UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE TECNOLOGIA
CURSO DE ENGENHARIA ELÉTRICA

MARCO ANTONIO RIOS

CONCEPÇÃO DE CÉLULAS DE MEMÓRIA SRAM EM TECNOLOGIA SEQUENCIAL 3D

Curitiba
2018

MARCO ANTONIO RIOS

CONCEPÇÃO DE CÉLULAS DE MEMÓRIA SRAM EM TECNOLOGIA SEQUENCIAL 3D

Trabalho de conclusão de curso apresentado ao Curso de Graduação em Engenharia Elétrica da Universidade Federal do Paraná como requisito parcial à obtenção do título de Engenheiro Eletricista.

Orientador: Prof. Ph.D. André Augusto Mariano
Co-orientador: Ph.D. Adam Makosiej

Curitiba
2018

TERMO DE APROVAÇÃO

MARCO ANTONIO RIOS

CONCEPÇÃO DE CÉLULAS DE MEMÓRIA SRAM EM TECNOLOGIA SEQUENCIAL 3D

Trabalho de conclusão de curso apresentado ao Curso de Graduação em Engenharia Elétrica da Universidade Federal do Paraná como requisito parcial à obtenção do título de Engenheiro Eletricista.

Prof. Ph.D. André Augusto Mariano
Departamento de Engenharia Elétrica, UFPR

Prof. Dr. Luis Henrique Assumpção Lolis
Departamento de Engenharia Elétrica, UFPR

Prof. M.Sc. Ademar Luiz Pastro
Departamento de Engenharia Elétrica, UFPR

Curitiba, 26 de novembro de 2018

Agradecimentos

Agradeço primeiramente ao meus pais, Maria e Antonio, por terem me dado todo o suporte necessário para que eu pudesse me focar ao máximo nos estudos. Agradeço todo o apoio e amor dado e por terem sempre me encorajado em minhas decisões. A eles o meu muito obrigado.

Agradeço enormemente a minha esposa, Angélica, por ter participado dessa longa caminhada comigo, onde sempre fomos muito companheiros. Agradeço por estar sempre ao meu lado, inclusive durante o intercambio, onde pudemos compartilhar experiencias incríveis.

Agraço as professores do GICS, especialmente aos professores Ph.D André Augusto Mariano pela orientação neste trabalho de conclusão de curso e todo o apoio durante a graduação, Ph.D Bernardo Leite pela orientação durante o trabalho de iniciação cientifica e a grande importância desse trabalho nas minhas conquistas e Ph.D Luis Lolis pelos momentos em que trabalhamos junto no PET e pelas palavras de incentivo. Vocês foram fundamentais em minha formação.

Finalmente, agradeço ao meu coordenador de estágio no CEA - LETI, Ph.D Adam Makosiej, que soube me orientar de forma que eu pudesse desenvolver o trabalho independentemente das dificuldades encontradas.

Resumo

Conforme a tecnologia de fabricação de transistores avançou, as barreiras foram superadas das mais diversas formas, porém essas soluções sempre tiveram uma coisa em comum: fabricação planar. Atualmente os transistores estão próximos da escala atômica, o que traz desafios que não têm perspectivas de serem superados ou a solução é muito cara. Um exemplo disso é a grande variabilidade do processo de fabricação devido às impurezas no material. Logo, afim de manter a evolução da eletrônica sob a lei de Moore, um novo paradigma é apresentado: Integração sequencial 3-D. Essa tecnologia é realidade, porém ainda muito cara para ser concorrente no mercado. Na esperança que um dia isso mude, circuitos que a muito tempo são conhecidos e já foram explorados exaustivamente podem agora voltar a superfície da pesquisa. Neste trabalho a célula da memoria SRAM foi explorada visando aplicações em tecnologias de integração 3D, são propostos configurações de tensão de corpo que são possíveis apenas com essa tecnologia.

Palavras chave: 3D CoolCube, sequencial, SRAM, Eldo, Tensão de Corpo.

Abstract

As transistor technology has advanced, the barriers have been overcome in many ways, but these solutions have always had one thing in common: planar fabrication. Currently the transistors are close to the atomic scale, which brings challenges that have no prospect of being overcome or the solution is very expensive. An example of this is the great variability of the manufacturing process due to impurities in the material. Thus, in order to maintain the evolution of electronics under Moore's law, a new paradigm is presented: 3-D sequential integration. This technology is a reality, but still very expensive to be a competitor in the market. In the hope that one day this will change, circuits that have long been known and have been explored extensively can now return to the surface of the search. In this work the SRAM memory cell was exploited for applications in 3D integration technologies, it is proposed body voltage configurations that are possible only with this technology.

Keywords: 3D CoolCube, sequencial, SRAM, Eldo, Body Bias.

Lista de Figuras

Lista de siglas

SRAM Static Random Access Memory

CEA Alternative Energies and Atomic Energy Commission

IoT Internet of Things

SoC System-on-Chip

SRAM Static random access memory

SNM Static Noise Margin

WNM Write Noise Margin

DRV Data Retention Voltage

VTC Voltage Transfer Characteristic

ULP Ultra Low Power

MC Monte Carlo

PDC Polarização Direta de Corpo

PRC Polarização Reversa de Corpo

Conteúdo

1 Introdução

Considerando a evolução da tecnologia, especialmente a miniaturização de chips e sensores eletrônicos, as melhorias na comunicação sem fio e seus preços relativamente baixos, tudo isso levou à enorme expansão de pequenos dispositivos eletrônicos. Na verdade, dispositivos conectados (inteligentes), chamados de Internet of Things (IoT), podem ser encontrados em todos os aspectos da vida humana moderna, tais como: celulares, aparelhos móveis e portáteis voltados para a saúde, automação residencial, controle de ambientes em edifícios inteligentes e transporte. Além disso, nos próximos anos espera-se que o número total desses dispositivos aumente drasticamente. Não apenas em aplicações personalizadas, mas ainda mais em negócios e manufatura, por exemplo: para rastrear estoques, gerenciar máquinas e aumentar a eficiência de fábricas. Na verdade, é impossível prever todas as aplicações potenciais nessa área, à medida que novas soluções criativas surgem a cada dia, alterando fundamentalmente a maneira como interagimos com nosso ambiente físico. [1]

Um dos maiores desafios no projeto de IoT é a criação de sistemas com eficiência energética que podem operar em baixo consumo de energia (Ultra Low Power (ULP)). Esses dispositivos geralmente são alimentados por baterias e passam a maior parte do tempo no modo de espera. Portanto, há uma grande necessidade do menor consumo possível no modo de espera, sem prejudicar o desempenho no modo ativo. O consumo no modo de espera é dominado pelas correntes parasitas nos transistores. Sabendo que com a miniaturização do CMOS as potências dissipadas pelas correntes parasitas aumentam proporcionalmente, o problema da eficiência energética em tecnologias avançadas torna-se ainda mais crítico. Como representado na fig.1, em tecnologias mais primordiais, como por exemplo no processo de 400 nm, a porcentagem da potência dissipada na forma de corrente de fuga em relação a potência total consumida nos transistores é praticamente nula. Conforme o tamanho do canal dos transistores fica menor, essa porcentagem se torna mais importante, podendo chegar a 20% no processo de 130 nm. Em processos ainda menores, esses valores podem se tornar muito preocupantes.

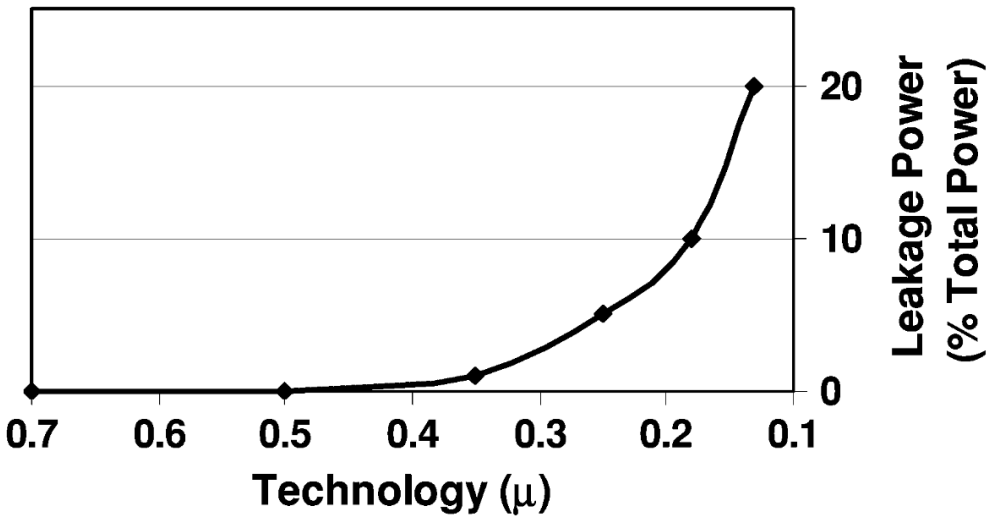


Figura 1: Aumento da participação de corrente de fuga em relação a potência total dissipada em tecnologias avançados.

Fonte: Conception de circuits memoire statique a access aleatoire tres basse tension en technologie UTBB-FDSOI - Adam Makosiej

A indústria de semicondutores tem pressionado por inovações no design de circuitos e na tecnologia de fabricação, a fim de seguir a lei de Moore. Essa lei afirma que o número de transistores dentro do mesmo sistema-em-um-chip - *System-on-Chip (SoC)* deve dobrar a cada 18 meses, como mostra a fig. 2. Essa lei vem com o principal desafio de aumentar o desempenho e reduzir o consumo de energia. Atualmente, enfrentamos grandes problemas, pois o tamanho dos transistores está atingindo um nível atômico, o que leva ao surgimento de inevitáveis fenômenos parasitários, como efeito de canal curto, variabilidade do processo e aumento das correntes de fuga. Como consequência, diferentes métodos de empilhamento são desenvolvidos para reduzir gargalos no design planar convencional e permitir densidades mais altas.

Esta tendência pode ser comparada a um arquiteto que constrói um edifício empilhando apartamentos para usar eficientemente uma pequena terra. Na microeletrônica, o equivalente é o projeto de SoCs empilhando circuitos ou transistores para aumentar

a densidade para a mesma área 2-D. Isso é chamado de design 3-D, e a fabricação é geralmente denominada integração 3-D. Entre as tecnologias monolíticas 3-D, a tecnologia LETI 3-D CoolCube sequencial oferece uma interconexão 3-D muito fina, oferecendo possibilidades reais de roteamento em 3-D. [2]

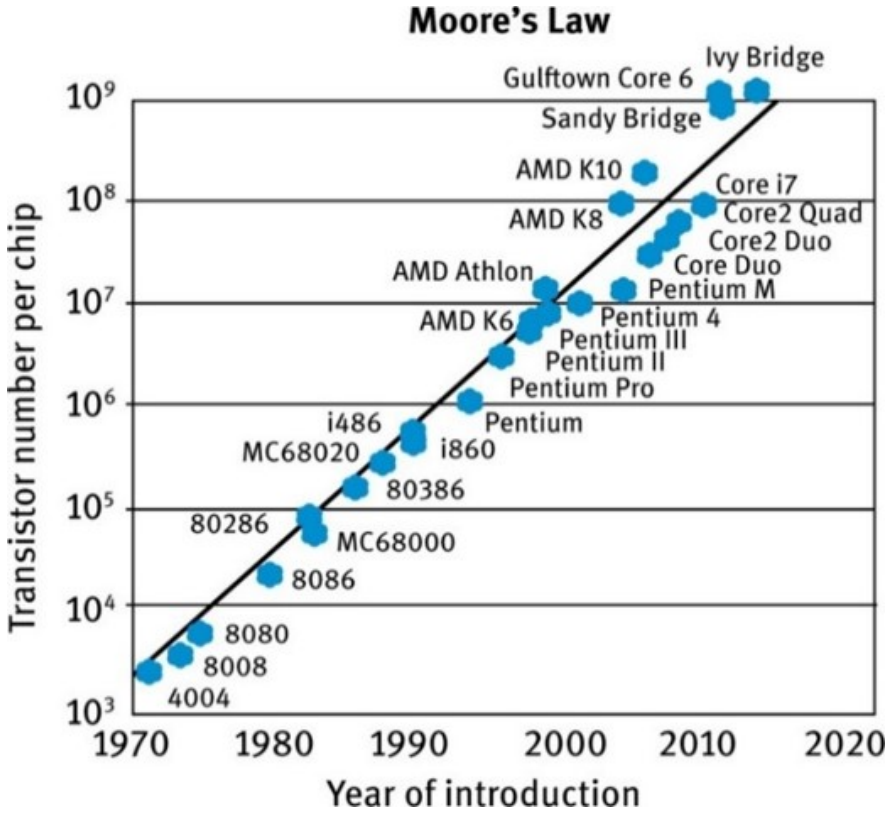


Figura 2: Lei de Moore representada por número de transistores dos processadores lançados a cada ano.

Fonte: Electroiq.com

Memórias estáticas de acesso randômico - *Static random access memory (SRAM)* ocupam a área mais significativa em sitema-em-um-chip [3], como pode ser visto na fig. 3. Em processos mais antigos, como 300 nm a porcentagem da área ocupada pela memória cache (na grande maioria das vezes SRAM) é muito pequena, conforme a tecnologia avança os chips se tornam mais potentes e necessitam de mais memória. No processo de fabricação de 130 nm, os microprocessadores fabricados nesse processo

possuem em média 30% de área ocupada por memória cache. Em processos ainda mais avançados, como 14 nm, esse valor pode chegar em 90%. Portanto, muito esforço é colocado na indústria de semicondutores para otimizar as memórias SRAM afim de manter os sistemas-em-um-chip dentro da lei de Moore.

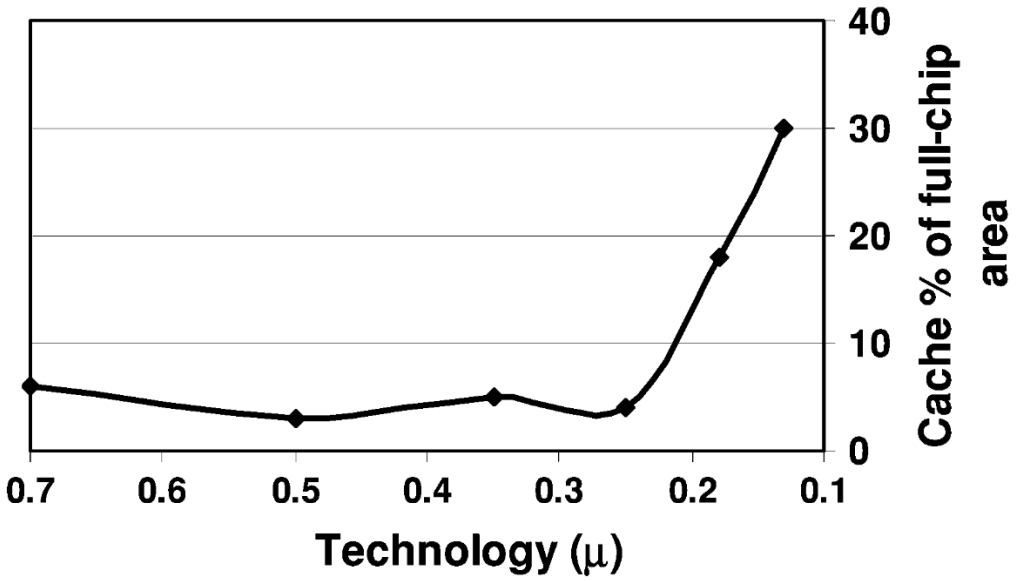


Figura 3: Participação crescente da memória SRAM na área total de um SoC.

