Realisation of a neural network, based on pulse width modulation

de Rooij, M.

Award date:
1997

REALISATION OF A NEURAL NETWORK, BASED ON PULSE WIDTH MODULATION

Master thesis of M. de Rooij

Supervisor	:	Prof. Dr. Ir. W.M.G. van Bokhoven
Coach	:	Dr. Ir. J.A. Hegt
Period	:	May 1 1996 - August 28 1997
At	:	Systems for Electronic Signal Processing (SES) Faculty of Electrical Engineering Eindhoven University of Technology

Abstract

A biological neural network (for example the human brain) can very quickly and accurately process information. It is able to recognize patterns and to identify incomplete patterns, even if they are incomplete and buried in noise, or even when certain neurons have failed. Because of these facts, it is very desireable to have an electronic circuit that can do the same. The two most important units of a neural network are the synapse and the neuron. This work deals with the electronic imitation of these two units.

The synapse imitation is done by a four-quadrant multiplier, with an input consisting of a pulse with variable duration and a weighting factor between -1 and +1. The synapse has a low power dissipation (13.2μW), excellent linearity in the case of output voltage versus input pulse width and less linearity in the case of output voltage versus weight. Also, it is small in circuit size, and has a large weight input range (2V).

The neuron imitation is done by an integrator/sample & hold part (relatively low power dissipation of 42μW, ability to adjust the circuit to the amount of connected synapses by changing the integrator capacitor), an inverse sigmoid part to realise a saturation in the neuron's response (shape can be adjusted), and a comparator part to generate an output pulse with a duration dependent on the comparison between the sample & hold circuit and the inverse sigmoid circuit (offset voltage ≤ 0.4mV, propagation delay time ≤ 15.6ns with a 10pF load capacitance and ≤ 18.7ns with a 20pF load capacitance).

The complete circuit realises a saturation in the neuron's response: a sigmoid relation between the output pulse width and the inputs (input pulse width and weight) of the circuit (synapse unit).

Contents

Introduction

The human brain processes information very quickly and very accurately. It is able to recognize patterns and to identify incomplete patterns, even when certain neurons have failed. For example, when we speak with a person in a crowded and noisy room, we are able to filter out his or hers voice. This ability of the brain to recognize information, literally buried in noise, and retrieve it correctly is one of the amazing processes that we wish could be duplicated by a machine.

Over the past few decades, a serious attempt has been made to design electronic circuits that closely resemble biological neural networks. There is some difference between a biological neural network and an artificial neural network. The first one is the way a signal passes through the network. In the biological neural network this is done by chemical processes and in an artificial neural network by electronic processes. Another difference is the speed of processing. In het human brain, the neurons are communicating with frequencies in the order of 100Hz. In the artificial neural network we will be using frequencies in the order of 1MHz, so the speed in the neural networks to build is much higher than in an biological neural network.

There are two main approaches to build an artificial neural network; the first one is a digital approach with high accuracy, medium speed and very flexible, and the second one is an analog approach with medium accuracy, high speed and rigid. One can also combine these two approaches to use the best properties of both.

The two most important units of a neural network (for example the human brain) are the synapse and the neuron. Consequently, to build a device that resemble a biological neural network, the first step is to design a synapse and a neuron. This work deals with the electronic imitation of these two units. The synapse imitation will be done by a four-quadrant mulitplier. The input of the multiplier consists of a pulse with variable duration and a weight value. Dependent on the weight value, the multiplier will supply a current to, or withdraw a current from, the neuron unit. The input of the neuron consists of one or more synapses. The neuron converts the current into a voltage, which will be compared with a non-linear signal voltage (an inverse sigmoid signal). The output of the neuron consists of a pulse with a duration dependent on the above mentioned comparison. This output can now be connected to another neuron. In this way, a large artificial neural network can be build.

1. Biological background of neural networks

Because this work deals with the synapse and neuron of an artifical neural network, it is desireable to know how thes units work in a biological neural network. Therefore, the working of these units will be explaned. The works of [5] and [16] have contributed to this chapter, and one can find more details in there.

1.1. The neuron

1.1.1. Description of the neuron

The human brain is the example of a nervous system. One of its most important units is the neuron. The neuron receives and combines signals from other neurons, and transports it further to other neurons. At this way, a signal travels through the nervous systems. In figure 1-1, a representation of a neuron is given. The signals are received in the dendrites, which are grouped into dentritic trees. At this way a very large total surface area will be accomplished. The soma is the main body of the neuron, and the dendritic trees are conntected with it. The interior of the soma is filled with intracellular fluid. The outher boundary of the soma is called the membrane. Outside the membrane is the extracellular fluid.

The signals travels further through the neuron along it's transmission line, the axon. It consists of a row of Schwann-cells, and between these cells are nodes, known as nodes of Ranvier. The Schwann-cells are covered with an insulating material called myelin. The membran capacitance is then reduced and this will cause an increase in the signal propagation speed. This increase in speed is approximately 20 times faster than in a transmission line without myelin. The axon ends in axonic endings, which are connected with dendrites of other neurons.

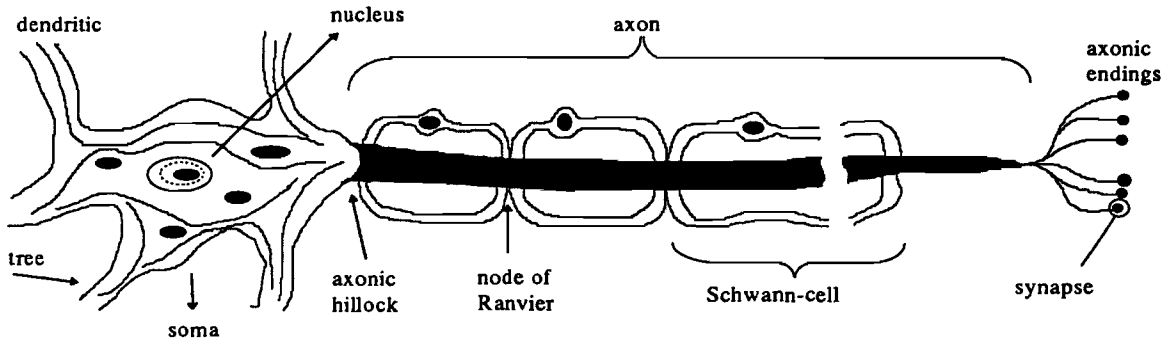


Figure 1-1 Representation of a neuron.

1.1.2. Working of a neuron

Every nerve cell has a threshold voltage, which must be overcome to start a nerve impulse. This impulse starts at the axon hillcock (origin of the axon). When the threshold voltage has been exceed (depolarization), the voltage difference across the axon's membrane is locally lowered. As a result of this voltage difference reduction, channels in the membrane ahead will open and sodium (Na) ions flows into the axon. This will cause a reduction of the voltage difference across this membrane region, and more sodium channels in the

membrane (a little further away) will open, and the proces starts all over again. So, when a nerve cell is triggered to start an impulse, it becomes self-stimulated and the puls continues until it reaches the axonic endings. After the depolarization, the sodium channels in the membrane will close in a few milliseconds. This closing is known as sodium inactivation. A pump mechanism restores the balance of the ion concentration of the membrane. Obviously, this balancing begins at the origin of the axon, because the depolarization has started there first.

The action potential is repeated at fixed intervals. The myelin in the axon is interrupted every few milimeters, forming gaps. These gaps (nodes of Ranvier) will cause a repeating or regeneration of the signal. So the signal is periodically restored, and it's possible to carry up signals to 1 m in length.

At the end of the axon (synapse) the signals will be decoded by means of temporal summation and spatial summation (figure 1-2). In temporal summation, the voltage potential of an impulse is added to previous impulse voltage potentials. So the total sum depends on the amount of impulses and their amplitude. In spatial summation, the summation consists of integration of excitations or inhibitions by all neurons at the target neuron. The summations of voltage potentials from temporal and spatial (spatiotemporal) summation will cause a potential charge, and this charge is encoded as a nerve impulse, transmitted to other cells. The synapse of a neuron integrates the received impulses further over a short time as the charge is stored in the cell membrane. So, the membrane acts first as a capacitor, and when the chemical processes take place, it acts like a messanger. At the soma, all integrated signals are combined. When this combined signal exceeds the threshold voltage of the neuron, it will start a firing process; a signal is produced and transmitted along the axon to it's axonic endings.

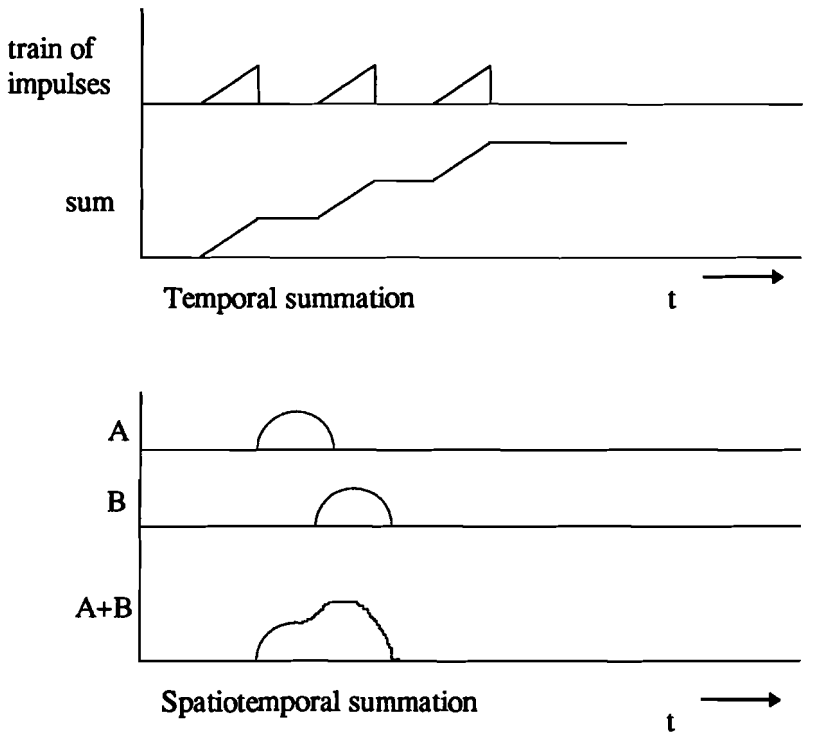


Figure 1-2 Temporal and spatiotemporal summation.

1.2. The synapse

1.2.1. Description of the synapse

The connection (or junction) between the axonic endings of a neuron and the dendrite's of other neurons is called a synapse. On the left (figure 1-3) there is the terminal of an axonic ending, or presynaptic terminal. On the right, one can see the receptor part of another neuron, or postsynaptic terminal. Between the presynaptic and postsynaptic terminals lies the synaptic cleft (approximately 200 nm thin). Within the presynaptic terminal there are an amount of extremely small bladders, called the vesicles. These vesicles store chemical neurotransmitters, which are produced by mitochondria. The fact that a signal, to travel from the presynaptic terminal to the postsynaptic terminal, has to cross this cleft, implies that immediate electrical stimuli-transmission isn't possible. This crossing is done through a chemical diffusion proces. The cleft is filled with extracellular material. One of the containing ions within this material is calcium (Ca), which is very important for transmitting signals between neurons.

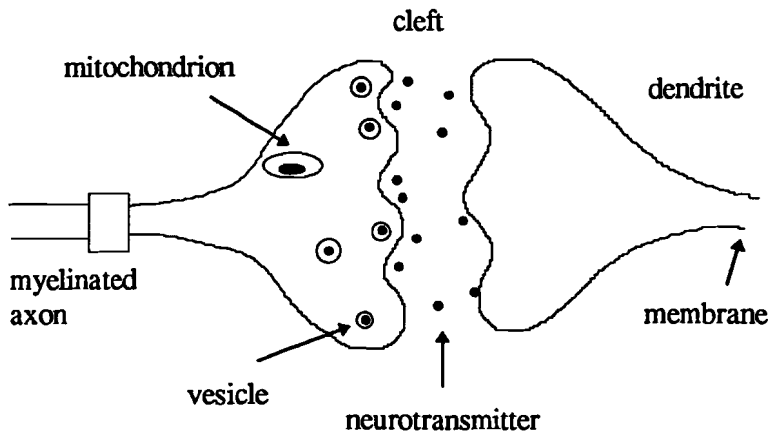


Figure 1-3 Synapse in detail.

The synapse is not only transmitting signals between neurons, but also can change its synaptic efficiency. This synaptic efficiency is a result of synaptic activity, strength and pulse frequency. Thus, action potentials not only encode information, but they can also alter network parameters over time.

1.2.2. Working of a synapse; excitatoire

The stimulitransmission in the synapse is as follows: along the length of the axon of the presynaptic neuron arrives an action potential at the presynaptic terminal. Voltage-activated calcium channels in the membrane are opened by this electrical impulse, and calcium ions pour into the presynaptic terminal. Then the vesicles close to the presynaptic membrane are attracted by the calcium ions, and aid the fusion of the vesicles to the membrane. The vesicles then burst and the neurotransmitter inside the vesicles is released and diffuses through the presynaptic membrane and the cleft till the postsynaptic membrane, which is impenetrable for this neurotransmitter. This process is called exocytosis. This diffusion has a time span of approximately 0.5 msec. Once the ions accomplisch their job, they are neutralized by an as yet indeterminate mechanisme, so that the ion concentration in the presynaptic terminal returns to normal. The vesicles are quickly refilled with a new neurotransmitter (they contain about 10,000 molecules of the same neurotransmitter).

The neurotransmitter act upon the postsynaptic membrane and causes a permeability increase of all ions. The membrane can not longer separate the charges, so the potential difference accross the membrane is lowered; the membrane potential (e_m) is more positive than the resting potential. The value e_m=0 will never be reached because in the cleft there is an enzym that quickly will break down the neurotransmitter. By this, the permeability of the membrane will return to it's former value and the voltage difference accross the membrane will be restored. So, an action-potential in the presynaptic neuron will cause a small and short depolarization wave in the postsynaptic neuron, and will be transported as an subliminal stimuli. Such a wave is called an excitatory postsynaptic potential (E.P.S.P.). Figure 1-4 shows this wave. This type of synapse is called an excitatoire synapse.

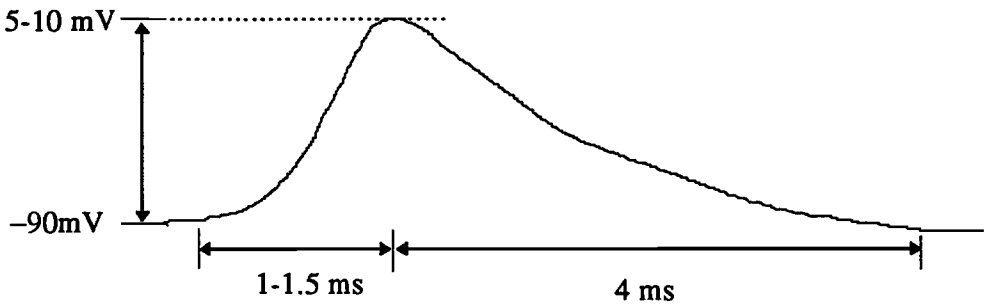


Figure 1-4 Excitatory postsynaptic potential wave.

1.2.3. Working of a synapse; inhibitoire

Besides the excitatoire synapse there is the inhibitoire synapse. They look the same but the difference between them lies in the working. The neurotransmitter in the vesicles and transmitted at the arrival of an action potential is a different one than that of the excitatoire synapse. This neurotransmitter only increases the permeability of the potassium (K) and chloride (Cl) ions. As a result of this, the potential of the inner-cell becomse more negative in accordance with the extracellulare fluid; the potential accross the membrane will rise (hyperpolarisation). The neurontransmitter will be neutrolized and the ion concentration in the presynaptic terminal returns to normal. An action potential in the presynaptic neuron will cause a small and short hyperpolarisation wave in the postsynaptic neuron. Figure 1-5 shows this wave. Such a hyperpolarisation wave is called inhibitory postsynaptic potential (IPSP).

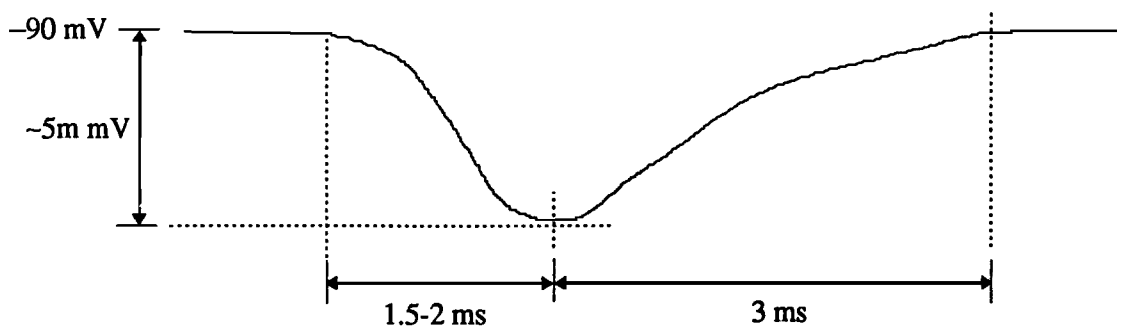


Figure 1-5 Inhibitory postsynaptic potential wave.

1.2.4. Some properties of synapses

The E.P.S.P.'s and the I.P.S.P.'s are allowed to be summated; the local membrane potential equals the resting resting potential plus the summation of the E.P.S.P.'s and the I.P.S.P.'s. The resting potential is about -70mV and the threshold potential -55mV. An action potential will start when at a certain moment in a certain membrane region, the following occurs:

$$\sum \text{E.P.S.P.} + \sum \text{I.P.S.P.} \geq 15\text{mV} \qquad (1.1)$$

For example:

1) 4 E.P.S.P.'s of 5mV, 2 I.P.S.P.'s of −2mV, then (1.1) becomes : 4(5mV) + 2(−2mV) = 16mV ≥ 15mV, so there will be an action potential; $e_m = (-70\text{mV}) + 16\text{mV} = -54\text{mV}$
2) one E.P.S.P. of 8mV, tail of a former E.P.S.P. of 2mV and one I.P.S.P. of −3mV, then (1.1) becomes: 8mV + 2mV − 3mV = 7mV ≤ 15mV, so no action potential will start; $e_m = -70\text{mV} + 7\text{ mV} = -63\text{mV}$.

There are some properties of synapses. They are:

1) The synapse has an oneway traffic, only from presynaptic axon to a postsynaptic soma or den drietes;
2) If an amount of action potentials arrive at a synapse, then it is not certain that the synapse react with an analog row of impulses. The original decharge frequency can change in a positive or negative way at the synapse;
3) In one single nerve cell, all synaptic endings are inhibitoire or all excitatoire. This is not the case with the dendrites (input paths);
4) When a signal appears at a synapse, a charge is generated at the postsynaptic site. The strength of this incoming signal will cause a certain magnitude of the charge. Also, the strength is weighted by a factor associated with its input. The potassium ion flow is responsible for keeping the charge of a cell membrane well below the threshold potential. So, if the potassium ion flow is reduced, the weighting factors of electrical signals will change. This change of weight can last for many days, but is not constant over a longer time.

2. Signal processing in the artificial neural network

In this chapter an overview is presented how signals are processed in the artificial neural network. After an introduction, the complete neural network will be discussed, followed by the different stages of the network.

2.1. Introduction

This work deals with the implementation of the synapse and neuron of an artificial neural network. In figure 2-1, a model is presented of a neuron with N synapses. This model is earlier discussed in [3] and [11]. A neuron can be connected with a different amount of synapses. Every synapse performs a weighting (w_{ij}), which can have a positive or a negative value (excitation and inhibition). The input of a synapse is presented by X_{ij}. The weights are adjustable and their final value is obtained by training. The weights of each synapse is multiplied with the incomming signals and the neuron adds up these multiplications (S_i). These summation becomes now:

$$S_i = \sum_{j=1}^{n} w_{ij}X_{ij} \tag{2.1}$$

After the summation, the resulting voltage is compared with a nonlinearity function. This nonlinearity function is to ensure, that the neuron's response is bounded. The nonlinearity function is not necessarily a close replica of the biological one; often it is merely used for mathematical convenience [5].

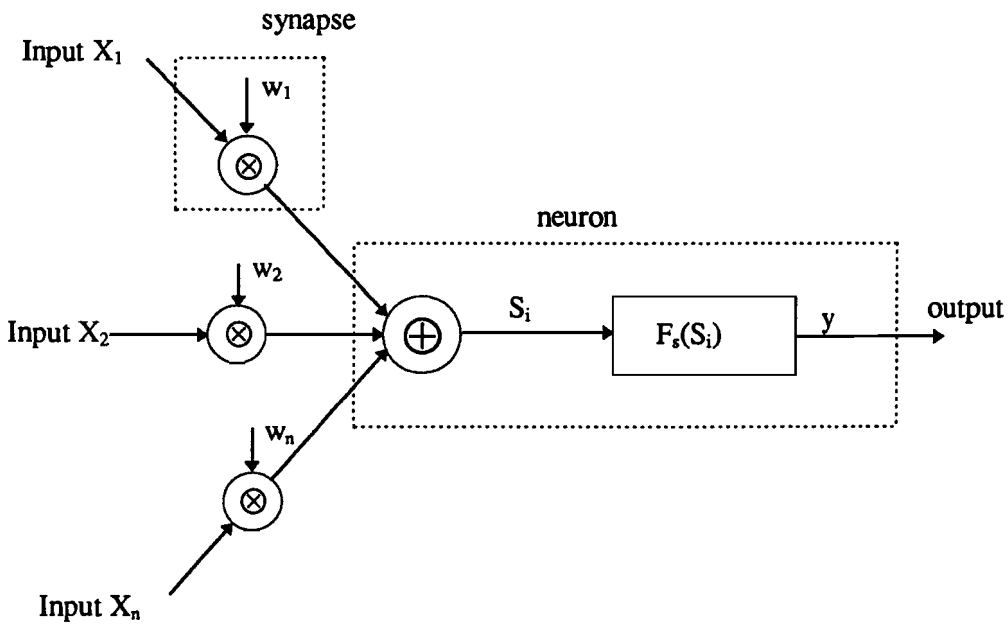


Figure 2-1 Model of a neuron with N synapses.

