UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO, DE CIÊNCIAS EXATAS E EDUCAÇÃO
DEPARTAMENTO DE ENG. DE CONTROLE, AUTOMAÇÃO E COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Júlia Ender

Estudo sobre o efeito das interconexões de cobre em circuitos de redes neurais baseadas em transistores MOS de tecnologia 16nm

Blumenau
2024

Júlia Ender

Estudo sobre o efeito das interconexões de cobre em circuitos de redes neurais baseadas em transistores MOS de tecnologia 16nm

Trabalho de Conclusão de Curso de Graduação em Engenharia de Controle e Automação do Centro Tecnológico, de Ciências Exatas e Educação da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Engenheira de Controle e Automação.
Orientadora: Janaína Gonçalves Guimarães, Dra.
Coorientadora: Beatriz Oliveira Câmara da Fé, Ma.

Blumenau
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Ender, Júlia
Estudo sobre o efeito das interconexões de cobre em circuitos de redes neurais baseadas em transistores MOS de tecnologia 16nm / Júlia Ender ; orientadora, Janaína Guimarães, coorientadora, Beatriz Oliveira Câmara Fé, 2024.
56 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Blumenau, Graduação em Engenharia de Controle e Automação, Blumenau, 2024.

Inclui referências.

1. Engenharia de Controle e Automação. 2. Interconexões. 3. Cobre. 4. Redes Neurais Artificiais. 5. Transistores MOS. I. Guimarães, Janaína. II. Fé, Beatriz Oliveira Câmara. III. Universidade Federal de Santa Catarina. Graduação em Engenharia de Controle e Automação. IV. Título.

Júlia Ender

Estudo sobre o efeito das interconexões de cobre em circuitos de redes neurais baseadas em transistores MOS de tecnologia 16nm

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Engenheira de Controle e Automação" e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Controle e Automação.

Blumenau, 15 de Julho de 2024.

Banca Examinadora:

Documento assinado digitalmente
Janaina Goncalves Guimaraes
Data: 20/07/2024 17:09:31-0300
CPF: ***.475.831-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Janaína Gonçalves Guimarães, Dra.
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Adao Boava
Data: 20/07/2024 18:21:22-0300
CPF: ***.640.969-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Adão Boava, Dr.
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Ciro Andre Pitz
Data: 20/07/2024 18:01:51-0300
CPF: ***.728.089-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Ciro André Pitz, Dr.
Universidade Federal de Santa Catarina

Dedico esta monografia aos meus pais, Marli Terezinha Zago Ender e Laércio Ender, e à minha irmã, Manoella Ender, que desde o princípio me apoiaram e acreditaram no meu potencial.

AGRADECIMENTOS

Expresso minha profunda gratidão aos meus pais, Marli T. Z. Ender e Laércio Ender, e à minha irmã, Manoella Ender, por estarem ao meu lado durante toda a minha trajetória, incentivando e apoiando diante das adversidades. Vocês sempre foram a minha base e o meu exemplo de determinação, persistência e ética. A presença constante e o amor incondicional de vocês foram fundamentais para que eu pudesse alcançar meus objetivos.

Agradeço ao meu namorado e melhor amigo, Felipe de Souza Kalume, por sempre me amparar e manter a positividade diante das dificuldades, além de me escutar e aconselhar com tanto carinho.

Agradeço à minha orientadora, Janaína Gonçalves Guimarães, por todo o conhecimento compartilhado e por toda a paciência durante a construção desta monografia, por todas as palavras de carinho e motivação diante dos desafios encontrados, e por toda a prestatividade, independentemente do horário ou dia da semana.

Agradeço à minha coorientadora, Beatriz Oliveira Câmara da Fé, por toda a disponibilidade, auxílio e gentileza ao compartilhar seus conhecimentos no decorrer deste estudo.

Agradeço ao meu amigo, Luiz Augusto Scheuermann França, por todas as risadas e auxílio durante esse período.

Agradeço à Universidade Federal de Santa Catarina e aos professores por proporcionarem um ambiente propício ao aprendizado. A disponibilidade de conteúdos e ferramentas relevantes foi essencial para a realização deste trabalho.

Agradeço ao CNPq pelo suporte financeiro para o desenvolvimento desta monografia.

RESUMO

As redes neurais são cruciais para aplicações de inteligência artificial e aprendizado de máquina, devido à sua capacidade de emular o comportamento do cérebro humano e processar grandes volumes de dados de maneira eficiente. No entanto, a implementação dessas redes enfrenta desafios significativos devido aos altos níveis de conectividade exigidos por redes neurais densas, além de questões de dissipação de potência, tempo de atraso, largura de banda e escalabilidade. Neste contexto, o presente trabalho investiga o impacto das interconexões nos parâmetros de desempenho em circuitos de redes neurais baseadas em transistores MOS de tecnologia 16 nm. Para isso, o estudo categoriza as interconexões em locais, intermediárias e globais, examinando seus efeitos específicos no consumo de potência, no atraso e na largura de banda do sistema. A análise visa identificar combinações de interconexões que podem ser inseridas sem comprometer o desempenho do sistema além dos limites estipulados, garantindo assim a eficiência em ambientes baseados em tecnologia de transistores MOS de 16 nm. Esse esforço contribui para o avanço contínuo da computação neuromórfica e suas aplicações práticas.

Palavras-chave: Interconexões; Redes Neurais Artificiais; Cobre; Transistor MOS.

ABSTRACT

Neural networks are crucial for artificial intelligence and machine learning applications due to their ability to emulate human brain behavior and process large volumes of data efficiently. However, implementing these networks faces significant challenges due to the high levels of connectivity required by dense neural networks, as well as issues related to power dissipation, delay, bandwidth, and scalability. In this context, this study investigates the impact of interconnections on the performance parameters of neural network circuits based on 16 nm MOS transistors. To address these challenges, the study categorizes interconnections into local, intermediate, and global, examining their specific effects on power consumption, delay, and system bandwidth. The analysis aims to identify combinations of interconnections that can be implemented without compromising system performance beyond stipulated limits, thus ensuring efficiency in environments based on 16 nm MOS transistor technology. This effort contributes to the ongoing advancement of neuromorphic computing and its practical applications.

Keywords: Interconnections; Artificial Neural Networks; Copper; MOS Transistor.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADC	*Analog-to-Digital Converter*
CMOS	*Complementary Metal-Oxide Semiconductor*
DAC	*Digital-to-Analog Converter*
FF	*Feed-forward*
HNN	*Hardware Neural Network*
MLP	*Multi-Layer Perceptron*
MLPs	*Multi-Layer Perceptrons*
MOS	*Metal-Oxide Semiconductor*
MOSFET	*Metal-Oxide-Semiconductor Field Effect Transistor*
NMOS	*N-channel Metal-Oxide Semiconductor*
PMOS	*P-channel Metal-Oxide Semiconductor*
ReLU	*Retified Linear Unit*
RNAs	Redes Neurais Artificiais
VLSI	*Very Large Scale Integration*

SUMÁRIO

1 INTRODUÇÃO

A análise do impacto das interconexões nos parâmetros de desempenho de circuitos em redes neurais artificiais baseadas em transistores MOS (*Metal-Oxide Semiconductor*) de tecnologia 16 nm [1] é essencial para melhorar a eficiência e funcionalidade desses sistemas. Redes neurais artificiais desempenham um papel fundamental em aplicações de inteligência artificial e aprendizado de máquina devido à sua habilidade de emular o funcionamento do cérebro humano e processar grandes volumes de dados de maneira eficaz. Contudo, a implementação dessas redes em hardware apresenta desafios significativos, como altos níveis de conectividade, aumento da dissipação de potência, incremento do tempo de atraso, redução da largura de banda e problemas de escalabilidade. O uso de tecnologia menores visa estudar a possibilidade de implementar essas redes neurais em escalas de integração cada vez maiores. Quanto menor o transistor, maior o número de neurônios que podem ser construídos na mesma área. Além disso, dentre os modelos completos de transistores disponíveis publicamente para uso, o de 16 nm já foi utilizado e validado em trabalhos anteriores.

Neste contexto, as interconexões se tornam elementos críticos para a comunicação eficiente entre os elementos das redes neurais. O cobre, devido às suas propriedades elétricas e térmicas superiores, além de sua alta resistência à eletromigração [2], se destacou como o material preferido para essas interconexões em circuitos integrados. Desde sua introdução como alternativa à metalização com liga de alumínio em 1997, houve um crescimento significativo no entendimento dos requisitos de processo e nas questões de confiabilidade associadas ao uso de cobre nas interconexões (MERCHANT *et al.*, 2001).

A presente monografia categoriza as interconexões em três classes segundo Baohui Xu *et al.* (2022), as quais são as interconexões locais, intermediárias e globais. Cada categoria é examinada quanto aos seus efeitos específicos no consumo de potência, tempo de atraso e largura de banda do sistema. A meta é identificar combinações de interconexões que possam ser implementadas sem comprometer o desempenho global do sistema, assegurando assim a eficiência em circuitos baseados em transistores MOS de 16n m. São realizadas simulações empíricas variando as cargas para determinar os valores mínimo e máximo de variação de frequência da rede neural. Os resultados deste estudo fornecem uma base para futuras pesquisas e melhorias em interconexões de circuitos integrados, destacando a importância de considerar diferentes tipos de interconexões na construção de sistemas eficientes e robustos.

---

[1] Transistores MOS com uma largura de canal de 16 nanômetros. A redução da largura do canal permite uma maior densidade de transistores, melhorando o desempenho e a eficiência energética.

[2] Processo pelo qual átomos em um condutor metálico são deslocados devido ao fluxo contínuo de corrente elétrica. Esse fenômeno pode causar a formação de vazios e aglomerações, levando à degradação do desempenho das interconexões e eventualmente à falha do circuito. O efeito é mais pronunciado em condições de alta densidade de corrente e temperaturas elevadas, afetando a confiabilidade dos dispositivos eletrônicos.

1.1 OBJETIVOS

Nas seções a seguir estão descritos o objetivo geral e os objetivos específicos deste TCC.

1.1.1 Objetivo Geral

Esta monografia tem como objetivo geral realizar um estudo e análise sobre o impacto das interconexões físicas no desempenho de redes neurais baseadas em transistores MOS de tecnologia $16n$ m.

1.1.2 Objetivos Específicos

Visando atingir o objetivo geral citado anteriormente, esta monografia tem como objetivos específicos:

1. Investigar como as redes neurais podem ser eficientemente interconectadas;
2. Classificar as interconexões de cobre físicas em redes neurais como locais, intermediárias e globais, e estudar os efeitos dessas interconexões no consumo de potência, na frequência máxima e no atraso do sistema;
3. Identificar combinações de interconexões que assegurem a comunicação eficiente entre neurônios, mantendo a dissipação de potência e o tempo de atraso dentro dos limites estipulados.

1.2 ORGANIZAÇÃO DO TRABALHO

A presente monografia está organizada em cinco capítulos, conforme descrito a seguir:

1. **Introdução**: O presente capítulo discorre sobre o tema e o contexto do estudo, além de apresentar o objetivo geral e os objetivos específicos;
2. **Revisão Bibliográfica**: Fornece as bases conceituais essenciais para a compreensão da pesquisa, abordando temas como redes neurais artificiais, redes em chip, tecnologia MOS e interconexões;
3. **Metodologia:** Descreve as metodologias adotadas para a obtenção dos dados e avaliação dos resultados. Demonstra a utilização do *software LTspice* (ANALOG DEVICES, 2023) para simulações, abordando a coleta de dados do circuito ideal, simulações de potência, frequência de corte e tempo de atraso, bem como quais são as ponderações para a análise de desempenho do circuito;
4. **Análise e Resultados:** Esse capítulo apresenta uma análise dos dados obtidos e discute algumas combinações de interconexões cabíveis, avaliando o desempenho e a eficiência de cada uma;

5. **Conclusão:** Apresenta uma síntese dos principais pontos discutidos ao longo do estudo, destacando as contribuições da pesquisa para o campo das redes neurais artificiais e das interconexões em circuitos integrados.

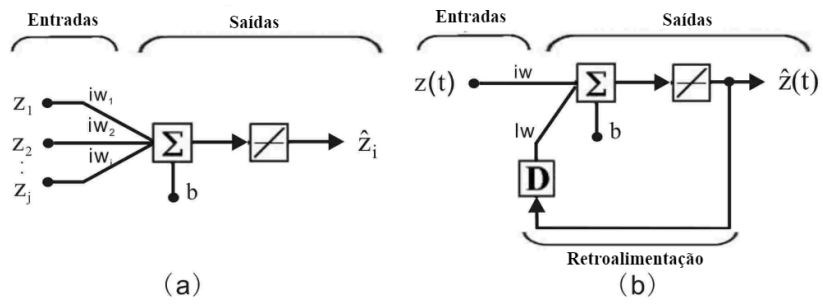
2 REVISÃO BIBLIOGRÁFICA

2.1 REDES NEURAIS ARTIFICIAIS

As RNAs (Redes Neurais Artificiais) se destacam como um paradigma computacional poderoso e versátil. Inspiradas na complexa rede de conexões entre neurônios, as RNAs processam informações e resolvem problemas de forma autônoma. Com o desenvolvimento de novas técnicas e o aprimoramento das redes neurais, podemos esperar avanços ainda mais significativos em áreas como processamento de linguagem natural, visão computacional, robótica, saúde, entre outros.

No aspecto estrutural, as RNAs podem ser classificadas em modelos de arquitetura. O primeiro são as redes FF (*Feed-forward*), onde o processamento ocorre camada por camada, da entrada até a saída da rede. Já o segundo são as redes recorrentes que possuem uma estrutura de rede interconectada com ciclos. A principal aplicação da primeira classe de redes é a classificação supervisionada, realizada por meio de um algoritmo de *Perceptron* (FORSSELL, 2014). A segunda categoria de redes é mais ampla e abrange aplicações como mapas auto-organizáveis, memória associativa, como as redes de Hopfield, e redes de Boltzmann. Essas redes possuem uma estrutura mais complexa e são utilizadas em diversos contextos de aprendizado de máquina e inteligência artificial (MOERLAND; FIESLER, 2020). A Figura 1 mostra os modelos das conexões das redes recorrente e FF.

Figura 1 – Modelos de conexão. (a) Rede FF (b) Rede recorrente.



Fonte: Adaptado de Cheng, Wang e Li (2008).

Um exemplo notável de rede FF é a rede neural multicamada, MLP (*Multi-Layer Perceptron*). Essa rede consiste em múltiplas camadas de neurônios artificiais, onde cada camada está completamente conectada à camada seguinte. Isso significa que todos os neurônios de uma camada enviam sinais para todos os neurônios da próxima camada, sem *feedback*. As MLPs são aplicadas em problemas de aprendizado supervisionado, como classificação e regressão. Cada neurônio realiza uma combinação linear das entradas, seguida por uma função de ativação não linear, como a sigmoidal ou a ReLU (*Retified Linear Unit*). As camadas intermediárias permitem que o aprendizado de representações

complexas dos dados de entrada, tornando-o uma ferramenta poderosa para diversas tarefas de aprendizado de máquina (MOHAMMADZADEH *et al.*, 2022).

A estrutura das RNAs, independentemente de sua aplicação específica, possui dois componentes principais que se assemelham aos sistemas biológicos. Estes componentes são os neurônios e as sinapses, que correspondem aos vértices e arestas de um grafo, respectivamente.

