UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Engenharia Elétrica

Wellington de Oliveira Avelino

Implementação de Redes Neurais por Pulsos a partir de Sinapses Memristivas

Belo Horizonte
2022

Wellington de Oliveira Avelino

Implementação de Redes Neurais por Pulsos a partir de Sinapses Memristivas.

Versão final

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais como requisito para obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Prof. Gilberto Medeiros Ribeiro

Belo Horizonte – Minas Gerais

2022

Avelino, Wellington de Oliveira.
A948i Implementação de redes neurais por pulsos a partir de sinapses memristivas [recurso eletrônico] / Wellington de Oliveira Avelino. - 2022.
1 recurso online (112 f. : il., color.) : pdf.

Orientador: Gilberto Medeiros Ribeiro.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Inclui Apêndices.

.

Bibliografia: f. 92-99.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Inteligência artificial - Teses. 3. Redes neurais (Computação) - Teses. I. Ribeiro, Gilberto Medeiros. II. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.
CDU: 621.3(043)

Ficha catalográfica elaborada pela Bibliotecária Roseli Alves de Oliveira CRB/6 2361
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
COLEGIADO DO CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FOLHA DE APROVAÇÃO

"Implementação de Redes Neurais Por Pulsos A Partir de Sinapses Memristivas"

WELLINGTON DE OLIVEIRA AVELINO

Tese de Doutorado defendida e aprovada, no dia 20 de maio de 2022, pela Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais constituída pelos seguintes professores:

Prof. Dr. Jose Alexandre Diniz

Departamento de Semicondutores, Instrumentos e Fotônica (DSI)

Prof. Dr. Janaína Guimarães

Engenharia Eletrica (UFSC))

Prof. Dr. Wagner Nunes Rodrigues

DF (UFMG)

Prof. Dr. Jhonattan Cordoba Ramirez

DELT (UFMG)

Prof. Dr. Gilberto Medeiros Ribeiro - Orientador

UFMG

Belo Horizonte, 20 de maio de 2022.

Documento assinado eletronicamente por **Gilberto Medeiros Ribeiro**, **Presidente de comissão**, em 25/05/2022, às 15:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site

https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1469409** e o código CRC **5DBCF777**.

Referência: Processo nº 23072.229271/2022-11 SEI nº 1469409

Este trabalho é dedicado à minha família e a minha mulher, sempre presentes.

AGRADECIMENTOS

Agradeço à Deus por tudo.

Ao meu orientador Prof. Dr. Gilberto Medeiros pela oportunidade, paciência e pela parceria durante o desenvolvimento de minha tese.

Aos membros da banca os quais contribuíram com desenvolvimento do trabalho: Prof. Dr. Jose Alexandre Diniz, Profa. Dra. Janaína Guimarães, Prof. Dr. Wagner Nunes Rodrigues, Prof. Dr. Jhonattan Cordoba Ramirez.

Aos professores do PPGEE: Prof. Davies William, Profa. Luciana Pedrosa, Prof. Hugo Daniel, Prof. Frank Sill, Prof. Luís Antônio Aguirre, pela contribuição em suas disciplinas.

Aos meus pais Maria Edilva e Orlando André, e minha a tia Socorro por todo apoio e incentivo.

Agradeço à minha esposa Júnia Graciela pelo apoio e paciência durante esse período de formação.

Aos meus irmãos William, Wallison e Werbeson pelo companheirismo.

Aos colegas Fabiano Santana, Rafael Gonçalves, Leonardo Fonseca, Doug Ohlberg e Bernardo Lavall, por compartilhar momentos de discussão importantes no ambiente de laboratório.

Aos colegas do 3iT: João Henrique, Phillippe Drolet, Mathieu Valdenaire, Patrick Dufour por todo suporte técnico. Em especial aos professores Yann Beillard, Fabien Alibart e Dominique Drouin por terem me inserido em um ambiente magnífico de colaboração científica.

Aos coordenadores Prof. Dr. Frederico Gadelha Guimarães e Prof. Dr. Eduardo Mazoni Mendes por todo suporte ao longo do curso.

Aos membros do colegiado, corpo docente, secretariado e alunos do PPGEE.Aos representantes discentes. Em especial, ao Marcos Alves M.Sc. por todo suporte e dedicação a função.

RESUMO

As aplicações de inteligência artificial (IA) estão cada vez mais presentes e necessárias, principalmente as redes neurais (RN). A limitação no escalonamento da tecnologia CMOS (do inglês *complementary metal-oxide-semiconductor*) e a crescente complexidade computacional dessas aplicações exigem implementações de hardware mais energeticamente eficientes e escaláveis. As principais primitivas de computação de RNs são operações de multiplicações e acumulações que levam a um movimento significativo de dados entre memória e unidade de processamento nos sistemas computacionais baseados na arquitetura de von Neumann. Uma alternativa promissora é a mimetização da computação baseada em eventos, como em sistemas neuromórficos, colocalizando memória e processamento. Novos elementos de circuito inspirados no funcionamento neurológico representam uma nova alternativa para atingir a tão desejada eficiência computacional do cérebro, entre eles, uma série de dispositivos nanoescalares, conhecidos como memristores, foram propostos para serem usados como elementos fundamentais na construção de sinapses e neurônios artificiais. Nesse cenário, os esforços desse trabalho visam impulsionar a implementação de redes neurais por pulsos (RNP) com memristores à maturidade tecnológica. Essa tese foca em aspectos construtivos das redes, destacando metodologias para o acoplamento entre os elementos da rede estabelecendo condições satisfatórias para maximizar a eficiência no processamento da informação e implementação de técnicas de treinamento locais. Com esse propósito, uma plataforma de testes e um ambiente gráfico de interface com o usuário foram especialmente desenvolvidos e participaram na demonstração de uma rede neural inteiramente implementada em hardware a partir de sinapses memristivas, circuitos de neurônios a partir de dispositivos de NDR (do inglês *negative differential resistance*) e circuitos complementares. Ainda, experimentos prototípicos foram demonstrados para validar inferência e aprendizagem em redes neurais a partir desses componentes.

Palavras-chave: Inteligência Artificial, Sistemas Neuromórficos, Memristores, Redes Neurais por Pulsos, Treinamento Local, Transportadores de Corrente de Segunda Geração.

ABSTRACT

Artificial intelligence (AI) applications are increasingly present and necessary, especially neural networks (NN). The limited scalability of CMOS (complementary metal-oxide-semiconductor) technology and the increasing computational complexity of these applications require more energy efficiency and scalable hardware implementations. The main computational primitives of NNs are multiply-and-accumulate operations that lead to a significant data movement between memory and processing unit on von Neumann-based computational architectures. A promising alternative is the mimicry of event-based computing, as in neuromorphic systems, co-locating memory and processing. New neurologic-inspired circuit elements represent a new alternative to achieve the much-desired computational efficiency of the brain, among them, a series of nanoscale devices, known as memristors, were proposed to be used as fundamental elements in the creation of artificial synapses and neurons. In this scenario, the efforts of this work aim to boost the implementation of memristor-based spiking neural networks (SNN) to technological maturity. This thesis focuses on constructive aspects of networks, highlighting methodologies for network element coupling, establishing satisfactory conditions to maximize efficiency in information processing and implementation of local training techniques. For this purpose, a testing platform and a graphical user interface environment were specially developed for a demonstration of a fully hardware neural network based on memristive synapses, neuron circuits from NDR devices (negative differential resistance) and complementary circuits. In addition, prototypical experiments were demonstrated to validate inference and learning in neural networks from these components.

Keywords: Artificial Intelligence, Neuromorphic Systems, Memristors, Spiking Neural Networks, Local Training, Second-Generation Current Conveyors.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT: Associação Brasileira de Normas Técnicas
ADC: Analog-Digital Converter
ALU: Arithmetic Logic Unit
API: Application Programming Interface
CC: Corrente Contínua
CiFET: Carrier Injection Field Effect Transistor
CMOS: Complementary Metal-Oxide-Semiconductor
CNN: Convolutional Neural Network
CPU: Central Processing Unit
DAC: Digital-Analog Converter
DRAM: Dynamic Random-Access Memory
ECM: ElectroChemical Metalization
FCCC: Fonte de Corrente Controlada por Corrente
FTCC: Fonte de Tensão Controlada por Tensão
FTCT: Fonte de Tensão Controlada por Corrente
GPU: Graphics Processing Unit
H-H: Hodgkin e Huxley
HP: Hewlett-Packard
HRS: High Resistance State
IA: Inteligência Artificial
IF: Integrate-and-Fire
IoT: Internet of Things
LIF: Leaky Integrate-and-Fire
LRS: Low Resistance State
LTD: Long-term Depression
LTP: Long-Term Potentiation
MAC: Multiply-Accumulate
NDR: Negative Differential Resistance
NVM: Non-Volatile Memory
PCI: Placa de Circuito Impresso
PCM: Phase Change Memories

PE: Processing Element

PECVD: Plasma-Enhanced Chemical Vapor Deposition

ReLU: Rectifier Linear Unit

RDN: Resistência Diferencial Negativa

RF: Register File

RN: Redes Neurais

RNA: Redes Neurais Artificiais

RNP: Rede Neurais por Pulsos

SRAM: Static Random-Access Memory

TBJ: Transistor Bipolar de Junção

TCSG: Transportador de Corrente de Segunda Geração

TUJ: Transistor de Unijunção

UJT: Unijunction Transistor

VCM: Vacancies Change Memories

VLSI: Very Large-Scale Integration

SUMÁRIO

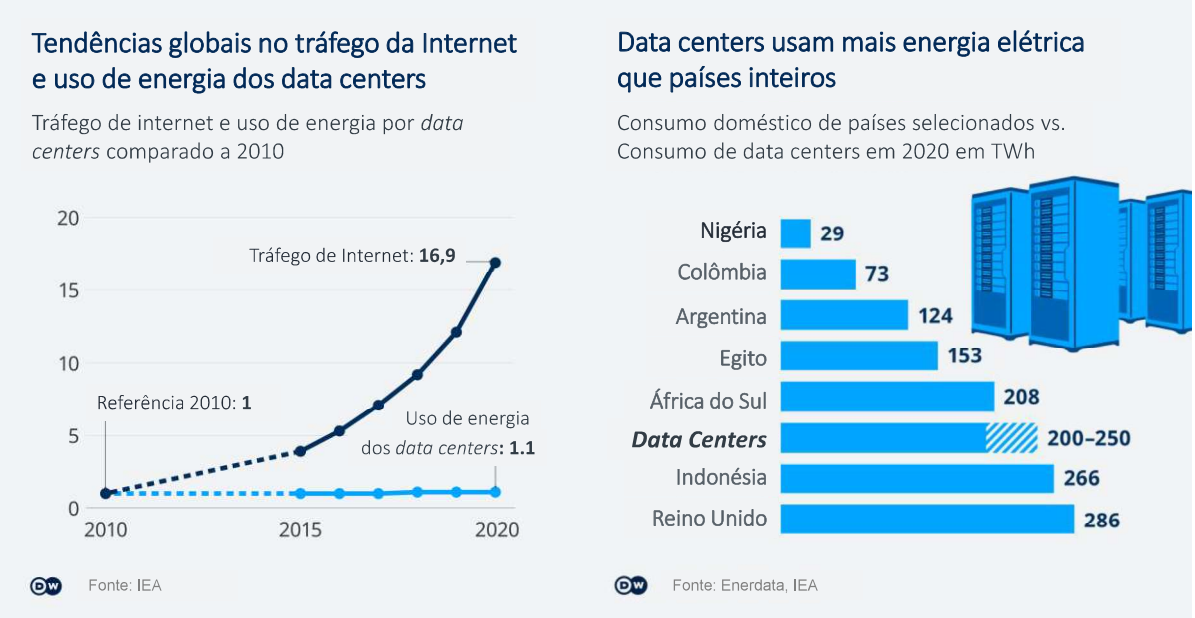
CAPÍTULO 1

INTRODUÇÃO

1.1. ASPECTOS ENERGÉTICOS SOBRE A INTELIGÊNCIA ARTIFICIAL

A inteligência artificial está cada vez mais presente no cotidiano, em diversas aplicações. Sob uma perspectiva de sustentabilidade, por exemplo, a implementação dessa tecnologia pode ajudar principalmente o setor energético, provendo previsões climáticas mais precisas, detectando perdas de energia ou ajudando a reduzir as emissões de CO_2 para diferentes tipos de indústrias. Sob outra perspectiva, a própria tecnologia tem como característica um elevado consumo energético e sua demanda crescente aponta para preocupantes impactos ecológicos (LEE; CHEN; CHAO, 2022). *Data centers* e grandes modelos de inteligência artificial (IA) usam grandes quantidades de energia e representam uma parte substancial da produção global de eletricidade. Embora esses *data centers* já se tornaram muito mais eficientes nos últimos anos e usam parcialmente energia renovável, o impacto em nosso planeta é considerável. Os *data centers*, que atualmente usam cerca de 200 TWh de energia por ano, cerca de 1,1% do consumo global, ver Fig. 1.1, possuem uma previsão de crescimento de cerca de uma ordem de magnitude até 2030(ROOKS, 2022). As consequências ambientais são geralmente negligenciadas e não é dada a devida importância para o crescimento da demanda energéticas para IA. Frequentemente, energia não é uma métrica considerada sobre o uso de IA, pois o foco está sempre na funcionalidade: performance, precisão, velocidade de resposta, etc.

Fig. 1.1 - Comparativos do consumo de energia para datacenters. Extraído de (ROOKS, 2022).

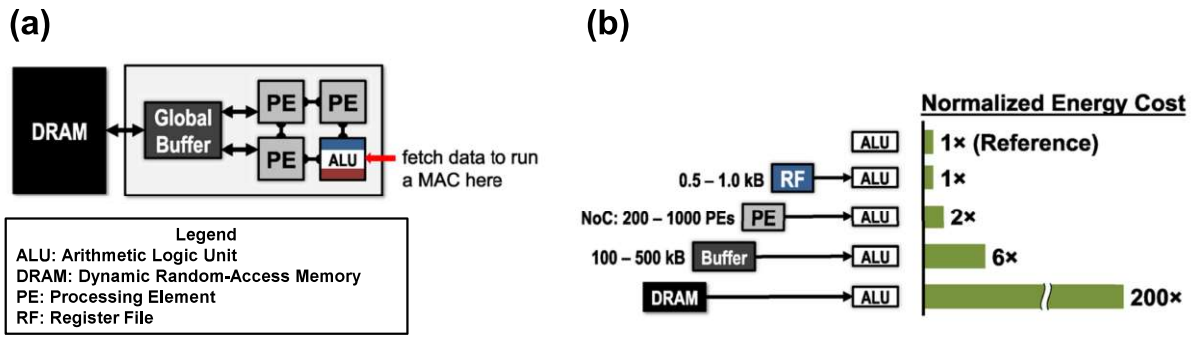


Ademais, as estimativas citadas não compreendem o surgimento de uma tendência em crescimento: aparelhos eletrônicos inseridos a redes de comunicação para controle e compartilhamento de dados. Uma nova geração de dispositivos conectados à internet (IoT do inglês *Internet of Things*) compõem o tráfego de dados remoto(FARHAN et al., 2017) e representam fontes de dados distribuídas que alimentam as aplicações de IA. A inserção desses dispositivos pode resultar em um aumento no consumo global de energia de cerca de 1,03% em 2030(KOOT; WIJNHOVEN, 2021). Considerando as aplicações de IoT, os dispositivos conectados geram altas taxas de dados, sob a exigência de serem analisadas em tempo real. Para cumprir os requisitos energéticos são estudadas novas arquiteturas computacionais que atuem como base para atender essas crescentes demandas.

No cenário de IA, as redes neurais (RN) são modelos computacionais mais comumente usados. Inspirados no funcionamento do cérebro, elas são capazes de realizar atividades complexas como encontrar padrões e relações entre dados brutos, agrupar e classificar dados, etc. RNs são definidas como cadeias de nós de processamento, chamados neurônios, interconectados por elementos que atuam na conexão de informação entre eles, as sinapses. Porém, grande parte das RNs são executadas em arquiteturas completamente diferentes dos sistemas biológicos, e por isso, o seu treinamento se torna uma atividade muito ineficiente. O treinamento de uma RN é baseado na definição dos pesos sinápticos, que são calculados e armazenados em memória. A cada atualização, o valor do peso é retirado da memória,

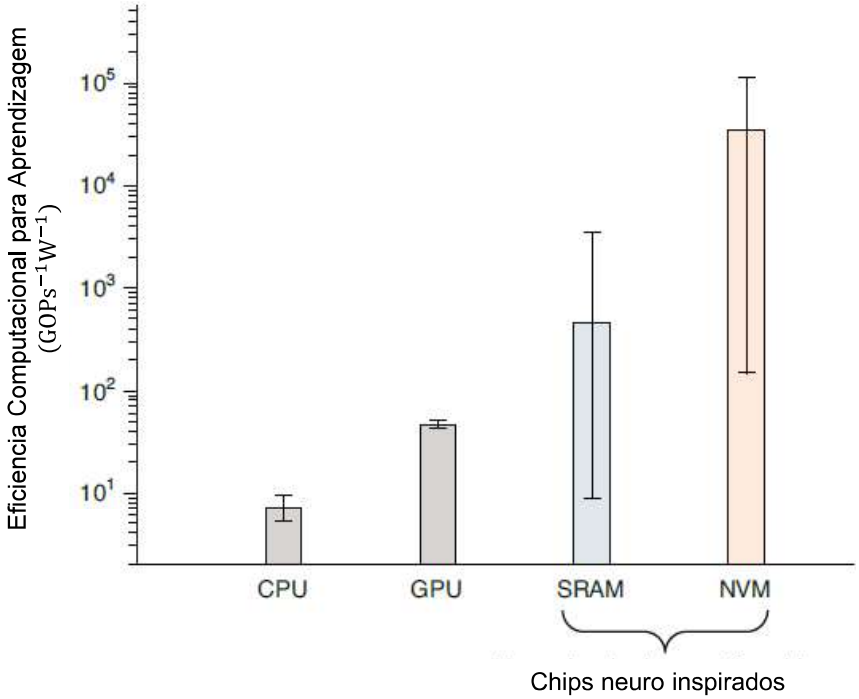
levado para o cálculo em uma unidade de processamento e regressado novamente a memória com valor atualizado. Considerando isso, uma grande quantidade de energia é usada para a movimentação de dados, que se torna cada vez mais grave quando a memória se distancia fisicamente da unidade de processamento, ver Fig. 1.2.

Fig. 1.2 – (a) Arquitetura computacional e (b) Comparativo dos custos energéticos para operações de redes neurais. Extraído de (SZE et al., 2017).



Ordens de grandeza em energia podem ser economizadas pela implementação física de sinapses e neurônios em chips neuromórficos. A razão é a aproximação do modo de processamento de informação realizado no cérebro, que pode ser reproduzido em circuitos integrados. Circuitos especialmente dedicados podem ser condicionados a representar algumas características dos sistemas computacionais biológicos como: paralelismo, computação em memória, técnicas de treinamento eficientes, dispositivos de baixo consumo e alta densidade. Sob a ótica da implementação com tecnologia CMOS e com a capacidade de implementação de processamento paralelo, é possível ganhar uma ordem de grandeza de eficiência energética comparada a CPUs ou GPUs, como mostra o gráfico da Fig. 1.3. Para esses últimos, onde os pesos sinápticos são armazenados em memórias CMOS como SRAM, esse tipo de implementação ocupa uma grande área sob um alto custo de fabricação(KIM et al., 2018).

Fig. 1.3 - Valores típicos de eficiência de energia para a fase de aprendizagem para algumas arquiteturas. Extraído de (ZHANG et al., 2020).



Problemas de escalabilidade são encontrados em implementações de circuitos neurais somente a partir de tecnologia CMOS. Como exemplo, para classificação de imagens, é preciso algo na ordem de centenas de milhões de neurônios e sinapses(DU et al., 2018). Tendo em vista a necessidade de colocar todos esses circuitos em chip de $1cm^2$, é preciso que cada elemento ocupe uma área menor que $1\mu m^2$. O que implica numa procura de representação de circuitos de tecnologias mistas e que cada vez mais se aproxima a escala nanométrica(ZHANG et al., 2020). A Fig. 1.3 mostra ainda que a existência da complementação entre nanodispositivos de memória, os memristores, com tecnologia CMOS, permite ganhar ainda mais duas ordens de grandeza em eficiência energética. Essa abordagem permite que novos dispositivos tenham um caráter multifuncional, e mimetizem muitos recursos computacionais presentes em sinapses e neurônios biológicos. Além disso, tecnologias emergentes de memória têm mostrado um enorme avanço sobre os processos de fabricação, demonstrado possibilidade de integração monolítica a tecnologia CMOS(GARBIN et al., 2015; SURI et al., 2011; VINCENT et al., 2015). Explorar novos dispositivos, circuitos e técnicas de treinamento, visando maximizar eficiência energética e densidade, são abordagens chaves para atender à crescente demanda energética de aplicações de IA.

1.2. IMPEDIMENTOS AO ESCALONAMENTO DE SISTEMAS COMPUTACIONAIS

Adequar as aplicações de redes neurais a uma arquitetura computacional eficiente, é a chave para atender à crescente demanda energética neste segmento. Os computadores convencionais ainda atuam sobre uma organização que se iniciou na década de 40. A arquitetura de von Neumann, que representa base para a implementação de sistemas computacionais, possui em essência recursos lógicos e aritméticos separados fisicamente dos recursos de armazenamento. O processamento de informações é dependente da comunicação entre CPU e memória, interligados por um barramento de dados, que limita a velocidade e eficiência energética das máquinas computacionais que utilizam essa arquitetura, o gargalo de von Neumann(BACKUS, 1978). O consumo energético para sistemas computacionais von Neumann tem sido estudado já há muito tempo, como em (HOROWITZ, 2014) que mostra que, experimentalmente demonstrado para tecnologia $45nm$, uma ordem de magnitude de energia a mais é gasta somente para transferência e acesso de dados à memória, em relação ao custo de energia para operações fundamentais. A arquitetura de von Neumann dá estrutura para toda a cadeia de computação impactando até no desenvolvimento de novos sistemas. O gargalo de von Neumann não é apenas uma restrição de movimentação de dados, trata-se de um conceito mais abrangente que influencia no modo de pensar em como se resolve problemas por meio da computação, separando as atividades em instruções que são executadas sequencialmente. Elaborar programas é basicamente planejar um enorme tráfego de palavras através de um gargalo, e apesar de atualmente existir sistemas computacionais que possuem múltiplos núcleos processadores que atuam simultaneamente, não abre possibilidade de paralelizar e redistribuir completamente o modo de processar informação(HILL; MARTY, 2008).

Vencer as barreiras até aqui apresentadas requer que pesquisadores pensem em fazer computação através de estratégias alternativas. Se exploradas com sucesso, soluções visionárias podem resolver simultaneamente limitações relacionados às exigências de energia crescentes, dependência excessiva de sistemas centralizados e dependência excessiva de arquiteturas baseada na movimentação de dados.

1.3. ABORDANDO PARADIGMAS DA COMPUTAÇÃO

Os modelos computacionais de redes neurais são fundamentados por conjuntos de operações de multiplicação-acumulação (MAC do inglês *multiply-accumulate*), cujas arquiteturas de von Neumann concentram cada parte desse par em unidades fisicamente distintas. Os sistemas de computação propostos nesta tese abordam técnicas de computação em memória (comumente conhecido na literatura pelo termo em inglês *in-memory computing*) que cointegra diretamente funções de computação e armazenamento(IELMINI; WONG, 2018). No nível de dispositivo, os sistemas computacionais propostos em escala nano compõem o uso de uma nova geração de nanodispositivos de memória, os memristores, que detém propriedades que melhoram aspectos energéticos e de performance para redes neurais. Características não-lineares encontradas em dispositivos nanoescalares reproduzem propriedades de neurônios promovendo melhorias em escalabilidade e eficiência energética pela mimetização desses elementos em escala nano. Por fim, uma arquitetura altamente propícia, dispondo dos dispositivos de memória em organização matricial proporciona vantagens funcionais e de fabricação(KIM et al., 2021) e viabiliza MACs via leis de Kirchoff e Ohm.

A busca pela implementação de redes neurais em hardware resultou em várias propostas de dispositivos, topologias de circuitos e técnicas de treinamento. Contudo, uma questão essencial diz respeito às restrições que uma arquitetura computacional enfrenta quando esses blocos de construção estão conectados a fim de formar circuitos. As implementações de sistemas computacionais em hardware compreendem um conjunto de elementos ativos e passivos interconectados onde as informações são codificadas na forma de sinais de tensão ou corrente que variam no tempo. A informação durante o processamento experimenta obstáculos como: atenuação de sinal, descasamento de impedância, pontos de operação distintos para diferentes dispositivos, *fan-in, fan-out*, para citar alguns problemas. Esses desafios criam dificuldades no dimensionamento de implementações de redes neurais.

Este trabalho mostra a implementação de uma arquitetura computacional que visa a maximização da eficiência energética, projetando máquinas computacionais bioinspiradas. Esta tese tende a promover uma análise aprofundada na implementação de redes neurais neuromórficas sob uma visão de tornar mais eficiente o transporte de informação ao longo da rede, concordando com requisitos de implementação de regras de aprendizado para memristores.

1.4. ORGANIZAÇÃO DESSE TRABALHO

Um dos objetivos desta tese versa sobre o uso de memristores como uma nova classe de dispositivos que emulam sinapses de redes neurais. Suas relações de impedância com os demais circuitos é o tema central, visando maior eficiência no processamento de dados. O estudo traz uma perspectiva de eficiência na transmissão de informação entre neurônios e destaca benefícios para o treinamento local em redes neurais baseadas em memristores. Esse trabalho é organizado da seguinte forma:

No capítulo 2, os elementos de circuitos das redes neurais são apresentados, dando uma perspectiva de operação e interação entre eles. A mimetização de algumas propriedades intrínsecas dos elementos que compõem uma rede neural biológicas por parte de dispositivos eletrônicos são discutidas.

No capítulo 3, é investigado como fluxo de informação se dá ao longo do seu processamento. Nesse aspecto, a relação entre as impedâncias dos elementos da rede é estudada a fim de destacar configuração de maior eficiência na transmissão de energia.

No capítulo 4, detalhes sobre a implementação de redes neurais são explorados, apresentando o projeto de redes neurais em torno de uma topologia de circuito de interface sinapse-neurônio que agrega as funcionalidades cruciais na maximização da eficiência das redes e implementação de regras de aprendizado.

No capítulo 5, resultados experimentais de implementações físicas de redes neurais são apresentados, destacando um teste de validação de aprendizagem associativa realizando ciclos de potenciação e depressão em um experimento prototípico de Pavlov.

No capítulo 6, uma recapitulação das contribuições neste trabalho é apresentada, destacando trabalho futuros.

CAPÍTULO 2

FUNDAMENTOS DA COMPUTAÇÃO NEUROMÓRFICA

Implementações físicas de sistemas neuromórficos surgem como candidatas a máquinas computacionais eficientes que desempenham papel importante para atender a demanda energética crescente. As redes neurais exibem capacidade de realizar atividades complexas como: predição, classificação, e filtragem de dados sob baixos orçamentos energéticos, estabelecendo-se condições para que a computação possa estar cada vez mais próxima de onde os dados são gerados. Neste capítulo, elementos de circuitos formadores de redes neurais são apresentados, tal como as especificidades de sua operação, destacando como a interconexão deles possam maximizar a eficiência computacional.

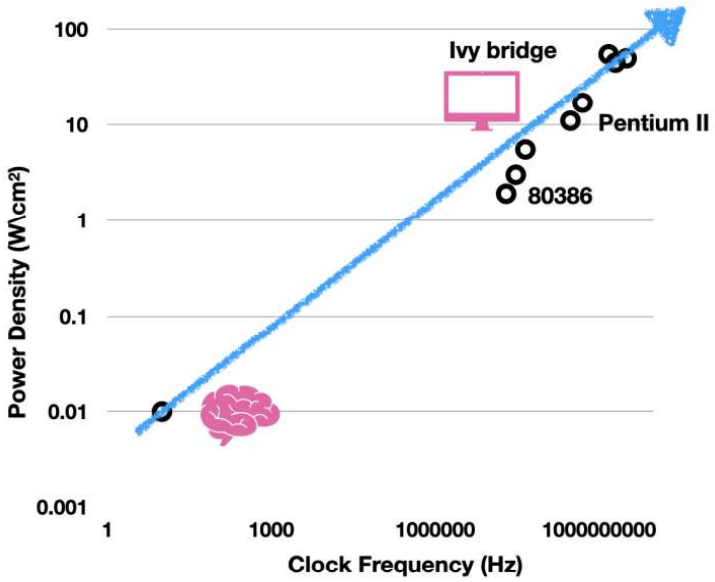
2.1. EFICIÊNCIA ENERGÉTICA INSPIRADA NO CÉREBRO

A exploração de sistemas computacionais inspirado no cérebro teve uma crescente destaque na década de 80, impulsionada pelas publicações de Carver Mead (MEAD; MAHOWALD, 1988; MEAD; ISMAIL, 1989). As projeções de Mead baseavam-se na eficiência energética dos sistemas computacionais convencionais, que mesmo sob uma condição de escalonamento extremo, não se equipariariam a capacidade de processamento de um cérebro humano. Por exemplo, Mead notou que o custo de uma simples operação computacional em um chip era estimado em $10^{-7}J$,

enquanto o custo de uma operação considerando um sistema completo (considerando todos os recursos agregados como periféricos, vídeo etc.) ficaria na ordem de $10^{-5}J$. Devido ao aprimoramento tecnológico esses números seriam menores hoje. Para a época, técnicas de escalonamento de circuitos integrados apontam na direção de uma redução dos parâmetros de consumo dos circuitos, apoiados pela diminuição dos níveis de tensão e corrente nos transistores. A visão de Mead baseia-se na suposição de que, mesmo assumindo "escala final" (densidade 100 vezes mais alta), ultrapassando $10^{-9}J$ por operação elementar no chip e $10^{-7}J$ por operação no sistema, os sistemas computacionais convencionais permaneceriam, na melhor das hipóteses, cerca de 10 milhões de vezes menos eficientes que o cérebro humano(MEAD, 1990).

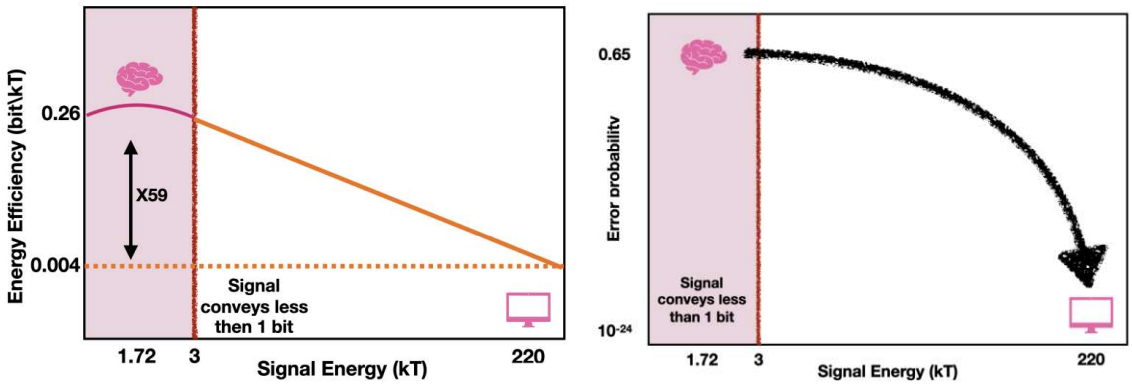
Estimativas da eficiência e orçamento energético do cérebro foram feitas a partir de medições de fluxo sanguíneo, como destacado em (MERKLE, 2007). O cérebro necessita para poder realizar todas as suas funções aproximadamente $20W$. Esse orçamento de potência é processado por um sistema com 10^{16} sinapses, interligando aproximadamente 10^{23} neurônios, compondo uma eficiência energética de $10^{-16}J$ por operação(KOCH; POGGIO, 1984). O cérebro opera em frequência mais baixa, $10^{1}Hz$, em comparação com os computadores digitais modernos, $10^{9}Hz$. Ademais, este "processador biológico" conserva uma densidade de potência melhor de $\approx 0{,}01\ W/cm^2$ em comparação com $50\ W/cm^2$ de computadores digitais, ver Fig. 2.1.

