Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the

Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

Dipl.-Phys. Sebastian Millner

born in Witten, Germany

Date of oral examination: November 6, 2012

Development of a Multi-Compartment Neuron Model Emulation

Referees: Prof. Dr. Karlheinz Meier
Prof. Dr. Peter Fischer

Development of a Multi-Compartment Neuron Model Emulation

This work describes the design of an analog circuit emulating a multi-compartment neuron model on a microchip. Initially, the single-compartment adaptive exponential integrate-and-fire neuron model is implemented as a hardware model. Therefor, the differential equations describing the model dynamics are directly translated into an electronic circuit based on operational transconductance amplifiers. Consequently a close correspondence between model and circuit is achieved enabling references to experiments done with computer simulators. 512 of these neurons are implemented on a single micro-chip. Individual control of each neuron's biases is achieved by the use of analog floating-gate memory. In most cases, these biases directly correspondent to parameters of the model, hence simple translations are possible.

The single neuron implementation has been verified on a prototype chip in several experiments. Inter alia, its capabilities of reproducing biological neuron's behavior and the influence of fixed-pattern noise on the circuit are analyzed.

To step over to a multi-compartment circuit, the neuron has been enhanced by a resistive element and a routing network to build complex dendrite structures. Furthermore, the parameterization allows compartments of different sizes covering large somatic and small dendritic compartments. A dedicated test chip has been designed for the verification of the new model. Several simulations show the enhanced behavior of the multi-compartment emulation including dendritic attenuation and active spike propagation.

The neuron circuits are dedicated for a new kind of computer based on the cortex.

Entwicklung einer Multi-Kompartiment-Neuronenmodel-Emulation

Diese Arbeit beschreibt den Entwurf einer analogen Schaltung zur Emulation eines Multi-Kompartiment-Neuronenmodels auf einen Mikrochip. Zunächst beschränkt auf einzelne Kompartimente wird das Adaptive Exponential Integrate-and-Fire Neuronenmodell implementiert. Hierzu werden die Differentialgleichungen des Modells durch Transkonduktanzverstärker direkt in elektrische Schaltungen übersetzt. Folglich wird eine enge Korrespondenz zwischen Schaltung und Model erreicht, wodurch es möglich wird, Ergebnisse aus Computersimulationen als Referenzen zu verwenden. 512 dieser Schaltungen werden auf einem Chip integriert. Für jedes einzelne Neuron können Steuerspannungen und Ströme individuel durch analoge Floating-Gates konfiguriert werden. In der Regel sind die Schaltungsparameter in direkter Beziehung zu den Modellparametern.

Durch Testchips wurde die Schaltung mit mehreren Experimenten verifiziert. Unter anderem wird das Reproduzieren spezieller Verhaltensmuster biologischer Neuronen gezeigt. Ferner wird das Verhalten bezüglich Produktionsschwankungen analysiert.

Um eine Multi-Kompartiment-Emulation zu konstruieren, wird die Schaltung im Weiteren um ein resistives Element und ein Schaltnetzwerk erweitert. Dadurch wird es möglich komplexe Dendriten nachzubilden. Außerdem wird die Parametrisierung so erweitert, dass sowohl große Somakompartimente, als auch kleine Dendritenkompartimente nachgebildet werden können. Zur Verifikation wurde ein Testchip entworfen. Mehrere Simulationen zeigen das erweiterte Verhalten des Multi-Kompartiment-Models. So werden Dämpfung, sowie die Weiterleitung von Aktionspotentialen im Dendrit gezeigt.

Die Neuronen wurden zur Integration in einen neuartigen Computer entworfen, dessen Funktion auf den Prinzipien des Gehirns basiert.

"Analog design is art and science at the same time. It is art because it requires creativity to strike the right compromises between the specifications imposed and the ones forgotten. It is also science because it requires a certain level of methodology to carry out a design, inevitably leading to more insight in the compromises taken."

Willy M. C. Sansen in *Analog Design Essentials*, Springer 2006

"Thus, we are no longer confident as we were 18 years ago that simplicity will eventually emerge from the complexity. The extreme sophistication of cellular mechanisms will challenge cell biologists throughout the new century [. . .]."

From the preface of *Molecular Biology of the Cell*, fourth edition by Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Robert, and Peter Walter, 2002

Contents

Introduction **1**

1 Neuroscience, Circuits and Neuromorphic Engineering **5**
1.1 Biological Substrate . . . 5
1.1.1 Neurons . . . 5
1.1.2 Synapses . . . 6
1.1.3 Plasticity . . . 8
1.1.4 Typical Cortical Neuron Behavior . . . 8
1.1.5 Measurement Capabilities . . . 8
1.1.6 High-Conductance State . . . 9
1.2 Single Cell Modeling . . . 9
1.2.1 The Hodgkin Huxley Model . . . 9
1.2.2 The Integrate-and-Fire Model . . . 10
1.2.3 Adaptation . . . 11
1.2.4 Positive Feedback . . . 11
1.3 Technical Substrate . . . 11
1.3.1 Transistors . . . 12
1.3.2 Basic Transistor Circuits . . . 14
1.3.3 Simulation Techniques . . . 15
1.4 Rebuilding Biology - Neuromorphic Engineering . . . 16
1.4.1 Emulation . . . 16
1.4.2 Sensors, Neurons and Systems . . . 18
1.5 Single Cell Emulation . . . 18
1.5.1 The Design Approach Triangle . . . 18
1.5.2 Ion Channel Implementation . . . 20

2 Neuromorphic Environment **25**
2.1 The BrainScaleS Project . . . 25
2.1.1 Interaction . . . 25
2.1.2 Hybrid Multi-Scale Computing Facility . . . 25
2.2 The BrainScaleS Wafer-Scale System . . . 26
2.2.1 Communication . . . 27
2.3 The HICANN Microchip . . . 28
2.3.1 Life Time of an Action Potential . . . 29
2.3.2 High Input-Count . . . 30

3 Point Neuron Emulation **31**
3.1 The Adaptive Exponential Integrate-and-Fire Neuron Model . . . 31
3.1.1 Model Description . . . 31
3.1.2 Model Dynamics . . . 32
3.1.3 Synaptic Stimulation . . . 33
3.1.4 Conclusion . . . 33

3.2 Structure and Design Concept . . . 34
3.3 Operational Transconductance Amplifier and Leakage . . . 35
3.3.1 Ideal Operation . . . 35
3.3.2 Circuit . . . 36
3.3.3 Simulation Results . . . 38
3.3.4 Conclusion . . . 40
3.4 Membrane Capacitor . . . 40
3.5 Adaptation . . . 41
3.5.1 Circuit and Theory . . . 41
3.5.2 Real Circuit Behavior . . . 42
3.5.3 Conclusion . . . 43
3.6 Synaptic Input . . . 43
3.6.1 Conductance Shape: Theory . . . 43
3.6.2 The Resistive Element . . . 44
3.6.3 Conductance Shape: Simulated Circuit . . . 45
3.6.4 Weight Saturation . . . 47
3.6.5 Delays . . . 48
3.6.6 Conclusion . . . 48
3.7 Exponential Term . . . 49
3.7.1 Circuit Principle . . . 49
3.7.2 Voltage Divider . . . 51
3.7.3 Complete Circuit Simulation . . . 52
3.7.4 Conclusion . . . 54
3.8 Spike Detection . . . 54
3.9 Resetting . . . 55
3.10 Neuron Connectivity . . . 56
3.10.1 Circuit . . . 57
3.10.2 Simulation . . . 57
3.11 Readout and Stimulation . . . 58
3.11.1 Analog Readout . . . 59
3.11.2 Current Stimulation . . . 59
3.12 Parameterization . . . 61
3.12.1 Biological Parameter Ranges . . . 61
3.12.2 Parameter Translation . . . 63
3.12.3 Realization . . . 64
3.12.4 Hardware Parameter Summary . . . 65

4 Point Neuron Experiments **69**
4.1 Methods . . . 69
4.1.1 Evaluation Setups . . . 69
4.1.2 Wafer-Scale Setup . . . 72
4.1.3 ASICs . . . 73
4.2 Characterization of Output Capabilities . . . 73
4.2.1 Methods . . . 73
4.2.2 Results . . . 74
4.2.3 Conclusion . . . 76
4.3 Reference Emulation . . . 77
4.3.1 Methods . . . 78
4.3.2 Results . . . 78
4.3.3 Conclusion . . . 80

4.4 Characteristic Patterns . . . 80
4.4.1 Methods . . . 81
4.4.2 Results . . . 82
4.4.3 Conclusion . . . 87
4.5 Compartmental Effects . . . 88
4.5.1 Methods . . . 88
4.5.2 Results . . . 89
4.6 Fixed Pattern Noise . . . 90
4.6.1 Methods . . . 90
4.6.2 Results . . . 91
4.7 Simple Networks . . . 95
4.7.1 Methods . . . 95
4.7.2 Results . . . 96
4.8 Reproducing Computer Simulations . . . 96
4.8.1 Methods . . . 97
4.8.2 Results . . . 97
4.8.3 Conclusion . . . 99

5 Discussion: Single-compartment 101
5.1 Model Implementation . . . 101
5.2 Measurements . . . 102
5.3 Comparison to Other Implementations . . . 103

6 Multi-Compartment Emulation 105
6.1 Biological Concepts . . . 105
6.1.1 Cable and Compartments . . . 105
6.1.2 Passive Computational Power . . . 106
6.1.3 Active Channels . . . 107
6.1.4 Which Model to Use? . . . 108
6.1.5 Where to Cut? . . . 109
6.2 Multi-Compartment Implementations . . . 109
6.3 Circuit Structure and Concepts . . . 110
6.4 Inter Compartment Resistance . . . 111
6.4.1 Transconductors . . . 112
6.4.2 Resistive Element . . . 113
6.4.3 Conclusion . . . 115
6.5 Firing Modes - the Interface Module . . . 116
6.5.1 Spikes . . . 116
6.5.2 Implementation . . . 117
6.6 Dendrite Routing . . . 118
6.6.1 Building Neurons . . . 118
6.6.2 Pass Transistors . . . 119
6.6.3 Routing matrix . . . 119
6.7 Reset mechanism . . . 120
6.8 Parameterization . . . 121
6.8.1 Range extraction . . . 121
6.8.2 Parameter translation . . . 122
6.8.3 Realisation . . . 123
6.9 Additional Changes . . . 123
6.10 The Multi-Compartment Chip . . . 123

7 Multi-Compartment Experiments **127**
7.1 Four Compartment Reference Simulation . . . 127
7.1.1 Methods . . . 127
7.1.2 Results . . . 128
7.2 Action Potentials with Active and Passive Dendrites . . . 128
7.2.1 Methods . . . 128
7.2.2 Results . . . 129
7.2.3 Conclusion . . . 132

8 Discussion: Multi-Compartment **133**
8.1 Model Choice . . . 133
8.2 Implementation . . . 134
8.3 Simulation Results . . . 134
8.4 Other Implementations . . . 135
8.5 Conclusion . . . 135

9 Analog Floating-Gate Memory **137**
9.1 Cells . . . 137
9.1.1 Voltage Cells . . . 138
9.1.2 Current Cells . . . 138
9.2 Architecture . . . 139
9.2.1 Array . . . 140
9.2.2 Driver . . . 140
9.2.3 Decoder . . . 141
9.2.4 Controller . . . 141
9.3 HICANN Integration . . . 141
9.3.1 Biasing . . . 142
9.3.2 Level shifter . . . 142
9.3.3 Global Parameters . . . 143
9.3.4 Current Source . . . 143
9.3.5 Layout . . . 144
9.4 Digital Controller . . . 144
9.4.1 Programming Functions . . . 144
9.4.2 Additional Functions . . . 145
9.4.3 Detailed Implementation . . . 145
9.4.4 Test Environment . . . 145
9.5 Sources of Variance . . . 146
9.5.1 Parameter Drifts . . . 147
9.5.2 Crosstalk . . . 147
9.5.3 Output Settling and Strobe . . . 147
9.5.4 Programming Limits . . . 148
9.6 Improvements . . . 148
9.6.1 Control, Driver and Decoder Revision . . . 148
9.6.2 Cell Revision . . . 148
9.6.3 Biasing Revision . . . 149
9.7 Test Results . . . 150
9.7.1 General Functioning . . . 150
9.7.2 Programming Schemes . . . 151
9.7.3 Precision . . . 152
9.7.4 Crosstalk . . . 153

9.7.5 Stress Test 154
9.7.6 New Cells 155
9.8 Discussion 155

10 Final Remarks and Outlook **157**

Nomenclature **161**

Bibliography **163**

Acknowledgments **169**

Introduction

Apart from the big bang and the question of how the material world is composed, the comprehension of the brain is probably one of the most important matters of today's science or science at any time. Although the brain it-self is as close to each single human as anything can be, even simple networks of neurons are treated as black-boxes without really analytical knowing whats going on. However, even without knowing the sense of detailed network connections, computation can be done with those black boxes[1]. Nevertheless, it might be unsatisfying to work with black boxes. Each single connection probability might be there for a reason in the end. Although hard to model and analyze, intuition tells it will be there for a reason.

Macroscopic and microscopic modeling

Macroscopic modeling can be a successful method to describe a complex system. Probably the best example of this approach is the ideal gas described by properties like volume, temperature and pressure without accounting for the trajectory of each single molecule. However, information which might have been coded in the detailed molecule distribution is lost. For a system like the brain, macroscopic modeling can help but is apparently not sufficient as information is coded by microscopic elements. Macroscopic properties can be activities of complete brain areas or global connectivity probabilities for instance.

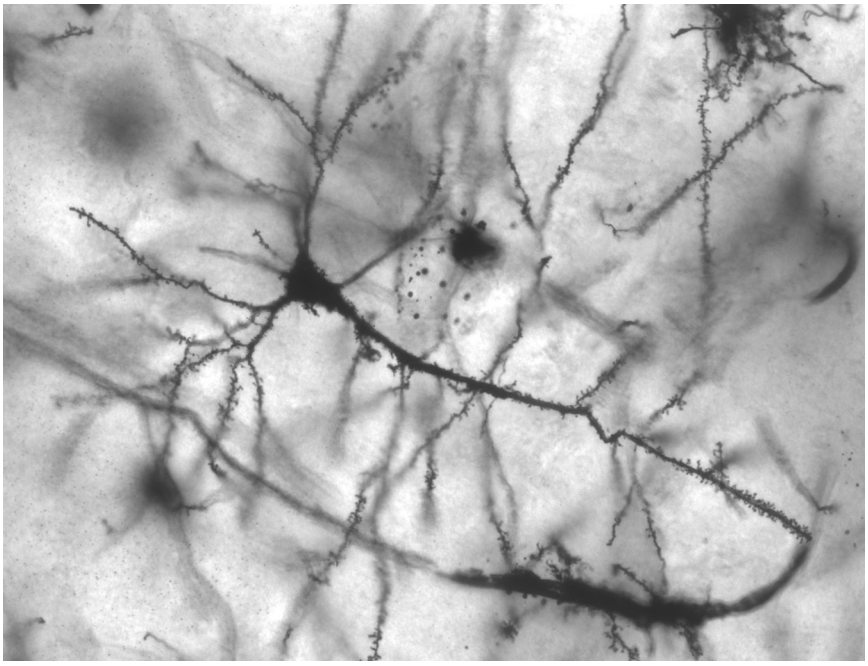


Figure 0.1: Pyramidal neuron from the hippocampus of an epileptic patient.(CreativeCommons Attribution-ShareAlike 2.5 licenced, created by MethoxyRoxy, from [2])

Single cells

This thesis is about microscopic modeling and emulation of single neurons, which are the

most important cells of the brain. A photograph of such a neuron cell can be found in Figure 0.1. Neurons themselves are complex structures indeed. This raises the question which complexity needs to be retained in a model of the cell. The question is rather similar to the macroscopic microscopic question. A simplified model might forget important features necessary for the function of a brain. A more detailed model on the other hand might enlarge the effort drastically while using potentially unnecessary information. In addition, information necessary for the creation of the complex model might not be available at-all. There is obviously a trade-off.

Point and compartmental-models

Two different models are used within this thesis. Using a point neuron model, the complete cell is assumed to be equipotential. Consequently, it is modeled as a single point. The second model assumes a relevance of the tree like structures of the neuron called dendrites. Here, the structure of the brain is modelled by the use of several interconnected so called compartments. Compartments are sections of the cell which are assumed equipotential.

Brains and computers

The brain is different to a digital computer indeed. A classic single-core of a micro processor sequentially computes in nano seconds. Data and instructions are fetched from a memory; a result is computed and written back into memory. The brain on the other hand computes massively parallel in micro seconds.

The total energy consumption of the Human Brain and a current digital computer chip is in the same order of magnitude. However, the actual computational power is hard to compare as both systems are best in different tasks.

Brains are more stable against errors.

Cutting a wire or adding one in a digital computer chip usually results in a wrong behaviour or a completely broken chip. In contrast, a single connections or cell from the brain will hardly cause any harm due to redundancy of the massively parallel system. In fact, dying cells or connections are very common in the brain. This tolerance against errors is a major advantage of the brain in comparison to a computer.

Saturation of single-core performance

In the last decade, Moore's law is saturating for single core performance due to the energy necessary for computation. However, the work-around is to use multi-core processors instead of single cores to achieve a continuous rise of computational power. Nevertheless, there will be limits for the classical van Neumann computer structure. A new kind of computer based on the structure and principles of the brain might be an alternative.

Construction plans

When building a computer, a complete schematic is necessary, to retain a certain working system. Computers are build as completed structures with the program and the memory status as its dynamical variables. The wiring of the brain is dynamic in contrast. There is no such thing as a complete construction plan. Connections are evolving. Learning occurs. There is no strict devision between hardware and software.

Brain like machines

When building a brain-like machines with dynamical connections, the complete function of the brain and its networks does not need to be known from the beginning as long as realistic learning mechanisms are implemented. Indeed, the machine can be a tool to understand the function of the brain. In fact, special hardware based on the structure of the brain can enable neuroscientific experiments which have not been possible so far due to a lack of computational power for computer simulations.

This thesis describes how to build neurons on a standard micro-chip using transistors. The approach is called emulation. Instead of simulating a model on a digital computer, the model dynamics are rebuild using analog circuits. The subject is called Neuromorphic Engineering[3].

Structure of this document

This dissertation is structured into 10 chapters. Here I will give a brief overview.

- **1 Neuroscience, Circuits and Neuromorphic Engineering:** This chapter gives an introduction into the interdisciplinary field of Neuromorphic Engineering. I start with a brief introduction into neurons, synapses, plasticity. Neuron models are introduced next. Subsequently a contrast is given by micro-chips and the description of transistors. After these basics, the step to an emulation and Neuromorphic Engineering is done. I conclude the chapter with the discussion of different design approaches and implementations from literature. This chapter is very important for nomenclature definitions.

- **2 Neuromorphic Environment:** I give an introduction into the system, the presented circuits are nested in. A top-down approach is taken during presentation. The HICANN microchip is introduced here for instance.

- **3 Point Neuron Emulation:** Starting with the introduction of the implemented model, now each single circuit part of the point-neuron implementation is described and discussed.

- **4 Point Neuron Experiments:** This chapter describes some experiments performed with the point-neuron circuit. I present a basic benchmark experiment and some typical patters produced by the circuit for instance. A focus is laid on the analysis of fixed pattern noise. The chapter is concluded by a small network experiment.

- **5 Discussion: Single-Compartment**

- **6 Multi-Compartment emulation:** At the beginning of this chapter, the concept of compartmental modeling and the model choice are discussed. Furthermore, implementations from literature are presented. Subsequently, the changes necessary for a multi-compartment implementation are presented first in an overview and in detail next.

- **7 Multi-Compartment Experiments:** Two small experiments are presented as simulations. The first experiments covers passive attenuation of dendritic stimulation. Subsequently, action-potential generation and propagation in the dendrite is observed.

- **8 Discussion: Multi-Compartment**

- **9 Analog Floating-Gate Memory** The immense parameterizability of the designed neuron circuits is only possible due to the use of analog floating-gate memory. These memory cells and the necessary periphery circuits are presented here. The chapter is concluded with some measurement results.

- **10 Conclusion and Outlook**

1 Neuroscience, Circuits and Neuromorphic Engineering

This chapter provides the foundation to the subjects of this thesis. I start with a very brief introduction of the brain and neurons in particular as its main components. Subsequently, I introduce the Hodgkin and Huxley neuron model followed by the leaky Integrate-and-Fire model which is enhanced to reproduce real neurons' behaviour. The next section gives background on the main device of a microchip which is the transistor. The concept of emulation is an alternative procedure to the analytical solution or the simulation of a system. It is introduced next. In the same section, I introduce the work field of Neuromorphic Engineering. At last, to get an overview about the status in literature, design approach considerations and different neuron implementations are discussed.

1.1 Biological Substrate

The understanding of the brain and its computational powers is is one of the main topics of the work presented in this thesis. However, what is the brain? On an abstract level, the brain is a collection of specialized cells called neurons which are interconnected by synapses to form a large and complex network.

Information exchange

Information is interchanged by electro-chemical signals called action potentials or spikes which are the binary time continuous output signal of a single neuron. The brain is electric. Neurons are the basic internal information processing units of the brain. Three dimensions can be used for the positioning of individual neurons and the routing of connections[1].

Information processing

Each neuron has an internal membrane voltage V created by differences in ion concentration inside and outside of the cell. This voltage is referred to as membrane potential. There can be differences along the membrane surface. When a neuron receives an action potential from another neuron via one of its synapses, channels are opened allowing ions to flux into the neuron. These ions raise or lower the membrane potential. If the voltage of a neuron reaches a certain threshold, it creates an own action potential which is transported to other neurons. An action potential is a large transient membrane voltage spike.

This is a very brief conceptional introduction only. For details see [4] for instance.

1.1.1 Neurons

A "closed" system

Neurons are cells. Cells are basically closed units surrounded by a lipid bi-layer. Lipids are molecules with a hydrophilic and a hydrophobic end which tend to build bi-layers with the hydrophobic end in the middle when exposed to water. This lipid bi-layer constitutes a membrane that divides the interior and the exterior of a cell. A lipid bi-layer alone close to an insulator. It can be modeled as a capacitor. However in biological cells, a variety of channels is present, which allow for passive and active transport through the cell.

[1] Indeed, the cortex is organised in layers which can be thought as two dimensional. However, these layers as well as neurons are spacial objects.

Ions and reversal potential

Inside and outside of the cell in the brain are ions. The most important ones for the work presented here are potassium (K^+), sodium (Na^+), calcium (Ca^{2+}) and chlorine (Cl-). The concentration of the these ions inside and outside of the cell is different. However, this concentration gradient tends balanced through open ion conducting channels. The balancing creates a voltage difference between the interior and the exterior due to the ion charge. Concentration gradients of ions are counterbalanced by this voltage differences. In equilibrium the dedicated voltage for an ion type is called Nernst Potential or reversal potential. A typical value for the reversal potential of the potassium channel of a cell is -89 mV [4] for instance.

Sodium concentration gradient

Due to large charged macro molecules within the cell the ion concentration inside the cell tends to be larger than outside the cell. However, a larger ion concentration would cause the water influx by osmosis. If no mechanisms where to counterbalance this this influx the cell could burst[4]. Nevertheless, we know that it does not burst. In animal cells the main mechanism are special active channels constantly pumping sodium ions out of the cell. Consequently, there is a large sodium concentration gradient between the exterior and the interior of the cell and hence a large tendency of sodium to enter the cell. The reversal potential of sodium is 50 mV.

Generating action potentials

For neurons, this omnipresent gradient is very important. Synapses can mediate Ion flux which changes the membrane potential. If the membrane voltage reaches a certain threshold voltage[2], special voltage-gated channels for sodium ions open and the membrane voltage sharply rises. Action-potentials are created in regions of high channel densities. The rise of the action potential is propagated avalanche. However, after a certain delay, the sodium channels close again and slower voltage-gates potassium channels open pulling the membrane down again. For a certain time, no further action potentials are possible as the equilibrium concentrations have to be maintained. This time span is called absolute refractory period. Due to their shape, action potentials are also called spikes. The creation of an action potential is referred to as spiking or firing. A spiking neuron fires spikes respectively action potentials.

Spacial structure

So far, our neuron did not have a special spacial shape. An exemplary schematic of a neuron can be found in Figure 1.1. Synapses from other neurons usually connect at the dendrites which are the tree like structures growing out of the cell body. The cell body is called soma. The axon hillock is the region with the highest concentration of voltage-gated channels. Action potentials are usually initiated in this region. They are propagated along the axon inducing ion influx into other neurons via synapses. A smaller action potential might also be back propagated into the dendrites. The axon can be surrounded by myelin sheath which accelerate the propagation by locally insulating the membrane. The section between the myelin sheaths is called node of Ranvier and is a region of high active channel density restoring the attenuated action potential.

1.1.2 Synapses

Synapses are the interface between neuron. There are electrical synapses directly interconnecting membranes and chemical synapses using molecules called neuron transmitter as mediator. Here I will concentrate on chemical synapses however.

Neurotransmitter

The synapse at the end of the presynaptic neuron reaches close to the membrane of the postsynaptic neuron. The space separating the two neurons is called synaptic cleft (See Figure 1.1). When an action potential arrives at the synapse, Ca^{2+} enters the synapse at the postsynaptic neuron. So called vesicles carrying neurotransmitters are Ca^{2+} triggered to combine with the postsynaptic neuron's membrane and release the neurotransmitters into

[2] The threshold can be variable.

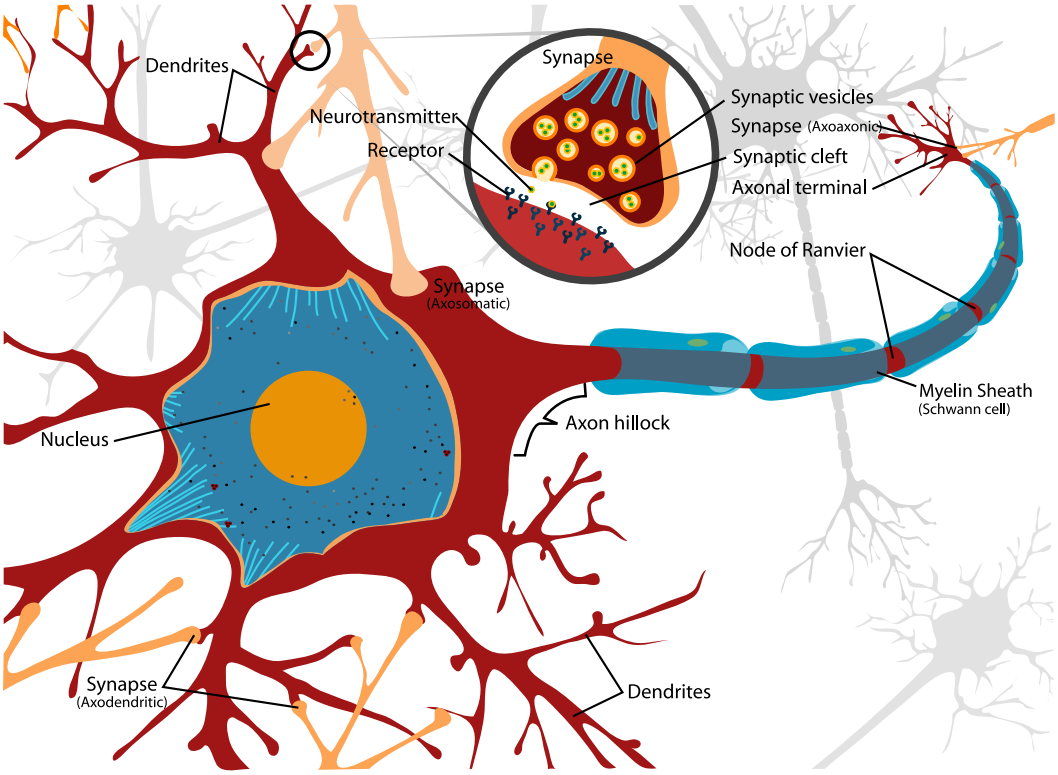


Figure 1.1: Schematic of a neuron. (Reduced version of public domain figure from [2])

the synaptic cleft.

Transmitter-gated channels

In the synaptic cleft at the postsynaptic neurons site are transmitter-gated-channels in a high density. These channels open a conductance if a neurotransmitter molecule is received. The conductance is selective for different types of ions and can inhibit or excite the postsynaptic neuron. Excitatory synapses mainly conduct Na^+ for instance while inhibitory synapses conduct Cl- for instance. The voltage response of the postsynaptic neuron is referred to as postsynaptic potential(PSP).

Different synapses

The neurotransmitter GABA[3] is the main transmitter for fast inhibitory synapses [5]. The conductance can be described by a rapid opening conductance which is exponentially decaying with a time constant of 5 ms. Excitatory transmitters are NMDA[4] and AMPA[5] for instance. AMPA synapses can open a conductance with a sharp exponential rise with a time constant below 10 µs which is followed by a decay with 1.5 ms [5]. The behavior of NMDA synapses is more complicated however as they are membrane voltage dependent. They are stronger for higher membrane potentials [5]. A role of NMDA synapses can be the amplification of synaptic signals in apical dendrites [6].

A synapse's capability of changing the membrane potential of the presynaptic neuron is referred to as synaptic efficacy, or strength.

[3] γ-aminobutyric acid
[4] N-methyl-D-aspartate
[5] α-amino-3-hydroxy-5-methyl-4-isoxalone propionic acid

1.1.3 Plasticity

In fact, the synaptic connections and even the morphology of the neuron is not static. This temporal change is referred to as plasticity. Plasticity mechanisms can change the synaptic efficacy – which is a measure of the effect of a synapse on the postsynaptic neuron. Furthermore, new synapses are constantly created while other synaptic connections disappear. Here I will discuss the phenomenology of plasticity occurring on short time scales (hundreds of microseconds) [7, 8] and spike-timing dependent plasticity (STDP).

Short term plasticity

The temporary change of synaptic efficacy depending on the previous reception of action potentials is called short term plasticity. Typically, the synaptic efficacy is increased or decreased at each single spike. If no further action potentials are received, the efficacy returns to a steady state efficacy. These mechanisms can for instance amplify low frequency stimulus or attenuate high frequency stimulus [7, 8].

Hebbian Learning

A very intuitive kind of plasticity is given by Hebbian Learning [9]. If a synapse is responsible for the firing of the postsynaptic neuron, its efficacy is increased. On the other hand, if the postsynaptic neuron fires acausally, meaning without any influence of the corresponding synapse, the efficacy is decreased.

STDP

A Hebbian Learning mechanism with real neurons has been confirmed by Bi and Poo [10]. It is called spike-timing dependent plasticity(STDP). The change of synaptic efficacy is weighted with an exponential function depending on the timing difference of the post- and presynaptic spikes. Causal events result in a strengthening of the synapse. In addition, beyond Hebbian Learning, acausal events decrease its efficacy.

1.1.4 Typical Cortical Neuron Behavior

Here, I will discuss some typical neuron behaviors which are often referred in this thesis. In particular spike-frequency adaptation and bursting and derived behaviors. Indeed, there are many more possible patterns. For a comprehensive collection see [11] or Figure 4.11 in this work. The nomenclature follows [11] and [12].

Spike-frequency adaptation

When a neuron is stimulated by an adequate current, it starts creating action potentials. After a short delay, another action potential will be created. This delay however can be adapted. At the beginning, the neuron spikes with a higher frequency which decreases with each additional spike until an equilibrium is reached. The frequency is adapted. This effect is called spike-frequency adaptation. An extreme case of spike-frequency adaptation would be the creation of one single spike at the beginning an no further action potentials. This effect is called phasic spiking. A neuron spiking with a regular frequency is called tonic spiking.

Bursting

Some neurons tend to fire small groups of spikes with a high frequency. These groups are called bursts. The corresponding behavior is referred to as bursting. If bursts are created with a constant frequency, it is called tonic bursting. A single burst is a phasic burst. When bursts and single spikes are created, the behavior is called mixed-mode.

Bursting as well as single spikes can occur with a delay after the onset of stimulus.

1.1.5 Measurement Capabilities

In vivo, in vitro

Measurements can generally be done *in vivo*, or *in vitro*. An *in vivo* measurement is carried out in an living animal. *In vitro* experiments are done with insolated cell cultures in a test-tube for instance. While the access to neurons with a defined stimulus is not possible *in vivo* by virtue of network activity , *in vitro* measurements miss the surrounding input of the network. When working *in vivo*, the model animals have to be anesthetized, which has an effect on the observed neurons.

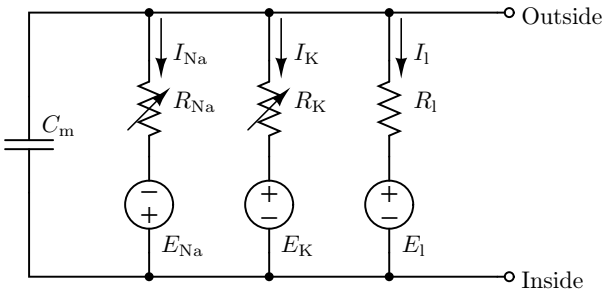


Figure 1.2: Schematic of the Hodgkin Huxley neuron model[19].

Some methods

Several neurons can be measured *in vivo* using arrays of electrodes [13]. However, the measured neurons a random sample of all neurons. Detailed recordings and stimulations *in vivo* and *in vitro* are possible using the patch-clamp method [14] for instance. Furthermore, special molecules reflecting the membrane voltage in their florescence behaviour when excited can be used to measure the mean voltage of groups of neurons or even local voltages in dendrites[15].

Simulation

A common approach to understand single cell or defined network behaviour is carried out using simulations in a neural simulator like the simulators NEST[16] or Neuron [17] .

1.1.6 High-Conductance State

The high conductance state, as introduced in [18] for instance, is the state of a neuron receiving large amounts of synaptic input from other neurons. It is the suspected typical state of neurons measured *in vivo* due to the high number of connections from other neurons. A contrast is given by a neuron measures *in vitro*, where surrounding network activity is not given. The high synaptic input opens conductances enlarging the total conductances of the neuron. Consequently the nomenclature is high-conductance state.

1.2 Single Cell Modeling

Here I discus different neuron models. Starting with the complex mother of all neuron models - the Hodgkin Huxley Model[19], the simple phenomenological Integrate-and-Fire model is presented and enhanced by adaptation and positive feedback.

1.2.1 The Hodgkin Huxley Model

The Hodgkin Huxley Model (HHM), published by Alan Lloyd Hodgkin and Andrew F. Huxley in [19], is a is a model of the squid giant axon. The size of this axon allows for better experimental access. In particular, it has a diameter between 0.5 mm and 1 mm and a length of several centimeters[4].

Observed subject

In their experiments, Hodgkin and Huxley cut the axon from the cell body and retain an axon tube. To ensure a defined experimental setup the plasma inside the tube is replaced by a defined solution. The outside solution is usually sea water. This way ion concentration dependency measurements are possible[4].

To measure current and voltage dependencies, a long electrode is placed inside the axon. Additional electrodes are placed outside the axon.

Schematic

A schematic of the resulting model is presented in Figure 1.2. The I_l branch is the passive leakage current which is ascribed to chloride and other ions in [19]. The adjustable conductances are actually voltage dependent conductances with a complex temporal deviation. Action potentials are created by these channels as described in Section 1.1. The description of the voltage gated Sodium and Potassium channels is the major accomplishment of the HHM.

Hodgkin and Huxley describe the conductances using an approach of so-called gating variables n, m, and h. The resulting conductances are:

$$g_{\mathrm{K}} = \bar{g}_{\mathrm{K}} n^4, \text{ and} \tag{1.1}$$

$$g_{\mathrm{Na}} = \bar{g}_{\mathrm{Na}} m^3 h. \tag{1.2}$$

Gating variables

The gating variables and there exponentiation can be understood as different conditions necessary for an open channel. They are time varying and described by differential equations according to

$$\frac{dn}{dt} = \alpha_n(1-n) - \beta_n n. \tag{1.3}$$

The alphas and betas are voltages dependent but time independent properties. They are determined by measurements.

Action potential generation

The gating variable h is one for small membrane voltages and decreases at about 50 mV. m on the other hand rises for voltages above -25 mV and reaches one close to 50 mV. However, the key is the temporal deviation. h reaches its final value much slower than m. Nevertheless, in an action potential, h blocks sodium influx enabled by m after a short time. The actual temporal course is complicated as α and β are voltage dependent.

n opens for large membrane voltages with a slow time constant and pulls down the membrane back to the resting potential.

Limitations

The HHM describes axons of special neurons of squids. When analysing cortical neurons from the human brain, further channels have to be added for realistic modeling. In particular, calcium channels and calcium-concentration-dependent potassium channels can be important.[5].

1.2.2 The Integrate-and-Fire Model

HHM not optimal for simulations

Each gating variable adds another differential equation to the HHM. It is neither optimized for computer simulations, nor for model analyses. In [5], the authors describe how to retain the HHM dynamics using only two variables. I skip this step and directly come to the reduced phenomenological leaky Integrate-and-Fire neuron model (lIaF) [20] which is a drastic simplification.

Removing active channels

Basically, this model assumes that all spikes are equal. Consequently, there is no information in the shape of a spike and it can be omitted. In the lIaF, the potassium and the sodium channels have been removed from the HHM retaining only the leakage conductance. The result is a leaky integrator – a capacitor with a conductance to a leakage potential connected in parallel. It can be described by a single dynamic variable - the membrane voltage:

$$C_m \frac{dV}{dt} = g_l(E_l - V) + I. \tag{1.4}$$

Reseting

However, there is still the stimulus current. If the membrane voltage crosses a threshold Θ it is reset to a certain reset potential V_{reset}, which is usually equal to the leakage potential. Thus, the membrane time course is not continuous.

1.2.3 Adaptation

Calcium channels

High-threshold voltage-gated calcium channels in cortical neurons allow the influx of Ca^{2+} into the neuron. In addition, there are low-threshold voltage-gated calcium channels allowing calcium influx at smaller voltages. Those can create a pull-up of the membrane if inhibiting input is removed[5].

Calcium modulated K^+

Calcium concentration in a neuron is low and strongly interfered by calcium influx, while global sodium and potassium concentration stays nearly constant. Furthermore, there are calcium-concentration-gated K^+ channels [5]. Consequently, rising calcium concentration can induce a negative feedback onto the rising membrane potential.

This results in effects like spike-frequency-adaptation. When stimulated with a current pulse, the spike frequency is lowered with each action potential due to the enlarged calcium concentration.

A second variable

To account for this effect, a second variable can be added to the lIaF. The adaptation variable, or slow variable w [21, 22]. The new pair of equations is:

$$C_m \frac{dV}{dt} = g_l(E_l - V) + w + I \tag{1.5}$$

$$\tau_w \frac{dw}{dt} = a(E_l - V) - w. \tag{1.6}$$

In the steady state, w add another conductance a to the leakage potential. However, if the membrane is released from a lower potential very fast, the adaptation variable might not be able to follow due to its large time constant τ_w (Its magnitude is 100 µs.). This way, the membrane reacts different to fast changing signals which is a behavior observed in cortical neurons [11]. Another effect caused this way is called inhibitory rebound. The removal of an inhibitory signal can cause an action potential. In addition to subthreshold effects, w needs to be enlarged at every action potential – this results in spike-frequency adaptation.

Phenomenological approach

However, the addition of the adaptation variable is a phenomenological approach. It enhances the capabilities of the lIaF model to reproduce the behaviour of cortical neurons. Nevertheless, the direct biological correspondence with the calcium concentration is weak. In [21] it is referred to account for activation of K^+ and the deactivation of Na^+ channels.

1.2.4 Positive Feedback

Re-adding Na^+-channels

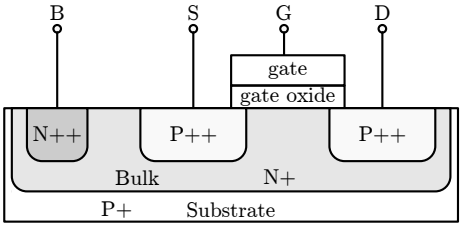
Due to the removal of the voltage-gated channels, the membrane dynamics of the lIaF model cannot produce real spikes. Spikes might look similar. However, the onset of the spike can change the neurons dynamics completely as positive feedback is added by the voltage-gated sodium channels. The positive feedback can amplify stimulus close to the threshold for instance. In addition, the spiking threshold is more smooth this way.

Izhikevich used a quadratic feedback of the membrane voltage for instance to improve the capabilities of his model[21].

1.3 Technical Substrate

Microchips

A microchip is the technical substrate of this thesis. Microchips are semi-conductor circuits. Indeed, microchips are apparently electric. Their main computational elements are transistors which have a far less complexity and function in comparison to a neuron. In addition to the active devices, components like fixed resistors or capacitors are available for circuit



(a) Cross section of a PMOS transistor

Figure 1.3: Cross-section of a PMOS transistor. P+, P++, N+, and N++ denote positive respectively negative doping concentrations of different levels.

implementation. Capacitors can be implemented most precisely using two opposing metal plates for instance.

Production Usually microchips are produced on silicon wafers which are cut into the individual chips. These wafers are referred to as substrate. Devices are created using doting, evaporating and etching technologies. Fine structures can be produced using lithography techniques. Although, several different layers exist on a microchip, the basic structure of chips is usually two dimensional. Each device on a microchip is different as transistors are real physical devices exposed to production variations. These variations are called fixed-pattern noise.

ASIC The microchips used within this thesis are Application Specific Integrated Circuits (ASIC). These chips usually produced in a small small volume to solve special problems. For prototyping, our chips are produced with other designs on a single silicon wafer to save production cost. This design approach is called Multi-Project Wafer (MPW) prototyping.

Here I will give a very short introduction on transistors and the schematic nomenclature and some basic circuits used within this dissertation. Basic common circuits like logic gates, transmission gates are assumed as known and not introduced hence. For further introductions into circuits see e.g. [23–26].

1.3.1 Transistors

MOSFET The transistors discussed here and used in the circuits presented in this thesis are metal-oxide-semiconductor field-effect transistors (MOSFET). There are two basic devices PMOS and NMOS (p respectively n-type metal-oxide-semiconductor device). The beginning letter denotes the charge of the conducting charge carriers in the corresponding device.

MOSFET structure The cross-section of a p-type MOSFET in a positive doted substrate is presented in Figure 1.3. Current flux between source (S) and drain (D) is maintained by the Gate (G) potential. The bulk contact (B) can be ignored for now. The isolating gate oxide is usually very thin (several atom layers), so there is a strong capacitive coupling between the gate and the area below the gate. The area below the gate oxide is called channel area. In the N areas, the charge carriers are given by additional electrons of the doting atoms which do not fit into the structure of the silicon. On the other hand, missing electrons are the positive virtual charge carriers in the P areas. They are called holes.

PN-junctions At first, I describe the case where, S, G and D are at the same voltage level. There are two opposing PN-junctions between source and drain. At these junctions, electrons from the negative doping atoms of the N+ area diffuse into P++ area and recombine with the atoms used for P++ doping. This way, positively charged ions are left in the N+ area while negatively charged ions are left in the P++ region. The ionized areas are called depletion zone. There is a barrier voltage between the ionized areas.

Diffusion

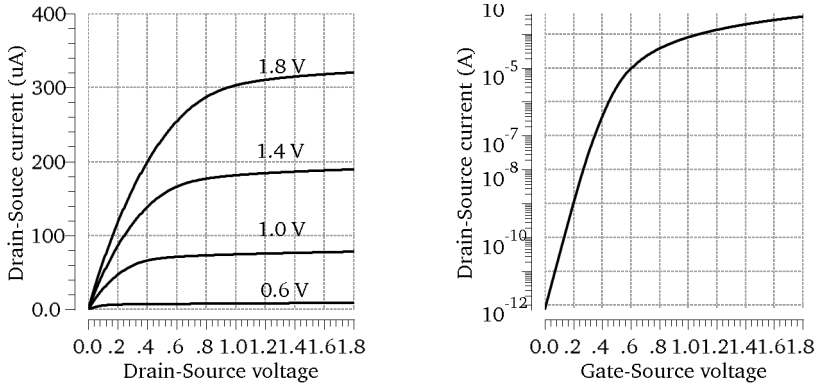
Now, we set the drain potential D at a negative voltage e.g. -1.8 V. The depletion zone at the drain junction grows and blocks electrons from D from entering the channel area. In addition, the depletion zone is enlarged. However, some electrons have enough energy to diffuse into the channel. Concentration gradients enhance the diffusion process in the channel. Electrons are pulled to the source by the source junction. A very small diffusion current can flux.

Subthreshold

Adding a small negative voltage at the gate created a depletion zone below the gate. The ionized atoms in the depletion zone balance the negative gate voltage. Below a certain threshold voltage V_t, the diffusion current in the channel rises exponentially with the falling gate potential:

$$I_{\mathrm{DS}} = I_0 e^{\frac{-\kappa V_{GS}}{u_t}}. \tag{1.7}$$

Here I_0 is a constant current, u_t is the thermal voltage and κ is the subthreshold slope factor. I_{DS} is the current between the terminals D and S and V_{GS} is the voltage between the terminals G and S. For a detailed derivation see [24]. The exponential rise of the drain-source current of an NMOS transistor below the threshold voltage can be seen in Figure 1.4 b)



(a) V_{DS} characteristic for different values of V_{GS}

(b) V_{GS} characteristic for different values of $V_{DS} = 1.8$ V

Figure 1.4: Simulation of an NMOS transistor.

Inversion

For negative gate voltages below the threshold voltage, inversion occurs. The positive charges of the ionized doting atoms cannot compensate the gate potential anymore. Although, the charge carrier of negative doted silicon are electrons, free holes become available in the inversion layer. Now, current can be conducted by drifting holes in the channel now. The current is roughly proportional to the square of the gate-source voltage. This operating region is called strong inversion. It is the usual stable operation region of a MOSFET.

The bulk potential

The contact B in Figure 1.3 is the connection to the bulk of the transistor. The bulk potential influences the effective threshold voltage drastically. This issue is called body effect. In some analog applications the bulk potential needs to be maintained to reduce the body effect. In Figure 1.3, the bulk is an n-well which is a negative doping area on the p-doted substrate. N-wells are used for the creation of PMOS devices. Several PMOS devices can share an n-well to allow a more compact design.

NMOS bulk

N-mos devices are usually created directly on the positive doted substrate. Consequently, they share a bulk potential. However, special devices using an isolated p-well inside an n-well are available but area inefficient.

V_{DS} characteristic

So far, the drain-source voltage has been kept constant. I will discus the drain-source voltage characteristic of a MOSFET now. It is shown in Figure 1.4 b).

Ohmic region

For absolute drain source voltages below $|V_{GS} - V_t|$, the velocity of the moving holes in the PMOS respectively electrons in the NMOS transistor can still be increased with rising drain source voltage. In this region, called ohmic region or triode region, the drain-source current is roughly proportional to the drain-source voltage:

$$I_{DS} = K' \frac{W}{L} \left((V_{GS} - V_t) V_{DS} - \frac{V_{DS}^2}{2} \right) \tag{1.8}$$

See [26] for instance. K' is a process parameter, W is the channel width, and L is the length of the channel. However, in fact there is a smooth transition to the saturation region which comes next. A linear dependency can only be assumed for small values of V_{GS}.

Saturation region

In the saturation region, the current can hardly be enlarged by larger drain-source voltages as interaction with the fixed atoms decelerated the charge carriers. However, larger voltage can shrink the channel itself and cause a small increase of the current this way. This effect is called channel length modulation. In the saturation region, the current can be described by:

$$I_{DS} = \frac{K'}{2} \frac{W}{L} (V_{GS} - V_t)^2 (1 - \lambda V_{DS}) . \tag{1.9}$$

For further details, compare [26]. λ is called channel length modulation parameter. For small transistors channel length modulation has a drastic influence. Due to the flat characteristic in saturation, a transistor biased in this region can be used as a current source.

Nomenclature

The used schematic symbols for MOSFET devices are shown in Figure 1.5. When they are used, the bulk connection is omitted in most cases however. If no bulk connection is shown, the bulk of NMOS devices is connected to the ground potential while the bulk of PMOS devices is connected to the corresponding power supply.

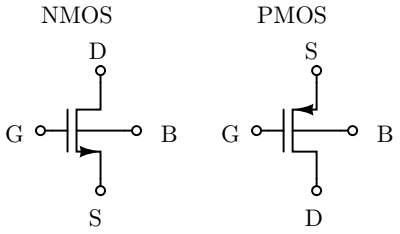


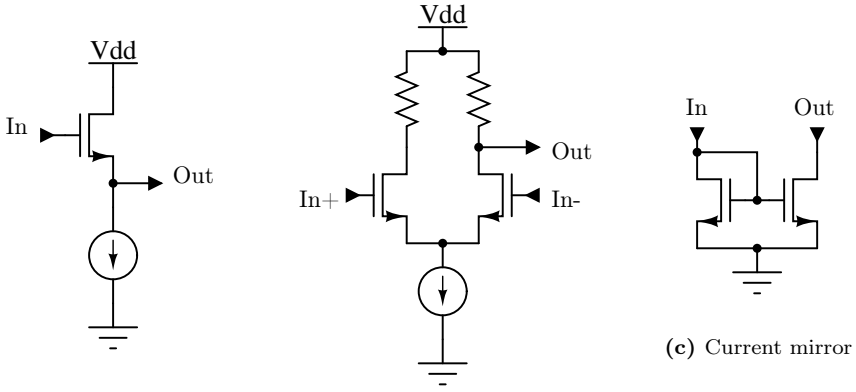
Figure 1.5: MOSFET schematic symbols

1.3.2 Basic Transistor Circuits

Here, I will briefly introduce source followers, differential pairs and current mirrors in the following. The circuits are shown in Figure 1.6.

Source follower

A source follower consists of a single transistor and a current source, which is usually implemented by an additional transistor operated in strong inversion. A schematic is presented in Figure 1.6 a). The output voltage follows the input voltage as the gate source voltage is mainly defined by the cross current of the transistor. An application of the circuit is in impedance converter. The input voltage might be driven by a weak driver which is amplified by the circuit to drive larger loads. In this thesis, the circuit is used for the creation of biasing voltages in addition.



(a) Source follower

(b) Differential pair

(c) Current mirror

Figure 1.6: Basic transistor circuits

Differential pair

The next circuit is a differential pair (Figure 1.6 b). Differential input voltages are amplified. When both input voltages are equal, half of the current from the current source will flow through each of the two transistors causing a voltage drop at the resistors. Subsequently, if a differential voltage is applied, less current will flow through one branch, while the same amount more current will flow through the other. For small differential voltages, the input voltage difference is linearly amplified. Indeed, the total current is limited by the current supplied by the current source. Furthermore, the characteristic of a MOSFET is not linear. Consequently, the linear range is limited. A differential pair will be discussed in more detail in Chapter 3.

Current mirror

A current mirror is an essential circuit in analog circuit design. The circuit can be found in Figure 1.6 c). It mirrors an input current to the output. A single mirror can have several outputs to create further incarnations of a biasing current. The gate-source voltage of both transistors is given by the input transistors. Consequently, if both transistors are equal and without channel length modulation, the output current matches the input current.

1.3.3 Simulation Techniques

To analyze the behavior of a circuit before the production of a microchip, simulations are carried out using a circuit simulator. Here I give a very brief overview on the simulation techniques of analog devices. All circuit simulations shown in this thesis have been carried out with the SPICE [27] simulator spectre [28].

DC simulation

The most basic simulation is a direct current (DC) simulation. For a DC analysis, components creating time-dependent signals like capacitors or inductors are removed from the circuit. Next, at a constrained parameter set, the circuit's equations according to Kirchhoff's laws are composed. Where necessary, the equations are linearized or numerically simplified. Finally, the remaining system of equations can be solved by any common method.

Sweeping parameters

However, so far there is only a single solution for a set of parameters. This solution is called DC operating point and is the initial starting point of each simulation. To investigate dependencies of different quantities, single parameters can be swept in a DC-simulation. In the DC simulation executed for Figure 1.4 the gate-source voltage respectively the drain-source voltage are swept for instance. DC-simulations are carried out for most characterizations in

this work.

Transient simulation

A more complex simulation technique is a transient simulation. Starting with the DC operating point, the system is evolved in time. Therefore, the system is linearized in time by differentiating time dependent properties. The subsequent state of the system is calculated assuming a certain time step. If the changes are too large, a smaller time step is chosen dynamically. A transient simulation of a fast spiking neuron is much more time consuming than the simulation of a silent membrane due to time step adaptation.

AC simulation

Another important simulation approach is an alternating current (AC) simulation. At the DC-operating point, the frequency dependency of the system is analyzed with a linearized version of the circuit. This simulation technique could also be referred to as small-signal simulation, as only the first derivative around the operating point is used. Consequently, the solution is only correct for small changes.

Monte-Carlo simulation

Monte-Carlo simulation has become available during the course of this thesis. In a Monte-Carlo simulation circuit parameters are randomly changed according to measured deviations supplied by the chip producer. Several (hundreds) Monte-Carlo simulations are carried out to gain statistics on circuit behavior. This way, variation of circuits can be taken into account.

Example: miss-match of a single transistor

Sampling the curve of Figure 1.4 b) at 200 mV, which is below the threshold results in a relative standard deviation of 20% of the drain-source current. However, sampling at 1 V, which is above the threshold voltage results in 1%. Nevertheless, the transistor has been dimensioned large which causes less fixed-pattern noise. Repeating the same simulation with a minimums size transistor results in standard deviations of 50% respectively 5%.

1.4 Rebuilding Biology - Neuromorphic Engineering

Limits of simulations

When complex systems are to be understood, simulation approaches can reach their limit depending on the available computational resources. Smaller networks with a low detail level can be simulated on standard computers. When network size [29] or the complexity of the model are enlarged [30], however, super computer architecture or special simulation hardware [31] become necessary. Introducing local plasticity mechanisms drastically enlarges the communication effort as each synaptic connection needs to have information about the post and the presynaptic neuron. Scaling of the necessary power consumption with the network complexity can completely inhibit biologically realistic simulations of larger cortical areas or even the complete human brain.

Nevertheless beyond the concept of simulation, an emulation, described next, can bride the gab.

1.4.1 Emulation

Concept

The concept of emulation relies on rebuilding a system's dynamics instead of simulating them. When a physicist wants to understand how a dynamical system behaves, he develops a model of the system. To comprehend this model and to make predictions, in the best case, the model can be solved analytically. However, a simulation can be required for complex systems. Emulation is an alternative approach to simulation. Instead of simulating the model, the dynamics of the model are rebuilt in a physical model. Consequently the dynamics are not analytically ore numerically solved but the system behaves according to them. The model's results can be read out by measurements. The concept of emulation is often referred to as analog computing [32]. A nice aspect of an emulation is the real existence of a physical incarnation of the model.

Newton's apple

The the concept of emulation can be explained by a nice example with Newton and an apple. Newton sees an apple hanging on a tree and wants to know the time it takes to drop on the ground once released. He could do a calculation. A computer simulation can surly be excluded here. However, he could pick another fallen apple from the ground, lift it to the height of the apple under observation and let it drop. The measured fall time is the solution given by the physical model or the emulation.

A renaissance

Indeed, using electronics for emulation is an old concept. In [32], John R. Ragazzini et al. introduce the operational amplifier as a basic module for analog computation for instance. The application is analog computation of airplane dynamics by rebuilding differential equations. However, due to the enormous increase of digital computational resources in the last decades, analog computation of smaller systems based on few equations nearly died out. Nevertheless, in parallel systems like neural networks in combination with modern microchip production, emulation is having a renaissance [3].

Electronic neurons

Looking at a neuron, an emulation approach can be directly followed as models are based on the use of electronic components. The membrane capacitance, for instance, can be implemented by a real capacitor. Passive channels, like the potassium leakage channel can be build using a real resistor and a voltage source. An Integrate-and-Fire neuron has been implemented. More effort has to be spent, however, when it comes to active channels.

Very Large Scale Integration (VLSI) techniques allow the integration of large numbers of neurons on a microchip. This way, a large amount of differential equations can be solved continuously in parallel.

Property miss-match

However, VLSI techniques have been developed to implement computer chips or analog devices like amplifiers and receivers. Consequently, there is a large gap between the available conductances and capacitances in VLSI and biology, if the emulation is done in real time and the electronic devices are operated in their destined regimes.

Real-time neurons

Nevertheless, real-time emulation is possible by leaving the "secure" operation regime of strong inversion and moving over to subthreshold implementations. The great advantage is a greatly reduced power consumption as smaller conductances mean smaller currents. Furthermore, depending on the desired level of model accuracy, complex designs are possible. On the other hand, the fixed-pattern noise effects are worse in this regime and a large variation between individual components is to be expected. In addition, the sensitivity to cross-talk between individual components and from digital circuit on the chip is much larger according to a higher gate voltage change sensitivity in the subthreshold regime.

Scaled time neurons

Another approach, is to operate the electronic devices of VLSI in their destined operating regimes while scaling the model parameters. This design is carried out in the neurons presented in this thesis. Take a simple Integrate-and-Fire neuron model for instance. With a membrane capacitance of 200 pF and a membrane conductance of 20 nS the time constant of the R-C circuit would be 10 ms. A realistic hardware capacitance would be 2 pF with a conductance of 2 µS. This results in an emulation time constant 1 µs. Consequently, the dynamics of the emulated model are 10^4 times faster than real time. The factor is called acceleration factor.

Communication gap

However, the drawback of a so-called accelerated model its inability to communicate with real time systems. Hence, real world applications e.g. sensors in robots are not directly implementable for instance. In addition, the communication bandwidth needs to be scaled with the acceleration factor

Less fixed-pattern noise

Nevertheless, the advantages of an accelerated implementation outbalance the issues. An implementation using the transistors in the destined strong inversion region is much more stable to fixed pattern noise and cross-talk. This way a much more reliable model is possible and digital circuits with high clock speeds can be used for control structures.

More computational power

Most importantly, the acceleration factor is a real feature indeed. As the emulation is

accelerated, its computation speed is accelerated equally. The emulation of long long as needed for the evaluation of learning mechanisms is possible. Single experiments can be carried out several time in a loop. In addition, the energy used for a single action potential is reduced this way, as the total system power can be scaled with the time scaling factor.

1.4.2 Sensors, Neurons and Systems

Carver Mead

Using electronic circuits (usually VLSI) to emulate parts of the nervous system like neurons or sensors is called neuromorphic engineering. The concept has been given birth by Carver Mead [3, 33, 34].

