FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Desenho de um neurónio analógico com recurso a tecnologia de filmes finos

Teresa Margarida Pereira Pias

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Vítor Manuel Grade Tavares

29 de outubro de 2022

Resumo

Recentemente, tem havido um aumento no uso de redes artificiais para identificação de objetos e reconhecimento de padrões. Estas redes neuronais tendem a ter um grande número de camadas e estruturas complexas que consequentemente consomem um grande nível de potência. O mesmo não ocorre para sistemas biológicos, como o cérebro humano, pois possui uma grande capacidade de processamento e desempenho com baixo gasto energético. A dinâmica do cérebro torna-o altamente eficiente, por isso, a inspiração neste sistema para o desenvolvimento de sistemas eletrónicos permite melhorar a eficiência energética dos mesmos.

Existem vários modelos que replicam o comportamento dos neurónios para integrar em redes neuronais e perante diversas implementações pode-se constatar que a tecnologia predominante é CMOS. Apesar do uso de CMOS ser amplamente utilizado para redes neuronais, TFTs são uma alternativa atraente para projetar o neurónio, pois trazem algumas vantagens em relação ao CMOS. Algumas das considerações englobam o facto de TFTs baseados em a-GIZO serem uma tecnologia muito promissores por apresentarem elevado grau de transparência e mobilidade alta que se traduz num bom desempenho. Além disto, são fabricados a temperatura ambiente e processados a temperaturas mais baixas (cerca de 150ºC), permitindo uma maior diversidade de substratos, contrariamente a CMOS. Contudo, não possui pares complementares, uma vez que apenas foi possível obter bons valores de desempenho para transístores do tipo n. Como consequência, é necessário fazer mais manipulações nos circuitos para compensar esta limitação.

No processo de desenvolvimento desta dissertação, o foco inicial foi modelizar o TFT fornecido para prevenir futuros problemas com a operação do transístor numa região possivelmente problemática. Posteriormente à validação do modelo, este foi utilizado para o desenvolvimento de circuitos. Sendo assim, primeiramente foi implementado um circuito do neurónio baseado num existente, com tecnologia CMOS. No decorrer desta etapa, foram estudadas algumas opções para construir partes integrantes desse circuito com TFTs, nomeadamente, foram propostos alguns inversores para constituir a parte comparadora do circuito e uma forma alternativa de implementar o período refratário devido a limitações do TFT. Atingida a proposta final e após verificar que o neurónio gerava os impulsos esperados face a uma corrente de entrada, procedeu-se à fase seguinte do desenvolvimento. Como o objetivo do neurónio é ser integrado numa rede neuronal, é necessário o elemento de ligação entre neurónios: a sinapse. Sendo assim, inicialmente foi proposta uma sinapse simples que gerava um impulso de acordo com o peso atribuído. Mais tarde, surgiu a opção de usar um multiplicador para representar a sinapse, onde as entradas representariam um sinal com uma determinada amplitude e um sinal para definir o peso associado. Como este multiplicador tem a saída diferencial, houve a necessidade de desenvolver um circuito complementar que gere uma saída única. Posteriormente foi verificada a gama de operação destes circuitos conectados. Numa etapa final do desenvolvimento, procedeu-se à fase de *layout* de cada elemento escolhido para participar na construção da rede neuronal.

Abstract

Recently, there has been an increase in the use of artificial networks for object identification and pattern recognition. These neural networks tend to have many layers and complex structures that consume a large amount of power. The same does not occur for biological systems, such as the human brain, as it has a great processing capacity and performance with low energy expenditure. The dynamics of the brain make it highly efficient, so the inspiration in this system for developing electronic systems allows them to improve their energy efficiency.

Several models replicate the behavior of neurons to be integrated into neural networks. In the face of different implementations, it can be seen that the predominant technology is CMOS. Despite CMOS being widely used for neural networks, TFTs are an attractive alternative to designing the neuron, as they bring some advantages over CMOS. Some considerations include that a-GIZO-based TFTs are an up-and-coming technology because they present a high degree of transparency and high mobility that translates into good performance. In addition, they are manufactured at room temperature and processed at lower temperatures (about 150ºC), allowing for a greater diversity of substrates, contrary to CMOS. However, it does not have complementary pairs since it was only possible to obtain good performance values for n-type transistors. Consequently, it is necessary to do more manipulations in the circuits to compensate for this limitation.

In the development process of this dissertation, the initial focus was to model the provided TFT to prevent future problems with the transistor operation in a possibly problematic region. After the validation of the model, it was used to develop circuits. Therefore, a neuron circuit was implemented based on an existing one with CMOS technology. During this stage, some options were studied to build integral parts of this circuit with TFTs. Namely, some inverters were proposed to constitute the comparator part of the circuit and an alternative way to implement the refractory period due to TFT limitations. Having reached the final proposal and verifying that the neuron generated the expected impulses in the face of an input current, we proceeded to the next stage of development. As the purpose of the neuron is to be integrated into a neural network, the connecting element between neurons is needed: the synapse. Therefore, a simple synapse was initially proposed that generated an impulse according to the assigned weight. Later, the option arose to use a multiplier to represent the synapse, where the inputs would represent a signal with a specific amplitude and a signal to define the associated weight. As this multiplier has a differential output, there was a need to develop a complementary circuit that generates a single output. Subsequently, the operating range of these connected circuits was verified. In the final stage of development, we proceeded to the *layout* phase of each element chosen to participate in the construction of the neural network.

Agradecimentos

A concretização do trabalho desenvolvido ao longo de todo o projeto de dissertação não seria possível sem o apoio e incentivo de algumas pessoas.

Quero primeiramente agradecer ao meu orientador, Prof. Doutor Vítor Manuel Grade Tavares, por todos os conhecimentos transmitidos e pelo apoio incansável prestado ao longo de todas as etapas.

Agradeço também ao Eng. Guilherme Luís Leitão Teixeira Guia de Carvalho pela disponibilidade e prontidão no esclarecimento de duvidas no decorrer do desenvolvimento da dissertação.

Uma palavra de apreço à minha família e ao meu namorado pela força transmitida e compreensão por todos os momentos em que estive menos disponível.

Também não podia deixar de agradecer aos meus amigos e colegas Diogo Remião, Maria Inês Durão e Pedro Martins, que sempre estiveram presentes no decorrer do projeto.

Margarida Pias

"Well done is better than well said"

Benjamin Franklin

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

AMLCD	Active-Matrix Liquid-Crystal Display
AMOLED	Active-Matrix Organic Light-Emitting Diode
ANN	Artificial Neural Network
CC	Constant Current
CMOS	Complementary Metal–Oxide–Semiconductor
CSV	Comma-Separated Values
DC	Direct Current
DRC	Design Rule Checking
FET	Field Effect Transistor
HH	Hodgkin-Huxley
IF	Integrate-and-Fire
LCD	Liquid Crystal Display
LIF	Leaky Integrate-and-Fire
LTD	Long Term Potenciation
LTP	Long Term Depression
LVS	Layout Versus Schematic
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MSE	Mean Squared Error
NMSE	Normalized Mean Squared Error
SNN	Spiking Neural Network
STDP	Spike-timing-dependent plasticity
TAOS	Transparent Amorphus Oxide Semiconductor
TFT	Thin Film Transistor
VLSI	Very large-scale integration

Ca^{2+}	Ião de Cálcio
K^{+}	Ião de Potássio
Na^{+}	Ião de Sódio

a-GIZO	Amorphous Gallium–Indium–Zinc Oxide
a-Si:H	Hydrogenated Amorphous Silicon
GIZO	Gallium–Indium–Zinc Oxide
ITO	Indium tin oxide
poly-Si	Polycrystalline silicon
Ti/Au	Titanium Gold

Capítulo 1

Introdução

O uso de redes artificiais tem vindo a crescer e podemos perceber isso através de múltiplas aplicações, desde simples atos quotidianos, como desbloquear o telemóvel com reconhecimento facial ou até mesmo para viagens quando fazemos uso de tradução em tempo real.

De forma a perceber o que está por detrás destas aplicações, é necessário esclarecer o conceito de redes neuronais e o que envolvem. O *deep learning* é um subconjunto do *machine learning* e são essencialmente distinguidos na forma como os seus algoritmos de aprendizagem funcionam. No *machine learning* são necessários alguns dados para treinar o algoritmo juntamente com intervenção humana para corrigir os resultados e assim fazer a aprendizagem. Relativamente ao *deep learning*, este utiliza *artificial neural networks* (um tipo de rede neuronal) para reproduzir o processo de aprendizagem do cérebro humano. Para isto, necessita de muitos dados para aprender mas não tem tanta necessidade de intervenção humana como o anterior, pois aprende com erros passados [10]. Estes algoritmos de aprendizagem, podem consumir muita energia e portanto surge a necessidade de ter uma resposta mais eficiente do *hardware* para processar redes neuronais.

1.1 Motivação

O cérebro humano é altamente eficiente em termos de energia, consumindo cerca de 20 W de potência [11]. Para atingir este nível de desempenho e eficiência energética nos circuitos integrados, faz sentido procurar inspiração neste modelo biológico e sua forma de funcionamento.

Os neurónios processam a informação, enquanto que as sinapses fornecem o acesso à memória dos neurónios. Existem vários modelos neuronais que aproximam, de forma bastante realista, o funcionamento do neurónio. Estes modelos tendem a ter estruturas cada vez mais complexas, o que implica a utilização de um número maior de transístores. Quando integramos uma rede neuronal com este tipo de modelos, o circuito integrado tende a ocupar uma grande área, utilizando assim um número muito elevado de transístores. Estes têm um impacto direto no nível de potência, que é superior ao desejável. Por outro lado, o uso de modelos mais simples que consequentemente requerem um número menor de transístores, tornam as redes neuronais escaláveis e também são uma aproximação igualmente boa e funcional do cérebro humano.

A computação neuromórfica tenta perceber dinâmica do cérebro humano e replicá-lo nos sistemas eletrónicos. Atente-se que, não pretende reproduzir fielmente as interações existentes no cérebro, apenas o seu funcionamento e aplicá-lo de uma forma prática com a finalidade de ter uma maior eficiência. Sendo assim, a procura pelo realismo não constitui uma prioridade no desenvolvimento de circuitos. Esta área de pesquisa é bastante promissora, pois se for construído *hardware* com um comportamento que retrata a dinâmica entre as sinapses e neurónios, isto abre a possibilidade de perceber e reagir a eventos do mundo real. Um exemplo disso é o uso de eletrónica neuromórfica na constituição de sensores, permitindo que a informação extraída por estes, possa ser processada no próprio sensor [12].

1.2 Objetivos e definição do problema

De forma a responder ao problema da eficiência energética e ao mesmo tempo ter uma boa capacidade de computação, o objetivo deste projeto é desenvolver um circuito eletrónico analógico de integração em *Very-Large-Scale-Integration* (VLSI), que visa imitar elementos do cérebro, neste caso, o neurónio. O circuito será projetado com Transístores de Filme Fino (TFT).

Como já referido, o desenvolvimento do neurónio é interessante para integrar em redes neuronais. Para encontrar uma solução é necessário rever os neurónios apresentados em tecnologias alternativas com custo-beneficio mais atrativos e que permitam implementações em larga escala.

A implementação de um neurónio com recurso a TFTs permite a utilização de substratos de baixo custo devido à ampla escolha dos mesmos, pelo facto de estes transístores terem uma temperatura de processamento mais baixa. Para além disto, os TFTs têm um processo de fabrico mais simples e possuem potencial para construir grandes redes pois são reconhecidos em eletrónicas de grandes áreas.

1.3 Estrutura da dissertação

A estrutura deste documento está estruturada em cinco capítulos. No presente capítulo, é realizada uma introdução abordando também o contexto, motivação e o problema a resolver. Nos restantes capítulos são abordados vários temas de acordo com a seguinte distribuição:

- No Capítulo 2, são apresentados alguns fundamentos teóricos relativamente à parte biológica do cérebro humano, para uma melhor compreensão da sua dinâmica. Posteriormente, são apresentados alguns modelos neuronais que imitam o comportamento dos neurónios.

- Relativamente ao Capítulo 3, segue-se uma revisão sobre a tecnologia a usar (TFTs), bem como uma apresentação do modelo fornecido. Neste capítulo também é feita uma modelização do TFT, que será utilizado para o desenvolvimento dos circuitos no próximo capítulo.

- O Capítulo 4 começa com um breve enquadramento dos circuitos existentes relativos a cada modelo neuronal visto no Capítulo 2. De seguida, é descrito todo o processo de desenvolvimento dos circuitos dos elementos do cérebro: neurónio e sinapse. Por fim, procedeu-se ao *layout* desses elementos.
- No Capítulo 5 é feita uma conclusão relativa ao trabalho elaborado, bem como a satisfação dos objetivos e possível trabalho futuro.

Capítulo 2

Sistema Neuronal

De forma a ter uma boa compreensão do desenvolvimento do projeto, é necessária a introdução a alguns conceitos relativos ao funcionamento do cérebro. Posteriormente serão vistos alguns modelos que imitam as interações cerebrais aos quais servirão de base para a elaboração dos circuitos.

2.1 Revisão dos Modelos Sinápticos e Neuronais

O cérebro humano é altamente complexo e contém milhares de milhões de células neuronais, designadas de neurónios, que comunicam entre si. A interação entre neurónios é feita ao nível das sinapses. Estas modelam a ação do neurónio pré-sináptico sobre neurónio alvo com o qual comunica (neurónio pós-sináptico) [11].

2.1.1 Neurónio

O neurónio é a unidade base da estrutura do cérebro. Este é responsável pela propagação do impulso e é essencialmente constituído por três partes: dendrites, corpo celular e axónio. A estrutura de um neurónio está representada na Figura 2.1. Quando um neurónio recebe um estímulo, a direção de propagação do sinal dá-se no sentido das dendrites para a ramificação terminal do axónio. No momento em que o impulso chega à ramificação terminal, este é transmitido às dendrites do neurónio seguinte através da sinapse.

Para perceber como acontece a propagação do impulso ao longo do neurónio, é importante saber que a membrana do mesmo é constituída por proteínas que formam canais, representados na Figura 2.2. Sendo assim, existem canais de potássio (K^+), canais de sódio (Na^+) e um canal designado bomba sódio-potássio. O ultimo bombeia 3 iões de Na^+ para fora do neurónio e 2 iões K^+ para o meio intracelular, trabalhando para sustentar o potencial de repouso (aproximadamente -65 m V). Este valor de potencial significa que no meio extracelular encontra-se uma maior quantidade de Na^+ tornando o ambiente fora do neurónio com carga positiva em relação ao interior, que por ter uma menor quantidade de iões positivos considera-se com carga negativa.

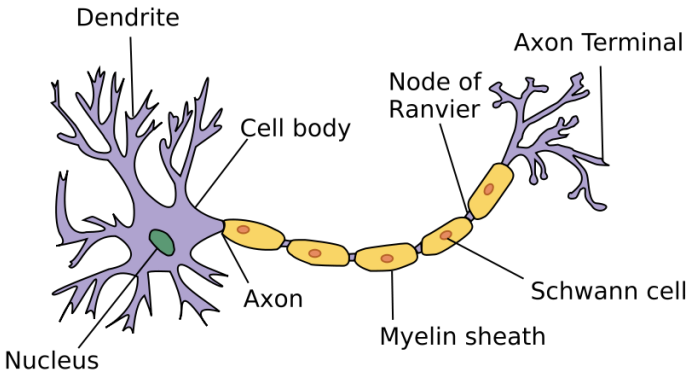


Figura 2.1: Esquema de um neurónio, reproduzido de [1]

O impulso nervoso é transmitido como uma onda que faz uma mudança de polarização na membrana. Inicialmente o neurónio encontra-se no potencial de repouso mantido pela bomba sódio-potássio, como já mencionado. Nesta fase, os iões Na^+ e K^+ podem voltar a mover-se para o meio onde estavam anteriormente, através dos canais de K^+ e Na^+, seguindo o gradiente de concentração. Como a membrana é menos permeável ao Na^+, ou seja, tem menos canais, apenas uma pequena quantidade de Na^+ retorna ao neurónio, mantendo-se a carga positiva fora.

Devido a esta troca de iões entre os meios, quando o valor limite (valor de *threshold*) dentro do neurónio é alcançado (em torno de $-50\,\mathrm{m\,V}$), os canais Na^+ permitem um fluxo contínuo para o ambiente interno aumentando as cargas positivas. Quando o potencial atinge $30\,\mathrm{m\,V}$, os canais K^+ abrem-se por um tempo, restabelecendo-se assim o potencial de repouso. A esta troca de cargas dentro do neurónio, chama-se despolarização (mudança de potencial negativo para positivo) e repolarização (restabelecer o potencial negativo dentro do neurónio).

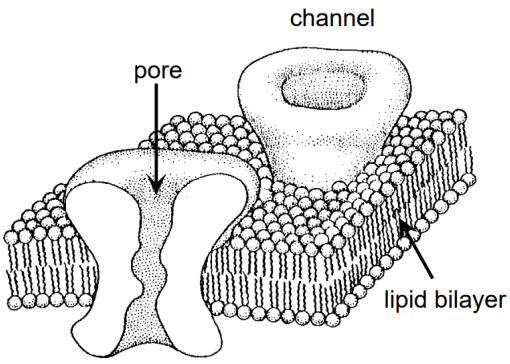


Figura 2.2: Canais na membrana do neurónio, reproduzido de [2]

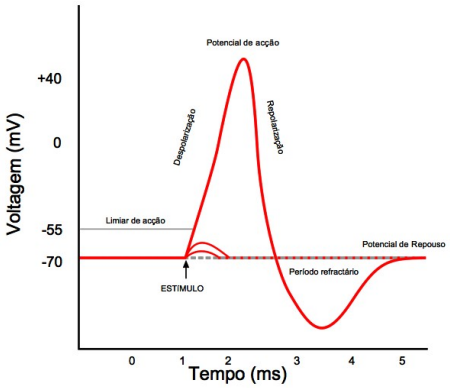


Figura 2.3: Impulso Nervoso, reproduzido de [3]

2.1.2 Sinapse

Existem dois tipos de sinapses: química e elétrica. Na sinapse química, o impulso é sempre transmitido do neurónio pré-sináptico para o neurónio pós-sináptico. Quando o potencial de ação chega aos terminais do axónio do neurónio pré-sináptico, são libertados neurotransmissores na fenda sináptica. Esta fenda é definida como o espaço entre neurónio pré-sináptico e pós-sináptico. Neste local, os neurotransmissores conectam-se a recetores específicos do neurónio pós-sináptico desencadeando o impulso nervoso nesse mesmo neurónio, continuando assim a propagação do sinal.

Na sinapse elétrica, o fluxo de corrente é transmitido de um neurónio para o outro, podendo ocorrer nos dois sentidos (do neurónio pré-sináptico para o pós-sináptico e vice-versa). Este fluxo é feito diretamente de um neurónio para outro através que canais de junção, permitindo a transmissão de uma corrente de iões. Este tipo de sinapses são mais rápidas em comparação com as sinapses químicas, embora possam ocorrer em simultâneo. Geralmente ocorrem quando é necessária uma resposta do corpo muito instintiva e por isso ocorrem com menos frequência.

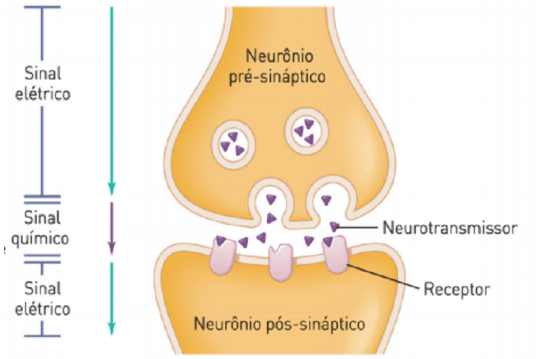


Figura 2.4: Sinapse, reproduzido de [4]

Plasticidade Sináptica

A plasticidade sináptica altera a intensidade sináptica entre neurónios. Se dois neurónios despoletam ao mesmo tempo, a conexão entre eles é fortalecida e, por isso, aumenta a probabilidade de ativarem juntos no futuro. Caso os neurónios despoletem em alturas diferentes, a conexão é enfraquecida e torna-se mais provável ativarem de forma independente.

O fortalecimento da ligação entre neurónios designa-se por *Long Term Potenciation* (LTP) e acontece quando o neurónio pré-sináptico é acionado antes do neurónio pós-sináptico. O enfraquecimento da mesma conexão é chamado *Long Term Depression* (LTD) e ocorre quando neurónio pós-sináptico despoleta antes do neurónio pré-sináptico (Figura 2.5). Esta dinâmica é designada por *Spike-Timing-Dependent Plasticity* (STDP).

Quando o neurónio pré-sináptico despoleta primeiro, há processos químicos e biológicos que aumentam consideravelmente o fluxo de cálcio (Ca^{2+}) no neurónio pós-sináptico. Pelo contrário,

quando o neurónio pós-sináptico despoleta primeiro repercute-se num baixo fluxo do ião Ca^{2+} nesse neurónio.

Podemos concluir então que grandes fluxos de Ca^{2+} levam a LTP e baixos fluxos de Ca^{+} levam a LTD.

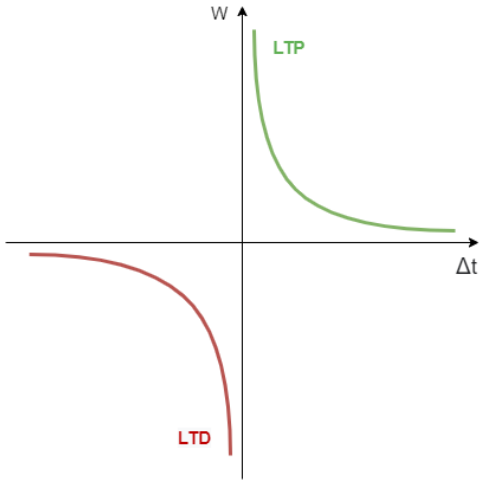


Figura 2.5: Descrição da STDP. Quando o neurónio pré-sináptico despoleta antes do neurónio pós-sináptico $\Delta t > 0$, quando o neurónio pós-sináptico ocorre primeiro que o neurónio pré-sináptico $\Delta t < 0$. W representa o respetivo peso sináptico

Integração de *Spiking Neural Networks*

A dinâmica de transmissão de impulsos nos neurónio inspirou a criação de uma topologia de rede neuronal: *Spiking Neural Network* (SNN). As redes neuronais são compostas por múltiplas camadas de neurónios organizadas. Tradicionalmente a transmissão de informação entre camadas dá-se a cada ciclo de relógio, mas nas SNN a informação é transmitida quando o valor de *threshold* do neurónio é atingido e consequentemente ocorre o impulso.

Estas SNNs melhoram a capacidade de aprendizagem do sistema explorando a plasticidade dependente do tempo de pico (STDP) [10]. Devemos ter em mente que o pico pré-sináptico simboliza o impulso transmitido por um neurónio e o pico pós-sináptico está relacionado com o impulso recebido por um neurónio. Sendo assim, a plasticidade é explorada através do seguinte: a ocorrência de um pico pré-sináptico antes do pico pós-sináptico, aumenta a intensidade sináptica. Ao contrário, um pico pós-sináptico antes do pico pré-sináptico, diminui a força sináptica. Quanto menor o tempo entre os picos, maior a mudança e quanto maior o tempo entre os picos, menor a alteração.

Portanto, podemos ver que os SNNs têm uma grande aproximação com a atividade sináptica real, que é exatamente o que se pretende.

2.2 Modelos Neuronais

Os modelos de um neurónio visam imitar o comportamento dos neurónios corticais, pois estes têm um comportamento bastante diversificado e há vários tipos de neurónios descobertos [13].

Estes modelos descrevem com linguagem matemática a comunicação entre neurónios. Existem vários modelos que simulam diferentes aspetos biológicos dependendo das características que se pretende. Os modelos mais relevantes abordar são o *Integrate-and-Fire*, *Hodgkin-Huxley* e *Izhikevich*, que serão elucidados a seguir.

2.2.1 *Integrate-and-Fire*

Os modelos *Integrate-and-Fire* (IF) são dos mais usados para integrar SNNs. Nestes modelos, consideramos os potenciais de ação como eventos que acontecem num preciso momento no tempo. O propósito não é descrever realisticamente a forma do impulso, mas sim caracterizar o comportamento do potencial de membrana e o como os picos são gerados [14]. Isto torna este tipo de modelos bastante precisos em gerar eventos precisamente cronometrados no tempo (designados pulsos - *spikes*).

