CENTRO DE INVESTIGACION Y DE ESTUDIOS AVANZADOS DEL I. P. N.

DEPARTAMENTO DE INGENIERIA ELECTRICA

SECCION DE ELECTRONICA DEL ESTADO SOLIDO

TITULO DE LA TESIS:

"ESTRUCTURAS CMOS-VLSI DE COMPUERTA FLOTANTE PARA SISTEMAS NEURONALES"

QUE PRESENTA:

M. en C. Mario Alfredo Reyes Barranca*

Para obtener el Grado de
Doctor en Ciencias
Especialidad en Ingeniería Eléctrica

ASESORES:

Dr. José Antonio Moreno Cadenas
Dr. Ciro Falcony Guajardo

*Becario de CONACYT

Septiembre de 1998

CONTENIDO

Abreviaciones

ADALINE	Neuronas Lineales Adaptativas.
ALU	Unidad Lógica Aritmética.
AR	Recombinación Auger.
ART	Redes de Resonancia Adaptativa.
BAM	Memoria Asociativa Bidireccional.
CG	Compuerta de control.
CMOS	Tecnología MOS de transistores complementarios.
CHE	Electrones calientes de canal.
DAHC	Portadores calientes por avalancha de drenador.
DIBL	Disminución de la Barrera Inducida por Drenador.
EEPROM	Memoria programable de sólo-lectura eléctricamente borrable.
EPROM	Memoria programable de sólo lectura borrable.
FEEPROM	Memorias EEPROM Flash.
FG	Compuerta flotante.
FGMOS	Transistor MOS de Compuerta Flotante.
FGNMOS	Transistor NMOS de compuerta flotante.
FGPMOS	Transistor PMOS de compuerta flotante.
FN	Tunelamiento Fowler-Nordheim.
LASER	Amplificación de Luz por Emisión Estimulada de Radiación.
LSB	Bit menos significativo.
MOS	Metal-Oxido-Semiconductor.
MOSFET	Transistor MOS por Efecto de Campo.
MOSIS	Sistema de Implementación MOS.
MSB	Bit mas significativo.
NMOS	Transistor MOS canal N.
P1	Polisilicio 1.
P2	Polisilicio 2.
P3	Polisilicio 3.
PMOS	Transistor MOS canal P.
RNA	Red Neuronal Artificial.
SGHE	Electrones calientes generados secundariamente.
SHE	Electrones calientes de substrato.
SOM	Memorias Auto-organizadas.
TMOS	Transistor MOS.
VLSI	Muy alta escala de integración.
WTA	Redes Ganador-Toma-Todo.

Símbolos

A	Area de traslapamiento entre polisilicios.
A	Constante para la ecuación de FN.
a_{ij}	Elemento de la matriz A.
A_k	Vector k de la capa A.
Atun	Area de tunelamiento.
B	Constante para la ecuación de FN.
b_{ij}	Elemento de la matriz B.
B_k	Vector k de la capa B.
C	Valor del capacitor en una red de capacitores conmutados.
C'd	Capacidad equivalente de la compuerta de programación.
C_D	Capacidad total del óxido de tunelamiento y cualquier capacidad parásita.
C_{FD}	Capacidad de acoplamiento del drenador.
Cfd	Capacidad debida al traslapamiento entre compuerta flotante y área de tunelamiento.
CFG	Capacidad entre compuertas de silicio policristalino.
Cfgb	Capacidad debida al traslapamiento entre compuerta flotante y el substrato.
$Cinyector_1$	Capacidad formada entre compuerta flotante e inyector 1.
$Cinyector_2$	Capacidad formada entre compuerta flotante e inyector 2.
Cox	Capacidad de compuerta.
C'ox	Capacitancia del óxido de compuerta por unidad de área.
Cpar	Capacitancia parásita.
Cpp	Capacidad entre P1 y P2.
Ctot	Capacidad total en el modelo del FGMOS.
d	Ciclo útil de un interruptor en la red de resistencias conmutadas.
E	Campo eléctrico a través del óxido de tunelamiento.
E	Valor de error en la Regla Delta.
$Ec_{1\rightarrow2}$	Campo eléctrico en la dirección Poly1 a Poly2..
$Ec_{2\rightarrow1}$	Campo eléctrico en la dirección Poly2 a Poly1.
e_{ox}	Espesor del óxido de silicio.
Etun	Campo eléctrico en el óxido de tunelamiento.
f(x)	Función de transferencia de una neurona.
f_{clk}	Frecuencia de reloj.
$H(A_i,A_j)$	Distancia de Hamming entre el vector A_i y el vector A_j.
$H(B_i, B_j)$	Distancia de Hamming entre el vector B_i y el vector B_j.
I	Entrada de los pesos a la red neuronal.
Ia	Corriente de entrada en la sinapsis de corriente diferencial.
Ib	Corriente de entrada en la sinapsis de corriente diferencial.
Id	Diferencia entre la corriente de un NMOS y de un PMOS en un inversor CMOS.
ID(x)	Corriente en drenador en función de la posición.
Idiff	Diferencia entre las corriente Ia e Ib.
Idn	Corriente a través de un transistor NMOS.
Idp	Corriente a través de un transistor PMOS.
Iij	Corriente entre la neurona i y la neurona j.
Io	Corriente de salida del arreglo del resistor controlado por voltaje.
Ipol	Corriente de polarización.
$I_{túnel}$	Corriente de tunelamiento.
Jt	Densidad de corriente FN.
k'	Parámetro de transconductancia (μo Cox)
Kcg	Coeficiente de acoplamiento entre la compuerta de control y la compuerta flotante.
Kp	Parámetro de transconductancia.
L	Longitud de canal del transistor MOS.
m_{ij}	Elemento de la matriz M.
Mk	Matriz de correlación de la BAM.
n	número de resistencias empleadas para el arreglo en escalera de resistencias conmutadas.

n	Número de vectores.
Na	Concentración de aceptores de un sustrato tipo P.
p	Número de vectores.
Q'_B	Densidad de carga en la región de deserción.
Qfg	Carga en la compuerta flotante.
Q'_I	Densidad de carga en la región de inversión.
Q'si	Densidad de carga en la superficie de canal.
Qt	Carga total en la compuerta.
R	Resistencia de salida.
Rds	Resistencia de canal del transistor MOS.
Ro	Resistencia fija en la red de resistencias conmutadas.
$S(a_i)$	Función de transferencia de la neurona A.
$S(b_j)$	Función de transferencia de la neurona B.
T	Actividad mínima requerida por la neurona para generar una salida positiva.
t	Salida deseada de la red neuronal supervisada.
t	Tiempo.
Tij	Conductancia entre la neurona i y la neurona j.
tox	Espesor del dióxido de silicio.
V(con carga)	Potencial efectivo debido a la carga en la interfaz Si/SiO_2.
V_{BB}	Voltaje de substrato.
Vc	Voltaje flotante en el arreglo del resistor controlado por voltaje.
Vcg	Voltaje en la compuerta de control.
V_{DB}	Voltaje drenador-substrato.
V_{DD}	Voltaje de polarización positivo de un inversor CMOS.
Vent	Voltaje de entrad de una neurona.
Vfb	Voltaje de bandas planas.
Vfg	Voltaje en la compuerta flotante.
V_{G2}	Voltaje en compuerta del FET para el arreglo de un resistor controlado por voltaje.
V_{GB}	Voltaje compuerta-substrato.
Vgs	Voltaje compuerta-substrato del transistor MOS.
Vin	Voltaje de entrada a un inversor CMOS.
Vin	Voltaje de entrada para el arreglo del resistor controlado por voltaje.
Vo	Voltaje de salida de un inversor CMOS.
Vo	Voltaje en las terminales diferentes a la compuerta de control del modelo del FGMOS.
$Vout_j$	Voltaje de salida de la neurona j.
Vp	Potencial de la compuerta flotante.
$Vp_{1\to2}$	Potencial del inyector 2.
$Vp_{2\to1}$	Potencial del inyector 1.
Vpol	Voltaje de polarización.
Vsal	Voltaje de salida de una neurona.
V_{SB}	Voltaje fuente-substrato.
V_{SS}	Voltaje de polarización negativos de un inversor CMOS.
Vt	Potencial térmico (kT/q)
Vt(flotox)	Voltaje de umbral de una estructura EEPROM.
Vt(MOS)	Voltaje de umbral de una estructura MOS.
Vta	Voltaje de entrada en el transistor Ma de la sinapsis de corriente diferencial-
Vtb	Voltaje de entrada en el transistor Mb de la sinapsis de corriente diferencial.
Vtd	Diferencia entre los voltaje de umbral Vta y Vtb.
Vth	Voltaje de umbral del transistor MOS.
Vth_n	Voltaje de umbral de un NMOS.
Vth_p	Voltaje de umbral de un PMOS.
Vti1	Voltaje en el inyector 1.
Vti2	Voltaje en el inyector 2.
Vtun	Voltaje aplicado al óxido de tunelamiento.
W	Ancho de canal del transistor MOS.
wi	Vector de pesos.

w_{ij}	Elemento de la matriz de pesos.
w_{new}	Actualización de los pesos.
w_{old}	Peso anterior a la actualización.
x	Voltaje variable de entrada a una neurona.
xi	Vector de entrada.
X_{tun}	Espesor del óxido de tunelamiento.
y	Salida del Perceptrón.
α	Constante para la ecuación de tunelamiento FN.
α	Razón de acoplamiento del drenador.
α	Razón de aprendizaje en la Regla de Hebb.
β	Constante para la ecuación de tunelamiento FN.
β	Factor de validez del perceptrón.
β	Parámetro de transconductancia del transistor MOS.
β	Razón de acoplamiento del canal.
ΔVth	Variación del voltaje de umbral del transistor MOS.
ε	Campo eléctrico a través del óxido de tunelamiento.
εo	Permitividad del vacío.
ε_{ox}	Permitividad del dióxido de silicio.
ε_{si}	Permitividad del silicio.
ϕ_B	Potencial interconstruído.
ϕ_F	Potencial de Fermi del semiconductor.
ϕ_{MS}	Potencial de contacto entre el metal y el semiconductor.
γ	Coeficiente de efecto de cuerpo.
λ	Parámetro de modulación de canal del transistor MOS.
μn	Movilidad de los electrones.
ψ_s	Potencial superficial.
ψ_{SD}	Potencial de superficie en el drenador.
ψ_{SS}	Potencial de superficie en la fuente.
σ	Parámetro de ganancia para la función de transferencia.

Prefacio.

Con la motivación de contribuir al desarrollo de circuitos analógicos que se puedan aplicar al estudio de las Redes Neuronales Artificiales, dado que el crecimiento de los circuitos en esta rama no se ha desenvuelto con la misma rapidez como lo han hecho los programas, es que se desarrolla este tema de tesis.

Uno de los problemas que enfrenta la implementación de las Redes Neuronales Artificiales y por lo que su desarrollo como circuito no ha sido tan rápido como los programas, es la tecnología disponible. Esto está relacionado con el hecho de que se requiere una alta interconectividad entre neuronas, por lo que reproducir la configuración de un tejido biológico con su equivalente electrónico, implica la utilización de una gran cantidad de área en silicio; ya que la sinapsis electrónica (el elemento de interconexión) se forma mediante un circuito analógico. Por lo tanto, esto limita la integración de una función compleja en un circuito integrado con menor área. Esto está conduciendo a que el desarrollo del estado del arte de las Redes Neuronales Artificales, se enfoque en buscar tanto tecnologías adecuadas, como elementos electrónicos que permitan aumentar la densidad de integración, así como diseñar sistemas neuronales complejos.

Esta tesis tiene como objetivo principal el diseño, fabricación y caracterización de estructuras de almacenamiento analógico utilizando compuerta flotante, con tecnología CMOS estándar, y su posible aplicación como elemento de interconexión programable (sinapsis) en arquitecturas de *Redes Neuronales Artificiales* (RNA).

El diseño concierne lo relativo a la concepción física y geométrica del dispositivo con *compuerta flotante* (FGMOS), incluyendo la simulación de su funcionamiento como elemento individual y como parte de un circuito. La fabricación, realizada con *tecnología estándar*, en nuestro caso implica la utilización de los servicios de una fábrica de silicio externa, por lo que es necesario cumplir con sus normas de diseño, y la caracterización permitirá obtener los parámetros típicos para configurar un *modelo*, así como corroborar el funcionamiento de los circuitos fabricados. Para alcanzar los objetivos anteriores, fue necesario resolver ciertos problemas que se mencionan a continuación.

Uno de los principales problemas para simular dispositivos de compuerta flotante, como elementos de un circuito de baja, mediana o alta complejidad, es la carencia de un modelo compatible con programas de simulación de circuitos. Dado que el trabajo desarrollado depende en gran medidad de la simulación, fue requisito indispensable establecer un modelo que permitiera representar eléctricamente el comportamiento del dispositivo.

Una vez fabricados los dispositivos, en este caso con tecnología estándar, se tiene la limitante de que no existe una manera sencilla de extraer uno de los parámetros de suma importancia, como lo es el coeficiente de acoplamiento. Esto llevó a desarrollar un método que permitiera extraer dicho parámetro con un montaje simplificado.

Como vehículos de prueba, validación del modelo propuesto y de la metodología indicada para extraer los parámetros, hubo necesidad de diseñar diversas estructuras, en las cuales el fenómeno de tunelamiento que se presenta en el proceso de inyección de carga, fue del tipo Fowler-Nordheim. Las estructuras diseñadas comprenden un grupo de inversores CMOS y dispositivos individuales, con los cuales se realizaron las pruebas eléctricas presentadas en este trabajo.

Otro elemento diseñado fue el correspondiente a una Memoria Asociativa Bidireccional (BAM). Esta es una de las arquitecturas con mayor atractivo para su implementación en circuitos integrados, dada la alta simetría de su arquitectura, lo que permite la integración de una Red Neuronal Artificial en áreas pequeñas y lo ilustrativo de su comportamiento como memoria asociativa, representada por una matriz sencilla, capaz de aprender una serie de patrones, en el marco del reconocimiento de patrones.

En resumen, en esta tesis se exponen las soluciones propuestas, estableciendo una metodología para el diseño, la fabricación y la caracterización de memorias analógicas de compuerta flotante.

En cuanto a la estructura que tiene el presente manuscrito, se tiene que en el Capítulo 1 se introduce una comparación breve de la computación digital con la analógica, específicamente enfocada a su aplicación en Redes Neuronales Artificiales. Se presenta a uno de los elementos mas atractivos para la implementación de sinapsis, como lo es la memoria analógica de compuerta flotante. En base al modelo simple de una neurona, se explica el papel que tiene la sinapsis en la respuesta de la neurona y las condiciones que debe cumplir para lograr que la red almacene información y sea capaz de aprender, en similitud al cerebro humano. Una red se forma conectando varias neuronas entre sí y se dan algunos ejemplos de redes con los que se pueda comprender el papel de la sinapsis, de manera global. De manera mas particular, se presentan las diferentes alternativas prácticas que existen para la configuración tanto de la neurona, como de la sinapsis, explicando sus ventajas y desventajas.

Ya que de la comparación de los elementos empleados como sinapsis, las memorias analógicas de compuerta flotante surgen como una alternativa prometedora, en el Capítulo 2, se hace una comparación dentro del estado del arte de este dispositivo, para elegir a su vez, aquel que cumpla con características relevantes, como no volatilidad, fabricación accesible, poco costo y área pequeña. Una vez justificadas estas características, a continuación se puede proponer la implementación de una red para el análisis del desempeño de este elemento, empleado como sinapsis. Esta red es la Memoria Asociativa Bidireccional.

En el Capítulo 3, se elige la configuración de la Memoria Asociativa Bidireccional y se presentan los cálculos, en base al modelo propuesto por Kosko, para obtener su matriz de correlación. Esta matriz corresponde a los pesos de interconexión y es la que se relaciona mas adelante, con el voltaje de umbral de los dispositivos de compuerta flotante empleados en la sinapsis. Partiendo de estos cálculos y de las configuraciones electrónicas de la sinapsis y la neurona, se hace el diseño eléctrico de la red, incluyendo los circuitos de control de la misma.

Una vez especificados los parámetros de diseño eléctrico, en el Capítulo 4 se definen los parámetros de diseño físico a partir de las condiciones de programación (cambio del voltaje de umbral) ahí propuestas, y del coeficiente de acoplamiento, para una estructura cuya inyección de carga hacia o desde la compuerta flotante sea mediante el fenómeno de tunelamiento Fowler-Nordheim. Se establece una metodología para el cálculo y caracterización de los parámetros de importancia y se propone un modelo compatible con el programa de simulación. Finalmente, se realiza una simulación circuital utilizando el programa PSpice, para determinar la matriz de correlación en función de los voltajes de umbral y para comprobar el desempeño de la red, en cuanto al reconocimiento de patrones, empleando las memorias analógicas.

El Capítulo 5, muestra el resultado del diseño geométrico del circuito propuesto y simulado, empleando el programa L-Edit. Las reglas de diseño empleadas corresponden a la fábrica de silicio Orbit y de su tecnología de 2 μm, pozo n, doble metal y doble polisilicio. Se incluyen también las estructuras de prueba diseñadas para la caracterización de los parámetros eléctricos y de programación.

La caracterización de las estructuras de compuerta flotante y sus resultados, se discuten en el Capítulo 6. Se comparan los resultados medidos con los calculados y se comprueba la validez del modelo propuesto. Además. se presentan las condiciones de programación obtenidas para las estructuras de compuerta flotante diseñadas, haciendo uso de estas condiciones para obtener las características de salida (I-V) de una sinapsis y de una fuente de corriente programable.

Finalmente, en el Capítulo 7 se presentan las conclusiones del trabajo.

CAPITULO 1.

INTRODUCCION.

Para el análisis y procesamiento de datos por medio de computadora, se están haciendo esfuerzos que ayuden a simplificar y optimizar su manejo, tomando criterios como velocidad, área y precio. La técnica que ha tenido gran aplicación es la digital, basada sobre todo en circuitos digitales **CMOS**, donde se manejan valores discretos que pueden ser interpretados, almacenados, procesados y presentados de manera rápida y sencilla. Esto dio origen a un rápido desarrollo de tecnología que permitiera la fabricación de manera accesible de circuitos con inmediata aplicación en la electrónica digital.

En la década de los 60's, se comenzó con la idea de realizar circuitos analógicos para fines de computación y aprendizaje, pero utilizando componentes discretos que limitaban la capacidad y hacían que el volumen ocupado fuera poco práctico. Sin embargo, se ha alcanzado un nivel tecnológico que permite tener circuitos para aplicación analógica, integrados en un área muy pequeña. Algunas de las aplicaciones encontradas se dirigen a *redes neuronales* utilizadas como *memorias analógicas*, *almacenamiento ponderado* o *ajuste*.

Lo anterior hace que las implementaciones analógicas sean promisorias debido a que se ha visto que el área en silicio sea considerablemente menor que la contraparte digital, no sólo desde el punto de vista de circuitos, sino también por la parte del alambrado. La alta conectividad de la arquitectura de las redes neuronales, hace que la transmisión de información en forma analógica sea particularmente atractiva. Esto ha hecho que últimamente se dedique investigación al desarrollo de electrónica para redes neuronales.

Un ejemplo de esto, son los trabajos hechos para encontrar un elemento de memoria analógico [1, 2, 3] implementado con técnicas de almacenamiento digital, con las cuales se necesita electrónica adicional, métodos restringidos de conversión de analógico a digital, tener un compromiso entre el número de bits de resolución ponderada y el área, y en algunos casos, el uso de tecnología muy especial para su fabricación, como lo es para las memorias **EEPROM**. Dada la dependencia de los elementos analógicos con respecto a parámetros físicos y eléctricos, no se puede lograr tener una buena resolución con los mismos para una función dada, por lo que pretender cumplir con el objetivo de tener alta resolución con circuitos analógicos sería una mala elección. Un ejemplo ilustrativo de esto se puede ver comparando un multiplicador digital con uno analógico, donde con el primero el resultado es exacto (aunque el tamaño de palabra determina el número de componentes) y con el segundo, no es posible tener un resultado exacto, aún cuando la tecnología sea muy buena.

Los elementos de memoria más comunes en la actualidad son la **EPROM** y la **EEPROM**. Sin embargo, se ha visto que se está llenando un nicho existente entre estas, con el desarrollo de memorias analógicas. Cada una de ellas tiene sus ventajas y desventajas, según la utilidad y propósito destinado. En este caso, la justificación para estudiar las memorias analógicas, se da por la aplicación que se le puede dar a las redes neuronales, como es el caso del presente trabajo.

Enfocándose entonces hacia los circuitos CMOS analógicos, cuya precisión depende del acoplamiento de transistores, para disminuir, por ejemplo, el voltaje de corrimiento (offset), la no linealidad y el error de ganancia, se tiene que el diseño teórico indica que el tamaño de los dispositivos se debe aumentar para evitar los posibles efectos de desacople entre transistores. En este caso se tiene un compromiso entre la función y el elemento a usar, por lo que una mayor precisión impone la utilización de electrónica digital por ser más conveniente, pero cuando no se requiere precisión, como es el caso de la mayoría de las aplicaciones en redes neuronales artificiales, la opción analógica es la mejor. Al respecto, una solución es la aplicación de *memorias MOS analógicas de compuerta flotante*, para el ajuste del voltaje de corrimiento a un valor mínimo, dejando de lado, técnicas como arreglos de centroide común, ajuste por LASER o programación de redes de resistencias. La fabricación de estas memorias, es compatible con la tecnología CMOS estándar, derivando costos similares. En trabajos iniciales en circuitos analógicos,

usando memorias no volátiles, se llegó a tener un desplazamiento en el voltaje de umbral del 40 % en cuatro días. Actualmente, se han logrado realizar memorias cuyo valor extrapolado de variación en 10 años, alcanza el 1 % [4] y también se pudo reducir el voltaje de corrimiento en un circuito analógico de 10 mV a 0.5 mV [5].

Las memorias están basadas en la compuerta flotante de un MOSFET, diseñado para el proceso CMOS de doble polisilicio. Su comportamiento depende de la carga almacenada en la compuerta flotante. Esta compuerta está totalmente aislada de las capas de óxido que la rodean y actúa como un capacitor con enorme capacidad de retención. La cantidad de carga atrapada en la compuerta flotante puede ser cambiada mediante carga y descarga de electrones a través del óxido vía tunelamiento **Fowler-Nordheim**.

Esto último permite la escritura o borrado de la memoria, propiedad que puede ser utilizada para su aplicación en redes neuronales con características de aprendizaje. Se define a una red neuronal artificial, como la simulación de un sistema nervioso real que contiene una colección de neuronas unitarias, comunicadas unas con otras, a través de un axón. La información fluye a través del axón hacia el elemento de decisión, que es la neurona. Por lo tanto, el modelo artificial simple del sistema nervioso constará de los elementos de interconexión, llamados sinapsis, y del elemento de procesamiento, llamado neurona. La forma y cantidad de interconexiones entre los elementos básicos de la neurona artificial dará lugar a las arquitecturas para desempeñar una función en particular. La exactitud y tipo de la función final dependerá de la complejidad de la arquitectura, así como del procesamiento de la señal de entrada, es decir, el algoritmo.

Para comprender la importancia de la sinapsis dentro de las redes neuronales, a continuación se presentan algunos ejemplos básicos y los conceptos fundamentales que clasifican las diferentes arquitecturas y algoritmos.

