Karlsruhe Institute of Technology

Mixed-Signal Circuit Implementation of Spiking Neuron Models

zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN

der Fakultät für Elektrotechnik und Informationstechnik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

SYED AHMED AAMIR

Referent: Prof. Dr. Marc Weber
Korreferent: Prof. Dr. Karlheinz Meier
Tag der mündlichen Prüfung: 2018-01-10

Abstract

Inspired by the nervous system, analog neuromorphic systems integrate computational models of neural elements to capture the rich temporal dynamics of the neuronal membrane. For the development of the BrainScaleS neuromorphic hardware, this thesis implements spiking neuron models with accelerated dynamics in a 65 nm CMOS process. Compact, low-power, highly-tunable continuous-time analog circuits are developed and characterized over three prototype chips. The first design emulates a Leaky Integrate-and-Fire (LIF) model implemented as an array of neurons. The measured results demonstrate the availability of a vast range of time constants and a one-to-one correspondence with the dynamics of the mathematical model. The LIF neuron circuit is enhanced to the Adaptive-Exponential Integrate-and-Fire (AdEx) model where the circuits for exponential and adaptation are designed. The AdEx implementation results demonstrate a variety of firing patterns typically known from cortical neurons. The neuron circuit will provide the end-users with biologically plausible spiking dynamics and is to be integrated as the fundamental computational element in the second-generation BrainScaleS hardware.

Zusammenfassung

Inspiriert vom Nervensystem spiegeln analoge neuromorphe Systeme die vielseitige Dynamik biologischer Neuronen und Synapsen wider. Für die Weiterentwicklung des neuromorphen Systems „BrainScaleS“ wird in dieser Arbeit ein beschleunigtes, spikendes Neuronenmodell in einem 65-nm-CMOS-Prozess umgesetzt. Kompakte, hoch konfigurierbare, zeitkontinuierliche analoge Schaltungen mit niedrigem Energieverbrauch werden entwickelt und mit Hilfe von drei Prototypen-Chips charakterisiert. Das erste Design emuliert ein Leaky-Integrate-and-Fire-Modell (LIF). Die Messergebnise zeigen, dass ein grosser Bereich von Zeitkonstanten eingestellt werden kann und eine gute Übereinstimmung mit dem mathematischen Modell erreicht wird. Das LIF-Modell wird dann zu einem Adaptive-Exponential-Integrate-and-Fire-Modell (AdEx) erweitert, wofür zusätzliche Schaltungen für den Exponential- und Adaptionsterm implementiert werden. Die Ergebnisse des AdEx-Prototyp-Chips zeigen, dass eine Vielzahl typischer Feuermuster, wie sie bei kortikalen Neuronen beobachtet wurden, mit der Schaltung emuliert werden können. Der in dieser Arbeit entworfene Neuronenschaltkreis wird den Benutzern ein der Biologie ähnliches, spikendes Neuron bieten und als fundamentale Einheit in die zweite Generation des BrainScaleS-Systems eingehen.

Contents

Chapter 1

Introduction: Neurons and Synapses

The human brain is a massively parallel and highly flexible organ consisting of neurons that are interconnected through synaptic connections. Weighing just about three pounds, the brain contains on the order of 10^{11} neurons and 10^{14} synapses [1, 2]. Organized in specialized regions and as ensembles, each neuron within a population receives thousands of synaptic inputs through neighboring cells [3, 4], and communicates using electrical signals called action potentials or spikes. At the system level, the different patterns of interconnections give rise to perceptions and motor actions [5]. The communication between neurons is modified by experience, leading towards learning and development. The foremost goal towards the understanding of the nervous system has been to elucidate, how such neural ensembles compute and lead to the cognitive states and behavior that we experience as individuals.

The digital microprocessor processes information, just like the brain does. The accuracy of numerical calculations it may achieve is 32 or 64 significant figures. In contrast, a neuron signaling in terms of average firing rate, offers at best a few significant figures [6]. Further, while the brain consumes only 20 W of average power [7, 8], similar information processing abilities require orders of magnitude higher power consumption in modern microprocessors [9, 10]. The architecture of the microprocessors have traditionally been based on a so-called *von Neumann* architecture, where the CPU accesses data and program memory using shared resources [11]. Further, they utilize the digital logic gates as elementary primitives and have little to no fault tolerance.

A fundamentally different kind of computing architecture has evolved from the late eighties that takes strong inspiration from the architecture of the nervous system. Termed neuromorphic systems [12, 13], they radically depart from von-Neumann architectures by collocating memory close to the computational elements. They are massively parallel and flexible with high fan-in and fan-out capabilities. They offer a high degree of reconfigurability and are far more energy effi-

cient than conventional computing architectures. Above all, they integrate computational models of neural elements as computational units. They are realized using standard CMOS technologies, allowing integration at a very large scale [12, 14].

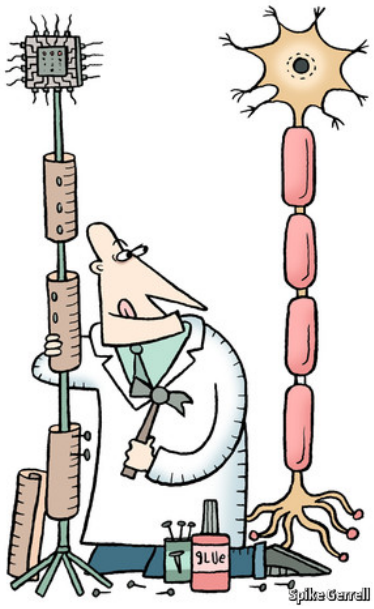


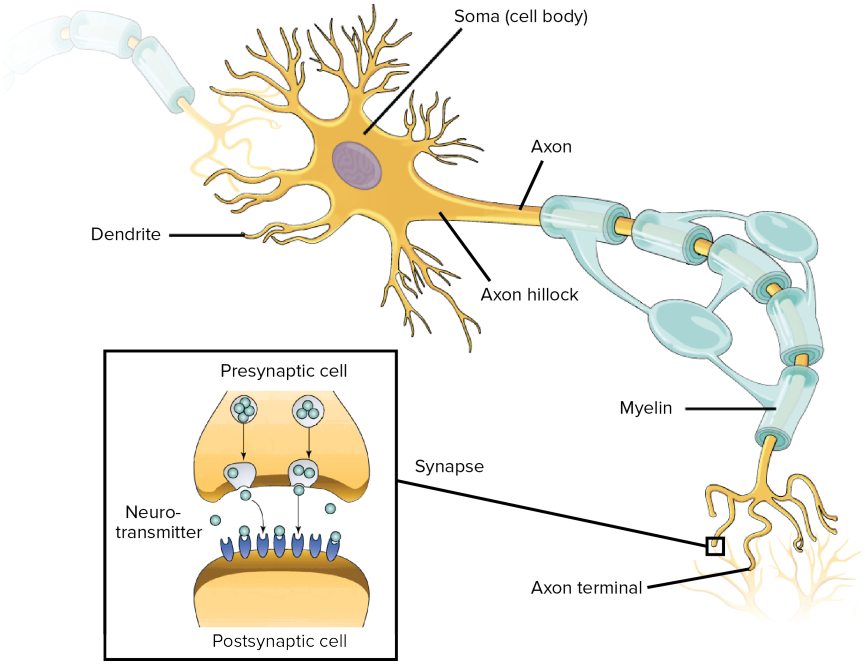
Figure 1.1: An illustration of neuron emulation. Image © Spike Gerrell, published in Economist [15].

This thesis is about the emulation of the neuron circuit – the primary computational element of any neuromorphic substrate. The neuron is designed as a continuous-time processing element that emits a binary event (spike) in analogy to the nervous system. The neuron circuit emulates biophysically inspired spiking neuron models through the use of analog and mixed-signal circuit techniques. It features high tunability, a modular architecture and an implementation realized for large-scale integration. An artist's view of neuron emulation is depicted in Fig. 1.1.

Before proceeding with this emulation, we review the basic architecture and models of neurons and synapses.

1.1 Neurons

The main anatomical and computational units in the nervous system are the neurons. The structure of a typical neuron is depicted in Fig. 1.2. The morphological form and this discrete cell structure was not known until the late 1800's when Ramón y Cajal discovered it with the aid of a staining method developed by Camillo Golgi [16, 17]. The neuron structure can be divided into four distinct regions: the cell body, the dendrites, the axon and the synaptic terminals. At the center is the cell body or *soma* that contains the nucleus. From the soma emerge tree-like branches that are the *dendrites*. These let the cell receive synaptic inputs and integrate them on the cell body. When the potential at the *axon hillock* exceeds a threshold, an action potential is initiated, that propagates down along the *axon*. Axon is a specialized structure that transmits the action potential over a long distance and is typically covered with *myelin sheath*. Myelin is an insulating material that reduces the capacitance between cytoplasm and the extra-cellular fluid. The sheath is interrupted at regular intervals by the *nodes of ranvier* that help regenerate and restore the action potential repeatedly. The axon further divides into fine branches, where it makes contact with other neurons at specialized zones of communication called *synapses* [5]. Fig. 1.2 shows this structure with dendrites at the top and axon terminating with synaptic terminals at the bottom.



.

Figure 1.2: The structure of a typical neuron. Image from OpenStax College, Biology (CC BY 4.0) [18], as modified by [19].

The rapid change in neuron's cell membrane that leads to an action potential

is mediated by ion channels and pumps. The cell membrane that separates the extra-cellular fluid from the interior is a thin phospholipid bilayer. Of the various ion species, the permeable ions responsible for electrical signaling are Na^+, Cl^- and K^+. The distribution of these ions varies across the membrane, and Na^+ and Cl^- have higher concentration outside the membrane, whereas inner cytoplasm has higher concentration of K^+ compared to the outside.

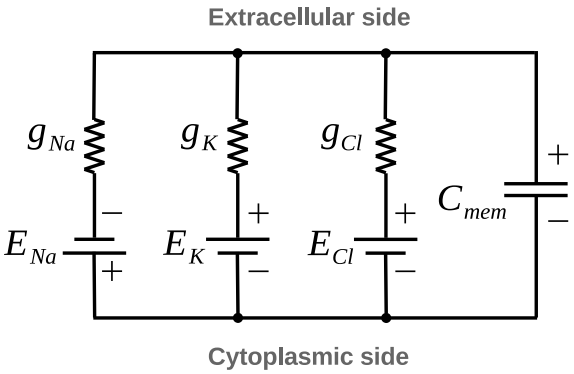


Figure 1.3: An equivalent circuit schematic of the neuronal membrane [5].

The equilibrium potential for any permeable ion present across the membrane can be calculated via the Nernst equation, named after the German chemist Walther Nernst. This is the Nernst potential and given by $E_{\text{rev}} = \frac{-kT}{q} ln \frac{N_{in}}{N_{ex}}$. Where N_{ex} and N_{in} are the external and internal ion densities, k is the Boltzmann constant and T is the absolute temperature. A positive current flows into the cell, if the potential inside the membrane is below the Nernst potential. Conversely, an outward positive current flows if the membrane potential is greater than E_{rev}. Since the direction of current is reversed, this potential difference is also referred to as the *reversal potential* of a specific ionic current. At rest the membrane potential is approximately –70 mV and the cell membrane is said to be in a polarized state. This resting potential is maintained by the sodium-potassium (Na^+/K^+) pump that flushes Na^+ out and takes K^+ in, as well as the membrane's selective permeability to K^+ which leaves behind a net negative charge.

The electrical properties of the neuron membrane help derive an equivalent circuit schematic, shown in Fig. 1.3. The lipid bilayer endows the membrane with an electrical capacitance, whereas the conductances reflect the membrane permeability of a given ion channel. The batteries represent the reversal potential of a particular ion. For example, considering the sodium current, by virtue of Ohm's law, it is expressed as $I_{\text{Na}} = g_{\text{Na}} \cdot (V_{\text{mem}} - E_{\text{Na}})$. The contribution of various ions to the resting membrane potential can be quantified using the Goldman equation [20–22]. Alternatively, by solving the equivalent circuit schematic one can calculate the resting potential of the cell membrane [5].

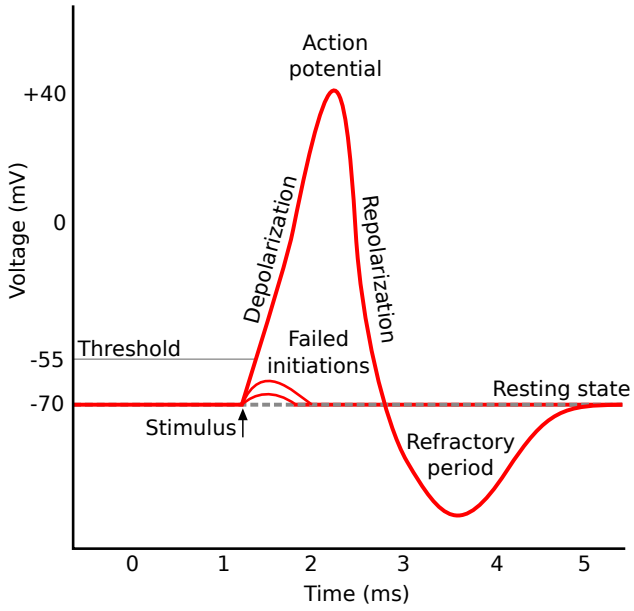


Figure 1.4: A typical shape of an action potential generated in a resting neuronal membrane as a result of input stimulus. Figure taken from [23].

If the membrane is charged more negatively than its resting potential, it is said to be *hyperpolarized* and its input electrical signal is *inhibitory*. Conversely, an *excitatory* input *depolarizes* the membrane increasing the likelihood of evoking an action potential. When the cell membrane is depolarized, resulting in a rise from resting potential, it is more permeable to Na^+ than to K^+. The resulting influx of positively charged Na^+ neutralizes the negative charge inside, causing a sharp rise in membrane to about +40 mV. This is the action potential, which lasts a short interval of approximately 1 ms and is drawn in Fig. 1.4. The membrane then returns to its resting state and back to the higher permeability to K^+.

The foregoing description briefly outlines the passive electrical properties of the neuronal membrane. A more comprehensive discussion can be found in [5,24].

1.2 Neuron Models

The computational models of neurons are typically classified into two categories. The phenomenological models that include the threshold models [25], and the biophysical models. The McCulloch-Pitts neuron [26] and the Perceptron [27] are examples of the first, since they only model the I/O behavior of the neuron unit using simple math. The biophysical models on the other hand, describe ion channels with the goal to model the electrophysiological behavior. Hodgkin and Huxley model [28] and others [29, 30] are prime examples. The threshold models or integrate-and-fire models [31–36] are a good compromise between the complexity of biophysical models and the simplification of linear threshold units. They are

mathematically tractable [37], easier to tune than the biophysical models and lend themselves well for hardware implementations [38].

This section reviews the dynamics of the leaky integrate-and-fire model as well as its extended two variable variant, the Adaptive Exponential Integrate-and-Fire (AdEx) model.

1.2.1 The Leaky Integrate and Fire Model

The Leaky Integrate-and-Fire (LIF) model [31, 32, 39] is one of the simplest computational models, that is described with two separate components. First, is the subthreshold behavior of the cell membrane described as

$$C\frac{\mathrm{d}V_{\mathrm{mem}}}{\mathrm{d}t} = -g_{\mathrm{leak}} \cdot (V_{\mathrm{mem}} - V_{\mathrm{leak}}) + I \tag{1.1}$$

where C_{mem} is the membrane capacitor that integrates the input current I, and $g_{\mathrm{leak}} \cdot (V_{\mathrm{mem}} - V_{\mathrm{leak}})$ is the current that leaks out of the membrane, making it an imperfect integrator. V_{leak} models the potential towards which the membrane leaks away in the absence of input activity, while g_{leak} is the leak conductance. The input current I is the sum of synaptic excitatory (I_{synExc}) and inhibitory (I_{synInh}) currents as well as externally injected current I_{stim}, such that, $I = I_{\mathrm{synExc}} + I_{\mathrm{synInh}} + I_{\mathrm{stim}}$. This LIF model is visualized in Fig. 1.5.

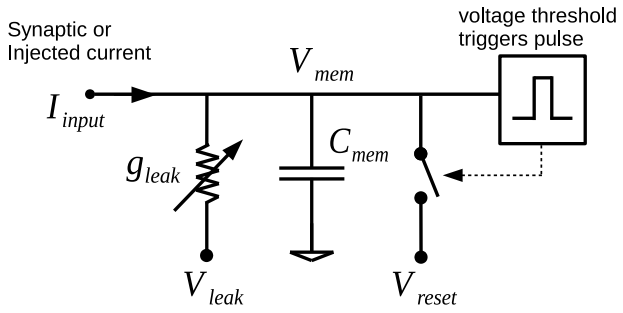


Figure 1.5: The ideal model of the Leaky Integrate-and-Fire model. Adapted from [40].

The second part of the model describes the output event generation. When the membrane potential described by Eq. 1.1 reaches a well-defined threshold V_{thresh}, the neuron outputs a binary signal and the membrane V_{mem} is reset to a potential V_{reset}. Since LIF model does not incorporate the detailed time-course of an action potential, information is contained in the presence or absence of an event output that marks the spike occurrence. This digital output pulse marks the model's all-or-none spike events, followed by a relative refractory period – a time duration when the neuron is less excitable.

1.2.2 Adaptive Exponential Integrate and Fire Model

The AdEx neuron model [36] adds an exponential activation mechanism and a second variable for adaptation to describe the membrane dynamics of the integrate-and-fire neuron. The model builds upon the exponential integrate-and-fire model [34] and the two variable Izhikevich model [41]. Along with spike-triggered adaptation, the model can reproduce electrophysiologically known firing patterns such as the fast and regular spiking, phasic and tonic bursting, post inhibitory spiking and bursting, delayed spike and burst initiation [42]. It is defined by the set of equations

$$\begin{aligned} C_{\text{mem}}\frac{\mathrm{d}V_{\text{mem}}}{\mathrm{d}t} &= I - w - g_{\text{leak}}(V_{\text{mem}} - V_{\text{leak}}) \\ &\quad + g_{\text{leak}}\Delta_T \exp\left(\frac{V_{\text{mem}} - V_T}{\Delta_T}\right) \end{aligned} \tag{1.2}$$

$$\tau_{\text{w}}\frac{\mathrm{d}w}{\mathrm{d}t} = a(V_{\text{mem}} - V_{\text{leak}}) - w \tag{1.3}$$

where V_{mem} is the membrane potential, C_{mem} is the membrane capacitance, g_{leak} is the leak conductance, V_{leak} is the resting or leak potential, Δ_{T} is the slope threshold and V_{T} is the effective threshold potential. I the input current, w is the adaptation current, and τ_{w} is the adaptation time constant. As the input current pulls the membrane beyond the threshold V_{T}, the exponential non-linearity triggers, leading to

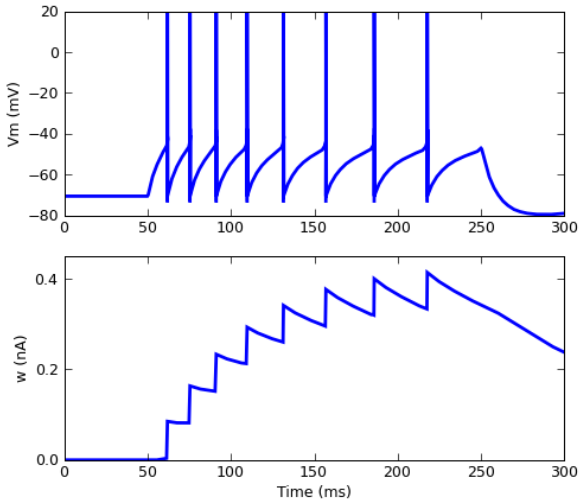


Figure 1.6: The membrane potential (top) in an AdEx model showing exponential spikes coupled with an adapting behavior. The adaptation occurs as a result of the evolution of the second variable w (bottom). Image taken from [43].

the upswing of the action potential. The downswing is replaced by a fixed reset. When spike is triggered the membrane is reset to V_{reset}, while the adaptation variable is increased by an amount b, such that $w \rightarrow w + b$. The second equation describes the evolution of the adaptation current with a decaying time constant τ_{w}. Voltage is coupled to the second adaptation variable w. The update of w to $w + b$ is referred to as spike-triggered adaptation, while the linear coupling of voltage via a is referred to as subthreshold adaptation [43] – since the current is active regardless of a spike. When $a > 0$, the membrane gets increasingly hyperpolarized, leading to a decelerating spike behavior. The coupling with $a < 0$ leads to a depolarizing current and eventually to an accelerating spiking response.

The response to an input stimulus of the AdEx neuron model is shown in Fig. 1.6. The top figure shows the exponentially rising membrane followed by repeated resets. The firing adapts with a decelerating response as a result of hyperpolarizing membrane due to spike-triggered adaptation. The evolution of the variable w is shown in the lower plot.

Out of the nine parameters, five are classified as scaling parameters, while the remaining four are bifurcation parameters. The scaling parameters are C_{mem}, g_{leak}, V_{leak}, Δ_{T} and V_{T}, as they scale the time axis or stretch the state variables [42]. The bifurcation parameters are a, b, τ_{w} and V_{reset}. One can modify these to evoke different spike patterns. The role of bifurcations is best explained in texts for dynamic systems and phase plane analysis, such as [44] and are not covered here. The work from [42] shows at least eight firing patterns known from cortical neurons that can be reproduced using the AdEx model. Fig. 1.7 reproduces these firing patterns along with the phase space representation against each of them. This is a two-dimensional space of variables – membrane voltage and the adaptation current. The nullclines of both variables represent the set of points where their time derivative is zero, i.e., V-nullcline represents $\frac{\text{d}V_{\text{mem}}}{\text{d}t} = 0$ and the w-nullcline are the set of points where $\frac{\text{d}w}{\text{d}t} = 0$. In the figure, the membrane is shown as a blue trace, while the V-nullcline is shown as black dashed curve (prior to input stimulus) and as solid line (after current stimulation). The w-nullcline is shown in green. As the firing trajectories are plotted the blue cross indicates the resting state, while the blue square denotes the sequence of reset values one after the other. The intersection of the two nullclines define stable or unstable fixed points. When a bifurcation occurs these fixed points change, leading the system to show a different (spiking) behavior.

Of the shown patterns tonic spiking occurs when spike-triggered adaptation is not playing any role, i.e., $a = b = 0$; a case like that of the simple LIF model. The AdEx model can produce sharp and broad spike after-potential (SAP). In initial bursting, spiking starts with one or more sharp resets, followed by a broad reset (long low curvature). In delayed acceleration, a stimulus close to the spiking threshold (rheobase) is injected and the value of $a < 0$. This eventually leads the neuron to spike after a time interval. After the initial spike, the negative a increases the spike rate. Transient spikes occur in response to a sudden increase in

current, where the adaptation current is slow enough to fully compensate for the sharp change in current – eventually leading to a spike. The behavior has a close resemblance to rebound spikes. Irregular spiking is an aperiodic change of sharp and broad resets – and according to [42], is valid for a limited set of parameters. More details on the dynamics of AdEx model can be found in [36, 38, 42, 43].

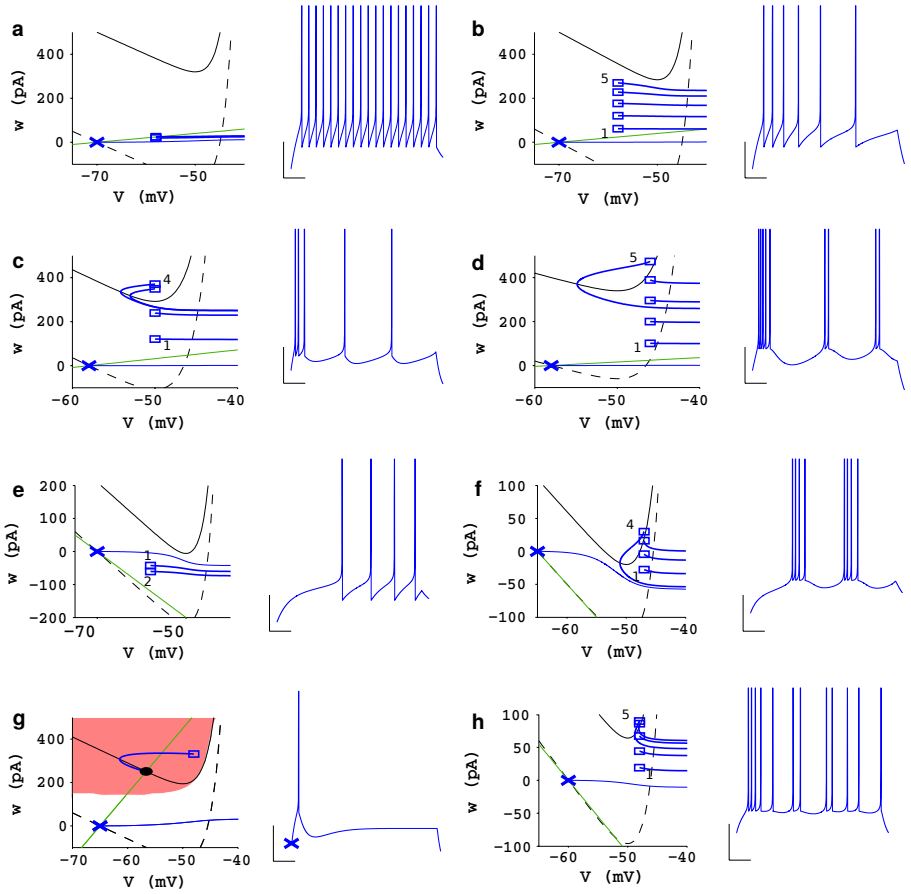


Figure 1.7: The eight spike patterns known from the AdEx model upon current stimulation, together with their phase plane representations (2D space of membrane voltage and the adaptation current shown at left side of each pattern). These are classified as: a) tonic spiking, b) adaptation, c) initial bursting, d) regular bursting, e) delayed accelerating, f) delayed regular bursting, g) transient spiking, h) irregular spiking. Image taken from [42].

1.3 Synapses

Synapses are specialized structures responsible for information transfer between any two neurons in the central nervous system. Most synapses are chemical in

nature, i.e., an action potential from a pre-synaptic neuron activates ion channels in the post-synaptic neuron, essentially changing the excitability of the later.

The inset in Fig. 1.2 shows the synapse formed between a pre- and post-synaptic cell. As an action potential invades a synapse, a series of biochemical processes lead to the release of neurotransmitter in the *synaptic cleft* – a gap between the terminals of a pre-synaptic and post-synaptic cell membranes. The neuroreceptors in the post-synaptic cell membrane detects this neurotransmitter, leading specific channels to open, that causes an inflow of ions into the cell. This ion-flow brings about a transient change in the post-synaptic neuron membrane voltage – the *post-synaptic potential*(PSP).

If the PSPs increase the likelihood that the post-synaptic cell evoke an action potential they are excitatory PSPs or (EPSP), and if they decrease this likelihood, they are inhibitory PSPs (IPSP). The EPSPs have a reversal potential E_{rev} more positive than the threshold, while the IPSPs have it more negative compared to the firing threshold. The exact nature of a synapse depends upon the neurotransmitter and receptors it activates. Excitatory synapses have typically glutamate as the neurotransmitter, and the receptors are either AMPA where channels open faster, or the voltage-gate N-Methyl-D-Aspartat (NMDA) which are typically much slow. Examples of inhibitory synapses are the fast GABA_A and the slower GABA_B, both of which use Gamma-Aminobutyric Acid (GABA) as the neurotransmitter [38].

1.3.1 Models of Synaptic Interaction

The ion-channels activated by the neurotransmitters are defined by the time dependent conductance $g_{\text{syn}}(t)$, which opens upon arrival of a pre-synaptic spike [38]. The current that is put out onto the membrane in Eq. 1.1 from the synapses is described in terms of the conductance as

$$I_{\text{syn}} = g_{\text{syn}}(t) \cdot (V_{\text{mem}} - E_{\text{syn}}) \tag{1.4}$$

where E_{syn} is the reversal potential. The equation expresses that the current in a conductance-based model depends on the difference of reversal potential and the membrane potential. The difference of these two potentials also defines if the type of synapse is excitatory or inhibitory. The time dependent conductance g_{syn} is typically expressed as the superposition of exponentials, where its time course is an exponential decay

$$g_{\text{syn}}(t) = \sum_{\text{f}} w_{\text{syn}} e^{-(t-t_{\text{f}})/\tau_{\text{syn}}} \Theta(t - t_{\text{f}}) \tag{1.5}$$

where w_{syn} denotes the weight that quantifies the amplitude of the post-synaptic response, τ_{syn} is the synaptic time constant of the decay, t_{f} denotes the arrival time of the pre-synaptic spike and Θ is the Heaviside step function. By substitution we obtain

$$I_{\text{syn}}(t) = \sum_{\text{f}} w_{\text{syn}} e^{-(t-t_{\text{f}})/\tau_{\text{syn}}} \Theta(t - t_{\text{f}})(V_{\text{mem}} - E_{\text{syn}}) \tag{1.6}$$

Alternatively, the synaptic interaction models can also be expressed as current-based kernel, which is expressed as

$$I_{\text{syn}}(t) = \sum_{\text{f}} w_{\text{syn}} e^{-(t-t_{\text{f}})/\tau_{\text{syn}}} \Theta(t - t_{\text{f}}) \tag{1.7}$$

which indicates that a linear sum of PSPs is possible, as opposed to the conductance based sum of Eq. 1.6 where the difference from reversal potential makes it non-linear [24].

1.3.2 Plasticity

Learning and memory in the nervous system is widely attributed to synaptic plasticity. Synapses are able to modulate their strength depending upon the activity. Typically, the induction of such changes are classified over different timescales and referred to as Short-Term Plasticity (STP) or Long-Term Plasticity (LTP). The changes brought about during STP last only a few hundred milliseconds [45] and the successive pre-synaptic spikes evoke smaller (*depression*) or larger responses (*facilitation*) in the post-synaptic cell [46]. A recovery to normal amplitudes occur within a second. A popular phenomenological model implementing short term dynamics is the Tsodyk-Markram model, where depression [47] as well as its extension for facilitation [45] is modeled.

Long term plasticity occurs in the form of *potentiation* or *depression*, and the changes are more persistent – spanning from minutes to hours or longer. One example is the Spike-Timing-Dependent Plasticity (STDP) [48–51], where the induction time can still be brief, e.g., a few seconds, but change is persistent for more than an hour [46]. Additionally, homeostatic plasticity where the activity of synapses is regulated, is also on the timescales that extend from minutes to hours [52].

1.4 Dendrites

Neurons compute by transforming a complex set of dynamical inputs into a sequence of output spikes [53]. McCulloch and Pitts argued that, by adding memory to a network of linear threshold units, all fundamental operations of a digital computer can be computed [54]. Similarly, the threshold-based models (discussed in Sec. 1.2) introduce a non-linearity provided by the threshold, and can compute logical functions, such as the AND operation. However, these models provide a simplified behavior of the computational aspects, for example, by assuming that the synaptic inputs do not interact with each other, or by not capturing the properties of the dendritic tree. In biology the dendrites appear in diverse shapes and sizes, with morphologies varying widely from the depiction shown in Fig. 1.2. For example, some well-studied neuron types are shown in Fig. 1.8 [40, 55].

The development of linear cable theory by Wilfrid Rall in the late 50's [57] together with experimental findings showed, that neuronal dendrites are electrically

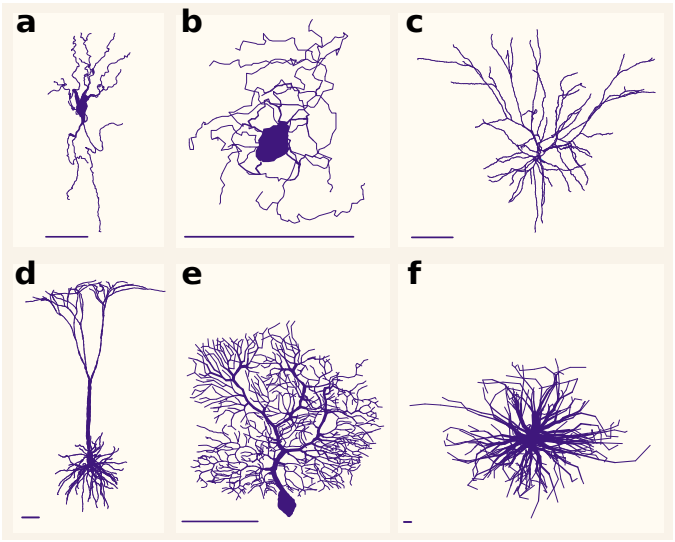


Figure 1.8: Neurons from different brain regions with varying morphological shapes and sizes: a) Vagal motoneuron, b) Olivary neuron, c) L2/L3 pyramidal cell, d) L5 pyramidal cell, e) Purkinje cell, f) α motoneuron. Scale bars are 100 µm long. Image taken from [40, 55].

distributed rather than being isopotential elements, and that post-synaptic potentials undergo voltage attenuation and significant temporal delay in the dendritic tree [40]. In passive dendrites, dependent upon the initiation site, the sum of post-synaptic potentials of two synapses can be less than the linear sum of their separate responses. The non-linear effects are contributed by *shunting inhibition*, where the synaptic reversal potential is close to the cell's resting potential – and the channel increases the local conductance, thereby reducing the effect of subsequent incoming EPSPs. In the equivalent schematic of the cell membrane shown in Fig. 1.3, the chloride ions are one such example, since their reversal potential are typically close to the resting potential. More examples of passive dendritic computation include, for example, the auditory neurons in the chicken brainstem and barn owls. Their bipolar dendrites form coincidence detectors [58–60], firing strongly only if the input from the two ears coincide in time – with a time window of 10 – 100 µs. Another example are the direction selective retinal ganglion cells, that respond to the movement of the stimulus in only one direction and not the opposite way [61, 62].

The active properties of dendrites contribute with a backpropagating action potential that goes back from the soma into the dendrites. This implies that single neurons provide an internal feedback – not restricting it only as a network property [63, 64]. This backpropagating action potential induces LTP in CA1 and Layer 5 neurons [49, 65]. Further, it evokes a broad Ca^{2+} spike in apical dendrites and multiple somatic action potentials [66] when dendritic and somatic input coincide. In this way, the active properties of dendrites help amplify the synaptic inputs attenuated by the passive dendritic tree [64]. This is further supported since

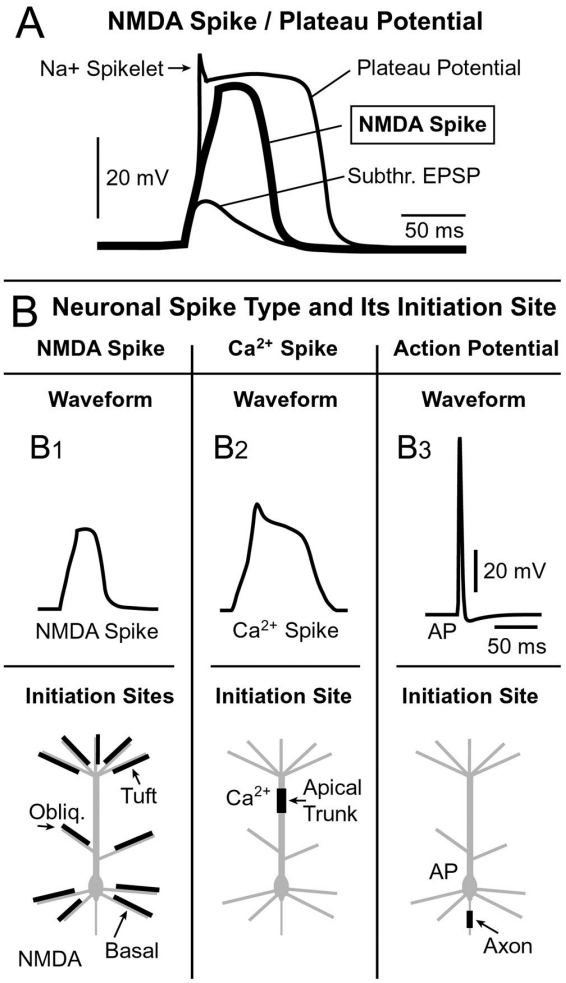


Figure 1.9: Somatic and dendritic spikes and their regions of initiation in a cortical pyramidal neuron. A) NMDA spike and plateau potential is shown together with the subthreshold EPSP evoked in the thin dendrites. Upon sufficient stimulation the NMDA spike transforms into a plateau potential, initiating with an Na^+ spikelet and followed by a plateau phase and an abrupt collapse. B1–B3) Neuronal spike types and their corresponding initiation region. The nominal action potential is initiated in the axon, whereas the Ca^{2+} spike typically occurs in the apical dendrites. The NMDA spikes are elicited in the basal/oblique/tuft regions. Figure taken from [56].

studies [67, 68] suggest that conductances of distal synapses are scaled according to their distance from the soma. The active properties also evoke local dendritic spikes triggered by the co-activation of synaptic inputs and mediated by voltage-gated ion channels [69, 70].

So far we have mostly described the dendritic properties related to thick apical dendrites. The thin basal, oblique and tuft dendrites in cortical pyramidal cells receive a high density of glutamatergic synaptic inputs, whose subset of 10–50 synchronous activation can trigger a local dendritic regenerative potential, referred to as the NMDA spike or plateau potential [56]. The name NMDA refers to the ionic contributor which is the NMDA receptor current [71]. In contrast to the calcium spikes, these are truly local to the dendritic branch [72, 73]. Compared to sodium spikes they are characterized by significant amplitude (40–50 mV) and duration [71]. Models predict that they are evoked by an initial fast local sodium spikelet, giving rise to a slower calcium-mediated regenerative response, which elicits a full-blown NMDA spike [71]. These three spike shapes as well their initiation region in cortical pyramidal neuron is depicted in Fig. 1.9.

Upon sufficient (glutamatergic) stimulation the step-like depolarizing NMDA spikes broaden into plateaus, without an effect on the amplitude – demonstrating their strong non-linear behavior [56, 74, 75]. The local NMDA dendritic spikes/-plateau potentials have been implicated to have a role in cortical information processing and memory consolidation [56, 76].

1.5 Outline of this work

This thesis is about the implementation of two spiking neuron models – namely the LIF and the AdEx, defined for integration in the second-generation BrainScaleS hardware. This work is carried out as part of the design and development of the BrainScaleS analog neuromorphic computing hardware. The BrainScaleS hardware design project has previously been carried out in European projects such as the *FACETS* and the *BrainScaleS*. Since late 2013, the project is funded under the *Human Brain Project*, where development of the second generation neuromorphic platform has been initiated. The project is strongly driven by previous design experience and end-user requirements – which in large part is the computational neuroscience community. The BrainScaleS design approach is to emulate the computational models of neural elements as analog and mixed-signal circuit implementations. The thesis is structured as follows:

Chapter 2 explains the system architecture of the second generation neuromorphic platform, which is still under design phase. The architectural description is an update only until the time of the compilation of this thesis. Since the second-generation wafer-scale system is not yet existent, a description of the existing first generation system is provided for overview.

Chapter 3 starts with the design approach. The requirements and target specifications, the used CMOS technology, the description and measurement framework

of the prototype chips are explained.

Chapter 4 and 5 describe the detailed design and implementation of the two spiking neuron models, their individual circuits, their measured and simulated results.

Chapter 6 summarizes the achieved results and concludes the thesis. The limitations of the current implementation together with improvement suggestions are listed.

The author worked in a team of analog and digital designers, where close collaboration especially during chip design runs was necessary. This resulted in a group effort for the targeted goal. The production of successful mixed-signal chips is therefore an achievement of all designers involved during the different phases – from specification, design, verification as well as lab measurements.

Two theses have been supervised during the course of this work. The first Bachelor thesis carried out by Gerd Kiene, was just prior to the design of the first prototype chip. It evaluated the performance of the synaptic input circuit of the first generation HICANN neuron, with the help of circuit simulations [77]. The second Bachelor thesis carried out by Yannik Stradmann measured and characterized the neuron array on the second prototype of the HICANN-DLS chip [78]. The work started after the initial measurements on the designed neuron circuit looked promising and a detailed characterization was necessary. Yannik Stradmann also worked as a scientific assistant later under the supervision of the author, and calibrated the neuron circuit over multiple dies. The multi-chip results shown from the second prototype of the chip are a result of this work and has been submitted for a publication [79].

Apart from thesis supervision, the author also assisted for teaching an electronics course for three semesters.

1.5.1 Publications

Most of the work compiled in this thesis is either already published or is under review. The following is a list of journal papers and conference contributions made by the author:

- S. A. Aamir, P. Müller, A. Hartel, J. Schemmel and K. Meier, “A Highly Tunable 65-nm CMOS LIF Neuron for a Large Scale Neuromorphic System”, in *Proceedings of the 42nd European Solid-State Circuits Conference*, September 2016, pp. 71-74.

- S. A. Aamir*, P. Müller*, L. Kriener, G. Kiene, J. Schemmel and K. Meier, “From LIF to AdEx Neuron Models: Accelerated Analog 65 nm CMOS Implementation” in *Proceedings of the 13th IEEE Biomedical Circuits and Systems Conference*, October 2017, pp. 1-4.

- S. A. Aamir*, Y. Stradmann*, P. Müller, C. Pehle, A. Hartel, A. Grübl, J. Schemmel and K. Meier, “An Accelerated LIF Neuronal Network Array for

a Large Scale Mixed-Signal Neuromorphic Architecture", article under review, submitted to *IEEE Transactions for Circuits and Systems I: Regular Papers*.

- S. A. Aamir, P. Müller, G. Kiene, L. Kriener, Y. Stradmann, J. Schemmel and K. Meier, "A Mixed-Signal Structured AdEx Neuron for Accelerated Neuromorphic Cores", article under review, submitted to *IEEE Transactions on Biomedical Circuits and Systems*.

- J. Schemmel, S. A. Aamir, S. Billaudelle, T. Demirci, A. Grübl, A. Hartel, G. Kiene, Y. Leblebici, C. Pehle, K. Schreiber, Y. Stradmann and K. Meier, "An Analog Neuromorphic Hardware System Combining Structured Neurons with Hybrid Learning", article in preparation.

Chapter 2

Second Generation BrainScaleS Hardware

The BrainScaleS hardware is a mixed-signal wafer-scale neuromorphic system with an analog physical neuron model implementation. In analogy to the nervous system the local neural computation is analog, whereas the spike communication over the network is digital. The system operates faster than biological real-time – with an acceleration factor of 10^3 to 10^5 times. Since the neuromorphic substrate is wafer-scale, the CMOS wafer is not diced into individual dies. Instead, the entire post-processed wafer is used for large-scale integration by interconnecting multiple, identical on-wafer dies through vertical and horizontal connections [80]. This allows for the large neuronal count, not realizable within the geometry of a single ASIC.

The first generation BrainScaleS hardware has a 180 nm CMOS wafer of 20 cm diameter with a total of 48 functional reticles. Each reticle contains eight HICANN neuromorphic dies as the basic building blocks, resulting in a total of 384 on-wafer dies. Within each die, the neuron array and synapse matrix are arranged in a columnar architecture referred to as Analog Network Core (ANC). On the wafer-scale substrate, the adjacent ANCs are interconnected through horizontal and vertical lanes of the Layer 1 (L1) bus that spreads out over the entire wafer. The L1 routing is responsible for wafer-wide event communication, whereas the Layer 2 (L2) bus provides high-speed external routing to host FPGAs. Fig. 2.1 highlights a single reticle with eight first-generation HICANN chips as well as the vertical and horizontal routing. Fig. 2.2 shows a single HICANN chip bonded directly on a test measurement setup.

The second generation BrainScaleS hardware – currently under development, will replace the HICANNs with the 65 nm HICANN-DLS[1] chips. HICANN-DLS is designed in a way that it fits into the existing hardware and software framework – namely the main wafer PCB as well as being compatible with the software stack. Compared to the HICANN architecture, the HICANN-DLS also integrates a digital

[1] High Input Count Analog Neural Network with Digital Learning System.

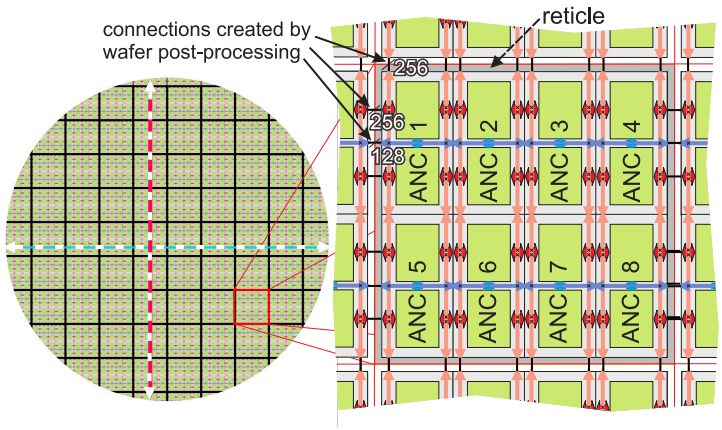


Figure 2.1: A drawing of the 180 nm CMOS wafer with highlighted reticle boundaries. A single reticle is zoomed-in to show the arrangement of eight on-wafer dies. Vertical (red) and horizontal (blue) connections pass through individual ANCs and created during the post processing stage. Image taken from [81].

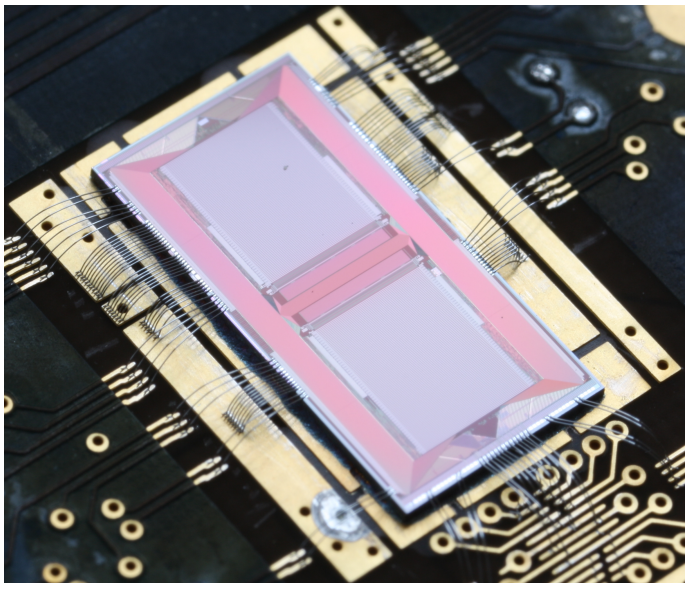


Figure 2.2: The first generation 5 mm $\times$ 10 mm HICANN die bonded on a measurement board. Two different ANC quadrants are visible in the two halves of the chip. Photo by Matthias Hock.

Plasticity Processing Unit (PPU), specialized for learning and plasticity [82–84]. Since the chip is a mixed-signal design, it is divided into the digital logic core and the PPU units on one hand, and the ANC and analog peripherals on the other. The acceleration factor in the HICANN-DLS is reduced (and fixed) to 10^3 – this entails new analog architectures for the ANC circuits that include neuron, synapses, etc.

2.1 The HICANN-DLS Chip

HICANN-DLS is a mixed-signal 65 nm CMOS neuromorphic system on-chip solution with analog and digital cores. The chip architecture accelerates all biological timescales with a fixed factor of 1000 times compared to real-time. The chip is currently under development phase[2] and several smaller prototypes have tested the features and individual circuit blocks. A first prototype of the larger HICANN-DLS ASIC is planned as a 4 µm $\times$ 8 µm die whose finalized version will replace the HICANN ASIC of Fig. 2.2. The HICANN chip had been designed with two ANC quadrants whereas HICANN-DLS is tiled into four quadrants.

The ANC is a columnar architecture of an edge-connected neuron array and synapse matrix and forms the core of the BrainScaleS computational units. A generalized architecture of the ANC is sketched in Fig. 2.3. Within each column a single neuron compartment is connected to M synapses which forms the dendritic input of each neuron. The input pre-synaptic network events arrive at the synapse drivers via the L1 buses in the larger system, and directly via Synchronous Parallel Layer 1 (spL1) in the single-chip prototypes from the left or right edges. Each synapse driver drives two synapse rows, one in each adjacent quadrant. Each synaptic column of M synapses provides the neuron compartment with excitatory and inhibitory synaptic input pulses of 4 ns duration on two separate lines. The neuron integrates the input current on its membrane, and produces a digital output event when the membrane reaches a threshold. This output is routed to the synapse as the *post* event via the digital neuron control. Each of these digital neuron blocks get output spikes from N/2 neurons. A total of eight such digital blocks are therefore integrated. To serialize the output data from N/2 inputs, a priority encoder arbitrates access to the output bus inside each of them. Every neuron in the column is provided with 24 dedicated capacitive storage parameters for analog configuration. These are physically placed between the neuron compartments and the digital control blocks.