O modelo *Leaky Integrate-and-Fire* (LIF) é a versão mais simples, portanto não podemos esperar uma representação completa da bioquímica e biofísica dos neurónios. O LIF é baseado na dinâmica dos canais Na^+ e K^+ explicada anteriormente. O potencial de ação da membrana ocorre quando é atingido o valor de *Threshold* (V_{Th}) e posteriormente, este potencial é restabelecido ao valor de *reset* (V_{reset}). Esta aproximação combina todas as condutâncias ativas da membrana numa única fuga passiva, resultando no circuito elétrico representado por Figura 2.6 [2].

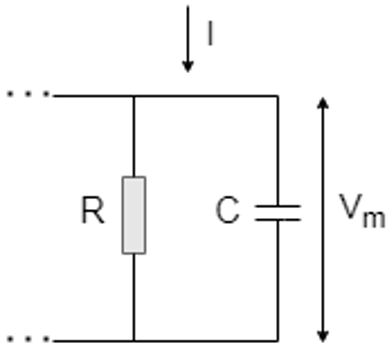


Figura 2.6: O neurónio que recebe uma corrente de entrada I e a membrana age como um condensador C em paralelo com uma resistência R

As equações do LIF baseadas na lei das correntes de *Kirchhoff*, são representadas por:

$$\tau_m \frac{dV_m}{dt} = E_L - V_m + R_m I \tag{2.1}$$

Quando $V_m \geq V_{Th}$

$$V_m = V_{reset} \tag{2.2}$$

A variável E_L representa o potencial de repouso, que tem o mesmo valor que o *reset* (V_{reset}), perto de $-65\,\mathrm{m\,V}$. Quando o potencial de membrana V_m atinge a tensão de limiar V_{Th} correspondente a $-50\,\mathrm{m\,V}$, a tensão V_m volta a ter $-65\,\mathrm{m\,V}$. O valor da resistência da membrana R_m é tipicamente $10\,\mathrm{M\Omega}$ e a constante de tempo da membrana τ_m é $10\,\mathrm{m\,s}$.

Podemos determinar o potencial de membrana resolvendo a equação (2.1)

$$V_m[t] = V_m[t-1] + \frac{E_l - V_m[t-1] + IR_m}{\tau_m} dt \tag{2.3}$$

Esclarecido o que é cada elemento das equações, a simulação resultante de (2.2) e (2.3) é representada na Figura 2.7. Podemos ver então que, como já descrito, o potencial de membrana começa em $-65\,\mathrm{m\,V}$ e vai aumentando até atingir a tensão de *threshold* onde sofre um *reset* ao valor inicial.

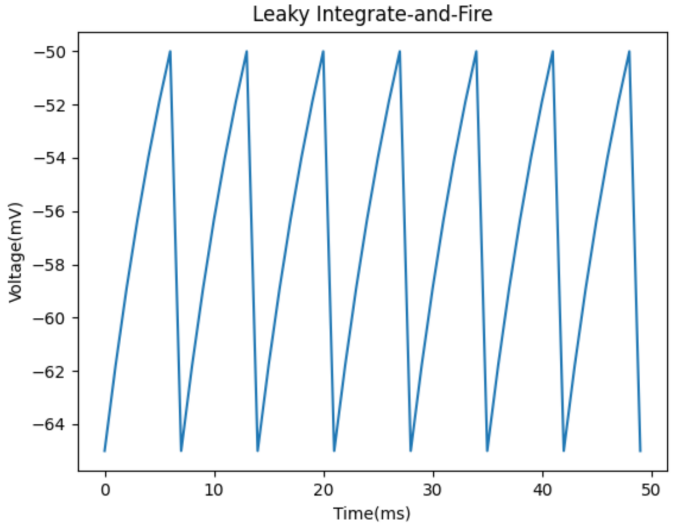


Figura 2.7: Simulação do modelo LIF

O LIF até agora não tem em conta a refratariedade ou adaptação de longa duração, uma vez que, descrevemos o modelo com uma aproximação muito simplificada. O potencial de ação é reduzido a uma abordagem muito simples e o potencial de membrana tem uma aproximação a um comportamento linear. A aproximação simplificada pode ser mantida para situações onde não é importante detalhar o que acontece durante o potencial de ação, mas é possível adaptar a dinâmica

deste modelo de forma a ter capacidade para prever com precisão os tempos de picos dos neurónios corticais.

2.2.2 *Hodgkin-Huxley*

O modelo *Hodgkin-Huxley* (HH) é um modelo mais realista comparando com o IF, pois podemos simular uma variedade de dinâmicas do comportamento dos neurónios. Devido a isto, o HH acaba por ser mais complicado logo implica um circuito mais complexo que requer um maior número de transístores. Este modelo é baseado nas condutâncias da membrana, e por isso, descreve 3 correntes principais presentes: correntes de K^+, Na^+ e outras correntes em quantidades menores que são agrupadas numa única designada corrente de fuga. Estas correntes atravessam a membrana através dos canais de K^+, Na^+ e fuga (*leakage*), respetivamente. A formação do potencial de ação pode ser representada pela sua relação com a corrente da membrana descrita como a soma das correntes de fuga, K^+ e Na^+.

$$I_m = \overline{g}_L(V - E_L) + \overline{g}_K n^4 (V - E_K) + \overline{g}_{Na} m^3 h (V - E_{Na}) \tag{2.4}$$

Para determinar a corrente que atravessa os canais na membrana é preciso especificar as mudanças que ocorrem nas condutâncias de cada canal $\overline{g}$. Para isto foram associadas as variáveis m, n e h necessárias a cada $\overline{g}$, como podemos perceber na equação (2.4).

O m consiste na ativação da condutância de Na^+, o h representa o grau da desativação da condutância de Na^+, por último, o n retrata a ativação da condutância de K^+.

Para um x que representa m, n ou h

$$\tau_x(V)\frac{dx}{dt} = x_\infty(V) - x \tag{2.5}$$

onde

$$x_\infty(V) = \frac{\alpha_x(V)}{\alpha_x(V) + \beta_x(V)} \tag{2.6}$$

Os parâmetros α e β propostos por [2] para cada m, n e h têm as seguintes expressões.
Para n:

$$\alpha_n = \frac{0.01(V + 55)}{a - e^{-0.1(V+55)}} \tag{2.7}$$

$$\beta_n = 0.125e^{-0.0125(V+65)} \tag{2.8}$$

Para m:

$$\alpha_m = \frac{0.1(V + 40)}{a - e^{-0.1(V+40)}} \tag{2.9}$$

$$\beta_m = 4e^{-0.0556(V+65)} \tag{2.10}$$

Para h:

$$\alpha_h = 0.07e^{-0.05(V+65)} \tag{2.11}$$

$$\beta_h = \frac{1}{1+e^{-0.1(V+35)}} \tag{2.12}$$

Para simulação foram usados os valores de $\overline{g}_L = 0{,}003\,\mathrm{mS/mm^2}$, $\overline{g}_K = 0{,}036\,\mathrm{mS/mm^2}$, $\overline{g}_{Na} = 1{,}2\,\mathrm{mS/mm^2}$, $E_L = -54{,}402\,\mathrm{mV}$, $E_K = -77\,\mathrm{mV}$ e $E_{Na} = 50\,\mathrm{mV}$ [2]. O resultado deste modelo através da computação das equações (2.4) a (2.12) é apresentado na Figura 2.8.

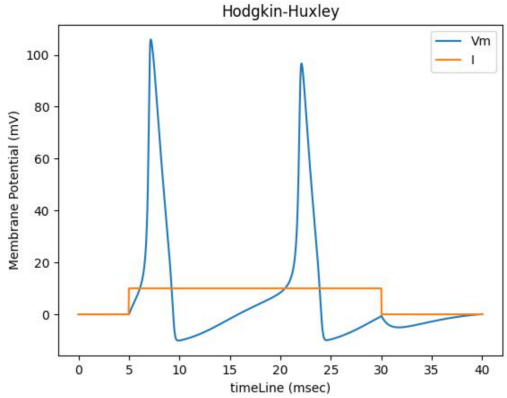


Figura 2.8: Simulação do modelo HH

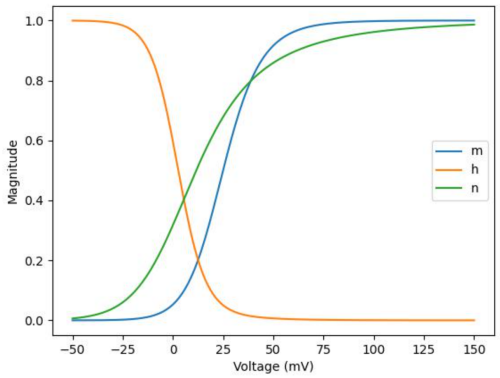


Figura 2.9: Comportamento das variáveis m, h e n durante a ocorrência do potencial de ação

Analisando o gráfico da simulação do modelo HH (Figura 2.8), podemos perceber que quando é injetada uma corrente positiva I, esta despoleta o potencial de ação.

Apesar de não estar explícito, o reestabelecimento do potencial de repouso requer que os valores de m, h e n (Figura 2.9) sejam repostos ao valor que inicialmente tinham.

2.2.3 *Izhikevich*

O modelo *Izhikevich* apresentado em [5] é uma opção que combina os dois modelos apresentados anteriormente. Esperamos que um modelo neuronal seja o mais realista possível, mas ao mesmo tempo que seja computacionalmente simples. O modelo proposto por *Izhikevich* é plausível como o HH, mas eficiente como o IF. Para além disto, representa a forma dos impulsos neuronais biologicamente bem aproximados, pois é oferecida a possibilidade de alterar o valor de uns parâmetros, que modelam o impulso ao desejado.

Este modelo é caracterizado por duas equações diferenciais (2.13) e (2.14) e um mecanismo de *reset* após o pico do impulso.

$$v' = 0.04v2 + 5v + 140 - u + I \tag{2.13}$$

$$u' = a(bv - u) \tag{2.14}$$

Se $v \geq 30$ mV então

$$v = c \tag{2.15}$$

$$u = u + d \tag{2.16}$$

Este modelo pode reproduzir vários padrões de impulsos (Figura 2.10) dos neurónios corticais apenas alterando os valores de a, b, c e d. O v representa o potencial de membrana do neurónio e o u é variável de recuperação da membrana responsável por ativar os canais de K^+ e desativar os canais de Na^+. A variável I representa as corrente injetada (corrente sináptica).

Tendo em conta isto, o valor a descreve o tempo da variável de recuperação u, o parâmetro b a sensibilidade de u, o c representa o valor de *reset* da membrana após o pico e por ultimo o d descreve o u após o pico.

Uma breve descrição de cada padrão de disparo é apresentado em [5].

2.3 Conclusões

Como o tema da dissertação está em torno do desenvolvimento de um neurónio analógico para integração de redes neuronais, neste capítulo foram abordados alguns conceitos acerca do cérebro e a sua dinâmica. Isto permite perceber de que forma o neurónio e a sua interação com outros neurónios por meio de sinapses podem servir de inspiração para a construção de *Spiking Neural Networks*.

Depois desta revisão, foram estudados alguns modelos maioritariamente usados para reproduzir o comportamento do neurónio. O primeiro modelo abordado foi o *Integrate-and-Fire*, que é dos mais usados para integrar redes neuronais pela sua simplicidade e por isso não descrevem de forma muito realista o impulso nervoso. A variação mais simples dentro deste tipo de modelos é o *Leaky Integrate-and-Fire*.

De seguida foi visto o modelo *Hodgkin-Huxley*, que é bastante realista mas acaba por ser mais complicado implementar resultando num circuito bastante complexo.

O último modelo é o proposto por *Izhikevich*. Este modelo combina os dois modelos anteriores, pois esperamos que um modelo neuronal seja o mais realista possível (como o caso do HH) mas ao mesmo tempo seja computacionalmente simples (como o IF).

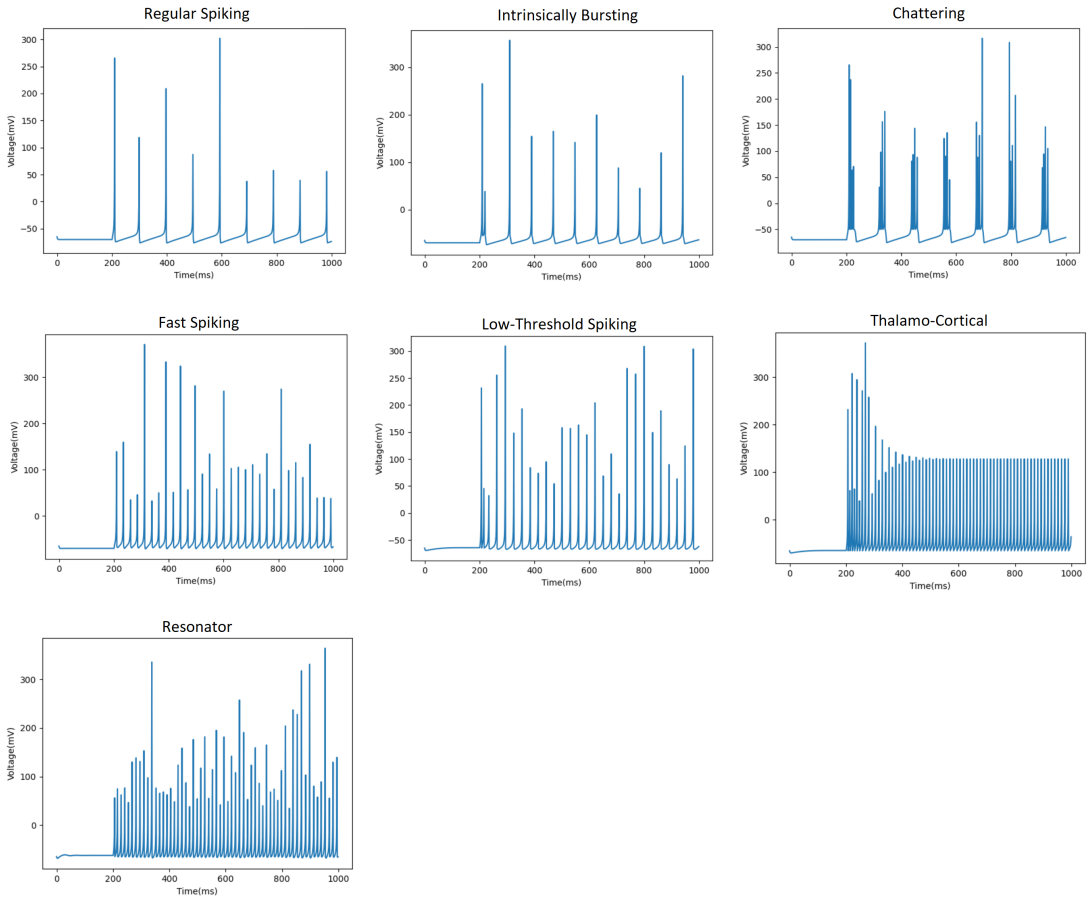


Figura 2.10: Simulação do modelo *Izhikevich* para vários parâmetros de a, b, c e d propostos em [5]

Capítulo 3

Introdução a Tecnologias de Filmes Finos e Modelização de Transístores

Os *Thin Film Transistors* (TFTs) a serem utilizados nesta dissertação, constituem um dos temas principais para o desenvolvimento do circuito do neurónio. Inicialmente, foi necessário ver vários circuitos dos modelos neuronais estudados de forma a organizar informação e ter um melhor entendimento de como implementar um neurónio. Durante este processo, foram alvo de estudo várias manipulações dos modelos neuronais e secções de cada circuito que possam ter relevância para a fase de desenvolvimento. No entanto, pôde-se perceber que apesar de haver muita variedade nos circuitos encontrados, apenas uma pequena quantidade incluía o uso de TFTs. Na maioria dos circuitos encontrados, os neurónios estavam implementados com tecnologia *Complementary Metal–Oxide–Semiconductor* (CMOS), que apesar de não ser o que se pretende, não se descartavam para futura inspiração, pois já tinham um funcionamento mais completo e bem estudado.

3.1 *Thin Film Transistors*

A tecnologia CMOS é bem conhecida e muito utilizada, mas TFTs semicondutores baseados em óxido trazem algumas vantagens sobre CMOS que devem ser consideradas.

Os TFTs são conhecidos pela sua dominância no mercado de *Liquid Crystal Displays* (LCDs), onde cada pixel é controlado por um TFT permitindo grandes resoluções. Isto deve-se ao facto de estes serem transístores muito pequenos (porém bastante maiores que os *CMOS*), acrescentando que também por isso, a energia para os controlar é proporcionalmente reduzida, trazendo grande eficiência energética para estes monitores. Este tipo de transístores também são usados em sensores de imagens radiográficas [15] e mais recentemente nos ecrãs AMOLED, que incluem uma camada de TFTs na sua composição.

O substrato tipicamente usado para TFTs é o vidro pela sua popular aplicação em LCDs, mas há necessidade de um processo dedicado para depositar o semicondutor pois o vidro não tem propriedades elétricas para a operação do dispositivo. No caso de MOSFETs este processo não é necessário uma vez que normalmente é usado uma *wafer* de silício como substrato, que também

serve de semicondutor. Contudo, dado que as temperaturas de processamento para CMOS são muito elevadas, não permitem tanta variedade de substratos como vidro ou plástico, como no caso de TFTs.

3.1.1 Evolução

Um dos primeiros TFTs que surgiu tinha como semicondutor o silício amorfo hidrogenado (a-Si:H). Este material, apesar de ter uma mobilidade baixa ($\mu = 1\,\mathrm{cm}^2/\mathrm{Vs}$), a sua estrutura amorfa permitia a fabricação para grandes áreas o que juntamente com a capacidade de rápida comutação destes transístores, criou grande impacto pela sua aplicação em LCDs, nomeadamente em *Active-Matrix LCD* (AMLCD) [16].

Na procura por uma mobilidade melhor, surgiram TFTs com silício policristalino (poly-Si) como semicondutor. Sendo assim, este material permitia implementar circuitos com elevado desempenho, mas com um elevado custo de produção e temperatura de processamento muito alta. Aqui apareceram também os TFTs orgânicos, que resolviam a questão da temperatura de fabricação nos TFTs de poly-Si, pois não necessita de temperaturas elevadas de processamento, mas em contrapartida os dispositivos não possuíam grande estabilidade nem desempenho [16].

Mais tarde surgiram os TFTs baseados em óxido que trouxeram grandes vantagens como implementação para grandes áreas, baixas temperaturas de processamento, bom desempenho elétrico e ainda ofereciam a possibilidade de transparência. Estes transístores surgiram porque grande parte dos TFTs não são transparentes, apenas têm partes que os constituem que podem ser. Sendo assim, houve a procura pela transparência a partir de materiais Semicondutores de Óxido Amorfo Transparente (TAOS), até que surgiu o primeiro TFT transparente baseado em óxido de zinco, produzido entre 450 a 600°C [16]. Na procura por outros materiais com melhor mobilidade que o silício amorfo tipicamente usado como semicondutor, surgiu a proposta por *Nomura* de um componente de óxido amorfo designado de gálio índio óxido de zinco (GIZO) [17]. Este material era produzido inicialmente a 1400°C, exibindo grande mobilidade quando integrado na estrutura do TFT. Posteriormente conseguiu-se fabricar GIZO a temperatura ambiente, que resultou num material amorfo (a-GIZO) [16].

3.1.2 *TFTs* com Semicondutores de Óxido Amorfo Transparente

Os TFTs baseados em TAOS, em particular a-GIZO, apresentam elevado grau de transparência e mobilidade elétrica alta para este tipo de transístores ($\mu > 10\,\mathrm{cm}^2/\mathrm{Vs}$) que se traduz num desempenho elétrico muito eficiente. Este material tem tido especial atenção como alternativa a a-Si:H e poly-Si especialmente para monitores de grande área AMOLED, pois como tem uma mobilidade cerca de 10 vezes maior que o a-Si:H, tem mais facilidade em fornecer as correntes necessárias e permite a comutação mais rápida dos "*pixels*". Contudo, tem uma mobilidade menor que poly-Si, mas em contrapartida apresenta melhor uniformidade por ser um material amorfo [18].

A alta mobilidade neste material amorfo é uma consequência da sua estrutura de ligação iónica, que também proporciona uma baixa densidade de estados de aprisionamento próximo da fronteira

da banda de condução, permitindo uma ativação destes transístores mais rápida do que os TFTs de a-Si:H [18].

Isto torna esta tecnologia bastante promissora, pois estes transístores são fabricados à temperatura ambiente, como já referido, onde a deposição de a-GIZO é feita através de pulverização contínua e processados a temperaturas mais baixas (cerca de 150°C). Sendo assim, isto permite uma ampla escolha de substratos de construção, como vidro e uma diversidade de materiais.

3.1.3 Estrutura

Os TFTs são um tipo de transístores de efeito de campo (FETs) que têm como foco o processamento de grandes áreas e baixas temperaturas, contrariamente aos MOSFETs que tem como objetivo o alto desempenho mas com custo de processamento maior e a altas temperaturas.

Relativamente à estrutura dos TFTs, a maior diferença em relação a MOSFETs reside no facto de que nos MOSFETs o substrato usado, tipicamente, é uma *wafer* de silício que ao mesmo tempo serve de semicondutor. No caso dos TFTs, geralmente é usado um substrato não condutor como o vidro.

Os TFTs podem ter várias estruturas dependendo da posição das camadas e a localização da porta ser em cima ou em baixo. Desta forma, se a fonte e o dreno estiverem do mesmo lado que a porta em relação ao semicondutor, então terão um estrutura *coplanar*, caso contrário designa-se por *staggered*. Se a porta estiver no topo então será *top-gate* senão então denomina-se *bottom-gate*. As diferentes estruturas podem ser consultadas na Figura 3.1.

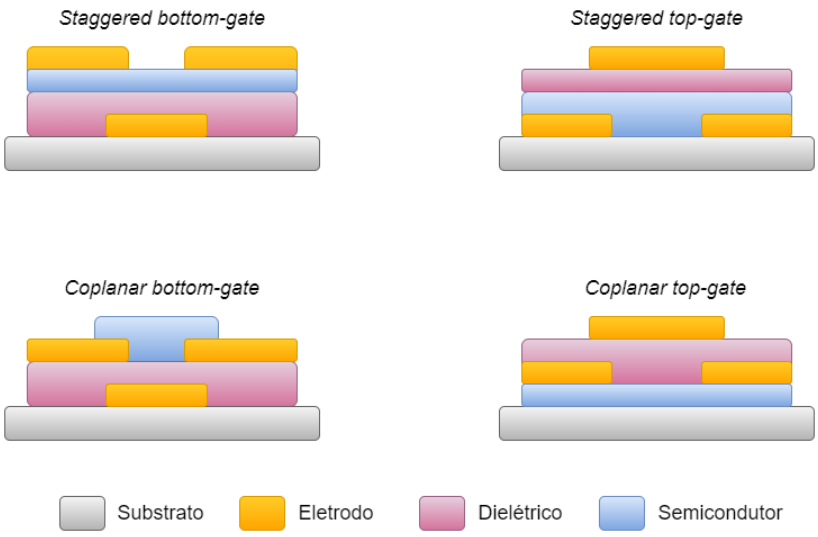


Figura 3.1: Estruturas do TFT

Cada configuração tem as suas vantagens e desvantagens e a escolha de uma estrutura para fabricação depende dos processos de deposição bem como temperaturas de processamento e o numero de mascaras litográficas [16]. Geralmente quando o semicondutor ou o dielétrico são feitos de materiais que necessitam de maiores temperaturas, são escolhidas configurações que

tenham essas camadas no nível mais baixo possível. Por exemplo, quando a camada de dielétrico necessita de temperaturas elevadas, normalmente é escolhida a *staggered bottom-gate*.