1.1 Conceptos fundamentales de las Redes Neuronales Artificiales (RNA).

Las Redes Neuronales Artificiales (**RNA**), son intentos de reproducir, por lo menos parcialmente, la estructura y funciones del cerebro y el sistema nervioso. El cerebro humano contiene billones de neuronas cuya manera de interconectarse, nos permite razonar, memorizar y computar. Los avances en la tecnología de circuitos VLSI y la demanda de máquinas "inteligentes", ha creado un creciente interés en emular sistemas neuronales para aplicaciones reales.

Las redes neuronales son entrenadas por ejemplos sucesivos en un ambiente de tiempo real. Estas se adaptan a los cambios de su ambiente y desarrollan sus propias reglas internas. Una de las ventajas de estas redes, es su habilidad de manejar datos difusos o incompletos. Se considera que las RNA son modelos computacionales viables para la solución de una gran variedad de problemas, que incluye los campos de clasificación de patrones, síntesis y reconocimiento del habla, interfaces adaptativas entre el humano y sistemas físicos complejos, compresión de imágenes, memorias asociativas, agrupamientos, optimización, modelado de sistemas no lineales y control. Los modelos usuales de redes neuronales incluyen:

- Redes de Hopfield
- Redes de Hamming
- Adaline de Widrow
- Perceptrones mono-capa de Rosenblatt
- Retropropagación de error de Werbos para Perceptrones multicapa
- Teoría de Resonancia Adaptativa de Carpenter y Grossberg
- Máquinas Boltzmann de Hinton y Sejnowski
- Mapa de auto-organización de Kohonen
- Neocongnitrón de Fukushima
- Memoria Asociativa Bidireccional de Kosko

En particular, las redes neuronales emplean una gran cantidad de eslabones de comunicación entre los elementos de procesamiento para realizar *procesamiento paralelo distribuido*. Dada la gran tolerancia al

error de las redes neuronales, la operación global de ésta, no se verá afectada si existen algunos elementos de procesamiento degradados o no funcionales.

Una de las características más importantes de las redes neuronales artificiales, es que realizan un gran número de operaciones numéricas en paralelo. Estas incluyen operaciones aritméticas simples así como mapeos no lineales y cálculo de derivadas. Casi todos los datos almacenados en la red son usados en un cálculo. De hecho, el procesamiento neuronal distribuido es realizado típicamente dentro de todo el arreglo compuesto por neuronas y pesos. Esto indica que la neurocomputación hace un uso más eficiente de los datos almacenados (pesos) y los datos de entrada (estímulos).

Las computadoras convencionales operan con un número relativamente limitado de datos al mismo tiempo. La unidad lógica aritmética (ALU) manipula sólo dos palabras recientemente extraídas de memoria. Aunque los procesadores convencionales pueden realizar rápidamente una compleja variedad de instrucciones, muchos megabytes de datos permanecen inactivos en un ciclo de instrucción. Durante el procesamiento, las operaciones realizadas con los datos cargados son realizadas durante el ciclo útil, sin embargo, al mismo tiempo, quedan almacenados muchos datos que quedan esperando su turno para ser procesados, lo cual impide que sean aprovechados durante el ciclo. Esto explica el cuello de botella de las computadoras convencionales (de operación secuencial) cuando manejan una gran cantidad de datos de entrada/salida y realizan rutinas como reconocimiento de visión o de voz. Por lo tanto, al ser las redes neuronales artificiales, sistemas de procesamiento en paralelo que evitan el cuello de botella, da como resultado que pueden manejar cantidades masivas de datos de entrada/salida de manera más eficiente.

La unidad de procesamiento neuronal, consiste de un nodo de procesamiento (*neurona*) con pesos (*conexiones sinápticas*), como se muestra en la Fig. 1.1. La unidad de procesamiento necesita memorizar el dato almacenado, para producir un valor de activación de la neurona y computar la salida de la unidad. Esta característica hace posible que la unidad pueda recordar la función. Además, sería deseable que el nodo pudiera almacenar una regla de aprendizaje y adaptar los pesos de acuerdo a tal regla. Las señales procesadas y funciones pueden ser analógicas, digitales o una combinación de las dos y es importante la implementación de un elemento electrónico que apoye esta característica, para la realización de circuitos.

Basado en una función de activación no lineal, la neurona “se dispara” si la suma de las señales de entrada exceden un cierto umbral (excitatoria) o “se apaga” si no lo exceden (inhibitoria). En general, se tienen tres funciones de transferencia no lineales: de alta ganancia, lineal y sigmoide, mostradas en la Fig. 1.2.

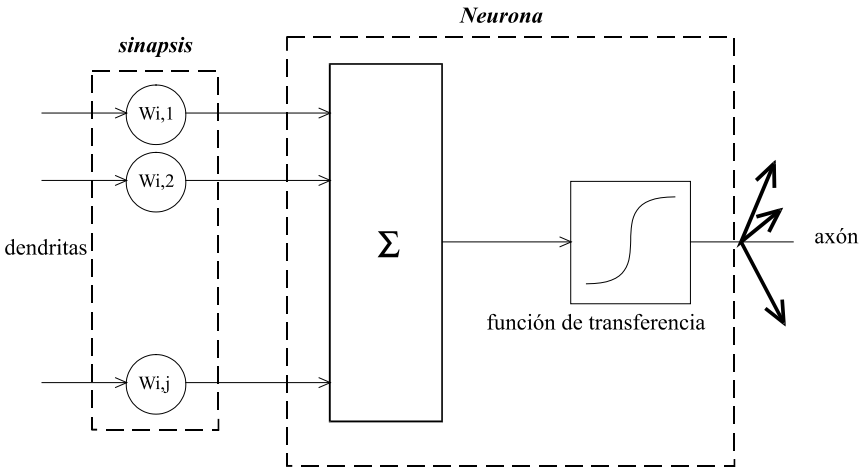


Fig. 1.1. Modelo de una Red Neuronal Artificial.

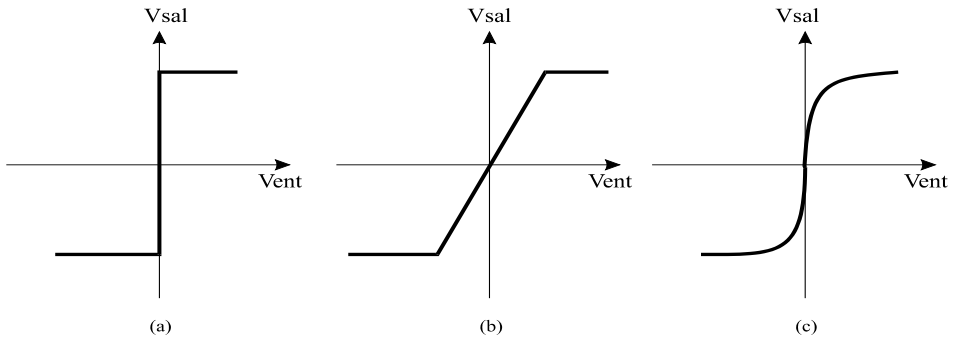


Fig. 1.2. Función de transferencia de una neurona artificial: a) alta ganancia; b) lineal; c) sigmoide.

En el aspecto de aprendizaje, se han desarrollado una gran variedad de algoritmos para este fin [6]. En una red neuronal, el aprendizaje se lleva a cabo adaptando cambios a los pesos de interconexión (ver Fig. 1.3), es decir, las sinapsis, que será el objeto de estudio de esta tesis. De esta manera, se puede construir un clasificador, no programándolo sino presentándole un cierto número de ejemplos y permitiendo que la red haga la discriminación automáticamente. La capacidad de aprendizaje de redes multicapas es una de las áreas más activas de la integración en redes neuronales.

En el esquema de la Fig. 1.3, se presenta un arreglo de aprendizaje supervisado. A la red, se le presenta un conjunto de datos de entrenamiento. Para cada ejemplo, la salida de la red se compara con el objetivo y se realizan ligeros ajustes a los pesos de interconexión de la red mediante el cálculo de error, dando como resultado un cambio de los pesos, aplicado a las sinapsis, representadas simplemente por las líneas de interconexión entre neuronas[1]. Con un algoritmo apropiado de ajuste de peso, como el de **Retropropagación**, una presentación abundante de datos de entrenamiento, da como resultado una red que presente una relación entrada/salida para datos de entrenamiento. Si la red tiene la arquitectura apropiada y si se tuvo suficiente entrenamiento, la red será capaz de generalizar y podrá dar una salida correcta para entradas que nunca ha visto. Las **Redes de Hopfield** pueden ser un ejemplo de redes con aprendizaje supervisado y las **Redes de Kohonen** pueden representar a algunas redes no supervisadas.

En la Fig. 1.4 se muestra otra forma de operación de las redes neuronales, donde se tiene un patrón de referencia que se va a comparar con la entrada. A la salida de los inversores (neuronas) se tiene una retroalimentación que se dirige a la red donde se determina qué salida se aproxima más al patrón de entrada, inhibiendo la salida de la red que menos se acopla. Si estas conexiones inhibitorias son más fuertes que las excitatorias, en la red del patrón de referencia, sólo una neurona tendrá salida estable. Si varias neuronas están encendidas, se tratarán de inhibir unas a otras y sólo una quedará encendida.

De lo anterior se desprende que los algoritmos de aprendizaje tienen gran importancia para las redes neuronales. Estos pueden ser aplicados de manera digital o analógica. Los algoritmos de aprendizaje que requieren una alta resolución en los pesos de interconexión, necesitan electrónica de alta precisión para ajustar los pesos. Los circuitos que cumplan con tal característica, requieren mayor área de silicio. Los circuitos analógicos se usan solo donde se requiere una precisión moderada, mientras que los circuitos digitales se usan para algoritmos con gran resolución.

[1] *La simbología para las sinapsis en lass arquitecturas de las RNA, es una línea de interconexión, entendiéndose que involucra un valor de peso o conductancia; a las neuronas con un círculo.*

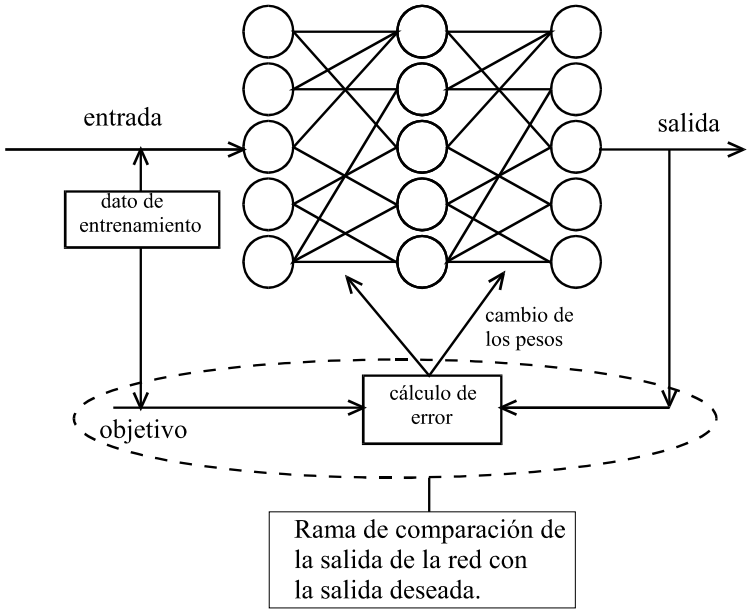


Fig. 1.3. Arreglo de Red Neuronal Artificial para aprendizaje supervisado.

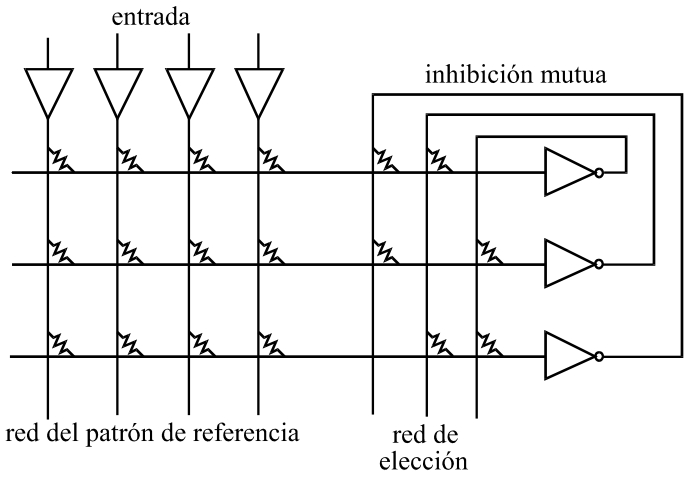


Fig. 1.4. Red de Hopfield con pesos fijos.

Comparando los circuitos analógicos con los digitales, se podrían identificar varias ventajas de los primeros:

1. El número de operaciones efectuadas, evaluadas en términos de interconexiones por segundo (ejecutar la suma de productos), es mucho más grande para los circuitos analógicos que para los digitales [1].
2. Se tiene una mejor aproximación del comportamiento analógico de las redes neuronales biológicas.
3. Ya que para realizar una misma función, se requieren menos componentes si se usa la electrónica analógica, comparada con la digital.

Implementación de los pesos sinápticos.

Se vio anteriormente que la neurona se puede representar por los pesos sinápticos que entran a un sumador, que a su vez llegan a un amplificador que es encendido o apagado, dependiendo de si la señal es excitatoria o inhibitoria.

En la Fig. 1.5, se puede ver que se tiene una entrada a la neurona i proveniente de la neurona j, a través de una conductancia T_{ij}. Esta conductancia es el llamado *peso de conexión*. Si el voltaje de entrada se mantiene a tierra (tierra virtual), las señales de las otras neuronas serán corrientes de valor [1]:

$$I_{ij} = V_{outj} * T_{ij} \quad . \tag{1.1}$$

Todas estas corrientes se suman a la entrada y por lo tanto, el voltaje a la salida de la neurona i será función de la corriente total:

$$V_{outi} = f\left(\sum I_{ij}\right) = f\left(\sum V_{outj} * T_{ij}\right) \quad . \tag{1.2}$$

La ecuación anterior muestra que la suma de productos es una operación clave de la red y por lo tanto el diseño del circuito se debe orientar a lograrlo de manera eficiente. En una red neuronal, una simple resistencia puede realizar una multiplicación usando la ley de Ohm y la suma de corrientes en un alambre se expresa por medio de la ley de Kirchhoff. Por lo tanto, un circuito analógico que calcule sumas de productos puede ser construido de manera más compacta que un circuito digital.

Una resistencia fija puede realizar la suma de productos, sin embargo, no se podría aplicar un algoritmo de aprendizaje, ya que, como se mencionó anteriormente, se requiere que los pesos sinápticos se ajusten a medida que se presenten los ejemplos de entrenamiento, además de que puedan ser "recordados" los valores de la sinapsis. Esto conduce a uno de los problemas centrales en el desarrollo de las RNA: hacer que los pesos sean continuamente ajustados, en respuesta a una señal analógica de control, además de que el diseño y control de su valor, no requiera de muchos transistores. También los pesos deben ser capaces de aprender bajo una regla de aprendizaje.

Se pueden encontrar varios circuitos que cumplan con la cualidad de tener una resistencia variable [1, 7, 8], como son: capacitores conmutados, resistencias conmutadas, resistencias en escalera conmutadas, o resistencias controladas por voltaje (pares FET acoplados), sin embargo, la tecnología del transistor MOS de compuerta flotante es la mejor alternativa en la actualidad. Este dispositivo combina el almacenamiento no volátil y el elemento de conexión (peso) en un solo dispositivo, es decir, el peso está determinado por la carga almacenada en la compuerta flotante.

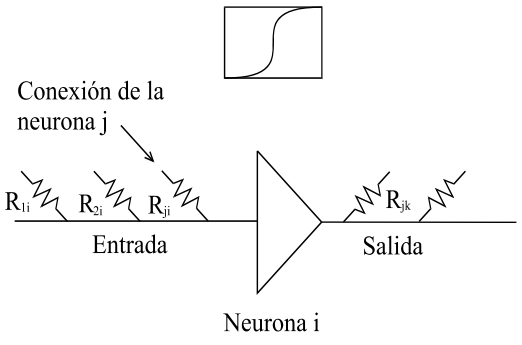


Fig. 1.5. Esquema de una Red Neuronal Artificial.

Memoria de compuerta flotante.

Actualmente se tienen muchas versiones de transistores MOS con compuerta flotante, similares a las ya conocidas EPROM y EEPROM. Según se puede recopilar en la literatura, las investigadas y desarrolladas para aplicar en redes neuronales, tienen características que las ubican por arriba de las EPROM y por debajo de las EEPROM en cuanto a costo. Estas últimas tienen la desventaja de requerir tecnología poco accesible a los investigadores, por constar de procesos muy especiales como un óxido ultrafino (20-40 nm) y superficies texturizadas para aumentar el campo eléctrico. La tendencia de las memorias para redes neuronales, es utilizar tecnologías estándar como la de doble polisilicio de 2 m que es accesible y no muy cara, para facilitar la fabricación y hacerlas competitivas contra otras memorias.

Las memorias de compuerta flotante están basadas en un transistor MOS, con una compuerta flotante y una compuerta de control, ambas de silicio policristalino. Mediante una polarización adecuada de las terminales de fuente, drenador y compuerta de control, se logra la inyección de electrones a través del óxido entre drenador y compuerta flotante, para almacenar carga en el régimen de escritura o eliminación de carga en el régimen de borrado.

Las principales preocupaciones en el diseño y fabricación de las memorias, se centran en la optimización del mecanismo de inyección de electrones, para disminuir la degradación del óxido de tunelamiento, el cual es provocado por tunelamiento Fowler-Nordheim (FN) o electrones calientes de canal. A medida que se avanza en la investigación de los fenómenos de inyección, se va logrando aumentar el número de ciclos de escritura/borrado, lográndose un orden de 10^4 ciclos útiles. También se desea aumentar el tiempo de almacenamiento de la carga que actualmente llega a ser del orden de 10 años. Esto se debe hacer disminuyendo los mecanismos de fuga de carga.

1.2. Arquitecturas de RNA.

Desde que se comenzaron a estudiar las RNA, se han ideado y desarrollado varias arquitecturas que permiten realizar diferentes funciones (como las mostradas en las Figs. 1.3 y 1.4) o tener ciertas habilidades, tales como clasificación, reconocimiento y control. Pueden encontrarse tres tipos de arquitecturas estudiadas en el campo de las RNA: Red de Capa Sencilla (**Single Layer Net**), Red Multicapa (**Multilayer Net**) y de Capa Competitiva (**Competitive Layer**). En cada una de las categorías anteriores se pueden encontrar variaciones en la interconexión y en el algoritmo de aprendizaje utilizado para el establecimiento de los pesos. La aplicación o el problema a resolver será lo que determine la arquitectura y complejidad de la misma. Lógicamente, entre más difícil sea el problema, más compleja deberá ser la arquitectura que se deba utilizar en su solución. La habilidad de ajustar los pesos se puede conseguir también por diferentes métodos de entrenamiento de la red. En este sentido, se puede tener entrenamiento *supervisados* o *no supervisados*, dependiendo de la arquitectura y del algoritmo usado.

1.2.1. Redes Supervisadas y No Supervisadas.

Se llama entrenamiento supervisado al proceso de presentar una secuencia de vectores o patrones de entrenamiento que tienen asociado un vector de salida de referencia (*target*), con lo que se realiza un ajuste de pesos sinápticos de acuerdo a un algoritmo de aprendizaje, es decir, en cada presentación de un patrón en particular, la red evoluciona cambiando el valor de las interconexiones entre neuronas ajustándose según el error existente al comparar el vector de salida con el vector de referencia.

La función de este tipo de redes puede ser tan simple como la clasificación de patrones pertenecientes o no a una determinada categoría: salida bivalente (1 si pertenece; -1 si no pertenece). Esto es posible con redes de una sola capa utilizando el *Algoritmo de Aprendizaje de Hebb* o la *Regla Delta*. El grado de dificultad en la clasificación puede aumentar y por consiguiente es necesario aumentar la complejidad de la red, en este caso, con una arquitectura multicapa en la cual se aplica el *Algoritmo de Aprendizaje de Retropropagación* (Backpropagation).

Las redes con aprendizaje supervisado encuentran aplicaciones como clasificación de patrones o como asociación de patrones. Esta última categoría asocia un grupo de vectores de entrada con otro grupo de vectores de salida, lo que hace que se le aplique el nombre de *memorias asociativas*. A su vez, las memorias pueden ser *autoasociativas*, cuando se desea que el vector de salida sea el mismo que el de entrada; o bien pueden ser *heteroasociativas*, cuando el vector de salida es diferente al de entrada. Cuando las memorias asociativas son entrenadas, estas pueden "reconocer" al patrón almacenado aún cuando se le presenta a la entrada un patrón muy similar, es decir, es posible que reconozca un patrón a pesar de encontrarse distorsionado o escaso de datos.

Las memorias asociativas se encuentran dentro del grupo de redes con aprendizaje supervisado y sus arquitecturas (la manera de interconectar las neuronas) pueden ser del tipo de *avance* (feedforward) o *recurrente* (iterativa). Como ejemplo de la arquitectura de avance están la **Red Neuronal Heteroasociativa** y la **Red Neuronal Autoasociativa**, que se muestran en la figura 1.6; la **Red de Hopfield Discreta** y la **Memoria Asociativa Bidireccional** (BAM), son ejemplos de arquitecturas recurrentes o iterativas, que se muestran en la figura 1.7. Todas ellas se pueden diseñar, por ejemplo, para reconocimiento de caracteres como letras. Mediante un algoritmo se consigue que la red memorice determinado número de letras (todas las redes tienen un límite de aprendizaje) y posteriormente debe ser capaz de reconocer alguna de las letras almacenadas cuando se le presenta una de ellas con alguna variante.

La representación mostrada en las figuras es general pero ilustrativa: los círculos representan a las neuronas, que colocadas en línea forman una capa, y las líneas representan a las sinapsis o pesos, que cambian su valor para almacenar los vectores deseados. Se consideran de capa sencilla o monocapa cuando existe sólo una capa después de la capa de entrada o multicapa cuando existen más de dos capas después de la de entrada.

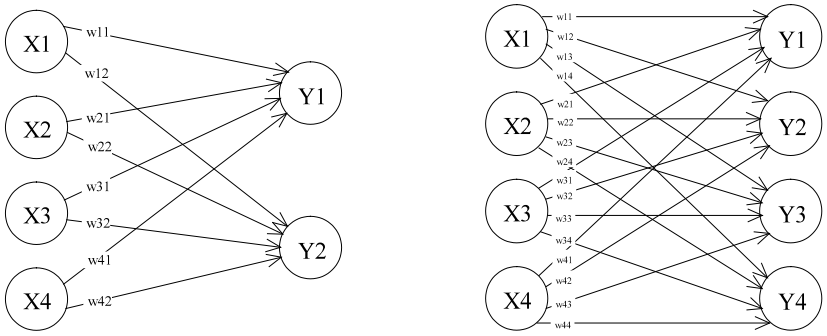


Fig. 1.6. a) Red Neuronal Heteroasociativa; b) Red Neuronal Autoasociativa.