Fig. 2.1 - Comparação eficiência energética. Extraído de (TSUR, 2021).



O cérebro alcança extraordinária eficiência energética por realizar atividades computacionais através de dados sobre a representação por pulsos, e o registro de informações no domínio analógico, explorando propriedades físico-químicas de seus componentes, em vez de funções de base lógica digital como nos sistemas computacionais tradicionais. Ainda, os recursos de aprendizado, realimentação e adaptação permitem que um sistema computacional biológico otimize-se a partir de um comportamento estocástico, mantenha a robustez e a tolerância a erros, apesar da variabilidade intrínseca dos elementos, garantindo uma operação eficiente(SHARPESHKAR, 2010). Isso tudo faz com que o cérebro trabalhe no regime de $peta$ a $exaflops$. Comparativamente, o cérebro supera em muitas ordens de grandeza os sistemas convencionais, como apresentado no quadro comparativo da Fig. 2.2.

Fig. 2.2 - Comparação entre níveis de potência de sinal e tolerância a erro para sistemas biológicos e computadores convencionais. Extraído de (TSUR, 2021).



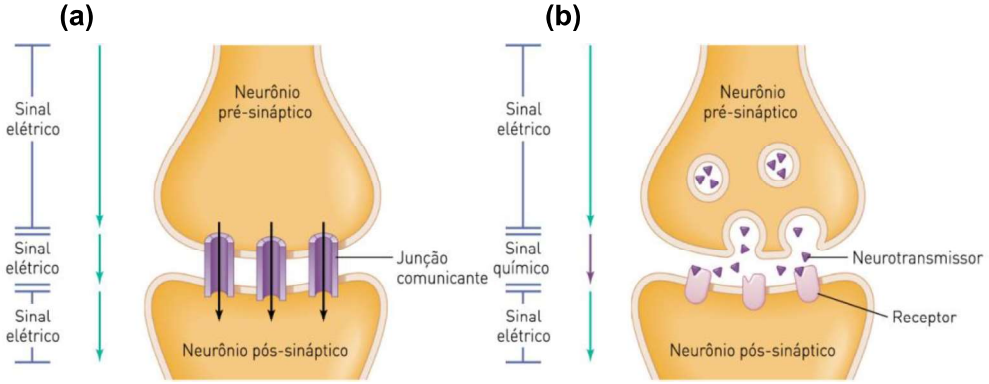
2.2. AS SINAPSES.

Em sistemas biológicos, a sinapse está localizada na interface entre neurônios e estabelece um elo para a transmissão de informação entre eles. Existem dois tipos de sinapses: a elétrica e a química. Na sinapse elétrica o potencial de ação é transmitido através de um fluxo direto de íons de um neurônio a outro, através de junções comunicativas, ver figura

Fig. 2.3-a. Na sinapse química, a transmissão do potencial de ação se dá através de uma agente químico, conhecido como neurotransmissor, ver fig.

Fig. 2.3-b. Ele atua na abertura dos canais iônicos que permitem a passagem de íons presentes na fenda sináptica, alterando o potencial interno ao neurônio. O peso sináptico entre dois neurônios pode ser representado pelo fluxo iônico através deles e acredita-se que a adaptação desses pesos atue no funcionamento do processamento da informação e está associado a aprendizagem dos sistemas biológicos (ZOMAYA, 2006).

Fig. 2.3 - Comparativo entre os tipos de sinapses: (a) elétrica e (b) química.



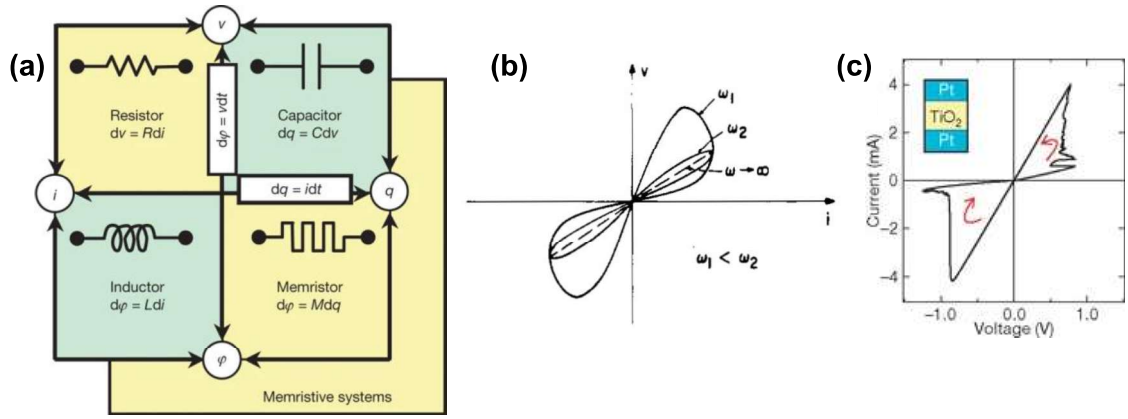
Em sistemas computacionais implementados em hardware, células de memória posicionaram-se como candidatos para emular sinapses. Uma série de dispositivos de armazenamento volátil representam os pesos sinápticos por meio do acúmulo de carga, como em memórias SRAM e DRAM(SIVAKUMAR; MALATHI, 2014), ou a base de transistores de porta flutuante estabelecendo armazenamento não-volátil, (RAMAKRISHNAN et al., 2013). Um outro conjunto de dispositivos emulam os pesos sinápticos na forma de resistência, como os memristores e suas classes de dispositivos: RRAM(JO et al., 2010), PCM(SURI et al., 2011), STT-MRAM(VINCENT et al., 2015). Todos esses dispositivos de memória possuem a mesma estrutura básica, e em particular, permitem a reprodução de protocolos de para a modulação sináptica biologicamente plausíveis.

2.2.1 Memristor como Sinapse

O Memristor foi um conceito proposto por Leon Chua em 1971, (CHUA, 1971), com a definição inicial do quarto elemento elétrico, relaciona carga e fluxo magnético, como mostrado na Fig. 2.4-a. Pelo postulado de Chua, das quatro variáveis, corrente i, tensão v, carga q e fluxo φ, era possível estabelecer correspondência entre elas, e

estabelecer relação com propriedades de dispositivos eletrônicos como: resistor, indutor e capacitor. Porém ainda não se observara uma relação entre carga e fluxo magnético incorporado em um dispositivo.

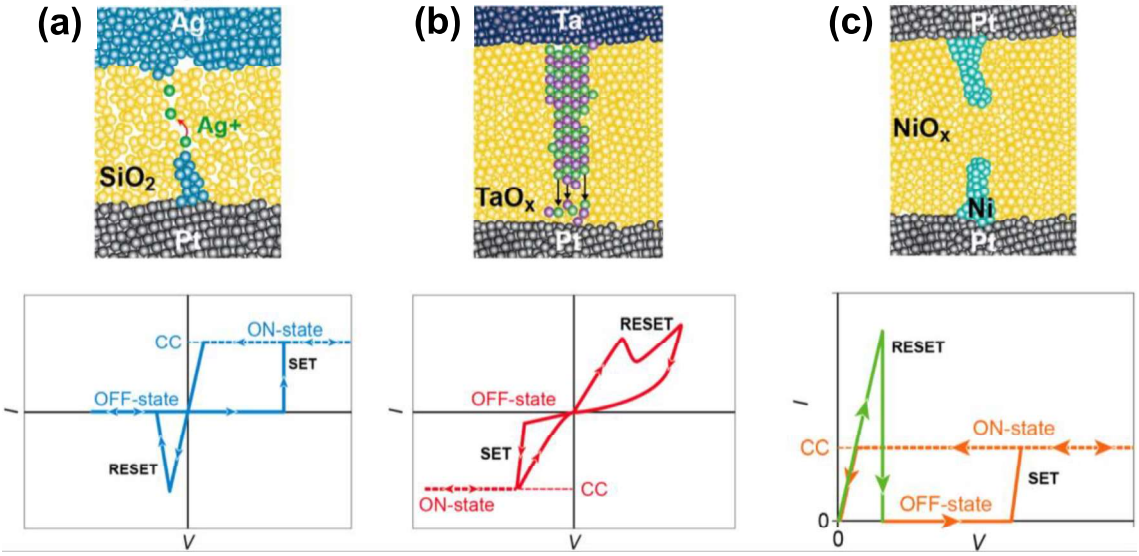
Fig. 2.4 – (a) O elemento postulado por Chua relaciona carga e fluxo magnético, (b) Curva IV do modelo foi observado em estruturas de óxido metálico. (c) em 2008 por pesquisadores da HP. Extraído de (CHUA; KANG, 1976; STRUKOV et al., 2008).



Em 2008, pesquisadores da Hewlett-Packard (HP) Labs publicaram que o elemento teórico havia sido fabricado na forma de células de comutação de dióxido de titânio (STRUKOV et al., 2008). O comportamento do dispositivo teórico, mostrado na Fig. 2.4-b, era observado em resultados experimentais, como na Fig. 2.4-c. Memristores são dispositivos passivos de dois terminais cuja resistência corresponde a história dos estímulos elétricos aplicados a ele. Esse dispositivo é constituído por eletrodos metálicos separados por um material isolante. A mudança de resistência é referente à dinâmica iônica promovida pela polarização de tensão e tem caráter não-volátil.

O chaveamento resistivo é ativado por campo elétrico a partir de um limiar de tensão. Então, para memristores, existe uma tensão limiar em que se estabelece as regiões de operação. A faixa de tensão abaixo da tensão característica se apresenta como uma região ausente de chaveamentos (região de leitura). E uma outra acima da tensão limite, é onde o dispositivo está propício a mudar sua resistência (região de escrita). O chaveamento resistivo pode ter como mecanismo físico principal o movimento de vacâncias de oxigênio, movimentação eletroquímica ou mudança de fase do material, como apresentada na Fig. 2.5.

Fig. 2.5 - Tipos de chaveamentos resistivos em memristores. Em (a) Comutação eletroquímica, em (b), por deslocamento de vacâncias e em (c), por mudança de fase. Extraído de (VALOV, 2017).



O efeito memristor por metalização eletroquímica (ECM do inglês *ElectroChemical Metallization*) baseia-se na oxidação de um eletrodo ativo (de Cu, Ag ou Ni) e a movimentação desses íons por difusão e deriva através de uma camada dielétrica (óxidos, selenetos ou sulfetos) em direção a um eletrodo inerte (Pt, W, Au) sob ação de campo elétrico de valor elevado (VALOV, 2017). Os íons dos eletrodos são reduzidos e formam um filamento condutor que liga os dois eletrodos e por consequência, diminuem a resistência da camada ativa passando para o estado de baixa resistência, Fig. 2.5-a Pelo fato do deslocamento ter como caráter iônico, a modulação da resistência é bipolar.

O efeito memristor da Fig. 2.5-b acontece pela polarização do dispositivo produzir o deslocamento vacâncias de oxigênio ao longo do óxido metálico (são classificados como VCM do inglês *Vacancies Change Memories*), permitindo comutar a resistência do dispositivo para estados de baixa e alta resistência. Esse efeito é não-volátil e por se tratar de deslocamento íons de oxigênio ao longo do óxido, a modulação da resistência é bipolar.

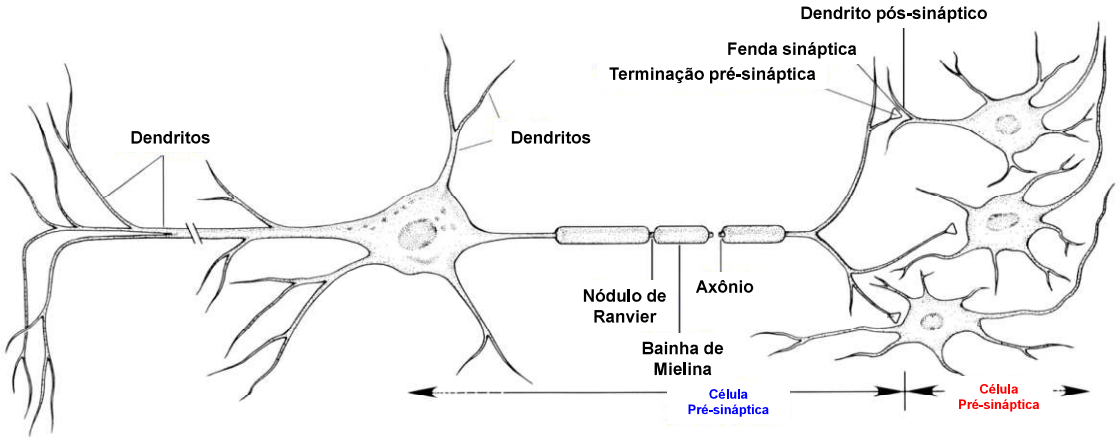
No mecanismo de mudança de fase (classificado como PCM do inglês *Phase Change Memory*), a camada isolante, geralmente um material a base de calcogeneto, sofre alteração de fase por aquecimento de joule: é inicialmente amorfo com um estado de alta resistência e depois se torna cristalino com estado de baixa resistência, ver Fig. 2.5-c. É um fenômeno reversível, com um pulso rápido de alta tensão e altos níveis de corrente, a camada ativa atinge seu ponto de fusão e depois é extinta para

o estado amorfo, estado de alta resistência. Com um pulso de tensão contínuo, respeitando uma lenta rampa de resfriamento, o material pode recristalizar e alternar para o estado de baixa resistência, conforme mostrado na Fig. 2.5-c.

2.3. OS NEURÔNIOS

Uma representação do neurônio é mostrada na Fig. 2.6. O estímulo de um neurônio para os neurônios adjacentes acontece por meio das terminações sinápticas, que atuam como pontos de transmissão, e de seus dendritos, que atuam como estruturas de recepção. Através deles, acontecem as trocas iônicas que modificam a diferença de potencial entre os meios intra e extracelular (conhecido por potencial de membrana). A região da fenda sináptica está populada por canais iônicos que são dependentes de tensão. Para as sinapses químicas, a aberturas dos canais permitem receber íons ou evacuá-los para o exterior. A flutuação de tensão ocasionada pelo movimento iônico é a base para a formação de pulsos que realizam a computação em sistemas biológicos.

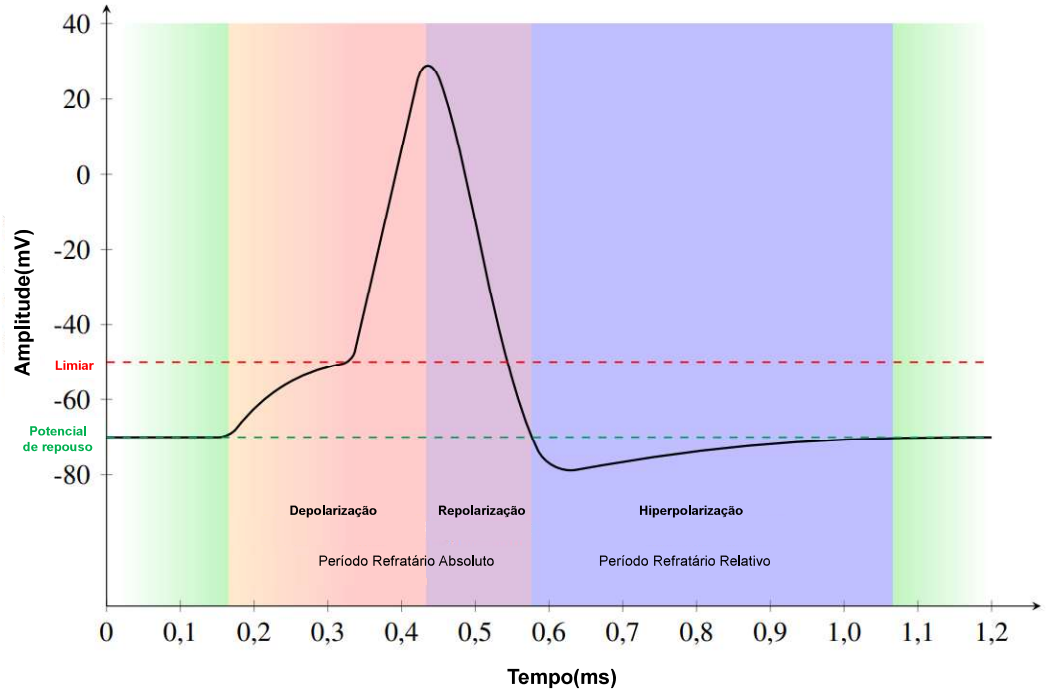
Fig. 2.6 - Representação de um neurônio, destacando suas partes.



A resposta de um neurônio à estimulação é um potencial de ação que percorre o axônio até as terminações sinápticas, ver Fig. 2.7. O potencial de membrana de repouso é cerca de $-70mV$. A estimulação proveniente de uma sinapse excitatória gera um aumento do potencial de membrana (despolarização). Se ultrapassar o limiar de $-50mV$, a despolarização torna-se mais abrupta, podendo chegar a $30mV$. O neurônio então inicia uma fase de repolarização e logo após hiperpolarização, onde o potencial se torna menor do que o potencial de repouso. O valor típico de a

hiperpolarização é $-80mV$. Por fim, o potencial de membrana retorna ao seu valor de repouso. Essas três fases descrevem um potencial de ação e abrangem um intervalo de aproximadamente $1ms$. Em situação contrária, quando o neurônio é estimulado através de uma sinapse inibitória seu potencial de membrana hiperpolariza, e, portanto, não estabelecendo chances de gerar um potencial de ação. Durante as fases de despolarização e repolarização, o neurônio está em um período refratário absoluto, ou seja, nenhuma estimulação irá gerar um novo potencial de ação. Por outro lado, durante a fase de hiperpolarização, diz-se que o neurônio está no período refratário relativo, pois se estimulado o suficiente durante este período, pode gerar uma novo potencial de ação.

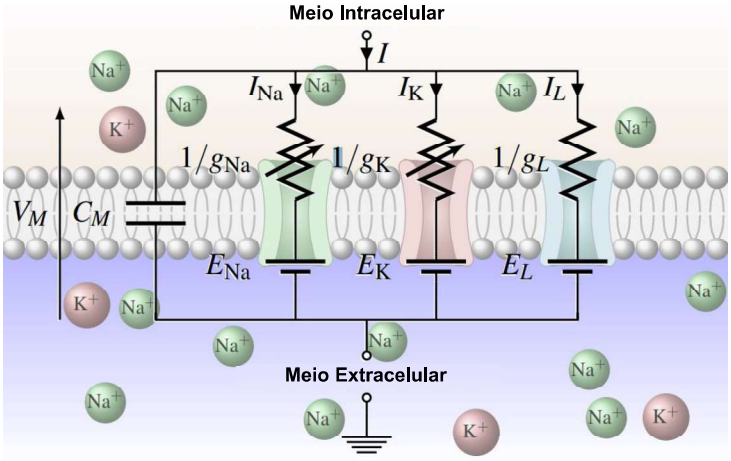
Fig. 2.7 - Evolução do potencial de ação neuronal



2.3.1 Modelos Neuronais

Redes neurais por pulsos empregam modelos de neurônios diferentes. Com base em observações biológicas, Hodgkin e Huxley (HODGKIN; HUXLEY, 1952), criaram um modelo levando em consideração as condutâncias dos canais iônicos presentes na interface dos neurônios. A Fig. 2.8 mostra o circuito elétrico equivalente sobreposto à membrana e canais iônicos de uma célula neuronal. Cada canal iônico é seletivo para um tipo específico do íon.

Fig. 2.8 – Representação esquemática do neurônio HH, modelando a membrana, os canais iônicos e a interface com o mundo extracelular.

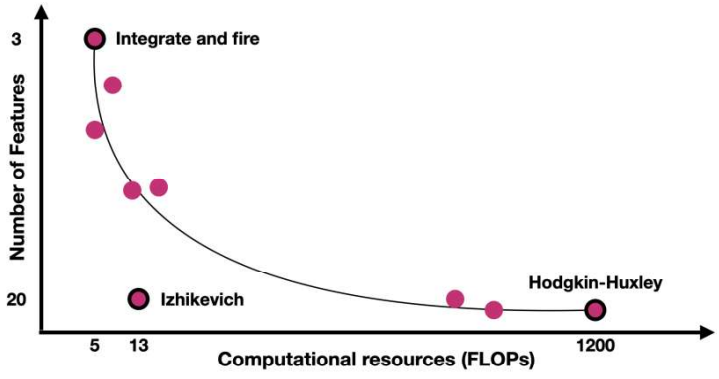


O modelo Hodgkin-Huxley (HH) foi elaborado a partir de três canais: o canal iônico de sódio, de potássio e um canal de vazamento (refere-se a L do inglês *leak*). A membrana é representada por uma capacitância elétrica. A equação diferencial resultante é:

$$C_M \frac{dV_M}{dt} = I - g_{Na}(V_M - E_{Na}) - g_K(V_M - E_K) - g_L(V_M - E_L) \quad (2.1)$$

O modelo FitzHugh-Nagumo, (FITZHUGH, 1961; NAGUMO; ARIMOTO; YOSHIZAWA, 1962), é uma simplificação matemática do modelo anterior, cujo funcionamento baseia-se em um oscilador de relaxação, e representa muitos dos fenômenos dinâmicos neuronais. Outros modelos de neurônio, principalmente investigado por Izhikevich (IZHIKEVICH, 2003), são abstrações matemáticas que omitem recursos dos neurônios a fim de reduzir os requisitos computacionais para simulações de maior escala, como mostra a Fig. 2.9.

Fig. 2.9 – Comparativo entre recursos de um neurônio e requisitos computacionais para modelização de RNP. Extraído de (TSUR, 2021).



Porém o modelo de neurônio mais utilizados é o integra-e-dispara, e sua variante com fuga, (LIF do inglês *leaky integrate-and-fire*)(GERSTNER; KISTLER, 2002). Trata-se de um modelo de neurônio de redes pulsadas bem simples que consiste em um processo de integração e uma detecção de limiar. Se a quantidade integrada, chamada potencial de membrana, exceder o limiar, o neurônio responde com um evento de saída e o potencial de membrana é reiniciado. Em sua versão no domínio dos circuitos, o integrador é representado por uma capacitância, emulando assim a o efeito de acumulação iônica desempenhada pela membrana lipídica em neurônios biológicos(HODGKIN; HUXLEY, 1952). A detecção limiar é representada por circuitos comutadores disparados por tensão que descarregará o potencial de membrana. Para esses últimos circuitos, a propriedade histerese é necessária a fim de promover condições suficientes para o descarregamento e compor parte do período refratário do neurônio. Os neurônios LIF são amplamente utilizados por sua simplicidade de implementação, ademais o emprego de uma classe de dispositivos comutadores é propicia para permitir a representação desses neurônios em uma versão compacta para redes neurais extremamente densas.

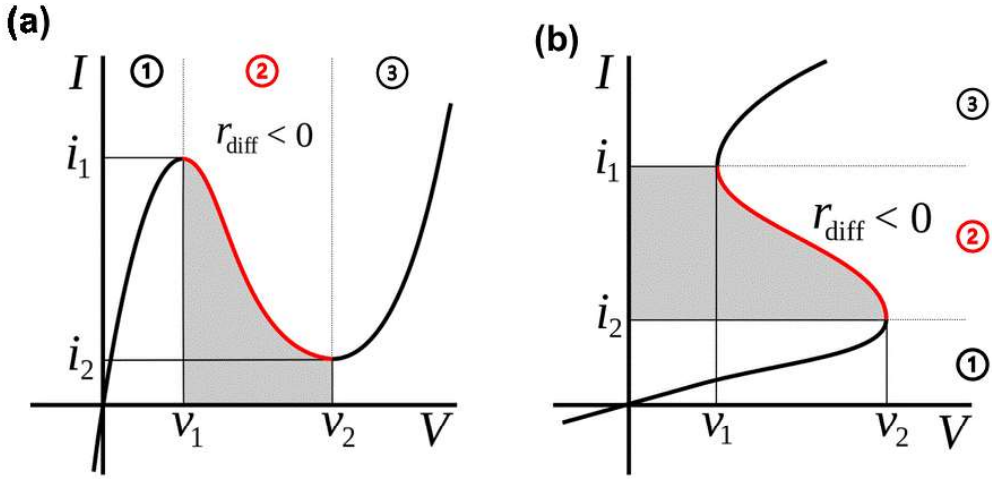
2.3.2 Neurônios Artificiais

Sob a ótica do funcionamento dos circuitos neuronais, o ponto de operação em que o neurônio está inativo (ou integrando) geralmente apresenta alta impedância. Circuitos que operam em modo corrente oferecem vantagem adicional para garantir o carregamento da capacitância de membrana, em oposição aos circuitos controlados por tensão (YI et al., 2018). Durante o disparo, um estado provisório de baixa impedância é apresentado a fim de produzir condições suficientes para o descarregamento da capacitância de membrana. A resposta dos neurônios durante o seu processamento é feita por meio de pulsos de tensão, onde a impedância de saída dessas implementações é geralmente alta, podendo ser suficiente o emprego de um *buffer* simples.

Sob uma perspectiva de circuitos, muitos das representações de neurônios artificiais compactas são implementados em uma topologia de multivibrador astável utilizando dispositivos de resistência diferencial negativa (NDR terminologia do inglês *Negative Differential Resistance*). Dois tipos de resistência diferencial negativa podem ser distinguidos: NDR controlada por tensão (VC-NDR) e NDR controlada por corrente

(CC-NDR). Ambas possuem aspectos diferentes na curva IV, como mostrado na Fig. 2.10.

Fig. 2.10 - a) NDR controlada por tensão e b) NDR controlada por corrente



A curva da Fig. 2.10-a é conhecido como NDR do tipo “N” que vem do formato da curva sobre o gráfico IV. Em tais dispositivos, a corrente é uma função de valor único e contínua da tensão; entretanto, a tensão é uma função multivalorada da corrente(KUMAR, 2008). Para melhor entendimento, a curva é dividida em três regiões 1, 2 e 3. Na região 1 da Fig. 2.10-a, à medida que a tensão aumenta, a corrente também aumenta (resistência positiva) até atingir um máximo de i_1 em v_1. Na segunda região da Fig. 2.10-a (marcada em vermelho) que representa a região da resistência diferencial negativa, à medida que a tensão aumenta, a corrente começa a diminuir até atingir um mínimo i_2 em v_2. E por fim, na terceira região da Fig. 2.10-a, à medida que a tensão aumenta, a corrente também aumenta apresentando novamente uma resistência positiva. Diodos túnel(KHAN, 2018), diodos de tunelamento ressonantes(KIDNER et al., 1990) e diodos de Gunn(YILMAZOGLU et al., 2007) são dispositivos que exibem VC-NDR.

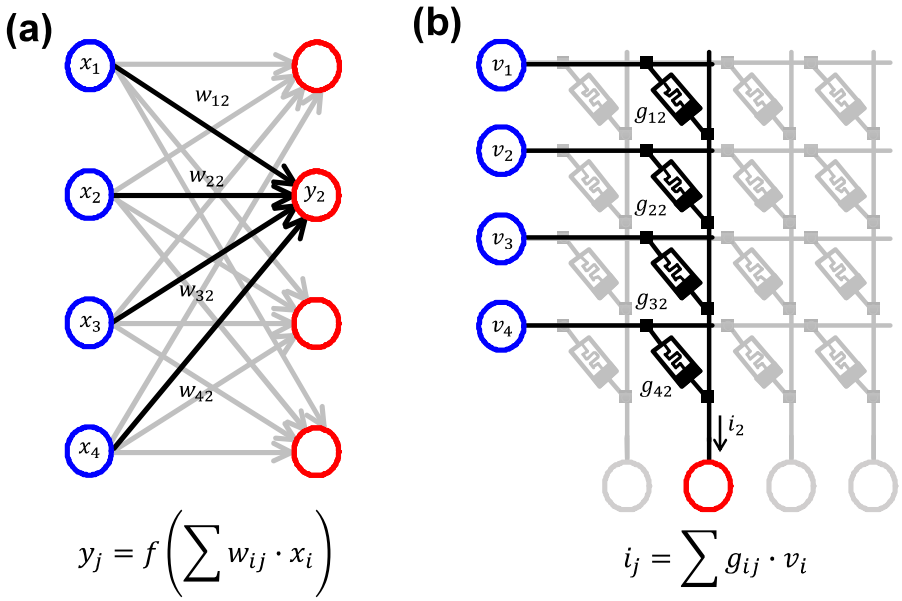
Na Fig. 2.10-b, o dual do comportamento de NDR controlada por tensão é apresentado, conhecido como resistência diferencial negativa tipo “S” que também se origina do formato sobre a curva IV. Ao contrário do primeiro tipo, a tensão é uma função de valor único da corrente, enquanto a corrente é uma função multivalorada da tensão(KUMAR, 2008). Na Fig. 2.10-b, regiões 1 e 3 que mostram resistência positiva, e a região 2 exibe resistência diferencial negativa. Este comportamento de CC-NDR é visto em dispositivos Mott(PICKETT; MEDEIROS-RIBEIRO; WILLIAMS,

2013), dispositivos de óxido metálico(WU et al., 2019), retificadores controlados de silício(ROZENBERG; SCHNEEGANS; STOLIAR, 2019), lâmpadas de descarga a gás(LOZNEANU; POPESCU; SANDULOVICIU, 2002) e transistores de unijunção(ASKEW, 1972), como os neurônios implementados na parte experimental deste trabalho.

2.4 REDES NEURAIS NO DOMÍNIO ANALÓGICO

A eficiência dos sistemas biológicos de computação, como mostrada na Fig. 2.2, inspira a criação de redes neurais sobre o domínio analógico. Essas redes formam uma cadeia interconectada de elementos de circuitos que exploram as características físicas dos dispositivos. O processamento analógico é realizado com operações em memória de funções de multiplicação-acumulação MAC (terminologia do inglês *multiply-accumulate*) diretamente em hardware, como apresentado na Fig. 2.11. Com dispositivos de memória analógica dispostos matricialmente em barras cruzadas, interligando representações neuronais periféricas, esta operação pode ser processada naturalmente usando sinais de tensão, gerados nos circuitos periféricos, convertidos em sinais de corrente sob as leis de Ohm e Kirchhoff (CHEN et al., 2021). Como apresentado na Fig. 2.11-b.

Fig. 2.11 – Comparação entre algoritmo(a) e configuração de matriz de barras cruzadas(b) para realizar operações de multiplicação-acumulação em redes neurais.