In the neural network, a pulse stream modulation (CPWM) is used. This modulation has several advantages above other modulation [15]. One of them is that the design of CPWM systems is less complicated to build.

Also the noise sensitivity is much lower that that of analog systems. Furthermore, CPWM presents the lowest computation energy and response time. In the next paragraph this modulation technique will be discussed.

2.2. Pulse stream approach

Pulse stream modulations are based on 'quasi-periodic' binary waveforms. In other modulations, the information is usually contained in the amplitude of a waveform. In pulse stream modulations, the information is contained in the timing. This is the reason why these modulations are used to encode analog values, using binary signals. An overview of several pulse stream modulations is already presented by others ([3], [15], [11]). In this work only one type of pulse stream modulation, Coherent Pulse Width Modulation (CPWM), will be discussed.

CPWM is a modulation technique where the information is encoded in the width of a pulse. All these pulses have a known phase relationship with each other. The whole system (artificial neural network) has an additional reference clock (CCK) that keep these pulses in phase (figure 2-2) . The pulses ($X_{1j}...X_{Nj}$) have a constant frequency $f_0 = 1/T_0$, while their width is proportional to the activation value:

$$T_i = T_{max}\left(\frac{1+\alpha_i}{2}\right) \tag{2.2}$$

Activation values α_i are normalized so that $\alpha_i \in [-1...+1]$, while $T_{max} < T_0$. If an activation value $\alpha_i = 0$, then the width of the pulse becomes $T_{max}/2$.

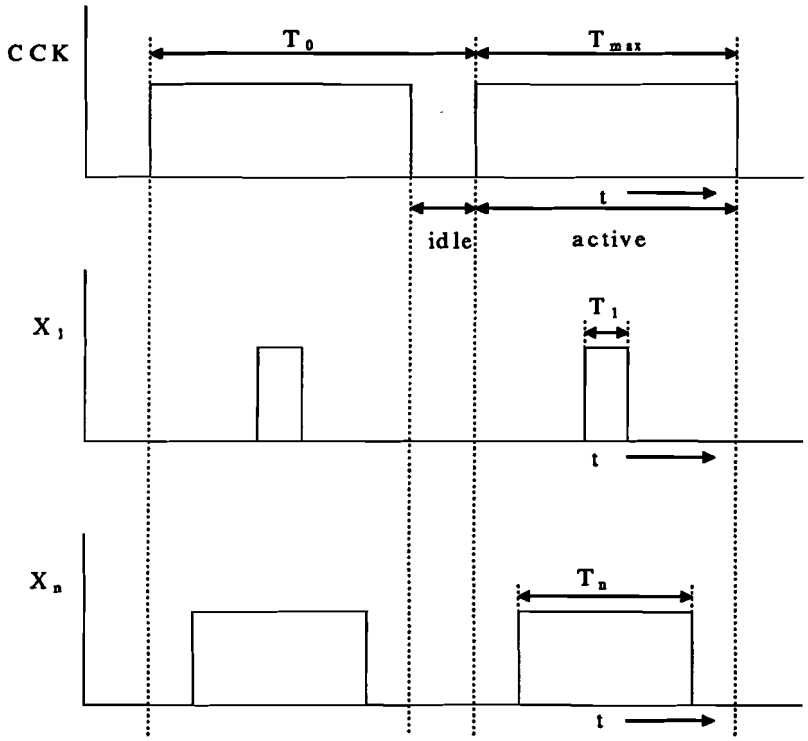


Figure 2-2 Timing diagram of CPWM modulation.

From figure 2-2, one can see that the reference clock defines two phases: an active phase ($T_i \leq T_{max}$) and an idle phase. Only during the active phase, CPWM signals are allowed to be '1', and during the idle phase they must be '0'. This is because during the idle phase, the whole artificial neural network is been reset. Furthermore, it is often convenient to avoid having $T_i = 0$ and $T_i = T_{max}$, since these conditions can increase the influence of charge injection effects and so decreasing the overall accuracy of the system. The pulses are centered within the active phase. For example, if an active phase is between 0ns and 800ns, and the width of a pulse is 300ns, then the pulse begins at 250ns and ends at 550ns.

To improve the speed of the artificial neural network, a frequency f_0 = 1Mhz will be used (in former implementations these frequencies where about 100kHz). Also, the active phase of the clock reference has a time span of 800ns; this leaves 200ns for the idle phase. These times are chosen, so that the artificial neural network has enough time for a reset.

Due to their digital nature and coherence, the CPWM signals can be multiplied more easily than analog signals. Furthermore, CPWM incorporates the main benefit of synchronous digital circuits; in each cycle, new information can be transmitted or processed. Another benefit with respect to synchronous circuits is the reduction of crosstalk noise towards the analog parts. This is done by the spreading of ground and power supply current spikes, caused by state switching. Also, CPWM is not affected by any frequency error, nor any phase uncertainty (provided that phase errors stay small enough, so that they do not exceed the active phase).

2.3. Signal processing in the synapse

The input of the synapse (figure 2-3) consists of a pulse with a duration between a minimum of 0ns and a maximum of 800ns, and a weighting factor w_{ij} between -1 and +1. There is also a reference clock input, which controls the active or idle phase. The output of the synapse delivers a current to, or withdraws a current from, a neuron. Thus, the synapse is actually a converter; it converts a puls duration into a current. Because the synapse is only active in the active phase of the reference clock, the power dissipation will be reduced (about 20 %). In chapter 3, the synapse will be discussed in detail.

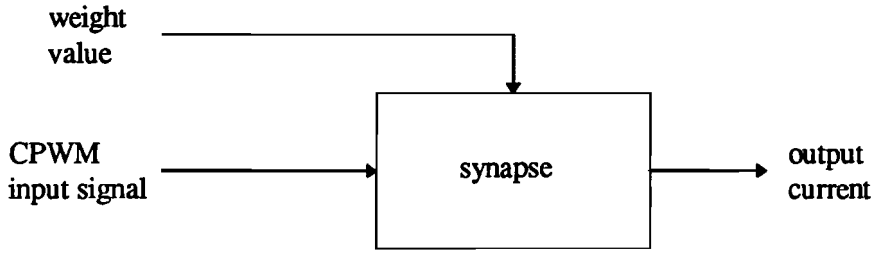


Figure 2-3 Model of a synapse.

2.4. Signal processing in the neuron

The input of the neuron consists of one or more synapses. The neuron adds/withdraws the currents by means of integration. Here, a current is converted into a voltage. This voltage, then will be compared with a non-linear function to ensure a boundary of the neuron's response. The output of the neuron consists of a pulse with a duration dependant on the comparison with the nonlinearity function. The duration of this pulse is between 0ns and 800ns. In chapter 4, the neuron unit will be discussed in detail.

3. The synapse unit

This chapter deals with the synapse unit in detail. The basic synapse unit which is used to develop the final synapse unit, is earlier discussed in [15].

3.1. Introduction

In figure 3-1, a model of the synapse is presented.

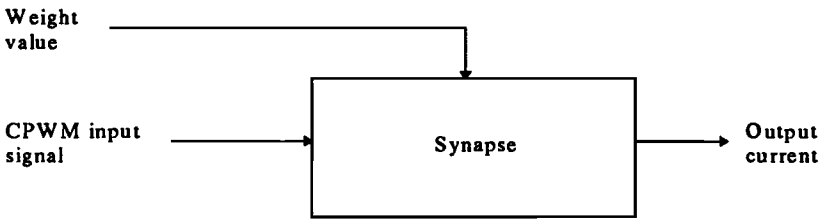


Figure 3-1 Model of a synapse

The input of the synapse consists of a CPWM signal, which is actually the output of a neuron. Together with this signal, a weighting factor is added to the synapse. The output of the synapse consists of a current. The relation-ship between the output of the synapse and its two inputs is as follows: the CPWM input signal determines a period of time in which the synapse is active (see next paragraph), and the weigthing factor determines the direction of the output current (positive or negative) and the magnitude of this current. To obtain a multiplicative relationship between those two inputs, the output current is used to charge or discharge a capacitor. This capacitor is part of the neuron unit and will be discussed in the next chapter. Because of the multiplicative relationship between the two inputs, we can treat the synapse unit as a CPWM multiplier, and in this case, as a four-quadrant CPWM multiplier (see next paragraph).

A very important issue to take in consideration in designing a synapse unit, is its power dissipation. The complete neural network contains (just as in a biological network) a large amount of synapses. So these synapses take a significant part in the total power dissipation. Thus, to realise an overall power dissipation reduction, one can reduce the power dissipation of the synapse unit. Another important aspect is the linearity of the synapse (multiplier) in the relationship between the output current and the input pulse width. This linearity has its origin in the linearity of the transfer characteristic with respect to the input continuous signal, if this signal is only used for switching (PDM, PPM and PFM techniques).

3.2. Four-quadrant multiplier

The basic four-quadrant multiplier design which was used as a starting point, is shown in figure 3-2. It consists of a voltage-to-current converter (VCC) , corresponding with the transistors M1 to M4, four switches (M7-M10) which switch the weight voltage (V_{in}-V_{ref}), a switch M6 to switch the power supply of the VCC, and a current mirror M5a-M5. The supply voltage Vdd is 5.0V, and the used process is the MIETEC 2.4 µm N-well CMOS process. The weight is presented by the voltage difference V_{in} - V_{ref}. In this case V_{ref} = 1.1V and V_{in} shifts between 0.1V and 2.1V, so the weight can vary between -1 and +1. The four switches M7-M10 (see figure 3-3) are controlled by the input pulse X_i (a pulse with a variable duration). From this point on, the index i, in combination with X_i and T_i, stands for a certain input signal respectively input pulse width of a synapse. We can distinguish two states of the input pulse; a high and a low state. When the input pulse is high (X_i = 5V) then M8 and M9 are conducting, and M7 and M10 are off. V_{ref} is transferred to V_1 and V_{in} to V_2. If the input pulse is low (X_i = 0V) then M7 and M10 are conducting and M8 and M9 are off. V_{ref} is now transfered to V_2 and V_{in} to V_1. Dependent on the input pulse X_i, the voltage V_{ref}

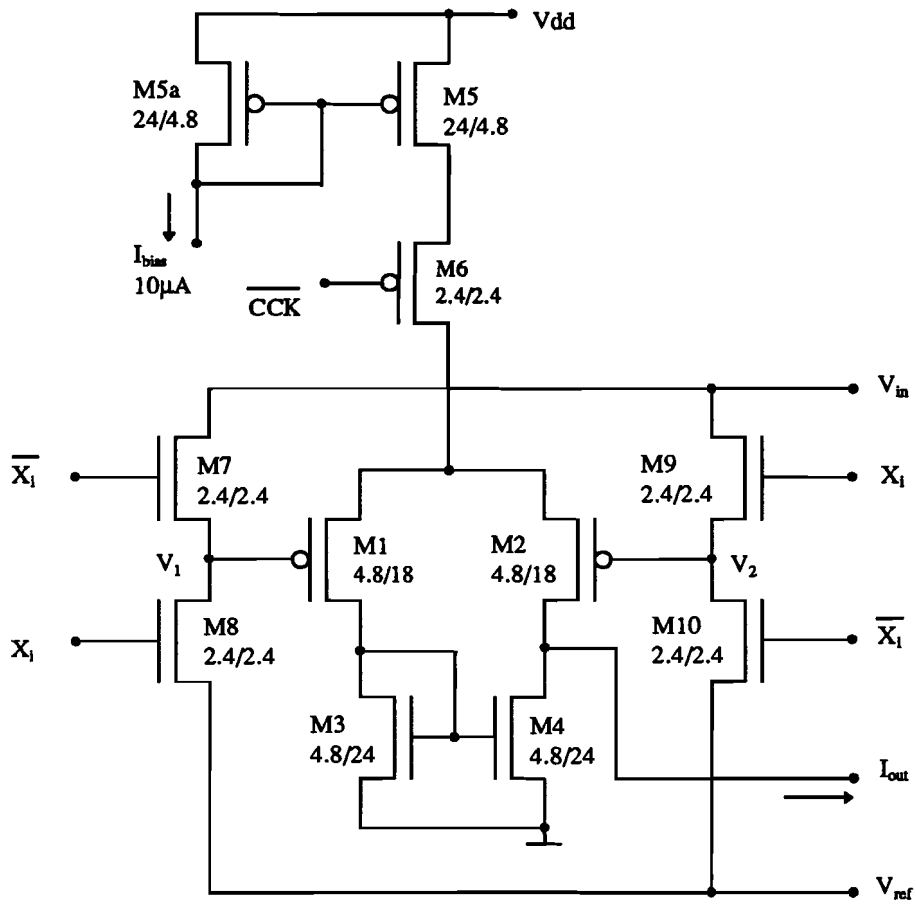


Figure 3-2 Four-quadrant multiplier

or V_{in} is offered to V_1 or V_2, so the weight voltage V_{in} - V_{ref} is switched between the inputs of the VCC by the CPWM input signal X_i. In the active phase of the synapse (X_i is high), the input of the VCC is V_1-V_2 = V_{ref} - V_{in}, and in the non-active phase (X_i is low) V_1 - V_2 = V_{in} - V_{ref}. In figure 3-4 a timing diagram is given of the synapse. In this timing diagram, an input pulse of 600ns is taken and V_{in} = 0.6V. The weight is therefore 0.5.

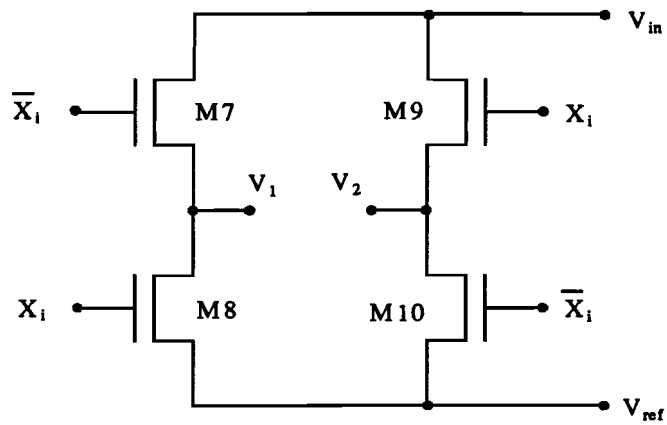


Figure 3-3 Switch controle

When the power supply is switched to the VCC (CCK is high), and the synapse receives an input pulse, then the voltage difference V_d of the VCC in the active phase of the synapse, is $V_d = V_{ref} - V_{in}$, which results in a positive current I_{out} In the non-active phase of the synapse, the voltage difference $V_d = V_{in} - V_{ref}$, and this results in a negative output current. The VCC has no power supply when CCK is low, and therefore no current can flow through the VCC and there is no output current. It is clear that, when the synapse receives an input pulse with a duration of $T_{max}/2$, that the period of time in which the output current is positive (X_i high), equals the period of time in which the output current is negative (X_i low), and therefore this input pulse represents a zero input value.

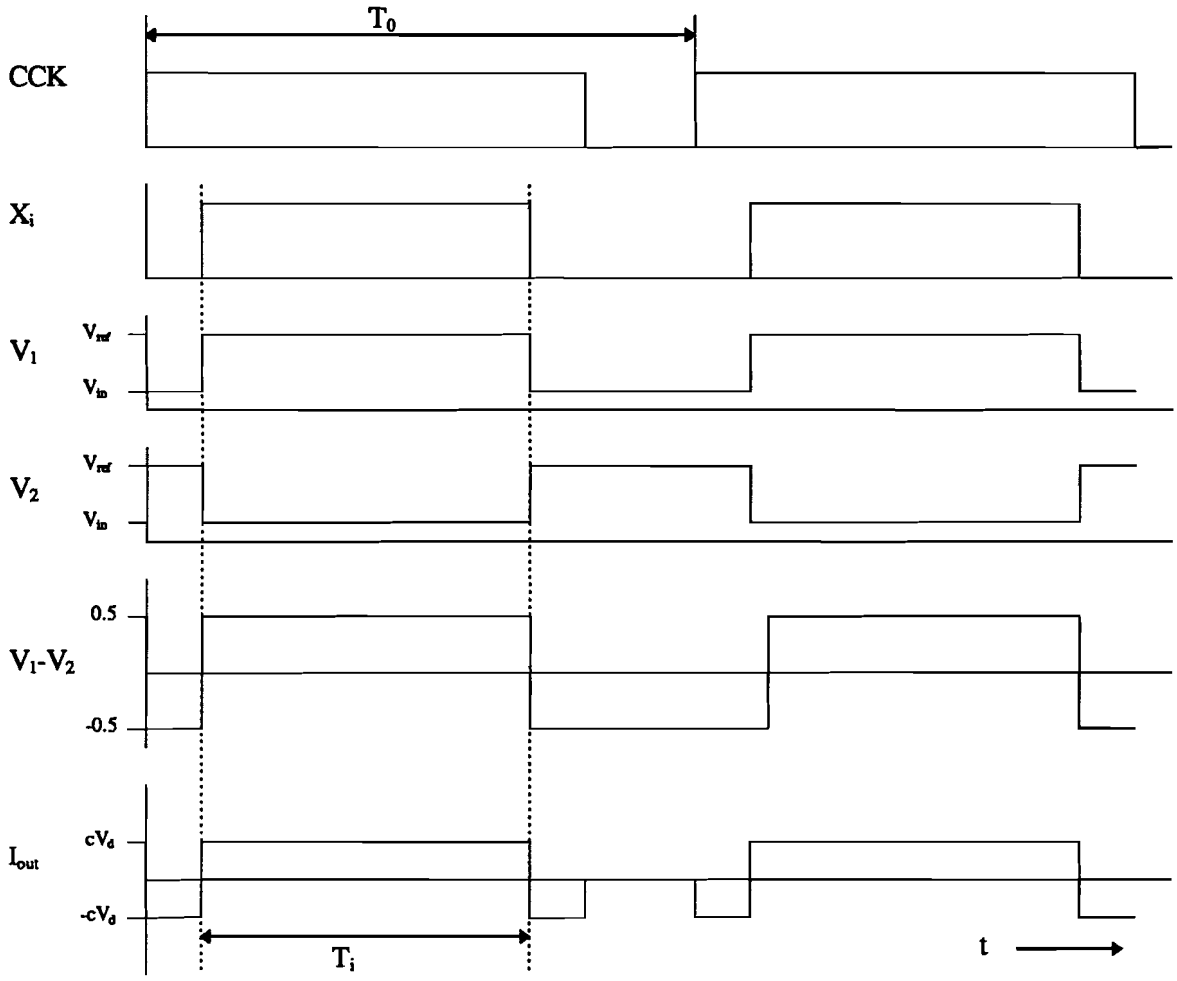


Figure 3-4 Timing diagram of the synapse; $V_{in} = 0.6V$, $V_{ref} = 1.1V$

The VCC is actually a differential amplifier (see figure 3-5). The transistors M1 to M4 operate in their saturation region. The following (simplified) relation exists between the current through the transistor and its gate-source voltage V_{gs} [1]:

$$I_{d_i} = \frac{\beta_i}{2}\left(-V_{gs_i} + V_{T_i}\right)^2 \qquad \text{PMOS} \qquad (3.1a)$$

$$I_{d_i} = \frac{\beta_i}{2}\left(V_{gs_i} - V_{T_i}\right)^2 \qquad \text{NMOS} \qquad (3.1b)$$

with:

$$\beta_i = \mu_i C_{ox} \frac{W_i}{L_i} = k_i \frac{W_i}{L_i} \tag{3.2}$$

The index i stands for a certain transistor Mi. The voltage difference between the inputs of the VCC is :

$$V_d = V_1 - V_2 \tag{3.3}$$

Furthermore:

$$I_o = I_{bias} = I_1 + I_2 \tag{3.4}$$

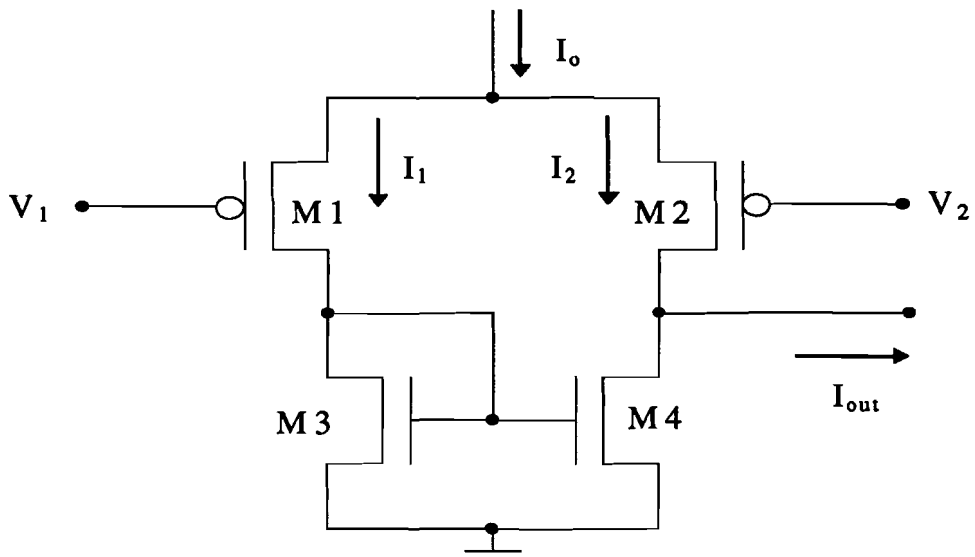


Fig 3-5 Voltage to Current Converter (VCC)

Because transistors M1 and M2 are identical (size and type), one can write:

$$\beta_1 = \beta_2 = \beta \qquad \text{and} \qquad V_{T_1} = V_{T_2} = V_T \tag{3.5}$$

With this and with (3.1a), (3.3) can be rewriten in:

$$V_d = V_1 - V_2 = \sqrt{2\frac{I_2}{\beta}} - \sqrt{2\frac{I_1}{\beta}} \tag{3.6}$$

Together with (3.4), one can write for the currents through M1 and M2:

$$I_1 = \frac{I_o}{2} - \frac{I_o}{2}\sqrt{\frac{\beta V_d^2}{I_o} - \frac{\beta^2 V_d^4}{4I_o^2}} \tag{3.7a}$$

$$I_2 = \frac{I_o}{2} + \frac{I_o}{2}\sqrt{\frac{\beta V_d^2}{I_o} - \frac{\beta^2 V_d^4}{4I_o^2}} \tag{3.7b}$$

Now, the output current can be written as a function of the voltage difference V_d:

$$I_{out} = I_2 - I_1 = \frac{\beta}{2} V_d \sqrt{\frac{4I_o}{\beta} - V_d^2} \tag{3.8}$$

It is clear, that there is a linear relationship between the output current and the voltage difference, if the next requirement is met:

$$V_d \ll \sqrt{\frac{4I_o}{\beta}} \tag{3.9}$$

If so, then the output current (3.8) becomes:

$$I_{out} = V_d \sqrt{\beta I_o} \tag{3.10}$$

Thus, linearity can be accomplished for a wide range of V_d (and, naturally, a wide range of the weighting factor), for a large current I_0 and/or large lenghts of the transistors M1 and M2).

