Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Caracterização de Memórias Analógicas implementadas com transistores MOS *Floating Gate*

Autor: André Luis do Couto

Orientador : Prof. Dr. Carlos Alberto dos Reis Filho

Tese de Mestrado apresentada à faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.
Área de Concentração: **Microeletrônica**

Banca Examinadora

Carlos Alberto dos Reis Filho, Prof. Dr. ..... DSIF/FEEC/UNICAMP
José Alexandre Diniz, Prof. Dr. ..... DSIF/FEEC/UNICAMP
Furio Damiani, Prof. Dr. ..... DSIF/FEEC/UNICAMP
Fernando Chavez Porras, Dr. ..... FREESCALE SEMICONDUTORES DO BRASIL

Campinas, SP
Novembro/2005

Este exemplar corresponde à redação final da tese defendida por: ANDRE LUIS DO COUTO
e aprovada pela Comissão
Julgada em 28/11/05
Orientador

UNIDADE BC
Nº CHAMADA
T | UNICAMP
C 837c
V EX
TOMBO BC/ 71017
PROC. 16.123-06
C D
PREÇO 11,00
DATA 20/12/06
ID: 395073

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

C837c

Couto, André Luis do
Caracterização de memórias analógicas implementadas com transistores MOS *Floating Gate* / André Luis do Couto. --Campinas, SP: [s.n.], 2005.

Orientador: Carlos Alberto dos Reis Filho
Dissertação (Mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Memória cache. 2. Sistemas de memória de computadores. 3. Transistores de efeito de campo. 4. Transistores. 5. Circuitos integrados. 6. Microeletrônica. I. Reis Filho, Carlos Alberto dos. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Analog memories characterization implemented with floating gate MOS transistors
Palavras-chave em Inglês: Microelectronic, Integrated circuit, Transistor, Field effect transistor, Computer memory system
Área de concentração: Microeletrônica e Optoeletrônica
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: José Alexandre Diniz, Furio Damiani, Fernando Chavez Porras
Data da defesa: 28/11/2005

Resumo

A integração de memórias e circuitos analógicos em um mesmo *die* oferece diversas vantagens: redução de espaço nas placas, maior confiabilidade, menor custo. Para tanto, prescindir-se de tecnologia específica à confecção de memórias e utilizar-se somente de tecnologia CMOS convencional é requisito para tal integração. Essa pode ser tanto mais eficiente quanto maior a capacidade de armazenagem de dados, ou seja, maior a densidade de informação. Para isso, memórias analógicas mostram-se bem mais adequadas, posto que em uma só célula (um ou dois transistores) podem ser armazenados dados que precisariam de diversas células de memórias digitais e, portanto, de maior área. Neste trabalho, transistores MOS com porta flutuante mostraram-se viáveis de serem confeccionados e resultados de caracterização como tipos de programação, retenção de dados e *endurance* foram obtidos. O trabalho apresenta as principais características dos FGMOS (*Floating Gate MOS*) e presta-se como referência à futuros trabalhos na área.

2006351821

BIBLIOTECA CENTRAL
CÉSAR LATTES
DESENVOLVIMENTO DE
COLEÇÃO
UNICAMP

Abstract

Monolithic integration of memories and analog circuits in the same die offers interesting advantages like: smaller application boards, higher robustness and mainly lower costs. Today, a profitable integration of these kind of circuit can only be possible using conventional CMOS technology, which allows efficiently extraordinary levels of integration. Thus, the possibility of integrating analog memories looks more suitable since one single cell (usually use one or two transistors) serves for storing the same data stored by few digital memory cells, therefore, they requiring less area. In this work, it was implemented different memory cells together with few devices using floating gate MOS transistors and manufactured by a conventional CMOS technology. Different sort of programming, data retention, and endurance were characterized as well as the main characteristics of the FGMOS (*Floating Gate MOS*) were obtained. The results of their characterization reveal that is possible to make and to program *floating gate* MOSFETS analog memories and must serve as starting-point and reference for new academic studies.

BIBLIOTECA CENTRAL
CÉSAR LATTES
DESENVOLVIMENTO DE
COLEÇÃO
UNICAMP

A minha família, a minha esposa e ao meu filho.

Agradecimentos

Primeiramente, pela tese concluída e por cada dia de minha vida, agradeço a Deus.

Agradeço a oportunidade concedida por meu Orientador, Prof. Carlos Reis, de ter podido realizar meu trabalho em sua equipe e sob sua orientação.

Não posso deixar de agradecer aos meus amigos que me ajudaram na realização deste trabalho, cuja amizade é sem dúvida muito mais importante do que qualquer resultado obtido na tese, sobretudo a Leandro Ferrari, o qual dispensou grande parte de seu tempo a me ajudar no *set-up* de caracterização das estruturas. Agradeço também a Fernando Castaldo, (o mano) pelas discussões filosóficas, científicas e pelos resultados de dispositivos que realizamos em conjunto.

Agradeço a minha família, minha querida esposa Kamilla e meu filho querido, André, que sempre tripulam meus pensamentos, pela confiança em mim.

À confecção de *layouts* e discussões sobre o tema, agradeço a Saulo Finco do Cenpra, sempre solícito. Agradecimentos também ao Instituto de Pesquisas Eldorado pelo apoio financeiro dispensado durante período parcial de realização de meu mestrado.

Por fim, agradeço a Banca Examinadora pela atenção e interesse dispensados ao meu trabalho.

"A arrogância pode se transformar
em uma arma tão letal para nossa
capacidade de gerar mudanças produtivas
quanto a carência de confiança"

"A luta contra o erro
tipográfico tem algo de
homérico. Durante a revisão
os erros se escondem, fazem-
-se positivamente invisíveis.
Mas assim que o livro sai,
tornam-se visibilíssimos..."
(Monteiro Lobato)

Sumário

LISTA DE FIGURAS

LISTA DE TABELAS

Lista de Símbolos

AMS	Austria Mikro Systeme
BIOS	Memória de parâmetros básicos de entrada e saída de um computador.
C_B	Capacitância entre o *control-gate* e o *floating-gate* para programação unipolar
C_{fg}	Capacitância entre o *floating-gate* e o substrato.
C_G ou C_g	Capacitância entre o *floating-gate* e o substrato.
CHE	Carregamento/programação por elétrons quentes
C_{inj} ou Ci	Capacitor de tunelamento
CMOS	Tecnologia MOS complementar
CTAT	Valor (de tensão ou corrente) complementar ao de Temperatura Absoluta.
DRC	*Design Rule Check*
dV_t	Delta da tensão de limiar
E_1 e E_2	Campo elétrico através de dielétricos 1 e 2 ($V.cm^{-1}$)
E_{max}	Intensidade de campo elétrica aplicada ao *control gate*
EPROM	ROM programável eletricamente
eV	Elétron-Volt
FAMOS	*Floating-gate Avalanche-injection MOS*
FG	*Floating gate*
FGMOS	Transistor MOS *floating-gate*
FN	Fowler-Nordheim
h	Constante de Plank
J ou j	Densidade de corrente
j_1	Densidade de corrente através do óxido entre as camadas de polissilício-1 e 2.
j_2	Densidade de corrente através do óxido entre o *floating-gate* e o substrato
Ki, Kb e K0	Acoplamentos capacitivos entre *control gate* e *floating gate.*
LPM	Laboratório de Pesquisas Magneti-Marelli
LDD	Low doped drain transistor
m	Massa efetiva do elétron no vácuo
$m_1, m_2,...,m_n$	Acoplamentos capacitivos entre *control gate* e *floating gate.*
MIFG	Dispositivo floating-gate de múltiplas entradas
MOS	Metal Oxido Semicondutor
MOSFET	Transistor de Efeito de Campo tipo MOS
m_{ox}	Massa efetiva do elétron no dielétrico (SiO_2)
nFGMOS	Floating-gate Mos Transistor tipo N
NMOS	Transistor MOS canal N
OTP	Memória programável uma única vez.
PMOS	Transistor MOS canal P
PWM	Modulação por largura de pulso
ROM	Memória não volátil apenas de leitura (programável uma única vez)
Q	Carga elétrica
q	Carga do elétron

Q_{fg} ou Q_F	Carga no *floating gate* (Coulomb.cm^{-1})
SAMOS	Tipo de memória não-volátil programável por avalanche/injeção eletrônica.
Si/SiO_2	Óxido de silício (tipo de dielétrico)
t	Tempo
UV	Radiação ultravioleta
V1	Tensão Aplicada ao *control gate* 1
V2	Tensão aplicada ao *control gate* 2
V_{C1}	Tensão sobre o *control gate* 1
V_{C2}	Tensão sobre o *control gate* 2
VDD	Tensão de alimentação positiva
V_{DS}	Tensão dreno-fonte
V_g	Tensão aplicada ao *control gate*
V_{GB} e V_{GA}	Tensões aplicadas aos *control gates* A e B respectivamente.
V_{GS}	Tensão *gate-source*
V_{lect}	Tensão aplicada ao gate do transistor de leitura para obtenção do dado.
VLSI	Integração em larga escala
V_{out}	Tensão de saída
V_{pp}	Valor de pulso de tensão
VSS	Tensão de alimentação negativa.
V_t ou V_T	Tensão de limiar
Vt1	Tensão de Limiar do dispositivo "virgem" (vindo da *foundry*).
Vt2	Tensão de Limiar do dispositivo após pulso de programação.
Vt3	Tensão de Limiar do dispositivo após tratamento térmico para estudo de retenção.
V_{target}	Tensão a ser programada na memória
V_{TH}	Tensão de limiar efetiva
ΔV_T	delta da tensão de limiar
ε_1 e ε_2	Constantes dielétricas das camadas de óxido de silício (F.cm^{-1})
ϕ_b	Barreira de potencial entre o dielétrico e o silício
σ	Condutância elétrica

Glossário Inglês-Português

Chip	Circuito integrado
Control gate	Terminal a acoplado capacitivamente ao *floating gate*
CTAT	*Complementary To the Absolute Temperature* Complementar ao valor de temperatura absoluta
Dummy Capacitors	Capacitores com função de melhorar casamento entre dispositivos.
Design Rule Check	Programa que verifica a concordância do layout com as regras ditadas pela *foundry*.
Endurace	Durabilidade
Erase ou ***erasing***	Operação de apagamento da memória
Floating gate	Diz-se do terminal porta quando não está conectado fisicamente a nenhum outro terminal condutor.
Foundry	Indústria que produz circuitos integrados
Gate	Terminal porta do transistor MOSFET
Hot-electron Injection	Injeção de elétrons quentes
Low doped drain transistor	Transistor com dreno levemente dopado.
Memory Window Collapse	Colapso da janela de memória – Não é mais possível a programação (escrita/apagamento) da memória.
Multiple input	Múltiplas entradas
Pinch-off	Região de estrangulamento do canal próximo ao dreno
Proto-board	Placa para montagem de protótipos de circuitos
Push-buttons	Botões que funcionam como chaves: fecham o circuito somente enquanto pressionados pelo o usuário.
Offset voltage	Desvio aleatório do nível de tensão
Overerased	MemóriaFGMOS tipo enriquecimento que se comporta como do tipo depleção por haver sido apagado em excesso.
Read	Leitura de dado (tensão de limiar) de memória.
Select gate	O mesmo que *control gate*
Self-limiting	Auto-limitante

Threshold Voltage	Tensão de limiar do transistor MOS.
Traps-up	Armadilhas (defeitos) no dielétrico capazes de reter carga
Trimming	Ajuste
Write ou ***writing***	Escrita na memória

Trabalho Publicado pelo Autor

1. André Luis do Couto, João Paulo C. Cajueiro, Carlos A. dos Reis Filho "*Temperature-Compensated Voltage Using Floating-Gate MOS Transistor*", IMAPS-Brazil Proceedings, 6-8 August, 2003. Campinas, Brazil.

Capítulo 1

Estruturas de Memória e Dispositivos Floating Gate

1.1.Introdução

1.1.1 Breve histórico

Data de 1967 o primeiro artigo, "*A floating-gate and its application to memory devices*" [1], tratando de estruturas com portas flutuantes como um mecanismo para armazenamento não-volátil de dados: uma memória bi-estável foi obtida ao se conseguir armazenar carga no *floating gate*.

Em 1971, o primeiro produto comercial foi anunciado e tornou-se conhecido como EPROM [20]. Usava um transistor FAMOS (*Floating-gate Avalanche-injection MOS*). Desde então os dispositivos *floating-gate* têm sido largamente utilizados em sistemas digitais. A memórias *Flash*-EPROM, presentes nos computadores atuais, que armazenam parâmetros básicos de entrada e saída do sistema (*BIOS*), armazenam os dados de maneira não volátil utilizando-se, para tanto, de estruturas *floating-gate*.

Nos anos 80, com a pesquisa crescente em redes neurais, houve necessidade de estruturas para armazenamento analógico não-volátil e, obviamente, as estruturas *floating-gate* figuraram como fortes candidatas para ocupação de parte deste novo nicho de pesquisas. Isto porque a natureza do dado armazenado no *floating-gate* é, fundamentalmente, analógica. Contudo, o controle preciso de tal armazenamento mostrava-se bastante complicado.

Em 1989, Richard Carley publicou artigo intitulado "*Trimming Analog Circuits Using Floating Gate Analog MOS Memory*" [22]. Este foi um dos primeiros artigos em que se relatava a utilização de tecnologia CMOS para a confecção de memórias *floating gate*. Trata-se de uma das principais referências neste tema.

Em 1994, Katsuhiko Ohsaki et al, *"A Single Poly EEPROM Cell Structure for Use in Standard CMOS Processes"* [16], propuseram estruturas FGMOS com tecnologia CMOS com sem dupla camada de polisssilício. Neste trabalho, os *gates* de controle não eram camadas de polisssilício, mas poços N.

Em 1999, Abouchi et al, publicaram o artigo *"Analog EEPROM in standard process AMS 0.8 μm CMOS"*, do qual muitas estruturas deste trabalho de tese foram baseadas.

Em 1992, Shibata e Ohmi [2] observaram que em componentes *floating gate* contendo múltiplos *gates* de controle, a corrente de canal era controlada pela soma ponderada das tensões aplicadas em cada um dos *gates*, sendo os pesos desta somatória inversamente proporcionais às capacitâncias de cada *gate* . Baseados na similaridade destes dispositivos com células do sistema nervoso, Shibata e Ohmi denominaram tais dispositivos **neuron MOS (neuMOS** ou **υMOS)**.

Yang e Andreou [3] referem-se a tais dispositivos como **FGMOS**. Ramírez-Ângulo [8] preferiu outra denominação: ***multiple-input floating gate transistors (MIFG)***.

1.1.2. Motivação

A gama potencial de aplicações de dispositivos e memórias analógicas *floating gate* é bem ampla: redes neurais [2], circuitos adaptativos em que informação analógica de alta densidade é requerida, sistemas VLSI, compensação de envelhecimento de sensores, filtragem [26], multiplicadores analógicos [27], sistemas de computação analógica, de controle automático de ganho e *offset* [25].

O estudo de confecção de estruturas de memória analógicas *floating gate*, que é o foco deste trabalho, visa a compatibilização de tais estruturas à tecnologia CMOS, sem prescindir das qualidades de não-volatilidade dos dados e da versatilidade conferida pela programação elétrica. As vantagens inerentes à utilização da tecnologia CMOS são seu custo reduzido comparado à tecnologias específicas de memória (tais como, ROM, EPROM, EEPROM, SAMOS e OTP), à possibilidade de integração monolítica de memórias a circuitos e sistemas e, por conseguinte, da alta densidade de integração [9].

Os processos de programação em tecnologias específicas de memória são otimizados através de meios como camadas de polissilício microtexturizado, de camadas

ultrafinas de dielétrico (óxido de silício) e de transistores não auto-alinhados [22]. Contudo, mesmo à custa de menor eficiência, os processos atuais CMOS demonstram-se adequados a confecção de memórias não-voláteis. A razão disso é a combinação das finas camadas de óxido combinadas com a obtenção possível (através de *charge-pumps*) de tensões relativamente altas necessárias aos processos de tunelamento e razoáveis tempos de retenção.

1.1.3.Memórias Analógicas e Memórias Digitais

As vantagens de memórias analógicas sobre memórias digitais incluem menor área, menor consumo de potência, maior faixa dinâmica e a compatibilidade com processos CMOS. Em contrapartida, como desvantagens, requer maior tempo de programação, altas tensões para programação e menor confiabilidade. A utilização de memórias analógicas ainda é restrita. Uma das razões para isto é a simplicidade do projeto de memórias digitais [22][26].

As memórias digitais podem tolerar perdas expressivas de carga antes que a detecção do bit armazenado torne-se ambígua ao circuitos externos conectados (alta faixa de ruído). Assim, os pulsos de tensão para programação, aos *control gates* de memórias digitais, podem ser altos e sem controle muito preciso da quantidade de cargas injetada. A programação de memórias analógicas, todavia, deve ser bem precisa e controlada, de forma a manter-se uma alta faixa dinâmica (grande precisão).

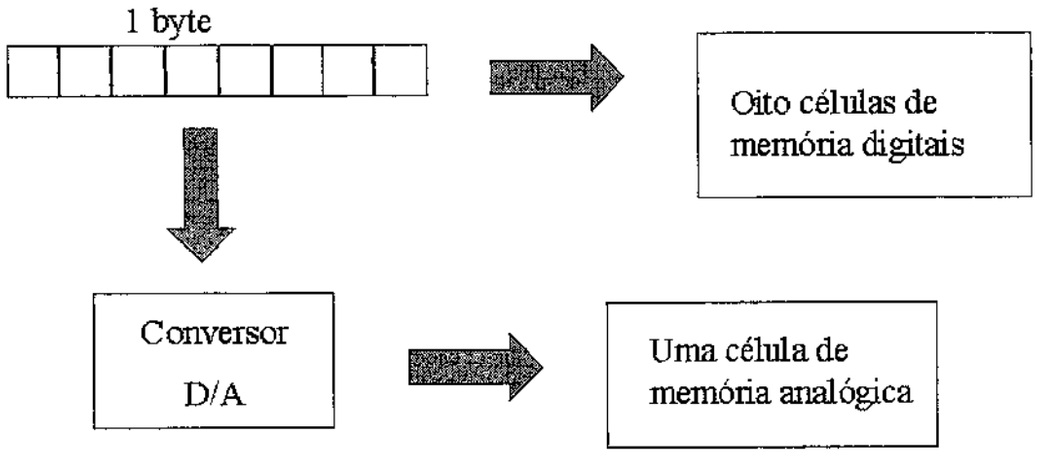

Fig.1.1)Esquema de Memória Analógica x Digital (Economia em área de Si)

Em geral, para cada bit é necessária uma célula de memória digital, fig.1.1, (1 ou 2 transistores), conforme esquema acima. A memória analógica constitui-se numa forma de compactar dados e área no circuito integrado. A utilização progressiva de memórias analógicas tornar-se uma maneira eficiente de aumentar substancialmente a capacidade e a densidade de armazenamento de dados. Além disso, as memórias analógicas podem simplificar sistemas de processamento de sinal implementados com dispositivos periféricos também analógicos [9].

1.2.Organização da dissertação

O capítulo 1, apresenta uma breve retrospectiva histórica, a motivação da tese e alguns aspectos básicos de dispositivos *floating gates*. O capítulo 2 a seguir apresenta a teoria básica para a compreensão das estruturas com dispositivos FGMOS, tais como: modelo de bandas de energia, tunelamento e equações essenciais, os cuidados na confecção dos *layouts* de tais estruturas e o seu modelamento básico. Além disso, as principais figuras de mérito e características de dispositivos de memória *floating gate* são destacadas no capítulo 3, dentre elas, tem-se: tempo de retenção, *endurance* e tempo de programação. O capítulo 4 apresenta a utilização de FGMOS como dispositivo multi-*gates*, caso em que não são usados como dispositivos de memória e uma aplicação prática: compensação de tensão em temperatura com dispositivo FGMOS. Já o capítulo 5 apresenta os resultados práticos de medidas e programação realizadas nas estruturas de memória e compensação em temperatura. Por fim, a conclusão ressalta que importância o estudo em memórias com dispositivos FGMOS em tecnologia digital CMOS apresenta.

Capítulo 2

Princípio de Operação e Métodos de Programação de Memórias *Floating Gate*

O princípio básico de operação de uma memória FGMOS consiste no aprisionamento de cargas em torno de um *gate* de polissilício completamente envolto por SiO_2. Tais cargas são armazenadas permanentemente, pois o dielétrico que envolve o polissilício, geralmente óxido de silício, por ser de boa qualidade, mantém as cargas na região do *gate*. A carga Q_{fg} armazenada no *floating gate* altera a tensão efetiva de limiar (*threshold voltage*) necessária ao acionamento do transistor de uma quantidade $dV_t=Q_{fg}/C_{fg}$, como mostrado na fig.2.2. Entre o gate de controle (*select gate* ou *control gate*) e o *floating gate* há SiO_2 como dielétrico [9]. Aplicando-se um pulso de tensão positiva ao *control gate*, os elétrons são eletrostaticamente atraídos ao *floating gate*, aumentando a quantidade total de carga. Após a supressão da tensão aplicada ao *control gate*, as cargas permanecem aprisionadas sob o dielétrico (SiO_2) [22].

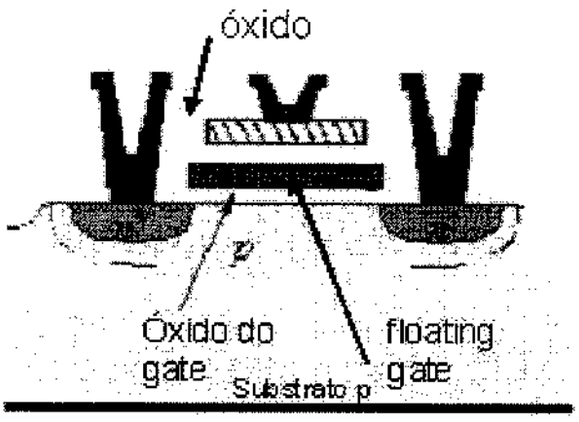

Fig.2.1)Seção Transversal

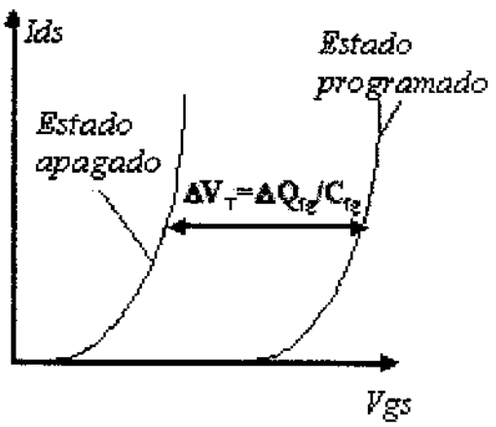

Fig.2.2)Operação de programação

O *floating gate* é uma camada de polisilício que não possui contato com nenhuma outra camada condutora, fig.2.1. Em termos de circuitos, um gate está flutuando (*floating*)

quando não possui caminho para um potencial fixo. A inexistência de tal caminho DC implica que há somente acoplamentos capacitivos a este *gate* [1].

2.1.Modos de Operação de Memórias a Transistores Floating Gates

Há quatro modos de operação com memórias *Floating Gate*:

1- Modo de Leitura: Neste modo, o dispositivo atua como um transistor MOS convencional, sendo sua corrente de dreno modulada pela tensão aplicada ao *gate* de entrada (neste caso, pela soma ponderada dos valores de tensão aplicados aos *control gates,* caso haja mais de um) e pela carga aprisionada na estrutura. Por ter sua tensão de limiar convenientemente ajustável, o FGMOS tanto pode atuar como dispositivo de enriquecimento, como de depleção (dependendo da quantidade de cargas armazenada) [30].

2- Modo de Escrita: Neste modo, um pulso de tensão aplicado ao *control gate*, com amplitude suficiente provoca a condução através de um ou dois dielétricos, SiO_2. O *floating gate,* imerso entre os dielétricos, carrega-se. Com este procedimento pode-se incrementar o valor da tensão de limiar para valores desejados[30].

3- Modo de Apagamento A fim de decrementar o valor da tensão de limiar procede-se ao modo de apagamento. Um pulso de polaridade oposta aplicada ao *control gate* atua removendo as cargas do *floating gate*, reduzindo o seu potencial e conseqüentemente a tensão de limiar. Um método alternativo para remoção de cargas do *floating gate* é através de estruturas com capacitores de *bootstrap* e capacitores injetores de carga, que atuam também removendo cargas do *floating gate* [26].

4- **Modo de Armazenamento:** Neste caso, sem alimentação, o dado (tensão analógica armazenada) é armazenado. Por se tratar de dielétricos bem espessos (16 – 40 nm), as correntes de fuga através deles tendem a ser bem pequenas, permitindo grandes tempos de retenção. Esse aspecto caracteriza as memórias baseadas em FGMOS como não-voláteis.