As sinapses atuam como dispositivos que interligam os neurônios, provendo armazenamento dos pesos sinápticos e simultaneamente a desempenhando a modulação do sinal processado. Esse mecanismo de operação elimina a movimentação de dados, evitando assim o gargalo de von Neumann. O dispositivo de memória nesse contexto realiza armazenamento e processamento, em comparação com as arquiteturas convencionais, onde essas duas operações estão fisicamente separadas. Um dispositivo promissor para a representação de sinapses artificiais é o memristor (JO et al., 2010), o qual pode naturalmente desempenhar esse papel de computação na memória em sistemas neuromórficos. A capacidade de alocar no mesmo elemento características de dispositivo de memória não volátil e realizador de operações de computação, pela conversão corrente em tensão, faz dos memristores componentes mais adequados para implementar redes neurais eficientes. Na Fig. 2.11-b, os elementos que constituem a periferia da matriz são os neurônios. Eles concluem o processamento de informação recebendo os estímulos de outros neurônios modulados pelas sinapses. Para as redes neurais pulsadas, o neurônio basicamente permanece inativo até que seu estado interno alcance um limiar, e gere uma resposta em forma de pulso.

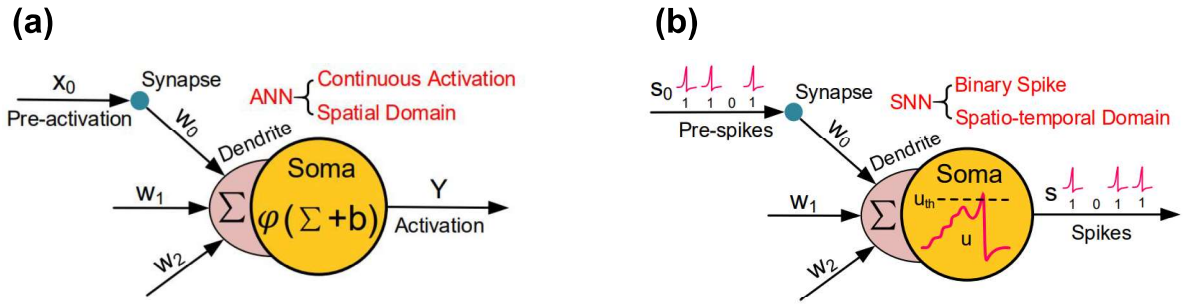
Sob o ponto de vista sistêmico, os blocos de construção realizam funções distribuídas e paralelas, o que contribuem para tornar mais eficiente as atividades computacionais. Porém, sob um ponto de vista individual, é preciso considerar que as mimetizações de sinapses e neurônios, distinguem-se no que concerne às condições de polarizações, quadrantes de funcionamento, etc. É preciso considerar que para o funcionamento das redes neurais, cada elemento apenas atua de forma apropriada sob condições de operação específicas, e por isso é preciso “desacoplar” o funcionamento de sinapses e neurônios. Sob um ponto de vista de processamento de sinal, as informações serão representadas por sinais elétricos (ora representados em tensão, ora representados em corrente), que enfrentaram diferenças e variações de impedâncias, na transmissão do sinal de um elemento a outro. Logo, para condicionar o funcionamento das redes neurais, circuitos que viabilizem o ajuste de impedância a fim de conservar a integridade do sinal a ser processado são também necessários. E por isso, o entendimento do comportamento dos dispositivos é crucial para a composição das redes.

2.5 REDES NEURAIS: UM COMPARATIVO ENTRE 2° E 3° GERAÇÕES

Modelos computacionais de redes neurais datam de mais de 50 anos, (PHAM, 1970), e a evolução nos seus modos de processamento foi evoluindo à medida que esse tipo de computação foi inserido para resolução de atividades cada vez mais complexas. As mudanças na representação dos dados, como: do domínio discreto ao contínuo, e após no espaço-temporal, vieram a ponto de se adaptar-se à demanda escalar de sistemas de inteligência artificial, e a diversidade de problemas a que estão propostas a resolver. Avanços no modo de processamento e representação dos sinais neuronais caracterizam as gerações de redes neurais (MAASS, 1997), e incorporam recursos que se assemelham a redes neurais biológicas. A primeira geração de redes neurais consistiu em neurônios que atuam baseado em limiares como no modelo neuronal de McCulloch-Pitts(MCCULLOCK; PITTS, 1956). Conceitualmente simples, essa rede reponde com valores binários lógicos se a soma de seus sinais ponderados estiver acima de um limiar. Mesmo organizada a partir de funções discretas e bem simplificadas, essa geração fez parte de diversas implementações de redes neurais artificiais, como *perceptrons* de múltiplas camadas e redes Hopfield(HOPFIELD, 1982).

As redes neurais de segunda geração são aquelas cujos valores de saída são representados sob escala contínua que utilizam modelos de neurônios com funções de ativação, como por exemplo, a função sigmoidal(SHUKLA; TRIPATHI, 2016). Tais redes são mais conhecidas como redes neurais artificiais (RNA) e foram adaptadas com sucesso a muitas aplicações de engenharia, que provaram ser eficazes na modelagem de alguns processos cognitivos. Sua representação esquemática é mostrada na Fig. 2.12-a onde as entradas x são ponderadas pelas sinapses w, e a soma desses valores é aplicada aos neurônios. Normalmente, um neurônio aplica uma função de transferência sigmoide que gera números com valores reais no intervalo (0,1). As regras de aprendizado para essas redes neurais determinam como ajustar os pesos para melhorar o desempenho da tarefa.

Fig. 2.12 - Comparação entre redes da 2° geração(esquerda) e 3° geração(direita). Extraído de (DENG et al., 2020).



A terceira geração é caracterizada por um grupo de redes que não possuem um ponto quiescente de operação, e são fundamentadas pela informação codificada por pulsos. A representação das redes neurais por pulsos (RNP) pode ser vista na Fig. 2.12-b que descreve principalmente o funcionamento. Os pulsos de entrada s são ponderados e integrados ao longo do tempo, comparando o resultado a um limiar. Se o limite for atingido, um pulso é emitido pelo neurônio de saída como resultado do processamento. Assim, uma vantagem significativa das redes neurais por pulsos é que tempo é inerentemente uma variável utilizada na computação. Por exemplo, latências de pulso, períodos refratários e frequência do trem de pulsos que geram uma capacidade intrínseca de processar dados que variam no tempo. Redes neurais pulsadas implementadas em hardware estabelecem naturalmente máquinas computacionais eficientes por ausência de consumo estático. A representação da informação por pulsos pode estabelecer ainda uma característica adicional de densidade, reduzindo a uma única linha necessária para a transferência da informação em vez de um barramento representando várias linhas.

Outro aspecto importante sobre a densidade está na implementação dos neurônios. Para uma classe de neurônios da segunda geração, esses são dependentes de funções de ativação, em que se baseiam basicamente de transformações não lineares sobre uma representação numérica do sinal de entrada. Para o campo de redes neurais por pulsos, os neurônios se apresentam mais simples compostos por dispositivos que mimetizam o comportamento neural biológico como foi primariamente estudado por Hodgkin e Huxley(HODGKIN; HUXLEY, 1952). Na Tab. 2.1, é apresentado um quadro comparativo sobre as propriedades das redes de 2° e 3° geração.

Tab. 2.1 - Comparativo entre propriedades das redes neurais de 2° e 3° geração.

	RNA (2° geração)	**RNP (3° geração)**
Rep. de dados	Numérica (ponto fixo ou flutuante)	Trem de Pulsos
Domínio das operações	Digital	Analógico
Funções de Ativação	ReLU, Sigmoidal	LIF, IF
Modelos	RNN, CNNs	Único

No campo de implementação de RNAs, as representações neuronais são modeladas, em sua grande maioria, por circuitaria digital estabelecendo a aplicação de conversores Analógico-Digital (ADC) e Digital-Analógico (DAC) (CAI et al., 2019; HEMSOTH, 2018), como componentes essenciais dos circuitos periféricos. Sistemas compostos por neurônios implementados sob a ótica de eletrônica digital são implementados em tecnologia CMOS, aumentando assim a área e uso de energia. Em contrapartida, no espaço de implementações de RNPs, tecnologias emergentes em mimetismo de neurônios e sinapses podem fornecer mais miniaturização e melhorias de eficiência energética, por não haver consumo quiescente, assim se aproximando da computação inspirada em sistemas biológicos (BOAHEN, 2017; MEROLLA et al., 2014).

2.6.REGRAS DE APRENDIZAGEM EM DISPOSITIVOS MEMRISTIVOS

Em 1949, o psicólogo Donald Hebb apresentou uma teoria sobre como a plasticidade funciona em sinapses (HEBB, 1949). Quando um neurônio repetidamente ou persistentemente participa do disparo do seu adjacente, algum processo de crescimento ou mudança metabólica ocorre em ambas as células (ou mais especificamente no elo entre elas), de modo que a eficácia de acionamento de um neurônio para o outro é aumentada. Isso indica que existe uma correlação entre o reforço sináptico e a atividade dos neurônios, o que levou o psicólogo a criar a célebre frase: *“Neurons that fire together, wire together”*. A conexão sináptica é flexível, e a experiência muda a eficácia de transmissão de sinal entre neurônios. A proposição de Hebb foi verificada alguns anos depois pelo neurocientista Eric Kandel(BRUNELLI; CASTELLUCCI; KANDEL, 1976).

Com este princípio, considera-se que a informação é armazenada nas sinapses. As leis de variação de plasticidade foram classificadas de acordo com quatro

categorias principais: habituação, sensibilização, potenciação e depressão. A habituação é uma resposta degradada ao longo do tempo aos mesmos estímulos, e tem um caráter de armazenamento volátil. A consciência, por outro lado, é uma resposta cada vez mais rápida a um determinado estímulo. A potenciação e a depressão são dois tipos de aprendizagem associativa complementar. Elas correspondem a como o elo entre os neurônios podem ser incrementados ou decrementados, dependentes dos estímulos aplicados e do aprendizado prévio. Nas seções anteriores foram introduzidas algumas condições operacionais para neurônios e sinapses artificiais, que derivam de excitação por tensão ou corrente. Nessa seção, serão introduzidos conceitos da implementação de regras de aprendizagem sobre dispositivos memristivos. As técnicas de aprendizagem empregadas impactam diretamente na performance e escalabilidade das redes neurais.

2.6.1 Aspectos sobe a Localidade do Treinamento

Dispositivos memristivos podem ser usados em sistemas de treinamento "*off-line*" (ou *ex-situ*). No aprendizado *off-line*, os pesos são obtidos usando um modelo computacional de treinamento e, em seguida, importados diretamente para o sistema matricial sináptico, para compor as operações de inferência(BOQUET et al., 2021). Separando a operação da rede em duas etapas não simultâneas, inferência e aprendizagem, as técnicas *off-line* necessitam de esquemas especiais ciclos de leitura e escrita ajustadas para vencer problemas clássicos de correntes parasitas em matrizes memristivas(ZIDAN et al., 2013). Circuitos adicionais são implementados junto à rede com intuito de vencer esses obstáculos e prejudicam a escalabilidade das redes neurais. Esse tipo de abordagem é usado em aplicações de modelos muito complexos e de difícil treinamento, como para grandes redes neurais convolucionais(GARBIN et al., 2015). No entanto, o uso mais atraente está em sistemas capazes de desenvolver seu aprendizado somente a partir de dados submetidos a rede, "*on-line*" (ou *in-situ*) sem depender de um sistema computacional adicional para estabelecer a atualização sináptica(ALIBART; ZAMANIDOOST; STRUKOV, 2013). Na abordagem de aprendizagem *online*, as arquiteturas neuromórficas utilizam o comportamento adaptativo das sinapses para realizar a aprendizagem. Em geral, essas arquiteturas têm elevado grau de imunidade à variabilidade intrínseca entre dispositivos (QUERLIOZ et al., 2013).

Existem duas variedades de aprendizado *on-chip*: supervisionado e não supervisionado. No aprendizado supervisionado *on-chip*, o sistema tem acesso a um sinal de professor e/ou função de custo. No contexto de atividades de classificação, por exemplo, o professor teria a responsabilidade de gerenciar o treinamento, definindo rótulos para as classes de dados e decidindo a sequência a ser apresentada a rede, como exemplo do sistema apresentado em (PREZIOSO et al., 2015). Em contraste, sistemas não-supervisionados aprendem sem conhecimento direto do que é 'certo' e 'errado'; em seu lugar, uma variedade de dinâmicas locais e efeitos de plasticidade empurram a rede para uma compreensão nativa dos dados que estão sendo alimentados nela (PEDRETTI et al., 2017).

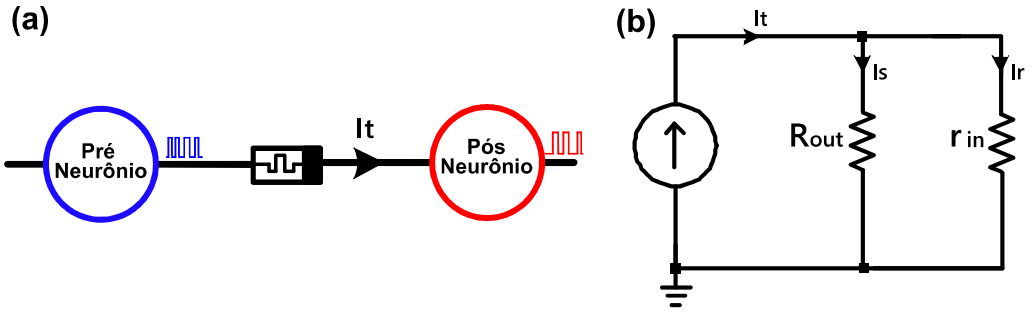
CAPÍTULO 3

EFICIÊNCIA NA TRANSFERÊNCIA DE INFORMAÇÃO

3.1 PROCESSAMENTO ANALÓGICO DE INFORMAÇÃO EM REDES NEURAIS

O processamento analógico em redes neurais pode ser interpretado como comunicação interneuronal cuja informação é codificada por sinais variantes no tempo de tensão ou corrente através de um canal sináptico, Fig. 3.1-a. A necessidade do casamento entre impedâncias dos elementos de circuito da rede é importante para tornar a transmissão de informações mais eficiente. O casamento de impedância é usualmente conhecido e refere-se à maximização da transferência de energia de uma fonte para uma carga. No entanto, para implementações de redes neurais, eficiência é mais importante do que energia transmitida. Para realizar uma análise do processamento de informação no domínio analógico, avalia-se a equivalência da comunicação entre dois neurônios como um circuito equivalente Norton, relacionando a diferença das impedâncias do circuito, como destaca na Fig. 3.1-b.

Fig. 3.1 - Transferência de informação interneuronal a) e um circuito equivalente Norton b).



A transmissão de sinais que variam no tempo através de elementos passivos enfrenta dificuldades de propagação devido a componentes parasitas que não estão expressos no circuito equivalente da Fig. 3.1-a. Partes reativas de impedância são omitidas dessa análise para fins de simplicidade, mas os efeitos parasíticos estão contemplados nos cálculos em termos de largura de banda BW, como será observado a seguir. Para fins de simplificação adicional, assume-se também que o tempo de comutação dos memristores é muito menor do que as constantes de tempo oriundas dos elementos reativos.

Para sinapses, os valores de impedância armazenam os pesos sinápticos no domínio analógico e realizam atividades computacionais, uma vez que convertem informações de neurônios (sinais de tensão constantes para RNA ou pulsos de tensão para RNP) em corrente como parte das operações MAC. Para os neurônios, as variações de impedância estão associadas ao estado deles, representadas pelo carregamento de capacitância de membrana e mimetismo da condutância dos canais iônicos, como no modelo de H-H(HODGKIN; HUXLEY, 1952), que operam a partir da corrente recebida das sinapses. A dinâmica das sinapses e dos neurônios dependem da polarização destes (estado quiescente) e dos estímulos recebidos(operação). Especificamente para a conexão entre esses dois elementos, um circuito de ajuste de impedância deve ser posicionado entre eles para permitir a maximização da eficiência na transferência de informações.

Neste trabalho, considera-se η como eficiência de transferência de informação interneuronal definida pela razão entre capacidade de transmissão e potência no canal sináptico, avaliada em $bits/J$. A abordagem de examinar η sobre custo computacional foi discutida anteriormente por Boahen em (BOAHEN, 2017). No entanto, a análise de eficiência sob um ponto de vista focado na impedância de entrada e saída entre elementos ainda não foi discutida e é aplicável a qualquer arquitetura neuromórfica, essencial para o projeto de circuitos.

3.2 ANÁLISE SOBRE A TRANSFERÊNCIA DE INFORMAÇÃO INTERNEURONAL

Analisando a capacidade de transmissão através de um canal sináptico pela equivalência de um circuito de Norton, diagrama da Fig. 3.1-b, considera-se o casamento de impedância entre sinapse e neurônios como uma fonte de corrente com impedância de saída R_{out} (representado a impedância de saída do neurônio pré-sináptico + resistência sináptica) conectada a uma carga com uma impedância r_{in} (representando a impedância de entrada de qualquer circuito receptor) e uma corrente sináptica I_t (sinal de corrente ponderada usado no processo de inferência).

C é a capacidade do canal definida por Shannon, (SHANNON, 1948), expressa pela expressão (3.1) em função da largura de banda BW.

$$C = BW \cdot log_2\left(1 + {}^{S}/_{N}\right) \tag{3.1}$$

Onde S é a potência do sinal, em V^2, representada por:

$$S = \left(\frac{R_{out}}{R_{out} + r_{in}} \cdot I_t\right)^2 \cdot {r_{in}}^2 \tag{3.2}$$

Enquanto N, é potência associada ao ruído térmico no elemento receptor, em V^2, definido por:

$$N = 4kTBW \cdot r_{in} \tag{3.3}$$

Logo a relação sinal-ruído é dada por:

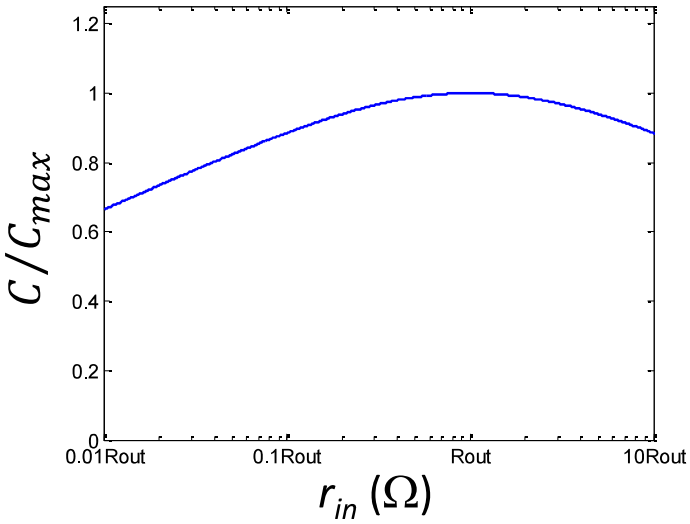
$${}^{S}/_{N} = \frac{{R_{out}}^2 \cdot r_{in}}{4kTBW} \cdot \frac{{I_t}^2}{(R_{out} + r_{in})^2} \tag{3.4}$$

Considerando o equivalente Norton da Fig. 3.1, a capacidade de transmissão sináptica é calculada por:

$$C = BW \cdot log_2\left(1 + \frac{{R_{out}}^2 \cdot r_{in}}{4kTBW} \cdot \frac{{I_t}^2}{(R_{out} + r_{in})^2}\right) \tag{3.5}$$

Então, deduz-se que a máxima capacidade de transmissão ocorre para $R_{out} = r_{in}$, como mostra a Fig. 3.2.

Fig. 3.2 - Capacidade normalizada do canal sináptico em função da impedância de entrada r_{in}, expressa em termos de R_{out}.



À medida em que o critério de casamento de impedâncias é de fundamental importância para a transmissão do sinal, deve-se considerar a eficiência máxima para a transferência do sinal, expressa em termos de $bits/J$. Em implementações biomiméticas neuromórficas, esta é uma condição procurada e, como tal, dividimos C pela dissipação de energia P, como calculada em (3.6).

$$P = \frac{R_{out} \cdot r_{in}}{R_{out} + r_{in}} \cdot {I_t}^2 \tag{3.6}$$

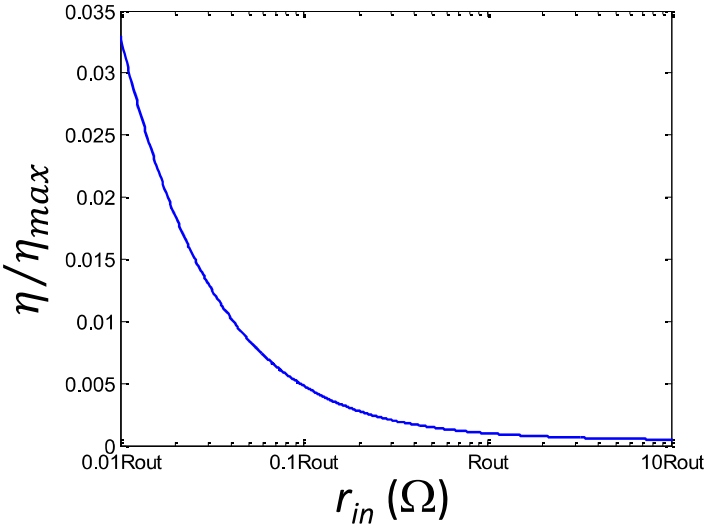
Nomeando η como eficiência na transmissão pelo canal sináptico, obtém-se a seguinte expressão:

$$\eta = \frac{C}{P} \tag{3.7}$$

Pela inspeção da Fig. 3.3, notamos que no caso de uma representação equivalente Norton, a máxima eficiência é obtida para quando $r_{in} \to 0$ (para a representação equivalente Thevenin, $r_{in} \to \infty$). A impedância r_{in} precisa ser

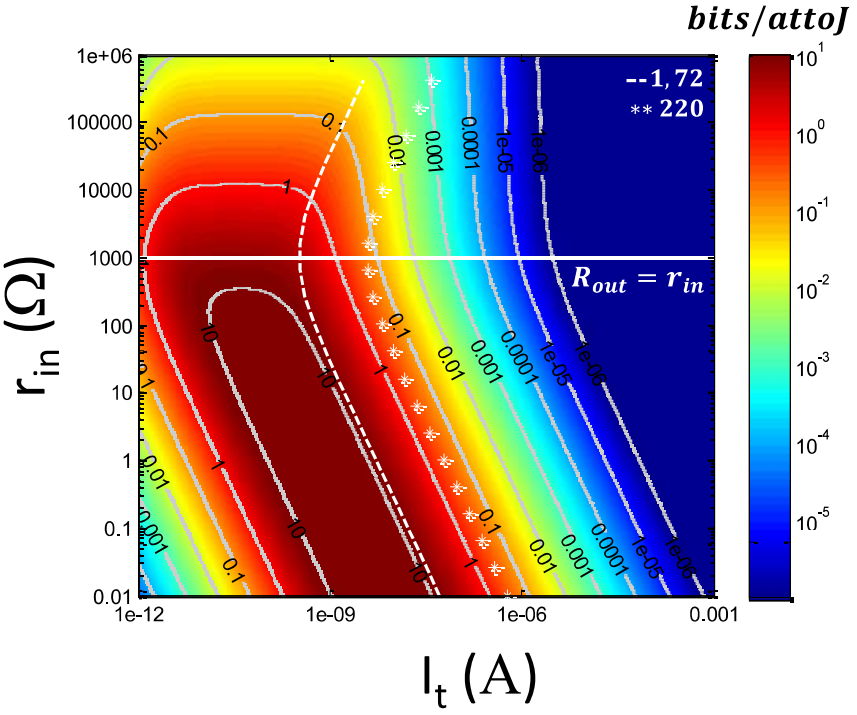
considerada no projeto de redes neurais, portanto, passível de maximizar a eficiência do canal.

Fig. 3.3 - Eficiência do canal sináptico em função da impedância de entrada r_{in}.



Condições de operação eficientes podem ser melhor visualizadas por mapas de cores, como mostrado na Fig. 3.4. A barra de cores representa em escala logarítmica a eficiência em função da corrente sináptica I_t e impedância de entrada r_{in}. Na Fig. 3.4, a eficiência de transmissão é apresentada para $BW = 1kHz$, e a resistência $R_{out} = 1k\Omega$.

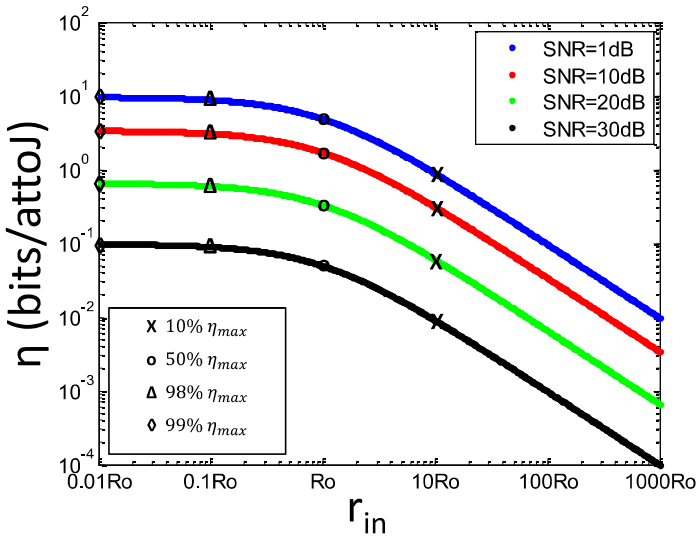
Fig. 3.4 – Eficiência na transmissão de informação no domínio analógico.



O mapa na Fig. 3.4 orienta projetistas de circuitos neuromórficos a selecionar melhores parâmetros operacionais da rede, como amplitude do sinal (associada a SNR e corrente I_t), formas de onda, ciclo de trabalho e frequência do sinal (incorporados ao termo BW), a fim de maximizar a eficiência na transferência de sinal entre neurônios. As linhas de contorno são espaçadas por uma década de eficiência expressa em $bits/attoJ$. A eficiência mostrada na Fig. 3.4, quando comparado com a literatura publicada (BORKAR, 2013), são ordens de magnitude maiores, principalmente devido à frequência de operação acima de $1GHz$ (e ao uso da fórmula $P = CV^2f$) para os sistemas convencionais, enquanto neste trabalho optou-se por uma largura de banda de $1kHz$. A região abaixo da linha horizontal $R_{out} = r_{in}$, na Fig. 3.4, define uma região onde as melhorias na eficiência são predominantemente relacionadas ao SNR, e pouco relevantes ao melhoramento do casamento de impedância. Para o mesmo conjunto de parâmetros R_{out}, BW, plota-se um gráfico acompanhando condições de SNR constante como apresentado na

Fig. 3.5.

Fig. 3.5 - Eficiência para diferentes condições de casamento de impedância. $R_{out} = 1k\Omega$, $BW = 1kHz$.



A mensagem chave a ser aprendida com os gráficos anteriores é que se adotarmos casamento de impedância para máxima transferência de energia com $R_{out} = r_{in}$, obteremos 50% do máximo teoricamente possível para a eficiência em $bits/J$. O aumento da eficiência com a diminuição de r_{in}, ou aumento da razão R_{out}/r_{in}, desacelera a partir de $R_{out} = r_{in}$, atingindo 98% em $R_{out} = 10r_{in}$. As condições de $\eta/\eta_{max} > 98\%$ podem ser extraídas através da escolha dos parâmetros de rede adequados, e η maiores ocorrem a partir de $R_{out} > 10r_{in}$, para uma dada corrente sináptica e BW. A partir disso, depois de atingido esta relação R_{out}/r_{in} de casamento de impedância, a adição de um fator de melhoramento no casamento de impedância resultará em aumentos menores na eficiência.

Através do mapa da Fig. 3.4 e do gráfico na

Fig. 3.5, é possível observar que, após estabelecido uma razão entre impedâncias de entrada de saída apropriada, as regiões operacionais mais eficientes para a transferência de informação interneuronal se orientam pela redução da potência do sinal (redução de I_t e consequentemente SNR). A ação de reduzir a potência do sinal transmitido é bastante promissor e vai ao encontro da implementação de sistemas neurais em larga escala. As duas linhas destacadas, representam a relação sinal-ruído para as condições de $1{,}72$ e 220, correspondentes ao cérebro humano e um processador convencional respectivamente, como descrito em (TSUR, 2021). Percebe-se que para os sistemas biológicos a redução da potência do sinal permite que o cérebro opere sobre condições de maior eficiência em

contrapartida a cálculos mais propensos a erros(KOCH, 2004). As probabilidades de erro são de $0{,}65$ para o cérebro enquanto 10^{-24} para os computadores.

O que se deseja é que redes neurais no domínio analógico possam operar com sinais de intensidades cada vez menores. Porém, o estado da arte da tecnologia de sinapses e neurônios artificiais limitam o funcionamento para baixos valores de SNR. Sistemas com baixa energia (que envolvam diminutos valores de intensidade de sinal) podem ser limitados pela atuação do neurônio, que precisam de energia suficiente para fazer ativar seu dispositivo de chaveamento. Operar sob pequenas SNR significa operar com valores cada vez menores de tensão e corrente. Isso pode ser um limitante, quando comparados a tensão de chaveamento dos memristores. Por exemplo, sob um regime de SNR de 10 com eficiência de $10\ bits/attoJ$ ($R_{out} = 1k\Omega$, $r_{in} = 10\Omega$, $BW = 1kHz$), a amplitude dos pulsos ficaria em torno de $100uV$, o que corresponde a pelo menos 4 ordens de grandeza menores do que as tensões de chaveamento para a tecnologia atual de memristores, que estão na faixa de ±1 a ±2 V, como apresentado em (BEILLIARD et al., 2020; PICKETT et al., 2009). Uma grande diferença nesses valores pode acarretar dificuldades de implementação de regras de aprendizagem locais, que utilizam os pulsos de excitação para gerar condições de plasticidade sináptica(SERRANO-GOTARREDONA et al., 2013).

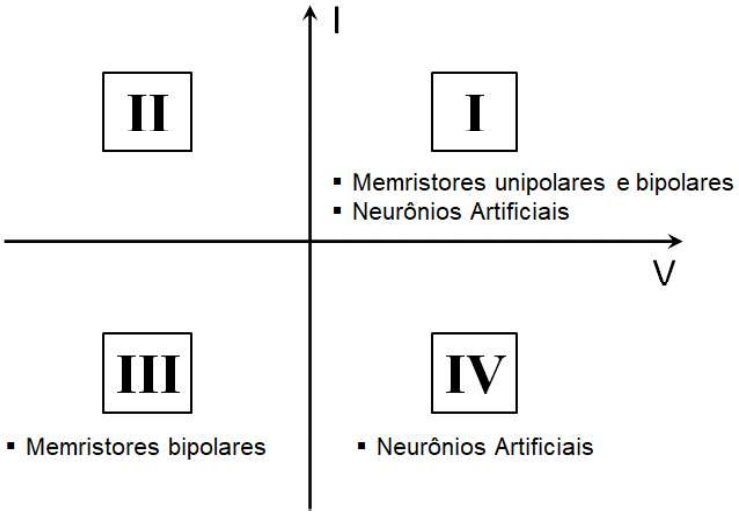
Diferenças de impedância entre os elementos também é algo existente em redes neurais de sistemas biológicos, e o mecanismo de casamento de impedância também está presente na forma de comunicação interneuronal química provida pela ação dos neurotransmissores (JOHNSTON; WU, 1994). No caso em que células pós-sinápticas são muito maiores do que a células pré-sinápticas, a transmissão de um potencial elétrico de ação através da fenda sináptica falhará. A célula pré-sináptica menor não produzirá corrente suficiente para despolarizar a célula pós-sináptica maior além do limiar. Como exemplo das células musculares que são muito maiores do que as células nervosas que as controlam (STEIN et al., 1999). Para vencer tal impedimentos inerentes as características físicas dos neurônios, uma substância mensageira química liberada da célula pré-sináptica se difunde através da fenda sináptica e se liga a receptores na célula pós-sináptica. De maneira análoga, a utilização de um circuito que intercomunique sinapse e neurônio, atendendo os requisitos para a operação eficiente, precisa compor mecanismos de casamento para que permita alimentar o circuito neuronal com a corrente sináptica independente de suas características de operação. Ainda, esse circuito precisa estar apto a prover a

implementação de regras de treinamento de acordo com o funcionamento da sinapse. Uma topologia de circuito com esses recursos será explorada no capítulo seguinte.