3.1.4 Funcionamento

Quando o dispositivo está em corte, pode ser considerado que está desligado, porque a quantidade de corrente entre a fonte e o dreno é desprezável. Isto acontece quando a tensão de v_{GS} aplicada é inferior à tensão mínima de v_{GS} que garante a acumulação mínima de cargas para formar um canal (v_{ON}), pois apenas pode existir um fluxo de corrente se existir um canal que o permita. Para a formação deste canal, a fonte pode ser ligada ao terra, a porta tem de ser polarizada com uma tensão positiva e é necessário aplicar uma tensão no dreno. Com isto, é alcançada uma tensão positiva de v_{DS}, que na prática provoca uma libertação de eletrões do elétrodo da fonte para a camada onde é formado o canal e depois esses eletrões são extraídos no dreno. Este transporte de cargas entre a fonte e o dreno que formam um canal que é responsável por originar um fluxo de corrente I_{DS}.

Se for verificada a condição que promova a formação de canal, então o transístor pode operar em duas regiões. Caso o $v_{DS} < v_{GS} - V_T$ então o TFT está a operar em tríodo e o seu comportamento pode ser descrito por:

$$I_{DS} = C\mu \frac{W}{L}\left[(v_{GS} - V_T)v_{DS} - \frac{v_{DS}^2}{2}\right] \tag{3.1}$$

A variável C representa a capacitância da porta, μ representa a mobilidade dos eletrões no canal enquanto que W e L representam a largura e comprimento do transístor, respetivamente. Nesta região o canal existe em toda a extensão entre o dreno e a fonte do dispositivo.

Na situação em que $v_{DS} > v_{GS} - V_T$, o transístor está na região de saturação e a sua corrente é dada por

$$I_{DS} = \frac{1}{2}C\mu \frac{W}{L}(v_{GS} - V_T)^2 \tag{3.2}$$

Relativamente a esta região, o canal não existe junto ao dreno, contrariamente ao anterior. Conforme o aumento de v_{GS}, a parte do canal próximo ao elétrodo do dreno sofre um estreitamento, pois deixa de ter portadores suficientes de eletrões e por isso, a corrente torna-se constante pois o canal, idealmente, não se altera com o aumento de v_{DS}. Isto acontece porque a diferença de potencial entre o canal e o dielétrico não é alta o suficiente para manter uma camada de inversão

3.1.5 Materiais

Embora a estrutura tenha influência no desempenho do transístor, não é o único fator a ter em consideração. Os materiais juntamente com os processos de deposição têm grande responsabilidade no desempenho. Não é apenas pensado como depositar as camadas, mas também as proporções dos materiais que compõem o semicondutor, bem como a sua espessura e a estrutura

do transístor. Os materiais escolhidos para elétrodo e o dielétrico também constituem uma parte importante.

Como já referido, a descoberta do material a-GIZO trouxe grandes oportunidades para os TFTs. Para tecnologias como a-GIZO, TFTs é altamente relevante a escolha de um dielétrico com uma constante dielétrica (k) alta. Isto porque, como estes transístores são produzidos a temperaturas mais baixas estão mais propensos a uma densidade de defeitos maior e uma capacidade de porta reduzida, que pode ser compensada com uso de um dielétrico com alto k [16].

Relativamente aos elétrodos, no artigo [19] é apresentado que para os elétrodos de dreno/fonte que usam o material Ti/Au, apresentam um bom desempenho, mas não apresentam transparência, enquanto que elétrodos ITO, apesar de serem transparentes, têm desempenhos mais fracos comparado com os anteriores.

3.1.6 Limitações

Relativamente aos TFTs baseado em a-GIZO, apesar de apresentarem aspetos muito positivos para a implementação de circuitos, há uma parte menos vantajosa, pois apenas é possível obter bons valores de desempenho em transístores do tipo n. Consequentemente, mais manipulações de circuito são necessárias para compensar a falta de transístores do tipo p [20]. Por sua vez, os circuitos integrados implementados com estes TFTs são mais lentos e consomem mais energia do que os circuitos de pares complementares disponíveis com TFTs de poly-Si [18].

A mobilidade afeta o I_{DS} máximo e a velocidade de comutação, portanto é um parâmetro que deve ser considerado. Existem algumas formas de determinar o μ, entre as quais (3.4) para valores baixos de v_{DS} [16]. Deve ser notado que a mobilidade afeta a transcondutância (g_m) dos transístores, nomeadamente, o g_m aumenta com o aumento da mobilidade.

$$g_m = \mu_{FE} C \frac{W}{L} v_{DS} \tag{3.3}$$

$$\mu_{FE} = \frac{g_m}{C \frac{W}{L} v_{DS}} \tag{3.4}$$

Apesar de os TFTs de a-GIZO terem uma mobilidade alta, esta é mais baixa comparado a CMOS. Isto tem repercussão no ganho obtido dos circuitos, pois é mais fácil ter ganhos elevados com CMOS do que com TFTs. Uma forma de controlar o ganho em TFTs pode ser feito através da implementação de uma carga, normalmente, ativa.

O uso de uma resistência para este efeito não é uma opção viável de implementar, pois como é necessário uma resistência grande, ocuparia uma grande área no circuito. Como não está disponível o dispositivo complementar (geralmente usado para circuitos com altos ganhos), uma alternativa é implementar a carga com o transístor do tipo n com realimentação positiva entre a fonte e a porta.

3.2 Modelização do *Thin Film Transistor*

Numa primeira fase há uma grande importância em avaliar a tecnologia a ser utilizada para o desenvolvimento do circuito do neurónio. Sendo assim, nesta parte do desenvolvimento do projeto, foi estudado o modelo de TFT fornecido.

3.2.1 Modelo Inicial

Quando desenhamos no *Cadence Virtuoso* um circuito com o modelo do TFT providenciado, este é internamente descrito com uma fonte de corrente entre o dreno e a fonte, de acordo com uma função dependente das tensões v_{DS} e v_{GS} ($f(v_{GS}, v_{DS})$). Esta função, juntamente com os valores das tensões à qual depende, são usados por uma rede neuronal artificial (ANN) para produzir o i_D. Sendo assim, este algoritmo ANN tem a função de encontrar o melhor mapeamento entre os dados experimentais e os dados que esperados de i_D, processo designado por treinamento da ANN (Figura 3.2).

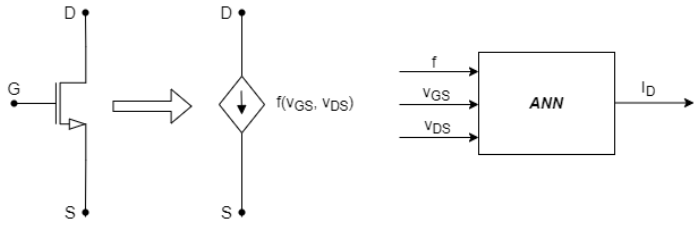


Figura 3.2: Modelo do TFT

Na observação da curva caraterística i_D-v_{GS}, verificou-se um problema com este modelo, os cruzamentos por zero não são bem definidos. Apesar de a distância ser pequena, o cruzamento dá-se um pouco antes de zero. Isto significa que a ANN tem dificuldade em impor zero quando o $v_{DS} = 0\,\text{V}$ e $v_{GS} = 0\,\text{V}$ ou quando $i_D = 0\,\text{A}$ e $v_{GS} = 0\,\text{V}$.

Para maioria dos circuitos analógicos, isto não é considerado um problema, uma vez que operam afastados da região problemática. No que toca a circuitos digitais e circuitos *on-off*, por vezes pode haver problemas de convergência. Como no género de circuitos que vão ser implementados é comum o uso de interruptores, e por isso os transístores por vezes têm de estar ligados e noutras ocasiões desligados, podem surgir os problemas de convergência mencionados. De forma a evitar estas complicações, o objetivo passa por modelizar de forma diferente, a fonte de corrente que descreve o transístor. Para isto, recorreu-se a outro modelo denominado de modelo *Alpha-Power Law*, para modelização do TFT, pois à partida este passa em zero na região explicada anteriormente, permitindo que o transístor opere nessa região sem problemas. Por outro lado, o modelo em si é mais compacto. Para o efeito, primeiramente foram extraídas as curvas características I-V do transístor e posteriormente realizou-se uma aproximação das equações que descrevem o modelo *alpha-power law* às curvas com o valores reais do modelo.

3.2.2 Extração das Curvas Características I-V

De modo a obter as curvas que caracterizam o comportamento do TFT, isto é, as curvas que relacionam a corrente no transístor (i_D) com a tensão entre a porta e a fonte (v_{GS}) e a tensão entre o dreno e a fonte (v_{DS}), foi feito um esquema no *Cadence Virtuoso*, representado na Figura 3.3.

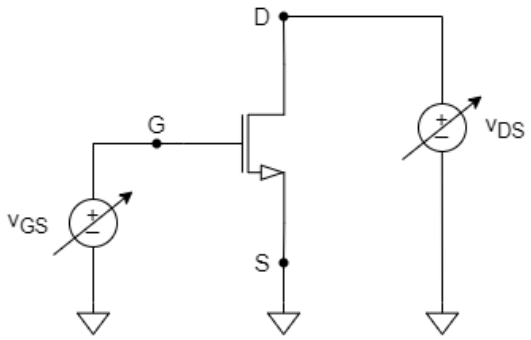


Figura 3.3: Esquemático para extrair curvas caraterísticas I-V

Uma vez que o modelo fornecido estava validado, foram fornecidos alguns valores muito relevantes para avaliar o TFT. Na operação do transístor, este deve ter valores de v_{GS} de 0 a 5 V e v_{DS} de 0 até 10 V .

Perante isto, usando o simulador do *Cadence Virtuoso*, foi realizada uma simulação onde era medida a corrente i_D em função do v_{DS} para múltiplos valores de v_{GS} (Figura 3.4) e também uma simulação com a relação de i_D e v_{GS} para várias tensões de v_{DS} (Figura 3.5).

Posteriormente, os valores dos gráficos foram extraídos para um tipo de ficheiro *Comma-Separated Values* (CSV), com intenção de serem utilizados na etapa seguinte.

3.2.3 Modelo *Alpha-Power Law*

O modelo *alpha-power law* foi proposto por *Sakurai* e *Newton* [21], para estimar a corrente no transístor (i_D). Este modelo concede mais um parâmetro α, que representa o índice de velocidade de saturação. A variável extra oferece flexibilidade para controlar o formato da curva permitindo assim uma boa aproximação aos valores obtidos.

De acordo com artigo mencionado, o comportamento deste modelo é descrito de acordo com as seguintes equações.

Para a região de saturação temos

$$i_D = i_{DSAT}(1+\lambda v_{DS}) \tag{3.5}$$

A região linear é descrita como

$$i_D = i_{SAT}(1+\lambda v_{DS})\left(2-\frac{v_{DS}}{v_{DSAT}}\right)\frac{v_{DS}}{v_{DSAT}} \tag{3.6}$$

Temos também que

$$v_{DSAT} = K(v_{GS}-V_T)^m \tag{3.7}$$

$$i_{DSAT} = \frac{W}{L_{EFF}}B(v_{GS}-V_T)^\alpha \tag{3.8}$$

Não esquecendo a condições que separam as regiões de operação do transístor são caraterizadas por

$$\begin{aligned} v_{GS} &\leq V_T && \text{, região de corte} \\ v_{DS} &< v_{DSAT} && \text{, região de tríodo} \\ v_{DS} &\geq v_{DSAT} && \text{, região de saturação} \end{aligned}$$

3.2.4 Aproximação do Modelo *Alpha-Power Law* ao *Thin Film Transistor*

Esta parte da modelização foi desenvolvida em *Python* baseada no modelo *alpha-power law* da referência [22] e é constituída por 3 etapas principais: aproximação na região de saturação, aproximação na região linear e junção das aproximações das diferentes regiões.

Sendo assim, os valores guardados no CSV na etapa de extração da curva caraterística I-V, foram usados para fazer uma aproximação das equações (3.5) e (3.6) à forma da curva que era determinada pelos valores extraídos.

Como se pode reparar, a equação que descreve a região linear (3.6) tem mais variáveis desconhecidas comparado à equação da região de saturação (3.5), entre as quais incluem o V_T, α, B, λ, K e m. Desta forma, há uma vantagem maior em começar a aproximação pela equação correspondente à saturação, pois assim já seriam conhecidas mais variáveis quando se fizesse a aproximação da equação da região de tríodo.

Como a corrente em saturação depende apenas de v_{GS}, essa dependência é melhor representada pelas curvas características i_D-v_{GS}. Assim sendo, através de uma regressão linear, foi feita a aproximação da equação (3.5) à região de saturação das curvas i_D-v_{GS}, isto é, para tensões de v_{DS} superiores a 4 V (Figura 3.4). Um ponto a ter em conta para a melhor aproximação possível é garantir que a aproximação seria comum a todas as curvas, para não obter diferentes valores das variáveis para diferentes curvas. Os resultados desta aproximação para as variáveis V_T, α, B e λ estão representados na Tabela 3.1. A estimativa de V_T é feita de forma própria e será discutida na secção seguinte. Deve ser mencionado que nas características do transístor W e L, foram usados os valores 320 µm e 20 µm, respetivamente, que correspondiam às medidas predefinidas pelo modelo fornecido.

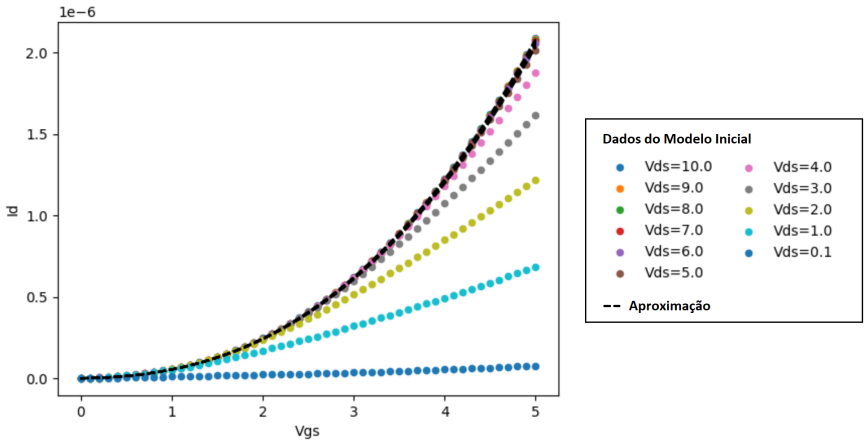


Figura 3.4: Aproximação na região de Saturação. Na regressão foram usados apenas os valores para $v_{DS} > 5\,\text{V}$ por estar garantida a saturação do transístor em toda a gama de v_{GS} considerado

Neste ponto, já eram conhecidos alguns valores, por isso o passo seguinte seria fazer uma aproximação da equação na região de tríodo com as variáveis conhecidas, para extrair as que estavam em falta (K e m).

Deste modo, a equação (3.6) juntamente com os valores obtidos, foi aproximada às curvas i_D-v_{DS} mas apenas na região linear. Assim, foram dados um número específico de pontos para cada v_{DS} aos quais a aproximação deveria ser feita para esta ser o mais precisa possível, resultando no gráfico da Figura 3.5. Os valores de K e m obtidos nesta fase podem ser consultados na Tabela 3.1.

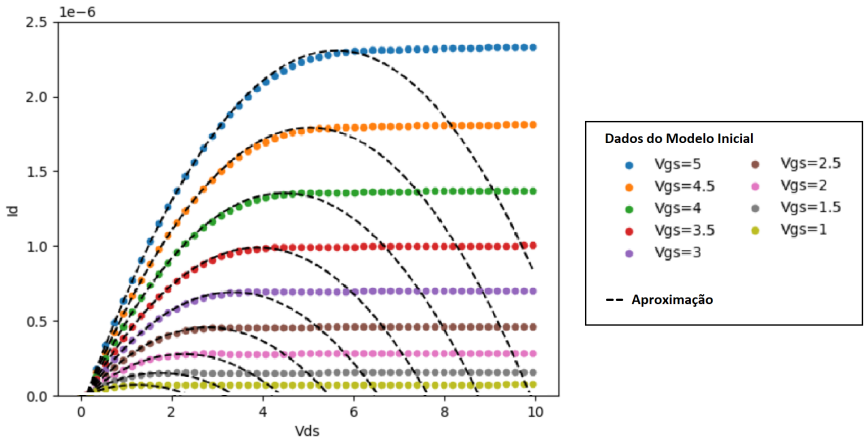


Figura 3.5: Aproximação na região de Tríodo

Agora que temos todos os valores das variáveis das equações do modelo *alpha-power law*, a equação característica I-V final está ilustrada na Figura 3.6. Para isto, apenas foi necessário definir

Tabela 3.1: Resultados da extração dos parâmetros na modelização do *TFT*.

$V_T(V)$	α	B	$\lambda(V^{-1})$	K	m
-0.2286	2.5091	$2.0027e^{-7}$	$2.5562e^{-3}$	0.8625	1.1249

as condições que separam a região linear da de saturação. Caso o v_{GS} fosse menor que V_T ou v_{GS} inferior ao v_{DSAT}, então seria usada a equação da região linear (3.6), caso contrário seria a equação de saturação (3.5).

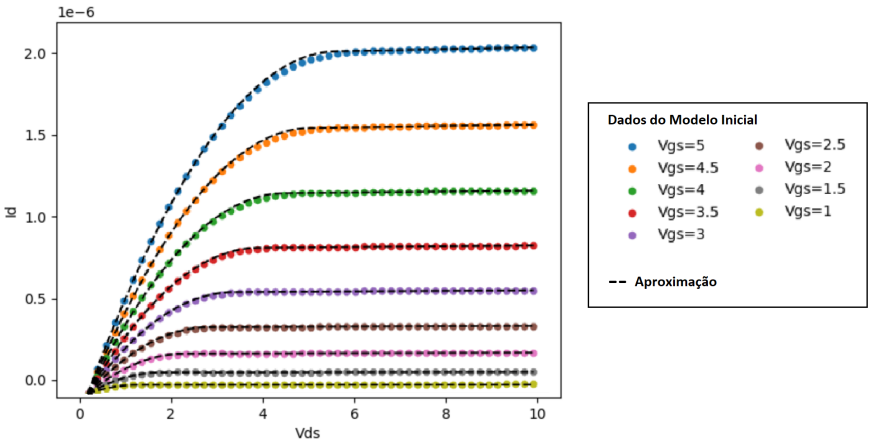


Figura 3.6: Aproximação do modelo *alpha-power law* aos valores extraídos da curva caraterística I-V do TFT

Por fim, depois de extraídos os parâmetros das equações que descrevem o modelo do transístor, surgiu a necessidade de perceber se estes valores retratam eficientemente o modelo pretendido, pois as equações, condições e parâmetros serão transcritos para *Verilog-A* para descrever o TFT. Posteriormente, este será usado para desenvolver os circuitos pretendidos, por isso é necessário ter a máxima confiança no modelo prevendo que se comportará como esperado pela modelização, caso contrário terão de ser feitas as correções necessárias.

3.3 Extração do V_T

A tensão de limiar (V_T), é um parâmetro importante na modelização de transístores, uma vez que é o que melhor descreve a operação dos mesmos, por isso torna-se fundamental para avaliar a confiabilidade da aproximação. Este fator pode ser referido como a tensão mínima necessária na porta do transístor que permite um fluxo de corrente ($I_D \neq 0\,\text{A}$), dando assim o ponto de transição entre o funcionamento do transístor em acumulação fraca e acumulação forte.

3.3.1 Métodos de Extração do V_T

Existem várias técnicas de extração do parâmetro V_T desenvolvidas para MOSFETs, nos quais grande parte dos procedimentos baseiam-se na medição de valores a partir da curva caraterística i_D-v_{GS}. Como a informação sobre esta relação da corrente no dreno e a tensão na porta foi anteriormente extraída, de modo a dar continuidade ao trabalho, o método de extração do V_T é feito através destas curvas. Como estamos a procurar uma extração do V_T de TFTs, os métodos convencionais aplicados a MOSFETs podem resultar em valores pouco precisos. Isto deve-se ao facto de a corrente no dreno (i_D) de saturação ser modelada com um expoente diferente de 2 [23], como no caso do modelo usado, onde podemos ver por (3.8) que o exponente é α que segundo a aproximação realizada, toma o valor de 2.5091. Desta forma, o método a ser usado deve ter em consideração o expoente na obtenção do V_T

De acordo com [23], uma parte considerável dos métodos usa a região de inversão forte, enquanto que apenas alguns consideram o funcionamento do transístor em inversão fraca. Maioria das extrações são feitas usando V_g baixos de modo a que o transístor opere na região linear, mas também existem alguns exemplos que usam a operação do transístor em saturação. De modo a escolher um procedimento que resulte com o modelo, primeiro foi necessário conhecer alguns métodos e a forma como estes são executados com fim a avaliar a sua aplicabilidade neste caso.

Os métodos estudados foram desenvolvidos para MOSFETs de monocristais. Entre vários procedimentos, os que são realizados na região linear são: corrente constante, extrapolação na região linear, extrapolação da transcondutância na região linear e da segunda derivada. No que diz respeito à região de operação na saturação foi considerado: a raiz de I_D. Com exceção do último método, todos foram fundamentados por [23].

Método da Corrente Constante

Neste método, a extração do valor de V_T é dado como o valor da tensão na porta (v_G) quando é aplicada uma corrente i_D arbitrária e constante conforme na Figura 3.7. Geralmente o valor da corrente no dreno escolhida é W/L x 1 nA [23]. Para um valor de v_{DS} na região que pretendemos, vemos a que valor de v_G corresponde o valor de i_D escolhido.

Dada a sua simplicidade, este método é um dos mais usados, mas tem a grande desvantagem de ser dependente de um valor arbitrário (i_D) que acaba por se repercutir nos resultados. Como o V_T será para avaliar a credibilidade da aproximação, este método não seria o mais adequado.

Método da Extrapolação na Região Linear

Como o nome sugere, este método é realizado na região linear e usa a relação da corrente no dreno e a tensão na porta (i_D-V_g) para se fazer a extrapolação linear da curva no ponto máximo da derivada, isto é, com maior inclinação (Figura 3.8). Este ponto também pode ser designado o ponto de máxima transcondutância (g_m). O valor de V_T é extraído através da interceção do resultado da extrapolação com o eixo V_g.

O principal problema deste método é o facto de o ponto de máxima inclinação poder ser incerto. No caso do modelo deste projeto, este método não pode ser aplicado porque o comportamento do TFT modelizado foge da linha reta ideal. Sendo assim, não se encontra facilmente o ponto com valor máximo da derivada para a situação que se pretende.

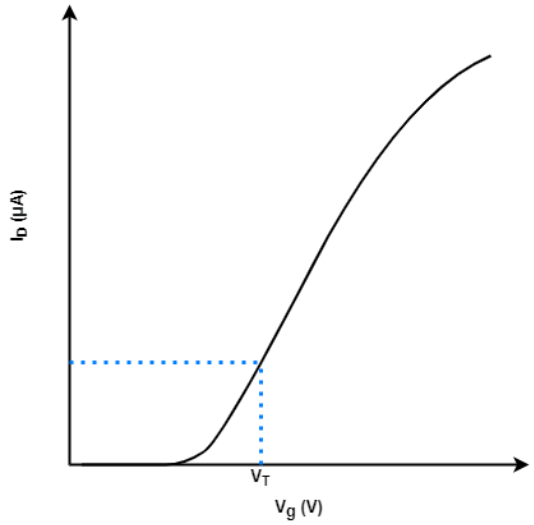


Figura 3.7: Método da corrente constante

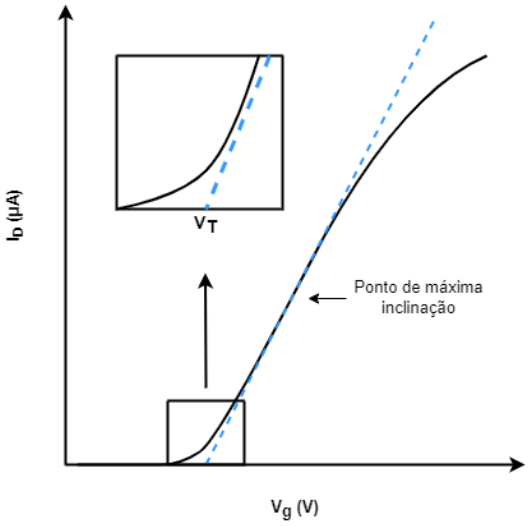


Figura 3.8: Método da extrapolação na região linear

Método da Extrapolação da Transcondutância na Região Linear

Este método usa a curva que resulta da característica g_m-V_g. O V_T corresponde à interceção da extrapolação linear da curva mencionada no ponto de derivada máxima com o eixo correspondente à tensão da porta.

