Testing of a neuromorphic Short Term Plasticity circuit

Johannes Weis
Supervised by: Sebastian Billaudelle, Yannik Stradmann

November 2017

1 Introduction

In the Electronic Vision(s) Group at Heidelberg University, we are designing an Application Specific Integrated Circuit (ASIC) containing analog neurons and synapses, as part of the Human Brain Project. In this report, we will show testing results of parts of HICANN-DLS 3, mainly the synapse driver circuit and its Short Term Plasticity feature.

The HICANN-DLS 3 prototype chip consists of 32 neurons and 1024 Synapses, arranged in an array of 32 rows and 32 columns. An array of 16 synapse drivers is located at the left side of the synapse matrix, each instance driving two synaptic rows. Every synapse column is connected to a neuron. The drivers distribute a set of digital signals along the row consisting of an address and a `dacen`-signal. Synapses will listen only to their configured address, which is stored in a 6-bit SRAM. The synaptic output amplitude is changed by the synaptic weight parameter and the length of the `dacen`-pulse.

Short Term Plasticity (STP) is processed in the synapse driver. In biology, forwarding action potentials through a synapse means emitting neurotransmitters. Many action potentials in a short timeframe have an impact on the concentration of neurotransmitters. The received input on the following neuron can get larger if there are still neurotransmitters present, or may get smaller if the synapse runs out of available vesicles containing the transmitters. We will call this facilitation and depression for increase and decrease of received amplitudes, respectively. While in biology both processes can happen at the same time, the chip is only able to process one exclusively. Further information can be found in [ZR02] and [Sch+07], which refers to the Spikey chip.

For a depressing configuration, introducing an *inactive* partition I and a *recovered* partition R of neurotransmitters, the implemented STP mechanism can be described with a set of differential equations, (1) and (2). With every action potential, some of the recovered transmitters are utilized and added to the inactive partition. This is controllable by the *utilization* of synaptic efficacy factor, U_{SE}. Over time, neurotransmitters recover into the recovered partition. The recovery is an exponential process with a time constant τ_{rec}.

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

$$\frac{dI}{dt} = -\frac{I}{\tau_{\mathrm{rec}}} + U_{\mathrm{SE}} \cdot R \cdot \delta(t - t_{\text{action potential}}) \tag{2}$$

To implement this behaviour in hardware, the voltage stored on a capacitor represents the current state of neurotransmitters. There are 64 capacitors in every synapse driver, one for each address. Receiving a spike, the according capacitor is connected with an update circuit holding a voltage V_{charge}. Sharing charge, the voltage V_{STP} on the storage capacitor changes. The capacity of the update capacitor is configurable, this allows setting the utilization factor.

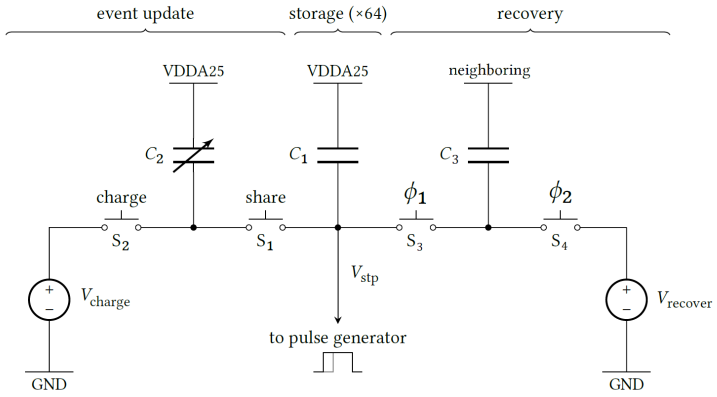


Figure 1: Schematic of the STP update and recovery switching mechanism. The capacity C_2 is configurable to set the utilization parameter. The capacitor C_3 transmits charge between V_{STP} and V_{recover}. Its capacity is much smaller and the switching continuous. Figure adapted from [Bil17].