3.3 CIRCUITOS DE INTERFACE PARA ACOPLAMENTO SINAPSE-NEURÔNIO

A análise da transferência de informação a partir do mapa da Fig. 3.4, foi realizada de forma a criar um entendimento da relação entre eficiência energética e a razão entre impedâncias sinapse-neurônio. Porém, aquele gráfico considera que apenas um elemento do circuito equivalente possui impedância variável, r_{in}. Isso não se aplica para arquiteturas de redes neurais com memristores e neurônios artificiais como mostrado aqui. Os valores de resistência dos pesos sinápticos são dependentes do treinamento. Além disso, os dispositivos memristivos, podem exigir condições de polarização (em corrente ou tensão) para atualização da sua resistência restritos aos $1°$ e $3°$ quadrantes, ver Fig. 3.6.

Fig. 3.6 - Quadrantes de operação destacando a operação de memristores e memristores e neurônios

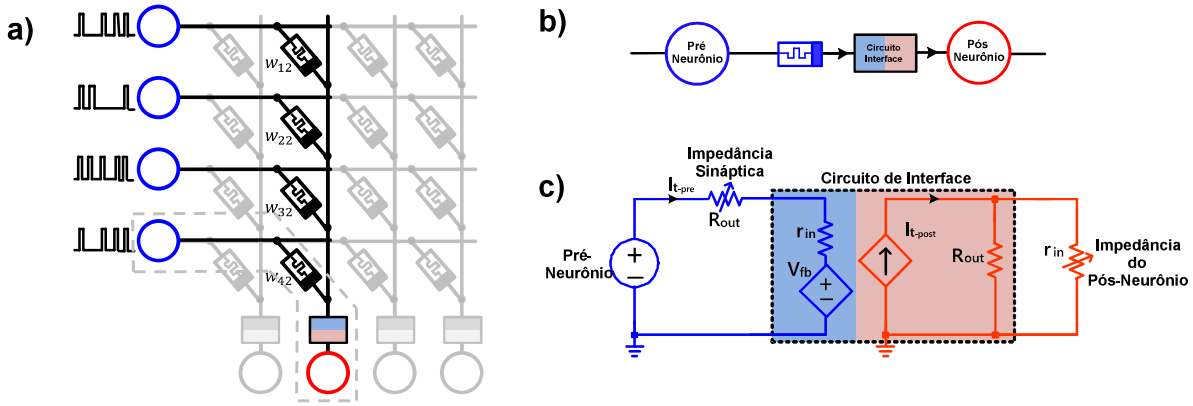


Tal como as sinapses, os neurônios possuem especificidades para o acionamento, associadas ao carregamento da capacitância de membrana e o funcionamento do dispositivo de disparo e descarga. Grande parte das topologias de circuitos neuronais são acionados por corrente, operando nos $1°$ ou $4°$ quadrantes, ver Fig. 3.6. A fim de garantir o funcionamento de sinapses e neurônios, um circuito de interface é necessário, que cumpra os seguintes requisitos:

- Estabelecer casamento de impedância para garantir alta eficiência da informação que flui ao longo da rede;
- Manter neurônio e sinapses operando sob suas respectivas condições de polarização;
- Prover condições suficiente para implementar as regras de aprendizagens locais a partir das condições de operação da sinapse.

Dessa forma, a arquitetura da rede neural pode ser expressa como mostrado na Fig. 3.7-a, com uma representação funcional da comunicação entre dois neurônios como expresso em Fig. 3.7-b. Os elementos de circuitos: neurônios, sinapse e circuito de interface podem ser representados por circuitos equivalentes, como apresentado na Fig. 3.7-c.

Fig. 3.7 - a). Diagrama simplificado de uma RNP. b) Conexão entre dois neurônios. c) circuitos equivalentes pré e pós-sinápticos correspondem à equivalência de Thevenin e Norton.



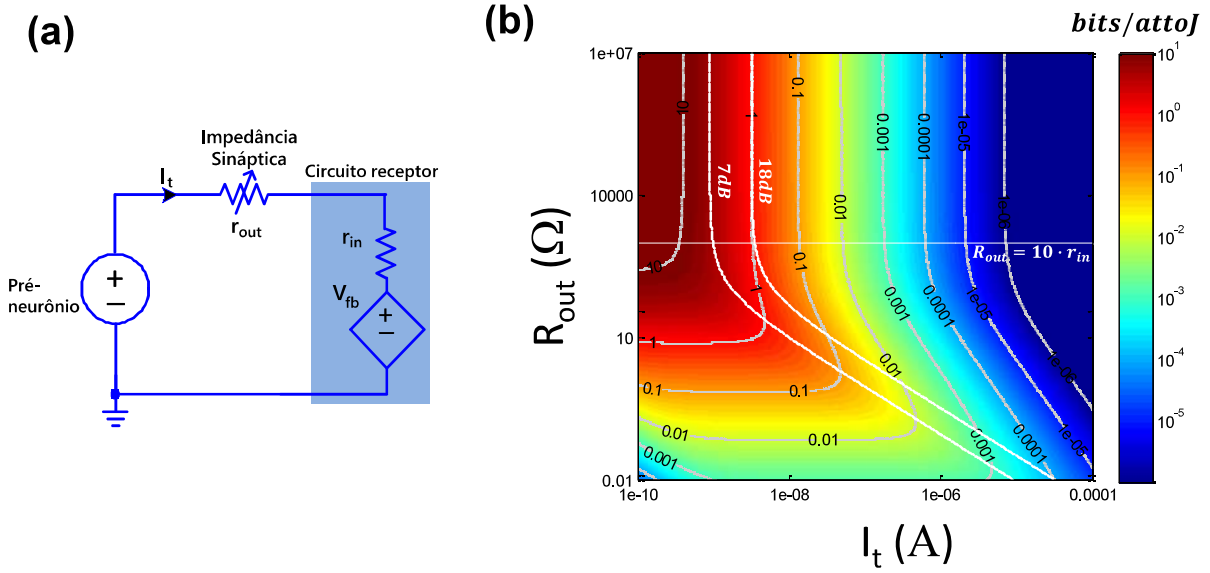
Primeiramente, o circuito de interface opera de forma a desacoplar sinapse e pós-neurônio a fim de gerar dois circuitos: pré e pós-sinápticos. O circuito que intermedia sinapse e neurônio garante que ambas as partes funcionem dentro de suas condições de operação e que variações de impedância de sinapses e neurônios não interfiram no funcionamento de cada um. Como na comunicação entre neurônios a informação é codificada como sinais de corrente, é preciso garantir que a corrente sináptica I_t seja transportada do neurônio pré-sináptico ao neurônio pós-sináptico sem atenuação. Um outro requisito para o circuito de interface é estabelecer o casamento de impedância a fim de prover eficiência na transferência de informação considerando as particularidades do funcionamento na mimetização de sinapses e neurônios. Dessa forma, a representação da comunicação entre dois neurônios, como mostrado na Fig. 3.7-b, pode ser representado por um circuito equivalente, descrito Fig. 3.7-c. A fim de

prover treinamento local, o circuito receptor precisa compor características de fonte de tensão a fim de produzir sinais de tensão para potenciação e depressão nos dispositivos sinápticos.

3.3.1 Circuito de casamento pré-sináptico

O circuito pré-sináptico é definido como mostrado na Fig. 3.8-a. Um circuito equivalente Thevenin é representado, onde o receptor apresenta impedância fixa r_{in}, e observa-se variação de impedância para a sinapse R_{out}. Uma fonte de tensão representa o neurônio pré-sináptico, a resistência variável representa a sinapse memristiva e o bloco destacado em azul representa o circuito receptor, o qual é composto por uma fonte de tensão controlada (apenas para representar a capacidade de aplicar sinais que possam modular a resistência da sinapse) e sua impedância de entrada, ver Fig. 3.8-a. O mapa de eficiência pré-sináptico para $r_{in} = 100\Omega$, $BW = 1kHz$, e uma variação de resistência de R_{out}, que incorpore a faixa de operação dos dispositivos usados neste trabalho (de $1k\Omega$ a $10k\Omega$, mais detalhes no cap. 5 e apêndice C), é apresentado na Fig. 3.8-b.

Fig. 3.8 - a) Circuito equivalente do circuito pré-sináptico. b) Mapa de eficiência para a interligação entre pré-neurônio, memristor e circuito receptor.



Para valores de R_{out} uma década maiores que r_{in}, linha horizontal na Fig. 3.8-b, a eficiência pouco depende dos valores da resistência sináptica, sendo a intensidade da corrente o fator determinante para definir condições mais eficientes de operação. Os níveis de corrente definem a potência do sinal usado nas operações

MAC do circuito pré-sináptico. O número de níveis distinguíveis de corrente I_t tem como limitante a relação sinal-ruído que define também a probabilidade de erro do sistema, como apresentado em (TSUR, 2021). Com a potência do sinal se aproximando dos níveis de ruído, a comunicação entre neurônios se torna mais propícia a erros, quando pulsos de níveis diferentes e individuais não puderem ser distinguidos. A relação entre o número de níveis distintos M e SNR é dada por:

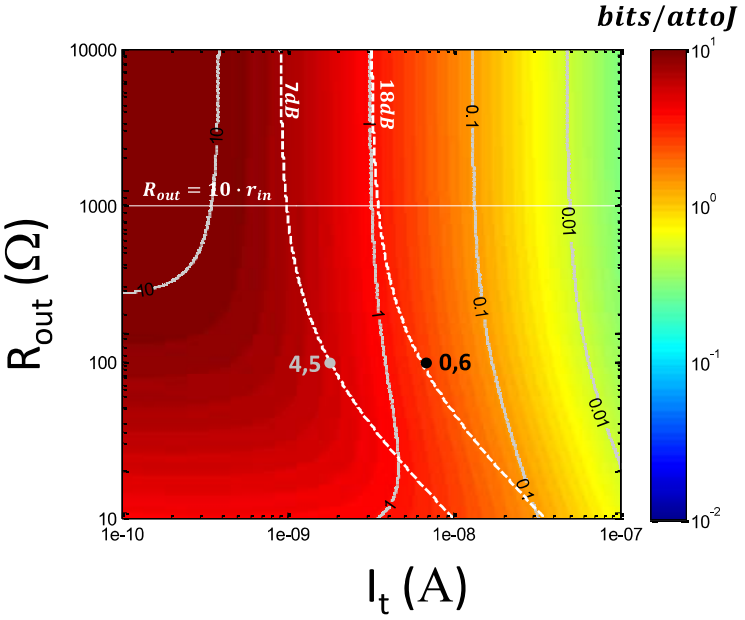
$$M = \sqrt{1 + SNR} \tag{3.8}$$

A probabilidade de erro a partir do SNR é definida pela expressão:

$$P_{erro} = \mathrm{e}^{-0{,}25 \cdot \mathrm{SNR}} \tag{3.9}$$

SNR e resolução sináptica são parâmetros considerados no projeto de redes neurais e projetistas definem esses critérios, de acordo com a necessidade para cada aplicação. O ensinamento sobre a transferência de informação entre neurônios, a partir da expressão (3.1), é que o número de bits de informação codificados em sinal diminui logaritmicamente com SNR, porém essa redução colabora para o aumento da eficiência η, de acordo com (3.7). Por exemplo, se o SNR reduzir de $18dB(\sim 63)$ para $7dB(\sim 5)$, a resolução de bits codificados em informação cairá de 3 para 2, ao passo de um aumento na eficiência energética de $0{,}6\ a\ 4{,}5$ ***bits/attoJ***, como mostra a Fig. 3.9, destacando central da Fig. 3.8-b. As probabilidades de erro são de para maior resolução $1{,}45 \cdot 10^{-7}$ e $0{,}28$ para a menor.

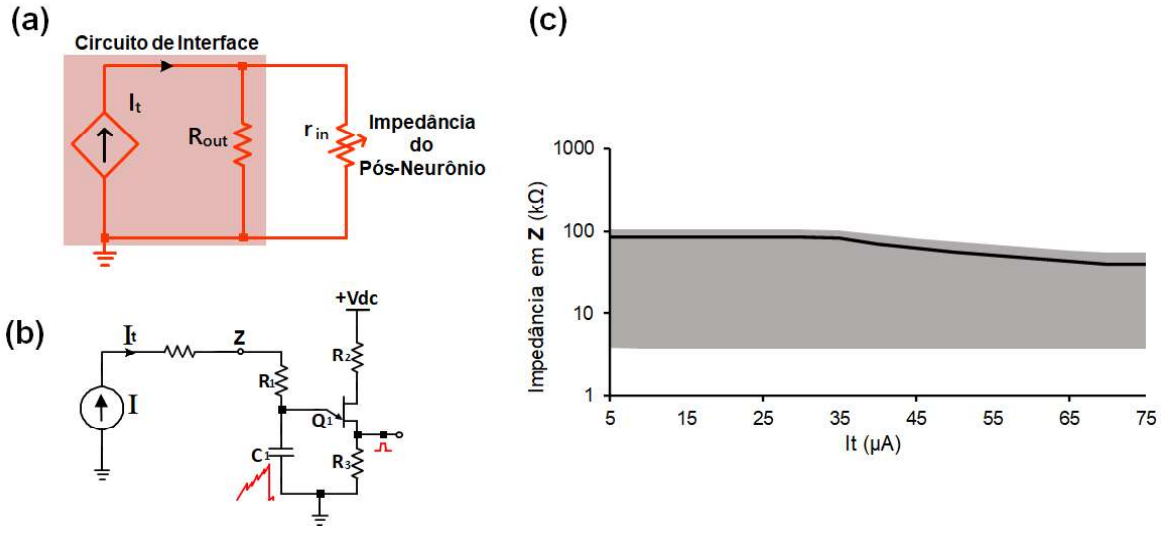
Fig. 3.9 - Mapa de eficiência para a interligação entre pré-neurônio, memristor e circuito receptor.



3.3.2 Circuito de casamento pós-sináptico

Analisando o circuito pós-sináptico (circuito de interface e pós-neurônio), a principal função é alimentar o neurônio pós-sináptico com a corrente "transportada" da parte pré-sináptica. A representação equivalente Norton é utilizada, como mostrada na Fig. 3.10-a. Através dessa analogia, o circuito de interface pós-sináptico possui um circuito independente e desacoplado da parte pré-sináptica. Ele é composto por uma fonte de corrente de valor I_t, uma impedância de saída fixa R_{out}, enquanto observa-se variação de impedância para o neurônio r_{in}.

Fig. 3.10 - a) Circuito equivalente pós-sináptico. b) Circuito de excitação do neurônio baseado em TUJ, e c) Perfil de impedância por corrente de excitação.



A impedância r_{in} está associado a impedância "vista" pelo circuito do bloco em vermelho, dependente da topologia de circuito do neurônio pós-sináptico. Topologias de neurônios para redes neurais por pulsos usualmente possuem como bloco de entrada capacitores que representam a mimetização do acúmulo de carga da membrana neuronal em sistemas biológicos. Conectado aos capacitores, circuitos responsáveis pelo disparo baseado em um limiar de tensão operam na descarga do potencial de membrana. Um exemplo do circuito pós-sináptico é apresentado na Fig. 3.10-b, onde um circuito de neurônio baseado em transistor de unijunção é acionado por uma fonte de corrente $\mathbf{I}$ que alimenta o neurônio com valores da corrente sináptica. Isso dá um caráter dinâmico à impedância de entrada do neurônio, proporcionando dificuldade de entregar a corrente ao neurônio em dependência do valor de tensão carregado no capacitor. Considerando os valores de corrente entregue pela fonte I_t e as variações de tensão no terminal $\mathbf{Z}$, o regime de impedância visto pelo circuito de interface pode ser plotado como na Fig. 3.10-c, onde a linha sólida representa os valores médio de impedância, e a área em cinza representa a variação de impedância para o mesmo valores de corrente I_t. Dessa forma, é desejável a utilização de circuitos que se comportem como fontes de corrente com alta impedância de saída a fim de seja indistinta a possibilidade de carregar a capacitância de membrana baseada nos valores de potencial de ação do neurônio. Também, os dispositivos comutadores são elementos que possuem resistência diferencial negativa, aptidão a serem controlados em corrente, e, portanto, favorável a conexão como mostrado na Fig. 3.10-a.

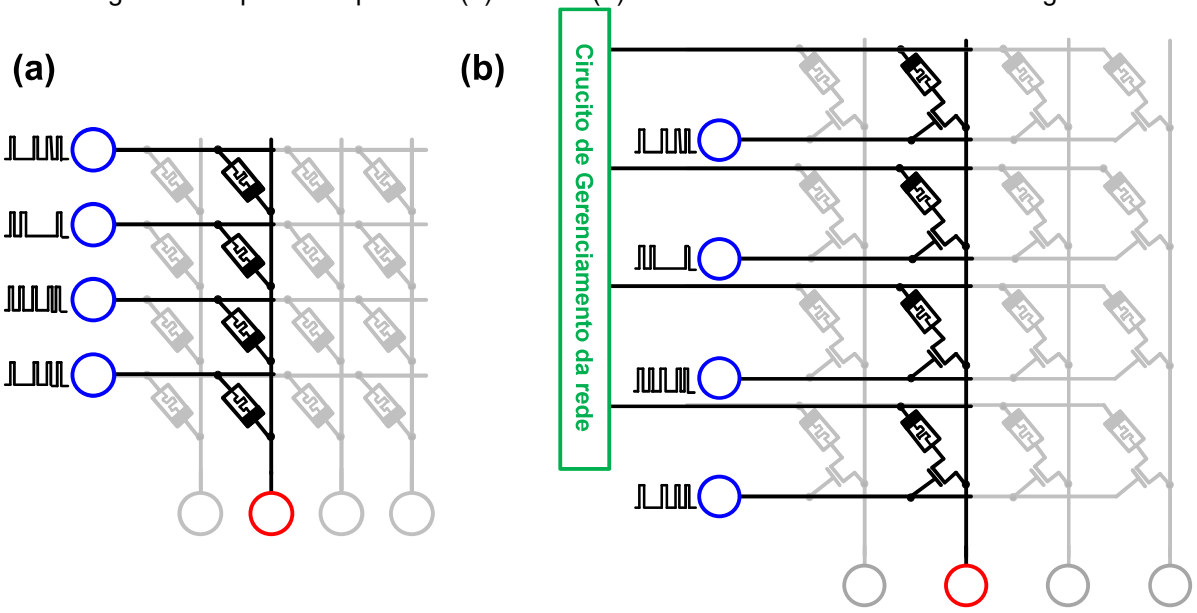
CAPÍTULO 4

CIRCUITOS DE INTERFACE PARA REDES NEURAIS ANALÓGICAS

4.1 ARQUITETURA DE REDES NEURAIS PULSADAS

O emprego de dispositivos de memórias resistivas organizadas matricialmente com eletrodos em barras cruzadas implementa naturalmente algoritmos de RNs pelas operações MACs, dado que as operações de multiplicação e acumulação do produto escalar podem ser realizadas no domínio analógico utilizando a lei de Ohm e a lei das correntes de Kirchhoff. Nesse contexto, os memristores são os elos entre as camadas de rede compostas por circuitos de periferias. Uma única coluna (ou linha), como destacado na Fig. 4.1, dessa configuração representa um processador independente. Todos os processadores operam de forma paralela nas operações computacionais e os circuitos periféricos provêm controle da operação da rede e eficiência no tráfego do sinal.

Fig. 4.1 – Arquiteturas passivas(a) e ativas(b) de redes neurais no domínio analógico.



Abordagem para MACs proposta na Fig. 4.1-a se baseia em células de memória totalmente passiva em configuração 1R, ou muitas vezes usando seletores em 1S1R (onde S descreve o seletor que pode ser uma camada depositada ou um diodo). Algumas arquiteturas analógicas de redes neurais propuseram o uso de sinapses ativas, construídas por mais de um dispositivo entre as barras cruzadas, Fig. 4.1-b. Transistores e memristores são usados como células de memória em configuração 1T1R(ZANGENEH; JOSHI, 2012), 2T1R(WANG et al., 2015), e outras. Todos elas usam transistores MOS (terminologia do inglês *metal-oxide semiconductor*) para garantir a o controle das correntes parasitas, e prover recursos de plasticidade. Essas implementações são escalar com N^2, com N o número de linhas ou colunas em uma matriz quadrada. Além disso, sinapses ativas exigem sistemas de controle para os terminais de porta que aumentam a complexidade, área de chip e o uso de energia com circuitos periféricos. Uma alternativa são as matrizes de barras cruzadas sinápticas totalmente passivas que fornecem implementações extremamente escalar pelo uso de apenas um único dispositivo memristor como célula de memória (KIM et al., 2021). Concentrando-se em circuitos periféricos que são escalados com N, arquiteturas baseadas em matrizes passivas por meio de circuitos de interface permitem incorporar primitivas de inferência e treinamento sem exigir recursos adicionais de gerenciamento da rede. Isso expande a escalabilidade por meio da simplificação da arquitetura.

Uma desvantagem de sistemas baseado matrizes passivas é a necessidade de alta capacidade de fornecimento de corrente para atualização de peso sináptico.

Durante a aplicação de tensão para atualização dos pesos sinápticos, o circuito de periferia fornece corrente para uma linha inteira (ou coluna), exigindo valores de corrente elevados para chavear todos os dispositivos conectados a ela. O requisito de alta corrente para circuitos periféricos é um problema inerente às implementações de barra cruzada passiva, o que limita o uso de grandes matrizes passivas. Os requisitos de corrente podem ser amenizados através de regra de treinamento otimizadas (DENG et al., 2016), ou sob uma abordagem fragmentada onde pequenos núcleos matriciais formam uma matriz maior(JAMES; CHUA, 2021).

4.2 ANÁLISE SOBRE OS REQUISITOS DOS CIRCUITOS DE INTERFACE

De acordo com as necessidades de escalabilidade, eficiência e treinamento local destacadas nas seções anteriores, é possível elencar requisitos funcionais aos circuitos que atuarão na interface entre sinapse e neurônio:

1) Para aumentar a eficiência na transferência de informação, é importante que o circuito acoplado à sinapse possua baixa impedância de entrada.

2) Para fornecer condições de modulação da resistência do memristor, no ponto de conexão com as sinapses, o circuito de interface precisa operar como uma fonte de tensão que pode ser controlada por tensão (FTCT) ou corrente (FTCC). Essa funcionalidade é necessária pelo fato de que um sinal de tensão é utilizado para atualização sináptica, cujo características são definidas pela regra de treinamento.

3) O circuito de interface precisa conduzir a corrente da matriz de sinapses ao neurônio de maneira unidirecional, e que opere independentemente das variações impedância desses dois componentes. Circuitos que operem como fontes de corrente controlada por corrente (FCCC) podem suprir essa função.

Em redes neurais de matrizes passivas, os requisitos funcionais se ampliam quando se leva em consideração o dispositivo usado como sinapse. Memristores são treinados pela modulação de sinais em amplitude, polaridade ou duração do sinal em seus terminais (SERRANO-GOTARREDONA et al., 2013). Tal ajuste exige condições para promover a comutação resistiva que pode ser restrita ao $1°$ quadrante (dispositivos unipolares como PCM(SURI et al., 2011) e memristores(HOWARD; BULL; DE LACY COSTELLO, 2015)), ou $1°$ e $3°$ quadrantes (para dispositivos bipolares, que compreendem a maioria dos memristores (WANG et al., 2018)). A

funcionalidade fundamental é fornecer modulação de tensão no dispositivo de memória durante os momentos de atualização sináptica.

Na literatura encontra-se muitas propostas de circuitos de interface que operam em modo corrente, a partir de espelhos de corrente no ponto de contato com a matriz sináptica, como em (ISHII et al., 2019). Circuitos que operam em modo corrente, que usam a corrente como informação ou variável de controle, muitas vezes podem superar significativamente a frequência de operação de circuitos tradicionais que operam em modo de tensão(TOUMAZOU; LIDGEY; HAIGH, 1990), de modo que pequenas oscilações de tensão podem ser mantidas permitindo grandes oscilações de corrente. Espelhos de corrente se apresentam como estruturas simples e de ampla faixa de frequência de operação, se encaixando às condições de implementação em circuitos integrados. Porém, necessitam de recursos adicionais para regulação de tensão e aplicação de pulsos de feedback.

Uma outra abordagem para circuitos de interface sinapse-neurônio é fundamentada em estruturas baseada em amplificadores com realimentação negativa, que incluem uma estrutura clássica de amplificador de transimpedância (TIA terminologia do inglês *transimpedance amplifier*), como nas implementações em (MUÑOZ-MARTIN et al., 2020; PEDRETTI et al., 2017; WU et al., 2015). Esta é uma topologia de circuito bem aceita e geralmente é uma escolha padrão devido à sua baixa impedância de entrada Z_{in} (sendo ela dependente do ganho), conjuntamente com a possibilidade de modulação de tensão nos pontos de contato com as sinapses. Porém, uma saída com características de fonte de tensão dificulta a integração com circuitos neuronais, pois estes são essencialmente alimentados em corrente, e exigem a implementação de circuitos complementares. As soluções por amplificadores ou espelhos de corrente são incompletas em termos de modulação de tensão sobre a sinapse e transporte de corrente aos neurônios, e por isso, circuitos complementares são necessários para garantir as fases de inferência e treinamento. Essas abordagens reduzem os benefícios da escalabilidade oportunizado pelas matrizes passivas, exigindo sistemas adicionais para controlar a operação da rede.

Uma topologia de circuito que engloba em único bloco os todos requisitos funcionais é o transportador de corrente de segunda geração (TCSG) (SEDRA; SMITH, 1970). Esse circuito opera com características de FCCC na saída e FTCT na entrada se adaptando aos requisitos de acionamento com neurônios artificiais e necessidades de treinamento local. Ademais, o TCSG é projetado por circuitos que

operam em modo corrente, e assim suprime a dependência ganho-banda passante constante presente em amplificadores (TOUMAZOU; LIDGEY; HAIGH, 1990). O TCSG pode ser implementado com possibilidade de alteração no sentido da corrente exportada aos neurônios, compondo assim um mecanismo excitatório e inibitório presente em redes neurais(LECERF; TOMAS; SA\"\IGHI, 2013).

Além de requisitos funcionais, também precisarão ser considerados requisitos de viabilidade de implementação. O reduzido número de componentes (área) e compatibilidade com a tecnologia de memória proposta são especialmente importantes. Baixo consumo de potência na operação, principalmente para circuitos que possua um consumo estático reduzido, é um requisito almejado pelos projetistas. As redes neurais pulsadas se mostram uma alternativa eficiente por se aproximarem mais ainda a sistemas neuromórficos biológicos e o consumo acontecer somente duração do pulso. Circuitos que possuem elevado consumo estático de potência podem comprometer o ganho de eficiência provido pela implementação de redes neurais analógicas.

A Tab. 4.1, apresenta um comparativo de atributos para as estruturas de circuitos elencadas como circuitos de interface.

Tab. 4.1 - Tabela de atributos para circuitos de interface.

	Espelho de corrente	**TIA**	**TCSG**
Compatível a matrizes passivas	**Sim**	**Sim**	**Sim**
Ganho-BW independente	--	**Não**	**Sim**
Casamento Imp. – Sinapse	**Sim**	**Sim**	**Sim**
Saída como FCCC	**Sim**	**Não**	**Sim**
Excitatório / Inibitório	**Não**	**Não**	**Sim**
Pulsos retroativos	**Não**	**Sim**	**Sim**
Compatível com sinapses bipolares	**Não**	**Sim**	**Sim**

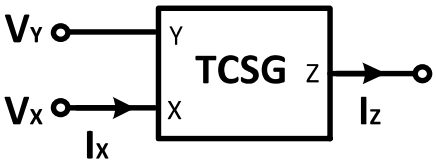
4.3 TRANSPORTADOR DE CORRENTE DE 2° GERAÇÃO COMO BLOCO DE INTERFACE VERSÁTIL

O TCSG é um circuito de três terminais, designados por X, Y e Z, onde a tensão no terminal X é igual a tensão aplicada no terminal Y. Qualquer corrente que flua para o terminal X é transportada para o terminal Z que representa uma fonte de corrente com uma alta impedância de saída (SMITH; SEDRA, 1968). Os TCSG

dispõem de polaridade. No subtipo TCSG+ (na literatura conhecido como CCII+, da terminologia em inglês *current conveyor second generation*), a corrente em X produz corrente com sentido contrário no terminal Z; em um TCSG- (ou CCII-), a corrente em X resulta em uma corrente de mesmo sentido no terminal Z (SEDRA; SMITH, 1970). O funcionamento do transportador de corrente e a simbologia são apresentados na Fig. 4.2. A constante P define a polaridade do transportador.

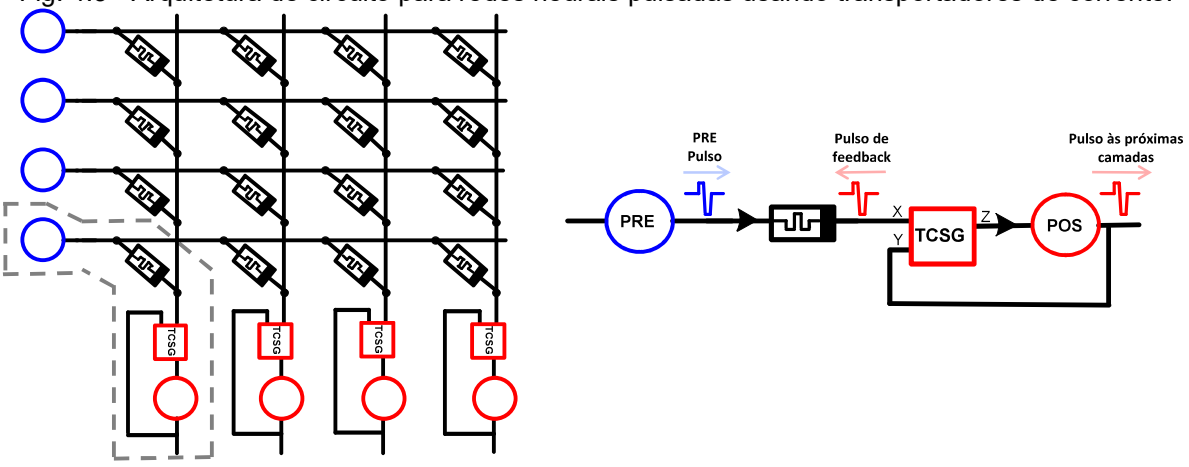
Fig. 4.2 – Representação simbólica e modelagem matemática para o transportador de corrente segunda geração.

$$\begin{bmatrix} i_Y \\ v_X \\ i_Z \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \pm 1 & 0 \end{bmatrix} \begin{bmatrix} v_Y \\ i_X \\ v_Z \end{bmatrix}$$



Os transportadores de corrente diferem-se por gerações, onde um ramo de baixa impedância que liga os terminais X e Y é característica da primeira geração. Para a segunda geração (SEDRA; SMITH, 1970), o terminal Y tem alta impedância e o terminal X tem baixa impedância. A terceira geração é semelhante a primeira, exceto que a corrente no terminal X é invertida com relação a corrente no terminal Y (FERRI; GUERRINI, 2003). Para aplicação em redes neurais a segunda geração é de interesse, pelo fato da entrada Y em alta impedância proporcionar facilidade para conexão com circuitos geradores de pulsos para treinamento. Na Fig. 4.3, é apresentado uma configuração esquemática da rede neural destacando as interconexões entre os blocos formadores.