Dentre os desafios de implementação das

Já no componente do neurônio, os desafios incluem o soma do neurônio, onde deve ser realizada a somatória das entradas ponderadas e a função de ativação, que costuma ser altamente não linear (FORSSELL, 2014; MOERLAND; FIESLER, 2020; SUNDARARAJAN; SARATCHANDRAN, 1998).

2.1.1 Regras de Aprendizado

O processo geral da maioria das RNAs envolve duas etapas principais, aprendizado e recuperação. Durante a etapa de aprendizado, os pesos da rede são ajustados conforme a aplicação. Para o *Perceptron*, isso envolve o uso do algoritmo de retropropagação em um conjunto de dados de treinamento classificado. No caso das memórias associativas, os pesos são configurados para que as memórias desejadas se tornem atratores locais. Na etapa de recuperação, novos dados são introduzidos e a rede processa esses dados, ajustando-se conforme necessário. Enquanto a recuperação sempre ocorre na rede física, o aprendizado pode ser feito antecipadamente. Ainda segundo Forssell (2014), existem três estratégias de treinamento:

1. **Aprendizado fora do chip:** consiste em realizar a fase de treinamento em uma rede simulada em *software*. Esta abordagem permite que os cálculos de treinamento sejam realizados de maneira mais rápida e precisa do que seria possível utilizando *hardware* físico. No entanto, essa metodologia não considera as variações que podem ocorrer durante a fabricação do *hardware*;
2. **Aprendizado *chip-in-the-loop*:** consiste na utilização de uma rede de *hardware* juntamente com computação externa em *software*. O *software* executa o algoritmo de aprendizado, enquanto a rede de *hardware* realiza os cálculos. No caso do algoritmo de retropropagação, por exemplo, a passagem direta é realizada pela rede de *hardware*, e as atualizações de peso são efetuadas pelo *software*. Isso permite que a precisão dos cálculos não seja limitada pelas capacidades do *hardware*, ao mesmo tempo que o comportamento real da rede é considerado;
3. **Aprendizado *on-chip*:** consiste apenas na utilização do chip de *hardware* para realizar todo o processo de aprendizado. Embora este método seja mais lento e menos preciso nos cálculos de peso em comparação com outros métodos, ele não

depende de manipulação externa durante a fase de aprendizado. Isso o torna mais adequado para *hardware* embarcado, permitindo que as redes atualizem seu aprendizado ao longo do tempo. No entanto, o design é inerentemente mais complexo e menos flexível, pois o algoritmo de aprendizado precisa ser implementado diretamente no *hardware*.

2.1.2 Tipos de implementação

A rede neural pode ser implementada em *software* ou em *hardware*, e a escolha de implementação depende de diversos fatores, como o tamanho e a complexidade da rede, a necessidade de desempenho em tempo real, o consumo de energia e o custo. Cada abordagem oferece vantagens e desvantagens que devem ser consideradas.

Esta monografia está focada nas implementações em *hardware*, i.e., nos dispositivos de *hardware* projetados para realizar arquiteturas de RNAs e nos algoritmos de aprendizado associados, aproveitando o paralelismo inerente no processamento neural. Esses dispositivos são conhecidos como redes neurais de *hardware*, HNN (*Hardware Neural Network*) (MISRA; SAHA, 2010). Suas vantagens em relação à implementação em *software* estão relacionadas à algumas condições como velocidade, custo e degradação gradual. O *hardware* especializado entrega um poder computacional com preços limitados, alcançando velocidades muito maiores do que computadores ou estações de trabalho, como por exemplo no caso de implementações de sistemas VLSI (*Very Large Scale Integration*) que podem atingir velocidades de até vários *teraflops* (HÄNGGI; MOSCHYTZ, 2000). As implementações em *hardware* podem reduzir o custo do sistema ao diminuir a contagem de componentes e os requisitos de energia (MISRA; SAHA, 2010). Por fim, enquanto aplicações baseadas em processadores sequenciais podem parar de funcionar com falhas no sistema, a implementações em *hardware* oferecem tolerâncias a falhas (ZHANG *et al.*, 2007).

Os desafios na implementação de HNN englobam principalmente temas como mapeamento de topologias de interconexões complexas em superfícies bidimensionais regulares, restrições de *hardware* que podem introduzir erros de computação degradando o aprendizado e a precisão de resultados e a não-linearidade das funções de ativação das HNN (MISRA; SAHA, 2010).

2.1.2.1 Implementações físicas

Nesta parte são discutidas as implementações físicas via circuitos digitais e via circuitos analógicos.

2.1.2.1.1 Circuitos Digitais

As RNAs podem ser implementadas utilizando circuitos digitais (FRYE; RIETMAN; WONG, 1991; JUNG; KIM, 2007; HIKAWA, 2005; MEROLLA *et al.*, 2014), cujas vantagens são fácil construção e fácil projeto. Esses circuitos se baseiam em elementos lógicos já existentes e os pesos sinápticos podem ser implementados com células de memória digital ou *latches*. A quantidade de bits usados para armazenar os pesos é fundamental para a precisão do algoritmo, especialmente durante a fase de aprendizado (MOERLAND; FIESLER, 2020). Embora o aprendizado fora do chip, utilizando *software* de alta precisão e ajustando o resultado final para a precisão do chip, possa atenuar essa questão, é essencial contar com uma implementação eficaz dos pesos sinápticos. A soma dos estados dos neurônios pode ser implementada com multiplicadores comuns e estágios de somadores. No entanto, com um grande número de neurônios de entrada, a quantidade total desses elementos pode se tornar significativa.

Implementar a função de ativação pode ser mais desafiador devido à sua necessidade de ser altamente não linear. Enquanto uma função de limiar simples pode ser implementada com facilidade, suas capacidades são limitadas (FORSSELL, 2014). Por outro lado, funções de ativação mais sofisticadas, como a sigmoide, exigem tabelas de consulta, o que pode reduzir a velocidade dos cálculos e demandar uma área considerável do chip, além de aumentar o consumo de energia para garantir alta precisão.

Embora a adaptação da lógica digital para RNAs resulte em projetos relativamente simples, esses projetos não são ideais em termos de eficiência energética e de espaço. Em contrapartida, uma vantagem significativa de usar tecnologia digital é a facilidade com que pode ser integrada a outros circuitos padrão, possibilitando a fabricação com o mesmo processo.

2.1.2.1.2 Circuitos Analógicos

De maneira generalizada, o *design* de circuitos integrados é muito mais complexo no caso de circuitos analógicos quando comparado com circuitos digitais, especialmente ao se escalar para um grande número de portas. Essa mesma lógica se aplica aos circuitos RNAs. Embora o *design* analógico possa resultar em circuitos mais eficientes em termos de energia e área (HOLLIS; PAULOS, 1990; INDIVERI; HORIUCHI, 2011), seu *design* não é facilmente automatizado como os circuitos digitais.

Esse problema pode ser mitigado através da abordagem híbrida, com partes da rede projetada com circuitos digitais. Por exemplo, os valores dos pesos sinápticos podem ser digitalizados e armazenados em memórias digitais, semelhante ao que é feito para sinapses digitais. Para realizar o cálculo, serão necessários elementos conversores ADC (*Analog-to-Digital Converter*)/DAC (*Digital-to-Analog Converter*), o que pode introduzir atrasos indesejáveis na computação. Além disso, a escalabilidade em termos de energia e

área dependerá da precisão dos pesos sinápticos. Alternativamente, os pesos podem ser armazenados utilizando elementos analógicos, como resistores ou capacitores, o que é viável se os pesos forem fixados no projeto (FORSSELL, 2014). Existem métodos para armazenar pesos de forma programável usando elementos analógicos reconfiguráveis (HOLLIS; PAULOS, 1990), que podem ser mais eficientes do que armazená-los em memória digital. No entanto, isso ainda exigirá conversores ADC/DAC durante a etapa de aprendizado, ou seja, para a atualização dos pesos.

A soma das entradas de um neurônio não costuma ser um problema em um circuito analógico, uma vez que as entradas e saídas do neurônio são representadas como correntes ou tensões. As correntes podem ser somadas conectando-se os ramos do circuito em paralelo, enquanto as tensões são somadas conectando-os em série. Já a função de ativação é geralmente implementada utilizando um amplificador, que evidencia uma forte não linearidade no regime de saturação. Funções de ativação essencialmente arbitrárias podem ser construídas usando circuitos relativamente simples (FORSSELL, 2014).

As variações de fabricação inerentes aos circuitos analógicos limitam a precisão alcançável das RNAs analógicas. Para atingir um certo nível de precisão, os circuitos precisam ser dimensionados suficientemente grandes para minimizar os efeitos das tolerâncias. No entanto, redes neurais que utilizam memória associativa são conhecidas por sua tolerância a falhas (FORSSELL, 2014), permitindo alguma variação nos elementos básicos do circuito sem comprometer significativamente o desempenho.

2.1.3 Rede Neural *Perceptron* multicamada baseada em transistores CMOS (*Complementary Metal-Oxide Semiconductor*)

Uma rede neural *Perceptron* multicamada MLP baseada em transistores CMOS é uma implementação de *hardware* de RNAs utilizando a tecnologia de semicondutores complementar metal-óxido-silício (CMOS). Esta abordagem combina as vantagens das redes neurais, como capacidade de aprendizagem e processamento paralelo, com a eficiência e a escalabilidade da tecnologia CMOS.

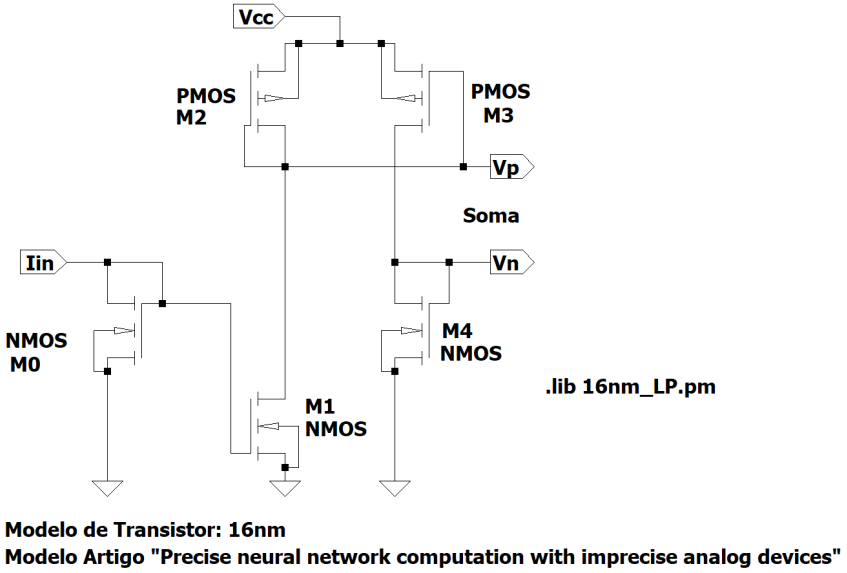
As redes neurais *Perceptron* multicamada são compostas por múltiplas camadas de neurônios artificiais, incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio em uma camada está conectado a todos os neurônios da camada seguinte através de pesos sinápticos (LECUN; BENGIO; HINTON, 2015).

A tecnologia CMOS é a base para a maioria dos circuitos integrados modernos devido à sua baixa dissipação de potência e alta densidade de integração. Implementar redes neurais nessa tecnologia envolve projetar circuitos que emulam o comportamento dos neurônios e sinapses (MEAD, 1989).

O estudo realizado nesta monografia é baseado no circuito de rede neural *Perceptron* multicamada utilizando o modelo de neurônio analógico proposto por Binas *et al.* (2020) e Binas *et al.* (2016). Em um trabalho anterior, este circuito foi implementado e analisado

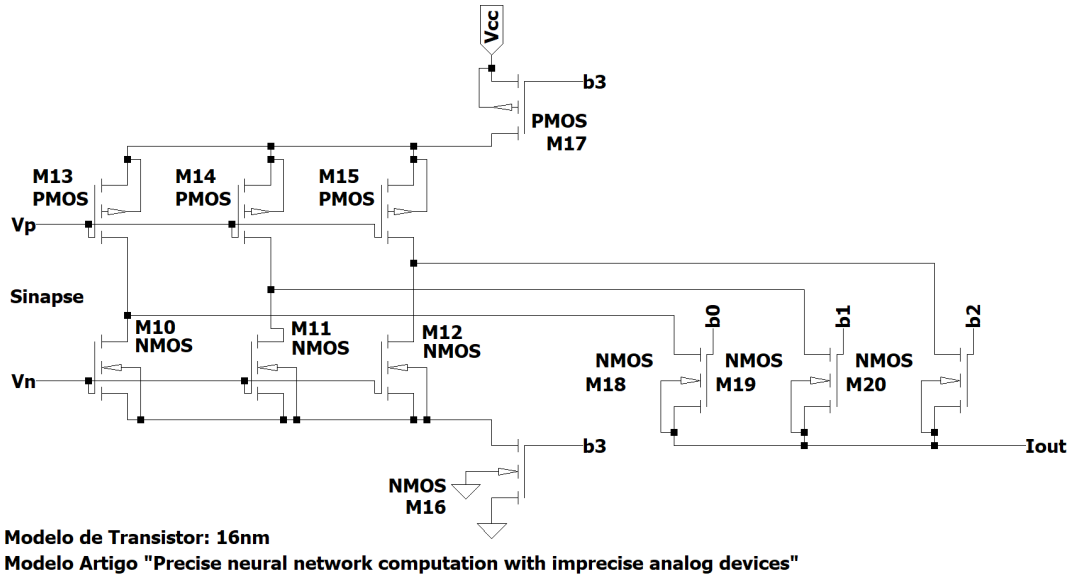
por França (2023) utilizando a aplicação na classificação dados Iris de Fisher para avaliar a implementação em tecnologia de transistores CMOS $16n$ m com 4 neurônios na camada de entrada, 7 neurônios na camada oculta e 3 neurônios na camada de saída, conforme as Figuras 2, 3, 4, 5 e 6. A implementação mostrou-se bem sucedida, demonstrando que a rede neural baseada em Binas pode operar eficientemente em tecnologia de $16n$ m, alcançando alta precisão e eficiência energética. A partir deste projeto de rede neural é realizada a avaliação de desempenho junto à inserção das interconexões pertinentes.

Figura 2 – Circuito de soma.



