Faculty of Physics and Astronomy

University of Heidelberg

Bachelor Thesis

in Physics,

submitted by

Sebastian Billaudelle

born in Heidelberg, Germany

September 2014

Characterisation and Calibration of Short Term Plasticity on a Neuromorphic Hardware Chip

This bachelor thesis has been carried out by

Sebastian Billaudelle

at the

Kirchhoff-Institute for Physics,

University of Heidelberg

under the supervision of

Prof. Dr. Karlheinz Meier

Abstract

The Short Term Plasticity (STP) implementation of the neuromorphic High Input Count Analog Neural Network (HICANN) chip is characterised. In this thesis, a high-level approach is adopted by analyzing a neuron circuit's response to a presynaptic stimulus. By recording analog traces of the membrane voltage, the STP parameters are extracted.

Starting from the theoretical model, protocols for characterising its parameters were developed. Corresponding measurements are discussed in this thesis. The experiments have shown that the hardware's characteristics lie within ranges suitable for biology-inspired network models. Deviations between multiple instances of the circuit could be observed. In order to compensate for these deviations, calibration methods were developed and successfully verified. As a conclusion, the measured operating range is compared to data from biological systems and possible limitations of the hardware design are discussed.

Zusammenfassung

Im Rahmen dieser Arbeit wird die Implementierung von Kurzzeitplastizität auf dem neuromorphen HICANN-Mikrochip untersucht. Durch Beobachtung und Analyse postsynaptischer Potentiale auf der Membranspannung als Reaktion auf einen presynaptischen Stimulus, können die Parameter des Platizitätsmechanismus bestimmt werden.

Ausgehend von einem theoretischen Modell werden Messprotokolle für die Charakterisierung des Schaltkreises entwickelt. In dieser Arbeit werden die Ergebnisse der zugehörigen Messreihen diskutiert. Es kann gezeigt werden, dass das Verhalten des Chips die Realisierung biologisch inspirierter Netzwerke unterstützt. Allerdings fallen starke Schwankungen zwischen verschiedenen Instanzen der Schaltung auf. Deshalb werden Kalibrationsmethoden zum Ausgleich dieser Abweichungen entwickelt und erfolgreich getestet.

Abschließend werden die Eigenschaften der Kurzzeitplastizität des HICANN-Chips mit aus der Biologie stammenden Daten verglichen und die möglichen Grenzen der Schaltung dargelegt.

Contents

1 Introduction

The human brain is estimated to consist of over 20 billion neocortical neurons each with approximately 7000 synaptic connections on average (Drachman, 2005). In order to understand the communication in this network, large-scale numerical simulations of neural models are conducted. Approaches like these require huge computational resources provided by supercomputers (EPFL and IBM, 2008). Within the BrainScaleS Project (BSS) (BrainScaleS, 2012), a very different approach is taken: Analog neuromorphic hardware implements physical models with similar dynamics in silicon. Systems following this design promise a low energy footprint as well as a high execution speed mostly independent of the model's size. In the context of the BSS, the mixed-signal Hybrid Multi-Scale Facility (HMF) is developed, primarily at the Kirchhoff-Institute for Physics in Heidelberg and the TU Dresden. For the HMF, a high-performance computing cluster is combined with highspeed communication links based on Field Programmable Gate Arrays (FPGAs) and a neuromorphic core. The latter consists of wafer-scale integrated High Input Count Analog Neural Network (HICANN) chips (HBP SP9 partners, 2014).

This chip does not only provide configurable analog neuron circuits and software-defined routing capabilities, but also features implementations of Spike-Timing Dependent Plasticity (STDP) (Bi and Poo, 1998, Schemmel et al., 2006) and Short Term Plasticity (STP) (Tsodyks and Markram, 1997, Schemmel et al., 2006). While the former is believed to at least partially explain memory storage and Hebbian learning (van Rossum et al., 2000), the latter regulates a synapse's response based on presynaptic activity (Zucker and Regehr, 2002). This thesis will entirely focus on the characterization of the STP mechanism on the HICANN microchip.

An implementation of STP has been available for the predecessor chip *Spikey*. On that platform, other publications have already examined STP (Bill, 2008). It has shown to allow for a compensation of inhomogeneities in neuromorphic hardware systems due to variations in the production process (Bill et al., 2010). In particular, STP enables the implementation of self-stabilizing neural network models (Bill, 2008). In order to lay the groundwork for similar networks on the current hardware generation, it is required to characterise its properties.

Each HICANN contains 224 individual instances of the STP circuitry. While it is possible to write down a model of the implementation's dynamics, hardware measurements show deviations from these predictions. For a reliable STP mechanism, these variations must be characterised and compensated by calibration routines.

At the beginning of this thesis, the STP circuit's dynamics are examined and compared to the biology-inspired Tsodyks-Markram model (Tsodyks and Markram, 1997). Then, measurement protocols for the characterization of the parameters are presented. The results of these measurements are shown for one HICANN chip exemplarily and possible calibration strategies are discussed. As a conclusion, similarities between the hardware implementation and the Tsodyks-Makram model are pointed out, as well as potential limitations of the design.

2 Materials and Background

The HMF is a mixed-signal neuromorphic hardware developed primarily at the Kirchhoff-Institute in Heidelberg and the TU Dresden in the context of the BSS funded by the European Union (BrainScaleS, 2012). Core of the HMF wafer-scale neuromorphic hardware system is the HICANN chip.

2.1 The HICANN Chip

The following paragraphs are meant to serve as a short introduction to the hardware components and most importantly to the terminology used in this thesis. For more detailed information refer to HBP SP9 partners (2014) and Schemmel et al. (2010).

Due to shorter intrinsic time constants, an accelerated execution of neural networks is inherent in the design of the HICANN microchip. Compared to biological real time, a speedup of $\approx 10^4$ can be observed. In this thesis, all time scales are given in the Hardware Time Domain (HTD) as opposed to the Biological Time Domain (BTD). Exceptions to this are clearly marked.

2.1.1 Analog Neurons

A single HICANN chip contains 512 analog *neuron* circuits – also referred to as *dendritic membranes* – implementing the Adaptive Exponential Integrate-and-Fire model (AdEx) model (Brette and Gerstner, 2005). Each neuron can receive input via two synaptic input circuits – usually configured for excitatory and inhibitory stimulation, respectively. Multiple dendritic membranes can be connected together to form a larger neuron in order to increase the number of inputs. This feature was not used for this thesis.

When the voltage on a neuron's membrane reaches a configurable threshold, a digital spike event is emitted and the membrane voltage is pulled to a configurable reset voltage.

2.1.2 Digital Event Network

Events carry a digital 6 bit address which can be assigned freely to each neuron. The same applies to external stimulus that is injected via the FPGA's playback memory.

Spike events are transmitted over the asynchronous, serial *Layer 1 bus* (Schemmel et al., 2008). By configuring static switch matrices, events can be passed to different components of the chip or forwarded to neighbouring HICANNs. *Repeaters* located at the chips' edges restore signals in order to increase the bus system's transmission reliability, especially across longer distances.

2.1.3 Synapse Drivers

Furthermore, each HICANN contains 224 *synapse drivers*, which represent a special type of repeaters. They serve as an interface between the digital Layer 1 bus system and the analog synapses. Figure 1 shows a schematic view of a synapse driver and the synapse array.

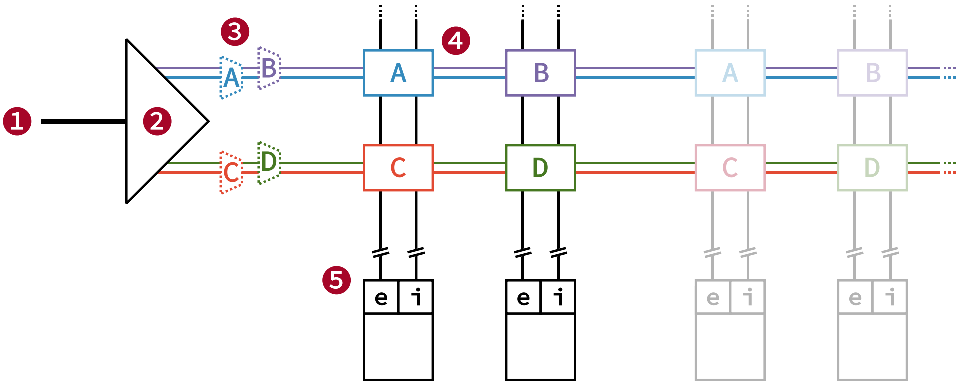

Figure 1: A schematic view of a synapse driver and the synapse array. An event arrives on a driver bus line (1) and enters the synapse driver (2). Its two MSBs are then matched against a preconfigured mask by the event decoders (3). The synapse driver translates the event into an analog pulse which can be forwarded to the synapse circuits (4) via two synapse lines (top bottom). Within the synapses, the 4 LSBs are again matched to a configurable mask. Valid events are then fed into one of the two synaptic inputs of the neuron circuit (5).

Layer 1 events enter the synapse driver via *driver lines* where an *address decoder* compares the two Most Significant Bits (MSBs) of its address against a preconfigured mask. Matching events are then translated into current pulses which are forwarded to the synapse circuits. A pulse's height depends on a configurable base weight, while its width is set by the STP circuit which is part of the synapse driver.

Each synapse driver has four individually configurable address decoders. Two decoders each account for the top and bottom *synapse line*, respectively. While one decoder per line forwards signals to neurons with even x-coordinates, the other is connected to neurons with odd ones. This topology results in the *A, B, C, D* pattern shown in figure 1.

2.1.4 Synapses

One synapse line contains 256 synapses allowing to forward events to the connected neuron. Here, the remaining 4 Least Significant Bits (LSBs) of the address are compared to another mask. The current pulse received from the synapse driver is then scaled proportionally to a 4 bit weight. This value is stored locally together with the address mask.

The routing topology presented above allows the processing of spike events either originating from spiking neurons or external input. By configuring static switch matrices and address masks, events with specific source addresses can be forwarded to one or more neurons. By configuring a synapse driver's address decoders correctly, all 64 addresses can be processed by a single driver. With address 0 being assigned to the background event generators ensuring the correct locking of the synapse drivers to the Level 1 bus signal (Schemmel et al., 2008), 63 addresses are available for experiments.

2.2 Short Term Plasticity

STP or Short Term Depression and Facilitation (STDF) represents a concept describing the change in synaptic efficacy depending on the recent history of presynaptic activity. In biology, the typical timescale of STP ranges from hundreds of milliseconds to seconds (Regehr, 2012).

Both, short term *depression* and *facilitation* can be explained by the dynamics of neurotransmitters. On one side, dense stimulus can lead to a depletion of neurotransmitters at the presynaptic axon terminals, in turn weakening the synaptic efficacy and leading to a depression of Postsynaptic Potentials (PSPs). On the other hand, the release probability of neurotransmitters is increased by calcium influx caused by presynaptic action potentials. This can be observed as facilitation of PSPs. In biology, both effects can be observed individually or – more commonly – in combination (Hennig, 2013).

The *Tsodyks-Markram model* represents a phenomenological model describing STP (Tsodyks and Markram, 1997). In this model, neurotransmitters at the synaptic cleft are divided into the recovered partition R, effective partition E and inactive partition I. The filling levels of these partitions range from 0 to 1. Their dynamics are given by

$$\frac{dR}{dt} = \frac{I}{\tau_{\text{rec}}} - U_{\text{SE}} \cdot R \cdot \delta(t - t_{\text{AP}})\,, \tag{1}$$

$$\frac{dE}{dt} = -\frac{E}{\tau_{\text{inact}}} + U_{\text{SE}} \cdot R \cdot \delta(t - t_{\text{AP}})\,, \tag{2}$$

$$I = 1 - R - E\,. \tag{3}$$

In these differential equations, t_{AP} denotes the time of a presynaptic action potential. The *utilization of synaptic efficacy* U_{SE} represents the fraction of available neurotransmitters released per incoming spike. τ_{rec} and τ_{inact} specify the time constants for the recovery and inactivation process, respectively. In figure 2, a simulation of short term depression following this model is shown.

On the occurrence of a presynaptic spike, the fraction U_{SE} of the neurotransmitters stored in the recovered partition R is transferred into the effective partition E. The level of E decays with a time constant of τ_{inact}, while R is recovered with τ_{rec}. The synapse's weight defining the strength of a synaptic connection is given by

$$w \propto E\,. \tag{4}$$

Other models have been developed covering specific aspects of STP or cell-specific behaviour (Varela et al., 1997, Pan and Zucker, 2009). However, the Tsodyks-Markram model presented above is supported by different neural simulators (Goodman and Brette, 2009, Gewaltig and Diesmann, 2007) and is the only available STP model in the common interface PyNN (Davison et al., 2008).