The system features an implementation of the STDP learning rule for which it stores the temporal correlation of pre- and post-synaptic events as voltages on two capacitors inside each synapse. This correlation data (voltage on the two capacitors) is digitized by two Analog-to-Digital Converter (ADC) channels per column and then read by the PPU which implements the learning rule. These ADC channel are located at the top half of each synapse matrix. The chip implements 512

[2] as of December 2017

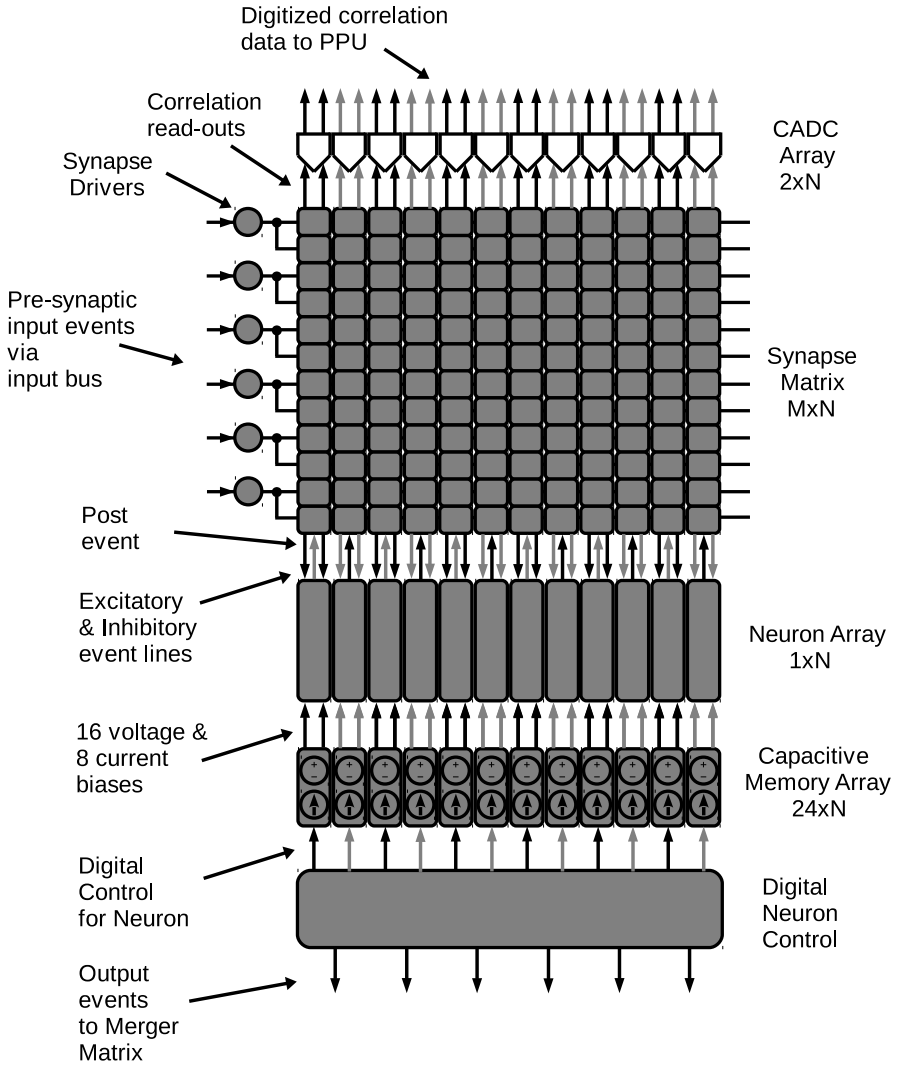


Figure 2.3: The columnar architecture of the analog network core.

neurons in four separate quadrants as shown in Fig. 2.4. Each of the four quadrants implements $N = 128$ columns taking input from $M = 256$ synapse rows.

The chip contains two PPU units [82] located at the top and bottom edges of ANC. The PPU is a general purpose microprocessor implementing a 32-bit Power Instruction Set Architecture (ISA) with a specialized vector processor in a parallel Single Instruction Multiple Data (SIMD) organization. It can modify the synaptic weights based on the implemented learning rule and is geared to implement STDP [85].

The chip features a separate on-chip Membrane ADC (MADC) to digitize the neuron's membrane prior to external read-out. A Phase-Locked Loop (PLL) gen-

erates four clocks from an external input clock of 50 MHz. It provides 750 MHz to the MADC, 500 MHz to the PPU units, 1 GHz to the Serializer/Deserializer (SerDes) and 250 MHz to the spL1 bus respectively. A JTAG interface can configure the PLL control registers. Four SerDes channels are realized to ensure high speed serial communication. For reading out analog voltages externally two 50Ω output buffers are integrated. The left/right and top/bottom edges of the chip are endowed with horizontal and vertical L1 repeaters. They restore the signal levels and timing in the larger system, since the L1 bus lane length in each chip can be as long as 8 mm and signal quality due to crosstalk, etc. can degrade. Fig. 2.4 shows the potential floorplan containing the edge L1 repeaters, the MADC, the PLL as well as the SerDes units.

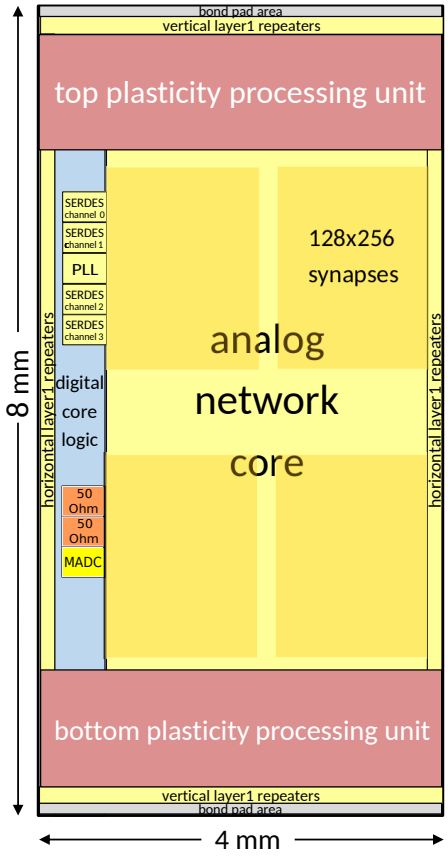


Figure 2.4: The preliminary sketch of the floorplan of HICANN-DLS chip. Image modified and used with permission from [86].

The event communication interfaces and various components of the ANC are described below, followed by a brief overview of the first-generation wafer system.

2.1.1 Communication Interfaces

In the absence of L1 routing, digital event communication in the single chip prototypes occurs via the spL1 interfaces. A merger matrix merges the input events from the digital neuron control and the off-chip interface, and generates either local events for the synapse drivers or for the host via the L2 interface. Host (external) communication is performed via the L2 interface which communicates via eight serial transceivers with the FPGA. The total communication bandwidth is shared between spL1 events, the OMNIBUS and the PPU's memory interfaces and is arbitrated by the L2. All slow control, for example, the configuration of neuron and Capacitive memory (Capmem) occurs via the OMNIBUS, including the inputs to ANC from the PPU. This event communication routing controlled by the merger matrix is sketched in Fig. 2.5. The synapse drivers receive input events from the Parallel Debug Input (PADI) bus (not shown) in the absence of L1. The PADI arbitrates between the inputs from merger matrix or from the OMNIBUS.

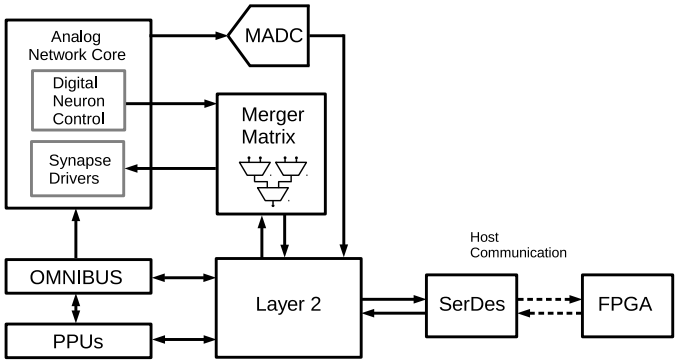


Figure 2.5: A simplified architectural sketch of event communication interfaces. Image adapted from [86].

2.1.2 Synapse Drivers

Pre-synaptic input events enter the synapse matrix via the synapse drivers from its left edge or right edge [87]. The input arrives either from the L1 bus, or alternately from the OMNIBUS. Synapse drivers either relay the input events directly or add a dynamic behavior governed by STP dynamics. When enabled, the circuit implements Short-Term Facilitation (STF) or Short-Term Depression (STD) according to a modified Tsodyk-Markram model [47] resulting in a facilitating or depressing pre-synaptic input. The synapse driver achieves this by modulating the pulse width of the signal that enables the synapse. A detailed description of synapse drivers can be found in [87].

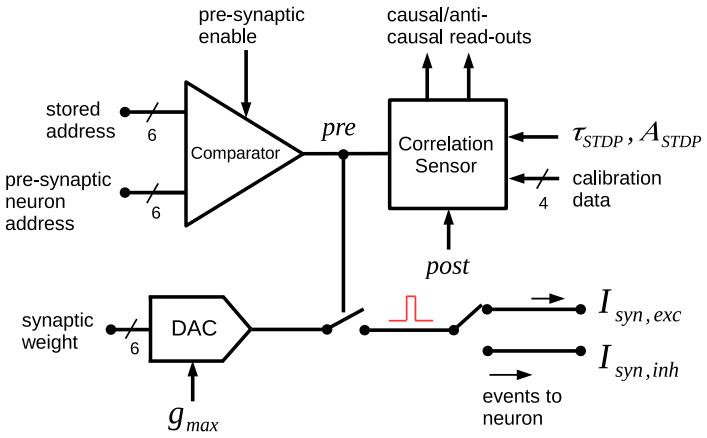


Figure 2.6: A simplified block-level schematic of a single synapse. Adapted from [83].

2.1.3 Synapse Matrix

Each synapse stores a pre-synaptic address in its local SRAM that determines the neuron it responds to. The synapse circuit consists of an address comparator, a 6-bit DAC and a correlation sensor circuit, all with associated SRAM memories. The comparator compares the locally stored 6-bit address to that of incoming received address. Upon address match, the comparator enables the output of a 6-bit DAC for a 4 ns duration (modulated by the STP circuit). This marks an output current pulse event, whose amplitude is modulated by a 6-bit synaptic weight (DAC code). The synapse schematic is shown in Fig. 2.6. The amplitude of each event can be up to 10 µA. The generated output pulse can be taken to either the excitatory synaptic input line or the inhibitory one, common to all synapses in a column.

The comparator output *pre* (pre-synaptic event) is compared with a *post* (post-synaptic event) signal provided by the (post-synaptic) neuron via the digital control. The correlation sensor finds the temporal correlation between a pre and post events. It determines the time difference of the two signals and depending upon the polarity, stores an exponentially weighted time-difference on the causal or anti-causal storage capacitors. The stored values can be simultaneously read-out for all synapses in a row. The PPU iterates over all rows of synapses sequentially and performs the weight updates.

2.1.4 Correlation ADCs

These are 8-bit single-slope ADC converters which digitize the analog voltage on causal and anti-causal storage capacitors of the synapse columns. Also known as integrating ADCs, the architecture provides good accuracy for slow moving signals [88]. Each column within the ANC contains two such ADC channels and each

ANC quadrant on the chip has 256 channels. A global ramp generator distributes a voltage ramp to all channels. Inside the array the ramp voltage is compared with the correlation input from synapses. The time it takes for the voltage ramp to reach the input voltage is simultaneously counted by an 8-bit counter in terms of clock cycles. Upon a comparator hit, the counter value is latched as a converted digital value. An 8-bit digital register is used to compensate for the input offset of each comparator. The ADC runs within the clock domain of the PPU and therefore has a maximum conversion time of about $2^8 \cdot 1/f_{\text{clk}} = 0.5$ µs corresponding to a rate of 2 MSps.

2.1.5 Capacitive Memory

Every single neuron circuit in the columnar ANC architecture is endowed with a block of 24 Capmem cells [89, 90]. Each one of them can store tunable and reconfigurable voltage or current bias to tune the neuron circuit. The 24 cells are distributed as 16 current biases and 8 voltage biases. These are the dedicated or local biases, since they are individually tunable for every neuron. In addition to this, there exists a global block of 24 $\times$ 2 biases that may not be tuned individually, but are meant to be common to all neurons (or other circuits in the ANC). These are the global or shared biases. Each bias can also be read out via a debug interface.

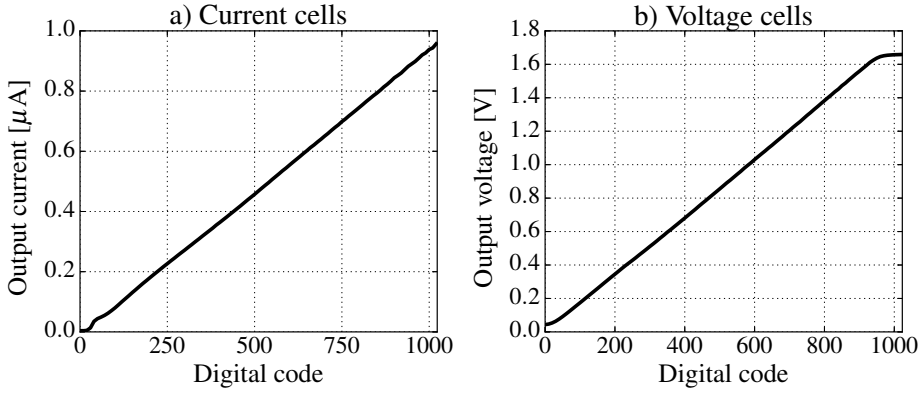


Figure 2.7: Example traces for the current and voltage biases, being programmed with a 10-bit digital code as their corresponding analog output is read-out.

Within each voltage bias cell a digital target value is programmed in a 10-bit SRAM memory. A global voltage ramp is generated that rises from 0 to 1.8 V and is distributed along the entire Capmem array. A 10-bit digital counter starts counting simultaneously, and every time it hits the maximum value, it resets the voltage ramp. A comparator compares the digital counter with that of the target SRAM value and upon a match, the ramp voltage is sampled on a capacitor. In the current cells, the voltage ramp is converted to generate a linearly rising current, and

using a circuit similar to the voltage cells, the sampled voltage generates current from an output PMOS transistor.

The Capmem provides a bias storage from 200 mV to over 1.6 V for the voltage cells and 15 nA to 1 µA in the current cells. This range is tunable with a 10-bit resolution. The current cells have a PMOS output stage tied to a 2.5 V supply. This means additional mirroring if the receiver circuit (that uses the bias) does not have an NMOS based input current mirror. This may entail more variation due to device mismatch. Further, if the receiver circuit has a thin-oxide circuit directly biased with such current cells, then the design should ensure that no low-voltage nodes exceed 1.2 V. Due to the high output impedance, in most cases this should not be a problem. From the software interface the user programs a digital 10-bit DAC value, against which an equivalent voltage and current bias is programmed. Once a target value is programmed, it may take one or more ramp refresh cycles until the target value is reached. The refresh rate is typically set at 1–2 kHz.

Fig. 2.7 shows plots for voltage and current cells where the digital programmed value is swept for both types of cells. The analog output of both voltage and current cells are read-out for every LSB increase over the entire 10-bit range and is plotted for all cells. More details on the design and specifications of the capacitive memory can be found in [89, 90].

2.1.6 Membrane ADC and PLL

The high-speed ADC that digitizes the membrane (MADC) is a 10-bit time-interleaved successive approximation register architecture. It has a 125 MSps maximum conversion rate with two channel time-interleaving, each with a rate of 62.5 MSps and consuming about 2 mW power. The membrane input signal is converted to a differential signal and connected to both ADC channels with different sample and hold phases. An on-chip phase generator generates separate phases from an externally supplied input clock.

The PLL is a two channel all-digital architecture with three output clocks per channel. It takes an input clock $f_{\text{clk,ext}}$ of 50 MHz to provide a maximum of $\frac{N \cdot P_0 \cdot f_{\text{clk,ext}}}{P_2}$ from the first output, and $\frac{N \cdot P_0 \cdot f_{\text{clk,ext}}}{P_1 M_{0,1}}$ from the other two outputs. Where N is the loop divider and $P_{0,1,2}$ are the pre-dividers. These pre-dividers can be set between 2, 3 or 4, whereas $M_{0,1}$ can be set between 0 and 31. The loop divider N can have values between 2 and 31. The PLL can generate a maximum output clock frequency of 1 GHz.

The PLL and the MADC have been designed by project partners at TU-Dresden and EPFL.

2.1.7 Neuron Array

Each column within an ANC quadrant integrates a single neuron circuit and a total of 128 columns lead to a 128 $\times$ 4 neurons on a single HICANN-DLS die. The neuron circuit emulates an AdEx neuron model as a point neuron [91] that can be

extended to multiple compartments. Along with the equivalent of Sodium (Na^+) spikes, it can be configured to evoke broad spikes, such as NMDA plateau potentials and calcium (Ca^{2+}) spikes [92]. The point neuron which takes a synaptic fan-in of 256, can merge its membrane with other neuron compartments to realize larger neurons with a higher fan-in of greater than 10 thousand. A larger neuron (combined membrane) however decreases the size of individual Post Synaptic Potential (PSP)s. Each point neuron circuit is configured via 16 current biases and 8 local voltage biases for analog control, as well as 40-bit SRAM for digital configuration. The neuron circuit can be reduced to the LIF model due to its modular architecture [91, 93]. The configurable pulse intervals, for example, the tunable refractory period, the adaptation pulse-width are controlled by a digital block [94].

The neuron integrates the current pulses from the dendritic input provided by excitatory and inhibitory synaptic input lines. Each excitatory input causes the membrane potential to rise and decay with a time constant towards a resting potential. It evokes a digital event once the integrated analog voltage reaches a specified threshold. This output digital pulse *fire* marks a single spike, that is routed to other pre-synaptic inputs in the network through the digital neuron control and the merger matrix. The analog voltage can be read-out using a voltage buffer. Fig. 2.8

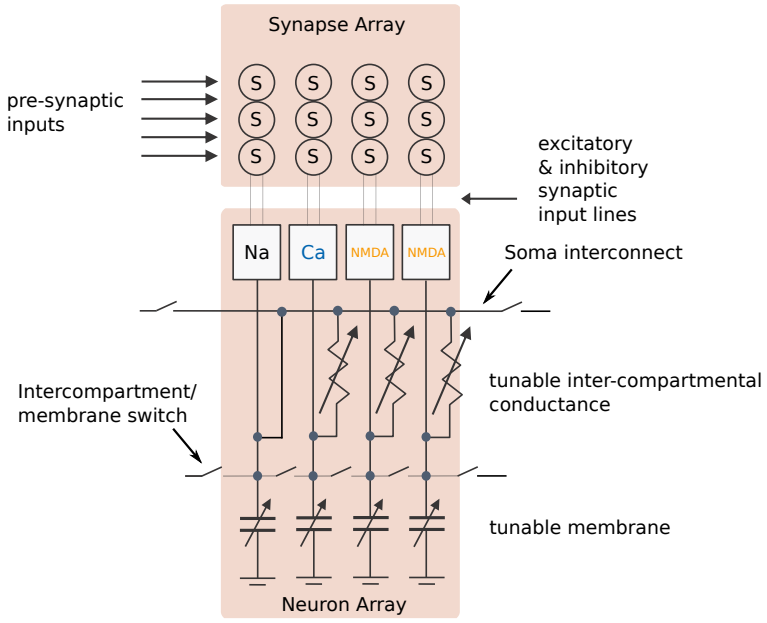


Figure 2.8: An illustration of the realized multi-compartment columnar array as highlighted in [92]. Each neuron compartment in the array can be configured to elicit different types of spike responses. The Na^+ compartment realizes a high-conductance path (direct connection) to the soma. The soma forms connections to other compartments via a tunable inter-compartment conductance.

shows a block diagram of the synaptic array connected in the columnar architec-

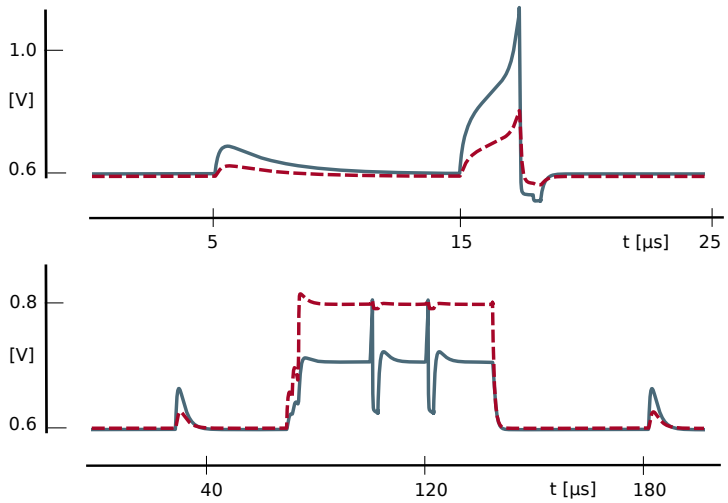


Figure 2.9: Spiking response in different compartments and ion channels upon stimulation. *Top)* A sodium exponential spike (blue) and its coupling effect in a neighboring compartment (red). *Bottom)* Na^+ spikes (blue) at 100 µs and 120 µs and an NMDA plateau potential (red). Figure adapted from [92].

ture to the neuron compartmental array. Four columns are shown, where only Na^+ compartment is connected directly to the soma. The other compartments, configured to evoke Ca^{2+} spikes/NMDA Plateau Potentials, are connected to the soma via a tunable conductance. Every single compartment may also be configured as a single point neuron, or alternatively, as a single large neuron where the membrane capacitance of all columns/compartments is connected via the membrane switch. The membrane capacitor of each compartment is in turn tunable.

Fig. 2.9 (top, blue line) shows a circuit simulation where an exponential spike is evoked in one neuron compartment, as a result of strong input stimulus applied at 15 µs. A neighboring connected compartment (top, red trace) passively follows the main compartment and responds with a subthreshold behavior. The bottom figure shows example spike shapes and their duration. Two Sodium spikes (bottom, blue trace) are shown at about 100 µs and 120 µs respectively and a broad NMDA plateau potential (red trace) is evoked between 78 µs and 145 µs when stimulated by input synaptic current.

This thesis covers the implementation of single compartment neuron models. The emulation and analysis of multi-compartment emulation is not covered. Interested readers for multi-compartment realization are directed to read [92].

2.2 Existing Wafer-Scale System

As mentioned above, the BrainScaleS system does not dice the CMOS wafer and chip-to-chip communication can occur by staying on the wafer. Compared to a standard PCB solution with multiple chips, this approach helps address the high data rate requirement, reduces the switching energy due to lower line capacitance (shorter on-wafer traces), and simplifies the signal integrity and matching issues [80].

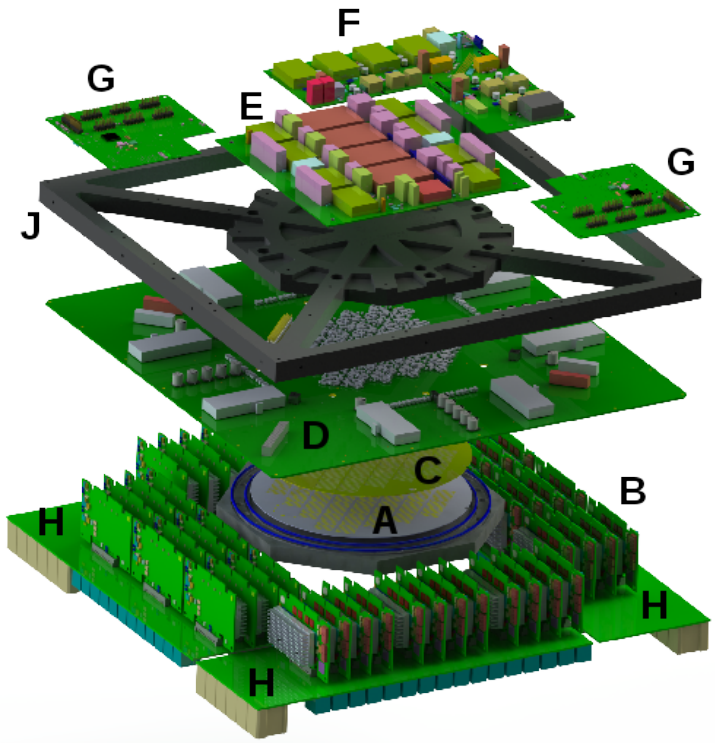


Figure 2.10: A 3D rendered drawing of individual parts that constitute the wafer module [80].

Fig. 2.10 shows a rendered drawing of the integrated wafer module in the first-generation wafer system (A). The foundry CMOS wafer is post-processed to create links between adjacent dies as well as to provide mating structures to connect the wafer to its main PCB (D) via the elastomeric connectors. An opening in passivation layer is created to connect top metal layer to two copper routing layers – the fine-pitched routing for interconnection between the adjacent reticles, and the coarse intermediate routing to connect to the reticle pads. Two elastomeric connectors are shared between every two dies, and are used to communicate high-speed clock and data signals, JTAG I/O and read-out data as well as the supply voltages. To align the 384 elastomeric connectors during the assembly stage, a positioning mask (C) with 384 slots is cut out from an FR4 sheet [80]. Beneath the wafer is

another PCB that hosts up to 48 Xilinx Kintex-7 FPGA boards, one per reticle (B). This FPGA board communicates with eight HICANN dies via the Low-Voltage Differential Signaling (LVDS) interface, as well as to the host PC via the Gigabit Ethernet interface. It can also communicate to other FPGA boards using the Xilinx Gigabit Transceivers (GTX). The board provides an I/O interface for configuration and spike data from the host PC and contains DDR3 memory for storage of Ethernet frames, stimulation pulses and output pulse activity during an experiment. The Gigabit Ethernet connectors (H) are available at the bottom edges and provide connectivity to the compute cluster as well as to other wafers. An aluminum frame (J) provides mechanical stability to the system. On the top there are separate PCBs (E, F, G) that provide power supply to the system, as well as the analog read-out capability. A photograph of the fully assembled wafer module is shown in Fig. 2.11.

Figure 2.11: A photograph of a single fully assembled wafer module [80, 95].

An output event generated from one neuron on a chip, takes a path that is specified in Fig. 2.12. A neuron marked *N1* fires and a priority encoder arbitrates its access to the network. The neuron with highest (and fixed) priority is selected. The 6-bit neuron identifier is streamed out by the serializer, and sent out to the L1 bus via an output driver that caters for L1 voltage levels. As the signal traverses the vertical and horizontal L1 buses, each chip inserts repeaters at its boundaries for signal and timing restoration. A repeater consists of an input differential amplifier to restore signal levels. Timing restoration is done with a serializer and a Delay-Locked Loop (DLL). A crossbar switch connects the horizontal and vertical L1 lanes. The event may go through several repeaters before being received at the

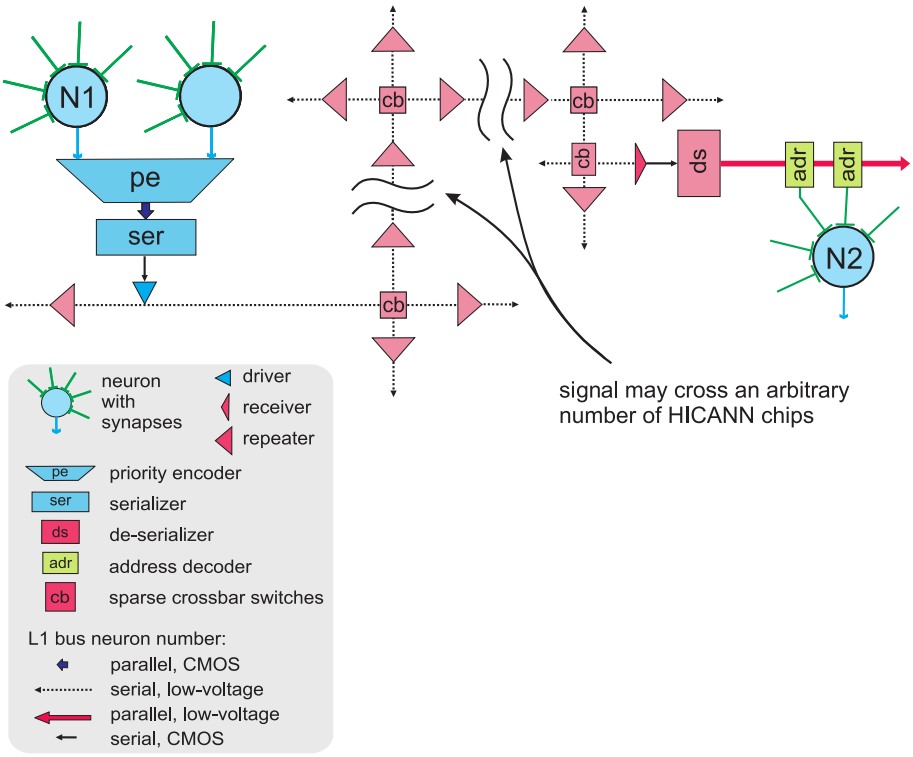


Figure 2.12: The event route over the horizontal and vertical L1 lanes in the wafer-scale system [81].

target site where it is de-serialized. The resulting output address is decoded and enters the target neuron through the respective synapse driver. More information on BrainScaleS wafer-scale system integration can be found in [80, 81].

Chapter 3

Design and Measurement Framework

This chapter introduces the pre-considerations, design environment and the measurement framework available for the HICANN-DLS neurons. In this context, we review the implemented neuron models and the parameter ranges they target to motivate the circuit specifications. An overview of the CMOS process technology, operating characteristics of MOS devices used in this work, introduction to the fabricated chips as well as their measurement setups are provided.

3.1 Models for Hardware Implementation

The choice of a hardware neuron model is dictated by the trade-off between the rich template and computational power of a biological neuron, versus the design complexity, power consumption and on-chip area requirements. This implies that the neuron models should be able to replicate most computational studies, yet be simple enough such that a prototype ANC integrates a sizable number of neurons to prove and realize small functional networks. This opinion has also been endorsed, e.g., in [96, 97] as a means to envision large-scale neuromorphic systems. We therefore rely on low-dimensional integrate-and-fire neuron models, described previously in Sec. 1.2.

The LIF model can be directly realized in hardware, as the membrane equation and the reset condition (reproduced in Eq. 3.1 and Eq. 3.2) require current integration on a capacitor and a non-linear pulse generation circuit that resets the membrane as it reaches a threshold. Mathematically this can be expressed as

$$C_{\text{mem}} \frac{\mathrm{d}V_{\text{mem}}}{\mathrm{d}t} = -g_{\text{leak}} \cdot (V_{\text{mem}} - V_{\text{leak}}) + I \tag{3.1}$$

and if $V_{\text{mem}} \geq V_{\text{thresh}}$

$$V_{\text{mem}} \rightarrow V_{\text{reset}} \tag{3.2}$$

where V_{mem} is the membrane potential, C_{mem} is the membrane capacitance, V_{reset} and V_{thresh} are reset and threshold potentials, V_{leak} models the leak potential and g_{leak} is the leak conductance. I is the sum of the externally injected current (I_{stim}), the synaptic excitatory ($I_{\text{syn,exc}}$) and inhibitory ($I_{\text{syn,inh}}$) currents. The synaptic inputs integrating these currents are exponentially decaying current-based inputs, whose time course can be defined as

$$I_{\text{syn}}(t) = \sum_i \sum_f w_i e^{-\left(\frac{t-t_i^f}{\tau_{\text{syn}}}\right)} \Theta(t - t_i^f) \tag{3.3}$$

where w_i is the weight of the synapse connecting a pre-synaptic neuron to a post-synaptic neuron, t_i^f denotes the fth spike at a synapse i, $\Theta(x)$ is the Heaviside step function and τ_{syn} is the synaptic time constant.

The first implementation in the prototype chips has featured the LIF model, extended to the AdEx model [91] in the last revision. The AdEx model previously described in Sec. 1.2.2 is defined by

$$\begin{aligned} C_{\text{mem}} \frac{\mathrm{d}V_{\text{mem}}}{\mathrm{d}t} &= I - w - g_{\text{leak}}(V_{\text{mem}} - V_{\text{leak}}) \\ &\quad + g_{\text{leak}} \Delta_T \exp\left(\frac{V_{\text{mem}} - V_T}{\Delta_T}\right) \end{aligned} \tag{3.4}$$

$$\tau_{\text{w}} \frac{\mathrm{d}w}{\mathrm{d}t} = a(V_{\text{mem}} - V_{\text{leak}}) - w \tag{3.5}$$

where in addition to the Eq. 3.1 we have w as the adaptation current, and the last term in Eq. 3.4 models the exponential current. Where a is the subthreshold conductance, V_{T} is the exponential threshold and Δ_{T} is its slope factor. At spike time, the membrane is reset to a specified reset potential like the LIF model, but additionally the adaptation variable w is updated by a current b, such that $w \rightarrow w + b$.

To simplify the hardware realization of the adaptation term [98, 99], its output current is equated as

$$w = a(V_{\text{w}} - V_{\text{leak}}) \tag{3.6}$$

whose substitution modifies Eq. 3.5 as

$$-\tau_{\text{w}} \frac{\mathrm{d}V_{\text{w}}}{\mathrm{d}t} = a(V_{\text{w}} - V_{\text{mem}}) \tag{3.7}$$

Further, as $\tau_{\text{w}} = C_{\text{w}}/g_{\text{w}}$, one can solve to obtain

$$-C_{\text{w}} \frac{\mathrm{d}V_{\text{w}}}{\mathrm{d}t} = g_{\text{w}}(V_{\text{w}} - V_{\text{mem}}) \tag{3.8}$$

In the realized adaptation term, Eq. 3.6 and Eq. 3.8 are implemented. Similarly, for the exponential term (last term of Eq. 3.4), the circuit exploits the subthreshold MOS dynamics to generate the exponential current dependent on the membrane potential V_{mem}. The exponential slope factor Δ_{T} as well as the scaling parameter $g_{\text{leak}}\Delta_T$ in the model are determined by the transistor dynamics.

3.2 Specifications and Parameter Ranges

The target specifications of the designed neuron circuit and the tunable range of its parameters have been identified from a selected set of computational studies [36, 42, 99–109]. The selection has been done on the basis of model networks that are expected to be realized on the prototypes and wafer-scale systems. They have been compiled by Paul Müller [110] and reproduced here in Table 3.1.

variable	min.	max.	unit
τ_{mem}	7	50	ms
τ_{syn}	1	100	ms
τ_{ref}	0	10	ms
V_{leak}	-100	-56	mV
V_{thresh}	-57	-40	mV
V_{reset}	-72.5	-46	mV
$E_{\text{rev,E}}$	0	0	mV
$E_{\text{rev,I}}$	-90	-70	mV
a	-11	56	nS
b	0	250	nA
τ_{w}	16	600	ms
Δ_{T}	0.8	5.5	mV

Table 3.1: The selected set of parameter ranges collected from a number of computational modeling studies [110]. These define the target specifications for the neuron circuit.

In the BrainScaleS model the hardware dynamics are accelerated. The voltage level is dictated by the used supply voltage and designed circuits. It can be scaled with a chosen factor α_{v} and shifted with an offset ω_{v}. Denoting the speed-up factor as α_{t}, the hardware voltages can be related to the biological potentials as

$$V_{\text{hw}}(t) = V(\alpha_t) \cdot \alpha_{\text{v}} + \omega_{\text{v}} \tag{3.9}$$

The target time constants in the hardware domain are scaled by α_{t} fixed to one thousand times. From Eq. 3.9 one can derive the hardware-based conductances leading to

$$g_{\text{hw}} = \frac{C_{\text{hw}}}{C_{\text{bio}}} \alpha_{\text{t}} + g_{\text{bio}} \tag{3.10}$$

The equivalent hardware conductances, e.g., the leak conductance g_{leak} as well as the subthreshold adaptation conductance g_{a} are obtained from the biological quantities (C_{bio}, g_{bio}) using Eq. 3.10.

3.3 MOS Devices and the 65 nm CMOS Process

The neuron circuits have been designed in a low-K 1P9M 65 nm low power digital CMOS process. The technology offers 1.2 V thin oxide (core) devices as well as 2.5 V thick-oxide (I/O) transistors. For the core devices, a number of variants are available. For example, devices with high, low and standard threshold voltages V_{th}, as well as those for high speed are provided. The technology features 9 metal layers and a single poly layer and offers two different metal capacitors, namely the Metal-Insulator-Metal Capacitors (MIMCAP) and the Metal-Oxide-Metal (MOMCAP). Monte Carlo and Corner models are available to simulate device mismatch and the process corners.

The previous neuron designs for BrainScaleS hardware used a 1.8 V 180 nm CMOS process [99]. The neuron circuits were designed with an acceleration factor of 10^4 or faster. With a change in technology node and different time constants (due to a different acceleration factor) new circuit architectures were to be explored and evaluated in the current 1.2/2.5 V 65 nm CMOS process.

The neuron circuit in this thesis evaluates new circuits inspired from previous neuron designs and targets the specification ranges listed in Table 3.1. Wherever the design allows, a low-voltage (1.2 V) solution is targeted with the additional benefit of reduced silicon area (thin-oxide devices). As a whole, it remains a combination of both supplies (1.2/2.5 V) designed using thin- and thick-oxide devices.

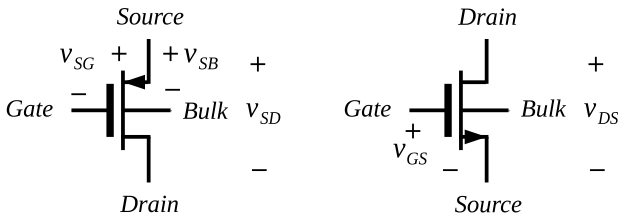


Figure 3.1: The voltage designations and symbols for PMOS (left) and NMOS (right) used in this thesis.

Fig. 3.1 shows a PMOS (left) and NMOS device (right) where the voltage designations are marked, as used in the course of this thesis. The four terminals are marked together with their potential difference labels. A distinction between 1.2 V (core) or 2.5 V (I/O) devices is not explicitly shown by the transistor symbols and is to be understood from the power supply lines, unless otherwise stated. Bulk terminals in schematics are only shown, when they are *not* tied to ground (for NMOS) and voltage supply (for PMOS).

The output characteristic curves of the minimum length transistor with a device threshold (V_{th}) of about 0.52 V is shown in Fig. 3.2a. For $V_{\text{GS}} > V_{\text{th}}$, if $V_{\text{DS}} < V_{\text{GS}} - V_{\text{th}}$, the device is in linear (ohmic/triode) region where the current is expressed by

$$I_{\mathrm{D}} = \mu_{\mathrm{n}} C_{\mathrm{ox}} \frac{W}{L} \left[(V_{\mathrm{GS}} - V_{\mathrm{th}}) V_{\mathrm{DS}} - \frac{V_{\mathrm{DS}}^2}{2} \right] \tag{3.11}$$

where μ_{n} is the carrier mobility of an NMOS device, C_{ox} is the gate oxide capacitance per unit area and W/L is the ratio of channel width and length. As $V_{\mathrm{DS}} \ll V_{\mathrm{GS}} - V_{\mathrm{th}}$, the transistor enters a deep triode region where the drain current is

$$I_{\mathrm{D}} = \mu_{\mathrm{n}} C_{\mathrm{ox}} \frac{W}{L} (V_{\mathrm{GS}} - V_{\mathrm{th}}) V_{\mathrm{DS}} \tag{3.12}$$

which emphasizes a more linearized relationship between the drain current and the drain potential with respect to the grounded source. For the sake of simplicity only NMOS case is discussed. The threshold voltage V_{th} of a device is modeled as

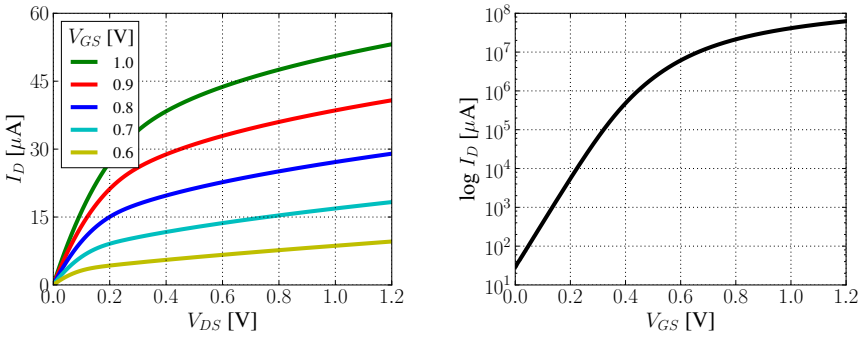
$$V_{\mathrm{th}} = V_{\mathrm{th0}} + \gamma(\sqrt{|2\phi_{\mathrm{F}} + V_{\mathrm{SB}}|} - \sqrt{|2\phi_{\mathrm{F}}|}) \tag{3.13}$$

where V_{th0} is the threshold voltage at V_{SB} = 0, and is a function of the manufacturing process. $\gamma = \sqrt{2q\varepsilon_{\mathrm{Si}} N_{\mathrm{A}}/C_{ox}}$ is the body-effect coefficient and $\phi_{\mathrm{F}} = \frac{kT}{q} ln \frac{N_{\mathrm{A}}}{n_{\mathrm{i}}}$ is the Fermi potential, q is the electron charge, N_{A} is the doping concentration of the substrate, $\varepsilon_{\mathrm{Si}}$ is the dielectric constant of silicon and n_{i} is the carrier concentration of intrinsic silicon.

For analog design we mostly bias the transistors in saturation region where $V_{\mathrm{DS}} \geq V_{\mathrm{GS}} - V_{\mathrm{th}}$. The drain current there can be expressed as

$$I_{\mathrm{D}} = \frac{\mu_{\mathrm{n}} C_{\mathrm{ox}}}{2} \frac{W}{L} (V_{\mathrm{GS}} - V_{\mathrm{th}})^2 (1 + \lambda \cdot V_{\mathrm{DS}}) \tag{3.14}$$

where λ is the channel length modulation parameter inversely proportional to the



(a) I_{D} vs. V_{DS} with swept V_{GS}.

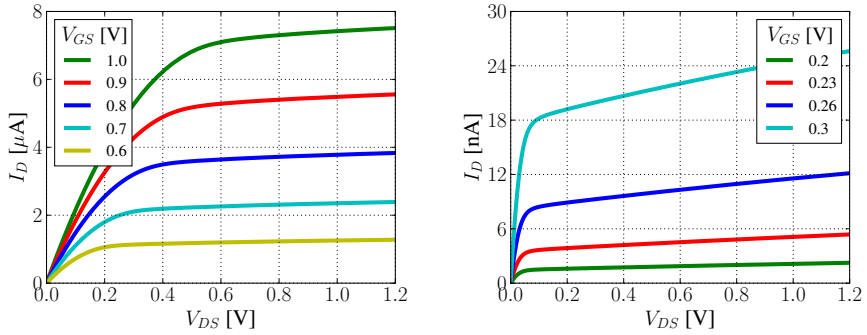
(b) log I_{D} vs. V_{GS} with fixed V_{DS}.

Figure 3.2: The simulated output and input characteristics of a short-channel (minimum length) NMOS device.

channel length ($\lambda \propto \frac{1}{L}$). The larger λ is evident from the slope of the curves for various values of V_{GS} shown in Fig. 3.2a. Fig. 3.2b shows the input characteristic

at a fixed value of V_{DS} = 0.5 V and plotted on the log-scale. The V_{th} of the device is 0.52 V, below which the device exhibits an exponential increase in current with an increase in gate-source potential V_{GS}.

As the channel length is increased from 60 nm to 1 µm, λ decreases and the saturated output characteristic curves show a reduced slope, as shown in Fig. 3.3a. A longer than minimum channel length is therefore a typical choice for analog circuits.



(a) Biased above threshold, $V_{GS} > V_{th}$. (b) Biased below threshold, $V_{GS} < V_{th}$.

Figure 3.3: The output characteristics of a long-channel NMOS device.

The output characteristics of a long-channel NMOS device biased below threshold voltage ($V_{GS} < V_{th}$) is further shown in Fig. 3.3b. The device threshold here is 0.38 V and the plot shows four curves with V_{GS} between 0.2 V and 0.3 V. It shows an exponential relationship of the drain current as a function of gate-source potential. Notice how the current increases linearly for up to a U_T of V_{DS}, where it is proportional to $\frac{V_{DS}}{U_T}$ for a fixed V_{GS}. As V_{DS} is increased more than 3 – 4U_T, the current saturates. In subthreshold or weak inversion the drain current [111] of a MOS device is given as

$$I_D = I_0 e^{\frac{V_{GB}-V_{th0}}{nU_T}} (e^{\frac{-V_{SB}}{U_T}} - e^{\frac{-V_{DB}}{U_T}}) \tag{3.15}$$

where $I_0 = 2n\mu C_{ox}\frac{W}{L}U_T^2$. The parameter n is a technology parameter equivalent to $\frac{C_{ox}+C_{j0}}{C_{ox}}$. Here C_{j0} is the junction-depletion capacitance per unit area of a reversed bias diode (0 V bias), specified in units of fF/µm^2. It increases as the technology nodes are scaled and is usually between 1.5 (old technologies like 0.8 µm CMOS) to 1.85 (45-nm CMOS) [88]. For the 65 nm CMOS process node, the estimated value is around 1.8. For details one can look into the predictive technology model cards [112].

If the bulk and source terminals are shorted, i.e., V_{BS} = 0, Eq. 3.15 is reduced to

$$I_D = I_0 e^{\frac{V_{GS}-V_{th}}{nU_T}} (1 - e^{\frac{-V_{DS}}{U_T}}) \tag{3.16}$$

when $V_{\text{DS}} > 4U_{\text{T}}$, this approximates to

$$I_{\text{D}} = I_0 e^{\frac{V_{\text{GS}}-V_{\text{th}}}{nU_{\text{T}}}} \tag{3.17}$$

Since a MOS device acts as a voltage controlled current source, its transconductance in saturation region is defined as

$$g_{\text{m}} = \frac{\partial I_{\text{D}}}{\partial V_{\text{GS}}}\Big|_{V_{\text{DS}}} \tag{3.18}$$

$$= \mu_{\text{n}} C_{\text{ox}} \frac{W}{L}(V_{\text{GS}} - V_{\text{th}}) \tag{3.19}$$

$$= \sqrt{2\mu_{\text{n}} C_{\text{ox}} \frac{W}{L} I_{\text{D}}} = \frac{2I_{\text{D}}}{V_{\text{eff}}} \tag{3.20}$$

where $V_{\text{eff}} = V_{\text{GS}} - V_{\text{th}}$. Similarly the output conductance g_{DS} is defined as

$$g_{\text{DS}} = \frac{\partial I_{\text{D}}}{\partial V_{\text{DS}}} = \lambda I_{\text{D}} \tag{3.21}$$

In the triode region the transconductance and output conductance are

$$g_{\text{m}} = \mu_{\text{n}} C_{\text{ox}} \frac{W}{L} V_{\text{DS}} \tag{3.22}$$

$$g_{\text{DS}} = \mu_{\text{n}} C_{\text{ox}} \frac{W}{L}(V_{\text{eff}} - V_{\text{DS}}) \tag{3.23}$$

as the transistor moves to deep triode region, the negative V_{DS} in Eq. 3.23 drops out, giving us a linearized relationship.

The MOS gate oxide capacitance contribution varies with the biasing conditions. When $V_{\text{GS}} > V_{\text{th}}$ a channel is formed between the drain and source terminals and the total gate capacitance is

$$C_{\text{ox}} = C'_{\text{ox}} \cdot \; WL \tag{3.24}$$

where WL is the device area and $C'_{\text{ox}} = \varepsilon_{\text{ox}}/t_{\text{ox}}$ is the oxide capacitance per area with $\varepsilon_{\text{ox}} = \varepsilon_{\text{r}}\varepsilon_0$. The relative dielectric constant of SiO_2 is ε_{r} equal to 3.9, and ε_0 is the vacuum permittivity equal to 8.85×10^{-18} F/μm. Fig. 3.4 shows the capacitance contribution of a core (thin-oxide) device, sized to match an ideal capacitor of 1 pF. Note the gate-oxide contribution varies non-linearly with the biasing conditions. However, when biased in inversion region the capacitance contribution is nearly linear. For an I/O device, due to its greater oxide thickness, the capacitance contribution is about half of the core devices. Further discussion on MOS gate capacitors can be found in Sec. 4.7.

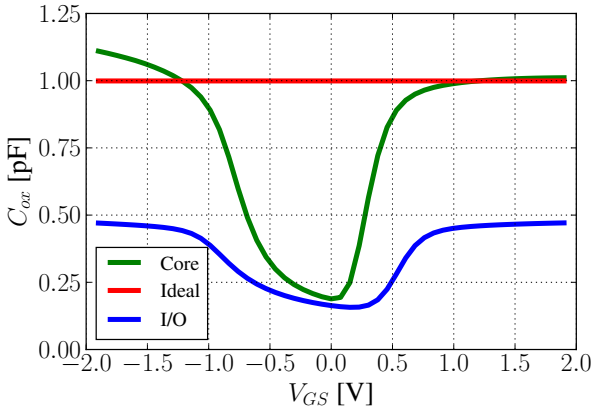


Figure 3.4: The capacitance to gate voltage curves for core (thin-oxide) device compared to an ideal capacitor, and an I/O (thick-oxide) device of the same size.

3.4 Prototype Chips

A total of three prototype chips have been designed in this thesis to test the neuron circuit arrays. In the course of this thesis, the chips are labeled *DLS-1*, *DLS-2* and *DLS-3* – where the number indicates the tape-out sequence. The first two chips implement the LIF neuron model, while the third one implements the AdEx model. All three prototypes have a digital backend implemented together with the ANC and the PPU, and an upstream software programmability for chip parameters (biases) has been provided. All prototypes have had routing capabilities to realize small networks – and therefore the basic functionality of the larger enhanced chip per se.

The *DLS-1* chip is a 1.7 mm $\times$ 2.2 mm MPW run, with a total of 64 neurons in the array. Being the first prototype in 65 nm CMOS technology, the designed circuits take inspiration from the HICANN neuron circuits. A notable shortcoming is however the architecture of the synaptic input, which targets a novel architecture. The circuit is not robust to device mismatch, which make the synaptic inputs unusable. The other sub-circuits within the neuron are qualitatively tested. A die photograph of the *DLS-1* chip is shown in Fig. 3.5. In order to correct the errors encountered in *DLS-1*, a revised synaptic input architecture has been designed in a second mini@sic tape-out, which also fixes other circuits of the ANC. A 1.9 mm $\times$ 1.9 mm chip is designed with an overall similar chip architecture and no new major components. The chip features 32 columns within the ANC, and therefore 32 neurons in the array. The circuit has been measured in considerable detail [79], described further in Chapter 4.

