Department of Physics and Astronomy

University of Heidelberg

Bachelor Thesis

in Physics.

submitted by

Aron Leibfried

born in Bad Friedrichshall, Germany

August 2018

On-chip calibration of analog neuromorphic circuits

This Bachelor Thesis has been carried out by

Aron Leibfried

at the

Kirchhoff-Institute for Physics

Ruprecht-Karls-Universität Heidelberg

under the supervision of

Prof. Dr. Karlheinz Meier

Abstract

In the framework of this thesis a collection of approaches is investigated to calibrate various analog neuromorphic circuits, that are part of the HICANN-DLSv3 ASIC implemented in a 65 nm process. Only resources and observables available to the on-chip Plasticity Processing Unit (PPU) are used.

The chip contains 32 neurons based on the Adaptive Exponential Integrate-and-Fire model (AdEx), whose parameters are subject to mismatch. Some of the algorithms calibrating the analog circuitry are based on the neurons' spike events, while others make use of the parallel on-chip CADC. Furthermore a calibration for the pre-synaptic synapse drivers is presented, which implement Short-Term Plasticity (STP).

The different approaches are investigated with respect to their precision, runtime, and scalability, especially in the light of future chip generations which will feature a larger number of neuromorphic circuits. All calibration algorithms presented in this thesis can be executed on the order of seconds for a whole chip and – due to their scalable nature – their runtime is expected to stay approximately constant even for an increased number of neurons or synapse drivers, respectively.

Zusammenfassung

Im Rahmen dieser Arbeit wurden verschiedene Methoden untersucht um neuromorphe Schaltungen zu kalibrieren, wie sie auf dem HICANN-DLSv3 Prototypen realisiert sind. Lediglich die beobachtbaren Größen und Rechenkapazitäten des auf dem Chip implementierten Prozessors (PPU) werden genutzt.

Der Chip enthält 32 Neuronen, die auf dem Adaptive-Exponential-Integrate-and-Fire Modell (AdEx) basieren und deren Parameter Fertigungstoleranzen unterliegen. Um die analogen Schaltungen zu kalibrieren, nutzen Algorithmen die Spike-Events von Neuronen oder den implementierten CADC. Desweiteren wird eine Kalibration der Synapsentreiber präsentiert, durch welche Short-Term Plasticity (STP) realisiert wird.

Die verschiedenen Ansätze werden bezüglich ihrer Präzision, Laufzeit und Skalierbarkeit evaluiert. Der Grund sind vor allem zukünftige Chip Generationen, welche deutlich mehr neuromorphe Schaltkreise enthalten werden. Alle Kalibrationsalgorithmen, die in dieser Arbeit präsentiert werden, benötigen wenige Sekunden um einen gesamten Chip zu kalibrieren. Aufgrund ihrer Skalierbarkeit sollte die Laufzeit selbst für eine erhöhte Anzahl von Neuronen und Synapsentreibern ungefähr auf dem selben Level bleiben.

Contents

1 Introduction

Modern science would not be possible without traditional computers, as they are able to execute large-scale simulations or managing Petabytes of data like in the ATLAS Detector at CERN [Borodin et al., 2015]. State of the art supercomputers like the *Summit* are able to perform $2 \cdot 10^{17}$ floating-point operations per second [Feldman, 2018]. One can neglect using the human brain for such arithmetic operations, as a normal person needs quite some time to calculate the product of two floating numbers. On the other hand, the brain excels at cognitive tasks like face recognition or understanding a language. But with complex algorithms also cognitive tasks like traffic sign recognition [Stallkamp et al., 2012] can be done on computers. However they are limited to a small number of tasks and are not even close to the flexibility of the brain. Also the power consumption of the brain is much smaller, which consumes around 20 W, while supercomputers used to perform such algorithms are consuming power in the order of megawatts. Also traditional computers, which are used for personal use are consuming several hundred watts. That is the reason why it is desireable to create machines with the computational abilities of the brain combined with its low power consumption [Meier, 2017].

Several projects with different approaches are currently running to implement the behavior of the brain on hardware. The *SpiNNaker* platform is a purely digital approach, which is optimized for highly parallel tasks to simulate the behavior of the brain [Furber et al., 2013]. Another approach is to emulate the structure of the brain on an integrated analog circuit, like it is done on the *BrainScaleS* system [Schemmel et al., 2010]. These systems make it possible for brain researchers to conduct experiments on hardware.

In the Electronic Vision(s) group at the Kirchhoff-Institute for Physics Heidelberg we are currently developing a successor for the *BrainScaleS* system, called High Input Count Analog Neural Network (HICANN) with Digital Learning System (DLS). The HICANN-DLS is an Application-Specific Integrated Circuit (ASIC) which contains analog and digital parts. Because of its prototype status it contains a reduced number of all relevant building blocks, but it is fully functional and can be used for experiments.

This thesis will present new calibration methods for the third prototype of these chips, called HICANN-DLSv3, using the general-purpose on-chip processor, called *Plasticity Processing Unit* (PPU). A calibration of the chip is necessary due to manufacturing tolerances. It is evaluated which observables can be used to calibrate the neurons on the chip. Also the calibration of Short Term Plasticity (STP), which is suspected to play an important role in neural information filtering, which was previously implemented on the host computer [Weis, 2018], will be extended and ported to the PPU.

The motivation behind this thesis is to find algorithms which reduce the mismatch between the respective circuits' parameters. It would be desireable to find highly scalable algorithms that require a short runtime. Implementing these algorithms on the PPU reduces the communication overhead to the host system. Furthermore, for a future wafer-scale system, the amount of PPUs will scale with the number of neurons, such that the runtime should stay approximately constant for these larger systems.

2 Principles

2.1 Biological background

The computational power of the human brain arises from small cells in the brain which are called neurons and their connections. A neuron consists of three parts: dendrites, a soma and an axon. The dendrites can be understood as input of the neuron. The cell soma connects to the dendrites and forms a membrane, which receives signals from the dendrites. Once the a critical membrane potential is reached, an action potential is created and send along the axon [Eyzaguirre and Kuffler, 1955]. Via synapses the output axon is connected to the dendrites of other neurons. Usually synapses transfer their signals via neurotransmitters [Pereda, 2014]. In total it is estimated that the human brain contains 10 - 22 billion neurons and with an estimate of 20000 synapses per neuron [Dicke and Roth, 2016], this would add up to 200 - 440 trillion synapses.

2.2 Leaky Integrate and Fire model

The Leaky Integrate and Fire model (LIF) is one of the most suitable neuron models for analog hardware, because it is based on first order differential equations, as found in basic analog circuitry. It can be extended to the Adaptive Exponential Integrate-and-Fire model (AdEx) [Brette and Gerstner, 2005] to describe the action potentials in a more realistic way. It is also able to reproduce biological firing patterns such as adapting, bursting, and delayed spiking. The model is described by two differential equations

$$C\frac{\mathrm{d}V_\mathrm{m}}{\mathrm{d}t} = -g_\mathrm{l}\left(V_\mathrm{m} - E_\mathrm{l}\right) + g_\mathrm{l}\Delta_\mathrm{t}\exp\left(\frac{V_\mathrm{m} - V_\mathrm{t}}{\Delta_\mathrm{t}}\right) - w + I, \tag{1}$$

$$\tau_w\frac{\mathrm{d}w}{\mathrm{d}t} = a\left(V_\mathrm{m} - E_\mathrm{l}\right) - w \tag{2}$$

and a reset condition which sets V_m to a reset potential V_reset for the duration of the refractory time τ_ref, when it is crossing the threshold potential V_thresh.

In this case V_m describes the membrane potential and the first term in equation 1 describes the leakage term. The second and third term are the extensions of the AdEx model, together with equation 2. The potential V_t describes the soft threshold for the exponential term, while Δ_t is its slope factor and w marks the adaptation current. The dynamics of the adaptive extension are calculated within the second term. The last term I describes all other currents on the membrane. It includes both excitatory and inhibitory synaptic inputs, the excitatory and inhibitory current, and a direct current stimulus.

2.3 Short Term Plasticity

Processing the synaptic input is called Short Term Plasticity (STP) [Fioravante and Regehr, 2011]. It is based on synaptic changes which are a result of prior synapse activity, lasting to short term effects of couple minutes [Zucker and Regehr, 2002]. A set of two effects can occur at the postsynaptic response, which are called Short Term Depression (STD) and Short Term Facilitation (STF) [Hennig, 2013]. While the postsynaptic input decreases over the course of repeated stimulation for STD,

it leads to higher inputs for STF. Both effects seem to exclude each other, but they can occur at the same time [Hennig, 2013].

A first attempt to describe the synaptic depression was done by Tsodyks and Markram [Tsodyks and Markram, 1997]. The *Tsodyks-Markram model* was improved by dividing synaptic neurotransmitters into three different states, recovered, active and inactive [Tsodyks et al., 1998]. With this the dynamics of the synaptic input over time can be described by three differential equations [Tsodyks and Wu, 2013].

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{E}{\tau_{\text{faciliation}}} + U_{\text{SE}} \cdot \left(1 - E^{-}\right) \cdot \delta\left(t - t_{\text{AP}}\right) \tag{3}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{1 - R}{\tau_{\text{depression}}} - E^{+} \cdot R^{-} \cdot \delta\left(t - t_{\text{AP}}\right) \tag{4}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{I}{\tau_{\text{syn}}} + A \cdot E^{+} \cdot R^{-} \cdot \delta\left(t - t_{\text{AP}}\right) \tag{5}$$

In this case E describes the effective partition and R the recovered one, while I is the current onto the neuron membrane. E and R range between 0 and 1, as it is linked to the amount of neurotransmitters in the respective partition. The time of the action potential is given by t_{AP} and states before an action potential are marked with an upper -, states afterwards with an upper +.

If the synapse is idle, the effective partition E will decay to 0 with the time constant $\tau_{\text{faciliation}}$ while the recovery partition R decays to 1 with it is time constant $\tau_{\text{depression}}$. These time constants are in the order of tens of milliseconds to seconds [Regehr, 2012]. After an event in form of an action potential a fraction U_{SE}, which is called *utilization* is added to the effective partition E. This leds to depression as the recovery partition will shrink due to the enlarged effective partition E^{+}. The actual current onto the neuron membrane is 0 before the action potential. But if such an event occurs, I is given as the product of a maximum amplitude A and the product of E and R, which leads to depression as well as facilitation. At the end it decays back to 0 with a time constant τ_{syn}.

2.4 The HICANN-DLSv3 ANNCORE

The analog circuitry corresponding to the discussed brain model is realized in the Analog Neural Network Core (ANNCORE), which is the heart of the chip. A sketch of the ANNCORE is shown in figure 2.1. The HICANN-DLSv3 contains a total of 32 neurons and 1024 synapses, which are arranged in a 32 x 32 array with a neuron connected at the bottom of each row. So every neuron can receive the signals of 32 synapses in its column. Every synapse has an individual weight and address, which are both 6-bit values.

To realize STP, 16 synapse drivers are added on the left side of the synapse array. Each driver is connected to two synapse rows. Signals from the drivers always contain one of the 64 source addresses, which allows to enable the according synapses.

The current onto the neuron is defined by the signal from the synapse driver and the weight of the synapses. Spiking neurons will send out a digital signal, which can be fed back to the synapse drivers. These signals can reach other neurons by traveling through the synapse array. Drivers can also receive external inputs.

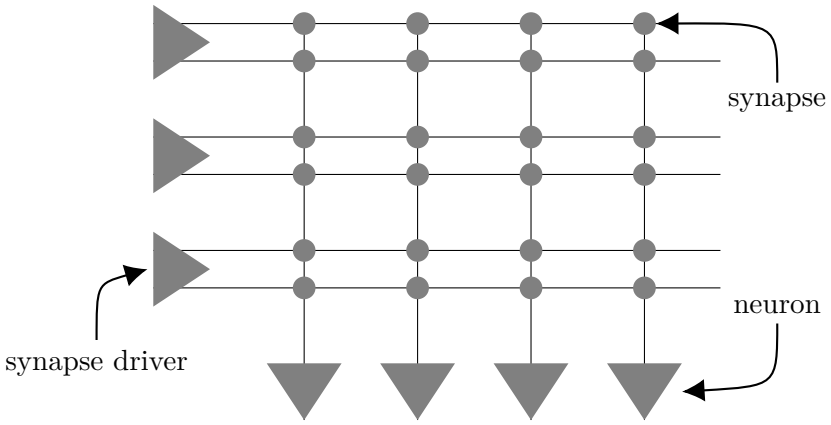


Figure 2.1: ANNCORE of the HICANN-DLSv3. On the left side the synapse drivers which process STP are located, each connected to two synapse rows. Every column of the synapse array is connected to a neuron at the bottom, which can receive signals form all synapses in the column. Figure adapted from [Weis, 2018].

The ANNCORE is sped up by a factor of 10^3 compared to biological time. This gives the possibility to emulate long-term processes in a shorter time. An hour in biological time equals to 3.6 s of chiptime. In this thesis all times are given in the actual chip time, if not specified otherwise.

Tunable voltages and currents which are used in the ANNCORE as parameters are generated in the capacitive memory (capmem) [Hock et al., 2013]. The capmem cells are arranged in an array of 34 columns and 24 rows. Every row consists of 8 voltage cells and 16 current cells, which can be configured with an indivdual 10-bit value called LSB. This led to voltages of 0.2 V - 1.8 V and to currents of 15 nA - 1000 nA [Aamir et al., 2018b]. Every column belongs to a single neuron, which leaves two columns for different global currents and voltages.

2.5 Plasticity Processing Unit (PPU)

The Plasticity Processing Unit (PPU) is a general-purpose processor based on PowerPC architecture with a vector unit to process 16 bytes at one time. A thorough introduction and motivation can be found in [Friedmann et al., 2017]. It can either be programmed in Assembler- or in C-Code. Its general purpose is to change the synaptic weights according to complex algorithms (e.g. STDP) by reading out the Correlation Analog to Digital Converter (CADC). The CADC has two channels for each synapse column and converts an analog voltage to an 8-bit digital value.

The PPU is also able to read and write SRAM memory on the chip. On this memory are the different changeable parameters of the chip saved. So it is possible to read and write these parameters with the PPU, making it possible to use the PPU calibration algorithms aswell. The main goal of this thesis is to test different methods to calibrate parameters with the PPU.

2.6 Neuron implementation

As explained in section 2.2, the differential equations enable an easy implementation on analog hardware. This is achieved by emulating the neurons membrane using a capacitor. While on the older prototype HICANN-DLSv2 just the LIF model was used as neuron model [Aamir et al., 2016], an adaptation circuit and an exponential circuit were added to the neurons of HICANN-DLSv3 [Aamir et al., 2017]. This hardware neuron follows the model as explained in section 2.2 very accurately. A schematic overview can be found in figure 2.2.

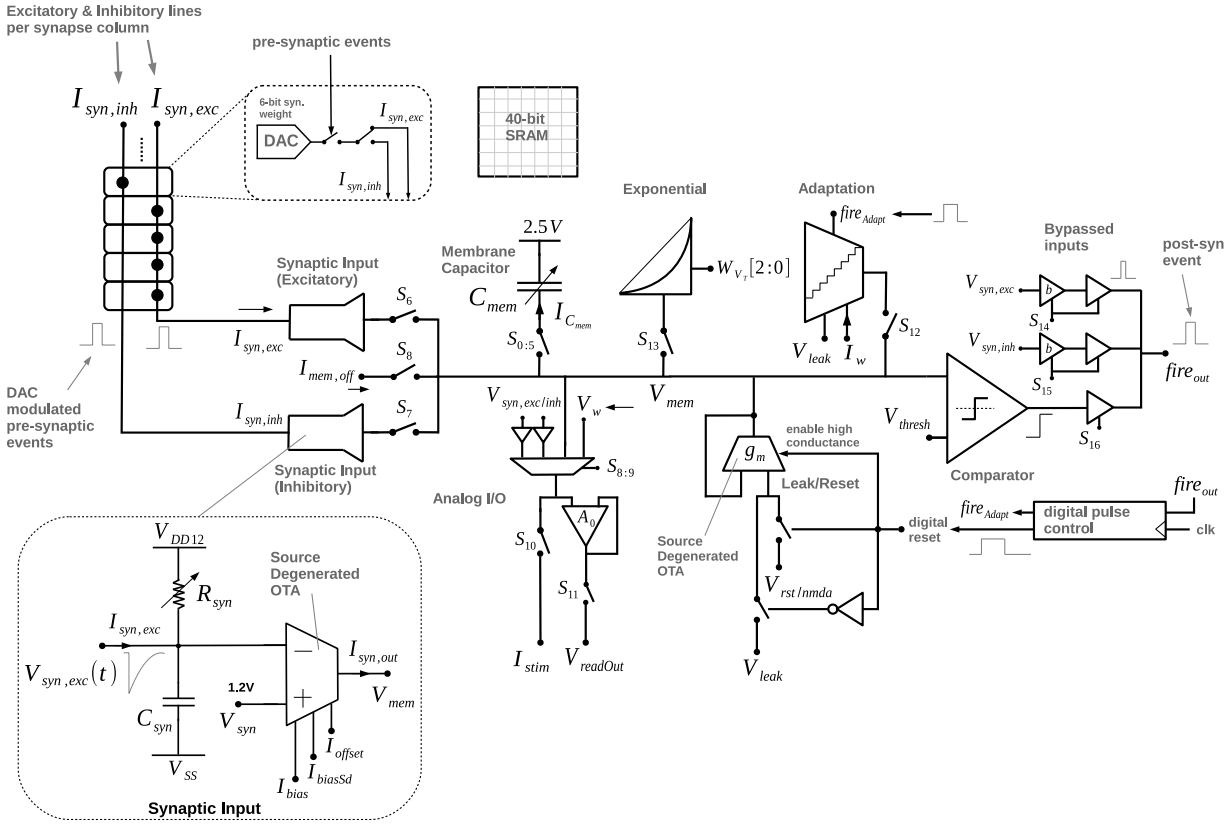


Figure 2.2: Schematic of the neuron circuit realized on HICANN-DLSv3. The membrane is connected to the synaptic inputs on the left side, the exponential circuit, the adaption circuit, the I_{stim} current and to the leakage circuit. On the right side there is a comparator which realizes the threshold in the LIF model. Figure adapted from [Aamir et al., 2018a].

The membrane C_{mem} is a 6-bit tunable capacitor and integrates all input currents. On the left side are the synaptic inputs sketched. Every pre-synaptic event enables a 6-bit DAC which is the synaptic weight and it modulates the amplitude of the pulse event.

The pulses are either processed in the excitatory or inhibitory input. The pulses are integrated on C_{syn} and will decay back exponentially because of the adjustable R_{syn}. The voltage $V_{\text{syn,exc}}(t)$ or $V_{\text{syn,inh}}(t)$ is compared to V_{syn} via an OTA, which outputs a current proportional to the difference of the compared voltages ΔV. Its

slope can be adjusted by I_{bias} which sets the transconductance. To drive positive currents to the membrane for the excitatory input, while the inhibitory input should act as a current sink, the polarity of the OTA is changed for the two different inputs.

The membrane is also connected to the leakage circuit. It is modelling the first term of equation 1 and basically describes a resistor which is connecting the membrane and the leakage voltage V_{leak}.

Also an exponential circuitry and an adaption circuit are connected to the membrane. These extension are modelling the second and third term of equation 1. Also an Analog I/O input is added, which provides debug read-outs as well as the possibility to inject manually a current into the membrane.

On the right side of the sketch is a comparator. The membrane voltage V_{mem} is compared to a threshold voltage V_{thresh}. The comparator outputs a signal if V_{mem} reaches V_{thresh}, resulting in triggering a counter based delay circuit, which is resetting the membrane potential to V_{reset} via the leakage circuit for an adjustable time.

This thesis will discuss different methods of calibrating V_{syn}, I_{bias}, R_{syn} V_{res} and V_{thresh} (compare figure 2.2) using the PPU.

2.7 Short Term Plasticity implementation

Short Term Plasticity is completely processed in the synapse drivers. The synapse driver outputs an address for the synapses and the `dacen` pulse. The charge on the capacitor on the synaptic line will be proportional to the width of the `dacen` pulse as it is the integral of it. Thus higher `dacen` pulses are causing "stronger" input spikes. To modulate STP, the width of the `dacen` pulse is changed according to the level of depression or faciliation. If STP is disabled the `dacen` pulse will reach his maximum duration of 4 ns (with the intended chip clock of 250 MHz).

The state of neurotransmitters is stored on a capacitor as voltage V_{STP} for all available addresses. But with just one capacitor per address just one parameter can be stored, while equation 3 and 4 require two parameters to process depression and faciliation at the same time. Because of this, one have to decide which mode one wants to use.

The synaptic input w received at the neuron is proportional to the recovered partition R for depression and proportional to the inactive partition I for faciliation [Schemmel et al., 2007], what is shown in the following equations:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{I}{\tau_{\text{rec}}} + U_{\text{SE}} \cdot R \cdot \delta\left(t - t_{\text{AP}}\right) \tag{6}$$

$$R + I = 1 \tag{7}$$

$$w \propto \begin{cases} R & \text{for depression} \\ I & \text{for faciliation} \end{cases} \tag{8}$$

As these equations show, faciliation is based on the same equations as depression but with an inverted role. That is the reason why switching between the two modes can be done by inverting the dacen pulse. This results in pulses which get shorter for depression, while they get longer for faciliation.

The STP circuit can be seen in figure 2.3. V_{STP} is stored on the capacitor C_{storage} and gets updated every time an action potential is forwarded. At every update the

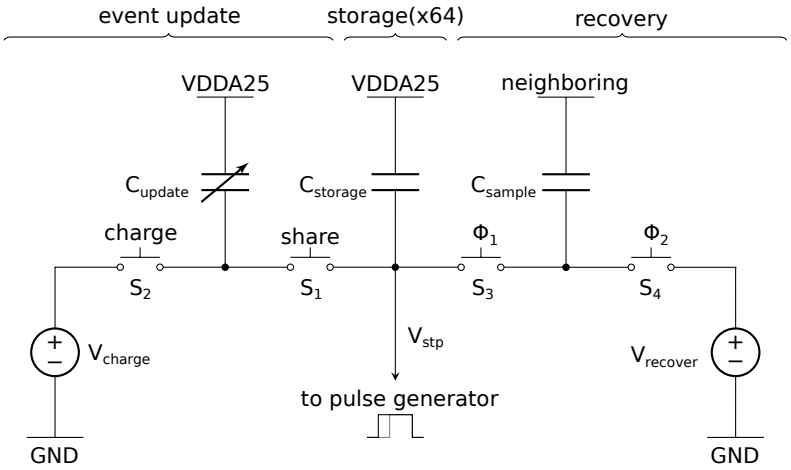


Figure 2.3: Schematic of the STP circuit. The state of the neurotransmitters is stored as a charge on C_{storage}. The left side updates this charge after an action potential and the right side is responsible for recovery. Figure adapted from [Weis, 2018], originally from [Billaudelle, 2017].

charge of C_{storage} and C_{update} are shared by closing S_1 causing V_{STP} to change by

$$V_{\text{STP},f} = V_{\text{charge}} + \left(V_{\text{STP},i} - V_{\text{charge}}\right) \cdot \frac{C_{\text{storage}}}{C_{\text{storage}} + C_{\text{update}}}. \tag{9}$$

During this process S_2 gets opened while it is normally closed to charge C_{update} for a next update of V_{STP}. The ratio of C_{storage} and $(C_{\text{storage}} + C_{\text{update}})$ is the utilization parameter U_{SE}.

V_{STP} also decays exponentially towards V_{recover} with a select-able timing constant τ_{rec}. This is achieved by a pseudo-resistor with a resistance $R = (C_{\text{sample}} \cdot f)^{-1}$, with f as switching frequency. This allows configuring τ_{rec} by changing f. Switches S_3 and S_4 are therefor switched continously with the switching frequency.

