is done for current cells as well, shown in figure 3.2b. A *Keithley 2635B SYSTEM SourceMeter* was used here, applying $0.5\,\mathrm{V}$ of reverse voltage to cell (32, 8+1) on chip 8, which is the STDP *ibias_correlation_ramp* current. The graph shows a linear dependency as well. For low settings below $100\,\mathrm{LSB}$, the measurement was verified using the multimeter and oscilloscope (with their internal resistances), where the results fit perfectly to the ones shown here. Current cells are used to set several bias currents such as the synapse bias current, which scales configured synapse weights into amplitudes. Therefore, it is relevant during characterization.

3.1.3 Flyspi ADC

The signal path when measuring synaptic inputs of neurons consists of an on-chip source follower that drives the signal off the chip [Kiene, 2017], an amplifier on the baseboard, and the 12-bit $96\,\mathrm{MHz}$ ADC on the Flyspi board. The source follower shifts all voltages by a constant amount, so without characterizing it, absolute measurements of chip voltages are not suitable. However, relative changes in voltage are mapped accurately and can be read out on the ADC. Since we are mainly interested in amplitudes of synaptic input, this is no problem at all and we do not quantify the shift. Still, we need to characterize the ADC and find out how the digital output corresponds to the measured voltage. Using the *Keithley 2635B SYSTEM SourceMeter*, the digital value is measured in steps of $0.05\,\mathrm{V}$ input voltage. The amplifier is configured with the

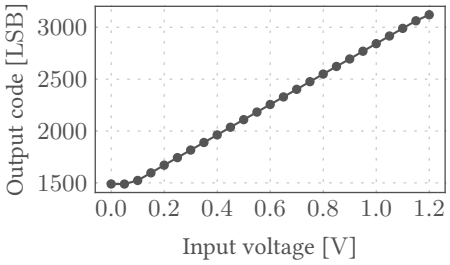


Figure 3.3: The onboard ADC digital output code as a function of the applied input voltage.

parameters `preamplify` 0 and `attenuation` 5. The settings are identical in all measurements, therefore this characterization is sufficient.

A plot of the data is shown in figure 3.3. A linear fit is used to convert digital signals from the ADC into analog voltages. When measuring voltage differences, only the slope of this linear function is relevant. Given the digital signal S that ranges from 0 to 4095, the original voltage V_0 is given by

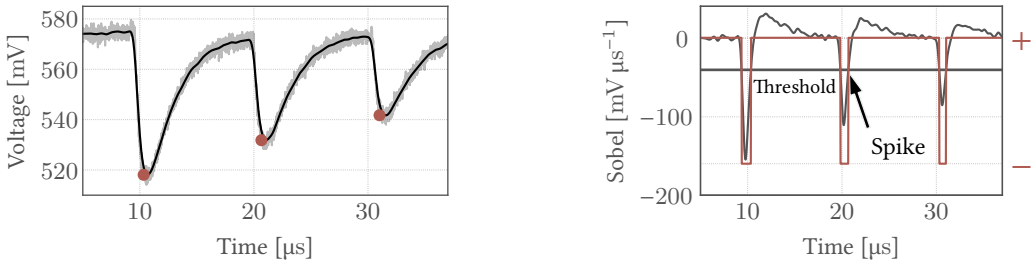
$$V_0 = 6.83 \times 10^{-4}\,\mathrm{V} \cdot S - 0.9399\,\mathrm{V}. \tag{3.1}$$

Therefore, 1 LSB corresponds to $0.683\,\mathrm{mV}$ of input voltage. The ADC readout is subject to gaussian noise with a standard deviation of 2.5 LSB, which equates to $1.7\,\mathrm{mV}$, if connected to the synaptic input of a neuron, which is the usual configuration in this thesis.

3.1.4 Evaluation of amplitudes

When an experiment is executed on the chip, the onboard ADC records a predefined number of samples. It is usually connected to the synaptic input of neurons, where, receiving a spike, the voltage drops proportional to its weight. The voltage drop is the spike amplitude, which has to be extracted from the ADC trace. This is done automatically using several methods from the python module *SciPy* [Jones et al., 2014]. The biggest problem concerning automatic evaluation of the voltage trace is the bad signal to noise ratio. Averaging multiple measurements to increase the signal to noise ratio is usually done but not always possible. In case a noise-related edge in the signal is considered a spike or a spike is considered noise, corrupted data is produced. This is especially problematic in longer experiments like the characterization of U_{SE}, where we are interested in decrease of amplitudes and not constant amplitudes alone.

At first, a gaussian filter is applied to the data, the width is set to 20 samples. The signal is now low-pass-filtered, see figure 3.4a. A sobel filter, resembling the first derivation, is used to find edges in the trace. A drop is considered a spike when its amplitude is $3\,\sigma$, where σ is the standard deviation of the sobel-filtered trace. In figure 3.4b, the sobel signal and spike threshold is plotted. Where the sign of the difference, plotted in red, jumps from -1 to 1, a spike is noted. In a small timeframe of $0.5\,\mu\mathrm{s}$ around the found spike time, the minimum of the low-pass-filtered trace is used to extract the amplitude. In figure 3.4a, the red dots indicate the time found with the sobel filter, the voltage is the minimum of the low-pass filtered trace. The baseline is given by the mean voltage in a timeframe of typically $2\,\mu\mathrm{s}$, shortly before the spike. All the given numbers, especially this timing of the pre-spike baseline voltage measurement, are changed according



(a) Original ADC samples (gray), low-pass-filtered trace (black) over time. Found spikes are marked as red dots.

(b) Sobel-filtered trace and spike detection threshold (black). The sign of the difference (red) yields spiketimes.

Figure 3.4: Extracting spike times and amplitudes from a voltage trace.

to the individual experiment requirements, allowing us to confidently use the acquired data. However, amplitudes are systematically under-estimated however by the low-pass filtering and due to the fact that the ongoing exponential decay back to the baseline voltage just before the spike is neglected. Therefore, low inter-spike-intervals make amplitude extraction difficult. For our experiments, this should not be of major concern, especially not for the calibration, where only comparing amplitudes of different drivers is necessary, so systematic effects do not matter.

This method to evaluate the ADC traces is CPU-intensive. The required time increases further with additional averaging that may be necessary to find spike times correctly. In this case, the same experiment is repeated several times and the mean voltage trace is used for amplitude extraction. This worked reliably. Consistency of spike timings across multiple identical experiments was no problem. Especially for experiments at low spike amplitudes, averaging was necessary.

3.2 Synaptic weights

Every synapse in the synapse array can have an individual weight ranging from 0 to 63. Received amplitudes at the neuron should be proportional to this weight. The synapse bias current scales the received amplitude at maximum weight, it is the proportionality constant. Since the synapse is basically a DAC, consisting of transistors, it underlies imperfections due to manufacturing tolerances as well. The mismatch between synapses as well as the dependency of amplitudes of the synaptic weights shall be investigated here.

In this experiment, all synapses in the first column are investigated, thus all amplitudes are measured at the same neuron. Synapses receive a spike from their driver one after another. Weights for the measured synapse are sweeped from 0 to 63, all other 1023 synapses are set to address and weight 0. The experiment is repeated for synapse bias current settings of 300 LSB, 400 LSB, and 500 LSB which equate to currents of 267 nA, 366 nA, and 466 nA. The latter setting on weight 63 will be used in the further part of this thesis since higher amplitudes yield better signal to noise ratios. The experiment has been repeated 10 times and mean values have been calculated in order to minimize statistical variations. Using the mechanism above to evaluate spike amplitudes, very small ones ($< 5\,\mathrm{mV}$) could not be measured. The results are shown

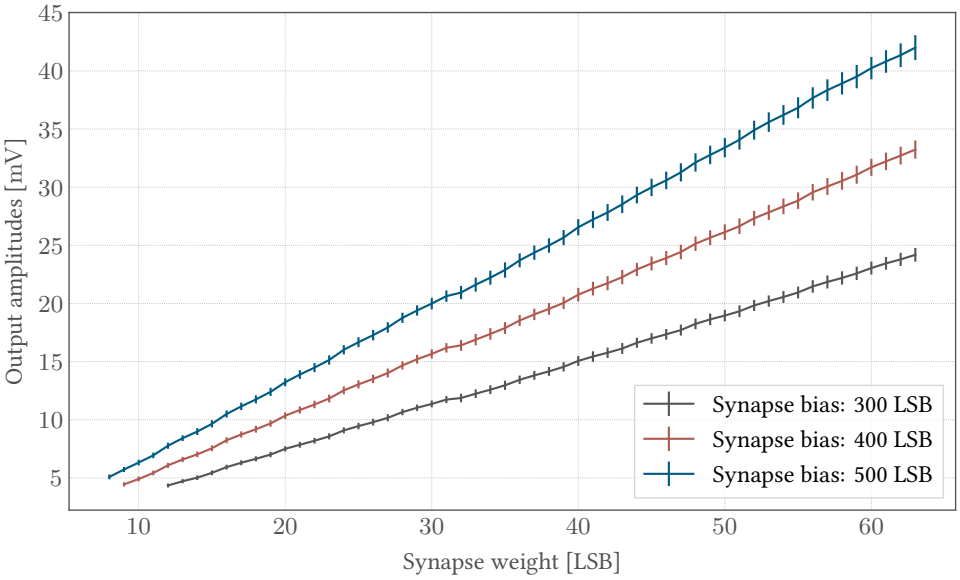


Figure 3.5: Characterization of synaptic weights. The received input amplitude is plotted over the digital weight setting for different bias currents. Error bars indicate the standard deviation between different synapses in a column.

in figure 3.5, error bars indicate the standard deviation between amplitudes of the 32 individual synapses in the column.

In the plot, amplitudes rise almost linearly with increasing weights, there are no major problems visible. The lines are not perfectly straight however, which means the synaptic conductances controlled by individual bits are not matched perfectly. This effect is best visible at the flat spot between weights 31 and 32, where all of the 6 LSB are switched. The conductance switched by the MSB has to be a little higher in order to increase synaptic amplitudes linearly. Slight mismatches like the one shown here must be expected. When updating weights with STDP and assuming a linear dependency of configured weights and amplitudes, the flat spot can pose a problem. Since there is no mechanism to calibrate synaptic amplitudes further, one has to accept the mismatch shown in figure 3.5. For the data point at 500 LSB current and a weight of 63, the observed relative standard deviation between the synapses is 2.55 %.

3.3 Recovery

3.3.1 Characterization reading out V_{STP} directly

While a synapse is idle, its neurotransmitters recover from depressed states until they are all available again. On HICANN-DLS 3, it is possible to read out the V_{STP} voltage indicating the current state of neurotransmitters directly, using the Flyspi ADC. This is only possible for address 15. In order to investigate other addresses later, a different measurement protocol will be used. Reading V_{STP}, the exponential recovery can simply be observed in the voltage trace. Since the sample rate of the ADC is known, we only need to fit an exponential growth to the trace. Its time constant is the recovery time constant τ_{rec}. An example is shown in figure 3.6, where V_{STP} is

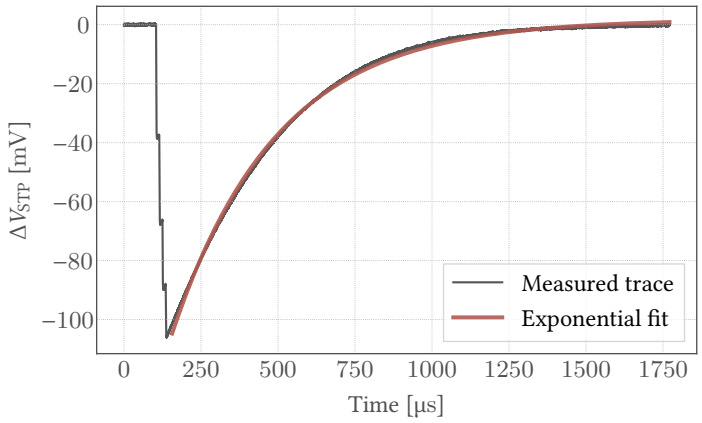


Figure 3.6: Example measurement showing the extraction of the recovery time constant τ_{rec}. Measured using driver 3 on chip 8, `prescaler` 5, `recovery` 0. In black, the mean of 100 V_{STP} traces is plotted, in red the exponential fit yielding the time constant.

plotted in black and the exponential fit yielding τ_{rec} in black. It was measured using driver 3 on chip 8. After depression with 4 spikes, the recovery is visible. It is not perfectly exponential since the source follower's output voltage is close to the supply voltage and therefore in saturation. This introduces some error affecting the time constants.

In order to configure the speed of recovery, two settings are available. First, the global recovery clock divider can be used to select faster or slower recovery on all drivers. From a clock signal that is intended to be 250 MHz, but in this setup is running at 200 MHz due to timing issues, the divider selects lower frequencies in steps of 2^{-k}. This setting k is referred to as `prescaler`, ranging from 0 LSB to 15 LSB [Billaudelle, personal communication, January 2017]. Additionally, there is a local property of the drivers. It is a 4-bit counter that counts downwards from 15 at every recovery clock cycle, compares it to the `recovery` setting and toggles the switches in the recovery circuit when the set value is reached. It then resets the counter. This means the recovery time constant grows linearly with smaller `recovery` settings. Please note that this way of implementation also affects the duty cycle between the two switches.

We sweep the `prescaler` setting from 2 (lowest time constant) to 8 (highest time constant) and the `recovery` setting from 0 (highest time constant) to 15 (lowest time constant). The data is plotted in figure 3.7 above the `recovery` setting. The different plotted lines are the `prescaler` settings. Error bars indicate the deviations of the 16 drivers. It was acquired on chip 3 using capmem parameters $V_{recover} = 300\,\text{LSB}$, $V_{charge} = 200\,\text{LSB}$ and a synapse driver bias current $I_{bias} = 500\,\text{LSB}$.

In the plot, we observe the linear dependency of τ_{rec} on the `recovery` setting. The same data is plotted on a logarithmic time axis in figure 3.8, where we can observe the influence of the `prescaler` more easily. Furthermore, the configurable recovery time range using these parameters can be estimated here: using the mean of all drivers, we can configure τ_{rec} from 2.38 µs to 2120 µs on chip 3. The error bars consist of readout noise accounting for roughly 2.4 % of deviations and a fixed pattern for the drivers, showing mismatch of 8.2 %. Thus, recovery time constants cover three orders of magnitude in the range of the global recovery clock that was used here.

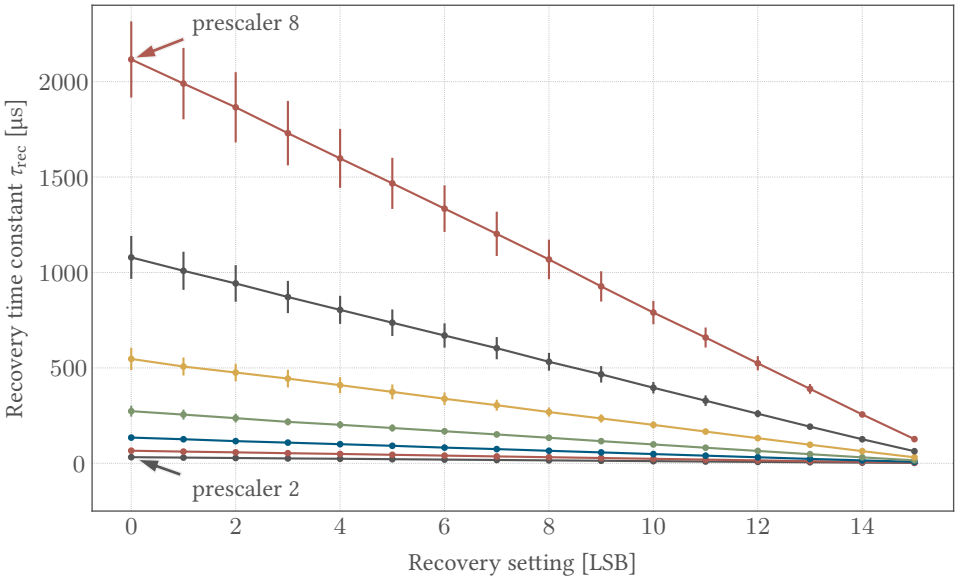


Figure 3.7: Available configuration range for the recovery time constant τ_{rec} on linear scale. Error bars indicate deviations between drivers. Measured on chip 3. The top line shows a `prescaler` setting of 8, for the bottom one it is set to 2.

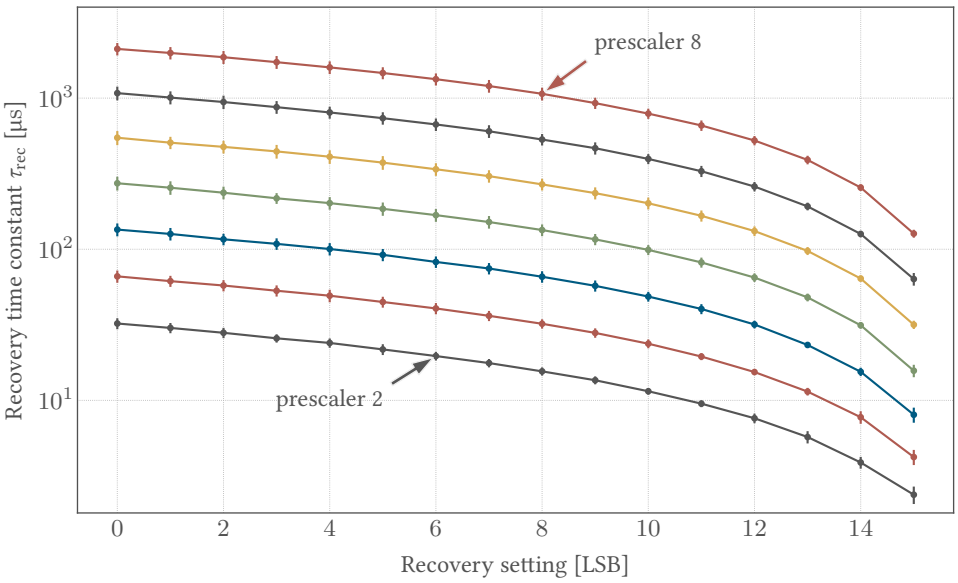


Figure 3.8: Available configuration range for the recovery time constant τ_{rec} on *logarithmic* scale. Error bars indicate deviations between drivers. Measured on chip 3. The top line shows a `prescaler` setting of 8, for the bottom one it is set to 2.