The third prototype chip *DLS-3* enhances the neuron circuit significantly, emulating the AdEx model, multiple compartments and a conductance-based reset. It features digital control for the configurable timing of refractory period

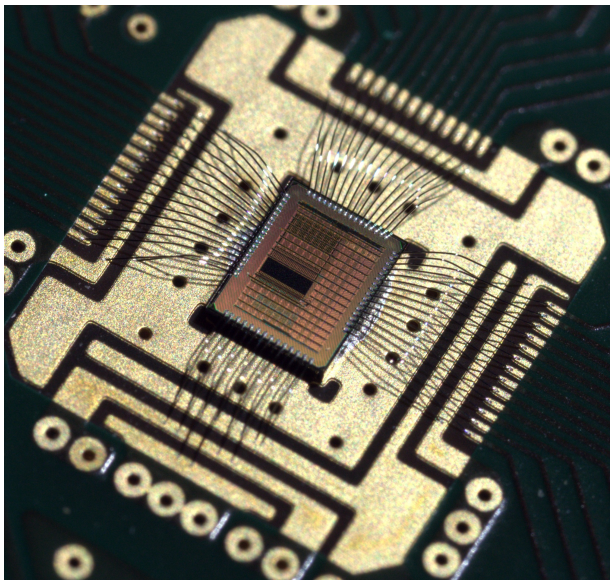


Figure 3.5: The first prototype of the HICANN-DLS chip bonded on a setup daughterboard. Photo by Matthias Hock.

and adaptation pulses. The author implemented the AdEx analog neuron, while multi-compartment/conductance-based reset have been implemented by Johannes Schemmel. The digital control for the neuron is an implementation from Gerd Kiene. During the measurement of the *DLS-3* chip, significant crosstalk has been detected that affects the neuron operation during certain firing regimes. Details are provided in Chapter 5.

3.5 Measurement Framework

The measurement framework comprises of a daughterboard that directly bonds the chip die and is mounted atop the main setup carrier PCB. The PCB hosts a Xilinx Spartan-6 XC6SLX150T FPGA board equipped with DDR3 SDRAM memory. The FPGA communicates with the digital chip interface over a SerDes link for control and event data. The FPGA also communicates with a host PC via a USB 2.0 link. The data communicated with the chip is first buffered into the DDR3 memory to maintain precise timing control. The system can be operated up to a clock frequency of 500 MHz [83]. The FPGA board also hosts a 12-bit 125 MSps ADC. The schematic diagram in Fig. 4.2 shows the arrangement where the digital interface of the chip communicates with the FPGA on the PCB. The description of the chip architecture is described in Chapter 4.

On the setup PCB power supplies of 2.5 V and 1.2 V are derived from low drop-out (LDO) voltage regulators. The debug outputs from the chip are selected by three SP3T analog switches, and are digitized by the Flyspi ADC via an on-

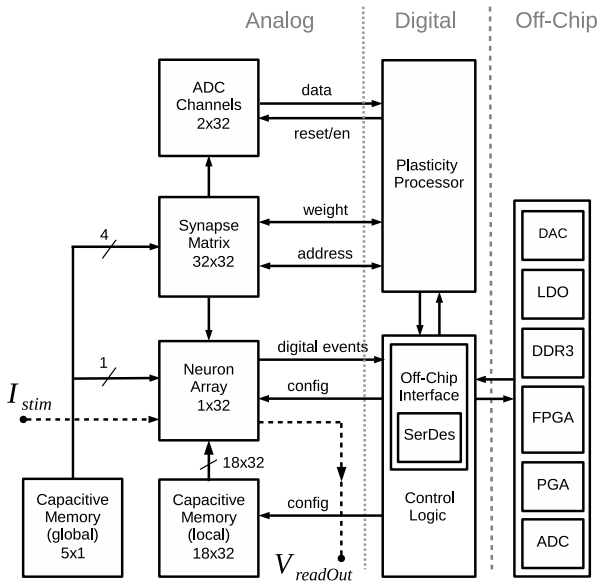


Figure 3.6: Architecture of the *DLS-2* chip together with the off-chip components integrated on the PCB.

board programmable gain amplifier. The board hosts sixteen 12-bit DAC channels that are interfaced to the Flyspi FPGA via the I2C interface for input settings. The voltage outputs of three DAC channels are used to generate bias currents for the capacitive memory global current biases. The board further hosts LVDS transceivers for communication between the chip's digital backend and the FPGA board. The board has been designed by Matthias Hock and shown in Fig. 3.7.

In the revised board for the third prototype, more features have been added. This board has been designed by Korbinian Schreiber and is shown in Fig. 3.8. A JTAG interface is featured to program the on-chip PLL. A fully differential buffer is added to test the MADC directly using the external interface. To support network access, a Gigabit Ethernet transceiver is integrated onboard. For characterization of the chips, the Keithley 2635B sourcemeter is used to measure or source nano to microampere currents. This is, for example, useful to record OTA output currents, their residual output offsets on the membrane or inject external stimulus current in the membrane. The analog voltage measurements are taken using LeCroy Wavesurfer 44Xs 400 MHz oscilloscope in conjunction with LeCroy active probes, e.g., ZS1000 with a specified impedance of 0.9 pF$\|$1 MΩ. The Hewlett Packard 50 MHz HP8116A is used as a signal generator for the large and small signal measurements, for example to test the read-out buffer.

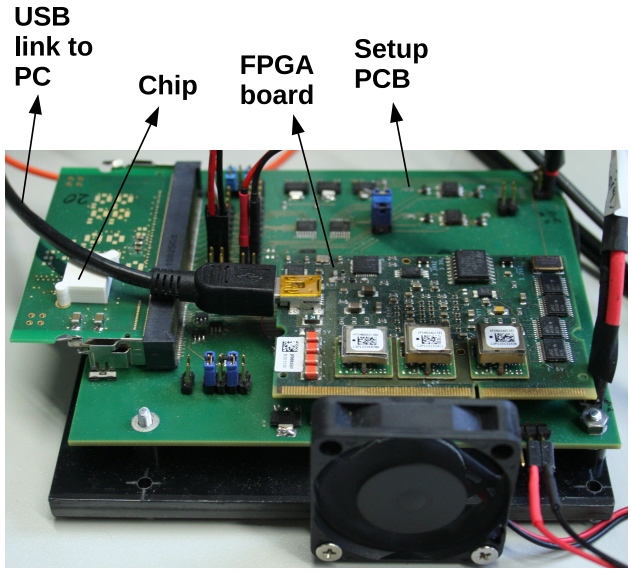


Figure 3.7: The chip measurement setup used to measure the first two prototypes of the chip. Shown in the figure is a setup PCB that hosts the chip carrier board as well as the FPGA board that communicates with the PC over a USB link [93].

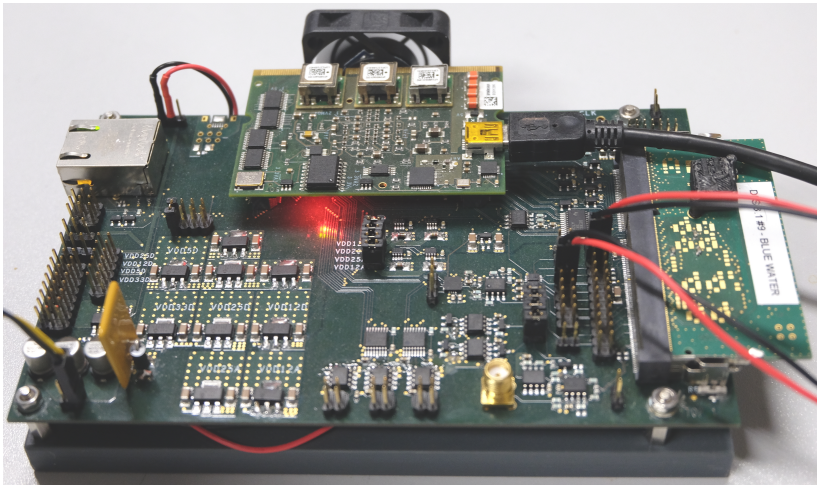


Figure 3.8: The enhanced measurement board designed for the third prototype. Photo by the Author.

3.6 Calibration

The design of the neuron ensures that all individual sub-circuits can be calibrated [113] against the non-ideal effects, caused for example due to device mismatch or corner effects. This is possible due to the presence of tunable Capmem cells and it increases the accuracy of individual circuits in the neuron array. Calibration is performed at two levels. Initially, the pre tape-out netlist is calibrated using Monte-Carlo device models, for example as described in [110]. Furthermore, during measurements the entire array is calibrated and a database is maintained for every available die under use. For the chip prototypes described in this thesis, the pre tape-out calibration has been verified by Paul Müller, whereas the post tape-out calibration for *DLS-2* chip has been done by Yannik Stradmann. Calibration is done by taking polynomial fits for the individual response of the circuit blocks in the entire array, as described in [79].

Chapter 4

Emulation of the Leaky Integrate and Fire Model

This chapter describes the emulation of the leaky integrate-and-fire model. The implemented neuron circuit as well as the design of individual subcircuits are described and the results from chip measurements as well as circuit simulation are presented.

As already described in the previous section, the subthreshold dynamics of the leaky integrate and fire model with current based synapses is described as

$$C_{\text{mem}} \frac{\mathrm{d}V_{\text{mem}}}{\mathrm{d}t} = -g_{\text{leak}} \cdot (V_{\text{mem}} - V_{\text{leak}}) + I_{\text{synExc}} + I_{\text{synInh}} + I_{\text{stim}} \tag{4.1}$$

where C_{mem} represents the membrane capacitor, g_{leak} and V_{leak} are the leak conductance and leak potential, I_{synExc} and I_{synInh} represent the incoming integrated excitatory and inhibitory currents from the synaptic inputs, while I_{stim} denotes the possibility of an externally injected current which can stimulate the membrane.

4.1 Neuron Circuit

The circuit conceived to implement the LIF neuron model of Eq. 4.1 is shown in Fig. 4.1. The left half of the schematic shows the synapse array which is external to the neuron circuit, while the right side sketches the neuron circuit. In the columnar ANC architecture of the HICANN-DLS chip, a multitude of synapses give out current to a single neuron circuit. The schematic shows a single synapse column feeding short current pulse events on two synaptic lines labeled I_{synExc} and I_{synInh}. These lines relay input events to a single neuron circuit in every column. The output stage of the synapse is a 6-bit Digital-to-Analog Converter (DAC) that modulates the size (amplitude) of these pulse events, 4 ns in duration. At the neuron side, the two synaptic input subcircuits, an excitatory and an inhibitory one integrate these current pulses, before integration onto the membrane capacitor C_{mem}. Every incoming pulse event increases the integrated membrane potential V_{mem}. A

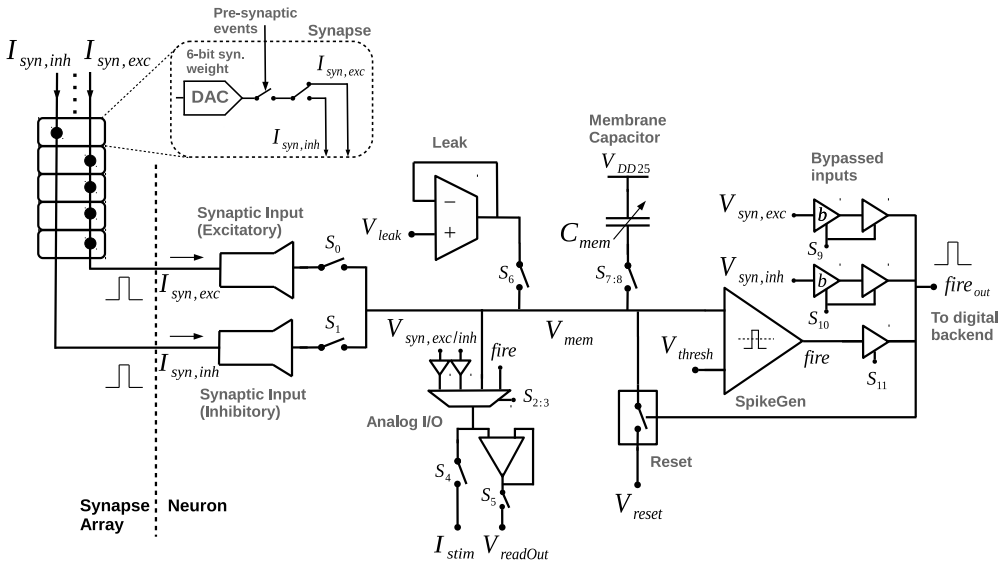


Figure 4.1: The full circuit schematic of the implemented LIF neuron model.

transconductor in unity gain feedback models the leak circuit and pulls the membrane towards a leak potential V_{leak}. A pulse generating circuit labeled SpikeGen generates a digital output pulse *fire*, as the membrane potential reaches a voltage threshold V_{thresh}. The pulse marks the emission of a single spike event, and at the same time resets the membrane V_{mem} to the reset potential V_{reset} via the Reset circuit. Along with membrane reset, this reset circuit also adds the refractory period duration τ_{refr}, during which the membrane is clamped to V_{reset}.

A debug buffer amplifier reads out the input synaptic activity or the membrane potential externally, at the pin labeled V_{readOut}. Another pin labeled I_{stim} can stimulate the membrane with an externally applied current, or hold it at a certain specified potential from the external environment. The debug block is labeled Analog I/O in Fig. 4.1.

Notice the presence of twelve different switches labeled S_{0-11} in the neuron schematic. These either disconnect the individual circuit blocks from the membrane V_{mem} or act as select lines, and are controlled digitally through the digital back-end of the chip. Finally the output fire_{out} pulse that goes to the digital back-end, may also arrive from the *bypassed* synaptic inputs via tri-state inverters controlled by switches $S_{9,10}$. This happens when a bypass mode is enabled, where the analog current integration is disabled and the incoming synaptic input pulses from synapse array (column) directly drive the tri-state buffers to generate output pulse events.

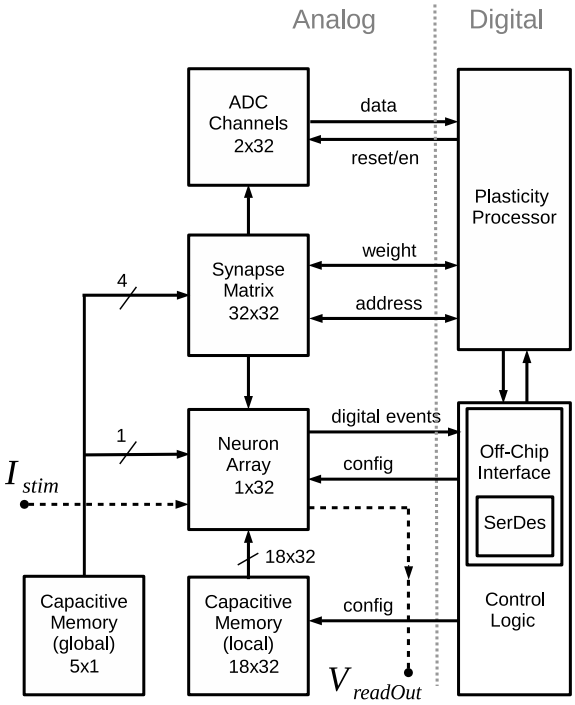


Figure 4.2: The simplified architecture of the second HICANN-DLS prototype.

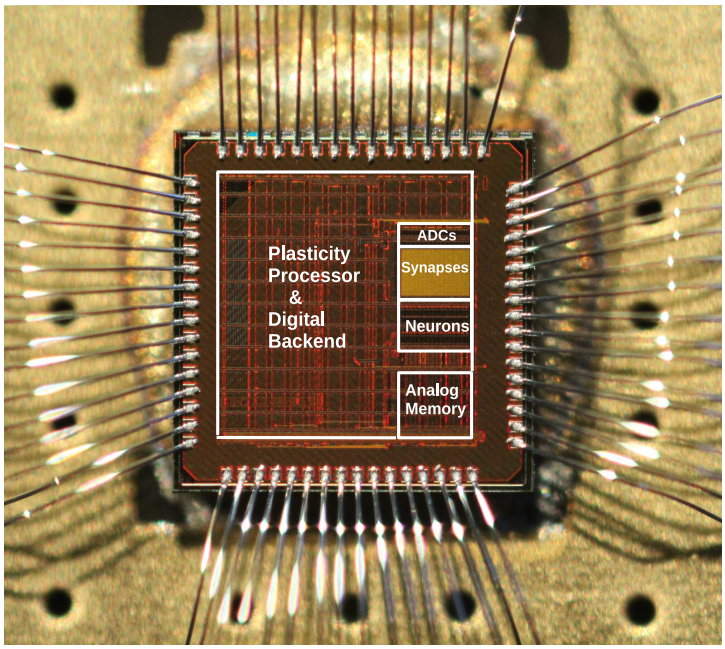


Figure 4.3: The micrograph of the second HICANN-DLS prototype chip (*DLS-2*). Photo by the Author.

4.2 Prototype Chips

The emulated leaky integrate-and-fire model circuit described in Sec. 4.1 has been developed over two prototype chips. The schematic of Fig. 4.1 has been identical for both prototypes, however, internally the sub-circuits have had significant differences. The second tape-out improved and fixed the circuits of the first version. The following changes have been made between the two prototypes:

- The architecture and sub-circuits of the synaptic input
- The circuit designed for the leak OTA
- The number of integrated neurons on the chip
- Re-adjustment of achievable refractory period range

The architecture of the synaptic input circuit was completely revised, whereas the leak OTA was improved. The refractory period range adjustment was a minor change. The number of integrated neurons were changed mainly because *DLS-1* was an multi-project wafer (MPW) run, while *DLS-2* was a mini@sic where the allowed die area is limited.

In both prototypes, the neuron array was embedded in the ANC that comprised of a synapse matrix, the capacitive memory parameter arrays as well as correlation ADCs. This architecture in its simplified form is shown in Fig. 4.2. An array of 32 neurons is connected to a 32×32 synapse matrix. The ANC has in total 32 columns, one per neuron, in which each neuron takes 18 individual (local) parameters and one globally tunable parameter. The capacitive memory that stores these biases is located right at the bottom of the neuron array. At the top each column in synapse matrix gives out current pulse events on two synaptic lines, each for excitatory and inhibitory current. The system also features an implementation of the STDP rule, for which it stores two analog voltages inside each synapse on two capacitors. These two voltages are digitized by the Correlation ADC (CADC) channels shown on the top edge of synapse matrix. The digitized weights are read by the plasticity processor which implements the learning rule and modifies the synaptic weights stored in a 6-bit SRAM inside each synapse accordingly. In this prototype, the digital output events generated by the neuron were taken off-chip via the SerDes to the FPGA, from where they were routed back in the synapse matrix. All data transfer is carried out in the form of packets using OMNIBUS [82]. The packets encode synaptic addresses, synapse enables as well as spike events. The die micrograph of the taped-out chip is shown in Fig. 4.3.

The following sections describe the individual sub-circuits of the taped-out neuron circuit, their design, simulation and measurement results as well as the modifications between the prototype chips.

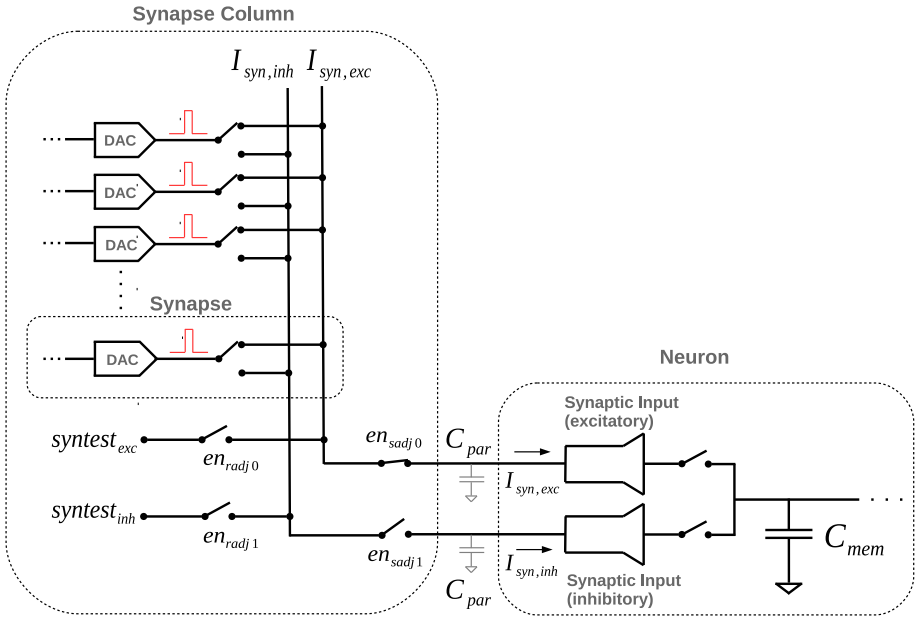


Figure 4.4: The synaptic event pathways between the synapse array and the synaptic inputs of each neuron.

4.3 Synaptic Input

The synaptic input circuit provides exponential synaptic dynamics to incoming excitatory or inhibitory current pulse events from shared synaptic lines, prior to their integration on the neuron membrane. The synaptic event path between the synapse array and the synaptic inputs of the neuron is visually illustrated in Fig. 4.4. The output stage of each synapse circuit within the synapse array consists of a 6-bit DAC, whose analog output controls the maximum amplitude of the pulse event that each synapse emits. The 6-bit input DAC code is hence the *weight* of the synapse circuit and the amplitude (size) of each pulse is directly modulated by it. The DAC output is connected to either of the two synaptic input lines within the synapse, and depending on the connection the input events are designated as excitatory or inhibitory events. The voltage on these lines (labeled I_{synInh}, I_{synExc}) is 1.2 V, unless the DAC emits a pulse event which pulls it lower. In the latter case, the integrator architecture at the neuron side recovers the voltage back with a time constant. These lines also have substantial parasitic capacitance (labeled C_{par} in Fig. 4.4), which grows with the number of synapses in the column. Further, shown in Fig. 4.4 are switches in the input event path – these are transmission gate switches and are controlled digitally by signals en_{sadj0}, en_{sadj1} for the excitatory and inhibitory inputs. Together with two other switches en_{radj0}, en_{radj1}, these switches help debug the synaptic interface, e.g., in measuring the synaptic current at the output pins $\text{syntest}_{\text{exc}}$ and $\text{syntest}_{\text{inh}}$. Alternately, they facilitate measurements of the input stage of synaptic inputs. In the nominal setting, the switches en_{radj0},

en_{radj1} are kept open, whereas the switches en_{sadj0}, en_{sadj1} are kept closed.

The following subsections first describe the initial architecture of *DLS-1* briefly from the first chip, followed by a more detailed discussion on the modified solution of *DLS-2* chip.

4.3.1 Initial Architecture

The synaptic input architecture needs to first integrate the current pulses using an integrator, before integrating equivalent current onto the neuron membrane. A straightforward architecture to realize the current-based synaptic input, is to first integrate using an opamp based integrator circuit and then convert the output voltage into an equivalent current. The integration stage can simply be a leaky integrator with a floating tunable resistor R_{syn} parallel to the integrating capacitor C_{syn}. This is shown in Fig. 4.5.

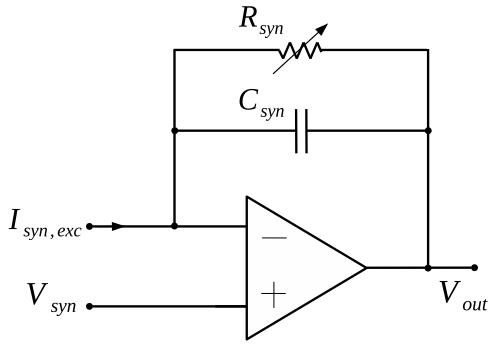


Figure 4.5: A tunable leaky integrator circuit used in many applications.

Such an integrator architecture would however require a metal capacitor or metal-finger capacitors, which are more area consuming on the chip compared to MOS capacitors. Furthermore, to realize the desired time constants either large capacitors or large values of tunable resistor would be required. Typical on-chip capacitors such as MIMCAP or MOMCAPs require a large die area[1], and since the two synaptic inputs may consume only up to 25% of the final neuron area, use of larger metal capacitors is kept to a bare minimum. An alternate is to use MOS based capacitors, which usually reduce substantial area compared to a metal capacitor, but they have certain known non-ideal effects, for example

- their capacitance value varies across different regions with the applied gate voltage,
- to bias them in inversion region, a gate-voltage greater than the MOS threshold voltage V_{th} is required.

[1]This CMOS process utilizes the metal layers 3-to-5 to form MOMCAPs, whereas MIMCAPs block metal layer 7 of the routing stack. To prevent reduction of routing resources, MOMCAPs were not used anywhere in the neuron.

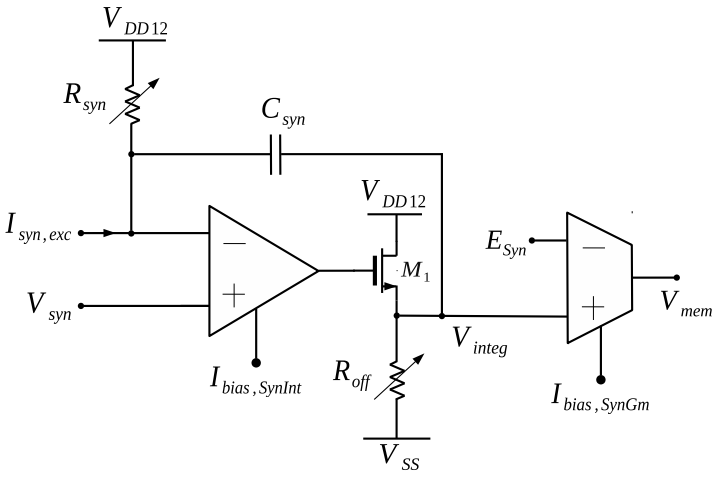


Figure 4.6: The schematic of the synaptic input circuit for *DLS-1* neuron.

These two properties have already been shown in Sec. 3.3. Further discussion on MOS capacitors is deferred until Sec. 4.7. These two considerations for utilization of MOS based capacitor give rise to the synaptic integrator architecture shown in Fig. 4.6. This integrator has several differences. First, it creates a voltage drop between the Gate-Source voltage of the MOS capacitor C_{syn}, since the voltage at node $I_{\text{syn,exc}}$ is 1.2 V, while node V_{integ} is dropped by the drain-source voltage of transistor M_1. Secondly, it uses two grounded tunable resistors, as opposed to a resistor floating between two nodes. The upper resistor R_{syn} tunes the synaptic time constant τ_{syn} (with a fixed capacitor), whereas the lower resistor helps set the DC offset at the output node V_{integ} of the integrator.

While the proposed integrator does work in the ideal case, it has an obvious flaw. The gate voltage of transistor M_1 is actually being driven by the output common-mode of the amplifier. The output common mode is typically set by the design at mid-range of the 2.5 supply, i.e., at about 1.2 V. This 1.2 V gate voltage is applied to the transistor M_1, which ensures a saturation region in the ideal case. However, since the amplifier is in open loop, the input-referred offset voltage between the two amplifier terminals will be amplified by the open loop amplifier gain. This will shift the output common-mode point towards either of the supply rails, depending on the offset polarity. This will in turn shift the output node V_{mem} of integrator towards either of the supplies. The resistor R_{off} in principle could compensate and tune this effect, but it does not cover all statistical samples. To effectively trim the input offset using the parameter V_{syn}, it requires higher resolution than the 10-bit available from the current implementation of voltage parameter cells. The circuit limitations were known from the simulation results prior to chip tape-out.

4.3.2 Amplifier

The amplifier designed for the initial synaptic input integrator was a current mirror OTA shown in Fig. 4.7. There are a few advantages in using this design. First, it does not require a compensation capacitor, which saves significant area. Second, if we derive expressions for gain, bandwidth and slew-rate, we find out how the mirroring factor plays a role on in determining these quantities. The open loop DC-gain of the amplifier is given by

$$A_{\mathrm{OL}} \approx \frac{g_{\mathrm{m1}}}{g_{\mathrm{m3}}} \cdot \frac{g_{\mathrm{m9}}}{g_{\mathrm{DS8}} + g_{\mathrm{DS9}}} = K \cdot \frac{g_{\mathrm{m1}}}{g_{\mathrm{DS8}} + g_{\mathrm{DS9}}} \tag{4.2}$$

where $K = g_{\mathrm{m9}}/g_{\mathrm{m3}}$ is the ratio of transconductances formed by the current mirror transistors M_3,M_9. The unity gain bandwidth is therefore $\omega_{\mathrm{u}} = K \cdot g_{\mathrm{m1}}/C_{\mathrm{L}}$, where C_{L} is the output load capacitance. Further, the slew rate $SR = K \cdot I_{\mathrm{tail}}/C_{\mathrm{L}}$, where I_{tail} is the current from tail current source formed by transistor M_5. The designed amplifier achieves 58.3 dB DC gain, a unity gain frequency of 1.1 MHz and an input-referred offset of 6.1 mV.

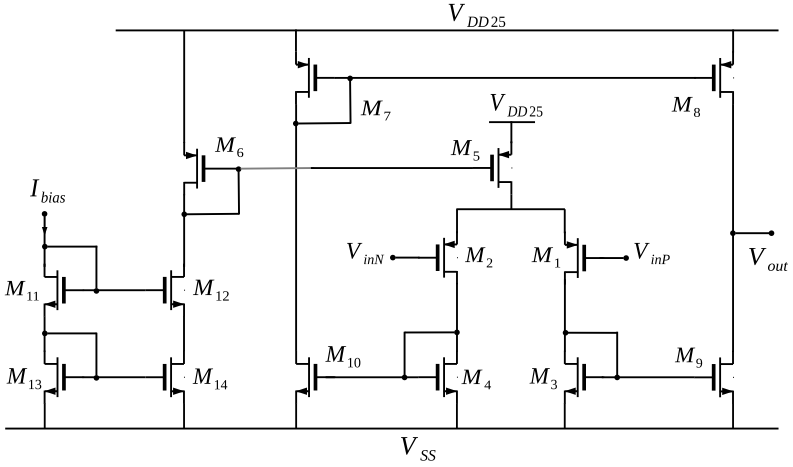


Figure 4.7: The amplifier used in the synaptic input circuit in *DLS-1* neuron.

4.3.3 Tunable Resistor Architectures

A literature study was carried out before implementing the MOS tunable resistors for HICANN-DLS prototypes. There is a plethora of tunable CMOS resistor architectures, most of which linearize the resistive range, as well as enhancing the available tuning range. Both floating resistors [114–118] as well as their grounded counterparts are proposed [119–121] with varying area and power consumption figures. For a neuromorphic large-scale integration application, a resistor which is ultra lower power and consumes minimal area is desirable. Linear tuning in the range of specified synaptic time constants is required. Note that grounded resis-

tor are simple in architecture due to their supply (ground) connection. The initial integrator architecture also requires a grounded resistor as shown in Fig. 4.6.

Implemented Resistor

The resistor architecture for the integrator is inspired from [121]. The architecture mainly relies on the resistance provided by a single transistor in triode region, and linearizes its non-linearity by adding a saturated transistor in parallel, as shown in Fig. 4.8. Transistor M_1 is the main triode device, while M_2, a diode configuration

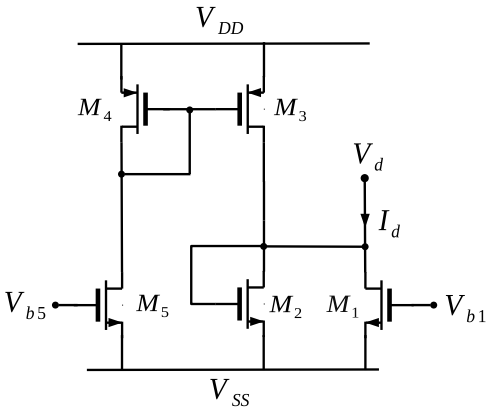


Figure 4.8: The tunable grounded resistor used in the synaptic input circuit for *DLS-1*.

is in parallel and obeys a squared current law equation. If we neglect the branch that steals the current through mirror M_3/M_4, then $I_d = I_{\text{d1}} + I_{\text{d2}}$, where I_{d1} and I_{d2} are currents through the triode transistor M_1 and the saturated transistor M_2. Adding them results in

$$I_{\text{d}} = V_{\text{in}} \left[K_1(V_{\text{b1}} - V_{\text{th1}}) - K_2 V_{\text{th2}}\right] + I_{\text{off}} \tag{4.3}$$

where $K = \mu C_{\text{ox}} \frac{W}{L}$ of each device and $I_{\text{off}} = \frac{K_2}{2 \cdot V_{\text{th}}^2}$ is the offset residual current. This offset current is further canceled by the presence of another saturated device M_5 which steals this current through the current mirror formed by M_3 and M_4, which was initially neglected. Therefore a linear characteristic is obtained. The resulting resistance is then given as

$$R = \left[K_1(V_{\text{b1}} - V_{\text{th1}}) - K_2 V_{\text{th2}}\right]^{-1} \tag{4.4}$$

The terminal V_{d} is the input of the resistor, while V_{b1} allows to tune the resistance. This resistor implements the two tunable resistors R_{syn} and R_{off} in the architecture (see Fig. 4.6). R_{off} is an NMOS based architecture, identical to Fig. 4.8 whereas R_{syn} is a PMOS based complementary implementation.

4.3.4 Modified Architecture

The synaptic input architecture is modified in the second chip prototype *DLS-2* to simplify the design, as well as to reduce area and possibly also power consumption.

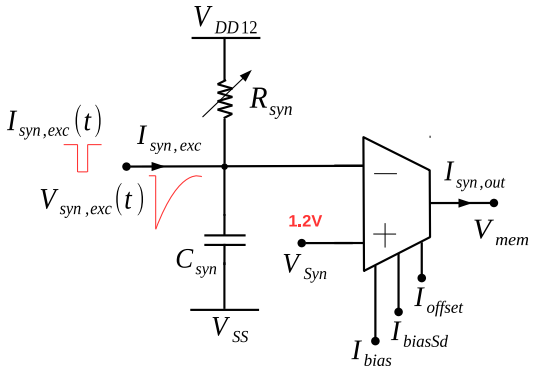


Figure 4.9: The synaptic input circuit schematic designed for *DLS-2*.

This architecture replaces the active integrator with a passive one, and utilizes the parasitic line capacitance to integrate incoming input current pulses. The voltage to current conversion is the same as in the last case, i.e., with the use of a linear source-degenerated transconductance amplifier. The resulting architecture is shown in Fig. 4.9. The parasitic capacitance multiplies with the number of input synapses in each row – which in the *DLS-2* prototype is 32. The total parasitic capacitance for 32 input synapses is approximately 50 fF. Therefore, to ensure the time constants[2], 950 fF per synaptic line is placed as a metal capacitor. The synaptic time constant τ_{syn} is the product of tunable resistor R_{syn} and this fixed (but lumped) capacitor C_{syn}. A final prototype of this chip will be scaled up and the number of input synapses will grow from 32 to 256 per neuron. In that case, the parasitic line capacitance will be at least four times larger contributing 400 fF. The tunable resistor designed for this synaptic input is a novel architecture inspired from the bulk-drain connected devices [122–126] described in the next subsection. The realized transconductance amplifier architecture in Fig. 4.9 also has additional current biases namely, I_{biasOff} and I_{biasSd}. The former cancels the effect of input offset at the output of the OTA, while the latter is used to provide a separate bias for the source degeneration devices, meant to linearize the response. This is further elucidated in Sec. 4.4 which describes the transconductance amplifiers.

Each incoming current pulse event on the synaptic input line I_{synExc} pulls the line voltage lower, proportional to the size of input synaptic pulse. Since the tunable resistor R_{syn} pulls the line up, it recovers the voltage level back with a time constant τ_{syn} to its initial potential of 1.2 V. An event of 10 µA (4 ns long) event drops the line potential by about 100 mV. The incoming input synaptic events and their resulting integrating response on the membrane is shown in the measured

[2]calculated with a 1 pF fixed capacitor

result of Fig. 4.10.

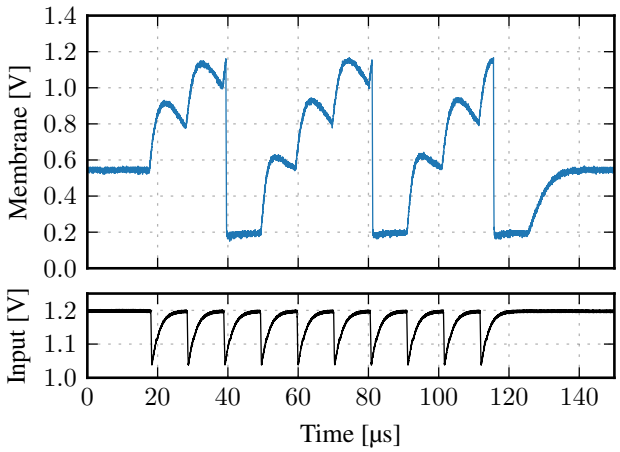


Figure 4.10: Measured results from the *DLS-2* chip [79]: a train of incoming synaptic input events (lower trace) pulls the $I_{\text{syn,exc}}$ low successively, recovered with a time constant τ_{syn}. The upper trace reflects the corresponding response on the membrane.

4.3.5 Bulk Drain Connected Devices

The resistor designed for the modified synaptic input architecture uses the bulk-drain connected devices [122] – where the bulks are connected to the drains, instead of the source terminals. When biased in weak-inversion regime, the output device characteristics of the nominal bulk-source connected PMOS devices express ohmic region only up to a few U_{T}. Compared to this, the bulk-drain connected devices, due to a finite (and controllable) output resistance, exhibit linear behavior up to a few hundred millivolts in their output device characteristics I_{DS}-V_{DS} [122, 124]. This allows for the realization of large value tunable resistive loads [123–126]. The output characteristics of a bulk-drain device are depicted in Fig. 4.11. Each of the curves are plotted for a different value of gate-source potential V_{GS} and therefore represent a controllable resistance. The channel length of this device is 2 µm. A conventional MOS device would have a much longer channel length to implement these large resistances.

The devices are usually implemented in a p-substrate CMOS technology where separate n-wells are embedded in a p-type substrate, which forms the ground plane. To realize these devices with "wrong" bulk connection, instead of the larger substrate, isolated wells are required. One can utilize either n-wells, or p-wells from the triple-well devices, which usually consume significant area given the process rules for triple-wells and the long channel lengths for resistor design. Note that MOS devices from physical design assembly are symmetric, and the convention of

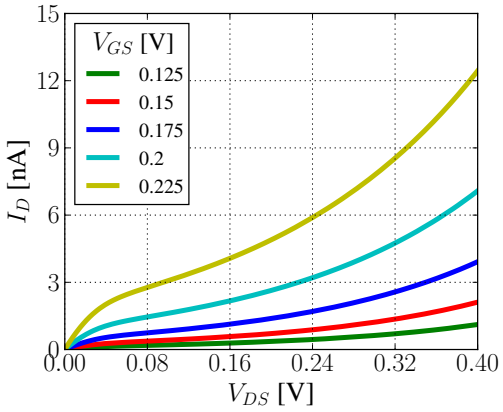


Figure 4.11: The output characteristics of the bulk-drain connected device.

source or drain terminals are based on their relative potential. The source terminal, for a PMOS transistor therefore needs to stay at higher potential than drain (connected to bulk), otherwise the source/drain terminals swap, and the device switches to a nominal mode with a source-bulk connection. Further, in the nominal (source-bulk) configuration, the PMOS (NMOS) are tied to highest (lowest) potential (supply voltage), which ensures the parasitic diode stays shut off. In the bulk-drain connection, the PMOS bulk is not tied to supply, but to the drain, therefore to prevent this diode from turning on, the drain voltage shall stay below the threshold of the parasitic diode. Mathematically $0 < V_{\text{SD}} < V_{\text{th,par}}$. Fig. 4.12 shows a cross section of a bulk-drain connected device where this parasitic diode between drain and bulk inside the n-well is shown. Note that the drain-bulk diode is off due to the shorted connection. As source-bulk voltage $V_{\text{SB}} \neq 0$, the device is prone to

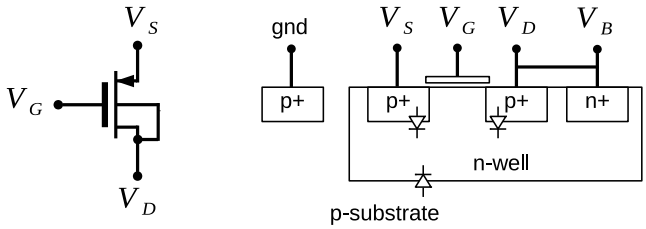


Figure 4.12: The bulk-drain connected PMOS device and its cross-section view during implementation.

body-effect due to a constantly changing V_{th} of the transistor. Using MOSFET device models based on BSIMv4.5 for the given technology, one can demonstrate the variation of threshold voltage V_{th} with varying source-bulk potential V_{SB}, shown in Fig. 4.13. The selected device is a thin-oxide PMOS transistor with dimensions W/L = 0.2/0.4. As V_{SB} increases, the threshold decreases – something the designer

should be aware of, to ensure proper biasing regions. Multiple bulk-drain devices may also be used in series so that the eventual source-drain drop V_{SD} can be reduced. Alternatively, one can arrange two PMOS bulk-drain devices back to back in series, such that the drain terminals of the two devices are directly connected. This prevents the source-drain drop limitation which can possibly swap terminals and allows to realize floating resistors[3]. The former approach has been adopted to realize the resistor in the modified synaptic resistor, while the latter is used to realize a much higher value resistor within the adaptation term, and explained later in Sec. 5.3.1.

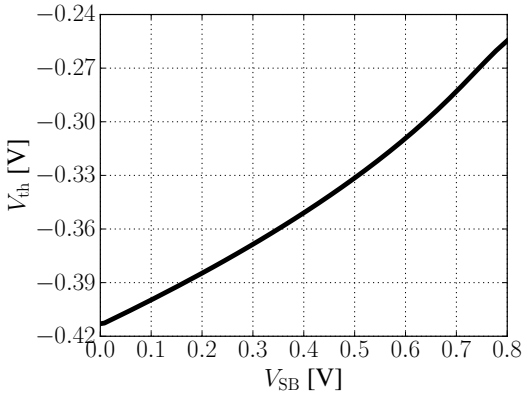


Figure 4.13: The device threshold voltage changes with increasing source-bulk voltage drop.

4.3.6 Synaptic Resistor

The resistor designed in the modified synaptic input architecture is based on a series of bulk-drain connected PMOS devices discussed in Sec. 4.3.5 and shown in Fig. 4.14. Between the two terminals V_{inP} and V_{inN} are four bulk-drain connected PMOS devices labeled $M_{1,2,3,4}$, each of them operating as a bulk-drain connected resistive device. The devices connect their drain to the bulk, instead of the nominal bulk to source connection. Compared to the original concept, they are all biased in linear region rather than subthreshold region, to contribute lower overall resistance. This was necessary to adjust for the requirements of synaptic time constants, given a 1 pF capacitor. All four devices are biased from cascode current mirrors formed by devices $M_{5,6}$ and $M_{1\text{b},1\text{a}}$, $M_{2\text{b},2\text{a}}$, $M_{3\text{b},3\text{a}}$, $M_{4\text{b},4\text{a}}$ respectively. The pull-up devices $M_{1\text{c},2\text{c},3\text{c},4\text{c}}$ set the bias points of each bulk-drain device. When simulated with a 200 mV voltage drop across the terminals and an I_{bias} of 100 nA, all four devices drop 43 to 57 mV across their terminals, each providing between 740 kΩ

[3] such that even if source-drain drop V_{SD} of first device becomes negative, the second device maintains a positive drop.

to 980 kΩ resistance. All devices stay in linear region, as the applied bias is varied from 1 µA to as low as 28 nA. Decreasing further shifts the device regions to subthreshold, eventually driving all four to subthreshold when biased below 20 nA. The measurement results of a single synaptic resistor tuned over its full resistive range are shown in Fig. 4.15. The family of curves plotted are the current versus the applied potential difference across the resistor terminals, by varying the resistor bias current I_{bias}. The equivalent bias current and the achieved resistances are shown. The measurements have been taken by applying a potential difference of 0.2 V, which is the worst case scenario, as the synaptic input line does not typically drop beyond 150 mV for strong synaptic events. Note the presence of finite offset voltage, as the plotted traces cut the 0 nA current at a potential difference of about 15 mV. This is likely due to the supply voltage drop, since a direct measurement through the debug pin measures a resistor pull-up voltage of 1.185 V in the chip. The resistance values start a linear increase as the bias is swept, and show more exponential increase with lower bias currents.

The entire synaptic array, both consisting of inhibitory and excitatory synaptic inputs are characterized next. Fig. 4.16a shows 64 traces of resistors from both synaptic inputs integrated on a single chip. The tunable bias is fixed at a mid-range value, which in this case provides a resistance of about 1.5 MΩ. The traces show a mismatch which is more evident beyond a potential difference of 100 mV. This is mainly due to the variations among individual bias stages, where the current mirrors formed by $M_{\text{Xa,Xb}}$ together with the biasing devices M_{Xc} define the biasing points in individual bulk-drain devices (X represents the respective stage from 1 to 4).

The range of available synaptic time constants τ_{syn} as tuned by the resistor's bias are further measured and plotted in Fig. 4.16b. The figure plots traces from three separate dies for both synaptic input circuits, leading to a total of 192 samples. The bias current in the traces are being swept from a mid-range bias value of 500 nA down to 15 nA. It is evident that the increase in the time constant τ_{syn}

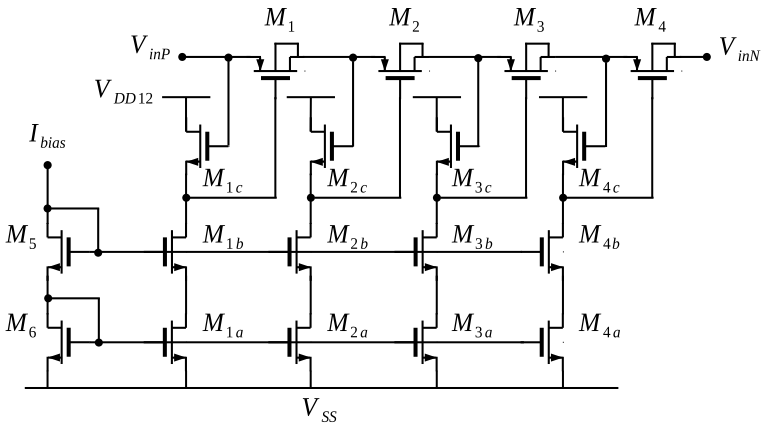


Figure 4.14: The tunable resistor designed for the synaptic input for *DLS-2*.

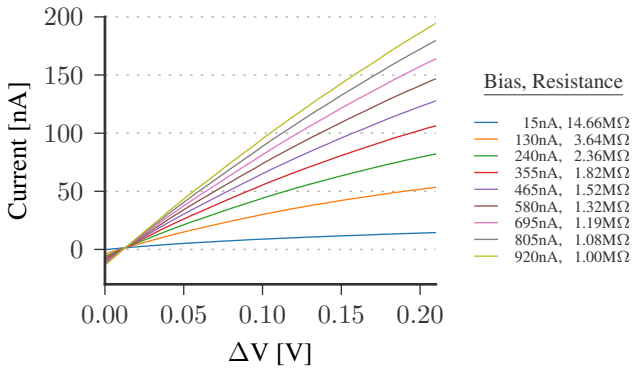
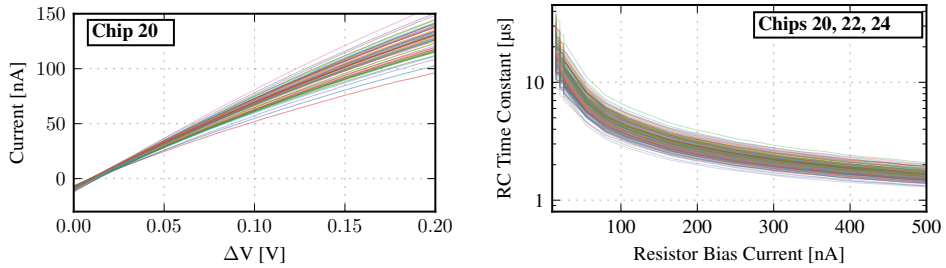


Figure 4.15: Tuning of the synaptic resistor by changing its current bias [79].



(a) A family of characteristic curves of all 64 synaptic resistors on a single chip tuned for their mid-range resistance [79].

(b) Tuning the synaptic time constants by varying the resistor bias current. Data has been acquired from 192 synaptic input circuits from three different dies [79].

Figure 4.16: Variation in synaptic resistor and available time constants among multiple samples.

is more linear up until 150 nA with decreasing bias current, followed by an exponential behavior in the low-current regime. This aligns well with the resistor tuning curves shown in Fig. 4.15, where the achieved resistances start with a linear increase with decreasing bias, followed by an exponential behavior.

In order to derive the range of available synaptic time constants τ_{syn}, a distribution of minimum and maximum synaptic time constants is plotted in Fig. 4.17. It can be seen that longer time constants (set by tuning very low bias currents) have more variation compared to shorter time constants which have a linear increase, as evident from Fig. 4.16b.