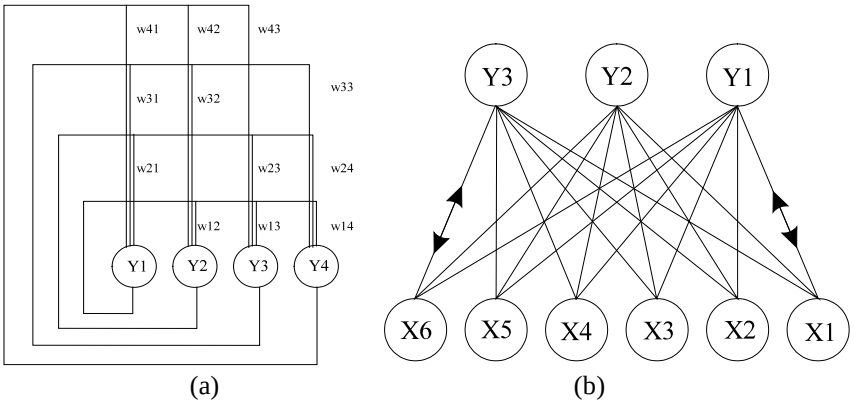
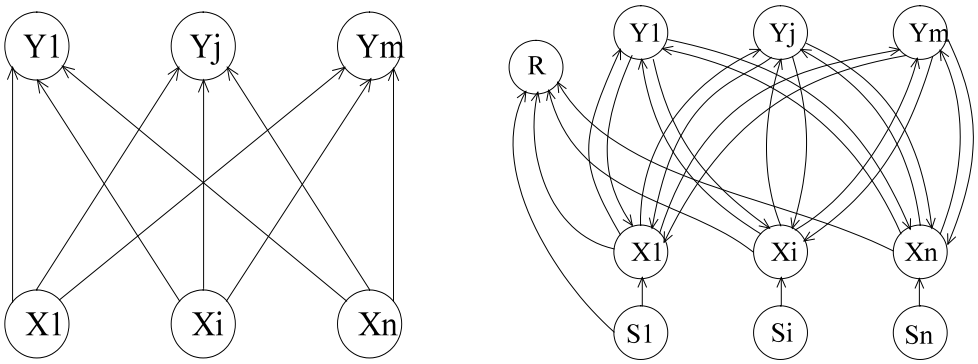


Fig. 1.7. a) Red de Hopfield discreta; b) Memoria Asociativa Bidireccional.

Las Redes No Supervisadas no requieren tener un patrón de referencia para comparar con la salida. A este tipo de redes, se les entrena presentándoles una secuencia de vectores de entrada y la red agrupa vectores muy similares en categorías correspondientes. Los pesos son modificados de tal manera que los vectores parecidos son asignados a una misma salida. Las Redes Auto Organizadas de Kohonen (**SOM**) y las Redes de Resonancia Adaptativa (**ART**) son ejemplos clásicos de redes no supervisadas y además son las que corresponden a la tercera categoría de redes llamada de Competencia.



(a) (b)

Fig. 1.8. a) Mapa Auto-organizado de Kohonen; b) Teoría de Resonancia Adaptativa.

En ocasiones, al clasificar varios patrones la red responde con un mismo agrupamiento para varios patrones y se desea que se separe la clasificación de manera más definida, por lo que se requiere forzar a que responda solo una neurona y esto se hace incluyendo otra estructura en la red para que sea tomada una decisión respecto al "conflicto". Es por esto que a este tipo de redes se les llama de **Competencia** o **Inhibición Lateral,** en las que varias neuronas entran en competencia y solo una es la que es entrenada para dar la respuesta correcta. Ejemplos de este tipo de redes se muestran en la figura 1.8, con el Mapa Auto-Organizado de Kohonen (Fig. 1.8a) y la red de Teoría de Resonancia Adaptativa, ART (Fig. 1.8b). Se puede decir entonces, que la diferencia entre la arquitectura de la Fig. 1.6 (a) y 1.8(a) es su función, donde la primera reconoce patrones (es una memoria) y la segunda los clasifica según la similitud de propiedades, lo cual se logra al aplicar diferentes algoritmos a una misma arquitectura. Esta puede ser mas compleja según el objetivo que se persiga.

1.2.2. Redes con algoritmo de Avance y de Retropropagación.

En la sección anterior, se hizo mención de los conceptos de avance y retropropagación y ahora se explicará la diferencia entre uno y otro. A grandes rasgos, se puede tener una Red de Avance o una Red de Retropropagación, donde la primera es una red en la cual el flujo de información se lleva a cabo únicamente desde las capa de entrada hacia la capa de salida y donde el algoritmo utilizado para cambiar y ajustar los pesos para aprender determinados patrones, se aplica según se le presenten los ejemplos de entrenamiento a la entrada de la **Red de Avance**, dándose a este proceso el nombre de *Aprendizaje por Avance*; la segunda también puede ser una red de avance pero entrenada con el algoritmo de retropropagación donde el ajuste de pesos se consigue enviando la información desde la capa de entrada hacia la capa oculta inmediata y de ahí hacia la capa de salida de donde la respuesta se regresa en sentido inverso hacia la capa oculta, llevando la información del error existente, cambiando los pesos anteriores para disminuir el error y enviando la información, a su vez, hacia la capa de entrada e iniciar el proceso de regreso. Este proceso de enviar y regresar la información, se realiza hasta que se cumpla con la tolerancia de error especificada, es decir, el entrenamiento de la **Red de Retropropagación** se lleva a cabo en tres fases: 1) enviar la información del patrón presentado a la entrada, 2) el cálculo y retropropagación del error asociado y 3) el ajuste de los pesos; en términos generales a este proceso se le conoce como *Aprendizaje por Retropropagación*.

1.3. Algoritmos de Aprendizaje.

La importancia de las RNA, radica en la capacidad que tienen de aprender a resolver un problema. Esto se logra mediante algoritmos con los cuales es posible cambiar el valor del peso de las interconexiones (sinapsis), conforme se entrena una red para resolver un problema específico. Estos pesos no siempre se ajustan, ya que existen redes que se aplican a problemas con soluciones conocidas de antemano y por lo tanto no es necesario cambiar el valor de los pesos; a estas se les llama *redes de pesos fijos*. Sin embargo, cuando se requiere que la red responda de manera correcta a excitaciones que nunca ha visto, sí se requiere que se modifiquen los pesos durante una fase de entrenamiento, con ejemplos típicos similares a los que se piensa usar durante la operación normal de la red. De esta forma la red es capaz de extrapolar, de la misma manera como lo hace el cerebro, a partir de unos cuantos datos recibidos hacia la información completa requerida.

Las redes neuronales deben ser entrenadas para poder aprender. De aquí surgen dos conceptos que es oportuno diferenciar: *Entrenamiento* es el procedimiento por el cual las redes aprenden; *Aprendizaje* es el resultado final del entrenamiento. El entrenamiento es un proceso externo a la red y el aprendizaje es una actividad interna de la red. Se pueden encontrar diferentes procesos o algoritmos de aprendizaje que están relacionados con el tipo de arquitectura empleada [9]. A continuación se explican los más generales, de los cuales se pueden derivar algunas variantes.

Perceptrón.

Al final de la década de los 50's, Frank Rosenblatt introdujo la idea del Perceptrón en las redes neuronales artificiales, cuyo procesamiento es muy similar al expuesto por McCulloch-Pitts, a quienes se les considera como los iniciadores de las redes neuronales. El procedimiento bajo el cual se realiza este proceso consiste simplemente en sumar las señales de entrada a la neurona y comparar el peso neto con respecto a un umbral, dando una señal de +1 cuando la suma es mayor al umbral definido y -1 cuando es menor a éste. La función de transferencia del perceptrón es como sigue:

$$I = \sum_{i=1}^{n} wi\, xi \quad , \tag{1.3}$$

$$y = \begin{pmatrix} +1, & \text{si} \quad I \geq T \\ -1, & \text{si} \quad I < T \end{pmatrix} \quad , \tag{1.4}$$

donde *I* es la entrada de los pesos a la red, *wi* se refiere al vector de pesos y *xi* al vector de entrada; *y* es la salida del perceptrón. T es la actividad mínima requerida por la neurona para generar una salida positiva y puede tener cualquier valor. El entrenamiento establecido por Rosenblatt para su red estaba diseñado para separar varios patrones en dos categorías. El algoritmo usado en el perceptrón se expresa a continuación:

$$w_{new} = w_{old} + \beta\, y\, x$$

$$\beta = \begin{pmatrix} +1, & \text{si la respuesta del perceptrón es correcta} \\ -1, & \text{si la respuesta del perceptrón es incorrecta} \end{pmatrix} ,$$

$$y = \text{salida del perceptrón} ,$$

donde *w* representa al vector de pesos y *x* al vector de entrada. Así, cada vez que se presente un patrón a la entrada de la red, el nuevo peso será función de la suma o resta del peso anterior con la respuesta dada por la neurona para cada patrón. Con el algoritmo anterior, se puede clasificar a un patrón como perteneciente o no a una categoría dada.

Regla Delta.

Uno de los investigadores que ayudaron a impulsar las RNA fue Bernard Widrow, quien junto con su estudiante Ted Hoff, desarrollaron la llamada Regla Delta, también conocida como Regla de Mínimos Cuadrados o Regla de Widrow-Hoff y que está muy relacionada con la Regla del Perceptrón. Esta última, ajusta los pesos cuando la respuesta de la neurona es incorrecta y la Regla Delta ajusta los pesos para reducir la diferencia entre la entrada de la red y su respuesta a un patrón y se considera un proceso iterativo. Con esto se logra que la red sea optimizada y pueda generalizar al ser capaz de responder correctamente a entradas similares para las que fue entrenada.

El sistema que usa este algoritmo, el cual aplica una técnica de mínimos cuadrados, fue bautizado por sus creadores como *Adaline* (*ada*ptive *line*ar). Su salida es bipolar, generando un +1 cuando la suma de los pesos es mayor a cero (>0) y -1 cuando la suma de los pesos es igual o menor a cero (≤ 0). Con esto, es posible una clasificación por categorías; por ejemplo, sea A la categoría cuando se tiene +1 y B cuando se tiene un -1. Al igual que con el perceptrón, se aplican las ecuaciones (1.3) y (1.4) para calcular la suma de los pesos. La salida es comparada con la salida deseada, calculando el error entre uno y otro, de la siguiente manera:

Error = <salida deseada> - <salida calculada> .

Cuando el error es calculado, entonces el valor es utilizado para ajustar los pesos del adaline mediante la regla de aprendizaje llamada la Regla Delta, cambiando los pesos de la siguiente manera:

$$w = w_{old} + \frac{\beta E x}{|x|^2} \quad , \tag{1.5}$$

donde es una constante de aprendizaje ($0 \le \beta \le 1$), E es el valor del error , *x* es el vector de entrada y *w* es el peso nuevo. La regla es aplicada iterativamente hasta que se obtenga el valor correcto para un patrón, lo cual indica que no habrá cambio en los pesos e inmediatamente, se introduce un nuevo patrón a aprender con la misma secuencia de la regla. Sin embargo, también se debe verificar si los pesos ajustados reconocen también al primer patrón antes de introducir un tercer patrón, y así sucesivamente.

Regla de Hebb.

Donald Hebb descubrió en 1949 que existían cambios celulares en los tejidos animales, cuando estos aprendían y estableció una regla al respecto, conocida ahora como Regla de Hebb. Esta regla establecía que cuando una neurona estimula a otra neurona estando esta última “encendida”, la conexión entre ambas se reforzaba. La descripción de este comportamiento, fue la clave del descubrimiento del proceso de aprendizaje en modelos biológicos y que fue posible extender a las RNA. Sin embargo, dado que la Regla de Hebb únicamente contemplaba incremento en el reforzamiento sin indicar además la magnitud de este incremento, se hicieron modificaciones a la regla original para adaptarla a los modelos de tal manera que pudieran ser simulados o aplicados. A esta modificación se le llama Regla de Hebb extendida y es la que se describe en la literatura de las RNA’s.

Con esta regla, la salida de la neurona se establece igual que la salida deseada:

$$y = t \ ,$$

donde *t* es la salida deseada, y los pesos se ajustan de acuerdo a lo siguiente:

$$w_{new} = w_{old} + x\, y \ , \tag{1.6}$$

donde w_{new} es el peso nuevo, w_{old} es el peso anterior, *y* es la salida de la neurona y *x* es el vector de entrada.

Los tres algoritmos anteriores, son métodos de entrenamiento de redes de una sola capa usadas para clasificación de patrones y, como se puede observar, las modificaciones son ligeras entre una y otra regla respecto a la modificación de los pesos. El Perceptrón y el Adaline están muy ligadas entre sí y cronológicamente, la Regla de Hebb fue la primera de las tres. Comúnmente, la Regla de Hebb y la Regla Delta son utilizadas también para fines de asociación de patrones mediante memorias asociativas como las descritas en la sección 1.2.1 y cuya arquitectura puede ser de avance (feedforward) o recurrente (iterative). Otro aspecto que vale la pena destacar con los algoritmos anteriores, es el hecho de que se aplican en redes con aprendizaje supervisado, es decir, aquellas redes en las que es necesario tener de antemano la respuesta deseada o tener una retroalimentación.

Sin embargo, los algoritmos de aprendizaje no están limitados a la supervisión ya que también es posible tener algoritmos aplicados a redes no supervisadas, como los Mapas Auto-Organizados de Kohonen (SOM), los cuales caen dentro de la categoría de aprendizaje competitivo, donde se aplica una característica del cerebro, como lo es la inhibición lateral.

Los Mapas Auto-Organizados, son redes que modifican el peso de interconexión entre las neuronas, basadas en las características del patrón de entrada para elegir mediante competencia a solo una de las neuronas con las que está formada la capa, para ser entrenada modificando sus pesos y hacer una clasificación del patrón presentado; uno de los métodos de competencia más populares es el llamado "Ganador Toma Todo" (*Winner Take All* -WTA). En este tipo de redes, que contienen más de una capa, el arreglo puede ser en una dimensión o en dos dimensiones y se consideran "vecindades" en cada una de las neuronas que forman la capa; dependiendo de la cercanía o la lejanía de las demás con respecto a una de ellas, se tendrá excitación o inhibición, respectivamente, de los pesos de interconexión con respecto a la neurona de referencia, de tal forma que sólo una tendrá un valor de peso mayor a todos los demás.

La forma como se cambian los pesos para este tipo de redes, es como sigue:

$$y = \begin{pmatrix} +1, & \text{si} \quad I = \sum_{i=1}^{n} w_i x_i \text{ es grande} \\ 0, & \text{cuando no es así} \end{pmatrix},$$

$$w_{new} = w_{old} + \alpha\left(x - w_{old}\right) \quad , \tag{1.7}$$

donde x es el vector de entrada, w_{new} es el peso nuevo, w_{old} es el peso anterior y α es la razón de aprendizaje, cuyo valor disminuye conforme avanza el aprendizaje. La tabla 1.1 resume los cuatro algoritmos presentados.

De lo anterior, se puede ver entonces la influencia de los pesos en el funcionamiento de las RNA para cumplir con el objetivo de almacenamiento y aprendizaje.

1.4. Elementos y configuraciones usados como sinapsis y neuronas artificiales.

Uno de los principales atractivos de las redes neuronales es su capacidad de aprendizaje gracias a la posibilidad de adaptación sináptica de acuerdo a un algoritmo que puede ser supervisado como el de Hopfield, o no supervisado, como el de Kohonen. En cuanto a las RNA's es necesario adquirir un compromiso en cuanto a la precisión requerida y el área utilizada por el circuito adaptado como sinapsis. Esto se debe a que si se requiere una alta resolución en el valor del peso de interconexión, implica tener una electrónica de gran precisión que se vuelve compleja y en consecuencia, ocupa mucha área de silicio y por lo tanto, es una alternativa cara. Esta opción corresponde a los circuitos digitales, que permiten tener pesos con una alta precisión. Tal podría ser el caso de su utilización en el algoritmo de retropropagación, donde se requiere una resolución para el peso de al menos 8 bits en un problema de interés práctico.

Por otro lado, existen algoritmos que permiten un alto porcentaje de tolerancia al error o imprecisión y es suficiente con usar circuitos analógicos, cuya característica es de tener una precisión moderada pero con la gran ventaja de ocupar menor espacio de integración en el silicio. Por lo tanto, dependiendo del problema a resolver, se utilizan sinapsis digitales o analógicas.

Tabla 1.1. Algoritmos de Aprendizaje.

Red	Algoritmo
Perceptrón	$I = \sum_{i=1}^{n} w_i x_i$ $y = \begin{pmatrix} +1, & \text{si } I \geq T \\ -1, & \text{si } I < T \end{pmatrix}$ $w_{new} = w_{old} + \beta\, y\, x$ $\beta = \begin{pmatrix} +1, & \text{si la respuesta del perceptrón es correcta} \\ -1, & \text{si la respuesta del perceptrón es incorrecta} \end{pmatrix}$
Adaline	Error = <salida deseada> - <salida calculada> $w = w_{old} + \frac{\beta E x}{\lvert x \rvert^2}$
Hebb	$y = t$ $w_{new} = w_{old} + x\, y$
SOM	$y = \begin{pmatrix} +1, & \text{si } I = \sum_{i=1}^{n} w_i x_i \text{ es grande} \\ 0, & \text{cuando no es asi} \end{pmatrix}$ $w_{new} = w_{old} + \alpha(x - w_{old})$

1.4.1. Elementos de resistencia variable.

Como se explicó en la sección 1.1, el peso de interconexión se representa por un resistencia de conductancia T_{ij} (ver ecuación 1.1 y Fig. 1.5). Para tener la posibilidad de aplicar un algoritmo de aprendizaje que simule el comportamiento del cerebro, es necesario que dicha conductancia (resistencia equivalente) sea variable a cada paso de entrenamiento hasta que tome un valor estable que indica el final de la etapa de entrenamiento. Esto ha llevado a la búsqueda de circuitos o elementos que permitan tener una variación de resistencia en un intervalo de voltaje o corriente atractivos para ser utilizados como pesos variables o sinapsis. A continuación se explican algunos de los circuitos y elementos que se han utilizado con la finalidad de tener un resistor variable [1 ,7, 8].

Capacitores conmutados.

Se puede realizar un circuito a base de capacitores conmutados, donde la resistencia de salida (R) depende de la frecuencia del reloj f_{clk} y del valor del capacitor (C) ya que:

$$R = \frac{1}{f_{clk}C} \quad . \tag{1.8}$$

En la Fig. 1.9, se presenta la simulación de un circuito de capacitor conmutado con una f_{clk} = 1 kHz y C = 1 nF y se puede ver que la resistencia que se obtiene con estos parámetros es de aproximadamente 1 M en un intervalo de voltaje de entrada de -5.0 a +5.0 V.

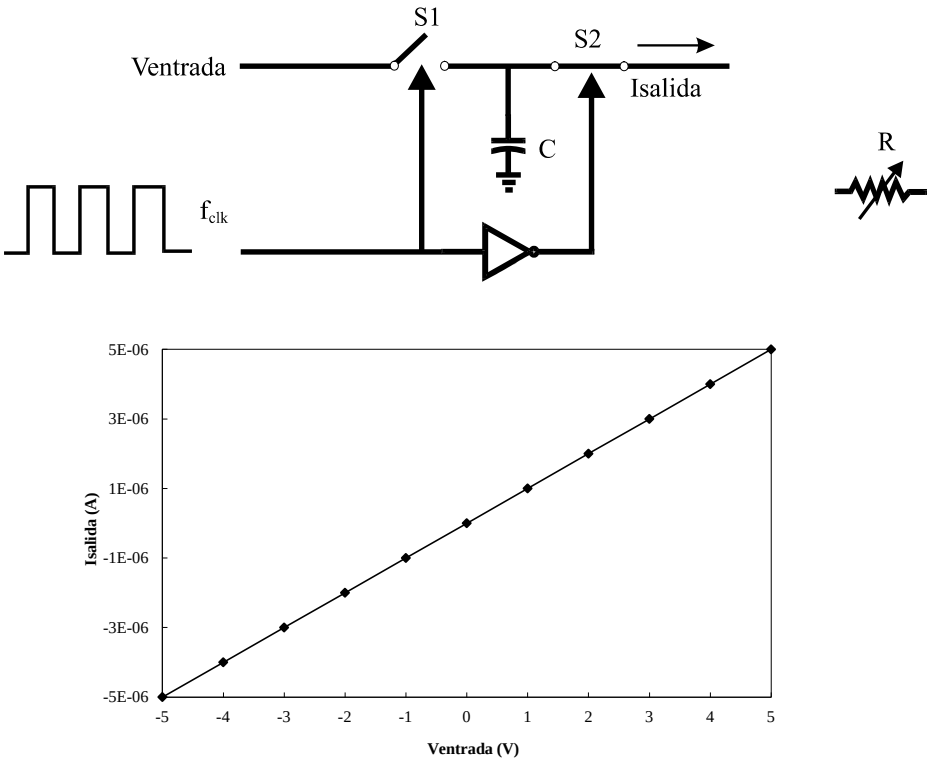


Fig. 1.9. Circuito de capacitores conmutados.

Resistencias conmutadas.

Este circuito consta de una resistencia R_o fija, un interruptor analógico y un capacitor, como se muestra en la Fig. 1.10. El valor de la resistencia de salida se determina como sigue:

$$R = \frac{R_o}{d} \quad , \tag{1.9}$$

donde d es el ciclo útil del interruptor; así, a menor d, R aumenta. Para el ejemplo mostrado en la Fig. 1.10, los valores son: R_o = 1 k y d = 0.5, con lo que se tiene una resistencia de aproximadamente igual a 2 k en un intervalo de voltaje de entrada de -5.0 a +5.0 V.

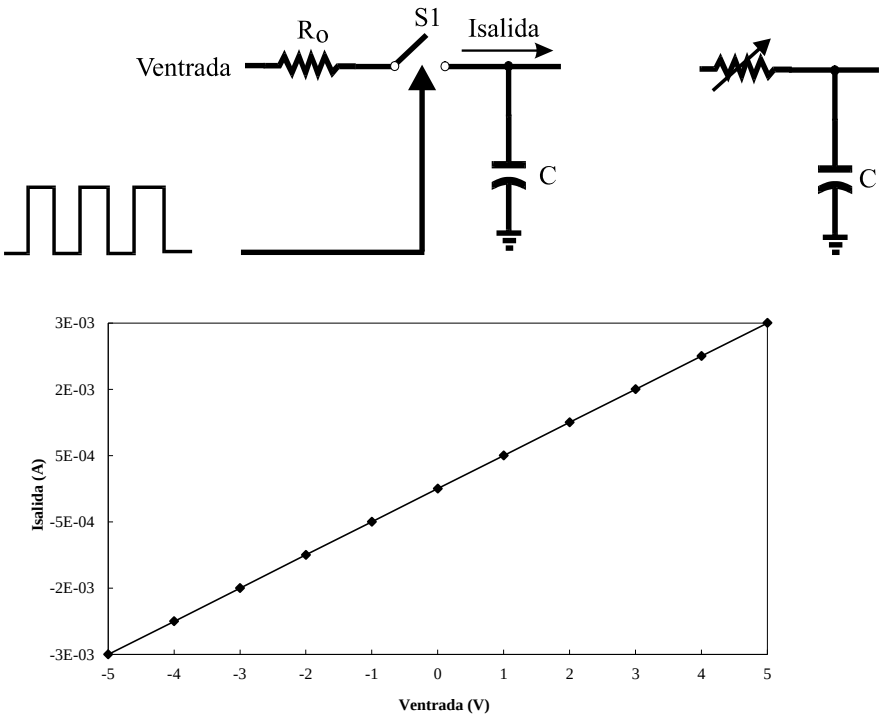


Fig. 1.10. Circuito de resistencia conmutada.

