UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

EDUARDO CONRAD JÚNIOR

Técnicas de redução de potência estática em memórias CMOS SRAM e aplicação da associação de MOSFETs tipo TST em nano-CMOS

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Sergio Bampi
Orientador

Porto Alegre, Março de 2009.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Conrad Jr, Eduardo; Bampi, Sergio

Técnicas de Redução de potência Estática em Células de Memória para SRAMs / Eduardo Conrad Júnior – Porto Alegre: Programa de Pós-Graduação em Computação, 2008.

112 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2008. Orientador: Sergio Bampi.

1. Microeletrônica. 2. SRAMs 3. Métodos de Economia de Energia 4. T-Shaped Transistors. 5. Low Power Design I. Bampi, Sergio. II. Título.

Agradecimentos

Primeiramente agradeço a Deus pelas oportunidades dadas a mim durante minha maravilhosa vida.

A minha esposa, Cíntia, pelo suporte e terno amor dados a mim durante o mestrado e desenvolvimento de minhas atividades estudantis.

A meus pais, Eduardo e Altiva, por terem me incentivado e me provido condições de desenvolvimento e da busca por novos alvos pessoais. Aos meus irmãos, Guilherme e Juliana, que mantiveram um elo com a realidade e apoiaram-me durante o decorrer da minha vida. A toda minha família, vocês todos são muito importantes para mim.

Ao prof. Dr. Sergio Bampi, orientador e amigo, que primeiramente acreditou nas minhas qualidades e potencial e guiou-me durante esta ultima jornada para o termino de mais esta etapa. Agradeço pelas horas dedicadas a conversas, não somente apreciando assuntos estudantis, mas também sobre o futuro e sobre a busca por novos alvos na minha vida.

Aos amigos e companheiros do Laboratório 110, Fernando, Dalton, Luff, Juan, Alessandro, Luciano, Guilherme, Giovano e Felipe pelo incentivo, amizade e ajuda. Agradeço as horas dedicadas a brincadeiras, trabalho duro, incentivos mútuos e companheirismo. Agradeço especialmente o apoio e incentivo que vocês deram para a realização e conclusão de meu Mestrado. Vocês todos são amigos muito especiais para mim.

Aos amigos e companheiros do GME e do CT1, pela amizade e pelas longas horas de trabalho e estudo compartilhadas durante este período.

Ao Brasil, por proporcionar-me uma Universidade pública, uma pós-graduação de altíssima qualidade e uma bolsa de estudos; acredito que na ausência destas condições provavelmente este momento não estaria acontecendo.

A todos vocês que acreditaram, preocuparam-se e apoiaram-me durante mais esta etapa de minha vida, vocês são co-merecedores desta vitória.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ATD	Address Transition Detection
CMOS	Complementary Metal-Oxid Semiconductor
Corners	Modelos de transistores limites de um processo
DC	Direct Current
DIBL	Drain Induced Barrier Lowering
Die	A superfície onde o circuito de um chip é fabricado
DLC	Dynamic Leakage Cut-off
DRAM	Dynamic Random Access Memory
DRV	Data Retention Voltage
DSM	Deep Sub Micron
DVS	Dynamic Voltage Scaling
DWL	Divided Word Line
ECB	Electron Conduction-Band Tunneling
EVB	Electron Valence-Band Tunneling
Fanout	Carga capacitiva aplicada a saída de uma porta
Foundry	Fabrica de semicondutores
HVB	Hole Valence-Band Tunneling
I/O Ring	Conjuntol de entradas e saidas de um chip
Leakage	Corrente de Fuga
MPW	Multi-Project Wafer
Pad	Ponto de comunicação entre mundo externo e projeto dentro do die
PCB	Printed Circuit Board
RAM	Random Access Memory
RC	Resistance - Capacitance
SA	Sense Amplifier
SEU	Single event Upset
SNM	Static Noise Margin
SOC	System On a Chip
SRAM	Static Random Access Memory
Stand-by	Modo de espera
TAT	Trapezoidal Association of Transistors
Threshold	Tensão do limiar de condução do transistor
TST	T-Shsaped Transistor
UDSM	Ultra Deep Sub Micron

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Em nossos dias a crescente busca por portabilidade e desempenho resulta em esforços focados na maximização da duração de bateria dos equipamentos em fabricação, ou seja, busca-se a conflitante solução de circuitos com baixo consumo e ao mesmo tempo com alto desempenho. Neste contexto usualmente na composição de equipamentos portáteis empregam-se SOC´s (*Systems On Chip*) o que barateia o custo de produção e integração destes circuitos. SOC´s são sistemas completos que executam uma determinada função integrados em uma pastilha de silício única, normalmente possuem memórias SRAM como componente do sistema, que são utilizadas como memórias de alta performance e baixa latência e/ou também como caches. O grande desafio de projeto em memórias SRAMS é a relação de desempenho versus potência consumida a ser otimizada. Basicamente por sua construção estes circuitos apresentam alto consumo de potência, dinâmica e estática, relacionada a primeira diretamente ao aumento de freqüência de operação. Um dos focos desta dissertação é explorar soluções para a redução de consumo de energia tanto dinâmica como estática, sendo a redução de consumo estático de células de memória em *standby* buscando desempenho, estabilidade e baixo consumo de energia.

No desenvolvimento de técnicas para projeto de circuitos analógicos em tecnologias nanométricas, os TST´s (*T-Shaped Transistors* – Transistor tipo T) surgem como dispositivos com características potenciais para projeto analógico de baixa potência. TSTs / TATs (*Trapezoidal Associations of Transistors* – Associação Trapezoidal de transistores) são estruturas *self-cascode* que podem tornar-se uma boa escolha por apresentar redução do *leakage*, redução na área utilizada e com incremento na regularidade do layout e no casamento entre transistores, propriedade importantíssima para circuitos analógicos. Sendo este o segundo foco deste texto através do estudo e análise das medidas elétricas dos TSTs executadas para comprovação das características destes dispositivos. Também apresenta-se uma análise das possibilidades de utilização dos TSTs em projeto analógico para tecnologias nanométricas.

Palavras-Chave: Memória Cache, SRAMs, Células de memória, Redução de Consumo Estático, Redução de consumo dinâmico, Análise de consumo estático, Associações de Transistores, MOSFET.

Static energy reduction techniques for CMOS SRAM memories and TST MOSFET association application for nano-CMOS

ABSTRACT

Nowadays the increasing needs for portability and performance has resulted in efforts to increase battery life, i. e., the conflicting demands for low power consumption and high performance circuits. In this context using SOC´s (System On Chip) in the development for portable equipments composition, an integration of an entire system for a given function in a single silicon die will provide less production costs and less integration costs. SOC´s normally include a SRAM memory as its building block and are used to achieve memories with low latency and short access time or (and) as caches. A performance versus power consumption analysis of SRAM memory building blocks shows a great challenge to be solved. The electrical design aspects of these blocks reveal high power consumption, dynamic and static, and the former is directly proportional to the operating frequency. The design space exploration for dynamic and leakage consumption reduction in these circuits is one of the focus of this work. The main contribution of this topic is the leakage reduction techniques based in performance, stability and low energy consumption for the memory cell stand-by mode.

Among the electrical techniques developed for analog circuits at the 20-100 nanometer scale, the TST (T-Shaped Transistors) rises with potential characteristics for analog low power design. TST /TAT (Trapezoidal Associations of Transistors) are self-cascode structures and can be turning into a good alternative for leakage and area reduction. Another point is the increment in mismatch and layout regularity, all these characteristics being very important in analog designs. The TST electrical measurements study and analysis are developed to show the device properties. An analysis of the TST desired properties and extrapolation for nanometer technologies analog design are also presented.

Keywords: Cache Memory, SRAMs, Cell memory, Leakage consumption reduction, Dynamic consumption reduction, Leakage consumption analysis, Association of MOSFET transistors.

1 INTRODUÇÃO

Em nossos dias verifica-se a multiplicação do conhecimento e da tecnologia, em continuo desenvolvimento, tornando-se cada vez mais acessíveis e influenciando visivelmente o modo de vida através dos aparelhos eletrônicos, dos sistemas de comunicação, dos computadores pessoais, etc. Assim o usuário se acostumou com as facilidades proporcionadas e torna-se cada vez mais crítico quanto às qualidades e funcionalidades de um aparelho que deseja adquirir. Os equipamentos devem agregar as características desejadas, a não aceitação das limitações de operação impostas pelo fabricante é o palco atual onde a insatisfação é gerada principalmente pela disponibilidade de produtos concorrentes equivalentes que podem atender a esta demanda. Este comportamento do consumidor leva a uma competição cada vez mais acirrada no mercado de eletrônica pois passa-se ao próximo produto quando o ora manuseado não atende as ansiedades impostas. *Time to Market* e agregação de novas funções nunca foram tão importantes como agora e imbuídos neste panorama de competição acirrada temos a produção de circuitos dedicados como um dos pontos diferenciadores entre as empresas. A procura por incorporar no produto todos os anseios do consumidor, este cada vez mais exigente, leva aos projetistas de circuitos integrados (CI´s) o requisito de superar barreiras e desafios no processo de projeto e de integração. Neste contexto de desenvolvimento de CI´s dedicados temos a crescente demanda por mobilidade e miniaturização, ou seja, acesso ao que se deseja a qualquer hora e lugar, utilizando sofisticados dispositivos móveis.

Notadamente a vida moderna caminha no sentido da mobilidade e à agilidade agregada a esse conceito, a vida não se passa somente em um escritório, temos relações, trabalhos e funções que são globais não se fixando mais a tempo e local determinados. Assim percebe-se que a portabilidade está intimamente ligada à durabilidade das fontes de alimentação, que é tempo de operação sem recarga, com desprendimento de ponto de alimentação sendo esta uma das características crescentemente almejada pelos consumidores. Assim a portabilidade se torna um fator determinante onde anteriormente o desempenho imperava sozinho, claro ainda existindo nichos específicos para equipamentos com uma só destas características. Os circuitos de alto desempenho e consequente alto consumo e os de baixo desempenho e consequente baixo consumo tem mercado em nichos próprios e não mais no âmbito do consumidor comum que busca um equilíbrio entre estas características. Percebe-se ainda que a miniaturização dos circuitos tornou-se um caminho sem retorno quando se pensa em mobilidade, não existindo mercado para soluções obsoletas com tamanho e peso elevados. Desta forma buscam-se hoje soluções conflitantes de projeto, que são circuitos com baixo consumo, alto desempenho e ao mesmo tempo tamanho reduzido. Todos estes fatores implicam em miniaturização do circuito, da fonte de energia, do tamanho da placa de circuito impresso (PCB) empregada e nas demais escolhas de projeto levando a busca constante

por soluções inovadoras. A superação dos limitantes ora impostos pela tecnologia gera a demanda por novas técnicas e soluções, especialmente as que gerem baixo consumo de energia agregada a baixo valor de produção. Como se sabe há uma necessidade crescente por maior poder de processamento nos equipamentos modernos, este ligado diretamente ao aumento de consumo de energia e de área de circuito, a necessidade de menores custos de produção são alguns dos fatores que impulsionaram e impulsionam a indústria microeletrônica a desenvolver-se seguindo a "Lei de Moore". O caminho de miniaturização das tecnologias e a consequente redução da área consumida proveniente da maior capacidade de integração por unidade de área tem sua contrapartida no aumento crescente de transistores por unidade de área, exigindo um equilíbrio do consumo de energia através da redução da tensão de operação e das capacitâncias associadas.

Nas implementações em tecnologias submicrométricas temos circuitos cada vez mais complexos tanto por empregar um maior número de transistores, bem como por exigir maior conhecimento pelo projetista da tecnologia em utilização. O hiato criado pelo fator humano no projeto é um problema no acompanhamento e aprendizado das novas gerações de tecnologias de fabricação, estas cada vez mais complexas e com redução no intervalo de tempo de atividade entre tecnologias. As gerações de tecnologias recentes impõem novos desafios e características anteriormente negligenciadas tornam-se fatores primordiais que dificultam a adaptação dos projetistas a este novo modo de projetar. Desta forma o estudo de novas soluções e técnicas para minimizar o consumo e sua consequente aplicação em circuitos analógicos e digitais objetivando circuitos mais eficientes geram a necessidade de um alto fator de atualização e reciclagem. Nota-se ainda que em circuitos submicrométricos existe o acréscimo de consumo estático decorrente da miniaturização da porta do transistor, com tamanho e isolação reduzidos devidos a miniaturização criam-se correntes de fuga entre os terminais do transistor. Levando em conta da maior capacidade de integração produzem-se circuitos cada vez mais complexos que empregam um maior número de transistores e consequentemente uma maior potência consumida associada a uma unidade de área, esta tanto dinâmica como estática.

Técnicas de projeto de baixo consumo podem levar a economia de energia significativa, acima de 30%, quando se acrescenta a isto o aumento da vida útil do circuito e o menor custo de ventilação e de encapsulamento evidencia-se o aperfeiçoamento. Assim um produto projetado para menor consumo de energia tem por consequência menor emissão de calor que leva a um menor custo de produção (processos mais baratos) e integração (integração mais barata sem tantas exigências térmicas, por exemplo) criando assim um produto com maior confiabilidade, durabilidade e mais barato para produção e integração. A partir da alta capacidade de integração em projetos modernos deve-se levar em conta a distribuição adequada dos módulos de um circuito que gerem calor excessivo, pois através da alta taxa chaveamento em uma área concentrada gerar calor suficiente para danificar permanentemente os dispositivos ali integrados. Evidencia-se assim que a utilização de técnicas de baixo consumo propiciam reduções de potência necessárias para tornar circuitos complexos passíveis de integração em CIs de tecnologias modernas de fabricação. Do ponto de vista econômico a utilização de encapsulamentos especiais para dissipação do calor e (ou) a escolha por ventilação forçada encarecem o produto final e minimizam possibilidades de seu emprego. Técnicas para distribuição adequada dos circuitos geradores de calor, e redução da energia consumida nos mesmos criam a

possibilidade de utilizar um encapsulamento menos complexo, mais barato, e da não utilização da ventilação forçada que também contribui com a economia de energia de todo o sistema integrado.

Todas as técnicas desenvolvidas caminham no sentido de obter menores custos de fabricação e maior confiabilidade dos circuitos projetados e fabricados, sendo que a aplicação destas técnicas gera um impacto tanto na tecnologia embarcada de novos produtos bem como na economia de recursos naturais em nosso planeta. Notadamente hoje este é assunto de alta relevância, amplamente discutido e muito apropriado no desenvolvimento de técnicas em vários dos níveis de projeto de um CI, com o intuito de trazer produtos competitivos e diferenciados ao mercado. Este panorama levou a criação do conceito da introdução de sistemas completos em uma única pastilha de silício como solução alternativa a clássica composta por vários CI´s discretos interligados através de um PCB. O conceito de SOC (*System on Chip*), sistema em um chip, é decorrente dessa necessidade de miniaturização, compactação, barateamento dos custos de produção, da facilitação da integração dos dispositivos e consequente necessidade de economia de energia. Desta forma a grande maioria dos equipamentos ora produzidos são compostos por SOC´s projetados especificamente para uma determinada função. Este caminho explorando as fronteiras do espaço de projeto na busca por soluções que agreguem um somatório de características normalmente conflitantes criam produtos específicos para uma determinada função e operação, especializados. O desempenho, área consumida, facilidade de integração e consumo de energia normalmente partem em eixos desconexos na exploração do espaço de projeto sendo impossível a exacerbação de todas estas características em um único projeto. Agregado a toda essa problemática tem-se ainda o cliente, este interessado no maior número de funções possíveis agregadas em um mesmo circuito. Através da capacidade de polivalência de um projeto pode-se atingir um amplo espectro de consumidores atraídos pelo status ou justamente pela funcionalidade. Um exemplo típico são os nossos celulares com função mp3, rádio, máquina fotográfica, agenda, etc. Assim o menor tamanho de canal das tecnologias CMOS (*Complementary Metal-Oxide Semiconductor*) modernas possibilita a construção de grandes circuitos em uma pequena área, criando a possibilidade de economizar interligações externas justamente produzindo os subcircuitos do sistema em uma mesma pastilha de silício. A redução de interligações entre CIs externos barateia o produto final, acrescido a este fato temos a redução do consumo pela diminuição das componentes RC (Resistências / Capacitâncias) parasitas na composição do sistema completo, sendo os SOC´s a perfeita aplicação desta ideia como solução a este problema. Dessa forma obtêm-se menor consumo de energia, maior miniaturização, menor número de circuitos auxiliares na pcb, menor número de ligações externas e o principal foco, menor custo de fabricação e integração.

SOC´s são formados por um conjunto de subcircuitos, percebe-se que com o crescente desenvolvimento frequentemente empregam alguma memória na sua composição. As diferenças de velocidade entre os subcircuitos geraram a necessidade de implementação de memórias para manter o fluxo de dados constante entre os componentes do sistema, ou seja, reduzir a espera por dados a serem trocados. Através do estudo de Ariz Phoenix baseado na *Semico Research* temos uma estimativa da utilização crescente de memórias na composição de SOC´s, tornando-se mais de 50% da área dos mesmos a partir do ano de 2008 conforme o estudo citado. Acrescido a isso temos a progressão do mercado de SOC´s com estimativas de crescimento dos atuais 37,4 bilhões de dólares no ano de 2007 para 56 bilhões de dólares em 2012, um crescimento de aproximadamente 9% ao ano. A figura 1.1 traz uma estimativa dos

blocos componentes dos atuais e futuros SOC´s segundo o estudo acima citado. Neste contexto encontra-se o principal problema das memórias, por construção apresentam deficiência na relação potência consumida versus desempenho requerido. Buscam-se soluções para compensar esta constante elevação do consumo de energia devido ao aumento do desempenho necessitado a cada nova geração de tecnologia, sendo este um nicho de produção cada vez mais em foco. Os estudos de técnicas e soluções para viabilizar economia de energia nestes circuitos mantendo um compromisso com um desempenho cada vez mais elevado, geram um processo antagônico no espaço de projeto de memórias. A necessidade de um conjunto de circuitos de baixo consumo e ao mesmo tempo alto desempenho para memórias RAM (*Random Access Memory*) CMOS evidenciando a necessidade do desenvolvimento de novas soluções.

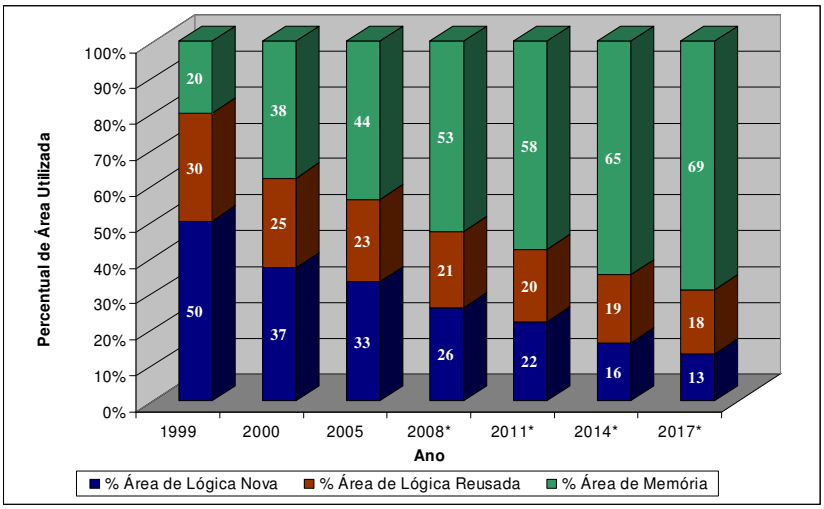


Figura 1.1 – Aumento da área em SOC´s de memória crescente

Memórias SRAMs (*Static Random Access Memory*) em SOC´s podem ter diferentes configurações e arquiteturas para obter diferentes níveis de desempenho ou economia de energia conforme seu nível na arquitetura. Sendo as vantagens principais de sua utilização o baixo tempo de acesso e a baixa latência associados à estabilidade do dado na presença de alimentação. SRAMS são memórias que por seu consumo de área apresentam alto custo de produção, em contrapartida tem maior desempenho quanto à velocidade de escrita e leitura. Sendo memórias rápidas SRAM são utilizadas no nicho de memórias de alto desempenho e através do acréscimo de um controlador apropriado vem a ser utilizadas como caches. Memórias Cache tem a função de transparecer ao circuito que o acesso a dados externos acontece em menor tempo que o requerido nesta operação, funcionando como ponte entre estas trocas de dados. Gerenciando as trocas de dados entre o circuito e a memória externa e/ou periféricos, compensando as diferentes velocidades de acesso sendo assim hoje amplamente utilizadas. Outro emprego das memórias SRAM encontra-se no mercado de sistemas com necessidade de memórias de alto desempenho. Assim por conta de seus aspectos construtivos e alta freqüência de operação em SRAMs haverá um grande consumo potência resultante da necessidade da elevação do desempenho. Em circuitos modernos para maximização do desempenho são utilizados diferentes níveis de memória colocados estrategicamente na arquitetura, para obter um melhor aproveitamento da capacidade máxima de processamento ou redução da energia necessária nestas buscas por dados. A demora em acessar os meios externos é um fator determinante no desempenho final e na energia consumida nas operações; os

periféricos externos são centenas, milhares quando não milhões de vezes mais lentos. As diferenças de tempo de acesso levam ao aumento do tempo de computação e consequente aumento do consumo de energia nos periféricos e no circuito principal. As diferenças de velocidade podem levar a espera do circuito principal pelos dados externos, desperdiçando capacidade de processamento e energia. Assim a implementação de memórias rápidas de menor capacidade, estas utilizadas para que haja informações e dados sequenciais a processar no menor tempo de acesso possível cria o ambiente para um menor consumo de energia sendo esta principal função das memórias cache. A redução do tempo total de espera do circuito por dados, tanto de entrada como de saída, que estão alojados em meios mais lentos traz um incremento de desempenho como demonstrado na figura 1.2 através do exemplo de um processador.

Na produção de circuitos modernos buscam-se soluções arquiteturais que possibilitem maior desempenho, este aspecto torna-se bastante apreciável principalmente no desenvolvimento de novos processadores. O gargalo existente na troca de dados entre o processador e os periféricos normalmente mais lentos, gerou a necessidade de um mitigador para estas diferenças de tempo de acesso que é a memória cache. No intuito de atender as necessidades crescentes de quantidade de processamento e focando no desempenho principalmente, em processadores modernos são utilizados vários níveis de memória cache. Na figura 1.3 vê-se a limitação das memórias DRAM (*Dynamic Random Access Memory*) por seu desempenho não acompanhar em mesma taxa o dos processadores gerando um hiato que justifica a aplicação das caches na arquitetura. A figura 1.4 demonstra os diferentes níveis de memória, a velocidade de operação e o custo de fabricação relacionado. Estas caches com diferentes velocidades e tamanhos, adequados a sua posição na arquitetura, possibilitam que existam dados a processar no menor tempo de acesso possível levando a um menor consumo de energia e maior desempenho sendo essa sua a principal função. Finalmente, segundo Margala (1999) técnicas de projeto de baixo consumo para memórias embarcadas podem levar a economia de energia próxima a 50% em um processador // circuito dedicado. Quando acrescido a isto há o aumento da vida útil dos circuitos e ainda mais a obtenção de menor custo de ventilação e encapsulamento visualizam-se os verdadeiros ganhos.

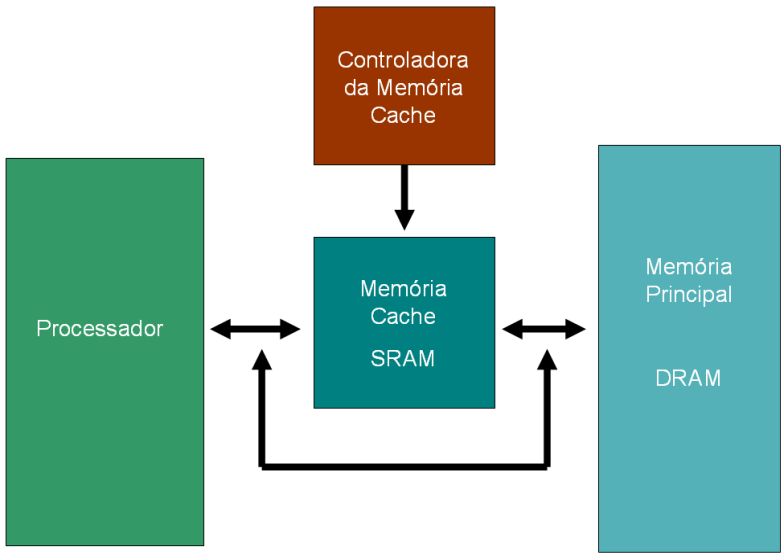


Figura 1.2 – Localização da cache em relação ao processador e memória principal

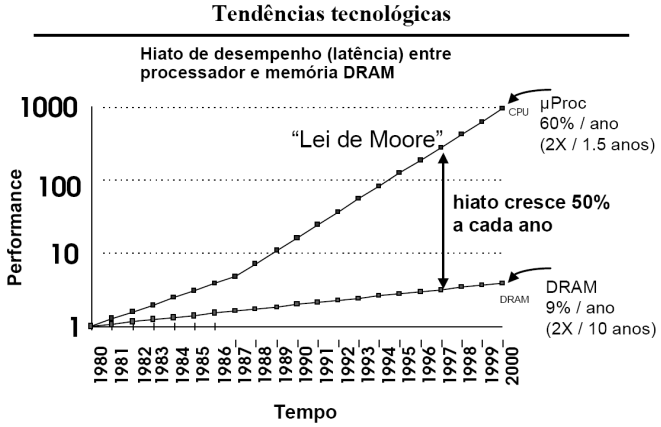


Figura 1.3 – Estrutura das memórias, dados de velocidade, tamanho e custo (Wagner)

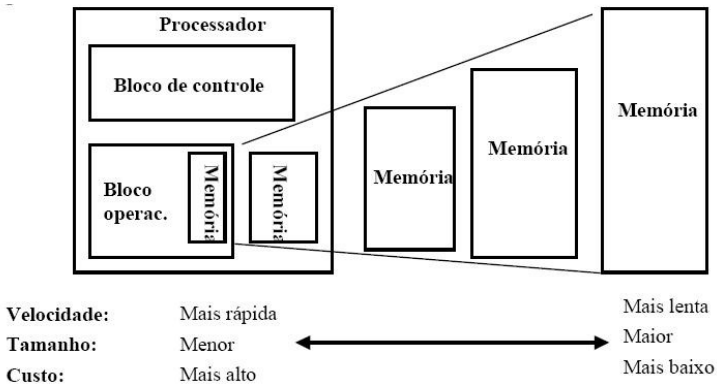


Figura 1.4 – Estrutura das memórias, dados de velocidade, tamanho e custo (Wagner)

Hoje há um consenso nas práticas aplicadas a novas tecnologias de fabricação que para obter um baixo consumo de energia é necessário agir em todos os níveis de hierarquia do projeto. Desta forma as melhorias são necessárias na tecnologia de fabricação, nos dispositivos, nos circuitos, na lógica, na estrutura da arquitetura, no comportamento do algoritmo e nos níveis de sistema; sendo estas condições demonstradas separadamente em níveis na figura 1.5. Em dispositivos para aplicações portáteis observa-se a necessidade crescente do desenvolvimento de técnicas focadas na obtenção de baixo consumo de energia e alto desempenho conjuntamente quando este é solicitado pelo usuário.

A necessidade e busca por equipamentos portáteis que obtenham a maior autonomia e também maior número de funcionalidades é um dos principais focos do mercado atual. Neste nicho de produção verifica-se a necessidade do estudo de soluções arquiteturais e técnicas para viabilizar economia de energia levando em conta um compromisso com o desempenho, este cada vez mais elevado. Para isso durante o tempo em que não há acesso externo há a necessidade de desligar as partes que geram continuamente consumo estático excessivo, sendo este conhecido como modo de espera (*stand-by)* é um aspecto fundamental para qualquer dispositivo portátil. As demais técnicas decaem em um conjunto de conceitos sendo a dissipação de potência é minimizada com a redução da tensão de alimentação, redução da variação de tensão nas linhas, redução da capacitância física, redução da atividade de chaveamento, ou uma combinação destas reduções assumindo sempre que uma redução de desempenho não controlável não é aceitável.

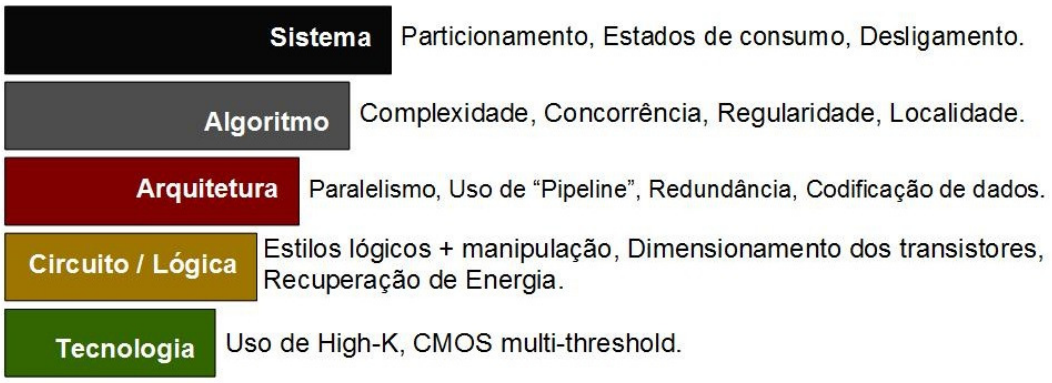


Figura 1.5 – Uma metodologia para baixo consumo requer otimizações em todos os níveis de abstração do projeto.

Como SRAMs são amplamente utilizadas apresentam um desenvolvimento continuo de técnicas para baixo consumo, soluções de projeto tanto analógicas como digitais. São vistas geralmente como simples circuitos digitais, mas considerando que nestes circuitos circulam uma variedade de sinais analógicos e também a alta frequência de operação, podem ser consideradas circuitos mistos analógico-digitais complexos. No estudo de técnicas de baixa potência aplicadas a memórias SRAM temos fatores conflitantes entre redução do consumo versus desempenho exigido da mesma na exploração de seu espaço de projeto. Por construção estes circuitos apresentam deficiências na relação potência consumida versus performance, o aumento de desempenho requer maior consumo dinâmico e hoje mais visivelmente também o estático. A elevação de frequência necessária ao aumento de desempenho leva a maior consumo de energia nas linhas, maior corrente de fuga nos dispositivos provocada pela consequente necessidade da elevação da tensão de operação e maior corrente consumida nos blocos para atingir as exigências deste modo de operação. Desta forma o projeto de SRAMs eficientes energeticamente e com desempenho elevado é um problema de trabalho atual e que promove a criação de novas técnicas possivelmente aplicadas a outros nichos de projeto, demonstrando assim a contribuição e necessidade desta área de estudo no panorama da microeletrônica atual. A motivação principal é a necessidade de atender as especificações de consumo de energia e desempenho das novas aplicações. As memórias se dividem em diferentes categorias, por exemplo dinâmicas e estáticas, sendo que estas tem aspectos funcionais e construtivos diferenciados. As memórias DRAM tem menor custo em área ligado a uma menor performance, SRAMs tem maior performance, maior custo em área e maior consumo de energia. O interessante é que muitas das técnicas e circuitos para construção de RAMs de baixo consumo podem ser compartilhadas entre memórias estáticas e dinâmicas, o que é um aspecto empolgante no que tange o reaproveitamento de novos circuitos e soluções. No entanto existem algumas diferenças entre os circuitos das memórias dinâmicas e estáticas que são apresentados na literatura na forma de uma comparação sistemática de blocos construtivos, sendo que uma análise da capacitância nos nós, da tensão de operação e da corrente estática dos circuitos já propicia a diferenciação das mesmas. Memórias estáticas e dinâmicas têm aspectos diferenciados quanto à abordagem de projeto, um dos principais pontos de diferenciação é o consumo estático que é muito mais elevado em SRAMs, sendo um dos pontos fundamentais a ser reduzido em tecnologias CMOS modernas.

O *scaling* das tecnologias veio como a solução para todos os problemas dinâmicos propiciando a redução do consumo por redução da tensão de operação e das capacitâncias e resistências associadas ao circuito, todavia, obteve o fator adverso do aumento do consumo estático. A redução do comprimento efetivo de canal abaixo de 100nm, bem como da espessura do óxido de porta, gerou um consumo estático nos blocos de um circuito, sempre presente por ser uma característica física advinda da redução dos dispositivos e interconexões. Através dos dados de previsões, por exemplo o ITRS (ITRS, 2007), comprova-se que o consumo estático poderá ser de cerca de 60% da energia consumida por um circuito integrado. O consumo estático torna-se assim, um problema de grande relevância técnica nos nossos dias e soluções para reduzi-lo são amplamente estudados e de alta relevância no projeto moderno de CIs. Estas soluções estudadas têm guiado e guiarão os caminhos de projeto dos novos dispositivos e circuitos produzidos hoje e num futuro próximo, sendo que a necessidade de reciclagem dos projetistas se torna mais perceptível através das novas aptidões exigidas. A redução da largura e comprimento de canal e óxido de porta a cada nova geração de tecnologia de fabricação leva a menor tensão de operação e consequente menor tensão de *threshold* (tensão de limiar dos transistores) e aos problemas de corrente de fuga e dificuldades nas relações sinal ruído, sendo que destes problemas decorrentes da miniaturização há a elevação da razão potência ativa por área medido em Watts/m^2. Notadamente a diminuição da tensão de operação é uma necessidade tanto para redução do consumo bem como para propiciar que o fator de potência ativa por área seja sustentável e não danifique o circuito por sua própria operação. Por outro lado obriga a redução das tensões de limiar dos transistores e gerando os problemas do projeto moderno, entre eles a dificuldade da implementação de circuitos analógicos em tecnologias UDSM (*Ultra Deep Sub-Micron*). O consumo estático decorrente da miniaturização é constante, sendo um agravante na potência consumida e também ao circular nas linhas de dados provinda dos transistores de acesso, podendo ocasionar desequilíbrios no funcionamento que levam ao aumento do atraso e consequente aumento de consumo.

A dissertação está focada na redução de potência estática consumida em células de memória SRAMs em tecnologias com canal sub 100nm. Sendo que através deste estudo constrói-se um sólido alicerce para projeto de SRAMs e circuitos auxiliares para aplicações diversas. Finalmente exporemos a utilização de associações de transistores em tecnologias manométricas como uma proposta viável de redução de consumo estático e área. Primeiramente analisaremos a arquitetura e os blocos constituintes de uma SRAM, após isso as soluções para redução de consumo de energia dinâmica e estática, seguido da análise do problema de elevação do consumo estático em tecnologias modernas e finalmente a análise da redução do consumo estático em memórias estáticas, foco deste texto. Neste tópico apresentaremos comparações entre diferentes soluções encontradas na literatura e a partir desta análise se gerará a escolha da arquitetura a ser implementada. Finalmente apresentam-se as associações de transistores tipo T-shaped (TST) com uma breve explanação, medidas experimentais e novas perspectivas aplicadas ao projeto em tecnologias manométricas.

2 ECONOMIA DE ENERGIA DINÂMICA EM SRAM´S

O mundo moderno trouxe a necessidade de portabilidade para os equipamentos e com isto a duração da bateria de um equipamento se torna determinante na funcionalidade de um circuito integrado e na sua utilização em novos equipamentos. Segundo Margala (1999), técnicas de projeto *low-power* para memórias embarcadas podem levar a economias próximas de 50% em processadores. Ao mesmo tempo lembra-se que esta redução de consumo de potência gerará circuitos com maior durabilidade e menor custo de encapsulamento advindo da redução do calor gerado pelo CI. SRAM´s apresentam um rápido desenvolvimento de técnicas para operação em baixa potência e baixa tensão devido ao aumento da demanda por notebooks, equipamentos portáteis e cartões de memória. Neste contexto há uma contribuição crescente de métodos específicos para memórias SRAM´s, sendo esse desenvolvimento aplicado nas técnicas de baixo consumo para circuitos em geral.

Existem diferentes fontes de consumo em uma SRAM pode-se separá-las em estáticas, normalmente retenção de dados, e consumo dinâmico. No consumo dinâmico encaixam-se o consumo dos: - Decodificadores, da matriz de memória em troca de dados, dos circuitos de entrada/saída, dos buffers de escrita e dos de saída, dos *sense amplifiers*, dos comparadores, etc. A Potência dinâmica é proporcional ao número de linhas e colunas, a corrente das células de memória ativas e a perda das não ativadas, da capacitâncias do decodificador, das lógicas envolvidas, da bufferização, da tensão de alimentação, e durante um intervalo de tempo também dos *sense amplifiers*. Não esquecendo a relação direta com a frequência de operação da memória, ou seja, o número de vezes que o ciclo se repete em uma base de tempo. Mostra-se claramente que mesmo havendo a possibilidade de circuitos em espera sempre haverá algum consumo em uma memória.

Pode-se conter o consumo de uma memória via sua controladora ou via alterações diretas nos blocos da parte operativa. Ao tentar-se reduzir o consumo via estratégias na controladora trabalhar-se-á em um nível mais elevado, pensando principalmente em como e quando acionar os bancos para modo ativo ou inativo. Estes métodos de operação gerenciam os blocos da memória buscando um menor consumo, sendo uma forma de projetistas de mais alto nível de abstração obter redução de consumo em memórias. No âmbito da parte operativa de memórias SRAM se agirá diretamente nos blocos componentes da memória, sendo uma forma de melhorar as relações consumo versus performance, e promover a possibilidade de novas estratégias a controladora de memória. As chaves para estas técnicas de redução de consumo tanto no modo ativo baseiam-se em: - Redução da capacitância valendo-se de *bitlines* hierárquicas, ativação

transversal de uma única *bitline*, operação pulsada com gerador ATD (*Analog Transition Detection* – Detector de transições analógicas), redução da variação dos sinais nas linhas de alta capacitância de pré-decodificação, escrita em linhas de barramento e nas linhas de dados, redução da corrente dinâmica usando decodificação multi-estágio, redução da tensão de operação, método de leitura de baixa potência utilizando um amplificador de transferência de cargas, esquema de *word-line* amplificada ou esquema de corrente máxima nas operações de leitura e escrita, técnica de *Auto-Backgate* controlando duplo Vth e finalmente técnicas para reduzir perdas provocadas por corrente dinâmica.

2.1 Técnicas de Operação em Baixa Potência para Redução de Consumo Dinâmico

Como objetivar a redução do consumo em uma SRAM significa conseguir uma redução do consumo global de alguma forma, logo existe um trabalho constante de desenvolvimento de novas tecnologias e formas de reduzir o consumo em memórias, globalmente ou em determinados blocos componentes da mesma. Este estudo continuado a soluções eficazes na redução de potência tais como:

- Redução de tensão de operação;
- Redução de capacitância das *word-lines* e do número de células conectadas a ela; de mesma forma nas linhas de dados, nas linhas de E/S, e nos decodificadores;
- Redução da corrente dinâmica pelo uso de novas técnicas de decodificação;
- Redução da corrente contínua pelo uso de técnicas de operação pulsada em *word-lines, sense amplifiers* e circuitos de periferia;
- Redução de consumo dinâmico pelo uso de múltiplos *thresholds* e/ou *thresholds* variáveis;
- Pré-carga em Vdd/2;
- Redução do chaveamento, ou frequência de operação;

Seguem-se maiores explanações das soluções e empregabilidade das mesmas em soluções para memórias SRAM que são o foco deste trabalho, transcorrendo sobre os assuntos abordando mais especificamente cada um dos subitens apresentados na lista acima.

2.1.1 Redução de Tensão de Operação

Sabe-se que o consumo de uma memória aumenta com o aumento da capacidade da mesma, que é o aumento de bits armazenados, e também eleva-se a potência consumida pelo aumento da frequência de operação que aumenta o número chaveamentos dentro de uma base de tempo definida. A redução de tensão de operação em circuitos integrados tem um efeito grande pois tem correlação quadrática com a potência, exponencial com o *leakage* e linear com o desempenho. Estabelecendo um compromisso entre estas variáveis geram-se diferentes modos de operação. Então visto que pode-se definir as grandezas abaixo como:

- Potência Dinâmica: $P_{dinâmica} = CV^2F$, onde C = capacitância, V = tensão e F = Frequência;

- Atraso: $D = \frac{CV}{k(V - Vth)^2}$, onde C = capacitância, V = tensão e Vth = tensão de limiar dos transistores;

Para minimizar o produto atraso-potência, tem-se:

$$P_{dinâmica} \cdot D = \gamma \frac{V^3}{(V - Vth)^2} \Rightarrow \frac{\partial P_{dinâmica} \cdot D}{\partial V} = \gamma \frac{V^3 (V - 3Vth)}{(V - Vth)^3} = 0 \Rightarrow V = 3 \cdot Vth$$

Desta forma, o melhor compromisso para tensão de operação encontra-se na tensão de alimentação de aproximadamente 3Vth. Visualiza-se que com a redução das portas a tensão de operação tem-se reduzido e por isto há um agravamento desta técnica que depende de margens para correta operação. Para que tal redução aconteça temos de ter transistores menores que possibilitem *thresholds* mais baixos e por consequência menores tensões de operação global. Pode-se notar esta busca intensa continua hoje no desenvolvimento de tecnologias de menor canal, para manter a Lei de Moore válida, e por consequência uma menor tensão de operação. O *scaling* de tecnologias proporciona justamente essa redução de *threshold* e a elevação da capacidade de integração. Como consequência desta redução vê-se a constante elevação da corrente de fuga e por consequência a necessidade de novas soluções para suprimi-la. Outro problema inerente ao aumento de integração é o aumento da dissipação de calor, pois haverá maior chaveamento em menor área e por consequência elevação do calor gerado, demonstrando na segunda variável que influi para que haja redução dos *thresholds* e da tensão de operação. Finalmente vale acrescentar que a redução do Vth se demonstra na redução do desempenho e abaixo de 2Vth de alimentação implicam-se em efeitos negativos para a operação do circuito.

2.1.2 Redução de Capacitâncias

A redução de capacitância tem como principais objetivos a redução de potência e elevação da performance, em ambas as relações interage linearmente. Para isso o projeto utilizando transistores mínimos é recomendado, pois mantem a capacitância baixa nos nós e assim a potência pode ser reduzida. Por outro lado o aumento da área da porta do transistor é utilizada quando a capacitância de carga ou da linha exige o incremento da corrente para acompanhar o *fanout*.