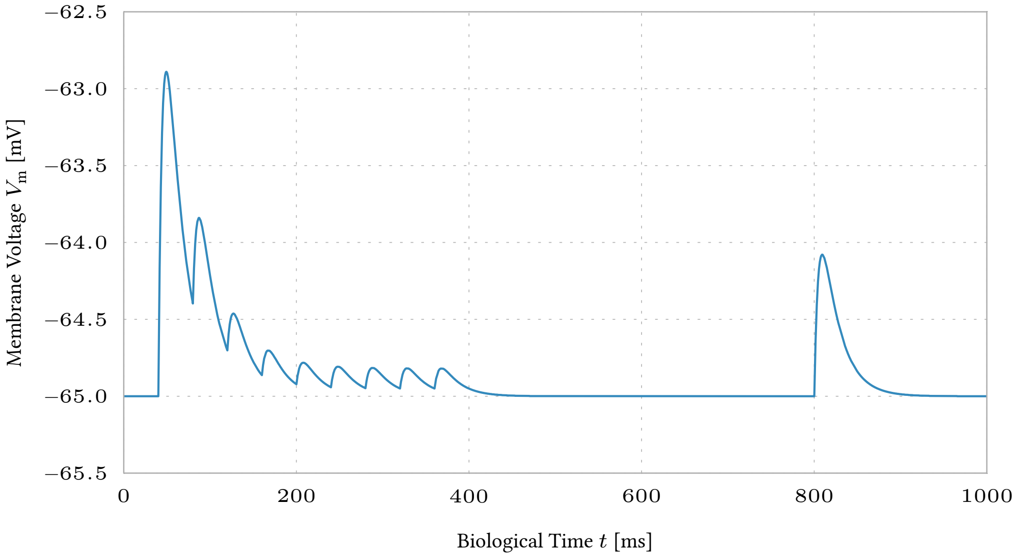

Figure 2: Simulation of short term depression following the Tsodyks-Markram model (Tsodyks and Markram, 1997). The depression of synaptic efficacy can be seen in the decrease of the PSPs' amplitudes. After a period with no presynaptic activity, the PSPs' heights get restored. A stimulus of 25 Hz (BTD) was used and short term depression was set to $U_{\mathrm{SE}} = 0.67$, $\tau_{\mathrm{rec}} = 800$ ms (BTD) and $\tau_{\mathrm{inact}} = 0$. The simulation was implemented using PyNN with the NEST a backend (Davison et al., 2008, Diesmann and Gewaltig, 2002).

2.3 Hardware Model

An STP implementation was developed for the analog neural network chip *Spikey* (Schemmel et al., 2006). The implementation for the chip's successor HICANN mostly follows the same design. Despite being inspired by the Tsodyks-Makram model, depression and facilitation can not be applied simultaneously in hardware. This results in only two partitions I and R being necessary and leads to the simpler model (Schemmel et al., 2006)

$$\frac{dI}{dt} = -\frac{I}{\tau_{\mathrm{rec}}} + U_{\mathrm{SE}} \cdot R \cdot \delta(t - t_{\mathrm{AP}})\,, \tag{5}$$

$$R = 1 - I\,. \tag{6}$$

However, the actual implementation does not exactly follow equation 5. An exponential recovery process would require discharging a capacitance over an ohmic resistor. Furthermore, this resistor would have to be adjustable in order to allow configurable time constants. In integrated CMOS designs, resistors can be realised using *poly elements*. Such implementations take up large areas and, per se, are not adjustable (Aamir, 2014). Thus, the recovery term had been implemented using a configurable but static current I_{rec}. This results in a linear recovery process with slope M. Equation 5 is adjusted accordingly:

$$\frac{dI}{dt} = -M + U_{\text{SE}} \cdot R \cdot \delta(t - t_{\text{AP}})\,. \tag{7}$$

For short term *depression*, the inactive partition reduces a synapse's effective weight. It is given by

$$w_{\text{dep}} \propto 1 - \lambda \cdot (I - N) \tag{8}$$

with λ scaling the impact of short term depression on the synaptic weight and an offset parameter N. Since λ has no equivalent in the biological model presented in 2.2, it should be calibrated to $\lambda = 1$. In the absence of a recovery current, this would result in a full depression for continuous spike input. Higher values of λ lead to *premature depression*. The offset offers control over the absolute amplitudes of the PSPs. In case of a functional neuron calibration incorporating the calibration of PSP amplitudes, the offset parameter would be set to $N = 0$. This fixates the first PSP's height to the same amplitude as for disabled STP. It has to be noted, that for Spikey, the offset parameter was not configurable in depression mode. The model was extended accordingly.

In *facilitation* mode the inactive partition increases the effective weight. Now, the weight can be described by

$$w_{\text{fac}} \propto 1 + \lambda \cdot (I - N)\ . \tag{9}$$

As for depression, λ and N allow to control the amplitudes of the first PSP as well as the steady state amplitude for successive input spikes. These parameters have no equivalents in the Tsodyks-Makram model where they are implicitly set to $\lambda = 1$ and $N = 1$. Later in this thesis, a calibration of the hardware to these values will be discussed.

2.4 Hardware Implementation

The model presented above is implemented in an analog circuit located in each synapse driver. A reduced schematic is shown in figure 3. A more detailed description of the circuit can be found in Schemmel et al. (2006).

The design is centred around two MOS capacities (Nicollian et al., 1982). C_1 represents the inactive partition I. As a Layer 1 event enters the synapse driver, capacitor C_2 is charged to the voltage difference $V_{\text{dda}} - V_{\text{stdf}}$ by temporarily closing switch S_2. Then, transiently connecting both capacities via S_1 initiates a charge sharing process between C_2 and C_1. Thus, a single event increases the voltage across C_1 by

$$\Delta V = \left(V_{\text{stdf}} - V_{\text{cap}}\right) \cdot \underbrace{\frac{C_2}{C_1 + C_2}}_{=U_{\text{SE}}}\ . \tag{10}$$

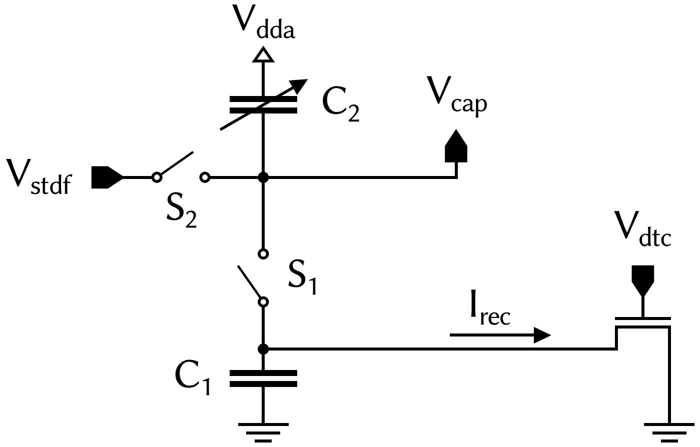

Figure 3: Reduced schematic of the STP circuit. Presynaptic stimulus leads to a stepwise increase of the voltage on C_1, which represents the inactive partition I. C_2 is implemented as a 3 bit configurable capacitance allowing to manipulate the utilisation of synaptic efficacy U_{SE}. The recovery process is implemented with a constant current I_{rec} controlled by V_{dtc}. For detailed description of the circuit refer to Schemmel et al. (2006).

In order to allow configuration of the utilisation of synaptic efficacy U_{SE}, C_2 has been designed as a configurable capacitance. Three bits, accessible through the digital setting cap, are used for adjusting its value.

In a second stage, the pulse emitted by the synapse driver is modulated. For this purpose, V_{cap} is applied to a comparator circuit with a reference voltage of V_{dep} or V_{fac}, depending on the STP mode. The voltage difference $V_{\mathrm{cap}} - V_{\mathrm{dep/fac}}$ is then used to scale the width of the pulse. With the reference voltage, an offset can be configured.

The recovery current is controlled by a floating gate cell. This current is forwarded to the STP circuit by two current mirror stages. The gate-source voltage is called V_{dtc}, hence the name of the corresponding floating gate parameter. Since low recovery currents are required, the current mirrors are biased in the deep subthreshold region (Schemmel, 2014). The level of the inactive partition needs to be stored and processed individually per input address. The corresponding capacity C_1 is therefore multiplexed 64 times within each synapse driver. While the first current mirror is used to transfer the current to the individual synapse driver instances, the second stage mirrors I_{rec} to these multiplexed capacities. Therefore, the end point of the second mirror is replicated with each instance of C_1. This has to be considered during the characterization of the recovery process, since deviations across different source addresses are expected.

2.5 Hardware Parameters

The STP circuit is configured through multiple digital parameters as well as analog floating gate voltages and currents (Lande et al., 1996).

As part of the synapse driver SRAM, the digital settings shown in table 1 consist of five bits. One is used for enabling STP, a second one for setting the mode (depression or facilitation). Three bits are reserved for choosing capacity C_2. These parameters are uniquely adjustable for each synapse driver instance.

Furthermore, the circuit can be configured through five floating gate parameters, which are presented in table 5. Four voltage cells are available to control the biasing and reference voltages, while a current cell is used to set the recovery current. Synapse drivers located at the top left (1, 3, ..., 111), top right (0, 2, ..., 110), bottom left (112, 114, ..., 222) and bottom right (113, 115, ..., 223) quarter of the chip share the same analog parameters, respectively. These parameters are located in the shared floating gate cells, allowing only four different configurations per HICANN chip at the same time. This results in high demands for calibration algorithms, since only the mean value of a parameter can be calibrated for 56 synapse drivers each. Within these groups, deviations from this average can not be compensated for.

Voltage cells cover a range of $0\,\mathrm{V}$ to $1.8\,\mathrm{V}$, current cells can be configured for up to $2.5\,\mu\mathrm{A}$. These parameters can be set with a precision of $10\,\mathrm{bit}$. In this thesis, voltages as well as currents are given in DAC values representing the $10\,\mathrm{bit}$ range (0 to 1023) of the floating gate cells. This choice has been made since it represents the impact of a configuration parameter on an observable more directly.

Name	Description
cap	Size of capacitor C_2. With $cap = 0 \dots 7$, the actual capacity is given by $C_2 = {}^{cap}/_{30} \cdot C_1 \approx cap \cdot 8\,\mathrm{fF}$
dep	STP mode (0: facilitation, 1: depression).
enstdf	Enable STP for the synapse driver.

Table 1: Digital STP parameters as an excerpt of the synapse driver SRAM bits. These settings are specific to each synapse driver. For a complete documentation of the synapse driver configuration refer to HBP SP9 partners (2014).

Parameter	Description
V_{stdf}	Maximum voltage, the storage capacitors C_1 are being charged to. V_{stdf} scales the impact of STP on the effective weight.
V_{bstdf}	Bias voltage for the comparator circuit.
V_{dep}	Offset voltage for depression mode.
V_{fac}	Offset voltage for facilitation mode.
V_{dtc}	Voltage for setting the recovery current. In fact, V_{dtc} is not set directly but corresponds to the gate-source voltage of a current mirror connected to a current cell within the floating gate block.

Table 2: Shared floating gate parameters for the short term plasticity implementation of the synapse drivers. Since they are located in the shared floating gate blocks, synapse drivers can not be configured independently.

3 Methods

3.1 Measurement Setup

Investigating the properties of a synapse driver's STP circuit requires the stimulation of an arbitrary neuron via that specific synapse driver. The stimulus consists of predefined spike trains activating the STP mechanism. A series of input events leads to multiple Excitatory Postsynaptic Potentials (EPSPs) on the neuron's membrane. Recording its voltage allows the analysis of the response to the stimulus and thus a characterisation of the STP parameters. Per HICANN, two Analog Digital Converters (ADCs) can be used to record analog traces.

3.1.1 Triggered Recording of Membrane Voltage Traces

The ADC's trigger mechanism allows synchronisation of the analog readout traces with the FPGA's playback memory. This enables correlating Layer 1 input spikes with EPSPs on the recorded membrane voltage. The trace length is limited to approximately 4 ms.

3.1.2 Averaging of Voltage Traces

Membrane voltage traces are subject to both voltage fluctuations on the membrane and readout noise. Averaging over multiple traces is important not only to fulfil statistical requirements but also to obtain smooth traces that allow further analysis. The speedup factor of 10^4 results in a timescale of 100 µs for most of the STP measurements on the HICANN chip. This allows repeating the input spike pattern until reaching the 4 ms limit for triggered recordings in order to retrieve an averaged trace with a single recording procedure. In practice it has shown to be possible to repeat a pattern with a duration of 120 µs up to 50 times in a single recording. This allows to reduce measurement overhead.