Fonte: Conception de circuits memoire statique a access aleatoire tres basse tension en technologie UTBB-FDSOI - Adam Makosiej

1.1 Motivação

Este assunto de dissertação foi escolhido devido primeiramente a experiência que tive como estudante de iniciação cientifica do GICS (Grupo de circuitos e sistemas integrados da UFPR). Durante dois anos o Prof. Ph.D Bernardo Leite me orientou na área de amplificadores de potência analógicos. Como fruto desse trabalho o artigo: *"Linearity characterization of a CMOS Power Amplifier for IEEE 802.15.4, IEEE 802.11n and LTE signals"* (M. A. Rios, E. L. Santos, B. Leite, L. Lolis, A. Mariano) foi publicado em um congresso internacional (XXII Iberchip Workshop, Florianópolis, 2016).

Após esse período, realizei um intercambio de dupla-diplomação em Grenoble, França. O estágio de fim de curso, foi realizado no CEA-LETI, um laboratório muito importante na França. Destaco aqui a importância do período de iniciação cientifica no GICS para conseguir entrar nesse laboratório.

O estágio se desenvolveu no Comissariado da Energia Atômica e das Energias Alternativas (CEA), organismo público francês de pesquisas com caráter técnico e industrial. O CEA é referência de pesquisa de ponta, desenvolvimento e inovação em âmbito internacional. Fundado em 1945 com a criação da primeira bateria nuclear em território francês, o CEA nasceu com a intenção de garantir a produção energética nuclear francesa, e, ao passar das décadas, se adaptou à realidade tecnológica mundial, atuando em diversas áreas, de maneira mais notável, na defesa nacional e microeletrônica.

Contando com uma experiência científica e técnica reconhecida em todo o mundo, o CEA estabelece vários projetos em colaboração com parceiros acadêmicos e industriais. Além disso, o CEA é reconhecido como um dos laboratórios que depositam o maior número de patentes no mundo. Em 2016, a agência de notícias Thomson Reuters classificou a CEA como a organização de pesquisa pública mais inovadora do mundo.

O CEA está presente em todo o território francês, como pode ser visto na fig. 4. Em Grenoble o CEA é dividido em 3 laboratórios, sendo o LETI especializado no desenvolvimento da eletrônica. O LETI é dividido em vários departamentos, dentro do LETI trabalhei no DACLE, repartição responsável pelo desenvolvimento de circuitos digitais.

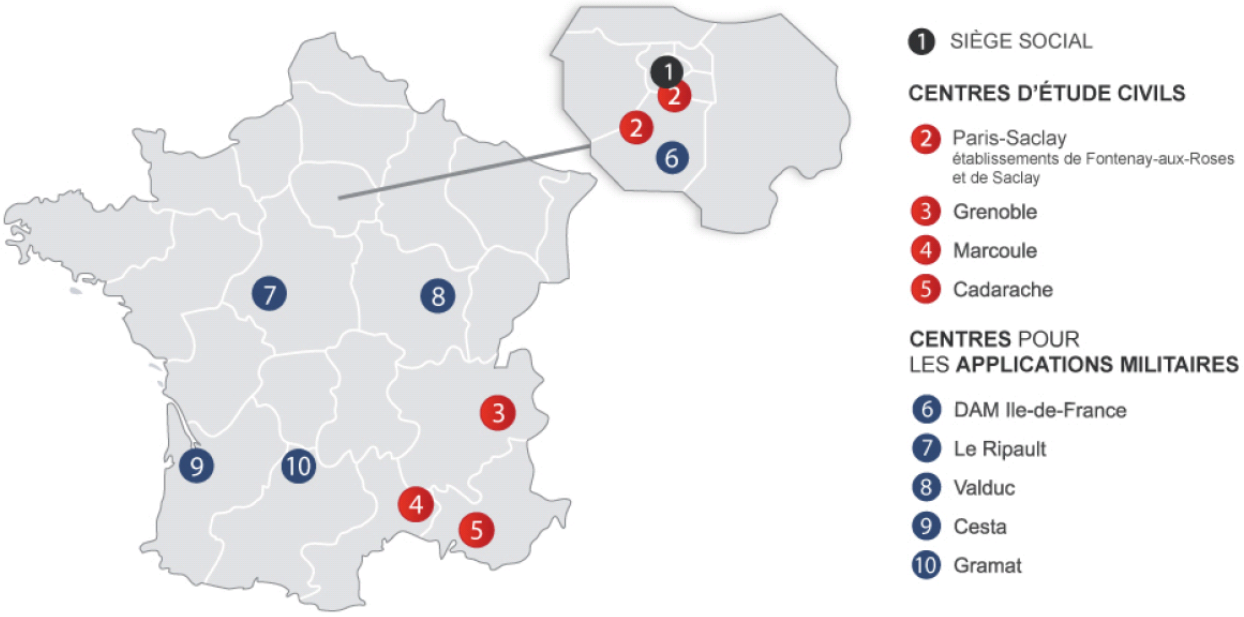


Figura 4: Mapa dos centros de pesquisa do CEA.

Fonte: CEA

A missão do DACLE é desenvolver soluções de hardware e software para sistemas digitais complexos, chamados system-on-chip (SoCs). O departamento cobre todo o processo, passando pelo design de funções e circuitos em matrizes avançadas, até o projeto de arquiteturas para sistemas embarcados confiáveis e robustos.

Vale ressaltar que CEA-LETI teve grande participação no desenvolvimento e amadurecimento da tecnologia FD-SOI juntamente com a STMicroeletronics e Global Foundries. O CEA-LETI também é responsável pela tecnologia que é o foco desse trabalho: 3-D CoolCube Sequencial.

1.2 Objetivos

1.2.1 Objetivo geral

Esse trabalho de pesquisa e conclusão de curso tem como objetivo buscar novas possibilidades de operação para a memória SRAM com tecnologia 3-D CoolCube, operações

que ainda não foram publicadas e exploradas na literatura. Dessa forma esse trabalho tem um forte carácter inovador. Diferentes formas de controle da tensão de corpo dos transistores que só são possíveis graças ao processo de fabricação do 3-D CoolCube serão explorados.

Neste contexto, a memória SRAM será trabalhada em nível celular para otimizar o consumo de energia no modo de espera e o desempenho no modo ativo. As principais métricas que definem o bom funcionamento de uma célula SRAM serão simulados tomando em conta as principais variações que podem ocorrer no processo de fabricação, o que faz com que esse trabalho também tenha um carácter industrial. Além disso, é proposto a realização de layouts celulares de SRAM em processo FD-SOI para 3D CoolCube, que se destina a ser fabricado para validação dos resultados propostos neste trabalho.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho consistem em:

- Estudo bibliográfico para compreensão da memória SRAM e os periféricos mínimos para a construção de uma matriz de SRAM.
- Criação de netlists para simulações em ELDO, cada netlist deve possuir a descrição da célula, os estímulos e extrações para cada métrica.
- Extração de parasitas via Calibre LVS.
- Estudo de ferramentas para simulações de Monte Carlo avançadas;
- Simulações em grande escala afim de realizar um trabalho exploratório;
- Identificar as melhores formas de se realizarem tensões de corpo inovadoras;
- Design de layouts de células SRAM em FD-SOI para tecnologia 3D CoolCube.

1.3 Justificativa

Devido a mudança na forma de fabricação dos circuitos integrados que a integração sequencial 3-D proporciona, vários circuitos clássicos, que já foram massivamente explorados devido a simplicidade, como a memória SRAM, podem agora ser trazidos novamente ao campo de pesquisa e serem explorados de formas totalmente novas.

No processo de fabricação convencional, os transistores são construídos em cima de poços de silício dopado, também conhecidos como PWELL e NWELL, para NMOS e PMOS, respectivamente. As tensões aplicadas nesses poços, chamadas de tensão de corpo, influenciam diretamente o comportamento dos transistores, tornando a polarização do corpo muito interessante para o controle de circuitos. No entanto, NWELL e PWELL são construídos lado a lado, formando um diodo e limitando a faixa de operação das tensões de corpo. No processo de fabricação do 3D CoolCube, como será explicado futuramente, não existe a limitação do diodo formado pelos poços NWELL e PWELL, abrindo novas possibilidades para a tensão de corpo.

1.4 Resultados esperados

Os resultados esperados terão como base uma matriz de SRAM em que as tensões de corpo são divididas pelas células da mesma linha. Também será considerado que conforme a operação a tensão de corpo aplicada a essas linhas possa mudar, ou seja, essa matriz terá dependência de operação. Então, espera-se que ao final do trabalho, seja definido certas configurações para as tensões de corpo que favoreçam as 3 operações da SRAM individualmente: leitura, escrita e estado de repouso.

1.5 Estrutura do trabalho

Este trabalho está dividido em seis capítulos:

1. No primeiro capítulo é introduzido o assunto da dissertação, juntamente com as

experiencias pessoais que motivaram esse trabalho. Os objetivos propostos e finalmente os resultados esperados ao final do trabalho.

2. No segundo capítulo é apresentado toda a fundamentação teórica necessária para a compreensão e análise do trabalho. É definido as métricas estáticas e dinâmicas da memória SRAM, juntamente com as principais operações. Além disso, é explicado os processos físicos de fabricação que convém para a dissertação.

3. No terceiro capítulo é mostrado a metodologia utilizada no trabalho, juntamente com as ferramentas que foram fundamentais durante o desenvolvimento.

4. No capítulo seguinte, são discutidos os resultados obtidos através das simulações e em seguida são discutidos os seus significados.

5. No capítulo 5, é trazido o desenho da memória SRAM genérico para a explicação dos principais fundamentos do seu design para circuitos integrados, e em seguida são mostrados os designs de SRAM realizados para a tecnologia 3-D CoolCube.

6. Finalmente, o ultimo capítulo sintetiza os resultados de forma a responder se o objetivo foi alcançado, concluindo com possíveis aplicações futuras para continuidade da pesquisa.

2 Fundamentação Teórica

2.1 memória Estática de Acesso Randômico - SRAM

Em um chip semicondutor de memória, cada digito de informação (bit) é armazenado em um pequeno circuito chamado célula de memória, consistindo de um a vários transistores. As células de memória são dispostas em matrizes retangulares na superfície do chip. As células de memória de 1 bit são agrupadas em pequenas unidades chamadas *palavras* que são acessadas juntas como um único endereço de memória. O conjunto das palavras de memória constrói a matriz de memória e define o seu tamanho (ex: uma memória com 256 palavras de 256 bits possui capacidade de armazenamento de 64kb). As duas operações básicas executadas por um chip de memória são ”leitura”, nas quais o conteúdo de uma palavra de memória é lido e ”escrita”em que dados são armazenados em uma palavra de memória, substituindo todos os dados armazenados anteriormente em um dado endereço.

Existem dois tipos principais de memórias em material semicondutor, volátil e não volátil. Sua classificação depende se o conteúdo da memória é perdido quando desligada a alimentação. Além disso, é possível dividir as memórias em estáticas ou dinâmicas, esta divisão depende se a memória precisa atualizar as informações armazenadas para não perdê-las. A memória de acesso aleatório (RAM) geralmente é volátil e leva o mesmo tempo para acessar qualquer local de memória, então os dados podem ser acessados com eficiência em qualquer ordem aleatória.

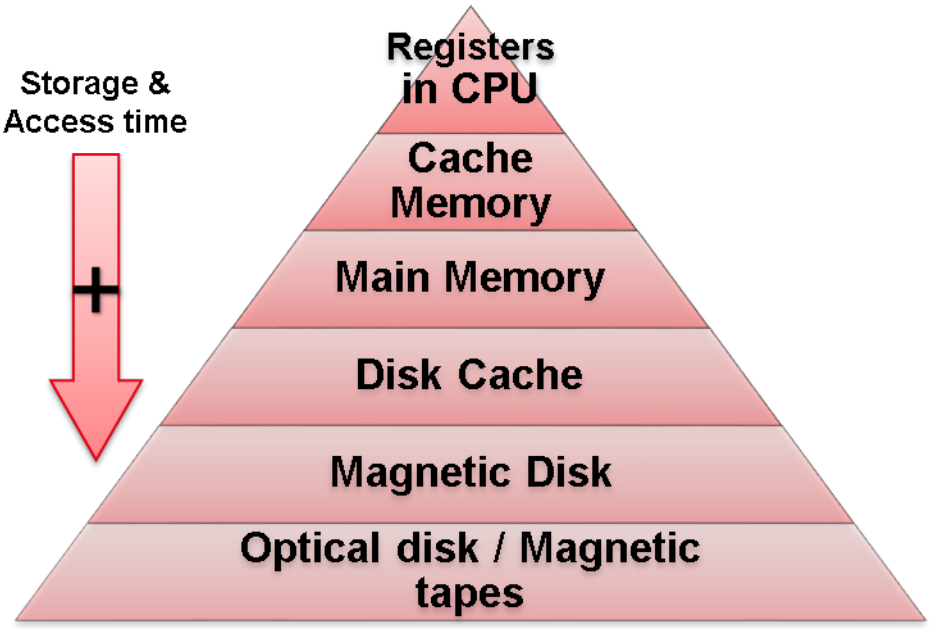


Figura 5: Hierarquia de memória.

Fonte: O Autor.

A hierarquia de memória organiza as memórias em armazenamento e tempo de acesso, como mostrado na fig. 5. Quanto maior a posição da memória na pirâmide, mais rápido ela será, porem com menor capacidade de armazenamento. A base da pirâmide é composta de memórias cujos dados podem ser permanentemente armazenados, geralmente uma unidade de disco rígido (HDD) ou disco de estado sólido (SSD).

Em seguida, no meio, há a memória principal que contém a maioria dos dados e instruções necessários para os programas atualmente em execução. A memória principal geralmente é a DRAM (Dynamic Random Access Memory), que armazena cada bit em um capacitor. A carga elétrica nos capacitores lentamente vaza e isso pode resultar em perda de dados, portanto, essa memória precisa de atualização constante. O processador não pode acessar os dados da DRAM quando ela está sendo atualizada, diminuindo a velocidade da memória. DRAM é o tipo mais barato de RAM, mas também é aquele que consome mais energia por causa das constantes atualizações.

Um nível acima na pirâmide, encontra-se a memória cache, normalmente SRAM (Static Random Access Memory), essa memória é usada pela CPU para reduzir o custo médio (tempo ou energia) para acessar dados da memória principal. A cache é uma memória menor e mais rápida, mais próxima de um núcleo de processamento, que armazena cópias dos dados dos locais de memória principal usados com frequência. A SRAM usa um circuito biestável (latch) para armazenar cada bit. As principais desvantagens dessa memória são o preço do circuito e a área ocupada.

Normalmente, um módulo SRAM consome menos energia do que um módulo DRAM. Isso ocorre porque a SRAM requer apenas uma pequena corrente constante, enquanto a DRAM requer rajadas de energia a cada poucos milissegundos para atualizar os dados armazenados. Essa corrente de atualização é várias ordens de grandeza maior que a corrente constante da SRAM. Assim, a SRAM é mais usada em equipamentos portáteis e operados por bateria. No entanto, o consumo de energia da SRAM depende da frequência de operação. Quando a SRAM é usada em um ritmo mais lento, ela consome energia praticamente insignificante quando está inativa. Por outro lado, em frequências mais altas, a SRAM pode consumir tanta energia quanto a DRAM.