The current I_0 is determined by the current mirror created by the transistors M5 and M5a, and the bias current I_{bias}, which is externally defined. This is done to minimize the susceptibility of the circuit to parameter variations.

The switch M6 is added to the circuit to give the neuron (see next chapter) the opportunity to discharge the integration capacitor (in the idle phase). Furthermore the switch M6 will accomplish a power dissipation reduction. Because the total neural network works with an idle and active phase (see chapter 2), it's not necessary to let the synapse in action in the idle phase. So the synapse is shut down in the idle phase, done by M6. The output current I_{out} will be integrated over a capacitor C_{int} in the neuron unit. To obtain a simulation of the synapse unit, this part of the neuron will be connected to the synapse unit, and then used for simulations. Between the synapse and the neuron, a switch is connected which is off during the idle phase, but this will also be discussed in the next chapter. The model of an integrator in figure 3-6 can be used for calculation purpose. The voltage of V_{plus}, which is used in the simulations, is 2V.

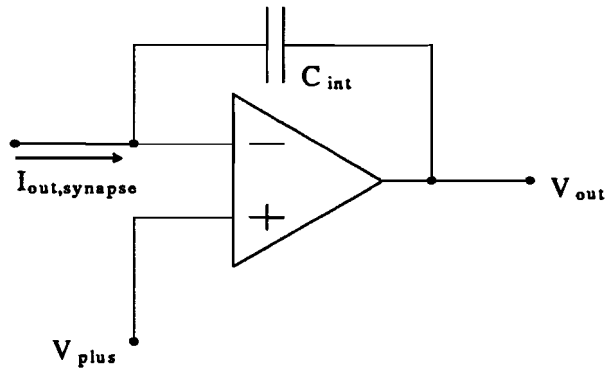


Figure 3-6 Model of an integrator

For the two inputs of the amplifier, one can write:

$$V_- = V_+ = V_{plus} \tag{3.11}$$

Furthermore, the following equation can be written for the integrator:

$$V_{plus} - V_{out} = \frac{1}{C_{int}} \int_0^t I_{out} dt + V_{C_{int}}(t=0) \tag{3.12}$$

Because the capacitor C_{int} will be discharged every period in the idle phase, so the voltage over the capacitor will be zero each time at the beginning of a period. (3.12) becomes now:

$$V_{out} = V_{plus} - \frac{1}{C_{int}} \int_0^t I_{out} dt \tag{3.13}$$

To determine the time dependence of the integration, it must determined when there is an output current I_{out}. Therefore a closer look is been taken at the timing diagram of the four-quadrant multiplier. In figure 3-7, the timing is given in the case of $V_{ref}>V_{in}$ and with an input pulse width>400ns. This is done to match the sign of the output current with (3.13). Later on, this will be further explained.

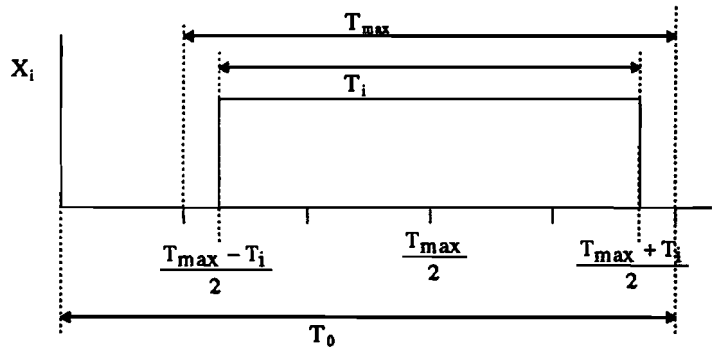


Figure 3-7 Timing diagram of the synapse in detail

From this figure, and together with (3.13), the multiplicative relationship between the output voltage V_{out} and the two inputs of the synapse (T_i and the weight w_i) can be derived. If (3.10) is used for I_{out}, it can be determined that

$$V_d = V_1 - V_2 = \begin{cases} V_{ref} - V_{in} & X_i \text{ high} \\ V_{in} - V_{ref} & X_i \text{ low} \end{cases} \tag{3.14}$$

Which means that I_{out} has a positive sign, when X_i is high ($V_d = V_{ref} - V_{in}$) and a negative sign when X_i is low ($V_d = V_{in} - V_{ref}$). I_{out} can now be written as:

$$I_{out}(t) = \begin{cases} -V_d\sqrt{\beta I_0} & 0 \le t < \dfrac{T_{max} - T_i}{2} \\ +V_d\sqrt{\beta I_0} & \dfrac{T_{max} - T_i}{2} \le t \le \dfrac{T_{max} + T_i}{2} \\ -V_d\sqrt{\beta I_0} & \dfrac{T_{max} + T_i}{2} < t \le T_{max} \end{cases} \tag{3.15}$$

Now (3.13) can be rewritten as:

$$V_{out}(T_0) = V_{plus} - \frac{1}{C_{int}} \int_0^{T_0} I_{out}\,dt = V_{plus} - \frac{V_d\sqrt{\beta I_0}}{C_{int}}(2T_i - T_{max}) \tag{3.16}$$

Voltage difference V_d can be rewritten in such a manner that it is a function of the maximum voltage difference (V_1 - V_2) and a factor w_i, wich determines the magnitude of V_1-V_2:

$$-(V_1 - V_2)_{max} \le V_d \le +(V_1 - V_2)_{max} \tag{3.17a}$$

$$V_d = w_i(V_1 - V_2)_{max} \qquad \text{with} \qquad -1 \le w_i \le +1 \tag{3.17b}$$

So actually the voltage V_{in} is now replaced by a factor w_i (see formula 3.17). Also T_i can be rewritten in such a manner, that T_i is a function of T_{max} and a factor α_i (see formula (2.2)). If formula (2.2) is used to rewrite T_i, and (3.17) to rewrite V_d, then the output voltage has a multiplicative relationship with the two inputs of the synapse :

$$V_{out} - V_{plus} = -w_i\alpha_i \frac{(V_1 - V_2)_{max} T_{max}\sqrt{\beta I_0}}{C_{int}} = -w_i\alpha_i p \tag{3.18a}$$

$$\text{with} \quad p = \frac{(V_1 - V_2)_{max} T_{max} \sqrt{\beta I_0}}{C_{int}} \quad \text{and} \quad -1 \leq \alpha_i \leq +1 \qquad (3.18b)$$

This relationship is also a linear one, so linearity is maintained. We can now explain why the synapse is also called a four-quadrant multiplier. In figure 3-8, a diagram is given with four quadrants. In each quadrant, the requirement is given for the input values of the synapse.

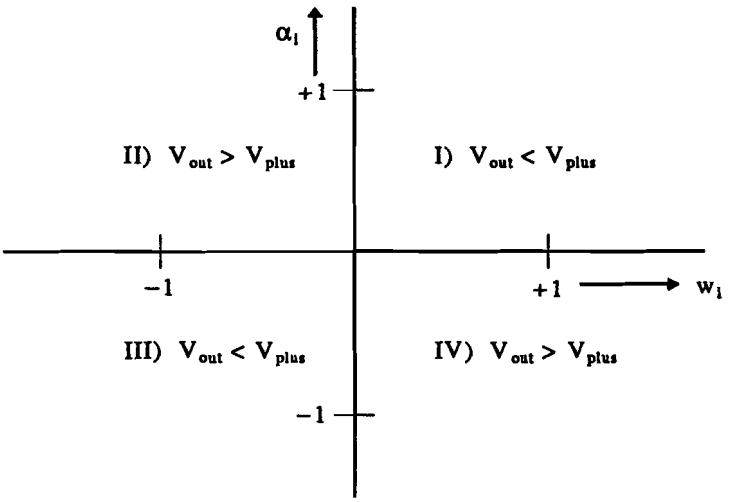


Figure 3-8 The four quadrants of the synapse

For example, if the weight $w_i < 0$ (then $V_d < 0$) and $\alpha_i > 0$ (input pulse width > 400ns), then the output voltage $V_{out} > V_{plus}$, and this is in quadrant II.

Now the working of the synapse is explained, its simulation results can be described. In figure 3-9, the setup for the simulation is given. A sample & hold circuit is connected to the integrator, because the output voltage must be held for further processing of the input signals (this will be further explained in chapter 4). The switch, integrator and the sample & hold circuit are a part of the neuron unit and are here only used for simulation purposes; they will be discussed in more detail in the next chapter. In figure 3-10 the simulation results are shown; in figure 3-10a the simulation of the output voltage versus the weight (with several input pulse widths), and in figure 3-10b the simulation of the output voltage versus the input pulse width (for several weight values) are depicted.

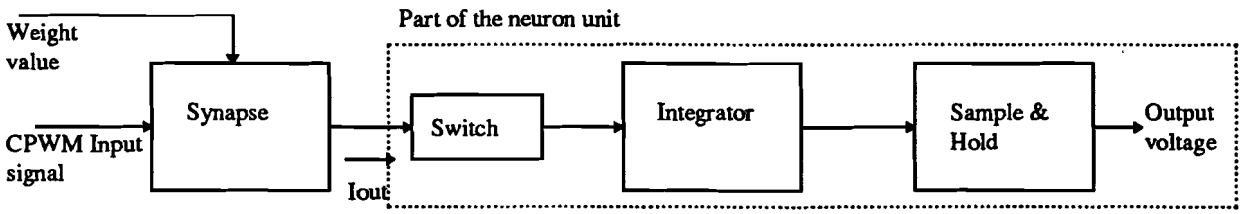
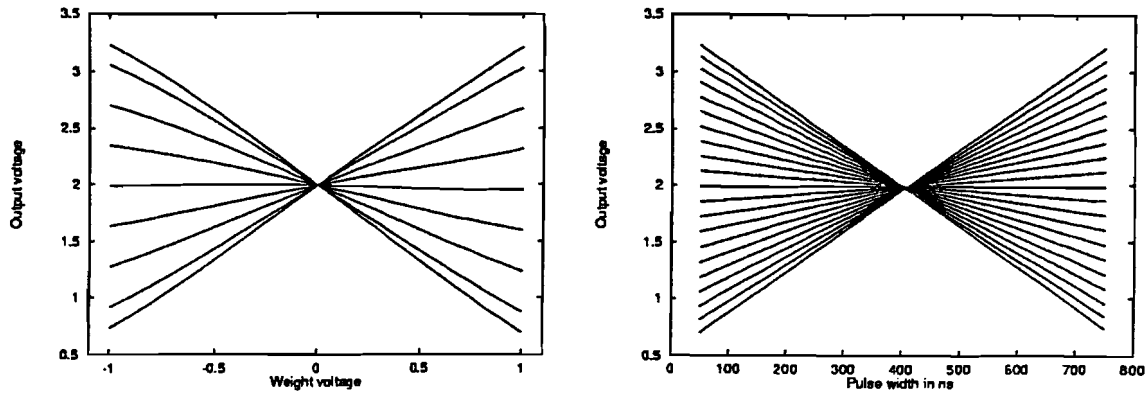


Figure 3-9 Simulation setup for the synapse

The simulation of the synapse unit shows formidable linearity and furthermore, the output voltage is equal to the V_{plus} voltage (2V) at two occasions: when there is an input pulse width of 400ns (figure 3-10b), and when there is a weight value of 0V (figure 3-10a).

The power dissipation is not dependent on the CPWM input signal, but is constant during the active phase. In the idle phase, the synapse is shut down and therefore the power dissipation in the idle phase is zero. In the active phase the power dissipation is $P = I_{bias} \cdot Vdd = (10\mu A)5V = 50\mu W$. The average power dissipation (or total power dissipation) is then $P_{tot} = (0.2)0\mu W + (0.8)50\mu W = 40\mu W$. These results are already discussed in [15].



a) Output voltage versus weight value

b) Output voltage versus input pulse width

Figure 3-10 CPWM characteristics of the synapse unit for the 2.4μm process

The forementioned simulation results were obtained for the 'old' process (MIETEC 2.4μm NWELL, $W_{min} = L_{min} = 2.4\mu m$, Vdd = 5V). The purpose of the work described in this thesis however, was a circuit design in the 'new' process (c05md and c05ma, $W_{min} = 0.8\mu m$, $L_{min} = 0.5\mu m$ and Vdd = 3.3V). As a first attempt, direct rescaling of all transistors sizes by a factor of 0.8/2.4 was tried. Furthermore, to accomplish a power dissipation reduction, the bias current of the multiplier was reduced from 10μA to 5μA. However, the new circuit does not meet the requirement of linearity in simulations, and the current mirror (M3 and M4) has a small error. Therefore, some adaptions were made. The first one is improve the current mirror (M3 and M4). The slight error in the current mirroring is a result of the output-resistance effect [1]. The higher the value of the output resistance, the lower the error will be. So the output resistance needs to be increased. This is done by replacing the current mirror in a Wilson current mirror [1] (see figure 3-11).

The Wilson current mirror uses the principle that an output resistance can be increased through the use of negative, current feedback [2].

The second improvement was an optimization of all individual transistor sizes, in such a way that the result of the simulations were satisfactory. The final circuit, with the dimensions of the transistors, is shown in figure 3-11. V_{ref} has a voltage of 1.2V and V_{in} is variable between 0.2V and 2.2V, so the weight value can again vary between -1 and +1 V.The simulations are shown in figure 3-12. In figure 3-12a the simulation of

the output voltage versus the weight (for several input pulse widths), and in figure 3-12b the simulation of the output voltage versus the input pulse width (for several weight values) are depicted.

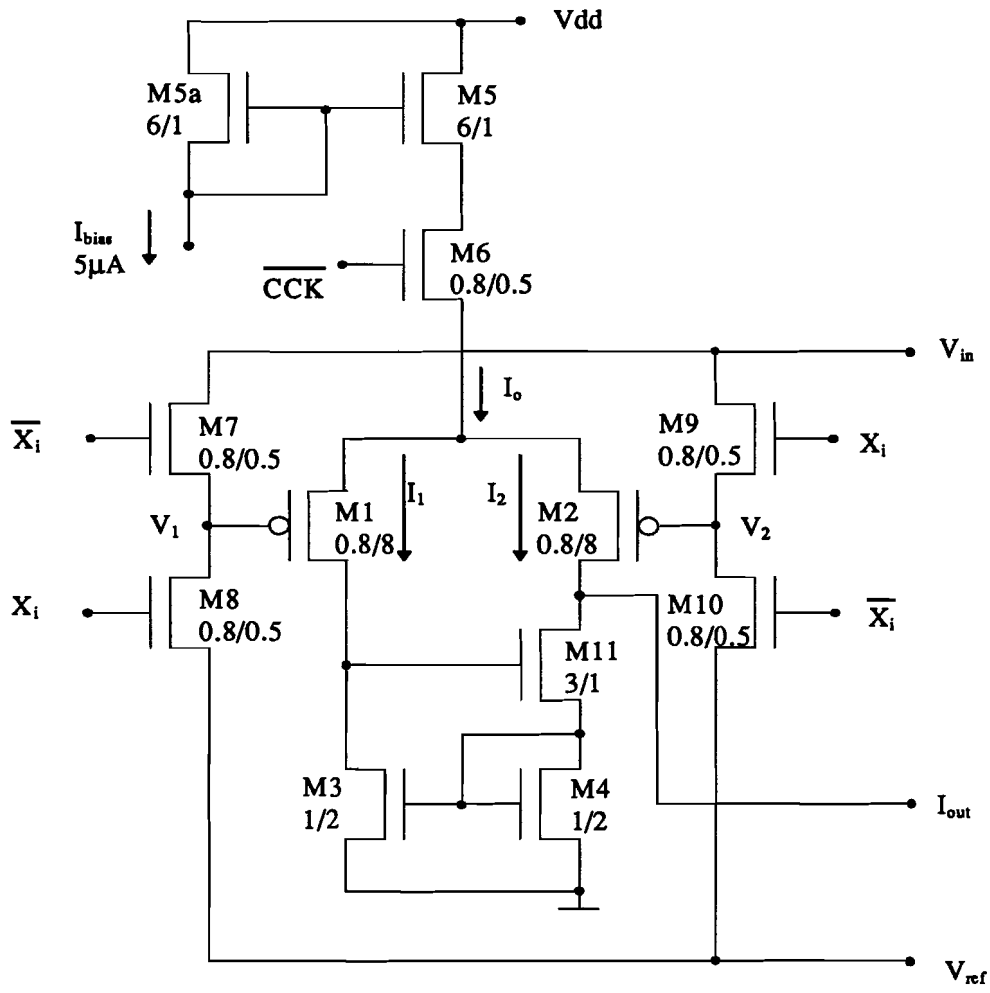
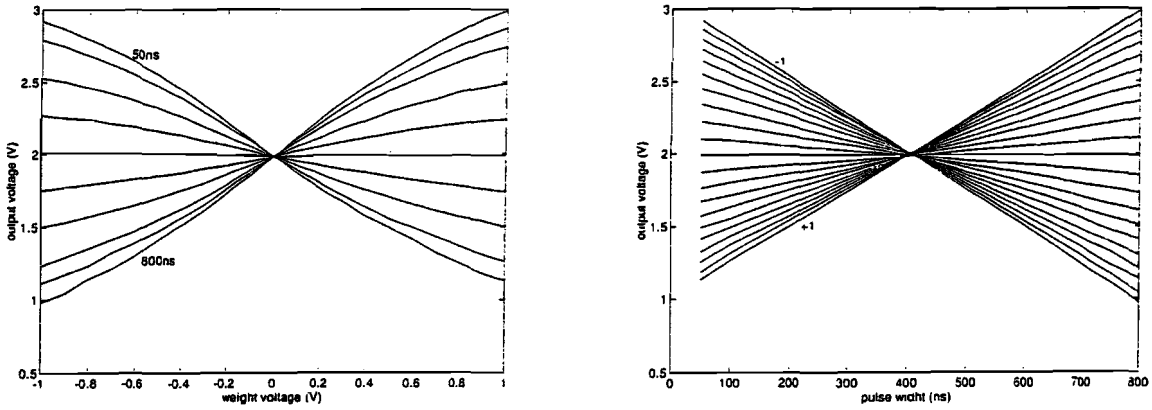


Figure 3-11 New four-quadrant multiplier/synapse unit

In the 'new' process design, it can be seen that the linearity between the output voltage V_{out} and the (variable) input pulse width of the synapse unit is excellent (almost perfect). The linearity in the case of variable weight is less excellent (even compared with the 'old' process simulation results), but this is not so important, because this means only that the weight value is slight different than it would be in a perfect linear multiplier, but it has no effect on the total network behaviour. This non-linearity is due to the fact that in the case of a large weight (large V_d), the specification, formed by formula (3.9), not is met, and therefore formula (3.8) must be used in stead of formula (3.10). Furthermore, the weight range is maintained while the supply voltage is reduced from 5V to 3.3V. The power dissipation of the final circuit is considerable lower then the power consuption of the 'old' version. The total power dissipation is now P_{tot} = 0.2·0µW + 0.8·I_{bias}·Vdd = 0.8·5µA·3.3V = 13.2µW. This means a reduction by 67%. Also, the new circuit is much smaller than the old circuit, in spite of the addition of one transistor in the Wilson current mirror, and it remains its simplicity.



a) Output voltage versus weight value

b) Output voltage versus input pulse width

Figure 3-12 CPWM characteristics of the final synapse unit

4. The neuron unit

This chapter deals with the neuron unit. One can distinguish four different parts of the neuron unit: an integrator, a sample & hold circuit, a non-linear circuit and a comparator. These four parts will be discussed in detail in this chapter.