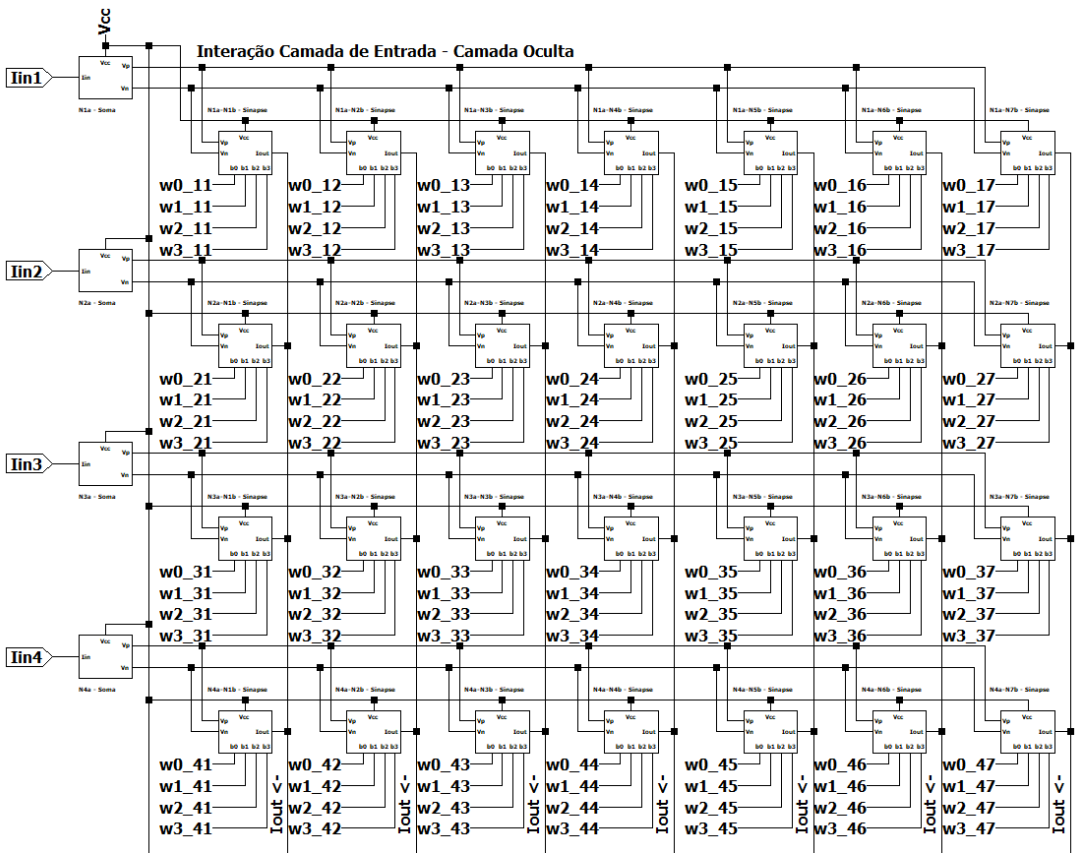
Fonte: França (2023).

Figura 3 – Circuito de sinapse.



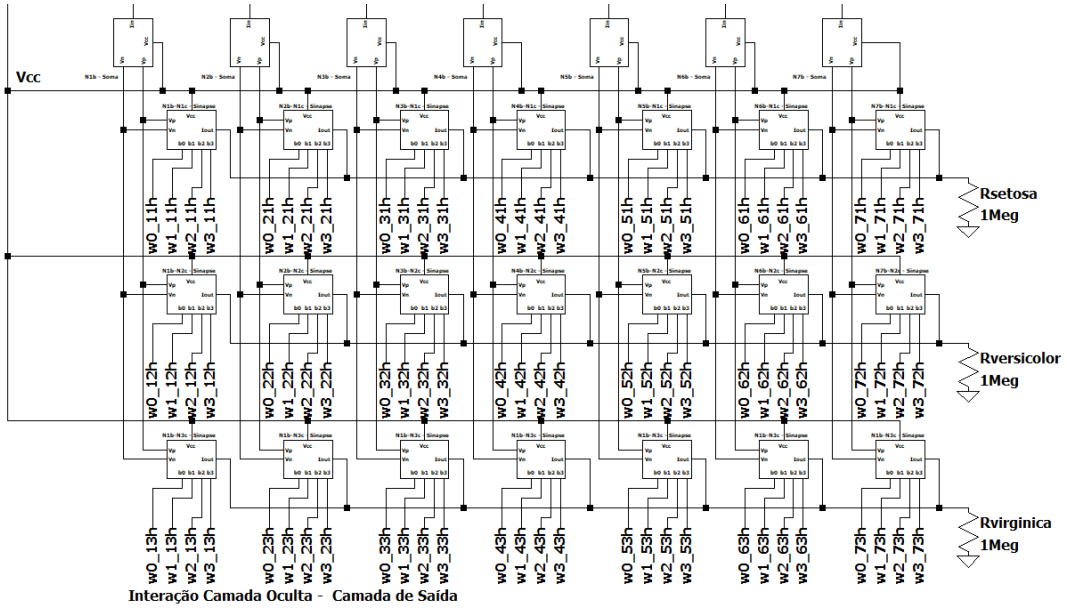
Fonte: França (2023).

Figura 4 – Estrutura de conexão entre neurônios da camada de entrada e camada oculta.



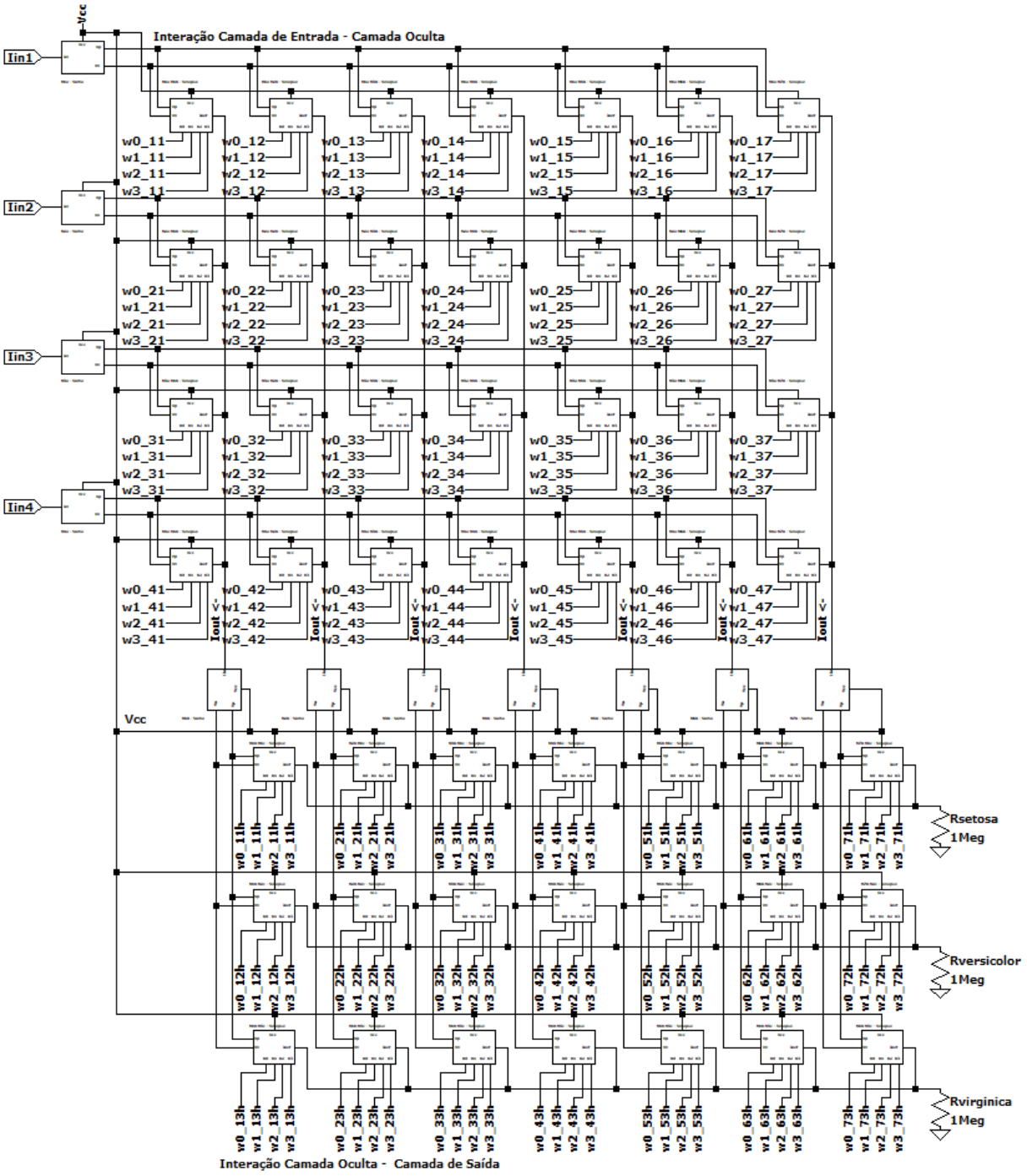
Fonte: França (2023).

Figura 5 – Estrutura de conexão entre neurônios da camada oculta e camada de saída.



Fonte: França (2023).

Figura 6 – Estrutura de conexão em rede entre neurônios da camada oculta e camada de saída - Modelo completo.



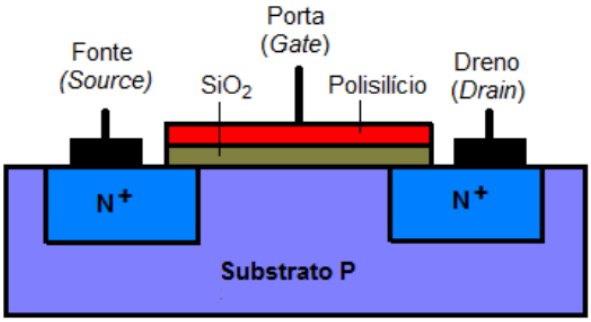
Fonte: França (2023).

2.2 TRANSISTORES MOS

O transistor MOSFET (*Metal-Oxide-Semiconductor Field Effect Transistor*), também conhecido como MOS, é composto por quatro terminais, sendo eles o dreno, a fonte,

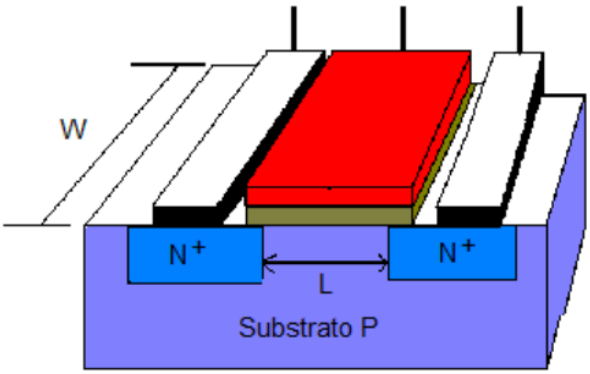
a porta e o substrato. Sua estrutura básica inclui três camadas fundamentais: o metal, o óxido de silício, que isola a porta do restante do transistor, e o semicondutor de silício monocristalino que forma o substrato. Essas camadas são mostradas na Figura 7. Já na Figura 8 são apresentadas as dimensões do canal do transistor, em que W refere-se à largura que está relacionada com o fluxo de corrente passa pelo transistor quando está conduzindo e L refere-se ao comprimento do canal, medidas essas cruciais para definir suas características elétricas e garantir a eficiência energética (ZIMPECK; MEINHARDT; BUTZEN, 2014).

Figura 7 – Camadas transistor MOS.



Fonte: Zimpeck, Meinhardt e Butzen (2014).

Figura 8 – Dimensões transistor MOS.



Fonte: Zimpeck, Meinhardt e Butzen (2014).

A evolução da tecnologia dos transistores MOSFET tem sido guiada pela Lei de Moore, que prevê a duplicação do número de transistores em um chip aproximadamente a cada 18 a 24 meses (ZHAO; CAO, 2007; WU *et al.*, 2013). A miniaturização contínua levou ao desenvolvimento da tecnologia de $16n$ m, um marco significativo na indústria de semicondutores. A redução do comprimento do canal L para $16n$ m reflete a capacidade de fabricar transistores com distâncias extremamente pequenas entre as difusões, aumentando

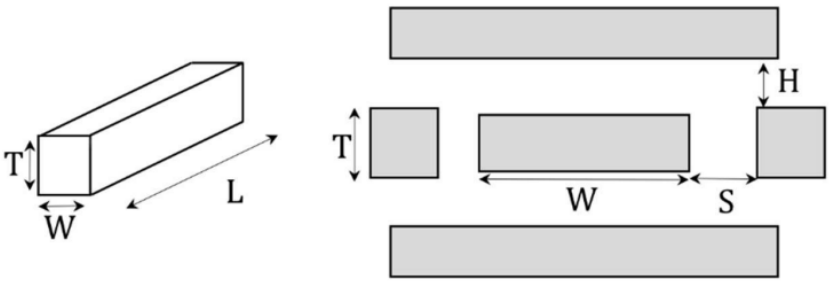
assim a densidade de transistores em um chip. A transição para a tecnologia de $16n$ m foi possível através de avanços significativos em materiais e técnicas de fabricação, permitindo o desenvolvimento de sistemas mais poderosos e eficientes, mantendo os princípios da Lei de Moore e continuando a inovação na microeletrônica (SHEU *et al.*, 1987; KUHN, 2012).

A tecnologia CMOS é uma tecnologia de construção de circuitos integrados que combina transistores MOSFET de canal N, NMOS (*N-channel Metal-Oxide Semiconductor*), e de canal P, PMOS (*P-channel Metal-Oxide Semiconductor*), em que os canais dos tipos N ou P são resultantes de um processo denominado dopagem do semicondutor de silício. A dopagem consiste na adição controlada de impurezas químicas na estrutura cristalina do material (DUARTE *et al.*, 2017). Semicondutores do tipo N (*negative*) possuem maior quantidade de elétrons livres circulando através do material, aumentando a negatividade do mesmo. Em contrapartida, semicondutores do tipo P (*positive*) possuem uma quantidade maior de lacunas no fragmento semicondutor, aumentando a positividade do mesmo (JAEGER; BLALOCK, T. N.; BLALOCK, B. J., 1997; FRANÇA, 2023).

2.3 INTERCONEXÕES DE COBRE

As interconexões possuem dimensões associadas à sua seção transversal, conforme Figura 9, em que T é a sua espessura, W é a sua largura, L é o seu comprimento, S é a distância entre os condutores pertencentes a uma mesma camada e H é a espessura do isolante entre camadas.

Figura 9 – Dimensões da interconexão.



Fonte: Rangel (2018).

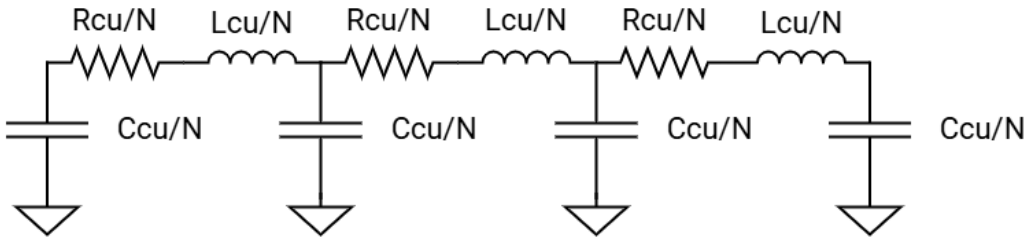
É possível ajustar os parâmetros de largura, espessura e espaçamento entre os fios. A dificuldade atrelada ao uso das interconexões se dá pelo fato de que ao aumentar a largura e o espaçamento dos fios, tem-se o aumento do tempo de atraso de propagação. Todavia, isso impacta diretamente na redução da largura de banda global do sistema, podendo esse parâmetro de desempenho se tornar um impeditivo da implementação. É necessário encontrar uma harmonia entre as dimensões das interconexões e os parâmetros de desempenho avaliados (REEHAL, 2012).

Essas interconexões podem ser produzidas com vários materiais condutores, com cobre, alumínio, prata e nanotubos de carbono.Para esta monografia é utilizado modelo de cobre.