2.2 Célula SRAM

A célula de bit SRAM padrão é composta de 6 transistores (6T) (fig. 7). Quatro transistores (M1, M2, M3, M4) formam um par de inversores de acoplamento cruzado usados para manter a informação dentro da célula, como mostrado na fig. 6. Dois transistores (M5, M6) funcionam como transistores de acesso. A ativação das linhas das palavras (WL - Word Line) conectam os nós internos $\overline{Q}$ e Q às linhas de bits, $\overline{BL}$ e BL (Bit Line), permitindo as operações de leitura/escrita nas células.

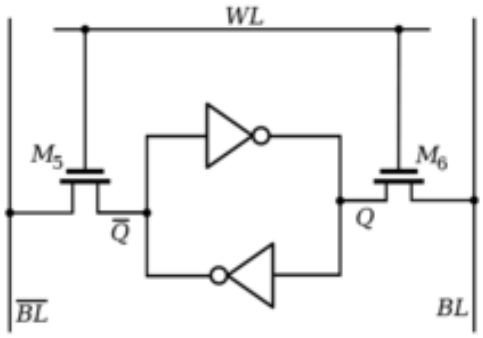


Figura 6: Célula SRAM a nivel dos inversores.

Fonte: O Autor.

Além disso, existem outras variações da célula de bit SRAM, com 4 até 9 transistores. A célula de bit 4T substitui um par de transistores por resistores, este circuito busca mais densidade de integração ao custo da estabilidade da célula. Por outro lado, as células bit com mais de 6 transistores buscam melhorar o desempenho das células em algum aspecto particular, como o tempo de acesso ou a estabilidade da célula, com os transistores adicionais.

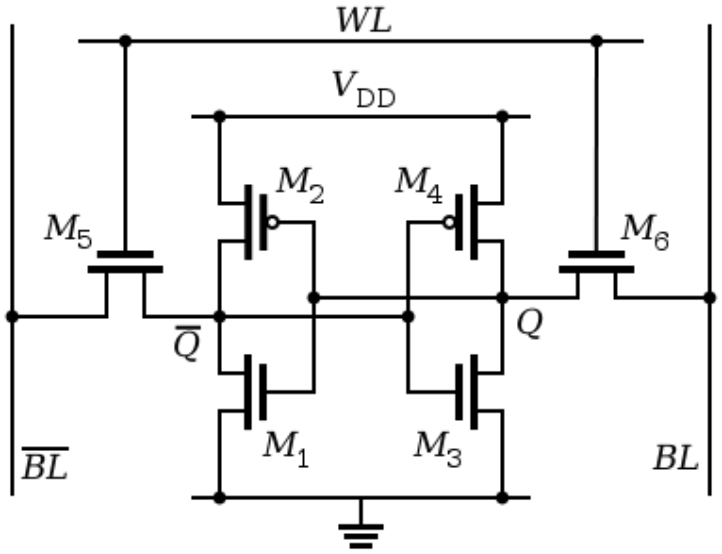


Figura 7: Arquitetura da célula SRAM com 6 transistores.

Fonte: O Autor.

Essas células são conectadas em uma matriz, compartilhando as linhas de palavras e linhas de bits. O endereço de memória é composto de duas partes, uma contendo o endereço da linha e a outra o endereço da coluna, essa abordagem permite acesso exclusivo a uma palavra específica.

2.3 Operações e Métricas da célula SRAM

Nos próximos capítulos serão definidas as operações e métricas da célula SRAM. As operações básicas da memória são: Escrita, Leitura e Retenção. As métricas são divididas conforme a natureza: estática ou dinâmica.

Durante a explicação será usado o circuito da fig. 7.

2.3.1 Margem de Ruído Estática - *SNM*

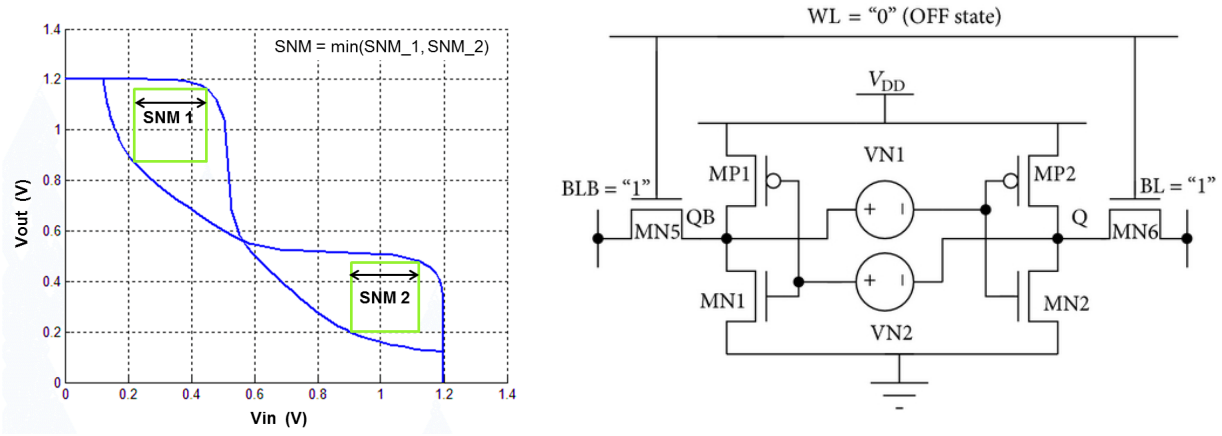


Figura 8: **(a)** Curva borboleta. **(b)** Circuito para a medida de SNM

Fonte: O Autor.

Margem de ruído estático (Static noise Margin (SNM)) é uma métrica que reflete a estabilidade da célula SRAM. Essa métrica é descrita como o maior valor de tensão de ruído entre os dois inversores na célula de memória, para os quais a célula ainda pode reter seus dados [4]. Em uma representação gráfica, o SNM pode ser extraído desenhando e espelhando as curvas características dos inversores, isso é chamado de curva borboleta e é representado na fig 8 (a). SNM corresponde ao tamanho do menor entre os dois maiores quadrados que podem ser inscritos entre as curvas borboleta [5].

Para medir o SNM, os inversores devem ser divididos e duas fontes de tensão de ruído são colocadas nas conexões cruzadas (VN1 e VN2), como mostrado na fig 8 (b).

A importância do SNM repousa no fato de que a retenção de dados na célula SRAM, tanto no modo de espera quanto durante o acesso para a leitura da célula, é uma restrição crítica nos processos mais finos. Devido ao fato de que a célula se torna menos estável (menor o valor do SNM) conforme a tensão de alimentação é mais baixa, tendência observada em processos de fabricação mais avançados. A menor estabilidade é devido ao impacto da variabilidade e à mudança no comportamento dos transistores, já que em al-

gumas aplicações de IoT eles operam com valores de tensão abaixo da tensão de limiar do transistor (subthreshold).

Deve-se destacar também que a célula é mais vulnerável a ruídos durante um acesso de leitura, já que o nó interno que armazena o ”0”sobe para um valor de tensão acima da terra, o que, em certos casos onde a estabilidade é baixa, pode alterar o estado da célula. [6] Esse fenômeno será explicado no próximo capítulo.

2.3.2 Operação de leitura

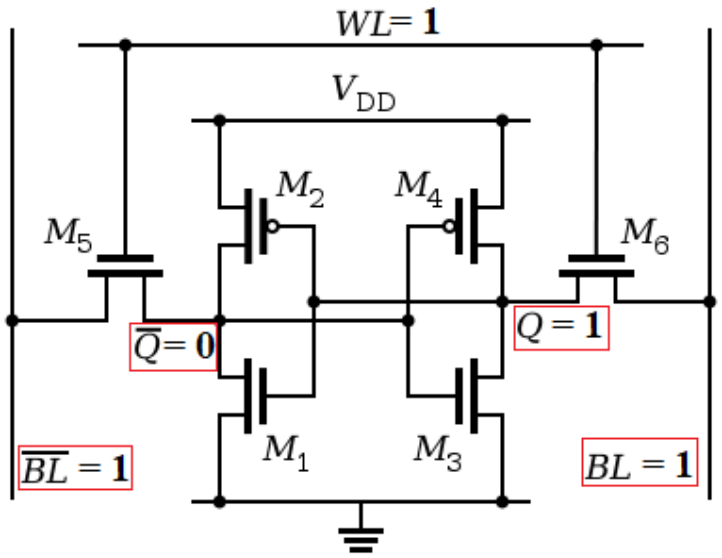


Figura 9: Circuito para operações de leitura.

Fonte: O Autor.

Para realizar a operação de leitura, inicialmente a memória deve ter algum valor armazenado. Portanto, considerando que a memória tem $\overline{Q}$ = 0 e Q = 1 e a WL esta ativada WL = 1 para executar a operação de leitura.

$\overline{BL}$ e BL atuam como saídas que transmitirão a informação do que esta armazenado para o amplificador de sentido, $\overline{BL}$ e BL são inicialmente pré-carregados em V_{DD}. Como mostra a fig. 9 (a). Como Q e BL estão em nível alto, não haverá descarga neste lado do

circuito. Como $\overline{Q}$ é 0, haverá uma diferença de tensão entre o nó $\overline{Q}$ e a tensão do nó $\overline{BL}$.

A corrente de leitura I_{READ} irá, portanto, descarregar o nó $\overline{BL}$ em direção a GND através dos transistores M5 e M1. No momento em que a corrente I_{READ} flui, o valor da tensão no nó $\overline{Q}$ sobe levemente a um valor acima do GND, caso esse valor seja maior do que SNM, a célula sofrerá uma inversão de valor.

BL e $\overline{BL}$ estão conectados ao amplificador de sentido (sense amplifier) através de multiplexadores, o amplificador de sentido detectará uma pequena diferença entre as linhas de bit (BL e $\overline{BL}$) e ligará essa diferença a um estado lógico (VDD ou GND).

O atraso de leitura é definido a partir do momento em que a WL sobe para uma tensão acima de 50% V_{DD} até o momento em que a diferença entre o BL e $\overline{BL}$ é menor que um determinado limite V_{TH}. Esse limite pode variar, mas normalmente é 100mV, como mostra a fig. 10.

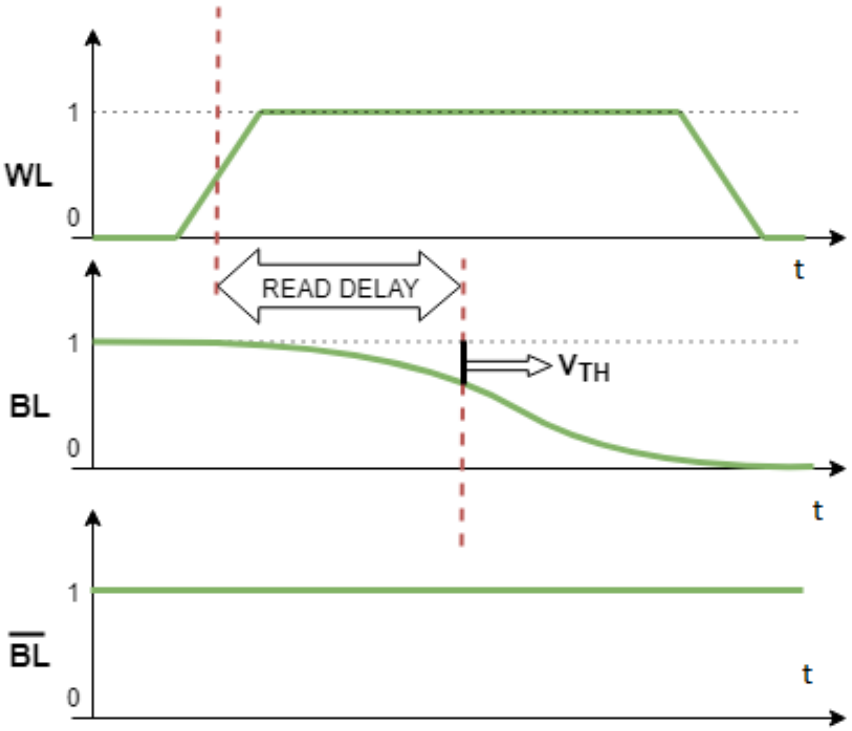


Figura 10: Atraso de leitura

Fonte: O Autor.

A operação de leitura pode ser degradada pela corrente de fuga presente nas outras células da mesma coluna. O pior caso, onde uma célula tem Q = 1 e $\overline{Q}$ = 0, enquanto todas as outras células da mesma coluna têm Q = 0 e $\overline{Q}$ = 1 é representado na fig. 11. Na operação de leitura, a tensão em $\overline{BL}$ é descarregada pela corrente I_{READ} que flui através da célula de bit. No outro lado, BL apresentará uma certa queda de tensão devido à corrente de fuga $\sum I_{LEAK}$ que flui de BL = 1 para Q = 0 em todas as células não-acessadas naquela coluna.

Normalmente, o I_{READ} é muito maior que uma corrente de fuga I_{LEAK}. Em tensões mais baixas e altas temperaturas a relação I_{on}/I_{off} dos transistores é degradada. Isso pode levar a um impacto de margem na operação de leitura tanto na velocidade da operação quanto na estabilidade. Isto é verdade em particular para matrizes com longas colunas onde a corrente de fuga pode ser muito alta.

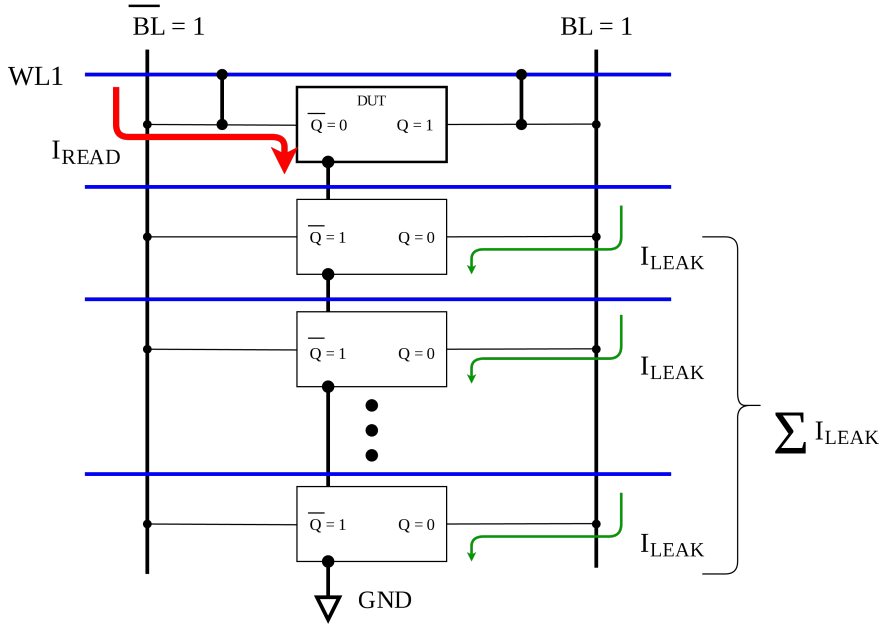


Figura 11: Coluna SRAM onde a primeira célula tem Q = 1 e as outras células têm Q = 0.

Fonte: O Autor.

2.3.3 Margem de Ruído de Escrita - *WM*

Como o SNM é a métrica de estabilidade para a célula de bit na operação de retenção ou leitura, a Margem de Ruído de Escrita - *Write Noise Margin (WNM)* é o critério de estabilidade para a operação de escrita. O WNM é definido como o menor quadrado inscrito que pode caber entre as curvas características da célula durante uma operação de escrita, enquanto que SNM é o maior quadrado. Em relação ao circuito de medição mostrado na fig. 12 (b), WL = 1 e uma linha de bit esta em 1 e outra em 0.