Neurons and Sensors

With great success, sensors like silicon retinas [35] have been developed. Cochleas are presented in [36] and [37] for example. Several different neurons have been designed. Some examples can be found in [38–45]. A review of different neuron implementations is given in [46]. Different implementations will be discussed in the next section.

Communication

Communication can be accomplished through digital representations of action potentials. Most implementation use a real time spike propagation mechanism called AER[6] (See [47]). Each digital event representing an action potential is an address corresponding to the source or destination neuron of the action potential. Usually no time-stamp, i.e. coding the time of action potential initiation, is implemented. Signals from several neurons share one bus. The AER protocol is a standard in the community and allows the interconnection of different devices like retinas and neuron chips for instance [48].

The spike propagation mechanisms of the systems presented in this thesis are similar for low level communication. However, a more complex AER protocol is used for higher level communication [49].

Systems

A single neuron does not build a network. Indeed, usually several silicon neurons are embedded into a chip to build networks. However, there is a great difference from the scalability point of view. The reduction of power consumption at single neuron level can be meaningless, if all synaptic connections have to be routed through an FPGA[7]. Larger plastic networks need large amounts of plastic synapses for each single neuron.

In [48] Indiveri describes how to construct systems using several individual chips based on real-time AER. An accelerated system is presented in [40]. The system the neurons presented in this thesis are integrated in, is shown in [50]. It will be introduced in Chapter 2.

1.5 Single Cell Emulation

Here I will first talk about different design approaches when designing a neuron circuit. Next different implementations of ion channels from literature are discussed.

1.5.1 The Design Approach Triangle

"I'm building neurons not differential equations."

John Arthur, designer of the neuron presented in [41] at the Telluride Neuromorphic Cognition Engineering Workshop 2009.

This quote points at different design approaches argued when creating a neuron or a neuromorphic system. I use an image I call Design Approach Triangle for explanation (See Figure 1.7). When designing a neuron, a reference is necessary. The question is whether

[6] Address Event Representation
[7] Field Programmable Gate Array

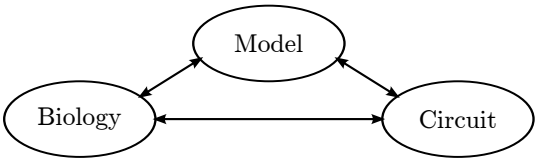


Figure 1.7: The Design Approach Triangle.

this reference should be the designer's interpretation of biology or a model from theoretical neuroscientists. The biological realism of the latter models has usually been investigated. I will call the first approach circuit-driven design and the latter model-driven design. Both approaches are justified. However, the best solution will be something in between.

Circuit-driven design

When directly connecting the circuit with "biology", the circuit itself is the model of biology. The great benefit is, that very compact designs are possible in a circuit-driven design. Circuit designers have much more freedom in their work. In addition, the models of theoretical neuroscientists are driven by the purpose to be simulated. Different models might allow a much easier hardware implementation while simulation is inconvenient. In [39] for instance single transistors are used as ion channels and the authors argue Hodgkin and Huxley would have used transistors if they only had been available at their time.

Losing correspondence?

However, the biological relevance of a circuit-driven design needs to be carefully proven. The danger of loosening the correspondence to biology is apparent. In addition, a new circuit model might not be used by modelers as theories from simulations can hardly be referenced. Who will tell it is not a special feature or bug of the circuit if new outcomes are measured with the circuit? Another danger is the tendency to sell noise or other imperfections directly as a feature. Indeed, both systems are noisy. However, an analytical comparison of the occurring noise figures is necessary before setting them equal.

Model-driven design

Model-driven design can lead to larger circuits as the implemented model might be optimized for simulations. In addition, special design techniques like a current-mode design can be inhibited. The freedom of the designer reduced. Furthermore, direct correspondence of individual circuit parts to biology is not necessarily given especially when phenomenological models are designed. Indeed, a perfect match between model and circuit is not possible – there will always be trade-offs. An emulated model will not reach the perfection of a software simulation. Nevertheless, full floating-point accuracy is not given in biology either so models relying on a high accuracy might not be realistic.

No perfection

Linking model and circuit

However, the link between the model and the circuit can be proven much easier. Individual differential equations can be directly implemented by circuits. In addition, the model itself and the circuit can be simulated and directly compared. The model is a benchmark for the circuit.

Neuroscientists can transfer their experiments onto the hardware implementation and use it as an analysis tool for new theories. On the other hand, outcomes of hardware experiments can be double checked in scaled down simulations.

Limits

When working on the multi-compartment implementation of the neuron presented in this thesis, I discovered a lack of simple phenomenological multi-compartment implementations in literature. Model-driven design approaches its limit.

Next I will discus different implementations of ion-channels in analog VLSI.

1.5.2 Ion Channel Implementation

Voltage-mode

The direct way of designing a neuron circuit is to code the membrane voltage directly as a voltage. This approach is followed in [38–40,42,51] and in the work presented in this thesis[44]. Ion channel channels are implemented using Operational Transconductance Amplifiers(OTA) [40, 44], single transistors [39, 42] or even switched capacitors [51].

Current-mode

In contrast, Arthur [41] and Indiveri [45] decode the membrane voltage as a current. Integrator behavior is achieved using subthreshold current-mode low-pass filter. Very low power consumption is achievable using this approach.

In section, I discuss the different implementations. A review of different neuron types can be found in [46] if more detail is desired.

Operational Transconductance Amplifier

Higher level circuit

Operational Transconductance amplifiers are higher level circuits directly implementing a (small-signal) conductance which is usually adjustable by a biasing current. The devices have two input terminals and output a current which is proportional to the differential input voltage multiplied by a conductance. Using feedback, these devices can be used directly to implement a leakage conductance for instance (See Section 3.3). The simplest OTA is a single differential pair.

Small-signal circuit

However, OTAs are usually small-signal devices with a small linear range when larger signals are applied. There are techniques to enlarge the linear range (See 3.3). Nevertheless, a limitation remains.

In comparison to other implementations except for switched-capacitors, OTA based implementations tend to need more transistors. When kept within the linear range, OTAs allow a close model correspondence. Individual parameters like the individual reversal potential and the conductances are easily parameterizable. Fixed-pattern noise exists indeed, but the effects here are smaller than in comparable designs relying on subthreshold biased transistors.

Current Channels

Compact, little link to biology

The work presented in [42] by Jayawan H.B. Wijekoon and Piotr Dudek, builds a very compact (only 14 transistors) neuron working in an accelerated time domain. It is based on the Izhikevich model presented in [21]. However, ion channels are not implemented as conductances but as currents - they can be implemented by single transistors. Accordingly the biological relevance is disputable. Basically the only set-able parameters are the reset voltage and the maximum spike height. Without stimulus, the membrane voltage would converge to the chips ground level. Compactness has its price.

Nevertheless, the neuron can achieve biological spiking behavior like bursting and spike frequency adaptation for instance. It might be a good model if no close match with biological parameters is necessary.

Single Transistors

Replacing conductances by transistors

A very interesting approach is taken by Farquhar and Hasler in [39]. The authors suggest that it is sufficient to use a single transistor as ion-channel. Transistors are nonlinear indeed – linear approximations can only be done for small drain source voltages. It is discussed, however, that a perfect linear conductance is not biophysically realistic. In an ion channel, a diffusion process is happening, which has a similarity to a MOSFET in subthreshold region. The characteristic of a single transistor might be even closer to biology than a perfect conductance.

Realistic action potentials

Using only six transistors and four capacitors, the authors manage to implement a HHM like behavior. Opening and closing potassium and sodium channels have been implemented. A realistic continuous action potential is achieved. Due to local floating-gate circuits used for biasing, effects of miss-match can be counterbalanced and all time constants and reversal potentials can be adjusted.

Disputable model correspondence

However, although this approach might lead to the best neuron implementation from a circuit point of view, the taken assumptions can conflict with models used in simulations. The circuit's model is similar but not equal to a Hodgkin and Huxley neuron. Indeed, it is not equal to a Hodgkin and Huxley neuron as the gating variables are different and transistors are used for conductances. The authors argue that a fit to the equations from Hodgkin and Huxley would be nothing but adding another layer of abstraction.

Switched Capacitor

Function

Switched capacitor conductances are implemented in [51] for instance. A basic switched capacitor conductance element can be found in Figure 1.8. The principle can be described as

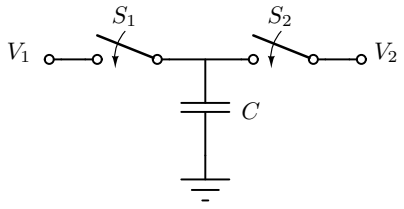


Figure 1.8: Implementing a resistor with a switched capacitor.

follows: Assuming S_2 is closed at the beginning, the capacitor is loaded to V_2. If S_2 is opened now and S_1 is closed, the charge $Q = (V_1 - V_2)C$ is conducted from V_1 onto the capacitor. Going back to the state at the beginning, the same charge is flowing to V_2.

Opening and closing the switches with a frequency f results in a current of $I = f(V_1 - V_2)C$. Consequently, the circuit behaves like a conductance if the frequency is high enough.

The value of a capacitance can be relatively well defined on a microchip in comparison to resistors or the threshold voltage of a transistor. Accordingly, a switched capacitor conductance is well defined.

High price

However, the price is high. Periodically switching with a high frequency causes noise and more current consumption. The switches themselves should have a high conductance. Consequently, they will have a large switching capacitance. If different conductance values are needed, either the clock frequency or the size of the capacitor must be switchable. Several different clocks might be required. The additional clocks will interfere with the analog signal by cross-talk. Furthermore, large capacitors might be required increasing the size of the circuit.

Subthreshold and Current-Mode

In the models designed by John Arthur[41] and Giacomo Indiveri[45], the membrane voltage of a neuron is represented by a current instead of a voltage. Although the intuitive correspondence is lost this way, many advantages are gained.

Easy distribution

Currents can be distributed and copied easily using only the two transistors of a current mirror. Complex designs are possible. Due to the characteristic of the MOSFET, larger current changes require only small changes at the gate voltage. This is especially the case when

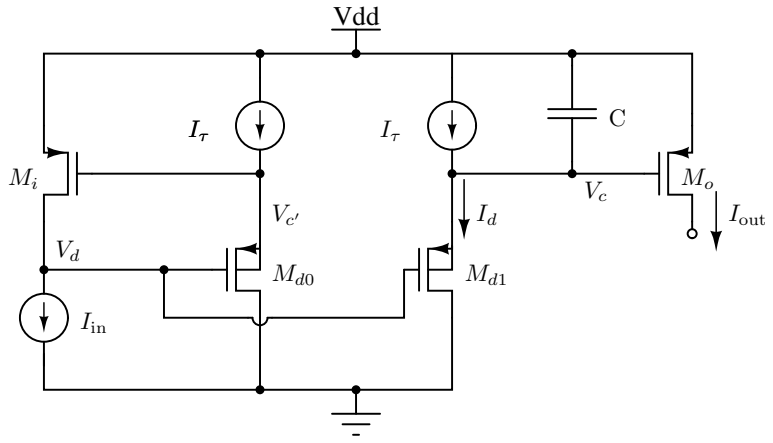


Figure 1.9: Current-mode leaky low pass filter as shown similar in Figure 2 in [52].

working in the subthreshold regime as the characteristic is exponential then. Consequently, crosstalk is reduced as voltage levels are reduced. The relative value range of the membrane current can be large.

Log domain

I will discuss a log domain current-mode integrator circuit which is close to the one used in [41]. A more evolved version is used in [45]. It is discussed in [52]. The circuit shown in Figure 1.9 has been published first in [53]. The discussion here is based on its version shown in [52] however.

When working subthreshold, the output current can be written as

$$I_{\text{out}} = I_0 e^{\frac{\kappa(V_c - V_{dd})}{u_t}}. \tag{1.10}$$

(See [52].) Applying Kirchhoff's current law at V_c results in:

$$C\frac{dV_c}{dt} = I_\tau - I_d. \tag{1.11}$$

However, by differentiating Equation 1.10 by V_c, dV_c in Equation 1.11 can be replaced and we retain:

$$-\frac{Cu_t}{\kappa I_\tau}\frac{dI_{\text{out}}}{dt} = I_{\text{out}}\left(1 - \frac{I_d}{I_\tau}\right) \tag{1.12}$$

The equation already looks like an equation describing the dynamics of an temporal integration circuit. However, the input current is still missing.

The voltage V_d is the sum of the gate source voltages of M_i and M_{d0} and the sum of the gate source voltages of M_o and M_{d1}. As we are working in the subthreshold regime, these voltages correspond to the logarithm of the currents of their transistors. Consequently, we can set: $I_{\text{in}}I_\tau = I_{\text{out}}I_d$ which can be induced in Equation 1.12. In addition, we use $\tau = \frac{Cu_t}{\kappa I_\tau}$. The result is:

$$-\tau\frac{dI_{\text{out}}}{dt} = I_{\text{out}} - I_{\text{in}} \tag{1.13}$$

Finally the current-mode low-pass filter is finished. The derivation is not intuitive however.

The concept is elegant. Nevertheless, when more complex behavior is to be achieved, analytical solution can rely on complex assumptions like an equal κ for NMOS and PMOS

devices [52]. To implement adaptation, or an exponential feedback as sodium channel emulation, Circuits adding or subtracting additional currents are directly connected to the capacitor c. They would have to be added in Equation 1.13. This procedure is only possible if the neuron itself is assumed as a linear system however. In [45], for instance, the adaptation current is not influenced by the membrane current beyond an action potential.

A danger of the concept is the reliance on stable voltages as the current is exponential to the voltages. Although not producing much noise, the circuit itself can be noise sensitive. In addition, the influence of fixed pattern noise is drastic in the subthreshold regime. Hence the circuit cannot be controlled easily – especially if no individual bias parameters are available. Another issue is the temperature dependency, as there is a dependency on $\exp(1/u_t)$.

2 Neuromorphic Environment

This chapter describes the environment for which the neuron of this work are designed. A top-down approach is followed. The chapter starts with a very brief description of the BrainScaleS project. Subsequently, the complete wafer-scale system and its concepts are presented. Finally the chip HICANN which is the analog neuron network chip of the system is discussed.

2.1 The BrainScaleS Project

A multi national project

This dissertation has been performed within the neuroscientific project BrainScaleS. The project is a collaboration of 18 work groups distributed to ten different European countries. The goal of this project is to observe and comprehend different scales of the brain and the interaction of mechanisms occurring on different scales [54]. Scales can be temporal or spacial.

Temporal an spacial scaling

Scales in space or complexity reach from the detailed modelling of single cells to a functional modeling of complete cortical areas. Time-scales reach from milliseconds when observing voltage traces of single neurons and fast synaptic adaptation effects to hours or days when development and long term learning effects have to be taken into account [54]. Research involves biological measurements, modeling and simulation on – if necessary – super computers and the construction of specialized neuromorphic hardware. Apparently, this thesis is located in the neuromorphic part. Additionally some modeling aspects are discussed.

2.1.1 Interaction

Interconnection different levels of accuracy

Interaction between different scales can arise from having different levels of detail in a complex simulation or emulation. Interaction is given by information exchange. A macroscopic phenomenological model of retina could create action potentials out of a video stream for instance. Those are fed into a network of neurons. Neural circuits under test or circuits relying on a high level of detail for proper functioning are simulated or emulated in detail. Different neurons might even need a special model. Other surrounding areas of the circuit are simulated with functional models interacting with the detailed models via action potentials for instance. This can involve macroscopic models of complete areas of the brain. There can be a fluent transition between scales. Without emulatin, the concept is similar to approaches used in engineering to allow complete system simulations without unusable simulation times (Look for Virtuoso UltraSim Full-Chip Simulator in [28]).

Closing the loop

The holy grail would be to have closed loop experiments where the system can interact with a (virtual) environment. Environmental changes could occur as consequence of actions of the system. Subsequently, the system could react on these changes.

2.1.2 Hybrid Multi-Scale Computing Facility

System overview

To allow large scale multi-scale modeling from a hardware point of view, a system called Hybrid Multi-Scale Computing Facility(HMF) is under development. The most recent state

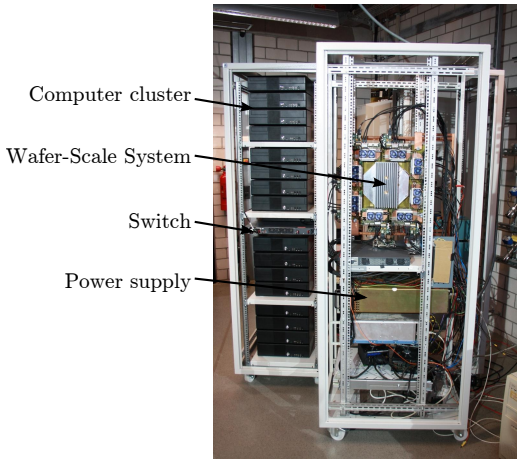


Figure 2.1: Two racks forming the current state of the Hybrid Multi-scale Computing Facility. A computer cluster is connected with a Wafer-Scale System via Giga=Bit Ethernet. The final system will have 6 inter connected Wafer-Scale Systems.

of the implementation can be found in Figure 2.1. The final system will consists of six BrainScaleS Wafer-Scale Systems(BWS) interconnected with each other and a computer cluster.

Neuromorphic computation is performed on the wafer-scale system discussed in the next section, while the cluster allows the use of conventional models. Connection between the BWS and to the cluster is done via Giga-Bit Ethernet. Each BWS emulates up to 200 000 neurons with 224 synaptic connections each. The emulation time of the neurons is between 10^3 and 10^5 shorter in comparison to biological real-time (See 1.4.1). Accordingly, biological days can be emulated in a minute or less. The cluster will support an environment using macroscopic functional models.

2.2 The BrainScaleS Wafer-Scale System

Four basis components

The BrainScaleS Wafer-Scale System (BWS) – see Figure 2.2 a) and Figure 4.5 – can emulate networks on a complete silicon wafer. The BWS is described in [50] for instance. The concepts are presented in [55] and [56]. The system consists of four components which are connected in a hierarchical structure. A custom FPGA[1]-board [57], the Digital Network Chip (DNC) [58], the system PCB[2] [59] and the wafer.

Communication group

An FPGA-board with four DNC form the communication group. The communication group is responsible for interfacing the wafer-scale system to the cluster or other BWS.

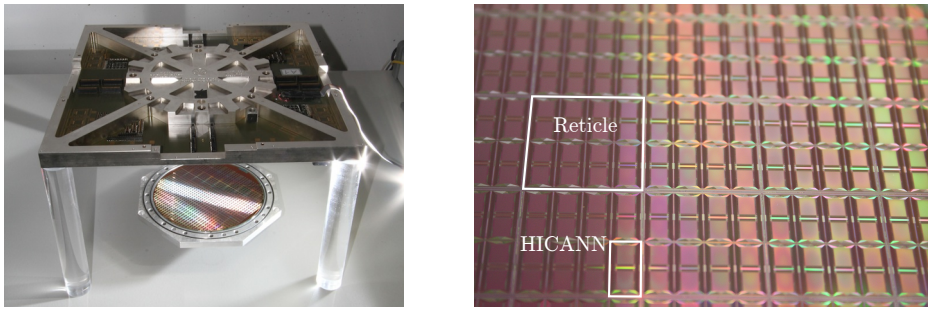
System PCB

The system PCB routes the signals between the communication group and the wafer. Furthermore, power is supplied through this PCB. Current consumption is monitored and single components are switched of when consuming to much power [59]. Special elastomer connectors interface the wafer [60].

Wafer and reticles

A wafer consists of 48 reticles – see Figure 2.2 a). A reticle is the maximum sized chip production unit given by the size of the masks used for lithography. The mask is stepped over the wafer during production, so the structure is repeated on the silicon substrate. In

[1] Field Programmable Gate Array
[2] Printed Circuit Board



(a) Disassembled wafer-scale system, communication group excluded

(b) Detail of uncut wafer without post-processing. 8 HICANNs form a reticle.

Figure 2.2: Wafer-scale integration

the uses manufacturing process, the reticle size is a square of 2 cm edge length. A DNC is responsible for one individual reticle. Each reticle consists of 8 interconnected HICANN[3]. A HICANN is a rectangle shaped ASIC[4] of 5 mm times 10 mm. The size is given by MPW[5] prototyping. The HICANN will be presented in greater detail in the next section. However, communication will be discussed first.

HICANN

2.2.1 Communication

Chip production techniques prevent reticle interconnections

Wafer-scale integration with an uncut wafer as implemented in the BWS in only useful if individual chips can be interconnected directly on the wafer. This way the bottleneck of PCB connections can be skipped. However, when producing ASICs, there is no electrical connection between reticles. In addition, the factories place special characterisation structures in the scribe-line where reticles are usually cut.

Post-processing

Consequently, additional layers of metal have to be added onto the wafer to interconnect reticles in a post-processing step. Furthermore, large connection pads for the elastomer connectors are created this way.

Layer 1

Vertical and horizontal serial buses are routed over the complete wafer to transport digital action potential events between individual HICANNs to form large networks. Communication via these serial buses is called layer 1 communication (L1).

Layer 2

L1 signals cannot be directly interconnected to external components. However, each HICANN has one layer 2 (L2) interface which can be fed by up to 8 L1 buses. This L2 connections are connected to the DNC and through the DNC to the FPGA and to the outside world. L2 is realized via a packet network using time stamps. Hence, delays can be added to the digital spike events.

Longer distance smaller bandwidth

The bandwidth of a L2 bus is similar to a L1 bus. However, a L2 connection can be fed by 8 L1 connections. Consequently there is a bottle neck. Nevertheless, the same accounts for long range connections in biology which are less dense than local connections as they are much more expensive. The hierarchical design of the BWS reflects this issue.

[3]High Input-Count Analog Neural Network
[4]Application Specific Integrated Circuit
[5]Multi-Project Wafer. A reticle is shared among different projects to save prototyping costs.

2.3 The HICANN Microchip

The HICANN ASIC is the basis of the wafer of the BWS. It is specified in [61]. First concepts are published in [55] and [56]. A newer publication of the complete ASIC showing some simulation results can be found in [50]. First neuron measurements have been published in [44]. A photograph of the chip can be found in Figure 2.3

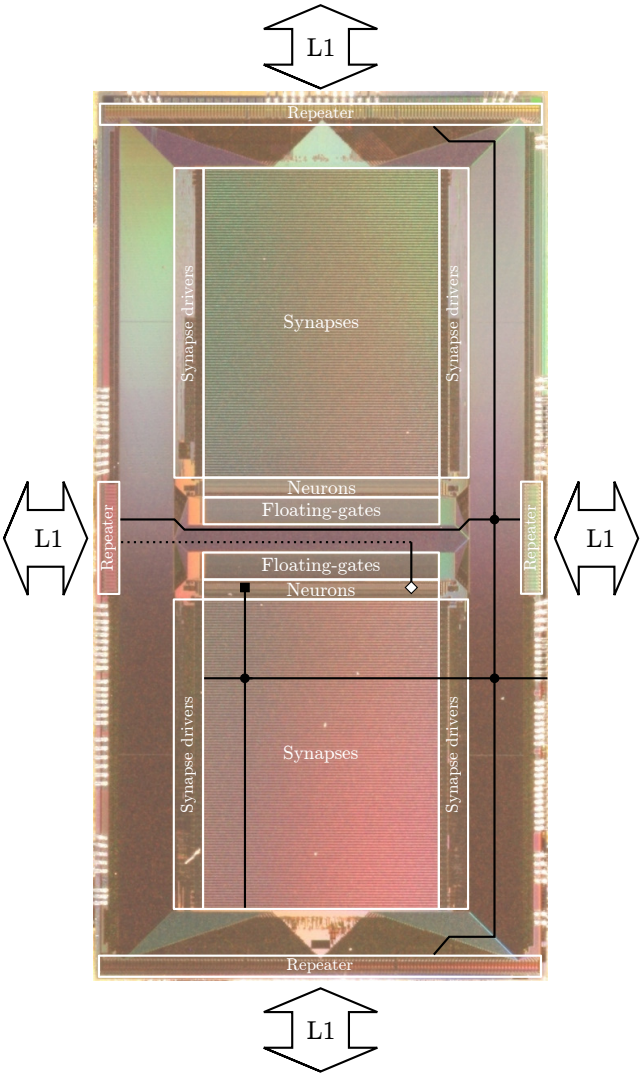


Figure 2.3: Photograph of the HICANN microchip. Important analog circuits are marked. The black line denotes the path of an action potential starting at the white diamond and ending at the black square. Digital spike events can be transported to neighbouring HICANNs via L1.

Synapse array

The most dominant features of the photograph are the two synapse array of 256 x 224 synapses. Each synapse has a four bit address and a four bit weight. A mechanism measuring the exponentially weighted time differences between spikes for STDP is implemented locally

in each synapse.

Synapse driver

The synapse arrays get input from the synapse drivers on the sides. Short-term plasticity mechanism are implemented in these circuits. The implementation is analogue to [62].

Neurons

Two rows of 256 neurons receive synaptic stimulus from the synapse arrays. These neurons implement the Adaptive-Exponential Integrate-and-Fire Neuron Model (AdEx). Their design is the topic of this thesis.

Floating-gates

Analog floating-gate memory cells are used to supply individual biases for each neuron and to create all other necessary biases of the chip. These cells and their control are discussed in Chapter 9.

L1 repeater

The repeater circuits receive and transmit serial L1 data. This procedure is necessary to restore signal quality. In addition, eight repeaters can feed data onto the L1 bus.

L1 buses

The serial L1 buses themselves are are situated around the analog circuits. They are directly visible in Figure 2.3 as they are routed on the topmost metal layer. Horizontal an vertical buses can be interconnected via pass transistors located in a cross bar matrix below the crossing at the bottom and top middle of the chip.

Between the two floating-gate arrays, there is circuitry generating digital spike events with a 6 bit address from single one bit signals created by the neurons. This circuit is called Neuron L1 Interface (NL1).

Digital part

Below the L1 bus – except for the area between the floating-gates – are standard cells forming the digital control of the HICANN. The different digital control module are interfaced via a bus based on the OCP[6] standard [63]. The HICANN has two main clocks. A fast clock which is usually set to 200 MHz and a slow clock which is the fast divided by five.

2.3.1 Life Time of an Action Potential

Here, the journey of an action potential, created at the diamond neuron in Figure 2.3 and received by the black square neuron, is presented.

Neuron

When the diamond neuron detects an action potential, it creates a time continuous digital signal. This signal is driven to NL1[7]. Each neuron has a 6 bit address in NL1. When a firing signal is received, NL1 sends this 6 bit address into the digital part. This 6 bit address is the spike event which is transported in the system.

Digital part

In the digital part, the event can be send to a repeater connecting to L1 or to an interface connecting to L2 to send the event to the corresponding DNC. The latter is used if the event needs to be red out for measurements.

Repeater

However, in this case, the event is send to a repeater only feeding the event onto a serial L1 bus. This vertical (in relation to Figure 2.3) bus crosses the complete chip and connects to neighboring chips. A pass-transistor from the crossbar switch connects this vertical bus to a horizontal bus.

Synapse driver

The vertical bus is connected to a synapse driver by another pass-transistor. In the synapse driver, the serial signal is deserialized. Two bits of the address are used to chose which column of synapses is to be driven. For details see [56]. The remaining four address bits are driven onto the four address lines of the corresponding column. In addition, a digital pulse with a defined length is send into the array. The length of this pulse can be adapted for short term plasticity.

Synapse

If the remaining 4 bits of the neuron address match the address of a synapse, the synapse transforms the digital pulse into a weighted current pulse. In addition, its STDP mechanism is triggered. The strengths of the current pulse is determined by a 4 bit weight.

[6] Open Core Protocol

[7] In addition it is fed back into the synapse array for triggering STDP

Neuron

The current pulse is driven onto a line connecting to the synaptic input circuitry of a neuron where it is transformed into a conductance. This open conductance might excite the post synaptic neuron to create another action potential. The circle is closed.

2.3.2 High Input-Count

10 000 inputs in biology

The acronym HICANN stands for Hight Input-Count Analog Neural Network. This has to be understood from a historical point of view, as the predecessor chip, the SPIKEY chip [64] was limited to 256 inputs [40]. In contrast, biological neurons can have 10 000 or even more inputs from different neurons. The number of synapses on a chip is limited due to the size of necessary memory bits and local STDP circuitry. How to overcome this hard constraint?

Building large neurons

The number of synapses per neuron circuit on the HICANN is even smaller as addresses and thereby more memory cells have been added to the synapses. However, neurons can be interconnected by switching their membranes together. This way, large neurons are constructed of up to 64 neurons circuits. Consequently, the total maximum input-count of a larger neuron 14 336.

Wafer-scale integration is necessary

This high-input count needs to supplied with neural events. Especially as the system operates between 10^3 and 10^5 times faster than biology, high bandwidth are necessary. Assuming the 14 336 inputs would fire with a biological rate of 20 Hz at 10^5, a bandwidth of 172 Gbit/s results. However, a much higher bandwidth is be required as not all neurons have the same set of pre-synaptic neurons. Assuming the HICANN would be equipped with a bond pads with a pitch of 100 µm on all edges and all these bond pads would be used for differential buses, 150 buses could be implemented. With a bandwidth of more than 1 Gbit/s a realization would be possible. However, a chip needs more pins for operation – in particular power supply pins. The technological limit of wire bonding techniques in our technology is approached. Furthermore, power consumption would be a great issue with this high bandwidth of off-chip communication as the load of the buses is larger when leaving the chip. Wafer-scale integration solves this problem.

3 Point Neuron Emulation

This chapter describes the design of the actual neurons of the HICANN chip. At first, the implemented model - the Adaptive exponential integrate-and-fire neuron model - is introduced. Subsequently, after presenting general design concepts and the complete circuit, each circuit component is analyzed. Theoretical concepts of circuits are discussed followed by simulations showing the behaviour of the real circuit. The chapter is concluded by a presentation of the circuits parameter ranges and the relationship between biological parameters and technical.

3.1 The Adaptive Exponential Integrate-and-Fire Neuron Model

A two dimensional neuron model with positive feedback

The Adaptive Exponential Integrate-and-Fire Neuron Model (AdEx)[22] has been developed by Romain Brette and Wulfram Gerstner within the FACETS project[65]. Similar to the quadratic adaptive Integrate-and-fire Model from Izhikevich [21](called Izhikevich Model in the following), it is a two variable model enhancing the classic Integrate-and-fire neuron (See [66] for instance) by an adaptation variable and a positive feedback term. In contrast to the Izhikevich Model, the AdEx uses an exponential function as positive feedback. The positive feedback is essential for burst generation for instance as the model needs to have points in phase plane where spiking is inescapable. In [22], Brette and Gerstner prove that the AdEx is capable of reproducing biological neuron behavior.

3.1.1 Model Description

The AdEx is defined by the following two equations for the membrane voltage V and the adaptation variable b completed by the reset conditions shown in Equations 3.3 and 3.4:

$$-C_{\mathrm{m}}\frac{dV}{dt} = g_{\mathrm{l}}(V - E_{\mathrm{l}}) - g_{\mathrm{l}}\Delta_{\mathrm{t}}e^{\left(\frac{V-V_{\mathrm{t}}}{\Delta_{\mathrm{t}}}\right)} + g_{\mathrm{e}}(t)(V - E_{\mathrm{e}}) + g_{\mathrm{i}}(t)(V - E_{\mathrm{i}}) + w; \tag{3.1}$$

$$-\tau_w\frac{dw}{dt} = w - a(V - E_{\mathrm{l}}). \tag{3.2}$$

Here C_{mem} is the membrane capacitor; g_{l}, $g_{\mathrm{e}}(t)$, and $g_{\mathrm{i}}(t)$ are the conductances for leakage and the excitatory respectively inhibitory synapses. E_{l}, E_{e}, and E_{i} are the corresponding reversal potentials. The second term of Equation 3.1 is the so called exponential term. Here V_{t} and Δ_{t} are the effective threshold potential and the threshold slope factor. a is adaptation parameter and has the dimension of a conductance. Last but not least, τ_w is the time constant of the adaptation variable.

When V reaches a certain threshold Θ, a spike or action potential is triggered and both variables V and b are set to new values:

$$V \to V_{\text{reset}}; \tag{3.3}$$
$$w \to w + b. \tag{3.4}$$

w is increased by b at each action potential generating a model behavior called Spike-Triggered Adaptation. The mechanism results in a decrease of the spiking frequency if the model is stimulated by by a constant pulse and is one of the main features of the model. A sample for Spike-Frequency Adaptation can be found in Figure 3.1.

Indeed, the exact value of the threshold Θ is uncritical if the exponential term is active[22]. The large derivative in the exponential limits the effect of Θ on the detected spike time. Furthermore, the independence to the exact value of Θ can be seen in Figure 3.1. Due to the limited resolution of the simulator, the spike heights differ from spike to spike.

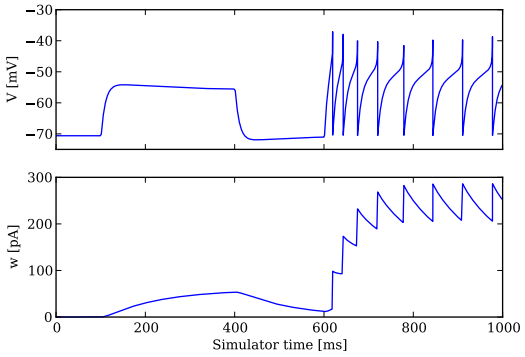


Figure 3.1: NEST[67] simulation: Membrane voltage and adaptation variable w of AdEx neuron stimulated by two current pulses. The first pulse is to small to reach the spiking threshold. During the second pulse Spike-Frequency Adaptation can be observed. Neuron parameters are equivalent to the parameters shown in [22].

3.1.2 Model Dynamics

Phase planes and nullclines

A phase plane plot can be used to visualize the dynamics of the two variables V and w. Figure 3.2 shows the phase plane of the model using parameters from [12]. The nullcline of a variable is the trace where the derivative in time is 0. Below their nullclines, w and V are growing. The crossing on the left is a stable fix point. In contrast, the crossing on the right is unstable. When the neuron is stimulated by a constant current, the V-nullcline is shifted in w direction.

The phase plane of the stimulated model can be found in Figure 3.3. Here, the Nest simulation is the same as in Figure 3.1 with two different current pulses as stimulus. The trajectory starts in the lower left corner. The first circle is the first stimulus which is still too small for spiking. After the first pulses is done, the neuron moves back to the steady state - the stable fix point. Subsequently, the second current pulse forces the neuron to spike and w is enlarged with each spike. At last a constant frequency is reached and w decreases the same amount between each spike, it is enlarged at each spike.

By changing the models parameters, the model is capable of reproducing different neuron behaviors like Bursting, Initial Bursting, Tonic Spiking, and more [12]. The versatility is similar to the capabilities of the Izhikevich model shown in [11]

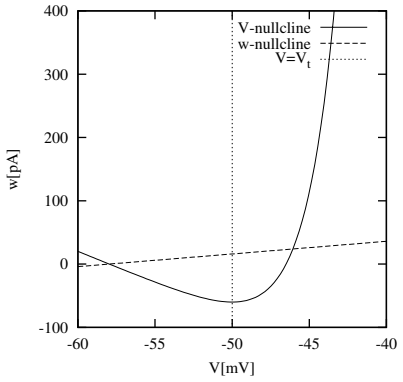


Figure 3.2: Phase plane of the AdEx model with parameters according to Figure 4 d) from [12], stimulus excluded. V and w will be rising below their nullclines and falling above. The figure has been published in [44]

3.1.3 Synaptic Stimulation

Interconnecting neurons

In network operation, the neuron model is not stimulated by current pulses, but by synaptic input from other neurons. This stimulus is modeled by the time dependent conductances $g_{\mathrm{e}}(t)$ and $g_{\mathrm{i}}(t)$. These simplified conductances can be described by the following equation[5] for incoming spikes at times $t^{(f)}$:

$$g_{\mathrm{e/i}}(t) = \sum_f \bar{g}_{\mathrm{syn}} e^{-\left(t-t^{(f)}\right)/\tau_{\mathrm{syn}}} \Theta\left(t - t^{(f)}\right) \tag{3.5}$$

Here Θ is the so-called Theta-function which is one for values above 0 and 0 below. Hence the conductance value for a single for a single input spike starts with the conductance g_{syn} and decays with τ_{syn}. The latter is called synaptic time constant.

A more complex model could include a exponential rise of the conductance especially for excitatory synapses. Using this approach and going to the limit of small rise times results in so called Alpha-functions as conductance shape[5] for a single incoming spike:

$$\alpha(x) = \frac{x}{\tau^2} e^{\left(\frac{x}{\tau}\right)} \Theta(x), \tag{3.6}$$

with $x = t - t^{(f)}$.

This conductance shape is commonly used in modelling.

3.1.4 Conclusion

A flexible and realistic model

Due to its flexibility and biological relevance, the AdEx fits perfectly for an implementation in biologically inspired neuromorphic hardware. With the proper parameterization capabilities, the model is backward compatible to less complex models. Complex behavior can be switched off, if not needed. Using a model like the AdEx, neuromorphic hardware is not limited to a single type of neuron model.

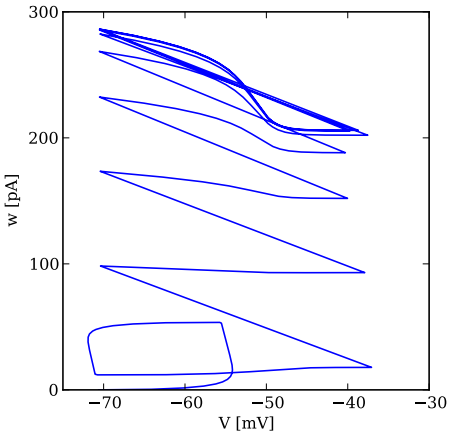


Figure 3.3: Nest simulation: Corresponding phase plane of the traces shown in Figure 3.1

3.2 Structure and Design Concept

OTAs implement ion channels

The most important question when designing a neuron circuit is how the individual conductances are implemented. Operational transconductance amplifiers (OTA) have been chosen here. Different ion channel concepts and the OTA have been introduced in 1.5.2. The OTA ion channel is the straight forward solution here, allowing direct ion channel implementation with full parameterizability without requiring special technology like local floating-gates. OTAs are possible as our neuromorphic hardware device is not supposed to work in real time, but with an accelerated time scale. Accordingly, the conductance ranges can be in the ranges "natural" for the used CMOS process and subthreshold dynamics are rarely used.

Circuit keeps model structure

The structure of the circuit neurons keeps the structure of the model implementing each term in a dedicated circuit. An overview of the neuron is given in the schematic shown in Figure 3.4. The use of analog floating-gates for parameterization allows individual values for nearly all neuron parameters. Consequently, each neuron has its own set of individual parameters. Here we give a short overview of the parts of the neuron circuit.

Each hardware neuron has two synaptic input circuits receiving current pulses from the Synapse Array. The parameterization of these circuits chooses the type of connected synapses by setting the time constants and the corresponding reversal potentials. The Leak Circuit is basically a single OTA implementing the Leakage Term of Equation 3.1. An operational amplifier with a voltage divider and a single transistor is used to implement the Exp circuit. The adaptation variable w is basically stored on a capacitor in the Adapt circuit. The Adapt circuit receives spike signals to enable Spike-Triggered Adaptation. Spike detection is done by a comparator in the Spiking/Connection circuit. Furthermore this circuit is responsible for connecting the membrane of the neuron to neighbour neurons and to propagate and occasionally receive spiking signals. Subsequently, the Reset circuit pulls the membrane potential V to the reset potential V_{reset} Finally, the In/Out circuit includes a buffer connecting the membrane voltage to the readout line. In addition, the membrane can be stimulated by a current through this circuit.

Neurons are designed as pairs

For layout and routing reasons (see 3.10), respectively two neighbouring neurons are combined in a pair and share the spiking and connection circuit. Additionally, internal control memory is shared.

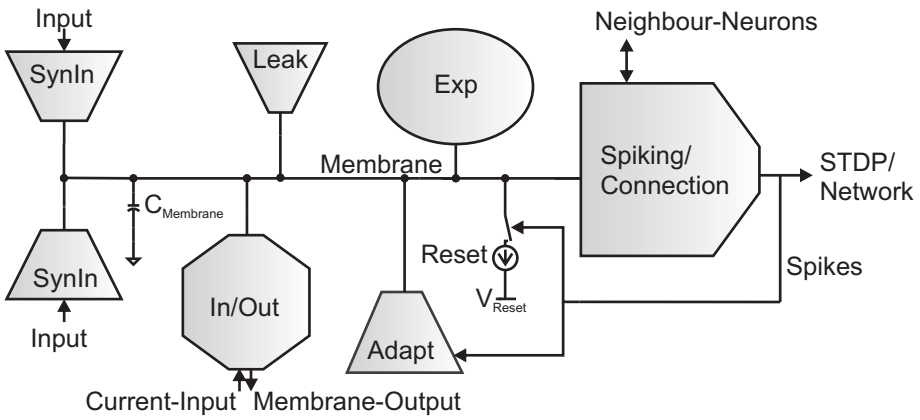


Figure 3.4: Simplified schematic of the AdEx implementation. Each term from the model equations is a dedicated circuit and can be controlled individually. Parts of the model can be switched off to emulate less complex models. The figure has been published in [44]

At the beginning of this thesis, some schematics have allready been available. These circuits included the OTA, the adaptation term, and the synaptic input. They have been designed by Johannes Schemmel - their integration, analysis and verification is done in this thesis.

3.3 Operational Transconductance Amplifier and Leakage

As the design methodology is based on operational transconductance amplifier(OTA) emulated ion channels, the OTA is the key circuit of the neuron. It appears 7 times in the complete neuron schematic.

3.3.1 Ideal Operation

Direct conductance emulation

The ideal OTA is a 3 terminal device with two Inputs and one output. The current is the product of the difference voltage at the inputs and a conductance which is proportional to a biasing current. Accordingly, as the name suggests, an OTA emulates a conductance. Figure 3.5 shows an OTA connected to emulate the Leakage Term of the AdEx. In addition to the conductance emulation, the terminal E_l can be at high impedance.

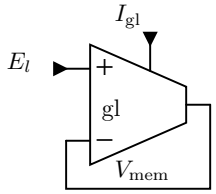


Figure 3.5: The leakage term circuit: simple and direct usage of OTAs

3.3.2 Circuit

A symmetrical CMOS OTA

A schematic of the complete real circuit of the neuron's OTA can be found in Figure 3.6. Basically, the circuit is a symmetrical CMOS OTA(see [26] for an introduction). The differential pair $M_{\mathrm{i}+/-}$ is loaded by two current mirrors. The current from the left branch is mirrored back to the current of the right branch. To compensate the load at the output, M_1 has been introduced in the left branch. All current mirrors except $M_{\mathrm{c}1,2}$ have a one to one current relation.

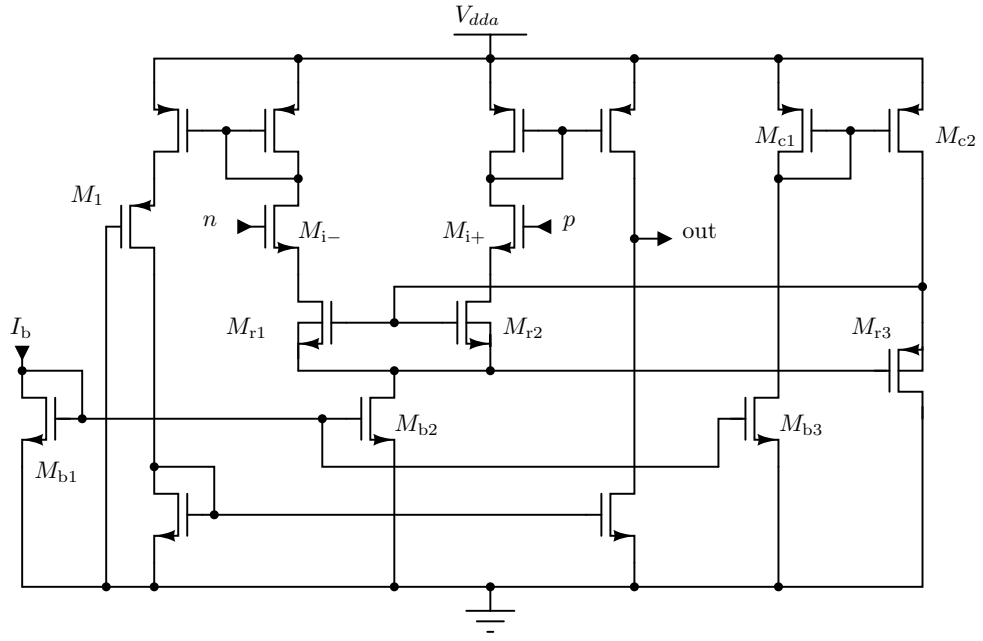


Figure 3.6: OTA of the neuron circuit

Using a small signal device for large signals

Usually an OTA is used to filter AC signals. Indeed, the linear range of a differential pair is quite limited. Without the Transistors $M_{\mathrm{r}1,2,3}$, the maximum differential input voltage is directly limited by the gain of $M_{\mathrm{i}+/-}$ and the biasing current (See 4-7-2 in [23] or below, setting $R = 0$ a detailed derivation). All second order effects like channel length modulation are ignored in the following for simplicity. If the differential input voltage V_d is larger than

$$V_d = V_p - V_n = \sqrt{\frac{I_b}{K'_p(W_i/L_i)}}, \tag{3.7}$$

the complete biasing current I_b flows through $M_{\mathrm{i}+}$ and $M_{\mathrm{i}-}$ is turned off. $M_{\mathrm{i}+}$ is operated as a source follower hence and the OTA directly mirrors the biasing current to the output.

Less gain for larger linear range

The linear range can be enlarged by using a smaller gain. This is achieved in the neuron OTA by the transistors $M_{\mathrm{r}1,2}$. The source follower $M_{\mathrm{r}3}$ sets the gate-source voltages of $M_{\mathrm{r}1}$ and $M_{\mathrm{r}2}$. Both are supposed to be resistive biased in normal operation and lower the gate-source voltages of the input transistors. Notice the bulk connections of $M_{\mathrm{r}1,2,3}$. All three source potentials are connected directly to the corresponding bulk(For $M_{\mathrm{r}1,2}$ triple well transistors have to be uses for this purpose). Accordingly, the common mode dependency of their conductances is minimized to achieve a better linearity.

For simplifications, we assume M_{r1} and M_{r2} to be ideal resistors in the following calculations for the complete circuit. We use I_d as differential output current and V_d as differential input voltage; I_n is the crosscurrent of transistor $M_{\mathrm{i}-}$ while I_p is the crosscurrent of transistor $M_{\mathrm{i}+}$:

$$I_d = I_p - I_n; \tag{3.8}$$
$$I_b = I_p + I_n. \tag{3.9}$$

Consequently:

$$I_p = \frac{I_b + I_d}{2};\ I_n = \frac{I_b - I_d}{2} \tag{3.10}$$

Looking at the differential voltage, we get:

$$V_d = V_{\mathrm{GSp}} + RI_p - (V_{\mathrm{GSn}} + RI_n) \tag{3.11}$$
$$= V_{\mathrm{GSp}} - V_{\mathrm{GSn}} + RI_d. \tag{3.12}$$

$V_{GSn,p}$ can be described using the equation of saturation region without channellength modulation (see 1.3.1) for transistors $M_{\mathrm{i}+}$ and $M_{\mathrm{i}-}$:

$$V_{\mathrm{GSp,n}} = V_t - \sqrt{\frac{I_{p,n}}{K'(W_{\mathrm{i}}/L_{\mathrm{i}})}} \tag{3.13}$$

If we plug this result in Equation 3.12 the threshold voltage V_t gets eliminated:

$$V_d = \sqrt{\frac{1}{K'(W_{\mathrm{i}}/L_{\mathrm{i}})}}\left(\sqrt{I_p} - \sqrt{I_n}\right) + RI_d. \tag{3.14}$$

Now we include Equations 3.10 we obtain:

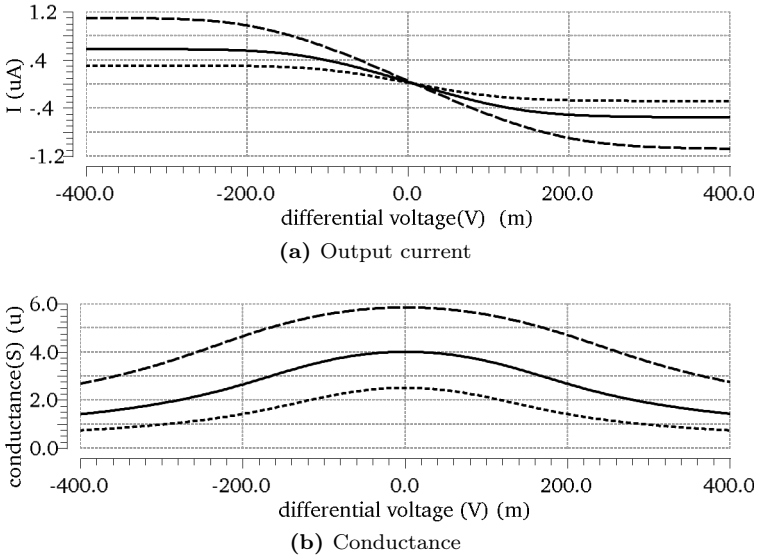
$$V_d = \sqrt{\frac{1}{2K'(W_{\mathrm{i}}/L_{\mathrm{i}})}}\left(\sqrt{I_b + I_d} - \sqrt{I_b - I_d}\right) + RI_d. \tag{3.15}$$

The maximum absolute value of the differential current is the biasing current. It is achieved at the maximum absolute value of the differential input voltage. Without loss of generality, we choose $I_d = I_b$. Consequently, we get an upper border for V_d:

$$V_d \leq \sqrt{\frac{I_b}{K'(W_{\mathrm{i}}/L_{\mathrm{i}})}} + RI_b. \tag{3.16}$$

For $R = 0\,\Omega$ this is similar to the border given in Equation 3.7. Including R, the valid input range of the OTA, is increased by RI_b.

Indeed, the maximum differential input voltage still linearly depends on I_b. Here, we have to take into account that the resistor is only a simplification. In the real circuit (schematic in Figure 3.6), it is implemented by M_{r1} and M_{r2}. The source follower generating the bias for M_{r1} and M_{r2}, M_{r3} is biased by one third of I_b itself. Consequently, simplified R is proportional to $\sqrt{1/I_b}$. Summarized, the maximum V_d is proportional to $\sqrt{I_b}$.



(a) Output current

(b) Conductance

Figure 3.7: DC simulation of neuron OTA for different biasing currents (long dashes: 1 µA, solid: 500 nA, and dashes: 250 nA) . Terminal n is kept at 900 mV while the voltage at terminal p is swept from 750 mV to 1050 mV.

3.3.3 Simulation Results

Linear range is larger than 150 mV

Figure 3.7 shows output current and conductance of the OTA. The current is saturating at values close to the biasing current I_b for differential voltages between 250 mV and 150 mV. To get a better impression of the maximum the maximum differential input voltage still allowing a monotonic rising/falling output current Figure 3.8 marks the differential voltage where 90 % of the maximum current are reached. The scale has been chosen logarithmic to better cover small biasing values. For small biasing currents the voltage is nearly constant at 150 mV. Subsequently a linear rise can be observed.

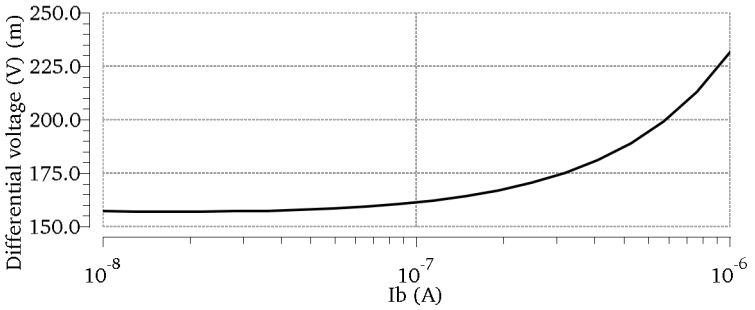


Figure 3.8: Differential input voltage where 90 % maximum output current is reached obtained from DC simulation of neuron OTA with different biasing currents. Terminal n is kept at 900 mV while the voltage at terminal p is swept.

Indeed, the value does not show a square root behavior like suggested in the last subsection. The assumptions are not fulfilled at the borders as the resistive biasing region of $M_{\mathrm{r1,2}}$ is

left. In addition, the used simplified models ignore effects like channel length modulation for instance. Nevertheless, due to the saturation for small biasing currents, we do not have an issue.

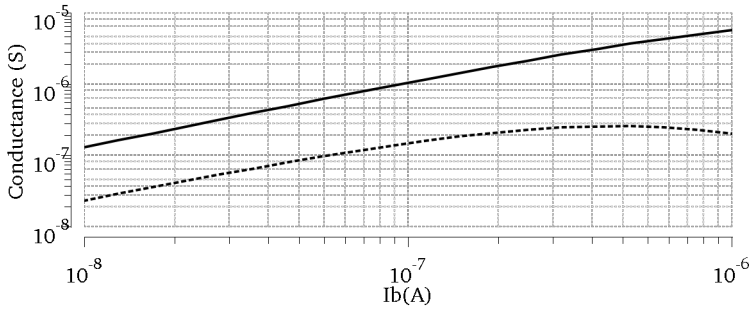


Figure 3.9: Average large signal conductance(solid) and standard deviation (dashed) for a 150 mV sweep of differential input voltage. Obtained from DC simulation of neuron OTA for different biasing currents. Terminal n is kept at 900 mV while the voltage at terminal p is swept.

Better linearity for higher biasing currents

Looking at the conductance in Figure 3.7, a great deviation can be observed. Indeed, this is not surprising do to the saturation of the output current. If we cut the figure to maximum differential voltages of 150 mV, Figure 3.9 gives the average conductance and its deviation for a large sweep of biasing currents. The standart deviation is given as am measure of linearity here. Zero deviation results in a constant conductance as desired. The deviation is one roughly order of magnitude smaller than the average value. In numbers it is estimated 20 % for small biasing currents and estimated 5 % for large currents. Nevertheless, a constriction to smaller voltages will result in a smaller deviation.

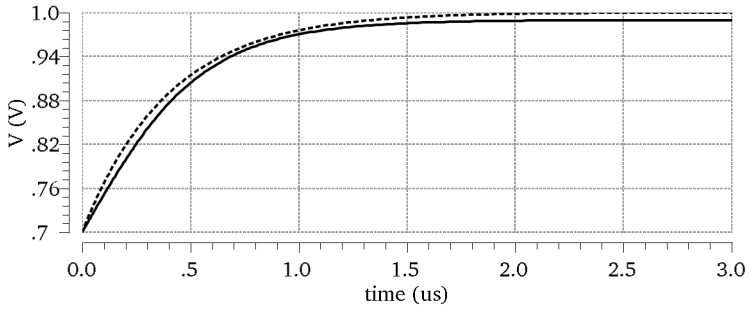


Figure 3.10: Transient simulation: A 2 pF capacitor is charged from 700 mV to 1 V through the leakage OTA at $I_b = 1$ µA (solid) and an ideal conductance of 5 µS (dashed)

Close matching in comparison to real R-C-circuit

Better results can be obtained by simulating the circuit in a more realistic environment and in a transient simulation. The OTA connected as leakage term as shown in Figure 3.5 is simulated in Figure 3.10. Correspondence between both curves is apparent although the OTA is saturated below roughly 800 mV. Indeed, the derivative of both curves is different at the beginning of the curve. Smaller ranges lead to an even better result. The final values in Figure 3.10 differ as the OTA has a small offset.

Limited input voltage range

Another nonlinear effect of the OTA is the limited input range due to the power supply

rail. Simulations have shown that the maximum input voltage resulting in a change at the output if the OTA is connected as a buffer is 1.3 V for 1.8 V power supply. Consequently, the actual limit is 500 mV below the power supply rail. The current mirrors at the load of the differential pair M_n and M_p (Figure 3.6) need roughly one threshold voltage gate-source voltage for operation. Although M_p and M_n are Low-V_t transistors[1], the input voltage is limited. The Low-V_t transistors are used to achieve a better linearity and to enlarge the input range close to the ground rail.

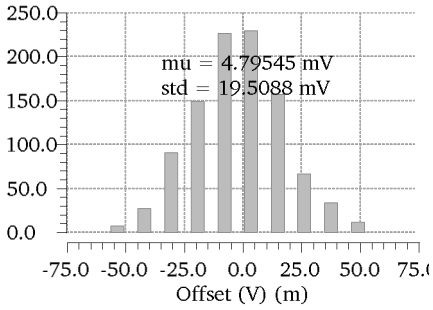


Figure 3.11: Offset at 900 mV with OTA connected as buffer. Obtained from typical DC-simulation with Monte-Carlo-Sampling on Miss-match-data using 1000 samples.

Input stage offset

Last but not least, a major imperfection of differential input stages is the offset created by transistor miss-match especially concerning the threshold voltage. For illustration the Monte-Carlo simulation shown in Figure 3.11 shows the offset of the buffer connected OTA. A sigma of 20 mV is large in an operating range limited to 300 mV to 400 mV. Nevertheless, in most circuits it can be removed by calibration.

3.3.4 Conclusion

The small linear range is not a problem for emulation as long as the OTA is operated in the monotonic range ore close to it. The neuron's operating region will be close to threshold as in biology the relevant usual region would be the high-conductance state[68]. Furthermore, work done by Marc-Olivier Schwartz[69] has shown that the complete neuron circuit is capable of fitting biological membrane traces if calibrated correctly. Nevertheless, the nonlinear side effects destroy the mathematical equivalence between model and emulation over the complete operating range.

3.4 Membrane Capacitor

Metal-metal capacitors as membrane capacitor

The membrane capacitor C_m can be implemented directly using a real capacitor. A process option allows the use of special metal-metal capacitors called MIM-caps. Here an additional metal layer is added between the two top most metal layers having a small distance to the second last metal to enlarge capacitance. Accordingly, these capacitors are close to ideal plane-parallel capacitors.

Switching resistance

Nevertheless, even with the membrane capacitor are some imperfection. Firstly, the capacitor is actually implemented as two capacitor, where one capacitor can be switched off the

[1] This is a special option in our process enabling a threshold voltage below 300 mV

membrane cap. This has to be done for parameter scaling reasons(see 3.12.3). The imperfection added here is the impedance of the switch for the second capacitor. It is approximated 1 kΩ and supposed to be negligible.