Arreglo en escalera de resistencias conmutadas.

Este arreglo consiste de n resistencias e interruptores en paralelo, como se muestra en la Fig. 1.11. La resistencia total de salida es controlada por los interruptores analógicos y, en consecuencia, se tienen $(2^{n+1} - 1)$ valores posibles en un intervalo que va desde 0 hasta $(2^{n+1} - 1)R_o$.

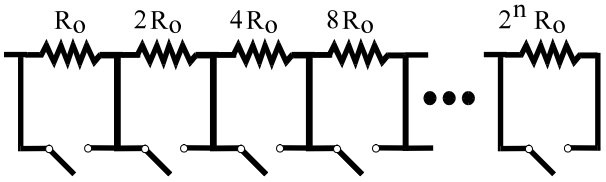


Fig. 1.11. Escalera de resistencias conmutadas.

Resistor controlado por voltaje

Una opción precisa pero que paga el precio de utilizar mucha área y consumo de potencia, es el resistor lineal controlado por voltaje. Este circuito se muestra en la Fig. 1.12 y usa dos transistores FET apareados. El voltaje de compuerta del transistor T2 (V_{G2}) determina la resistencia lineal, junto con el voltaje flotante Vc en la compuerta del transistor T1. La corriente de salida se encuentra de la siguiente manera:

$$I_o = \beta V_{in}\left(V_C - V_{G2}\right) \quad , \qquad (1.10)$$

donde β es el parámetro de transconductancia. Como se puede observar en la Fig. 1.12, la linealidad se presenta en el intervalo de voltajes de entrada, que va desde 0 hasta 2 V, con un valor de resistencia cercano a 1 M .

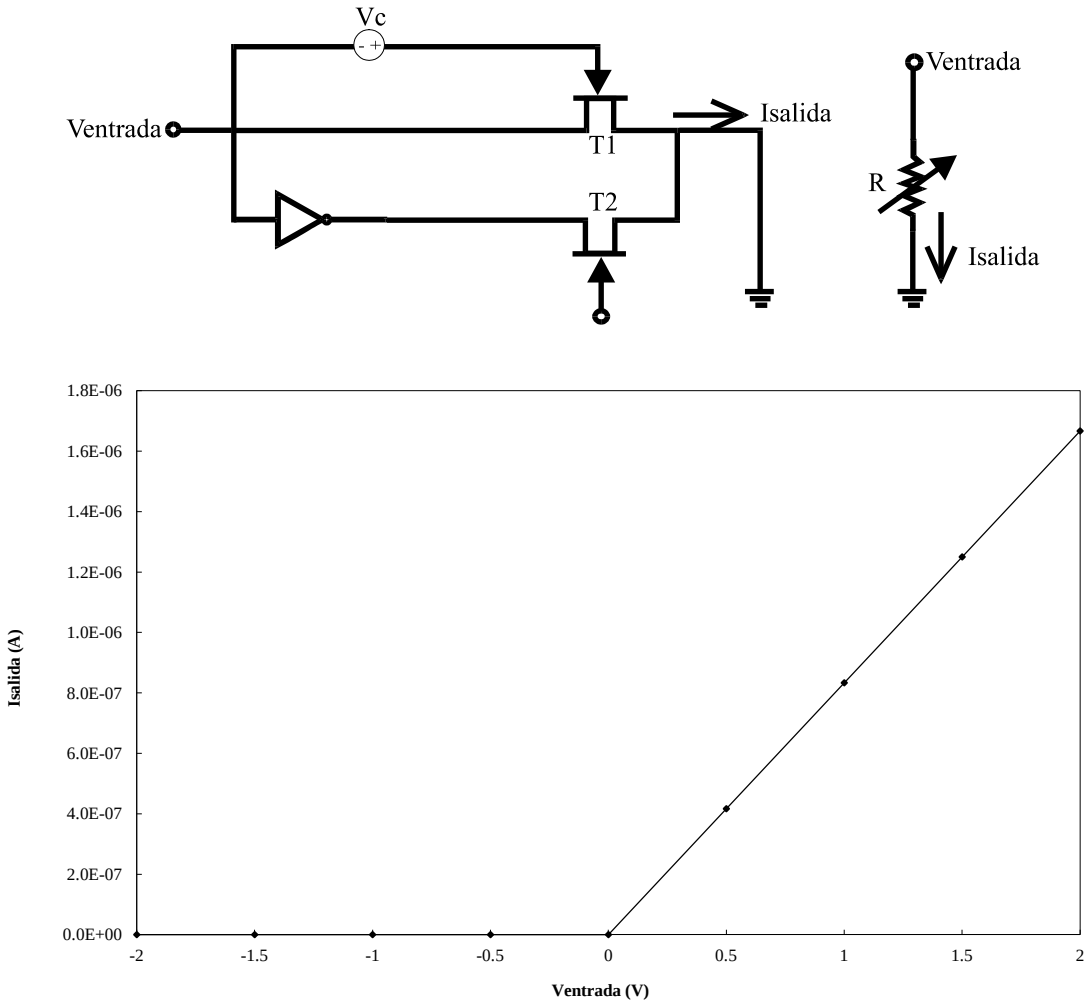


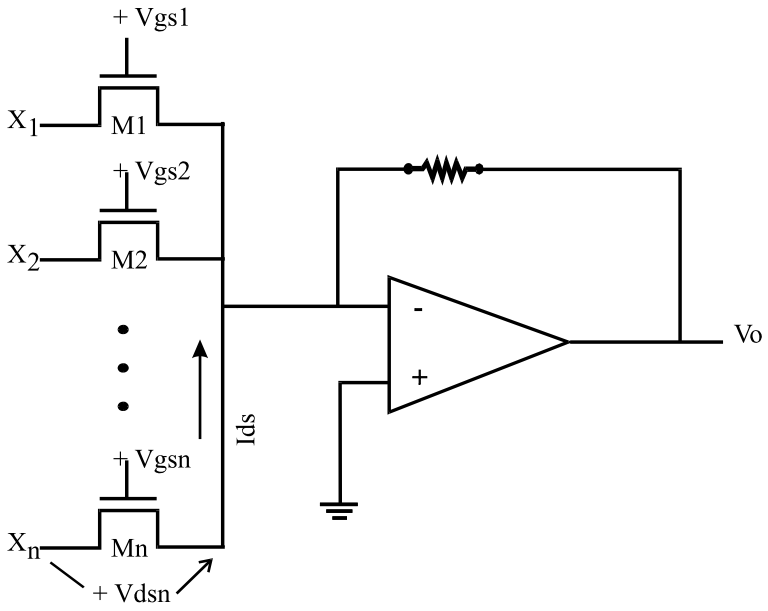
Fig. 1.12. Resistor lineal controlado por voltaje.

Transistor MOS.

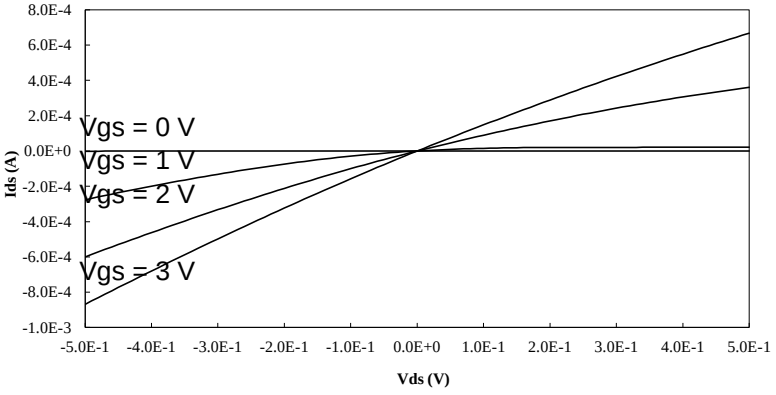
Otro resistor controlado por voltaje puede ser aprovechado a partir del transistor MOS cuando es operado en su región ohmica, donde la resistencia puede ser controlada por medio del voltaje aplicado a la compuerta del transistor; esta corresponde a la resistencia de canal dada por la siguiente ecuación:

$$R_{ds} = \frac{L}{k'W\left(V_{gs} - V_{th}\right)} \quad , \tag{1.11}$$

siendo válida para voltajes pequeños entre drenador y fuente (V_{ds}), donde k' es el parámetro de transconductancia, W es el ancho de canal, L es la longitud de canal, V_{gs} es el voltaje de control aplicado entre compuerta y fuente y V_{th} es el voltaje de umbral del transistor MOS. La Fig. 1.13(a) muestra un circuito en el que cada transistor MOS representa a una resistencia variable, controlada por el respectivo voltaje de compuerta y en la Fig. 1.13(b) se presenta el cambio de la resistencia Rds. En la Fig. 1.13(b), el intervalo de resistencias presentado es de 650 Ω a 340 kΩ.



(a)



(b)

Fig. 1.13. Transistores MOS como resistores controlados por voltaje.

Cuando V_{gs} es menor que V_{th}, el transistor presenta una resistencia muy alta, pero a medida que V_{gs} sea mayor, la resistencia va disminuyendo. Una ventaja de esta configuración, es que además de ser sencilla, la resistencia se puede escalar dependiendo de la razón W/L del transistor. A esto se le puede agregar el hecho de que con este dispositivo, se logra realizar intrínsecamente, el producto del peso por la señal de entrada, como se especifica en la ecuación (1.3), ya que la corriente de salida será el resultado de aplicar un voltaje a través de una resistencia controlada por la compuerta del transistor MOS.

Los circuitos anteriores, permiten cambiar la resistencia de salida mediante una señal de voltaje de control que es calculado externamente (off-line) mediante un algoritmo. Este método no contempla el cambio de los pesos de manera interna (on-line), a medida que el aprendizaje va evolucionando y lo único que se hace, es ajustar la resistrencia sin la posibilidad de almacenar su valor dentro del mismo circuito ni de ser ajustada a cada paso del entrenamiento.

Idealmente, los pesos se deben ajustar internamente al aplicar el algoritmo de aprendizaje, además de tener un medio de almacenar el valor preferentemente en un dispositivo no volátil y que sea fácilmente alterable, es decir, poder ser cambiado y almacenado. Todo lo anterior se debe conjugar con la propiedad de utilizar una área pequeña y ser fácilmente integrable con la tecnología VLSI estándar.

Un dispositivo clásico de almacenamiento de voltaje, es el capacitor que, configurado junto con un arreglo de transistores MOS, como el mostrado en la Fig. 1.14, permite guardar el valor de peso. El voltaje Vc almacenado en el capacitor C, es aplicado a la compuerta del transistor MOS M1 y proviene de un circuito muestreador formado por M2, que funciona como interruptor dependiendo de la señal Vg aplicada en su compuerta para dejar pasar la señal al capacitor. La resistencia se controla con el voltaje Vgs1 de M1.

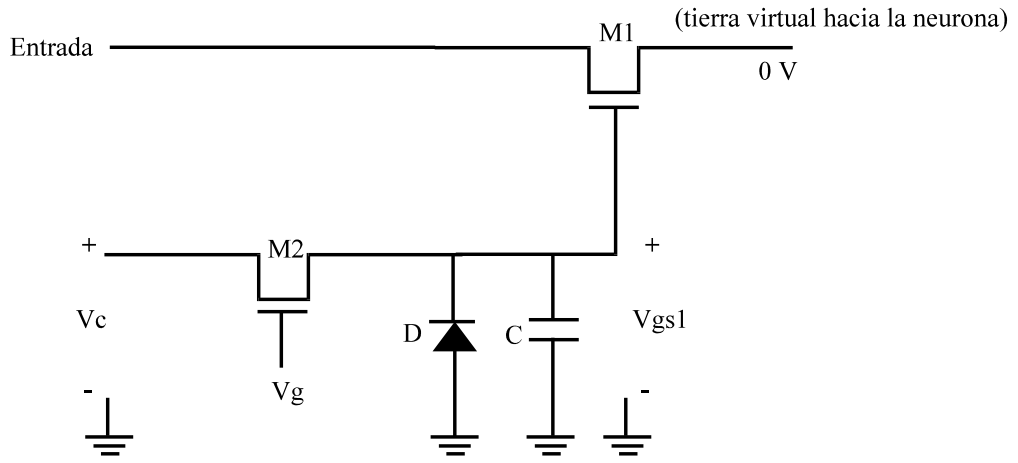


Fig. 1.14. Almacenamiento de pesos con capacitores.

Esta aproximación, sin embargo, tiene limitaciones dada la corriente de polarización inversa de las uniones asociadas al drenador o fuente de los transistores MOS de paso empleados, lo que lleva a diseñar circuitos de refrescamiento para recobrar el nivel del voltaje almacenado, agregando área al circuito. Otra forma de realizar las sinapsis, es mediante circuitos digitales, ya que se tiene la facilidad de programar y almacenar los pesos. Cuando se requiere gran precisión en el aprendizaje, por ejemplo, con el algoritmo de retropropagación o WTA, los pesos se calculan externamente y son programados en la RNA a través de una interfaz digital-analógica y son almacenados en memorias RAM o en los registros de una computadora digital. La Fig. 1.15 muestra un circuito en el cual el voltaje de entrada x_i es descompuesto en componentes binarios (i_1, i_2, ..., i_n) proporcionales a x_i, correspondiendo cada corriente a un peso, que a su vez son sumadas y posteriormente convertidas a voltaje mediante un convertidor corriente-a-voltaje.

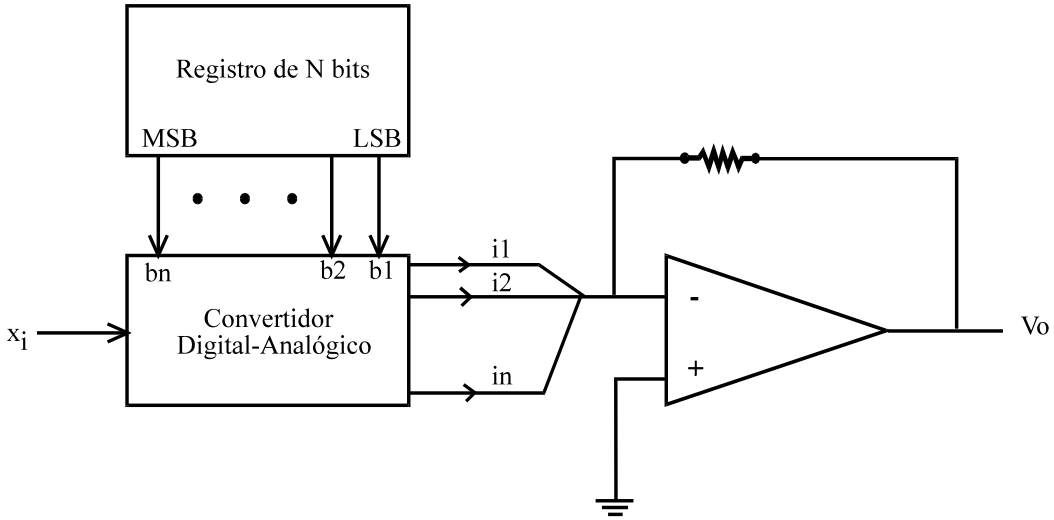
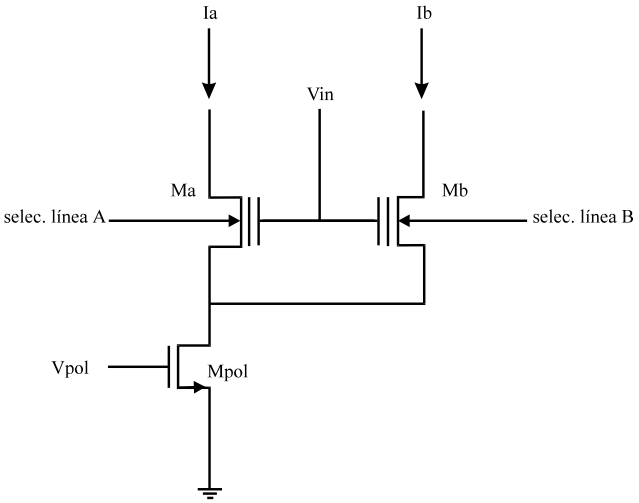


Fig. 1.15. Circuito digital para variación y almacenamiento de pesos.

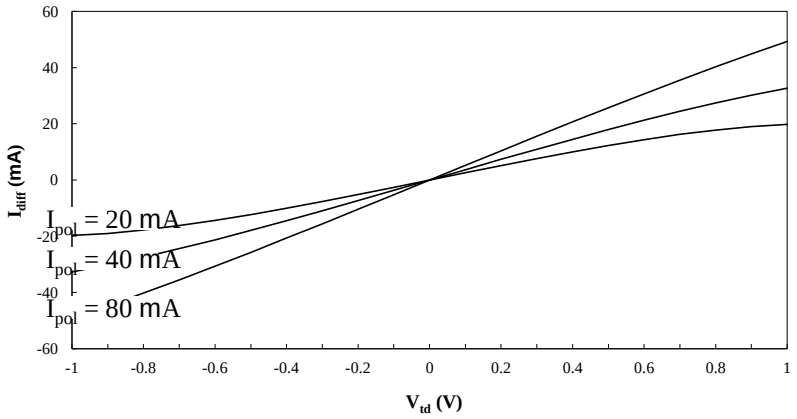
Transistor MOS de compuerta flotante (FGMOS).

Un dispositivo que ha causado mucho interés en años recientes, por su potencialidad en cuanto a la capacidad de almacenamiento y de realizar la multiplicación analógica necesaria para el procesamiento en las redes neuronales, es el transistor MOS de compuerta flotante (FGMOS). Este dispositivo abre la posibilidad de realizar el producto escalar de una manera simple y almacenar el peso, todo en un mismo dispositivo. Por lo tanto, cuenta con tres propiedades deseadas para una sinapsis artificial que son: 1) poder alterar eléctricamente el valor del peso, 2) realizar la multiplicación analógica, y 3) tener la capacidad de almacenamiento analógico. Dado que en la compuerta flotante se almacena carga, de la misma forma como se hace en un capacitor, este dispositivo puede tomar su lugar con la ventaja adicional de que la pérdida de carga llega a ser de 0.1 % en 26 años comparada con la carga originalmente almacenada [4], con la tecnología actual. De ahí el esfuerzo que se ha dedicado al estudio de diferentes estructuras de compuerta flotante para tener la factibilidad de ser incluidas en una RNA de regular complejidad, cumpliendo sobre todo con características como la de tener poca área, consumir baja potencia y ser compatible con tecnologías estándar. El FGMOS, es un transistor MOS que tiene una compuerta flotante, es decir, sin acceso físico a ella por estar aislada completamente por óxido de silicio a su alrededor, y que se encuentra colocada por encima de la región de canal separada por un óxido delgado, y por debajo de la compuerta de control del MOS, separado por un óxido más grueso.

Un ejemplo dentro de la gran variedad de circuitos que aprovechan al FGMOS para implementar una sinapsis integrada, se presenta en la Fig. 1.16, junto con la respuesta I-V de la misma, obtenida a partir de la diferencia de las corrientes de entrada y la diferencia de los voltajes de umbral de los transistores Ma y Mb. El intervalo de resistencia que se obtiene en esta gráfica, con los parámetros utilizados es desde 20 kΩ hasta 45 kΩ.



(a)



(b)

Fig. 1.16. a) Sinapsis que utiliza FGMOS.; b) Respuesta I-V.

Este arreglo consta de tres transistores canal n, donde M_{pol} forma una fuente de corriente cuyo valor depende del voltaje V_{pol}; los transistores Ma y Mb tienen compuerta flotante, lo que permite programarlos independientemente, conectados en configuración par diferencial. Gracias a que se puede cambiar por separado el voltaje de umbral de cada transistor mediante inyección por avalancha, las corrientes Ia e Ib pueden ser diferentes, característica que de no ser transistores programables, no podría ser así. La programación de estos transistores se hace con ayuda del voltaje de entrada, Vin, los voltajes de sustrato, *selec. línea*, y el voltaje de drenador en cada uno. De esta manera, una vez programados, el peso efectivo es proporcional a la diferencia de las corrientes Ia e Ib, que a su vez dependen del voltaje de umbral existente en el transistor respectivo:Vta es el voltaje de umbral del transistor Ma y Vtb es el voltaje de umbral del transitor Mb. Estas corrientes se expresan analíticamente, de la siguiente manera:

$$I_{a,\,b} = \frac{I_{pol}}{2} \pm V_{td}\sqrt{\frac{\beta I_{pol}}{2}}\sqrt{\left(1 - \frac{\beta V_{td}^{\,2}}{4I_{pol}}\right)} \quad , \tag{1.12}$$

donde:

$$Vt_d = Vt_b - Vt_a \quad , \tag{1.13}$$

$$\beta = \frac{\mu_n \varepsilon_{ox}}{2t_{ox}} \frac{W}{L} \quad , \tag{1.14}$$

μ_n : movilidad de los electrones.

ε_{ox} : permitividad del óxido de silicio.
t_{ox} : espesor del óxido de compuerta.
W : ancho del canal del transistor MOS.
L : longitud del canal del transistor MOS.
I_{pol} : corriente a través de M_{pol}.

La ecuación (1.12) representa la corriente cuando los transistores Ma y Mb se encuentran en la región de saturación. Como se mencionó anteriormente, el peso corresponderá a la diferencia de las corrientes Ia e Ib, que se expresa de la siguiente manera:

$$I_{diff} = Ia - Ib = V_{td}\sqrt{\beta I_{pol}}\sqrt{\left(1 - \frac{\beta V_{td}^2}{4I_{pol}}\right)} \quad . \tag{1.15}$$

Esta ecuación es la que se utiliza para obtener de forma analítica las curvas que se presentan en la Fig. 1.16(b) y se observa que el intervalo de cambio de la diferencia de voltaje de umbral va desde -1 V hasta 1 V, lo que se podría considerar un intervalo muy reducido y la resistencia de salida depende de la corriente I_{pol}.

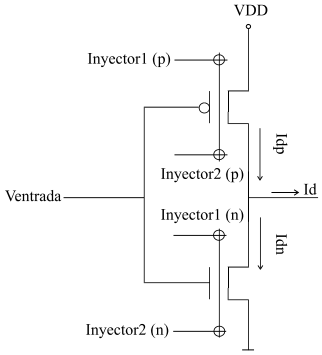
Otro circuito que es usado como sinapsis analógica, se puede construir en base a un inversor CMOS, como el mostrado en la Fig. 1.17, donde ambos transistores incluyen en su estructura, una compuerta flotante, también con la posibilidad de ser programados independientemente para cambiar el voltaje de umbral en cada transistor mediante inyección o extracción de carga por tunelamiento Fowler-Nordheim, utilizando los dos inyectores que tiene cada transistor. Mediante una polarización adecuada entre los inyectores y la compuerta de control, se crea un campo eléctrico de tal magnitud y sentido que es posible inyectar o extraer electrones de la compuerta flotante de una manera controlada, para cambiar el voltaje de umbral.

La magnitud y dirección de la corriente depende del voltaje de umbral de los transistores y del voltaje de entrada. El intervalo de resistencias que cubre el arreglo de la Fig. 1.17(b) va desde 5 kΩ hasta 60 kΩ.