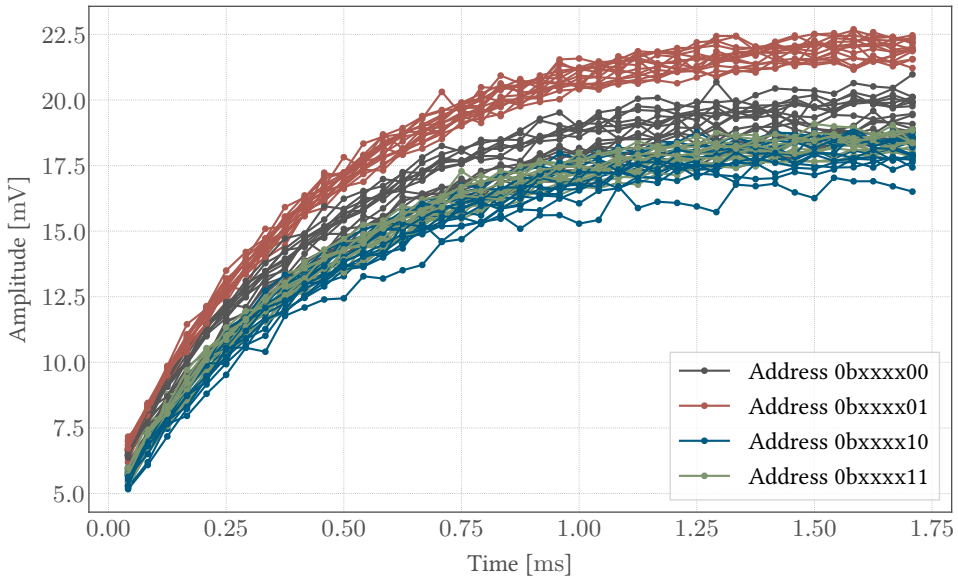


Figure 3.9: Recovery traces for all addresses of a driver. Measured on chip 3. Colors indicate the 2 LSB of the addresses. We clearly see that traces for addresses ending with 01 recover differently.

3.3.2 Amplitude-based analysis of recovery

Now, instead of only address 15, every address of a single driver is investigated. Since the direct readout of V_{STP} is not possible for the other addresses, simple traces of spike amplitudes are acquired. A burst of spikes is used to depress amplitudes. To characterize the recovery, after a pause, an additional recovery spike is sent. Depending on the duration of the pause, the amplitudes get larger. Of course it is not possible to have more than one recovery spike in one experiment run, since the processing of this spike utilizes neurotransmitters as well. Instead, the idle times are swept and amplitudes recorded for every run.

Plotting the results of many experiments in one figure, we can see an exponential increase in amplitudes - the recovery. This is shown in figure 3.9 for chip 3. Previous experiments have shown that for very slow settings of the recovery, leakage currents are visible, pulling V_{STP} towards varying potentials. In order to minimize their influence on this experiment, V_{charge} was set high and V_{recover} low [Weis, 2017]. The STP mode bit was then set to 0 instead of 1 to obtain the usual depressing configuration.

Amplitudes get depressed by forwarding 20 spikes with inter-spike-intervals of $\tau_{\text{ISI}} = 40\,\mu\text{s}$. Then, there is a pause of $t = k \cdot \tau_{\text{ISI}}$, with k ranging from 1 to 41. After the pause, the single recovery spike is sent. This way, we can plot a recovery trace consisting of 41 spike amplitudes. Using neuron 0 and driver 0, we choose these parameters: $V_{\text{charge}} = 300\,\text{LSB}$, $V_{\text{recover}} = 100\,\text{LSB}$, $V_{\text{offset}} = 50\,\text{LSB}$, $I_{\text{ramp}} = 620\,\text{LSB}$, $I_{\text{bias}} = 200\,\text{LSB}$, `mode` $= 0$, `recovery` $= 0$, `prescaler` $= 5$, $I_{\text{bias\_syn}} = 500\,\text{LSB}$, weight 63.

In the plot, we can see the recovery traces splitting into four groups, forming a symmetry of 4 addresses. While recovery time constants are very similar, the target voltage differs between the traces. This can be caused by neighboring nets in the layout: implementing a switched-capacitor recovery results in digital signals being close to the sample capacity. Additionally, the digital

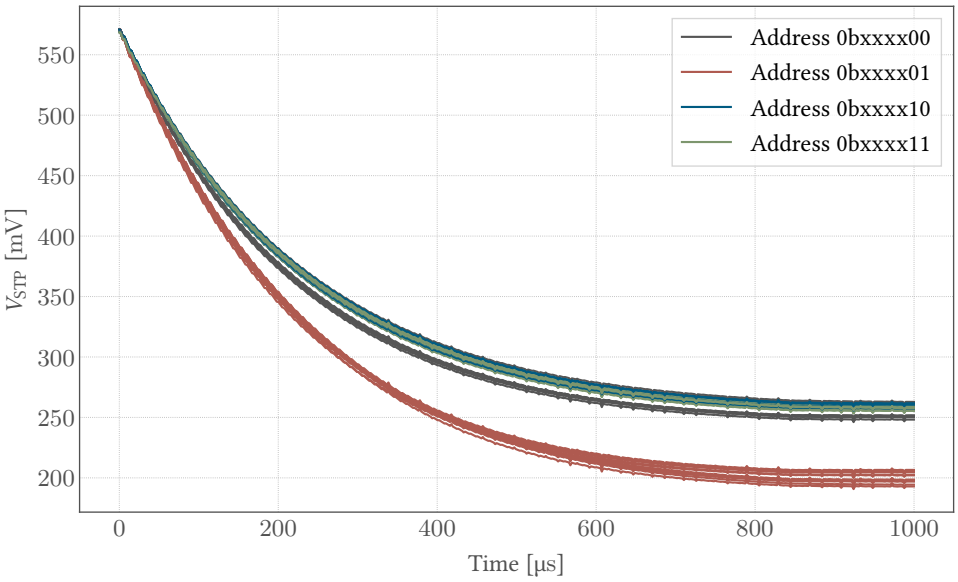


Figure 3.10: Results of a post-layout-simulation of the recovery circuit (voltages inverted). The simulation shows the same behaviour as the experiment (figure 3.9).

signals gating the switches have unequal duty cycles, what may contribute to the observation. In the layout, four capacities each are placed in a repetitive pattern [Billaudelle, 2017, fig. 21-22]. Thus, the symmetry of the observed effect fits to the symmetry in the layout.

The effect is well visible in the conditions used here. We set up a simulation using identical parameters and we got similar results: the recovery traces split similarly into four groups, which is shown in figure 3.10. The ability to reproduce the problem in a simulation is always appreciated. The chip designers are able to change the layout and observe the effects.

3.4 Utilization of Synaptic Efficacy

In a synapse with short term depression, the amplitudes of subsequent spikes are lowered. Since recovery is exponential, amplitudes drop exponentially towards a steady value when inter-spike-intervals are constant. There, the amount of neurotransmitters released at a spike is equal to the recovered neurotransmitters in between two spikes, thus amplitudes are constant. The time constant of the exponential amplitude decay lets us calculate the utilization U_{SE}. The lower U_{SE}, the more spikes it takes to approach the steady state.

3.4.1 Characterization on chip 8

In an experiment, recovery is turned off and the `utilization` parameter is varied from 0 to 15. Additionally, the `enshare` parameter can be set false, which decreases U_{SE} further. There, charge in the STP update circuit is only shared with parasitic capacities of the routing lines and no additional capacitor C_{update} (see figure 2.2). 30 spikes in intervals of 40 µs are evaluated for one driver after another and read out using neuron 12 on chip 8. In order to use a broader

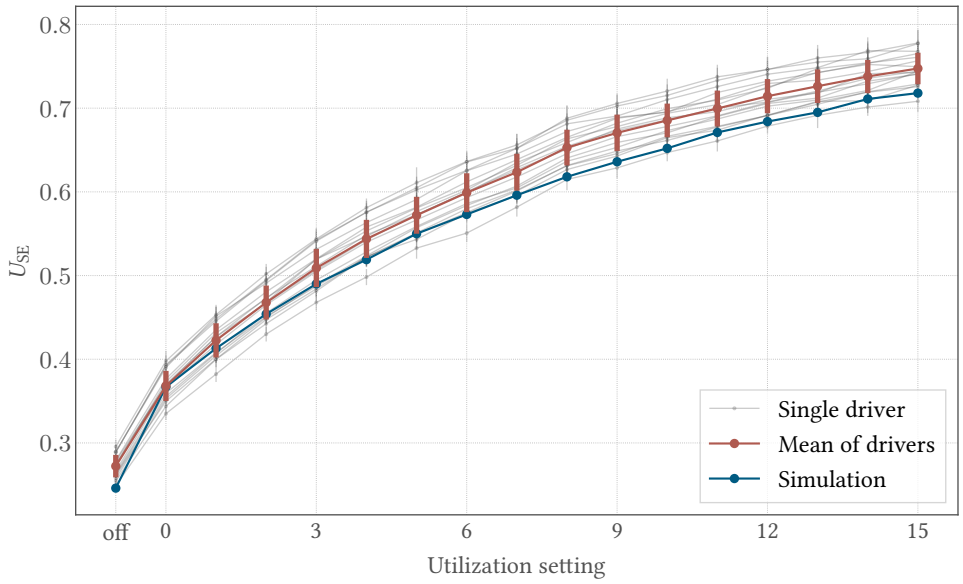


Figure 3.11: Available configuration range for the Utilization of Synaptic Efficacy U_{SE}. Red: Mean of all drivers, error bars indicate deviations between drivers. The individual driver's results are plotted in light gray in the background, their errors indicate statistical variations. Simulation results [Billaudelle, 2017] are plotted in blue.

voltage range and still retrieve usable data, an amplitude-based offset calibration is loaded, which will be explained in chapter 4. The acquired amplitude traces look like these in figure 4.10, the configuration is not much different: the inter-spike-interval is increased to read out amplitudes more precisely and recovery is off, thus the baselines are lower here. An exponential function is fitted to the spike amplitudes in order to find the "time constant" of depression, τ_{dep}.

In order to calculate the U_{SE} value, we use equation 3.2. Let the inter-spike-interval be t_{ISI} and the time constant of exponential amplitude depression mentioned above be τ_{dep}. In general, the utilization also depends on the recovery time constant τ_{rec}. From the deduction in appendix A follows that it is given by

$$U_{SE} = 1 - \frac{e^{-t_{ISI}/\tau_{dep}}}{e^{-t_{ISI}/\tau_{rec}}}. \tag{3.2}$$

Since during this experiment the recovery is turned off, only leakage currents can flow onto the STP capacitor. The time constant of this leak-recovery is some orders of magnitude longer [Weis, 2017], so it can be neglected here. Using an infinite recovery time constant, the denominator of equation 3.2 is therefore simply 1. Without the recovery, the V_{STP} voltage might be higher than desired and amplitudes may be saturated. The target voltage of leak-recovery is not necessarily $V_{recover}$. Although therefore it would be best to exclude the first spike from the exponential fits, it is almost impossible to fit an exponential decay to the remaining trace and extract a reliable time constant from there. Especially for high utilizations, the trace following the second spike shows a bad signal-to-noise ratio due to the small amplitudes. The results did not change significantly when including the first spike, however, statistical variances are reduced drastically. Therefore, we include the first spike in all fits.

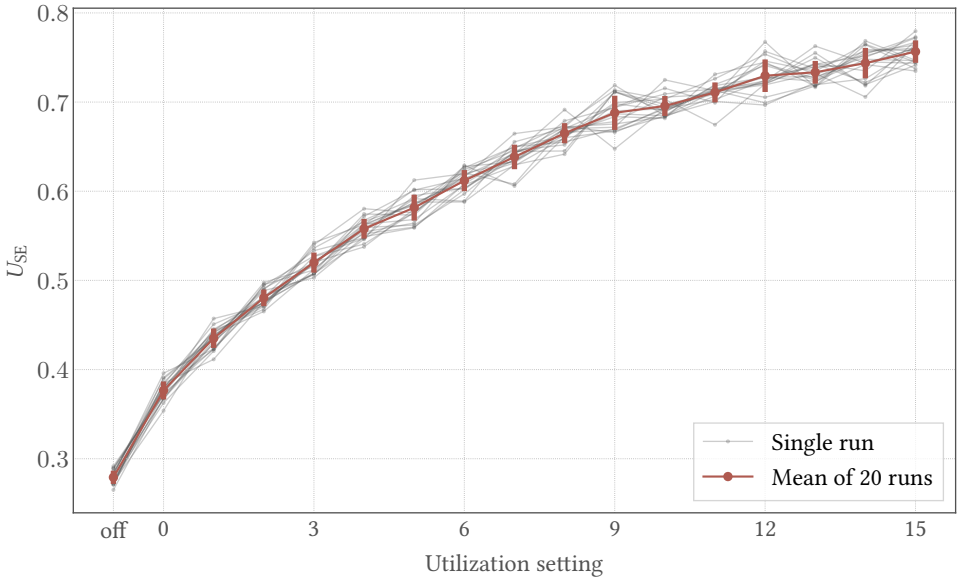


Figure 3.12: Results calculating U_{SE} for driver 0 only. Red: Averages of all 20 measurements, error bars indicate the standard deviation between them. The results of every run are plotted in light gray in the background.

Plotting the results over the digital settings yields the plot in figure 3.11. The thick red line shows the average of all drivers over 20 measurements, error bars indicate the standard deviation between the drivers. It fits well to the expectations in [Billaudelle, 2017, figure 27b], these simulation results are plotted in blue in the same figure. Parasitic capacities have been slightly underestimated in the simulation. The curves in a light gray in the background correspond to individual drivers' results, averaged over 20 measurements as well. Their error bars show the statistical variations between the 20 runs. The slight underestimation of amplitudes introduced by the low-pass filtering during their evaluation affects high amplitudes stronger, but should only yield a multiplicative error that does not affect the time constant of amplitude depression. The long inter-spike-interval used here minimizes the error resulting from reading out the baseline voltage, so this result should be precise. The data was measured in depressing mode, but the factor U_{SE} stays identical for facilitation since the `dacen` pulse just gets inverted (see figure 2.3).

An in-depth view of a single driver is shown in figure 3.12, including the variations between the individual runs. The thick red line in the foreground shows the mean utilization values for driver 0 only, it is one of the gray lines in figure 3.11. Here, the lines in the background show the individual measurements. This is the data the average was computed from. We can see that the trial-to-trial variations are very notable and averaging of multiple measurements has to be done in order to characterize the function properly. The mean relative statistical variations between multiple runs across all drivers are 2.2 %.

For chip 8, using the average of all drivers, the available configuration range for U_{SE} is from 0.272 to 0.747. The mean relative mismatch between the drivers across all utilization settings is 3.7 %. This, especially with the huge range of available recovery times, should be a usable configuration environment for biological experiments using STP. However, the mismatch between drivers is higher than expected and will be investigated further.

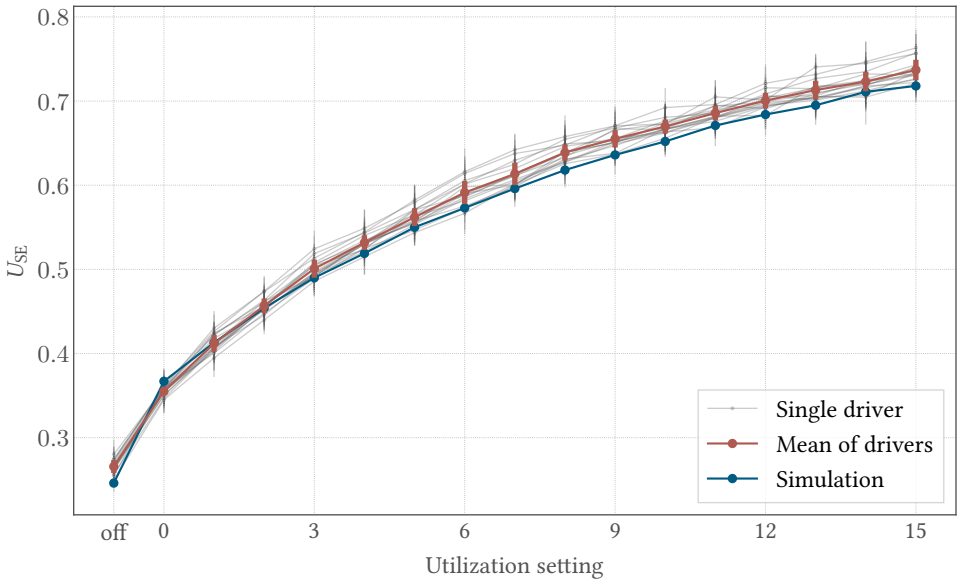


Figure 3.13: Available configuration range for the utilization U_{SE}. Red: Mean of all drivers, error bars indicate deviations between drivers. Measured on *chip 10*. The measurements of single drivers (background lines) show less variations than on chip 8. Simulation results [Billaudelle, 2017] are plotted in blue.