Series resistance

Secondly, the membrane capacitor is not a point capacitor as suggested in the model. Actually, is membrane voltage part is routed through the complete neuron and the capacitors are located at the one end for layout reasons. In fact, this add the line impedance in series to the capacitor (100 Ω estimated from layout).

Parasitic capacitance

Additionally, the membrane voltage is connected to all input stages of the neuron circuit components. This would not have any effect in the model, but here, input capacitances have to be add up. Summarized, a parasitic membrane capacitor of estimated 150 fF has to be added to the membrane capacitor.

3.5 Adaptation

After some parameter transformation, the adaptation term can be implemented straight forward using two OTAs and a capacitor.

3.5.1 Circuit and Theory

Looking at the model Equation 3.1, the adaptation variable w can be identified as a current. The following transformation replaces w by an equivalent adaptation voltage V_w via the conductance a:

$$w = a(V_w - E_l). \tag{3.17}$$

If we include this transformation in Equation 3.2, we retain:

$$-\tau_w \frac{dV_w}{dt} = V_w - V. \tag{3.18}$$

The time constant τ_w can be generated by a capacitance C_w and a conductance g_w. C_w stores V_w

$$-C_w \frac{dV_w}{dt} = g_w(V_w - V). \tag{3.19}$$

This equation can be directly implemented using OTAs (see Schematic shown in Figure 3.12). What is still missing is spike-frequency adaptation. b has to undergo the same transformation:

$$b = aV_b = a\frac{q_b}{C_w} \tag{3.20}$$

V_b is the voltage change on the capacitor C_w necessary to achieve a current change b. Consequently, an enlargement of w by b is equivalent to adding the charge q_b on the capacitor C_w. In the circuit this is done by a short current pulse at each single spike:

$$q_b = t_{\text{fire}} I_{\overline{\text{fire}}}. \tag{3.21}$$

Here t_{fire} is the pulse length of the digital pulse used for spike propagation. $I_{\overline{\text{fire}}}$ is a biasing current. In the circuit (Figure 3.12), this is implemented a current source whose source is cut off by an additional transistor.

In addition to the shown circuit, the implemented circuit has the capability to short cut the membrane voltage and V_w to get a defined initial voltage.

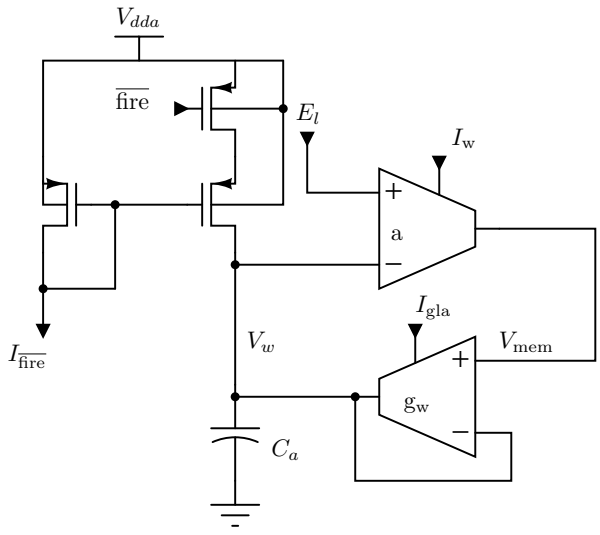


Figure 3.12: Circuit implementing the adaptation term of the AdEx. $\overline{\text{fire}}$ is a digital the digital pulse signaling the spiking of the neuron. All other inputs are biasing voltages and currents.

3.5.2 Real Circuit Behavior

The real circuit is exposed to the imperfections of OTAs shown in Section 3.3.

Saturation limits adaptation

Firstly, the saturation causes the OTA a (Figure 3.12) to output a constant current for large differential input voltages. Consequently, larger values of V_w will not have any more influence. The current pulses at future spikes will not strengthen adaptation anymore. Indeed, the model does not limit the impact of w and the spike frequency will decrease if w is enlarged. Especially for neuron behaviors like Bursting, this can be a problem. In fact, the valid parameter range for Bursting in the model is quite narrow (See 4.4).

As V_w follows the membrane potential, the differential voltage at OTA a is no issue.

Input offset at both OTAs

Secondly, input offset can be a problem in the shown circuit as the offset of both circuits is added. In addition, the potential E_l is shared between the adaptation and the leakage circuit adding another offset. There can easily be a static voltage difference of 20 mV between the membrane Voltage and V_w. If OTA a in the adaptation term and the leakage OTA have opposing offsets, the total difference would be 60 mV, assuming that the membrane potential has been pulled to E_l by the leakage term. Due to parameter sharing, these offsets can not be removed by calibration. A direct solution of this issue is to remove the E_l parameter sharing which has been done for the next chip revision[2].

Small currents

The time constant τ_w for the adaptation term is the largest of the model, so the OTA a of the term has to be operated with very low biases. At the beginning of this thesis, this was expected to be a potential problem. Nevertheless, those solicitudes have not been confirmed as the circuit performs good for small time constants.

[2]HICANN v3

3.5.3 Conclusion

A direct translation

The presented adaptation implementation offers a close link to the equations of the AdEx. In contrast to the neuron circuits presented in [45] which are claimed to implement the AdEx by the author in [46], we designed the complete adaptation term including subthreshold adaptation by direct translation from the equations. However, the main limitation of the presented circuit is the limited impact due to the linear range of the OTA a.

3.6 Synaptic Input

Translate current pulses to decaying conductances

The signals, the neuron receives from the synaptic array are short(5 ns pulse length at 200 MHz clock speed excluding short time plasticity, see 2.3.) current pulses, whereas the strength of the efficacy of the synaptic connection is determined by the length and height of these pulses. Furthermore, the temporal shape of the synaptic conductance is generated at the neurons side at the synaptic input circuit. Figure 3.13 presents a schematic of the synaptic input circuitry. The current pulses are integrated using an operational amplifier connected as integrator using a capacitor and a resistor. The integrator's output voltage is translated into a current by an OTA, which is used as bias for a second OTA which generates the synaptic conductance.

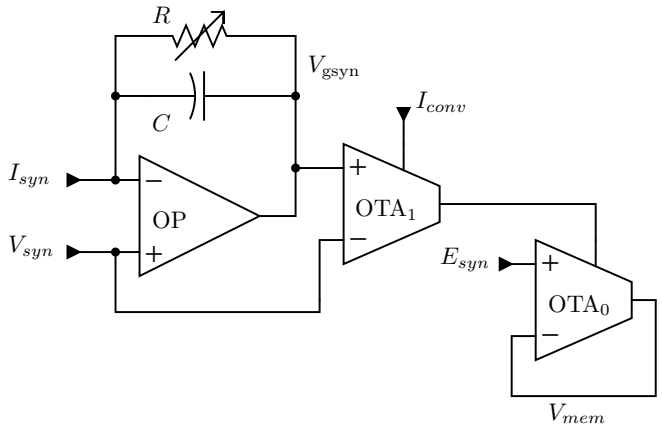


Figure 3.13: Simplified schematic of synaptic input circuit. This circuit is two times present in the neuron to allow different types of synapses. I_{conv} can be used to set the maximum influence of the input. The adjustable resistor R is used for selection of the time constant. The circuit receives rectangle shapes current pulses from the Synapse Array through I_{syn}. Those are transfered into exponential decays by the integrator OP. The OTAs transfer the integrator's output voltage signal into a conductance.

3.6.1 Conductance Shape: Theory

Ideal transformation

The shape of the integrators output signal and so the shape of the generated conductance has a sharp rise during the input current pulse and an exponential decay afterwards. As the rise time is very short in comparison to time constant of the integrator, this model is close to the model for synaptic conductances, used in [5] for instance. Expressed in equations, using an input current pulse height of I_{syn} and pulse length of t_{psyn}, the ideal circuit behaves

according to:

$$V_{gsyn}(t) = \begin{cases} I_{syn}R\left(1 - e^{-\frac{t}{RC}}\right) & : \quad t \leq t_{psyn} \\ I_{syn}R\left(e^{-\frac{t-t_{psyn}}{RC}} - e^{-\frac{t}{RC}}\right) & : \quad t > t_{psyn} \end{cases} \tag{3.22}$$

Scaling with R

Indeed, this suggests the maximum conductance scales with R, which is adjusted to obtain the desired synaptic time constant. Accordingly, this would result in a linear connection between the synaptic weights and the time constant which would be crucial. Moreover, necessary calibration methods would get much more complicated this way.

Looking at the border case

Nevertheless, the minimum synaptic time constant RC aimed at during design is 50 ns. Consequently, if we account the limit $t_{psyn} \ll RC$, we obtain:

$$V_{gsyn}(t) \approx \begin{cases} \frac{I_{syn}t}{C} & : \quad t \leq t_{psyn} \\ \frac{I_{syn}t_{psyn}}{C}e^{-\left(\frac{t-t_{psyn}}{RC}\right)} & : \quad t > t_{psyn} \end{cases} \tag{3.23}$$

Accordingly, the leaky integrator behaves like a real integrator now.

3.6.2 The Resistive Element

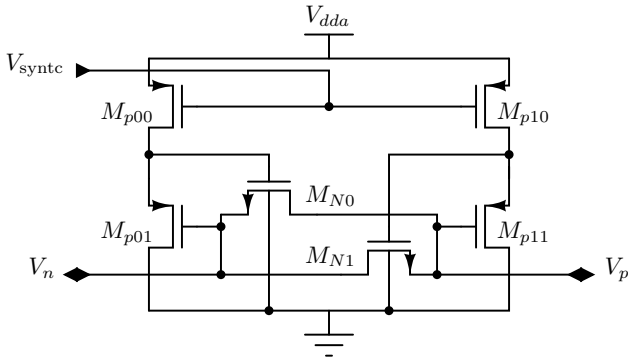


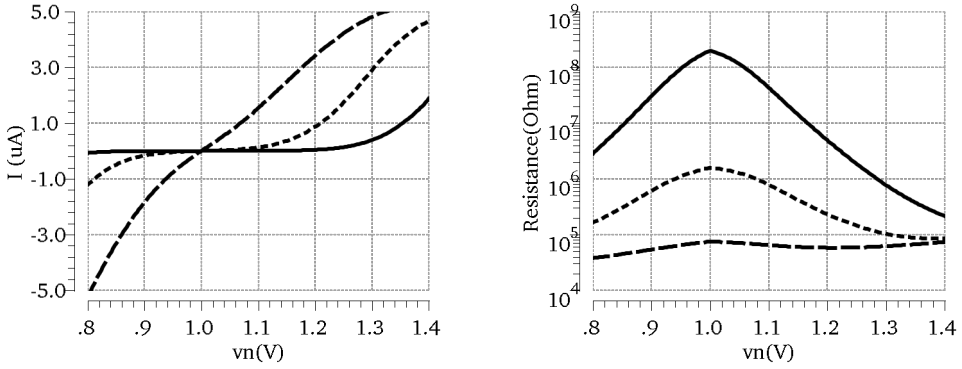
Figure 3.14: Implementation of the adjustable resistor for the synaptic input circuit. Two source follower generate biasing voltages for two parallel NMOS transistors operating in ohmic region. The complete circuit needs to be voltage biased to achieve an adequate value range.

Adjustability needed

However, the resistor R used in Equations 3.22 and 3.23 has to be adjustable, and cannot be build using the available resistors of the process. The circuit shown in Figure 3.14 emulates the resistor. The transistors $M_{N0,1}$ are supposed to be biased in in ohmic region by the source followers on the sides of the schematic. In contrast, although biased resistive, the whole circuit does not have linear current voltage characteristic for the needed parameter range (100 kΩ to 40 MΩ according to the needed synaptic time constants (see 3.12). The gate-source voltage of one of the two NMOS $M_{N0,1}$ is swept during operation

No linear characteristic

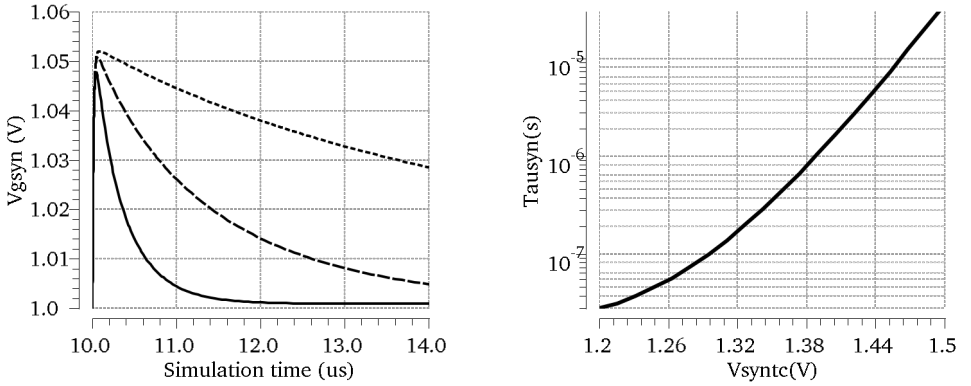
Simulation results can be found in Figure 3.15. The resistance curve for the upper biasing voltages are nearly symmetric, as the circuit switches between dominating M_{N1} and M_{M0} when V_n is crossing $V_p = 1$ V. The lower curve is not symmetric.The gate-source voltage of M_{N0} gets too close to V_{dda} and the left source follower stops operation. Indeed, the resistance change in relation to V_n is drastic. Accordingly, the circuit cannot be talked of as a real linear resistor. Instead, average values or the resulting time constants of the complete circuit have to be used for characterization.



(a) Cross current **(b)** Resistance

Figure 3.15: Typical DC simulation of resistive element shown in Figure 3.14; V_P is kept at 1 V while V_N is swept for different for 1.2 (long dashes), 1.35 (short dashes), and 1.5 V V_{syntc} (solid)

3.6.3 Conductance Shape: Simulated Circuit



(a) V_{gsyn} for 1.25 (solid), 1.35 (long dashes) and 1.45 V V_{syntc} (short dashes)

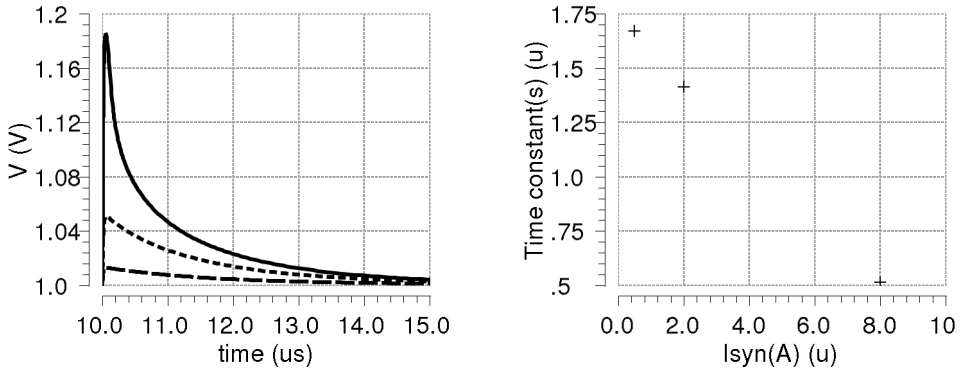
(b) Time constant τ_{syn} when sweeping V_{syntc}.The time constant has been obtained as the time between the peak and the decrease to $1/e$.

Figure 3.16: Transient simulation: time constant and V_{gsyn}. The circuit is stimulated by 5 ns current pulse of 2 µA at 10 µs.

Height nearly independent from time constant

Figure 3.16 shows the conductance equivalent internal voltage V_{gsyn} for different values of V_{syntc} and resulting time constant. As predicted by Equation 3.23, the maximum peak hight has only small dependency on the time constant. However, things will get worse if longer pulses are used. The relationship between the time constant and V_{syntc} is roughly exponential, as the resistive circuit needs to be biased sub threshold or at least close to the threshold to achieve the needed time constants. Accordingly, the parameter range of this voltage is narrow and calibration of this parameter is a challenge in comparison to other parameters. The voltage bias was necessary to achieve a proper time constant range. Indeed, direct current biasing would result in a square root relationship between parameter and time

constant.



(a) V_{gsyn} trace. Stimulus: long dashes: 500 nA, short dashes: 2 µA, and solid 8 µA

(b) Resulting time constants obtained as the peak to $1/e$ decrease time in figure left

Figure 3.17: Transient simulation: The circuit is stimulated by a 5 ns current pulse of different values at 10 µs, V_{syntc} is kept at 1.4 V.

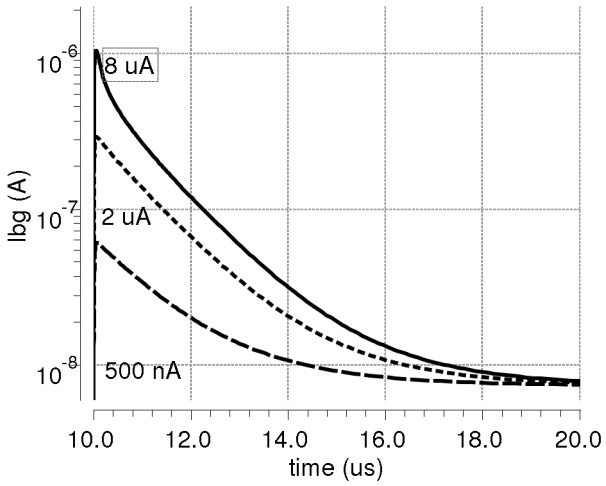


Figure 3.18: Transient simulation: Biasing current I_{bg}. The circuit is stimulated by a 5 ns current pulse of different values at 10 µs; V_{syntc} is kept at 1.4 V.

Smaller time constants for larger pulses

Looking at the dependency between the time constant and the size of the input current pulse, which is equivalent to the synaptic weight, the imperfection of the resistive element is clearly visible (Figure 3.17 and Figure 3.19). As the resistance is smaller for larger values of V_{gsyn}, the time constant of V_{syntc} is decreasing in a similar way for larger input currents.

Fitting an exponential is more accurate

The time constants obtained in (Figure 3.17 and Figure 3.19 differ due to the different methods. The method used in Figure 3.17 only looks at the first part or the curve. Indeed, the real time constant is a function of V_{gsyn} and so this method is inaccurate. Although the method of Figure 3.19 also assumes a constant time constant, the complete curve is taken into account for the fit.

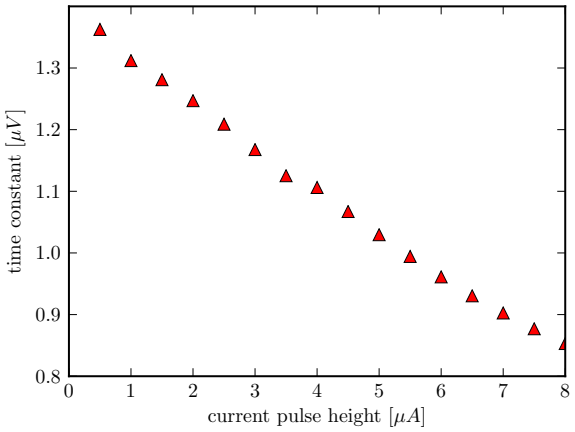


Figure 3.19: Time constant resulting from an least square fit of an exponential to the shape of I_{bg}.

Fits on PSPs achieve better results

Nevertheless, for calibration of the circuit, Marc-Olivier Schwartz is using a different way of obtaining the time constant. Instead of directly measuring the decay of V_{gsyn} which is not directly available for measurement in the real circuit, the post synaptic potential is used. An alpha function can be fitted on this potential and the time constants for the synaptic input can be directly extracted. Indeed, this method results in a much smaller dependency between the synaptic time constant and the synaptic weight. Results can be found in Figure 3.20.

Concluded, the impact of the imperfection of the resistive element is smaller than the voltage dependency of the resistance suggests. In addition, a variation of roughly less than 20 % from the mean, as obtained in Figure 3.20 is not too bad for an analog implementation.

3.6.4 Weight Saturation

Weights are limited by R and saturation of the OTAs

The resistance of the resistive element decreases drastically with in larger voltage differences. In addition, Equation 3.22 suggest a smaller impact of the current pulse for smaller time constants. Moreover, OTA_1 in Figure 3.13 will saturate for voltage differences larger than roughly 200 mV. Consequently, there is a maximum weight or current pulse height for a linear weight conductance relationship. Indeed, looking at Figure 3.17a), the peak for 8 µA is already smaller than 4 times the peak for 2 µA.

No additional effect on membrane above maximum weight

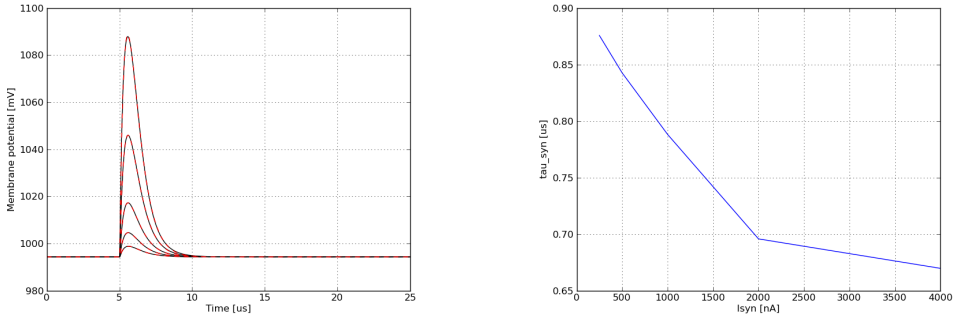
Figure 3.21 expresses weight saturation in explicit way, as higher currents are used here. For a pulse of 16 µA the peak in V_{gsyn} is still close to twice the peak of the 8 µA peak. Nevertheless, at I_{bg} the difference is already smaller as the linear range of OTA_1 is left. Additionally, the peak is suffering a small decay constant. Accordingly, the curves for 8 µA and 16 µA are nearly similar after 200 ns. However, hardly any effect of the different weights can be investigated looking at the shape of the PSP. Both pulses have the same impact on the membrane voltage.

Miss-match of parameter ranges

In the actual HICANN version 2 the parameter range of the pulse height is far larger than 8 µA. In fact, it is designed for values between 400 nA and 160 µA. In addition, the clock speed during first experiments has always been set to 100 MHz, shrinking the usable input current pulse height even further. Consequently measurements done in Kononov's diploma thesis [70] achieve linearity only close to minimal biasing parameters in the floating-gate array. The next HICANN version is planned to counterbalance the parameter range miss-matches.

Single conductance does not scale with neuron size

When neurons are combined to larger neurons to enhance the number of individual synaptic



(a) Alpha functions fitted to PSPs

(b) Dependency between synaptic time constant and height of current pulse.

Figure 3.20: Transient simulation results. PSP fits and the corresponding time constants in dependency to the height of the stimulation current pulse. The figure has been created by Marc-Olivier Schwartz.

connections, the maximum possible synaptic conductance of a single connection does not scale with the neuron size if only one synaptic input circuit is used for a single connection. However, the total synaptic conductance scales as the number of synaptic inputs scales with the neuron number. Furthermore, it is possible to connect one post-synaptic neuron to several synaptic inputs of the pre-synaptic neuron to achieve a larger impact of the connection.

3.6.5 Delays

Each computation step adds delay

The three analog computation steps in the synaptic input induce a delay between the onset of the pulse from the synapse array and the actual rising of the PSP. Typical simulation results suggest a total delay of maximal 25 ns. Firstly, the integrating OP introduces delay. The rise of V_{gsyn} starts 8 ns after the onset of stimulus, while the maximum peak is reached roughly after 50 ns. Subsequently, the onset of the I_{bg} pulse is 10 ns with a maximum after 50 ns. The greatest delay can be observed actual output current of OTA_2 which is equivalent to the total delay. The beginning can be more than 25 ns after the stimulus starts, depending on weight. The bias current of OTA_2 has to charge internal capacitances of the OTA before it can operate. Hence a small biasing current creates a larger delay. The total delay shrinks down to 19 ns for larger weights.

25 ns hardware delay is equivalent to 250 µs in biological time scales at 10^4 speedup. Nevertheless, the total delay between the synapse driver output and the onset of the PSP has been measured to be 60 ns by Andreas Grübl and Alexander Kononov. This adds up to the digital delay of at least 60 ns. Consequently, the total delay in biological time scales gets in the region of 1 ms. Accordingly, the delay starts to get biological relevance.

3.6.6 Conclusion

The presented circuitry is a straight forward implementation using analog filter techniques. However, the imperfections of the resistive element and the linear range the OTAs limit the direct translation between between model and circuit. Nevertheless, it is possible to fit alpha function shaped post synaptic potentials onto the membrane response generated by the circuit. The final weight dependency of the τ_{syn} is not desirable. However, it remains

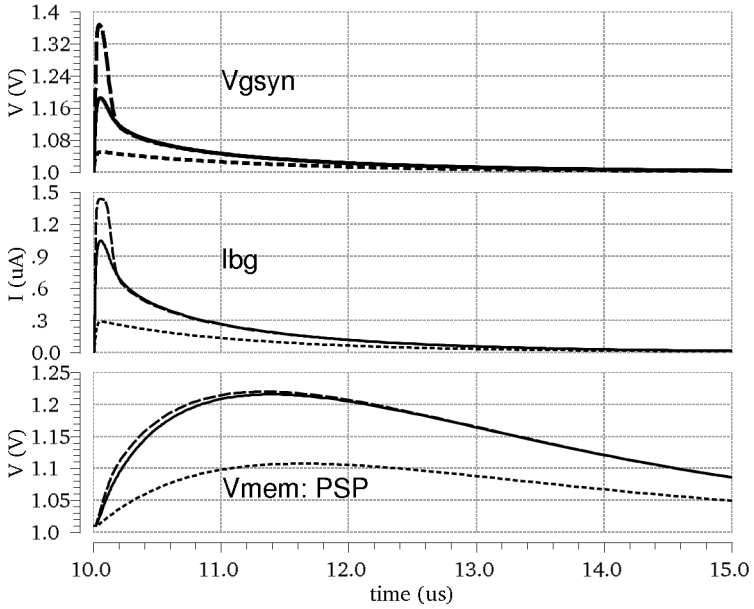


Figure 3.21: Transient simulation: Conductance equivalent measures V_{gsyn} and I_{bg} and membrane voltage of idealized neuron with 2 pF capacitance and 1 µS leakage conductance pulling to a leakage potential of 1 V. The circuit is stimulated by a 5 ns current pulse of 2 µA(short dashes), 8 µA(solid) or 16 µA(long dashes) at 10 µs. Due to weight saturation the membrane voltage impact of the latter pulses is close to similar

in the margins I would expect from an analog implementation. The small parameter range of V_{syntc} is expected to be a challenge in calibration. Here, an improvement remains on the wish-list for future implementations.

3.7 Exponential Term

The exponential term is the last model circuit designed for the implementation of the continuous equations of the AdEx. In contrast to the circuits discussed before its operation is not based on OTAs.

3.7.1 Circuit Principle

Subthreshold characteristic as exponential function

The basic circuit of the term can be found in Figure 3.22. Exponential feedback is realized by sweeping the subthreshold characteristic of M_0. The current characteristic, which needs to be implemented as extracted from Equation remeqn:adexvis:

$$I_{\mathrm{exp}} = -g_{\mathrm{l}}\Delta_{\mathrm{t}}\exp\left(\frac{\mathrm{V}-\mathrm{V_t}}{\Delta_{\mathrm{t}}}\right) \tag{3.24}$$

Normalization

Here, the pre-factor is normalizing the model to a minimum of the V-nullcline (see Figure 3.2). It can be hidden by adding a constant to V_t in the equation. With constant c and $V_{\mathrm{t}}' = V_{\mathrm{t}} + c$, we retain:

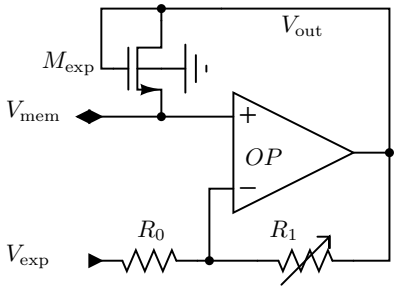


Figure 3.22: Simplified schematic of exponential term circuit. The subthreshold characteristic of M_0 creates the exponential dependency.

$$I_{\text{exp}} = -\exp\left(\frac{\mathrm{V} - \mathrm{V}'_{\mathrm{t}}}{\Delta_{\mathrm{t}}}\right) \tag{3.25}$$

Consequently, we eliminated the parameter g_{l}. The same transformation can be used to remove any constant factor in this equation.

Simplified, the subthreshold of the drain-source current I_{DS} of a MOSFET is:

$$I_{\text{DS}} = I_{d0}\exp\left(\frac{\mathrm{V_{GS}} - \mathrm{V_{th}}}{\mathrm{n u_t}}\right) = \mathrm{I_{d0th}}\exp\left(\frac{\mathrm{V_{GS}}}{\mathrm{n u_t}}\right) \tag{3.26}$$

Here, V_{th} is the threshold voltage of the MOSFET device and $u_t = \frac{k_b T}{q} \approx 25\,\text{mV}$ is the thermal voltage. I_{d0} is a constant. I_{d0th} includes the threshold voltage. We assumed a drain-source voltage much large than u_{t} and hardly any leakage. n is called subthreshold swing parameter and depends on the channel length and state density in the gate-oxide[71]. The complete subthreshold model used in simulations can be found in [71].

The operational amplifier in Figure 3.22 keeps the voltage at its inputs at the same potential. Furthermore, the resistors R_1 and R_2 form a voltage divider. This way we can calculate V_{out}:

$$V_{\text{mem}} - V_{\text{exp}} = (V_{\text{out}} - V_{\text{exp}})\left(\frac{R_2}{R_1 + R_2}\right) \tag{3.27}$$

$$\Leftrightarrow\ V_{\text{out}} = (V_{\text{mem}} - V_{\text{exp}})\left(\frac{R_1 + R_2}{R_2}\right) + V_{\text{exp}}. \tag{3.28}$$

To get to V_{GS} of M_0 we have to subtract V_{mem}:

$$V_{\text{GS}} = V_{\text{out}} - V_{\text{mem}} = \frac{R_1}{R_2}(V_{\text{mem}} - V_{\text{exp}}) \tag{3.29}$$

This can be included in Equation 3.26:

$$I_{\text{DS}} = I_{d0th}\exp\left(\frac{\mathrm{V_{mem}} - \mathrm{V_{exp}}}{\mathrm{n u_t}}\frac{\mathrm{R_1}}{\mathrm{R_2}}\right) \tag{3.30}$$

Accordingly, Equation 3.25 can be emulated if the voltage V_{exp} and the relationship between R_1 and R_2 are adjustable parameters.

3.7.2 Voltage Divider

Adjustable, large resistor needed

The critical elements in the exponential term circuit are the resistors R_1 and R_2. Indeed, to achieve a proper efficiency, the current through the resistors should be much smaller than the current through M_{exp} In contrast, large resistors are much less area efficient. In Addition, noise will be much larger for larger resistors. Furthermore, as if there was not enough trouble, at least one of these resistors needs to be adjustable.

The influence of the exponential term is small if the membrane voltage is below V_{t}. Consequently, exact exponential behavior is only needed above this threshold.

Channel length modulation implements the resistor

The circuit shown in Figure 3.23 implements the correct dependency between V_- (equivalent to the membrane voltage V) and $V_{\text{out}} - V_-$ using the channel length modulation of the transistor M_1. Obviously, channel length modulation can only be used if M_1 is operated in saturation, so a minimum voltage difference is necessary.

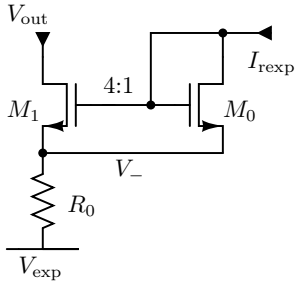


Figure 3.23: Circuit implementing the voltage divider of the exp term. The terminals V_- and V_{out} connect to the OP in Figure 3.22. The parameters I_{exp} and V_{exp} represent $V_{\text{t}'}$ and Δ_{t}

Using the equation for a MOSFET in saturation region (see 1.3.1) we obtain:

$$V_- - V_{\text{exp}} = R_0 I_{\text{R0}} = R_0 \left(I_{\text{M0}} + I_{\text{M1}}\right) \tag{3.31}$$

$$= R_0 I_{\text{rexp}} \left(1 + 4 + 4\lambda(V_{\text{out}} - V_-)\right). \tag{3.32}$$

In Figure 3.22 the gate-source voltage of M_{exp} is $V_{\text{out}} - V_-$ as the OP drives V_- to V_{mem}. Consequently, we have to regard this difference:

$$V_{\text{out}} - V_- = \left(\frac{V_- - V_{\text{exp}}}{R_0 I_{\text{rexp}}} - 5\right) \cdot \frac{1}{4\lambda} \tag{3.33}$$

$$= \frac{V_- - V_{\text{exp}} - 5R_0 I_{\text{exp}}}{R_0 I_{\text{rexp}} 4\lambda} \tag{3.34}$$

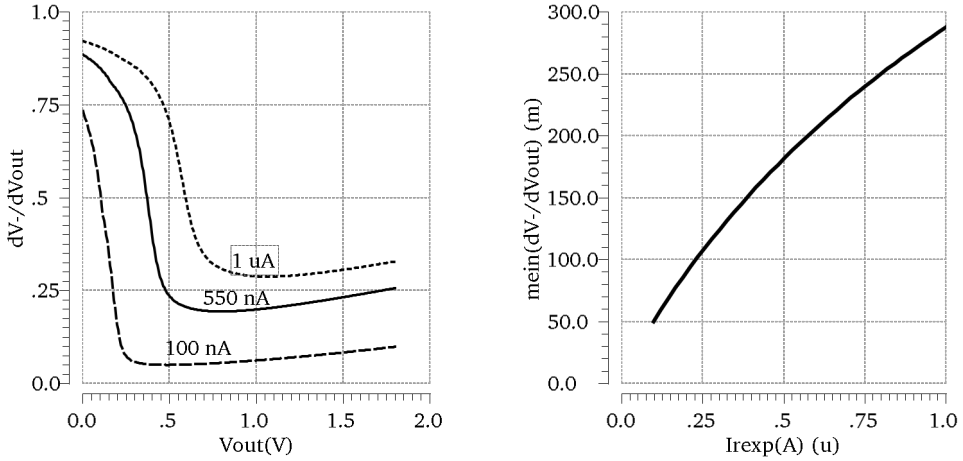
Consequently, Δ_{t} can be adjusted by the parameter I_{rexp}, while V_{t} needs both, I_{rexp} and V_{exp}.

Looking at Equation 3.26 we have to ad a factor u_{t}.

$$\Delta_{\text{t}} = u_{\text{t}} \left(\frac{d\left(V_{\text{out}} - V_-\right)}{dV_-}\right)^{-1} \tag{3.35}$$

$$= u_{\text{t}} R_0 I_{\text{rexp}} 4\lambda \tag{3.36}$$

Indeed, the direct dependency to u_t results in a direct dependency to the temperature.



(a) Derivative for different values of I_{rexp} **(b)** Minimum of derivative

Figure 3.24: Typical DC-simulation of voltage divider of exponential circuit (Figure 3.23). V_{out} is swept

Figure 3.24 shows the derivative of V_- in V_{out} which is similar to $R_0 I_{\text{rexp}} 4\lambda/(1+R_0 I_{\text{rexp}} 4\lambda)$. Here the denominator is close to one. Once the saturation regime is reached, the derivative changes only little for voltages. However, in this case, a smaller linear range would have be sufficient. The valid range of the difference between V_{out} and V_{mem} is shrunk as we are mapping on an exponential. Larger values of the derivative shrink the operating range as the value of V_- is enlarged. The second plot shows the minimum of the derivative in a sweep of I_{rexp}. A dependency close to linear can be observed.

3.7.3 Complete Circuit Simulation

When simulating the complete circuit, the current through transistor M_{exp} has to be observed. Figure 3.25 shows some simulation results with a sweep of V_{mem}.

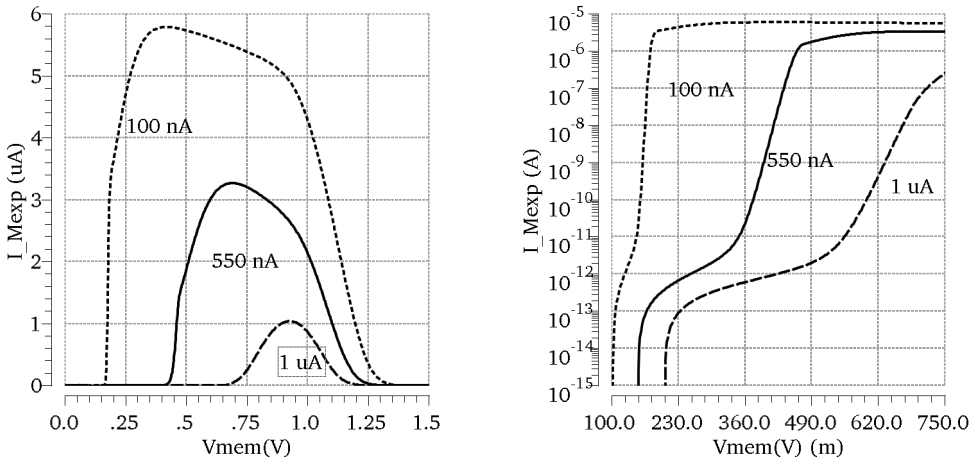
Smaller impact for higher biasing currents

In the curve for 100 nA and 550 nA, a bend can be identified. At this point, V_{out} cannot be enlarged enough by the operational amplifier. Consequently, V_- cannot follow V_{mem} and the rise of the gate source voltage of M_{exp} collapses. Subsequently, for larger membrane voltages, the current output of the circuit decreases again as the body effect of M_{exp} shrinks its conductance. Additionally, the amplifier reaches its power rails. For larger biasing current, the collapse happens earlier as the amplifier has to source the mirrored biasing current. Indeed, adding $4 \cdot 550$ nA to the 550 nA curve results in similar current ranges as the 100 nA curve. For larger biasing values, the output of the exponential circuit vanishes.

The point of the exponential rise is shifted a V_- is shifted for different values of I_{rexp} (see Figure 3.26). At V_{cross} the current through M_{exp} reaches $I_{\text{cross}} = 100$ nA. This visualizes the dependency V_t on I_{exp}.

Looking at Equation 3.30, Δ_t can be obtained from I_{Mexp} by

$$\Delta_t = \frac{I_{\text{Mexp}}}{\frac{dI_{\text{Mexp}}}{dV_{\text{mem}}}}, \tag{3.37}$$



(a) linear scale **(b)** log scale

Figure 3.25: Current through M_{exp} obtained by DC-simulation of exponential circuit(Figure 3.22) for different values of I_{rexp}. The membrane Voltage is swept. The exponential rising current can be observed in the log scale plot(right). V_{exp} is set to 100 mV here.

if I_{Mexp} is measures in the exponential rise. This calculation has been done in Figure 3.26. As Δ_{t} does only need to be scaled with the voltage scaling factor an not with the time scaling factor, a smaller value range is sufficient.

An approximation of V_{cross} can be done through the voltage drop at R_0 from Figure 3.22:

$$V_{\text{cross}} \approx V_{\text{exp}} + R_0 \cdot 5 \cdot I_{\text{rex}} \tag{3.38}$$

The value of the resistor is 121.5 kΩ. The factor 5 is created by the current mirror in Figure 3.23.

In summary, the following translation between the parameters V_{exp} and I_{rexp} can be suggested:

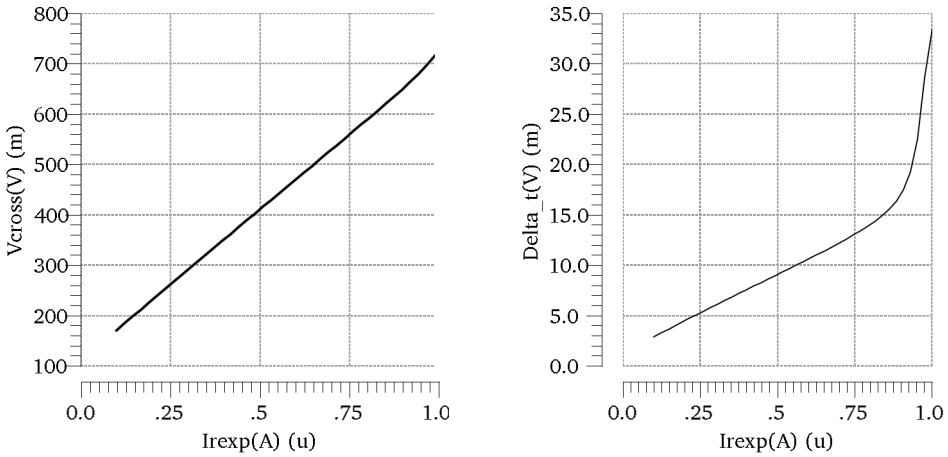
1. Choose I_{rexp} according to the given Δ_{t}
2. Calculate the necessary V_{cross} (compare Equation 3.1:

$$I_{\text{cross}}\exp\left(\frac{\mathrm{V}-\mathrm{V}_{\text{cross}}}{\Delta_{\text{t}}}\right) = \Delta_{\text{t}}\mathrm{g_l}\exp\left(\frac{\mathrm{V}-\mathrm{V_t}}{\Delta_{\text{t}}}\right) \tag{3.39}$$

$$\Rightarrow\ V_{\text{cross}} = \Delta_{\text{t}} ln\left(\frac{I_{\text{cross}}}{\Delta_{\text{t}} g_{\text{t}}}\right) + V_{\text{t}} \tag{3.40}$$

3. Shift V_{exp} to reach the necessary V_{cross}

Indeed, the lower bound of V_{t} is limited for larger values of Δ_{t}. As V_{t} is located close to the upper limit of the neurons operation regime in the circuit implementation, this is not an issue.



(a) Membrane voltage V_{cross} at which the current through M_{exp} (see Figure 3.22) reaches 100 nA. V_{exp} is 100 mV.

(b) Δ_{t} calculated at V_{cross}

Figure 3.26: Same simulation as in Figure 3.25 with a sweep of the bias current I_{exp}

3.7.4 Conclusion

I have shown an implementation of the exponential term which allows adjustment of V_{t} and Δ_{t}. Indeed, both parameters are linked. However, the proposed translation scheme allows obtaining necessary hardware parameters from biological parameters.

3.8 Spike Detection

Spike detection, or the detection of crossing of Θ is done by a comparator circuit called Spike Amp. As a basis of this circuit, the comparator of the SPIKEY chip[64] has been used. Nevertheless, some changes have been applied. A schematic can be found in Figure 3.27

An OTA as comparator

Basically, this circuit is a differential OTA with an additional feedback. As long as V_{mem} is far below Θ, only the left branch of the pair is conducting. Consequently, the voltage at node N1 is close to V_{dda} and M_4 is high ohmic. At this point, we assume node N3 to be at 1.8 V which is the static case of the circuit. When V_{mem} approaches Θ, the voltage at N1 decreases and M_4 starts conducting. Subsequently, the voltage level at node N2 is increased and finally the feedback through M_0 gets active pulling N1 to ground. Node is pulled close V_{dda} accordingly.

A spike has been triggered and driven to the output by the inverter I_1. The negative polarity is necessary for spike routing (see Section 3.10).

Delaying the signal

The two buffers are implementing a delayed version of the spike signal which is feed back into the circuit. Each is build off two asymmetric inverters with either the PMOS or the NMOS half with a long channel length to achieve a larger delay. The delay is critical as is used for STDP for instance.

After the delay, N3 is at ground. M_1 is switched high ohmic cutting the feedback. In addition, M_2 is pulling N1 to the power rail.

Preventing multiple spikes

Without the latter transistor which has been added, two things could happen. Firstly,

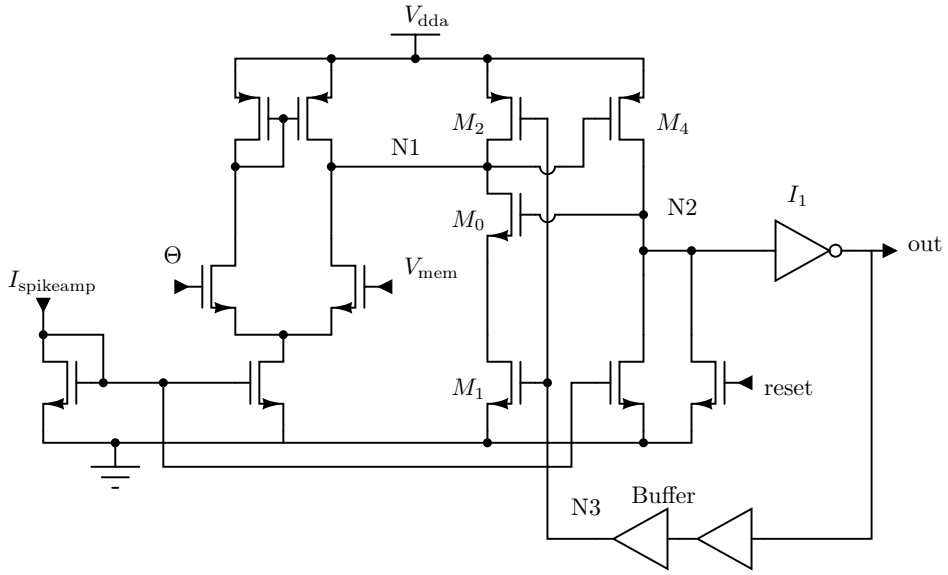


Figure 3.27: Spike amp circuit.

if V_{mem} is still above the threshold, N1 would still be pulled by the differential pair. Consequently, the length of the spike signal would not be determined by the delay created by the two buffers. Secondly, if the even if the membrane voltage is already low enough the impedance of the differential pair is high in comparison to M_2. N1 would only have a small slope. This could cause another activation of the feedback causing two spike signals out of one crossing of Θ

The reset signal is a global neuron signal for each chip half disabling the firing of neurons.

3.9 Resetting

The reset mechanism is responsible for setting the membrane voltage to the reset potential V_{reset} after a detected spike. In contrast to the reset of the model equation, the reset in a circuit cannot break continuity.

When a spike is detected, the reset mechanism pulls the membrane to V_{reset} by a globally adjustable current I_{reset} for an adjustable time frame. Both, the globally adjustable current and adjustable time frame are extensions and are not included in the standard AdEx model. The latter extension should implement a refractory period by disabling spiking. Indeed, the biological definition of the refractory period is different. The biological "absolute refractory period" [72] would disable channels necessary for spiking, preventing any action potential. On the other hand, "relative refractory period"[72], just disables some channels and inhibits spiking. Consequently, the biological relevance of the implemented refractory period is weak. Nevertheless, some models like "Neural Sampling" [73, 74] rely on a dead time after spiking. Accordingly its implementation was good intuition.

Implementing refractoriness

The adjustable current has been implemented to allow broader action potentials. Those broader action potentials can be observed in dendrites[75]. However, the implementation via the slope during the pull-down to the reset potential is a very rough translation.

Broad action potentials

Figure 3.28 shows a simplified schematic of the circuit. M_{C} is connected to work as a

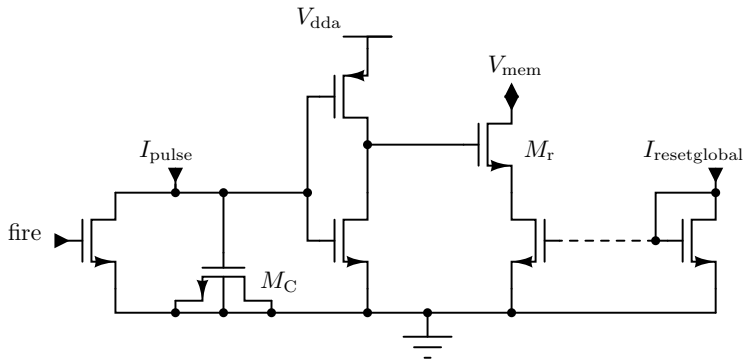


Figure 3.28

capacitor. It is discharged when a firing signal is received. Consequently a current pulls the membrane to the reset voltage V_{reset} through M_{r} This current itself can be adjusted by a distributed current mirror which gets its bias from the floating gate array.

The current I_{pulse} charges M_{C} and once the voltage is high enough the inverter is triggered and resetting stops.

3.10 Neuron Connectivity

Routing firing signal and membrane voltage

As described in 2.3 single neurons can be interconnected to combine their synaptic inputs, allowing neurons with up to 14 thousand dedicated synaptic input connections. This interconnection includes two signals: the membrane voltage V and the firing signal. The membrane voltage connections are directly realized using transmission gates. The firing signal cannot be routed this way, as it is a digital signal. Voltage drop on the transmission gates would inhibit triggering of other signals after a chain of neurons. Consequently buffers are needed here. Figure 3.29 visualizes how neurons can be connected. To achieve a structure with fewer buffers in the fire lines, respectively two neurons form a pair which can be interconnected by a transmission gates. Additionally all connections between top and bottom are realized using transmission gates. This way, buffers are only needed if larger neurons have to be constructed.

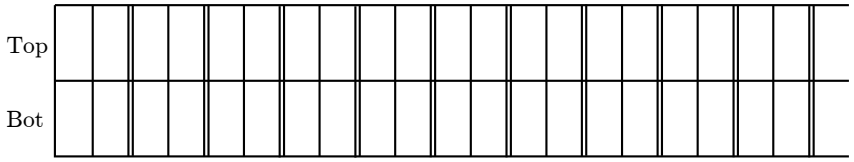


Figure 3.29: Simplified neuron connectivity overview. Each rectangle corresponds to one neuron. There are two rows - on in the each half of the HICANN. Membrane voltages of neighbouring neurons can be interconnected using transmission gates. The firing signal is routed through transmission gates(single lines) or switchable buffers

3.10.1 Circuit

Transmission gates and buffers

Firing is directed

The implementation of the fire routing can be found in Figure 3.30 which displays the Spiking/Connection circuit (see Figure 3.4) of a neuron pair. Membrane routing is not displayed here as it only uses transmission-gates. A firing signal $fire_{top/bot}^{<i>}$ can have four different sources: the neuron itself, or each of its 3 neighbours. Only one source is allowed to be active. Except for the transmission gates to the other chip half, a neuron pair chooses locally, where to receive spikes from by activating or deactivating input buffers to neighbouring neurons. A pair itself outputs the inverted spiking signal to the input buffer of its neighbours on the left and right. The same signal is inverted again and fed back into the synapse array for triggering STDP.

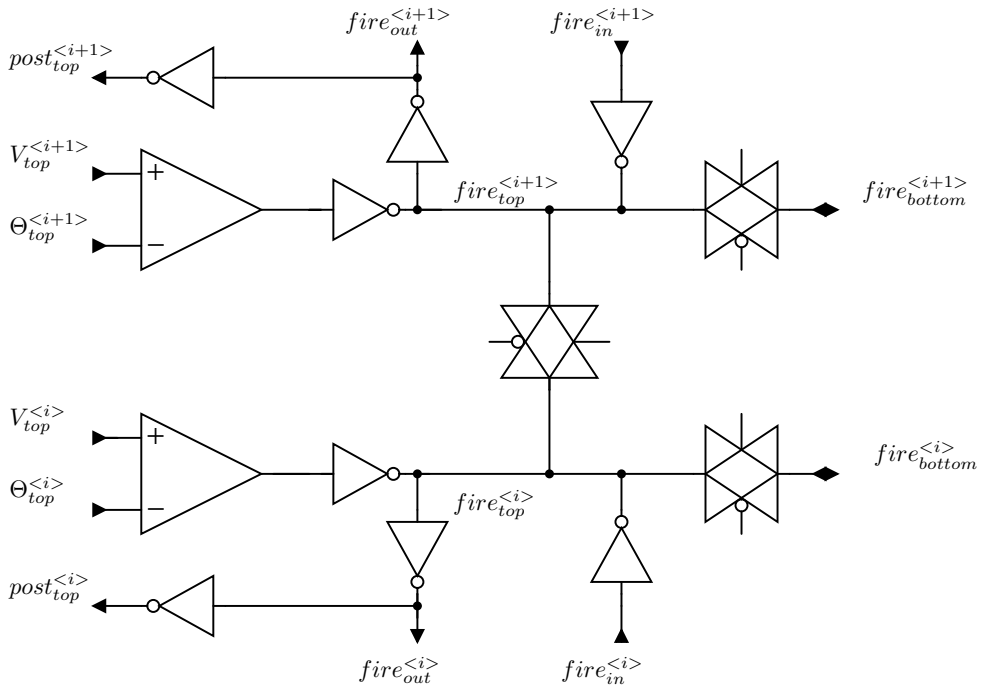


Figure 3.30: Routing of spike signals in a neuron pair in the top chip half. The operational amplifier is the spike detection comparator of a denmem. Grey inverters can be switched on or off. Connectivity between bottom and top and pair internal is realized by transmission gates; other connections need to be buffered. Connections to neighbour pairs are done by connecting $fire_{in}^{<j-1>}$ to $fire_{out}^{<j>}$ and $fire_{in}^{<j>}$ to $fire_{out}^{<j-1>}$, where j is any even neuron number in this case.

3.10.2 Simulation

Two simulations have been performed here to show the natural imperfections of the circuit implementation. The first observes the conductance of the transmission gates, while the second concentrated on the delay of the fire signal.

A real switch, or a real transmission gate always has a finite resistance in open and closed mode. Table 3.1 shows the conductance values of the closed and open transmission gates used for spike signal and membrane routing. The implementation goal was 1000 µS here.

This conductance already needs quite large transmission gates indeed. Nevertheless, work done for the multi compartment neuron has shown that a conductance of this size is in the order of magnitude of inter compartment conductances. This suggest a biological relevance of the neuron interconnection in the single compartment implementation.

corner	Conductance			
	open[µS]		closed[pS]	
	average	σ	average	σ
typical	846.8	11.82	5.164	0.276
fast	1106	14.34	11.86	0.850
fnsp	931.5	13.64	6.267	0.421
slow	607.0	11.38	2.999	0.076
snfp	766.9	12.15	5.114	0.247

Table 3.1: Open and closed conductance of transmission gate responsible for interconnecting denmems. The results have been obtained by a Monte-Carlo-Simulation with 100 samples for each corner for currents between 0.1 µA and 20 µA for the open case and a voltages of 1.8 V for the closed case

Biological relevance of delays

The propagation delay is observed in Table 3.2. A delay of 1 ns is equivalent to a biological delay of 10 µs at time scaling factor 10^4. When connecting optimum maximum size neurons including 64 neuron circuits, the delay of $fire_{bot}^{<i-1>}$ has to be applied 15 times to estimate the maximum delay. Adding the delay inside the first neuron pair, this results in roughly 19 ns technical or 190 µs biological delay with a time scaling factor of 10^4. However, at 10^5 it would be 1.9 ms. Indeed, a delay of this size has biological relevance. Nevertheless, spike propagation delays introduced by digital buffering and the synaptic input out rule this delay. In addition, the usual operation mode of the chip is time scaling factor 10^4 reducing the biological impact of the delays.

Position	Delay [ps]	σ [ps]
$fire_{top}^{<i-1>}$	840.2	3.356
$fire_{bot}^{<i-1>}$	1229	3.620
$fire_{top}^{<i-2>}$ wc	1679	4.572
$post_{top}^{<i>}$	1773	11.39
$post_{top}^{<i>}$ end	2203	11.23

Table 3.2: Propagation delay of internal digital continuous time spike signal at different positions in the circuit (see Figure 3.30). The input signal is $fire_{top}^{<i>}$ For $fire_{top}^{<i-2>}$, the signal is not directly routed, but connected through the bottom pair - three transmission gates are needed. The point $post_{top}^{<i>}end$ is measured at the end of the synapse array. Values are obtained with a typical corner Monte-Carlo Simulation with 100 samples.

3.11 Readout and Stimulation

Membrane voltage readout and current stimulation

For single neuron measurements, each neuron can be stimulated by a programmable current source. Furthermore, the membrane voltage V_{mem} of a neuron can be read out. The responsible circuit is called In/Out in Figure 3.4. Two pairs consisting of a current stimulation line

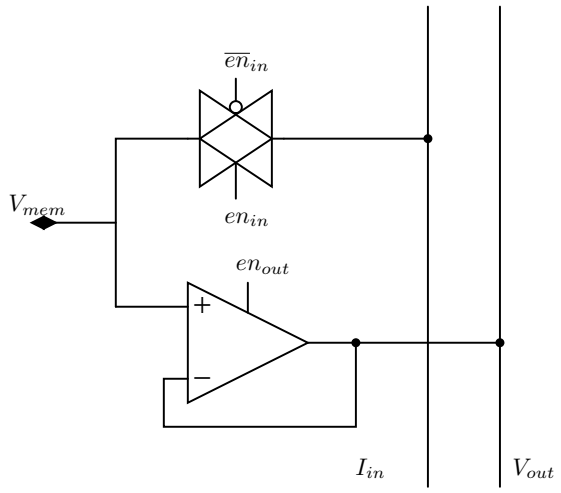


Figure 3.31: Simplified schematic of the input and output module of a neuron. The lines V_{out} and I_{in} are connected to every second neuron to enable readout and stimulation of neighbours. Input and output can be enabled or disabled - for details see [61].

and an output line are routed through each of the two neuron rows. Every second neuron can be connected to every second line. Consequently, it is possible to stimulate and readout two neurons in each chip half. The total number of readouts is limited to two by the two output amplifiers of the HICANN chip.

Figure 3.31 gives an overview of the In/Out circuit. Switching of the output is done by setting the OP high ohmic. A neuron can be selected for current stimulation by closing the transmission-gate.

3.11.1 Analog Readout

Switchable OP

The output amplifier of a neuron is a standard CMOS Miller OTA (see [23] for instance). It is identical to the operational amplifier used for the neurons of the SPIKEY[64] microchip. To allow a switch-off of the circuit, it is equipped with a transmission gate at the output to be able to set it high ohmic. Additionally, the biasing of the OP is switched.

A typical simulation of the amplifier with realistic output load including line impedances and periphery results in an bandwidth of 18.5 MHz at the input of the chip output amplifier.

3.11.2 Current Stimulation

A simplified schematic of the current source used for neuron stimulation can be found in Figure 3.32. The used operational amplifier is the rail-to-rail amplifier developed in [76]. V_{ref} is translated into a current by the voltage drop at the resistor R. This current is mirrored into the neuron.