O problema para recorrer a este método está relacionado novamente com o facto de o modelo não ter um comportamento como esperado, pois como referido anteriormente o modelo foge da linha reta sendo sempre crescente. Como tal, o ponto de derivada vai ser sempre incerto, pelo que não é muito fiável a aplicação deste método.

Método da Segunda Derivada

Para a obtenção do V_T, é necessário ver o ponto máximo da derivada da transcondutância, ou seja, a dupla derivada de I_D. Relembrando que $g_m = dI_D/dV_g$ e por isso $dg_m = d^2I_D/d^2V_g$. Deve ser usado um valor de V_d para que o transístor opere na região linear.

Este método não é possível aplicar porque como sabemos pela aproximação, o V_T tenderá a ser negativo e como apenas extraímos a curva para V_g de 0 a 5V, não iríamos ter um resultado realista face ao modelo usado.

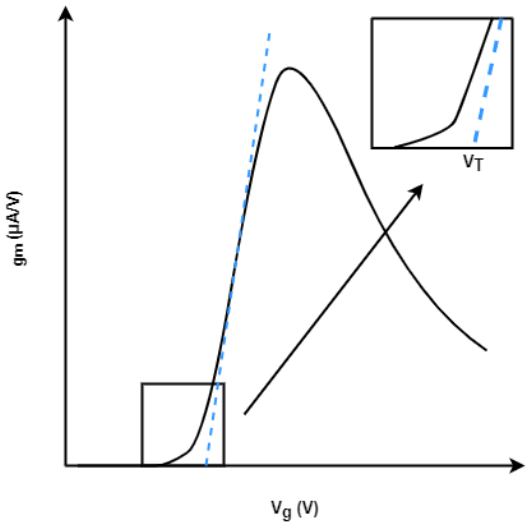


Figura 3.9: Método da extrapolação de transcondutância na região linear

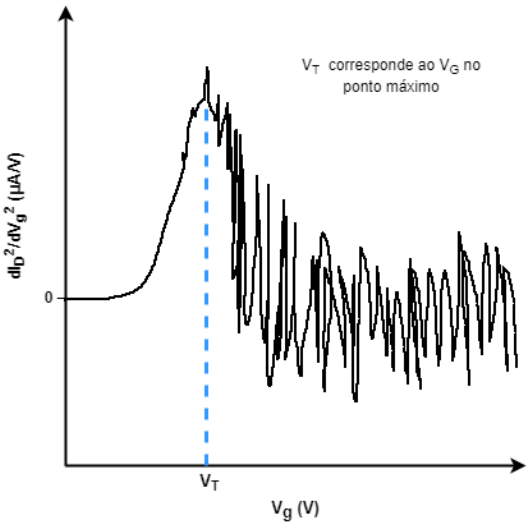


Figura 3.10: Método da segunda derivada

Método da Raiz de i_D

O valor de V_t é dado pela interceção da extrapolação linear da caraterística $\sqrt{i_D}$-v_g na parte linear do gráfico com o eixo v_g. Este método é baseado na aproximação da equação da corrente no dreno do MOSFET na saturação a

$$i_D = K(v_{GS} - V_T)^2 \tag{3.9}$$

Aplicando a raiz quadrada em ambos os termos

$$\sqrt{i_D} = \sqrt{K}(v_{GS} - V_T) \tag{3.10}$$

Considerando equação que descreve uma reta $y = mx + b$, onde $y = \sqrt{(i_D)}$, $m = \sqrt{K}$, $x = v_{GS}$ e $b = -\sqrt{K}V_T$, a forma da curva da caraterística $\sqrt{i_D}$-v_g está representado na Figura.

Este método tem em consideração o expoente da expressão da corrente na saturação. Neste caso, o método apresentado é relativo a MOSFETs, mas pode ser adaptado para os TFTs usando a respetiva equação da corrente na saturação. Sendo assim, este foi o método escolhido, por ter em consideração o expoente α, que já vimos ser uma consideração importante para ter um resultado o mais preciso possível.

3.3.2 Aplicação do Método de Extração de V_T

De forma a aplicar o método escolhido ao modelo definido, foram usadas as curvas caraterísticas i_D-v_{GS} em saturação, que neste caso corresponde a valores onde $v_{DS} > 4\,\text{V}$. Atendendo à equação (3.8), podemos ver que esta tem expoente α, portanto recorrendo à representação gráfica da relação $^{\alpha}\sqrt{i_{DSAT}}$-v_{GS} verificamos o formato linear na região de acumulação forte (Figura 3.12).

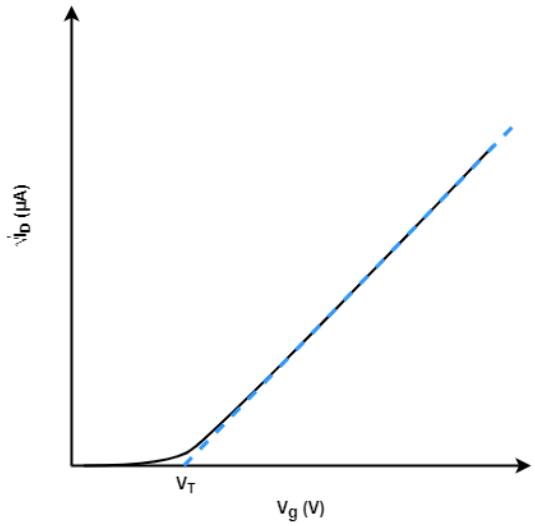


Figura 3.11: Método da raiz de I_D

De seguida, usando os valores representados graficamente, foi realizada uma aproximação da parte linear a uma reta. Sendo assim, o valor de V_T foi determinado através da interceção dessa reta com o eixo V_{gs}, onde se obteve o resultado de −0,236 V. O valor extraído dos dados originais resulta em −0,183 V. Confirma-se que os transístores são de depleção, representando uma diferença relativa a este último valor, de cerca de 29 %. De qualquer modo, o método de de extração a partir dos dados é feita assumindo uma relação quadrática o que em parte será responsável pela diferença.

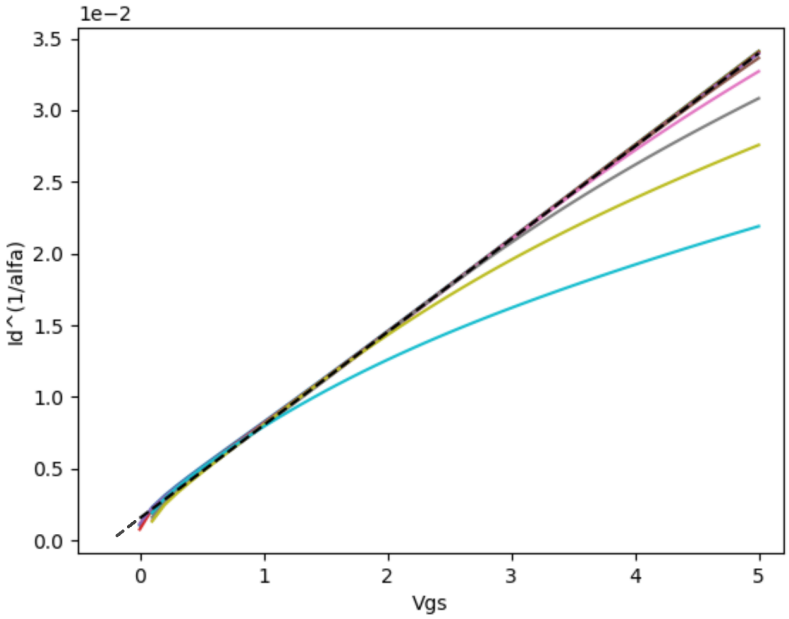


Figura 3.12: Aplicação do método da raiz de I_D

O valor V_T dos transístores do tipo n tipicamente é positivo, mas neste caso, como se pode perceber, o V_T tem um valor negativo. Isto significa que para v_{GS} nulo, haverá corrente a passar no transístor ($i_D \neq 0\,\text{A}$). Desta forma, o TFT torna-se vantajoso para fazer polarizações, pois com $v_{GS} = 0\,\text{V}$, o transístor funciona como uma fonte de corrente constante. Estes tipos de transístores designam-se por transístores de depleção, pois necessitam de uma tensão na porta para que o transístor deixe de conduzir. Normalmente este tipo de transístores são escolhidos para minimizar a dissipação de energia nos circuitos dimensionados.

3.4 Precisão da Aproximação

De forma a descobrir o resultado a aproximação do modelo alfa ao TFT, podemos fazer uma comparação entre as curvas extraídas do modelo do transístor inicial e as curvas resultantes da aproximação. Como referido anteriormente o valor V_T é muito importante para avaliar a confiabilidade na modelização do TFT, uma vez que descreve a sua operação.

À partida pode-se considerar que a aproximação está bastante satisfatória e com resultados perto do pretendido. Esta afirmação sustenta-se na comparação entre os valores de V_T extraídos pela aproximação e pelo método da raíz de I_D, relembrando que o V_T obtido pela aproximação realizada neste capítulo foi de $-0{,}229\,\text{V}$ e o resultado da extração deste parâmetro a partir do método da raíz de i_D foi de $-0{,}236\,\text{V}$.

Através da normalização do erro quadrático médio (NMSE), pode ser feita uma comparação direta do erro da aproximação aos valores usados. Esta parte foi dividida pelo cálculo do NMSE da região linear e de saturação. Novamente, com recurso ao *Python*, foram divididos os pontos para cada v_{GS} que pertenciam às diferentes regiões. Posteriormente, esses pontos foram aplicados à formula que define o NMSE.

$$NMSE = \frac{MSE}{\frac{1}{N}\sum_{i=1}^{N}(x_i)^2} \tag{3.11}$$

Tendo em consideração que MSE é o erro quadrático médio, este é definido por

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{3.12}$$

O parâmetro y representa os valores da aproximação a x, por isso x serão os valores observados. O N representa o tamanho do conjunto de valores de x.

Finalizando, o resultado do NMSE está representado na Tabela 3.2, onde podemos ver que o erro é muito reduzido, tornando a aproximação apta para ser transcrita em *Verilog-A* e descrever o TFT a ser utilizado no desenvolvimento dos circuitos dos elementos do cérebro.

Tabela 3.2: Resultados do NMSE

V_{GS}	Tríodo	Saturação
1	1.0×10^{-3}	1.4×10^{-3}
1.5	3.9×10^{-4}	1.2×10^{-4}
2	2.8×10^{-4}	1.0×10^{-5}
2.5	1.9×10^{-4}	4.1×10^{-6}
3	1.5×10^{-4}	2.6×10^{-6}
3.5	1.3×10^{-4}	1.3×10^{-6}
4	1.4×10^{-4}	1.0×10^{-6}
4.5	1.3×10^{-4}	1.1×10^{-6}
5	1.2×10^{-4}	1.5×10^{-6}

3.5 Modelo Final

Após a validação do resultado da modelização, este foi transcrito para *Verilog-A* juntamente com os parâmetros obtidos. Neste processo, foram implementadas capacidades parasitas no modelo, para ter uma descrição mais aproximada do que acontece no dispositivo real.

3.5.1 Descrição do Modelo

Na fase de descrição do modelo em *Verilog-A*, foi bastante direto pois sabendo as condições que separam as regiões e as respetivas equações (3.6) e (3.5), apenas se teve de transcrever esse comportamento.

Foi necessário ter uma atenção e verificar se a tensão da fonte do transístor é superior à tensão no dreno, pois caso isso não aconteça, é necessário trocar os dois terminais para se verificar essa condição. Assim, é assegurado que o transístor é sempre do tipo n.

3.5.2 Descrição das Capacidades Parasitas

No processo de simulação do circuito, o simulador constrói um modelo matemático do mesmo, maioritariamente através de uma análise nodal modificada. O modelo matemático mencionado é constituído por um sistema de equações que posteriormente é resolvido de forma a antever o comportamento do circuito. O simulador resolve a lei da corrente de *Kirchhoff* em cada nó para descobrir a tensão no mesmo, mas existem alguns componentes que não se enquadram diretamente neste formato, que é o caso de fontes de tensão e bobinas. Para estes componentes é necessário descobrir o valor da corrente através do elemento e incluir uma equação no sistema que tenha a relação da tensão para o elemento. Pela necessidade destas adições, resulta a analise nodal modificada utilizada pelos simuladores.

Sendo assim, de forma a seguir a formulação de nós utilizada pelo simulador e com o propósito de evitar o acrescento de variáveis, os modelos devem ter como base elementos de corrente controlados por tensão ($I(V)$) e a derivada temporal de elementos de carga controlados por tensão ($\partial q(V)/\partial t$). Há que se ter em atenção que o modelo esteja devidamente representado, permitindo assim que o simulador consiga resolver as equações de forma a chegar a uma solução única e garantir que a função e as suas derivadas sejam numericamente bem comportadas para todos os valores dos nós [24].

Tendo isto em consideração, existem algumas formas de implementar capacidades, de acordo com [24].

```
q = f(V(b_cap));
I(b_cap) <+ ddt(q);

C = f(V(b_cap));
I(b_cap) <+ C*ddt(V(b_cap));

C = f(V(b_cap));
I(b_cap) <+ ddt(C*V(b_cap));
```

Numa primeira vista, qualquer uma destas opções pode ser usada mas há algumas considerações a ter. Em relação à segunda forma, esta tem uma variável extra no ramo ddt da equação e devido a isso, não deve ser usada porque não se enquadra na formulação nodal feita pelo simulador, criando um sistema extra. Em relação à terceira forma, não se deve recorrer a este modo se a capacidade for não linear (pode resultar em situações onde o princípio da conservação da carga não se verifica). Como neste caso as capacidades têm um valor constante, a terceira proposta acaba por não ser um problema.

Como a carga pode ser vista como $q = C \cdot V$, a lógica para implementação do código foi baseada diretamente na carga, representada pela primeira opção proposta no artigo. Como resultado, surgiu a seguinte lógica implementação

```
if (V(D,gnd) > V(S,gnd)) begin
        I(D,S) <+ ids;
        qgs = cgs * V(G,S);
        qgd = cgd * V(G,D); end

I(G,S) <+ ddt(qgs);
I(G,D) <+ ddt(qgd);
```

Nesta proposta, os argumentos da carga estão nos blocos condicionais, tal como sugerido na referência [24]. Isto evita que sejam criadas mais equações, simplificando o processo. Juntamente a isto, de acordo com a região em que o transístor opera, foram definidos os respetivos valores para as capacidades C_{gs} e C_{gd}. Estes valores foram definidos de acordo com o Modelo de *Meyers* com base em valores já caracterizados destas capacidades e *overlaps*.

As capacidades introduzidas equivalem a adicionar condensadores no esquemático do transístor como representado da Figura 3.13.

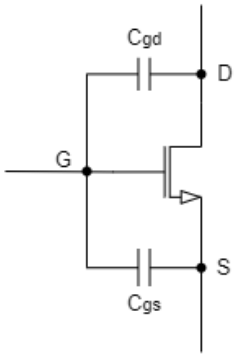


Figura 3.13: Capacidades parasitas

3.5.3 Resultado

De forma a confirmar que o transístor estava a ter o comportamento esperado, foi novamente retirada a característica I-V do TFT tendo como resultado a Figura 3.14.

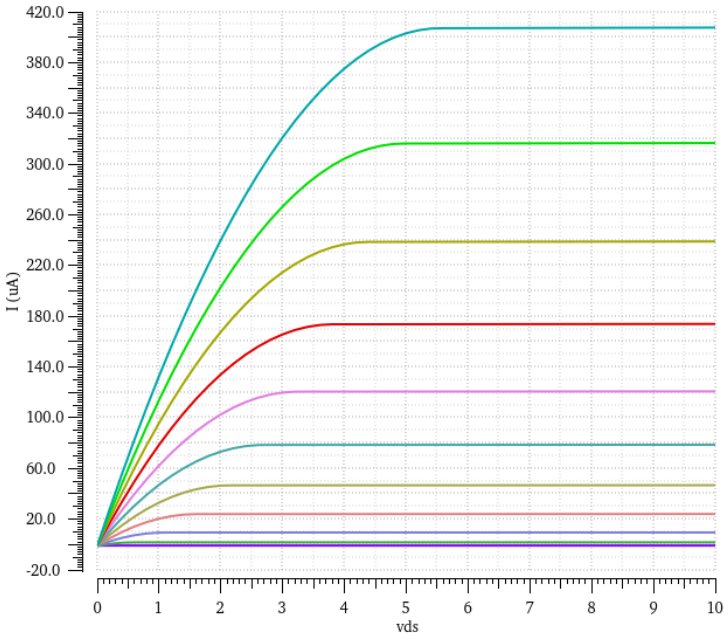


Figura 3.14: Curva característica I-V do novo modelo do TFT

Pode ser confirmado que este modelo passa por zero e o formato das curvas corresponde ao esperado do modelo, portanto à partida não haverão problemas relacionados com a operação do

transístor, especialmente perto de zero por ser a preocupação inicial. Sendo assim, este modelo foi o utilizado para a fase de desenvolvimento dos circuitos.

3.6 Conclusões

Este capítulo começa com um estudo à tecnologia a ser usada nesta dissertação: TFTs. Assim sendo, foi revista a evolução destes transístores até ao surgimento de TFTs baseados em a-GIZO, bem como a possíveis estruturas destes transístores, modo de funcionamento, materiais e algumas limitações.

Conhecida a tecnologia, foi importante avaliar o modelo de TFT inicialmente fornecido. Como se observou um problema relevante para a implementação de circuitos que exigem a comutação do transístor, foi necessário fazer uma nova modelização do transístor. Sendo assim, foi escolhido o modelo *alpha-power law* e foi feita uma aproximação deste modelo ao TFT. Posteriormente, foi avaliada a confiabilidade da aproximação através da extração do V_T do modelo inicial e a precisão da aproximação feita através do calculo do $NMSE$.

Por fim, este modelo foi transcrito para *Verilog-A* e juntamente, foram acrescentadas capacidades parasitas. De forma a confirmar que o modelo ficou como previsto do ponto de vista estático, foi retirada uma ultima simulação para confirmar se o transístor tem o comportamento esperado.

Capítulo 4

Arquiteturas CMOS de estruturas neuronais e arquitetura proposta em TFT

4.1 Implementações

Considerando os 3 modelos neuronais abordados no Capítulo 2, esta parte de desenvolvimento passa por ver implementações já realizadas neste âmbito de forma a escolher um modelo que melhor se adeque aos requisitos procurados para os circuitos a serem trabalhados. Sendo assim, os pontos principais a ter em conta serão a eficiência energética e a complexidade das propostas.

Grande parte das implementações revistas são feitas com CMOS, mas também serão discutidas implementações com TFTs encontradas e a sua relevância para a conceção deste projeto.

4.1.1 Implementações baseadas em *Integrate-and-Fire*

Os circuitos fundamentados no modelo IF tipicamente integram pequenas correntes através de um condensador até ao valor de limiar (V_{Th}). Quando este ponto é atingido, é desencadeado um impulso e o condensador é descarregado.

Existem algumas implementações com fabricação de chip de variações do modelo IF. O modelo adaptado do LIF na implementação proposta por [6], faz considerações importantes, que diminuem a potência consumida pelo circuito. O circuito do neurónio (Figura 4.1) entra num período refratário (controlado por M8 a M11) depois de ocorrer um impulso. Isto significa que não consegue despoletar outro potencial de ação pois toda a corrente injetada no circuito é absorvida. Este circuito também tem um mecanismo de adaptação da frequência do impulso (executado por M15 a M19), que é controlada por um transístor (M18) pois aumenta a constante de tempo efetiva do integrador. Isto permite escolher transístores de menor tamanho para o circuito que consequentemente têm capacidades parasitas menores. Para além da capacidade de estabelecer um período refratário e o mecanismo de adaptação da frequência do impulso este modelo usa 20 transístores,

pelo que ocupa uma pequena área de circuito. Todos estes fatores referidos em conjunto participam na redução do consumo de potência geral do circuito.

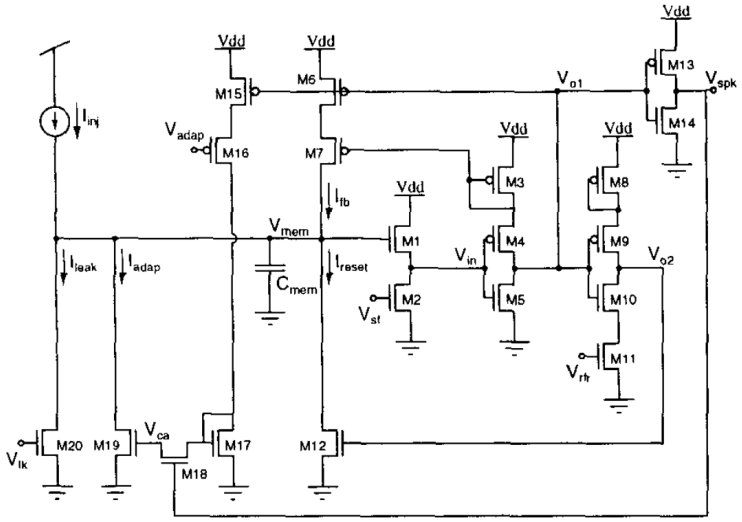


Figura 4.1: Diagrama do circuito, reproduzido de [6]

Por outro lado, a abordagem da referência [25] para reduzir a potência consumida apoia-se em decisões como o material que constitui os transístores (óxido fino e espesso), uso de condensadores de porta MOS e opções de alimentação dupla. Há a introdução de uma memoria capacitiva (Capmem) analógica *on-chip* que é interessante por oferecer STDP e assim otimizar as operações de sinapse. Esta técnica também tem outro ponto positivo, pois favorece o circuito em termos de consumo de potência pelas baixas correntes de polarização (na ordem dos 15 nA) permitindo que o circuito trabalhe em inversão fraca e moderada.

Os autores de [7] apoiam-se na escolha do tipo de transístores usados para reduzir o consumo de potência. Depois de feita uma simulação ao circuito o resultado apresentou alguns problemas, marcando partes limitantes do circuito (Figura 4.2): o período refratário (N19 e N20), o input (P3, P4, N3 e N4) e a primeira parte do comparador de corrente (P14, P15, N5, N6), estas podem ser otimizadas para variações aumentando o comprimento e largura dos transístores nessas partes dos circuito.

Todas as implementações têm um base semelhante que consistem num bloco de *input*, bloco comparador e um bloco que representa o período refratário. De acordo com as especificações que se espera de cada implementação, são adaptadas algumas partes do circuito para satisfazer os requisitos.

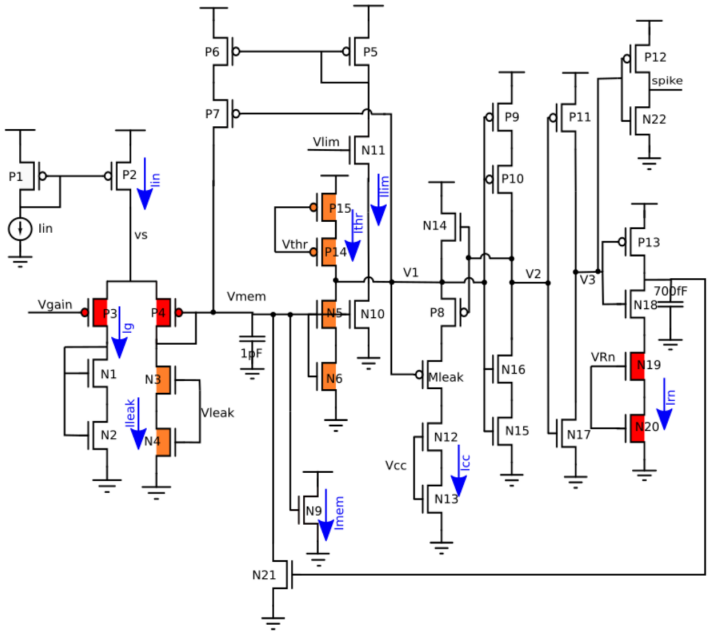


Figura 4.2: Esquema do circuito, reproduzido de [7]

4.1.2 Implementações baseadas em *Hodgkin-Huxley*

Analisando a referência [26] vemos que dada a complexidade do modelo que por sua vez tem uma árdua implementação, ocupa uma grande área de circuito pelo numero elevado de transístores necessários.