3.1.3 Improving Trace Quality

Besides averaging, other steps can be adopted to further improve the recorded traces' quality. Primarily, the membrane time constant τ_m can be minimised. This leads to a better separation of consecutive PSPs.

Furthermore, STP measurements do not require all features of the hardware neurons. Minimizing the impact of the exponential and adaptive terms of the neuron by tuning floating gate parameters has shown to improve trace quality and suppress unwanted behaviour like overshooting membrane voltages. Additionally, applying the neuron calibration currently being developed increases reproducibility of the results.

3.1.4 Extraction of PSP Heights

The membrane voltage is composed of membrane effects on the timescale of τ_m on the one hand and multiple PSPs as a response to the input stimulus on the other hand. The height of a PSP is proportional to the effective weight of the specific synapse. Extracting

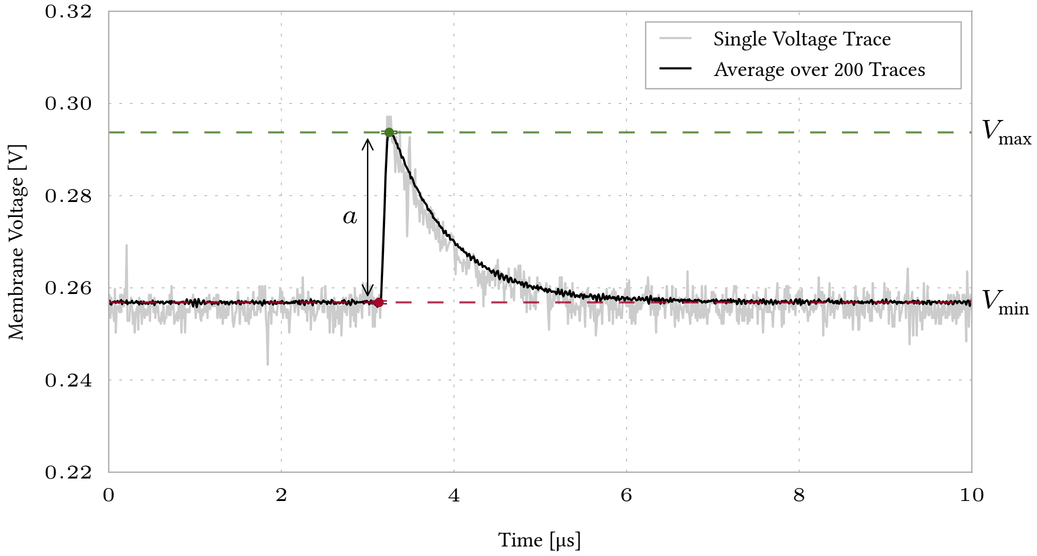

Figure 4: Extraction of a PSP's amplitude from an averaged voltage trace following the procedure explained in 3.1.4. The average results from 200 single traces, one of which is shown for illustration. V_{min}, V_{max} and the resulting amplitude a are marked within the plot.

this information out of membrane voltage traces is important for conducting STP measurements. During the work for this thesis, two approaches to extract the PSPs' heights from an averaged trace were considered.

Fitting one or multiple alpha functions to the PSPs' expected positions seems to be the most comprehensive solution. Unfortunately, the superposition of multiple PSPs and membrane effects leads to unstable fit results.

Defining the height as the difference of the PSPs' maximum and minimum has shown to yield reliable results with sufficient precision. This method is shown in figure 4. It is

$$V_{\text{max}} = max\, V_{\text{m}}(t) \quad \text{with } T_n < t < T_{n+1}\,, \tag{11}$$

$$V_{\text{min}} = min\, V_{\text{m}}(t) \quad \text{with } t \in S_\epsilon(T_n) \tag{12}$$

with T_n and T_{n+1} being the rise time of the specific PSP and its successor, respectively, and S_ϵ representing a neighbourhood with a small radius ϵ of approximately 5 to 10 ADC samples. The amplitude a of the PSP results as

$$a = V_{\text{max}} - V_{\text{min}} \qquad \Delta a = \sqrt{(\Delta V_{\text{max}})^2 + (\Delta V_{\text{min}})^2} \tag{13}$$

with the $\Delta V_{\text{min/max}}$ resulting from the error of the mean of the averaged trace.

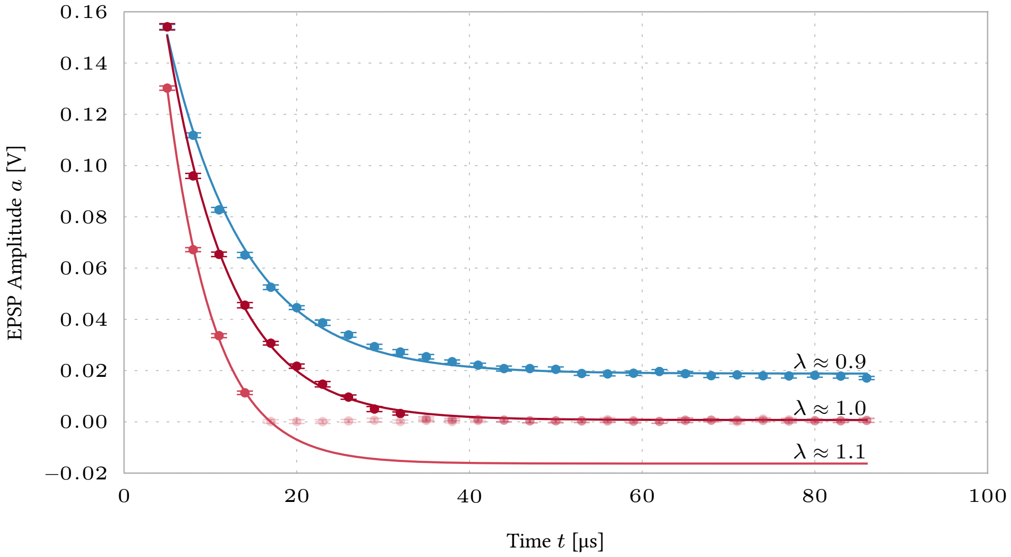

Figure 5: Protocol for measuring parameters λ and U_{SE} in depression mode. Stimulating a neuron with an equidistant spike train – here, with $0.3\,\mathrm{MHz}$ – results in the time course shown above. The parameters can be extracted by fitting an exponential decay to the measured amplitudes. Here, three fits are shown for $\lambda < 1$, $\lambda = 1$ and $\lambda > 1$, respectively. As visible for the latter two cases, samples that are not significantly deviating from zero need to be excluded from the fitting range in order to represent the correct exponential time course. Note that floating gate variations explain the difference in the amplitudes' absolute scaling, especially observable for the first EPSP.

3.2 Characterisation of the STP Parameters

The parameters of the model presented in 2.3 can be measured by using the extracted PSP amplitudes. The following measurements are grouped into depression, facilitation and the recovery phase.

3.2.1 Depression

In order to characterise the depression, parameters λ, N and U_{SE} are measured. The protocol presented in the following has shown to yield stable and reasonable results. Exemplarily, figure 5 presents some of the measurements.

The impact of the recovery process needs to be extinguished for these measurements. Therefore, the recovery current I_{rec} is minimised by setting $V_{\mathrm{dtc}} = 0$. Now, stimulating a neuron with an equidistant spike train with period Δt will result in a series of EPSPs. From equation 7 it follows with $M = 0$ that

$$R_0 = 1\,, \tag{14}$$

$$R_{i+1} = R_i + \Delta R_i = R_i - \Delta I_i \tag{15}$$

$$= R_i \cdot (1 - U_{\mathrm{SE}}) \tag{16}$$

$$\Rightarrow R_i = (1 - U_{\mathrm{SE}})^i \ . \tag{17}$$

Keeping equations 6 and 8 in mind and assuming that an EPSP's amplitude is proportional to the synaptic weight, the time course of the EPSPs can be expressed as

$$a_i = \hat{a} \cdot [1 - \lambda \cdot (I_i - N)] \tag{18}$$

$$= \hat{a} \cdot [1 - \lambda \cdot (1 - N) + \lambda \cdot R_i] \tag{19}$$

$$= \hat{a} \cdot \left[1 - \lambda \cdot (1 - N) + \lambda \cdot (1 - U_{\mathrm{SE}})^i\right] \tag{20}$$

with $\hat{a}$ being the reference amplitude. The latter is measured directly before running the depression protocol without rewriting the floating gate voltages. This ensures that the same configuration is used for both parts of the measurement. Index i is given by $i = {}^{(t_i - t_0)}/_{\Delta t}$ with t_0 being the time of the first EPSP.

The measured amplitudes are used to fit the exponential in equation 20. For $\lambda < 1$ all data points can be included within the fit. However, the EPSPs' amplitudes can not completely represent the time course for $\lambda \geq 1$, since negative values are not possible. Therefore it is necessary to omit amplitudes that do not deviate from zero significantly ($a \not> 3 \cdot \Delta a$). This of course will reduce the fit's quality, since only few or no data samples are available in the steady state region. Thus, larger confidence intervals have to be expected for this case.

This method was verified for a simulation of short term depression. Please refer to appendix A.

3.2.2 Facilitation

The facilitation process's parameters can be measured with a protocol similar to the one presented for short term depression. With equations 9 and 17 it follows that

$$a_i = \hat{a} \cdot [1 + \lambda \cdot (I_i - N)] \tag{21}$$

$$= \hat{a} \cdot \left[1 + \lambda \cdot (1 - N) - \lambda \cdot (1 - U_{\mathrm{SE}})^i\right] \ . \tag{22}$$

with the offset N, scaling parameter λ and the reference amplitude $\hat{a}$. This again is fitted to the extracted amplitudes. In figure 6, this protocol is shown.

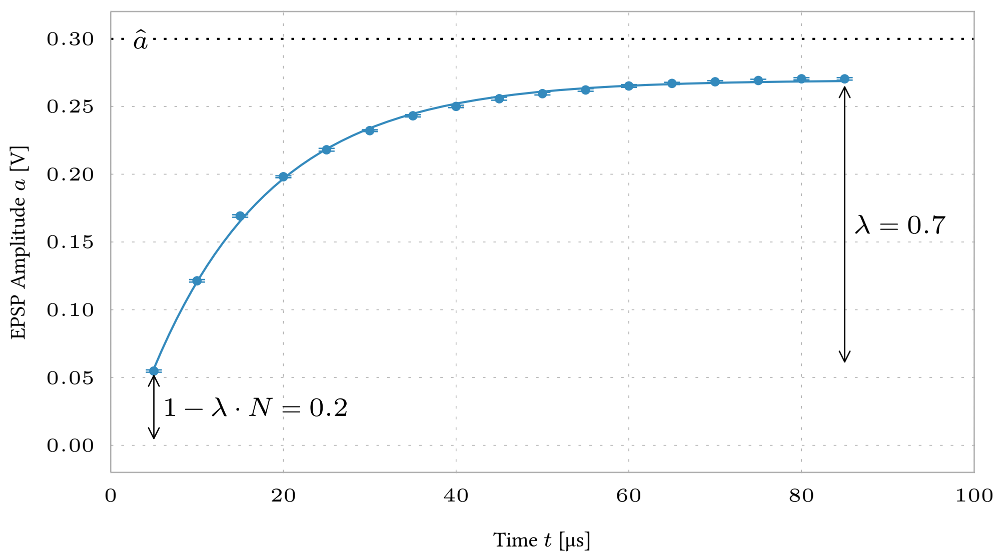

Figure 6: The measurement protocol for facilitation mode is shown for an exemplary measurement. As for short term depression, a neuron is stimulated with an equidistant spike train. Here, 0.2 MHz have been used. As the synapse changes its synaptic efficacy, the PSP amplitudes are facilitated. The data can been fitted to equation 22 in order to extract the model's parameters. In contrast to λ which defines the scaling of facilitation relative to the reference amplitude $\hat{a}$, N is not a concrete quantity. Instead, $1 - \lambda \cdot N$ representing the height of the first PSP is annotated in this plot.

3.2.3 Recovery

In order to characterise the recovery phase, the neuron is stimulated with a series of equidistant input spikes, which saturate the inactive partition I and thus drive the amplitudes into a steady state. This burst is followed by another *probe spike* after a time difference δt_i. By recording the trace for different δt_i and combining these measurements as shown in figure 7, the recovery process can be observed. For the individual samples, only the input stimulus is changed without reconfiguring the rest of the chip. Therefore, all probe EPSPs are recorded for exactly the same floating gate voltages.

