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

ANDRÉ LUÍS RODEGHIERO ROSA

# Projeto de Células e Circuitos VLSI Digitais CMOS para Operação em Baixa Tensão

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

Orientador: Prof. Dr. Sergio Bampi

Porto Alegre
2015

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Rosa, André Luís Rodeghiero
Projeto de Células e Circuitos VLSI Digitais CMOS para operação em baixa tensão / André Luís Rodeghiero Rosa. – 2015.
93 f.:il.
Orientador: Sergio Bampi;
Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2015.
1. CMOS digital. 2. Ajuste de tensão e frequência. 3. Eficiência energética em CMOS. 4. *Near-threshold*. 5. Transistores MOSFET multi-VT.
I. Bampi, Sergio. II. Projeto de Células e Circuitos VLSI Digitais CMOS para operação em baixa tensão

## AGRADECIMENTOS

A Deus por permitir a minha existência, prover inspiração e tranquilidade, principalmente, nos momentos difíceis.

Aos meus queridos pais, Rossana e José Luís, pelo carinho incondicional, apoio, exemplo, dedicação, motivação, educação e orientações. Amo vocês.

Ao meu irmão César pela amizade, companheirismo e motivação. Obrigado, também, pelos conselhos profissionais, financeiros e pessoais.

A minha amada Vanessa, pelo carinho, cumplicidade, paciência, motivação e apoio nos bons e maus momentos.

*In memoriam* aos meus saudosos avós Clair, Jovelina e José por colaborarem com a minha existência e pelo exemplo de humildade e idoneidade.

Ao meu avô Neri pela amizade e compartilhamento de memórias saudosas sobre suas jornadas nas estradas deste país.

Ao meu padrinho Gianni, minha madrinha Erlaine e minha tia Luzia pelo apoio em diversos momentos de minha vida.

Aos demais familiares que, de alguma forma, contribuíram para esta conquista.

A Gilda Ione Santos Goulart pelas doações de utensílios domésticos e móveis para uma permanência mais confortável em Porto Alegre.

A Marli Goulart de Moura por me hospedar em Porto Alegre até encontrar um local para residir.

Ao meu amigo e "irmão de Porto Alegre" Leonardo Bandeira Soares pela divisão do apartamento durante os anos iniciais, pelo companheirismo, cumplicidade, conselhos, colaboração na produção científica e desenvolvimento deste trabalho.

Ao meu orientador Prof. Sergio Bampi por me selecionar como orientando no PPGC, pela confiança, paciência, suporte e orientação durante todas as etapas do presente trabalho.

Aos meus amigos e ex-professores Sebastião Cícero Pinheiro Gomes, Vitor Irigon Gervini e Vagner Santos da Rosa pelo exemplo e incentivo para seguir na carreira acadêmica.

Ao meu amigo Dalton Colombo pelo companheirismo, apoio, conselhos e motivação para dar continuidade ao presente trabalho.

Ao meu amigo Luciano Timm Gularte pelo coleguismo, companheirismo, diligência, conselhos e motivação. Agradeço, também, pela parceria agradável nas viagens entre Porto Alegre e Pelotas.

Ao meu amigo e ex-colega de laboratório Kleber Hugo Stangherlin, por iniciar a pesquisa sobre operação em *near*-VT no âmbito da UFRGS e prover o ferramental para continuidade do presente trabalho. Agradeço, também, pela presteza e apoio incondicional.

Aos amigos de laboratório Cláudio Diniz, Eduarda Monteiro, Mateus Grellert, Leandro Ávila, Daniel Palomino, Dieison Silveira, Bruno Vizzotto e Felipe Sampaio pelo coleguismo, compartilhamento de experiências e suporte incondicional.

Aos demais amigos da UFRGS, especialmente ao Wiliam Guareschi, Eduardo Souza, Juan Brito, Henrique Pimentel, Antonio David Souza, David Cordova, Pedro Toledo, Sandro Binsfeld, Alexandre Simionovski pela parceria e coleguismo em diversos momentos de minha permanência em Porto Alegre.

À equipe de TI da UFRGS/NSCAD pelo excelente suporte, especialmente a Marcia Silva.

A Camila Mendonça Rabassa por conceder seu apartamento em Porto Alegre para que eu pudesse concluir este trabalho.

Aos meus amigos Tiago Férsula, Matheus Figueira, Daniel Buchweitz, Juliano Cipili, Wesley Castelluber, Lucas Hlenka, Lizandro Oliveira, Daniel Martins, Tiago Gonçalves, Paulo Correia, Márcio Neves pelo companheirismo e pelos momentos de descontração. Desculpem minha ausência em prol da conclusão deste trabalho.

Aos colegas do IFSUL/Pelotas que de alguma forma colaboraram com a conclusão deste trabalho, especialmente ao Paulo Fernando Aranalde Morales, Alexandre De Pauli Bandeira, Patrícia Borges Barcellos e Fernanda Pereira Teixeira De Mello.

**CMOS Digital Cells and VLSI Circuits Design for Ultra-Low Voltage Operation**

## ABSTRACT

This work proposes a strategy for designing VLSI circuits to operate in a very-wide Voltage-Frequency Scaling (VFS) range , from the supply voltage at which the minimum energy per operation (MEP) is achieved, at the Near-Threshold regime, up to the nominal supply voltage for the processes, if so demanded by applications workload. This master thesis proposes the sizing of transistors for three library cells using MOSFETs with different threshold voltages: Regular-VT (RVT), High-VT (HVT), and Low-VT (LVT). These libraries have five combinational cells: INV, NAND, NOR, OAI21, and AOI22 with multiple strengths. The sizing rule for the transistors of the digital cells was an adapted version from related works and it is directly driven by requiring equal rise and fall times at the output for each cell in order to attenuate variability effects in the low supply voltage regime. Two registers were also included in the RVT library cell. This library cell was characterized for typical, fast, and slow processes conditions of a CMOS 65nm technology; for operation at -40ºC, 25ºC, and 125ºC temperatures, and for supply voltages varying from 200 mV up to 1.2V, to include the region of interest, for $V_{DD}$ near the MEP. Experiments were performed with ten VLSI circuit benchmarks: notch filter, 8051 compatible core, four combinational and four sequential ISCAS benchmark circuits. From the energy savings point of view, to operate in MEP results on average reduction of 54.46% and 99.01% when compared with the sub-threshold and nominal supply voltages, respectively. This analysis was performed for 25ºC and typical process. When considered the performance, the very-wide VFS regime enables maximum operating frequencies varying from hundreds of kHz up to MHz/GHz at -40ºC and 25ºC, and from MHz up to GHz at 125ºC. This master thesis results, when compared with related works, showed on average an energy reduction and performance gain of 24.1% and 152.68%, respectively, for the same circuit benchmarks operating with $V_{DD}$ at the minimum energy point (MEP).

**Keywords**: Digital CMOS. Voltage-frequency scaling. CMOS Energy-efficiency. Near-threshold. Multi-VT MOSFET Transistors.

## RESUMO

Este trabalho propõe uma estratégia para projeto de circuitos VLSI operando em amplo ajuste de tensão e frequência (VFS), desde o regime em *Near-threshold*, onde uma tensão de $V_{DD}$ caracteriza-se por permitir o funcionamento do circuito com o mínimo dispêndio de energia por operação (MEP), até tensões nominais, dependendo da carga de trabalho exigida pela aplicação. Nesta dissertação é proposto o dimensionamento de transistores para três bibliotecas de células utilizando MOSFETs com tensões de limiar distintas: *Regular-VT* (RVT), *High-VT* (HVT) e *Low-VT* (LVT). Tais bibliotecas possuem cinco células combinacionais: INV, NAND, NOR, OAI21 e AOI22 em múltiplos *strengths*. A regra para dimensionamento dos transistores das células lógicas foi adaptada de trabalhos relacionados, e fundamenta-se na equalização dos tempos de subida e descida na saída de cada célula, objetivando à redução dos efeitos de variabilidade em baixas tensões de operação. Dois registradores também foram incluídos na biblioteca RVT e sua caracterização foi realizada considerando os parâmetros de processo CMOS 65 nm *typical*, *fast* e *slow*; nas temperaturas de operação de -40°C, 25°C e 125°C, e para tensões variando de 200 mV até 1,2V, para incluir a região de interesse, próxima ao MEP. Os experimentos foram realizados utilizando dez circuitos VLSI de teste: filtro digital *notch*, um núcleo compatível com o micro-controlador 8051, quatro circuitos combinacionais e quatro sequenciais do *benchmark* ISCAS. Em termos de economia de energia, operar no MEP resulta em uma redução média de 54,46% em relação ao regime de sub-limiar e até 99,01% quando comparado com a tensão nominal, para a temperatura de 25°C e processo típico. Em relação ao desempenho, operar em regime de VFS muito amplo propicia frequências máximas que variam de centenas de kHz até a faixa de centenas de MHz a GHz, para as temperaturas de -40°C e 25°C, e de MHz até GHz em 125°C. Os resultados desta dissertação, quando comparados a trabalhos relacionados, demonstraram, em média, redução de energia e ganho de desempenho de 24,1% e 152,68%, respectivamente, considerando os mesmos circuitos de teste, operando no ponto de mínima energia (MEP).



**Palavras-chave**: CMOS digital. Ajuste de tensão e frequência. Eficiência energética em CMOS. *Near-threshold*. Transistores MOSFET multi-VT.

# LISTA DE FIGURAS

# LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AD | *Analog-to-Digital* |
| CMOS | *Complementary metal-oxide-semiconductor* |
| CPL | *Complementary Pass-transistor Logic* |
| CPU | *Central Processing Unit* |
| CUT | *Circuit Under Test* |
| DC | *Direct Current* |
| DCVSL | *Differential Cascode Voltage Switch Logic* |
| DWPT | *Discrete Wavelet Packet Transform* |
| EDA | *Electronic Design Automation* |
| EDP | *Energy-Delay Product* |
| E/S | Entrada/Saída |
| FFT | *Fast Fourier Tranform* |
| FO2 | Fanout-of-2 |
| FO4 | *Fanout-of-4* |
| HVT | *High-VT* |
| ITRS | *International Technology Roadmap for Semiconductors* |
| LER | *Line-Edge Roughness* |
| LVT | *Low-VT* |
| MEP | *Minimum Energy Point* |
| MMMC | *Multi-mode Multi Corner* |
| MOS | *Metal-oxide-semiconductor* |
| MOSFET | *Metal-oxide-semiconductor Field Effect Transistor* |
| NMOS | *N-Channel Metal-oxide-semiconductor* |
| NTC | *Near-Threshold Computing* |
| NTV | *Near Threshold Voltage,* ou *Near-Threshold* |
| PDK | *Process Design Kit* |

| | |
|---|---|
| PDN | *Pull-down Network* |
| PDP | *Power-Delay Product* |
| PMOS | *P-Channel Metal-oxide-semiconductor* |
| PTM | *Predictive Technology Model* |
| PUN | *Pull-up Network* |
| PV | *Process Variation* |
| PVT | *Process, Voltage and Temperature* |
| RDF | *Random Dopant Fluctuations* |
| RVT | *Regular-VT* |
| SIMD | *Single Instruction, Multiple Data* |
| SNM | *Static-Noise Margin* |
| SoC | *System-on-Chip* |
| SPICE | *Simulated Program with Integrated Circuits Emphasis* |
| SRAM | *Static Random Access Memory* |
| STA | *Static Timing Analysis* |
| TG | *Transmission-Gate* |
| UFRGS | Universidade Federal do Rio Grande do Sul |
| $V_{DD}$ | Tensão de alimentação positiva |
| VFS | *Voltage-frequency Scaling* |
| $V_{GS}$ | *Gate-Source Voltage* |
| VLSI | *Very Large Scale Integration* |
| VT | *Threshold Voltage* |
| $t_r$ | *Rise Time* |
| $t_f$ | *Fall Time* |
| $t_p$ | *Propagation Delay* |
| $t_{pLH}$ | *Propagation Delay for a low-to-high output transition* |
| $t_{pHL}$ | *Propagation Delay for a high-to-low output transition* |
| TSMC | *Taiwan Semiconductor Manufacturing Company* |
| WBAN | *Wireless Body Area Network* |

**SUMÁRIO**

# 1 INTRODUÇÃO

Na década de 60 foi previsto que a densidade de transistores dobraria a cada geração tecnológica (MOORE, 1965). Tal constatação ainda é válida e, por sua vez, continua servindo como referência para a indústria de semicondutores no desenvolvimento de seus processadores multinúcleos. Estes processadores surgiram como uma alternativa aos processadores de um único núcleo, em função da necessidade de reduzir a potência dissipada proveniente das altas frequências de operação dos processadores *singlecore* de alto desempenho. Normalmente, processadores multinúcleos trabalham em frequências menores, exploram a concorrência para alcançar desempenho alto, e contam com a simplicidade de projeto e operação de cada núcleo para alcançar a eficiência energética (HIJAZ; KHAN, 2014). Entretanto, enquanto a Lei de Moore continua provendo mais transistores, restrições térmicas e de potência limitarão o número de núcleos que poderão estar simultaneamente ligados, bem como as suas frequências de operação (CHO; MAHLKE, 2012). Baseado em dados tecnológicos do ITRS e da Intel, no nó de 8 nm, mais de 50% da área do *chip* será desligada em função das restrições mencionadas anteriormente (SHAFIQUE et al., 2014).

Existe três principais fontes de dissipação de potência nos circuitos CMOS: potência dinâmica de chaveamento (*switching power*), potência dinâmica de curto-circuito (*short-circuit power*) e potência estática (*static power*) (CHANDRAKASAN; BRODERSEN, 1995). As fontes de dissipação citadas são representadas pela equação:

$$P_{méd} = P_{chav} + P_{cc} + P_{est} = \alpha C_L V_{DD}^2 f_{clk} + \beta I_{cc} V_{DD} + I_{est} V_{DD} \quad (1)$$

onde $P_{chav}$ é a potência dinâmica de chaveamento de capacitores, $P_{cc}$ é a potência dinâmica decorrente da corrente de curto-circuito, e a potência estática (DC) é denotada como $P_{est}$. É possível observar a partir de equação (1) que a potência dissipada depende quadraticamente da tensão de alimentação do circuito CMOS. A potência tem uma dependência linear com $V_{DD}$ nas parcelas estáticas e de curto-circuito, e uma relação quadrática na componente

dinâmica de chaveamento. Portanto, a redução da tensão de alimentação é fundamental para que o consumo seja atenuado. A potência dinâmica de chaveamento está relacionada com o produto do fator de atividade de chaveamento (α) nas entradas das portas lógicas pela carga/descarga das capacitâncias internas inerentes ao processo de fabricação CMOS e à capacitância de carga ($C_L$) pelo quadrado da tensão de alimentação ($V_{DD}$) e pela frequência de chaveamento do circuito ($f_{clk}$). A potência de curto-circuito é definida pelo produto da tensão de alimentação pela corrente de curto-circuito ($I_{CC}$), que surge quando há um caminho direto entre a linha de alimentação e o terra, no momento que os transistores PMOS e NMOS estão conduzindo simultaneamente (CHANDRAKASAN; SHENG; BRODERSEN, 1992). A corrente de curto-circuito, por sua vez, depende da frequência de chaveamento do circuito, do tempo de transição da tensão (tempo de *Slew*) nas entradas das portas, da tensão de alimentação e da tensão de limiar dos transistores. O termo produto $\beta I_{CC}$ na equação (1) deve incluir as dependencias mencionadas acima. Modelos mais detalhados da corrente de curto-circuito (VEENDRICK, 1984; DA COSTA et al., 2000) demonstraram que a potência média de curto-circuito depende não linearmente em ($V_{DD}$ - $2V_T$). A corrente estática ($I_{est}$) é composta por três parcelas, de acordo com o fenômeno físico que gera uma corrente estática entre $V_{DD}$ e terra (*ground*): i) a corrente por tunelamento do óxido de porta, ii) a corrente dreno-fonte de sub-limiar dos transistores MOSFETs, e iii) a corrente de fuga (corrente reversa) nos diodos parasitas presentes em cada porta lógica. É comum na literatura de projeto digital CMOS que autores referenciem a corrente estática como corrente de *leakage*, genericamente. A parcela de corrente estática de tunelamento torna-se relevante em nós tecnológicos mais recentes devido à diminuição (abaixo de 3nm) da espessura do óxido de silício (ou outro dielétrico) de porta que isola o eletrodo de *gate* do canal. A probabilidade de portadores (elétrons ou lacunas) tunelarem aumenta exponencialmente com essa diminuição de espessura (STANGHERLIN, 2013). As correntes de sub-limiar dos MOSFETs e de *leakage* dos diodos têm uma dependencia exponencial com a temperatura. A corrente de sub-limiar ocorre pela difusão de portadores entre fonte e dreno quando a polarização de *gate* está abaixo da tensão de limiar do transistor, onde a corrente de difusão é dominante (CHANDRAKASAN; SHENG; BRODERSEN, 1992). Esta parcela tem uma dependência exponencial com a tensão entre porta e fonte ($V_{GS}$). Por fim, a potência estática é obtida através do produto da corrente estática total pela tensão de alimentação.

Segundo Virga et al. (2014), as duas principais abordagens para reduzir potência e energia são a redução da tensão de alimentação e a diminuição das geometrias dos dispositivos no circuito integrado, ou seja, a redução do nó tecnológico CMOS. Entretanto,

parâmetros intrínsecos dos materiais como potencial de junção, potencial de barreira, função trabalho e tensão de limiar impuseram uma barreira prática para a diminuição da tensão de alimentação, à medida que um processo CMOS mais avançado é introduzido. Adicionalmente, a partir da tecnologia de 65 nm, a redução do nó tecnológico já não fornece os ganhos de energia que impulsionaram a indústria de semicondutores das últimas décadas (DRESLINSKI et al., 2010). Paralelamente a este dilema de projeto, existe uma demanda crescente por dispositivos alimentados por baterias, cuja carga deve durar na ordem de dias, enquanto os requisitos de funcionalidade são extremos, como, por exemplo, vídeo de alta definição, reconhecimento de voz, juntamente com uma série de padrões de redes sem-fio (DRESLINSKI et al., 2010). Adicionalmente, existem aplicações que necessitam de ultra-baixo consumo, como por exemplo, implantes biomédicos e redes de sensores autônomos (MARKOVIC et al., 2010). Segundo Chang e Haensch (2012), como as restrições atuais de refrigeração do *chip* e vida útil de baterias impõem limitações severas no desempenho de um produto, a eficiência energética será a chave para a sustentação continuada do aprimoramento de desempenho em sistemas VLSI futuros.

Uma abordagem há muito tempo conhecida da indústria e do meio acadêmico, para trabalhar em regime de ultra-baixo consumo de energia, é reduzir a tensão de alimentação de modo a operar na região de inversão fraca do transistor MOS. No trabalho de Swanson e Meindl (1972) foi utilizada implantação iônica de Boro para ajustar a tensão de limiar dos transistores de modo a operá-los em uma tensão mínima, bem abaixo de VT, de 200 mV a 27ºC. Desde então, houve um grande interesse no desenvolvimento de aplicações operando em condições de sub-limiar (sub-VT ou *sub-threshold*), como por exemplo, os trabalhos de Wang e Chandrakasan (2005) e Zhai et al. (2006). No primeiro, foi desenvolvido um processador FFT com comprimento variável de pontos (128 a 1024), com precisão dupla ( 8 e 16 bits), operando a 180 mV (VT de 450 mV). O processador foi fabricado em uma tecnologia de 180 nm e o ponto de mínima dissipação de energia, 155 nJ/FFT, foi atingido na tensão de 350 mV para uma FFT de 16 bits / 1024 pontos a uma frequência de 10 kHz. No trabalho de Zhai et al. (2006) foi desenvolvido um processador de propósito geral para aplicações baseadas em sensores, o qual foi prototipado em 130 nm, consumindo 2.6 pJ por instrução em seu ponto de maior eficiência energética, na tensão de 360 mV, a uma frequência de 833 kHz. Entretanto, o circuito é totalmente operacional a 200 mV. Os autores salientam que o núcleo do processador apresenta uma economia de energia da ordem de 10 vezes quando comparado a processadores semelhantes. Apesar de operar em tais condições de alimentação ser factível, trabalhar neste limite inferior de tensão resulta num aumento

exponencial indesejável em atraso (ASHRAF; ALZAHRANI; DEMARA, 2014) devido à redução de $V_{DD}$, ocasionando um aumento na mesma proporção na energia de *leakage*. Adicionalmente, neste regime de operação, há uma série de desafios que impactam na funcionalidade das portas lógicas, como, por exemplo, aumento de variabilidade ambiental, em função da temperatura, e de variabilidade de processo, em função de fenômenos como RDF (*Random Dopant Fluctuations*), LER (*Line-Edge Roughness*) e variação da espessura do óxido (KUHN, 2007). Desta forma, a operação na região de sub-limiar tem uma aplicabilidade limitada (ASHRAF; ALZAHRANI; DEMARA, 2014), ficando confinada a um conjunto menor de mercados, como relógios de pulso, aparelhos auditivos (DRESLINSKI et al., 2010) e de blocos lógicos de importância limitada em circuitos CMOS comerciais.

Para obter expressiva redução de potência e energia dissipadas, ao invés de operar os circuitos digitais CMOS no limite inferior de tensão, em regime de sub-VT, apresenta-se como alternativa viável, de melhor desempenho, a operação dos transistores em inversão moderada, próximo ao limiar dos transistores. Esta abordagem, denominada de *near*-VT (ou *near-threshold*), por vezes chamada de NTC (*Near-Threshold Computing*), por outros autores denominada NTV (*Near-Threshold Voltage*), entre outras, resulta em grandes benefícios em termos de economia de energia. Entretanto, não há um consenso na literatura em qual faixa de tensões de alimentação podemos considerar como *near*-VT. Segundo Dreslinski et al. (2010), a faixa estará muito próxima ou acima da tensão de limiar. Por outro lado, De (2013) afirma que a mesma está tipicamente acima da tensão de limiar, mas que o autor em questão prefere definir NTV como sendo a faixa que contém a tensão $V_{DD}$ e frequência onde a energia consumida por operação alcança um mínimo. De acordo com Dreslinski et al. (2010), o ponto de mínimo consumo por operação está localizado abaixo da tensão de limiar. Segundo Chandrakasan et al. (2010), o ponto de mínima energia por operação (MEP) não possui uma tensão fixa para um dado circuito, e pode variar amplamente dependendo da sua carga de trabalho, condições ambientais, como temperatura, e do balanço entre lógica combinacional, registradores e SRAM (*e.g.* caches) presentes no circuito integrado. Consequentemente, é admissível considerar que o regime de *near*-VT pode ser expandido para valores abaixo da tensão de limiar. As definições de De (2013) e Chandrakasan et al. (2010) serão adotadas neste trabalho. Os resultados obtidos ao cabo desta dissertação demonstrarão que, para os circuitos projetados neste trabalho, o MEP determinado por simulação para cada circuito sintetizado tende a situar-se abaixo da tensão de limiar média dos transistores. A operação em NTV apresenta menor sensibilidade às variações de processo e temperatura quando comparadas ao regime de sub-VT. Entretanto, os impactos destas variações na potência e nos

atrasos ainda são apreciáveis, muito relevantes e inerentes à operação em regimes de tensão reduzida em inversão moderada ou fraca. Uma forma de minimizar os efeitos da variabilidade é desenvolver uma biblioteca de células focada na operação a baixo $V_{DD}$. Segundo Stangherlin (2013), tal biblioteca deve levar em consideração todos os efeitos que surgem quando em operação em tensões reduzidas, como, por exemplo, amplitudes reduzidas de tensão e degradação das margens estáticas de ruído.

Uma abordagem bem estabelecida em projetos de sistemas dedicados a baixo consumo é denominada de ajuste dinâmico de tensão e frequência (VFS - *Voltage-frequency Scaling*). Esta técnica tem por função variar os referidos parâmetros de acordo com as demandas da carga de trabalho, em tempo de execução, atendendo de forma dinâmica às restrições de desempenho e consumo de energia (DRESLINSKI et al., 2010). O estado da arte em circuitos complexos comerciais (como microprocessadores, SoCs, memórias, etc), permite um ajuste em frequência de três a cinco vezes, no máximo, e uma redução de $V_{DD}$ até em torno de 0,7 V. Entretanto, Stangherlin (2013) propôs em seu trabalho uma redução da tensão nominal até o regime de NTV, denominando esta extrapolação das técnicas convencionais de VFS como *very wide Voltage-Frequency Scaling*. Apesar dos benefícios em termos de energia quando em *near*-VT, reduzir a tensão de alimentação até tais condições causa um aumento exponencial nos atrasos e nos tempos de transição nos caminhos lógicos em CMOS digital.

Outra técnica CMOS muito disseminada e efetiva para redução de consumo é a utilização de transistores multi-limiar (ou multi-VT). Normalmente, são oferecidos nas tecnologias CMOS mais recentes em três categorias: *Regular*-VT (RVT), ou *Standard*-VT, *High*-VT (HVT) e *Low*-VT (LVT). Os transistores LVT são os que apresentam os menores atrasos lógicos, em função de possuírem uma tensão de limiar inferior aos outros dois tipos. Entretanto, possuem a maior corrente de sub-limiar (ou *leakage*) entre os três transistores, por volta de 10 a 20 vezes superior em relação ao transistor de menor corrente sub-limiar (HVT), para uma tecnologia de 65 nm (LUO; NEWMARK; PAN, 2008). Por outro lado, os transistores HVT detêm os maiores atrasos em função do VT superior aos outros dois. O transistor RVT, possui consumo e desempenho intermediários. Segundo Stangherlin (2013), os transistores multi-limiar fornecem otimizações em nível de circuito, sob o ponto de vista de economia de energia e desempenho: transistores HVT podem ser usados em caminhos não críticos em termos de restrições de *timing*, enquanto transistores LVT, por serem os mais rápidos, podem ser utilizados nas células lógicas que estão nos caminhos críticos.

## 1.1 Objetivos

Em virtude dos conceitos anteriormente descritos, os objetivos deste trabalho são:

• Desenvolver uma biblioteca de células com circuitos combinacionais e sequenciais, utilizando transistores RVT, para operação a baixo $V_{DD}$, focando o mínimo ponto de energia por operação, levando em consideração a redução dos efeitos de variabilidade, inerentes em tal regime de trabalho. Tal biblioteca poderá operar em VFS amplo, se necessário;

• Utilizar a mesma metodologia de projeto da biblioteca RVT, no desenvolvimento de duas bibliotecas de células com circuitos combinacionais, utilizando transistores HVT e LVT, com o intuito de compará-las em termos de atraso e área ocupada;

• Determinar o MEP para a biblioteca RVT, em três temperaturas distintas e processo típico, para um determinado conjunto de circuitos de teste.

• Analisar o efeito de uma diversidade um pouco maior de células combinacionais no fluxo de síntese, para o mesmo conjunto de circuitos de teste;

• Comparar os resultados obtidos em termos de energia e frequência de operação, para os mesmos circuitos de teste, com um trabalho relacionado;

## 1.2 Organização da dissertação

A sequência desta dissertação está organizada da seguinte forma: O capítulo 2 apresenta uma série de técnicas no nível elétrico (circuito), arquitetural e de sistema para operação de CMOS digital em regimes de alimentação próximos à tensão de limiar do transistor encontradas na literatura com o intuito de aumentar a eficiência energética. Adicionalmente, são apresentadas algumas implementações utilizando transistores HVT e LVT sob o mesmo foco. O capítulo 3 apresenta a metodologia de projeto de três bibliotecas de células lógicas, com transistores RVT, HVT e LVT, para operação em VFS dinâmico desde o regime de inversão forte até tensões próximas ao limiar do transistor. O capítulo 4, primeiramente, demonstra a variação do ponto de mínima energia em função da temperatura para a biblioteca com transistores RVT, quando aplicada a um conjunto de circuitos de teste. Posteriormente são discutidos os efeitos de uma diversidade maior de células no fluxo de síntese em termos de número total de instâncias de células, energia consumida e frequências alcançadas. Por fim, os resultados de energia e desempenho obtidos neste estudo são comparados a trabalho relacionado, utilizando a mesma tecnologia CMOS, os mesmos circuitos de teste e parâmetros de simulação. O capítulo 5 conclui o trabalho e indica possibilidades para o seu aprimoramento.