2.2.Mecanismos Físicos de Programação de memórias FGMOS

Em tecnologia CMOS três mecanismos físicos propiciam a injeção ou a extração de carga elétrica do *floating gate*:

1-Tunelamento Fowler-Nordheim, FN.
2-Injeção de elétrons quentes (*Charging by Hot-electron*), CHE.
3-Radiação por luz ultravioleta, UV.

2.2.1.Tunelamento Fowler-Nordheim

Uma célula de memória *floating gate* consiste em um *gate* que é completamente imerso num dielétrico sobre a região ativa de um transistor CMOS. Pelo fato deste *gate* não ter conexão elétrica direta à qualquer outro condutor, ele é freqüentemente denominado *floating gate*. O potencial elétrico no *floating gate* pode ser modulado por condutores adjacentemente dispostos de forma a haver acoplamento capacitivo entre eles (*control gates*). A corrente resultante através do canal do transistor será função da tensão induzida no *floating gate*. Esta tensão pode ser alterada mudando-se a quantidade de cargas no *floating gate*, operação esta denominada *escrita* [30].

BIBLIOTECA CENTRAL
CÉSAR LATTES
DESENVOLVIMENTO DE
COLEÇÃO
UNICAMP

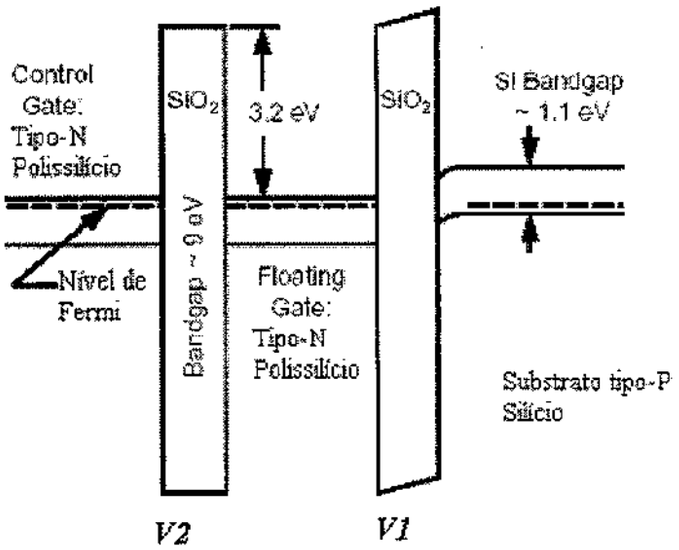

Fig.2.3) Diagrama de bandas de energia de nFGMOS. A esquerda tem-se o diagrama de bandas do *control gate*. Na região central o diagrama do *floating gate* e a direita o substrato de Si. Entre cada uma destas regiões interpõem-se camadas do dielétrico SiO_2 [30].

A fig.2.3 apresenta o diagrama de bandas de uma estrutura nFGMOS. A grande barreira de potencial entre as bandas de condução do Si e do SiO_2 impede, sob condições normais, a passagem de elétrons do Si através do SiO_2 [30].

Com a aplicação de uma tensão, V_g, ao *control gate* (referida ao substrato) do dispositivo, um campo elétrico é estabelecido em cada um dos dois dielétricos. Pela Equação de Gauss, obtém-se: [30]

$$V_g = V_1 + V_2 = d_1 E_1 + d_2 E_2 \quad (2.1)$$

Onde: Vg = tensão aplicada ao *control gate*,

V1 = tensão através do dielétrico entre o *floating gate* e o substrato (região ativa)

V2 = tensão entre o *floating gate* e o *control gate*

E_1 e E_2 são, respectivamente, os campos elétricos aplicados aos óxidos (dielétrico) de espessura d_1 (entre o *floating gate* o substrato) e d_2 (entre o *control gate* e o *floating gate*).

Durante a aplicação de V_g, a carga no *floating gate*, Q_{fg}, varia, dado que as correntes pelos dielétricos entre o *control gate* e o *floating gate* e o *floating gate* e o substrato não

são iguais. A partir da lei pontual de Ohm, j = σE, em que j é a densidade de corrente, σ é a condutividade do material e E é campo elétrico aplicado, tem-se:

$$j_1 = \sigma_1 E_1$$

$$j_2 = \sigma_2 E_2$$

$$\Rightarrow \quad \frac{dQ_{fg}}{dt} = j_1 - j_2$$

Em que j_1 e j_2 são as densidades de corrente através dos óxidos entre as camadas de polissilício e entre o *floating gate* e o substrato, respectivamente[1].

Quando Q_{fg} muda, a tensão de limiar, V_T, do transistor MOS, também muda pela quantidade:

$$\Delta V_T = -\frac{d_2}{\varepsilon_2}\Delta Q_{fg} = -\frac{1}{C_2}\Delta Q_{fg} \quad (2.2)$$

A relação acima é fundamental ao entendimento das estruturas de memória *floating gate*. Ela significa que a tensão de limiar, V_T, pode ser ajustada através da inserção de cargas na região de *floating gate*, ou seja, um ΔV_T pode ser convenientemente somado ou subtraído. O controle da quantidade de cargas injetada no *floating gate* está diretamente relacionada a possibilidade de se programar um valor de tensão analógico. Quanto menos preciso este controle, menos preciso é o valor analógico armazenado. Em tecnologia CMOS AMS com espessura de óxido de até 40nm, a tensão mínima para que se inicie o tunelamento é de 11V (obtido experimentalmente).

O tunelamento Fowler-Nordheim é um fenômeno quântico que permite a elétrons atravessarem barreiras de potencial em dielétricos [1]. Constitui-se em um tipo de ruptura *(breakdown)* não destrutivo [31] e é descrito pela equação (2.3).

$$j = aE^2 \exp\left(-\frac{b}{E}\right) \quad (2.3)$$

[1] j_1 ou j_2 podem ser iguais a zero, mas não simultaneamente. j_1 ou j_2 terão valor nulo se a mínima tensão necessária ao estabelecimento do efeito de tunelamento não for estabelecida sobre o óxido relacionado às densidades de correntes. Sobre o óxido deve haver uma queda de tensão mínima de 11V.

Em que a e b são constantes dependentes do material e da geometria da região de tunelamento e são obtidos experimentalmente. Teoricamente, a e b têm as seguintes expressões (equações 2.4 e 2.5) [31]:

$$a = \frac{q^3 m}{8\pi h m_{ox} \phi_b}, b = \frac{8\pi\sqrt{2m_{ox}}\phi^{\frac{3}{2}}}{3hq} \quad (2.4 \text{ e } 2.5)$$

onde : q é a carga do elétron, h é a constante de Planck, ϕ_b é a barreira de potencial entre o dielétrico e o silício, m é a massa efetiva do elétron no vácuo, m_{ox} é a massa efetiva do elétron no óxido (SiO_2)

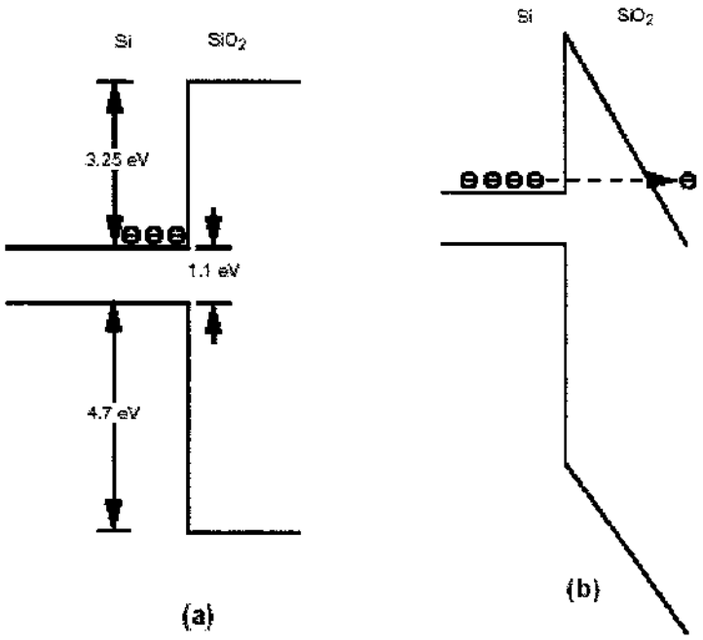

Fig.2.4) Ilustração do processo de tunelamento FN através do óxido [30]

Através da fig.2.4 pode-se observar que : há uma barreira de potencial de aproximadamente 3.2 eV que impede a entrada de elétrons provenientes do Si (substrato) ou do polissilício-1 (*gate*) de penetrarem no óxido de silício, SiO_2, fig2.4.a. Em temperatura ambiente, os elétrons têm energia cinética suficiente que lhes confere a probabilidade de tunelamento de aproximadamente 5 nm SiO_2 adentro. Se o potencial dentro destes 5 nm tem valor inferior a 3.2V (definindo-se o potencial do Si a 0V), então os elétrons que penetram no SiO_2 retornam ao Si não havendo portanto fluxo de carga. Entretanto, cf. fig.2.4.b, se o campo elétrico no SiO_2 for forte o suficiente (>3.2V/5nm = 6.4 $.10^8$ V/m), os elétrons que tunelarem os 5nm no dielétrico serão impulsionados pelo campo constituindo, assim, uma

corrente elétrica. O aumento da intensidade de campo elétrico diminui a distância que os elétrons têm de tunelar, aumentado ainda mais a corrente estabelecida. A distância diminui porque há um estreitamento da barreira de potencial imposta pelo dielétrico, fig.2.4.b [22].

2.2.1.1.Programação de Memórias Floating Gate através de Tunelamento FN

Dispositivos que se utilizam de tunelamento Fowler-Nordheim para injeção e extração de carga (elétrons) têm a vantagem de, apesar da alta tensão necessária, precisar de pouquíssima corrente (pico a nano-ampéres). Além disso, é relativamente fácil a confecção de *charge-pumps*, para obtenção das altas tensões necessárias ao tunelamento, em tecnologia CMOS [25][26].

O processo de programação consiste na introdução e extração de cargas na região do *gate* flutuante. Disto resulta a alteração do valor efetivo da tensão de limiar da estrutura do transistor. Os terminais da estrutura devem ser polarizados de forma diversa à polarização utilizada durante a etapa de leitura dos dados armazenados. Pulsos de tensão devem ser convenientemente aplicados de forma a promover o tunelamento e injeção de cargas. A amplitude de pulsos de tensão aplicados deve ser superior a da tensão de limiar de tunelamento, abaixo da qual nenhuma injeção de portadores ocorre. Além disso, a duração destes pulsos é função da espessura do óxido, de sua constante dielétrica, de sua composição, de *traps-up* (armadilhas) e contaminantes no óxido e dos valores inicial e final armazenados para se atingir um dado valor de tensão programado [22].

A corrente no MOSFET aumenta logaritmicamente com o número de pulsos de tensão aplicados ao *control gate*. A estrutura de camadas é apresentada na fig.2.5. A variação é grande para os primeiros pulsos, porém rapidamente decresce para os demais pulsos. Isto ocorre porque as cargas injetadas mudam o potencial do *gate* causando um decréscimo na tensão de tunelamento FN (Fowler-Nordheim). Quando o decréscimo na tensão efetiva no *gate* é tal que não seja mais suficiente para que ocorra o tunelamento, a injeção de cargas cessa e a corrente no MOSFET não mais se altera [1] [22]. Nestas condições:

$$\varepsilon_1 E_1 = \varepsilon_2 E_2 + Q_{fg} \qquad (3.1)$$

Onde : ε_1 e ε_2 são as constantes dielétricas das camadas de óxido de silício ($F.cm^{-1}$)

E_1 e E_2 são campos elétricos nas camadas de óxido ($V.cm^{-1}$)

Q_{fg} é a densidade de carga no *floating gate* (Coulomb.cm^{1})

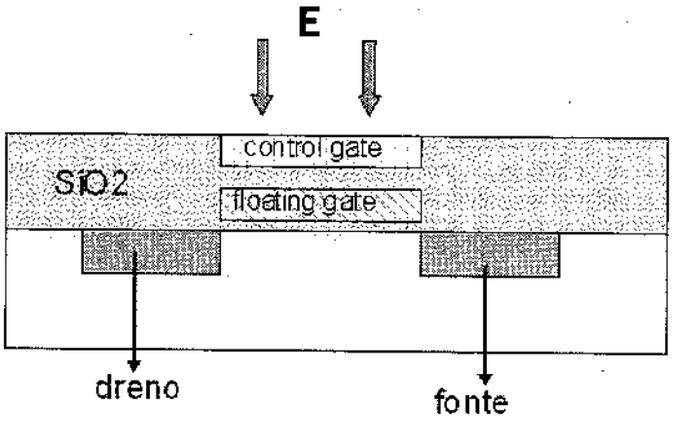

Fig.2.5)Estrutura FGMOS com dupla camada de polissílicio

Este tipo de comportamento do processo de tunelamento denomina-se auto-limitante (*self-limiting*). Portanto, a quantidade de carga incremental não é proporcional a quantidade de pulsos aplicados ao *control gate*, dado que a intensidade do campo elétrico através do óxido é função não somente da amplitude do pulso de programação, mas também da quantidade de cargas armazenadas no *floating gate*. Tais aspectos tornam difícil o controle das cargas armazenadas com alta precisão: ao se utilizar pulsos de alta tensão[2], a aplicação de poucos pulsos é utilizada para um ajuste fino da quantidade de cargas. Sob tensões de valores menores, obtém-se maior precisão em detrimento de maior tempo de programação exigido. A obtenção de um controle fino e preciso pode ser obtido aplicando-se pulsos de alta tensão em intervalos de tempo extremamente curtos ou sob tensões moderadas com controle da largura dos pulsos aplicados [1].

De acordo com o gráfico da fig.2.6, nota-se que a carga armazenada inicialmente aumenta linearmente com o tempo e depois satura. A corrente através da estrutura (dielétricos) permanece quase constante durante certo período, mas em seguida decresce rapidamente. O campo elétrico no óxido de *gate* decresce sensivelmente com o transcorrer do tempo [22] (o campo elétrico na fig.2.6 refere-se ao campo elétrico aplicado externamente).

[2] Tensões altas aplicadas podem causar *trap-up* (defeitos), reduzindo a confiabilidade.

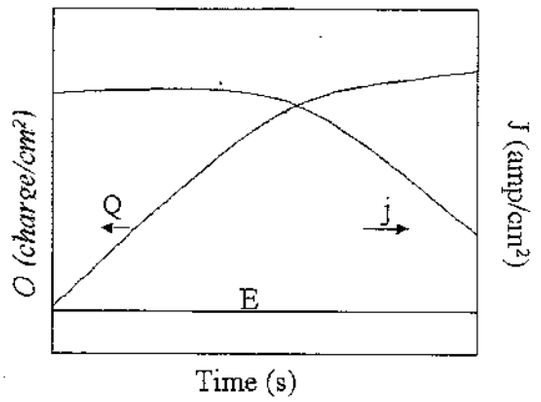

Fig.2.6) Variação de Carga (Q), densidade de corrente (J) e de campo elétrico externo (E) em função do tempo de programação [1].

Tal comportamento de carga, corrente e campo pode ser explicado: no instante inicial, quando um pulso de tensão é aplicado, a carga inicial é zero e o campo elétrico inicial através do óxido de *gate* tem seu valor máximo, E_{max} (não representado na fig.2.6). Com o passar do tempo, Q aumenta linearmente, já que nesta fase, como o valor de Q é relativamente pequeno, o campo elétrico permanece praticamente constante. Sob tais condições a corrente também mantém-se constante (corrente através do óxido), $Q = j_{(E_{MAX})}t$. Assim, quando Q torna-se suficientemente grande para que haja redução substancial de E, a corrente decresce rapidamente com o tempo e Q aumenta lentamente [1].

A maior limitação de precisão de dados de células de memórias *floating gate* programadas por tunelamento FN provém de acoplamentos capacitivos parasitas entre o nó de tunelamento e o *floating gate*. Mesmo sendo tal acoplamento pequeno, comparado a capacitância total do FG, a variação de tensão através de tais capacitâncias parasitas pode ser significativa quando a tensão de tunelamento é chaveada. Essa variação acopla-se ao FG, provocando um grande valor de tensão de *offset* na saída da célula de memória. Desde que os capacitores parasitas conectados ao *floating gate* armazenam cargas, a tensão de *offset* não decai, mas, em vez disso, permanece durante o período de tunelamento, o que provoca imprecisão do valor tensão (dado analógico) a ser memorizado [9].

2.2.1.2.Métodos de Programação de Memórias Floating Gate

Quando se deseja injetar elétrons no *floating gate*, escrita (*write)*, aplica-se um pulso de tensão V_{pp} (>11V) ao *control gate*, este pulso de tensão pode ser de qualquer tipo, desde que o valor de pico ultrapasse a tensão de limiar de tunelamento (> 11 V para tecnologias 0.6 e 0.8 μm AMS). Para se retirar elétrons do *floating gate*, apagamento (*erase*), ao *control gate* aplica-se $-V_{pp}$. Este procedimento tem o inconveniente da necessidade de fonte de duas polaridades dentro de um mesmo circuito integrado, fig.2.7.

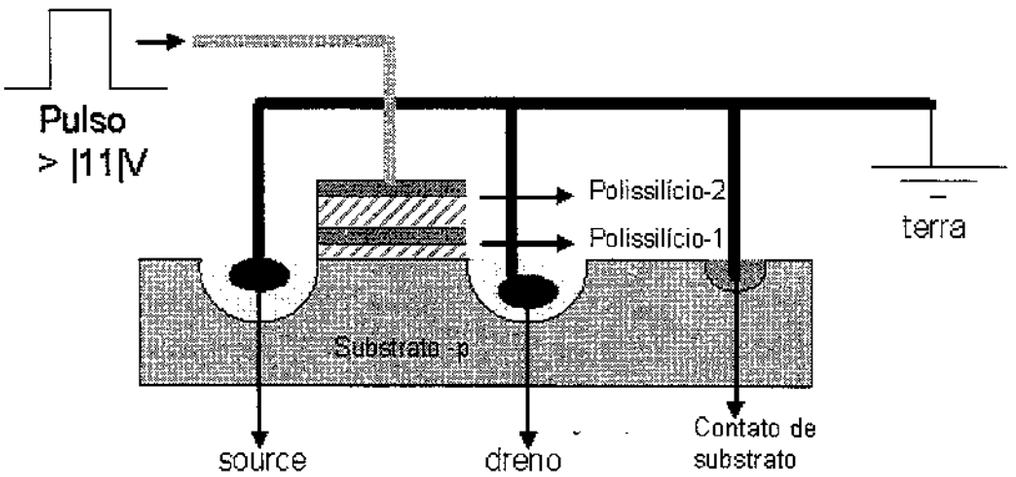

Fig.2.7) FGMOS com dreno, fonte e substrato aterrados. Ao gate são aplicados pulsos de tensão positivos ou negativos.

2.2.1.2.1.Mecanismo de tunelamento através de fonte bipolar

Ao se aplicar o pulso de programação ao *control gate* (tunelamento Fowler-Nordheim), a tensão divide-se de forma inversamente proporcional aos valores das capacitâncias entre as camadas de polissilício-1 (*floating gate*) e polissilício-2 *(control gate)*, (C_B), e entre polissilício-1 e substrato (C_g), configurando assim um divisor capacitivo de tensão, fig.2.8.

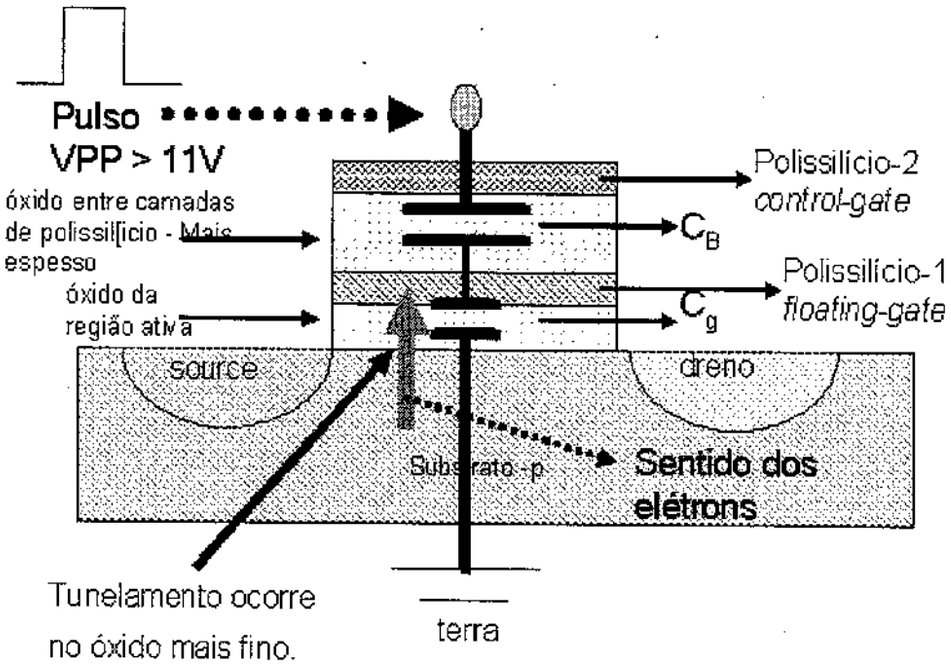

Fig.2.8) Escrita : esquema de capacitâncias e processo de tunelamento.

Sobre o capacitor de menor valor de capacitância estará a maior queda de tensão. Dessa forma, o tunelamento tende a ocorrer através do dielétrico que suporta a maior queda de tensão, ou se houver tensão suficiente, pode ocorrer nos dois dielétricos simultaneamente. Geralmente, o capacitor de menor valor de capacitância está entre o *floating gate* e o substrato. Atraídos pela polaridade positiva do pulso (para células nFGMOS), elétrons dirigem-se para a região de *floating gate* (polissilício-1), permanecendo aí aprisionados pela grande barreira de potencial dos dielétricos que contornam o *floating gate*. Elétrons aprisionados nesta região criam um campo elétrico reverso ao pulso externamente aplicado e, ao mesmo tempo, depletam elétrons da região de canal, aumentando, assim o valor efetivo da tensão de limiar.

A fig.2.9 ilustra o processo de apagamento. De forma similar ao processo de escrita, a tensão externa subdivide-se entre capacitores em série. Contudo, desta vez, os elétrons deixam a região de *floating gate*, fazendo o caminho inverso, e reduz-se o valor efetivo de Vt.

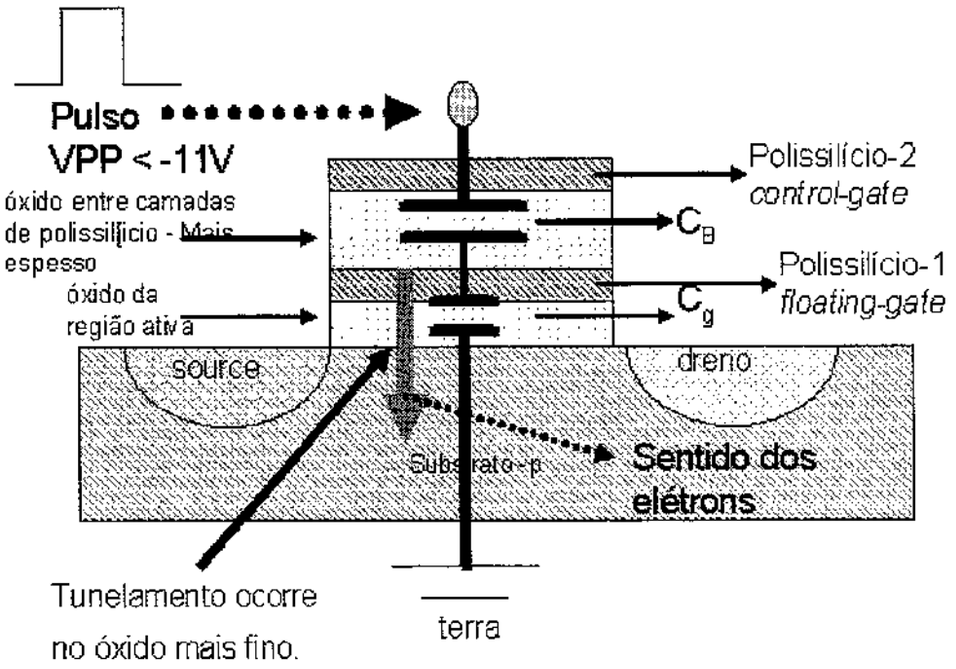

Fig.2.9) Apagamento: esquema de capacitâncias e processo de tunelamento.

A fig.2.10 ilustra um resultado de aplicação de pulsos positivos e negativos ao *control gate* de um transistor com *gate* flutuante.