A minimal utilization is given by the capacity of the switched lines during readout of the V_{STP} voltage. This parasitic capacity is originally charged to V_{charge}. Recovery is accomplished using a similar mechanism but switching continuously with a high frequency and a small capacity. This makes V_{STP} recover towards V_{recover} exponentially. The STP update and recovery switching circuit is shown in figure 1.

The impact of STP on post-synaptic output amplitudes is transmitted by the length of the `dacen` pulse. The voltage is compared to a voltage ramp produced in the driver, and the `dacen` pulse is triggered once the voltages cross each other. This should map V_{STP} linearly to output amplitudes. Since the comparison mechanism, like the rest of the chip, is subject to manufacturing inaccuracy, the starting voltage of the ramp can be controlled by a 4 bit offset parameter. This allows the STP amplitudes of different drivers to be calibrated in order to match each other. Details on the circuitry are presented in [Bil17].

2 Available configuration range

For the STP functionality, there are two main configurable parameters for each driver: The utilization of synaptic efficacy U_{SE} and the recovery time constant τ_{rec}. There is a 4 bit configuration range available for both values. To make the circuit available to the end users, it is important to find a translation between the model's and implementation's parameters. The offset parameter however should be invisible to the user and is used for calibration only. We will later present an offset calibration and further investigations on the recovery process. The U_{SE} parameter was looked at first, still acquiring voltage traces with an oscilloscope. Therefore, only 10 measurements for each utilization setting were taken, each using the mean of 10 traces to readout spike amplitudes. This was necessary due to large noise levels making it difficult for an automated process to determine spikes in the trace. Averaging can decrease the observed noise significantly.

To measure the utilization, we send 20 spikes into the driver with recovery turned off completely. We send a spike every 40 µs. While STP amplitudes are fully recovered at the beginning, they will be fully depressed at the end of the spike burst. Since recovery is turned off, the amplitude ratio of the first spikes (while they are being depressed) yields the

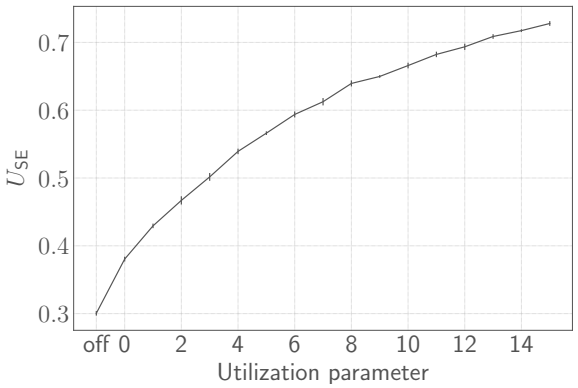


Figure 2: Configurable range for U_{SE} parameter. Each data point shows statistics from 10 measurements averaging over 10 traces each. Measured on driver 5, neuron 13, address 63 on chip 8.

used synaptic efficacy U_{SE}. This can be repeated for utilization settings from 0 to 15 and utilization turned off completely: Disabling the `enshare` parameter on the synapse driver will reduce U_{SE} to a minimal possible value. The charge on the STP capacitor is shared only with necessary lines which mean parasitic capacity, not with an additional capacity in the STP update handler. In [Bil17] (figure 27b), an available range from about 0.25 to 0.72 is expected.

The results from our experiment (figure 2) fit well to the expected behaviour. Regarding the small number of measurements, this shows a well-working circuit.

3 Offset calibration

The offset parameter allows a constant shift of output amplitudes, for all V_{STP} voltages. To reduce mismatch effects between synapse drivers, this offset has to be calibrated. This is done using an automated calibration script. For the following measurements, the onboard-ADC is used.

3.1 Calibration algorithm

To investigate the mismatch, 25 spikes are sent to a driver. STP parameters are set at medium levels in order to reduce variations in amplitudes, i.e. V_{charge} relatively high and V_{recover} low. This is necessary to prevent the `dacen` pulse being saturated at 0 or 4 ns. Furthermore, to find the spikes, a good signal to noise ratio is desired. The used script uses a low-pass and a sobel filter to find edges in the recorded voltage trace. A steep decrease in voltage (edge) is a spike event. In a small timeframe around an edge, the deviation between the baseline and the minimum of the low-pass-filtered trace is treated as the spike amplitude. This method turned out to be effective and efficient.