# 2 REVISÃO BIBLIOGRÁFICA

## 2.1 Introdução

Neste capítulo, basicamente, serão apresentadas algumas técnicas de baixa potência implementadas em circuitos CMOS, utilizando diferentes estilos lógicos e escalamento da tensão de alimentação até regimes de operação próximos à tensão de limiar para redução de consumo. Adicionalmente melhorias em circuitos de memória operando em NTV serão discutidas. Posteriormente, serão apresentadas algumas técnicas arquiteturais como redundância e paralelismo, de forma a lidar com variabilidades de processo e atrasos inerentes à operação em NTV. Implementações de técnicas no nível de sistema considerando trabalhar em regimes de *near*-VT em blocos de banda base digital também são apontadas. Por fim, no âmbito das aplicações utilizando transistores multi-limiar, serão apresentados trabalhos que os utilizam combinadamente com técnicas de posicionamento, dimensionamento de portas, algoritmos para troca de transistores e novas metodologias de síntese com o intuito de reduzir o consumo de energia.

## 2.2 Implementações de técnicas no nível elétrico (circuitos)

Em (VIRGA et al., 2014) é realizada uma comparação de somadores completos (*full adders*) de um *bit* em lógica estática complementar (CMOS) e em lógica diferencial, como a *Differential Cascode Voltage Switch Logic* (DCVSL). No trabalho, foi explorada a variação da tensão de limiar (de 0 a 20% em sete passos), e o escalamento (redução) da tensão de alimentação (de 1,0 a 0,3 V em passos de 100 mV). Em função da metodologia adotada, foram determinados os pontos ótimos de operação para as duas lógicas, considerando o compromisso entre energia consumida e frequência de operação. Considerando a tensão de alimentação de 0,8 V, a lógica CMOS foi, no mínimo, 2 vezes mais rápida que a lógica diferencial, consumindo 50% menos. Tal condição foi sendo invertida à medida que a tensão de alimentação foi aproximando-se da tensão de limiar. Abaixo de VT, a lógica DCSVL apresentou um desempenho superior ao dobro do desempenhado pelo somador CMOS. Entretanto, por volta de 0,3 V e, considerando uma variação de 10% na tensão de limiar, a

lógica diferencial retornou à segunda colocação no quesito desempenho e potência consumida. Tal fato ocorreu em função da lógica CMOS apresentar maior robustez relativa quando a variabilidade é introduzida. O autor justifica que esta perda de desempenho e consumo superior é responsabilidade (50 a 98%, dependendo da variação da tensão de alimentação e da tensão de limiar) dos transistores PMOS. Desta forma, foi apresentada uma análise de dimensionamento para rede *pull-up* com o intuito de atenuar os efeitos da variabilidade na mesma e, paralelamente, aumentar o desempenho. Os autores concluíram que existem duas razões (entre as redes *pull-up* e *pull-down*) ótimas para operação em baixas tensões: uma direcionada ao desempenho (2:1) e outra com maior robustez na presença de variabilidade (1:1). No caso da razão 1:1, a lógica diferencial foi até 27% mais rápida do que a lógica complementar. Em termos de consumo para uma mesma tensão, a lógica complementar apresentou, para todos os casos, maior eficiência energética. Entretanto, para situações críticas no ponto de vista de desempenho (*e.g.* circuitos digitais de alto desempenho para circuitos integrados, como as CPUs), o autor sugere o uso da lógica diferencial de modo a atenuar um dos principais problemas de operar no regime de *near-VT*: o aumento exponencial do atraso. Naquelas situações, um consumo adicional de energia pode ser tolerado, obviamente sendo ainda muito inferior do que em condições de operação em inversão forte. O modelo de transistores utilizado naquele trabalho foi um PTM (*Predictive Technology Model*) para aplicações de alto desempenho de 45 nm da (ASU, 2008).

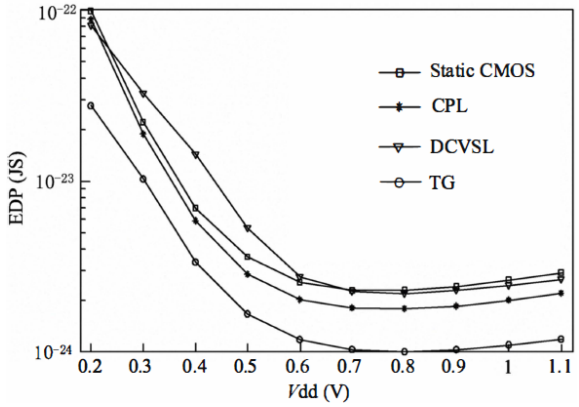
Outro trabalho que explora a utilização de somadores completos implementados em diversos estilos lógicos para aplicações de baixo consumo é realizado por Hu e Yu (2010). Basicamente, os autores apresentam equacionamentos para atrasos de propagação em regimes de inversão forte e fraca e combinam tais formulações na métrica de EDP (*Energy-Delay Product*) de modo a chegar à conclusão de que para reduzir o consumo, nada é mais eficiente do que reduzir a tensão de alimentação do circuito, uma vez que a energia de chaveamento é reduzida quadraticamente e a energia de *leakage* é reduzida exponencialmente com a redução da tensão. Segundo os autores, antagonicamente ao regime de operação em inversão forte, reduzir a tensão de alimentação até a condição de sub-limiar proporciona as melhores reduções de consumo. Entretanto, limita o desempenho (*performance*) para aplicações em uma faixa de 50 KHz a 5 MHz (BOL; FLANDRE; LEGAT, 2009). Posteriormente, Hu e Yu (2010) defendem um equilíbrio entre desempenho e economia de energia ao direcionar os seus somadores a operarem com tensões intermediárias. Tipicamente uma abordagem em concordância com a métrica de EDP.

Após uma breve análise das condições de operação em sub-limiar, super-limiar e da métrica de EDP, Hu e Yu (2010) apresentam uma revisão sobre os estilos lógicos utilizados em seus somadores completos: CMOS, DCSVL, CPL (*Complementary Pass-transistor Logic*) e TG (*Transmission-Gate*). O dimensionamento dos transistores para cada lógica é apresentado e brevemente discutido.

Posteriormente foram realizadas comparações entre as quatro lógicas através de simulações em HSPICE utilizado um modelo preditivo (PTM) de 65 nm, tendo a sua tensão de alimentação variada de 0,2 a 1,1 V em passos de 10 mV. Com o intuito de realizar uma comparação mais justa, os autores duplicaram a carga na saída das lógicas *single-rail* (CMOS e TG) de modo a apresentar capacitâncias de carga similares às experimentadas nas lógicas *dual-rail* (DCSVL e CPL), bem como submeteram o mesmo padrão de entrada às quatro lógicas do trabalho. Quanto aos atrasos de propagação, a lógica CMOS apresentou um atraso na ordem de 1,6 vezes superior às outras lógicas, enquanto que os *transmission-gates* foram os mais rápidos. A máxima frequência de operação variou muito pouco para cada lógica. Quanto ao consumo de energia, os autores constataram que a lógica de *transmission-gates* apresentou o menor consumo energético em todas as tensões de operação analisadas, enquanto que a DCVSL foi a lógica que consumiu mais energia em função de que em momentos de transição, as redes *pull-up* e *pull-down* estão ligadas simultaneamente, produzindo um curto-circuito entre $V_{DD}$ e o terra. No caso da métrica de eficiência adotada no referido trabalho, a lógica CMOS alcançou o menor EDP na tensão de alimentação de 700 mV, enquanto que as outras três lógicas atingiram a melhor eficiência em 800 mV. Tais valores ótimos representaram uma redução no EDP de 21,23%, 16,98%, 19,36% e 15% em relação a operação em 1.1 V para as lógicas CMOS, DCVSL, CPL e TG, respectivamente. A Figura 2.1 apresenta o comportamento do EDP para os quatros somadores completos em função da variação das tensões de alimentação. Adicionalmente, os autores salientam que o menor PDP (*Power-Delay Product*) foi alcançado pela lógica de *transmission-gates*. Por fim, os autores concluem que reduzir a tensão de alimentação é uma forma efetiva de reduzir a EDP para somadores completos, especialmente se mantiverem a tensão numa região média, entre 700 mV e 800 mV, de forma a prover a melhor eficiência em termos de EDP. Portanto, os autores não estão focados especificamente em eficiência energética e sim num ponto de melhor compromisso entre economia de energia e desempenho. Virga et al. (2014) acrescenta que os valores de energia apresentados no trabalho de Hu e Yu (2010) não são totalmente precisos porque não levam em consideração a variabilidade, por exemplo, da tensão de limiar,

fator que tem impacto significante nos atrasos, potências e energias consumidas pelos circuitos.

**Figura 2.1 - *Energy-Delay Product* para quatro somadores completos em quatro estilos lógicos**




**Fonte: (HU; YU, 2010)**

Em relação aos circuitos de memória, Chen et al. (2011) apresentam o projeto de uma SRAM em tecnologia CMOS 45 nm. Trata-se de um novo projeto com seis transistores (6T) que trabalha em regime de operação próximo a tensão de limiar (NTV) com o intuito de controlar a potência estática e manter um desempenho admissível. Em (SIL et al., 2008), os processos de leitura e escrita foram separados em dois blocos, de modo a reduzir a potência de chaveamento no momento da operação das linhas de bits (*bitlines*). Com esta abordagem, houve uma redução de atraso e potência no processo de escrita . Entretanto, o bloco separado para a leitura adicionou outro caminho para a corrente de *leakage*, que por sua vez, aumentou a potência dissipada. Em (WANG; LEE; LIN, 2007), foi proposto o esquema de linha de palavra negada (*negative word-line*) para reduzir a corrente de *leakage* no processo de leitura de uma SRAM, com o intuito de reduzir o consumo de potência em modo de espera (*standby*), principalmente. Basicamente, Chen et al. (2011) combinaram o trabalho de Wang et al. (2007) e de Sil et al. (2008) para compor a sua célula de memória (4T para escrita e 2T para leitura). Adicionalmente, um *sense amplifier* de tensão do tipo *latch* foi proposto para proporcionar um aumento adicional de desempenho na etapa de leitura. O projeto proposto foi comparado com uma célula 6T tradicional mediante simulações utilizando um modelo preditivo (PTM) CMOS de 45 nm em temperatura ambiente (25°C) com tensão de alimentação de 0,4 V. Os transistores do bloco de escrita foram dimensionados em tamanho mínimo e os do bloco de leitura com o dobro do tamanho mínimo. No caso da célula 6T

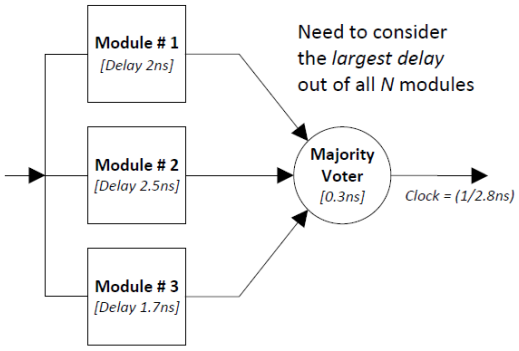
tradicional, tamanho mínimo para a rede *pull-up*, 2X para os transistores de passagem e 4X o tamanho mínimo para os transistores da rede *pull-down*. Após uma série de comparações, os autores salientaram que o projeto proposto reduziu o consumo de potência em 50%, corrente de *leakage* em 68%, atraso de escrita em 90% e atraso de leitura em 78%, quando comparados à célula tradicional de seis transistores em regime de operação próximo à tensão de limiar da tecnologia.

### 2.3 Implementações de técnicas no nível arquitetural (arquitetura e sistema)

Em (ASHRAF; ALZAHRANI; DEMARA, 2014) são discutidos os impactos da variabilidade de processo (PV - *Process Variation*), em sistemas que utilizam técnicas de redundância espacial de tamanho N (*N*MR - N *Modular Redundancy*) para mascarar *soft errors*[1] em caminhos lógicos (*logic paths*). Adicionalmente, os autores argumentam sobre o custo da redundância espacial em regime de NTV (*Near Threshold Voltage*). Segundo os autores, existem três mecanismos inerentes de mascaramento de um erro ao longo de um caminho lógico: mascaramento lógico, elétrico e de janela (*latching-window masking*). Entretanto, tais mecanismos tornam-se menos efetivos à medida que a tensão de operação diminui. Desta forma, a taxa de *soft errors* (SER - *Soft Error Rate*) pode ser reduzida utilizando-se técnicas como dimensionamento de portas ou múltiplos domínios de tensão. Contudo, estas técnicas aumentam área e consumo de potência e não garantem uma cobertura eficiente. Uma técnica que é muito utilizada, e eficiente, na redução de erros é a redundância espacial, especialmente a redundância modular tripla (TMR - *Triple Modular Redundancy*). Ela é utilizada em aplicações críticas como veículos autônomos, satélites e outros sistemas espaciais, bem como aplicações computacionais de alto desempenho. Basicamente, a técnica de redundância espacial envolve a replicação de N instâncias de um circuito e a saída é computada através de um elemento de votação. A Figura 2.2 apresenta um diagrama básico de uma TMR.

[1] Erros transientes que ocorrem, por exemplo, em células de memória fazendo com que um valor armazenado seja corrompido/alterado.

**Figura 2.2 - Ilustração de uma técnica de redundância modular tripla**




**Fonte: (ASHRAF; ALZAHRANI; DEMARA, 2014)**

A variação randômica na tensão de limiar é proeminente nos processos CMOS mais recentes e afeta severamente a estabilidade dos circuitos, bem como a distribuição de desempenho. Os principais contribuintes são as flutuações randômicas de dopantes (RDF - *Random Dopant Fluctuations*) e a rugosidade de bordas (LER - *Line-Edge Roughness*) (YE et al, 2011). Segundo Ashraf et al. (2014), o aumento de tais variações de processo implicam numa distribuição da tensão de limiar. Com o aumento da distribuição, haverá um aumento no tempo de chaveamento, o qual afetará o desempenho do circuito. Estes problemas são amplificados com a diminuição do nó tecnológico (YE et al, 2011). Adicionalmente, uma variação de 5% na tensão de alimentação (quando ela está próxima de VT) ou na tensão de limiar causam grandes impactos na frequência de operação de um circuito. Uma variação de 50% na frequência pode ser esperada quando em regime de operação em *near*-VT (KAUL et al., 2012).

No trabalho de Ashraf et al. (2014) somente os efeitos de variabilidade na tensão de limiar são analisados paralelamente com sistemas de redundância modular de tamanho N aplicados em caminhos de dados. Segundo os autores, sob condições nominais de operação, o consumo de energia de sistemas *N*MR é por volta de N vezes o consumo de sistemas sem redundância (N=1). Adicionalmente, é esperado que o pior caso em termos de atraso para sistemas de redundância múltipla exceda o pior atraso de sistemas simples. O estudo foi limitado à análise de sistemas redundantes de tamanho 3 e 5, comparados a sistemas sem redundância.

As variações de processo foram avaliadas através de simulações Monte Carlo no HSPICE para processos PTM de alto desempenho (nós de 22 e 45 nm) para um único *wafer* (*intra-die*). Os efeitos de RDF e LER foram modelados através da variação na tensão de limiar com desvio padrão de 25,9 mV para o processo de 45 nm e de 59,9 mV para o nó de 22

nm. Para uma cadeia de inversores, tomando-se como base um sistema sem redundância, houve uma redução de 10,6 vezes em *performance* na tensão de 0,5 V quando comparadas ao desempenho em tensão nominal. No caso de um acréscimo de 50 mV na tensão de alimentação, a redução de desempenho caiu para 6,29 vezes em relação à situação de inversão forte. No caso de sistemas com N=3 e N=5, não houve alterações significativas, apenas um leve aumento no atraso em função do aumento de redundância. Entretanto, houve um espalhamento na média dos atrasos entre sistemas simples e com redundância à medida que a variabilidade aumentou ao aproximar-se da tensão de limiar. No processo de 22 nm houve um aumento na diferença de desempenho entre sistemas redundantes e simples em função da variabilidade aumentar com a diminuição do nó tecnológico. Para N=3 o atraso médio foi de 1,16 vezes e, no caso de N=5, aumentou para 1,24 vezes quando comparado ao atraso médio de sistemas sem redundância para a tensão de 0,55 V. Como critério de comparação, no caso do processo de 45 nm, os atrasos médios foram de 1,06 e 1,09 para N=3 e N=5, respectivamente. É importante salientar que a quantidade de variabilidade depende do comprimento do caminho lógico (número de portas lógicas no caminho crítico). Quanto maior for o caminho de uma cadeia de inversores, por exemplo, menor será a variabilidade. Entretanto, em regimes de NTV, o comprimento do caminho de dados não pode ser muito longo em função da redução de desempenho inerente deste tipo operação. Adicionalmente, a quantidade de variação é dependente do tipo de lógica utilizada. Caso sejam utilizadas NAND2 (com suas entradas conectadas) em vez de inversores convencionais, haverá uma menor ocorrência de variação. No caso de uma redundância modular tripla, em 22 nm, houve uma redução de 13% na variação quando foram utilizadas portas NAND de duas entradas em vez de inversores. Em relação à energia consumida, sistemas com redundância quíntupla em 22 nm necessitam, em média, de 3,94% mais energia do que o mesmo sistema em 45 nm. Tais experimentos foram realizados numa cadeia de 26 inversores com um *fanout* de 4 inversores (FO4). A Tabela 2.1 apresenta a comparação de consumo entre sistemas sem redundância e sistemas com redundâncias triplas e quíntuplas para tensões próximas do limiar (NTV).

**Tabela 2.1 - Impacto da redução de tecnologia no consumo de energia em sistemas *N*MR mantendo o mesmo desempenho de sistemas sem redundância em tensões próximas do limiar**

| | 45nm | | 22nm | |
|---|---|---|---|---|
| $N = 1, V_{DD}$ (NTV) | N=3 | N=5 | N=3 | N=5 |
| $0.5V$ | 3.04X | 5.07X | 3.17X | 5.30X |
| $0.55V$ | 3.03X | 5.05X | 3.14X | 5.27X |
| $0.6V$ | 3.03X | 5.04X | 3.13X | 5.26X |
| $0.65V$ | 3.02X | 5.03X | 3.12X | 5.23X |
| $0.7V$ | 3.01X | 5.02X | 3.10X | 5.16X |

**Fonte: (ASHRAF; ALZAHRANI; DEMARA, 2014)**

Por fim, o trabalho de Ashraf et al. (2014) avalia o custo do aumento de confiabilidade em NTV. Os autores demonstram que um sistema de redundância tripla operando a 690 mV consome a mesma energia que um sistema sem redundância operando na tensão nominal (1.1 V), resultando numa diferença de atraso de 2,58 vezes. No caso de uma resiliência quíntupla, para as mesmas condições, é possível operar a 545 mV com uma redução de 7,15 vezes em desempenho. Se for adotada uma TMR operando em 550 mV, haverá uma redução de consumo de 38,4% em relação a um sistema sem redundância operando em inversão forte. Desta forma, haverá um mascaramento de *soft errors* operando em NTV enquanto ocorre uma redução substancial de consumo energético. É importante salientar que o trabalho de Ashraf et al. (2014) restringiu-se à análise da variação da tensão de limiar como a única fonte de variabilidade. Outras fontes como variação na tensão de alimentação, temperatura e envelhecimento (*aging*) serão investigadas em trabalhos futuros.

Em (KRIMER et al., 2010) é proposta uma família de processadores de *stream*[2] que se baseia em alto nível de paralelismo, incorporando cooperativamente técnicas de circuito e arquiteturais para tolerar as grandes variações de atraso inerentes à operação em NTV.

Tal família foi denominada de Synctium. Basicamente, trata-se de um processador de *stream* paralelo que opera em NTV e alcança eficiência energética próxima de seu valor ótimo, com alta taxa de transferência, baseando-se em circuitos paralelos, de baixa frequência, e próximos da tensão de limiar. Em função da baixa frequência de operação, este processador possui uma grande quantidade de unidades lógico-aritméticas. Sua arquitetura é tradicional, com multi-núcleos e *wide*[3] SIMD (*Single Instruction, Multiple Data*). Cada núcleo possui 16 elementos de processamento (PE - *Processing Elements*) compostos por um sequenciador de instruções e 16 faixas (*lanes*) de execução, que acessam cada uma, 16 KB da memória local.

[2] Conjunto de dados

[3] Capaz de realizar desvios, cargas e armazenamentos de 128 ou 256 bits de uma única vez

Os autores salientam que tal arquitetura não representa grandes inovações, entretanto, a maneira com que os desafios de lidar com extremas variações temporais estáticas e dinâmicas no regime de operação próximo à tensão de limiar são abordados, justificam o seu trabalho. São propostos dois mecanismos para redução de variação temporal em arquiteturas paralelas: DPSP (*Decoupled Parallel SIMD Pipelines*) e *Pipeline Weaving*. A primeira lida com especulação temporal em *pipelines* paralelos, provendo tolerância de variação dinâmica da entrada e otimizando a taxa de transferência média, enquanto a segunda fornece redundância espacial eficiente de grão fino dentro do *pipeline* paralelo para compensar variações estáticas.

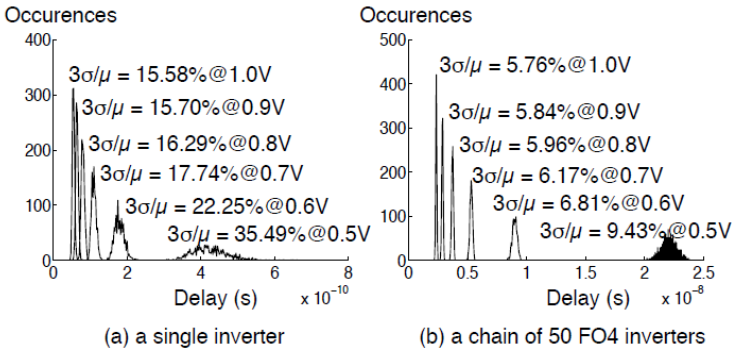
O trabalho de Krimer et al. (2010) é preliminar. Estimativas de área e consumo foram feitas para um futuro processador em 16 nm, ocupando 4 $mm^2$ de área com um desempenho de 30 bilhões de operações (16 bits) por segundo (aproximadamente 560 fJ / operação), consumindo menos que 17 mW. Adicionalmente, pretendem prototipar em 90 nm o mesmo processador, ocupando a mesma área e alcançando até 640 milhões de operações por segundo (1,2 pJ/op.), consumindo menos de 1 mW e trabalhando a 10 MHz. Maiores detalhes sobre a implementação proposta estão em (KRIMER et al., 2010).

Outro trabalho que explora a operação próxima à tensão de limiar juntamente com paralelismo em nível de arquitetura é apresentado em Seo et al. (2012). Os autores realizam um estudo aprofundado sobre as variações de atraso quando em operação em NTV e demonstram que técnicas como duplicação estrutural, tolerâncias na tensão de alimentação e nas frequências de operação são suficientes para redução destes atrasos em arquiteturas *wide* SIMD ao custo de um pequeno aumento de área e consumo de energia.

Primeiramente, foi realizado um estudo de variabilidade no nível elétrico (circuito) em regime de NTV. Desta forma, simulações Monte Carlo no HSPICE foram realizadas utilizando modelos de transistores para processos CMOS industriais de propósito geral (GP) de 90 e 45 nm e também realizadas em modelos preditivos PTM de alta *performance* (HP). As variações da tensão de limiar e rugosidade de bordas (LER), efeitos mais severos em nós tecnológicos avançados, foram representadas como distribuições normais e consideradas nos modelos preditivos. A Figura 2.3 apresenta as distribuições de atraso em função da variação da tensão de alimentação para um único inversor e para uma cadeia de 50 inversores com FO4. Percebe-se, por exemplo, para o caso da tensão 0,5 V, que a houve uma redução (de 35,49% para 9,43%) considerável na variação do atraso no caso da utilização da cadeia de inversores. Desta forma, o problema de variabilidade no atraso dos caminhos lógicos é de menor magnitude, quanto mais profundo for o caminho (em número de portas lógicas

CMOS). Entretanto, foi sugerido em (ASHRAF; ALZAHRANI; DEMARA, 2014) que a profundidade do *pipeline* não fosse elevada de modo a recuperar a taxa de transferência perdida em operações próximas à tensão de limiar. Adicionalmente, foi demonstrado em (SEO et al., 2012) que a variabilidade de atraso aumentou aproximadamente 2,5 X quando o nó tecnológico foi de 90 para 22 nm para a mesma cadeia de 50 inversores, operando a 550 mV. Além disso, a variação de atraso, quando a tensão é reduzida de 1 V para 0.5 V, é de somente 4% em 90 nm. Entretanto, para 22 nm, esta variação aumenta para 14% quando a tensão é reduzida de 0,8 (tensão nominal) para 0,5 V.

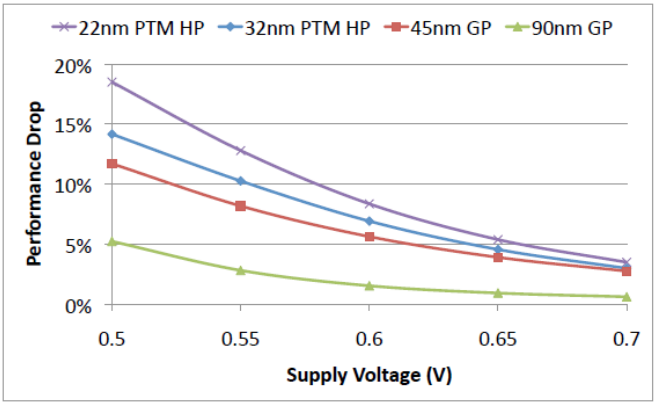
**Figura 2.3 - Distribuições de atraso de um único inversor (a) e de uma cadeia de 50 inversores com FO4 (b) em diferentes tensões de alimentação para um modelo de 90 nm**




**Fonte: (SEO et al., 2012)**

Além do estudo de variabilidade em nível de circuito, Seo et al. (2012) apresentou resultados sobre variabilidade em nível arquitetural. Utilizou como arquitetura alvo um processador para câmeras digitais denominado Diet SODA (SEO et al., 2010). Trata-se de uma arquitetura 128-*wide* SIMD. Como critério de simplificação, foi utilizado uma cadeia de 50 inversores (FO4) para emular o caminho crítico do caminho de dados (SEO et al., 2012). A Figura 2.4 ilustra a redução de desempenho para os quatro nós tecnológicos discutidos neste trabalho quando a tensão de alimentação aproxima-se da tensão de limiar. No caso do modelo de 90 nm, a queda de desempenho é da ordem de 5% em 0,5 V em relação à 1 V, enquanto que no modelo de 22 nm a redução é da ordem de 18% para a mesma tensão. Desta forma, os autores afirmam que não são necessárias grandes alterações arquiteturais para lidar com as variações de atraso.

**Figura 2.4 - Queda de desempenho (%) em NTV para a arquitetura 128-wide SIMD para quatro nós tecnológicos**




**Fonte: (SEO et al., 2012)**

A primeira abordagem de Seo et al. (2012) para amenizar as variações de atraso é a duplicação estrutural. Esta técnica consiste em adicionar unidades micro-arquiteturais sobressalentes com o intuito de servir como unidades reparadoras em caso de falhas em tempo de execução na unidade principal. Os autores analisaram o número ótimo de unidades sobressalentes para o regime de operação em NTV na sua arquitetura 128-*wide* SIMD. Para o caso do nó de 90 nm, seis unidades adicionais é o número ótimo. A Tabela 2.2 apresenta o número de unidades sobressalentes e seus respectivos aumentos de área e consumo para os quatro nós tecnológicos de forma a cumprir os requisitos de atraso.

**Tabela 2.2 - Número necessário de unidades sobressalentes e respectivos aumentos de área e consumo para a arquitetura 128-wide SIMD em quatro nós tecnológicos**

| | 90nm | | | 45nm | | | 32nm | | | 22nm | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Vdd** | **spares** | **area ovhd.** | **power ovhd.** | **spares** | **area ovhd.** | **power ovhd.** | **spares** | **area ovhd.** | **power ovhd.** | **spares** | **area ovhd.** | **power ovhd.** |
| **0.50V** | 28 | 12.1% | 4.6% | >128 | > 57.8% | > 25.0% | >128 | > 57.8% | > 25.0% | >128 | > 57.8% | > 25.0% |
| **0.55V** | 6 | 2.6% | 1.0% | 84 | 37.2% | 15.3% | >128 | > 57.8% | > 25.0% | 80 | 35.3% | 14.5% |
| **0.60V** | 2 | 0.9% | 0.3% | 26 | 11.2% | 4.3% | 48 | 20.9% | 8.2% | 22 | 9.5% | 3.6% |
| **0.65V** | 1 | 0.4% | 0.2% | 10 | 4.3% | 1.6% | 12 | 5.1% | 1.9% | 7 | 3.0% | 1.1% |
| **0.70V** | 1 | 0.4% | 0.2% | 4 | 1.7% | 0.6% | 6 | 2.6% | 1.0% | 3 | 1.3% | 0.5% |

**Fonte: (SEO et al., 2012)**

A segunda abordagem de Seo et al. (2012) para amenizar as variações de atraso sem aumentar o período de relógio baseia-se em adicionar tolerâncias (margens) na tensão de alimentação. Tal abordagem justifica-se pelo aumento exponencial do atraso à medida que a tensão de alimentação decresce. Desta forma, um pequeno aumento da tensão na região próxima à tensão de limiar pode contribuir com a diminuição das variações temporais. A Tabela 2.3 informa as tensões de alimentação, margens adicionadas à mesma e correspondentes sobrecargas de consumo para os quatro nós do referido trabalho. No caso do

maior nó, é necessário adicionar apenas 5,8 mV quando em operação à 500 mV. Entretanto, para o menor nó, é necessário adicionar aproximadamente 16,4 mV aos 500 mV. Os autores salientam que à medida que as variações temporais aumentam em regime de operação em NTV a abordagem de adicionar margens na tensão de alimentação é mais eficiente energeticamente do que a técnica de duplicação estrutural.