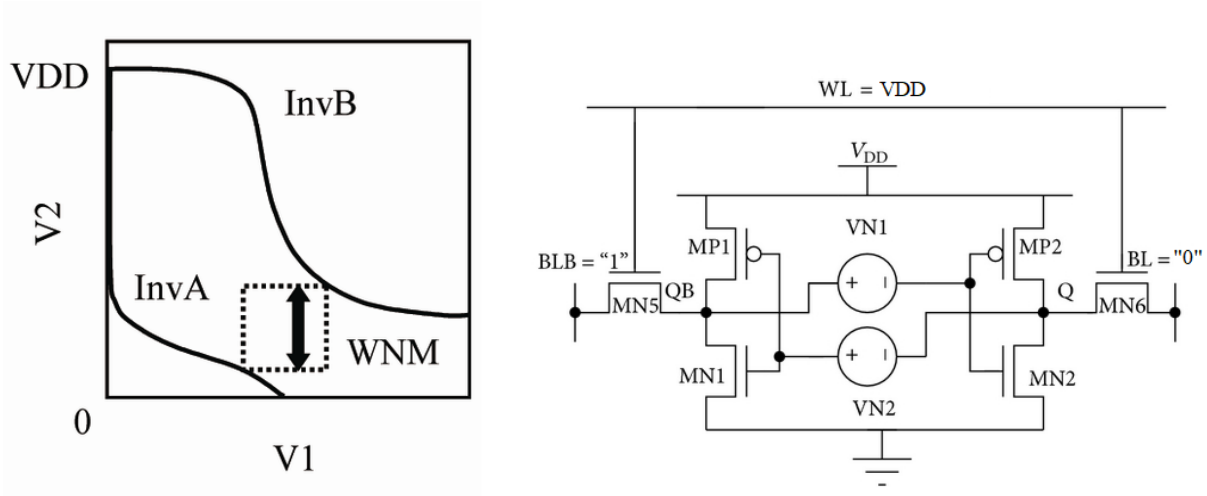


Figura 12: **(a)** Curva borboleta. **(b)** Circuito para medir WNM

Fonte: O Autor.

2.3.4 Operação de Escrita

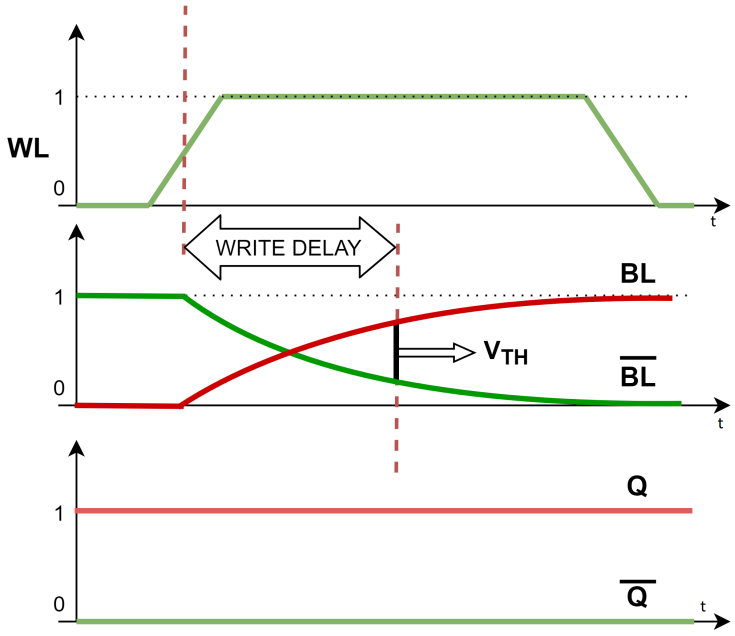


Figura 13: Operação de escrita

Fonte: O Autor.

Considerando uma operação de escrita com a finalidade de armazenar ”0”na célula (Q = 0 e $\overline{Q}$ = 1) e também considerando que a memória inicialmente tem ”1”armazenado (Q = 1 e $\overline{Q}$ = 0). Inicialmente, WL esta em estado alto WL = 1 e, portanto, a operação de gravação pode ser executada. Na operação de escrita BL e $\overline{BL}$ são agora as entradas do circuito, as linhas de bit são carregadas com o valor desejado BL = 0 e $\overline{BL}$ = 1. Dessa forma, Q vai para o estado baixo Q = 0 e $\overline{Q}$ é levado para o estado alto. Como mostra a fig. 13.

Resumindo, para realizar uma operação de escrita, uma das linhas de bit é carregada para o VDD, a outra permanece no GND, enquanto a WL é ativada. Em seguida, a célula inverte dependendo das informações armazenadas.

O atraso de gravação é definido a partir do momento em que o WL sobe para uma tensão acima de 50 % de V_{DD} para o momento em que a diferença entre Q e $\overline{Q}$ é inferior

a um determinado limiar. Esse limite pode variar, em alguns casos é 80% de V_{DD}.

2.3.5 Tensão de Retenção de Dados

A Tensão de Retenção de Dados - *Data Retention Voltage (DRV)* é a tensão operacional mínima no modo de retenção em que a célula SRAM pode manter seu estado [7]. Este modo é normalmente usado em dispositivos alimentados por bateria, com longos períodos de espera. Portanto, ao operar nesta tensão de alimentação, o circuito tem como objetivo consumir a mínima potência através da corrente de fuga com a maior parte do circuito desligado.

2.4 Produção de memória

Levando em consideração que um dos pontos mais fracos da SRAM em comparação com outras topologias de memorias da mesma classe na hierarquia, é a área ocupada no silício, a fabricação da SRAM sempre usará os menores tamanhos de transistores dentro do permitido da tecnologia, buscando a máxima densidade de integração. Somado a essa fato ao escalonamento tecnológico, a variabilidade intrínseca do processo de fabricação da memoria SRAM é alto. Essa variabilidade é vista em vários aspectos do circuito, mais notavelmente na tensão de limiar V_{TH} dos transistores.

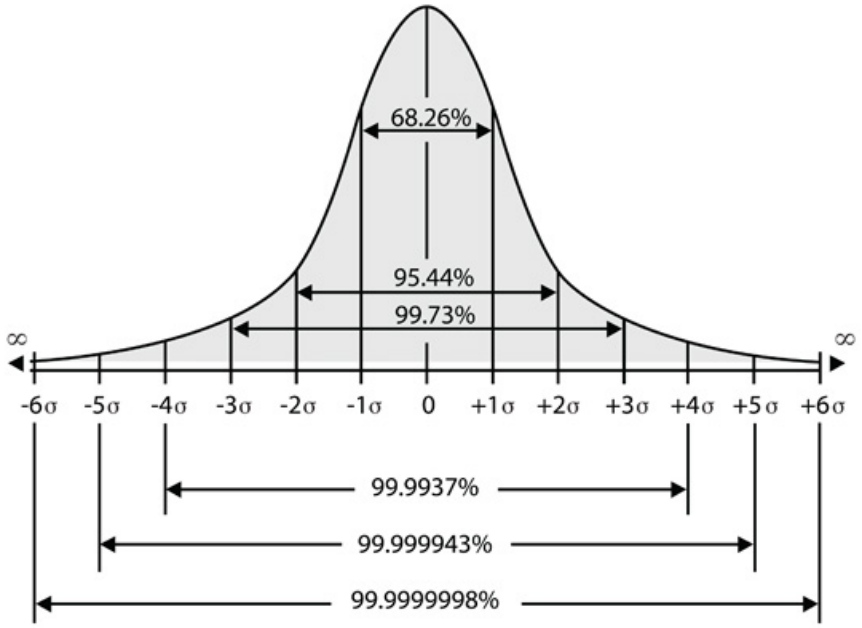


Figura 14: Distribuição Gaussiana.

Fig. 14 representa uma distribuição estatística gaussiana onde o desvio padrão (sigma - σ) representa a distância para o centro (média - μ). O desvio padrão é uma medida que quantifica a quantidade de dispersão de um conjunto de valores de dados. Um desvio padrão baixo indica que os dados tendem a estar próximos da média, também chamado de valor nominal do conjunto, enquanto um desvio padrão alto indica que os dados estão espalhados por um intervalo maior de valores.

Considerando a quantidade de sigmas em uma distribuição normal, $\pm 1\ \sigma$ inclui 65.26% do conjunto, enquanto $\pm 6\ \sigma$ representa 99.9999998 %. portanto, fora dos limites de $\pm 6\ \sigma$, encontra-se cerca de 2 amostras em um conjunto de 1 bilhão.

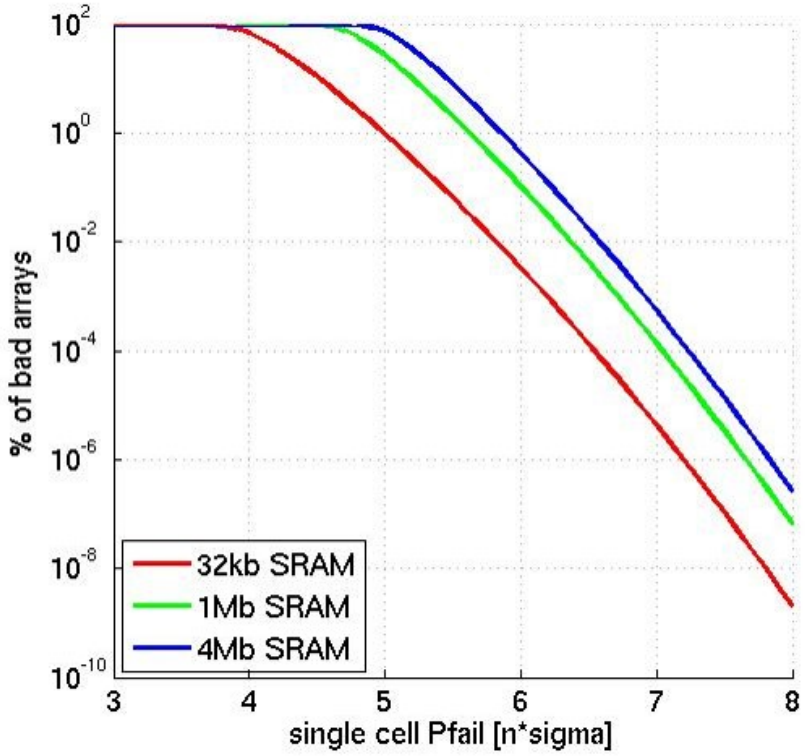


Figura 15: Relação entre σ e porcentagem de chips não utilizáveis.

Fonte: Han Ping Fung

O rendimento da memória é a porcentagem de chips funcionais, presumindo que, se uma única célula de bit falhar em qualquer operação, todo o chip falhará [3]. Fig. 15 mostra a relação entre o número de σ e a porcentagem de falhas para três tamanhos diferentes de SRAM. Células SRAM são encontradas aos milhões em cada circuito, logo esse dispositivo deve ter uma alta confiabilidade, se for assumido um sigma de 4 para as células (1 célula falhando em um conjunto de quase 16.000) um chip de 32kb terá quase 100% de rejeição. Logo, deve ser assumido um valor maior para σ.

Neste trabalho, será considerado o valor de margem padrão industrial: 6 σ. Con-

siderando uma SRAM de 32kb com margem de 6 σ, aproximadamente um chip não é utilizável em um conjunto de dez mil. Esse valor é um ótimo compromisso entre o custo de produção (chips perdidos) e complexidade de desenvolvimento do circuito.

2.4.1 Desenvolvimento e Simulação

O método que considera a variabilidade do processo é a simulação Monte Carlo (MC). Em uma simulação de MC, o software de simulação SPICE (eldo, neste caso) executará várias vezes a mesma simulação, aplicando uma distribuição normal para cada desvio (dimensão, quantidade de materiais de doping, etc.).

No entanto, considerando 6 σ, o custo computacional é alto, ou seja, demora muito tempo para simulação com os computadores que tem-se a disposição, cerca de 1 bilhão de simulações para cada cenário, e um MC pode levar horas. Assim, existem várias alternativas para o assunto, algumas serão exploradas neste trabalho.

2.5 Limites de Processo de Fabricação

Quando o CMOS é fabricado, há uma variabilidade na fabricação que resulta em alguns dispositivos com tensão limiar maior ou menor em comparação com o caso nominal. Os limites do processo representam os extremos dessas variações de parâmetros dentro dos quais um circuito que foi gravado no silício deve funcionar corretamente. Um circuito funcionando nos dispositivos fabricados nesses limites do processo pode ficar mais lento ou mais rápido do que o especificado e em temperaturas e tensões menores ou maiores, mas se o circuito não funcionar em algum desses limites do processo, o projeto é considerado como tendo margem de projeto inadequada.

Os limites do processo são descritos como uma sigla de duas letras, em que a primeira letra é a condição do NMOS e a segunda letra do PMOS. A letra T significa Típico (tensão de limiar V_{TH} nominal), F para Rápido (tensão de limiar V_{TH} abaixo do nominal) e S para Lento (tensão de limiar V_{TH} acima do nominal). Os limites estão ilustrados na fig. 16 em

um gráfico bidimensional, o eixo Y é a variabilidade da tensão de limiar V_{THp} do PMOS e o eixo X é a variabilidade da tensão de limiar V_{THn} do NMOS.

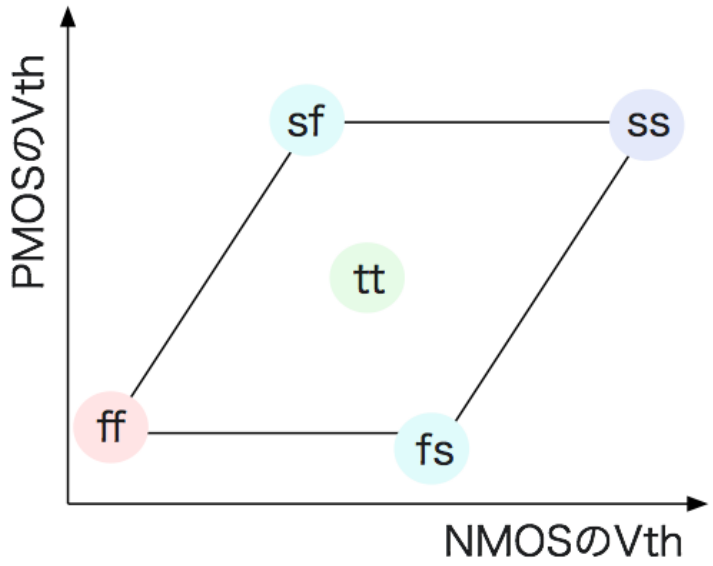


Figura 16: Desvio dos *corners* para PMOS e NMOS.

O limite TT apresenta NMOS e PMOS com tensão de limiar V_{TH} nominal, as simulações padrões para analisar um circuito consideram este limite. Uma vez que a variabilidade do processo precisa ser levada em conta, os limites que apresentam a mesma característica para NMOS e PMOS, como os limites FF e SS, e os limites transversais, SF e FS, também são simulados.

2.5.1 Influência dos limites na célula SRAM

A célula de bit SRAM padrão possui quatro transistores NMOS e dois transistores PMOS, portanto ambos os limites do processo tem grande influência neste circuito. O limite FF torna o circuito muito mais rápido em operações de leitura e escrita, e tem baixo impacto na estabilidade da célula, já que a tensão de limiar V_{th} de ambos NMOS e PMOS muda simultaneamente por uma margem similar. A desvantagem do limite FF é a corrente de fuga, que atinge seu valor mais alto. O limite oposto, SS, apresenta

características complementares, pior desempenho nas operações de leitura e escrita, porém com a menor corrente de fuga.