Programming memory for floating-gates is reused

V_{ref} is generated by the digital-to-analog converter(DAC) of one of the four floating-gate arrays(see Chapter 9). The memory of the digital controller of the array is used to generate a time variating signal. Values in the memory are connected to the DAC one after another. The loop of the memory values can be done for a certain number of repetitions or continuous.

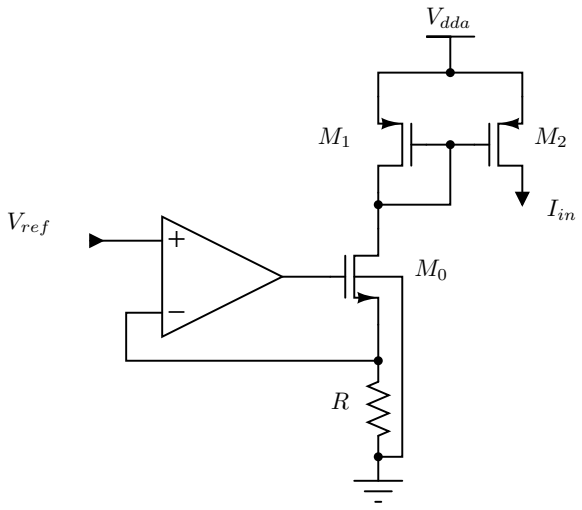


Figure 3.32: Current source used for neuron stimulation. The maximum output current is limited through the gate source voltages of M_0 and M_1

Translation

Linear relationship

The relationship between the DAC output values and the current is linear up to a maximum voltage close to 700 mV below the power supply. The limit is given through the necessary gate-source voltages of the transistors M_0 and M_1. The actual value of the current and the slope of the current is given by the value of the resistor. Resistors are exposed to large variations during production, so the value of the slope is different for each instance of the current source. Table 3.3 gives an overview of different slopes and maximum corners for different production corners.

corner	Current[µA]		Slope[µA/V]	
	At 1 V	Maximum	average	deviation
typical	1.95	2.071	1.95	0.014
fast	2.874	3.083	2.87	0.023
slow	1.408	1.462	1.408	0.009
snfp	1.955	2.088	1.955	0.015
fnsp	1.945	2.041	1.945	0.013

Table 3.3: Output current and slope (to DAC voltage) of current input. The deviation is given here to show the grade of nonlinearity of the characteristic

Settling Time and Maximum Switch Frequency

Maximum frequency is allowed for small changes

Simulations showed that the settling time of the DAC is smaller than the settling of the current output. Accordingly, the latter has to be taken into account for the maximum stimulation frequency. For large deviations between adjacent values, the maximum frequency should be limited to 5 MHz as a settling time of 25 ns is to be expected. In contrast, frequencies up to the maximum of 31.25 MHz are allowed if the digital DAC input value is only

incremented or decremented by one. The current settling is not monotonic - there will be some overshooting (Figure 3.33).

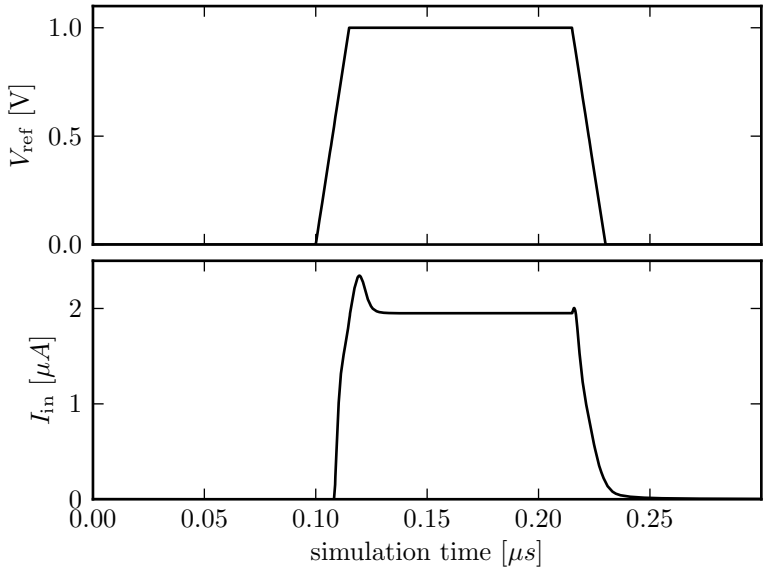


Figure 3.33: Settling and overshooting of current input in typical simulation. The rise and fall time has been chosen to imitate the DAC behavior

3.12 Parameterization

Acceleration

The parameters shown in this chapter differ drastically from the parameters existing in a biological neuron. Indeed, they differ by orders of magnitude. This is do to the fact that the design goal was not a real time neuron. Instead a so-called accelerated neuron emulation has been designed keeping the natural conductances of the used process when operated in strong-inversion. Hence, the dynamics of the implemented neuron are between 10^5 and 10^3 times faster than the biological dynamics.

3.12.1 Biological Parameter Ranges

Parameters extracted from publications

In the design phase it is crucially important to choose a proper parameter range covering biological parameters of common models. For this purpose parameters have been extracted from different publications from Romain Brette and Alain Destexhe. A start point can be given by the parameters directly available in the paper introducing the AdEx([22]). These values can be found in Table 3.4

Leakage

Nevertheless, Brette and Gerstner use [77] as a basis for their model parameters. This paper lists the leakage conductance density for four different types of neurons. For better comparison, leakage time constants C/g_{l} have been calculated during extraction. A variation of one order of magnitude can be observed. The results can be found in Table 3.5. We implement time constants between 5 ms and 50 ms.

Adaptation

The paper [78] by Prospishil et al. has been used for the ranges of the subthreshold adaptation parameter a and the adaptation time constant τ_w. Prospishil et al. map a

Model parameter	Value
C (membrane capacitance)	281 pF
g_{L}	30 nS
E_{L}	-70.6 mV
V_{t}	-50.4 mV
Δ_{t}	2 mV
τ_w	144 ms
a	4 nS
b	0.0805 nA

Table 3.4: Parameters of the AdEx model as presented in Table 1 in [22]. See Equations 3.1 and 3.2

Neuron type	Time constant[ms]
Pyramidal (AdEx [22])	9.30
Cortical pyramidal cells(Destexhe [77])	10
Cortical inter neuron cells(Destexhe [77])	6.67
Thalamus reticular cells(Destexhe [77])	20
Thalamocortical cells (Destexhe [77])	100

Table 3.5: Neuron leakage time constants obtained from [22] and [77]

Hodgkin Huxley type neuron model onto different biological cortical neurons. Those are regular spiking inhibitory and excitatory and fast spiking inhibitory cells. For each type a roughly 10 different individual cells have been fitted. Indeed, the model used in this paper is not an AdEx, but it is possible to transfer the parameters.

Adaptation time constant

The maximum spike-frequency-adaptation time constant $\tau_{\max}$ of [78] ranges from 500 ms to 3 s. Nevertheless, the actual adaptation time constant τ_p in [78] is strongly membrane voltage dependent and reaches much lower values below the spiking threshold. In fact, the adaptation time constant of the AdEx model is most important for subthreshold behavior as the update during a spike is maintained by the parameter b. Calculating τ_p for a membrane voltage results in time constants 5 times smaller than than $\tau_{\max}$. The resulting range for the AdEx parameter τ_w which has been used for design is 100 ms to 600 ms.

Subthreshold adaptation parameter a

The subthreshold adaptation parameter a can be identified with the $\overline{g_{\mathrm{M}}}$ in [78]. Indeed, as the parameter of [78] is conductance densities, it has to transformed to a time constant using the membrane capacitance density of 1 $\mu\mathrm{F/cm^2}$. Accordingly, we retain time constants between 5 ms and 59 ms. In the design, we set out upper limit to 50 ms. As this time constant belongs to the parameter a, we will refer to it as τ_a in the following.

Spike-triggered adaptation

To obtain the range of the spike-triggered adaptation parameter b, the value given in [22] has been chosen as benchmark. Later researches obtain relative b values up to 2.5 times larger([12]). Here relative should denote that the actual comparison in done through b/a as this is the relevant parameter for the circuit implementation. Actual chosen parameter range has been one order of magnitude around the model value [22] excluding scaling of the parameter a. However, during parameter translation in the design phase, voltage scaling (see below in 3.12.2) has been omitted for this parameter. Consequently, the maximum value of b is a factor 5 times smaller than the value from [22] if a voltage scaling factor of 10 is used. Nevertheless, for chip revision HICANN v3, the biasing has been increased by a factor of 5 so [22] can be reached.

Synaptic input

Next, to get a starting point for synaptic parameters, values have been extracted from [79] and [22]. For excitatory synapses, the time constants lie between 2.7 ms and 7.8 ms. Inhibitory synapses have values between 8 ms and 10.5 ms.

For validation, the time constant and conductance ranges have been discussed with the biologist Alain Destexhe in an email conversation.

Relation to leakage conductance

To get the parameters of the synaptic conductances, the total conductances presented in [22] can be taken into account. Here the conductance values are given in terms of the leakage conductance. The maximum total value of the excitatory synaptic conductance is $4/3 \pm 2/3$ times the leakage conductance. The value for the inhibitory conductance are much larger indeed. It is 3 ± 2 times the leakage conductance. Variation is added do to the change in the activity of stimulation neurons. Those conductances parameters set the neuron in the High Conductance State[18].

Hardware conductance

The actual achievable conductance depends on the chosen leakage conductance. A conductance three times larger than the leakage conductance con only be achieved if the latter is set to less than one third of the maximum possible conductance indeed. This is generally possible in 10^4 mode as the leakage conductance is limited to this margin in this case.

Exponential term

The parameter Δ_t is set to 2 mV in [22]. Accordingly, a value of the same size including a margin has been chosen during design. However, the real achievable factor depends on voltage scaling. The maximum adjustable value of Δ_t is 15 mV in the current HICANN. It might be enlarged in the next chip version.

Voltage level

The voltage level of AdEx in [22] lies between -75 mV which is the value of the inhibitory reversal potential and the 20 mV which is the threshold for spike detection. In our design, the spike detection threshold has to be set to lower voltages as the value output range of the exponential circuit is shrunk. Due to the sharp exponential rise, this is not a problem as it does not change the spike timing drastically. Furthermore, adaptation is hardly effected by the fast changing membrane voltage during a spike due to the long time constant τ_w. The highest voltage which needs to be implemented is the excitatory reversal potential which is set to 0 V in [22].

3.12.2 Parameter Translation

The parameters used in biology and the parameters available in the hardware implementation vary drastically. Here we give an overview of parameter translation needed during design. Detailed translation used in the current software framework will be discussed in [69].

Voltage scaling

To translate between both systems, at first sight the voltage levels have to be met. Here, the model covers a range of -75 mV while the available hardware range has an extend of more than1 V. The hardware ranges are set by the valid input range of the OTAs (up to 1.3 V, see 3.3.3) and a reasonable distance to the ground rail. The latter constraint is weak. At first, the biological operating range is multiplied by a scaling constant factor. E.g. ten. Next, the operating range needs to be shifted to fit into the hardware range. Depending on the needs of the model, different scaling factors can be used. Smaller scaling factors usually result in a better linearity of the OTAs while unwanted effects like noise and crosstalk are enhanced. In the following, we chose a factor of 10. The factor needs to be applied to all voltages respectively voltage differences including Δ_t, the adaptation voltage V_w and $V_b = b/a$

Time scaling

To achieve the acceleration of the model, basically all biological time constants have to be divided by the time scaling factor, which is 10^3, 10^4 or 10^5. We assume a time scaling factor of 10^4. Accordingly, the necessary hardware leakage time constant range is 500 ns and 5 µs. Including the membrane capacitor which is 2 pF in this case, we retain 400 nS to 4 µS. The same can be done with the parameter τ_a. This results in a hardware a of 400 nS up to to 4 µS.

Adaptation

The value of the conductance g_w can be obtained using the adaptation time constant τ_w. The scaled version of τ_w reaches from 10 µs to 600 µs. Subsequently, the adaptation capacitor C_w has to be taken into account. It is 2 pF. Consequently, the conductance range of g_w is 33 nS to 200 nS. The parameter b is closely linked to the model parameter a(see Equation 3.20). Time-scaling of the parameter b is done through a. Hence, the biological parameter a and b can be used for the calculation of the necessary circuit parameter $I_{\overline{\text{fire}}}$. The combination of the equations 3.20 and 3.21 results in:

$$I_{\overline{\text{fire}}} = \frac{b}{a}\frac{c_w}{t_{\text{fire}}} = 2\ \mu\text{A} \tag{3.41}$$

with the fire pulse length $t_{\text{fire}} = 20$ ns.

To allow for a scaled membrane voltage, w needs to be scaled, in addition, as dV/dt is directly scaled with the voltage scaling factor in Equation 3.1. This does not apply to the conductances, as here the voltage scaling is directly done through V and the reversal potentials. The scaling of w is done through a scaling of the V_w in the model implementation. As b directly impacts on it needs to be scaled in addition. This would result in a ten times larger current $I_{\overline{\text{fire}}}$ if ten is the voltage scaling factor. However, this translation was not included in the design phase, so the parameter range of b is shrunk in HICANN version 2. Nevertheless, this issue is solved for HICANN version 3.

The synaptic time constants can directly be multiplied with the time scaling factor.

3.12.3 Realization

Bias current and capacitor scaling

The parameter ranges given in the design constraints are huge indeed. One order of magnitude conductance scaling is possible using the range of the floating-gate cells which are sourcing the neuron's biases(See Chapter 9). This is only sufficient if the necessary parameter ranges are constrained for only one time scaling factor. To cover the timescales 10^3 and 10^5 a combination of bias current scaling and membrane capacitor scaling has been applied. When working in 10^5 mode, the membrane capacitor can be switched to 160 fF. This way, the OTA is able to achieve short time constants although the conductance is limited. In 10^4 and 10^3 mode, a large capacitor of 2 pF can be added to the C_{m}.

Switchable current mirrors

The translation between the floating-gate parameter ranges (100 nA to 2 µA) and the neuron parameters is done through current mirrors. Indeed, even the relative range of the floating gate parameters is to small, so a single mirror is not enough for parameters depending on the time scaling factor. However, the mirrors them-self can be scaled by adding or removing transistors in the branches of the current mirror by switches. Figure 3.34 shows the realization of the current mirror for g_{l} and a.

Current mirrors are no ideal devices of course. In addition to the finite output impedance, the real multiplicator of a current mirror is influenced by device miss-match. The importance of the latter is increased by the small currents needed for biasing. Some of the mirrors will have to be operated in the subthreshold region. The output impedance is no problem, as the mirrors are sourcing other mirrors in most cases. Thus, their output voltage is constant - no drastic current variations between different neurons will be caused.

To observe the miss-match sensitivity of the current mirror circuit, a Monte-Carlo simulation has to be done. The results for the current mirror in Figure 3.34 can be found in Figure 3.35. The variation is large indeed. On a wafer, there will be 200 000 neurons thus there will be neurons with a limited parameter range.

The biasing current mirror configurations can be found in Table 3.6.

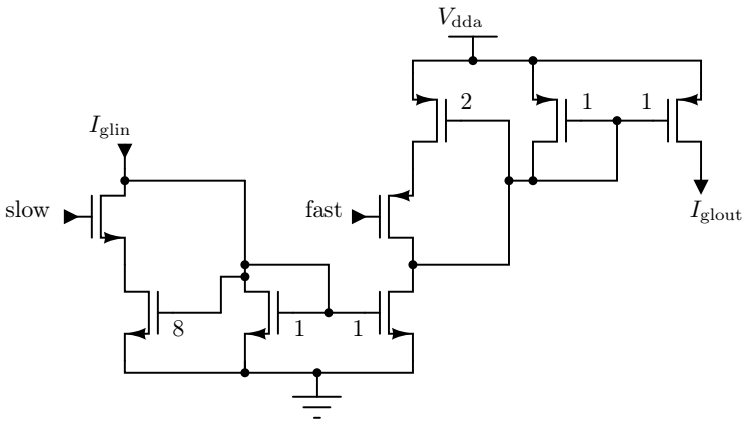


Figure 3.34: Current mirror used for switching the bias current for the parameters g_{l} and a. Switches set the translation factor.

Table 3.6: Model parameter bias scaling factors applied on on floating-gate currents between 100 nA and 2.5 µA

Parameter	default	fast	slow
I_{gl}	3:1	1:1	27:1
I_{gla}	3:1	1:1	27:1
I_w	32:1	8:1	640:1
I_{rexp}	3:1		
I_{conv}	1:1		
I_{int}	1:1		
I_{pl}	1:1		
I_{reset}	1:10		
$I_{\overline{\mathrm{fire}}}$	1:2 (1:10 in HICANN v3)		

3.12.4 Hardware Parameter Summary

Table 3.7 displays all parameters of the hardware neuron model with their functional meaning and ranges obtained from simulations. Therefore, the parameter dimensions correspond to the function and not to the corresponding bias. All ranges are given in 10^4 mode which is the default case in Table 3.6 and in real hardware dimensions. Accordingly, voltage and time scaling still need to be applied as described above. Hardware ranges are presented as the biological parameters might depend on two hardware parameters. The leakage membrane time constant depends on C_{m} and g_{l} for instance.

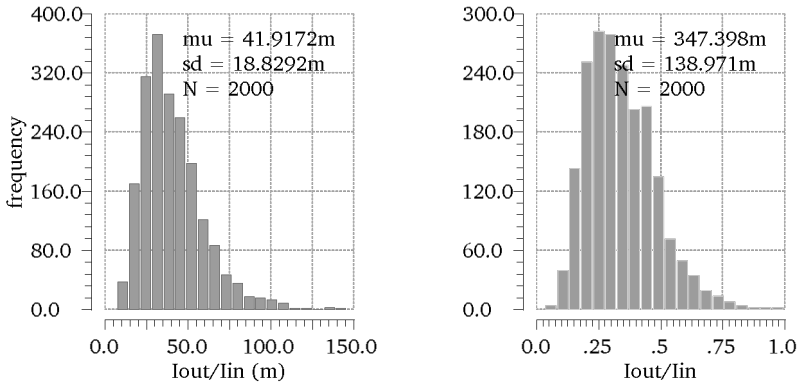
Functional ranges

The lower boundaries of current parameters are to be understood as margins of precise programmability. Indeed, smaller currents are possible as suggested by the current scaling factors. However if smaller values are used a larger variation can be expected between different experiments and in time. Nevertheless, some models need like "neural sampling" [73, 74] need to work in regions out of the specified regimes. In contrast to the variation for smaller currents, all currents can be set to zero precisely. Consequently complete circuit parts can be switched off.

Soft boundaries

Table 3.7: Neuron hardware parameter from HICANN v2. Ranges are given in the effected functional parameter. If not pointed out different, the given ranges are projections of the complete floating-gate parameter range from 100 nA to 2.5 µA or 0 to 1.8 V.

Param.	Function	Functional Range 10^4	Comment
I_{gl}	g_{l}	400 nS to 4 µS	
E_{l}	E_{l}	0 to 1.3 V	
I_{gla}	a	400 nS to 4 µS	
I_w	τ_w	33 nS to 200 nS	
I_{fire}	b/a	0 to 50 mV	voltage scaling applies
V_{exp}	V_{t}	0 to 1 V	larger Δ_{t} limits lower boundary
I_{bexp}	(V_{t},) Δ_{t}	2 to 15 mV	voltage scaling applies
E_{syn}	E_{syn}	0 to 1.3 V	
I_{conv}	maximum synaptic conductance	5 µS	
V_{syntc}	τ_{syn}	25 ns to 10 µs	technical range 1.25 to 1.45 V
V_{syn}	synaptic input integrator reference	fix at 1 V	technical
I_{int}	synaptic input integrator bias	fix at 2 µA	technical
I_{pl}	refractory period	50 ns to 500 ns	
V_{t}	Θ	0 to 1.3 V	
V_{reset}	reset voltage	0 to 1.3 V	global
I_{reset}	current used to pull down membrane at reset	1 to 25 µA	global
I_{bout}	bias for neuron output amplifier	2.5 µA	technical, global
I_{bexp}	bias for buffer of V_{exp}	2.5 µA	technical, global
C_{m}	C_{m}	2.16 pF or 160 fF	140 fF parasitics
C_w	Adaptation capacitor C_w	2 pF	



(a) Scaling 1/27 **(b)** Scaling 1/3

Figure 3.35: Typical Monte-Carlo DC simulation of parameter scaling current mirror for g_l and a (see Figure 3.34). The figure shows the output/input at 1 µA input current.

The minimum value of most biasing voltages is set to 0 V. This is a theoretical limit indeed, as hardly any of the presented circuits will properly at the ground rail. A distance of 200 mV to 300 mV is recommended.

Global parameters

Global parameters are shared between several neurons on one chip half. Two instances of V_{reset} connect to every second neuron for instance, while all neurons of one half share I_{bout}. See [61] for details. All other parameters correspond to individual floating-gate (See Chapter 9) values.

4 Point Neuron Experiments

This chapter describes neuron measurements done on HICANN v1 and v2. I start with an introduction to the used measurement setups and the used chips. Subsequently, different experiments are performed. Each experiment has its own methods and results subsection. Experiments start with the analysis of analog output capabilities and finish with a small network experiment using several neurons.

4.1 Methods

A zoo of different setups has been used for the experiments presented in this chapter as setups evolved in time. Generally, a division can be made between prototyping platforms for MPW chip evaluation and demonstration and the wafer-scale setup (See 2.2.).

4.1.1 Evaluation Setups

Setup evolution

The evaluation setup for the HICANN chip grew with with completion of the system. The idea is to have the complete chain of components tested in the evaluation setup. A schematic overview can be found in Figure 4.1.

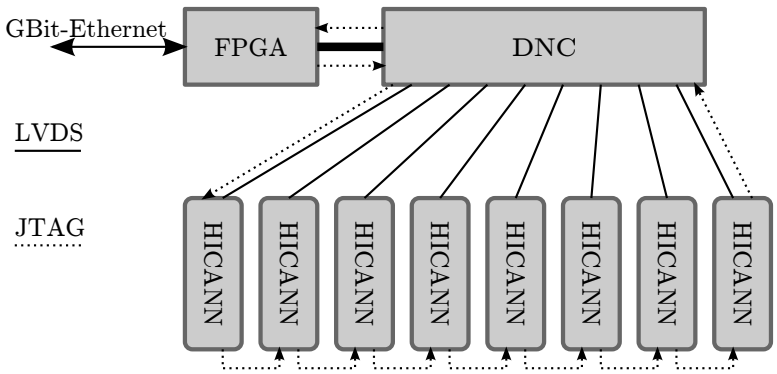


Figure 4.1: Overview of the chip chain implemented in the complete evaluation setups.

Interfaces

The final interface to the outside world is Giga-Bit-Ethernet. However, before this high speed connection was available, the system has been interfaced via the common boundary scan protocol JTAG[1]. The JTAG connection can be used to measure each element individually. Four dedicated printed circuit boards (PCB) are necessary to implement the setup while keeping independent testability for all components:

[1] Joint Test Action Group. The standard can be found in [80]

- The FPGA board is equivalent to the final FPGA board used in the wafer-scale system. Before its completion, a commercial FPGA board has been used as a place holder.
- The DNC module carrying the current version of the DNC.
- The System Emulator Board(SEB) supplies and interconnects the HICANN.
- 2 HICANNs are bonded on a small board called HICANN module.

Building up

First measurements started using only the SEB and a single HICANN module. The next step was to include the commercial FPGA board(This setup has been used in the measurements presented in 4.4 for instance.) Subsequently, the DNC has been added and finally(See Figure 4.3), the FPGA board has been replaced by the FPGA board of the wafer-scale system. A photograph of the complete setup can be found in Figure 4.2. I will concentrate

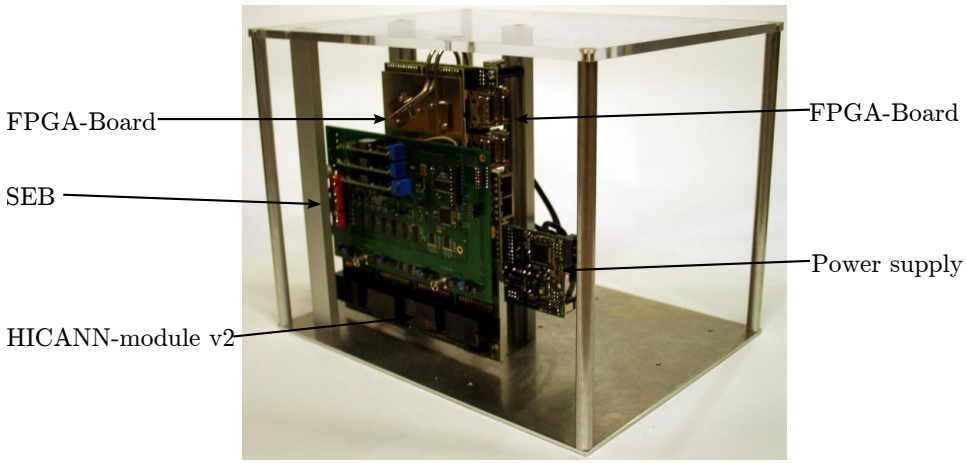


Figure 4.2: Photograph of the final evaluation setup

on the SEB and the HICANN modules now as those boards have been designed within this thesis and interfere the measurements most.

System Emulator Board

Cost efficient design

A photograph of the SEB can be found in Figure 4.3. The idea is to have all active components needed for testing of the HICANN chip located on the SEB to allow small and cheap HICANN modules. Three different types of the SEB have been designed v1, v2 and v2.1. v1 has been revised to v2 and v2.1 by Andreas Grübl. Here, I will focus on the description of v2 and close with the differences in v2.1.

Power supply

The SEB is designed to supply 8 HICANNs located on 4 four HICANN modules. DC-DC converters are used to generate the main power supplies as the worst-case current consumption of the digital and analog 1.8 V power of the HICANN is specified to 1 A each[61]. All special power supplies which might need calibration can be adjusted via the serial interface I2C[2]. This was especially important for the power supply of the floating gate array as earlier measurements on test ASICs showed sensitivity to the power up scheme[82].

[2] A serial communication protocol([81])

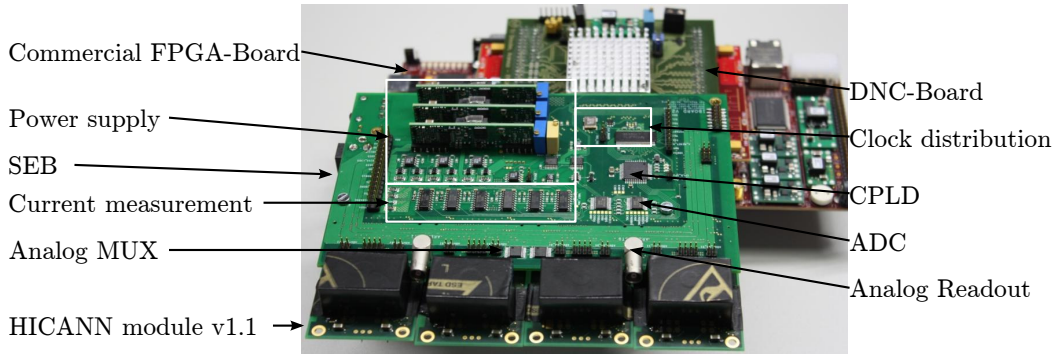


Figure 4.3: Photograph of the System Emulator Board v2 with commercial FPGA board and DNC Board

Periphery

The CPLD[3] is basically responsible for signal level transformation. HICANN and DNC v1 had different logic power levels. In addition it merges two control signals to a single signal as only few signal pins are available on HICANN due to wafer-scale integration. Furthermore, the CPLD allows to connect the control signals between DNC and the HICANNs on output pins for measurements or LEDs[4].

Current measurements

Shunt resistances in the power nets allow to measure current consumption via instrumentation amplifiers. However, this method has been hardly used due to inaccuracy. Current measurement of power supplies is usually done by external power supplies in this setup which is more reliable.

Analog output

The analog outputs of each HICANN can be switched to one of the two outputs of the SEB. Switching is done through an analog multiplexer which is accessed via I2C. In addition, an ADC enables direct measurement of DC-levels of the analog outputs. This feature has been used in the first experiment presented here and in [70].

Linear regulators for DC-DC converters

The DC-DC converters used for power supply have been necessary due to the high current constraints. However, when doing measurements with one or two HICANNs, linear regulators could have been used if the real power consumption of the HICANN is taken into account. Signal quality suffers the DC-DC converters. Consequently, a version v2.1 has been designed allowing to use only one DC-DC converter to regulate the power supply to a reasonable level to use linear regulators to generate better power supplies for the ASIC.

HICANN module

Two versions

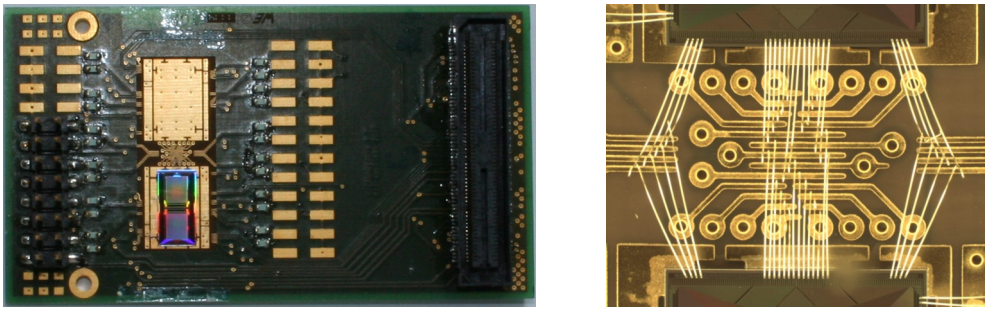
Two different versions HICANN modules have been produced. However due to bad production quality in the bonding area, the v1 had to be produced by two different companies. Here I will only describe version v2 which is close to identical to v1. It has been designed for HICANN v2. A photograph of the HICANN module can be found in Figure 4.4 a).

Besides the up to two HICANNs, the HICANN module holds blocking capacitances, termination resistors, a Zener diode and connectors to simplify measurements of L1 signals. The Zener diode protects the HICANNs against too large differences between the floating-gate powers which would destroy the chip. This inhibits problems which occurred in [82].

Two chip module

Two HICANNs can be mounted on the HICANN module. However, for first tests, usually only one chip is installed. The upper edges of the two HICANNs have been placed

[3] Complex Programmable Logic Device
[4] Light Emitting Diode



(a) Full view, single chip

(b) Detail of bonding area between two chips

Figure 4.4: HICANN module

as close together as possible, to allow direct wire bond between the too chips. Hence, it has been possible to verify inter-HICANN L1 connections before an uncut wafer existed. The placement constraint of the HICANNs results in the necessity of a complex bonding scheme(Figure 4.4 b)).

4.1.2 Wafer-Scale Setup

A photograph of the wafer-scale setup can be found in Figure 4.5. The FPGA-board are the same as in Figure 4.2. Only a single chip of the complete wafer is used here. Hence, there are

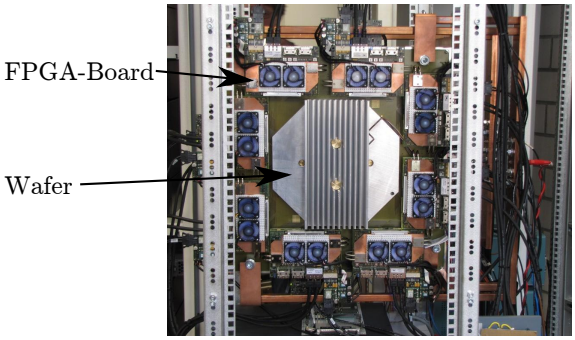


Figure 4.5: Detail of Wafer-Scale Setup. The wafer is located within the aluminium plate.

two main difference to the evaluation setup from a measurement point of view. The signal routing of analog signals differs and the power supply is stabilized better in virtue of much the blocking capacitors for each reticle.

The power supply is blocked with a much larger capacitance as each of the 48 reticles on a wafer. In addition the main supply is directly created by a laboratory power supply and not by DC-DC converters. The analog signals recorded from the wafer are less noisy than the signals recorded from the evaluation signals.

Better power supply

However, due to longer signal routing, they are sensitive to external coupling from switching signals of DC-DC converters of supply voltages on the BWS(See the small peaks in Figure 4.10

for instance). The crosstalk between two analog output lines on the wafer is more developed than on the evaluation setup. Hence only a single analog output should be used for precise analog measurements on a wafer HICANN. A reason for the larger cross talk could be the close routing of the analog signals on the post-processing layer on the reticle, or parallel routing in the system PCB. However, no parallel routing has been intended. Hence an observation of cross-talk of analog outputs from different reticles would be necessary for further characterization.

Crosstalk

4.1.3 ASICs

A list of the HICANNs which are used in the experiments presented this chapter can be found in Table 4.1. The given scaling factor has to be applied to the analog signals when 50 Ω DC-termination is used. Due to fixed-pattern noise of the series line impedance balancing resistor located in the chip and the series impedance of the output stage of the amplifier, the factor varies significantly from the factor 2 which would be created by a perfect voltage divider of two 50 Ω resistances. Only 6 chips are used in the experiments shown in this chapter. However,

Voltage divider at output

Name	Version	Wafer	Identifier	A0 scale	A1 scale
nips1	v1	MPW 4	1		
w0	v2	W 0	reticle 34 chip 0	2.23	2.32
iscas1	v2	MPW 14	5	2.17	2.26
iscas2	v2	MPW 15	1	2.18	2.15
d1	v2	MPW 15	13	2.14	2.17
d2	v2	MPW 14	1	2.15	2.20

Table 4.1: Individual HICANN chips used in the presented measurements. The measurement error of the scaling factors is below one per cent

the total number of bonded and tested HICANN v1 chips is 34. HICANN v1 had bonding problems due to too small spacing between bond pads and scribe line respectively some power nets on top metal. The bonding issue could be solved by manual bonding techniques. Hence 18 chips could be used finally. Among theses are 7 chip pairs on single HICANN modules.

Bad bonding yield for HICANN v1

28 HICANN v2 chips have been bonded. 26 are usable. 20 are bonded as pairs.

A complete wafer has been extensively digitally tested by Andreas Grübl using a wafer prober with a needle card. The resulting chip yield for digital tests was about 98 %[83]. Analog functionality has been measured for a single chip from the probed wafer. The number of chips on wafer w0 which have been tested for analog functionality is above 10[84].

4.2 Characterization of Output Capabilities

For correct interpretation of measurements, it is necessary to characterize the analog output devices. The output amplifiers[76] of the chip and especially the individual output buffers of each single neuron as those will generate an individual offset for all measurements. For these characterizations, two individual experiments are performed in this section.

4.2.1 Methods

These measurements have been performed on the SEB v2.1 with HICANN v2 d1. For better measurement quality, a SEB with linear regulators for digital and analog power supply has

been used. The ADC of the SEB is used for DC measurements. Each measurement is a tuple of the average of 5 individual ADC measurements and the standard deviation.

Offset Comparison of Analog Output

Exhausting hidden features

This measurement uses the feature that every neuron output line can be connected to each of the two output amplifiers. If the same output line is connected to both amplifiers, their outputs are connected and the offset can be measured. Nevertheless, for sweeping the input of the amplifiers, the membrane voltage has to be swept which is not intended by design. However, setting a neuron to a high leakage conductance g_l will fix the membrane voltage at the leakage potential E_l if all other conductances are off. Consequently, a sweep of E_l can be used to sweep the membrane voltage.

In this experiment, both output amplifiers are connected to the membrane of neuron 0, which is the left most neuron of the upper neuron row. The neuron's g_l is programmed to the maximum possible value. For this purpose, the g_l current mirror is set to a ratio of 1:1 and the bias is set to the maximum floating gate value.

Neuron Output Buffer Offset

Inter-connecting membranes

The same membrane voltage sweep is performed for a group of 64 neurons now with a maximum DAC value of 600. The DAC value limit has been chosen to remain in the operation regime of the OTAs. All transmission gates between the neurons are connected, so the neurons are expected to work as a single neuron with equal membrane potential. For each floating-gate DAC value, the membrane voltage is measured through the output amplifier of each of the 64 individual neurons.

4.2.2 Results

Output amplifier

The results of the offset difference measurements can be found in Figure 4.6. The voltage range measured lies between 111 ± 1 mV and 1202 ± 2 mV. In addition, an extrapolation of the linear range of the output voltage curve to DAC value 0 would result in roughly 50 mV.

Lower boundary

This value correspond to the chip internal ground level. The higher lower boundary in the curve can be explained by the used n-mos input stage in the leakage OTA partially. Indeed, obtaining the minimum DC-output voltage of the OTA via typical Monte-Carlo simulation with 100 samples results in 27 ± 6 mV. An explanation for the missing 35 mV could be explained by a higher ground level at the OTA.

Higher boundary

The higher boundary of the measured voltage should be around 1.3 V. Here, the explanation is a voltage drop on the power supply suggested by extrapolation of the linear part of the curve to higher values(interpretation of earlier measurements shown in [70] suggest a drop to 1.7 V) . In addition, the next measurements results in a negative offset of 40 mV for the used neuron output amplifier. With the assumption of an additional negative offset at the neurons leakage OTA, the boundary difference can be explained.

Output offset

The offset between the two output amplifiers is 4.5 ± 0.8 mV when averaging over the whole input range. Indeed, this offset is small. Nevertheless, it has to be taken into account that the design goal of the floating-gate precision [61] is within this range for instance. In addition, this is only a single sample. However, a typical DC Monte-Carlo simulation with 1000 samples reaches an output offset sigma of 3.7 mV which adds up to 5.2 mV using error propagation. Thus the measurement lies within the margins.

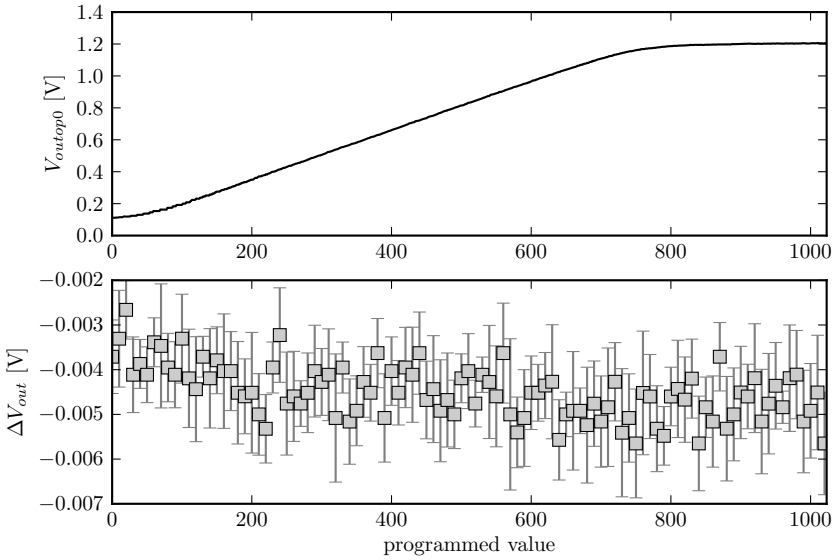


Figure 4.6: Analog Output 0 with swept membrane voltage(top) of neuron 0 using the leakage term (top) and offset between Output 0 and 1. The upper limit is given by the limit of the leakage OTA

Neuron output buffer offset

Larger measurement error

In comparison to the last measurement, the single measurements here suffer a higher measurement variation, which is voltage dependent. The maximum variation is 3 mV for DAC-values between 100 and 400. As all neurons are interconnected and each used leakage OTA has a different input offset some current flux is expected.

Figure 4.7 shows the measured output voltage V_{out} of all included neurons for two different values. The standard deviation of V_{out} is around 15 mV. Differences of 50 mV between neighbouring neurons have been observed.

Compensation

The input offset of the output buffer is supposed to be input voltage independent over a wide voltage range. Accordingly, a single measurement of the offset can be used to counterbalance the offset in other measurements. These measurements have been called aout[5] fingerprint of a chip as they are characteristic for each individual chip. The fingerprint has been obtained for a DAC value of 500 averaging over 10 different measurements at this value. The corrected values from Figure 4.7 are the measurements counterbalanced by the fingerprint. At DAC value 500 the mean is hardly dependent on the neuron number which corresponds to the location on the chip. However for higher or lower values, a dependency can be observed. The fingerprint corrected voltage for DAC value 650 is 15 mV higher at neuron 63 than at neuron 0. This dependency is getting worse if higher DAC values are observed, reaching a limit at the operating limit of the OTA. For smaller DAC value than 500, the slope according to the neuron number is counter wise. At DAC value 1023, the difference between neuron 0 and 63 is 40 mV. A voltage drop of 40 mV could be caused by a current of around 640 nA from neuron 63 to neuron 0. Indeed, a current of this size is not unrealistic

Voltage dependency

[5] analog output

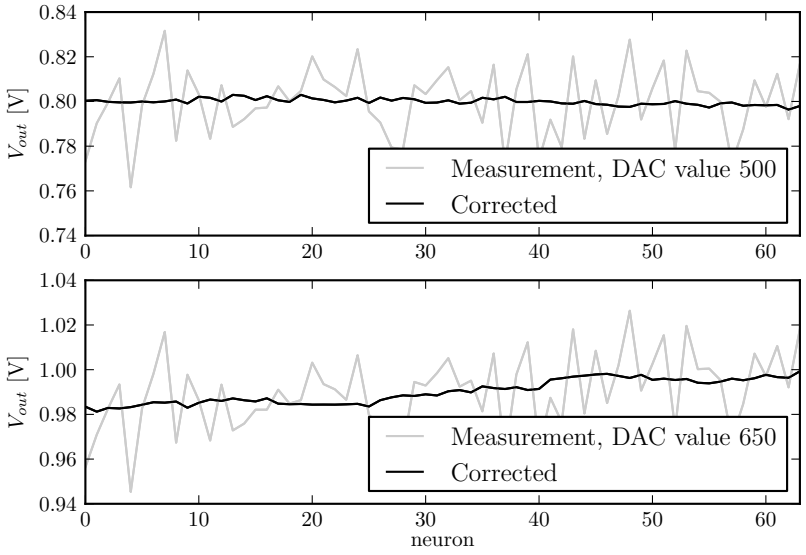


Figure 4.7: Red out membrane voltage in each of 64 interconnected neurons for two different programmed values and corrected by the aout fingerprint of the chip. Single measurements are interconnected for readability reasons only. An optimal correction would result in a straight line parallel to the x axis.

as the leakage OTA is not calibrated at all. Furthermore, chip production techniques can create location dependent gradients of transistor parameters. Another explanation for the slope could be chip internal voltage drop on the analog power supply.

Reduction matches results from Monte-Carlo simulation

The standard deviation according to the mean of the aout values of all neurons in dependency to the programmed membrane voltage can be found in Figure 4.8. All measurements have been balanced with the finger print for DAC value 500. The σ of the simulation trace has been obtained by a DC Monte-Carlo simulation. In contrast to the measurement, it does not show any voltage dependency in distance to the power rail. Indeed, this is no surprise, as the measured local gradients can not be included in this simulation. However, the simulated value roughly matches the difference between the sigma of the measured and the corrected values. Consequently, the offset compensation technique using the fingerprint is capable of drastically reducing offsets created by statistical process parameter variations.

4.2.3 Conclusion

Neuron output amplifiers must be included in calibration

Measurements and simulations have shown that the offset variation of the chip output amplifier is small. Nevertheless, as it matches the desired precision of the floating gates. Accordingly, precise measurements or calibrations should use either only one operational amplifier, or add a measured offset compensation.

The offset variation of neuron output amplifier is drastic indeed. However, the aout fingerprint can be used as offset compensation. It can be obtained with a single floating gate programming run and read out as DC-voltage by an ADC. Thus, its measurement is fast - in the order of magnitude of seconds.

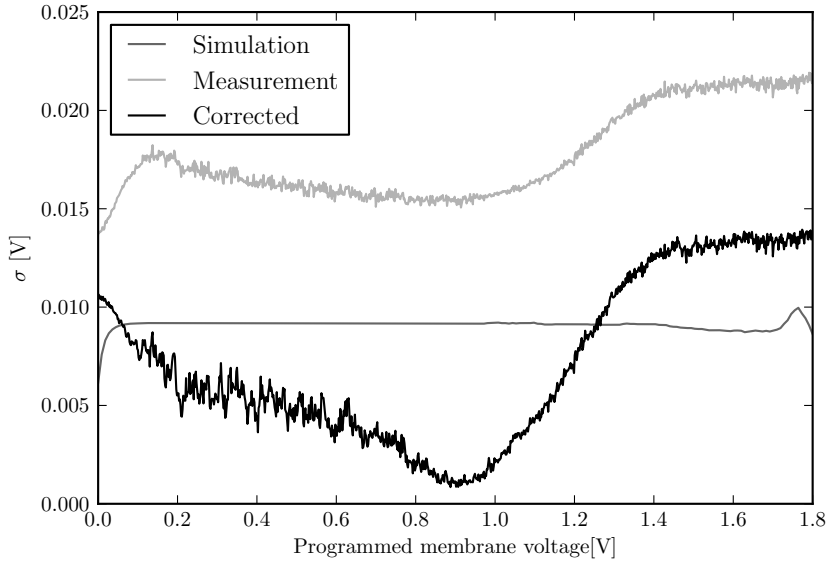


Figure 4.8: Standard deviation of the mean membrane voltage, read out in each of 64 interconnected neurons in comparison to results from Monte-Carlo simulation. The standard deviations from measurements rise above 1.1 V due to gradienst.

Reset voltage used as reference

Currently, offset compensation is done by referencing to the neuron reset voltage in calibration[85]. Indeed, the offset of the neuron output amplifier can be compensated this way. The offset of the reset amplifiers is already compensated as the amplifiers output voltage is used as a reference during programming the floating gates. Differences between the reset voltages can only be caused by the two different used DAC.

4.3 Reference Emulation

In [22], the publication introducing the AdEx, Brette and Gerstner show the models capability of reproducing the behaviour of a more complex Hodgkin-Huxley style model[19, 86]. In an exemplary simulation, the neuron is stimulated with two different current pulses. The first one is not high enough to generate an action potential. Nevertheless, it produced subthreshold membrane behavior. Subsequently, the second pulse excites the neuron an it starts spiking. The spiking frequency is adapted. The membrane trace and adaptation variable of a similar simulation using the parameters from [22] and the Nest simulator can be found in Figure 3.1

Benchmark during the design

During the design of the circuit implementation my first goal was to qualitatively reproduce this pattern. Due to the importance of this experiment, it has been repeated in simulation here with the final circuit of HICANN v2. Furthermore, the parameters of the simulation have been roughly mapped to real chip parameters to measure a real membrane voltage of this experiment.

Qualitative measurements

Measurements shown in this section are qualitative and use only a single neuron on a single HICANN (w0) on the wafer scale system. However these patterns have been used as a benchmark in chip's initial operation tho show general neuron function. Consequently, similar measurements have been performed on nearly any analog tested HICANN chip.

4.3.1 Methods

Simulation

Parameterization

The neuron has been simulated in a typical transient circuit simulation. As membrane capacitor, 2.16 pF have been used. Parameter have been mapped using hand calculations and the conductance characteristic of the OTA. A voltage scaling factor of 5 and a time scaling factor of 10^4 have been applied. However, as described in 3.12.2, the dimension of b/a is to small in the circuit. Consequently, the maximum possible b/a has been chosen and the missing factor two has been counter balanced by doubling a.

The height of the current pulses in [22] are 0.5 nA and 0.8 nA. This results in 0.5 respectively 0.8 times 384 nA.

Measurement

Correspondence to simulation

The measurements have been performed on HICANN w0 on the wafer scale setup. Parameters from the simulation have been transformed to DAC values assuming a maximum voltage of 1.8 V and a maximum current of 2.5 µA. No calibrations have been included. Accordingly, miss-match is not counterbalanced at all and large parameter variations can be expected. To find a neuron reproducing the behavior of Figure 3.1 best, several different neurons on the chip have been observed.

Two separate measurements

To generate the step current stimulus, the chip internal programmable current sources(3.11.2) are used in continuous mode. The loop has a default length of 32 µs [6]. Consequently, the experiment has been divided in two parts - one with the membrane below the firing threshold and one with a firing neuron. The height of the current pulse is adapted during the experiment to find the height for sub and above threshold behavior.

4.3.2 Results

Simulation

Similar results for NEST and hardware simulation

The membrane voltage trace, the V_w can be found in Figure 4.9. V_w in the emulation is equivalent to w in the model. Although parameter mapping has only been done using hand calculations, the similarity of the results of the Nest simulation of the AdEx model in Figure 3.1 is obvious. However, the number of produced spikes is different.

Scaling factor

The scaled final time of 100 µs instead of 1 s in the model simulation expresses the time-scaling factor of 10^4. Voltage scaling can be observed by comparing the height of the membrane voltage during the smaller current pulse. The voltage of the model rises about 15 mV, while the emulation reaches 50 mV. Consequently, a smaller voltage scaling factor could be assumed.

Smaller relative b

However, a direct explanation could be the enlarged value of a for counterbalancing the smaller b/a. The difference in b/a can be seen directly by observing the size of the b enlargement of w in Figure 3.1 in respect to the maximum height of w during the small current pulse. There is roughly a factor of two. In contrast, the amount V_w is enlarged at each single action potential is comparable to the maximum height of V_w during the small current pulse.

Measurement

Neuron picking

The resulting measurements are presented in Figure 4.10. The chosen sample neuron is neuron number 18. About 10 neurons have been observed - the number 18 is a result of

[6] There are parameters to generate longer loops but those parameters are usually fixed. See Chapter 9

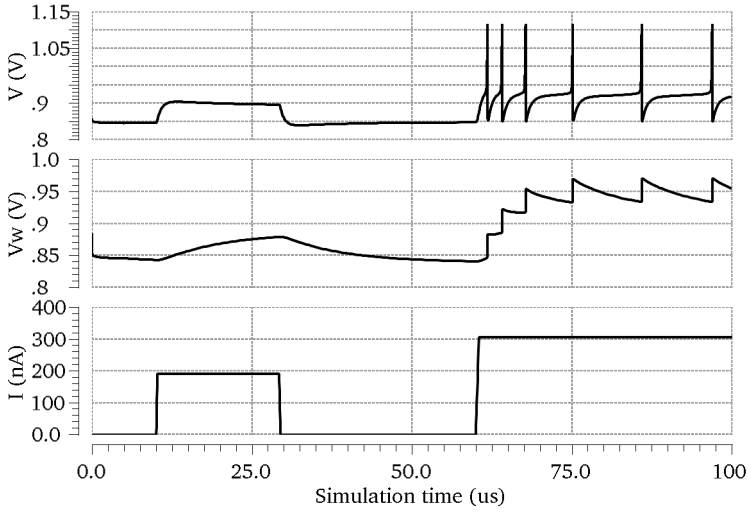


Figure 4.9: Typical transient circuit simulation of a single neuron of HICANN 2. The neuron is stimulated by two current pulses of different height. The second pulse excites the neuron, generating action potential. This figure is a circuit reproduction of Figure 3.1

random number picking. Given errors in this section are estimated from reading from the plots in Figure 4.10

Comparison to simulation

Indeed, the difference between the measurement and the simulation is large. However, after multiplying by the factor of 2.3 for the used analog output to compensate the division by the 50 Ω series termination, we retain a leakage potential of 820 ± 12 mV which is about 20 mV lower than the programmed potential. The value of Θ is 1.1 V in the circuit simulation. In contrast, the maximum height of an action potential is 1.143 ± 5 mV in the measurement. Variations in this order of magnitude can easily be explained by miss-match of transistors.

In simulation, the reset potential is equivalent to the leakage potential. With a value of 867 ± 5 it is 47 ± 11 larger than the measured leakage potential. This discrepancy is cause by the input voltage offset variation of the neuron's leakage OTA. The operational amplifier generating the reset potential is basically compensated during programming of the floating gates.

Spike density

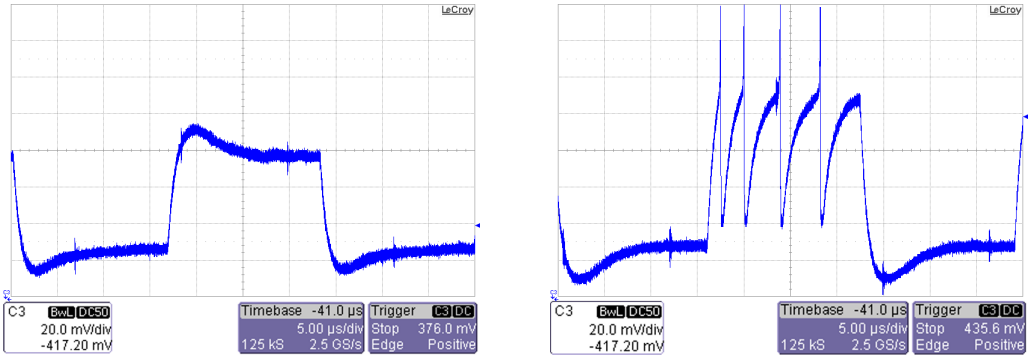
The length of the current pulse has to be much shorter in the experiment. Hence, the spike density is much higher in the measurement - 4 in 16 µs against 6 in 40 µs in the simulation. There are two reasons: Firstly, the stimulation current has been set to a higher value to see more spikes. Secondly, the effect of adaptation is much smaller in the measurement. Indeed, both reasons are correlated.

Voltage scaling

The exponential rise in the simulation starts at approximately 925 mV or 75 mV above the leakage potential. In contrast, the rise starts at 894 ± 12 mV or 208 ± 17 mV above the measures leakage potential. Accordingly, the effective voltage scaling factor of the measurement is much larger than the factor of the simulation. Therefore, much larger currents are necessary to achieve spiking and the impact of b is much smaller as it would have to be scaled by the effective voltage scaling which is not possible.

Point of exponential rise

The large deviation of the point of exponential rise can be explained by the two involved parameters V_{exp} and I_{rexp} which both are exposed to miss-match. Occasionally, other neurons even showed an exponential rise close to the leakage potential generating spikes with little or no stimulus. Other neurons' exponential threshold was above Θ so no rise could be observed



(a) Stimulus pulse height 1.03 µA: subthreshold

(b) Stimulus pulse height 1.20 µA: firing

Figure 4.10: Oscilloscope screen shots(setup parameters enlarged and recolored for readability) from measurement on wafer-scale setup: Membrane of an AdEx neuron stimulated by a periodic current pulse of different height. The voltage level is scaled down by a factor of 2.3 by 50 Ω termination. The bandwidth of the oscilloscope has been reduced to 200 MHz for noise reduction. The small regular peaks are external cross-talk and do not effect the membrane.

at all and the circuit behaved like and adaptive integrate-and-fire neuron.

4.3.3 Conclusion

Calibration needed

Straight forward parameter translations can lead to good results in circuit simulations. However, drastic variations of circuit parameters in the real silicon circuit inhibit direct translation in practise. Therefore, a calibration of neuron parameters is inevitable. However, it might be sufficient to calibrate the voltage levels of the neurons to achieve a reasonable behaving circuits.

Calibration is done by Marc-Olivier Schwartz in [69].

4.4 Characteristic Patterns

Reproduction of typical patterns

In [11] neuron models are characterized according to their capability of reproducing certain characteristic patterns occurring in biological neurons. The AdEx is not listed in this paper as its first publication is newer, but as most possible patterns are based on dynamical features like a second variable for adaptation and a non linearity causing positive feedback for spike generation, the AdEx should be capable of reproducing all patterns the Izhikevich model can. Firing patterns of the AdEx have been characterized in [12] However, inspiration for the experiments in these section was [11]. Figure 4.11 shows the different spiking patterns introduced by Izhikevich. For an explanation of the biological meaning of all patterns see [11]. In this part, I present and discuss simulations and measurements, I published in [44] at the NIPS[7] conference 2010. The presented patterns are tonic spiking, tonic bursting, phasic bursting and spike-frequency adaptation. However, simulations and measurements for phasic spiking and mixed/mode (called initial burst in [12]) have also been performed during the measurements for the paper. The pattern rebound-spike has been reproduce in simulation,

[7] Neural Information Processing Systems

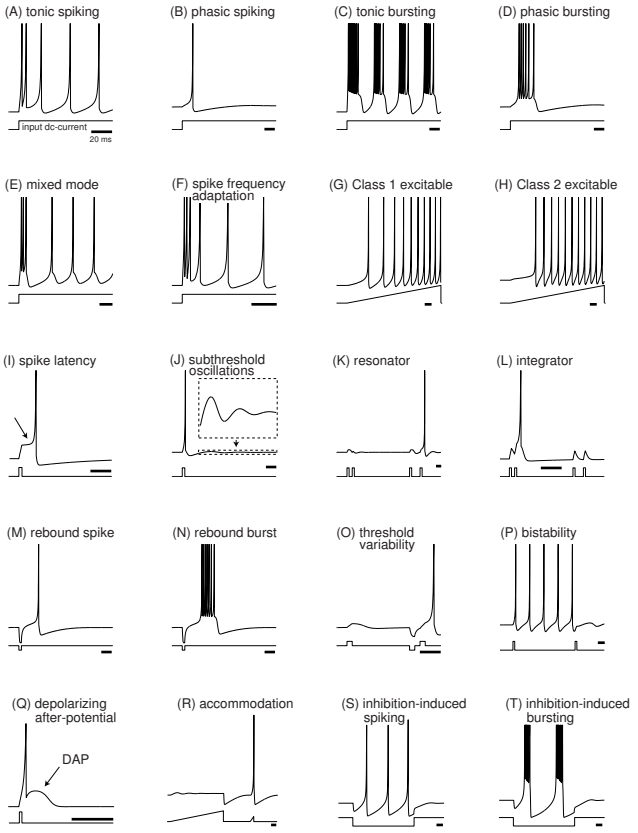


Figure 4.11: Neuron firing modes figure from[11](Electronic version of the figure and reproduction permissions are freely available at www.izhikevich.com)

but there was no intention to reproduce it in experiment due to the lack of a negative current source[8]. In addition Class 1 excitable, Class 2 excitable, resonator, integrator, subthreshold-oscillation, rebound burst and threshold variability have been reproduced in simulation by Marc-Olivier Schwartz during a supervised internship[85].

In contrast to the model, the parameter a is fixed positive in the circuit implementation. This inhibits patterns relying on a negative a like spike latency in [11] or delayed bursting in [12](equivalent pattern with a burst instead of a single spike). Furthermore, a depolarizing after-potential is not possible.