A comparator is used to convert V_{STP} into a dacen pulse. This is done by comparing a linear voltage ramp with V_{STP}. An ideal ramp should start at V_{charge} and end at V_{recover} after the maximum dacen pulse width. The dacen pulse starts when the ramp starts rising and will end when the ramp reaches V_{STP}.

Generating the ramp is sketched in figure 2.4. It starts with precharging the ramp capacitor to $V_{\text{precharge}}$ during the first 2 ns, with a global voltage V_{offset}. During another 2 ns an offset is added from a tunable capacitor with a resolution of 4 bit. That is done by charging it with a global voltage V_{zero} and connecting it to the ramp capacitor. The added offset to the ramp depends on the selected capacity and the ramp will be charged to $V_{\text{calibration}}$. After this the ramp gets charged for 4 ns with a constant current I_{ramp}, resulting in a linear rise of the ramp. A detailed overview of this circuit is given in [Billaudelle, 2017].

Because of manufacturing variations of the chip, it needs to be calibrated. The 4-bit capacitor (called `offset` parameter) needs to be calibrated to shift the start value of the ramp to get similar amplitudes for all drivers at similar STP states. This part of this thesis is based on the work of Johannes Weis [Weis, 2018] and will

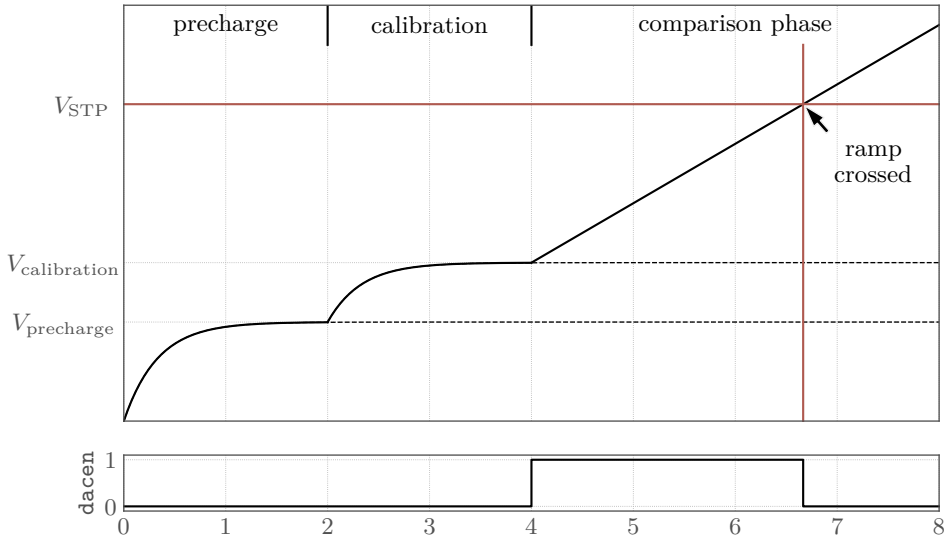


Figure 2.4: Sketch of the voltage ramp which is compared to V_{STP} to model the width of the dacen pulse. In the first 2 ns it gets precharged, following 2 ns of adding an individual offset. Then the ramp is generated by a constant current. Figure adapted from [Weis, 2018].

extend his results including to implement his algorithms to the PPU.

Equation 5 is processed in the neuron as R_{syn} (compare to figure 2.2), which will also be discussed in this thesis.

2.8 Experimental setup

All experiments have been done on a HICANN-DLSv3 setup, which is shown in figure 2.5. It is basically a baseboard, which provides the necessary voltages and currents by using Digital-Analog-Converters (DACs) and giving the possibility to acces pin headers for different analog parameters, e.g. the membrane potential or the synaptic input line. It connects to USB via the FlySpi-Board which contains an FPGA and memory to control experiments in realtime. The DLSv3 chip itself is bonded to a SODIMM module and can be inserted into a socked on the baseboard. A Host-Computer with frickel-dls software installed is necessary to describe experiments in Python.

For this thesis the system clock is set to 400 MHz instead of the intended system clock of 500 MHz because of different bugs reported in [Leibfried, 2018]. As Baseboard "Jack London" was used for this thesis and unless stated otherwise together with Chip 8: "Green Bamboo" (DLSv3.0).

If in this thesis it is meant that the data was collected via an oscilloscope, it means that a *LeCroy WaveSurfer 44Xs* was used together with *LeCroy ZS1000* active probes. Active probes are required for readout to maintain high amplitudes, as most signals are driven off the chip. As sourcemeter a *Keithley 2635B SYSTEM SourceMeter* has been used during this thesis.

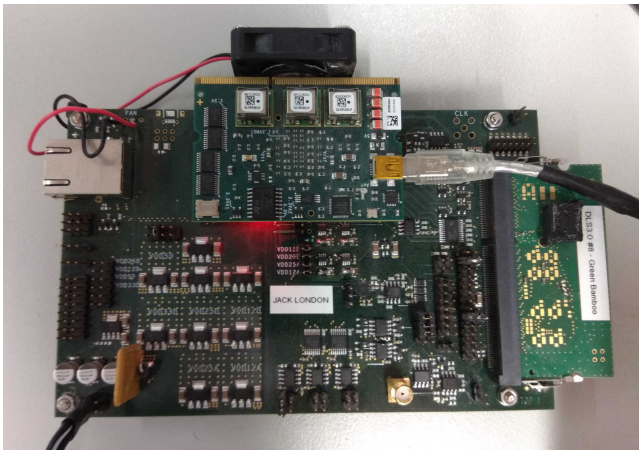


Figure 2.5: Photo of the HICANN-DLSv3 setup which was used during this thesis. The HICANN-DLSv3 chip can be seen on the right and is located below the black cover. Its module is connected to the baseboard. The FlySpiBoard is located in the upper part of the picture and is connected to USB.

2.9 Binary search

Most algorithms in this thesis are based on a binary search. Every parameter which will be calibrated is a digital value which can be saved as binary number. So setting every bit in a single run starting from the most significant bit and comparing the result with the desired result should give the perfect value. An example for a binary search with a 3 bit parameter is given in figure 2.6.

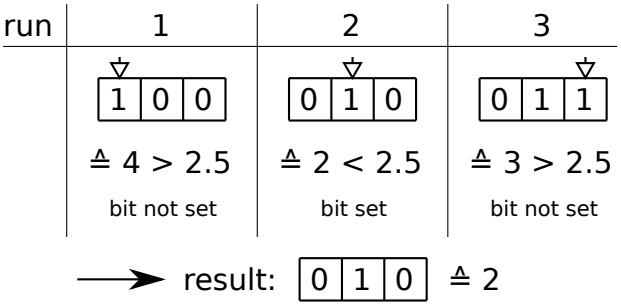


Figure 2.6: Example of a binary search with 3 bit. The goal value is 2.5, so the bit of the first run is not set because it is above 2.5. The second bit is set because now the value is below 2.5. The last bit is not set again.

The example shown in figure 2.6 describes a method to approximate a floating number to an integer number. But this also works for other obervables, which are changing according to the binary settings. So by setting each bit one can compare the observables to the desired one, resulting in a calibrated setting. Of course the translation function of setting-observable must be monotonic rising or falling, otherwise this is not working.

But as the example in figure 2.6 shows, there can be some deviations. An value

of 2.9 would be also assigned to a binary value of 2 with the binary search, while it is better to have a binary value of 3. This can be fixed by adding an additional run to the binary search, which raises or lowers the binary value by one according to the last setting and one tests which setting is better. In the example one would raise the binary value by one and would see that it is better than the old value.

For big binary numbers this should not be a problem. There can be also some mismatch in the observable which makes it hard to find the "perfect" value. For small values however this can make some difference, so for 4 bit values an additional run is really useful to get better results.

Sometime one wants to have a desired range within the 10-bit value. One reason can be a lower runtime. This can be done by using an offset and a smaller binary search. For example if one wants a capmem value of (450 ± 50) LSB it is possible to calibrate this range with 7 runs. The algorithm searches within a range of 128 LSB for 7 runs. To get the start of the calibration 64 LSB have to be subtracted from the 450 LSB. So the starting value is 386 LSB. By doing a 7 run binary search at maximum 127 LSB can be added to this starting value. This results in the the calibration range of 386 LSB to 513 LSB. The desired capmem value of 450 LSB is in the middle of this calibration range. This method is often used during this thesis.

A code snippet of the binary search used in this thesis can be found below. In this case the example with the 7-bit offset to a value of 386 LSB is included. The `measured_value` is measured with a `find_value()` function. The output of this function depends on the `capmem_value`, which was set with `set_capmem_value()`. In every run one bit is set and the `measured_value` is compared to a `mean_value`. If the mean is bigger than the measured value, the bit is not set.

```
1  uint16_t capmem_value = 386;
2  uint16_t binary = 64;  // 64 equals to 0b1000000
3
4  for(uint8_t i=0; i<7; i++) {
5      capmem_value += binary >> i;
6
7      set_capmem_value(capmem_value);
8      measured_value = find_value();
9
10     if(measured_value > mean_value) {
11         capmem_value -= binary >> i;
12     }
13 }
14 set_capmem_value(capmem_value);
```

3 Neuron Calibration

There are many analog parameters in the neuron, as explained in section 2.6. To reduce the mismatch between the different neurons are calibrations necessary. This could be achieved by learning on the one hand, but also by providing a functional starting point with a calibration algorithm. The weight of the synapses are limited to 6-bit and should not be wasted for calibration. So developing a calibration algorithm is important to use the chip.

Different neuron parameters are calibrated within this chapter. The observables available to the PPU are used to test different calibration approaches. The results should be a starting point for calibrating the neurons of the HICANN-DLS prototypes with the PPU. Every algorithm in this chapter is executed on the PPU.

The successor of HICANN-DLSv3 is called HICANN-X. The new system will contain 512 neurons and an algorithm should be able to calibrate all of them in a small runtime. Executing the calibration on the PPU has different advantages. It is located on the chip and once programs are stored on the PPU, it does not need frequent connection to the host computer making it also more energy efficient. Data and commands also do not have to be frequently exchanged making it independent of network delays. HICANN-X was taped-out during this thesis and it is desired to run experiments on it as fast as possible. The HICANN-DLSv3 is perfect for testing the calibration possibilities of HICANN-X because it features a very similar neuron implementation. This should result in a faster availability of calibration algorithms for HICANN-X.

3.1 Neuron calibration via spike rates

3.1.1 Synaptic input reference voltage (1)

As explained in section 2.6 the synaptic inputs are realized by an OTA, whose output current is proportional to the differential input voltage. This is the differential between the voltage of the synaptic line and V_{syn}. V_{syn} has to match the idle voltage of the synaptic line very well, otherwise an offset current is flowing onto or off the membrane. V_{syn} is provided by a capmem cell for each neuron twice, as excitatory and inhibitory inputs need a separate reference voltage. It is designed to have an idle voltage on the synaptic line of 1.2 V, which can vary due to variations caused by manufacturing or due to different temperatures. So the capmem value has to be calibrated that in the idle state the current onto or off the membrane is minimized.

If not specified otherwise the calibration methods presented in this thesis are applicable to both synaptic input circuits. That is because the only difference is the polarity of the OTA, as explained in section 2.6. So the only difference in the algorithms should be a sign. The methods are all tested for the excitatory input because this input will be important for the calibration of STP, which will be explained in section 4.1.

A first attempt to write an algorithm which calibrates the 10-bit value of the capmem cell providing V_{syn} is based on a manual search, which was done by [Weis, 2018]. The basic idea is to disable all possible inputs to the membrane besides the excitatory synaptic input. With an enabled spike comparator, a too low value for V_{syn} should result in a low membrane potential, while a too high value should result

in regular spiking, while the amount of spikes is higher with an higher V_{syn} value. The membrane potential for different settings of V_{syn} is shown in figure 3.1. At the beginning V_{syn} is below the voltage on the synaptic line causing a current off the membrane. The membrane stays at a voltage of 200 mV. With a higher V_{syn} a current onto the membrane is flowing, starting from 0.2 ms resulting in firing. Around 0.6 ms V_{syn} was increased again and more spikes in the same time window can be seen. Membrane potential recorded from neuron 12 on chip 8. The following algorithms should find the capmem value of V_{syn} where the spiking starts.

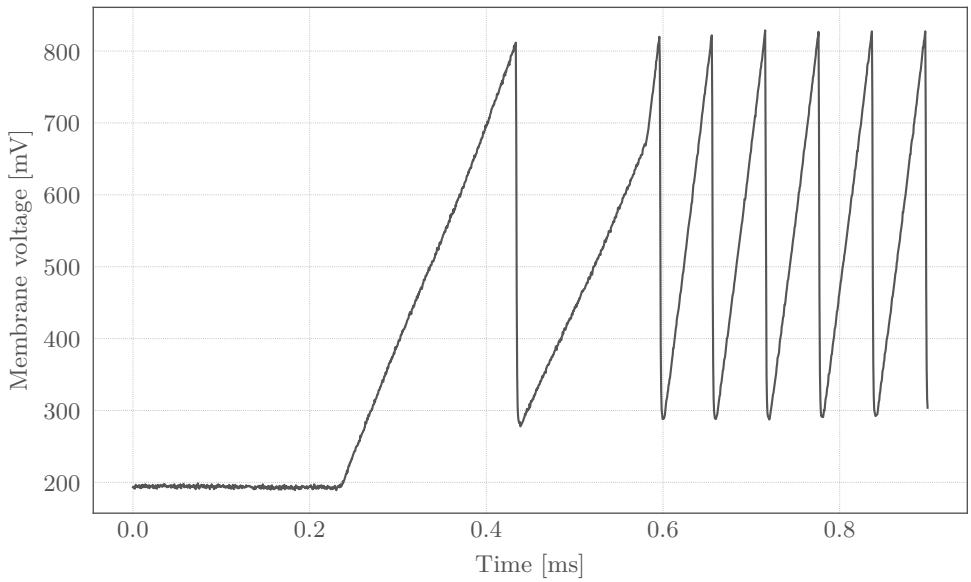


Figure 3.1: Membrane potential for different settings of V_{syn}. V_{syn} is set to 600 LSB at the beginning, 660 LSB from 0.2 ms to 0.6 ms and at the end to 700 LSB.

To register spikes every neuron has an individual 8-bit spike counter [Kiene, 2017] which can be read out and reset with the PPU. Waiting a certain amount of time and reading out the spike counter is therefore a possible observable for this calibration method.

To get an idea how the settings of V_{syn} are depending on the spike rate a plot is made, see figure 3.2. It shows the spike rate of all neurons depending on different settings of V_{syn}. Neuron 2 was removed from the plot because of a defect spike counter [Johannes Weis, 2018, personal communication]. It is falsely counting some spikes several times. Plots which contain the spikes of this neuron can be found in section 3.2.4.

Strictly speaking, the given values are not rates, because they are not divided by time. But as the PPU does not allow floating point operations, all spike rates in this thesis will be given in a total amount of spikes in a certain time window. In figure 3.2 this time window is 5 ms.

As shown in figure 3.2, the neurons start spiking for different settings of V_{syn}. The point where it starts spiking is the perfect value for V_{syn}, because it represents the zero-crossing of the OTA, where it is switching from a current off the membrane to

a current onto the membrane.

The overflow of the 8-bit spike counter can also be observed. So if 256 spikes occured during this timing window, one will readout zero spikes and for more spikes just the modulo of 256.

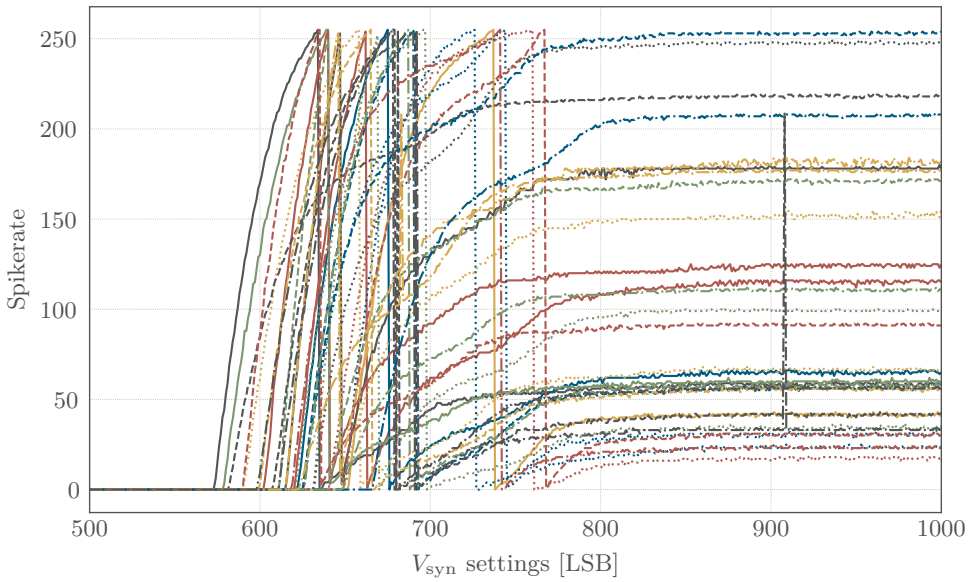


Figure 3.2: Different settings of V_{syn} with the resulting spike rate. Just the excitatory input is enabled.

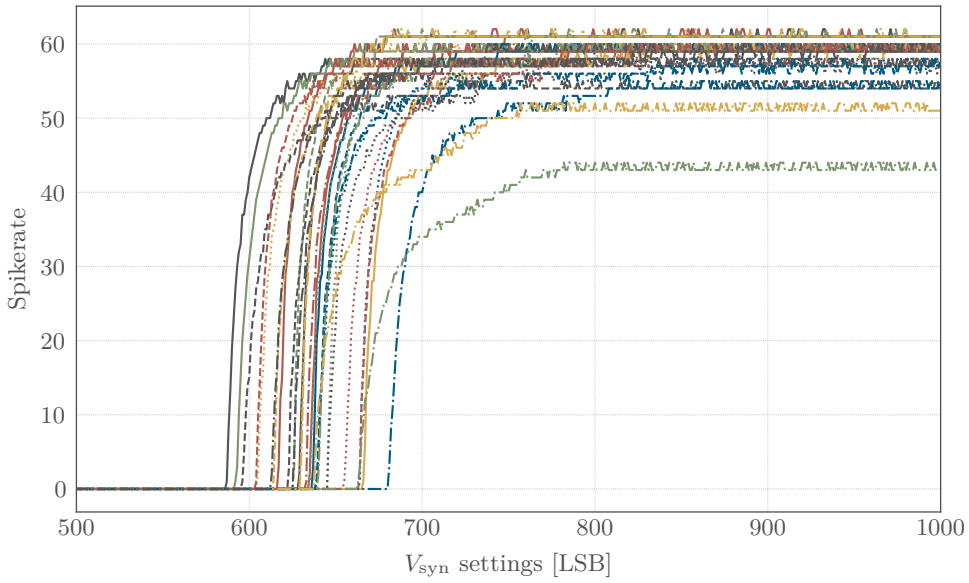


Figure 3.3: Different settings of V_{syn} with the resulting spike rate. Just the excitatory input is enabled and the reset time is increased.

This can be fixed by minimizing the timing window or raising the refractory times.

In figure 3.3 the same plot is shown with an increased refractory time. Now the overflow of the spike counters is prevented and the spike rate depends monotonically on the V_{syn} settings, which makes a binary search possible. A similar plot can be generated by measuring only a time window of 2 ms.

The algorithm is split into two parts: a rough, fast search for the capmem setting of V_{syn}, followed by a slower fine adjusting algorithm.

The rough and fast algorithm is based on a binary search, which takes 10 runs because of the 10-bit capmem value. Each run is setting one bit starting from the most significant bit and ending with the least significant bit. After resetting the spike counter of each neuron one waits 2 ms for spikes and counts all spikes in this time window. This time window was chosen to prevent an overflow of the counters. One will set the bit of the according run if 5 or less spikes are counted. Otherwise the bit will not be set. This algorithm is fast and takes approximately 50 ms in total. But it is not perfectly accurate and with its settings the neurons are regularly spiking, if one is looking at the membrane potential on the oscilloscope. On average the capmem value is 3 LSB too high. That is because for these settings the current onto the the membrane is too small, so the neurons do not reach the five spikes in this time window. However there is still an offset current onto the membrane. So the fine adjusting algorithm has to minimize the offset current by taking the values of the rough algorithm and shift the settings of V_{syn}.

The fine adjust algorithm takes the value from the rough algorithm reduced by 3 LSB. It counts spikes in a certain time window of 1 s. If 6 or more spikes are counted the capmem value will be reduced by 1 LSB and the neuron is marked as `spiked`. If 1 to 5 spikes are counted the neuron is marked as `calibrated` and it will not be changed anymore. If zero spikes are counted two things are possible. If the neuron is not marked as `spiked` it could be that V_{syn} is lower than the synaptic line voltage and therefore a current off the membrane occurs, which can not be detected with this approach. In this case 1 LSB is added to the capmem value. But if the neuron was marked as `spiked` one has found a well calibrated state of V_{syn}, because with one LSB more the neuron will spike. In this case the neuron is also marked as `calibrated`. The whole algorithm will end after 5 runs to set a stopping point. This algorithm takes around 5.1 s runtime.

By using both algorithms together, V_{syn} can be calibrated in under 6 s. It is important to note that this time is dominated by the PPU waiting for spikes. The computational time is small compared to this waiting time and it is possible to wait for all neurons at one time and compute the observables at once. That is the reason why the runtime should not change significantly on HICANN-X. The values which are found with a manual guess in [Weis, 2018] and the values from the algorithm are except for 3 LSB the same which proofs the functionality of this algorithm. Further investigation was done in section 3.1.5 and section 3.2.8 to quantify the calibration results.

3.1.2 Synaptic input current (1)

A second important parameter to calibrate the OTA is I_{bias}, which sets the transconductance $g_{\text{m}}(I_{\text{bias}})$ of the OTA. The charge onto the membrane however is proportional to two different parameters: $g_{\text{m}}(I_{\text{bias}})$ and $\tau_{\text{syn}}(R_{\text{syn}})$ as shown in the

following. The output of the OTA can be determined by

$$I_{\text{syn,out}}(t) = g_{\text{m}} \cdot \Delta V(t). \tag{10}$$

By sending one spike to the synaptic line, the voltage on the synaptic line will be an exponential decaying curve, with a decay time of τ_{syn}. With an calibrated V_{syn}

$$\Delta V(t) \propto \exp\left(-\frac{t}{\tau_{\text{syn}}}\right), \tag{11}$$

while the amplitude depends in this case on the synaptic weights and the STP state. The charge $Q_{\text{syn,out}}$ onto the membrane for one spike is given by

$$Q_{\text{syn,out}} = \int_0^\infty I_{\text{syn,out}}(t)\,\mathrm{d}t \quad \propto \quad g_{\text{m}} \cdot \int_0^\infty \exp\left(-\frac{t}{\tau_{\text{syn}}}\right)\mathrm{d}t = g_{\text{m}} \cdot \tau_{\text{syn}}. \tag{12}$$

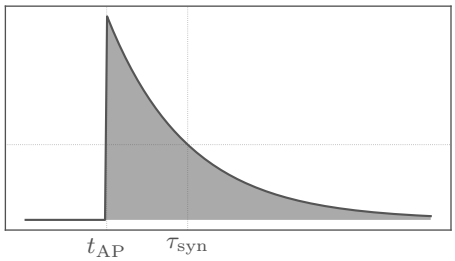


Figure 3.4: An incoming action potential is decaying back to the ground potential with τ_{syn}. On hardware the charge $Q_{\text{syn,out}}$ onto the membran is the integral.

Figure 3.4 shows the charge $Q_{\text{syn,out}}$ as integral. On hardware this is mirrored and the decay is a limited growth to 1.2 V.