3.4.2 Verification on chip 10

The results will now be verified using chip 10 (DLS 3b), which originated from a different manufacturing run. An amplitude-based offset calibration has been done prior to this experiment as well. The used voltages were adopted for the different chip. The plot showing the distribution of synapse drivers' U_{SE} parameters over the settings is displayed in figure 3.13. There, using the mean of all drivers, the available range for U_{SE} is from 0.266 to 0.737. The observed deviation between drivers on chip 10 is $1.8\,\%$. The statistical variations are $3.8\,\%$. With the deviation of the results between the chips being at most $1.5\,\%$, the results lay well in each others $1\,\sigma$ area of driver mismatch and therefore show no significant deviations.

For chip 10, we can see that the deviation between drivers is lower. While it is possible that the two chips actually differ in that manner, we want to investigate if other factors matter. The recovery is turned off and its time constant is assumed to be infinite, however, leakage currents flow onto the STP capacitors. The time constant and even the target voltage of this leakage can be very different, thus possibly yielding higher or lower utilizations. The fact that the relative mismatch is mostly constant across the configurable range of U_{SE} means that this mismatch is probably not introduced by the driver but rather during the experiment. When caused by the capacities in the driver, the mismatch would depend on the utilization settings presented on the horizontal axis, not the resulting parameters on the vertical axis. Since the function is not linear, we can differ those influences. Keep in mind that the experiment includes a noisy readout, passing many stages on the chip such as the STP comparator, the synapses, and source followers. The evaluation of ADC traces is certainly not perfect as well. We want to know whether the capacities C_{STP} and C_{update} are subject to mismatch, which we don't expect to that extent. We

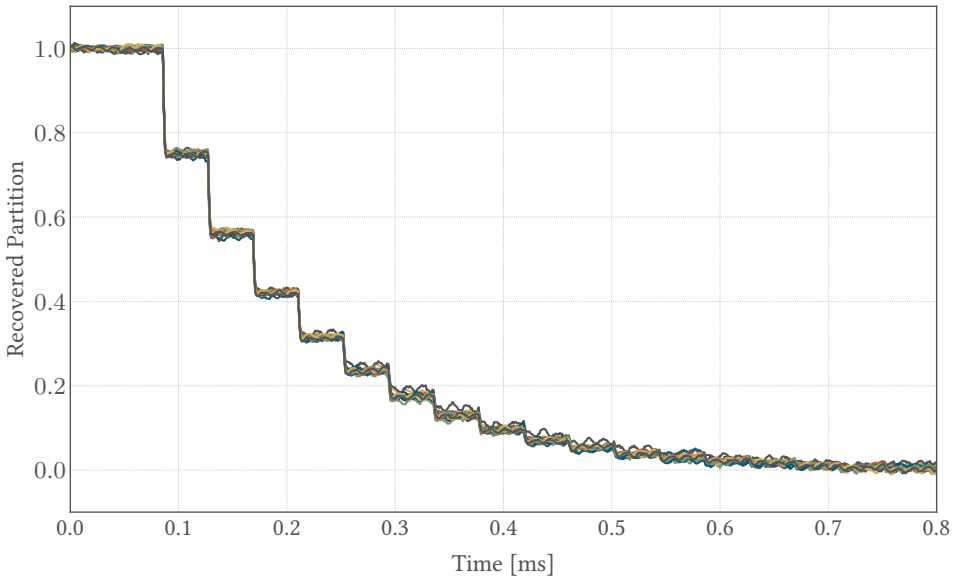


Figure 3.14: Traces reading out the V_{STP} voltage directly for all drivers on chip 8. The traces are low-pass-filtered and normalized between 0 and 1 and therefore represent the recovered partition R of neurotransmitters. There are no visible systematic deviations between the drivers. The utilization setting is the lowest possible.

expect the observed mismatch to be added by the readout chain or the measurement protocol including amplitude extraction and leakage currents.

To investigate the mismatch between drivers better, we change the readout: the voltage V_{STP} is recorded at the minimal utilization setting. There, we see drops that are proportional to U_{SE}, as shown in equation 2.4. On chip 8, we record V_{STP} for every driver during depression. The voltages differ between drivers, firstly because of the leakage currents when recovery is turned off, and secondly because the buffers in the readout chain differ. This is why we normalize all voltages into a range between 0 and 1. In this configuration, this can be interpreted as the fraction of neurotransmitters in the recovered partition R.

When drivers are plotted in individual colors, we get the plot shown in figure 3.14. The steps visible there happen when a spike is forwarded, since the recovered partition R then gets smaller. Mainly, we can see almost no deviations between the drivers, only some statistical noise. Doing the same for chip 10 yields a very similar plot. Extracting the U_{SE} parameter from this voltage trace can be done the same way as before, fitting an exponential decay as an envelope to the step function. Doing so, we get a value $U_{\mathrm{SE}} = 0.249 \pm 1.9\,\%$ for the lowest utilization on chip 8. The error given here is the sum of statistical variations and systematic variations. Those can not be distinguished after executing just a single experiment run. The systematic mismatch present between the drivers therefore is much lower using this readout method. This proves that the high variations observed above are mostly introduced by the experiment and not by the drivers.

We conclude that STP should be well usable with a configuration range for the Utilization of Synaptic Efficacy of approximately 0.27 to 0.74. The mismatch between drivers is hard to distinguish with the readout methods used, but it is less than 2 % based on the observations on chips 8 and 10.

4 Calibration of Short Term Plasticity

The STP circuitry, explained in section 2.3, includes a comparator which is subject to mismatch. Comparing the output of different synapse drivers, the amplitudes are shifted by a constant amount. A calibration parameter, `offset`, adds an opposite offset, counteracting the shift. A calibration algorithm should effectively reduce the mismatch between drivers, require a short runtime and be highly scalable. This is especially important since the upcoming HICANN-X chip will contain 256 synapse drivers instead of the 16 drivers on the prototype system. The number of neurons will be increased from 32 to 512 [Billaudelle, personal communication, February 2018].

4.1 Calibration algorithm

The algorithm that will be used to find the `offset` parameter that minimizes the deviation between drivers is based on a binary search. The calibration consists of multiple runs. Each time, the amplitudes of all drivers are measured. Comparing them to the mean amplitude of all drivers yields the change of the `offset` parameter: the setting is changed in a way that shifts amplitudes towards the mean. Since there are 4 configuration bits available, the binary search takes exactly 4 runs, setting one bit in each run. Since the dependency of amplitudes on the `offset` setting is not perfectly linear, we extended the binary search algorithm by an additional shift. Therefore, at least 3 consecutive settings are tested: the result from the binary search and both neighboring settings. As the final result, the algorithm chooses settings that minimize the spread around the mean amplitude at the end of the binary search. To visualize the algorithm, the individual drivers' results are later plotted over the course of calibration (see figure 4.4, which shows spike rates instead of amplitudes already).

In total, 6 runs are required for calibration. Since, in this thesis, the standard deviation of amplitudes after calibration is the most important parameter, in a seventh run the calibrated offsets are set and evaluated. Using the Flyspi ADC to measure the amplitudes, this calibration algorithm takes about 6 minutes to run since it checks all drivers sequentially [Weis, 2017]. Scaling this up to a larger chip the runtime grows linearly, limited by the number of parallel readout channels. Therefore, we test whether an alternative readout method using the neurons exclusively is possible and yields adequate results.

4.2 Neuron spike counter readout

The synapse drivers' output, the `dacen` pulse, gates a current in the synapses, which is sourced from a capacitor in the neuron synaptic input, decreasing the voltage V_{syn} on it. The voltage drop is proportional to the width of the `dacen` pulse. V_{syn} gets pulled up again to a baseline potential $V_{\text{syn},0}$ with a configurable time constant. The usual readout is sampling V_{syn} using the ADC, what we call amplitude readout. By design of the neuron, the mean current flowing onto the membrane is proportional to the with of the `dacen` pulse. If the membrane potential reaches a threshold, the neuron spikes. Thus, the rate of spiking can be a measured variable for the received input as well.

4.2.1 Spike rate measurement

Every neuron has an individual 8-bit spike counter that can be reset and read out externally [Kiene, 2017]. It is also possible to access data from the PPU. It is desirable to measure synaptic input amplitudes as precisely as possible using this counter. A linear dependency between spike rate and input amplitude is desired, but in theory any strictly monotonic dependency suffices calibration purposes. In the leaky integrate-and-fire model the membrane potential on a neuron which receives a constant synaptic input behaves like charging a capacitor over a resistor: the voltage rises, but saturates exponentially at a target voltage. The voltage stays constant when the leakage current and the input current are equal. Using discrete spikes instead of a constant input current, the neuron's membrane potential rises step-wise. Since the input is given regularly, the envelope of the potential is still a bounded growth.

However, this is not what we desire. A neuron receiving a small input may even not be able to cross the spiking threshold, and if it does, the events are extremely rare. In order to make the membrane potential rise linearly, we disable the leak term. Only the synaptic input is now integrated on the membrane capacitor, the membrane voltage therefore rises linearly. When it crosses the spike threshold voltage, the event counter increases and the membrane potential is set back to a reset voltage. So in theory, the number of spikes recorded in a given time interval is a linear function of the synaptic input amplitudes.

An experiment will consist of several bursts of spikes, where inter-spike-intervals are lower than when an amplitude measurement is desired. A very linearly rising membrane potential is ideal, since this will provide a better resolution in spike rate readout than a function that rises step by step. Fast spiking is preferred here. For the ADC measurement, the pre-spike voltage must be read correctly. For fast spiking, this voltage may still increase towards its baseline. Slow spiking with high weights is ideal there.

The event counter only has 8 bits and an overflow indicator, which limits the maximum size of a single burst. Between the bursts, the counter is read out and reset. Shortly before a new burst begins, the membrane potential is reset as well, since it may be floating without the leakage current. We use multiple bursts to do more averaging in-place and reduce statistical variations.

If not specified otherwise, we will send 40 bursts consisting of 300 spikes each to the drivers. A spike is sent every $10\,\mu s$, between the bursts is a pause of $500\,\mu s$. Drivers activate both their top and bottom attached synapse row. This yields higher amplitudes and decreases the error introduced by synapse mismatch. In order to not include possible differences in utilization or recovery in the calibration, we disable recovery and set the synapse drivers to `renewing`. This means that U_{SE} is set to 1 by connecting the STP storage capacitor $C_{storage}$ directly with the V_{charge} supply. During amplitude measurements using the ADC, the first 10 spikes of a burst are discarded since they include the depression phase, which may differ between drivers. The target driver is selected using the `enreceiver` setting, which makes the other drivers discard any input they receive. Using chip 8, the voltages $V_{recover}$ and V_{charge} are set to 210 LSB and 170 LSB, respectively. The synapse bias current is set to 500 LSB, the weight is 63. Concerning the STP calibration, as explained in section 2.3, the ramp is precharged to a capmem voltage of $V_{offset} = 50\,\text{LSB}$, the offset calibration capacitors are precharged to $V_{zero} = 300\,\text{LSB}$ and the current flowing onto the ramp capacitor is set to $I_{ramp} = 600\,\text{LSB}$.

4.2.2 Parameter search for synapse reference voltage

With the leakage current disabled, the membrane potential reacts sensitively to changes in the synaptic input reference voltage $V_{\text{syn},0}$. This voltage is the reference for the transconductance amplifier which converts the synaptic input voltage, V_{syn}, into a current flowing onto the membrane capacitor. The reference voltage has to match the idle synaptic input voltage very well. The voltage varies across different neurons, but also with temperature or other factors. If it is set poorly, a constant current is flowing onto or off the membrane, this means the spike rate at a given input can appear both over- or underestimated.

Watching the membrane potential on the oscilloscope, a manual guess of the right parameters was taken for every even neuron. For the chosen parameter, the membrane voltage seems to float, which means that rare spiking can occur. Also, the membrane voltage should not constantly be very low, since this could indicate a negative input. Of the 16 available even neurons, 5 were not used for various reasons. Therefore, the spike rate readout is tested using 11 neurons. Based on these manually adjusted values, an automatic parameter search is run before every experiment that includes spike rate readout. This ensures that the rates are neither too high nor too low. The script uses the setting for $V_{\text{syn},0}$ that is 1 LSB below regular spiking. The initial parameter is varied in steps of 1 LSB until the "threshold" for regular spiking is found. Although the script takes only a few seconds to run, its runtime is not included in the later presented runtime measurements.

Another setting that is individually set for all neurons is the `ibias_syn_gm_exc` parameter, which allows scaling of the synaptic input voltages. Here, the parameter is set so that the spike rates of all 11 neurons are approximately equal. This means that when averaging the spike rates, they contribute equally to the result. It is not important to have precise settings here, since all neurons will show the same dependency on changes in the synaptic amplitudes. This is why for this parameter, no additional automatic calibration was implemented, only the manually set values are used. Used settings range from 90 LSB to 160 LSB.

4.2.3 Feasibility analysis

Besides the faster runtime on this chip, the big advantage of a spike rate-based measurement is that it can use many neurons in parallel to acquire data for multiple drivers at the same time. Since the larger HICANN-X chip will not only hold more synapse drivers but also more neurons, the whole chip could be calibrated in the same time as the current prototype. In order to determine whether the spike rate readout method is suitable, the dependency of amplitudes and spike rates on the `offset` parameter is measured. This is done by reading out spike rates of the 11 available neurons via the counter, as presented above, and in parallel sampling the synaptic input voltage of neuron #12 with the ADC. The trace gets evaluated as usual in order to find spike amplitudes.

For every driver, amplitudes and spike rates are plotted as a function of the `offset` setting. The plot is shown in figure 4.1. Error bars indicate statistical deviations of the data over 20 runs. It is clearly visible that the black data points that were acquired with spike rates and the red data points that are ADC amplitudes show similar behaviour. However, the error bars of red (amplitude) data points are invisibly small. For the black (spike rate) data points, they are fairly large, even large enough that neighboring offset settings can not clearly be distinguished by using just one run of spike rate measurement. This is already a sign that calibration using spike rates may be less accurate.

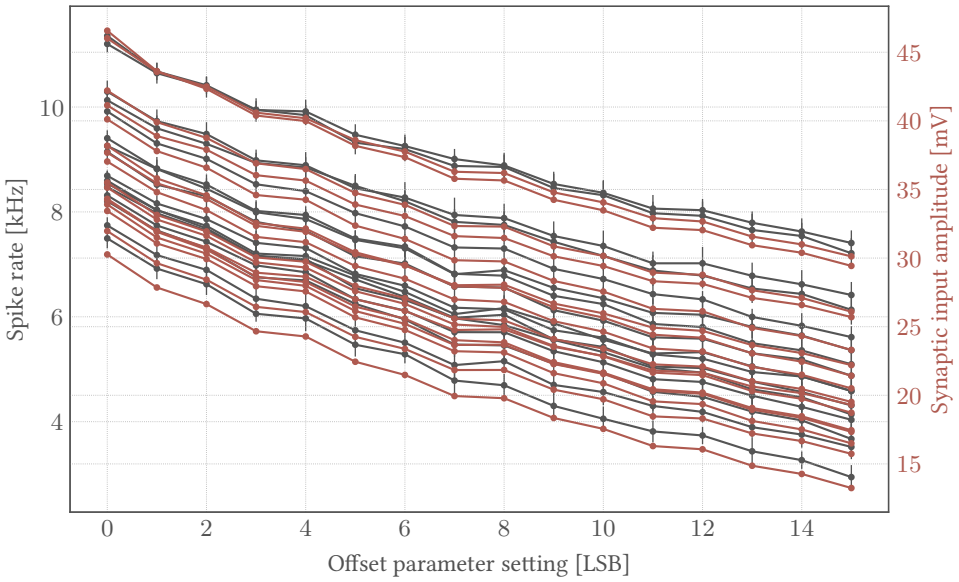


Figure 4.1: Comparison of spike rate (black) and amplitude (red) measurement as a function of the `offset` parameter. One line per driver, respectively. Error bars indicate statistical deviations across 20 runs. Measured on chip 8 using 11 neurons for spike rate readout and neuron 12 for amplitudes.

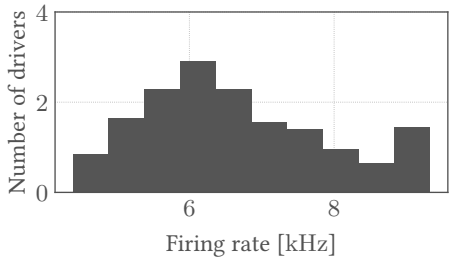
We can see that a calibration is possible, since all drivers reach identical amplitudes or spike rates. All plotted lines, representing drivers, intercept a common horizontal line. The voltages V_{zero} and V_{offset} are chosen well. Increasing V_{zero} further would enlarge the steps between `offset` settings, yielding higher mismatch after calibration.

The plot also justifies the calibration method presented above, checking both data points left and right of the result of the binary search. Since the amplitude dependency on the setting is not linear, this is an important step. The more bits switch, the larger is the observed nonlinearity. Between settings 7 and 8, where all of the 4 bits are changed, a flat spot is visible, similar to the synapse amplitude characterization. This is again caused by imperfections during chip manufacturing.