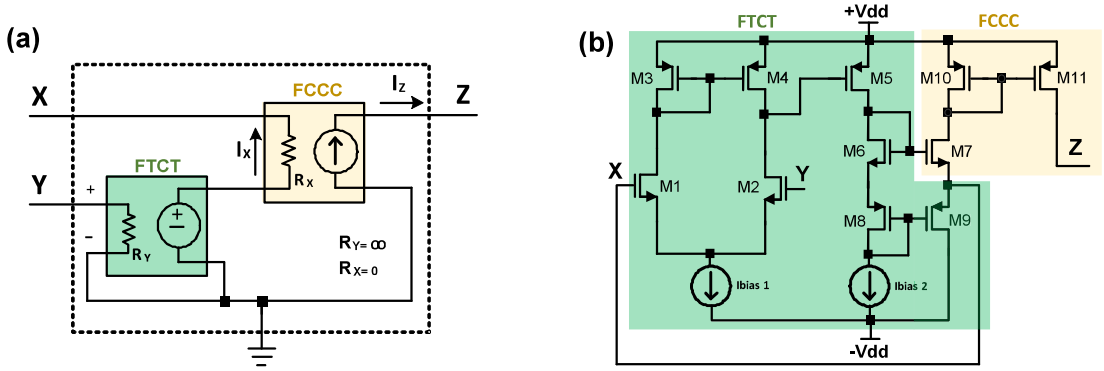
Fig. 4.3 - Arquitetura de circuito para redes neurais pulsadas usando transportadores de corrente.



O circuito implementado com transistores discretos desse trabalho é baseado no circuito proposto em (KURASHINA; OGAWA; WATANABE, 1998). Ele é composto

por uma topologia TCSG- Classe AB com estágio de saída modificado, conforme mostrado na Fig. 4.4. TCSG da Fig. 4.4 é composto por quatro blocos; um estágio diferencial ($M1 - M2\ e\ M3 - M4$) para sensoriamento das tensões nos terminais X e Y, um estágio *push-pull* ($M7\ e\ M9$) para fornecer corrente no terminal X, um espelho de corrente ($M10\ e\ M11$) para a fornecer a corrente de saída no terminal Z, e componentes complementares para polarização representado pelas fontes de corrente I_{bias1}, I_{bias1} e os transistores $M5, M6\ e\ M8$.

Fig. 4.4 – Detalhamento do transportador de corrente de segunda geração. (a) representação de blocos funcionais, e em (b) representação em nível de transistores.



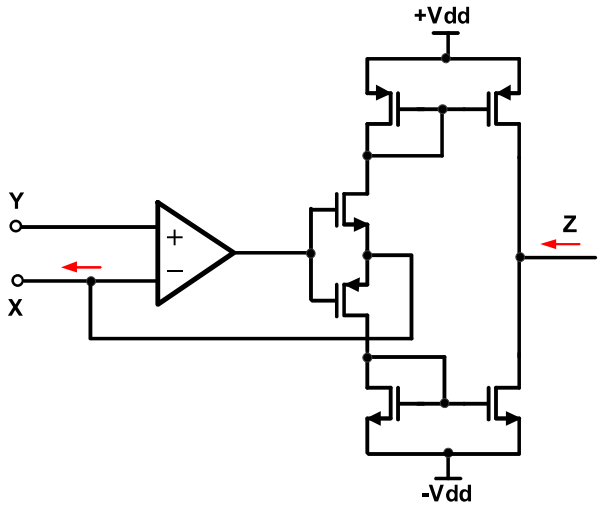
O transportador de corrente proposto foi projetado para que o ponto de conexão com a matriz sináptica opere nos $1º$ e $3º$ quadrantes, para permitir o uso de dispositivos memristivos bipolares como sinapses. A saída do amplificador diferencial conectada ao transistor $M5$ é usado para a regulação da tensão no terminal X. O transistor $M5$ atua acionando, através de $M6 - M8$, o estágio *push-pull,* que controla a tensão e fornecendo corrente no terminal X. Isso faz com que a tensão aplicada no terminal Y seja replicada no terminal X, operando como um curto-circuito virtual, porém, sem necessidade de realimentação negativa para alcançá-lo.

A corrente que flui do nó X é detectada através do espelho de corrente $M10$ e $M11$, espelhada no terminal Z de alta impedância. O estágio de saída, terminal Z, foi modificado. Em sua configuração proposta em (KURASHINA; OGAWA; WATANABE, 1998), existem dois espelhos de corrente interconectados em cascata a fim de suprir corrente no terminal Z para ambos os sentidos, e operação em 4 quadrantes. Porém, concordando com o modo de operação dos neurônios artificias, restrito ao $1°$ quadrante, e a fim de prover um circuito com menor número de transistores, apenas

um espelho de corrente foi implementado para fornecer apenas correntes positivas (excitatórias) para os neurônios.

Sob o ponto de vista de implementação com dispositivos discretos, o transportador de corrente segunda geração pode também ser implementado com a versão de polaridade negativa descrita na Fig. 4.5, apresentado por (TOUMAZOU; LIDGEY; HAIGH, 1990).

Fig. 4.5 - Topologia de circuito para o transportador de corrente segunda geração. Extraído de (TOUMAZOU; LIDGEY; HAIGH, 1990).



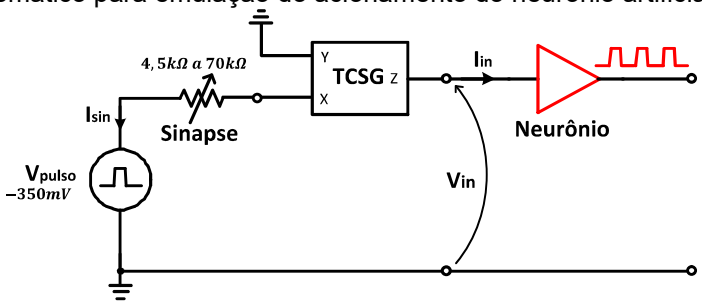
Para satisfazer a relação $V_X = V_Y$, um amplificador operacional é conectado como seguidor de tensão e para satisfazer a equação $i_Z = -i_X$, um circuito seguidor de corrente é conectado em cascata ao amplificador. Na ilustração, dois espelhos de corrente simples são usados para copiar a corrente i_X na entrada e gerar a corrente i_Z na saída. Essa versão compacta para a aplicações em placas de circuito impresso, foi utilizada para demonstração física de uma rede neural, e será apresentada com mais detalhes no capítulo de resultados experimentais.

Na seção seguinte, a topologia de circuito para o TCSG da Fig. 4.4 é analisada quanto a sua versatilidade e compatibilidade com topologias de circuitos para neurônios artificiais. O ambiente de simulação em SPICE foi escolhido a fim de prover facilidade na implementação de modelos matemáticos para os dispositivos comutadores utilizado no disparo do neurônio.

4.4 CONEXÃO DO TCSG COM NEURÔNIOS ARTIFICIAIS.

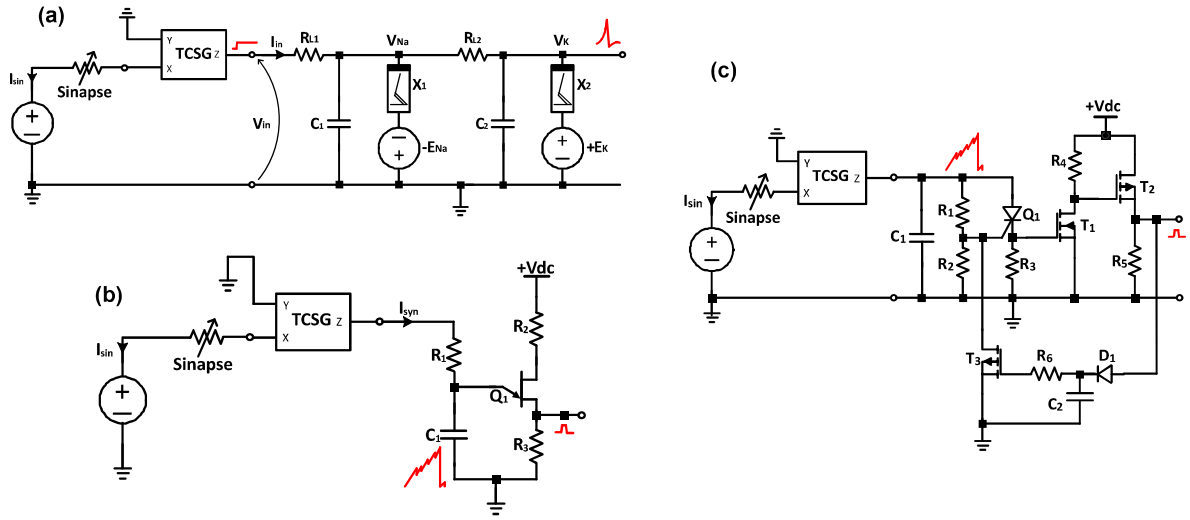
A conexão do TCSG a neurônios artificiais foi investigada pela emulação de pulsos excitatórios a diferentes topologias de circuitos de neurônios utilizado o simulador LTSPICE(MIKKELSEN, 2005). Primeiramente, a transferência de pulsos entre neurônios é emulada usando o circuito simplificado como descrito na Fig. 4.6. Uma fonte geradora de pulsos de tensão representa um neurônio pré-sináptico, que emula o pulso, e um resistor conectado ao terminal X do TCSG representa a sinapse que codifica a informação em corrente. O terminal Z é conectado ao neurônio pós-sináptico, suprindo-o com a corrente sináptica. O terminal Y é aterrado, pois para essa investigação não se necessita de modular a resistência da sinapse.

Fig. 4.6 - Esquemático para emulação do acionamento do neurônio artificial pelo TCSG



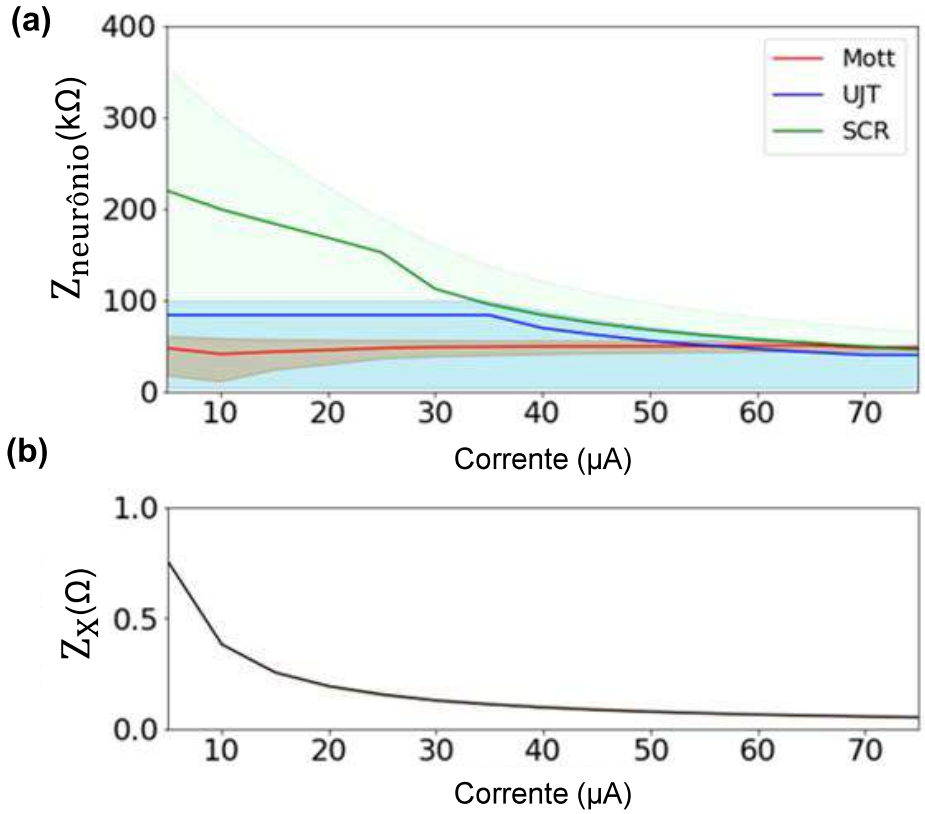
Para compor o ambiente de simulação 3 topologias de neurônios LIF (terminologia do inglês *leak integrate-and-fire*) foram investigadas, cujo dispositivo de comutação é baseado em: retificador controlado de silício (SCR terminologia do inglês *Silicon Controlled Rectifier*) (ROZENBERG; SCHNEEGANS; STOLIAR, 2019), baseado em dispositivos Mott (PICKETT; MEDEIROS-RIBEIRO; WILLIAMS, 2013) e transistores de unijunção (UJT terminologia do inglês *unijunction transistor*) (ASKEW, 1972). Como mostrada na Fig. 4.7.

Fig. 4.7 - Topologias neurônios analisadas. Em (a) neurônio a partir de dispositivos Mott. Em (b) o dispositivo comutador é um transistor de unijunção. E em (c) A comutação é feita através de um SCR.



A conexão direta entre circuitos de sinapses e neurônios afetam a eficiência energética. O circuito pré-sináptico operaria fora da região de maior eficiência pelo fato de sinapse e neurônio não representarem um bom casamento de impedância. A incompatibilidade de impedância entre a sinapse e o neurônio também produziria correntes sinápticas que não carregariam a capacitância da membrana em uma escala de tempo aceitável para aplicações (estimativa de segundos), se conectadas diretamente. Além disso, a conexão direta entre sinapse e neurônio não realiza a codificação tensão-corrente corretamente, pelo fato dos níveis de tensão e corrente no neurônio se alterarem durante sua operação (carregamento da capacitância de membrana). Tensão e corrente mudando ao longo do tempo representam um comportamento dinâmico de impedância sentido pelo circuito que o aciona, e desta forma, a impedância de entrada, para neurônios da topologia LIF, tem um comportamento variável seguindo o carregamento, fuga e descarregamento do potencial de membrana, como mostra os gráficos da Fig. 4.8.

Fig. 4.8 - Uma demonstração de casamento de impedância usando TCSG para 3 tipos diferentes de topologia de circuito para neurônios. (a) Variação de Z de entrada para os neurônios e (b) Z no terminal X do TCSG.



A linha sólida na Fig. 4.8-a representa o valor de impedância médio para os 3 neurônios, enquanto a região colorida representa a variação de impedância de entrada durante o funcionamento, considerando diferentes valores de corrente sináptica. A Fig. 4.8 também mostra que ao acionar os neurônios LIF, o TCSG atua fazendo eficiente a conversão IV, para resistências sinápticas que variam na faixa entre $4{,}5k\Omega$ a $10k\Omega$, enquanto que a impedância do terminal X é quase nula, Fig. 4.8-b, representando um bom casamento para transferência eficiente de informação como discutido no cap. 2. A implementação do circuito de interface via TCSG manteve o circuito dentro da faixa de alta eficiência para transferência de informações entre neurônios, apesar da variação de resistência sináptica e da ampla variação de impedância de entrada para as 3 topologias de circuito neuronais.

As análises mostradas nessa seção apresentam a versatilidade do TCSG sendo usado durante as operações de inferência, no transporte eficiente de informação entre neurônios. No capítulo seguinte, dispõe-se de análises

experimentais a fim de mostrar a utilização dessa topologia de circuito na atuando na fase de treinamento, provendo condições de plasticidade sináptica.

CAPÍTULO 5

INVESTIGAÇÃO EXPERIMENTAL DE REDES NEURAIS POR PULSOS

Com o propósito de investigar as funcionalidades do transportador de corrente de segunda geração para implantação de redes neurais por pulsos, uma plataforma de teste foi concebida servindo como bancada de testes para coleta de dados experimentais. As sinapses utilizadas nas implementações são baseadas em tecnologia de óxido de titânio. Uma demonstração de aprendizado associativo foi usada como experimento protótipo de aprendizado local, revisitando o experimento de Pavlov.

5.1 DETALHAMENTO SOBRE A SINAPSE MEMRISTIVA

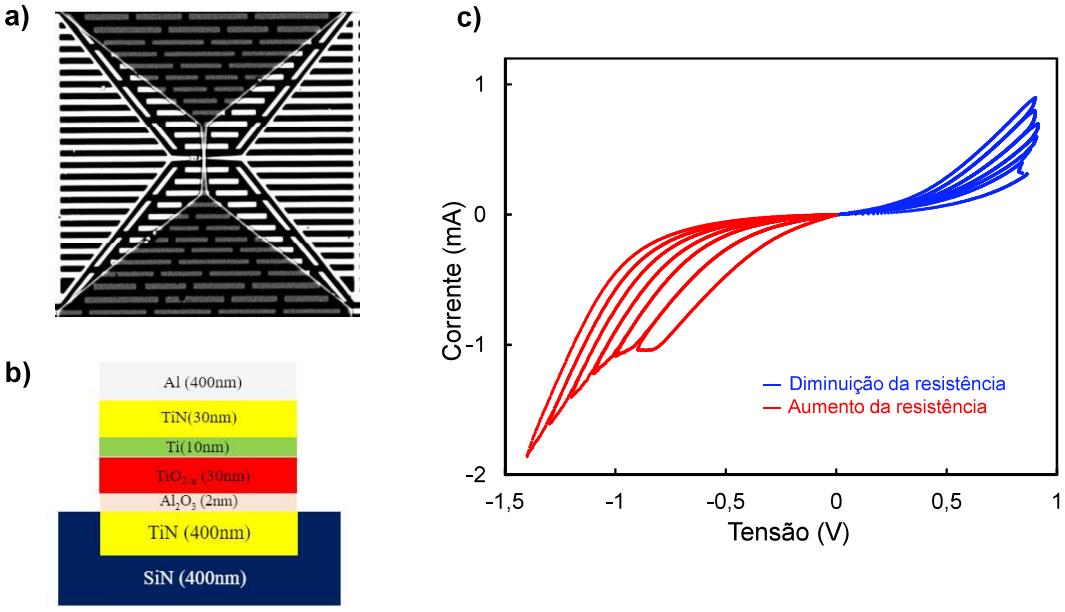
As sinapses deste trabalho consistem em dispositivos memristores arranjados em configuração de barras cruzadas que utilizam camada de chaveamento de óxido de titânio. Na

Fig. 5.1-a, uma imagem do dispositivo feita por microscopia eletrônica de varredura é apresentada. Os detalhamentos sobre o processo de fabricação estão apresentados no apêndice B. A sequência de camadas é esquematicamente apresentada na fig.

Fig. 5.1-b. Os eletrodos inferiores são de TiN e possuem $400nm$ de espessura. Uma camada de Al_2O_3 de $2nm$ foi depositada para melhorar as

características de chaveamento resistivo gradual e reduzir a corrente de fuga, como descrito em (BEILLIARD et al., 2020). Em seguida, uma camada não estequiométrica de $30nm$ de TiO_{2-X} é depositada. Uma camada de Ti de $10nm$ de espessura foi depositada sobre a camada de óxido de titânio para criar um gradiente de vacância de oxigênio, benéfico ao desempenho da comutação. Esta camada também atua como uma camada de retenção para os íons de oxigênio em difusão. Para compor os eletrodos superiores, camadas de TiN/Al de $30/400nm$ de espessura, respectivamente, foram depositadas.

Fig. 5.1 - Dispositivo memristor usado como sinapse. Em a) imagem de microscopia eletronica de varredura. Em b) sequência de camadas. Em c) caracterização elétrica quase-estática.

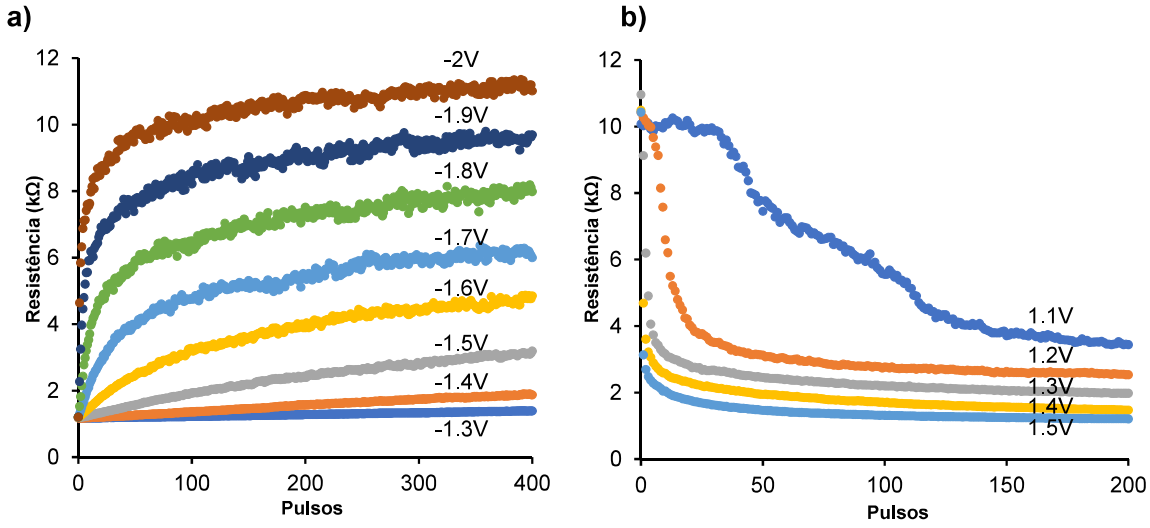


O comportamento característico do memristor pode ser observado nas curvas IV da

Fig. 5.1-c, para o chaveamento progressivo da resistência do dispositivo. Evidenciando a característica bipolar do dispositivo, as excursões de tensão positiva aplicadas reduzem a resistência gradualmente e enquanto para tensões negativas, a resistência aumenta gradualmente. A cada varredura de tensão, observa-se a sobreposição das curvas destacando o comportamento não volátil da resistência do memristor. O condicionamento da resistência também pode ser realizado através de excitação por pulsos de tensão. Este tipo de operação oferece uma maneira de reduzir a influência dos efeitos térmicos e possui condições operacionais simplificadas, mais condizente com o ambiente de operação interno aos circuitos integrados(ALIBART et

al., 2012). O incremento de resistência pode ser visto na Fig. 5.2-a. No total, 400 pulsos de tensão negativa foram aplicados ao dispositivo para diferentes amplitudes. A sequência de pulsos era composta por um pulso de escrita de $200ns$, valor típico de largura de pulso para dispositivos memristores de óxido de titânio, seguido por um pulso de leitura de $1ms.$, valor típico para ciclo de leitura de corrente para instrumentação comercial. Após cada conjunto de pulsos, uma varredura de tensão com polaridade oposta foi realizada para recuperar o estado inicial de alta resistência. A redução da resistência acompanha protocolo de teste similar, mudando a polaridade dos pulsos. Os resultados são apresentados na Fig. 5.2-b, onde testes para 200 pulsos foram realizados.

Fig. 5.2 - Protocolos de testes para incremento e redução da resistência.

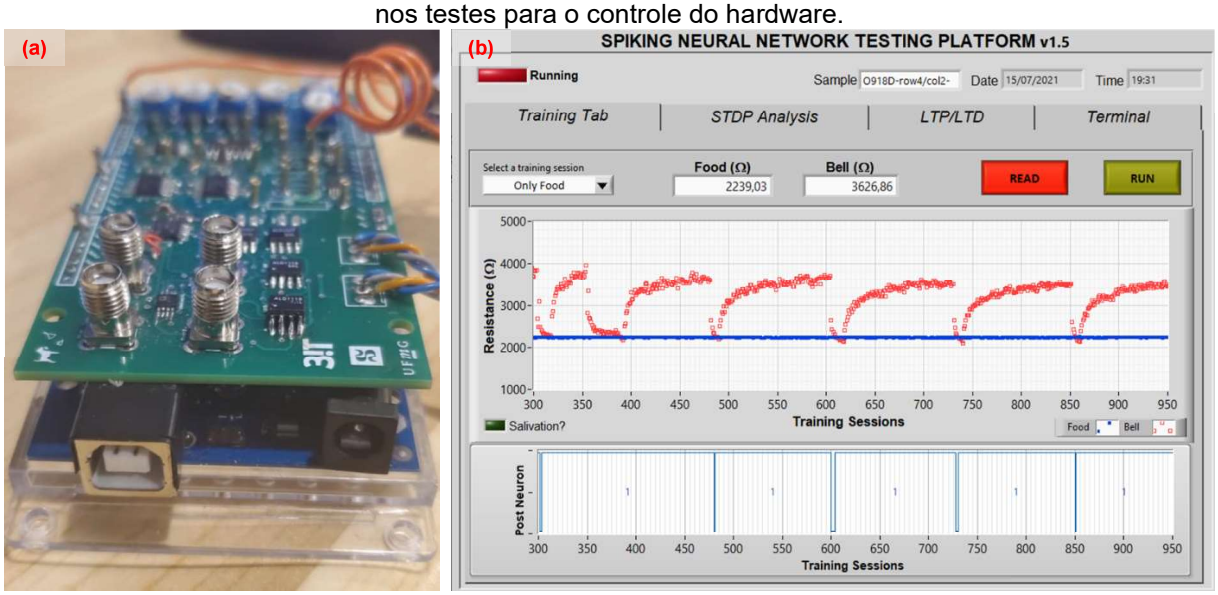


5.2 PLATAFORMA DE TESTES PARA REDES NEURAIS POR PULSOS

Uma plataforma de testes foi especialmente desenvolvida para explorar as funcionalidades do TCSG em redes neurais. Primeiramente, decidiu-se por uma implementação em hardware que portasse de comunicação com um computador a fim de fornecer recursos de coleta e pós-processamento dos dados gerados pela rede durante o seu funcionamento. Optou-se por uma arquitetura simples composta por uma placa customizada no formato de um *shield* (NAYYAR; PURI, 2016) compondo os circuitos da rede neural, que se conectam a uma unidade microcontrolada Arduino MEGA, como mostra a Fig. 5.3-a. A unidade microcontrolada Arduino (interface sensorial) emula sinais de estímulos externos, que podem ser gerados por sensores. Ela é responsável por gerar estímulos artificialmente, demonstrando ser versátil a

promover múltiplas formas de codificação de pulsos (AUGE et al., 2021). Essa opção implica em flexibilidade para adaptar a plataforma para processar dados reais. Sensores ou outros dispositivos, podem ser facilmente conectados a plataforma Arduino, que opera como interface como o mundo externo às redes neurais, como sugere a seguinte referência (BELL, 2014). A escolha de uma unidade microcontrolada para o gerenciamento dos testes amplia as possibilidades de análise e caracterização da rede. Na ausência de sensores, é possível permitir que todo o processo de aprendizagem da rede seja emulado, dados sejam gerados ficticiamente, que um usuário programe sessões de treinamento, pause o processo e analise o andamento do treinamento.

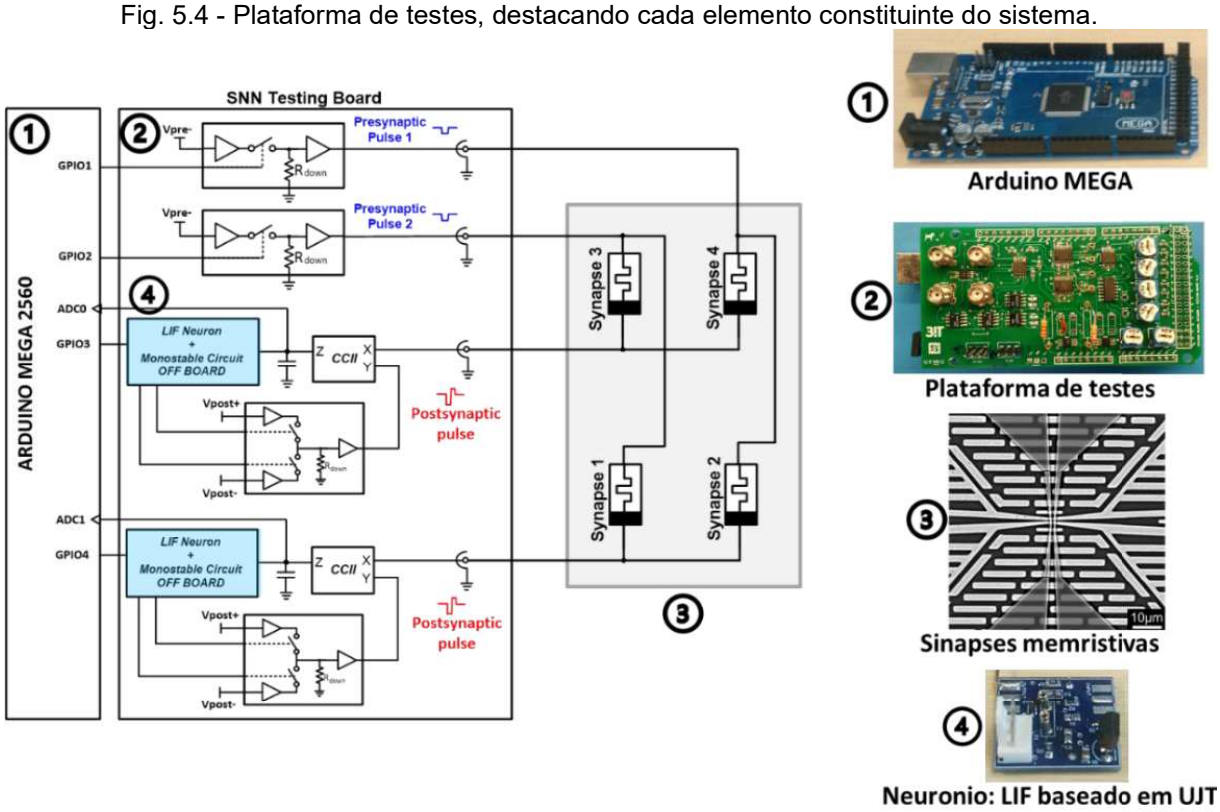
Fig. 5.3 – Em a) Placa customizada para os testes das redes neurais por pulsos. Em (b), API usada nos testes para o controle do hardware.



A interação com o computador é feita através de uma API (terminologia do inglês *Application Programming Interface*) desenvolvida em ambiente LabVIEW, que auxilia o usuário a usar os recursos disponíveis na placa, como apresentados na Fig. 5.3-b. As aplicações embarcadas na interface compõem um conjunto de funcionalidades que se resumem a caracterização do dispositivo, análise de regra de treinamento e emulação de rede neural. Mais detalhes sobre as funcionalidades da plataforma são apresentados no Apêndice B. A plataforma de testes permite a implementação física de redes com até 4 sinapses, numa estrutura 2x2, dispondo de conectividade para testes sobre o substrato, através de uma *probe station*. A plataforma, e todas as funcionalidades embarcadas são esquematicamente

apresentadas na Fig. 5.4. Cada componentes da plataforma é representado ao lado pelos rótulos.