As explained in 2.3 and shown in equation 7, the recovery process follows a linear rise. The recovery is complete, when the amplitude of the first EPSP is reached. Therefore, the composed function

$$f(t) = \begin{cases} m \cdot (t - t_0) + a, & \text{if } (t - t_0) \cdot m + a \leq b \\ b, & \text{otherwise} \end{cases} \tag{23}$$

can be used to perform a fit to the measured amplitudes of the probe spikes. Here, a represents the height of the first probe spike and b the fully recovered amplitudes. In figure

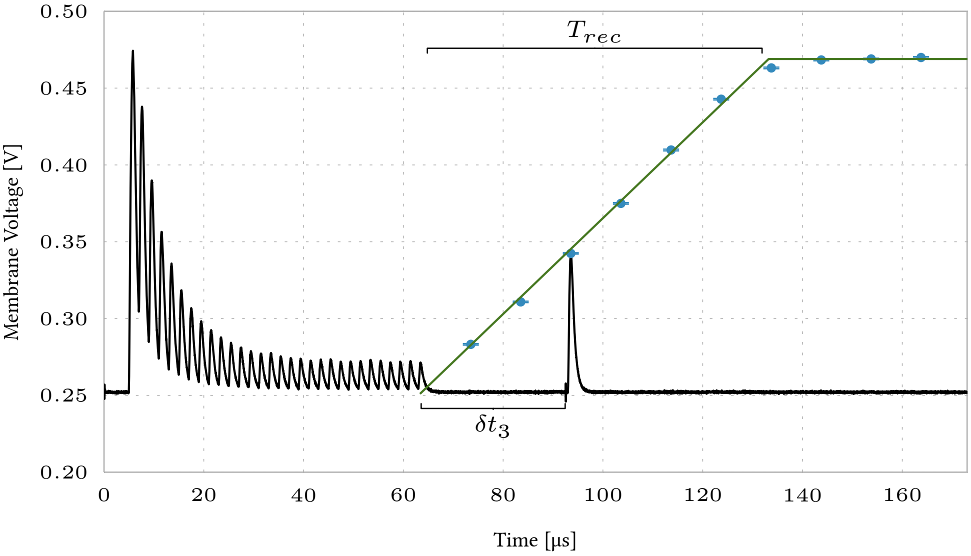

Figure 7: Protocol for measuring the recovery process' slope. The depression is initiated by a dense burst of input spikes. Then, probe spikes are injected after additional time periods δt_i. Here, the third probe spike is shown. By extracting the course of probe spike amplitudes, the linear recovery can be analysed. For illustration reasons, the EPSP amplitudes are shifted by the resting potential of the membrane. It has also to be noted that the fit deviates from the measured data at the beginning and end of the recovery, where the linearity does not apply.

7, the extracted amplitudes are overlayed on top of the measured traces. Also included is an exemplary fit.

The measured slope m is given in units of $^{\mathrm{V}}/_{\mathrm{s}}$ and is proportional to the absolute EPSP amplitudes. It turns out that due to slightly different configuration and calibration, every neuron shows an individual response and thus m is not transferable to other neurons. However, STP is a neuron-agnostic property and thus m does not represent a meaningful quantity. Instead, normalizing the slope to the inactive partition's maximum value of 1 is expected to be more concrete. With a previous calibration of $\lambda = 1$ and $N = 0$ (see 3.2.1), the inactive partition is directly represented in the EPSPs' amplitudes. A height of zero indicates a fully charged inactive partition, while the recovered state of $I = 0$ is represented in the fully recovered EPSPs. Now, the normalised slope can be calculated as

$$M = \frac{m}{b} \equiv \frac{1}{T_{\text{rec}}} \,. \tag{24}$$

T_{rec} represents the time to complete recovery from total depression. It can be used as an equivalent to the time constant τ_{rec} used in the Tsodyks-Markram model (see equation 1).

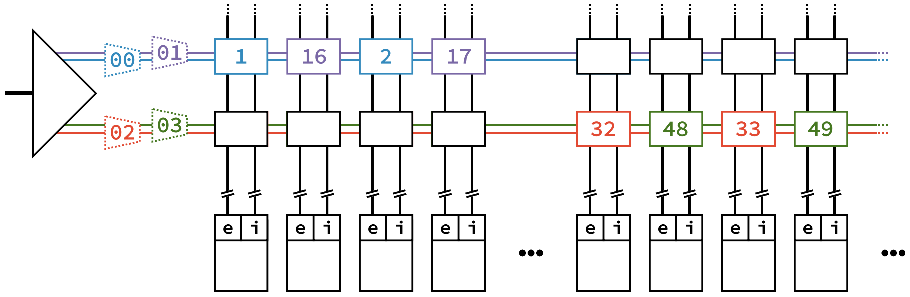

Figure 8: Schematic drawing of the routing scheme for the synapse driver defect detection tool. By assigning four different MSB masks to the address decoders of a synapse driver, all of the 64 L1 source addresses can be processed and forwarded. This enables a fast address-wise defect detection.

3.3 Functional Testing of Synapse Drivers

In this thesis, sweeps over all 224 synapse drivers of a HICANN chip were carried out. This did not only allow to perform the actual STP measurements but also to gather information about malfunctioning synapse drivers. With some modifications to the measurement software, a dedicated *synapse driver defect detection tool* was implemented. In contrast to the STP protocols, a spike based approach was chosen instead of analysing membrane traces recorded with ADCs. This enables a parallel evaluation of all source addresses per synapse driver, while an ADC based approach is limited by the number of available readout boards, typically to one neuron per experiment.

For this spike based method, the neurons are required to be configured for a binary spiking behaviour. A preliminary neuron calibration was used for this purpose. Furthermore, a naive blacklisting algorithm was applied to exclude inadequately behaving, e.g. constantly firing, neurons. Furthermore, STP was disabled in the synapse drivers.

A dedicated routing scheme was developed as presented in figure 8. By assigning disjunct MSB masks to the four address decoders on the “half synapse lines”, the whole range of 64 addresses can be forwarded to appropriately configured synapses. Thus, every Level 1 address can be assigned to a neuron which in turn is configured to emit spikes on the same address. This step allows to easily correlate recorded output spikes to the input addresses processed of the synapse driver.

Stimulating a neuron with a burst of dense input spikes ideally results in a high output activity during the same time window. Successively injecting these bursts for all source addresses creates a chain-like, diagonal pattern of recorded spikes. Analysing the latter allows to detect defects with a single-address resolution by scanning for either missing or incorrect activity. Thus, the developed software not only allows to extract a list of malfunctioning synapse drivers but also serves as an important tool for investigating issues with the digital event network.

4 Results

4.1 Measurement Results

In section 3.2, protocols for the characterisation of the STP parameters are presented. Measurements following these protocols were carried out during this thesis. The following section contains results for HICANN 84 of the first wafer-scale system. The methods' portability to other setups was verified.

4.1.1 Depression

The depression phase is characterised by parameters λ, N and U_{SE}. In hardware, the two floating gate voltages V_{stdf} and V_{dep} can be used to configure the first two quantities. The digital parameter *cap* controls the size of capacity C_2, which in turn influences the utilisation parameter U_{SE}.

Utilisation of Synaptic Efficacy U_{SE}

A sweep over *cap* and the 56 synapse drivers of a quarter of the chip was carried out. Its results are shown in figure 9. In order to understand the dependency of U_{SE} on the hardware parameter, a comparison of the experimental results to the theoretical model is required. As explained in 2.4, the increase of the inactive partition is implemented using a charge sharing process. In theory, the relative step size of the storage capacitor's charge can be described by

$$U_{\mathrm{SE,\,theo}} = \frac{C_2}{C_1 + C_2} \tag{25}$$

as presented in (Bill, 2008) and equation 10. Consequently, for $C_2 = 0$ a value of $U_{\mathrm{SE}} = 0$ would be expected. With an observed offset of 0.25, the hardware measurements clearly show a different behaviour. To account for these deviations, the model was extended with two additional parameters. C_0 represents parasitic capacities within the circuit. Especially the metal layer routing of V_{cap} reaching to the multiplexed storage capacities C_1 could contribute to these parasitics. U_0 accounts for an additional offset. Early measurements have shown that the introduction of U_0 improved the fit, though the origin of this offset is not yet fully understood. The fit model now reads as

$$U_{\mathrm{SE}} = \frac{C_0 + C_2}{C_0 + C_1 + C_2} + U_0\,. \tag{26}$$

For further analysis, transistor-level simulations of the STP circuitry were carried out. With the intention to keep the test bench simple, the time course of the inactive neurotransmitter partition was directly observed via V_{cap} instead of simulating a complete synapse driver. An exponential fit similar to the measurement protocol for short term

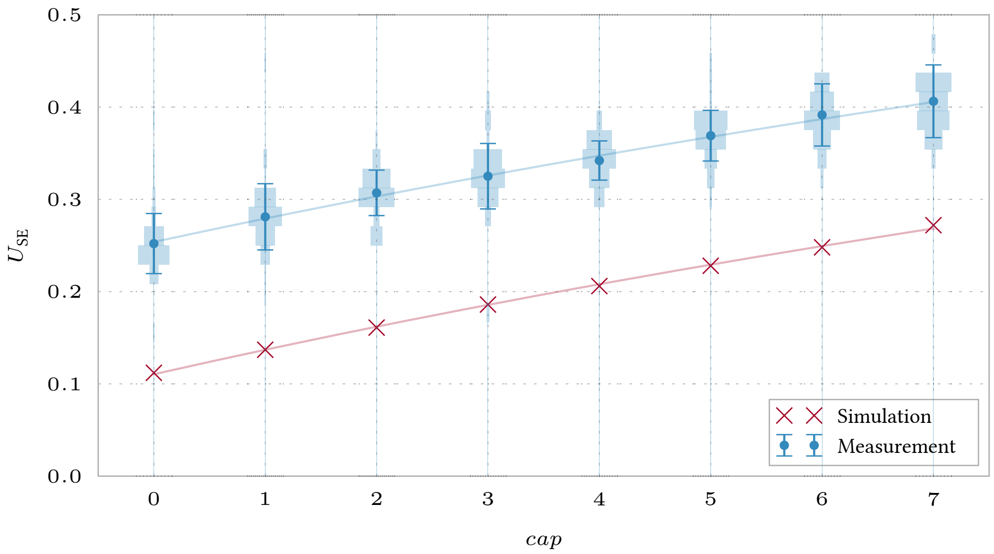

Figure 9: The utilisation of synaptic efficacy U_{SE} is shown depending on the digital parameter *cap* controlling the size of capacitor C_2. In this plot, hardware measurements across 56 synapse drivers are shown. Horizontal histograms are included as an overlay to give an indication of the statistical distribution of U_{SE} across multiple synapse drivers. Error bars indicate the standard deviation over those individual measurements. Also shown are the results of a transistor-level simulation of the circuit. The plot includes fits following equation 26 for both, the hardware measurement and the simulation. Table 3 contains the corresponding fit results.

depression was used to extract U_{SE}. The results of this simulation are also included in figure 9.

The fit results for both, the hardware measurement and the simulation are shown in table 3. The estimations for the parasitic capacity C_0 match for both scenarios. Furthermore, a parasitic extraction conducted by Andreas Hartel yielded very similar results. However, simulation and hardware measurements yield largely different results for U_0. Most likely, this difference can be explained by a non-linear translation of V_{cap} into the pulse width. This theory is supported by the observation, that U_{SE} increases for larger values of V_{stdf} and V_{dep} which influence the modulation of the output pulse. Additionally, it has to be considered that MOS capacities show a dependency on the gate voltage which might further distort the ratio of C_1 and C_2 and thus U_{SE}. Finally, only the part of the synapse driver specific to STP was simulated. For a more comprehensive comparison a complete synapse driver should be incorporated in the simulation.

To conclude, the measured dynamic range of $U_{\mathrm{SE}} \approx 0.25 \ldots 0.4$ allows for biology-inspired use cases. Compared to a natural range of $0.1 \ldots 0.95$ (Tsodyks and Markram, 1997), the hardware implementation can realise low to medium utilisation values. Furthermore, with deviations of 6 % to 13 % across multiple synapse drivers, the parameter turns out to be precisely adjustable.

Type	C_0	U_0
Hardware Measurement	(31.0 ± 4.8) fF	0.14 ± 0.01
Simulation	(25.0 ± 2.0) fF	0.02 ± 0.01
Parasitics Extraction	26.2 fF	–

Table 3: Results of the fits shown in figure 9 following equation 26 for hardware measurements as well as a transistor-level simulation of the circuit. Additionally, a parasitics extraction was conducted in software. While C_0 matches for hardware measurement, simulation and extraction of parasitics, this is not the case for U_0.