Existem algumas implementações que simplificam o HH, como o caso da proposta [27], mas este acaba por não descrever tão realisticamente nem ter tão bons resultados a nível de consumo, quando comparado com outros modelos mais simples. Portanto, reduzir um modelo complexo acaba por não ser um caminho a seguir, uma vez que conseguimos melhores resultados com um modelo simples bem estudado.

Apesar deste modelo ser bastante atrativo pela sua dinâmica realista sobre o potencial de ação, após um análise da literatura, as implementações encontradas deste modelo não são muito relevantes, como já justificado. Assim, o modelo HH deixa de ser um candidato interessante para o desenvolvimento do projeto e possível integração em SNNs.

4.1.3 Implementações baseadas em *Izhikevich*

Sabendo que o modelo *Izhikevich* simula o comportamento de vários padrões de impulsos, é esperado que padrões distintos tenham diferentes consumos. Na referência [13] isto é verificado depois de experimentar a capacidade do circuito para diferentes comportamentos, concluindo que o consumo de energia é aproximadamente proporcional à frequência média de picos, que para este caso é muito baixa.

Relembrando que a área de circuito é uma consideração importante para VLSI, a proposta de implementação é inteiramente baseada na proposta do modelo *Izhikevich*, usando apenas 14 transístores para o concretizar. Através de algumas avaliações, concluiu-se que esta proposta tem a capacidade de simular todos os comportamentos dos neurónios. A implementação não foi integrada numa rede neuronal com mecanismos de sinapse e comunicação necessários, por isso não sabemos o impacto energético real para incluir estas ações. Ainda assim, a abordagem apresentada em [13] é um circuito interessante para integração VLSI pela sua simplicidade de implementação, baixo consumo de potência e baixa área de circuito.

Os autores de [28] apresentam uma abordagem semelhante, mas em vez de ser controlado em modo tensão como a implementação anterior, é feita em modo corrente. Isto diz-nos logo que à partida esta abordagem terá uma eficiência energética ainda melhor. Apesar de uma pequena adaptação, esta proposta continua bastante semelhante ao modelo *Izhikevich* original e com uma simplicidade de implementação semelhante ao proposto em [13]. O circuito de [28] também é adequado para integração VLSI pelos mesmo motivos apresentados para [13], mas nas simulações foi detetado algum ruído nos impulsos.

A referência [29] apresenta um modelo adaptado que consegue simular quase todos os padrões de impulsos (17 de 20 totais). O circuito tem um integrador em modo corrente de u (representa a quantidade de corrente dos canais que participam para o potencial de membrana), em vez da abordagem no circuito todo, como discutido na referência [28]. O resultado das simulações conclui uma boa eficiência energética e por isso um possível candidato.

4.1.4 Implementações com *Thin Film Transistors*

Como já referido, há poucas aplicações de TFTs em eletrónica neuromórfica, mas já se encontram simulações com boa performance [20] e [30].

Na referência [31] é apresentado o desenvolvimento uma rede neuronal pouco comum, com posterior fabricação de chip. Esta é composta por elementos de processamento do cérebro simplificados (neurónio e sinapse) formados com recurso a TFTs de poly-Si. Nesta criação, o funcionamento do neurónio foi reduzido aos requisitos mínimos e propuseram um circuito do mesmo composto por 2 inversores e 2 *switches* ligados circularmente. Relativamente à sinapse, esta foi simplificada a um único transístor onde a fonte e o dreno serão ligados a neurónios pré e pós sinápticos e a porta usada para controlar a tensão de forma a transmitir o sinal de um neurónio para o seguinte. Apesar de não ter uma estrutura convencional e aproximada ao funcionamento real do cérebro, isto mostra resultados positivos a futuras implementações de redes neuronais completas com TFTs.

4.2 Circuito do Neurónio

Como já referido, a eficiência energética constitui uma das importantes considerações a ter para escolha do modelo neuronal. Os modelos candidatos até agora foram escolhidos com base no seu baixo nível de complexidade, uma vez que isto tem impacto direto no consumo de potência

e área de circuito. O modelo LIF particularmente é o modelo mais simples e apesar de não ser o mais realista, quando integrado numa rede neuronal tem uma aproximação igualmente boa da dinâmica do cérebro humano e ao mesmo tempo é simples e tem uma boa eficiência energética.

Desta forma, o desenvolvimento do circuito do neurónio tem como base de inspiração o circuito proposto por *Carver Mead* [8] (Figura 4.3), que tem um comportamento com essência do modelo IF.

4.2.1 Funcionamento do Circuito Proposto por *Carver Mead*

De forma a descrever o comportamento do circuito num contexto mais qualitativo, este começa com a entrada do circuito em corrente, que vai carregando os condensadores (que serve para simular o comportamento da membrana ao receber o impulso). Quando o ponto de comutação é atingido, isto é, o valor de limiar (V_{Th}), ocorre um pico e o circuito de *feedback* evita que os condensadores acumulem mais carga, fazendo um *reset* provocando um recomeço no ciclo que gera o impulso. Considerando a parte de *reset* desempenhada pelos transístores Q_1 e Q_2, esta acontece quando o V_{out} atinge uma tensão que faz com que Q_2 conduza (atuando como um *switch*) e deixe passar a corrente de descarga (I_{Pl}) controlada por Q_1.

Relembrando um pouco o funcionamento do neurónio, quando o valor de limiar é atingido há um fluxo contínuo de Na^+ para dentro do neurónio, tornando o seu potencial positivo (despolarização). Quando é atingida um determinado valor de potencial durante a despolarização (equivalente ao pico do impulso), este volta a ser reposto ao valor inicial pela ativação temporária dos canais K^+ (repolarização). De acordo com [8], num contexto biológico os transístores Q_1 e Q_2 representam os canais K^+ presentes na membrana que são ativados de forma restabelecer o potencial de repouso.

4.2.2 Desenvolvimento do Circuito

Numa tentativa inicial de replicar o circuito mencionado, uma das primeiras preocupações é encontrar formas de desenhar o circuito apenas com transístores do tipo n. Como já referido, ainda não há a possibilidade de reproduzir TFTs baseados em a-GIZO do tipo p. Neste caso, os transístores do tipo p em complementaridade com os do tipo n formam dois pares de inversores ligados, que desempenham a função de um amplificador/comparador. Já conhecidas as caraterísticas dos TFTs a serem usados, na parte de dimensionamento do circuito foi importante começar pela análise do par de inversores mencionado, uma vez que este produz o V_{out} que por sua vez controla a ativação do transístor Q_2.

Inversor 1

Uma prática comum para construir inversores apenas com transístores do tipo n passa por substituir o transístor do tipo p por um transístor do tipo n com ligação em díodo como representado na Figura 4.4a.

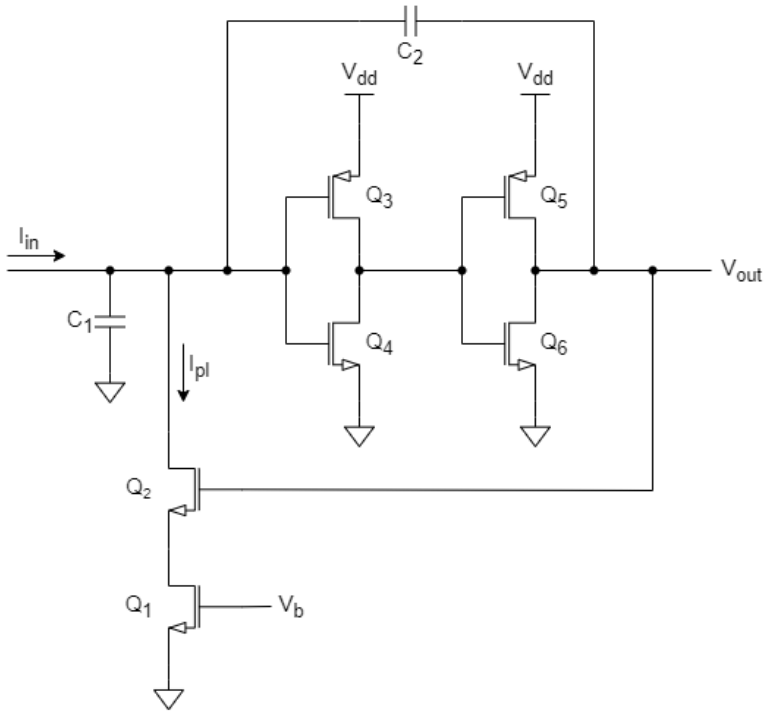
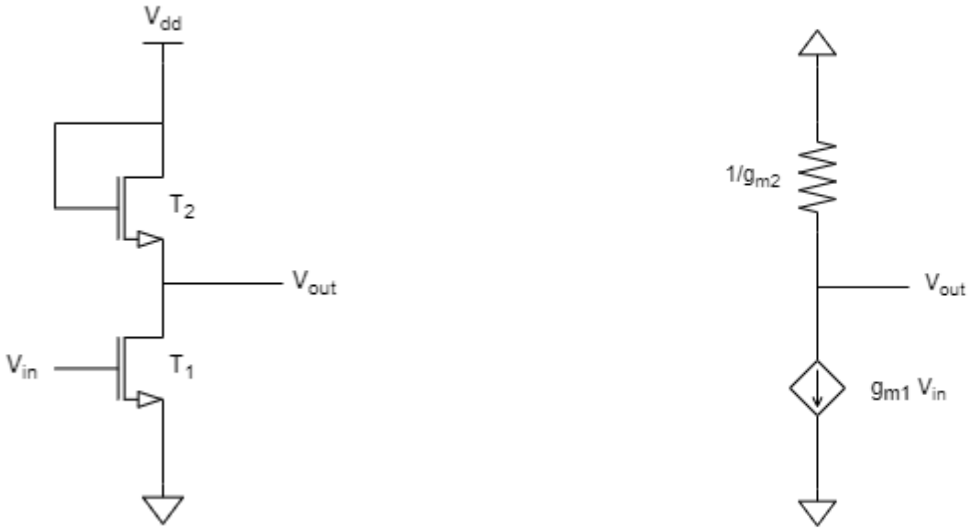


Figura 4.3: Esquema do circuito proposto por *Carver Mead*.



(a) Circuito

(b) Equivalente para sinal

Figura 4.4: Inversor 1

No ponto de vista do equivalente do circuito para sinal (Figura 4.4b), nesta configuração o transístor superior T_2 pode ser representado como uma resistência com valor aproximadamente de $1/g_{m2}$. Isto porque, para esta analise não será considerado o r_o, caso contrário seria $(1/g_{m2})//r_{o2}$.

Relativamente ao transístor inferior T_1 pode ser exprimido como uma fonte de corrente controlada por v_{in} por onde i_d terá como valor $g_{m1} \cdot v_{gs}$. Como o v_{gs} é igual a v_{in}, podemos definir a relação de v_{in} e v_{out} da seguinte forma

$$v_{out} = 1/g_{m2} \cdot (-g_{m1} v_{in}) \tag{4.1}$$

Portanto, o ganho desta configuração é dado aproximadamente por

$$A_{v1} \approx -\frac{g_{m1}}{g_{m2}} \tag{4.2}$$

Baseado nesta relação e na equação (3.3), pode-se perceber que o tamanho de T_1 terá que ser superior a T_2. Como a ideia é tentar replicar o circuito analisado, então primeiramente foi feito o dimensionamento para um par de inversores. De forma a achar a relação entre os *TFTs* que proporcione o melhor ganho possível resultante, foi realizada uma análise paramétrica com recurso ao simulador do *Cadence*, para determinar W e L dos transístores que dão os melhores resultados possíveis (Tabela 4.1). Deve ser considerado, que por limitações de fabricação o W tem de ser entre 20 µm a 320 µm e o tamanho do L deverá ser entre 10 µm a 20 µm. Atente-se que o valor de V_{DD} usado em todos os circuitos será sempre 10 V.

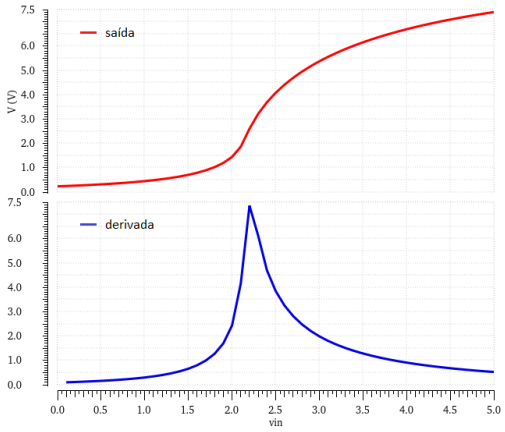


Figura 4.5: Tensão de saída para um par de inversores

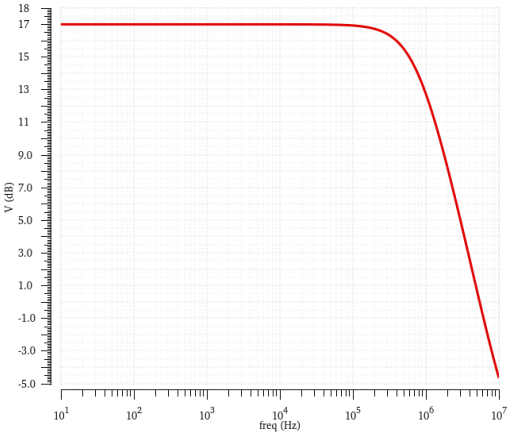


Figura 4.6: Resposta em frequência do inversor 1

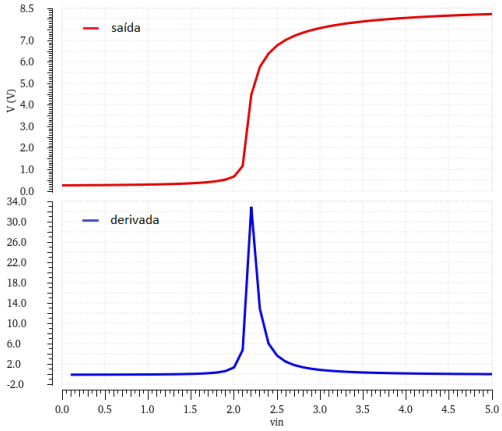
Tabela 4.1: Dimensionamento do inversor 1

W_1	220 µm	W_2	20 µm
L_1	10 µm	L_2	20 µm

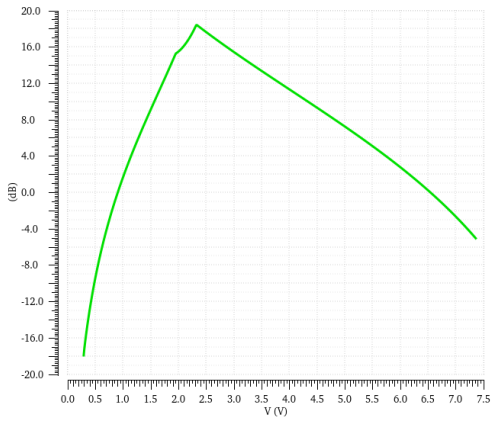
Do ponto de vista *DC*, o ganho dado por um par de inversores pode ser analisado através do calculo da derivada da tensão de saída dos inversores, representado na Figura 4.5, ou através da análise AC na Figura 4.6. Através da análise do gráfico da Figura 4.5, é possível ver que a derivada mostra o ganho máximo e o ponto de comutação (valor de V_{in} correspondente ao ganho máximo).

Estas informações são importantes para o desempenho do circuito do neurónio, pois informa o valor médio de operação no circuito. A análise *DC* foi feita usando dois inversores, pois como seriam usados sempre pares de inversores, pode-se ver os valores finais.

Como o ganho não é suficiente (aproximadamente 7,5 V) , de forma a permitir o funcionamento do circuito foi necessário acrescentar mais um par destes inversores para ter um ganho aceitável (pelo menos de 10). Não esquecendo que o ganho de um inversor é negativo, então tem de se utilizar sempre um numero par de inversores de forma a ter um ganho positivo. Posto isto, a característica obtida por dois pares de inversores em série, pode ser consultado na Figura 4.7a, onde se pode perceber que apesar de o ganho ser bom, a transição não é muito rápida nem linear. Uma forma de avaliar melhor como varia o ganho do amplificador formado pelos dois inversores foi extraído o gráfico da Figura 4.7b. Isto permite perceber que o ganho não é muito linear, pois verifica-se uma inversão do mesmo perto de 2,5 V.



(a) Tensão de saída e respetiva derivada para dois pares de inversores



(b) Derivada vs Vout

Figura 4.7: Análise DC para dois pares do inversor 1

Geralmente, o inversor com a configuração em díodo é feita para *MOSFETs* de enriquecimento e por isso, a característica desta configuração faz com que o transístor superior esteja sempre em saturação ($v_{DS} > v_{GS} - V_t$). Neste caso isso não acontece, pois estamos perante transístores de depleção e consequentemente o transístor T_2 do Inversor 1 encontra-se em tríodo. Isto pode ser verificado, pois como a porta e o dreno estão ligados pode-se reconhecer que $v_{GS} = v_{DS}$ e dado que o valor de V_T para estes *TFTs* é negativo então verifica-se que

$$v_{DS} < v_{GS} - V_t \tag{4.3}$$

Como o transístor T_2 se encontra em tríodo a equação que descreve a corrente nesse transístor é dada por

$$I_D = K\left[(v_{GS} - V_T)v_{DS} - \frac{v_{DS}^2}{2}\right] \tag{4.4}$$

Se v_{DS} for substituído por v_{GS}

$$I_D = K\left[\frac{v_{GS}^2}{2} - v_{GS}V_T\right] \tag{4.5}$$

Não esquecendo que

$$g_m = \frac{\partial I_D}{\partial v_{GS}} = K(v_{GS} - V_T) \tag{4.6}$$

De forma a tentar manter a saturação do transístor referido e ter uma melhor resposta do inversor 1, foi abordada uma segunda possibilidade.

Inversor 2

O inversor aqui proposto está representado na Figura 4.8. Este circuito é uma versão melhorada do Inversor 1, onde são acrescentados os transístores T_3 e T_4 de formar a assegurar a saturação do transístor T_2. Se este se encontrar a operar em saturação, à partida esta configuração inversora terá um ganho maior. Isto porque como nesta região o g_m é maior em relação à região linear e dado que o ganho é dependente dos g_m, então é esperado um comportamento melhor face ao que se pretende.

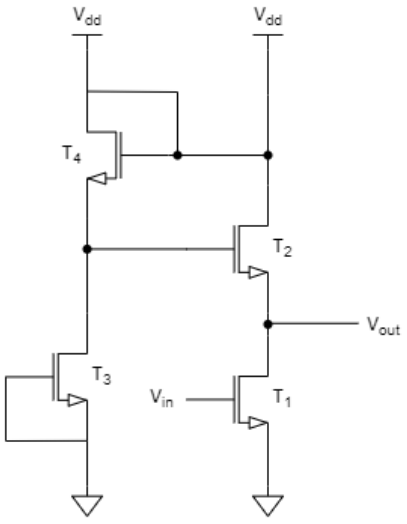


Figura 4.8: Circuito

Tendo em conta que v_{DS} pode ser obtido como

$$v_{DS2} = v_{DG2} + v_{GS2} \tag{4.7}$$

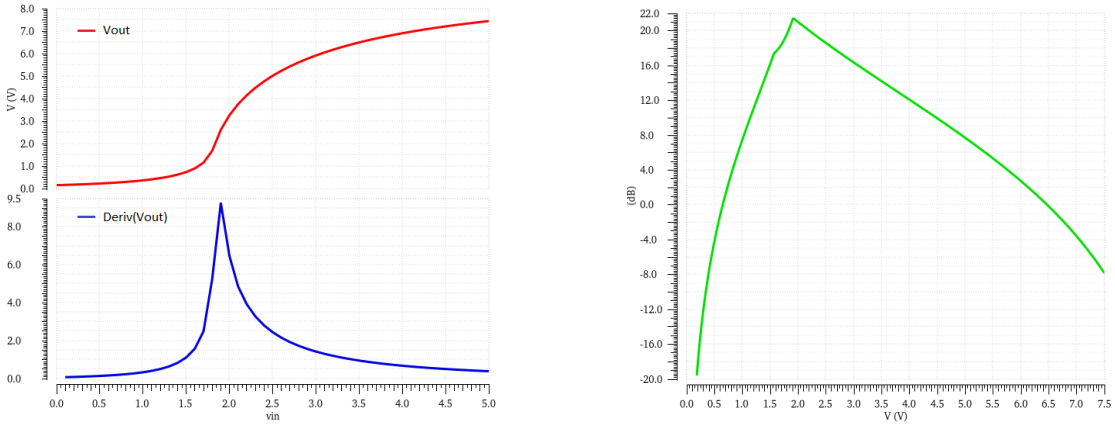
Para se manter T_2 em saturação tem de se verificar que $v_{DS2} > v_{GS2} - V_T$, portanto, substituindo v_{DS2} por (4.7) obtém-se $v_{DG2} + v_{GS2} > v_{GS2} - V_T$. Sendo assim tem de se verificar que

$$v_{DG2} > -V_T \tag{4.8}$$

Do ponto de vista do circuito para sinal, o transístor T_3 comporta-se como uma fonte de corrente, onde passa uma corrente I. Esta corrente também percorre T_4 e impõe o v_{GS4}. Como $v_{DG2} = v_{GS4}$, pode ser garantida a saturação de T_2 se a relação entre tamanhos de T_3 e T_4 seja suficiente para que $v_{DG4} > -V_T$. Para isto, a relação W/L de T_3 terá de ser bastante superior a T_4. Como o Inversor 1 já foi dimensionado e esta proposta é baseada nesse inversor, mas com o acrescento de dois transístores, os tamanhos de T_1 e T_2 foram mantidos. Após uma análise paramétrica, chegou-se à conclusão dos valores representados na Tabela 4.2 para esta proposta.

Tabela 4.2: Dimensionamento do inversor 2

W_1	220 µm	W_2	20 µm	W_3	280 µm	W_4	40 µm
L_1	10 µm	L_2	20 µm	L_3	10 µm	L_4	20 µm



(a) Vout e respetiva derivada

(b) Derivada vs Vout

Figura 4.9: Análise DC do Inversor 2

Após uma avaliação do sinal de saída de um par de inversores desta topologia (Figura 4.9) e da resposta em frequência do inversor (Figura 4.10), observa-se que efetivamente tem um ganho maior (próximo de 10) comparado ao primeiro inversor, contudo tem o ponto de comutação um pouco mais baixo. Como se pode perceber, o ganho obtido não é muito diferente do conseguido no inversor 1, isto porque se forem comparados os g_m do transístor T_2 em ambas as configurações, a expressão do g_m em tríodo (4.6) é igual à expressão na saturação (4.10).

$$I_D = \frac{1}{2}K(v_{GS} - V_T)^2 \tag{4.9}$$

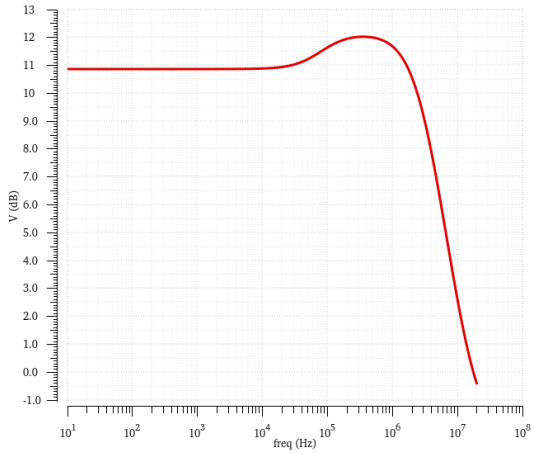


Figura 4.10: Resposta em frequência do inversor 2

$$g_m = K(v_{GS} - V_T) \tag{4.10}$$

A ideia seria apenas usar um par destes inversores em vez de dois pares do inversor 1, mas o ganho acaba por não ser tão bom comparado à segunda e o ponto de comutação é mais baixo. Consequentemente, a configuração do inversor 2 acaba por não ser uma proposta tão interessante como parecia inicialmente, pois apenas um par destes inversores não tem uma resposta muito melhor que dois pares do inversor 1, o que acabaria por consumir mais potência pelo número de transístores necessários para ter uma resposta superior.