**Tabela 2.3 - Margens na tensão de alimentação para tolerar erros de temporização em função da variabilidade para a arquitetura 128-wide SIMD para quatro nós tecnológicos**

| | 90nm | | 45nm | | 32nm | | 22nm | |
|---|---|---|---|---|---|---|---|---|
| **Vdd** | **Vdd margin** | **power ovhd.** | **Vdd margin** | **power ovhd.** | **Vdd margin** | **power ovhd.** | **Vdd margin** | **power ovhd.** |
| **0.50V** | 5.8 mV | 1.0% | 19.6 mV | 3.3% | 12.1 mV | 2.0% | 16.4 mV | 2.8% |
| **0.55V** | 4.1 mV | 0.6% | 18.2 mV | 2.8% | 11.1 mV | 1.7% | 17.6 mV | 2.7% |
| **0.60V** | 2.9 mV | 0.4% | 16.2 mV | 2.3% | 10.4 mV | 1.5% | 11.1 mV | 1.6% |
| **0.65V** | 2.2 mV | 0.3% | 14.0 mV | 1.8% | 8.9 mV | 1.1% | 11.5 mV | 1.5% |
| **0.70V** | 1.7 mV | 0.2% | 12.8 mV | 1.5% | 7.7 mV | 0.9% | 9.6 mV | 1.1% |

**Fonte: (SEO et al., 2012)**

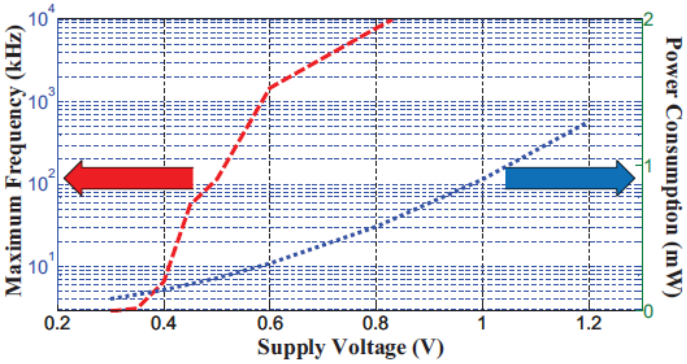
A última abordagem de Seo et al. (2012) para amenizar as variações de atraso em regime de operação em NTV baseia-se em adicionar tolerâncias na frequência de operação do sistema. Ou seja, basicamente, trata-se de aumentar o período de relógio em situações onde as restrições de temporização sejam mais relaxadas, entretanto, cumprindo os requisitos de tempo. No entanto, após experimentações, os autores detectaram a necessidade de um aumento de 20% nas margens de atraso em nós mais avançados. Tais valores inviabilizam a utilização desta técnica para lidar com a variabilidade de temporização.

Em suma, o trabalho de Seo et al. (2012) demonstra que para o modelo de 90 nm, apenas a técnica de duplicação estrutural é suficiente para lidar com os erros de variabilidade de temporização em arquiteturas *wide* SIMD. À medida que o nó tecnológico diminui, os autores recomendam a utilização combinada das técnicas de duplicação estrutural e o acréscimo de margens na tensão de alimentação para alcançar a menor sobrecarga de consumo.

Em (ZHAO et al., 2013) é apresentado um *System-on-Chip* (SoC) para aplicações em redes de área corporal (WBAN - *Wireless Body Area Network*). São integrados neste sistema um transceptor de RF, unidades de processamento digital, uma unidade micro-controlada, um conversor AD (*Analog-to-Digital*) de 10 bits, entre outros. Os autores apresentam uma série de aperfeiçoamentos no transceptor e nas técnicas de modulação com o intuito de reduzir o consumo de energia necessário em aplicações à que se destina este tipo de rede corporal: implantes de retina, cápsulas endoscópicas e sistemas de gravação neurais. Além das melhorias na parte de radiofrequência (RF), foram contempladas possibilidades de operação

em proximidade com a tensão de limiar (NTV) no projeto do bloco da banda base digital. Para lidar com os problemas de processo, tensão e temperatura inerentes à operação em NTV, os autores retiraram algumas *standard cells* de tamanho mínimo, bem como transistores empilhados (*multi-stack*), justificando que são vulneráveis às variações de PVT. A eficiência energética em condições normais de operação (tensão nominal de 1,2 V) é da ordem de 130 pJ/bit e atinge a maior eficiência energética na tensão de 0,55 V (34,8 pJ/bit). Adicionalmente, os autores afirmam que a parte digital pode operar de forma robusta em 0,4 V, em condições de baixo desempenho, consumindo cerca de 30 µW. A Figura 2.5 relaciona a frequência de operação e potência consumida em função da tensão de operação para a parte digital.

**Figura 2.5 - Consumo de potência e frequência de operação da banda base digital de um SoC para aplicações em WBAN**




**Fonte: (ZHAO et al., 2013)**

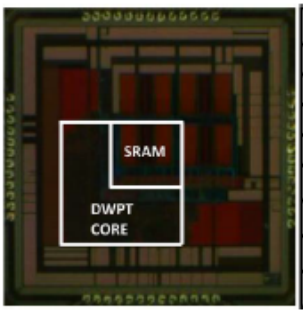
No bloco de RF, o receptor consome 2,14 mW (eficiência de 0,214 nJ/bit) com uma sensitividade de -65 dBm, enquanto o transmissor consome 2,85 mW (eficiência de 0,285 nJ/bit) a 10 Mb/s com uma potência de saída de -5,4 dBm.

O SoC foi fabricado em CMOS, 130 nm, com dimensões aproximadas de 3,4 mm x 2,5 mm, incluindo *buffers* de teste e *pads* de E/S. O sistema pode ser alimentado por uma bateria (*button battery*) de 1,5 V.

Em (WANG et al., 2015) é apresentado o projeto de um processador DWPT (*Discrete Wavelet Packet Transform*) para aplicações de monitoramento de saúde. Com o intuito de trabalhar com eficiência energética, primordial para a aplicação alvo, os autores utilizam diversas técnicas de projeto, em nível de algoritmo ao nível de circuitos, como: computação reconfigurável, esquema de *lifting*, processamento de *pipeline* em duas portas, operação próxima a tensão de limiar (NTV) e *clock gating*.

Uma tecnologia de 180 nm CMOS padrão foi utilizada para implementação do processador DWPT. O processador possui três domínios de alimentação: 3,3 V para os *pads* de E/S, 1,8 V (nominal) para memória e 0,6 V para o núcleo do processador. Foram selecionadas da biblioteca de *standard cells* padrão, células lógicas com atrasos e variações pequenas em tensões ultra baixas. Adicionalmente, foram utilizados conversores de nível (*level shifters*) otimizados para realizar conversões rápidas e energeticamente eficientes entre a lógica do núcleo do processador, memórias e *pads* de E/S. O circuito foi prototipado em 180 nm, ocupando uma área total de 0,95 $mm^2$, das quais 0,53 $mm^2$ são referentes à área ocupada pelo processador. O núcleo pode operar funcionalmente de 1,8 a 0,5 V, reduzindo o consumo de potência em aproximadamente dez vezes, até o mínimo de 26 µW. O consumo médio da memória SRAM em 1,8 V é de 102,5 µW, totalizando uma potência mínima de 128,5 µW. Os autores estimam que este consumo poderia ser reduzido para aproximadamente 40,6 µW se a memória comercial utilizada fosse substituída por uma SRAM trabalhando a 0,5 V. A Figura 2.6 apresenta a foto do *die* do processador DWPT, bem como apresenta um resumo das especificações do projeto.

**Figura 2.6 - Foto do *die* do processador DWPT + SRAM e resumo do chip de teste prototipado em 180 nm**




| Test Chip Summary | |
|---|---|
| Technology | 0.18µm |
| Voltage | 0.5~1.8V |
| Data width | 16-bit |
| Frequency | 1~20MHz |
| Gate count | 14K |
| SRAM | 512Byte |
| Power | 128.5µW |
| Area | 0.95mm² |

**Fonte: (WANG et al., 2015)**

Os autores integraram seu processador numa plataforma de testes baseada num SoC ARM Cortex-M0 onde obtiveram acelerações de três ordens de magnitude no processamento, com reduções de quatro ordens de magnitude em termos de consumo de energia quando comparadas com implementações baseadas apenas em CPU. Resultados de desempenho e consumo foram estimados para o caso de decomposições de um sinal randômico de 256 pontos. Para esta condição, o SoC pode operar a 20 MHz/1,8 V ou a 2,2 MHz em 0,6 V (tensão mínima de operação). Entretanto, o núcleo pode trabalhar em 0,5 V a uma frequência de 1 MHz. Após uma série de revisões da literatura e subsequentes comparações, os autores salientam que a sua implementação é inovadora (primeiro projeto de processador

reconfigurável que trabalha em regime de operação próximo a tensão de limiar) e eficiente tanto em consumo de energia quanto em área ocupada e que pode ser utilizada para aplicações de monitoramento de saúde.

## 2.4 Implementações utilizando transistores multi-limiar

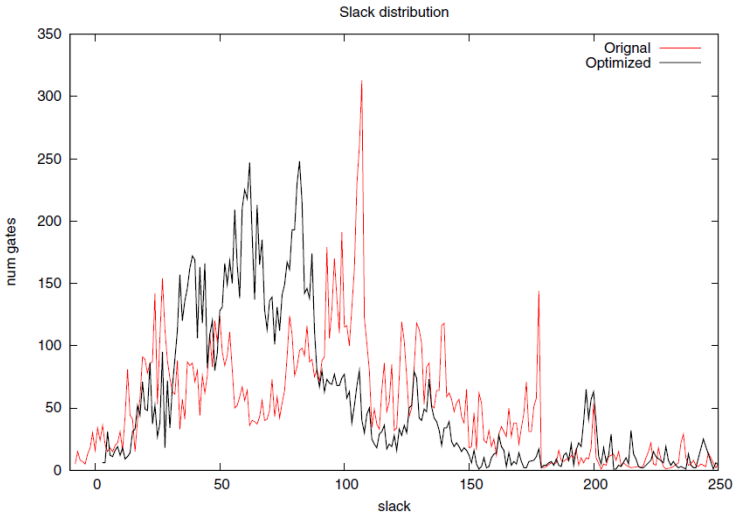
No trabalho de Luo et al. (2008) é proposto um novo fluxo de otimização de potência total sob restrição de desempenho. Para isto eles combinam técnicas de posicionamento (*placement*), dimensionamento de portas e dispositivos multi-limiar através do conceito de gerenciamento de distribuição de *slack* para maximizar a redução de potência durante a síntese física.

Conforme os autores, a etapa de posicionamento é usada, tradicionalmente, para otimizações de temporização, uma vez que potência e temporização são, normalmente, objetivos conflitantes no processo de otimização. Além disso, não existem metodologias de posicionamento que consideram redução da potência de *leakage*. Adicionalmente, as metodologias de dimensionamento de portas minimizam os piores casos de atraso ou minimizam potência sob restrições de desempenho. Entretanto, tais metodologias de dimensionamento não colaboram com o algoritmo de troca de transistores com limiares distintos. Estes algoritmos são efetivos na redução da corrente de *leakage*.

Basicamente, Luo et al. (2008) desenvolveram algoritmos de gerenciamento de distribuição de *slack* com o objetivo de vincular as etapas de posicionamento e dimensionamento de transistores para acelerar a técnica de troca de transistores com limiares distintos. Eles aumentam as somas de *slacks* nos caminhos críticos e também dos caminhos próximos aos críticos de modo a modificar a curva de distribuição de *slacks* para longe do crítico. Desta forma, os autores trocam um pequeno aumento na potência dinâmica (aumentando o número de transistores HVT, inclusive aumentando o seu tamanho superficialmente) por uma grande redução na potência de *leakage* (reduzindo a quantidade de transistores LVT). Adicionalmente, eles diminuem a potência reduzindo células que não estão no caminho crítico. O fluxo de otimização começa com a utilização de transistores padrão (RVT). Nos caminhos que apresentam violação de tempo, e são difíceis de otimizar, os transistores RVT são substituídos por transistores LVT. Todos os transistores RVT que estão em caminhos não-críticos com *slack* amplo são substituídos por transistores HVT para economizar energia. Ao reduzir o número de células em situação de *slack* crítico, menor será o número de transistores LVT utilizados e, consequentemente, maior será a probabilidade de utilização de transistores HVT. A Figura 2.7 apresenta um histograma com a distribuição de

*slacks* para um circuito baseado em transistores padrão (RVT) antes e depois do posicionamento mais dimensionamento de portas. É possível observar que após o processo de otimização, a distribuição de *slack* ficou mais estreita e ao redor de uma média menor. Com uma maior quantidade de células em situação de slack reduzido (não crítico), menor será a necessidade de células LVT, consequentemente, menor será o consumo devido a *leakage*.

**Figura 2.7 - Distribuição de slack antes e após otimização**




**Fonte: (LUO; NEWMARK; PAN, 2008)**

Luo et al. (2008) realizaram seus experimentos de otimização em uma série de circuitos de um micro-processador de 65 nm e constataram que as técnicas combinadas de posicionamento, dimensionamento de transistores e substituição de transistores com limiares distintos obtiveram os melhores resultados em relação às técnicas tradicionais, e que desta forma, colaboraram para uma redução de 63,8% da potência estática e 32,9% em termos de potência total.

Em (CALIMERA et al., 2008) é proposta uma nova metodologia de síntese para uma biblioteca de baixo consumo de energia, que leva em consideração o fenômeno ITD (*Inverted Temperature Dependence*). Este fenômeno causa uma redução nos atrasos de uma célula à medida que a temperatura aumenta. Os autores salientam que ferramentas tradicionais de síntese podem incorrer em erros de temporização por não contemplar os efeitos de temperatura no processo de otimização de dispositivos multi-limiar, uma vez que, tipicamente, tais ferramentas utilizam bibliotecas de células que foram caracterizadas em uma única temperatura, e assumem, para a maior temperatura possível no processo, a situação de pior caso de atraso. Os possíveis erros de temporização podem ser contornados pelo projetista,

que normalmente adota uma abordagem conservativa, relaxando os requisitos de temporização. Entretanto, segundo os autores, tal abordagem poderá aumentar a área e o consumo do projeto, desnecessariamente. Os autores demonstram como aproveitar com maior eficiência o *slack* disponibilizado pela ferramenta da síntese de modo a reduzir *leakage* sem deixar de garantir o cumprimento de temporização em função dos *corners* de temperatura. No referido trabalho, foram caracterizadas funções lógicas tradicionais em 25ºC e 125ºC, em uma biblioteca contendo somente transistores HVT e LVT de 65 nm da STMicroelectronics. Para validar a metodologia proposta, um novo fluxo de síntese baseado em ferramentas comerciais foi configurado e aplicado a um conjunto de circuitos combinacionais do ISCAS *Benchmark* (HANSEN; YALCIN; HAYES, 1999). Os resultados obtidos em seu fluxo de síntese resultaram numa redução média de 27% na potência estática em relação ao fluxo tradicional. Adicionalmente, os autores salientam que podem garantir se um determinado circuito cumprirá os requisitos de temporização em ambas condições de contorno de temperatura pré-determinadas.

# 3 DESENVOLVIMENTO DE BIBLIOTECAS DE CÉLULAS CMOS PARA OPERAÇÃO A BAIXO $V_{DD}$

## 3.1 Introdução

Pesquisa sobre biblioteca de células digitais a baixo $V_{DD}$, realizada na UFRGS por Stangherlin (2013), propôs uma nova metodologia de dimensionamento de portas lógicas para operação próxima da tensão de limiar (*near-VT*). Essa se baseia em ajustar a largura dos transistores de modo a equalizar os tempos de subida e descida, na saída das portas lógicas. O autor citado mostrou que este critério igualmente propicia a obtenção de margens estáticas de ruído (SNM) mais equilibradas (em valores lógicos *low* e *high*) e adequadas para baixa tensão, além de reduzir os efeitos da variabilidade, prejudiciais principalmente em $V_{DD}$ baixo. O critério de equalização dos tempos de transição (*Trise* e *Tfall*) é adotado também neste trabalho em função da importância destes tempos de transição na potência dinâmica de curto-circuito, a qual aumenta linearmente com a duração destas transições nas entradas das portas lógicas. Nos resultados de simulação das células lógicas obtidos neste capítulo, considerada a tensão de alimentação a 300 mV, os tempos de transição com *fan-out* 4 são apreciáveis (de 100 ns ou mais para os transistores RVT, de 1 a 5 µs para os transistores HVT e de dezenas de nanossegundos para os transistores LVT que foram utilizados no *design* das células lógicas). Adicionalmente, é permitido que as portas lógicas possuam qualquer *driving strength*, desde que a razão entre os tempos de subida e descida seja próxima a 1,0. Por fim, para gerar células de *strengths* maiores é necessário, tão somente, multiplicá-las por uma constante. A Tabela 3.1 apresenta a biblioteca de células desenvolvida por Stangherlin (2013). Para a operação em NTV, opta-se por empregar apenas células lógicas com no máximo dois transistores em série na rede PMOS e/ou na rede NMOS das portas CMOS.

**Tabela 3.1 - Biblioteca de células *near-VT* desenvolvida por Stangherlin (2013)**

| Célula | X1 | X2 | X3 | X4 | X8 |
|---|---|---|---|---|---|
| INV | • | • | • | • | • |
| NAND2 | • | • | | • | |
| NOR2 | • | • | | • | |
| DFFR | • | • | | • | |
| DFFS | • | • | | • | |

**Fonte: (STANGHERLIN, 2013)**

Na biblioteca de células do trabalho mencionado acima os transistores foram dimensionados para a tensão de operação de 450 mV. O autor demonstrou os benefícios em termos de redução de consumo quando comparados à operação muito acima de VT para alguns circuitos VLSI de teste com média complexidade (da ordem de 10K a 30K portas lógicas). Adicionalmente, tais células foram projetadas de tal maneira que poderiam operar desde condições de baixo consumo, em *near*-VT a 450 mV, até regimes de trabalho em tensão nominal, se necessário. Esta extrapolação das técnicas convencionais de ajuste de tensão e frequência (VFS) dinâmico é denominada de *very wide Voltage-Frequency Scaling* por Stangherlin e Bampi (2013). Entretanto, foi demonstrado que para os circuitos testados, o ponto de mínima energia por operação (MEP) situava-se entre 260 e 310 mV. Segundo De (2013), a operação em *near-threshold voltage* (NTV) de um projeto CMOS é definida como o ponto de tensão e frequência onde a energia consumida por operação computada atinge um mínimo, ou a eficiência energética atinge um pico. Desta forma, apesar da tensão de dimensionamento da biblioteca de Stangherlin (2013) estar próxima da tensão de limiar dos transistores de 65 nm utilizados, operar com $V_{DD}$ de 450 mV não resulta no ponto de maior eficiência energética, como foi demonstrado em (STANGHERLIN, 2013). Cabe salientar que os transistores utilizados no referido trabalho possuem tensão de limiar padrão (RVT).

Neste capítulo, serão introduzidos três aperfeiçoamentos no trabalho de Stangherlin (2013) para o desenvolvimento da biblioteca de células CMOS para operação a baixo $V_{DD,}$ a saber:

a) Redimensionamento dos transistores da biblioteca de células apresentada em (STANGHERLIN, 2013), tendo como alvo inicial uma outra tensão de operação: 300 mV, com o intuito de otimizar mais a operação do circuito próximo ao MEP;
b) Introdução de uma diversidade maior de células combinacionais, com a inclusão de células OAI21 e AOI22;

c) Projeto de duas bibliotecas de células combinacionais adicionais, com transistores HVT e LVT, utilizando a mesma metodologia de projeto e tensão de operação da biblioteca RVT.

Anteriormente às questões de dimensionamento para as três bibliotecas, será apresentada a metodologia de simulação utilizada e, após, informações sobre a caracterização da biblioteca de células com transistores RVT serão discutidas.

### 3.2 Metodologia de simulação das células

As simulações das células desenvolvidas neste trabalho, cujos resultados serão apresentados na próxima seção, foram realizadas através de uma série de *scripts* desenvolvidos na linguagem Python combinados ao simulador HSPICE©. Basicamente, este conjunto de *scripts* recebe como entrada:

• Um *netlist* SPICE contendo variáveis a serem substituídas;

• Definições de parâmetros como, por exemplo, faixas de temperatura, tensão de operação e largura de transistores;

• Definição da estimativa de capacitâncias parasitas de junções associadas a cada transistor;

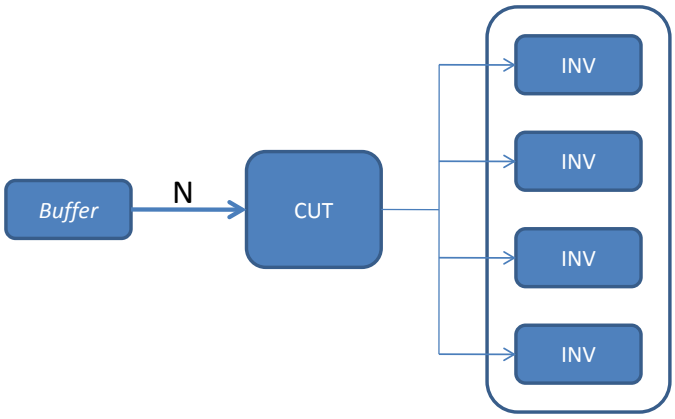
• Definições de medidas, como, tempos de subida e descida, etc.

Após o recebimento da entrada, as definições de parâmetros e de medidas são inseridas em todas as combinações possíveis no *netlist* SPICE, substituindo suas variáveis correspondentes e, gerando, consequentemente, uma série de novos *netlists* que, por sua vez, são simulados de forma paralela no HSPICE. Ao final das simulações, os dados são relacionados e coletados pelos *scripts* em Python que, por sua vez, possibilitam que tais resultados sejam plotados e/ou salvos para pós-processamento por ferramentas de cálculo numérico. Tal *framework* Python, apesar de não ser discutido em seu trabalho, foi desenvolvido por Stangherlin (2013).

A Figura 3.1 ilustra um *netlist* SPICE genérico para simulação de células combinacionais. No bloco central, o circuito em teste (*circuit under test*) é uma variável que é substituída em tempo de execução pelos *scripts* em Python. Esta variável pode, por exemplo, assumir o valor de uma porta NAND de duas entradas com um comprimento de canal mínimo e, em outro momento, tal comprimento pode ser incrementado em 50%, bem como, ter sua tensão de alimentação ou temperatura modificadas. O bloco central recebe seus sinais de entrada por um *buffer* de tamanho fixo. O termo *N* acima da interligação dos blocos representa a quantidade de entradas do circuito em teste. Consequentemente, se o circuito

possuir mais de uma entrada, elas estarão curto-circuitadas. O circuito em teste estará sempre conectado a uma carga de quatro inversores em paralelo (FO4). A mesma metodologia de simulação foi adotada por Stangherlin (2013).

**Figura 3.1 - Metodologia de simulação de células combinacionais**




## 3.3 Dimensionamento de transistores

Nesta seção serão apresentados e discutidos aspectos de dimensionamento e temporizações para três bibliotecas de células, cada uma utilizando transistores com tensões de limiar distintas. Em todos os casos, o dimensionamento será ditado pela mesma tensão de operação, 300 mV, visando atuar em regime de alimentação próximo à tensão de limiar dos transistores, ou *near*-VT, com o intuito de atingir alta eficiência energética. Problemas de variabilidade que podem afetar, por exemplo, o intervalo de variação de tensões (*voltage swings*) não serão exploradas neste trabalho, uma vez que foi demonstrado em (STANGHERLIN, 2013) que para operação na tensão de dimensionamento de células deste trabalho, 300 mV, não é necessário realizar um aumento das redes *pull-up* e *pull-down* (simultaneamente) para garantir um mínimo de variação de tensão entre 10% a 90% de $V_{DD}$ para os sinais lógicos.

No desenvolvimento das bibliotecas deste trabalho e no dimensionamento dos transistores foi utilizado o PDK (*Process Design Kit*) de tecnologia 65 nm CMOS *Bulk* (IBM, 2009), um processo comercial acessível à UFRGS através do serviço da empresa americana MOSIS Inc. Todas as células lógicas foram inicialmente dimensionadas para a tensão $V_{DD}$=300 mV, à temperatura de 25ºC e não considerando os casos limites (*corner cases* de 3-sigma) de variação de processo. O dimensionamento é feito para o caso TT (*Typical* NMOS/ *Typical* PMOS), enquanto a caracterização posterior será feita para cada célula considerando os casos limites. O modelo dos transistores utilizados é o BSIM4 (BSIM4, 2000). A tensão

nominal de operação para esta tecnologia, tal como recomenda a empresa detentora do processo 65 nm, é de 1,2 V.

### 3.3.1 Células com transistores Regular-VT

Nas subseções abaixo serão apresentados dimensionamentos de células combinacionais e sequenciais utilizando transistores com tensão de limiar convencional ou padrão, também conhecidos como *Regular*-VT, *Standard*-VT, ou RVT, para o PDK acima referido. A Tabela 3.2 apresenta os valores das tensões de limiar para transistores NMOS e PMOS de comprimento de canal mínimo.

**Tabela 3.2 - Tensão de limiar para transistores RVT de tamanho mínimo para o PDK de 65 nm CMOS Bulk**

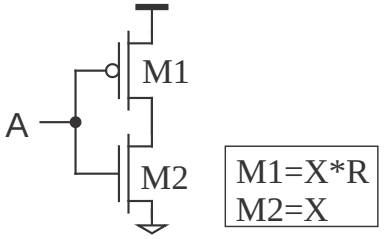
| | **NMOS** | **PMOS** |
|---|---|---|
| **RVT** | 428 mV | -400 mV |

**Fonte: (IBM, 2009)**

*3.3.1.1 INV*

A Figura 3.2 apresenta o esquemático de um inversor CMOS estático e as equações que determinam o dimensionamento de seus transistores para a biblioteca desenvolvida. Tanto o transistor NMOS quanto o PMOS apresentam a constante *X*, que representa o fator multiplicativo da largura efetiva do transistor, em relação à mínima largura (*W*) admitida no processo. Neste texto, referimos a *X* como "*transistor strength*", ou a sua capacidade de drenar corrente, considerando-se que o mesmo $L_{eq}$ seja utilizado como referência. Seu valor mínimo é um. O termo *R* representa a razão ($W_P$ / $W_N$) das larguras dos dois transistores.