4.4.1 Methods

All experiments have been carried out on HICANN v1 chip nips1. A neuron pair (neuron 0 and neuron 1) has been interconnected, therefore, all conductances and the membrane capacitor are doubled. The resulting membrane capacitance on HICANN v1 is 4 pF. The used setup is a single SEB with the commercial FPGA board. No explicit parameter mapping between biological parameters and hardware parameters has been used for parameter definition. In contrast, the parameters have been chosen to reproduce the desired patterns best while

No explicit parameter translation of automatic calibration

[8]However, a solution can be to virtually shift the zero level of current stimulus.

keeping as many parameters as possible between different patterns. In all simulations except for tonic spiking, the subthreshold-adaptation parameter a and the leakage conductance g_l have been set equal to facilitate nullclines. However, as the chip has not been calibrated. Consequently, effects by miss-match had to be counterbalance by hand. Thus the mapping between real hardware parameters and circuit simulation is only rough. The chosen current stimulus is a current pulse of 600 nA with a length of 16 µs. This pulse is repeated every 33 µs.

Bursting

To account for the findings about bursting dynamics, new simulations have been performed in addition. These simulations use parameters chosen to achieve a more stable regime for bursting. The HICANN v2 neuron is used for the simulations.

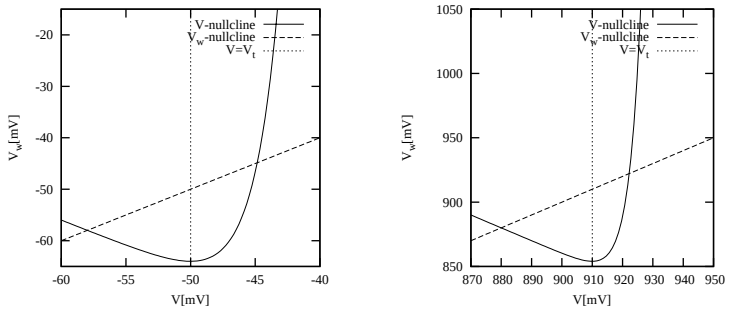
4.4.2 Results

The dynamics of a two dimensional model can be explained best, looking at a phase plane graph of the two variables their its nullclines(See 3.1.1). However, the in the circuit the w of the model equations has been transformed to V_w via $w = a(V_w - E_\mathrm{l})$. As new nullclines, we retain:

$$V-\text{nullcline}: \ \mathrm{V_w} = -\frac{g_\mathrm{l}}{a}(V - E_\mathrm{l}) + \frac{g_\mathrm{l}}{a}\Delta_T e^{\left(\frac{V-V_\mathrm{T}}{\Delta_\mathrm{T}}\right)} + E_\mathrm{l} + \frac{I_\mathrm{stim}}{a} \tag{4.1}$$

$$V-\text{nullcline}: \ \mathrm{V_w} = V; \tag{4.2}$$

Here I_stim is the stimulation current.



(a) Parameters according to figure 4 d) from [12] with $a = g_\mathrm{l}$

(b) Parameters used in hardware simulation for spike-frequency adaptation.

Figure 4.12: Phase planes of the AdEx model with w transformed to V_w, stimulus excluded.

Nullclines for spike-frequency adaptation

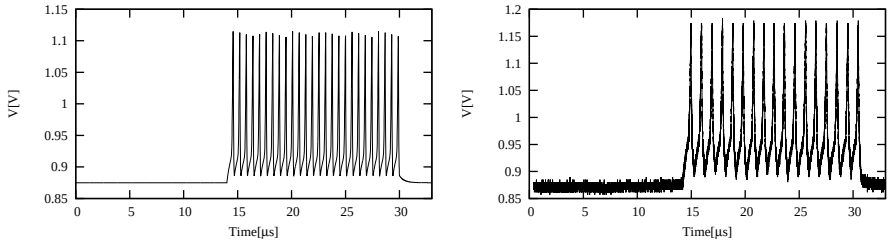
Figure 4.12 shows the phase plane corresponding to Equations 4.1 and 4.2 with $a = g_\mathrm{l}$ with biological model parameters and the parameters used for the hardware emulation of spike-frequency-adaptation in [44]. V_w and V are decreasing above their nullclines and increasing below. Still, the hardware nullclines are obtained from the model equations and not directly from the circuit. However, hardware parameters have been set into the equations.

Smaller Δ_t in circuit simulation

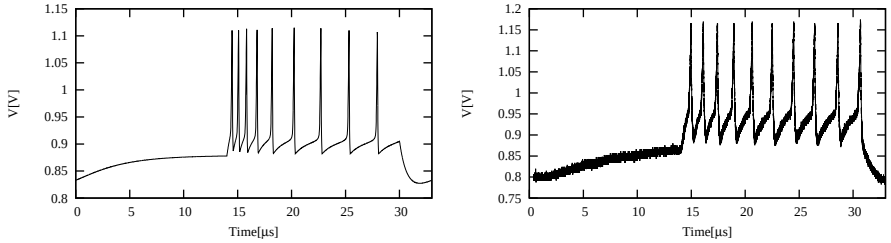
As conductances cancel out, only voltage-scaling needs to be applied to compare between the two phase plane plots. The distance between V_t and E_l is 8 mV in the model and 30 mV in the circuit, which lead to a voltage scaling factor of 3.75. However, the exponential slope is much larger in the simulation in Figure 4.12 b) is much steeper than the simulation in a). In fact, Δ_t in the simulation is about 4 mV in comparison to 2 mV in the model. Consequently,

a factor two is missing in the simulated Δ_t if the voltage scaling factor is included. Indeed, this miss-match is no surprise as no explicit parameter translation has been used for the simulation. Occasionally, it will have a strong effect of the different reproduced patters, especially on bursting. The margin between V_t and the V-nullcline is only 10 mV to 15 mV in the critical area.

Tonic Spiking and Spike-Frequency Adaptation



(a) Tonic spiking



(b) Spike-frequency adaptation

Figure 4.13: Membrane voltage trace from simulations and measurement. The neuron has been stimulated by a step current of 600 nA[44].

The results for the simulations and measurement on tonic spiking ans spike-frequency adaptation can be found in Figure 4.13. For tonic-spiking, a has been set to 0 to disable adaptation, while g_l has been doubled to counter balance the smaller conductance. Consequently, V_w has no effect for tonic spiking.

Spike after potentials

The AdEx model can produce two different types of *spike after potentials* (SAP) which are described in [12]. Firstly, sharp SAPs are reached if the trajectory is reset to point is below the V-nullcline in phase plane after an action potential. Secondly, a broad SAP will be created if the reset ends above the V-nullcline. The membrane voltage is pulled below the reset potential then. The patterns from Figure 4.13 all create sharp SAPs.

Accommodation index

On first sight, the measurements from Figure 4.13 look quite similar if the spike frequency is excluded. Due to a larger exponential threshold V_t in the measurement, the effect of adaptation is damped. However, to distinguish between , [12] and [87] use a metric they call accommodation index(A). It is a metric for the change of the length of the inter-spike interval (ISI):

$$A = \frac{1}{N-k-1} \sum_{i=k}^{N} \frac{\mathrm{ISI_i} - \mathrm{ISI_{i-1}}}{\mathrm{ISI_i} + \mathrm{ISI_{i-1}}} \tag{4.3}$$

To exclude transient behavior[87, 88] at the beginning of a spike train evaluation starts with the k^{th} spike of a pattern. According to [12], the k can be chosen as one fifth for small

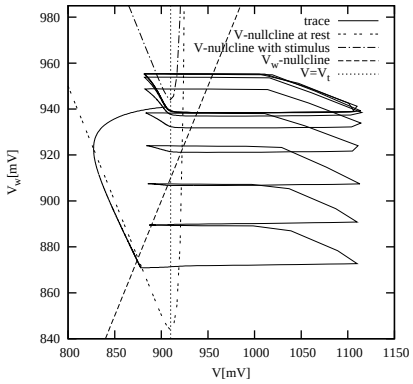


Figure 4.14: Phase plane plot for Figure 4.13 b). The plotted nullclines are according to the model nullcline Equations 4.1 and 4.2. Nevertheless, the correspondence between the model dynamics and the nullclines is apparent.(Enhanced version of figure published in [44])

number of ISIs, which is the case in my presented results. The index should be 0 for ideal tonic spiking behavior.

Comparison to literature

Applying the index to the tonic spiking results I retain 0 ± 0.0003 for the simulation and -0.0004 ± 0.001 in measurement. The Values for adaptation are 0.1256 ± 0.0002 for simulation and 0.039 ± 0.001 for measurement. Compared to the values calculated in [87], the resulting values are extreme as model parameters have been chosen to create the patterns evidently. The values from [87] are 0.0045 ± 0.0023 for fast spiking interneurons and 0.017 ± 0.004 for adapting neurons. Consequently, adaptation in the presented results is still strong.

Phase plane

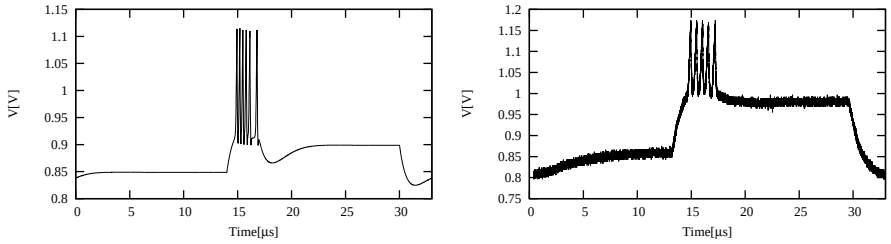
The phase plane of the spike-frequency adaptation simulation can be found in Figure 4.14. In comparison to Figure 4.12 b), the leakage potential has been shifted by -5 mV to account for the effective leakage potential obtained form the tonic spiking simulation. During stimulation by the current pulse, the V-nullcline is shifted $I_{\text{stim}}/a = 100$ mV with $a = 6$ µS and $I_{\text{stim}} = 600$ nA.

Trajectory discussion

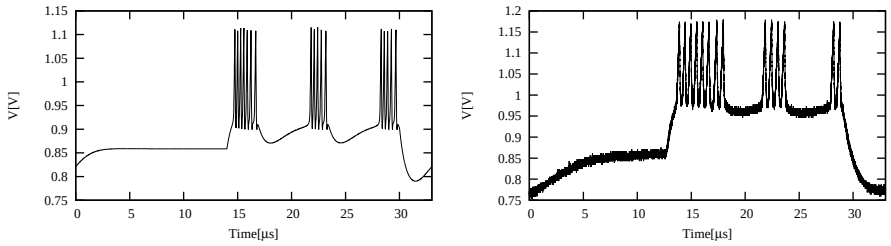
In the ideal case, the trajectory would settle at the crossing or the V_w-nullcline and the V-nullcline at rest before stimulus starts. The small offset can be explained by parameter translation and deviations to the ideal model used for the nullclines. At the onset of stimulus, the trajectory moves to larger values of V. After crossing, V_t, the positive feedback of the exponential term is taking over and the membrane is pulled to Θ. When Θ is reached, V is reset and V_w is enlarged by V_b. In contrast to similar phase plane plots[12] done with the simulated model, the trajectory of the circuit simulation is continuous, indeed. Above the V_w-nullcline, V_w decreases. The time constant of V_w is large in comparison to the membrane time constant. Hence, the change of V_w aside spiking is small, but rising with distance the V_w nullcline. When the trajectory approaches the V-nullcline it traverses tangential to it. A slightly large value of V_{reset} would have cause a broad SAP. The trajectory circles at a nearly constant spiking frequency until the stimulus is removed. Now the trajectory falls down onto the lowered V-nullcline and moves to the stable fix point at the crossing of the nullclines. Indeed, the matching between the simulated circuit dynamics and the nullclines obtained from the model equations with the circuit parameters is impressive.

Phase plane analysis of tonic spiking is not presented due to the unused adaptation variable V_w.

Tonic and Phasic Bursting



(a) Phasic burst



(b) Tonic burst

Figure 4.15: Membrane voltage traces from simulations and measurement of the neuron stimulated by a step current of 600 nA(Enhance version of figure from [44]).

Voltage trace and SAP

The membrane voltage traces for tonic and phasic bursting can be found in Figure 4.15. A phasic burst is a single burst, while tonic bursting is a regularly repeated burst. A burst can be defined by a group of spikes with sharp SAP followed by a broad SAP[12]. A definition only looking at the spike frequency is not sufficient, as there can be mix-ups with fast tonic spiking neurons. However, a broad SAP is achieved by crossing the V-nullcline.

Stable fix point at phasic bursting

The corresponding phase plane plots can be found in Figure 4.16. Comparing the plots, a smaller leakage potential E_l can be found for phasic bursting. The leakage potential is the crossing of the nullclines at rest. The smaller leakage potential shifts the V-nullcline to lower voltages creating a crossings between the nullclines during stimulation. Hence a stable fix point is generated below the spiking threshold. No further spikes are created after the first burst in in the phasic spiking pattern.

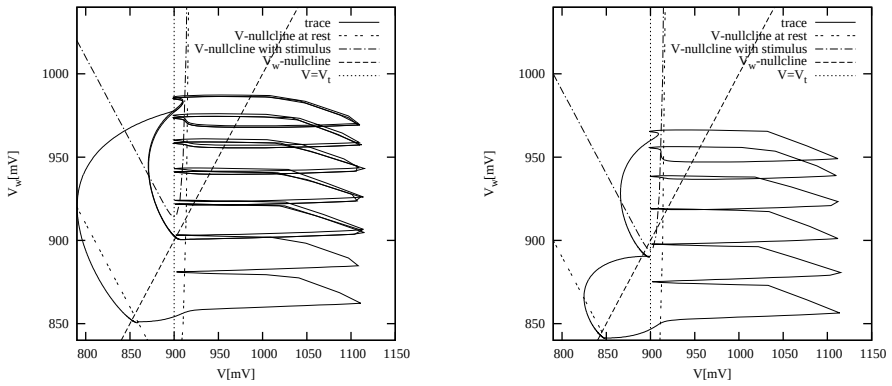
Small margin for bursting

As described above, the chosen value of Δ_t is small. In addition the voltage scaling of the burst simulations is larger due to a smaller E_l. Hence, the margin between V_t and the V-nullcline is small in the phase plane plots. Furthermore, even above the V-nullcline, the membrane voltage rises shortly due to different reaction times of the adaptation therm and adaptation and a weak reset. However, the short rise could be balanced by a longer refractory period. A larger, more realistic value of Δ_t could change the issue drastically.

The small margin of the reset was a challenge in simulation and especially in measurements. In addition, it makes the circuit noise sensitive, as small changes in the membrane voltage could result in one spike more, or less. Therefore, the exact number of spikes in a burst differs[44]. As more spikes correspond to larger values of V_w this results in different intervals between single bursts.

Voltage scaling too large

Another effect can be observed in the measurement. The value of V_t is obviously much higher in the measurement than in simulations. Indeed, the difference between the leakage potential and V_t is roughly 150 mV creating a voltage scaling factor of about 10. In fact, this



(a) Tonic burst **(b)** Phasic burst

Figure 4.16: Phase plane plots for Figure 4.15 b). The plotted nullclines are according to the model nullcline Equations 4.1 and 4.2. V_t and E_l for the nullclines have been shifted according to the effective values.

enlarges the Δ_t problematic. However, what adds is the saturation of the a OTA. This can be interpreted by the flat SAP in comparison to the simulation in tonic bursting and the not occurring overshoot in the phasic bursting pattern. As the OTA is already close to saturation, larger values of V_w have only small impact. Indeed, this issue could be removed completely by a proper calibration of V_t which has not been done in [44]. Furthermore, starting with a smaller operating range and using smaller currents as stimulus would simplify keeping the operating range of the OTA. However, this is a trade-off indeed, as a smaller operating range created a larger sensitivity to programming accuracy of the parameters.

Bursting Regime

New simulations with HICANN v2

As result from the outcomes of the simulations and measurements above, new bursting simulations using the neuron model of HICANN v2[9] have been performed to achieve a more stable regime. Therefore, a larger value of $\Delta_\mathrm{t} = 10$ mV and the operating range has been shrunk to keep the OTA a in saturation. Simulations have been done for different reset voltages. As an example, the membrane voltage and the phase plane for $V_\mathrm{reset} = 906$ mV is shown in Figure 4.16

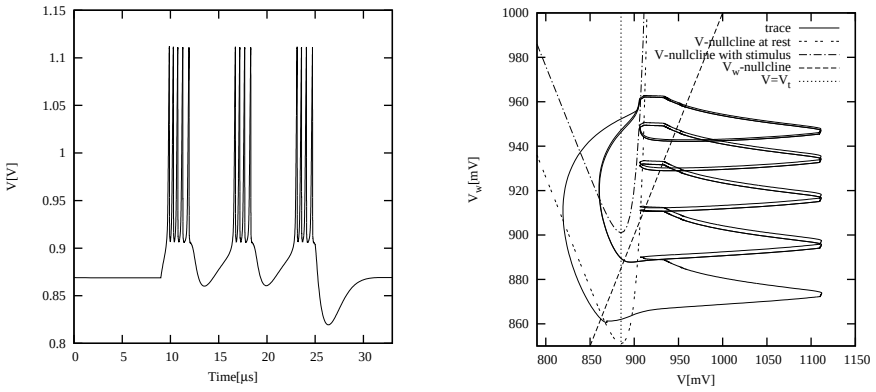
Larger margin

The phase plane plot in Figure 4.17 b) shows a much larger margin between the reset voltage and the V-nullcline than the phase plane plot from Figure 4.16. However, what remains is the exponential relation ship between the necessary V_w and a given reset voltage. A slight shift of the reset voltage will change the number of spikes in a burst if V_reset is large. In addition a too large value will inhibit adaptation from stopping the burst. Controlled bursting behavior could be observed between reset voltages of 885 mV and 908 mV. At $V_\mathrm{reset} = 910$ mV, the neuron will spike continuously during stimulation. When 915 mV are chosen, the neuron will even spike continuously without stimulus if V_reset is reached once.

Staying away from the exponential rise

The number of produced spikes dependent on the reset voltage can be found in Table 4.2. Below 900 mV, the spike number is stable. To achieve a larger stable spike number for small variations, smaller values of V_b can be used to achieve more spikes in a single burst. If only the last two spikes of a single burst reach above the minimum of the V-nullcline, the slope of

[9] The only difference relevant for this experiment is a stronger reset mechanism



(a) Membrane trace

(b) Phase plane, V_t and g_l have been chosen according to the simulations.

Figure 4.17: Simulation with more bursting friendly parameters with $V_\mathrm{reset} = 906$ mV. The exponential slope has been set to $\Delta_\mathrm{t} = 10$ mV and the operating range has been shrunk in comparison to Figure 4.16. The stimulation pulse height can be reduced to 300 nA

V_reset[mV]	First burst	other bursts
885	2	1
890	2	1
895	2	1
900	3	2
905	4	3
906	5	4
907	6	5
908	9	8

Table 4.2: Spikes in first and following bursts depending on the reset voltage.

the exponential is flatter and model and circuit are less sensitive to the reset voltage. This is consistent to the bursting phase plane plots shown in [12], indeed.

4.4.3 Conclusion

The designed hardware neuron is capable of reproducing most biological relevant patterns the AdEx model can reproduce, indeed. However, due to the small margin, limits are given when reproducing bursting behavior, depending on the quality of calibration and the total programming accuracy of parameters. Nevertheless, when staying close to the minimum of the V-nullcline, the variation sensitivity is reduced.

Real operation with synaptic stimulus

The neuron will not be stimulated by a current pulse in network operation. In contrast, stimulation is variating and consequently, the position of the V-nullcline is shifted. Different spiking patterns can be reproduced by a single neuron with one parameter configuration.

4.5 Compartmental Effects

Finite resistances and propagation delays

The connection elements described in 3.10 are observed here. When neurons are interconnected to build larger neurons, a finite conductance of 1 kΩ connects the membrane. In addition, the spiking signal is routed via transmission gates and buffers. Ideal switching of neurons would suggest an interconnection with 0 Ω and no delays for spike signal propagation. However, as the real interconnected hardware neuron behaves like a multi-compartment neuron, I talk of compartmental effects in this section.

The experiment has been executed on HICANN v2 d2 and d1. Shown oscilloscope traces are measured on HICANN d2. However, a similar experiment has been carried out on another HICANN v2 chip during evaluation of the interconnect ability of HICANN v2 neurons[10].

4.5.1 Methods

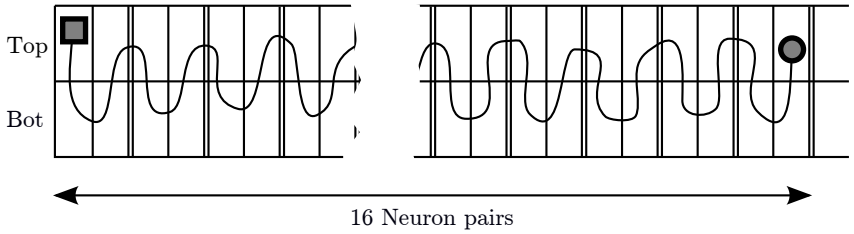


Figure 4.18: Simplified neuron connectivity overview showing connections for multi-compartment experiment. Each rectangle corresponds to one neuron. There are two rows - on in the each half of the HICANN. Membrane voltages of neighbouring neurons can be interconnected using transmission gates. The firing signal is routed through transmission gates(single lines) or switchable buffers. Connections are done longitudinal to the drawn line. The neuron is simulated and read out at the circle and read out at the square. Action potentials are detected at the square and propagated to the other tail of the neuron

Giant neuron

To see the results from the resistance and delay simulations done in refsec:nconn, 64 neurons are interconnected to form a large neuron. 64 is the maximum number of interconnected neurons aimed at network mapping during system design[56]. However, larger neurons are possible, but are not planned to be used.

Worst case routing

The interconnection scheme can be found in Figure 4.18. The neurons are connected as a line to create maximum effects of delay and resistance. This is a bad connection scheme for real operation indeed. For minimization of compartmental effects, the neurons of one chip half should be interconnected directly and all switches between top and bottom half should be set.

Spike detection and stimulation at opposing edges

Stimulation is done by a periodic current pulse at the right end of the neuron, while spike detection is done on the left end. Hence, the time impedance between stimulus and spike detection and the spike propagation time are maximized. The preferred arrangement in real operation would be to set the spike detection in the middle of the build up neuron.

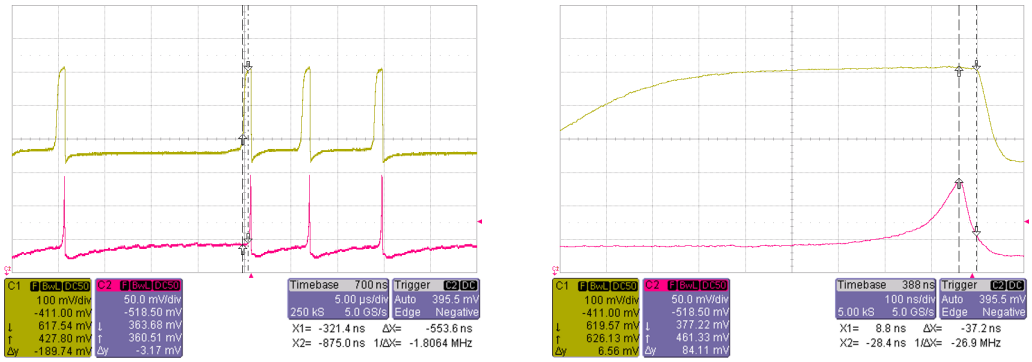
The readout of the membrane is done at both end of the neuron. Voltage traces are measures via an oscilloscope using a 20 MHz low-pass filter and 2-bit oversampling to reduce noise.

[10] Interconnection beyond pairs was not possible on HICANN v1

Neuron parameters are chosen to implement the complete AdEx model. No calibration has been applied. In addition, no dedicated parameter translation has been used. Measurements are carried out in HICANN d2.

For simplification, I will call the neuron on the left end square neuron and the one on the right circle neuron (See Figure 4.18).

4.5.2 Results



(a) One stimulus period **(b)** Detail of action potential

Figure 4.19: Measurement of multi-compartment effects on HICANN d2. C1 is the stimulated neuron(circle in Figure 4.18), while spike detection is done at C2(square). Notice the different voltage scales of both signals

Measurement results can be found in Figure 4.19. Large action potentials can be observed at the circle neuron, where the stimulus is injected. Indeed, the action potentials reach far beyond the Θ used for spike detection the square neuron. In fact, it is only limited by the operation range of the exponential term circuit.

However, further investigations have shown that the rise does not even start at the stimulated neuron, but some neurons more to the left in this case. The initial neuron for the exponential rise is given by device miss-match.

Voltage drop

The total resistance between the circle and the square neuron is supposed to be roughly 63 kΩ. An equal membrane voltage over the complete neuron in steady state, can be assumed. Hence, the voltage drop between the circle neuron and the square neuron is around 250 mV at max. This correspondents to a current of 4 µA which is close to the maximum current the exponential term can source. Nevertheless, the current is not injected by a single neuron's exponential term, but distributed between the terms of the large neuron.

Spike delay

The delay between the reset of the square neuron and the circle neuron can be used as a measure for the delay of digital spike propagation in the connected neuron. 37 ± 5 ns can be read form Figure 4.19 b)(The error is estimated). However, this measurement is very in accurate. A more precise measurement on HICANN d2 using the local maxima detection functions of the scope and higher resolutions results in 32.9 ± 0.4 ns. The same measurement has on HICANN d1 which results in 33.8 ± 0.3 ns.

Comparison to simulation

To compare with simulations, the worst-case value from Table 3.2, the worst-case delay (1679 ± 5 ps) has to be taken into account 16 times. Subsequently, one buffer delay(840 ± 2 ps) needs to be subtracted. We retain 26.02 ± 0.02 ns in the typical corner.

Simulations in the slow corner result in 32.57 ± 0.03 ns. Consequently, this result would suggest, the wafers would be in the slow corner.

However, process data is not available anymore. Nevertheless, it is unlikely that both wafers are in the slow corner. Furthermore, the simulation procedure could have been inaccurate as line resistances and capacitances and gate capacitances have been estimated. Summarized, the margins of the simulated and the measured delay are similar.

4.6 Fixed Pattern Noise

Fixed pattern noise of the output amplifiers has been analysed in 4.2. In addition, here the device miss-match created fixed pattern noise of the membrane capacitor, the current output of the leakage OTA, the reset voltage and the spiking threshold is observed.

4.6.1 Methods

Current stimulus as reference

The on-chip current stimulus is equal for every second neuron on one chip half. Hence it can be used for characterization. However, it can not be calibrated it self so far, so, only measurements in correspondence to a given current stimulus can be done.

Membrane capacitor measurement

The fist measurement sets all conductances of a neuron to low values to switch them off. A constant current is used as stimulus creating a saw tooth signal between the programmed Θ and the reset potential. With the measured amplitude A and the period T, the relationship C_m/I_{stim} can be obtained:

$$\frac{C_m}{I_{\text{stim}}} = A \cdot T. \tag{4.4}$$

However, as V_{reset} and Θ have to be measured, they are analyzed in addition. Due to the two different current stimuli, different relationships are expected for every second neuron as the two stimuli suffer fixed pattern noise, too.

Leakage biasing current

The second experiment measures the current sourced by the leakage term when the OTA is saturated in relationship to I_{stim}. The leakage potential is programmed above the spiking threshold. The distance to the threshold needs to be high enough to create a constant current output instead of a conductance behavior of the OTA. Another saw-tooth is generated by the mirrored current I_{gl} which is measured. The measured voltage slope is proportional to I_{gl}. It can be normalized by the capacitance measurements from the first experiment.

Both experiments sweep over all 256 neurons of a HICANN chip half. Measurements have been performed on HICANN d1 and d2, while most of the presented results correspond to HICANN d1. HICANN d1 and d2 belong to different wafers of the HICANN v2 MPW run.

Calibration software interface used

To facilitate automatic measurements, the experiment interface from the calibration software written by Marc-Olivier Schwartz is used to obtain membrane traces. This software interfaces the oscilloscope via python. Biological values are translated with a voltage scaling factor of 10 and a time scaling factor of 10^4. However, these scalings are not important in the measurement here.

The scaling switches of the I_{gl} biasing current is set to achieve a division of three.

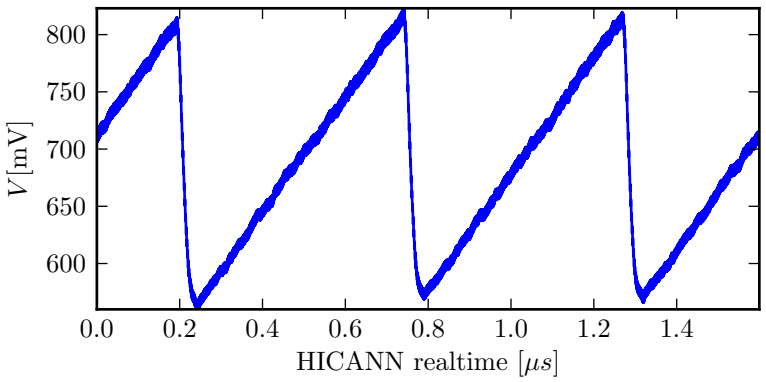


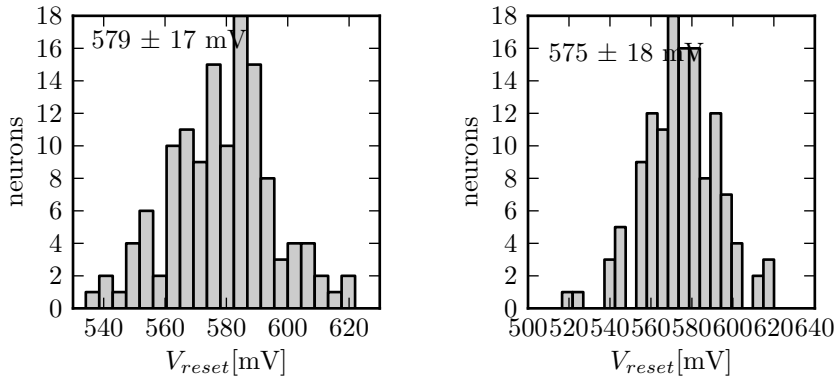
Figure 4.20: Detail of saw tooth signal generated by current stimulus, spike detection and membrane reset as used for capacitance measurements.(Neuron 0 on chip d1)

4.6.2 Results

Capacitance Measurement

Extracting amplitude and frequency

A sample saw tooth signal used for capacitance measurements can be found in Figure 4.20. It is measured with a 20 MHz low-pass filter, a sample rate of 2.5 Gigasamples and an ERes[11] noise filter of 2 bits. However, the signal remains noisy, so another averaging filter has been applied in data processing. The frequency is obtained by calculating the crossings between the signal and is mean value. The averaging filter size has been chosen to create only one crossing. Indeed, this procedure destroys the edges of the signal, so the raw signal has to be used for minima and maxima calculation. Minima and maxima are obtained for each period. Results are given as mean values over several periods.



(a) Even neuron numbers

(b) Odd neuron numbers

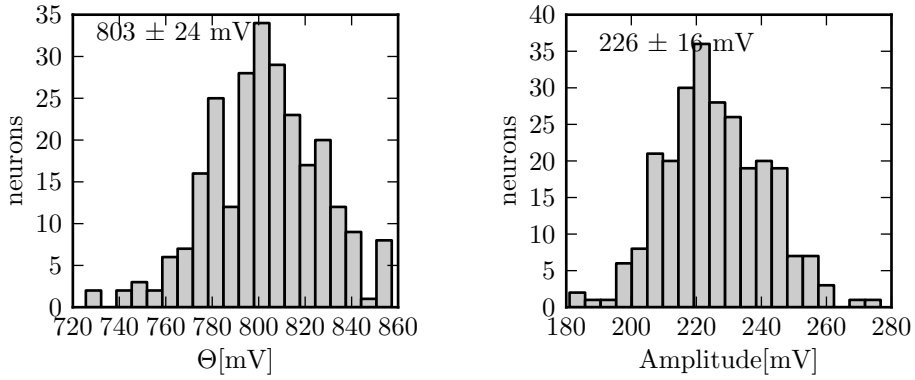
Figure 4.21: Distribution of V_{reset}

Distribution caused by neuron output amplifier

The distribution of V_{reset} can be found in Figure 4.21. Every second neuron is connected to one of the two global reset voltages. Hence the figure distinguishes between odd and even

[11] Enhanced Resolution, used oversampling to enhance the resolution by continuously averaging[89]

neuron numbers. The deviation of the measured values of V_{reset} is mainly caused by the deviations of the offset of the output amplifier of the neuron (4.2). Indeed, the difference of 4 mV is not very meaning full concerning the standard deviation around 17 mV. The reset voltage offset is already compensated during programming of the floating gates.



(a) Spike detection threshold **(b)** Membrane amplitude

Figure 4.22: Measured distribution of spike detection threshold Θ and membrane operating range

Elimination of fixed pattern noise by differencing

The spike detection threshold distribution is presented in Figure 4.22 a). Indeed, the standard deviation is large as it is created by basically by two different sources of fixed pattern-noise: The input voltage offset of the spike detection comparator and the input voltage offset of the neuron output amplifier. However, when calculating the difference between the Θ and the reset voltage, the most fixed pattern noise from the output amplifier is eliminated. Both measurements are influenced by the same output amplifiers.

The result can be found in the histogram of the amplitude in Figure 4.22 b). The reduction of the standard deviation is apparent.

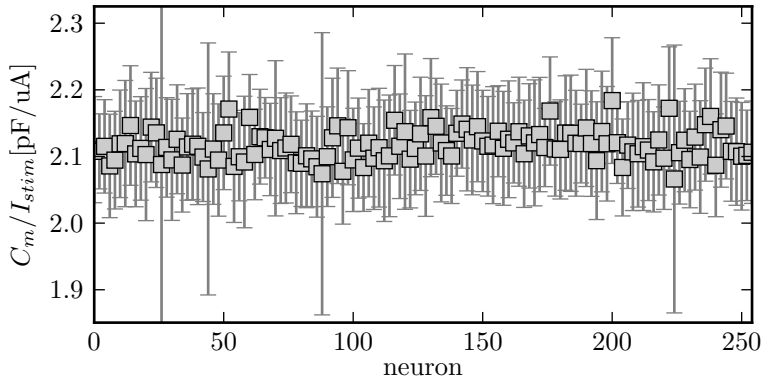
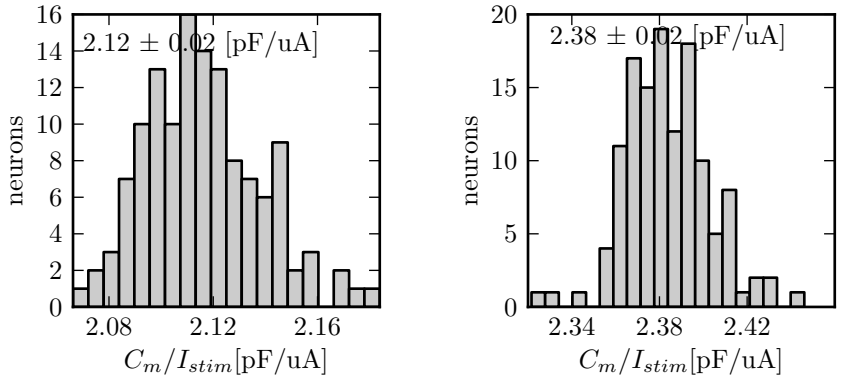


Figure 4.23: Membrane capacitance of even neurons in relation to current stimulus plotted against the neuron number

No gradients

Figure 4.23 and Figure 4.24 show the resulting relative capacitance. The trace in dependency to the neuron number points out(Figure 4.23) points out that there not gradients in



(a) Even neuron numbers **(b)** Odd neuron numbers

Figure 4.24: Histogram of membrane capacitance in relation to current stimulus

capacitance. A sweep of the neuron number is equivalent from sweeping from the left side of the chip to the right.

Precise results

Doe to the two different current stimuli, the histograms(Figure 4.24 divide between odd and even neuron numbers. The measurement of the capacitance values is more precise than expected. However, compared to the values given in the process documentation, it is still 10 times smaller. Hence the variation is basically given by measurement error. Capacitance measurements on HICANN d2 led to a similar precision. There the value for even neurons is 2.14 ± 0.02 pF/μA while odd neurons' capacitance can be determined 2.43 ± 0.02 pF/μA.

Absolute value larger than expected

The absolute value of the capacitance can be estimated by the given DAC value of 500 which would result in a current of roughly 1.7 μA. This leads to 3.60 pF respectively 4.0 pF as capacitance. However, this is a miss-match of a factor two in comparison to the designed 2 pF capacitors. Partially, this rise can be explained by parasitics. Internal gate and conductor capacitances of the neuron could add another 140 fF. Additionally, the capacitance of the line used for current stimulation can be estimated to 600 fF. Nevertheless, the value remains smaller than the measurement.

Significantly larger capacitances in the same amount on two wafers are unlikely. However, a systematic miss-match of the stimulation current would solve the problem.

The difference between the two capacitance value for odd and even neurons points at the fixed-pattern noise created miss-match of the stimulus current. Resistor matching is supposed to be worse than capacitance matching. The resistor used in the current source had to be laid out narrow, which is consuming less chip area. However, it degrades matching. Given the capacitances, the relative error of the current is 13 %.

Nevertheless, the matching of the relative capacitances for odd and even neurons between HICANN d1 and d2 is suspicious. Measurements on two different chips from two different wafers are close to equal within the error margins.

Better absolute matching than expected

Assuming that this is not a very unlikely random match, both, the absolute values of the resistor and the metal capacitors seem to be much better defined than expected by a chip designer. The variation of the capacitance supplied by the chip producer is 15 %.

Systematic miss-match

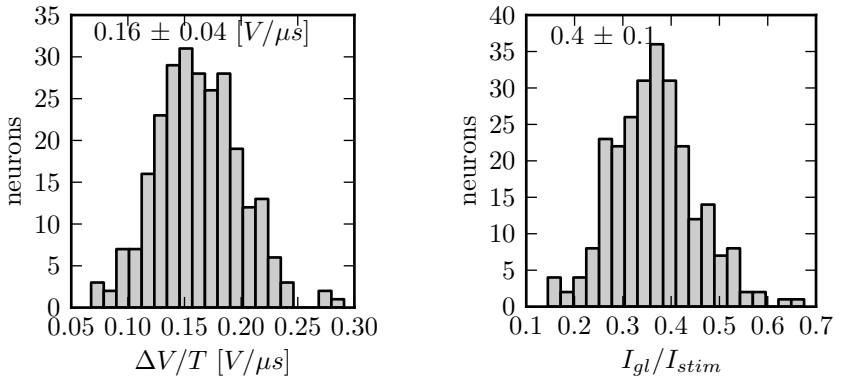
What remains is the miss-match between odd and even neurons. Large miss-match between the two resistors used for the two different current sources could be created by the fact that they are mirrored in layout. As the same machines are supposed to be used for the different

wafers, systematic miss-match of the current source could be created. Furthermore there could be a miss-match between odd and even capacitors due to very small differences in layout.

Voltage drop?

Another explanation for this result could be a voltage drop on the analog power supply between the left and the right side of the chip. This would result in a lower supply voltage for the DAC. Hence, the current would be smaller. However, it is unlikely to have the same voltage drop on two different chips. Furthermore the necessary voltage drop would have to be around 1/6 of the power supply to explain the miss-match. Indeed, voltage drops of this size would be fatal. However, measurements observing the floating-gates arrays, which included circuits supplied at the left and at the right side of the chip did not show any voltage drop in this scale[70]. Hence a large voltage drop can be excluded.

Current Measurement



(a) Voltage slope

(b) Corrected using capacitance measurement

Figure 4.25: Results from current measurements

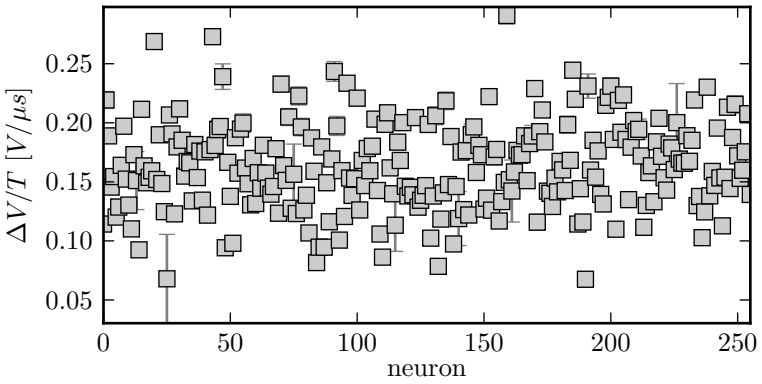


Figure 4.26: Voltage slope obtained from current measurements

Less variations than expected

Results from the I_{gl} current measurements can be found in Figure 4.25 and Figure 4.26. The corrected histogram is not divided into odd and even neurons as the standard deviation would only slightly change. The variation is large, indeed. However, Monte-Carlo simulations presented in Figure 3.35 suggest a larger variation even if only the parameter scaling current mirror was involved. In fact, this is not the case as the current is mirrored in the floating-gate cells, in the scaling current mirror and in the OTA it self. Consequently, a variation of 25 % is a good result in comparison to Monte-Carlo simulation.

No symmetry exhausting layout techniques in Monte-Carlo simulations

This miss-match to the simulation can be explained by applied layout techniques to reduce fixed-pattern noise. Those techniques cannot be included in Monte-Carlo simulation. Optimal regular structures are needed to generate data for a Monte-Carlo simulation. This optimal regular structure would inhibit layout techniques balancing gradients creating fixed-pattern noise.

Comparison of multiplicator value

Assuming perfect matching of the capacitors, the stimulus current used in the predecessor experiment was around 1 µA.Furthermore, the current programmed into the floating gates as source current for I_{gl} has is supposed to be 1 µA. Concluded, the relationship between I_{gl} and I_{stim} is a measure for the multiplicator of the bias scaling current mirror. It has been designed to 1/3 and Monte-Carlo simulations resulted in 0.34 ± 0.14. The measured value is 0.4 ± 0.1. Consequently, there seems to be a good matching between simulation and measurement. However, many assumption have been included in this comparison.

Conclusion

Fixed pattern noise characterization experiments have shown that the relative matching between devices is better than suggested by simulations, expectations and process documentation. Even the matching between capacitances from chips of different wafers could be shown which points on a good absolute matching of capacitors and resistors.

However, as expected miss-match is apparent and needs to be counterbalanced by calibrating the neuron circuits if close model connection is desired.

4.7 Simple Networks

A feasibility study

During the Capo Caccia Cognitive Neuromorphic Engineering Workshop 2012, I set myself the goal to create a simple network on a HICANN chip, as this had not been done before. As high level software is not able to build up larger networks so far, only networks using small neuron counts (2-8) have been used in earlier experiments. Right now, the network topology had to be hard-coded in the low level software which is a very inefficient, but working approach. Furthermore, unfinished calibration of the chip has been counterbalance by hand. However, once the software framework is evolved enough for larger networks and the calibration of the neurons is finished, the network demonstration presented here should be trivial. It can be understood as a feasibility study at this moment. The network experiment has been presented during a demonstration at the ISCAS[12] 2012[90].

4.7.1 Methods

Feed forward chain

A feed forward chain with four pools of eight neurons has been constructed(Figure 4.27). Each neuron fires on all neurons of the next group. Only the first 32 neurons of HICANN d1 are involved to simplify layer 1 routing configuration. Stimulation is done by the internal

[12]International Symposium on Circuits and Systems

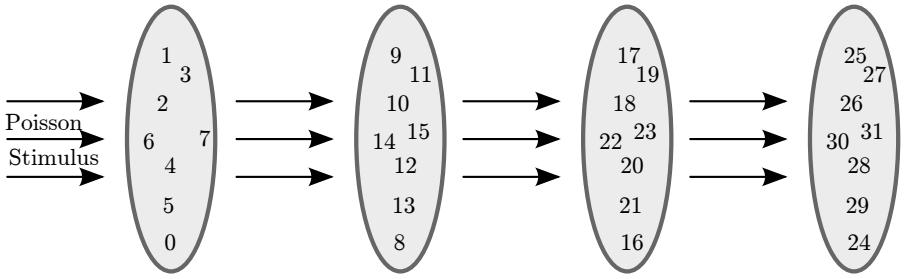


Figure 4.27: Network topology used in experiment. 4 Pools of 8 neurons have been interconnected feed forward. Each number corresponds to one neuron. Each neuron gets input from all neurons of the predecessor group(neurons have been renumbered for simplification.

digital spike sources on the chip, configured to Poisson behavior with a threshold[13] of 100. The chip is operated with a frequency of 200 MHz.

The experiment is carried out on a complete evaluation setup. Digital spike readout is done through the FPGA.

Hand calibration As calibration of the chip could not be completed for this experiment, continuously spiking neurons have been set to a large leakage potential by hand. This way, most continuously spiking neurons could be inhibited. The synaptic weights have been above the maximum possible weight of the synaptic input. Consequently, single input spikes might be sufficient to create an action potential at the post-synaptic neuron.

4.7.2 Results

Optimum result The results from the measurement can be found in the raster plot in Figure 4.28. The optimum result would be all neurons of a group firing in synchrony with a very short delay between the actions potential of the consecutive groups. In addition, the first group should fire as reaction to the stimulus.

Obtained result In the experiment some neurons show no response at all, while some spike without correlation to the stimulus. These unwanted spikes propagate to other neurons. Bad behaving neurons lack a proper calibration in this experiment.

Networks are feasible However, a possible observation is that there is some synchrony between the spikes of a group, between groups and to the input spike train. The neurons are obviously connected in a network. It is possible to generate larger networks on a HICANN chip.

4.8 Reproducing Computer Simulations

Although, it is not the task to compete with the precision of a computer simulation, a comparison has been done for the demonstration at the ISCAS 2012. It has been published in [90]. The presented measurement has been carried out by Alexander Kononov. Computer simulation scripts and calibrated neuron parameters have been provided by Marc-Olivier Schwartz.

[13] Poisson stimulus is generated by picking pseudo random numbers from a linear feedback shift register which are compared to a threshold.

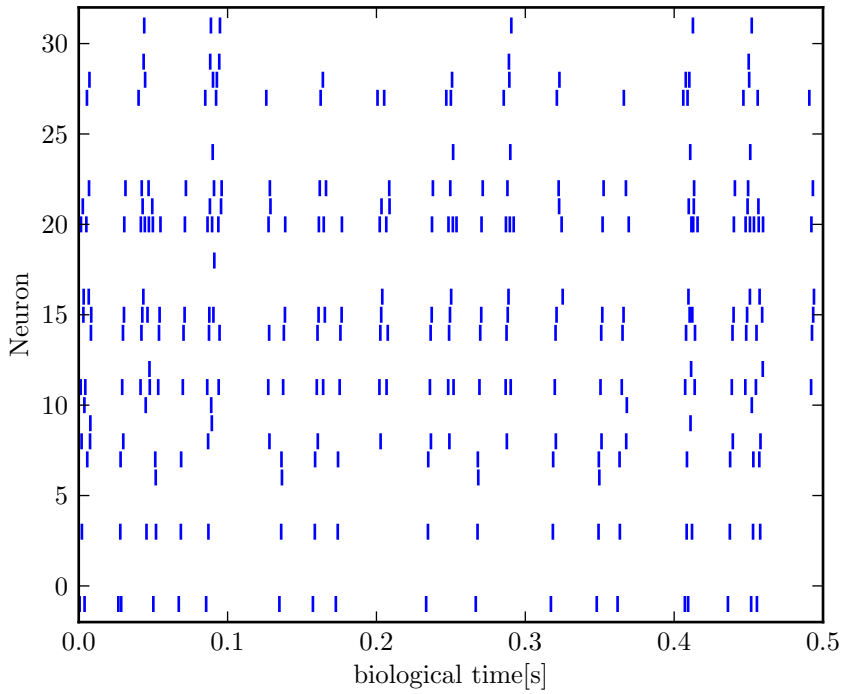


Figure 4.28: Raster plot of read out spike events of the programmed feed forward chain. Neuron number -1 is the stimulus. Neurons 0-7, 8-15, 16-23 and 24-31 are the individual groups. Each line corresponds to an action potential of the corresponding neuron.

4.8.1 Methods

For simplification, an integrate-and-fire neuron model (no adaptation, no exponential) has been emulated in this experiment. A PyNN[91] script is used to create the Poisson stimulus data and to simulate the model. As simulator back end, the neural simulator BRIAN[92] has been used. Neuron 5 on chip iscas2 has been examined for the presented result. The neuron number is a random pick from the calibrated neurons of this chip at the time of first experiment execution. As time-scaling factor 10^4 has been applied. All necessary parameters of this single neuron have been calibrated by Marc-Olivier Schwartz. All synaptic weights are equal. However, hand-tuning has been used to retain the weight. Interfacing the chip is done through the low-level software interface.

4.8.2 Results

Close matching of spike times

Figure 4.29 presents the results. The corresponding output voltage scaling factor of the used output amplifier is 2.15. The matching between the action potentials of emulation and simulation is apparent. However, large differences can be observed between the membrane voltage traces when looking at the details. After the spikes at roughly 40 µs and 45 µs, there is a large PSP[14] in the model and no PSP in the emulation. The hardware action potentials

[14]Post synaptic potential - membrane voltage reaction on synaptic input

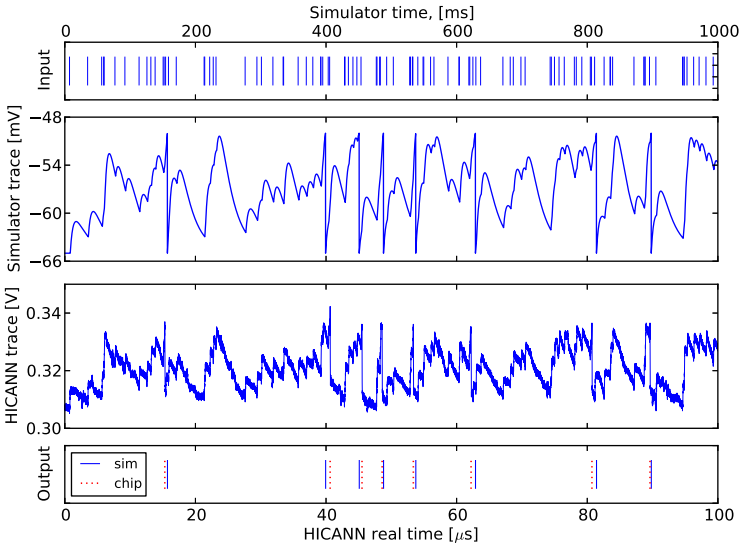


Figure 4.29: Comparison between software simulation(2nd. plot) and hardware emulation (3rd. plot) stimulated by Poisson stimulus (top) via synapses. There is a good timing match between the hardware action potentials and the ones created by the simulation (bottom). The time-scaling factor between emulation and simulation is 10^4. (modified version of figure published in [90])

are some time of at this point indeed and the PSPs are inhibited by the action potential and the refractory period.

Current-based synapses

Another difference is the shape of the individual PSPs which are α-function shaped in the simulation, while the rise in the emulation is much sharper. This is a sign for current based membraned dynamics instead of conductance based dynamics due to a large distance to the reversal potentials.

PSP height simulation

When comparing the size of the first PSP in hardware and simulation, a close matching can be observed. The total operating range of the simulation ($\Theta - V_{\text{reset}}$) is 15 mV. Subsequently, the height of the first PSP obtained from simulation data is rounded 3.96 mV. Hence, the relative PSP height of the simulation is 2.64.

PSP height measurement

Obtaining the operating range from the hardware measurement is harder, as the spike detection threshold Θ and V_{reset} differ between different spikes. This is supposed to caused by coupling on the output lines. The other output amplifier was used as output for the input spike representation in the chip. This signal is created from a digital signal, hence high signal level differences occur creating a large coupling. However, the operating range of the neuron can be estimated as 27 ± 2 mV which corresponds to 54 ± 4 mV with applied output voltage divider compensation. Next, the size of the first PSP in hardware can be estimated as 7.5 ± 1 mV which is 15 ± 2 mV with compensation. Concluded, the relative PSP height in hardware is 0.28 ± 0.05.

The relative height from simulation lies within the error margins.

4.8.3 Conclusion

The hardware emulation's capability of reproducing results from a computer simulation has been shown. However, this result is academic as explicit and accurate reproduction is not the goal of the emulation. In addition, the synaptic weight of emulation and simulation had to be adjusted. Synaptic weights will differ between different synapses of a neuron if the synaptic input is not driven into saturation. A calibration of all synapses is not realistic from experiment time frame as well as from synaptic weight resolution.

Nevertheless, the results nicely show the power of the implemented circuit.

5 Discussion: Single-compartment

A single compartment analog VLSI implementation of the Adaptive Exponential Integrate-and-Fire neuron model has been presented so far. The use of operational transconductance amplifiers allows a direct translation of model equations into circuits. Consequently, there is a strong correspondence between the neuron model and the emulation. I will reflect and discuss the outcomes of model implementation and measurements. A comparison to implementations from literature is performed.

5.1 Model Implementation

Linearity

The implementation of each single term of the model equations has been discussed in theory and in simulations. Imperfections of the circuit have been outlined. The main trade-off is the limited linear range of operational transconductance amplifiers. However, when staying in a shrunk operating range, mathematical model correspondence can be kept. In addition, comparison to ideal circuit elements have shown only little deviations for larger operating ranges than the OTAs linear range. Nevertheless, linearity issues have been a faithful companion when analyzing the neuron's circuits. When using the model, a trade-off between better voltage parameter accuracy with larger noise robustness and a better linear behaviour has to be taken.

Adaptation

The adaptation term is implemented straight forward using two OTAs. The model's variable adaptation variable w has been replaced by a new voltage V_w for this purpose. This new variable follows the membrane voltage with a large time constant and is exposed to the same linear range issues. In comparison to other implementations of adaptation[42,45], which focus on spike-frequency adaptation, the implementation includes subthreshold adaptation as implemented in the AdEx or in the Izhikevich model[21]. A lack of subthreshold adaptation would disable biological neurons behaviour like subthreshold oscillations or inhibitory rebound spikes created by the removal of inhibitory signals[5]. The strength of the spike-frequency parameter b has been dimensioned too small in the current chip version HICANN v2. However it can be compensated by choosing a larger total impact of adaptation by using a larger parameter a. In addition, the translation of parameter b has been adjusted in for the next chip version HICANN v3.

Synaptic input

Electronic current pulses created by synapses are translated into synaptic conductances in the synaptic input circuit using an integrator and two OTAs. The circuit is limited by a nonlinear resistance used in the integrator circuit which creates a weight dependency of the synaptic time-constant which is a strong deviation from the model. In addition, the linear rage of the used OTAs limits the conductance behavior of the circuit. Nevertheless, it is possible to fit post-synaptic potentials(PSP) created by conductance based synapses onto the PSP created by the circuit.

Delays

Three analog computations are performed in the synaptic input. Each creats a delay which needs to be taken into account. Measurements have shown delays of 60 ns for the complete analog path including synapses. This delay corresponds to a biological real-time delay of 0.6 ms in 10^4 operating mode. Additional delays created by digital event transportation are

in the same order of magnitude or larger.

Impact The impact of the synaptic input is indeed limited by the maximum conductance of the OTA and its maximum bias created by the first OTA. About 5 µS is the maximum peak conductance of the circuit. This conductance has to be compared to the total conductance of the neuron - a smaller leakage time constant results in less impact of the synaptic input. When several compartments are used the impact can be enlarged as several synaptic inputs are used while leakage conductances might be switched of in single compartments.

Resistor implementation The resistor implementation in the synaptic input is a clear candidate for an improvement in future chip revision. On the one hand because of the lack of linearity. On the other hand because of the necessary voltage biasing which is very sensitive to fixed-pattern noise and makes calibration inevitable. However, so far the neuron is operating and programmable with the current resistor implementation.

Exponential term Next, the exponential term has been presented. It uses the subthreshold characteristic of a MOSFET to create the exponential characteristic. Channel-length modulation is used to implement an adjustable voltage divider with low crosscurrent. In contrast to the exponential term implemented by Indiveri in [45], slope and exponential threshold are directly adjustable. The positive quadratic feedback of Wijekoon's circiut [42] (implementing the Izhikevich model [21]) is similar to the one from Indiveri, but above threshold and hence quadratic and not parameterizable.

Parameterization A key feature of the neuron circuit is its parameterizability according to the parameterization of the model. To achieve a large parameter range, the range of critical parameters has been enhance by the introduction of global switch-able current mirrors. Indeed, this is not the most elegant implementations but it full-fills the constraints given by the needed operating regimes for time-scaling factors between 10^3 and 10^5.

5.2 Measurements

In the measurement chapter 4 I have presented several single neuron experiments and a small network experiment. Fixed-pattern noise of the circuits has been analyzed in detail.

Fixed-pattern noise Experiments showed that there is indeed miss-match which needs to be counter balanced by a calibration procedure. On voltage parameters connected to the input stage of OTAs a standard deviation of 16 mV can be expected. It is created by the input voltage offset of the OTA. The input voltage offset of the neuron has been measured to 17 mV in the critical voltage range. Its impact could be reduced to below 5 mV by a straight forward calibration measurement.

The deviation of the membrane capacitance has been measured to less than 1 % which is a very good result given the measured signal. However, the deviation given by the chip producer is even smaller 0.1 %. Accordingly, the measured deviation is dominated by measurement errors.

Subsequently the biasing current deviation of the leakage OTA has been determined. The result was 25 % which is good in comparison to the 33 % obtained by a Monte-Carlo simulation.

Reference simulation The first biologic experiment reproducing the reference simulation used during design reflects fixed-patter noise measurements. An apparent deviation between simulation and measurement could be observed which, however, could have been explained by different parameters due to miss-match.

Characteristic patterns Results, we published at the NIPS conference ([44]) have been reflected and enhance by a more detailed phase plane analysis including nullclines according to simulation parameters. Experiments involved characteristic patterns like spike-frequency adaptation, tonic spiking

and different bursting patterns. An outcome of the additional analysis was a too small value of Δ_t during the experiments which complicates patterns like bursting. In the experiments done for [44] bursting behavior changed between runs in virtue of programming variation of the reset voltage.