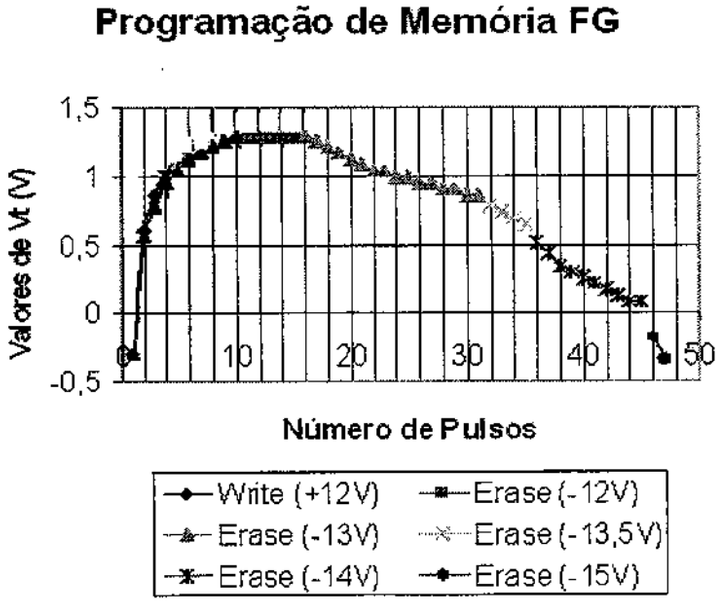

Fig.2.10) Resultados de programação através de tunelamento (Pulsos de tensão aplicados ao *control gate*). Resultados obtidos pelo autor desta tese.

Para valores positivos de pulso houve um incremento do valor de V_t, ao passo que para valores negativos, houve decremento, fig.2.10. Nota-se que para valores de V_t menores que 0, o dispositivo comporta-se como um transistor de depleção. Estes resultados foram obtidos em tecnologia 0.8 um CYE AMS CMOS.

2.2.1.2.2.Mecanismo de tunelamento através de fonte unipolar

Como dito anteriormente, o pulso de tensão subdivide-se entre as capacitâncias em série. No caso de se ter apenas um *control gate*, o acoplamento do capacitor formado entre o *control gate* e o *floating gate* será sempre o mesmo e, por conseguinte, a subdivisão de tensão manterá a mesma razão de proporcionalidade tanto para escrita quanto para apagamento. Ou seja, o tunelamento tenderá a ocorrer de forma predominante sobre apenas um dos dielétricos que compõe a estrutura *sandwiche* da fig.2.8. Assim, para a escrita terá de se aplicar sempre pulsos de polaridade oposta ao de apagamento, de forma a que elétrons ora entrem e ora saiam do *floating gate* através do mesmo dielétrico, fig.2.8 e fig.2.9. Apesar desse processo ser bastante eficaz, possui o inconveniente da necessidade de pulsos positivos e negativos [26].

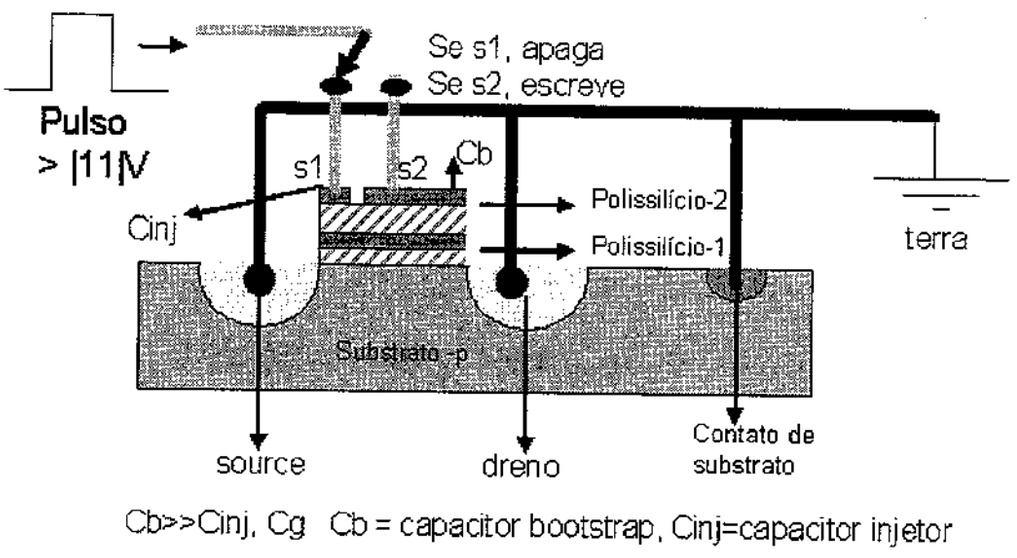

Fig.2.11) Esquemático de estrutura que prescinde de pulsos negativos para apagamento

Uma alternativa à aplicação de pulsos bipolares é a introdução de mais um *control gate* à estrutura, conforme a fig.2.11. A relação entre as capacitâncias deste novo *control gate* com o *floating gate* será diferente: neste caso, sendo ele confeccionado para ter menor valor de capacitância (capacitor de tunelamento, C_{inj}) a maior queda de tensão ocorrerá, portanto, sobre ele e o tunelamento se realizará preferencialmente através do dielétrico entre o *control gate* e o *floating gate* e não mais sobre o óxido entre o *floating gate* e a região de canal (substrato), fig.2.11.

Para operação de Apagamento, fig.2.12, aplica-se um pulso de tensão positiva ao capacitor C_{inj} e aterra-se o capacitor C_B, sendo $C_{inj} << C_B$. O valor de tensão aplicado em C_{inj} deve ser superior ao valor de tensão de limiar de tunelamento (> 11V). O tunelamento, neste caso, ocorrerá através do óxido entre as camadas de polissilício-1 e polissilício-2, capacitor C_{inj}, removendo assim cargas Q (elétrons) do *floating gate* (polissilício-1) e conseqüentemente aumentando o seu potencial.

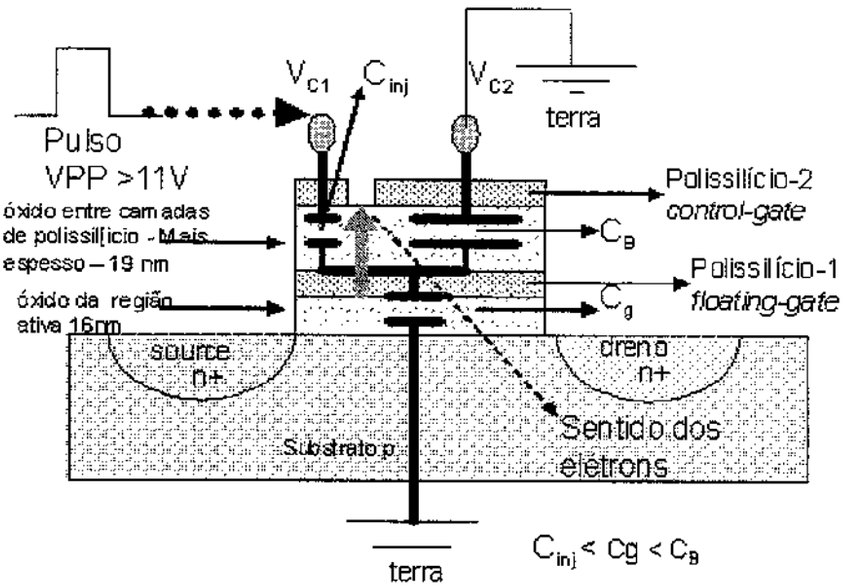

Fig.2.12) Apagamento : $V_{C1}=V_{PP}$, $V_{C2}= 0$ (terra), remove eletróns do gate (tunnel in C_2 oxide).

A capacitância equivalente $C_B//C_g$ é bem superior a C_{inj}, portanto, por divisão capacitiva de tensão, há maior queda de tensão sobre C_{inj}, onde o tunelamento acontece e cargas fluem do polissilício-1 (*floating gate*) para a fonte de pulsos de tensão.

Para Escrita, fig.2.13, permutam-se as tensões aplicadas sobre C_{inj} e C_B : aterra-se C_{inj} e aplica-se em C_B a tensão de tunelamento anteriormente aplicada a C_{inj}. O efeito de tunelamento, desta vez, ocorrerá noutro sentido: os elétrons (cargas) serão empurrados de

volta por V_{Cinj} e atraídos por V_{CB}, acumulam-se no *floating gate* (polissilício-1), o que decresce seu potencial.

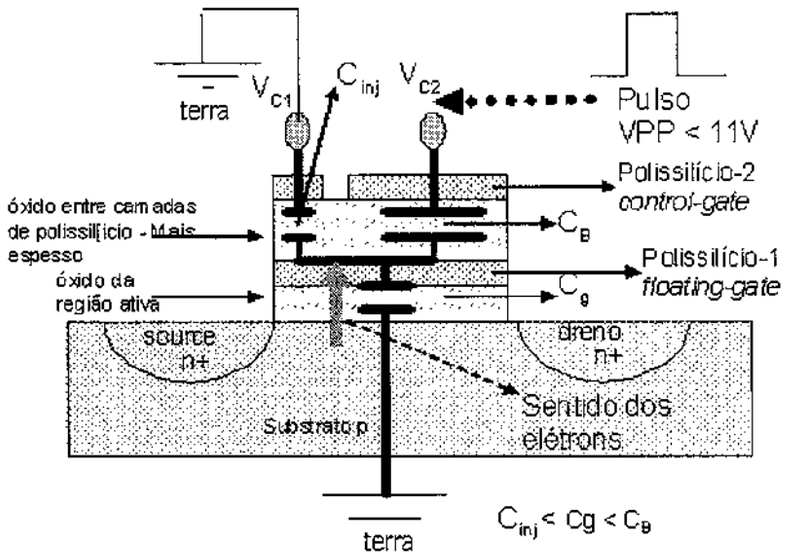

Fig.2.13) Escrita : $V_{C2}=V_{PP}$, $V_{C1}= 0$ (terra), tunelamento em C_{inj}.

Na escrita, C_B está em série com $C_{inj}//C_g$, ocorrendo portanto maior queda de tensão sobre a associação $C_{inj}//C_g$. Como o capacitor C_{inj} apresenta dielétrico mais espesso que C_g, é portanto através do dielétrico de C_g que o tunelamento ocorre.

A fig.2.14 mostra os resultados obtidos com a escrita e apagamento através dos capacitores de *bootstrap*, C_B, e de tunelamento, C_{inj}.

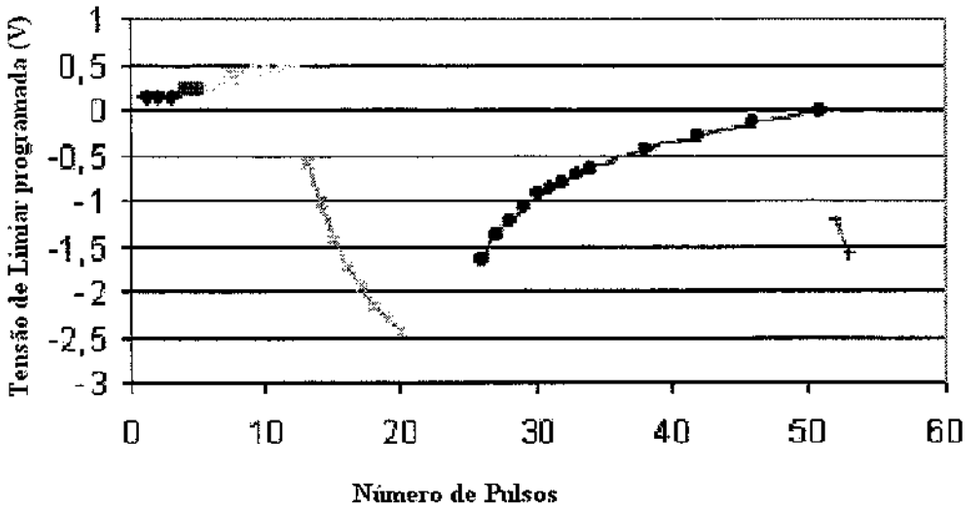

Fig.2.14) Resultados obtidos por programação (obtido em tecnologia AMS 0,6um) (ilustrativa).

2.2.1.3. Programação PWM usando tunelamento FN

Uma sugestão de esquema para programação PWM (modulação por largura de pulso) de memórias *floating gate* é apresentado na seguinte figura.

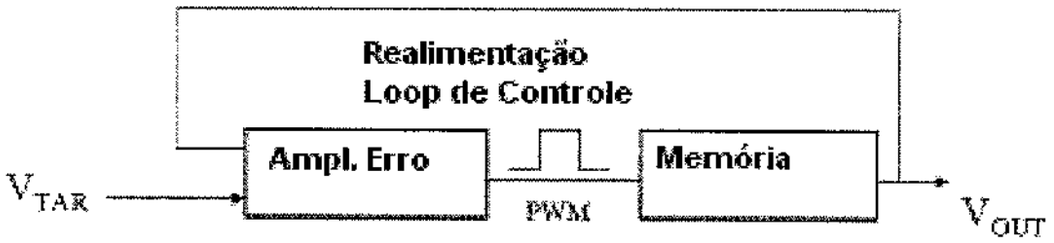

Fig.2.15) Esquema de programação PWM de memória analógica

O amplificador de erro obtém a diferença entre o valor de tensão armazenado na memória, V_{out}, e o valor de tensão a ser armazenado, V_{tar}, fig.2.15. O resultado da diferença entre V_{out} e V_{target}, na saída do amplificador de erro modula um sinal de largura de pulso, PWM, que contém informação de qual período de tempo será aplicada a tensão de tunelamento do *control gate* da memória [28].

Durante o pulso do PWM, o dispositivo FG altera o dado, ou seja, a quantidade de carga armazenada. Um pulso largo resulta quando há grande diferença entre V_{out} e V_{target}. Assim, comparativamente aos demais esquemas de programação, o tempo de programação pode ser reduzido, dado que o número de verificações necessárias para se atingir a nível de tensão de limiar a ser programado decresce. Por outro lado, para uma pequena diferença entre V_{out} e V_{target}, resulta um pulso estreito, já que os pulsos de escrita serão mais estreitos. Tempos maiores de largura de pulso de programação sob valores moderados de tensão conduzem a maior precisão. Em esquemas tradicionais, ao contrário, a precisão é limitada pela largura do pulso constante de escrita (em memórias digitais) [28].

BIBLIOTECA CENTRAL
CÉSAR LATTES
DESENVOLVIMENTO DE
COLEÇÃO
UNICAMP

2.2.2. Injeção de elétrons quentes (Hot-electron Injection)

A injeção de elétrons quentes é o mais limitado dos três tipos de programação. Este mecanismo envolve a imposição de determinados valores V_{DS} ao dispositivo assim como valores de V_{GS}. Nestas circunstâncias, elétrons que estejam no canal podem tunelar pelo óxido de *gate* na região próxima ao dreno (região de *pinch-off*), cf. fig.2.16.

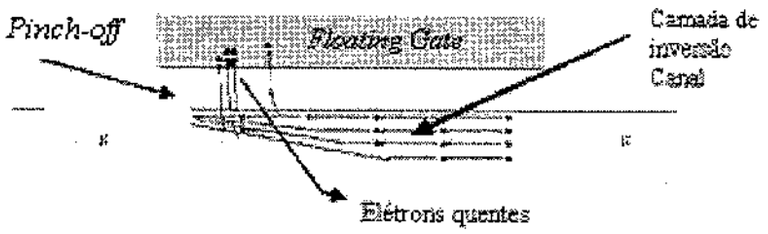

Fig.2.16) Ilustração do processo de programação por elétrons quentes na região de pinch-off.

Para o início do processo de programação por injeção de elétrons quentes, o dreno e o *floating gate* devem ser polarizados com tensões relativamente elevadas, porém inferiores às aplicadas para tunelamento FN. A polarização de dreno é realizada através da conexão direta do terminal de dreno com uma fonte de alimentação, ao passo que a polarização do *floating gate* depende do acoplamento capacitivo entre o *control gate* e o *floating gate*.

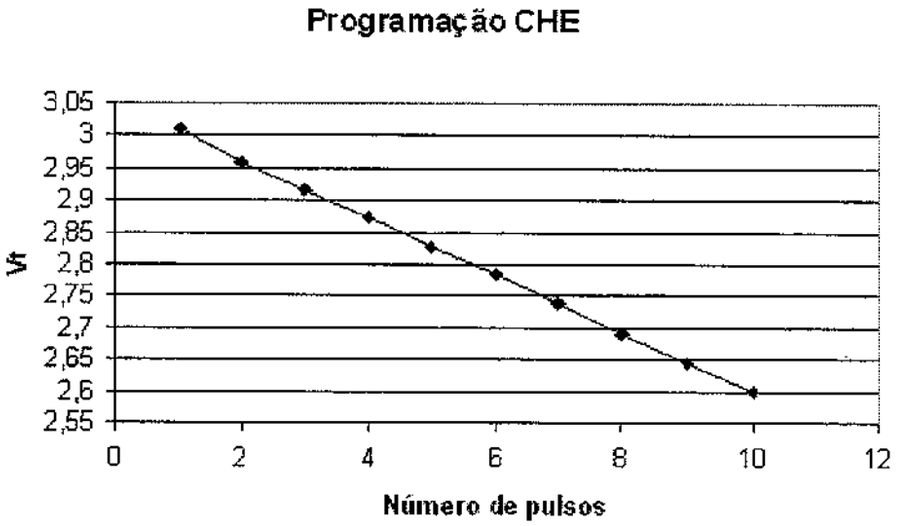

Fig. 2.17) Dados experimentais de apagamento de *floating gate* através de injeção de elétrons quentes.

A fig.2.17 mostra o resultado experimental de programação. Observa-se que a tensão de limiar é reduzida à medida que mais pulsos são aplicados.

Para que o processo de programação por elétrons quentes seja eficiente, o transistor deve ser polarizado em saturação, causando um intenso campo elétrico lateral criado na região de depleção perto do ponto de *pinch-off*. A fig.2.18 ilustra a situação descrita. A camada de inversão do canal é maior perto da fonte (*source*) e estreita-se a medida em que se aproxima do dreno, região de *pinch-off*. Quando os elétrons passam através da região de *pinch-off*, eles sofrem grande aceleração por conta do alto campo elétrico na região de depleção de dreno. Tal processo promove o espalhamento dos elétrons de forma a que alcancem a interface Si/SiO_2.

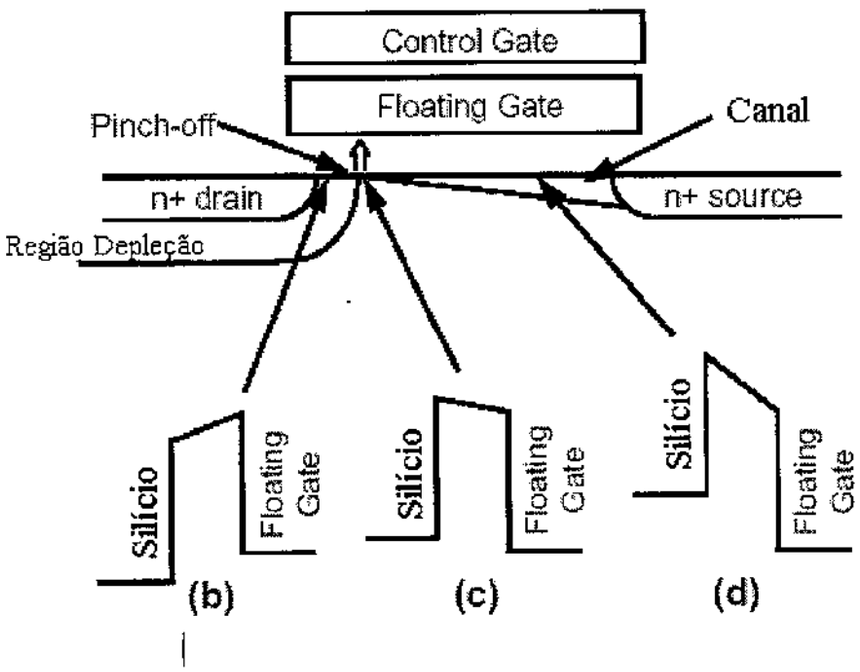

Fig.2.18) a) Ilustração do esquema de injeção de elétrons próximo a região de pinch-off com Transistor em Saturação [30]
(b)-(c)-(d) Apresentam os diagramas de potencial das bandas de condução nos pontos indicados pelas setas. Nota-se que, próximo ao dreno, o campo no óxido repele os elétrons injetados provenientes de substrato.

O formato da barreira de potencial do SiO_2 varia ao longo do canal em razão do potencial ao longo do canal também variar da fonte para o dreno enquanto o potencial do *gate* permanece constante.

A vantagem deste procedimento é a redução da tensão necessária ao tunelamento. Por outro lado apresenta a desvantagem de ser unidirecional e de possuir taxa de tunelamento variável, sendo um método que aumenta a quantidade de defeitos no óxido e, portanto, reduz o tempo de retenção de dados na memória. Adicionalmente, o mecanismo de injeção de elétrons requer corrente de canal da qual os elétrons a serem tunelados possam ser drenados. Assim, a necessidade de corrente e, portanto de consumo de potência, através deste mecanismo é muito maior que a do mecanismo de tunelamento FN [30].

2.2.3.Radiação de luz Ultra-Violeta UV

O mecanismo clássico utilizado para tornar o óxido de silício (SiO_2) condutivo é através de irradiação de luz ultra-violeta sobre sua superfície. O comprimento de onda usado neste procedimento é de 254 nm, ou seja, irradiação de alta energia [30].

A exposição do SiO_2 à luz UV gera pares elétron-lacuna com energia cinética suficiente para transpor a barreira de potencial que cerca a região de carga no *floating gate* [30].

Este método tem a vantagem de não requerer processamentos especiais e de ser amplamente difundido. Entre suas desvantagens há a necessidade de fonte de luz ultravioleta externa e a indução de condutâncias entre todas as camadas do circuito separadas por SiO_2, o que pode provocar curto-circuitos[21].

A eficiência de tal processo é baixa, posto que a corrente estabelecida no dielétrico é bem baixa. Em memórias EPROM, as células a serem apagadas têm de permanecer sob irradiação durante períodos de mais de meia hora.

Capítulo 3

Características de Memórias Analógicas *Floating Gate*

No projeto e caracterização de memórias baseadas em estruturas *floating gate*, vários fatores e parâmetros devem ser levados em conta. O tempo de retenção do dado na memória, por exemplo, confere certo grau de confiabilidade de precisão do dado armazenado em função do tempo, o que é muito importante durante a etapa de projeto do sistema onde a memória estará embarcada [11]. Assim, dependendo da finalidade do dado a ser armazenado, pode-se transigir que o dispositivo de memória apresente até mesmo baixo tempo de retenção. A repetibilidade do processo de programação (*endurance*) tem estreita relação com o tempo de vida da memória em função de seus ciclos de escrita/apagamento [30]. A amplitude do valor de tensão usado para programação é responsável pelo decréscimo da taxa de programação ao longo de vários ciclos de escrita, pois quanto menor o seu valor, menor tende a ser o grau de degradação sofrido pelo óxido através do qual se realiza o tunelamento [22]. Assim, este capítulo descreve os aspectos básicos mais relevantes à caracterização dos dispositivos *floating gate* e as implicações principais de cada um deles no projeto de tais estruturas.

3.1. Tempo de Retenção

Na operação **Escrita** (*Write*), o *floating gate* é carregado negativamente com elétrons. Isto é feito aplicando-se um pulso de tensão positiva ao *control gate*, mantendo-se fonte, dreno e substrato aterrados. A carga negativa no *floating gate* desloca a tensão de limiar, medida a partir do *control gate*, em direção a um valor mais positivo [30].

A operação de **Apagar** (*Erase*) remove elétrons do *floating gate* aplicando-se um pulso de tensão negativa ao *gate*, mantendo-se todos os outros terminais aterrados. A tensão de limiar então se desloca na direção de menores valores, podendo até mesmo se tornar um dispositivo de depleção (*overerased*) [30].

A operação de **Leitura** (*Read*), a priori, não deve alterar a quantidade de cargas no *floating gate*[30].

Um diagrama de potencial eletrônico de uma célula de memória apagada, com substrato e *control gate* aterrados é mostrado na fig.3.1. O campo elétrico através dos óxidos é devido à diferença entre as funções trabalho do *control gate* e substrato. No entanto, ao aplicar-se tensão positiva ao *control gate*, V_{CG}, obtém-se o diagrama da fig.3.2. Para níveis de tensão usados na operação de leitura, (V_{CG} > Vth) o campo na interface substrato/óxido é alto o suficiente para inverter a superfície e permitir o fluxo de corrente.