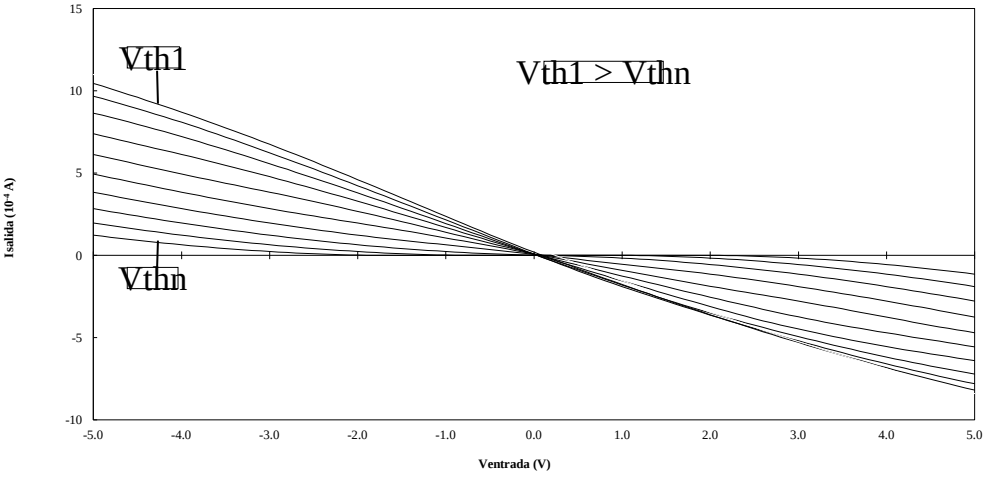
El comportamiento del inversor se rige por varias ecuaciones, dependiendo de la región de operación en la que se encuentran los transistores, como se indica en la tabla 1.2.

Tabla 1.2. Regiones de operación del inversor CMOS.

Región	Condición	PMOS	NMOS
A	$0 \le Vin \le Vth_n$	lineal	corte
B	$Vth_n \le Vin \le Vo+Vth_p$	lineal	saturación
C	$Vo+Vth_p \le Vin \le Vo+Vth_n$	saturación	saturación
D	$Vo+Vth_n \le Vin \le V_{DD}+Vth_p$	saturación	lineal
E	$V_{DD}+Vth_p \le Vin \le V_{DD}$	corte	lineal



(a)



(b)

Fig. 1.17. a) Sinapsis utilizando un inversor CMOS con transistores de compuerta flotante, b) respuesta I-V.

Los diferentes parámetros especificados en la tabla 1.2, son los siguientes:

Vin : Voltaje de entrada al inversor.
V_{DD} : Polarización del inversor.
Vo : Voltaje de salida del inversor.
Vth_n : Voltaje de umbral del NMOS.
Vth_p : Voltaje de umbral del PMOS.

La corriente total a la salida del inversor, Id, es la suma de la corriente que pasa por el transistor canal N y la corriente que pasa por el transistor canal P, como se indica en la siguiente ecuación:

$$Id = Id_p - Id_n \tag{1.16}$$

Las ecuaciones que rigen en cada región de operación se expresan como sigue:

Región A:

$$Id = -\beta p\left[\left(Vin - V_{DD} - Vth_p\right)\left(Vo - V_{DD}\right) - \frac{1}{2}\left(Vo - V_{DD}\right)^2\right] \quad , \tag{1.17}$$

Región B:

$$Id_p = -\beta p\left[\left(Vin - V_{DD} - Vth_p\right)\left(Vo - V_{DD}\right) - \frac{1}{2}\left(Vo - V_{DD}\right)^2\right] \quad , \tag{1.18}$$

$$Id_n = \frac{\beta n}{2}\left(Vin - Vth_n\right)^2 \quad , \tag{1.19}$$

Región C:

$$Id_p = -\frac{\beta p}{2}\left(Vin - V_{DD} - Vth_p\right)^2 \quad , \tag{1.20}$$

$$Id_n = \frac{\beta n}{2}\left(Vin - Vth_n\right)^2 \quad , \tag{1.21}$$

Región D:

$$Id_p = -\frac{\beta p}{2}\left(Vin - V_{DD} - Vth_p\right)^2 \quad , \tag{1.22}$$

$$Id_n = -\beta n\left[\left(Vin - Vth_n\right)Vo - \frac{1}{2}Vo\right] \quad , \tag{1.23}$$

Región E:

$$Id = -\beta n\left[\left(Vin - Vth_n\right)Vo - \frac{1}{2}Vo^2\right] \quad . \tag{1.24}$$

La región de mayor interés en el uso de un inversor CMOS como sinapsis, es cuando se tiene la transición y corresponde a las regiones B, C y D, donde se puede ver que la corriente de salida es proporcional a $(Vin-Vth)^2$, lo que justifica la no linealidad de la respuesta, sin embargo a pesar de eso, este tipo de respuesta puede seguir siendo útil como sinapsis.

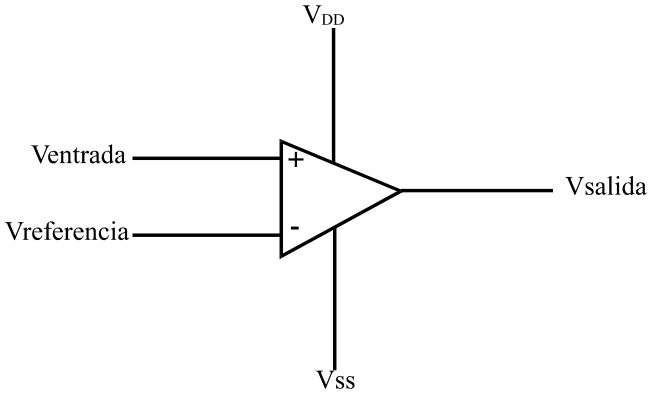
Como se puede ver, en la mayoría de las sinapsis presentadas, se tiene tanto corriente negativa como positiva, lo que permite la inhibición o excitación, respectivamente en el punto de suma de las corrientes, de manera análoga como lo hace una red neuronal biológica. Sin embargo, la última cumple ampliamente con los requerimientos de almacenamiento no volátil para fines de aprendizaje, procesamiento de señal (inhibición o excitación), configuración simple y modificación de conductancia. Por lo tanto, vale la pena analizar este elemento, desde el punto de vista de operación y fabricación, con la finalidad de incluirlo en las RNA.

1.4.2. Elementos de procesamiento.

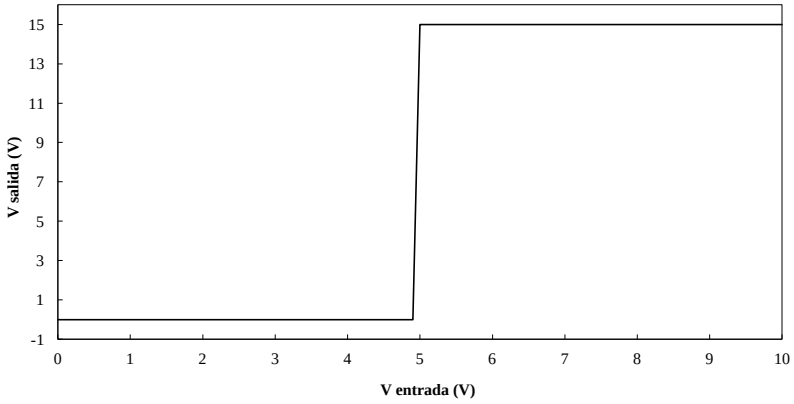
Después del punto donde se suman las corrientes provenientes de varias neuronas, la corriente total debe ser procesada por un elemento que responda según la magnitud de entrada, es decir, permanecer "apagado" si la suma está por debajo de un umbral, o "encendido" si lo sobrepasa. Este elemento está indicado en la Fig. 1.1 y los tipos de funciones que comúnmente se aplican, se presentan en la Fig. 1.2. Es necesario entonces, utilizar configuraciones que realicen cualquiera de las tres funciones básicas mencionadas con anterioridad: a) alta ganancia, b) lineal y c) sigmoidal. Al igual que con los elementos de resistencia variable, la función de transferencia de las neuronas tienen modelos matemáticos precisos de fácil aplicación en simulaciones por computadora, pero la obtención de estas funciones en circuitos integrados sólo es posible de manera aproximada. A continuación se presentan algunos de los circuitos que se reportan en la literatura para realizar tales funciones, según su función de transferencia:

Alta ganancia.

Este tipo de función, se utiliza para configurar memorias asociativas en las Redes de Hopfield, Redes de Hamming y Máquinas de Boltzman. Una manera sencilla de lograr esta función es mediante un comparador analógico, como el que se muestra en la Fig. 1.18(a), donde el voltaje de umbral depende del voltaje de referencia Vref.



(a)



(b)

Fig. 1.18. a) Comparador analógico de voltaje, b) respuesta del comparador.

La ecuación que expresa a esta función se indica a continuación.

$$f(x) = \begin{cases} 1 & x \geq \theta \\ 0 & x < \theta \end{cases} , \qquad (1.25)$$

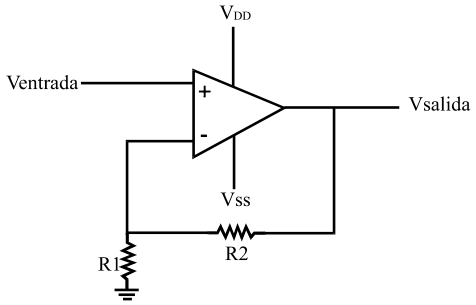
donde θ es el umbral deseado.

Lineal.

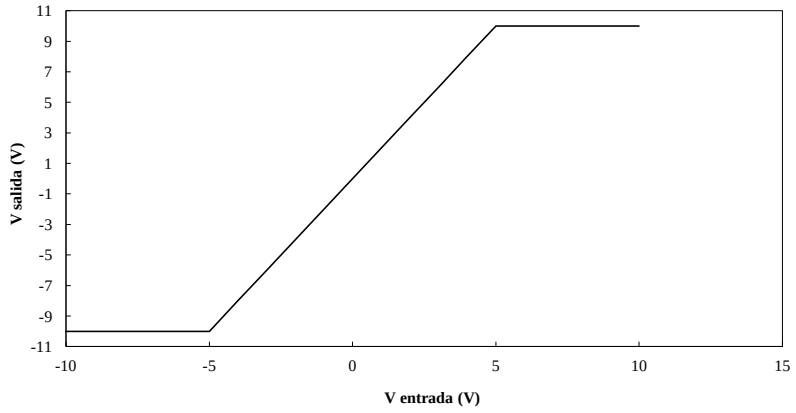
La función de activación lineal se usa comúnmente en redes de capa sencilla de avance como el ADALINE y el Perceptrón. La Fig. 1.19(a) muestra un circuito con el que se puede conseguir este tipo de función, mediante un amplificador operacional no inversor, cuya ganancia está determinada por las resistencias R1 y R2, de la siguiente manera:

$$\frac{V_{sal}}{V_{ent}} = \frac{R_1 + R_2}{R_1} \quad . \qquad (1.26)$$

Los límites superior e inferior, se deben a la saturación del circuito, que depende de la magnitud de la polarización que se use en V_{DD} y en V_{SS}.



(a)



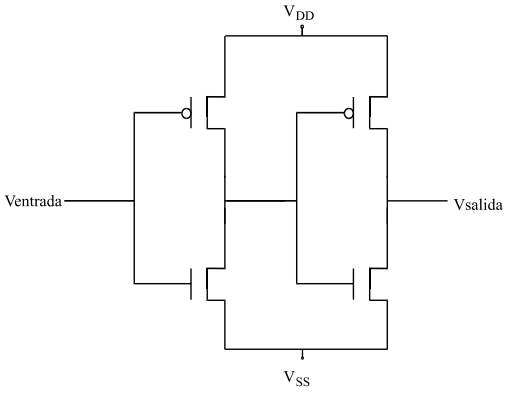
(b)

Fig. 1.19. a) Circuito con función de transferencia lineal, b) función de transferencia lineal.

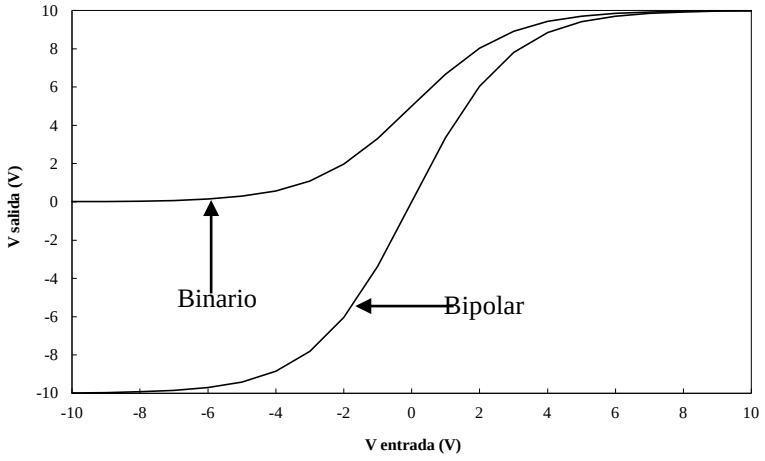
En el caso de la Fig. 1.19, la función de transferencia tiene un umbral fijo, pero existen circuitos con los que se puede ajustar el umbral de la función, es decir, desplazar la gráfica hacia la derecha o hacia la izquierda, según sea necesario para la operación de la red.

Sigmoidal.

En las Redes de Hopfield, también se puede utilizar la función de transferencia sigmoidal, así como en las redes asociativas multicapa de avance o en redes de retropropagación. En este tipo de redes, se requiere una función creciente y diferenciable para poder aplicar la regla delta y esto se consigue con la función sigmoidal. En la Fig. 1.20 (a) se muestra un circuito sencillo y de fácil integración, a base de dos inversores CMOS en serie, con los que se logra la sigmoide.



(a)



(b)

Fig. 1.20. a) Circuito para obtener la función sigmoidal, b) Respuesta del circuito.

En la Fig. 1.20(b) se presentan los dos tipos de función de transferencia que se pueden utilizar cuando se considera a una sigmoide: a) binaria y b) bipolar. A continuación se presentan las expresiones analíticas de cada una de ellas.

$$f(x) = \frac{1}{1 + \exp(-\sigma x)} \quad , \qquad \text{(binario)} \qquad (1.27)$$

$$f(x) = \frac{2}{1 + \exp(-\sigma x)} - 1 \ , \qquad \text{(bipolar)} \qquad (1.28)$$

donde σ es el parámetro de ganancia de la función alrededor del origen y el límite superior e inferior también están dados por la polarización V_{DD} y V_{SS}.

1.5. Sumario.

En este capítulo, se revisaron algunos conceptos fundamentales de las Redes Neuronales Artificiales, como una analogía de las redes biológicas, en las que son representadas como varias interconexiones alimentando a un elemento de procesamiento. Las señales que entran a este último, pueden llegar a ser inhibitorias o excitatorias, dependiendo del sentido que tenga el parámetro de la información, que en este caso, es una corriente. La corriente dependerá a su vez de la resistencia de interconexión, llamada sinapsis, y de la magnitud de la señal proveniente de la neurona alimentadora.

Se explicaron también, las diferentes clasificaciones de redes que existen, separándolas tanto por tipo de arquitectura como por algoritmos y se presentaron algunas de las mas usadas como las redes de capa sencilla, multicapa o competitivas, desde el punto de vista de arquitectura, o bien, supervisadas o no supervisadas, desde el punto de vista de aprendizaje, donde se aplican diferentes tipos de algoritmos como la Regla de Hebb, el Perceptrón, Adaline y Mapas Auto-Organizados.

Dada la importancia de la programabilidad de los pesos con la finalidad de poder aplicar un algoritmo de aprendizaje, algunos de los circuitos y elementos con los que se puede tener un comtrol en el cambio de la magnitud de la resistencia, son destacados en las diferentes aproximaciones que se han desarrollado en la búsqueda de encontrar una sinapsis que cumpla con los objetivos que plantea la integración VLSI de elementos con poca área, precisión y almacenamiento. De estos, el FGMOS presenta mayores ventajas con respecto a los demás por la propiedad de conjuntar en un mismo dispositivo, la realización de la multiplicación escalar y el almacenamiento del peso, lo que lo hace un dispositivo muy atractivo para su utilización como sinapsis en Redes Neuronales Artificiales integradas. También se presentan algunos circuitos empleados como elementos de procesamiento con los que se puede implementar las funciones de alta ganancia, lineal y sigmoide.

Lo presentado en este capítulo, sirve como antecedente de lo que se explicará en el siguiente capítulo, donde se hará referencia a las diferentes estructuras de compuerta flotante que se reportan en la literatura, como elemento de las RNA. Se hace una comparación de estas estructuras desde el punto de vista de problemas de programación y borrado, diseño topológico y métodos de programación.

Referencias.

1.- H. P. Graf and L. D. Jackel, “Analog electronic neural network circuits”, *IEEECircuits and Devices Magazine*, Vol. 5, pp. 44-45, Jul. 1989.

2.- M. Holler, S. Tam, H. Castro and R. Benson, “An electrically trainable artificial neural network (ETANN) with 10 240 floating-gate sinapses”, *Proc. IJCNN, Vol. II* (Washington, D. C.), Jun. 1989, pp. 191-196.

3.- T. C. Ong, P. K. Ko and C. Hu, “ The EEPROM as an analog memory device”, *IEEE Trans. Electron Devices*, Vol. 36, pp. 1840-1841, Sept. 1989.

4.- A. Thomsen and M. A. Brooke, “A floating-gate MOSFET with tunneling injector fabricated using a standard double-polysilicon CMOS process”, IEEE Electron Device Letterws, Vol. 12, No. 3, pp. 111-113, March 1991.

5.- E. Säckinger and W. Guggenbühl, “An analog trimming circuit based on a floating-gate device”, *IEEE Jour. of Solid State Circuits*, Vol. 23, No. 6, Dec. 1988, pp. 1437-1440.

6.- L. Fausett, *Fundamentals of Neural Networks. Arquitectures, Algorithms and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1994.

7.- J. M. Zurada, “Analog implementation of neural networks”, *Circuits and Devices Magazine*, pp. 36-41, Sept. 1992.

8.- S. Y. Foo, L. R. Anderson and Y. Takefuji, “Analog components for the VLSI of neural networks”, *Circuits and Devices Magazine*, pp. 18-26, Jul. 1990.

9.- M. Caudill and C. Butler, *Understanding Neural Networks. Computer Explorations. Vol. I: Basic Networks*, The MIT Press, Cambridge, Mass., 1992.

CAPITULO 2.

Estructuras de compuerta flotante.

2.1. Comparación de estructuras de compuerta flotante

El desarrollo de redes neuronales artificiales viene demandando la implementación de circuitos VLSI costeables, para la modificación del peso de conexión asociado con cada *sinapsis*, dada la gran cantidad requerida para interconectar a las neuronas. Esto se debe al objetivo de tener una red neuronal integrada, que cumpla lo más parecido posible con las funciones realizadas por el cerebro, como aprendizaje, memoria, reconocimiento de patrones, toma de decisiones, etc. Para esto, el cerebro cuenta con alrededor de 100 billones de neuronas, cada una de ellas interconectada con aproximadamente 10 000 neuronas; evidentemente, tal cantidad se ve reducida en la tecnología de integración dada la limitación en dos dimensiones de los circuitos fabricados en silicio. En la actualidad, uno de los circuitos neuronales más complejos, consta de una red de 1000 neuronas, lo que involucra alrededor de 1 millón de interconexiones [1, 2]. Por lo tanto, la principal preocupación de la tecnología, es tratar de alcanzar un escalamiento tal, que el área de los circuitos integrados se optimice, logrando un compromiso entre la complejidad de los circuitos y el tamaño del chip. Sin embargo, el escalamiento hacia menores dimensiones, trae como consecuencia la aparición de fenómenos que no se presentan en la tecnología de canal largo de los **MOSFET's** (dispositivo base de la sinapsis), y que hace necesario el desarrollo de nuevos modelos para la simulación de los dispositivos.

Esto último, junto con el desarrollo de tecnología de **VLSI** es el trabajo que ocupa a muchos investigadores dentro del área de las redes neuronales. De algunas de estas investigaciones, se deriva la llamada memoria **FLASH EEPROM (FEEPROM)**, llamada así, porque puede ser borrada completamente o en partes, con pulsos eléctricos. Estas memorias constan de transistores MOS que utilizan una compuerta flotante, en la cual se almacena o se elimina carga, para cambiar el voltaje de umbral y programar o borrar un valor deseado (cambio del peso).

Las FEEPROM ocupa un lugar entre las **UV-EPROM** y las **EEPROM** [3], con la propiedad de poder ser borradas eléctricamente como las EEPROM y ser tan densas y baratas como las UV-EPROM. Cuentan también con un tiempo de borrado de pocos segundos, comparado con el de las UV-EPROM, que es de alrededor de 20 minutos. Además, el tamaño es 20-30 % mayor que estas últimas, con lo que el costo por bit es comparable en ambas. Esto se debe a que el aumento en área, se ve compensado porque el encapsulado no requiere una ventana de cuarzo, como las UV-EPROM.

En términos de funcionamiento, las memorias FEEPROM deben tener corrientes de lectura, tiempos de programación y fuentes de alimentación, similares a las UV-EPROM, para complementar su bajo costo.

Respecto al mecanismo de inyección de carga, se emplean estrategias como la de evitar recurrir al borrado por el mecanismo de huecos calientes que causa grandes corrientes de substrato, ya que esto causa un aumento indeseado en el tiempo de borrado, por lo que se usa el **tunelamiento Fowler-Nordheim (FN)**. Por otro lado, con el fin de simplificar el proceso, la escritura se lleva a cabo por **inyección de electrones calientes de canal**, lo que evita el uso de un óxido especial de tunelamiento, como lo usan las EEPROM. También se consideran las magnitudes de las concentraciones de impurezas implantadas bajo el óxido delgado del inyector o la alineación de las ventanas del óxido delgado.

Se pueden encontrar en la literatura, muchos desarrollos de memorias de compuerta flotante, en los cuales se trata de optimizar el funcionamiento del dispositivo y se trata de reducir el área ocupada por este, para la integración en un circuito específico dentro del campo de las RNA. En esta sección, se señala el estado del arte de las memorias de compuerta flotante.

2.1.1. Algunos criterios para la construcción de EEPROM de compuerta flotante.

Es conveniente tomar en cuenta ciertos aspectos prácticos dentro de la fabricación y funcionamiento de las memorias de compuerta flotante para tener las mejores características eléctricas y aprovechar recursos tecnológicos existentes. Se puede mencionar el caso reportado por Concannon et. al. [4], donde se construyó una memoria de compuerta flotante (ver Fig. 2.1), utilizando un óxido de tunelamiento de aproximadamente 10 nm y un óxido de compuerta de 50 nm.