4.1. Introduction

A model of the neuron unit is given in figure 4-1. In chapter 3, it was shown that the output current of a synapse is converted into a voltage. This was done by an integrator and a sample & hold circuit. However, chapter 3 dealt with a situation with only one synapse connected to the neuron (the integrator part). If more synapses are connected to the neuron, then the output voltage of the integrator will exceed certain bounderies of operation (maximum and minimum output voltage) when the capacitor Cint is not adjusted to the amount of synapses. So scaling of this capacitor is absolutely necessary.

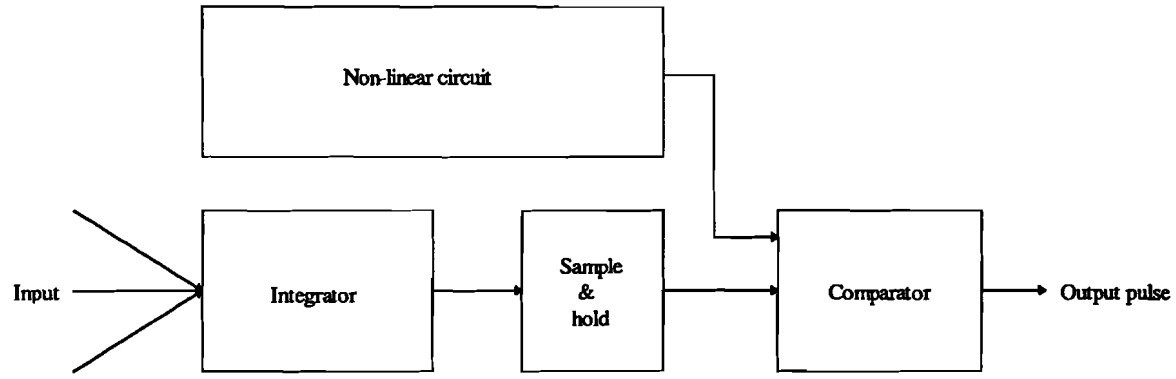


Figure 4-1 Model of the neuron unit

The non-linear circuit is embedded into the neuron unit to simulate a certain saturation of the neuron's respons. The comparator will compare the output voltage of the sample & hold circuit with the output voltage of the non-linear circuit. The output of the neuron unit will be a pulse with a duration dependent on that comparison. So the width of the output pulse is dependent on the output voltage of the sample & hold circuit.

4.2. The integrator/sample & hold circuit

A model of the integrator and the sample & hold circuit is given in figure 4-2, together with switches to ensure appropriate behaviour of the circuit. To explain the working of the switches, the timing of the integrator/sample & hold circuit will be described in more detail. In figure 4-3 the timing is given of the switches, together with the system clock CCK. In the active phase of the synapse (or synapses), switch S1 is closed ($V_{connect}$ is High) and switches S2 and S3 are open ($V_{discharge}$ and V_{hold} are Low). The current I_{in} charges capacitor C_{int}. At the end of the active phase, S1 is opened ($V_{connect}$ is Low) and the voltage across the capacitor represents now the sum of weighted inputs.
In the idle phase, the output voltage of the integrator is stored by the sample & hold circuit by closing S3 (V_{hold} is High). Before discharging the capacitor C_{int}, S3 is being opened to keep $V_{S\&H}$ unaffected by the discharging of C_{int}. After S3 is opened, C_{int} can be discharged by closing S2 ($V_{discharge}$ is High). Before the

active phase begins again, S2 has to be opened again. Then the integrator/sample & hold circuit is ready for the next period T_0.

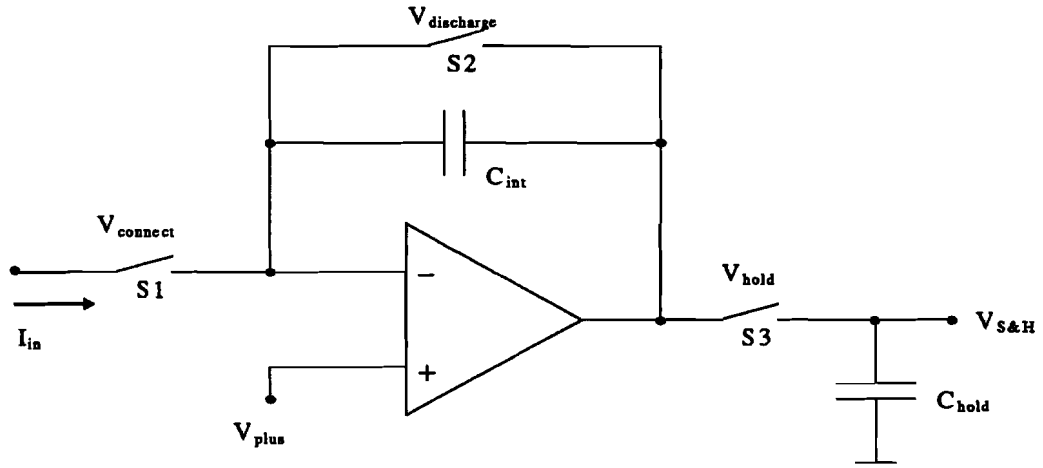


Figure 4-2 Model of the integrator/sample & hold circuit

It can be seen that the output pulse of the neuron unit lags actually one period behind the real-time period of the input. The circuit of the integrator/sample & hold, is given in figure 4-4. This circuit was designed in the 'old' process (see chapter 3).

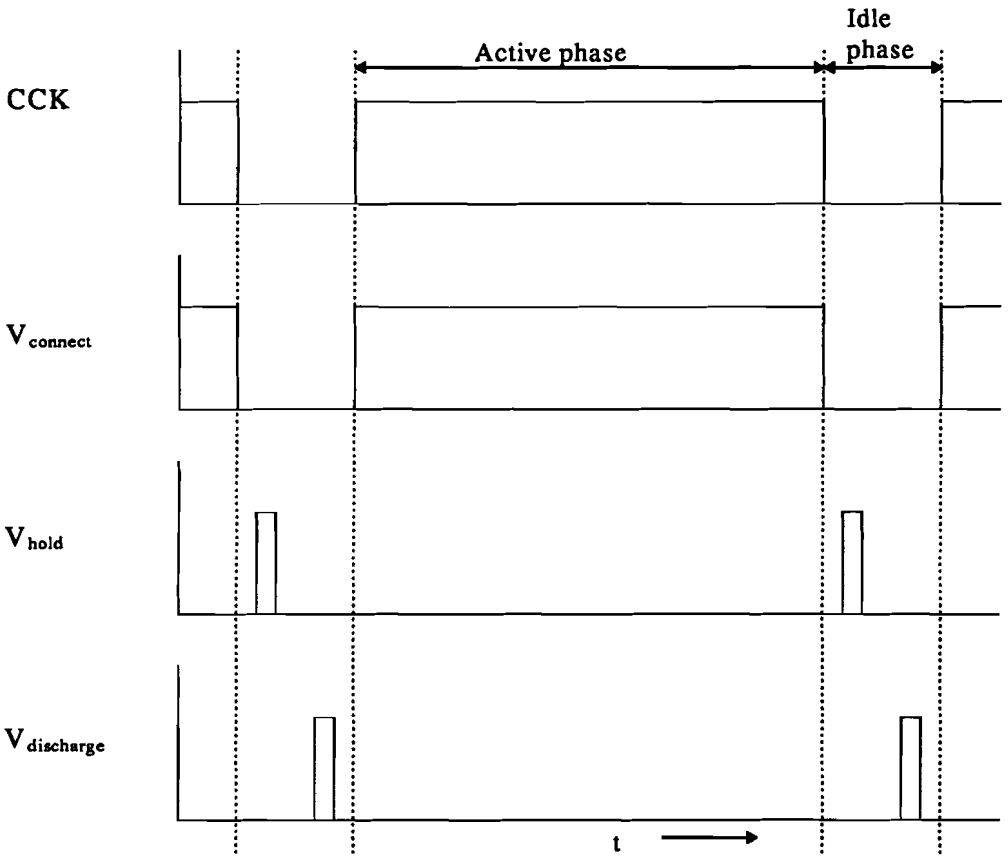


Figure 4-3 Timing diagram of the integrator/sample & hold switches

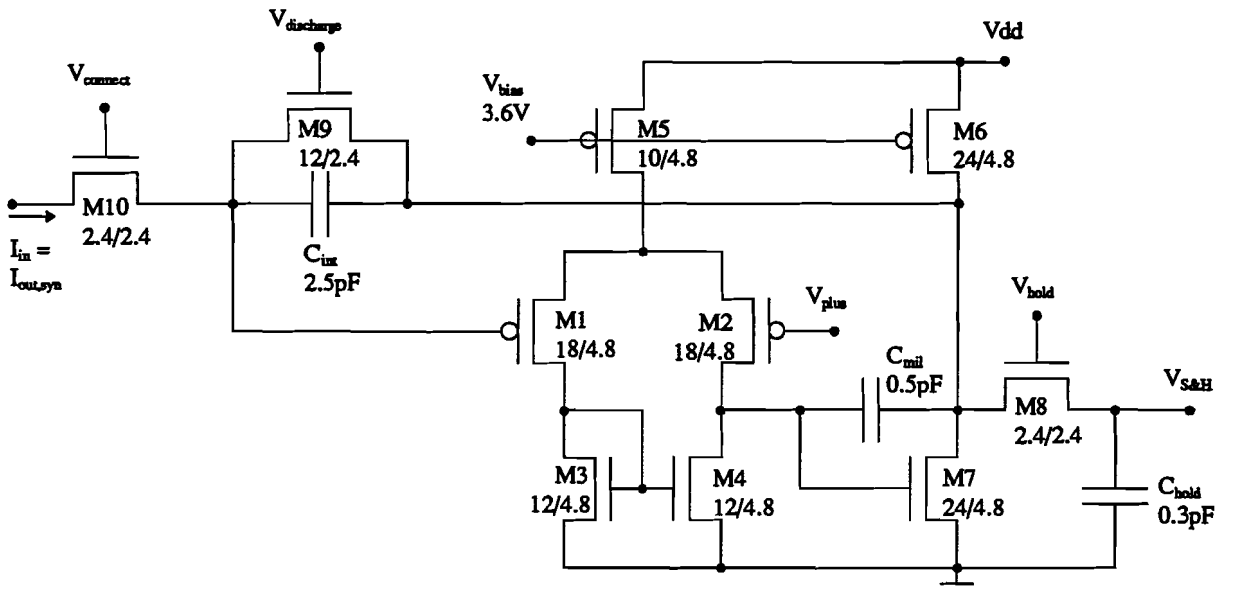


Figure 4-4 Design circuit of the integrator/sample & hold at starting point

Transistor M8 represents switch S3, and is on when V_{hold} is high. When V_{hold} is low, transistor M8 is off, and the voltage over C_{hold} is held. Transistor M9 represents switch S2, and when $V_{discharge}$ is high, the transistor is on, and the capacitor C_{int} will be discharged, so the voltage across C_{int} is zero. When $V_{discharge}$ is low and $V_{connect}$ is high, then a current charges the capacitor C_{int}. Transistor M10 represents switch S1, which connects the synapse(s) with the neuron unit.

C_{mil} represents the Miller capacitor, which is needed to avoid oscillations in the circuit. These oscillations are a side-effect when using the operational amplifier with a feedback.
The current through the differential amplifier is approximately 4.5μA, and through the output stage approximately 12μA. The simulation of this circuit is already described in the former chapter, and the results were satisfactory.

To redesign the integrator/sample & hold circuit for the new process, the first step is (just as it was with the synapse) to rescale the transistors. The result is that all transistor are a factor 3 smaller than before. For example, transistor M1 is now 6/1.6. Also take notice of the fact that k (see formula 3.2) of the NMOS transistor is not the same as the k of the PMOS transistor. For both transistors the values are:

$$k_p = 4.37\ 10^{-5}\ F/Vs$$
$$k_n = 17.35\ 10^{-5}\ F/Vs \qquad (4.1)$$

It can be seen that k_n is approximately four times larger than k_p. For example, if the same current must flow through a NMOS and a PMOS transistor with having the same node voltages, then (W/L) of the PMOS transistor must be four times larger than the (W/L) of the NMOS transistor. In the output stage of the operational amplifier, the (W/L) of transistor M6 is four times larger than the (W/L) of transistor M7, for symmetry purposes only.

The seconds step is to replace the current source M5, together with a bias voltage, into a current source M5, with a current mirror M5a. The use of the current mirror is to ensure that the same current flows through the circuit as it is specified by the bias current, in spite of parameter variation. The bias current is set to 4.5μA. Because the current $I_{out,syn}$ is different with respect to the old process circuit, the integrator capacitor C_{int} must also be adjusted.
The results of the simulation of the integrator/sample & hold circuit were not satisfactory. The output voltage will not reach the desired maximum output voltage of 3V. This is due to the fact that the saturation voltage of transistor M6 is larger than the maximum available voltage of 3.3V-3.0V=0.3V. So, to maintain an output voltage swing from a maximum of 3V to a minimum of 1V (with centre V_{plus} = 2V), is to replace the p-channel input operational amplifier by a n-channel input operational amplifier. The PMOS transitors will be replaced by NMOS transistors and vica versa (not the switches). To do so, care must be taken that the (W/L) of a NMOS transistor is four times larger if the transistor is replaced by a PMOS transistor, and the (W/L) of a PMOS transistor is four times smaller if it is replaced by a NMOS transistor. To minimize the Miller capacitor C_{mil} and the current through the output stage, a final adjustment has been made in the sizing of M6 and M7. Furthermore, switches M8 and M9 are replaced by a PMOS transistor for better results in switching. The voltage V_{hold} and $V_{discharge}$ are to be reversed as a result of the transistor replacement. The final timing diagram and circuitry are given in figure 4-5 and 4-6. From this diagram, it is clear that CCK and $V_{connect}$ can be connected to each other, so no new input signal is needed.

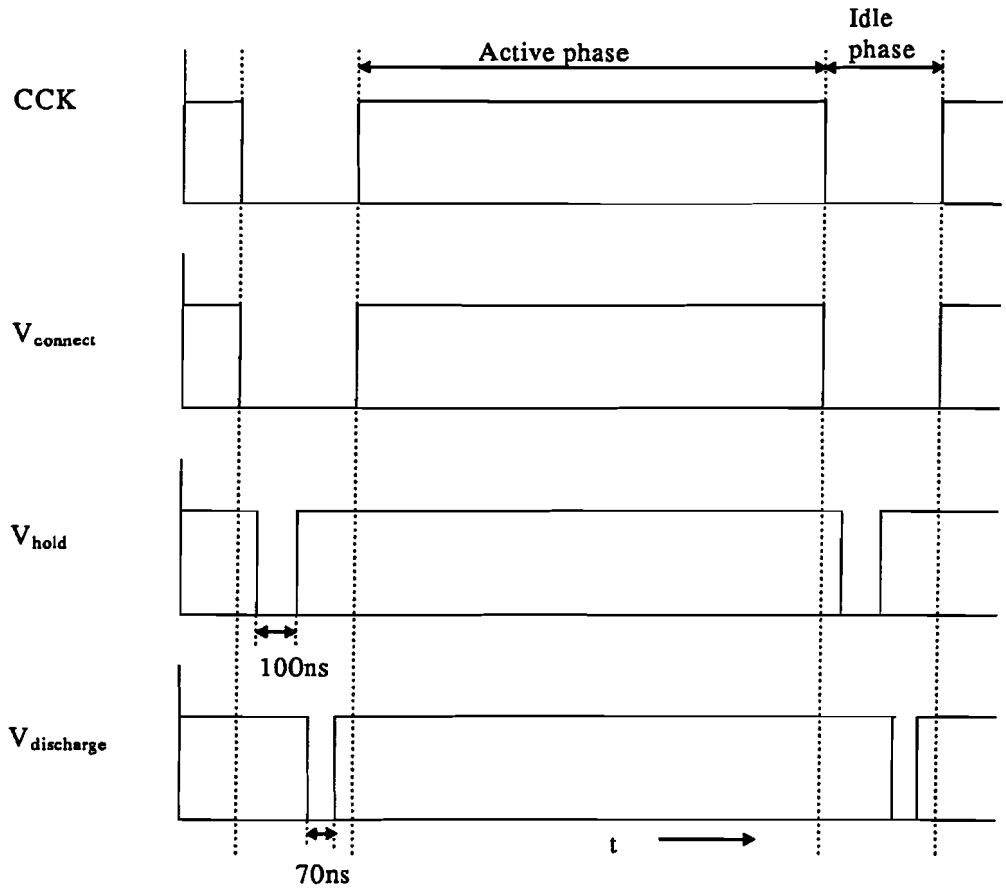


Figure 4-5 Final timing diagram of the integrator/sample & hold circuit

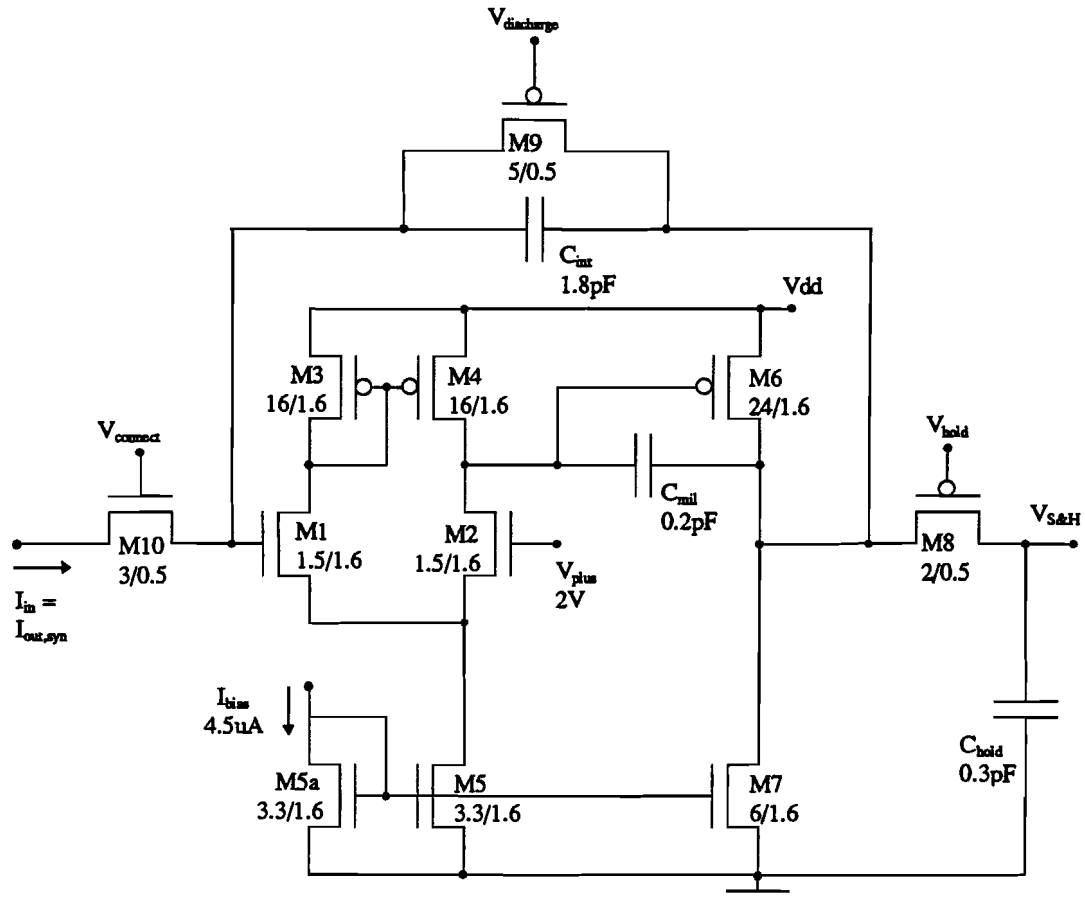


Figure 4-6 Final circuit of the integrator/sample & hold

The current through the input stage of the final circuit is 4.5μA, and the current through the output stage is approximately 8.2μA. The total power dissipation of the integrator/sample & hold circuit is therefore approximately 42μW (in the first design 83μW). Also the integrator capacitor C_{int} and the Miller capacitor C_{mil} are smaller, so less area is needed. The simulation of this circuit, together with the synapse, was already described in chapter 3 and the results came up to the expectations.