As interconexões podem ser modeladas a partir das suas características físicas como resistência, indutância e capacitância.

Para esta monografia será utilizado o modelo de cadeia em π de 3 segmentos conforme Figura 10. Este modelo fornece uma medida de erro inferior à 3% (BAKOGLU, 1990).

Figura 10 – Modelo de Interconexão de Cobre π de 3 segmentos.



Fonte: Da Autora.

Neste modelo R_{Cu} é a resistência do cobre, L_{Cu} é a indutância do cobre, C_{Cu} é a capacitância do cobre e N é número de segmentos do modelo.

2.3.1 Resistência do Cobre

A Resistência do cobre é dada por $R_{Cu} = \frac{\rho \cdot L}{W \cdot T}$, onde ρ é a resistividade do cobre. A resistividade do cobre é influenciada pela combinação dos fenômenos de espalhamento superficial e de espalhamento de contorno, quando tratada em escala nanométrica (RAYCHOWDHURY; ROY, 2004; XU, Y.; SRIVASTAVA, 2010). Estes fenômenos correspondem aos parâmetros ρ_{FS}, proposto por Fuchs e Sondheimer e dado por $\rho_{FS} = \rho_o \left[1 + \frac{3}{4} \cdot \frac{\lambda_o}{W} (1 - p_F)\right]$, e ρ_{MS}, apresentado por Mayadas e Shatzkes e dado por $\rho_{MS} = \frac{\rho_o}{3} \left[\frac{1}{3} - \frac{\alpha}{2} + \alpha^2 + \alpha^3 \cdot \ln\left(1 + \frac{1}{\alpha}\right)\right]$ (DAS; RAHAMAN, 2010; KOO, 2011).

O coeficiente α é calculado como $\alpha = \frac{\lambda_o}{D} \cdot \frac{R}{1-R}$. A variável ρ_o representa a resistividade do cobre, desconsiderando os fenômenos citados anteriormente, λ_o é o caminho médio livre de elétrons do cobre, p_F é o parâmetro de espalhamento de Fuchs, R é o coeficiente de reflexão no contorno que assume valores entre 0 e 1, e D é o tamanho médio da região de depleção do contorno. A resistividade do cobre pode ser calculada a partir da soma dos parâmetros ρ_{FS} e ρ_{MS}. Assim, tem-se a resistência do cobre dada por $R_{Cu} = \frac{(\rho_{FS} + \rho_{MS})L}{W \cdot T}$.

2.3.2 Indutância do Cobre

A indutância do cobre é composta pela soma da indutância própria em escala nanométrica dada por $L_{Cu} = \frac{\mu_o \cdot L}{2\pi}\left[\ln\left(\frac{2L}{W+T}\right) + \frac{1}{2} + \frac{0{,}22(W+T)}{L}\right]$ e da indutância mútua em escala nanométrica dada por $M_{Cu} = \frac{\mu_o \cdot L}{2\pi}\left[\ln\left(\frac{2L}{S}\right) - 1 + \frac{S}{L}\right]$, onde μ_o é a permeabilidade magnética do vácuo.

2.3.3 Capacitância do Cobre

Para fins de simulação, a capacitância do cobre pode ser modelada a partir de quatro capacitores de placas paralelas e pela capacitância de borda. A capacitância total do fio pode ser obtida por $C_T = C_a + 2C_b + \frac{C_c}{S}$, que relaciona todas as parcelas citadas anteriormente. Onde C_a é a capacitância de borda, C_b é a capacitância de placas paralelas e C_c é a capacitância de acoplamento.

A capacitância de placas paralelas é obtida a partir de $C_b = \xi\frac{L}{H} \cdot W$ (REEHAL, 2012), onde ξ é a permissividade elétrica para uma dada constante dielétrica. A capacitância de borda para fins de modelagem e simulação pode ser aproximada por $C_a = \xi \cdot L\left[\left(\frac{W}{H}\right) + 0{,}77 + 1{,}06\left(\frac{W}{H}\right)^{0{,}25} + 1{,}06\left(\frac{T}{H}\right)^{0{,}5}\right]$ e a capacitância de acoplamento por $C_c = \xi \cdot L \cdot S\left[\left(0{,}03\frac{W}{H}\right) + 0{,}83\left(\frac{T}{H}\right) - 0{,}07\left(\frac{T}{H}\right)^{0{,}222}\right]\left(\frac{H}{S}\right)^{0{,}75}$ (REEHAL, 2012).

2.3.4 Classificação das interconexões

As interconexões podem ser categorizadas em interconexões locais, interconexões intermediárias e interconexões globais de acordo com o seu comprimento. A Tabela 1 mostra o intervalo de comprimento de cada uma das classes de acordo com a classificação de Baohui Xu *et al.* (2022).

Tabela 1 – Classificação das interconexões.

Tipo de interconexão	**Dimensão**
Local	$< \sim 2\mu$m
Intermediária	$2 \sim 100\mu$m
Global	$> \sim 100\mu$m

Fonte: Adaptado Baohui Xu *et al.* (2022).

As interconexões locais geralmente estão em escala nanométrica e estão conectadas próximas dos transistores e blocos lógicos. Já as interconexões intermediárias costumam estar em escalas milimétricas e servem em sua maioria para interligar blocos lógicos e núcleos. Por fim, as interconexões globais são utilizadas para conectar os componentes mais distantes no chip, bem como são responsáveis pela distribuição de energia e sinal de *clock* (XU, B. *et al.*, 2022).

2.3.5 Interconexões aplicadas às redes neurais

Em um projeto de uma rede neural, as interconexões são fundamentais na comunicação e conforme comentado anteriormente podem impactar significativamente nos parâmetros de desempenho geral do sistema. Dentre os desafios, pode-se citar o consumo total de energia das interconexões, a garantia da largura de banda necessária e a garantia de que o tempo de atraso agregado pelas interconexões não comprometa a performance sistemática. Com o avanço da tecnologia em direção à escala nanométrica, aumentar a largura de banda dos canais de comunicação torna-se uma tarefa complexa.

2.3.5.1 Parâmetros de desempenho

Nesta subseção serão abordados os parâmetros de desempenho das interconexões que serão utilizados na análise proposta.

2.3.5.1.1 Tempo de Atraso

O tempo de atraso (t_d) é causado por conta do tempo de carregamento e de descarregamento da capacitância de carga, bem como ao tempo de chaveamento dos transistores (SEDRA; SMITH, 2007). O tempo de atraso é obtido a partir da diferença entre o tempo medido a 50% da transição dos sinais de entrada e de saída (BAKOGLU, 1990).

Para os circuitos lógicos, este atraso na propagação é calculado pela média aritmética do tempo de propagação do sinal de alto para baixo (t_{PHL}) e do tempo de propagação do baixo para alto (t_{PLH}) como $t_d = \frac{t_{PHL}+t_{PLH}}{2}$ (SEDRA; SMITH, 2007).

2.3.5.1.2 Velocidade Máxima

A velocidade máxima sem distorção do sinal de entrada é uma métrica importante para avaliar o desempenho das interconexões em circuitos integrados (SRIVASTAVA; XU, Y.; SHARMA, 2010; KURUVILLA; RAINA, 2009). Esta análise pode ser realizada pela frequência máxima de operação da interconexão, que representa a largura de banda na qual o sinal começa a decair em -3 dB em relação ao sinal de referência (SRIVASTAVA; XU, Y.; SHARMA, 2010; SEDRA; SMITH, 2007).

2.3.5.1.3 Potência dissipada

A dissipação de potência é constituída da soma de duas variáveis, a potência estática e a potência dinâmica. A potência estática é dada por $P_{Est} = V_{DD} \cdot I$ e ocorre devido ao efeito da resistência do circuito (SEDRA; SMITH, 2007). Já a potência dinâmica é dada por $P_{Din} = f \cdot V_{DD}^2 \cdot C$ e ocorre por causa do efeito da capacitância de carga do

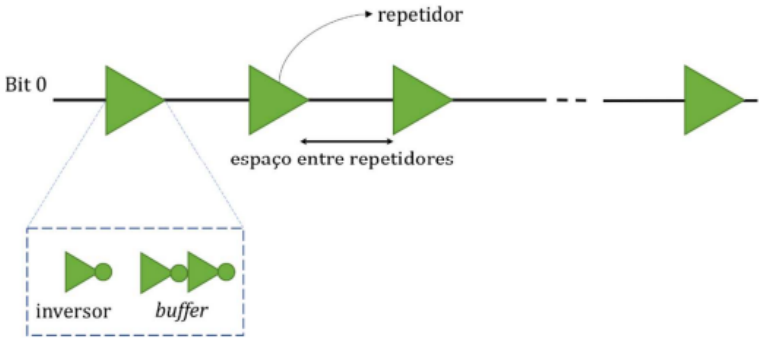
circuito. A variável V_{DD} é a tensão de alimentação, f é a frequência do circuito e C é a capacitância de carga do circuito (BAKOGLU, 1990; SEDRA; SMITH, 2007).

2.3.6 Repetidores

A inserção de repetidores nas interconexões é uma técnica amplamente utilizada para melhorar a performance dos sistemas em termos de largura de banda e tempo de atraso. Entretanto, o uso de repetidores pode levar a um aumento no consumo de potência e no tempo de atraso. Esse aumento pode ser gerenciado e minimizado por meio de técnicas de gerenciamento de energia e roteamento eficiente (BENINI; DE MICHELI, 2002).

Os repetidores são essencialmente amplificadores, conforme mostra a Figura 11, que regeneram o sinal em intervalos regulares, permitindo que ele percorra distâncias maiores sem degradação significativa por atenuação e interferência (DALLY; TOWLES, 2001). A regeneração dos sinais mantém a largura de banda efetiva comparável à das interconexões menores (HU; MARCULESCU, 2003; BJERREGAARD; MAHADEVAN, 2006; MURALI; BENINI; DE MICHELI, 2005). Apesar da inserção dos repetidores aumentar a latência do circuito, a regeneração do sinal é um benefício maior do que o atraso inserido.

Figura 11 – Repetidores.



Fonte: Rangel (2018).

3 METODOLOGIA

Para a simulação e coleta de dados é utilizado o *software LTspice*. O processo é dividido em três etapas: levantamento de dados das interconexões, levantamento de dados referentes ao circuito ideal, e inserção das interconexões no circuito ideal.

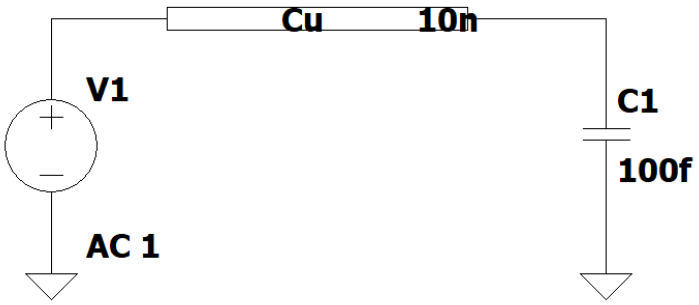
3.1 LEVANTAMENTO DE DADOS DAS INTERCONEXÕES

As interconexões foram simuladas na temperatura padrão do*LTspice* que é equivalente a 27^o C. A seguir são elucidadas as simulações realizadas.

3.1.1 Dissipação de potência

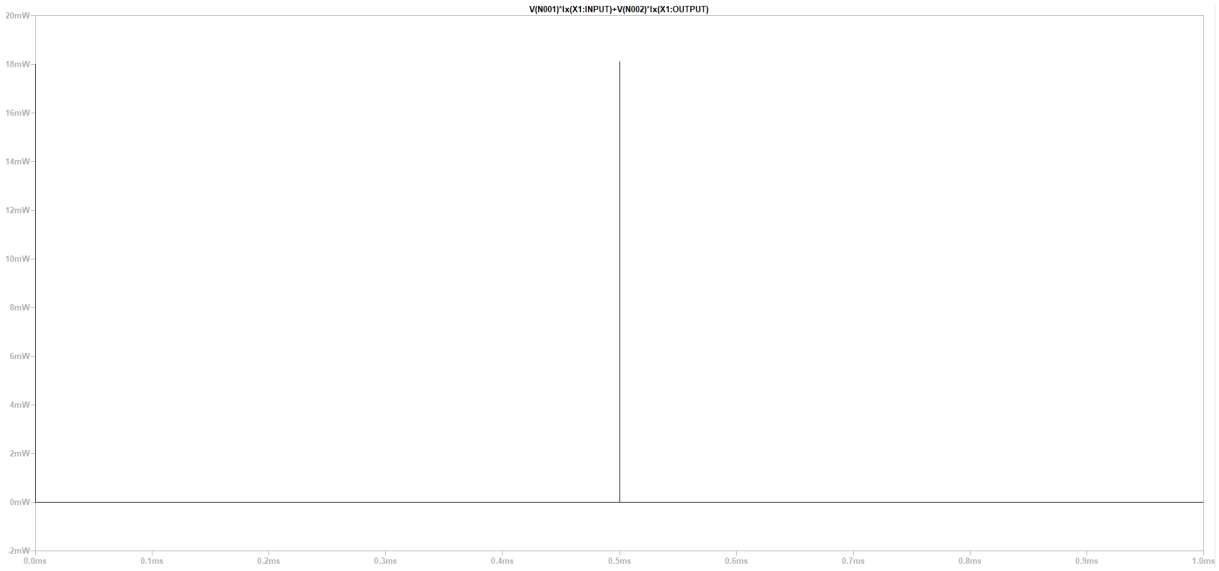
A dissipação de potência das interconexões isoladas é obtida aplicando um pulso com período de 1 ms variando entre 0 e 1V na entrada interconexão e conectando na sua saída um capacitor de $100fF$ para simular a conexão a um transistor 16 nm, conforme Figura 12. O bloco de cobre 10 nm apresentado na Figura 12 contém a interconexão de cobre de 10 nm de acordo com o modelo apresentado na Figura 10. O valor do capacitor foi escolhido de forma a simular a conexão a entrada de um transistor de 16 nm. A dissipação de potência é determinada plotando o pulso de tensão sobre a interconexão como mostra a Figura 13 e calculando sua média, como ilustrado na Figura 14.

Figura 12 – Simulação da interconexão unitária no *LTspice*



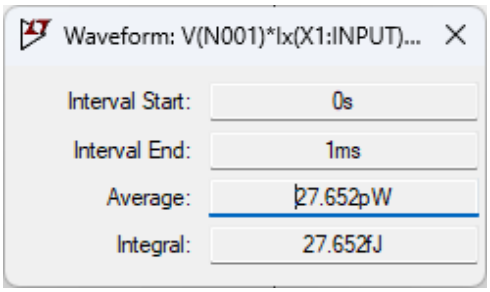
Fonte: Da autora.

Figura 13 – Obtenção do pulso de dissipação de potência no *LTspice*



Fonte: Da autora.