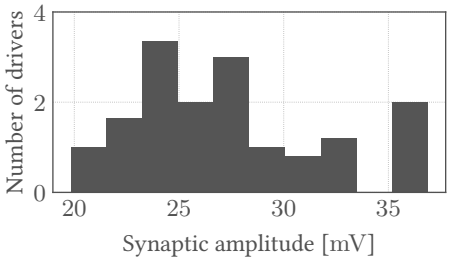
A calibration using amplitudes can not use parallel readout. This means an increase of runtime from 6 min on DLS 3 to roughly 1.5 h on HICANN-X. Using spike rates, the calibration could be done in about a minute if sufficient neurons are available, which is the case on HICANN-X. We conclude that a runtime advantage of a factor 100 justifies slightly worse results of the calibration, as long as everything stays usable. Thus, we keep working on the spike rate-based calibration, in order to investigate how much the standard deviation of drivers' amplitudes really is and how much time is required.

4.3 Spike rate-based calibration results

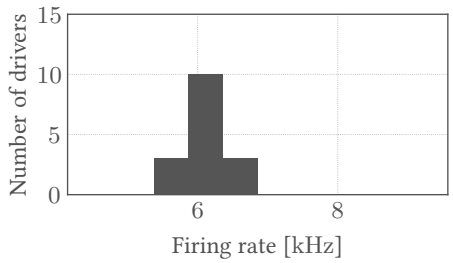
Combining the newly implemented readout mechanism with the calibration algorithm used before, a spike rate-based calibration is available. In this section, we directly compare the results of amplitude-based and spike rate-based calibration.



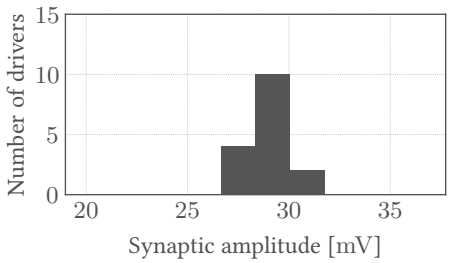
(a) Uncalibrated spike rate



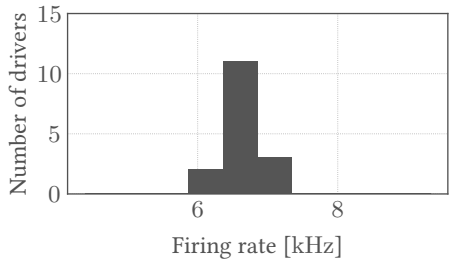
(b) Uncalibrated amplitude



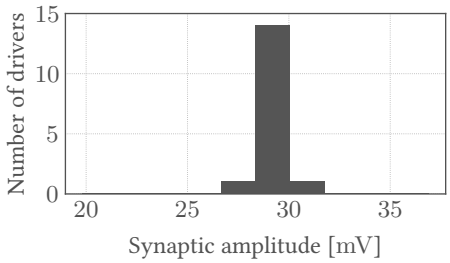
(c) Spike rate-calibrated spike rate



(d) Spike rate-calibrated amplitude

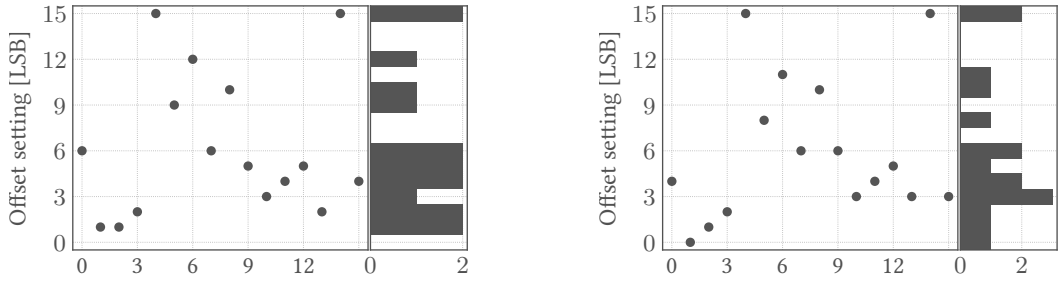


(e) Amplitude-calibrated spike rate



(f) Amplitude-calibrated amplitude

Figure 4.2: Histograms showing the distribution of spike rates and amplitudes of all drivers. The first row shows the original uncalibrated state, a calibration based on spike rate readout is displayed in the middle, a calibration based on amplitude readout using the ADC is at the bottom.



(a) After calibration using spike rates

(b) After calibration using amplitudes

Figure 4.3: For every driver, the `offset` parameter resulting from the calibration is marked as a dot. A small histogram on the right of each plot shows the distribution of used settings. While the settings are alike for both calibration methods, they are not identical.

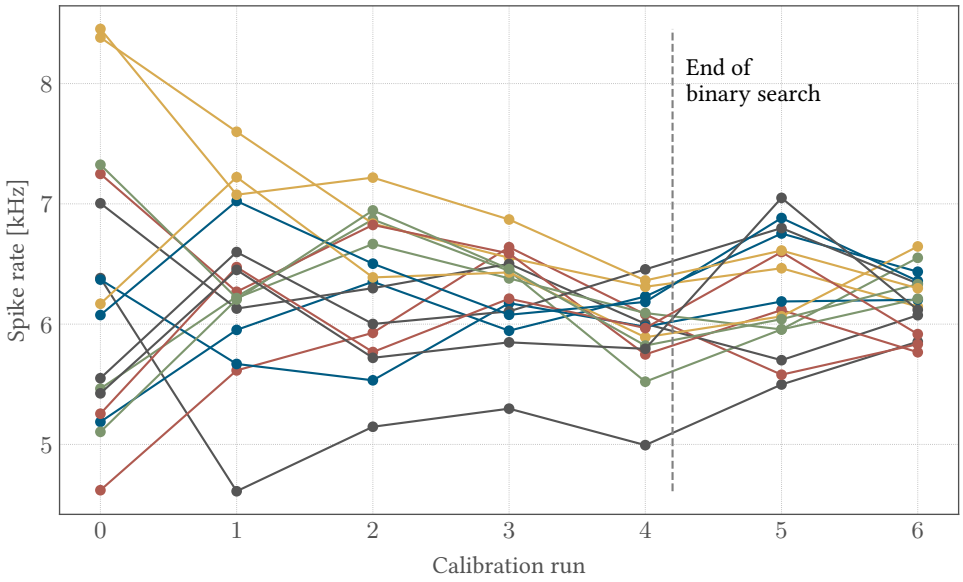


Figure 4.4: Spike rate measurements during the whole calibration plotted for every driver. On the horizontal axis, the calibration iterations are plotted. After run 4, the initial binary search is finished and, after an additional shift, calibrated amplitudes are available on the far right.

In figure 4.2, histograms of the distribution of drivers are displayed. In the left column, we see spike rate distributions while on the right column, amplitude distributions are shown. The first row shows uncalibrated data that was measured using an `offset` setting of 7 for all drivers. Data from 20 runs were taken and averaged to plot the histogram. This allows the distribution to depict a higher resolution than only steps of one. The second row shows the distribution of the same drivers after the `offset` parameter was calibrated using spike rates. For the calibrations, results of only one run are displayed. In the bottom row, the distributions after a conventional calibration using the analog readout is plotted. The scaling of the horizontal axis and the position of bins was kept constant for a whole column. The upper two histograms share the vertical axis. The lower four plots, all showing results of a calibration, are plotted on equal vertical axes as well.

Looking at the top two histograms (figures 4.2a and 4.2b), we can see that the spike rate distribution of uncalibrated drivers is a relatively symmetric function that rises towards a mean value and decreases again. The amplitude distribution looks less symmetric, although the basic shape is preserved: the distribution seems centered around a mean value. It is more difficult to tell where exactly this mean value is located. The two drivers putting out very high amplitudes seem to be more distinct. However, it might be the bin limits of the histograms that make the figures look differently.

Looking at the calibration results below (figures 4.2c to 4.2f), we can see that the relatively large range of values covered by uncalibrated drivers is shrunken down to only three bins. This is a satisfying result, meaning that the calibration algorithm works fine. However, the histograms showing amplitude-based calibration (bottom row) still show smaller deviations than the plots containing spike rate-based calibration in the row above. This is especially true when comparing both amplitude measurements, figures 4.2d and 4.2f. For the amplitude calibration, only two synapse drivers are not in the main bin, one is off to the left and another to the right. For the spike rate calibration, there are six drivers that are not in the center bin. While we can not expect that both calibration mechanisms use the same target for all drivers since the means of the spike rate- and amplitude-distribution are different, in figure 4.2d, even the bin to the right is larger. Two drivers are in there, one more than before, while judging by the spike rate measurements, the calibration target has shifted to the left.

In order to investigate the parameter sets resulting from the calibrations, we compare the `offset` parameter settings after calibration for individual drivers. The values are plotted over the synapse driver numbers. Since we are using the exact same drivers, the parameters should be the same. Identical parameters should produce identical results however, so judging from the histograms, there may be deviations. In figure 4.3, the used `offset` settings are shown for both calibration methods: on the left using spike rates, on the right using extracted amplitudes.

While the basic shape of both plots is the same, looking closely, drivers 0, 1, 5, 6, 9, 13 and 15 all have a different setting. This is best visible looking at the small histograms included on the right of each plot, which differ clearly. There are seven drivers with different settings, which is nearly half of them. The number of different settings is higher than we expected, however, the parameters differ by only 1 LSB, apart from driver 0. It is possible that both used settings do not match the target value very well, so even a small readout noise could decide which setting gets selected. However, this observation confirms the theory stated already in the previous section: due to the higher statistical deviations of the spike rate measurements, calibration results will be less accurate and also less reproducible than the amplitude calibration.

The third figure presented in this section visualizes the path taken by the spike rate-based calibration to find the just displayed `offset` settings. Figure 4.4 shows the spike rate measure-

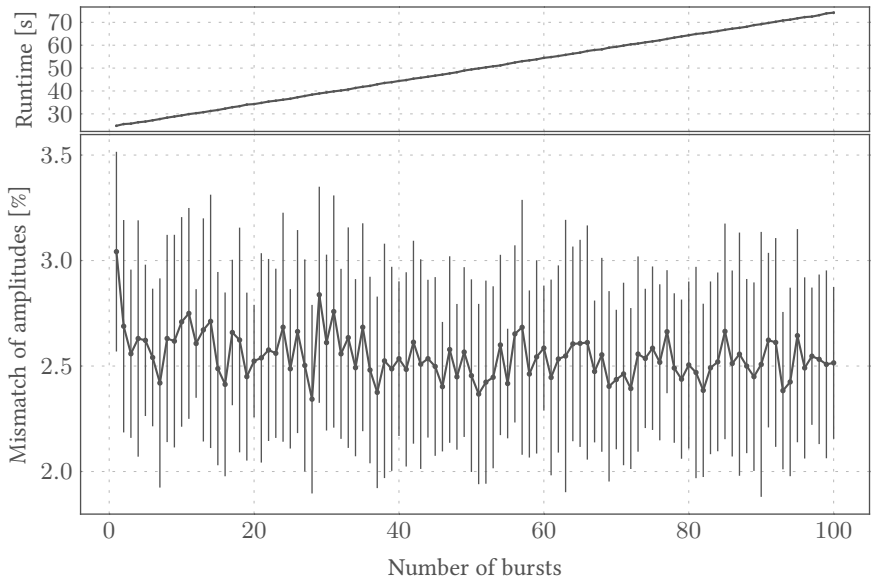


Figure 4.5: Varying the number of bursts used in the experiment, the lower part of the plot shows relative standard deviation of amplitudes. The data is averaged from 20 runs, error bars indicate the statistical variations when repeating the calibration using the same settings. In the upper part, the runtime of the whole calibration is plotted.

ments during calibration plotted over its 7 runs. Each line represents a driver. In run 4, the results of the binary search algorithm that starts the calibration are evaluated. In run 5, the yet untested neighboring setting of the binary search result is tested. Looking at the bottom driver of run 4 in figure 4.4, the additional step proves useful, potentially reducing the spread further. The final run shows the calibrated state, which is also shown in figure 4.2c.

Before closing this section, the spread of the synapse drivers' amplitudes depicted in the histograms shall be expressed by numbers. The calibration results presented show a standard deviation of amplitudes of $0.5\,\mathrm{mV}$ ($1.7\,\%$ of the mean amplitude) for the amplitude calibration (figure 4.2f) and $0.9\,\mathrm{mV}$ ($3.1\,\%$) for the spike rate calibration (figure 4.2d). It was $4.6\,\mathrm{mV}$ ($17\,\%$) before the calibration (figure 4.2b). Note that the amplitude calibration used here was not fully optimized. When, e.g., synapse weights are tuned first to compensate for their mismatch, it is possible to achieve an even lower spread. The spike rate readout is less susceptible to unequal amplitudes from the synapses, since using multiple neurons in parallel enables averaging rates from multiple synapses. The amplitude readout uses only one synapse column and therefore is more susceptible to synapse mismatch. Also, the used inter-spike-interval is very low for reliable amplitude readout, as explained in section 3.1.4. While the measurements in [Weis, 2017] show lower numbers, they were done at a higher V_{charge} setting of $240\,\mathrm{LSB}$ which decreases relative deviations. Thus, the numbers are not comparable. Additionally, a different chip was used.

4.4 Runtime versus mismatch

The statistical variations in spike rate measurements seem to be a significant problem in order to achieve less mismatch of amplitudes after calibration. One could expect that by changing the

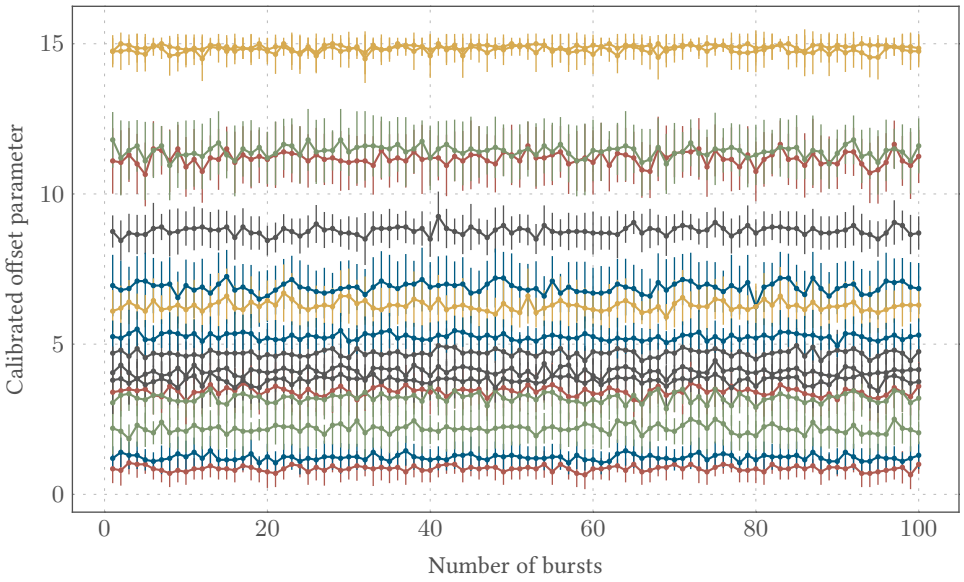


Figure 4.6: The obtained `offset` parameters from the calibrations are shown depending on the number of bursts used. The data is averaged from 20 runs, error bars indicate the statistical variations. One line per driver is plotted. No trend is visible, so experiments that use many bursts to average from yield the same parameters as ones with very few bursts.

number of bursts that are sent during one experiment, these variations differ. Since sending more bursts means that more averaging is done in-place, spike rate variations can be interpreted as the error of the mean spike rate and should decrease proportionally to the square root of the number of bursts. Sending more bursts obviously takes time, so the potentially increased calibration quality is at the expense of a longer runtime. Since for the 40 bursts used in figure 4.1 variations are quite high, we expect that by increasing the number of bursts we get lower relative standard deviations of calibrated amplitudes.

To investigate how the number of bursts affects mismatch and runtime, we sweep it from 1 to 100 and run the calibration at each setting. We disable the amplitude readout, requiring the majority of time. The runtime measurement is started after the parameter search for the synaptic input reference voltage V_{syn} is complete and stops after run 6, when the calibration results were tested once. Afterwards, amplitudes in the calibrated state are acquired.

This way we are able to plot the standard deviation of amplitudes and the pure runtime of the spike rate calibration over the number of bursts used, which is shown in figure 4.5. In the plot, data is averaged from 20 measurements, error bars indicate the statistical deviations during those. The relative amplitude mismatch is constant for almost the whole plotted range, statistical deviations are multiple times larger than the systematic decrease we expected. Only for the first data point using only 1 burst, a higher average mismatch is observable. Hence, we can reduce the number of bursts to low numbers, such as 5.