Bursting

To improve the bursting behaviour, a more reasonable value of Δ_t has been chosen and a simulations have been redone. Bursting behavior seemed to be more stable to the reset voltage. However, the expected remaining model created sensitivity to precise values of the reset potential has been shown by an analysis of the spike count in each burst. Nevertheless, a different approach of bursting staying below the v-nullcline until the last action potential has been suggested according to [12]. This way, the behaviour is less dependent on the reset voltage.

Network operation possible

The last two experiments, done for demonstrations at the ISCAS, showed simulator-like behavior of the model and a small network experiment. They can be understood as feasibility proofs since the complete network architecture is involved. Neurons receiving synaptic input and creating action potentials which are transported via L1 to other neurons or read out via L2. L2 is used for stimulation in addition.

5.3 Comparison to Other Implementations

The design approach of accurately implementing the AdEx makes a large difference in comparison to other existing neuron emulations.

Transistor count

Indeed, the number of used transistors and the circuit complexity is huge. Each single OTA consists of 17 transistors and there are 7 OTAs in the full neuron circuit. [43] uses only 14 transistors for the complete neuron circuit for instance. Especially the current-mode circuits from [41, 45] and the HHM neuron implementation from [39] have the beauty of reduced complexity. In total, around 200 transistors are used for the designed AdEx implementation. Apparently, a large number of different circuits enlarges design effort drastically.

Nevertheless, the actual size of the neurons is small in comparison to the total size of the HICANN chip. The size-critical elements in our implementation are the synapses. Indeed, this less constraints reduce design effort. However, it shrinks comparability to less complex neuron implementations.

Current consumption

In addition, the use of voltage-mode circuits, usually working in strong inversion, consumes more power than a current-mode design as implemented in [41, 45]. For the spike-frequency adaptation experiment from Section 4.4 the current consumption on the analog power supply is 50 mA for instance. Accordingly, the necessary power is about 10 µW. In contrast, current-mode subthreshold neurons can achieve 50 nW to 1 µW [46]. Nevertheless, if the power consumption is divided by the acceleration factor, similar consumptions are reached with the presented circuit.

Chip power consumption

Looking at the whole HICANN chip, the current consumption would result in 25.6 mA which is small in comparison to the maximum current consumption of the chip, specified to 1 A in [61]. Neurons take only a small part in the total chip's power consumption. Power consumption depends on parameterization.

More transistors for a reason

The price of a larger transistor count and more power consumption has not been paid without a reason. A circuit allowing direct translation between model and circuit has been gained. Indeed reducing parameterizability, function and model correspondence would reduce the circuit's size and current consumption. None of the discussed models from literature can compete with the parameterizability of the presented circuit in any way. This way, the circuit can directly be referenced by the model and basically all complex behaviours of the model can be emulated using the circuit.

Realistic, scalable networks are necessary

Another very important aspect of the other implementations which is not a direct neuron issue is scalability and integrability. Although, single neurons might be scalable, they are usually not developed in a scalable system like the BWS which limits network sizes. Reasonable network sizes with plastic configurable synapses are usually not possible or not even targeted at. This clearly limits biological relevance and possible applications for a deeper understanding of the brain.

Model correspondence is necessary

In my opinion, understanding the brain using analog neuron emulations is only possible if a close model correspondence is kept and biologically realistic networks can be constructed. Comparison to experimental results from modeling or biology would hardly be possible otherwise. Indeed, this contradicts a circuit-driven design methodology. When designing an analog neuron, lots of compromises have to be taken. However, if those are taken without a model in mind, biological correspondence might be lost.

Circuit-driven designs in the far future

Nevertheless, once the neuron and the brain are understood completely, it might be a good approach to design a specialized analog neuron to exhaust the computational power of the brain. In this case, the work presented in [39] is the best approach in my opinion if it would be implemented in a scalable network. However, I do not expect this to happen in the next 10 years.

In the next part of this dissertation, the presented single compartment implementation will be extended to a multi-compartment model implementation.

6 Multi-Compartment Emulation

This chapter enhances the AdEx emulation presented above to a multi-compartment emulation. I start with an introduction into theoretical concepts followed by a collection of implementations currently available. Next an overview of the new circuit concept is presented. Subsequently, each circuit part is discussed in detail. Finally the ASIC Multi-Compartment Chip is presented.

6.1 Biological Concepts

Here I give a very brief introduction to multi-compartment modeling and its consequences in comparison to point-neuron modeling. For further reading, I can recommend the book "Biophysics of Computation" by Christoph Koch[6] and the references given in this section.

6.1.1 Cable and Compartments

Point neurons

Figure 0.1 displays an example of biological pyramidal neuron. Indeed, it sound like a joke on theoretical physicists (spherical cow) to map this complex structure onto a single point looking at the photograph. Nevertheless, models like the initial HHM are usually imlementing point neurons. However, the assumption seems definitely warrantable if only the soma is taken into account. Due to smaller longitudinal resistances and the active channels creating action potentials, the soma can be assumed as an equipotential sphere. Hence the soma could be modeled by a single point or compartment. Compartments are equipotential sections. A reduced model using only a single compartment is much easier to analyze, understand, parameterize and simulate. Reduction of complexity by the creation of a model is one of most basic principles of science. However, the reduction must be kept in mind. The apparent complexity of a single neuron cannot be ignored if its complete function is to be discovered.

Modelling the dendrite as a cable is a model closer to "biology". First cable analysis have been done by Wilfrid Rall[93, 94].

Cable equation

Let the dendrite be a single line of infinite length and constant diameter. An ansatz similar to transmission line analyses is possible. I follow a derivation similar to the one presented in [5]. With a constant transversal leakage conductance g_l per length, a constant membrane capacitance c per length and a stimulus current density $i_s(x)$ the system can be divided into compartments of infinitesimal length (See Figure 6.1).

Applying Kirchhoff's current law for the single infinitesimal compartment of Figure 6.1 results in (longitudinal currents flowing to the right; u is time dependent):

$$\frac{u(x-dx)-u(x)}{r_l dx} - \frac{u(x)-u(x+dx)}{r_l dx} = u(x) g_l dx + \frac{\partial u}{\partial t} c dx - i_s dx \tag{6.1}$$

$$\frac{\partial u(x-dx)}{\partial x} - \frac{\partial u(x)}{\partial x} = u(x) g_l r_l dx + \frac{\partial u}{\partial t} c r_l dx - i_s r_l dx \tag{6.2}$$

$$\frac{\partial^2}{\partial x^2} u(x) = u(x) g_l r_l + \frac{\partial u}{\partial t} c r_l - i_s r_l. \tag{6.3}$$

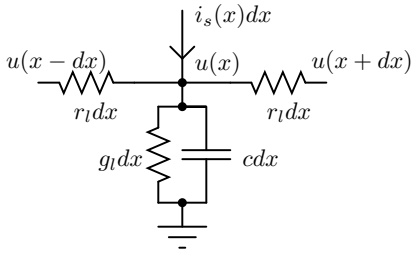


Figure 6.1: Infinitesimal dendritic compartment

Introducing the electronic length scale $\lambda = \sqrt{1/(glrl}$ and the membrane time constant $\tau = c/gl$ results in

$$\tau \frac{\partial u}{\partial t} = \lambda^2 \frac{\partial^2}{\partial x^2} u(x) - u(x) - i_s/g_l. \tag{6.4}$$

This is the voltage cable equation for a dendrite modeled as a cable. The term with the squared derivative is responsible for longitudinal current diffusion. Setting the longitudinal resistance to infinity would cut it off completely and the equation transforms to an integrator's equation with an exponential decay.

For a delta stimulus $i_s(x) = \delta(x)A/g_l$, the stationary solution of this equation is[5]:

$$u(x) = \frac{1}{2} e^{|x/\lambda|}. \tag{6.5}$$

Hence the electrical length scale λ is the distance where a signal drops down by $1/e$.

With a Greens function approach, the cable equation is analytical solve able for the dynamic case. Solution and derivation can be found in [5].

Limits of analytical solutions

However, the assumptions made for easier mathematical description are not given in a real neuron. The neuron in Figure 0.1 has branches dendrites with decreasing diameter with distance to the soma. Furthermore, a solution of the cable equation is not practical, if active channels like synapses or voltage controlled conductances are added.

Multi-compartment modeling

The solution is to go one step back. Instead of working with infinitesimal sections, larger compartments are used for modeling. This compartments can have different sizes and branches can be constructed. Exaggerated the solution is to work with several interconnected points instead off a single point neuron. Indeed, each reduction has consequences for the computational power of the neuron. However, a trade-off is necessary at some point.

6.1.2 Passive Computational Power

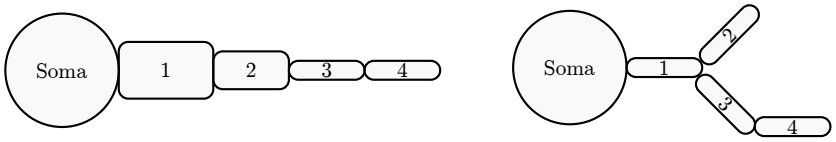
Equalizing of structures

At this point I want to discuss the influence of dendritic morphology on the computational properties of the neuron. For this purpose, two sample neurons are shown in Figure 6.2. When mapped two a point neuron model, both neurons would be identical if the membrane size is matched.

Attenuation and delaying

Looking at Figure 6.2 a), synaptic input is weighted by the dendrite. The impact of excitatory input at compartment 1 on the soma is larger than the impact of input at the dendritic compartment compartment 4. The first experiment done in the next chapter will demonstrate this effect[1]. The dendrite attenuates and delays the synaptic input. Delays

[1] Indeed without active channels, apical dendrites of some neurons would not have any impact at all [6].



(a) Linear **(b)** Branched

Figure 6.2: Two artificial sample neurons. The dendrites are passive except for conductance based synapses.

are in the order of magnitude of the membrane time constant and hence significant[95]. In addition, similar open synaptic conductances will have a larger impact on the voltage level of smaller apical compartments than on larger compartments. Hence, the total amount of moved charge is smaller for an apical dendrite for an equal conductance.

Silent inhibition

A simple example of compartmental effects is silent inhibition. If a neuron's membrane is close to the resting potential, inhibitory synaptic input hardly changes the membrane voltage. However, the conductance of the membrane changes and incoming excitatory input would be inhibited. Assuming both, inhibitory and excitatory input would be induced at compartment 1 excitation might be shunted by inhibition. On the other hand, if the inhibitory stimulus would be located at compartment 4, the longitudinal resistance weakens the effective inhibitory conductance and the excitatory effect would be larger.

Coincidence detection

The neuron in Figure 6.2 b) could work as a coincidence detector. The work from [95] gives an example with bipolar neurons with two dendritic branches in chickens for coincidence detection of auditory signals. However, the same effect might occur in neurons with branched dendrites as shown in Figure 6.2 b)(See [96]). With conductance based input, the transfered charge of two coincident inputs in a single compartment is smaller than the transfered charge of delayed synaptic inputs or inputs in different compartments. Consequently, a coincident input at compartments 2 and 3 Figure 6.2 b) might result in an action potential while the same input concentrated in compartment 1 does not. By setting the inputs to compartments 2 and 4, coincidence of the input at compartment 2 and a delayed version of the input of compartment 4 could be measured. Improvement of coincidence detection for bipolar dendrites has been shown in [96].

In more complex dendritic trees, local computation might occur[95]. The results of local computations would be summed at the branching points of the dendrite.

Local memory

Another possible effect of passive dendrites is created the lack of a pull down of the dendritic membrane potential after an action potential due to missing or not triggered voltage-gated potassium channels. In contrast to a single-compartment model without an additional adaptation variable, the information condensed in the membrane voltage of the dendrite can be kept in a multi-compartment model. This memory effect can result in behavior like bursting for instance (See 7.2). In addition, local computation results could be kept.

6.1.3 Active Channels

Voltage-gated channels

Voltage gated active channels can be found in dendrites. The zoo of ion-channels includes voltage-gated sodium channels, high and low-voltage-gated calcium channels and potassium channels. However, local channel density differs between different sections of the dendrites. For fast sodium channel density in dendrites is orders of magnitudes lower then the density at the soma (See [6]). Voltage-gated channels are necessary as influence of synaptic input at

apical dendrites would vanish due to attenuation in some larger neurons[6, 97]. Nevertheless, active dendritic channels are common to be ignored in modeling due to the enlargement of complexity and the missing information of channel densities[5].

Several reviews discuss the effects of active channels on dendritic computation[95, 97, 98].

Amplifying distal input

The distance to the soma attenuates the efficacy of synaptic input indeed. In some large pyramidal neurons(layer 5 cortical pyramidal neurons [97]) the apical dendritic sections are so far from the soma that synaptic input would not have any influence. To counter balance, this effect, synaptic input can be amplified by voltage-gated channels. This amplification can be subthreshold enlarging the size of a single PSP or above threshold creating global or local dendritic action potentials. Further amplification can be achieved through the use of NMDA synapses[97].

Dendritic action potentials

Dendritic action potential are usually calcium spikes which are broad in comparison to the sodium spikes at the soma. The calcium channels naming those spikes are slower than the active sodium channels. They open close to the edge of an action potential and remain open during the pull down of the membrane voltage. Pull-down is realized by potassium channels which depend on the concentration of calcium[75]. Global dendritic action potentials are strong enough to be transmitted to the soma and create a somatic spikes. Local spikes only amplify synaptic input and do not necessarily result in an action potential at the soma. Large neurons can have a second spike initiation zone with a high density of calcium channels at the main branch of the apical dendrite[97]. An calcium action-potential would result in a somatic spike.

Dendritic spines

Local spike amplification could even occur in single so called dendritic spines[6]. Spines are small outgrowth orthogonal to the main dendritic branches(Have a close look at Figure 0.1. Several or single synapses connect directly to spines which connects to the dendrite. Apparently, having spines as single compartments is to complex for modelling of complete neurons. In addition, realistic parameterization of such models would be a hard task. However, modelling of single spines might be necessary for a complete comprehension.

Back-propagating acton potentials

Another important effect of active channels is the active back-propagation of action potentials which could be the missing messenger for triggering of STDP in synapses. This propagation could be realized by fast deactivating voltage-gated sodium channels in the dendrite[95].

6.1.4 Which Model to Use?

Looking at potential users

I ask this question in analogy to the same question asked by Eugene Izhikevich in[11]. It is necessary not only to have a reference for emulation. In addition, the step to use a multi-compartment implementation will be easier for modellers if the implemented model is common. The discrepancy lies in the level of complexities. The neuron model implemented in the HICANN is the AdEx, which is a reduced phenomenological model. On the other hand, multi-compartment modeling enhances the complexity of models drastically.

Few publications with less complex single compartments

The only publication known to me using the AdEx as a basis for multi-compartment modeling is work done by Claudia Clopath[99]. A single passive compartments is added to the AdEx. Larger multi-compartment simulations rely on Hodgkin and Huxley style neuron models as a more complex single compartment model would be the first logical step to enhance biological realism. Reference models used in this dissertation are [100] and [101].

AdEx as basis

However, the question is rhetorical as the HICANN is the basis of the multi-compartment implementation. The final necessary choice is to use the AdEx as basis for each single compartment. In the first implementations, no additional active channels will be added. Nevertheless, the exponential term is available in each compartment.

Complex models for dendrite structure

Complex models are used to extract the size of individual compartments and the dendrite structure only. This information can be extracted easily, as axial resistance and compartment capacitance are directly given even in complex models[100].

6.1.5 Where to Cut?

Level of complexity

Indeed, even a multi-compartment model is only a reduction of a real neuron. With increasing compartment count, realism of the model might be enhances. However, parameterization complexity and the danger of building something which does the right things for the wrong reasons rises. Over fitting is an issue in multi-compartment modeling. In addition, the complete distribution of ion channels along a dendrite is unknown.

Hardware constraints

At least, when building a hardware emulation, the choice of complexity is limited by hard constraints. Usually compartment capacitances have to be implemented using real capacitances and the size of real capacitances cannot be variated without limits in a chip. Although generally, complex models using up to 512 compartments would possible on a multi-compartment HICANN, compartment size scaling limits the realistic number. As educated guess I would propose 32 compartments as a reasonable compartment count for a hardware neuron.

Possibly high compartment number needed

The work done in [102] observes the influence of compartment and branch reduction of a complex model. Especially when working with unbranched models active channel dependent effects like a back-propagating action potentials disappear. The outcome is basically that reduction is hardly possible.

Active propagation between compartments

However, when action potentials are propagated on the membrane of a real neuron, voltage-gated channels are opened one after another like an avalanche. Indeed, with an R-C delay between those channels, this avalanche effect would be attenuated. This effect seems to be a weak point of current multi-compartment modeling. Nevertheless, adding voltage-gated channels depending on the voltage level of neighbour compartments would solve the issue. However I stick to conventional multi-compartment modeling in this work to keep correspondence.

6.2 Multi-Compartment Implementations

Several different implementations of multi-compartment neurons on ASICs exists in literature. A collection is presented in [46] for instance. The main difference of the concepts is the implementation of the inter-compartment conductance as the implementation of adjustable resistors is nontrivial on a micro-chip. However, not all concepts implement an adjustable resistor at all.

Switched capacitor

Early work by Christoph Rasche and Rodney Douglas [103] which is based on concepts of [104] use switched capacitor circuits to implement the inter compartment conductance and OTAs as transversal conductances. Different dendrite structures are implemented to prove the concept of multi-compartment emulation in analog VLSI. They show the effect of attenuated synaptic input at different dendrites. However, the scalability of switched capacitor implementations is limited indeed. Different clocks are required to implement different switched capacitor resistors. The clock frequency needs to be high enough to reduce noise. The concept of [103] is hardly configurable, however, it is a proof of concept.

Floating-gates

The group around Jennifer Hasler developed a programmable neuron array based on their neuron presented in [39]. As longitudinal dendrite conductance, the work in [105–107] uses PMOS transistors biased in subthreshold regime. These transistors are implemented as local

floating-gates[2]. Dendritic compartments are arranged in a two dimensional matrix. Next neighbour connections can be used to create complex dendritic tree structures. The use of local floating-gates allows a high configurability. Nevertheless, different compartments sizes can only modeled by different conductances.

Current-mode

The concept presented in [108] by John Arthur and Kwabena Boahen uses current-mode low-pass filters as basic neuron element. Calcium concentration and voltage dependent active calcium channels and calcium-dependent potassium channels are implemented. The dendrite topology it-self is fixed with non adjustable compartment sizes. In addition, no conductances between dendritic compartments are implemented - dendrites directly interfere the soma unidirectional.

Bidirectional, current-mode

However, based on the work of Arthur and Boahen, Yingxue Wang and Shih-Chii Liu developed a new circuit presented in [109]. This is the most modern implementation in literature. The conductance between compartments is emulated by mirroring the output current of the current-mode low-pas filter onto the next compartment. Indeed, this connection would be unidirectional, but a current feedback from the next compartment allows bidirectional connections. The cable constant λ can be adjusted. However, it needs to be constrained to assure stability. Spikes from somatic compartments are feed back into dendrites to allow back-propagating action potentials. Dendritic compartments are arranged in an array of 3 times 32. Each compartment can be connected to 3 of 32 somatic compartments. This structure is a trade-off between a larger number of neurons and more complex neuron structures. Routing of dendritic tree compartments is limited to neighbour compartments. The size of individual compartments cannot be adjusted. In [110], Wang and Liu present first biologic deductions from experiments with chip. The authors try to map the concept on biology by the similarity that both system have sigmoid functions as transfer function.

6.3 Circuit Structure and Concepts

The arrangement of the neurons in the HICANN is already predestined for a multi-compartment implementation. In addition, some compartmental effects like a longitudinal resistance are already apparent (See 3.10 and 4.5). However, those effects are parasitic, unwanted and hardly controllable.

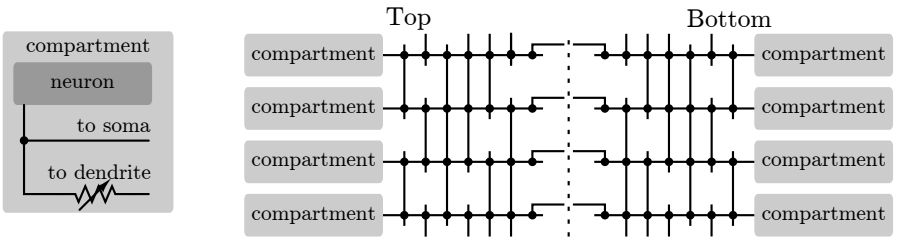
To enable realistic multi-compartment emulation in a HICANN-like ASIC, several steps have to be performed:

1. Add an directed interface module with an adjustable conductance to connect to other compartments.
2. Develop a routing scheme to allow complex tree structures.
3. Enhance parameterization to account for size differences between compartments (e.g. soma and dendrite).
4. Implement active dendritic channels (back-propagating action potentials).

Compartment interface

In contrast to the neurons of the HICANN which correspond to points, each compartments has two ends. One end pointing to the dendrite and one end pointing to the soma. At least one of these interfaces to other compartments needs to implement a controllable conductance(See Figure 6.3 a)). Furthermore, the compartment interfacing circuit a new spike routing scheme.

[2] In contrast to the floating-gates presented in this thesis, which create biases for analog circuits, their floating-gate transistors are the transistors of the biased analog circuit.



(a) Compartment interface

(b) Dendrite routing. Each horizontal line corresponds to somatic and dendritic line. Each dot corresponds to two switches

Figure 6.3: Interconnecting compartments

New routing scheme

Routing of the connections between two compartments has to be done with both compartment ends. The membrane and fire routing of the HICANN neurons has been removed completely. To reduce the number of necessary interconnections, the firing is not propagated on a dedicated line any more. In contrast, it is directly propagated on the membrane either by active channels, or by an additional pull-up mechanism.

Beyond nearest neighbour routing

In addition to the nearest neighbour routing capabilities of the HICANN neurons, connections between every fourth neuron in a chip half have been implemented(Figure 6.3 b)). Furthermore, some long range connections have been implemented.

Different compartment sizes

The size of individual compartments can differ orders of magnitude(Compare compartment sizes in [100]). Especially if one compartment is a soma compartment and one is dendritic. To account for theses size differences, the switches of all conductance parameters have been realized with individual parameter memories in contrast to the global switches in the HICANN. In addition, the membrane capacitor can locally be switched to different sizes.

Active channels

Active channels take a major part in multi-compartmental effects. However, due to computation complexity they are usually ignored or clustered in a single compartment([5]). Indeed, synaptic input is available in all neurons of the HICANN and hence in all compartments of the new implementation. Moreover, the positive feedback of the exponential term is available in all compartments to model voltage-gated channels. Nevertheless, the action potentials created in the HICANN are usually sharp action potentials and do not model wide dendritic action potentials. To allow for different shaped action potentials, the slope of the reset can be set individually for each single neuron. Furthermore, each compartment can chose between four different reset potentials. In addition, the bias of the exponential term's amplifier can be used to adjust the influence of the exponential term accounting for the different number of active channels in individual compartments.

Spike shape

In total, the new functions and increased parameterizability increased the number of digital control bits per neuron from four to 41. The number of individual current biasing parameters increased from 12 to 16.

6.4 Inter Compartment Resistance

The inter compartment resistance has been the most demanding single circuit of the multi-compartment implementation. Sections 3.3, 3.6 and 3.7 showed different implementations of resistors in CMOS. However, only an OTA (3.3) or the resistive element element from 3.6 can be used as bidirectional devices.

The absence of a proper linear characteristic for larger differential input voltages in both

devices led to the search for a new kind of device. Classical differential pairs could be excluded. However, in [111] a device called transconductor is introduced by the authors. Finally, I will present a resistive element which uses resistive biased transistors to implement the conductance.

Nevertheless, although the transconductor is not used as inter compartment resistance in the end, I do a small excursus on transconductors now as results might be useful in other circuit parts and the results have been promising.

6.4.1 Transconductors

Transconductors and OTAs

The transconductor is defined as a device linking the output current directly to the differential input voltage by a controllable conductance. In contrast to the OTA, linearity and controllability are directly included in the definition. Consequently the device matches the constraints needed for an inter compartment resistance.

Inconsistent definitions

Different transconductors using CMOS devices are shown in [112] for instance. However, the definition in literature is not completely clear and some authors call transconductors linear OTAs[113]. In addition, transconductor from literature[112] show deviations from linearity if only CMOS devices are used. Hence the difference between OTAs and transconductors nearly vanish. Nevertheless, CMOS OTAs are usually based on differential pairs, while the CMOS transconductors from [112] do not.

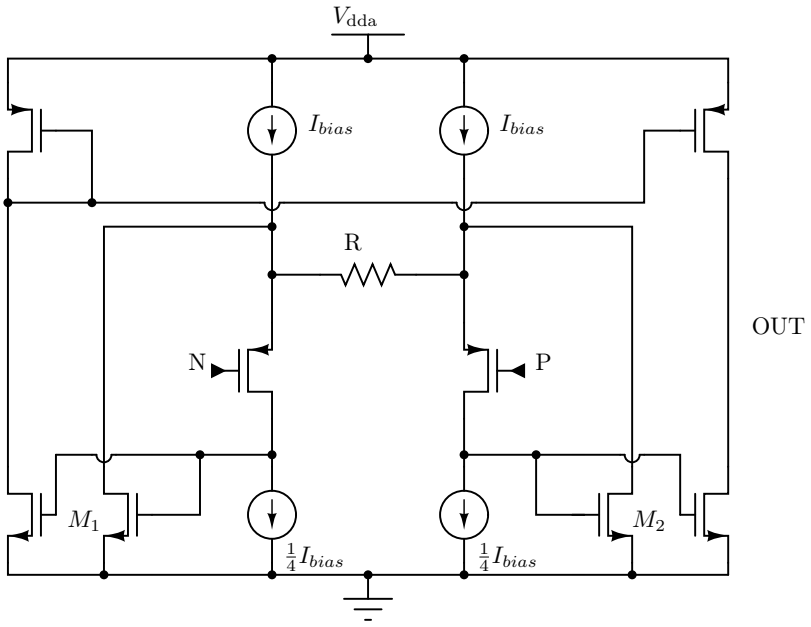


Figure 6.4: Simplified schematic of transconductor studied for multi-compartment emulation. The circuit is basically a simplified of version of the transconductor shown in [113] except for the implementation of R. Indeed, the resistance R is bias dependent and implemented by MOS transistors.

After studying several different implementations, a transconductor based on a simplified version of the transconductor presented in [113] has been designed. A simplified schematic

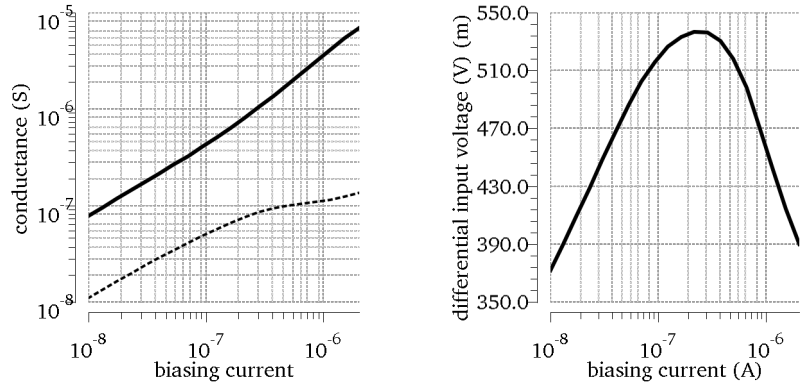
can be found in Figure 6.4. However, the resistor R is implemented by transistors and will be discussed in the next subsection.

Basic operation

When both terminals are at the same level, no current flows through the resistor. 3/4 of the biasing current is flowing through transistors M_1 and M_2. If a differential voltage u_d is applied, simplified the current u_d/R flows through the resistor. The current difference has to be balanced by transistors M_1 and M_2. Consequently, the difference is mirrored to the output.

The resistance

The conductance of the circuit is set by controlling the resistance R with the biasing current. Indeed, the ideal current sources in the schematic are implemented by current mirrors. To allow an adjustable conductance, with real resistors for R, parallel resistors could be made digitally switchable. In addition, the translation of the current mirrors could be made digitally switchable. This solution is tracked in [113].



(a) Average conductance(solid) and standard deviation(dashed) for positive terminal voltages between 750 mV and 1.05 V

(b) Difference voltage where 0.9 of the maximum output current is reached.

Figure 6.5: Results from typical parameterized simulation of transconductor. Negative terminal and output are fixed at 900 mV. The positive terminal voltage is swept from 0 to 1.8 V furthermore, the biasing current is swept. Compare to Figures 3.8 and 3.9

Better results than neuron OTA

Simulation results of the complete designed transconductor can be found in Figure 6.5. The performance is much better than the performance of the OTA used in the single compartment neuron design. However, the complete circuit is expected to be larger.

Resistive element used for multi-compartment circuits

The transconductor circuit has been abandoned due to a miss-match between the realistic conductances implementable and the necessary conductances for the inter compartment conductance. Indeed, they have to be orders of magnitudes larger than the transversal conductances of the compartments. Nevertheless, the resistive element and its biasing circuitry designed for the transconductor have been used. They are presented next.

6.4.2 Resistive Element

The transconductors only shifted the problem if a continuous adjustability is needed, as they need resistors as reference for perfect linear operation. However, a resistive element has been designed for this purpose. This resistive element is used as inter compartment conductance.

Figure 6.6 is a simplified schematic of the resistive element. The actual resistance is implemented by the transistors M_{r1} and M_{r2}. All other transistors are needed for biasing.

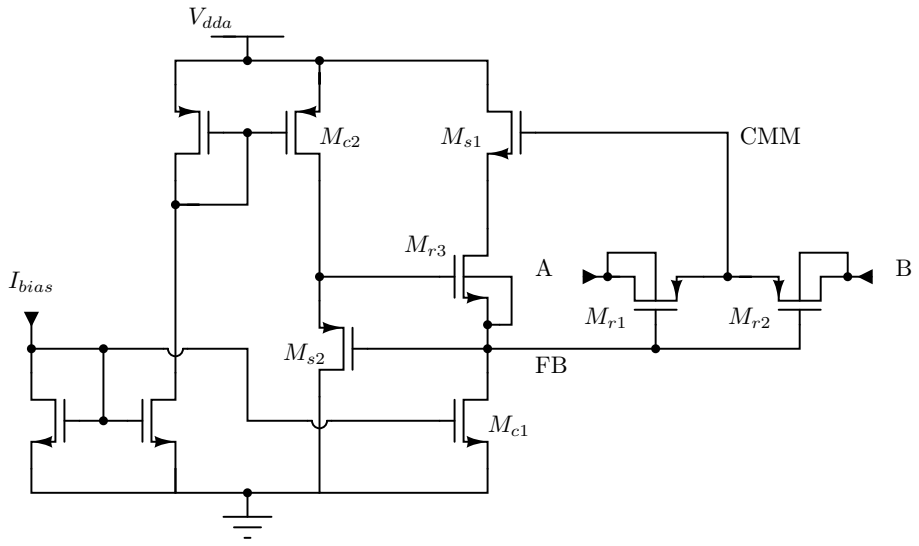


Figure 6.6: Resistive element used as inter compartment resistance. The resistance is between terminals A and B.

Bidirectional resistance

Indeed, a single transistor biased resistive would be sufficient. However, the problem is to set the proper bias (Node FB). A real resistor is bidirectional. Consequently, it is not defined which of the terminals A and B is at higher voltage and source and drain of the transistors M_{r1} and M_{r2} can be permuted. Nevertheless, a node is needed as reference for biasing. Accordingly, two transistors are used and the reference is taken at Node CMM[3] in between[4]. Alternative, two MOSFET in parallel could have been used. However, this approach would have doubled the biasing effort.

Basic operation

Without loss of generality, I assume a higher voltage at terminal B now. In addition I assume a constant voltage above the PMOS threshold voltage between nodes FB and CMM. In this case, gate source voltage of M_{r1} is larger than the voltage of M_{r2} which result in a larger conductance for M_{r1}. Hence, the potential at node CMM is closer to the potential at the terminal A. The resistive characteristic is dominated by transistor M_{r2}.

Biasing

The straight forward solution to set a constant voltage between nodes FB and CMM is to use a source follower. Firstly ignoring M_{r3} this source follower is build by transistors M_{s1} and M_{c1} which is the corresponding current source. The four transistors used so far would be basically sufficient to implement a resistor. However, adjustability is the next issue.

Linear vs quadratic characteristic

To control the value of the resistor, the differential voltage between FB and CMM need to be modifiable. A biasing current needs to be used here to build a robust circuit. Using only a source follower, the voltage difference would be proportional to the square root of the biasing current assuming the source follower is biased in strong inversion. In contrast, the resistance of a resistive bias MOS transistor linearly depends on its gate-source voltage. Accordingly, a very large parameter range would be needed for the biasing current to achieve a sufficient range for resistance values. Indeed, this parameter range would not be implementable with out side-effects like a worse miss-match and more area consumption.

[3] CMM stands for common mode. I actually want to measure the common mode of terminals A and B.

[4] Another approach would be to use two transistors in parallel. However, this would double the biasing circuitry.

Another resistance

The transistor M_{r3} is added in the biasing branch to achieve a more linear current voltage characteristic between the voltage difference and the biasing current. It is bias resistive by the source follower build of M_{s2} and M_{c2}. Consequently the voltage drop on M_{r3} is roughly linear to the biasing current.

Common mode sensitivity

The bulk potentials of transistor M_{r1}, M_{r2} and M_{r3} are not set to V_{dda} respectively ground. The body effect changes the effective threshold voltage of a MOS transistor. Hence, a fixed bulk potential would directly change the resistance value which would result in less linearity. Consequently, the bulk potentials must be moved with the common mode of terminal A and B. However, the trade-off results in a larger layout due to harder spacing constraints.

Additional features

The implemented capabilities for bias current scaling and a switch-off are not shown here. The latter is implemented by shutting down the biasing current. Additionally, the potential at note FB is pulled to V_{dda}.

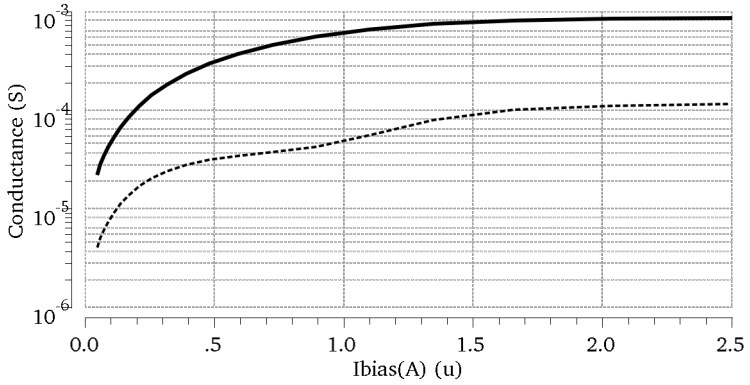


Figure 6.7: DC-simulation with swept voltage at terminal A and terminal B at 900 mV. Average conductance(solid) and standard deviation(dashed) for positive terminal voltages between 700 mV and 1.10 V in relation to the biasing current

Conductance range

Conductance values in dependency to the biasing current can be found in Figure 6.7. The standard deviation is given as a measure for linearity again here. With biasing currents between 50 nA and 2.5 µA, resistance values between 43 kΩ and 1.2 kΩ are reachable. However, the conductance starts saturating above 1 µA. Below, the relation between conductance and biasing current is nearly linear. Larger resistance values could be achieved by using different divisors for bias current scaling. Smaller resistances are senseless due to the resistance of switching pass-transistors used for routing.

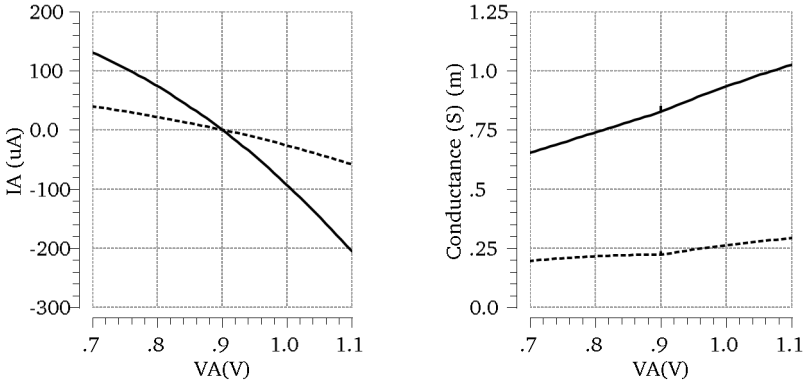
Linearity

Crosscurrent and conductance for two different biasing currents are shown in Figure 6.8. Perfect linearity would result in two flat parallel lines in Figure 6.8 b). The slope is created by the moving common mode when one terminal voltage is swept while the other is kept constant. However, this situation is more realistic than a sweep of the differential voltage which fixed common mode.

6.4.3 Conclusion

The presented resistive element can implement the inter compartment resistance. Linearity is better than the linearity of the neuron OTA or the resistive element of the synaptic input. However, there is still a common mode dependency.

The presented transconductor cannot be used to interconnect compartments. However, a decoupling of input and output impedance is not needed here. It could be a candidate for



(a) Current

(b) Conductance. The peaks are simulation artefacts.

Figure 6.8: DC-simulation withs terminal A voltage swept between 700 mV and 1.1 V. The voltage at terminal B is fixed at 900 mV. Crosscurrent and conductance for different biasing currents: 2 µA(solid) and 400 nA(dotted)

future implementations of more precise transversal conductances.

A simplified version of the resistive element could replace the restive element of the synaptic input in the AdEx neuron implementation (See Section 3.6). However, this would be a new circuit requiring intense simulations for verification.

Next, the resistive element will be put in context with the complete inter compartment connection module.

6.5 Firing Modes - the Interface Module

Two tails

The compartment interface module is the part of a compartment which interfaces other compartments via the routing network which will be presented in the next section. In soma direction, the membrane is directly connected to the routing network. However, in addition to the resistive element, spike propagation circuitry is added in the branch connecting in dendrite direction.

6.5.1 Spikes

Spike propagation on the membrane

The digital spike propagation mechanism between connected neurons used in the HICANN chip has been removed in the multi-compartment implementation to shrink the number of necessary routing and switching circuitry by a factor two. Spikes are propagated between compartments in a more biological realistic fashion now directly on the membrane. Triggering of a digital spike is done by the spike detection comparator in each individual compartment. Two different modes of spike propagation are implemented.

Active mode

In active mode, the compartments are constantly connected through the resistive element or a pass-transistor. When a spike is initiated at a compartment, it is propagated by pulling the next compartments membrane to higher values. The exponential threshold in this compartment will be activated once its exponential threshold is reached. This mechanism implements dendritic spikes and back-propagating action potentials.

Increased current during action potentials

To allow larger and broader action potentials the reseting of the membrane is delayed by the length of a digital spike[5]. This increases the current propagated to other compartments during a spike Nevertheless, the spike detection threshold in dendrites must be carefully chosen if the creation of a digital spike should be warranted.

Consequences for STDP

An annotation needs to be made that this way of spike propagation can have interesting effects on STDP which is the only mechanism using the digital spikes if the dendritic compartments are not reset. The digital spike signal is only created for triggering STDP if the threshold is reached in the dendrite or if a spike has been back-propagated into the dendrite. This mechanism is realistic as as information needs to be propagated to the synapse in a real neuron to trigger change of efficacy. In addition, the creation of a dendritic spike is more probable if the dendritic membrane voltage is already higher. Hence excited dendrites are more probable to trigger STDP. STDP would be membrane voltage dependent in a realistic way. Membrane voltage dependency of STDP is assumed in modern models[114].

Passive mode

The second mode of spike propagation is called passive mode. Now the propagation of spikes to the dendritic compartments is guaranteed. However, the exponential term is not allowed in compartments except for the soma compartment.

When a spike is detected in a compartment, the compartment capacitance is cut from the next compartment in dendrite direction. At the same time, this dendritic compartment is pulled to the power supply. This way, spike detection can be assured. Nevertheless, all dendritic compartments are reset this way and all information which might have been coded in the dendritic membrane voltage is lost.

6.5.2 Implementation

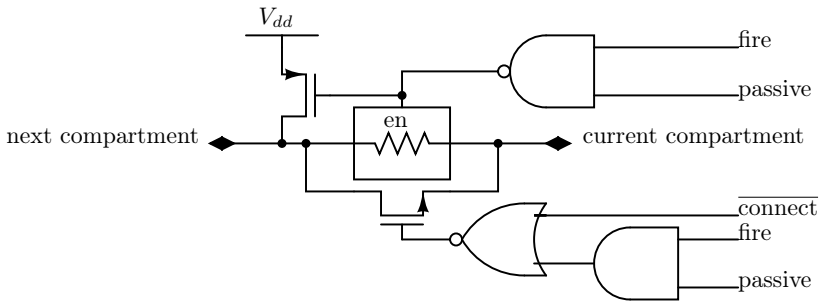


Figure 6.9: Simplified schematic of compartment interface

Figure 6.9 displays a schematic of the interface; memory cells are excluded. The connect signal and the passive signals are connected to local memories. The pass-transistor is able to connect the compartments with a small resistance of 120 Ω. This way, larger soma compartments can be constructed. The pass-transistor is activated by setting the connect signal.

In passive mode, the resistive element and pass-transistor are deactivated during a fire signal. In addition, the next compartment's membrane is pulled up while fire is active. The cut off of the two compartments is necessary to prevent current flux onto the current compartment's membrane.

[5] Adjustable in the multi compartment implementation

To pull up the next dendritic compartment, a compartment needs to have access to the membrane of this compartment. Accordingly, the connecting elements can only be located in the compartments tail leading to the dendrite. This is a weak point of the concept and has consequences for branching in dendrite routing which will be discussed next.

6.6 Dendrite Routing

(a) Symmetrical compartment with longitudinal in each tail(not implemented)

(b) Compartment with longitudinal resistance in one tail only (implemented)

Figure 6.10: Symbols used for different types of compartments. The round corner indicates a longitudinal resistance.

In contrast to the neurons on the HICANN which allow interconnection of neighbouring neurons, complex tree structures are possible with the multi-compartment implementation. The elementary element is a single compartment with its two tails. Figure 6.10 shows the symbols used for compartments.

If passive mode(6.5) should be usable, the resistive element has to be always in the tail pointing away from the soma. However, passive mode is not used, no constraint is given for the resistive element location.

6.6.1 Building Neurons

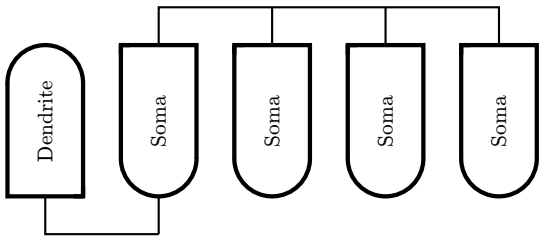


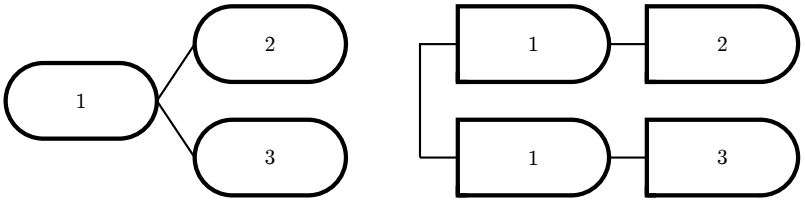
Figure 6.11: Implementing a large soma compartment with a single dendritic compartment.

Large single compartments

The size of a single compartment in a model can be too large for an implementation with a single compartment in the emulation. Hence, several compartments can be interconnected directly to form a single larger compartment. Figure 6.11 displays a two compartment model with a large soma. Implementation is done using four directly interconnected hardware compartment as soma compartment and a single compartment for the dendrite.

Branching

The location of the resistive element in the dendritic tail of a compartment complicates branching. Figure 6.12 shows a three compartment branch structure which has to be implemented using four compartments. However, there are still differences as the model branch is equivalent to three resistances connecting to a single point. A more precise implementation



(a) Symmetrical compartments **(b)** Implemented compartments

Figure 6.12: Implementation of branches.

would be to use a 5 compartment hardware structure with additional compartments between the compartments 1 and 2 respectively 1 and 3. Nevertheless, this would add additional parasitic capacitance and more compartments would be necessary.

If the neuron is constricted to active mode, the hardware compartments can be mirrored and 3 compartments would be sufficient to implement the branching structure similar to Figure 6.12 b).

6.6.2 Pass Transistors

Pass transistors realize connections

Connections between compartments are routed through pass-transistors. The use of pass-transistors is possible as it is not necessary to route analog voltages close to the power supply. In addition the process offers so called low-vt transistors with a reduced threshold voltage.

Better performance than transmission gates

NMOS pass-transistors achieve a larger conductance value for less area than a transmission gate. The implemented pass-transistors achieve resistance values smaller than 120 Ω in the relevant operating region. Around 11 μ^2m chip area are needed. This compares to 1 kΩ used for the transmission gates in the HICANN which have about half the size in layout. However, when routing dendrites, at least two pass-transistors have to be interconnected in a row.

6.6.3 Routing matrix

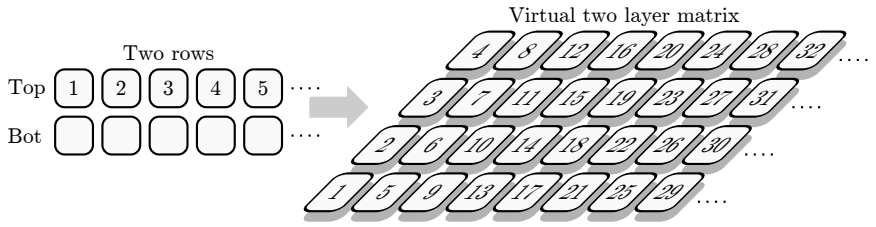


Figure 6.13

Two rows to double layered matrix

The neurons in the HICANN are arranged in two rows. Based on this constrain a routing matrix has been developed. This matrix keeps the connections between neighbouring neurons. In addition, connections between every fourth neuron of a row are possible. Hence, the two rows of neurons are transformed into two matrix layers(See Figure 6.13).

Not optimized

However, this is not an optimized structure, but a structure which allows intuitive routing. To find a better routing scheme, a routing and mapping algorithm needs to be developed,

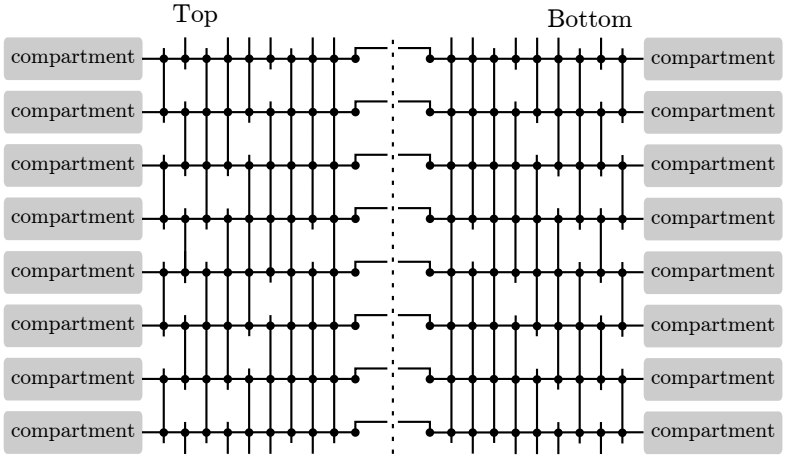


Figure 6.14: Dendrite routing matrix. Each horizontal line connecting to a compartment corresponds to both compartment tails. Vertical lines are single lines. Each dot consists of two pass-transistors and the necessary memory cells. Long endings of vertical lines should denote their continuation in the next, not shown compartment. The dashed line is divides between top and bot and should visualize that a longer distance has to be crossed.

to map biological multi-compartment models automatically onto a dendrite routing matrix. Mapping needs to be done for several different neurons for different routing matrices and the best evolving structure needs to be chosen. Nevertheless, although this approach would result in a better routing scheme, it cannot be said the result would be optimal. The routing matrix presented here is sufficient to prove the concept of the multi-compartment emulation and the necessary circuits. My benchmark was to implement the 9 compartment model from [100]. The connectivity structure can be found in Figure 6.15.

The number four has been chosen as a trade-off between necessary routing lines in layout and routing capabilities. The next reasonable number would be to interconnect every eights neuron.

Long range connections

The final routing matrix not only includes connections between every second and every fourth neuron. Sparsely, every eighth neuron can be interconnected in addition. Moreover, groups of 16 neurons can be interconnected. Furthermore, compartments which are crossed by the interconnection lines can be connected to them. This allows to build larger compartments and branches using fewer switches. However, it enlarges the line capacitance and coupling. The complete dendrite routing scheme can be found in Figure 6.14

6.7 Reset mechanism

Spike shape

The shape of an action-potential can be different in the dendrite. Hence, dendritic action potentials might require different reset potential. Furthermore, the strength of the reset can be different in the dendrite, in virtue of less potassium channels.

Delayed reset

The reset mechanism presented in 3.9 has been modified and enhanced for the multi-compartment implementation. First of all, the reset is delayed by the length of a fire pulse now. This allows larger membrane voltages during a spike. Furthermore, it assures that all compartments connected to a larger soma compartment can detect a spike.

4 blocked reset voltages

In the HICANN, there are two global reset lines, each driven by a buffer[76] which is capable to supply a DC-current of 2 mA. Odd neurons connect to one line and even neurons connect to the other. However, this is not very flexible. Furthermore, the HICANN includes one spare buffer for global voltages per floating-gate block which is not used. Consequently, each compartment of a chip half can be chosen to reset to one of four different reset potentials now. In addition, the reset voltages are locally blocked by 200 fF PMOS transistor capacitances in each compartment.

An adjustable strength of the reset has already been implemented in the HICANN. However, this was a global parameter. Consequently, all odd or even neurons would get a similar current.

Individually adjustable reset current

In the compartment implementation, one of the four reset voltages can be chosen to be implemented with an individually adjustable current. This allows different reset slopes in somatic and dendritic compartments for instance. Only one of the four reset voltages is implemented with an adjustable current as individual floating-gate parameters are an expensive resource on the ASIC. Its a trade-off. All four voltages can be used as reset through low ohmic pass-transistors.

6.8 Parameterization

6.8.1 Range extraction

9 compartment reference model

To get a starting point for the necessary parameter ranges of each compartment the work from [100] has been used. The authors supply the multi-compartment models they used as scripts for the neural simulator neuron which enables extraction of single parameters. The chosen reference neuron model is an nine compartment model of a disassociated[6] thalamocortical relay neuron. Figure 6.15 gives a schematic of the compartmental model including transversal resistances and compartment capacitances.

Membrane capacitor

Extraction from [100] results in membrane capacitance of 23.5 pF for the soma and 0.2 pF for the smallest dendritic compartment. In contrast to the single compartment parameterization, the membrane capacitor has to be used directly to account for compartment sizes. However, what matters is the relationship between the capacitors. The membrane time constant of the model is 23 ms for all compartments. It lies perfectly within the margins for single compartment emulation.

Inter compartment resistance

The axial resistance of the compartments has been determined between 240 kΩ in the soma and 102 MΩ in an apical dendrite. Nevertheless, what matters is the resistance between the single compartments. Assuming, the compartment capacitance would be in the middle of the compartments, the value can be calculated as the sum of half the axial resistances of two adjacent compartments. However, branches in the dendrite have to be treated differently. As described in Section 6.6, branches have to be emulated by additional compartments with minimum capacitance. Consequently, only half of the axial resistance has to be taken into account. The inter compartment resistance has to be between 2.35 MΩ and 51 MΩ. Consequently, the value spread is much smaller than the spread of the axial resistances. Nevertheless, it must be possible to vary the resistance by a factor 20.

Treating branches

The currently implemented resistance range in the multi-compartment implementation is 1.3 kΩ up to 40 kΩ. Consequently the factor 20 is meet.

[6] The axon has been cut from the soma

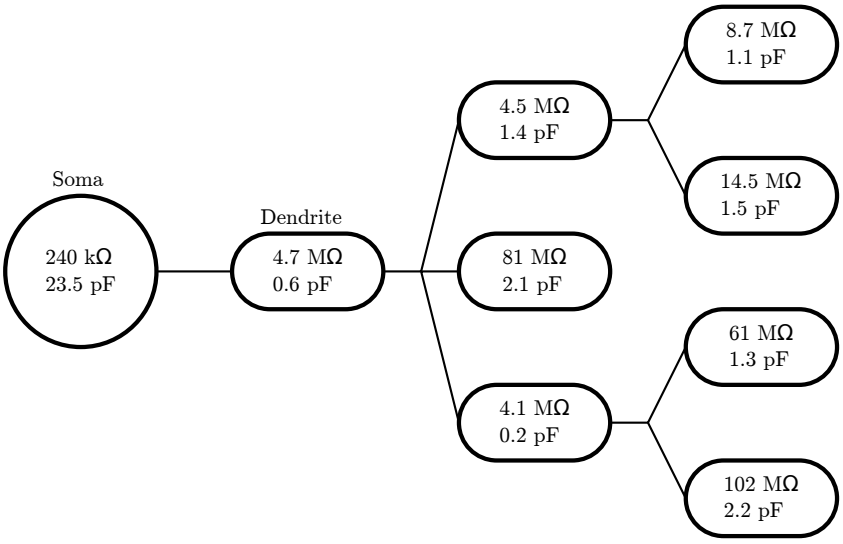


Figure 6.15: Reference compartment model extracted from [100]. The given parameters are the compartmental axial resistance and the compartment capacitance.

6.8.2 Parameter translation

The time scaling and voltage scaling can be done similar to 3.12. However, to account for the compartment sizes, the membrane capacitance has to be determined first.

Capacitance scaling comes first

The maximum available capacitance of a hardware compartment is 3.5 pF parasitics excluded. However, several compartments can be interconnected directly to build larger compartments. The minimum capacitance of a compartment is the parasitic capacitance which is supposed to be around 160 fF.

The membrane capacitances from the nine compartment model described above differ by a factor 100. Consequently, the choice would be to choose 5 compartments as soma compartment.

Inter compartment resistance

The next step is to determine the inter compartment resistance. Here, it is essential to take into account the compartment capacitance and calculate a longitudinal time constant which needs to be kept constant in one time scale. Hence, additional to the time scaling factor the resistance directly scales anti proportional with the membrane capacitor.

In the model, the inter compartment resistance between the soma and the first dendritic compartment is 2.47 MΩ. After scaling with the membrane capacitor, this results in a resistance 3.6 MΩ at real time with a hardware soma capacitance of 16 pF. At time scaling factor 10^4, this would be 360 Ω which is below the resistance of the pass-transistors used for routing[7]. Consequently, the has to be operated in 10^3 mode or at most $3.6 * 10^3$. However, here, to better match the conductance ranges in the compartments, I assume $3.6 \cdot 10^3$ mode and hence a resistance of 1 kΩ The use of a smaller soma capacitance would allow larger inter compartment resistance and hence a higher time scaling factor.

All other resistances must be scaled the same way. This results in a maximum inter

[7]Remind the 1 kΩ resistance of the transmission gates used for inter connecting membranes in the single compartment neuron (See 3.10). Due to the smaller membrane capacitance the 360 Ω would correspond to 2.9 kΩ. Hence, 1 kΩ is still good. However, in 10^5 the transmission gate resistance has biological relevance.

compartment resistance of 20.6 kΩ. Consequently, the model is implementable with the multi-compartment circuit.

6.8.3 Realisation

Conductance parameter

The main difference of the parameterization of the compartments in comparison to the single compartment neurons is the relative size difference of the single compartments. The conductance parameter range of the single compartment is already huge in the single compartment solution if the scaling switches are included. Accordingly, the conductance ranges are kept. Nevertheless, to allow individual scaling of the biasing currents, local memory bits are added at each switchable current mirror to allow individual switching.

Capacitor switching

To allow variable membrane capacitances, the MIM membrane capacitor has been divided int to parts which can be switched. The smallest unit is 125 fF. There are for switchable capacitors of one, two, four and eight units. This way a maximum metal capacitor membrane capacitance of 1.875 pF can be switched with a resolution of 4 bits.

However, 1.875 pF is only a factor 15 larger than 125 fF and a factor 100 has been constrained by design. Consequently, two switchable gate capacitors of 800 fF have been added in each compartment. These capacitors will not match as good as the MIM caps and should only be used if large capacitances are needed. If larger capacitors are needed, the several compartments have to be interconnected with the path transistors to construct a single large compartment.

6.9 Additional Changes

In addition to the changes interfering the multi-compartment implementation, several improvements have been done to the neuron implementation from chapter 3.

Individual addressing

In the HICANN, neuron pairs are sharing 8 memory bits and hence a memory address. Due to the drastic enlargement of the number of memory bits - 41 are used for a compartment - the address sharing has been removed to allow fewer word lines. Mapping of necessary 4 bits for input and output configuration on three available bits in the HICANN reduces the possible input output configurations(See [61] for details). This constraint has been removed by having a single enable memory bit for each current input and membrane voltage read out.

Modifiable firing pules length

The length of the digital firing pulse created in the HICANN neuron (see 3.8) is fixed. However this pulse is used for STDP and an adjustable length would enhance the ability to calibrate the chip. In addition, the pulses of the pre-synaptic neuron vary with the clock cycle and according to short time plasticity in the HICANN chip. Furthermore, the resulting height of a spike on the membrane depends on the firing pulse length due to the new reset implementation(See 6.7). Consequently, the firing pulse length of a compartment is adjustable now.

6.10 The Multi-Compartment Chip

To verify the multi-compartment concept in silicon, a small test ASIC has been designed. The ASIC is called Multi-Compartment Chip (MCC). It has been developed a collaboration with Andreas Hartel, who did the back-end and the main part of the digital code.

A scaled HICANN

The MCC is basically a small version of the HICANN with the multi-compartment features described above. This design approach has been chosen to be able to transfer all circuits into the HICANN after verification. Furthermore, it is possible to build small networks on the MCC this way. Two rows of 32 compartments have been implemented. Each compartment

can receive synaptic input from 16 configurable synapses. In contrast to the HICANN, spike event routing on L1 has been fixated to reduce complexity. Only a single layer one bus is used.

Input/Output The high-speed serial L2 IO of the HICANN has been replaced by an easier to handle parallel interface, as bond pads are no critical resource on this chip. Digital spike event IO is done via this bus. All control commands are send via JTAG.

Clocking The HICANN uses a PLL[8] to generate the digital 250 MHz clock out of a slow 50 MHz clock. However, in the MCC, chip area is limited and a PLL has not been possible. Consequently, the chip is directly fed by a 200 MHz external clock. The differential clock input of the HICANN which used the general purpose operational amplifier from [76] has been replaced due to limited bandwidth. A specialized symmetrical OTA with biasing similar to the circuit presented in 9.6.3 has been designed for this purpose.