The calibration script aims at reducing the spread of the individual drivers' output amplitudes. The offset parameters are initially set to 8, which is right in between the available range, 0 to 15. Based on a binary search algorithm, the offset parameters are changed in steps of 4, 2, 1 and 1 in runs 0 to 3. A high output amplitude will be corrected with a higher offset (shifting the amplitude trace towards the baseline), whereas too small signals are set

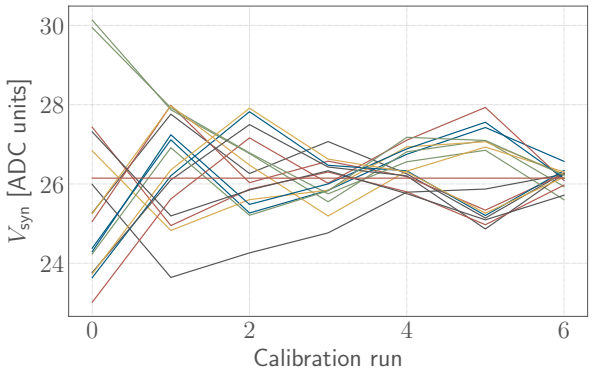
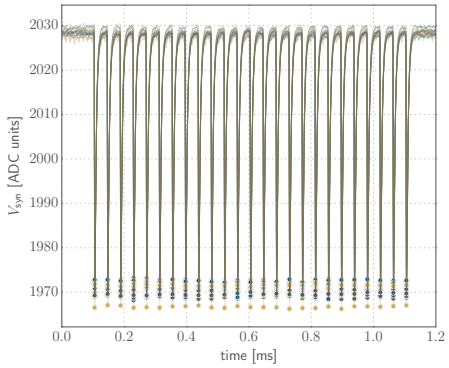


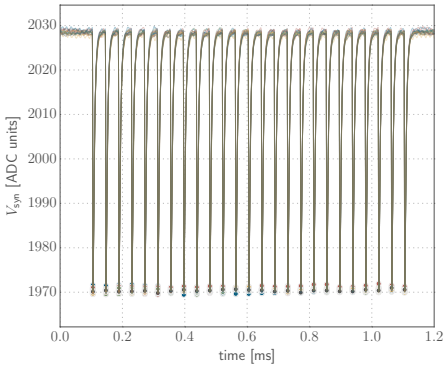
Figure 3: Change of output amplitudes during offset calibration, one track per driver is shown. The final mean amplitude is plotted as a straight line.

to a lower offset setting. In run 4, the results of the binary search are ready and investigated. The last change in offset values was only a step by one from run 3 to 4. In an additional fifth run, the offset parameter is shifted towards the same direction again. This means that in runs 3 to 5, 3 consecutive offset settings have been tested. Since the step size of variations is not equal everywhere, this additional run did improve calibration results for some drivers.

The final, *calibrated* result is chosen from results of runs 2 to 5, the setting that is closest to the mean amplitude of all drivers is used. For testing purposes, those now calibrated offset parameters are set up again in run 6, so we can get a plot of all traces with ideal settings. The shifting of amplitudes during the calibration process is shown in figure 3.