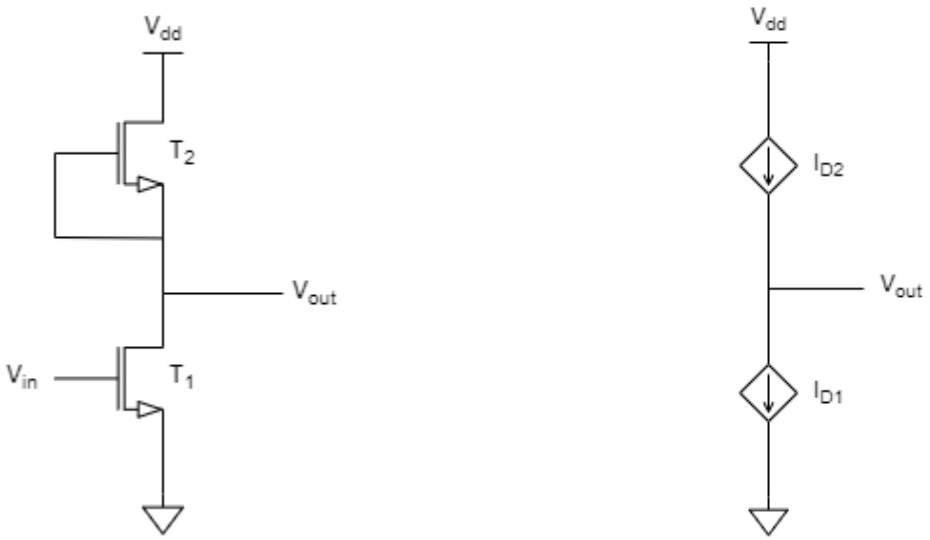
No sentido de chegar a um inversor com uma resposta melhor que as configurações vistas até ao momento, foi pensada numa terceira possibilidade, que aparentemente é simples mas tem um comportamento muito diferente.

Inversor 3

A terceira proposta de inversor é bastante parecida com o Inversor 1, mas em vez de ter a porta e o dreno ligados, tem a porta e a fonte conectadas como sugerido na Figura 4.11a.

Esta pequena alteração faz toda a diferença porque como já mencionado, estes *TFTs* tratam-se de transístores de depleção e, por isso, quando temos um v_{GS} nulo, que é o caso, o transístor comporta-se como uma fonte de corrente constante. Assim sendo, o circuito pode ser visto como duas fontes de corrente em série, representado na Figura 4.11b. Como passa a mesma corrente em ambas as fontes e como I_D pode ser dado aproximadamente por (3.8) visto anteriormente, as correntes nas duas fontes podem ser relacionadas como

$$I_{D1} = I_{D2} \tag{4.11}$$



(a) Circuito (b) Equivalente do circuito

Figura 4.11: Inversor 3

$$B\frac{W_1}{L_1}(v_{GS1}-V_T)^{\alpha} = B\frac{W_2}{L_2}(v_{GS2}-V_T)^{\alpha} \tag{4.12}$$

Como $v_{GS2} = 0\,\mathrm{V}$

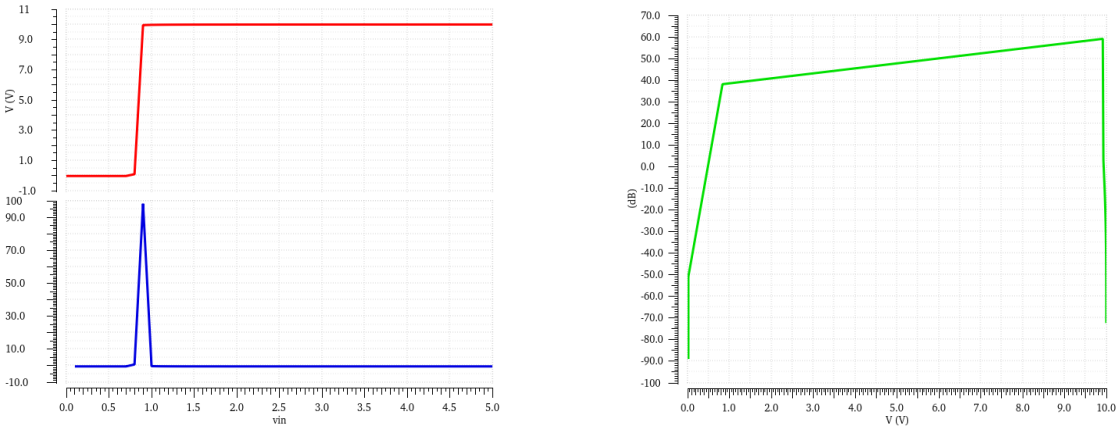
$$\frac{W_1}{L_1}(v_{GS1}-V_T)^{\alpha} = \frac{W_2}{L_2}(-V_T)^{\alpha} \tag{4.13}$$

Como V_T é um valor conhecido e negativo, a relação $(v_{GS1}-V_T)^{\alpha}$ vai ser sempre maior que $(-V_T)^{\alpha}$. Consequentemente, para a igualdade se manter, a relação W_2/L_2 terá de ser maior do que W_1/L_1. Novamente com recurso ao *Cadence* foi feita uma análise paramétrica dos valores W e L de ambos os transístores, para obter a relação que dá o melhor ganho de um par deste tipo de inversores (Tabela 4.3).

Tabela 4.3: Dimensionamento do inversor 3

W_1	20 µm	W_2	320 µm
L_1	20 µm	L_2	10 µm

Para este inversor foi acrescentado um transístor em paralelo a T_2, com o intuito de subir um pouco o ponto de comutação, que mesmo assim é bastante baixo. Observando a Figura 4.12a e 4.13, constata-se que esta configuração tem um ganho muito superior apenas com um par de



(a) Vout e respetiva derivada

(b) Derivada vs Vout

Figura 4.12: Analise DC do Inversor 3

inversores, contudo o ponto de comutação é demasiado baixo (em torno de 0,9 V). Como já mencionado, este ponto dita o valor médio de operação do circuito do neurónio e por isso queremos que esteja o mais a meio possível de V_{DD}, que tem o valor de 10 V. Também pode ser observado que a variação do ganho (Figura 4.12b) é muito mais linear em relação aos inversores anteriores.

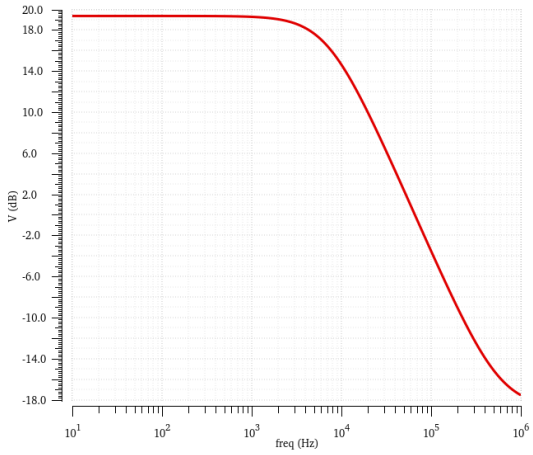


Figura 4.13: Resposta em frequência do inversor 3

Discussão dos Inversores

Perante a análise de cada inversor, uma forma de ter um circuito amplificador com o funcionamento pretendido é juntar diferentes inversores. Se for usado o inversor 2 juntamente com um par dos inversores 3, mas o ponto de comutação resultante da combinação referida é mais baixo. Para aumentar este ponto, seria necessário usar pelo menos mais um par destes inversores. Isto acabaria por consumir mais potência que a proposta anterior, pois são necessários mais transístores para obter o mesmo desempenho. Como o último inversor possui uma caraterística boa, mas

tem o problema do ponto de comutação ser próximo de zero. Uma solução é introduzir pares dos primeiros inversores para deslocar esse ponto para a direita.

De forma a ter uma comparação mais justa entre os inversores, foram medidos mais alguns valores. Primeiramente, foi feita uma avaliação relativa à rapidez de comutação dos inversores através da velocidade de propagação. Para obter o gráfico que permite fazer essa medição, foram impostas as condições em que iam operar. Sendo assim, temos as duas situações mais viáveis de operação: o conjunto formado por um par de inversores 1 mais um par de inversores 3 (opção A da Figura 4.14); o conjunto formado por dois pares de inversores 2 mais um par de inversores 3 (opção B da Figura 4.14).

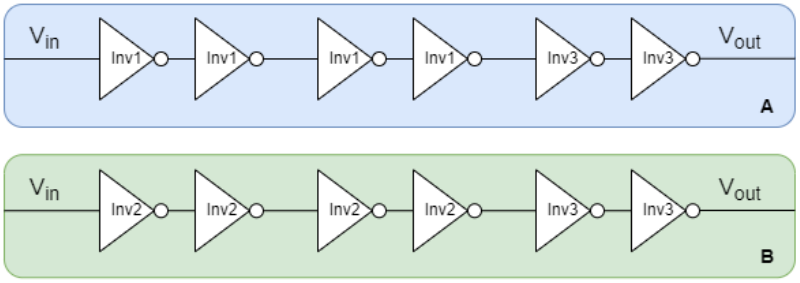
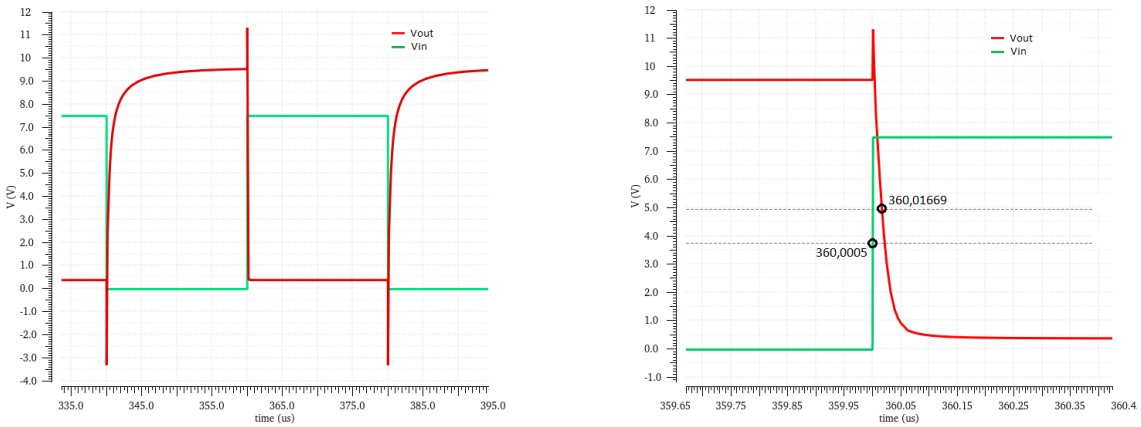


Figura 4.14: Esquemático com proposta da sequência de inversores

Estas duas sugestões resultam do que foi discutido inicialmente e como tal, de forma a definir as condições de operação para cada inversor, foram medidos os valores de tensão máxima e mínima que iriam receber na sua entrada. Para além disto, como nesta medição se mede a diferença entre a entrada e saída de um inversor, foi utilizado um inversor do mesmo tipo como carga, pois serão usados em pares como referido antes,

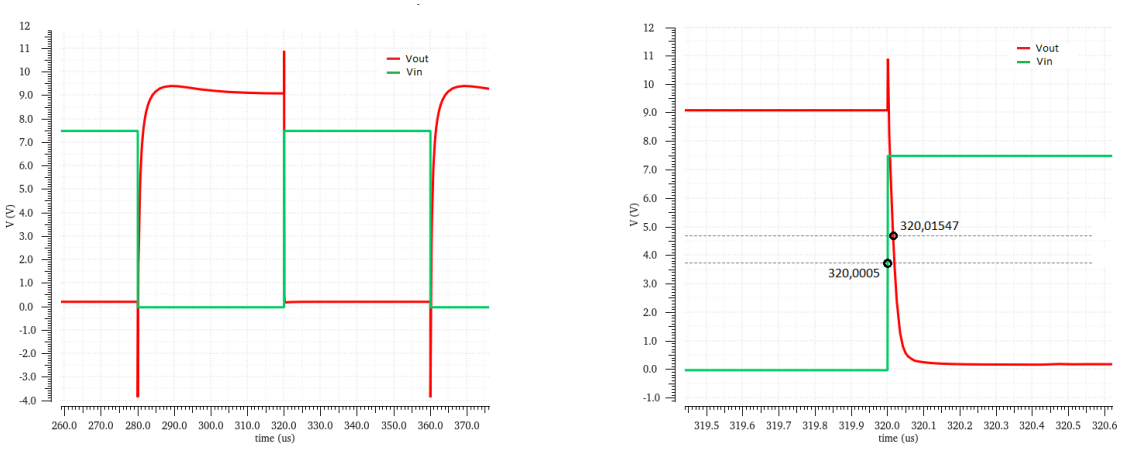
No caso do inversor 1, o gráfico obtido está representado na Figura 4.15b e medindo exatamente a 50% do sinal de entrada e de saída a diferença de tempo é aproximadamente 16,2 ns.



(a) (b)

Figura 4.15: Velocidade de propagação do Inversor 1

Relativamente ao inversor 2, pode-se ver que a velocidade de propagação é semelhante ao anterior, com valor cerca de 15 ns (Figura 4.16b). Pode-se notar que é ligeiramente mais rápido, mas nada muito significativo.

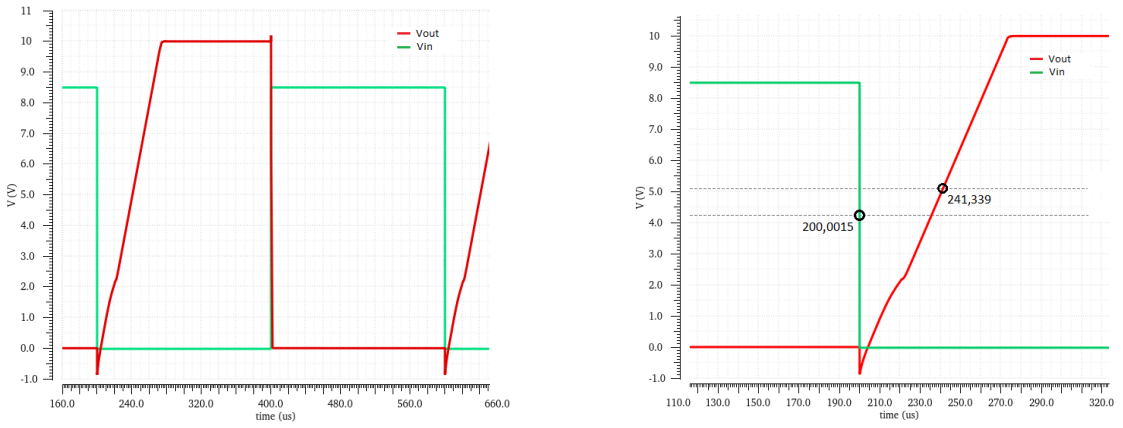


(a) (b)

Figura 4.16: Velocidade de propagação do Inversor 2

Por último, no inversor 3 obteve-se o atraso de propagação mais alto, embora na resposta DC seja a melhor entre os três. Para este caso a velocidade de propagação tem o valor de 41,34 µs como pode ser consultado na Figura 4.17b.

Como pode ser percebido, há uma pequena diferença no gráfico retirado para medir o atraso de propagação neste inversor comparativamente aos inversores previamente analisados. Na transição do sinal de 0 para 10 V pode-se perceber que essa passagem tem o formato de uma rampa em vez de ser mais espontâneo. Isto deve-se ao facto de inicialmente ambos os transístores estares a conduzir, mas chega a um ponto onde T_1 entra em corte e fica apenas T_2 a carregar com uma inclinação I/C.



(a) (b)

Figura 4.17: Velocidade de propagação do Inversor 3

De forma a perceber o resultado do último inversor, foram medidos os valores da corrente quando ocorre o ponto de comutação. Observou-se que o último inversor tem uma corrente bastante menor que as restantes configurações (cerca de 64,3 µA), por isso, as capacidades carregam mais lentamente resultando num atraso de propagação maior.

Novamente, foram medidas as velocidades de propagação, mas agora com uma carga igual para todos os inversores de forma a ser uma comparação mais justa, portanto, foi adicionado um condensador nas três configurações. O formato dos gráficos obtidos foram idênticos aos anteriores e as suas respetivas velocidades de propagação podem ser consultadas na Tabela 4.4, juntamente com os restantes valores. Pode-se perceber que nesta situação o inversor 3 continua a ter uma velocidade de propagação maior, portanto o tipo de carga não fez muita diferença.

Já é conhecido que o inversor 3 tem o seu ganho elevado, que é uma característica importante para a construção do esquema de inversores e por isso mesmo, terá de ser introduzido no final, caso contrário este definirá o ponto de comutação. Sendo assim, a duvida estava em utilizar o inversor 1 ou 2 no início da sequência. Na Tabela 4.4 podemos observar, de forma geral, que o inversor 2 é ligeiramente melhor que o inversor 1 em termos de ganho e atraso de propagação, contudo o ponto de comutação é inferior. Como o ganho vai ser decidido pelo inversor 3, interessa neste momento subir o ponto de comutação e com apenas um par de inversores 1 ou 2, não é suficiente. Como precisamos de 2 pares, o inversor 1 é uma escolha melhor, dado que necessita de menos transístores e o valor no pronto de transição é um pouco superior ao do inversor 2.

Tabela 4.4: Comparação entre os três inversores

Inv	Nº TFTs	Ponto Com.	Ganho	Vprop (carga inv)	I no Ponto Com.	Vprop (carga C)
1	2	2,2 V	7.5	16,2 ns	1,70 m A	8,2 ns
2	4	1,9 V	9.5	15 ns	1,88 m A	8,1 ns
3	2	0,8 V	$>> 10$	41,34 µs	64,3 µA	31 µs

Desta forma, a parte amplificadora do circuito do *Carver Mead* fica como esquematizada na Figura 4.18. O resultado desta sequência conjuga as características que são necessárias para o funcionamento desta componente. Através da Figura 4.19), é possível ver que o ponto de comutação foi inicialmente definido pelo inversor 1 e o ganho bastante elevado pelo inversor 3 como previsto. Neste ponto, pode-se passar à próxima fase do circuito do neurónio e ver como este se comporta.

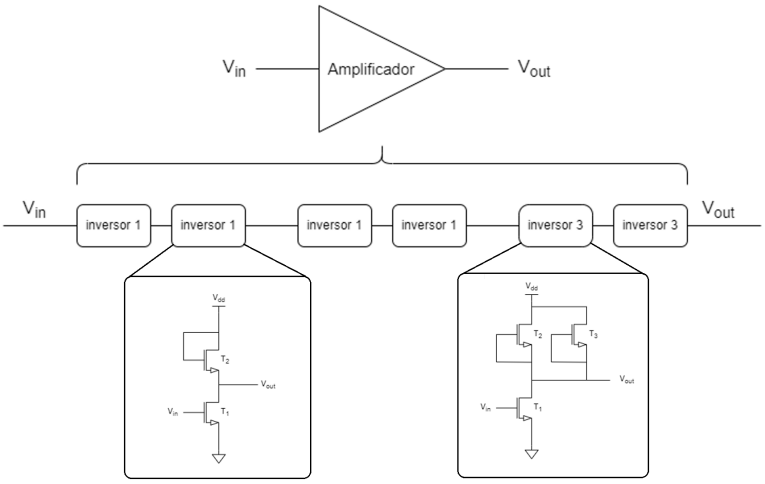


Figura 4.18: Esquemático com proposta de amplificador para substituir o par de inversores do circuito original

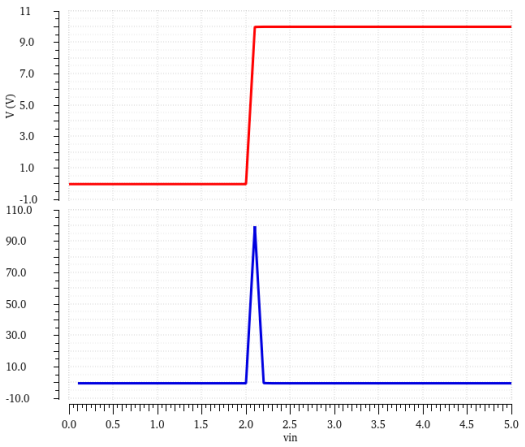


Figura 4.19: Simulação do amplificador proposto

Proposta do Circuito Final

Perante o estudo feito, idealmente o circuito do neurónio ficaria como representado na Figura 4.20.

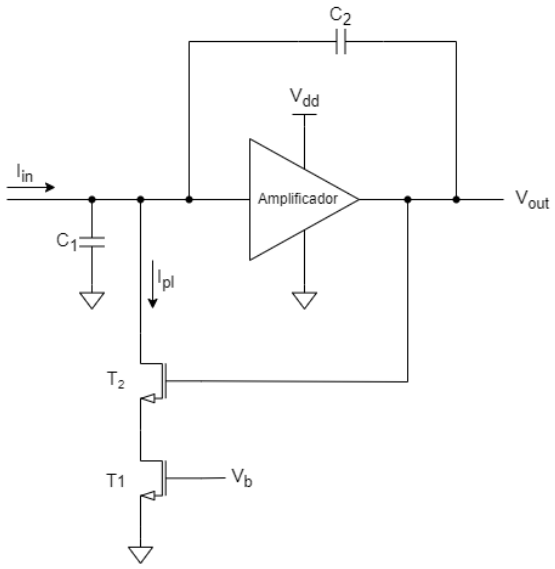


Figura 4.20: Primeira proposta de replicação do circuito proposto por *Carver Mead*

De forma a dimensionar a parte restante do circuito, procedeu-se à analise das equações que descrevem o comportamento deste neurónio de acordo com [8].

Considerando que ΔV representa a forma como a tensão varia na entrada do circuito amplificador (Figura 4.21), a sua relação é dada por

$$\Delta V = V_{DD} \cdot \frac{C_2}{C_1 + C_2} \tag{4.14}$$

Na parte de despolarização do neurónio, a equação que descreve essa transição no circuito é dada por

$$t_{high} = \frac{C \cdot \Delta V}{I_{Pl} - I_{in}} = \frac{C_2 \cdot V_{DD}}{I_{Pl} - I_{in}} \tag{4.15}$$

onde

$$C = C_1 + C_2 \tag{4.16}$$

Em relação à repolarização, esta é descrita como

$$t_{low} = \frac{C \cdot \Delta V}{I_{in}} = \frac{C_2 \cdot V_{DD}}{I_{in}} \tag{4.17}$$

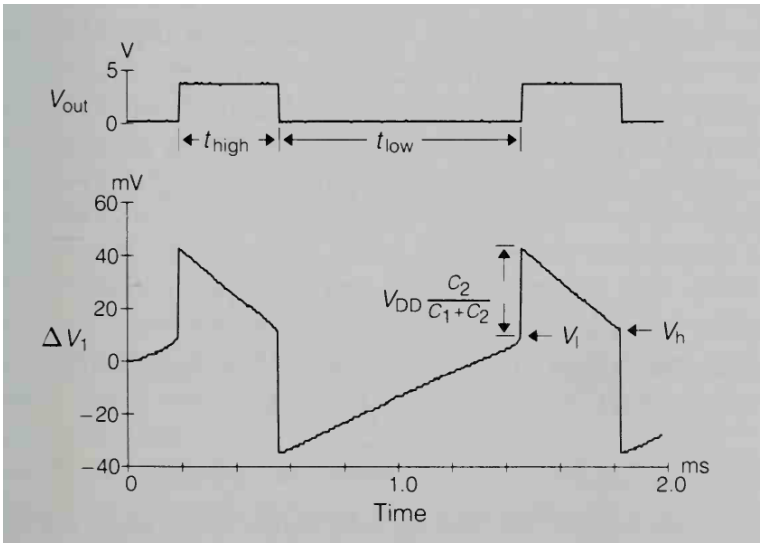


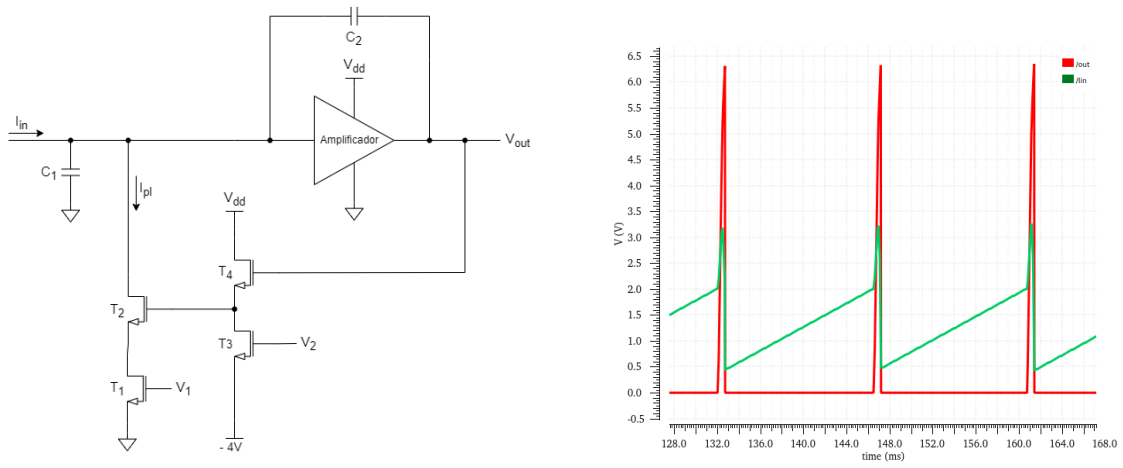
Figura 4.21: Variação da tensão na entrada do circuito amplificador, reproduzido de [8]

Já é sabido pelo amplificador dimensionado que o ponto de comutação é em torno de 2V, então nesse ponto é esperada a ocorrência de um pico em ΔV, que também desencadeia um pico na tensão de saída do circuito.