Figura 14 – Média da dissipação de potência no *LTspice*

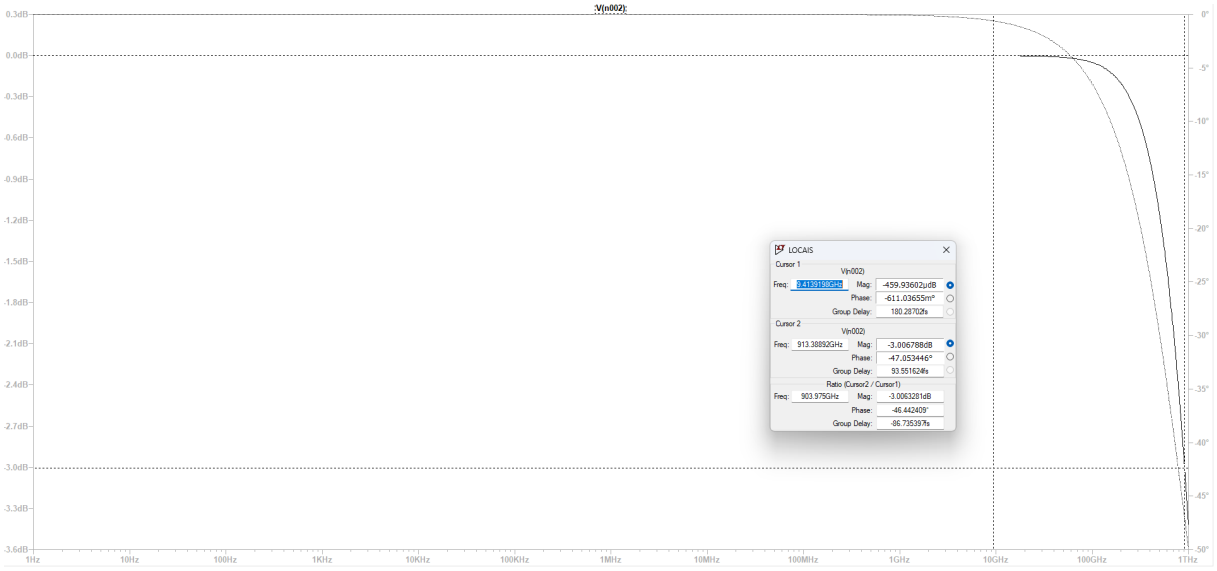


Fonte: Da autora.

3.1.2 Frequência crítica

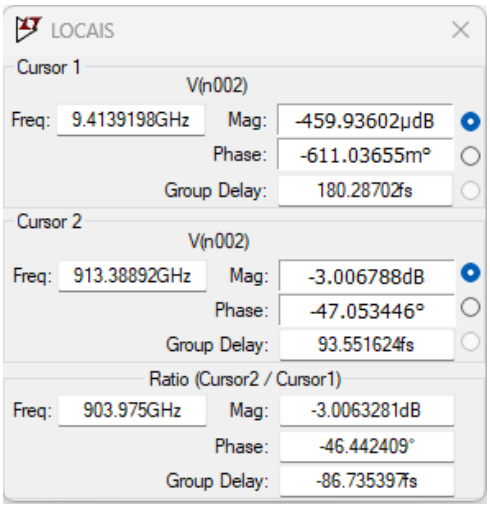
A frequência crítica é identificada através do diagrama de Bode, coletando dados na frequência de corte de -3 dB, sendo ela equivalente a $903,975$ GHz, conforme as Figuras 15 e 16. Para obter o diagrama de Bode, a simulação é realizada na frequência, simulação do tipo AC. É utilizada a mesma conexão do circuito da Figura 12.

Figura 15 – Obtenção do diagrama de Bode no *LTspice*



Fonte: Da autora.

Figura 16 – Identificação da frequência de corte em −3 dB no *LTspice*

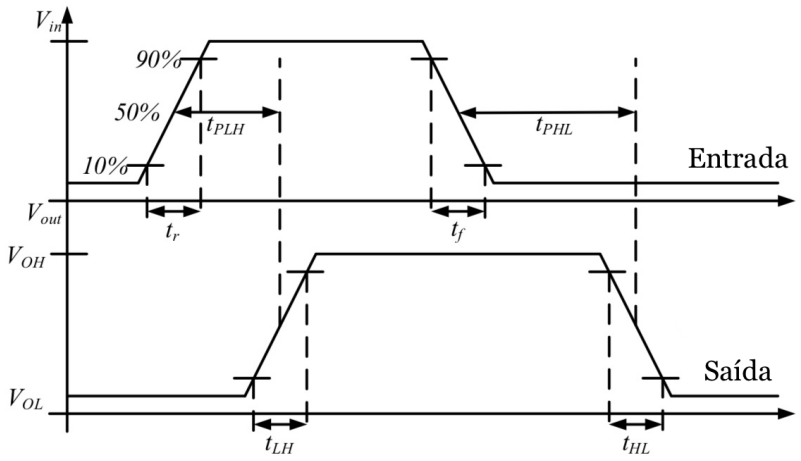


Fonte: Da autora.

3.1.3 Tempo de atraso

O tempo de atraso é medido aplicando uma onda do tipo degrau ou quadrada na entrada do circuito criado com a interconexão e medindo o tempo necessário para que o sinal de saída responda à transição de entrada, similar à Figura 17.

Figura 17 – Definição do tempo de atraso



Fonte: Adaptado de Cisneros *et al.* (2014).

3.2 LEVANTAMENTO DE DADOS REFERENTES AO CIRCUITO IDEAL

Para analisar a interferência das interconexões no desempenho da rede neural são coletados dados referentes ao circuito ideal conforme Seção 2.1.3, sem a inclusão das interconexões. Esse levantamento serve como base de referência para comparar os efeitos das interconexões nas simulações subsequentes. As simulações são similares às apresentadas na Seção 3.1.

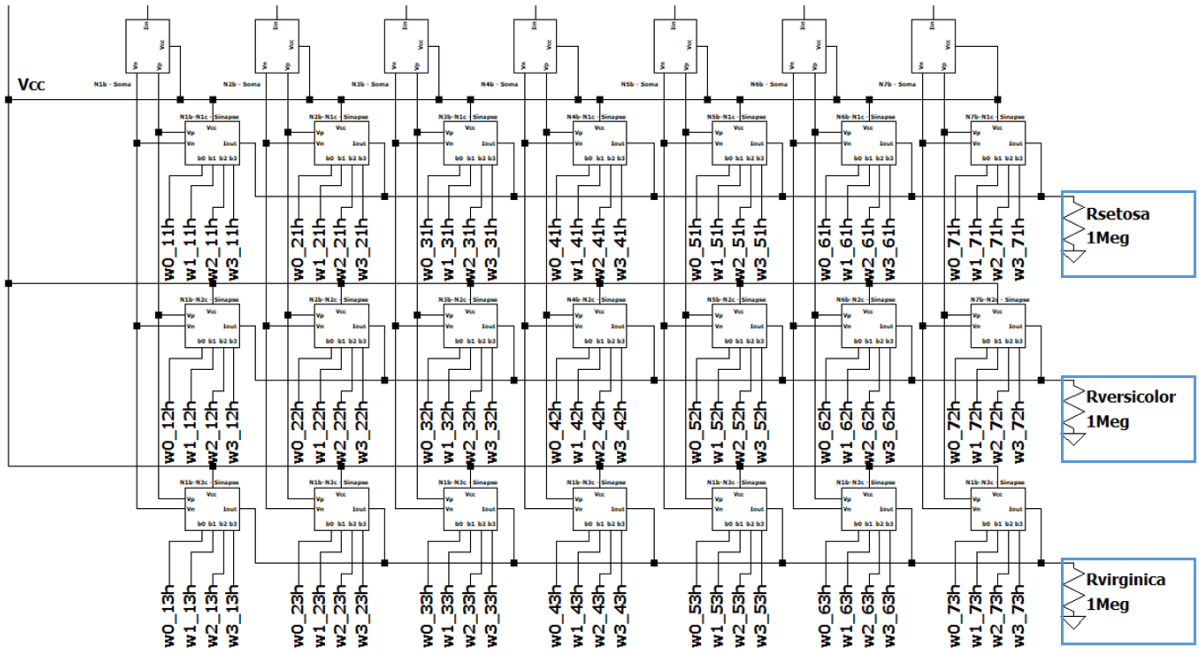
3.2.1 Dissipação de potência

Para o levantamento de dados a respeito da dissipação de potência, o circuito é simulado no tempo e os valores dos circuitos de soma de entrada, sinapse de entrada, soma da camada escondida e sinapse da camada escondida são somados. É considerado o valor máximo de potência de cada bloco para obter o valor total do circuito.

3.2.2 Frequência crítica

A frequência crítica foi identificada utilizando o diagrama de Bode das cargas de saída da rede, coletando dados na frequência de corte de -3 dB e escolhendo o menor valor entre eles. Para se obter o diagrama de Bode, foi realizada a simulação na frequência. As cargas de saída são as resistências denominadas *Rsetosa*, *Rversicolor* e *Rvirginica* conforme Figura 18.

Figura 18 – Cargas de saída no circuito da rede neural



Fonte: Adaptado de França (2023)

3.2.3 Tempo de atraso

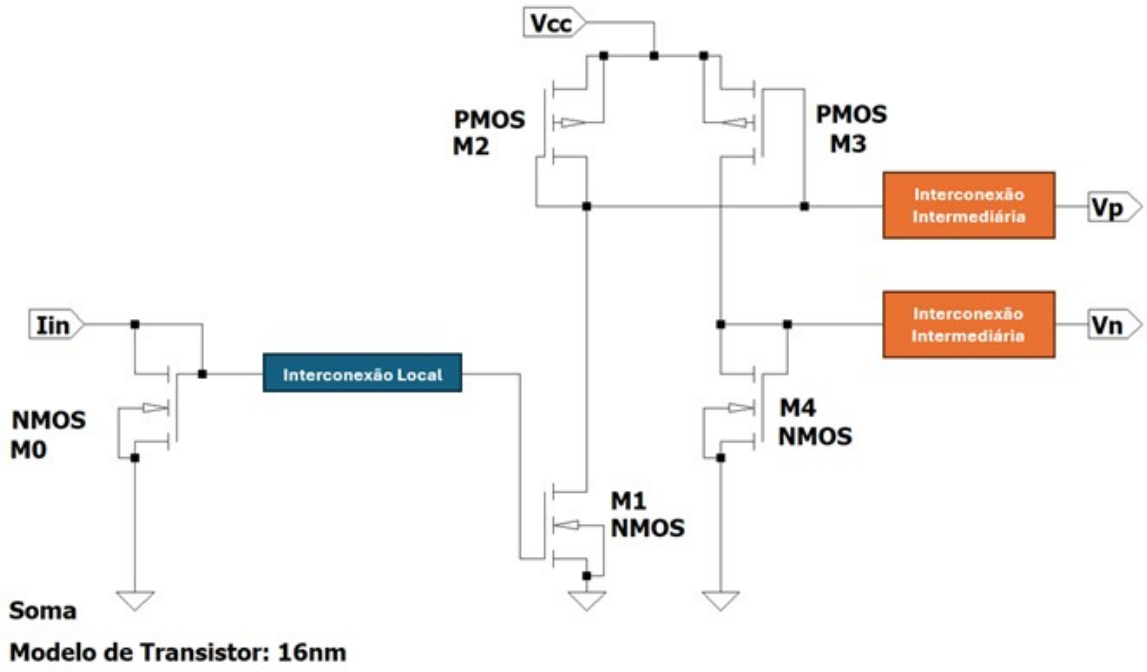
Para medir o tempo de atraso, foi aplicada uma onda do tipo degrau ou quadrada na entrada do circuito, medindo-se o tempo necessário para que o sinal de saída respondesse à transição de entrada.

3.3 INSERÇÃO DAS INTERCONEXÕES NO CIRCUITO IDEAL

Para o posicionamento das interconexões no circuito ideal é utilizada a classificação de Baohui Xu *et al.* (2022) apresentada na Seção 2.3.4.

A inserção das interconexões no circuito de soma é mostrada pela Figura 19. É válido citar que a rede neural na qual o estudo é realizado possui 11 circuitos de soma.

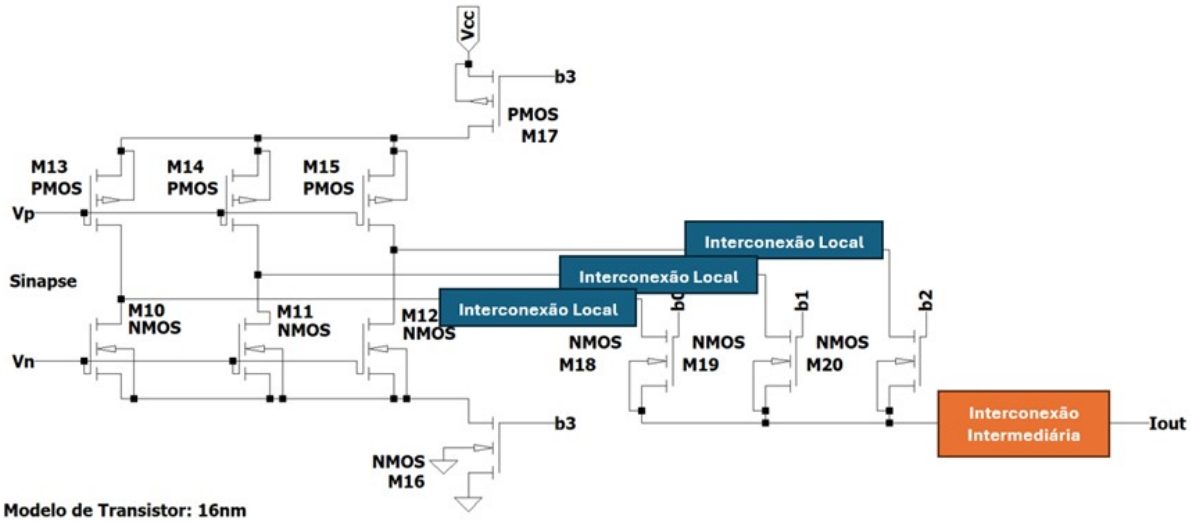
Figura 19 – Inserção das interconexões no circuito de soma



Fonte: Adaptado de França (2023)

A inserção das interconexões no circuito de sinapse é mostrada pela Figura 20. A rede neural em que o estudo é realizado possui 49 circuitos de sinapse.