The `offset` parameters that were the result of the calibrations are plotted in figure 4.6. There, one plotted line corresponds to a single driver. Error bars still show the standard deviation across 20 runs. We can see that the lines are constant, except for statistical variations, in particular not even the amount of variation shows a trend. Neither the size of the error bars nor the parameters change with the number of bursts. This makes us even more confident in using low burst counts:

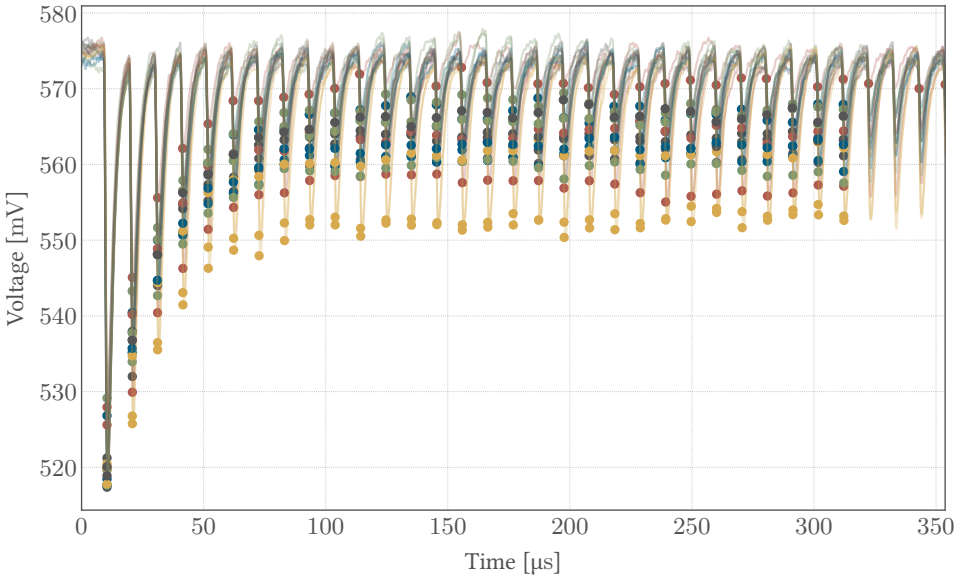


Figure 4.7: STP traces in usable range. One colored low-pass-filtered trace per driver, spike peaks are marked in the same color. The `offset` parameter is *not calibrated*.

if the resulting set of parameters stays the same, the relative standard deviation of the drivers has to be constant as well. The set of `offset` parameters is the only actual result of the calibration.

The second advantage of the spike rate calibration, besides its scalability, is shown in the upper graph in figure 4.5: the runtime. Here, the minimum runtime of the 20 executions is plotted. This represents best what runtime can be expected by selecting the run that was the least disturbed by other processes. The data shows that the calibration can be done in roughly 30 s, when low burst counts like 5 are used. For the calibration shown above, using 40 bursts, a runtime of about 45 s and an average standard deviation of amplitudes of 2.5 % are expected. Looking at the error bar however, the previously obtained 3.1 % relative mismatch is only slightly outside the 1 σ area and thus not unusual.

The result of this experiment is that we can use the new readout confidently: using about 30 s of runtime, we get an offset parameter calibration that reduces amplitude mismatch of the synapse drivers down to (2.5 ± 0.5) %. Since the larger HICANN-X chip contains more synapse drivers but also proportionally more neurons that can be used for parallel readout, the runtime should not change. Keep in mind that here, the experiment is still controlled by the FPGA. Potentially, the runtime changes when implementing the calibration locally on the PPU.

4.5 STP example traces

In this section, we want to verify that the calibration can be applied to real world STP usage. During calibration, the difference between V_{charge} and V_{recover} is very small in order to reach constant spike amplitudes fast and have all drivers' amplitudes in a measurable range. In this section, we present examples of STP traces that use a wider range of STP voltages and thus change amplitudes by a larger amount. We will also compare results for the uncalibrated state and both calibration methods. All plots presented here are measured using the ADC in order

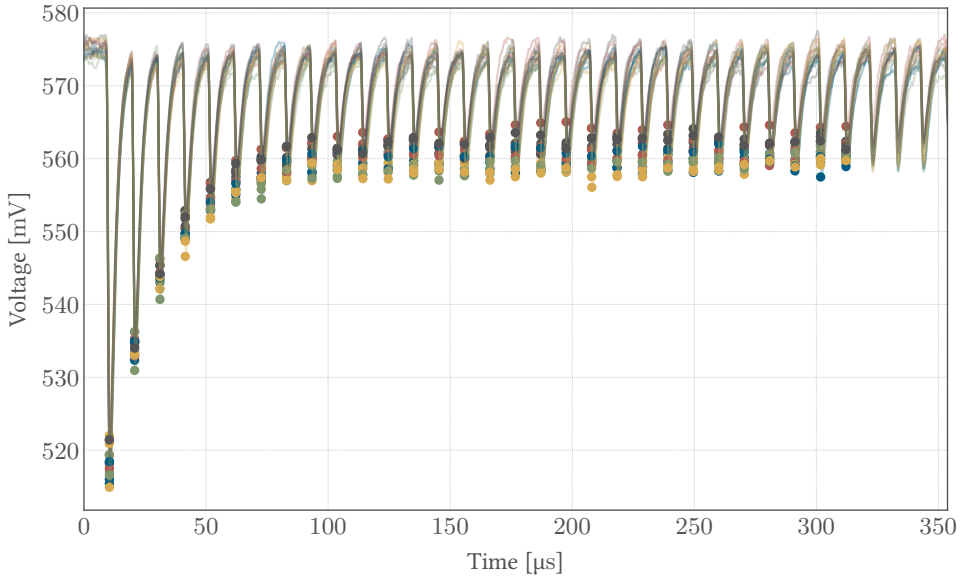


Figure 4.8: STP traces in usable range. One colored low-pass-filtered trace per driver, spike peaks are marked in the same color. The `offset` parameter is *calibrated using the spike rate readout.*

to acquire voltage traces. Comparing the synapse drivers, all their amplitudes are plotted in the same figure. For every driver, the full low-pass-filtered trace is plotted and spikes found by the usual algorithm are marked as colored dots. The used voltages are $V_{\text{charge}} = 80\,\text{LSB}$ and $V_{\text{recover}} = 320\,\text{LSB}$. The V_{offset} and V_{zero} parameters are unchanged in order to use the available calibration.

The experiment is executed using neuron 12 on chip 8. A burst of spikes is sent to the drivers, similar to the calibration scenario. Of the 300 sent spikes, only the first 30 are plotted, since the decrease in amplitudes during the depression phase is what we are interested in. In order extend this depression phase, we do not disable the recovery and no longer use the `renewing` setting of the synapse drivers. It is the opposite: they are now configured to have a low U_{SE} parameter. We use the settings `utilization` $= 0$, `recovery` $= 0$, `prescaler` $= 5$. The `enshare` setting is still enabled.

The original uncalibrated traces are shown in figure 4.7. We can see that spike amplitudes are strongly spread, which seems so due to the lower amplitudes here. The amplitude offsets between the drivers are constant for all depression states, which means that in this state with low amplitudes, the relative deviations are high. For one driver, the driver on the top marked with a red amplitude trace even showed amplitudes that are too small for the amplitude extraction algorithm to work properly: some of the spikes were not recognized. If, during calibration, an amplitude or spike rate is out of range and reads zero, the `offset` parameter still gets shifted into the right direction, but this will strongly shift the mean of all drivers, which is the calibration target. Therefore, we avoid this large range of amplitudes during the calibration.

While the traces in the uncalibrated figure impressively show why a calibration is inevitable, the traces calibrated by spike rates lay much closer. They are displayed in figure 4.8. The spread of the marked spike peaks is partially caused by the spread of the baselines. Comparing the spikes, we can see that the measured voltages are not constant either. This is why plotting these traces for an amplitude-based calibration yields a very similar figure. In order to evaluate amplitudes

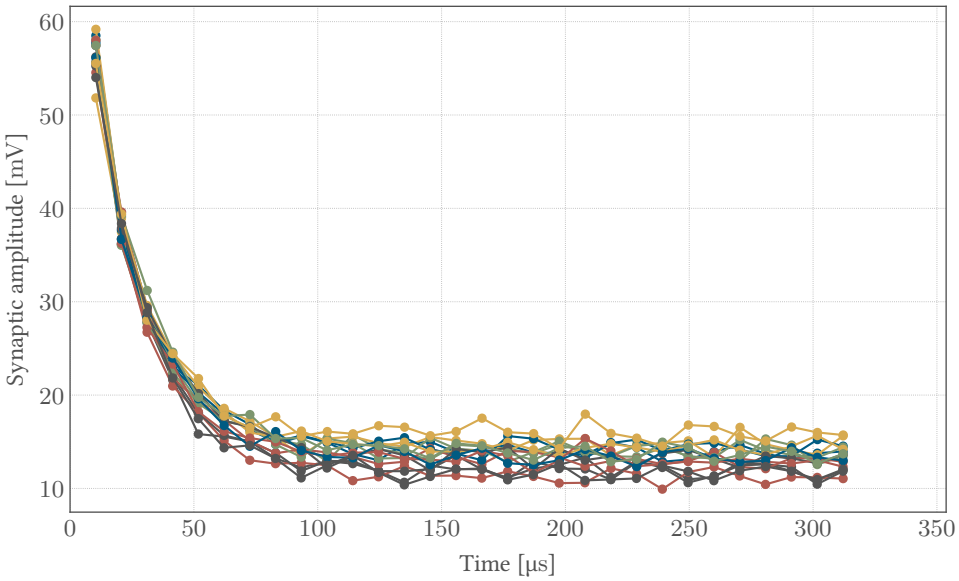


Figure 4.9: STP amplitudes *calibrated using spike rates*, extracted from the shown trace (figure 4.8) subtracting the baseline. Each line represents a single driver.

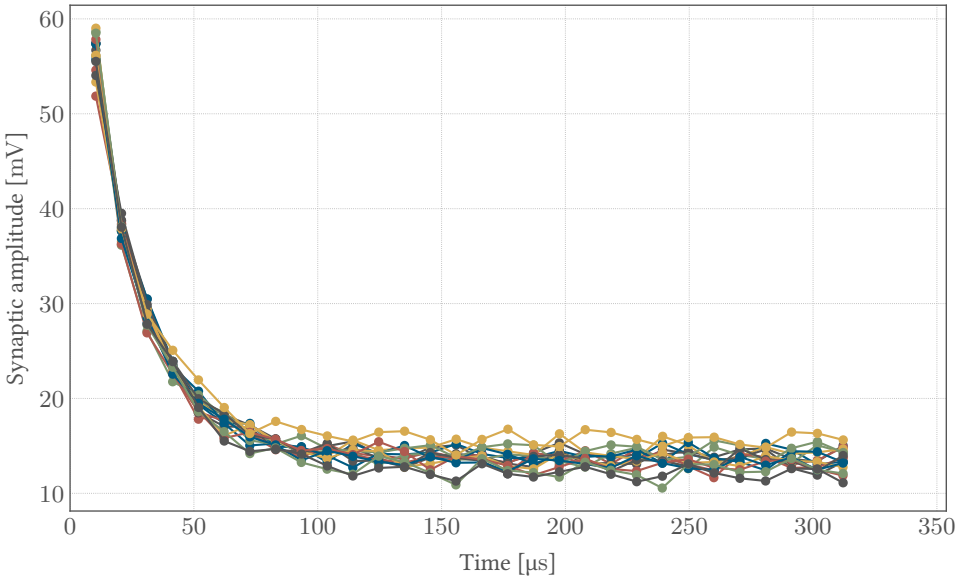


Figure 4.10: STP amplitudes *calibrated using the ADC for amplitude readout*, extracted subtracting the baseline. Each line represents a single driver.

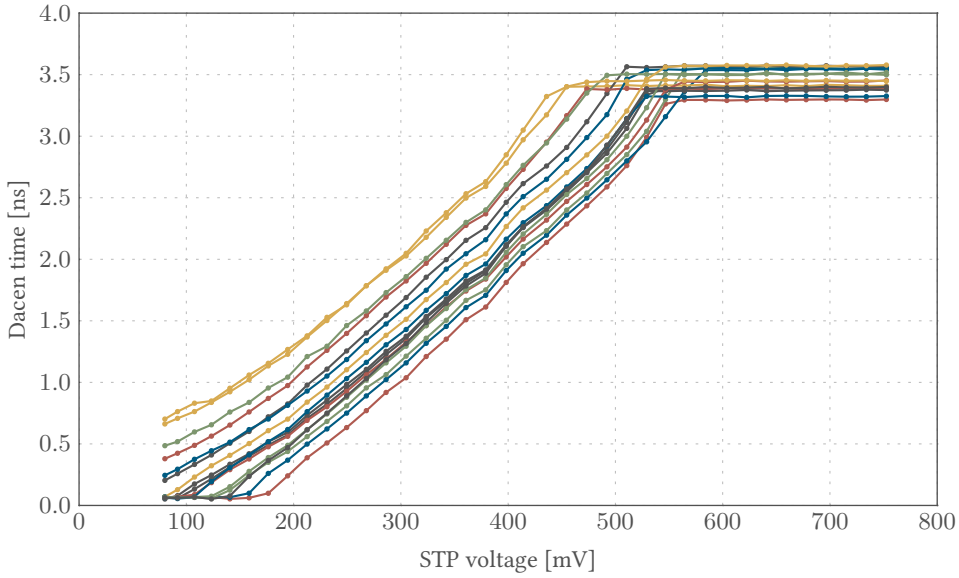


Figure 4.11: `dacen` time plotted over the voltage on the STP storage capacitor V_{STP}. The `offset` parameters are *uncalibrated* and set to 7 for all drivers. Data is read out using ADC amplitude measurements.

more precisely and compare the two calibrations, we plot only the amplitudes over time. This means we use the amplitude extraction algorithm in order to subtract the individual baseline at every spike.

The extracted amplitudes for the spike rate calibration are plotted in figure 4.9. We can see that in relation to the amount of depression, from $55\,\text{mV}$ to $14\,\text{mV}$, the spread of the individual drivers looks small. The STP mechanism seems to be usable with this calibration applied. In comparison to the amplitude trace with a calibration using ADC amplitudes, which is shown in figure 4.10, the amplitude calibration manages to decrease the spread. Therefore, we can confirm that, for the best results, an amplitude calibration is the best approach. As previous data has already shown, the mismatch after a spike rate calibration is simply larger, there is no doubt about that. One will have to decide whether the spike rate calibration is good enough or the decreased mismatch is worth the long runtime of the amplitude calibration. Usually, the spike rate calibration should be sufficient.

4.6 STP comparator ramps

The presented calibration mechanism is able to compensate for constant offsets. Other deviations such as linear terms can not be equalized. Looking at the STP example shown in figure 4.9, we can see that some drivers cross traces of other drivers. The spread is not constant across the whole range of amplitudes. There are two major sources which can introduce a deviation that is not constant: the generation of V_{STP}, including utilization and recovery, and the comparator ramps, possibly varying in steepness.

Calibrating on a specific voltage V_{charge} means that all drivers generate the same amplitudes at this voltage, but not necessarily at other voltages. Assuming a configuration as used here where

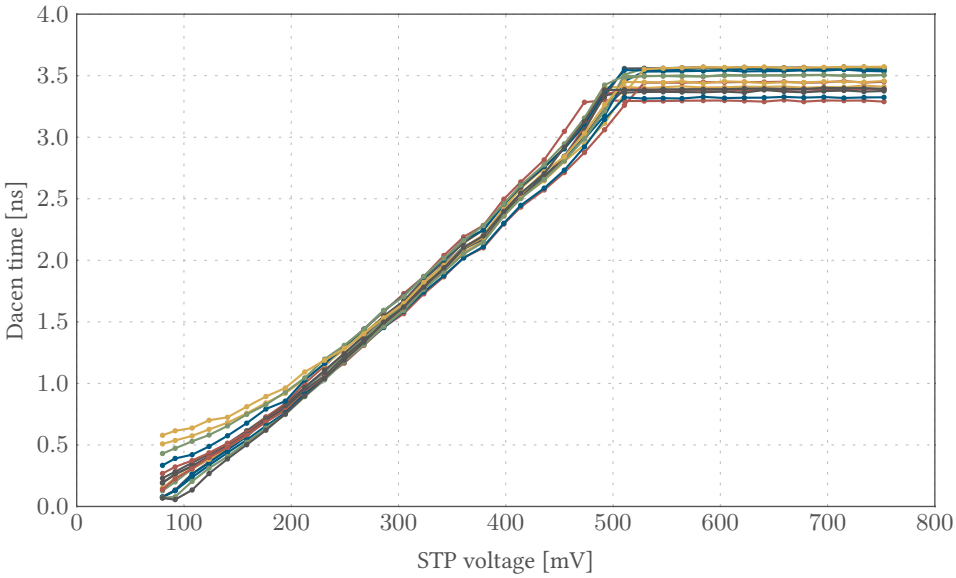


Figure 4.12: `dacen` time plotted over the voltage on the STP storage capacitor V_{STP}. The `offset` parameters are *calibrated using amplitudes*. Data is read out using ADC amplitude measurements.