The spike rate is related to the charge, which is proportional to g_{m} and τ_{syn}. So it is not possible to calibrate g_{m} and τ_{syn} by sending in spikes with the synapse drivers and counting the spikes of the neuron. One have to find an algorithm which is independent from one of these parameters to do this.

However, for calibrating the STP circuitry it is just important that the spike rate is the same for every neuron with the same input, compare to section 4.1. So it is possible as done in [Weis, 2018] to calibrate just the settings of I_{bias}, while the settings of R_{syn} are fixed for all neurons. This algorithm should find settings for I_{bias} that the spike rate is equal for all neurons, so the spread of the different neuron parameters is calibrated with I_{bias}. This algorithm will be discussed first. In section 3.1.4, another algorithm will be presented to calibrate I_{bias} independent of R_{syn}.

The algorithm is also based on a binary search with 7 runs. A 7-bit offset is added to the capmem value of 61 LSB to get a range from 61 LSB to 188 LSB. These values were chosen, because the used settings range of [Weis, 2018] was 90 LSB to 160 LSB on chip 8. Because this algorithm should run on all different chips a bigger offset value was chosen. A 6-bit offset would just allow a range of 63 LSB to 157 LSB. In

both cases a setting of 125 LSB is in the middle of both ranges.

For this algorithm spikes are send in from synapse driver 0 to all neurons. A total of 5 bursts consisting of 300 spikes each are used to gain a certain amount of statistics, numbers from the STP calibration of [Weis, 2018]. A spike is sent every 10 µs with a pause between the bursts of 500 µs. Before every burst the neuron counter is reset and read out afterwards. First of all, the mean of the spike rate of all neurons is determined to get a target value. As explained above the binary search starts by setting the individual bits starting with the most significant bit. The bit will be set if the mean rate is bigger than the measured rate from the individual neuron.

This calibration just can be used for the excitatory synaptic. It is not possible to calibrate the inhibitory synaptic input with this algorithm because it will take charge off the membrane instead of loading it. However, for calibrating STP just the excitatory input is necessary. That is the reason why this calibration should allow for STP calibration, compare to section 4.1. To quantify the results of this calibration, compare to section 3.1.3.

3.1.3 Suitable neuron configuration for STP calibration

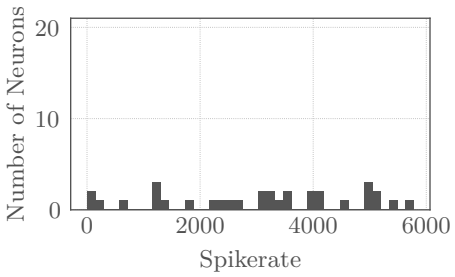
To get a neuron configuration which is suitable for the STP calibration V_{syn} and I_{bias} have to be calibrated. The calibration of V_{syn} should be independent from other neuron parameters and will be the first algorithm to be executed. The calibration of I_{bias} depends on the calibration of V_{syn}, because otherwise there would be a current onto or off the membrane which would lead into a higher/lower spike rate.

The complete calibration starts with the rough algorithm to calibrate V_{syn}, followed by the fine adjusting algorithm. Then I_{bias} is calibrated. Because the operating point of the OTA is shifted by changing I_{bias}, another complete search of V_{syn} is done, followed by another calibration of I_{bias}. This complete calibration takes less than 12 s.

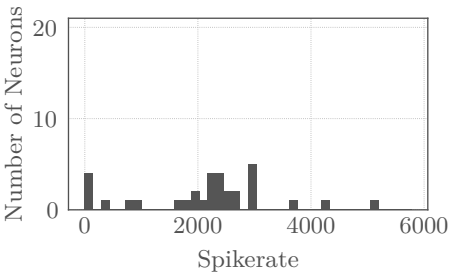
To check the calibrated values, a function which measures the spike rates of the neurons is used. This function uses another synapse driver to cross check the calibration. It uses synapse driver 1. To get a high resolution of the distribution, the function sends a total of 30 burst, each with 300 spikes. The pause between the bursts is 500 µs while a spike is sent every 10 µs. The function adds all spike numbers after a burst of a neuron. This number is called spike rate in this case.

The progress of the different calibration steps can be seen in figure 3.5. Each histogram is showing the spike rate distribution of all neurons for a certain state. Starting with figure 3.5a, which is the uncalibrated state, one can see a vast distribution of the spike rates ranging from zero spikes up to 6000 spikes. Every capmem value is the same for all neurons, showing that manufacturing variances have a big influence.

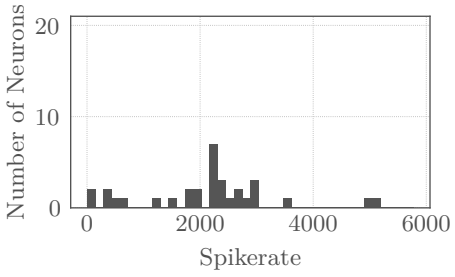
By calibrating V_{syn} roughly, the spike rates look more calibrated but they are not perfect at all as figure 3.5b shows. 22 neurons lie within an area of 1500 LBS. Also the fine adjusting algorithm for V_{syn} does not give a better result (figure 3.5c). But it is expected that the spike rates are randomly distributed, because the charge onto the membrane for a spike is not calibrated. Just currents onto or off the membrane with no spikes are minimized.



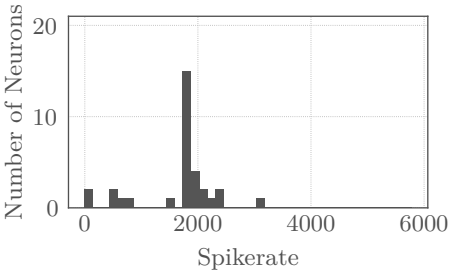
(a) Uncalibrated



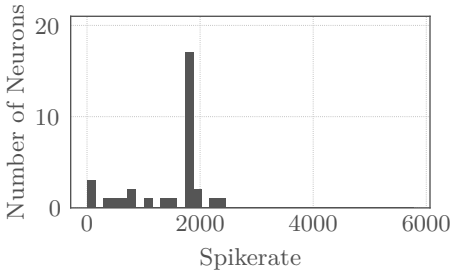
(b) After the rough V_{syn} calibration



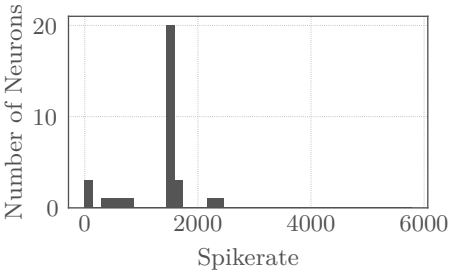
(c) After the fine adjust calibration of V_{syn}



(d) After calibration of I_{bias}



(e) V_{syn} calibrated again with the rough and the fine adjusting algorithm



(f) Result of the complete algorithm

Figure 3.5: Each histogram is showing the distribution of spike rates of all neurons, starting with the uncalibrated state and ending with the calibrated one it is showing the progress of the different calibration steps.

Figure 3.5d is showing the distribution after the calibration of I_{bias}. As one can see 21 neurons lie within three bins around a spike rate of 2000. Compared to the uncalibrated state there are now neurons with a similar spike rate, which should make the calibration of STP possible, compare to section 4.1.

To set the working point of the OTA one has to calibrate V_{syn} again. By using both alogrithms for V_{syn} again (figure 3.5e), followed by another calibration of I_{bias} one gets the final result of the complete calibration showed in figure 3.5f. With this calibration 23 neurons lie within two bins, which is a satisfying result showing that the calibration for spike rates is working for some neurons. But there are 9 neurons left which could not be used for STP calibration. This problem is also shown in [Weis, 2018], where out of 16 neurons, 5 were not used for calibrating STP for various reasons. A reason why this calibration is not working for all neurons will

be discussed in section 3.1.5.

Before closing this section, the spread of the spike rates depicted in the histograms shall be expressed by numbers. The standard deviation for the uncalibrated state is 1679 spikes with a mean of 3051 spikes. The relative deviation does not change for figure 3.5b and 3.5c. The deviations are 1197/1198 spikes with a mean of 2141/2147 spikes. By calibrating I_{bias} the deviation is lowered to 678 spikes with a mean of 1666 spikes. The relative deviation is higher for figure 3.5e with a deviation of 675 spikes and a mean of 1448 spikes. This result improved again for figure 3.5f with a standard deviation of 591 spikes and a mean of 1337 spikes. The relative deviation is the lowest for figure 3.5d, but this includes all neurons. By using the 21 neurons around a spike rate of 2000 and determining the deviation of these neurons one gets a standard deviation of 77 spikes with a mean of 1871 spikes, a relative deviation of 4.1 %. This can be improved for figure 3.5f, the 23 neurons around a spike rate of 1600 are having a standard deviation of 28 spikes with a mean of 1559 spikes. This is a relative deviation of 1.8 %. The lower deviation and more neruons which are useable justifies the second run of the I_{bias} calibration.

3.1.4 Synaptic input current (2)

To calibrate I_{bias} properly for all neurons one has to disable the influence of R_{syn}. So sending in spikes from the synapse drivers as done before is not a possibility to calibrate this parameter. By setting the synaptic input line to a chosen voltage would cause the OTA to output a constant current and would disable the influence of R_{syn}. With synaptic input line the line connected to R_{syn} and C_{syn} in figure 2.2 is meant. Of course V_{syn} also has to be calibrated, otherwise the current onto the membrane would be shifted. Fortunately the synaptic input line can be connected to a pin on the baseboard. So it is possible to connect a DAC to the synaptic input to force the line onto a certain voltage.

By disabling the leak term and just connecting the excitatory input to the membrane, the spike rate should be related to the constant current onto the membrane. By reading the amount of spikes in a certain time window should allow for calibrating on a desired spike rate.

The synaptic input line has to be below a voltage of 1.2 V for the excitatory input, to get a current onto the membrane for a calibrated V_{syn}. This algorithm should also work for the inhibitory input, but the voltage at the synaptic line must be above 1.2 V, otherwise there would be a current off the membrane. In the following just the excitatory input will be calibrated to test the method.

The algorithm is based on a binary search with 9 runs as a 9-bit offset is added to the capmem value of 511 LSB to get a range from 511 LSB to 1022 LSB. According to common theories it is desired to keep this value high [Sebastian Billaudelle, 2018, personal communication], so this range was chosen.

The OTA should be linear for $\Delta V = 200\,\text{mV}$ [Aamir et al., 2018b] before saturation. It is desired to stay in the linear range of the OTA to get an output current proportional to the g_{m} value. That is the reason why a DAC on the baseboard is used by the algorithm to set the synaptic line onto 1.15 V.

For every neuron the spike counter is reset and read out after 3 ms. This time windows was chosen after some investigation on the oscilloscope. To gain some statistics it is better to have a big time window. The counter however just has 8 bit,

so with the highest value of I_{bias} no neuron should reach the 255 spikes in the time window. This is the case for 3 ms and 1.15 V. For different voltages at the synaptic input also another time window have to be chosen.

In every run the individual spikes are compared to a mean rate. If there were more spikes as the mean, the bit of the run is not set. For this calibration a mean rate of 166 spikes was chosen. That is because the g_{m} value can be different for different models and experiments. There is not the "one and only" calibration as it is for V_{syn}.

The PPU however is not able to measure the value directly, so it is not possible to give the value and calibrate to it. On the other hand, the PPU is able to find the relative differences of the observables. These deviations can be minimized by algorithms to calibrate the parameters. To get a desired value for all neurons it would be possible to choose one neuron and select the parameters as desired. Then an algorithm could minimize the relative difference between the chosen neuron and all other different neurons. After such a calibration the deviations should be minimized an all neurons should have equal values.

The algorithm is extended with the fine adjusting algorithm of V_{syn} in every run to set the operating point of the OTA for the different values of I_{bias}. It works as described above but it does not take the former values reduces by 3 LSB to get a correction range of ± 5 LSB for every run, which should be enough to correct the operating point.

3.1.5 Calibration of the synaptic input (1)

Now the OTAs at the excitatory synaptic input of all neurons will be calibrated. With a rough search of V_{syn} at the beginning and the search for I_{bias} with fine adjusting V_{syn} in every run the whole calibration takes around a minute. The problem is the fine adjusting algorithm of V_{syn}, which takes around 5 s and is executed ten times during the whole calibration, taking the most time. Another approach to calibrate this parameter faster is done later in section 3.2.2.

To check this calibration of the OTA the sourcemeter can be used. The membrane is clamped to the sourcemeter and 0.8 V are applied. By varying the voltage on the synaptic input line between 0.9 V and 1.4 V the current of the OTA can be measured to get the operating characteristics of every single OTA of the neurons. For every OTA 50 single measurements are taken.

In figure 3.6 the different operating characteristics of all 32 excitatory OTAs in an uncalibrated state are plotted. It is clearly visible that the V_{syn} value is not calibrated, because of the wide distribution around 1.2 V. Calibrated curves should cross the x-axis at 1.2 V, because the idle synaptic input line voltage should be 1.2 V, which have to equal to V_{syn} for a calibrated OTA. In this case V_{syn} is set to 650 LSB and I_{bias} to 750 LSB for all neurons.

Due to supply drop the voltage at the synaptic input can be lower than the actual 1.2 V [Aamir et al., 2018b]. So making histograms of the distribution at 1.2 V will not deliver the true distribution of V_{syn}. Further investigations in section 3.2.8 are showing that this supply drop is about 20 mV. Histograms of the distribution at 1.18 V can be also found there.

By calibrating V_{syn} with the rough and fine adjusting algorithm this is better, as shown in figure 3.7. Most curves cross the x-axis in a similar point. Around

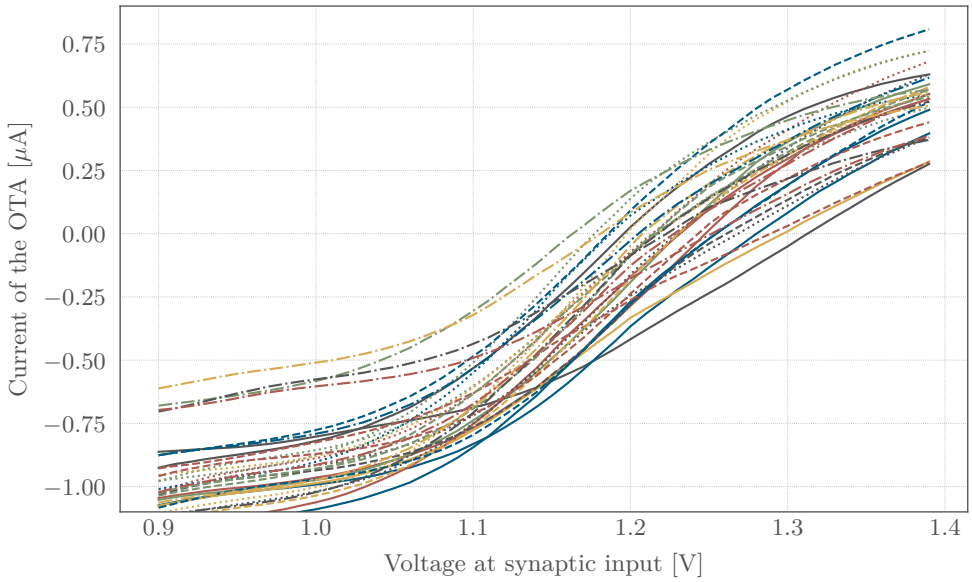


Figure 3.6: Operating characteristics of all 32 excitatory OTAs in the uncalibrated state. Every OTA has a individual color and linestyle.

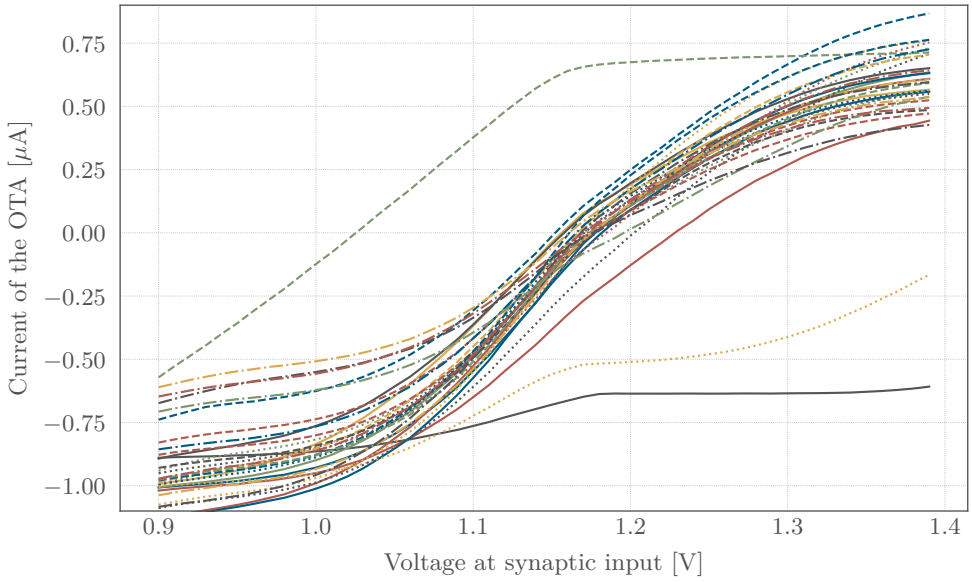


Figure 3.7: Operating characteristics of all 32 excitatory OTAs with spike rate calibration of V_{syn}. Every OTA has a individual color and linestyle.

five OTAs however show a different behavior, crossing the x-axis in a completely different point or do not even cross it. This indicates a wrong calibrated value of V_{syn}, because the crossing of the x-axis marks the value.

By manually adjusting the neurons it was possible to set the "right" V_{syn} value. Further investigation showed that some neurons start spiking at different values of V_{syn}, depending from which side one is changing the value. For example if one is

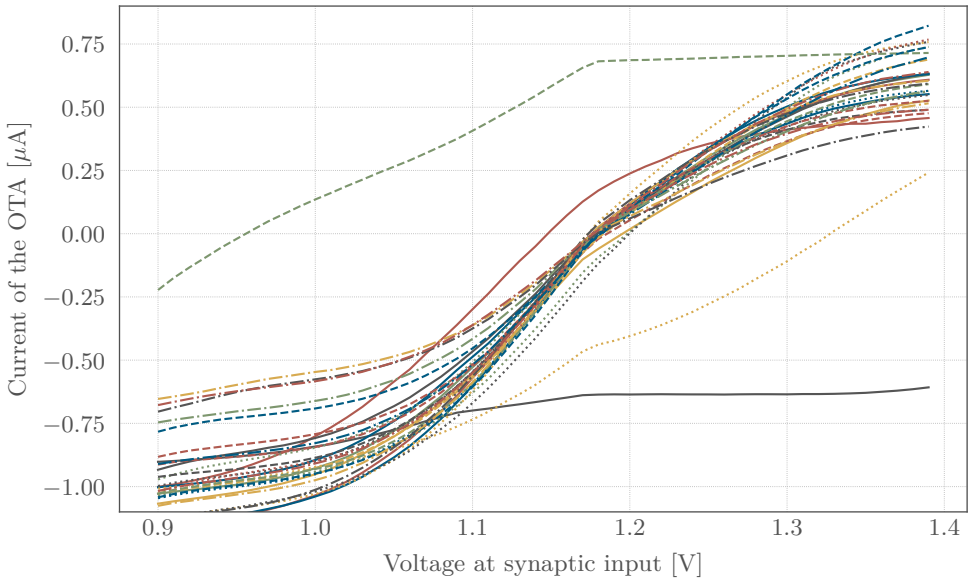


Figure 3.8: Operating characteristics of all 32 excitatory OTAs with spike rate calibration of V_{syn} and I_{bias}. Every OTA has a individual color and linestyle.

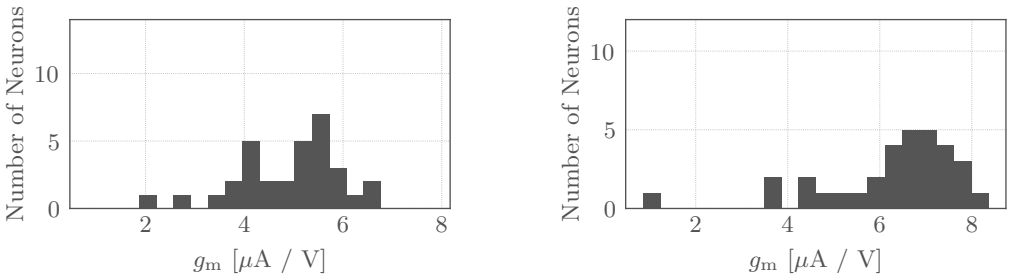
starting from 1000 LSB and going down until the neuron stops spiking one is reaching 550 LSB. But if one is starting from 0 LSB and going up until it is spiking one is reaching 650 LSB. This hysteresis problem is causing the algorithm to find wrong values for neurons which have this problem. But the algorithm proves its theoretical functionality for most neurons on the chip. It must be investigated on HICANN-X if this hysteresis is still there for some neurons. This problems led to a CADC based calibration of V_{syn} presented in section 3.2.1.

This investigation can also explain the histogram in figure 3.5f. Some neurons could not be calibrated to have a similar spike rate. Neurons with a spike rate of zero have a low value of V_{syn}, taking current off the membrane. Incoming spikes do not have an effect and the neuron is never spiking. Also high spike rates can be explained with too high values of V_{syn}, because a constant current is added to the incoming spikes onto the membrane causing a high spike rate.

In figure 3.8 the operating characteristics after a complete calibration of the OTAs is shown. Compared to figure 3.7 most curves are crossing the x-axis in a similar point. But there are again some neurons with a wrong calibrated V_{syn}. But the result is not convincingly because the curves are still spread somehow. One reason for the spread of I_{bias} will be discussed in section 3.2.2.

It is also possible to get the g_{m} value by determining the slope of each trace. According histograms are shown in figure 3.9. Figure 3.9a shows the distribution for an uncalibrated I_{bias}. The standard deviation is 1.03 µA/V with a mean of 4.93 µA/V. By calibrating I_{bias} with spike rates one gets a distribution shown in figure 3.9b. The standard deviation is 1.55 µA/V with a mean of 6.27 µA/V.

It can be concluded that it is no feasible to use spike rates to calibrate all neurons. A majority however can be calibrated in V_{syn}. But one wants to use all neurons on the chip and not a part of them. That is the reason why other methods are tested



(a) g_m distribution for uncalibrated I_{bias}

(b) g_m distribution for the spike rate calibration of I_{bias}

Figure 3.9: Comparison of the g_m values for figure 3.6 and figure 3.8.

to calibrate the OTAs of the neuron input, see section 3.2.

3.2 Neuron calibration via CADC

3.2.1 Synaptic input reference voltage (2)

All calibrations presented in section 3.1 were done with the leak term disabled, allowing for spike rate calibration. Perturbations on the membrane were a problem, because with a calibrated V_{syn}, the membrane potential is fluctuating. Enabling the leak term should stabilize the membrane potential, because the leak is realized with an OTA in unity gain feedback (see figure 2.2). This is modelling the conductance g_l from equation 1.

As explained in section 2.5 the PPU is able to read out the CADC, whose channels can be connected to each membrane potential. First of all the CADC must be calibrated with a reference voltage, which was done in [Weis, 2018] and the calibration is used in the following. It is based on connecting a known reference voltage from one of the DACs to the CADC and calibrating every channel by doing a linear fit for different voltages.

V_{syn} can be calibrated with an enabled leak and the CADC. By disabling all inputs besides the leak term, the membrane potential will be at V_{leak} plus an offset of the leak OTA and there will be no leakage current I_{leak}. So the potential difference at g_l is zero. By enabling one of the synaptic inputs, a current onto or off the membrane starts flowing for a wrong calibrated V_{syn}. This results in a higher/lower membrane potential, because the leakage current I_{leak} is not zero anymore resulting in a potential difference at g_l. By calibrating V_{syn} the current onto or off the membrane should be reduced. Perfectly calibrated the membrane potential should be at V_{leak}. So calibrating to the former membrane potential without synaptic input is another possibility to calibrate V_{syn}. The potentials can be read out with the CADC.