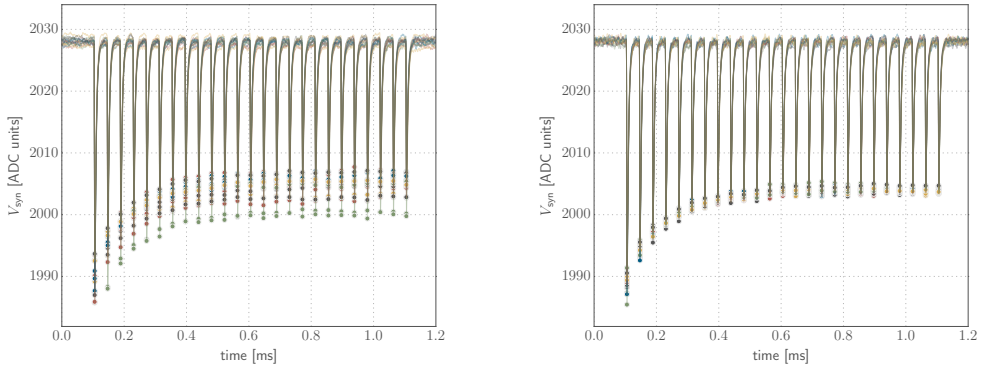


(a) Original synaptic outputs. Mean of 20 traces.



(b) Synaptic outputs after setting weights. Mean of 30 traces.

Figure 4: Recorded traces before and after synapse equalization. Traces are low-pass filtered and were recorded on chip 3, column 0.



(a) Uncalibrated amplitudes. Mean of 20 traces.

(b) Amplitudes after offset calibration. Mean of 30 traces.

Figure 5: Recorded traces before and after offset calibration, $V_{\text{zero}} = 260$. Traces are low-pass filtered and were recorded on chip 3, column 0.

3.2 Synapse equalization

To be able to neglect neuron-specific mismatch, we record all synapse drivers' output pulses via the synaptic input of the same neuron. For each synapse driver, we choose a single synapse out of the respective column. However, the weights of the individual synapses are subject to mismatch, too. Appearing as a multiplicative deviation, it can not be calibrated for via the synapse drivers' offset parameter. Additionally, using another synapse column later would then possibly void the purpose of an offset calibration.

In order to minimize impact of synaptic mismatch on the calibration, synapse weights in the used column are tuned for equal amplitudes. This is done before starting the offset calibration. With STP disabled, an algorithm similar to the one described above is used to find optimal synapse weights, based on a binary search. We got good results here: Figure 4 shows measured amplitudes before and after tuning synapse weights. Mismatch standard deviations are below 1 %–2 % when tuned, $0.3/58 = 0.5\%$ in this plot versus 2.1% original.

To maintain a good signal-to-noise ratio during offset calibration, only weights of 48 and higher were used. Allowing for smaller weights would decrease synapse mismatch further, but could increase statistical fluctuations in measured amplitudes during offset calibration on the other hand.

3.3 Calibration results

Running the offset calibration script takes about 6 minutes. Two plots, before and after offset calibration, are shown in figure 5. We can see that mismatch effects are drastically reduced. The standard deviation of calibrated driver amplitudes is expected to be less than 3% in general. This statement is based on observations on 3 different chips only, there are no proper statistics available yet. For the shown plot, which was measured on synapse column 0 on chip 3, it is only $0.26/26 = 1\%$ while it was 8% before calibration.

While the depressed part of the trace, the steady state, is matched pretty well, the first spikes show greater deviations. Some drivers seem to have a steeper depression curve than others. This issue will be investigated later. Using another synapse column with the existing calibration shows higher deviations, since synaptic mismatch is present for both the original and other column.

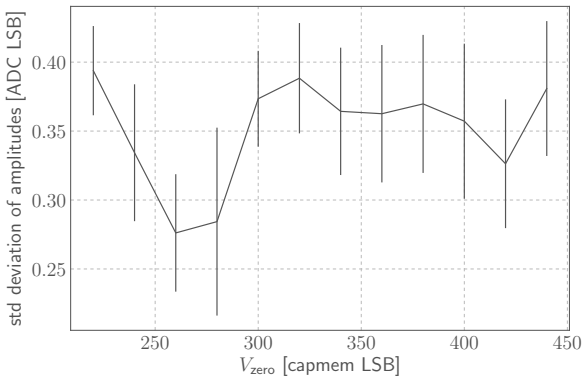


Figure 6: Standard deviation of calibrated amplitudes plotted over V_{zero}. Measured on neuron 0 on chip 3. Error bars indicate standard deviations from 10 calibrations with identical parameters.