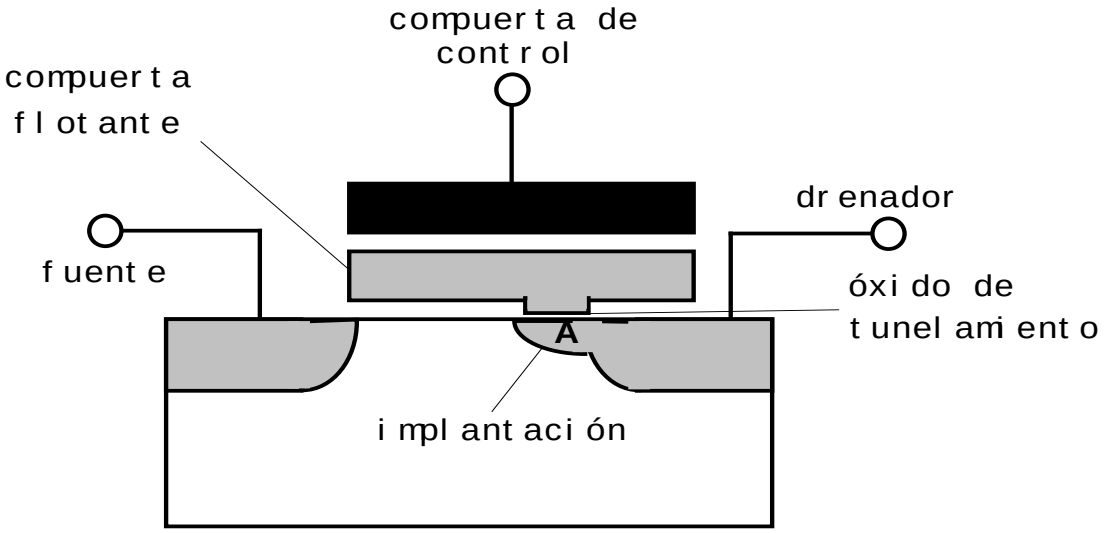


Fig. 2.1. Corte transversal de la EEPROM de compuerta flotante.

La región A ilustrada en la Fig. 2.1, se refiere a la concentración superficial bajo la región de tunelamiento, la cual es una implantación y que tiene una magnitud inferior a la concentración del drenador, para minimizar los problemas relacionados al crecimiento de óxido de silicio en regiones de alta concentración.

En este caso particular, tanto la programación como el borrado de la memoria, se hacen a través de tunelamiento FN. Para programar, se aplica una rampa de voltaje a la compuerta de control con el drenador conectado a tierra y la fuente abierta, para inyectar electrones del drenador a la compuerta flotante. Para borrar, es decir, remover los electrones de la compuerta flotante, se aplica una rampa de voltaje al drenador, con la compuerta de control aterrizada.

La expresión que caracteriza la corriente hacia la compuerta flotante, por inyección FN es:

$$Jt = A\mathcal{E}^2 \exp\left(-\frac{B}{\mathcal{E}}\right) \quad , \qquad (2.1)$$

donde A y B son constantes y son función de la masa efectiva y altura de la barrera de los electrones y es el campo eléctrico a través del óxido de tunelamiento.

Si se requiere modelar el tunelamiento al programar y borrar la memoria, se deben utilizar valores diferentes de A y B para cada proceso, ya que para el primero, la superficie emisora es una interfaz óxido de silicio/silicio monocristalino y en el segundo caso, la superficie emisora, es una interfaz óxido de silicio/silicio policristalino (material con el que se fabrica la compuerta flotante).

Durante la operación del borrado, se puede presentar un comportamiento anormal, caracterizado por un pico, como se muestra en la Fig. 2.2. En esa región, el óxido de tunelamiento se ve sujeto a un gran esfuerzo por la alta corriente, teniéndose el peligro de que se presente una ruptura prematura del óxido. Este pico es notable cuando la concentración de impurezas bajo el óxido de tunelamiento es bajo (~$5x10^{18}$ cm^{-}), sin embargo, si se aumenta la concentración ($5x10^{19}$ cm^{-3}), esta región de alta corriente, desaparece.

El pico anómalo durante el borrado, se debe a la recuperación de una condición de *deserción profunda* presente en ese estado, mediante la generación por tunelamiento *banda-banda* de los portadores bajo el óxido de tunelamiento. Cuando la concentración superficial de impurezas se aumenta a tal nivel, que se evite la presencia de la condición de deserción profunda, la anomalía no se presenta.

Por otro lado, la alineación de óxido de tunelamiento con respecto a la implantación, también presenta consecuencias en el caso de no ser realizada de una manera óptima. En la referencia [4], se demuestra que un alineamiento desplazado 0.2 μm del lugar adecuado, provoca un aumento de la corriente de substrato, en tres órdenes de magnitud, con lo que se ve reducida la eficiencia de la memoria.

De todo lo anterior, se pueden deducir los siguientes criterios para la fabricación de la memoria.

1. Evitar concentraciones superficiales mayores a 10^{20} cm^{-3} en la implantación de la zona de tunelamiento para no inducir efectos indeseables durante el crecimiento del óxido de tunelamiento en la zona del inyector (< $5x10^{19}$ cm^{-3}).

2. Tener una concentración superficial lo suficientemente alta ($5x10^{19}$ cm^{-3}) en la región de tunelamiento, para evitar que se tenga deserción profunda, que al relajarse, provoca un rápido aumento del campo eléctrico con la latente posibilidad de reducir el tiempo de vida útil del óxido de tunelamiento.

3. Alinear adecuadamente el óxido de tunelamiento con la implantación, con la finalidad de mantener baja la corriente de substrato.

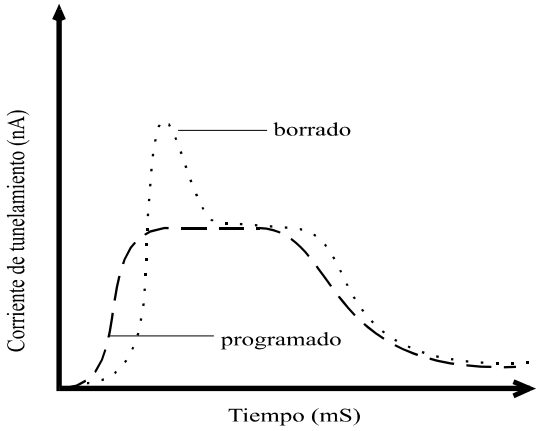


Fig. 2.2. Corriente de tunelamiento con borrado y programación mediante una rampa de 10^{4} V/s y una concentración de $5x10^{18}$ cm^{-3} en la región A.

Todo lo anterior, tiene como fín el resolver problemas detectados durante transitorios ocurridos en la medición de las características de las memorias.

Otros parámetros de importancia en el funcionamiento de las FEEPROM, que son estudiados con la finalidad de optimizarlos dentro de un compromiso *costo-área-funcionamiento*, son: densidad de integración, corrientes de lectura, tiempos de programación y de borrado. Estos fueron estudiados por Amin [3], quien propone varias estructuras y su caracterización, con la finalidad de tener elementos objetivos de

juicio, que sirvan para seleccionar alguna de ellas, aprovechando las mayores ventajas de rendimiento, para un circuito en particular.

Para comenzar, la estrategia utilizada por Amin, es no recurrir al mecanismo de borrado por huecos calientes, para evitar grandes corrientes de substrato, por lo que se utiliza únicamente el borrado por tunelamiento FN. Además, para evitar el uso de un óxido especial de tunelamiento, como en las EEPROM, se programan las memorias de la misma manera que las EPROM, esto es, con electrones calientes de canal [5], lo que simplifica el proceso de fabricación. En este momento cabe comentar una diferencia importante entre el borrado por rayos UV, que es muy lento y el borrado eléctricamente, que toma pocos segundos: el primero es *autolimitado*, es decir, una vez removidos los electrones, no existe ningún fenómeno adicional, a comparación del segundo, con el que dependiendo del voltaje y el tiempo, se puede tener un borrado inapropiado, ya sea incompleto o en exceso, donde la carga no fue completamente removida, o dejar una carga neta positiva, respectivamente. Esto también puede ser evitado, con consideraciones de diseño adecuadas.

Fueron cinco las estructuras fabricadas, abarcadas estas en dos grupos: Tipo 1, aquellas en las que el polisilicio 1 (arriba del semiconductor) forma la compuerta flotante y el polisilicio 2 (arriba del polisilicio 1), que forma la compuerta de control; y Tipo 2, donde se invierte la colocación de las terminales, es decir, el polisilicio 1 corresponde a la compuerta de control y el polisilicio 2 a la compuerta flotante. En todas ellas la inyección o extracción de carga se hace entre la compuerta flotante y drenador o fuente. Dentro de las celdas Tipo 1, se tienen la Celda A y la Celda B; no se indica el área de cada celda. A continuación se explica su estructura.

Celda A.

Es parecida a las EPROM, sin embargo la programación se hace a través del drenador, mientras que el borrado se hace por la fuente. La unión de la fuente se difunde de tal forma que incrementa el voltaje de ruptura de diodo correspondiente, y por otro lado la unión del drenador se implanta independientemente para hacer más eficiente la programación.

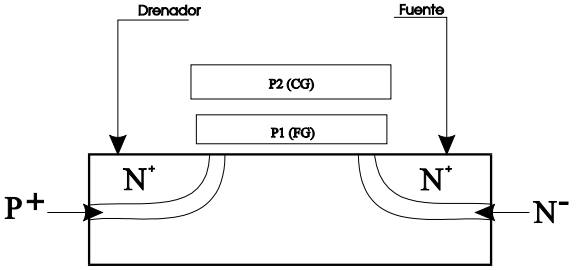


Fig. 2.3. Celda A.

Celda B.

Este tipo de celdas, evitan el problema de borrado excesivo, problema que se presenta con las celdas tipo A. En la celda tipo B, la compuerta flotante se traslapa únicamente con el drenador, por lo que tanto la programación como el borrado se hacen por esta terminal. Esto implica una optimización de la implantación del drenador para lograr un compromiso entre el voltaje alto necesario para la programación y el voltaje bajo requerido para el borrado.

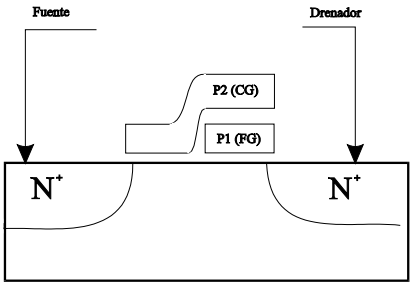


Fig. 2.4. Celda B.

Las celdas del Tipo 2, son tres: Celda C, Celda D y Celda E. Estas, como ya se mencionó anteriormente, tienen la compuerta flotante arriba de la de control.

Celda C.

En esta celda, también la compuerta flotante se traslapa con el drenador, realizándose por tanto, la programación y el borrado por la misma terminal.

Una de las grandes ventajas de esta celda, es que la razón de acoplamiento en el programado de la memoria, es mayor, debido a la línea metálica (bit-line) que va por encima de la compuerta flotante y a la fácil resolución de la geometría de la compuerta flotante, con lo que se reduce la porción de la compuerta flotante que cubre la sección de programación del canal del transistor, y esto a su vez reduce la capacitancia entre la compuerta flotante y el substrato y con esto, el área de la celda.

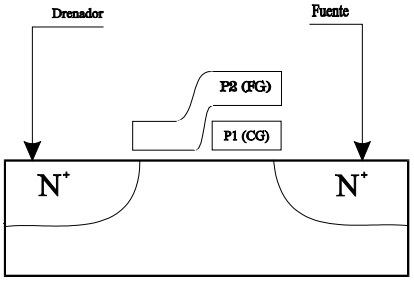


Fig. 2.5. Celda C.

Celda D.

Esta celda es similar a la celda C, excepto que utiliza un tercer electrodo de silicio policristalino, para la función de borrado. Esto evita el problema de optimizar la implantación del drenador para cumplir con el borrado y la programación de una manera eficiente. Este tercer electrodo cubre la unión de fuente y traslapa mínimamente parte de la compuerta flotante para lograr una mejor razón de acople durante el borrado. Con una oxidación a baja temperatura, la superficie de polisilicio de la compuerta flotante se texturiza y con esto se logra incrementar el campo eléctrico durante el borrado, provocando una disminución de la barrera de energía de los electrones y en consecuencia, se tiene una mayor corriente FN. La ventaja que esto tiene, es que los óxidos son menos delgados (60-80 nm) y fuentes de voltaje no son altas (~12 V).

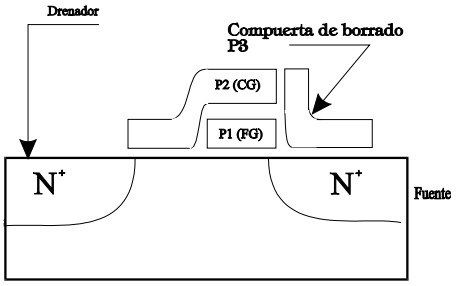


Fig. 2.6. Celda D.

Celda E.

La celda E es parecida a la celda C, sólo que la compuerta flotante cubre no sólo al drenador, sino también a la fuente. De esta manera, las uniones de drenador y fuente son independientemente optimizadas para la programación y borrado de la memoria, respectivamente. Este tipo de celda tiene un proceso de fabricación muy simple.

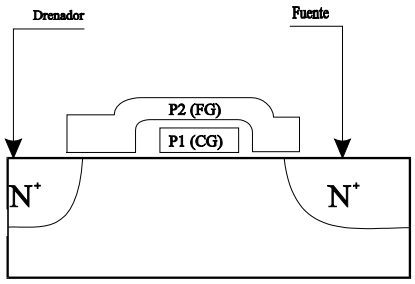


Fig. 2.7. Celda E.

En cuanto al proceso tecnológico utilizado en la fabricación de las celdas, se diseñaron en base al pozo **N CMOS** para EPROM de 1.5 μm, con doble capa de silicio policristalino y una sola capa metálica. La definición por decapado de las geometrías del polisilicio 1 y el polisilicio 2, varía según la estructura. Si cada capa de silicio policristalino es grabada independientemente, la alineación entre ambas controlará la longitud de canal de la celda. Por lo tanto, un desalineamiento entre el polisilicio 1 y 2 con respecto a la línea metálica del bit, causará que algunas celdas de una hilera dada tengan diferentes longitudes de canal, con respecto a la hilera adjunta. Esto no sucede si se tiene un proceso de auto-alineado, es decir, cuando ambas capas de silicio policristalino se graban a la vez.

Se usaron dos implantaciones $\mathbf{N^+}$ enmascaradas, una después del depósito del polisilicio 1 y la otra después del depósito del polisilicio 2. La primera implantación fue usada para definir el drenador de la celda B. y la fuente de las celdas A, C, D y E. La segunda implantación $\mathbf{N^+}$ se usó para definir la fuente de la celda B y el drenador de las celdas A, C, D y E.

Con respecto al desempeño funcional de las estructuras anteriores, se utilizaron tres criterios para compararlas entre sí: 1) corriente de lectura, 2) velocidad de programación y 3) velocidad de borrado. Todas las celdas las diseñaron con el mismo ancho de canal, sin embargo la corriente de lectura se controla con la longitud efectiva de canal y en menor grado, por la razón de acople durante la programación. Esto último debido a que entre mayor razón de acople exista, mayor voltaje se tendrá en la compuerta flotante y en consecuencia, existirá mayor corriente de lectura.

Como conclusión de la comparación hecha por Amin, se puede decir que la celda A tiene la mejor densidad de integración, mayor corriente de lectura y buena característica de programación, pero cuenta con un borrado poco útil. La celda B mejora la característica de borrado, aún sobre la celda C, pero se requiere optimizar la unión del drenador para cumplir eficientemente con la programación y el borrado. La celda D tiene mejores características de borrado, pero una tecnología más compleja. Por último, la celda E es de muy bajo rendimiento y gran área. Todas estas celdas están en desarrollo por el Laboratorio de National Semiconductor y por lo tanto, son exclusivas de esta compañía.

Un desarrollo mas, se encuentra en [4], cuya celda corresponde a la mostrada en la Fig. 2.1. En este caso, se emplea el tunelamiento Fowler-Nordheim tanto para inyectar como para remover carga de la compuerta flotante; con esto se tiene mayor simetría del borrado con respecto a la escritura. El óxido de tunelamiento se coloca por encima del drenador, por lo que el proceso de programación se hace creando un campo eléctrico entre la compuerta de control y el drenador. El hecho de tener que polarizar el drenador para la programación, lleva a tener tres operaciones del transistor: 1) *borrado* (voltaje positivo en el drenador y compuerta de control a cero volts o con menor voltaje positivo); 2) *escritura* (drenador aterrizado y compuerta de control polarizada con voltaje positivo); y 3) *lectura* (polarizar al transistor en su punto de operación).

El óxido de tunelamiento de esta estructura tiene un espesor de 10 nm, lo que se puede hacer con tecnologías muy especiales y caras. Por otro lado, el análisis hecho en [4] concluye que se necesita una optimización de los procesos de difusión e implantación para mejorar tanto el borrado como la escritura de esta memoria. En la misma referencia [4] hacen una simulación con las modificaciones estudiadas y demuestran la viabilidad de operación del dispositivo, haciendo énfasis en la posibilidad de estudiar una posible modificación a las reglas de diseño de la tecnología de fabricación. Esto último lo hacen también en [5] con el propósito de analizar el escalamiento de estructuras a menores dimensiones.

Otra estrategia interesante en el diseño y fabricación de memorias de compuerta flotante, es la adoptada por Montalvo y Paulos [6], con la que lograron tener dispositivos muy rápidos, de poca área y fabricados con tecnología **CMOS estándar**, de 2 μm, doble polisilicio, doble metal y pozo **N.** Estos dispositivos también usan la inyección por electrones calientes para programar la memoria y tunelamiento FN para borrarla.

Para una aplicación que requiera alta densidad de integración, la unión de fuente puede ser compartida por celdas vecinas, lo que reduce el área. Dado que una razón de acoplamiento adecuada es necesaria para obtener buenas características, se agregan dos capacitores de acoplamiento: uno de valor grande para cumplir con un porcentaje de acoplamiento mayor al 50 % durante la programación, y otro pequeño para generar una razón de acoplamiento menor al 10 % durante el borrado. Al requerir de dos capacitores de acoplamiento, se requiere un compromiso entre ambos para tener un funcionamiento y tamaño adecuado; también, se debe polarizar al drenador cuando se quiere borrar, haciendo estos dos últimos factores que el diseño se complique. Los tiempos de escritura son de aproximadamente cientos de microsegundos mientras que los tiempos de borrado son de decenas de milisegundos, con cambios en el voltaje de umbral de alrededor de 3 V con bajos voltajes de programación.

Esta estructura tiene propiedades que la hacen un buen candidato para ser empleada en los propósitos planteados. Quizá la única característica que repercute en la operación, sin ser necesariamente definitiva, es el hecho de tener que polarizar al drenador, por un lado para su lectura y por otro lado para su programación.

La referencia [7] presenta un diseño similar al presentado en [4], donde tamabién se emplea el tunelamiento Fowler-Nordheim tanto para escribir como para borrar, a través de un óxido de 10 nm. Se necesitan diseñar dos coeficientes de acoplamiento y se debe polarizar al drenador (junto con la compuerta) para la programación de la memoria. A pesar de tener un buen desempeño, es preciso emplear tecnología exclusiva para tener un óxido de tunelamiento delgado.

También en la referencia [8], proponen una estructura como la mostrada en la Fig. 2.8, a partir de la cual diseñan y construyen una memoria Flash EEPROM de 128 Kbyts. Emplean inyección de electrones calientes de canal para la escritura y tunelamiento Fowler-Nordheim para el borrado, a través de un óxido fino de 20 nm y entre compuerta flotante y drenador. El dieléctrico que forman entre los dos polisilicios (compuerta de control y compuerta flotante) está hecho a base de una capa de óxido-nitruro de silicio-óxido. Por lo tanto, es una tecnología especial la que se emplea para la fabricación de esta estructura.

Con anterioridad, ya se había mencionado que el borrado mediante pulsos eléctricos podía provocar un sobreborrado en la memoria, esto es, dejar carga positiva en la compuerta flotante, convirtiendo al transistor de la memoria, en un transistor de deserción. Esto se puede evitar con un transistor de enriquecimiento puesto en serie con el transistor de compuerta flotante [8]. El dispositivo completo consta de la implantación de fuente, sobre la que se traslapa la compuerta de control, que a su vez cubre también a la compuerta flotante, como se puede ver en la Fig. 2.8. La compuerta de control, junto con la implantación de drenador forman el transistor de enriquecimiento. El agregar este dispositivo, es lo que hace que las FEEPROM sean 20 % más grandes que las EPROM, pero se logran mayores ventajas como mayor corriente de lectura y mejores características de programación. Al emplear óxido de tunelamiento delgado, la fabricación de esta memoria requiere de tecnología especial.

El empleo de una tecnología estándar para la fabricación de una estructura de compuerta flotante, se reporta en [9]. En este caso, un metódico estudio del funcionamiento de las estructuras en función de las características geométricas de diseño, da como resultado que la inyección o extracción de carga se ve más favorecida cuando el tunelamiento es a través del óxido que separa a los dos polisilicios, ya que la misma cantidad de carga se puede transferir con menor campo eléctrico, a comparación del aplicado entre el polisilicio de la compuerta flotante y el drenador. Este resultado es interesante, ya que de esta manera, se pueden diseñar inyectores independientes de la terminal de drenador o fuente, como lo hacen las demás estructuras. En este caso, sólo se necesita un capacitor de acoplamiento.

Una estructura similar es la estructura reportada en [10], la cual consta de dos inyectores que sirven respectivamente para inyectar y extraer carga mediante tunelamiento Fowler-Nordheim, lo que hace a este proceso mas simétrico en sus características de programación. Estos inyectores se forman al traslapar transversalmente dos polisilicios, separados por un óxido de tunelamiento relativamente grueso, de aproximadamente 70 nm. Su fabricación es posible con tecnología estándar y los voltajes empleados en el proceso de inyección están por debajo de los 20 V, a pesar de no usar óxidos ultrafinos; los tiempos de borrado y escritura son de algunos milisegundos. Una desventaja relativa consiste en el área que ocupa el capacitor de acoplamiento, que es mediante el cual se tiene en la compuerta flotante, una fracción del voltaje aplicado en la compuerta de control, sin embargo, también se emplea únicamente un capacitor de acoplamiento.

Con esta estructura, se puede tener un rango de voltajes de umbral por encima de los 10 V sin perjuicio del funcionamiento del dispositivo. Otra ventaja que se puede mencionar, es que no se requiere polarizar al drenador o a la fuente para la inyección; esto se hace exclusivamente polarizando al inyector y a la compuerta de control.

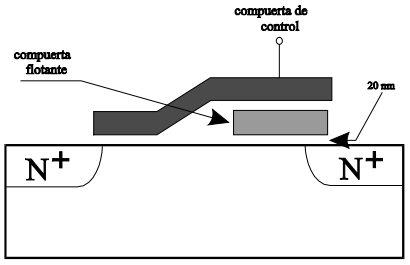


Fig. 2.8. Memoria FEEPROM.

Los desarrollos encontrados en la literatura tratan de dar énfasis al uso de tecnología convencional, en donde no se requieren procesos especializados como óxidos ultrafinos de buena calidad ó texturización de superficies, siendo estos últimos, procesos poco accesibles a los diseñadores, por ser tecnologías exclusivas, a diferencia de la tecnología convencional, a la cual se puede recurrir con facilidad para la fabricación en tecnología MOS, por ejemplo, por MOSIS[2] y en compañías establecidas como ORBIT. Esta compañía ofrece un proceso CMOS de 2 μm, pozo P y doble polisilicio, sobre el cual se han basado muchos desarrollos de memorias para redes neuronales [6, 8, 9, 10, 11]. Entre los diseños reportados, existen quienes utilizan el tunelamiento FN tanto para programar como para borrar, o quienes prefieren programar por inyección de electrones calientes de canal y borrar por tunelamiento FN.

Los circuitos que usan este último fenómeno para ambos propósitos, tienen la ventaja de que aunque se necesitan voltajes mayores a los empleados con tecnologías de óxidos ultrafinos, se genera poca corriente y entonces es relativamente fácil integrar la fuente en el circuito integrado [11].