Such a method was already developed by Yannik Stradmann [Yannik Stradmann, 2018, personal communication]. But his method did not calibrate the last three bits. So just steps of 8 LSB are possible to calibrate with this algorithm. That is too inaccurate to use this as calibration, so a new algorithm was developed.

The algorithm is based on a binary search with 7 runs. It is calibrating a 7-bit offset to a value of 586 LSB leading to a range of 586 LSB up to 713 LSB. Former

research showed that this range should work for all neurons. Again just the excitatory input will be calibrated, but the algorithm should work perfectly fine also with the inhibitory input without changes.

First of all the leak gets enabled and all other inputs disabled. V_{leak} is set to 330 LSB and the g_{m} value for the leak OTA is set to 1000 LSB. Also spiking is disabled to prevent spiking for high settings of V_{syn} causing wrong results. All membrane voltages are read out with the CADC and get saved. After this the excitatory input is enabled and the binary search starts. In every run the value of V_{syn} is changed and one is waiting 10 ms for the membrane to recover. The CADC value of the current membrane potential is compared to the saved one. If this value is higher than the saved one it indicates a current onto the membrane is flowing and the bit is not set.

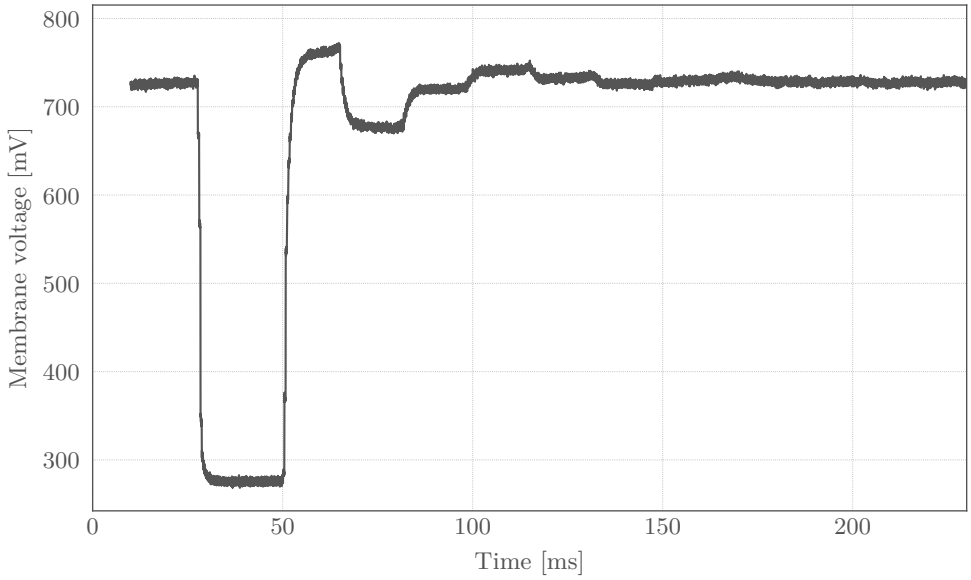


Figure 3.10: Progress of the calibration of V_{syn} with the CADC. The membrane potential changes according to the selected capmem value and is adjusted to fit to the voltage at the beginning of the trace.

In figure 3.10 the progress of the calibration is shown for neuron 14. The membrane potential during the calibration of V_{syn} with the CADC is recorded with the FlySpi-Board. Starting with V_{leak} at 0 ms, the voltage drops after 25 ms because the excitatory input is enabled. The former setting of V_{syn} was 586 LSB. At 50 ms the first significant bit is set to a total value of 650 LSB causing the membrane potential to rise above the first value. Thats why this bit is not set. At 65 ms the potential drops because just the second bit is set to a total value of 618 LSB. Because now the membrane is lower the bit get set and the voltage is rising again at 80 ms. So the membrane potential is rising and falling according to V_{syn} and at the end it is on the same level as at the beginning. The potentials are not changing every 10 ms because the PPU have to readout the CADC and set the capmem values which is extra time. The runtime of the complete algorithm is around 250 ms.

3.2.2 Calibration of the synaptic input (2)

To calibrate the whole OTA one can use the algorithm explained in section 3.1.5 by replacing the spike rate based algorithm of V_{syn} with the CADC based algorithm and compare the results.

One big advantage is the better runtime of 5 s for the whole calibration. The operating characteristics are shown in figure 3.11. Compared to figure 3.8 the result is also better, because V_{syn} was calibrated for all neurons successfully. By using the fine adjusting spike rate algorithm for V_{syn} at the end of the complete calibration the result can be improved, compare to figure 3.12. The runtime however is 10 s now.

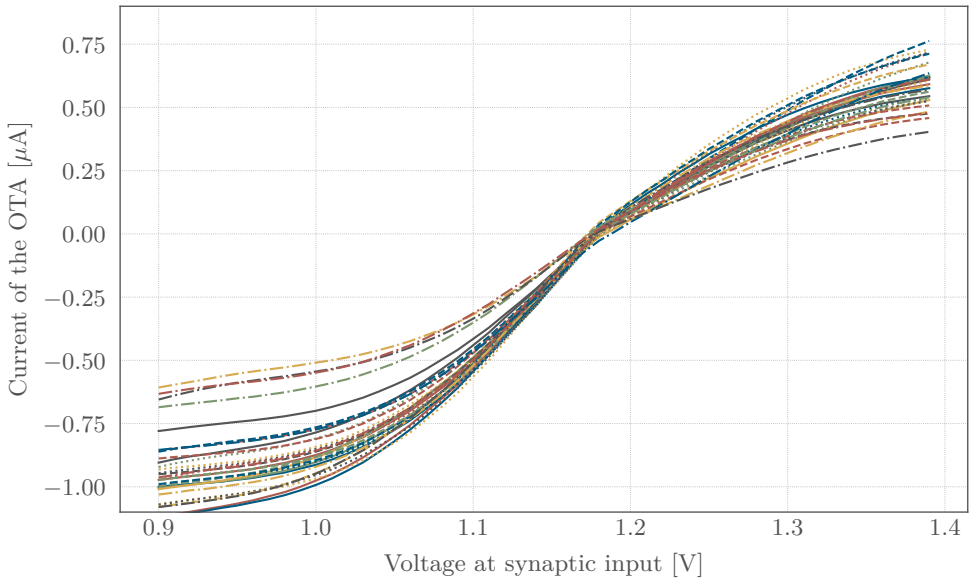


Figure 3.11: Operating characteristics of all 32 excitatory OTAs with CADC calibration of V_{syn} and spike rate based calibration of I_{bias}. Every OTA has a individual color and linestyle.

But figure 3.12 also shows some deviation in I_{bias}, as discussed in section 3.1.5. One reason for this is the deviation of V_{thresh} and V_{res}, which is sketched in figure 3.13. In two plots a model of the membrane potential during the calibration of I_{bias} is sketched. Because of the constant current onto the membrane its voltage is rising linearly resulting in regular spiking by reaching V_{thresh} and getting reset to V_{res}. In figure 3.13a the difference $\Delta V = V_{\text{thresh}} - V_{\text{res}}$ is smaller than in figure 3.13b. With the same I_{bias} one gets different spike rates in the same time windows, resulting in a calibration, which depends also on ΔV.

Figure 3.21a shows the distribution of the g_{m} compared to distribution with a calibrated ΔV. The standard deviation is 0.87 µA/V with a mean of 6.16 µA/V. The result improved compared to figure 3.9b.

Quantitatively one can write this in a formula. The charge Q on a capacitor can be calculated with its capacity C and the voltage U with $Q = C \cdot U$. By deriving it in time one gets the current I onto the capacitor as function of $\dot{U}$ with $I = C \cdot \dot{U}$. Let t be the time for a complete flank, starting at V_{res} and ending at V_{thresh}. Now

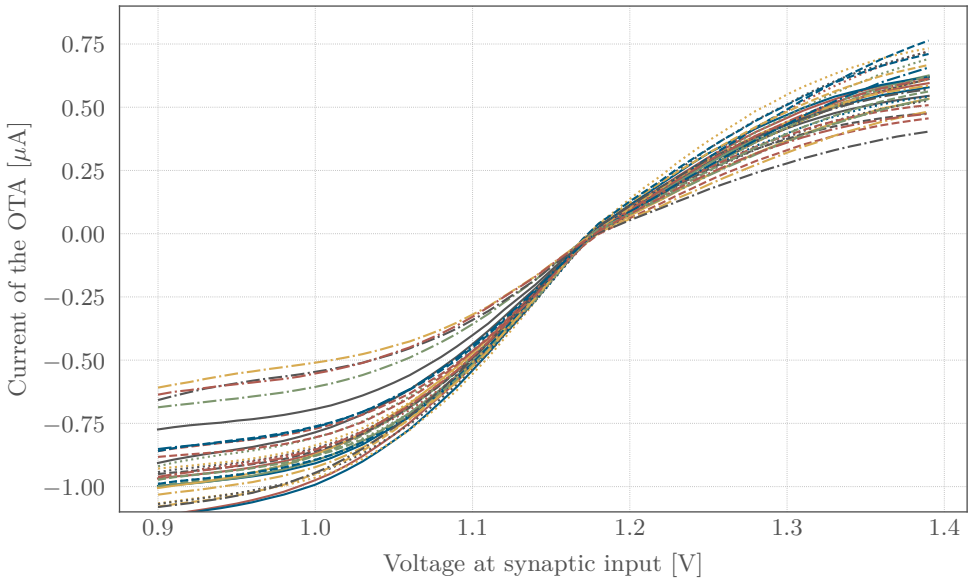
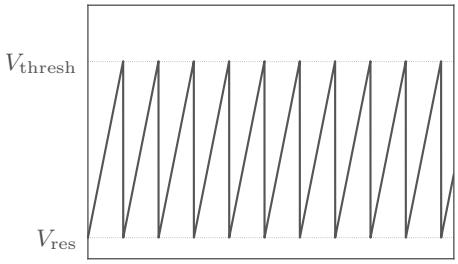
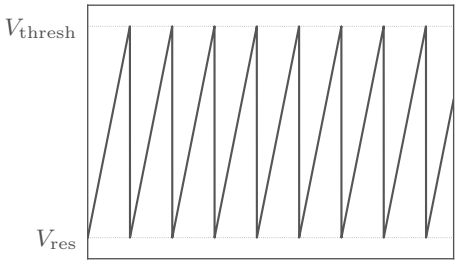


Figure 3.12: Operating characteristics of all 32 excitatory OTAs with CADC calibration of V_{syn}, spike rate based calibration of I_{bias} and fine adjusting spike rate algorithm for V_{syn} at the end. Every OTA has a individual color and linestyle.



(a) Small ΔV resulting in a higher spike rate compared to a bigger ΔV.



(b) Higher ΔV resulting in a lower spike rate compared to a lower ΔV.

Figure 3.13: Model of the membrane potential in a certain time window during the I_{bias} calibration. Because of the constant current on the membrane the voltage is rising linearly to V_{thresh} resulting in spiking and getting reset to V_{res}. The different plots show different values for V_{thresh}.

the voltage change can be calculated:

$$\dot{U} = \frac{\Delta V}{t} = \frac{I}{C}. \tag{13}$$

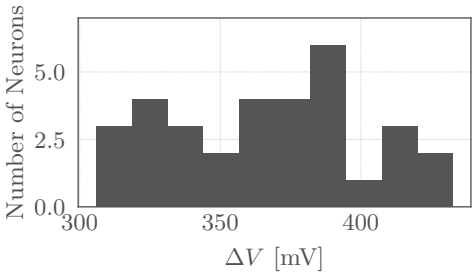
The spike rate s is now proportional to $1/t$ and with equation 13 it is

$$s \propto \frac{1}{t} = \frac{I}{\Delta V \cdot C}. \tag{14}$$

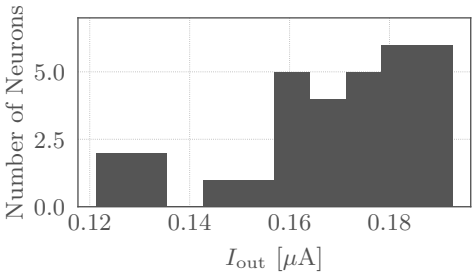
Let s_1 and s_2 be the I_{bias} calibrated spike rates of two different neurons. Because they are calibrated they should be equal. The membrane capacities should be equal by design, this results in

$$\frac{\Delta V_1}{\Delta V_2} = \frac{I_1}{I_2} \tag{15}$$

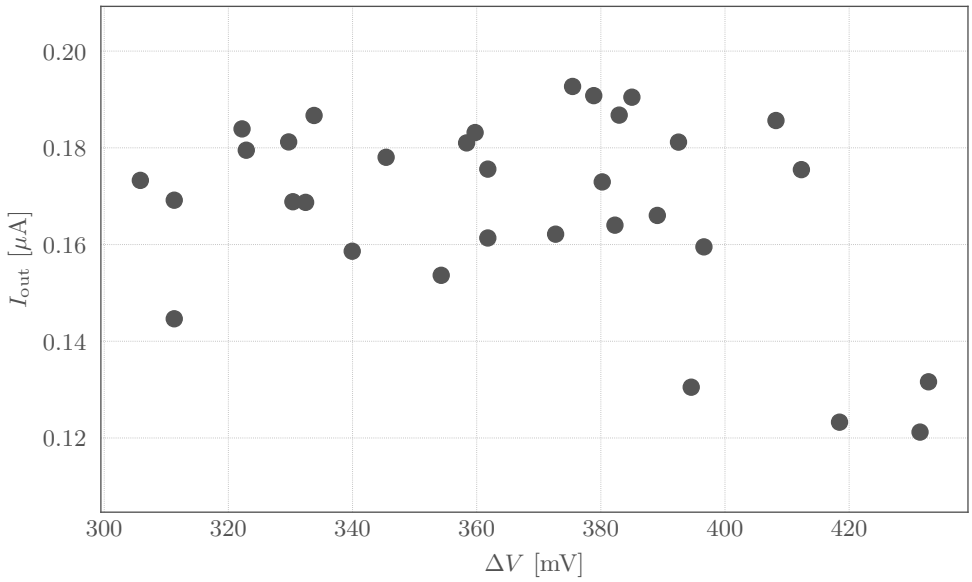
showing that the mismatch of I_{bias} after the calibration is related to the mismatch of the difference of V_{thresh} and V_{res}.



(a) Distribution of the difference of V_{thresh} and V_{res} for an uncalibrated state.



(b) Distribution of the current I_{out} onto the membrane with the synaptic input set to 1.15 V.



(c) Scatter plot for ΔV and the according I_{out} for all neurons.

Figure 3.14: Distributions of ΔV and I_{out}.

That is the reason why ΔV was investigated to find out if the mismatch of it causes such variations. This was done by recording the membrane potential with the FlySpi. To get V_{thresh} and V_{res}, 1.15 V were applied to the synaptic input to make the neurons spike. The curves look like the models in figure 3.13. By reading out the maximum and the minimum of each trace one can get the threshold and reset potential for every neuron. ΔV can be calculated for every neuron and the distribution is shown in figure 3.14a.

The current I_{out} onto the membrane can be determined by reading the current at 1.15 V from figure 3.12 for each neuron. The distribution of I_{out} is shown in figure 3.14b.

Direct comparison of both plots show a vastly distribution. In figure 3.14c ΔV is compared to I_{out} for every neuron to investigate the correlation of the two parameters. The four points in the lower part of the right are the four curves from figure 3.12 which have a lower linear slope. Reasons for that can be found in section 3.2.5. The correlation factor ρ for the remaining neurons is $\rho = 0.22$. ΔV and I_{out} are just weakly correlated. According to equation 15 they should be correlated. The equation however is based on different assumptions, for example the reset time of the membrane potential was neglected. Nevertheless V_{res} and V_{thresh} are calibrated in the following and it is tested again if the calibration of I_{bias} can be improved with a calibrated ΔV.

3.2.3 Reset potential

To calibrate V_{res} with the PPU it is not possible to record the trace of the membrane potential with the CADC and find its minimum points as it is done above with the FlySpi and the host computer during regular spiking. That is because the PPU has not the computational power and memory to do this. Also reading out the CADC is way to slow to record this trace.

But it is possible to trigger a reset manually and also the reset time can be increased. With this method the membrane potential is set to V_{res} and can be read out with the CADC. So calibrating to a given CADC value could calibrate the according capmem values.

In figure 3.15 the CADC measured V_{res} is compared to the according capmem value. The 8-bit CADC value was converted to the according calibrated voltage. The reset time is increased and all neurons are forced to the reset value and the CADC reads out the membrane potential.

Starting from 200 LSB all curves show a linear behavior, while they seem to be saturated for values below 100 LSB. This plot shows that a calibration of V_{res} above 0.4 V is possible with a maximal difference of 80 LSB for the capmem values.

The calibration algorithm uses the function described above to readout the membrane voltage with the CADC for a forced reset.

A 6-bit offset would just allow a range of 64 LSB, which is to small to calibrate all neurons to the same V_{res} as figure 3.15 shows. That is the reason why the algorithm is based on a binary search with 7 runs to add a 7-bit offset to a value of 86 LSB. This results in a calibration range from 86 LSB to 213 LSB. It is desired to get capmem values around a value of 150 LSB.

The CADC can be calibrated differently, so the values of different CADC readouts can not be compared if the CADC was calibrated between these readout. As the calibration depends on CADC values, a fixed calibration value as done with the I_{bias} calibration was not chosen. Instead the mean of all CADC values for a setting of 150 LSB is taken. In every run the V_{res} potential is determined and compared to this mean value. If the actual value is bigger than the mean, the bit is not set.

To compare the uncalibrated with the calibrated state the same measurement technique as used to find ΔV was used by using the FlySpi and evaluate the data on the host computer. The comparison is shown in figure 3.16.

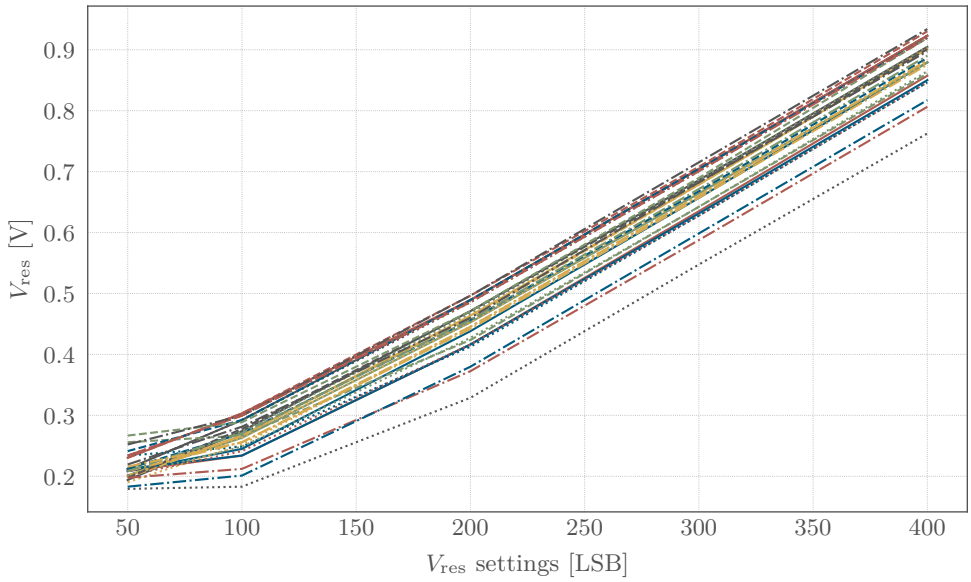
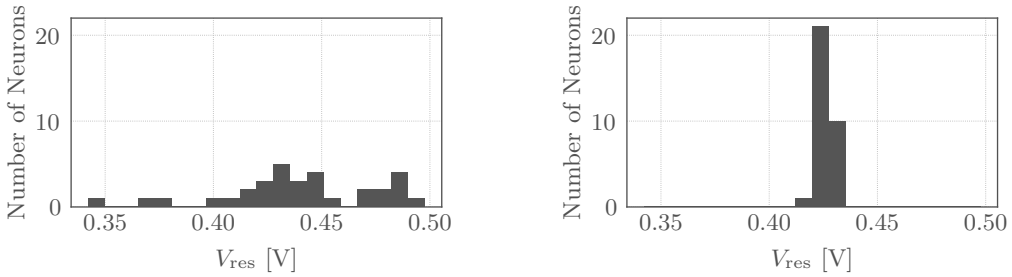


Figure 3.15: The capmem settings of V_{res} are plottet against the measured voltage of V_{res}.



(a) Uncalibrated V_{res}.

(b) Calibrated V_{res}.

Figure 3.16: Both histograms showing the distribution of V_{res}, measured by the FlySpi and evaluated with the host computer.

As figure 3.16b shows, the calibration works fine as all values of V_{res} lie within three bins. This is a big improvement compared to the uncalibrated distribution shown in figure 3.16a. The uncalibrated V_{res} has a standard deviation of 35.5 mV with a mean of 439.3 mV. Calibrated the standard deviation is lower by an order of magnitude, it is 3.6 mV with a mean of 425.9 mV. The relative deviation can be lowered from 8.1 % to 0.9 %. The runtime is also below 1 s and should not change for HICANN-X significantly.

3.2.4 Threshold potential

The calibration of V_{thresh} ca not be done by reading the whole trace and finding the maximum by applying 1.15 V to the synaptic input for the same reason discussed in the context of the V_{res} calibration. So a new method has to be developed to find

the threshold with the PPU.

One possible method is using the leak to calibrate the threshold. By just connecting the leak term to the membrane, the potential of the membrane is V_{leak} plus an offset of the leak OTA. So by raising the capmem value until it is spiking will indicate that V_{leak} reached V_{thresh} and is above it. This can be seen in figure 3.17. The spike rate is measured over a time of 500 µs. For all neurons V_{thresh} is set to 450 LSB and the g_{m} value for the leak OTA is set to 1000 LSB.

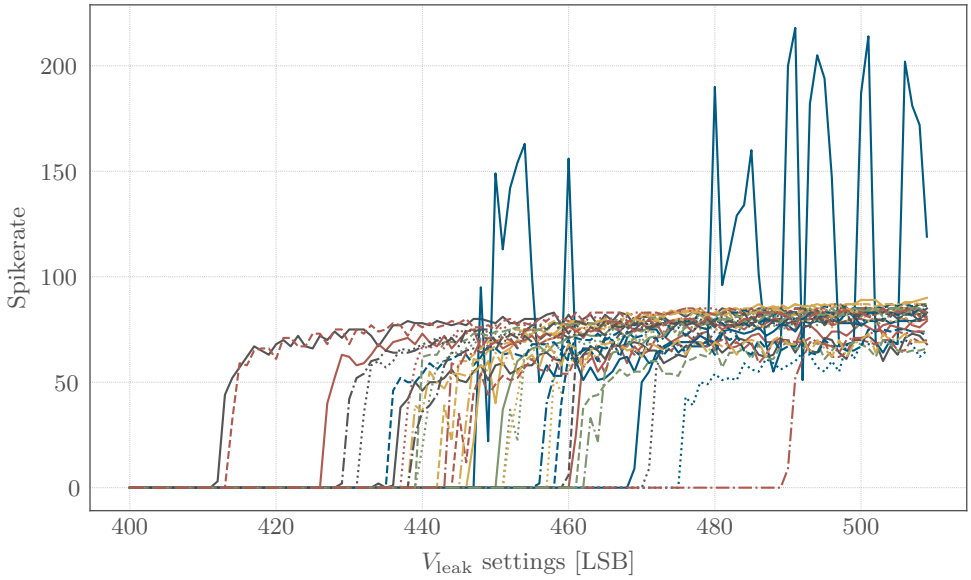


Figure 3.17: Different settings of V_{leak} with the corresponding spike rate. Spiking is enabled for the neurons, but all other membrane inputs are disabled.