4.3.7 Calibration

The input-referred offset of the synaptic OTA can result in unwanted output synaptic current in the absence of input synaptic events. If not compensated, this can cause the membrane to integrate this input ”leakage” current, and lead the neuron to eventually spike. This input offset therefore is the prime candidate to be com-

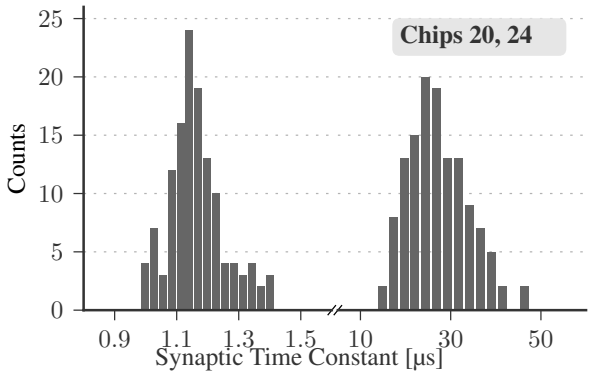


Figure 4.17: A distribution of the minimum and maximum range of the achieved synaptic time constants [79].

pensated for. Since the second OTA terminal is a tunable voltage parameter (from Capmem cells), this can be reasonably trimmed within the given 10-bit resolution. Secondly, the synaptic OTAs are also equipped with output offset cancellation that trims this residual current. These biases are labeled I_{biasOff} and shown in Sec. 4.4. The distributions of Fig. 4.18 show the synaptic input leakage current as a result of input offset before and after calibration.

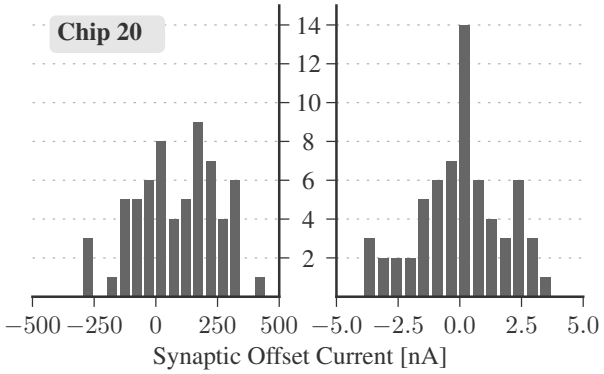


Figure 4.18: Calibration results of 64 synaptic inputs on a single die. Left histogram: offset current with V_{syn} at 1.2V for all OTAs; Right histogram: residual current after individually adjusting V_{syn} and I_{biasOff} [79].

The input offset is calibrated in a two stage process. First, I_{biasOff} is set half way of the bias tuning range and V_{syn} is tuned to minimize the output leakage current. The residual leakage is then fine-tuned for, by using I_{biasOff} by taking a linear fit. The family of curves showing the achievable time constants shown in Fig. 4.16b are next calibrated for. In this case, the only calibratable parameter is the resistor tuning bias, so the curves are fitted with a polynomial which lead to a reduced spread. The pre- and post-calibration results for three different time constants are shown in the distributions of Fig. 4.19. Although the distribution in the 5 µs time

constant is not too large compared to longer ones, as a result of polynomial fitting, the overall spread is reduced among all three time constants. This data is taken from [79] from measurements performed by Yannik Stradmann.

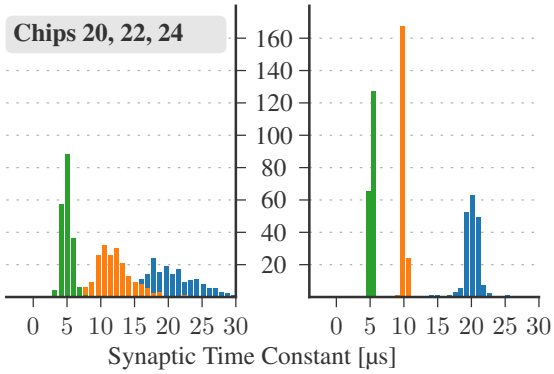


Figure 4.19: Spread of synaptic time constants with a mean of 5 µs, 10 µs and 20 µs. Left: The resistor bias sets three different time constants and plots the statistical variations from three different dies. Right: The resulting time constants are processed through individual polynomial fits, resulting in a reduced spread [79].

4.4 Transconductance Amplifier

The transconductor designed for synaptic input as well as the leak circuit is a source-degenerated OTA architecture. The output current of the OTA has a lin-

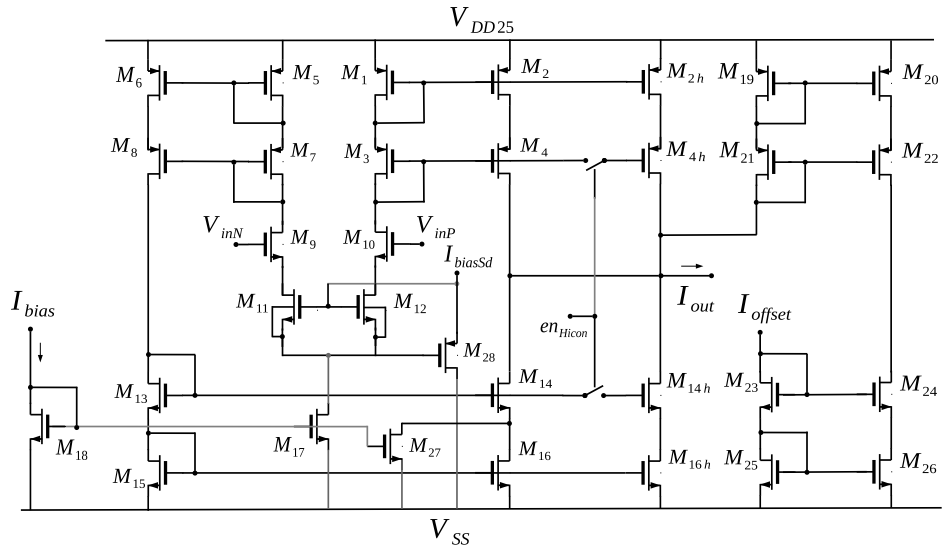


Figure 4.20: The generic schematic of the operational transconductance amplifier architecture used within synaptic input and leak circuits.

Feature	Leak OTA	Syn. OTA	Realized by
Output stage multiplier	✓	-	$M_{2h,4h,14h,16h}$
Input level shifter	✓	-	not shown
Output offset tuning	-	✓	M_{19-26}
Forced offset	-	✓	M_{27}

Table 4.1: The architectural differences and feature set of the two OTAs.

ear dependence on input differential voltage within a limited operational range, as $I_{out} = G_m(V_{in+} - V_{in-})$, where G_m is the OTA transconductance.

The OTA schematic in its generic form is shown in Fig. 4.20. The designed OTA not only strives to widen the input operational range by reducing gain, but also features output offset compensation, output current multiplication, and forced-offset in one branch. The input differential pair is formed by devices $M_{9,10}$, together with the cascode current mirror loads formed by M_{1-4} and M_{5-8}. The devices $M_{11,12}$ act as degeneration resistors controlled by a separate bias current labeled I_{biasSd}. They are implemented in a deep n-well to avoid the body effect. The mirror M_{13-16} completes the symmetric design on the lower side, while $M_{17,18}$ forms the tail current source biased by input current I_{bias}. In parallel to the output stage formed by devices $M_{2,4,14,16}$, there are wider devices labeled $M_{2h,4h,14h,16h}$, which when enabled via en_{Hicon}, give out approximately ten times more output current. This output stage multiplication is a useful feature for high-conductance mode when the membrane needs short time constants. The cascode current mirrors formed by devices M_{19-26}, trim the residual offset current at the OTA output by tuning the parameter bias I_{offset}. Note the presence of another device M_{27} that makes the output stage asymmetric, as it sinks more current in one branch. This is the forced-offset mentioned above, and its purpose is to create output offset in one direction, such that at zero differential voltage input, the output current is positive and non-zero.

The architectures of the two OTAs differ as far as the implementation of these features are concerned. Table 4.1 summarizes the features and their implementation in both OTAs. Note the presence of an input level shifter in the leak OTA. Compared to the synaptic input OTA, which has an input voltage range between 1 V – 1.2 V, the typical leak values are centered around 600 mV. Therefore for a 2.5 V supply, it is reasonable to shift the range up by using a source-follower based level shifter (not shown in Fig. 4.20). The output offset tuning and forced offset is implemented in synaptic input OTA only.

The input referred offset of the OTAs have been simulated using Monte Carlo device models shown in Fig. 4.22. The 1σ variation is 30.9 mV from both OTAs with I_{bias} = I_{biasSd} = 1 µA. Note that the mean for the synaptic OTA is centered around 33.4 mV due to forced offset. The mean of the leak OTA is -1.5 mV. The input offset increases as I_{biasSd} is lowered. For example, at a bias current of 0.4 µA the input offset increases to approx. 36 mV. Lowering further however, makes the distribution non-normal with wider spread and the offset uncalibratable.

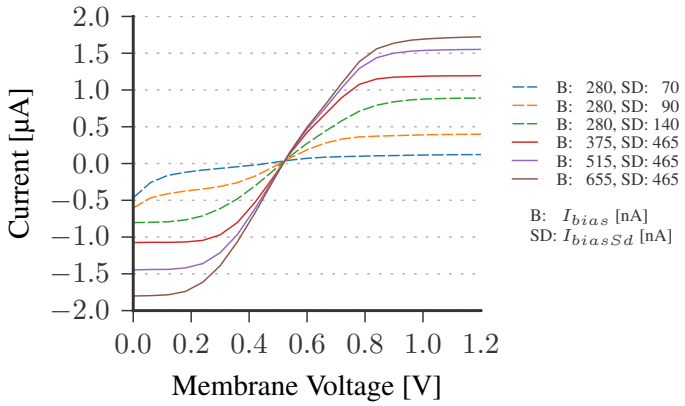


Figure 4.21: The measured traces showing the OTA output current from the leak term versus swept membrane voltage with V_{leak} fixed to 0.55 V [79].

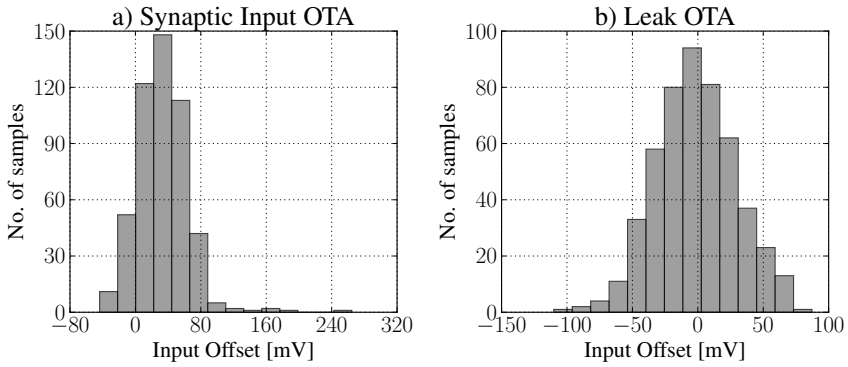


Figure 4.22: The input offset of OTAs used within the synaptic input as well as to realize the leak term.

4.5 Leak Circuit

The leak term in the neuron has been realized using an OTA in unity gain feedback as shown in Fig. 4.23. In this configuration the transconductance G_{m} of the OTA acts as the leak conductance of the neuron circuit. The membrane time constant can then be defined as $\tau_{\text{mem}} = C_{\text{mem}}/G_{\text{m}}$. As described earlier in Sec. 4.4, the two biases I_{bias}, I_{biasSd} help tweak the resulting transconductance. I_{bias} sets the maximum OTA current where its output current saturates, while I_{biasSd} alters the gain to linearize the response. Fig. 4.21, plots measured traces of the output current as a function of membrane voltage, when the leak potential V_{leak} is set at 0.55 V. The traces sweep both biases in the nanoampere range one at a time, i.e., first three curves sweep I_{biasSd} while keeping I_{bias} constant, and the next three sweep I_{bias} while keeping I_{biasSd} constant. It can be seen that small values of degeneration bias result in very flat and wide ranged curve due to little gain. This is a case

where the degeneration resistors formed by transistors M_{11}, M_{12} are pushed into subthreshold region where they contribute a large resistance. The transconductance of the degeneration stage approximates to $G_{\mathrm{m}} = g_{\mathrm{m}}/(1 + g_{\mathrm{m}}R_{\mathrm{SD}})$, where g_{m} is the input pair's transconductance and R_{SD} is the resistance contributed by the source-degeneration transistors. With large degeneration resistors the transconductance curve linearizes itself as $G_{\mathrm{m}} \approx 1/R_{\mathrm{SD}}$. This case is highlighted in Fig. 4.21 when I_{biasSd} is set to its lowest setting.

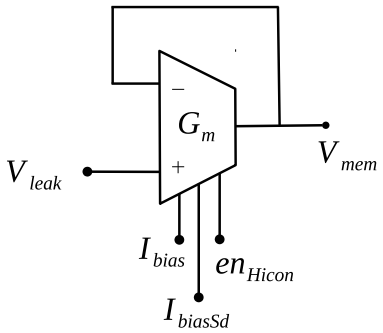


Figure 4.23: The leak term realized using a transconductance amplifier.

Fig. 4.24a shows the transconductance G_{m} as a function of I_{bias}, as I_{biasSd} is swept. At the highest setting of I_{biasSd} = 1 µA, the transconductance varies near-linearly. However, this leads to a reduced linear voltage range, as shown in Fig. 4.21. Decreasing I_{biasSd} below 0.5 µA makes the traces non-linear, such that the transconductance only rises up to $2{\cdot}I_{\mathrm{biasSd}}$ – and further I_{bias} increase reduces the transconductance. In the unity-gain configuration with an applied potential ΔV = V_{mem} - V_{leak} between the input and output, current flows through one branch of the input differential pair. With small I_{biasSd}, the increase in output current can go only up to a limit - and further increase in I_{biasSd} entails a current flow from both branches of the differential pair. As I_{bias} is increased further, current in the second branch increases to the point that net output current is eventually zero. Since G_{m} is proportional to the bias current for a fixed potential difference, it follows the same behavior. This is for example shown in Fig. 4.24a when I_{biasSd} is 208 nA. Fig. 4.24b shows the performance of the source-degeneration biasing stage, comparing V_{GS} of source degeneration transistor as I_{biasSd} is varied. It can be seen that it is almost linear above the threshold V_{th}. Below threshold, the resistance R_{SD} gets large (as expected) and varies with I_{bias}. Fig. 4.24d shows this effect, where as I_{bias} is lowered, the transistor shifts from linear to saturation, and then to subthreshold contributing a few MΩ resistance. Finally Fig. 4.24c shows the transconductance G_{m} as a function of I_{biasSd} as I_{bias} is swept. With R_{SD} varying in linear region, we obtain the flat part of the traces, which shifts to saturation and subthreshold upon decreasing I_{biasSd}. The traces can in principle be used for tuning the leak conductance G_{m}, if wide linear range is required, however less than 0.4 µA of I_{biasSd} will entail large offset, which may not be calibratable – and the available range of G_{m}, above 0.4 µA in Fig. 4.24c is not too large.

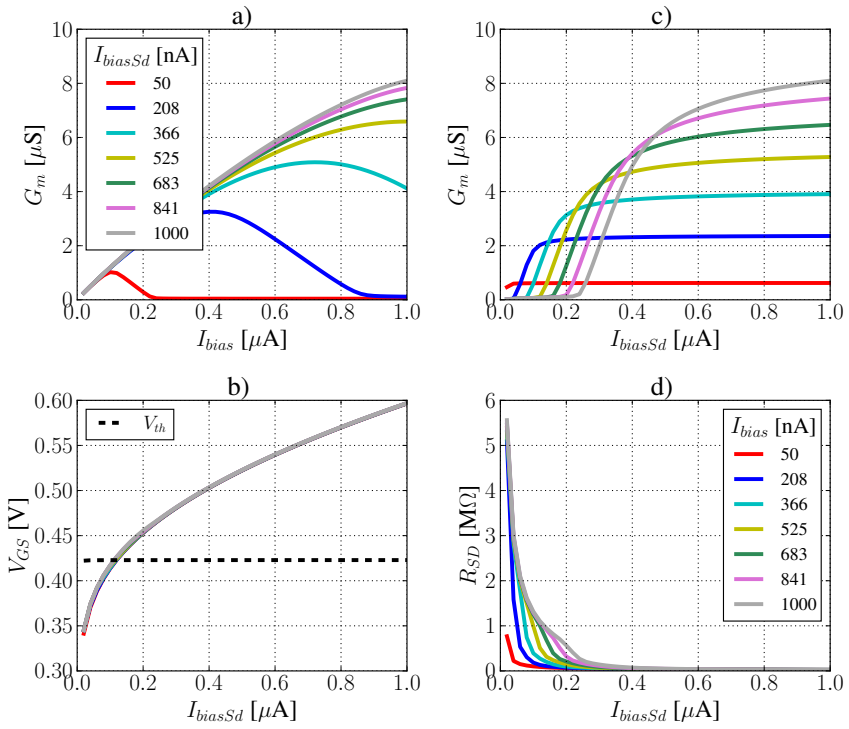


Figure 4.24: a) The leak conductance as a function of the OTA bias I_{bias}. b) The gate source potential across the source degenerating MOS transistors as a function of their control bias I_{biasSd}. c) The leak conductance as a function of the OTA source degeneration bias I_{biasSd}. d) The resistance contributed by the source degeneration (SD) MOS transistors, as a function of their control bias I_{biasSd}.

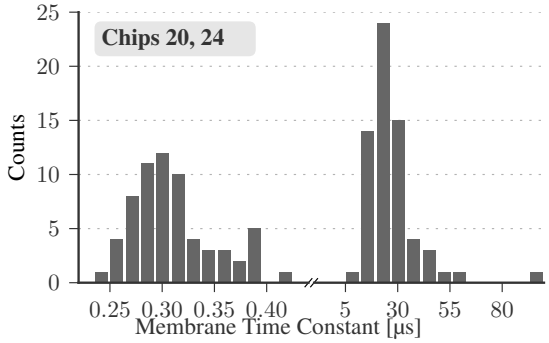


Figure 4.25: A distribution of minimum and maximum achievable membrane time constants [79].

The membrane time constant is therefore tuned by varying the adjustable capacitor C_{mem}, and by tuning the bias I_{bias}. We trade linear-range with robustness by not decreasing I_{biasSd} below 0.4–0.5 µA. The measured results of the minimum and

maximum achieved time constants from two separate dies are shown in Fig. 4.25. The neuron can set the time constants between 0.35 µs, up to 16.6 µs, given a 1σ variation. The circuit has been calibrated by measuring the relation between

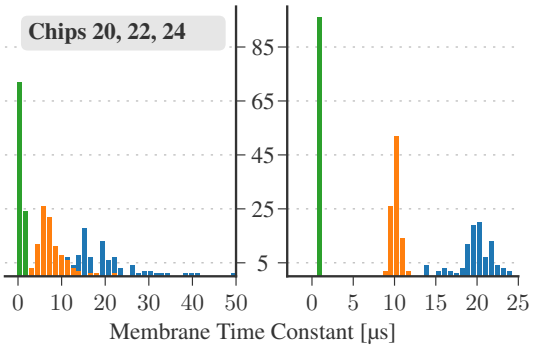


Figure 4.26: The pre- and post-calibration distribution of τ_{mem} for settings of 1 µs, 10 µs and 20 µs respectively.

membrane time constant τ_{mem} and I_{bias}, for a voltage decay from 0.6 V to its leak potential of 0.4 V [79]. The resulting curves are fitted by second degree fractional polynomials. The pre- and post-calibration results are depicted in Fig. 4.26 for a setting of 1 µs, 10 µs and 20 µs respectively. The post-calibration results show that while shorter time constants have very small spread, not all neurons can achieve the maximum 20 µs time constant.

4.6 Spike Generator and Reset Circuits

The spike generator evokes a digital pulse event once the membrane potential reaches a specified voltage threshold. This digital event further triggers a refractory reset circuit, initiating the refractory duration of the membrane trace. The circuit to achieve this is shown in Fig. 4.27. A two-stage comparator compares the voltages V_{mem} and V_{thresh} and asserts the logic levels V_{OH} or V_{OL} depending upon the input levels. These output logic levels are delayed by a programmable finite time t_{delay}, before resetting the output stage of the very comparator. This delayed resetting creates a pulse event *fire* that indicates a single spike event.

The left half of Fig. 4.27 shows the refractory reset circuit. In this circuit, the membrane potential V_{mem} can be connected to a fixed reset potential by a pass-transistor switch S_1. The control of this switch is triggered by an inverter, which is in turn controlled by the voltage on a 110 fF capacitor C_{refr}. The capacitor integrates a current I_{refr} constantly, due to which the inverter input is at a saturated voltage, eventually disconnecting the switch S_0. The inverter output is toggled when a fire pulse resets the voltage on the capacitor. This connects the membrane to the reset potential V_{reset} for a duration that lasts until the inverter toggles again. The input current linearly charges the capacitor and as soon as its trip point is reached, the switch S_0 turns off, indicating the end of refractory period. The membrane is

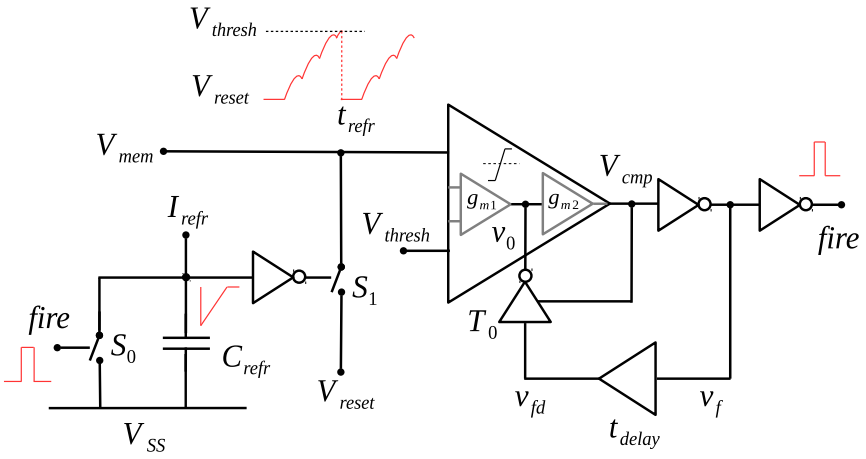


Figure 4.27: The spike event generator and the refractory circuit.

then free to respond to input activity again. The measured trace of Fig. 4.10 show the refractory period duration of about 10 µs during which the membrane is reset to 0.2 V.

4.6.1 Delay Element

The delay element used inside the spike generator circuit is a current starved inverter circuit whose schematic is shown in Fig. 4.29. The transistors M_3, M_4 operate as an inverter, while M_2, M_5 operate as current sources. They are meant to limit the current that is available to the inverter and are controlled by the current mirrors formed by M_7, M_5 and M_1, M_2. The capacitor C_{delay} is a 58 fF MOS gate-oxide capacitor, that will be charged and discharged by the current sources M_2 or M_5 depending upon the input signal. The incoming digital signal can then be delayed by a time interval equal to $t_{\text{rise}} = C_{\text{tot}} \cdot V_{\text{mid}}/I_{\text{M5}}$. Where V_{mid} is the voltage to be reached, e.g., $V_{\text{DD}}/2$ and C_{tot} = C_{delay} + C_{par}. C_{par} is the parasitic capacitance contribution at the drains of the inverter transistors M_3, M_4. I_{M5} is the current through the source formed by M_5. As the two current sources multiply the main bias current $I_{\text{biasDelay}}$, it directly tunes the delay interval as shown in Fig. 4.28.

The time delay within the spike generator circuit has a direct impact on the maximum firing rate of the neuron. If the neuron is to fire at higher rates of 1 MHz (assuming a 1 kHz maximum biological firing rate), then long delay times should be avoided. The capacitor C_{delay} has been sized considering its charging/discharging time and mismatch in delays, such that the circuit works in all cases.

Looking back at the pulse generation mechanism of Fig. 4.27, one may notice the presence of T_0, whose output is connected to intermediate node, between the two stages of the comparator circuit. The circuit T_0 is an inverter whose NMOS stage is enabled by the comparator's output node V_{cmp}, shown in Fig. 4.30. This ensures that the pulse is evoked only when $V_{\text{mem}} \geq V_{\text{thresh}}$. This further makes

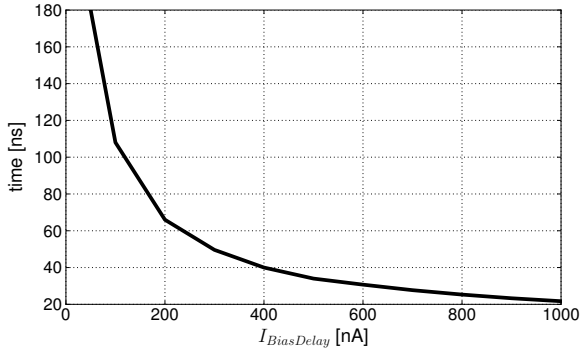


Figure 4.28: Simulated data showing how the bias current can be changed to tune the time delay.

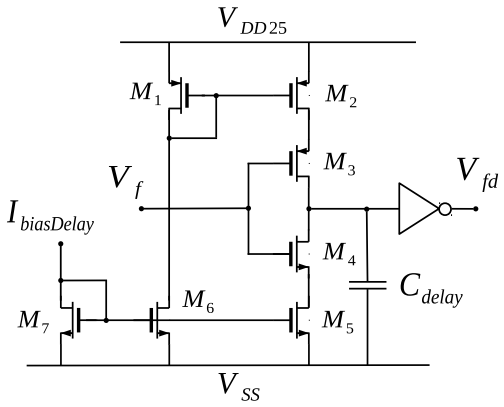


Figure 4.29: The schematic of the current starved delay element used inside the spike generator circuit.

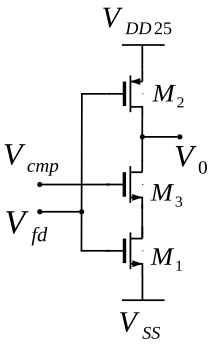


Figure 4.30: The intermediate resetting stage within the spike generator circuit.

it evident that the falling edge t_{fall} of the delay element t_{delay} matters and not the rising edge delay. The duration of the falling edge delay will therefore determine the pulse width of the generated *fire* output pulse.

4.6.2 Refractory Period

The refractory period circuit (shown in Fig. 4.27) in *DLS-2* implementation allows for a much longer time duration compared to *DLS-1*, first due to a slightly larger capacitor C_{refr}, but mainly due to the pre-scaling of input bias current that divides I_{refr} by a factor of ten (not shown in schematic). This essentially reduces the 15 nA input bias from the Capmem to 1.5 nA. The integration of smaller current on the capacitor allows longer refractory times, as its takes longer to reach the trip point of the inverter that toggles the switch S_1 (see. Fig. 4.27). The range of tunable refractory times as the bias current I_{refr} is swept in terms of equivalent 10-bit digital

code is plotted in Fig. 4.31.

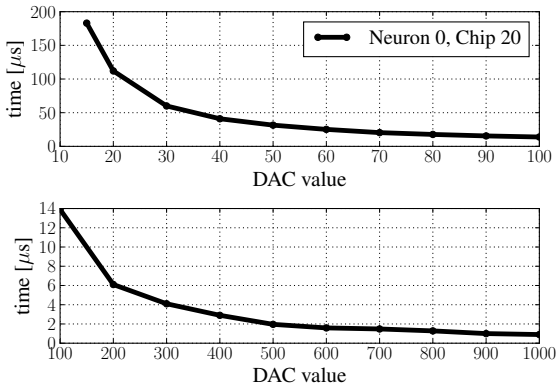


Figure 4.31: Measured results from a single neuron showing the available refractory times as a function of its bias current.

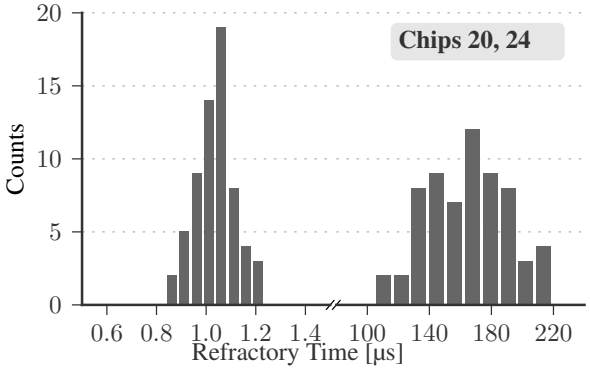


Figure 4.32: A distribution of minimum and maximum measured refractory times [79].

Note that at very low currents, one can get very long refractory periods. Fig. 4.32 shows measured results showing distribution of minimum and maximum possible refractory periods for all neurons on a single die. The distribution therefore comprises of 32 samples. Note that the small currents and consequently long refractory periods have more mismatch, compared to large current with short refractory periods. Yannik Stradmann has calibrated the refractory periods duration by applying a second order fractional polynomial fit for each neuron. The pre- and post-calibration results are shown in Fig. 4.33. It can be seen that the residual spread for the refractory periods is reduced considerably for ranges below 5 and 15 µs, our specified range from Table 3.1. A maximum refractory period of about 104.5 µs is available from the measurement data from two dies, if calculated with 3σ single-sided quantiles [79].

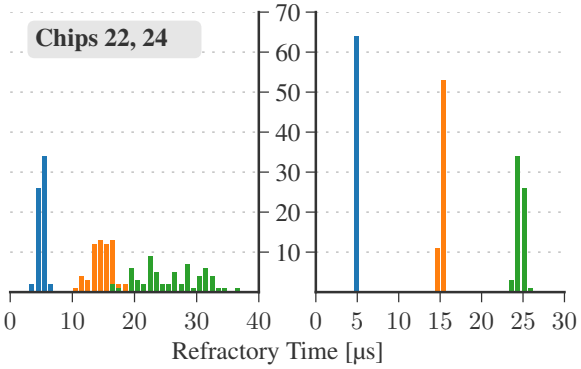


Figure 4.33: Pre- and post-calibration refractory times for three different time constants [79].

4.7 Membrane Capacitor

The membrane capacitor in the *DLS-2* prototype is realized as two-bit adjustable MOS gate oxide capacitor contributing a total of 2.36 pF. This is implemented as four parallel and equal transistors of size W/L = 9.9/10.4 µm allowing 590 fF each. In this configuration, the user may use them as 590 fF, 1.77 pF or an accumulated 2.36 pF. Together with leak conductance, the membrane time constant is therefore also configurable via this capacitor since $\tau_{\text{mem}} = C_{\text{mem}}/G_{\text{m}}$. Eq. 3.24 relates

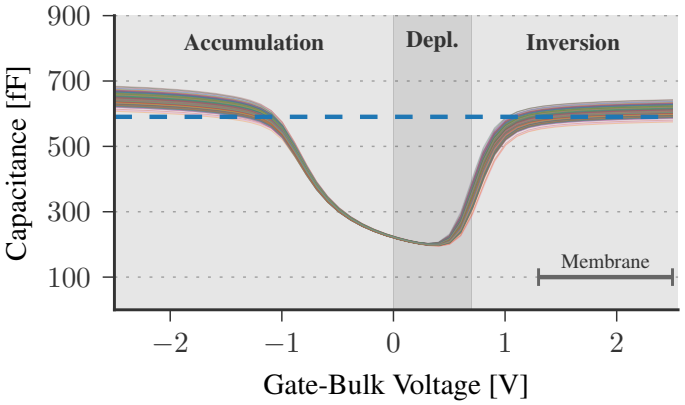


Figure 4.34: Capacitance vs. gate voltage (CV) of MOS gate-oxide capacitor simulated over several Monte-Carlo samples.

the contribution of a MOS gate capacitor per unit area. The C'_{ox} values for thick-oxide transistors is, to a rough estimate between 5.5 and 6.5 fF/µm^2 for PMOS and NMOS, and 12 to 13.5 fF/µm^2 for core PMOS and NMOS respectively[4]. The capacitance contribution of a device with a dimension W/L = 9.9/10.4 can be verified

[4] In the used process technology, the oxide thickness for thick-oxide transistors is approx. 5.5 – 6 nm, whereas for the thin oxide it is 2.5 – 3 nm, with a relative dielectric model parameter of 3.9.

to a first order by plugging in numbers, such that $C'_{\text{ox}} \cdot\ W\ L = 5.75 \times 9.9/10.4 =$ 597 fF. One can further simulate the capacitance to gate-voltage (CV) curve of the MOS capacitor using foundry device models. This is shown in Fig. 4.34 where 500 Monte Carlo samples are included to check the extent of possible mismatch. The figure also highlights the various MOS capacitor regions. Note that thick-oxide transistors are utilized, although they provide half as much capacitance compared to the core (thin-oxide) counterparts. This ensures that there is always a sufficient V_{GS} drop of 1.3 V or more and the device is biased in inversion region, as shown in Fig. 4.34. Note that despite inversion region, the transistor mismatch can lead up to 5% variation in the eventual capacitance. The area benefit one obtains, as opposed to using metal capacitors is still more than twofold, hence this is a minor concern, and the mismatch arising is to be calibrated for.

Fig. 4.34 shows the three device regions determined by the biasing condition. Although accumulation and inversion regions seem to contribute approximately equal capacitance, inversion is preferred because the gate-bulk capacitance in accumulation region can have large series parasitic substrate resistance [127]. However, the use of accumulation region has also been reported in literature [128]. While designing the integrator architecture of Sec. 4.3.1, the depletion mode compensation schemes were visited [129, 130]. Parallel and series compensation schemes broaden the depletion region of the MOS CV curve by utilizing the substrate bias. However, in order to avoid additional biasing complexity as well as bulk-driven compensation schemes, their implementation was not pursued.

4.8 Analog Input/Output

The analog I/O block shown in Fig. 4.1 enables the read-out of internal node voltages externally or inject stimulus directly from the pad interface in the first two LIF prototype chips.

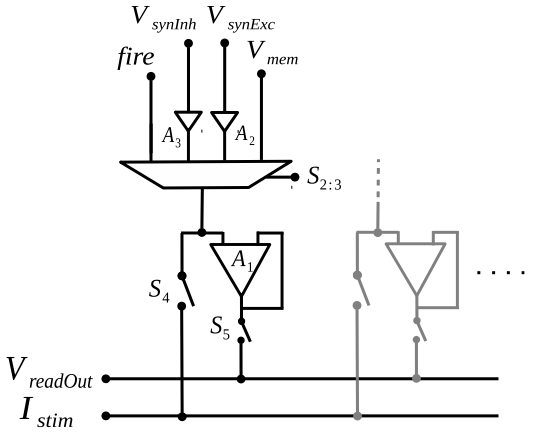


Figure 4.35: The architecture of the analog I/O block used as read-out and debug interface, terminating at shared output lines.

The two pins labeled I_{stim} and V_{readOut} terminate directly at the chip pads and are shared between all integrated neurons. The block schematic highlighting this is shown in Fig. 4.35. A multiplexer reads out either of the four different voltage signals from among membrane voltage V_{mem}, the voltage on the two input synaptic lines V_{synExc}, V_{synInh}, as well as the digital fire pulse. $V_{\text{syn,exc}}$ and $V_{\text{syn,inh}}$ are buffered via source followers labeled in Fig. 4.35 as A_2 and A_3. The source followers are NMOS based and share their biasing circuit with the two-stage read-out amplifier. The schematic of the source follower, together with the shared bias is shown in Fig. 4.36. The input transistor M_1 is embedded in a deep n-well to prevent body-effect and keep the follower gain close to unity at the expense of more area. In the presence of body-effect the follower gain is given as:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{g_{\text{m1}}}{g_{\text{m1}} + g_{\text{DS1}} + g_{\text{DS2}} + g_{\text{mb1}}} \approx \frac{g_{\text{m1}}}{g_{\text{m1}} + g_{\text{mb1}}} \tag{4.5}$$

where g_{m1} and g_{mb1} are the transconductance and body-effect transconductance ($\frac{\partial I_{\text{D}}}{\partial V_{\text{BS}}}$) of the input transistor M_1, whereas g_{DS1}, g_{DS2} are the output conductances of transistors M_1 and M_2 . Note that $g_{\text{mb}} = \eta\ g_{\text{m}}$, where η is between 0 (no body effect) and 0.5. The source follower creates a drop of about 0.63 V as the shared cascode bias sets a drain current of 1.7 µA. The static synaptic input line voltage of 1.2 V therefore drops to about 0.57 V after the read-out. Note that when the

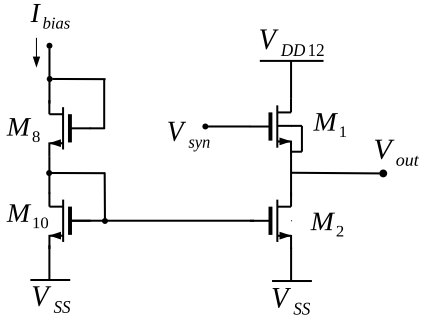


Figure 4.36: The source follower used to read out the synaptic input lines, and its shared bias circuit.

read-out amplifier is disabled digitally, it grounds the gates of transistors $M_{10,2}$ and M_8 (not shown in schematic). This disables the operation of the source follower as well as the read-out amplifier.

The multiplexer is a 4x1 transmission gate multiplexer with two digital select lines. The output of multiplexer is either sent out to the pad pin V_{readOut} via an opamp based buffer A_1 and S_5 or is directly connected to output pin I_{stim} via another transmission gate switch labeled S_4. The direct connection can inject current into the membrane or hold it to a reference voltage potential. The realized opamp is described in Sec. 4.8.1. The switches S_4, S_5 are thick-oxide (I/O) transmission gates as they directly connect to the pads, whereas those within the multiplexer are

thin-oxide (core) transistors since the internal nodes are not meant to rise above 1.2 V.

4.8.1 Two-Stage Opamp

The read-out buffer A_1 used within the analog I/O block of Fig. 4.35 is a two-stage opamp whose schematic is shown in Fig. 4.37. The amplifier was originally designed for *DLS-1* where the maximum input bias current from capacitive memory was specified at 2 µA. The maximum bias in *DLS-2* Capmem is reduced to 1 µA – however the design of the amplifier has remained unaltered. The initial buffer is designed for a total current consumption of 100 µA. When the bias was scaled the total consumption halved to approximately 50 µA. The opamp schematic in Fig. 4.37 shows this bias being fed as I_{bias} to an n-type input cascode current mirror, since the capacitive memory only has PMOS output stage and must be mirrored. The current mirror formed by M_{8-11} multiplies the input current 6 times, which is further multiplied twice and five times in mirrors formed for tail current source $M_{12,5}$ and the output stage $M_{12,7}$ respectively. The opamp is a two-stage architecture with indirect compensation scheme. It realizes a p-type input stage to sense lower common-mode levels, and a standard class-A output stage. It uses indirect compensation scheme to stabilize itself using split length transistors.

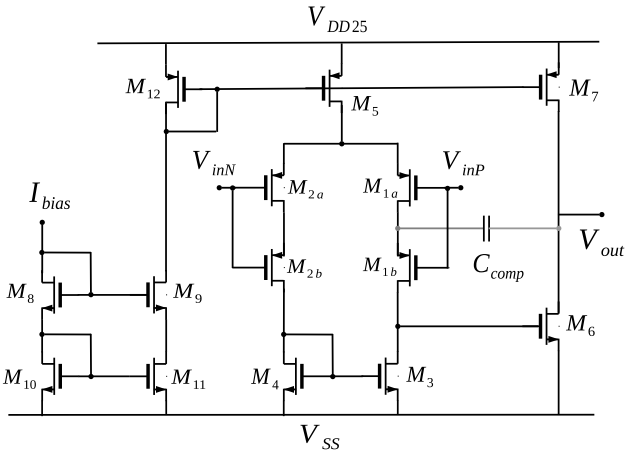


Figure 4.37: The two-stage amplifier designed for the membrane read-out buffer.

Two-stage amplifiers are typically compensated by Miller compensation where a compensation capacitor connects the outputs of the two gain stages [131]. However, as a result of this connection a Right Hand Plane (RHP) zero appears in the frequency response which decreases the phase margin. This is due to the feedforward current that flows through this compensation capacitor. Several circuit techniques have been proposed in literature that compensate for this RHP zero. These include, for example, dominant pole compensation, where higher phase margin is achieved by pushing the dominant pole towards the origin, splitting the dominant

Supply	1.2 V
Power	127.6 µW
DC Gain	62.6 dB
CMRR	65.1 dB [1]
PSRR	67.8 dB [1]
Compensation Cap	600 fF
UGF (f_u)	149.6 MHz
Phase Margin [2]	89 °

[1] evaluated at 5 kHz
[2] C_L=3 pF || 10 MΩ

Table 4.2: Open loop opamp specifications.

and non-dominant poles at the expense of bandwidth. Lead compensation is another common compensation scheme that uses a series resistor to either eliminate the zero altogether or to cancel the LHP non-dominant pole. Active techniques include a feedback voltage buffer [132] or current buffers [133, 134] that block the feedforward component that passes directly from input to output avoiding gain stage inversion. Such buffers therefore allow only feedback current. Alternatively one can embed such buffer stages within the input stage, e.g., using common gates with cascode topology [135, 136]. For low-voltage design an even better technique is the use of splitting the channel length of the input differential pair [137], thereby creating low impedance nodes. An indirect connection to the internal node by a connection through an internal low impedance node forms the basis for indirect compensation schemes. Indirect compensation can be very useful in stabilizing multi-stage low-voltage amplifiers, for example as shown previously by the author in [138, 139].

A MOS device can split its length into half, as shown in Fig. 4.37 (see $M_{1a,1b}$). Of the two input devices, the upper PMOS is then in triode region, while the lower one is in saturation. Their mid-node where the two devices connect is a low-impedance node, first because the connection is to a source terminal, but also since one device is in triode region. The low impedance node can hence be used to feed the compensation current back to the output of the first gain stage. In the designed amplifier, a single transistor with $2\cdot L_{\text{min}}$ was initially designed for the input differential pair, and later the device was split-up for compensation with each split device having a minimum channel length (L_{min}) – 280 nm for the I/O thick-oxide transistors in the given technology.

The simulated achieved specifications of the amplifier are listed in Table 4.2. The listed results are with a bias setting of 1 µA. Since the amplifier is designed from thick-oxide transistors, a 2.5 V supply is used with a total power consumption of about 127 µW. The open loop DC-gain in the current implementation is about 62.6 dB. The common mode rejection, defined typically as a ratio of differential to

common-mode gain, $\frac{A_{\text{diff}}}{A_{\text{cm}}}$ is 65.1 dB in the ideal case. The negative and positive supply rejection are also simulated. PSRR is typically defined as min($\frac{A_{\text{diff}}}{A_{\text{Vdd}}}$, $\frac{A_{\text{diff}}}{A_{\text{Vss}}}$) and the lesser of the two is 67.8 dB for the current architecture. These results are summarized at a signal frequency of 5 kHz and the frequency response is shown in Fig. 4.38c. Table 4.3 shows the pole and zero locations of the uncompensated amplifier, as well as the results after compensation. The two uncompensated poles given by $\frac{g_{\text{DS2}}+g_{\text{DS4}}}{C_{\text{par}}}$ and $\frac{g_{\text{DS6}}+g_{\text{DS7}}}{C_{\text{L}}}$ are located close to each other, thereby deteriorating the phase margin. With indirection compensation the dominant pole is pushed lower on the frequency axis, while two conjugate poles typically appear higher up the frequency axis. Where $g_{\text{DS2}} = g_{\text{DS1}}$ and $g_{\text{DS3}} = g_{\text{DS4}}$, C_{par} is the parasitic capacitance at the output node of first gain stage and C_{L} is the output load capacitance at node V_{out}.

The compensated split pole locations can be approximated [137] to be at $\frac{-2}{g_{\text{mII}}R_2R_1C_{\text{comp}}}$, where $R_1 = \frac{1}{g_{\text{DS2}}+g_{\text{DS4}}}$, $R_2 = \frac{1}{g_{\text{DS6}}+g_{\text{DS7}}}$, and the real part of the conjugates poles as $\text{Re}(p_{2,3}) = \frac{g_{\text{mII}}}{C_{\text{L}}}\sqrt{\frac{g_{\text{mp}}C_{\text{L}}}{g_{\text{mII}}C_{\text{par}}}}$. The node impedance at the mid-node is $1/g_{\text{mp}}$, where $g_{\text{mp}} = \sqrt{2}g_{\text{mI}}$, and g_{mI} is the equivalent transconductance from the first stage. Further, instead of an RHP zero, a Left Hand Plane (LHP) zero appears at a frequency $\frac{2\sqrt{2}}{3}\omega_{\text{u}}$, which in the implemented amplifier actually helps improve the phase margin. The bode plots in Fig. 4.38a,b show the compensated gain and phase plots overlaid on the uncompensated ones. Note that for the compensated phase plot, the phase shift from –90° decreases due to the zero, followed by the sharp cutoff. The resulting bandwidth is therefore increased and the two conjugates poles are pushed much forward, so that they appear almost at the UGF – and therefore do not deteriorate the phase margin. The unity gain frequency for the compensated case is $f_{\text{u}} \approx \frac{2g_{\text{mI}}}{2\pi C_{\text{comp}}}$, which is further increased due to the zero that appears at about $0.94f_{\text{u}}$.

In the close loop configuration, the buffer encounters a low pass filter right at the output of close loop amplifier. This is formed by the output transmission gate switch together with the large off-chip load capacitance, since the pin labeled V_{readOut} directly drives the output pad. An off-chip parasitic is estimated using

	Uncompensated	Compensated
Phase Margin [1]	14 °	89 °
f_{u}	131 MHz	149.6 MHz
f_{p1}	–321 kHz	–17.7 kHz
$f_{\text{p2,3}}$	– 39 M	– 63.1 MHz $\pm$ j141 MHz
f_{z1}	-	– 28.6 MHz

[1] C_{L} = 3 pF $\parallel$ 10 MΩ

Table 4.3: Pole and zero locations of the uncompensated and compensated two-stage open loop opamp.

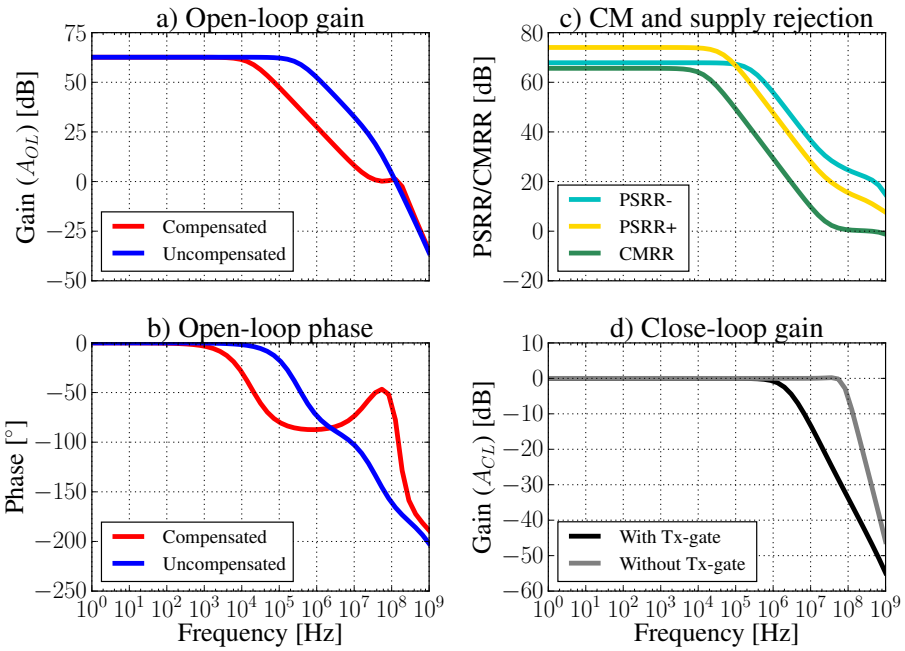


Figure 4.38: Frequency response of the open-loop opamp. a) The uncompensated and compensated gain curves. b) The respective phase plots for the two gain curves. c) Common-mode and power supply rejection. d) The close loop buffer bandwidth with and without the output transmission gate (shown as S_5 in Fig. 4.35). The output load here is 16 pF ∥ 10 MΩ.

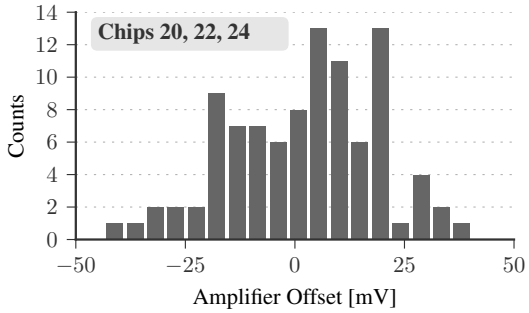


Figure 4.39: The input-referred offset measured from 96 amplifiers buffers over three chip dies [79].

direct measurement on the PCB using an LCR meter as well as from the slope of the large signal. It is concluded that around 16 pF is encountered on the *DLS-2* measurement board. Fig. 4.38d shows the close loop amplifier bandwidth with C_{L} of 16 pF with and without the transmission gate. It can be seen that the close loop –1 dB bandwidth reduces from 63.6 MHz to 1.15 MHz. However, this low pass

Output range	0.1–2.1 V
Load	$\geq$ 16 pF $||$10 MΩ[1]
–3-dB bandwidth	2.4 MHz
–1-dB bandwidth	1.15 MHz
Input offset	14 mV
Slew rate	2 V/µs

[1] Estimated off-chip load

Table 4.4: Measured results of the read-out buffer [79].

filter helps increase the stability in the presence of large off-chip load. The buffer has been characterized during chip measurements and the results are summarized in Table 4.4. The output range and input offset correlate very well with the simulated results. The slew rate of the amplifier is limited by the charging of the output node via the source transistor M_7 and given by I_7/C_L. A slew rate of 2 V/µs is deemed sufficient for readout measurements although it can be improved by tweaking the current source formed by M_7. Fig. 4.39 shows the distribution of input offset of the buffer, measured from 96 amplifier instances on three different chips and summarized in Table 4.4.

4.9 Switches

The neuron circuit uses switches to interconnect individual subcircuits from the neuron membrane and spike generator, as well as inside the analog I/O debug subcircuit. These switches are transmission gate switches made by a parallel combination of NMOS and PMOS devices. Those that connect directly to the membrane are realized with 1.2 V core devices (these also include those inside debug multiplexer in the analog I/O). The switches that connect the neurons to the external interface, e.g., the two pins V_readOut and I_stim are 2.5 V I/O devices.