ESD[9]-protection has been improved by using ESD-clamp circuits designed by Marc-Olivier Schwartz.

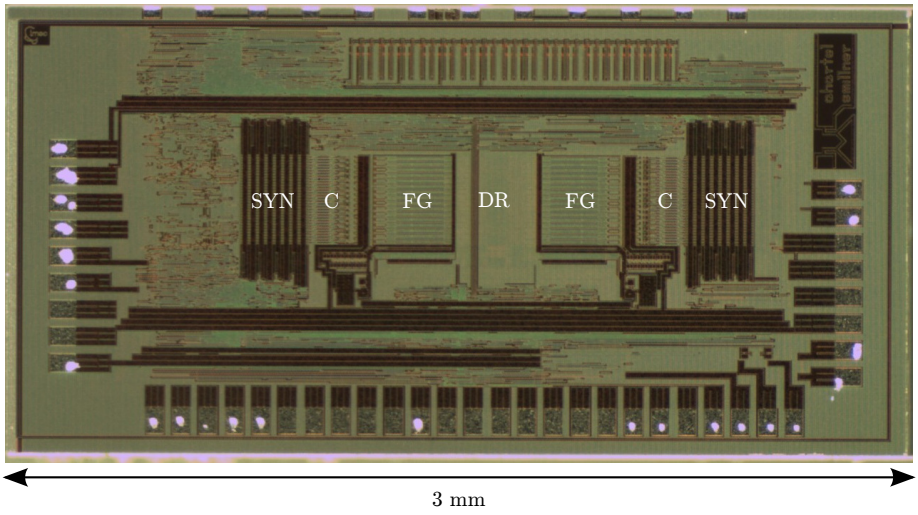


Figure 6.16: Photograph of the Multi-Compartment Chip with marked synapses(SYN), compartments (C), parameter memory (FG) and dendrite routing (DR). The spike routing is bisect in the middle by digital spike event creation and distribution circuitry.

A photo graph of the MCC can be found in Figure 6.16. The analog circuitry in the middle is surrounded by standard cells. In contrast to the HICANN chip which is dominated by synapses, the MCC is dominated the compartments, the parameter memory and dendrite routing.

Used die area In comparison, the length of a neuron has been increased from 240 µm to 300 µm. Furthermore, 80 µm have to be added for dendrite routing and the inter compartment interface. Accordingly, the new compartment structure would not fit into the area of the HICANN neurons. However, the size of the floating-gate array could be shrunk. Its length is 425 µm in the HICANN which could be reduced to 300 µm although the total cells used per compartment has been increased by 1/6. Consequently, the compartment structure would nearly fit

[8] Phase-Locked Loop
[9] Electro-static discharge

into the HICANN. The missing micro meters could be retained by either removing spare[10] floating-gates or reducing the routing matrix.

[10] I added two additional cells per compartment as reserve for replacing the synaptic input's voltage biasing by current biases

7 Multi-Compartment Experiments

This chapter presents two simulated experiments showing effects created by the compartmentalization of the model. The first experiment shows the reaction of a four compartment model with passive dendrites to synaptic stimulus. The second observes action potentials with active and passive dendrites.

The MCC is not ready for neuron measurements, hence simulation results are presented here.

7.1 Four Compartment Reference Simulation

The experiment presented here reproduces results from a four compartment neuron model simulation from [101]. This publication uses the neuron model commonly used in network experiments by the BrainScaleS modeling group around Anders Lansner. The shown results from circuit simulations have been published at the ESANN[1] 2012 in [115].

7.1.1 Methods

The model presented in [101], called Lansner Model in the following is a four compartment conceptual model with no detailed branched dendritic tree. A schematic of the model can be found in Figure 7.1. It consists of three compartments which are passive except for synaptic input. These compartments modeling the dendrite. A single active compartment as soma which uses Hodgkin-and-Huxley type channels to implement action potentials. An axon is only included using spike propagation delay mechanisms. Each dendritic compartment has

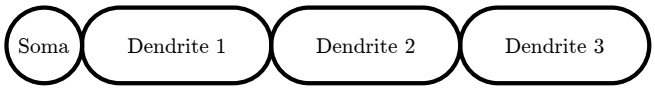


Figure 7.1: Four compartment arrangement used in [101]

a leakage conductance g_l which drains current to a leakage potential E_l. Compartments are connected in a line with a conductance g_core in between. Do to its simplicity the model is a perfect candidate for a first benchmark of the implemented circuit.

The model implements AMPA and NMDA based synapses for neuron interconnections. However, here I only compare to the AMPA based synapses. Theses are implemented by a conductance which is active for a certain time after a post-synaptic action potential. It connects to a reversal potential at biologically 0 mV for excitatory synapses and −85 mV for inhibitory synapses.

In the circuit simulation done here, the passive properties of the model have been transfered to the circuit dividing the capacitance by advanced 20 and multiplying the conductances by advanced 500(Table 7.1). Therefore, the time scaling factor is roughly 50 000. The different conductances of dendrite and soma are reached by using the scaling switches only and keeping

[1] European Symposium on Artificial Neural Networks

Parameter	Model	Hardware
E_l	-70 mV	880 mV
g_l Soma	0.003 µS	1.7 µS
C_m Soma	0.03 nF	250 fF
g_l Dendrites	0.01 µS	5 µS
C_m Dendrites	0.3 nF	2.675 pF
g_core	0.04 µS	20 µS

Table 7.1: Passive parameters of the model presented in [101] and parameters used in circuit emulation.

the bias current constant. The granularity of the hardware capacitance inhibits a perfect matching between model and simulation parameters.

No dedicated calibration scheme has been applied - the bias values are chosen directly from the characteristic of the circuit. The voltage level of the simulation has not been scaled at all. Only the leakage potential has been used as a reference. Accordingly, no voltage scaling is necessary.

The excitatory synapses used in the model are conductance based. However, the size of the PSPs in the model is below 6 mV. Hence, the membrane is far away from the excitatory reversal potential of 0 V and the synapses behave similar to current based synapses.

Synaptic input is emulated by a short current pulse in the circuit simulation. The size of the current pulse has been chosen, to reproduce the PSP size of the model.

In model and circuit simulation, the neuron is stimulated consecutively in each single dendritic compartment. The resulting PSP at the soma is observed.

7.1.2 Results

Figure 7.2 shows the results from [101] overlaid by the circuit simulation results. Although the mapping of the parameters has not been perfect, model and circuit simulation traces are similar. Even better results would be expected with a calibration.

However, only passive properties have been taken into account. Hence this is basically a matching between two R-C-chains. Matching with active channels would be much greater challenge.

Nevertheless, the effect of the compartment structure is apparent. Reducing this compartment simulation to a point neuron would normalize the size of all PSPs in Figure 7.2 and hence change the synaptic efficacy.

7.2 Action Potentials with Active and Passive Dendrites

The simulations shown here illustrated the effect of passive and active dendrites on action potential creation and propagation. Although the compartment arrangement is a simple four compartment chain, the effects are drastic even for passive dendrites. All results shown here are qualitative.

7.2.1 Methods

The compartment arrangement used here is identical to the structure shown in Figure 7.1 which is also used in [101]. However, the parameterization of the compartments is completely

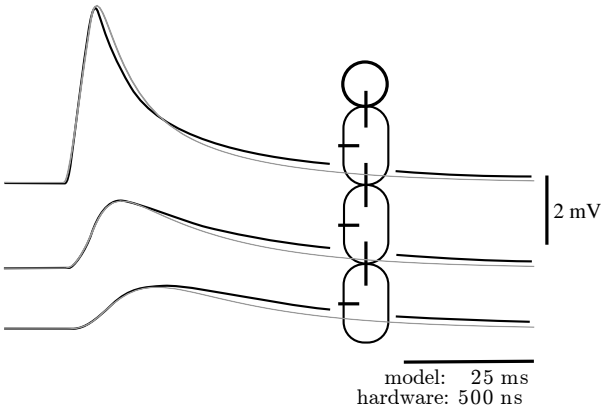


Figure 7.2: Post synaptic potentials. Simulation results extracted from Figure 4 B) in [101] overlaid by the results form circuit simulations(gray). The neuron is stimulated at different parts of the dendrite (oval). The shown membrane traces are recorded at the soma(circle). The figure has been published at [115]

different to [101]

Except for the positive feedback of the exponential term, all compartments are parameterized equally. No dedicated parameter mapping has been applied, however, the time scaling is supposed to be between 10^5 and 10^4. All fast and slow switches in each compartment have been set to active. The leakage conductance is two times smaller than the adaptation conductance. Accordingly, strong adaptation effects are to be expected.

The membrane capacitor has been set to 1 pF for all compartments. The longitudinal resistance between the compartments is programmed to the minimum possible value which is about 40 kΩ. The large resistance has been chosen to maximize compartmental effects. The inter compartment resistance is much smaller for larger compartment numbers in a more realistic model. All compartments are capable of detecting spikes at a threshold Θ of 1.1 V.

To achieve a passive dendrite with an active soma, the bias current for the operational amplifier of the exponential term is set to zero in the dendritic compartments. Subsequently, when emulating active dendrites, the bias is set to 800 nA which needs to be compared to the 2 µA used in the soma. The smaller current limits the total impact of the exponential term in the dendrite to model a lower density of active channels.

Three different simulations are performed here. I start with passive dendrite and stimulate the neuron at the soma. Next, the neuron is stimulated at the end of the dendrite. At last, the dendrite the exponential term in the dendrite is activated. The neuron is still stimulated at the dendritic end again.

7.2.2 Results

Results using passive dendritic compartments can be found in the Figures 7.3 and 7.5. Both figures show a short burst of action potentials. Although the stimulus continues beyond the figures margins, no further bursts can be observed during on stimulus pulse. Hence, this behavior could be labeled as phasic respectively transient bursting (See 4.4). Figure 7.3 gives details of the action potential.

When stimulating at the soma(Figure 7.3), a stimulus current of 600 nA is sufficient to trigger the exponential term of the soma. The voltage drop between the compartments can

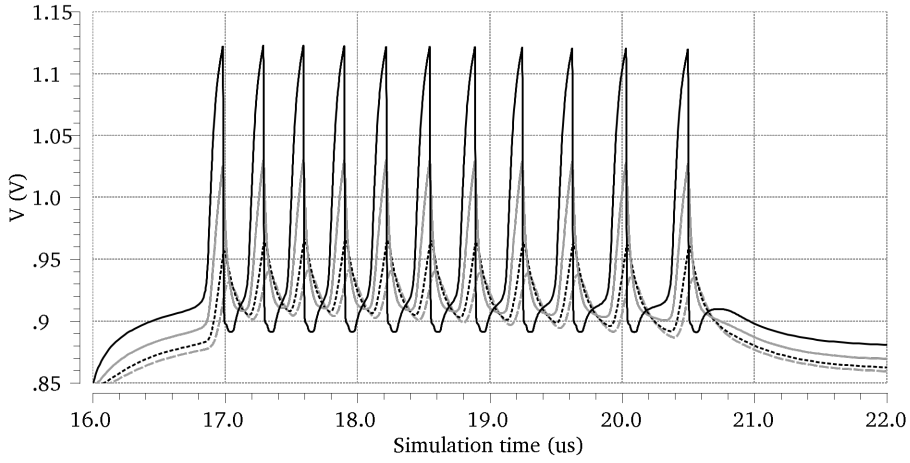
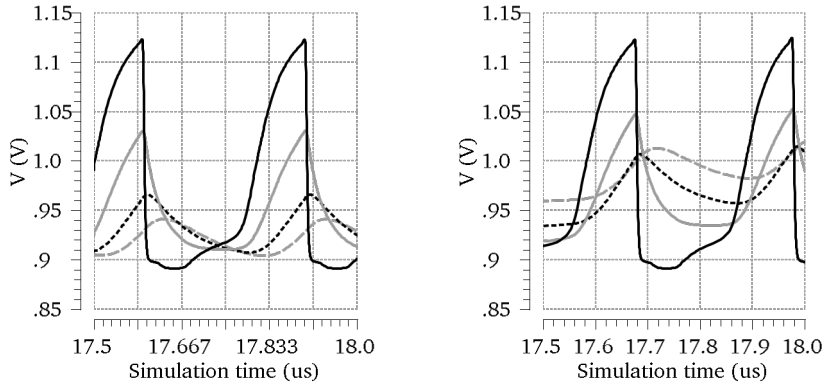


Figure 7.3: Membrane potentials of four compartment model stimulated by a step current of 600 nA (onset at 16 µs) at the soma. Soma:black,1st dendritic compartment:gray, 2nd black dotted, 3rd gray and dashed.

be observed before the onset of the first action potential at the soma. The soma pulls up the dendritic compartments.

A closer look at the action potential (Figure 7.4 a)) shows a flattening above above 1.1 V. In the multi compartment neuron, the membrane is not reset directly after spike detection. Hence voltages above the set spiking threshold (1.1 V) are possible. The flattening of the action potential is caused by reaching the maximum current of the exponential term of the soma.



(a) Stimulus at soma **(b)** Stimulus at dendrite end

Figure 7.4: Detail of membrane potentials of four compartment model with passive dendrites stimulated by a step current. Soma:black,1st dendritic compartment:gray, 2nd black dotted, 3rd gray and dashed.

The dendritic compartments do not reach the spiking threshold. They are pulled down by the soma compartment's reset mechanism before. However, the reset voltage is only reached

at the soma.

After the reset is released, the soma is pulled up by the dendritic compartments which are still above the reset potential. Indeed, this effect enables bursting although the reset voltage is above the exponential threshold(See 4.4). The burst finishes due to adaptation.

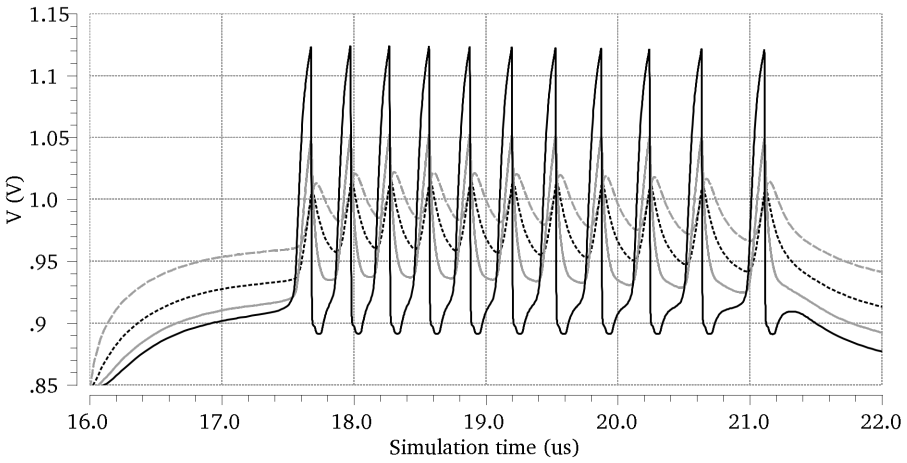


Figure 7.5: Membrane potentials of four compartment model stimulated by a step current of 1 µA (onset at 16 µs) at the last dendritic compartment. Soma:black,1st dendritic compartment:gray, 2nd black dotted, 3rd gray and dashed.

To achieve spikes in the neuron stimulated at the last dendritic compartment(Figure 7.5), a much larger current is necessary for stimulation as the voltage drop between soma and dendrite is reversed now. The capacitance of all compartments needs to be charged until the membrane voltage of the soma reaches a voltage high enough to activate the positive feedback of the exponential term.

The pattern observed can be characterised as delayed transient burst[12]. The burst is finished by adaptation again.

With a closer look at the spike (Figure 7.4 b)) a dominance of current from the exponential term of the soma can be observed once the exponential threshold is crossed by the soma's membrane. The first dendritic compartment reaches higher voltages than the last compartment although the last is stimulated. Only the membrane of the soma crosses the spiking threshold.

More complex behaviour can be achieved by activating the exponential term in the dendritic compartments(Figure 7.6). The feedback at the soma is still the strongest. Consequently, the soma spikes first. Detected action potentials can be identified by the reset of the of a compartment in Figure 7.6 Looking at the first spike,the reset of the soma manages to inhibit the first and second dendritic compartment from spiking. However, the resistance to the third compartment is too large, hence it creates an action potential.

The current flowing from the dendrite into the soma is much larger. Accordingly, the spike frequency in a burst is higher than the frequency observed in the prior simulations. The frequency is given by the refractory period in this case. After a first burst of 5 somatic spikes, the neuron produces three bursts of two spikes.

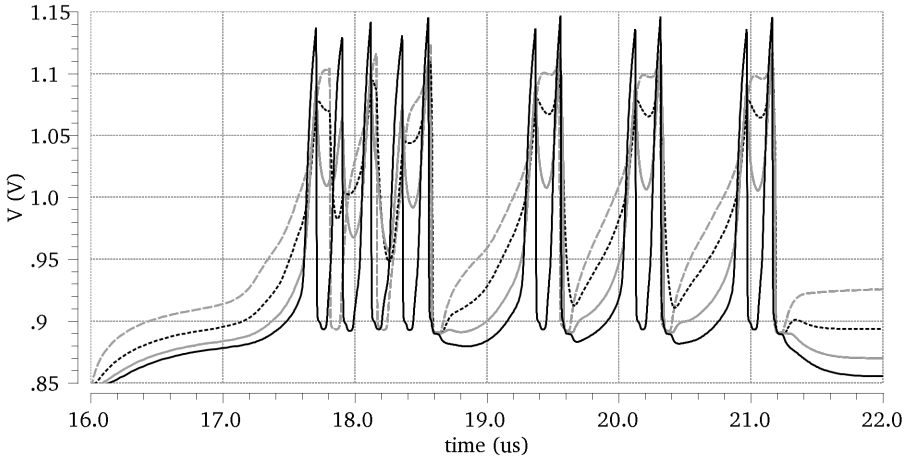


Figure 7.6: Membrane potentials of four compartment model stimulated by a step current of 600 µA (onset at 16 µs) at the last dendritic compartment with active dendrites. Soma:black,1st dendritic compartment:gray, 2nd black dotted, 3rd gray and dashed.

7.2.3 Conclusion

Simulations showed that the multi-compartment emulation is capable of producing more complex behavior than the single compartment AdEx emulation. Delayed bursting has only been possible introducing a negative a in the AdEx in [12] for instance. In addition, bursting seems to be much more stable.

However, only a very small parts of the circuits modeling capabilities have been used in this experiment. More complex structures will allow for many more effects. Individual spike detection thresholds and reset voltages could be used in the dendrite to model dendritic action potentials more precise.

A reduction to a point model seems to be a very drastic step after these simulations. Nevertheless, the cut needs to be taken at some place. Although multi-compartment models are more accurate than point neurons, models using more compartments should be more accurate than models with less compartment.

8 Discussion: Multi-Compartment

Multi-compartment emulation is the next step to enhance the power of analog emulation in comparison to simulations as compartments are build in a natural way. Furthermore, single-compartment modeling seems to be a rough simplification real neurons which can hardly be sufficient to understand and exhaust the computational power of neurons. Consequently, the HICANNs neurons have been enhanced and modified to allow a realistic multi-compartment implementation in this thesis. Here I will recapitulate the modeling decisions and different design-parts. A direct comparison to the HICANN neuron concerning single compartment performance is done. After discussing the simulation results, the circuit is compared to other solutions of a multi-compartment emulations.

8.1 Model Choice

AdEx compartments

The chosen model is a multi-compartment implementation of the Adaptive Exponential Integrate-and-Fire neuron model(AdEx). However, it is optimistic to talk of a choice in this case, as the model has been given by the HICANN neuron. Nevertheless it is important to point out the use of the AdEx as most multi-compartment model neurons are Hodgkin Huxley Model(HMM) based. Indeed, this is no surprise as the AdEx is a phenomenological simplification. To get a more realistic model, the first step would be to use a HMM. Afterwards, a multi-compartment emulation would be used.

Active and passive channels

However, as passive dendritic properties can be extracted from a multi-compartment HMM, the modeling discrepancy is no problem for dendritic morphology. Talking about active channels, there is an issue indeed as the only active voltage-gated channel in the AdEx is the exponential term. Its strength can be modified to account for different ion-channel densities in dendrites. No additional voltage-gated calcium channels have been added. However there is some correspondence between calcium concentration and the adaptation variable of the AdEx.

Reset mechanism

Another aspect is the reset mechanism. In the AdEx, continuity is broken after a detected spike which is not realistic as the membrane is pulled down by active potassium conductances in "biology". Implementing multi-compartment neurons, the reset mechanism is obviously very important, as dendritic compartments might not be reset if they detect no spike. In this case they are pulled down by the reset of the soma compartment and the real length of the reset pulse becomes important (Compare to simulations in Section 7.2). In the model, the reset-mechanism breaks continuity - the pulse length is zero.

No strictly model-based design possible

Concluded, the design approach of an explicit model-base design needed to be broken here due to the lack of AdEx models used as complex multi-compartments models. The comfort-zone of using an approved model is left looking at active dendritic channels. However it is still kept for each single compartment and the way neurons are build interconnecting compartments. Nevertheless, direct comparison to HHM based multi-compartment models would have to be done for AdEx compartments to approve biological relevance concerning active channels.

8.2 Implementation

Compartment Interface

The HICANN neurons have been transformed into compartments by adding two tails to the point neuron model. One end is a direct connection to the membrane while the other can be equipped with a new designed resistive element. This resistive element is configurable between 43 kΩ and 1.3 kΩ with a much better linear behaviour than the resistive elements of the AdEx implementation. However, to achieve passive spike propagation to dendritic compartments, the next dendritic can be pulled up by the current compartment. It is reset by its own reset mechanism then. This way correspondence to passive AdEx implementations is kept. Nevertheless, it is a major change which needs to be compared carefully. In active mode, spikes are only propagated by current flux and the local exponential terms in each single compartment.

Dendrite routing

A routing matrix has been implemented to build complex dendritic structures. This way, the two neuron rows can be transformed into a virtual double layered matrix. In addition long range connections have been implemented. The new routing scheme allows a lower ohmic connection when building single-compartment neurons than the current HICANN implementation. Hence, compartmental effects in single-compartment neurons are reduced in the multi-compartment implementation. In addition, the construction of a single-compartment neuron should be simplier with the new circuit.

Compartment scaling

A great challenge in multi-compartment emulation is the different size of the individual compartments which is apparent giving the structure of a neuron (See Figure 0.1). Indeed, size scaling capabilities are limited as each capacitor is implemented by a real capacitor. This limits the compartment numbers usable for an implementation. To account for different compartment sizes, the parameterizing of the AdEx implementation has been enhanced by local switching capabilities of the bias current mirrors. Furthermore, the membrane capacitor can be digitally set for each single compartment. The enlarge parameterizability is even a great advantage if the model is used as a single compartment model.

Active channels

In addition to the exponential term which can be used as active voltage-gated channel in the dendrites, the reset mechanism of the AdEx implementation has been improved. The strength of the pull down after a detected spike can be set locally in each single compartment now. Furthermore the reset voltage is much better stabilized than the reset potential of the single compartment implementation. Additional buffers and blocking capacitors have been used.

8.3 Simulation Results

So far only simulation results exist from the multi-compartment implementation. Complete functionality needs to be proven in measurement indeed.

Nevertheless, results are promising so far. Simulation showed the circuit's capability of reproducing results from a biologically realistic model with passive dendrites. Indeed, the dendritic stimulus has been attenuated by the dendrite.

Furthermore, complex compartmental effects could be observed when working with the circuit creating action-potentials. Much more stable bursting behavior could be achieved without intention. The behavior is indeed much richer than that of a single-compartment model.

These simple results of the experiments make it hard to believe in the sufficiency of a single-compartment model to understand the brains functionality. However, the shown simple simulations are far away for exhausting the complete capabilities of the circuit.

8.4 Other Implementations

Comparing compartment features

The multi-compartments implementation from literature ([103, 106, 108, 109]) are usually based on circuit-driven designed single neuron circuits. Single neuron implementations have been discussed in Chapter 5. Consequently, I focus on issues concerning multi-compartment implementations here.

Routing

The compartment structures from [103, 108] are rather fixed. Implementations allowing flexible dendrite routing are [109] and [106]. The arrangement of the individual compartments is a 2d-matrix structure with usually allowed next neighbour connections by a resistive element. Indeed, possible branched tree structures are limited. More routing freedom is given by the new implementation of this thesis using a virtual double layer 2d routing matrix as basis and long range connections skipping 8 of 16 compartments. A 2d matrix arrangement of compartments is not usable in our implementation as it conflicts with the HICANN structure. Furthermore, it is inefficient if synapses are the area consuming elements of a system.

Resistive element

The most accurate implementations of the inter compartment resistance is given by the switched-capacitor implementation. However, this solution is impracticable in virtue of the necessary additional high-speed clock enlarging noise and power consumption. In addition, good Metal-Metal capacitors consume expensive area at the top most metal layers. The implementation of [106] is a similar approach to the one implemented here as a single transistor is used for the implementation of the resistance. However, due to the available local floating gates, their implementation is much more compact. The current-mode low-pass filter implementation from [109] uses two opposing low-pass filters in parallel. This is a good and compact solution. However, a current-mode subthreshold neuron is needed. Furthermore, the adjustment of the cable constant λ is directly affected by the MOSFET threshold voltage and its fixed-pattern noise.

Compartment sizes

None of the models takes the different compartments sizes into account. This is a great difference in comparison to biology indeed. It is always possible to implement single large compartments by strongly coupled small compartments however. Nevertheless, this approach is limited by the total number of compartments and can be inefficient. Reaching a scaling of 100 between different compartments is unrealistic this way.

None of the designs is prepared for an integration into a large scalable system.

8.5 Conclusion

The designed multi-compartment implementation can compete with implementations in literature. Regarding the parameterizability and the routing capabilities, it allows more evolved mapping of biological neurons on the hardware compartment structure.

In virtue of the model-driven design of the single compartments, a closer model correspondence is kept. Looking at the implementation of active channels, this direct correspondence is left, however, due missing AdEx-based multi-compartment models. Modeling with AdEx compartments needs to be done to achieve a correspondence.

In comparison to the single compartment implementation of this thesis, the multi-compartment implementation offers several improvements even concerning single compartment usage. However, the complexity has been enlarged drastically. Carefully parameter mapping is necessary when working with multi-compartment emulations. For the mapping of dendritic trees, a dedicated mapping algorithm needs to be developed.

The designed multi-compartment neuron is fully integrable in to the architecture of the HICANN. However, before a possible integration, a full verification with the MCC is necessary. The presented design can enhance the BWS for biologically realistic multi-compartment

emulation.

A reduced single-compartment model of the complex structure of a neuron is not biologically realistic. Using a multi-compartment emulation is one step towards a realistic emulation of biological neural networks.

Next, I will present the floating-gate memory cells which are inevitably necessary for flexible the parameterizability of the presented neuron circuits.

9 Analog Floating-Gate Memory

A major advantage of the BrainScaleS Wafer-Scale System in comparison to other neuromorphic devices is its configurability regarding its network connectivity[56] as well as the parameterizability of its analog components[44]. The latter is provided by analog floating-gate memory developed by Jan-Peter Lock and André Srowig[116, 117]. In this work, the floating-gate array designed by Srowig and Lock has been integrated into the HICANN microchip. A controller has been written in System Verilog, connecting the array to the HICANN bus interface. For the multi-compartment chip MCC described in chapter6 a complete revision of the array has been done to gain better programming performance and to save area as multi-compartment neurons need more parameters.

Flash memories are floating-gates

Nearly everyone carries floating-gate cells his pocket, as floating-gates are the basic technology behind flash memory. Basically, a floating-gate is a transistor gate which is not connected to any nets with a pathway to ground or power.

In industrial production a second gate above the floating-gate is used to capacitively couple the floating-gate to a dedicated voltage when it is charged or discharged through Fowler-Nordheim tunneling or Channel Hot Electron Injection. An introduction to floating-gate devices can be found in [118].

No second gate available storing analog voltages

The cells described below use additional transistors instead of the second gate. Only a single poly silicon gate is available in our process. Instead of normally used digital floating-gates, our floating-gate cells are analog and capable of storing any voltage between 0 V and 1.8 V[1].

9.1 Cells

Capacitive coupling and Fowler-Nordheim-tunneling maintain the floating-gate voltage

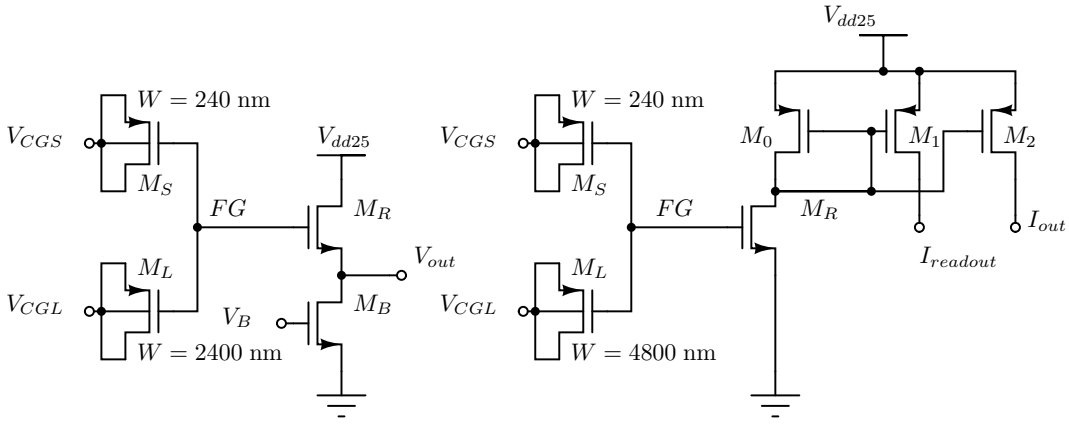
Schematics of the floating-gate cell are presented in Figure 9.1. Transistors M_S and M_L are used for charging and discharging the floating-gate FG while M_R is used as readout transistor only. M_S and M_L are connected as MOS transistor capacitors (called C_S and C_L in the following). The control transistor M_L is 20 times as large as the tunnel transistor M_S. Simplified (without the readout transistor), they form a capacitive voltage divider. For $V_{FG} = 1\,\text{V}$ and $V_{CGL} = V_{CGS} = 1.8\,\text{V}$, which is the static case, let us now set V_{CGL} to 9.5 V[2] while setting V_{CGS} to 0 V. This results in:

$$V_{FG} = V_{CGL}\frac{C_L}{C_S + C_L} - 0.8\,\text{V} = V_{CGL}\frac{20}{21} - 0.8\,\text{V} = 8.24\,\text{V} \tag{9.1}$$

The -0.8 V is the floating-gates voltage when V_{CGS} and V_{CGL} are set to zero. It can be added as the capacitive coupling is linear. The charge on the floating-gate remains constant for this calculation.

[1] Equivalent voltage at the output of the readout transistor, see below.
[2] Some voltage drops in the driving circuits reduce v_{dd11} on the way to the floating-gate.

The voltage V_{FG} is sufficient for Fowler-Nordheim-tunneling to occur[116] at M_S, so the floating-gate would be discharged. If V_{CGL} and V_{CGS} are swapped, the floating-gate is coupled to -0.35 V and the high difference to $V_{CGS} = 9.5$ V causes Fowler-Nordheim-tunneling in the other direction; electrons tunnel from the floating-gate.



(a) Voltage cell **(b)** Current cell

Figure 9.1: Floating gate cells as implemented in HICANN version 1 and 2

Static drift

When V_L and V_S are kept at 1.8 V, direct tunneling and probably ohmic leaking caused by impurities or trapped charge carriers in the gate oxide occur at M_L, M_S and M_R. Consequently the floating-gate voltage slowly drifts towards a voltage that is higher than 0.9 V, as the sum of the gate ares of M_L and M_S are much larger than the gate area of M_R. The floating-gates cells on the described systems come in two variants - voltage cells and current cells.

9.1.1 Voltage Cells

A source follower drives the voltage parameters

Voltage cells (see Figure 9.1a) can output voltages between nearly zero and more than 1.8 V[3] The source follower formed by M_R and M_B drives the voltage to the neurons and the readout line. The Voltage V_B is generated globally (see 9.3.1) for each floating-gate array. M_B works as part of a distributed current mirror.

9.1.2 Current Cells

Current mirrors for the output of current cells

In the current cells of HICANN v1 and v2 (schematic in Figure 9.1b), the source of the readout transistor M_R is directly connected to ground. M_R is a current source now. The generated current is mirrored through M_0 and M_2 to the neurons. M_1 is a separate mirror transistor for the array internal readout. This transistor is four times larger than M_2 to allow larger currents, and therewith shorter readout times. The output current is static, while the readout current is switched.

Unintentional tunnelling through M_R

The main problem of this design lies in the ground connection of M_R as it allows high voltages between gate and ground of M_R. Accordingly Fowler-Nordheim-tunneling can occur. This is especially an issue when cells are discharged. In contrast to Equation 9.1 we cannot

[3]The 2.5 V power supply is needed to keep the readout transistor in saturation region

ignore the readout transistor M_R, as its source is fixed to a static potential. Furthermore, M_R has twice as large gate area as M_S. We will use C_R as equivalent capacitance in the following equations.

If we try to discharge the floating-gate like in Equation 9.1 by setting V_{CGL} to 9.5 V and V_{CGS} to 0 V, we acquire:

$$V_{FG} = V_{CGL}\frac{C_L}{C_S + C_L + C_R} - 0.64\text{ V} = V_{CGL}\frac{20}{23} - 0.64\text{ V} = 7.62\text{ V} \tag{9.2}$$

This voltage would still be sufficient for the process. However, if we now set V_{CGS} to 5 V as we would do to deselect a cell for programming, V_{FG} is coupled to a higher voltage.

$$\begin{aligned} V_{FG} &= V_{CGL}\frac{C_L}{C_S + C_L + C_R} + V_{CGS}\frac{C_S}{C_S + C_L + C_R} - 0.64\text{ V} && (9.3)\\ &= V_{CGL}\frac{20}{23} + 0.21\text{ V} - 0.64\text{ V} && (9.4)\\ &= 7.83\text{ V} && (9.5) \end{aligned}$$

Main source of inner line crosstalk

This is only a problem for current cells as M_R is kept at 0 V and so the discharging of unselected cells is stronger than the discharging of selected cells. This issue has been workaround in HICANN version 1 and 2 by inverting the cell selection during discharging of current cells - still the normal discharging occurs for unselected current cells. The main source of cell to cell crosstalk during discharging (see Section 9.5.2) is generated this way.

9.2 Architecture

The cells are framed by support structures

To allow programming and usage of the floating-gates, they have to be integrated in a support structure. The necessary analog floating-gate architecture is described here. All circuits have been developed by Lock and Srowig. Figure 9.2 gives an overview of the structure. The complete architecture is referenced as floating-gate array.

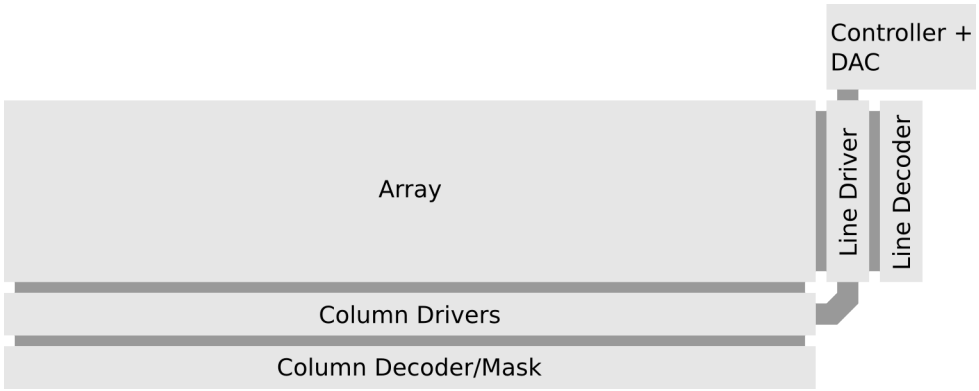


Figure 9.2: Schematic of the floating-gate architecture. Dark grey areas symbolize data and control connections

9.2.1 Array

V_{CGS} and V_{CGL} are used for selection

The floating-gate cells are arranged in an array, where every second line contains voltage respectively current cells. The voltage V_{CGS} is connected column wise and V_{CGL} is connected line wise (Figure 9.3). Cells are selected for programming by setting both, column voltage V_{CGS} and line voltage V_{CGL}, to valid programming voltages. The V_{CGS} or V_{CGL} of lines respectively currents that are not selected, are normally[4] set to 5 V to achieve a small voltage difference between floating-gate and tunnel transistor.

In Addition to the programming voltage lines, a readout line is necessary to check the voltage respectively the current of the cells to control the programming process and for measurement purposes. This connection is routed line wise. Each cell is equipped with an additional switch transistor to switch its output onto this readout line. These transistors are switched column wise, so a single cell can be selected for readout if the correct readout line is connected to the array output.

In both `HICANN` versions, this array contains 24 lines - including 12 voltage and 12 current lines - and 129 columns. 128 Columns are used as individual neuron parameters while one column defines global parameters.

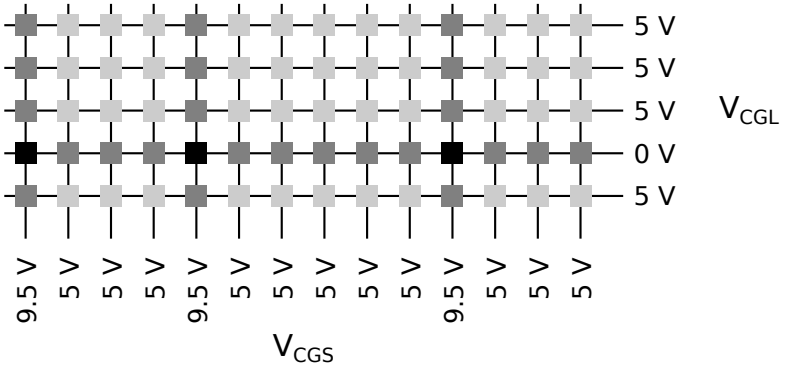


Figure 9.3: Floating Gate Array. The squares represent the cells. Black cells are charged by selection, dark grey cells are charged by crosstalk

9.2.2 Driver

The floating-gate column- and line drivers are responsible for generating the control voltages. A schematic of the core circuit can be found in Figure 9.4.

Static gate potentials at median voltages ensure save gate-source voltages

The key role is taken by the transistor pair M_{*0}. As both gates are connected to static V_{dd5}, the voltage between the gates and the sources can never exceed the break down voltage. Inputs of the circuit are nGH nGL which are generated by logic gates in the driver circuit. Both are 5 V logic signals. V_{bp} is a global biasing voltage for the current source connected transistor M_b. V_{low} gives the lower bound of $V_{CGL/S}$, it is at 1.8 V when no programming occurs or when low $V_{CGL/S}$ voltages are needed during programming[5].

If V_{CGL} should be at high voltage, nGL is set to 0 V. Accordingly, M_{p2} is conducting, and M_{n0} charges $V_{CGL/S}$ to nearly 5 V until it is closed as both, gate and source are at 5 V. Subsequently, nGH is switched to 0 V and the source of M_{p0} is pulled to 11 V through M_{p3}.

[4] For discharging current cells of HICANN version 2 and below have to be treated different (See 9.1.2).
[5] For V_{CGS} when discharging and V_{CGL} when charging

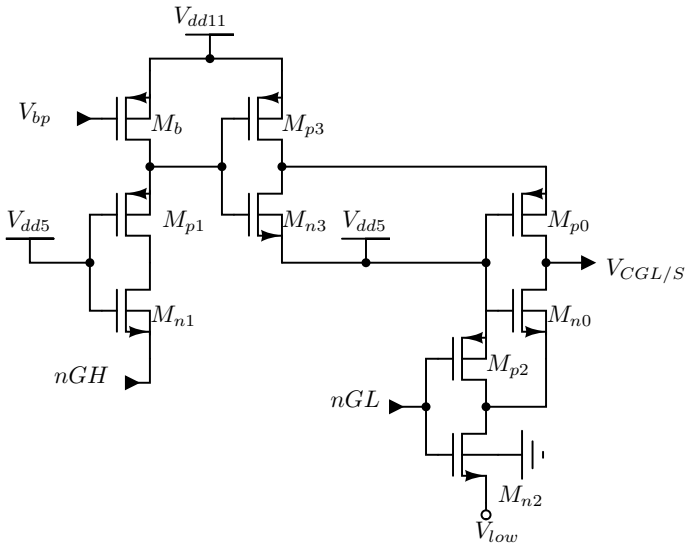


Figure 9.4: simplified schematic of the control voltage driving circuit. All unconnected bulks are connected to the respective transistors source

M_{p0} pull $V_{CGL/S}$ to near by 11 V. After programming, nGH is set to 5 V first and $V_{CGL/S}$ is pulled to 5 V by M_{p0}.

For low V_{CGL} all inputs are kept at 0 V while for deselection of a column or row, only nGI is set to 5 V resulting in a 5 V output at $V_{CGL/S}$

9.2.3 Decoder

Cell selection by shift registers and mask registers

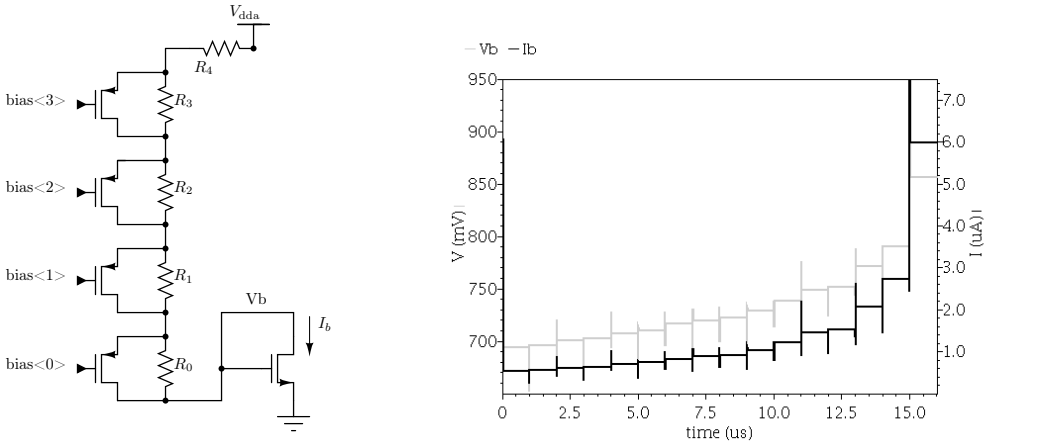
The decoders are responsible for line and column selection and are implemented as shift registers in HICANN v1 and v2 using manually placed standard cells. The column decoder additionally includes a mask register to be able to program a complete line of cells in parallel. Cells which have to be programmed are marked in this register. After each programming step, the cells are checked and the register is reset if the process is finished.

9.2.4 Controller

The analog controller part contains a digital-to-analog converter(DAC) to generate analog voltages from a digital 10 bit word, and a comparator to compare the floating-gates output voltage with the DAC voltage. The DAC is realized as R2R-DAC; the comparator uses the switching threshold of an inverter.

9.3 HICANN Integration

To integrate the array, designed by Jan-Peter Lock and André Srowig[116, 117] described above, several changes and additions had to be implemented.



(a) Schematic

(b) Transient simulation: The bias vector is decremented every microsecond. At 16 µs, the maximum current defined by R_4 is reached.

Figure 9.5: Biasing generator for V_b

9.3.1 Biasing

External biasing removed

The floating gate array in its state at the beginning of this thesis needed 2 biasing voltages respectively currents. V_b is the gate voltage of the biasing NMOS current source of the readout source follower of each voltage cell (see Figure 9.1a). V_bp is a biasing voltage for the floating-gate driver (see Section 9.2.2). This voltage needs to be close to the high voltage V_dd11.

Manually tuned resistors allow bias current modification

Figure 9.5a is the schematic of the lower biasing generator. Resistors directly connect between the input of a current mirror and V_dda power. To be able to change the biasing current, the resistor is divided into five parts, where four of them can be bridged by transistor switches. The fifth cannot be switched to prevent high current flux. The input voltage of the current mirror changes if the current changes, so it is not sufficient to scale the resistors binary to achieve the full 4 bit adjustability. Simulations have been used to tune the resistors to achieve a monotonic behavior when switching and to gain maximum parameter space. Results can be found in Figure 9.5b.

additional circuits needed for V_{bp}

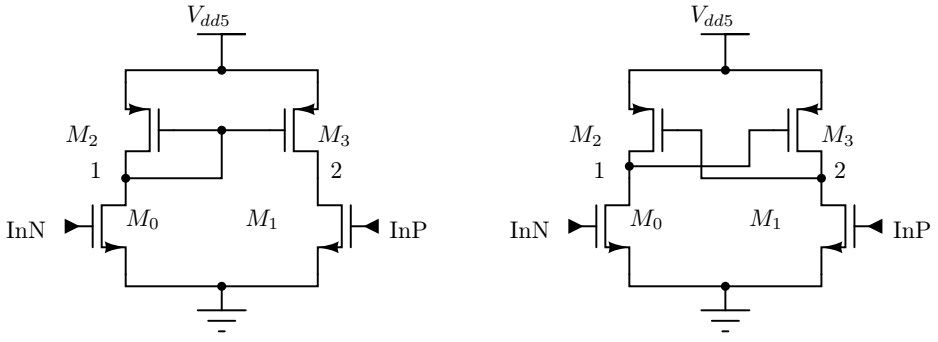
The circuit above can be directly used to generate V_b. For the high voltage bias, the same circuit can be used if it is enhanced by another current branch. Several PMOS and triple well NMOS are needed to avoid critical voltage drops on a single device as the complete chain has to connect V_dd11 and ground.

These circuits are implemented in the facets_fg microchip (A prototype for evaluation of the Floating-Gate-Array. See [82]) and both HICANN prototypes. As there are several potential problems(power supply rejection for example) with these circuits and better solutions exist in literature, new biasing circuits have been designed for the MCC(6).

9.3.2 Level shifter

Translation between 1.8 V and 5 V digital power domain

Level shifter are used in several parts of floating-gate array control circuitry as the default digital power supply is at 1.8 V while 5 V are needed in the column- and line drivers. Furthermore, 5 volt switching is needed at one point to achieve full 1.8 V swing during readout



(a) Old: high cross current

(b) New: only leak current

Figure 9.6: Input stage of level shifter. Next stage is connected at node 2.

of the floating-gates without being forced to use transmission gates (Those are much less area efficient.).

High cross current on HICANN version 1

A circuit diagram of the input stage can be found in Figure 9.6a. In valid operation, InN is always the inverse of InP. The problem occurs if InN is at 1.8 V. As M_2 and M_0 are conducting, static current will flow from V_{dd5} to ground. In fact this results in 80 mA current on V_{dd5} in HICANN v1 instead of less than 1 mA as specified during design[6]. The power routing for V_{dd5} has not been laid out for this amount of current, so this was a severe problem. Luckily, measurements with HICANN v1 have still been possible, but voltage drop could lead to unforeseen behavior(such as a different programming accuracy for instance) especially for the arrays on the left side of the chip, where the distance to the bonding pad is at maximum.

In HICANN v2, the level shifter has been replaced by the circuit in Figure 9.6 b). Now M_2 is off when M_0 is conducting and the total chip static current on V_{dd5} is less than one mA.

9.3.3 Global Parameters

Global Voltages are driven by operational amplifiers

The array described in Section 9.2 supplies 128 neurons with individual parameters and reserves 12 currents and 12 voltages for global parameters. As the cells are not built to drive high capacitive loads of global parameters, an array of operational amplifiers[76] has been integrated into the array. The output of amplifiers has been fed back into the floating-gate controller circuit for comparison with the wanted value instead of the actual floating-gate value of the global parameter. This way, offsets of the differential input stage of the parameter drivers are compensated automatically for instance. The reset potential of the neurons is created this way as it needs to be driven by a low impedance.

Global currents are not distributed as currents but via the control voltages of global current mirrors. These voltages are blocked by large MOST capacitors.

9.3.4 Current Source

Neurons can be stimulated by a variating current

Each floating-gate array has been enhanced by an operational amplifier from [76] connected as a current source (See Figure 3.32 in Section 3.11.2). This current source is used for

[6]This makes v_{dd5} the maximum power draining supply in most experiments

stimulating neurons in most experiments in this thesis and can be used for calibration in addition. The input voltages of the source come directly from the DAC and thus can be digitally set.

9.3.5 Layout

The changes that needed to be made a new placement of the controller, as the initial placement did not fit to the metal constraints, and a modification of the cell pitch. Synapses needed to be wider than expected before layout, so the floating-gate cells had to be realigned.

9.4 Digital Controller

Breaking the time scale gap between array and slow control

To program the floating-gates, pulses of several hundreds of nano seconds are necessary. Before comparison of the floating-gate value to the wanted value, long waiting times are needed for settling due to RC-delays. The clock period of the slow control clock of the HICANN is, on the other hand, 16 ns. To interface the array digitally and to optimize the writing process, a controller has been written in the hardware description language System Verilog. Each array instance has its own controller which enables parallel writing of all four arrays to counterbalance the slow speed of the programming.

9.4.1 Programming Functions

programming in general

As described above, all values of a floating-gate row are programmed in parallel (See decoders in Section 9.2). At first, 10 Bit values for each desired value of a row are loaded into one of the two[7] memory banks of the floating-gate controller. Then a write process is initiated. During the write process, the controller needs to be polled to check if the writing has finished. When the state machine has finished, a signal marks if some cells did not reach their desired value. By polling the controller, these cells can be identified.

During programming, all cells are compared with their desired values first and marked if their values are already reached. Finished cells are excluded in the following. In the next step, the cells are charged or discharged. Both steps are repeated until either all floating-gates are marked as finished or the maximum number of writing cycles is reached.

Differentiating Current and Voltage Cells

The floating-gate controller is capable of charging and discharging current and voltage cells with dedicated writing pulse lengths. This is necessary because current cells are much more sensitive during programming. While the output voltage of the current cells is equivalent to the voltage of the floating-gate because a source follower is applied, the current cells use the characteristic of a MOSFET for translation. The parameter range for floating-gate voltages is very small for the current cells as the characteristic is exponential (sub threshold) and quadratic (strong inversion). The programming pulse directly effects the floating-gate voltage. Consequently current and voltage cells need to be treaded differently.

adaptive writing

During writing, the difference between the floating-gate voltage and the programming voltage decreases and therewith the effect of the programming pulse. To gain good writing performance for smaller differences (high voltages during charging and low voltages during discharging) while the programming accuracy is kept for large differences, the length of a programming pulse is adapted.

[7] There are two banks to be able to fill a bank with values from the slow control, while the controller uses the other one for programming

9.4.2 Additional Functions

analog readout

Each floating-gate array has the option to to connect one cell to the analog readout for debugging and calibration. This function is supported by the controller with a dedicated controller command.

neuron stimulation

The memory used for programming can also be used as playback memory for neuron current stimulation (see 9.3.4). With each internal clock cycle (OCP clock slowed down by `pulselength`, see Table 9.1), the next RAM value is connected to the DAC. There are two different modes. One is a continuous loop over of the playback of one memory bank while the other just repeats the loop for a settable number of times.

9.4.3 Detailed Implementation

architecture overview

The floating-gate controller consists of three components. The module `fgateCtrl` is the top level module and connects the other modules to the OCP interface. The controller state machine is located in the module `fgateCtrlSlave` which is the core module of the design. The module `FG_RAM` encloses a single-port-write-dual-port-read latch memory[119] from the *Synopsis Designware* library. Its interface is translated to fit the 32 Bit data width of the OCP-Interface and the 10 Bits floating-gate DAC word - 20 Bit or two DAC words are used here. Furthermore the two banks are implemented here to allow reloading of the RAM while the floating-gate controller is running.

`fgCtrlSlave` is the core state machine of the design

The state machine of the module `fgCtrlSlave` is implemented using the three process state machine architecture introduced in [120]. This state machine design uses one process for state switching, one for the next state generation logic and one for registering outputs. The state change mechanism has an integrated counter which allows to slow down the state change by a factor of up to 32 to the slow OCP clock rate. This is done to allow smaller counters in the rest of the machine. The machine is started by a request signal generated in `fgCtrl`. At first the instruction is decoded and everything is prepared for command execution. Next, the command is executed. All time scaling parameters such as the write times for current and voltages are implemented by counters.

`fgCtrl` connects the OCP bus to RAM and state machine

The shell module `fgCtrl` multiplexes the OCP interface between the memory and the controller using address bit 8. All static parameters for the array and `fgCtrlSlave` are mapped on three registers: the instruction register, the operation register and the bias register(see Table9.1) . To trigger the state machine, the register 256, called instruction register has to be written. This register holds the instruction code, the active ram bank number and the column- and line number of the dedicated cells. Every time the register is written, a request signal is generated for the state machine.

9.4.4 Test Environment

FPGA verification

During the design of several ASICs the test environment of the floating-gate controller has been optimized. While first test always needed manual control of wave form charts, the test for HICANN v2 and MCC use a complex behavioral[8] giving feedback of the success of a programming step. Before being integrated into the HICANN chip, the controller has been checked by programming it onto an FPGA to control the microchip facets_fg, a prototype, designed within this thesis, for testing the floating-gate architecture. (Results can be found in [82]).

Simulation on module and chip level

[8] A behavioral is a model of a circuit reproducing the behavior of the circuit. This way, simulation effort can be shrunk drastically.

parameter	register	function
fg_bias	bias	driver bias
fg_biasn		cell source follower bias
pulseLength		clock cycle multiplicator to slow down the state machine
groundVm		short cut the parameter V_m to ground
calib		set array to calibration mode
maxCycle	operation	maximum number of programming steps
readTime		number of state machine clock cycles waited for settling of the readout line
acceleratorStep		number of programming steps between doubling of programming time (adaptive programming)
voltageWriteTime		initial length of programming pulse for voltage cells
currentWriteTime		initial length of programming pulse for current cells
columnNumber	instruction	line (column and line are switched internally)
lineNumber		column
bankNumber		ram bank number

Table 9.1: Parameter of `fgCtrl`

For the submissions of the different versions of the HICANN microchip, the controller has been checked checked on two scopes. The chip test bench allows to directly control the simulated module integrated into the complete HICANN design together with all other components. This test is done for final verification. For component tests a test bench is used which controls the module via the OCP interface. This bench is used during development to quickly verify changes. Both test benches use the same behavioral for floating-gate array simulation and implement similar command chains.

Behavioral

The floating-gate behavioral simulates the real floating-gate array as a mixed signal simulation is usually too elaborate. The behavioral includes an array of 10 Bit values equivalent in size to a real floating-gate column. These numbers are decremented or incremented during strobe. A value dependent behavior is not implemented. For MCC verification, this behavioral has been enhanced.

9.5 Sources of Variance

Systematic sources of variance can be removed by calibration

During the design phase the Floating-Gate Array has been constrained to have an accuracy of 4 mV [61] for voltage cells. If we define this accuracy as reproduction of a written value, the constraint can be achieved. However, this definition excludes all systematic caused by miss match in the DAC [9] for example. It is sufficient as those systematic effects are eliminated by neuron calibration. Here I will list the different error sources causing variance when writing parameters to the floating-gate array. A complex analysis has been done by Alexander Kononov in his diploma thesis [70].

[9] The resistors in the R2R-DAC for comparing the floating-gates to the desired values are exposed to process variation causing a systematic variation when programming a dedicated array.

9.5.1 Parameter Drifts

3.3 V transistors with thick gates have been chosen to improve the durability of the programmed values [116] to allow storing of parameters for a reasonable time. Nevertheless, even through the thick gate, tunneling or leaking through impurities and trapped charge carriers occurs. Parameter drifting limits the length of an experiment if a certain level of accuracy is necessary. Typical emulations of 4 hours should be possible with the array.

9.5.2 Crosstalk

Cell selection via programming lines results in crosstalk

Crosstalk occurs when the floating-gate cells are written as they are arranged in an array and the cell selection is done via the programming lines (see Figure 9.3). A row is selected by putting its programming line to high or low voltage while keeping all others at intermediate voltage. The same is done for valid cells. This way not selected cells in a selected column or row still see the row's or column's programming voltage. Of course, the differences are smaller as one gate is set to the intermediate voltage, but it is still sufficient for tunneling. This way the programming of a cell changes other cells in the same row or column in a parasitic fashion. One way to eliminate this easily is to program the complete array two times in a row as the effect is much smaller in the second step. The cell are close to their desired values than and fewer programming pulses are necessary.

Crosstalk is especially a problem when discharging current cells (Section 9.1.2).

9.5.3 Output Settling and Strobe

Charging of the readout line

The most critical parameter when programming the floating-gates is the settling time of the array readout line as every cell has to be read out individually after every strobe pulse. In the worst case, this line has a length of 1.6 mm which results in a maximum resistance of 570 Ω and a total capacitance of 280 fF[10] at max. The measurement capacitor of the comparator itself has a value of 500 fF. A simple model of a transmission line assumes two third of the line capacitance at the end of the line and one third at the beginning[68]. Consequently we have to charge about 700 fF via the line every time we read out, resulting in a time constant of 0.4 ns. We want to achieve a accuracy of 4 mV over a value range of 1.8 V, which takes about 6 time constants or 2.4 ns for readout.

The output resistance of the readout source follower is the bottleneck

This result does not fit to the measurements, so a simulation has been set up including all components of a voltage cell's output line including the floating-gate readout transistor M_R. Not the ohmic resistance of the output but the bias current of the floating-gate readout it self is the settling time defining bottle neck. Simulation results in more than 1.5 µs settling time for a relatively high bias of 1 µA when switching from 1.8 V to 0 V. Settling time for rising edges is much smaller (less than 200 ns) as the maximum current here is not limited by the biasing current. Unfortunately the line is pre charged to 1.8 V before readout, so the falling edge settling time has to be accounted every time a cell is read out.

If a cell has not been properly detected as finished it will be exposed to another strobing pulse. Depending on the length of this pulse the induced error of this cell can be much larger than the desired 4 mV. To achieve higher accuracy at a certain value, smaller strobe pulses should be chosen.

[10] Calculated using conductor between adjacent planes model from foundry design kit.