One can see that the neurons start spiking at a certain value for V_{leak}. The spike rate goes fast up to a range of 60 - 90 spikes in the selected time window and stays at the same level for higher values. Just one neuron is irregular spiking, compare to figure 3.17. It is again neuron 2 which is reported to count some spikes several times, as already mentioned in section 3.1.1. But it also starts spiking at a certain V_{leak}, so this should not be a problem for this calibration.

The idea of the algorithm is to raise the leak to it is maximal value that no spiking occurs. So V_{leak} is a little bit smaller than V_{thresh}. By reading out the membrane potential with the CADC on can readout a good approximation of V_{thresh} for all neurons. The value is a little bit shifted downwards for all neurons, so this should not be a problem. By calibrating to a given CADC value the capmem values of V_{thresh} can be calibrated.

The search for the settings of V_{leak} is done with a binary search. As figure 3.17 shows, a 7-bit search is necessary to get the whole range. The start depends on V_{thresh}, because the curve in figure 3.17 will be shifted to the right for a higher V_{thresh}. Because the mean of V_{thresh} should be 450 LSB in this case, a 7-bit offset is added to a V_{leak} value 386 LSB to get the whole range. This results in a calibration range of 386 LSB to 513 LSB. The spike counters are read out within a time window of 5 ms. If at least one spike is detected the bit is not set. The high waiting should

make sure that no spiking occurs, which would led to wrong results by reading it out with the CADC.

In figure 3.18 the CADC measured V_{thresh} is compared to the according capmem value. The 8-bit CADC value was converted to the according calibrated voltage. The CADC was read out after the algorithm described above was executed. Steps of 50 LSB for V_{thresh} were recorded.

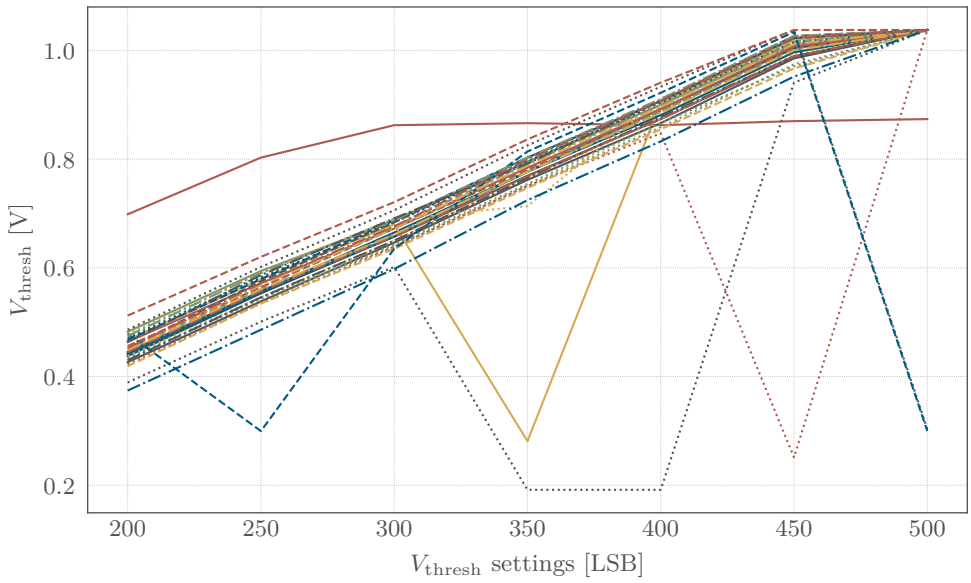


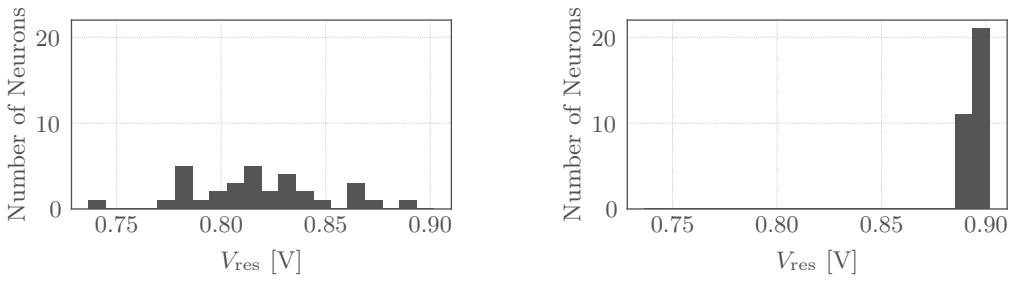
Figure 3.18: The capmem settings of V_{thresh} are plottet against the measured voltage of V_{thresh}, determined by the algorithm describes above.

Most curves in the whole interval are showing a linear behavior saturating at 500 LSB. So calibrating these around 450 LSB should not be a problem. A higher V_{thresh} however is not possible with this CADC calibration. This has to be changed to get the linearity for higher voltages, compare to [Weis, 2018]. But there are also some outlier which differ from the linear behavior for one measurement point. That is because spiking can occur even if it does not spike during the earlier 5 ms. V_{leak} is too close to V_{thresh} that small influences result in spiking. This can be fixed by subtracting 5 LSB at the end from V_{leak} to be sure the neuron is not spiking.

The main algorithm uses the algorithm described above in each run to find V_{thresh}. The main algorithm is based on a binary search with 7 runs adding an offset to a value of 386 LSB. This results in a range of 386 LSB to 513 LSB to calibrate V_{thresh}. As done in the calibration for V_{res}, a mean value of the CADC is determined for a capmem setting of 450 LSB to be independent of the CADC calibration. In every run the V_{leak} value close before spiking is determined by the algorithm described above and reduced by 5 LSB. This does not make a difference for the calibration because all values are corrected and V_{leak} is linearly connected to its capmem settings below 450 LSB, compare to figure 3.32. The bit is not set if the current membrane potential is above the mean membrane potential. The complete calibration of V_{thresh} can be done in under 1 s and should not change for HICANN-X.

To compare the uncalibrated with the calibrated state the FlySpi was used to find

the value of V_{thresh}. Both histograms are shown in figure 3.19.



(a) Uncalibrated V_{thresh}.

(b) Calibrated V_{thresh}.

Figure 3.19: Both histograms showing the distribution of V_{thresh}, measured by the FlySpi and evaluated with the host computer.

Figure 3.19b shows the functionality of the whole algorithm because all values lie within two bins. The standard deviation is 3.9 mV at a mean of 894.8 mV. The distribution of the uncalibrated values is shown in figure 3.19a and the standard deviation is 33.6 mV with a mean of 817.3 mV. Compared to the calibration of V_{res} in section 3.2.3, the standard deviations of the uncalibrated and calibrated voltages are the same. But because of the higher mean the relative deviation is 4.1 % for the uncalibrated V_{thresh} and 0.4 % for the calibrated one.

3.2.5 Calibration of the synaptic input (3)

With the calibrated V_{res} and V_{thresh} also ΔV should be calibrated better. So the whole spike rate based calibration of I_{bias} was redone to check the results from section 3.2.2. Again during the calibration of I_{bias} the settings for V_{syn} were calibrated with the CADC algorithm to set the working point and at the end the fine adjusting algorithm was used. The calibration results are shown in figure 3.20.

Compared to figure 3.12 the operating characteristics still have some deviation but the whole result improved, showing that ΔV was playing a role in calibrating I_{bias} with spike rates. This is also shown in figure 3.21 for their g_{m} values. Without calibrated ΔV (figure 3.21a) the relative deviation is 14.1 % with a mean of 6.16 µA/V. With a calibrated ΔV the standard deviation is 0.61 µA/V with a mean of 6.28 µA/V. The relative deviation is 9.6 %. But for both values the outliers with a smaller g_{m} value are causing these high standard deviations.

These outliers are four neurons which differ from the rest and can also be seen in the other operating characteristic curves in this thesis. This led to an extra investigation of this phenomena. The current from the OTA on the membrane was recorded for different settings of I_{bias}. This was done by connecting the sourcemeter with 0.8 V to the neuron and connecting the excitatory synaptic input to the DAC with 1.15 V. All other inputs were disabled and for every setting of I_{bias} the CADC calibraiton of V_{syn} was executed to set the working point.

The result is shown in figure 3.22. One can see that that the current depends linear on the settings for I_{bias}, but there are different slopes for all neurons. Four neurons are having a lower slope resulting in small currents even for high settings of I_{bias}. Their current output at 1000 LSB equals the output at 600 LSB for other

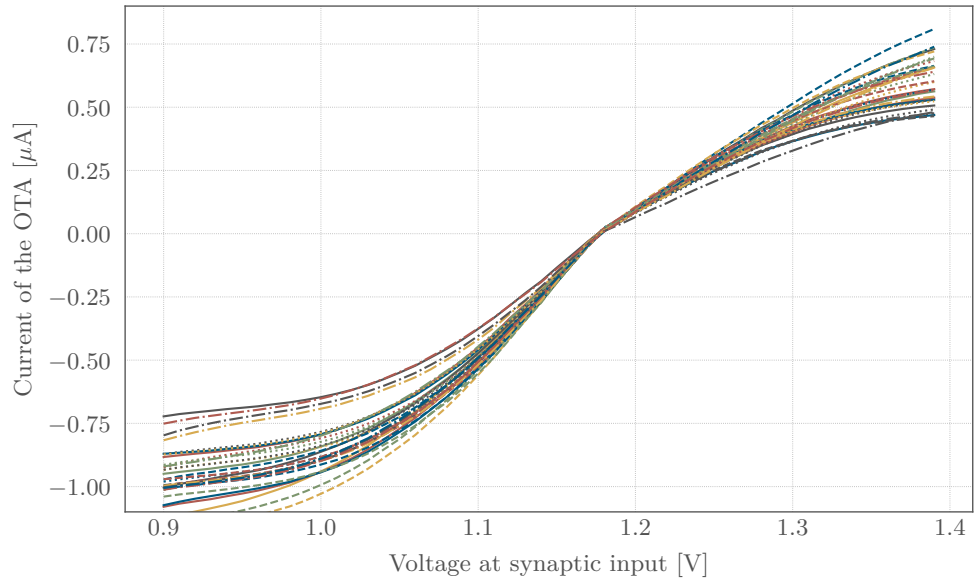
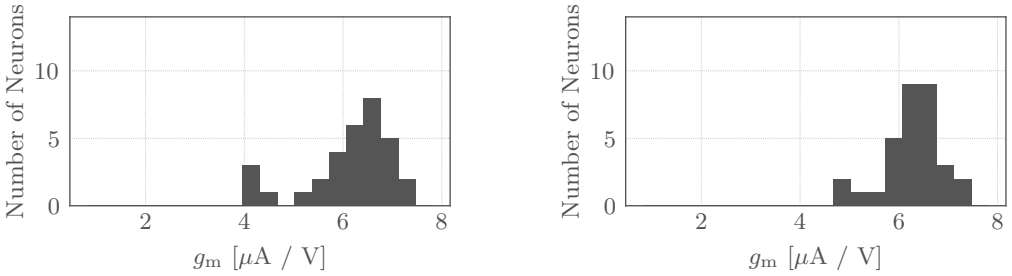


Figure 3.20: Operating characteristics of all 32 excitatory OTAs, with precalibrated V_{res} and V_{thresh}. OTAs were calibrated with CADC calibration of V_{syn}, spike rate based calibration of I_{bias} and fine adjusting spike rate algorithm for V_{syn} at the end. Every OTA has a individual color and linestyle.



(a) Using spike rates to calibrate I_{bias}, but V_{res} and V_{thresh} are not calibrated

(b) V_{res} and V_{thresh} are calibrated and I_{bias} is calibrated with spike rates

Figure 3.21: Comparison of the g_{m} values for figure 3.12 and figure 3.20.

neurons. This makes it nearly impossible to calibrate them for a range of 511 LSB to 1022 LSB, which was done in the spike rate based calibration of I_{bias}. But it explains the deviations of some neurons in figure 3.20, which are the same neurons mentioned here. They have been setted to 1022 LSB, but they output a current which is too low.

The plot however shows that a calibration of the OTAs is possible, but a full 10-bit calibration of I_{bias} is necessary. That is because the mismatch of different OTAs is a factor of two. While one OTA outputs 0.7 µA with I_{bias} set to 1000 LSB another OTA outputs 1.4 µA with the same settings on the synaptic input line. To fulfill the expectations of the common model this mismatch have to be lowered to

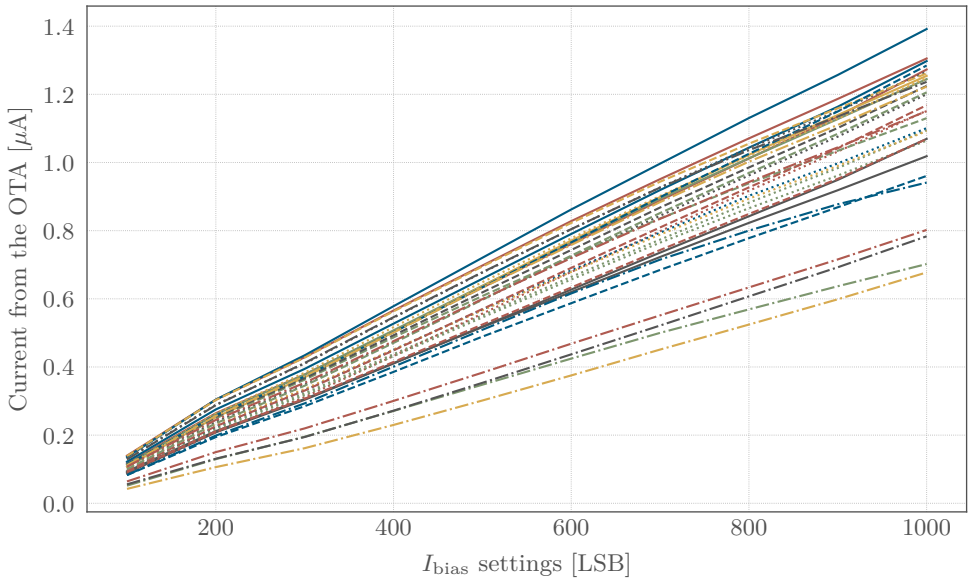


Figure 3.22: Different capmem settings of I_{bias} are plottet against their corresponding current onto the membrane from the OTA. Measured with a sourcemeter.

get high input currents for all neurons.

In the current state there are two options. The first option is to keep these high settings and exclude the neurons with weak input OTAs to fit to the common model. Another method is to use lower settings for I_{bias} to get all neurons calibrated. With this method I_{out} is lower for all neurons.

Figure 3.23 shows the calibration result for a full 10-bit calibration of I_{bias} with calibrated V_{res} and V_{thresh}. It is the same algorithm as used in section 3.1.4. However the time windows was increased to 4 ms and the mean rate was set to 170 spikes. With this changes a smaller value for I_{bias} will be set. The g_{m} value will be discussed in section 3.2.6.

Now there are two neurons which vary from the other calibrated ones. Closer investigation showed that this are neuron 0 and 31. The same problem appeared in section 3.2.6, where this issue will be discussed.

3.2.6 Synaptic input current (3)

The spike rate based calibration of I_{bias} has calibrated the OTA successfully. But by inspecting figure 3.20 there is still room for improvement. That is the reason why another method of calibrating I_{bias} with the CADC was tested.

It is using the same effect as the spike rate calibration of I_{bias} by disabling all terms beside the excitatory synaptic input and setting it to 1.15 V. Like the spike rate based algorithm this calibration can also be used for the inhibitory input by setting the input line to a voltage higher than 1.2 V. There is a constant current onto the membrane. The current onto the membrane is proportional to the slope of the membrane voltage, because of $I = C \cdot \dot{U}$. Determining the slope of the trace allows for calibrating I_{bias}.

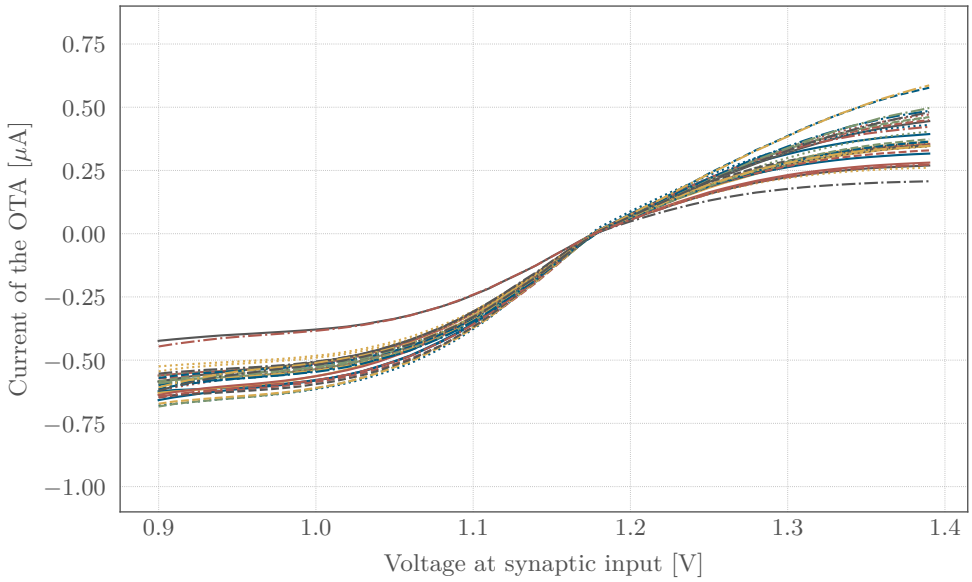


Figure 3.23: Operating characteristics of all 32 excitatory OTAs, with precalibrated V_{res} and V_{thresh}. OTAs were calibrated with CADC calibration of V_{syn}, spike rate based calibration of I_{bias} and fine adjusting spike rate algorithm for V_{syn} at the end. I_{bias} is calibrated over the whole 10-bit range.

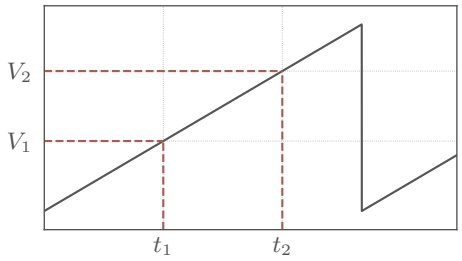


Figure 3.24: Model of the membrane potential in a certain time window during the I_{bias} calibration. The slope can be determined by reading out two points of the membrane potential. Two points are enough because it is rising linearly.

The algorithm is based on finding the slope with the CADC. The basic idea is sketched in figure 3.24. After triggering a reset the membrane potential will start rising again. By reading out the CADC and saving the different values in arrays one can get V_1 and V_2. To get the slope of the trace one normally must know $\delta t = t_2 - t_1$. But reading out the CADC can be done in parallel for all neurons. So δt is the same for all and the slope is proportional to $\delta V = V_2 - V_1$. So calibrating all neurons to have the same δV will calibrate I_{bias}.

First of all the algorithm is tested for high values of I_{bias}. To compare the algorithm with the spike rate algorithm also the same range from 511 LSB to 1022 LSB

is used by using a binary search with 9 runs. It is expected that the four neurons from figure 3.22 also cannot be calibrated with this algorithm. It starts by determine a mean slope for a setting of 767 LSB. In every run V_{syn} gets calibrated with its CADC calibration. Afterwards the slope is determined after a triggered reset by reading out two points of the rising potential. If the slope is bigger than the mean slope the bit is not set. Afterwards the fine adjusting algorithm for V_{syn} is used. In particular this algorithm is the same as the spike rate bases algorithm just with the slope as condition, the rest is the same.

The operating characteristics can be seen in figure 3.25. It also shows the problem of the four neurons which cannot be calibrated in the range of the algorithm. But it seems that this approach improves the calibration of I_{bias} compared to the spike rate approach (figure 3.20). This is also verified by figure 3.26. Also the runtime does not change and the whole calibration can be done in under 10 s and should not change for HICANN-X significantly.

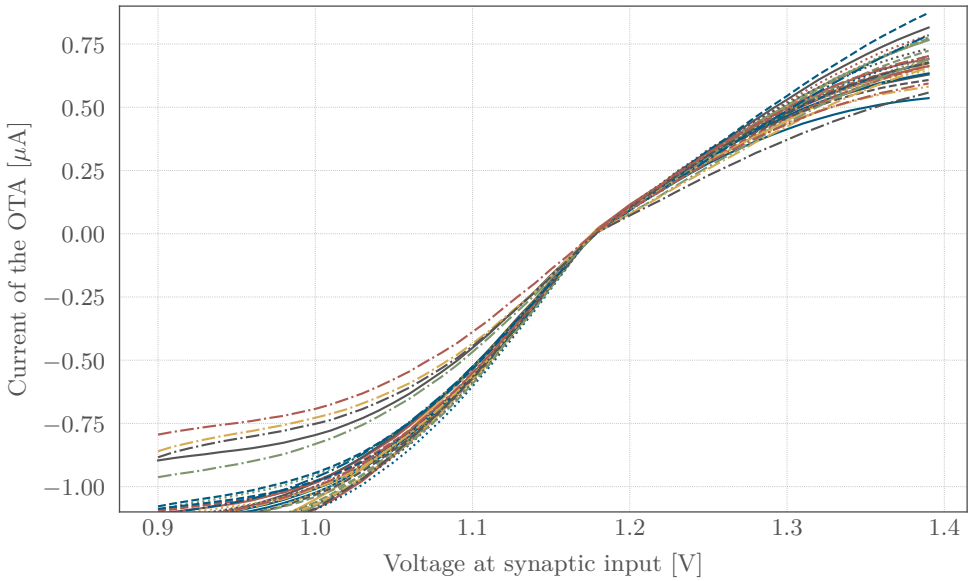
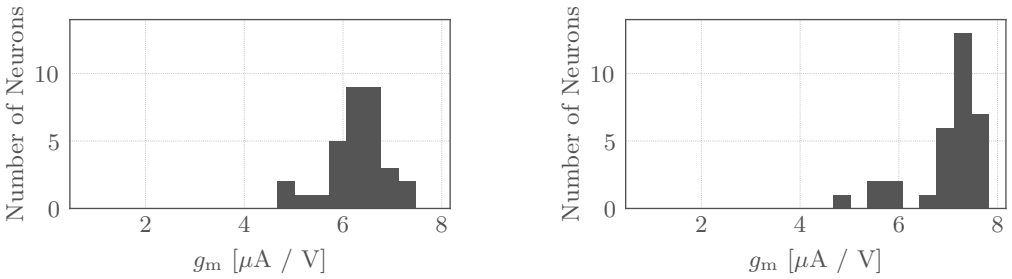


Figure 3.25: Operating characteristics of all 32 excitatory OTAs with CADC calibration of V_{syn}, CADC based calibration of I_{bias} by slopes and fine adjusting spike rate algorithm for V_{syn} at the end. Every OTA has a individual color and linestyle.

Figure 3.26 shows the direct comparison between both calibration approaches of I_{bias}. Figure 3.26a is the same as figure 3.21b. But it is compared to figure 3.26b. The standard deviation of the latter is 0.70 µA/V, which is 10.0 % of the mean. The spike rate approach has a relative deviation of 9.6 %. With the slope approach however 26 neurons lie within three bins and the five neurons with a lower g_{m} are causing the higher standard deviation. For the 26 neurons the standard deviation is lower.

To compare the results to section 3.2.5, where I_{bias} also was calibrated over its full 10-bit range, another algorithm also should be possible to calibrate lower values for I_{bias}. Therefore the same slope algorithm was used but the mean slope was



(a) V_{res} and V_{thresh} are calibrated and I_{bias} is calibrated with spike rates

(b) I_{bias} calibration with the CADC slope approach

Figure 3.26: Comparison of the g_m values for figure 3.20 and figure 3.25.