The on-resistance of a MOS pass transistor is given by

$$R_\mathrm{on} = \frac{L}{\mu C_\mathrm{ox} W (V_\mathrm{GS} - V_\mathrm{th})} \tag{4.6}$$

which indicates that larger resistance requires longer channel length. The equation also indicates that the maximum voltage an NMOS transistor can allow is $V_\mathrm{DD} - V_\mathrm{th}$, whereas the lowest voltage a PMOS can pass is limited by its V_th. A transmission gate chooses both in parallel to enable a rail-to-rail swing, and with an overall lower resistance over the entire range, such that $R_\mathrm{eq} = R_\mathrm{onP} \parallel R_\mathrm{onN}$. It can be shown that if $\mu C_\mathrm{ox} W/L$ of both transistors are the same, and the variation in V_th is neglected, then R_eq is independent of the input level [140]. The channel resistance of various switches used in the neuron design are plotted in Fig. 4.40. The top trace simulates the resistance for the 1.2 V core transistors, while the bottom trace

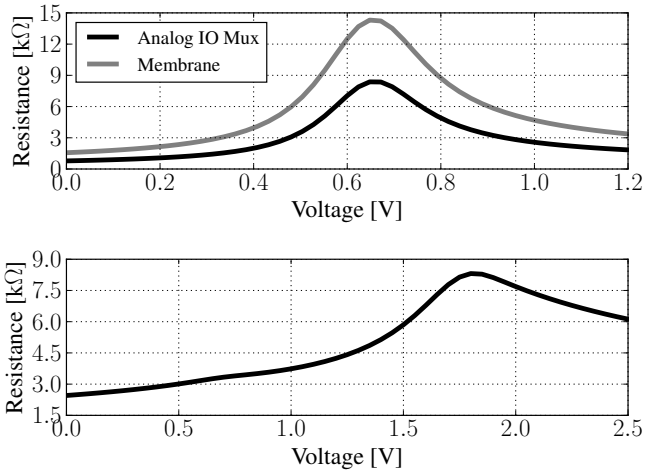


Figure 4.40: The channel resistance of various switches used in the design. The top figure shows the core transistor switches (gray curve for membrane switches, black curve for debug multiplexer), while bottom shows the 2.5 V thick-oxide transistors used at the two output pins.

plots the same for 2.5 V transistors. Note the gray curve, which is for the switches that connect the individual circuits to the membrane capacitor has slightly higher resistance, compared to the black curve (top subplot) for switches in the debug multiplexer (Analog I/O). This was done to minimize off-state leakage, especially from circuits like synaptic inputs, where leakage current is more crucial. Since the membrane voltage in the neuron is restricted to 1.2 V, the maximum resistance seen by the I/O switches is approximately 4 kΩ.

All switches are enabled by the digital backend configuration. The digital signals for a 2.5 V I/O transistors are up-converted to 2.5 V with a level-shifter. Appendix A enlists these digital configuration bits as well as their description.

4.10 Bypass Mode

The neuron includes an option to bypass the analog integration and evoke a single spike per input synaptic event. This mode is mainly integrated for debugging the system without worrying to configure the neuron circuit, for example, to test event routing. The short input pulse event arriving at the synaptic input pulls the synaptic line voltage lower, in proportion to the pulse amplitude (or total equivalent current). If the input event is made strong enough, the line drop increases to the extent that it triggers the bypass buffer. The schematic of the bypass link is shown in Fig. 4.41, together with a tri-state inverter. Two inverters are cascaded, where the first one is raised by a diode-connected transistor. This increases the trip point of the inverter by a few hundred millivolts. The input line is pulled up by another transistor to

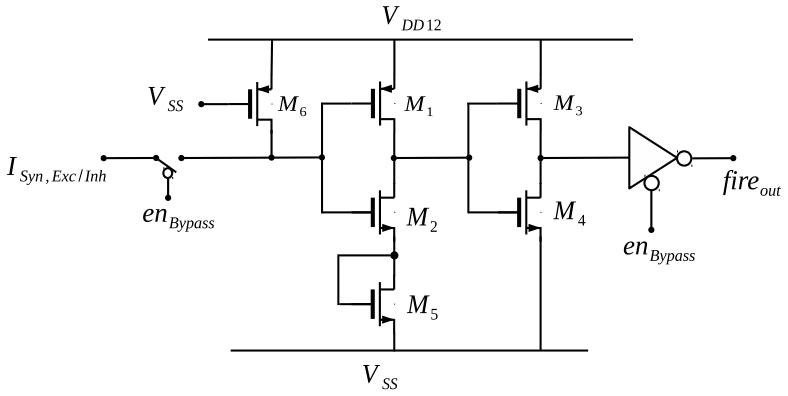


Figure 4.41: Schematic diagram of the bypass link together with a cascaded tri-state inverter.

avoid a floating connection when the link is disabled. The bypass part of the circuit was conceived by Johannes Schemmel.

The simulated behavior of the bypass link is shown in Fig. 4.42. The strength of incoming input event is increased to the point that it triggers the bypass link. In the simulation setup, up to 10 µA amplitude with a minimum pulse-width of 29 ns is required to achieve this. The plot shows that the synaptic input line drops to about 920 mV to trigger the shifted inverter trip point. The lower plot shows the output spike event (*fire*) evoked not through spike generator, but the bypass link.

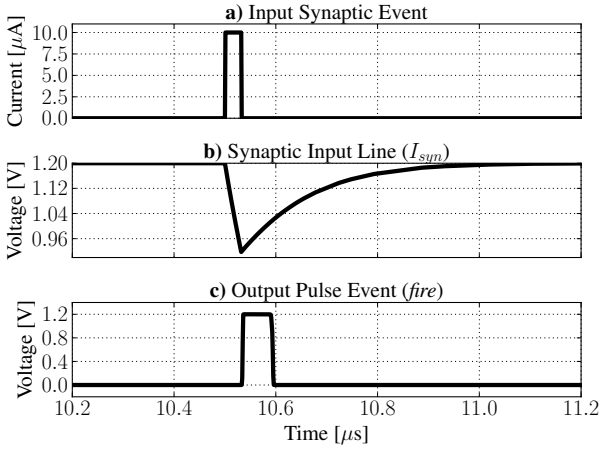


Figure 4.42: Simulation results of the bypass-mode of the neuron circuit. a) A large incoming synaptic current pulse on the synaptic input line. b) The resulting voltage drop during the pulse interval on the input line, followed by a recovery with a time constant. c) The voltage drop triggers the bypass link, which evokes a digital output event.

4.11 Power Consumption

The neuron circuit is simulated to estimate its total power consumption. The full circuit draws its current mainly from 2.5 V supply line, but also from 1.2 V supply. The total static current it draws (outside the spiking interval) is about 6 µA for a typical set of biases, given both synaptic inputs are enabled, and debug amplifier is disabled. The static supply current from the 2.5 V and 1.2 V supply are around 5.6 µA and 400 nA respectively - leading to a power consumption of about 14.4 µW. When the inhibitory input is disabled (since its not used) the power consumption is reduced to 10 µW. The dynamic power is increased during operation

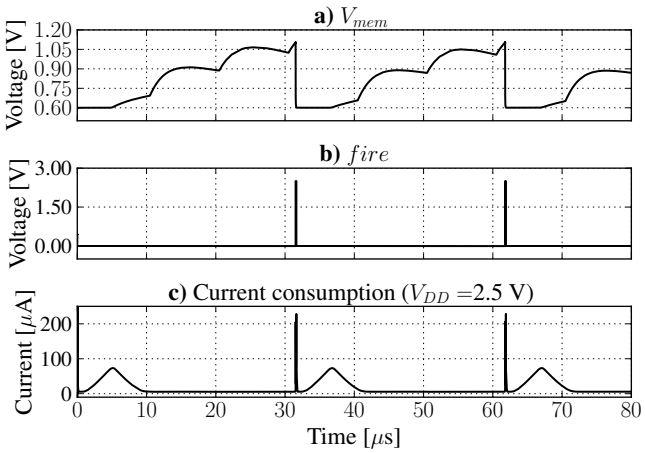


Figure 4.43: a) Membrane potential as a result of input synaptic activity. b) Corresponding 2.5 V output digital spikes that initiate the refractory period. c) The current consumption from the 2.5 V supply – notice the spike due to the switching in SpikeGen circuit and the increase/decrease in consumption due to the slow inverter in refractory period circuit.

due to inverter leakage, especially when its input is a slowly increasing voltage. During the refractory period, the leakage current due to the slow inverter (see inverter of S_1 in Fig. 4.27) increases until the refractory period ends – after which it decreases. This leakage is therefore proportional to the duration of refractory period. Similarly the second stage (g_{m2}) of the spike comparator as well as the two inverters in Fig. 4.27 contribute to leakage during the pulse *fire* interval or at edges. This is a strong reason to not use inverters for slow signals when power consumption is important. The dynamic current consumption is shown in Fig. 4.43. The membrane in Fig. 4.43a builds up due to the input synaptic activity, evoking spikes twice. The 2.5 V digital *fire* signal is plotted in Fig. 4.43b, which initiates the refractory period duration. Fig. 4.43c shows the current consumption from the 2.5 V supply. Note the initial spikes that are due to the leakage from *SpikeGen* circuit, followed by the leakage in the refractory period that lasts as long as neu-

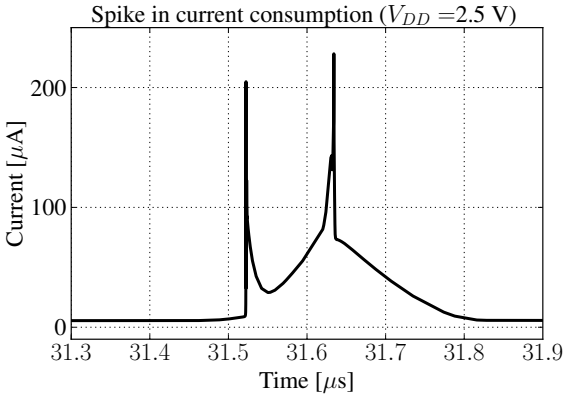


Figure 4.44: A zoomed-in version of a current spike in Fig. 4.43c.

ron is in refractory state. After refractory period ends, the leakage decreases. The slow inverter leakage mentioned above also gives extra current consumption within the *fire* pulse duration in the delay element circuit (Fig. 4.29), due to the presence of output inverter. This is visible in Fig. 4.44 which is a zoomed-in version of a current spike in Fig. 4.43c. Notice that along with the two spikes the base current consumption increases followed by a decrease.

The energy per output spike can be determined by stimulating a neuron with strong synaptic input current, setting a short refractory period of 1 µs and dividing the average power consumption from the two supplies by the output spike rate. For an output rate of 14.2 kEvents/sec, such that 71 events are evoked per 5 ms, and 7 input events (each of 1 µA amplitude, 32 ns long) lead to a single output event, with disabled inhibitory synaptic input, the energy consumption is 907 pJ per spike. This decreases with high output rate resulting in 193 pJ per spike for a rate of 290 kEvents/sec.

4.12 Physical Neuron Implementation

On the physical level, the neurons are arranged vertically in an array embedded in the analog network core of the prototype chip. Each neuron in the *DLS-2* chip is 11.76 µm wide and 200 µm in height. The prototype chip integrates a total of 32 neurons in an array occupying 11.76 µm $\times$ 32 $=$ 376 µm. This neuron array arrangement from external routing perspective and top-level view is further sketched in Fig. 4.45. The local parameters from the capacitive memory enter each neuron from the bottom of the array, while the global parameters enter from the left edge. The global input/output pins I_{stim} and V_{readOut} enter and leave the array from the left edge. The array is edge-connected at the top with the synapse matrix where the lines I_{synExc}, I_{synInh} from a single synapse column enter the neuron. The signals postIn and postOut enter (leave) each neuron from the bottom and top respectively

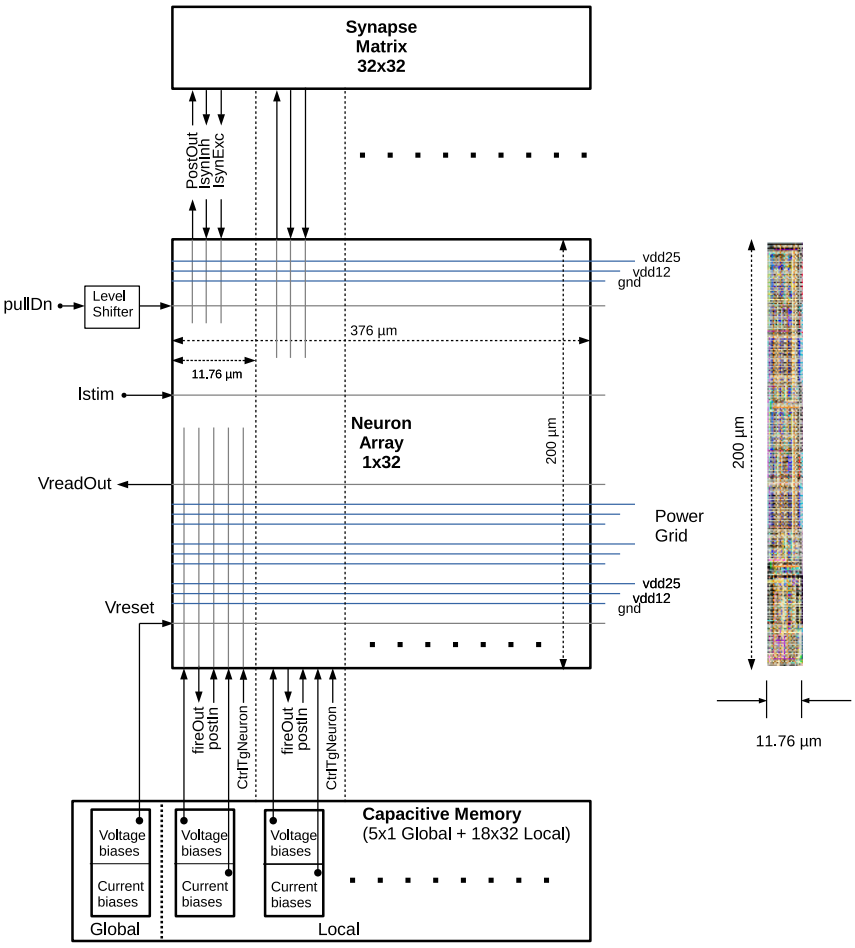


Figure 4.45: The physical architecture of the neuron array, together with vertical and horizontal routing lines. On the right is the layout view of a single physical instance.

too. The output fire_{out} as well as the input configuration bus ctrlTgNeuron coming from the digital backend leaves and enters each neuron from the bottom too. All global circuit blocks are placed on left of the array – e.g., in Fig. 4.45 a level shifter for the comparator reset sits on the left side and is common to all neurons. The power distribution lines enter the array from the right side and further distribute inside each neuron in vertical lanes, together with parameter biases and digital signals. The right inset in Fig. 4.45 shows the physical layout view of a single neuron circuit.

4.13 Bias Parameters

As described earlier, the neuron circuit is configured by a set of global and local voltage and current parameters, provided by the capacitive memory. The complete set of these voltage and current biases are listed in Table 4.5. The table lists every

Parameter	Circuit	Type	Typical Values
$\mathtt{V_{thresh}}$	SpikeGen	local/voltage	0.6 V – 1.2 V
$\mathtt{V_{leak}}$	Leak	local/voltage	0.2 V – 1.2 V
$\mathtt{V_{reset}}$	Reset	global/voltage	0.2 – 1 V
$\mathtt{I_{biasSpkCmp}}$	SpikeGen	local/current	0.6 µA
$\mathtt{I_{biasDelay}}$	SpikeGen	local/current	100 nA
$\mathtt{I_{biasLeak}}$	Leak	local/current	15 nA – 1 µA
$\mathtt{I_{biasLeakSd}}$	Leak	local/current	0.5 µA – 1 µA
$\mathtt{I_{biasReadOut}}$	Analog I/O	local/current	1 µA
$\mathtt{I_{refr}}$	Reset	local/current	15 nA – 1 µA
$\mathtt{V_{synExc}}$	Syn. Input (Exc.)	local/voltage	1.05 – 1.25 V
$\mathtt{I_{synResExc}}$	Syn. Input (Exc.)	local/current	15 nA – 1 µA
$\mathtt{I_{biasSynGmExc}}$	Syn. Input (Exc.)	local/current	15 nA – 1 µA
$\mathtt{I_{biasSynSdExc}}$	Syn. Input (Exc.)	local/current	0.5 µA – 1 µA
$\mathtt{I_{biasSynOffExc}}$	Syn. Input (Exc.)	local/current	15 nA – 1 µA
$\mathtt{V_{synInh}}$	Syn. Input (Inh.)	local/current	1.05 – 1.25 V
$\mathtt{I_{biasSynResInh}}$	Syn. Input (Inh.)	local/current	15 nA – 1 µA
$\mathtt{I_{biasSynGmInh}}$	Syn. Input (Inh.)	local/current	15 nA – 1 µA
$\mathtt{I_{biasSynSdInh}}$	Syn. Input (Inh.)	local/current	0.5 µA – 1 µA
$\mathtt{I_{biasSynOffInh}}$	Syn. Input (Inh.)	local/current	15 nA – 1 µA

Table 4.5: A summary of tunable analog neuron parameters and their operating range. A total of 18 local (individual) parameters and 1 global parameter tune the neuron.

parameter, the circuit it is used in, its type, as well as the typical tuning range. Note that the range of allowed values sometimes depend mutually on a set of parameters, and the given parameter range does not cover possible corner cases. The values of V_{syn} parameters are set together with $I_{\text{biasSynOff}}$ biases (excitatory/inhibitory) to cancel the synaptic offset, therefore they typically depend on the calibration algorithm. The voltage values are dependent on the dynamical behavior one wants to reproduce. The source degeneration values are advised to be set higher, since low $I_{\text{biasLeakSd}}$ values give large input offset and the leak conductance varies non-linearly with I_{biasLeak}. The parameters $I_{\text{biasSpkCmp}}$, $I_{\text{biasDelay}}$, $I_{\text{biasReadOut}}$ are analog circuit parameters (i.e., they do not change the biological behavior) and should be set only once.

4.14 Full Circuit Characterization

The complete neuron circuit is characterized next by stimulating the membrane with incoming synaptic input pulses. Fig. 4.10 showed one case, where strong incoming events drop the synaptic input line by approximately 170 mV in the lower trace, recovered back with a short time constant of a few µs. The upper trace showed the resulting membrane potential, which requires three or four incoming events to spike. Spike threshold and reset potential are set at roughly 0.58 V and 1.2 V.

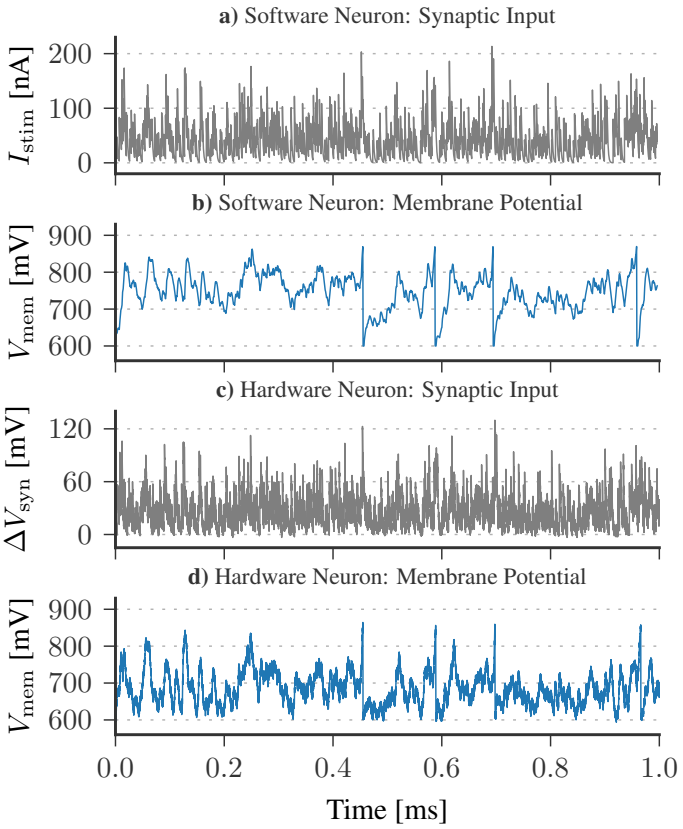


Figure 4.46: A comparison of the on-chip neuron vs. software simulation [79]. Both neurons are stimulated with a random spike train stimulus and their membrane voltages are plotted. The corresponding synaptic input current (for software neuron) and the proportional voltage drop for synaptic input line (for the hardware neuron) is plotted.

This section demonstrates the use of random spike train stimulus as input and neuron modeling parameters from a computational study [107]. Further a side-by-side comparison is made with the ideal response of the LIF model simulated in Brian spiking neural network simulator [141]. The on-chip neuron here is uncalibrated and is only manually tuned to cancel the effect of non-ideal effects, e.g., by

tuning V_{synExc}. Further, time constants are tuned to match the spike times of software simulation. The synaptic and membrane time constants (τ_{syn}, τ_{mem}) are set as 1.5 µs and 10 µs respectively, with a refractory period (τ_{ref}) of 2 µs (1000 times faster than biological real-time). The spiking threshold is 870 mV, while leak and reset potentials are set at 600 mV. The random spike train stimulus containing 32 $\times$ 20 events is injected into both the on-chip neuron as well as the software model neuron. The event number here indicates that 20 events are evoked by each synapse connected to a single neuron. The input stimuli as well as the resulting membrane response is plotted in Fig. 4.46. The gray traces are the inputs, while the blue ones are the membrane responses. For the software model the synaptic input current is shown, whereas for the emulated hardware neuron, the synaptic input line is monitored, since the line voltage drop is proportional to the charge of each incoming synaptic event. For the on-chip neuron, therefore, $\Delta V_{\text{syn}} = V_{\text{syn}} - V_{\text{synExc}}$ is plotted. The analog I/O circuit allows us to trace this voltage line or membrane potential one at a time. A comparison of the membrane response of the two neurons shows that both neurons evoke exactly four spikes, having similar spike times as well as the time course that develops as a result of various time constants. This demonstration has originally been done for [93] together with Andreas Hartel and reproduced with Yannik Stradmann [79]. The network parameters are provided by Paul Müller.

4.15 Discussion

This chapter presented the silicon implementation and design of the first spiking neuron model targeted for the HICANN-DLS's ANC architecture. The LIF emulation is characterized in detail, realized as a modular circuit architecture. Individual subcircuits have been characterized and presented with measurements from a single or multiple chip dies. The measured statistical data from various subcircuits conformed to the pre-taped-out Monte Carlo samples acquired from simulation results. This verifies the accuracy of the technology model files as well as the statistical variation models.

These prototypes also tested the ANC architecture shown in Fig. 4.2. More specifically, the arrangement of local voltage and current cells distributing biases as well as digital bus in vertical columns in a narrow space of 11.76 µm is prototyped and tested. Although the bias lanes are closely packed, no noticeable crosstalk detrimental to neuron operation has been noticed. Further a comparison of the on-chip neuron with a software model neuron shows a close correspondence in the spiking dynamics. On the circuit level, the capacitive memory's 10-bit tunable resolution has been found to be beneficial for tuning various circuits within the neuron. The reduction of bias currents available from current cells to sub 60 nA range (up to 15 nA) has remarkably improved the tuning ability of time constants for various subcircuits, e.g., short synaptic time constants. For the voltage cells, the possibility of stable neuron potentials have been verified as well as the available

range for reliable neuron operation. In the work towards [79] a winner-take-all network has also been demonstrated.

On the subcircuit level, the neuron presented the synaptic input which reduces the capacitor area by realizing it from line parasitics alone. An architecture which realizes a passive RC integrator with a transconductance stage made manifold improvement compared to the first generation BrainScaleS implementation [99]. First, the architecture uses a grounded resistor which is simpler in terms of design complexity, compared to a floating resistor. Secondly, it saves the area, power, and eventually the unwanted input-offset that would come from an integrator amplifier in the active RC solution. The use of bulk-drain connected devices allow a compact and high-value tunable resistor, with moderate power consumption compared to other solutions. The current implementation however relies on current-based synapses to avoid a complex design in the first phase. Conductance-based synapses have been explored in the next revision (although not implemented in the *DLS-3* chip). Linearized transconductance amplifiers are used both in synaptic input as well as the leak term. However, due to the limited input differential range at synaptic input (maximum 0.2 V) and the less critical role of synaptic OTA in determining the synaptic time constant (compared to the leak OTA for tuning τ_{mem}), it turns out to be a more suitable solution.

The enhanced OTA provides a separate bias for source-degeneration MOS transistor. An independent control of the device is to evaluate the results from the decreasing gain, trading it off for wide linear range. From the results presented one can conclude that very wide linear range comes at the cost of increased offset and inability to calibrate the leak OTA. Therefore a revised architecture is required for wide linear operating range. The achieved neuron circuit specifications including the range of time constants are summarized in Table 4.6.

While a very large range of refractory period τ_{refr} has been implemented, the range of achieved synaptic time constant τ_{syn} is limited compared to the targeted specifications of Table 3.1. The residual statistical spread for longer synaptic time constants after calibration is still large, which can lead to large variation at the network level. The range of synaptic time constants (τ_{syn}) and this variation needs a design improvement in a future revision. The synaptic output offset cancellation, however, has been demonstrated using V_{syn} and the offset correction bias $I_{\text{biasSynOff}}$, reducing the residual output offset below 4 nA from each synaptic input.

The read-out amplifier in the LIF prototypes directly drives the pad. Since it is a 2.5 V amplifier, the output-range is more than desired. On the lower side it achieves 0.1 V which is lower than the specified 0.2 V. Its large power consumption is less of a concern, since it is usually switched-off and is only enabled for neurons whose voltages are to be read-out. The same holds for the source followers that buffer the voltage levels from the synaptic input lines. The read-out amplifier has an input offset of 14 mV, which can be improved by slight design optimization, e.g., the amplifier in synaptic input of *DLS-1* trims the input-offset to 6 mV. To ensure stability under most conditions, the two-stage amplifier uses a large 600 fF compensation capacitor with indirection compensation scheme. The

Neuron model	Leaky I&F
No. of neurons	32
Voltage supply	2.5/1.2 V
Process	65-nm CMOS
Speed-up (acceleration) factor	$\times 1000$
Global parameters	1 voltage bias [1]
Local (individual) parameters	18 (14 I-bias, 4 V-bias) [1]
Configurability	15-bit digital bus
Membrane capacitor (max.)	2.36 pF (2-bit configurable)
Input synaptic event (max.)	10 µA, 4 ns pulses [2]
Area (single neuron)	11.76 µm $\times$ 200 µm
Area (array of 32 neurons)	376 µm $\times$ 200 µm
τ_{refr} (min./max. range) [3]	[1.11, 1.24] µs – [137.5, 104.5] µs
τ_{syn} (min./max. range) [3]	[1.24, 1.41] µs – [20.5, 13.6] µs
τ_{mem} (min./max. range) [3,4]	[0.35, 0.39] µs – [16.4, 14.1] µs

[1] available from on-chip tunable capacitive memory cells
[2] amplitude and length of each current pulse emitted by the synaptic circuit
[3] single-sided $[1\sigma, 3\sigma]$ quantiles of 84.13% (99.86%)
[4] measured using C_{mem} = 2.36 pF; the min./max. τ_{mem} $(1\sigma, 3\sigma)$ for C_{mem} = 570 fF is estimated to be 0.08, 0.09 µs and 4.11, 3.54 µs respectively

Table 4.6: The achieved specifications of the LIF neuron array.

large compensation capacitor is kept for large off-chip loads. Indirect compensation reduces power and increases bandwidth, with a potential to reduce area by choosing a smaller compensation capacitor. The revision in *DLS-3* restricts the size of this capacitor to 92 fF with the introduction of two stage buffering scheme.

The pulse generation architecture for digital spike generation is mostly similar to the one reported in [99], except for the delay element. The slow moving signal at the input of inverters results in leakage current during output pulse generation interval. This is similar to what is shown in the refractory reset circuit, which causes leakage current consumption, reducing the reported energy efficiency. This power consumption therefore increases with long refractory times. The same holds for the output inverter inside the delay element. Slow inverters should therefore be avoided to conserve energy. The revised refractory period circuit in *DLS-3* is therefore a digital counter based implementation, where no slow analog integration occurs. Finally, the addition of membrane integration bypass as reported in Sec. 4.10 adds a very useful debug feature in the system – especially for the initial digital tests for event routing, for example, during the evaluation of STDP mechanism together

with the PPU.

Chapter 5

Emulation of the Adaptive Exponential I&F Model

This chapter describes the implementation of the AdEx circuit designed as an extension of the existing LIF architecture. The modular architecture of the LIF neuron allows us to seamlessly add the adaptation and exponential circuits.

5.1 Neuron Circuit

The emulated neuron circuit in the third prototype of the chip (*DLS-3*) is not only an upgrade to the AdEx model, but also incorporates a number of other enhancements and modifications to the LIF circuits. They include:

- adaptation circuit
- exponential circuit
- integrated SRAM array
- 6-bit tunable membrane capacitor
- extended analog I/O
- fixed current bias distributor
- spike comparator and membrane offset input

The circuit also goes through the following modifications which are implemented outside the AdEx neuron array by other designers:

- conductance-based reset by using merged leak/reset concept and multi-compartment extensions (designed by Johannes Schemmel)
- digital neuron control for configuration of refractory period, adaptation pulses as well as SRAM decoder (designed by Gerd Kiene)

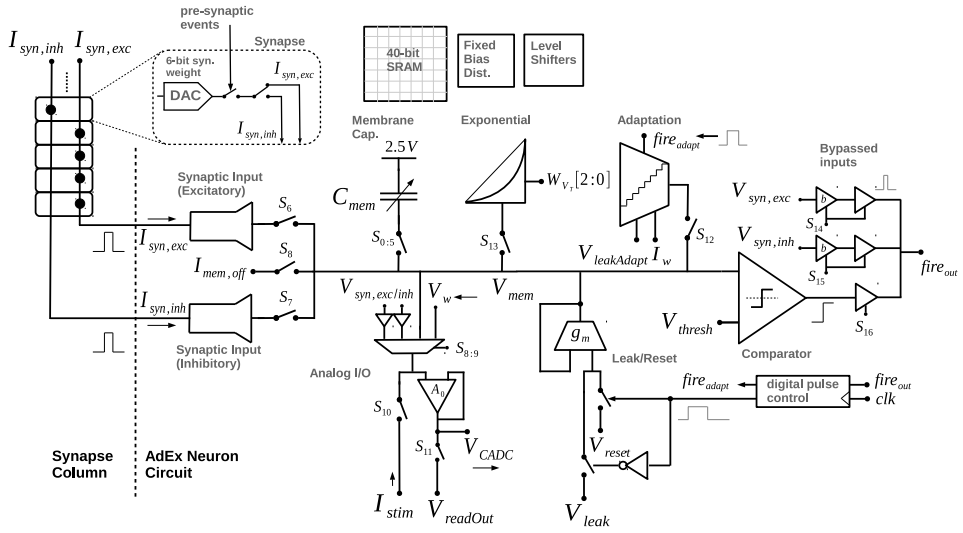


Figure 5.1: The full circuit schematic of the implemented AdEx neuron model.

The schematic of the neuron circuit designed for the *DLS-3* prototype is shown in Fig. 5.1. In addition to the main circuit components of the LIF neuron, the circuit integrates the exponential and adaptation circuits. The exponential circuit is controlled by a single 3-bit digital V_{T} parameter (called W_{V_T}) and connected with a switch S_{13} to the membrane. The adaptation circuit uses a current source I_{w} to integrate charge on the adaptation capacitor during the input pulse duration $\mathrm{fire}_{\mathrm{adapt}}$. It is endowed with its separate leak potential $V_{\mathrm{leakAdapt}}$. A 6-bit tunable capacitor selected by $S_{0\text{-}5}$ forms the neuron membrane. On the left side, the circuit integrates input synaptic current pulses (excitatory, inhibitory) as in the LIF architecture. Between the two is the $I_{\mathrm{mem,off}}$ current bias switched by S_8. This is an output offset trimming input previously implemented as an output offset canceling bias inside the synaptic OTAs. This has been moved outside onto the membrane, where additionally, it may inject input current to the membrane. The analog I/O circuit reads out the membrane V_{mem} and the input synaptic potentials $V_{\mathrm{syn,exc}}$ (excitatory) and $V_{\mathrm{syn,inh}}$ (inhibitory) via a source follower. Additionally, in this revision, it can read out V_{w}, the voltage on the adaptation capacitor. The read-out buffer A_0 designed for the LIF implementation reduces its size, since it does not drive the off-chip load directly anymore. The output V_{readOut} is buffered via a read-out chain and described later in Sec. 5.5. This neuron circuit (and the entire neuron array) is therefore only connected to the output with a single pin I_{stim}. I_{stim} is a shared input to all neurons in the array. The output of the read-out buffer can also be connected to the correlation ADC. The output labeled V_{CADC}, prior to the transmission gate, terminates at the input of the correlation ADC and allows, for example, to digitize the neuron membrane on a per-column basis. A voltage comparator circuit replaces the spike (pulse) generator circuit of *DLS-2* implementation, where it outputs voltage levels V_{OH} or V_{OL}. This output level $\mathrm{fire}_{\mathrm{out}}$ is interfaced directly to a digital neuron

control block, where it triggers a counter-based programmable delay. This digital input is fed back from the digital interface and is used to reset the membrane to the reset potential V_{reset}. The delay interval of the programmable counter, initiated by the comparator toggle, therefore implements the refractory mechanism. The details of the circuit can be found in [94]. The *DLS-3* neuron circuit introduces a merged leak and reset circuit. During neuron integration interval, the counter input is low, which connects the neuron membrane to the V_{leak} potential via the leak term. Once the input from the counter is high – an indication of a digital spike event, the neuron is reset to V_{reset} via the same transconductor. This transconductor is introduced in the reset path to realize a conductance-based reset. To enable a high conductance path to reset potential, the OTA enables its high-conductance mode during the reset interval (see high-conductance mode in Sec. 4.4). The merged leak/reset circuit is not covered in this thesis and the details can be found in [92, 110].

The digital configuration bits of the neuron circuit in this implementation are stored locally in an SRAM array. A 4×10 array has been provided in each neuron circuit. The analog control is provided by 6 local voltage biases as well as 14 local current biases. The current biases in the neuron that do not require tuning, have been generated from a fixed analog bias called $I_{\text{refAnalog}}$ to save current parameters. The input $I_{\text{refAnalog}}$ is itself a tunable bias that generates and distributes current biases for fixed bias circuits. These include the read-out buffer, the buffer in the adaptation circuit and the spike comparator. The circuit utilizes level shifters to shift the voltage levels from 1.2 V to 2.5 V, for example, to translate 1.2 V digital enable signals for 2.5 V transistors.

Before describing the individual circuits, we take a brief look at the chip architecture.

5.2 Chip Architecture

A simplified schematic of the *DLS-3* chip architecture is shown in Fig. 5.2. The schematic is divided into two parts – analog full custom implementation and the digital logic core including the PPU. On the top left half, we see synapse drivers which feed the pre-synaptic input events from the digital OMNIBUS into the 32 column ANC array. Each synapse driver controls two rows in the synapse matrix. Each of the synapse circuits in the matrix modulates events through an output DAC and feeds them on either of the two shared excitatory and inhibitory event lines. At the top end, correlation ADCs read-out two correlation voltages per synapse for implementation of STDP together with the PPU. Each of the AdEx neurons in the array (edge-connected to the synapse matrix) integrates input current from the two lines in each column. The membrane of the neuron is reset via the digital neuron control which sits beneath the Capmem arrays. The digital neuron control transmits the digital event via a priority encoder to the digital backend for event routing and controls the duration of adaptation pulses. The merged leak/reset circuit as well as inter-compartment switches are implemented in the leak/MC extension of the

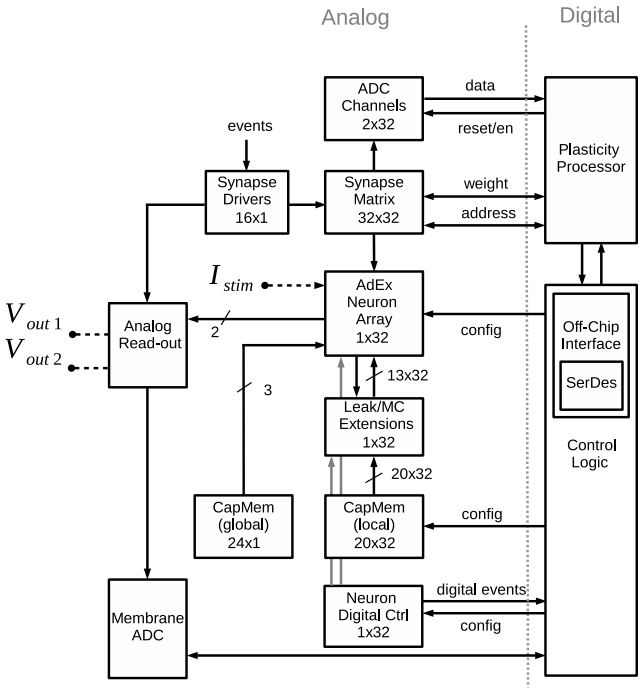


Figure 5.2: The architecture of the third prototype of the HICANN-DLS chip (*DLS-3*). The block level implementation is simplified to provide a general overview.

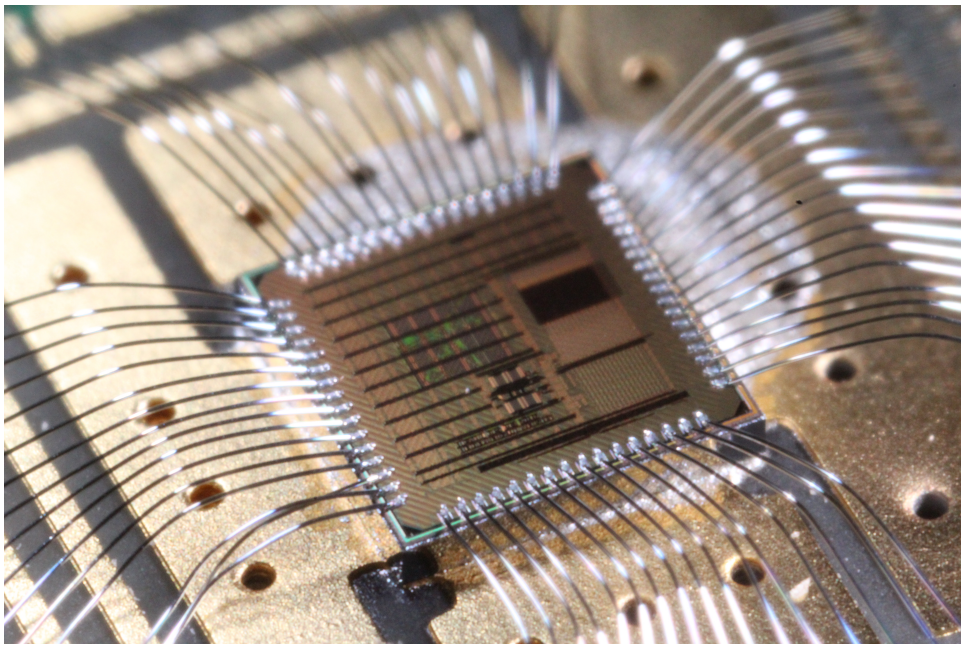


Figure 5.3: The third prototype of the HICANN-DLS chip. Photo by the Author.

neuron and the AdEx analog neuron provides it with SRAM configuration bits and routes the membrane voltage line. The chip features an extended read-out scheme and two simultaneous voltages from the neuron array can be read out off-chip. This is implemented in the analog read-out block, which multiplexes debug inputs from the neuron array as well as the synapse drivers. The neuron membrane can be digitized by the high-speed MADC and can in turn be read by the PPU via the digital interface. Out of the twenty local biases per neuron, seven are used in the leak/MC block, while the remaining 13 tune the AdEx analog neuron circuit. Three current biases are tunable globally (shared among all neurons).

5.3 Adaptation Circuit

The adaptation circuit implements the spike triggered adaptation and subthreshold conductance given by Eq. 3.8 and Eq. 3.6 and is shown in Fig. 5.4. The circuit is inspired by the first generation BrainScaleS neuron [99]. However, it modifies the architecture significantly and realizes both positive and negative values of the model parameters a and b. In the implemented circuit, the adaptation current w is generated as the output of a source degenerated OTA, with its input terminals sensing the difference between V_{leak} and V_{w}. The OTA conductance g_{a} implements the subthreshold adaptation conductance a. The two input terminals of the OTA can be flipped using a digital enable bit en_{Va} to realize negative values of a. This is implemented by adding two 2×1 transmission gate multiplexers at the OTA inputs with en_{Va} as their select line.

Eq. 3.8 is realized by connecting a tunable conductance g_{w} between the buffered membrane and the node V_{w}, the voltage across the adaptation capacitor C_{w}. This is shown in the lower right half of Fig. 5.4. A tunable floating resistor realizes the conductance g_{w} and a compact 1.2 V OTA has been designed for the buffer A_{w}. When a spike occurs, the adaptation circuit gets a digital input $\text{fire}_{\text{adapt}}$ with an adjustable pulse-width duration $t_{\text{fireAdapt}}$. This triggers the spike frequency adaptation part of the circuit implemented by a charge pump [142, 143]. Every incoming pulse integrates a small charge q equivalent to $I_{\text{w}} \cdot t_{\text{fireAdapt}}$ on the adaptation capacitor. The current I_{w} is a local Capmem parameter that can be tuned to set the Δq. Additionally, the charge q can either be integrated or removed from the membrane, depending upon the digital configuration bit en_{Vw}. When en_{Vw} bit is asserted, the current I_{w} is integrated on the capacitor C_{w} via the switch S_{p}, while S_{n} is held low. In case of charge removal (en_{Vw} is low), the current I_{w} is mirrored via the cascode current mirror (transistors $M_{1\text{-}4}$) and sinks the charge via the enabled switch S_{n}. This implementation lets us enable both accelerating and decelerating spike triggered adaptation.

The presence of a level shifter LS indicates that a 1.2 V $\text{fire}_{\text{adapt}}$ input pulse is shifted up to 2.5 V, to trigger the thick-oxide pass-transistor switches $S_{\text{p}}/S_{\text{n}}$. Note that the Capmem output stages are 2.5 V circuits and when biasing 1.2 V core (thin-oxide devices) overvoltage protection needs to be considered to prevent drain

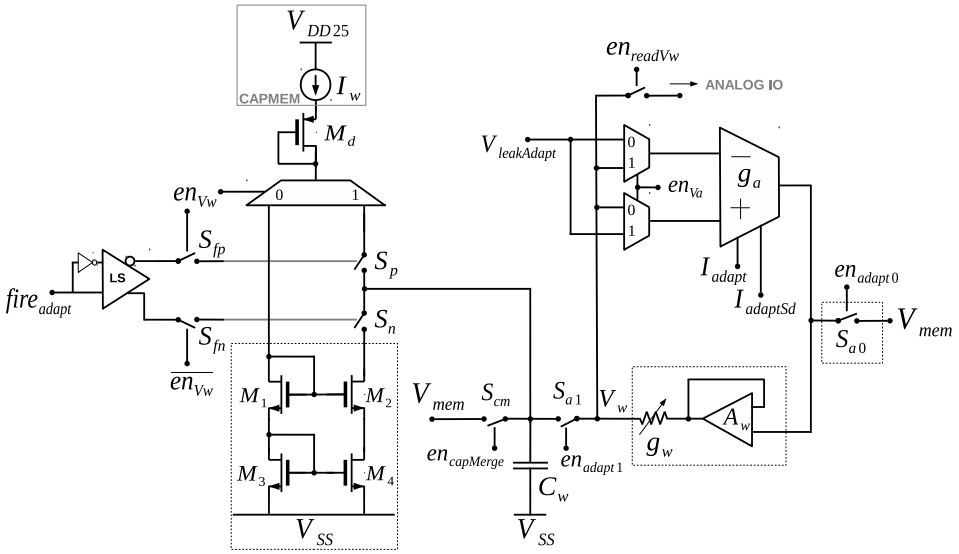


Figure 5.4: The schematic of the implemented adaptation circuit.

leakage. Due to the large output impedance of the Capmem current cells, nominally this is a not a problem, except when the input impedance of the receiver circuit is considerably high. In the adaptation circuit, the presence of the diode M_{d} drops the output potential of the current cells (two diodes in series, one shown). Alternately, one can mirror the Capmem current source twice, and tie it to 1.2 V supply in the second mirror. The solution however has not been adopted to reduce the additional current mismatch. In the schematic, all circuits are 2.5 V (thick-oxide) designs except those marked by the dashed boxes. All switches are thick-oxide transmission gates, except $S_{\mathrm{p,n}}$ which are 2.5 V pass transistors, and the membrane connecting S_{a0} which is 1.2 V transmission-gate. The buffer, the conductance and the cascode current mirror of the charge pump are 1.2 V thin-oxide designs. The sizing and resistance of transmission gates are identical to those of Sec. 4.9. The OTA used for buffering the membrane is a current-mirror OTA with n-type input-stage. A p-type input stage of the same architecture has previously been described in Sec. 4.4.

Note the presence of enable switches driven by digital configuration bits $\mathrm{en}_{\mathrm{adapt0}}$ and $\mathrm{en}_{\mathrm{adapt1}}$, as well as a third switch driven by $\mathrm{en}_{\mathrm{capMerge}}$. This switching scheme was thought out together with Paul Müller during the design phase. The switch S_{a0} is an output enable switch. Switches S_{a1}, S_{cm} are meant to disconnect the capacitor from the adaptation, when the latter is not used. In that case, $\mathrm{en}_{\mathrm{adapt0,1}}$ will both be low, whereas $\mathrm{en}_{\mathrm{capMerge}}$ will be asserted to connect the C_{w} to the membrane V_{mem}. In the nominal case (when adaptation is to be used), $\mathrm{en}_{\mathrm{capMerge}}$ is kept low, while the enables $\mathrm{en}_{\mathrm{adapt0,1}}$ are asserted. Since the two enables $\mathrm{en}_{\mathrm{adapt0,1}}$ are activated and deactivated together they can be combined in a single enable bit. Note that when adaptation is disabled, there is still a provision to inject/sink additional current using the charge pump's current source I_{w}. This is

suggested to sink additional current from the membrane, for example, in case the conductance-based reset is not strong enough.

5.3.1 Tunable Floating Resistor

The conductance g_{w} of Eq. 3.8 in the adaptation term has been implemented as a tunable bulk-drain connected resistor. Previously, Sec. 4.3.5 introduced the bulk-drain connected devices and an implementation of a single-sided resistor for synaptic input circuit was discussed in Sec. 4.3.6. The adaptation resistor, on the other

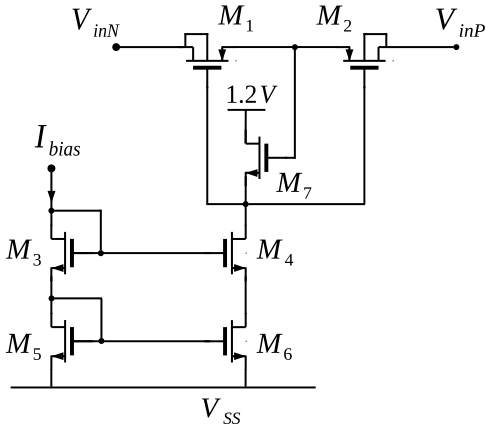


Figure 5.5: The simplified schematic of the floating tunable bulk-drain resistor used inside the adaptation term.

hand is implemented as a floating resistor – which means that either of the two terminals can be at a potential higher than the other. Another major difference from design point of view is the tunable resistive range. According to Table 3.1, the range of adaptation time constant τ_{w} is between 16 µs and 600 µs (equivalent hardware time). This is indeed a large tuning range and for practical implementation more importance is given to cover the lower half of this range. Compared to this, the synaptic resistor implements a lower resistance range (τ_{syn} are of the order of 1 µs – 19 µs), and to accommodate for lower resistance range, it is biased in linear region. The adaptation resistor is instead biased in sub-threshold region as proposed in [122, 125].

The circuit designed for the dual-sided bulk-drain connected resistor is shown in Fig. 5.5. Between the two terminals of the resistor, V_{inP} and V_{inN}, are two series PMOS bulk-drain connected devices labeled M_1 and M_2. Both devices are connected back-to-back, such that their source terminals are connected. This arrangement lets the resistor be dual-sided and depending upon the potential difference, one device provides higher resistance than the other. The total resistance contributed by the resistor is hence R_{SD1} + R_{SD2}. Where R_{SD} denotes the resistance between the drain and source terminal of each PMOS device. When $\Delta V > 0$ V, $R_{\mathrm{SD2}} \gg R_{\mathrm{SD1}}$, whereas for $\Delta V < 0$ V the role flips, and $R_{\mathrm{SD1}} \gg R_{\mathrm{SD2}}$. In this

way, only one of them contributes to the total resistance in the two conditions. The gates of both bulk-drain connected devices are controlled by a biasing circuit – a cascode current mirror formed by devices $M_{3,4}$ and $M_{5,6}$, as well as M_7 that sets the eventual bias point.