**Figura 3.2 - Dimensionamento do inversor**

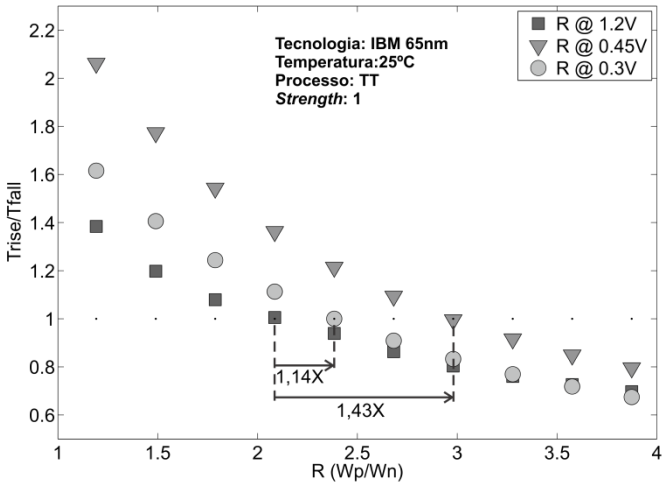



Unidades de tamanho mínimo;
X: *Transistor Strength*;
R: Razão Wp/Wn;

**Fonte: Figura modificada de (STANGHERLIN, 2013)**

A Figura 3.3 ilustra a variação das razões de tempos de subida (*Trise*) / tempos de descida (*Tfall*), como função das razões *R* (*Wp/Wn*) para três tensões $V_{DD}$ distintas. A tensão de 1,2 V refere-se à tensão nominal do processo (PDK) utilizado, a tensão de 450 mV foi a escolhida para o desenvolvimento da biblioteca de células de Stangherlin (2013) e a tensão de 300 mV foi adotada para o desenvolvimento da biblioteca deste trabalho. Como esperado, para equalizar os tempos *Trise*/*Tfall* (linha horizontal pontilhada no gráfico), existem três razões *Wp/Wn*, uma para cada tensão. Nota-se que, há um incremento da razão *R* do inversor de 1,14X da tensão nominal para a tensão de 0.3 V, e um aumento de 1,43X para a tensão de 0.45 V. Desta forma, conclui-se que a equalização dos tempos de subida e descida não é factível para situações onde há grandes variações na tensão de alimentação dos circuitos.

**Figura 3.3 - Tempos de subida/descida x Razão de larguras de um inversor X1 em três tensões distintas**




De acordo com Rabaey et al. (2003), reduzir a tensão de alimentação tem impacto positivo na dissipação de energia, entretanto é absolutamente prejudicial em relação ao desempenho de uma porta lógica CMOS. Tal constatação, um aumento exponencial dos tempos de resposta, pode ser facilmente observada na Figura 3.4. Levando-se em consideração o valor da razão *Wp/Wn* onde o inversor é simétrico para cada tensão, ou seja, onde a razão *Trise*/*Tfall* é igual a 1, o tempo de subida é reduzido em 96,7 % de 300 mV para 450 mV e diminuído em 99,96 % quando a tensão de *near-VT* é comparada ao ponto de inversão forte. Os valores do tempo de subida são 108,5 ns, 3,541 ns e 38,6 ps para 0,3 V, 0,45 V e 1,2 V, respectivamente. Com relação ao atraso de propagação[4] ($t_p$), as reduções são

[4] Representa a média do atraso para uma comutação "baixo para alto" ($t_{pLH}$) e "alto para baixo" ($t_{pHL}$) na saída de uma porta lógica

próximas às encontradas na avaliação do *Trise*, e são 47,51 ns para 300 mV, 1,738 ns para 450 mV e 20,9 ps para 1,2 V. A Tabela 3.3 resume o dimensionamento e os atrasos de propagação para o inversor INVX1 para a tensão de 0,3 V à 25ºC. Cabe salientar que para o processo utilizado, a largura mínima dos transistores é de 120 nm e seu comprimento mínimo é de 60 nm.

**Figura 3.4 - Tempos de subida e atrasos de propagação do INVX1 para três tensões distintas**

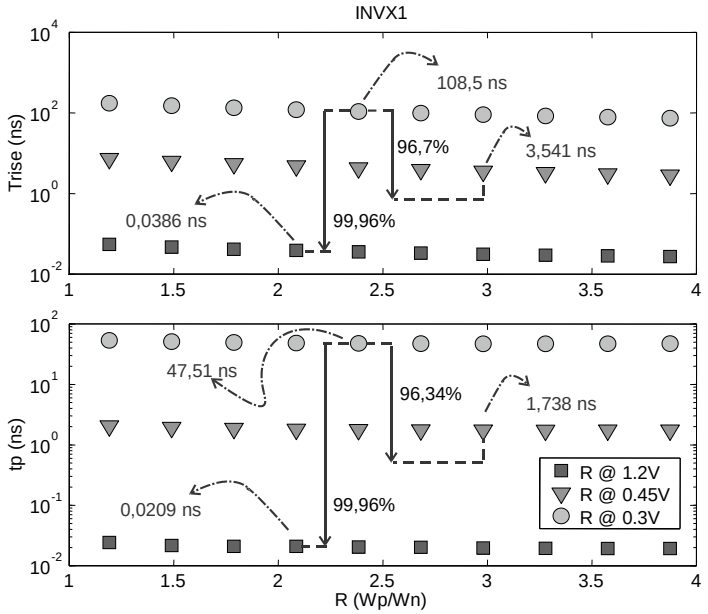
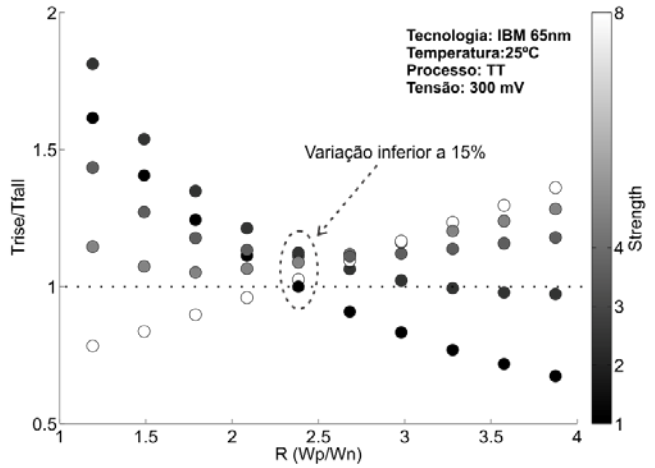



**Tabela 3.3 - Dimensionamentos e temporizações para INVX1 em função da razão Wp/Wn adotada @ 0.3 V**

| Wp / Wn | Wp (nm) | Wn (nm) | Lp=Ln (nm) | $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|---|---|---|---|
| 2,38 | 286 | 120 | 60 | 108,5 | 47,5 | 48,1 | 47 | 25 |

A Tabela 3.3 apresentou as dimensões e as características temporais para o inversor de *strength* 1 (INVX1). Entretanto, da mesma maneira que Stangherlin (2013), foram desenvolvidos mais quatro inversores, INVX2, INVX3, INVX4 e INVX8, aplicando-lhes as mesmas regras de dimensionamento apresentadas na Figura 3.2. Na Figura 3.5, é possível observar que os inversores de *strength* superior a 1 apresentam um Trise/Tfall também superior a 1. Entretanto, esta variação não ultrapassa 15%.

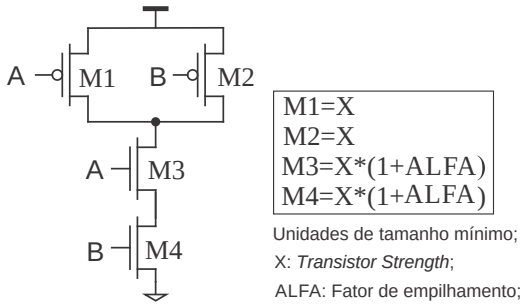
**Figura 3.5 - Trise / Tfall x Wp/Wn x *Strengths1-8* para inversor @ 300mV / 25ºC**




*3.3.1.2 NAND*

A Figura 3.6 apresenta o esquemático de uma porta NAND de duas entradas e o dimensionamento dos seus transistores. Os transistores PMOS, M1 e M2, da rede *pull-up*[5] (PUN), somente serão maiores que o tamanho mínimo quando a constante *X* for maior do que um. Por outro lado, os transistores NMOS em série, M3 e M4, da rede *pull-down*[6] (PDN), apresentam além do termo *X*, a variável *ALFA*. Ela representa o "fator de empilhamento", que nada mais é do que uma possibilidade de aumentar a largura dos transistores em série de modo a evitar a redução de desempenho inerente a este tipo de conexão (RABAEY; CHANDRAKASAN; NIKOLIC, 2003).

**Figura 3.6 - Dimensionamento da porta NAND**




**Fonte: Figura modificada de (STANGHERLIN, 2013)**

[5] A finalidade da rede PUN é promover um caminho de baixa resistência entre a saída de uma porta lógica e sua linha de $V_{DD}$ (tensão de alimentação). Ela irá prover este caminho em função dos níveis lógicos presentes em suas entradas.

[6] A rede PDN tem um propósito exatamente oposto. Ou seja, tem por função proporcionar um caminho de baixa resistência entre a saída de uma porta lógica e o ponto de referência (potencial nulo), de acordo com a combinação dos níveis lógicos de entrada.

Neste projeto, foram analisadas portas NAND de até quatro entradas e, tal como ocorreu no dimensionamento à 450 mV (STANGHERLIN, 2013), as três portas apresentaram um fator *ALFA* inferior a 10. Precisamente, 1,28 para a NAND2X1, 3,84 para a NAND3X1 e 6,4 para a NAND4X1. Neste caso, os transistores da rede PDN da NAND4X1 experimentariam um aumento de 7,4X em relação à largura mínima desta tecnologia (120 nm). Na lógica CMOS estática complementar, o número de transistores para implementar uma porta de N entradas é 2N (RABAEY; CHANDRAKASAN; NIKOLIC, 2003). Desta forma, a área ocupada por uma porta de quatro entradas, seria consideravelmente grande.

Outro fator problemático é que o atraso de propagação deste tipo de lógica aumenta rapidamente em função do número de entradas (RABAEY; CHANDRAKASAN; NIKOLIC, 2003). O atraso para uma comutação na saída de "alto para baixo" ($t_{pHL}$) é de 71,13 ns para a porta de duas entradas, 114,2 ns para a NAND3X1 e 183,5 ns para a NAND4X1. Em outras palavras, isto representa um aumento de 60,55% no $t_{pHL}$ da NAND2X1 para a NAND3X1 e de 157,98% da porta de duas entradas para a porta de quatro entradas. Adicionalmente, quanto maior for o número de transistores empilhados, maior será sua vulnerabilidade à variações de processo, tensão e temperatura, inerentes à operação próximo a tensão de limiar (ZHAO et al., 2013). Desta forma, assim como em Stangherlin (2013), apenas a porta de duas entradas em três *strengths* distintos (X1, X2 e X4), foi implementada para a operação em VFS muito amplo. O comportamento *Trise/Tfall* e $t_{pHL}$ versus *ALFA* são ilustrados na Figura 3.7 e informações de dimensionamento e temporizações podem ser encontrados na Tabela 3.4.

**Figura 3.7 - *Trise/Tfall* e $t_{pHL}$ versus *ALFA* para a NAND2X1 em 300mV/25ºC**

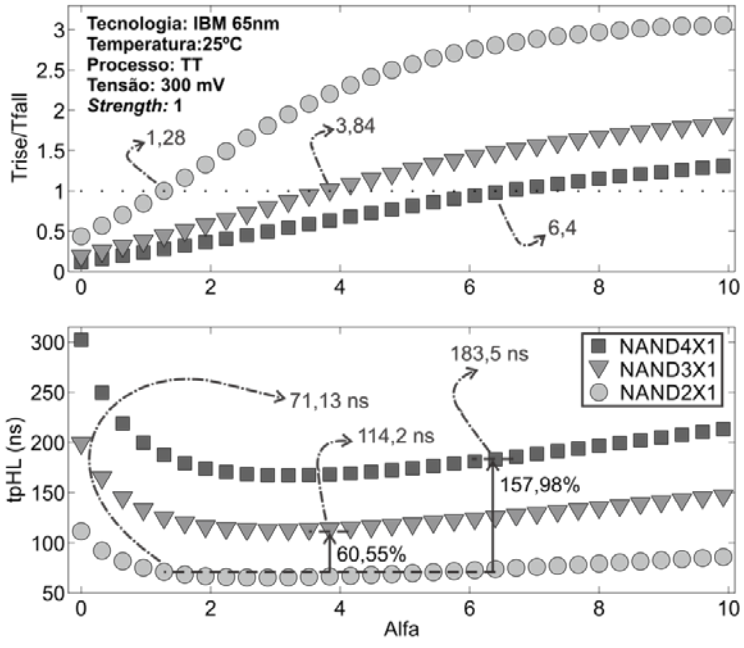


**Tabela 3.4 - Dimensionamentos e temporizações para a NAND2X1 em função do fator *ALFA* adotado**

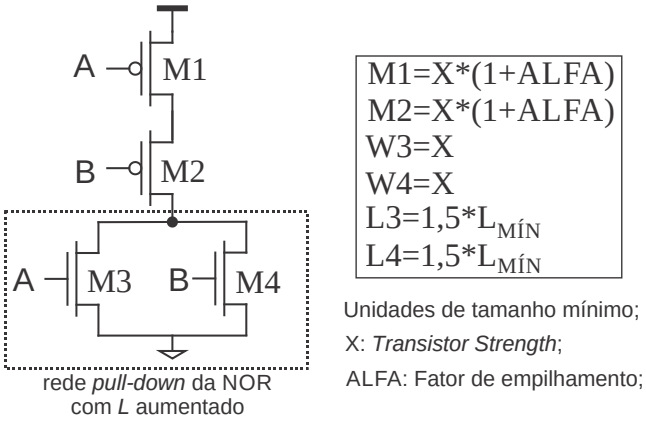
| ALFA | Wn3,4 (nm) | Wn3,4 / Wmin | Wp1,2 (nm) | Wp1,2 / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|
| 1,28 | 274 | 2,28 | 120 | 1 | 60 |

| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 112,5 | 65,6 | 60,1 | 71,13 | 25 |

*3.3.1.3 NOR*

A Figura 3.8 ilustra o esquemático de uma porta NOR de duas entradas, bem como o seu dimensionamento. Esta porta, semelhantemente à porta NAND, possui a variável *ALFA* que refere-se ao fator de empilhamento. Entretanto, obviamente, a referida variável está presente de modo oposto à porta NAND, uma vez que na NOR, os transistores empilhados localizam-se na rede *pull-up*. O termo *ALFA*, diferentemente da constante *X*, é definido individualmente para cada célula. O comprimento do canal dos transistores NMOS foi aumentado em 50% em relação ao comprimento mínimo com o intuito de reduzir a largura dos transistores PMOS. Esta abordagem será explicada e justificada posteriormente nesta seção. A largura dos transistores da rede *pull-down* continua sendo um múltiplo do tamanho mínimo pela constante *X*.

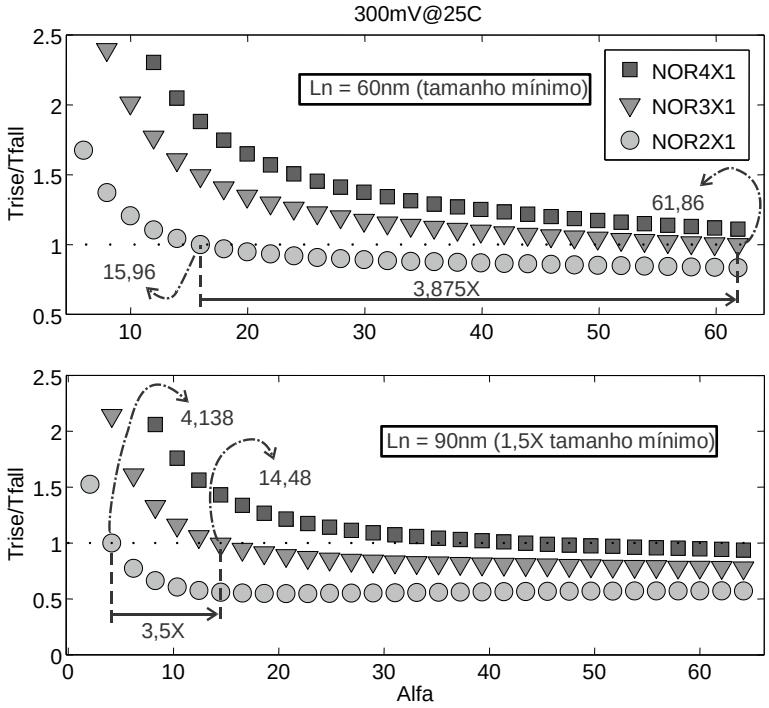
**Figura 3.8 - Dimensionamento da porta NOR**




Como pode-se observar na parte superior da Figura 3.9, referente a uma rede PDN com *L* mínimo, há um aumento de 3,875X da NOR de duas entradas para a NOR de três entradas. Além disto, o *ALFA* referente à NOR3X1 é de aproximadamente 61,86, o que resultaria em transistores 63X superiores ao tamanho mínimo. Nesta figura, também é possível observar que o *ALFA* da NOR4X1 é, obviamente, superior ao da NOR3X1 e que não chega a cruzar a linha de simetria de Trise/Tfall no último ponto considerado (61,86). A parte inferior da Figura 3.9 apresenta, também, o comportamento de três portas NOR com número

de entradas distintas, mediante a variação de *ALFA* por Trise/Tfall. Entretanto, nota-se que o ponto de cruzamento de *ALFA* com o eixo de simetria horizontal foi antecipado. Aproximadamente, de 15,96 para 4,138 no caso da NOR2X1, de 61,86 para 14,48 referente à NOR3X1 e que o valor de *ALFA* para a NOR4X1 tornou-se aparente (por volta de 43). Em outras palavras, isto representa uma redução de 74,1% para a NOR de duas entradas e 76,6% relativo a NOR de três entradas. Houve também, uma redução da diferença entre o *ALFA* da NOR2X1 para a NOR3X1. Estas reduções foram obtidas adotando-se o aumento do comprimento do canal dos transistores NMOS para 90 nm (1,5X o tamanho mínimo da tecnologia).

**Figura 3.9 - Trise/Tfall x Alfa para porta NOR com Ln=60 nm e Ln=90 nm @ 300mV/25ºC**




Considerando o *ALFA* (15,96) que equaliza os tempos de subida e de descida para a NOR2X1, o atraso de propagação, referente a uma porta NOR com comprimento mínimo de canal (60 nm) para todos os transistores, para uma comutação na saída do tipo "baixo para alto" ($t_{pLH}$) é de 177,4 ns. Nas mesmas condições (*ALFA*=61,86), este atraso de propagação para a porta de três entradas é de 622,3 ns. Isto representa um acréscimo de 250,79% quando comparado ao atraso da porta de duas entradas. No caso da NOR4X1, percebe-se claramente que este atraso é superior a 900 ns. Estas informações podem ser encontradas na Figura 3.10 (a). Repetindo-se esta análise para a situação onde o comprimento do canal dos transistores da

rede *pull-down* é aumentado para 90 nm, o $t_{pLH}$ para a NOR2X1 é de 126,5 ns (redução de 28,69% em relação a NOR2X1 com *L* mínimo). Na NOR3X1 este valor é de 296,5 ns (52,35% de redução) e, para NOR4X1, o atraso é de 762,8 ns. Apesar das reduções nos atrasos em função do aumento do *L* mínimo, estes valores representam um aumento de 134,4% no $t_{pLH}$ da NOR2X1 para a NOR3X1 e de 503% da porta de duas entradas para a porta de quatro entradas. Porcentagens superiores às encontradas na situação de pior atraso de *swing* ($t_{pHL}$) para a porta NAND. Obviamente, em função da diferença de mobilidade entre os transistores PMOS e NMOS. A análise anterior está associada a Figura 3.10 (b). Em virtude dos fatos mencionados anteriormente, e, também pela menor vulnerabilidade à variações de PVT, apenas a porta NOR de duas entradas foi implementada em três *strengths*: X1, X2 e X4. A redução da capacidade de corrente dos transistores NMOS em função do aumento de seu comprimento não impactou substancialmente o desempenho da rede PDN. A Figura 3.10 (c) ilustra o comportamento do $t_{pHL}$ para a NOR2X1 nos dois comprimentos analisados em função da variação de *ALFA*. Dois pontos estão evidenciados: 92,35 ns e 109 ns. O primeiro indica o atraso de "alto para baixo" em relação ao *ALFA* de 4,138 da Figura 3.10 (b) e o segundo indica o $t_{pHL}$ referente ao *ALFA* de 15,96 da Figura 3.10 (a). Portanto, a redução de *ALFA* em "detrimento" do aumento do *L* do NMOS, resultou, na realidade, em uma redução no tempo de propagação da porta. Em termos percentuais, houve uma redução de 15,28% no $t_{pHL}$. Tal resultado demonstrou-se em desacordo com Stangherlin (2013), que havia concluído que o referido aumento impactaria no desempenho da célula. O aumento do comprimento de canal dos transistores M3 e M4 (Figura 3.8) mostrou-se adequado para esta diminuição de atraso (Figura 3.10 (c)) devido ao efeito de *Reverse Short-Channel* (KIM et al., 2007), pelo qual há uma redução na tensão de limiar dos transistores com o comprimento de canal maior que o mínimo (*Lmin*).

Informações de dimensionamentos e temporizações para a NOR2X1 podem ser encontradas na Tabela 3.5.

**Figura 3.10 - $t_{pLH}$ versus *ALFA* para NOR2X1, 3X1 e 4X1: (a) Ln=60 nm; (b)Ln=90 nm; (c)$t_{pHL}$ versus *ALFA* para NOR2X1 com Ln=60 nm e Ln=90 nm @ 300mV/25°C**

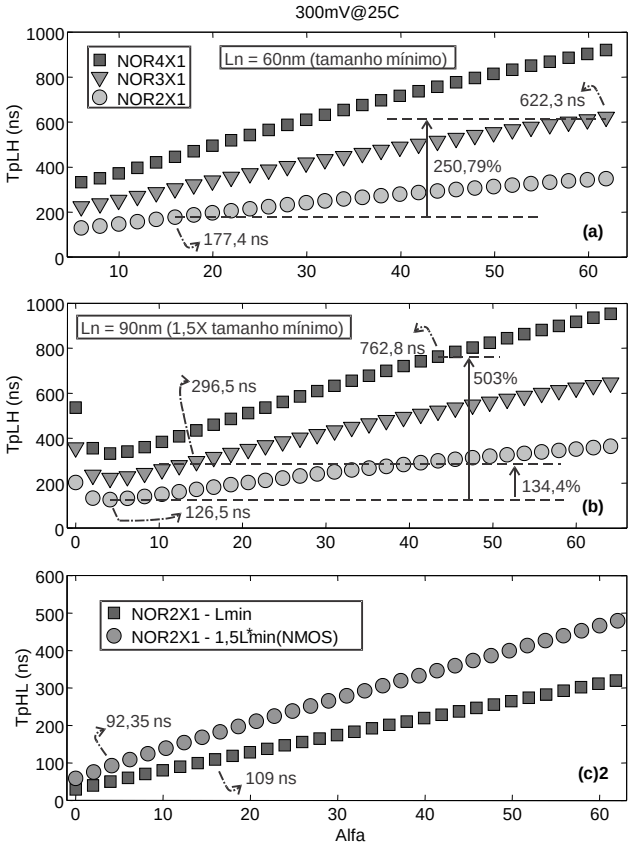



**Tabela 3.5 - Dimensionamentos e temporizações para a NOR2X1 referentes ao ALFA e L do NMOS adotados**

| ALFA | Wp1,2 (nm) | Wp1,2 / Wmin | Wn3,4 (nm) | Wn3,4 / Wmin | Lp (nm) | Ln (nm) |
|---|---|---|---|---|---|---|
| 4,138 | 617 | 5,138 | 120 | 1 | 60 | 90 |

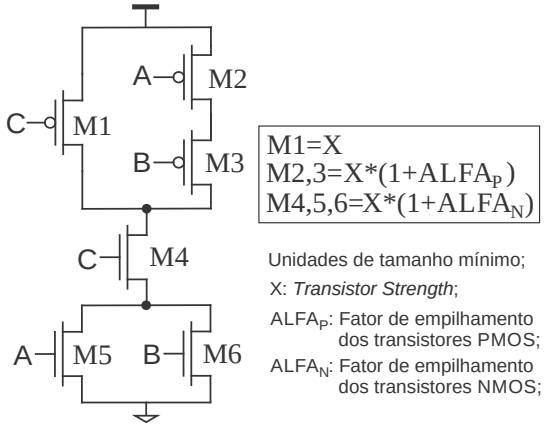
| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 142,8 | 109,4 | 126,5 | 92,35 | 25 |

*3.3.1.4 OAI21*

Com o intuito de prover maior flexibilidade para a ferramenta de síntese lógica, duas portas adicionais foram incluídas na biblioteca para operação em NTV: OAI21 e AOI22. Tais portas são partes integrantes, por exemplo, de diferentes topologias de somadores projetados em uma abordagem CMOS invertida (Harris; Sutherland, 2003). A Figura 3.11 apresenta o esquemático da porta OAI21 e as equações que ditam seu dimensionamento. No caso da rede PUN, o transistor M1 está em paralelo com os transistores M2 e M3, que por sua vez, estão empilhados. Desta forma, o transistor M1 possui seu tamanho ditado apenas pela constante *X*, enquanto que os transistores M2 e M3 dependem também do termo $ALFA_P$ (fator de empilhamento dos transistores PMOS), por estarem em série. No caso da rede PDN,

independentemente do padrão de entrada, haverá sempre dois transistores empilhados. Desta forma, o dimensionamento dos transistores M4 a M6 depende de *X* e de $ALFA_N$ (fator de empilhamento dos transistores NMOS). As variáveis representantes do "fator de empilhamento", $ALFA_P$ e $ALFA_N$, são dimensionadas separadamente.

**Figura 3.11 - Dimensionamento da porta OAI21**




A Figura 3.12 ilustra o comportamento de Trise/Tfall em função da variação de $ALFA_P$ para três valores distintos de $ALFA_N$. É possível observar que se os transistores da rede PDN fossem dimensionados com tamanho mínimo, isto é, $ALFA_N = 0$, não haveria uma simetria entre os tempos de subida e descida. Adicionalmente, se o valor de $ALFA_N = 1{,}6$, o valor de $ALFA_P$ seria muito superior ao valor adotado. As informações de dimensionamento e temporizações para X=1 podem ser encontradas na Tabela 3.6. Esta tabela refere-se à OAI21 otimizada, com $ALFA_P = 2{,}46$ e $ALFA_N = 0{,}8$. Além do *strength*=1, foram projetadas OAI21 com X=2 e X=4.

**Figura 3.12 - Trise/Tfall x ALFAp x ALFAn para porta OAI21X1**

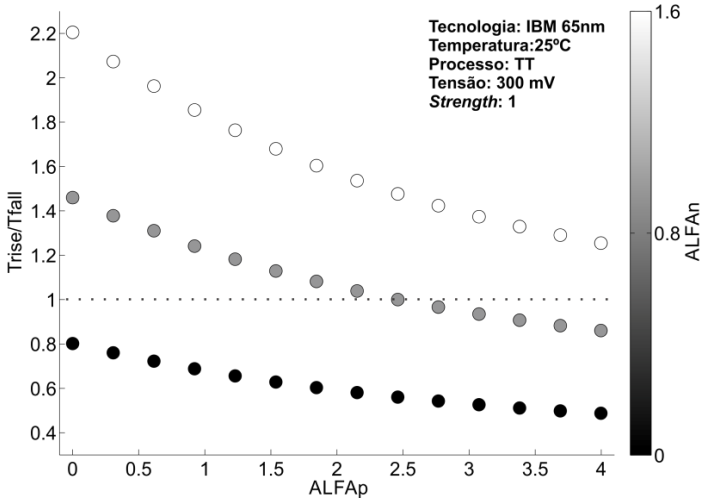


**Tabela 3.6 - Dimensionamentos e temporizações para a OAI21X1**

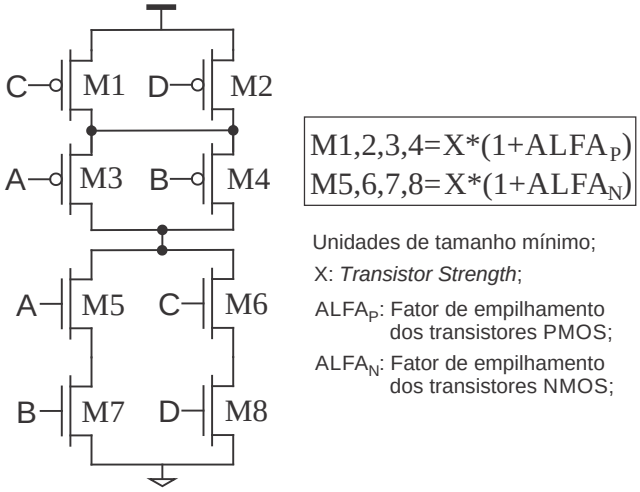
| ALFAp | Wp1 (nm) | Wp2,3 (nm) | Wp2,3 / Wmin | ALFAn | Wn4-6 (nm) | Wn4-6 / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|---|
| 2,46 | 120 | 415 | 3,46 | 0,8 | 216 | 1,8 | 60 |

| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 113,9 | 91,3 | 97 | 85,7 | 25 |

*3.3.1.5 AOI22*

Outro exemplo de aplicação para uma porta AOI é apresentado no trabalho de Hsu et al. (2012). Os autores desenvolveram uma máquina de permutação vetorial SIMD reconfigurável de 4 até 32 fluxos para cargas de trabalho comuns em processamento de dados, multimídia e aplicações gráficas. Em sua implementação, foi utilizado um multiplexador AOI estático para integrar duas células de memória. A Figura 3.13 apresenta o esquemático e as equações de dimensionamento de uma porta AOI22. Tanto na rede PUN quanto na rede PDN, independentemente do padrão de entrada, haverá sempre dois transistores empilhados. Desta forma, o dimensionamento de cada transistor dependerá de um dos dois fatores de empilhamento *ALFA*.