Fig. 5.4 - Plataforma de testes, destacando cada elemento constituinte do sistema.



Os pesos sinápticos são representados na forma de resistência, e, para análise do treinamento, requer a necessidade de acompanhar a progressão da resistência de cada célula de memória logo após cada etapa de treinamento, a fim de criar os perfis de resistências sinápticas ao longo das épocas. A técnica de sensoriamento de corrente é realizada usando a capacitância da membrana do pós-neurônio como elemento sensor. Após a capacitância ser completamente descarregada, no final de cada ciclos de treino, é então aplicado um pulso de tensão sobre a sinapse a ser lida. O microcontrolador, por meio de um conversor analógico-digital, registra a variação de tensão sobre a capacitância, medindo indiretamente a corrente sináptica, através da expressão 5.1.

$$R_{sin} = \frac{V_{pulso} \cdot \Delta t}{C_{mem} \cdot \Delta V} \tag{5.1}$$

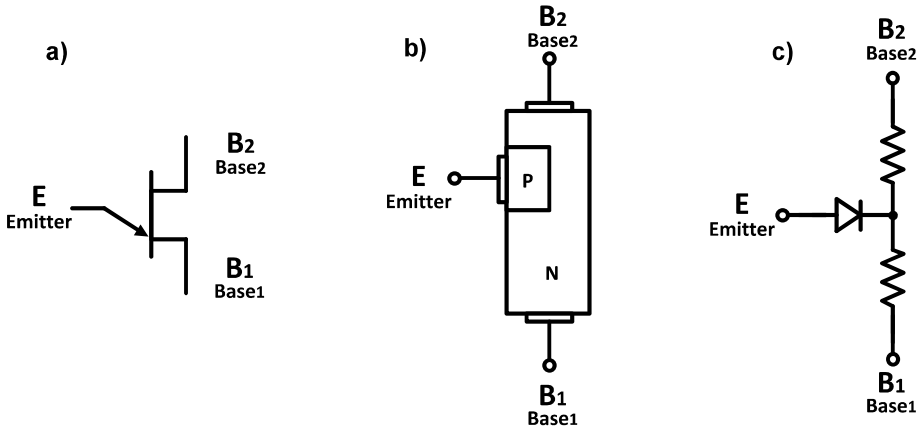
Onde R_{sin} é a resistência da sinapse, V_{pulso} é a amplitude do pulso, Δt é o tempo entre duas de amostragem de tensão do conversor AD, C_{mem} é a capacitância

de membrana e ΔV é a variação de tensão registrada pela Arduino. O valor de resistência sináptica é calculado internamente e apresentado por meio de uma interface que traça o perfil de resistência sináptica durante o treinamento.

5.3 NEURÔNIO ARTIFICIAL BASEADO EM TRANSISTOR DE UNIJUNÇÃO

O dispositivo usando como comutador no pós-neurônio é o transistor unijunção. Esse dispositivo possui 3 terminais que consiste em um substrato de silício do tipo n levemente dopado na qual existe uma região do tipo p fortemente dopado onde o terminal de EMISSOR está incorporado (HACHTEL; PEDERSON, 1962). Nas duas extremidades, existem contatos ôhmicos designados como BASE 1 e BASE 2 conforme mostrado na Fig. 5.5.

Fig. 5.5 - Transistor de unijunção usado como dispositivo comutador em neurônios. Em (a) A simbologia. Em (b) a representação esquemática construtiva e em (c) circuito equivalente.

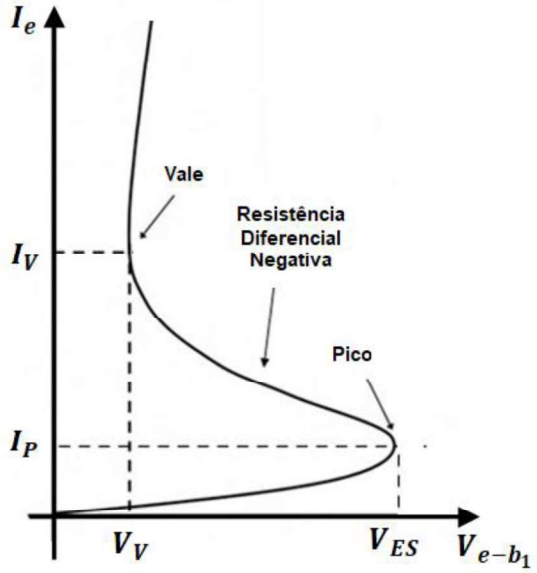


A resistência entre as duas bases está na faixa entre $5k\Omega \; a \; 10k\Omega$. O circuito equivalente consiste em um diodo de junção pn e a resistência interbase dividida em duas partes $RB1$ e $RB2$, ver Fig. 5.5-b. A transição de "desligado" para "ligado" é controlada pela tensão do emissor, Ve, quando o dispositivo é acionado por corrente. A curva IV do UJT é apresentada na

Fig. 5.6. Quando $Ve - b_1 < Ves$ (tensão de saturação do emissor), o diodo equivalente não está diretamente é polarizado. Este é o estado "desligado" do dispositivo como região de corrente muito baixa na curva IV. Quando a tensão do dispositivo excede o limiar, $Ve - b_1 > Ves$, o diodo torna-se polarizado diretamente. Este é o estado "ligado" do dispositivo. Devido ao fluxo de Ie através de $RB1$, o número

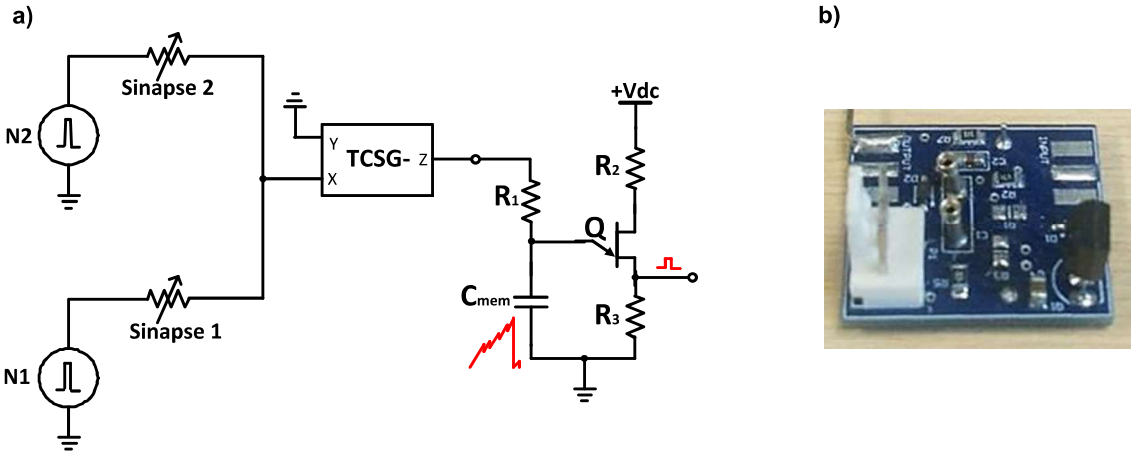
de portadores de carga em $RB1$ é aumentado, o que reduz sua resistência, o que, por sua vez, diminui a tensão incorporada em $RB1$. Isso faz com que o diodo se torne cada vez mais diretamente polarizado e Ie aumente ainda mais, levando a uma ação regenerativa. Ve diminui com o aumento de Ie, apresentando resistência diferencial negativa.

Fig. 5.6 - Curva IV para o transistor de unijunção.



Eventualmente, o ponto de vale será alcançado, após o qual não haverá mais diminuição de $RB1$. Após o ponto de vale, o dispositivo atingirá o estado de saturação. As características IV não lineares, NDR e histerese do transistor de unijunção são os principais recursos que permitem uma implementação simples de um neurônio LIF. A proposta de circuito para o neurônio, apresentada na Fig. 4.7-b, é fisicamente implementada com componentes discretos para os testes experimentais, como mostra a Fig. 5.7.

Fig. 5.7 - Em a) diagrama de circuito para uma RNP simples composto por 2 neurônios pré-sinápticos, TCSG- e um LIF a partir de TUJ. Em b), O neurônio implementado em uma pequena PCI.

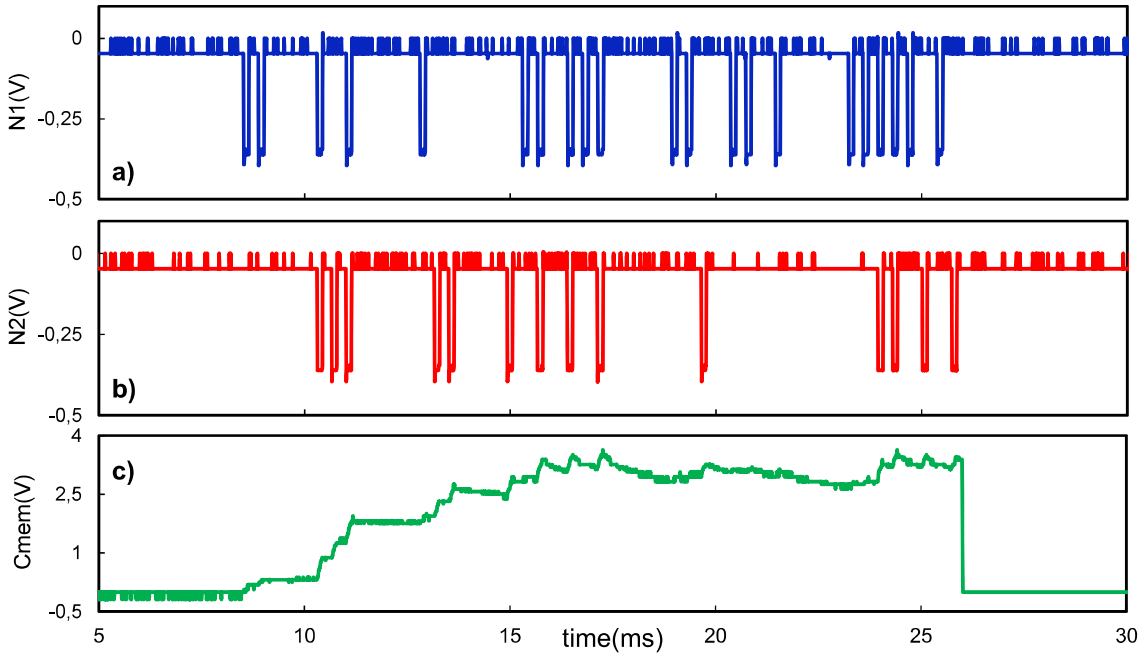


Quando o transportador de corrente injeta a corrente sináptica no $C1$ (representando a capacitância da membrana), o potencial da membrana aumenta. Quando $Vc1$ atinge o Vp (tensão de ponto de pico), a junção $E-B1$ é diretamente polarizada e o TUJ liga. O capacitor $C1$ descarrega rapidamente através de $R3$ e o TUJ desliga quando a tensão decai para a tensão de vale Vv. O processo de descarga do potencial de membrana está relacionado às constantes de tempo $R3$ e $C1$ e pode estar associado ao período de relaxamento ou refratário do neurônio. É importante notar que a tensão de saturação do emissor está relacionada ao potencial no ponto entre o diodo e os resistores interbase Vc. Esta tensão é definida pelos valores de componentes externos e tensão Vdc, conforme mostrado na Fig. 5.7, e definida pela expressão 5.2.

$$V_c = V_{dc}\left(\frac{R_{B1}+R_3}{R_2+R_{B1}+R_{B2}+R_3}\right) \qquad (5.2)$$

Onde Vc é o potencial entre os resistores de interbase, Vdc é a tensão de alimentação do neurônio, R_2 e R_3 são resistores de polarização do neurônio. O ajuste dos parâmetros da equação 5.2 podem ser usados como mecanismo para implementação de técnicas de homeostase baseadas no controle dos limiares dos neurônios (LIU et al., 2001). Ao controlar esses parâmetros, é possível controlar os neurônios que disparam em tempo de operação.

Fig. 5.8 - Formas de onda para o neurônio baseado em TUJ. Em a) formas de onda para os pulsos pré-sinápticos do neurônio 1 e b) do neurônio 2. Em c) A evolução do potencial de membrana durante a operação da RNP.

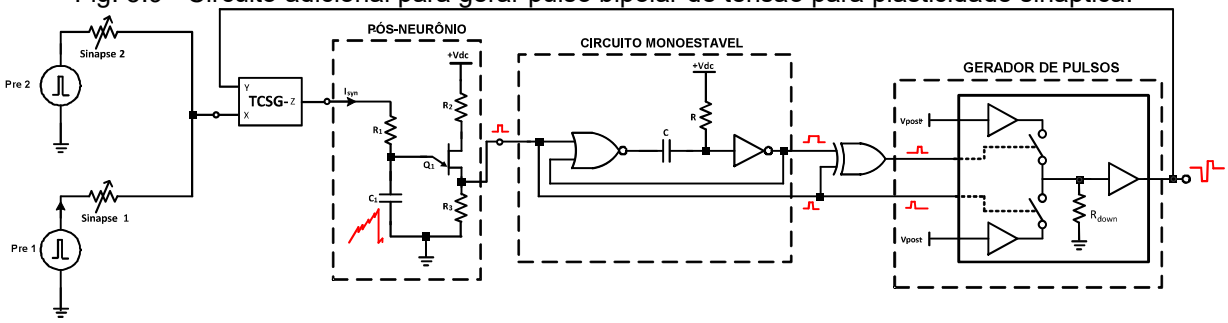


5.4 CIRCUITO GERADOR DE PULSOS RETROATIVOS

Compondo as funcionalidades para aplicação em redes neurais fisicamente implementadas, o TCSG pode prover capacidade de plasticidade sináptica, uma vez que esse circuito permite a propagação de pulsos de feedback. A regra de aprendizagem utilizada foi baseada na plasticidade dependente da temporização entre pulsos (STDP), fornecendo uma demonstração experimental sobre a utilidade do TCSG como uma primitiva de neurocircuito. A sinapse artificial é implementada por um dispositivo memristivo de óxido de titânio (MESOUDY et al., 2021) e, portanto, a modulação por pulsos de tensão e polaridade da excitação são importantes para definir a plasticidade. A sobreposição, ou não, dos pulsos pré e pós-sinápticos representam a causalidade entre dois neurônios. Para um pulso pós-sináptico que é disparado ainda sob a atuação do pulso pré-sináptico, isso representa uma resposta do neurônio estimulado perante o neurônio gerador do estímulo. Justifica-se assim, reforçar o elo entre eles com a potenciação sináptica (diminuição da resistência). Entretanto, se o neurônio pós-sináptico dispara, mas não pela ação do estímulo do pulso do neurônio pré-sináptico, o disparo não representa causalidade e por isso justifica-se que naquele elo entre os neurônios ocorra uma depressão sináptica

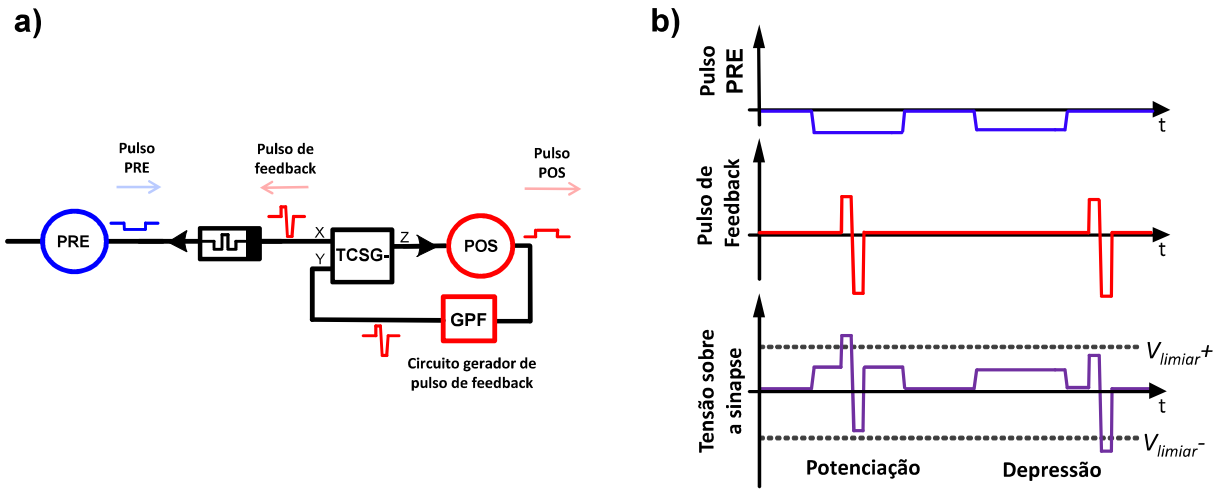
(aumento da resistência). O neurônio proposto neste trabalho tem como resposta pulsos unipolares positivos de dois níveis (ASKEW, 1972). No entanto, os memristores são dispositivos bipolares onde sinais de polarização bipolar são necessários para produzir condições para potenciação e depressão. Para tal necessidade, foi projetado um circuito de condicionamento, conforme mostra a Fig. 5.9, para gerar pulsos bipolares a partir do pulso de disparo do pós-neurônio.

Fig. 5.9 - Circuito adicional para gerar pulso bipolar de tensão para plasticidade sináptica.



Um circuito monoestável e uma porta lógica XOR foram usados para gerar pulsos defasados de alguns nanosegundos. Esses pulsos acionam circuitos de *pull-up-down* para a obtenção do pulso bipolar. Combinando os picos unipolares de excitação dos pré-neurônios com os picos bipolares do circuito gerador de pulso de feedback, é possível gerar condições de potenciação e depressão nas sinapses. Inspirado na regra de aprendizagem STDP apresentada em (QUERLIOZ et al., 2013), as formas de onda mostradas na Fig. 5.10-b resumem as condições de plasticidade a longo prazo para os dispositivos memristores de óxido de titânio. Uma representação esquemática das conexões entre uma sinapse e neurônios, é mostrado na Fig. 5.10-a.

Fig. 5.10 - Esquema de polarização para a STDP segundo (QUERLIOZ et al., 2013).



A razão para aplicação de tensões negativas para os pulsos pré-sinápticos está relacionada com a polaridade negativa do TCSG. O circuito TCSG- é mais simples, e possui número reduzido de componentes, representando um circuito mais compacto para a implementação física. Um sinal unipolar de tensão negativa do pré-neurônio acarretará uma corrente saindo do terminal Z, sentido correto para o carregamento do neurônio. O neurônio pós-sináptico, quando disparado, gera um pulso unipolar positivo, que pode ser adaptado a circuitos de *pull-up-down* e buffer de tensão, demonstrado na Fig. 5.4-②, a fim de se tornar o pulso compatível a ser aplicado em arquiteturas de redes neurais multicamadas. O neurônio pós-sináptico é conectado a um circuito gerador de pulso de feedback, representado na Fig. 5.10-a pela sigla **GPB**. A função desse circuito é de gerar um pulso de feedback compatível com as regras de aprendizagem e níveis de operação para a sinapse escolhida. Um pulso de feedback é um pulso bipolar de 3 níveis, como demonstrado na Fig. 5.10-b. A sobreposição, ou não, entre os pulsos pré e pós-sinápticos induz um aumento (depressão) ou uma diminuição (potenciação) na resistência, respectivamente.

Na Fig. 5.11, formas de onda coletadas com um osciloscópio são apresentadas, representando as condições para a plasticidade sináptica. Os parâmetros usados neste experimento foram: pulso pré-sinápticos de $500\mu s$ de largura e $-350mV$ de amplitude; os pulsos de feedback são $+1V/-1{,}6V$, com largura de pulso de $300ns$. Detalhes do pulso de feedback são mostrados na Fig. 5.11 sobre as condições de potenciação(a) e depressão(b).

Fig. 5.11 - Formas de onda sobre a sinapse para a) potenciação e b) depressão. A área dentro da curva pontilhada representa a ampliação, e é apresentada na figura seguinte.

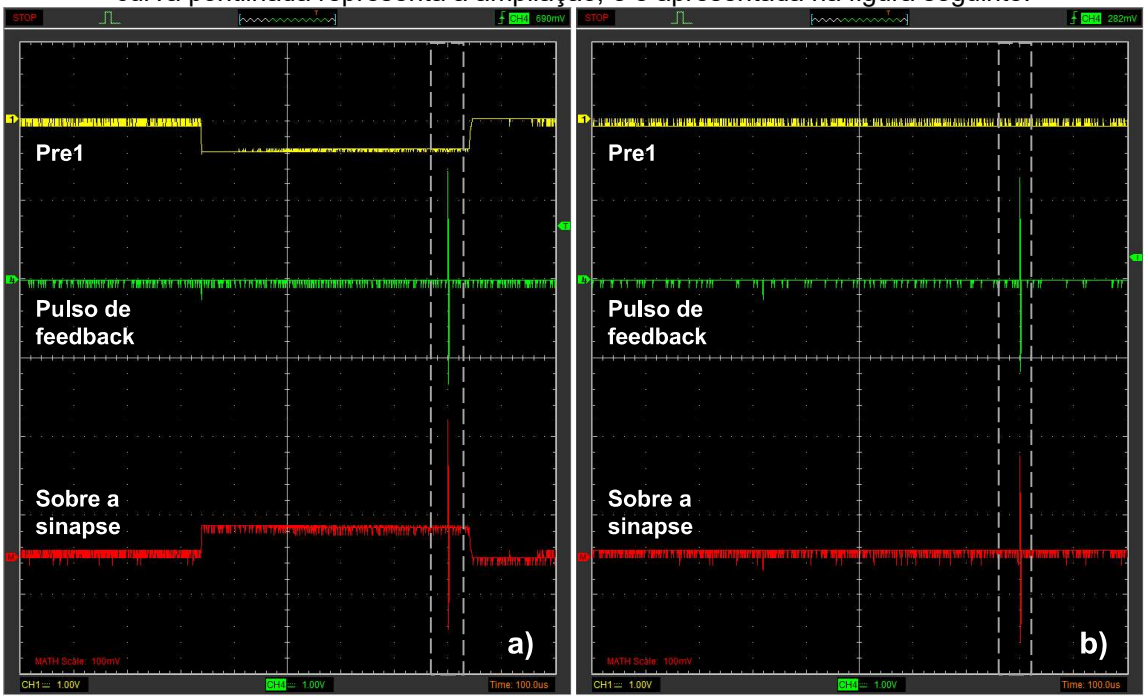
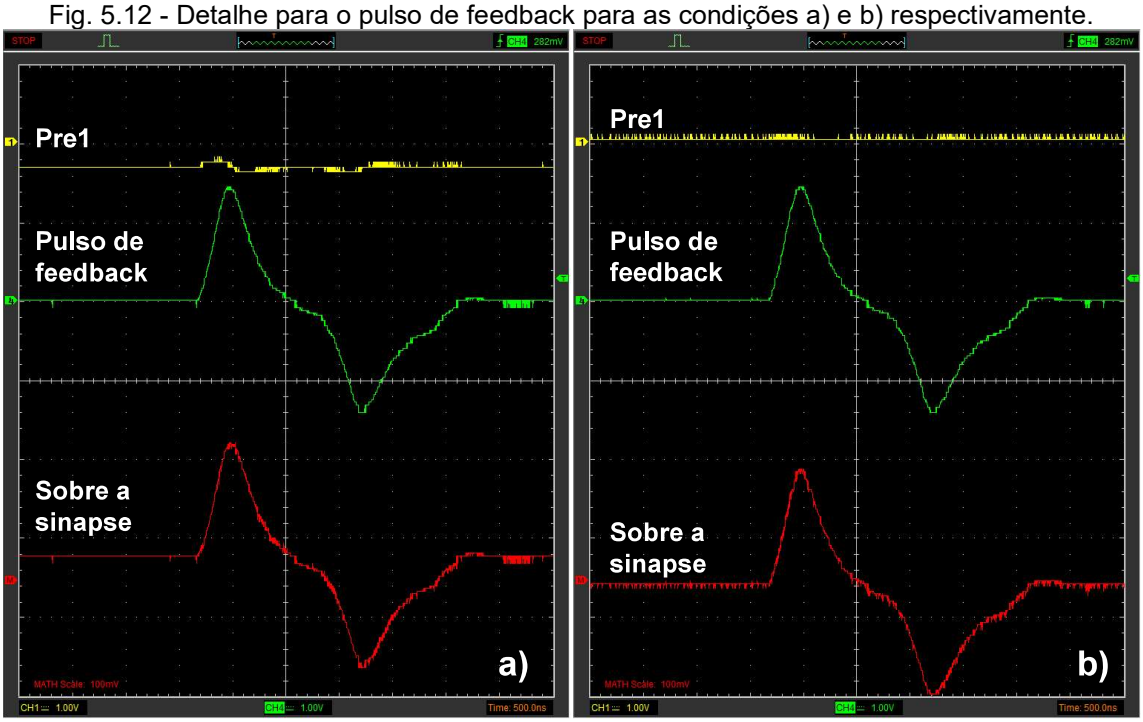
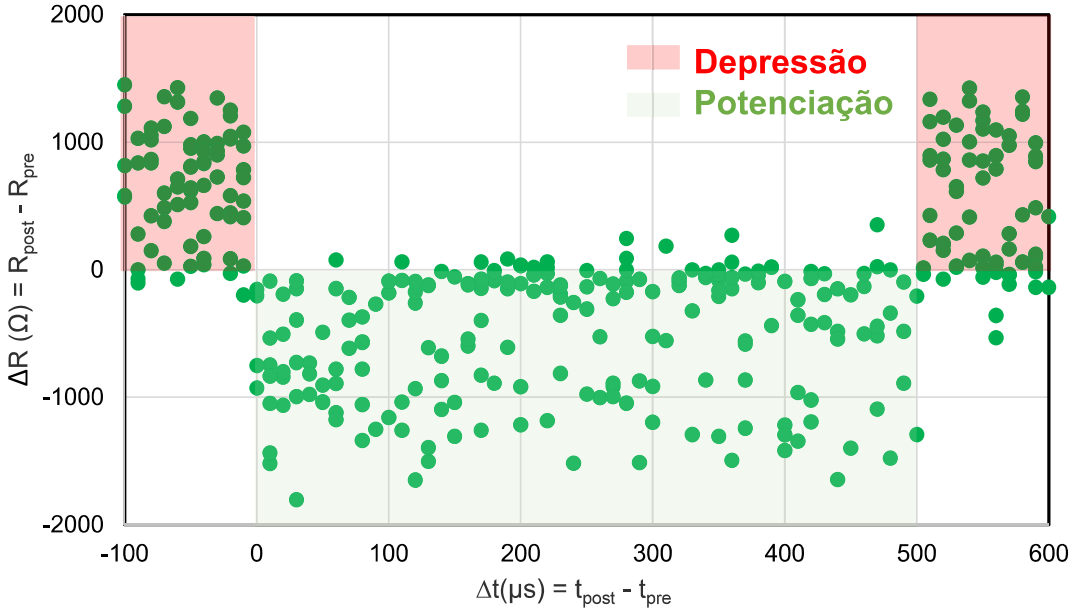


Fig. 5.12 - Detalhe para o pulso de feedback para as condições a) e b) respectivamente.



Dessa forma, a plataforma desenvolvida pode emular combinações de pulsos pré e pós-sinápticos a fim de caracterizar o dispositivo de memória segundo a lei de aprendizagem. Pulsos pré-sinápticos no eletrodo inferior e pulsos pós-sinápticos no eletrodo superior fornecem condições adequadas para a análise da plasticidade, conforme descrito na Fig. 5.13.

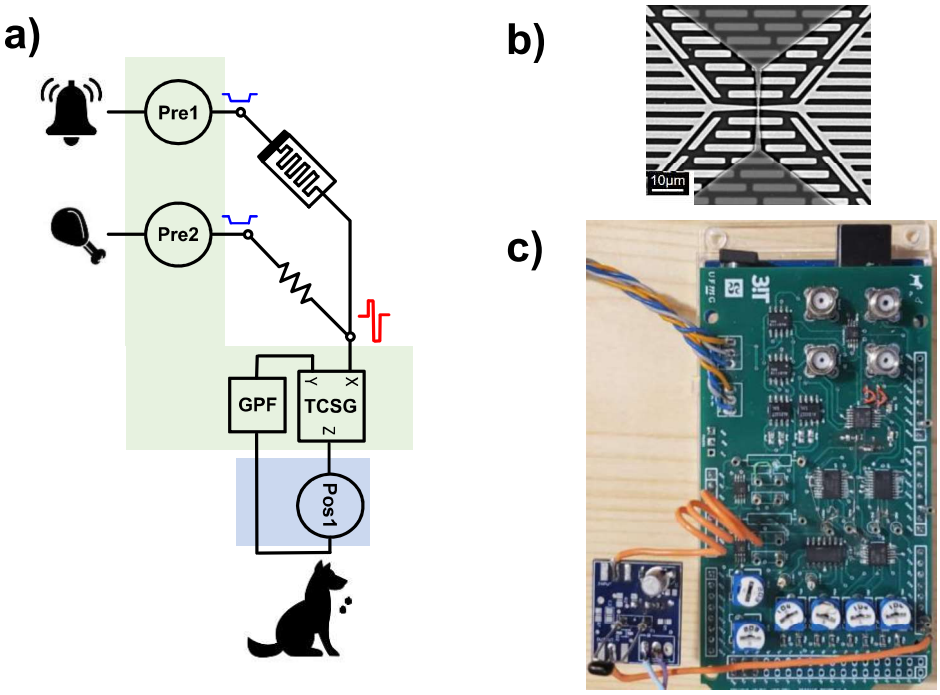
Fig. 5.13 - Gráfico STDP para esquema de polarização da Fig. 5.10.



5.5 DEMONSTRAÇÃO DE APRENDIZADO ASSOCIATIVO

Uma rede neuromórfica foi implementada usando a plataforma desenvolvida, revisitando o experimento de aprendizado associativo do cachorro de Pavlov. Neurônios pré-sinápticos representando $comida$ e $sino$ são circuitos geradores de pulsos implementados na plataforma. A codificação de pulso pré-sináptica é definida por taxa, com trem de pulsos de mesmos parâmetros da demonstração de treinamento da seção passada ($500\mu s$ de largura de pulso e $-350mV$ de amplitude). Os intervalos entre os picos variam de $500\mu s$ a $2ms$ seguindo uma distribuição de Poisson. Cada sessão de treino dura em torno de $20ms$, após isso o microcontrolador registra a quantidade de disparos e realiza um ciclo de leitura para cada sinapse para registrar a modificação da resistência durante o treino. Os ciclos de leitura de resistência realizado pela plataforma existem apenas para monitorar o processo de aprendizado, não para atuar na rede. Uma representação esquemática do experimento é apresentada na Fig. 5.14-a, destacando todo o hardware utilizado e suas conexões em Fig. 5.14-(b e c).