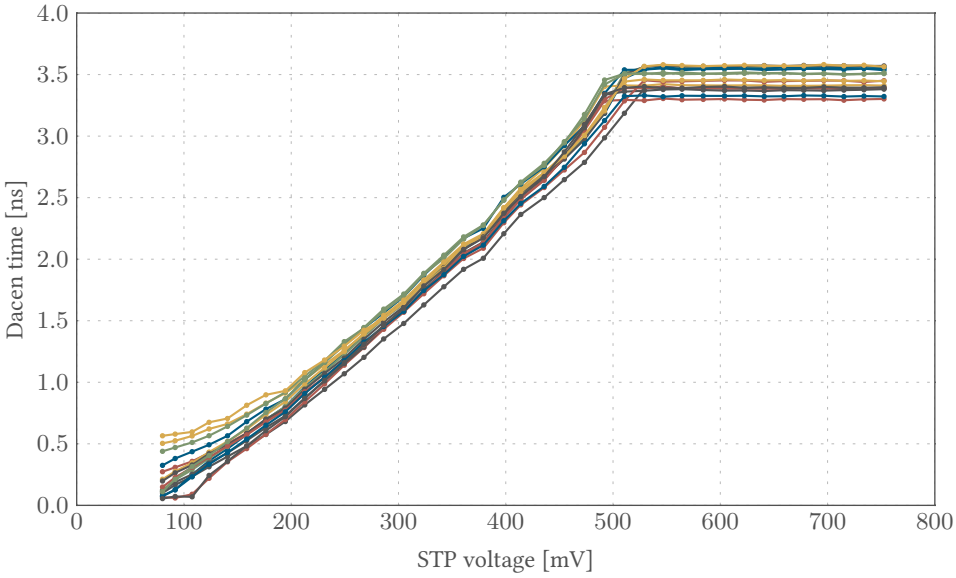


Figure 4.13: `dacen` time plotted over the voltage on the STP storage capacitor V_{STP}. The `offset` parameters are *calibrated using spike rates*. Data is read out using ADC amplitude measurements.

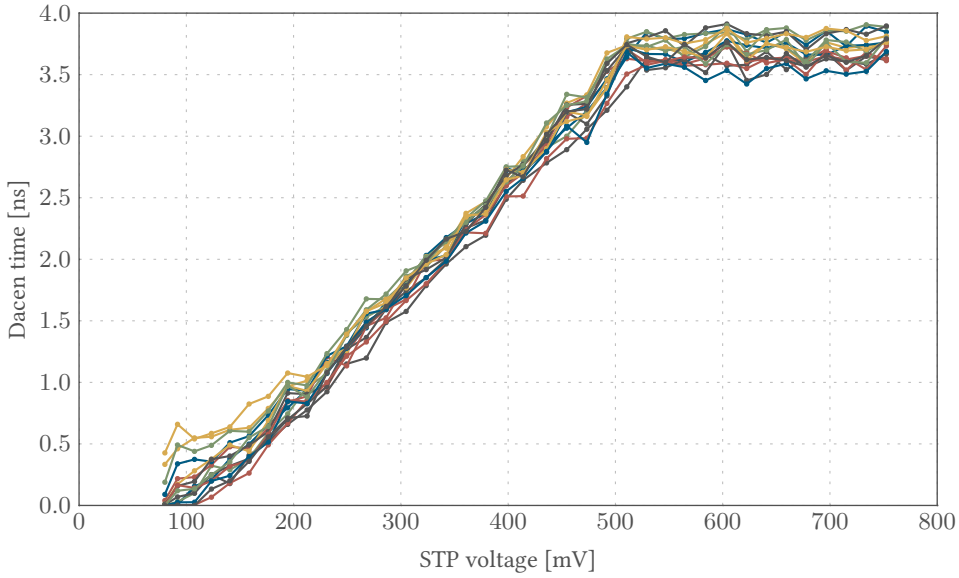


Figure 4.14: `dacen` time plotted over the voltage on the STP storage capacitor V_{STP}. The `offset` parameters are calibrated using spike rates. Data is read out using neuron *spike rate measurements*.

$V_{\text{recover}} > V_{\text{charge}}$ and `mode` $= 1$, changes in ramp steepness mean the following: if the current I_{ramp} responsible for generating the comparator ramp is lower for one driver, the amplitudes for depressed states are higher than usual while the amplitudes for a fully recovered state is lower. This leads to drivers crossing each others' curves.

In order to investigate the spread of the ramp slope across all drivers, we plot the `dacen` pulse widths over the voltage V_{STP} on the STP storage capacitor. In addition, with these plots we can estimate the usable range of STP voltages until the `dacen` pulse does not depend linearly on the voltage any more.

The length of the `dacen` pulse cannot be measured easily. We calculate it as the ratio of the measured synaptic input amplitude versus the synaptic input amplitude with STP disabled, which would equate to a `dacen` pulse of $5\,\text{ns}$ on this particular setup with reduced clock speeds. Therefore, at first, all the amplitudes with STP disabled are measured. Dividing by these means that synapse mismatch will not influence our measurement since we use an individual value for each synapse. Since we are reading out synaptic input amplitudes, we can do this either using the ADC or neuron spike rates. We tested both: the spike rate readout shows similar results but a higher statistical noise on the data, exactly as we would expect from previous experiments. We therefore prefer and use the ADC to read out the amplitudes. Nevertheless, a plot showing data read out using spike rates is included as well. The horizontal axis of the plots, showing the applied voltage on the STP storage capacitor, is translated from the capmem setting in LSB to mV using the capmem characterization presented in figure 3.2a. The actual voltage V_{STP} at every driver involves error as well.

In the experiment, we configure the synapse drivers like during calibration. This means we enable the `renewing` setting which discharges the STP storage capacitor completely to V_{charge}, and disable recovery. We send 40 bursts of 300 spikes for the spike rate measurement but measure amplitudes in parallel using only 200 of the available spikes in total. Again, the first 10

amplitudes are discarded. Sweeping V_{charge} from $30\,\text{LSB}$ to $400\,\text{LSB}$, we measure the ramps for both calibration methods and in uncalibrated state.

In figure 4.11, the uncalibrated ramps are shown. We can see the high spread of `dacen` times between different drivers at the same voltage V_{STP}. This is exactly the offset that we are calibrating for. So comparing with the plots showing the ramps after amplitude calibration (figure 4.12) and after spike rate calibration (figure 4.13), we can see that the mismatch is reduced drastically there. Again, the amplitude calibration shows a little less spread than the spike rate calibration. The usable range of voltages V_{STP} is from around $200\,\text{mV}$ to $480\,\text{mV}$. The optimal setting for V_{recover} is higher:

$$V_{\text{recover}} = V_{\text{charge}} + (V_{\text{STP,max}} - V_{\text{charge}}) \cdot (1 + U_{\text{SE,min}}), \tag{4.1}$$

with $V_{\text{STP,max}}$ being the maximum usable STP voltage and $U_{\text{SE,min}} \approx 0.27$ being the utilization of synaptic efficacy when `enshare` is disabled. The reason is that charge on C_{storage} is shared with parasitic capacity before it reaches the voltage comparator. Only the additional sharing in order to increase U_{SE} is done after the spike transmission [Billaudelle, 2017, figure 25]. This explains why in the STP example traces a voltage $V_{\text{recover}} = 320\,\text{LSB}$ can be used, while in the ramp plots, the maximum `dacen` times are reached at around $500\,\text{mV}$, which equates to roughly $250\,\text{LSB}$.

These first three figures show data that was acquired by using the ADC to find amplitudes in the voltage trace. For very low voltages V_{STP}, there are two visible effects: the spread between drivers increases, and even the `dacen` times for the lowest drivers never actually reach zero. The latter happens because the dacen time is directly calculated from the amplitude measurements. If there is only noise present during the extraction of amplitudes, some edges in the noise may be treated as spikes.

In figure 4.14, the STP comparator ramps are plotted using spike rate measurements. The drivers are configured to use offsets from the spike rate calibration and should therefore look identical to figure 4.13. The new plot using spike rate readout shows a much higher noise on the measurements itself, but we can see that for very low voltages V_{STP}, the `dacen` times actually drop to zero. When the neurons show no spikes, the spike rate reaches zero, in contrast to the amplitude measurement which still shows amplitudes different from zero and therefore shows a `dacen` time that is non-zero.

The increased spread between drivers at low amplitudes is partly caused by the STP comparator not being designed for voltages below $100\,\text{mV}$. This poses a problem when STP should be configured so that fully depressed spikes have an amplitude of zero. However, the offset voltage V_{offset} can be increased, so that the ramp start voltage is higher. Another factor is the edge steepness of the `dacen` pulse. For very low amplitudes, the falling edge starts before the pulse has fully settled. Some synapses might still detect the pulse, while others, due to mismatch, do not receive it at all. In biological use cases, the recovery normally prevents such strongly depressed amplitudes.

We conclude that the comparator ramps are not the cause for crossing curves in an STP experiment (figure 4.9). The crossing is most likely caused by differences in STP recovery.

5 Spike Timing Dependent Plasticity

5.1 Characterization of Correlation ADC

In order to collect data about the correlation of presynaptic input and postsynaptic neuron spike timings, the voltages stored on the STDP correlation capacitors are evaluated using the CADC (section 2.4). The CADC has two channels for each synapse column, one for anticausal and causal measurements, respectively. They can be connected to their synapse column or to a common external voltage using switches addressed as a virtual synapse row 33 [Wunderlich, 2016]. As part of the readout mechanism, the CADC is inspected first. Being subject to mismatch, it provides an offset setting for each channel that allows shifting of the readout. A calibration within 1 LSB is feasible.

The CADC compares the applied input voltage with a voltage ramp. An internal counter measures the time from the start of the ramp until the input voltage is reached. The calibration offset is subtracted from the original counter value before the result is transmitted. Since the voltage ramp is shared between all CADC channels, no further calibration is necessary. In order to find the correct offset parameters, it is sufficient to apply a constant external voltage at a medium level that does not clip on any ADC. The offsets are set as the respective differences to the minimum measured value. The voltage is supplied by the baseboard using a DAC that is connected to the synapse debug line. Since all the CADC channels are connected to this line as well, it poses no problem if the applied voltage is not precisely the desired voltage. It only has to be identical for all channels, which it is by design.

A characterization of the CADC output is shown in figure 5.1. There, the applied external voltage is swept from 0.1 V to 1.1 V in steps of 0.1 V, each line corresponds to one channel. Black lines represent anticausal channels, red lines represent causal ones. We can see that the linear range is calibrated well indeed, it shows only minimal mismatch. Depending on the configured offset, the channels clip at different levels. For voltages below the usable range, the digital output clips to 255 instead of 0. This is a known bug. The plot shows that the usable linear range is around 30 LSB to 200 LSB of digital output, so we will keep the correlation amplitudes in this range.

5.2 Characterization of Correlation amplitudes

5.2.1 Measurement setup

Correlation between synapitc inputs and the neurons' spike events can be causal or anticausal. To express configurations more clearly, we use the term "prespike" for the signals coming into a synapse from the synapse drivers, representing synaptic inputs; and the term "postspike" for the signals outputted by neurons that indicate it was spiking. If prespikes arrive before postspikes, we call that causal correlation, while postspikes arriving before prespikes are called anticausal correlation. The exponentially weighted time differences are correlation amplitudes, as shown in figure 5.2. When plotting the correlation amplitudes, the horizontal axis represents the time

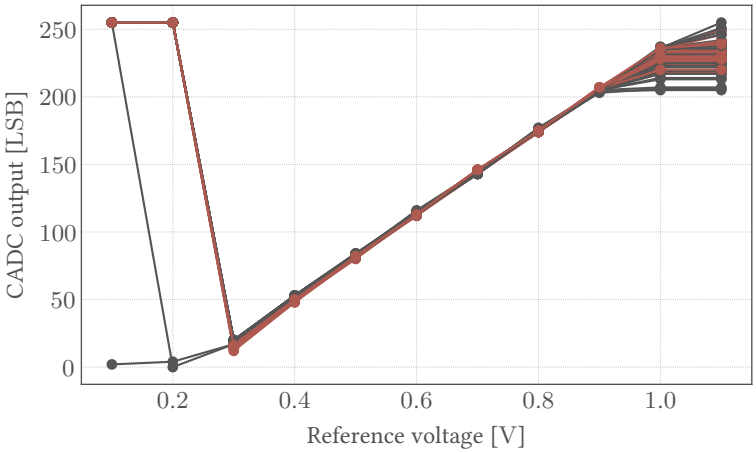


Figure 5.1: Characterization of the CADC output codes depending on the applied input voltage.

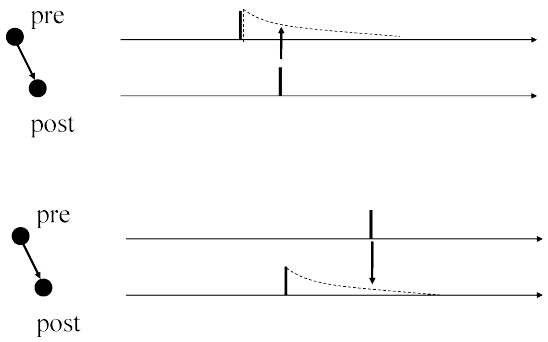


Figure 5.2: Correlation of pre- and postsynaptic events can be measured as amplitudes. The upper part shows causal, the lower part anticausal correlation. With higher time differences, the measured amplitudes decrease exponentially. Figure adapted from [Sjöström and Gerstner, 2010].

difference between pre- and postspikes. As a consequence, the negative branch represents anti-causal correlation, positive values correspond to causal correlation.

An experiment starts with resetting the correlation capacitors to a voltage V_{reset}. This voltage is chosen such that it does not exceed the linear range of the CADC measurements. It is generated using a DAC on the baseboard. From the available range of 4095 LSB, which corresponds to a voltage of 2.5 V, we use a setting of 3300 LSB. Since readout of this voltage via the CADC includes a buffer in the synapses, this voltage does not equate to the input voltage at the CADC. After the reset, the initial voltage on the capacitors is read out via the CADC and used as a baseline. Therefore, a calibration of the CADC is not crucial, only the changes in output signals are relevant. Then, 6 correlated pairs of pre- and postspikes are sent. That means depending on the desired time difference, either the synapse drivers or the neurons send a signal to the synapses first.

We send one pair of pre- and postspikes, sweeping the time difference in steps of $0.52\,\mu s$, which equates to 50 FPGA clock cycles. Time differences of up to $10.4\,\mu s$, or 1000 FPGA clock cycles, are measured. Before sending the next pair of spikes, we wait $417\,\mu s$ or $40\,000$ FPGA cycles. After all correlated spikes have been sent, the CADC is used to read out all correlation measurements again. This allows calculating the difference between the initial and the final measurements, which we will refer to as the STDP amplitude.

A big advantage on this chip compared to the previous generation is the ability to generate an artificial spike at the neuron which produces the postspike signal for the synaptic column. It is no longer necessary to use synapse rows on high weights for triggering the spiking of neurons. Therefore, all synapse weights are set to 0 in this experiment.

The configured voltages (see figure 2.4) are $V_{\text{resmeas}} = 4095\,\text{LSB}$, $V_{\text{ramp}} = 1100\,\text{LSB}$, $V_{\text{store}} = 1340\,\text{LSB}$ all generated on a $1.2\,\text{V}$ 12-bit DAC, and $V_{\text{reset}} = 3300\,\text{LSB}$, $V_{\text{coroutbias}} = 573\,\text{LSB}$ generated on a $2.5\,\text{V}$ 12-bit DAC.

5.2.2 Measurement results

The correlation amplitudes are evaluated as a function of the differences between spike times. Since we observe a dependency between the number of synapses that are configured on the correct prespike address and correlation amplitudes, the presented plots show correlation measurements of only the first synapse column. The remaining synapse columns listen to prespikes if enabled, but never receive postspikes.

For the first experiment, the whole synapse array is configured at address 63, where the prespikes arrive. For this setup, we see a strong asymmetry. The observations are plotted in figure 5.3. Anticausal amplitudes are plotted in blue, causal amplitudes in green. Error bars indicate the standard deviation between the 32 available synapses. The anticausal measurements reach amplitudes of about 71 LSB at a time difference of -100 FPGA clock cycles ($-1.04\,\mu s$), while causal amplitudes are only 29 LSB on average at a time difference of $+100$ FPGA cycles. A similar asymmetry was observed on HICANN-DLS 2. However there, the anticausal measurements showed lower amplitudes than the causal ones. The problem appears inverted.

Investigating the issue, we configure only one column of synapses to listen to the injected events. This way, only the first column receives prespikes and measures correlation, the remaining synapses are entirely disabled. The observed asymmetry can not be reproduced when enabling only a single synapse column. The measurement is plotted in black for the acausal amplitudes and in red for the causal amplitudes in figure 5.3. The error bars indicate the standard deviation between synapses. As a conclusion, the problem must lie outside of the individual synapses' STDP circuitry.

We see that causal amplitudes are much lower when all synapses listen to prespikes (green curve) than when only the first column receives prespikes (red curve). Otherwise, the plot shows that there are no causal amplitudes measured while the spike timing is anticausal and vice versa. The first measurement after switching from anticausal to causal, for some reason, shows a small anticausal amplitude and a lower causal amplitude than expected. This is also observable for the first anticausal measurement. We did not investigate this further, but discard the respective data points.