When more synapses are connected to the neuron unit, it is necessary to adapt the integration capacitor to the amount of synapses. In earlier works, the relation between the amount of synapses and the size of the integration has been established [3]. Let the initial capacitor of the integrator (in the case of one synapse) be C_{init}, then the relation between the amount of synapses (N) and the integration capacitor is:

$$C_{int} = C_{init}\sqrt{N} \tag{4.2}$$

In this case, the value of C_{init} = 1.8pF (see figure 4-6). For minimal use of area space, the integration capacitor will be realised as several capacitors in parallel. Every capacitor has a switch to control the connection of the capacitor in the parallel network. The values of the capacitors are related in a 1:2:4:8 row. To determine the value of the capacitor needed, it is necessary to introduce a function CEIL(x). This function (ceiling) rounds x up to the nearest higher integer value. If the fraction of the square root of N is larger than zero, then the function CEIL($\sqrt{N}$) converts $\sqrt{N}$ up to the nearest higher integer value. For example, if N=5 then CEIL($\sqrt{N}$) = 3 and if N=9 then CEIL($\sqrt{N}$) = 3.

With the use of this function, the capacitor, which is needed when more synapses are used, can be determined in the following way:

$$C_{int} = C_{init}\,CEIL\left(\sqrt{N}\right) \tag{4.3}$$

Because the capacitors are related in a 1:2:4:8 row (binairy with $2^0:2^1:2^2:2^3$), C_{int} can be realised by switching the capacitors in the parallel network. For example, if N=7 then C_{int} = 5.4pF and this is a combination of 1.8pF parallel to 3.6pF. In figure 4-7 the parallel network of this example is given.

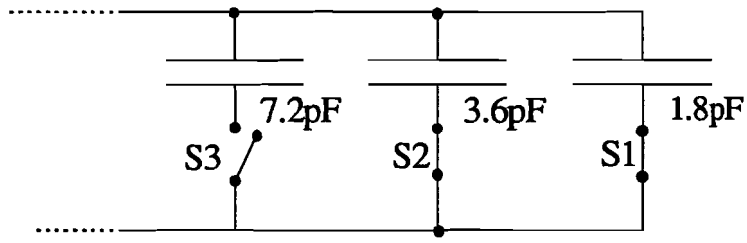


Fig 4-7 Parallel network with N=7

In the case of 7 synapses connected to the neuron unit, only two capacitors are needed to adjust the integrator capacitor. In this way, a flexibel neuron can be realised when it is not certain how many synapses are connected to it.

4.3. The nonlinear function

The output voltage of the neuron/sample & hold circuit will be compared with a non-linear function. This non-linear function is to ensure that the neuron's response is bounded; there is a certain saturation of the neuron's respons. A sigmoide function is such a non-linear function which has been used in [3] and [11]. The mathematical expression for a sigmoid function is:

$$S(x) = \frac{S_0}{1 + e^{-T(x - a)}} \tag{4.4}$$

with 'T' representing a temperature factor (steepness) and 'a' representing an offset. In figure 4-8, the shape of a sigmoid function is given. In [3], it is proposed to use the inverse of the sigmoid as a non-linear function. Let the inverse sigmoid function $S(x)^{-1}$ be a voltage dependent on time, then a sigmoid mapping between the output voltage of the integrator/sample & hold circuit, and the output pulse duration can be obtained. In figure 4-9, the shape of the inverse sigmoid function is given, together with the maximum (V_{max}) and minimum (V_{min}) output voltage of the integrator/sample & hold circuit. Because of parameter variations, it is important that the shape of the inverse sigmoid can be changed (for example height, time span). The inverse sigmoid circuit is a part of the neuron unit, but it is enough to implement it only once for the whole neural network and connect the output line with other neuron units (the timing of all neuron units will be the same). This will not only save area, but power dissipation as well. But connecting the inverse

circuit with other neuron units will cause a larger load capacity. This must be taken in account in designing the inverse sigmoid circuit.

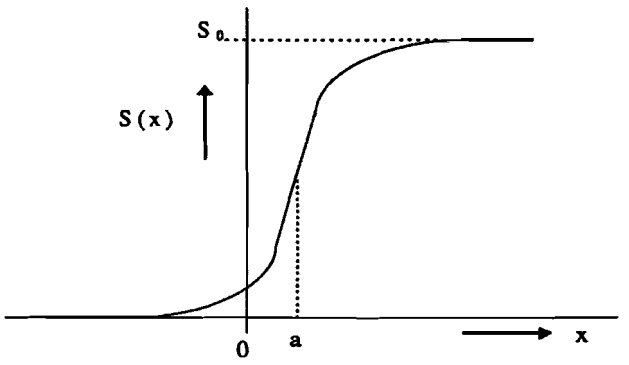


Fig 4-8 Shape of a sigmoid function

If the output voltage of the integrator/sample & hold circuit approaches the maximum (or minimum) voltage, then the output pulse duration stays almost the same, so a saturation relation exists between the output pulse duration and the voltage.

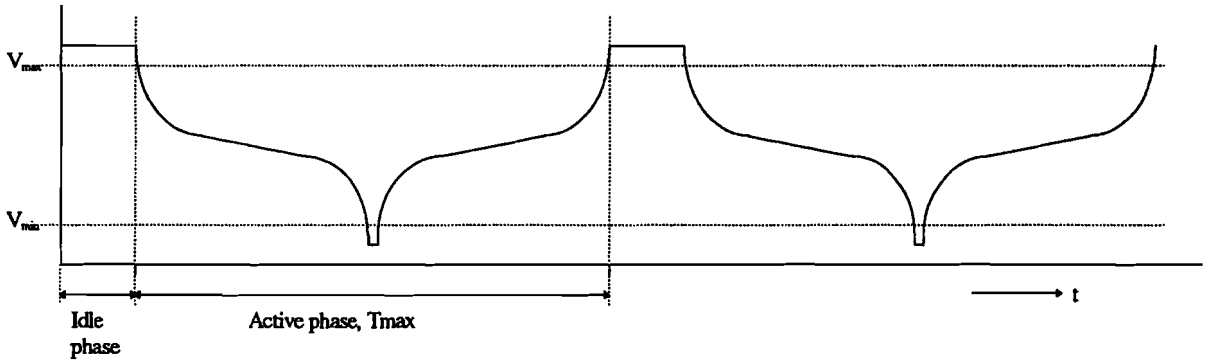


Fig 4-9 Non-linear function; the inverse sigmoid

4.3.1. Inverse sigmoid circuit; first approach

An inverse sigmoid signal can be obtained with the use of an operational amplifier. This principle is given in figure 4-10. A circuit S, which realizes a sigmoid characteristic, is placed between the output V_{out}, and the inverting input (V_) of the amplifier. A triangular signal is supplied to the non-inverting input of the amplifier (V+). In this figure, also the input and the output of the sigmoid circuit is given

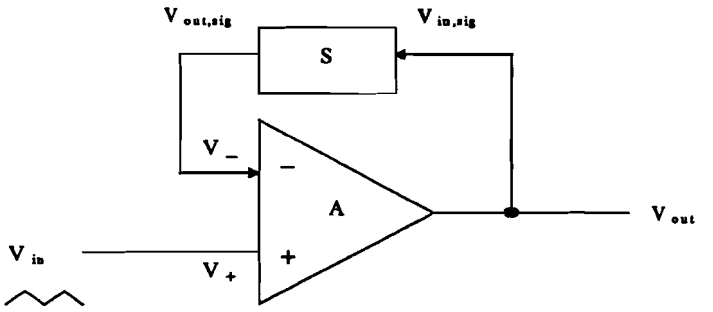


Fig 4-10 Principle of an inverse sigmoid circuit

Suppose the relation between the input and the output of the sigmoid function is as follows:

$$V_{out,sig} = SV_{in,sig} \tag{4.5a}$$

$$\text{and with} \begin{cases} V_{out,sig} = V_- \\ V_{in,sig} = V_{out} \end{cases} \tag{4.5b}$$

Then the relation (provided that the opamp is ideal) between the input and output of figure 4-10 is:

$$V_{out} = S^{-1}V_{in} \tag{4.6}$$

The output signal of the opamp of figure 4-10 has the characteristic of figure 4-9.
The realisation of the sigmoid circuit is already described in [3] and [11]. A simplified version of that sigmoid circuit is given in figure 4-11.

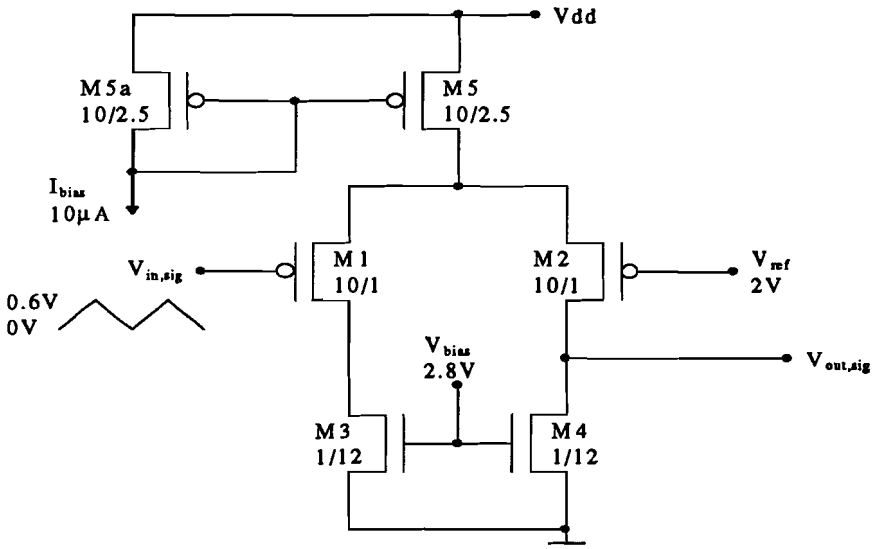


Fig 4-11 Simplified version of the sigmoid circuit

The principle of the sigmoid circuit is based on the shape of the currents through the differential input stage (see figure 4-12).

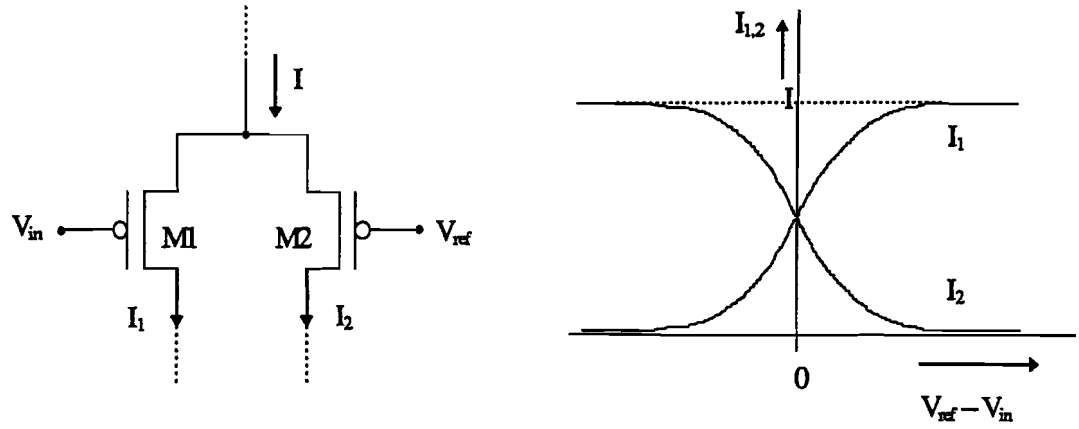


Fig 4-12 The differential input stage

The characteristic $I_{1,2} = f(V_{ref} - V_{in})$ is a sigmoid-like characteristic, and if these currents flow through resistors, then the voltage across these resistors also has a sigmoid shape. A resistor can be made of a CMOS transistor, operating in its linear region. In order to have a wide range of the output voltage with transistors M3 and M4 operating in their linear region, it is necessary to have a high V_{bias} voltage. This higher voltage will lead to a lower resistance value of the transistors and therefore a higher current I (and through that a higher I_{bias}) is needed. For maintaining the sigmoid width, it is also necessary to increase the widths of the differential input stage transistors. Another way to realize a larger resistor value is to increase the lengths of transistors M3 and M4, but this will lead to a higher drain-source voltage. This higher drain-source voltage causes transistors M3 and M4 operating earlier in their saturation region. The output characteristic of the sigmoid circuit, after proper tuning, is given in figure 4-13.

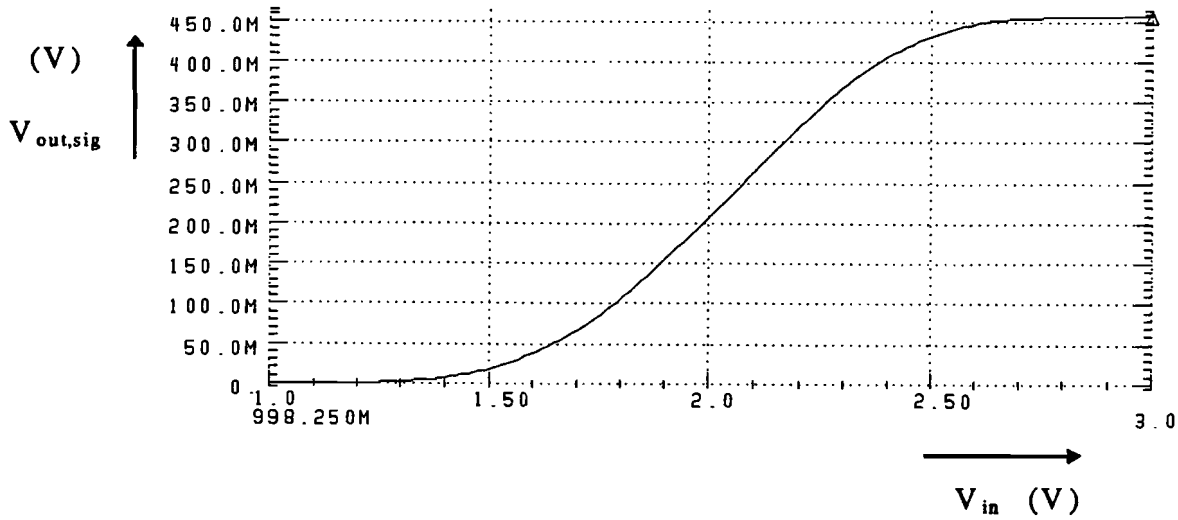


Fig 4-13 Output characteristic of the sigmoid circuit of figure 4-11

The output characteristic of the inverse sigmoid circuit of figure 4-10 with an ideal opamp (with maximum output of 3.3V and minimum output of 0V), and with the sigmoid circuit of figure 4-11, is given in figure 4-14. The non-inverting input of the opamp is also given in this figure (triangular shape). The characteristic of figure 4-14 resembles the characteristic of figure 4-9. The width of the inverse sigmoid shape can be varied by changing the magnitude of the input voltage. This effect is given in figure 4-15a. In this way, the width of the inverse sigmoid characteristic can be set to 800ns (T_{max}; the maximum time difference between the points where the characteristic crosses the 3V line). To control the signal width at the 2V line, the voltage of Vref can be changed. This is shown in figure 4-15b.

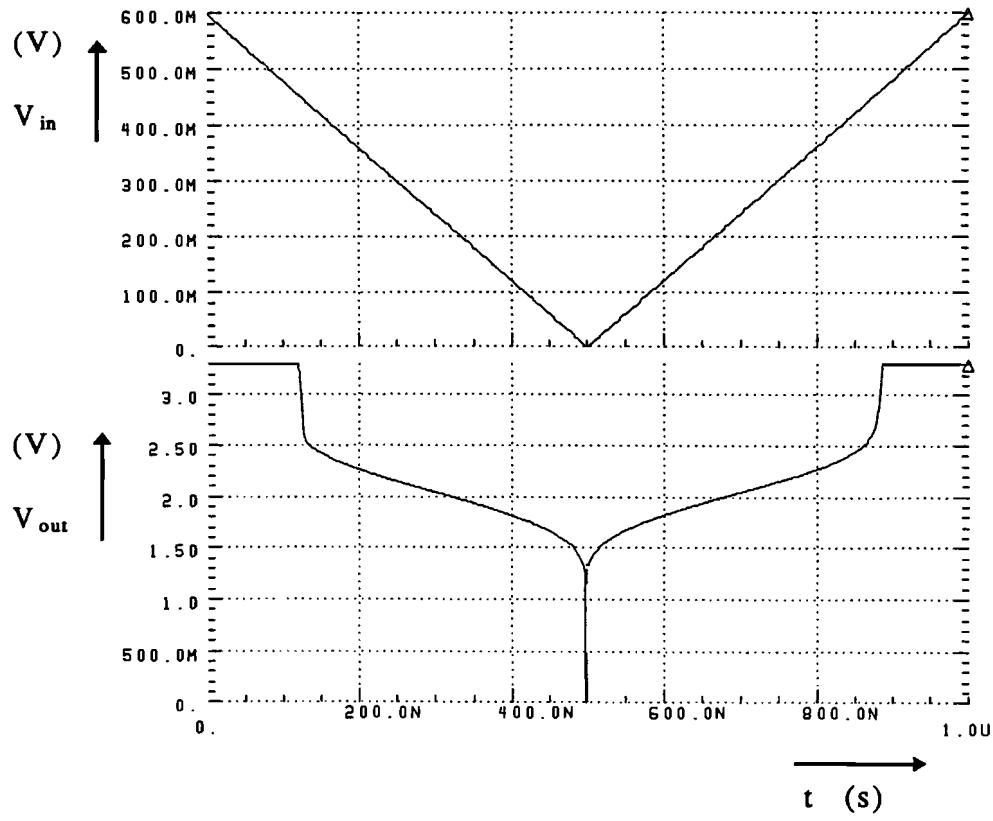


Fig 4-14 Output characteristic of the inverse sigmoid of figure 4-10 with the use of an ideal opamp

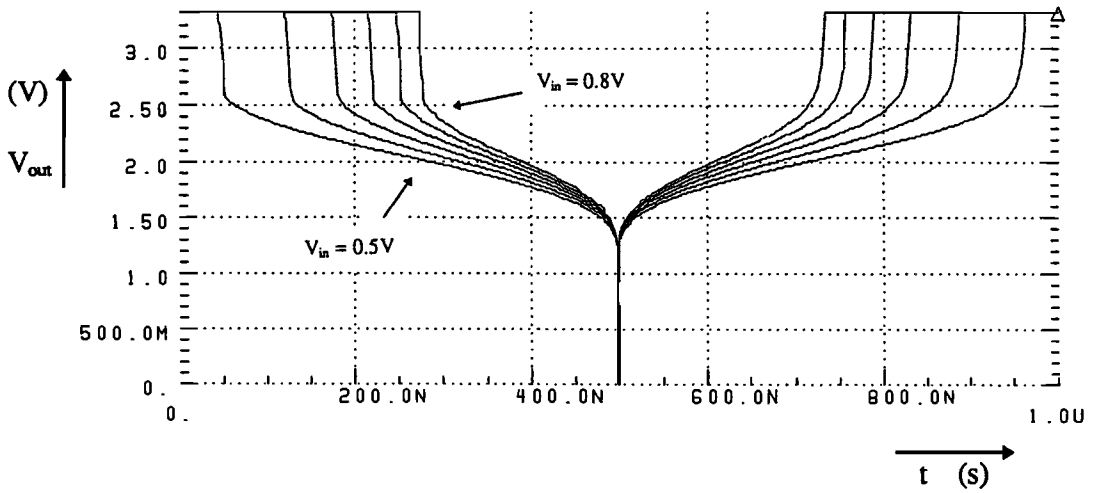
The ideal opamp must be replaced by a designed opamp. Two important parameters to take into account in designing an opamp, are the gain bandwidth frequency and the open loop gain (A_0). To obtain these two parameters, a bode diagram has been made of the sigmoid circuit (set at a dc value where the sigmoid shape has the largest slope). This bode diagram is given in figure 4-16. The transfer function of figure 4-10 has been made to examine (in-)stability of the inverse sigmoid circuit. The transfer function is:

$$H(jw) = \frac{A(jw)}{1 + A(jw)S(jw)} \tag{4.7}$$

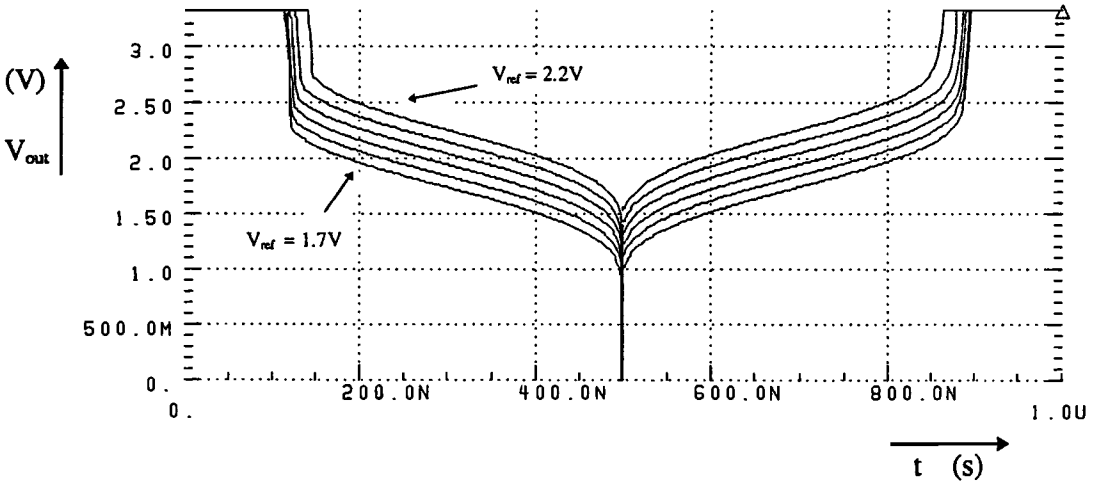
The circuit is stable when $|A(jw)S(jw)|$ curve crosses the 0dB point, before Arg[-A(jw)S(jw)] reaches 0 degrees. The S(jw) curve is already determined (figure 4-16), so if the 1/A(jw) curve, with a slope of 20dB/dec, is drawn in the same figure (intersection with the S(0) - 3dB point of the S(jw) curve), the cut-off frequency f_C of the opamp can be determined, dependent on the gain A_0 of the opamp. The gainbandwidth (GB) of the opamp is now:

$$GB = A_0 f_C \tag{4.8}$$

Usually the magnitude of the gainbandwidth is in the order of several MHz.