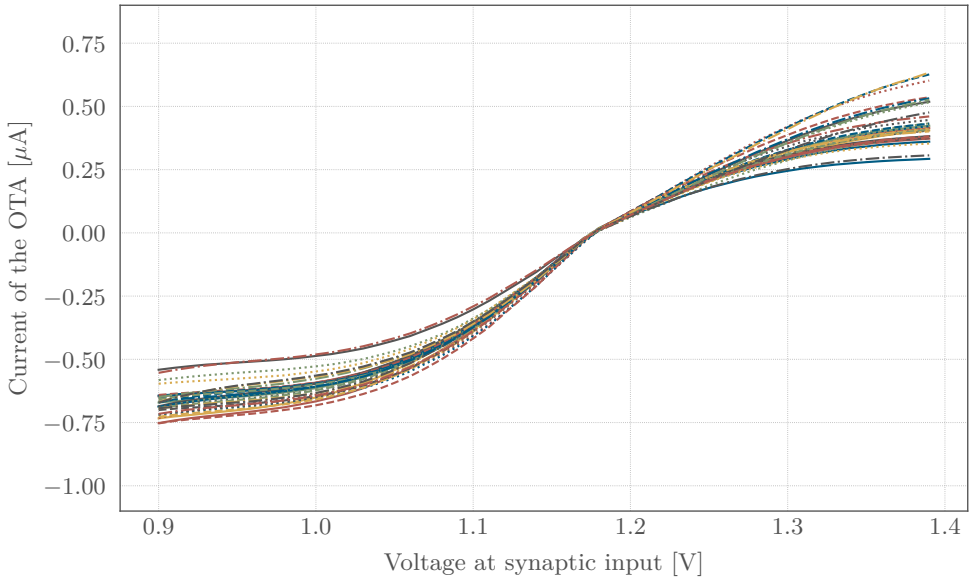
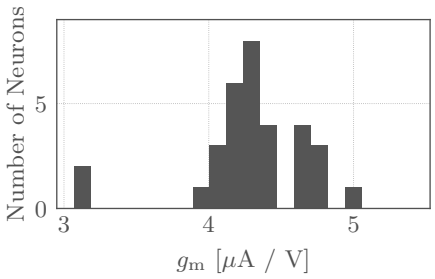


Figure 3.27: Operating characteristics of all 32 excitatory OTAs with CADC calibration of V_{syn}, CADC based calibration of I_{bias} by slopes and fine adjusting spike rate algorithm for V_{syn} at the end. I_{bias} is calibrated over the whole 10-bit range.

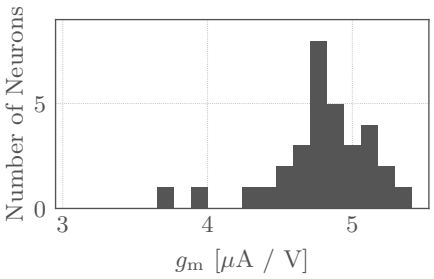
determined at 512 LSB. The other parts of the algorithm stays the same. The calibration result is shown in figure 3.27.

The g_m values for both calibrations with the whole 10-bit I_{bias} calibration are shown in figure 3.28. Both histograms show two outlier at low g_m values. As mentioned above these are neuron 0 and 31 for both calibrations. Figure 3.28a has a standard deviation of 0.40 µA/V with a mean of 4.29 µA/V. The standard deviation for figure 3.28b is 0.35 µA/V with a mean of 4.78 µA/V. By comparing the relative deviations the slope approach is better. With 7.4 % its relative deviation is lower than the 9.4 % from the spike rate approach.

Again neuron 0 and 31 are not calibrated as already mentioned in section 3.2.5. The reason for this is unknown, but both neurons are on the edge of the chip



(a) I_{bias} is calibrated with spike rates for calibrated V_{res} and V_{thresh}



(b) Slope approach used for the calibration of I_{bias}

Figure 3.28: Comparison of the g_m values for figure 3.23 and figure 3.27. In both cases the whole 10-bit search was done.

and cannot be calibrated with different algorithms. An assumption is that due to parasitic effects the membrane capacity is lower than for other neurons. For all calibrations of I_{bias} it was supposed that the membrane capacity is the same for all neurons. So the voltage change as shown in equation 13 is higher for lower capacities. As shown in equation 14 for same spike rates the current onto the membrane is lower for the same ΔV. So the g_m value should be also lower for neurons with smaller membrane capacities calibrated with all algorithms discussed in this thesis.

Mismatch in the neurons' membrane capacitances could also explain the spread observed for the g_m values. The synaptic input line was pulled to 1.10 V – instead of the previous 1.15 V – to investigate if the measurement resolution was the dominating factor contributing to the observed post-calibration variances. Since the received results were very comparable, the membrane capacitors are suspected to vary strongly.

3.2.7 Problems calibrating the synaptic input current

The best calibration of I_{bias} shows a relative deviation of 7.4 % for the g_m values. Compared to a deviation of 20.9 % for an uncalibrated I_{bias} this is not a significant improvement. That is the reason why one has to look at the possible observables critically.

Equation 14 shows that the spike rate depends on the current I_{out} onto the membrane, the voltage difference ΔV ($V_{thresh} - V_{res}$) and the membrane capacity C. By calibrating ΔV it was possible to improve the relative deviation of g_m with the spike rate based calibration for I_{bias} from 14.1 % to 9.6 %. However with neglected capacity mismatch the correlation factor between ΔV and I_{out} was just 0.22. This low correlation could be caused by the capacity mismatch.

The voltage change $\dot{U}$ depends on the current onto the membrane I and the capacity C as shown in equation 13. Also for the slope based calibration of I_{bias} the mismatch of the neuron capacities was neglected and it was assumed that it is the same for all neurons. But even with calibrating the whole range of I_{bias} it was just possible to get a relative deviation of 7.4 % for g_m.

As mentioned in section 3.2.6 it is expected that the membrane capacities are spread. This would cause such a deviation. But also the algorithm could be im-

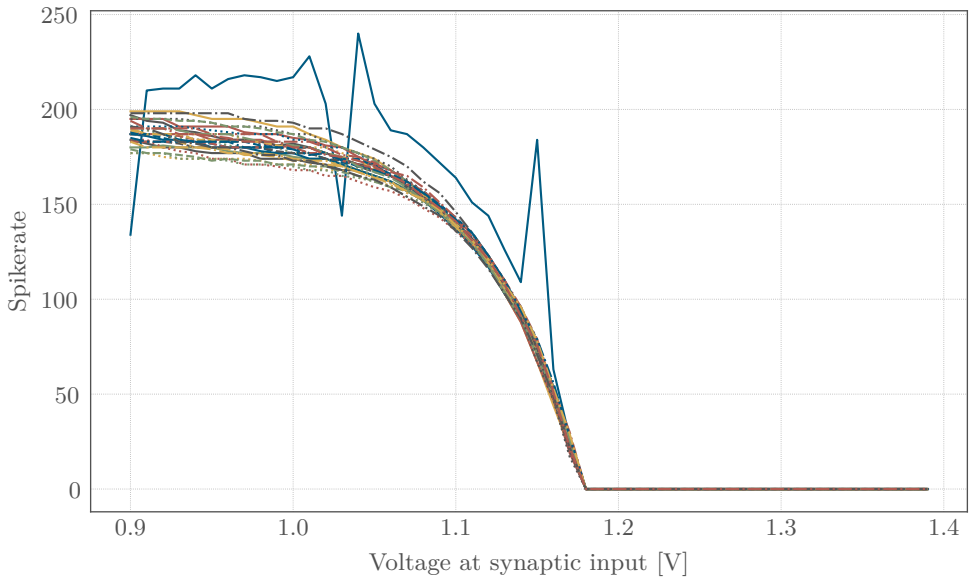


Figure 3.29: Different voltages at the synaptic input line are causing different spikesrates. Neuron 2 behaves because of the defect spike counter as outlier.

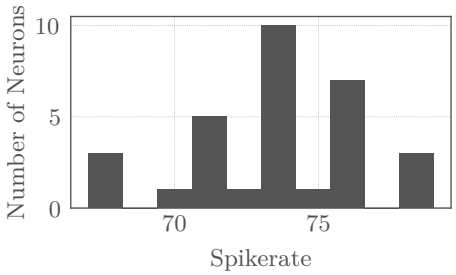


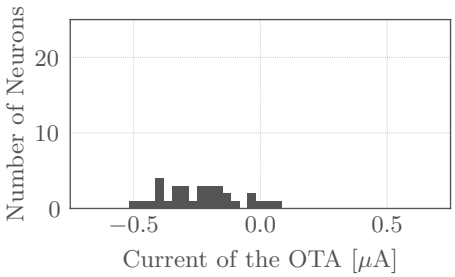
Figure 3.30: Distribution of the spike rates at 1.15 V from figure 3.29.

perfect, so the spike rates for different voltages at the synaptic input are recorded to test this, which is shown in figure 3.29. The different spike rates at a voltage of 1.15 V are shown in figure 3.30. With a standard deviation of 3.2 spikes this makes a relative deviation of 4.3 % when neuron 2 is excluded.

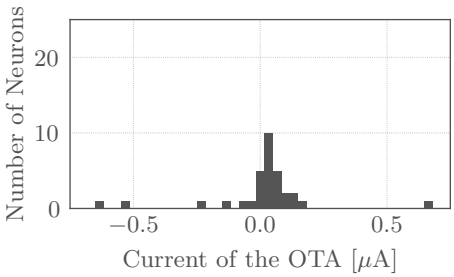
3.2.8 Characterization of the synaptic input voltage

As mentioned in section 3.1.5 due to a supply drop the voltage at the synaptic input line can be different from the desired 1.2 V. By evaluating figure 3.20 this voltage drop was determined to be 20 mV, because all calibrated curves cross the x-axis around 1.18 V. So for different methods of calibrating V_{syn} are the distributions at 1.18 V taken. The result can be seen in form of histograms in figure 3.31.

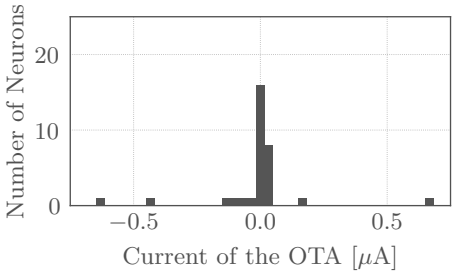
Figure 3.31a shows the distribution for an uncalibrated value of V_{syn}. For every neuron this value is set to 650 LSB. For most neurons this value is too small, because there is a current off the membrane. The mean is -0.238 µA with a standard



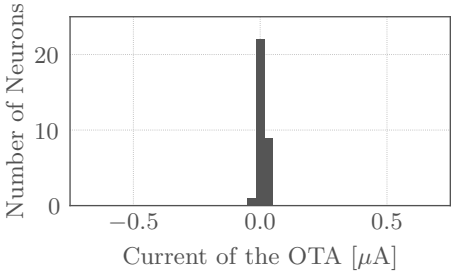
(a) Uncalibrated (figure 3.6)



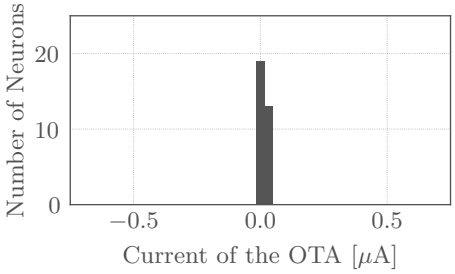
(b) spike rate calibration of V_{syn} (figure 3.7)



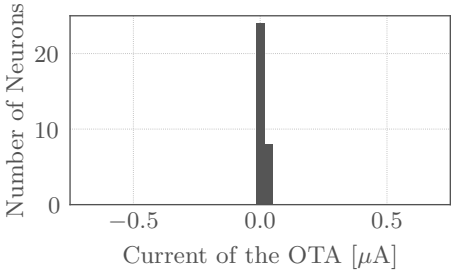
(c) spike rate calibration of V_{syn} and I_{bias} (figure 3.8)



(d) CADC calibration of V_{syn} and spike rate calibration of I_{bias} (figure 3.11)



(e) Same as figure 3.31d with fine adjusting V_{syn} algorithm at the end (figure 3.12)



(f) Same as figure 3.31e with calibrated V_{res} and V_{thresh} (figure 3.20)

Figure 3.31: Cross section of different OTA curves with 1.18 V.

deviation of 0.147 µA. The currents are within a range of −0.491 µA to 0.079 µA.

By calibrating V_{syn} with spike rates, 20 neurons lie within three bins around a current of 0 µA (figure 3.31b). This is a improvement compared to the uncalibrated state. There are some outlier making this method not usable to calibrate all neurons on the chip. Because of the outlier the standard deviation is 0.204 µA and the mean 0.012 µA. The minimal current is −0.638 µA and the maximum current is 0.666 µA.

Calibrating V_{syn} within every run of I_{bias} gives figure 3.31c. 24 neurons lie within two bins around 0 µA. So calibrating I_{bias} and set the working point with V_{syn} improves the distribution. But there are also some outlier making this method unusable. The standard deviation is 0.190 µA with a mean of −0.008 µA. The range of the currents is from −0.636 µA to 0.618 µA.

Using the CADC calibration of V_{syn} all neurons lie within three bins around a current of 0 µA. This is shown in figure 3.31d. So with this calibration all neurons

can be calibrated. The standard deviation is 0.014 µA with a mean of 0.009 µA. The minimal current is −0.027 µA and the maximum current is 0.44 µA.

With the fine adjusting algorithm at the end and the CADC based calibration of V_{syn} used during the calibration of I_{bias} the result can be improved. In figure 3.31e all neurons lie within two bins and the standard deviation is 0.010 µA with a mean of 0.016 µA. For this calibration method the minimal current is −0.002 µA and the maximal current is 0.039 µA.

Figure 3.31f shows the distribution of figure 3.20 at 1.18 V. The standard deviation is 0.004 µA with a mean of 0.015 µA. The currents range from 0.008 µA up to 0.024 µA.

The resolution of a voltage capmem cell is around 2 mV [Hock et al., 2013]. The mean of the g_{m} parameter in figure 3.26b is 7 µA/V. So with a perfect calibration of V_{syn} just a resolution of 0.014 µA can be reached. The difference of the maximum and minimum current from figure 3.31f is 0.016 µA. These deviation is in the same order of magnitude as the possible resolution, so this calibration of V_{syn} can not be significantly improved.

These distributions show that the calibration of V_{syn} is possible. The rough and fast algorithm which is based on spike rates is not useful. The fine adjusting algorithm however improves the results of the V_{syn} calibration together with the CADC algorithm. That is the reason why the CADC algorithm should be used whenever a calibration of V_{syn} is needed and at the end one should execute the fine adjusting algorithm to get better calibrated settings for V_{syn}.

3.3 Further investigations

3.3.1 Leak potential

Calibrating the leak potential V_{leak} is really simple, because it is easy to set the membrane potential to it. That is because V_{leak} is connected with a conductance g_{l} realized by an OTA to the membrane. By disabling all inputs beside the leak the membrane potential will equal the leak voltage. In figure 3.32 the CADC measured V_{leak} is compared to the according capmem value. The voltage was determined with the CADC an converted to the calibrated voltage.

If the CADC is not in saturation, the capmem settings are depending linearly on the leak voltage. With this calibration of the CADC it would be possible to calibrate the leak in a range from 0.3 V to 1.2 V. This range can be shifted by using a different calibration for the CADC.

A binary search could be used to calibrate the leak to a desired voltage. In this case it is possible to input a desired voltage and convert it to the according CADC 8-bit value. Calibrating to this value would give the desired voltage to all neurons. But a calibration was not tested in this thesis due to the limited time and because it was not needed for further progression.

3.3.2 Synaptic time constant

With an calibrated I_{bias} it should be possible to calibrate the charge onto the membrane, which is shown in equation 12. With a calibrated reset and threshold potential it should be possible to calibrate R_{syn} with spike rates. But this will just

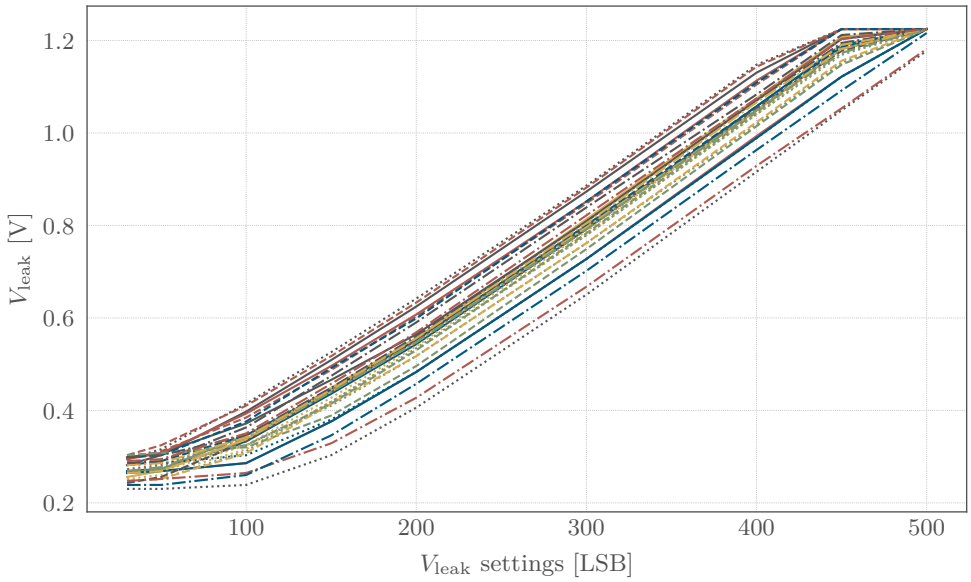


Figure 3.32: The capmem settings of V_{leak} are plotted against the measured voltage of V_{leak}.

work for the excitatory input, because a current onto the membrane is needed for such a calibration. This can be achieved by connecting the excitatory input to the membrane to charge it, while all other terms are disabled.

To measure τ_{syn} the voltage on the synaptic input line has to be investigated. This can be done by recording the trace with the FlySpi and processing the data with the host computer. τ_{syn} is independent of STP, which is disabled for an easier evaluation. The synaptic input line of neuron 14 and $R_{\text{syn}} = 200\,\text{LSB}$ is shown in figure 3.33 during this process. Every 50 µs a spike is send in from the drivers with disabled STP. So all spikes are having the same amplitude and they decay exponentially back to the origin voltage. By doing a exponential fit to every decaying curve τ_{syn} can be determined for every fit individually. The output τ_{syn} is the mean of all five fits.

With this method it is possible to research the dependency of the settings of R_{syn} with the real measured τ_{syn}. This should allow setting a proper range of τ_{syn} to calibrate R_{syn}. Every neuron has to be calibrated that the synaptic time constant is the same for all neurons. This plot is shown in figure 3.34 for a range of 20 LSB up to 900 LSB.

Some curves are not starting at a R_{syn} setting of 20 LSB. That is because sometimes it was not possible for the computer to find a fit due to different reasons. A big problem was to set initial values to get a fit during the search. That is the reason why the fit failed sometimes. These failed fits are not shown in the figure.

This measurement was already done in [Stradmann, 2016] for HICANN-DLSv2. This part of the circuitry was not changed so the results should be comparable. That is the reason why both results are just compared qualitatively. The minimum synaptic time constants are around 1.15 µs in [Stradmann, 2016]. Figure 3.34 shows also time constants in the same order of magnitude for high values of R_{syn} confirming the results from Yannik Stradmann. With a setting of 20 LSB for R_{syn}, the

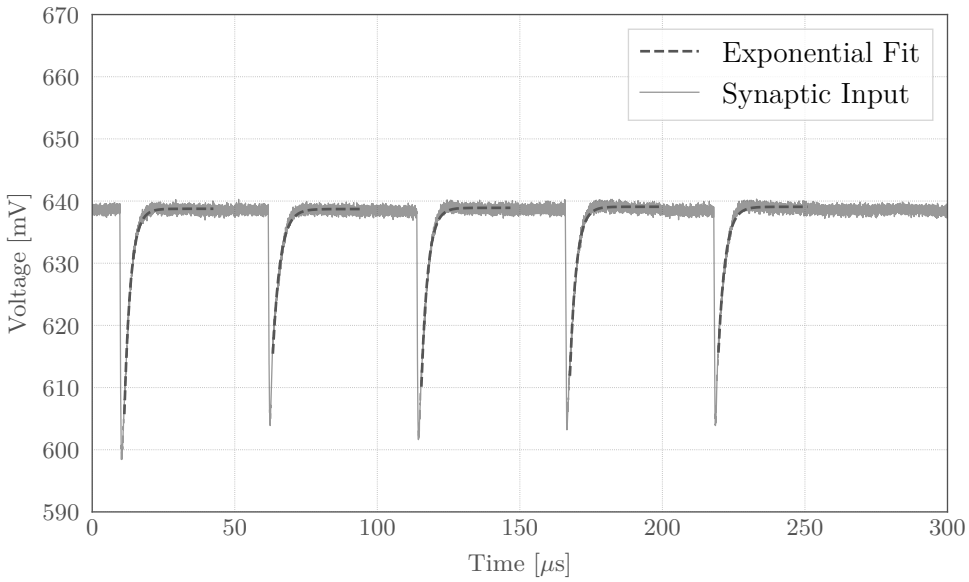


Figure 3.33: Synaptic input with five incoming spikes with STP disabled. Every spike is decaying back to the ground voltage exponentially with τ_{syn}. To determine the latter exponential fits are made to every rising flank.

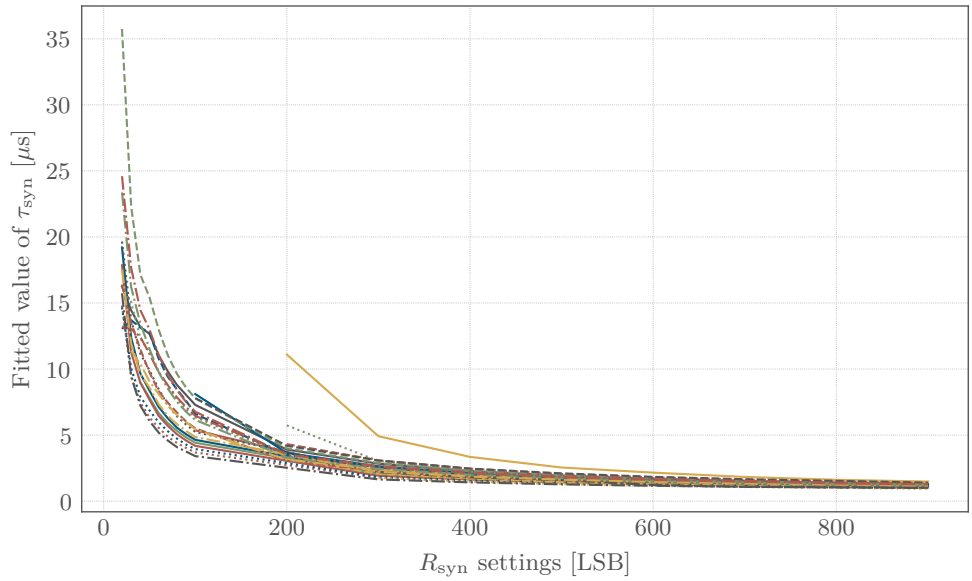


Figure 3.34: Different capmem settings of R_{syn} are plotted against the corresponding τ_{syn} which was measured with the FlySpi and evaluated with a host computer.

time constants vary within 10 µs up to 40 µs for different neurons in [Stradmann, 2016, Figure 3.16]. This range can be confirmed for HICANN-DLSv3 as shown in figure 3.34.