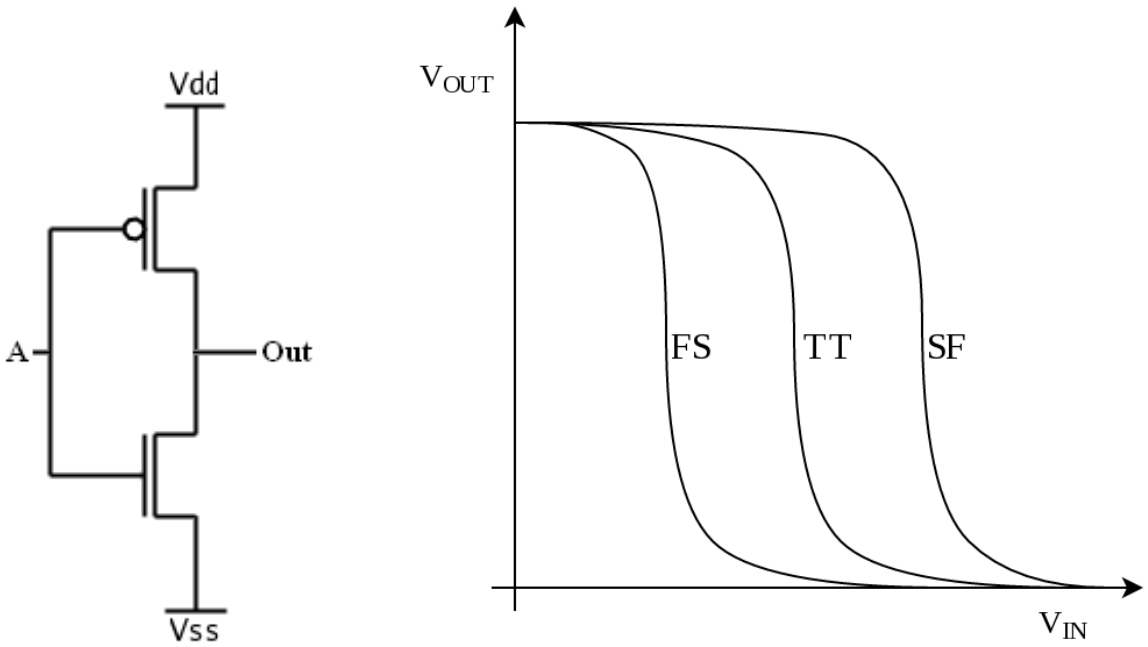


Figura 17: **(a)** Inversor **(b)** VTC

Os limites transversais desempenham um papel mais importante na estabilidade da célula. Tomando um inversor como exemplo, como mostra a fig. 17 e considerando que os transistores são dimensionados para ter a curva de transferência - Voltage Transfer Characteristic (VTC) nominal no limite TT. O limite SF apresenta um NMOS mais fraco e um PMOS mais forte, forçando o VTC para o lado direito, porque o PMOS, também chamado de Pull Up, dificulta que o circuito troque do estado alto para o estado baixo. O limite oposto, FS, tem o NMOS (Push Down) mais forte, puxando VTC para o lado esquerdo.

Lembrando a definição de SNM, que diz que a métrica corresponde ao tamanho do menor dos dois maiores quadrados que podem ser inscritos entre a curva borboleta, ao retirar o VTC do centro certamente irá degradar o SNM, como mostrado na fig. 18.

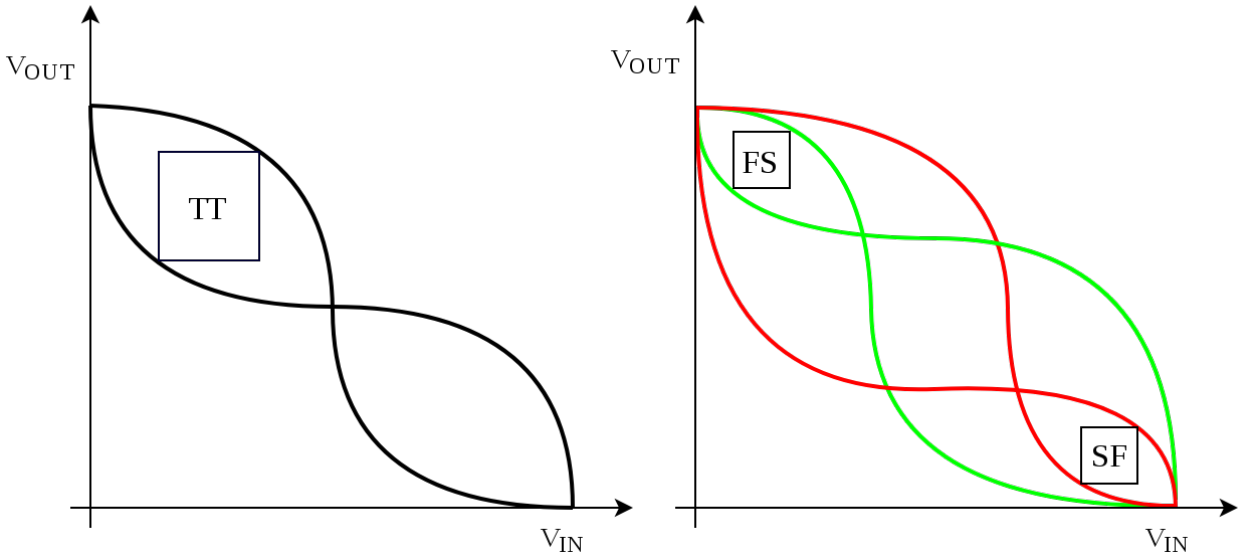


Figura 18: VTC para diferentes limites.

2.6 3-D CoolCube e tensão de corpo

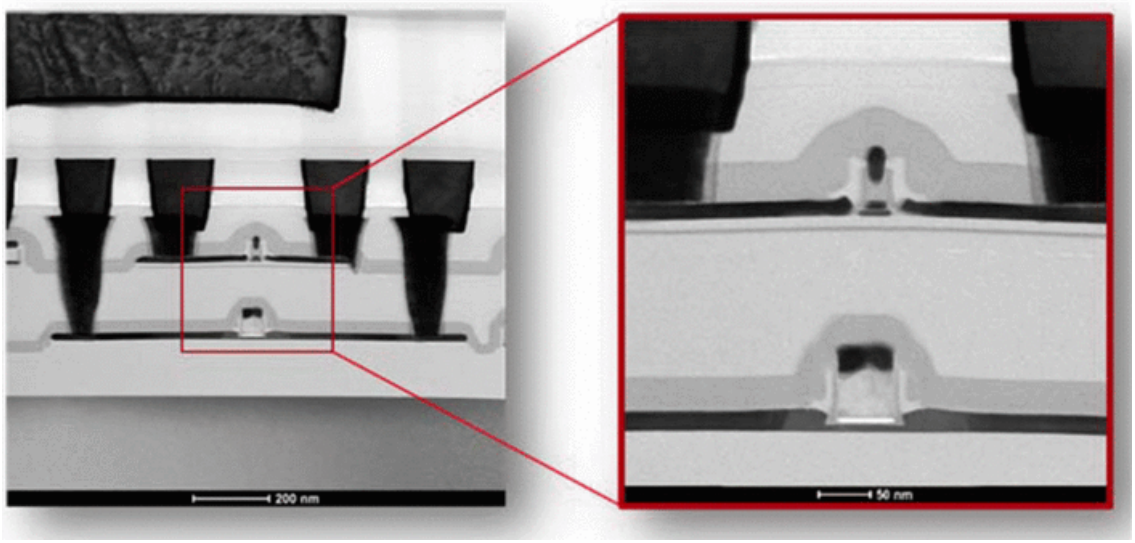


Figura 19: Micrografia 3-D CoolCube.

Desde o início do projeto dos circuitos integrados, os semicondutores foram projetados de maneira planar, em um único nível de transistores. Essa abordagem foi considerada

satisfatória, uma vez que o aumento na densidade alcançado com a miniaturização dos transistores foi suficiente. No entanto, os transistores já são fabricados em escala molecular, colocando assim um possível fim à era da miniaturização. Para contornar essa barreira de aumento da densidade de transistores por área, uma solução que está sendo explorada é o empilhamento de transistores, também chamado de fabricação monolítica 3D ou sequencial.

O laboratório LETI introduziu a tecnologia 3-D CoolCube a alguns anos atrás, a fig. 19 apresenta uma micrografia de uma fabricação bem sucedida dessa tecnologia. Este processo sequencial traz um enorme ganho em densidade, desempenho e custo. Essa tecnologia é baseada no empilhamento de níveis de transistores, mas realizar esse processo não é simples. O processo de fabricação padrão submete o silício a temperaturas que podem chegar a 1000 ℃. Se o mesmo processo for executado nos níveis superiores de transistores, o nível inferior certamente seria danificado. Portanto, um processo com temperaturas mais baixas é necessário nos níveis superiores. Esses dois conceitos dão nome à tecnologia: o método de fabricação de baixa temperatura (”COOL”) e a integração 3-D (”CUBE”).

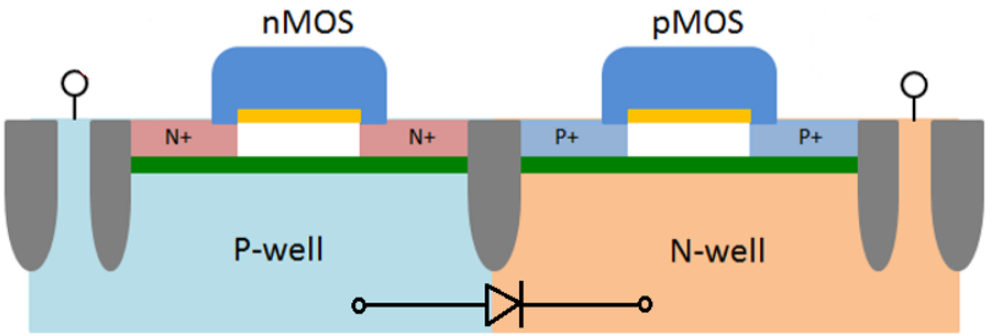


Figura 20: Diodo formado pelos poços NWELL e PWELL em uma fabricação convencional de FD-SOI.

Na primeira camada, os transistores são construídos em cima de poços de silício dopado, também conhecidos como PWELL e NWELL, para NMOS e PMOS, respectivamente. A tensão aplicada nesses poços, chamada de tensão de corpo, influencia direta-

mente o comportamento dos transistores, tornando a polarização do corpo muito interessante para o controle de circuitos. No entanto, NWELL e PWELL são construídos lado a lado, formando um diodo, como mostrado na fig. 20, e limitando a faixa de operação da tensão de corpo, já que no caso de $V(PWELL) > V(NWELL)$ isso representa um curto-circuito.

Devido ao inovador método de fabricação do 3D CoolCube, a última camada metálica do primeiro nível de transistores desempenha o papel de base do segundo nível, como pode ser visto na fig. 21, abrindo a possibilidade de isolar planos para construir transistores com uma granularidade muito fina, assim possibilidades de tensão de corpo que não eram possíveis no método convencional, podem agora ser realizadas.

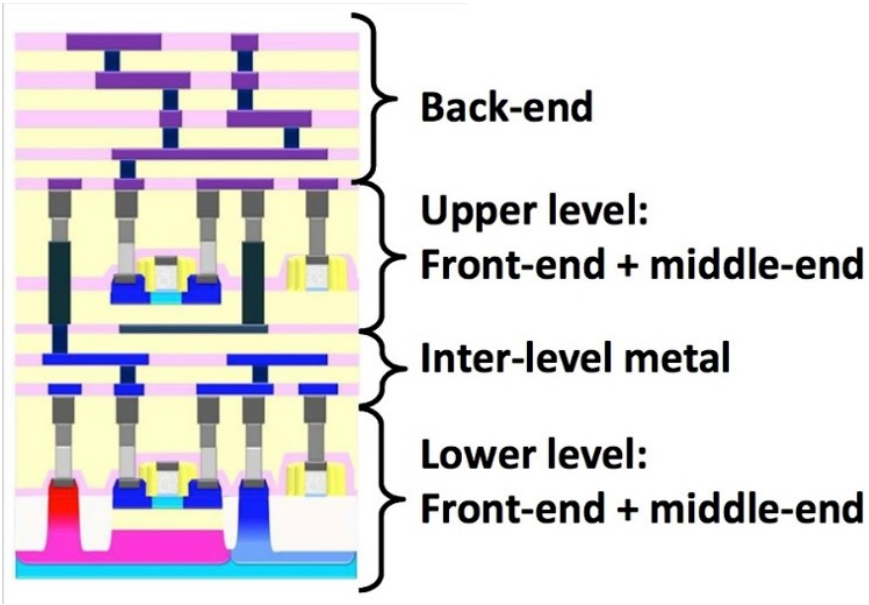


Figura 21: Esquema 3-D CoolCube.

2.6.1 Efeito da tensão de corpo nos transistores

A polarização do corpo altera a tensão de limiar (V_{TH}), influenciando também a corrente de fuga dos transistores. Podemos então diferenciar duas formas de polarização corporal:

- Polarização Direta de Corpo (PDC): Diminui o VT do dispositivo, aumentando o desempenho do transistor e permite operações em tensões mais baixas.
- Polarização Reversa de Corpo (PRC): Aumenta o VT do dispositivo, diminuindo a corrente de fuga.

A fig 22 resume a influência da tensão do corpo na operação dos transistores.

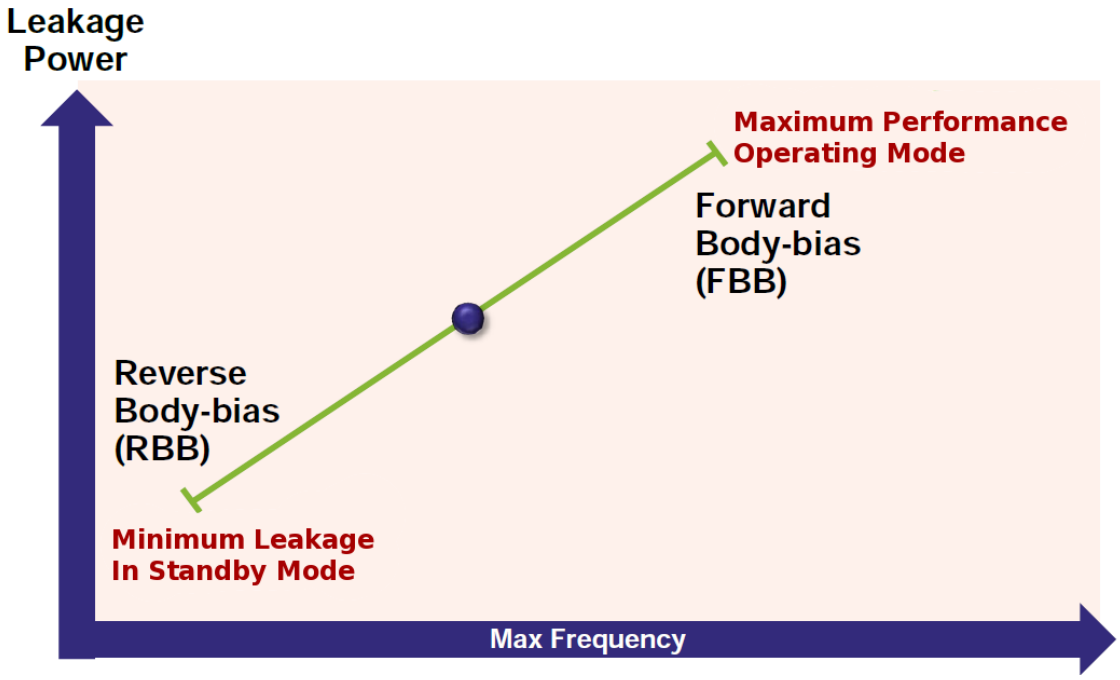


Figura 22: Efeito da tensão do corpo no comportamento dos transistores.