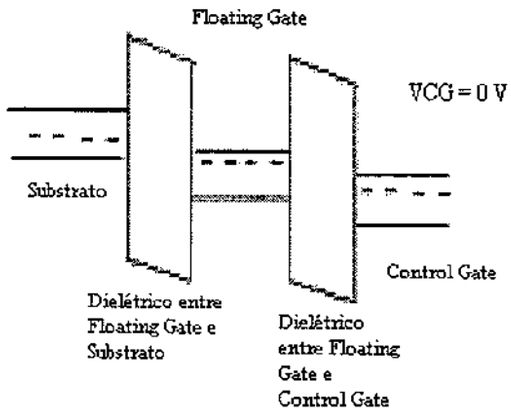

Fig.3.1.) Estado Apagado com VCG=0V [30]

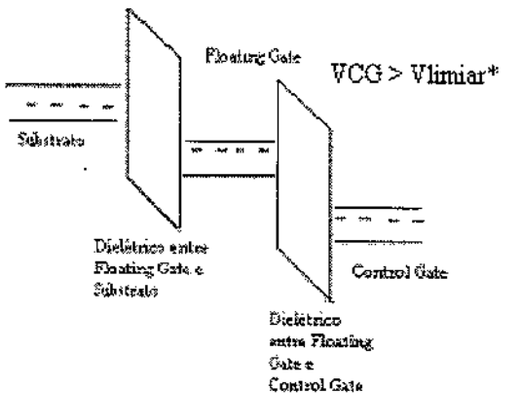

Fig.3.2)Estado Apagado comVCG >0 [30]

3.2.Endurance e envelhecimento

A injeção de portadores no *floating gate* não é repetitiva em cada ciclo. Impurezas e armadilhas (*traps*) no óxido que viabilizam o tunelamento Fowler-Nordheim sob baixa tensão (>12V) possuem distribuição aleatória no tempo. Isso motivou o desenvolvimento de circuitos de programação adaptativos (comparações sucessivas, iterativas, entre a tensão memorizada e a tensão desejada) [29].

No caso de memórias analógicas, a *endurance* é determinada pela quantidade de ciclos total permissível para escrita e apagamento. Em memórias digitais, o processo de escrita/apagamento injeta/retira uma quantidade fixa de carga no/do *floating gate*. Para memórias analógicas, no entanto, a quantidade de cargas para se escrever/apagar não é sempre a mesma, posto que o número de armadilhas (*trap-ups*) de cargas no óxido é aleatório, exigindo dessa forma um processo iterativo de programação: escritas e

apagamentos sucessivo até se alcançar o valor de tensão desejado (dentro de uma faixa de precisão obtida através da caracterização dos dispositivos e da tecnologia) [29].

Os *trap-ups* podem ser reduzidos ao se decrementar o máximo campo elétrico no óxido, aumentado-se o tempo de subida (*rise time*) do pulso de programação. Os elétrons presos no óxido promovem a redução da taxa de programação porque os elétrons que fluem através do óxido encontram pelo caminho campos elétricos repulsivos a sua passagem (devido aos *trap-ups*). Assim, como resultado do campo estabelecido pelos elétrons aprisionados a cada ciclo, a tensão de limiar (tensão de *threshold*) então programada não atinge o valor idealmente previsto (considerando-se que a largura e amplitude do pulso sejam as mesmas durante os ciclos). Tal efeito, que é denominado **colapso da janela de memória** (*Memory Window Collapse*), fig.3.3, é resultado de envelhecimento da memória em função de vários ciclos de programação da memória (*endurance*). Há um número limite de ciclos de programação, em memórias digitais, depois do qual a *Threshold Window* entre os estados programados e apagado colapsam-se rapidamente. O decaimento dos estados programado e apagado é indicativo da degradação do óxido em virtude da corrente de tunelamento através dele. Em memórias analógicas o *threshold window* não tem sentido, já que o tempo de programação é ajustado automaticamente a cada ciclo. Em memórias digitais tal tempo é fixo e não adaptativo. Contudo, mesmo memórias analógicas possuem *endurance* limitada, já que a largura do pulso ou a amplitude dos pulsos de tensão não poderão ser aumentados ilimitadamente [30].

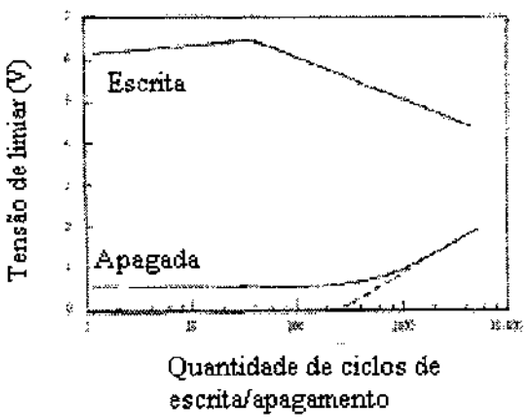

Fig.3.3) Tensão de limiar em estado escrito e apagado após de ciclos de programação. Após > 1000 ciclos, verifica-se o fenômeno de colapso [31].

Os *trap-ups* não constituem um mecanismo de falha da memória se um procedimento de programação iterativo com tensão de programação variável é utilizado: aumentado-se a tensão de programação e/ou a largura de pulso da tensão da mesma (até que atinja o valor máximo possível).

3.3. Tempo de Programação e Erasing

O tempo necessário para uma memória analógica ser programada para um valor desejado é função de seu estado inicial. Tal aspecto constitui uma diferença essencial entre memórias analógicas e memórias digitais. Nas digitais o tempo de programação é aproximadamente o mesmo entre atualizações sucessivas.

Nas memórias analógicas, por serem os tempos de programação variáveis e adaptivos, não ocorre o fenômeno de *endurance* denominado colapso da janela de memória (*Memory Window Collapse*), fig.3.3.

3.4. Tempo de Programação x Precisão

Precisão e velocidade de programação são importantes parâmetros no projeto e estudo de memórias analógicas. Em geral, tempos reduzidos de programação requerem correntes de tunelamento relativamente mais elevadas, das quais as magnitudes são de difícil controle e mensuração. Por outro lado, correntes mais baixas são mais fáceis de controlar de modo a se obter maior precisão, porém a custa de maior tempo de programação. Algoritmos de programação podem ser otimizados para reduzir o tempo de programação, as expensas, portanto, de menor resolução (faixa dinâmica). O contrário também é verdadeiro: maior resolução em vez de menor tempo.

A literatura registra a utilização de estruturas *floating gate* para a confecção de memórias analógicas. Tal prática apresenta três inconvenientes [15]:

1-Os valores de tensão no *floating gate* no momento da programação e no momento da leitura são diversos um do outro. Isso se deve a erros causados por acoplamentos capacitivos. Para contornar tal problema, faz-se necessário usar um processo iterativo

de escrita e leitura para memorização de algum dado analógico. Muitos ciclos de escrita, por sua vez, resultam em maior tempo de programação.

2-Aumento drástico no tempo de programação quando alta resolução é almejada. Devido ao fato de o comportamento da corrente de tunelamento ser logarítmico, (Fowler-Nordheim), ou seja, com conseqüente queda da taxa de programação (Efeito auto-limitante). Inicialmente a quantidade de cargas injetadas é grande, porém à medida que cargas vão sendo armazenadas no *floating gate* um campo reverso se opõe ao externamente aplicado, reduzindo assim a taxa de programação de forma logarítmica.

3-A máxima corrente de tunelamento é muito maior que o valor médio de corrente necessário para programação, o que, inevitavelmente, acelera a degradação do óxido pelo qual ocorre o tunelamento.

3.5. Carga remanescente pós-processo

Durante a etapa de fabricação dos dispositivos e memórias *floating gate* é natural que, em decorrência do processamento térmico e/ou químico, haja incorporação de quantidade aleatória de cargas na região do *floating gate*. Tais cargas alteram significativamente o valor da tensão de limiar dos dispositivos. Para a remoção desta carga remanescente a literatura propõe [9], a exemplo de memórias EPROM comerciais, a exposição do dispositivo à luz ultra-violeta (UV).

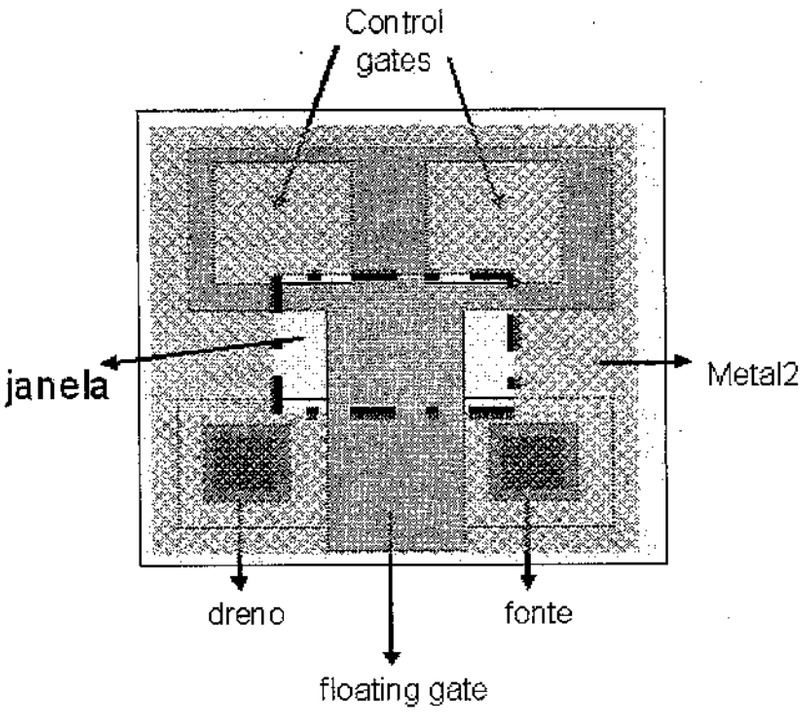

Fig.3.4) Layout do FGMOS com camada de metal-2. A região hachurada é a janela onde incidirá luz UV [6]

O procedimento de iluminação UV promove o descarregamento do *floating gate*. No entanto, em tecnologia CMOS padrão há de se remover o óxido de passivação para que se viabilize tal método. Para tanto, como a fig.3.4 mostra, é preciso, durante a etapa de confecção de *layout*, colocar uma camada de *PAD* sobre a região a ser iluminada. Com a retirada o óxido de passivação pode haver danos ao transistor FGMOS devido a etapa de corrosão para a retirada do óxido. Em substituição à proteção fornecida pela passivação retirada, coloca-se sobre toda a região a proteger uma camada de Metal-2, fig.3.4, o que adiciona ainda mais capacitâncias e acoplamentos parasitas à estrutura [6].

Outras maneiras de alteração de Vt são através de tunelamento Fowler-Nordheim, injeção de elétrons quentes (*hot-electrons*) e à exposição a altas temperaturas sobre longos períodos em que as correntes de fuga intensificam-se e descarregam o *floating gate* [30].

3.6. Características Estruturais de dispositivos FG

Todas as estruturas implementadas em tecnologia CMOS que estão separadas por finas camadas de óxido são potenciais candidatas à transferência de carga para um *floating gate*. O capacitor padrão feito com uma fina camada de óxido entre duas camadas de polissilício pode ser usado (em tecnologias com duas camadas de polissilício disponíveis), fig.3.5 e fig.3.6.

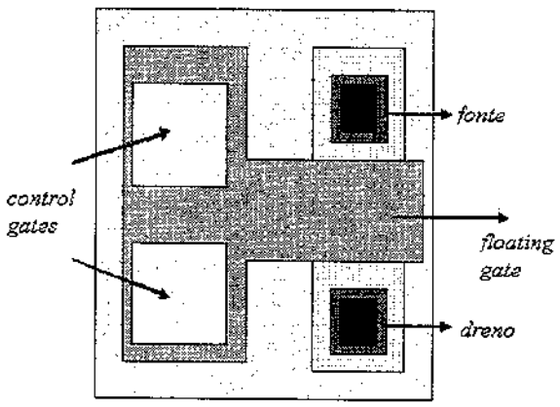

Fig.3.5) Layout de nFGMOS

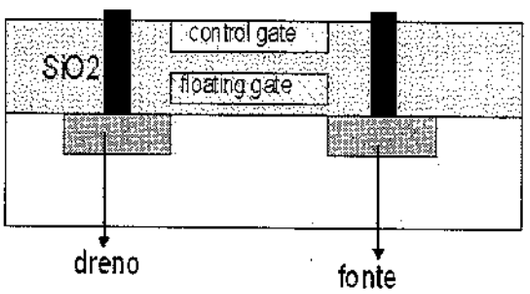

Fig.3.6) Estrutura de camadas de nFGMOS

3.7.Modelagem de capacitâncias do transistor FGMOS

A estrutura física de um transistor MOS inclui vários acoplamentos capacitivos como indicado na fig.3.7. Como é bem conhecido da modelagem do transistor MOS, o valor das capacitâncias é variável com as condições de polarização, sendo usualmente mais pronunciado, em condições normais de operação, o acoplamento capacitivo entre *gate* e fonte. Quando, no entanto, não há canal formado sob o *gate*, o acoplamento capacitivo entre este e o substrato (*bulk*) é predominantemente maior que os demais. Além disso, há sobreposição, ainda que pequena, entre o *gate* e as regiões de dreno e fonte. Todas estas estruturas capacitivas podem ser utilizadas para a transferência de carga para o *floating gate*.

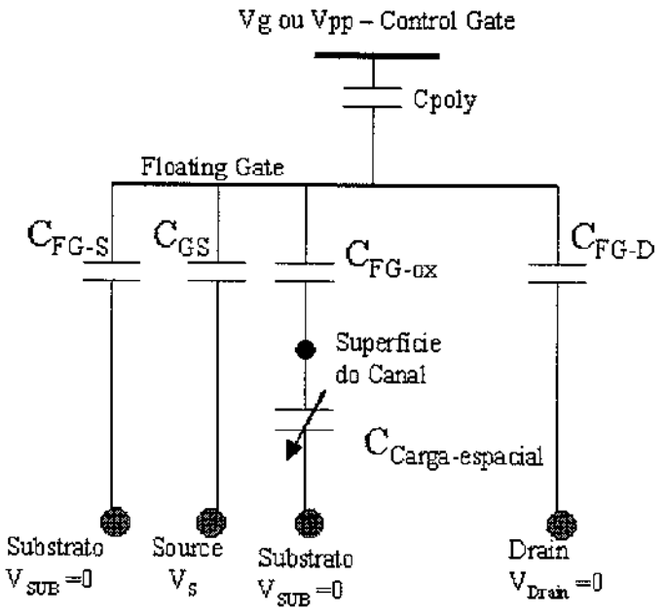

Fig.3.7) Modelo de acoplamentos capacitivos num transistor FGMOS. Cpoly = capacitância inter-polissilício entre o *control gate* e o *floating gate*, C_{FG-S} = capacitância entre o *floating gate* e o substrato, C_{GS} = capacitância *Gate*-Fonte, C_{FG-ox} = Capacitância entre poly1 e o canal. $C_{Carga-espacial}$ = capacitância espacial da região de canal e C_{FG-D} = capacitância entre o *floating gate* e o dreno [30]

3.8.Como obter o valor analógico memorizado

Através dos modos de programação (escrita e apagamento) pode-se ajustar a tensão de limiar, Vt, ao valor desejado. O valor desta tensão analógica, que resulta da quantidade de cargas armazenadas no *floating gate* e da capacitância do óxido, constitui o dado armazenado, a memória.

Contudo, como se pode obter o valor da tensão de limiar armazenada? Parece, a primeira vista, inviável, a todo momento que se desejar efetuar uma operação de leitura na memória, extrair-se a tensão de limiar do componente, uma vez que isso implica na repolarização do dispositivo (os terminais de dreno e *control gate* são curto-circuitados e o terminal fonte aterrado para extração do valor de V_T para um nFGMOS). O valor memorizado pode ser obtido indiretamente através de duas maneiras:

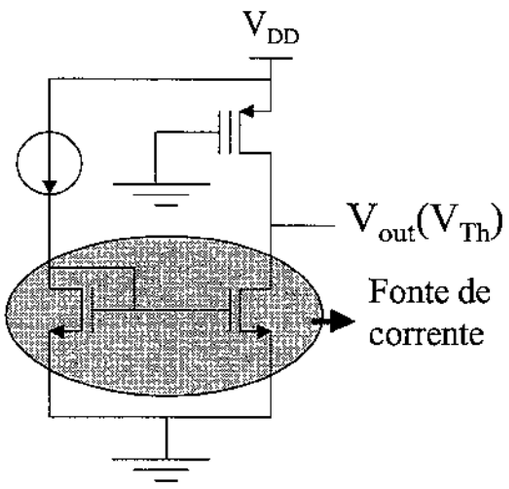

Fig.3.8) Variação da tensão de saída

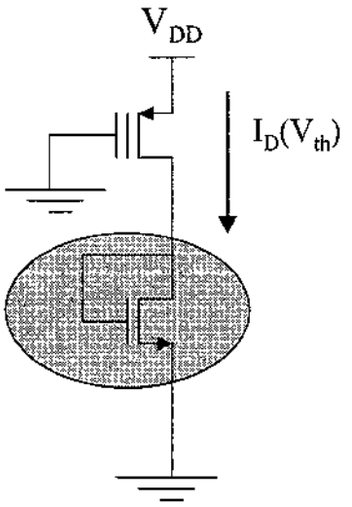

Fig.3.9) Variação da corrente de saída

i. Através da variação da tensão de saída, fig.3.8. A tensão Vout está diretamente relacionada ao valor da tensão de limiar programado. Para uma dada corrente de polarização através do FGMOS, uma curva de correlação entre Vout e a tensão de limiar pode ser obtida

ii. Através da variação da corrente de saída no dreno do dispositivo, fig.3.9. A corrente resultante no dreno do FGMOS é resultante da tensão de limiar programada, estando o dispositivo na região linear ou na de saturação.

3.9.Layout de Transistores Floating Gate

O cuidado com o *layout* de estruturas *floating gate* em tecnologia CMOS deve ser redobrado a fim de se evitar acoplamentos capacitivos indesejados. A quantificação dos valores de capacitâncias parasitas deve ser realizada, de modo a que sejam consideradas posteriormente na operação do FGMOS.

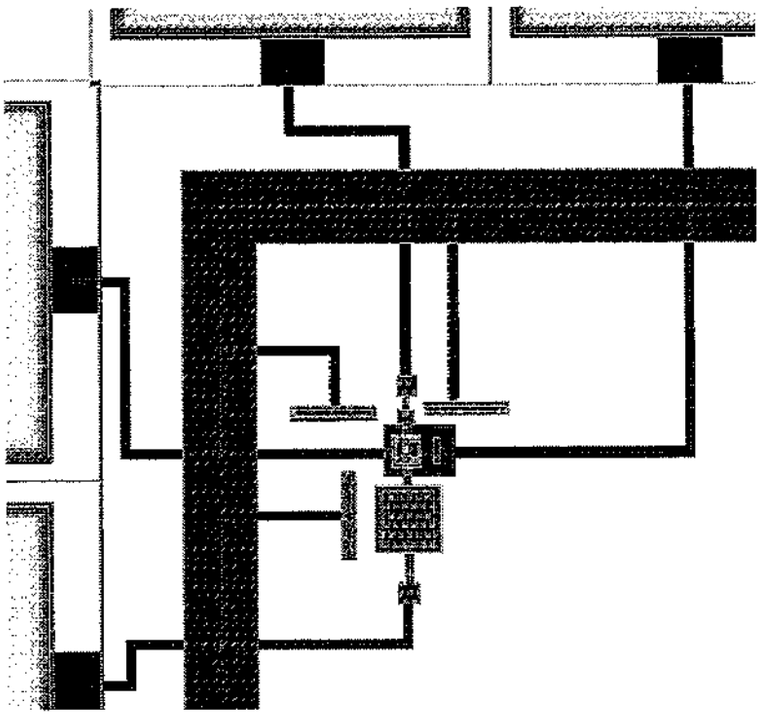

Fig.3.10) Layout de pFGMOS, 0.8 µm AMS

Cuidados a serem observados no *layout* de um FGMOS:

i. Não passar camadas de metal sobre o *floating gate* (polissilício-1 da área ativa), posto que haveria inserção de mais um acoplamento capacitivo, figs.3.11 e 3.12.

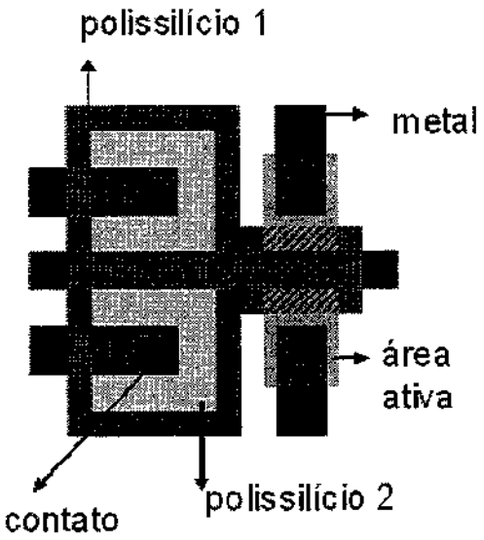

Fig.3.11) Layout, forma incorreta : Deve-se evitar trilhas de metal sobre a estrutura, evitando-se acoplamentos indesejados.

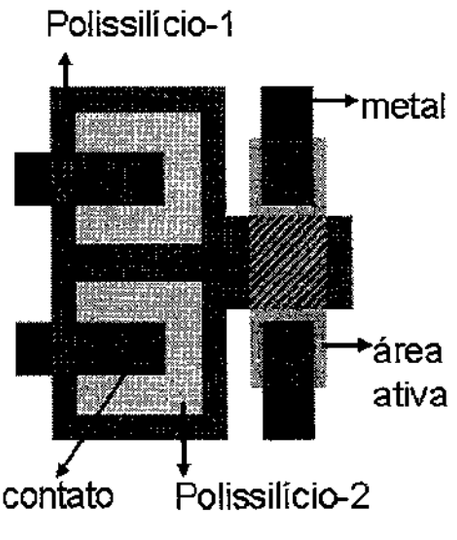

Fig.3.12)Layout, forma Correta: sem cruzamentos com trilhas de metal.

ii. O *floating gate* não deve estar sobre fronteiras de poço com substrato. Deve estar inteiramente contido ou sobre o poço, ou sobre a região de substrato, figs.3.13 e 3.14.

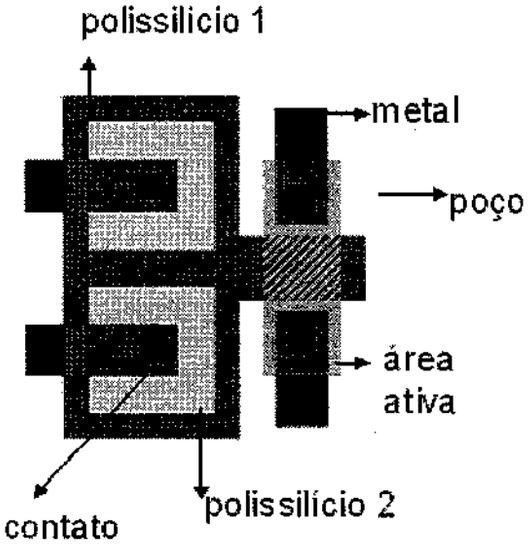

Fig.3.13)Layout, forma incorreta : *floating gate* sobre fronteira de poço com substrato. Neste caso pode haver acoplamento com a polarização do poço (VDD).

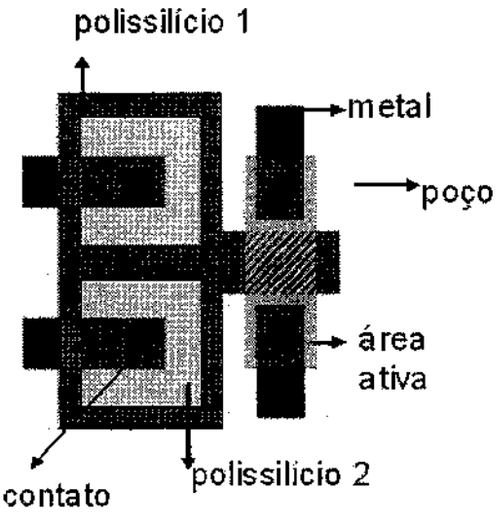

Fig.3.14)Layout, forma correta: O *floating gate* deve estar inteiramente sobre o poço, de forma que em sua totalidade ele esteja sujeito aos mesmos eventuais acoplamentos.

iii. Garantir boa polarização de substrato (de forma a evitar *Latch-ups* ou polarização indevida de junções).

iv. O *floating gate* deve ser inteiramente constituído por polissilício-1, ou seja, não se deve ligar duas partes de um *floating gate* por pontes de metal (a inserção de contatos na estrutura do *gate* provocaria a perda de carga armazenada). Os óxidos que envolvem o metal são tipicamente óxidos depositados e não óxidos crescidos (este de melhor qualidade), fig.3.15.