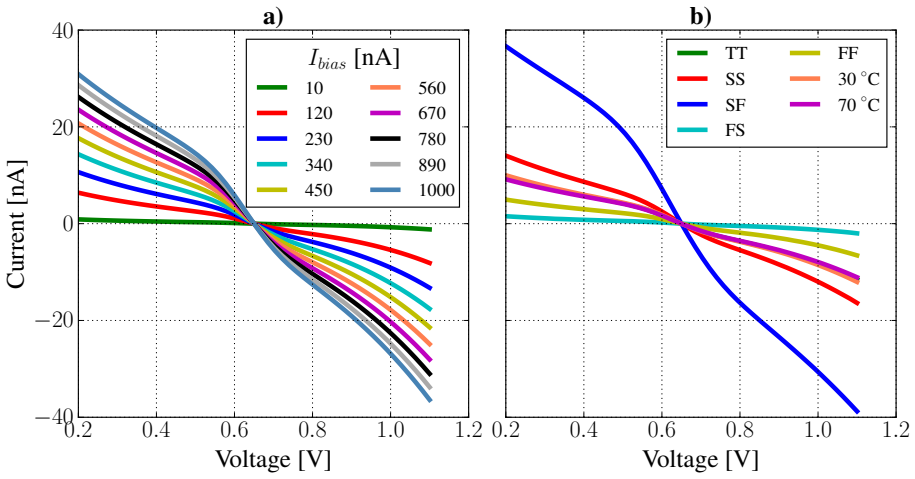


Figure 5.6: Simulated I-V characteristics of the adaptation resistor: a) Nominal corner with swept I_{bias}. b) Process corner and temperature sweep at I_{bias} = 0.5 µA.

The resistance R_{SD} realized by the given PMOS device can be derived [124] using the EKV model [111] as

$$R_{\text{SD}} = \frac{nU_T}{I_{\text{SD}}} \left[\frac{\left(e^{\frac{V_{\text{SD}}}{U_{\text{T}}}} - 1 \right)}{(n-1)\left(e^{\frac{V_{\text{SD}}}{U_{\text{T}}}} + 1 \right)} \right] \tag{5.1}$$

where the resistance is controlled by V_{SG} that varies the current I_{SD}. For a negligible voltage drop V_{SD}, one can write

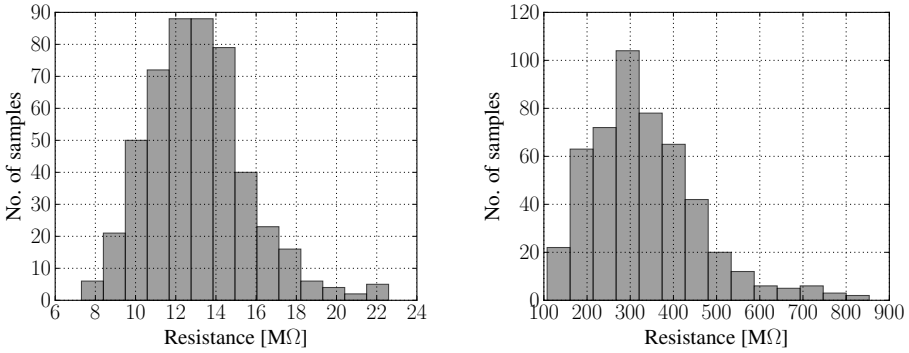
$$R_{\text{SD}}\Big|_{V_{\text{SD}}=0} = \frac{U_T}{I_0} \cdot e^{\frac{-V_{\text{SG}}-V_{\text{th0}}}{nU_{\text{T}}}} \tag{5.2}$$

where the current $I_0 = 2n\mu C_{\text{ox}} \frac{W}{L} U_T^2$, which highlights the dependence of resistance on temperature. The simulated I-V characteristics of the designed resistor is shown in Fig. 5.6a. The bias current I_{bias} which tunes the resistance is swept between 10 nA and 1 µA in equal intervals. One terminal is kept fixed at 0.65 V, while the other is swept from 0.2 V to 1.1 V. The resistance increases more linearly at higher bias currents than the lower values, which results in a very large resistance.

The circuit is further simulated for corner variations and temperature differences in Fig. 5.6b for a fixed bias current of 0.5 µA corresponding to 20.5 MΩ resistance. Note that the results from skewed process corners (slow-fast and fast-slow) are at extremes and contribute a very large (125.3 MΩ) or smaller resistance

(5.7 MΩ) compared to other corners. This is likely due to the presence of an all NMOS biasing stage that sets the common-mode of the series PMOS transistors. It needs to be seen how much calibration can help to compensate for the aforementioned variations. Compared to process corners, the resistor is more immune to temperature changes, as visible by temperature variations to 30°C or 70°C in Fig. 5.6b.

The resistor has been designed to cover a range between 13 MΩ up to at least 331 MΩ (mean values). The histograms of Fig. 5.7a and Fig. 5.7b show the statistical variation due to device mismatch whose 1σ spread is 2.5 MΩ and 127.1 MΩ. The reported resistances are evaluated at bias currents of 1 µA and 20 nA respectively.



(a) Minimum resistance.

(b) Maximum resistance.

Figure 5.7: Statistical variation in the minimum and maximum resistance that can be tuned in the adaptation resistor.

5.3.2 Low Voltage Buffer

The adaptation circuit needs an amplifier to buffer the membrane potential as shown in Fig. 5.4. The circuit is meant to be low-power and compact, and therefore a current mirror OTA (CM-OTA) is designed using thin-oxide 1.2 V transistors only. The schematic of the OTA is shown in Fig. 5.8. The OTA is fully symmetric and has an n-type input stage to allow the range of membrane potential – typically between 1.2 V and 400 mV. The circuit is different than that of Sec. 4.3.2, which is a 2.5 V p-type input stage design with low input offset requirement. The input bias current supplied to the amplifier is distributed via the bias current distributor block outlined in Sec. 5.7. As described in Sec. 4.3.2, the design does not require a compensation capacitor, thereby saving area. The amplifier open-loop DC-gain can be mathematically expressed as

$$A_{OL} = \frac{V_{\text{out}}}{V_{\text{in}}} \approx \frac{g_{\text{m2}} \cdot g_{\text{m6}}}{g_{\text{m4}}(g_{\text{DS6}} + g_{\text{DS7}})} = \frac{K \cdot g_{\text{m2}}}{(g_{\text{DS6}} + g_{\text{DS7}})} \tag{5.3}$$

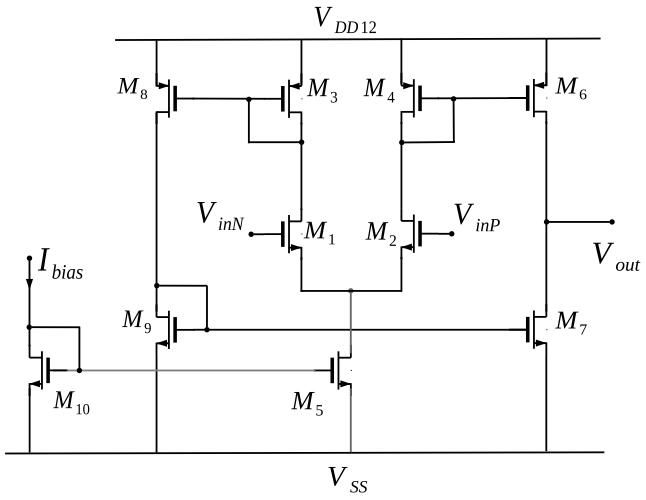


Figure 5.8: The schematic of the opamp circuit used inside the adaptation circuit.

where K is the ratio of transconductances g_{m6} and g_{m4} and is kept at 5 times. The total achieved gain in the amplifier is 31.9 dB. To compare how good the approximation of Eq. 5.3 is, one can plug in the values from the simulator, which evaluates to 32.6 dB of open loop gain. The expressions therefore provide a reasonably good estimate. Due to the usage of thin-oxide transistors (low intrinsic-gain) the achievable gain is low. The low gain comes from input stage mirror loads, which reduce input stage gain to approx. $\frac{g_{\mathrm{m2}}}{g_{\mathrm{m4}}}$. A typical solution to increase gain is the use of cross-coupled input-stage in parallel to the mirror load $M_{3,4}$. For the current version this is however not implemented. High gain in buffers is desirable since $A_{\mathrm{CL}} = \frac{A_{\mathrm{OL}}}{1+A_{\mathrm{OL}}}$, where A_{CL} is the close-loop gain.

The unity gain bandwidth of the OTA is $\frac{K \cdot g_{\mathrm{m2}}}{C_{\mathrm{L}}}$ and the achieved value with a capacitive load of 100 fF is 23.5 MHz. In unity-gain configuration the amplifier achieves a –1 dB bandwidth of 22.5 MHz when loaded with a 100 fF output capacitor. In the same configuration it achieves a positive slew rate of 15.7 V/µs and a negative rate of 11.3 V/µs. This corresponds well to $K \cdot I_{\mathrm{tail}}/C_{\mathrm{L}} = 5 \cdot 260$ nA/100 fF = 13 V/µs. The amplifier compensates itself using the dominant pole, which entails that a larger output load makes it more stable, at the expense of decreased bandwidth. An output load of 100 fF forms the dominant pole ($\frac{g_{\mathrm{DS6}}+g_{\mathrm{DS7}}}{C_{\mathrm{L}}}$) and reducing it makes the amplifier less stable – as it comes closer to the non-dominant pole ($\approx \frac{g_{\mathrm{m4}}}{C_{\mathrm{par}}}$). The non-dominant pole is contributed by the parasitics C_{par} at the gate of transistors $M_{4,6}$ (mostly C_{GS}) and drains of $M_{2,4}$. For a capacitive load of 50 fF, the open loop amplifier achieves a phase margin of 57.5°, which increases to 70.5° for a 100 fF capacitive load. Note that when placed inside the adaptation term the amplifier sees the RC filter formed by the tunable resistor and the 2 pF adaptation capacitor at its output, which improves the phase margin.

Fig. 5.9 shows the input-referred offset voltage of the amplifier simulated using the device Monte Carlo models. The offset is 15.6 mV estimated for a 1σ (single

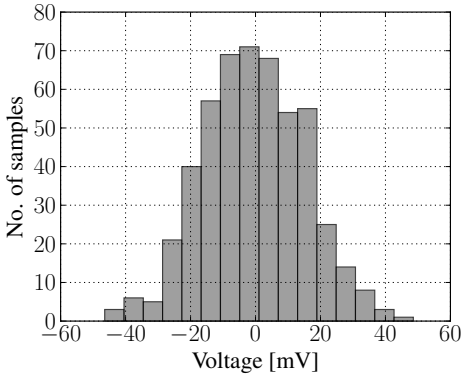


Figure 5.9: The distribution showing input offset voltage of the low voltage amplifier simulated across 500 samples using Monte Carlo device models.

sided quantile) with a mean of 531 µV. In the adaptation circuit, the effect of this offset can be compensated with the voltage $V_{\text{leakAdapt}}$. The amplifier is biased with 200 nA input current that is multiplied inside. The total current consumption is about 1.5 µA, corresponding to 1.82 µW. The achieved specifications of the amplifier are summarized in Table. 5.1.

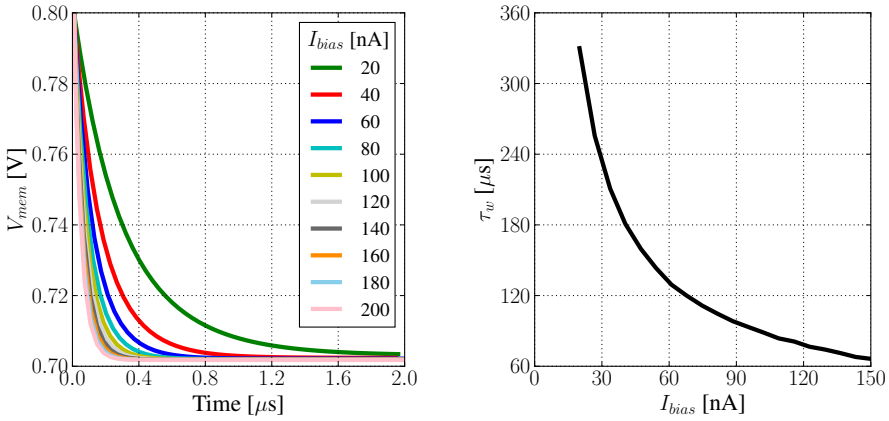
–3 dB bandwidth[1]	36.4 MHz
–1 dB bandwidth[1]	22.5 MHz
DC Gain[2]	31.9 dB
Current consumption	1.5 µA
Power consumption	1.82 µW
Input offset	15.6 mV
Slew rate	11.32 V/µs
Phase margin[2] @C_{L}=100 fF	70.5 °
Phase margin[2] @C_{L}=50 fF	57.5 °
Compensation capacitor	none
Voltage supply	1.2 V
MOS devices	core thin-oxide (std./low-V_{th})

[1] C_{L} = 100 fF
[2] Open loop configuration

Table 5.1: The achieved specifications of the OTA used inside the adaptation term.

5.3.3 Full Circuit Characterization

Having reviewed the internal design of conductance g_w and the buffer A_w, the full circuit performance of the adaptation circuit is discussed in this section.



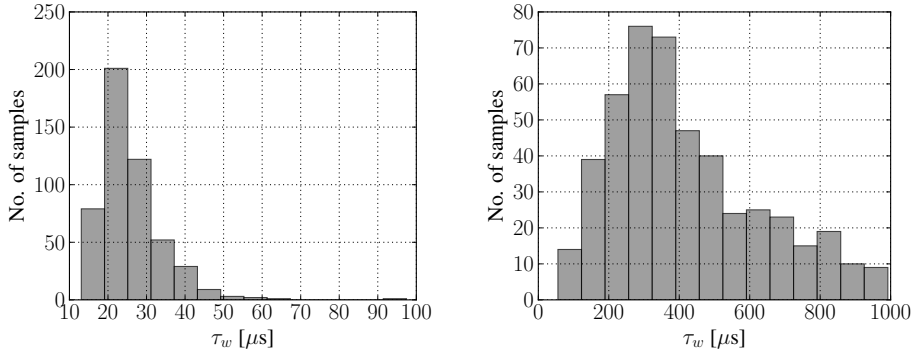
(a) Decaying traces showing the adaptation time constant τ_w.

(b) τ_w as a function of resistor's bias.

Figure 5.10: Simulating the adaptation time constant, controlled with the bias of the floating resistor.

The adaptation time constant τ_w is given by $C_w/g_w = R_w C_w$. Certain deviations could however be expected due to the dependency of resistance on the voltage applied across its terminals, as well as the corner and mismatch variations. As shown earlier, the adaptation resistance can be tuned to set a range of time constants. Fig. 5.10a shows the voltage curves decaying with a time constant τ_w from a held potential of 800 mV towards the resting potential of 700 mV. The bias is swept between 20 nA and 200 nA which yields time constants between 324 µs and 56 µs. The ideal tuning curve is plotted in Fig. 5.10b. The minimum and maximum time constants are further evaluated for device mismatch at the resistor bias of 15 nA and 1 µA and shown in Fig. 5.11a and Fig. 5.11b respectively. The time constants are again evaluated by a decay of 100 mV, from 800 mV towards a leak potential of 700 mV. The mean (μ) and 1σ spread for the shorter bias is 25.8$\pm$8.2 µs, whereas for the longest time constant it is 418.5$\pm$210.8 µs respectively. The large variation in longer time constants comes from the variation shown for resistor curves, e.g., in Fig. 5.6 and Fig. 5.7.

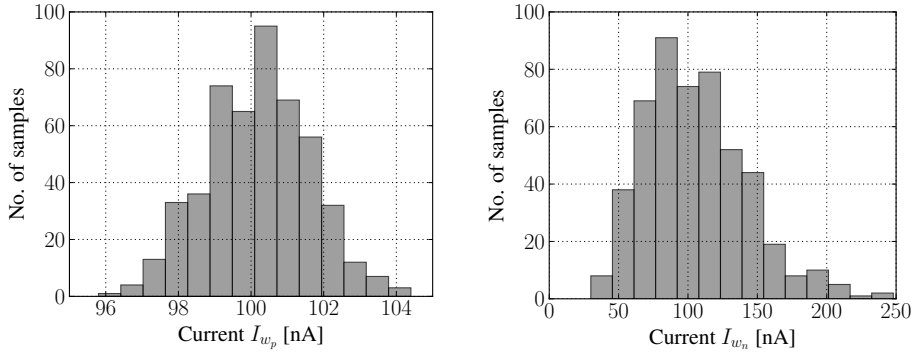
In the adaptation circuit architecture, accelerating and decelerating spike frequency adaptation is achieved by a charge pump, where the current source I_w integrates on, or removes the charge from the capacitor C_w. Since the current source is the same, the discharge is via a mirrored current source through the switch S_n. This leads to more variable amount of charge removal compared to integrated charge due to device mismatch. Shown in Fig. 5.12 is a distribution of mismatch among



(a) Minimum τ_{w} at I_{bias} = 1 µA. (b) Maximum τ_{w} at I_{bias} = 15 nA.

Figure 5.11: A distribution showing minimum and maximum values of adaptation time constants.

both mirrors. A direct current from capacitive memory having a mean and standard deviation [$\mu \pm 1\sigma$] of 100.2$\pm$1.42 nA, yields 105.7$\pm$36.5 nA for the mirrored negative current source. Fig. 5.13 shows the adapting membrane response of the neuron when stimulated with a constant fixed dc input current. Fig. 5.13a shows how the membrane adapts over time as a result of the growing voltage on the adaptation capacitor C_{w} with each spike event (shown in Fig. 5.13b). Once the configuration bit $\mathrm{en}_{\mathrm{Vw}}$ is switched off, with all other settings unchanged, the neuron membrane responds with accelerating adaptation, as shown in the Fig. 5.13c. The figure shows an accelerating membrane as a result of more net positive current on the membrane, due to decreasing voltage V_{w} (Fig. 5.13d).



(a) Variation in positive I_{w}. (b) Variation in negative I_{w}.

Figure 5.12: A distribution showing variations of the large negative source variation in comparison to the positive current source, for a fixed I_{w} of 100 nA.

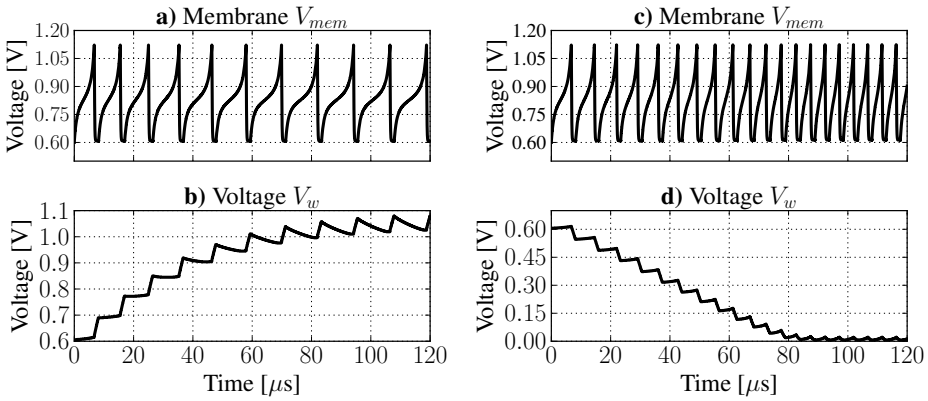


Figure 5.13: Membrane response of the neuron showing spike frequency adaptation in both directions, in response to a fixed DC input current. a) Decelerating membrane and its positively growing adaptation voltage (b). c) Accelerating membrane and its corresponding decreasing adaptation voltage (d).

Note that the positive and negative V_{w} in Fig. 5.13 are not symmetric. The step size ΔV_{w} between the two varies. This is first due to the non-negligible current that flows through the resistor and raises the voltage level outside the pulse interval, for both positive and negative V_{w} (with the chosen parameters). It is more pronounced at small time constants, compared to large ones, due to smaller resistance. Moreover, it is dependent on the potential difference across the resistor – which increases, when membrane is raised high by exponential term, close to the spike threshold and is the expected model behavior. As a result, during charging integration (positive increase in V_{w}), it increases the voltage level of V_{w} by up to a few millivolts right before a spike. During charge removal (decreasing V_{w}), it reduces the effective voltage drop ΔV_{w}, attained during the last pulse interval. This is visible more clearly in Fig. 5.14f and is discussed in the following text.

More importantly, during the pulse interval $\mathrm{fire}_{\mathrm{adapt}}$ (integration or charge removal), a short glitch occurs on the node voltage V_{w} as shown in the simulation results of Fig. 5.14b,e. The glitch is predominantly due to clock feedthrough (but also due to charge injection) through the pass transistor switches (S_{n} and S_{p}), driven by the pulse input of short transition period of 0.5 – 1 ns. The peak glitch is slightly higher in amplitude during charging in comparison to discharging, due to the mismatch of current sources (p-type near-ideal source from Capmem in comparison to a n-type cascode current mirror for the discharge path). The effect of this glitch is also visible in Fig.5.14c as an initial pedestal at the start of every pulse increasing V_{w}. Matched current sources [144] can be employed to remove such non-ideal effects.

In the linear region, the gate-source and gate-drain parasitic capacitances of a

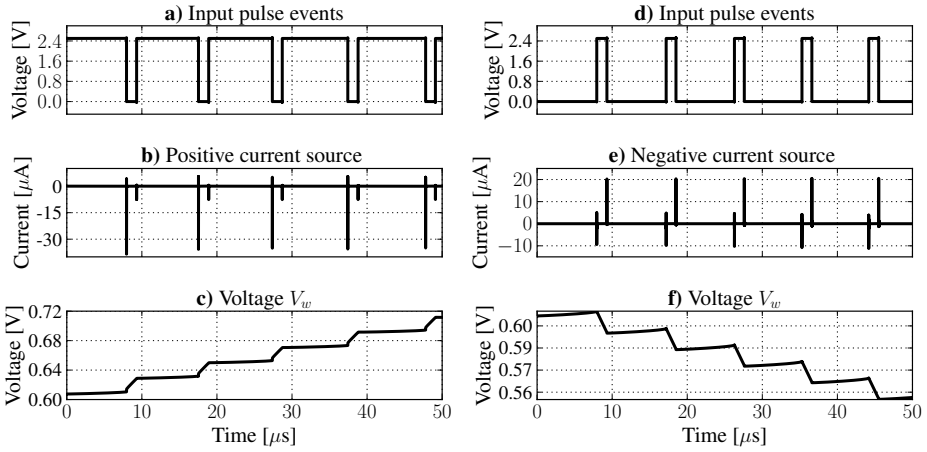


Figure 5.14: a) and d) The incoming pulses that enable positive and negative current sources via $S_{\text{p,n}}$. b) and e) The glitches due to the incoming clocks on the two current sources (sourcing/sinking 20 nA) enabled by $S_{\text{p,n}}$. c) and f) Resulting positive and negative increase on V_w changing with every input pulse.

pass switch is

$$C_{\text{GS}}/C_{\text{GD}} = \frac{1}{2}WLC_{\text{ox}} + WC_{\text{ov}} \tag{5.4}$$

where C_{ov} is the overlap capacitance between gate and source/drain (sometimes referred to as C_{GSO}). When the input pulse to the switch makes a transition from logic high to low, the clock voltage (2.5 V) makes a voltage divider between this parasitic and the adaptation capacitor C_{w}. If the transistor parasitic C_{GS} is as small as 1 fF, a maximum voltage pedestal of up to 1.25 mV can appear as error on the node V_{w}. Mathematically it is expressed as

$$V_{\text{err,cft}} = 2.5 \cdot \frac{C_{\text{GS}}}{C_{\text{GS}} + C_{\text{w}}} \tag{5.5}$$

Similarly, the error voltage due to channel charge injection can be written as

$$V_{\text{err,inj}} = WLC_{\text{ox}}(2.5 - V_{\text{w}} - V_{\text{th}}) \tag{5.6}$$

where W and L are the channel width and length of the transistor. In the charge pump implementation, both the pass transistors are prone to body-effect. Therefore, the error has a non-linear dependence upon the input voltage.

Fig. 5.15 shows the simulation results of swept current source I_{w} between 20 nA and 200 nA and incremented by 60 nA each time. The results for both positive (left) and negative V_{w} (right) are plotted. The small steps correspond to 20 nA, while the largest ones to 200 nA. V_{leak} is 0.6 V and all parameters are kept constant between the two runs. The plot shows that while positive V_{w} increase does saturate, the negative counterpart continues to decrease until it reaches the ground

potential. This is clearly due to the membrane that toggles between 0.6 V to 1.1 V, which in the positive case balances out the currents after multiple spikes, whereas a large potential difference between the membrane and the adaptation voltage V_{w} exists for the negative case. The saturating adaptation voltage V_{w} shown in Fig. 5.15a

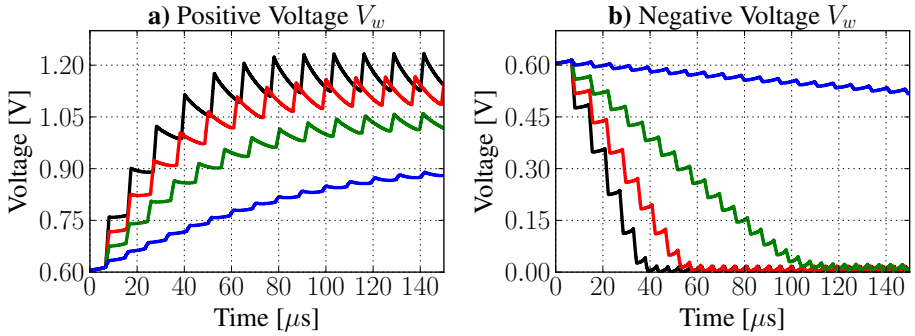


Figure 5.15: The positive and negative V_{w} voltage with swept input adaptation current I_{w} of 20, 80, 140, 200 nA ({blue, green, red, black}), where τ_{w} is fixed.

rises above 1.2 V depending upon the parameter settings. Fig. 5.16 shows two measured traces showing the read-out voltage on the adaptation capacitor in two different cases – one with higher input event rate (gray trace) and another with low input rate (black trace). While the black trace peaks at about 1.05 V, the gray trace saturates at 1.36 V. A higher than 1.2 V drain voltage may lead to drain leakage in 1.2 V thin-oxide transistors. The voltage rise however effects only during charge integration via S_{p}, where the only 1.2 V design is the resistor, whose resistance is usually hundreds of MΩ. In any case, the circuit should limit the voltage to 1.2 V by realizing the charge-pump as a 1.2 V circuit.

Several techniques are proposed in literature to compensate for the effects of charge injection and clock feedthrough [145]. One should avoid sharp clock transitions. Further, a quick solution could be to replace the switches with transmission-gates, where complementary signals will act to cancel the effect of each other. The usage of a half-sized dummy switch can also be employed [146]. Finally, one can also improve the circuit by matching the p- and n-type current sources, in order to have more symmetric curves. Low-glitch, matching current sources techniques are proposed in literature, for example in [147, 148] which could be employed too. The potential build-up on the capacitor C_{w} is limited on the lower end by the correct operation of the current mirror – since it also determines the minimum output voltage (common-mode) of the current mirror that will guarantee saturation regions. This lower limit is approximately 250 mV. The upper limit is that it may not go significantly beyond 1.2 V potential, as that would cause leakage in thin-oxide transistors of the tunable resistor or reduce its resistance. The off-state isolation of the pass transistors $S_{\mathrm{n,p}}$ is high, with a maximum leakage current of few tens of picoamperes, since they are both thick-oxide transistors.

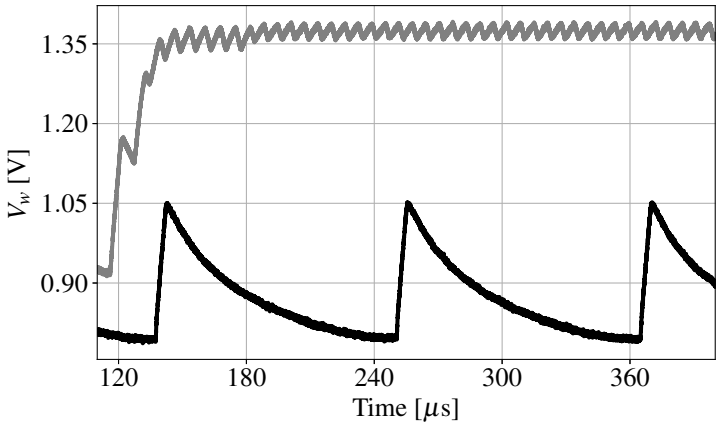


Figure 5.16: Measured results showing V_{w} traces in two different neurons. The neuron with higher input rate saturates the adaptation voltage at approx. 1.36 V (gray trace). The neuron with low input rate peaks at 1.05 V (black trace).

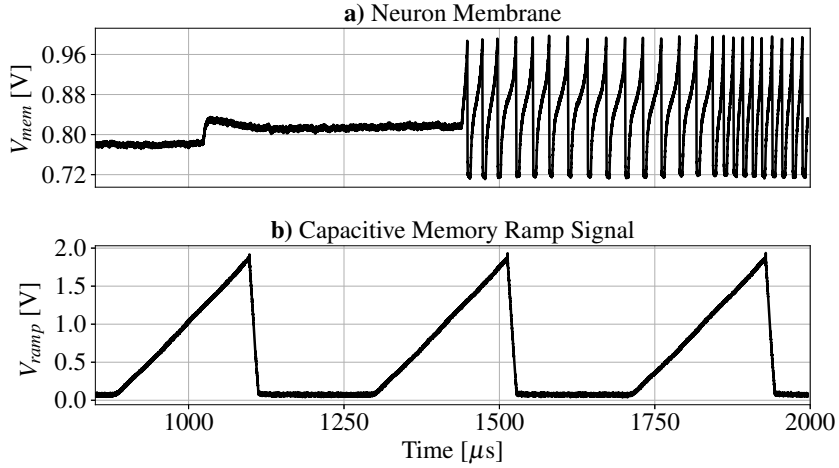


Figure 5.17: Measured results showing an erroneous behavior of the neuron, due to the crosstalk linked to the Capmem update cycle. a) The membrane potential changing its behavior as a result of the Capmem ramp update. b) The corresponding Capmem ramp measurement.

During chip measurements, significant crosstalk induced by the capacitive memory's refresh ramp has been encountered. The effect, as measured by simultaneous monitoring of the neuron membrane as well as the Capmem ramp signal is shown in Fig. 5.17. The Capmem ramp updates the stored capacitor value in voltage cells as well as the current cells (using a voltage-to-current converter). The figure shows that when the ramp updates the stored current bias value, the

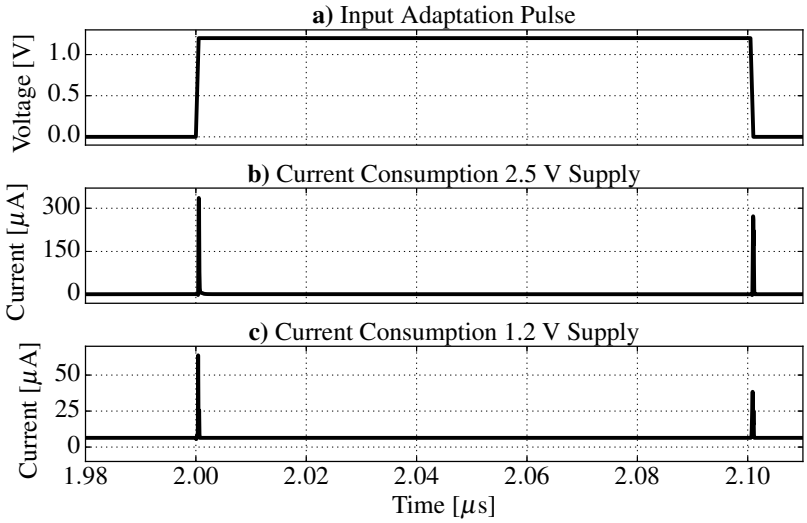


Figure 5.18: The current consumption of the implemented adaptation circuit. a) Incoming pulse event from the digital neuron control. b) The current consumed over time from the 2.5 V supply. c) The current consumed over time from the 1.2 V supply.

membrane behavior changes. The hypothesis for this behavior is the sensitivity of the current bias that tunes the leak/reset OTA to the Capmem ramp signal. When the value of the bias is refreshed by the incoming voltage ramps, the leak conductance changes as a result. This happens also when the ramp is reset. Since the exponential circuit is active, the abrupt perturbance in the leak conductance can lead the neuron to a different spiking behavior, than what it is configured for. The leak/reset circuit is not designed by the author and is outside the AdEx implementation described in this work. At the time of this writing, the circuit schematic has already been revised.

The static power consumed in the adaptation circuit is mainly distributed among the 2.5 V OTA, the 1.2 V adaptation buffer and the 1.2 V resistor. During the chip layout phase, as a mistake the bias distributed for the amplifier was set at 1 µA, rather than 200 nA. That caused the specified total current consumption of 1.5 µA to increase to 7.1 µA. The OTA current consumption is highly dependent on the bias settings, with 2.1 µA for $I_{\text{bias,biasSd}}$ of 200 nA/500 nA. The resistor typically consumes twice its I_{bias} setting – approximately 200 nA for I_{bias} = 100 nA. The dynamic current consumption of the circuit increases at the time of incoming pulse's rising and falling edges, as shown in Fig. 5.18. This is due to dynamic leakage in the level-shifters that convert the 1.2 V pulse signal to 2.5 V pulses (see LS in Fig. 5.4). The total static current consumed from 1.2 V and 2.5 V supplies in this case is 7.5 µA and 2.2 µA respectively.

5.4 Exponential Circuit

The exponential circuit emulates the exponential term given by Eq. 3.4. The circuit generates the exponential current by exploiting the weak inversion characteristics of a MOS transistor. When biased in the subthreshold domain, i.e., $V_{\mathrm{GS}} < V_{\mathrm{th}}$, and the drain-source drop (V_{DS}) is greater than 3–4 U_{T}, i.e., $V_{\mathrm{DS}} > 75$ mV (with $V_{\mathrm{S}} = 0$ for NMOS, V_{DD} for PMOS), the drain current is given [88] as

$$I_{\mathrm{D}} = I_0 \frac{W}{L} \exp\left(\frac{V_{\mathrm{GS}} - V_{\mathrm{th}}}{n U_{\mathrm{T}}}\right) \tag{5.7}$$

where n is the slope factor, U_{T} is the thermal voltage at room temperature, W and L are the width and length of the device and V_{th} is the device threshold voltage.

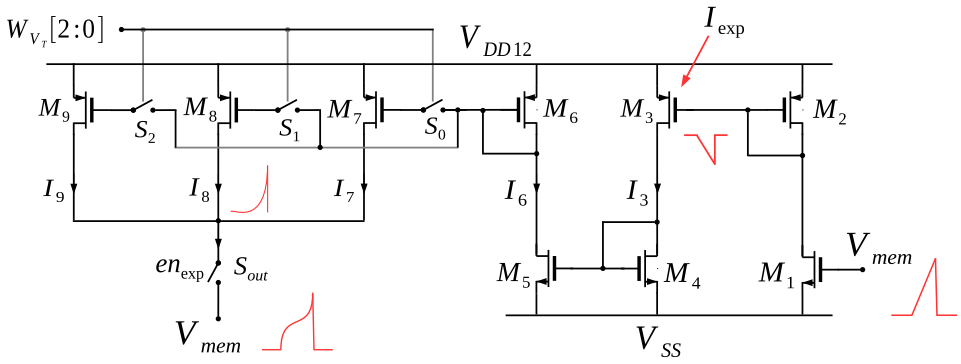


Figure 5.19: The schematic of the exponential cicuit.

The schematic implementing the exponential circuit is shown in Fig. 5.19. In the circuit the exponential current is generated by the device M_3. The neuron membrane potential V_{mem} is sensed by the device M_1, which together with the diode-connected pull-up device M_2, inverts its input. This inverted output is fed to the exponential device M_3, in a way that for maximum range of V_{mem}, the transistor M_3 stays within the subthreshold domain. This is maximized by choosing M_3 as a high-threshold (V_{th}) device. The exponential output current of M_3 is copied by a current mirror formed by $M_{4,5}$, before being subjected to three binary-scaled current mirrors formed by transistor M_6 and $M_{7,8,9}$. The mirroring ratio of $M_{4,5}$ is 1:1, so no current amplification is done. However, when the current is scaled using the scaling bits W_{V_T}, the design maintains a 1:1 ratio between transistors $M_{6,9}$, and decreases in a binary fashion between mirrors $M_{6,8}$, and $M_{6,7}$. The 3-bit tuning weight alters the AdEx exponential threshold V_{T}, and is therefore labeled as W_{V_T}. The output of all current mirrors are connected and the accumulated current is integrated directly on the neuron membrane. An output transmission-gate switch controlled by the digital configuration bit $\mathrm{en}_{\mathrm{exp}}$ toggles the exponential output on the membrane. Note that when exponential circuit is enabled, the source-drain drop (V_{SD}) of the output transistors of each current mirror is $V_{\mathrm{DD12}} - V_{\mathrm{mem}}$. This specifies the maximum limit up to which the exponential circuit may operate, since

at V_{mem} = 1.2 V, the effective source-drain drop and output current is zero. The current starts to decrease above 1.1 V. In order to ensure correct operation of the exponential circuit, it is suggested to limit the membrane potential to 1.05 V. Another rather soft limit comes from the threshold of the transistor M_3, since higher membrane beyond 1 V, cannot ensure a subthreshold operation of the device M_3. An exponential nature of the current, driven by subthreshold MOS dynamics will be compromised in that case.

The switches $S_{0\text{-}2}$ in the schematic of Fig. 5.19 are realized as PMOS-based pass transistors. When a mirror is disabled, the isolated gate nodes are pulled up towards the 1.2 V supply using another pass transistor (not shown in schematic) driven by configuration bits complementary to those which drive $S_{0\text{-}2}$. To demonstrate the functionality, we subject the circuit to a linearly increasing voltage input, which, given the subthreshold biasing, generates an output current that increases exponentially [13, 149]. This is shown in Fig. 5.20. The circuit output in Fig. 5.20

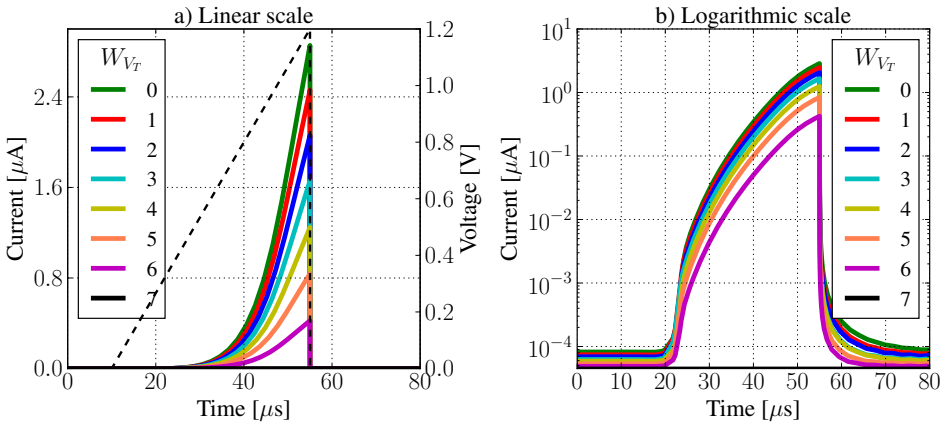


Figure 5.20: a) The exponential term is exposed to a slow voltage ramp (dashed signal). The resulting output current is swept for its digital input W_{V_T}. The output current is evaluated on a separate load. b) The exponential current output of a) is shown here on a logarithmic scale.

is evaluated on a separate load instead of the membrane. Therefore, in principle it can be taken higher than 1.05 V up to 1.2 V. The simulation setup sweeps 3-bit W_{V_T} which is a PMOS input, and is therefore active low. A weight equivalent to 000_2 is the highest setting as depicted in the figure. Note that up to approximately 0.45 V the circuit does not contribute with significant output current. Further, the exponential threshold varies with each setting of W_{V_T}. The logarithmic plot shows the near-exponential region above 10 nA output current that extends to more than 1 µA.

The circuit is prone to device mismatch and the output current indeed varies as a result. Fig. 5.21 shows the variation in peak output current as simulated using device Monte-Carlo models. The output current has a mean around 1.7 µA, with

a single-sided 1σ deviation of about 0.57 µA. The variation in exponential drain

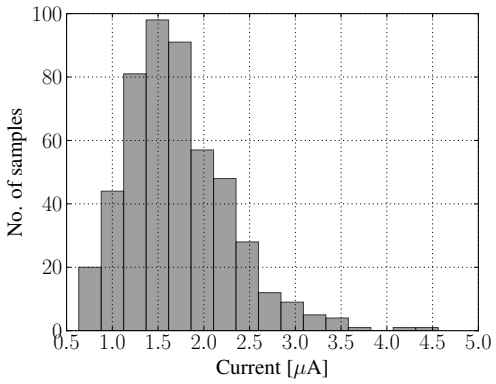


Figure 5.21: Variation in the peak output current of the exponential circuit with the maximum setting of W_{V_T}.

current is expected due to the subthreshold biasing. Further, it has a significant dependence on the mobility and the device threshold as well as on absolute temperature. The circuit is simulated for variations due to process corners, as shown

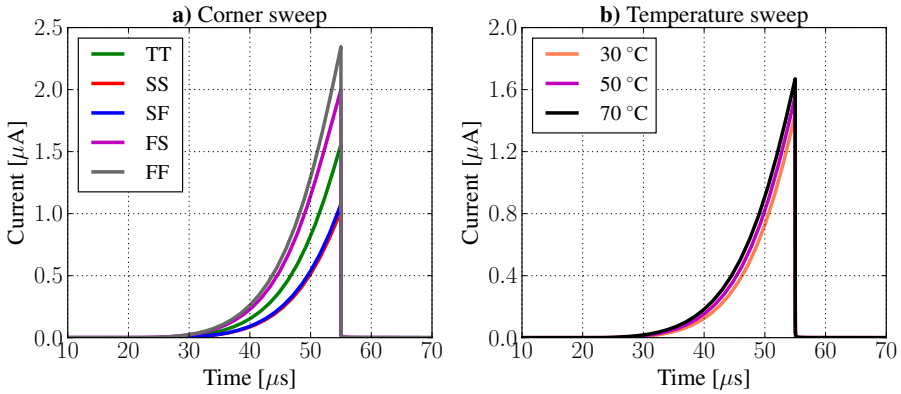


Figure 5.22: Corner and temperature sweeps of exponential circuit for the highest setting of W_{V_T}.

in Fig. 5.22a. As expected the fast corners generate higher currents as compared to typical or slow corners. The variations caused as a result of temperature change are depicted in Fig. 5.22b. Note how higher temperatures cause output currents to increase as seen by the 70 °C trace. The variations caused by the process and temperature variations as well as device mismatch can eventually be tuned for by the digital configuration bits (W_{V_T}). The current version of the circuit is tunable by a 3-bit configuration. This resolution needs to increase with controlled mismatch to compensate for the mentioned variations.

Taking a look at the emulated parameters of the AdEx exponential term, besides the exponential threshold V_{T}, the slope factor ΔT is emulated by the transistor weak inversion parameter nU_T (see Sec. 3.3). The slope threshold ΔT is therefore not meant to be a tunable parameter but is a fixed value. The value of slope threshold parameter ΔT is estimated during [91], by fitting the output current with an exponential function. The typical value is around 156 mV and does not vary with values of W_{V_T}. The plot in Fig. 5.23b shows ΔT as a function of the W_{V_T}.

The circuit has been simulated to estimate its current consumption. The static current consumption when the membrane is at resting potential of 600 mV is about 955 nA. To estimate the dynamic power a slowly rising input voltage is subjected to the input once again. Fig. 5.23a shows the current dissipation as a function of membrane potential. With W_{V_T} = 000_2 the current consumption rises from 955 nA for the membrane at 600 mV to its peak value of 8.7 µA for membrane at 1.05 V. The two curves correspond to minimum and maximum W_{V_T} code and the difference in their peak current values is the current flowing onto the membrane. This highlights that during the brief time interval when the membrane peaks at 1.05 V, the current consumption shoots up to 7.5 µA, independent of the W_{V_T} value. If the membrane threshold is set lower, indeed a reduction of dynamic power can be achieved. About 80% of this current is consumed in the input inverting stage via transistor M_1 due to its sizing and the rest in the current mirror formed by $M_{4,5}$.

Fig. 5.24 plots measured traces from a single on-chip neuron that was tuned to evoke tonic spiking with exponential term enabled. The leak and reset potential are set around 0.78 V. The exponential term pulls the membrane up exponentially until the neuron spikes around the specified threshold of 1.03 V. The parameter W_{V_T} is swept during successive runs and the plot overlays the tonic spikes on top of each other. Note that the order in which the exponential strength increases is as expected,

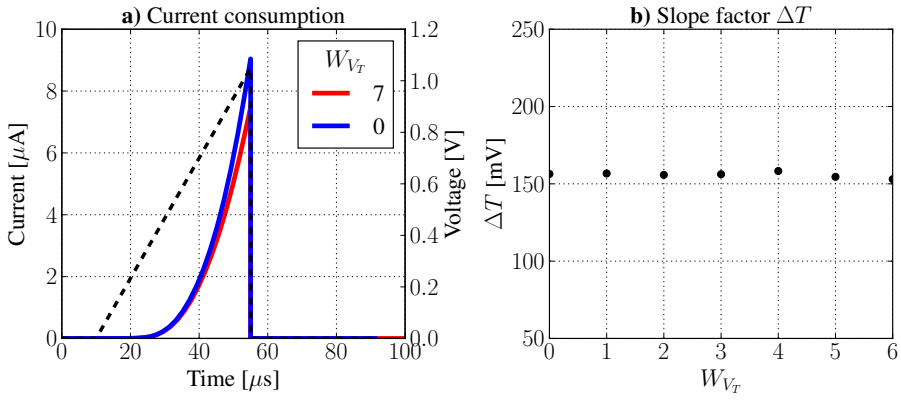


Figure 5.23: a) The dynamic current consumption of the circuit as a function of time-varying input voltage, as well as the minimum and maximum values of W_{V_T}. b) The slope threshold ΔT as a function of the digital parameter W_{V_T}.

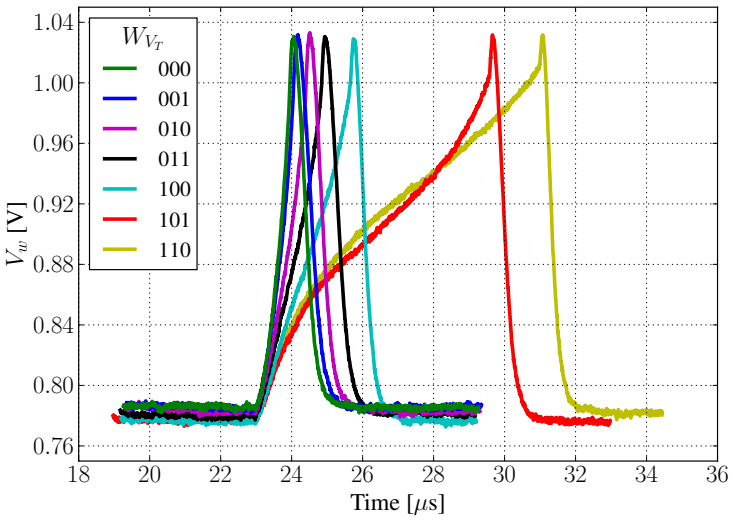


Figure 5.24: Measured results from the exponential term on a single neuron. The 3-bit digital W_{V_T} is swept from its minimum value 110_2 to the maximum 000_2.

starting from the weakest value 110_2 and growing with every increasing bit. One needs to expect considerable mismatch in the successive samples, as visible by the significant difference between 101_2 and 100_2. The traces have been acquired after applying the oscilloscope's built-in linear-phase FIR low-pass filter to cancel out the high frequency noise.

In comparison to the implementation of [99], the exponential term is realized without using an OTA/Opamp and therefore saves significant area. The exponential circuit consumes 47.7 µm^2 compared to approximately 358 µm^2 area of [99]. The power consumption however increases during dynamic activity and can be further improved by optimizing the input stage. The circuit has been designed with thin-oxide transistors and with 1.2 V supply only. The modular architecture and the digital control allows to completely turn the circuit off, and the exponential current switches-off by itself before reaching 1.2 V. The implementation in [99] does allow for the tunability of ΔT, which is not possible in the current implementation. The tunability using a digital parameter has been preferred and if more bits are allocated for W_{V_T}, calibration of corner variations could be possible. This may however be limited by the maximum achieveable Integral Nonlinearity (INL) in that case. The digital W_{V_T} saved an analog (Capmem) parameter as an additional benefit. In the thesis work of Mitja Kleider [150] it is shown that the exponential current range in [99] lasts only about 100 mV, to pull the membrane up only in the initial duration. This circuit, on the other hand, gives an exponential current for a few hundred millivolts of the membrane. Lastly, a circuit that utilizes weak inversion dynamics will be prone to variations nonetheless, although at a low implementation cost.