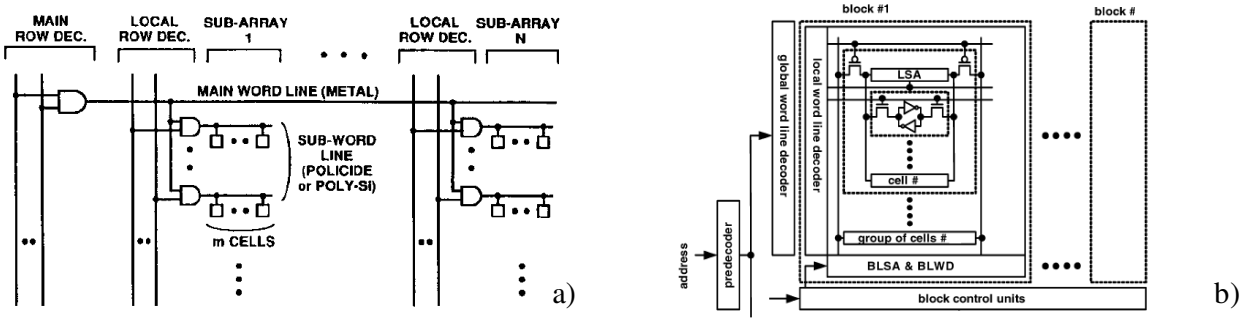


Figura 2.1 – a) Estrutura DWL (MARGALA, 1999) e b) Esquema de *bitlines* hierárquicas (YANG, 2005)

Assim, as maiores capacitâncias associadas a nós em uma memória são as *word-lines*, as *bit-lines* e as *data-lines* dependendo diretamente do número de células ligadas a elas e também seu comprimento. Um método efetivo de diminuir a capacitância é quebrar estas linhas em conjuntos menores e assim consequentemente diminuir o

consumo. O método DWL (*Divided Word Line*) é um método comum em grandes memórias. Também é chamado de método de *Hierarquical Bit-lines*, como demonstrado nas figura 2.1a) e figura 2.1b). Apresenta um decodificador de endereço global e um decodificador de endereços local que assim diminuindo o comprimento das linhas reduz a capacitância associada a ela. Outro fator é que a área do decodificador fica distribuída diminuindo grandemente a capacitância associada a ele e a potência consumida. Assim em uma única *word-line* global o número mais comum de *word-lines* locais conectadas é de quatro, este método de quebra da seleção de uma *word-line* em duas etapas reduz grandemente a capacitância associada a linha de endereço e ao decodificar a coluna também minimiza o atraso RC da *word-line.*

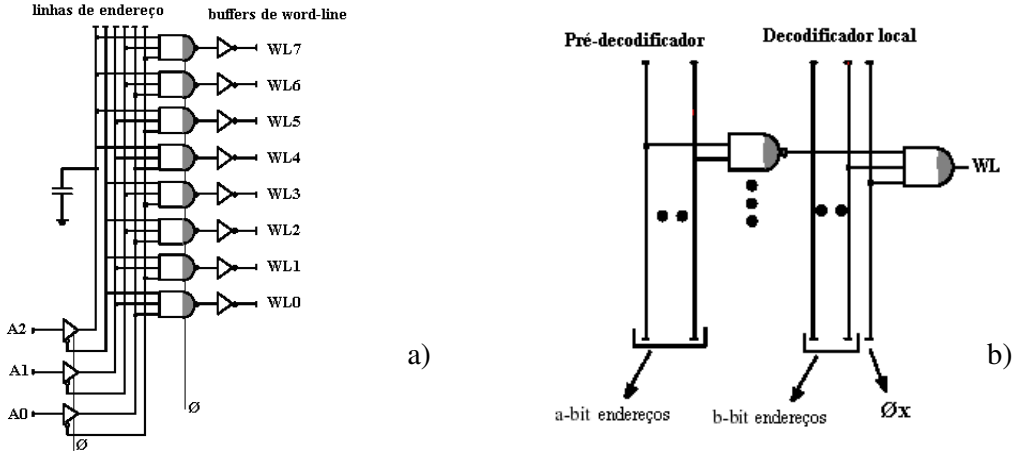


Figura 2.2 –a)Decodificador de três bits com lógica dinâmica e b)Decodificador com dois estágios e lógica dinâmica

Aplicando um sinal Ø de ativação pulsada, tempo determinado para decidir-se a posição do decodificador e fixa-se assim o valor nas *word-lines*. Com o desdobramento em uma decodificação em dois estágios o número de *gates* a carregar diminui, assim a carga a chavear (*fanin)* e o tempo de carga dos buffers de endereçamento são reduzidos. Como resultado há diminuição do consumo de potência e também aumento do desempenho da memória, na figura 2.2 apresenta um exemplo da quebra em dois estágios de um decodificador.

Outra solução é usar o método SCPA, *Single Bitline Cross-point Cell Activation*, ou ativação cruzada da célula de memória como demonstrado na figura 2.3. Esta arquitetura gera a menor corrente, pois aumenta a divisão de blocos reduzindo a área do decodificador. O fato da redução de consumo se deve justamente a somente uma célula de memória ser ativada por vez pelo cruzamento dos endereços X e Y. A desvantagem dessa configuração é a necessidade de amplificação nas linhas X e Y de endereço durante a escrita. Também do valor devido justamente a mais uma queda de tensão no caminho ao gravar ou ler esta memória, pois a mesma apresenta maior resistência nos transistores de passagem que a interligam a *bitline,* dois onde comumente só há um transistor de passagem, ou seja, maior componente RC.

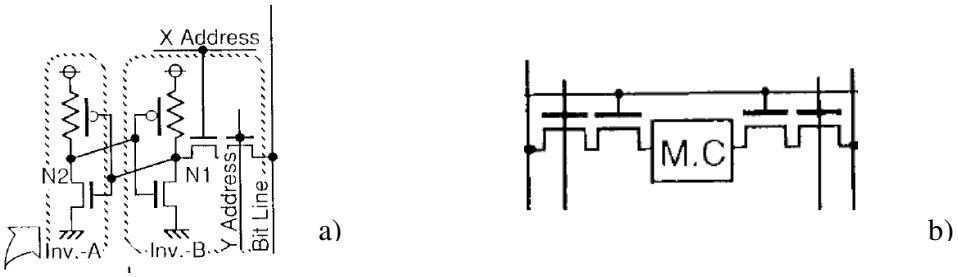


Figura 2.3– a) Estrutura *memory core* com transistores de acesso e b) Esquema de decodificação SCPA (MARGALA, 1999)

Estas técnicas, tanto DWL como SCPA, apresentam a desvantagem de gerar área adicional, maior lógica de controle e também maior roteamento. A vantagem desta divisão em blocos menores em uma memória é a possibilidade de diminuir ainda mais o consumo estático, pois estas técnicas permitem a ativação de pequenos blocos somente no momento de seu uso e de uma espécie de modo de espera no restante da memória. Estas soluções relacionadas a circuitos em espera e redução de consumo estático serão apresentadas no próximo capítulo deste trabalho.

2.1.3 Técnicas de Operação Pulsada

O objetivo principal da operação pulsada é manter ativos na memória somente os circuitos necessários à operação realizada e assim economizar energia estática. Assim podemos ativar uma *word-line,* um decodificador ou mesmo um *sense amplifier* somente no momento que ele se faz necessário. Para tal função é utilizado um gerador de pulsos, ATD (*address transition detection)*, o circuito demonstrado na figura 2.4 é a chave para possibilitar redução de potência ativa em memórias, são constituídos de um circuito de atraso (por exemplo uma cadeia de inversores) e uma porta XOR. O circuito de ATD gera um pulso Ø toda vez que há uma transição de nível alto para baixo ou de nível baixo para alto, assim estes pulsos cortados podem gerar todos os sinais de ativação de uma SRAM. Em algumas memórias são utilizadas matrizes de atraso para composição de todos os sinais a partir de um mesmo gerador, desta forma a variação de processo se dará em todos os sinais ao mesmo tempo reduzindo as possibilidades de falha advindas da variação não correlacionada em sinais com geradores independentes.

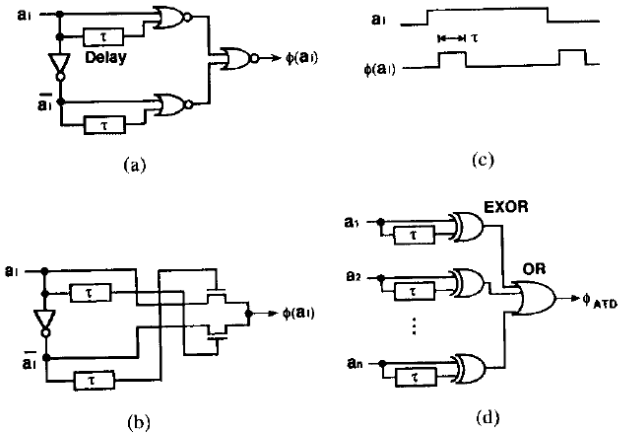


Figura 2.4 - Circuitos de detecção da transição do endereço; a) e b) geradores de pulso ATD, c) Formas de onda do pulso ATD, d) Gerador de pulsos ATD a partir das transições dos endereços (MARGALA, 1999).

2.1.4 Modo de operação Half Swing

A operação pulsada é usada para diminuir o consumo de potência pela redução da amplitude de tensão na variação dos sinais em canais de alta impedância, por exemplo, *bitlines* sem perda de performance. Muito dessa economia vem da operação das *bitlines* em variações máximas de Vdd/2, baseado nas portas *half swing pulse mode.* A redução de consumo se dá pela redução da potência ativa usada nas transições destes sinais ($P=CV^2F$) onde obteremos $(½*V)^2= ¼*V^2$, ou seja, haverá uma redução a ¼ da potência utilizada originalmente.

A figura 2.5a) e figura 2.5b) demonstram dois circuitos que implementam uma porta E *half-swing pulse-mode.* O principio de funcionamento é a união dos níveis de tensão através de um E lógico através de uma meia transição positiva, de Vdd/2 para Vdd e voltando a Vdd/2, outra meia transição negativa, de Vdd/2 para Gnd e voltando a Vdd/2, combinados com uma porta de restauração que leva a saída excursão total. Não há perda alguma de desempenho no receptor por ter entradas com meia transição.

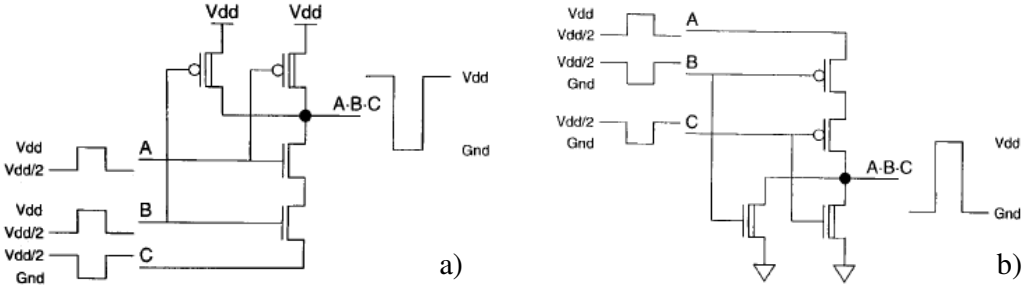


Figura 2.5 – Porta E *Half-swing Pulse-mode* a) Tipo NMOS e b) Tipo PMOS (MARGALA, 1999)

Esta estrutura é combinada com uma porta auto resetável com carregador PMOS para melhorar a margem de ruído e a velocidade da transição do reset na saída. A figura 2.6 mostra uma porta auto resetável *half-swing pulse-mode*.

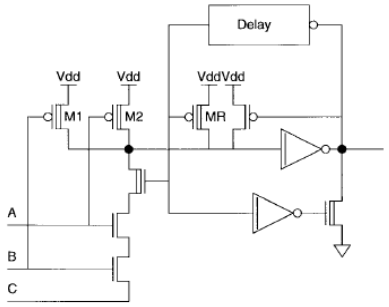


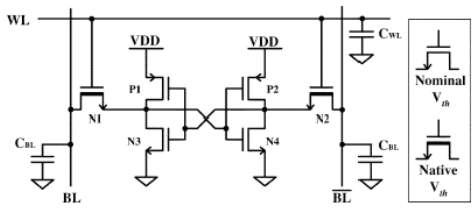
Figura 2.6 – Porta auto resetável *Half-swing Pulse-mode* (MAI, 1998)

Adicionalmente pode-se economizar energia utilizando reciclagem de carga. A carga usada para provocar uma transição positiva pode também provocar uma transição de reset em um pulso negativo. Se as capacitâncias dos pulsos negativos e positivos forem iguais não será drenada corrente da fonte Vdd/2. Esta fonte de Vdd/2 é interna ao chip formada por um conversor DC. Com a combinação destas duas técnicas, *half-swing*

pulse-mode e reciclagem de carga, pode-se obter até 75% de economia de energia nas linhas de alta capacitância.

2.1.5 Uso de múltiplos thresholds e/ou thresholds variáveis

A ideia básica é obter transistores com menor corrente de fuga para a retenção de dados e transistores com maior desempenho para a passagem de corrente, esta característica é obtida através de tecnologia com múltiplos Vth onde apresentam-se transistores de com *threshold* de maior tensão (*High* Vth) ou Nominal, Vth Nativo de *threshold* em tensão média entre os extremos e o *low* Vth com tensão de *threshold* mais baixa. Utilizando os transistores de alto Vth na retenção e os nativos ou *low* Vth para transistores de passagem, multiplexadores (mux), etc, onde haverá maior necessidade de condutividade. Isto traz ao circuito economia de energia e maior desempenho pela redução das componentes RC nas transições dinâmicas. A figura 2.7 mostra um exemplo de uma célula de memória fabricada com esta técnica em tecnologia apropriada.



THRESHOLD VOLTAGES OF NOMINAL nMOS/pMOS AND NATIVE nMOS TRANSISTORS (NATIVE pMOS IS NOT AVAILABLE IN THE 0.25-μm CMOS PROCESS)

	Nominal V_{th}	Native V_{th}
PMOS	V_{thNoP} = -0.53 V	V_{thNaP} = N/A
NMOS	V_{thNoN} = 0.53 V	V_{thNaN} = 0.21 V

COMPARISON BETWEEN HIGH AND LOW THRESHOLD VOLTAGE TRANSISTORS

V_{th}	Characteristic	Advantage
Nominal (0.53 V)	low leakage current	data retention
Native (0.21 V)	high output current, fast switching time	driving capability

Figura 2.7 – Célula de memória com Vth Duplo (WANG, 2003)

O problema desta solução é que os transistores nativos apresentarão sempre maior consumo estático devido seu baixo *threshold* e assim esta é uma solução um tanto quanto mais focada no aumento de desempenho, aumentando o desempenho dos transistores de passagem. Mas aproveitando esta ideia de aumento de desempenho por que não aplicá-la no formato de DLC (*Dynamic Leakage Cut-off*) ou VT (*Virtual Threshold*) CMOS obtendo o melhor destes dois métodos. Pode-se justamente variar a tensão de corpo do transistor para obter esta solução apenas em momentos adequados de acesso para leitura e escrita aumentando o desempenho da memória e assim obter o melhor de ambos os mundos com maior velocidade de escrita e gravação, ao mesmo tempo consumo estático minimizado pela operação para obter um *threshold* elevado. Uma amostra disso pode ser vista em Kawaguchi et. al., conforme a figura 2.8.

Aproveitando a característica de duplo poço da tecnologia empregada no projeto a ser desenvolvido para poder obter a funcionalidade desejada. Acredita-se em ganho de desempenho e redução de consumo estático com o uso dessa abordagem, contudo haverá um aumento de complexidade de controle para que tal técnica seja implementada.

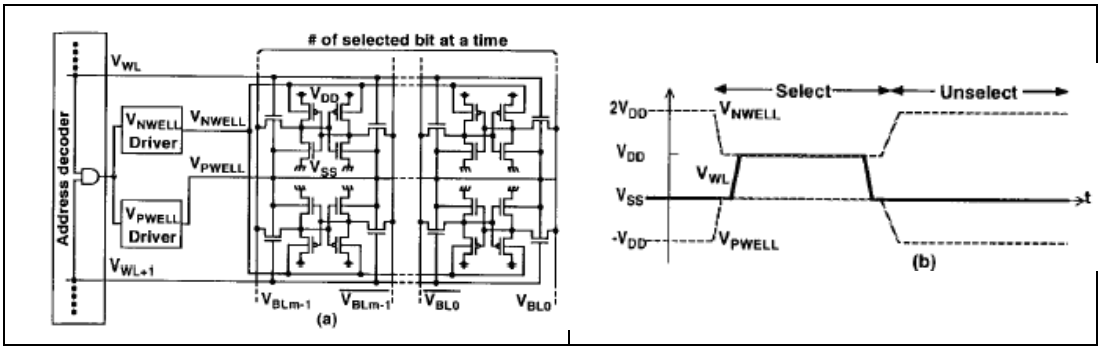


Figura 2.8 – Funcionamento da técnica DLC com a) Esquemático do circuito e b) Operação das tensões de poço (MARGALA, 1999)

2.1.6 Redução de Chaveamento nas linhas

Para redução da potência dinâmica consumida pode-se reduzir a frequência de operação que provoca redução de forma linear. Para a redução do consumo precisa-se de uma avaliação do fator de atividade de um sub-bloco, para então partirmos para soluções específicas como guia das alterações que resultarão em ganhos reais. Em memórias temos uma gama de sub-blocos que trabalham independentemente e podem ser ativados de forma a estarem inativos durante uma parcela do tempo e assim decrescer essa potência dinâmica e estática durante este tempo. Um segundo aspecto é valer-se de diferentes tipos de lógica MOS buscando a redução da potência consumida na operação do mesmo. O terceiro aspecto é o redimensionamento dos caminhos lógicos com o objetivo de redução da carga capacitiva a ser chaveada. Finalmente, um quarto aspecto é reduzir potência em barramentos *tristate* / pré-carregados.

2.1.6.1 Fator de atividade

O fator de atividade (α) é a forma de avaliar quanto um circuito é ativado durante uma base de tempo definida, por exemplo, o ciclo de uma retirada de dado da memória. Desta forma pode-se avaliar o montante de operação neste ciclo e estabelecer a relação: α = (Tempo_Operação / Tempo_do_Ciclo). Esta relação se torna importante alicerce na definição do método de economia de energia a ser aplicado no referido sub-bloco do sistema por ter relação linear com a potência consumida. Reestrutura a potência dinâmica consumida com um fator multiplicativo de correção, assim:

$$P_{dinâmica} = \alpha \cdot (CV^2F)$$

2.1.6.2 Tipo de lógica empregada

A primeira análise a ser tomada em consideração tem relação à função do circuito dentro da arquitetura. Este é um circuito de *clock* ou uma lógica para outras funções como por exemplo decodificadores, multiplexadores, etc. Como segundo ponto qual é o fator de atividade desta lógica, utilizando o valor de α como valor de definição quanto ao tipo de lógica a ser utilizada levando em conta principalmente a corrente de fuga gerada.

Circuitos de Clock

No caso de circuitos de *clock* se o α >30% deve-se utilizar lógica Dominó pois apresentam menor área, menor capacitância de entrada e são mais rápidos. Quando temos α<30% deve-se usar lógica estática CMOS, com ajustes para menor *leakage* como por exemplo o *gate* ser 1,5*Lmin pois reduz-se o *subthreshold leakage.*

O dimensionamento correto dos circuitos no caminho crítico são muito importantes para redução de potência ativa e estática. Utilizando *latches* e *flip-flops* mínimos no caminho de *clock* e na lógica de sequenciamento reduzir-se-á a corrente de curto circuito na mudança de estado e o *leakage* de *gate* por tunelamento que será dispendido por um *gate* maior. Estas lógicas dissipam em torno de 70% da potência total no projeto do processador IBM Power4(RABAEY, 1996). Não há necessidade da lógica de sequenciamento ser superdimensionada, isto leva a uma consequência de elevação das cargas capacitivas, se necessário utilizar buffers nas saídas da lógica de sequenciamento de forma a atingir os requerimentos de atraso. Utilizando o α e as regras de atraso redimensione toda a lógica fazendo com que a capacitância se reduza tanto nos *gates* bem como nas interconexões, balanceando a redução de consumo de energia e o aumento de atraso por acréscimo de resistência na utilização de dispositivos menores.

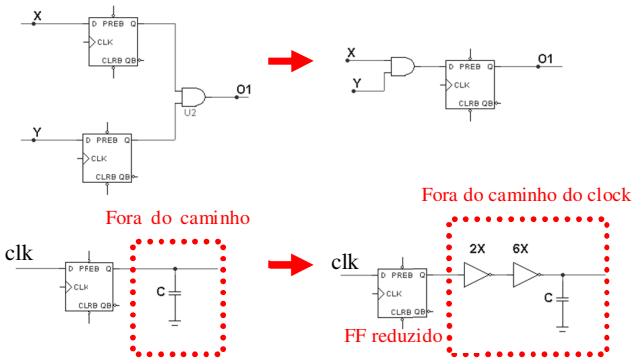


Figura 2.9 – Exemplos de redução de lógica de sequenciamento e redução de capacitância de carga no caminho do clock

A figura 2.9 demonstra alguns exemplos de redução de lógica de sequenciamento reduzindo a energia gasta nos circuitos de relógio (*clock*), produzindo o mesmo resultado lógico. Outra ideia importante é a utilização de ativação de clock para blocos do projeto somente no momento da sua ativação como exemplificado na figura 2.10. Esta abordagem tem maior complexidade de validação e teste, também pode implicar em dificuldades de avaliação do atraso e do *timming* de ativação dos subcircuitos, sendo a principal vantagem deste método é a redução do chaveamento de relógio em um barramento quando este não é necessário o que leva a uma redução de consumo.

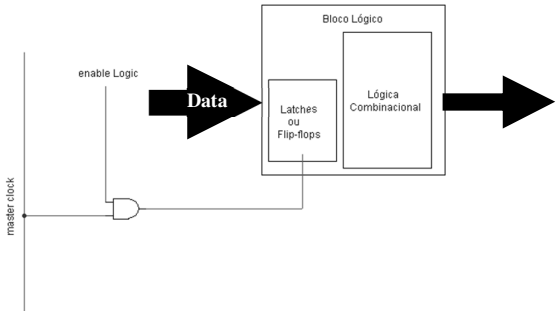


Figura 2.10 – Exemplo de ativação de *clock*

Finalmente em linhas de *clock* deve-se tentar usar linhas o mais largas possíveis, para reduzir sua resistência, com o maior distanciamento possível, buscando menor

capacitância parasita. No que se trata das trocas de metal utilizar o maior número possível de vias buscando a menor resistência nestes caminhos. Estes cuidados nas linhas de *clock* reduzem substancialmente a quantidade de energia dispendida para chaveamento dentro dos limites de atraso permitidos.

Blocos Lógicos

Reduza os dispositivos ao mínimo em blocos de alto fator de atividade e aumente o tamanho dos dispositivos nos de baixa atividade (provendo maior capacidade de corrente) aumentando a velocidade das transições. Garantir um estado de baixo consumo aos circuitos quando colocados em *power-down*, utilizando circuitos com maior eficiência no consumo de energia e obtendo ganhos onde isto se fizer possível.

O ordenamento dos sinais também é importantíssimo, pois pode reduzir *glitches* de ativação e a partir daí obter melhoras na redução do consumo dinâmico. Um exemplo desta técnica é demonstrada na figura 2.11, como regra geral deve-se colocar os sinais de maior AF em um estágio mais ao fim da lógica de ativação, reduzindo a quantidade de *glitches* criados no tempo de avaliação deste sinal.

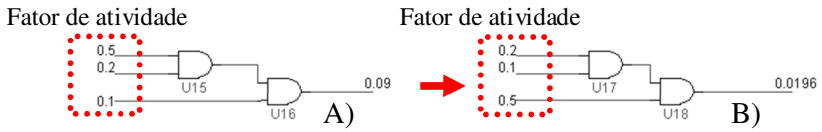


Figura 2.11 – Exemplo de ordenamento de sinais para redução de *glitches*

Outro ponto importante nas lógicas é tentar minimizar a capacitância dos nós de saída dos circuitos. Toda redução de capacitância parasita nos nós de transmissão de sinal podem trazer ganhos em velocidade e consumo de energia, um exemplo é demonstrado na figura 2.12.

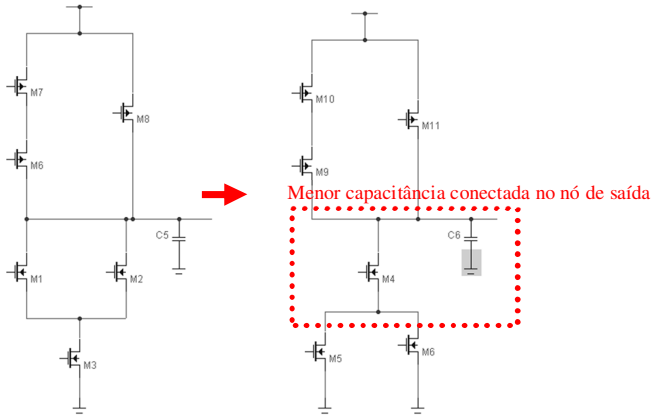


Figura 2.12 – Exemplo de ordenamento de dispositivos para redução da capacitância no nó de saída

2.1.6.3 Ativação somente dos caminhos operantes

A busca pela redução de parasitas leva a análises da arquitetura adotada nos projetos, novamente relevando o fator de atividade de cada um dos blocos do sistema. Na figura 2.13 temos o exemplo de um sistema composto por três blocos, onde nesta arquitetura um bloco é menos acessado e está conectado ao barramento através de um buffer *tri-state* que desconecta o ramo do barramento deste bloco do barramento

principal, fazendo com que a capacitância parasita seja reduzida na maior parte da operação, reduzindo a potência necessária ao chaveamento.

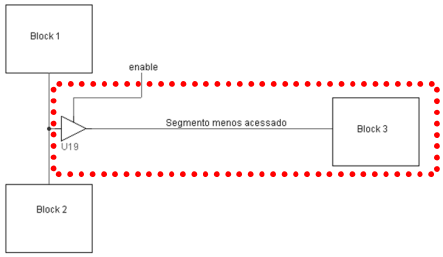


Figura 2.13 – Exemplo de quebra de barramento para redução de capacitância chaveada

A figura 2.14 traz outro exemplo de quebra de barramento, desta vez utilizando um multiplexador como modo de controlar o acesso e assim reduzir a capacitância total chaveada por cada um dos blocos em comunicação. Soluções arquiteturais promovem a redução da energia utilizada no chaveamento das capacitâncias e o tamanho dos dispositivos utilizados nos buffers de saída. Por outro lado estas alterações incrementam a dificuldade da verificação de todo o circuito pois aumentam o número de quebras de caminhos e o número de *test-cases*.

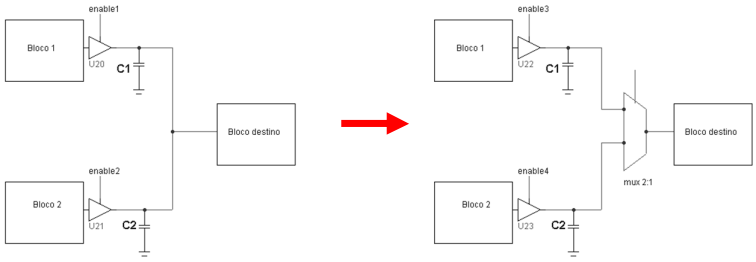


Figura 2.14 – Exemplo de utilização de multiplexador para redução de capacitância chaveada em barramento

2.1.6.4 Redimensionamento dos caminhos lógicos

O redimensionamento dos caminhos lógicos leva principalmente em consideração a potência de curto circuito desperdiçada em portas lógicas complexas. Quando temos portas complexas com grande numero de entradas e dimensionadas para chavear a capacitância do barramento de saída, implicamos em diversos transistores de grande W que levam a uma utilização de maior área e mais energia. A solução mais acertada é a utilização de porta lógica com transistores mínimos e então usar buffers na saída desta porta obtendo assim menor capacitância de saída. Desta forma há menor desperdício de energia em potência de curto circuito, pela redução das capacitâncias a serem carregadas durante a operação normal.

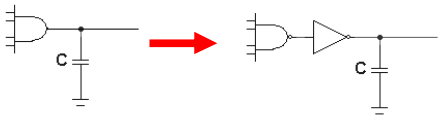


Figura 2.15 – Exemplo de utilização de buffer para reduzir de capacitância na lógica

2.1.7 Técnicas de Leitura e Escrita de Baixa Potência

Um modo eficiente de reduzir a tensão alternada nas *bitlines* e *data-lines* é valer-se de métodos de escrita e leitura baseados em corrente. A ideia básica é reduzir ao mínimo a variação de tensão nas linhas, ou seja, aumentar a corrente transmitida, mas com a tensão dentro de pequenas variações, logo reduzindo a potência alternada consumida. O modo de corrente completo para escrita e leitura consome 70% menos potência que uma metodologia em que somente se emprega a leitura por corrente (MARGALA, 1999).

Apresenta-se então uma nova célula de memória com sete transistores, um transistor de equalização (Meq) foi adicionado a célula de memória SRAM tradicional com seis transistores, quando aberto observa-se a mesma funcionalidade de uma célula de memória comum, sendo apresentado um exemplo desta célula de memória na figura 2.16. Esta célula SRAM apresenta funcionalidades extras para a escrita, primeiramente pelo acionamento de Meq equalizando ambas entradas dos inversores da célula de memória com a redistribuição das cargas ali armazenadas levando a memória a um meta-estado, estado instável próximo ao valor médio entre Vdd e Vss em que o estado lógico esta indefinido. Em um segundo momento mantendo-se o transistor Meq acionado e aí acionando os transistores de passagem da memória tem-se acesso a *bitline* previamente equalizada e já ligada ao buffer de escrita. Observa-se que na escrita por corrente o transistor funcionará como um resistor e pela circulação da corrente através do mesmo haverá uma queda de tensão, ao desativar-se Meq os nós da memória irão automaticamente para o valor entrante sem ocorrerem grandes variações de tensão nas *bitlines*. A redução da variação da tensão é a principal vantagem deste método e por consequência diminui-se a potência consumida com uma maior velocidade de escrita (WIECKOWSKI, 2004).

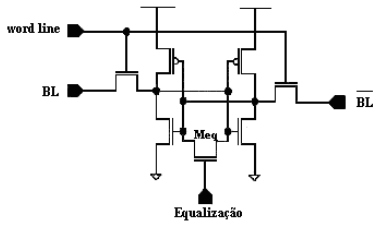


Figura 2.16 – Célula de memória por corrente com sete transistores (MARGALA, 1999)

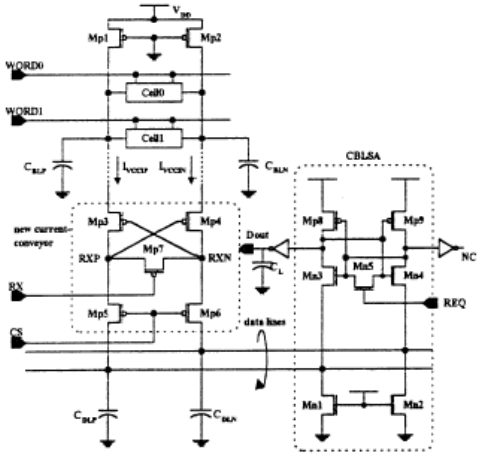


Figura 2.17 – *Sense Amplifier* de leitura por corrente (MARGALA, 1999)

Em *sense amplifiers (SA)* sabe-se que o processo de leitura de um valor por corrente promove economia de energia e também maior velocidade de leitura. Da mesma forma que na escrita da célula de memória, mantem as *bitlines* com valores baixos de variação de tensão e por consequência menos potência consumida. Os *sense amplifiers* são circuitos componentes de todos os tipos de memória e por isso continuamente estudados e apresentando novas soluções. A figura 2.17 mostra um *sense amplifier* sem consumo estático, comparando-o com um sense amplifier construído com par diferencial, este projeto apresenta redução de consumo de energia entre 60 e 90%. Outro ponto que se pode ressaltar é a inclusão de um transistor de equalização que da mesma forma acelera o processo de leitura pois equaliza as *bitlines* anteriormente fazendo com que qualquer variação seja notada pelo *SA* mais rapidamente.

2.1.8 Circuitos auxiliares

Uma memória SRAM é constituída de diversos sub-blocos principais, pode-se acrescer a estes circuitos auxiliares que podem ser desenvolvidos paralelamente com objetivo de maior performance ou menor consumo. Alguns exemplos destes são os equalizadores de *bit-lines* ou *data-lines*, os *queenchers* (limitadores de variação de tensão nas *bitlines*), entre outros que serão apresentados a seguir.

2.1.8.1 Equalização de Bitlines

Os circuitos de equalização de *bitlines* podem tambem ser chamados de transistores de equilíbrio sendo parte integrante do circuito de pré-carga para manter valores idênticos em ambas as *bitlines* no inicio de uma leitura ou gravação acelerando a operação. Até aí nenhuma novidade, mas o interessante que até um simples transistor de equilíbrio pode ter sua funcionalidade alterada fazendo com que aconteça a equalização de forma mais rápida e assim aumentando o desempenho. Conhecido como *body equalization* a técnica consiste em aplicar uma tensão menor ao corpo do transistor aumentando sua condutividade, ou seja, variar seu *threshold* para obter uma equalização como se houvesse um curto entre as *bitlines*. Podendo também ser executada após o mux que liga algumas *bitlines* a um mesmo SA. Esta equalização traz maior desempenho também ao *sense amplifier*, pois suas entradas estarão em valores idênticos capacitando-o a perceber a menores variações na *bitline* mais rapidamente. A figura 2.18 mostra o acesso da pré-carga e os transistores de equilíbrio ligados entre as *bitlines,* e a figura 2.19 traz uma imagem do circuito de equalização com *body equalization*.

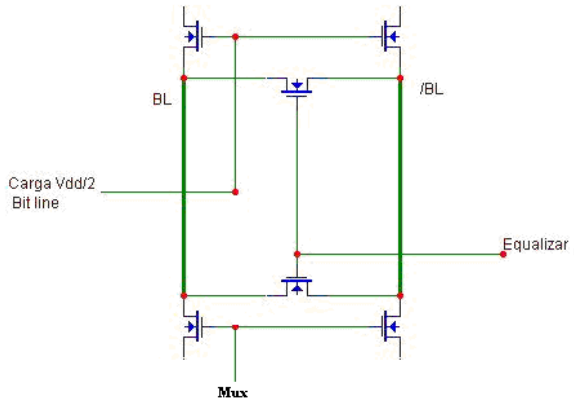


Figura 2.18 – Circuito de pré-carga com transistores de equalização

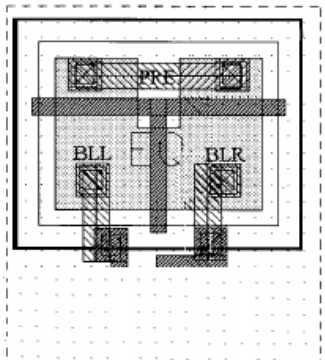


Figura 2.19 – Layout do circuito de pré-carga com transistores de equalização e possibilidade de *body equalization* (TSIATOUHAS, 2000)

2.1.8.2 Quenchers

Os quenchers são utilizados como limitadores de tensão nas *bitlines*, fazendo com que o valor limítrofe definido reduza a potência dinâmica consumida. Estes circuitos auxiliares são tipicamente diodos MOS ligados entre as *bitlines* e desta forma limitam a diferença de tensão a um Vth entre as mesmas. Este tipo de operação é tipicamente útil quando em conjunto com *sense amplifiers* de corrente pois assim garante-se a não elevação da tensão nas *bitlines* e a manutenção da economia de energia e redução do atraso. A figura 2.20 é um exemplo de localização dos mesmos nas *bitlines.* Na figura 2.21 demonstra-se sua implementação elétrica e na figura 2.22, uma comparação entre a corrente nas *bitlines* com e sem *quenchers*.

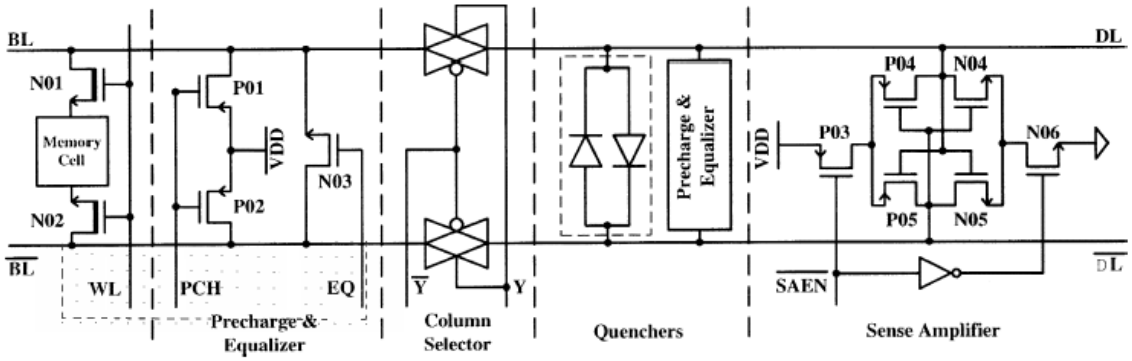


Figura 2.20 – Célula de memória com *quenchers* nas *bitlines* (WANG, 2003)

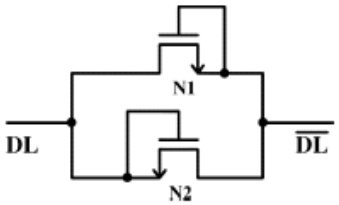


Figura 2.21 – *Quenchers* NMOS (WANG, 2003)

A contribuição básica deste dispositivo é a impossibilidade de oscilação da corrente nas *bitlines* conforme demonstrado na figura 2.22. Esta não oscilação leva a menores

atrasos para leitura, através da definição de valores na entrada dos SAs. De mesma forma diminui o consumo de corrente na célula de memória e a possibilidade de um *bit-flip,* que é a inversão do valor armazenado na mesma.

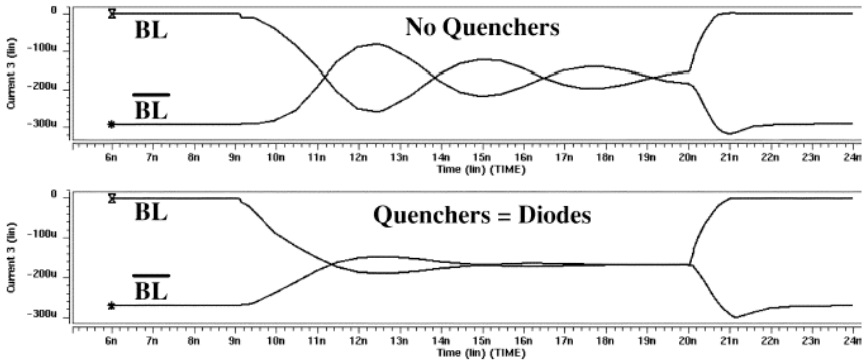


Figura 2.22 – Comparação de corrente em funcionamento com e sem *queenchers* (WANG, 2003)

3 ECONOMIA DE ENERGIA ESTÁTICA EM SRAM´S

O mundo moderno trouxe a necessidade de portabilidade para os equipamentos e com isto a duração da bateria de um equipamento e seu tamanho se tornaram tão determinantes quanto suas funcionalidades e utilização. As melhorias na capacidade de armazenamento de energia não acompanharam a evolução da tecnologia CMOS na mesma taxa de desenvolvimento. Assim para obter durabilidade de bateria, estas cada vez menores, novos métodos de projeto de baixa potência devem ser empregados. Os picos de potência consumidos devem ser controlados para que bateria tenha maior autonomia, o que gerará melhores características para o circuito e maior usabilidade do mesmo. Neste panorama, SRAM´s necessitaram um rápido desenvolvimento de técnicas para baixa potência devido ao aumento da corrente de fuga advinda da miniaturização dos dispositivos e do aumento de área utilizada por elas nos projetos modernos. O *leakage* de uma memória deve ser controlado para que se possa permitir a operação das baterias pelo maior tempo possível, não é mais aceitável um projeto com um consumo estático elevado enquanto o circuito se encontra em modo de espera (*stand-by*).

Existem formas de conter o consumo estático de uma memória via sua controladora ou via alterações diretas nos blocos da parte operativa, que geram composições para atender as especificações exigidas a cada nova geração de produtos. Ao tentar-se reduzir o consumo via estratégias na controladora trabalhar-se-á em um nível mais elevado, pensando principalmente em como e quando acionar as estratégias e em que parte dos bancos de memória serão ativados ou desativados na busca por obter ganhos em redução de *leakage* estático na memória acessada. No âmbito da parte operativa de memórias SRAM se agirá diretamente nos blocos componentes da memória, sendo esta uma forma de melhorar as relações consumo versus desempenho e também de promover a possibilidade de novas estratégias de redução de consumo à controladora de memória. As chaves para estas técnicas de redução de consumo estático baseiam-se em: - Incremento da largura de canal, variação de tensão entre operação e stand-by, redução da corrente de fuga utilizando *footed transistors* para Vdd ou Vss, redução de consumo estático pelo uso de múltiplos *thresholds* e/ou *thresholds* variáveis e otimização dos dispositivos utilizados na fabricação do CI.

3.1 Variação do consumo de potência entre tecnologias

Atualmente um dos principais pontos de ponderação em projeto de CIs é maximizar a performance dentro de um limite de potência pré-estipulado ou imposto fisicamente. Para explorar o espaço de projeto pode-se, ou deve-se, reduzir o Vth dos dispositivos até um limite onde a potência estática seja tolerável e finalmente elevar o Vcc até

atingir-se o limite da potência, o *scaling* de tecnologia. Esta tem sido a realidade presente no desenvolvimento de novas tecnologias em busca de cada vez mais performance.

Com a redução das tecnologias os efeitos não programados do *scaling* tornam-se cada vez mais notórios e por isto necessitam-se cada vez mais de circuitos auxiliares que procurem compensar os erros criados. Desta forma nota-se que no passar das tecnologias a distância das conexões tem aumentado, aumentando a capacitância parasita, e desta forma ao trocar-se uma tecnologia ou mantem-se ou eleva-se o consumo de potência. Segundo Gielen (2005), pode-se aproximar a variação de potência no *scaling* através da relação:

$$\frac{P1}{P2} = \frac{1}{m} \cdot \frac{tox_1}{tox_2} \qquad \text{(eq. 3.1)}$$

Sendo P1 e P2 as potências consumidas na tecnologia 1 e 2 respectivamente, tox_1 e tox_2 a espessura do óxido em ambas as tecnologias e m a relação entre as tensões de alimentação. Demonstra-se assim que não há um real beneficio em utilizar tecnologias nanométricas neste caso, além disso, aumentam a dificuldade de implementar circuitos analógicos por redução na tensão de alimentação pois esta redução inviabiliza técnicas como transistores de corte de alimentação, etc.

Quando acrescido estes fatores temos as necessidades de desvio de *clock* em sistemas complexos, tem-se a necessidade de incremento de área e redução de comprimento de conexões como demonstrado na figura 3.1.

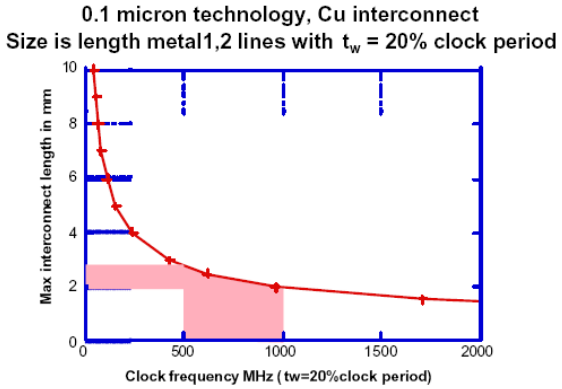


Figura 3.1 – Tamanho máximo da interconexão em Metal 1 ou Metal2 para um *clock skew* menor que 20% em função da frequência de *clock* (GIELEN, 2005)

O projeto analógico tem sofrido com as características dos dispositivos nas tecnologias UDSM pois criaram-se problemas de linearidade, de variação de ganho, casamento, redução de *headroom* que dificultam os projetos. Acrescido a isso temos a elevação do *leakage* que leva a dificuldades na operação dos circuitos. Na figura 3.2 temos um exemplo da variação da potência estática e dinâmica ao longo das tecnologias de fabricação usando como exemplo um processador. O incremento da potência estática torna-se cada vez mais um fator a ser relevado no projeto de um novo circuito, assim pode-se considerar a nova formulação da potência total consumida como:

$$P_{tot} = P_{dinâmica} + P_{estática} + P_{Curto\_Circuito} \qquad \text{(eq. 3.2)}$$

As componentes da potência estática serão apresentadas no próximo tópico da dissertação, esta por ser uma energia que está presente em todos os modos de operação de um circuito integrado influencia diretamente na capacidade de *stand-by* do mesmo por aumentar o consumo durante este modo de operação. Em circuitos modernos projetados em tecnologias de 65nm a potência consumida em corrente de fuga no modo de espera é da ordem de 1mW, sendo que a durabilidade da bateria para SOCs estado da arte é da ordem de 150 a 250 horas.