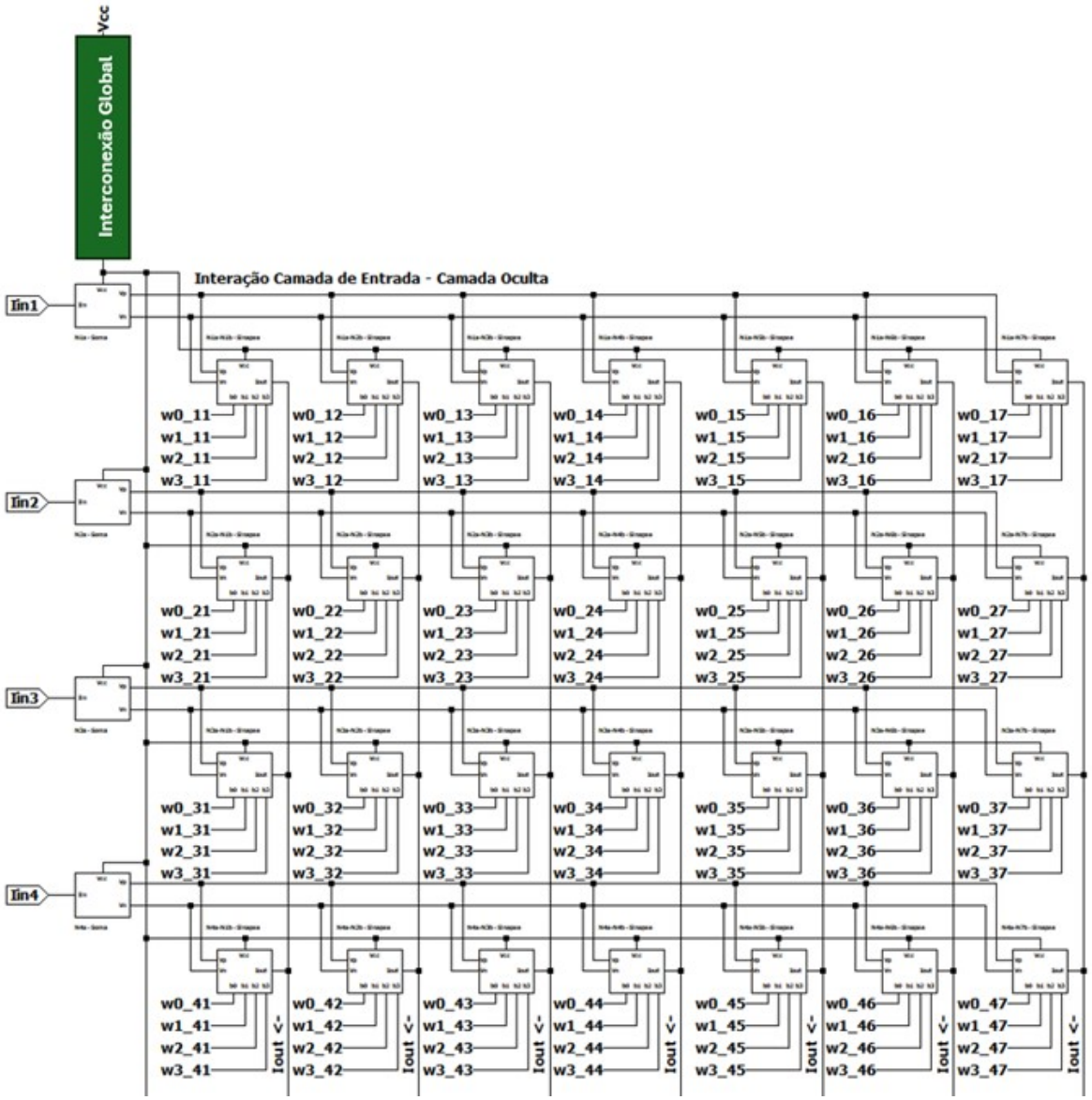
Figura 20 – Inserção das interconexões no circuito de sinapse



Fonte: Adaptado de França (2023)

Por fim, a inserção da interconexão global no circuito da rede neural é apresentada pela Figura 21.

Figura 21 – Inserção das interconexões no circuito da rede neural



Fonte: Adaptado de França (2023)

Com os dados das simulações em mãos, é realizada uma análise detalhada dos resultados, agora incluindo as interconexões. As comparações são feitas com o circuito ideal para avaliar o impacto das interconexões em termos de potência, frequência e tempo de atraso. Essa metodologia permite uma avaliação abrangente dos efeitos das interconexões no desempenho do circuito, fornecendo dados essenciais para futuras otimizações.

3.4 DEFINIÇÃO DA TOLERÂNCIA NOS PARÂMETROS DE DESEMPENHO

Alguns parâmetros são estabelecidos para a inserção das interconexões no circuito da rede neural MLP. Admite-se que o desempenho da frequência de -3 dB não pode diminuir abaixo de 70% do valor do circuito ideal. Além disso, o consumo de potência das interconexões e repetidores não pode ser maior que 15% do valor do circuito ideal. O mesmo limiar é usado para o tempo de atraso, que com o efeito das interconexões não deve ser superior a 115% do atraso do circuito ideal. A Tabela 2 mostra os limiares adotados.

Tabela 2 – Limiares dos parâmetros de desempenho ao inserir interconexões.

Parâmetro	**Limiar** [%]
Frequência de corte -3 dB	-30
Consumo de potência	+15
Tempo de atraso	+15

Fonte: Da Autora.

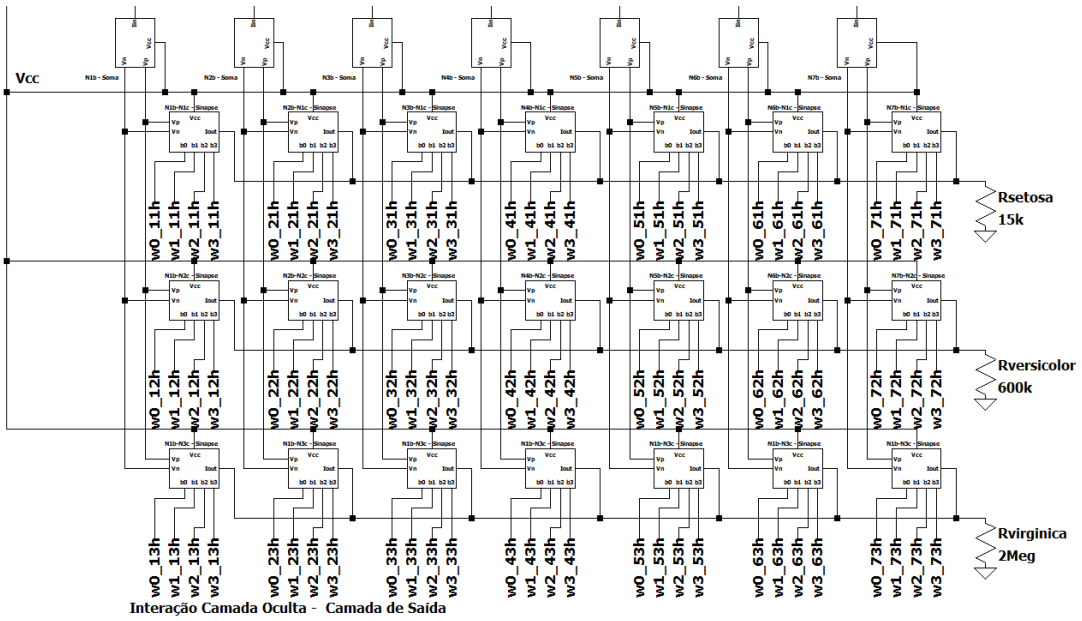
4 ANÁLISE E RESULTADOS

4.1 DADOS DO CIRCUITO SEM A INSERÇÃO DE INTERCONEXÕES

Ao avaliar a resposta em frequência do circuito da rede neural, constatou-se que a largura de banda depende significativamente das cargas de saída consideradas para o circuito. Com isso, foram realizadas simulações empíricas variando as cargas para determinar os valores mínimo e máximo dessa variação da frequência de corte. A partir dessas frequências extremas, foi realizada a análise das interconexões, possibilitando um entendimento mais aprofundado sobre o comportamento do circuito em diferentes condições de carga e sua influência na performance da rede neural.

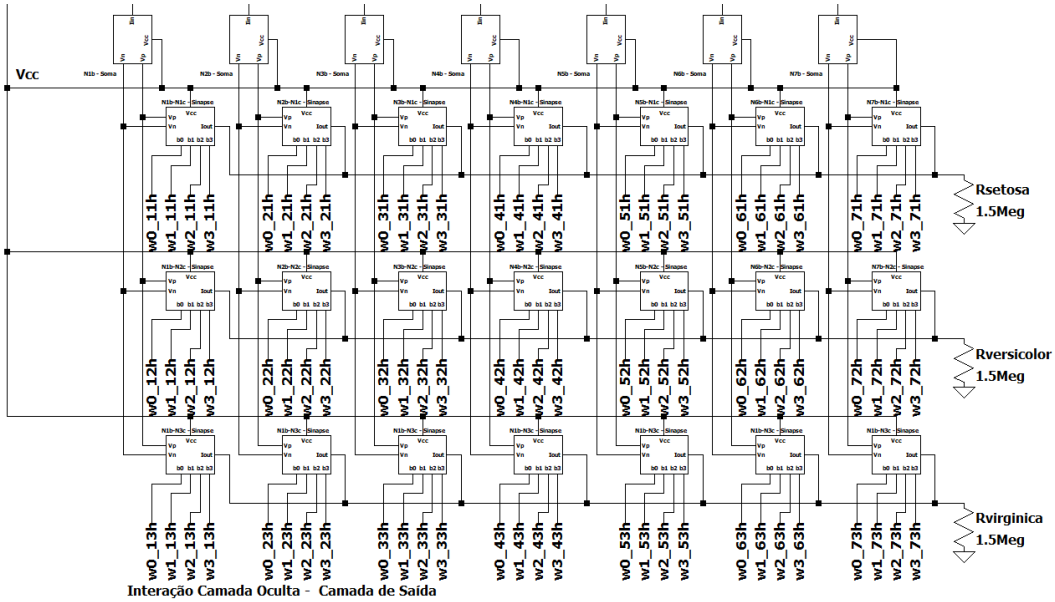
Sendo assim, foram considerados dois conjuntos de cargas de saída que geram os extremos das frequências críticas encontradas. O primeiro conjunto maximiza a frequência crítica e consiste nas cargas de saída dadas como $Rsetosa = 15\ k\Omega$, $Rversicolor = 600\ k\Omega$ e $Rvirginica = 2\ M\Omega$, conforme Figura 22. Já o segundo conjunto minimiza a frequência crítica e consiste nas cargas dadas como $Rsetosa = Rversicolor = Rvirginica = 1,5\ M\Omega$ conforme Figura 23.

Figura 22 – Conjunto 1 de cargas de saída para frequência crítica máxima.



Fonte: Da autora.

Figura 23 – Conjunto 2 de cargas de saída para frequência crítica mínima.



Fonte: Da autora.

A partir dessas duas possibilidades extremas fixadas, simulou-se a rede neural tanto no tempo, quanto na frequência e foram obtidos os resultados presentes nas Tabelas 3, 4 e 5.

Tabela 3 – Dissipação de potência: Rede Neural sem interconexões - Conjunto 1

Parte do circuito	**Potência [μW]**
Somas camada de entrada	0,047
Somas camada oculta	2,882
Sinapses camada de entrada	0,366
Sinapses camada oculta	12,532
Total	15,827

Fonte: Da Autora.

Tabela 4 – Dissipação de potência: Rede Neural sem interconexões - Conjunto 2

Parte do circuito	**Potência [μW]**
Somas camada de entrada	0,047
Somas camada oculta	2,882
Sinapses camada de entrada	0,366
Sinapses camada oculta	3,271
Total	6,566

Fonte: Da Autora.

Tabela 5 – Parâmetros de desempenho: Rede Neural sem interconexões

Cargas de saída	**Atraso** [n**s**]	**Frequência crítica** [M**Hz**]	**Potência total** [μ**W**]
Conjunto 1	666,667	190,429	15,827
Conjunto 2	624,578	78,052	6,566

Fonte: Da Autora.

O conjunto 1 refere-se às cargas que geram a maior frequência crítica encontrada empiricamente e o conjunto 2 às cargas que geram a menor frequência crítica encontrada empiricamente.

Ao analisar os dados de dissipação de potência apresentados nas Tabelas 3 e 4, observa-se que o conjunto 1 de cargas apresenta um consumo de potência significativamente maior nos circuitos de sinapses da camada oculta em comparação ao conjunto 2. Esta diferença é explicada pelo fato de que cargas mais baixas exigem uma corrente maior, resultando em um aumento no consumo de potência, bem como pelo fato de que os circuitos de sinapses possuem mais conexões na camada oculta e as cargas estão diretamente conectadas a eles.

Por fim, ao analisar a Tabela 5, é possível inferir que as variações das cargas praticamente não impactaram o tempo de atraso do circuito. No entanto, conforme esperado, essa variação no conjunto de cargas trouxe os dois extremos empiricamente possíveis para a frequência de corte do circuito, sem a inclusão das interconexões.

4.2 DADOS OBTIDOS SOBRE AS INTERCONEXÕES DE COBRE

A Tabela 6 mostra os dados obtidos a partir das simulações das interconexões locais conforme a metodologia escolhida.

Tabela 6 – Parâmetros de desempenho: Interconexões locais.

Comprimento [n**m**]	**Atraso** [p**s**]	**Frequência crítica** [G**Hz**]	**Potência** [p**W**]
10	0,684	903,975	27,652
25	0,895	362,556	52,399
50	1,143	177,828	69,057
100	1,252	89,125	67,314
250	1,850	35,476	69,057
500	6,765	18,034	110,830
1000	13,297	8,811	115,100

Fonte: Da Autora.

As Tabelas 7 e 8 demonstram o quanto estas interconexões locais impactam individualmente nos parâmetros de análise da rede neural.

Tabela 7 – Impacto das interconexões locais nos parâmetros de desempenho - Conjunto 1 de cargas

Comprimento [n**m**]	**Atraso** [%]	**Frequência crítica** [%]	**Potência** [%]
10	+0,00010	0	+0,00017
25	+0,00013	0	+0,00033
50	+0,00017	0	+0,00044
100	+0,00019	0	+0,00042
250	+0,00028	0	+0,00044
500	+0,00101	0	+0,00070
1000	+0,00199	0	+0,00072

Fonte: Da Autora.

Tabela 8 – Impacto das interconexões locais nos parâmetros de desempenho - Conjunto 2 de cargas

Comprimento [n**m**]	**Atraso** [%]	**Frequência crítica** [%]	**Potência** [%]
10	+0,00011	0	+0,00042
25	+0,00014	0	+0,00080
50	+0,00018	0	+0,00105
100	+0,00020	0	+0,00102
250	+0,00030	0	+0,00105
500	+0,00108	0	+0,00169
1000	+0,00213	0	+0,00175

Fonte: Da Autora.

Ao analisar as Tabelas 6, 7 e 8 é possível perceber que o consumo de potência das interconexões locais analisadas separadamente praticamente não impacta na dissipação total de potência da rede neural. Da mesma forma, os tempos de atraso isoladamente são tão pequenos que também não impactam muito no tempo de atraso total da rede neural. Entretanto, há uma ressalva, a rede neural possui 11 circuitos de soma e 49 circuitos de sinapse. Avaliando por esse ponto de vista, há a possibilidade das interconexões impactarem de forma razoável estes parâmetros quando inseridas nos circuitos conforme a metodologia apresentada.

Por fim, as frequências de corte são superiores à frequência crítica do circuito original, logo não apresentam impacto. Em resumo, qualquer uma das interconexões locais pode ser inserida sem prejudicar drasticamente o desempenho global da rede neural.

Repetindo o processo, a Tabela 9 explana os dados obtidos a partir das simulações das interconexões intermediárias.

Tabela 9 – Parâmetros de desempenho: Interconexões intermediárias.

Comprimento [μ**m**]	**Atraso** [p**s**]	**Frequência crítica** [G**Hz**]	**Potência** [p**W**]
2	26,050	4,399	129,670
5	64,152	1,730	103,460
10	139,211	0,864	105,930
25	359,629	0,339	113,420
50	730,858	0,158	125,650
75	1148,491	0,100	122,570

Fonte: Da Autora.

As Tabelas 10 e 11 demonstram o impacto unitário das interconexões intermediárias nos parâmetros de análise da rede neural.