5.5 Analog Input/Output

The neurons within the array of the *DLS-2* chip directly drive the shared output line as highlighted in Fig. 4.35. This is replaced with a distributed, multiplexed read-out scheme in *DLS-3* chip as shown in Fig. 5.25. The top half of the figure shows the neuron array where it highlights the analog I/O circuits of each neuron. In addition to the neuron membrane V_{mem} and the voltage on the two synaptic lines $V_{\text{syn,exc/inh}}$, the voltage V_{w} on the adaptation capacitor can be read out via the multiplexer. The

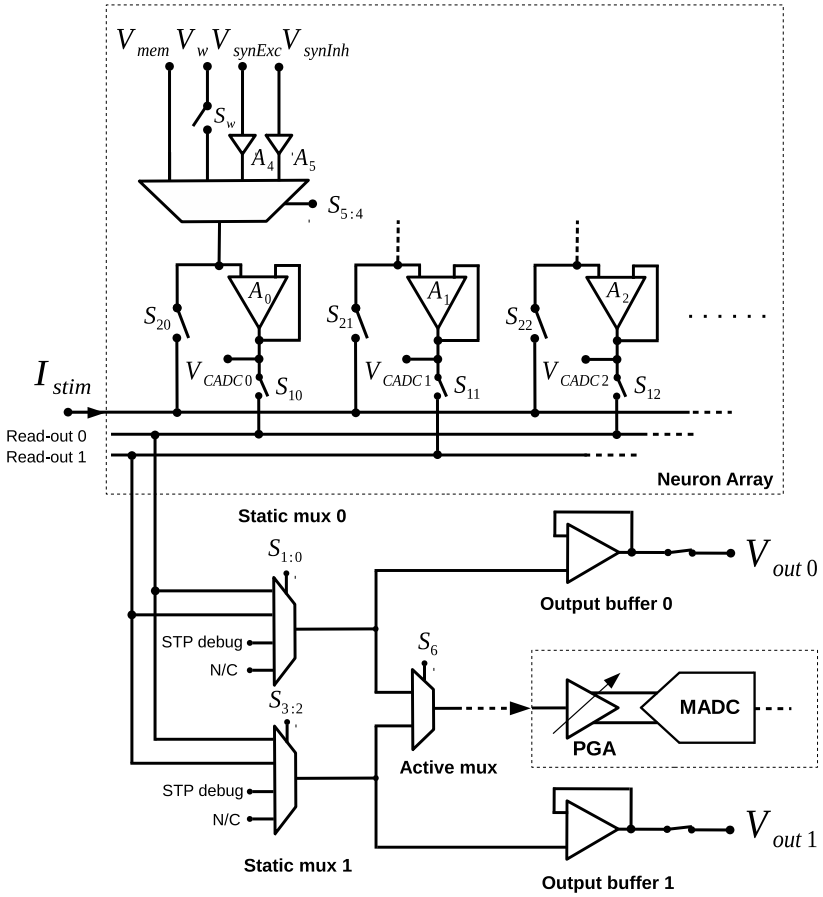


Figure 5.25: The distributed membrane read-out interface of the *DLS-3* chip.

size of the neuron buffer is reduced due to a smaller compensation capacitor, possible by virtue of the chosen compensation scheme and a decreased output load. This is described in the next subsection. The output of the reduced buffers (marked $A_{0,1,2}$ in Fig. 5.25), are connected to two separate read-out lines. Out of the 32 neurons in the array, the odd-numbered buffers are all connected to the first line, while the even ones to the second. The two read-out lines are again multiplexed through two transmission-gate multiplexers. These multiplexers are identical to those used inside the analog I/O of the neuron circuit. With this multiplexing, alongside the

neuron inputs, the debug input from the STP circuit can also be selected. Outputs of both multiplexers can be read-out simultaneously through two read-out buffers connected to the pads (through an always-on switch). These output-buffers are the ones designed for *DLS-2* neuron described in Sec. 4.8.1. The output of the two multiplexers are also routed to the MADC through a Programmable Gain Amplifier (PGA) and active multiplexers, described further in [94]. Note that the output of the reduced buffer is also directly routed to the CADC via the synapse matrix (see output V_{CADC} in Fig. 5.25). This provides us with digitized membrane voltage

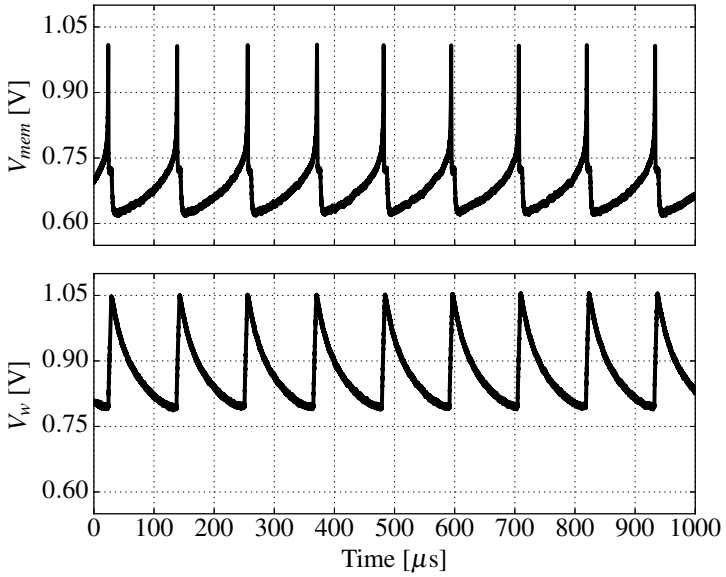


Figure 5.26: Superposition of measured results from a single on-chip neuron. The membrane potential V_{mem} (top) is being read-out simultaneously together with the adaptation voltage V_{w} (bottom) of the neighboring neuron circuit firing with the same dynamics. The spiking membrane shows the broad SAP possible from the AdEx models.

per column (neuron), directly readable via the PPU. Notice a switch S_{w} between the multiplexer input and V_{w}. This is a thick-oxide transmission-gate placed in the read-out path to prevent potential leakage, since V_{w} voltage rises above 1.2 V and 4×1 multiplexer is made up of 1.2 V core transistors. The I_{stim} input is common to all neurons and terminates at the pad, as in the *DLS-2* implementation.

Fig. 5.26 shows the membrane potential V_{mem} (top) read simultaneously from the chip together with the adaptation voltage V_{w} (bottom). The adaptation voltage in this case is that of an adjacent neuron, which fires with the same dynamics.

5.5.1 Read-Out Buffer

As described in Sec. 4.8.1, an advantage of using indirection compensation scheme is the reduction in the size of compensation capacitor without any extra circuits, such as the nulling resistor or feedback buffer. The neuron read-out amplifier designed in the *DLS-3* scaled the compensation capacitor of *DLS-2* read-out amplifier from 600 fF to 92 fF, without any further change in the design. This impacts primarily the frequency response of the circuit with a change in phase margin and the unity gain bandwidth. The DC-gain, common-mode and supply rejection at low frequencies remain unchanged. Table 5.2 summarizes the phase margin and unity gain bandwidth as the output capacitive load is decreased from C_L = 3 pF to 100 fF. Note that the current amplifier has a lower phase margin than that achieved in *DLS-2* read-out amplifier for C_L = 3 pF $\|$ 10 MΩ. The implemented circuit is in a buffer

C_L [1]	Phase Margin	f_u
3 pF	57.4°	115.2 MHz
1 pF	68.7°	217.4 MHz
100 fF	83.3°	714.6 MHz

[1] $\|$ 10 M Ω

Table 5.2: Phase Margin and unity gain bandwidth of the open loop OTA, without the output transmission gate.

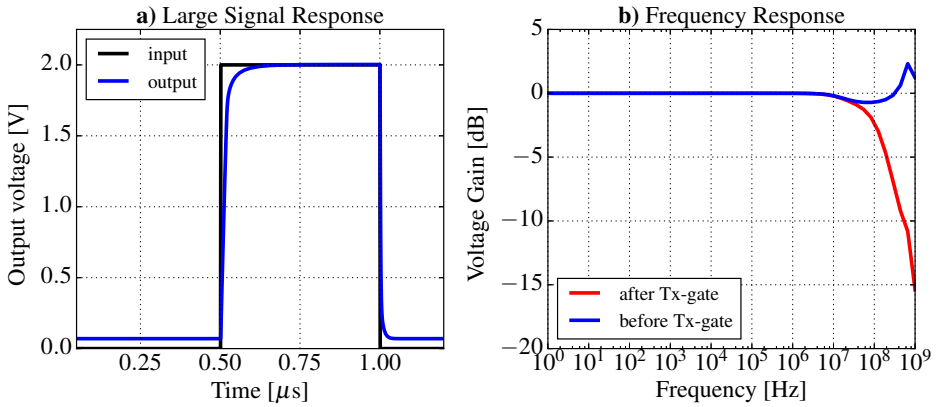


Figure 5.27: a) Transient large signal response of the *DLS-3* read-out buffer. b) Frequency response of the buffer before and after the transmission gate.

configuration, where it drives the input line connected to CADC (connected prior to transmission gate), and a connection to the read-out chain (through the transmission gate) as shown in Fig. 5.25. Assuming the parasitic load imposed by these two connections (the CADC line and the read-out line that terminates at static multiplexers) is 100 fF and 250 fF, the transient and frequency response of the close loop

amplifier is shown in Fig. 5.27. The transient settling response shown in Fig. 5.27a shows high stability of the amplifier for the chosen load. Note that the unfiltered bandwidth (Fig. 5.27b) shows some peaking at frequencies over 600 MHz. This is load-dependent and typically occurs in close loop amplifiers when conjugate poles are shifted towards the origin.

5.6 Membrane Capacitor

The membrane in the *DLS-3* neuron is a 6-bit adjustable MOS capacitor. The schematic of its implementation is shown in Fig. 5.28. The capacitors are realized

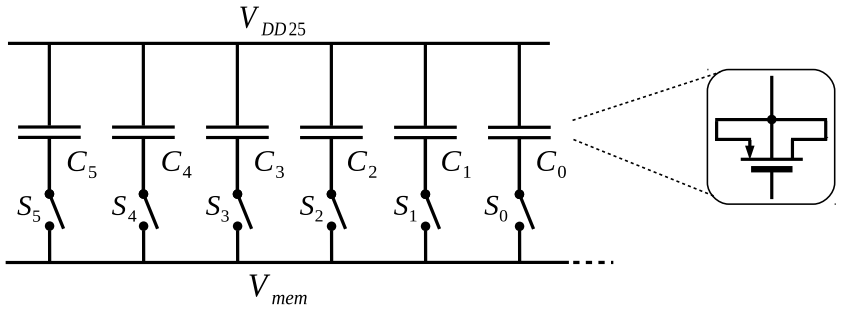


Figure 5.28: The schematic of the 6-bit selectable membrane capacitor.

using standard 2.5 V thick-oxide transistors as described previously in Sec. 4.7 and shown as inset in Fig. 5.28. The capacitance contributed by the MOScaps are summarized in Table 5.3. For layout reasons, capacitors C_5 and C_3 are chosen as $2\times$ and $0.5\times$ the size of C_4 (10.4 µm $\times$ 9.94 µm). Similarly, the sizes of C_2 and C_0 are twice and half of C_1 (3.55 µm $\times$ 3.55 µm). The total membrane capacitance sums up to 2.36 pF. The variation of MOScap when biased in inversion region has been previously shown in Fig. 4.34. Note that when adaptation is not used, an additional 2 pF of adaptation capacitor (MIMCAP) can be added in parallel, which will increase the membrane to 4.36 pF.

C_5	C_4	C_3	C_2	C_1	C_0
1.18 pF	590 fF	295 fF	148 fF	74 fF	37 fF

Table 5.3: The capacitance of the individual MOScaps that realize the membrane capacitor in *DLS-3* neuron.

5.7 Fixed Bias Distribution

The circuits within the *DLS-3* AdEx neuron that do not require tunable current biases have been provided with a fixed bias. This is essential to utilize the Capmem

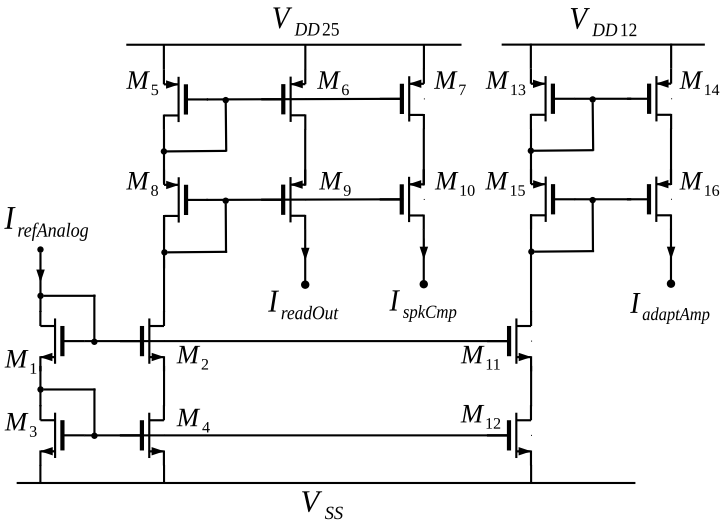


Figure 5.29: Schematic of the circuit that distributes current biases to readout amplifier, spike comparator and the adaptation buffer.

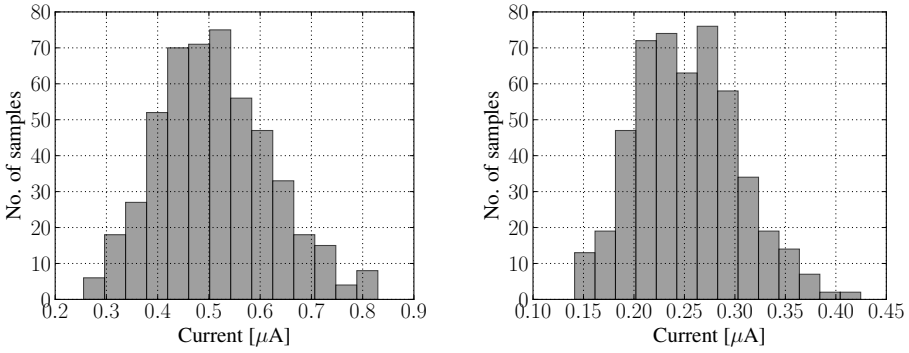
parameters efficiently. Since all such biases are current-based, no voltage fixed voltage biases are generated. The circuit to produce fixed biases is shown in Fig. 5.29. A cascode current mirror (devices $M_{1\text{-}4}$) takes its input reference current $I_{\text{refAnalog}}$ from a tunable Capmem based current cell. The mirror provides three different outputs for three different fixed bias circuits in the neuron. These include the read-out opamp, the spike comparator and the low voltage buffer used inside the adaptation term.

Current/Voltage	μ	1σ
I_{spikeCmp}	511 nA	110 nA
I_{adaptAmp}	252 nA	50 nA
I_{readOut}	1.01 µA	199 nA
V_{outSyn}	570.4 mV	27 mV

Table 5.4: The mean and 1σ variation due to the device mismatch in current biases produced. Since the output of the synaptic read-out source follower depends on I_{readOut}, the variation on its output voltage level is also listed.

The typical (fixed) value for $I_{\text{refAnalog}}$ is reduced to 250 nA, which is multiplied twice to provide 500 nA for spike comparator, and four times to provide 1 µA bias current for read-out amplifier. The membrane buffer inside the adaptation term is a 1.2 V circuit, therefore the cascode mirror ($M_{13\text{-}16}$) giving out I_{adaptAmp} = 250 nA is a 1.2 V circuit.

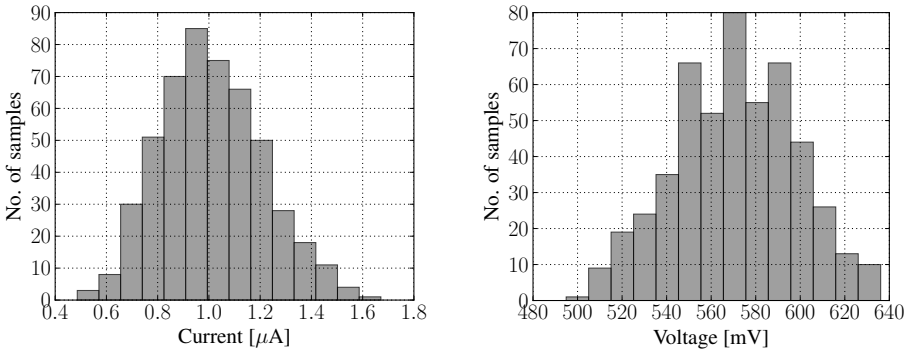
Note that the Capmem current cells have a PMOS based output stage – as a result the current needs to be mirrored twice, which introduces more variation in



(a) Bias current of spike comparator.

(b) Bias current of adaptation buffer.

Figure 5.30: Variation in generated input bias currents for spike comparator and adaptation buffer.



(a) Bias current of read-out amplifier.

(b) Output DC level of synaptic read-out follower.

Figure 5.31: Variation in generated bias current of readout amplifier as well as its result on output voltage of read-out source follower that reads synaptic input activity.

the output bias current. This variation due to device mismatch simulated using Monte Carlo models is shown in Fig. 5.30 for I_{spkCmp} and I_{adaptAmp}. Note that the synaptic activity on the synaptic lines is sensed by the source followers, as previously described in Sec. 4.8. As shown in the circuit of Fig. 4.36, the bias current for read-out amplifier I_{readOut} also provides a shared bias through a cascode current mirror (devices $M_{8,10}$ in Fig. 4.36) to the source follower. This determines the common-mode of the source follower bias load device M_2 (see Fig. 4.36). A variation in I_{readOut} will therefore result in a change of gate voltage, which in turn will alter the output DC-level V_{outSyn} of the source follower. The variation of the

read-out bias as well as its influence at the output level of source follower is shown in Fig. 5.31. Table. 5.4 summarizes the result of device mismatch histograms in tabular form as mean (μ) and 1σ single sided quantiles. Note that the output bias currents show about 20% 1σ variation from their mean value. The circuits in question are in any case robust and the variation does not alter their functionality.

5.8 SRAM Array and Level Shifters

The digital configuration bits used within the neuron circuit are stored in a 4×10 SRAM array. Four words, each containing 10 bits are integrated per neuron, for which four word-line signals are generated by the decoder implemented in the digital neuron control block. The arrangement of the SRAM array is shown in Fig. 5.32. The word-lines are shown entering from the left side labeled WL, while the bit-lines BL/BL^{B} from the top. Each SRAM bit-cell is a standard 6T cell equipped with a minimum-sized output buffer to improve the driving strength. Of the 40 bits, 30 are used in the main neuron circuit, while the last 10 bits are routed outside for use by the merged leak/reset/MC circuit. Five out of these 10 bits are dual polarity as indicated in the figure (DATA/DATA^{B}<39:35>).

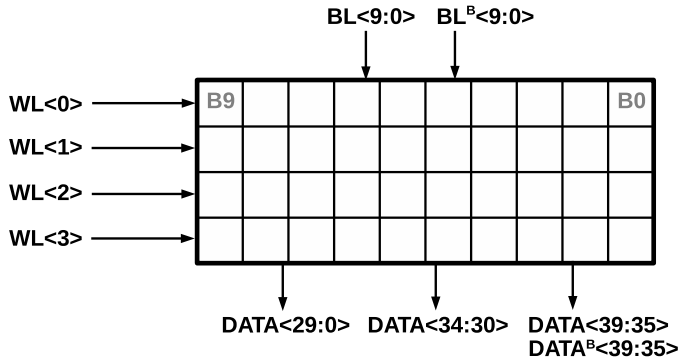


Figure 5.32: The 4×10 SRAM array implemented within each neuron circuit.

The configuration bits stored in the SRAM array have logic levels V_{OH} of 1.2 V. To drive an input of a 2.5 V transmission-gate or pass transistor, level conversion is achieved using the circuit shown in Fig. 5.33. The circuit operates as follows: when it encounters logic high (1.2 V) at the input of M_1, it turns on M_1 and turns off M_2. The input node to M_4 is pulled down as a result of M_4 turning on, pulling the output node V_{out} to V_{DD25}. As V_{out} rises to V_{DD25}, M_3 is turned off. The opposite happens when V_{input} is at logic low (0 V) turning M_2 on, eventually turning M_4 off. This causes V_{out} to output logic level low (V_{OL}). An inverter formed by $M_{5,6}$ produces the opposite polarity, since the transmission gates need both signals.

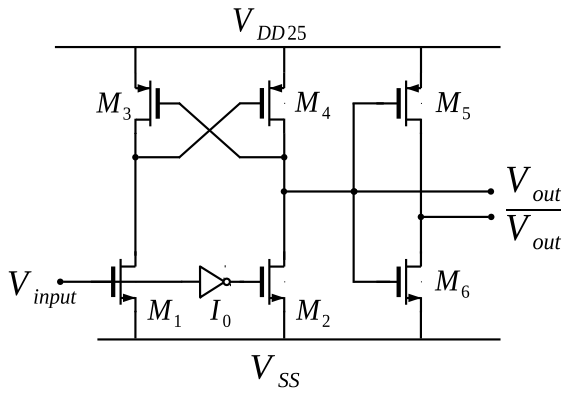


Figure 5.33: The level shifter used to convert 1.2 V signals to 2.5 V.

5.9 On Conductance-Based Synaptic Input

In order to extend the synaptic input to conductance-based inputs, an architecture is conceived that realizes both current-based and conductance-based inputs in parallel. The idea is to let the end-user select which type of synaptic input circuit he wants to use in the neuron circuit, one at a time. As shown in Fig. 5.34, in addition to the *DLS-2* synaptic input, another OTA with the conductance g_{m2} converts the voltage on synaptic input line into an equivalent current. The output current of this OTA controls a conductance element g_{syn}, such that output current integrated on the membrane is $g_{syn}\cdot(V_{mem}$ - $E_{syn})$ (see Eq. 1.4). A digital enable input *en* selects between the outputs of the two synaptic inputs, as well as switching the input biases I_{bias} and I_{biasSd} between the two variants. The latter saves current parameters by sharing them between the two circuits.

The circuit realizes a source-follower buffer to shift input voltages V_{rev} to the synaptic reversal potential E_{syn} – which can be excitatory reversal potential or the inhibitory reversal potential. In the excitatory case the source-follower is p-type, such that V_{rev} is shifted up, and E_{syn} is tunable between 1.2 V to 2 V. This provides the relevant range of E_{syn} in the excitatory case and reversal potentials higher than the range of Capmem voltage cells can be provided. In the inhibitory case, an n-type source follower shifts the voltage range lower. Due to the lack of design and verification time, the circuit is however not implemented in the *DLS-3* neuron. The maximum range of the conductance g_{syn} has been determined (by Paul Müller) to be from 10 nS to 9.5 µS (biological scale) in various modeling studies. Since the current-based input has already been silicon tested, the dual architecture targets a completely separate path (via OTA g_{m2}) to test the initial conductance-based design.

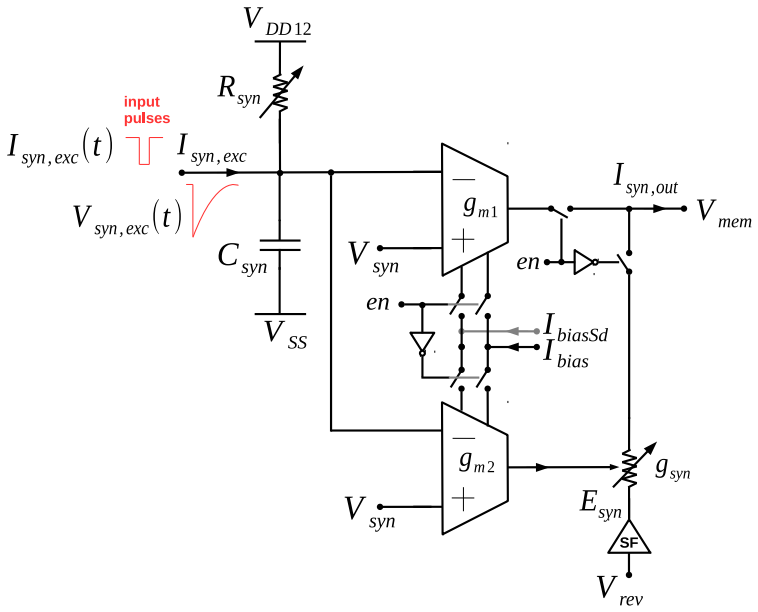


Figure 5.34: A dual synaptic input architecture that can switch between current-based and conductance-based synaptic inputs.

5.10 Spike Comparator and Membrane Offset

The *SpikeGen* circuit described in Fig. 4.27 is modified in the AdEx implementation. The circuit here acts as a voltage comparator that compares the membrane V_{mem} against the spiking threshold V_{thresh}, and the feedback loop formed via t_{delay} and T_0 has been removed. The fire output is therefore only a voltage level, that is sensed by the digital control. The digital control implements a programmable delay (refractory period) and resets the membrane via the conductance-based reset. Further, the synaptic input OTA does not implement output offset cancellation (Sec. 4.3.4). The offset is instead directly canceled at the membrane using a Capmem current source. The bias parameter $I_{\text{mem,off}}$ tunes the offset via a 2.5 V thick-oxide transmission-gate. Additionally, it may be used to stimulate the membrane.

5.11 Spike Patterns

Being a two-variable neuron model the AdEx circuit can reproduce a wide variety of spiking behaviors. The AdEx firing patterns have been introduced in Sec. 1.2.2. These are mostly evaluated for by the neuronal response of a current step stimulus, while tuning the hardware bifurcation parameters. During the design phase of the neuron, the circuit netlist has been simulated to evaluate for firing regimes by Laura Kriener [151], in conjunction with known model behavior from [42]. Due to the crosstalk during measurements (Fig. 5.17), all firing patterns are not evaluated on

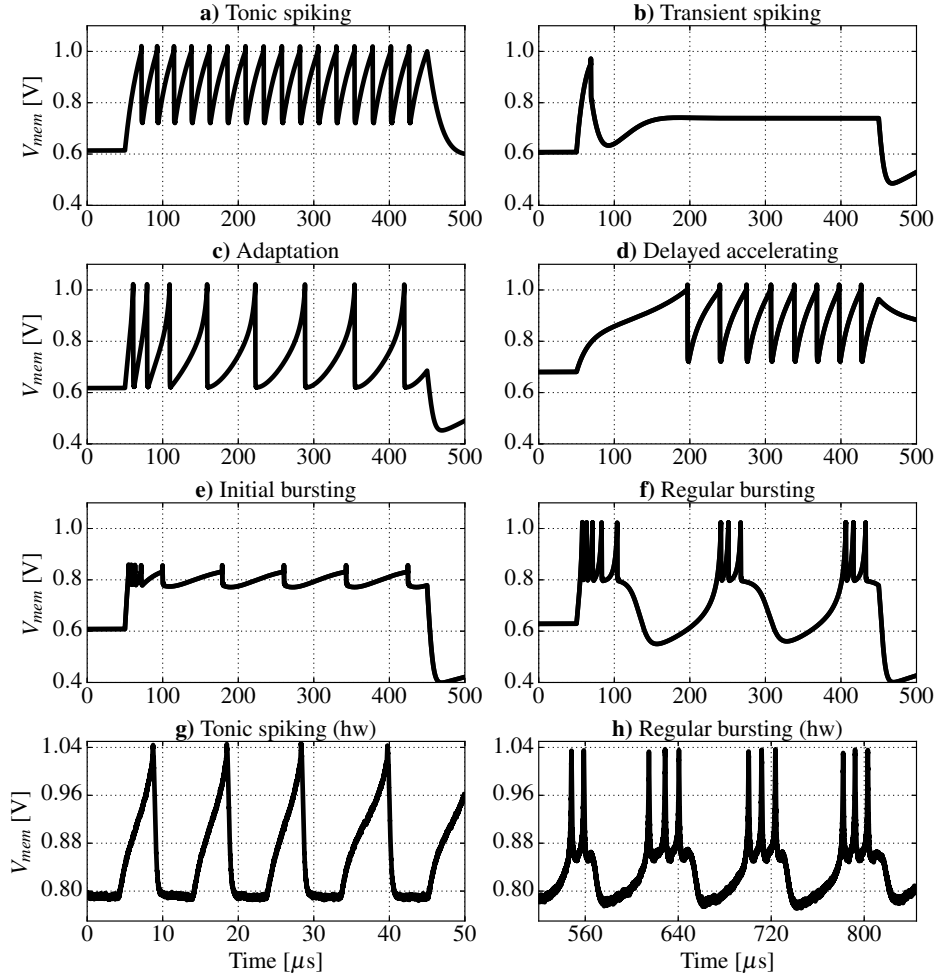


Figure 5.35: Firing patterns of the designed AdEx neuron. From a) to f) are those simulated on a circuit netlist [91,151], whereas g) and h) are measured results from the prototype chip.

the prototype chip since the firing regimes switch with every Capmem update cycle. However, tonic spiking and regular bursting are demonstrated. Fig. 5.35 shows the six simulated firing patterns from [91,151] together with the measured results from the chip. Fig. 5.35a-f demonstrate the tonic spiking, transient spiking, adaptation, delayed accelerating as well as initial and regular bursting from the simulations. Fig. 5.35g-h show tonic spiking and regular bursting as measured on the prototype chip. Fig. 5.26 shows the broad SAPs together with the adaptation variable V_{w}. Appendix C enlists the parameters used to reproduce the firing regimes in hardware shown in Fig. 5.35g as well as in Fig. 5.26. The parameters for Fig. 5.35a-f are listed in [151].

5.12 Bias Parameters

A description of current and voltage parameters used in the neuron circuit is given in Table 5.5. The SRAM bit-line bus is labeled 10-bit configSramIn together with

Parameter	Circuit	Type	Typical Range
V_{thresh}	SpikeCmp	local/voltage	0.6 V – 1.1 V
V_{leak}	Leak/Reset[1]	local/voltage	0.3 V – 1 V
V_{reset}	Leak/Reset[1]	global/voltage	0.2 – 1 V
$I_{biasLeak}$	Leak/Reset[1]	local/current	15 nA – 1 µA[2]
$I_{biasLeakSd}$	Leak/Reset[1]	local/current	15 nA – 1 µA
$I_{biasReset}$	Leak/Reset[1]	local/current	15 nA – 1 µA [2]
$I_{biasResetSd}$	Leak/Reset[1]	local/current	15 nA – 1 µA
$I_{refAnalog}$	Ampl./SpkCmp	local/current	0.25 µA
I_{memOff}	–	local/current	15 nA – 1 µA
V_{synExc}	Syn. Input (Exc.)	local/voltage	1.05 – 1.25 V
$I_{biasSynResExc}$	Syn. Input (Exc.)	local/current	15 nA – 1 µA
$I_{biasSynGmExc}$	Syn. Input (Exc.)	local/current	15 nA – 1 µA
$I_{globSynSdExc}$	Syn. Input (Exc.)	global/current	0.5 µA – 1 µA
V_{synInh}	Syn. Input (Inh.)	local/voltage	1.05 – 1.25 V
$I_{biasSynResInh}$	Syn. Input (Inh.)	local/current	15 nA – 1 µA
$I_{biasSynGmInh}$	Syn. Input (Inh.)	local/current	15 nA – 1 µA
$I_{globSynSdInh}$	Syn. Input (Inh.)	global/current	0.5 µA – 1 µA
I_{AdaptW}	Adaptation	local/current	15 nA – 1 µA
$V_{leakAdapt}$	Adaptation	local/voltage	0.2 V – 1.2 V
$I_{globAdapt}$	Adaptation	global/current	15 nA – 1 µA
$I_{biasAdaptSd}$	Adaptation	local/current	0.5 µA – 1 µA
$I_{biasAdaptRes}$	Adaptation	local/current	15 nA – 1 µA
configSramIn<9:0>	SRAM array	digital bits	–
configSramInB<9:0>	SRAM array	digital bits	–
writeSram<3:0>	SRAM array	digital bits	–

[1] Circuits are implemented externally to the AdEx circuit. See [92, 110] for details
[2] Typically set at 1 µA

Table 5.5: A summary of *DLS-3* neuron parameters and their typical operating range.

its inverted counterpart. The word-line is the 4-bit wide writeSram. The range of all synaptic input biases are similar to the LIF circuit of *DLS-2* neuron, except for

its source-degeneration bias $I_{\text{globSynSd}}$ (both excitatory/inhibitory) which is halved due to the intermediate scaling in the fixed bias mirror. The adaptation OTA bias current follows the range of the leak OTA (*DLS-2* circuit), while the adaptation resistor bias $I_{\text{biasAdaptRes}}$ and the adaptation current I_{AdaptW} take a full-scale current range. In addition to the two OTA leak biases, two separate reset OTA biases are also listed, and the *DLS-3* implementation switches between them, depending upon if its refractory interval or leak interval. The bias for analog components that do not require tuning is $I_{\text{refAnalog}}$, designed to be kept fixed at 250 nA. The membrane offset canceling (current injection) input is the I_{memOff}. A separate leak potential $V_{\text{leakAdapt}}$ parameter in the adaptation is provided, in case the user wants to decouple it from the main V_{leak} potential. Appendix B summarizes the corresponding digital configuration stored in the SRAM for each neuron circuit.

5.13 Power Consumption

With increasing number of OTAs and overall more circuits compared to the *DLS-2* neuron, it is pertinent to mention that power consumption is highly dependent on the bias settings within each neuron subcircuit. For a low power setting where all circuits would reliably work, the neuron draws approximately 18 µA of current from the 2.5 V supply and 9.5 µA from the 1.2 V supply. This sums up to approxi-

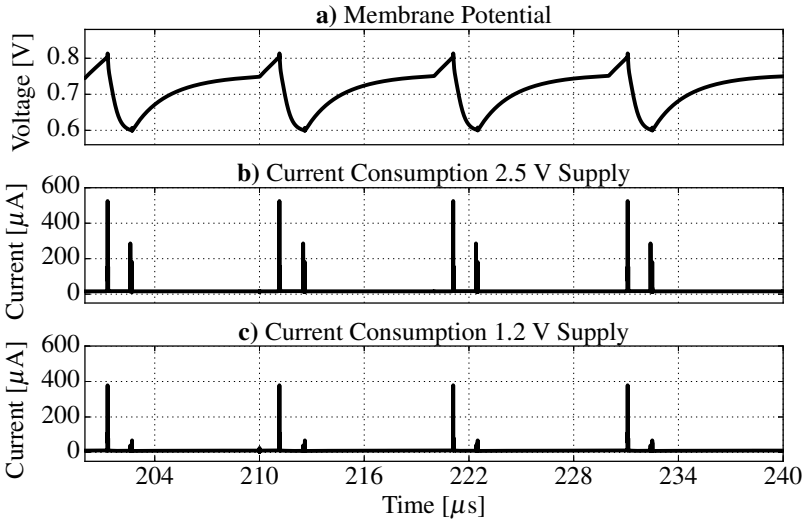


Figure 5.36: The current consumption of the implemented neuron circuit. a) The integrating neuron membrane, evoking spikes as a result of input events. b) The current consumed over time from the 2.5 V supply. c) The current consumed over time from the 1.2 V supply.

mately 46.4 µW of static power consumption. Fig. 5.36 shows the average current consumption over time from both supply lines, as the membrane toggles and spikes

due to incoming synaptic events (not shown). The membrane here is set for a weak conductance-based reset as seen in Fig. 5.36a. Spikes in current consumption are both visible during spike time as well as at the end of the refractory period. The latter is due to the OTA switching from high-conductance reset to low-conductance leak.

The spike-triggered level shifters in adaptation and merged leak/reset circuit cause a spike in current consumption from the 2.5 V supply. Further, the average current consumption from 2.5 V supply increases by 2.5 µA for the chosen biases during reset interval, since the (leak/reset) OTA switches to high-conductance mode. The fixed bias distribution and current mirrors also contribute to quiescent power consumption due to mirroring. The 1.2 V power consumption is higher than expected due to the high fixed bias in the adaptation buffer, which causes approximately 4–5 µA additional current consumption. Further consumption from 1.2 V comes from the membrane dependent current consumption of the exponential circuit, which, in the current setup is less pronounced due to the threshold being set at 0.8 V.

5.14 Physical Neuron Implementation

The total area occupied by the physical implementation of each neuron is 243.5 µm $\times$ 11.76 µm. The array of 32 neurons takes 243.5 µm $\times$ 403.6 µm. This space does not include the merged leak/reset block which occupies approx. 16.38 µm $\times$ 11.76 µm per neuron. Similarly, the digital neuron part is also not included which takes approximately 26.9 µm $\times$ 11.76 µm per neuron. The laid out view of the 1 $\times$ 32 array of neuron circuits is shown in Fig. 5.37.

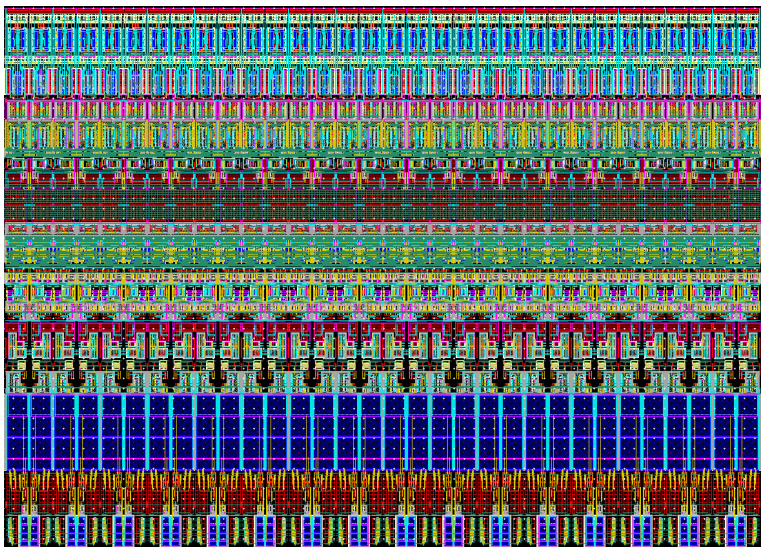


Figure 5.37: The layout view of the implemented array of 32 AdEx neuron circuits.

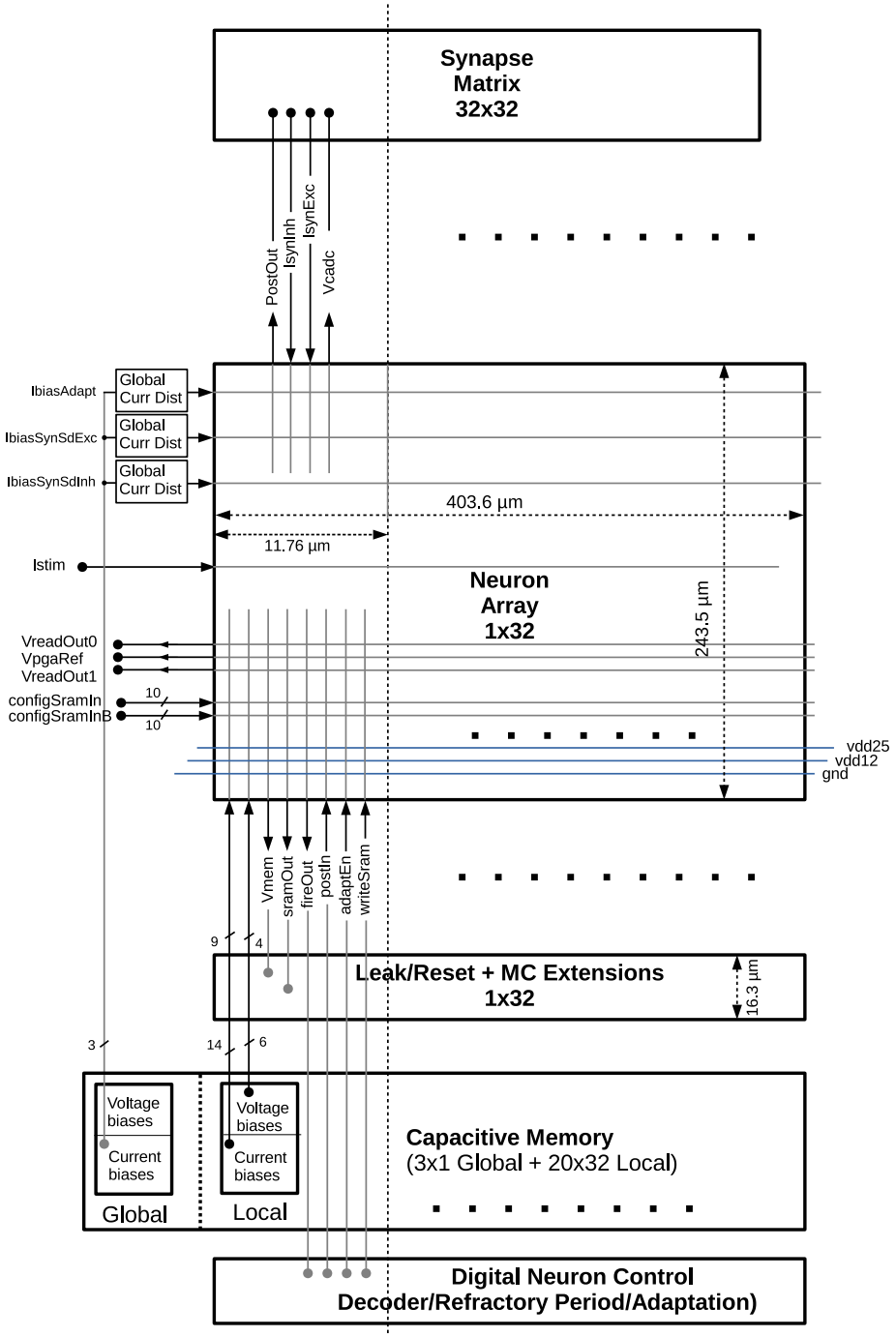


Figure 5.38: A schematic of the physical implementation of the AdEx neuron array along with its interface with the neighboring circuit blocks.

Shown in Fig. 5.38 is the implementation diagram of the entire array together with the interconnecting interface on the physical level with neighboring circuits. The array forms edge-connections at the top with the synaptic matrix and sends the read out (membrane) to the CADC via the synaptic array. At the bottom, it is edge-connected to the merged leak/reset circuits and multi-compartment extension block. It sends the analog membrane voltage as well as SRAM storage bits to the leak/reset array. Below the leak/reset is the capacitive memory which provides all local biases from the bottom and global biases from the left edge. The neuron array sends the fire_{out} pulse event to the digital neuron and receives the programmable $\text{fire}_{\text{adapt}}$ pulse back, used to trigger adaptation. Together with the post pulse that is routed to the synapse matrix, the SRAM address decoder input is also received from the digital neuron. Through the digital backend, neuron configuration bits are set via a 10-bit bus (labeled configSramIn). On the left edge, we also have global current stimulation pin I_{stim} as well as the two read-out voltages that are further routed to external buffers as shown in Fig. 5.25. A PGA reference is provided from the middle of the two read-outs (see [94] for details) and three global current biases are distributed inside the array using a current mirror from the left edge. The entire neuron array (including leak/reset circuit) is configured using 14 local current bias parameters, 6 local voltage parameters, and 3 global current parameters as highlighted in the schematic.

5.15 Discussion

This chapter presented the design and implementation of the AdEx neuron model, which is the targeted model of the BrainScaleS hardware. The circuit is designed on top of the modular LIF architecture implemented in the initial prototypes. It allows to easily integrate exponential and adaptation circuits without further changes. Along with the enhancement to the AdEx model, the circuit also realizes an array of SRAM bit-cells and stores the digital configuration within each neuron circuit. As in the LIF circuit, the AdEx circuit specifications are motivated from parameters suggested from a number of modeling studies.

The adaptation circuit implements the hardware model discussed in Sec. 3.1, similar to the approach adopted by the first generation design [99]. However, the design enhances the implementation by adding positive and negative subthreshold adaptation conductance ($\pm a$). It realizes a charge pump to have positive and negative spike-triggered adaptation ($\pm b$). The design in [99] only realized positive a and b. Further, at the component level [99] uses an OTA bias to tune the adaptation time constant, which, on one hand, has very small conductance range, and secondly suffers from limited linear range of the OTA. The current implementation uses a dual-sided resistor for a large linear range and provides a fairly wide conductance (resistive) range. Additionally, it requires only a single current bias parameter for tuning the time constant. This conductance implementation via a resistor is also a low-power solution since source-degenerated OTAs typically con-

sume more power. The designed membrane buffer, however, needs to reduce its current bias input to its specified 200 nA in the next revision.

Besides the OTA that models subthreshold conductance a (g_{a} for hardware), the rest of the adaptation circuit should be implemented by 1.2 V transistors only. As shown in Fig. 5.4, the circuits marked with dashed lines are already 1.2 V core transistor implementation. Therefore, to shift the rest of the circuit to 1.2 V, one can mirror the capacitive memory current I_{w} via a 1.2 V supply, before feeding it as positive current I_{w} in the circuit. This will remove the 2.5 V transmission gates as well as level-shifter(s) that cause dynamic leakage. Further, this will also limit the saturation level of voltage V_{w} to 1.2 V. The solution has in fact been considered during the design phase – however, it has not been implemented due to the additional mismatch when I_{w} is kept small (in the order of a few tens of nanoamperes). The additional mismatch could come from double mirroring in that case, since Capmem cells have a PMOS output only. However, given the implementation cost, mirroring the Capmem source twice and feeding the positive mirrored source via 1.2 V supply is still a viable solution (see Fig. 5.4). This will also reduce dynamic power consumption caused by leakage current of the level-shifter and reduce all core transistors to a maximum 1.2 V. The circuit could be realized with a charge-pump circuit where the individual current sources are well matched. Alternate charge pump architectures are reviewed in [152] and [153]. A dynamic current-matching charge-pump implementation suggested in [154, 155] is a good candidate. It reports a balanced positive and negative current source architecture that compensates for channel length modulation via negative feedback. The circuit is compact, feasible for 1.2 V implementation, and reduces the effects of charge-sharing.

The exponential circuit design is a 1.2 V implementation, motivated by a simple and area-efficient circuit that is digitally tunable and intended to reduce power consumption. To keep it simple and compact, it is designed as an OTA-less circuit in comparison to the implementation in [99]. Both [99] and the current implementations exploit the sub-threshold MOS characteristics to generate exponential current. The model parameter V_{T} is realized as 3-bit digitally tunable parameter, which saves an analog parameter. If extended to 6-bits this may also compensate for mismatch and process variations, although this is yet to be demonstrated. The circuit in [99] can also tune the slope factor, which is fixed in the current implementation. Further, [99] generates exponential current for only an initial 100 mV range [150]. The current implementation extends it to multiple hundreds of mV. The input stage of the exponential circuit (see Fig. 5.19) needs improvement to save dynamic power at higher membrane voltage. This can be achieved either by more efficient transistor sizing or a re-design of the input stage for better efficiency.

The circuit can produce the number of firing regimes as demonstrated by simulated and measured traces. However, in order to properly traverse the parameter-space for firing regimes, the sensitivity of the revised leak circuit to Capmem update cycle needs to be fixed. The leak circuit is not implemented by the author and is external to the current work done as part of this thesis. The conductance-based

Neuron model	AdEx I&F
No. of neurons	32
Voltage supply	2.5/1.2 V
Process	65 nm CMOS
Speed-up (acceleration) factor	$\times 1000$
Global parameters	3 current biases
Local (individual) parameters[1]	20 (14 currents, 6 voltages)
Configurability	40-bit SRAM per neuron
Membrane capacitor (max.)[2]	2.36 pF (6-bit configurable)
Area (single AdEx neuron)	2863.5 μm^2
Area (array of 32 AdEx neurons)	243.5 µm $\times$ 403.6 µm
Area (extended neuron)[3]	3372 μm^2
Power Consumption[4]	46.5 µW
τ_w (min./max. range) [5]	[25.8 $\pm$ 8.2] µs – [418.5$\pm$ 210.8] µs

[1] 13 biases are used in the Ad-Ex neuron circuit (9 currents, 4 voltages)
[2] can be increased to 4.36 pF if adaptation is switched-off
[3] this is the cumulative area of the larger neuron with AdEx implementation (243.5 µm $\times$ 11.76 µm), multi-compartment and conductance-based reset (16.3 µm $\times$ 11.76 µm) and the digital control (26.9 µm $\times$ 11.76 µm)
[4] approx. for the extended neuron circuit; varies with parameter settings
[5] with 1σ variation based on simulation results

Table 5.6: A summary of achieved specifications of the implemented AdEx neuron array.

reset is not modeled in the AdEx model.

Several other changes have been summarized for marching on towards a final chip. This includes a small neuron read-out buffer, a 6-bit tunable membrane, fixed bias distribution, etc. A preliminary architecture that looks very plausible for conductance-based synapses is finalized and needs an implementation in the next revision. Table 5.6 summarizes the achieved specifications of the AdEx neuron circuit.

Chapter 6

Conclusion and Outlook

This thesis reported the circuit design and implementation of spiking neuron models for the development of second-generation BrainScaleS mixed-signal neuromorphic hardware. Given the BrainScaleS approach, biophysically-inspired integrate-and-fire neuron models are selected for implementation to reproduce the temporal neural dynamics. The second-generation hardware implementation is in 65 nm CMOS, a step ahead from the previous HICANN chip implementation in 180 nm CMOS. Along with the technology difference, the new hardware features a reduced acceleration factor of 10^3 times compared to biology, as opposed to the previous factor of 10^4–10^5. This shifts the available range of neuron time constants and new circuit architectures have been explored and implemented for the 10^3 acceleration factor.

The first model implemented is the leaky integrate-and-fire model. The circuit explores a novel synaptic input concept that simplifies the integrator design compared to the implementation in [98, 99]. The integrator uses parasitic capacitance of the synaptic input line and implements a novel compact tunable resistor architecture. The leak term in the circuit is realized from the legacy approach of a transconductance amplifier, but adds additional features such as high-conductance mode and output offset compensation (synaptic OTA only). A spike pulse generator is designed to evoke a digital output event and a read-out buffer scheme is integrated to read-out both the membrane and synaptic input activity. The buffer is based on a two-stage opamp with indirect compensation scheme and directly drives the output pad. All circuits are interconnected using transmission-gate switches to devise a modular architecture.

The AdEx model is a mixed-signal implementation which integrates adaptation and exponential circuits to the modular LIF architecture. The adaptation circuit realizes accelerating and decelerating spike-triggered adaptation using a charge-pump based circuit. It integrates a dual-sided compact floating tunable resistor to tune the large range of adaptation time constants. An OTA within the adaptation circuit models the subthreshold adaptation conductance. The exponential circuit exploits the MOS subthreshold region to generate exponential current, dependent

on the membrane voltage. The circuit is compact with a digitally tunable exponential threshold and switches itself off before reaching the 1.2 V supply rail. The neuron circuit further implements a 6-bit tunable membrane and a 40-bit SRAM array per neuron. The AdEx implementation explored in this work is further integrated with a digitally tunable refractory period as well as a conductance-based reset.