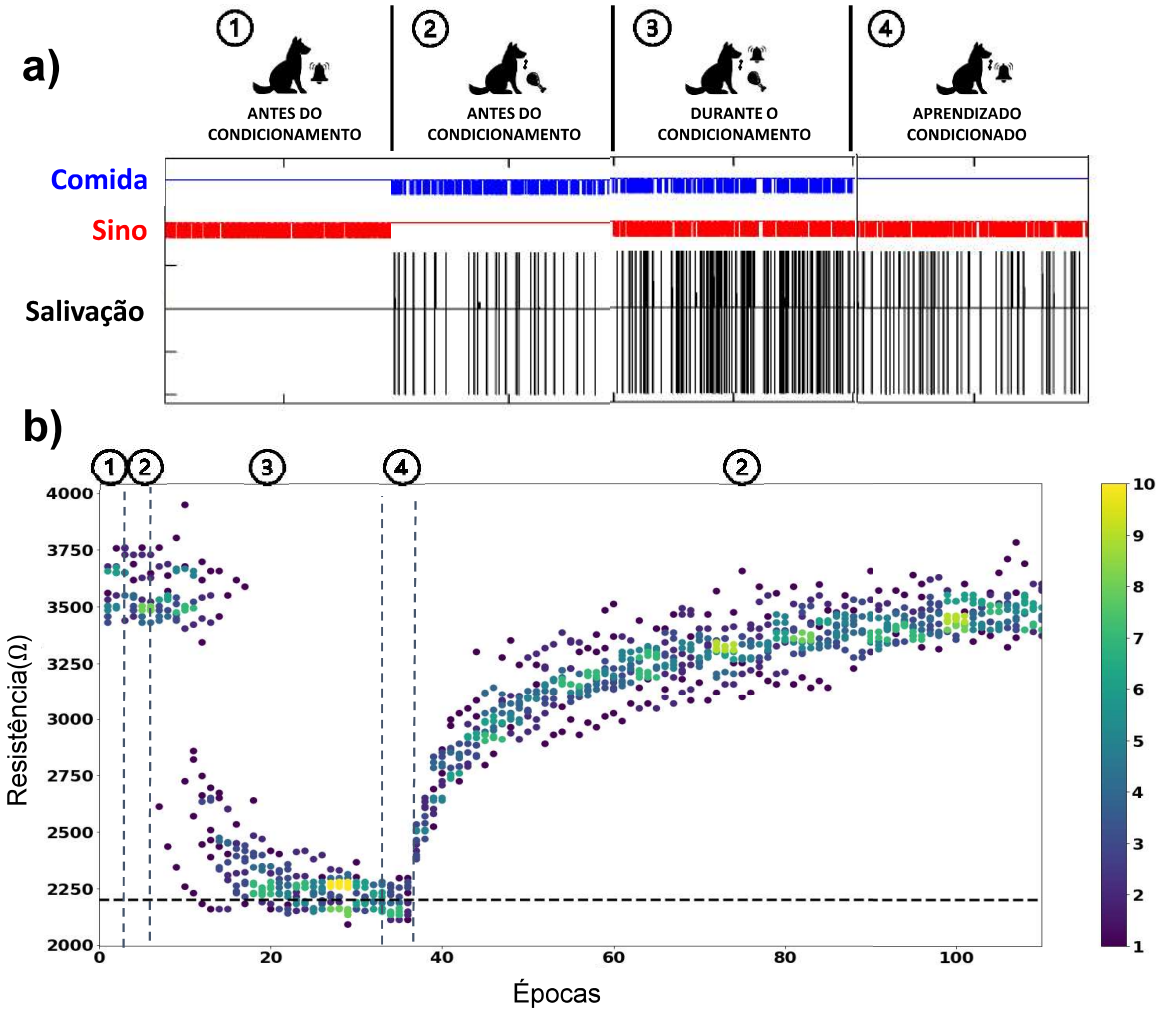
Fig. 5.14 - Aprendizado associativo demonstrado em hardware. Em a), a representação esquemática do experimento. Em b), imagem da sinapse usada. Em c), Conexões entre plataforma e neurônio utilizado.



O neurônio pós-sináptico de $salivação$ está conectado ao circuito diretamente ao terminal Z do TCSG, conforme mostrado na Fig. 5.14. A sinapse de $comida - salivação$ é representada por um resistor de 2200Ω e a sinapse $sino - salivação$ é representada por um memristor de óxido de titânio, que é inicializado em seu estado de alta resistência de $\sim 3600\Omega$. O memristor foi conectado diretamente ao terminal X por meio de sondas discretas. O valor de resistência sináptica $comida - salivação$ é escolhido para produzir um estímulo não-condicionado, e portanto o neurônio pós-sináptico $salivação$ apresenta resposta quando o neurônio pré-sináptico $comida$ é estimulado, como mostrado na Fig. 5.15(sessão 2). Como a sinapse $sino - salivação$ tem maior resistência, o estímulo condicionado não induz nenhuma resposta do neurônio de saída Fig. 5.15(sessão 1), pois os estímulos ponderados pela sinapse *sino-salivação* não são suficientes para carregar o potencial de membrana do neurônio pós-sináptico até seu limiar de disparo. Então, quando submetido ao treinamento, Fig. 5.15(sessão 3), com estímulos de $comida$ e $sino$, a resistência da sinapse do $sino - salivação$ diminui de acordo com a regra de aprendizagem STDP estabelecida. Isso leva a uma resposta do neurônio $salivação$ ao estímulo condicionado do $sino$ Fig. 5.15(sessão 4).

Após a aprendizagem associativa, foi aplicado um conjunto de sessões apenas com estímulo $comida$, ver Fig. 5.15(sessão 2). Este teste foi feito para promover a depressão sináptica a fim de dissociar o aprendizado adquirido anteriormente. A sequência entre as sessões foi programada, e, portanto, a operação da rede e o processo de aprendizado ocorreram naturalmente nos circuitos implementados. A plataforma não interveio no treinamento, apenas realizou ciclos de leitura após cada sessão de treino. Como análise de repetibilidade foi adicionando um gráfico de recorrência, na Fig. 5.15-b.

Fig. 5.15 - Demonstração de aprendizado associativo. Em a), um perfil temporal para os pulsos segundo os estímulos "comida" e "sino", e a resposta de "salivação" para cada etapa das sessões de treinamento. Em b) O perfil de resistência sináptica, destacando a resiliência, para 10 ciclos de treinamentos.



CAPÍTULO 6

CONCLUSÕES E APONTAMENTOS FUTUROS

Este trabalho abordou fundamentos e proposições sobre pesquisas orientadas a circuitos com memristores de comutação resistiva. Há uma grande promessa na utilização de memristores não apenas no armazenamento de memória digital, mas também principalmente como elemento emulador de sinapses artificiais, usufruindo de sua capacidade de armazenamento não-volátil e analógico. Até este momento, grande parte das investigações apresentadas na literatura consistem em demonstrações envolvendo modelos de dispositivos idealizados ou fora da realidade prática. Este trabalho concentrou-se no projeto de circuitos memristivos e neurônios, implementando módulos que permitam alcançar os limites fundamentais para a maximização da eficiência energética de sistemas neuromórficos, ponto chave da escalabilidade de redes neurais. No aspecto de projetos de redes neurais fisicamente implementadas, este trabalho elencou as funcionalidades para os elementos de circuitos, e apresentou alternativas para blocos formadores, antes de implementar circuitos práticos e abordar demonstrações experimentais. Em síntese, esta tese cobre implementações de circuitos neuromórficos com memristores e neurônios artificiais, reunindo detalhes construtivos e funcionais, contribuindo para a maturidade da tecnologia.

Inspirando-se em sistemas computacionais biológicos, que correspondem o auge da eficiência energética, um aspecto pouco levado em consideração em sistemas neuromórficos foi abordado: implementar eficiente circuitos processadores

analógicos de informação se refere a propor adequadas razões de impedância entre elementos. O casamento de impedância entre os blocos formadores e a necessidade de um circuito de interface sinapse-neurônio, promove condições de operação e maximização do tráfego do sinal ao longo da rede. Uma analogia entre os mecanismos de casamento de impedância pode ser feita quando comparamos as sinapses químicas com o sistema com transportadores de corrente. A interação da substância mensageira neurotransmissor estabelece condições para que a informação seja transmitida entre dois neurônios, de maneira a vencer as barreiras físicas impostas no meio. A utilização de um agente seletivo que age na abertura e fechamento de canais iônicos de um neurônio, permite que um sinal de baixa intensidade possa excitar o potencial de membrana de um outro neurônio adjacente independendo de suas diferenças físicas de tamanho e funcionalidade. De maneira análoga opera o memristor conectado a um TCSG. Circuito pré-sináptico permite que a corrente seja eficientemente ponderada e o circuito pós-sináptico, com características de fonte de corrente, faz com que essa corrente seja transportada ao neurônio independente de suas características topológicas e elementos de circuitos, que se refletem em uma impedância vista pelo circuito transportador.

Para aprofundar no trabalho experimental de sistemas neuromórficos, uma plataforma versátil de caracterização e implementação de redes neurais foi projetada, construída e validada com dispositivos reais. O instrumento foi direcionado a um conjunto de testes que contribuem para a investigação do comportamento de dispositivos memristivos sob a abordagem de regras de treinamento, e investigação de topologia físicas de neurônios. A **SNN testing platform v1.5**, é uma plataforma na qual uma ampla gama de procedimentos pode ser facilmente executada. Isso inclui forçar um memristor a uma combinação específica de estímulos pré e pós-sinápticos (para fins de estudo de plasticidade), ou aplicar protocolos básicos para a funcionalização e caracterização do dispositivo: eletroformação, chaveamentos de SET e RESET, testes de ciclagem para potenciação e depressão. O sistema foi projetado para permitir a composição de redes neurais com diferentes tipos e topologias de circuitos para neurônios e memristores, aproveitando-se de características importantes do transportador de corrente de segunda geração na composição de circuitos de interface sinapse-neurônio. Esta plataforma também pode ser utilizada para emular redes neurais, como demonstrado no experimento de aprendizagem associativa, e facilmente integrada à sensores tornando factível

implementações reais. Um sumário detalhado das principais contribuições da tese é elencado a seguir.

6.1 CONTRIBUIÇÕES DESTE TRABALHO

O Capítulo 2 forneceu uma visão geral sobre a importância da eficiência energética para o escalonamento de redes neurais. A inspiração em sistemas computacionais biológicos conduz para que as implementações físicas de sistemas neuromórficos se aproximem da representação de dados e do modo de processamento de informação que o cérebro realiza, o que influencia a utilização de uma série de dispositivos e circuitos candidatos a mimetizar os elementos formadores das redes neurais. Começando com os dispositivos de comutação resistiva, os quais foram inicialmente teorizados por Chua, e posteriormente sintetizados pela HP. A tecnologia atual de memristores dispõe de diferentes mecanismos fundamentais de chaveamento resistivo, porém, este trabalho focou em um tipo baseado em filmes de óxido de titânio. Também, foram elencadas as características fundamentais dos circuitos neuronais, e principalmente os dispositivos de resistência diferencial negativa, que regem as funcionalidades desses circuitos. Em vista disso, é observado a necessidade de um circuito que compatibilize as condições operacionais de sinapses e neurônios, principalmente sobre as suas variações de impedância. Visando a escalabilidade das redes neurais, é levantado a importância de desenvolver sistemas de aprendizado online e cada vez mais compactos, sem a necessidade de sistemas adicionais para cálculos dos pesos sinápticos, e sem supervisão.

O Capítulo 3 forneceu uma metodologia de projetos de sistemas neuromórficos baseado nos limites fundamentais da transmissão de informação interneuronal, comparando-a à comunicação através de um canal sináptico. A análise da variação de parâmetros através de um mapa de cores facilita projetistas a escolha de parâmetros, indica onde aplicar melhorias nos circuitos, e quais são as mais relevantes. Equivalências entre circuitos Thevenin e Norton com os circuitos que conectam sinapses e neurônios foram apontadas, e a importância de estabelecer determinadas razões entre impedâncias foi explicitada, destacando condições para que o sistema neuromórfico atue dentro de uma região de mais alta eficiência energética. Ademais, na representação do circuito equivalente pré-sináptico, uma funcionalidade importante para as redes neurais foi indicada e que se destina ao

treinamento. Existe a necessidade de modulação de tensão para os circuitos conectados a sinapses, a fim de estabelecer condições da atualização da resistência dos memristores.

Dado as condições para garantir eficiência na transmissão de informação e treinamento local, o capítulo 4 apresenta opções de circuitos de interface, indicando o transportador de corrente de segunda geração como topologia de circuito mais adequada e versátil. Funcionalmente, a versatilidade do TCSG no emprego de redes neurais pode ser garantida devido a quatro pontos:

1) independentemente do tipo de neurônios artificiais, o terminal Z com características de fonte de corrente com alta impedância de saída garante o carregamento da capacitância da membrana e operação de mimetismo de canal iônico. Mecanismos de excitação inibitória ou excitatória também podem ser representados, pela polaridade do transportador de corrente.

2) independentemente dos requisitos para o chaveamento resistivo (quadrante de operação, tipo de modulação, ou regras de aprendizagem), o terminal X com características de fonte de tensão conectada à sinapse fornece capacidade de prover plasticidade, permitindo modulação de um sinal de tensão. A forma como é habilitada a geração do sinal de feedback pode definir a regra de aprendizado aplicada, a qual, para este trabalho, foi focado nas características de temporização entre pulsos pré e pós-sinápticos.

3) Os terminais X (baixa impedância) e Z (alta impedância) permitem que o TCSG se acople a neurônios e sinapses provendo uma faixa operacional energeticamente eficiente para transferência de informações.

4) A forma desacoplada do transporte de corrente entre sinapse e neurônio faz com que ambos funcionem de maneiras autônomas demonstrando uma versatilidade na escolha dos elementos que foram a rede.

No capítulo 5, uma demonstração de aprendizado associativo em uma implementação totalmente em hardware foi realizada, destacando todos os aspectos de projetos concebidos em uma plataforma para testes e caracterização de redes neurais. O dispositivo de memória (memristor de óxido de titânio) e o dispositivo de comutação do neurônio (transistor unijunção) foram utilizados para constituir uma rede neural que fornece aprendizado e processamento de inferência. Aspectos de projetos também foram abordados, apresentando um circuito gerador de pulso de feedback, que pode ser conectado a circuitos de neurônios para implementação de regra de

aprendizado STDP. Exclusivamente para este trabalho, foi dada ênfase para a dependência da temporização entre disparos para definir a plasticidade sináptica, porém, a flexibilidade dessa arquitetura abre precedentes para implementação de várias outras regras de aprendizados que se resumem ao projeto de circuitos geradores de pulsos de feedback.

6.2 RECOMENDAÇÃOES PARA CONTINUIDADE DO TRABALHO

As análises sobre o processamento em memória e o postulado de Shannon para transmissão de informações através de um canal ruidoso aponta para métricas de projeto que definem condições eficientes na implantação de redes neurais. Definir uma razão de impedâncias entre blocos conectados é importante para a maximização da eficiência, porém, após isso, as melhorias na eficiência estão atreladas a outros fatores, especialmente a redução de SNR. À medida que os níveis de potência dos pulsos ficam menores, um estudo mais aprofundado sobre os o ruído intrínseco dos diferentes tipos de sinapses precisa ser considerado. Os diferentes tipos de transporte de carga, para os diferentes tipos de sinapses memristivas, apontam a diferentes abordagens de ruído, e essas complexidades precisam ser incluídas nas equações 3.3 e 3.4 para que o cálculo da eficiência se apresente mais completo.

Em comparação com redes neurais implementadas fisicamente, o TCSG permitiu implementar mecanismos de inferência e treinamento, e não depender criticamente da combinação de componentes externos. O experimento prototípico do cão Pavlov demonstrado por meio de uma plataforma personalizada é uma boa exemplificação de como a arquitetura baseada em TCSG pode contribuir para implementações de redes neurais, agindo como validação de todos os circuitos implementados. Porém, o escalonamento para operação com matrizes maiores é a mais importante indicação de continuidade do trabalho a fim de habilitar a plataforma a resolução de problemas mais complexos, e mais próximos das necessidades computacionais de hoje. Essa plataforma foi inicialmente projetada para testes sobre *on-wafer*, e possui como recurso de conexão elétrica, o acesso por sondas discretas. Essas restrições limitam os testes a, no máximo, 4 sinapses em uma organização matricial 2x2. Para versões posteriores da plataforma, espera-se um projeto mais adaptado ao uso de *probecard* ou dispositivos encapsulados em soquetes.

Uma expectativa é apresentada no apêndice B. Embora versátil e útil em sua forma atual, a plataforma **SNN Testing Platform v1.5** pode ser ainda mais avançada, pela acomodação de operações de escrita e leitura mais complexas em matrizes de barra cruzadas. Uma característica imediata que pode ser implementada em versões futuras é a programação de ciclos de leituras com esquemas de redução da influência das correntes parasitas, como esses destacados em (GÜL, 2019; SHI et al., 2020). Dessa forma, seria possível usar a plataforma para programar qualquer dispositivo individualmente na matriz para valores específicos de resistência, permitindo programar a matriz memristivas para quaisquer condições iniciais. Recursos como esses, exigem que a plataforma adapte a amplitude dos pulsos a serem aplicados o funcionamento. Isso implica na necessidade de adicionar um DAC, ou versão de unidade microcontroladora que conste de um conversor digital-analógico embarcados, a fim de que a mudança das amplitudes dos pulsos possa ser realizada em tempo real, ou controlada pelo usuário através da interface.

Em vista da versatilidade do TCSG aplicado a redes neurais no domínio analógico, espera-se que esse circuito siga a tendência de dimensionamento presente em neurônios e sinapses, ampliando a capacidade de processamento exigida por aplicações de inteligência artificial. Uma abordagem para estudos futuros compreende a análise dos TCSG implementadas em circuitos integrados, onde questões sobre densidade e consumo energético são extremamente importantes. Sobre o consumo energético, sinapses memristivas e neurônios artificiais têm dissipação de energia apenas durante a geração ou propagação do pulso, porém, os TCSG são circuitos de consumo de energia estática. A potência CC dissipada no TCSG está essencialmente associada às correntes de polarização do amplificador diferencial, representadas por $Ibias1$ e $Ibias2$ na Fig. 4.4, elementos fundamentais na estrutura da FTCT. Técnicas de projeto de circuitos VLSI podem ser implementadas para reduzir a corrente de polarização, e extensamente discutidas em (FERRI; GUERRINI, 2003). Sobre densidade, outro tópico importante relacionado à otimização do projeto de redes neurais diz respeito à modificação do estágio de saída dos transportadores de corrente, principalmente sobre o circuito da FCCC. Reduzindo proporcionalmente a corrente sináptica que alimentará os neurônios, propicia a redução da área usada pelos capacitores de membrana, provendo artifícios de otimização em área. A modificação de espelhos de corrente, como em (WIDLAR, 1967), pode reduzir os níveis de corrente “espelhados” por uma degeneração da tensão de polarização dos

transistores. A diminuição da corrente, pelo uso de apenas alguns componentes adicionais, tem como desvantagem a distorção da corrente. Outra abordagem da redução da corrente simpática é implementada usando uma sequência de divisores de corrente e atenuadores de corrente, como em (AHMADI-FARSANI; LINARES-BARRANCO; SERRANO-GOTARREDONA, 2020; MOHAN et al., 2019). A corrente é discretamente dividida por circuitos divisores, ao preço de complexidade e número de componentes adicionais.

A implementação escalar de redes neurais leva a sinapses e neurônios artificiais representadas por dispositivo únicos e extremamente pequenos, para compor redes neurais extremamente densas (KENDALL; KUMAR, 2020; KUMAR; WILLIAMS; WANG, 2020). Comparativamente, os transportadores de corrente são blocos mais complexos, compostos por partes como FCCC e FTCT, que os tornam mais difíceis de seguir essa abordagem de miniaturização. Os transistores bipolares de junção (TBJ) podem representar FCCC com um único dispositivo. Os TBJs são dispositivos ativos nos quais um sinal de corrente é transferido em corrente na saída, tornando-os adaptáveis a compor FCCCs, porém, questões como baixa impedância de saída restringem a integração deste tipo de dispositivo à implementação de TCSG. Em relação às FTCTs representados por um único dispositivo, trata-se de dispositivos ativos ausente na literatura. Uma solução possível solução pode ser a investigação de circuitos baseados em trancitores(LEE, 2018), uma base não convencional de eletrônica como transistores, que compõem um novo paradigma da eletrônica pela proposição de dispositivos ativos que operam como FTCT. Porém, dispositivos trancitores fisicamente implementados ainda não estão disponíveis. Outra estratégia para implementação de TCSG em pequenas dimensões são os circuitos baseados na tecnologia CiFET (terminologia do inglês *Carrier Injection Field Effect Transistor*)(SUGAWARA et al., 1995). Esses dispositivos podem habilitar entradas de tensão ou corrente e produzir saída de tensão ou corrente e são totalmente compatíveis com a tecnologia CMOS (SCHOBER; SCHOBER; HUDRLIK, 2019). Essencialmente, qualquer circuito digital ou analógico pode ser construído usando CiFETs, tornando essa tecnologia um candidato promissor para integrar transportadores de corrente de segunda geração em redes neurais densas.

REFERÊNCIAS

AHMADI-FARSANI, J.; LINARES-BARRANCO, B.; SERRANO-GOTARREDONA, T. **A Current-Attenuator for Performing Read Operation in Memristor-Based Spiking Neural Networks**. 2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS). **Anais**...2020.

ALIBART, F. et al. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. **Nanotechnology**, v. 23, n. 7, p. 75201, 2012.

ALIBART, F.; ZAMANIDOOST, E.; STRUKOV, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. **Nature communications**, v. 4, n. 1, p. 1–7, 2013.

ASKEW, W. J. **Unijunction transistor artificial neuron**Google Patents, , 1972.

AUGE, D. et al. A survey of encoding techniques for signal processing in spiking neural networks. **Neural Processing Letters**, v. 53, n. 6, p. 4693–4710, 2021.

AVELINO, W. et al. **Opportunities for transition metal oxide devices in solid state random number generators**. Low-Dimensional Materials and Devices 2019. **Anais**...2019.

BACKUS, J. Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. **Communications of the ACM**, v. 21, n. 8, p. 613–641, 1978.

BEILLIARD, Y. et al. Conductive filament evolution dynamics revealed by cryogenic (1.5 K) multilevel switching of CMOS-compatible Al2O3/TiO2 resistive memories. **Nanotechnology**, v. 31, n. 44, p. 445205, 2020.

BELL, C. **Beginning sensor networks with Arduino and Raspberry Pi**. [s.l.] Apress, 2014.

BOAHEN, K. A neuromorph's prospectus. **Computing in Science & Engineering**, v. 19, n. 2, p. 14–28, 2017.

BOQUET, G. et al. **Offline training for memristor-based neural networks**. 2020 28th European Signal Processing Conference (EUSIPCO). **Anais**...2021.

BORKAR, S. Role of interconnects in the future of computing. **Journal of Lightwave Technology**, v. 31, n. 24, p. 3927–3933, 2013.

BRUNELLI, M.; CASTELLUCCI, V.; KANDEL, E. R. Synaptic facilitation and behavioral sensitization in Aplysia: possible role of serotonin and cyclic AMP. **Science**, v. 194, n. 4270, p. 1178–1181, 1976.

CAI, F. et al. A fully integrated reprogrammable memristor--CMOS system for efficient multiply--accumulate operations. **Nature Electronics**, v. 2, n. 7, p. 290–299, 2019.

CHEN, J. et al. Multiply accumulate operations in memristor crossbar arrays for analog computing. **Journal of Semiconductors**, v. 42, n. 1, p. 13104, 2021.

CHUA, L. Memristor-the missing circuit element. **IEEE Transactions on circuit theory**, v. 18, n. 5, p. 507–519, 1971.

CHUA, L. O.; KANG, S. M. Memristive devices and systems. **Proceedings of the IEEE**, v. 64, n. 2, p. 209–223, 1976.

COOK, G. **Clicking Clean: Who is winning the race to build a green internet**. [s.l: s.n.].

DENG, L. et al. Energy consumption analysis for various memristive networks under different learning strategies. **Physics Letters A**, v. 380, n. 7–8, p. 903–909, 2016.

DENG, L. et al. Rethinking the performance comparison between SNNS and ANNS. **Neural networks**, v. 121, p. 294–307, 2020.

DU, Y. et al. An analog neural network computing engine using CMOS-compatible charge-trap-transistor (CTT). **IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems**, v. 38, n. 10, p. 1811–1819, 2018.

FARHAN, L. et al. **A survey on the challenges and opportunities of the Internet of Things (IoT)**. 2017 Eleventh International Conference on Sensing Technology (ICST). **Anais**...2017.

FERRI, G.; GUERRINI, N. C. **Low-voltage low-power CMOS current conveyors**. [s.l.] Springer Science & Business Media, 2003.

FITZHUGH, R. Impulses and physiological states in theoretical models of nerve membrane. **Biophysical journal**, v. 1, n. 6, p. 445–466, 1961.

FONSECA, A.; KAZMAN, R.; LAGO, P. A manifesto for energy-aware software. **IEEE Software**, v. 36, n. 6, p. 79–82, 2019.

GARBIN, D. et al. HfO 2-based OxRAM devices as synapses for convolutional neural networks. **IEEE Transactions on Electron Devices**, v. 62, n. 8, p. 2494–2501, 2015.

GERSTNER, W.; KISTLER, W. M. **Spiking neuron models: Single neurons, populations, plasticity**. [s.l.] Cambridge university press, 2002.

GÜL, F. Addressing the sneak-path problem in crossbar RRAM devices using memristor-based one Schottky diode-one resistor array. **Results in Physics**, v. 12, p. 1091–1096, 2019.

HACHTEL, G.; PEDERSON, D. O. **Integrated, unijunction transistor oscillators**. 1962 International Electron Devices Meeting. **Anais**...1962.

HEBB, D. O. **The organisation of behaviour: a neuropsychological theory**. [s.l.] Science Editions New York, 1949.

HEMSOTH, N. **A Mythic Approach to Deep Learning Inference**. [s.l: s.n.].

HILL, M. D.; MARTY, M. R. Amdahl's law in the multicore era. **Computer**, v. 41, n. 7, p. 33–38, 2008.

HODGKIN, A. L.; HUXLEY, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. **The Journal of physiology**, v. 117, n. 4, p. 500, 1952.

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. **Proceedings of the national academy of sciences**, v. 79, n. 8, p. 2554–2558, 1982.

HOROWITZ, M. **1.1 computing's energy problem (and what we can do about it)**. 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). **Anais**...2014.

HOWARD, D.; BULL, L.; DE LACY COSTELLO, B. Evolving unipolar memristor spiking neural networks. **Connection Science**, v. 27, n. 4, p. 397–416, 2015.

IELMINI, D.; WONG, H.-S. P. In-memory computing with resistive switching devices. **Nature Electronics**, v. 1, n. 6, p. 333–343, 2018.

ISHII, M. et al. **On-chip trainable 1.4 M 6T2R PCM synaptic array with 1.6 K stochastic LIF neurons for spiking RBM**. 2019 IEEE International Electron Devices Meeting (IEDM). **Anais**...2019.

IZHIKEVICH, E. M. Simple model of spiking neurons. **IEEE Transactions on neural networks**, v. 14, n. 6, p. 1569–1572, 2003.

JAMES, A. P.; CHUA, L. O. Analog neural computing with super-resolution memristor crossbars. **IEEE Transactions on Circuits and Systems I: Regular Papers**, v. 68, n. 11, p. 4470–4481, 2021.

JO, S. H. et al. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. **Nano Letters**, v. 10, n. 4, p. 1297–1301, 14 abr. 2010.

JOHNSTON, D.; WU, S. M.-S. **Foundations of cellular neurophysiology**. [s.l.] MIT press, 1994.

KAMIYA, G.; KVARNSTRÖM, O. Data centres and energy--from global headlines to local headaches. **International Energy Agency https://www. iea. org/commentaries/data-centres-and-energy-from-global-headlines-to-local-headaches**, 2019.

KENDALL, J. D.; KUMAR, S. The building blocks of a brain-inspired computer. **Applied Physics Reviews**, v. 7, n. 1, p. 11305, 2020.

KHAN, M. A. Study of Magnetization Switching in Coupled Magnetic Nanostructured Systems using a Tunnel Diode Oscillator. 2018.

KIDNER, C. et al. **Potential and limitations of resonant tunneling diodes**. Proc. of the First International Symposium on Space Terahertz Technology. **Anais**...1990.

KIM, H. et al. 4K-memristor analog-grade passive crossbar circuit. **Nature communications**, v. 12, n. 1, p. 1–11, 2021.

KIM, J. et al. Efficient synapse memory structure for reconfigurable digital neuromorphic hardware. **Frontiers in neuroscience**, p. 829, 2018.

KOCH, C. **Biophysics of computation: information processing in single neurons**. [s.l.] Oxford university press, 2004.

KOCH, C.; POGGIO, T. Biophysics of computation: neurons, synapses and membranes. 1984.

KOOT, M.; WIJNHOVEN, F. Usage impact on data center electricity needs: A system dynamic forecasting model. **Applied Energy**, v. 291, p. 116798, 2021.

KUMAR, A. A. **Pulse and Digital Circuits**. [s.l.] PHI Learning Pvt. Ltd., 2008.

KUMAR, S.; WILLIAMS, R. S.; WANG, Z. Third-order nanocircuit elements for neuromorphic engineering. **Nature**, v. 585, n. 7826, p. 518–523, 2020.

KURASHINA, T.; OGAWA, S.; WATANABE, K. **A high performance class AB current conveyor**. 1998 IEEE International Conference on Electronics, Circuits and Systems. Surfing the Waves of Science and Technology (Cat. No. 98EX196). **Anais**...1998.

LECERF, G.; TOMAS, J.; SA\"\IGHI, S. **Excitatory and inhibitory memristive synapses for spiking neural networks**. 2013 IEEE International Symposium on Circuits and Systems (ISCAS). **Anais**...2013.

LEE, D.; CHEN, Y.-T.; CHAO, S.-L. Universal workflow of artificial intelligence for energy saving. **Energy Reports**, v. 8, p. 1602–1633, 2022.

LEE, S. A Missing Active Device--Trancitor for a New Paradigm of Electronics. **IEEE Access**, v. 6, p. 46962–46967, 2018.

LIU, S.-C. et al. Homeostasis in a silicon integrate and fire neuron. **Advances in Neural Information Processing Systems**, p. 727–733, 2001.

LOZNEANU, E.; POPESCU, V.; SANDULOVICIU, M. Negative differential resistance related to self-organization phenomena in a dc gas discharge. **Journal of applied physics**, v. 92, n. 3, p. 1195–1199, 2002.

MAASS, W. Networks of spiking neurons: the third generation of neural network models. **Neural networks**, v. 10, n. 9, p. 1659–1671, 1997.

MALMODIN, J.; LUNDÉN, D. **The electricity consumption and operational carbon emissions of ICT network operators 2010-2015**, 2018.

MCCULLOCK, W. S.; PITTS, W. A logical calculus of ideas immanent in nervous activity. archive copy of 27 november 2007 on wayback machine. **Avtomaty [Automated Devices] Moscow, Inostr. Lit. publ**, p. 363–384, 1956.

MEAD, C. Neuromorphic electronic systems. **Proceedings of the IEEE**, v. 78, n. 10, p. 1629–1636, 1990.

MEAD, C. A.; MAHOWALD, M. A. A silicon model of early visual processing. **Neural networks**, v. 1, n. 1, p. 91–97, 1988.

MEAD, C.; ISMAIL, M. **Analog VLSI implementation of neural systems**. [s.l.] Springer Science & Business Media, 1989. v. 80

MERKLE, R. C. Energy limits to the computational power of the human brain. 2007.

MEROLLA, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. **Science**, v. 345, n. 6197, p. 668–673, 2014.

MESOUDY, A. EL et al. Fully CMOS-compatible passive TiO2-based memristor crossbars for in-memory computing. **arXiv preprint arXiv:2106.11808**, 2021.