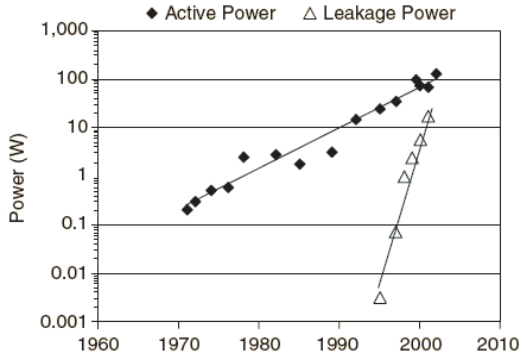


Figura 3.2 – Potências dinâmica e estática em um processador (MOORE, 2003)

3.2 Mecanismos de Consumo Estático:

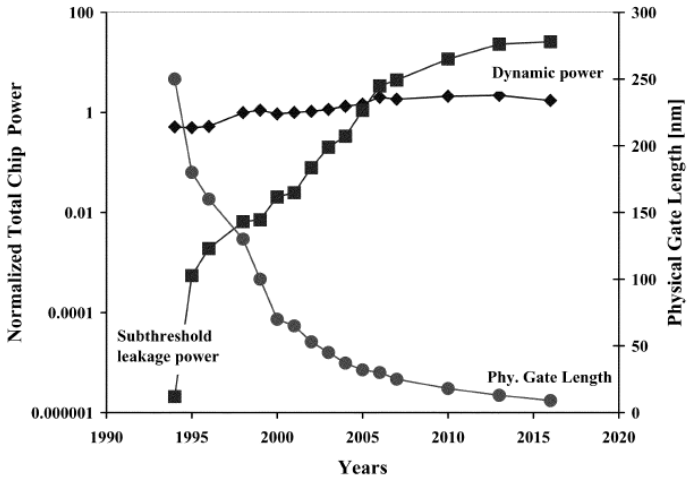


Figura 3.3 – Potências dinâmica e estática normalizadas para dispositivo de W/L=3 (KIM, 2004).

Para alcançar maior densidade de integração, melhor desempenho e alcançar menores custos de produção os dispositivos CMOS tem sofrido *scaling*, segundo a lei de Moore. A cada nova geração tecnológica os fatores negativos do *scaling* tem sido cada vez mais sentidos, o incremento do consumo estático, ganho menos linear, grande variação de Vth, etc. Antes ignorado como um fenômeno irrelevante, o consumo estático hoje torna-se um dos focos mais importantes de pesquisa e desenvolvimento na busca de sua redução ou minimização. A figura 3.3 apresenta a normalização das potências dinâmica e estática baseado no ITRS 2001, demonstrando o incremento do consumo estático ao longo dos anos e a planaridade do consumo dinâmico com pequeno

incremento. Os circuitos modernos utilizam *gates* de poucas dezenas de nanômetros, sofrendo um grande numero de efeitos físicos dantes desprezíveis, gerando uma nova preocupação no modelamento e estudo dos mecanismos físicos vigentes nestes dispositivos.

Em dispositivos fabricados em tecnologias UDSM o *leakage* é composto dominantemente pela corrente de *subthreshold*, *leakage* de *gate* e *leakage* de tunelamento na junção. Podem ser visualizadas as componentes do *leakage* na figura 3.4. Nesta seção serão apresentados os mecanismos de maior relevância na composição do *leakage*.

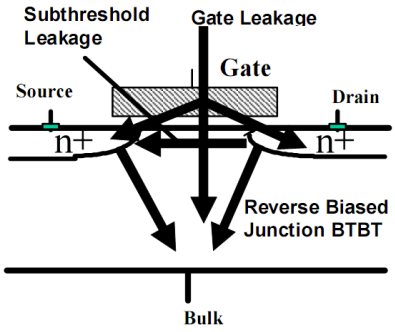


Figure 3.4 – Mecanismos de *leakage* em transistor MOS (CHEN, 2007).

3.2.1 Corrente Sub-Limiar

A corrente de *subthresholds* depende exponencialmente da tensão Vgs, do Vth do transistor e da temperatura. Devido a efeitos de canal curto há um acréscimo de corrente de *subthresholds* juntamente com a polarização de dreno (efeito de DIBL – *Drain Induced Barrier Lowering*) e com a redução da largura do canal (Vth–roll off). Através do efeito de corpo há uma redução da corrente de *subthresholds* com a aplicação de uma polarização reversa no corpo. Assim pode-se modelar a corrente de *subthresholds*, seguindo Roy (2003), como:

$$I_{sub} = I_{sub0} \cdot e^{\left(\frac{V_{gs} - \eta_{DIBL}(VDD - V_{BS}) + \lambda_{body} V_{BS}}{\frac{m \cdot kT}{q}} \right)} \qquad \text{(eq. 3.3)}$$

Onde, I_{sub0} é a corrente de sub-limiar quando o transistor está polarizado com Vgs=0V e Vds=Vdd, η_{DIBL} é o coeficiente de DIBL, λ_{body} é o coeficiente de efeito de corpo, m é o fator de declividade da característica log(I_D) versus Vg na região de sub-limiar do *subthresholds* e o potencial térmico VT= kT/q.

3.2.2 Corrente de Tunelamento Direto pelo Gate

O *leakage* de *gate* em transistores de óxido ultra-fino é devido ao tunelamento direto de elétrons (ou lacunas) através do dielétrico de *gate*. O tunelamento dominante é denominado de Fowler-Nordheim e depende do campo elétrico no óxido e da espessura do óxido. O *leakage* de gate tem redução exponencial com a espessura do óxido e incrementa exponencialmente com o aumento do campo elétrico através do óxido (i.e. Vgs). Os componentes predominantes no MOSFET são a corrente de *gate* para as regiões de sobreposição fonte/dreno, corrente do *gate* para o canal e corrente do *gate* para o substrato (ROY, 2003). A corrente de tunelamento na sobreposição fonte/dreno

domina o *leakage* quando o transistor está cortado e na condução a corrente de *gate* para o canal é a componente dominante. Assim modela-se o *leakage* de *gate* como (CHEN, 2007):

$$I_{g\_OFF} = I_{g\_OFF0} \cdot e^{\left(-\alpha_{g\_OFF} \cdot (VDD - |V_{GD}|)\right)} \qquad \text{(eq. 3.4)}$$

$$I_{g\_ON} = I_{g\_ON0} \cdot \left[e^{\left(-\alpha_{g\_ON} \cdot (VDD - |V_{GD}|)\right)} + e^{\left(-\alpha_{g\_ON} \cdot (VDD - |V_{GS}|)\right)} \right] \qquad \text{(eq. 3.5)}$$

Onde, I_{g_OFF0} é a corrente de tunelamento na sobreposição S/D quando o transistor está cortado e polarizado com |Vgd|=Vdd e I_{g_ON0} é a corrente de tunelamento de *gate* para o canal quando o transistor está conduzindo e polarizado com |Vgs|=Vdd. α_{g_ON} e α_{g_OFF} são coeficientes de *leakage* de *gate* para condução e corte.

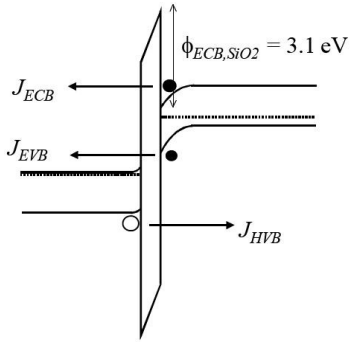


Figure 3.5 – Os três mecanismos de fuga através do dielétrico do *gate* (CAO, 2000).

A figura 3.5 apresenta três componentes do *leakage* de *gate*, sendo a primeira componente o tunelamento de elétron na banda de condução (ECB - *Electron Conduction-Band tunneling*) devido ao tunelamento de elétrons do *gate* na banda de condução para o substrato ou vice-versa. A segunda componente de tunelamento de elétron na banda de valência (EVB - *Electron Valence-Band tunneling*) devido ao tunelamento de elétrons da camada de valência do substrato para a camada de condução do *gate*. A terceira componente é o tunelamento de lacuna da camada de valência (HVB - *Hole Valence-Band tunneling*) devido ao tunelamento de lacunas do *gate* na camada de valência para a camada de valência do substrato, ou vice-versa.

Segundo Roy (2003) pode-se equacionar o *leakage* de Fowler-Nordheim que contribui para a corrente DC de *gate* como:

$$I_g = W.L.A\left(\frac{V_{ox}}{t_{ox}}\right)^2 \exp\left(\frac{-B\left(1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right)}{\frac{V_{ox}}{t_{ox}}}\right) \qquad \text{(eq. 3.6)}$$

Onde W e L são o comprimento e largura efetivos do transistor, respectivamente, $A = q^3 / 16\pi^2 h\phi_{ox}$, $B = 4\pi\sqrt{2m_{ox}}\,\phi_{ox}^{3/2} / 3hq$, m_{ox} é a massa efetiva da particular em tunelamento, ϕ_{ox} é o valor da diferença de potencial sobre o óxido, t_{ox} é a espessura do óxido, h é $1/2\pi$ vezes a constante de Planck e q é a carga do elétron.

3.2.3 Corrente de tunelamento na Junção

Este *leakage* é devido ao tunelamento entre bandas dos elétrons em junções p-n altamente dopadas e polarizadas reversamente. Em UDSM MOSFETS devido ao uso de junções altamente dopadas há a criação de grandes junções BTBT (*band-to-band tunneling*) entre dreno conectado no VDD e substrato conectado ao terminal de terra. O efeito da junção BTBT aumenta exponencialmente com o valor da polarização de dreno para o substrato. Segundo Chen (2007) modela-se a corrente de fuga por tunelamento na junção:

$$I_{JN} = I_{JN0} \cdot e^{(-\beta_{JN} \cdot (VDD - |V_{DB}|))} \qquad \text{(eq. 3.7)}$$

Onde, I_{jn0} é o *leakage* de junção quando o transistor está polarizado com Vdb = Vdd e β_{JN} é o fator de ajuste empírico da dopagem. Estes fatores I_{jn0} e β_{JN} podem ser extraídos de simulações.

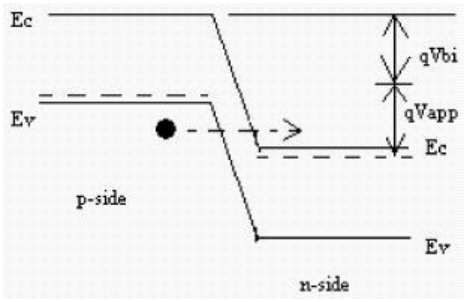


Figure 3.6 - BTBT em junção pn reversamente polarizada (ROY, 2003).

Como demonstrado na figura 3.6 um campo elétrico forte sobre uma junção pn reversamente polarizada causa um fluxo de corrente através da junção por tunelamento de elétrons da banda de valência da região P para a banda de condução da região N.

Quando se tem aplicada sobre a junção uma tensão reversa maior que a tensão de *band-gap*, a densidade de corrente de tunelamento pode ser modelada segundo Roy (2003) como:

$$J_{BTBT} = A \frac{E \cdot V_{app}}{(E_g)^{1/2}} \exp\left(- B \frac{(E_g)^{3/2}}{E} \right) \qquad \text{(eq. 3.8)}$$

Onde, $A = \sqrt{2m^* q^3} / 4\pi^3 h^2$, e $B = 4\sqrt{2m^*} / 3hq$ (m^* é a massa efetiva do elétron); Eg é o gap entre as bandas de energia, V_{app} é a polarização reversa aplicada; E é o campo elétrico na junção; q é a carga do elétron e h é $1/2\pi$ vezes a constante de Planck.

3.2.4 Implicações da temperatura

O acréscimo da temperatura faz com que a corrente de fuga seja incrementada, criando novas preocupações na hora de se projetar um circuito integrado. Se houver uma elevação de temperatura no *die* para, por exemplo, 110ºC haverá um acréscimo nas componentes do *leakage* quando comparadas à temperatura base de 30ºC. Com esta elevação de temperatura de trabalho a corrente de *subthresholds* fica incrementada em cinco vezes, o *leakage* de *gate* em uma vez e meia.

Tendo-se em conta estes fatores a distribuição correta de sub-blocos de alta taxa de chaveamento que provocam aquecimento deve ser planejada de forma a se trabalhar o circuito para haver uma menor elevação de temperatura do *die* e consequentemente menor consumo estático.

3.3 Controle de Consumo de Potência Estática em SRAM´s via Controladora de Memória:

A controladora é a entidade que controla a parte operativa de uma memória, sendo a parte da memória que gerencia os aspectos de atividade e inatividade da mesma colocando-a em modos de operação adequados. Dentre outros fatores controlados pela mesma temos o tempo de validade de uma linha na memória, bem como tempo de operação em modo de *stand-by* ou tempo de espera para ativação deste mesmo modo. Outro aspecto será o tempo para um *reset* de todas as posições da memória ou qual o tempo para que a memória adormeça. O conjunto de soluções aplicadas a parte operativa da cache fornece as ferramentas para que a controladora possa habilitar os modos de baixo consumo assim como ativar os caminhos dos dados justamente no momento em que serão utilizados. A somatória destas estratégias, da divisão de blocos e de ativação aplicadas aos sub-blocos de uma memória cria o ambiente para o projetista da controladora de memória poder desenvolver estratégias em alto nível para economia de energia.

O conjunto de técnicas ou mesmo a aplicação de uma somente pode levar as boas reduções da energia consumida por uma memória SRAM. Estes métodos de economia de energia se fazem mais comuns para controladoras de memórias cache, estas comuns em diversos circuitos dedicados, assim sendo um bom teste para medir as melhoras ocorridas. Aproveitando dados coletados da literatura em Flautner (2002), agarwal (2002) e Li (2004), pode-se comparar e demonstrar as vantagens e desvantagens da aplicação destas técnicas e os ganhos inerentes das mesmas. Apresentam-se brevemente os tipos de controladoras de cache e comparações entre as mesmas, obtendo assim uma boa análise comparativa dos principais gêneros de controladoras e dos ganhos inerentes a cada uns dos métodos de controle aplicados a caches comerciais.

3.3.1 Panorama na comparação dos principais gêneros de controladoras:

Os principais gêneros de controladoras para redução de consumo estático estão divididas nas que preservam e nas que não preservam o estado do dado na memória, respectivamente *drowsy* e *decay* caches conceitos apresentados a seguir no texto. Estas controladoras aplicam diferentes estratégias para controlar o *leakage*, na cache tipo *drowsy* há um chaveamento entre duas alimentações sendo uma a de operação normal e outra de modo de *stand-by*. As vantagens deste tipo de controladora é que a mesma mantem o dado armazenado na memória durante o período de espera e apresentam uma troca rápida entre as alimentações de retenção de dado e a de acesso para possibilitar leitura ou escrita. Suas principais fragilidades são a baixa margem sinal ruído (SNM – *Signal to Noise Margim*) quando em modo de espera e o valor da tensão de retenção tem dependência das variações de processo para sua escolha dentro de margens seguras de operação.

O segundo tipo principal de controladoras é a *decay*, esta controladora através de um transistor de corte procura eliminar o caminho do *leakage* para o terra do sistema, sendo uma tática bem eficaz. Também apresenta variação rápida entre o modo de *stand-by* e o de operação normal, sendo o método de maior economia de energia. Sua fragilidade é a

inserção de muitos sinais de perda (*misses*) de dados caso seu tempo de *decay* não seja adequado ou ajustado à aplicação rodada. Pode por aumento de dados não válidos, indução de sinais de perda de dados (*misses*), não ser a escolha mais adequada.

3.3.1.1 Decay Cache:

A figura 3.7 demonstra uma célula de uma matriz de memória do tipo *decay*, sua principal característica é o transistor de corte entre o terra e o terra virtual. As *decay* caches devem ter tempos de *decay* para as células de memória adequados a aplicação desejada, se aplicação alvo for, por exemplo, uma cache de instruções o tempo de validade do dado é totalmente diferente da de uma cache de dados. Se os tempos de *decay* não forem adequados poderá haver maior consumo de energia que economia, pois a perda de um dado pode gera uma busca no nível acima da memória, ocasionando assim maior consumo de energia que o economizado na operação de stand-by.

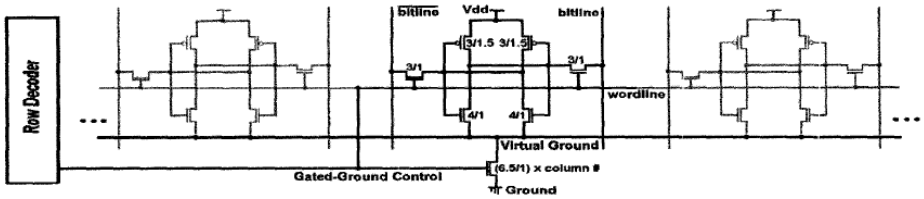


Figure 3.7 – Implementação de uma linha de uma *decay cache* (AGARWAL, 2002)

3.3.1.2 Drowsy Cache:

A principal característica de uma cache do tipo *drowsy* é o uso de duas tensões de alimentação, uma normal de operação e outra mais baixa ativada enquanto a célula não é acessada. A figura 3.8 demonstra o controle de tensões de uma célula de memória do tipo *drowsy*, caracterizada pelos transistores de chaveamento entre as duas alimentações. Desta forma, esta memória ao mesmo tempo em que mantem os dados nela armazenados, também proporciona economia de energia enquanto não é acessada.

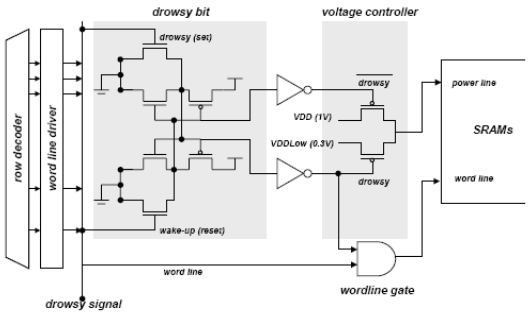


Figure 3.8 – Implementação de uma linha de uma *drowsy* cache (FLAUTNER, 2002)

A tensão de *stand-by* de *drowsy* caches deve ser adequadamente calculada, para operar com certa folga sobre o pior caso de variação de processo nas células de memória. O tempo de *stand-by* em tensão reduzida possibilita a redução do consumo de energia, pois os mecanismos de corrente de fuga tem correlação exponencial com o valor da tensão aplicada aos dispositivos. A figura 3.9 apresenta uma simulação de acesso a uma célula para escrita, com a alimentação em modo normal, estando uma

vizinha em modo *drowsy*, demonstrando pelo teste que não ocorrerá troca do valor do armazenado pela ação da capacitância C_{DS}.

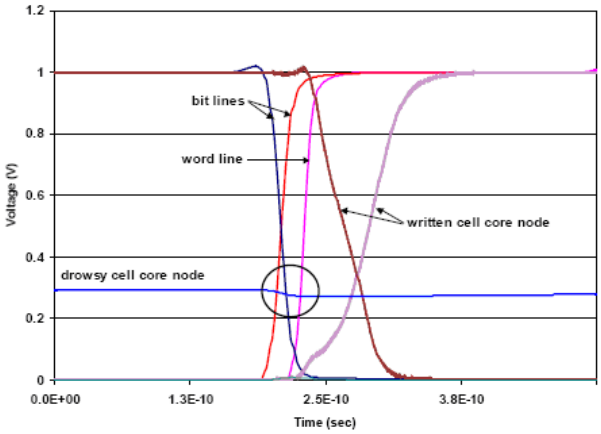


Figure 3.9 – Escrita em célula demonstrando a imunidade das células vizinhas (FLAUTNER, 2002)

3.3.1.3 Cache modo Híbrido:

Um terceiro tipo de cache foi idealizado unindo as propriedades mais interessantes de ambos os modos *drowsy* e *decay*. A principal ideia é manter durante um tempo ajustável a cache em modo *drowsy*, o dado continua ali armazenado se não houver acesso durante este tempo de espera, só então a célula de memória entra em modo decay. Se a célula é colocada em modo *decay* presume-se que o dado ali armazenado já perdeu a sua validade tornando assim a economia de energia mais efetiva. Kaxiras (2005) demonstra que a utilização de tempos adequados ou a variação destes tempos através da utilização de um sensor de temperatura que varia os tempos de *decay* e *drowsy* de acordo com a temperatura de operação pode alcançar melhor desempenho em economia de energia que o modo *drowsy* somente. A figura 3.10 apresenta a comparação entre o modo *drowsy* e os modos híbridos implementados no citado *paper*.

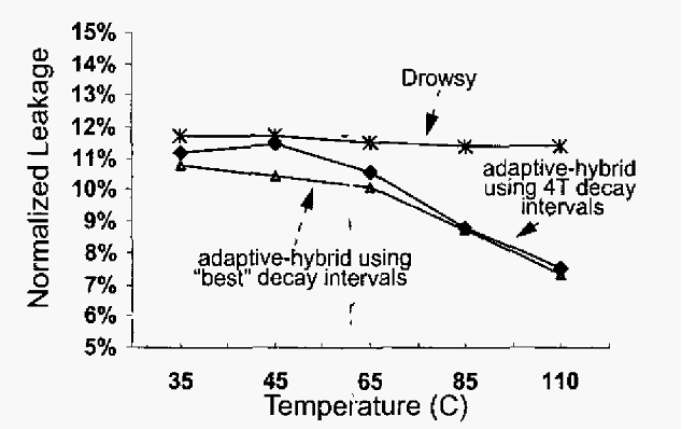


Figure 3.10 – Comparação de redução de *leakage* entre controladora hibrida e *drowsy* versus a temperatura (KAXIRAS, 2005)

3.4 Técnicas para Redução de consumo de Potência Estática

Objetivando a redução da corrente de fuga, especialmente para SRAMs, foco deste trabalho também sendo este um bloco que tem sido muito utilizado e cada vez em maior número e área a cada nova tecnologia de fabricação. A partir deste panorama a busca por soluções que tratem esta realidade de forma a criar circuitos com técnicas de redução deste consumo indesejado e que possam atingir os diferentes blocos empregados na construção de um (a) CI / SRAM.

Sendo que a corrente de fuga é um fator de extrema relevância atualmente, pois com o *scaling* das tecnologias tem se tornado uma fonte de consumo relevante, maior que o dinâmico muitas vezes, levaram a busca por técnicas que minimizem o consumo estático de um circuito moderno. Existem formas de fazê-lo trabalhando com o Vth dos transistores, já exposto anteriormente nesta dissertação, relembrando-as: - Vth duplo, Vth variável, método DLC. Temos também os métodos VT CMOS, MT CMOS e ABC MT CMOS, sendo que todos estes referem-se a variações dinâmicas de Vth. Quando juntamos a isso as soluções empregadas pelo tipo de controladora, apresentados em tópico prévio da dissertação, o consumo em modo de espera reduz-se pois a maioria dos circuitos auxiliares estão desligados. Numa SRAM não acessada o consumo resume-se a uma corrente de fuga multiplicada pelo tamanho da matriz de memória.

Como as SRAMs estão maiores e mais rápidas a cada nova geração de processadores, exige-se um trabalho constante de desenvolvimento de novas técnicas que solucionem os fatores adversos, sendo estas algumas soluções eficazes na redução de potência estática :

- Incremento da largura de canal;
- Variação de tensão entre operação e stand-by;
- Redução da corrente de fuga utilizando transistores de corte para Vdd ou Vss;
- Redução de consumo estático pelo uso de múltiplos *thresholds* e/ou *thresholds* variáveis;
- Otimização dos dispositivos utilizados na fabricação do CI.

Seguem-se maiores explanações das soluções e empregabilidade das mesmas, transcorrendo sobre os assuntos abordando mais especificamente cada um dos subitens apresentados na lista acima.

3.4.1 Redução da corrente de fuga por utilização de múltiplos *thresholds*

Sabe-se que o *scaling* de tecnologias por redução do óxido de porta leva a redução da tensão de alimentação e consequente redução do *threshold* para que a operação seja consistente dentro deste novo ambiente. Com isto a corrente de sub-limiar aumenta justamente pelas características deste canal cada vez mais curto, assim para obter melhor relação entre a corrente no modo ativo e modo de espera há um método fundamental conhecido como técnica de circuitos com múltiplos Vths ou com Vths variáveis. Um exemplo desta técnica é mostrado na figura 3.11, os transistores de alto Vth têm baixa corrente de fuga e podem ser utilizados para reter tensões durante o modo de espera, enquanto que transistores de baixo Vth têm melhor desempenho dinâmico e consequente maior corrente de fuga.

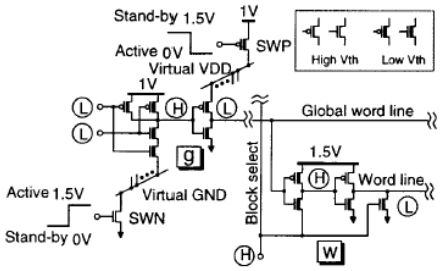
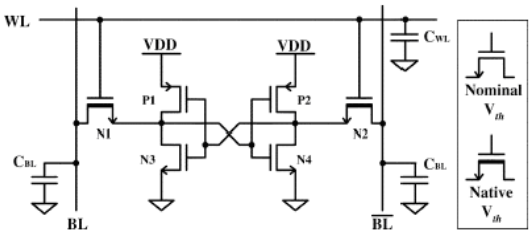


Figura 3.11 – Circuito projetado com Vth duplo (MARGALA, 1999)

Nesta abordagem temos soluções similares com a mesma ideia básica que é obter transistores com menor corrente de fuga para a retenção de dados e transistores com maior desempenho para a passagem de corrente. Esta característica é obtida através do método de Vth variável onde apresentam-se transistores de Vth Nominal com *threshold* de maior tensão e Vth Nativo de menor tensão, variando a dopagem dos transistores. A ideia básica é utilizar os transistores nominais na retenção e os nativos como transistores de passagem, multiplexadores, etc. Assim onde empregam-se os transistores adequados à retenção, ao corte de corrente, ou onde é necessária maior condutividade; trazendo ao circuito economia de energia dinâmica e maior desempenho. A figura 3.12 mostra um exemplo de uma célula de memória fabricada com a aplicação desta técnica.



THRESHOLD VOLTAGES OF NOMINAL nMOS/pMOS AND NATIVE nMOS TRANSISTORS (NATIVE pMOS IS NOT AVAILABLE IN THE 0.25-μm CMOS PROCESS)

	Nominal V_{th}	Native V_{th}
PMOS	V_{thNoP} = -0.53 V	V_{thNaP} = N/A
NMOS	V_{thNoN} = 0.53 V	V_{thNaN} = 0.21 V

COMPARISON BETWEEN HIGH AND LOW THRESHOLD VOLTAGE TRANSISTORS

V_{th}	Characteristic	Advantage
Nominal (0.53 V)	low leakage current	data retention
Native (0.21 V)	high output current, fast switching time	driving capability

Figura 3.12 – Célula de memória com Vth Duplo (WANG, 2003)

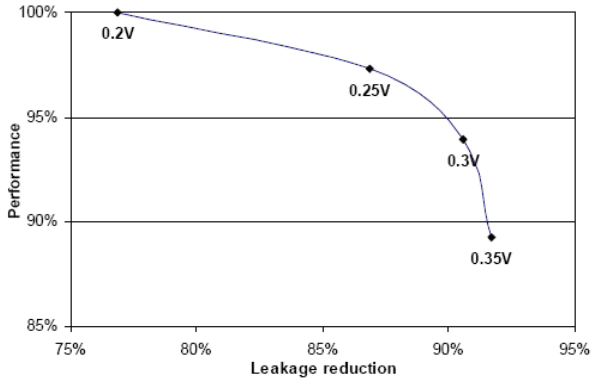


Figura 3.13 – Variação de tensão de limiar versus *leakage* e desempenho (FLAUTNER, 2002)

Como toda solução, há um ônus relativo à utilização dos transistores nativos, pois apresentarão sempre maior consumo estático devido seu baixo *threshold* e assim sendo

uma solução um tanto quanto mais focada no aumento de desempenho que na redução da corrente de fuga no *latch* da memória. A figura 3.13 apresenta a variação entre aumento de *leakage*, o valor de Vth e o desempenho do transistor mostrando justamente a característica destas variáveis como plano de fundo da escolha dos tipos de transistores adequados as diferentes funções no projeto de um CI.

Como qualquer nova solução esta abordagem tem seus defeitos, pela redução do *threshold* dos transistores de passagem há uma elevação da corrente de fuga através das bit-lines. Criam-se novas complicações para a compensação destas componentes nos sense amplifiers e na equalização das *bitlines*. Outra forma de emular as mesmas variações de Vth é a utilização do *back bias*, ou seja, a variação da tensão de *bulk* criando uma variação de Vth dinâmica no transisitor como demonstrado na figura 3.14.

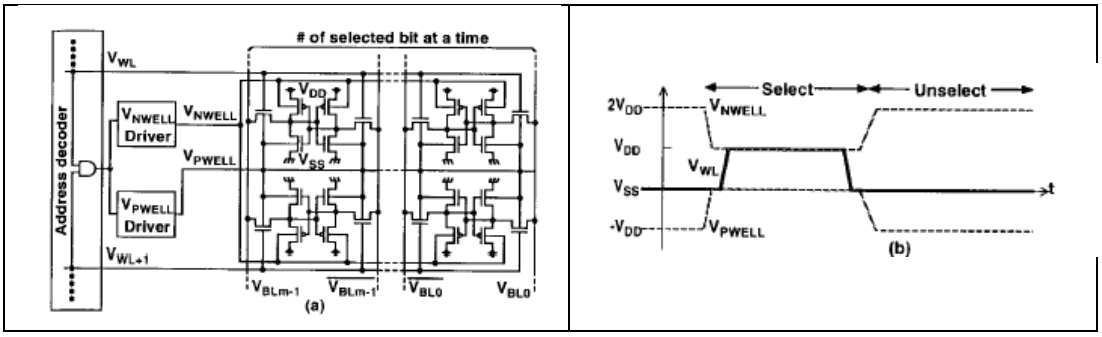


Figura 3.14 – Funcionamento da técnica DLC com a)Esquemático do circuito e b)Operação das tensões de poço (MARGALA, 1999)

Kawagushi et. al. (MARGALA, 1999) introduziu uma nova técnica denominada DLC (D*ynamic Leakage Cut-off*) as formas de onda estão demonstradas na figura 3.14. O funcionamento desta arquitetura é a alteração da tensão dos poços N e P, para a célula selecionada ou em operação os poços vão respectivamente para Vdd e Vss. As células em repouso recebem aproximadamente 2*Vdd para o poço N e aproximadamente –Vdd para o poço P e assim operam em Vth mais alto reduzindo a corrente de *subthreshold*. Esta técnica é similar a VT CMOS (V*ariable Threshold CMOS*) a diferença é o sinal de sincronização da polarização dos poços. Na técnica de VT CMOS a polarização do poço é sincronizada pelo sinal de *stand-by*, na técnica de DLC a sincronização é feita pelo sinal de *word-line*.

O grande revés deste modo de operação é a necessidade de poço duplo ou têm-se somente a aplicação em transistores PMOS o que reduz a sua efetividade a uma parcela dos transistores implementados. Um segundo fator é a instabilidade gerada pela variação das tensões de poço, ao mesmo tempo esta solução se torna menos efetiva com a miniaturização dos transistores, isto segundo Prince (2007).

Nii et. al. (MARGALA, 1999) melhorou a técnica MT-CMOS e ainda propôs o método de *Auto-Backgate Controled* (ABC) MT-CMOS. A vantagem do ABC MT-CMOS é a redução significativa da corrente de fuga durante o tempo de espera, a figura 3.15 mostra um exemplo desta técnica. Os transistores de alto Vth Q1 a Q4 agem como chaves que cortam a corrente de fuga. O circuito mais interno é constituído de transistores de baixo Vth. No modo ativo SL é setado baixo e SLB, SL barrado, é elevado a nível alto. Assim Q1, Q2 e Q3 estão ativos e Q4 e cortado, assim a fonte virtual VVDD e a polarização do substrato se tornam 1V. Durante a operação em modo de espera SL é setado alto e SLB setado em nível baixo e Q1 a Q3 são cortados enquanto Q4 é ativado e assim BP tem 3.3V. A corrente de fuga circula de Vdd/2 ao

GND por D1 e D2 determinando Vd1, Vd2 e Vm. Vd1 é a polarização entre a fonte e o substrato dos transistores PMOS, Vd2 é a polarização dos transistores NMOS e VM é a tensão da fonte virtual entre VVDD e VGND, sendo que segundo Nii este método reduz o consumo para 20pA por célula de memória.

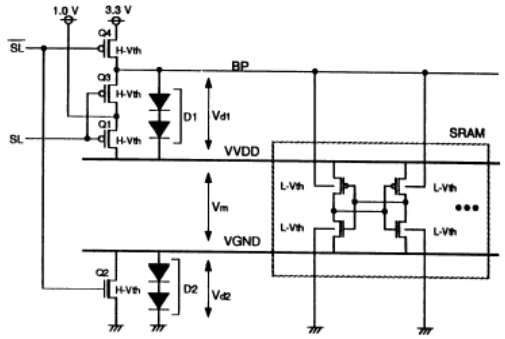


Figura 3.15 – Diagrama esquemático de um circuito ABC MT- CMOS (MARGALA, 1999)

Na figura 3.16 tem-se a variação das diferentes componentes da corrente de fuga versus a tensão de polarização do substrato de um transistor NMOS. Demonstra-se que o *leakage* de *gate* é praticamente estável, com variação nos *leakages* de *subthresholds* e de junção. Como se visualiza os ganhos em *leakage* de *subthresholds* pela redução da tensão de substrato não se traduzem em ganho real, pois há a elevação da corrente de fuga pela junção que equilibra ou torna o *lekage* total maior que o na tensão normalmente aplicada ao bulk do transistor, comprovando visualmente Prince (2007).

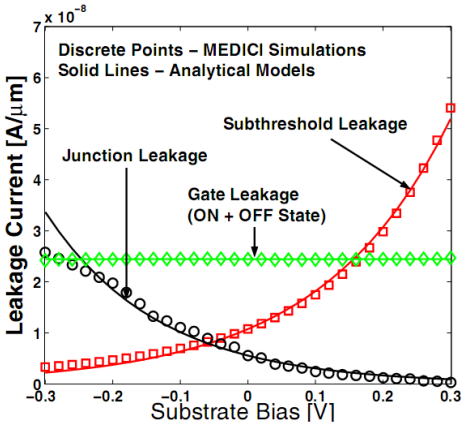


Figure 3.16 – Tensão de polarização do substrato versus componentes das correntes de *leakage* em transistor MOS (CHEN, 2007).

3.4.2 Redução da corrente de fuga por incremento de comprimento de canal

O *scaling* das tecnologias tem nos levado a transistores com dimensões cada vez menores chegando hoje a ser mensurada em numero de átomos de silício o que torna as imperfeições de canal curto cada vez mais pronunciados e relevantes. Na figura 3.17 apresenta-se o corte de um transistor com o caminho de fuga da corrente. A redução do canal do transistor faz com que o mecanismo de corrente de fuga de sub-limiar se torna mais pronunciado e por isso o acréscimo do comprimento de canal provoca o aumento da barreira de potencial entre os terminais de dreno e fonte, reduzindo assim o tunelamento. Segundo Chattopadhyay (2007) o incremento de uma vez e meia a duas

vezes o tamanho mínimo da tecnologia no comprimento do canal da lógica provoca uma queda de três a oito vezes na corrente de fuga, claro isto dependente da temperatura, da tensão e do processo.

Esta é uma solução aplicável a circuitos de baixo fator de atividade e que podem trabalhar com maior capacitância agregada. A redução de tensão de operação em modo de espera leva a redução também no *leakage* do *gate* que se incrementará por estar diretamente relacionado à área do *gate*. Esta técnica de aumento de canal é muito útil na fabricação de SRAMs que em tecnologia 65nm e nós abaixo já não devem utilizar transistores mínimos na estrutura da célula de memória.

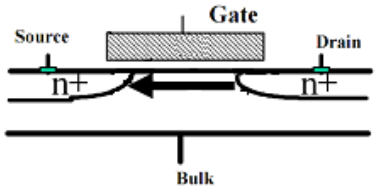


Figure 3.17 – Corrente de fuga Sub-limiar em transistor MOS

O aumento do comprimento do canal dos transistores em caminhos não críticos de lógica reduzindo a corrente de fuga é uma técnica conhecida como *Power Gating*. Baseia-se em análise de *timing* nos caminhos críticos reduzindo assim controladamente o desempenho do bloco funcional. Nesta técnica também trabalha-se com a remoção de transistores de baixo Vth para redução do *leakage*.

3.4.3 Redução da corrente de fuga por utilização de estruturas empilhadas

Estruturas empilhadas são uma forma de obter economia de energia através do corte do caminho de fuga em direção a Vdd ou a Vss. *Stacked structures* como definido em Butzen (2007) são uma solução viável para redução de *leakage* em circuitos digitais e analógicos. A figura 3.18 traz um exemplo de controle de blocos utilizando transistor de corte para o Vss, este controle se torna mais efetivo se utilizarmos transistores de alto VT que proporcionam um corte mais efetivo da corrente de fuga. O particionamento em blocos de um circuito proporciona que a ativação do bloco a ser utilizado só aconteça no momento que este se faz necessário reduzindo o consumo tanto estático como dinâmico do mesmo.

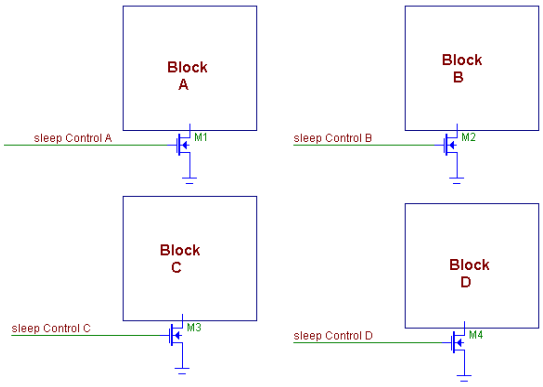


Figure 3.18 – Transistores de corte de Vss em diversos sub-blocos de um sistema

Em memórias pode-se aplicar o transistor de corte tanto no Vcc quanto no Vss, sendo que segundo demonstrado na figura 3.19, o transistor de corte de Vss se mostra

mais efetivo para eliminação de *leakage* em tecnologias nanométricas pois a pré-carga se dá em Vdd e desta forma quebram-se todos os caminhos de corrente de fuga para o terra.

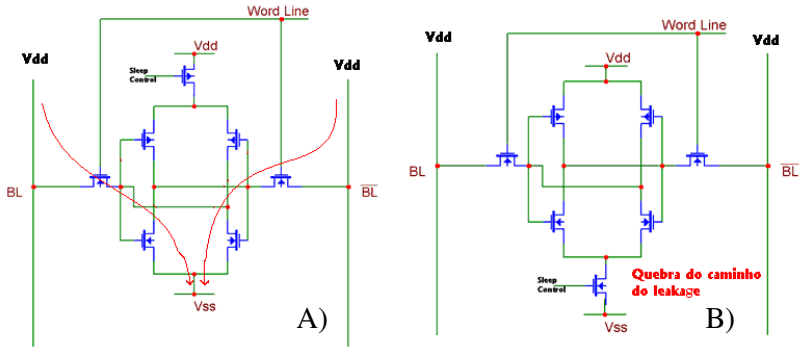


Figure 3.19 – Células de memória com (A)Transistores de corte de Vdd e (B) Transistores de corte de Vss

Neste circuito, durante o modo de espera, os blocos são retirados da alimentação através de transistores de corte. Utilizando-se transistores de baixo Vth para o acesso a célula de memória, estes com alta capacidade de corrente e menor queda de tensão têm como contrapartida possuir maior corrente de *subthreshold*. Assim como em tecnologias nanométricas as pré-cargas são feitas para Vdd, pois leva a maior desempenho, com o corte de Vss temos maior redução da corrente de fuga.

Finalmente este é o sistema utilizado para economia de energia em memórias do tipo *decay* que apresentamos anteriormente que com o estudo cuidadoso dos tempos de corte das alimentações proporciona economias de energia da ordem de 70%.

3.4.4 Redução da corrente de fuga por variação da tensão de operação e *stand-by*

As componentes da corrente de fuga são dependentes da tensão de operação e tem relação de queda exponencial com a redução mesma. Assim a criação de sub-blocos que operem em tensões reduzidas provocará o mesmo efeito dos transistores de corte, com a vantagem de o tempo de retorno do modo de espera ser menor e de o circuito manter o ultimo valor. O particionamento em sub-blocos de um circuito proporciona que a ativação do sub-bloco a ser utilizado só aconteça no momento em que este se faz necessário. Haverá redução do consumo tanto estático como dinâmico e se o fator de atividade deste circuito for baixo pode-se utilizá-lo com tensão de alimentação mais baixa, desta forma obter-se um circuito com consequente menor desempenho e frequência, e como resultado disto, obter também menor *leakage*.

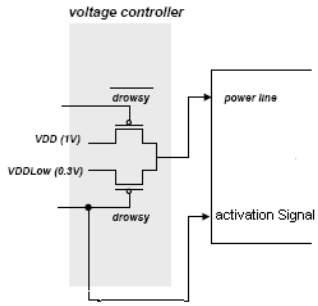


Figure 3.20 – Controle de tensões de modo normal e modo de *stand-by*

A figura 3.20 apresenta o sistema de controle de tensão, onde se visualiza as chaves para a tensão normal de operação e a tensão de modo de espera. Este é o sistema utilizado para economia de energia em memórias do tipo *drowsy*, e com o estudo cuidadoso da tensão de modo de espera proporciona economias de energia da ordem de 70% e com acréscimo de circuitos e compensações da ordem de 98%.

3.4.4.1 DRV (Data Retention Voltage)

Em células de memória através da redução da alimentação há redução do SNM, como demonstrado na figura 3.21 A). A redução da alimentação abaixo de um valor denominado DRV provoca a perda do dado armazenado. A estabilidade é indicada pela SNM, margem estática de ruído. Com a redução da tensão há redução do SNM até um ponto onde não há mais decisão entre os valores lógicos, o SNM é representado graficamente com o maior quadrado entre as curvas do VTC (*Voltage transfer Characteristic* - Característica de transferência de tensão). Idealmente na tensão de DRV a SNM é zero, o que não pode acontecer em uma experiência prática levando em consideração a variação de processo.

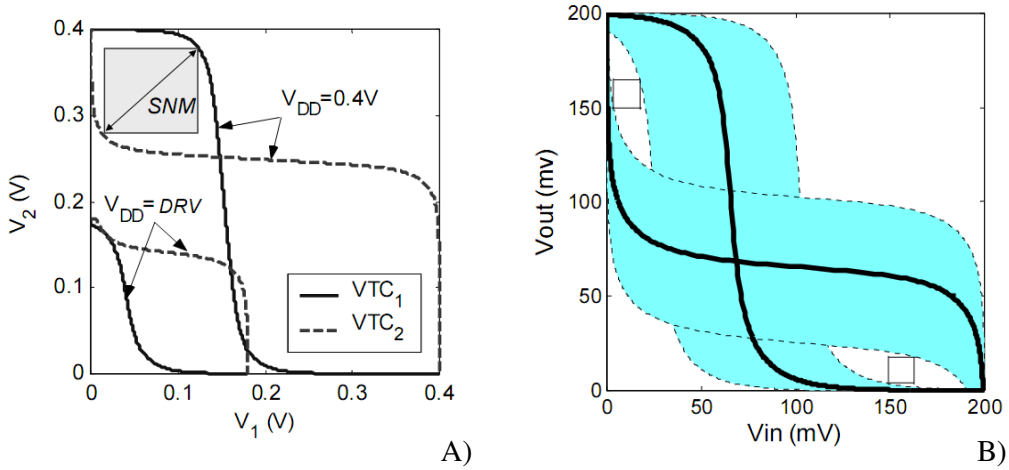


Figure 3.21 – (A) Degradação do SNM com a redução da tensão de alimentação e (B) Variação de SNM por variação de 3σ no comprimento de canal e no V_T (QIN, 2007)

Segundo Qin (2007) a tensão de DRV ideal pode ser definida como:

$$DRV_{ideal} = 2 \cdot V_T \cdot \ln(1+n) \qquad \text{(eq. 3.9)}$$

Onde V_T é o potencial térmico ($V_T = kT/q$), e n é o parâmetro que modela o fator de *subthreshold*. Em uma tecnologia CMOS ideal n=1 (i.e., 60mV/dec, AC como variação de tensão) chegaria a um resultado de 36mV para DRV. Para uma tecnologia típica 90nm com n=1.5 o DRV se eleva para 50mV, o que pode ser confirmado por simulações SPICE segundo Qin (2007). Esta é uma avaliação não realística da DRV pois não assume variações de processo, tornando este modelo fraco para aplicação a uma situação realística. Como pode ser visto na figura 3.21 B) onde o SNM varia através das variações de processo de 3σ na largura de canal e no Vth, sendo as linhas sólidas o modelo operando no caso típico.

O descasamento (*mismatch*) tem impacto direto e causa impacto direto no SNM como pode ser visto na figura 3.21 B). Assim pode-se concluir que para memórias o maior impacto negativo se dá pela variação local, pois estas afetam diretamente o casamento dos transistores da célula de memória provocando variações no SNM. As

variações globais afetam todos os transistores da estrutura de mesma forma e por isso não impactam tão significativamente o SNM como demonstrado pela figura 3.22. Finalmente a influencia da temperatura no DRV é relativamente fraca, sendo na ordem de 10% da variação de 3σ do *mismatch* quando se passa de 27ºC para 100ºC segundo Qin (2007).