Tabela 10 – Impacto das interconexões intermediárias nos parâmetros de desempenho - Conjunto 1 de cargas

Comprimento [μ**m**]	**Atraso** [%]	**Frequência crítica** [%]	**Potência** [%]
2	+0,00390	0	+0,00082
5	+0,00962	0	+0,00065
10	+0,02088	0	+0,00067
25	+0,05394	0	+0,00072
50	+0,10962	-17,029	+0,00079
75	+0,17227	-52,513	+0,00077

Fonte: Da Autora.

Tabela 11 – Impacto das interconexões intermediárias nos parâmetros de desempenho - Conjunto 2 de cargas

Comprimento [μ**m**]	**Atraso** [%]	**Frequência crítica** [%]	**Potência** [%]
2	+0,00417	0	+0,00197
5	+0,01027	0	+0,00158
10	+0,02229	0	+0,00161
25	+0,05758	0	+0,00173
50	+0,11702	0	+0,00191
75	+0,18388	0	+0,00187

Fonte: Da Autora.

Ao analisar as Tabelas 9, 10 e 11 é possível perceber novamente que o consumo de potência e o tempo de atraso das interconexões intermediárias analisadas separadamente praticamente não impacta na dissipação total de potência da rede neural. Entretanto, a ressalva a respeito da quantidade de circuitos de soma e de sinapse permanece.

Por fim, as frequências de corte podem ser problemáticas a partir de 50μ m dependendo do conjunto de cargas de saída escolhido. Para o caso do conjunto 1 de cargas, devem ser utilizadas as interconexões intermediárias de até 50μ m para respeitar até

30% de queda na frequência de corte em −3 dB. Já para o caso do conjunto 2, todas as interconexões intermediárias podem ser utilizadas sem impactar no desempenho geral do sistema.

Finalizando as simulações, a Tabela 12 apresenta os dados obtidos a partir das simulações das interconexões globais.

Tabela 12 – Parâmetros de desempenho: Interconexões globais.

Comprimento [μ**m**]	**Atraso** [n**s**]	**Frequência crítica** [M**Hz**]	**Potência** [p**W**]
100	1,638	72,235	153,080
125	2,165	55,358	165,210
250	6,260	22,300	222,220
375	10,640	12,529	281,650
500	16,654	8,040	341,580

Fonte: Da Autora.

As Tabelas 13 e 14 demonstram o impacto das interconexões globais isoladas nos parâmetros de análise da rede neural.

Tabela 13 – Impacto das interconexões globais nos parâmetros de desempenho - Conjunto 1 de cargas

Comprimento [μ**m**]	**Atraso** [%]	**Frequência crítica** [%]	**Potência** [%]
100	+0,24570	-62,067	+0,00097
125	+0,32475	-70,930	+0,00104
250	+0,93910	-88,290	+0,00140
375	+1,59510	-93,421	+0,00178
500	+2,49810	-95,778	+0,00216

Fonte: Da Autora.

Tabela 14 – Impacto das interconexões globais nos parâmetros de desempenho - Conjunto 2 de cargas

Comprimento [μ**m**]	**Atraso** [%]	**Frequência crítica** [%]	**Potência** [%]
100	+0,26226	-7,453	+0,00233
125	+0,34663	-29,075	+0,00252
250	+1,00228	-71,429	+0,00338
375	+1,70355	-83,948	+0,00429
500	+2,66644	-89,699	+0,00520

Fonte: Da Autora.

Analisando as Tabelas 12, 13 e 14 é perceptível que as interconexões globais são as mais críticas por conta do maior comprimento e maior degradação do sinal. É possível notar que as potências dissipadas aumentam bastante em relação às interconexões locais, por exemplo. Os atrasos aumentam ainda mais quando comparados às interconexões menores,

justamente pela distância que o sinal necessita percorrer nas interconexões. Ainda assim, esses valores são muito pequenos ao observar o impacto que geram na rede completa. Para esse caso, a ressalva expressa anteriormente é inválida, haja vista haver apenas 1 circuito em que essa interconexão é utilizada.

O maior problema dessas interconexões se dá na largura de banda que elas impõem ao sistema. A implementação da interconexão global no caso do conjunto 1 só é possível com o auxílio de repetidores, dividindo a interconexão global em algumas interconexões intermediárias. Já no caso do conjunto 2, tem-se duas opções de interconexões globais que respeitam os limites impostos para avaliação do desempenho, a interconexão de 100 μm e a interconexão de 125 μm.

4.3 SUGESTÕES DE INSERÇÕES PARA O CONJUNTO 1 DE CARGAS

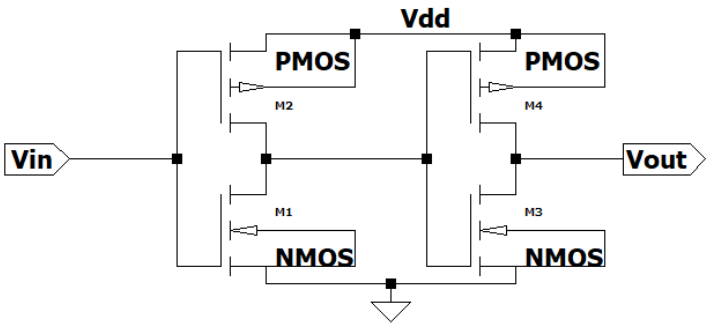
Nesta seção são apresentadas duas possibilidades de escolha de interconexões para a inserção na rede com o conjunto 1 de cargas de saída atendendo aos limiares propostos.

O tempo de atraso referente à inserção das interconexões pode ser calculado somando os atrasos do maior caminho possível do sistema em que o sinal pode passar. Analisando o circuito, percebe-se que o maior caminho possível, tanto para o circuito de soma quanto para o circuito de sinapse, inclui uma interconexão local e uma interconexão intermediária para cada caso. Já para o circuito da rede completa, tem-se apenas uma interconexão global. É importante lembrar que a rede neural em questão possui 11 circuitos de soma e 49 circuitos de sinapse.

Similarmente ao tempo de atraso, a dissipação de potência é calculada somando a dissipação de potência de todas as interconexões inseridas em todos os circuitos da rede neural e não apenas as do maior caminho possível. Por fim, a frequência crítica da rede neural com as interconexões é a menor frequência de corte em -3 dB entre todas.

Os valores dos parâmetros necessários dos repetidores estão inclusos na Tabela 15 e foram obtidos por simulação similarmente às interconexões. A Figura 24 mostra a implementação do repetidor no *software LTspice*.

Figura 24 – Implementação do repetidor no *LTspice*



Fonte: Da autora.

Tabela 15 – Parâmetros de desempenho: Repetidores

Atraso [ps]	**Potência total** [pW]
32,154	232,54

Fonte: Da Autora.

4.3.1 Interconexão local 1000 nm, Interconexão intermediária 25 μm e interconexão global dividida em interconexões de 25 μm associadas a repetidores

Para o cálculo do tempo de atraso, considera-se apenas o maior caminho possível a ser percorrido pelo sinal. Neste caso, a interconexão global de 100 μm é dividida em quatro interconexões intermediárias de 25 μm, por conta disso, deve-se substituir a interconexão global por quatro interconexões intermediárias e somar os valores de três repetidores. Tem-se então o tempo de atraso dado por $t_{d\mathbf{Total}} = 11 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + 49 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + 4 \cdot t_{d\mathbf{I}} + 3 \cdot t_{d\mathbf{R}} = 23,910$ ns, onde $t_{d\mathbf{L}}$ refere-se ao tempo de atraso da interconexão local, $t_{d\mathbf{I}}$ ao tempo de atraso da interconexão intermediária, $t_{d\mathbf{G}}$ ao tempo de atraso da interconexão global, $t_{d\mathbf{R}}$ ao tempo de atraso do repetidor e, por fim, $t_{d\mathbf{Total}}$ ao tempo de atraso do total da rede neural.

A dissipação de potência é calculada de forma similar considerando todas as interconexões de todos os circuitos como $P_{\mathbf{Total}} = 11 \cdot (P_{\mathbf{L}} + 2 \cdot P_{\mathbf{I}}) + 49 \cdot (3 \cdot P_{\mathbf{L}} + P_{\mathbf{I}}) + 4 \cdot P_{\mathbf{I}} + 3 \cdot P_{\mathbf{R}} = 27,381$ nW. A variável $P_{\mathbf{L}}$ refere-se à dissipação de potência da interconexão local, $P_{\mathbf{I}}$ à dissipação de potência da interconexão intermediária, $P_{\mathbf{G}}$ à dissipação de potência da interconexão global e, por fim, $P_{\mathbf{R}}$ à dissipação de potência do repetidor.

Para essa combinação a frequência de corte dominante é a das interconexões intermediárias de 25 μm, haja visto que os repetidores mantêm a largura de banda das interconexões intermediárias escolhidas. Logo, a frequência de corte dessa combinação é dada por $f = 339$ MHz.

A Tabela 16 mostra o impacto da inserção de todas as interconexões nos parâmetros da rede neural em estudo.

Tabela 16 – Limiares dos parâmetros de desempenho ao inserir interconexões.

Parâmetro	**Limiar** [%]
Frequência de corte −3 dB	0
Consumo de potência	+0,173
Tempo de atraso	+3,586

Fonte: Da Autora.

Com essa combinação de interconexões os limites dos parâmetros foram respeitados. O consumo de potência total e o tempo de atraso acrescidos na rede neural não representaram 15% dos valores do circuito original. Da mesma forma, a largura de banda não interferiu no valor original por ser inclusive maior do que o da rede neural sem interconexões.

4.3.2 Interconexão local 1000 nm, Interconexão intermediária 50 μm e interconexão global dividida em interconexões de 50 μm associadas a repetidor

Para essa combinação, a interconexão global de 100 μm é dividida em duas interconexões intermediárias de 50 μm e, por esse motivo, deve-se substituir a interconexão global por duas interconexões intermediárias e somar os valores de um repetidor. O tempo de atraso é dado por $t_{d\mathbf{Total}} = 11 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + 49 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + 2 \cdot t_{d\mathbf{I}} + t_{d\mathbf{R}} = 46,143$ ns. A dissipação de potência é calculada de forma similar por $P_{\mathbf{Total}} = 11 \cdot (P_{\mathbf{L}} + 2 \cdot P_{\mathbf{I}}) + 49 \cdot (3 \cdot P_{\mathbf{L}} + P_{\mathbf{I}}) + 2 \cdot P_{\mathbf{I}} + P_{\mathbf{R}} = 27,591$ nW.

Para essa combinação, a frequência de corte dominante é a das interconexões intermediárias de 50 μm, pois os repetidores mantém a largura de banda das interconexões intermediárias. Logo a frequência de corte desta combinação é dada por $f = 158$ MHz.

A Tabela 17 mostra o impacto da inserção de todas as interconexões nos parâmetros da rede neural em estudo.

Tabela 17 – Limiares dos parâmetros de desempenho ao inserir interconexões.

Parâmetro	**Limiar** [%]
Frequência de corte −3 dB	-17,029
Consumo de potência	+0,174
Tempo de atraso	+6,921

Fonte: Da Autora.

Similar à combinação apresentada na Seção 4.3.1, os limites dos parâmetros foram respeitados. O consumo de potência das interconexões e do repetidor, e o tempo de atraso acrescido pelos mesmos não representaram 15% dos valores do circuito original. Já em

relação à largura de banda é perceptível que houve uma queda considerável, no entanto, ainda dentro das tolerâncias impostas, se mantendo em $82,971\%$ do valor prévio.

Fazendo uma análise global, é possível notar que a inserção de interconexões maiores impacta muito mais no tempo de atraso do circuito e na frequência crítica do que no consumo de potência. Avaliando também as duas combinações propostas, a primeira apresentada demonstra um desempenho melhor quando comparada com a presente combinação. Para a primeira combinação, o tempo de atraso do sistema é consideravelmente menor, o consumo de potência é praticamente o mesmo e a frequência de corte em -3 dB do circuito da rede neural original não é impactada pelas interconexões inseridas.

4.4 SUGESTÕES DE INSERÇÕES DAS INTERCONEXÕES PARA O CONJUNTO 2 DE CARGAS

Nesta seção são apresentadas, analogamente à seção anterior, duas possibilidades de escolha de interconexões para a inserção na rede com o conjunto 2 de cargas de saída atendendo aos limiares propostos. Para estas combinações não é necessário fazer o uso de repetidores.

Novamente, o tempo de atraso referente à inserção das interconexões pode ser calculado somando os atrasos do maior caminho possível do sistema em que o sinal pode passar. De modo semelhante ao tempo de atraso, a dissipação de potência é calculada somando a dissipação de potência de todas as interconexões inseridas em todos os circuitos da rede neural e não apenas as do maior caminho possível. Por fim, a frequência crítica da rede neural com as interconexões é a menor frequência de corte em -3 dB entre todas.

4.4.1 Interconexão local 1000 nm, Interconexão intermediária 75 μm e interconexão global 100 μm

O tempo de atraso é calculado por $t_{d\mathbf{Total}} = 11 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + 49 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + t_{d\mathbf{G}} = 71,345$ ns. A dissipação de potência é calculada de forma similar, $P_{\mathbf{Total}} = 11 \cdot (P_{\mathbf{L}} + 2 \cdot P_{\mathbf{I}}) + 49 \cdot (3 \cdot P_{\mathbf{L}} + P_{\mathbf{I}}) + P_{\mathbf{G}} = 27,041$ nW. Por fim, a frequência crítica que prevalece é a menor entre as interconexões. Para essa combinação a frequência de corte dominante é a da interconexão global de 100 μm, dada por $72,235$ MHz.

A Tabela 18 mostra o impacto da inserção de todas as interconexões nos parâmetros da rede neural em estudo.

Tabela 18 – Limiares dos parâmetros de desempenho ao inserir interconexões.

Parâmetro	**Limiar** [%]
Frequência de corte -3 dB	-7,453
Consumo de potência	+0,412
Tempo de atraso	+11,423

Fonte: Da Autora.

Os limiares previstos dos parâmetros foram respeitados. O consumo de potência e o tempo de atraso agregados pelas interconexões inseridas não representam 15% dos valores do circuito original. A frequência crítica em -3 dB se manteve em $92,547\%$ do valor inicial.

Analisando outro ponto, é notável que o tempo de atraso aumenta muito conforme as interconexões também aumentam, ou seja, o impacto começa a ser tornar alarmante ao chegar perto dos limites propostos. O consumo de potência praticamente não varia com o tamanho das interconexões; no entanto, o percentual do impacto aumenta por conta de que a rede neural com o conjunto de cargas 2 de saída possui um consumo de potência muito menor do que com o conjunto de cargas 1. Neste caso, a dissipação de potência inclusive ficou menor do que as duas combinações anteriores apresentadas na Seção 3 por conta dos repetidores terem um consumo de potência considerável agregado.

4.4.2 Interconexão local 1000 nm, Interconexão intermediária 75 μm e interconexão global 125 μm

O tempo de atraso é calculado por $t_{d\mathbf{Total}} = 11 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + 49 \cdot (t_{d\mathbf{L}} + t_{d\mathbf{I}}) + t_{d\mathbf{G}} = 71,872$ ns. Novamente, a dissipação de potência é calculada de forma similar, $P_{\mathbf{Total}} = 11 \cdot (P_{\mathbf{L}} + 2 \cdot P_{\mathbf{I}}) + 49 \cdot (3 \cdot P_{\mathbf{L}} + P_{\mathbf{I}}) + P_{\mathbf{G}} = 27,053$ nW. Por fim, a frequência crítica que prevalece é a menor entre as interconexões. Para essa combinação a frequência de corte dominante é a da interconexão global de 125 μm, dada por $f = 55,358$ MHz.