Calibration results can be reproduced very good: A deviation of 1 LSB in offset values can be seen for some drivers, but this can be expected if both values yield amplitudes that deviate from the target by an equal absolute value. The quality of calibration is therefore limited by the minimal step size of the offset parameter, which can be set using the V_{zero} voltage. There is an optimal setting for this: Setting V_{zero} too low makes some drivers not reach the mean amplitudes even when set to an offset of 0 or 15, while when setting it too high the outer values of the offset range are never used. When sweeping V_{zero}, a minimum in amplitude standard deviation was visible. The plot is shown in figure 6. A setting of around 240–260 (10-bit capacitive memory setting) accounted for good calibration results here.

4 Testing synapse driver power supply

We did investigate post-synaptic output amplitudes (PSPs) depending on the number of enabled synapse drivers. The `dacen` pulse however has to be driven along the synapse row and edge steepness needs to be maintained in order to not shorten pulses and therefore decrease amplitudes. Since the power supply of all drivers is shared, one could explain drops of amplitudes this way. The main power consuming activity of the synapse driver is driving the `dacen` pulse, therefore the STP feature is disabled here.

We observe the amplitudes of the post-synaptic currents by reading out the internal integrator voltage of the synaptic input circuit inside the neuron. The trace can be read out via an 96 MSamples/s ADC located on the baseboard, or an external oscilloscope. Here, the oscilloscope with a LeCroy passive probe is used to minimize capacity. For this experiment, 70 spikes are sent into the chip. We can then calculate the integral of the whole trace after subtracting its baseline voltage. Out of the whole array, only a single synapse – the one used to stimulate the observed neuron – is configured for the stimulus' address. We activated a step-wise increasing number of "silent" synapse rows to investigate the effect of the increasing power consumption in the buffer stages of the synapse driver circuit. This yields the plot in figure 7. We can see no drops of amplitudes there.

We conclude that the available power supply is well suited here. The statistical deviations of measured amplitudes are much higher than any systematic effect.

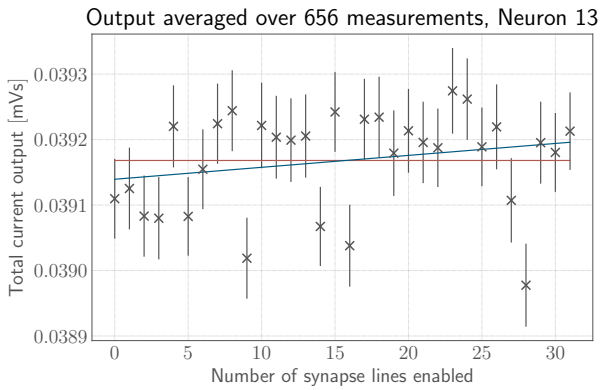


Figure 7: Post-synaptic amplitudes over number of enabled driver lines, testing neuron 13 on chip 8. Error bars on the data points represent errors of the mean ($\sigma/\sqrt{n}$). There is no systematic trend visible. A linear fit shows a small incline here ($(1.8 \pm 1.4) \times 10^{-9}$ mVs), the slope is 0 within 3σ for every tested neuron however (1.28σ here). Even when arranged differently (e.g. plotted over time), data points look alike.

5 Variations in STP behaviour

One of the results of the offset calibration is the fact that different synapse drivers depress their amplitudes with a different steepness. We invested quite some time trying to understand what is happening. There could be mismatch effects in ramp steepness or the comparator itself could cause problems. In particular, we investigated if there is a coupling between V_{zero} and V_{offset} voltages. The ramp capacitor is charged to V_{offset} in the first 2 ns of the 8 ns event processing window in the synapse driver. By sharing charge with a configurable capacity on V_{zero}, the initial ramp voltage can be offset individually per driver (offset calibration). The timeframe 2 ns–4 ns is used for that. Afterwards, from 4 ns to 8 ns, a current I_{ramp} flows on the ramp capacity, the linearly rising voltage on there will cross V_{STP}, so the comparator toggles the `dacen`-pulse.