**Figura 3.13 - Dimensionamento da porta AOI22**




Diferentemente da porta OAI21, o tamanho mínimo pode ser utilizado nos transistores da rede PDN de forma que a simetria de temporizações seja mantida. Na Figura 3.14 é apresentado o comportamento da razão dos tempos de subida pelos tempos de descida em função da variação de $ALFA_P$ para três valores de $ALFA_N$. Se o fator de empilhamento adotado fosse o maior dos três casos, haveria um aumento de 83% nos transistores PMOS e de 50% nos transistores NMOS. Desta forma, os menores $ALFA_P$ e $ALFA_N$ que fazem *Trise/Tfall*=1 foram adotados. A exploração de dimensionamento ilustrada na Figura 3.14

conduziu às escolhas de $ALFA_P$ = 1,35 e $ALFA_N$ = 0. Da mesma forma que nas portas combinacionais anteriores, foram projetadas portas AOI22 com três *strengths* distintos: X1, X2 e X4. As informações de dimensionamento e temporizações para a AOI22X1 podem ser encontradas na Tabela 3.7.

**Figura 3.14 - Trise/Tfall x ALFAp x ALFAn para porta AOI22X1**

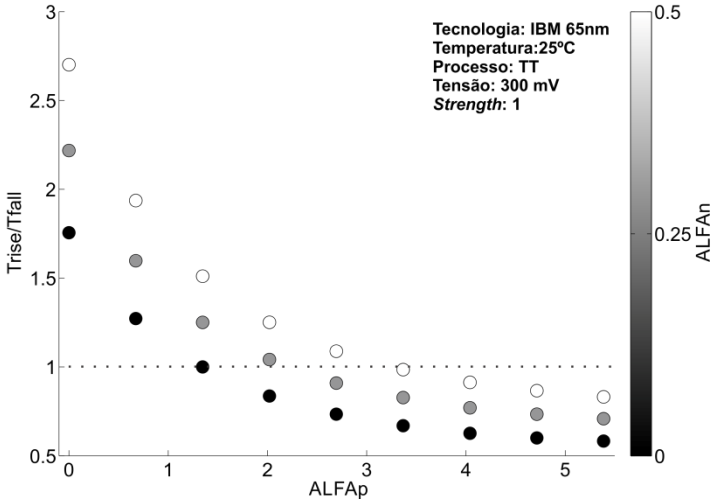



**Tabela 3.7 - Dimensionamentos e temporizações para a AOI22X1**

| ALFAp | Wp1-4 (nm) | Wp / Wmin | ALFAn | Wn5-8 (nm) | Wn / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|
| 1,35 | 282 | 2,35 | 0 | 120 | 1 | 60 |

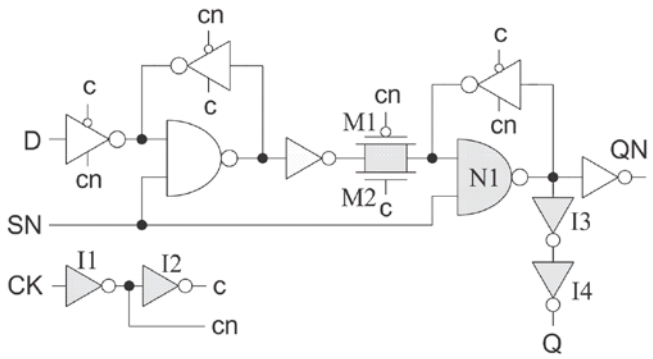
| $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 143,3 | 131,9 | 148,2 | 115,5 | 25 |

*3.3.1.6 FFs*

Foram projetados dois registradores mestre-escravo baseados em *transmission-gates*: o primeiro possui um sinal de SET ativo em nível baixo (DFFS) e o segundo possui um sinal de RESET, do mesmo modo, ativo em nível baixo (DFFR). A metodologia de desenvolvimento, simulações e análise de resultados foi a mesma apresentada em (STANGERLHIN, 2013), entretanto, alterando o dimensionamento para uma tensão de alimentação de 300 mV. A Figura 3.15 apresenta a arquitetura do registrador com sinal de SET ativo em baixo, bem como as equações que ditam seu dimensionamento. O projeto baseia-se em variar o dimensionamento dos transistores e portas evidenciados em cinza, enquanto as portas em branco permanecem em tamanho mínimo (STANGERLHIN, 2013). Os inversores I3 e I4, bem como a porta NAND N1 do bloco escravo foram dimensionadas previamente neste capítulo e só variam de tamanho de acordo com o *strength* do registrador em simulação (X1, X2 ou X4). Os *drivers* de *clock*, I1 e I2, foram projetados para manter um atraso no sinal de *clock* para uma condição de FO2 (STANGERLHIN, 2013). Os *transmission-gates* M1 e M2

foram dimensionados, também, de acordo com a variação de *strength* do registrador. Entretanto, cabe salientar, que o transistor PMOS foi dimensionado com o dobro do NMOS nas três configurações de tamanho projetadas.

**Figura 3.15 - Arquitetura e dimensionamento para o registrador mestre-escravo com SET ativo em nível baixo. Portas em cinza foram otimizadas via simulação através das faixas de variação**




**Faixas de variação de dimensionamento dos transistores/portas (portas em branco estão em tamanho mínimo):**

| | | | |
|---|---|---|---|
| M1=[1; I4*3] | I1=FO2 | I3=[1;I4] | N1=[1;I4] |
| M2=[1; I4*1.5] | I2=FO2 | I4={1,2,4} | |

Unidades de tamanho mínimo; I4 é ajustado de acordo com o *strength* do registrador;

**Fonte: Figura modificada de (STANGHERLIN, 2013)**

A Figura 3.16 apresenta os resultados de consumo de energia em função do atraso de propagação ($t_{c\text{-}q}$) do registrador para os três *strengths* projetados, considerando o melhor tempo de setup ($t_{su}$), isto é, o dado está na condição inicial (STANGERLHIN, 2013). Tais resultados são, na realidade, médias em função da energia consumida pelas transições de subida e de descida na saída do registrador. É possível observar que o espaço de projeto é vasto. Entretanto, apenas alguns registradores são eficientes em termos de consumo (STANGERLHIN, 2013), os quais foram evidenciados na referida figura. No caso do DFFSX1, todos os 15 pontos foram considerados. O ponto de menor energia (1,159 fJ), detém o maior atraso, 499,75 ns, enquanto que o ponto de maior consumo energético (1,197 fJ) possui o menor atraso (477,6 ns). Tal comportamento é facilmente percebido na figura, entretanto, o mesmo padrão não ocorre em relação aos outros dois registradores. Por exemplo, no caso do DFFSX2, apenas 9 pontos dos 216 são energeticamente eficientes, dos quais o menor valor é da ordem de 1,218 fJ para 510,15 ns de atraso, enquanto que o ponto de maior consumo energético (1,664 fJ) está para 548 ns de atraso. Entretanto, o menor atraso é da ordem de 478,45 ns para um consumo de 1,319 fJ e o maior atraso é de 618,8 ns para um consumo de 1,465 ns. Com relação ao registrador de *strength* 4, somente 7 dos 216 pontos são

energeticamente eficientes. O ponto de menor consumo é de 1,345 fJ para um atraso de 538,7 ns e o ponto de menor atraso é de 492,05 ns para um consumo de 1,508 fJ. A Tabela 3.8 resume as relações entre consumo e atraso para os seis registradores projetados, em três situações distintas: maior eficiência energética, menor atraso e registrador escolhido. Os registradores escolhidos localizam-se numa região intermediária entre consumo e atraso e, obviamente, estão entre os valores energeticamente eficientes. A partir da referida tabela, é possível perceber, que, por exemplo, no caso de maior eficiência energética, o atraso de propagação dos registradores com sinal de RESET (DFFR) é no mínimo o dobro dos registradores com sinal de SET e consomem, no mínimo, 59,7% a mais. Basicamente, isto ocorre porque os registradores com sinal de RESET são implementados com portas NOR em substituição às portas NAND dos DFFS. Uma implementação com NANDs é claramente preferível a uma implementação com portas NOR em função da diferença de mobilidade relativa entre os transistores PMOS e NMOS (RABAEY; CHANDRAKASAN; NIKOLIC, 2003). Entretanto, com o intuito de prover maior flexibilidade para a ferramenta de síntese, os registradores com sinal de RESET foram mantidos.

Por fim, cabe salientar que, da mesma forma que os circuitos combinacionais dimensionados anteriormente neste capítulo, quatro cargas idênticas foram conectadas em paralelo na saída de cada registrador (FO4).

**Figura 3.16 - Espaço de projeto para o DFFS em seus três strengths projetados. Os registradores energeticamente eficientes estão evidenciados**

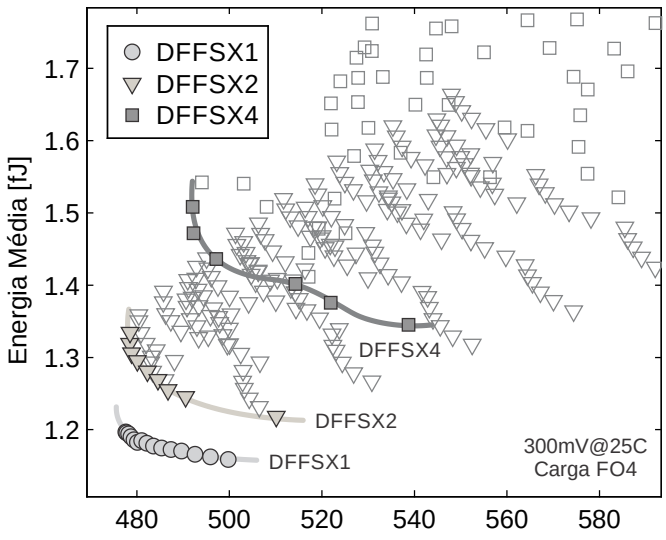


**Tabela 3.8 - Relações entre energia média e atraso de propagação para os seis registradores projetados**

| DFF | Maior eficiência energética | | Incluído na biblioteca | | Menor atraso | |
|---|---|---|---|---|---|---|
| | Energia [fJ] | $t_{c-q}$ [ns] | Energia [fJ] | $t_{c-q}$ [ns] | Energia [fJ] | $t_{c-q}$ [ns] |
| DFFSX1 | 1,159 | 499,75 | 1,175 | 485,35 | 1,197 | 477,6 |
| DFFSX2 | 1,218 | 510,15 | 1,255 | 486,75 | 1,319 | 478,45 |
| DFFSX4 | 1,345 | 538,7 | 1,436 | 497,15 | 1,508 | 492,05 |
| DFFRX1 | 1,942 | 1009,05 | 1,949 | 979,3 | 1,959 | 898,65 |
| DFFRX2 | 1,999 | 1091,65 | 2,035 | 898,2 | 2,157 | 825,6 |
| DFFRX4 | 2,148 | 1131,65 | 2,286 | 845,65 | 2,44 | 825,4 |

### 3.3.2 Células com transistores *High-VT*

Nas subseções a seguir serão apresentados dimensionamentos e temporizações para as mesmas células combinacionais da seção anterior, utilizando transistores com a tensão de limiar acima da convencional, também conhecidos como *High-VT*, ou HVT, para o mesmo processo de fabricação. A Tabela 3.9 informa os valores das tensões de limiar para os transistores de dimensões mínimas do processo de fabricação comercial CMOS 65 nm.

**Tabela 3.9 - Tensão de limiar para transistores HVT de tamanho mínimo para o PDK de 65 nm CMOS Bulk**

| | **NMOS** | **PMOS** |
|---|---|---|
| **HVT** | 585 mV | -587 mV |

**Fonte: (IBM, 2009)**

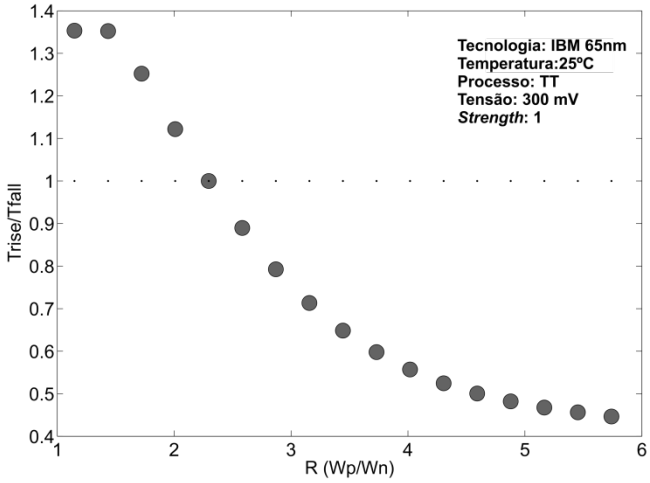
De modo a fazer uma comparação justa, as células foram dimensionadas para a tensão de 300 mV, temperatura de 25°C e variação de processo 3-sigma TT (*Typical* NMOS/ *Typical* PMOS), do mesmo modo que as células dimensionadas para o limiar padrão (RVT), na subseção anterior. A operação destas células em 300 mV, em condição de sub-limiar e inversão fraca, conduz a baixíssimo desempenho em transientes nestas células com transistores HVT. Como se verifica nos resultados a seguir, as células lógicas, mesmo otimizadas, apresentam atrasos lógicos muito altos, superiores a 1 µs. A utilização de células lógicas HVT é extremamente prejudicial ao desempenho dos circuitos a baixas tensões, já que estas células operam em inversão muito fraca.

#### *3.3.2.1 INV*

As equações que ditam o dimensionamento deste inversor são as mesmas utilizadas para os inversores da biblioteca de transistores RVT, apresentados na subseção anterior. Tais equações, se necessárias, podem ser revistas na Figura 3.2. A Figura 3.17 ilustra a variação de

Wp/Wn em relação à razão entre os tempos de subida e descida para o INVX1 HVT. É possível observar que o *R* (Wp/Wn) para equalização dos tempos de subida e descida não variou significativamente quando comparado ao *R* do INVX1 RVT, resultando, inclusive, numa redução de aproximadamente 3,5% na largura do transistor PMOS em relação ao referido inversor.

**Figura 3.17 - Tempos de subida/descida x Razão de larguras de um inversor X1 com transistores HVT**




A Tabela 3.10 resume os resultados de dimensionamento e temporizações para o inversor X1. Como esperado, os atrasos aumentaram significativamente em relação às temporizações do inversor RVT. Os tempos de subida e descida aumentaram 1630,88%, enquanto que no $t_{pHL,}$ $t_{pLH}$ e $t_p$ houve um incremento de 1667,02%, 1596,67% e 1633,3%, respectivamente. Por fim, analogamente ao inversor RVT, foram desenvolvidos quatro inversores adicionais: X2, X3, X4 e X8.

**Tabela 3.10 - Dimensionamentos e temporizações para o INVX1 HVT em função da razão Wp/Wn adotada @ 0.3 V**

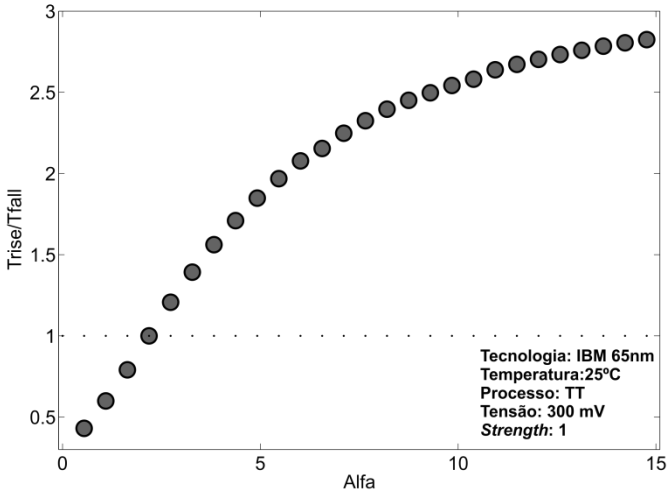
| Wp / Wn | Wp (nm) | Wn (nm) | Lp=Ln (nm) | $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|---|---|---|---|
| 2,3 | 276 | 120 | 60 | 1878 | 823,3 | 816,1 | 830,5 | 25 |

*3.3.2.2 NAND*

Da mesma forma que os inversores HVT seguiram as regras de dimensionamento dos inversores com transistores RVT, as portas NAND com transistores HVT seguem o dimensionamento das NANDs RVT. As equações de dimensionamento podem ser revistas na Figura 3.6. Cabe salientar que, também, foram mantidos apenas dois transistores empilhados na rede PDN. Portanto, somente portas de duas entradas em três *strengths*, X1, X2 e X4,

foram incluídas na biblioteca de células HVT. Os transistores PMOS foram mantidos em tamanho mínimo, igualmente aos transistores da NAND RVT, enquanto que os transistores NMOS experimentaram um aumento de aproximadamente 3,18 vezes em relação ao mínimo. Este acréscimo de largura nos transistores HVT foi superior em aproximadamente 39,8% quando comparado ao *Wn* dos transistores RVT. A Figura 3.18 ilustra o comportamento de trise/tfall em relação à variação de *ALFA.*

**Figura 3.18 - Trise/Tfall x ALFA para a NAND2X1 HVT em 300mV/25ºC**




As informações completas de dimensionamento e temporizações para a NAND de duas entradas, em *strength* 1, com transistores HVT são condensadas na Tabela 3.11. Do mesmo modo que ocorreu na comparação entre as temporizações dos inversores RVT e HVT, houve um aumento na ordem de 1000% nos atrasos de propagação e tempos de subida/descida da NAND2X1 HVT em relação à NAND2X1 RVT. Os atrasos e tempos de transição em condições TT ultrapassam 1 microssegundo nesta NAND2X1 a 300 mV.

**Tabela 3.11 - Dimensionamentos e temporizações para a NAND2X1 HVT em função do fator *ALFA* adotado**

| ALFA | Wn3,4 (nm) | Wn3,4 / Wmin | Wp1,2 (nm) | Wp1,2 / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|
| 2,19 | 383 | 3,19 | 120 | 1 | 60 |

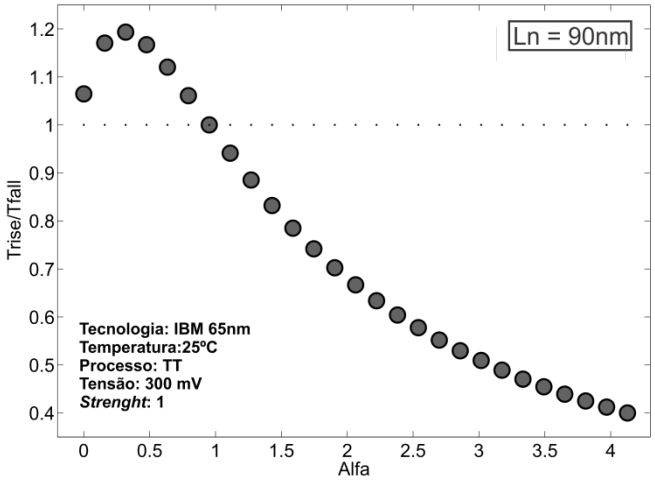
| $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 1325 | 1007,6 | 927,2 | 1088 | 25 |

*3.3.2.3 NOR*

Seguindo o mesmo raciocínio de dimensionamento das portas anteriores, a porta NOR HVT manteve as relações de dimensionamento da sua porta equivalente com transistores

RVT. Foram mantidos somente dois transistores empilhados na rede PUN e transistores de tamanho mínimo na rede PDN. Desta forma, somente portas NOR de duas entradas (X1, X2 e X4), foram incluídas na biblioteca de células HVT. Adicionalmente, de modo a reduzir o tamanho dos transistores PMOS, o comprimento dos transistores NMOS foi aumentado em 50%, como ocorreu na porta NOR com transistores RVT. Tais equações de dimensionamento foram apresentadas na Figura 3.8. É importante mencionar que a largura dos transistores de canal *n* foi mantida em tamanho mínimo enquanto que o *Wp* dos transistores de canal *p* aumentou aproximadamente 95% em relação ao mínimo (120 nm). Quando comparados à largura da porta NOR2X1 RVT, houve uma redução de aproximadamente 62%. A Figura 3.19 apresenta a variação de comportamento da razão dos tempos de subida e descida em relação à variação de *ALFA* para a porta NOR2X1 HVT com o *Ln* = 90 nm .

**Figura 3.19 - Trise/Tfall x Alfa para porta NOR2X1 HVT com Ln=90 nm @ 300mV/25ºC**




As informações sobre dimensionamentos e temporizações para a porta NOR2X1 HVT foram agrupadas na Tabela 3.12. Os atrasos nas temporizações aumentaram consideravelmente para a porta NOR2X1 HVT em relação à sua função lógica equivalente implementada com transistores RVT. Por exemplo, o tempo de subida=descida aumentou aproximadamente 3590%.

**Tabela 3.12 - Dimensionamentos e temporizações para a NOR2X1 HVT referentes ao ALFA e L do NMOS adotados**

| ALFA | Wp1,2 (nm) | Wp1,2 / Wmin | Wn3,4 (nm) | Wn3,4 / Wmin | Lp (nm) | Ln (nm) |
|---|---|---|---|---|---|---|
| 0,95 | 234 | 1,95 | 120 | 1 | 60 | 90 |

| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 5266 | 2613,5 | 2751 | 2476 | 25 |

*3.3.2.4 OAI21*

As equações de dimensionamento desta porta também são ditadas pela sua função lógica equivalente, implementada com transistores RVT. Tais equações foram apresentadas na Figura 3.11. Analogamente à implementação em RVT, a largura do transistor M1 foi mantida em tamanho mínimo, enquanto que os transistores M2 e M3 experimentaram uma redução de 46,02% em relação aos transistores RVT. Na rede *pull-down*, da mesma forma que na OAI2X1 RVT, os transistores não poderiam ser dimensionados em tamanho mínimo por não respeitarem a premissa de equalização dos tempos de subida e descida. Quando comparados à largura dos transistores RVT, não houve uma redução significativa (inferior a 3%). A Figura 3.20 apresenta a razão dos tempos de subida e descida em função da variação de $ALFA_P$ para três valores de $ALFA_N$. Analogamente à implementação com transistores RVT, foram incluídas na biblioteca de células com transistores *High*-VT, três versões de *strength* para a porta OAI21: X1, X2 e X4.

Com relação às questões de temporização, os atrasos desta implementação, comparados à OAI21X1 RVT, aumentaram entre 1231,96% a 1536,52%. Tanto as informações de dimensionamento quanto temporizações para a OAI21X1 podem ser encontradas na Tabela 3.13.

**Figura 3.20 - Trise/Tfall x ALFAp x ALFAn para porta OAI21X1 HVT**

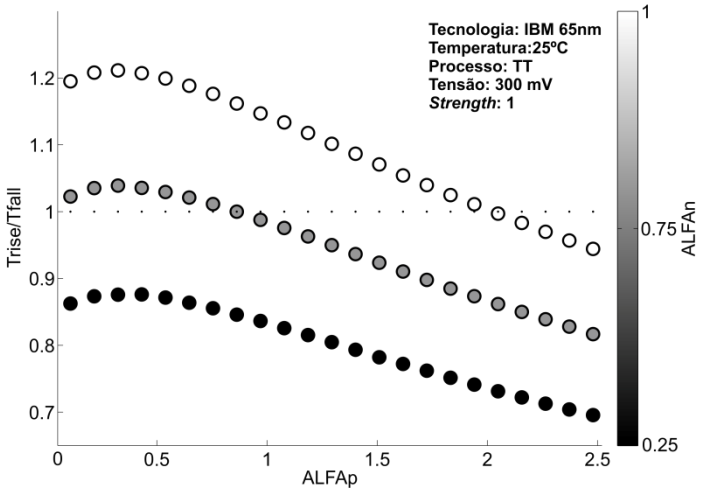



**Tabela 3.13 - Dimensionamentos e temporizações para a OAI21X1 HVT**

| ALFAp | Wp1 (nm) | Wp2,3 (nm) | Wp2,3 / Wmin | ALFAn | Wn4-6 (nm) | Wn4-6 / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|---|
| 0,87 | 120 | 224 | 1,87 | 0,75 | 210 | 1,75 | 60 |

| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 1864 | 1284 | 1292 | 1276 | 25 |

*3.3.2.5 AOI22*

O esquemático do circuito desta porta, bem como as equações de dimensionamento, foram apresentados na Figura 3.13. Tal figura refere-se à implementação da porta AOI22 com transistores RVT, adotada, também, para esta porta com transistores HVT. Os transistores da rede *pull-down* foram mantidos em tamanho mínimo e os da rede *pull-up* aumentaram aproximadamente 2,5 vezes quando comparados ao mínimo. Em relação ao *Wp* da porta com transistores RVT, houve um aumento de 6,38%. A Figura 3.21 apresenta os relacionamentos entre *Trise/Tfall*, $ALFA_P$ e $ALFA_N$ para a tensão de 300 mV à 25ºC. Na Tabela 3.14 as informações de dimensionamentos e temporizações são informadas para a AOI22X1. Os atrasos desta porta quando comparados à implementação com transistores RVT experimentaram um acréscimo entre 1386,5% ($t_{pLH}$) e 1693,44% (*Trise=Tfall*). Adicionalmente, foram incluídas na biblioteca mais duas portas: AOI22X2 e AOI22X4.

**Figura 3.21 - Trise/Tfall x ALFAp x ALFAn para porta AOI22X1 HVT**

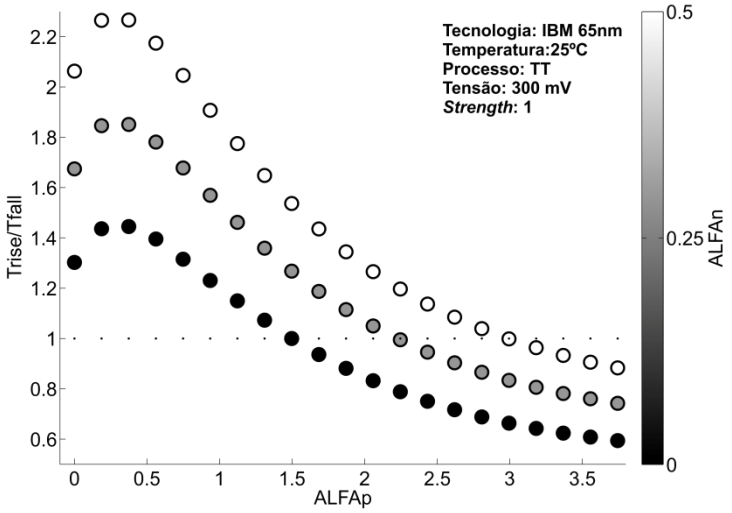



**Tabela 3.14 - Dimensionamentos e temporizações para a AOI22X1 com transistores HVT**

| ALFAp | Wp1-4 (nm) | Wp / Wmin | ALFAn | Wn5-8 (nm) | Wn / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|
| 1,5 | 300 | 2,5 | 0 | 120 | 1 | 60 |
| $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) | | |
| 2570 | 2039 | 2203 | 1875 | 25 | | |

### 3.3.3 Células com transistores *Low-VT*

Nas subseções abaixo serão apresentados dimensionamentos e temporizações para as células INVX1, NAND2X1, NOR2X1, OAI21X1 e a AOI22X1, utilizando transistores com a tensão de limiar abaixo da convencional, também conhecidos como *Low-VT*, ou LVT, para o

mesmo processo de fabricação de 65 nm. A Tabela 3.15 expõe tal tensão para os transistores PMOS e NMOS de tamanho mínimo.