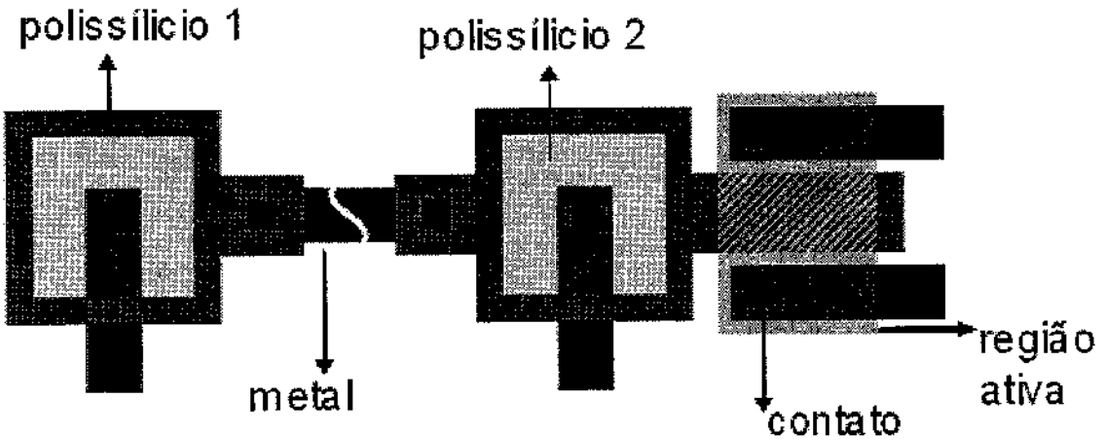

Fig.3.15) *Floating gates* conectados através de polissilício-1

v. Havendo necessidade de estruturas multi-*gates* é recomendável a inserção de capacitores e estruturas *Dummy*, fig.3.16, de forma a garantir a simetria e bom casamento entre as estruturas.

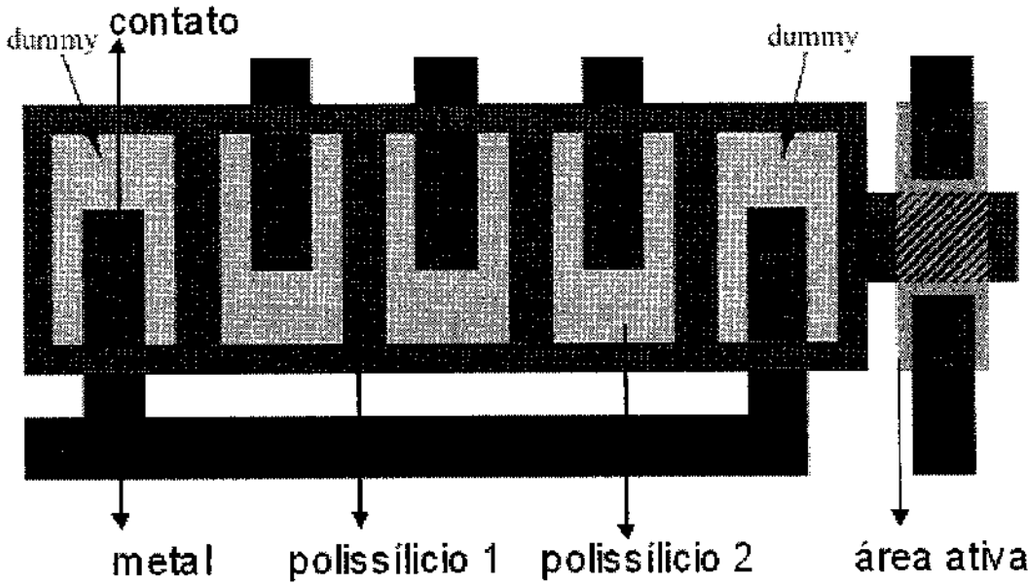

Fig.3.16) Capacitores dummies em FGMOS

vi. Para a caracterização de estruturas *floating gate* de testes faz-se necessária a retirada dos diodos de proteção do *pads*, já que o processo de programação exige a aplicação de tensões bem superiores aos ±5V da tecnologia.

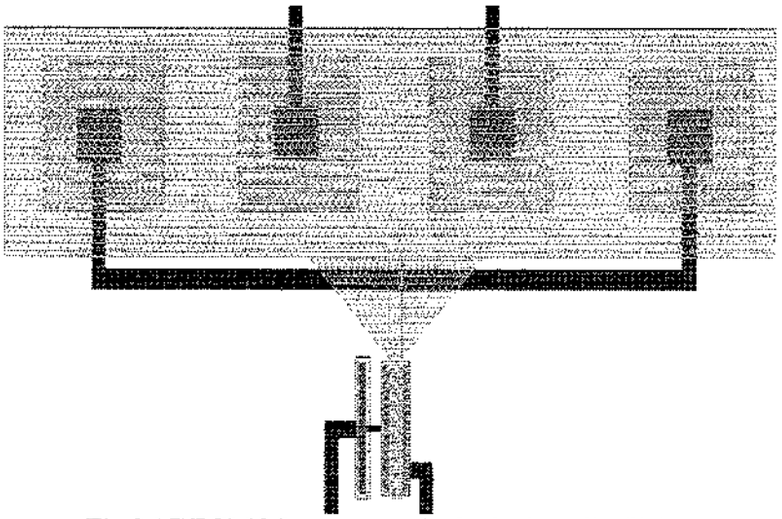

Fig.3.17)FGMOS com camada dupla de polissílicio

A fig.3.17 apresenta o *layout* de implementação de dispositivo FGMOS: camadas de polissício sobrepostas e intercaladas com dielétrico fora da região ativa do transistor. Pelo *layout* da fig.3.17, identificam-se quatro *control gates*, sendo dois deles *Dummy Capacitors*. A introdução destes últimos visa garantir que os efeitos de diferenças de geometria sobre os capacitores sejam minimizados, já que enxergam bordas idênticas (simetria).

Capítulo 4

Operação de Transistores FGMOS

Os transistores FGMOS, além de armazenarem carga, prestando-se, portanto, como dispositivos de memória, podem ser utilizados como dispositivos de múltiplos *gates* (*control gates*). Diversas aplicações analógicas e blocos funcionais podem ser implementados com FGMOS de múltiplos *gates*, tais como: pares diferenciais [2], espelhos de corrente [2] [8], conversores D/A [14]. A multiplicidade de *gates* confere versatilidade aos circuitos : enquanto a um dos *gates* é aplicada uma tensão de polarização DC do dispositivo, os outros gates ficam disponíveis para a aplicação de sinal AC, ou seja, polarização do dispositivo e injeção de sinal são aplicadas a entradas diferentes. Neste capítulo é apresentada uma referência de tensão em temperatura com dispositivos FGMOS. Tal circuito é inédito na literatura.

4.1. Estrutura dos dispositivos FGMOS multi-gates

A estrutura básica e símbolos de nFGMOS apresentados nas figs.4.1 e 4.2 representam dispositivos com n-entradas. Cada uma destas entradas, *control gate*, é capacitivamente acoplada ao *floating gate* [2].

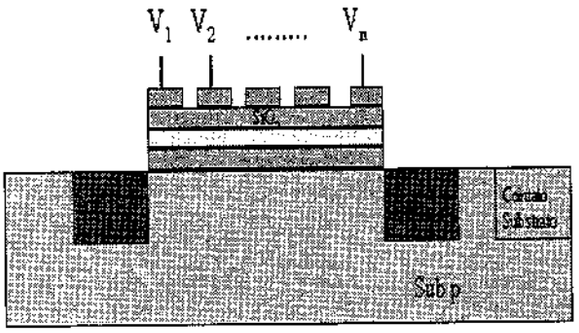

Fig.4.1) Esquema estrutural de nFGMOS

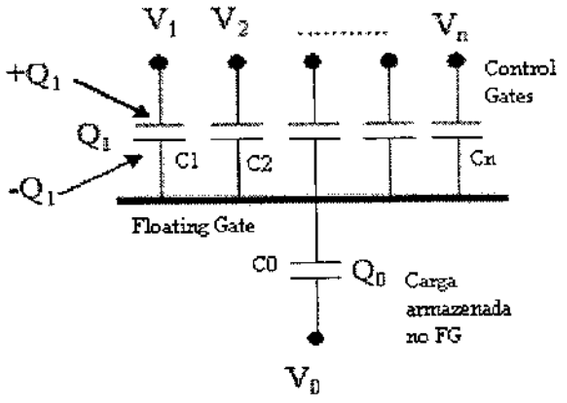

Fig.4.2) Esquema de acoplamentos [2].

BIBLIOTECA CENTRAL
CESAR LATTES
DESENVOLVIMENTO DE
COLEÇÃO
UNICAMP

As tensões terminais e os coeficientes de acoplamento capacitivos são definidos na fig.4.3, onde ϕ_F é o potencial do *floating gate*, V_1, V_2, ..., V_n são os valores de tensão de entrada, C_1, C_2,.., C_n, as capacitâncias entre o *floating gate* e cada um dos *control gates*. C_0 é o valor de capacitância entre o *floating gate* e o substrato. Q_1, ...Q_n são as cargas armazenadas em cada um dos capacitores. Q_0 é a carga inicialmente armazenada no *floating gate*. [1].

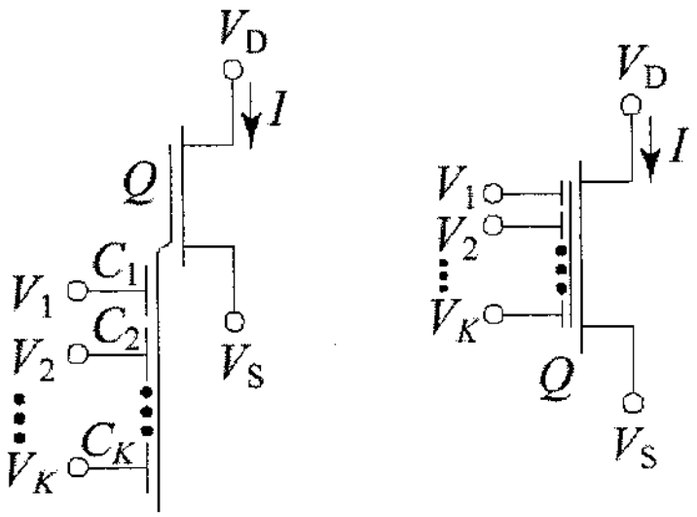

Fig.4.3) Símbolos usuais para FGMOS de K entradas [5].

Considerando os elementos definidos na fig.4.3, seja Q_F a carga armazenada no *floating gate*, calculada como:

$$Q_F = Q_0 + \sum_{i=1}^{n}(-Q_i) = \sum C_i(\phi_F - V_i) = \phi_F \sum_{i=0}^{n} C_i + \sum_{i=0}^{n} C_i V_i \qquad (4.1)$$

Assume-se, inicialmente, que não ocorre injeção de carga durante o funcionamento do dispositivo, sendo assim Q_F é igual à carga inicialmente armazenada *no floating gate*, Q_0, que é assumida ser zero (por simplicidade). Tal situação não impede a generalização, como será visto mais adiante. Todas as tensões são tomadas relativamente à terra. V_S e V_{Sub} denotam os potenciais de fonte e substrato respectivamente e, neste caso, V_0=V_S=0. A equação (4.1) reduz-se a:

$$\phi_F = \frac{C_1V_1 + C_2V_2 + ... + C_nV_n}{C_{TOT}} \qquad (4.2)$$

$$\text{Onde:} \quad C_{TOT} = \sum_{n=1}^{n} C_i$$

A equação (4.2) revela claramente um dos aspectos mais importantes destes dispositivos: O potencial do *floating gate* ϕ_F é expresso como a soma de todas as entradas ponderadas pelos valores de capacitâncias dos *control gates*. Os sinais de tensão são diretamente ponderados e somados ao nível do gate sem qualquer dissipação de potência, posto que o divisor de tensão é capacitivo e não resistivo. Tal característica é uma das mais importantes da técnica de acoplamento capacitivo comparada aos métodos tradicionais de divisores resistivos (*wired sum*).

O valor de ϕ_F é determinado unicamente por (4.2), tendo em vista que nenhuma das capacitâncias C_i varia durante a operação do dispositivo. Dentre todos os C_i's, somente C_0 pode variar dependendo da condição de operação do transistor. O valor de C_0 pode ser considerado como sendo o valor da capacitância do óxido do *gate* quando o transistor se encontra no limiar da condução (tensão entre gate e fonte acima do limiar, V_{th}).

A partir daqui pode-se introduzir o parâmetro γ, definido como:

$$\gamma = \frac{C_1 + C_2 + ... + C_n}{C_{TOT}} = \frac{C_{TOT} - C_0}{C_{TOT}} \qquad (4.3)$$

γV_{DD} representa o máximo potencial de *floating gate* obtido quando todas as entradas estão conectadas ao V_{DD}. Desde que γ é um ganho de tensão do *floating gate* como resultado do acoplamento capacitivo de todos os *input gates*, γ é denominado ***fator de floating gate***, um dos parâmetros chave no projeto de circuitos e dispositivos.

Seja V_{TH}^* a tensão de limiar do transistor vista do *floating gate*. Então o transistor inicia a condução ("liga") quando satisfaz-se a condição $\phi_F > V_{TH}^*$, isto é:

$$\frac{C_1V_1 + C_2V_2 + ... + C_nV_n}{C_{TOT}} > V_{TH}^* \qquad (4.4)$$

Como base na equação (4.4), pode-se concluir que quando a soma ponderada de todos os sinais de entrada excede o valor da tensão de limiar, o transistor inicia a condução.

Ajuste da tensão de limiar V_{TH}

Rearranjando-se a relação (4.4) em função de V_1, tem-se:

$$V_1 = \frac{C_{TOT}}{C_1} V_{TH}^* - \frac{C_2}{C_1} V_2 - \frac{C_3}{C_1} V_3 - ..... - \frac{C_n}{C_1} V_n \qquad (4.5)$$

Se um transistor FGMOS de n entradas for considerado como um MOSFET de entrada simples em que o *gate* 1 é somente um sinal de entrada e os outros *gates* são para controle da tensão de limiar, então a tensão de limiar do MOSFET vista do *gate* 1 é dada por:

$$V_{TH}^{(1)} = \frac{C_{TOT}}{C_1} V_{TH}^* - \frac{C_2}{C_1} V_2 - \frac{C_3}{C_1} V_3 - ..... - \frac{C_n}{C_1} V_n \qquad (4.6)$$

De (4.6) observa-se a dependência de $V_{TH}^{(1)}$ aos sinais de controle V_2, V_3, ..., V_n.

Considerando-se agora o caso mais simples de um FGMOS com duas entradas, com C_1=C_2, obtém-se:

$$V_{TH}^{(1)} = \frac{C_{TOT}}{C_1} V_{TH}^* - V_2 \qquad (4.7)$$

Portanto, a tensão de limiar do MOSFET é controlada por um sinal analógico V_2.

4.2.Circuitos elementares com dispositivos FGMOS

Para um transistor FGMOS de k entradas, a tensão de *floating gate* é dada pela seguinte expressão, eq.4.8:

$$V_{FG0} = \frac{m_1 V_{g1} + m_2 V_{g2} + ... + m_n V_{gn}}{m_1 + m_2 + ... + m_n} \quad (4.8)$$

Onde os parâmetros m_1, m_2,.....,m_n são ponderações (coeficientes) de acoplamento capacitivo. Tal expressão é válida quando a capacitância total de acoplamento é maior que a capacitância MOS entre o *gate* e o substrato (caso contrário, haveria um divisor de tensão capacitivo e a tensão resultante sobre a capacitância MOS seria menor que a tensão de limiar, V_{Th}).

Desde que o MOSFET esteja operando na região de saturação, a corrente de dreno I_D pode ser descrita por:

$$I_d = K(V_{FGS} - V_T)^2 \quad (4.9)$$

Onde K é o parâmetro de transcondutância, V_{FGS} é a tensão *floating gate-Source* e V_T é a tensão de limiar.

A possibilidade de se controlar a tensão de limiar relativa a um dos terminais de entrada, ou através do tunelamento/injeção de elétrons no *floating gate* pode tornar o uso destes dispositivos adequado à implementação de circuitos *low-power*. Além disso, a soma ponderada de todas as tensões de entrada nos *control gates* é implementada por divisores que se utilizam de acoplamento capacitivo e não resistivo, e essencialmente nenhum outro fluxo de corrente, que não seja o de carregamento e de descarregamento dos capacitores, está presente. **Tal natureza *low power* do FGMOS é crucial na realização de integração de alta densidade.**

Para a implementação de dispositivos elementares, a idéia básica de utilização de disposistivos FGMOS reside na separação entre polarização e injeção de sinal: a um dos *control-gates* uma tensão de polarização é aplicada e, a outro *control gate*, o sinal, para processamento do sinal propriamente dito. Tal manobra permite reduzir ou mesmo cancelar a tensão de limiar, de tal maneira as que outras entradas (*control-gates*) dispensem sinais de polarização para o funcionamento do transistor.

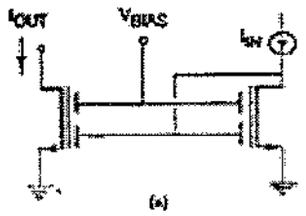

Fig.4.5)Espelho de corrente [13]

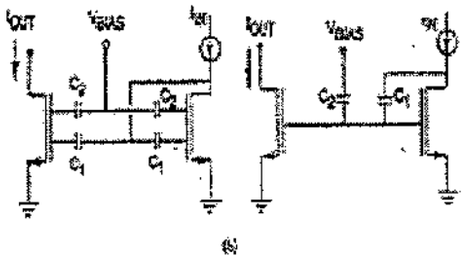

Fig 4.6)Implementações diversas de espelhos [13].

A fig.4.5 mostra um espelho de corrente de baixa tensão que utiliza dois FGMOS de duas entradas. A fig.4.6 mostra duas maneiras diferentes de se implementar espelhos de corrente. Na primeira delas, os terminais de entrada, assim como C_1 e C_2, são compartilhados por ambos os transistores. Nestes circuitos, um dos terminais de entrada de cada transistor é conectado a uma fonte de polarização DC, V_{Bias}; enquanto os outros terminais são usados como entradas e saídas convencionais de um espelho de corrente. Nota-se que toda a faixa de tensão de alimentação, VDD-VSS, está disponível para polarizar o espelho de corrente.

4.3.Compensação de Tensão em Temperatura com Dispositivos FGMOS[3]

Diversas são as possibilidades de implementação de Referências de Tensão, entre elas:

1- As que se utilizam do princípio de compensação térmica entre a tensão V_{be} de transistores bipolares denominado *Bandgap* [17].

2- As que se utilizam de transistores CMOS com dopagens opostas *de gate* de forma que a diferença de tensões *gate-source* dos transistores possa ter coeficiente de temperatura oposta a de um transistor simples, permitindo portanto a compensação térmica [18]

3- Combinação de transistores MOS de depleção e enriquecimento de forma a haver diferença de tensão que compense a tensão V_{GS} de um transistor em relação a temperatura [18].

A utilização de transistores *floating gates* (FGMOS) [19], ao contrário das opções citadas acima, é compatível com a tecnologia CMOS padrão, não exigindo passos extras no processo de fabricação dos mesmos. O protótipo foi confeccionado em tecnologia CMOS padrão 0.6 μm AMS.

4.3.1.O mecanismo de compensação

A equação de um transistor polarizado na região de saturação é a seguinte:

$$I_D = \frac{\mu_n C'_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{th})^2 \quad (4.10)$$

A partir desta equação obtém-se o valor de V_{GS} que quando aplicado ao *gate* do transistor mantém a corrente no dreno, I_D, constante (derivação em função da temperatura). Este valor de V_{GS} é definido a partir de agora como V_{GS0}:

$$V_{GS0} = V_{th} + 2\mu_n \frac{\partial V_{th}/\partial T}{\partial \mu_n/\partial T} - 2\mu_n \frac{\partial V_{GS}/\partial T}{\partial \mu_n/\partial T} \quad (4.11)$$

[3] Resultado de trabalho apresentado no IMAPS-Brazil, 6-8 Agosto, Campinas-SP [19]

A mobilidade, que é função da temperatura, tem a seguinte expressão empírica [18]:

$$\mu_n(T) = V_{th}(T_0)\left(\frac{T}{T_0}\right)^{\alpha\mu} \quad (4.12)$$

A expressão da tensão de limiar, por sua vez, é a seguinte:

$$V_{th}(T) = V_{th}(T_0) - \alpha_{Vth}(T - T_0) \quad (4.13)$$

Onde α_{Vth} é o coeficiente térmico de variação de V_{th} com a temperatura.

Com as equações (4.14) e (4.15), obtém-se:

$$\frac{\partial V_{GS}}{\partial T} = \sqrt{\frac{2I_{D0}}{\mu_n(T)C'_{ox}}\frac{L}{W}}\frac{T^{\alpha\mu-1}}{T_0^{\alpha\mu}} + \alpha_{vt} \quad (4.16)$$

Substituindo-se (4.16), (4.15) e (4.14) em (4.13) :

$$V_{GS0} = V_{th}(T_0) - \alpha_{Vth}\left(1 + \frac{2}{\alpha\mu}\right)T + \alpha_{Vth}T_0 - \frac{2}{\alpha\mu}\left[\sqrt{\frac{2I_{D0}}{\mu_n(T)C'_{ox}}\frac{L}{W}}\frac{T^{\alpha\mu}}{T_0^{\alpha\mu}} + \alpha_{vt}T\right] \quad (4.17)$$

Quando o coeficiente de mobilidade térmica $\alpha\mu$ é –2 (que é um parâmetro dependente do processo), um ponto de polarização V_{GS}-I_D independente da temperatura pode ser encontrado. Assim, o transistor MOS conectado como diodo (isto é, com seu *gate* conectado ao dreno), sob um ponto particular de polarização, pode constituir-se em uma referência de tensão e ao mesmo tempo uma referência de corrente [23]. Todavia, esta técnica é aplicável somente a tecnologias que tenham coeficiente $\alpha\mu$ igual a –2.

Quando $\alpha\mu$ é maior que –2, a derivada de V_{GS} com respeito a temperatura tanto pode ser negativa quanto positiva, dependendo do valor de I_D. Escolhendo-se um valor apropriado do nível de corrente de dreno, I_{D0}, os valores de V_{GS0} necessários para a obtenção de corrente constante em temperatura segue uma função CTAT (*Complementary To the Absolute Temperature*).

No caso de um transistor FGMOS com dois *control gates*, fig.4.7, com sua fonte conectada ao terra, sua tensão gate-fonte é dada por:

$$V_{GS} = (K_1V_{GA} + K_2V_{GB}) = K(V_{GA} + V_{GB}) \quad (4.18)$$

Em que K é o coeficiente de acoplamento capacitivo dos *control gates*. Neste caso $K=K_1=K_2$.

Com a utilização do FGMOS em substituição a um MOS ordinário, o Gate-A pode ser conectado ao dreno do transistor, configuração de diodo, e o gate B, por sua vez, atuará como um gate de controle, como na fig.5.1 abaixo:

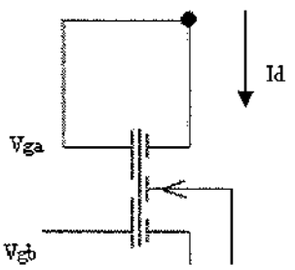

Fig.4.7) Configuração Proposta do FGMOS

Em tais condições de polarização, a aplicação de uma tensão V_{GB} CTAT com coeficiente térmico apropriado e uma corrente constante I_{D0} força V_{GA} manter-se praticamente constante com a temperatura.

$$V_{GB} = V_{GB0} - \alpha_G T \quad (4.19)$$

A eq.(8) é a expressão da tensão CTAT que deve ser aplicada ao gate-B. Neste caso, V_{GS} iguala-se a V_{GS0}. Nestas condições é valida a seguinte relação:

$$V_{GS} = V_{GS0} = K\left(V_{GA} + V_{GB0} - \alpha_G T\right) \quad (4.20)$$

Considerando-se uma junção PN diretamente polarizada e que portanto produza uma tensão CTAT com coeficiente térmico α_{VD}, uma fração desta tensão, $G = \alpha_{VD}/\alpha_G$, pode ser aplicada ao gate-B para produzir uma tensão independente da temperatura no dreno do transistor FGMOS.

Os resultados experimentais são apresentados ao final do capítulo de resultados, capítulo 5.

Capítulo 5

Resultados do Estudo

de Dispositivos e Memórias *Floating Gate*

De posse dos chips confeccionados, dispositivos FGMOS puderam ser estudados e propriedades interessantes e úteis puderam ser observadas. A possibilidade de programação e apagamento utilizando-se somente de pulsos de tensão positivos, sem a necessidade de pulsos negativos, é uma destas propriedades. Além disso, pôde-se perceber que, com a aplicação sucessiva de pulsos, o Vt programado seguia uma curva exponencial mostrando uma característica essencial ao projeto de circuitos de programação de memórias analógicas. Quanto maior a amplitude do pulso aplicada, maior faixa dinâmica pôde ser alcançada e em menos tempo. Em contrapartida, sob pulsos de menor amplitude e larguras de pulso controladas, maior precisão pôde ser alcançada ao estabelecimento de um determinado valor de tensão de limiar almejado. A propriedade mais significativa verificada nos experimentos, todavia, foi identificada como a possibilidade de armazenamento não-volátil de cargas em estruturas de porta flutuante (*floating gate*) em tecnologia CMOS digital, objetivo primeiro deste trabalho.

As figuras 5.1 e 5.2 seguintes são fotografias dos chips e estruturas FGMOS realizadas. Na realização destas, para o *layout* dos *floating gates*, houve detecção de erro de DRC (*Design Rule Check*): *Floating Gate Error*. Obviamente, tal "erro" era esperado, pois a presença dos *floating gates* era justamente o objeto do estudo.