Scaling λ and Offset N

Scaling λ and offset parameter N of the depression mode result from the same measurement protocol. As already mentioned, λ and N are configured through V_{stdf} and V_{dep}. With a coarse sweep, the usable range was narrowed down. A finer measurement is shown in figures 10 and 11. While trends are observable for both, λ and N, no one-to-one correlation between the voltage and model parameters are visible. Furthermore, in this specific range of V_{dep}, the measurements behave very unstable. Large trial-to-trial variations of up to $\Delta\lambda = 0.09$ and $\Delta N = 0.05$ (standard deviation of ten measurements) were observed. This is caused by the fact that the floating gate cells do not show a linear behaviour in the range below 100 DAC values (Koke, 2014). Unfortunately, this corresponds to the range required for V_{dep}.

Nevertheless, areas with the desired values, $\lambda = 1$ and $N = 0$ respectively, can be found for both variables. By calculating the overlap of these two independent sets, suitable parameters can be found for V_{stdf} and V_{dep}. However, this approach can not be used efficiently for all synapse drivers, since the required number of measurements scales quadratically. Instead, an iterative algorithm was developed. For this method, two helper variables are defined as

$$\alpha = -\lambda \cdot N \,, \qquad \beta = 1 - \lambda \cdot (1 - N) \;. \tag{27}$$

According to equation 8, α represents the difference of the first EPSP and the reference amplitude $\hat{a}$ and β the deviation of the steady state amplitudes from 0, respectively.

The algorithm is initialised with two predefined values for V_{stdf} and V_{dep}. After measuring λ and N, the parameters are changed according to

$$\Delta V_{\text{stdf}} = c_1 \cdot \beta \,, \qquad \Delta V_{\text{dep}} = c_2 \cdot \alpha \,. \tag{28}$$

Constants c_i can be used to control the speed of the algorithm. By applying these corrections to the voltage parameters, the algorithm follows the gradients of λ and N. The algorithm terminates, when one of the two following exit conditions is met. An upper limit for the number of iterations is defined, for cases where the calibration does not converge.

Otherwise, the calibration is complete when α and β are within a predefined neighbourhood of zero. It is to be noted that this algorithm does not necessarily terminate with the same results for multiple runs. Due to floating gate variations, the path taken by this algorithm might vary. Furthermore, the operating point is not unique, several correct settings can exist.

With this algorithm, a preliminary calibration was generated for the top left and top right quarters of the chip. Since both V_{stdf} as well as V_{fac} are shared parameters, a common value must be inferred from the values of multiple individually calibrated drivers. Instead of measuring all 56 synapse drivers per side, the calibration was carried out for only eight drivers each, for this proof-of-concept. This sample size was chosen in order to reduce the execution time. Since a random distribution of deviations is expected, this should not significantly influence the quality of the resulting calibration. However, a future calibration should include all synapse drivers.

In order to quantify the precision of this method, a measurement across the drivers located in the top half was taken. For each side separately, V_{stdf} and V_{dep} were set to an average of the calibration results for the eight drivers. Despite the low number of calibrated individual instances, the calibration yielded $\lambda = 1.00 \pm 0.19$ and $N = 0.01 \pm 0.07$ for the complete top half. Unfortunately, a calibration is not able to decrease the variations between synapse drivers since a shared value has to be set. However, most network models will not use all synapse drivers of a chip for STP and blacklisting of outliers can be considered for a better precision.

4.1.2 Facilitation

For facilitation mode, the same parameters λ, N and U_{SE} are available. They can be measured following the protocol presented in 3.2.2. While the utilisation of synaptic efficacy behaves similarly to the depression mode, the other two quantities require further investigation.

They are controlled through floating gate voltages V_{stdf} and V_{fac}. These voltages were swept coarsely to find a usable operating range. In figures 14 and 15, a second, finer sweep is shown for λ and N, respectively. Since V_{fac} is configured to values in the medium range of floating gate voltages, the trial-to-trial variations decreased compared to the depression mode measurements. Noticeably, the measurements contain a much lower level of noise. However, as for the depression mode, a one-on-one assignment of the model's parameters to the floating gate voltages is not possible.

In order to calibrate $\lambda = 1$ and $N = 1$, an iterative algorithm for finding valid configuration values was implemented. The algorithm is similar to its equivalent for the depression mode. Now the helper variables are defined as

$$\alpha = 1 - \lambda \cdot N \,, \qquad\qquad \beta = -\lambda \cdot (1 - N) \ . \tag{29}$$

The first value stands for the relative amplitude of the first EPSP while the second one represents the difference between $\hat{a}$ and the steady state heights. For each iteration, these

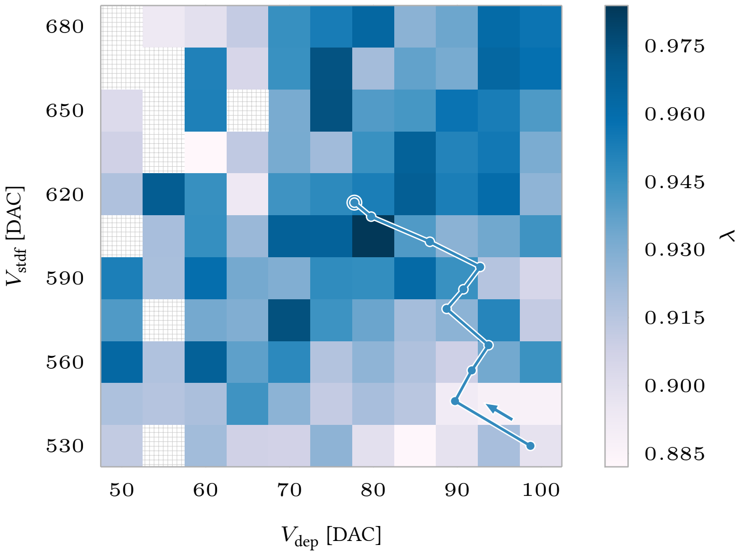

Figure 10: In this plot, λ is shown for depression mode as a dependency of V_{stdf} and V_{dep}. It can be observed that the variable depends on both parameters. As an overlay, the path taken by the iterative calibration algorithm is shown as it converges to an optimised value. Note, that outliers caused by large trial-to-trial variations were removed from this plot in order to optimise the dynamic range of the color plot. This sweep takes about 5 h with approximately 150 s per sample.

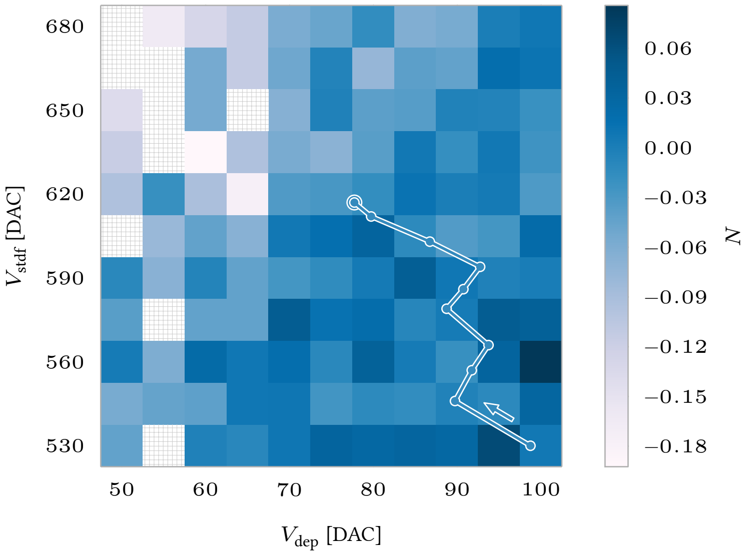

Figure 11: This plot shows the offset N for depression mode depending on V_{stdf} and V_{dep}. The data originates from the same measurement displayed in figure 10. As above, the path taken by the calibration method is included. Outliers were removed.

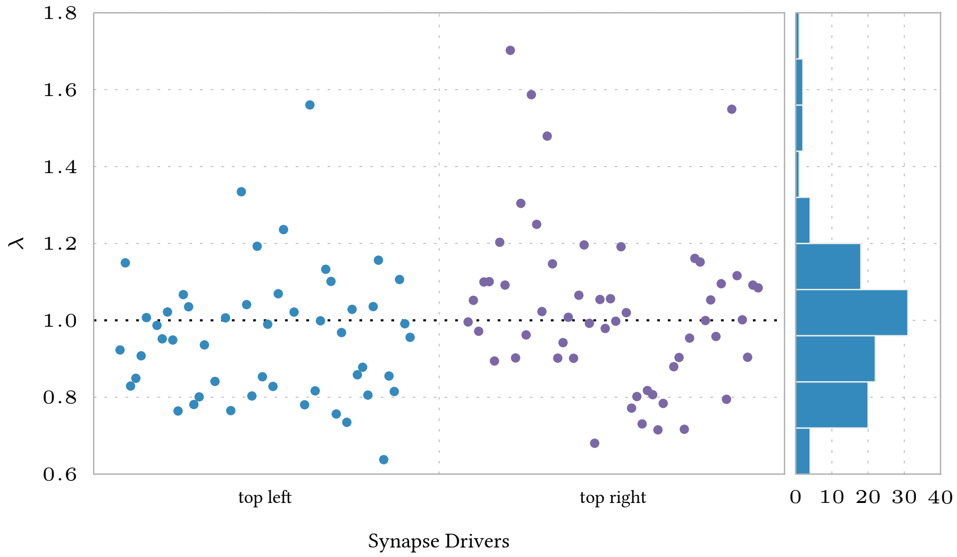

Figure 12: Result of a proof-of-concept calibration of the top half to $\lambda = 1$ in depression mode. Despite only including data from eight synapse drivers per side, the calibration can be considered successful with a resulting value of $\lambda = 1.00 \pm 0.19$. The left side was configured with $V_{\text{stdf}} = 597$ and $V_{\text{dep}} = 89$, while the calibration resulted in $V_{\text{stdf}} = 579$ and $V_{\text{dep}} = 98$ for the right half. Please note that only the average across a quarter's drivers can be calibrated. The variance can not be decreased, since both voltages are shared parameters. Measuring a single sample takes approximately 150 s.

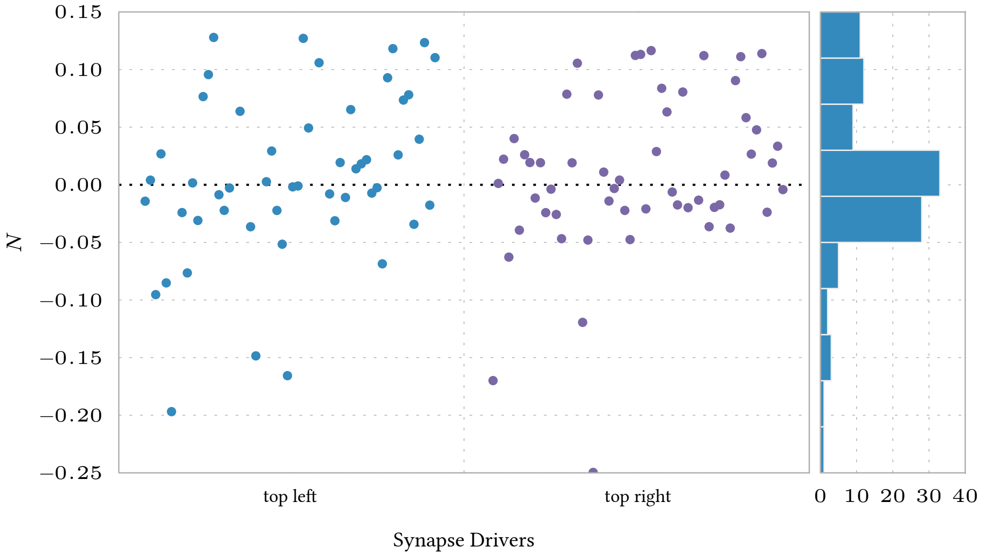

Figure 13: Calibration results to $N = 0$ for depression mode originating from the same data already shown in figure 13. The calibration resulted in $N = 0.01 \pm 0.07$.