Las estructuras y fenómenos relacionados con los dispositivos de compuerta flotante vistos anteriormente, abarcan la generalidad del estado del arte en cuanto a lo reportado en la literatura. Se pueden encontrar otras estructuras, pero cuyo desarrollo es realizado exclusivamente en los laboratorios de los intitutos de investigación, perdiendo por lo tanto, uno de los puntos importantes para su elección, como lo es el de tener una tecnología accesible. Se puede hacer, por lo tanto, una comparación que arroje los puntos sobresalientes de estas estructuras y elegir la que más convenga a los propósitos de este trabajo.

Los elementos base para hacer la elección, en función de su posible aplicación en RNA, son:

1. Tecnología estándar y compatible con VLSI.
2. Capacidad de almacenamiento.
3. No volátil.
4. Tiempos rápidos para escritura y borrado.
5. Area pequeña.

Todos estos requisitos los cumplen en diferente medida las estructuras reportadas en [9] y [10] y el hecho de que la programación se realice independiente de la terminal de drenador o fuente, daría ventajas al estar sensando la corriente de operación del dispositivo simultaneamente con la aplicación de un algoritmo de ajuste en línea en el voltaje de umbral. Al incluir dos inyectores, cuya área es mínima (2 μm x 2 μm), la estructura de la referencia [10] se puede emplear para un ajuste de programación más controlado. Por lo tanto, esta será la estructura que se utilizará para incluirla en una RNA y se explicará mas extensamente.

En la tabla 2.1 se resumen las comparaciones entre los resultados de las mediciones hechas en [3] así como de [4], [5], [6], [7], [8], [9] y [10], junto con las dificultades tecnológicas asociadas.

Como se pudo apreciar en la presentación de las estructuras de compuerta flotante, se manejan dos conceptos comunes e importantes para las mismas: tunelamiento Fowler-Nordheim y coeficiente de acoplamiento. Ya que son conceptos que intervienen en el funcionamiento del dispositivo, vale la pena una explicación mas amplia para que sean comprendidos. El enfoque se hará particularmente hacia el tipo de estructura elegida: *la memoria de compuerta flotante de dos inyectores.*

[2] *MOSIS es un acrónimo de MOS Implementation System. Este servicio proporciona la fabricación del chip en fábricas establecidas.*

Tabla 2.1. Comparación de las diferentes celdas.

Celda	Corriente de lectura (μA)	Velocidad de prog. (s)	Velocidad de borrado (s)	Comentarios
A	100-110	Vcg=13 V 10^{-6} @ Vd=7.5 V 10^{-5} @ Vd=6.5 V 10^{-2} @ Vd=5.5 V	20 @ Ie=1000 nA 80 @ Ie=400 nA >100 @ Ie=100 nA	Tiene la mayor corriente de lectura y la menor área de todas; la mejor característica de programación, pero con una velocidad de borrado poco útil. Su borrado puede convertir a la celda en un transistor de deserción. El drenador y la fuente se difunden con diferente perfil. Compuertas autoalineadas. Programación y borrado por drenador y fuente, respectivamente.
B	75-80	Vcg=13 V 10^{-6} @ Vd=10 V 10^{-5} @Vd=9 V 10^{-3} @ Vd=8 V	10 @ Ie=1000 nA 15 @ Ie=600 nA 40 @Ie=200 nA 100 @ Ie=100 nA	Características de programación similares a la celda A con mayores voltajes de drenador. Riesgo de borrar por debajo de su Vt nativo. Mejores características de borrado que la celda A, mejoradas con drenador de perfil gradual. Compuertas autoalineadas. Programación y borrado por drenador.
C	75-80	Vcg=13 V 10^{-6} @Vd=10 V 10^{-5} @ Vd=9 V 10^{-3} @ Vd=8 V	10 @ Ie=1000 nA 15 @ Ie=600 nA 40 @ Ie=200 nA 100 @ Ie=100 nA	Características de programación y de borrado similares a la celda B. Mayor acople capacitivo y área pequeña. Grabado independiente del polisilicio 1 y polisilicio 2. Programación y borrado por drenador. No se tiene borrado excesivo. Compuertas no autoalineadas.
D	75-80	Vcg=13 V 10^{-6} @ Vd=10 V 10^{-5} @Vd=9 V 10^{-3} @ Vd=8 V	10^{-1} @ Ve=13 V 1 @ Ve=12 V 10 @ Ve=11 V	Utiliza una tercer compuerta para borrado. Características de programación similares a la celda B. No se tiene borrado excesivo. Permite el uso de óxidos más gruesos. Programación por drenador. No tiene borrado excesivo. Fabricación compleja. Compuertas no autoalineadas.
E	60-65	malas caract. de programación	10 @ Ie=1000 nA 15 @ Ie=600 nA 40 @ Ie=200 nA 100 @ Ie=100 nA	Fabricación sencilla, área grande. Programación por drenador y borrado por fuente. Características de borrado similares a la celda B. Compuertas no autoalineadas.
[4]	-	Vcg=16 V 10^{-3} @ Vd=0 V	Vd=16 V 10^{-3} @Vcg=0 V	Programación y borrado mediante FN. Oxido de tunelamiento de 10 nm. Emplea tecnología de fabricación especial.
[5]	-	Vcg=12 V 10^{-3} @ Vd=5 V	-	Programación mediante electrones calientes de canal. Estructuras de canal corto (0.5-0.7 μm). Modelo para validación de reglas de diseño.

[6]	-	Vcg=16 V 10^{-3} @Vd=10 V	Vd=16 V 10-2 @Vcg=0 V	Fabricada con tecnología estándar. Programación por electrones calientes y borrado por FN. Se necesitan dos capacitores de acoplamiento.

Continuación de la Tabla 2.1.

[7]	-	Vcg=20 V 10^{-3} @Vd=0 V	Vd=20 V 10^{-3} @Vcg=0 V	Oxido de tunelamiento delgado (10 nm), fabricada con tecnología especial. Programación y borrado con tunelamiento FN. Se requieren dos capacitores de acoplamiento.
[8]	60-90 μA	Vcg=16 V 10^{-3} @Vd=9 V	Vd=19 V 1 @Vcg=0 V	Oxido de tunelamiento delgado (20 nm) y dieléctrico entre poly1 y poly2 a base de óxido-nituro-óxido. Tecnología de fabricación especial. Programación mediante electrones calientes y borrado mediante FN.
[9]	-	Vi=-21 V 1-3 @Vcg=-4 V	Vi=-17 V 1-3 @Vcg=4 V	Fabricación con tecnología estándar. Se emplea un solo capacitor de acoplamiento. Programación y borrado mediante tunelamiento FN.
[10]	-	Vi=-18 V 3-4 @Vcg=0 V	Vi=22 V 3-4 @Vcg=0 V	Fabricación con tecnología estándar. Se emplea un solo capacitor de acoplamiento. Programación y borrado mediante tunelamiento FN. Emplea dos inyectores: uno para borrar y otro para escribir.

2.1.2. Modelos.

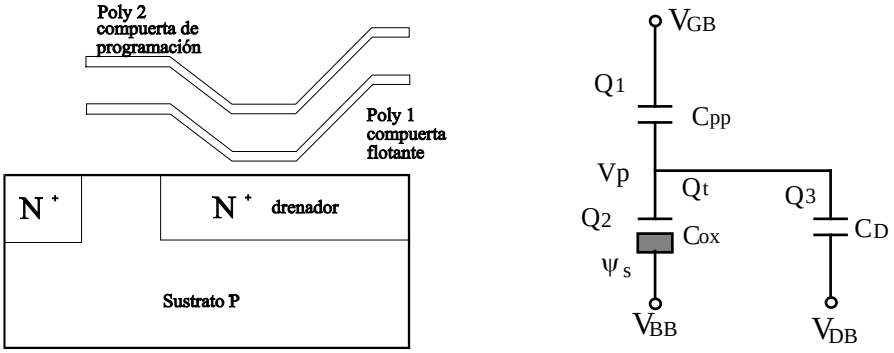
Un complemento al diseño topológico de las memorias, es el modelado de sus características I-V. Se requiere su conocimiento para lograr una buena interfaz entre la memoria y el circuito periférico y optimizar el funcionamiento del sistema.

Uno de estos modelos, es el establecido por Liong y Liu [12], basado en el modelo de *disminución de la barrera inducida por drenador*, (DIBL), la cual predice un aumento de la corriente de drenador, con respecto al modelo simple del transistor MOS de canal corto. El circuito equivalente de una estructura EEPROM sin la terminal de fuente, del cual se parte para el desarrollo de la ecuación de corriente de drenador, en función del voltaje V_{DS}, se muestra en la Fig. 2.9(b).

Los parámetros mostrados en la Fig. 2.9(b), son los siguientes:

Cpp: capacidad entre el polisilicio 1 y el polisilicio 2.
Cox: capacidad de compuerta (polisilicio 1 y canal).
C_D: capacidad total del óxido de tunelamiento y cualquier capacidad parásita.

s: potencial superficial.
Vp: potencial de la compuerta flotante.



(a) (b)

Fig. 2.9. a) Estructura EEPROM; b) Circuito eléctrico equivalente.

Para la derivación del modelo, se establecen las siguientes suposiciones:

i) Qt es la carga atrapada en la compuerta flotante.
ii) No existe carga en todas las interfaces.
iii) No se tienen efectos de borde (fringing).
iv) El potencial de contacto y las cargas atrapadas en el óxido, se asumen nulas, de tal forma que se tiene condición de bandas planas.
v) No hay tunelamiento de huecos o electrones en las condiciones de operación.
vi) El substrato está contaminado uniformemente y no degenerado.
vii) Todas las impurezas están completamente ionizadas.

La distribución de carga en el silicio, para un capacitor MOS se puede encontrar resolviendo la ecuación de Poisson. Con concentraciones normales de substrato, tal que se cumpla que $2\ _F$ » Vt y cuando el canal está en deserción o inversión, de tal manera que se tenga que $_s$»V_t, se puede dar una ecuación aproximada para la densidad de carga en la superficie del canal:

$$Q'_{si} = -\left(F\sqrt{N_a}\right)\sqrt{\psi_s + V_t e^{\frac{\psi_s - 2\phi_F}{V_t}}} \quad , \qquad (2.2)$$

donde $F = \sqrt{2 e_{si} q}$, $_F$ es el potencial de Fermi del volumen y Vt es el voltaje térmico.

Cuando se aplica un voltaje al drenador, la ecuación (2.2) se modifica introduciendo V_{DB} en el argumento de la exponente. Se tiene entonces:

$$Q'_{si} = -\left(F\sqrt{N_a}\right)\sqrt{\psi_s + V_t e^{\frac{\psi_s - 2\phi_F - V_{DB}}{V_t}}} \quad . \qquad (2.3)$$

Asumiendo que se aplican los voltajes como se muestra en la Fig. 2.9, se tiene que el potencial resultante en la compuerta flotante es Vp con una carga atrapada en ella de Q_t, por lo tanto, con un potencial de superficie resultante, ψ_s.

La carga total en la compuerta se debe conservar y, se tiene entonces que:

$$Q_t = Q_1 + Q_2 + Q_3 \quad , \tag{2.4}$$

$$Q_1 = C_{pp}(V_p - V_{GB}) \quad , \tag{2.5}$$

$$Q_2 = C_{ox}(V_p - \psi_s) \quad , \tag{2.6}$$

$$Q_3 = C_D(V_p - V_{DB}) \quad . \tag{2.7}$$

Resolviendo las cuatro ecuaciones anteriores, se puede encontrar la carga por unidad de área de la superficie inferior de la compuerta flotante, Q'2, la cual está por encima del canal:

$$Q'_2 = \left(\frac{C'_{ox}}{C_{ox} + C_D + C_{pp}}\right)\left[Q_t + V_{GB}C_{pp} + C_D V_{DB} - (C_D + C_{pp})\psi_s\right] \quad , \tag{2.8}$$

Se requiere que exista conservación de carga en el canal y la superficie inferior de la compuerta flotante justo arriba del canal, por tanto:

$$Q'_2 + Q'_{si} = 0 \quad . \tag{2.9}$$

Sustituyendo (2.3) y (2.8) en (2.9) y resolviendo para V_{GB}, se tiene una relación del potencial superficial con el voltaje de la compuerta de programación:

$$V_{GB} = -\frac{Q_t}{C_{pp}} + (1+\alpha)\psi_s - \alpha V_{DB} + \gamma\sqrt{\psi_s + V_t e^{\frac{\psi_s - 2\phi_F - V_{DB}}{V_t}}} \quad , \tag{2.10}$$

donde:

$$\alpha = \frac{C_D}{C_{pp}} \quad , \tag{2.11}$$

$$\beta = \frac{C_{ox}}{C_{pp}} \quad , \tag{2.12}$$

$$\gamma = \frac{F\sqrt{N_a}}{C'_d} \quad , \tag{2.13}$$

$$C'_d = \frac{C'_{ox}}{1+\alpha+\beta} \quad , \tag{2.14}$$

siendo la razón de acoplamiento del drenador, la razón de acoplamiento de canal, el coeficiente de efecto de cuerpo y C'd se puede interpretar como la capacidad equivalente de la compuerta de programación.

Considerando en general que existen cargas atrapadas en las interfaces y también potencial de contacto, no se tiene la condición de bandas planas con cero polarización, por lo que introduciendo estos factores en la ecuación (2.10), a través de V_{fb}, se tiene:

$$V_{GB} = V_{fb} + (1+\alpha)\psi_s - \alpha V_{DB} + \gamma\sqrt{\psi_s + V_t e^{\frac{\psi_s - 2\phi_F - V_{DB}}{V_t}}} \quad , \tag{2.15}$$

$$V_{fb} = -\frac{Q_t}{C_{pp}} + \phi_{MS} + V(\text{con carga}) \quad , \tag{2.16}$$

donde $_{MS}$ se refiere al potencial de contacto y V(con carga) representa el potencial efectivo debido a carga en la interfaz. V_{fb} se puede interpretar como el voltaje de compuerta de programación, necesario para tener la condición de bandas planas con cero V_{DB} y V_{SB}. Los potenciales superficiales del canal cerca de drenador y fuente, están dados por:

$$\psi_{SD} = \frac{1}{1+\alpha}\left(V_{GB} - V_{fb} + \alpha V_{DB} - \gamma\sqrt{\psi_{SD} + V_t e^{\frac{\psi_{SD} - 2\phi_F - V_{DB}}{V_t}}}\right) \quad , \tag{2.17}$$

$$\psi_{SS} = \frac{1}{1+\alpha}\left(V_{GB} - V_{fb} + \alpha V_{DB} - \gamma\sqrt{\psi_{SS} + V_t e^{\frac{\psi_{SS} - 2\phi_F - V_{DB}}{V_t}}}\right) \quad . \tag{2.18}$$

A continuación se hacen las siguientes suposiciones para calcular la corriente de drenador de la estructura EEPROM:

i. Longitud de canal *corto* y ancho de canal *largo*.
ii. V_{DB} es mayor que V_{SB}.
iii. Se hace uso de la aproximación gradual de canal.

La corriente de drenador está compuesta de dos términos: *difusión* y *arrastre*. En cualquier punto de la capa invertida, denominada Q'_I, se tiene que:

$$I_D(x) = \mu W(-Q'_I)\frac{d\psi_s}{dx} + \mu W V_t \frac{dQ'_I}{dx} \quad . \tag{2.19}$$

Si se integra (2.19) a lo largo de la longitud de canal L y suponiendo la movilidad superficial de electrones como constante, se obtiene:

$$I_D = \mu\frac{W}{L}\int_{\psi_{SS}}^{\psi_{SD}}(-Q'_I)d\psi_s + \mu\frac{W}{L}V_t\int_{Q'_{IS}}^{Q'_{ID}}dQ'_I \quad . \tag{2.20}$$

Tomando las aproximaciones de *deserción* y *carga de hoja*, entonces $Q'_{si}=Q'_I+Q'_B$, donde Q'_B es la densidad de carga de deserción. Q'_I está expresada por:

$$Q'_I = -C'_d\left(V_{GB} - V_{fb} - (1+\alpha)\psi_s + \alpha V_{DB} - \gamma\sqrt{\psi_s}\right) \quad . \tag{2.21}$$

Usando (2.20) y (2.21) e integrando, se encuentra la corriente de drenador:

$$I_D = I_{D1} + I_{D2} \quad , \tag{2.22}$$

$$I_{D1} = \mu \frac{W}{L} C'_d \left[(V_{GB} - V_{fb} + \alpha V_{DB})(\psi_{SD} - \psi_{SS}) - \frac{1}{2}(1+\alpha)(\psi^2_{SD} - \psi^2_{SS}) - \frac{2}{3}\gamma(\psi^{1/2}_{SD} - \psi^{1/2}_{SS}) \right], \quad (2.23)$$

$$I_{D2} = \mu \frac{W}{L} C'_d \left[(\psi_{SD} - \psi_{SS}) V_t + \gamma(\psi^{1/2}_{SD} - \psi^{1/2}_{SS}) V_t \right] \quad . \quad (2.24)$$

A partir de la ecuación (2.22), se pueden encontrar las características I-V de la EEPROM, tanto en la región lineal, como en saturación. En la Fig. 2.10, se muestran estas características para una estructura con un voltaje de umbral Vth = 0.274 V después de programada, Kp = 59 A/V^2, = 1.063 A/V$^{1/2}$, = 0.291 y Vfb= -1.13 V.

$$K_p = m\frac{W}{L} C¢_d \quad . \quad (2.25)$$

Para poder simular el comportamiento de la estructura a partir del modelo, se tienen que medir los parámetros para comparar las mediciones con la teoría. Los voltajes de umbral de la estructura EEPROM y de un transistor MOS, se definen como sigue:

$$V_t(\text{flotox}) = V_{fb} + (1+\alpha)\phi_B + \gamma_{\text{flotox}}\sqrt{\phi_B + V_{SB}} \quad , \quad (2.26)$$

$$V_t(\text{MOS}) = V_{fb} + \phi_B + \gamma_{MOS}\sqrt{\phi_B + V_{SB}} \quad . \quad (2.27)$$

Si se cumple que $_B$ y C'ox son iguales para ambos transistores, aplicando el mismo V_{SB}, se obtiene la siguiente relación:

$$(V_t - V_{fb})_{\text{flotox}} = (1+\alpha+\beta)(V_t - V_{fb})_{MOS} - \beta\phi_B \quad , \quad (2.28)$$

$$\frac{\gamma_{\text{flotox}}}{\gamma_{MOS}} = 1 + \alpha + \beta \quad . \quad (2.29)$$

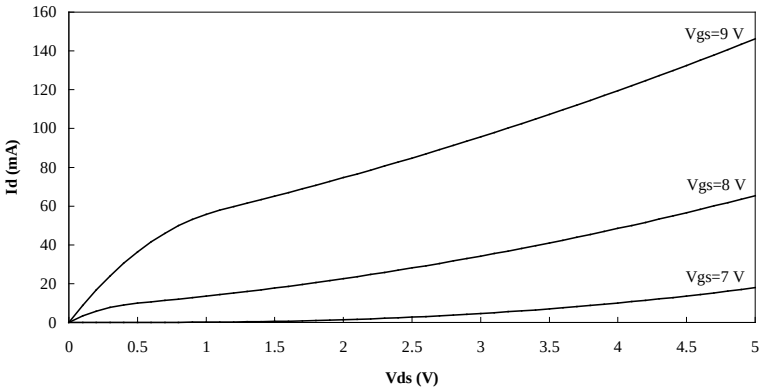


Fig. 2.10. Características I-V de una estructura EEPROM a partir de (2.22).

Dado que las ecuaciones de corriente y voltaje de la memoria EEPROM son similares a las del transistor MOS, se pueden usar las técnicas utilizadas en éste último para medir el *voltaje de umbral*, *voltaje de bandas planas*, *potencial superficial* y el *coeficiente de efecto de cuerpo* en una capa fuertemente invertida, para extraer los parámetros de la memoria EEPROM.

Los diferentes valores de Vt(flotox) se obtienen graficando I_D vs V_{GS} para varios voltajes V_{SB}. De graficar Vt(flotox) vs $(V_{SB}+\phi s)^{1/2}$ se encuentra γflotox de la pendiente de la recta obtenida y V_{fb}(flotox) se obtiene del cruce con el eje *y*. Posteriormente, de graficar $(V_t-V_{fb})_{flotox}$ vs $(V_t-V_{fb})_{MOS}$ se encuentra β del cruce con el eje *y* de la recta graficada. Este es un método iterativo partiendo de un valor arbitrario de α propuesto y después calculado con (2.29) a partir de los otros valores medidos (ver Figs. 2.11, 2.12 y 2.13).

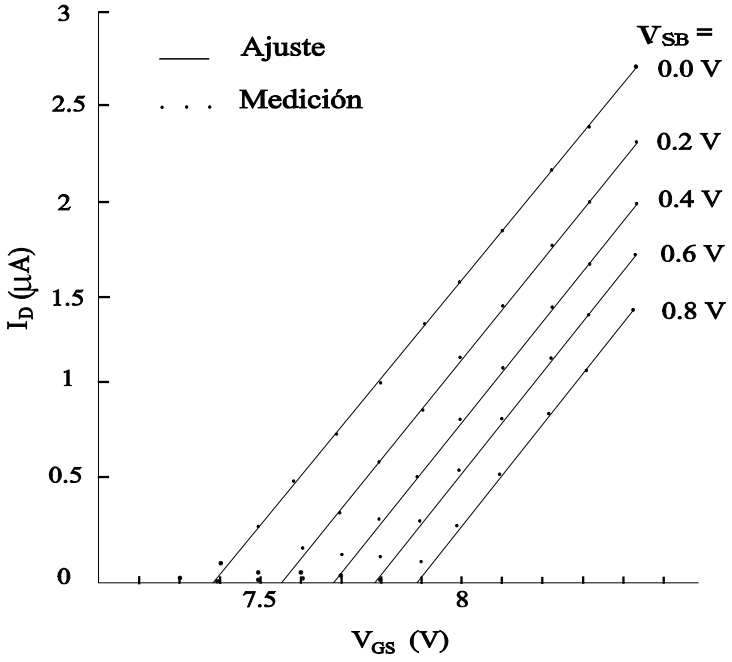


Fig. 2.11. I_D vs Vgs.

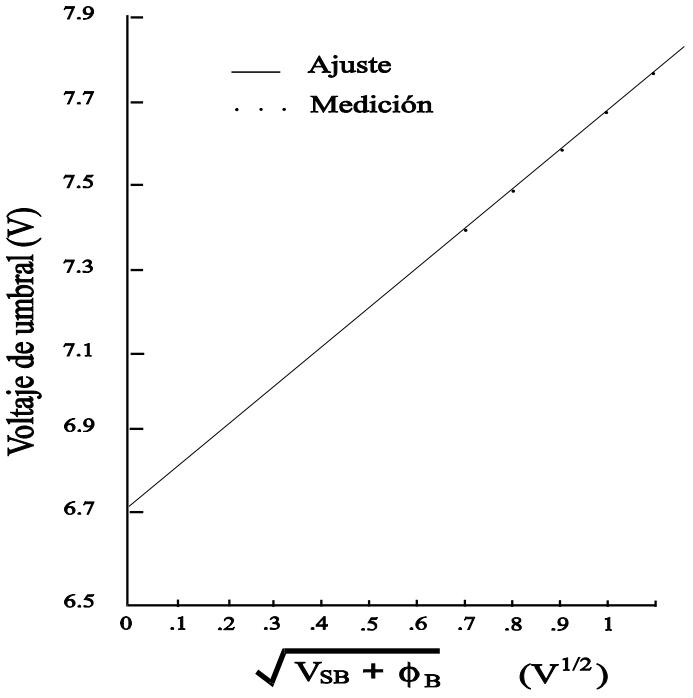


Fig. 2.12. V_T vs. $(V_{SB} +$ $_B)^{1/2}$ para obtener y V_{fb}.