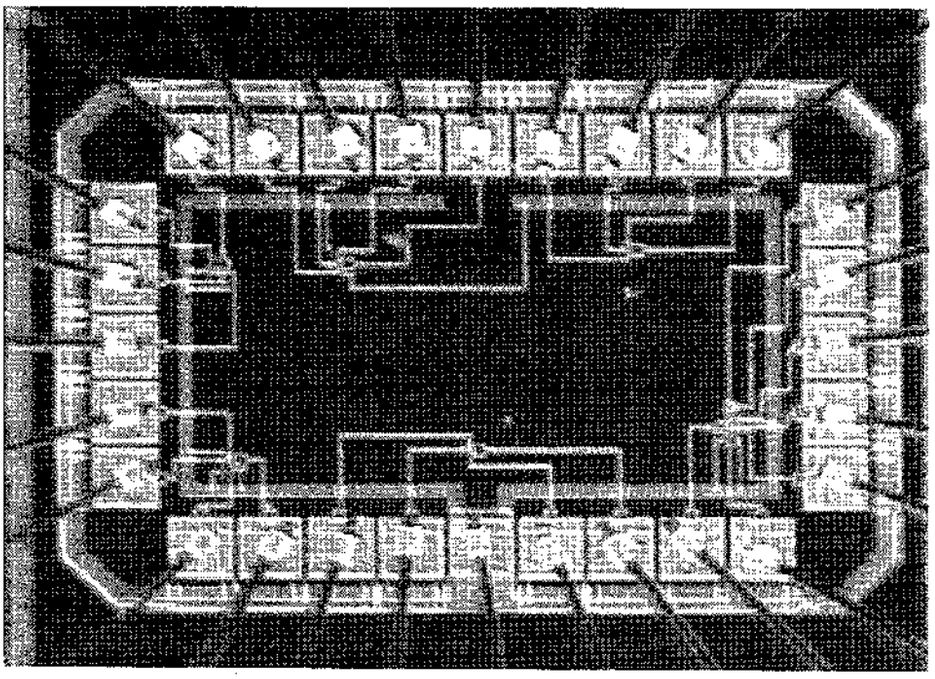

Fig.5.1) Aspecto geral do chip floatcye

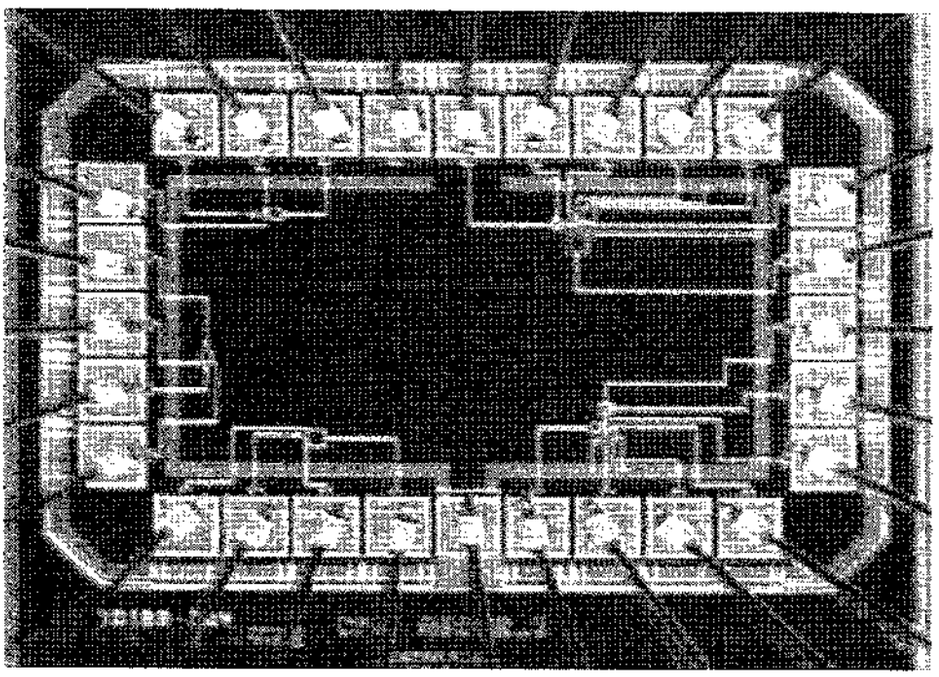

Fig.5.2) Aspecto geral do chip floatcye2

As fotografias anteriores (fig.5.1 e fig.5.2) apresentam a estrutura interna dos chips e foram obtidas com apoio do Cenpra (Centro de Pesquisas Renato Archer).

As figuras 5.3 e 5.4 apresentam detalhes das estruturas realizadas.

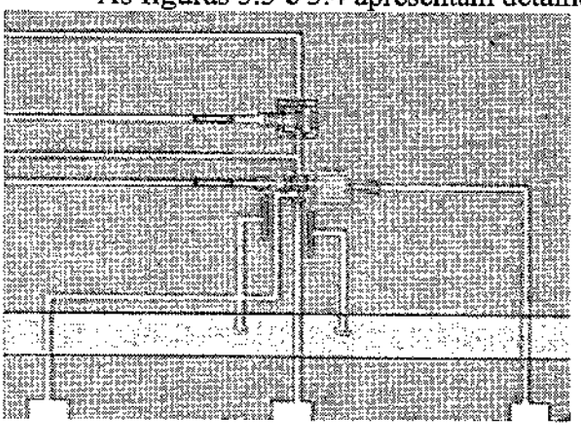

Fig.5.3)Detalhe de pFGMOS

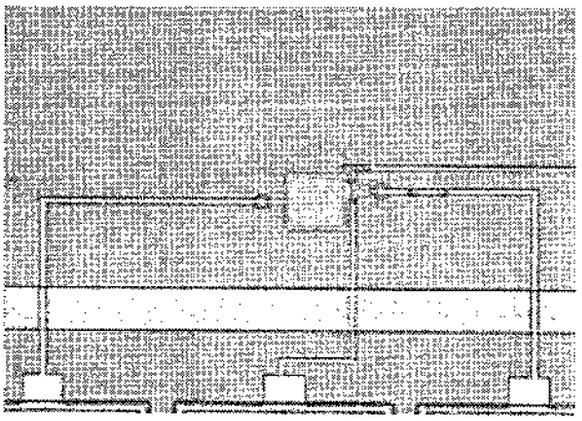

Fig.5.4)Detalhe de nFGMOS

No chip floatcye2, fig.5.2, quatro diferentes FGMOS estruturas foram estudadas: est1, est2, est3 e est6. No chip floatcye, fig.5.1, três estruturas foram estudadas, est1, est2 e est3. Suas dimensões e características são apresentadas na Tabela I.

Tabela I – Dimensões e acoplamentos capacitivos dos dispositivos

Chip	CYE - 0,8 um AMS			Acoplamentos		
floatcy2	Tipo	W(um)	L (um)	Ki	Kb	K0
est1	PMOS	7	2	0,062	0,613	0,325
est2	NMOS	2	7	0,062	0,613	0,325
est4	NMOS	12	2	0,051	0,497	0,452
est6	PMOS	2	7	0,062	0,613	0,325
floatcye	Tipo	W(um)	L (um)	Ki	Kb	K0
est1	NMOS	2	14	0,047	0,463	0,490
est2	NMOS	2	7	0,037	0,769	0,194
est5	NMOS	12	2	0,051	0,497	0,452

Os acoplamentos capacitivos Ki, Kb e K0 (descritos no capítulo 3 como parâmetros m) têm importância crucial nos processos de programação e apagamento das estruturas. Ki é o acoplamento do menor dos capacitores da estrutura, Ci, capacitor Injetor. Kb é relativo ao maior capacitor, capacitor *Bootstrap*. K0, relaciona-se ao capacitor da área ativa do MOSFET. Tais acoplamentos são obtidos a partir da razão entre a capacitância do *control gate* para o *floating gate* pelo somatório das capacitâncias de todos os outros *control gates* mais a capacitância de canal.

5.1.Aparato de caracterização

Para a realização das medidas e da programação dispôs-se do seguinte aparato: HP 4155 (analisador de parâmetros) provido de gerador de pulsos, um PC, uma Placa HPIB, matriz de comutação, câmara térmica para ensaios de tempo de retenção.

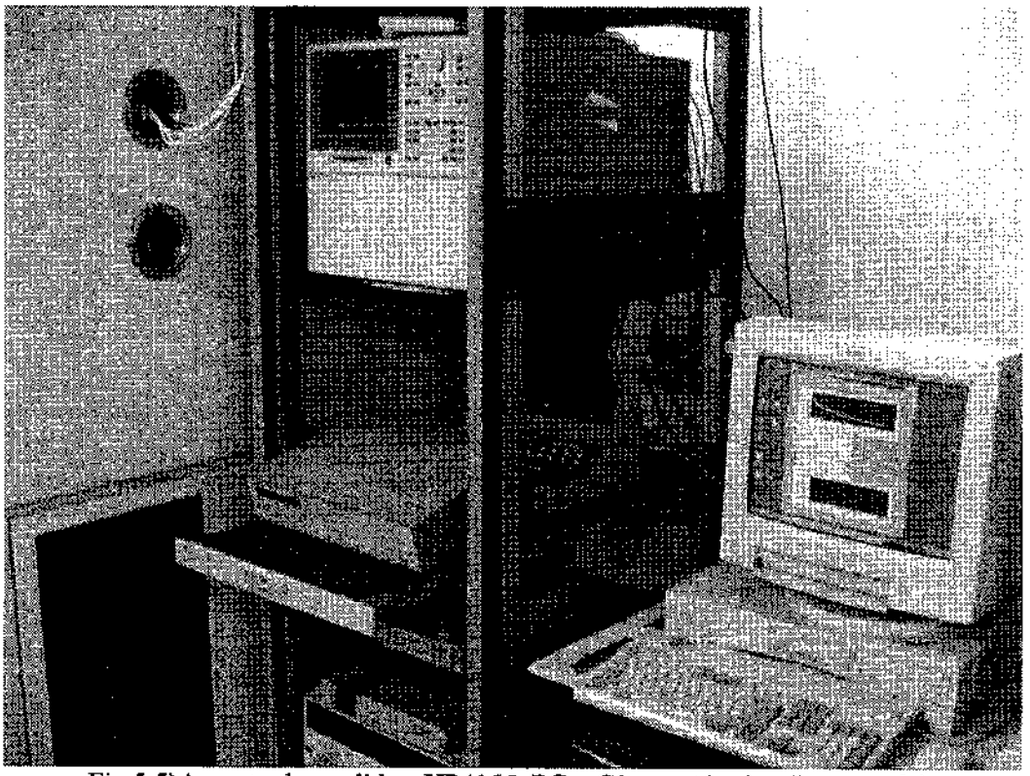

Fig.5.5)Aparato de medidas: HP4155, PC e Câmara térmica (à esquerda).

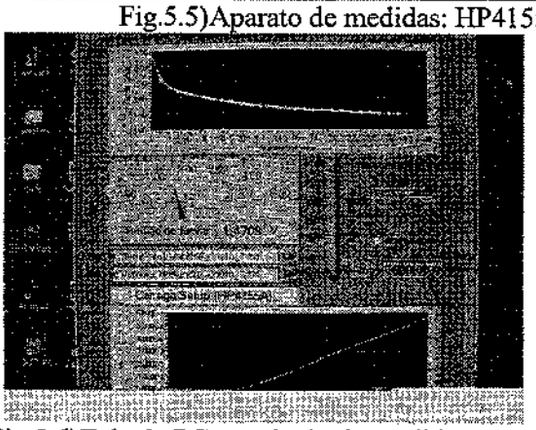

Fig.5.6)Tela do PC: resultado de medidas
Controle realizado com Labview-HPIB

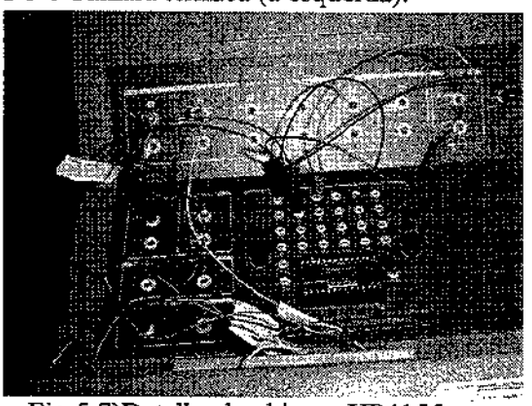

Fig.5.7)Detalhe do chip no HP4155

5.2.Extração de Vt

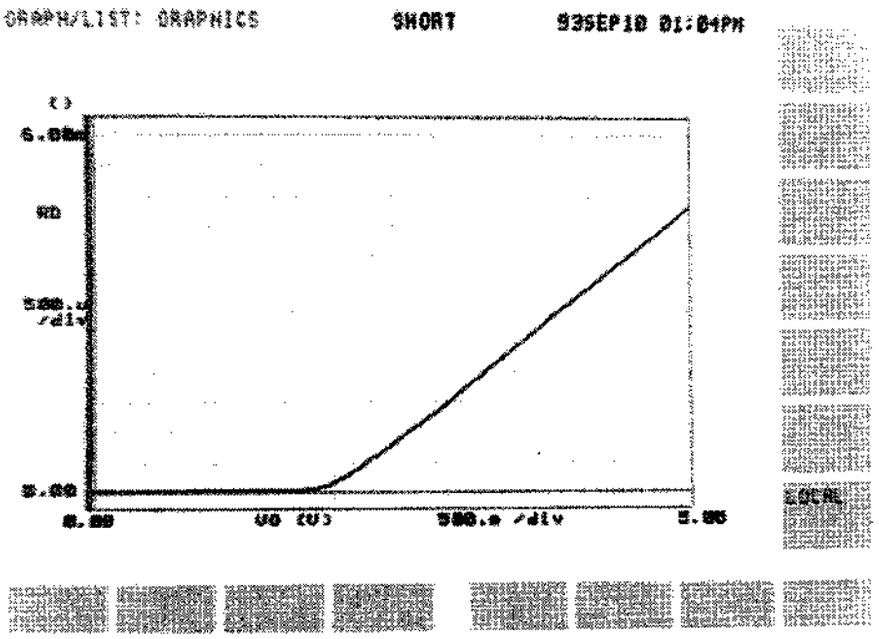

Fig.5.8) Tela do HP4155 e curva sqrt(Id) x Vgs (foto negativa)

A extração dos valores de tensão de limiar a partir das medidas realizadas seguiu o procedimento da segunda derivada descrito na referência [32]. Este método consiste na obtenção da segunda derivada da corrente de dreno, I_D, em função da tensão *gate*-fonte, V_{GS}. A tensão de limiar, V_T, é obtida no ponto onde a segunda derivada tem valor máximo. A fig.5.8 apresenta resultado de medida corrente de dreno versus tensão *gate*-fonte da qual é extraído o valor da tensão de limiar.

5.3.Resultados de programação Fowler-Nordheim de estruturas de Memória

5.3.1.Procedimento Experimental:

Através da aplicação de pulsos de tensão aos *control gates* das estruturas FG, introduzem-se cargas à estrutura (polissilício-1/óxido/polissilício-2). As amplitudes dos pulsos de tensão variam entre ±11 e ±15 V. A eficiência de programação aumenta à medida em que se aumentam as amplitudes dos pulsos: menores tempos de programação e maior

faixa dinâmica de memória constituem-se na métrica de tal eficiência. Por outro lado, ajustes finos podem ser obtidos com a aplicação de pulsos de amplitudes menores e variação da respectiva largura de pulso. A fig.5.9 mostra que a faixa dinâmica é tanto maior quanto maior a amplitude usado para a programação, isto é, maior é a faixa de tensões que podem ser programadas. Sob amplitudes de tensão maiores, maior é a quantidade de cargas que pode ser acumulada no *floating gate* e, conseqüentemente, maior é a faixa de programabilidade.

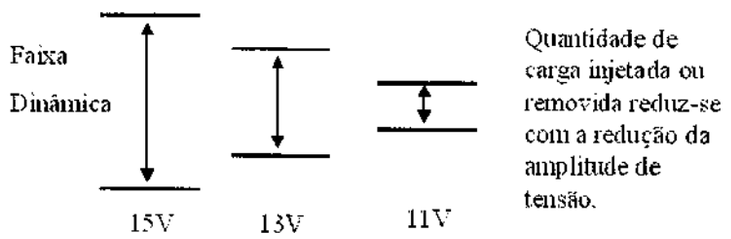

Fig.5.9) Faixa dinâmica ⇒ Faixa de programabilidade dos valores de Vt (tensão de limiar).

Pulsos de amplitudes negativas removem cargas da estrutura (elétrons aprisionados no *floating gate*). O resultado é a redução do Vt efetivo do FGMOS. Assim, o procedimento adotado consistiu na aplicação de pulsos de diversas amplitudes, positivas e negativas, aos *control gates* da estrutura FGMOS. Os dados obtidos revelam total concordância com a teoria e literatura (vide bibliografia).

5.3.2.Resultados do estudo de Escrita em memórias *floating gate*

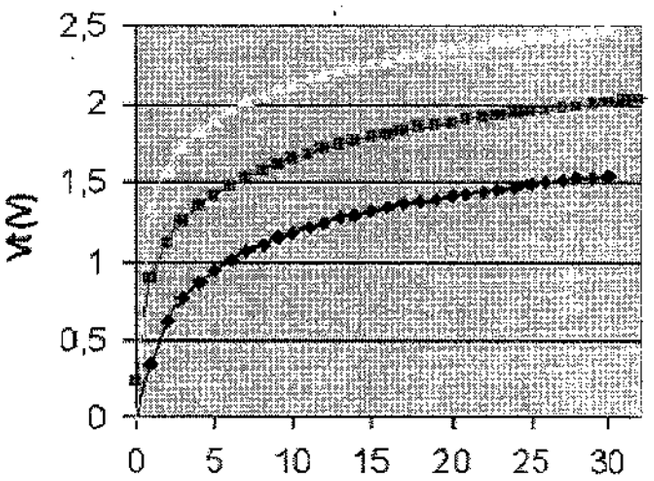

Fig.5.10.) Resultado de Programação: Vt (V) x Número de Pulsos. Curva com marcas triangulares, pulsos de +13V. Curva com quadrados pulsos de +12,5V e curvas com losangos, pulsos de +12V (W=2um, L=14um).

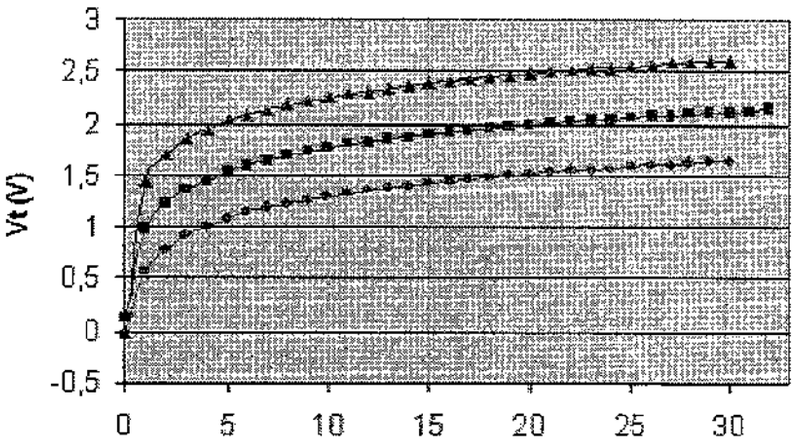

Fig.5.11) Resultado de Programação: Vt (V) x Número de Pulsos. Curva superior (triângulos), pulsos de +13V, curva do meio (quadrados), pulsos de +12,5V e curva inferior (círculos), pulsos de +12V ; W=2um, L=7um.

Às estruturas *est1* e *est2* (chip floatcye2, vide tabela I) foram aplicados pulsos de 12, 12,5 e 13V. As curvas das figs.5.10 e 5.11 revelam um aspecto importante do processo de programação: inicialmente, para os primeiros pulsos, a quantidade de cargas injetadas é bem maior que a quantidade injetada em pulsos posteriores até o ponto em que a o processo de programação (ou apagamento) cessa. Neste ponto, a quantidade de cargas aprisionadas no dielétrico entre o *floating gate* e o *control gate* produz um campo elétrico suficientemente forte que se opõe ao pulso elétrico externamente aplicado.

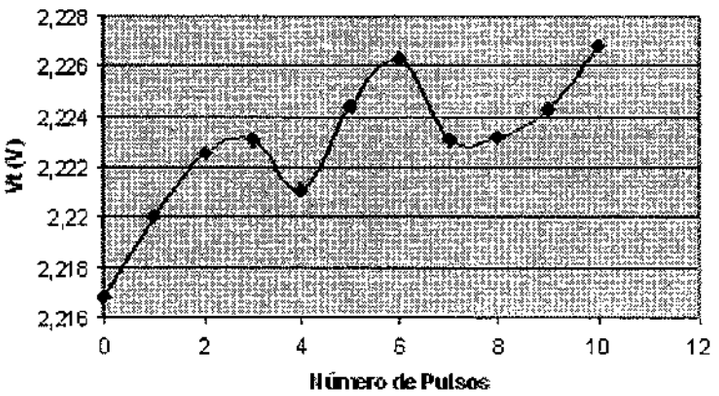

Fig.5.12) Ilustração do efeito auto-limitante *(obtido com +12V de pulsos)*

Isto é conhecido como efeito **auto-limitante** (*self-limiting effect*) [9] [1], fig.5.12. A partir deste ponto, caso se queira prosseguir com a programação, será necessário aumentar a amplitude do pulso externo aplicado, de forma que a diferença entre este e o campo interno da estrutura esteja acima do valor mínimo de tensão de limiar de tunelamento (±11V). (vide capítulo 2). A fig.5.12 mostra a ineficácia de pulsos externos aplicados quando o FGMOS já se encontra saturado de cargas: o Vt é muito pouco alterado e pode até mesmo ter seu valor variado positivamente ou negativamente, porém não de forma significativa.

Tabela II – Faixa Dinâmica x Amplitude de Pulsos, est1 floatvce2[4]

Amplitude Pulsos (V)	Estrutura 1 – Faixa Dinâmica			
	Vt inicial (V)	Vt final (V)	No de Pulsos	Faixa Dinâmica
12	-0,0246	1,5468	30	1,5714 V
12,5	0,2338	2,0153	30	1,7815 V
13	-0,0188	2,4891	30	2,5079 V

[4] est1floatcye2 - estrutura constante na Tabela I

Tabela III – Faixa Dinâmica x Amplitude de Pulsos, est2 floatcye2[5]

Amplitude Pulsos (V)	Estrutura 2 – Faixa Dinâmica			
	Vt inicial (V)	Vt final (V)	No de Pulsos	Faixa Dinâmica
12	-0,0197	1,6695	30	1,6892 V
12,5	0,1165	2,1293	30	2,0128 V
13	-0,0186	2,6205	30	2,6391 V

Segundo as tabelas I e II e a fig.8.9 , observa-se que a faixa dinâmica de programação é tanto maior quanto maior é a tensão do pulso de programação : além da programação ser mais rápida, os intervalos entre valores máximos e mínimos de tensões de limiar, Vt, são maiores.

5.3.3.Resultados do estudo de Apagamento em memórias *floating gate*

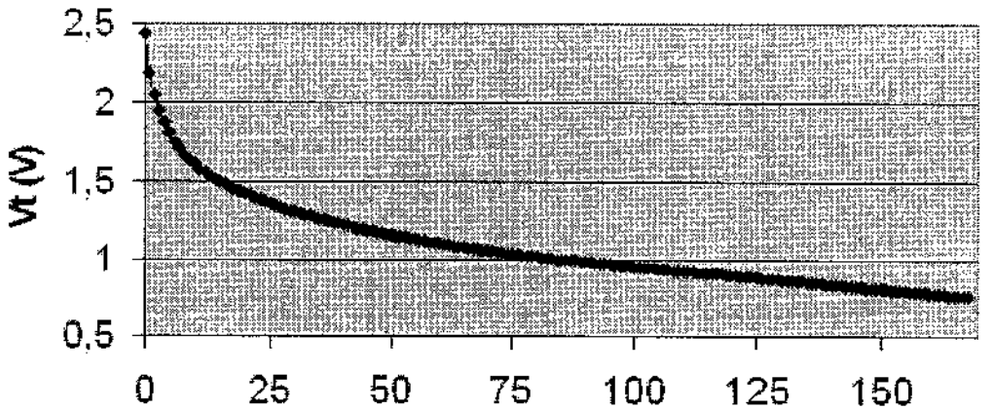

Fig.5.13) Característica do processo de apagamento, Est1, chip floatcye (tabela I). Vt(V) x Número de pulsos (-13 V)

O processo de tunelamento segue o mesmo comportamento do de escrita: inicialmente para os primeiros pulsos a eficiência do processo é maior. Para pulsos sucessivos, o processo vai sendo limitado pela relação de campo elétrico que se estabelece através das barreiras de potencial dielétricas até o momento em que cessa totalmente, cf. fig.5.13.