9.5.4 Programming Limits

Biasing can limit programming ranges

The natural limits of the readout values of the floating-gate array are the power rails. Nevertheless, a small bias can lead to the lower rail not being reachable. On the other hand, a high bias in combination with short strobe pulses leads to a lower upper limit. Especially the upper limit can cause an accuracy deficit as the actual maximum depends on the individual cell.

9.6 Improvements

Changes were implemented for the MCC

Especially during the development of the multi compartment chip, several improvements have been introduced. Those involve programmability, crosstalk, control, biasing and area consumption. These changes were necessary to gain area for the new multi-compartment structures and to allow more parameters.

9.6.1 Control, Driver and Decoder Revision

Replacing standard cells in full custom logic

The supporting modules of the FGA, the controller, the decoders and line- and column drivers described above are designed using manually placed old standard cells which is a very area inefficient approach in full custom design. In a first step all standard cells have been replaced by smaller full custom equivalents.

Shrinking layout

In layout, spacing needed between wells of non equal potential results in a larger area consumption. To avoid this problem, wells have been shared whenever possible during redesign of the layout.

Replacing shift registers

Selection of rows and columns has been done using large full custom shift registers. In the digital controller, counters send a certain number of pulses for selection. As row- and column number have to exist in the digital controller anyhow, there is no reason not to use them for directly addressing rows and columns, so the shift registers have been replaced by small decoders.

At all the revision of the controlling structures shrank their vertical extension from 144 µm to 60 µm which would be equivalent to nearly 6 synapse rows more on each side of the chip.

9.6.2 Cell Revision

Much more current cells than voltage cells

In contrast to global parameterization, individual parameterization of the neuron circuits needs more current parameters than voltage parameters. The multi-compartment design continued this trend and finally 15 current parameters and 10 voltage parameters were needed. This conflicts with the interlacing of current and voltage cells in layout as here 4 voltage cells would have to be integrated without being used. The layout of the cells had to be redone.

New optimized cells laid out

The new layout used symmetries and n-well-sharing to shrink the area consumption of the complete array drastically. The array of the MCC has a vertical extension of 236 µm for 10 voltage and 18 current rows. This competes to of 283 µm for 12 current and 12 voltage rows of the array in HICANN version 1 and 2.

Cutting current cell's M_R off ground
Using source degeneration for a better characteristic of the current cells

A major problem of the floating-gate array described above is the discharging of current cells. As described in Section 9.1.2, tunneling through the readout transistors of not selected cells dominates the tunneling through the tunneling transistors of selected cells, as the readout transistors' sources are fixed at ground. This ground connection has been cut by a switch NMOS transistor in the new design. Additionally, the gate potential of the switch transistor can be externally set to a dedicated voltage to obtain a voltage drop via source degeneration. This way, the quadratic characteristic of the readout transistor can be counterbalanced to

allow a larger valid range for floating-gate voltages in current cells to improve programming accuracy.

9.6.3 Biasing Revision

Power supply sensitivity

The biasing circuitry described in 9.3.1 is relatively large as large resistors are needed and the power supply rejection ratio is improvable because V_{dda} is directly used as reference for current generation.

Self-Biasing

In the MCC a new circuit using a technique called Self-Biasing has been integrated. The technique is introduced in [25] for example. A schematic of the circuit can be found in Figure 9.7.

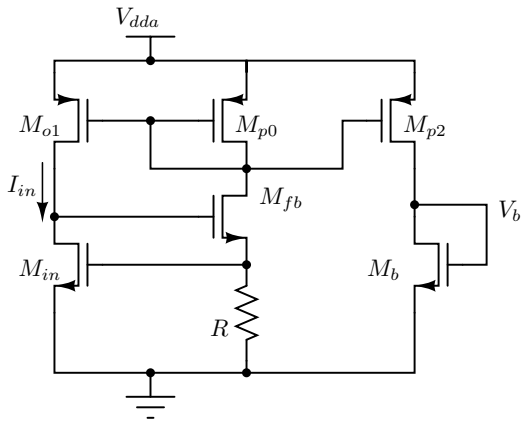


Figure 9.7: New biasing circuit using Self-Biasing technique, startup circuit excluded

The circuit consists of three parts: the current source formed by M_{in}, M_{fb} and R, the current Mirror $M_{p0,1}$ and the output circuit which is formed by M_{p2} and M_b. M_{in} shall be operated in strong inversion; if we ignore channel length modulation, we obtain:

$$I_{in} = \frac{k'}{2}\frac{W}{L}\left(V_{GSM_{in}} - V_t\right)^2 \tag{9.6}$$

Here $k' \approx 125\,\mu\mathrm{A/V^2}$ and the threshold voltage $V_t \approx 0.7\,\mathrm{V}$ are process parameters. For simplification reasons I introduce $K = {}^{k'W}/_{2L}$. Due to the chosen dimensions for M_{in}, $K \approx 350\,\mu\mathrm{A/V^2}$. If we assume ideal mirroring with a factor of 1 for the current mirror $M_{p0,1}$ we can substitute $V_{GSM_{in}}$ in Equation 9.6:

$$I_{in} = K\left(I_{in}R - V_t\right)^2 \tag{9.7}$$

Completing the square and doing some algebraic transformations, results in:

$$I_{in} = \frac{1}{R}\left(V_t - \frac{1}{2RK} \pm \frac{\sqrt{4V_tRK - 1}}{2RK}\right) \tag{9.8}$$

$$\approx \frac{1}{R}\left(V_t + \sqrt{\frac{V_t}{RK}}\right) \tag{9.9}$$

I excluded the negative solution of Equation 9.8, as the therm in brackets is equivalent to $V_{GSM_{in}}$, which is operated above the threshold voltage and assumed $2RKV_t$ is larger than 1 for simplification. I_{in} depends only on the threshold voltage of M_{in} - it is independent of the power supply voltage which is the main advantage in comparison to the old circuit from section 9.3.1. Of course the threshold voltage will vary from chip to chip and from circuit to circuit, but those variations can be counter-balanced by switching R to different values.

Good temperature performance in simulations

Still, temperature dependency could be an issue, but DC simulations showed, that the expected current change at the voltage cell's current source M_B (see Figure 9.1b) is less than 5 % when sweeping the temperature from 20°C to 80°C.

Start-up critical

Equation 9.9 is not the only DC solution of the circuit. I assumed, that M_{in} is operated in Strong Inversion. Indeed, a DC solution of the circuit would be no current flowing at all. Leaking currents usually will allow this solution to be stable. To prevent the zero current solution, a start up circuit is needed to push the circuit in the desired region. The chosen circuit pulls down the gate of M_{p0} when $V_{GSM_{in}}$ is too low.

Danger of instability

A danger in this circuit lies in the AC and transient behavior if a large capacitor is added at the gate of M_{in}, which would typically be the case if this node is used to drive a large distributed current mirror like the current sources of the floating-gate array. The node at M_{in}'s drain will be able to change much faster then, and the circuit will develop an oscillating behavior. This is why the output circuit with M_{p2} and M_b have been added.

Monte-Carlo simulation of the circuit at minimum current adjustment results in a sigma of 5 % for the internal current I_{in} while the sigma for the crosscurrent of the readout source follower of the voltage cell is about 20 % if Monte-Carlo Simulation is done only for the cell's current source. This static variation is not a problem as it is counterbalanced during programming of the cells.

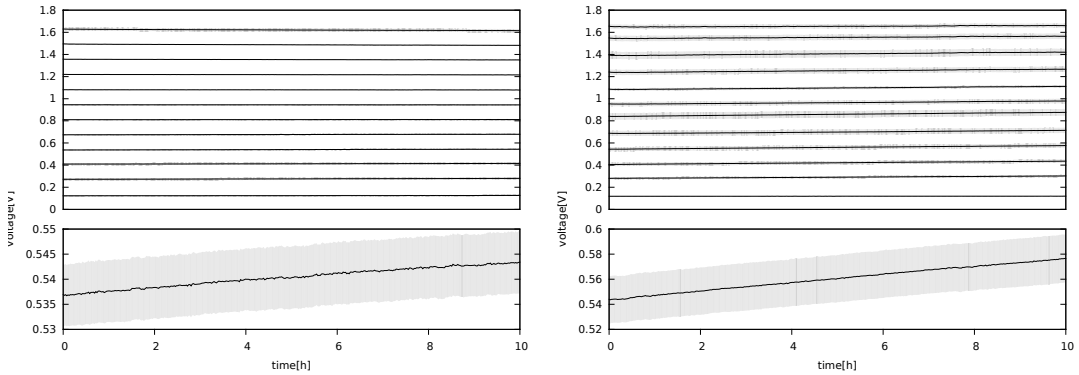
9.7 Test Results

This section begins with measurements which have been done in the first few months of HICANN v1 testings, or reproductions of those measurements on HICANN version 2 to obtain an overview about the floating-gates performance. Some of these measurements are presented in the FACETS Deliverable D-7-7[121]. Detailed measurements concerning precision, crosstalk of the programmed floating-gate values and a stress test are presented next. Here, each measurement is done for both generations of the floating-gate cells. The detailed precision, crosstalk and stress test measurements on HICANN have been carried out by Alexander Kononov, a supervised student, and are published in his diploma thesis[70].

9.7.1 General Functioning

To show general functioning of the array in [121], all floating-gate cells of a line in one array in one HICANN version 1 have been programmed to the same value. With rising line number, this value has been incremented to cover the whole value range which is externally limited by the power rails of the analog output. All cells have been read out automatically using a digital multimeter. Its interface has been integrated into the system software. A minimum deviation of 4 mV was achieved during these measurements with a maximum programming time of 100 ms per line. The progressing of the medians and the deviations of the line is shown in Figure 9.8. Voltages are plotted for current cells as the currents are read out via the voltage drop at the readout resistor.

Voltage drift is much stronger for current cells; 3 mV against 20 mV in 10 h. This is due to the translation of the floating-gate voltage into a current by the characteristic of the read



(a) Voltage cells **(b)** Current cells

Figure 9.8: Deviation and value propagation with time for different programming values as done for [121]. Each line corresponds to the average of a complete floating-gate line. The values have been read out periodically with a multi meter. The lower graph is a zoom into upper one. The drifts do not effect the standard deviation.

out transistor which is quadratic at least. The characteristic of the readout source follower of the voltage cell is linear in comparison. The internal drift of the current cells is only quarter of the one in Figure 9.8 as the current is quadrupled for faster readout.

Photo Electric Effect

Illumination changes floating-gate values

At the beginning of the floating-gate measurements, parameter drift results have been much worse with HICANN v1. Exemplary measurement results can be found in Figure 9.9. Large drifts of 100 mV have been observed. In some sections, the change was even 0.8 V for 7 h. These sections occurred periodically every 24 h at day time (Measurements have been done in January or February). The cap protecting the chip has been transparent for this first chip and strong drifts took place at daylight time. Electrons photons and gained enough energy to cross the gate oxide barrier. Depending on the angle of the light the effect dominates for electrons added to the gate or removed from the gate due to metal structures above the cells. Consequently, some cells are drifting to higher voltages and some to lower. Furthermore, the slope is different due to the angle. The lower boundary can be explained as 1.8 V minus the threshold voltage of the readout transistor. Photoelectric effect dominate at the programming transistors in this case. The upper boundary, which is close to the chips power supply, is given by the 2.5 V supplying the readout transistor minus its threshold voltage. In diffuse light at night, the angle dependency nearly disappears and all cell's voltages are drifting upwards.

The strong drifts could be eliminated by the use of black caps.

9.7.2 Programming Schemes

Differential programming did not produce reproducible results

Kononov distinguishes between two programming schemes[70]: *sequential* - and *differential programming*. The latter is the programming scheme, which has been aimed during design: cells are first discharged as far as necessary and than charged if needed. This way the writing time would be minimal. Unfortunately, especially during the work for [44] this scheme did not

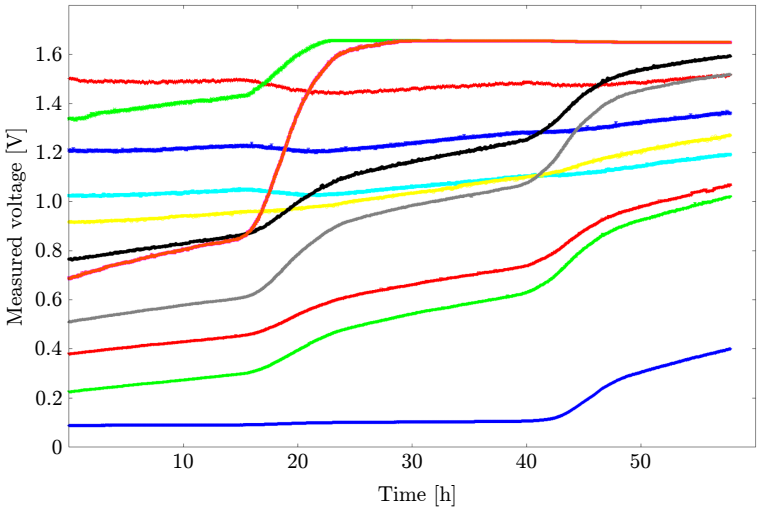


Figure 9.9: Photo Electric Effect in one of the first measurements of the floating-gate arrays of HICANN v1. 11 cells are programmed to different values and read out periodically using a multimeter. (measurement error of each measurement can be estimated to below 5 mV. However, this measurement is qualitative.)

create reproducible results due to the crosstalk problematic described above. Changing one parameter influenced all others. This was especially the case when cells have been charged first and are subsequently discharged (crosstalk is the worst for discharging current cells).

Sequential programming as workaround

The workaround was to fall back to *sequential programming*. All cells have been programmed to zero followed by charging each single line of floating-gates to the desired values. This way, reproducible results could be produced for the neuron measurements presented in [44].

Single differential programming is only better for writing a single line

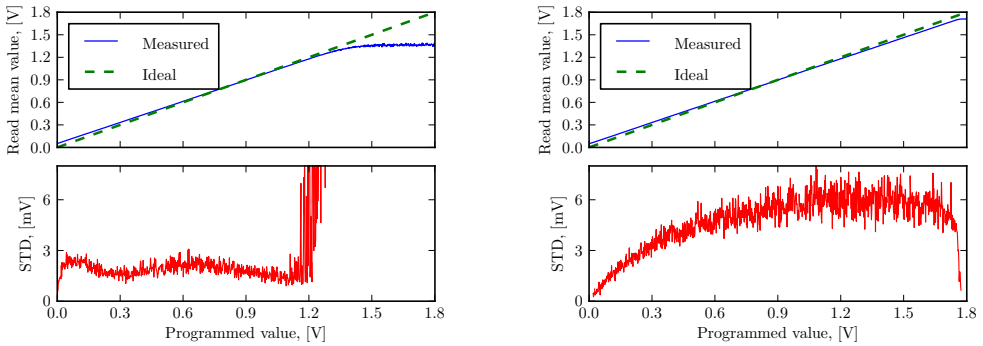
In [70] the *differential programming* is producing the best results when programming a single line with random numbers. These measurements only look at a single row. It is evident, that this programming scheme is the best here, as fewer and smaller programming steps are needed and so the crosstalk is reduced. This is correct if the other lines can be ignored. If all lines need to be programmed to achieve the same neuron behavior from a parameter set, each time it is programmed, a single *differential programming* step is not sufficient due to crosstalk. All other lines need to be reprogrammed right after the changed line. However, this could make the programming scheme complicated as any number of parameters could change. The failsafe method is to program all cells twice each time (See crosstalk measurements in [70]).

9.7.3 Precision

For measuring the possible writing precision for current and voltage cells, Kononov has developed a measurement method which is more precise than the complete line programming method described above. The method used in [121] does not include internal line crosstalk as all cells are programmed to the same value. Nevertheless, it is not unrealistic as values are expected to be in the same region in many experiments at least for HICANN v1 and v2.

Kononov programs random numbers into the array and defines the *response function* of the array as the projection between these numbers and the corresponding readout voltage.

Several of these functions are recorded to gain a median *response function* and a standard deviation. In his precision measurements, Kononov uses ned optimal values for the controller's parameters that were determined beforehand and the *differential programming* scheme.



(a) Voltage cells. No voltages larger than 1.35 V could be programmed with the applied parameter choice. The standard deviation goes up as the finally reached value differs from cell to cell.

(b) Current cells. The standard deviation goes down for large values as the maximum voltage is given by the power supply.

Figure 9.10: Programming accuracy. Copy of Figures 4.14 and 4.15 from [70]. Due to a lower power supply respectively a higher ground level, the (extrapolated) traces to not reach 1.8 V respectively 0 V.

Results can be found in Figure 9.10. While the current cells have a larger deviation, the voltage cells do not reach values higher than 1.8 V, although longer writing pulses are used for voltage cells. The current cells are capable of creating currents larger than the measurable input range of the analog floating-gate controller[11]. In addition, the maximum output-able voltage of the HICANN is given by the power supply. The output amplifiers can reach DC values up to several millivolts close to the power rail. Due to this border, the standard deviation is going down in these regions. However, the measured value is not the floating-gate value but the power supply instead. The saturation below the programmed value of 1.8 V points to a lower digital power supply at the output amplifier than the analog power supply at the DAC used for comparison of the cell values.

The current cells are expected to have a better performance in the MCC because the readout transistor's source has been cut from ground during programming.

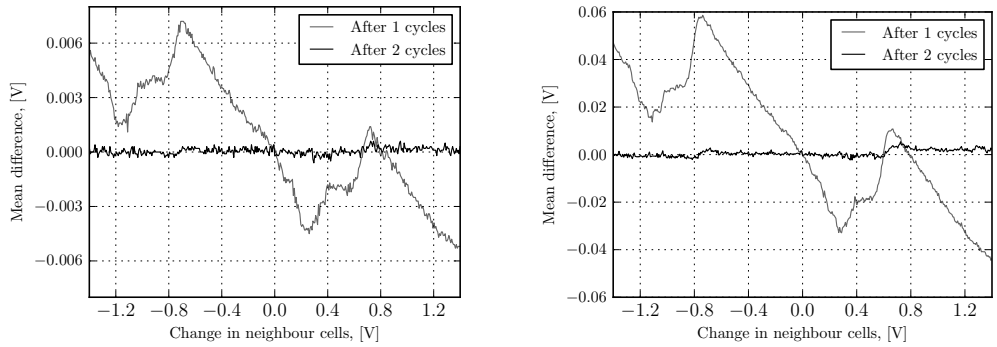
9.7.4 Crosstalk

Line-to-line crosstalk

Crosstalk here means line-to-line crosstalk as inner line crosstalk is already covered in the precision measurements. The measurements described by Kononov in [70] use the following scheme to determine crosstalk: All cells of the examined line, called *control line* are programmed to 900 mV, a voltage in the middle of the programming range. The start value of all other cells depends on the rest of the experiment; if crosstalk while charging cells is to be determinated, a voltage of 180 mV is chosen. For the other directions, the disturbing cells are initially programmed to 1.62 mV. After this initial conditions are set, the mean and the standard deviation of the *control line* are measured and calculated. Subsequently, all

[11] Currents causing a drop larger than 1.8 V on the measurement resistor.

other lines are charged or discharged to achieve a voltage change of ΔV. At last, the mean and standard deviation of the *control Line* are determined again. The complete experiment is repeated for different ΔV and the dependency of mean and crosstalk on ΔV is plotted. Additionally, the programming of the *control line* is repeated after each experiment step with differential reprogramming of all other cells afterwards. This imitated the effects of several differential programming steps. Kononov's results can be found in Figure 9.11.



(a) Voltage cells **(b)** Current cells

Figure 9.11: Crosstalk. The shown traces are average values of complete floating-gate lines. The standard deviation of the line values did hardly change due to crosstalk - it is around 3 mV. However, indeed change can be expected, if different values are programmed to the other cells. (Modified versions of Figures 4.12 and 4.13 from [70].)

The expected trace after a single cycle should be monotonic. Indeed, if not all lines are written, monotonic decreasing values can be observed (See [70]). Occasionally, effects can be worse than the presented, when monotonic crosstalk is observed. Nevertheless, after eleven lines are written, the current trace looses monotony and transfers into the shape presented in Figure 9.11 b). To understand this effect completely, further investigations would be necessary. Actually the most proper explanation could be a systematic measurement error. Especially the similarities between voltage and current measurement figure and within a figure are suspicious. The good news is, that crosstalk affects can be reduced to a level below the necessary programming accuracy by applying two write cycles of the complete array as only small changes have to be programmed in the second cycle.

Cross talk affects current cells much more than voltage cells. However, the large current cell crosstalk described above is not visible here as this crosstalk mainly occurs within a line. It can be counterbalanced by programming a line two times first down and than up.

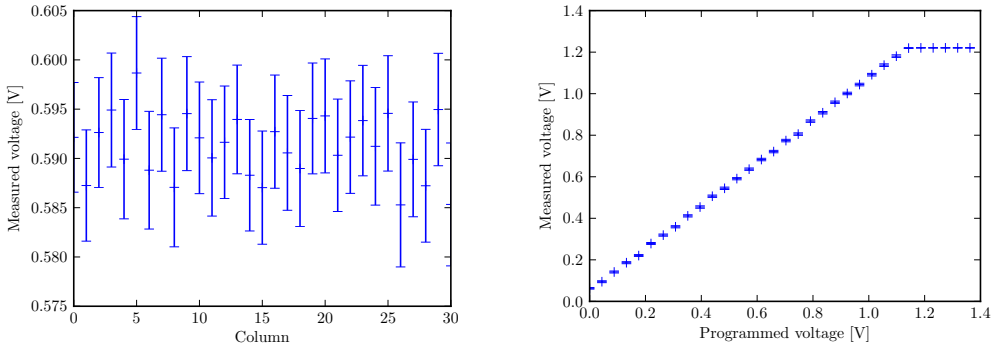
9.7.5 Stress Test

During the first measurements of the chip HICANN version 1, some floating-gates, including complete columns and lines could not be programmed after a while. This is why the suspicion arose that continuous programming could destroy cells, or the circuitry in line - or column drivers. In earlier experiments, the destruction could not be reproduced, so Kononov did an extensive stress test to try to destroy cells or drivers. Luckily these tests did not show any destruction although V_{dd} was even set to 11 V.

9.7.6 New Cells

New voltage cells are working

Currently only the voltage cells of the new floating-gate array in the MCC can be measured. Results pointing out value range and programming accuracy can be found in Figure 9.12. The upper voltage limit is given by measurement setup constraints, however. Indeed, programming accuracy of this measurement is worse than the accuracy of the old floating-gate array. This is probably caused by too large currents on the 5 and 11 V power supply which are larger than the setup specification. Those have been removed in the next revision, hence better precision is to be expected. The main result of this measurement is that is is possible to program the newly designed voltage cells. Current cells cannot be used in the current version of the MCC.



(a) All cells of the line are read out for a single programming step.

(b) Several programming steps are performed. The same cell is read out while the programmed voltage is incremented. Saturation occurs due limits of the measurement setup.

Figure 9.12: Measurement of new voltage cells of the MCC. All cells of a line are programmed to the same value.

9.8 Discussion

The presented floating-gate solution offers individual current and voltage biasing parameters for analog circuits. The specified standard deviation of 4 mV for voltage cells between different programming trials would correspond to biological differences between 0.2 mV and 0.4 mV depending on the chosen neuron voltage scaling factor. Stability of the values allows for experiments running several our without reprogramming depending on the needed parameter accuracy.

I presented changes to reduce crosstalk, shrink chip area usage, simplify addressing and stabilize biasing. However, the improvements done for the MCC still have to be proved in measurements. So far only measurements with the voltage-cells have been possible. So far, the array currently implemented in the HICANN is doing a good job although not everything is completely out bidden.

Crucial in current technology

The Floating-Gate Array is crucial for the parameterizability and calibration of the system. Other solutions like digital memory based individual ADCs as used in the SPIKEY microchip consume too much area in a 180 nm process to implement the necessary amount of parameters. However, things might change when moving to a process with smaller feature

sizes. In particular, floating-gate cells might not be implementable on a more modern process at all.

High effort

The cost of using floating-gates is huge indeed. Special power voltages are needed that are above the design constraints of our given process. These high voltages have to be treated with care inside the ASIC as they can easily destroy any device. In addition, the floating-gates cannot be implemented without additional test ASICS. The complete architecture presented here took estimated 5 man years and there is still lots of room for optimizations.

Still high potential

However, the potential behind floating-gates is huge indeed and far from being exploited by the presented design. Other groups have developed devices that are able to hold precise values for years(Cells used in [122] for instance). In addition to the vast amount of parameters build in the BrainScaleS Wafer-Scale System, it is possible to integrate floating-gates directly into analog circuits instead of using source followers and current mirrors for parameter propagation. This way, completely different neuron designs are possible[39].

Floating-gates remain a very exciting device for an analog designer.

10 Final Remarks and Outlook

This thesis described the development of point-neuron model emulation which is enhanced to a multi-compartment model emulation. By all means, the implementation has been kept as close to common models as possible. However, with multi-compartment emulation, the model-based design approach reached its limits. Here I will conclude my work and give an outlook on the future of the implementations. A detailed discussion of the two circuits can be found in the corresponding discussion chapters.

Single-Compartment Implementation

Direct translation

The single neuron circuit uses OTAs to directly implement the model equations of the adaptive-exponential integrate-and-fire neuron model. A close model correspondence allows mathematical translation between model and circuit if a limited operating range is kept.

Linearity

The main limitation of the circuit is the natural finite linearity of MOSFET differential pairs. It is enhanced in the presented circuit. Nevertheless limitations remain. To overcome this issue circuits not based on differential pairs would be necessary. An example has been given with the transconductor in the multi-compartment chapter. Nevertheless, when sticking to a limited range of hundreds of millivolts, the single neuron can be calibrated to reproduce the AdEx.

Synaptic input

The synaptic input circuit has the most potential for future changes as the linearity of the used resistive element can be improved. In addition the voltage based biasing is no nice solution. Circuits similar to but simpler than the resistive element used for compartment interconnection are candidates. Another improvement might be a stronger output stage of the synaptic input circuit to allow a larger impact when working with single neuron circuits. However the circuit is working with the current implementation. Nevertheless, it needs to be verified in larger network experiments.

Next, I will summarize all circuit changes which have been implemented for the next chip revision HICANN v3.

HICANN v3

The next iteration of the HICANN chip, HICANN v3 is about to be submitted for production in late October 2012. Several small changes concerning the neuron circuits have already been included in schematics and layouts of this chip:

- A dedicated leakage potential has been introduced for the adaptation term to enable input-offset removal by calibration. In HICANN v2, the adaptation and the leakage circuit share a common leakage potential (See 3.5).
- The parameter bias scaling for the spike-frequency adaptation parameter b has been changed to allow larger relative values b/a (See 3.12).

Depending on the time frame, changes concerning the synaptic input, as described above, will be included. Another small change might be a scaling of the bias adjusting the exponential slope Δ_t.

Replacement of floating-gates by MCC circuits?

Another planned major change in HICANN v3 is the replacement of the floating-gate arrays by the new arrays designed for the MCC. This way, the needed silicon area would be reduced drastically. Furthermore, less crosstalk during writing and a better writing precision are expected. However, it is not sure if a replacement without using the multi-compartment neurons would be efficient. In particular as the area freed by the downsized floating-gates is needed by the multi-compartment implementation and cannot be used by other components. Indeed, the new floating-gate cells need to be verified on the MCC before an integration into HICANN v3.

Multi-Compartment Implementation

A competitive circuit

The AdEx implementation has been enhanced to a circuit being capable of emulating multi-compartment neuron models. For this purpose, a new resistive element and a routing network have been designed. Moreover, the parameterizability of the AdEx circuit has been enhanced by local switchable bias current scaling mirrors and a locally switchable membrane capacitor. Consequently, large soma compartments as well as small dendritic compartments can be implemented. Sized differences of compartments are not considered in any implementation from literature so this is an unique feature. Furthermore, the routing capabilities are more flexible than the routing networks from literature.

Improvements concerning single-neuron operation

The multi-compartment circuit has several improvements in comparison to the AdEx implementation. Enhanced parameterization capabilities allow a better miss-match compensation. In particular, the scalable membrane capacitance can be useful. The enhanced parameterization is reached by the introduction of new local memory cells in each compartment. Indeed, in contrast to the values of the floating-gate cells, they can be changed during an experiment to implement structural plasticity effects. Moreover, the resistance of the compartment connection elements is smaller than the connection used for the single-compartment implementation. The spike routing on the membrane itself is a more evolved mechanism than the use of an additional network in the single-neuron implementation. Last but not least, the reset potential is better stabilized now.

Measurements needed and prepared

However, the circuits need verification in silicon. Nevertheless, results from simulations using the complete circuits are promising so far. The second revision of the MCC is in production and expected to be available for measurements in the end of September. Necessary software and circuit boards have already been developed for the first MCC.

After verification, the next step for the multi-compartment implementation would be its integration into a new revision of the HICANN microchip.

Multi- or Single-Compartment?

Replacing the AdEx circuit

The multi-compartment circuit is apparently a huge improvement in comparison to the AdEx circuit. Even if no multi-compartment features are used, the circuits function is enhanced. The functions of the AdEx circuit are a subset of the functions of the new circuit.

Area loss compensated by shrunk floating-gates

The only drawback would be, however, the larger chip area necessary for the enhanced function. Nevertheless, it has been compensated by shrinking the floating-gate array. Consequently, the multi-compartment circuit can replace the AdEx circuit without the need of

additional area. From a circuit point of view, the multi-compartment circuit should clearly be implemented into the HICANN once verified.

Models need to be improved

From a modeling point of view, the question is harder to answer. Although the reduced function when using a single compartment model is apparent, the biologically realistic implementation of a multi-compartment model is disputable. In particular, the propagation of action potentials in the dendrite is questionable in a compartmental model (See Section 6.1). Furthermore, the unknown distribution of ion channels on the dendrite and the complexity of compartmental models raise the problem of over fitting. Nevertheless, a compartmental model is the most realistic reference to a biological neuron at the moment. It is much more realistic than a single-compartment model. As multi-compartment emulation is possible in a natural way on a micro-chip, it should be enabled by a system claimed biologically realistic.

Calibration

Calibration nearly finished

To counterbalance impairments caused by miss-match a calibration of the presented circuits is possible and necessary. These calibrations are being developed by Marc-Olivier Schwartz in his dissertation [69]. Currently, all parts of the AdEx implementation can be calibrated in circuit simulations. Except for the method for the exponential term, all calibration methods have been verified on real hardware[85].

Enhancements for multi-compartment circuits in development

Some efforts for enhancing the single-compartment calibration on a multi-compartment calibration have already been taken. In particular, a method for calibrating the resistive element has been developed[85].

Network Operation

Scalable network operation necessary

A single neuron does not make a network. The designed neuron circuits are only reasonable if they are integrated in a scalable system implementing a plastic network. Indeed, this scalable system is the BWS with its analog microchip HICANN.

BWS infrastructure close to network operation

For the implementation and usability of a complex system like the BWS, however, elaborate hardware and software structures beyond neurons are necessary[1]. At the current status of the system, larger network experiments are only possible using low level hardware testing software as done within this thesis. This approach, however, is very inefficient and therefore hardly carried out. Nevertheless, the current system software is close to be able to implement larger networks. Calibrated network operation will be the baptism of fire for the implemented circuits.

Horizon

Within the BrainScaleS project, the HMF will be assembled and first experiments will be carried out. This will allow for experiments with up to 1.2 million neurons in a network. This system will enable completely new neuroscientific experiments and probably enhance the comprehension of the brain.

On a longer time frame, however, even larger system appear on the horizon. Plans for the Human Brain Project[123] involve systems of a scale of 10.000 BWS. This would correspond to 2 billion neurons if a similar architecture is used, which competes with about 12 -15 billion cells in the human cerebrum[124]. Brain-like computing seems to be near at hand.

[1] This is a fact, most other neuron implementations not caring about scaling seem to forget. There is more than neurons.

By all means, the respect for the human brain, or brains at all, must remain. The complete human brain contains about 85 billion cells[124] and those cells are complex spacial structures whose computational function is not completely understood. In addition about 70 billion of the 85 billion are so called cerebellar granule cells maintaining input counts orders of magnitude beyond the capabilities of the BWS.

Abstraction and complexity reduction might lead to brain-like numbers. However a similar function might not be reached. Nevertheless it is always one step further to a full comprehension of the brain which is still a far but important and exciting way to go.

List of Abbreviations

AC	Alternating current
AdEx	Adaptive Exponential Integrate-and-Fire Neuron Model
AMPA	α-amino-3-hydroxy-5-methyl-4-isoxalone propionic acid, an excitatory neurotransmitter
ASIC	Application Specific Integrated Circuit. Custom made chips to solve special problems.
BWS	BrainScaleS Wafer-Scale System. Neuromorphic system implementing up-to 200 000 AdEx neurons using wafer-scale integration techniques.
DAC	Digital-to-Analog Converter
DAC	Digital-to-analog converter
DC	Direct current
DNC	Digital Network Chip. Chip responsible for action potential event transportation.
FPGA	Field Programmable Gate Array, a special microchip with programmable logic cells.
GABA	γ-aminobutyric acid, an inhibitory neurotransmitter
HHM	Hodgkin Huxley Model
HICANN	High Input-Count Analog Neural Network. The analog ASIC of the BWS.
HMF	Hybrid Multi-Scale Computing Facility
L1	Layer 1 communication. Digital spike event transportation on the wafer via serial buses.
L2	Layer 2 communication. Digital spike event transport in a packet based network through DNC and FPGA.
lIaF	leaky Integrate-and-Fire neuron model
MOSFET	Metal-oxide-semiconductor field-effect transistor. The transistor type used within this thesis.
MPW	Multi-Project Wafer. A reticle is shared among different projects to save prototyping costs.
NL1	Interface creating digital spike events from the fire pulses created by a neuron
NMDA	N-methyl-D-aspartate, an excitatory neurotransmitter
OCP	Open Core Protocol. Bus standard used in the digital core of the HICANN.
OTA	Operational Transconductance Amplifier
PCB	Printed Circuit Board
PMOS,NMOS	P respectively n-type metal-oxide-semiconductor device
PSP	Postsynaptic potential. Voltage response of the postsynaptic neuron on synaptic input
SEB	System Emulator Board
STDP	Spike-timing dependent plasticity
VLSI	Very Large Scale Integration

Bibliography

[1] H. Jaeger, W. Maass, and J. Principe, "Special issue on echo state networks and liquid state machines," *Neural Networks*, vol. 20, no. 3, pp. 287–289, Apr. 2007.

[2] Wikimedia commons, "A database of 13,455,000 freely usable media files to which anyone can contribute," commons.wikimedia.org, 2012.

[3] C. A. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, pp. 1629–1636, 1990.

[4] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watsonx, *Molecular Biology of the Cell, third edition.* Garland Publishing, Inc, 1994.

[5] W. Gerstner and W. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity.* Cambridge University Press, 2002.

[6] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons.* Oxford University Press, 1999.

[7] 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*, vol. 94, pp. 719–723, Jan. 1997.

[8] H. Markram, Y. Wang, and M. Tsodyks, "Differential signaling via the same axon of neocortical pyramidal neurons." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, pp. 5323–5328, Apr. 1998. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/9560274

[9] D. Hebb, *The Organization of Behavior: A Neuropsychological Theory.* Taylor & Francis, 2002.

[10] 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*, vol. 18, no. 24, pp. 10 464–10 472, Dec. 1998. [Online]. Available: http://www.jneurosci.org/content/18/24/10464.abstract

[11] E. M. Izhikevich, "Which Model to Use for Cortical Spiking Neurons?" *IEEE Transactions on Neural Networks*, vol. 15, pp. 1063–1070, 2004. [Online]. Available: http://www.izhikevich.org/publications/whichmod.htm

[12] R. Naud, N. Marcille, C. Clopath, and W. Gerstner, "Firing patterns in the adaptive exponential integrate-and-fire model," *Biological Cybernetics*, vol. 99, no. 4, pp. 335–347, Nov 2008. [Online]. Available: http://dx.doi.org/10.1007/s00422-008-0264-7

[13] G. W. Gross, "Multielectrode arrays," *Scholarpedia*, vol. 6, no. 2, p. 5749, 2011.

[14] B. Sakmann and E. Neher, Eds., *Single-channel recording.* Plenum press, 1995.

[15] H. Tsutsui, S. Karasawa, Y. Okamura, and A. Miyawaki, "Improving membrane voltage measurements using fret with new fluorescent proteins," *Nature methods*, vol. 5, no. 8, pp. 683–685, 2008.

[16] M. Diesmann and M.-O. Gewaltig, "NEST: An environment for neural systems simulations," in *Forschung und wisschenschaftliches Rechnen, Beiträge zum Heinz-Billing-Preis 2001*, ser. GWDG-Bericht, T. Plesser and V. Macho, Eds. Göttingen: Ges. für Wiss. Datenverarbeitung, 2002, vol. 58, pp. 43–70.

[17] R. Brette, M. Rudolph, T. Carnevale, M. Hines, D. Beeman, J. M. Bower, M. Diesmann, A. Morrison, P. H. Goodman, F. C. Harris Jr, M. Zirpe, T. Natschlager, D. Pecevski, B. Ermentrout, M. Djurfeldt, A. Lansner, O. Rochel, T. Vieville, E. Muller, A. P. Davison, S. E. Boustani, and A. Destexhe, "Simulation of networks of spiking neurons: A review of tools and strategies," *Journal of Computational Neuroscience*, vol. 3, no. 23, pp. 349–98, December 2006.

[18] A. Destexhe, M. Rudolph, and D. Pare, "The high-conductance state of neocortical neurons in vivo," *Nature Reviews Neuroscience*, vol. 4, pp. 739–751, 2003.

[19] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve." *J Physiol*, vol. 117, no. 4, pp. 500–544, August 1952. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12991237

[20] R. Stein, "Some Models of Neuronal Variability," *Biophysical Journal*, vol. 7, no. 1, pp. 37–68, Jan. 1967. [Online]. Available: http://dx.doi.org/10.1016/S0006-3495(67)86574-3

[21] E. M. Izhikevich, "Simple Model of Spiking Neurons," *IEEE Transactions on Neural Networks*, vol. 14, pp. 1569–1572, 2003. [Online]. Available: http://www.izhikevich.org/publications/spikes.htm

[22] R. Brette and W. Gerstner, "Adaptive exponential integrate-and-fire model as an effective description of neuronal activity," *J. Neurophysiol.*, vol. 94, pp. 3637 – 3642, 2005.

[23] K. R. Laker and W. M. C. Sansen, *Design of Analog Integrated Circuits and Systems*. McGraw-Hill,Inc, 1994.

[24] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbrück, and R. Douglas, *Analog VLSI: Circuits and principles*. The MIT press, 2002.

[25] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, "Analysis and design of analog integrated circuits, fourth edition." John Wiley & Sons, 2001.

[26] W. M. C. Sansen, *Analog Design Essentials (The International Series in Engineering and Computer Science)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[27] SPICE, "Website," http://bwrc.eecs.berkeley.edu/classes/icbook/spice/, 2012.

[28] I. Cadence Design Systems, "Virtuoso multi-mode simulation," www.cadence.com, 2012.

[29] M. Djurfeldt, M. Lundqvist, C. Johansson, M. Rehn, O. Ekeberg, and A. Lansner, "Brain-scale simulation of the neocortex on the ibm blue gene/l supercomputer," *IBM Journal of Research and Development*, vol. 52, no. 1.2, pp. 31–41, January 2008.

[30] EPFL and IBM, "Blue brain project," Lausanne, 2008. [Online]. Available: http://bluebrain.epfl.ch/

[31] S. B. Furber, D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple, and A. D. Brown, "Overview of the spinnaker system architecture," *IEEE Transactions on Computers*, vol. 99, no. PrePrints, 2012.

[32] J. Ragazzini, R. Randall, and F. Russell, "Analysis of problems in dynamics by electronic circuits," *Proceedings of the IRE*, vol. 35, no. 5, pp. 444–452, 1947.

[33] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison Wesley, 1989.

[34] C. A. Mead and M. A. Mahowald, "A silicon model of early visual processing," *Neural Networks*, vol. 1, no. 1, pp. 91–97, 1988.

[35] T. Delbrück and S. C. Liu, "A silicon early visual system as a model animal." *Vision Res*, vol. 44, no. 17, pp. 2083–2089, 2004.

[36] R. Lyon and C. Mead, "An analog electronic cochlea," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 7, pp. 1119–1134, 1988.

[37] V. Chan, S. Liu, and A. van Schaik, "Aer ear: A matched silicon cochlea pair with address event representation interface," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 54, no. 1, pp. 48–59, 2007.

[38] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, no. 6354, pp. 515–518, Dec 1991. [Online]. Available: http://dx.doi.org/10.1038/354515a0

[39] E. Farquhar and P. Hasler, "A bio-physically inspired silicon neuron," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 52, no. 3, pp. 477 – 488, march 2005.

[40] 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.

[41] J. Arthur and K. Boahen, "Silicon neurons that inhibit to synchronize," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 27-30 2007, pp. 1186 –1186.

[42] J. H. Wijekoon and P. Dudek, "Compact silicon neuron circuit with spiking and bursting behaviour," *Neural Networks*, vol. 21, no. 2-3, pp. 524 – 534, 2008, advances in Neural Networks Research: IJCNN '07, 2007 International Joint Conference on Neural Networks IJCNN '07. [Online]. Available: http://www.sciencedirect.com/science/article/B6T08-4RFSCV3-5/2/c005fcc0c2482bf724210a079932484e

[43] S. Mihalas and E. Niebur, "A generalized linear integrate-and-fire neural model produces diverse spiking behaviors," *Neural computation*, vol. 21, no. 3, pp. 704–718, 2009.

[44] S. Millner, A. Grübl, K. Meier, J. Schemmel, and M.-O. Schwartz, "A VLSI implementation of the adaptive exponential integrate-and-fire neuron model," in *Advances in Neural Information Processing Systems 23*, J. Lafferty *et al.*, Eds., 2010, pp. 1642–1650.

[45] G. Indiveri, F. Stefanini, and E. Chicca, "Spike-based learning with a generalized integrate and fire silicon neuron," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 30 2010-june 2 2010, pp. 1951 –1954.

[46] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Frontiers in Neuroscience*, vol. 5, no. 0, 2011. [Online]. Available: http://www.frontiersin.org/Journal/Abstract.aspx?s=755&name=neuromorphicengineering&ART_DOI=10.3389/fnins.2011.00073

[47] D. Fasnacht, A. Whatley, and G. Indiveri, "A serial communication infrastructure for multi-chip address event systems," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, may 2008, pp. 648 –651.

[48] G. Indiveri, "Modeling selective attention using a neuromorphic analog vlsi device," *Neural computation*, vol. 12, no. 12, pp. 2857–2880, 2000.

[49] S. Scholze, S. Schiefer, J. Partzsch, S. Hartmann, C. G. Mayr, S. Höppner, H. Eisenreich, S. Henker, B. Vogginger, and R. Schüffny, "VLSI implementation of a 2.8GEvent/s packet based AER interface with routing and event sorting functionality," *Frontiers in Neuromorphic Engineering*, vol. 5, no. 117, pp. 1–13, 2011.

[50] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proceedings of the 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2010, pp. 1947–1950.

[51] F. Folowosele, A. Harrison, A. Cassidy, A. Andreou, R. Etienne-Cummings, S. Mihalas, E. Niebur, and T. Hamilton, "A switched capacitor implementation of the generalized linear integrate-and-fire neuron," in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on.* IEEE, 2009, pp. 2149–2152.

[52] C. Bartolozzi, S. Mitra, and G. Indiveri, "An ultra low power current-mode filter for neuromorphic systems and biomedical signal processing," in *Biomedical Circuits and Systems Conference, 2006. BioCAS 2006. IEEE.* IEEE, 2006, pp. 130–133.

[53] D. Frey, "Future implications of the log domain paradigm," in *Circuits, Devices and Systems, IEE Proceedings-*, vol. 147, no. 1. IET, 2000, pp. 65–72.

[54] BrainScaleS, "Research," http://brainscales.kip.uni-heidelberg.de/public/index.html, 2012.

[55] 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.

[56] J. Fieres, J. Schemmel, and K. Meier, "Realizing biological spiking network models in a configurable wafer-scale hardware system," in *Proceedings of the 2008 International Joint Conference on Neural Networks (IJCNN)*, 2008.

[57] S. Hartmann, S. Schiefer, S. Scholze, J. Partzsch, C. Mayr, S. Henker, and R. Schuffny, "Highly integrated packet-based aer communication infrastructure with 3gevent/s throughput," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, dec. 2010, pp. 950 –953.

[58] S. Scholze, H. Eisenreich, S. Höppner, G. Ellguth, S. Henker, M. Ander, S. Hänzsche, J. Partzsch, C. Mayr, and R. Schüffny, "A 32 GBit/s communication SoC for a waferscale neuromorphic system," *Integration, the VLSI Journal*, 2011, in press.

[59] M. Güttler, "Konzeptoptimierung und Entwicklung einer hochintegrierten Leiterplatte," Diploma thesis (German), University of Heidelberg, HD-KIP-10-68, 2010.

[60] H. Zoglauer, "Entwicklung und testergebnisse eines prototypensystems für die wafer-scale-integration," Diploma thesis (German), University of Heidelberg, HD-KIP-09-28, 2009.

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

[62] J. Schemmel, D. Brüderle, K. Meier, and B. Ostendorf, "Modeling synaptic plasticity within networks of highly accelerated I&F neurons," in *Proceedings of the 2007 IEEE International Symposium on Circuits and Systems (ISCAS).* IEEE Press, 2007, pp. 3367–3370.

[63] "Open core protocol specification 2.2," 2008. [Online]. Available: http://www.ocpip.org/home

[64] A. Grübl, "VLSI implementation of a spiking neural network," Ph.D. dissertation, Ruprecht-Karls-University, Heidelberg, 2007, document No. HD-KIP 07-10. [Online]. Available: http://www.kip.uni-heidelberg.de/Veroeffentlichungen/details.php?id=1788

[65] FACETS, "Fast Analog Computing with Emergent Transient States – project website," http://www.facets-project.org, 2010.

[66] A. Destexhe, "Conductance-based integrate-and-fire models," *Neural Comput.*, vol. 9, no. 3, pp. 503–514, 1997.

[67] M.-O. Gewaltig and M. Diesmann, "NEST (NEural Simulation Tool)," *Scholarpedia*, vol. 2, no. 4, p. 1430, 2007.

[68] J. Schemmel, "personal communication," Kirchhoff Institut für Physik, Universität Heidelberg, Deutschland.

[69] M.-O. Schwartz, *PhD thesis*, University of Heidelberg, in preparation, 2012.

[70] A. Kononov, "Testing of an analog neuromorphic network chip," Diploma thesis (English), University of Heidelberg, HD-KIP-11-83, 2011.

[71] C. Hu, W. Liu, and X. Jin, *The BSIM3v3.2 MOSFET Model*, Dec 1998.

[72] J. Yeomans, "The absolute refractory periods of self-stimulation neurons," *Physiology & Behavior*, vol. 22, no. 5, pp. 911–919, 1979.

[73] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons," *PLoS Computational Biology*, vol. 7, no. 11, p. e1002211, 2011.

[74] M. A. Petrovici, *PhD thesis*, University of Heidelberg, in preparation, 2012.

[75] B. Bean, "The action potential in mammalian central neurons," *Nature Reviews Neuroscience*, vol. 8, no. 6, pp. 451–465, 2007.

[76] S. Millner, "An integrated operational amplifier for a large scale neuromorphic system," Diploma thesis, University of Heidelberg, HD-KIP-08-19, 2008.

[77] A. Destexhe, D. Contreras, and M. Steriade, "Mechanisms underlying the synchronizing action of corticothalamic feedback through inhibition of thalamic relay cells," *Journal of Neurophysiology*, vol. 79, pp. 999–1016, 1998.

[78] M. Pospischil, M. Toledo-Rodriguez, C. Monier, Z. Piwkowska, T. Bal, Y. Frégnac, H. Markram, and A. Destexhe, "Minimal hodgkin–huxley type models for different classes of cortical and thalamic neurons," *Biological Cybernetics*, vol. 99, no. 4, pp. 427–441, Nov 2008. [Online]. Available: http://dx.doi.org/10.1007/s00422-008-0263-8

[79] A. Destexhe, M. Rudolph, J. M. Fellous, and T. J. Sejnowski, "Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons." *Neuroscience*, vol. 107, no. 1, pp. 13–24, 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11744242

[80] "IEEE Standard Test Access Port and Boundary-Scan Architecture," *IEEE Std 1149.1-2001*, pp. i –200, 2001.

[81] NXP Semiconductors, "I2C-bus specification and user manual," 2012.

[82] M. Hock, "Test of components for a wafer-scale neuromorphic hardware system," Diploma thesis, University of Heidelberg, HD-KIP-09-37, http://www.kip.uni-heidelberg.de/Veroeffentlichungen/details.php?id=1935, 2009.

[83] A. Grübl, "personal communication," Kirchhoff Institut for Physics, University of Heidelberg, Germany, 2012.

[84] A. Kononov, "personal communication," Kirchhoff Institut für Physik, Universität Heidelberg, Deutschland, 2012.

[85] M.-O. Schwartz, "personal communication," Kirchhoff Institut for Physics, University of Heidelberg, Germany, 2012.

[86] D. McCormick, Z. Wang, and J. Huguenard, "Neurotransmitter control of neocortical neuronal activity and excitability," *Cerebral Cortex*, vol. 3, pp. 387–398., 1993.

[87] S. Druckmann, Y. Banitt, A. Gidon, F. Schürmann, H. Markram, and I. Segev, "A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data," *Front Neurosci*, vol. 1, no. 1, pp. 7–18, Nov 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2570085/?report=abstract

[88] H. Markram, M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu, "Interneurons of the neocortical inhibitory system," *Nat Rev Neurosci*, vol. 5, no. 10, pp. 793–807, Oct 2004. [Online]. Available: http://dx.doi.org/10.1038/nrn1519

[89] LeCroy, "Enhanced resolution," LeCroy Corporation, 700 Chestnut Ridge Road, Chestnut Ridge, NY 10977-6499, Tech. Rep. AN006A. [Online]. Available: http://lecroygmbh.com

[90] J. Schemmel, A. Grül, A. Kononov, K. Meier, S. Millner, M.-O. Schwartz, S. Scholze, S. Schiefer, S. Hartmann, J. Partsch, C. Mayer, and R. Schüffny, "Live demonstration: A scaled-down version of the BrainScaleS wafer-scale neuromorphic system," 2012.

[91] 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.*, vol. 2, no. 11, 2008.

[92] D. Goodman and R. Brette, "Brian: a simulator for spiking neural networks in Python," *Front. Neuroinform.*, vol. 2, no. 5, 2008.

[93] W. Rall, "Branching dendritic trees and motoneuron membrane resistivity," *Experimental neurology*, vol. 1, no. 5, pp. 491–527, 1959.

[94] ——, "Electrophysiology of a dendritic neuron model," *Biophysical journal*, vol. 2, no. 2 Pt 2, p. 145, 1962.

[95] C. Koch and I. Segev, "The role of single neurons in information processing." *Nat Neurosci*, vol. 3 Suppl, pp. 1171–1177, Nov. 2000.

[96] H. Agmon-Snir, C. Carr, and J. Rinzel, "The role of dendrites in auditory coincidence detection," *Nature*, vol. 393, no. 6682, pp. 268–272, 1998.

[97] M. London and M. Häusser, "Dendritic computation," *Annu. Rev. Neurosci.*, vol. 28, pp. 503–532, 2005.

[98] R. Silver, "Neuronal arithmetic," *Nature Reviews Neuroscience*, vol. 11, no. 7, pp. 474–489, 2010.

[99] C. Clopath, R. Jolivet, A. Rauch, H. Lüscher, and W. Gerstner, "Predicting neuronal activity with simple models of the threshold type: Adaptive exponential integrate-and-fire model with two compartments," *Neurocomputing*, vol. 70, no. 10, pp. 1668–1673, 2007.

[100] A. Destexhe, M. Neubig, D. Ulrich, and J. R. Huguenard, "Dendritic low-threshold calcium currents in thalamic realy cells," *Journal of Neuroscience*, vol. 18, pp. 3574–3588, 1998.

[101] Ö. Ekeberg, P. Wallén, A. Lansner, H. Travén, L. Brodin, and S. Grillner, "A computer based model for realistic simulations of neural networks," *Biological Cybernetics*, vol. 65, pp. 81–90, 1991, 10.1007/BF00202382. [Online]. Available: http://dx.doi.org/10.1007/BF00202382

[102] E. Hendrickson, J. Edgerton, and D. Jaeger, "The capabilities and limitations of conductance-based compartmental neuron models with reduced branched or unbranched morphologies and active dendrites," *Journal of computational neuroscience*, vol. 30, no. 2, pp. 301–321, 2011.

[103] C. Rasche and R. Douglas, "Forward- and backpropagation in a silicon dendrite," *Neural Networks, IEEE Transactions on*, vol. 12, no. 2, pp. 386 –393, mar 2001.

[104] J. G. Elias and D. P. M. Northmore, *Building silicon nervous systems with dendritic tree neuromorphs*, 1st ed. The MIT Press, Nov. 1998, ch. 5, pp. 135–156.

[105] E. Farquhar, D. Abramson, and P. Hasler, "A reconfigurable bidirectional active 2 dimensional dendrite model," in *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, vol. 1, may 2004, pp. I–313 – I–316 Vol.1.

[106] E. Farquhar, C. Gordon, and P. Hasler, "A field programmable neural array," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, may 2006, pp. 4 pp. –4117.

[107] P. Hasler, S. Kozoil, E. Farquhar, and A. Basu, "Transistor channel dendrites implementing hmm classifiers," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, may 2007, pp. 3359 –3362.

[108] J. Arthur and K. Boahen, "Recurrently connected silicon neurons with active dendrites for one-shot learning," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 3, july 2004, pp. 1699 – 1704 vol.3.

[109] Y. Wang and S.-C. Liu, "A two-dimensional configurable active silicon dendritic neuron array," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 58, no. 9, pp. 2159 –2171, sept. 2011.

[110] ——, "Multilayer processing of spatiotemporal spike patterns in a neuron with active dendrites," *Neural Computation*, vol. 28, pp. 2086–2112, 2011.

[111] D. Johns and K. Martin, *Analog integrated Circuit*. John Wiley and Sons, Inc, 1997.

[112] A. Carusone and D. Johns, "A 5th order gm-c filter in 0.25 mu;m cmos with digitally programmable poles and zeroes," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, vol. 4, 2002, pp. IV–635 – IV–638 vol.4.

[113] Y. Deng, S. Chakrabartty, and G. Cauwenberghs, "Three-decade programmable fully differential linear ota," in *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, vol. 1, may 2004, pp. I – 697–700 Vol.1.

[114] A. Morrison, M. Diesmann, and W. Gerstner, "Phenomenological models of synaptic plasticity based on spike timing," *Biological Cybernetics*, vol. 98, no. 6, pp. 459–478, June 2008.

[115] S. Millner, A. Hartel, J. Schemmel, and K. Meier, "Towards biologically realistic multi-compartment neuron model emulation in analog VLSI," in *Proceedings ESANN 2012*, 2012.

[116] J.-P. Loock, "Evaluierung eines floating gate analogspeichers für neuronale netze in single-poly umc 180nm CMOS-prozess," Diploma thesis (English), University of Heidelberg, HD-KIP-06-47, 2006.

[117] A. Srowig, J.-P. Loock, K. Meier, J. Schemmel, H. Eisenreich, G. Ellguth, and R. Schüffny, "Analog floating gate memory in a 0.18 μm single-poly CMOS process," *FACETS internal documentation*, 2007.

[118] P. Pavan, L. Larcher, and A. Marmiroli, *Floating gate devices.* Boston [u.a.]: Kluwer Academic, 2004, includes bibliographical references and index.

[119] Synopsis, Inc., "Write-port, dual-read-port ram (latch-based),dw_ram_2r_w_a_lat," DesignWare Building Block IP, 2002.

[120] C. E. Cummings, "Synthesizeable finite state machine design techniques using the new SystemVerilog 3.0 enhancements," in *Synopsis User Group*, 2003.

[121] A. Grübl, S. Millner, and J. Schemmel, "Design the final ASIC for the wafer-scale system," FACETS Deliverable D7-7, 2010, university Heidelberg.

[122] A. Basu, S. Brink, C. Schlottmann, S. Ramakrishnan, C. Petre, S. Koziol, F. Baskaya, C. Twigg, and P. Hasler, "A floating-gate-based field-programmable analog array," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 9, pp. 1781 –1794, sept. 2010.

[123] H. Markram, "The human brain project," *Scientific American*, vol. 306, no. 6, pp. 50–55, 2012.

[124] R. Williams and K. Herrup, "The control of neuron number," *Annual Review of Neuroscience*, vol. 11, no. 1, pp. 423–453, 1988.

Acknowledgements

Finally, I want to acknowledge all the people who supported me in my work. In particular, I express my gratitude to:

- Prof. Dr. Karlheinz Meier for supervising this dissertation and his strongly motivating attitude in each conversation.
- Dr. Johannes Schemmel for his inspirational and critical feedback.
- Prof. Dr. Peter Fischer for being the second referee of this thesis.
- My former student Alexander Kononov for extensively measuring the floating-gate circuit for his diploma thesis and for being always very helpful in any concern.
- Andreas Hartel for his collaboration in the design of the MCC. With out his work on the digital code and the back-end in particular, the chip would not have been possible. Furthermore I would like to thank him for all the kilometers we left on the road while running.
- Marc-Olivier, Alexander Kononov, Paul Müller, Andreas Hartel, Simon Friedmann and Sven Schrader for proofreading parts of this thesis. In particular, I want to thank Marc-Olivier Schwartz who fought his way through all core chapters.
- My lovely Vera for supporting me in any case and for inspiring me to sometimes have a brake in months without weekends or holidays.
- Marc-Olivier Schwartz for doing all the calibration work necessary for the usage of the neuron circuits.
- Andreas Grübl and Erik Müller for being always supportive when any problem occurred.
- Mihai Petrovici for organizing lots of social events keeping the group together besides working. His cheerful attitude is a great asset for the group's atmosphere.
- Ralf Achenbach for spending frustrating weeks trying to find a way to bond HICANN v1 and finally succeeding.
- All members of the vision group not mentioned above. You guys are awesome.
- My parents.
- The developers of the open sources software tools used for the creation of this thesis. In particular the developers of Xcircuit, Inkscape, Gimp, Vim, Latex, Gnuplot and Matplotlib.