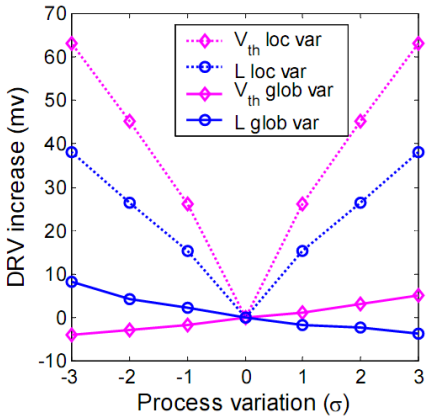


Figure 3.22 – Degradação da tensão de DRV por variações de processo locais e globais (QIN, 2007)

Finalmente Qin (2007) remodela o valor de DRV para considerar as variações de processo globais e locais como:

$$DRV = DRV_{mathed} + \sum_i a_i \frac{\Delta\beta_{i,local}}{\beta} + \sum_i b_i \cdot \Delta V_{thi,local} + c\Delta T \qquad \text{(eq. 3.10)}$$

Onde $DRV_{matched}$ é a tensão de DRV somente com as variações globais na temperatura ambiente; a_i , b_i e c são coeficientes de ajuste para cada um dos transistores. Os termos $\Delta\beta_{i,local}$ e $\Delta V_{thi,local}$ são os termos que representam as variações locais dos transistores e ΔT é a variação de temperatura de todo o chip. Segundo Qin (2007) os parâmetros a_i e b_i são extraídos de simulações SPICE e o fator c tem o valor de 169mV/ºC incrementando-se em 12,3mV quanto a temperatura sobe de 27ºC para 100ºC.

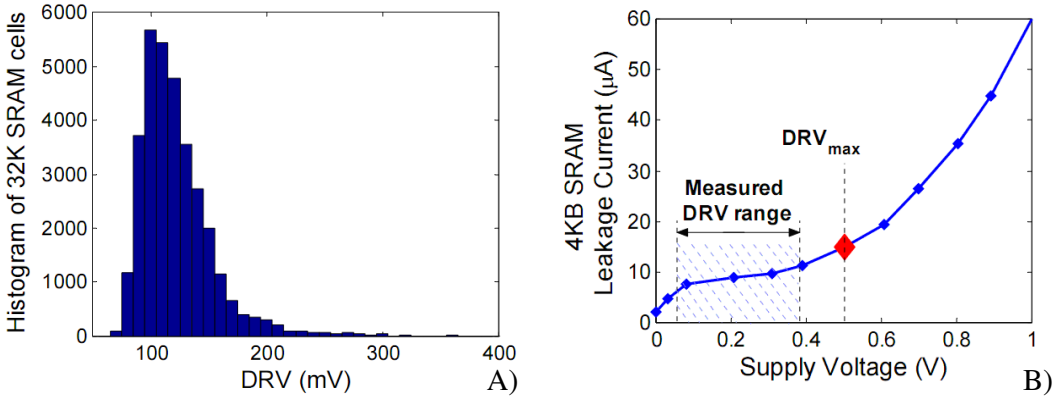


Figure 3.23 – (A) Tensão de DRV mínima para as 32K células de uma SRAM prototipada em 130nm e (B) Corrente de fuga medida na mesma SRAM para blocos de 4K células (QIN, 2007)

A figura 3.23 A) apresenta as tensões mínimas de DRV para as células de memória prototipadas por Qin(2007) e a figura 3.23 B) apresenta o *leakage* e a tensão de DRV mínima aplicada a todos os blocos com fator de segurança para operação de todas as células dentro de uma variabilidade de 3σ. Finalmente, percebe-se que a correção dos fatores adversos e ajuste das células de memória proporciona a queda da tensão mínima de DRV para um valor mais próximo do limite teórico com exposto na figura 3.24 B); ao mesmo tempo cada uma das correções proporciona o aumento do SNM da memória com exposto na figura 3.24 A). O ponto a ser analisado é o compromisso entre custos em área e melhoras de desempenho, lembrando-se que a célula será replicada milhares de vezes.

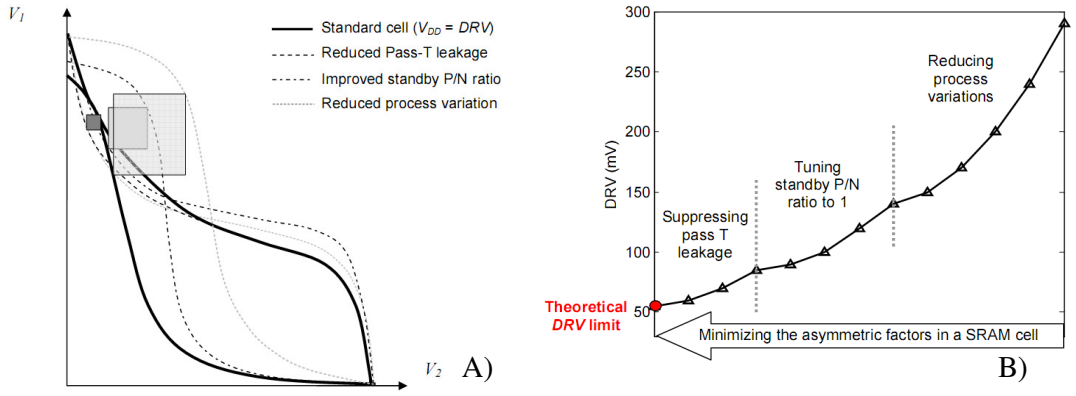


Figure 3.24 – (A) Variação do SNM durante retenção e (B)Variação da tensão de DRV mínima em uma SRAM prototipada em 90nm através da aplicação técnicas para compensação de erros de fabricação (QIN, 2007)

3.4.4.2 DVS *(Dynamic Voltage Scaling)*

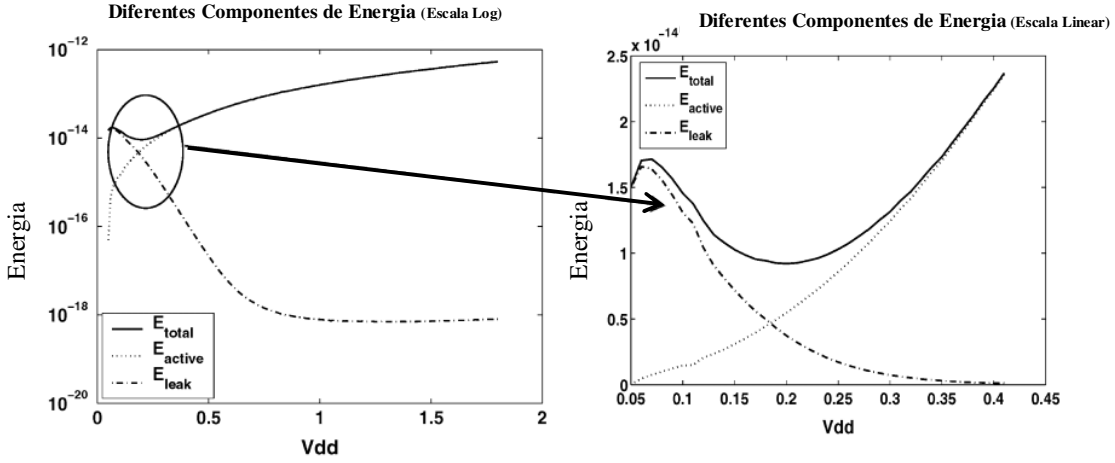


Figure 3.25 – Energia consumida como função da tensão de alimentação (ZHAI, 2005)

A técnica de DVS, ajuste dinâmico de tensão (*Dynamic Voltage Scaling*), é grandemente utilizada nos dias de hoje em processadores e circuitos integrados para redução da potência consumida. A ideia principal é reduzir a tensão quando não há necessidade de máxima capacidade de processamento, reduzindo a frequência de operação o que reduz a energia utilizada no sistema. Através dos experimentos de Zhai (2005) pode-se constatar que há uma tensão onde a energia consumida em *leakage*

torna-se o fator determinante na energia consumida, através dos testes executados com 50 inversores pode-se visualizar o valor mínimo interessante para redução da tensão de operação como demonstrado na figura 3.25.

Esta técnica apresenta dados interessantes quanto a tensão de retenção de inversores demonstrando que abaixo de um valor de tensão a elevação da energia consumida em fugas se torna dominante, sendo assim interessante permanecer próximo ao ponto de mínimo consumo de energia na escolha na tensão de DRV de uma memória do tipo *drowsy*.

3.4.5 Dispositivos e sua variação com dopagem:

Outra maneira de conseguir melhoras em economia de energia é através da variação do processo com o objetivo de alcançar as características desejadas. Criam-se então os processos especiais que são ajustados de forma a terem as características desejadas no projeto ali processado, um exemplo típico são processos ajustados para produção de SRAMS. A figura 3.26 apresenta a dependência das correntes de *leakage* da dopagem, demonstrando como é possível controlar estas características, já na figura 3.27 A) apresenta-se a variação do tempo de acesso versus a dopagem e na figura 3.27 B) temos a corrente de fuga em uma célula de SRAM versus a dopagem.

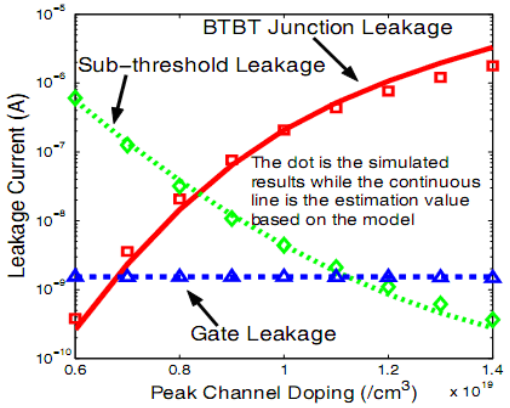


Figure 3.26 – Componentes do *leakage* versus a dopagem no canal (CHEN, 2007).

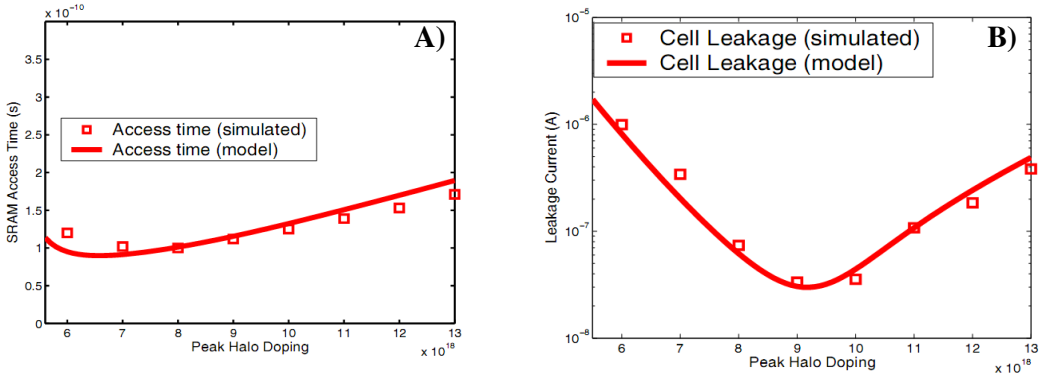


Figure 3.27 – Dopagem do canal versus A) Tempo de acesso de uma SRAM e B) Corrente de fuga (CHEN, 2007).

4 PROJETO DE UMA MATRIZ DE MEMÓRIA SRAM E RESULTADOS SIMULADOS

Com o *scaling* das tecnologias e o incremento do consumo estático, torna-se cada vez mais importante valer-se de estratégias para obter circuitos com menor consumo tanto dinâmico como estático. A composição de técnicas para economia de energia dinâmica devem ser aplicadas no circuito de uma SRAM para que se reduza o consumo dinâmico e, se possível, haja maior performance. Ao mesmo tempo a aplicação das técnicas de economia de energia estática devem ser colocadas em uso de forma a possibilitar um estado de baixo consumo para a memória, quando não é acessada.

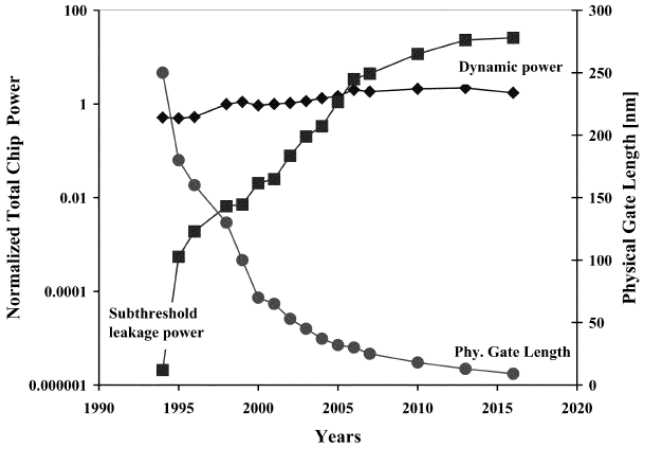


Figura 4.1 – Potências dinâmica e estática normalizadas para dispositivo de W/L=3. (ITRS, 2001)

Como é demonstrado na figura 4.1, com o passar dos anos, temos *gates* cada vez menores e, consequentemente, tensões de operação menores. Segundo Li (2004) em óxidos muito finos, a cache do tipo *drowsy* proporciona uma melhor economia de energia. Qin (2007) apresenta conclusões expressivas, demonstrando que as caches do tipo *drowsy* têm se aproximado na ordem de economia de energia das caches do tipo *decay*, até seu trabalho, ambas na ordem dos 70%. Ao fim de seu estudo, provou que uma cache do tipo *drowsy*, com controle de erros, pode chegar a 98% de economia de energia, demonstrando que esta é a melhor escolha para este projeto. Assim após uma avaliação criteriosa da literatura, pode-se demonstrar que uma cache do tipo *drowsy* apresenta diversas qualidades superiores a uma cache do tipo *decay*.

Um primeiro fator a ser avaliado é a validade do dado na memória, numa cache *decay* ele é perdido no momento que a memória entra em modo de economia de energia. No caso de uma *drowsy* o dado permanece salvo e disponível, assim pode ser utilizada tanto para dados bem como para instruções, tendo assim melhor aplicabilidade sem alterações estruturais de projeto. No caso de uma cache do tipo *decay*, a mesma se torna um fator negativo se não houver um ajuste do intervalo de entrada em modo de economia de energia.

Um segundo fator a ser avaliado é a possibilidade de ajuste da tensão de modo de espera, o que gera mais um ponto de ajuste para maior economia de energia. A queda do *leakage* tem relação exponencial com a tensão de alimentação, pode-se obter um modo de grande economia de energia como demonstrado na figura 4.2, com a redução da corrente sub-limiar de fuga. A queda do valor de alimentação leva a redução das componentes do *leakage*, apresentadas na figura 4.3, estas já equacionadas e relacionadas no tópico 3.2 deste texto.

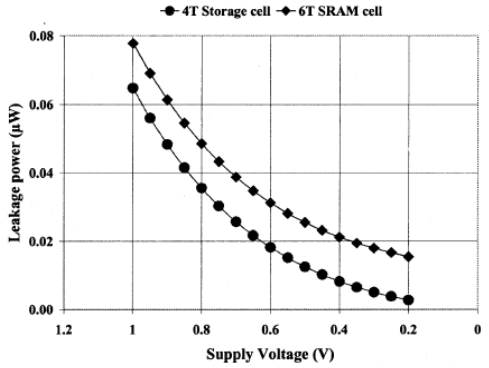


Figura 4.2 – Célula de SRAM em modo drowsy demonstrando a redução do *subthresholds leakage* com o DVS em tecnologia 70nm. (KIM, 2004)

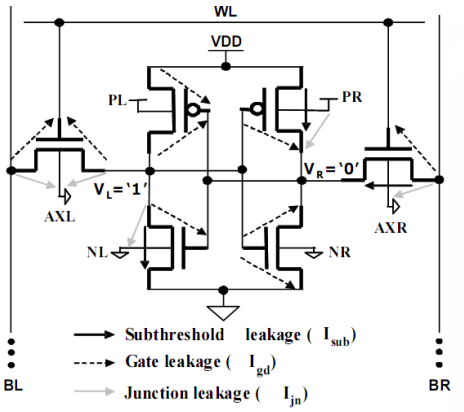


Figura 4.3 – Componentes da corrente de fuga em uma célula SRAM de 6 Transistores (CHEN, 2007)

A utilização de transistores de acesso de diferentes Vths pode ocasionar o incremento de desempenho com o ônus da elevação do *leakage* da célula para as bit-lines ou das bit-lines para a célula de memória. A figura 4.4 demonstra uma célula de memória de seis transistores tradicional com transistores de acesso que quando tem seu Vth variado apresentam uma relação entre tempo de acesso e quantidade de redução de *leakage*.

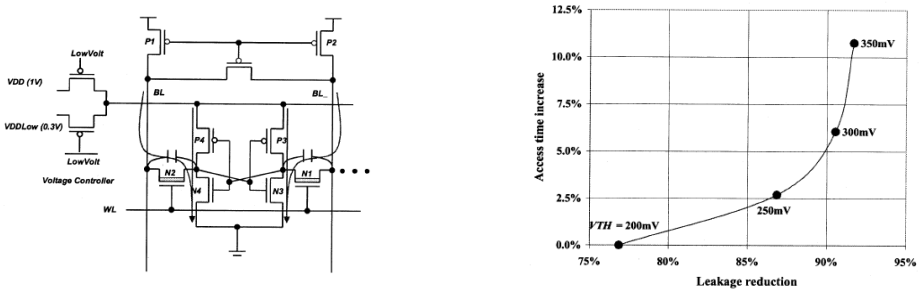


Figura 4.4 – Avaliação do aumento do atraso por utilização de transistores de acesso, N1 e N2, na SRAM com *threshold* mais elevado (KIM, 2004).

Outro fator demonstrado por Sanchez-Sinencio (2008) em um estudo de uma tecnologia 65nm, é a variação dinâmica de Vth por efeitos de canal curto em transistores de comprimento mínimo de canal como demonstrado na figura 4.5. Com o *scaling* dos transistores tornam-se mais sensíveis aos efeitos de canal curto, criando-se a necessidade de circuitos robustos a variações de processo e que suportem cada vez maior variação, imposto pelas novas tecnologias. Na figura 4.6 Kwai (2006) apresenta o efeito em uma célula SRAM, demonstrando-se que ao reduzir-se a tensão de alimentação a variabilidade do *threshold* reduz-se, ou seja, a variância é reduzida. Nos testes os resultados simulados apresenta-se a corrente pelo transistor de acesso indo ao Vss através do transistor N de um dos inversores da memória.

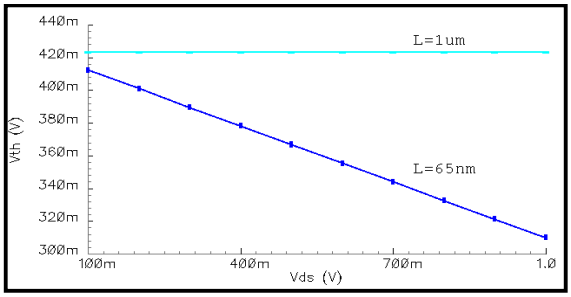


Figura 4.5 – Tensão de *threshold* versus tensão de dreno para L=65nm e 1um, W=10µm, Vgs=1V e Vds=0V (SANCHEZ-SINENCIO)

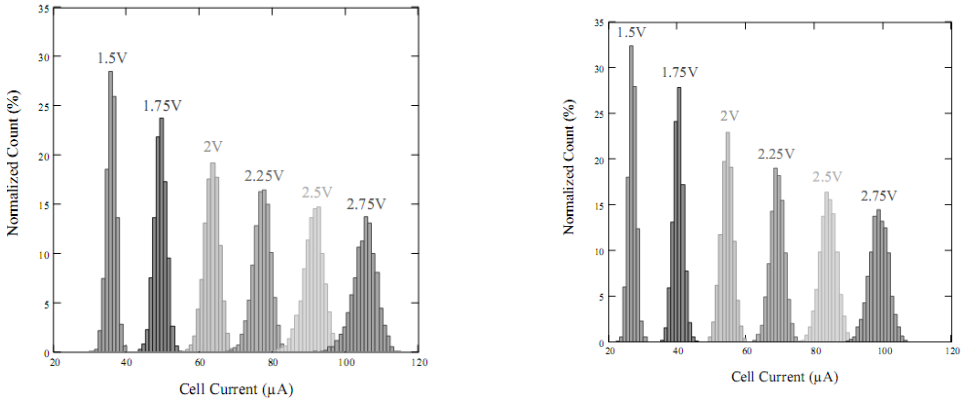


Figura 4.6 – Variação tecnologia 0,25µm Tensão de *threshold* versus tensão de dreno para L=65nm e 1um, W=10µm (KWAI, 2006)

O incremento de área ativa nos transistores reduz a variabilidade do Vth dos transistores, mas em células de memória qualquer aumento de área representa aumento

indesejado de custos. Por outro lado olhando-se o processo de produção de um CI constata-se que o aumento da área também leva a um incremento da possibilidade de incidência de erros na área da memória, encarecendo assim os CIs funcionais produzidos. Estes fatores econômicos incidem sobre o projeto de SRAMs fazendo com que o numero de transistores aceitáveis para uma SRAM comercial varie entre seis a oito, segundo Prince (2007) na tabela 4.1. A utilização de áreas acima destes limites acontece principalmente em pesquisas, onde células com até dez transistores estão publicadas com fins específicos de utilização e aplicação. A tabela 4.1 apresenta o panorama comercial da fabricação de memórias para SRAMs, DRAMs e Flash; já na figura 4.7 apresenta-se a área média de uma célula de memória através das tecnologias de produção.

Tabela 4.1 – Característica das Memórias MOS *Standalone* (PRINCE, 2007)

	SRAM	DRAM	Flash
Velocidade de Leitura	Rápida (ns)	Média (ns)	Média (ns)
Velocidade de Escrita	Rápida (ns)	Média (ns)	Lenta (ms / s)
Não Volatilidade	Não	Não	Sim
Tamanho da Célula	6 tr - 8 tr	1.5 tr (1T1C)	1 (1T)
Tipo de Célula	Latch CMOS	Capacitor	Gate Flutuante
Densidade	32 Mb	4 Gb	8Gb

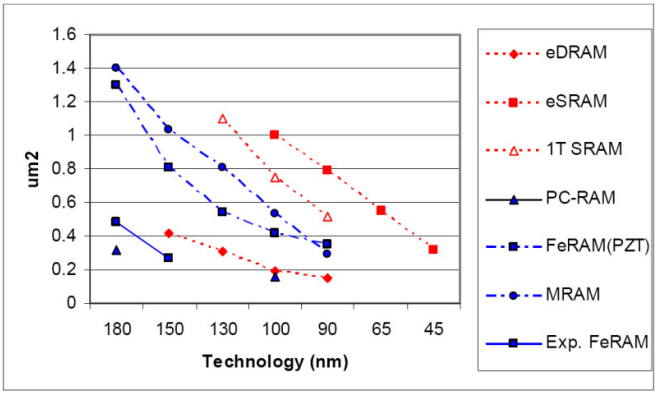


Figura 4.7 – Tamanho das células de memória em tecnologias Nanométricas (PRINCE, 2007)

Alguns dos problemas potenciais provenientes do *scaling* são: - Redução da tensão de alimentação e consequente redução da estabilidade da célula (SNM); - Incremento das correntes de fuga e da dissipação de potência estática; - Instabilidades nas operações dinâmicas decorrentes dos efeitos de canal curto. Pelo incremento da variabilidade criam-se a instabilidade no *latch* pela redução do SNM, técnicas já consagradas como *back bias* tornam-se menos efetivas (GIELEN, 2005) e cria-se a necessidade de novas técnicas para atingir os requerimentos de projeto dos CIs modernos.

4.1 Resultados de Simulação de Corrente DC de Fuga

Para avaliar os transistores da tecnologia alvo foram implementados inversores mínimos utilizando os diferentes modelos disponíveis na tecnologia ST 65nm, Lmin = 60nm, sendo eles hvtgp (*High Vth*), svtgp (*Standard Vth*) e lvtgp (*Low Vth*). Avaliando-se as correntes de fuga em cada um dos inversores para obter seu comportamento frente às variações de alimentação e temperatura. As simulações foram efetuadas utilizando os modelos disponibilizados pela tecnologia que utiliza modelos elétricos BSIM 4.1. A figura 4.8 apresenta os circuitos implementados e as figuras 4.9, 4.10, 4.11 e 4.12 os resultados simulados.

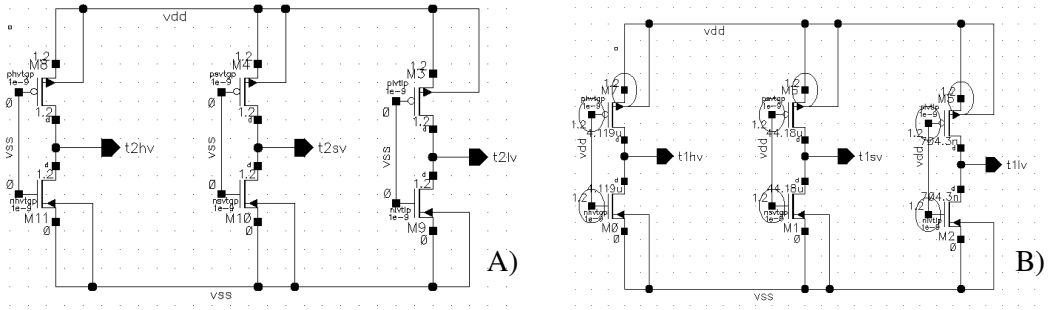


Figura 4.8 – Inversores Mínimos com entradas em (A) Vss e (B) Vdd para medição do *leakage* através dos transistores

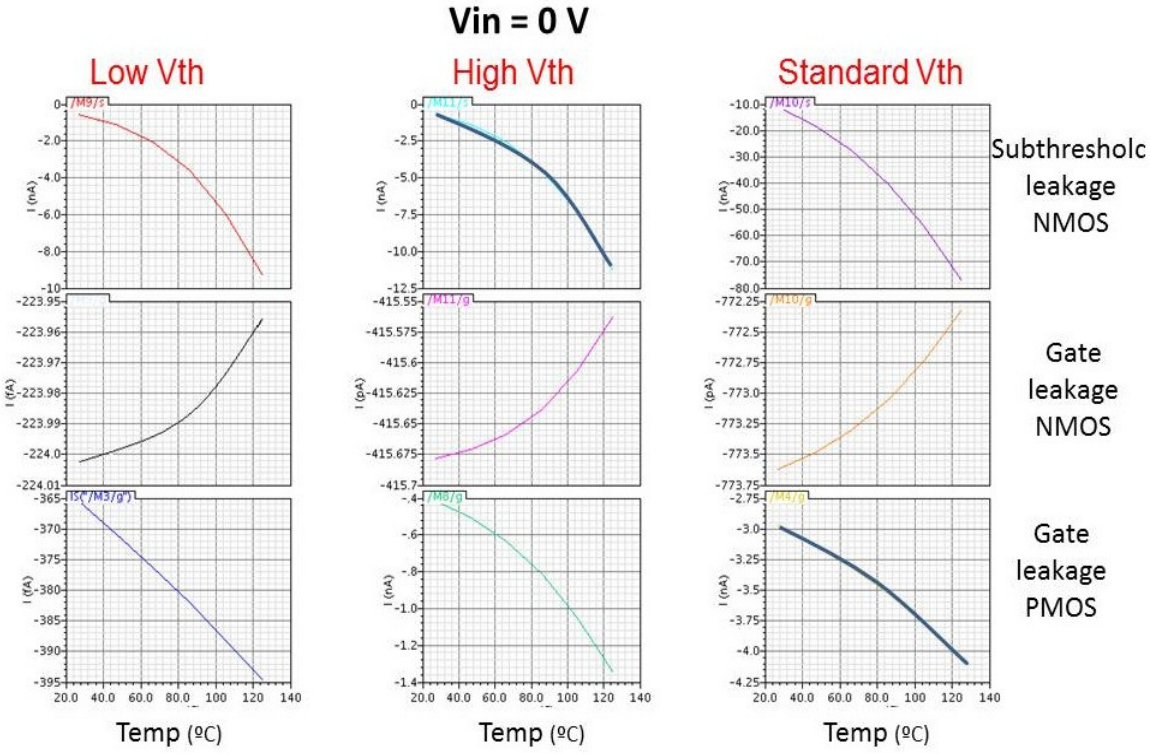


Figura 4.9 – Variação das componentes da corrente de fuga através dos inversores mínimos com entradas em Vss e variação de temperatura de 27ºC a 125ºC

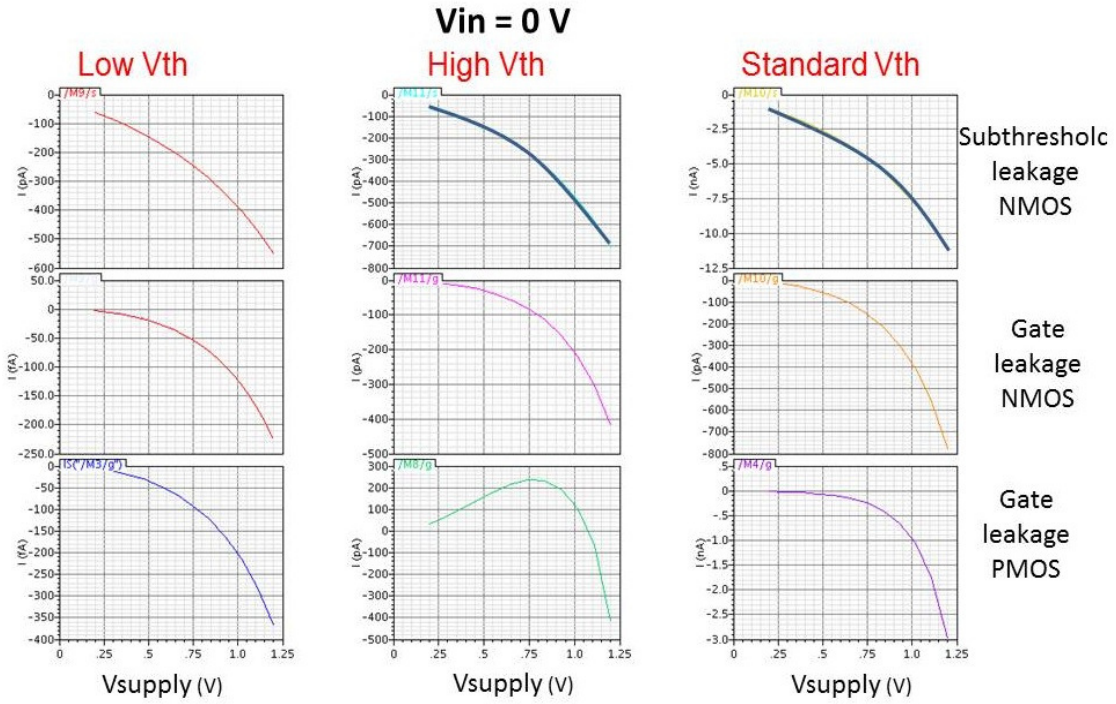


Figura 4.10 – Variação das componentes da corrente de fuga através dos inversores mínimos com entradas em Vss e variação de Vdd de 0,2V a 1,2V

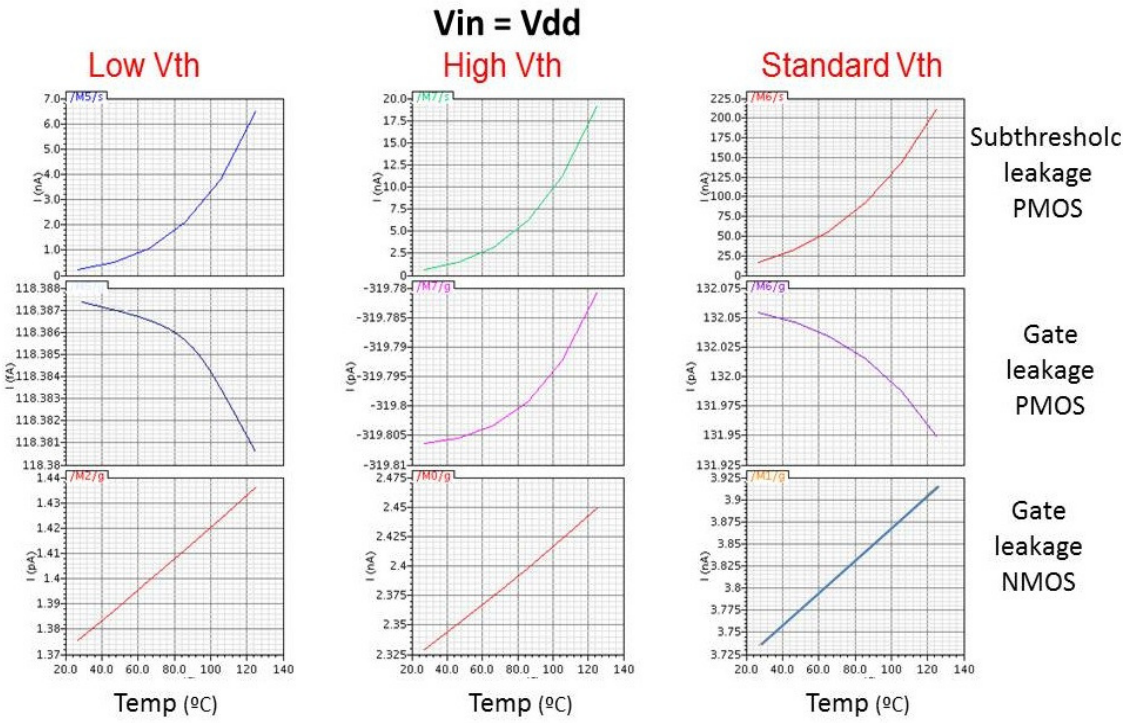


Figura 4.11 – Variação das componentes da corrente de fuga através dos inversores mínimos com entradas em Vdd e variação de temperatura de 27ºC a 125ºC

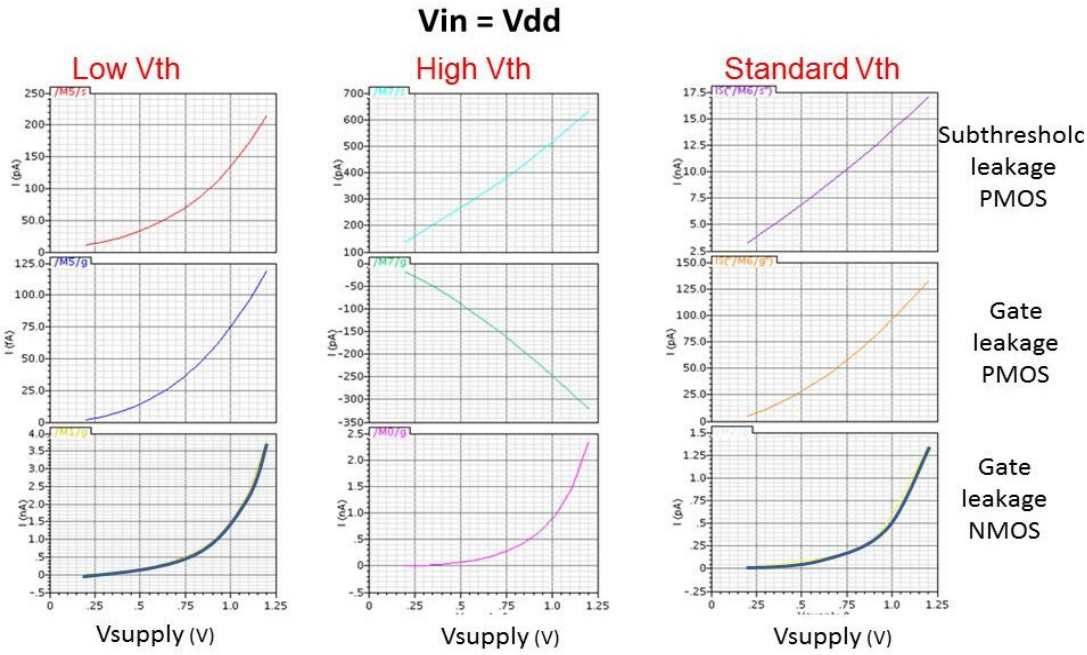


Figura 4.12 – Variação das componentes da corrente de fuga através dos inversores mínimos com entradas em Vdd e variação de Vdd de 0,2V a 1,2V

Percebe-se que a sensibilidade dos transistores PMOS é menor para variações de tensão, sendo a influência da temperatura similar para ambos os transistores. Para investigar a variação de corrente de fuga decidiu-se por testar os transistores através de simulações Monte Carlo e assim adquirir o valor da variação estatística do *leakage* nos inversores.

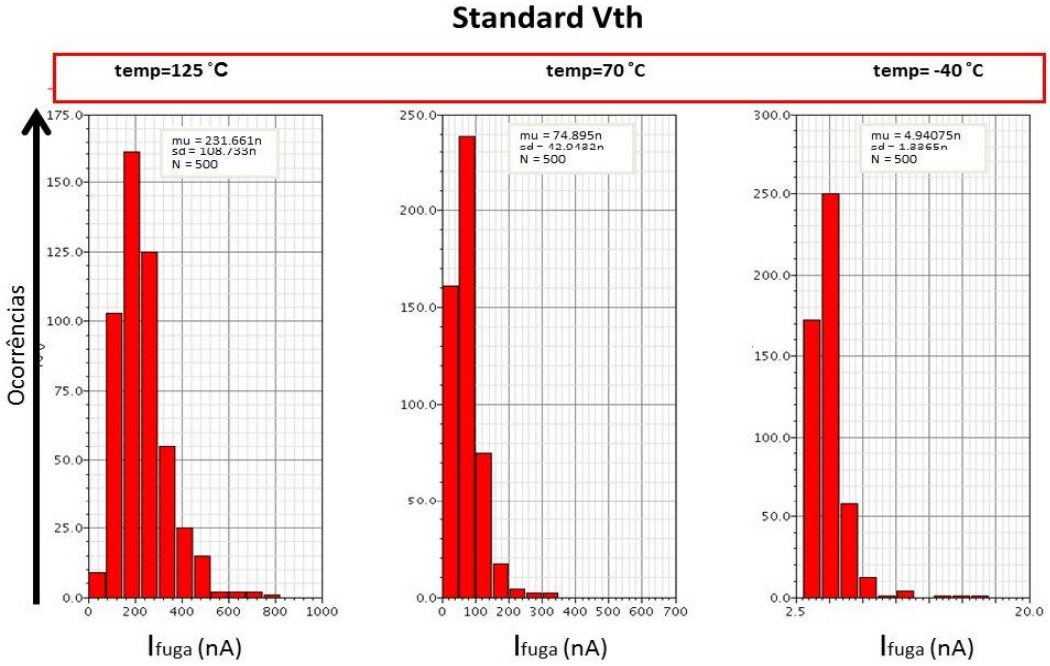


Figura 4.13 – Variação das componentes da corrente de fuga através dos inversores standard Vth com variação da temperatura

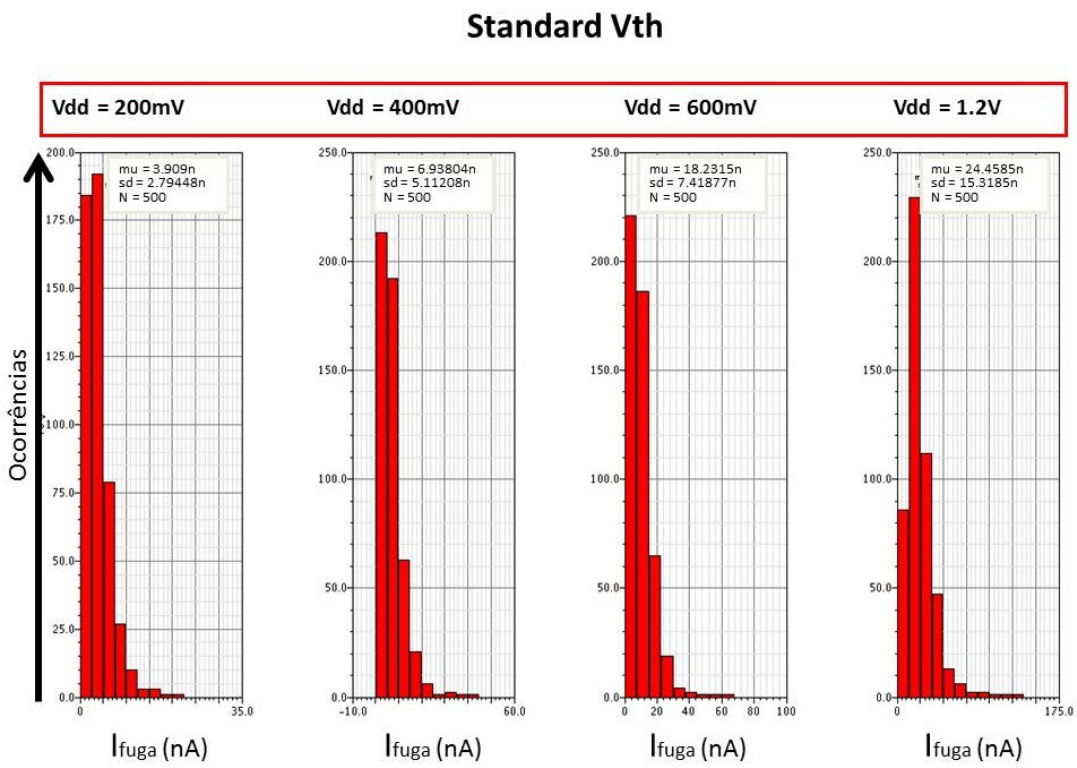


Figura 4.14 – Variação das componentes da corrente de fuga através dos inversores standard Vth com variação da tensão de alimentação

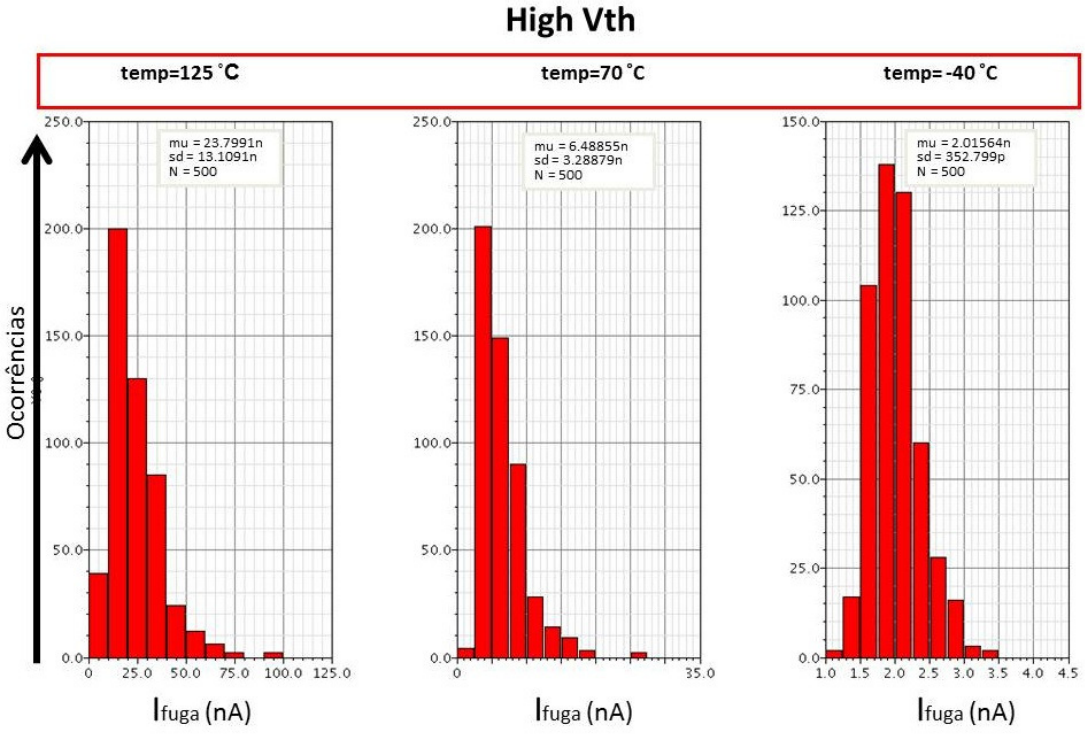


Figura 4.15 – Variação das componentes da corrente de fuga através dos inversores *high* Vth com variação da temperatura