Tabela 19 – Limiares dos parâmetros de desempenho ao inserir interconexões.

Parâmetro	**Limiar** [%]
Frequência de corte -3 dB	-29,075
Consumo de potência	+0,412
Tempo de atraso	+11,507

Fonte: Da Autora.

Por fim, com a última sugestão de combinação de interconexões, os limites dos parâmetros também foram respeitados. O consumo de potência das interconexões e do repetidor e o tempo atraso acrescido pelos mesmos não representaram 15% dos valores do circuito original. A frequência crítica em -3 dB teve uma queda maior, ainda assim mantendo-se em $70,925\%$ do valor original, o que é extremamente próxima do limite estipulado.

Novamente fazendo uma análise mais generalizada, é possível notar que a inserção de interconexões maiores impacta muito mais no tempo de atraso do circuito e na frequência crítica do que no consumo de potência. Avaliando também as duas combinações propostas, a primeira apresentada demonstra um desempenho melhor quando comparada com a presente combinação, por conta da frequência de corte em -3 dB ser levemente impactada, enquanto na presente combinação esta é drasticamente impactada.

Observando todos os resultados obtidos, pode-se afirmar que as interconexões de fato influem na degradação do sinal propagado ao notar que a largura de banda decai drasticamente com o aumento do comprimento da interconexão. Dentre os parâmetros avaliados, o segundo que é mais alarmante é o tempo de atraso do sinal, que também se intensifica com o uso de interconexões muito compridas. A dissipação de potência foi impactada de maneira muito mais branda. Os resultados encontrados na presente monografia estão de acordo com resultados obtidos em trabalhos anteriores como os de Ratnam e Srinivasarao (2015), de Dhillon e Singh (2014) e de Alam *et al.* (2009) considerando a tecnologia de $16n$ m empregada e o comprimento das interconexões.

Os impactos podem ser reduzidos não apenas escolhendo as interconexões adequadas para o projeto em questão, mas também projetando adequadamente as cargas de saída do sistema. O impacto da escolha das cargas de saída fica muito claro ao se comparar os resultados obtidos para ambos os conjuntos estudados. Avaliando a questão da largura de banda, outra solução é o uso de repetidores para as interconexões maiores serem fracionadas em interconexões menores, assim garantindo que o sinal seja regenerado e a largura de banda se preserve conforme a frequência de corte das interconexões mais curtas. A ressalva do uso de repetidores é que os mesmos agregam tempo de atraso e consomem potência, isso deve ser balanceado para um desempenho satisfatório da rede neural. No entanto, os *layouts* de circuitos integrados de tecnologia $16n$ m possuem menor chance do uso de interconexões globais, pois a área total ocupada pelo circuito não demanda conexões tão extensas.

5 CONCLUSÃO

Este estudo investigou o impacto das interconexões nos parâmetros de desempenho de circuitos de redes neurais baseadas em transistores MOS de tecnologia $16n$ m. Foram analisadas as interconexões de cobre categorizadas em interconexões locais, intermediárias e globais, observando seus efeitos no consumo de potência, tempo de atraso e largura de banda do sistema. Os resultados indicaram que as interconexões locais têm um impacto mínimo no desempenho, enquanto as intermediárias e globais aumentam significativamente o tempo de atraso e a dissipação de potência, além de reduzir drasticamente a largura de banda devido ao maior comprimento e degradação do sinal. Esse estudo é crucial para avaliar as melhores formas de projetar o *layout* do circuito integrado da rede neural de forma a minimizar os efeitos negativos das interconexões.

A regeneração dos sinais por meio de repetidores mostrou-se essencial para manter a largura de banda com valor próximo à das interconexões menores. Isso é crucial para aplicações que requerem altas taxas de transferência de dados e baixa latência, assegurando uma comunicação eficiente dentro da rede neural. As simulações de variação de carga ajudaram a identificar os valores extremos de variação de frequência da rede neural, proporcionando uma visão clara sobre como otimizar o design dessas redes.

Em síntese, esta monografia forneceu uma base sólida para futuras pesquisas e melhorias em interconexões de circuitos integrados, destacando a importância de se considerar diferentes tipos de interconexões na construção de sistemas eficientes e robustos. Para trabalhos futuros, sugere-se realizar a análise de circuitos com outras formas de tecnologias de transistores e de interconexões e avaliar a aplicação desse estudo no conceito de redes em chip. Este estudo é parte de um projeto que objetiva chegar à topologia de redes em chip, visando otimizar ainda mais a interconexão e o desempenho de redes neurais.

REFERÊNCIAS

ALAM, Naushad; KURESHI, Abdul Kadir; HASAN, Mohd; ARSLAN, Tughrul. Carbon nanotube interconnects for low-power high-speed applications. *In*: IEEE. 2009 IEEE International Symposium on Circuits and Systems. [*S.l.*: *s.n.*], 2009. P. 2273–2276.

ANALOG DEVICES, Inc. **LTspice Simulator**. [*S.l.*: *s.n.*], 2023. Disponível em: https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html.

BAKOGLU, H Buman. Circuits, interconnections, and packaging for VLSI, 1990.

BENINI, Luca; DE MICHELI, Giovanni. Networks on chips: A new SoC paradigm. **computer**, IEEE, v. 35, n. 1, p. 70–78, 2002.

BINAS, Jonathan; NEIL, Dan; INDIVERI, Giacomo; LIU, Shih-Chii; PFEIFFER, Michael. Precise neural network computation with imprecise analog devices, p. 22, fev. 2020.

BINAS, Jonathan; NEIL, Daniel; INDIVERI, G.; LIU, Shih-Chii; PFEIFFER, Michael. Precise deep neural network computation on imprecise low-power analog hardware. **ArXiv**, abs/1606.07786, 2016.

BJERREGAARD, Tobias; MAHADEVAN, Shankar. A survey of research and practices of network-on-chip. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 38, n. 1, 1–es, 2006.

CHENG, Tao; WANG, Jiaqiu; LI, Xia. Space-time series forecasting by artificial neural networks. *In*: SPIE. INTERNATIONAL conference on earth observation data processing and analysis (ICEODPA). [*S.l.*: *s.n.*], 2008. v. 7285, p. 1057–1064.

CISNEROS, Susana Ortega; PANDURO, Juan Jose Raygoza; BARON, Jose Roberto Reyes; ARANDA, B Daniel Tonali; CASILLAS, Z Antonio. Characterization technique to implement self-timed cells for VLSI design blocks. *In*: IEEE. 2014 11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). [*S.l.*: *s.n.*], 2014. P. 1–6.

DALLY, William J; TOWLES, Brian. Route packets, not wires: on-chip inteconnection networks. *In*: PROCEEDINGS of the 38th annual design automation conference. [*S.l.*: *s.n.*], 2001. P. 684–689.

DAS, Debaprasad; RAHAMAN, Hafizur. Timing analysis in carbon nanotube interconnects with process, temperature, and voltage variations. *In*: IEEE. 2010 International Symposium on Electronic System Design. [*S.l.*: *s.n.*], 2010. P. 27–32.

DHILLON, Gurleen; SINGH, Karamjit. Evaluation and Comparison of Single-Wall Carbon Nanotubes and Copper as VLSI Interconnect. **The International Journal of Innovative Research in Computer and Communication Engineering**, v. 2, n. 5, 2014.

DUARTE, Renan Rodrigo *et al.* Estudo comparativo entre semicondutores de silıcio e nitreto de gálio em circuitos de acionamento de leds. Universidade Federal de Santa Maria, 2017.

FORSSELL, Mats. Hardware implementation of artificial neural networks. **Information flow in networks**, v. 18, n. 2014, p. 1–4, 2014.

FRANÇA, Luiz Augusto Scheuermann. **Computação neuromórfica em nanoescala**: desenvolvimento de arquitetura NoC baseada em tecnologia MOS para redes neurais. 2023. Universidade Federal de Santa Catarina.

FRYE, Robert C; RIETMAN, Edward A; WONG, Chee C. Back-propagation learning and nonidealities in analog neural network hardware. **IEEE transactions on neural networks**, IEEE, v. 2, n. 1, p. 110–117, 1991.

HÄNGGI, Martin; MOSCHYTZ, George S. **Cellular neural networks: analysis, design and optimization**. [*S.l.*]: Springer Science & Business Media, 2000.

HIKAWA, Hiroomi. FPGA implementation of self organizing map with digital phase locked loops. **Neural Networks**, Elsevier, v. 18, n. 5-6, p. 514–522, 2005.

HOLLIS, Paul W; PAULOS, John J. Artificial neural networks using MOS analog multipliers. **IEEE Journal of Solid-State Circuits**, IEEE, v. 25, n. 3, p. 849–855, 1990.

HU, Jingcao; MARCULESCU, Radu. Energy-aware mapping for tile-based NoC architectures under performance constraints. *In*: PROCEEDINGS of the 2003 Asia and South Pacific Design Automation Conference. [*S.l.*: *s.n.*], 2003. P. 233–239.

INDIVERI, Giacomo; HORIUCHI, Timothy K. Frontiers in neuromorphic engineering. **Frontiers in neuroscience**, Frontiers, v. 5, p. 13375, 2011.

JAEGER, Richard C; BLALOCK, Travis N; BLALOCK, Benjamin J. **Microelectronic circuit design**. [*S.l.*]: McGraw-Hill New York, 1997. v. 97.

JUNG, Seul; KIM, Sung su. Hardware implementation of a real-time neural network controller with a DSP and an FPGA for nonlinear systems. **IEEE Transactions on Industrial Electronics**, IEEE, v. 54, n. 1, p. 265–271, 2007.

KOO, Kyung Hoae. **The Comparison Study of Future On-chip Interconnects for High Performance VLSI Applications**. 2011. Tese (Doutorado) – Stanford University.

KUHN, Kelin J. Considerations for ultimate CMOS scaling. **IEEE transactions on Electron Devices**, IEEE, v. 59, n. 7, p. 1813–1828, 2012.

KURUVILLA, Nisha; RAINA, JP. Carbon Nanotubes-A Solution for Tera Hertz's IC Interconnect. **International Journal of Recent Trends in Engineering**, Academy Publisher, v. 1, n. 4, p. 32, 2009.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

MEAD, Carver. Analog VLSI and neutral systems. **NASA STI/Recon Technical Report A**, v. 90, p. 16574, 1989.

MERCHANT, Sailesh M; KANG, Seung H; SANGANERIA, Mahesh; SCHRAVENDIJK, Bart van; MOUNTSIER, Tom. Copper interconnects for semiconductor devices. **Jom**, Springer, v. 53, p. 43–48, 2001.

MEROLLA, Paul A *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. **Science**, American Association for the Advancement of Science, v. 345, n. 6197, p. 668–673, 2014.

MISRA, Janardan; SAHA, Indranil. Artificial neural networks in hardware: A survey of two decades of progress. **Neurocomputing**, Elsevier, v. 74, n. 1-3, p. 239–255, 2010.

MOERLAND, Perry D; FIESLER, Emile. Neural network adaptations to hardware implementations. *In*: HANDBOOK of neural computation. [*S.l.*]: CRC Press, 2020. e1–2.

MOHAMMADZADEH, Ardashir; SABZALIAN, Mohammad Hosein; CASTILLO, Oscar; SAKTHIVEL, Rathinasamy; EL-SOUSY, Fayez FM; MOBAYEN, Saleh. Multilayer Perceptron (MLP) Neural Networks. *In*: NEURAL Networks and Learning Algorithms in MATLAB. [*S.l.*]: Springer, 2022. P. 5–21.

MURALI, Srinivasan; BENINI, Luca; DE MICHELI, Giovanni. Mapping and physical planning of networks-on-chip architectures with quality-of-service guarantees. *In*: PROCEEDINGS of the 2005 Asia and South Pacific Design Automation Conference. [*S.l.*: *s.n.*], 2005. P. 27–32.

RANGEL, Edylara Ribeiro. **Estudo sobre o consumo de energia em redes-em-chip baseadas em dispositivos nanoeletrônicos**. 2018. Diss. (Mestrado) – UNIVERSIDADE DE BRASÍLIA.

RATNAM, Ridhi; SRINIVASARAO, K. Comparison of Cu Interconnects With CNT Interconnect For High Performances Applications, 2015.

RAYCHOWDHURY, Arijit; ROY, Kaushik. A circuit model for carbon nanotube interconnects: comparative study with Cu interconnects for scaled technologies. *In*: IEEE. IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004. [*S.l.*: *s.n.*], 2004. P. 237–240.

REEHAL, Gursharan Kaur. **Designing Low Power and High Performance Network-on-Chip Communication Architectures for Nanometer SoCs**. 2012. Tese (Doutorado) – The Ohio State University.

SEDRA, Adel S; SMITH, Kenneth Carless. **Microeletrônica**. [*S.l.*]: Pearson Prentice Hall, 2007.

SHEU, Bing J; SCHARFETTER, Donald L; KO, P-K; JENG, M-C. BSIM: Berkeley short-channel IGFET model for MOS transistors. **IEEE Journal of Solid-State Circuits**, IEEE, v. 22, n. 4, p. 558–566, 1987.

SRIVASTAVA, Ashok; XU, Yao; SHARMA, Ashwani K. Carbon nanotubes for next generation very large scale integration interconnects. **journal of nanophotonics**, SPIE, v. 4, n. 1, p. 041690, 2010.

SUNDARARAJAN, Narasimhan; SARATCHANDRAN, P. **Parallel architectures for artificial neural networks: Paradigms and implementations**. [*S.l.*]: IEEE Computer Society Press, 1998.

WU, Jerry; SHEN, Yin-Lin; REINHARDT, Kitt; SZU, Harold; DONG, Boqun. A Nanotechnology Enhancement to Moore s Law. **Applied Computational Intelligence and Soft Computing**, Wiley Online Library, v. 2013, n. 1, p. 426962, 2013.

XU, Baohui; CHEN, Rongmei; ZHOU, Jiuren; LIANG, Jie. Recent progress and challenges regarding carbon nanotube on-chip interconnects. **Micromachines**, MDPI, v. 13, n. 7, p. 1148, 2022.

XU, Yao; SRIVASTAVA, Ashok. A model for carbon nanotube interconnects. **International Journal of Circuit Theory and Applications**, Wiley Online Library, v. 38, n. 6, p. 559–575, 2010.

ZHANG, Lei; HAN, Yinhe; LI, Huawei; LI, Xiaowe. Fault tolerance mechanism in chip many-core processors. **Tsinghua Science and Technology**, TUP, v. 12, S1, p. 169–174, 2007.

ZHAO, Wei; CAO, Yu. Predictive technology model for nano-CMOS design exploration. **ACM Journal on Emerging Technologies in Computing Systems (JETC)**, ACM New York, NY, USA, v. 3, n. 1, 1–es, 2007.

ZIMPECK, Alexandra L; MEINHARDT, Cristina; BUTZEN, Paulo F. **Análise do comportamento de portas lógicas CMOS com falhas Stuck-On em nanotecnologias**. [*S.l.*]: ICCEEg, 2014.