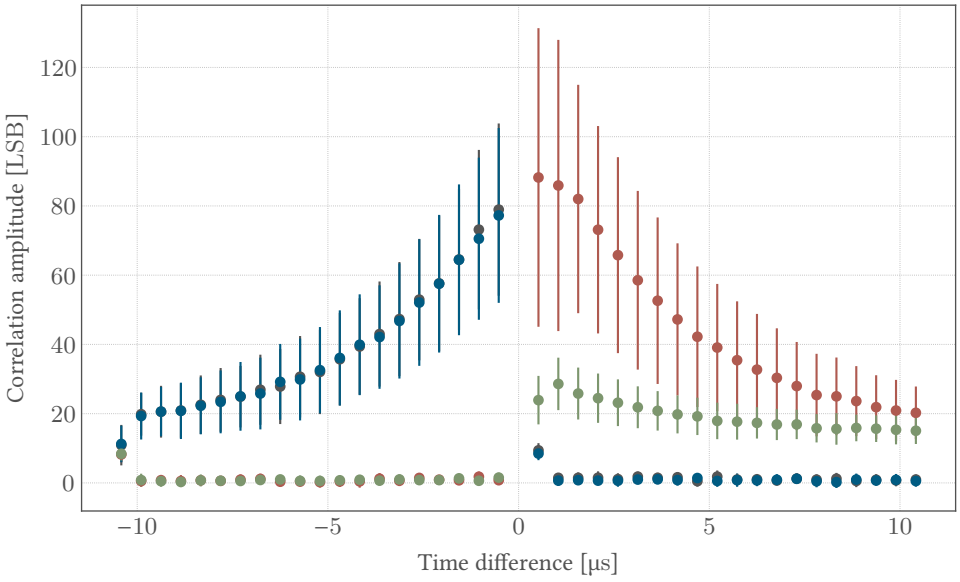


Figure 5.3: Measured correlation amplitudes as a function of the time difference between pre- and postspike. The data points are average amplitudes of the first synapse column, error bars show their spread. Black: anticausal measurement with addresses set in 1 column, red: associated causal measurement. Blue: anticausal measurement with addresses set in 32 columns, green: associated causal measurement. A strong asymmetry between anticausal and causal measurements is visible when enabling 32 columns of synapses.

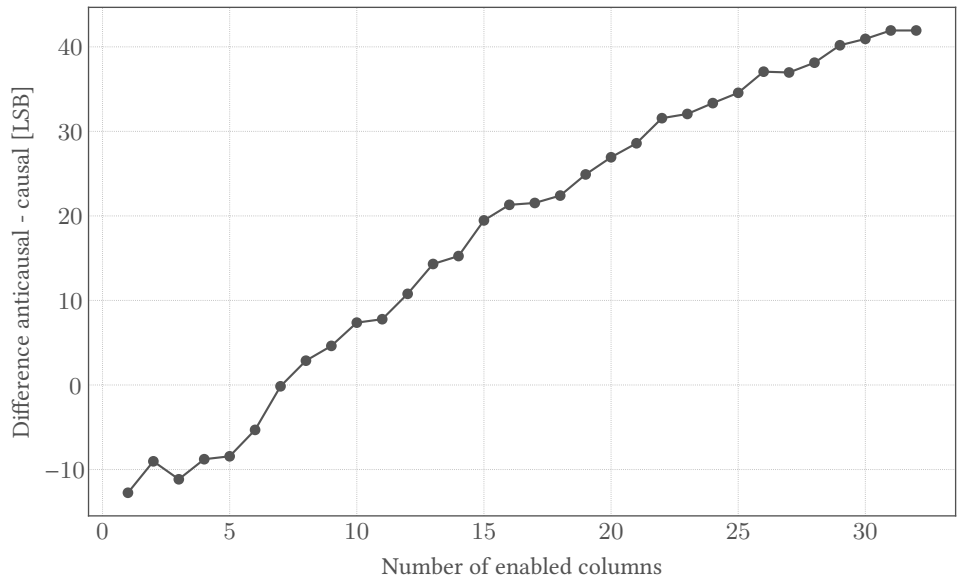


Figure 5.4: Dependency of the asymmetry of causal and anticausal STDP amplitudes on the number of synapse columns that are configured to the correct address.

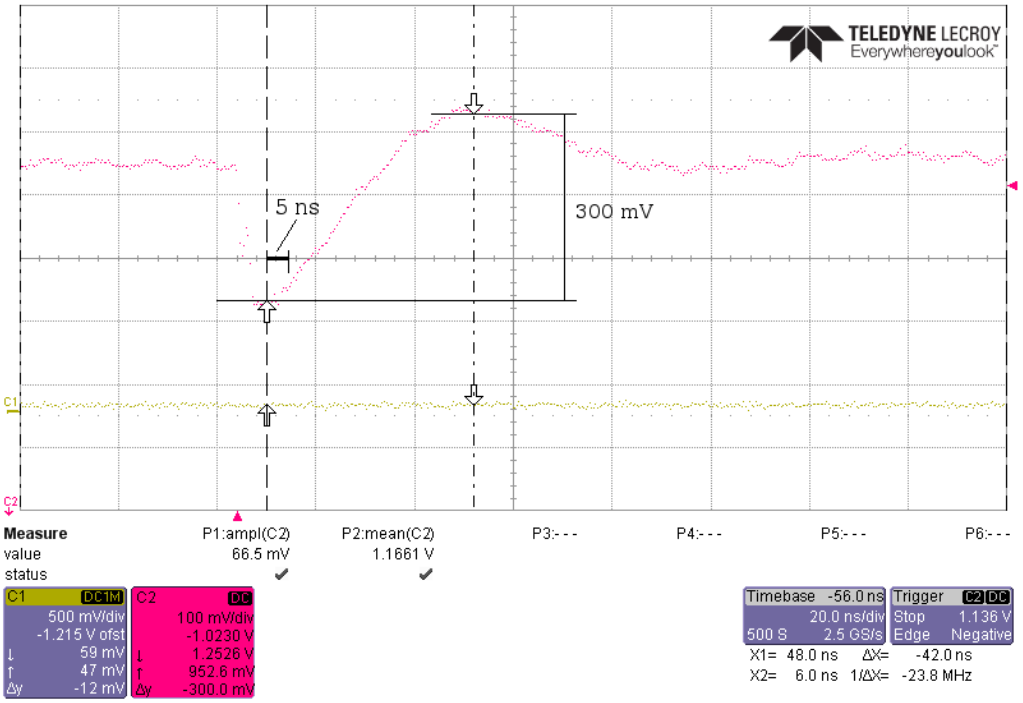


Figure 5.5: Oscilloscope readout of the used voltage: Channel `C2` shows a voltage trace of V_{resmeas} during a prespike with 31 synapse columns listening to the address. A significant drop is observable.

5.3 Troubleshooting asymmetry

To investigate the asymmetry, we sweep the number of enabled synapse columns. The amplitudes of the causal measurements decrease monotonically with the number of synapse columns that listen for prespikes at the correct address, as shown in figure 5.4. There, the difference of anticausal and causal amplitudes at a time difference of 100 clock cycles each is plotted over the number of synapse columns that have the correct address set. The curve does not seem to settle at the shown 32 enabled columns: if there were more synapses on the chip, amplitudes would get even lower and asymmetry would grow even stronger. This effect made us suspect the reset voltage for the measurement capacitors, V_{resmeas}, to be insufficient [Schemmel, personal communication, December 2017].

When a prespike arrives, a measurement is started. This means the measurement capacitor is connected to V_{resmeas} in order to be charged. When only one synapse column receives prespikes, V_{resmeas} does not drop significantly. However, when all synapses are configured at the same address, all 1024 synapses charge their respective capacitor at the same time. This effect only occurs for causal measurements, since for the anticausal branch, measurements are started with the postspikes. In this experiment, these are never sent to all synapses at once.

A drop of V_{resmeas} would explain the observed behaviour. When setting the voltage from the original 4095 LSB down to 3200 LSB on the 1.2 V DAC, causal amplitudes are about as high as in the asymmetric measurement, even though only one synapse column is used. The latter DAC setting corresponds to a voltage of 938 mV, assuming the DAC works perfectly linear. As a next step in troubleshooting the problem, V_{resmeas} is recorded during experiments as close to the chip as possible, using the oscilloscope and an active probe. We see that indeed the voltage

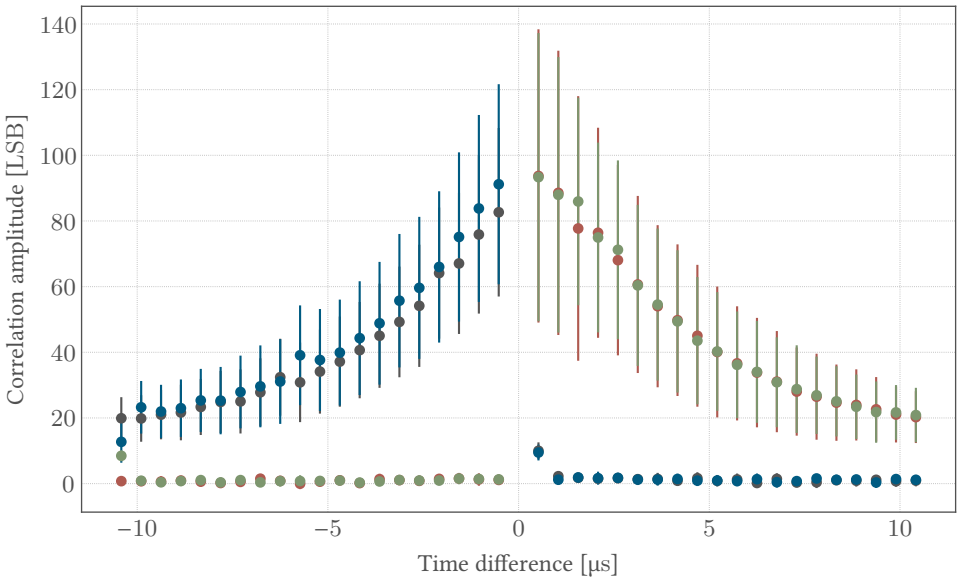


Figure 5.6: Characterization of the STDP mechanism after including a capacitor next to the chip. The anticausal and causal amplitudes now show only minimal asymmetry. The experiment is identical to the plot in figure 5.3, colors are chosen identical as well.

is dropping significantly. At its lowest point, it was found to be $953\,\mathrm{mV}$, which approximately corresponds to our observations above, considering we expect the voltage to be $1200\,\mathrm{mV}$.

In figure 5.5, the oscilloscope measurement is shown. The two cursors indicate the lowest point during the drop and the highest point afterwards, that lays higher than the baseline voltage. Between the minimum and the maximum of overdrive lay $42\,\mathrm{ns}$ of time and $300\,\mathrm{mV}$ of voltage difference. The time during which the voltage is low is much longer than the $5\,\mathrm{ns}$ time that synapses receive the prespike for. Also, the drop amplitude is strong enough to explain the observed effect.

The voltage drop can be reduced by placing a block capacitor next to the chip that holds sufficient charge to supply simultaneous charging of all STDP measurement capacitors. The DAC on the baseboard is able to drive a capacitive load of $500\,\mathrm{pF}$ [Schreiber, personal communication, December 2017]. Therefore, a $470\,\mathrm{pF}$ capacitor is soldered onto the chipboard. With this fix applied, no voltage drops are observable on the oscilloscope.

Using the same setup of first setting addresses at only the first synapse column and in a second run synapses in 32 columns, a new plot is produced. As shown in figure 5.6, the observed asymmetry has now vanished. Plotted using the same colors as in figure 5.3, we see that causal measurements are no longer influenced by the number of enabled synapses. We learn from this that the power draw for the reset of multiple STDP measurement capacitors C_{measure} is not to be underestimated. However, in a biological experiment with synapses configured at many different addresses, the case that the entire synapse array receives a prespike at the same time is unusual. In most experiments, problems would be much smaller than in this artificial test scenario.

With the help of Benjamin Cramer, we tried to apply the same fix for the asymmetric amplitudes observed on the previous chip generation, HICANN-DLS 2. Before soldering a capacitor, a drop of V_{resmeas} was observed as well. While the voltage drop indeed vanishes using the capacitor, the STDP characterization measurement on the older setup still does not show symmetric

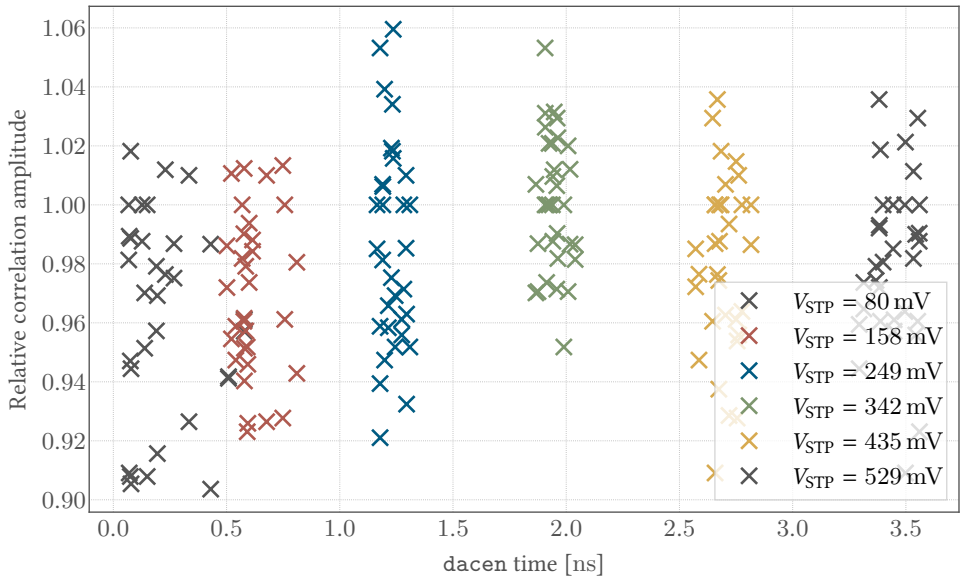


Figure 5.7: Dependency of STDP measurements on the STP state. On the vertical axis, the ratio of the STDP amplitude at the given STP state to the amplitude with STP disabled is plotted. On the horizontal axis, data from the STP ramp experiments is used to translate the used voltages V_{charge} (see legend) into lengths of the `dacen` pulse.

amplitudes. Certainly it changed, amplitudes now show less deviations than before. In particular, without the capacitor, it was not possible to have both anticausal and causal amplitudes in a usable range for many synapses. When choosing parameters so that causal measurements were readable, anticausal measurements read 0. When choosing parameters that yield higher amplitudes in general, anticausal measurements become readable, at the cost of causal measurements clipping at the maximum value. This problem is now less severe and multiplying anticausal measurements by a constant factor seems to work. The correlation measurements are usable [Cramer, personal communication, January 2018].

5.4 STDP in combination with STP

The synapse correlation circuit uses the address enable signal `addren` as a prespike. This makes sense since the synapse has to check whether the address sent out by the synapse driver matches its configured address. With the dependency of STDP amplitudes on the number of columns with matching event addresses, we already verified that synapses check the addresses. In this section, we want to enable STP while measuring the correlation amplitudes. We do not expect any influence of the STP states on the STDP measurements, since the `dacen` pulse does not have any influence on triggering the measurement.

The experiment we conduct uses the same chip configuration as for the measurement of the STP ramps (section 4.6). This means we configure STP on the synapse drivers to be renewing, so that V_{STP} reaches the applied V_{charge} voltage at the first spike. V_{recover} is set 20 mV above that, with recovery turned off. Before sending the 6 correlated spikes, 10 spikes are sent to the configured address in order to reach $V_{\text{STP}} = V_{\text{charge}}$. This is especially important during the STDP

measurement: we can not throw away the first 10 spikes during evaluation, like we did when measuring amplitudes before. Even the first spike has to be in the correct STP state.

Since we know from the ramp measurements how the length of the `dacen` pulses depends on the V_{STP} settings for every driver, we can now plot the amplitude of the STDP measurements as a function of the `dacen` time. For the correlation amplitude, we use the measurement at a time difference of +100 FPGA clock-cycles, corresponding to causal correlation. Since we measure data from the whole first column of synapses and we have data for every individual synapse driver concerning the pulse lengths, we can plot a data point for every single synapse of the first column. Since the absolute STDP amplitude measurements are scattered a lot (see error bars in figure 5.6), we plot the ratio of the measured amplitude to the original amplitude with STP disabled. We sweep V_{charge} in steps of $50\,\text{LSB}$, equivalent to changes of about $100\,\text{mV}$. The settings correspond to every fifth data point plotted in figure 4.12. However, since the time between spikes is 40 times larger compared to the recording of the ramps, the `dacen` times can be slightly higher in this experiment due to the increased influence of leakage currents, which are pulling up V_{STP} [Weis, 2017].

Loading a set of `offset` parameters that was obtained using the amplitude-based calibration, we run the experiment. The results are plotted in figure 5.7. Every data point represents a single synapse, the different V_{charge} settings are indicated by colors. We can see point clouds representing measurements with identical parameters. For low lengths of the `dacen` pulses, the horizontal spread increases. This is caused by some synapse drivers still outputting longer pulses than they should when we assume the ramps to be perfectly linear. This is the exact same effect that we saw in figure 4.12 for low V_{STP} settings. For the vertical spread, the amplitudes of the STDP measurement do not vary significantly, just as we expected. Correlation amplitudes are not lower when synaptic depression occurs, no prespikes are missed. Repeating this measurement shows that the seemingly lower amplitudes at the two lowest V_{charge} settings are purely incidental.

Therefore, we conclude that both synaptic plasticity topics discussed in this thesis, STDP correlation measurements and STP, work fine in this experiment.

6 Discussion and Outlook

In this thesis, synaptic plasticity mechanisms, STP and STDP, were tested and verified to work properly on HICANN-DLS 3. For the depletion of neurotransmitters, the utilization of synaptic efficacy ranges from 0.27 to 0.74 for the tested chips. The recovery of neurotransmitters, which increases the amplitudes of transmitted action potentials, covers three orders of magnitude: time constants range from $2.38\,\mathrm{ms}$ to $2120\,\mathrm{ms}$ biological time on the tested chip. These parameters allow for a wide range of configurations: STP can recover fast enough to be used in recurrent networks. Typical networks spike at rates lower than $100\,\mathrm{Hz}$ [Amit and Brunel, 1997], which equates to inter-spike-intervals of $10\,\mathrm{ms}$ biological time. Besides the usage of inhibitory and excitatory connections at the same time, STP will provide an additional method to stabilize the spike rates in such applications [Bill et al., 2010]. Concerning recovery, a dependency of time constants on the address is visible (figure 3.9). STP works as expected from simulations, both concerning the effects of configured parameters but also regarding this address pattern during recovery.