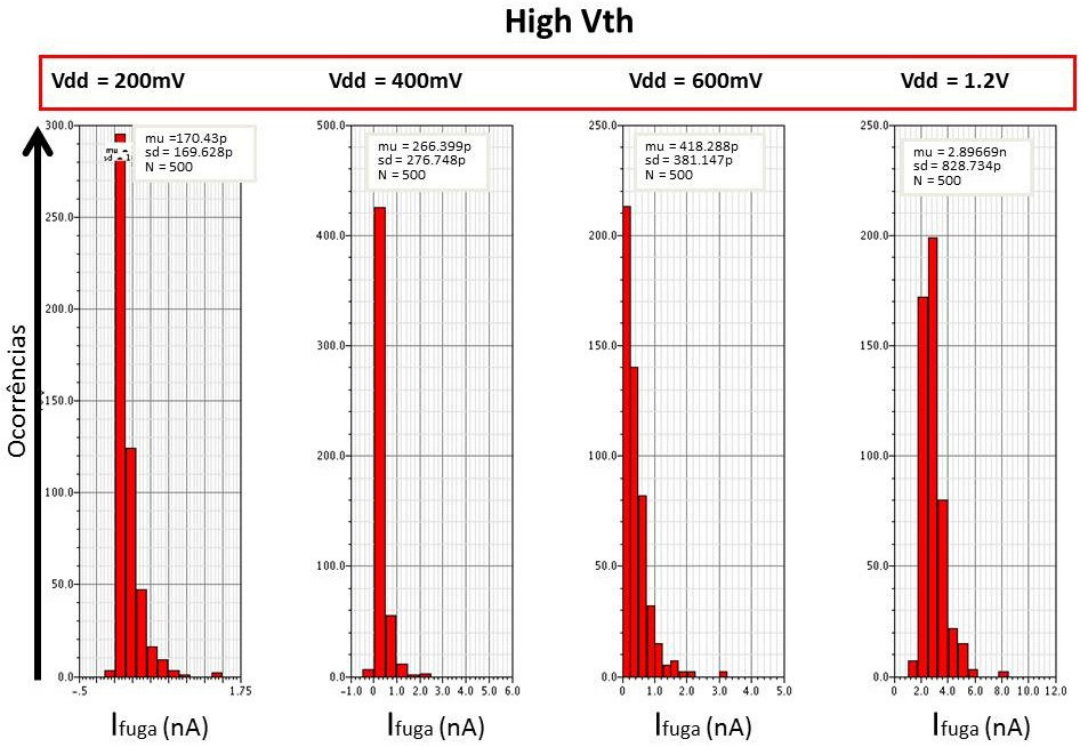


Figura 4.16 – Variação das componentes da corrente de fuga através dos inversores *high* Vth com variação da tensão de alimentação

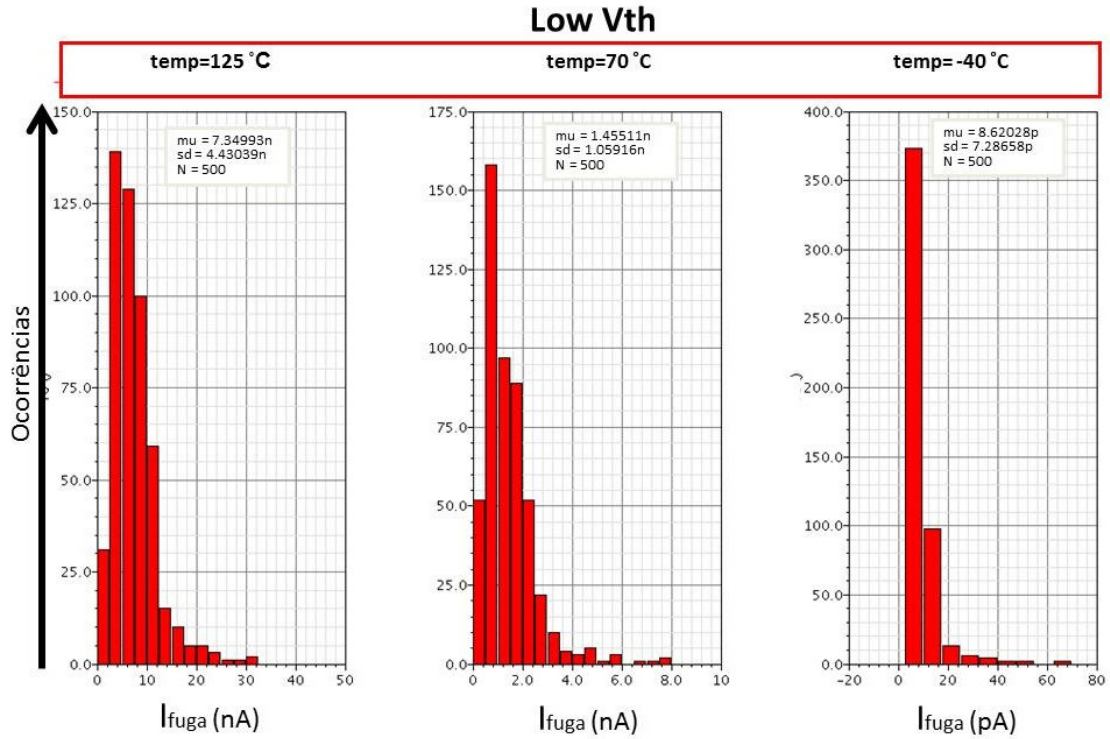


Figura 4.17 – Variação das componentes da corrente de fuga através dos inversores *low* Vth com variação da temperatura

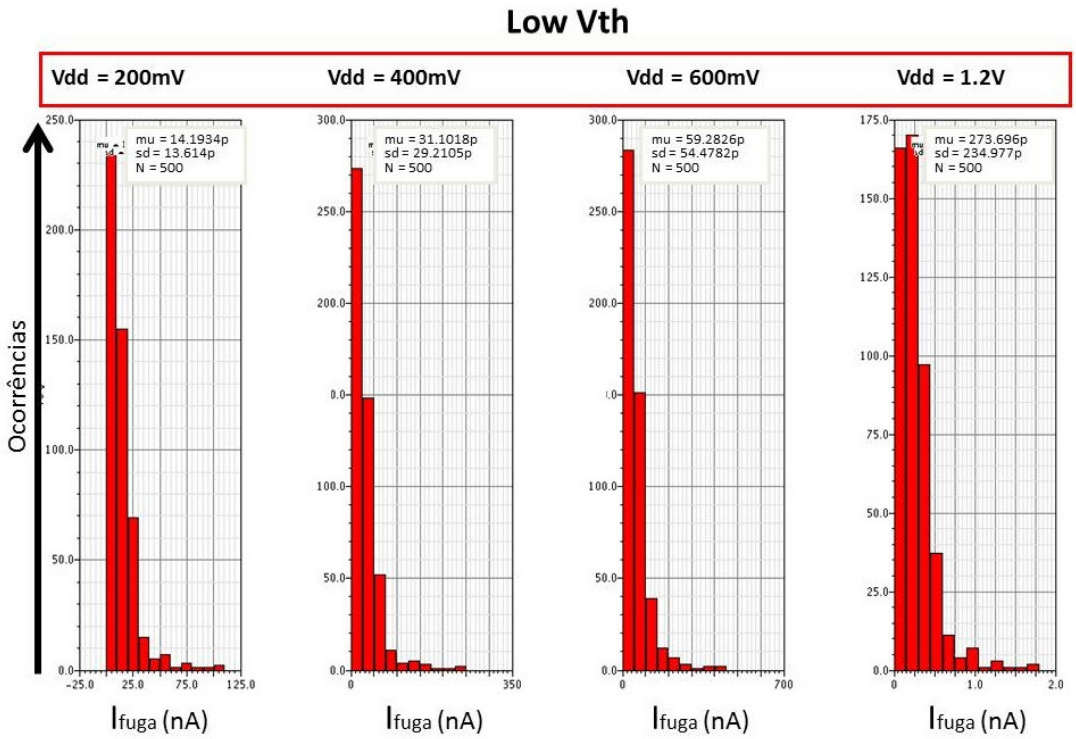


Figura 4.18 – Variação das componentes da corrente de fuga através dos inversores *low* Vth com variação da tensão de alimentação

De acordo com as previsões a variação do *leakage* reduz-se ao diminuir-se a tensão de *stand-by,* quando comparada a corrente de fuga do modo ativo com o modo de stand-by há uma redução entre 5% e 15% da corrente original. Ao mesmo tempo a temperatura incrementa o *leakage* entre 10 e 1000 vezes entre - 40ºC e 125ºC dependendo do tipo de modelo do transistor.

4.2 Decisões de projeto

Para os testes a ideia é dividir toda a matriz de memória em *bit-lines* hierárquicas e desta forma manter o maior numero de células em *stand-by*. Proporcionando a maior economia de energia possível, seguindo o raciocínio de Zhai (2007) que desenvolveu uma memória 6T-SRAM capaz de funcionar entre 1.2V e 193mV em tecnologia CMOS 0,13µm. Para obter a matriz de memória divide-se os bancos como exemplificado na figura 4.19. Desta forma habilita-se o menor número de células possível e desta forma obtêm-se uma formatação que se tem a potência estática da memória multiplicando as linhas em *stand-by* por banco vezes o numero total de bancos. As células habilitadas estão trabalhando em Vdd e desta forma apenas adiciona-se a potência estática já calculada a potência consumida por estas células ativas chegando ao valor total de potência consumida por toda a memória.

Assim quebrando-se a memória chega-se a uma *bit-lines* de 16 células, valor que resulta em um compromisso entre capacitância parasita e quebra de banco satisfatório e funcional para *bit-lines* hierárquicas. Para os testes utilizou-se uma *bit-lines* de 16 células sendo estas testadas para o seu consumo de corrente estática sobre temperatura e variação de tensão em modo de *stand-by*. A figura 4.20 mostra o setup utilizado para o teste das células de memória.

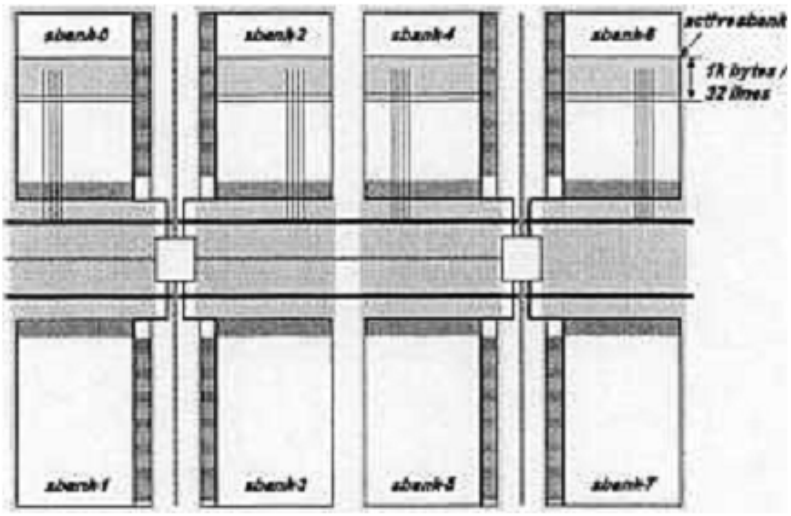


Figura 4.19 – Quebra da Matriz de dados de forma a habilitar a menor quantidade de células para um acesso (KIM, 2002)

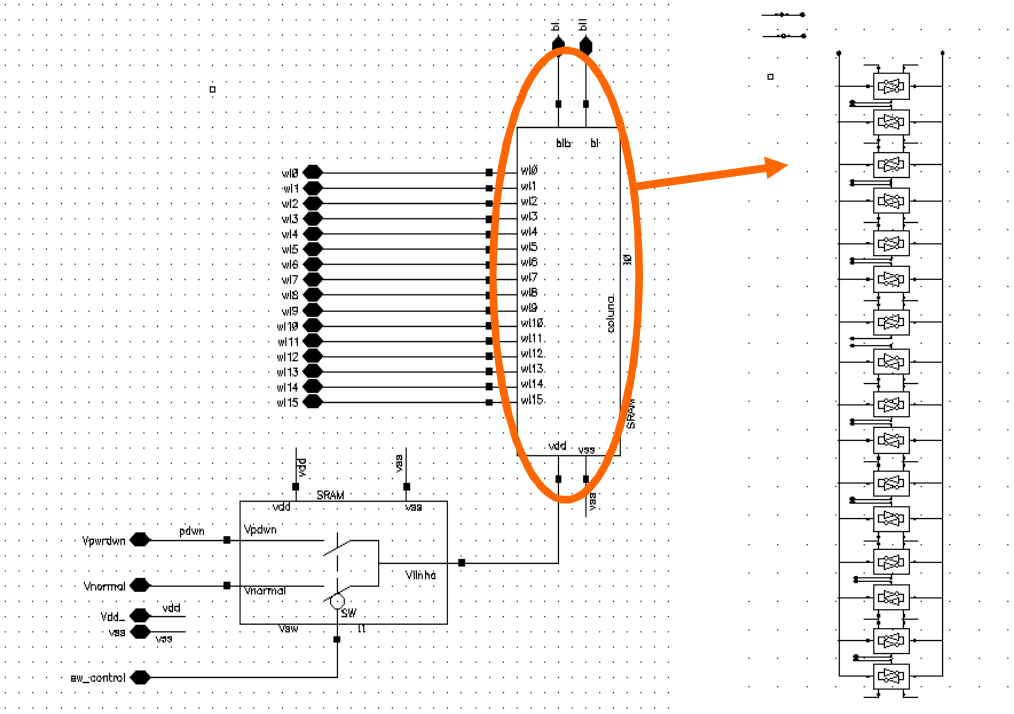


Figura 4.20 – Coluna de célula de memória com 16 células e chave para troca entre modo normal e modo de baixo consumo

Como já previamente exposto decidiu-se pelo modo *drowsy* de redução *de leakage* e por isto definiu-se uma tensão de *stand-by* de 200mV onde tem-se estabilidade do dado gravado. Para fabricação seria interessante que esta tensão possuísse um pequeno ajuste de compensação a eventuais variações de processo.

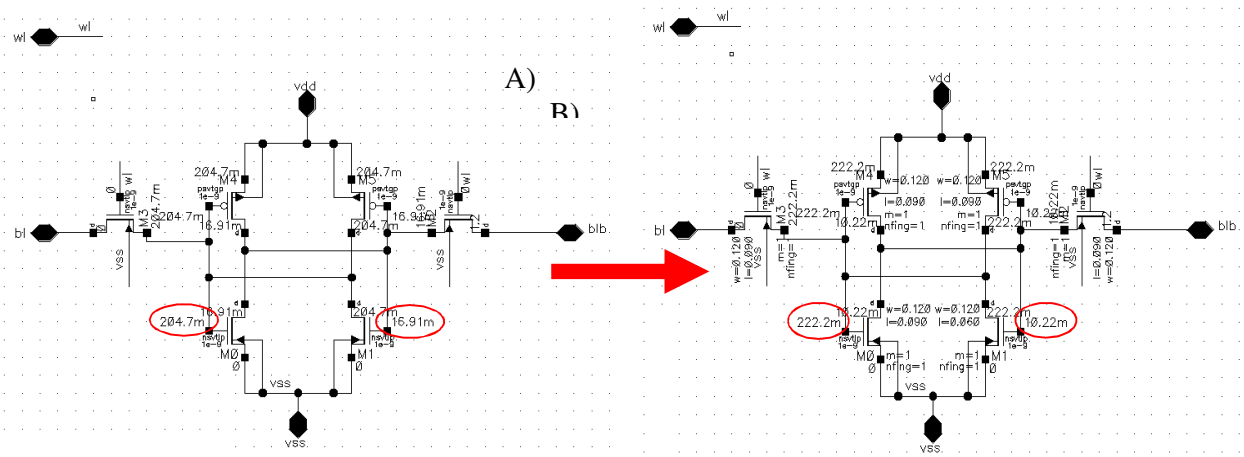


Figura 4.21 – Células de memória *standard* Vth trabalhando em (A) 200mV e (B) 220mV

4.2.1 *Leakage* em Células de memória

A célula de memória tem sua corrente de fuga reduzida por todos os transistores operarem em região de *subthresholds*. Na figura 4.21 tem-se a célula de memória de transistores Standard em 0,2V e 0,22V demonstrando certo incremento de estabilidade pelo acréscimo de 20mV; o valor do zero e do um lógico se distanciam sendo assim mais estáveis. As variações de processo sobre a temperatura e por variação de tensão estão expostas nas figuras 4.22 e 4.23, ambas as simulações Monte Carlo para tais variações.

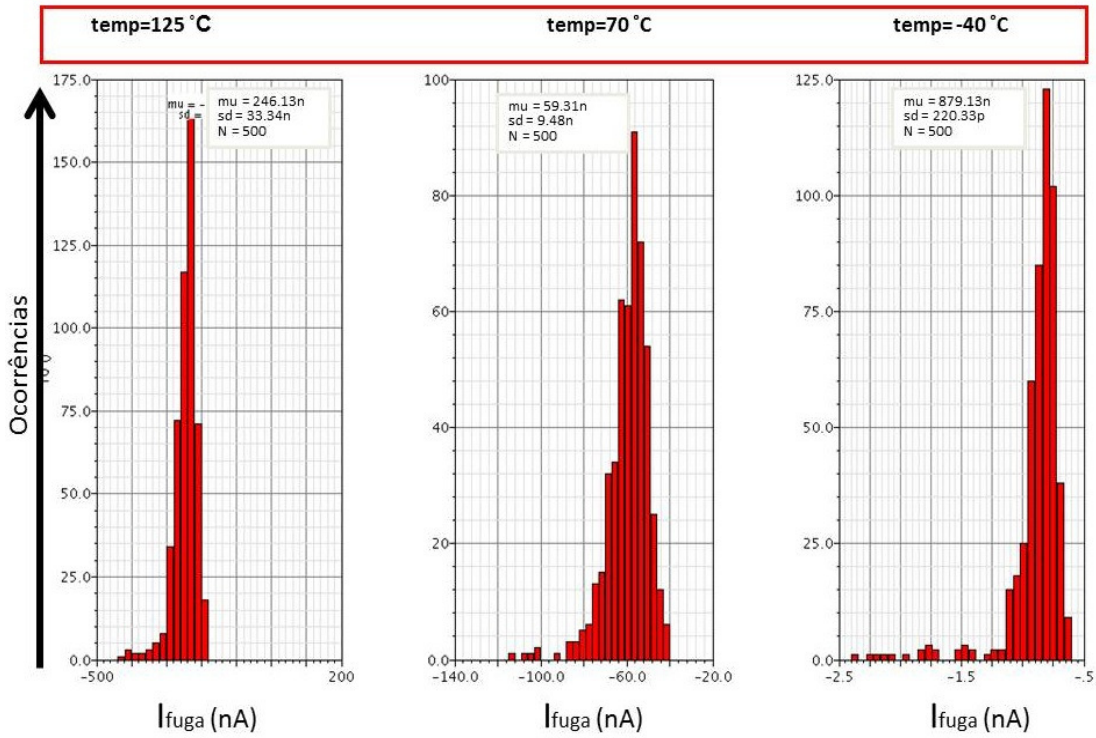


Figura 4.22 – Variação da corrente de fuga através da matriz de células de memória *standard* Vth com alimentação 220mV e variação de temperatura aplicada

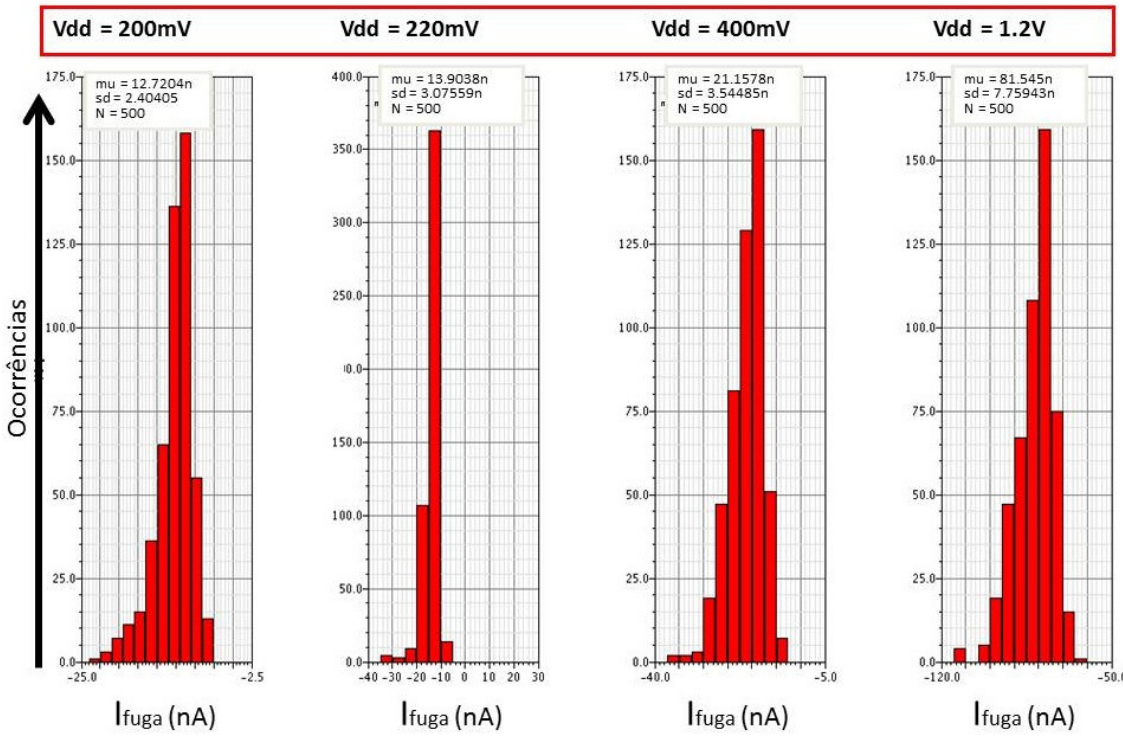


Figura 4.23 – Variação da corrente de fuga através da matriz de células de memória *standard* Vth com temperatura de 27ºC e variação de tensão aplicada

Seguindo as otimizações já apresentadas anteriormente no texto, a troca dos transistores do latch por transistores de *high* Vth e os de passagem por *low* Vth, obtêm-se uma melhora na redução do consumo estático e uma piora nas tensões de retenção do modo de *stand-by* como demonstrado nas figuras 4.24, 4.25 e 4.26.

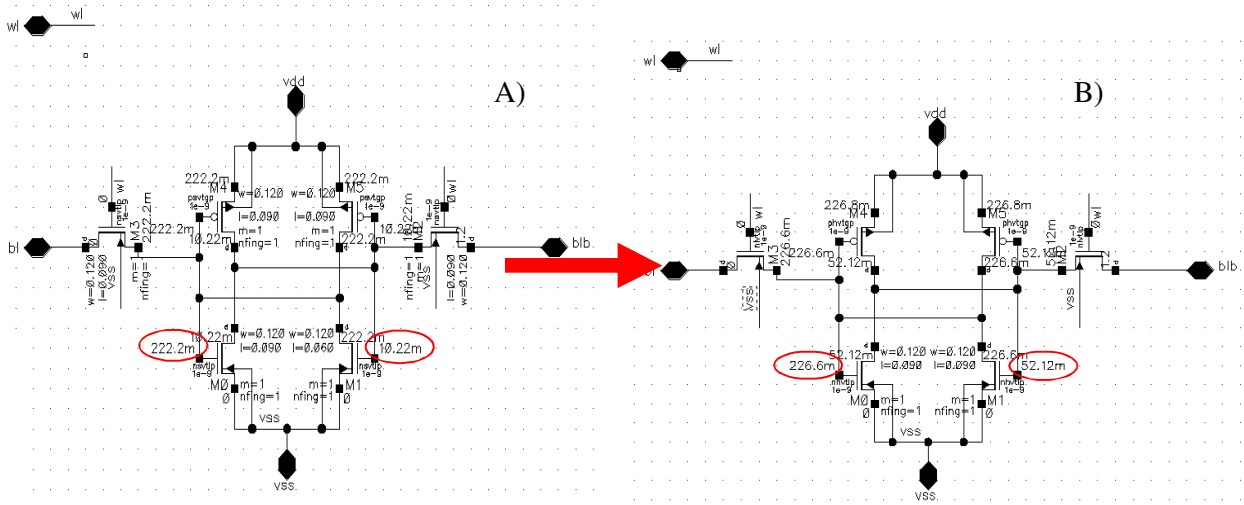


Figura 4.24 – (A) Células de memória *standard* Vth trabalhando em 220mV e (B) Células de memória otimizada para redução do *leakage* trabalhando em 220mV

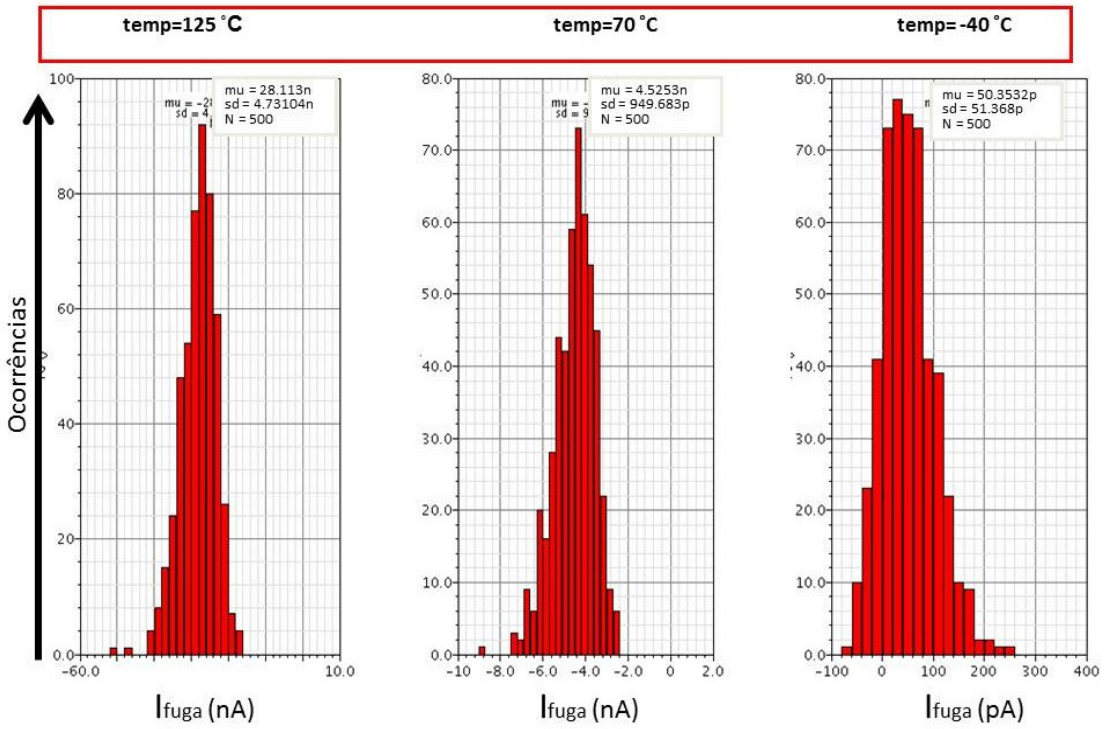


Figura 4.25 – Variação da corrente de fuga através da matriz de células de memória otimizada com alimentação 220mV e variação de temperatura aplicada

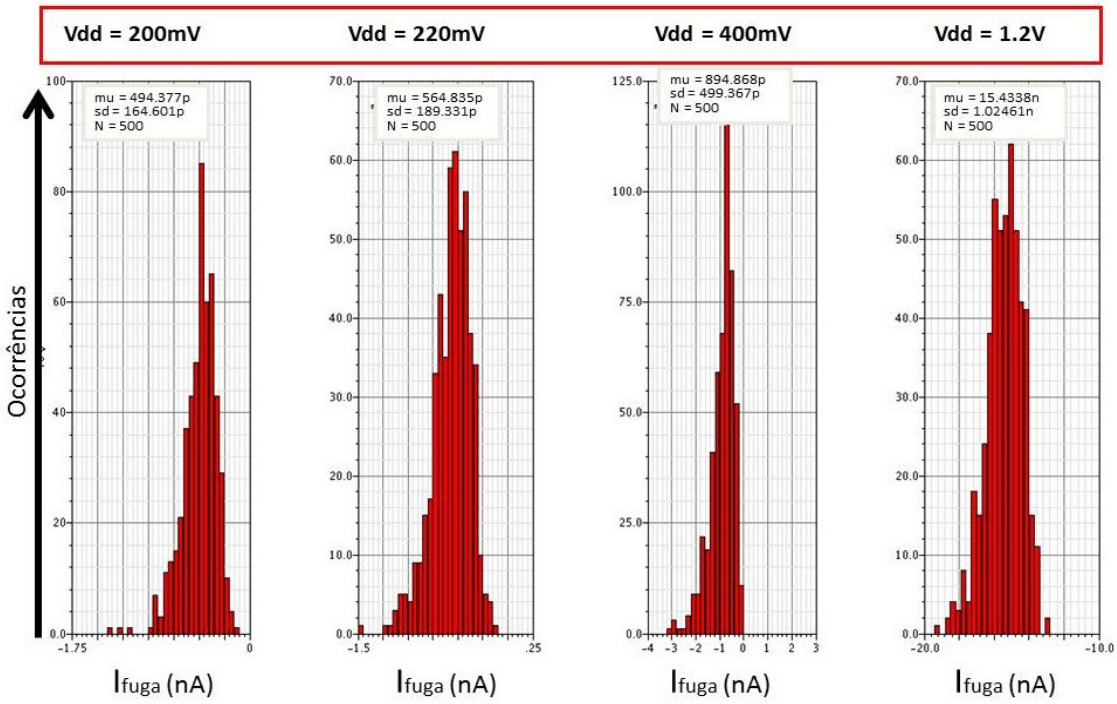


Figura 4.26 – Variação da corrente de fuga através da matriz de células de memória otimizada com temperatura de 27ºC e variação de tensão aplicada

Percebe-se que a célula otimizada realmente proporciona maior economia de energia como previsto anteriormente, sendo a partir deste ponto utilizada como padrão.

4.2.2 Comparação célula otimizada 1.5xLmin versus Lmin

Seguindo as otimizações a troca dos transistores do *latch* por transistores de *high* Vth e os de passagem por *low* Vth, consegue-se uma melhora na redução do consumo estático com o custo do incremento do Lmin, 60nm, para 1,5 x Lmin, 90nm, finalmente investiga-se se realmente há ganho em dispender este aumento de área. Apresentam-se estes resultados nas figuras 4.27, 4.28 e 4.29.

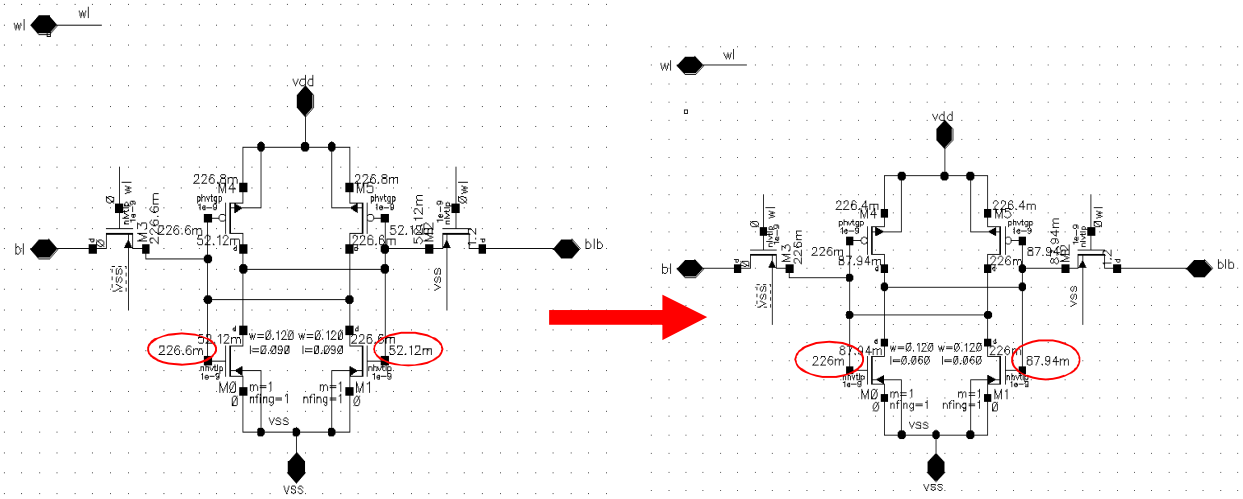


Figura 4.27 – (A) Células de memória otimizada 1,5xLmin trabalhando em 220mV e (B) Células de memória otimizada Lmin trabalhando em 220mV

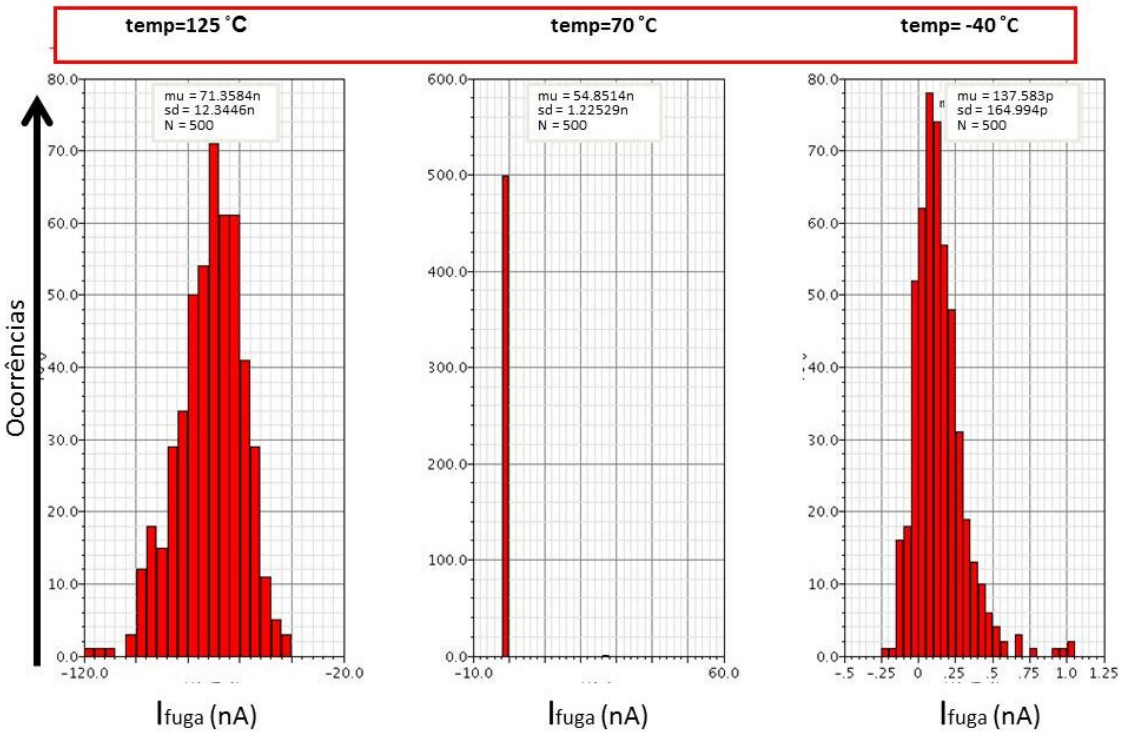


Figura 4.28 – Variação da corrente de fuga através da matriz de células de memória otimizada Lmin com alimentação 220mV e variação de temperatura aplicada

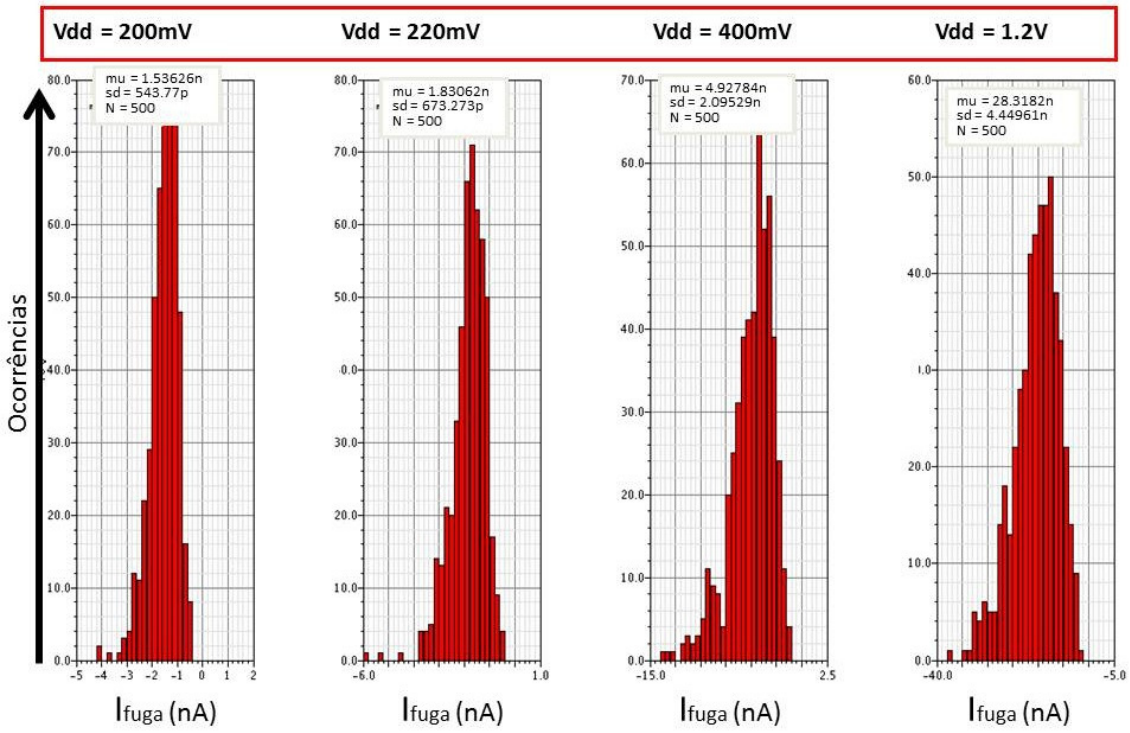


Figura 4.29 – Variação da corrente de fuga através da célula de memória otimizada Lmin com temperatura de 27ºC e variação de tensão aplicada

4.3 Conclusão

Como utilizamos transistores de comprimento de canal mínimo à variação de largura de canal não se faz expressiva em área e por isso não provoca grandes variações na corrente de fuga. O incremento do tamanho dos transistores é um fator inapropriado, pois se reflete na área total da memória e no custo de produção levando-se em consideração o DFT, *Design For Manufacturability*, a análise cuidadosa dos dados obtidos via simulação habilita o projetista a tomar as decisões apropriadas quanto ao incremento de área versus melhora de desempenho ou de redução de *leakage* em modo de *stand-by*.