MIKKELSEN, J. H. Ltspice--an introduction. **Technical report, Institute of Electronic Systems, Aalborg University, Aalborg**, 2005.

MOHAN, C. et al. **A current attenuator for efficient memristive crossbars read-out**. 2019 IEEE International Symposium on Circuits and Systems (ISCAS). **Anais**...2019.

MUÑOZ-MARTIN, I. et al. **A SiOx RRAM-based hardware with spike frequency adaptation for power-saving continual learning in convolutional neural networks**. 2020 IEEE Symposium on VLSI Technology. **Anais**...2020.

NAGUMO, J.; ARIMOTO, S.; YOSHIZAWA, S. An active pulse transmission line simulating nerve axon. **Proceedings of the IRE**, v. 50, n. 10, p. 2061–2070, 1962.

NAYYAR, A.; PURI, V. **A review of Arduino board's, Lilypad's & Arduino shields**. 2016 3rd international conference on computing for sustainable global development (INDIACom). **Anais**...2016.

PEDRETTI, G. et al. Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity. **Scientific reports**, v. 7, n. 1, p. 1–10, 2017.

PHAM, D. T. Neural networks in engineering. **WIT Transactions on Information and Communication Technologies**, v. 6, 1970.

PICKETT, M. D. et al. Switching dynamics in titanium dioxide memristive devices. **Journal of Applied Physics**, v. 106, n. 7, p. 74508, 2009.

PICKETT, M. D.; MEDEIROS-RIBEIRO, G.; WILLIAMS, R. S. A scalable neuristor built with Mott memristors. **Nature Materials**, v. 12, n. 2, p. 114–117, 2013.

PREZIOSO, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. **Nature**, v. 521, n. 7550, p. 61–64, 2015.

QUERLIOZ, D. et al. Immunity to device variations in a spiking neural network with memristive nanodevices. **IEEE Transactions on Nanotechnology**, v. 12, n. 3, p. 288–295, 2013.

RAMAKRISHNAN, S. et al. Neuron array with plastic synapses and programmable dendrites. **IEEE transactions on biomedical circuits and systems**, v. 7, n. 5, p. 631–642, 2013.

ROOKS, T. **Big data centers are power-hungry, but increasingly efficient**, 2022. Disponível em: <https://p.dw.com/p/45cO4>

ROZENBERG, M. J.; SCHNEEGANS, O.; STOLIAR, P. An ultra-compact leaky-integrate-and-fire model for building spiking neural networks. **Scientific reports**, v. 9, n. 1, p. 11123, 31 jul. 2019.

SCHOBER, S. M.; SCHOBER, R. C.; HUDRLIK, T. R. **Low noise sensor amplifiers and trans-impedance amplifiers using complementary pair of current injection field-effect transistor devices**Google Patents, , 2019.

SEDRA, A.; SMITH, K. A second-generation current conveyor and its applications. **IEEE Transactions on Circuit Theory**, v. 17, p. 132–134, 1970.

SERRANO-GOTARREDONA, T. et al. STDP and STDP variations with memristors for spiking neuromorphic learning systems. **Frontiers in neuroscience**, v. 7, p. 2, 2013.

SHANNON, C. E. A mathematical theory of communication. **The Bell system technical journal**, v. 27, n. 3, p. 379–423, 1948.

SHARPESHKAR, R. **Ultra low power bioelectronics: Fundamentals, biomedical applications, and bio-inspired systemCambridge University Press**, 2010.

SHI, L. et al. Research progress on solutions to the sneak path issue in memristor crossbar arrays. **Nanoscale Advances**, v. 2, n. 5, p. 1811–1827, 2020.

SHUKLA, M.; TRIPATHI, B. K. Second Generation Neural Network for Two Dimensional Problems. **International Journal of Hybrid Information Technology**, v. 9, n. 11, p. 47–56, 2016.

SIVAKUMAR, V.; MALATHI, M. **Programmable synaptic memory with spiking neural network in VLSI**. International Conference on Information Communication and Embedded Systems (ICICES2014). **Anais**...2014.

SMITH, K. C.; SEDRA, A. The current conveyor—A new circuit building block. **Proceedings of the IEEE**, v. 56, n. 8, p. 1368–1369, 1968.

STEIN, P. S. G. et al. **Neurons, networks, and motor behavior**. [s.l.] MIT press, 1999.

STRUKOV, D. B. et al. The missing memristor found. **nature**, v. 453, n. 7191, p. 80–83, 2008.

SUGAWARA, Y. et al. **350 V carrier injection field effect transistor (CIFET) with very low on-resistance and high switching speed**. Proceedings of International Symposium on Power Semiconductor Devices and IC's: ISPSD'95. **Anais**...1995.

SURI, M. et al. **Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction**. 2011 International Electron Devices Meeting. **Anais**...2011.

SZE, V. et al. **Hardware for machine learning: Challenges and opportunities**. 2017 IEEE Custom Integrated Circuits Conference (CICC). **Anais**...2017.

TOUMAZOU, C.; LIDGEY, F. J.; HAIGH, D. **Analogue IC design: the current-mode approach**. [s.l.] Presbyterian Publishing Corp, 1990. v. 2

TSUR, E. E. **Neuromorphic Engineering: The Scientist's, Algorithm Designer's, and Computer Architect's Perspectives on Brain-Inspired Computing**. [s.l.] CRC Press, 2021.

VALOV, I. Interfacial interactions and their impact on redox-based resistive switching memories (ReRAMs). **Semiconductor Science and Technology**, v. 32, n. 9, p. 93006, 2017.

VINCENT, A. F. et al. Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems. **IEEE transactions on biomedical circuits and systems**, v. 9, n. 2, p. 166–174, 2015.

WANG, R. et al. Bipolar Analog Memristors as Artificial Synapses for Neuromorphic Computing. **Materials (Basel, Switzerland)**, v. 11, n. 11, p. 2102, 26 out. 2018.

WANG, Z. et al. A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems. **Frontiers in neuroscience**, v. 8, p. 438, 2015.

WANG, Z. et al. In situ training of feed-forward and recurrent convolutional memristor networks. **Nature Machine Intelligence**, v. 1, n. 9, p. 434–442, 2019.

WIDLAR, R. J. Low-value current source for integrated circuits. **US Patent**, n. 03320439, 1967.

WU, J. et al. Current-controlled negative differential resistance in small-polaron hopping system. **AIP Advances**, v. 9, n. 5, p. 55223, 2019.

WU, X. et al. A CMOS spiking neuron for brain-inspired neural networks with resistive synapses and in situ learning. **IEEE Transactions on Circuits and Systems II: Express Briefs**, v. 62, n. 11, p. 1088–1092, 2015.

YI, W. et al. Biological plausibility and stochasticity in scalable VO 2 active memristor neurons. **Nature communications**, v. 9, n. 1, p. 1–10, 2018.

YILMAZOGLU, O. et al. Measured negative differential resistivity for GaN Gunn diodes on GaN substrate. **Electronics Letters**, v. 43, n. 8, p. 480–482, 2007.

ZANGENEH, M.; JOSHI, A. **Performance and energy models for memristor-based 1T1R RRAM cell**. Proceedings of the great lakes symposium on VLSI. **Anais**...2012.

ZHANG, W. et al. Neuro-inspired computing chips. **Nature electronics**, v. 3, n. 7, p. 371–382, 2020.

ZIDAN, M. A. et al. Memristor-based memory: The sneak paths problem and solutions. **Microelectronics journal**, v. 44, n. 2, p. 176–183, 2013.

ZOMAYA, A. Y. **Handbook of nature-inspired and innovative computing: integrating classical models with emerging technologies**. [s.l.] Springer Science & Business Media, 2006.

APÊNDICE A

TRABALHOS DESENVOLVIDOS DURANTE A TESE

O projeto de doutorado desenvolveu-se durante minha experiência como profissional da pesquisa e uma série de trabalhos foram gerados paralelamente. Todos contribuíram para a formação do conhecimento técnico e o desenvolvimento dos resultados alcançados até aqui. Entre eles, é possível citar:

Trabalhos divulgados em revista

- ✓ Ohlberg, D. A., Tami, D., Gadelha, A. C., **Avelino, W. O.**, Santana, F. C., Miranda, D., ... & Medeiros-Ribeiro, G. (2021). The limits of near field immersion microwave microscopy evaluated by imaging bilayer graphene moiré patterns. *Nature communications*, *12*(1), 1-6.

Apresentações em congressos

- ✓ AVELINO, Wellington O. et al. Demonstration of Memristor-based Spiking Neural Network. In: **Memrisys 2021 – 4th International Conference on Memristive Materials, Devices and Systems**. November 2021.
- ✓ AVELINO, Wellington O. et al. Near-field microwave microscopy and processing of nanostructured materials. In: **Low-Dimensional Materials and Devices 2020**. SPIE, 2020. p. 114650K.
- ✓ AVELINO, Wellington et al. Opportunities for transition metal oxide devices in solid state random number generators. In: **Low-Dimensional Materials and Devices 2019**. International Society for Optics and Photonics, 2019. p. 110850L.

Projetos de Pesquisa e Desenvolvimento

- ✓ Estágio de Pesquisa (3iT/Université de Sherbrooke): As atividades se concentraram no desenvolvimento de circuitos neuromórficos e no estudo de aplicações de aprendizado de máquina embarcadas em hardware. Outras tarefas incluíram caracterizações elétricas e estudos de confiabilidade de dispositivos de memória usando instrumentação comercial e plataformas de teste especialmente customizadas.
- ✓ Projeto Plataforma KeyBits para telecomunicação segura: As atividades concentraram-se no desenvolvimento de um gerador de números puramente aleatórios para compor sistemas criptográficos. Chaves criptográficas eram geradas, a partir de fontes físicas de ruido, em especial à flutuação de fótons quanto à polarização de um laser. Porém, durante esse período, o aprendizado sobre a utilização de fontes físicas de ruído despertou a atenção utilização do memristor como elemento gerador de ruído elétrico.

Sob específicas condições de polarização, o memristor pode ser uma fonte específica escalável e de baixo consumo para sistemas criptográficos. Parte dos resultados experimentais foram divulgados em congressos(AVELINO et al., 2019).

- ✓ Projeto Microwave Energy Deliver: Trata-se de um projeto de colaboração HPe e UFMG para desenvolvimento de ferramentas para transporte de energia em frequência de 2,4GHz. Durante esse período, atividades como projeto de circuitos de micro-ondas, testes e caracterização de óxidos metálicos, e análise de dados de caracterização foram realizadas. O desenvolvimento dessas atividades serviu como formação para o desenvolvimento de circuitos de alta banda passante, utilizado posteriormente em circuitos de driver para memristores.

- ✓ Projeto Memristor: Projeto de colaboração entre HP Labs e CTI Renato Archer para caracterização de dispositivos memristivos. Esse projeto teve duração de 3,5 anos, e foi inspirador a direcionar a meu campo de pesquisa para dispositivos memristivos e topologias de circuito que o utilizam. Durante esse período, o objetivo era realizar caracterização elétrica em dispositivos memristores de óxido de tântalo utilizando instrumentação comercial. Para caracterização de memristores em matrizes, as funcionalidades disponíveis na instrumentação não se mostravam suficientes, e por isso desenvolvi uma série de instrumentos que fazem testes específicos de caracterização em memristores.

APÊNDICE B

PLATAFORMA DE TESTES PARA REDES NEURAIS

As redes neurais por pulsos são estruturas computacionais muito promissoras para aplicações de inteligência artificial, e novas topologias de neurônios artificiais e dispositivos de sinapses estão sendo constantemente testados. No entanto, uma grande diversidade de elementos de circuito é proposta para compor as unidades primitivas das redes, e testar combinações dessas partes torna-se um gargalo para a análise de implementação. Uma plataforma de teste compacta e versátil foi projetada para fornecer versatilidade de integração entre diferentes combinações de neurônios e sinapses compondo uma rede fisicamente implementada. ***SNN Testing Platform v1.5*** é um sistema de teste/caracterização integrado a uma interface computacional que realiza a emulação de pequenas redes neurais que possuem sinapse artificial como dispositivos memristivos.

O projeto desta plataforma é baseado em estrutura modular que acomoda um escopo de recursos:

- 2 pré-neurônios que são controlados por unidade microcontroladora que funciona como interface sensorial.
- 2 transportadores de corrente de segunda geração fazem a interface neurônio-sinapse e fornecem condições para aplicação de sinais de retroativos de plasticidade.
- 2 neurônios LIF baseados em transistores de unijunção.
- 1 unidade de controle representada por um Arduino que faz a transferência de dados para o computador.
- 1 Interface de programação de testes (API) baseada em LabVIEW para programação de protocolos de teste automático e emulação de sessões de treinamento.

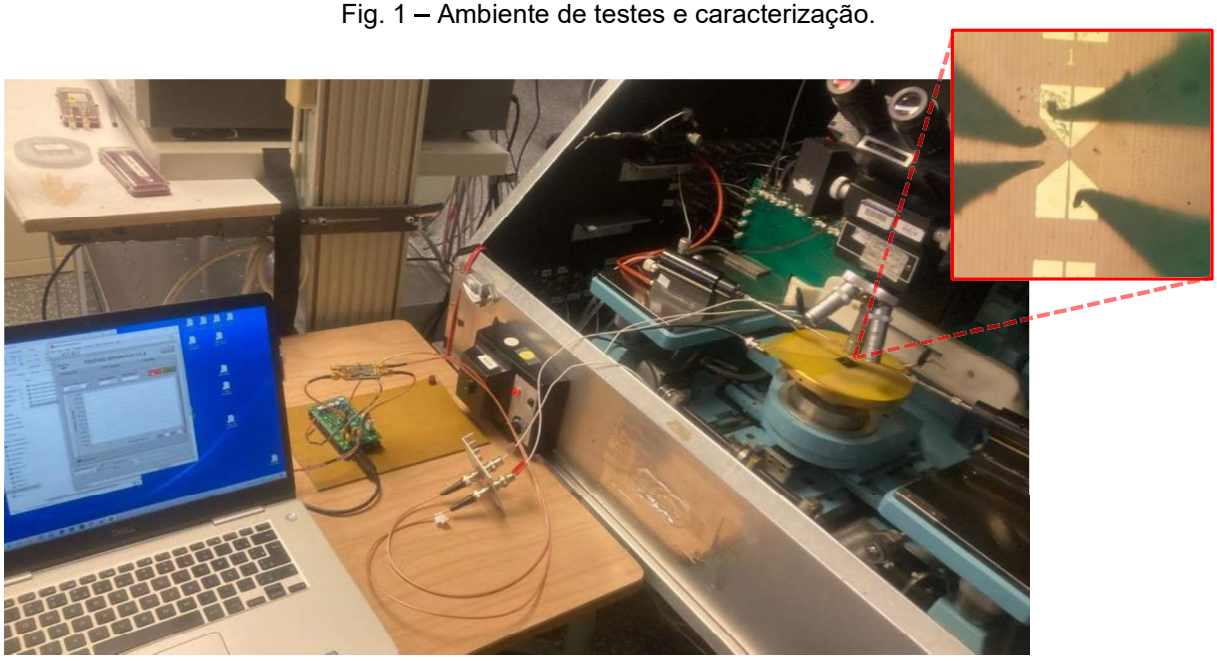
Essa infraestrutura permite flexibilidade na aplicação de testes e emulação de aprendizado compatíveis para diversos tipos de sinapses, tornando-se uma ferramenta importante na investigação de RNPs implementadas fisicamente, além de fazer parte de uma importante ferramenta para teste e validação de uma tecnologia de dispositivo de memória. A *SNN Testing Platform v1.5* possui as especificações técnicas conforme apresentado na tabela 1.

Tabela 1 - Especificações da plataforma

Especificação	**Faixa de operação**
Máxima tensão do pulso	-4V a +4V
Tensão de leitura	-4V a +4V
Largura de pulso	300ns
Mínimo intervalor entre pulsos	1µs (com passos de 1µs)
Tempo dos ciclos de leitura	500µs
Faixa de medida de resistência	10^{-2} a 10^{-5} A (±3% de erro)
Matriz sináptica	4 sinapses, no máximo 2x2
Dimensões	101,6mm x 53,35mm
Consumo de potência	3W

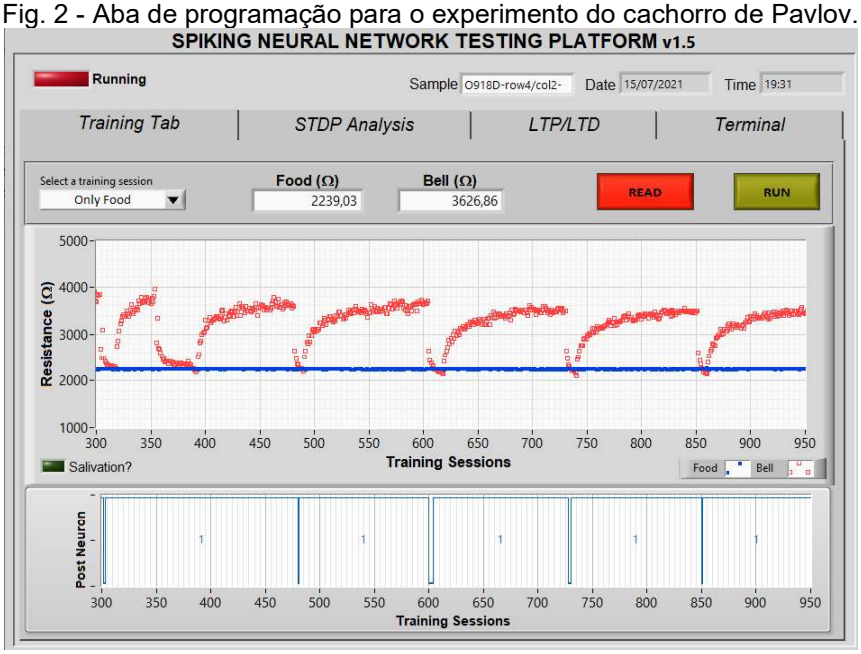
O acesso às sinapses é feito através de sondas discretas diretamente conectadas aos eletrodos dos memristores. Como mostra a Fig. 1.

Fig. 1 – Ambiente de testes e caracterização.



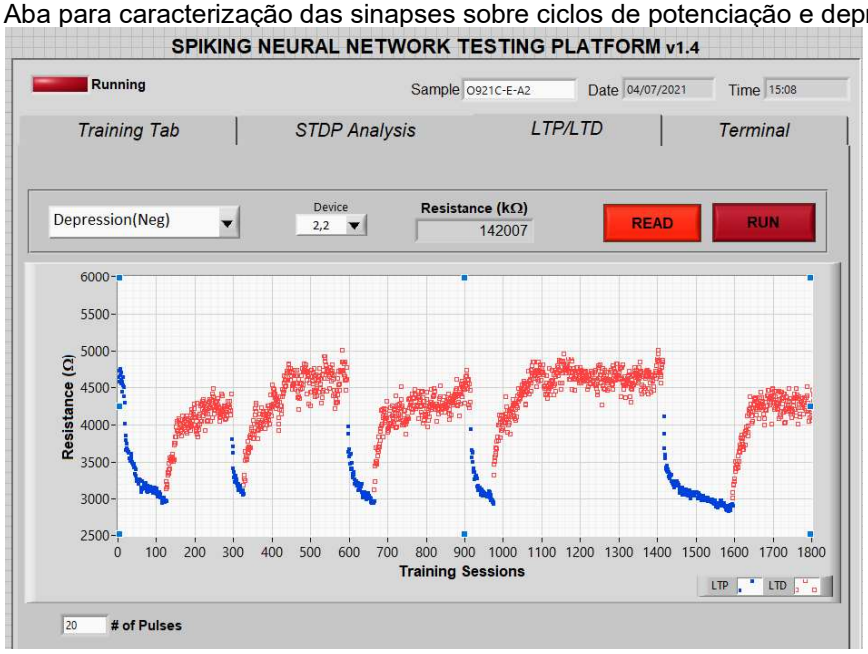
A interface gráfica usada para programar e apresentar as medidas elétricas realizadas pela plataforma foi desenvolvida em LabVIEW. Funcionalidades são divididas em sessões que possuem conjunto de parâmetros e gráficos para os seguintes experimentos: 1) Experimento do cachorro de Pavlov, 2) Análise de plasticidade STDP, 3) Protocolo de testes para potenciação e depressão de longo prazo, e 4) Terminal para depuração. As etapas de treino para a emulação do experimento do cachorro de Pavlov, cujos dados experimentais foram apresentados no capítulo V, foram programadas na interface, ver Fig. 2.

Fig. 2 - Aba de programação para o experimento do cachorro de Pavlov.



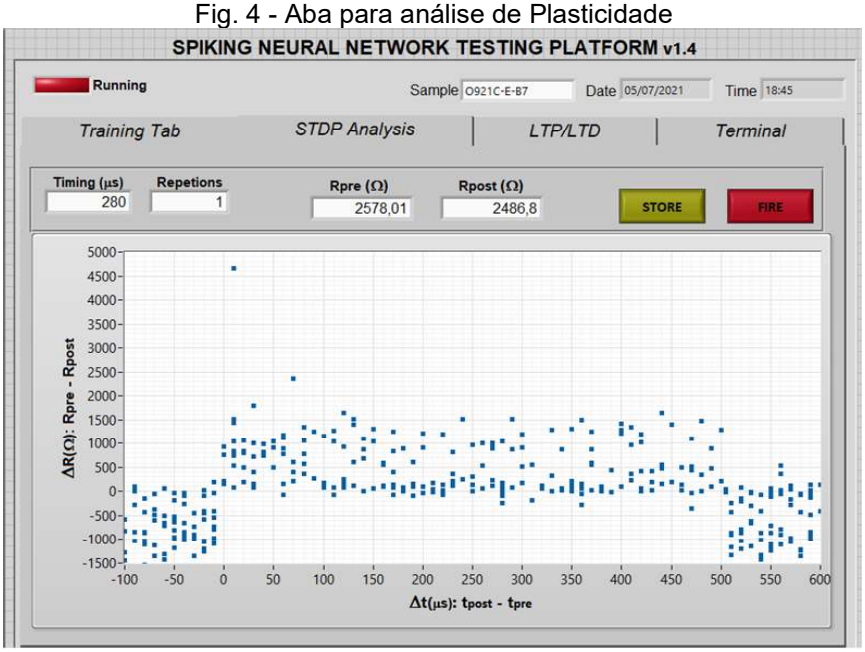
A caracterização da sinapse também pode ser feita através da plataforma, pela aplicação de pulsos de tensão, emulando ciclos de potenciação e depressão, como mostrado na Fig. 3.

Fig. 3 - Aba para caracterização das sinapses sobre ciclos de potenciação e depressão.



A plasticidade sináptica também é analisada pela emulação das condições de potenciação e depressão sobre o dispositivo de memória conectado a plataforma. Na aba de análise STDP, é possível programar a temporização entre os pulsos, e simulação o encontro dos pulsos pré-sinápticos e pulsos de feedback, ver Fig. 4.

Fig. 4 - Aba para análise de Plasticidade

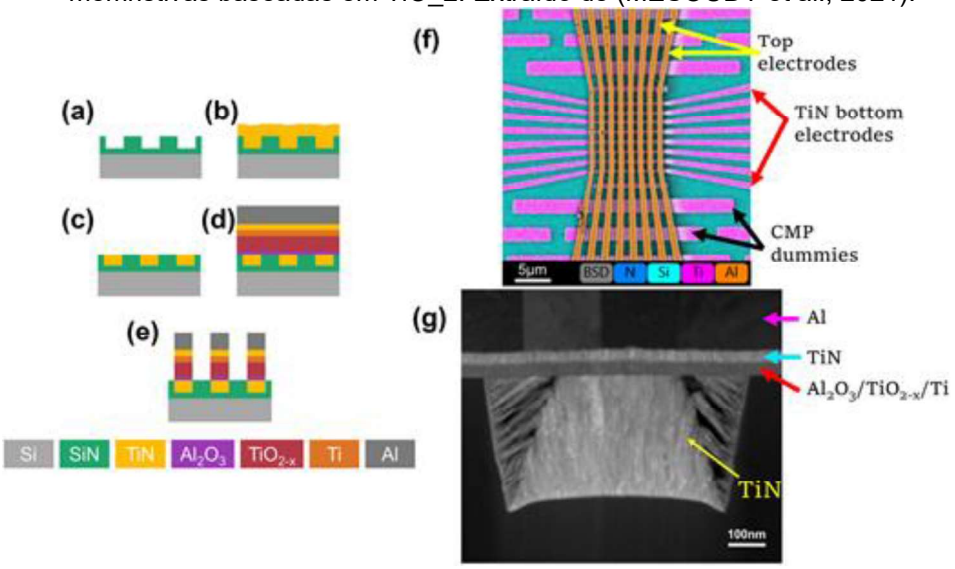


APÊNDICE C

FABRICAÇÃO DE MEMRISTORES DE ÓXIDO DE TITÂNIO

Os resultados experimentais discutidos nesse trabalho foram extraídos a partir de memristores de óxido de titânio. O processo de fabricação para matrizes em barra cruzada é apresentado na Fig. 5.

Fig. 5 - Fluxo de processo e caracterizações morfológicas de matrizes de barras cruzadas memristivas baseadas em TiO_2. Extraído de (MESOUDY et al., 2021).



Amostras de silício de $22 \times 22\ mm^2$ são cobertas por uma camada de nitreto de silício (SiN) de $600\ nm$ de espessura, depositada pela técnica de deposição por vapor químico melhorada por plasma (PECVD). Os eletrodos inferiores ($700nm$ de espessura e $1{,}4\ \mu m$ de espaçamento) de Nitreto de titânio foram fabricados usando 4 diferentes passos.

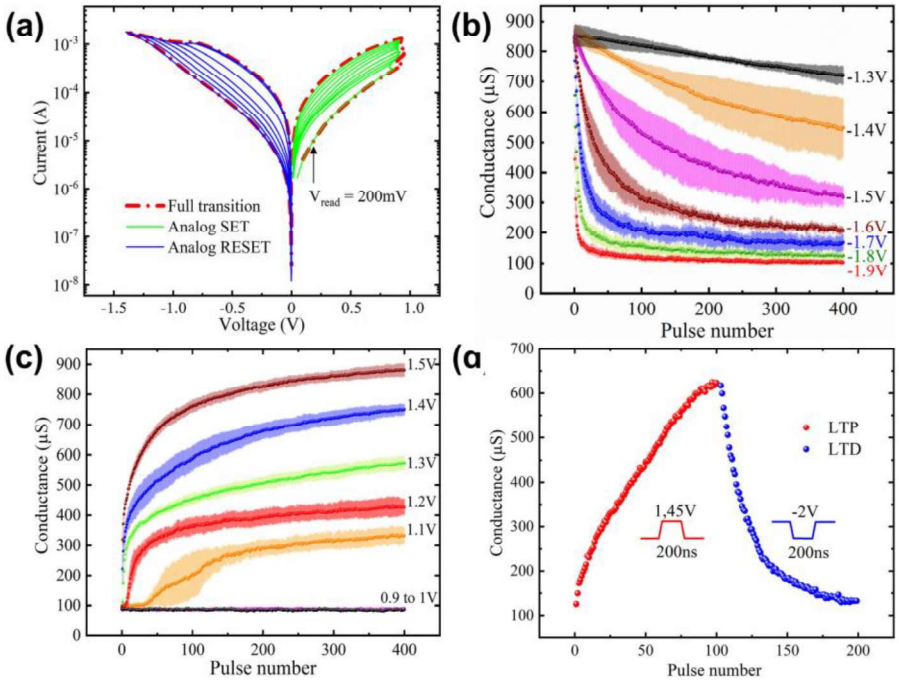
- Litografia por feixe de elétrons realizado com RAITH150-Two.
- Os padrões foram transferidos a camada de SiN através de corrosão por plasma usando CF_4, H_2 e He com uma proporção de $140/12/14$ sccm e potência de $100/50W$ para bobina e placa respectivamente (Fig. 5-a). O tempo de corrosão do calibrado para obter trincheiras de $400nm$ de profundidade.
- Deposição de $600nm$ de TiN por pulverização reativa (reactive sputtering), exibindo uma resistividade de $160u\Omega.cm$ (Fig. 5-b)
- Remoção do excesso de TiN e planarização através de Polimento mecânico químico (CMP terminologia do inglês *Chemical Mechanical Polishing*) (Fig. 5-c).
- A fim de reduzir as correntes de vazamento nos dispositivos, uma camada de $1{,}4nm$ de Al_2O_3 foi depositado por deposição de camada atômica melhorada por plasma (plasma-

enhanced ALD) com sistema Picosun R-200 advanced. O trimetilalumínio (TMA) foi usado como precursor.

- Para a região de chaveamento, uma camada sub estequiométrica de TiO_{2-x} foi depositada através de pulverização usando Plasmonique SPT320 sputtering tool.
- O eletrodo superior é composto de Ti/TiN, foram depositadas camadas no mesmo ambiente de pulverização do passo anterior. O titânio é usado como reservatório a fim de gerar vacâncias de oxigênio enquanto a camada de nitreto atua como eletrodo inerte.
- Posteriormente, com o intuito de diminuir a resistência dos eletrodos superiores, uma camada de $400nm$ de Al foi depositada por evaporação.

Os dispositivos memristivos fabricados são caracterizados para evidência de sua funcionalização. Um comportamento característico de "histerese comprimida" pode ser observado nas curvas IV, ver Fig. 6-a, evidenciando a característica bipolar do dispositivo, limitada aos quadrantes I e III. Para tensão positiva aplicada, a condutância do dispositivo aumenta gradualmente e enquanto para uma tensão negativa, a condutância diminui gradualmente. A corrente representada no gráfico da Fig. 6-a é o modulo do valor obtido nas medidas, para permitir a apresentação do perfil IV em uma escala logarítmica. Múltiplas varreduras quase-estáticas positivas ou negativas são aplicadas, enquanto observa-se a condutância do dispositivo continua a aumentar, exibindo sobreposições entre os loops de histerese, confirmando que a resistência do dispositivo possui caráter não-volátil. O incremento ou decremento da resistência do dispositivo apresenta valores de resistência máximo (HRS terminologia do inglês *High Resistance State*) e mínimo (LRS terminologia do inglês *Low Resistance State*).

Fig. 6 - Caracterização elétrica para os dispositivos construídos. Extraído de (MESOUDY et al., 2021).



O condicionamento gradual da resistência também pode ser realizado através de excitação por pulsos de tensão. Os resultados dos testes pulsados são apresentados nas Fig. 6-b e Fig. 6-c, onde o incremento de resistência, comumente associado a sistemas neuromórficos com depressão de longo prazo, e o decremento de resistência, associado a potenciação de longo prazo, respectivamente, podem ser observados para diferentes valores de amplitude do pulso. A sequência de pulsos era composta por um pulso de escrita de $200ns$ seguido por um pulso de leitura de $1ms$.

No total, 400 pulsos de escrita foram aplicados ao dispositivo para cada amplitude de tensão. Após cada sequência de pulso, uma varredura de tensão é realizada para recuperar o estado inicial de alta (ou baixa) condutância. O protocolo foi executado cinco vezes para diferentes valores de amplitude do pulso, a fim de avaliar a variabilidade ciclo a ciclo, Fig. 6-b,c. A curva sólida representa o valor médio de condutância para cada tensão, e a faixa colorida representa o desvio padrão. Na Fig. 6-d, um teste de potenciação e depressão para as tensões de $1{,}45V$ e $-2V$.