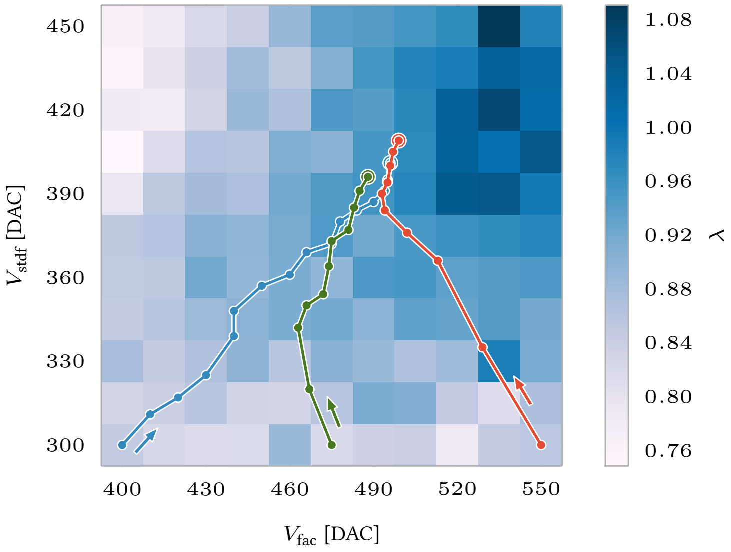

Figure 14: Scaling parameter λ depending on V_{stdf} and V_{fac} in facilitation mode. A direct one-to-one correlation of the observable to one of the voltage parameters is not possible, instead λ shows a dependency on both parameters. For three different initialisation values, the paths taken by the calibration algorithm are drawn as an overlay. It can be observed how they converge to similar final states. The measurements were all carried out for the same synapse driver. As can be seen in figure 16, the results vary for different synapse drivers.

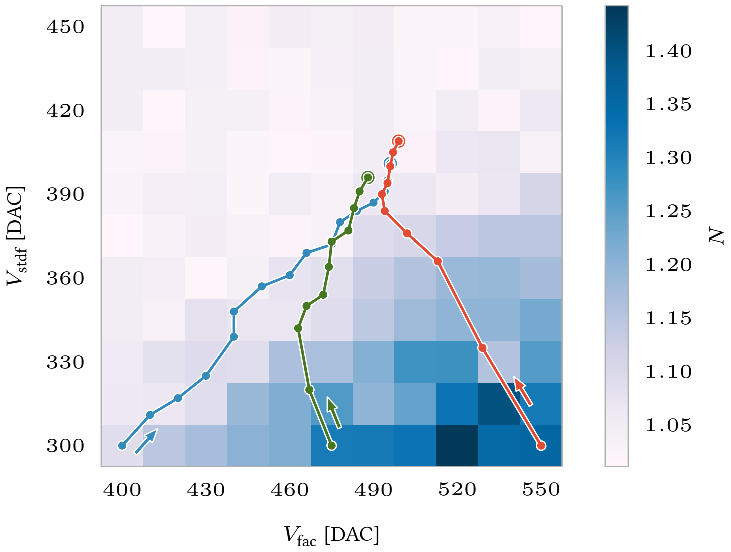

Figure 15: The same sweep as in figure 14 is shown for parameter N. While a gradient can be observed in the bottom right area, a plateau is reached slightly above $N = 1$.

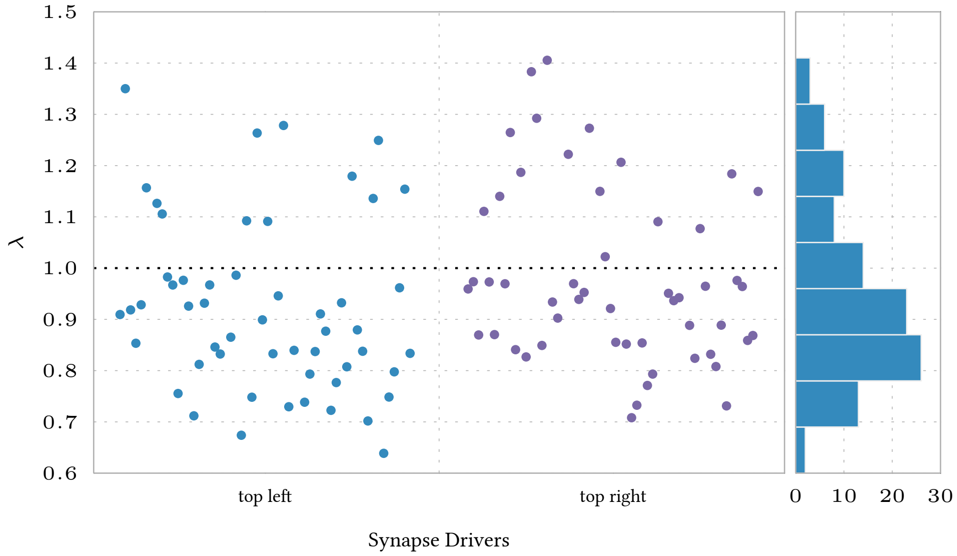

Figure 16: Results of the calibration of the top half of the chip to $\lambda = 1$ in facilitation mode. The top left and top right quarters were configured to $V_{\text{stdf}} = 440$, $V_{\text{fac}} = 494$ and $V_{\text{stdf}} = 425$, $V_{\text{fac}} = 472$ (DAC values), respectively. With $\lambda = 0.95 \pm 0.17$ the calibration did yield a usable result, despite being slightly shifted towards lower values.

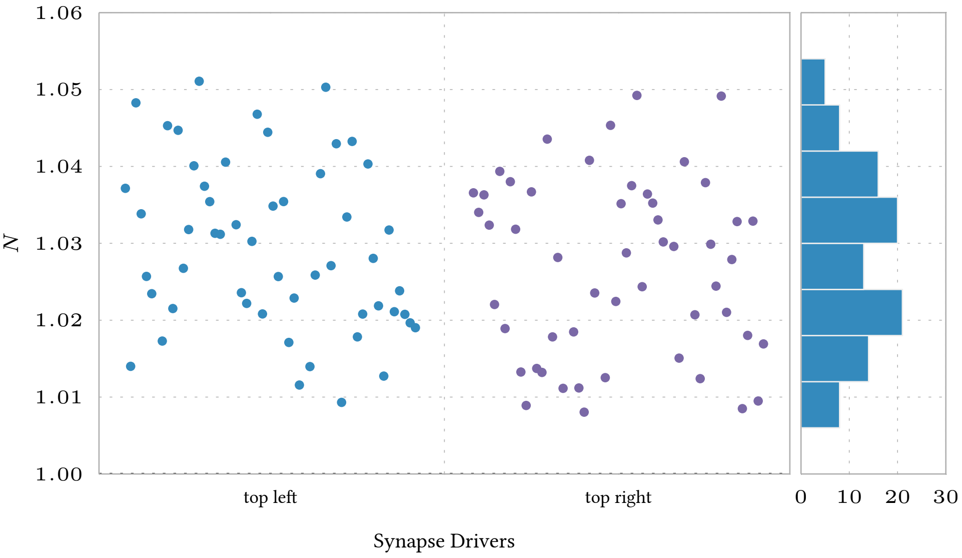

Figure 17: Results of the calibration to $N = 0$ from the same measurements as shown in figure 16. With $N = 1.03 \pm 0.01$, the calibration yielded usable results. As can be observed in figure 15, a plateau is reached for values slightly above 1. Thus, configuring the synapse drivers to a lower value is not possible.

quantities are calculated from the measurement results and then applied to the voltage parameters with constants c_1 and c_2 as defined by

$$\Delta V_{\text{stdf}} = c_1 \cdot \beta , \qquad \Delta V_{\text{fac}} = c_2 \cdot \alpha . \tag{30}$$

For each iteration, V_{stdf} and V_{fac} follow the gradients of λ and N and eventually converge to values suitable for a calibration. As for short term depression, these operating points are not necessarily unique. Exemplarily, paths taken by this algorithm are drawn on top of figures 14 and 15 for three different initialisation values. As can be seen, the paths converge to approximately the same final state.

As for depression, the calibration was tested for eight randomly chosen synapse drivers located on the top left and top right quarter of the chip, respectively. Both sides were then configured with the average of the eight individually acquired values and a sweep over all 112 synapse drivers of the top half was conducted. The results of this measurement are shown in figures 16 and 17. With $\lambda = 0.95 \pm 0.17$ the calibration again yielded usable results. Still, the calibration did not perfectly converge to the desired value. Similarly, the mean is slightly off for $N = 1.03 \pm 0.01$. As observable in figure 15, a plateau is reached for $N = 1$. Thus, lowering the mean value is not possible. It is expected that the calibration can be improved by taking all synapse drivers into account.

4.1.3 Recovery

As presented in equation 7, the recovery is characterised by its slope M which is configured through the floating gate current V_{dtc}. A sweep over this parameter is shown in figure 18 for a single synapse driver and source address. The fit indicates a linear dependency allowing for a straight-forward calibration method.

The two-staged current mirror design forwarding I_{rec} to the multiplexed capacities introduces large deviations. The first mirror leads to variations between different synapse drivers, while the second one causes further deviations across the 64 Level 1 addresses. The above sweep was extended to additionally quantify variations across multiple synapse drivers as well as source addresses. In figure 19, the results are shown. The linear dependence on V_{dtc} is still observable, while the slope now shows a wide, asymmetric distribution with deviations of 25 % to 67 %.

This distribution is inherent in the hardware implementation. Local within-die variations, primarily random dopand fluctuations, lead to a normally distributed threshold voltage V_t of the MOSFETs. Since the involved MOSFETs are biased in the deep subthreshold region, their drain current depends exponentially on V_t. Therefore, the recovery current I_{rec} is expected to feature a *lognormal* distribution. This is not only the case for deviations across synapse drivers. The second mirroring stage introduces further lognormally distributed fluctuations for different source addresses. The latter can be observed in figure 20a. Measuring the slope for different source addresses and synapse drivers leads to a sum of lognormal distributions which can be approximated by yet another lognormal distribution (Gubner, 2006). This behaviour can be observed in figure 20b. In both plots, a lognormal distribution was fitted to the data. While for a single synapse driver the num-

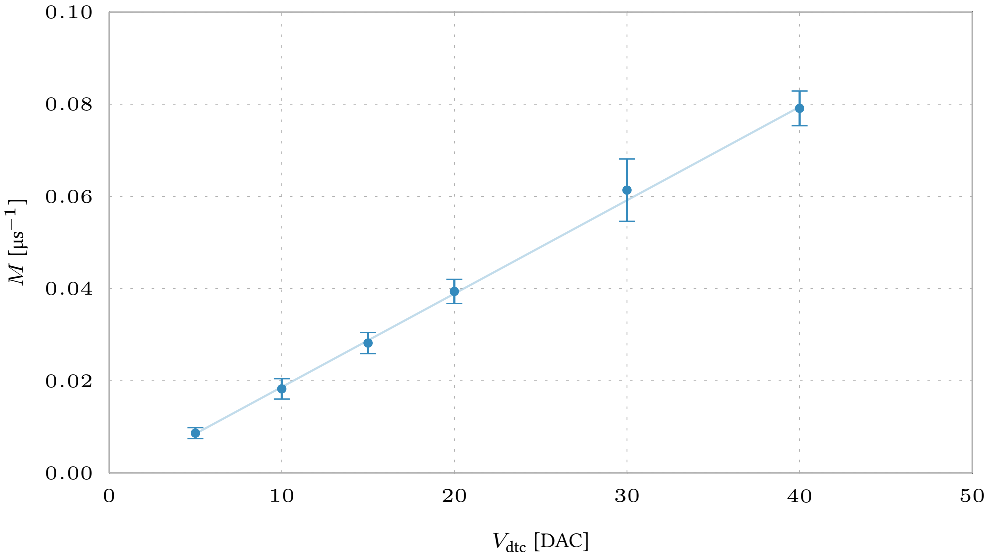

Figure 18: The slope of the recovery is plotted against V_{dtc} for a single synapse driver and source address. The average of ten measurements is shown, error bars indicate the corresponding trial-to-trial variations. A fit is included to highlight the linear dependency.

ber of samples is limited by the 63 usable Level 1 addresses and thus the statistics are not significant, the second histogram clearly follows the predicted distribution.

A comparison of the hardware results to the Tsodyks-Makram model is complicated by the fact that the latter implements an exponential recovery as opposed to the linear one in hardware. At least, the recovery time $T = {}^{1}/_{M}$ might represent a less abstract quantity in comparison to the slope itself. From the measurements shown in figure 19, a maximum recovery time of approximately $104 \pm 26\,\mu\text{s}$ can be calculated. With a speedup of 10^4, this results in $1040\,\text{ms}$ in the BTD, allowing to configure values similar to the behaviour of different biological cell types (Varela et al., 1997, Losonczy et al., 2002).