[5] est2 floatcye2 - estrutura constante na Tabela I

5.3.4.Utilização de pulsos unipolares para programação e apagamento de dados

Utilizando-se de estruturas com duplo *control gate*, onde um atua como *gate* de programação e o outro como *gate* de apagamentos, pode-se prescindir de pulsos bipolares de programação, vide capítulo 2. Aplicando-se pulsos de tensão ao *gate* de maior área (capacitância), obteve-se a programação da memória, ou seja, incremento em Vt. Contrariamente, ao se aplicar pulsos de tensão ao *control gate* menor, produziu-se a redução do valor de Vt.

A fig.5.14 apresenta o resultado da programação na estrutura Est1 do chip floatcye2 (vide Tabela I). Esta estrutura possui um transistor com largura W=7um , comprimento de canal de L=2um e acoplamentos Ki=0.062 (gate menor) e Kb=0.613 (gate maior). Um pulso de tensão (rise time = 1s e largura de 5s) com amplitude de 13V aplicado ao gate maior resultou no apagamento da memória (diminuição da tensão de limiar, V_t). Enquanto que aplicação de um pulso idêntico ao gate maior, resultou em escrita na memória.

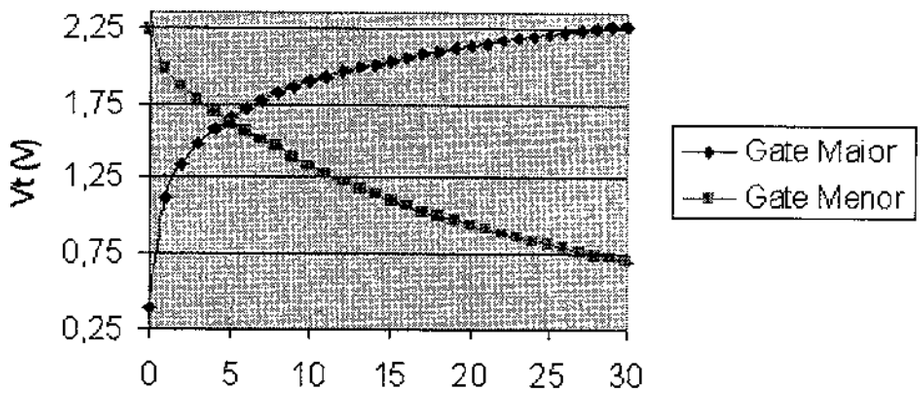

Fig.5.14) Programação e apagamento unipolar: curvas de tensão de limiar resultantes de pulso de +13V aplicados ao *gate* maior e ao *gate* menor. Estrutura Est1, chip floatcye2 (tabela I)

Resultados bem semelhantes ao anterior podem ser observados na fig.5.15. A estrutura Est2 do chip floatcye possui um transistor com largura W=2um , comprimento de canal de L=7um e acoplamentos Ki=0.062 (gate menor) e Kb=0.613 (gate maior).

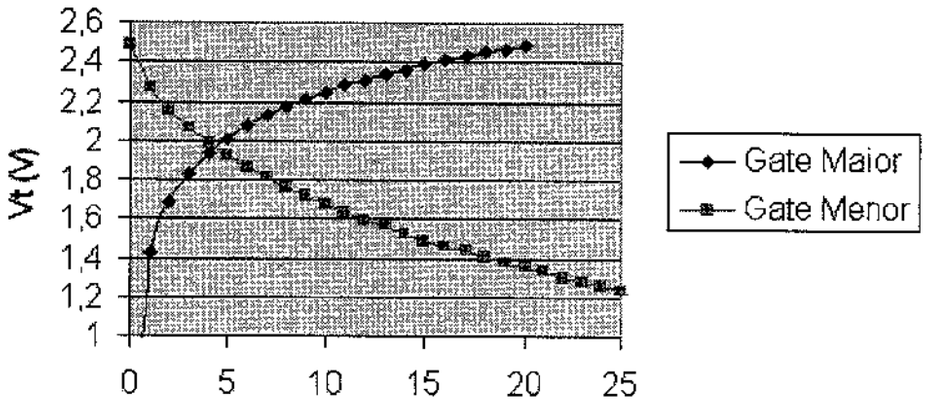

Fig.5.15) Programação e apagamento unipolar: Curvas de tensão de limiar resultantes depulso de +13V aplicados ao *gate* maior e ao *gate* menor. Estrutura Est2, floatcye2 (tabela I).

5.4.Estudo de retenção de carga

Os valores típicos de tensão de limiar (Vt, *Threshold*) de transistores NMOS e PMOS são 0,72V e 0,78V, respectivamente (AMS, 0,8um CYE). Contudo, em estruturas FGMOS o Vt relativo ao *control gate* é maior, visto que se faz necessária a aplicação de tensão superior àquelas, posto que há queda de tensão sobre o dielétrico entre o *control* e o *floating gate*. Vide tabela abaixo:

Tabela IV – Valores de Vt para dispositivos MOS e FGMOS

Chip A	Tipo	Vt(in)	Vt(ex)
cyeest1	PMOS	0,78	1,39
cyeest2	NMOS	0,72	1,28
cyeest4	NMOS	0,72	1,33
cyeest5	PMOS	0,78	1,39
cyeest6	PMOS	0,78	1,39
Chip B	Tipo	Vt(in)	Vt(ex)
cye2est1	NMOS	0,72	1,34
cye2est2	NMOS	0,72	1,28
cye2est5	NMOS	0,72	1,33

Os valores de Vt (in) são valores típicos dos dispositivos NMOS e PMOS, enquanto os valores Vt(ex) são os novos valores de Vt para os FGMOS (considerando-se que nenhuma carga tenha sido injetada sob nenhuma forma, carga nula).

Depois do processamento dos *wafers*, entretanto, os valores dos recém confeccionados FGMOS podem ser diferentes dos previamente esperados (Vt(ex)). Durante

o processamento, cargas e/ou defeitos podem ser introduzidos na estrutura FGMOS, alterando o valor efetivo de Vt(ex).

Para o estudo de capacidade de retenção de cargas nos FGMOS, algumas amostras foram submetidas à temperatura de 100°C por período de 5 horas. Recozimentos de estruturas como memórias aceleram os processos de degradação do óxido, constituindo assim uma técnica de estudo de longevidade de estruturas e, neste caso específico, de volatilidade de cargas em estruturas FGMOS. Sob tal condição, houve boa retenção de cargas, conforme resultados a seguir.

5.4.1.Procedimento experimental:

Inicialmente procedeu-se à obtenção dos valores de Vt (tensão de limiar) de estruturas. Em seguida, algumas amostras, tiveram seus valores de Vt alterados por pulsos de programação. Na etapa final, depois de já terem sido inseridas em câmara térmica por 5 horas a 100 °C, os valores de Vt foram novamente medidos para comparação com os valores previamente medidos.

Os dispositivos nFGMOS apresentaram boa retenção de carga sob tratamento térmico de 100 °C durante 5 horas. Muitas das amostras perderam pouquíssima quantidade de carga e outras, quantidade moderada. Outras, no entanto, perderam toda a carga que tinham. Tal fato, contudo, não significa, no caso das estruturas desta tese, que as estruturas não sejam apropriadas para o armazenamento não volátil de cargas. A perda de cargas, em alguns ciclos de programação e/ou de escrita, foi observada quando por algum motivo o equipamento HP 4155 recalibrava-se automaticamente ou quando se extraia o *chip* do soquete de testes. Tais estruturas, para que se viabilizasse a programação em tensões superiores a 11V, tiveram os diodos e resistores de proteção extraídos dos *pads*, ficando portanto susceptíveis a descargas eletrostáticas e conseqüente perda de carga. Num circuito onde estruturas como estas estivessem embarcadas com fontes internas de alta tensão (*charge-pumps*) tal fenômeno seria menos provável de ocorrer, posto que não haveria necessidade de retirar a proteção oferecida pelos diodos e resistores dos *pads*. Outra possível fonte de erro entre os dados antes e após a programação é o passo de amostragem de correntes e tensões (50 mV) do equipamento de medida (HP4155).

Vt1- Tensão de Limiar do dispositivo "virgem" (vindo da *foundry*)

Vt2- Tensão de Limiar do dispositivo após pulso de programação

Vt3- Tensão de Limiar do dispositivo após tratamento térmico para estudo de retenção.

Tabela V – Retenção de nFGMOS - floatcye

		floatcye						
		amostra 1	amostra 2	amostra 3	amostra 4	amostra 5	amostra 6	amostra 7
est2	Vt1	4,00	-0,25	1,65	4,77	-0,03	-0,03	-0,03
NMOS	Vt2	(sem pulso)	3,23	3,18	0,73	3,25	3,27	3,25
W=2u, L=7u	Vt3	4,00	2,91	3,16	0,88	3,00	3,12	2,60
est4	Vt1	4,50	-0,03	4,51	3,00	4,84	3,99	4,61
NMOS	Vt2	0,68	(sem pulso)	0,67	(sem pulso)	0,56	4,02	0,59
W=2u, L=12u	Vt3	0,84	-0,02	0,84	2,99	0,74	4,01	0,73

A tabela V apresenta dados de teste de retenção realizados com dispositivos nFGMOS do chip floatcye. Da diferença entre as tensões Vt2 (ou Vt1 para os casos onde não foi aplicado pulso de tensão) e Vt3, depreendem-se aspectos de conservação de cargas na estrutura FGMOS. Para a maioria dos casos, Vt3 e Vt2 permaneceram bem próximos.

Tabela VI – Retenção de nFGMOS – floatcye2

		floatcye2	
		amostra 1	amostra 2
est1	Vt1	0,88	0,75
NMOS	Vt2	2,92	2,90
W=2u, L=14u	Vt3	2,87	2,99
est2	Vt1	3,52	1,71
NMOS	Vt2	(sem pulso)	3,35
W=2u, L=7u	Vt3	3,50	3,27
est5	Vt1	3,44	3,67
NMOS	Vt2	(sem pulso)	3,78
W==12u, L=2u	Vt3	0,21	3,73

Na tabela VI observa-se que a amostra1 (est5) perdeu completamente a carga após o tratamento térmico, exemplo de possível fuga de carga provocada pela extração e ou colocação do chip no equipamento para medida de sua tensão de limiar.

Tabela VII – Retenção de p-FGMOS – floatcye2

		floatcye1			
		amostra 1	amostra 2	amostra3	amostra4
est1	Vt1			-0,1927	-0,1581
PMOS	Vt2	-5,0246	-5,0247	-5,0248	-4,5845
W=7u, L=2u	Vt3	-5,0241	-5,0243	-5,0238	-4,3206
est5	Vt1	-1,6504	-0,4303	-0,4303	-0,3494
PMOS	Vt2		-0,9366	-0,9366	-1,1098
W=2u, L=7u	Vt3	-1,6518	-0,9215	-0,9215	-1,0949

A tabela VII apresenta resultados de estudo de retenção para dispositivos p-FGMOS. Para estes dispositivos houve menor propensão à perda de cargas e maior capacidade de retenção de cargas, as quais variaram muito menos que para os dispositivos NMOS.

5.5. Variação de largura de pulso

Evidentemente, ao se variar a largura dos pulsos de programação aplicados, a quantidade de cargas injetadas através do dielétrico também varia.

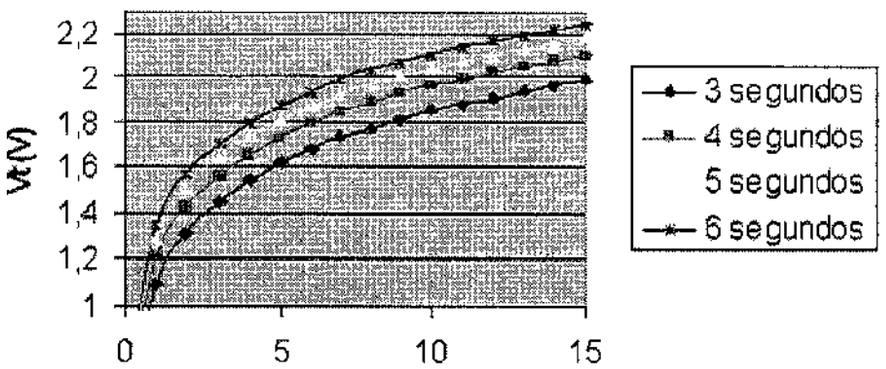

Fig.5.16) Característica de programação sob diferentes larguras de pulso :
Vt(V) x Número de pulsos (+13V)

Ao se aumentar a largura de pulso, a quantidade de pulsos aplicada para se alcançar um determinado valor de Vt é reduzida, fig.5.16, ao passo que a redução da largura de pulso pode ser utilizada para um ajuste mais preciso da quantidade de cargas a ser injetada.

BIBLIOTECA CENTRAL
CESAR LATTES
DESENVOLVIMENTO DE
COLEÇÃO
UNICAMP

5.6.Estudo de longevidade (*endurance*)

Idealmente, a programação e o apagamento dos valores de Vt, em diversos ciclos, deveria seguir uma mesma curva, isto é, partindo-se de um valor inicial Vto para se chegar a um valor final, a mesma quantidade de pulsos deveria ser aplicada. Contudo, em dispositivos reais, a quantidade de pulsos pode variar de ciclo para ciclo devido à degradação.

Procedendo-se a ciclos sucessivos de programação e apagamento das estruturas FGMOS sob teste, não se observou variação sensível da *endurance*.

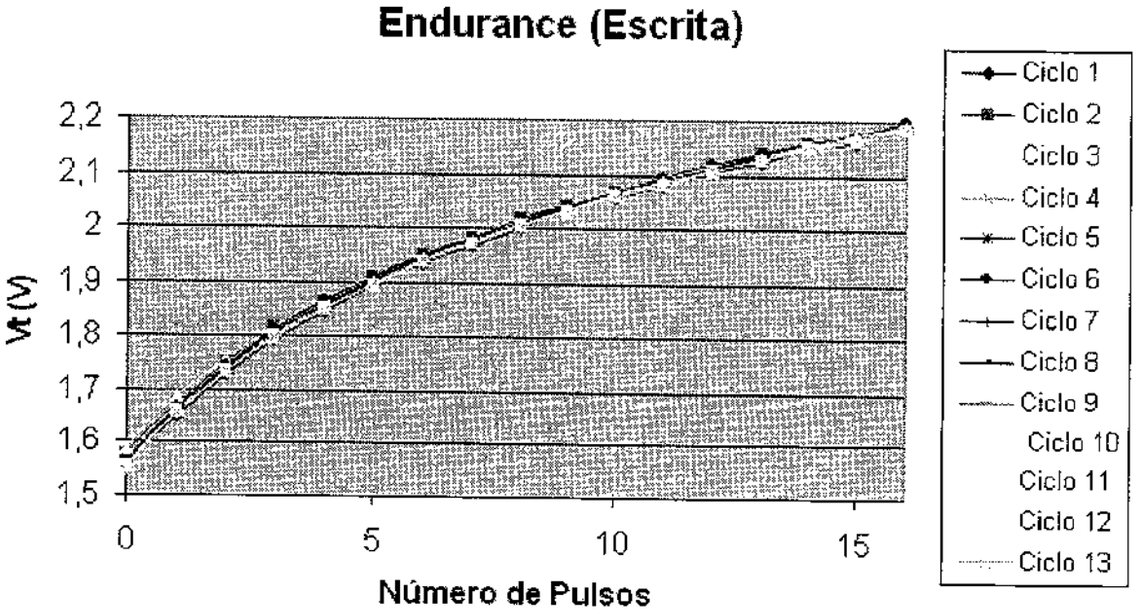

Fig.5.17) Curvas de programação em 13 ciclos, de endurance da escrita.

Para mais de uma dezena de ciclos de programação e apagamento, figuras 5.17 e 5.18, não se verificaram diferenças sensíveis de *endurance* (<1%). Era de se esperar que ao longo de vários ciclos de programação e apagamento se fizesse necessária a aplicação de mais pulsos para se programar um Vt que já fora obtido previamente em outro ciclo.

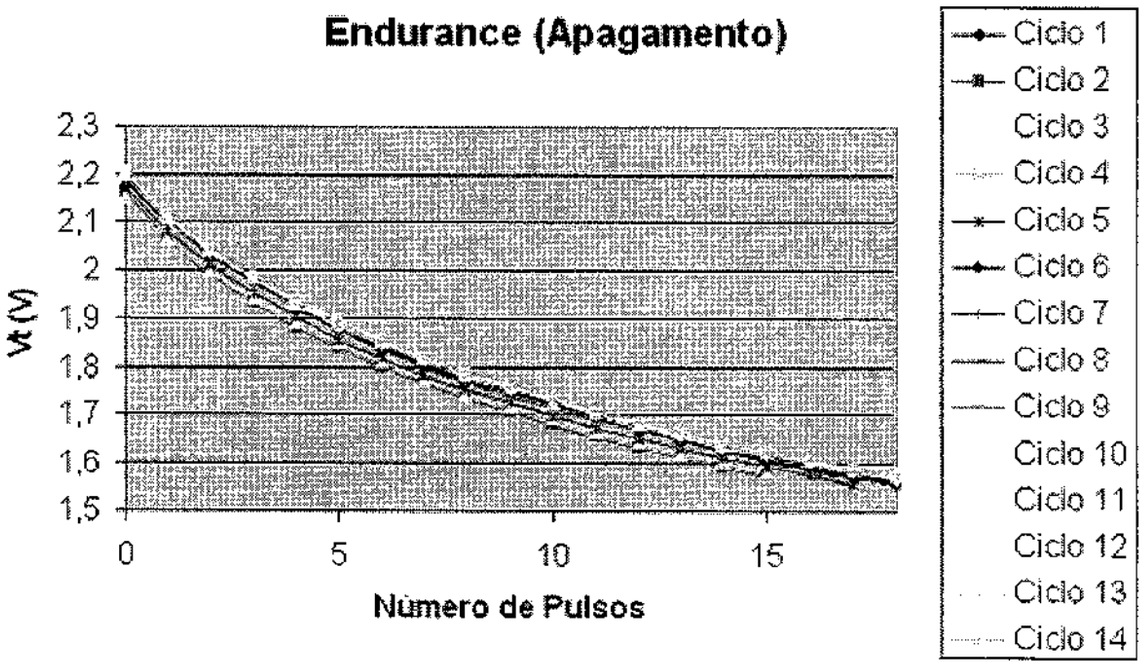

Fig.5.18) Curvas de apagamento em 13 ciclos, endurance do apagamento.

As pequenas diferenças observadas, no entanto, são decorrentes de imprecisão na tomada dos valores de Vt (precisão de casas decimais dos equipamentos) e por não se partir e chegar a valores iguais de Vt para todos os ciclos (seria muito demorado programar sempre os mesmos valores de Vt para todos os ciclos).

5.7. Trimming analógico

Quando da integração de memórias FGMOS, uma fonte externa de pulsos de tensão ou interna (*charge pump*) precisa ser utilizada para programação e apagamento das estruturas. Para tanto, dispositivos que suportem altas tensões devem ser utilizados.

Os dispositivos FGMOS têm a capacidade de variação dinâmica dos valores de suas tensões de limiar. Variações na tensão de limiar implicam necessariamente em variações na corrente de dreno do transistor, podendo esta, portanto, ser convenientemente ajustada, fig.5.19.

A tecnologia AMS 0,6 μm tem como espessura de óxido entre as camadas de polissilício-1 e polissilício-2, 40nm. Tal espessura oferece uma grande barreira de potencial para tunelamento de elétrons : é muito espessa (em tecnologia 0,8 μm, tal espessura é de 19nm). Diante desta contingência, para a implementação do *trimming* em tecnologia 0,6

μm, foram utilizados *control gates* diferentes : *control gates* formados entre o polissilício-1 e poço-N. Assim, o óxido de tunelamento passou a ter 16nm de espessura, necessitando deste modo de menor tensão para tunelamento (menor barreira de potencial dielétrica).

Na confecção deste *trimming* agregaram-se transistores LDD, os quais são dispositivos que suportam tensões maiores que os MOSFETs da tecnologia AMS. A introdução destes dispositivos é um passo precursor à inserção de *charge-pumps* ao circuito de programação da memória integrada. Os *charge-pumps* são circuitos que provêem tensões internas aos chips suficientemente altas para o processo de tunelamento. Para tanto, os transistores, que compõem os últimos estágios de circuitos como este, devem ser capazes de suportar, neste caso, tensões na região de dreno da ordem de ±15V. Os FGMOS eram do tipo p-FGMOS (melhor característica de retenção de carga).

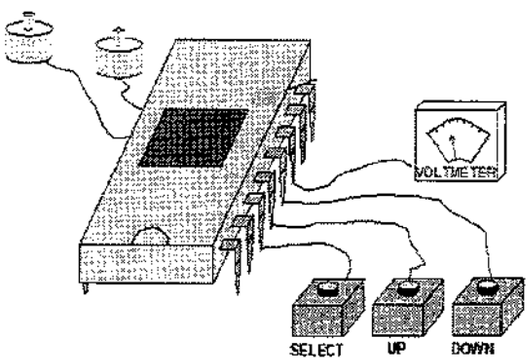

Fig.5.19)Esquema de trimming por endereçamento [9]

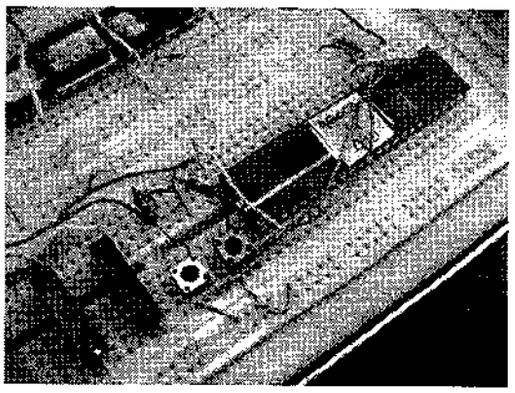

Fig.5.20)Protótipo de trimming realizado

De posse do chip, fig.5.21, montou-se o aparato de teste em *proto-board*. Dispôs-se de dois botões *push-buttons* : um atuava sobre a programação e o outro sobre o apagamento, ou seja, atuavam na variação dos valores da tensão de limiar. O comportamento de variação dos valores de tensão de limiar em função da quantidade e duração dos pulsos aplicados via *push-buttons* foi idêntico aos das outras estruturas FGMOS.

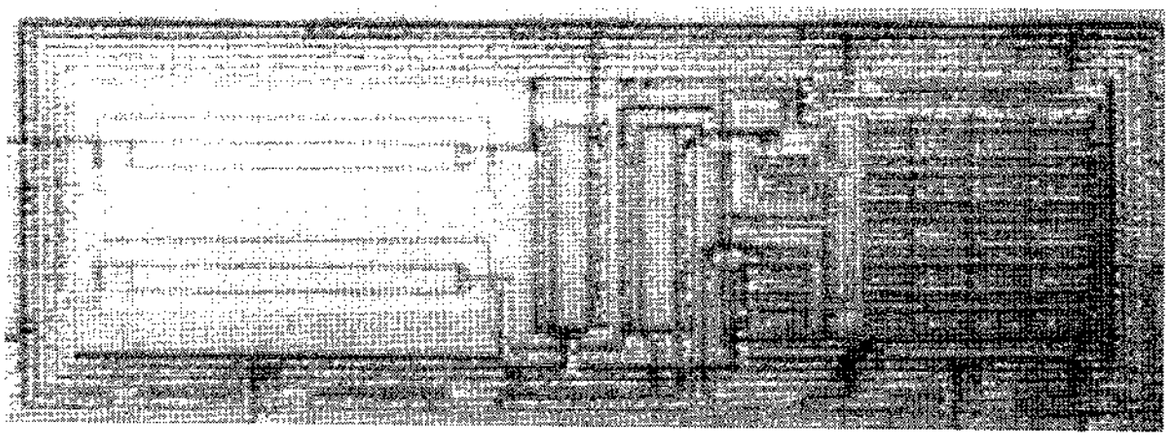

Fig.5.21) Fotomicrografia do triming em tecnologia 0,6 μm AMS

Este tipo de dispositivo é bem útil em circuitos onde se deseja variar os valores de corrente ou de Vt. Em circuitos mais complexos, alguns pinos do chip podem ser utilizados para endereçamento de células de memória que se deseja programa. Outros dois pinos, externos, podem ser utilizados, depois de endereçado o dispositivo desejado, serem utilizados para a programação da célula de memória.

5.8.Programação por elétrons quentes

Para programação através do processo de elétrons quentes o transistor FGMOS precisa estar polarizado na região de saturação. Assim, aplicou-se tensão de 8V ao dreno de dispositivos nFGMOS e pulsos de tensão foram aplicados ao *control gate*.

A eficiência do processo de programação por elétrons quentes é bem menor que a do processo de tunelamento, no entanto, por esse fato, o controle do ajuste da tensão de limiar é mais preciso e possui característica mais linear, fig.5.22.