a) by changing V_{in}; from 0.5V to 0.8V with step 0.1V



b) by changing V_{ref}; from 1.7V to 2.2V with step 0.1V

Fig 4-15 Possible adjustments of the inverse sigmoid shape of figure 4-14

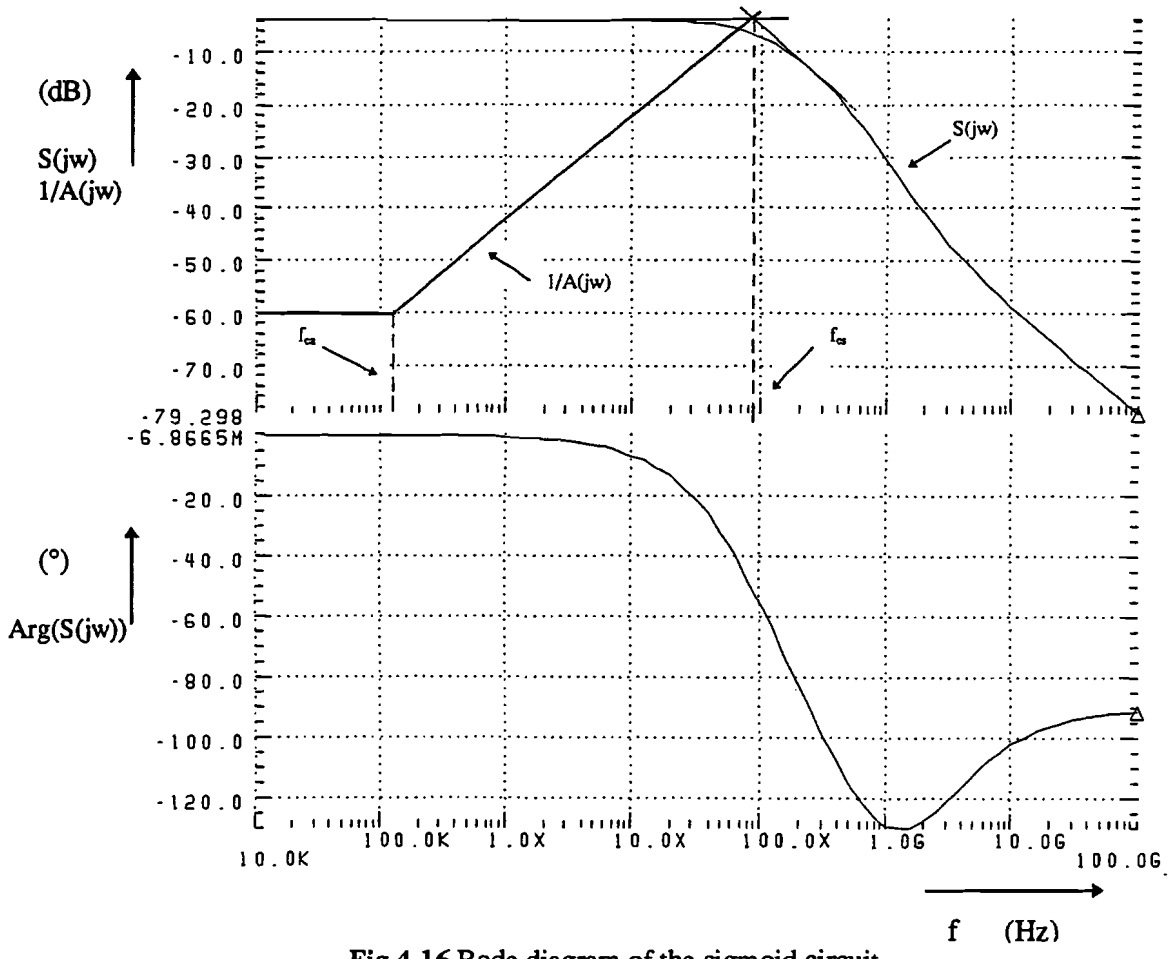


Fig 4-16 Bode diagram of the sigmoid circuit

The cut-off frequency of the sigmoid is f_{CS} = 90MHz. The desired gainbandwidth (GB) of the opamp is approximately 150MHz (A_0 = 60dB, f_{CA} = 150kHz), which is much larger than several MHz and in practice hard to realize. Therefore, the 1/A curve in figure 4-16 will be moved to the left until the 1/A(jw) curve crosses the 0dB point, with a frequency f_{CA} and gain A_0 in such a way that the gainbandwidth is 1MHz. So the gainbandwidth is set to GB = 1MHz, the amplification is chosen A_0 = 20dB and the cut-off frequency f_{CA} = 100kHz. In figure 4-17, a replacement of the ideal opamp is given to simulate the gain and the first order role-off frequency of the opamp to design.

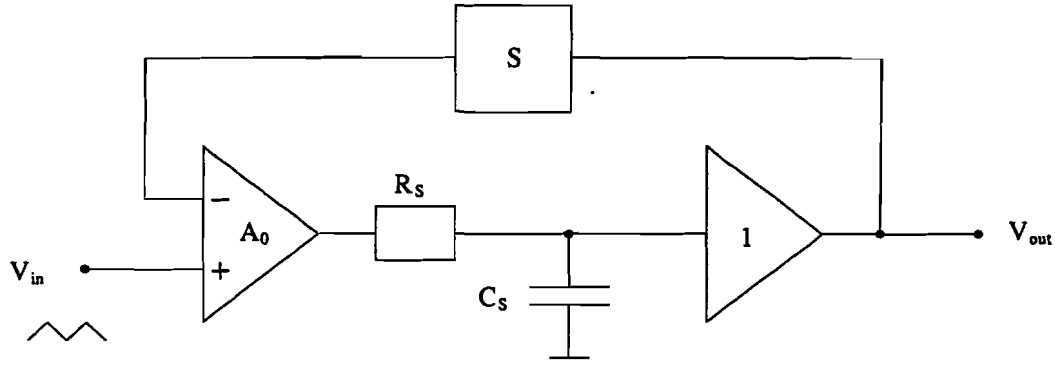


Fig 4-17 Replacement of the ideal opamp of figure 4-10

The RC value determines the first order role-off frequency of the opamp, and A_0 the gain. R_s = 1.6MΩ and C_s = 1pF to obtain the first order role-off frequency. In figure 4-18 the simulation of the circuit of figure 4-17 is given. The shape of the inverse sigmoid does not resemble the desired shape. A change of the gain will not improve the shape and a satisfactory improvement of the shape through increasing the gainbandwidth does not occur until the gainbandwidth reaches an unrealisable large value. It appeared to be unfeasible to combine the desired speed of the total circuit, with an acceptible and realisable gainbandwidth product of the opamp. Therefore a new approach must be made.

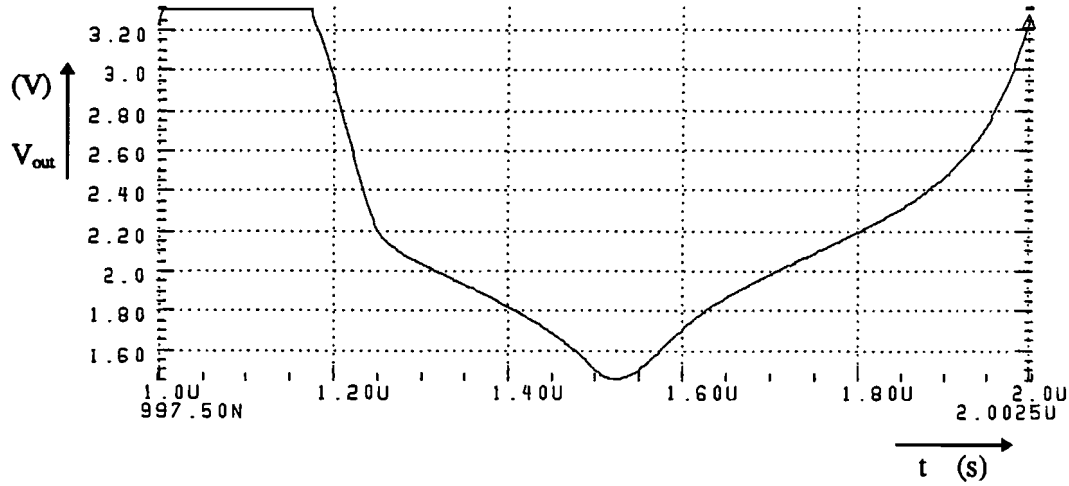
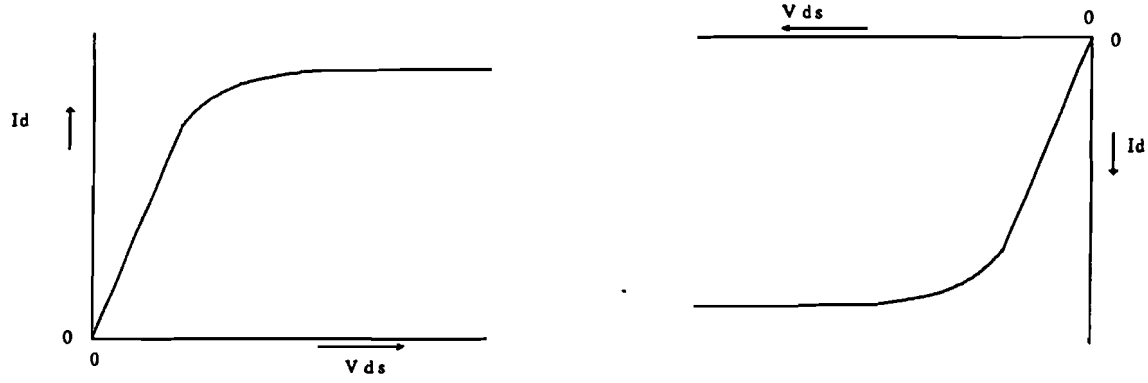


Fig 4-18 Inverse sigmoid shape of the circuit of figure 4-17

4.3.2. Inverse sigmoid circuit; the second approach

The principal idea for the final inverse sigmoid circuit, lies in the drain current versus the drain-source-voltage of a CMOS transistor. In figure 4-19a this characteristic of a NMOS transistor is given, for only one gate-source voltage to show that the shape resembles one half of the sigmoid shape.



a) NMOS b) PMOS

Fig 4-19 Characteristics of a CMOS transistor

To realise a complete sigmoid shape, the other half can be made with the aid of a PMOS transistor. The characteristic of a PMOS transistor is given in figure 4-19b. The complete sigmoid function can now be

made by connecting the NMOS transistor to the PMOS transistor. In figure 4-20, the circuit is given, which creates a sigmoid current shape (see also figure 4-21).

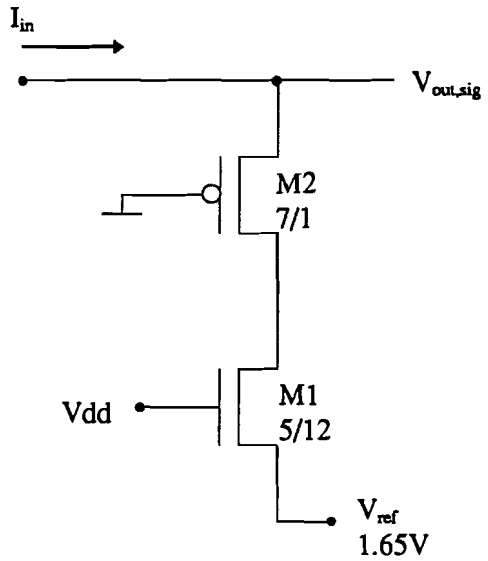


Fig 4-20 Sigmoid circuit

The gate of the PMOS transistor is connected to ground, and the gate of the NMOS transistor is connected to the supply voltage of 3.3V. This is done to minimize the necessary number of input pins of the total circuit. The voltage V_{ref} is added to the circuit to establish a symmetrical shape around a reference voltage, which is needed for appropriate working of the circuit. The sizes of the transistors where obtained by optimisation of the sigmoid circuit.
If the current Iin varies lineary from 0 to a positive current, then the drain-source voltage of transistor M2 is small, and if we neglect M2, the output voltage equals $V_{ref} + V_{ds,M1}$. If the current Iin varies lineary from 0 to a negative current, then the drain-source voltage of transistor M1 is small, and the output voltage (with neglection of M1) equals now $V_{ref} - V_{ds,M2}$. In figure 4-21, the current through the circuit as a function of the output voltage $V_{out,sig}$ is given.

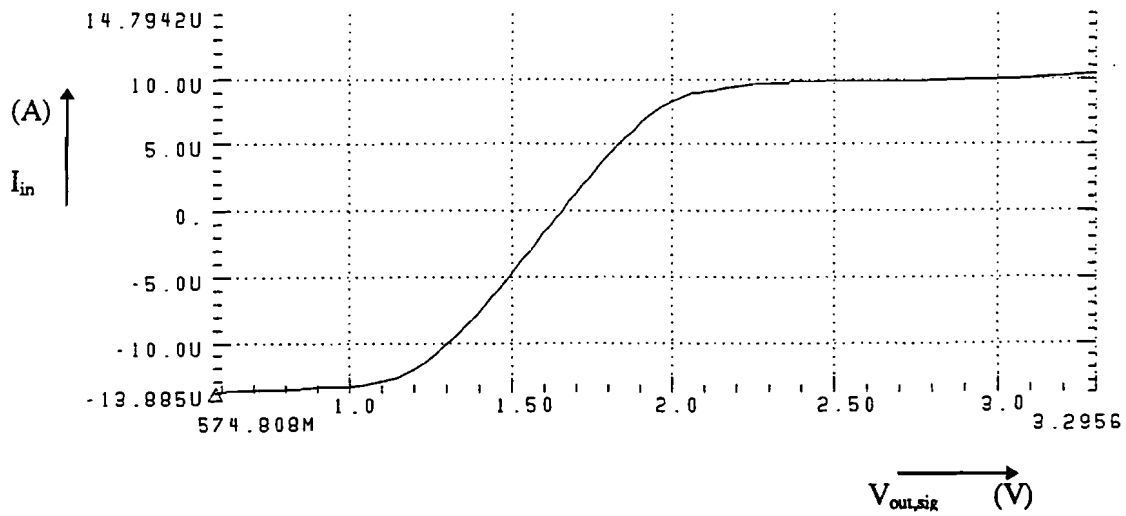
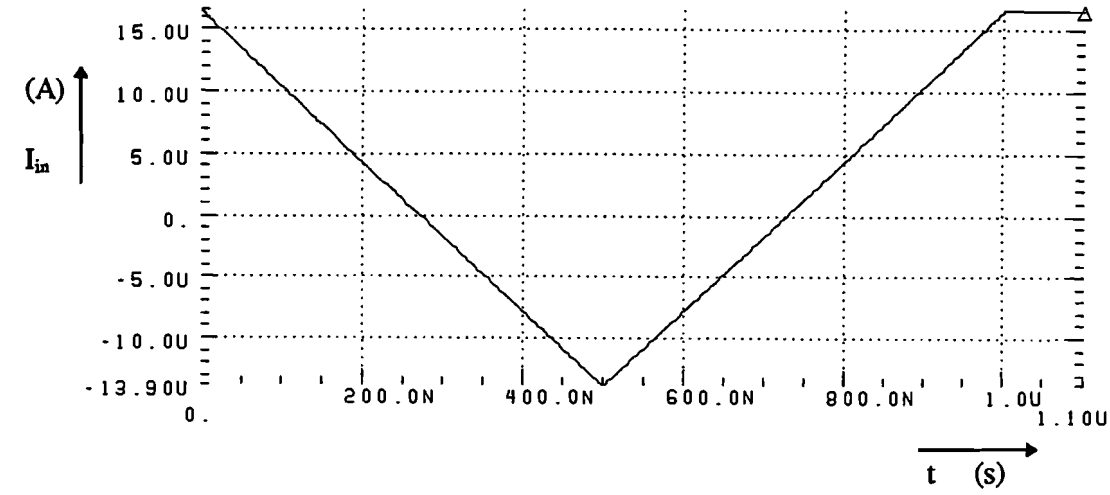
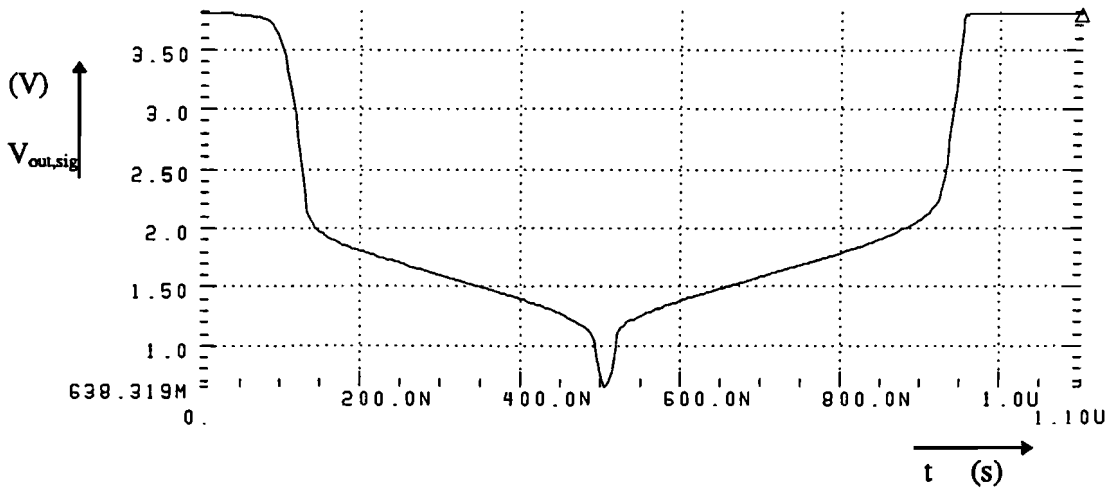


Fig 4-21 Sigmoid shape of the current through the sigmoid circuit; V_{ref} = 1.65V

If the current is linearly varied in time, and the output voltage $V_{out,sig}$ is placed on the vertical axis, then an inverse sigmoid shape can be established. The time then, is placed on the horizontal axis. This is given in figure 4-22; in figure 4-22a, the current I_{in} is given and in figure 4-22b, the inverse sigmoid shape is given. The current has a triangular shape. Looking at figure 4-22b, it is clear that the sigmoid circuit of figure 4-20, is actually an inverse sigmoid circuit.



a) Input current



b) Output voltage

Fig 4-22 Input current and output voltage of the (inverse) sigmoid circuit

A requirement of the non-linear function circuit is, that the output voltage must vary between 1V and 3V, with a 2V centre. From figure 4-22, it is clear that the centre is approximately 1.65V (as expected, because Vref is at 1.65V in the (inverse) sigmoid circuit). There are three possibilities to solve this problem: the first one is to higher the reference voltage, but this will lead to an asymmetrical shape of the inverse sigmoid, the second one is to adjust the synapse unit and the neuron unit, but this requires redesigning of these circuits, and the third possibility is to add a circuit which sets the centre to 2V. The third solution was chosen, because the shape will be maintained and in the extra circuit, adjustments can be made to solve problems, due to parameter variations. Furthermore, the circuit of figure 4-20 also needs a buffer circuit, and this can be combined with this third solution.

The basic idea of the third solution is given in figure 4-23.

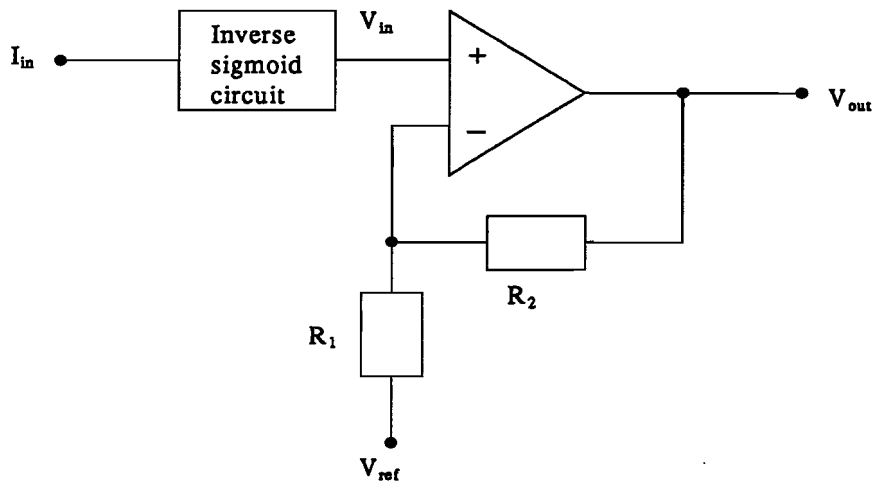


Fig 4-23 Adjustment circuit for the inverse sigmoid circuit

The inverse sigmoid circuit will not be affected by the opamp because of it's high input impedance. So the shape of the inverse sigmoid characteristic will be maintained.
The output voltage of the inverse sigmoid circuit will be amplified, and the output characteristic can be verticaly shifted by changing V_{ref}. The relation between the output V_{out} and the inputs V_{in} and V_{ref} is:

$$V_{out} = \left(1 + \frac{R_2}{R_1}\right)V_{in} - \frac{R_2}{R_1}V_{ref} \tag{4.7}$$

The simulation of the circuit of figure 4-23 is given in figure 4-24, with R_1=100kΩ, R_2=40kΩ and a V_{ref} variation from 0.3V to 1.1V, with step 0.2V.