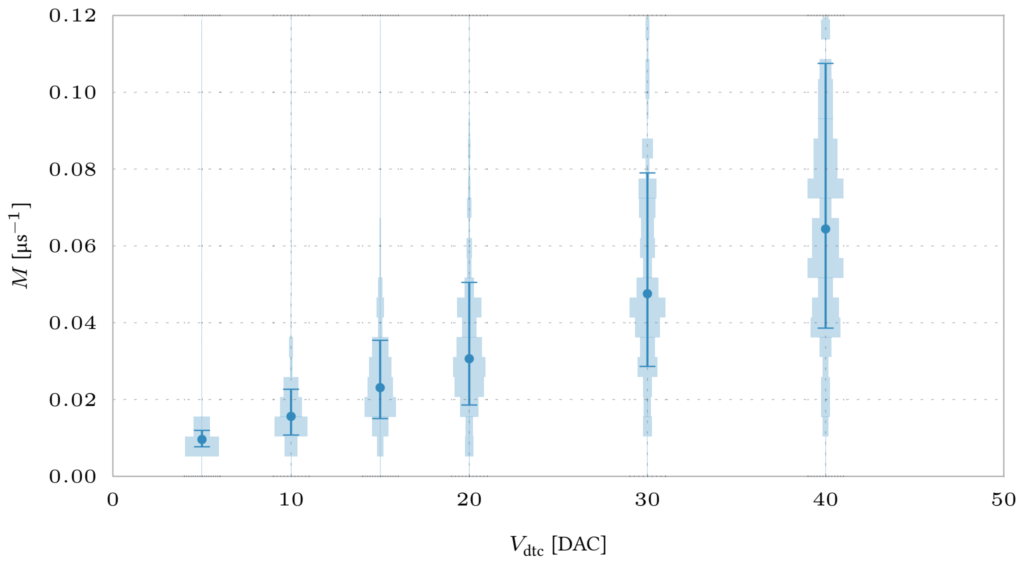

Figure 19: The slope of the recovery process dependending on V_{dtc}. The underlying data is based on a sweep over eight addresses for 19 synapse drivers each. Horizontal histograms are included as an overlay to indicate the statistical distribution. The asymmetric error bars show a confidence interval of 34 % on both sides of the geometric mean which is equal to the median for lognormal distributions.

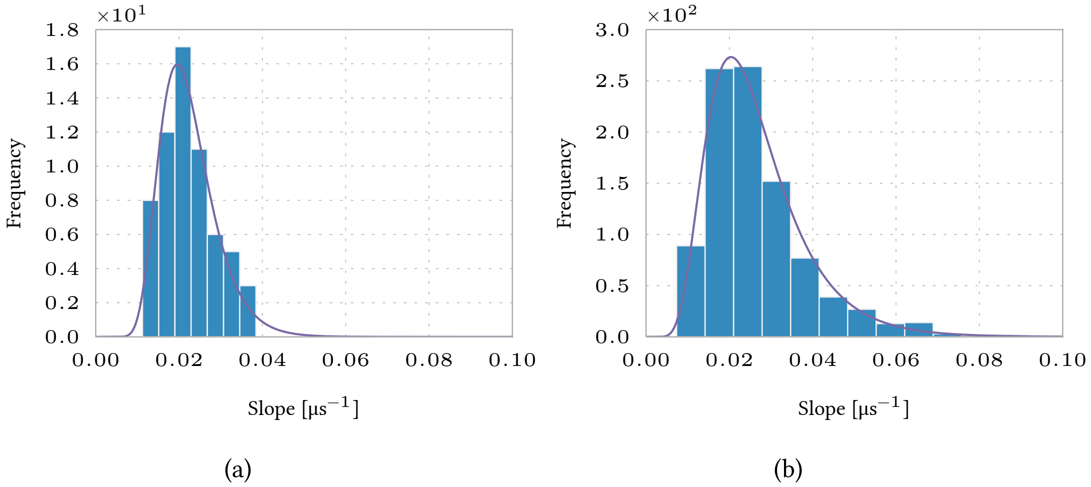

Figure 20: Distribution of the recovery process' slope M for multiple synapse drivers and source addresses. In (a), these deviations are shown for 63 input addresses of a single synapse driver, while (b) contains data of 63 source addresses for 15 drivers. In both plots, a lognormal distribution was fitted to the data. The measurements were taken for $V_{\text{dtc}} = 25$ DAC values.

4.2 Functional Testing of Synapse Drivers

The synapse driver defect detection tool introduced in 3.3 was implemented and then used to analyse several HICANNs. While first test runs were carried out by the author, the work was continued by Sebastian Schmitt. The following paragraphs will present and discuss preliminary findings.

As could be expected, events with address 0, originating from the background event generators (Schemmel et al., 2010) could be observed in all response patterns. Similarly, most responses contained incorrect spikes due to some neurons either firing continuously or showing no response at all. However, these erroneous spikes can be attributed to a non-ideal neuron calibration rather than defect synapse drivers.

Mainly, three typical response patterns could be observed: Approximately 80 % of the synapse drivers showed to operate correctly. This was indicated by a proper response pattern as shown in the topmost plot in figure 21. About 5 % of the drivers did not show a response pattern at all or forwarded random spike events to a subset of the connected neurons. This case is represented in the bottommost plot in figure 21. No spatial correlation for the defect drivers could be found. 15 % of the synapse drivers showed erroneous handling of addresses ≥ 32, which have a MSB of 1. The latter case is represented by the centre plot in figure 21.

Multiple experiments were conducted to examine the source of error (Schmitt, 2014). Firstly, the frequency of the background event generators was varied with the intention of significantly improving the synapse drivers' locking. This did not lead to lower fault rates. Furthermore, the influence of the slow bit of the DNC mergers (Schemmel et al., 2010) was investigated. It could be shown that deliberately disabling this bit setting resulted in a higher number of defects. Interestingly, the count of MSB faults could be reduced by disabling all but one sending repeaters, despite them being completely independent. Since a definite explanation is yet to be found, further investigations, e.g. tuning the Layer 1 signalling voltages V_{OH} and V_{OL}, are required to pinpoint the exact source of error.

Assuming all configuration problems are solved, the remaining malfunctioning synapse drivers can be blacklisted. This is feasible, since most network models do not utilise all synapse drivers available on a chip. In future revisions, the results of the synapse driver defect detection tool can be merged into the *redman* defect database (Klähn, 2013).

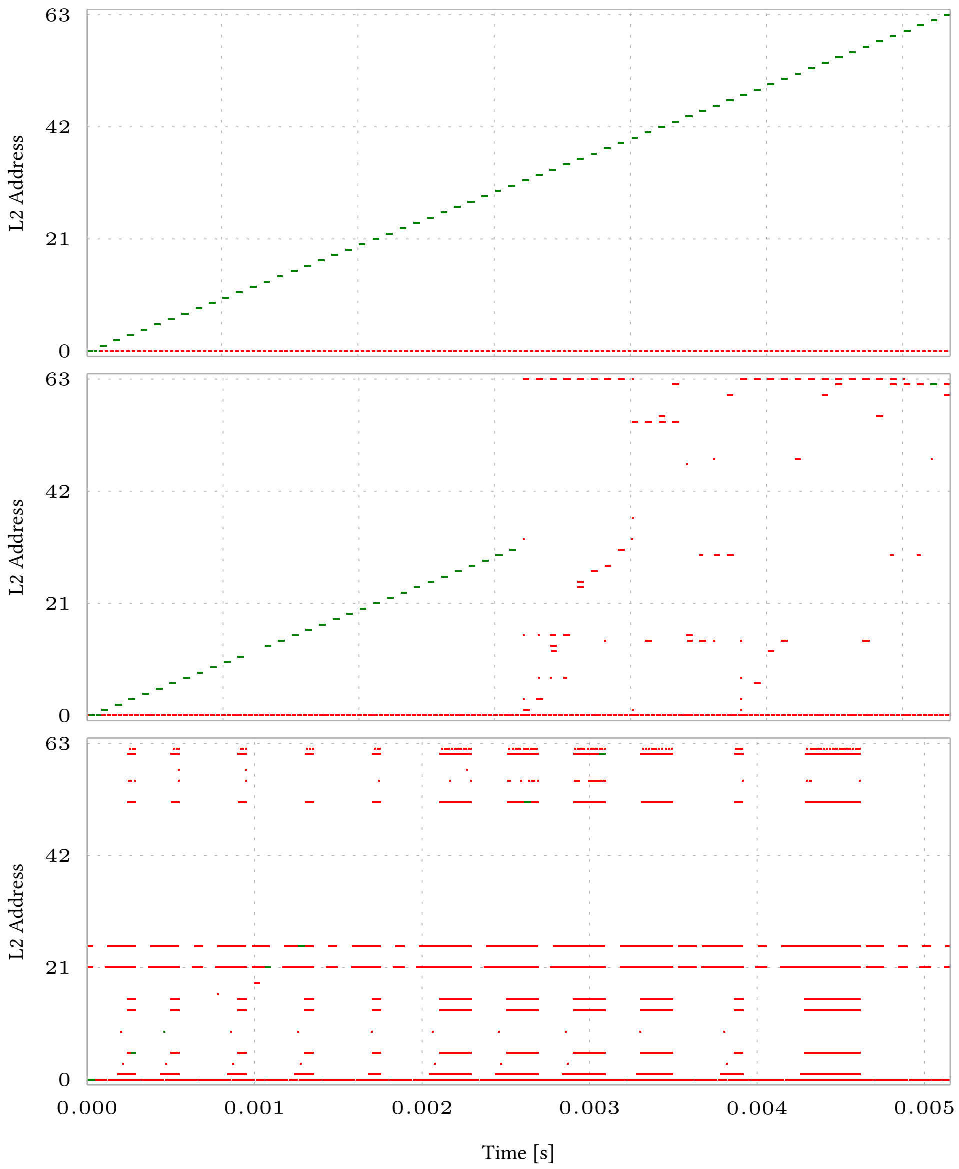

Figure 21: Results from the defect detection presented in section 3.3 exemplarily presented for three synapse drivers. In these plots, the neurons' responses to the input are shown. The *topmost* plot shows the result for a fully functional synapse driver, forwarding events with every possible L1 address. In the *centre*, a common defect scheme is displayed. Addresses ≥ 32 are not processed correctly, while all other events are not affected. Defect synapse drivers do not show the required response pattern but forward random events as shown in the *bottommost* plot.

5 Discussion and Outlook

The STP mechanism of the HICANN chip was examined during this thesis. Starting from a theoretical model based on the phenomenological description of STP by Tsodyks and Markram, measurement protocols for the characterisation of the neuromorphic implementation were developed and implemented. Measurements on the HICANN chip were carried out which proved the suitability of the circuit for short term depression as well as short term facilitation. Furthermore, configurations yielding biology-inspired parameter ranges were demonstrated. For the utilisation of synaptic efficacy U_{SE}, a dynamic range of $0.25 \dots 0.4$ was found and the recovery term showed to be configurable up to recovery times of approximately 1 s in the biological time domain. These values at least partly match observations in biology (Tsodyks and Markram, 1997, Varela et al., 1997, Losonczy et al., 2002).

Being inherent in the production of integrated circuits, large variations between different synapse driver instances were observed. Particularly, this applies to the recovery process's implementation. Here, deviations of 25 % to 67 % have to be expected for the recovery times. Similar observations were made for the hardware specific offset and scaling parameters. Since the floating gate voltages are shared for groups of 56 synapse drivers each, the variability of the measured parameters can not be compensated by calibration. However, algorithms for calibrating the mean across different drivers to the desired values were discussed. Proof-of-concept implementations showed that a calibration with a precision of approximately $\pm 20\,\%$ is feasible for the scaling and offset parameters. The utilisation of synaptic efficacy U_{SE} showed to be precisely adjustable. Still, deviations from the theoretical model were observed by simulating the corresponding circuitry on transistor-level.

Currently, the configuration through shared parameters represents the limiting factor for a calibration. Allowing an independent configuration would significantly reduce variations between synapse drivers. Furthermore, the sizing of the adjustable capacitance could be improved in order to enable larger values of the utilisation of synaptic efficacy.

Exposing the hardware's STP features to the end-user still requires further work. First and foremost, the scalability of the calibration methods developed in this thesis needs to be improved. Currently, the execution speed is limited by hardware configuration times. By improving the algorithms to reduce the configuration overhead and parallelise the characterisation of multiple synapse drivers, the calibration can be optimised. With the planned change to the ARQ communication protocol, a drastic decrease in hardware configuration times can be expected (Karasenko, 2014) allowing for large-scale characterisation of STP on the wafer-scale system incorporating a higher number of HICANNs. Additionally, the algorithms presented in this thesis sould be integrated with the current neuron calibration stack.