**Tabela 3.15 - Tensão de limiar para transistores LVT de tamanho mínimo para o PDK de 65 nm CMOS Bulk**

| | NMOS | PMOS |
|---|---|---|
| **LVT** | 270 mV | -280 mV |

**Fonte: (IBM, 2009)**

*3.3.3.1 INV*

O inversor LVT utilizou as mesmas regras de dimensionamento do inversor RVT, apresentadas na Figura 3.2. A Figura 3.22 demonstra a variação de *Trise/Tfall* em função da variação de *Wp/Wn* para o INVX1. De modo a equalizar os tempos de subida e descida, o *Wp* teve um aumento de aproximadamente 4,58 vezes em relação ao tamanho mínimo utilizado no transistor NMOS. Este incremento na razão de larguras representa um aumento substancial de aproximadamente 92% quando comparado ao *Wp* do inversor RVT. Entretanto, quando os atrasos de propagação são comparados, a situação é invertida. O inversor com transistores LVT possui atrasos muito inferiores aos inversores RVT. Por exemplo, quando comparados os tempos de subida (na situação de igualdade aos tempos de descida), há uma redução de 92,35% à favor do INVX1 LVT. Os dimensionamentos e temporizações para o referido inversor podem ser encontrados na Tabela 3.16. Da mesma forma que na biblioteca de transistores RVT e HVT, foram incluídos mais quatro inversores na biblioteca de transistores LVT: X2, X3, X4 e X8.

**Figura 3.22 - Tempos de subida/descida x Razão de larguras de um inversor X1 com transistores LVT**

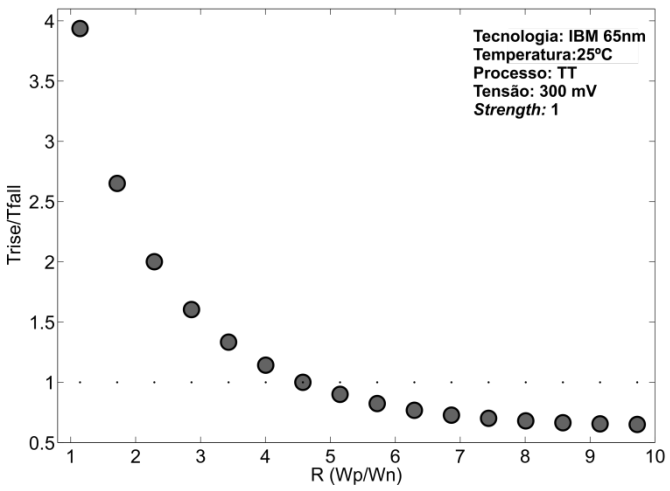


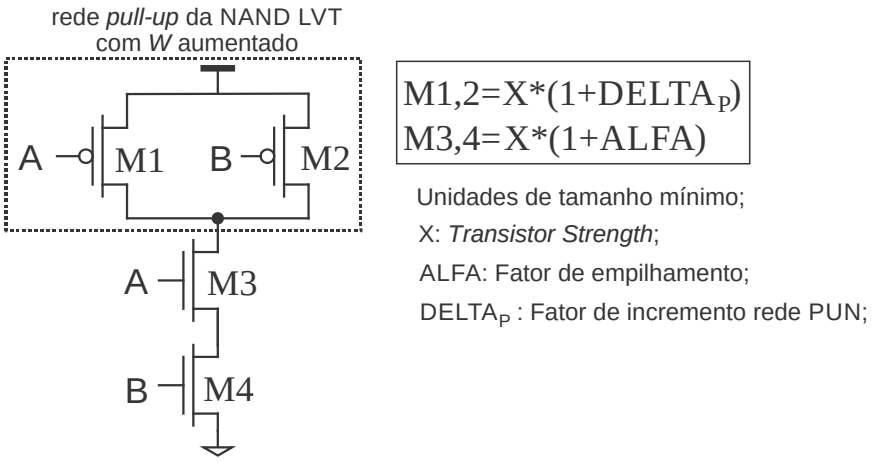
**Tabela 3.16 - Dimensionamentos e temporizações para o INVX1 LVT em função da razão Wp/Wn adotada @ 0.3 V**

| Wp / Wn | Wp (nm) | Wn (nm) | Lp=Ln (nm) | $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|---|---|---|---|
| 4,6 | 549 | 120 | 60 | 8,3 | 4,6 | 4,7 | 4,6 | 25 |

*3.3.3.2 NAND*

Nas equações de dimensionamento das portas NAND2 RVT e NAND2 HVT, a largura dos transistores da rede *pull-up* era determinada somente pela força da porta. No caso de um *strength* igual a 1, o *Wp* seria de 120 nm, portanto, tamanho mínimo para esta tecnologia de 65 nm. No caso do dimensionamento dos transistores da rede *pull-down*, além do fator de força, era considerado o fator de empilhamento de transistores em série. Entretanto, no caso da porta NAND2 LVT, tal equacionamento tem que ser melhorado, de modo a manter a metodologia de dimensionamento de células pela equalização dos tempos de subida e descida. Esta atualização é direcionada à rede *pull-up*, onde foi necessário, apesar de não possuir transistores em série, a utilização de um fator de incremento na largura dos transistores PMOS, denominado de $DELTA_p$. A Figura 3.23 apresenta o esquemático da porta NAND2X1 com as equações atualizadas. Cabe salientar que tentativas de aumento no comprimento do canal dos transistores NMOS, assim como utilizadas nas portas NOR, não apresentaram melhorias significativas.

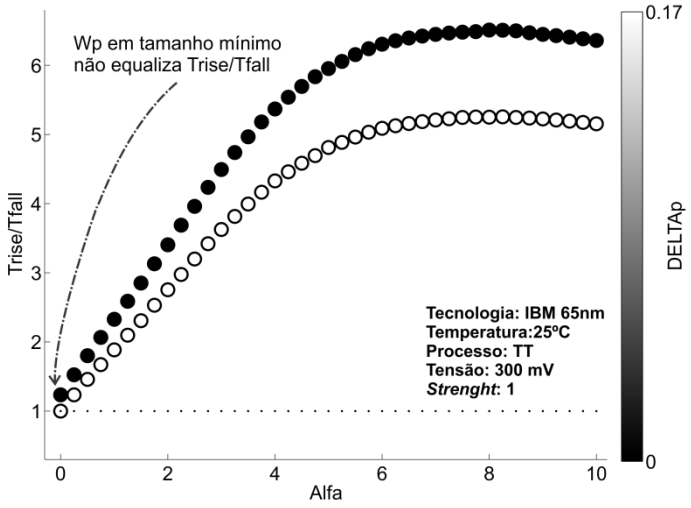
**Figura 3.23 - Dimensionamento da porta NAND2X1 LVT**




Na Figura 3.24 pode ser observado o comportamento dos tempos de subida/descida em função da variação de *ALFA*, para dois valores de $DELTA_P$, considerando uma NAND2X1. Torna-se evidente que mesmo com os transistores NMOS em tamanho mínimo, *ALFA*=0, a rede *pull-up*, em tamanho mínimo, $DELTA_P$=0, não consegue fazer o transiente da

saída para nível alto no mesmo tempo gasto pela rede *pull-down* no caso de levar a saída para nível baixo. Tal situação é representada pelos círculos escuros na Figura 3.24. Os círculos claros representam o comportamento da porta NAND com um pequeno incremento em *Wp* (cerca de 17%), mantendo-se os transistores NMOS em tamanho mínimo.

**Figura 3.24 - Trise/Tfall x ALFA x DELTAp para NAND2X1 LVT**




Na Tabela 3.17 as informações de dimensionamento e temporizações para NAND2X1 LVT são apresentadas. O atraso de equalização dos tempos de subida e descida é da ordem de 30,8 ns para a NAND de duas entradas com *strength* igual a um, com transistores *Low*-VT. Tal valor representa uma diminuição de 72,62% em relação ao mesmo atraso na NAND2X1 RVT. Por fim, analogamente às outras portas NAND, foram incluídas na biblioteca de transistores LVT mais duas células: NAND2X2 e NAND2X4.

**Tabela 3.17 - Dimensionamentos e temporizações para a NAND2X1 LVT em função do fator *ALFA* e *DELTAp* adotado**

| ALFA | Wn3,4 (nm) | Wn / Wmin | DELTAp | Wp1,2 (nm) | Wp / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|
| 0 | 120 | 1 | 0,17 | 140 | 1,17 | 60 |
| $t_r$=$t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | | | Temp (ºC) |
| 30,8 | 12,3 | 11 | 13,5 | | | 25 |

*3.3.3.3 NOR*

Analogamente às portas NOR RVT e HVT, a porta NOR LVT possui o comprimento do canal dos transistores NMOS aumentado em 50% de forma a reduzir o tamanho dos transistores PMOS em série na rede *pull-up*. Da mesma forma que nas outras implementações, os transistores de canal *n* estão com a largura em tamanho mínimo. A Figura 3.25 apresenta a

variação dos tempos de subida/descida em relação à variação de *ALFA* para a NOR2X1 LVT. O valor do fator de empilhamento que equaliza os tempos de *rise* e *fall* é de aproximadamente 10,58, o que resulta em um *Wp* de 1390 nm. Esta largura representa um aumento de 125,28% em relação à largura dos transistores em série na porta NOR com transistores RVT. Entretanto, a redução mínima nos atrasos de propagação foi de 87,66% no atraso de alto para baixo quando comparados ao mesmo atraso na NOR2X1 RVT. As informações de dimensionamento e temporizações para a NOR2X1 LVT estão condensadas na Tabela 3.18. Por fim, foram incluídas duas portas NOR adicionais na biblioteca LVT: NOR2X2 e NOR2X4.

**Figura 3.25 - Trise/Tfall x Alfa para porta NOR2X1 LVT com Ln=90 nm @ 300mV/25ºC**

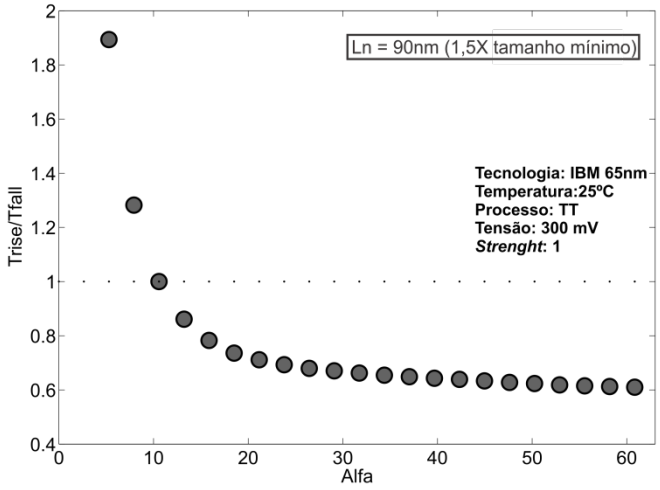



**Tabela 3.18 - Dimensionamentos e temporizações para a NOR2X1 LVT referentes ao ALFA e L do NMOS adotados**

| ALFA | Wp1,2 (nm) | Wp1,2 / Wmin | Wn3,4 (nm) | Wn3,4 / Wmin | Lp (nm) | Ln (nm) |
|---|---|---|---|---|---|---|
| 10,58 | 1390 | 11,58 | 120 | 1 | 60 | 90 |

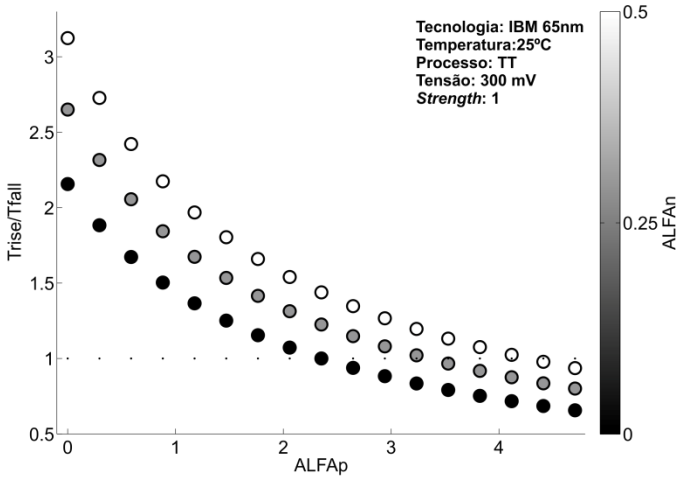
| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 12,2 | 13,1 | 14,7 | 11,4 | 25 |

*3.3.3.4 OAI21*

Esta implementação com transistores *Low*-VT adota os mesmos equacionamentos apresentados na Figura 3.11 para a OAI21 RVT. Desta forma, o transistor M1 foi mantido em tamanho mínimo enquanto que os transistores em paralelo com o referido, M2 e M3, foram reduzidos em 3,13% em relação aos seus equivalentes na OAI2X1 RVT. Diferentemente das outras duas implementações desta porta, no caso do uso de transistores LVT, é possível utilizar transistores de tamanho mínimo na rede *pull-down*. Com o intuito de reduzir área e

consequentemente, dissipação de potência, tal dimensionamento foi adotado. A Figura 3.26 apresenta a razão dos tempos de subida e descida em função da variação de $ALFA_P$ para três valores de $ALFA_N$. Os valores de $ALFA_N$ superiores à zero foram plotados com o intuito de ilustrar a necessidade de incrementar o tamanho da rede *pull-up* de modo a equalizar os tempos de subida e descida.

**Figura 3.26 - Trise/Tfall x ALFAp x ALFAn para porta OAI21X1 LVT**




No quesito atrasos de propagação, houve uma redução, obviamente, em todas as temporizações quando comparadas à OAI com transistores RVT. A menor redução foi de 78,67% no caso dos tempos de subida em condições de simetria ($t_r = t_f$). A Tabela 3.19 resume os dimensionamentos e temporizações para a OAI21X1 LVT. Além desta versão, foram adicionadas outras duas versões: OAI21X2 e OAI21X4.

**Tabela 3.19 - Dimensionamentos e temporizações para a OAI21X1 LVT**

| ALFAp | Wp1 (nm) | Wp2,3 (nm) | Wp2,3 / Wmin | ALFAn | Wn4-6 (nm) | Wn4-6 / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|---|
| 2,35 | 120 | 402 | 3,35 | 0 | 120 | 1 | 60 |

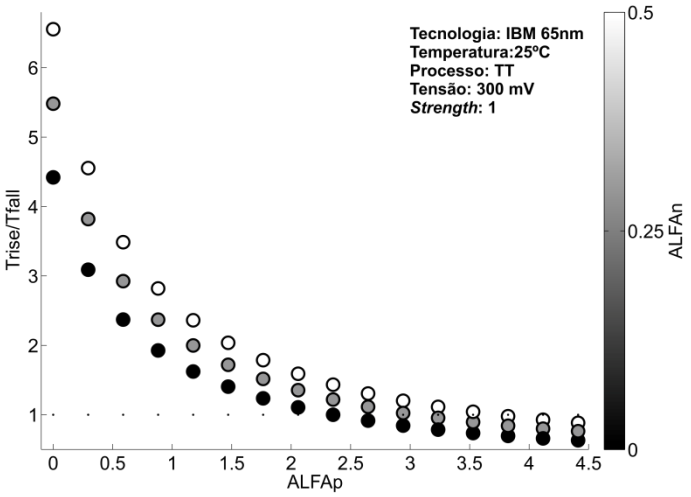
| $t_r=t_f$ (ns) | $t_p$ (ns) | $t_{pLH}$ (ns) | $t_{pHL}$ (ns) | Temp (ºC) |
|---|---|---|---|---|
| 24,3 | 13,1 | 13,5 | 12,7 | 25 |

*3.3.3.5 AOI22*

A versão da AOI22 com transistores LVT possui as mesmas regras de dimensionamento utilizadas nas implementações com transistores RVT e *High*-VT. Consequentemente, os transistores da rede PDN foram mantidos em tamanho mínimo enquanto que os transistores da rede PUN experimentaram um aumento de 3,35 vezes em relação à largura mínima, de forma a equalizar os tempos de subida aos tempos de descida.

Este aumento na largura dos transistores LVT supera em 42,55% o *upsizing* dos transistores da rede *pull-up*, no caso da AOI22X1 RVT. A Figura 3.27 ilustra os relacionamentos entre *Trise/Tfall*, $ALFA_P$ e $ALFA_N$. Da mesma forma que no caso da OAI21X1 LVT, os valores de $ALFA_N$ superiores à zero foram plotados com o intuito de ilustrar a necessidade de incrementar o tamanho da rede *pull-up* de modo a equalizar os tempos de subida e descida.

**Figura 3.27 - Trise/Tfall x ALFAp x ALFAn para porta AOI22X1 LVT**




Nas questões referentes à temporização, os atrasos desta implementação, comparados aos atrasos da AOI22X1 RVT, reduziram entre 86,67% a 89,54%. Tanto as informações de dimensionamento quanto temporizações para a AOI22X1 podem ser encontradas na Tabela 3.20. Por fim, duas células adicionais foram incluídas nesta biblioteca: AOI22X2 e AOI22X4.

**Tabela 3.20 - Dimensionamentos e temporizações para a AOI22X1 com transistores LVT**

| ALFAp | Wp1-4 (nm) | Wp / Wmin | ALFAn | Wn5-8 (nm) | Wn / Wmin | Lp=Ln (nm) |
|---|---|---|---|---|---|---|
| 2,35 | 402 | 3,35 | 0 | 120 | 1 | 60 |
| $t_r$=$t_f$ (ns) | | $t_p$ (ns) | $t_{pLH}$ (ns) | | $t_{pHL}$ (ns) | Temp (ºC) |
| 19,1 | | 14,7 | 15,5 | | 13,9 | 25 |

## 3.4 Células implementadas e resumo de dimensionamentos e temporizações

Neste trabalho foram implementadas 57 células distribuídas em três bibliotecas (RVT, HVT e LVT) utilizando transistores com diferentes tensões limiar (VT). Destas bibliotecas, a que possui a maior quantidade de células, 23, é a que utiliza transistores com tensão de limiar intermediária (RVT). O diferencial da referida biblioteca é a presença de dois registradores, cada um, em três *strengths* distintos. As outras duas bibliotecas, HVT e LVT, possuem

somente células combinacionais, 17 em cada, com funções lógicas idênticas. A Tabela 3.21 resume as células implementadas.

**Tabela 3.21 - Células incluídas em três bibliotecas com transistores multi-limiar para operação em *near*-VT**

| Bibliotecas | Células | X1 | X2 | X3 | X4 | X8 |
|---|---|---|---|---|---|---|
| RVT, HVT e LVT | INV | • | • | • | • | • |
| RVT, HVT e LVT | NAND2 | • | • | | • | |
| RVT, HVT e LVT | NOR2 | • | • | | • | |
| RVT, HVT e LVT | OAI21 | • | • | | • | |
| RVT, HVT e LVT | AOI22 | • | • | | • | |
| RVT | DFFR | • | • | | • | |
| RVT | DFFS | • | • | | • | |

Analisando as células combinacionais, com funções lógicas idênticas, distribuídas nas três bibliotecas, sob o ponto de vista de dimensionamento, percebe-se que não houve uma tendência de aumento ou redução na largura dos transistores HVT e LVT quando comparados a implementação com transistores RVT. Por exemplo, no caso das células HVT, as portas INVX1, NOR2X1 e OAI21X1 tiveram seu *Wp* reduzido em relação às implementações com transistores RVT. Entretanto, a largura foi incrementada nas funções NAND2X1 e AOI22X1. Nas células LVT, a NAND2X1 e a OAI21X1 tiveram reduções em *Wn* e *Wp*, respectivamente. Mas, no caso das células INVX1, NOR2X1 e AOI22X1, houve um aumento em *Wp*.

Entretanto, nas questões de temporização, a tendência é clara. As implementações com transistores de VT convencional (RVT) experimentaram atrasos intermediários em relação às outras duas bibliotecas. Obviamente, em função de possuir uma tensão de limiar intermediária em comparação aos transistores HVT e LVT. As células com transistores HVT, em função de possuírem limiares maiores em comparação às outras duas bibliotecas, obtiveram os maiores atrasos. Por exemplo, no caso dos tempos de subida, na situação de equalização com os tempos de descida, os atrasos foram, em média, 1905,26% superiores aos tempos das células com transistores RVT. Por outro lado, as células implementadas com transistores LVT obtiveram, em média, 84,35% de redução nos atrasos em relação aos MOSFETS RVT. A Tabela 3.22 resume os principais dimensionamentos e características de *timing* a $V_{DD}$=300 mV e 300°K, das células combinacionais projetadas neste capítulo, indicando as porcentagens de aumento/redução das implementações em HVT e LVT quando comparadas às suas respectivas funções projetadas com transistores RVT.

**Tabela 3.22 - Resumo comparativo dos principais dimensionamentos e temporizações para as células combinacionais projetadas**

| Porta | Transistor | Wp ou Wn * (nm) | ↑↓ | % | tr=tf (ns) | ↑↓ | % | tp (ns) | ↑↓ | % | tpLH (ns) | ↑↓ | % | tpHL (ns) | ↑↓ | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| INVX1 | HVT | 276 | ↓ | 3,5 | 1878 | ↑ | 1630,88 | 823,3 | ↑ | 1633,3 | 816,1 | ↑ | 1596,67 | 830,5 | ↑ | 1668,02 |
| | RVT | 286 | | | 108,5 | | | 47,5 | | | 48,1 | | | 47 | | |
| | LVT | 549 | ↑ | 91,96 | 8,3 | ↓ | 92,35 | 4,6 | ↓ | 90,32 | 4,7 | ↓ | 90,23 | 4,6 | ↓ | 90,21 |
| NAND2X1 | HVT | 383 | ↑ | 39,78 | 1325 | ↑ | 1077,78 | 1007,6 | ↑ | 1435,98 | 927,2 | ↑ | 1442,76 | 1088 | ↑ | 1429,59 |
| | RVT | 274 | | | 112,5 | | | 65,6 | | | 60,1 | | | 71,13 | | |
| | LVT [7] | 120 | ↓ | 56,2 | 30,8 | ↓ | 72,62 | 12,3 | ↓ | 81,25 | 11 | ↓ | 81,7 | 13,5 | ↓ | 81,02 |
| NOR2X1 | HVT | 234 | ↓ | 62,07 | 5266 | ↑ | 3587,68 | 2613,5 | ↑ | 2288,94 | 2751 | ↑ | 2074,7 | 2476 | ↑ | 2581,1 |
| | RVT | 617 | | | 142,8 | | | 109,4 | | | 126,5 | | | 92,35 | | |
| | LVT | 1390 | ↑ | 125,28 | 12,2 | ↓ | 91,46 | 13,1 | ↓ | 88,03 | 14,7 | ↓ | 88,38 | 11,4 | ↓ | 87,66 |
| OAI21X1 | HVT[8] | 224 | ↓ | 46,02 | 1864 | ↑ | 1536,52 | 1284 | ↑ | 1306,35 | 1292 | ↑ | 1231,96 | 1276 | ↑ | 1388,91 |
| | RVT[9] | 415 | | | 113,9 | | | 91,3 | | | 97 | | | 85,7 | | |
| | LVT | 402 | ↓ | 3,13 | 24,3 | ↓ | 78,67 | 13,1 | ↓ | 85,65 | 13,5 | ↓ | 86,08 | 12,7 | ↓ | 85,18 |
| AOI22X1 | HVT | 300 | ↑ | 6,38 | 2570 | ↑ | 1693,44 | 2039 | ↑ | 1445,87 | 2203 | ↑ | 1386,5 | 1875 | ↑ | 1523,38 |
| | RVT | 282 | | | 143,3 | | | 131,9 | | | 148,2 | | | 115,5 | | |
| | LVT | 402 | ↑ | 42,55 | 19,1 | ↓ | 86,67 | 14,7 | ↓ | 88,86 | 15,5 | ↓ | 89,54 | 13,9 | ↓ | 87,97 |

* Informa apenas a largura que foi aumentada em relação à largura mínima de 120 nm. Em termos de comprimento do canal, a maioria das portas utilizaram o mínimo, 60nm, excepcionalmente as portas NOR, das três bibliotecas, tiveram o seu valor aumentado em 50% com o intuito de reduzir a largura dos transistores da rede *pull-up*.

[7] *Wp* acima do mínimo

[8] *Wn* acima do mínimo

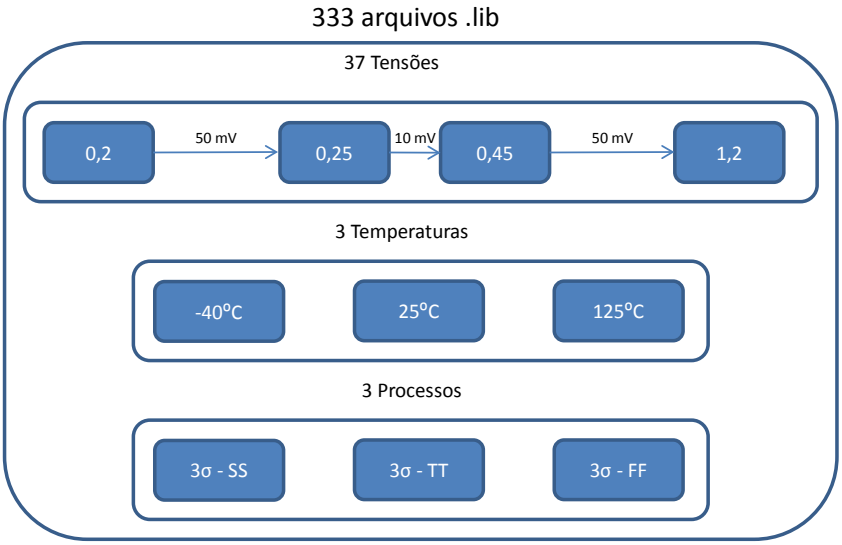
[9] *Wn* acima do mínimo

### 3.5 Metodologia de caracterização da biblioteca

Os circuitos combinacionais e sequenciais utilizando transistores RVT, dimensionados para operação a baixo $V_{DD}$, foram caracterizados através do *framework* em linguagem Python desenvolvidos por Stangherlin (2013), em conjunto com o ELC (*Encounter Library Characterizer*®). Em um primeiro momento, parâmetros como tempos de transição do sinal de entrada (*input slopes*) e cargas de saída (*output loads*) foram estimados por simulações SPICE para cada tensão a ser caracterizada, uma vez que os atrasos e até mesmo capacitâncias de entrada da porta são modificadas de acordo com a tensão de alimentação (STANGHERLIN, 2013). Estes valores foram tabelados e utilizados para determinar o tempo e o passo de cada caracterização realizada pelo ELC para cada célula da biblioteca.

No caso da biblioteca de células dimensionada para 450 mV desenvolvida no trabalho de Stangherlin (2013), a caracterização foi realizada para tensões de 150 mV a 1,2 V, com um passo de 10 mV, para três condições diferentes de processo: *slow* (3σ - SS), *fast* (3σ - FF) e *typical* (3σ - TT) à temperatura de 25°C. Neste trabalho foram consideradas as mesmas condições de processo. Entretanto no caso das tensões, as caracterizações foram realizadas de 200 mV a 1,2V com passo variável. No caso do trabalho de Stangherlin (2013), o ponto de mínima energia por operação (MEP) para os circuitos analisados situava-se entre 260 e 310 mV. Com o intuito de determinar o MEP com a mesma precisão, analisando os mesmos circuitos, para a biblioteca com transistores RVT dimensionada à 300 mV, o passo de caracterização de 10 mV foi mantido apenas para o intervalo de interesse, com certa margem de segurança: 250 mV a 450 mV. Abaixo e acima destas tensões, o passo foi aumentado para 50 mV. Adicionalmente, foram consideradas variações de temperatura: -40°C, 25°C e 125°C. Portanto, trinta e sete tensões foram contempladas, para três processos e três temperaturas, resultando em trezentos e trinta e três arquivos de biblioteca (.lib). Em cada um destes arquivos, informações de tensão, temperatura, temporizações, potências, entre outras, são registradas. A Figura 3.28 resume a metodologia de caracterização para a biblioteca de células com transistores RVT.

**Figura 3.28 - Metodologia de caracterização da biblioteca de células com transistores RVT**



# 4 RESULTADOS DA SÍNTESE LÓGICA DE CIRCUITOS CMOS "NEAR-VT"

## 4.1 Introdução

Neste capítulo serão apresentados os resultados de síntese da biblioteca desenvolvida com transistores RVT, em um PDK de 65 nm comercial, para dez circuitos de teste VLSI de média complexidade. Tais circuitos serão exercitados desde condições de sub-VT até super-VT, de modo a avaliar os relacionamentos entre energia consumida e frequência de operação, dando ênfase ao ponto de mínima energia por operação, para as três temperaturas e os três processos caracterizados. É importante salientar que, apesar da biblioteca ter sido dimensionada para a tensão de 300 mV, ela pode trabalhar em regimes de operação muito acima da tensão de limiar, se for necessário. Desta forma, caracterizando um regime de VFS dinâmico amplo, denominado por Stangherlin e Bampi (2013) de *very wide* VFS.