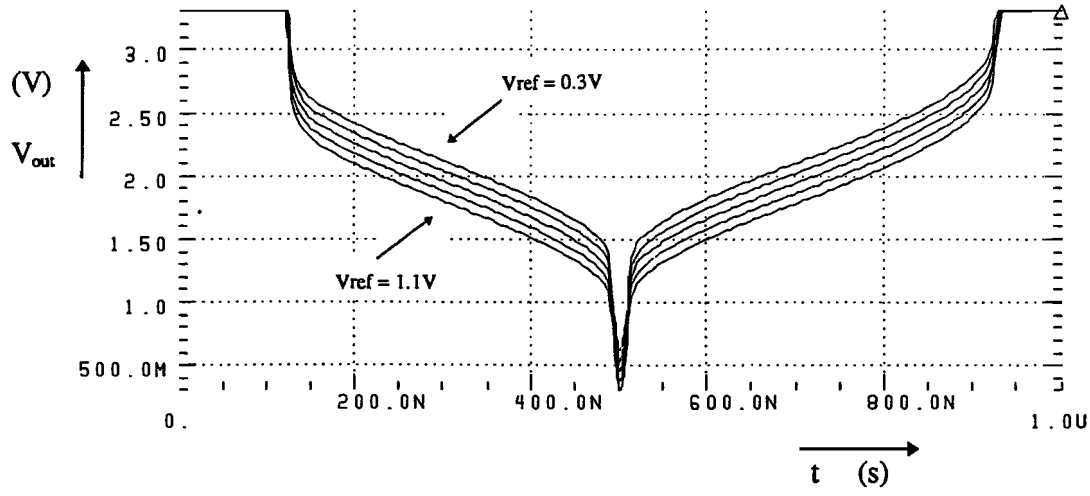
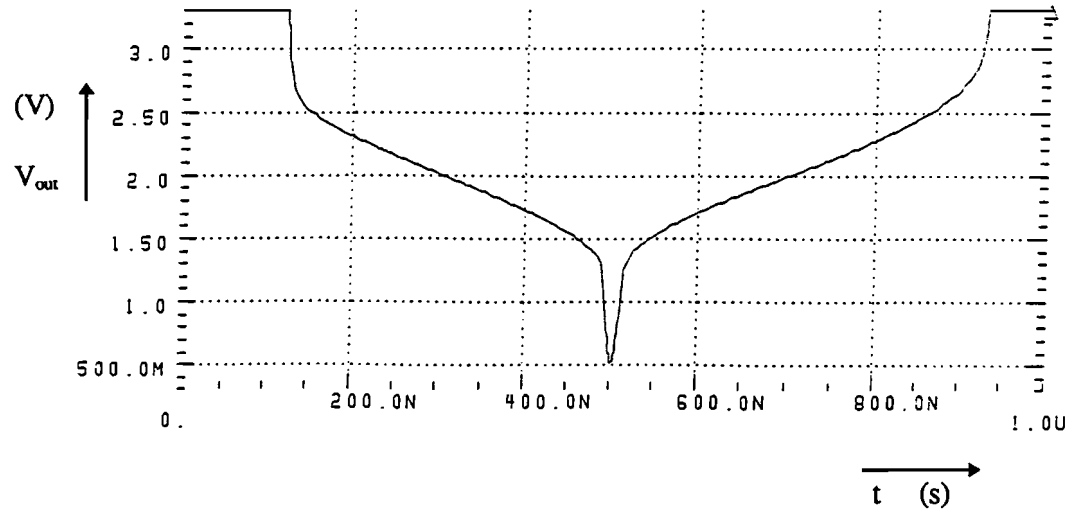


Fig 4-24 Output characteristic of the simulation of the circuit in figure 4-21, with V_{ref} changing from 0.3V to 1.1V with step 0.2V

These values are only taken to show the principle of the former circuit, with the exception of the relation between R_1 and R_2: with V_{in} = 1.6V (centre of the output voltage of the inverse sigmoid circuit) set to a 2V centre, and a variation possibility accomplished by V_{ref}, the relation between R_1 and R_2 is $R_2 = 0.4R_1$. The reference voltage is then (see (4.7)) : $V_{ref} = 0.6V$.

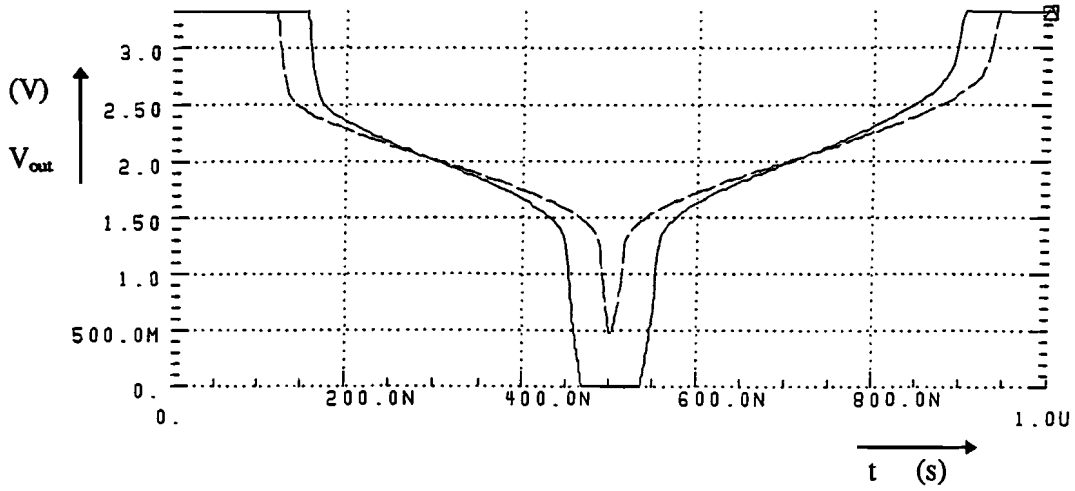
It is also very interesting to know if the shape of the inverse sigmoid characteristic will be maintained by parameter variation. One has the ability to check this by the use of two models, a FAST- and a SLOW model. These simulations are given in figure 4-25; in figure 4-25a the normal output characteristic, in figure 4-25b the SLOW-model output characteristic, unadjusted and adjusted (dotted line) by changing the input current (16,5µA set to 13.2µA and -13.9µA to -11.4µA) and in figure 4-25c the FAST-model output characteristic, unadjusted and adjusted (dotted line) by changing the input current (16.5µA set to 20.6µA and -13.9µA to -17.7µA). With the adjustment of the input current, the shape of the inverse sigmoid can be maintained.

To translate the circuit of figure 4-23 into a circuit at CMOS level, the formulas for designing a two-stage n-channel input operational amplifier discussed in [1], are used and afterwards, the sizes of the transistors are adjusted for better permormances. After examining the output characteristic of a n-channel input opamp in comparison with a p-channel input opamp, the last one was choosen because of its better performance (shape of the inverse sigmoid). For the translation of R_1 and R_2 into CMOS transistors, care must be taken with the current through these transistors. In order to maintain the shape of the inverse sigmoid characteristic, the current through these transistors must be small enough to let the output current of the opamp, charge (and discharge) a load capacitance of 2pF (total capacitance of all connected neuron units to the inverse sigmoid). Furthermore, a 'resistor' is added in series with the miller capacitor for stability purposes (R_{mil}). In figure 4-27, the final circuit of the inverse sigmoid is given.

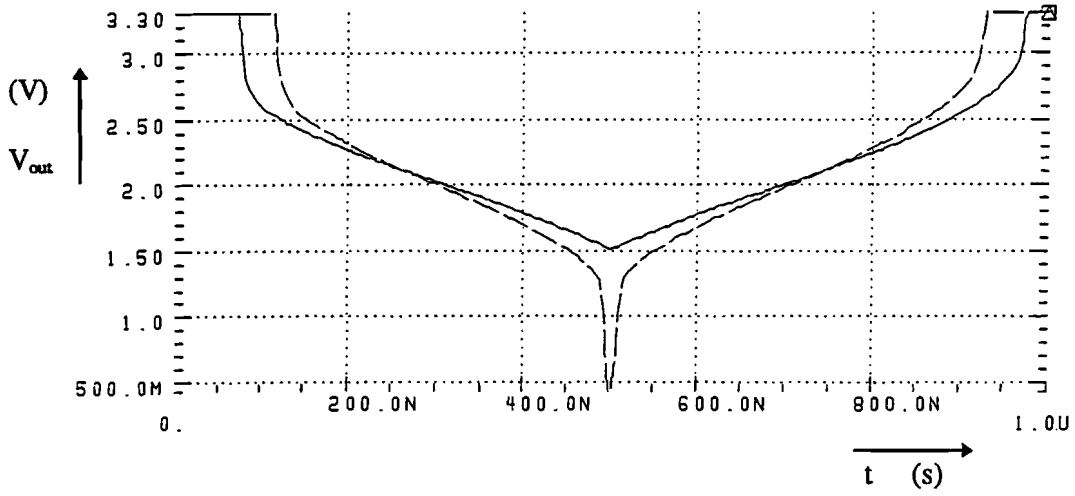


a) normal output characteristic

Fig 4-25 Output characteristic of the inverse sigmoid circuit by using a SLOW- and a FAST-model



b) output characteristic SLOW-model; input current unadjusted and adjusted (dotted line)



c) output characteristic FAST-model; input current unadjusted and adjusted (dotted line)

Fig 4-25 Output characteristic of the inverse sigmoid circuit by using a SLOW- and a FAST-model

Some specifications used in designing the opamp, are :

Vdd = 3.3V
GB (gain bandwidth) = 1MHz
C_{load} = 2pF
SR (slew rate) = 40 V/µs
V_{in} : $0.5V < V_{in} < 3V$
V_{out} : $0.5V < V_{out} < 3.3V$

The total power dissipation of the inverse sigmoid circuit is $P_{diss} = (I_{bias} + I_{output\ stage,\ opamp})Vdd = (20\mu A + 80\mu A)3.3V = 330\mu W$. This seemes a large power dissipation, but this circuit will only be used once, and is connected to all neurons on the chip.

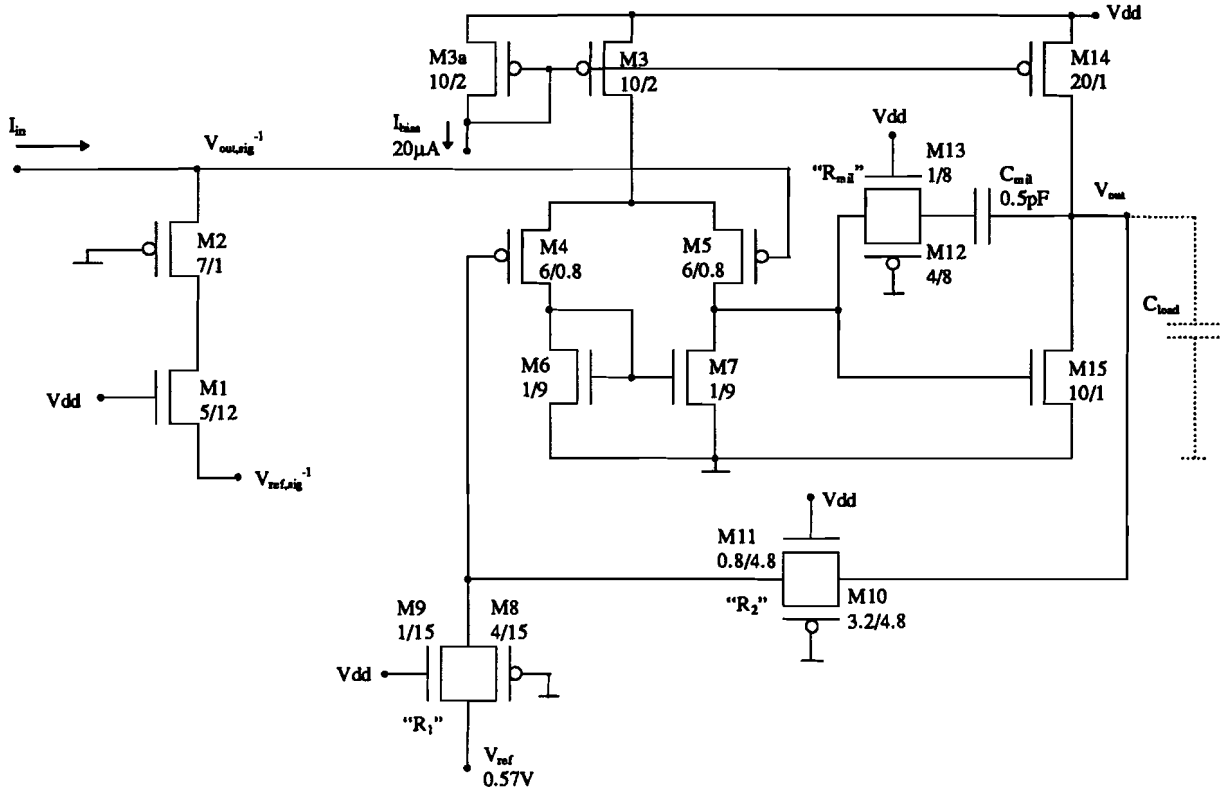
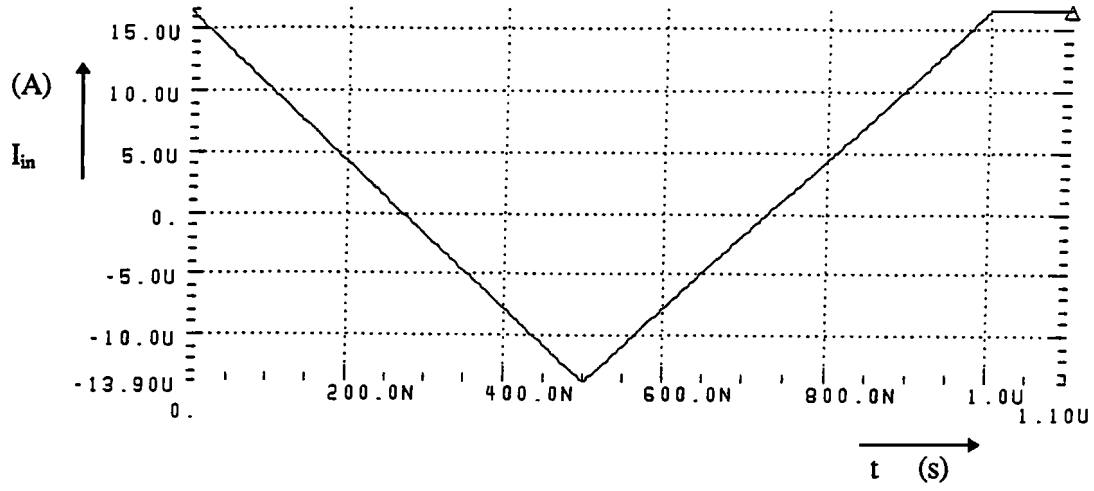


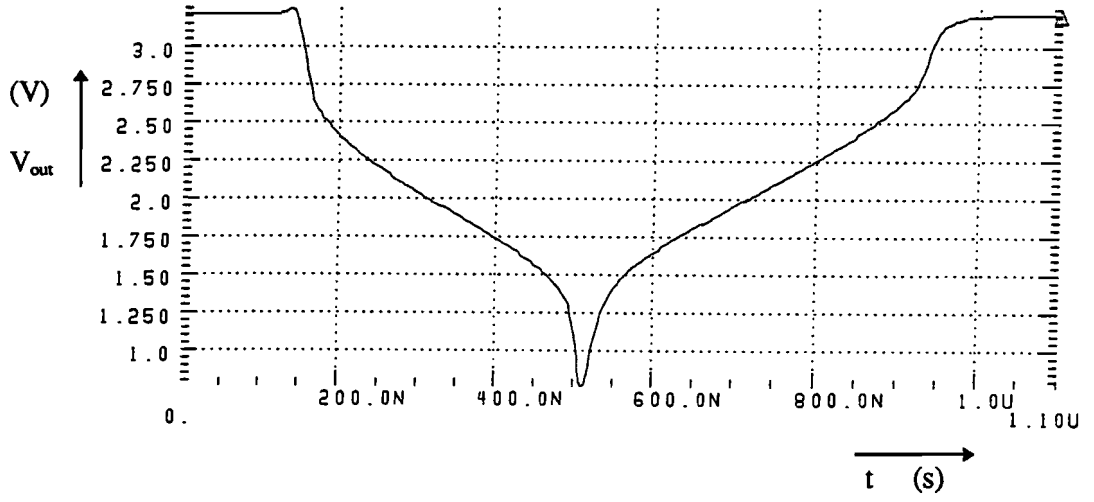
Fig 4-26 Final circuit of the inverse sigmoid

The simulation of the circuit in figure 4-26 is given in figure 4-27, together with the input current characteristic (figure 4-27a)

From figure 4-27, it can be seen that the centre output voltage is approximately 2V (time axis : 300ns and 700ns), the time duration where the output voltage is lower than 3V (T_{max}; see also chapter 3) is approximately 790ns and the time duration where the output voltage is lower than 1V (T_{min}) is approximately 20ns. Furthermore, the minimum output voltage is at 510ns (ideal case: 500ns) and the time duration of the first half of the inverse sigmoid characteristic is smaller (355ns) than the right half (431ns). So there are a few 'minor' non-idealities due to parasitic effects. If necessary, these non idealities can be solved by adjusting the timing of the input current. The output characteristic of the inverse sigmoid with adjustment of the input current is given in figure 4-28. In this figure, the timing is already adjusted to the timing of the synapse (see figure 3-7) and V_{ref} is set to 0.57V, to make the duration, where the output voltage is lower than 2V, 400ns.
The other time durations (T_{max} and T_{min}) stay the same as before the input current adjustment. The time duration of the left half equals approximately the right half of the output characteristic



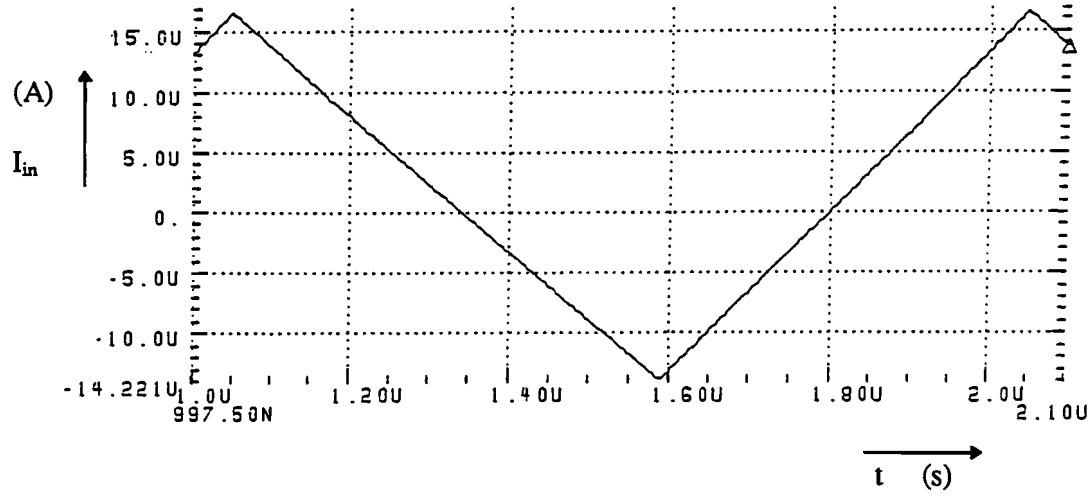
a) input current characteristic



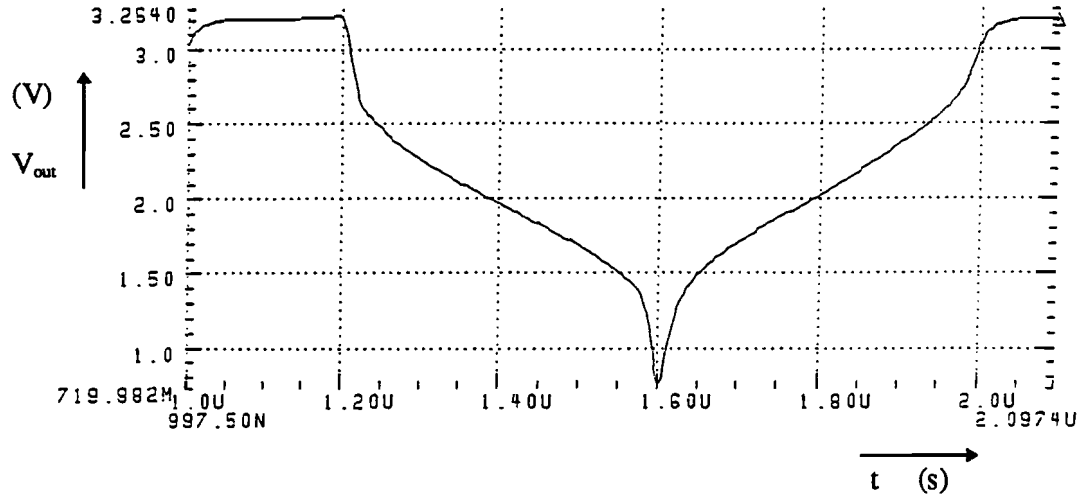
b) output characteristic

Fig 4-27 Output characteristic of the inverse sigmoid circuit given in figure 4-24

The conclusion of the designed inverse sigmoid circuit is that it is a flexible circuit; the output characteristic can be adjusted for use with several specifications.



a) input current with adjustment



b) output characteristic with input current adjustment

Fig 4-28 Output characteristic of the inverse sigmoid circuit with input current adjustment

4.4. The comparator

The last part of the neuron unit (see figure 4-1) is the comparator. The purpose of the comparator is to compare the output voltage of the sample & hold circuit with the output voltage of the non-linear circuit (inverse sigmoid circuit). The output of the comparator consists of a pulse with a time duration dependent on that comparison. The basic circuit of a comparator used, is an operational amplifier.