Before using STP, the synapse drivers need to be calibrated. The output amplitudes are encoded by the length of a `dacen` pulse reaching the synapses. This pulse is generated by comparing the voltage V_{STP}, representing the current state of neurotransmitters, with a ramp. The comparators located in different instances of the circuit are not identical, so amplitudes of different drivers would vary in a fixed pattern if they were not calibrated. We add an offset voltage to the ramp, fixing the effects of this mismatch. The offsets have to be determined individually for every driver. Until now, an external ADC was used in order to read out the amplitudes. This takes 6 minutes on DLS 3 [Weis, 2017], and the runtime scales linearly with the number of synapse drivers.

In this thesis, a new readout mechanism is tested. The amplitudes at the synaptic input of neurons are measured directly in the neuron, using only its core circuitry. By design, the neuron integrates inputs on the membrane capacity. However, in the leaky integrate-and-fire model, leakage introduces a strong nonlinearity between spike rates and PSP amplitudes. By eliminating the leakage current, membrane potentials rise linearly and neuron spike rates can be used as a linearly mapped observable representing the received synaptic input. A big advantage of this readout method is that neuron spike events can be stored in local counters in every neuron. This increases the number of available parallel readout channels to the number of neurons. Currently, after the experiment, the stored counts are read out using the FPGA. This approach allows calibrating the HICANN-DLS 3 prototype in approximately $30\,\mathrm{s}$. The runtime can be reduced further by adding parallel readout: since the HICANN-X chip will feature twice as many neurons as synapse drivers, the number of required runs to acquire every driver's data can be cut in half, using all neurons in parallel. This is possible while maintaining 16 neurons to average spike rates from, like it is done in the presented experiments.

The readout of spike rates unfortunately introduces higher statistical variations than the conventional method using the ADC. This means that the calibration script may regularly find an `offset` setting that is $1\,\mathrm{LSB}$ off the optimal setting determined by the ADC readout. Thus, the standard deviation of different synapse drivers' amplitudes is greater using the fast spike rate method compared to the slow ADC method. After a calibration using spike rates on the tested chip, the amplitude mismatch is approximately $1\,\mathrm{mV}$, which equates to $3\,\%$ of the amplitudes at

medium depression. All results are compared at the end of section 4.4. To put it in a nutshell, the remaining mismatch between drivers is higher using spike rate readout, but about 5 times lower than without calibration. The results are still very usable. The mismatch is about the same as the mismatch between individual synapses, which is $2.6\,\%$. Since the route of an action potential travelling between neurons includes the synapse drivers as well as the synapses, their mismatch effects are added squared. Therefore, while having lower individual mismatch numbers is always better, reducing one far below the other yields no advantages.

Future steps include implementing the calibration using the spike rate readout on the system's on-chip microprocessor, the PPU. It can access the neuron spike counters and write settings into synapse drivers. This was not tested during this thesis. Without frequent communication to the host computer, an entirely local calibration consumes less power and resources. Additionally, the calibration needs to be integrated into the calibration stack for neuron parameters. This is especially important since for the spike rate readout to work reliably, we need the synaptic input to be configured correctly. This also replaces the parameter search based on manually found values, which was used during this thesis.

The actual results after calibrating STP using spike rates are shown in figure 4.9, using a more realistic range of amplitudes than during calibration. The mismatch between drivers and synapses observed during these experiments should not be obstructive regarding the emulation of neural networks, or especially learning algorithms: the latter should be able to compensate a systematic pattern in connection strengths by tuning the synapse weight matrix accordingly.

The second synaptic plasticity mechanism covered by this thesis is STDP, which is an important feature concerning learning [Bi and Poo, 2001]. Synaptic weights are changed based on correlation of spikes, where causal and anticausal correlation between synaptic inputs and neuron spike events are responsible for both positive and negative weight changes.

In this thesis, the circuit measuring correlation for every single synapse was tested. Varying the time difference between the synaptic input and the spiking of the neuron, correlation measurements should show the expected exponential dependency. This was indeed the result, however, we observed a strong asymmetry between anticausal and causal measurements, with the causal amplitudes being smaller by a factor of 3 to 4. After observing a dependency of the asymmetry on the number of listening synapses, we found an insufficient voltage supply combined with the artificial test-scenario during characterization, where all synapses receive input simultaneously, to be causing the asymmetry. The voltage can be stabilized by placing a capacitor next to the chip, which holds sufficient charge. Resetting the measurement capacitors of all synapses at the same time no longer causes asymmetric results. Therefore, a nice symmetric plot showing the characterization of correlation amplitudes is the result of this thesis (figure 5.6). Asymmetric correlation measurements should not be a problem anymore, the amplitudes show the expected exponential function on a configurable timescale.

Together with the PPU being able to process complex algorithms in order to tune synaptic weights for successful learning, the system provides ideal requirements for complex experiments. This is why we look forward to see people using the upcoming HICANN-X chip, that is based on the same circuitry concerning STDP, for their network emulations. Concerning STP, the implementation of the comparator ramp generation will be different, including the calibration bits. However, the function of the configurable parameters will stay the same. Therefore, results from this thesis should still be applicable. With many problems solved, HICANN-X will not only be a larger chip targeted at biological use cases, but also an improvement over the current hardware available as prototypes.

A Deduction of utilization

In this appendix section, formula 3.2 will be deducted. It allows calculating U_{SE} from a fit to regular spiking when the recovery time constant is known. The deduction is based on personal communication with Sebastian Billaudelle in September 2017.

Let a_i be a spike amplitude. The amplitude of the following spike a_{i+1}, which is sent after a time difference Δt, is subject to utilization (with a parameter U_{SE}) and recovery (with a time constant τ_{rec}). Assuming a depressing configuration, the amplitude gets smaller proportional to utilization:

$$\Delta a_{i,\mathrm{util}} = -a_i \cdot U_{\mathrm{SE}}. \tag{A.1}$$

The recovery has a positive influence on the amplitude:

$$\Delta a_{i,\mathrm{rec}} = (1 - (a_i - a_i \cdot U_{\mathrm{SE}})) \cdot \left(1 - e^{-\frac{\Delta t}{\tau_{\mathrm{rec}}}}\right). \tag{A.2}$$

For the amplitude of the following spike a_{i+1}, this gives

$$\begin{aligned}
a_{i+1} &= a_i + \Delta a_{i,\mathrm{util}} + \Delta a_{i,\mathrm{rec}} && \text{(A.3)}\\
&= a_i - a_i \cdot U_{\mathrm{SE}} + (1 - (a_i - a_i \cdot U_{\mathrm{SE}})) \cdot \left(1 - e^{-\frac{\Delta t}{\tau_{\mathrm{rec}}}}\right) && \text{(A.4)}\\
&= (1 - U_{\mathrm{SE}}) \cdot a_i + (1 - (1 - U_{\mathrm{SE}}) \cdot a_i) \cdot \left(1 - e^{-\frac{\Delta t}{\tau_{\mathrm{rec}}}}\right) && \text{(A.5)}\\
&= (1 - U_{\mathrm{SE}}) \cdot a_i \cdot \left(1 - \left(1 - e^{-\frac{\Delta t}{\tau_{\mathrm{rec}}}}\right)\right) + \left(1 - e^{-\frac{\Delta t}{\tau_{\mathrm{rec}}}}\right) && \text{(A.6)}\\
&= a_i \cdot (1 - U_{\mathrm{SE}}) \cdot (1 - \alpha) + \alpha, && \text{(A.7)}
\end{aligned}$$

with $\alpha := 1 - e^{-\frac{\Delta t}{\tau_{\mathrm{rec}}}}$.

Applying the result to another spike a_{i+2} yields

$$\begin{aligned}
a_{i+2} &= a_{i+1} \cdot (1 - U_{\mathrm{SE}}) \cdot (1 - \alpha) + \alpha && \text{(A.8)}\\
&= [a_i \cdot (1 - U_{\mathrm{SE}}) \cdot (1 - \alpha) + \alpha] \cdot (1 - U_{\mathrm{SE}}) \cdot (1 - \alpha) + \alpha && \text{(A.9)}\\
&= [a_i \cdot \beta + \alpha] \cdot \beta + \alpha && \text{(A.10)}\\
&= a_i \cdot \beta^2 + \alpha + \alpha \cdot \beta, && \text{(A.11)}
\end{aligned}$$

with $\beta := (1 - U_{\mathrm{SE}}) \cdot (1 - \alpha)$.

Inductively, this yields a series for the i-th spike amplitude:

$$a_i = a_0 \cdot \beta^i + \alpha \cdot \sum_{j=0}^{i-1} \beta^j \tag{A.12}$$

$$= a_0 \cdot \beta^i + \alpha \cdot \frac{1-\beta^i}{1-\beta} \tag{A.13}$$

$$= \left(a_0 - \frac{\alpha}{1-\beta}\right) \cdot \beta^i + \frac{\alpha}{1-\beta}. \tag{A.14}$$

During regular spiking, this allows fitting an exponential function $A \cdot e^{-\frac{t}{\tau_{\text{dep}}}} + C$ to the spike amplitudes. The "time constant" of the depression is τ_{dep}. We can now set β^i equal to the fitted exponential function and insert the definitions of α and β again:

$$\beta^i = e^{-\frac{\Delta t \cdot i}{\tau_{\text{dep}}}} \tag{A.15}$$

$$\Leftrightarrow \quad (1 - U_{\text{SE}}) \cdot (1 - \alpha) = e^{-\frac{\Delta t}{\tau_{\text{dep}}}} \tag{A.16}$$

$$\Leftrightarrow \quad (1 - U_{\text{SE}}) \cdot \left(e^{-\frac{\Delta t}{\tau_{\text{rec}}}}\right) = e^{-\frac{\Delta t}{\tau_{\text{dep}}}} \tag{A.17}$$

$$\Leftrightarrow \quad U_{\text{SE}} = 1 - \frac{e^{-\frac{\Delta t}{\tau_{\text{dep}}}}}{e^{-\frac{\Delta t}{\tau_{\text{rec}}}}}. \tag{A.18}$$

This is equation 3.2.

Bibliography

S. A. Aamir, Y. Stradmann, P. Müller, C. Pehle, A. Hartel, A. Grübl, J. Schemmel, and K. Meier. An accelerated LIF neuronal network array for a large scale mixed-signal CMOS architecture. Submitted to IEEE Transactions on Circuits and Systems I: Regular Papers, 2018.

D. J. Amit and N. Brunel. Dynamics of a recurrent network of spiking neurons before and following learning. *Network: Computation in Neural Systems*, 8(4):373–404, 1997.

G.-q. Bi and M.-m. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, 18(24): 10464–10472, 1998.

G.-q. Bi and M.-m. Poo. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual review of neuroscience*, 24(1):139–166, 2001.

J. Bill, K. Schuch, D. Brüderle, J. Schemmel, W. Maass, and K. Meier. Compensating inhomogeneities of neuromorphic VLSI devices via short-term synaptic plasticity. *Frontiers in computational neuroscience*, 4:129, October 2010.

S. Billaudelle. Characterisation and calibration of short term plasticity on a neuromorphic hardware chip. Bachelor's thesis, Universität Heidelberg, 2014.

S. Billaudelle. Design and implementation of a short term plasticity circuit for a 65 nm neuromorphic hardware system. Master's thesis, Universität Heidelberg, 2017.

R. Brette and W. Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of neurophysiology*, 94(5):3637–3642, 2005.

D. A. Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, 2005. ISSN 0028-3878. doi: 10.1212/01.WNL.0000166914.38327.BB. URL `http://n.neurology.org/content/64/12/2004`.

S. Friedmann, J. Schemmel, A. Grübl, A. Hartel, M. Hock, and K. Meier. Demonstrating hybrid learning in a flexible neuromorphic hardware system. *IEEE Transactions on Biomedical Circuits and Systems*, 11(1):128–142, 2 2017. ISSN 1932-4545. doi: 10.1109/TBCAS.2016.2579164.

S. B. Furber, D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple, and A. D. Brown. Overview of the SpiNNaker system architecture. *IEEE Transactions on Computers*, 62(12):2454–2467, 12 2013. ISSN 0018-9340. doi: 10.1109/TC.2012.142.

D. O. Hebb et al. *The organization of behavior: A neuropsychological theory*. New York: Wiley, 1949.

M. Hennig. Theoretical models of synaptic short term plasticity. *Frontiers in Computational Neuroscience*, 7:45, 2013. ISSN 1662-5188. doi: 10.3389/fncom.2013.00045. URL `https://www.frontiersin.org/article/10.3389/fncom.2013.00045`.

M. Hock, A. Hartel, J. Schemmel, and K. Meier. An analog dynamic memory array for neuromorphic hardware. In *2013 European Conference on Circuit Theory and Design (ECCTD)*, pages 1–4, 9 2013. doi: 10.1109/ECCTD.2013.6662229.

E. Jones, T. Oliphant, and P. Peterson. SciPy: open source scientific tools for Python, 2014.

N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 1–12. ACM, 2017.

G. Kiene. Mixed-signal neuron and readout circuits for a neuromorphic system. Master's thesis, Universität Heidelberg, 2017.

S. Kunkel, G. Masumoto, T. Fukai, J. M. Eppler, H. E. Plesser, J. Igarashi, M. Diesmann, A. Morrison, M. Schmidt, M. Helias, et al. Supercomputers ready for use as discovery machines for neuroscience. In *10th Meeting of the German Neuroscience Society*. Computational and Systems Neuroscience, 2013. FZJ-2013-03827.

W. W. Lytton. *From computer to brain: foundations of computational neuroscience*. Springer Science & Business Media, 2007.

K. Meier. A mixed-signal universal neuromorphic computing system. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 4.6.1–4.6.4, 12 2015. doi: 10.1109/IEDM.2015.7409627.

L.-G. Pang, K. Zhou, N. Su, H. Petersen, H. Stöcker, and X.-N. Wang. An equation-of-state-meter of quantum chromodynamics transition from deep learning. *Nature Communications*, 9(1):210, 2018.

G. E. Pugh. *The biological origin of human values*. Basic Books, New York, 1977.

W. G. Regehr. Short-term presynaptic plasticity. *Cold Spring Harbor perspectives in biology*, 4(7): a005702, 2012.

J. Schemmel, A. Grübl, K. Meier, and E. Müller. Implementing synaptic plasticity in a VLSI spiking neural network model. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1–6, 2006. doi: 10.1109/IJCNN.2006.246651.

J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on*, pages 1947–1950. IEEE, 2010.

J. Sjöström and W. Gerstner. Spike-timing dependent plasticity. *Scholarpedia*, 5(2):1362, 2010. doi: 10.4249/scholarpedia.1362. revision #184913.

C. F. Stevens and J. F. Wesseling. Augmentation is a potentiation of the exocytotic process. *Neuron*, 22(1):139–146, 1999.

D. Stöckel. Exploring collective neural dynamics under synaptic plasticity. Master's thesis, Universität Heidelberg, 11 2017.

Y. Stradmann. Characterization and calibration of a mixed-signal leaky integrate and fire neuron on HICANN-DLS. Bachelor's thesis, Universität Heidelberg, 2016.

M. Tsodyks and S. Wu. Short-term synaptic plasticity. *Scholarpedia*, 8(10):3153, 2013. doi: 10.4249/scholarpedia.3153. revision #182489.

M. V. Tsodyks and H. Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences*, 94 (2):719–723, 1997.

J. Weis. Testing of a neuromorphic short term plasticity circuit. Internship report, Universität Heidelberg, 2017.

T. Wunderlich. Synaptic calibration on the HICANN-DLS neuromorphic chip. Bachelor's thesis, Universität Heidelberg, 2016.

R. S. Zucker and W. G. Regehr. Short-term synaptic plasticity. *Annual Review of Physiology*, 64 (1):355–405, 2002. doi: 10.1146/annurev.physiol.64.092501.114547. URL `https://doi.org/10.1146/annurev.physiol.64.092501.114547`. PMID: 11826273.

Acknowledgements

At first, I want to thank Prof. Dr. Karlheinz Meier for giving me the opportunity to work in the Electrionic Vision(s) group. Researching neuromorphic hardware would not be done at Heidelberg without you. Also thanks to Dr. Johannes Schemmel, keeping the group together, always waiting for new measurement results to be ready. Thank you for sharing your ideas of how to improve an experiment and all your knowledge about the hardware.

The persons that had to explain by far the most are my supervisors Sebastian Billaudelle and Yannik Stradmann. Despite having only little time with the tapeout of HICANN-X coming up, you calmly explained the circuits implemented, the parameters available, the methods to be used, and kept me motivated. Of course my thanks goes to the whole hardware office including Gerd Kiene, Korbinian Schreiber and Christian Pehle who helped me out when problems occurred, but also the whole group, who regularly share their thoughts, with Benjamin Cramer, David Stöckel and Eric Müller being frequent visitors.

Thanks to Sebastian Billaudelle, Yannik Stradnamm, Benjamin Cramer, David Stöckel, Timo Wunderlich, Arne Emmel, Sebastian Frank and Mathis Römer for proofreading this document and providing many helpful comments. Thanks to all my friends and classmates and, of course, my family for your support.

Statement of Originality (Erklärung)

I certify that this thesis, and the research to which it refers, are the product of my own work. Any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

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

Heidelberg, February 26, 2018

.....
(signature)