To check the possibility of a spike rate based calibration, the influence of different

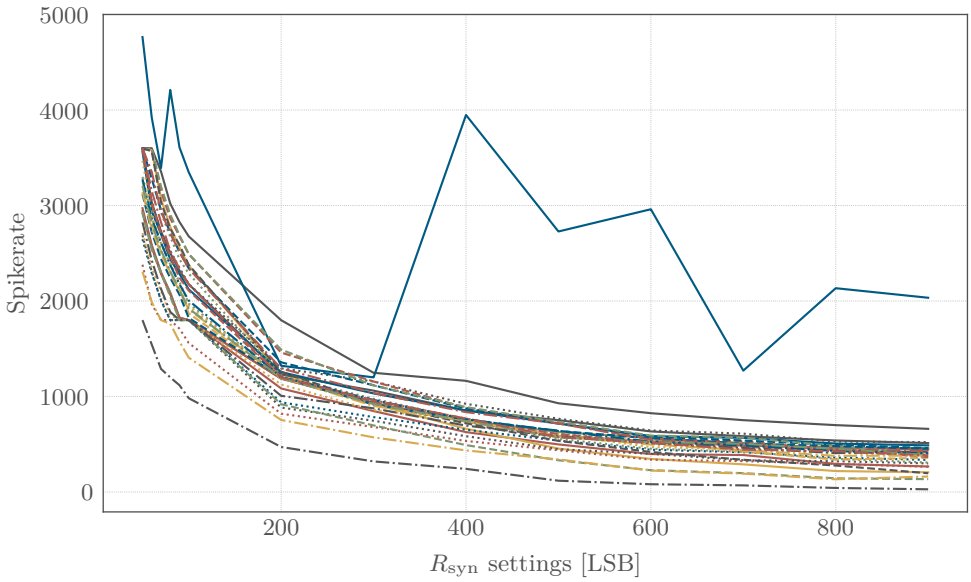


Figure 3.35: Different capmem settings of R_{syn} are plotted against a measured spike rate for each neuron.

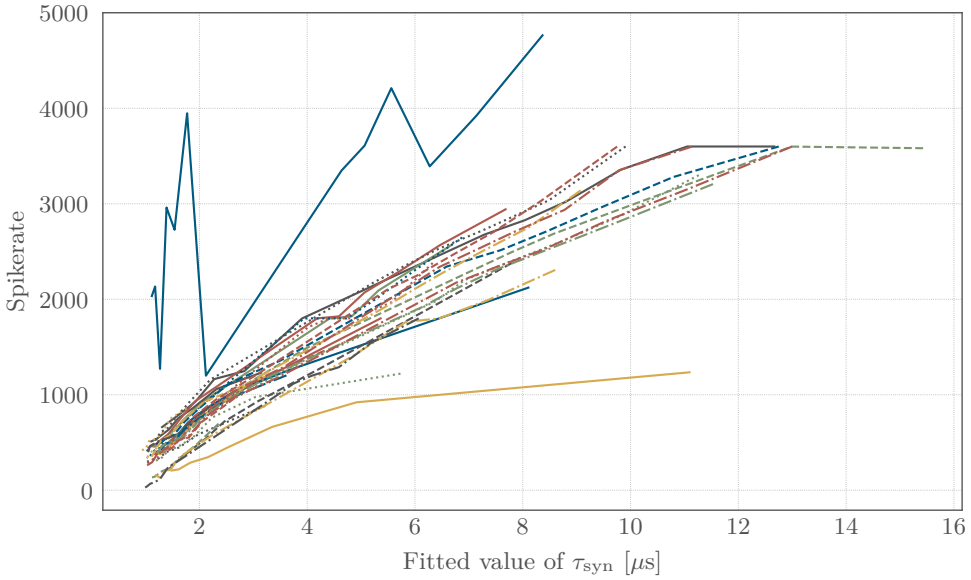


Figure 3.36: The measured τ_{syn} is plotted against the spike rate for the same values of τ_{syn} for each neuron.

R_{syn} settings to the spike rates is plotted in figure 3.35. For this plot a total of 50 bursts containing 120 spikes from synapse driver 0 with an interval of 10 µs between each spike are send in. An interval of 500 µs is between each burst. The spike rate of each burst is summed up to the displayed value.

The dependency of R_{syn} with the spike rate is similar to the dependency of R_{syn}

with τ_{syn} in figure 3.34. It shows that it is possible to calibrate R_{syn} for different spike rates. Again the defect spike counter of neuron 2 shows up.

But the distribution depends hardly on the spike rate. By calibrating on a spike rate of 2000, all values of R_{syn} are located in a range of 0 LSB to 200 LSB. This is changing for a spike rate of 1000, where the settings of R_{syn} are within a range of 100 LSB to 400 LSB. That is the reason why a binary search with a total of 10 runs should be done to calibrate R_{syn} with spike rates.

To check equation 12 the fitted values of τ_{syn} are plotted against the spike rate, which should be a linear dependency. This is shown in figure 3.36. The longer the synaptic time constant, the more charge is going onto the membrane resulting in more spiking. Again neuron 2 is showing up and the fitting faults as discussed above. Also measurement points of failed fits are not shown.

In conclusion the calibration of R_{syn} should be possible with a proper calibrated I_{bias}, V_{res} and V_{thresh}. But one has to be sure that the desired τ_{syn} can be reached for all neurons. Also the search for the right parameters should be done for all 10 bits to get R_{syn} calibrated for all kinds of selected τ_{syn}.

To calibrate the inhibitory input it would be theoretical possible to use a calibrated excitatory input. V_{syn} and I_{bias} have to be calibrated for both synaptic inputs to use this method. Every synapse driver has two output rows, one row can be chosen as excitatory row and the other one as inhibitory row. All synapse weights have to be same. By sending in spikes from the same driver, the different inputs should overlay each other. If both synaptic time constants are the same, the two currents should compensate each other, resulting in no change of the membrane potential. For a higher excitatory τ_{syn} however the membrane potential should rise. Such a calibration can be possible with the PPU.

4 Calibration of Short Term Plasticity

As explained in section 2.7 the `dacen` pulse have to be calibrated with the 4-bit `offset` parameter due to variations in the manufacturing process of the chip, because it is desired to let all 16 synapse drivers on HICANN-DSLv3 process STP in the same way.

A calibration routine should be fast and highly scalable. A first attempt to calibrate STP on HICANN-DLSv3 within 6 min was done with an amplitude based calibration [Weis, 2017]. By using a spike rate based calibration it was possible to lower the runtime by a order of magnitude to 30 s [Weis, 2018]. Both algorithms are processed on a host computer off chip.

Within this chapter the spike rate based calibration of STP will be ported to the PPU and it is limited computational power. It is evaluated whether a faster runtime can be achieved with no loss in accuracy of the calibration. Also research is done regarding HICANN-X with its 128 synapse drivers and 512 neurons. Also the synapse drivers will stay the same for it. It is also tested how the bigger amount of synapse drivers can be calibrated with a smaller runtime.

4.1 Getting started

The calibration depends on spikes sent from the synapse drivers. The total charge onto the membrane from one spike sent from the driver depends on four parameters. One is the length of the `dacen` pulse Δt, because this selects the time how long the capacitor on the synaptic input is charged. Another parameter is the synaptic weight w, which modulates the amplitude of the current onto the synaptic capacitor. Both parameters determine the amplitude $A(\Delta t, w)$ of the voltage at the synaptic input. Also g_m and τ_syn are playing a role, shown in equation 12. The resulting spike rate depends on different neuron parameters like V_thresh, V_res or the membrane capacity.

The mismatch of the synaptic weight w can be neglected, because the mismatch of different synapses with the same weight is below 3 % [Weis, 2018]. Δt is the parameter which have to be calibrated for the STP states, so this parameter should change the spike rate to read it out indirectly. All other parameters are specific for every neuron. But to get the same spike rate for the same Δt for all neurons one have to calibrate the other parameters for spike rates. This was done for I_bias as described in section 3.1.3, while the others were not calibrated. That it is possible to use such calibration for this was proven in [Weis, 2018].

With this algorithm (section 3.1.3) it was possible to have 23 neurons with similar spike rate as shown in figure 3.5f on the experimental setup, which are usable for the STP calibration. On different setups this amount can differ. To calibrate the `offset` parameter one needs to find the neurons which are usable, which is done with an algorithm.

This algorithm searches 16 usable neurons. This amount was chosen because of two reasons. One is that the amount of neurons which have similar spike rates can differ on different chips, but it should be possible to have at least 16 neurons which are usable. Another important reason is that there are 16 synapse drivers and it is tested in section 4.4 if the neurons can be read out in parallel.

The algorithm sends 100 spikes from synapse driver 0 and the spike counters are read out afterwards. Then the rate which the highest occurrence is determined to be the "mean rate". Thats the reason why just 100 spikes are send in that the spikecounts are not to vastly spread and some neurons have the same spikecount. The mean is not taken because otherwise the true mean would be lower than the actual perfect calibrated rate as figure 3.5f shows. Afterwards the difference from the "mean rate" is determined for every neuron. This is done 10 times and every deviation of a neuron is summed up. At the end the 16 neurons which are having the lowest summed up deviation are chosen to be usable.

Compared to the manual search of [Weis, 2018], the classification in usable/not usable neurons equals. Of course there are now more neurons declared to be not usable, because the algorithm just searches for 16 neurons. By doing multiple runs the results still stays the same, so the algorithm should work. If not specified otherwise, this neuron configuration was used to calibrate the `offset` parameter in the STP circuitry. Also the calibration algorithms presented in [Weis, 2018] are using this configuration.

4.2 Basic algorithm

For the calibration all inputs to the neurons are disabled besides the excitatory synaptic input. As explained in section 4.1, the spike rate of the calibrated neurons should be proportional to the `dacen` pulse. So calibrating to a mean rate should be a possible way to calibrate the `offset` parameter.

Just one driver is enabled at once and all spikes of the usable neurons are read out. This should minimize the mismatch of a single neuron by using more statistics. To get the spike rate of a driver, five bursts containing 300 spikes are send in with 10 µs between every single spike and 500 µs between each burst. This are the optimal parameters to have enough statistics, small runtime and no spike counter overflow [Weis, 2018]. Both rows of each synapse driver are enabled to get higher amplitudes and to reduce the synapse mismatch. The spike counter of every usable neuron is summed up for every neuron and burst.

The algorithm itself is based on a binary search with four runs. First off all the mean rate of all drivers is determined. This is done for an `offset` parameter of 8, because the binary search of a 4-bit value is starting with 8. In every run the spike rate of each driver is determined. If this spike rate is below the mean rate the according bit of the run is not set. An additional run is added as done in [Weis, 2018] and also described in section 2.9. In this run the `offset` parameter is raised or lowered by one whether the spike rate is below or above the mean rate. Then the spike rate with this settings is compared to the old one and the configuration with the smallest deviation to the mean rate is used.

The whole STP calibration can be done in under 2 s which is a big improvement compared to a calibration off chip. It is tested in section 4.3 if the smaller runtime leds to a bigger mismatch or not. In this case the runtime should increase on HICANN-X because for every driver all usable neurons are used. That is the reason why in section 4.4 the possibility of reading out neurons parallel is tested to get a better runtime.

The algorithm was not a direct copy of the algorithm presented in [Weis, 2018]. One difference was the start. In this thesis the algorithm starts with 8, because of

the method with the binary search. In the algorithm of Johannes Weis the search was not a standard binary search and it started with 7. Another difference is the mean rate. In this thesis the mean rate is determined at the beginning and is fixed during the whole algorithm. Johannes Weis determined the mean rate of every run and calibrated to this rate. Because of the fixed mean rate just one extra run was added in this algorithm instead of three extra runs.

During the calibration all synapse weights are set to 63. One sets $V_{\text{charge}} = 170\,\text{LSB}$ and $V_{\text{recover}} = 210\,\text{LSB}$. The ramp is precharged with $V_{\text{offset}} = 50\,\text{LSB}$ and the offset capacitors are charged with $V_{\text{zero}} = 300\,\text{LSB}$. The current flowing onto the ramp is $I_{\text{ramp}} = 600\,\text{LSB}$.

4.3 Testing the calibration algorithm

The algorithms are tested by recording real STP traces. The mismatch between each trace should be minimized. Every trace is recorded with the FlySpi and several spikes are send in every 10 µs. The curves of all different synapse drivers recorded on the synaptic input of neuron 12 are plotted in one diagram. The found spikes are marked with a dot in the same color. Every curve was low passed filtered. For recording one used $V_{\text{charge}} = 80\,\text{LSB}$ and $V_{\text{recover}} = 320\,\text{LSB}$ to get nice curves. Other important voltages and currents were not changed to use the former calibration result.

The STP trace for an uncalibrated `offset` parameter is shown in figure 4.1. The spike amplitudes are strongly spread for the depressed state. Some amplitudes in the depressed state are different by a factor of two, which would have a great influence when executing experiments.

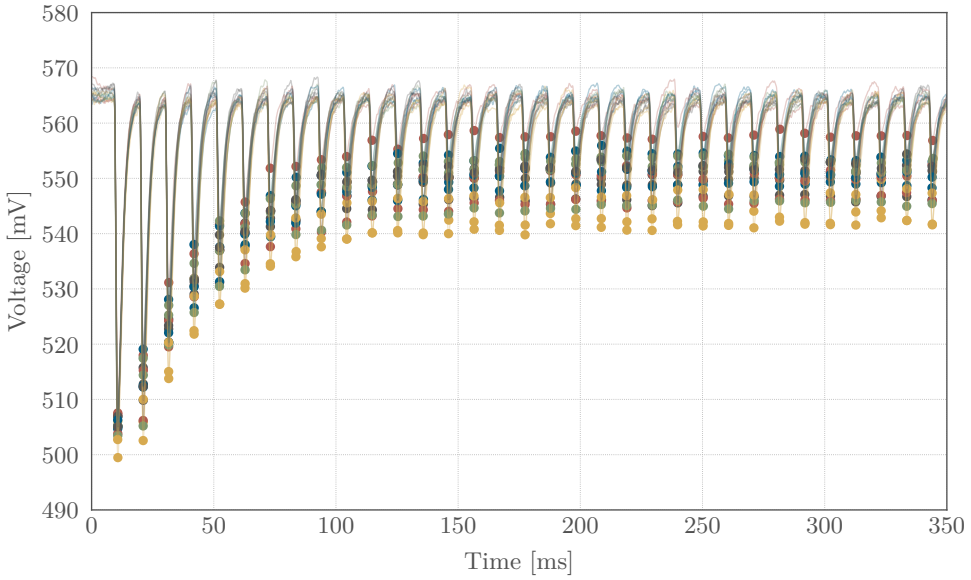


Figure 4.1: Low pass filtered STP trace of all synapse driver, spike peaks are marked with a dot in the same color. The `offset` parameter is not calibrated.

Looking at figure 4.2, the result improved alot. It shows the STP curve after the STP calibration with the PPU. The amplitudes of the different spikes lie much

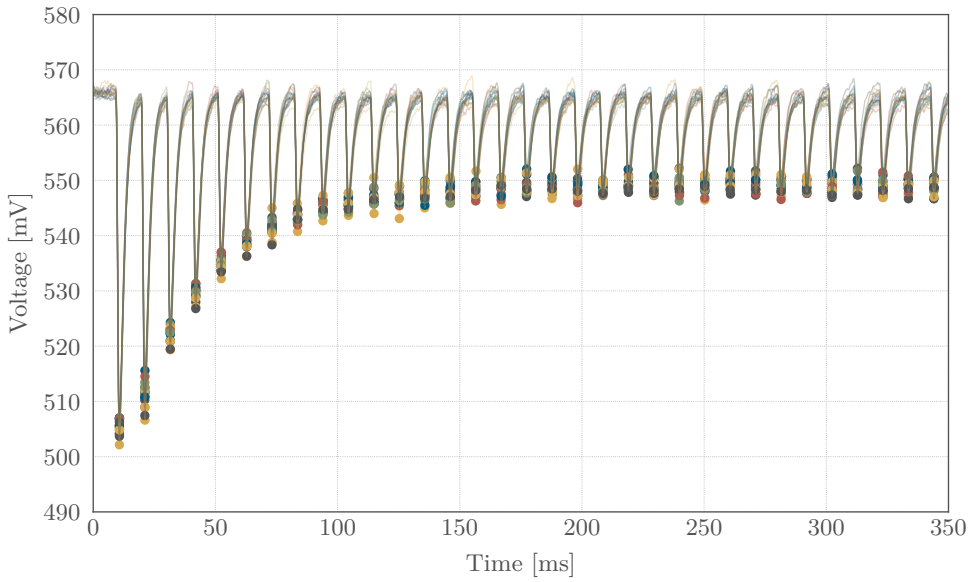


Figure 4.2: Low pass filtered STP trace of all synapse driver, spike peaks are marked with a dot in the same color. The `offset` parameter is calibrated via spike rates with the PPU.

closer. Some spread of the different dots could also come from the spread of the different baselines, which does not have the same voltage either.

To compare the different plots and to get a observable to check the calibration, the amplitudes of the different spikes are plotted. To check the quality of the calibration one can fit an exponential decay to the amplitudes with

$$f(x) = a \cdot \exp(-b \cdot x) + c. \tag{16}$$

The offset of the exponential function c should be the same for all neurons if the `offset` parameter is calibrated.

The uncalibrated amplitudes are shown in figure 4.3. The amplitudes are strongly distributed. Some amplitudes in the depressed state are below 10 mV, while others are above 25 mV. Also the amplitudes at the start are vastly distributed.

The amplitudes of the spike rate based algorithm which was executed on the PPU is shown in figure 4.4. Compared to the uncalibrated curves this looks alot better. In the depressed state the amplitudes are within 15 mV to 18 mV. Compared to the amplitudes at the beginning, which are around 60 mV, the spread is low compared to the amount of depression.

In figure 4.5 the spike rate based algorithm presented in [Weis, 2018] was used to calibrate the `offset` parameter. Compared to the PPU based algorithm there is not a big difference. But this is not surprising because both algorithms are based on the same idea. The only problem with the PPU is that it does not support floating numbers. So the mean rate at the beginning is determined by integer division, which makes it more inaccurate. But the comparison of the two plots show that this is not a big factor.

Figure 4.6 shows the different fitted exponential offsets of different calibrations.

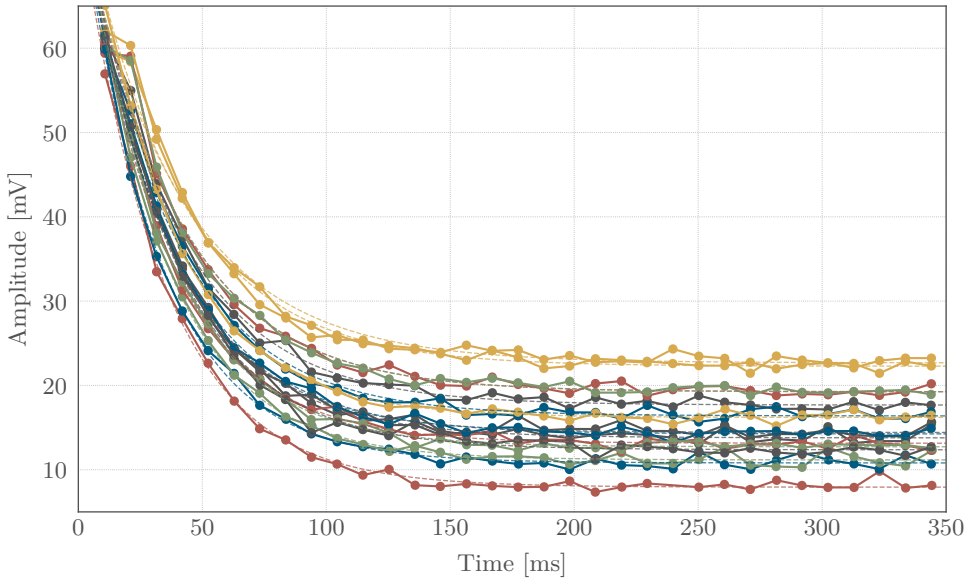


Figure 4.3: Amplitudes of the STP traces of all drivers. The dashed line is the result of the exponential fit. The `offset` parameter is not calibrated.

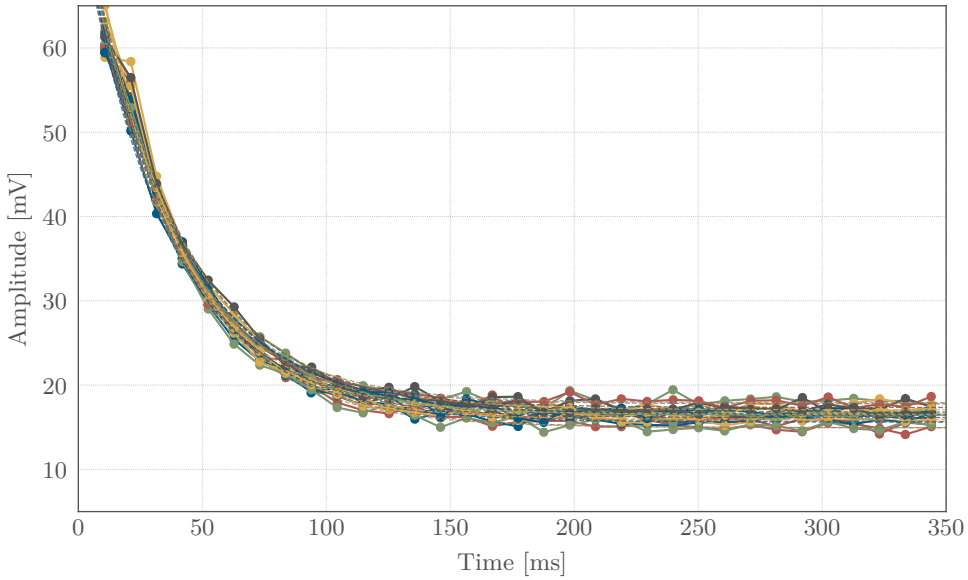


Figure 4.4: Amplitudes of the STP traces of all drivers. The dashed line is the result of the exponential fit. The `offset` parameter is calibrated via spike rates with the PPU.

This parameter can be used to check the quality of the calibration, because it is desired that this parameter is the same for all drivers. Figure 4.6a shows the distribution for an uncalibrated `offset` parameter. It is strongly spread and some drivers are below 10 mV and some above 20 mV, which is a factor of 2. Figure 4.6b however shows the distribution with the `offset` parameter calibrated with the spike

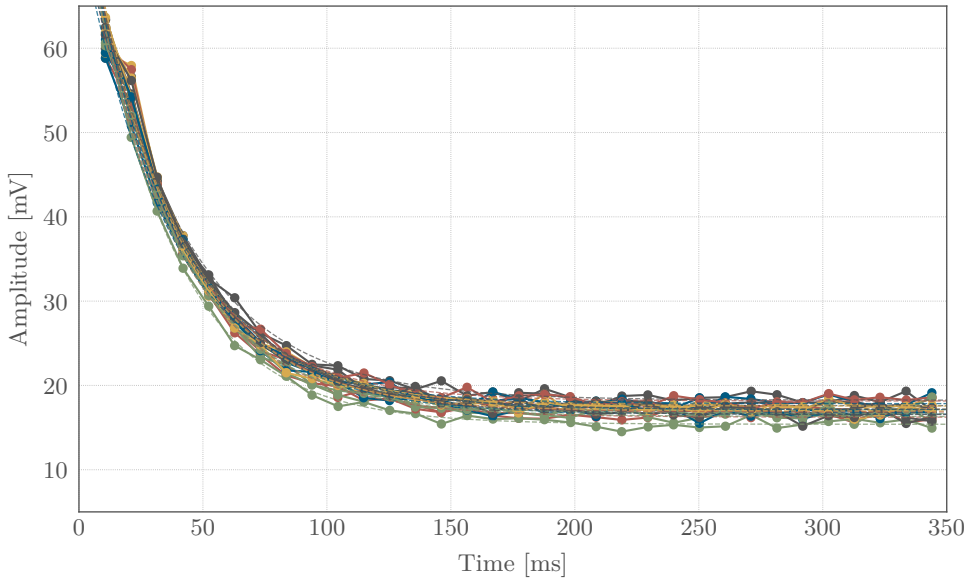


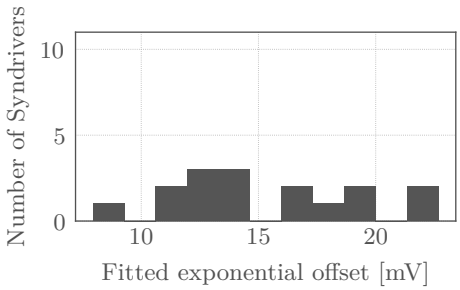
Figure 4.5: Amplitudes of the STP traces of all drivers. The dashed line is the result of the exponential fit. The `offset` parameter is calibrated via spike rates with the algorithm from [Weis, 2018].

rate approach on to the host computer and the neuron calibration manually found in [Weis, 2018]. Figure 4.6d shows the distribution with the calibration executed on the PPU, while figure 4.6c shows the distribution with the calibration executed on the host computer. Both had the same initial conditions because the neuron configuration was determined by the algorithm described in section 4.1. All calibrated exponential offsets are within three bins and the results are nearly equal. So it is possible to calibrate STP on the PPU by observing spike rates. The possibility to calibrate STP shows that the algorithms to calibrate the neuron parameters, shown in section 4.1, is working and also the right usable neurons are chosen.