We designed an experiment to send spikes with increasing distances in time. Spikes are sent $1, 2, 3, 4, ...$ times 8000 FPGA-cycles after the previous one. Starting at time 0, this means spikes at times of $0, 80\,\mu s, 240\,\mu s, 480\,\mu s, 800\,\mu s$ and so on. This is enough that the STP voltage is fully recovered at any spike time, recovery time constants are short (prescaler 4, `recovery` 15). Conducting the experiment on address 15, the STP status is visible by reading out V_{STP}, which behaves as expected.

Depending on the comparator bias I_{bias} and the number of synapse rows (and therefore drivers) enabled, a drop in output amplitudes is present. Overall spike amplitudes decrease when enabling more lines, which is expected considering all synapse weights are set to 63 here. However, amplitudes also decrease when many spikes are transmitted consecutively. This is shown in figure 8. Amplitudes over time look similar to depressing STP, however the STP voltage is always at V_{recover} and should furthermore not depend on the number of enabled drivers. For high I_{bias} settings like 800 LSB, even a small number of enabled drivers is enough, for small I_{bias} values like 200 LSB the effect is only visible when many synapse rows are turned on. The issue can be observed independently from the configured synaptic weight.

Having all drivers turned on, we investigate amplitudes for every synapse row. We plot the standard deviation of amplitudes during the spike-pattern in a colored image representing

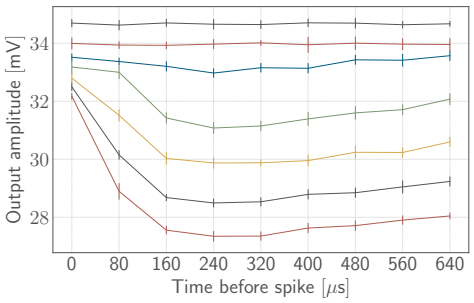


Figure 8: Spike amplitudes when increasing time distance. Drops increase with enabled synapse rows: 0 enabled in top trace, 30 enabled in bottom trace, steps of 5. Chip 03, $I_{\text{bias}} = 500$. Error bars from standard deviations in 10 runs.

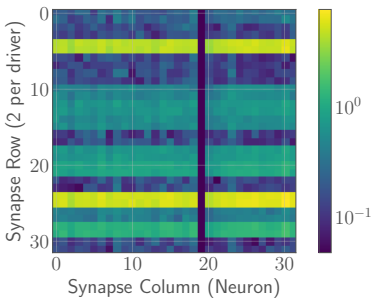


Figure 9: Standard deviations between spikes in a pattern like the one on the left. 30 synapse rows enabled on chip 03, $I_{\text{bias}} = 200$. Color scale given in [mV] (logarithmic scale). No data measurable for neuron 19.

the chip, with drivers on the left and neurons at the bottom (figure 9). This shows that it is a driver-related issue, we see similarities in lines. Testing another chip shows a different pattern, some chips contain much more "bad" drivers than others however. The problem seems to be controllable by changing STP mode and voltages V_{charge} and V_{recover}, the effect looks similar to the according STP behaviour. The observed deviations are the greatest at the first spike. Sending one additional spike before the experiment, even a long time before it and on another address, can reduce the effect drastically, whereas sending multiple spikes can even over-compensate it.

We have not fully understood the problem right now. There might even be influences of a small supply of V_{offset}. Output amplitudes can get twice as large when all drivers are enabled via the `enreceiver` setting, compared with only one driver active. Since the whole offset mechanism and ramp generation is solved differently in the upcoming tapeout (HICANN-X), the whole issue might change or be solved completely.

6 STP recovery

The recovery of neurotransmitters is on the chip implemented as an exponential decay of V_{STP} back to V_{recover}. Since V_{STP} is stored on a capacitor, the easiest way to recover would be charging this capacitor via a resistor. Regarding the very small capacities used on the chip, resistance would have to be very high, in the range of GΩ, which would require a lot of space. As described in [Bil17], recovery is instead accomplished by switching a small capacity between the STP capacitor and the V_{recover} potential. Changing the switching frequency allows for setting recovery time constants.