3 Metodologia

Para realizar as simulações necessárias para as análises do circuito o primeiro passo foi escrever as netlists. Foi necessário uma netlist para cada operação e então validar os resultados apresentados por elas. Os resultados obtidos com as simulações foram comparadas com os valores fornecidos pelo fabricante.

Considerando que o trabalho é exploratório e que o interesse repousa em tendências e não em valores exatos, foram usados células SRAM padrões na tecnologia 28 nm FD-SOI nas simulações, apenas considerando que elas serão construídas na camada superior de transistores e que não há limitações em relação a tensão de corpo.

Foram exploradas quatro células da STmicroelectronics em FD-SOI 28 nm, três células padrão (seis transistores) com diferentes densidades de integração e uma célula com oito transistores. As métricas destinadas à análise e comparação foram: SNM, WM, atraso de leitura, atraso de escrita e DRV.

O software de simulação SPICE usado foi o Eldo. Criado pela Mentor Graphics, a Plataforma Eldo é a plataforma de verificação de circuito mais avançada e comprovada do setor para circuitos analógicos. Oferece soluções diferenciadas para verificação de confiabilidade, análise e diagnóstico de circuitos.

As netlists escritas para a simulação em SPICE devem ser compostas dos seguintes elementos:

- Opções de simulação;
- Descrição do circuitos;
- Formas de onda que alimentam o circuito;
- Descrição das extrações e medidas.

As opções de simulação são usadas para selecionar os arquivos que serão criados, o tipo de simulação (DC ou AC), a temperatura e opções mais específicas do Eldo.

A descrição do circuito é composta pela célula e pelos parasitas. A forma de onda é responsável pela dinâmica do circuito e pela definição da operação. Por fim, a extração fornece a informação se a operação foi realizada corretamente, se sim, o valor medido.

3.1 Preparando netlists

A descrição da memória é feita instanciando metade do circuito, 3 transistores, em um sub-circuito e, em seguida, conectando dois sub-circuitos formando uma célula de bit completa.

As capacitâncias parasitas foram extraídos no Cadence Virtuoso com a ferramenta CalibreLVS. Primeiramente foi realizada uma matriz de memória de 128x128 com os nós curto-circuitados (Todas as células compartilhando a mesma Word Line, Bit Line e etc.) e então extraído as capacitâncias totais dessa matriz por nó.

Esta abordagem foi escolhida porque as capacitâncias parasitas de uma célula são definidas pela geometria da própria célula e das células ao seu redor. Utilizando o CalibreLVS para extração das capacitâncias da matriz, uma vez tendo os valores para os nós, eles foram divididos por 128x128. O resultado disso é um valor médio individual de capacitâncias parasitas.

Depois que as netlists foram concluídas e depuradas, as simulações avançadas de Monte-Carlo foram feitas. Dois métodos foram propostos para serem explorados:

1. Fast Fail Prob

2. Importance sampling Monte Carlo (ISMC)

Ambas as abordagens podem ser implementadas nas netlists do Eldo, devido ao tempo de simulação que o Fast Fail Prob (FFPROB) apresentou esse método foi escolhido. FFPROB é a extensão do método ISMC, proporcionando um tempo de simulação mais rápido, mantendo uma boa precisão.

ISCM e FFPROB são métodos para encontrar as características de uma distribuição aleatória de forma mais optimizada, que leva em conta o conceito de que alguns valores

tomados por uma escolha aleatória tem mais efeitos do que outros. Se essa escolha for bem realizada, o número de simulações de Monte-Carlo necessários será menor.

Dois modos de operação com FFPROB são possíveis, em ambos há dois argumentos obrigatórios: o nome da extração (métrica) e o lado da distribuição, ou seja, se o valor da probabilidade de falhar for maior ou menor que um determinado valor. O primeiro modo consiste em passar a probabilidade de falhar desejada (normalmente 6σ ou 1e-9) e a ferramenta retornará o valor específico que realiza essa probabilidade de falha. O segundo modo é o inverso, em que algum valor é passado e a probabilidade de falha é retornada.

Como o FFPROB é uma nova ferramenta da Mentor Graphics, foi entrado em contato para resolver as dúvidas e problemas que ocorreram durante a análise.

3.2 Custo Das Simulações

Com todas as netlists configuradas, foram realizadas simulações de ampla faixa alterando o PVT (processo, Tensão de alimentação e Temperatura) e a configuração da tensão do corpo. Para este fim, dois códigos em python foram desenvolvidos, um para o lançamento das simulações Eldo e outro para coletar os dados. Sabendo que o FFPROB é um algoritmo iterativo que executa várias simulações para chegar à solução, o tempo de execução de cada simulação não é insignificante.

Todas as simulações foram feitas com FFPROB, portanto o custo computacional é muito alto, um conjunto completo de simulações levou cerca de um mês para ser feito. Depois de identificar essa limitação, foi proposto usar um cluster de computadores composto de 10 máquinas rodando remotamente, o que aumentou muito a produtividade. Por exemplo, uma ampla gama de simulações pode ser composta de:

- 4 células;
- 5 limites de processo (TT, SS, FF, SF, FS);
- 3 Temperaturas (-40℃, 25℃, 125℃);

- 3 Tensões de alimentação (0.3V, 0.4V, 0.5V);
- 4 métricas (WM, SNM, Atraso de leitura, Atraso de escrita).

Apenas para esses elementos são necessários 720 simulações, este valor ainda deve ser multiplicado pelo número configuração de polarização do corpo diferentes. Além disso, outros argumentos específicos para cada operação, como o tamanho da coluna e valores diferentes para o FFPROB. Cada simulação precisa de mais de 10 minutos para ser feita.

A análise de dados e o gerenciamento de um grande número de simulações também se mostraram um grande desafio. O MS Excel foi usado para esta tarefa.

3.3 Explorando o 3D CoolCube para células de bit SRAM

Analisando a célula de memória SRAM, a mudança mais notável quando fabricado com 3-D CoolCube é as operações com tensão de corpo que não tem limitações. A célula utilizada foi uma das quatro fornecidas pela STmicroelectronics em FD-SOI 28 nm, a escolhida foi a que apresentou densidade média de integração.

Células planares foram usadas para a avaliação, pois a tecnologia 3-D CoolCube está em um certo nível de maturidade que a segunda camada de transistores é feita em 65 nm FD-SOI enquanto que a primeira é 28 nm FD-SOI. Logo, não é viável criar uma célula que esteja presente nas duas camadas.

Dessa forma, o objetivo é alinhar os dispositivos 3-D com os planares. Então, sob essa suposição, usar a célula planar para a análise é aceitável e deve projetar corretamente as análises para célula 3D assim que a tecnologia amadurecer.

Para as simulações, considerou-se que o layout da célula SRAM possui três backplanes diferentes unindo transistores com as mesmas características. As tensões de corpo são:

1. PU: Transistores Pull Up (M2, M4)

2. PD: Transistores Pull Down (M1, M3)

3. ACC: Transistores de acesso (M5, M6)

SETUP	PU (M2 M4) (V)	PD (M1 M3) (V)	ACC (M5 M6) (V)
1	0	0	0
2	-2	2	2
3	2	-2	-2
4	0	-2	-2
5	0	2	2
6	-2	2	0
7	-2	-2	2
8	0	-2	2

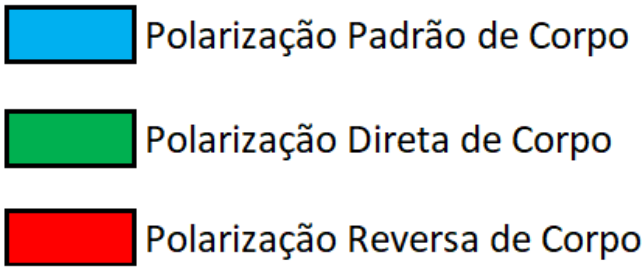


Figura 23: Tensões de corpo usadas na simulação.

Todas as configurações possíveis de tensão de corpo foram simuladas, as que trazem algo interessante para alguma métrica são apresentadas na fig. 23. Nesta tabela pode observar que a polarização de corpo padrão de 0V para NMOS e PMOS é representada em azul. A Polarização direta de corpo (PDC) é -2V para PMOS e 2V para NMOS e é representado em verde. A polarização reversa de corpo (PRC) é 2V para PMOS e -2V para NMOS e é mostrada em vermelho.

2V para polarização do corpo é bastante alto, mas é possível ser feito. Este trabalho foi focado em valores altos de polarização do corpo visando uma otimização mais extrema, os valores de polarização intermediários fazem parte de um trabalho futuro.

Os tópicos a seguir apresentarão as melhores configurações para a tensão de corpo para cada métrica discutida. Considerou-se que a tensão de corpo tem dependência de operação, o que significa que, em função da operação que está sendo executada, uma configuração de polarização de corpo diferente pode ser aplicada.

4 Resultados e Discussões

A seguir serão apresentados os principais resultados e discussões. A estabilidade da célula deve ser satisfeita em todos os cenários, portanto, uma abordagem estática é analisada primeiramente para então considerar as medidas dinâmicas e o atraso das operações. Todas as simulações foram realizadas nos piores casos para cada métrica.

4.1 Retenção

A característica comum entre os dispositivos IoT é que eles são alimentados por baterias, e alguns deles passam a maior parte da vida em modo de espera, transformando o consumo neste período (correntes de fuga) em um aspecto de grande importância. A memória SRAM é volátil, portanto, sempre é necessário ter um valor mínimo de tensão nessa memória para reter as informações.

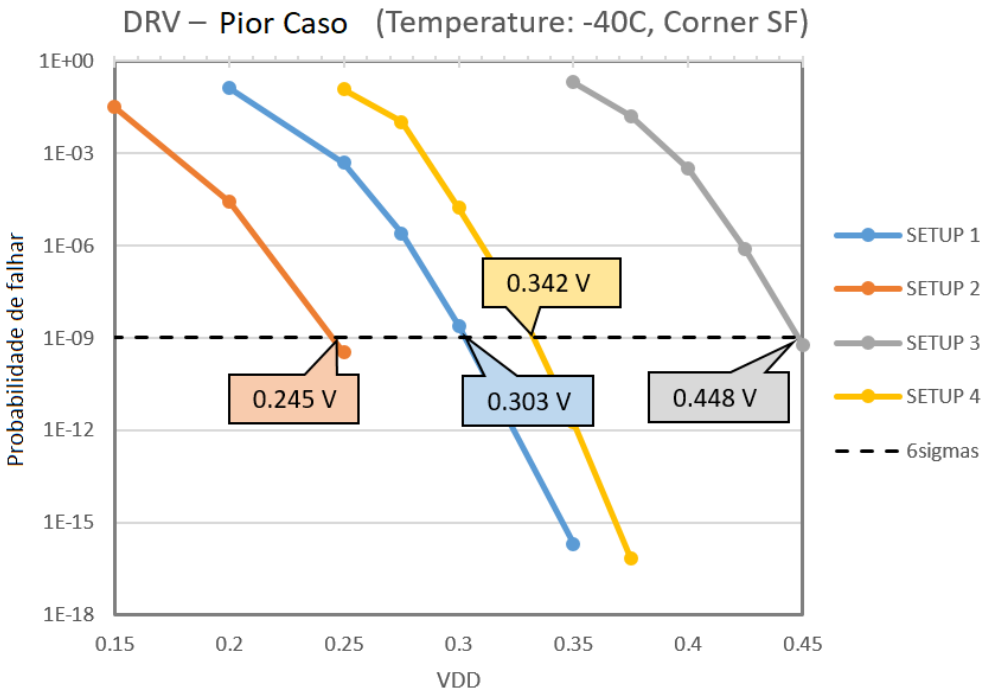


Figura 24: Simulações de DRV.

A tensão de Retenção de Dados (DRV) é a tensão mais baixa que a SRAM ainda

pode reter a informação dentro da célula. A fig. 24 mostra as simulações de DRV no pior caso e as configurações de tensão de corpo selecionadas. O pior caso é o limite SF com temperatura -40° C. Ambos os limites transversais são ruins para a estabilidade da célula e degradam o SNM da mesma forma. Por outro lado, a temperatura baixa faz com os transistores tenha uma queda de tensão maior nos seus terminais, limitando a operação de retenção.

As configurações de tensão de corpo são:

- 1. Todos os transistores em polarização padrão de corpo;
- 2. Todos os transistores em polarização direta de corpo;
- 3. Todos os transistores em polarização reversa de corpo;
- 4. Transistores Pull-Up com polarização padrão de corpo e transistores de Acesso e Pull-Down com polarização reversa de corpo.

O gráfico mostra que, com o aumento do VDD, a probabilidade de falha diminui para a operação de retenção. O DRV é o valor de VDD para quando a probabilidade de falhar cai abaixo de 1e-9 (6 σ).

A tabela da fig. 25 apresenta o valor para o pior caso para a corrente de fuga de uma matriz de 16kb, que é o mesmo limite simulado (SF) porém com a temperatura de 125 ℃ para maximizar o valor da corrente de fuga. A potência dissipada é calculado como:$I_{LEAK} \times DRV$.

O configuração de tensão de corpo 1 é o padrão e apresenta DRV = 0,303V. A configuração 2 (polarização direta de corpo completa) apresenta a menor tensão DRV = 0,245V, mas a maior potência de fuga, P_{LEAK} = 19766 pW. Por outro lado, a configuração 3 apresenta a maior tensão, DRV = 0,448 V, e a menor potência de fuga, P_{LEAK} = 67 pW.

Portanto, a configuração 3 quase o dobra o valor de DRV, mas reduz o P_{LEAK} em aproximadamente 300 vezes. Como o P_{LEAK} é a métrica crítica para retenção, a configuração 3 é a mais adequada. Dependendo se aplicação requerer um valor mais baixo para a

SETUP	PU (M2 M4)	PD (M1 M3)	ACC (M5 M6)
1	0	0	0
2	-2	2	2
3	2	-2	-2
4	0	-2	-2

SETUP	DRV	LEAKAGE (A)	P (pW)
1	0.303	3.26E-10	987.6
2	0.245	8.07E-09	19766.8
3	0.448	1.51E-11	67.6
4	0.342	4.50E-11	154.0

Figura 25: Configurações de DRV.