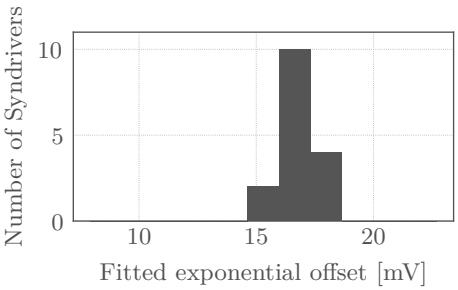
The uncalibrated exponential offset has a standard deviation of 4.14 mV, which is a relative deviation of 27.1 %. The spike rate based calibration shown in [Weis, 2018] has a standard deviation of 0.91 mV, which is 5.5 % of the mean. With calibrating the neurons as described in section 4.1 with the algorithm, the calibration of STP as done in [Weis, 2018] has now a standard deviation of 0.75 mV. The relative deviation is 4.4 % now. The algorithm for finding the settings of the neuron is better than manually adjusting them. By using also the PPU for calibrating STP one gets a standard deviation of 0.87 mV, which is 5.3 %. This is also a big improvement compared to the uncalibrated drivers.

So it is possible to calibrate STP on every HICANN-DLSv3 chip in under 14 s. 12 s are needed to get the calibrated neuron parameters and the actual STP calibration takes 2 s. The neuron parameters can be calibrated in under 2 s by using the CADC V_{syn} calibration, but it is recommended to use the slower method as described in section 3.1.3, because it is more precise.

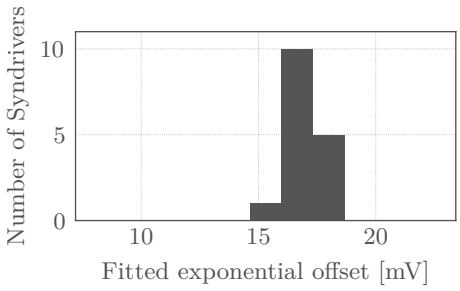
The runtime of the neuron calibration will not change on HICANN-X, because the most time one waits for incoming spikes and evaluating these spikes can be done



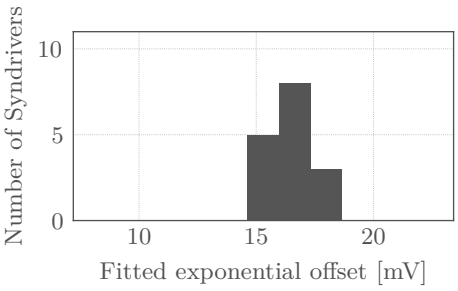
(a) Uncalibrated



(b) Spike rate based calibration from host computer with neuron parameters found in [Weis, 2018].



(c) Spike rate based calibration from host computer with neuron parameters determined by algorithm (section 4.1)



(d) PPU based spike rate calibration with neuron parameters determined by algorithm (section 4.1)

Figure 4.6: Each histogram is showing the fitted exponential offset of each driver for different calibrations and calibration conditions.

pretty fast. But the STP calibration will take more time if the same algorithm is used because of the bigger amount of synapse drivers. That is the reason why in section 4.4 the parallel readout of different neurons for different drivers is tested, to get a faster calibration of STP on HICANN-X.

To check the general functionality of this calibration different chips were used to confirm the results from chip 8. For this purpose chip 3 and chip 9 were calibrated and the STP curves were recorded after a calibration. Similar results were achieved on both chips, confirming the functionality of the whole STP calibration algorithm.

4.4 Parallel neuron readout

In section 4.1, 16 neurons were declared as useful. The number of 16 was not random chosen, because there are 16 synapse drivers. So there was the idea of reading out the neurons in parallel to get a better runtime. Every neuron belongs to one driver and all rates are determined by this neuron. In the current algorithm spikes are send in just from one driver at once and all usable neuron counters are read out to get enough statistic and minimize the error. With parallel readout all synapse drivers would send in spikes at once and the belonging neurons are evaluated. This should be possible to do in under 1 s.

This can be done by changing the address of the synapses for the desired neuron/-

driver combination. So in the column of every usable neuron are the addresses of two synapses changed, which are in the belonging synapse driver row. This should look like a diagonal in the synapse array. This new address have to be the same for all these synapses, as long it is different from the addresses of the other synapses. By sending in spikes with this address, every usable neuron receives just the spikes from its belonging driver. It would also be possible to set all weights to zero except for the synapses described above.

This was implemented as calibration by determining the mean rate for all neurons with its drivers with an `offset` value of 8. Also the rest of the algorithm is similar to the algorithm described in section 4.2, the only difference is that the spike rate is now just determined from one neuron instead of summed up from different neurons.

The decreased runtime however did not justify the result of this calibration, because the `offset` parameter was not well calibrated. The result was better than uncalibrated, but was not close as precise as the former algorithm. The mismatch of a single neuron is too big to use the parallel readout of a single neuron for the STP calibration. To use such an algorithm all neurons have to behave the same, but the neuron calibration is not exact enough to reach this.

Because there are more neurons on HICANN-X, a total of 512 neurons, it was investigated in figure 4.7 how many neurons are needed for a parallel readout to get good results. One does not want all usable neurons on HICANN-X to calibrate a single driver at once, which would take alot more time because there are 128 synapse drivers on HICANN-X.

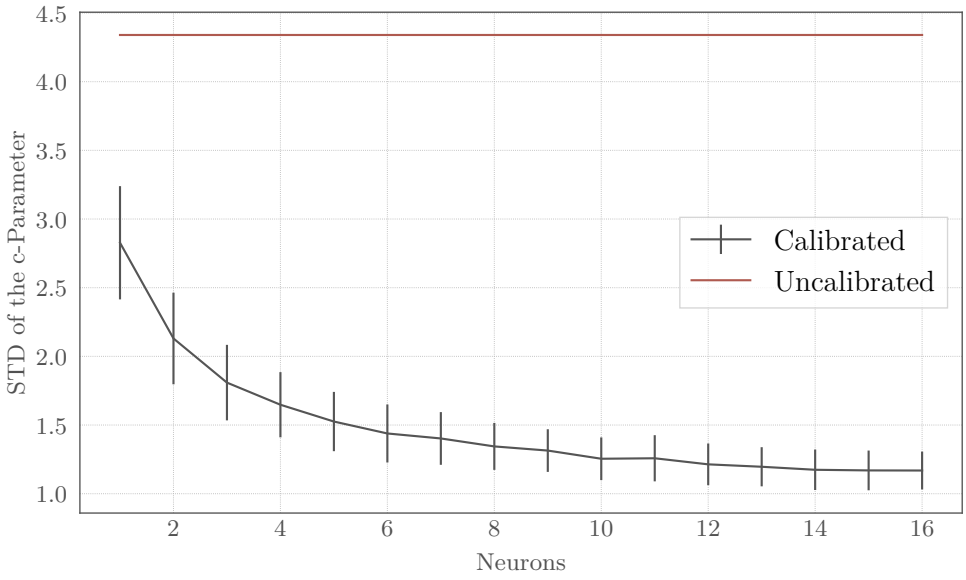


Figure 4.7: The amount of neurons which are read out together to calibrate the `offset` parameter is plotted against the standard deviation of the offset of the exponential fit to estimate how many neurons can be read out in parallel on HICANN-X to get a sufficient STP calibration.

In the figure the determined standard deviation of the offset from the exponential fit of the amplitudes is determined for every calibration. The x-axis is showing the

number of neurons which are read out in parallel. Because of its small size it was not possible to really read out the amount of neurons in parallel. Therefore for every driver and amount random neurons out of the usable are chosen and every driver is calibrated with them. The mean rate is determined with all usable neurons. For every amount of neurons are calibration 100 runs made and the standard deviation of the standard deviation of the different exponential fit offsets is also added.

One can see in figure 4.7 that even the readout of one single neuron is better than letting it uncalibrated. But the more neurons are read out in parallel, the better the result. It seems to approach to a limit. For HICANN-DLSv3 a readout of 8 neurons in parallel should be a sufficient calibration.

This plot have to be redone for HICANN-X to decide how many neurons should be readout in parallel. But by reading out an amount of 8 or 16 neurons in parallel, the runtime of 2 s should not be exceeded. By doing this, also the whole STP calibration with its belonging neuron configuration can be done in under 14 s for every HICANN-X chip.

5 Discussion and Outlook

In this thesis, different neuron parameters and the STP circuitry were calibrated with the PPU. The only observables used for these calibrations were spike rates and the CADC.

It was possible to calibrate V_{syn}, with two different approaches. By using only spike rates it was possible to calibrate this value in under 6 s. The runtime should not change significantly on HICANN-X. However it was not possible to calibrate all neurons with this approach because of a hysteresis effect. By using the CADC to calibrate V_{syn}, this problem was solved. Also, the complete calibration can be executed in 250 ms. The mismatch however of the different neurons is a little bit higher than with the spike rate calibration. A two-stage approach, applying a spike rate based fine-tuning after executing the CADC based algorithm, is recommended for maximal precision. Both methods can be used for the calibration of excitatory and inhibitory input, but in this thesis it was just tested for the excitatory input. The calibration results are summarized in table 1.

Also I_{bias}, which determines the g_{m} value of the OTA, was calibrated by using three different methods. It was the desire to reach two goals. The first goal was to get a suitable calibration for the STP circuit which was previously manually searched [Weis, 2018]. It was possible to reproduce the results of the manual search with an algorithm. This was done together with the calibration of V_{syn} in 12 s, within which the V_{syn} calibration took around 10 s. For HICANN-X the runtime should be approximately the same. The other goal was to calibrate the OTA to equalize the individual neurons' responses to the same stimuli. A DAC on the baseboard was used to clamp the synaptic input line to a fixed voltage, resulting in a constant current onto the membranes. Two approaches were tested, one based on the readout of spike rates and one determining the slope of the membrane voltage with the CADC. With both approaches the whole OTA can be calibrated in under 10 s including the calibration of V_{syn}. Again, on HICANN-X this algorithms should have an equal runtime. Both approaches performed well and no calibration approach outstands the other one. Also here just the excitatory input was calibrated, but it is also easily possible to convert the two algorithms which are calibrating the OTA to the inhibitory input. The used observables however are depending on the mismatch of the membrane capacitors, making both methods dependent of this mismatch. Table 1 collects the standard deviations of the different calibrations.

The reset potential V_{res} was also calibrated with the PPU. The CADC was used to calibrate this parameter. The whole calibration worked (results in table 1) and the runtime is below 1 s. V_{thresh} is another parameter which was calibrated. By using the leak, the spike counters and the CADC it was possible to calibrate the threshold in under 1 s. On HICANN-X it should be possible to run both calibrations within a similar runtime. Both parameters also had an influence on the spike rate based calibration of I_{bias}.

Also, the possibility of calibrating R_{syn}, to get a calibrated τ_{syn} for all neurons, was investigated. Because of the formerly calibrated I_{bias}, V_{res} and V_{thresh} it should be possible to calibrate this parameter with spike rates only. Because of the limited time it was not possible to implement an algorithm and to prove this hypothesis. But in this case just the excitatory R_{syn} can be calibrated with this method, because

spikes send in from the synapse drivers are necessary for this calibration. For the inhibitory input another method was discussed based on a calibrated R_{syn} for the excitatory input.

All of these calibration methods can be executed within seconds and are working, as table 1 shows. The runtime should not change significantly on HICANN-X for all neuron calibrations. This shows that some observables are enough to calibrate different parameters on the chip.

Calibrating STP with spike rates can be done in 2 s with the PPU instead of 30 s from the host computer [Weis, 2018]. The results of both calibrations are very comparable, so calibrating STP should be done on the PPU. Just the amplitude-based calibration [Weis, 2017] is better, but its runtime of 6 min is higher by two orders of magnitude. This trade-off must be decided based on the experimenters use-case.

Things are also changing on HICANN-X, with its higher amount of neurons and synapse drivers. The runtime of the amplitude based calibration will increase drastically [Weis, 2018]. It scales with the amount of neurons. But by using the spike rate calibration on the PPU with parallel readout, which is possible as shown in this thesis, the whole STP calibration on HICANN-X should be also possible in 2 s. This runtime advantage definitely justifies the slightly imperfect results of the spike rate based calibration compared to the amplitude-based calibration.

Including the matching neuron calibration, STP can be calibrated in 14 s for HICANN-DLSv3 and HICANN-X. This was tested for different DLSv3 chips. And if the calibration of τ_{syn} is working properly, the whole individual neuron calibration for STP can become unnecessary.

However there are also some disadvantages to execute calibration algorithms on the PPU. It is not possible to give the parameters of the neuron as absolute values, as a fit can not be executed on the PPU. Instead, the PPU can be used to calibrate all neurons to behave like a chosen neuron. This chosen neuron can be calibrated as desired, while a PPU based calibration reduces the mismatch between each neurons.

The final results of the different calibration algorithms performed during this thesis are collected in table 1. For the calibration of V_{syn} the standard deviation and the min-max gap ΔI are shown. For the g_{m} value are the relative deviations given. For V_{res} and V_{thresh} is the standard deviation for the uncalibrated values given compared to the calibrated ones. For the neuron calibration of the STP circuit are the relative deviations of the incoming spike rates given. Last but not least the standard deviation of the exponential offset is given for different calibration methods of the `offset` parameter in the STP circuitry.

In a future step more calibrations should be developed for this neuromorphic circuits. One example is the synaptic time constant. Theoretical methods were developed during this thesis to calibrate τ_{syn} for both synaptic inputs. But also the transconductance g_{l} of the leak OTA should be calibrated to set the membrane time constant τ_{m}.

For future chip generations the OTAs of the synaptic input have to be revised. According to the specification the linear range should be given for $\Delta V = 200\,\text{mV}$, but measurements show that the OTA starts to saturate even for lower deviations. Also the design on-chip have to be investigated to lower the parasitic deviations of the membrane capacity for boundary neurons.

Parameter	Uncalibrated	Method	Calibrated	Comment
V_{syn}	$\sigma = 0.147\,\mu\text{A}$ $\Delta I = 0.570\,\mu\text{A}$	spike rate based calibration	$\sigma = 0.190\,\mu\text{A}$ $\Delta I = 1.304\,\mu\text{A}$	Outlier making this calibration not usable
		CADC based calibration	$\sigma = 0.014\,\mu\text{A}$ $\Delta I = 0.071\,\mu\text{A}$	No outlier anymore
		CADC and spike rate	$\sigma = 0.004\,\mu\text{A}$ $\Delta I = 0.016\,\mu\text{A}$	Resolution of V_{syn} reached
g_{m}	20.9 %	spike rate based calibration	24.7 %	Outlier making this calibration not usable
		CADC based V_{syn} spike rate I_{bias}	14.1 %	ΔV uncalibrated (high I_{bias} values)
		CADC based V_{syn} spike rate I_{bias}	9.6 %	ΔV calibrated (high I_{bias} values)
		slope CADC calibration	10.0 %	(high I_{bias} values)
		CADC based V_{syn} spike rate I_{bias}	9.4 %	ΔV calibrated (lower I_{bias} values)
		slope CADC calibration	7.4 %	(lower I_{bias} values)
V_{res}	$\sigma = 35.5\,\text{mV}$	-	$\sigma = 3.6\,\text{mV}$	-
V_{thresh}	$\sigma = 33.6\,\text{mV}$	-	$\sigma = 3.9\,\text{mV}$	-
Neuron STP calibration	55.0 %	One search for I_{bias}	4.1 %	For 21 neurons with similar rate
		Two searches for I_{bias}	1.8 %	For 23 neurons with similar rate
STP calibration	27.1 %	Spike rates on host computer	5.5 %	Manually found capmem values
		Spike rates on host computer	4.4 %	Capmem values found with PPU
		Spike rates on PPU	5.3 %	Capmem values found with PPU

Table 1: Collection of different calibration approaches executed during this thesis. Their deviations are also collected in this table.

References

S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, and K. Meier. A highly tunable 65-nm cmos lif neuron for a large scale neuromorphic system. In *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, pages 71–74, Sept 2016. doi: 10.1109/ESSCIRC.2016.7598245.

S. A. Aamir, P. Müller, L. Kriener, G. Kiene, J. Schemmel, and K. Meier. From lif to adex neuron models: Accelerated analog 65 nm cmos implementation. In *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4, Oct 2017. doi: 10.1109/BIOCAS.2017.8325167.

S. A. Aamir, P. Müller, G. Kiene, L. Kriener, Y. Stradmann, A. Grübl, J. Schemmel, and K. Meier. A mixed-signal structured adex neuron for accelerated neuromorphic cores. *IEEE Transactions on Biomedical Circuits and Systems*, pages 1–11, 2018a. ISSN 1932-4545. doi: 10.1109/TBCAS.2018.2848203.

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 neuromorphic architecture. *IEEE Transactions on Circuits and Systems I: Regular Papers*, pages 1–14, 2018b. ISSN 1549-8328. doi: 10.1109/TCSI.2018.2840718.

Sebastian Billaudelle. Design and implementation of a short term plasticity circuit for a 65 nm neuromorphic hardware system. Masterarbeit, Universität Heidelberg, 2017.

Mikhail Borodin, Kaushik De, Jose Garcia Navarro, Dmitry Golubkov, Alexei Klimentov, Tadashi Maeno, David South, and Alexandre Vaniachine. Big data processing in the atlas experiment: Use cases and experience. *Procedia Computer Science*, 66:609 – 618, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.11.069. URL `http://www.sciencedirect.com/science/article/pii/S1877050915034183`. 4th International Young Scientist Conference on Computational Science.

Romain Brette and Wulfram Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, 94(5):3637–3642, 2005. doi: 10.1152/jn.00686.2005. URL `https://doi.org/10.1152/jn.00686.2005`. PMID: 16014787.

Ursula Dicke and Gerhard Roth. Neuronal factors determining high intelligence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 371(1685), 2016. ISSN 0962-8436. doi: 10.1098/rstb.2015.0180. URL `http://rstb.royalsocietypublishing.org/content/371/1685/20150180`.

Carlos Eyzaguirre and Stephen W. Kuffler. Further study of soma, dendrite, and axon excitation in single neurons. *The Journal of General Physiology*, 39(1):121–153, 1955. ISSN 0022-1295. doi: 10.1085/jgp.39.1.121. URL `http://jgp.rupress.org/content/39/1/121`.

Michael Feldman. Summit Up and Running at Oak Ridge, Claims First Exascale Application, 2018. URL https://www.top500.org/news/summit-up-and-running-at-oak-ridge-claims-first-exascale-application/. called on: 2018-07-30.

Diasynou Fioravante and Wade G Regehr. Short-term forms of presynaptic plasticity. *Current Opinion in Neurobiology*, 21(2):269 – 274, 2011. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2011.02.003. URL http://www.sciencedirect.com/science/article/pii/S0959438811000298. Synaptic function and regulation.

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, Feb 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, Dec 2013. ISSN 0018-9340. doi: 10.1109/TC.2012.142.

Matthias 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, Sept 2013. doi: 10.1109/ECCTD.2013.6662229.

Gerd Kiene. Mixed-signal neuron and readout circuits for a neuromorphic system. Masterthesis, Universität Heidelberg, 2017.

Aron Leibfried. Characterization of a pll circuit used on a 65 nm analog neuromorphic hardware system, 05 2018.

K. Meier. Special report : Can we copy the brain? - the brain as computer. *IEEE Spectrum*, 54(6):28–33, June 2017. ISSN 0018-9235. doi: 10.1109/MSPEC.2017.7934228.

Alberto E. Pereda. Electrical synapses and their functional interactions with chemical synapses. *Nature Reviews Neuroscience*, 15:250 EP –, Mar 2014. URL http://dx.doi.org/10.1038/nrn3708. Review Article.

Wade G. Regehr. Short-term presynaptic plasticity. *Cold Spring Harbor Perspectives in Biology*, 4(7), 2012. doi: 10.1101/cshperspect.a005702. URL http://cshperspectives.cshlp.org/content/4/7/a005702.abstract.

J. Schemmel, D. Bruderle, K. Meier, and B. Ostendorf. Modeling synaptic plasticity within networks of highly accelerated i amp;f neurons. In *2007 IEEE International Symposium on Circuits and Systems*, pages 3367–3370, May 2007. doi: 10.1109/ISCAS.2007.378289.

J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 1947–1950, May 2010. doi: 10.1109/ISCAS.2010.5536970.

J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323 – 332, 2012. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2012.02.016. URL `http://www.sciencedirect.com/science/article/pii/S0893608012000457`. Selected Papers from IJCNN 2011.

Yannik Stradmann. Characterization and calibration of a mixed-signal leaky integrate and fire neuron on hicann-dls. Bachelorarbeit, Universität Heidelberg, 2016.

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

Misha Tsodyks, Klaus Pawelzik, and Henry Markram. Neural networks with dynamic synapses. *Neural Computation*, 10(4):821–835, 1998. doi: 10.1162/089976698300017502. URL `https://doi.org/10.1162/089976698300017502`.

Misha V. Tsodyks and Henry 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. ISSN 0027-8424. doi: 10.1073/pnas.94.2.719. URL `http://www.pnas.org/content/94/2/719`.

Johannes Weis. Testing of a neuromorphic short term plasticity circuit, 11 2017.

Johannes Weis. Characterization and calibration of synaptic plasticity on neuromorphic hardware. Bachelor, Universität Heidelberg, 2018.

Robert S. Zucker and Wade 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

First of all, I want to thank Prof. Dr. Karlheinz Meier for giving me the chance to work in this group on this interesting topic.

I also want to thank Dr. Johannes Schemmel for his great leadership and for providing this great hardware.

Also thanks to my supervisors Sebastian Billaudelle and Yannik Stradmann for supporting me during this thesis. You calmly explained the usable methods, the implemented circuitry and you provided helpful solving approaches if I was stuck.

Another thanks goes to Johannes Weis for his nice groundwork in the STP calibration and taking his time to explain me his used code, his ideas and explaining me the influence of different parameters.

Thanks to Sebastian Billaudelle, Yannik Stradmann, Simon Rosenkranz, Malte Prinzler and Oliver Leibfried for proofreading this thesis and the helpful feedback.

I also want to thank the whole Electronic Vision(s) group for the great atmosphere and willingness to help.

Thanks to my friends and my familiy for supporting me during the course of my studies.

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, August 21, 2018 .....
(signature)