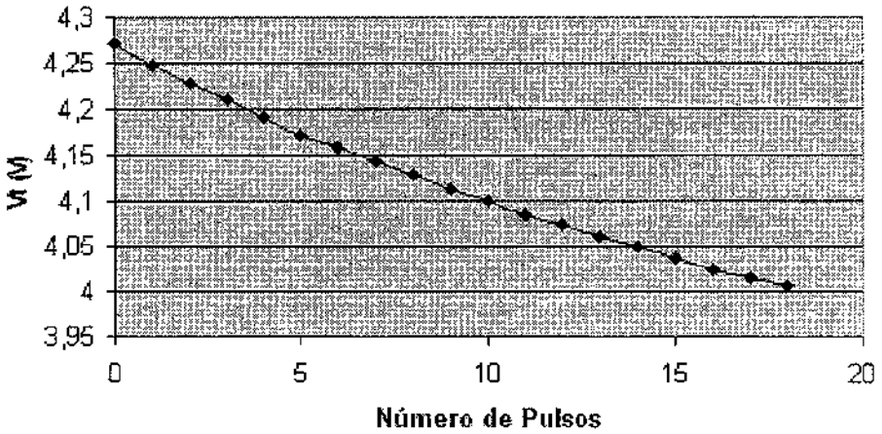

Fig.5.22) Variação de Vt em função do número de pulsos de programação por elétrons-quentes.

Neste processo de programação, para dispositivos nFGMOS, somente o apagamento é possível. Não é possível a escrita. O contrário é válido para pFGMOS.

5.9. Referência de tensão em Temperatura com dispositivo FGMOS

Ao final do capítulo 4 foi apresentada a teoria pertinente a uma referência de tensão constante com a variação de temperatura. Aqui são apresentados os resultados experimentais obtidos.

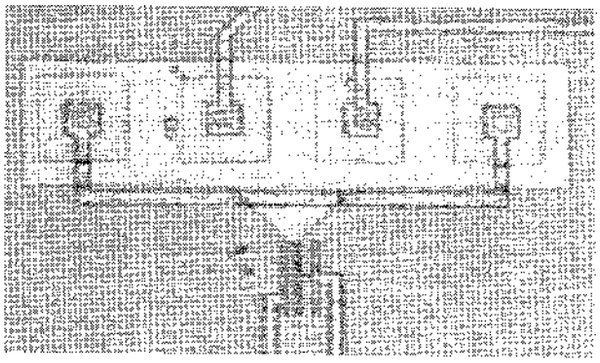

Fig.5.23) Fotografia do nFGMOS utilizado (AMS 0.6 μm).

5.9.1. Procedimento Experimental

Dispôs-se de uma câmara térmica dentro da qual o aparato da fig.5.24 foi montado. As fontes de corrente I1 e I2 provinham de circuito externo (não estavam submetidos a temperatura). Inicialmente, porém, somente o dispositivo FGMOS foi colocado dentro da câmara: a ele foi aplicada uma corrente I1 (corrente de dreno), o Gate-A foi conectado ao dreno do FGMOS e ao Gate B, para cada nova temperatura a que era submetido o dispositivo FGMOS, uma rampa de tensão de 0 a 5V era aplicada. Dessa maneira, em passos de 20 °C, e com a aplicação sempre constante da corrente I1 através do FGMOS, sua tensão de dreno era constantemente monitorada. Ao final do experimento, cuja temperatura variou na faixa de –40 a 120 °C, obtiveram-se 8 pontos de tensão que aplicados ao Gate-B mantinham a tensão de dreno sempre constante em 2.69V. Como previsto pela teoria precedente, a curva descrita pelos oito pontos era uma reta CTAT. O diodo D e os resistores R1 e R2 foram posteriormente adicionados ao circuito. A função destes novos dispositivos era a de reproduzir uma curva CTAT que serviria de entrada ao Gate-B do FGMOS de modo a manter a tensão de saída constante. A corrente I2 foi estabelecida através do diodo D como polarização.

5.10.Resultados Experimentais

O circuito da fig.5.24 foi implementado usando-se um transistor FGMOS canal N com W/L= 25μm/3μm e dois *control gates*.

A tensão aplicada ao gate-B na fig.5.24 é uma fração da tensão sobre o diodo D através de um divisor resistivo R1-R2. As correntes I1 e I2 são aplicadas através de fonte externa e têm valores constantes. Os coeficientes α_G and α_{VD} foram experimentalmente determinados e os resistores R1-R2 ajustando de forma a se obter os valores adequados de tensão CTAT:

$$\frac{R2}{(R1+R2)} \cong \alpha_G/\alpha_{VD} \quad (5.1)$$

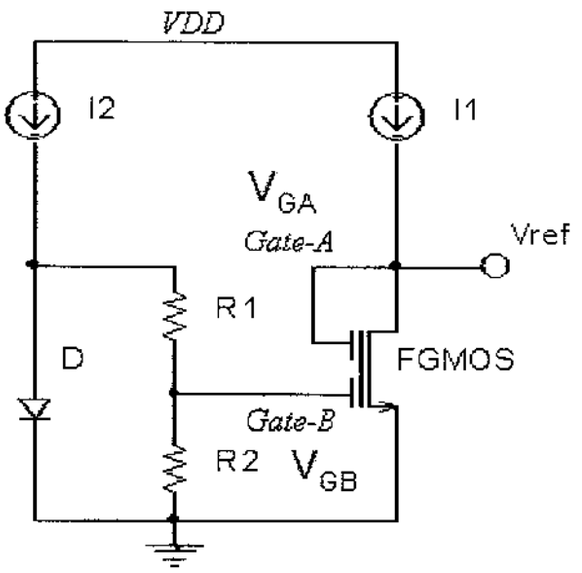

Fig.5.24) Circuito esquemático da referência de tensão FGMOS

As fontes de corrente I1 e I2,na fig.5.24, têm valores de corrente de 50μA.

V_{GA} e V_{GB} foram medidos para variações de temperatura entre –40 a 120 °C. A partir dos valores obtidos, V_{GS} pôde ser calculado através da relação (7) do capítulo 4, cujo resultado é apresentado na fig.5.25.

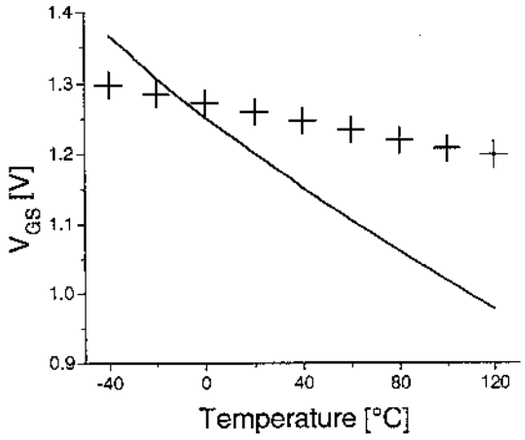

Fig.5.25) Valores de VGS x Temperatura.

Na fig.5.25 a linha contínua é o resultado dos valores calculados de V_{GS} através da relação (6) com os valores nominais dos parâmetros da tecnologia fornecidos pelo fabricante (AMS) [24]. Os valores de V_{GS} obtidos experimentalmente são os pontos em cruz na fig.3.

Entre os dados experimentais e teóricos há, entretanto, diferença, a qual pode ter resultado pela variação de alguns parâmetros fornecidos pelo fabricante e pelo fato de a variação de tensão na junção não corresponder exatamente à curva CTAT necessária à manutenção da tensão constante no dreno do FGMOS.

A tensão no dreno do FGMOS da fig.5.24 é o valor de saída do circuito, o qual é apresentado na fig.5.26 em função da temperatura. Na faixa de temperatura de –40 a 120°C o coeficiente de temperatura correspondente foi de 19 ppm/ °C.

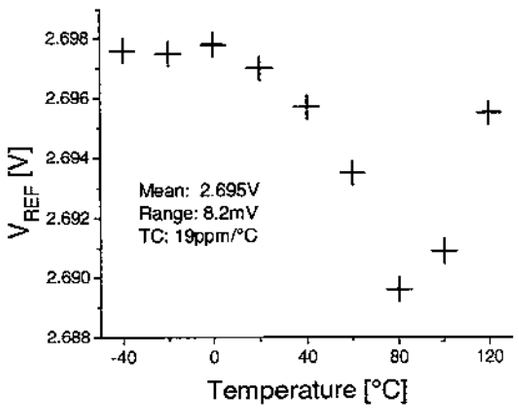

Fig.5.26) Tensão de Saída VREF x Temperatura (T).

Uma nova topologia de fonte de referência que se utiliza de transistor FGMOS com duplo *control gate* foi apresentada e descrita. Ao aplicar-se uma tensão CTAT a um dos *control gates* enquanto mantêm-se o outro *control gate* conectado ao dreno, a tensão de dreno, V_{REF}, mantém-se independente da temperatura. A literatura[6] não registra nenhuma aplicação de FGMOS para referências de tensão, cujo princípio de funcionamento seja igual a este.

[6] A empresa Xicor Inc. fabrica referência de tensão utilizando-se também de dispositivos floating gates. No entanto, o princípio de funcionamento baseia-se no armazenamento de cargas na região do *floating gate*. O dispositivo aqui apresentado possui quantidade de carga na região de óxido nula ou constante, sendo a tensão de referência mantida através de realimentação e não através de memorização de tensão através de cargas na região do *floating gate*.

Capítulo 6

Conclusão - Relevância do tema de tese

O objetivo do tema de tese proposto era saber se, de fato, era possível a implementação de memórias *floating gate* em tecnologia CMOS digital. E, sendo possível, estudar a viabilidade de confecção de memórias *floating gate* analógicas. O objetivo, portanto, exploratório, era o de aquisição de conhecimento em memórias analógicas *floating gate* pela equipe encabeçada pelo Prof. Carlos Reis.

Inicialmente, procedeu-se ao estudo bibliográfico. A vasta literatura sobre memórias concentrava-se sobretudo em memórias digitais e em tecnologias específicas para a confecção deste tipo de dispositivo, poucos artigos referiam-se diretamente a utilização de tecnologia CMOS digital para tal fim [9] [2] [25-26].

A simulação SPICE de tais dispositivos não era possível, posto que fenômenos de injeção de carga e tunelamento não são modelados neste tipo de simulador. Era preciso propor um modelo. No entanto, devido a carência de literatura a esse respeito e a complexidade dos fenômenos quânticos envolvidos nos processos de programação e operação dos dispositivos FGMOS, os quais estavam fora do escopo de Mestrado, resolveu-se implementar estruturas de teste simples: transistores nMOS e pMOS com *gates* flutuantes e dois *control gates*. Apesar de simples, havia riscos de comprometimento do funcionamento de tais dispositivos por dois motivos:

1. Violação da regra de DRC (*Design Rule Check*) de que os gates não podiam flutuar: necessariamente deveriam ter conexão elétrica a outro terminal. Durante as etapas de deposicão de camadas de dielétrico sobre o polissilício (*gate* flutuante) pode se armazenar muita carga na região de *floating gate*. A quantidade de cargas aprisionada pode ser suficiente para romper o dielétrico ou danificá-lo se o *gate* não puder drenar tais cargas para outro lugar.
2. Os diodos de proteção de ESD dos PADs foram removidos, pois precisaria-se de tensões superiores a 5V (>11V) para programação das estruturas. Dessa forma, imaginava-se que os dispositivos durante o processamento do

wafer ou durante o manuseio poderiam sofrer ruptura do dielétrico entre o *gate* e o substrato.

Assim, decidiu-se em violar uma regra de cada vez: a primeira regra violada foi, logicamente, a de *gates* flutuantes e os diodos de proteção foram mantidos na primeira rodada do projeto. Pretendia-se então estudar os FGMOS como dispositivos de dois *gates* (*control gates*) e não como dispositivos de memória. Como resultado desta primeira rodada, obteve-se o artigo apresentado no IMAPS: "*Temperature-Compensated Voltage Using Floating-Gate MOS Transistor*". Nesta aplicação, não havia necessidade de programação elétrica, o acoplamento capacitivo entre *control gate* e *floating gate* era o único responsável pelas características de tal dispositivo. Desse primeiro experimento descobriu-se que os FGMOS funcionavam. A tecnologia utilizada foi a AMS 0.6 μm, com espessura de dielétrico, entre o control gate e o *floating gate*, de 40 nm, ou seja, muito espessa para um ensaio exploratório de programação elétrica.

Procedeu-se ao redesenho de novos FGMOS sem os diodos e resistores anti-ESD dos PADs. Dessa forma seria possível aplicar pulsos elevados de tensão diretamente aos *control-gates* para a programação das estruturas FGMOS. A programação elétrica de memórias, em geral, em produtos comerciais, não é realizada através de pulsos externos, mas de pulsos de tensão internos gerados dentro do próprio circuito integrado através de *charge-pumps*. Todavia, a introdução de tais circuitos poderia mascarar características dos FGMOS, as quais eram o objetivo maior do estudo. As estruturas foram realizadas em tecnologia 0.8 μm com dielétrico entre *control gate* e *floating gate* de 19 nm de espessura, ou seja, fina o suficiente para um teste exploratório inicial de injeção de carga e a implementação dos dois tipos de programação dos FGMOS: unipolar e bipolar.

Mediu-se a tensão de limiar dos transistores recém chegados da *foundry*. Aterraram-se os terminais de dreno, fonte e substrato e aplicou-se uma tensão de 12 aos control gates e, novamente, mediu-se a tensão de limiar. Os FGMOS deram o primeiro sinal que funcionavam como dispositivos de memória, as tensões de limiar inicial e final eram diferentes. Aplicando-se mais pulsos verificou-se que a tensão de limiar seguia o comportamento logarítmico descrito na literatura [1] e que a partir de certa quantidade de pulsos aplicados aos control gates, o processo cessava devido ao campo elétrico criado

internamente ao FGMOS que se opunha ao campo elétrico externo aplicado, era o efeito auto-limitante sendo constatado.

A cada vez que pulsos de tensão eram aplicados ao *control gate* da estrutura, uma nova leitura do valor de tensão de limiar se fazia necessária. Tal processo era manual: tinha-se de trocar as conexões dos terminais do FGMOS para programá-lo: fonte, dreno e substrato eram aterrados. Durante a etapa de leitura, *gate* e dreno eram conectados juntos. Obtinha-se na tela do HP4155 a curva I_{DS} x V_{GS} , da qual, por extrapolação, calculava-se a tensão de limiar. Tal processo era cansativo e impreciso, e constantemente, por ESD, os FGMOS perdiam totalmente a carga injetada no *floating gate*. Era preciso automatizar o processo.

Utilizou-se o Labview para o controle de uma matriz de comutação de conexões dos terminais dos FGMOS da etapa de leitura de tensão de limiar para a de programação e vice-versa. Além disso, os dados, via interface HPIB, vinham do analisador de parâmetros HP4155 para um programa na memória de um computador PC. No computador os dados eram processados e a tensão de limiar era obtida por método numérico mais preciso que o da extrapolação: o método da segunda derivada [32].

De maneira mais rápida e precisa os resultados foram obtidos após a automatização do processo. Os resultados eram mais confiáveis e a quantidade de informação era maior e menos susceptível à falhas manuais.

Pôde-se variar a largura dos pulsos de tensão aplicados para programação e verificar que quanto mais largos eles eram, maior era a variação da tensão de limiar. A tensão de limiar variava até o ponto em que saturava: não era mais possível a programação da memória, pois o campo elétrico interno opunha-se ao externamento aplicado. A única maneira de continuar a programação era aumentado-se o valor do pulso de tensão, mas novamente a tensão de limiar, após certa quantidade pulsos, iria saturar. O pulso de tensão não podia ter seu valor indefinidamente aumentado: havia de se evitar a ruptura destrutiva dos dielétricos da estrutura.

Novos testes exploratórios podiam ainda ser realizados com as estruturas: os de tempo de retenção e os de durabilidade (*endurance*). Graças a boa infra-estrutura do Laboratório de Pesquisa LPM coordenado pelo Prof. Reis, pôde-se dispor de uma câmara

térmica para a realização de tais experimentos. Os experimentos revelaram que os FGMOS tinham boa capacidade de retenção dos dados e grande *endurance*.

Já no final do processo de caracterização dos FGMOS aproveitou-se a oportunidade e implementou-se um *trimming* analógico. Com ele, tensões internas ao circuito integrado poderiam ser ajustadas através de lógica externa. Neste experimento, os pulsos de tensão externos não eram aplicados diretamente aos *control gates*, mas a transistores de potência LDD (Low doped Drain). Tais LDD são os precursores da introdução de *charge-pumps* para geração de pulsos internos. Os LDD, em geral, situam-se nos últimos estágios dos *charge-pumps*.

Utilizando-se de tecnologia CMOS digital convencional, a integração de memórias analógicas mostrou-se viável. As estruturas de memória propostas, baseadas em transistores com porta flutuante (*floating gate*), podem ser programadas eletricamente em ambas as direções: escrita (positiva) ou apagamento (negativa). Esta característica confere versatilidade aos circuitos aos quais elas estejam embarcadas. Sua confecção em tecnologia CMOS convencional prescinde de processamentos especiais como filmes de óxido ultrafinos, camadas texturizadas de polissilício ou radiações UV.

Não obstante a gama potencial de aplicações para memórias analógicas, há grandes dificuldades impostas pelos circuitos adicionais usados para a programação: *Charge Pumps*, amplificadores de erro, PWM, chaves, etc, o que, de forma alguma, inviabiliza seu uso, posto que os circuitos adicionais são bem compreendidos e exaustivamente descritos na literatura.

Maior compreensão dos transistores com portas flutuantes (*floating gate* transistors), memórias analógicas, fenômenos físicos envolvidos na programação, cuidados de *layout*, estudo de viabilidade, tipos de memória, dimensões e de possibilidades de programação foram contemplados: estas são as principais contribuições aportadas por este trabalho a formação de seu autor e a comunidade acadêmica da qual fez parte durante o trabalho, transformando-o em uma boa referência para futuros estudos e artigos.

Futuros trabalhos neste tema podem abordar aspectos de integração de memórias *floating gates* a circuitos analógicos, geração interna de pulsos de tensão através de *charge pumps*, circuitos com memorização de tensão de *off-set*, entre outras.

Referências Bibliográficas

1. D. Kahng and S.M. Sze, "*A Floating Gate and its Application to Memory Devices*", The Bell System Technical Journal, vol. 46, no. 4, pp. 1288-1295, 1967.

2. T. Shibata, T. Ohmi, "*A Functional MOS Transistor Featuring Gate-Level Weighted Sum and Threshold Operations*", IEEE Transactions on Electron Devices, Vol.39, no.6, June 1993

3. K. Yang, A.G. Andreou, "*A Multiple-Input Differential Amplifier Based on Charge Sharing on a Floating Gate MOSFET*", Journal of Analog Integrated Circuits and Signal Processing, 6{3}, 1994.

4. Andreou A.G., "*A Multiple-Input Differential-Amplifier Based on Charge Sharing on a Floating Gate MOSFET*", Yang K.W., Analog Integrated Circuits and Signal Processing, vol. 6, No. 3, 1994, pp. 167-179.

5. Bradley A.Minch, Paul Hasler and Chris Diorio, "*Multiple-Input Translinear Elements Networks*", IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, Vol.48, no1, January 2001.

6. Naess. O., "*Continuous-Time Filter Design Using Floating Gate Circuits*", Ph.D. Thesis, University of Oslo, Depart. of Informatics, August 1999

7. C. Mead, "*Analog VLSI and Neural Systems*", Reading, MA: Addison Wesley, 1989

8. J. Ramirez-Ângulo, S.C. Choi and G. Gonzalez-Altamirano, "*Low-Supply Voltage OTA Architectures Using Floating Gate Transistors*", *IEEE Transactions on Circuits and Systems*, vol. 42, No. 12, pp.971-974, November 1995.

9. Paul Hasler, Bradley A. Minch and Chris Diorio, "*Floating Gate Devices : They Are Not Just For Digital Memories Anymore*", IEEE International Symposium on Circuits and Systems, Volume II, pages 391-399, Orlando, Florida, 1999.

10. A.F.Murray and L.W. Buchan, "A *User's Guide to Non-Volatile On-Chip Analogue",* Memory Electronics and Communication Engineering Journal, April, 1998.

11. "*Floating Gate Memory Arrays, Retention Issues*", Application Note, Mosaic Semiconductor Inc, February 1999.

12. Vincent F. Koosh and Rodney Goodman, "*Dynamic Charge Restoration of Floating Gate Subthreshold MOS Translinear Circuits*", Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS-2001), May 2001, Sydney, Australia, Vol. 1, pp. 33-36.

13. J.Ramirez-Ângulo, G. Gonzalez-Altamirano, and S.C. Choi, "*Modeling Multiple-Input Floating-Gate Transistors For Analog Signal Processing*", IEEE International Symposium on Circuits and Systems, June 9-12, 1997, Hong Kong.

14. Liming Vin, S.H.K. Edgar Sánchez-Sinencio, "*A Floating Gate MOSFET D/A Converter*", IEEE International Symposium on Circuits and Systems, June 9-12, 1997, Hong Kong.

15. Kyu-Hyoun Kim, Kwyro Lee, Tae-Sung Jung and Kang-Deog Suh, "*An 8-Bit-Resolution, 360us Write Time Nonvolatile Analog Memory Based on Differentially Balanced Constant-Tunneling-Current Scheme (DBCS)*", IEEE Journal of Solid-States Circuits, Vol. 33, No 11, November, 1998.

16. Katsuhiko Ohsaki, Noriaki Asamoto, and Shumichi Takagaki "*A Single Poly EEPROM Cell Structure for Use in Standard CMOS Processes*", IEEE Journal of Solid-State Circuits, Vol. 29, No 3, March 1994.

17. Hilbiber, D. F., "*A New Semiconductor Voltage Standard*", Digest of Technical Papers, IEEE-ISSCC, 1964, pp:32-33.

18. Robert A. Blauschild, Patrick A. Tucci, Richard S. Muller And Robert G. Meyer, "*A New NMOS Temperature-Stable Voltage Reference*", IEEE Journal of Solid-State Circuits, VOL. SC-13, No. 6, December 1978, pp: 767-774.

19. André Luis do Couto, João Paulo C. Cajueiro, Carlos A. dos Reis Filho "*Temperature-Compensated Voltage Using Floating-Gate MOS Transistor*", IMAPS-Brazil Proceedings, 6-8 August, 2003. Campinas, Brazil.

20. Frohman-Bentchkowsky-D. "*A Fully Decoded 2048-bit Electrically Programmable FAMOS Read-Only Memory*", IEEE Journal of Solid State Circuits, Vol. SC-6, No 5; Oct. 1971; p.301-6

21. Yngvar Berg, Tor S. Lande "*Programming Floating Gate Circuits With UV-Activated Conductances*", IEEE Transaction On Circuits and Systems II: Analog and Digital Signal Processing, Vol 48, NO. 1, January 2001.

22. L. Richard Carley "Trimming *Analog Circuits Using Floating-Gate Analog MOS Memory*", IEEE Journal of Solid-State Circuits, Vol 24, No 6, December, 1989.

23. M. Filanovsky "*Mutual Compensation of Mobility and Threshold Voltage Temperature Effects with Applications in CMOS Circuits*",*IEEE Transactions on Circuits and Systems*, Vol. 48, No 7, pp. 876-884, July 2001

24. 0.6um CMOS Joint Group Process Parameters #9933011, Rev.B. Austria Mikro Systeme, October 1988

25. P. Hasler, B.A. Minch and C. Diorio "*Adaptative Circuits using pfet floating gate devices*" in Proceedings of the 20th Anniversary Conference on Advanced Research in VLSI, Atlanta, GA, March 1999, pp. 215-229.

26. Abouchi, N.; Gallorini, R.; Vinard, C.; Grisel, R.; "*Analog EEPROM in standard process AMS 0.8 μm CMOS" Circuits and Systems*, 1999. 42nd Midwest Symposium on Volume 1, 8-11 Aug. 1999 Page(s):153 - 156 vol. 1

27. Matt Kucic, AiChen Low, Paul Hasler, and Joe Neff, "*Programmable continuous-time floating gate fourier processor*", IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, Vol. 48, pp.90-99, January 2001.

28. Hamid Reza Mehrvarz and Chee Yee Kwok "*A Novel Multi-Input Floating-Gate MOS Four-Quadrant Analog Multiplier*", IEEE Journal of Solid-State Circuits, Vol. 31, No 8, August 1996.

29. Shigeo Kinoshita, Takashi Morie, Makoto Nagata and Atsushi Iwata "*New Non-Volatile Analog Memory Circuits Using PWM Methods*", IEICE Trans. Electron, Vol E82-C, No 9, September 1999.

30. IEEE 1005-1998."*IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays*" New York, The Institute of Electrical and Electronics Engineer, Inc.

31. Z.A. Weingberg, "*On tunneling in metal-oxide-silicon structures*", J.Appl. Phys, 53(7), july 1982.

32. A.Ortiz-Conde, F.J. García Sánchez, J.J. Liou, "*A Review of Recent MOSFET threshold voltage extraction methods*", Microelectronic Reliability, 42(2002), 583-596