Furthermore, software support for STP must still be improved throughout the stack for the HMF. The top level layer, a PyNN-compatible software interface for the configuration of neural network models called *PyHMF*, contains support for STP following the Tsodyks-

Markram model (Billaudelle, 2014). The need for a custom configuration interface incorporating hardware-specific behaviour such as the linear recovery has to be evaluated. On a lower level, appropriate handling of STP related parameters needs to be implemented in the mapping and routing software *marocco* (Jeltsch, 2014). Algorithms for the assignment of synapse drivers were partly reimplemented by the author in order to take these additional settings into account. However, further efforts are still required.

As a side effect, the results from this thesis might prove valuable for an enhanced support of STP on the Spikey chip (Pfeil, 2014). As the implementations are nearly identical on both hardware platforms, the existing, yet rudimentary calibration could be improved with the findings and algorithms from this thesis.

Backed by a complete software stack, the realisation of network models requiring STP will be possible. The measured parameter ranges outlined above allow for an implementation of different neural networks on the HICANN chip. For example, the self-tuning network developed for Spikey (Bill, 2008) can be adopted without the need for changes in the STP configuration. The same applies to the cortical attractor-memory network (Breitwieser, 2011), which is currently under development as a benchmark for neuromorphic hardware systems (Rivkin, 2014). The influence of the linear recovery – as opposed to the exponential behaviour featured in the Tsodyks-Markram model – on such models still needs to be investigated.

To conclude, the implementation of STP on the HICANN chip showed to be operational. Assuming the existence of a feature-complete neuron calibration, the realisation of first network models incorporating STP can be approached.

A Verification of the Measurement Protocol for Depression

The measurement protocol for the depression phase presented in 3.2.1 has been tested with a simulation of short term depression. For this simulation, PyNN has been used in connection to the NEST backend. This STP implementation follows the Tsodyks-Makram model. Therefore, only U_{SE} can be configured for the depression phase, while $\lambda = 1$ and $N = 0$ are fixed.

In figure 22, the simulation is shown including the protocol's fit. The amplitudes have been extracted following the method presented in 3.1.4. The fit results as well as the configured values are shown in table 4.

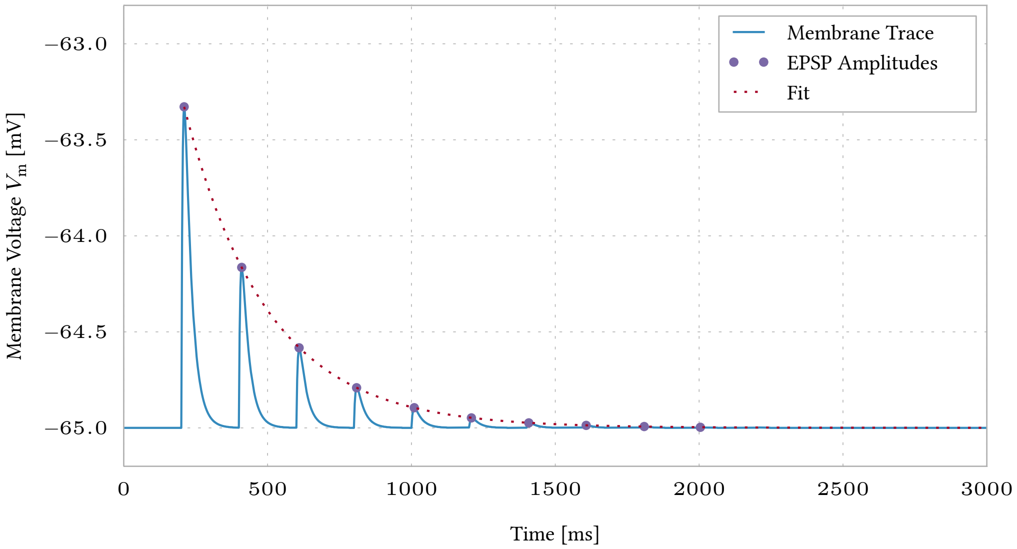

Figure 22: A simulation of STP including the extracted PSP amplitudes as well as a fit following the protocol for the characterization of the depression phase.

Parameter	Configured Value	Extracted Value
Utilization of Synaptic Efficacy U_{SE}	0.5	0.5003 ± 0.0002
Scaling Parameter λ	1 (fixed)	0.9999 ± 0.0003
Offset Parameter N	0 (fixed)	0.0000 ± 0.0002

Table 4: Fit results of the depression protocol for a simulation of STP. Parameters λ and N are not configurable in the Tsodyks-Markram model.

B Floating Gate Parameters

In the table 5, the floating gate parameters used as a starting point for the measurement presented in this thesis are shown. While unusual values were chosen e.g. for E_l, the measurement principles were tested with other parameters as well.

Type	Name	Value [DAC]	Comment
shared	V_fac	400...1023	Swept for facilitation measurements.
	V_dep	0...200	Swept for depression measurements.
	V_stdf	400...800	Most measurements taken with 200 DAC values.
	V_reset	200	
	V_dtc	0...100	Swept for recovery measurements.
	V_gmax0	50	
	V_bstdf	400...800	
neuron	V_t	500	
	I_gl	1023	Minimised in order to increase separation of PSPs.
	V_syntcx	800	
	V_syntci	800	
	V_synx	100	
	E_l	100	Set to low level in order to maximise PSP amplitudes.

Table 5: Common floating gate parameters used as a starting point for most measurements. Optimised for large PSP amplitudes and thus well suited for STP measurements.

The adaptive and exponential terms were disabled through floating gate settings shown in table 6.

Type	Name	Value [DAC]
neuron	V_exp	1023
	I_rexp	1023
	I_bexp	1023

Type	Name	Value [DAC]
neuron	I_gladapt	0
	I_fire	0
	I_radapt	1023

Table 6: Common floating gate parameters used as a starting point for most measurements. Optimised for large PSP amplitudes and thus well suited for STP measurements.

References

S. A. Aamir. personal communication, 2014.

G. Q. Bi and M. M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 18(24):10464–10472, Dec. 1998. ISSN 0270-6474. URL `http://www.jneurosci.org/content/18/24/10464.abstract`.

J. Bill. Self-stabilizing network architectures on a neuromorphic hardware system. Diploma thesis (English), University of Heidelberg, HD-KIP-08-44, 2008.

J. Bill, K. Schuch, D. Brüderle, J. Schemmel, W. Maass, and K. Meier. Compensating inhomogeneities of neuromorphic VLSI devices via short-term synaptic plasticity. *Front. Comp. Neurosci.*, 4(129), 2010.

S. Billaudelle. PyHMF – eine PyNN-kompatible Schnittstelle für das HMF-System, 2014.

BrainScaleS. Research. `http://brainscales.kip.uni-heidelberg.de/public/index.html`, 2012.

O. Breitwieser. Investigation of a cortical attractor-memory network. Bachelor thesis, Ruprecht-Karls-Universität Heidelberg, 2011. URL `http://www.kip.uni-heidelberg.de/Veroeffentlichungen/details.php?id=2637`. HD-KIP 11-173.

R. Brette and W. Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. Neurophysiol.*, 94:3637 – 3642, 2005. doi: NA.

A. P. Davison, D. Brüderle, J. Eppler, J. Kremkow, E. Muller, D. Pecevski, L. Perrinet, and P. Yger. PyNN: a common interface for neuronal network simulators. *Front. Neuroinform.*, 2(11), 2008.

M. Diesmann and M.-O. Gewaltig. NEST: An environment for neural systems simulations. In T. Plesser and V. Macho, editors, *Forschung und wisschenschaftliches Rechnen, Beiträge zum Heinz-Billing-Preis 2001*, volume 58 of *GWDG-Bericht*, pages 43–70. Ges. für Wiss. Datenverarbeitung, Göttingen, 2002.

D. A. Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, 2005.

EPFL and IBM. Blue brain project, 2008. URL `http://bluebrain.epfl.ch/`.

M.-O. Gewaltig and M. Diesmann. NEST (NEural Simulation Tool). *Scholarpedia*, 2(4):1430, 2007.

D. F. Goodman and R. Brette. The brian simulator. *Frontiers in neuroscience*, 3(2):192, 2009.

J. A. Gubner. A new formula for lognormal characteristic functions. *Vehicular Technology, IEEE Transactions on*, 55(5):1668–1671, 2006.

HBP SP9 partners. *Neuromorphic Platform Specification.* Human Brain Project, Mar. 2014.

M. H. 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 `http://www.frontiersin.org/computational_neuroscience/10.3389/fncom.2013.00045/abstract`.

S. Jeltsch. *A Scalable Workflow for a Configurable Neuromorphic Platform.* PhD thesis, Universität Heidelberg, 2014.

V. Karasenko. A communication infrastructure for a neuromorphic system. Master's thesis (English), University of Heidelberg, 2014.

J. Klähn. Untersuchung und Management von Synapsendefektverteilungen in einem großskaligen neuromorphen Hardwaresystem. Bachelor thesis, Ruprecht-Karls-Universität Heidelberg, 2013. URL `http://www.kip.uni-heidelberg.de/Veroeffentlichungen/details.php?id=2825`. HD-KIP 13-36.

C. Koke. personal communication, 2014.

T. Lande, H. Ranjbar, M. Ismail, and Y. Berg. An analog floating-gate memory in a standard digital technology. In *Microelectronics for Neural Networks, 1996., Proceedings of Fifth International Conference on*, pages 271 –276, 12-14 1996. doi: 10.1109/MNNFS.1996.493802.

A. Losonczy, L. Zhang, R. Shigemoto, P. Somogyi, and Z. Nusser. Cell type dependence and variability in the short-term plasticity of epscs in identified mouse hippocampal interneurones. *The Journal of physiology*, 542(1):193–210, 2002.

E. H. Nicollian, J. R. Brews, and E. H. Nicollian. *MOS (metal oxide semiconductor) physics and technology*, volume 1987. Wiley New York et al., 1982.

B. Pan and R. S. Zucker. A general model of synaptic transmission and short-term plasticity. *Neuron*, 62(4):539–554, 2009.

T. Pfeil. Personal communication., 2014.

W. 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`.

B. Rivkin. personal communication, 2014.

J. Schemmel. personal communication, 2014.

J. Schemmel, A. Grübl, K. Meier, and E. Muller. Implementing synaptic plasticity in a VLSI spiking neural network model. In *Proceedings of the 2006 International Joint Conference on Neural Networks (IJCNN)*. IEEE Press, 2006.

J. Schemmel, J. Fieres, and K. Meier. Wafer-scale integration of analog neural networks. In *Proceedings of the 2008 International Joint Conference on Neural Networks (IJCNN)*, 2008.

J. Schemmel, A. Grübl, and S. Millner. Specification of the HICANN microchip. FACETS project internal documentation, 2010.

S. Schmitt. personal communication, 2014.

M. Tsodyks and H. Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the national academy of science USA*, 94:719–723, Jan. 1997.

van Rossum, G.-Q. Bi, and G. Turrigiano. Stable hebbian learning from spike timing-dependent plasticity. *J Neurosci.*, 20:8812–21, 2000.

J. A. Varela, K. Sen, J. Gibson, J. Fost, L. Abbott, and S. B. Nelson. A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. *The Journal of neuroscience*, 17(20):7926–7940, 1997.

R. S. Zucker and W. G. Regehr. Short-term synaptic plasticity. *Annu. Rev. Physiol.*, 64: 355–405, 2002.

Acknowledgements

I would like to express my thanks of grattitude to Prof. Dr. Karlheinz Meier for giving me the opportunity to investigate this interesting topic and to conduct the oral examination. The same applies to Dr. Johannes Schemmel, not only for taking over the role as second revisor but also for his patience and willingness to help and explain the involved cicuitry. Special thanks also to my supervisors Sebastian Jeltsch and Christoph Koke who supported me during the work on this thesis. Furthermore, I would not have been able to complete this thesis without the help of Andreas Hartel helping me out with the transistor-level simulations.

Most importantly, my thanks go to Arno, Andreas, Christoph, Joscha, Laura, Mitja, Sebastian, Sebastian and especially Tobias, who helped me with proofreading this pile of paper (or pixels). I really appreciate your efforts!
Many thanks also to Eric, who tirelessly kept everything running (Did you know that one can login on every workstation in the container? Amazing!). Dear container people, I really enjoyed working with you – even without an AC. It was a lot of fun! Thanks to Kai who constantly filled up the *HBP Erfrischungskiste* with the elixir of life.
Last but not least, thanks to my friends and family, especially my mother Regina and my girlfriend Alexandra, who supported me during the last few weeks wherever it was possible!

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, October 31, 2014

....

(signature)