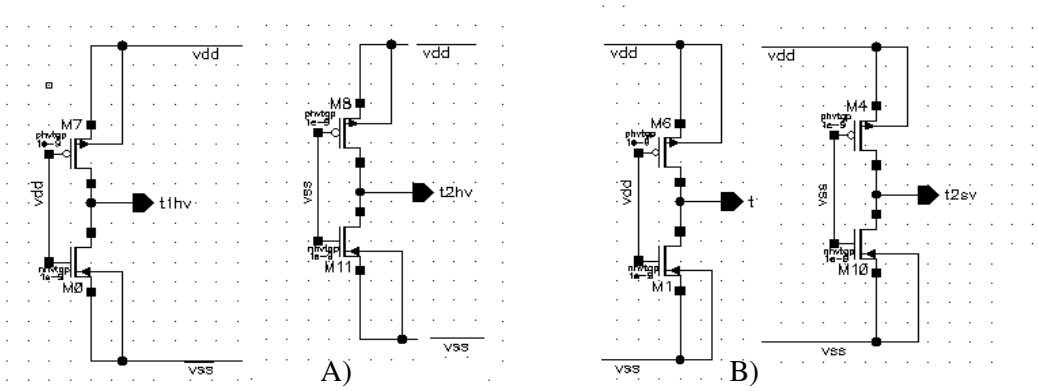


Figura 4.30 – (A) Inversores 1.5x Lmín com entradas em Vdd e Vss (B) Inversores Lmín com entradas em Vdd e Vss

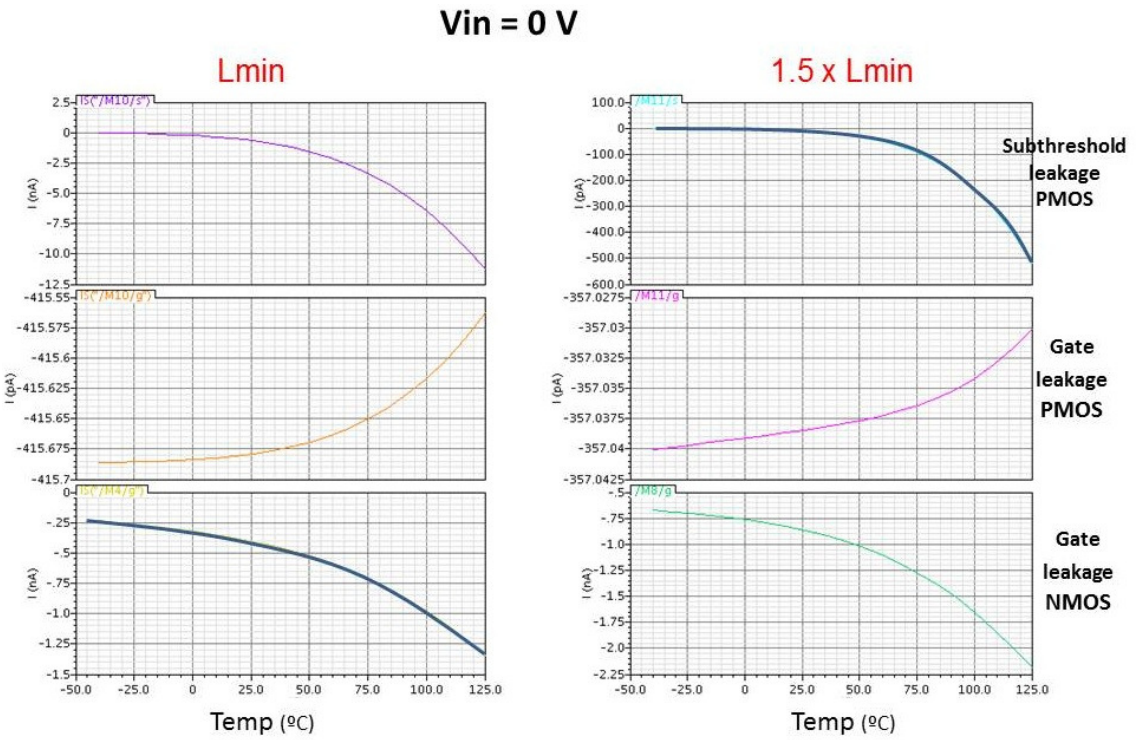


Figura 4. 31 – Variação das componentes da corrente de fuga através dos inversores de W mínimos com entradas em Vss e variação de temperatura de -40ºC a 125ºC

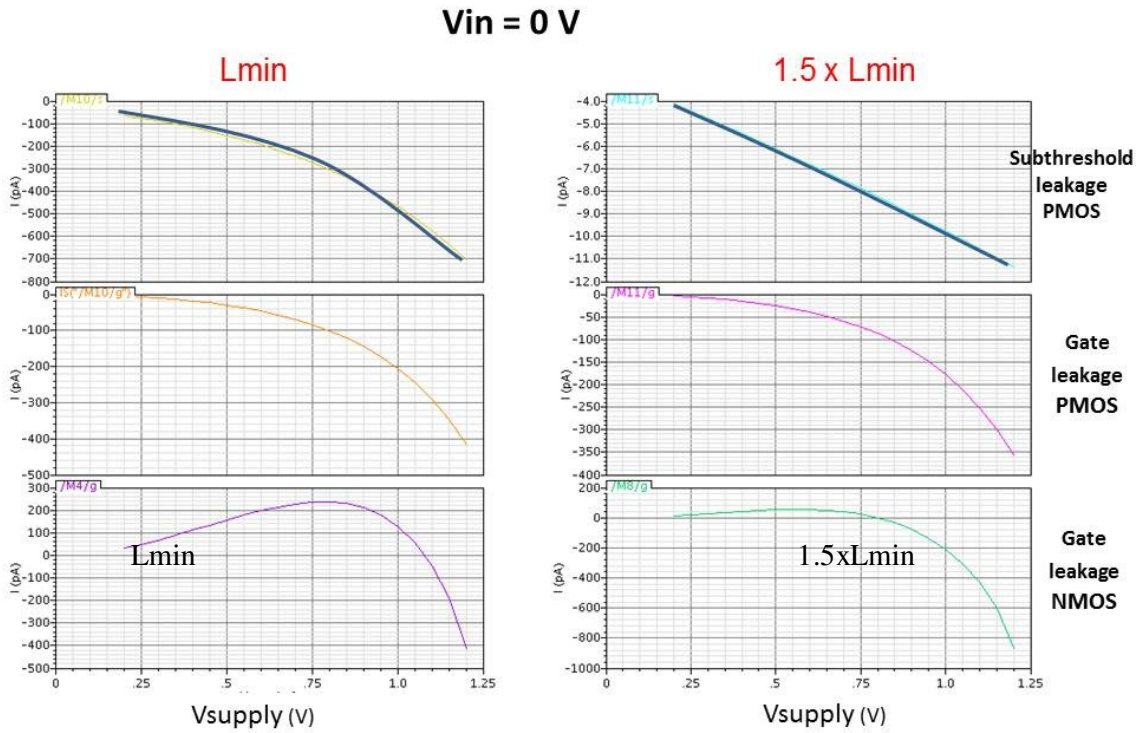


Figura 4.32 – Variação das componentes da corrente de fuga através dos inversores de W mínimo e entradas em Vss e variação de Vdd de 0,2V a 1,2V

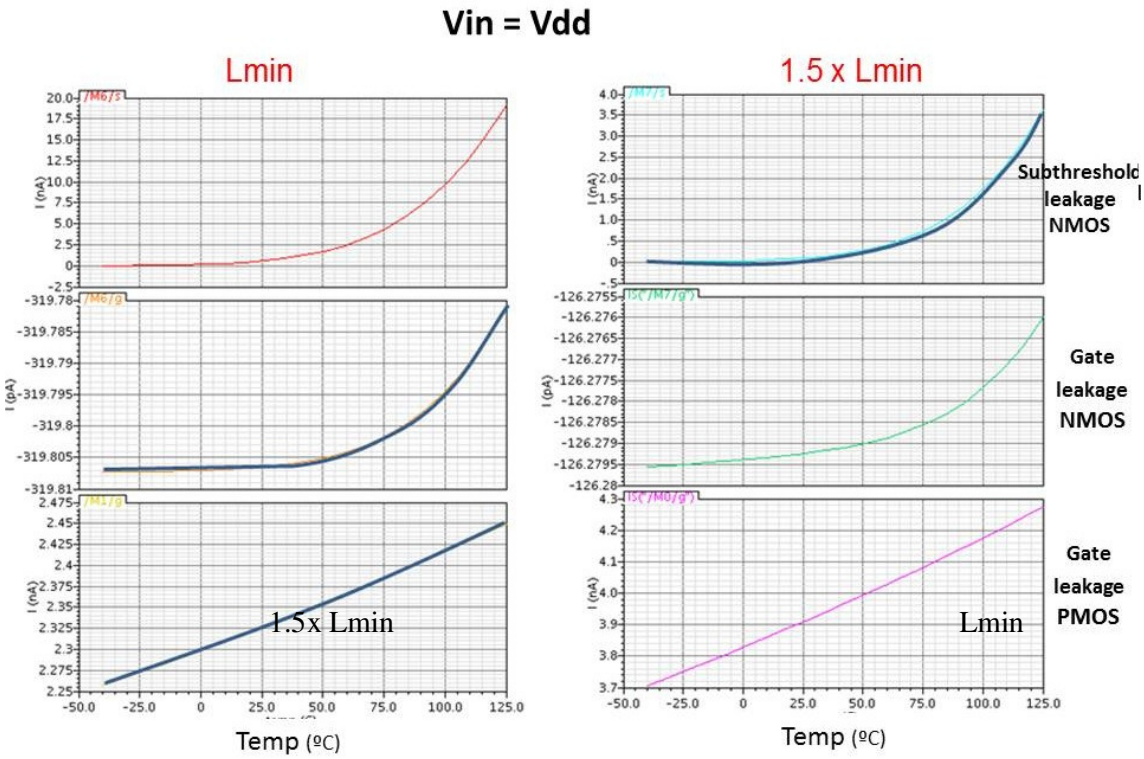


Figura 4.33 – Variação das componentes da corrente de fuga através dos inversores de W mínimo e entradas em Vdd com variação de temperatura de -40ºC a 125ºC

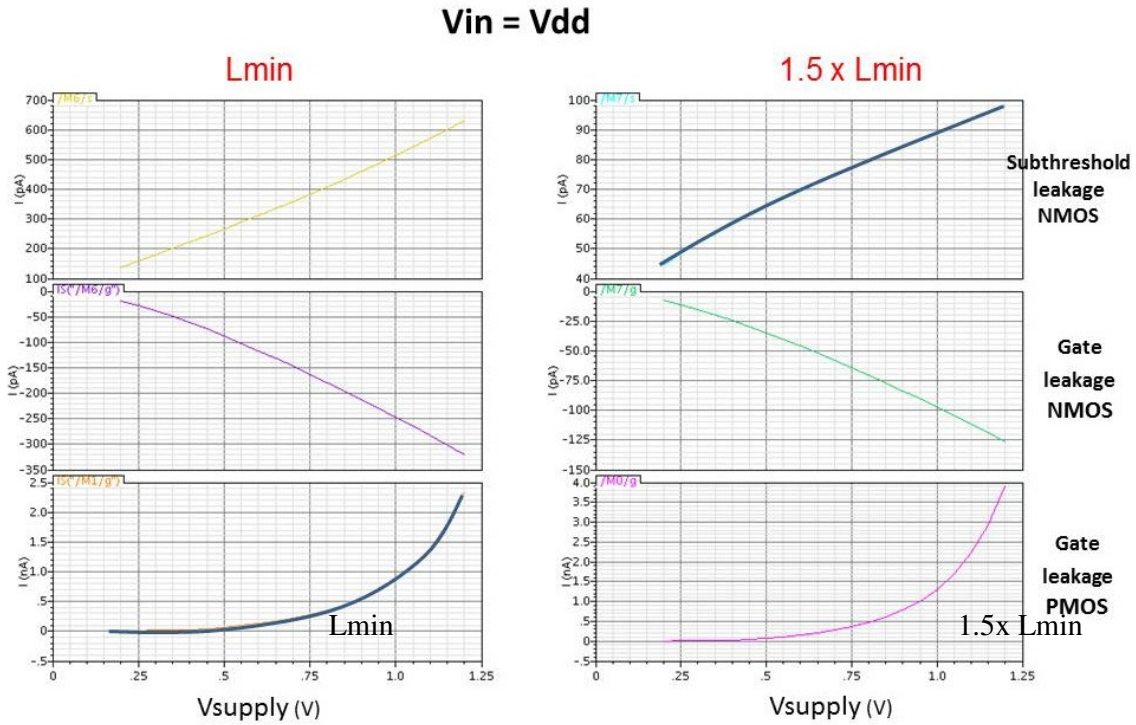


Figura 4.34 – Variação das componentes da corrente de fuga através dos inversores de W mínimo e entradas em Vdd com variação de Vdd de 0,2V a 1,2V

Como previsto anteriormente a largura de canal de 1,5x Lmin proporciona menor *subthresholds leakage* e assim torna-se mais apropriado para a utilização em células de memória. A redução do *leakage* que circularia pelas *bitlines* torna o projeto mais robusto e em contrapartida temos um incremento da corrente de *gate* (aumento de área) que não é expressivo percentualmente.

As avaliações do *leakage* na matriz de memória levaram justamente a utilização de uma célula de memória de Wmin = 120nm e L=90nm (1.5 x Lmin), demonstrando que as técnicas apresentadas nos capítulos de redução de energia estática e dinâmica consumidas são relevantes e devem ser aplicadas com o fim de obter circuitos ajustados para tecnologias nanométricas.

5 ASSOCIAÇÕES DE TRANSISTORES APLICADAS A PROJETO ANALÓGICO CMOS EM TECNOLOGIAS UDSM COMO SOLUÇÃO PARA REDUÇÃO DE LEAKAGE E AREA

O caminho para tecnologias UDSM fornece ferramentas principalmente para circuitos digitais obterem maior desempenho, ao mesmo tempo criaram a necessidade de circuitos analógicos gerarem compensações para que os digitais funcionem adequadamente (ajuste de frequência de relógio com variação de Vdd, de tensão de operação, etc). Por outro lado, o projeto de circuitos analógicos não obteve as mesmas vantagens de redução de área aplicada a circuitos digitais, entretanto manteve a quantidade de área aproximadamente idêntica devido justamente a especificações de casamento e ruído. As dimensões necessárias para que os circuitos analógicos funcionem adequadamente mantiveram-se aproximadamente iguais para que se alcance os requerimentos cada vez mais altos.

Nesse ínterim do desenvolvimento de circuitos analógicos para tecnologias UDSM ao avançar-se a pesquisa de TST´s (*T-Shaped Transistors*) visualiza-se um dispositivo com as características necessárias para projeto de baixa potência. TSTs / TATs (*Trapezoidal Associations of Transistors*) são estruturas *self-cascode* que podem tornar-se uma boa escolha por apresentar redução do *leakage*, redução na área utilizada e com incremento na regularidade do layout e no casamento entre transistores, propriedade importantíssima para circuitos analógicos. Através de revisão do estado da arte sabe-se que a corrente de *gate* tem um grande impacto no ruído, e cria dificuldades de projeto de integradores e circuitos de *sample and hold* (SANCHEZ, 2008).

Segundo Gielen (2005) e Nauta (2005) circuitos analógicos em tecnologias UDSM sofrem mais com a variação da dopagem, variação dos Vths, do L efetivo, redução da transcondutância gm, etc. O projeto de circuitos analógicos não tem reduzido em área, pois o ruído e o *mismatch* impedem esta redução (GIELEN, 2005), demonstrando o ponto de inserção dos TSTs como possibilidade desta quebra de paradigma.

Finalmente quando se avança para o projeto de circuitos em tecnologias UDSM obteremos menor corrente de fuga tanto pela porta, o *leakage* de *gate* é proporcional à área de gate do transistor. A redução do *subthresholds leakage* com a utilização de *cascades* ou *self cascodes* é comprovada no *paper* de Yan (2005) e Butzen (2007) por *stacked structures*. O segundo fator é a elevação do *matching* pela redução do *leakage* comprovado no *paper* de Annema (2005) e Nauta (2005), ainda mais obtêm-se um layout mais regular das estruturas (GIRARDI, 2003). Tudo isto contribui positivamente

para o aumento da confiabilidade e torna-se uma alternativa interessantíssima para o projeto de circuitos analógicos em tecnologias UDSM onde variações das propriedades físicas dos transistores MOS resultam das flutuações do processo intra-*die*.

Este capítulo visa reportar o modelo, a estrutura, os resultados obtidos e perspectivas no estudo das associações de MOSFETs tipo TST .

5.1 Chip Teste CMOS 180nm

O processo de fabricação moderno de CI´s necessita de modelos precisos que emulem o comportamento dos dispositivos básicos que os compõem, e neste contexto a caracterização destes dispositivos através de medidas elétricas é de fundamental importância. A comparação entre simulação e medida para transistores é primordial para a confiabilidade da funcionalidade de projetos desenvolvidos. Neste contexto normalmente projeta-se um circuito valendo-se de ferramentas de CAD apropriadas, cria-se o layout e posteriormente calcula-se por um modelo elétrico de extração automática as capacitâncias, indutâncias e resistências parasitas, valida-se o funcionamento elétrico do projeto utilizando o modelo enviado pela *foundry*. Todavia isto se torna insuficiente para comprovar a funcionalidade total de um circuito tanto digital como analógico, sendo o teste de real funcionalidade é feito após a fabricação. Sabe-se que os processos sofrem variações e que as mesmas são documentadas pela *foundry* e catalogadas como variações sobre o modelo típico (simulações monte carlo para variações locais) e simulações utilizando os modelos de *corners* (simulando as variações entre *dies* e entre áreas distintas de um mesmo *die*). A partir deste ponto de vista perguntemo-nos se estes dados fornecidos são realmente confiáveis, exprimem a funcionalidade do chip a ser fabricado? O único método realmente eficaz para aferição e controle dos modelos utilizados no chip produzido é a extração elétrica dos dados dos dispositivos diretamente, possibilitando o refinamento, avanço e documentação das técnicas utilizadas no circuito desenvolvido e que se tornarão fonte de informações valiosas para próximos desenvolvimentos. Para validar as metodologias de projeto empregadas no GME (Grupo de Microeletrônica da UFRGS), por exemplo a metodologia gm/Id e o estudo de TSTs / TATs, foi concebido um chip teste com diversos blocos RF e estruturas de medida. Este chip tem a função de validar a metodologia de projeto dos blocos implementados comparando-os com suas especificações e simulações. Neste teste chip os blocos RF são acessíveis via encapsulamento e os dispositivos de teste acessíveis via micropads. O chip foi prototipado na tecnologia IBM 0.18 µm CMOS por meio do serviço de fabricação *Multi-Project Wafer* (MPW) da MOSIS.

Foram executadas medidas de *leakage* nos transistores prototipados para verificação das características não modeladas pelo modelo BSIM3V3 fornecido pela *foundry*. Estas medidas tem por finalidade exemplificar se há necessidade ou não da avaliação destas características não previstas durante a fase de projeto, e que vem de encontro com o cerne desta dissertação. Finalmente a extensão do estudo dos TSTs / TATs, com a intenção de validar o conceito no tocante a sua aplicabilidade para projeto de baixa potência em tecnologias UDSM.

5.1.1 Visão Geral do Teste Chip

O chip teste é composto por blocos RF acessíveis via conexões ligadas a "*pads*" e também estruturas de teste situadas no centro do chip acessíveis via micropads. Foi prototipado na tecnologia IBM 0.18 µm CMOS por meio do serviço de fabricação

Multi-Project Wafer (MPW) da MOSIS. Foi utilizado um encapsulamento do tipo QFN 64 pinos. O serviço de fabricação via MPW retornou a UFRGS 10 amostras encapsuladas e 5 não encapsuladas para as medições e testes.

Como mostrado na figura 5.1, o chip possui uma área total de 6.8 mm², incluindo os 64 pads do *I/O ring* para conexão externa. A figura 5.2 apresenta uma fotografia ampliada do chip teste prototipado. A maior parte da área do chip é ocupada por micro-pads que dão acesso aos terminais dos dispositivos para a realização de medidas de caracterização e por 4 indutores integrados para a função de misturador de RF (CORTES, 2008). No projeto dos módulos RF incluídos no chip-teste, foi utilizada a metodologia baseada nas características gm/ID, apresentada em trabalhos anteriores (CORTES, 2003). Ainda nesse chip, vários circuitos analógicos (CORTES, 2008) (PAULA, 2007), assim como outras estruturas e circuitos de teste também foram implementadas. A figura 5.1 mostra o *layout* do chip teste visualizado na ferramenta Cadence Virtuoso®.

Block	Area (mm²)
Up Mixer FI=1.4GHz	0.7100
Down Mixer 1.4GHz to 40MHz	0.0210
VGA (Variable Gain Amplifier)	0.0350
RF Front-End	0.7700
VCO (Voltage Controlled Oscillator)	0.0220
Ring Oscillator	0.0073
Test Structures	1.2480
Pad Ring	1.7600
Unused Space	2.2260
Total Area	6.8000

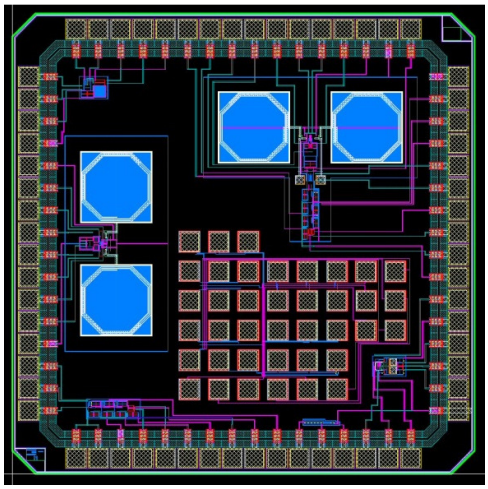


Figura 5.1 – Layout e área dos blocos prototipados no teste chip

O trabalho apresentado nas próximas seções concentra-se na medida das características DC e corrente de fuga dos transistores, incluindo associações de transistores (TAT´s).

Foram projetadas várias estruturas de teste que possibilitam a extração das características em todas as regiões de operação e para os diversos tamanhos de transistores prototipados. As estruturas de teste contem transistores de canais longo, largo, curto e estreito; inclui-se também novas geometrias como associações serie paralelo de transistores (TATs ou matriz de transistores *"self-cascode"*) de diferentes aspectos. Estas estruturas foram utilizadas para caracterizar corrente de dreno, transcondutância, condutância de saída, ruído, casamento e frequência intrínseca em termos de geometria das associações de transistores unitários e estratégias de *layout*. As estruturas foram medidas utilizando uma estação de microprovadoras, utilizando micropads para ter acesso aos terminais das estruturas. A figura 5.3A mostra o layout e a figura 5.3B apresenta a configuração do micropads para acesso dos terminais das estruturas de teste.

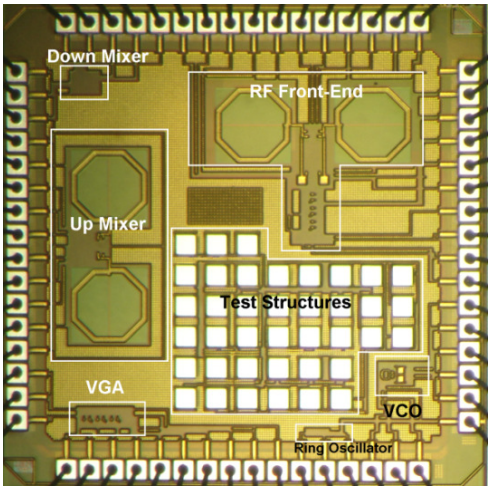


Figura 5.2 – Microfotografia do teste chip prototipado

5.1.2 Estruturas de Teste para Caracterização de Transistores e Arranjos de Transistores

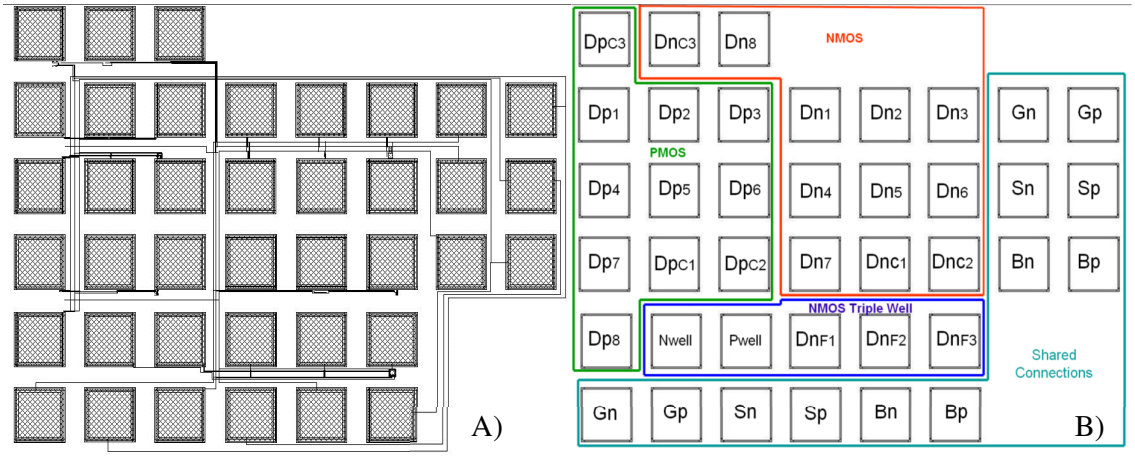


Figura 5.3 – A)Layout dos micropads e B) Configuração dos micropads de acesso

Na tabela 5.1 apresenta-se a lista das estruturas prototipadas com as dimensões físicas de cada uma delas. Contendo alguns transistores compostos por diferentes associações série-paralelo, estas compostas pela associação de transistores unitários também indicados na tabela. Estas associações foram selecionadas a partir dos dados obtidos no chip teste AMS 0,35µm prototipado em 2004 em trabalho prévio do grupo de pesquisa (GIRARDI, 2005). Os TATs são compostos por transistores de mesmo W e L, emulando um transistor comparável ao transistor de referência (Tref na tabela 5.1). Na matriz de transistores uma P e outra N temos os *gates*, fontes e *bulks* conectados de forma comum, ou seja, um contato de *gate*, outro de fonte e outro de *bulk* para toda a matriz. A individualização dos transistores se dá pelo contato individual de dreno o que possibilita a caracterização do mesmo.

Tabela 5.1 – Estruturas a serem medidas e pads de acesso

Tipo Transistor	Nome	Contato de Dreno	W1 (µm)	L1 (µm)	OBS
NMOS	N01	Dn1	0.22	0.18	Transistor Mínimo 180nm
NMOS	N02	Dn2	0.60	0.18	

NMOS	N03	Dn3	2.00	0.18	Transistor Unitário NMOS
NMOS	N04	Dn4	1.00	1.00	
NMOS	N05	Dn5	0.22	10.0	
NMOS	N06	Dn6	10.0	10.0	
NMOS	N07	Dn7	10.0	1.08	Tref N
NMOS	N08	Dn8	0.40	0.18	
NMOS	NC1	DnC1	10 (Weq)	1.08 (Leq)	TAT
NMOS	NC2	DnC2	10 (Weq)	1.08 (Leq)	TAT
NMOS	NC3	DnC3	10 (Weq)	1.08 (Leq)	TAT
PMOS	P01	Dp1	0.22	0.18	Transistor Mínimo 180nm
PMOS	P02	Dp2	0.60	0.18	
PMOS	P03	Dp3	2.00	0.18	Transistor Unitário PMOS
PMOS	P04	Dp4	1.00	1.00	
PMOS	P05	Dp5	0.22	10.0	
PMOS	P06	Dp6	10.0	10.0	
PMOS	P07	Dp7	10.0	1.08	Tref P
PMOS	P08	Dp8	0.40	0.18	
PMOS	PC1	DpC1	10 (Weq)	1.08 (Leq)	TAT
PMOS	PC2	DpC2	10 (Weq)	1.08 (Leq)	TAT
PMOS	PC3	DpC3	10 (Weq)	1.08 (Leq)	TAT
NMOS	NF1	DnF1	0.40	0.18	Triple Well Transistor
NMOS	NF2	DnF2	1.00	1.00	Triple Well Transistor
NMOS	NF3	DnF3	10.0	10.0	Triple Well Transistor

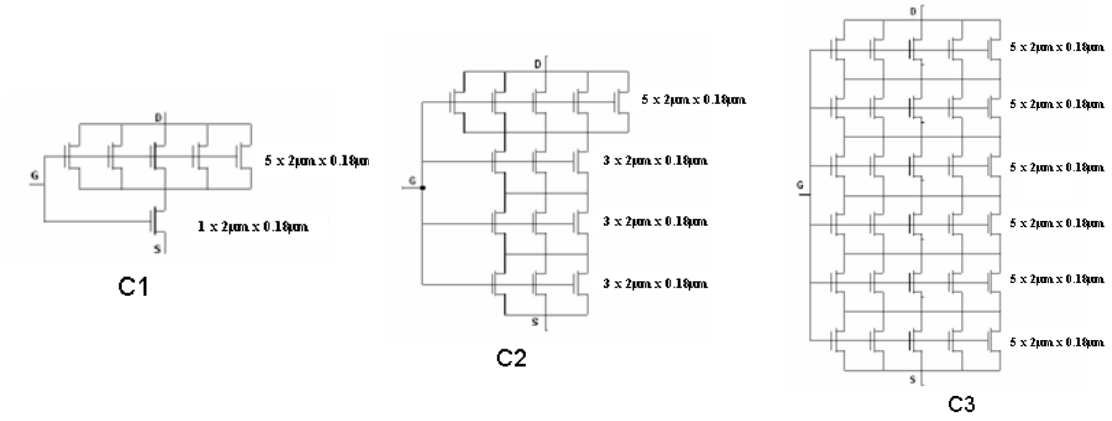


Figura 5.4 – Associações de transistores formato T prototipadas.

Os esquemáticos das associações C1 a C3 estão demonstrados na figura 5.4, tendo Weq = 10 µm e Leq = 1.08 µm, sendo Leq = 6 Lmin nas associações e no Transistor referência. A tabela 5.2 apresenta a lista das associações prototipadas e os tamanhos dos transistores inclusos nas associações, e também a área dos *gates* dos transistores. Sabidamente a operação analógica destas associações apresentará alguns efeitos de canal curto.

Tabela 5.2 – Associações de transistores de mesmo aspecto (usando a aproximação de primeira ordem) nas diferentes configurações.

Name	ND	NS	NL	W1	L1	W2	L2	Weq x Leq (µm)	Área de Gate (μm^2)
C1	5	1	1	5x 2µm	0.18 µm	1x 2µm	0.18 µm	10 x 1.08	2,16
C2	5	3	3	5x 2µm	0.18 µm	3x 2µm	3x 0.18 µm	10 x 1.08	5,04
C3	5	5	5	5x 2µm	0.18 µm	5x 2µm	5x 0.18 µm	10 x 1.08	10,8
Tref				10µm	1.08µm	-	-	10 x 1.08	10,8

Os transistores *T-shape* são uma alternativa para o projeto de circuitos integrados analógicos, incluindo uma nova variável de projeto para cada transistor. A possibilidade

de alterar os transistores unitários e escolher diferentes tamanhos é uma boa estratégia para obter melhora no desempenho do circuito. As maiores vantagens dos transistores compostos é a sua baixa condutância de saída, comparado ao Transistor de referência para o mesmo ponto de *bias*. É possível obter maiores Tensões de Early com transistores de menor área ativa, novamente comparando a associação ao transistor referência.

Finalmente quando se avança para o projeto de circuitos em tecnologias UDSM obteremos menor corrente de fuga tanto pela porta, pela menor área utilizada, bem como *subthresholds leakage* pelo canal. Obtêm-se ainda um layout mais regular das estruturas. Tudo isto contribui positivamente para o aumento da confiabilidade e torna-se uma alternativa interessantíssima para o projeto de circuitos analógicos em tecnologias UDSM onde variações das propriedades físicas dos transistores MOS resultam das flutuações de processo intra-*die*.

5.2 Associação Trapezoidal de transistores

Apresenta-se a seguir um breve resumo sobre o equacionamento e modelagem desenvolvido para os TSTs (GIRARDI, 2007), seguido de novos resultados obtidos, bem como novos pontos a serem explorados. As associações de transistores de formato T são caracterizadas pelo arranjo de transistores unitários em um formato trapezoidal, com o lado do dreno maior que o da fonte. A influência de campos elétricos mais intensos ocorre no dreno, composto por diversos transistores unitários ou por um transistor mais de canal largo que comportem a corrente requerida. Na associação de transistores o comprimento de canal é mantido o mínimo permitido pela tecnologia utilizada, por que não é necessário alto ganho em dispositivos em serie. Já o terminal de fonte possui transistores de comprimento de canal mais longo ou vários transistores mínimos em série para obter o comprimento de canal desejado, claro há uma limitação prática a ser considerada no número de transistores em série. Nomearemos o transistor superior de MD (lado do dreno) e os transistores em serie restantes como MS (lado da fonte). Finalmente sendo os gates unidos, de MD e MS, e considerando que MS tem comprimento de canal LS maior que LD o formato "T" estará caracterizado.

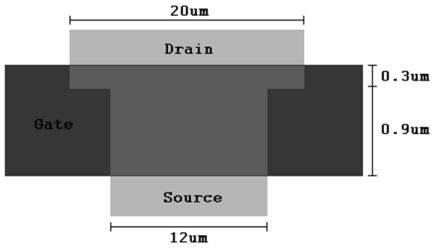


Figura 5.5 – *Layout* de um transistor formato-T intrínseco

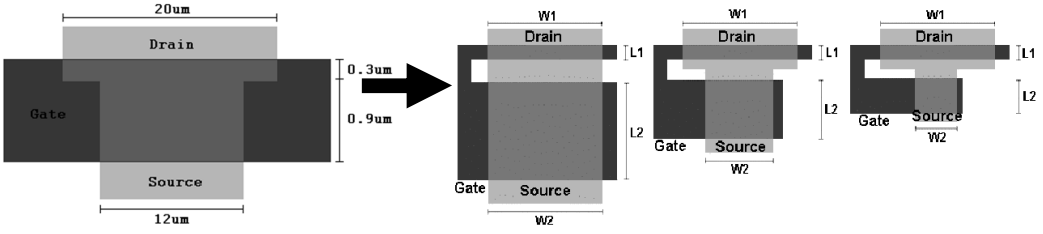


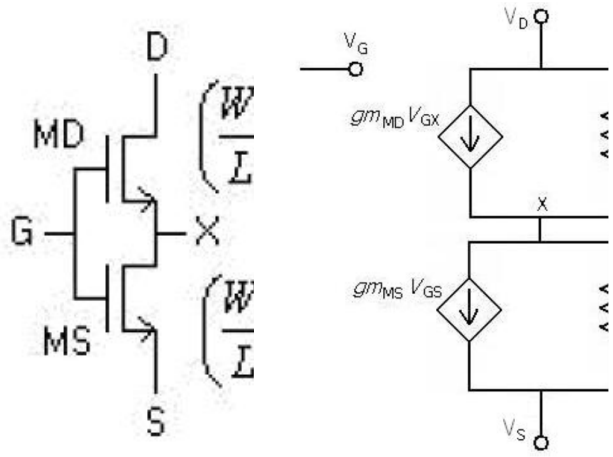
Figura 5.6 – Demonstração de layout economicamente viável de três configurações Formato T de mesmo aspecto W/L

A figura 5.5 mostra o layout de um transistor formato T customizado com o terminal de dreno mais largo que o de fonte. Este layout não é viável economicamente nem tecnicamente, pois desperdiça muita área de óxido de *gate* abaixo do estreitamento de fonte. Acrescido a isto o estreitamento do canal é uma péssima característica pois dificulta o fluxo de corrente no transistor MD.

A figura 5.6 exemplifica como decompor um Transistor formato T em um layout composto de transistores unitários, exibindo diferentes opções com mesmo aspecto W/L. Uma associação pode ser dita equivalente a um transistor referencia quando algumas características são idênticas em uma região de polarização. Quando uma associação tem, ou aproxima-se de ter, a mesma corrente de dreno para uma mesma tensão de *gate,* Vgs=1.8V e polarização diz-se que são equivalentes. A margem relativa do erro entre as correntes de dreno da associação e do transistor referencia definirão a equivalência. Usando o modelo ACM *(Advanced Compact MOSFET)* (CUNHA, 1998), pode-se deduzir equações que exprimam as características das associações formato T. Para transistores de canal longo, considerando apenas dois transistores em série (figura 5.7) e igualando as correntes de dreno que fluem quase estaticamente, tem-se:

$$\left(\frac{W}{L}\right)_{eq} = \frac{ND \cdot W_{UN(MD)}}{L_{UN(MD)} + \frac{W_{UN(MD)}}{W_{UN(MS)}} \cdot L_{UN(MS)} \cdot \frac{ND}{NS}} \qquad \text{(eq. 5.1)}$$

Sendo Wun, Lun são e largura e o comprimento do canal dos transistores unitários, ND e NS são o número de transistores unitários em paralelo em MD e MS, respectivamente.



A) B)

Figura 5.7 – (A) esquemático do TST e (B) modelo de pequenos sinais de um TST

5.2.1 Associação Trapezoidal de Transistores

O modelo de pequenos sinais de baixa frequência é importante para o calculo das características elétricas perto de um ponto de operação, e no caso do TST pode-se assumir um modelo de pequenos sinais da associação é composto por dois transistores em série. O efeito de corpo será ignorado nesta análise por não ser muito relevante, e desta forma obter-se-á equações mais intuitivas. A figura 5.7B mostra o modelo DC de pequenos sinais para o transistor composto (TSIVIDIS, 1999), com alguma

manipulação algébrica pode-se estimar as condutâncias e transcondutâncias equivalentes. A transcondutância de *gate* do transistor composto $gm_{TST} = \partial I_D / \partial V_{GS}$ é dada por:

$$gm_{TST} = \frac{gds_{MS} \cdot gm_{MD} + gm_{MS} \cdot gm_{MD} + gm_{MS} \cdot gds_{MD}}{gds_{MS} + gm_{MD} + gds_{MD}} \qquad \text{(eq. 5.2)}$$

De acordo com Choi (2001), a transcondutância de gate de um TST é menor que de um transistor comum em inversão forte, assim se considerarmos $gm_{MD(MS)} + gds_{MD(MS)}$, a equação 5.2 pode ser aproximada por:

$$gm_{TST} \approx gm_{MS} \qquad \text{(eq. 5.3)}$$

Esta aproximação só tem validade quando o transistor MD tem W/L muito maior que o do transistor MS, pode-se comparar este efeito com Veeravalli (2001), onde MS trabalhando na região linear age como um resistor controlado por tensão de *gate.* Entretanto na pratica, a transcondutância do TST será maior que a de MS e menor que a de MD.

$$gm_{MD} > gm_{TST} > gm_{MS} \qquad \text{(eq. 5.4)}$$

A redução do gm impacta na resposta de pequenos sinais, entretanto é compensada pelo aumento da resistência de saída DC. A condutância de do TST ($gds_{TST} = \partial I_D / \partial V_{DS}$) pode ser estimada através do modelo da figura 5.7B, sendo:

$$gds_{TST} = \frac{gds_{MD} \cdot gds_{MS}}{gm_{MD} - gds_{MS} - gds_{MD}} \qquad \text{(eq. 5.5)}$$

A condutância de saída é menor que a de um transistor de mesmo W/L em todas as regiões de operação, sendo essa uma das maiores vantagens dos TST que proporciona incremento em estágios de ganho de amplificadores podendo ocupar menor área.

5.2.2 Análise da Tensão de Saturação

A tensão no nó interno entre MD e MS de um TST, nó X (figura 5.7B), é geralmente a variável desconhecida para o projetista analógico; sendo relevante somente para o cálculo da corrente de dreno, transcondutância e transcapacitâncias da associação. Em inversão forte pode ser aproximada (ENZ, 1996) por:

$$V_X = \left[1 - \frac{1}{\sqrt{1 + \left(W/L\right)_{MD} \Big/ \left(W/L\right)_{MS}}} \right] Vp \qquad \text{(eq. 5.6)}$$

Vp é a tensão de "*pinch-off*" do canal do MOSFET. A tensão Vp, referida ao terminal de corpo do transistor é função da tensão de dreno-*bulk* e *source-bulk.* Em inversão fraca pode ser aproximada por:

$$V_X \cong \phi_t \cdot \ln\left(1 + \left(W/L\right)_{MD} \Big/ \left(W/L\right)_{MS} \right) \qquad \text{(eq. 5.7)}$$

A variação de tensão Vp da tensão Vx, portanto, nos terminais de um TST é quase linear com V_{GS}, especialmente em inversão forte e é constante com respeito a V_{DS} quando M_D está saturado. Φt é a *thermal voltage*. Outra importante característica dos TSTs é a tensão de saturação mais baixa quando comparada a um *cascode*. Se $(W/L)_{MD} > (W/L)_{MS}$ o transistor Md opera saturado e o MS em zona linear, assim a tensão entre dreno e fonte do transistor MS é menor e equivalente a tensão de saturação de um transistor composto e proporcional a tensão de saturação de MD. Assim considerando $V_{SB\_TST} = 0$ tem-se:

$$V_X \cong V_{DsatMD} + V_{DB_{MS}} \qquad \text{(eq. 5.8)}$$

A propriedade de um transistor composto faz com que estas estruturas *self-cascode* sejam ideais para aplicações de baixa potência substituindo as estruturas *cascode* convencionais, pois estas tem uma tensão de saturação mais alta (RAJPUT, 2002). A tensão de saturação de um TST é proporcional à tensão de *pinch-off*, que é idêntica para MD e MS, sendo proporcional somente a VG. Assim $V_{GS\_MD} < V_{GS\_MS}$, $V_{Dsat\_MS} < V_{DS\_MS}$ em uma análise de primeira ordem pode-se estimar $V_{Dsat\_MS} = V_{GS\_TST} - V_T$, em outras palavras, idêntica a de um transistor comum. Alguns trabalhos tem demonstrado a vantagem da utilização de *self-cascodes* em aplicações de baixa tensão e potência operando em inversão fraca e moderada principalmente, em espelhos de corrente com alta impedância de saída (CAMACHO-GALEANO, 2005).

A versão *cascode* de um espelho de corrente tem por objetivo prover menor condutância de saída e consequentemente menor diferença entre a cópia e a referência [Lee 2005], entretanto a menor tensão de saída é limitada a $V_{(out)\min} = V_T + 2 \cdot V_{on}$, sendo Von a quantidade de VGS que excede Vth. Obviamente $V_{out(min)}$ pode ser reduzido com o incremento da relação W/L dos transistores e do ajuste da relação de tensão *gate*-fonte. Entretanto utilizando TSTs pode-se obter uma melhor excursão mínima com menor área utilizada. Esta melhor excursão pode ser obtida porque para a versão TST temos $V_{out(\min)} = V_T + V_{on}$, idêntico $V_{out(min)}$ a versão com transistores simples. A vantagem de utilização da versão TSTs é que esta propicia uma menor condutância de dreno.

5.3 Medidas Realizadas

Com o intuito de executar a verificação dos transistores prototipados foram executadas medidas de caracterização DC, sendo elas: -Id X Vd, Id X Vg e Id X Vs. Estas medidas tem por objetivo a caracterização de diversas características dos transistores, bem como possibilitam a extração do modelo ACM. Em conjunto com essa parte da caracterização foram efetuadas medidas de *leakage* nos transistores com o objetivo de determinar a necessidade de sua avaliação ou não em projetos na tecnologia 0.18 µm IBM CMOS.

5.3.1 Medidas

Como citado anteriormente as medidas realizadas para a caracterização DC dos transistores foram Id X Vd, Id X Vg e Id X Vs, conforme demonstradas na figura 5.8 que exemplificam as configurações utilizadas para execução das mesmas.

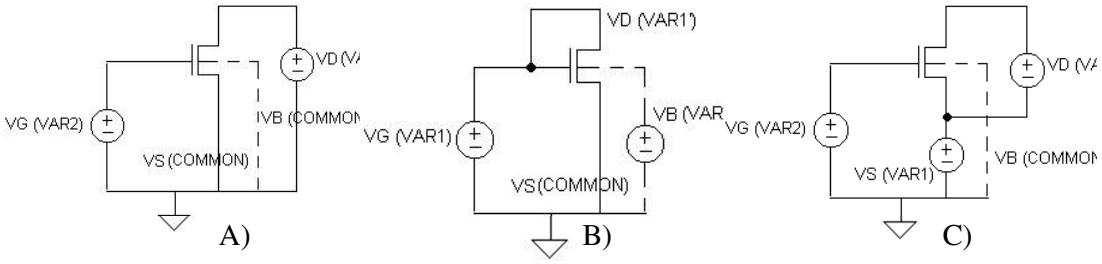


Figura 5.8 – Configuração para medida (A) de Id x Vd, (B) de Id x Vg e (C) de Id x Vs

Também executaram-se medidas para avaliação do *leakage* nos transistores, sendo que avaliaram-se as condições de transistor de passagem, transistor de retenção de dado, e variação de Vth através da variação da tensão de *Bulk*. Nestas configurações foram salvos os dados de tensão e corrente dos quatro terminais do transistor proporcionando a possibilidade de rastreabilidade dos valores. Na figura 5.9 apresentam-se as configurações utilizadas nas medidas.

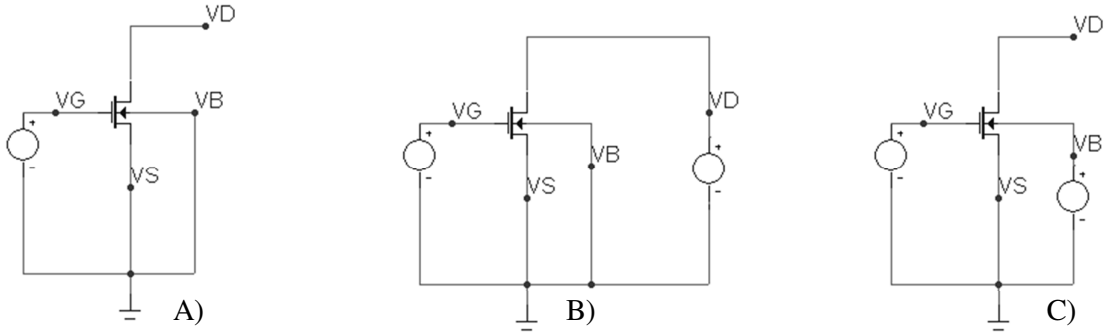


Figura 5.9 – Configuração para medida (A) transistor de retenção, (B) transistor de passagem e (C) Variação de Vth referida à variação da tensão de *Bulk* Vb.

As medidas de corrente de fuga, Ig e Is, foram prejudicadas pelo compartilhamento de terminais nas estruturas de teste. Com o compartilhamento dos terminais de *gate, source e bulk* fica impossível isolar somente um transistor, sendo a composição das contribuições destas dezenas de transistores visível nas medidas destes terminais.

5.3.2 Curvas I-V Medidas

A seguir mostraremos alguns dos gráficos obtidos das medições das estruturas no chip teste, sendo todas estas fabricadas com transistores *standard* Vth. A metodologia empregada é obter as curvas para canais curtos e longos demonstrando as diferenças entre estes. Demonstram-se também resultados de outras características obtidas através da manipulação dos dados em Matlab, sendo as planilhas de calculo desenvolvidas no decorrer das medidas. Todas as medidas referem-se a transistores standard Vth.