Arbitrando um valor de ΔV de 1V, e como também foi utilizado no circuito um V_{DD} de 10 V, fazendo o cálculo de (4.14) é obtida uma relação entre condensadores a ter em conta para o ΔV que se pretende. Sendo assim, o resto das variáveis dadas nas relações (4.17) e (4.15), podem ser obtidas arbitrando uma corrente I_{in} muito baixa (na ordem dos 10n) A e I_{Pl} bastante maior (200n A, por exemplo). Posto isto, falta escolher o tempo para a duração de t_{low} e t_{high}.

Após a análise de todas as tensões e correntes do circuito, foi detetado um problema na operação do mesmo. O transístor T_2 estava sempre em condução, fazendo com que a corrente de entrada passasse toda pelo ramo de T_2 e T_1. Isto impedia a formação de impulsos na saída, pois o neurónio nestas condições estará sempre a fazer a descarga da corrente (I_{Pl}). O motivo para isto acontecer está relacionado com o próprio transístor pois tem um valor de V_t negativo, então por mais que a tensão entre a porta e a fonte do T_2 fosse muito baixa, nunca era suficiente para o transístor estar em corte ($v_{GS} < V_T$). Isto aconteceu também em parte, porque o V_{out} nunca atingia o valor nulo devido à característica dos inversores.

Uma forma de contornar este problema foi acrescentar um *level shifter* composto por T_3 e T_4, com referência negativa (–4 V) no regulador de corrente T_3 de forma a que a tensão na fonte de T_4 seja bastante mais baixa e negativa. Como a porta de T_2 está ligada à fonte de T_4, esta também terá o mesmo valor negativo e T_1 recebe uma tensão muito pequena que mantém a fonte de T_2 próximo de zero. Com isto, é possível manter a condição de corte, por isso quando V_{out} sofre um pico o valor da fonte de T_4 vai subir e consequentemente a porta de T_2, permitindo assim a descarga da corrente I_{pl}. Estas alterações estão representadas na Figura 4.22a.



(a) Replicação do circuito proposto por *Carver Mead* com TFTs

(b) Tensão de saída V_{out} no neurónio em relação à tensão de entrada V_{in}

Figura 4.22: Proposta final do circuito do neurónio e respetiva simulação

Tabela 4.5: Dimensionamento do circuito do neurónio

W_1	150 µm	C_1	60 pF
W_2	260 µm	C_2	20 pF
W_3	280 µm	V_1	300 m V
W_4	280 µm	V_2	1 V
L	10 µm	V_{DD}	10 V

4.3 Circuito da Sinapse

De forma a criar uma rede neuronal é necessário implementar o elemento que permite a conexão entre neurónios: a sinapse. Assim, nesta secção é discutido o desenvolvimento de duas propostas que simulam este elemento.

4.3.1 Sinapse 1

A primeira proposta de sinapse tem como base o artigo [9]. Nesta referência são propostos vários circuitos que demonstram uma evolução, partindo de um circuito bastante simples até um bastante mais elaborado. A partir de uma determinada fase de evolução do circuito, é muito complicado replicar a sinapse devido as limitações dos TFTs. Consequentemente, o circuito desenvolvido tem como base o esquema da Figura 4.23.

De acordo com [9], na entrada é aplicado um pulso e quando este se encontra em cima, o nó V_{syn} diminui linearmente a uma velocidade imposta por $I_w - I_t$ e a corrente de saída I_{syn} aumenta exponencialmente. Quando o pulso se encontra em baixo, o V_{syn} é recarregado em direção à tensão de V_{DD} a uma velocidade imposta por I_t e a corrente I_{syn} diminui exponencialmente.

Os transístores M_{syn} e M_T juntamente com o condensador C_{syn} formam um integrador de espelho de corrente. Sendo assim, é possível perceber que I_t juntamente com C_{syn} controlam a forma

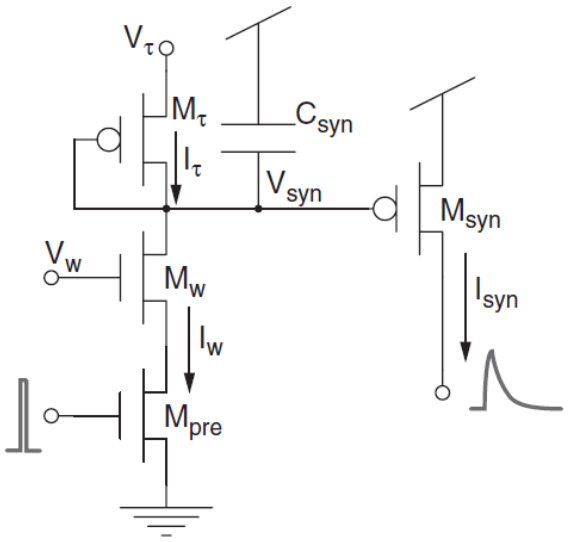
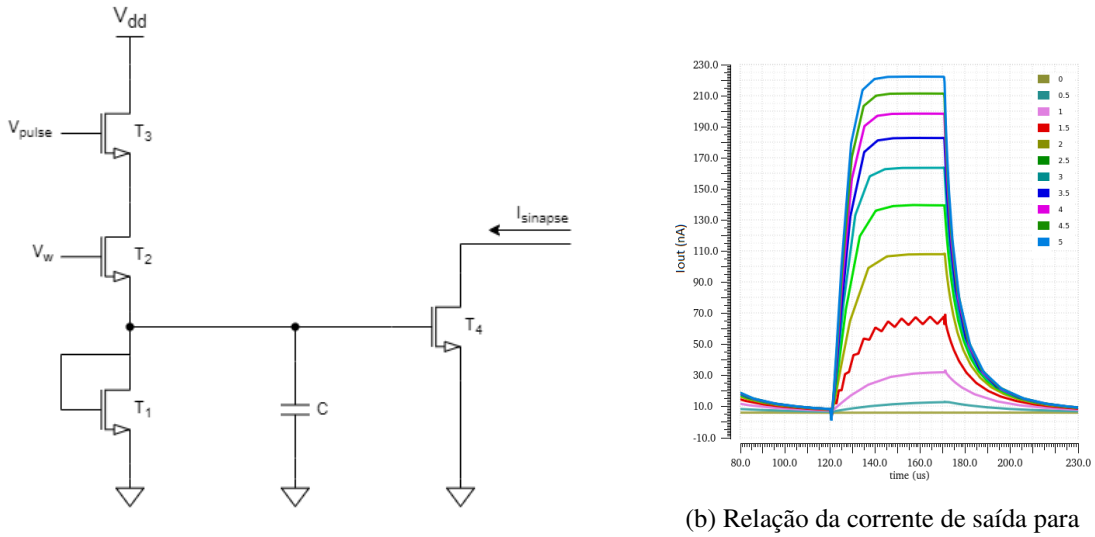


Figura 4.23: Circuito da sinapse, reproduzido de [9]

do carregamento e descarregamento do sinal. A tensão V_w permite controlar o peso (*weight*) da sinapse, isto é, a amplitude da corrente de saída.

Dado que o último transístor é responsável por estabelecer a corrente de saída, de forma a controlar a tensão na porta que impõe o v_{GS} que controla essa corrente é necessário inverter o circuito na replicação com TFTs. Isto significa que a parte superior composta pelos PMOS passa para a metade inferior do circuito só que em TFTs do tipo n com as alterações necessárias. Os NMOS passam para a parte superior mas com TFTs. Posto isto, o circuito proposto fica como representado na Figura 4.24a.

Caso não se invertesse o circuito, e apenas fosse feita uma conversão dos PMOS para o formato correspondente em TFTs, o circuito não funcionaria como esperado. Isto porque, na conversão dos PMOS, o M_t pode passar a ser um TFT chamado T_t com a porta e o dreno conectados e o M_{syn} apenas seria substituído por um TFT à qual podemos designar T_{syn}. Sendo assim, nesta configuração T_t atuaria como uma resistência de valor controlável a partir das sua relação W/L, impondo a corrente I_t desejada. Como este transístor estaria a ser alimentado por V_{DD}, a tensão na sua fonte, que por sua vez também é a tensão de porta no M_{syn} tem um valor muito próximo de V_{DD}. Para se obter o formado da corrente desejado que é controlado pelo v_{GS} de T_{syn}, é necessário ir de um v_{GS} relativamente baixo e quando o pulso de entrada estiver em cima, o v_{GS} aumenta de forma a gerar o pico. No caso da situação explicada, isso não acontece porque como mencionado a tensão na porta de T_{syn} seria muito alta podendo passar os limites de tensão em que o transístor opera corretamente (dependendo do valor de V_{DD}), e quando houvesse o impulso de entrada, essa tensão seria sempre mais baixa comparado quando o sinal de entrada é nulo. Isto tem o efeito contrário do que se pretende porque a corrente teria sempre um valor muito alto e quando na presença do pulso, esta ia diminuir.



(a) Circuito

(b) Relação da corrente de saída para diferentes V_w

Figura 4.24: Sinapse 1

Sendo assim concluiu-se que era necessário inverter o circuito como já descrito. Desta forma, a porta do T_4 tem uma tensão bastante mais baixa controlada por V_{pulse} e V_W e o transístor T_1 juntamente com C, integram a corrente que tem amplitude definida pelo peso representado por V_w. A relação dada pela variação de V_w com a amplitude do sinal está representado na Figura 4.24b.

De forma a perceber melhor o impacto que V_w tem na corrente de saída, foram medidos os valores da amplitude de I_{out} para diferentes valores de V_w. Os dados medidos não apresentam uma relação linear, tendo um erro máximo de aproximadamente 20 nA para V_w com valor de 5 V.

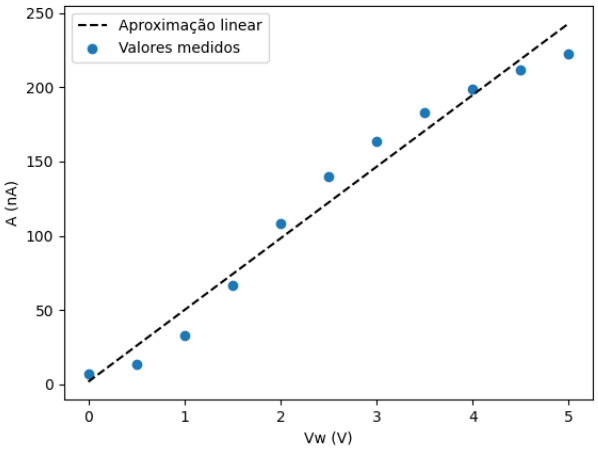


Figura 4.25: Relação de V_W com a amplitude da corrente

O neurónio está desenhado para receber na entrada correntes muito baixas. Sendo assim, a sinapse tem de produzir uma corrente com uma ordem de grandeza próxima, caso contrário

pode haver problemas de compatibilidade entre estes blocos eletrónicos. Posto isto, o circuito foi dimensionado de forma a ter como resultado uma corrente de saída baixa o suficiente para evitar o problema mencionado (Tabela 4.6).

Tabela 4.6: Dimensionamento da sinapse 1

W_1	320 µm	W_2	40 µm	W_3	50 µm	W_4	40 µm
L_1	10 µm	L_2	20 µm	L_3	20 µm	L_4	20 µm

4.3.2 Sinapse 2

A segunda sinapse proposta é um multiplicador diferencial [32]. De acordo com a referência, este multiplicador é também designado por *Gilbert Cell* com carga ativa, que conta com 3 pares diferenciais e dois transístores T_8 e T_9 que estão ligados em díodo atuando como resistências de valor $1/gm_{8,9}$ (Figura 4.26).

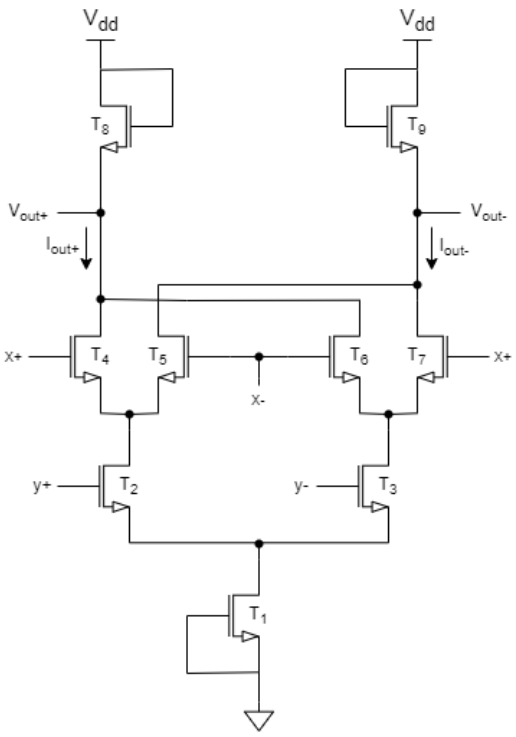


Figura 4.26: Circuito do multiplicador

O circuito é bastante semelhante a um dos representados do artigo, com exceção do transístor T_1. No circuito original, a corrente nesse transístor seria controlada por uma tensão externa, mas no circuito proposto o transístor T_1 atua como uma fonte de corrente. Como já visto, por se tratarem de transístores de depleção, quando o v_{GS} é nulo o transístor conduz. A corrente que passa em T_1 tem influência no resto do circuito, uma vez que esta é igual à soma das corrente em T_2 e T_3. Por

sua vez, cada uma destas correntes resultam da soma das duas correntes de cada par diferencial superior (T_4-T_5 e T_6-T_7).

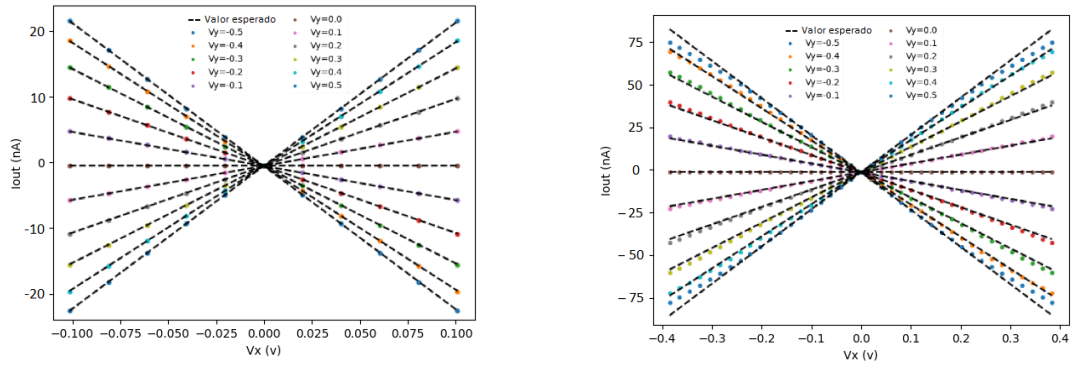
Se os transístores estiverem todos em saturação, de acordo com [32], para entradas diferenciais $V_x = x^+ - x^-$ e $V_y = y^+ - y^-$ pequenas, a corrente e tensão de saída podem ser dadas como

$$i_{out} = i_{out+} - i_{out-} \approx KV_xV_y \tag{4.18}$$

$$v_{out} = v_{out+} - v_{out-} = \frac{1}{gm_{8,9}} \cdot i_{out} \cdot \alpha V_xV_y \tag{4.19}$$

Para efeitos de simulação, foram então confirmados que os transístores se encontravam em saturação e portanto, a relação das correntes para as tensões aplicadas em V_x e V_y estão representadas na Figura 4.27. Foi imposto um V_x e V_y com valores de −0,5 V a 0,5 V. No gráfico, o V_y corresponde a cada uma das linhas ponteadas e no V_x, apesar de se ter importo valores $[-0.5, 0.5]$, como foram medidos os erros em relação à linearidade esperada (linha tracejada) até 1 e 10%, apenas são apresentadas as gamas de valores para esses erros. À partida o circuito apenas vai funcionar num quadrante do gráfico, para valores de *Iout* e *Vx* maiores ou iguais a zero.

Como já referido, a corrente em T_2 resulta da soma das correntes do par diferencial composto por T_4-T_5, e a corrente em T_3 segue exatamente a mesma lógica. Por este ponto de vista, os transístores $T_{2,3}$ devem ter o dobro do tamanho de T_{4-7}, assim como T_1 idealmente tem o dobro de $T_{2,3}$. Como existe a questão de a corrente ser numa determinada ordem de grandeza já mencionada, todos os transístores foram dimensionado de forma a que as correntes i_{out} estejam na ordem de valores necessários (Tabela 4.7). Os pares de transístores que formam os pares diferenciais devem ter o mesmo tamanho.



(a) Erro até 1% (b) Erro até 10%

Figura 4.27: Valor de I_{out} para uma gama de valores de V_x e V_y

Para uma melhor perceção de como este elemento pode ser usado no contexto de redes neuronais, se for aplicada uma tensão de entrada no circuito com o formato de pulso V_y e introduzida

Tabela 4.7: Dimensionamento do multiplicador

W_1	320 µm	$W_{2,3}$	40 µm	W_{4-7}	20 µm	$W_{8,9}$	40 µm
L_1	10 µm	$L_{2,3}$	20 µm	L_{4-7}	20 µm	$L_{8,9}$	20 µm

também uma tensão V_x, a corrente de saída (porque a sinapse tem a saída em corrente) terá valor KV_xV_y (Figura 4.28).

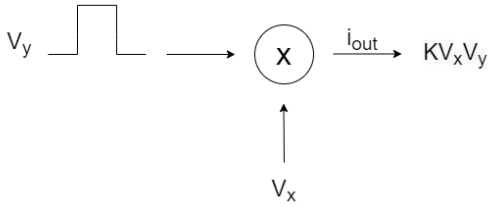


Figura 4.28: Aplicação do multiplicador

Como a saída deste circuito é diferencial, é necessário juntar o V_{out+} e V_{out-}. Para isto, pode-se recorrer a um amplificador, pois tem entrada diferencial mas uma saída única, complementando assim a sinapse 2 (Figura 4.29).

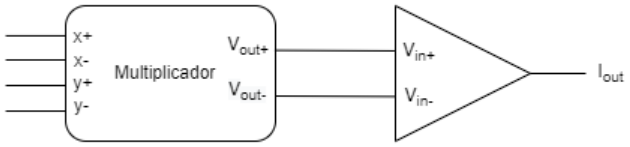
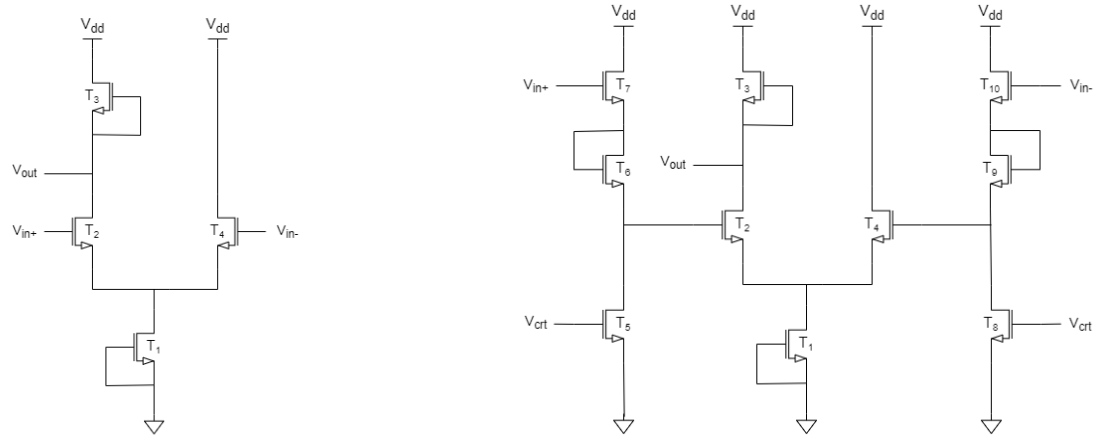


Figura 4.29: Esquemático da sinapse 2

A proposta para a parte do amplificador está representada na Figura 4.30a. Neste circuito o transístor T_1 atua como fonte de corrente, definindo as correntes do circuito. Como a corrente se divide igualmente no par diferencial se T_2 e T_4 tiverem o mesmo tamanho, o transístor T_3 deve ter metade do tamanho de T_1 para evitar *offsets*. Este circuito à partida resolvia o problema, mas como a entrada deste circuito é o V_{out} do multiplicador, torna-se necessário perceber o que se espera receber na entrada do amplificador, para tornar a ligação dos circuitos compatível. No ponto de operação DC do multiplicador, os V_{out+} e V_{out-} têm o mesmo valor de tensão muito próximo de V_{DD}, portanto é essa a tensão que entra em V_{in+} e V_{in-} do amplificador. Como o neurónio tem uma tensão média no nó de entrada perto dos 1,5 V (que será a tensão em V_{out} do amplificador) foi necessário reduzir a tensão que entra no amplificador para que seja possível conectar com o neurónio e não ter um v_{GS} muito elevado. Para isto acontecer, foi necessário introduzir um *level shifter* em cada entrada do amplificador como representado na Figura 4.30b. Dado que o v_{GS} não deve ser maior do que 5 V, foram necessários dois transístores para fazer a queda dessa tensão até aproximadamente 1,5 V na porta de T_2. O dimensionamento do circuito está representado na Tabela 4.8.

Tabela 4.8: Dimensionamento do amplificador

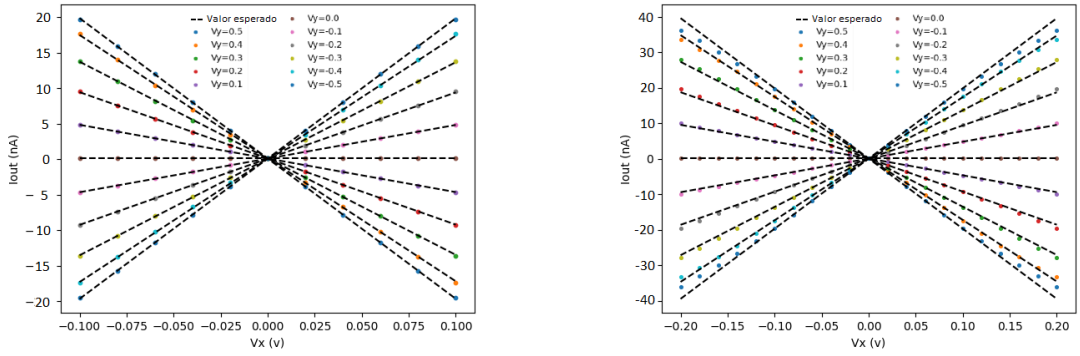
W_1	320 µm	$W_{2,4}$	80 µm	W_3	160 µm	W_{5-10}	200 µm
l_1	10 µm	$L_{2,4}$	20 µm	L_3	10 µm	L_{5-10}	10 µm



(a) (b)

Figura 4.30: Esquemático do amplificador para conectar com o multiplicador, em que (a) representa a proposta inicial e (b) o circuito final do amplificador

Desta forma, se este circuito for conectado ao multiplicador como demonstrado na Figura 4.29, a resposta é a que está representada na Figura 4.31. Para isto, foi aplicado o mesmo sinal no multiplicador como já feito para retirar a sua característica anteriormente, mas desta vez foi medido o valor da corrente no fim do amplificador. Sendo assim, os gráficos representam a corrente na saída da sinapse 2, com erro de aproximação até 1 e 10%.