For the design of the larger system the neuron circuits implement the following essential features of the system:

Tunability and Circuit Specifications

The neuron circuit specifications have been motivated from the parameter range used by known network models developed by computational neuroscience groups. Being a general-purpose computational element for a variety of network models the neuron circuit is a highly tunable element. This tunability is made possible by the availability of on-chip biases and tunable circuit architectures. For digital configuration an on-chip SRAM array has been integrated per neuron. For the identified parameter range Table 6.1 summarizes the achieved range of the two neuron circuits.

	Hardware [1]	Models	Units
τ_{mem}	0.35 to 16.4	7 to 50	ms
τ_{syn}	1.24 to 20.5	1 to 100	ms
τ_{ref}	1.11 to 137.5 [2]	0 to 10	ms
a	$\pm$1.9 to $\pm$650	–11 to 40	nS
b [3]	$\pm 0.12a$ to $\pm 51.8a$	0 to 250	mV/nA
τ_{w}	25.8$\pm$8.2 to 418$\pm$211	16 to 600	ms

[1] reported with a 1σ single-side quantile
[2] extended to a range of 0.1 ms – 1000 s in *DLS-3* chip
[3] not directly comparable since hardware is voltage (ΔV) and dependent on a, while model parameter b is current [91]

Table 6.1: Achieved tunable specifications in the neuron circuits.

While the range of refractory period, subthreshold adaptation and spike-triggered adaptation are achieved, the current implementation falls short in the case of membrane and synaptic time constants.

Scalability and Silicon Verification

The chip prototypes implement an array of neurons in a highly integrated ANC architecture aligned with the digital logic core. All three prototypes whose results

have been reported are miniature versions of the larger HICANN-DLS chip, where tight integration with capacitive memory cells, synapses and digital control has been verified. At the circuit level, individual circuits target scaled up specifications, e.g., the synaptic input design considers the line parasitic of the larger chip and the read-out amplifier drives a variable on-chip capacitive load (Sec. 4.3.4, Sec. 5.5.1). Further, through in depth testing of the prototypes, circuit statistics are demonstrated (Chapter 4 and [79]).

Testability

Although the entire neuron array is connected to the external pads with two pins, the design ensures the testability of most subcircuits individually. The implementation of multiplexed debug read-out scheme (Sec. 4.8, Sec. 5.5) as well as the modular architecture that disables other circuits, facilitates this evaluation. The read-out of input synaptic events as well as the adaptation voltage V_{w} let us characterize the neuron model dynamics (Sec. 4.3.4, Sec. 5.5). The extended read-out scheme in *DLS-3* prototype reads out two parallel voltages. The neuron circuit further integrates a membrane bypass-mode to test and debug the digital event routing.

Calibration

The subcircuits within the neuron ensure that they can be calibrated against non-ideal effects arising as a result of device mismatch or process corners. As a result of calibration (Sec. 4.3.7, Sec. 4.6.2) the time constants are set reliably and the residual offset currents are compensated for.

Power Consumption

The design reduces power consumption using 1.2 V supply, wherever possible. The circuits for the synaptic resistor, adaptation resistor, adaptation buffer, exponential circuit are a 1.2 V implementation. Most circuits can be switched off digitally. Power consumption is highly dependent on parameter settings. Depending upon the time constants and utilized circuits, the consumption varies. For an average setting and single synaptic input enabled, power consumption of the LIF neuron circuit is approximately 10 µW. It is emphasized that digital circuits, such as inverters, must not be subjected at the output of any slow moving voltage signal if dynamic power is to be conserved (Sec. 4.11). The static power consumption of the extended AdEx neuron (including conductance-based reset and the digital control) is approximately 46.5 µW. This can be reduced further by saving quiescent current consumption in merged leak/reset OTA, extra consumption in adaptation, as well as in the bias distributor (Sec. 5.13).

Silicon Area

With reduced 1.2 V supply the core transistor designs help reduce the occupied silicon area. The 1.2 V implementations are therefore compact, for example, the exponential circuit occupies only 47.7 μm^2 area, which is seven times less than the implementation in [99]. Area is further conserved by relying on reduced bias currents (Sec. 4.6.2) and small conductances, rather than increased capacitors for long time constants (Sec. 5.3.3). Further, MOS capacitors are used to restrict the area (Sec. 4.7, Sec. 5.6). Additionally, in the adaptation circuit, resource sharing is enabled, by merging the metal capacitor with membrane capacitor, when adaptation is not used. The LIF circuit utilizes 2352.0 μm^2, while the AdEx implementation uses 2863.5 μm^2. The extended AdEx circuit that includes conductance-based reset, multi-compartment extensions as well as the digital neuron control occupies 3372 μm^2 per neuron.

Future design improvements

Given the designs of the two neuron circuits, achieved specifications and the reported results, a number of design improvements are recommended:

- From Table 6.1 it is evident that the leak and synaptic time constants fall short of achieving their desired range. Their maximum range needs to be increased. This is a realistic target, since the resistor implementation in adaptation circuit has demonstrated the availability of a large tunable range for adaptation time constant. The variation in the implemented architecture is relatively low in the specified range of synaptic and leak time constants.

- The conductance-based parallel synaptic input architecture outlined in Sec. 5.9 needs to be implemented, to provide more biologically realistic synaptic dynamics [156]. The architecture facilitates the end-user to select between the current-based or conductance-based synaptic inputs.

- The exponential circuit should increase its W_{V_T} resolution to compensate for corner variations.

- The adaptation circuit should realize the charge pump circuit with core transistors only to limit V_{w} to 1.2 V, according to recommendations in Sec. 5.3.3.

- The fixed-bias distributor should correct the current bias input provided to the adaptation buffer.

- An overall static and dynamic power consumption is to be reviewed within the extended neuron with multi-compartment extension and digital control. In the AdEx implementation, the input stage of the exponential circuit requires improved sizing to conserve dynamic power consumption (Sec. 5.4). Both dynamic and static power is reduced once the adaptation charge pump is reduced to 1.2 V.

Comparison to other architectures

Several groups produce neuromorphic hardware with neuron model implementations ranging from software-definition [157, 158] to subthreshold analog [159–161] to fully-digital implementations [159, 162]. The SpiNNaker system [157, 158] based on arrays of many-core ARM-based microprocessor systems has a software-defined neuron model. A popular approach to design silicon neurons is the continuous-time implementation that exploits the subthreshold MOS dynamics, as proposed in [12, 13]. The systems adhering to this subthreshold analog integration [159–161] implement biophysically-inspired neuron models with real-time spiking dynamics. All-digital phenomenological model implementations typically capture the input/output neuron behavior and the digital nature benefits from the deep-submicron CMOS implementations [97, 162, 163]. Out of the above-mentioned implementations, state-of-the-art large-scale architectures are, for example, the IBM's TrueNorth system [97, 164, 165] in the digital domain, and Stanford University's Neurogrid architecture [159] for the analog subthreshold approach.

	TrueNorth	Neurogrid	HICANN-4	*DLS-2*	*DLS-3*
CMOS tech. [nm]	28	180	180	65	65
Architecture	digital	analog subthresh.	analog	analog	mixed-signal
Model	augmented LIF	quadratic I&F	AdEx	LIF	AdEx
Area [μm^2]	2900[a]	1800	3124[b]	2352	3372[c]
Area Est.[d] [μm^2]	–	–	–	1404	2847
Power [e] [μW]	N/A	N/A	100	10–15	40–46

[a] multiplexed 256 times per time step
[b] without the membrane capacitor (approx. 2800 μm^2), which overlaps (shares) the area occupied by floating gate array
[c] cumulative area including multi-compartment extensions, conductance-based reset and digital control. The AdEx design implemented in this work occupies 2863 μm^2
[d] estimated area if *DLS* neurons implement the membrane capacitor as MIM-CAPs overlapping another chip block and synaptic input capacitor is realized entirely from line parasitics
[e] dependent on parameter settings

Table 6.2: An overview of neuron model specifications in large-scale neuromorphic architectures.

The TrueNorth system integrates an augmented LIF implementation with an

objective of building a synthetic computational element for cognitive applications [97]. The classic LIF behavior is extended with various leak, threshold and reset modes with deterministic and stochastic operation. The neuron performs a number of synthetic, logical and fixed-point arithmetic operations. It operates in real-time with a possibility of 21× speed-up to real-time [164], and occupies an area of 2900 μm^2. Following a digital-approach, the neuron can be time-multiplexed for reuse. Compared to the HICANN/HICANN-DLS neuron, it not only produces the spiking behavior, but also synthetic, logical, signal processing, probabilistic and arithmetic functions. However, it requires a composition of 2 or 3 neurons to produce diverse behavior such as spike-frequency adaptation, tonic bursting, etc., which can be reproduced with a single element in bio-physical implementations (e.g., see Sec. 5.11). The stochasticity added to the system, though not directly comparable, but to a certain extent is inherent in integrated analog models of spiking neurons, pertaining to variations due to device mismatch. The TrueNorth system runs with a 1 ms time-step, which may limit the realization of shorter time constants. Further, the temporal dynamics such as the synaptic and refractory time constants are not tunable implementations. Compared to BrainScaleS neurons, the maximum fan-in and fan-out is limited to 256 per synaptic core (each containing 256 neurons and 64K synapses). On the other hand, compared to the current work, axonal delays are implemented.

Neurogrid is another large-scale system, which is more biologically-plausible hardware compared to TrueNorth. Compared to the BrainScaleS system which supports arbitrary connectivity, it is optimized for sparse long-range connections and dense local connectivity, such as for modeling neocortex [159]. The neuron implementation is a dimensionless quadratic integrate-and-fire model, where somatic and dendritic compartments are modeled. Dimensionless models reduce free parameters, whereas log domain analog subthreshold circuits realize real-time temporal dynamics. Synapses are shared among neurons and multiple spikes may superimpose in time on a single synapse circuit [166]. The larger architecture is designed using 16 neurocore chips, each having 256 × 256 neurons where all neurons share the parameter set. The HICANN and HICANN-DLS chips are highly tunable and configurable, since they provide individual parameters for every single on-chip neuron (Sec. 5.12 and Sec. 4.13). Neurogrid is implemented in a 180 nm CMOS process where the neuron occupies 1800 μm^2. Due to the subthreshold approach, the power consumption (as well as occupied area per neuron) is typically lower compared to the work presented here. Subthreshold current integration further allows for longer time constants. The circuits are calibrated [167] for variations arising as a result of transistor mismatch, similar to the approach in this work.

Table 6.2 summarizes the neuron specifications in TrueNorth and Neurogrid, in comparison to HICANN and *DLS* prototypes. Note that Neurogrid achieves less area in comparison as discussed above. This is due to less number of transistors, but typically subthreshold currents let scale the overall area occupied by the capacitors. In comparison, although the single instance of TrueNorth neuron is large, the

time-multiplexing scales the effective area per neuron to 14.3 μm^2 [164]. The area occupied by the first-generation HICANN-4 neuron is relatively large, and it utilizes the higher metal area (MIMCAP implementation) and places the large membrane capacitor over the floating gate array. Despite that, compared to 28 nm, an older CMOS process occupies more space. The *DLS* neurons are relatively smaller and use line parasitic and MOScaps for area efficiency. If chip implementation allows to overlap area as MIMCAP like HICANN-4, it may shrink its size further (Area Est. in Table 6.2). Compared to the HICANN neuron [98], the *DLS* neurons are power efficient despite their 2.5 V usage, but power consumption varies vastly with parameter settings.

At the system level, all three systems discussed here describe non-von Neumann architectures, where stronger biological inspiration is evident in Neurogrid and BrainScaleS design. The spike transmission is asynchronous digital in all three systems. The BrainScaleS system described here supports multiple levels of dedicated synaptic plasticity as opposed to Neurogrid and TrueNorth in general. A reasonable figure of merit to compare the performance of neuromorphic hardware is the energy per synaptic event. However, it is highly dependent on the realized network, the number of synapses and the finalized hardware. In general, studies suggest [163, 168, 169] that digital implementations are typically less energy-efficient, compared to continuous-time analog approaches. Since the implementations present very diverse benchmarks, a fair comparison is difficult. In future, other technologies, e.g., the memristive crossbars [170,171] will get mature enough, or novel emerging technologies will implement more scalable and energy efficient large-scale neuromorphic solutions.

Taking inspiration from the structure and dynamics of the nervous system, analog neuromorphic systems integrate physical neurons with spiking dynamics on a biologically-inspired substrate. This work presented the design and implementation of this physical 65 nm CMOS neuron, which endows the second-generation BrainScaleS hardware with highly tunable biologically plausible firing dynamics. The 65 nm system architecture will emerge as a powerful computational platform due to the integration of analog and digital cores (PPU), a distinct feature of BrainScaleS-2 hardware [86].

Applying nature's principles has its benefits – already in the last decade, deep learning [172, 173] enabled a technological stride in machine learning and pattern recognition, made possible by adopting biologically-inspired computing principles. The role of future neuromorphic systems towards the solution of real-world applications is yet to be determined. With the recent realization of advanced neuromorphic systems by industrial players [164,174,175], the prospects are high that neuromophic systems will play a central role in next-generation computing architectures.

List of Figures

List of Tables

List of Abbreviations

AdEx Adaptive Exponential Integrate-and-Fire 10

ADC Analog-to-Digital Converter 23

ANC Analog Network Core 21

Capmem Capacitive memory 26

CADC Correlation ADC 50

DAC Digital-to-Analog Converter 47

DLL Delay-Locked Loop 33

GABA Gamma-Aminobutyric Acid 14

INL Integral Nonlinearity 113

ISA Instruction Set Architecture 24

L1 Layer 1 21

L2 Layer 2 21

LDO low drop-out 43

Appendices

Appendix A

Digital configuration in the DLS-2 LIF Neuron

A description of the digital configuration bits used in the *DLS-2* neuron circuit is listed in Table 3.

Enable bit	Description
`ctrlTgNeuron<0>`	connects excitatory synaptic input to membrane
`ctrlTgNeuron<1>`	connects inhibitory synaptic input to membrane
`ctrlTgNeuron<2>`	enables the high conductance state in leak
`ctrlTgNeuron<3>`	enables the I_{stim} pin for debug inputs
`ctrlTgNeuron<4>`	enables the read-out debug amplifier
`ctrlTgNeuron<5>`	not connected
`ctrlTgNeuron<6>`	connects leak to the membrane
`ctrlTgNeuron<7>`	enables the excitatory bypass link
`ctrlTgNeuron<8>`	enables the inhibitory bypass link
`ctrlTgNeuron<9>`	connects SpikeGen as digital fire-out for backend
`enAnaOutMux<10:11>`	two-bit select line for debug multiplexer
`enAnaOutMux<12:13>`	two-bit select line for size of membrane capacitor
`enAnaOutMux<14:17>`	not connected
`pullDn`	global pull down signal resetting SpikeGen comparator.

Table 3: A summary of digital bus signals that configure the LIF neuron circuit.

Appendix B

Digital configuration bits in the DLS-3 AdEx Neuron

A description of the configuration bits stored in the SRAM array in the *DLS-3* neuron circuit is listed in Table 4.

Parameter	Circuit	Description
enSpkCmpB	SpikeCmp	comparator output reset
enSynBypExc	Bypass	enables excitatory bypass
enSynBypInh	Bypass	enables inhibitory bypass
enFireOut	–	enables *fire* output
enMemCap<5:0>	Memcap	membrane select lines
enSynIexc	Syn. Input (Exc.)	enables synaptic input (exc.)
enSynIinh	Syn. Input (Inh.)	enable synaptic input (inh.)
enAnaOutMux<1:0>	Analog I/O	select lines for debug read-out
enAnaIn	Analog I/O	enable for I_{stim}
enAnaOut	Analog I/O	enable for V_{readOut}
enOutAmp	Analog I/O	enable for read-out amplifier
enAdapt<1:0>	Adaptation	enables for adaptation
enCapMerge	Adaptation	adds C_{w} as membrane
enVa	Adaptation	enables positive conductance g_{a}
enVw	Adaptation	enables decelerating adaptation
enReadVw	Adaptation	enable for V_{w} read-out
enExpWeight<2:0>	Exponential	digital V_{T} bits (labeled W_{V_T})
enExp	Exponential	enable for exponential output

Table 4: A summary and description of digital configuration bits of the *DLS-3* neuron.

Appendix C

Hardware parameters for various firing patterns in AdEx Neuron

The hardware parameter settings for the broad spikes shown in Fig. 5.26 is listed in Table 5.

Parameter [1]	Value	Parameter	Value
$I_{refAnalog}$	250	enSpkCmpB	0
V_{leak}	360	enSynBypExc	1
$V_{leakAdapt}$	360	enSynBypInh	1
V_{reset}	400	enMemCap<5:0>	0x3F
V_{thresh}	600	enSynIexc	0
$I_{biasLeak}$	1022	enSynIinh	0
$I_{biasLeakSd}$	1022	enAnaOutMux<1:0>	0
I_{memOff}	400	enAnaIn	0
$I_{biasAdaptRes}$	150	enAnaOut	1
$I_{biasAdaptSd}$	280	enOutAmp	1
I_{AdaptW}	100	enAdapt<1:0>	0x03
V_{synExc}	780	enCapMerge	0
V_{synInh}	780	enVa	1
$I_{globAdapt}$	1022	enVw	1
$I_{globSynSdInh}$	0	enReadVw	0
$I_{globSynSdExc}$	0	enExpWeight<2:0>	0x0
enExp	1	enFireOut	1
enLeak	1	adaptConfig	0xE
enHiConReset	1	resetHoldOff	0xE
enHiConLeak	0	refrCounter	0xE0

[1] The voltage and current biases are given in terms of equivalent digital code

Table 5: The hardware parameters used to evoke broad SAP in the AdEx neuron.

In order to read-out two simultaneous inputs, the configuration of Table 5 for the second neuron (whose V_w is to be read-out) changes the bits enAnaOutMux<1:0>, enReadVw to 0x11 and 1 respectively. The parameter set for the tonic spiking shown in Fig. 5.35g is listed in Table 6. The parameters not listed in the table assume the default values (disabled for digital inputs).

Parameter	Value	Parameter	Value
$I_{refAnalog}$	250	enSpkCmpB	0
V_{leak}	430	enSynBypExc	1
$V_{leakAdapt}$	430	enSynBypInh	1
V_{reset}	430	enMemCap<5:0>	0x3F
V_{thresh}	600	enSynIexc	0
$I_{biasLeak}$	1022	enSynIinh	0
$I_{biasLeakSd}$	1022	enAnaOutMux<1:0>	0
I_{memOff}	0	enAnaIn	0
$I_{biasAdaptRes}$	100	enAnaOut	1
$I_{biasAdaptSd}$	200	enOutAmp	1
I_{AdaptW}	50	enAdapt<1:0>	0x03
V_{synExc}	780	enCapMerge	0
V_{synInh}	780	enVa	1
$I_{globAdapt}$	1022	enVw	1
$I_{globSynSdInh}$	0	enReadVw	0
$I_{globSynSdExc}$	0	enExpWeight<2:0>	0x5
enExp	1	enFireOut	1
enLeak	1	adaptConfig	0xE
enHiConReset	1	resetHoldOff	0xE
enHiConLeak	0	refrCounter	0xE0

Table 6: The hardware parameters used to evoke tonic spiking in the AdEx neuron.

Bibliography

[1] R. W. Williams and K. Herrup, "The control of neuron number," *Annual review of neuroscience*, vol. 11, no. 1, pp. 423–453, 1988.

[2] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, R. Lent, S. Herculano-Houzel *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532–541, 2009.

[3] D. O. Hebb, "The organization of behavior. 1949," *New York Wiley*, 2002.

[4] P. Dayan and L. F. Abbott, *Theoretical neuroscience*. Cambridge, MA: MIT Press, 2001, vol. 806.

[5] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*, 4th ed. New York: McGraw-Hill, 2000.

[6] T. J. Sejnowski, "The computer and the brain revisited," *Annals of the History of Computing*, vol. 11, no. 3, pp. 197–201, 1989.

[7] S. Furber, "To build a brain," *IEEE Spectrum*, vol. 49, no. 8, pp. 44–49, August 2012.

[8] K. Meier, "Special report: Can we copy the brain?-the brain as computer," *IEEE Spectrum*, vol. 54, no. 6, pp. 28–33, 2017.

[9] P. Kogge, "Next-generation supercomputers," *IEEE Spectrum, February*, 2011.

[10] M. F. Wehner, L. Oliker, and J. Shalf, "Low-power supercomputers," *IEEE Spectrum, October*, 2009.

[11] J. Backus, "Can programming be liberated from the von neumann style?: a functional style and its algebra of programs," *Communications of the ACM*, vol. 21, no. 8, pp. 613–641, 1978.

[12] C. A. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, pp. 1629–1636, 1990.

[13] C. A. Mead, *Analog VLSI and Neural Systems*. Addison Wesley, 1989.

[14] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[15] S. Gerrell, "The machine of a new soul," *Economist*, vol. 29, 2013.

[16] Ramón y Cajal, *Histologie du Systeme Nerveux de l'homme et des Vertebres*, 1909, vol. v. 1.

[17] O. Torres-Fernández, C. Golgi, and S. Ramón y Cajal, "The Golgi silver impregnation method: commemorating the centennial of the Nobel Prize in medicine (1906) shared by Camillo Golgi and Santiago Ramón y Cajal," *Biomedica*, vol. 26, pp. 498–508, Dec 2006.

[18] O. C. Biology. (2017) Neurons and glial cells, CC BY 3.0. [Online]. Available: http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.118

[19] T. K. Academy. (2017) The neuron and nervous system, CC BY-NC-SA 3.0 US. [Online]. Available: https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/overview-of-neuron-structure-and-function

[20] D. E. Goldman, "Potential, impedance, and rectification in membranes," *The Journal of general physiology*, vol. 27, no. 1, pp. 37–60, 1943.

[21] A. L. Hodgkin and B. Katz, "The effect of sodium ions on the electrical activity of the giant axon of the squid," *The Journal of physiology*, vol. 108, no. 1, pp. 37–77, 1949.

[22] C. S. Patlak, "Derivation of an equation for the diffusion potential," *Nature*, vol. 188, no. 4754, p. 944, 1960.

[23] Chris73. (2007) Wikimedia, CC BY-SA 3.0. [Online]. Available: https://commons.wikimedia.org/wiki/File:Action_potential.svg

[24] M. A. Petrovici, "Form vs. function: Theory and models for neuronal substrates," Ph.D. dissertation, Universität Heidelberg, 2015.

[25] W. Gerstner and R. Naud, "How good are neuron models?" *Science*, vol. 326, no. 5951, pp. 379–380, 2009.

[26] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[27] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[28] 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

[29] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical journal*, vol. 1, no. 6, pp. 445–466, 1961.

[30] C. Morris and H. Lecar, "Voltage oscillations in the barnacle giant muscle fiber," *Biophysical journal*, vol. 35, no. 1, pp. 193–213, 1981.

[31] L. Lapicque, "Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization," *Journal de Physiologie et Pathologie General*, vol. 9, pp. 620–635, 1907.

[32] 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

[33] G. B. Ermentrout and N. Kopell, "Parabolic bursting in an excitable system coupled with a slow oscillation," *SIAM Journal on Applied Mathematics*, vol. 46, no. 2, pp. 233–253, 1986.

[34] N. Fourcaud-Trocmé, D. Hansel, C. Van Vreeswijk, and N. Brunel, "How spike generation mechanisms determine the neuronal response to fluctuating inputs," *Journal of Neuroscience*, vol. 23, no. 37, pp. 11 628–11 640, 2003.

[35] R. Jolivet, T. J. Lewis, and W. Gerstner, "Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy," *Journal of neurophysiology*, vol. 92, no. 2, pp. 959–976, 2004.

[36] 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.

[37] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, 1999.

[38] W. Gerstner, W. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics*. Cambridge University Press, 2014.

[39] H. Tuckwell, "Introduction to theoretical neurobiology. vol. 1, linear cable theory and dendritic structure and stochastic theories," 1988.

[40] C. Koch and I. Segev, "The role of single neurons in information processing," *Nature neuroscience*, vol. 3, no. 11s, p. 1171, 2000.

[41] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on neural networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[42] 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.

[43] W. Gerstner and R. Brette, "Adaptive exponential integrate-and-fire model," *Scholarpedia*, vol. 4, no. 6, p. 8427, 2009, revision #90944.

[44] E. M. Izhikevich, *Dynamical systems in neuroscience*. MIT press, 2007.

[45] H. Markram, Y. Wang, and M. Tsodyks, "Differential signaling via the same axon of neocortical pyramidal neurons," *Proceedings of the National Academy of Sciences*, vol. 95, no. 9, pp. 5323–5328, 1998.

[46] 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, 2008.

[47] M. V. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proceedings of the National Academy of Sciences*, vol. 94, no. 2, pp. 719–723, 1997.

[48] H. Markram and B. Sakmann, "Action potentials propagating back into dendrites trigger changes in efficacy of single-axon synapses between layer v pyramidal neurons," in *Soc. Neurosci. Abstr*, vol. 21, no. 3, 1995, p. 2007.

[49] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps," *Science*, vol. 275, no. 5297, pp. 213–215, 1997.

[50] G.-q. Bi and M.-m. Poo, "Synaptic modification by correlated activity: Hebb's postulate revisited," *Annual review of neuroscience*, vol. 24, no. 1, pp. 139–166, 2001.

[51] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron*, vol. 32, no. 6, pp. 1149–1164, 2001.

[52] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature reviews. Neuroscience*, vol. 5, no. 2, p. 97, 2004.

[53] B. A. y. Arcas, A. L. Fairhall, and W. Bialek, "Computation in a single neuron: Hodgkin and huxley revisited," *Neural Computation*, vol. 15, no. 8, pp. 1715–1749, 2003.

[54] C. Koch, "Computation and the single neuron," *Nature*, vol. 385, no. 6613, p. 207, 1997.

[55] I. Segev, "Sound grounds for computing dendrites," *Nature*, vol. 393, no. 6682, p. 207, 1998.

[56] S. D. Antic, W.-L. Zhou, A. R. Moore, S. M. Short, and K. D. Ikonomu, "The decade of the dendritic nmda spike," *Journal of neuroscience research*, vol. 88, no. 14, pp. 2991–3001, 2010.

[57] W. Rall, "Theoretical significance of dendritic trees for neuronal input-output relations," *Neural theory and modeling*, pp. 73–97, 1964.

[58] S. R. Young and E. W. Rubel, "Embryogenesis of arborization pattern and topography of individual axons in n. laminaris of the chicken brain stem," *Journal of Comparative Neurology*, vol. 254, no. 4, pp. 425–459, 1986.

[59] H. Agmon-Snir, C. E. Carr, and J. Rinzel, "The role of dendrites in auditory coincidence detection," *Nature*, vol. 393, no. 6682, p. 268, 1998.

[60] C. Carr and M. Konishi, "A circuit for detection of interaural time differences in the brain stem of the barn owl," *Journal of Neuroscience*, vol. 10, no. 10, pp. 3227–3246, 1990.

[61] H. B. Barlow and R. M. Hill, "Selective sensitivity to direction of movement in ganglion cells of the rabbit retina," *Science*, pp. 412–414, 1963.

[62] W. R. Taylor, S. He, W. R. Levick, and D. I. Vaney, "Dendritic computation of direction selectivity by retinal ganglion cells," *Science*, vol. 289, no. 5488, pp. 2347–2350, 2000.

[63] G. Stuart, N. Spruston, B. Sakmann, and M. Häusser, "Action potential initiation and backpropagation in neurons of the mammalian cns," *Trends in neurosciences*, vol. 20, no. 3, pp. 125–131, 1997.

[64] M. London and M. Häusser, "Dendritic computation," *Annu. Rev. Neurosci.*, vol. 28, pp. 503–532, 2005.

[65] J. C. Magee and D. Johnston, "A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons," *Science*, vol. 275, no. 5297, pp. 209–213, 1997.

[66] M. E. Larkum, J. J. Zhu, and B. Sakmann, "A new cellular mechanism for coupling inputs arriving at different cortical layers," *Nature*, vol. 398, no. 6725, p. 338, 1999.

[67] R. Iansek and S. Redman, "The amplitude, time course and charge of unitary excitatory post-synaptic potentials evoked in spinal motoneurone dendrites," *The Journal of physiology*, vol. 234, no. 3, pp. 665–688, 1973.

[68] J. C. Magee, "Dendritic integration of excitatory synaptic input," *Nature Reviews Neuroscience*, vol. 1, no. 3, p. 181, 2000.

[69] S. Gasparini, M. Migliore, and J. C. Magee, "On the initiation and propagation of dendritic spikes in ca1 pyramidal neurons," *Journal of Neuroscience*, vol. 24, no. 49, pp. 11 046–11 056, 2004.

[70] N. L. Golding and N. Spruston, "Dendritic sodium spikes are variable triggers of axonal action potentials in hippocampal ca1 pyramidal neurons," *Neuron*, vol. 21, no. 5, pp. 1189–1200, 1998.

[71] J. Schiller, G. Major, H. J. Koester, and Y. Schiller, "Nmda spikes in basal dendrites of cortical pyramidal neurons," *Nature*, vol. 404, no. 6775, p. 285, 2000.

[72] B. A. Milojkovic, M. S. Radojicic, P. S. Goldman-Rakic, and S. D. Antic, "Burst generation in rat pyramidal neurones by regenerative potentials elicited in a restricted part of the basilar dendritic tree," *The Journal of physiology*, vol. 558, no. 1, pp. 193–211, 2004.

[73] M. E. Larkum, T. Nevian, M. Sandler, A. Polsky, and J. Schiller, "Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: a new unifying principle," *Science*, vol. 325, no. 5941, pp. 756–760, 2009.

[74] J. Schiller and Y. Schiller, "Nmda receptor-mediated dendritic spikes and coincident signal amplification," *Current opinion in neurobiology*, vol. 11, no. 3, pp. 343–348, 2001.

[75] T. Nevian, M. E. Larkum, A. Polsky, and J. Schiller, "Properties of basal dendrites of layer 5 pyramidal neurons: a direct patch-clamp recording study," *Nature neuroscience*, vol. 10, no. 2, p. 206, 2007.

[76] S. Diekelmann and J. Born, "The memory function of sleep," *Nature Reviews Neuroscience*, vol. 11, no. 2, p. 114, 2010.

[77] G. Kiene, "Evaluating the synaptic input of a neuromorphic circuit," Bachelor Thesis, Universität Heidelberg, 2014.

[78] Y. Stradmann, "Characterization and calibration of a mixed-signal leaky integrate and fire neuron on HICANN-DLS," Bachelor Thesis, Universität Heidelberg, 2016.

[79] S. A. Aamir, Y. Stradmann, P. Müller, C. Pehle, A. Hartel, A. Grübl, J. Schemmel, and K. Meier, "An accelerated LIF neuronal network array for a large scale mixed-signal neuromorphic architecture," *IEEE Transactions of Circuits and Systems I: Regular Papers*, 2018.

[80] G. M. Güttler, "Achieving a higher integration level of neuromorphic hardware using wafer embedding," Ph.D. dissertation, Ruprecht-Karls-Universität Heidelberg, 2017.

[81] 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.

[82] S. Friedmann, "A new approach to learning in neuromorphic hardware," Ph.D. dissertation, Ruprecht-Karls-Universität Heidelberg, 2013.

[83] S. Friedmann, J. Schemmel, A. Grübl, A. Hartel, M. Hock, and K. Meier, "Demonstrating hybrid learning in a flexible neuromorphic hardware system," *IEEE Transactions on Biomedical Circuits and Systems*, vol. PP, no. 99, pp. 1–15, 2016.

[84] S. Friedmann, "The nux processor v3.0," 2015. [Online]. Available: https://github.com/electronicvisions/nux

[85] S. Friedmann, N. Frémaux, J. Schemmel, W. Gerstner, and K. Meier, "Reward-based learning under hardware constraints - using a RISC processor in a neuromorphic system," *Frontiers in Neuromorphic Engineering*, September 2013. [Online]. Available: http://arxiv.org/abs/1303.6708

[86] J. Schemmel, "Brainscales 2: A novel architecture for analog accelerated neuromophic computing and hybrid plasticity," *Internal document, ASIC Lab., Kirchhoff-Institute for Physics*, 2017.

[87] S. Billaudelle, "Design and implementation of a short term plasticity circuit for a 65 nm neuromorphic hardware system," Masterarbeit, Universität Heidelberg, 2017.

[88] T. C. Carusone, D. A. Johns, and K. W. Martin, *Analog integrated circuit design*. John Wiley & Sons, 2012.

[89] M. Hock, "Modern semiconductor technologies for neuromorphic hardware," Ph.D. dissertation, Ruprecht-Karls-Universität Heidelberg, 2014.

[90] M. Hock, A. Hartel, J. Schemmel, and K. Meier, "An analog dynamic memory array for neuromorphic hardware," in *2013 European Conference on Circuit Theory and Design (ECCTD)*, Sept 2013, pp. 1–4.

[91] S. A. Aamir, P. Müller, L. Kriener, G. Kiene, J. Schemmel, and K. Meier, "From lif to adex neuron models: Accelerated analog 65 nm cmos implementation," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, October 2017.

[92] J. Schemmel, L. Kriener, P. Müller, and K. Meier, "An accelerated analog neuromorphic hardware system emulating nmda- and calcium-based nonlinear dendrites," in *International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 2217–2226.

[93] S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, and K. Meier, "A highly tunable 65-nm cmos lif neuron for a large scale neuromorphic system," in *Proceedings of 42nd European Solid-State Circuits Conference (ESSCIRC)*, Sept 2016, pp. 71–74.

[94] G. Kiene, "Mixed-signal neuron and readout circuits for a neuromorphic system," Masterthesis, Universität Heidelberg, 2017.

[95] S. Schmitt, J. Klähn, G. Bellec, A. Grübl, M. Güttler, A. Hartel, S. Hartmann, D. Husmann, K. Husmann, S. Jeltsch, V. Karasenko, M. Kleider, C. Koke, A. Kononov, C. Mauch, E. Müller, P. Müller, J. Partzsch, M. A. Petrovici, S. Schiefer, S. Scholze, V. Thanasoulis, B. Vogginger, R. Legenstein, W. Maass, C. Mayr, R. Schüffny, J. Schemmel, and K. Meier, "Neuromorphic hardware in the loop: Training a deep spiking network on the brainscales wafer-scale system," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 2227–2234.

[96] J. V. Arthur and K. A. Boahen, "Silicon-neuron design: A dynamical systems approach," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 5, pp. 1034–1043, 2011.

[97] A. S. Cassidy, P. Merolla, J. V. Arthur, S. K. Esser, B. Jackson, R. Alvarez-Icaza, P. Datta, J. Sawada, T. M. Wong, V. Feldman *et al.*, "Cognitive computing building block: A versatile and efficient digital neuron model for neurosynaptic cores," in *International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–10.

[98] S. Millner, A. Grübl, K. Meier, J. Schemmel, and M. Schwartz, "A VLSI implementation of the adaptive exponential integrate-and-fire neuron model," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 1642–1650.

[99] S. Millner, "Development of a multi-compartment neuron model emulation," Ph.D. dissertation, Ruprecht-Karls University Heidelberg, November 2012. [Online]. Available: http://www.ub.uni-heidelberg.de/archiv/13979

[100] A. Destexhe, "Self-sustained asynchronous irregular states and Up/Down states in thalamic, cortical and thalamocortical networks of nonlinear integrate-and-fire neurons." *Journal of Computational Neuroscience*, vol. 3, pp. 493 – 506, 2009.

[101] A. Destexhe and D. Contreras, "Neuronal computations with stochastic network states," *Science*, vol. 314, no. 5796, pp. 85–90, 2006.

[102] B. Nessler, M. Pfeiffer, L. Buesing, and W. Maass, "Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity," *PLoS Computational Biology*, vol. 9, no. 4, p. e1003037, 2013.

[103] T. P. Vogels and L. F. Abbott, "Signal propagation and logic gating in networks of integrate-and-fire neurons," *J Neurosci*, vol. 25, no. 46, pp. 10 786–95, Nov 2005.

[104] G. Deco and V. K. Jirsa, "Ongoing cortical activity at rest: criticality, multistability, and ghost attractors," *The Journal of Neuroscience*, vol. 32, no. 10, pp. 3366–3375, 2012.

[105] M. A. Petrovici, B. Vogginger, P. Müller, O. Breitwieser *et al.*, "Characterization and compensation of network-level anomalies in mixed-signal neuromorphic modeling platforms," *PloS one*, vol. 9, no. 10, 2014.

[106] M. A. Petrovici, I. Bytschok, J. Bill, J. Schemmel, and K. Meier, "The high-conductance state enables neural sampling in networks of lif neurons," *BMC Neuroscience*, vol. 16, no. Suppl 1, p. O2, 2015.

[107] J. Kremkow, L. Perrinet, G. Masson, and A. Aertsen, "Functional consequences of correlated excitatory and inhibitory conductances in cortical networks." *J Comput Neurosci*, vol. 28, pp. 579–594, 2010.

[108] M. Pospischil, Z. Piwkowska, M. Rudolph, T. Bal, and A. Destexhe, "Calculating event-triggered average synaptic conductances from the membrane potential," *J. Neurophysiology*, vol. 97, p. 2544, 2007.

[109] T. Masquelier and G. Deco, "Network bursting dynamics in excitatory cortical neuron cultures results from the combination of different adaptive mechanism," *PloS one*, vol. 8, no. 10, p. e75824, 2013.

[110] P. Müller, "Modeling and verification for a scalable neuromorphic substrate," Ph.D. dissertation, Ruprecht-Karls-Universität Heidelberg, 2017.

[111] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical mos transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog integrated circuits and signal processing*, vol. 8, no. 1, pp. 83–114, 1995.

[112] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov 2006.

[113] D. Brüderle *et al.*, "A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems," *Biological Cybernetics*, vol. 104, pp. 263–296, 2011.

[114] M. Banu and Y. Tsividis, "Floating voltage-controlled resistors in cmos technology," *Electronics Letters*, vol. 18, no. 15, pp. 678–679, July 1982.

[115] T. Shimmi, H. Kobayashi, T. Yagi, T. Sawaji, T. Matsumoto, and A. Abidi, "A parallel analog cmos signal processor for image contrast enhancement," in *Eighteenth European Solid-State Circuits Conference*, Sept 1992, pp. 163–166.

[116] H. Kobayashi, J. White, and A. Abidi, "An active resistor network for gaussian filtering of images," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 5, pp. 738–748, May 1991.

[117] S. Sakurai and M. Ismail, "A cmos square-law programmable floating resistor independent of the threshold voltage," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, no. 8, pp. 565–574, Aug 1992.

[118] L. Wang and R. Newcomb, "An adjustable cmos floating resistor," in *IEEE International Symposium on Circuits and Systems*, May 2008, pp. 1708–1711.

[119] I. Han and S. B. Park, "Voltage-controlled linear resistor by two mos transistors and its application to active rc filter mos integration," *Proceedings of the IEEE*, vol. 72, no. 11, pp. 1655–1657, Nov 1984.

[120] G. Wilson and P. Chan, "Novel voltage-controlled grounded resistor," *Electronics Letters*, vol. 25, no. 25, pp. 1725–1726, Dec 1989.

[121] Z. Wang, "Novel voltage-controlled grounded resistor," *Electronics Letters*, vol. 26, no. 20, pp. 1711–1712, Sept 1990.

[122] F. Cannillo, C. Toumazou, and T. S. Lande, "Bulk-drain connected load for subthreshold mos current-mode logic," *Electronics Letters*, vol. 43, no. 12, pp. 662–664, June 2007.

[123] F. Cannillo, C. Toumazou, and T. S. Lande, "Nanopower subthreshold mcml in submicrometer cmos technology," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 8, pp. 1598–1611, Aug 2009.

[124] A. Tajalli, E. J. Brauer, Y. Leblebici, and E. Vittoz, "Subthreshold source-coupled logic circuits for ultra-low-power applications," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1699–1710, July 2008.

[125] A. Tajalli, Y. Leblebici, and E. J. Brauer, "Implementing ultra-high-value floating tunable cmos resistors," *Electronics Letters*, vol. 44, no. 5, pp. 349–350, Feb 2008.

[126] A. Tajalli and Y. Leblebici, "A widely-tunable and ultra-low-power mosfet-c filter operating in subthreshold," in *2009 IEEE Custom Integrated Circuits Conference*, Sept 2009, pp. 593–596.

[127] R. J. Baker, *CMOS: circuit design, layout, and simulation*. John Wiley & Sons, 2008.

[128] H. Yoshizawa, Y. Huang, P. F. Ferguson, and G. C. Temes, "Mosfet-only switched-capacitor circuits in digital cmos technology," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 6, pp. 734–747, Jun 1999.

[129] T. Tille, J. Sauerbrey, and D. Schmitt-Landsiedel, "A 1.8-v mosfet-only sigma; delta; modulator using substrate biased depletion-mode mos capacitors in series compensation," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 7, pp. 1041–1047, Jul 2001.

[130] T. Tille, J. Sauerbrey, M. Mauthe, and D. Schmitt-Landsiedel, "Design of low-voltage mosfet-only sigma; delta; modulators in standard digital cmos technology," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 1, pp. 96–109, Jan 2004.

[131] P. R. Gray, P. Hurst, R. G. Meyer, and S. Lewis, *Analysis and design of analog integrated circuits*. Wiley, 2001.

[132] Y. P. Tsividis and P. R. Gray, "An integrated nmos operational amplifier with internal compensation," *IEEE Journal of Solid-State Circuits*, vol. 11, no. 6, pp. 748–753, 1976.

[133] R. D. Jolly and R. H. McCharles, "A low-noise amplifier for switched capacitor filters," *IEEE Journal of Solid-State Circuits*, vol. 17, no. 6, pp. 1192–1194, 1982.

[134] B. K. Ahuja, "An improved frequency compensation technique for cmos operational amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 18, no. 6, pp. 629–633, 1983.

[135] P. J. Hurst, S. H. Lewis, J. P. Keane, F. Aram, and K. C. Dyer, "Miller compensation using current buffers in fully differential cmos two-stage operational amplifiers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 2, pp. 275–285, 2004.

[136] V. Saxena, "Indirect feedback compensation technique for multi-stage operational amplifiers," Masterthesis, Boise State University, 2007.

[137] V. Saxena and R. J. Baker, "Compensation of cmos op-amps using split-length transistors," in *Midwest Symposium on Circuits and Systems*, Aug 2008, pp. 109–112.

[138] S. A. Aamir, P. Angelov, and J. J. Wikner, "1.2-v analog interface for a 300-msps hd video digitizer in core 65-nm cmos," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 4, pp. 888–898, April 2014.

[139] S. A. Aamir, P. Harikumar, and J. J. Wikner, "Frequency compensation of high-speed, low-voltage CMOS multistage amplifiers," in *International Symposium on Circuits and Systems (ISCAS)*, May 2013, pp. 381–384.

[140] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 2001.

[141] D. Goodman and R. Brette, "Brian: a simulator for spiking neural networks in Python," *Front. Neuroinform.*, vol. 2, no. 5, 2008.

[142] F. Gardner, "Charge-pump phase-lock loops," *IEEE Transactions on Communications*, vol. 28, no. 11, pp. 1849–1858, Nov 1980.

[143] D. K. Jeong, G. Borriello, D. A. Hodges, and R. H. Katz, "Design of pll-based clock generation circuits," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 2, pp. 255–261, Apr 1987.

[144] J.-S. Lee, M.-S. Keel, S.-I. Lim, and S. Kim, "Charge pump with perfect current matching characteristics in phase-locked loops," *Electronics Letters*, vol. 36, no. 23, pp. 1907–1908, 2000.

[145] G. Wegmann, E. A. Vittoz, and F. Rahali, "Charge injection in analog mos switches," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 6, pp. 1091–1097, 1987.

[146] C. Eichenberger and W. Guggenbuhl, "On charge injection in analog mos switches and dummy switch compensation techniques," *IEEE Transactions on Circuits and Systems*, vol. 37, no. 2, pp. 256–264, Feb 1990.

[147] S. Cheng, H. Tong, J. Silva-Martinez, and A. I. Karsilayan, "Design and analysis of an ultrahigh-speed glitch-free fully differential charge pump with minimum output current variation and accurate matching," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 9, pp. 843–847, Sept 2006.

[148] I. C. Hwang and S. G. Bae, "Low-glitch, high-speed charge-pump circuit for spur minimisation," *Electronics Letters*, vol. 45, no. 25, pp. 1273–1274, December 2009.

[149] S. Ethier and M. Sawan, "Exponential current pulse generation for efficient very high-impedance multisite stimulation," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 1, pp. 30–38, Feb 2011.

[150] M. Kleider, "Neuron circuit characterization in a neuromorphic system," Ph.D. dissertation, Ruprecht-Karls-Universität Heidelberg, 2017.

[151] L. Kriener, "Characterization of single-neuron dynamics in the development of neuromorphic hardware," Master's thesis, Heidelberg University, March 2017.

[152] W. Rhee, "Design of high-performance cmos charge pumps in phase-locked loops," in *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on*, vol. 2, Jul 1999, pp. 545–548.

[153] G. Palumbo and D. Pappalardo, "Charge pump circuits: An overview on design strategies and topologies," *IEEE Circuits and Systems Magazine*, vol. 10, no. 1, pp. 31–45, 2010.

[154] T.-H. Lin, C.-L. Ti, and Y.-H. Liu, "Dynamic current-matching charge pump and gated-offset linearization technique for delta-sigma fractional-n plls," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 5, pp. 877–885, 2009.

[155] M.-S. Shiau, H.-S. Hsu, C.-H. Cheng, H.-H. Weng, H.-C. Wu, and D.-G. Liu, "Reduction of current mismatching in the switches-in-source cmos charge pump," *Microelectronics Journal*, vol. 44, no. 12, pp. 1296–1301, 2013.

[156] A. Destexhe, "High-conductance state," *Scholarpedia*, vol. 2, no. 11, p. 1341, 2007, revision #88146.

[157] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[158] 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. 62, no. 12, pp. 2454–2467, 2013.

[159] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 699–716, 2014.

[160] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses," *Frontiers in neuroscience*, vol. 9, 2015.

[161] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," *IEEE transactions on neural networks*, vol. 18, no. 1, pp. 253–265, 2007.

[162] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. S. Modha, "A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*. IEEE, 2011, pp. 1–4.

[163] N. Imam, K. Wecker, J. Tse, R. Karmazin, and R. Manohar, "Neural spiking dynamics in asynchronous digital circuits," in *International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.

[164] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[165] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, 2015.

[166] B. V. Benjamin, J. V. Arthur, P. Gao, P. Merolla, and K. Boahen, "A superposable silicon synapse with programmable reversal potential," in *Engineering in Medicine and Biology Society (EMBC) Annual International Conference of the IEEE*, 2012, pp. 771–774.

[167] P. Gao, B. V. Benjamin, and K. Boahen, "Dynamical system guided mapping of quantitative neuronal models onto neuromorphic hardware," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 10, pp. 2383–2394, 2012.

[168] A. Joubert, B. Belhadj, O. Temam, and R. Héliot, "Hardware spiking neurons design: Analog or digital?" in *International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–5.

[169] S. Dytckov and M. Daneshtalab, "Computing with hardware neurons: spiking or classical? perspectives of applied spiking neural networks from the hardware side," *arXiv preprint arXiv:1602.02009*, 2016.

[170] F. Alibart, E. Zamanidoost, and D. B. Strukov, "Pattern classification by memristive crossbar circuits using ex situ and in situ training," *Nature communications*, vol. 4, 2013.

[171] K.-H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, "A functional hybrid memristor crossbar-array/cmos system for data storage and neuromorphic applications," *Nano letters*, vol. 12, no. 1, pp. 389–395, 2011.

[172] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[173] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[174] M. Mayberry, "Intel's New Self-Learning Chip Promises to Accelerate Artificial Intelligence," https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/, 2017, [Online; accessed 28-Nov-2017].

[175] S. Kumar, "Introducing Qualcomm Zeroth Processors: Brain-Inspired Computing," https://www.qualcomm.com/news/onq/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing, 2017, [Online; accessed 28-Nov-2017].

Acknowledgments

I would like to express my gratitude to Prof. Dr. Karlheinz Meier for the opportunity to pursue doctoral work in his group in Heidelberg and to Prof. Dr. Marc Weber for graciously accepting me as a PhD student at IPE in Karlsruhe.

I would like to extend my gratitude to Dr. Johannes Schemmel for the technical leadership and always being available for in depth discussions. Further, thanks to the first generation *hardies*, i.e., Andreas Grübl, Simon Friedmann, Matthias Hock and Andreas Hartel for the teamwork on chip development and many fruitful discussions.

My sincere thanks go to Yannik Stradmann for his bachelor work under my supervision, for the calibration/measurement of the *DLS-2* neuron and for our collaborative LIF paper. Thanks to Paul Müller for the long biological discussions, extensive verification, and the work towards the publications. For the teamwork on the third prototype, thanks to Korbinian Schreiber, Sebastian Billaudelle and Gerd Kiene. The EDA tools and measurement support from Ralf Achenbach and Markus Dorn is sincerely acknowledged.

Thanks to Eric Müller for the discussions spanning from popular culture to technology and for keeping the *espressomaschine* running! Thanks to Mihai Petrovici for exploring and organizing exotic sports, and to all the proof-readers including Sebastian Schmitt and Mitja Kleider. Thanks to all the rest of the Electronic Visions(s) group members (*hardies*, *softies* and *TMAs*) for being great colleagues and for the friendly atmosphere.

The kind administrative support extended by the staff at IPE/KIT as well as at KIP is gratefully acknowledged.

Thanks to Dr. J Jacob Wikner, my former supervisor and colleague whose tips still help me to this day. I thank all my Heidelberg friends for the wonderful times we have spent in this beautiful city.

Finally, I wish to thank my parents for their love, encouragement and support.