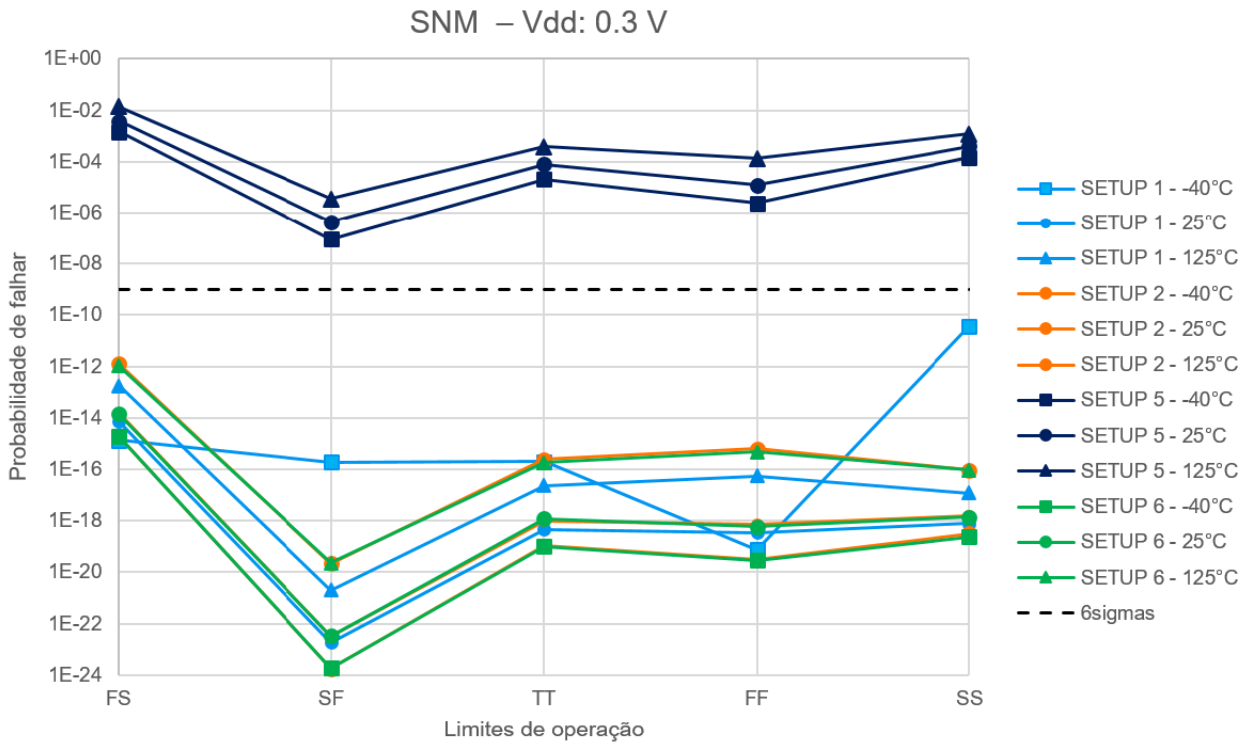
tensão DRV, a configuração de tensão de corpo 4 pode ser uma boa opção, pois apresenta um bom equilíbrio entre P_{LEAK} e DRV.

4.2 Operação de Leitura

A fig. 26 representa a simulação da Margem de Ruído Estático (SNM). O VDD foi fixado em 0,3V e a simulação considerou todos os limites (FS,SF,TT,FF,SS) e as temperaturas (-40 ℃, 25 ℃, 125 ℃) para 4 configurações diferentes de tensões de corpo. Para uma determinada configuração ser considerada adequada para a operação de leitura, a probabilidade de falhar em todos os limites e temperaturas deve ser inferior a 1e-9.

- 5. Transistores Pull-Up com polarização padrão de corpo e transistores de Acesso e Pull-Down com polarização direta de corpo.
- 6. Transistores Pull-Up e Pull-Down (Inversores) com polarização direta de corpo e transistores de Acesso com polarização padrão de corpo.

As configurações de tensão de corpo 1, 2 e 6 são adequadas para a legibilidade. Analisando a configuração 5, os inversores da célula são desequilibrado pela aplicação



SETUP	PU (M2 M4)	PD (M1 M3)	ACC (M5 M6)
1	0	0	0
2	-2	2	2
5	0	2	2
6	-2	2	0

Figura 26: Simulações SNM para todos os limites do processo e 3 temperaturas considerando as configurações de tensão de corpo 1, 2, 5 e 6.

de uma polarização direta do corpo apenas no NMOS que, nessa configuração, tem um grande impacto na estabilidade da célula. Por outro lado, quando é aplicada uma polarização direta ao PMOS ao mesmo tempo que no NMOS (Casos 2 e 6), a estabilidade da célula é recuperada e melhorada em comparação com o caso padrão.

As configurações de corpo usadas na operação de leitura que não foram apresentadas no capítulo anterior são:

Na fig. 27 é mostrada as simulações dinâmicas para operação de leitura considerando as configurações de tensão de corpo adequada para a operação: 1, 2 e 6. Essas

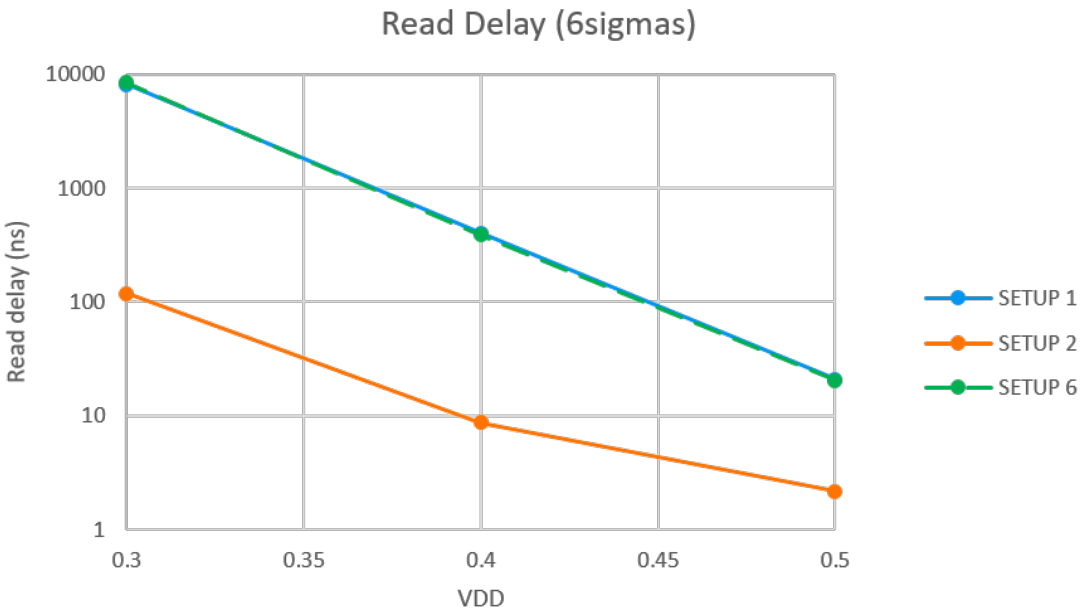


Figura 27: Simulação do atraso de leitura para as configurações de corpo 1, 2 e 6.

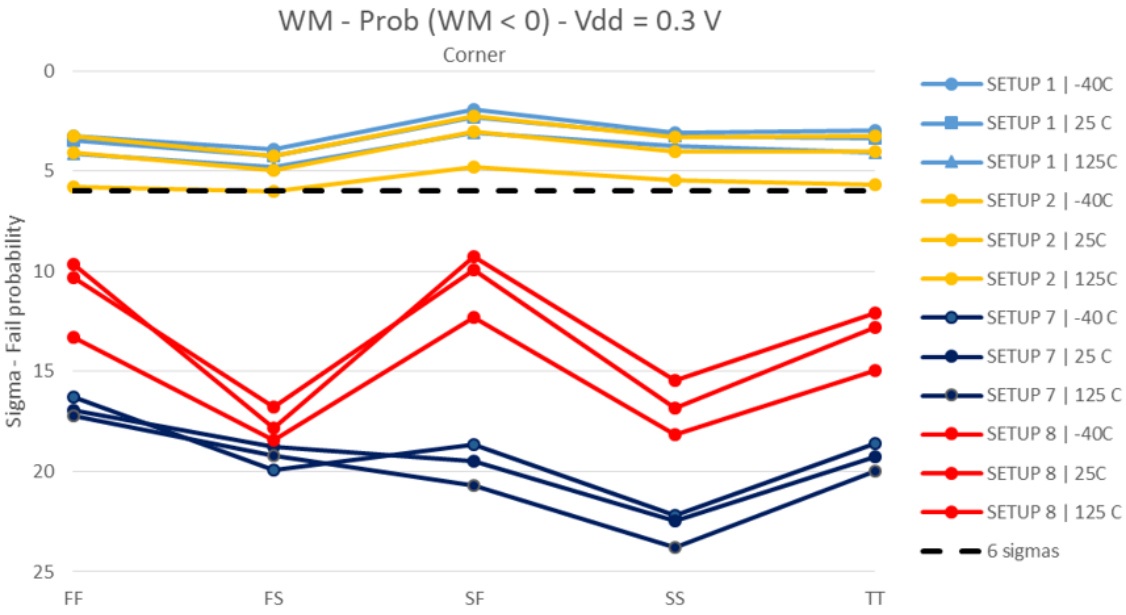
simulações foram realizadas com temperatura de 25 ℃ e o limite TT. O Atraso de leitura é o tempo no qual a Wordline deve permanecer ativada para que a probabilidade de falhar seja de 6 sigmas (1e-9). As configurações 1 e 6 apresentam a mesma velocidade de operação. Pode-se concluir que colocando uma polarização direta apenas nos transistores dos inversores e deixando os transistores de acesso com a tensão de corpo padrão não altera a velocidade de operação em relação a polarização de corpo padrão, logo os transistores de acesso desempenham um papel importante na operação de leitura.

A configuração 2 corresponde a todos os transistores em polarização direta de corpo e é o melhor para a operação de leitura, apresentando 100ns de atraso de leitura com 0,3V de tensão de alimentação (VDD) e 2ns para 0,5V de VDD.

4.3 Operação de Escrita

A fig. 28 representa a simulação para margem de escrita (WM). O VDD foi fixado em 0,3V e a simulação considerou todos os limites de processo e 3 temperaturas (-40 ℃, 25

℃, 125 ℃) para 4 configurações de tensão de corpo diferentes. Para uma determinada configuração ser considerada adequada para a operação de escrita, a probabilidade de falha em todos os limites de processo e temperaturas deve estar abaixo de 6 σ (1e-9).



SETUP	PU (M2 M4)	PD (M1 M3)	ACC (M5 M6)
1	0	0	0
2	-2	2	2
7	-2	-2	2
8	0	-2	2

Figura 28: Simulação para margem de escrita.

As configurações de corpo usadas na operação de leitura que não foram apresentadas nos capítulos anteriores são:

- 7. Transistores Pull-Up com polarização direta de corpo, transistores de Acesso com polarização reversa de corpo e Pull-Down com polarização direta de corpo.
- 8. Transistores Pull-Up com polarização padrão de corpo, transistores de Acesso com polarização reversa de corpo e Pull-Down com polarização direta de corpo.

As configurações 1 e 2 não são adequadas pois apresentam uma probabilidade de falhar acima de 6σ em praticamente todos os casos. Por outro lado, as configurações 7 e 8 são adequadas para a operação de escrita com VDD = 0,3V. Fazer esta operação com essa tensão é muito difícil, mas com as configurações 7 e 8, os transistores pull-Down estão em polarização reversa e Pull-Up estão em polarização direta ou padrão, a célula se torna muito instável, facilitando a rotação dos inversores.

Na fig. 29 é mostrado a simulação dinâmica para a operação de escrita para as configurações 7 e 8. Com -40 ℃ para os limites TT e SF. O atraso de leitura é o atraso é o tempo em que a WordLine deve ficar ativo para que se atingir 6σ (1e-9) de probabilidade de falha.

O Setup 7 apresenta a operação de escrita mais rápida entre as duas configurações. Com pouca diferença entre os limites de fabricação TT e SF. Essa configuração tem um tempo de leitura de 1μs com 0,3V e 1ns para 0,5V.

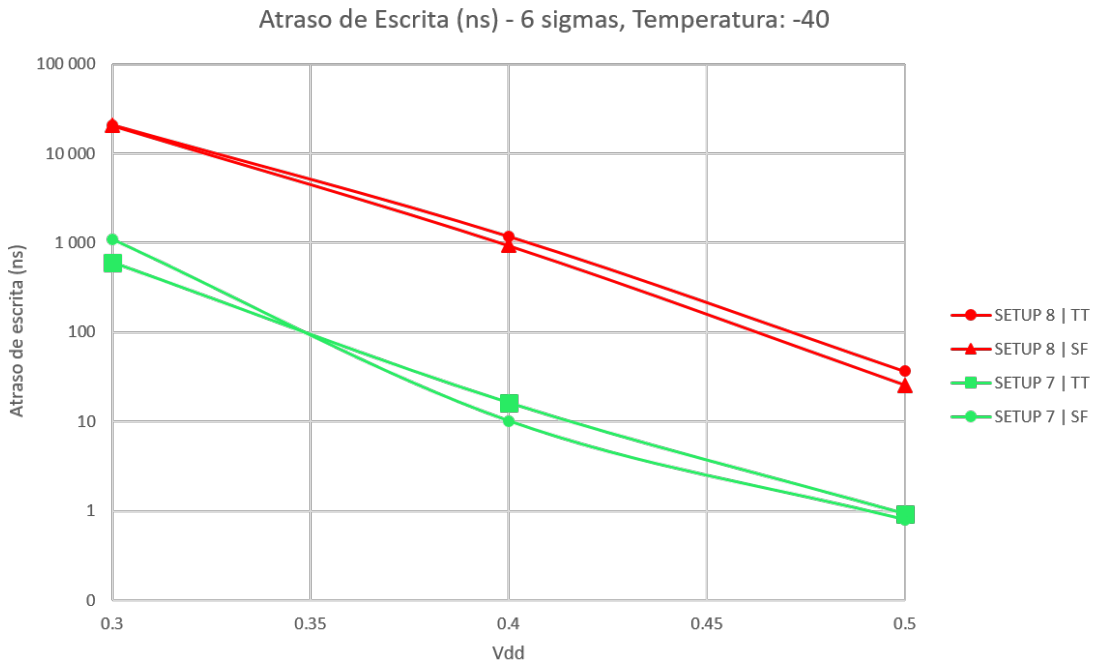


Figura 29: Simulação dinâmica para a operação de escrita.

A configuração responsável por ter a melhor operação de escrita, a configuração 7,

pode ser problemática. Simulações mostram que esta configuração coloca a célula de bit em um modo muito instável. O layout de uma matriz convencional de células SRAM dispõe as linhas de tensão de corpo, bem como a Wordline de forma horizontal. Dessa forma, aplicando a tensão de corpo para a operação de escrita em uma célula e ativando a Word Line, essa tensão será propagada para as outras células da mesma linha, podendo então destruir as informações armazenadas nas outras células.

Em relação a este problema, sem alterar o layout, com as seguintes configurações: PU = 1V, PD = -1V, ACC = 1V e VDD = 0.5V esta configuração não destruirá as informações das bitcells que compartilham a mesma linha e atingem 6 σ de probabilidade de falha.

Ainda é possível alterar o layout, colocando as linhas que alimentam as tensões de corpo de forma vertical compartilhando a mesma coluna. Isso resolveria esse problema, pois as células das outras linhas permaneceriam em estado de retenção e não sofreriam perigo de perder os dados. Porém esta missão permanece como trabalho futuro neste assunto.

4.4 Conclusão Sobre as Configurações de Tensão de Corpo

Foram apresentadas 3 configurações de tensões de corpo diferentes para as operações de célula SRAM: retenção, escrita e leitura. Esses modos permitem que a célula SRAM tenha um alto desempenho em tensões muito baixas e uma corrente de fuga muito pequena em modo de espera. Os modos estão resumidos na figura 30.