(a) Erro até 1% (b) Erro até 10%

Figura 4.31: Valor de I_{out} do amplificador

4.3.3 Discussão

Como foram desenvolvidas duas sinapses, foi necessário escolher uma delas para planear o *layout* para a construção de uma rede neuronal. Em relação à primeira sinapse, existe um problema para conectar com o neurónio. A ideia passa por ligar o dreno de T_4 à entrada do neurónio e assim ter o ponto de conexão com uma tensão estável, no entanto a corrente $I_{sinapse}$ está no sentido contrário ao desejado. Posto isto, é necessário um circuito para completar a sinapse e ter uma corrente para entrar no circuito do neurónio. Devido a questões de tempo, não foi possível desenvolver a parte complementar que integraria na sinapse 1.

Relativamente à sinapse 2, esta encontra-se completa e funcional como já visto na secção anterior. Neste ponto, a escolha torna-se evidente que a opção escolhida é a sinapse 2, uma vez que não há a possibilidade de fazer uma comparação devido à sinapse 1 estar incompleta.

4.4 *Layout*

4.4.1 Rede Neuronal

De forma a integrar os elementos projetados numa rede neuronal, primeiro é necessário definir como esta estrutura será concebida. O esquema em *crossbar* é bastante utilizado para este tipo de redes e tem o formato representado na Figura 4.33.

Começando pelo funcionamento do neurónio, este pode receber vários sinais de entrada, os quais têm um peso associado (w). O neurónio realiza a soma desses sinais e posteriormente faz uma integração, tendo como resultado a saída final deste elemento (Figura 4.32). Posto isto, torna-se mais intuitivo perceber a forma como o *crossbar* (Figura 4.33) é estruturado. Pode-se ver que esta rede tem vários neurónios representados, e cada um recebe as correntes provenientes das sinapses associadas aos diferentes inputs.

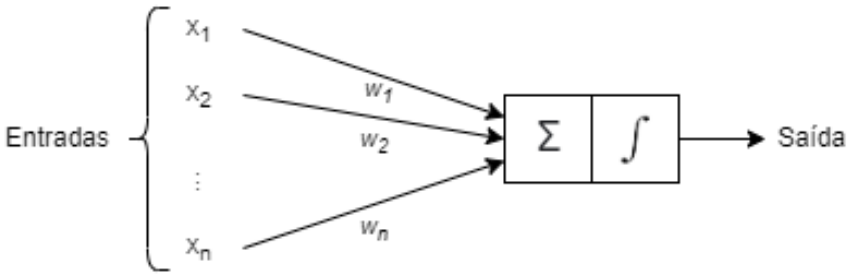


Figura 4.32: Diagrama do neurónio

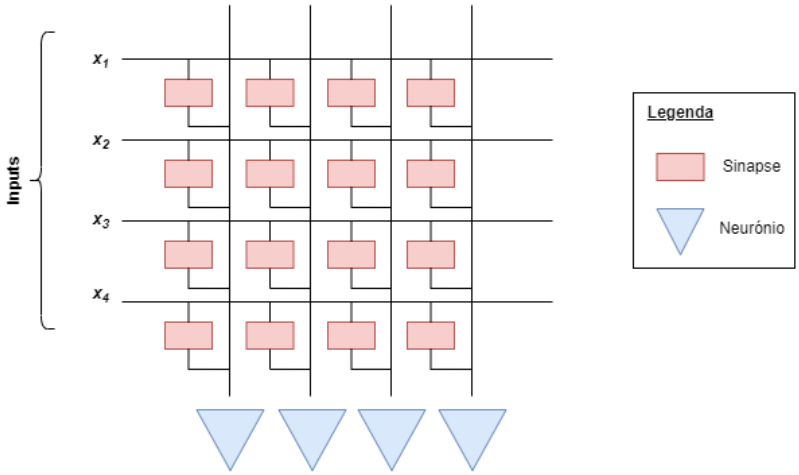


Figura 4.33: Esquema do *crossbar*

4.4.2 *Layout* dos Elementos

Como a mobilidade dos TFTs é mais baixa comparada a CMOS, o uso de transístores com um W grande é mais frequente, de forma a ter um ganho maior nas configurações. Geralmente, quando os transístores são muito largos faz-se um *fingered layout*, que consiste em dividir o transístor em n transístores mais pequenos, de largura W/n (o valor do L mantém-se). De acordo com [33], quando o *fingered layout* é feito para TFTs, este apresenta uma corrente no dreno maior comparativamente ao *layout* direto nas mesmas condições. Isto deve-se ao facto de o V_T sofrer alterações com esta estrutura, portanto para esta etapa não foi usado o *fingered layout*.

Os *layouts* realizados estão apresentado nas Figuras 4.34, 4.35 e 4.36. De forma a validar os *layouts*, estes tiveram de passar pela simulação *Design Rule Checking* (DRC) e *Layout Versus Schematic* (LVS), aos quais foram bem sucedidos.

4.5 Conclusões

Este capítulo foca-se essencialmente no desenvolvimento e avaliação de circuitos. Inicialmente foi feita uma revisão de algumas implementações em CMOS dos diferentes modelos neuronais já estudados, bem como as poucas implementações de TFTs em eletrónica neuromórfica encontradas até ao momento. Discutidas as diferentes considerações a ter para o desenvolvimento do neurónio, foi escolhido o modelo LIF como base de inspiração.

Como grande parte dos circuitos são implementados com CMOS, durante este capítulo são desenvolvidos circuitos com TFTs inspirados em existentes com CMOS. O primeiro foco foi o desenvolvimento do neurónio com base no circuito proposto por *Carver Mead*, que tem como estrutura o modelo LIF. No processo de desenvolvimento foram avaliadas possibilidades para substituir partes do circuito e no fim uma discussão que levou a uma proposta funcional. Dado que o neurónio serve para integração de redes neuronais, foi necessário desenvolver a sinapse. Posto

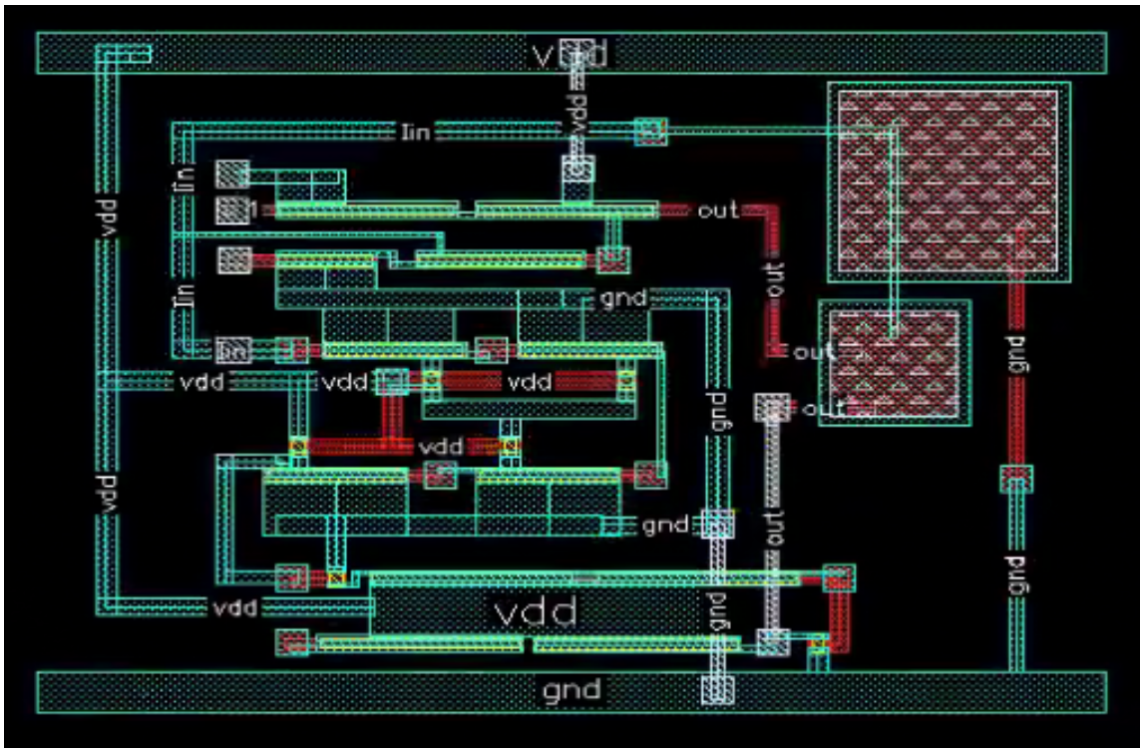


Figura 4.34: *Layout* do neurónio

isto, foram elaboradas duas propostas (sinapse 1 e sinapse 2), onde foi escolhida a sinapse 2 para a fase de *layout*. Isto porque, a sinapse 1 tinha um problema de compatibilidade com o neurónio que não foi possível ser resolvido.

Por fim, foi elaborado um esquema de ligações de cada elemento para *layout* e a sua posterior concretização.

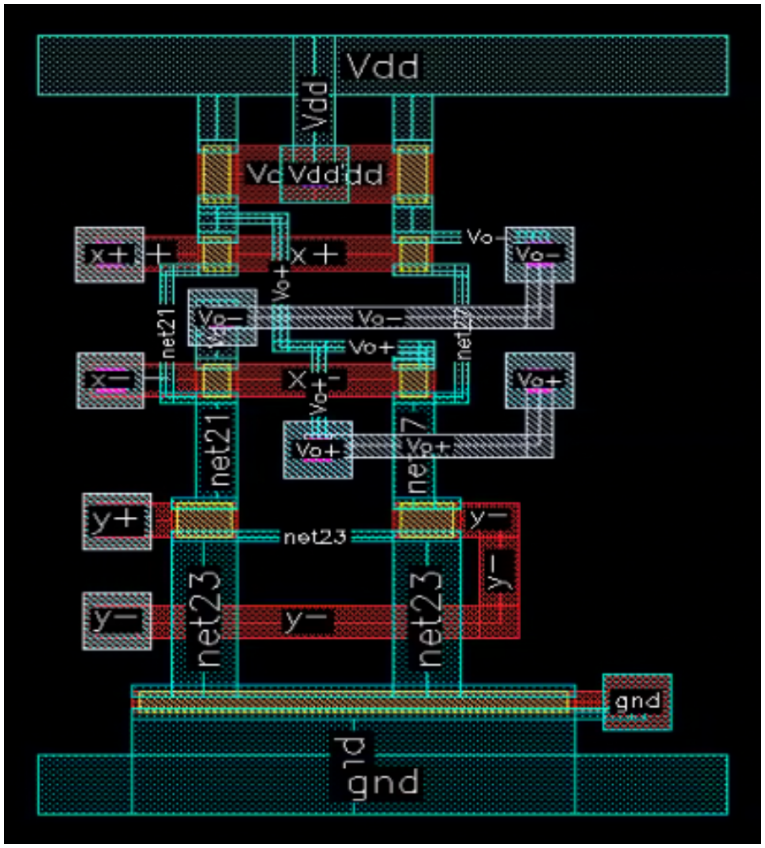


Figura 4.35: *Layout* do multiplicador

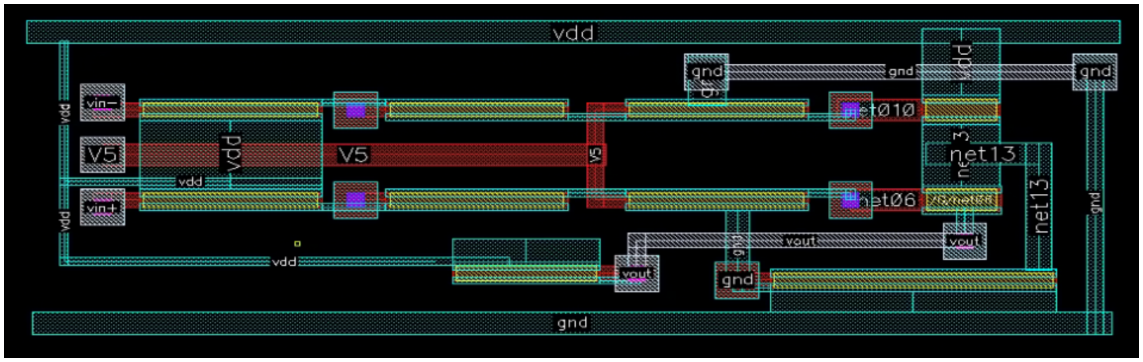


Figura 4.36: *Layout* do amplificador

Capítulo 5

Conclusão e Trabalho futuro

O foco principal deste trabalho é o projeto e implementação física de um neurónio eletrónico com TFTs. Alguns modelos de neurónios foram estudados, computadas e simulados. Foram revistas algumas simulações e implementações de circuitos integrados existentes relacionadas com os modelos apurados.

Seguiu-se uma revisão sobre TFTs, bem como o estudo do modelo já existente. Como haviam alguns problemas, então foi necessário fazer uma modelização do TFT para tentar evitar alguns problemas que podiam surgir na implementação de circuitos.

Como já referido, a eficiência energética constitui uma das importantes considerações a ter para escolha do modelo neuronal. Os modelos candidatos foram escolhidos com base no seu baixo nível de complexidade, uma vez que isto tem impacto direto no consumo de potência. O modelo LIF particularmente é o modelo mais simples e apesar de não ser o mais realista, quando integrado numa rede neuronal tem uma aproximação igualmente boa da dinâmica do cérebro humano e ao mesmo tempo é simples e tem uma boa eficiência energética. Sendo assim o circuito do neurónio foi desenvolvido com base num modelo proposto por *Carver Mead* que tem um comportamento semelhante ao modelo LIF. Durante o seu desenvolvimento surgiram alguns problemas e por isso foi feito um estudo minucioso do circuito e formas de contornar os obstáculos.

A integração do neurónio em redes neuronais apenas é possível se tivermos o elemento de ligação, designado de sinapse. Sendo assim, foi necessário desenvolver circuitos que se comportam como este elemento. Escolhida uma sinapse, chegou a fase de *layout* de cada circuito para construir uma rede neuronal em formato *crossbar*.

Sendo assim, de forma sucinta o trabalho realizado ao longo da dissertação conta com:

- Simulação de cada modelo neuronal ponderado;
- Modelização do TFT
- Desenvolvimento do circuito do neurónio
- Desenvolvimento de duas sinapses

- *Layout* do neurónio e da sinapse

5.1 Satisfação dos Objetivos

O objetivo principal da dissertação era a implementação de um neurónio. Como houve problemas com o modelo antigo do TFT, foi necessário acrescentar uma fase trabalho de forma a resolver as adversidades encontradas. Sendo assim, não foi possível finalizar pois não houve tempo para a fabricação do dispositivo.

Contudo, foram desenvolvidas duas sinapses que não estavam no plano de trabalho. Como o neurónio serve para integração de redes neuronais, fazia sentido investir algum tempo na elaboração de sinapses para esse objetivo.

Posteriormente, foi possível fazer alguns *layouts* embora a ideia de construir a rede neuronal não tenha sido concluída.

5.2 Trabalho Futuro

O trabalho que ficou por realizar conta com a fabricação do prototipo e posterior validação do resultado final.

Como o plano de desenvolver a sinapse era para integrar o neurónio numa rede neuronal, poderá ser feito o trabalho de construção desse esquema com os elementos escolhidos. Posteriormente, poderiam ter de ser feitos alguns ajustes no dimensionamento.

Tendo em consideração os circuitos desenvolvidos, futuramente podia-se avaliar a variabilidade paramétrica da tecnologia. Isto permitiria perceber se no caso de sinapse 2, a variação existente provocaria *offsets* muito elevados no circuito. Caso isso acontecesse, seria necessário algum trabalho de redimensionamento dos componentes.

Relativamente às capacidades parasitas implementadas, há uma parte que poderia ser desenvolvida. Como as capacidades foram implementadas segundo o modelo de *Meyer*, a transição entre a zona linear e de saturação é muito repentina. Sendo assim, poderia ser feita uma modelização para tornar esta transição mais suave.

Referências

[1] neuronio. https://commons.wikimedia.org/wiki/File:Neuron.svg. Accessed: 2022-02-20.

[2] Larry Abbott. Theoretical neuroscience.

[3] impulso. https://wikiciencias.casadasciencias.org/wiki/index.php/Impulso_nervoso. Accessed: 2022-02-20.

[4] o-que-e-sinapse. https://conhecimentocientifico.com/o-que-e-sinapse/. Accessed: 2022-02-20.

[5] Eugene M Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

[6] Giacomo Indiveri. A low-power adaptive integrate-and-fire neuron circuit. Em *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS'03.*, volume 4, páginas IV–IV. IEEE, 2003.

[7] Arianna Rubino, Melika Payvand, e Giacomo Indiveri. Ultra-low power silicon neuron circuit for extreme-edge neuromorphic intelligence. Em *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, páginas 458–461. IEEE, 2019.

[8] Carver Mead. Analog vlsi and neutral systems. *NASA STI/Recon Technical Report A*, 90:16574, 1989.

[9] Chiara Bartolozzi e Giacomo Indiveri. Synaptic dynamics in analog vlsi. *Neural computation*, 19(10):2581–2603, 2007.

[10] Taki Hasan Rafi. A brief review on spiking neural network-a biological inspiration. 2021.

[11] Jose M. Cruz-Albrecht, Michael W. Yung, e Narayan Srinivasa. Energy-efficient neuron, synapse and stdp integrated circuits. *IEEE Transactions on Biomedical Circuits and Systems*, 6:246–256, 2012. doi:10.1109/TBCAS.2011.2174152.

[12] Arash Fayyazi, Mohammad Ansari, Mehdi Kamal, Ali Afzali-Kusha, e Massoud Pedram. An ultra low-power memristive neuromorphic circuit for internet of things smart sensors. *IEEE Internet of Things Journal*, 5:1011–1022, 4 2018. doi:10.1109/JIOT.2018.2799948.

[13] Jayawan H.B. Wijekoon e Piotr Dudek. Integrated circuit implementation of a cortical neuron. páginas 1784–1787, 2008. doi:10.1109/ISCAS.2008.4541785.

[14] Wulfram Gerstner, Werner M Kistler, Richard Naud, e Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[15] Richard L Weisfield. Amorphous silicon tft x-ray image sensors. Em *International Electron Devices Meeting 1998. Technical Digest (Cat. No. 98CH36217)*, páginas 21–24. IEEE, 1998.

[16] Ana Paula Pinto Correia, PC Barquinha, e João Carlos da Palma Goes. *A Second-Order Σ ADC Using Sputtered IGZO TFTs*. Springer, 2015.

[17] P. Barquinha, L. Pereira, G. Gonçalves, R. Martins, D. Kuščer, M. Kosec, e E. Fortunato. Performance and stability of low temperature transparent thin-film transistors using amorphous multicomponent dielectrics. *Journal of The Electrochemical Society*, 156:H824, 2009. doi:10.1149/1.3216049.

[18] Stan D Brotherton. *Introduction to thin film transistors: Physics and Technology of TFTs*. Springer Science & Business Media, 2013.

[19] Pedro Barquinha, Anna M. Vila, Gonçalo Gonçalves, Luí Pereira, Rodrigo Martins, Joan R. Morante, e Elvira Fortunato. Gallium-indium-zinc-oxide-based thin-film transistors: Influence of the source/drain material. *IEEE Transactions on Electron Devices*, 55:954–960, 4 2008. doi:10.1109/TED.2008.916717.

[20] Ganga Bahubalindruni, Vítor Grade Tavares, Pedro Barquinha, Cândido Duarte, Rodrigo Martins, Elvira Fortunato, e Pedro Guedes De Oliveira. Basic analog circuits with a-gizo thin-film transistors: Modeling and simulation. páginas 261–264, 2012. doi:10.1109/SMACD.2012.6339389.

[21] AR T Sakurai. Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, 1990.

[22] Nishant Chandra, Apoorva Kumar Yati, e A. B. Bhattacharyya. Extended-sakurai-newton mosfet model for ultra-deep-submicrometer cmos digital design. páginas 247–252. IEEE Computer Society, 2009. doi:10.1109/VLSI.Design.2009.48.

[23] Adelmo Ortiz-Conde, FJ Garcıa Sánchez, Juin J Liou, Antonio Cerdeira, Magali Estrada, e Yaxing Yue. A review of recent mosfet threshold voltage extraction methods. *Microelectronics reliability*, 42(4-5):583–596, 2002.

[24] Colin C. McAndrew, Geoffrey J. Coram, Kiran K. Gullapalli, J. Robert Jones, Laurence W. Nagel, Ananda S. Roy, Jaijeet Roychowdhury, Andries J. Scholten, Geert D.J. Smit, Xufeng Wang, e Sadayuki Yoshitomi. Best practices for compact modeling in verilog - a. *IEEE Journal of the Electron Devices Society*, 3:383–396, 9 2015. doi:10.1109/JEDS.2015.2455342.

[25] Syed Ahmed Aamir, Yannik Stradmann, Paul Müller, Christian Pehle, Andreas Hartel, Andreas Grübl, Johannes Schemmel, e Karlheinz Meier. An accelerated lif neuronal network array for a large-scale mixed-signal neuromorphic architecture. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(12):4299–4312, 2018.

[26] Institute of Electrical, Electronics Engineers., IEEE Circuits, Systems Society., e Guo li Chenggong da xue. *2009 IEEE International Symposium on Circuits and Systems : circuits and systems for human centric smart living technologies, conference program, Taipei International Convention Center, Taipei, Taiwan, May 24-May 27, 2009*. IEEE, 2009.

[27] Institute of Electrical, Electronics Engineers., IEEE Circuits, e Systems Society. *2004 IEEE International Symposium on Circuits and Systems : proceedings : May 23-26, 2004, Sheraton Vancouver Wall Centre Hotel, Vancouver, British Columbia, Canada*. IEEE, 2004.

[28] Venkat Rangan, Abhishek Ghosh, Vladimir Aparin, e Gert Cauwenberghs. A subthreshold a vlsi implementation of the izhikevich simple neuron model. páginas 4164–4167, 2010. doi:10.1109/IEMBS.2010.5627392.

[29] Nobuyuki Mizoguchi, Yuji Nagamatsu, Kazuyuki Aihara, e Takashi Kohno. A two-variable silicon neuron circuit based on the izhikevich model. *Artificial Life and Robotics*, 16:383–388, 12 2011. doi:10.1007/s10015-011-0956-2.

[30] Pydi Ganga Bahubalindruni, Vítor Grade Tavares, Pedro Barquinha, Cândido Duarte, Nuno Cardoso, Pedro Guedes De Oliveira, Rodrigo Martins, e Elvira Fortunato. A-gizo tft neural modeling, circuit simulation and validation. *Solid-State Electronics*, 105:30–36, 2015. doi:10.1016/j.sse.2014.11.009.

[31] Mutsumi Kimura, Ryohei Morita, Sumio Sugisaki, Tokiyoshi Matsuda, Tomoya Kameda, e Yasuhiko Nakashima. Cellular neural network formed by simplified processing elements composed of thin-film transistors. *Neurocomputing*, 248:112–119, 7 2017. doi:10.1016/j.neucom.2016.10.085.

[32] Pydi Ganga Bahubalindruni, Vitor Grade Tavares, Jerome Borme, Pedro Guedes De Oliveira, Rodrigo Martins, Elvira Fortunato, e Pedro Barquinha. Ingazno thin-film-transistor-based four-quadrant high-gain analog multiplier on glass. *IEEE Electron Device Letters*, 37:419–421, 4 2016. doi:10.1109/LED.2016.2535469.

[33] Pydi Ganga Mamba Bahubalindruni. *Analog/Mixed Signal Circuit Design with Transparent Oxide Semiconductor Thin-Film Transistors*. Tese de doutoramento, Universidade do Porto (Portugal), 2014.