6.1 Leakage

The mentioned switches are transmission gates consisting of two parallel MOSFET transistors, one NMOS and one PMOS. In the closed state, there occurs leakage, depending on the present voltages. This is also true for transistors on the V_{charge} side, this potential may also affect leakage, especially since the transistors used are Low Voltage Threshold (LVT) transistors, having a low on-resistance but having higher leakage. The situation is shown in the schematic

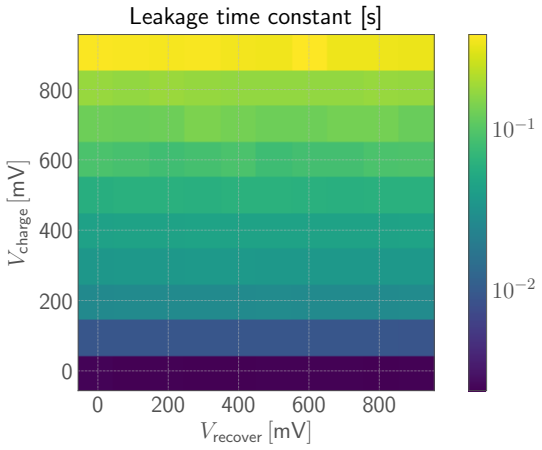


Figure 10: Time constants of voltage decay via leakage, depending on V_{charge} and V_{recover}. Measured on chip 3, driver 0, neuron 0, address 15. The dependency on V_{charge} dominates.

in figure 1.

To investigate leakage, recovery is turned off completely. Sending a few spikes and measuring V_{STP} on address 15, we can look at the "recovery" by leakage. We fit an exponential function and extract the baseline (leakage target), time constant (proportional to off-resistance) and the spike amplitudes, which give us an idea if the leakage potential is near V_{charge} or not.

Looking at time constants, we plot a picture sweeping over different V_{recover} and V_{charge} values. This is shown in figure 10. It is clearly visible that the leakage is dominated by V_{charge}, there is almost no dependency on V_{recover} visible. Higher settings for V_{charge} mean higher time constants, so less leakage current. However, setting V_{recover} to a low value should work fine. To minimize leakage effects, V_{charge} and V_{recover} can be swapped, which, changing the STP mode setting, still allows for the same STP behaviour.

As a result of this, for the next tapeout, the resistors used on the update circuit dealing with V_{charge} will be changed to be standard transistors instead of Low Voltage Threshold ones. This should decrease leakage currents by an order of magnitude.

6.2 Address crosstalk

To store individual STP states for each address, every single one needs an own STP capacitor. These 64 capacitors are arranged side-by-side in a repetitive pattern. Changing the charge for one of the addresses could possibly influence voltages on nearby other capacitors as well.

To test that, we measure V_{STP} on address 15, where this readout is possible. We observe the voltage, especially during recovery after sending some spikes, with and without spiking on addresses 12 to 14. Those should be nearby and form a block of 4 capacitors with address 15. In the plot, which is shown in figure 11, we can see no impact on the recovery of address 15 when other addresses process spikes (update circuit is active). The time of spikes on other addresses are shown as little bars below the trace, in groups of 3 there are addresses 12, 13 and 14 selected. Recovery is set to prescaler 5, `recovery` 0. The plotted trace is the mean of 100 measurements, not filtered however. While the shown plot was measured on chip 17 (DLS 3.1), this was similarly true on chip 3 (DLS 3.0). Firing does not have an observable influence on their neighbours' STP states.