Os circuitos avaliados são compostos de um filtro digital *notch* (SOARES et al., 2013), composto por 14kcells, um núcleo compatível com um micro-controlador 8051 de 14,5kcells, quatro circuitos de *benchmark* ISCAS (HANSEN; YALCIN; HAYES, 1999) combinacionais (C432, C1355, C3540 e C6288) e quatro circuitos sequenciais (S420, S1423, S9234 e S38584). Anteriormente à discussão do ponto de mínima energia para diferentes temperaturas e processos, será apresentada a metodologia de potência e análise de *timing* adotada. Por fim, serão discutidos efeitos da introdução de uma diversidade maior de células combinacionais nos resultados de síntese e uma comparação com os resultados de um trabalho relacionado será realizada.

## 4.2 Metodologia de Análise de Potência e de "Timing"

A metodologia de análise de potência e de atraso utilizada neste trabalho foi realizada através do *framework* em linguagem Python desenvolvido por Stangherlin (2013) em conjunto com as ferramentas comerciais de EDA (*Electronic Design Automation*) do ambiente Encounter™. Para cada circuito de teste, o mapeamento tecnológico foi feito para $V_{DD}$=1,2 V e o mesmo *netlist* foi usado para estimar potências e atrasos para todas as tensões de alimentação em um ambiente *multi-mode multi corner* (MMMC). Os dados de energia do

filtro digital *notch* foram extraídos através da computação de 2048 amostras de sinais de eletroencefalografia. Para o núcleo compatível com o 8051, tais informações foram obtidas através de 10 iterações de laço em ponto fixo do *benchmark* Dhrystone. No caso dos circuitos ISCAS, os dados de energia foram extraídos pela computação de 4096 valores randômicos de entrada. A máxima frequência de operação para cada $V_{DD}$ foi obtida através da extrapolação do tempo de *slack* do caminho crítico. Ou seja, para cada $V_{DD}$ (.lib caracterizado) um alvo de frequência foi estipulado e após a etapa de síntese o tempo de *slack* foi analisado. Se o mesmo for superior a zero (não negativo), significa que para aquela tensão de alimentação o circuito poderia operar a uma frequência maior. A avaliação de desempenho dos quatro circuitos combinacionais ISCAS foram realizadas através da inserção de registradores de saída no *netlist* original (STANGHERLIN, 2013). Cabe salientar que estas análises foram realizadas apenas para circuitos em cuja síntese foi feito o mapeamento para a biblioteca que contém células com transistores RVT, uma vez que apenas essa biblioteca por ora possui registradores nela incluídos.

## 4.3 Análise do Ponto de Mínima Energia

Nesta seção, serão apresentados os resultados para o ponto de mínima energia sob condições de variações extremas de temperatura, para os dez circuitos de teste avaliados. Primeiramente, será realizada uma análise para a temperatura de 25°C. Posteriormente, 125°C e por fim, os circuitos serão submetidos à -40°C.

### 4.3.1 MEP à 25°C

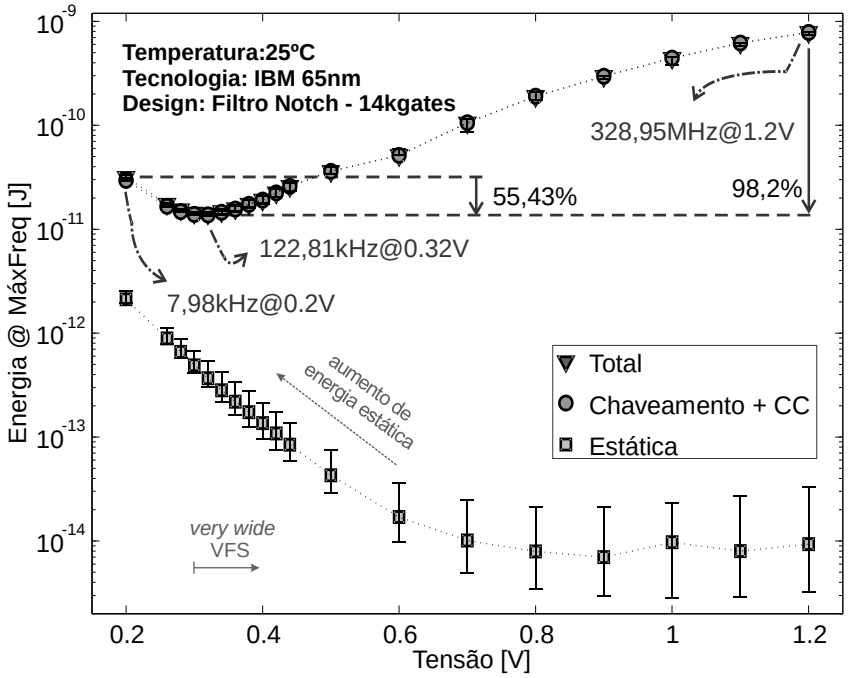
A Figura 4.1 apresenta os resultados de energia consumida na computação de 2048 amostras de sinais de eletroencefalografia, sob condições de máxima frequência em função da tensão de alimentação, para o filtro *notch*. As curvas superiores do gráfico correspondem a energia total consumida juntamente com a energia de chaveamento e curto-circuito. Percebe-se que há um aumento significativo da energia estática à medida que a tensão de alimentação diminui. Entretanto, para a temperatura de 25°C, tal energia não colabora substancialmente no consumo total (menos de 7% em 200 mV, onde alcança seu ápice para os pontos analisados). Neste gráfico, três pontos distintos de operação são marcados: sub-limiar, em 200 mV, ponto de mínimo consumo energético por operação, em 320 mV, e sob regime de super-limiar em tensão nominal, à 1,2 V. O consumo energético operando em *near*-VT é reduzido em torno de 55,43% quando comparado ao ponto de 200 mV. Adicionalmente, o ganho de desempenho é

substancial, ultrapassando 1400%. Em relação à tensão nominal, operar no ponto de mínima energia provê até 98,2% de economia de energia.

É importante notar que o MEP deslocou-se 20 mV para à direita do ponto de dimensionamento da biblioteca de células. Tal ponto ainda está consideravelmente abaixo da tensão de limiar dos transistores nesta temperatura e processo. Se o filtro operar em 300 mV, consumirá aproximadamente 2% a mais que o ponto de mínima energia. Entretanto, a variação de frequência é significativa, sendo reduzida em aproximadamente 37 % em relação ao MEP. Desta forma, evidencia-se que um pequeno aumento na tensão de alimentação resulta em grandes benefícios em desempenho.

Adicionalmente, para cada ponto de tensão discretizado, seja ele referente a energia total, chaveamento mais curto-circuito ou estática, existem barras de erro que indicam a variabilidade de energia quando considerados os *corners* de processo SS e FF. No caso da energia estática em tensões próximas à nominal, é possível observar uma grande variabilidade de processo. Entretanto, não representam grandes contribuições para a energia total. Cabe salientar que os pontos plotados no gráfico não representam todos os pontos caracterizados. Esta omissão foi intencional, com o intuito de facilitar a visualização dos dados importantes.

**Figura 4.1 - Energia sob condições de máxima frequência em função da tensão de alimentação para o filtro notch à 25ºC**



Na Tabela 4.1 são apresentados os resultados de energia, frequência e total de células lógicas para os dez circuitos de teste, em três condições de operação, para o processo típico à 25°C. Adicionalmente, para os pontos de tensão referenciados, existe uma coluna denominada de "Razão" que indica uma normalização da energia estática sobre a energia total. A partir desta informação, torna-se evidente o aumento daquela energia à medida que a tensão diminui, representando até 28,03% para o circuito sequencial S38584, sob operação em regime de sub-limiar. No ponto de mínima energia, sob condição de *near*-VT, o circuito que apresenta o maior consumo estático é o S1423, no qual atinge 10,9%. Entretanto, os dados mensurados não condizem com a afirmação de que o ponto de mínima energia é alcançado por um equilíbrio da energia de chaveamento com a energia estática (DE, 2013). Nas informações referentes ao MEP, é possível observar que para sete dos dez circuitos, o ponto de mínimo consumo energético deslocou-se 10 mV para a direita da tensão de dimensionamento da biblioteca de células, ou seja, 310 mV, portanto, abaixo de VT.

Adicionalmente, do ponto de mínima energia até o ponto de operação muito acima de VT, ou super-VT, a biblioteca projetada pode prover uma ampla faixa de escalamento em desempenho, com frequências máximas variando de centenas de kHz até a ordem de MHz/GHz, sendo, por exemplo, suficientemente superior às necessidades de desempenho do filtro *notch* para aplicações médicas (SOARES et al., 2013).

**Tabela 4.1 - Resultados de energia e frequência para os circuitos de teste em três condições de operação: sub, *near* e super-VT à 25ºC**

| TT @ 25ºC | | sub-VT @ 0.2 V | | | *near*-VT @ 0.3 V | | | MEP | | | | super-VT@ 1.2V | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CELLS | DESIGN | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | TENSÃO [V] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGY [J] | RAZÃO [%] | FREQ [Hz] |
| 14087 | ***Notch*** | 31,66p | 6,69 | 7,98k | 14,39p | 3,41 | 77,45k | 0,32 | 14,11p | 2,61 | 122,81k | 783,81p | 0,00 | 328,95M |
| 14538 | **8051** | 2,48p | 20,78 | 30,16k | 1,18p | 9,45 | 319,8k | 0,31 | 1,12p | 8,54 | 403,76k | 85,86p | 0,00 | 1,39G |
| 151 | **C432** | 90,24f | 8,56 | 27,51k | 41,91f | 4,17 | 277,12k | 0,31 | 40,79f | 3,72 | 349,31k | 5,27p | 0,00 | 1,19G |
| 354 | **C1355** | 422,58f | 4,9 | 30,81k | 208,83f | 2,23 | 307,44k | 0,31 | 205,02f | 1,96 | 386,84k | 16,36p | 0,00 | 1,28G |
| 840 | **C3540** | 273,93f | 9,38 | 19,82k | 130,57f | 4,33 | 205,5k | 0,31 | 127,26f | 3,83 | 259,7k | 10,08p | 0,00 | 829,88M |
| 1945 | **C6288** | 6,62p | 2,52 | 10,38k | 2,71p | 1,45 | 106,29k | 0,32 | 2,61p | 1,15 | 169,81k | 152,17p | 0,00 | 434,59M |
| 143 | **S420** | 29,4f | 22 | 28,71k | 14,86f | 10,06 | 306,22k | 0,31 | 14,5f | 8,98 | 387,54k | 5,55p | 0,00 | 1,35G |
| 535 | **S1423** | 242,8f | 27,27 | 15,29k | 115,36f | 12,32 | 164,65k | 0,31 | 112,03f | 10,9 | 208,73k | 26,31p | 0,00 | 681,2M |
| 913 | **S9234** | 267,13f | 17,98 | 36,49k | 144,05f | 7,53 | 379,42k | 0,31 | 140,48f | 6,71 | 479,29k | 45,04p | 0,00 | 1,53G |
| 8962 | **S38584** | 2,58p | 28,3 | 25,11k | 1,04p | 14,7 | 278,02k | 0,33 | 982,19f | 9,8 | 569,61k | 129,01p | 0,00 | 1,18G |

**Tabela 4.2 - Resultados de energia e frequência para os circuitos de teste em três condições de operação: sub, *near* e super-VT à 125ºC**

| TT @ 125ºC | | sub-VT @ 0.2 V | | | *near*-VT @ 0.3 V | | | MEP | | | | super-VT@ 1.2V | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CELLS | DESIGN | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | TENSÃO [V] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGY [J] | RAZÃO [%] | FREQ [Hz] |
| **14087** | ***Notch*** | 75,5p | 14,64 | 108,05k | 23,72p | 22,11 | 544,05k | 0,36 | 18,78p | 16,88 | 1,41M | 547,6p | 0,06 | 310,27M |
| **14538** | **8051** | 8,66p | 24,73 | 430,98k | 2,94p | 31,85 | 2,33M | 0,4 | 1,8p | 20,59 | 11,85M | 60,34p | 0,09 | 1,3G |
| **151** | **C432** | 237,72f | 15,44 | 388,94k | 72,58f | 23,93 | 1,96M | 0,35 | 56,27f | 20,08 | 4,39M | 3,75p | 0,03 | 1,12G |
| **354** | **C1355** | 976,61f | 10,18 | 409,9k | 301,15f | 15,42 | 2,03M | 0,34 | 255,9f | 12,85 | 3,85M | 21,75p | 0,01 | 1,19G |
| **840** | **C3540** | 640,54f | 18,45 | 266,98k | 200,21f | 26,7 | 1,37M | 0,34 | 166,55f | 22,03 | 2,65M | 14,71p | 0,02 | 796,18M |
| **1945** | **C6288** | 14,63p | 6,04 | 137,99k | 3,93p | 10,97 | 711,7k | 0,35 | 3,04p | 9,13 | 1,61M | 110,3p | 0,02 | 406,5M |
| **143** | **S420** | 92,85f | 30,03 | 409,52k | 30,91f | 42,66 | 2,21M | 0,38 | 22,8f | 27,4 | 8,16M | 10,56p | 0,01 | 1,26G |
| **535** | **S1423** | 738,45f | 38,09 | 216,69k | 251,65f | 48,36 | 1,18M | 0,39 | 182,89f | 27,66 | 5,21M | 47,53p | 0,02 | 632,11M |
| **913** | **S9234** | 839,34f | 24,17 | 51,55k | 280,4f | 33,99 | 2,66M | 0,38 | 212,36f | 21,4 | 9,79M | 75,3p | 0,01 | 1,47G |
| **8962** | **S38584** | 7,66p | 41,6 | 349,17k | 2,56p | 54,35 | 1,89M | 0,4 | 1,66p | 32,31 | 9,52M | 205,22p | 0,04 | 1,1G |

### 4.3.2 MEP à 125ºC

A Figura 4.2 ilustra a energia consumida em função da variação da tensão de alimentação para o filtro *notch* na temperatura de 125°C. Da mesma forma que na análise para 25ºC, os mesmos três pontos foram evidenciados no gráfico: sub-VT, *near*-VT e super-VT. O consumo energético, quando em operação no MEP, é reduzido em 75,13% em relação ao ponto de 200 mV, em inversão fraca, enquanto que os ganhos em frequência são da ordem de 1200%. Quando comparado à inversão forte, operar no MEP resulta em ganhos de economia energética na ordem de 96,57%.

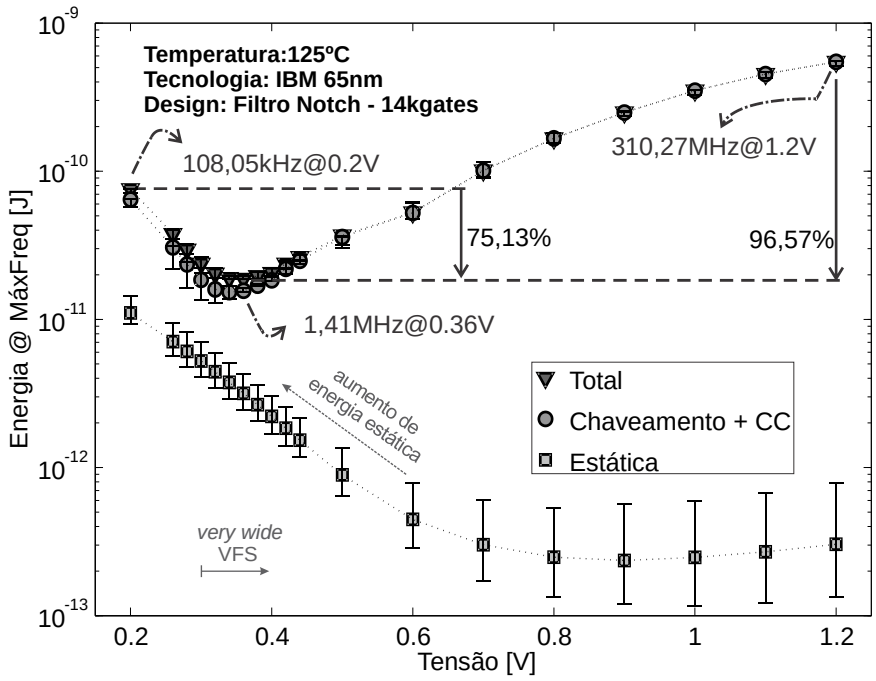
Entretanto, é importante notar dois aspectos: O aumento da frequência de operação para os regimes de inversão fraca e moderada, e o deslocamento do ponto de mínima energia para a direita, 60 mV, do ponto de dimensionamento da biblioteca de células. Os dois fenômenos estão relacionados com a redução da tensão de limiar  em função do aumento da temperatura, que, por sua vez, resulta em aumento de desempenho, e pode ser facilmente comprovado observando a Figura 4.1 e a Figura 4.2. Em 25°C, as frequências são de 7,98 kHz e 122,81 kHz para sub e *near*-VT, enquanto que para 125°C, aumentaram para 108,05 kHz e 1,41 MHz, respectivamente. Adicionalmente, observando a Tabela 4.2, torna-se evidente a variabilidade do MEP em função do aumento de temperatura. Existem seis tensões distintas, com variação de 60 mV, para o ponto de mínima energia e não há mais do que dois valores iguais para os dez circuitos de teste. Quando comparada a Tabela 4.1, referente à temperatura de 25°C, o MEP distribui-se em apenas três tensões, variando de 310 a 330mV , sendo que para a menor tensão, houve uma repetição de sete em dez casos. Por fim, houve uma pequena redução na frequência de operação no regime de tensão nominal, devido ao fato da mobilidade de portadores ser reduzida com o aumento da temperatura. Esse efeito é dominante em tensões acima de 1V (STANGHERLIN, 2013).

Outro fator que torna-se relevante com a elevação da temperatura é o aumento da energia estática. No caso do MEP, para o filtro *notch* em 25°C, tal parcela representa apenas 2,61% da energia total. Em 125°C, a energia estática contribui com 16,88% do total, no ponto de mínima energia. A Tabela 4.2 apresenta os resultados de energia e frequência para os dez circuitos de teste, no processo típico, à 125ºC. Para o caso do circuito sequencial S38584, a parcela estática da energia representa 41,6% do total no regime de sub-VT. No MEP, o referido circuito também possui a maior componente estática, representando 32,31% do total.

Em termos de frequência máxima de operação no MEP, todos circuitos possuem um desempenho máximo na ordem de MHz. Uma observação importante sobre a operação a

temperaturas altas: embora o MEP à 125°C represente um incremento médio de 34,5% em energia despendida por operação, a frequência máxima atingível no MEP é significativamente maior (tipicamente de 10X a 20X) do que à temperatura ambiente. Isto porque o coeficiente térmico da variação de VT favorece a operação a temperaturas mais altas.

**Figura 4.2 - Energia sob condições de máxima frequência em função da tensão de alimentação para o filtro *notch* à 125ºC**



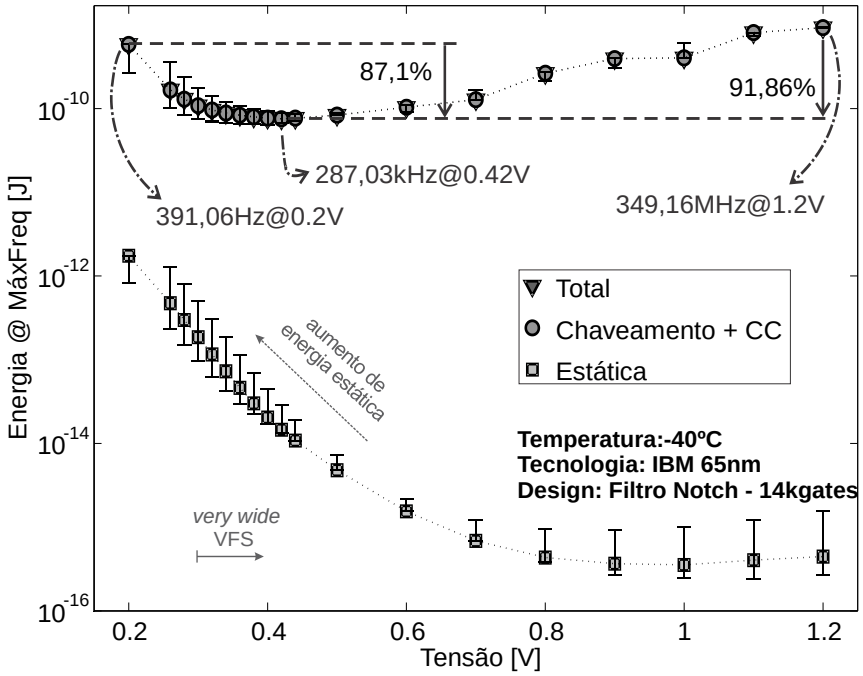

### 4.3.3 MEP à -40ºC

A Figura 4.3 apresenta o comportamento da energia em função da variação da tensão de alimentação para o filtro *notch* sob as mesmas condições analisadas nos casos anteriores, com exceção da variável temperatura, agora em -40°C. É possível observar que, graficamente, o ponto de mínima energia não representa mais um vale tão acentuado, como nos gráficos referentes às temperaturas de 25°C e 125°C. Adicionalmente, o ponto de mínima energia por operação deslocou-se ainda mais para à direita, superando a casa dos 400 mV, mais precisamente, alcançou os 420 mV. Neste ponto, a redução de consumo em relação ao regime de sub-VT é da ordem de 87,1% e os ganhos em frequência são da ordem 734X. Em relação à tensão nominal, os ganhos em energia ultrapassam 91%.

Cabe salientar que no ponto de sub-limiar, em 200 mV, a energia total consumida é superior ao consumo energético em regime de inversão forte, na tensão de alimentação de 1

V. Adicionalmente, a frequência de operação é drasticamente reduzida, sendo inviável, na operação de um filtro *notch*, por exemplo.

**Figura 4.3 - Energia sob condições de máxima frequência em função da tensão de alimentação para o filtro *notch* à -40ºC**




Os resultados de energia e frequência para os dez circuitos de teste no processo TT, à -40ºC são apresentados na Tabela 4.3. É importante observar que a redução de temperatura, bem como o aumento no caso de 125°C, impactou substancialmente na distribuição do ponto de mínima energia. A variação neste caso, distribui-se numa faixa de 190 mV, separadas em sete tensões distintas. Em duas situações, a tensão foi inferior ao ponto de dimensionamento: 0,26 e 0,28 V. Nelas a parcela de energia estática é bem superior aos outros casos, tanto operando em sub-VT quanto no MEP. Entretanto, obviamente, tal consumo é inferior quando comparado aos seus respectivos circuitos em 25°C. Para o circuito sequencial S9234, o ponto de mínima energia foi exatamente o ponto de tensão de dimensionamento da biblioteca, 300 mV, resultando em ganhos de energia da ordem 62% e de frequência acima dos 2200%, quando comparados ao ponto de sub-VT, em 200 mV.

**Tabela 4.3 - Resultados de energia e frequência para os circuitos de teste em três condições de operação: sub, *near* e super-VT à -40°C**

| TT @ -40°C | | sub-VT @ 0.2 V | | | *near*-VT @ 0.3 V | | | MEP | | | | super-VT@ 1.2V | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CELLS | DESIGN | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | TENSÃO [V] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGY [J] | RAZÃO [%] | FREQ [Hz] |
| **14087** | ***Notch*** | 584,87p | 0,3 | 391,06 | 108,33p | 0,17 | 8,3k | 0,42 | 75,47p | 0,02 | 287,03k | 926,68p | 0,00 | 349,16M |
| **14538** | **8051** | 19,71p | 2,49 | 1,52k | 7,59p | 0,66 | 33,55k | 0,41 | 6,28p | 0,06 | 918,87k | 194,64p | 0,00 | 1,47G |
| **151** | **C432** | 1,43p | 0,43 | 1,33k | 294,84f | 0,2 | 30,03k | 0,42 | 218,58f | 0,02 | 1,06M | 6,38p | 0,00 | 1,24G |
| **354** | **C1355** | 6,69p | 0,3 | 1,57k | 1,53p | 0,13 | 33,3k | 0,43 | 1,17p | 0,01 | 1,5M | 31,13p | 0,00 | 1,33G |
| **840** | **C3540** | 13,99p | 0,56 | 542,83 | 1,04p | 0,39 | 22,83k | 0,43 | 692,56f | 0,03 | 1,05M | 20,79p | 0,00 | 888,89M |
| **1945** | **C6288** | 135,59p | 0,14 | 497,6 | 21,47p | 0,09 | 10,87k | 0,45 | 14,28p | 0,01 | 835,29k | 184,03p | 0,00 | 458,93M |
| **143** | **S420** | 79,72f | 5,92 | 1,44k | 50,73f | 0,79 | 38,83k | 0,26 | 21,88f | 5,43 | 9,65k | 17,48p | 0,00 | 1,42G |
| **535** | **S1423** | 1,63p | 3,55 | 747,57 | 620,25f | 0,9 | 17,24k | 0,28 | 582,02f | 1,55 | 9,175k | 75,32p | 0,00 | 719,42M |
| **913** | **S9234** | 2,73p | 1,83 | 1,5k | 835,67f | 0,48 | 39,66k | 0,3 | 835,67f | 0,48 | 39,66k | 111,28p | 0,00 | 1,66G |
| **8962** | **S38584** | 25,14p | 2,62 | 1,26k | 6,16p | 1,02 | 29,2k | 0,45 | 4,964p | 0,04 | 2,236M | 310,75p | 0,00 | 1,25G |

**Tabela 4.4 - Resultados da inserção da OAI21 e AOI22 (colunas em branco) na biblioteca de células com transistores RVT operando a 300 mV**

| DESIGN | CELLS | CELLS | ↑↓ | % | ENERGIA [J] | ENERGIA [J] | ↑↓ | % | FREQ [Hz] | FREQ [Hz] | ↑↓ | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Notch*** | 14087 | 24405 | ↓ | 42,28 | 14,39p | 12,22p | ↑ | 17,76 | 77,45k | 65,9k | ↑ | 17,53 |
| **8051** | 14538 | 20420 | ↓ | 28,81 | 1,18p | 1,1p | ↑ | 7,27 | 319,8k | 353,22k | ↓ | 9,46 |
| **C432** | 151 | 182 | ↓ | 17,03 | 41,91f | 42,5f | ↓ | 1,39 | 277,12k | 262,96k | ↑ | 5,38 |
| **C1355** | 354 | 558 | ↓ | 36,56 | 208,83f | 151,16f | ↑ | 38,15 | 307,44k | 301,63k | ↑ | 1,93 |
| **C3540** | 840 | 1207 | ↓ | 30,41 | 130,57f | 138,85f | ↓ | 5,96 | 205,5k | 182,1k | ↑ | 12,85 |
| **C6288** | 1945 | 2769 | ↓ | 29,76 | 2,71p | 3,08p | ↓ | 12,01 | 106,29k | 91,5k | ↑ | 16,16 |
| **S420** | 143 | 188 | ↓ | 23,94 | 14,86f | 14,95f | ↓ | 0,60 | 306,22k | 306,48k | ↓ | 0,08 |
| **S1423** | 535 | 673 | ↓ | 20,51 | 115,36f | 111,55f | ↑ | 3,42 | 164,65k | 149,66k | ↑ | 10,02 |
| **S9234** | 913 | 1223 | ↓ | 25,35 | 144,05f | 149,13f | ↓ | 3,41 | 379,42k | 350,6k | ↑ | 8,22 |
| **S38584** | 8962 | 12702 | ↓ | 29,44 | 1,04p | 953,24f | ↑ | 9,10 | 278,02k | 271,28k | ↑ | 2,48 |

## 4.4 Comparações

Nesta seção, serão discutidos os efeitos da introdução de uma diversidade maior de células combinacionais nos resultados de síntese, com transistores RVT, dos dez circuitos de teste analisados neste trabalho. Posteriormente, será realizada uma comparação dos resultados de energia e frequência obtidos neste estudo em relação ao trabalho de Stangherlin (2013).