An output pulse must occur when the output voltage of the sample & hold circuit is higher than the output voltage of the inverse sigmoid circuit. So V_- represents the output of the inverse sigmoid circuit and the V_+

represents the output of the sample & hold circuit. There are two important specifications for the comparator, the first one is the offset voltage, and the second one is speed; the comparator (the output of the neuron unit) is connected to a large number of synapses, and these synapses together will cause a load capacitance at the output. This load capacitance will influence the reaction time (propagation delay) of the comparator. These specifications are :

offset voltage $V_{os} < 1mV$
propagation delay < 20ns , with a 10pF load

Another specification is that the output voltage is high and low enough for proper working of the connected synapses.

In figure 4-29, the comparator circuit is given. The comparator is a n-channel input operational amplifier because of its better performance (low offset voltage, largest input swing) compared to a p-channel input opamp. Furthermore, two inverters are connected to the opamp. This is to ensure maximum output voltage swing, and an increase of speed. The input swing (highest and lowest input values of both inputs) is 2V (between a minimum of 1V and a maximum of 3V). From simulation results it can be concluded that the offset voltage of the comparator is:

$$V_{os} \leq 0.4mV \quad \text{for} \quad 1V \leq V\_ \leq 3V$$

In figure 4-30a, the propagation delay is shown of the comparator with a load capacitor of 10pF, and in figure 4-30b, with a load capacitor of 20pF.

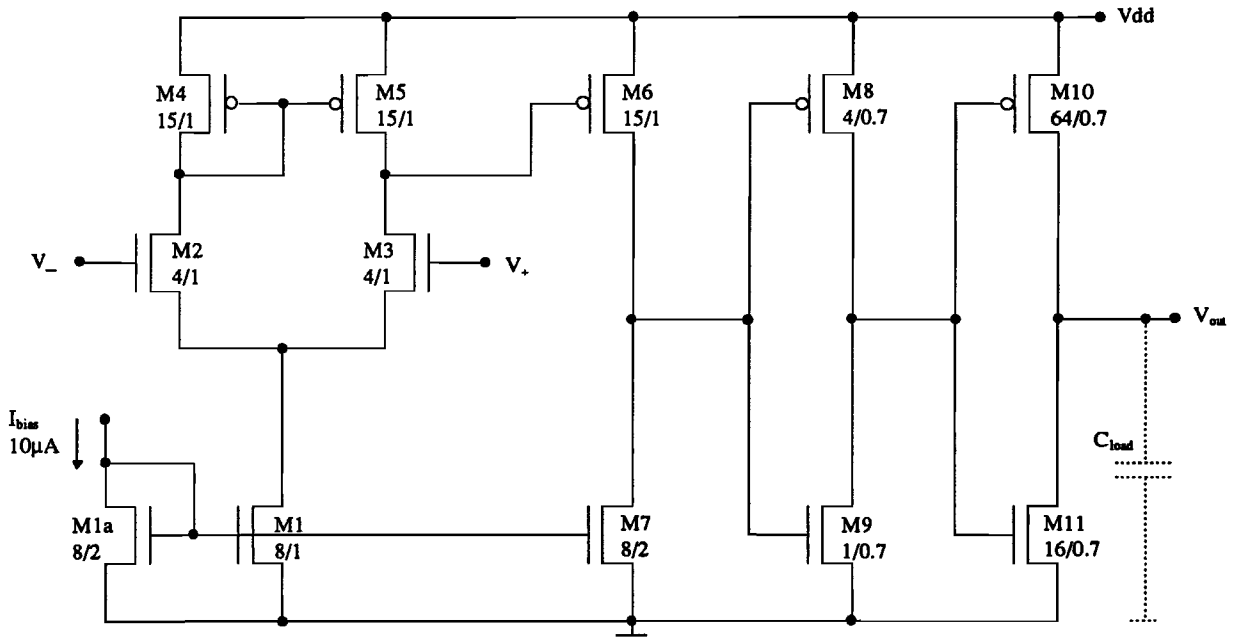
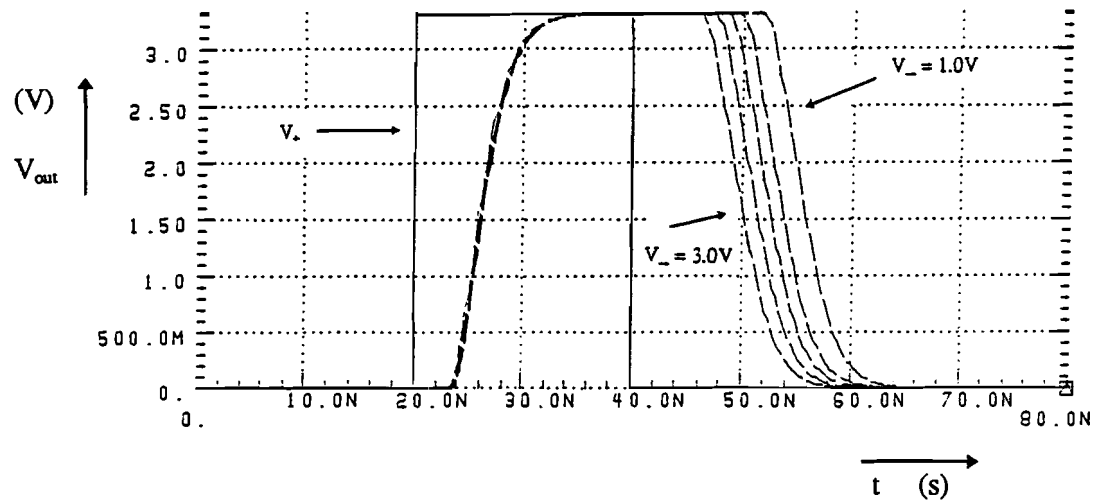
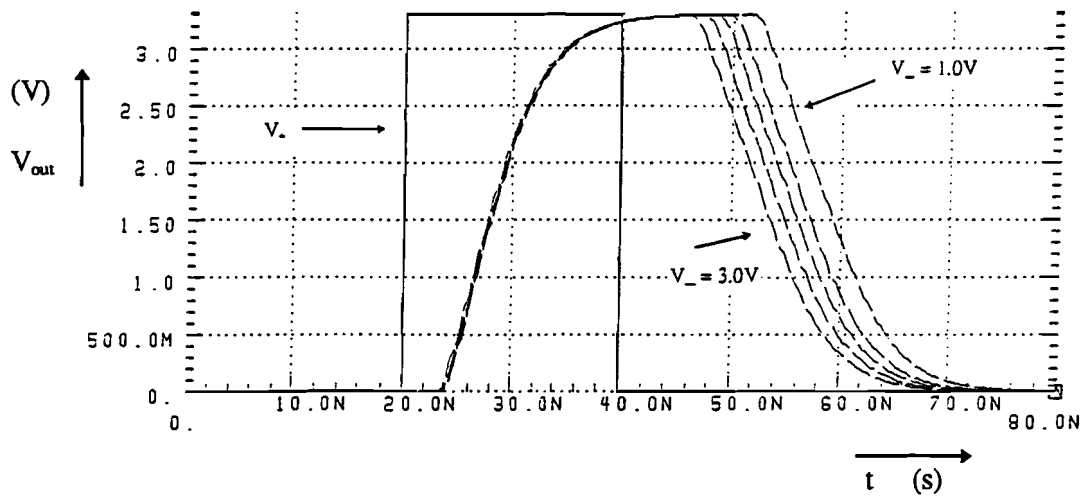


Fig 4-29 Comparator circuit

These propagation delay times can be split into a propagation delay time in the case the output voltage level goes from a low (L) to a high (H) level, and in the case the output voltage level goes from H to L. These maximum times are as follows: for a load of 10pF, $t_{P,LH} \leq 6.5$ns and $t_{P,HL} \leq 15.6$ns, and for a load of 20pF, $t_{P,LH} \leq 8.6$ns and $t_{P,HL} \leq 18.7$ns.
The rise- and fall-time (times when the output voltage is 10% and 90% of the maximum output voltage) are in the case of a 10pF load capacitance $t_{rise} \leq 5.1$ns and $t_{fall} \leq 6.2$ns, and in the case of a 20pF load capacitance $t_{rise} \leq 9.9$ns and $t_{fall} \leq 12.3$ns.
The minimum output voltage of the comparator is 44µV, the maximum output voltage is 3.3V.



a) with a load capacitance of 10pF



b) with a load capacitance of 20pF

Fig 4-30 Propagation delay characteristic of the comparator

The power dissipation of the comparator is, due to the two inverters, only high during the transition of the output voltage from a low- to a high level, or from a high- to a low level, and the duration time of the transition is dependent on the load capacitance.

It can be concluded that the comparator meets to the specifications (with no consideration of mismatch effects), certainly in the case of the offset voltage.

5. The complete circuit

In this chapter, alle the separate parts, discussed in chapters 3 and 4, will be connected together to form the complete circuit of the neuron and the synapse. In figure 5.1, these different parts are given (symbolic, not the circuitry in detail).

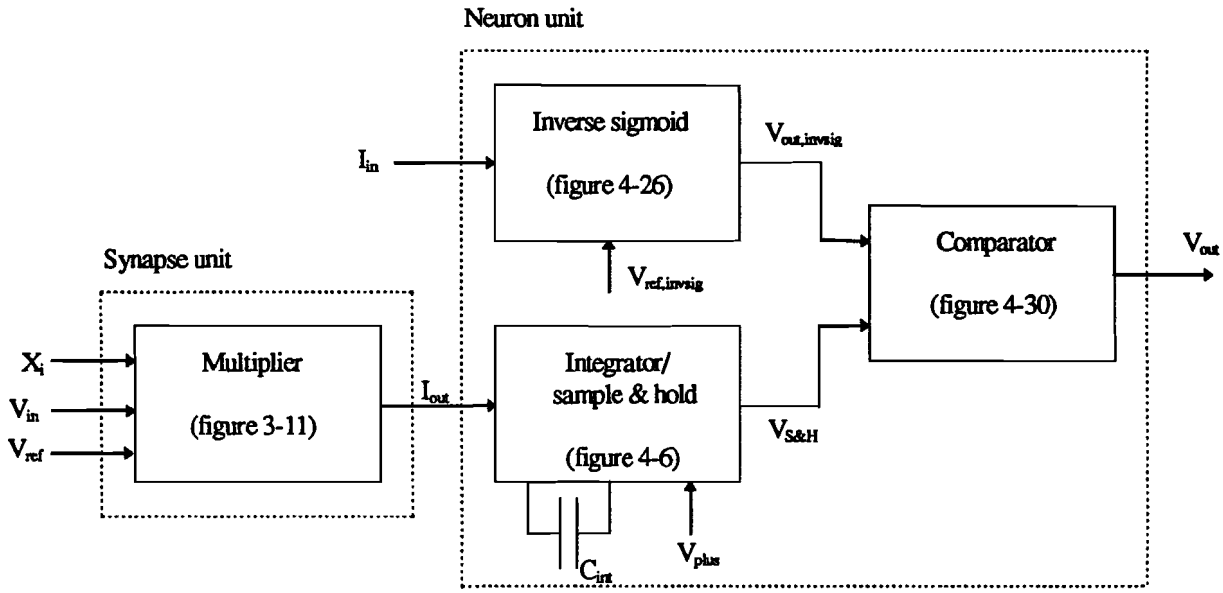
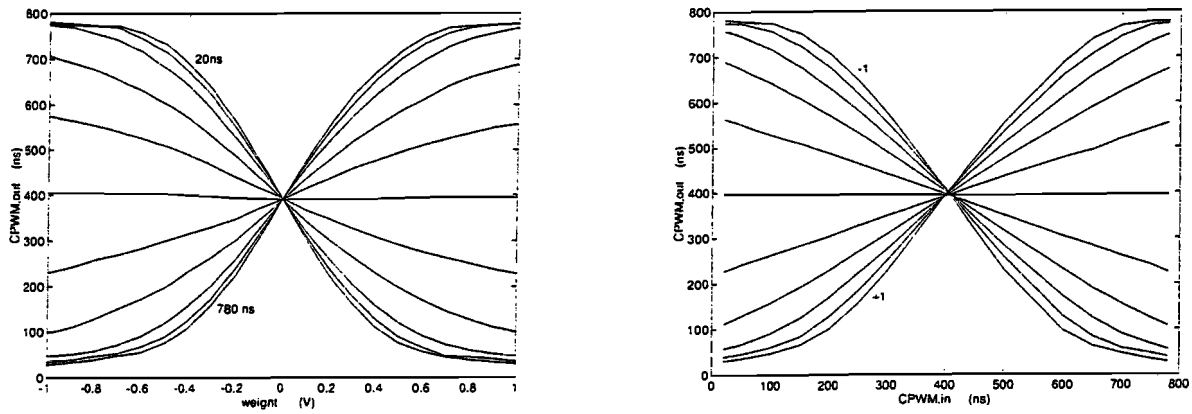


Fig 5.1 Schematic of the complete neuron- and synapse unit

The input of the synapse consists of the CPWM input signal X_i, a voltage V_{in} and a voltage V_{ref} : the difference of these two voltages represent the weight value. The output of the synapse is a current I_{out}, which forms the input of the neuron unit. The integrator part of the neuron unit can be adjusted to the number of connected synapses, by changing the capacitor value C_{int}. The output voltage $V_{S\&H}$ of the sample & hold circuit will be compared with the output voltage of the inverse sigmoid $V_{out,invsig}$. Dependent on this comparison, an output pulse will be generated. The inverse sigmoid can be adjusted by changing I_{in} and/or $V_{ref,invsig}$. The comparator compares a linear (or almost linear) input signal with a non-linear input signal (inverse sigmoid), so the output will be a pulse with a time duration which has a sigmoid relation with the input signal X_i.

The simulation is done in the case of one synapse connected to the neuron unit (C_{int} = 1.8pF), and with an inverse sigmoid shape of figure 4-28 (with an adjusted input current). Furthermore, the inverse sigmoid circuit and the comparator are simulated with a load capacitance of 2pF and 20pF respectively. The simulation results are given in figure 5.2; figure 5.2a represents the output pulse duration (CPWM,out) versus the weight value with several CPWM,in values, and figure 5.2b the output pulse duration (CPWM,out) versus the input pulse duration (CPWM,in) with several weight values. In figure 5.2a, these CPWM,in values are respectively 20, 50, 100, 200, 300, 400, 500, 600, 700, 750 and 780ns, and in figure 5.2b, the weight values are respectively -1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1.0 V. From these characteristics, it can be seen that they have a sigmoid shape; figure 5.2a even more than 5.2b because in the case of CPWM,out versus the weight, the input of the comparator, coming from the sample & hold circuit, is not a linear signal (see chapter 3.2 and figure 3-12a). So, the goal to accomplish a saturation in the neuron's response has been reached.



a) CPWM,out versus the weight

b) CPWM, out versus CPWM,in

Fig 5-2 CPWM,out characteristic of the complete neuron and synapse unit

The maximum output pulse duration is approximately 780ns, and the minimum output pulse duration is approximately 20ns.

6. Conclusions and recommendations

The linearity of output voltage versus input pulse width of the synapse unit is excellent, and in the case of output voltage versus weight less excellent, but this has no effect on the total network behaviour. Furthermore, the synapse unit has a low power dissipation (13.2μW), is small in circuit size and has a large weight input range (2V).

The integrator/sample & hold part of the neuron unit converts linearly the input current in an output voltage (centre at 2V with a 2V swing). It has a relatively low power dissipation (42μW), and it has the possibility (in the form of the integrator capacitor) to adjust the circuit to the amount of connected synapses, so a flexible neuron has been realised.

The shape of the inverse sigmoid of the non-linear part of the neuron unit can be maintained in the case of parameter variation. Also, the timing can be changed by changing its input current and/or reference voltage. The only drawback is its power dissipation of 330μW, but the inverse sigmoid circuit will be implemented once and can be used by several neuron units. In order to reduce the power dissipation of the inverse sigmoid, it will be interesting to examine the possibility of level shifting of the characteristic of figure 4-22, produced by figure 4-20. In the future, the input current of the inverse sigmoid circuit will be implementen on the chip, so it is desireable to have a current source, which generates a current in such a way, that the inverse sigmoid shape will be maintained in the case of parameter variation.

The comparator part of the neuron unit, has a small offsett voltage ($V_{os} \leq 0.4mV$) and a high speed (propogation delay time $\leq$ 15.6ns with a 10pF load capacitance, and $\leq$ 18.7ns with a 20pF load capacitance). This high speed might lead to a high power dissipation, but this has been reduced to approximately 100μW by the use of inverters in the comparator.

The complete circuit realises a boundery of the neuron's response and has a maximum output pulse duration of approximately 780ns, and a minimum output pulse duration of approximately 20ns.

Literature

[1] **Allen, P.E., Holberg, D.R.**
'CMOS Analog Circuit Design'
Dryden Press, 1987

[2] **Boorn, J.H. van den**
'Elektronica I', deel 2
Faculteit Electrotechniek, Vakgroep Elektronische Schakelingen EEB, TU Eindhoven, 1987
Collegedictaat nr. 4005603400000

[3] **Claasen-Vujcic T.**
'Implementation of a Multi-Layer Perceptron Using Pulse Stream Techniques'
Master Thesis, TU Eindhoven, Febr. 1993

[4] **Hodges, D.A., Jackson, H.G.**
'Analysis and Design of Digital Integrated Circuits'
McGraw-Hill International editions, 1988

[5] **Kartalopoulos, S.V.**
'Understanding neural networks and Fuzzy logic'
IEEE Press, 1996

[6] **Masa, P. et al.**
'A High-Speed Analog Neural Processor'
IEEE Micro, Vol. 14-15,1994-95, pp. 40-50

[7] **Murray, A.F., et al.**
'Pulse-Stream Arithmetic in Programmable Neural Networks'
IEEE International Symposium on Circuits and Systems, Vol.2, 1989, pp. 1210-1212

[8] **Murray, A.F.**
'Pulse Arithmetic in VLSI Neural Networks'
IEEE Micro, Vol. 9, No. 6, Dec. 1989, pp. 64-74

[9] **Murray, A.F. et al.**
'Pulse Stream VLSI Neural Networks'
IEEE Micro, Vol. 14-15, 1994-1995, pp. 29-39

[10] **Persoon, G.G.**
'Moderne Elektronika'
Faculteit Electrotechniek, Vakgroep Elektronische Schakelingen EEB, TU Eindhoven, 1990
Collegedictaat nr. 4005610100009

[11] **Petin, Y.A.,**
'Implementation of a Multi-Layer Perceptron including Back Propagation Training Algorithm'
Master Thesis, TU Eindhoven, Aug. 1993

[12] **Reyneri, L.M., Sartori, M.**
'A Neural Vector Matrix Multiplier using Pulse Width Modulation Techniques'
Proceedings of the Second International Conference on Microelectronics for Neural Networks, Munich, Oct. 16-18, 1991, pp. 269-272

[13] **Reyneri, L.M., et al.**
'A Comparison between Analog and Pulse Stream VLSI Hardware for Neural Networks and Fuzzy Systems'
Proceedings of the Fourth International Conference on Microelectronics for Neural Networks and Fuzzy Systems, Turin, Italy, Sept. 1994, pp. 77-86

[14] **Reyneri, L.M.**
'A Performance Analysis of Pulse Stream Neural and Fuzzy Computing Systems'
IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, Vol. 42, No. 10, Oct. 1995, pp. 642-660

[15] **Withagen, H.C.A.M.**
'Neural networks: Analog VLSI Implementation and Learning Algorithms'
Eindhoven: Technische Universiteit Eindhoven, 1997. Doctoral Dissertation.

[16] 'Metingen in de geneeskunde I'
Faculteit Electrotechniek, Vakgroep Medische Elektrotechniek, TU Eindhoven, 1991
Collegedictaat nr. 4005006400003