5.3.2.1 Id x Vd

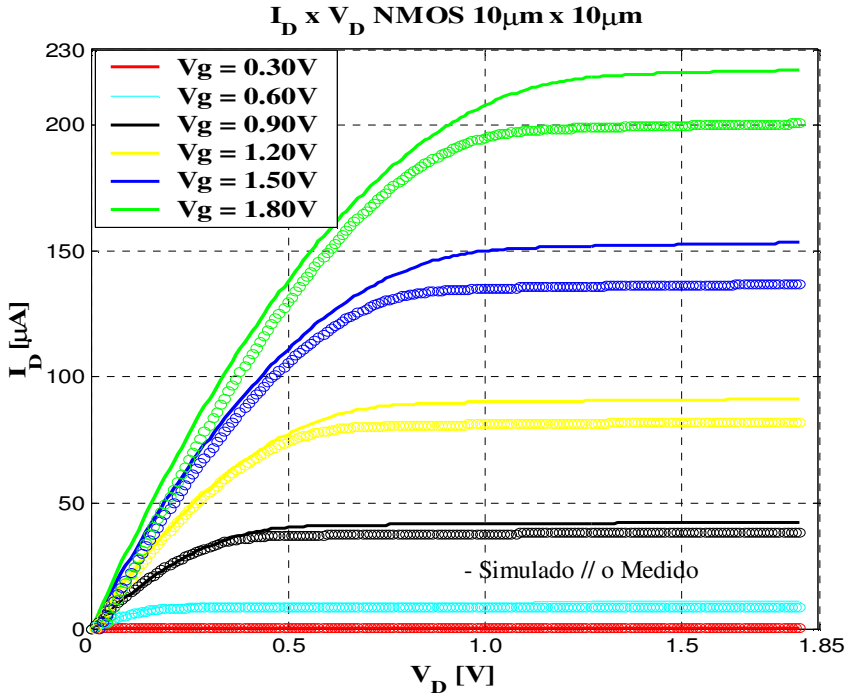


Figura 5.10 – Curvas medidas de Id x Vd para transistor NMOS 10µm x 10µm

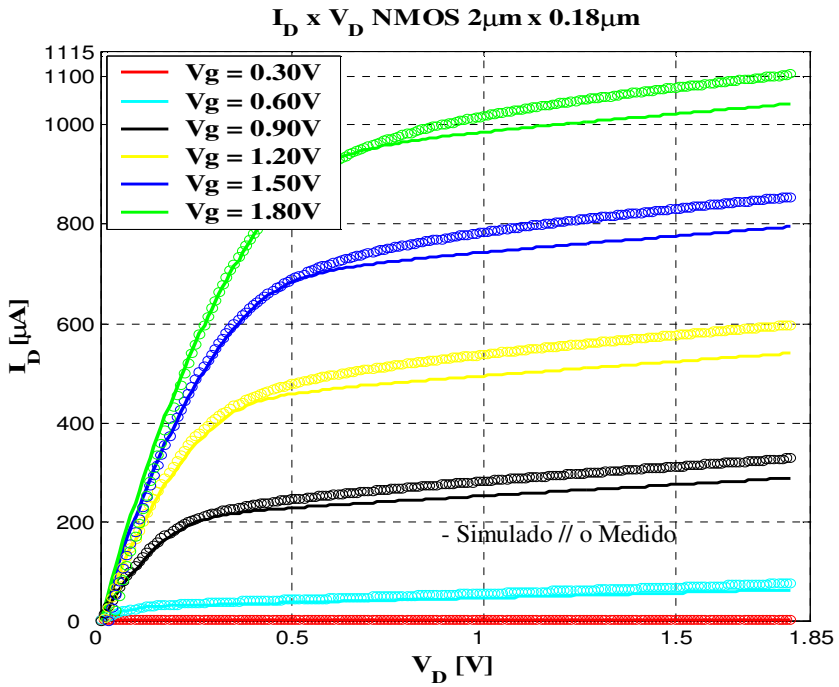


Figura 5.11 – Curvas medidas de Id x Vd para transistor NMOS 2µm x 0,18µm

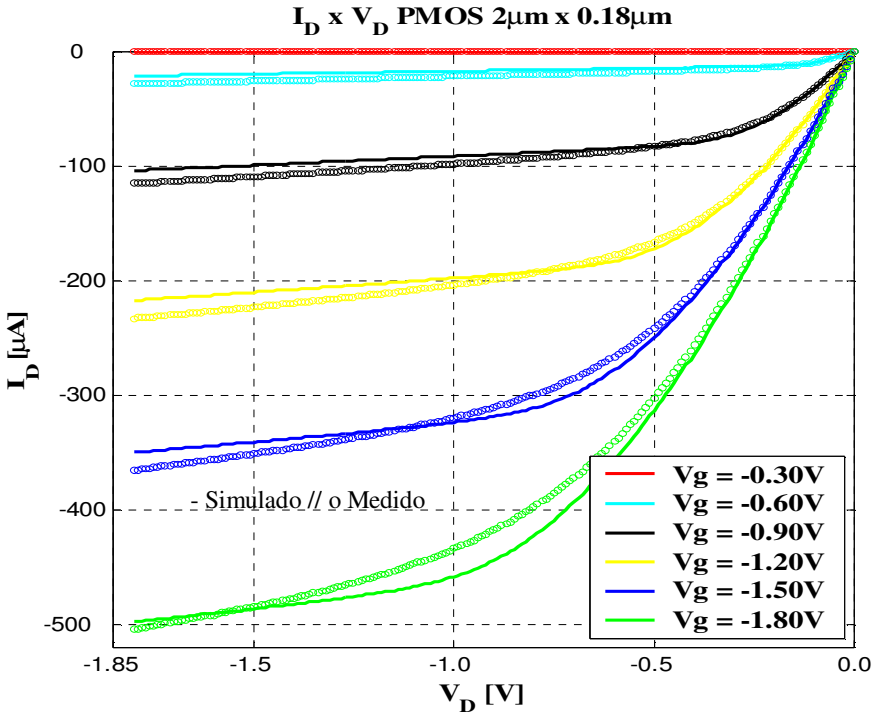


Figura 5.12 – Curvas medidas de Id x Vd para transistores PMOS 2µm x 0,18µm

5.3.2.2 Id x Vg

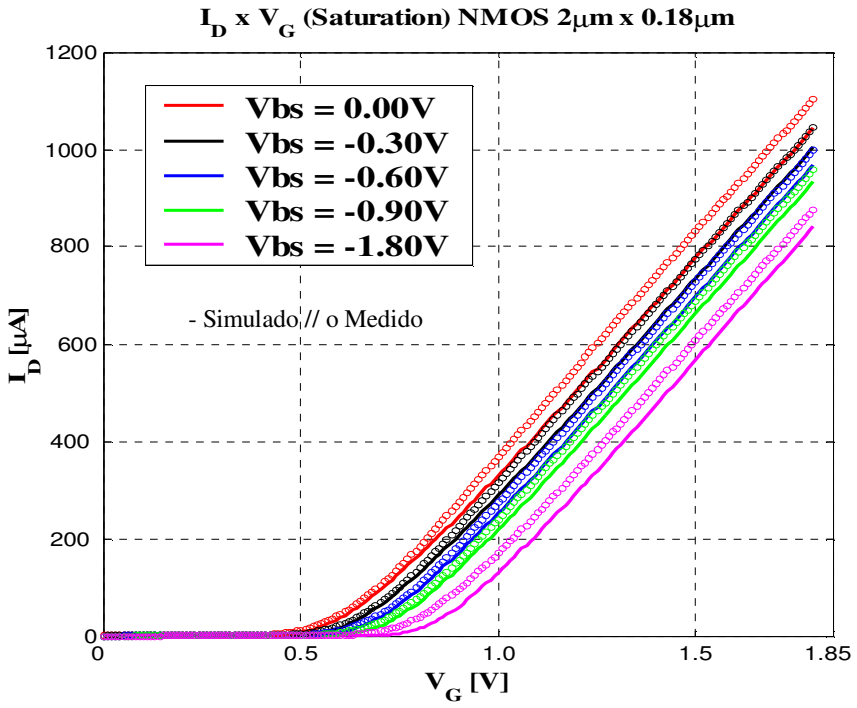


Figura 5.13 – Curvas medidas de Id x Vg para transistor NMOS 2µm x 0,18µm.

5.3.2.3 *Id x Vs*

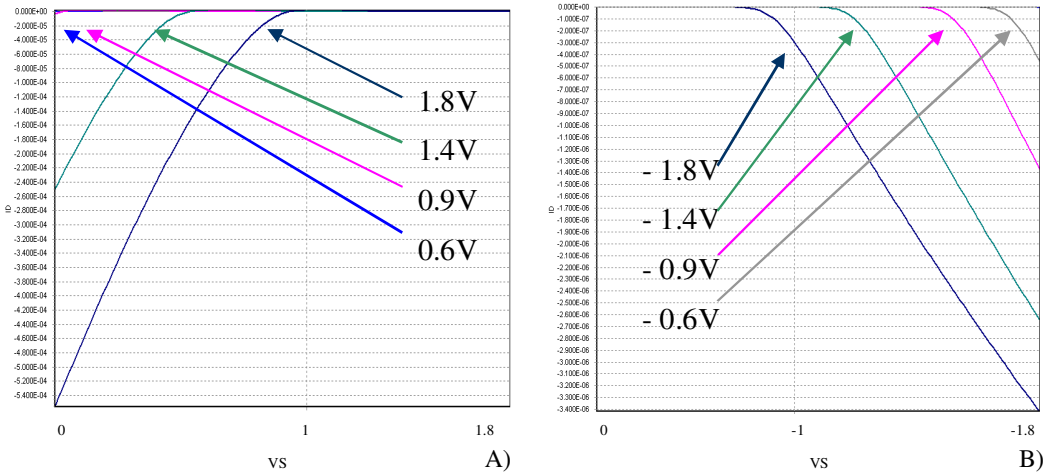


Figura 5.14 – Curvas medidas de Id x Vs para transistor A) NMOS 10µm x 0,18µm e B) PMOS 10µm x 10µm

5.3.2.4 *gm/Id*

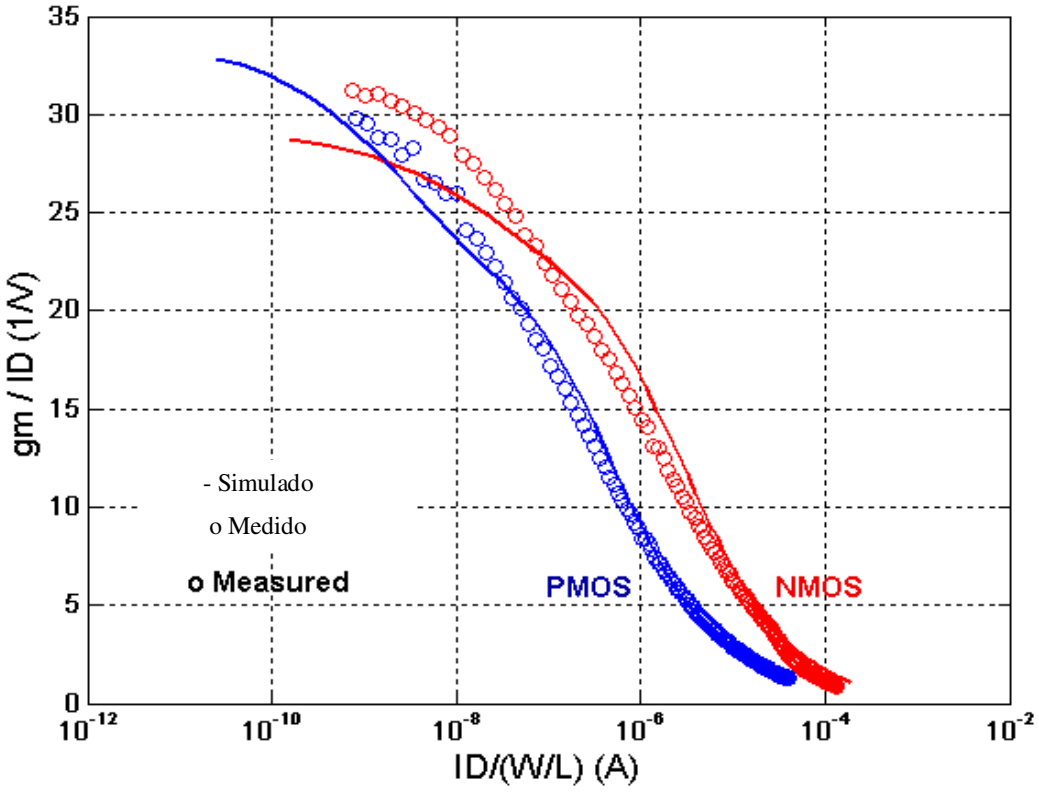


Figura 5.15 – Curvas medidas de gm/Id x Id normalizado para transistores prototipados

5.3.2.5 *Corrente de fuga*

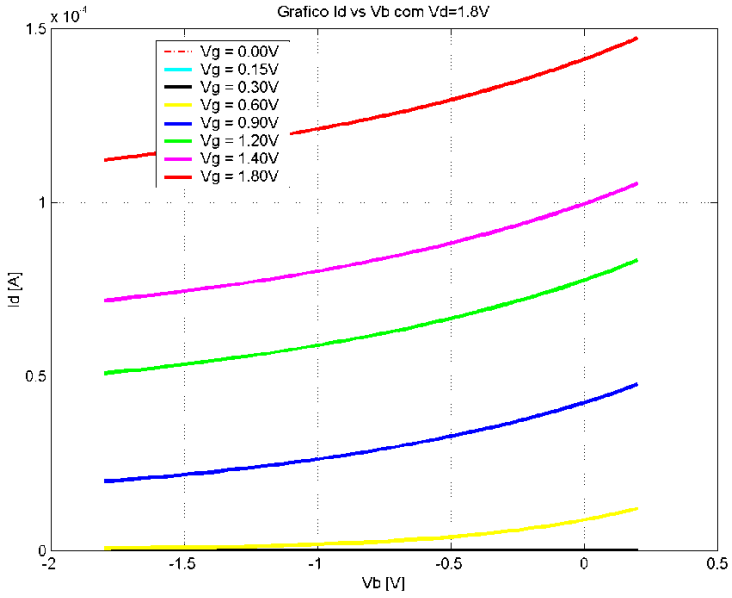


Figura 5.16 – Curvas medidas Id x Vb para transistor NMOS de W=0,22µm e L=0,18µm demonstrando o efeito da variação de Vth com V_{SB}

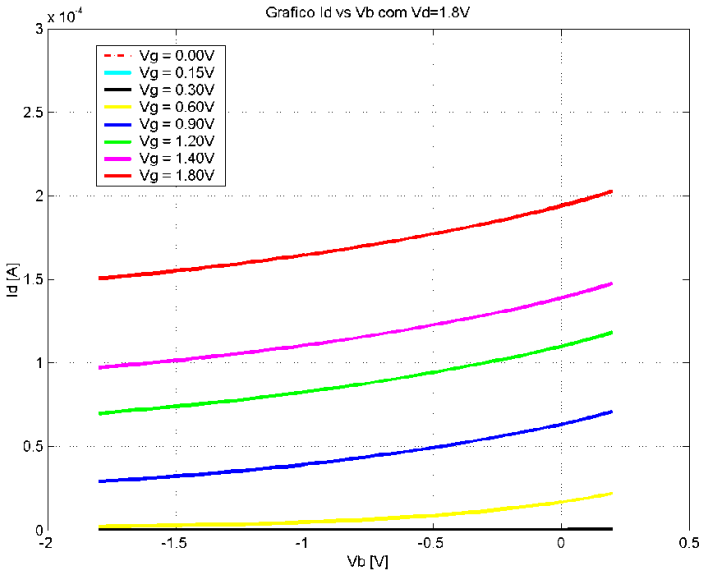


Figura 5.17 – Curvas medidas Id x Vb para transistor NMOS de W=0,40µm e L=0,18µm demonstrando o efeito da variação de Vth com V_{SB}

Avaliando-se as figuras 5.16 e 5.17 percebe-se o efeito de gmb pela variação de V_{BS} agindo no transistor variando a corrente I_D. Pode-se ressaltar a característica supra-linear de gmb vista através da característica I_D, diferentemente da característica gm que é exponencial. Utilizando-se V_{BS} como entrada de controle para o transistor proporciona-se maior linearidade na variação de Id dentro da faixa de operação, sendo esta técnica conhecida como "*bulk-driven*".

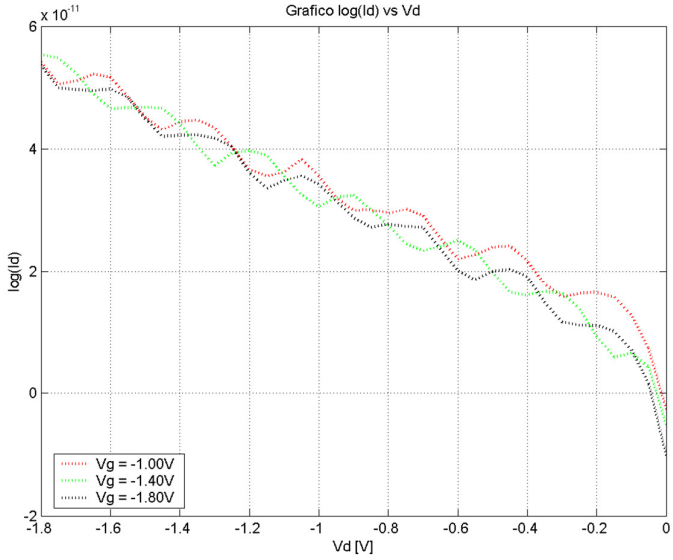


Figura 5.18 – Curvas Medidas log(Id) x Vd para transistor PMOS de W=0,22µm e L=0,18µm

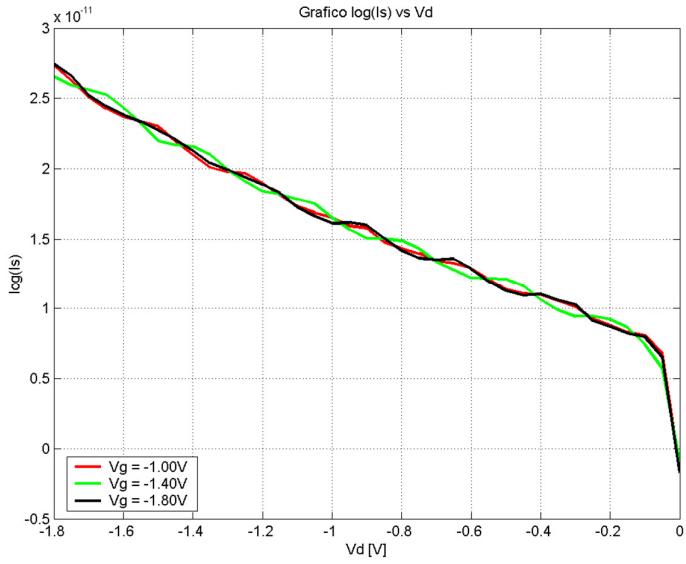


Figura 5.19 – Curvas Medidas log(Is) x Vd para transistor PMOS de W=0,22µm e L=0,18µm

Comparando-se as figuras 5.18 e 5.19 é visível que há algum mecanismo de fuga no transistor apresentado, nas medidas este mecanismo é visível em ambos os transistores PMOS e NMOS. O que dificulta a separação das componentes de fuga nestas medidas é a ligação comum de *sources*, *gates* e *bulks* de forma que torna-se impossível isolar o efeito apresentado por somente um transistor.

5.3.3 Comparações TATs versus Tref

Utilizando as estruturas de teste obtiveram-se as curvas de comparação entre as associações e o transistor simples equivalente prototipados, estas serão apresentadas a

seguir. Como anteriormente apresentado na seção 5.1.1 na tabela 5.2 temos a comparação entre os TATs de mesma razão de aspecto e na figura 5.4 vemos as diferentes configurações elétricas dos TATs.

As medidas da corrente de dreno para as três associações estão na figura 5.20 e pode-se perceber que a corrente de dreno das associações para um mesmo valor de VGS tem corrente degradada quando comparada ao transistor de referencia. Esta característica se deve aos efeitos de canal curto, como saturação de velocidade dos portadores, que é mais pronunciada nos transistores com largura mínima de porta (180nm). Como uma vantagem os TATs tem para uma mesma polarização uma condutância de saída menor como pode ser visto na figura 5.21. Outro fator importante é que a condutância de saída é proporcional a $[(W/L)_{MS}/(W/L)_{MD}]$, i.e., que é proporcional a assimetria da associação. Nota-se que para todas as associações o gds é menor em todas as regiões.

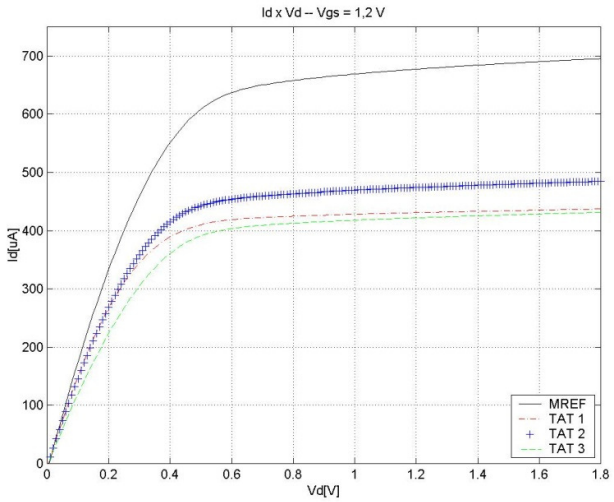


Figura 5.20 – Medida de corrente de dreno versus tensão de dreno

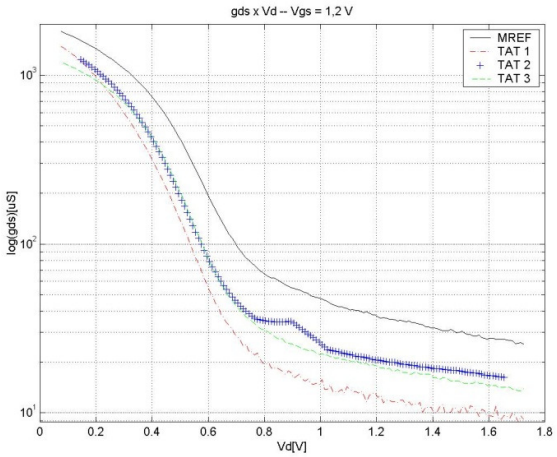


Figura 5.21 – Medida de condutância de saída versus tensão de dreno

Através da divisão da condutância de saída pela corrente de dreno (VA = gds/ID) obtêm-se a tensão de Early que indica a qualidade do dispositivo em região de

saturação. A figura 5.22 apresenta os resultados de medidas elétricas de VA versus VD (tensão de dreno), que demonstram um valor maior de tensão de Early para a associação TAT1(mais trapezoidal), e que as outras associações apresentam ainda maior VA que o do transistor equivalente. Isto se dá pelo incremento na largura do canal equivalente em associações trapezoidais.

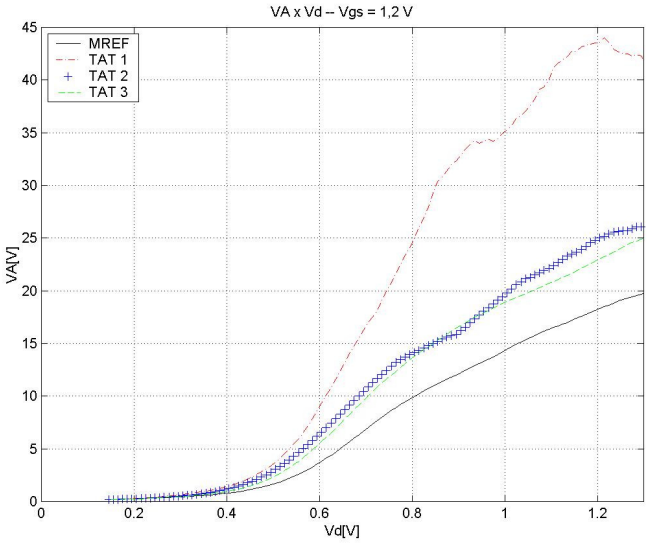


Figura 5.22 – Medida de tensão de Early versus tensão de dreno

Agora tratando da medida de transcondutância de *gate* a figura 5.23 apresenta a degradação dos TATs quando comparados ao transistor de referência. Novamente os efeitos de canal curto nos transistores unitários das associações contribuem para esse resultado, entretanto a figura mostra que o TAT2 tem um bom desempenho resultante de um correto dimensionamento do transistor MD que minimiza a degradação da transcondutância.

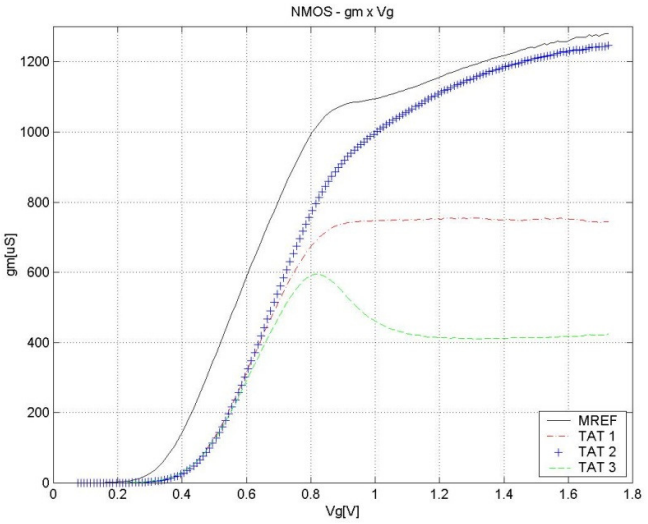


Figura 5.23 – Medida de transcondutância versus tensão de *gate* com VD = 1.8V

A figura 5.24 apresenta as curvas medidas de gm/ID versus ID/(W/L) que exprimem importantes características dos transistores medidos. Pode-se ver que em inversão fraca e moderada as associações tem gm/ID degradado, em inversão forte tem gm/ID similar ou maior que o transistor referencia. O mais importante no que se trata das associações é que para uma mesma região de inversão obtêm-se menor densidade de corrente o que é melhor quando tratamos de aplicações de baixa potência. Segundo em inversão forte temos praticamente a mesma operação demonstrando boas características do dispositivo a serem exploradas em novos projetos.

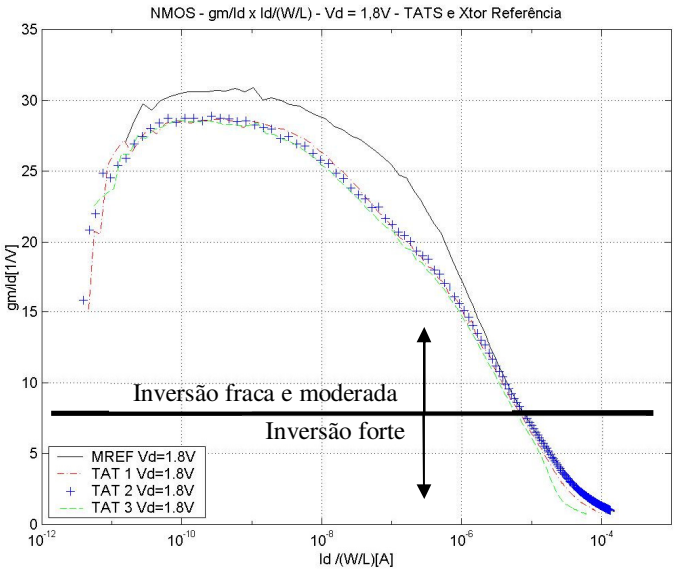
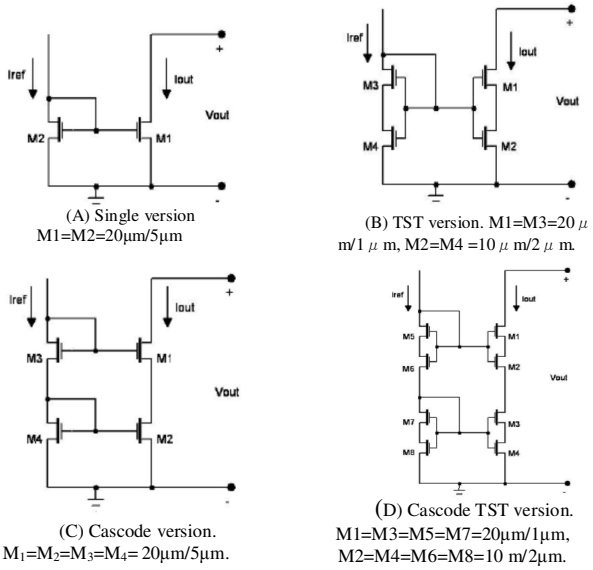


Figura 5.24 – Medida de [gm/ID] versus [ID/(W/L)] de um transistor NMOS

5.3.4 Espelhos de Corrente utilizando TSTs



(A) Single version M1=M2=20µm/5µm

(B) TST version. M1=M3=20 μ m/1 μ m, M2=M4 =10 μ m/2 μ m.

(C) Cascode version. $M_1=M_2=M_3=M_4=$ 20µm/5µm.

(D) Cascode TST version. M1=M3=M5=M7=20µm/1µm, M2=M4=M6=M8=10 m/2µm.

Figura 5.25 – Esquemático de 4 diferentes espelhos de corrente

Com o objetivo de visualizar o efeito de baixa saturação nos TSTs executaram-se simulações de espelhos de corrente em 4 diferentes versões, são elas: tradicional, TST, *cascode* e *cascode* TST. Os esquemáticos estão apresentados na figura 5.25 e na figura 5.26 apresentam-se as correntes de saída versus a tensão de saída para as quatro configurações utilizando uma corrente de referência de 50uA em tecnologia CMOS AMS 0,35µm. O efeito de redução da tensão de saturação nos TSTs é uma vantagem visível das associações comparadas com a configuração *cascode*. Na figura 5.26 o espelho TST apresenta uma tensão de saturação V_{Dsat1} menor que a das versões *cascode* C e D da figura 5.25. A versão *cascode* TST é menos viável, pois apresenta uma tensão de excursão menor mas mesmo assim apresenta resultados idênticos a versão *cascode*.

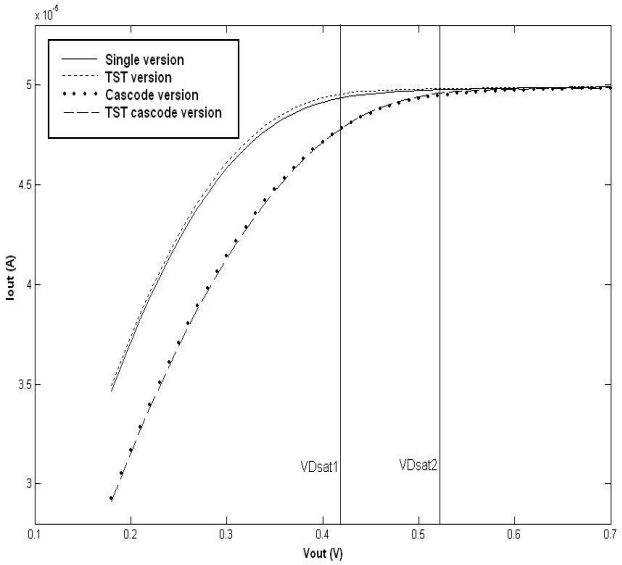


Figura 5.26 – Curvas Id x Vd simuladas para transistores de canal longo (GIRARDI, 2005)

5.3.5 Conclusões e Perspectivas Futuras

Neste capítulo apresentaram-se resultados de medidas e simulações elétricas de associações trapezoidais de transistores em tecnologias DSM (*deep sub-micron*). Sendo comprovada a redução da condutância de saída que proporciona fontes de corrente com melhor desempenho, a tensão de saturação é menor quando comparada a de uma configuração *cascode* convencional sendo esta uma característica importante para a utilização de TSTs em blocos de espelhos de corrente. A melhora na tensão de Early destes dispositivos demonstra que os mesmos apresentam menor efeito de corpo, emulando um transistor de canal mais longo. Ao mesmo tempo há um custo de redução de transcondutância de *gate*, que faz com que para se obter o mesmo ganho há de se dispender mais corrente. Desta forma com a redução da tensão de operação e os efeitos de canal curto cada vez mais notados os TSTs demonstram qualidades como uma alternativa para espelhar correntes utilizando tecnologias DSM CMOS.

As necessidades de casamento, ganho e ruído mantiveram, mesmo com o *scaling* de tecnologia, a área utilizada para os projetos analógicos quase que estável. Mesmo que algumas vezes consiga-se menor área no projeto em si, as soluções para sobrepujar as piores características dos transistores, estes sintonizados especialmente para o design digital, fazem com que a área dispendida seja praticamente a mesma. Ao mesmo tempo, há uma necessidade crescente por ferramentas de design analógico que executem o

projeto de forma rápida, precisa e assertiva de forma que o primeiro resultado obtido para um circuito seja garantido funcionalmente, reduzindo o tempo de trabalho. Fases de projeto semi-automotizado serão cada vez mais comuns em projetos analógicos, executando blocos construtivos menores de um sistema (ex: - Amplificadores operacionais de diversos, comparadores, etc) como GIRARDI (2007) previu e demonstrou. A possibilidade da execução do projeto destes blocos utilizando TSTs gera um grau a mais de liberdade ao projetista, pela escolha nos transistores unitários que compõem as associações do tamanho, largura, número, etc.

Ao mesmo tempo o design analógico vai se tornar mais complexo com o avanço do *scaling*. As variações estatísticas do processo fazem com que torne-se cada vez mais delicada a polarização e estabilidade dos circuitos sobre todos os *corners* de projeto. Tornou-se comum para transistores mínimos em sua tensão de *threshold* de aproximadamente 300mV com variações estatísticas de 50mV, comparado a *thresholds* de 700mV estas variações são muito mais pronunciadas em tecnologias UDSM que em tecnologias de canal mais longo, assim demonstra-se que será complicado obter circuitos estáveis com baixa área.

Nesse ínterim Nauta (2005) e Annema (2005) demonstram que com a redução da largura de canal do *gate* este torna-se predominantemente resistivo em baixa frequências o que resulta em maior consumo estático. A equação 5.7 calcula aproximadamente a frequência limite, sendo os valores de t_{ox} em [nm] e Vgs em [V].

$$\begin{aligned} f_{\text{gate}} &= \frac{g_{\text{tunnel}}}{2\pi C_{\text{in}}} \\ &\cong 1.5 \cdot 10^{16} \cdot v_{\text{GS}}^2 \cdot e^{t_{\text{ox}}(v_{\text{GS}}-13.6)} \quad \text{(NMOST)} \\ &\cong 0.5 \cdot 10^{16} \cdot v_{\text{GS}}^2 \cdot e^{t_{\text{ox}}(v_{\text{GS}}-13.6)} \quad \text{(PMOST)} \end{aligned} \qquad \text{(eq. 5.7)}$$

A figura 5.27 A) exprime graficamente a equação 5.7 para transistores NMOS demonstrando a divisão entre os domínios resistivos e capacitivos do *gate* em tecnologia 180nm através da frequência imposta ao dispositivo. A figura 5.27 B) mostra a variação desta função entre as tecnologias, demonstrando a relação inversa entre a largura mínima de canal da tecnologia e a frequência de separação entre domínio capacitivo ou resistivo na corrente de *gate* dos transistores.

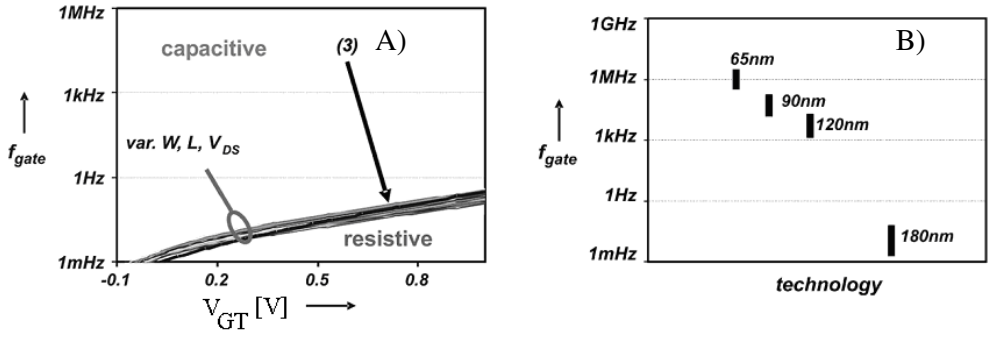


Figura 5.27 – A) f_{gate} como função da tensão de overdrive para diferentes transistores NMOS em tecnologia CMOS 180nm e B) Variações de f_{gate} em transistores NMOS de diferentes tecnologias CMOS para aplicações analógicas (ANNEMA, 2005)

O incremento da frequência limítrofe demonstra que cada vez mais a redução de área de *gate* proporcionará ganhos em economia de energia estática, pois reduzindo-se a área de *gate* minimiza-se *leakage* de *gate* e "*subthresholds leakage*" conjuntamente.

A redução de área de *gate* resulta também em menor ruído no dispositivo, o tunelamento de elétrons da porta para o canal resulta em ruído balístico, sendo que este é dependente da área. Se utilizarmos dispositivos com menor área de *gate* haverá menor tunelamento e consequente menor ruído intrínseco. Esta dependência cria mais um fator de interesse no estudo dos TSTs aplicados a projeto analógico buscando a redução de ruído gerado no dispositivo.

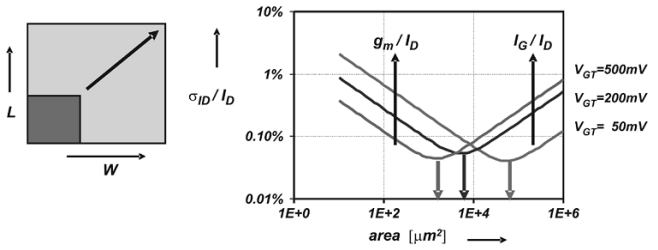


Figura 5.28 – Variação do *matching* em transistor NMOS 65nm com variação linear em W e L em função da área (NAUTA, 2005)

Também utilizando o modelo de Pelgron (ANEMMA, 2005) refere-se que o *matching* é afetado pelo *leakage* de *gate* sendo reduzido pelo aumento deste. Usualmente para aumentar o casamento necessita-se de aumento de área, em UDSM isto implica em aumento de consumo estático e aumento de capacitância parasita com consequente aumento da potência consumida. Nas tecnologias modernas necessita-se avaliar a redução de *matching* por incremento de *leakage* demonstrando-se que há uma área onde alcança-se um ponto ótimo de *matching*. Assim pelo *scaling*, a regra de incremento de área para melhorar o *matching* não é mais valida unicamente, há de se avaliar uma nova componente relacionada diferentemente com a área reduzindo o *matching*. Através das figuras 5.28 e 5.29 exemplifica-se esta propriedade em 65nm, assumindo que este valor continuará decrescendo em novas tecnologias nanométricas estendendo-se o conceito de TSTs pode-se utilizar o transistor de máximo *matching* e a partir deste criar associações que sobrepujem os transistores comuns em *matching*, *leakage* e área. Vê-se claramente na figura 5.29 que a partir de um ponto somente incrementa-se o casamento dos transistores a custo de maior consumo de potência e de redução de *headroom*.

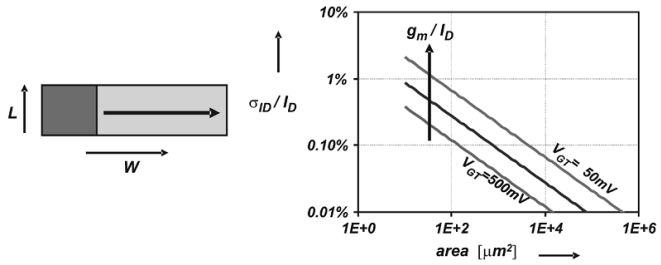


Figura 5.29 – Variação do *matching* em transistor NMOS 65nm com variação linear de W com L constante em função da área (NAUTA, 2005)

Na figura 5.30 tem-se um gráfico da função de ganho de corrente, levando-se em conta corrente de dreno sobre a corrente de fuga pelo *gate*. Vê-se claramente que os transistores são sintonizados para operação digital, ou seja, com canais mínimos temos o maior ganho sendo este decrescido com o aumento do L. Esta característica torna-se pior a cada nova tecnologia como demonstrado na comparação entre 90nm e 65nm. Analisando-se estes dados torna-se um fato que há incremento de ganho utilizando TSTs pois estes podem ser construídos com largura de canal mínimo, emulando larguras de canal maiores, para assim comporem uma associação gerando maior ganho.

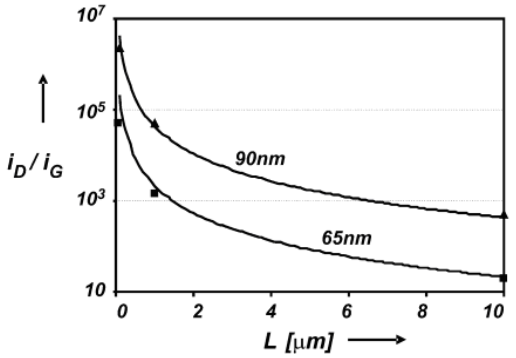


Figura 5.30 – Variação do ganho de corrente pela variação da largura de canal do transistor em tecnologia CMOS, Vgs=0,5V (ANNEMA, 2005)

Em projeto de circuitos digitais a utilização de TATs / TSTs reduz o *subthreshold leakage*, assim como o *stack effect* explorado também por Butzen (2007) entre outros, demonstrando ganhos na exploração de estruturas com *footed transistors* para redução do consumo estático. Em circuitos analógicos a configuração TAT oferece uma redução do *subthreshold leakage* como demonstrado por Lee (2004).

A previsão para os projetos em novas tecnologias nanométricas tem demonstrado que cada vez mais teremos circuitos *mixed-signal* analógico-digitais. Circuitos analógicos recebendo compensações de circuitos digitais e circuitos digitais recebendo compensação através de circuitos analógicos. Para a escolha do domínio onde a compensação será implementada deve-se sempre considerar a área consumida para a escolha do tipo de circuito de compensação. Finalmente avaliando as considerações já realizadas os TSTs criam uma nova dimensão de otimização disponível ao projeto *mixed-signal*.

Avançando-se para o projeto de circuitos em tecnologias UDSM utilizando o conceito de associações tipo T há menor corrente de fuga tanto pela porta bem como pelo canal. Sendo comprovado pela formulação do *leakage* de *gate* que é proporcional à área do dispositivo, e a redução do *subthresholds leakage* com a utilização de 2 transistores em série ou *self cascodes* é comprovada nos *papers* de (YAN, 2005) e (BUTZEN, 2007). Finalmente, a elevação do *matching* pela redução do *leakage* comprovado no *paper* (ANNEMA, 2005) e (NAUTA, 2005), e a obtenção de layout mais regular das estruturas (GIRARDI, 2003). Tudo isto contribui positivamente para o aumento da confiabilidade do projeto com TSTs e faz com que este torne-se uma alternativa viável para o projeto de circuitos analógicos em tecnologias UDSM, onde variações das propriedades físicas dos transistores MOS resultam das flutuações estatísticas locais dos transistores fabricados na mesma região do "*die*".

6 CONCLUSÃO

O projeto moderno de circuitos integrados em sua evolução agrega maior número de circuitos em uma mesma pastilha de silício, sendo que muitos sistemas digitais usam, ou valem-se, de módulos e circuitos analógicos em partes críticas ao seu funcionamento e vice versa. O projeto de circuitos integrados modernos implica em projeto *mixed-signal* sempre buscando vantagens em área e funcionalidade. Salienta-se que o estudo na área de economia de energia leva a um aprofundamento em eletrônica analógica CMOS, circuitos digitais dinâmicos, etc contribuindo no desenvolvimento de uma visão abrangente em termos de possibilidades de projeto, abrindo-se um leque de possibilidades em várias áreas de projeto e refinamento de circuitos integrados ao projetista. Observa-se a contribuição desta dissertação no sentido de prover um estudo aprofundado dos desafios de projetos CMOS com economia de energia dinâmica e estática. A implementação das técnicas apresentadas leva a circuitos mais robustos e mais confiáveis ante a prototipação, tanto por durabilidade bem como por serem mais viáveis economicamente.

O banco de memória implementado e simulado apresenta os desafios da variabilidade nas tecnologias modernas e principalmente quanto às soluções de redução de *leakage* são efetivas. A finalidade de comprovação da teoria pesquisada foi alcançada e demonstra um foco de pesquisa a ser melhor explorado através da implementação dos blocos dinâmicos ajustáveis de controle.

Finalmente apresentou-se o TST como uma solução viável a implementação de circuitos analógicos em tecnologias nanométricas provendo diversas características que apresentam-se cada vez mais apropriadas para o desenvolvimento de circuitos analógicos. A criação de uma ordem a mais de liberdade de desenvolvimento ao projetista provê novas possibilidades de economia de área e energia. Os testes efetuados comprovaram a validade do conceito quanto à aplicabilidade focado em projetos de sistemas de baixa potência.

REFERÊNCIAS

CAMACHO-GALEANO, E. M.; GALUP-MONTORO, C.; SCHNEIDER, M. C. A 2-nW 1.1-V **Self-Biased Current Reference in CMOS Technology. IEEE Transactions on Circuits and Systems - II**: Express Briefs, v.52, n.2, p.61–65, Feb. 2005.

RABAEY, J. M. **Digital Intregrated Circuits: A design perspective**, series editor, USA, Prentice Hall, 1996.

RABAEY, J. M.; PEDRAM, M. **Low Power Design Methodologies**, Boston, Kluwer Academic, 1996.

TSIVIDIS, Y. **Operation and Modeling of the MOS Transistor.** 2nd ed. Oxford: Oxford University Press, 1999. 620p.

ENZ, C. **Low-Power HF Microelectronics: a unified approach**. 1996. p.247–299.

BUTZEN, P. F. **Leakage Current Modeling in Sub-micrometer CMOS Complex Gates**, MsC. Thesis, PPGC – UFRGS, Porto Alegre, Brasil, 2007.