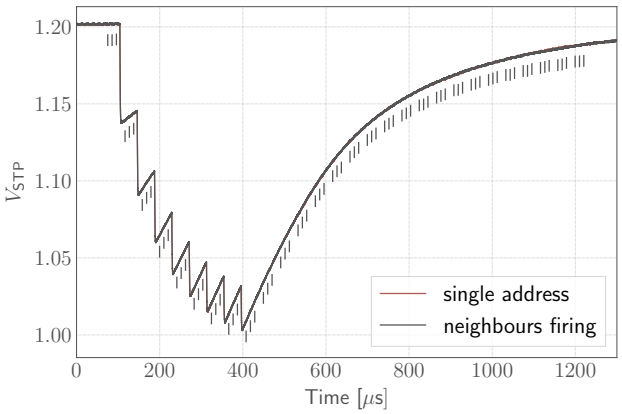


Figure 11: Voltage on STP capacitor 15, neuron 0, driver 0 on chip 17. Spiking only on address 15 in red, barely visible since overlayed by the trace with spiking on addresses 12 to 14 in between (black trace and spike indicators). No crosstalk visible.

7 Discussion and Outlook

Various experiments have been done to test whether the STP circuitry works as designed and to calibrate the offset parameter. As already stated above, most of the experiments yield the results we expect. The available configuration range for the U_{SE} parameter deviates only slightly from simulation results and covers a wide range of settings, making it well usable for emulating biologic models.

The synaptic output amplitudes when using STP can be shifted independent of the STP voltage by changing the offset parameter as shown in figure 5. An algorithm is now available that uses the analog output lines of the neuron to read out synaptic amplitudes. While the calibration results are great, this takes about 6 minutes of time for 16 drivers. On a larger system like the upcoming HICANN-X chip that contains more synapse drivers, the required time will scale linearly. Using the same algorithm, calibrating 512 drivers would take approximately 3 hours.

We are positive that a lot of time could be saved by changing the readout mechanism: Instead of using the baseboard ADC to measure amplitudes, the neurons themselves could be used. Spike rates can be used to represent the incoming synaptic signal amplitudes. Using this indirect measuring method, we avoid reading analog voltage traces, thus the calibration could be significantly faster. It would even be possible to use all neurons at the same time, eliminating the need for a synapse equalization before starting the actual synapse driver offset calibration. If the method works well, there even exists the possibility to control the calibration using the on-chip Plasticity Processing Unit (PPU). As a first step, assets and drawbacks of both methods will be looked at.

Concerning the observed variations in STP behaviour, some further testing will be done. To narrow down the problem, the comparator ramp will be investigated. If the deviations between drivers are due to changes in ramp steepness, the mapping of V_{STP} voltage to duration of `dacen` pulses could vary. The ramp generation could be equipped with a calibration possibility if large deviations are observed.

The observed higher spike amplitudes during the first spikes of a burst, like shown in figure 8, are not visible any more when a spike is sent to another address before. This might be a possible tactic for sensible experiments that heavily rely on precise STP amplitudes.

Investigating the STP recovery circuit, we found that leakage currents are high for low values of V_{charge}. As a conclusion, the used transistor type will be changed from Low Voltage Threshold transistors to standard transistors. Looking for possible crosstalk effects, working with multiple addresses at the same time showed no problems. To test the STP recovery circuit further, the actual recovery times and targets will be measured systematically for multiple addresses. This way we will know the available configuration range for recovery and will be able to notice possible problems.

For now, we conclude that the STP implementation on HICANN-DLS 3 works quite well and is usable as intended!

References

[Bil14] Sebastian Billaudelle. "Characterisation and Calibration of Short Term Plasticity on a Neuromorphic Hardware Chip". Bachelor Thesis. University of Heidelberg, 2014.

[Bil17] Sebastian Billaudelle. "Design and Implementation of a Short Term Plasticity Circuit for a 65 nm Neuromorphic Hardware System". Master Thesis. University of Heidelberg, 2017.

[Sch+07] Johannes Schemmel et al. "Modeling synaptic plasticity within networks of highly accelerated I&F neurons". In: *Proceedings of the 2007 International Symposium on Circuits and Systems (ISCAS)* (2007), pp. 3367–3370.

[ZR02] Robert S. Zucker and Wade G. Regehr. "Short-term synaptic plasticity". In: *Annual Review of Physiology* 64 (2002), pp. 355–405.