Retenção

SETUP	PU (M2 M4)	PD (M1 M3)	ACC (M5 M6)
3	2	-2	-2

Escrita

SETUP	PU (M2 M4)	PD (M1 M3)	ACC (M5 M6)
2	-2	2	2

Leitura

SETUP	PU (M2 M4)	PD (M1 M3)	ACC (M5 M6)
7	-2	-2	2

Figura 30: Configurações de tensão de corpo para cada operação.

Lembrando que na fabricação do CMOS tradicional, a limitação nos poços NWELL e PWELL que formam um curto circuito é:

$$V(PWELL) > V(NWELL)$$

Ou seja:

$$V_{Corpo}(NMOS) > V_{Corpo}(PMOS)$$

As configurações de corpo 2 e 7 não seriam possiveis sem a fabricação 3-D CoolCube.

5 Layout da célula SRAM

O layout de uma célula SRAM deve ser o mais compacto possível, visto que essa célula é encontrada aos milhões em um circuito, seu desenho é realizado de tal forma que elementos que são comuns nas células compartilham o mesmo espaço. Um exemplo de layout é mostrado na fig. 31. Nesta imagem esta indicado os terminais e os transistores. O circuito da célula SRAM é simétricos, logo pode-se identificar um plano espelho no layout indo do canto superior esquerdo ao canto inferior direito. Em vermelho está representada a camada que forma o portão dos transistores NMOS e PMOS, em laranja a área ativa dos transistores e em azul os contatos. Pode-se observar que M3 e M4 dividem o mesmo portão, formando o par inversor. Outro fator importante é que os contatos estão na periferia do circuito, logo serão compartilhados pelo seus vizinhos.

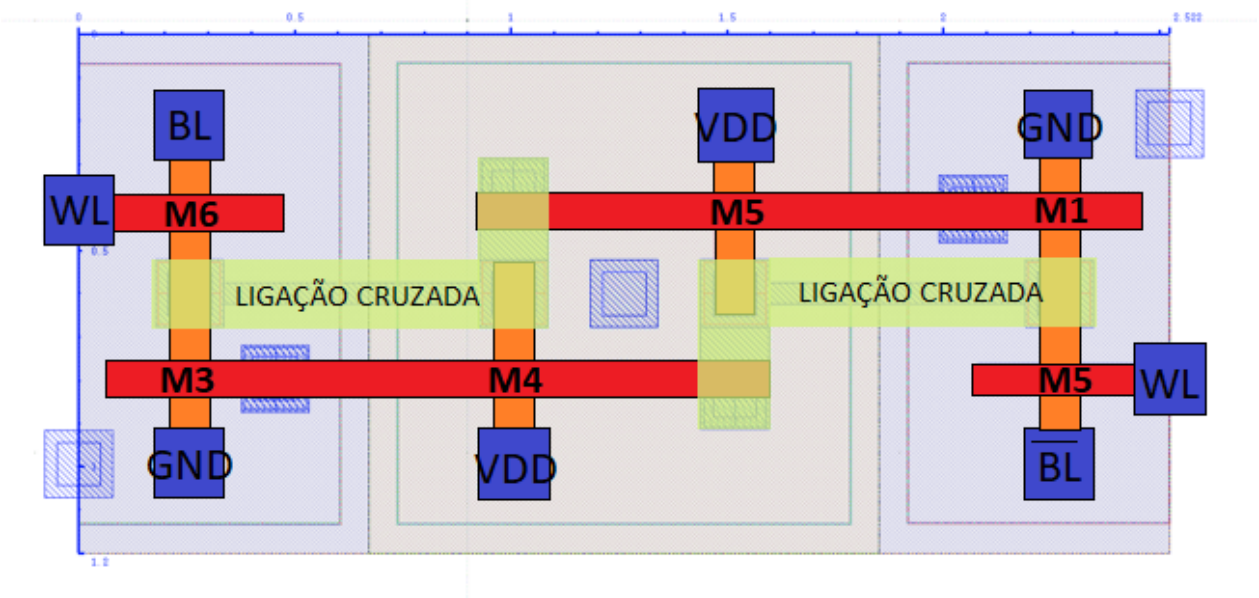


Figura 31: Exemplo de layout de mémoria SRAM.

Esta representado na fig. 32 uma matriz 2x2 afim de mostrar como são compartilhado os elementos das células. Da primeira coluna para a segunda, as células são espelhadas verticalmente, dividindo os terminais de Word Line e os poços PWELL com a célula vizinha. Por outro lado, as células da segunda linha são espelhadas horizontalmente em

relação as células da primeira linha. dessa forma, compartilhando os contatos de Bit Line, GND e VDD. Existem ainda mais camadas de metal, que não estão sendo representadas na figura, criando a matriz de contato que formam as palavras da memória.

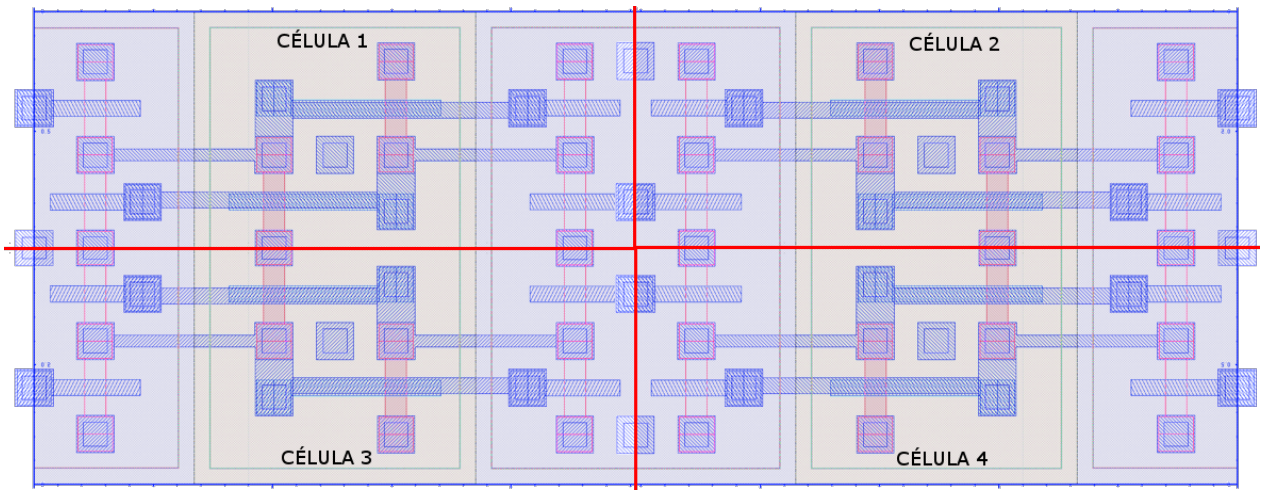


Figura 32: Exemplo de matriz de mémoria SRAM.

5.1 Design para 3-D CoolCube

A última missão do trabalho foi projetar dois layouts de célula SRAM para a tecnologia 3-D CoolCube em FD-SOI 65 nm. Os dois modelos são trazidos nas figuras. 33 e 34. O primeiro modelo apresenta 3 poços para cada célula enquanto que o segundo tem 5.

Conforme visto anteriormente, esses poços se unem quando as células são espelhadas. Assim, a primeira célula possui dois poços para a tensão de corpo, um para os transistores PMOS e outro para os transistores NMOS. O segundo modelo, na matriz, apresenta 3 poços. Este modelo é dividido como considerado nas simulações, uma tensão de corpo para os transistores de acesso, um poço para os transistores de Pull-Up e o último poço para os transistores Pull-Down.

O design das células SRAM busca a máxima densidade de integração, portanto, durante o desenvolvimento dessas células, foi entrado em contato com a DCOS, departamento responsável pelo Kit de Design (DK) 3D CoolCube, departamento que é uma

divisão do CEA - LETI. O contato foi realizado a fim de pedir permissão para violar algumas regras do DK. Para cada modelo foram realizados dois layouts. O segundo layout de cada modelo, apresentado nas imagens, foi desenhado infringindo algumas regras do DK que não terão efeitos negativos na fabricação.

O primeiro modelo após a violação de determinadas regras, encolheu cerca de 35% e o segundo 20%.

Os tamanhos do primeiro modelo são:

- Primeiro esboço: 3.02 μm^2
- Desenho final: 1.99 μm^2

Os tamanhos do segundo modelo são:

- Primeiro esboço: 3.11 μm^2
- Desenho final: 2.48 μm^2

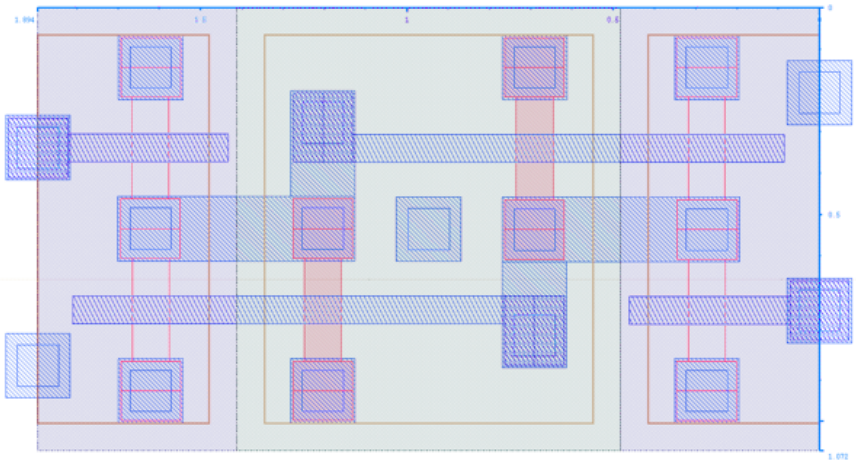


Figura 33: Layout for 2 SRAM bitcell.

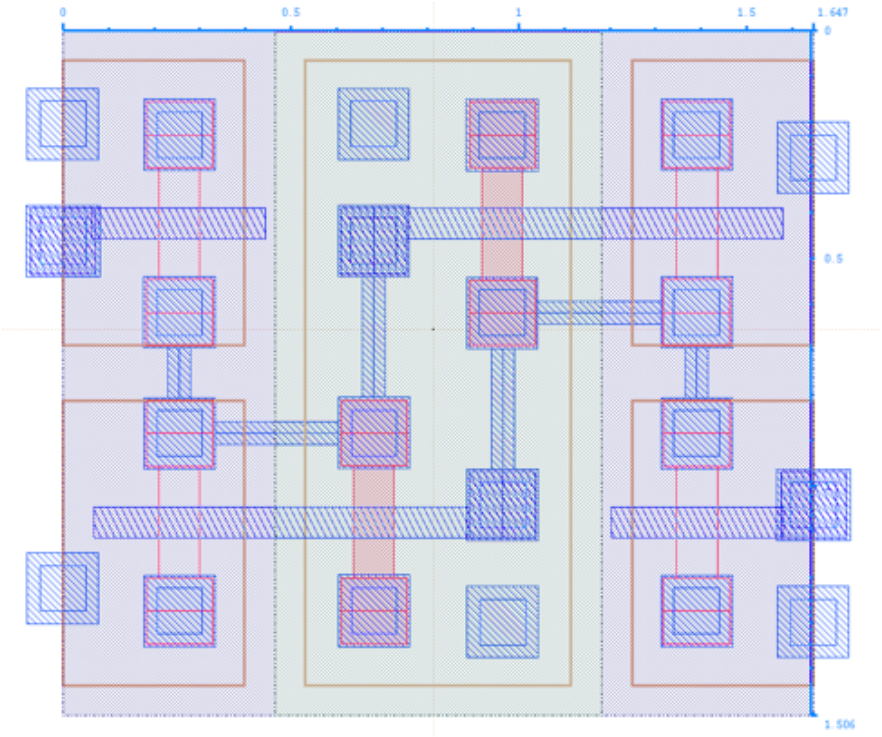


Figura 34: Evolução dos layouts.

6 Conclusão

Conforme a tecnologia de fabricação de transistores avançou, as barreiras foram superadas das mais diversas formas, porém essas soluções sempre tiveram uma coisa em comum: fabricação planar. Atualmente os transistores estão próximos da escala atômica, o que traz desafios que não têm perspectivas de serem superados ou a solução é muito cara. Um exemplo disso é a grande variabilidade do processo de fabricação devido às impurezas no material. Logo, afim de manter a evolução da eletrônica sob a lei de Moore, um novo paradigma é apresentado: Integração sequencial 3-D. Essa tecnologia é realidade, porém ainda muito cara para ser concorrente no mercado. Na esperança que um dia isso mude, circuitos que a muito tempo são conhecidos e já foram explorados exaustivamente podem agora voltar a superfície da pesquisa. Neste contexto, esse trabalho explorou células SRAM na tecnologia de integração sequencial 3-D CoolCube.

Através de simulações foi proposto um método de polarização de corpo dos transistores da célula SRAM que optimiza a velocidade e estabilidade das operações de leitura e gravação em tensões muito baixas, e que garante baixo consumo no modo de espera. Esse método supõem que a matriz SRAM possui dependência de operação, ou seja, conforme a operação realizada na palavra de memoria, uma configuração de tensão de corpo especifica será aplicada. Juntamente com essas configurações foi proposto um modelo para memoria SRAM para ser fabricado conforme a integração 3-D CoolCube, dessa forma esse trabalho cobre tanto o estudo e simulação de circuitos integrados quanto o seu desenvolvimento. Os designs apresentados são planares pois a tecnologia esta em um certo grau de maturidade que o processo de fabricação difere o tamanho dos transistores da segunda pra primeira camada. Porém, é possível realizar a matriz na camada superior e a lógica da memória na camada inferior.

Esse trabalho pode servir de base para trabalhos futuros, é necessário que se investigue se o valor proposto para a tensão de corpo, 2V, é ideal. Possivelmente valores intermediários possam entregar uma optimização ainda maior na célula SRAM. Também fica como trabalho futuro verificar a influência global das configurações de tensão de

corpo na matriz.

O saldo do trabalho foi muito positivo em todos os aspectos, pois apesar de grandes dificuldades na composição das simulações, compreensão do circuito e das métricas e o relativo curto tempo para a realização das simulações que possuem um alto custo computacional, todos os objetivos propostos no começo do trabalho foram cumpridos. Também deve-se ressaltar a importância desse trabalho de conclusão de curso para avaliar as competências que um engenheiro deve ter: organização a longo prazo, compromisso com um projeto de grande escala e auto-suficiência.

Referências

[1] Natalija Jovanovic. *Robust and reliable ReRAM-based non-volatile sequential logic circuits in deeply scaled CMOS technologies*. PhD thesis, Telecom ParisTech, 2016.

[2] R. Boumchedda, J. P. Noel, B. Giraud, K. C. Akyel, M. Brocard, D. Turgis, and E. Beigne. High-density 4t sram bitcell in 14-nm 3-d coolcube technology exploiting assist techniques. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2017.

[3] V. Gupta and M. Anis. Statistical design of the 6t sram bit cell. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2010.

[4] Adam MAKOSIEJ. *Conception de circuits memoire statique a access aleatoire tres basse tension en technologie UTBB-FDSOI*. PhD thesis, Ecole Doctorale Informatique, Telecommunications et Electronique, 2012.

[5] E. Seevinck, F. J. List, and J. Lohstroh. Static-noise margin analysis of mos sram cells. *IEEE Journal of Solid-State Circuits*, 22(5):748–754, Oct 1987.

[6] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene. Read stability and write-ability analysis of sram cells for nanometer technologies. *IEEE Journal of Solid-State Circuits*, 41(11):2577–2588, Nov 2006.

[7] K. Leochico and E. John. Data retention voltage analysis of various low-power sram topologies. In *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 913–916, Aug 2014.