CHOI, J. H. **Mixed-Signal Analog-Digital Circuits Design on the Pre-Diffused Digital Array Using Trapezoidal Association of Transistors**. 2001. PhD Thesis. Universidade Federal do Rio Grande do Sul, Porto Alegre-RS.

CORTES, F. P. **Analysis, design and implementation of baseband and RF blocks suitable for a multi-band analog interface for CMOS SOCs.,** Ph.D. Thesis, PGMICRO – UFRGS, Porto Alegre, Brasil, 2008.

GIRARDI, A. **Automação do projeto de modulos CMOS analógicos usando associações trapezoidais de Transistores,** Ph.D. Thesis, PGMICRO – UFRGS, Porto Alegre, Brasil, 2007.

GIRARDI, A. **Uma Ferramenta para automação da geração do leiaute de circuito sanalogicos sobre uma matriz de transistores MOS pré-difundidos**, MsC. Thesis, PPGC – UFRGS, Porto Alegre, Brasil, 2003.

QIN, H.; RABAEY, J. M. **Deep Sub-MICRON SRAM Design for Ultra-Low Leakage Standby Operation**, PhD. Thesis, University of Califórnia, Berkeley, USA, 2007.

AMELIFARD, B.; FALLAH, F.; PEDRAM, M. **Low-Leakage SRAM Design with Dual Vt Transistors**. ISQED '06 – 7th International Symposium on Quality Electronic Design, Proceedings, 2006

ANNEMA, A.; et al. **Analog Circuits in Ultra-Deep-Submicron CMOS**. IEEE Journal of Solid-State Circuits, v.40, n.1, January 2005.

BLOMSTER, K.; DELGADO-FRIAS, J. G. **High Peformance Memory Read Using Cross-Coupled Pull-up Circuitry**. 49th IEEE International Midwest Symposium on Circuits and Systems - MWSCAS '06. Volume 1, Page(s):332 – 336, Aug. 2006

CALHOUN, B. H.; CHANDRAKASAN, A. P. **A 256kb 65-nm Sub-treshold SRAM Design for Ultra-Low-Voltage Operation**. IEEE Journal of Solid-State Circuits, v.42, no. 3, March 2007.

CAO, K.M. et al. BSIM4 Gate Leakage Model Including Source-Drain Partition.In: INTERNATIONAL ELECTRON DEVICES MEETING, 2000. **Digest of Technical Papers.** [S.l.: s.n.], 2000. p. 815-818.

CHATTOPADHYAY, S. **Low Power Techniques for Nanometer Design Processes – 65nm and smaller.** Symposium on Integrated Circuits and Systems Design, SBCCI, 2007. Tutorial, 2007.

CHEN, Q.; et al. **Circuit-aware Device Design Methodology for Nanometer Technologies: A Case Study for Low Power SRAM Design.** IEEE Design, Automation and Test in Europe Conference and exhibition (DATE ´06), Proceedings, 2006

CORTES, F. P.; FABRIS, E.; BAMPI, S. **Applying the gm/ID method in the analysis and design of a Miller amplifier, a comparator and a Gm-C band-pass filter**., IFIP VLSI Soc 2003, Darmstadt, Germany, December 2003.

CUNHA, A. I. A.; SCHNEIDER, M. C.; GALUP-MONTORO, C. **An MOS Transistor Model for Analog Circuit Design**. IEEE Journal of Solid-State Circuits, v.33, n.10, p.1510–1519, Oct. 1998.

DEGALAHAL, V.; et al. **Soft Errors Issues in Low-Power Caches**. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 13, NO. 10, OCT. 2005

ELAKKUMANAN, P.; THONDAPU, C.; SRIDHAR, R. **A Gate Leakage Reduction Strategy For Sub-70 nm Memory Circuits**. IEEE Dallas/CAS Workshop Implementation of High Performance Circuits (DCAS-04). Proceedings, Page(s):145 – 148, Sept. 2004

FLAUTNER, K.; et al. **Drowsy Caches: Simple Techniques for Reducing Leakage Power**. IEEE 29th Annual International Symposium on Computer Architecture (ISCA ´02), Proceedings, 2002

GALUP-MONTORO, C.; SCHNEIDER, M.; LOSS, I. **Series-Parallel Association of FET's for High Gain and High Frequency Applications**. IEEE Journal of Solid-State Circuits, v.29, n.9, Sept. 1994.

GIELEN, G.; DEHAENE, W. **Analog and Digital Circuit Design in 65nm CMOS: end of the road?**. IEEE Design, Automation and Test in Europe Conference and exhibition (DATE ´05), Proceedings, 2005

GIRARDI, A.; BAMPI, S. LIT - **an automatic layout generation tool for trapezoidal association of transistors for basic analog building blocks.** DESIGN AUTOMATION AND TEST IN EUROPE, 2003. Proceedings… Piscataway, NJ: IEEE, 2004.

GIRARDI, A.; CORTES, F. P.; CONRAD JR., E.; BAMPI, S. **T-Shaped Association of Transistors: Modeling of Multiple Channel Lengths and Regular Associations**.

Symposium on Integrated Circuits and Systems Design, SBCCI, 2005. Proceedings… 2005.

HU, J. S.; et al. **Exploiting Program Hotspots and Code Sequentiality for Instruction Cache Leakage Management**. ISLPED '03 - International Symposium on Low Power Electronics and Design, Proceedings, Aug. 2003

JAIN, S. K.; AGARWAL, P. **A Low Leakage and SNM Free SRAM Cell Design in Deep Sub Micron CMOS Technology**. 19th IEEE International Conference on VLSI Design (VLSID´06), Proceedings. 2006.

JOUPPI, N. P.; REINMAN, G. CACTI 2.0: **An Integrated Cache Timing and Power Model,** Western Research Laboratory, USA, 2000.

JOUPPI, N. P.; SHIVAKUMAR, P. CACTI 3.0: **An Integrated Cache Timing, Power, and Area Model,** Western Research Laboratory, USA, 2001.

JOUPPI, N. P.; WILTON, S. J. **An Enhanced Access and Cycle Time Model for On-Chip Caches,** Western Research Laboratory, USA, 1994.

KAXIRAS, S.; XEKALAKIS, P.; KERAMIDAS, G. **A Simple Mechanism to Adapt Leakage-Control Policies to Temperature**. ISLPED '05 - International Symposium on Low Power Electronics and Design, Proceedings Page(s):54 – 59, Aug. 2005

KIM, N. S.; et al. **Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power**. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 12, Feb. 2004

KIM, N. S.; et al. **Drowsy Instruction Caches**. IEEE/ACM 35th Annual International Symposium on Computer Microarchitecture (MICRO-35), Proceedings, 2002

KIM, S.; et al. **On Load in Low-Power Caches**. ISLPED '03 - International Symposium on Low Power Electronics and Design, Proceedings, Aug. 2003

KIM, N. S.; BLAAUW, D.; MUDGE, T. **Quantitative Analysis and Optimization Techniques for On-Chip Cache Leakage Power**. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 13, NO. 10, FEB. 2005

KIM, N. S.; et al. **Single-Vdd and Single-Vt Super-Drowsy Techniques for Low-Leakage High-Peformance Instruction Caches**. ISLPED '04 - International Symposium on Low Power Electronics and Design, Proceedings, Aug. 2004

KIMURA, K.; ET AL. **Power Reduction Teechiniques in Megabit DRAM´s**, IEEE Journal of Solid-State Circuits, V. 21, Nº 3 , June 1986.

KWAI, P.; et al. **SRAM Cell Current in Low Leakage Design**. IEEE International Workshop on Memory Technology, Design, and Testing (MTDT´06). Proceedings, 2006

LEE, D.; BLAAUW, D.; SYLVESTER, D. **Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits**. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 12, NO. 2, FEB. 2004

LEE, H.; LEUNG, K. N.; MOK, P. K. T. **Low-Voltage Analog Circuit Techniques Using Bias-Current Re-tilization, Self-Biasing and Signal Superposition**. In: IEEE CONFERENCE ON ELECTRON DEVICES AND SOLID-STATE CIRCUITS, 2005. Proceedings.. . Piscataway: IEEE, 2005. p.533–536.

LEE, D.; BLAAUW, D.; SYLVESTER, D. **Static Leakage Reduction trough Simultaneous Vt/Tox and State Assignment**. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 24, NO. 7, July 2005

LI, Y.; et al. **State-Preserving vs. Non-State-Preserving Leakage Control in Caches**. IEEE Design, Automation and Test in Europe Conference and exhibition (DATE ´04), Proceedings, 2004

MARGALA, M. **Low-Power Circuits Design**, The 7th IEEE International Workshop on Memory Technology, Design, and Testing, Proceedings p.115, 1999.

MORITA, Y.; et al. **An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design under DVS Enviroment**. Symposium on VLSI Circuits Digest of Technical Papers, Proceedings, 2007.

NAUTA, B.; ANNEMA, A. **Analog/RF Circuit Design for Nanometerscale IC Technologies**. 31st European Solid-State Circuits Conference - ESSCIRC 2005. Proceedings, Page(s):45 - 53, Sept. 2005.

NOURIVAND, A.; et al. **An Adaptive Sleep Transistor Biasing Scheme for Low Leakage SRAM**. IEEE International Symposium on Circuits and Systems - ISCAS 2007. Proceedings, Page(s):2790 – 2793, May 2007

OHBAYASHI, S.; et al. **A 65-nm SoC Embedded 6T-SRAM Designed for Manufacturability With Read and Write Operation Stabilizing Circuits**. IEEE Journal of Solid-State Circuits, v.42, n.4, p.820–829, April 2007.

OKAZAKI, N.; et al. **A 30ns 256K Full CMOS SRAM** , ISSCC, february 1986.

PAULA, L. S.; FABRIS, E.; BAMPI, S. **A high swing low power CMOS differential voltage-controlled ring oscillator.**, 14th IEEE International Conference on Electronics, Circuits and Systems – ICECS 2007, Marrakech, Morocco, December, 2007.

PRINCE, B. **Embedded Non-Volatile Memories.** Symposium on Integrated Circuits and Systems Design, SBCCI, 2007. Tutorial, 2007.

PRINCE, B. **Nanotechnology and Emerging Memories.** Symposium on Integrated Circuits and Systems Design, SBCCI, 2007. Tutorial Notes, 2007.

RAJPUT, S. S.; JAMUAR, S. S. **Low Voltage Analog Circuit Design Techniques**. IEEE Circuits and Systems Magazine, v.2, n.1, p.24–42, 2002.

RAO, R. R.; et al. **Bus Encoding for Total Power Reduction Using a Leakage – Aware Buffer Configuration**. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 13, NO. 12, December 2005

ROY, K.; MUKHOPADHYAY, S.; MAHMOODI-MEIMAND, H. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. **Proceedings of the IEEE**, New York, v.91, n.2, p. 305-327, Feb. 2003

RAZAVIPOUR, G.; MOTAMEDI, A.; AFZALI-KUSHA, A. **WL-VC SRAM: A Low Leakage Memory Circuit for Deep Sub-Micron Design**. IEEE International Symposium on Circuits and Systems - ISCAS 2006. Proceedings. 2006

TAKEYAMA, Y.; et al. **A Low Leakage SRAM Macro with Replica Cell Biasing Scheme**. Symposium on VLSI Circuits Digest of Technical Papers, Proceedings, 2005.

TSIATOUHAS, Y.; et al. **New Memory Sense Amplifier Designs in CMOS Technology**, The 7th IEEE International Conference on Electronics, Circuits and Systems, 2000.

VEERAVALLI, A.; SÁNCHEZ-SINENCIO, E.; SILVA-MARTINEZ, J. **Transconductance Amplifier Structures With Very Small Transconductances: a comparative design approach**. IEEE Journal of Solid-State Circuits, v.37, n.6, p.770–775, June 2002.

WADA, T.; RAJAN, S.; PRYBYLSKY, S. A. **An Enhanced Access and Cycle Time Model for On-Chip Caches**, IEEE Journal of Solid-State Circuits, V. 27, Nº 8 , Aug 1992.

WANG, C. C.; LEE, P. M.; CHEN, K. L. **An SRAM Using Dual Threshold Voltage Transistors and Low-Power Quenchers**, IEEE Journal of Solid-State Circuits, V. 38, Nº 10 , Oct 2003.

WHANG, H.; et al. **Systematic Analysis of Energy and Delay Impact of Very Deep Submicron Process Variability Effects in Embedded SRAM Modules.** IEEE Design, Automation and Test in Europe Conference and exhibition (DATE ´05), Proceedings, 2005

WIECKOWSKI, M.; MARGALA, M. **A 32KB SRAM Cache Using Current Mode Operation And Asynchronous Wave-pipelined Descoders**, IEEE SOC Conference, Proceedings. IEEE International, 2004.

YAN, S.; SANCHEZ-SINENCIO, E. **Low Voltage Analog Circuit Design Techniques: A Tutorial**. IEICE Trans. Analog Integrated Circuits and Systems, Vol. E00-A, NO.2, February 2000

ZHAI, B.; et al. **A Sub-200mV 6T SRAM in 0.13µm CMOS**. IEEE International Solid-State Circuits Conference, Proceedings, 2007.

ZHAI, B.; et al. **Extended Dynamic Voltage Scaling for Low Power Design**. IEEE International SOC Conference, 2004. Proceedings, Page(s):389 – 394, Sept. 2004

ZHAI, B.; et al. **The Limit of Dynamic Voltage Scaling and Insomniac Dynamic Voltage Scaling**. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 13, Page(s):1239 – 1252, Nov. 2005

SANCHEZ-SINENCIO, E. **65nm Predictive Technology Model Characterization**. Class Notes TAMU, USA, Disponível em < http://amesp02.tamu.edu/~sanchez/607%20ref%2065nm%20Predictive%20Technology%20Model%20Characterization.doc >. Acesso em: Março. 2008.

WAGNER, F. R. **Organização de Computadores B: Memórias Cache, aulas 16, 17 e 18**, UFRGS,Porto Alegre. Disponível em < http://www.inf.ufrgs.br/inf113/crono.html >. Acesso em: set. 2005.

ANEXO A - ESTRUTURA DE MEMÓRIAS SRAM E CACHES

O estudo de memórias SRAMs e caches nos reporta a duas análises principais, uma da parte de controle e outra da parte operativa da memória. A análise da parte de controle se reporta principalmente a funcionalidade e controle da memória, já a análise da parte operativa se concentra principalmente no funcionamento da memória no nível elétrico. Na análise da parte operativa definem-se os alvos de projeto tanto em termos de frequência de operação, integração, compactação, aspecto de forma, entre outras características da memória alvo. Memórias SRAM são empregadas em sistemas com necessidade de memórias de alto desempenho e através do acréscimo de um controlador apropriado e algumas alterações na parte operativa podem vir a ser utilizadas como memórias cache. A cache tem a função de transparecer ao circuito que o acesso a dados externos se dá em menor tempo que o requerido nesta operação, funcionando como ponte entre estas trocas de dados. Gerenciando as trocas de dados entre o circuito principal e a memória e periféricos externos, compensando as diferentes velocidades de operação e troca de dados, as caches são amplamente utilizadas nos dias de hoje.

A análise do projeto de memórias SRAM ou cache inicia-se com o estudo da arquitetura, organização e hierarquia da memória desejada. Ao final destas definições passa-se ao projeto elétrico onde através da definição do funcionamento dos blocos funcionais da memória passa-se ao projeto das células básicas que os compõem sempre objetivando alcançar os níveis de desempenho determinados. Na figura A.1 vemos a comparação entre a parte operativa de uma SRAM e uma cache, visualiza-se a principal diferença que se dá na presença de um comparador acrescido aos demais blocos, este dependente da arquitetura e controle determinados a memória em construção.

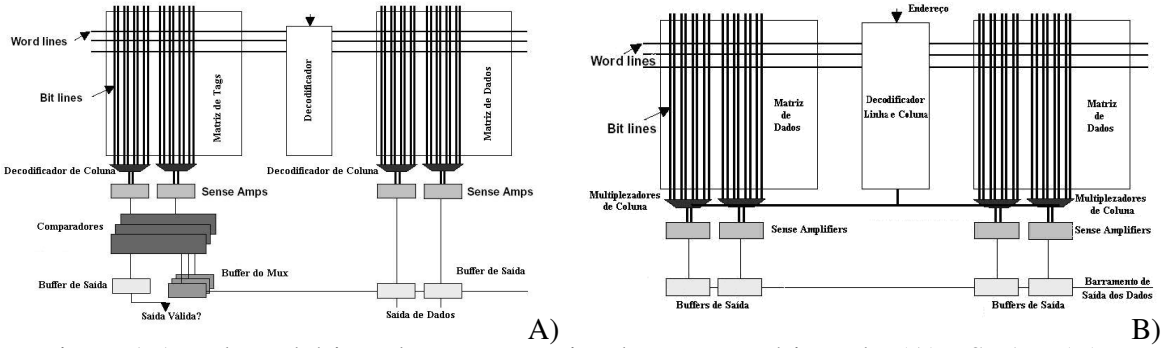


Figura A.1 – Blocos básicos da parte operativa de uma memória cache (A) e SRAM (B)

A análise e projeto de uma memória se facilita ao tratarmos a memória bloco a bloco, diminuindo o nível de abstração à medida que se trabalha no bloco, em seus subcomponentes e finalmente no nível do projeto elétrico. A função de cada um dos

blocos é específica e seu projeto tem suas peculiaridades, assim se fará um resumo dos blocos e de suas funções para partindo deste ponto iniciar-se a análise do foco do texto.

Arquitetura e organização de memórias SRAM

Ao tratar-se a arquitetura e organização de uma memória SRAM, deve-se inicialmente descrever as características básicas desta, memórias SRAM apresentam melhor desempenho comparativamente a outros tipos de memórias fabricadas em mesma tecnologia. São utilizadas principalmente como memórias de alto desempenho e amplamente utilizadas nos dias de hoje no intuito de compensar o crescente hiato de frequência máxima entre os circuitos digitais e memórias principais. A memória principal de um sistema digital normalmente é composta por DRAMs (*Dynamic Random Access Memory*). Estas apresentam menor custo de fabricação por seu baixo custo na relação megabyte armazenado por área ser melhor, sendo assim normalmente utilizadas na função de memória principal em sistemas digitais em geral. Por apresentarem menor desempenho comparativamente a SRAMs devido a seu maior tempo de acesso e a necessidade de *refresh* as mesmas são incompatíveis com funções que exijam alto desempenho.

As memórias SRAMs são randômicas que por definição significa que o tempo de acesso a todas as posições de uma memória é o mesmo, sendo esta característica imprescindível no quesito de desempenho tanto de leitura como de gravação que é algo primordial em memórias de alto desempenho. Por ser uma memória estática, seu conteúdo é mantido enquanto houver alimentação ativa sem a necessidade de um *refresh* das posições gravadas da matriz de memória. Uma de suas deficiências é a baixa densidade de integração que advêm de seus blocos construtivos básicos consumirem maior área, assim o número de células integradas por uma mesma unidade de área é menor que em outros tipos de memórias elevando seu custo de fabricação final. Outra deficiência é o alto consumo de potência, tanto dinâmica como estática que dificulta a sua utilização nas tecnologias UDSM e a aplicação em dispositivos *low power*.

Após esta análise inicial vê-se a necessidade de estabelecer as métricas da utilização dessas memórias em sistemas digitais, e assim partimos a hierarquia de memória utilizada. Hierarquia de memória pode ser tratada como uma forma de emular ao circuito principal que existe acesso aos dados dos meios externos em menor tempo que o realmente necessário ou despendido na operação. A ideia principal é emular ao requerente o máximo de memória com menor atraso e menor custo de fabricação. Dessa ideia derivam-se ideias subsequentes que são o método de acesso às camadas superiores de memória e o sistema de gravação e gerenciamento dos dados contidos em uma cache.

Principio de Localidade:

O princípio de localidade é empregado em sistemas digitais pois a ideia principal aplicada é que programas ou circuitos acessam repetidamente alguns dados ou leem uma sequência dos mesmos durante um bloco de instruções. Repetir trechos de código e/ou acessar repetidamente dados próximos caracterizam operações diferentes em sistemas digitais e desta forma originam diferentes estratégias de controle. Desta percepção do modo de acesso dos programas originou-se o conceito do principio de localidade, dividindo-se o mesmo em dois tipos: - Localidade temporal e localidade espacial. Algum destes deverá ser empregado pela controladora da cache, sendo que o foco deste texto a parte operativa da memória principalmente e não seu controle, omitiram-se assim maiores considerações e esclarecimentos referentes à metodologia a ser empregada.

Localidade temporal:

Normalmente mais adequado para arquivos de dados, onde com mais frequência acessa-se os mesmos dados. Estas posições de memória, uma vez acessadas, tendem a ser acessadas novamente no futuro próximo, como mostrado na figura A.2. Esta estrutura de dados pode evitar a ida aos meios externos obtendo economia em tempo e energia consumida.

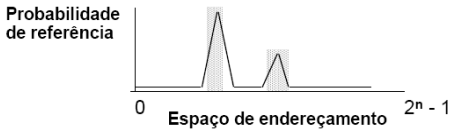


Figura A.2 – Principio de localidade temporal (Patterson/Hennessy, 1997)

Localidade espacial:

Normalmente mais adequado para arquivos de instruções, onde com mais frequência acessa-se uma lista encadeada de ações. Estes endereços que serão acessados tendem a ser próximos dos endereços de acessos anteriores, como mostrado na figura A.3. Esta estrutura pode aumentar o rendimento de uma sequência de dados evitando a ida aos meios mais lentos obtendo como consequência menor atraso e maior velocidade de operação.

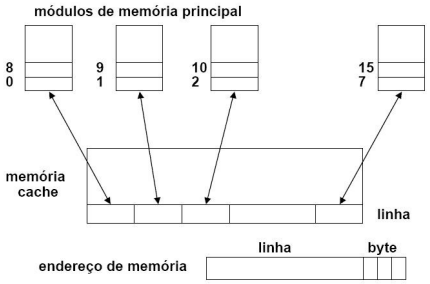


Figura A.3 – Principio de localidade espacial (Patterson/Hennessy, 1997)

Organização de memória

No modo de organização da memória é onde definem-se os modos de mapeamento entre a memória principal e a memória cache, esta um ou mais níveis abaixo da mesma. O mecanismo implementado será sempre similar entre os diferentes modos de mapeamento, onde o circuito principal gera um endereço de memória que é enviado à cache que deve verificar se há uma cópia da posição de memória correspondente ainda válida. O mapeamento entre endereços de memória principal e cache deve resolver três questões primordiais: - (1) Se existe cópia, (2) encontrar a posição da cache onde está a cópia; (3) se não existe uma cópia, buscar o conteúdo na memória principal escolhendo a posição da cache onde a cópia será armazenada. Estas operações devem ser executadas por hardware, acelerando ao máximo estas execuções para manter o desempenho da memória construída. Temos assim três diferentes estratégias de organização (mapeamento) da cache: - Mapeamento completamente associativo, mapeamento direto, mapeamento set-associativo.

Cada um destes métodos tem vantagens e desvantagens em algum dos diferentes aspectos do espaço de projeto. A organização tem impacto direto na parte operativa, o

método de mapeamento escolhido tem implicação direta sobre a quantidade de área utilizada na construção dos blocos.

Mapeamento Completamente Associativo:

A vantagem principal de seu emprego é a máxima flexibilidade, a controla pode no posicionamento de qualquer palavra ou linha da memória principal atrelá-la a qualquer palavra ou linha da memória cache. Desta forma não há a limitação por requisitos físicos, de mapeamento de regiões de memória para regiões determinadas da cache o que é uma vantagem quando analisa-se a potencialidade de alocação de dados na memória. Por outro lado há desvantagens no emprego deste tipo de mapeamento, principalmente o custo em hardware para a comparação simultânea de todos os endereços armazenados na cache, seguido pelo problema no algoritmo de substituição que controla a validade dos dados na memória e o método de substituição dos mesmos e finalmente o hardware para selecionar uma linha da cache para receber o dado como consequência de uma falta (miss). Em memórias associativas de tamanho reduzido pode-se usar tabelas, mas somente nestes casos devido ao aumento da memória necessária para manter esse registro ser elevado com o aumento da própria cache. Este método de mapeamento está exemplificado na figura A.4.

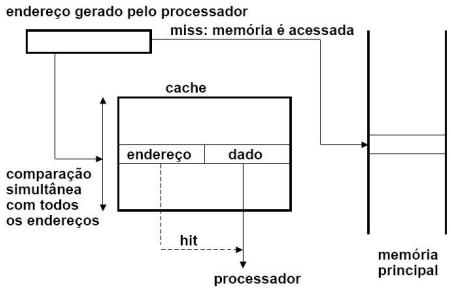


Figura A.4 – Exemplo mapeamento associativo (Patterson/Hennessy, 1997)

Mapeamento Direto:

Este método de mapeamento é o que provoca menos impacto no tamanho do hardware e na velocidade de procura de um dado na cache. A forma de operação deste hardware é na divisão do endereço utilizado em duas partes, a parte menos significativa é o índice, este é o endereço da cache que aponta o local onde será armazenado o dado. A parte mais significativa é o tag, armazenado na cache juntamente com o conteúdo da posição de memória como demonstrado na figura A.5.

Quando um acesso é feito, o índice é usado para encontrar palavra na cache, logo comparando o *tag* armazenado na cache com o *tag* requisitado se os mesmos forem idênticos houve um acerto (*hit*). Endereços de mesmo índice são mapeados sempre para a mesma posição da cache, as vantagens são que não há necessidade de um algoritmo de substituição o que se resulta em um hardware simples e de baixo custo com alta velocidade de operação.

As desvantagens consistem na redução do desempenho se acessos consecutivos são feitos a palavras com mesmo índice, nesta situação obtém-se um *hit ratio* inferior ao de caches com mapeamento associativo. Entretanto demonstra-se que o *hit ratio* aumenta com o aumento da cache e aproxima-se ao de caches com mapeamento associativo, sendo que a tendência atual é do uso caches com capacidade sempre crescente, assim compensando o *hit ratio* das diferentes caches.

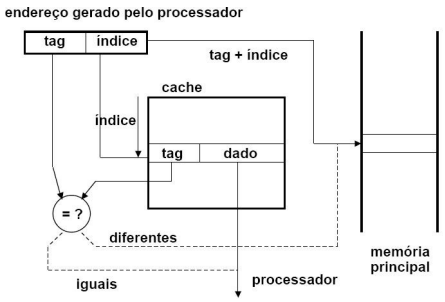


Figura A.5 – Exemplo mapeamento direto (Patterson/Hennessy, 1997)

Mapeamento Set-Associativo:

No mapeamento direto, todas as palavras armazenadas na cache preferencialmente devem ter índices diferentes, já no mapeamento associativo os dados podem ser colocados em qualquer posição da cache. A vantagens intrínsecas a ambos os tipos de mapeamento quando analisa-se a busca por maior desempenho, assim no mapeamento conjunto–associativo temos um compromisso entre um número limitado de linhas de mesmo índice mas de diferentes blocos de memória com possibilidade de *tags* distintos. Estas informações podem estar armazenadas na cache ao mesmo tempo, num mesmo conjunto, ou em conjuntos diferentes quando o índice das mesmas é idêntico. Assim o número de linhas de mesmo índice na memória é idêntico à associatividade da memória, como exemplificado na figura A.6.

A principal vantagem com relação ao mapeamento completamente associativo é a redução do número de comparadores, sendo estes compartilhados por todas as posições de um mesmo conjunto. O algoritmo de substituição necessita considerar somente linhas dentro de um mesmo conjunto, o que demonstra a o ganho de velocidade das mesmas por valer-se dos índices nos diferentes conjuntos. Pela flexibilidade acrescentada e desempenho elevado este tipo de mapeamento é bastante utilizado em microprocessadores.

As desvantagens da cache conjunto-associativo *N-way* versus mapeamento direto é a ocorrência de um atraso extra do multiplexador, assim o dado é disponibilizado somente depois da decisão Hit/Miss e da seleção do conjunto a saída. Quando consideramos uma cache com mapeamento direto o dado da cache estará disponível antes da decisão *Hit/Miss*, já na cache conjunto-associativa há o acréscimo de um atraso sendo possível assumir um *hit* e continuar o processo e recuperar-se depois se houver a ocorrência de um *miss*.

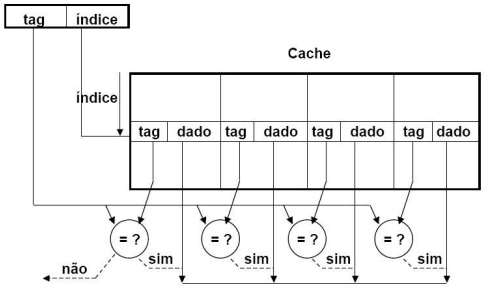


Figura A.6 – Exemplo mapeamento conjunto – associativo (Patterson/Hennessy, 1997)

Blocos Componentes Básicos de uma Memória SRAM

Pode-se desenvolver o projeto da memória através das características definidas de arquitetura, organização e desempenho; sendo que existe um conjunto de blocos básicos para composição de um banco de memória. A figura A.7 apresenta o esquema básico da parte operativa de uma memória SRAM, onde visualizam-se os blocos componentes básicos da mesma. O projeto de um banco completo fica simplificado se propusermos problemas menores a serem resolvidos individualmente, para então, através das replicações dos mesmos obter-se os blocos de composição com as características desejadas. Através da interligação dos bancos teremos a memória SRAM completa dentro do espaço de projeto determinado no inicio do projeto.

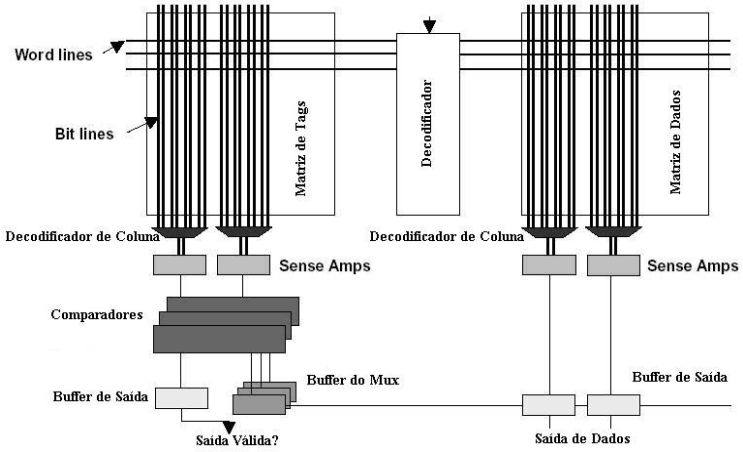


Figura A.7 – Blocos básicos da parte operativa de uma memória cache (JOUPPI, 2001)

De forma sistemática apresentar-se-á os blocos básicos construtivos componentes de uma memória através de uma breve análise dos mesmos. A partir deste ponto teremos o guia de passos a tomar para obter a estrutura completa de uma SRAM. Iniciando-se pela controladora, timing de controle das operações e finalmente passando aos sub-blocos da parte operativa os tópicos apresentarão a descrição da função e localização dos mesmos dentro da estrutura de um banco de memória. A interligação dos bancos obtidos resultará na construção de uma memória completa.

Controladora de Memória

A controladora é a entidade que na arquitetura controla a parte operativa de uma memória, sendo a parte da memória que gerencia os aspectos de atividade e inatividade da mesma colocando-a em modos de operação adequados. Outros fatores controlados pela mesma é o tempo de operação em modo de *stand-by* ou tempo de espera para este ativação deste mesmo modo em que linhas e sequência de linhas, qual o tempo para que a memória ou parte da mesma adormeçam. Como cache o principal objetivo da controladora de uma cache é fazer transparecer ao circuito principal que um acesso a memória principal tem menor tempo de espera, e que a gravação na mesma também leva menor tempo que o requerido. Funcionando como essa ponte entre os fluxos de dados a controladora é o bloco que estabelece as estratégias de que a parte operativa da memória irá valer-se, controlando os acessos em ambos os sentidos na memória principal. Mantendo também copias de dados recentemente acessados ou sequenciais que irão ser utilizados apresentam uma falsa impressão de que o tempo de acesso à memória acontece em menor tempo.

Aplicando alguma das estratégias de arquitetura e organização já apresentadas controla os tempos de validade dos dados na memória, bem como a temporização e acionamento da parte operativa. Dessa forma a controladora torna-se uma parte vital da estruturação de uma memória, sendo o ponto onde as estratégias de controle da memória serão aplicados, definindo os tempos de sonolência (*drowsy time*) e de desligamento de todas as posições de um banco (*decay time*). Estes dois conceitos dão origem aos dois modelos de controladoras que buscam low-power em caches, formando as *drowsy caches* e as *decay caches*. A controladora é responsável pelos sinais de ativação da memória SRAM a ela ligada, assim controlando a escrita e leitura e consequentemente a ativação de bancos para tais operações.

Temporização de Acesso à Memória

O *timing* de acesso à memória é o controle de eventos e tempos que coordena a parte operativa de uma memória ou uma cache. A definição da ordem de eventos que ocorrem quando a memória é acessada para leitura ou escrita seguindo uma sequência de eventos fixa de para cada operação, conforme a arquitetura da memória em operação. Na figura A.8 demonstram-se a sequência generalista de eventos na leitura de uma memória, esta sendo bem mais longa e com maior número de operações. A operação de escrita decorre na mesma sequência de passos até o ponto em que há uma tensão diferencial na *bit-lines* adequada ao valor a ser armazenado na célula de memória. Notadamente há encadeamento de eventos, se um dos mesmos possuir um atraso elevado pode limitar o desempenho global da memória projetada, assim os sub-blocos da parte operativa da memória devem ter atrasos adequados ao desempenho desejado. A geração dos sinais de acionamento se dá por uma cadeia de células de atraso atrelada a um circuito combinacional que efetua o controle dos tempos das operações na sequência adequado a arquitetura adotada. Manter os sinais de acionamento ativos durante um período possibilita o funcionamento da parte operativa da memória em maior frequência que a operada externamente a mesma. Quando atrelado a uma controladora, a memória e seus modos de operação são acionados e definidos por ela. Existe assim um controle superior que define o acesso e o modo de operação dos bancos de memória, controlando o timing da memória.

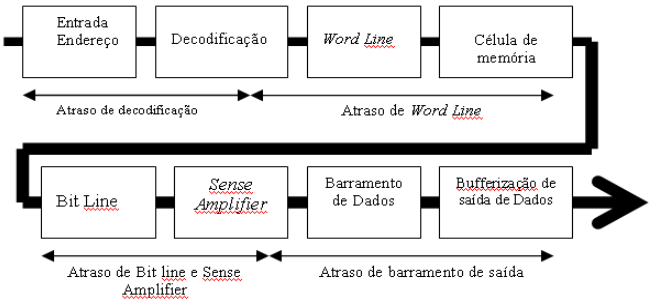


Figura A.8 – Sequência obedecida na leitura de uma memória SRAM

Célula de Memória Estática

A célula de memória tem por função a retenção do bit ali depositado em forma de tensão, sendo a granularidade mínima da matriz de memória. A utilização de área, consumo estático de potência e a margem de ruído estático (*static noise margin – SNM*) são pontos primordiais no projeto de uma memória, assim podem-se estabelecer os limites de operação para os modos ativo e de espera. Obedecendo aos limites impostos pela tecnologia utilizada e na busca pela melhor composição entre os eixos acima

citados, pode-se obter diferentes soluções que apresentam ganhos apreciáveis em um ou mais dos eixos. Imunidade a *bit-flips* é o que confere a confiabilidade ao valor salvo na posição de memória, assim o projeto deve garantir que entre as variações de alimentação acrescidas das variações de tensão nas *bit-lines* não criem *bit-flips*. A estabilidade da célula de memória deve ser mantida, tanto na operação de escrita como na de leitura, utilizando a mínima quantidade de energia possível. A célula formada por seis transistores é a mais utilizada no projeto de memórias SRAM, atribuído a ela temos em tecnologias acima de 100nm baixo consumo estático, sendo constituída por dois inversores e dois transistores para acesso ao bit da memória como mostrado na figura A.9. Pela literatura sabe-se da utilização de outros tipos de células de memória com fins e/ou características diferentes, sendo exemplos: - Células com resistores de *pull-up* em polisilício, células de quatro transistores, com acesso por transistores verticais, células de até 10 transistores para acesso único a uma posição de memória, operação por corrente, entre outras.

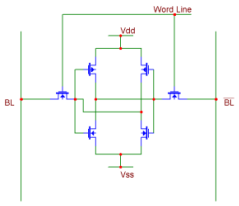


Figura A.9 – Célula de memória SRAM 6 Transistores

O aumento de performance de células de memória estática torna-se possível através da análise de dados provindos das especificações do projeto, corrente máxima através dos transistores de passagem na carga e descarga das tensões dos nós por eles amarrados, a capacitância da linha associada, a tensão de alimentação durante essa operação. Outro fator importante no emprego de uma célula de memória é método de implementação e o comprimento da *word line* refletindo-se na redução da sua resistência e consequentemente no atraso dessa linha. A replicabilidade da célula de memória e compactação do *layout*, provindo de um desenho compacto e do compartilhamento de contatos possibilita a minimização da área final da matriz de memória. A figura A.10 apresenta um projeto de uma matriz de células memórias de 6 transistores com bom aproveitamento de área pela inexistência de espaços vagos entre as células com grande regularidade na matriz formada.

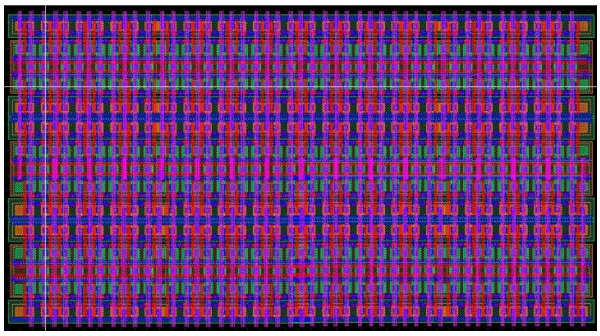


Figura A.10 – Exemplo de regularidade da matriz de células de memória

Sense Amplifier

O *sense amplifier* (SA) é um bloco importante na composição de uma memória pois define muitas das características da memória quanto a funcionamento, consumo

dinâmico/estático, confiabilidade e desempenho, tornando-se assim um dos blocos que merece uma atenção especial durante a fase de projeto. O princípio do SA é demonstrado na figura A.11.

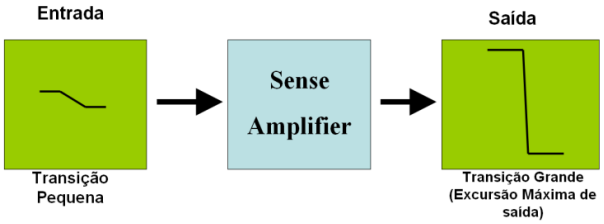


Figura A.11 – Principio de funcionamento do *sense amplifier*

Se o processo de detecção dos bits em uma memória for incorreto ou falhar seguidamente, mesmo que para um único bit somente, levará a erros encadeados e problemas de desempenho. Assim o *sense amplifier* influencia diretamente na confiabilidade da mesma, pois com uma pequena variação de tensão ou corrente em sua(s) entrada(s) através da(s) *bit-line(s)* que o interliga(m) a célula de memória executa a leitura do bit ali armazenado e indica-o em sua saída com o menos atraso possível. Dentre as características proporcionadas à memória pelos *sense amplifiers* citam-se:

- *Amplificação*: redução da excursão de tensão nas *bit lines* diminuindo atraso e dissipação de potência.
- *Aumento de velocidade*: compensação da capacidade da célula de memória de chavear capacitâncias, que leva a aceleração da transição em uma *bit-line*.
- *Restauração de Sinal*: Elevação do sinal ao máximo após a leitura, alimentando as capacitâncias dos barramentos com menor atraso.

Seu projeto requer um compromisso entre consumo e performance apresentando dificuldade por ser essencialmente de caráter analógico e de operação em alta frequência. A definição da arquitetura a ser empregada normalmente se dará no projeto elétrico justamente para atender as especificações requeridas pela memória, capacitância associada, velocidade, consumo de potência e corrente de fuga nas linhas.

Decodificador

Os decodificadores de linha e coluna são circuitos combinacionais que tem a função de interligar o endereço recebido à posição de memória a ele referente, armazenando o dado entrante ou o lendo. Assim, este é um bloco que detêm o controle de toda estrutura de blocos e sua metodologia de acesso, podendo quando partido ser o método para acionar apenas pequenas partes do circuito para economizar energia nas demais. Pode ser fabricado em diversos estilos de lógica no intuito de construir um balanço entre área consumida, atraso e consumo estático de correntes. Técnicas atuais com objetivo de economia de energia dividem o decodificador pela memória e dessa forma passam a constituir uma porcentagem significativa da área da mesma, pois geram também sinais de controle do estado das linhas da memória.

De forma reduzida a função principal do decodificador é habilitar corretamente as *word lines* e os multiplexadores relativos a posição de memória decodificada levando a informação contida ali ao barramento de saída. A figura A.12 apresenta um exemplo de um bloco decodificador.

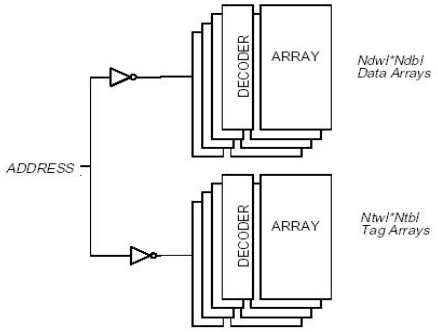


Figura A.12 – Exemplo de Esquema para decodificadores (JOUPPI, 1994)

Multiplexadores

Os blocos multiplexadores são circuitos combinacionais que são utilizados nos roteamentos de dados internos da memória, tanto para *bit-lines* para os *sense amplifiers*, bem como sinais de controle, etc. Seu emprego leva a um sistema onde há maior economia de energia, pois a capacitância atribuída ao ramo se reduz quando comparada à de barramentos longos. As aplicações e seu método de construção dependem diretamente da arquitetura a ser adotada no projeto, assim influenciando nas características finais da memória construída e seu funcionamento.

Pré Carga das Bit Lines

Este bloco é responsável pela carga das bit lines em uma tensão definida, por exemplo vdd/2, que antecede ao processo de leitura ou escrita de um bit. Esta operação de carregamento da capacitância das bit-lines deve ocorrer no menor tempo possível. A figura A.13 mostra um exemplo de circuito para pré-carga.

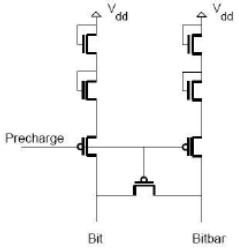


Figura A.13 – Circuito de pré-carga com transistor de equilíbrio (JOUPPI, 1994)

O circuito de pré-carga é necessário para equalizar-se ambas as bit lines a um valor intermediário o que resulta em duas condições importantes: (1)Ambas as entradas do *sense amplifier* em tensões idênticas tornando-o sensível a pequenas variações ao chavear-se a posição de memória e (2) resulta em menor variação para ambos os lados da excursão de alimentação ao chavear-se a posição de memória a *bit line* diminuindo a potência dissipada.

Comparador

O bloco de comparadores é um circuito combinacional utilizado para comparar o *tag* da posição da cache com a posição de memória requerida. Seu atraso é determinante no desempenho global da memória, pois é utilizado para comparação entre endereço provindo do processador e o *tag* lido na memória, se não houver um hit então será necessário buscar o dado pedido na memória do nível hierárquico acima (nível seguinte de cache, por exemplo), na memória principal ou disco. Será projetado para minimizar

as premissas já definidas de área, consumo estático e desempenho. A figura A.14 apresenta um exemplo de comparador utilizado na simulação pelo CACTI (JOUPPI, 1994).

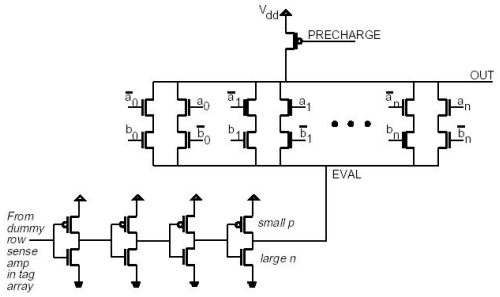


Figura A.14 – Exemplo de esquema elétrico comparadores (JOUPPI, 1994)