### 4.4.1 Efeitos da introdução de uma diversidade maior de células

Em um estudo preliminar, foi realizado um dimensionamento das três células combinacionais e duas sequenciais, e seus respectivos *strengths*, apresentadas no trabalho de Stangherlin (2013) para a tensão de operação em 300 mV, com transistores RVT. Os valores adotados para as variáveis *R* do inversor, *ALFA* da NAND e NOR, bem como o incremento de $L_N$ na porta NOR são idênticos aos valores adotados neste trabalho, para as respectivas portas. Posteriormente, tais células foram caracterizadas utilizando os mesmos *corners* de temperatura e processo deste trabalho, para a mesma faixa de tensões. Entretanto, o passo definido para a caracterização foi de 50 mV, com o intuito de reduzir o tempo de computação. Por fim, os mesmos circuitos de teste deste trabalho foram sintetizados utilizando a mesma metodologia. Os resultados do total de instâncias de células, de energia e de máxima frequência para a tensão de 300 mV, com transistores RVT, e processo típico à 25ºC, estão listados nas colunas em cinza, na Tabela 4.4. Na referida tabela, as colunas em branco à esquerda daquelas, representam os resultados deste trabalho, que incluem adicionalmente duas células combinacionais em três *strengths*, para as mesmas condições. Cabe salientar que a tensão de 300 mV resultou, no trabalho preliminar, no ponto de mínima energia por operação em nove dos dez circuitos. Entretanto, como a precisão da tensão não foi a mesma no processo de caracterização, o comparativo será realizado para a tensão de dimensionamento.

A partir da referida tabela, percebe-se que a quantidade total de instâncias de células foi reduzida substancialmente com a introdução das células OAI21 e AOI22. A maior redução foi no circuito de maior complexidade, o filtro *notch* (42,28%), e a menor ocorreu no combinacional C432, 17,03%, o mais simples dos circuitos de *benchmark*. Em média, a redução foi de 28,41%. Em termos de consumo de energia, foram obtidos resultados semelhantes. Entretanto, nas situações de maior consumo da nova biblioteca, com maior variedade de células, o consumo de energia está na mesma ordem de grandeza. No pior caso, bem acima das outras situações, o aumento foi de 38,15%. Com relação ao desempenho, a

nova biblioteca obteve maiores frequências em oito dos dez circuitos. O maior ganho foi no caso do filtro *notch*, 17,53%. Em média, a introdução de uma diversidade maior de células contribuiu com uma melhora de desempenho na ordem de 9,3%.

### 4.4.2 Comparação com (STANGHERLIN, 2013)

Como dito anteriormente, os dez circuitos testados nas subseções precedentes foram os mesmos utilizados no trabalho de Stangherlin (2013). No referido estudo, o ponto de mínima energia por operação, para o processo típico à 25°C, situava-se entre 260 e 310 mV. No presente trabalho, para as mesmas condições de processo e temperatura, a variação do MEP foi de 310 à 330 mV. Comparando-se circuito à circuito, a variação, para a maioria dos casos, não ultrapassou 30 mV. Somente para o circuito sequencial C432, a diferença alcançou os 50 mV. Tais valores foram relacionados na primeira metade da Tabela 4.5. Nas colunas em branco, estão os resultados do presente trabalho, enquanto que nas colunas em cinza, foram representados os valores obtidos por Stangherlin (2013). Além das tensões referentes ao MEP, foram incluídas: energia, razão entre energia estática e total, e frequência de operação. Tais valores foram considerados na segunda metade da Tabela 4.5, para as tensões de 450 mV. A referida tensão foi escolhida por um motivo: ser o valor de dimensionamento da biblioteca de células de Stangherlin (2013). Adicionalmente, pelo fato da biblioteca de células deste trabalho também poder operar em uma ampla faixa de tensões, desde condições de inversão moderada às tensões muito acima de VT.

Um complemento da Tabela 4.5 é apresentado na Tabela 4.6, onde são agrupadas e comparadas: energia e frequência para as duas situações apresentadas na primeira tabela. Desta forma, observa-se que para oito dos dez circuitos testados, tanto no caso do MEP quanto na tensão de 450 mV, houve redução de consumo energético. É importante salientar que os mesmos circuitos foram os mais econômicos, nos dois pontos analisados. Para o caso do mínimo ponto de energia por operação, a economia foi, em média, 24,1% menor. Em 450 mV, a média de redução foi de 20,57%. Os dois circuitos que apresentaram maior consumo, na implementação deste trabalho, foram o filtro *notch* e o combinacional C1355. No MEP, a média foi aproximadamente 33% superior e, na tensão de 450 mV, tal valor ultrapassou 44%. Entretanto, em todos os circuitos, nos dois pontos de operação analisados, houve menor consumo de energia estática, neste trabalho. No pior caso, tal parcela representou 10,9%, enquanto que no trabalho de Stangherlin (2013), alcançou 25,25%, para o mesmo circuito sequencial (S1423).

Em termos de desempenho, nas vinte situações analisadas, houve melhorias expressivas quando utilizada a biblioteca com maior diversidade de células combinacionais. Considerando apenas as situações onde houve melhorias em termos de consumo, a média de aumento na frequência de operação, para o MEP, é de 152,68%. Nas mesmas condições, o aumento é de 47,87% para a tensão de 450 mV. Portanto, para os circuitos analisados, se a prioridade for economia de energia, excepcionalmente desconsiderando o filtro *notch* e o circuito combinacional C1355, ou se a primazia for desempenho, para todos os casos, a biblioteca de células dimensionada para 300 mV é a recomendada para operação em NTV.

**Tabela 4.5 - Resultados de energia e frequência obtidos neste estudo *versus* resultados de Stangherlin (2013) para transistores RVT**

| TT@25ºC | MEP | | | | MEP (STANGHERLIN, 2013) | | | | *near*-VT @ 0.45 V | | | *near*-VT @ 0.45 V(STANGHERLIN, 2013) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DESIGN | TENSÃO [V] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | TENSÃO [V] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] | ENERGIA [J] | RAZÃO [%] | FREQ [Hz] |
| ***Notch*** | 0,32 | 14,11p | 2,61 | 122,81k | 0,3 | 10,27p | 7,95 | 51,26k | 28,02p | 0,27 | 2,01M | 18,79p | 0,72 | 1,27M |
| **8051** | 0,31 | 1,12p | 8,54 | 403,76k | 0,28 | 1,37p | 15,95 | 228,62k | 2,4p | 0,55 | 8,88M | 2,53p | 1,03 | 7,55M |
| **C432** | 0,31 | 40,79f | 3,72 | 349,31k | 0,26 | 43,16f | 15,15 | 76,1k | 89,11f | 0,27 | 7,41M | 93,48f | 0,54 | 4,69M |
| **C1355** | 0,31 | 205,02f | 1,96 | 386,84k | 0,29 | 159,59f | 7,12 | 162,68k | 429,03f | 0,14 | 8,21M | 307,1f | 0,53 | 5,06M |
| **C3540** | 0,31 | 127,26f | 3,83 | 259,7k | 0,28 | 288,91f | 11,19 | 100,12k | 269,97f | 0,26 | 5,62M | 596,2f | 0,56 | 4,05M |
| **C6288** | 0,32 | 2,61p | 1,15 | 169,81k | 0,29 | 2,69p | 2,81 | 72,36k | 4,91p | 0,12 | 2,82M | 5,03p | 0,22 | 2,3M |
| **S420** | 0,31 | 14,5f | 8,98 | 387,54k | 0,29 | 16,37f | 18,59 | 177,53k | 42,23f | 0,45 | 8,64M | 46,6f | 0,85 | 5,58M |
| **S1423** | 0,31 | 112,03f | 10,9 | 208,73k | 0,29 | 141,39f | 25,25 | 90,09k | 268,7f | 0,62 | 4,56M | 313,94f | 1,3 | 2,69M |
| **S9234** | 0,31 | 140,48f | 6,71 | 479,29k | 0,29 | 276,68f | 16,48 | 184,16k | 352,26f | 0,39 | 10,13M | 667,52f | 0,78 | 5,85M |
| **S38584** | 0,33 | 982,19f | 9,8 | 569,61k | 0,31 | 1,37p | 19,95 | 271,89k | 1,93p | 1,03 | 6,94M | 2,63p | 1,49 | 5,61M |

**Tabela 4. 6 - Resumo comparativo de energia e frequência deste estudo *versus* resultados de Stangherlin (2013) para transistores RVT**

| TT@25ºC | Energia e Frequência @ MEP x MEP (STANGHERLIN) em cinza | | | | | | | | Energia e Frequência @ 0.45 V x 0.45 V (STANGHERLIN) em cinza | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DESIGN | ENERGIA [J] | ENERGIA [J] | ↑↓ | % | FREQ [Hz] | FREQ [Hz] | ↑↓ | % | ENERGIA [J] | ENERGIA [J] | ↑↓ | % | FREQ [Hz] | FREQ [Hz] | ↑↓ | % |
| ***Notch*** | 14,11p | 10,27p | ↑ | 37,39 | 122,81k | 51,26k | ↑ | 139,58 | 28,02p | 18,79p | ↑ | 49,12 | 2,01M | 1,27M | ↑ | 58,27 |
| **8051** | 1,12p | 1,37p | ↓ | 18,25 | 403,76k | 228,62k | ↑ | 76,61 | 2,4p | 2,53p | ↓ | 5,14 | 8,88M | 7,55M | ↑ | 17,62 |
| **C432** | 40,79f | 43,16f | ↓ | 5,49 | 349,31k | 76,1k | ↑ | 359,01 | 89,11f | 93,48f | ↓ | 4,67 | 7,41M | 4,69M | ↑ | 58,00 |
| **C1355** | 205,02f | 159,59f | ↑ | 28,47 | 386,84k | 162,68k | ↑ | 137,79 | 429,03f | 307,1f | ↑ | 39,70 | 8,21M | 5,06M | ↑ | 62,25 |
| **C3540** | 127,26f | 288,91f | ↓ | 55,95 | 259,7k | 100,12k | ↑ | 159,39 | 269,97f | 596,2f | ↓ | 54,72 | 5,62M | 4,05M | ↑ | 38,77 |
| **C6288** | 2,61p | 2,69p | ↓ | 2,97 | 169,81k | 72,36k | ↑ | 134,67 | 4,91p | 5,03p | ↓ | 2,39 | 2,82M | 2,3M | ↑ | 22,61 |
| **S420** | 14,5f | 16,37f | ↓ | 11,42 | 387,54k | 177,53k | ↑ | 118,30 | 42,23f | 46,6f | ↓ | 9,38 | 8,64M | 5,58M | ↑ | 54,84 |
| **S1423** | 112,03f | 141,39f | ↓ | 20,77 | 208,73k | 90,09k | ↑ | 131,69 | 268,7f | 313,94f | ↓ | 14,41 | 4,56M | 2,69M | ↑ | 69,52 |
| **S9234** | 140,48f | 276,68f | ↓ | 49,23 | 479,29k | 184,16k | ↑ | 160,26 | 352,26f | 667,52f | ↓ | 47,23 | 10,13M | 5,85M | ↑ | 73,16 |
| **S38584** | 982,19f | 1,37p | ↓ | 28,31 | 569,61k | 271,89k | ↑ | 109,50 | 1,93p | 2,63p | ↓ | 26,62 | 6,94M | 5,61M | ↑ | 23,71 |

# 5 CONCLUSÃO

Este trabalho apresentou o desenvolvimento de circuitos combinacionais em três bibliotecas de células distintas, classificadas quanto ao tipo de transistor utilizado: *regular*-VT, *high*-VT e *low*-VT. Foram mensurados e comparados os atrasos de propagação de cada célula, comprovando a constatação de que os atrasos dos transistores HVT são muito superiores aos experimentados pelos transistores RVT, e portanto a biblioteca HVT seria muito limitante do desempenho a baixo $V_{DD}$ - ao passo que as células LVT reduzem os atrasos substancialmente. Para a biblioteca utilizando transistores de VT regular, foram desenvolvidos, adicionalmente, registradores. A metodologia de dimensionamento foi baseada em trabalhos prévios e baseia-se em ajustar a largura dos transistores de modo a equalizar os tempos de subida e descida, de modo a maximizar as margens estáticas de ruído e reduzir efeitos de variabilidade em cada $V_{DD}$, que são prejudiciais, principalmente, em baixas tensões de operação. Tais bibliotecas foram projetadas para operar em regime de amplo ajuste de tensão e frequência, desde o ponto de mínimo consumo energético, em condições de inversão moderada, até o ponto de inversão forte. Entretanto, somente a biblioteca com transistores RVT foi caracterizada, uma vez que, de acordo com a metodologia de análise de potência e *timing* utilizada, seria necessário o desenvolvimento de registradores, os quais não foram incluídos nas bibliotecas HVT e LVT.

A biblioteca com transistores RVT foi sintetizada para dez circuitos de teste: filtro digital *notch*, um núcleo compatível com um micro-controlador 8051 e oito circuitos *benchmark* ISCAS, quatro combinacionais e outros quatro sequenciais. Tais circuitos foram submetidos a variações extremas de temperatura, de forma a avaliar os resultados em três condições de operação: sub-VT, *near*-VT, onde situa-se o ponto de mínima energia por operação, e super-VT. Foi demonstrado que operar no MEP resulta em grandes economias de energia: em média, 54,46% quando comparada ao regime de sub-limiar e 99,01% em relação à tensão nominal, para a temperatura de 25°C e processo típico. Adicionalmente, do regime

de sub-VT para o MEP, houve, em média, um ganho de desempenho acima de 1300%. Tais ganhos em frequência e energia são ainda superiores em relação ao regime de sub-limiar, quando as temperaturas de -40°C e 125°C são consideradas. Nas mesmas condições, em relação à tensão nominal, a média de ganhos de energia é um pouco menor, entretanto ainda acima de 96%. Todavia, tanto a redução de temperatura quanto o aumento, impactaram substancialmente na localização do ponto de mínima energia, alcançando uma variação de 190 mV entre os dez circuitos analisados, no caso da temperatura negativa. Desta forma, torna-se claro que o MEP é muito sensível à variações de temperatura, em função da dependência da tensão de limiar com a mesma. Adicionalmente, em -40°C e 125°C, o consumo de energia para cada circuito foi superior quando comparado à temperatura ambiente. Portanto, é recomendável que os circuitos com ênfase em economia de energia, operando em *near*-VT, não trabalhem em ambientes hostis, sob o ponto de vista de variabilidade extrema de temperatura. A dissipação no regime NTV é tão baixa que o aquecimento do circuito pode ser mínimo, e portanto a amplitude de variação da temperatura do ambiente é que determinará as condições de variação do desempenho. Do ponto de vista de desempenho, a operação do MEP até o ponto nominal de tensão, em regime de VFS amplo, propicia frequências que variam de centenas de kHz até a faixa dos MHz/GHz para as temperaturas de -40°C e 25°C, e de MHz até GHz em 125°C.

Adicionalmente foram demonstrados os efeitos da introdução de uma diversidade maior de células combinacionais no fluxo de síntese para os dez circuitos testados. Em todos os casos, o número de *gates* equivalentes e de instâncias de células foi diminuído. Em média, houve uma redução de 28,41% no número de células. Em termos de energia, cinco circuitos beneficiaram-se da redução obtida no número de *gates*. Quanto ao desempenho, oito dos dez circuitos foram favorecidos, com uma média de 9,3% em ganho de *performance*.

Por fim, uma comparação dos resultados de energia e frequência obtidos neste estudo, para os dez circuitos de teste, à temperatura de 25°C e processo típico, foram confrontados com os resultados mensurados por Stangherlin (2013), baseando-se na mesma tecnologia CMOS, nas mesmas metodologias de simulação, caracterização das células e estilo de síntese lógica. A diferença fundamental foi a tensão escolhida para o dimensionamento da biblioteca de células e a introdução de duas células combinacionais. A comparação foi realizada em dois pontos: MEP e 450 mV. Os resultados demonstraram que para oito dos dez circuitos, nos dois pontos de comparação, houve redução de consumo de energia com a biblioteca deste trabalho. Em termos de desempenho, considerando somente os casos em que houve simultaneamente

redução de energia, a média de incremento de desempenho, na tensão de $V_{DD}$ para o MEP, foi acima de 150%.

A primeira contribuição deste trabalho foi a escolha de uma tensão de dimensionamento diferenciada, menor, visando otimizar os circuitos para funcionarem no ponto de mínima energia por operação, baseada nos resultados do trabalho de Stangherlin (2013) e definições de De (2013). Outra contribuição foi a introdução de uma diversidade maior de células combinacionais em relação a trabalho anterior, o que reduziu substancialmente a quantidade de instâncias de células lógicas utilizadas na síntese dos mesmos circuitos de teste, atenuando, na maior parte dos casos testados, a energia consumida e, simultaneamente, incrementando o desempenho do circuito como um todo. As demais contribuições foram:

• Aumentar o comprimento do canal dos transistores empilhados na rede *pull-down* das portas NOR com o intuito de reduzir a largura dos transistores PMOS. Para a implementação com transistores RVT, foi demonstrado que para o padrão de entrada utilizado na metodologia de simulação, não houve prejuízos e sim, benefícios, em termos de redução no atraso da célula;

• Adaptar as equações de dimensionamento para tratar o caso de transistores empilhados nas redes *pull-up* e *pull-down* simultaneamente, em função da introdução das células OAI21 e AOI22;

• Aplicar a metodologia de dimensionamento para transistores com limiares distintos do transistor convencional (RVT);

• Adaptar a equação de dimensionamento para a porta NAND com transistores LVT, adicionando um fator de incremento $DELTA_P$ para os transistores *single* da *rede pull-up*, de modo a respeitar a equalização de tempos de subida e descida;

Trabalhos futuros, podem explorar novas possibilidades. Primordialmente, finalizar as bibliotecas com transistores HVT e LVT, adicionando registradores otimizados de modo a avaliar os benefícios/prejuízos da utilização independente destas bibliotecas em termos de consumo de energia e desempenho. Paralelamente, investigar o MEP destes transistores para os circuitos de teste analisados, de modo a descobrir se a localização ficará abaixo das respectivas tensões de limiar, uma vez que isto ocorreu com os transistores RVT. Adicionalmente, verificar os benefícios da utilização combinada das três bibliotecas no processo de síntese de modo a aumentar o desempenho e/ou reduzir o consumo de acordo com os requisitos de projeto. É possível, também, aplicar as bibliotecas de células

desenvolvidas em outros circuitos de teste, como, por exemplo, somadores, inclusive, combinando técnicas de computação imprecisa de modo a reduzir o consumo energético. Uma outra possibilidade é investigar o uso de outros estilos lógicos CMOS (dinâmicos, diferenciais, em modo corrente, etc) e metodologias de dimensionamento visando a operação em *near*-VT.

# REFERÊNCIAS

ASHRAF, R. A.; ALZAHRANI, A.; DEMARA, R. F. Extending Modular Redundancy to NTV: Costs and Limits of Resiliency at Reduced Supply Voltage. In: WORKSHOP ON NEAR-THRESHOLD COMPUTING, 2014...**Proceedings**. Mineapolis: [s.n.], 2014. p. 1-7.

ASU. **Predictive Technology Model**. Tempe, 2008. Disponível em: <http://ptm.asu.edu/>. Acesso em: Jan. 2015.

BOL, D.; FLANDRE, D.; LEGAT, J.-D. Technology flavor selection and adaptive techniques for timing-constrained 45nm subthreshold circuits. In: INTERNATIONAL SYMPOSIUM ON LOW POWER ELECTRONICS AND DESIGN (ISLPED), 2009...**Proceedings**. [S.l.]: ACM, 2009. p.21-26.

BSIM4. **Berkley Short-channel IGFET Model**. Berkley, 2000. Disponível em: < http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4>. Acesso em: Mar. 2015.

CALIMERA, A. et al. Reducing leakage power by accounting for temperature inversion dependence in dual-Vt synthesized circuits. In: INTERNATIONAL SYMPOSIUM ON LOW POWER ELECTRONICS AND DESIGN (ISLPED), 2008...**Proceedings**. [S.l.]: ACM, 2008. p.217-220.

CHANDRAKASAN, A. P. et al. Technologies for Ultradynamic Voltage Scaling. **Proceedings of the IEEE**, v. 98, n. 2, Feb 2010, p. 191–214.

CHANDRAKASAN, A. P.; SHENG, S.; BRODERSEN, R. W. Low-power CMOS digital design. **IEICE Transactions on Electronics**, v. 75, n. 4, p. 371–382, 1992.

CHANDRAKASAN, A. P.; BRODERSEN, R. W. **Low Power Digital CMOS Design**. Dordrecht: Kluwer Academic Publishers, 1995.

CHANG, L.; HAENSCH, W. Near-threshold operation for power-efficient computing?: it depends.... In: DESIGN AUTOMATION CONFERENCE (DAC), 2012...**Proceedings**. [S.l.]: ACM, 2012. p. 1155-1159.

CHEN, Y. et al. Ultralow power SRAM design in near threshold region using 45nm CMOS technology. In: IEEE INTERNATIONAL CONFERENCE ON ELECTRO/INFORMATION TECHNOLOGY (EIT), 2011...**Proceedings**. [S.l.]: IEEE, 2011. p. 1-4.

CHO, H. K.; MAHLKE, S. Dynamic acceleration of multithreaded program critical paths in near-threshold systems. In: 45TH ANNUAL IEEE/ACM INTERNATIONAL SYMPOSIUM ON MICROARCHITECTURE WORKSHOPS (MICROW), 2012...**Proceedings**. [S.l.]: IEEE, 2012. p.63-67.

DA COSTA, E. A. C. et al. Modeling of short circuit power consumption using timing-only logic cell macromodels. In: 13TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI), 2000... **Proceedings**. [S.l.]: IEEE, 2000. p. 222-227.

DE, V. Near-Threshold voltage design in nanoscale cmos. In: DESIGN, AUTOMATION & TEST IN EUROPE (DATE), 2013...**Proceedings**. [S.l.]: IEEE, 2013. p. 612.

DRESLINSKI, R. G. et al. Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits. **Proceedings of the IEEE**, v. 98, n. 2, Feb 2010, p. 253–266.

HANSEN, M. C.; YALCIN, H.; HAYES, J. P. Unveiling the ISCAS-85 benchmarks: A case study in reverse engineering. **IEEE Design & Test of Computers**, v. 16, n. 3, 1999, p. 72–80.

HARRIS, D; SUTHERLAND, I. Logical Effort of Carry Propagate Adders. In: 37TH ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS, 2003. ..**Proceedings**. [S.l:s.n.], 2003. p. 873-878.

HIJAZ, F.; KHAN, O. Rethinking Last-Level Cache Management for Multicores Operating at Near-Threshold Voltages. In: WORKSHOP ON NEAR-THRESHOLD COMPUTING, 2014...**Proceedings**. Mineapolis: [s.n.], 2014. p. 1-6.

HSU, S. et al. A 280mV-to-1.1 V 256b reconfigurable SIMD vector permutation engine with 2-dimensional shuffle in 22nm CMOS. In: INTERNATIONAL SOLID-STATE CIRCUITS CONFERENCE (ISSCC), 2012...**Proceedings**. [S.l.]: IEEE, 2012. p. 178-180.

HU, J.; YU, X. Near-threshold full adders for ultra low-power applications. SECOND PACIFIC-ASIA CONFERENCE ON CIRCUITS, COMMUNICATIONS AND SYSTEM (PACCS), 2010...**Proceedings**. [S.l.]: IEEE, 2010. p. 300-303.

IBM. Industrial Business Machines: product manual. **CMOS10LPE Bulk**, [S.l.:s.n.], 2009.

KAUL, H. et al. Near-threshold voltage (ntv) design: opportunities and challenges. Design Automation Conference (DAC), 2012...**Proceedings**. [S.l.]: ACM, 2012. p. 1149-1154.

KIM, T. et al. Utilizing Reverse Short-Channel Effect for Optimal Subthreshold Circuit Design. **IEEE Transactions on Very Large Scale Integration (VLSI) Systems**, v. 15, n. 7, p. 821–829, 2007.

KRIMER, E. et al. Synctium: a Near-Threshold Stream Processor for Energy-Constrained Parallel Applications. **IEEE Computer Architecture Letters**, v. 9, n. 1, Jan 2010, p. 21–24.

KUHN, K. J. Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS. In: INTERNATIONAL ELECTRON DEVICES MEETING (IEDM), 2007...**Proceedings**. [S.l.]: IEEE, 2007. p. 471-474.

LUO, T.; NEWMARK, D.; PAN, D. Z. Total power optimization combining placement, sizing and multi-Vt through slack distribution management. In: ASIA AND SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE (ASPDAC), 2008...**Proceedings**. [S.l.]: IEEE, 2008. p. 352-357.

MARKOVIC, D. et al. Ultralow-Power Design in Near-Threshold Region. **Proceedings of the IEEE**, v. 98, n. 2, Feb 2010, p. 237–252.

MOORE, G. E. Cramming more components onto integrated circuits. **Electronics**, v. 38, n. 8, Apr 1965, p. 144.

RABAEY, J; CHANDRAKASAN, A; NIKOLIC, B. **Digital Integrated Circuits: a design perspective**. 2ª. ed. [S.l.]: Prentice Hall, 2003.

SEO, S. et al. Diet SODA: a power-efficient processor for digital cameras. In: 16TH INTERNATIONAL SYMPOSIUM ON LOW POWER ELECTRONICS AND DESIGN (ISLPED), 2010...**Proceedings**. [S.l.]: ACM, 2010. p. 79-84.

SEO, S. et al. Process variation in near-threshold wide SIMD architectures. In: 49TH ANNUAL DESIGN AUTOMATION CONFERENCE (DAC), 2012...**Proceedings**. [S.l.]: ACM, 2012. p. 980-987.

SHAFIQUE, M. et al. The EDA Challenges in the Dark Silicon Era: Temperature, Reliability, and Variability Perspectives. In: DESIGN AUTOMATION CONFERENCE (DAC), 2012...**Proceedings**. [S.l.]: ACM, 2014. p. 1-6.

SIL, A. et al. A novel high write speed, low power, read-SNM-free 6T SRAM cell. In: 51ST MIDWEST SYMPOSIUM ON CIRCUITS AND SYSTEMS (MWSCAS), 2008...**Proceedings**. [S.l.]: IEEE, 2008. p. 771-774.

SOARES, L. et al. 61 pJ/sample near-threshold notch filter with pole-radius variation. In: IEEE FOURTH LATIN AMERICAN SYMPOSIUM ON CIRCUITS AND SYSTEMS (LASCAS), 2013....**Proceedings**. [S.l.]: IEEE, 2013. p. 1-4.

STANGHERLIN, K. H. **Energy and Speed Exploration in Digital CMOS Circuits in the Near-threshold Regime for Very-Wide Voltage-Frequency Scaling**. 2013. 68 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

STANGHERLIN, K. H.; BAMPI, S. Energy-speed exploration for very-wide range of dynamic VF scaling. In: 26TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI), 2013...**Proceedings**. [S.l.]: IEEE, 2013.

SWANSON, R.; MEINDL, J. Ion-implanted complementary MOS transistors in low-voltage circuits. **IEEE Journal of Solid-State Circuits**, v. 7. n. 2, p. 146-153, 1972.

VEENDRICK, H. J. Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits. **IEEE Journal of Solid-State Circuits**, v. 19, n. 4, p. 468–473, 1984.

VIRGA, A. et al. Performance and Variation Robustness of Near-Threshold Differential Cascode Voltage Switch Logic. In: WORKSHOP ON NEAR-THRESHOLD COMPUTING 2014...**Proceedings**. Mineapolis: [s.n.], 2014. p. 1-6.

WANG, A.; CHANDRAKASAN, A. A 180-mV subthreshold FFT processor using a minimum energy design methodology. **IEEE Journal of Solid-State Circuits**, v. 40, n. 1, p. 310–319, Jan 2005.

WANG, C. et al. Near-Threshold Energy- and Area-Efficient Reconfigurable DWPT/DWT Processor for Healthcare-Monitoring Applications. **IEEE Transactions on Circuits and Systems II: Express Briefs**, v. 62, n. 1, p. 70–74, Jan 2015.

WANG, C.-C.; LEE, C.-L.; LIN, W.-J. A 4-kb Low-Power SRAM Design With Negative Word-Line Scheme. **IEEE Transactions on Circuits and Systems I: Regular Papers**, v. 54, n. 5, p. 1069–1076, Mai 2007.

YE, Y. et al. Statistical Modeling and Simulation of Threshold Variation Under Random Dopant Fluctuations and Line-Edge Roughness. **IEEE Transactions on Very Large Scale Integration (VLSI) Systems**, v. 19, n. 6, p. 987–996, Jun 2011.

ZHAI, B. et al. A 2.60 pJ/Inst subthreshold sensor processor for optimal energy efficiency. In: SYMPOSIUM ON VLSI CIRCUITS (VLSIC), 2006...**Proceedings**. [S.l.]: IEEE, 2006. p. 154-155.

ZHAO, B. et al. An energy efficient fully integrated OOK transceiver SoC for wireless body area networks. In: IEEE ASIAN SOLID-STATE CIRCUITS CONFERENCE (A-SSCC), 2013...**Proceedings**. [S.l.]: IEEE, 2013. p. 441-444.