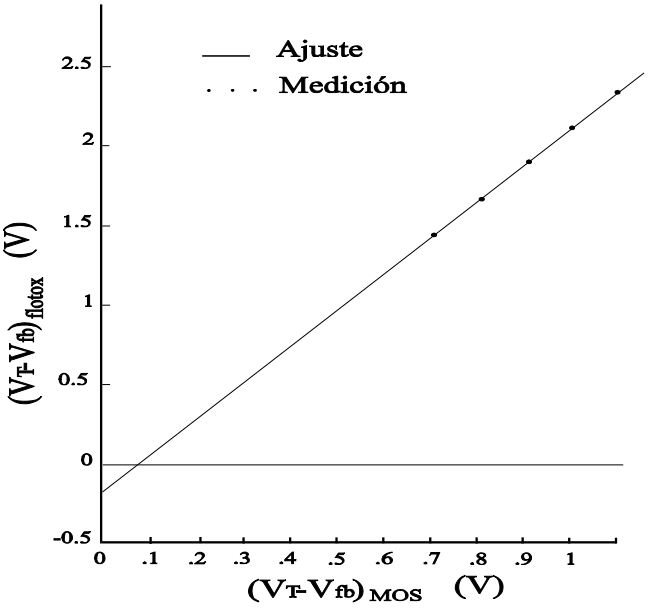


Fig. 2.13. $(V_T - V_{fb})_{flotox}$ vs. $(V_T - V_{fb})_{MOS}$, para obtener la razón de acople.

Este modelo tiene la ventaja de mostrar que el voltaje de umbral extrapolado es una cantidad conveniente para definir al voltaje de umbral de la memoria EEPROM, además que la técnica para la extracción de las razones de acoplamiento a partir de mediciones eléctricas, es insensible a la cantidad de carga atrapada en la compuerta flotante.

Otro modelo desarrollado en base a un circuito equivalente similar al de la Fig. 2.9(b), es el propuesto por Chi-Kai Sin et al. [13], donde parten de las ecuaciones para la corriente de drenador de un MOSFET convencional y relacionan el voltaje de la compuerta flotante, con el voltaje aplicado a las terminales.

Por lo tanto, el problema del modelado, se reduce a encontrar el voltaje de la compuerta flotante en términos de los voltajes aplicados, cantidad de carga almacenada en la compuerta flotante y otros parámetros del dispositivo.

Este modelo considera una programación y borrado mediante tunelamiento FN y el espesor del óxido delgado del diseño reportado es de 8-10 nm. La Fig. 2.15 muestra el diseño utilizado para la fabricación y modelado de una memoria EEPROM, fabricada con tecnología NMOS de doble polisilicio, óxido delgado de tunelamiento y una implantación N^- bajo esta región.

En base a la Fig. 2.14, las diferentes capacidades se definen a continuación

Capacidad entre compuertas de silicio policristalino (C_{FG}).

$$C_{FG} = \frac{\varepsilon_{ox} A_1}{t_{ox1}} \quad , \tag{2.30}$$

donde $_{ox}$ es la permitividad del dióxido de silicio, A_1 es el área de traslapamiento entre la compuerta superior y la compuerta flotante y t_{ox1} es el espesor de óxido entre ambas compuertas.

Capacidad entre la compuerta flotante y el canal (C_{ox}).

$$C_{ox} = \frac{\varepsilon_{ox} A_2}{t_{ox2}} = \frac{\varepsilon_{ox} WL}{t_{ox2}} \quad , \tag{2.31}$$

donde t_{ox2} es el espesor del óxido sobre el canal, W es el ancho efectivo de canal y L es el largo efectivo del canal.

Capacidad debido al traslapamiento entre compuerta flotante y el área de tunelamiento (C_{fd}).

$$C_{fd} = \frac{\varepsilon_{ox} A_4}{t_{ox3}} \quad , \tag{2.32}$$

donde A_4 es el área de traslapamiento entre la compuerta flotante y el área de tunelamiento y t_{ox3} es el espesor del óxido fino.

Capacidad debido al traslapamiento entre compuerta flotante y el volumen (C_{fgb}).

$$C_{fgb} \approx \frac{\varepsilon_{ox} A_3}{t_{ox2}} \quad , \tag{2.33}$$

donde A_3 es el área del traslapamiento entre compuerta flotante y el substrato.

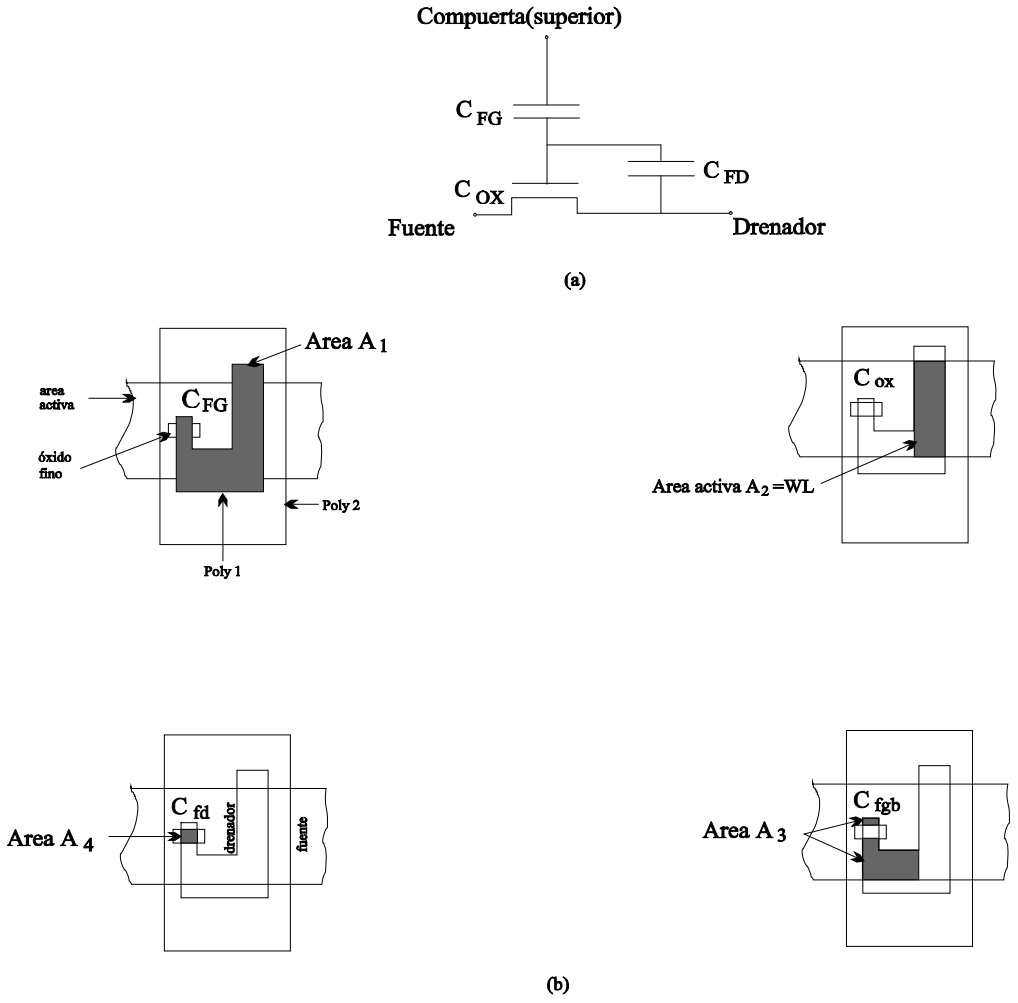


Fig. 2.14. a) Circuito equivalente de una memoria EEPROM; b) áreas que definen la capacidad entre compuertas, de óxido de compuerta y de acoplamiento de drenador.

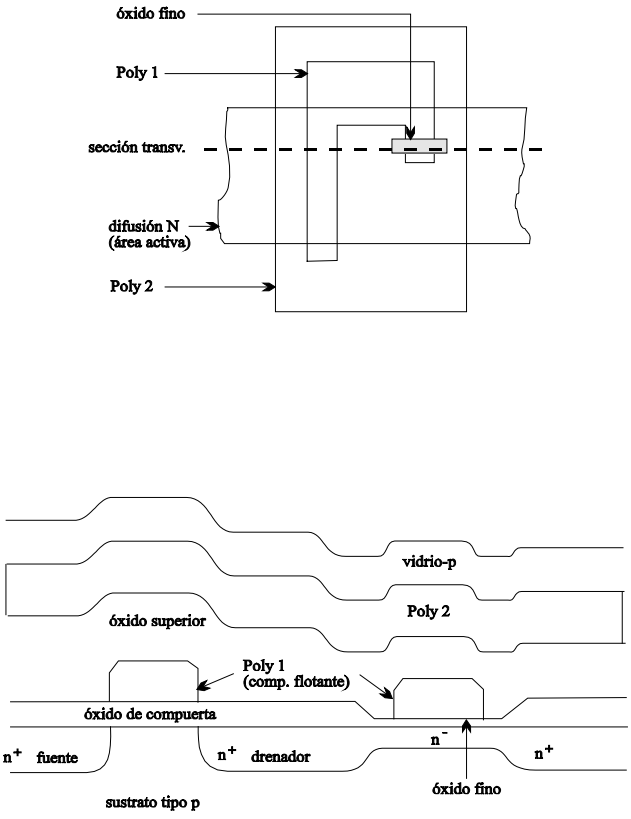


Fig. 2.15. Diseño y sección transversal de la memoria EEPROM.

De las dos capacidades anteriores, se puede encontrar la capacidad de acoplamiento de drenador (C_{FD}):

$$C_{FD} = \varepsilon_{ox}\left(\frac{A_3}{t_{ox2}} + \frac{A_4}{t_{ox3}}\right) \quad . \tag{2.34}$$

A diferencia del modelo tratado anteriormente, en que con una sola ecuación se modelizaba tanto la región lineal como la de saturación, este modelo separa ambas regiones, sin embargo, es posible modelar las características de corriente-voltaje para dispositivos de canal largo y dispositivos con longitud de canal menor a 5 m. Aquí se mostrará únicamente el segundo caso, donde se considera que los electrones llegan a una velocidad de saturación debido a un campo eléctrico alto. Resolviendo numéricamente un sistema de ecuaciones se encuentran los valores de corriente de drenador con su respectivo voltaje de drenador, para diferentes voltajes de compuerta aplicados, conociéndose al mismo tiempo, el voltaje de la compuerta flotante y la carga almacenada en ella. Con fines de simplicidad, se omiten las ecuaciones para obtener la curva de I_D-V_D, pero el desarrollo completo se puede consultar en la referencia [13]. En la Fig. 2.16 se presenta una curva obtenida a partir de resolver dichas ecuaciones.

Los modelos anteriores son adecuados para modelar las características I-V del dispositivo, pero aún no están incorporados como modelos compatibles con los programas de simulación, por lo que es necesario desarrollar uno que se pueda acoplar al programa de simulación que se emplea en este trabajo, como el PSpice.

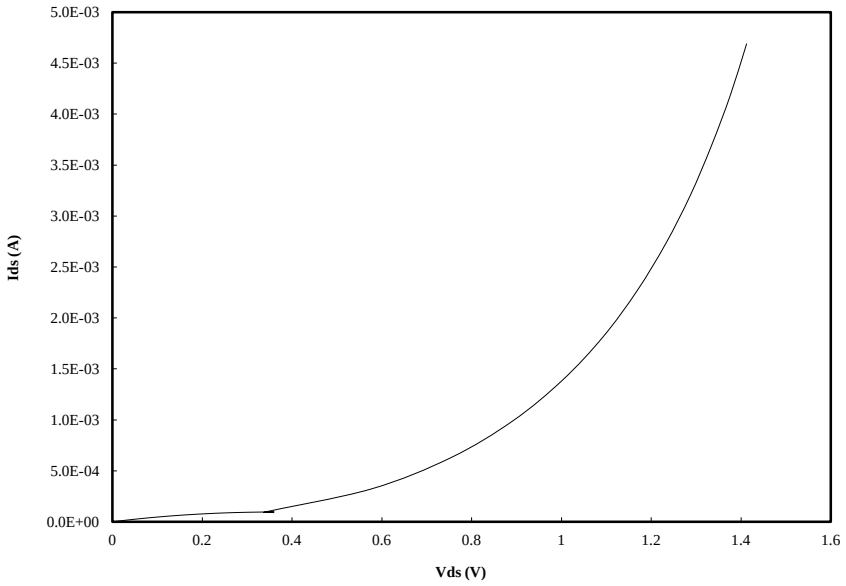


Fig. 2.16. Características I-V obtenidas a partir del modelo presentado en [13].

2.2. Tunelamiento Fowler-Nordheim.

Ya se ha mencionado que es necesario inyectar electrones en la compuerta flotante de un FGMOS para modificar su voltaje de umbral y tener de esa manera un incremento del peso requerido para el proceso de aprendizaje de una RNA. En esta sección, se explicará cómo es posible realizar dicha inyección de portadores (o también la eliminación de electrones, como el proceso de borrado) para introducir (extraer) carga en la compuerta flotante.

Varios son los mecanismos mediante los cuales un electrón puede ser introducido en la compuerta flotante [14]:

a) electrones calientes de substrato (SHE).
b) electrones calientes de canal (CHE).
c) portadores calientes por avalancha de drenador (DAHC).
d) electrones calientes generados secundariamente (SGHE).
e) recombinación Auger (AR).
f) tunelamiento Fowler-Nordheim (FN).
g) tunelamiento directo.

Los primeros cinco mecanismos involucran portadores calientes, es decir, estos no están en equilibrio con la red en el punto de inyección, ya que son capaces de absorber suficiente energía proveniente del campo eléctrico y la cual está por encima de la energía proporcionada por la temperatura. Los dos últimos casos (tunelamiento) involucran portadores que están en equilibrio con la red en el punto de inyección y se vuelven calientes hasta que llegan a la compuerta. Cada uno de ellos tiene sus ventajas y limitaciones y su elección dependerá de la factibilidad de su implementación así como de los efectos secundarios que pudiera causar cada uno en fases como la de programación y borrado del dispositivo.

Trabajos reportados en la literatura acerca de los FGMOS y donde indican el mecanismo usado [6, 8, 10, 11, 12, 13, 16, 17, 18], emplean preferentemente dos mecanismos para programación y borrado: portadores calientes por avalancha de drenador y tunelamiento Fowler-Nordheim. La razón principal por la que se usan, tiene que ver sobre todo con la facilidad tecnológica, es decir, el acceso a las fábricas de silicio que cuentan con la infraestructura para la realización de circuitos integrados VLSI.

Muchas de las estructuras reportadas, como las mostradas en la sección anterior, incluyen en su diseño, un óxido ultrafino (< 15 nm) que es difícil de depositar si no se cuenta con la tecnología adecuada, que entre otras cosas, es de uso restringido, ya que sólo grandes compañías especializadas y de amplio mercado pueden contar con ella, tecnología enfocada fundamentalmente a dispositivos digitales Esto hace que las investigaciones y desarrollos se orienten hacia la búsqueda de mecanismos que sean de fácil implementación con tecnologías accesibles como con la que cuentan fábricas de silicio cuyo producto es el servicio de manufacturación de circuitos integrados. En otras palabras, el costo y la facilidad de fabricación son un criterio importante para el estudio y concepción de circuitos utilizados en las RNA's, así como la compatibilidad con los procesos para la electrónica analógica.

En las estructuras de compuerta flotante, la programación ha sido realizada ya sea por electrones calientes o por tunelamiento FN [15], pero invariablemente, el borrado se ha realizado exclusivamente por tunelamiento FN. El hecho de utilizar inyección por electrones calientes, se debe a que el óxido de tunelamiento se encuentra colocado por encima del drenador y es suficiente con alcanzar las condiciones de avalancha en la región de deserción del drenador para crear electrones que se dirijan a la compuerta flotante debido al campo eléctrico presente en las terminales del dispositivo. Los voltajes utilizados son menores al voltaje de ruptura de la unión y de óxido y llega a ser atractivo como mecanismo de programación. Por otro lado, un fenómeno que se presenta paralelamente, es la creación de corrientes de huecos hacia el substrato, que en caso de ser excesiva, puede causar efectos no deseados como el efecto bipolar, fluctuaciones del voltaje de umbral [19] y ruptura inducida por avalancha [14], entre otros, además, que el espesor de óxido requerido sobre la región de drenador debe ser menor a 15 nm.

Durante el borrado, se requiere extraer los electrones de la compuerta flotante y por lo tanto no es necesaria la ionización presente en una avalancha, sino que es suficiente con crear un campo eléctrico de la magnitud y sentido adecuados para la extracción de los electrones. Esto sin embargo, implica campos eléctricos más intensos, lo que se traduce en mayores voltajes en las terminales del dispositivo, junto con el riesgo de degradar o incluso hasta destruir al óxido de tunelamiento, atrapamiento de carga y degradación de la ventana de programación. Se necesita también, un diseño correcto de las dimensiones de las compuertas de control y flotante, para tener un factor de acople capacitivo lo más alto posible, de tal modo que el voltaje aplicado a la compuerta de control, se vea reflejado lo más fielmente posible sobre la compuerta flotante. El factor de acople capacitivo es la razón de la capacitancias existente entre la compuerta de control y la compuerta flotante, y la suma de todas las capacitancias entre la compuerta flotante y las demás terminales; éste parámetro se explicará mas adelante.

Una estructura que ha demostrado un gran potencial en la aplicación de FGMOS en las RNA's, es la reportada en [10], en la cual se utiliza el mecanismo de FN tanto para inyectar como para extraer electrones de la compuerta flotante y es de la que se parte para la realización del circuito que se reporta en esta tesis. Debido a ello, se presentará una explicación mas amplia de este tipo de tunelamiento en la presente sección.

La Fig. 2.17, muestra un diagrama de bandas de un sistema metal-SiO_2-silicio, donde se puede apreciar que en equilibrio (Fig. 2.17a), existe una barrera de potencial de aproximadamente 3.2 eV que impide el paso de los electrones presentesen el silicio hacia el metal. Sin embargo, a temperatura ambiente, los electrones tienen la suficiente energía como para tener una alta probabilidad de recorrer aproximadamente 5 nm dentro del óxido, pero si el potencial existente en la interfaz SiO_2-silicio está por debajo de los 3.2 eV, regresarán al silicio sin la oportunidad de atravesar todo el óxido (se supone un óxido mayor a 5 nm). Pero cuando es aplicado un voltaje capaz de crear un campo eléctrico intenso, representado en la Fig. 2.17(b) y 2.17(c) por la inclinación de la banda de conducción del óxido, lo suficientemente grande como para permitir que el espesor de la barrera triangular o trapezoidal enfrente de la banda de conducción del silicio, sea de aproximadamente 5 nm, la probabilidad de tunelamiento se incrementa y se presentará una corriente de electrones hacia el metal. El valor mínimo del campo eléctrico para que esto suceda se puede encontrar de dividir la altura de la barrera que el óxido de silicio presenta a los electrones, entre el espesor mínimo de tunelamiento: 3.2 eV/5 nm = 6.4×10^8 V/m. Se ha reportado como posible el tunelamiento FN con espesores de óxido hasta de 70 nm.

Experimentalmente, este campo puede cambiar según la tecnología aplicada en la fabricación de las estructuras de compuerta flotante y el diseño de los inyectores, ya que el sistema de inyección puede estar formado por dos capas de silicio policristalino, correspondiendo en este caso a la compuerta de control y a la compuerta flotante en lugar del sistema explicado con anterioridad. Esto quiere decir que se puede elegir la inyección de electrones desde el substrato o bien desde el polisilicio 2, según sea conveniente. Para un sistema como el último, se ha visto que la inyección desde el polisilicio de la compuerta de control (polisilicio 1) hacia el polisilicio (polisilicio 2) de la compuerta flotante (inyección de electrones), requiere un campo eléctrico mínimo de aproximadamente 1.37 x 10^8 V/m y en sentido inverso (extracción de electrones), se necesita un campo eléctrico mínimo de aproximadamente 1.01 x 10^8 V/m [9]. Se especula que el cambio puede ser debido a la diferencia de textura en la superficie de ambos silicios policristalinos. Estos son los valores que se utilizarán en el cálculo de los voltajes que se requieren aplicar a la estructura empleada en esta tesis, ya que el sistema de inyección corresponde a este caso.

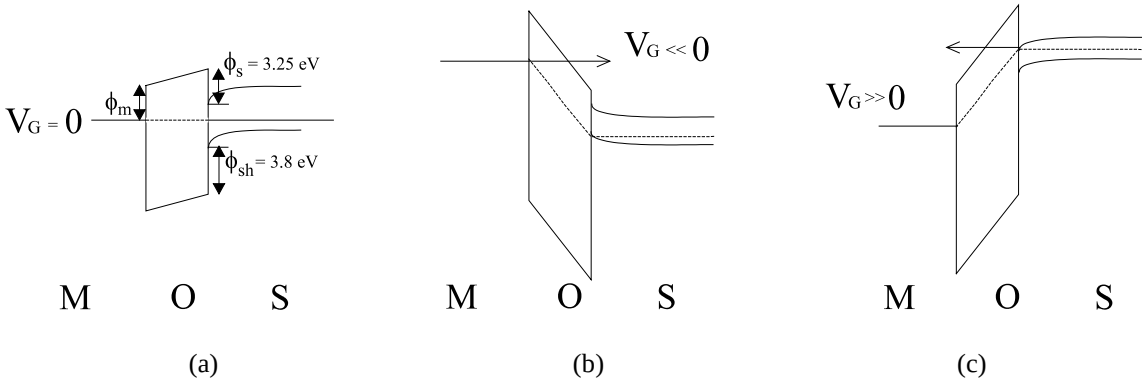


Fig. 2.17. Diagrama de bandas explicando el tunelamiento FN, a) en equilibrio (V_G=0), b) V_G<0, c) V_G>0.

Una característica que tiene el tunelamiento por FN, es que corresponde a un fenómeno autolimitado, es decir, cuando se tiene el campo eléctrico constante que produce la corriente de electrones a través del óxido, su magnitud es tal que permite el flujo de electrones para inyectar una cantidad de carga determinada por la magnitud del campo, pero a medida que éstos últimos quedan atrapados en el silicio policristalino, en un momento dado, la carga almacenada provoca que el campo eléctrico disminuya por debajo del necesario para el flujo de electrones, de manera análoga como sucede cuando se establece la región de deserción en una unión p-n en equilibrio. Esta característica es parecida a la presente cuando se borran las memorias con luz ultravioleta (UV), con la diferencia de que, según la cantidad de carga presente en la compuerta flotante inyectada por FN, se puede establecer de nuevo un campo eléctrico que aumente o disminuya la cantidad de carga, obviamente el voltaje aplicado deberá ser mayor.

Un modelo aproximado de la corriente de tunelamiento por Fowler-Nordheim, está indicado por la ecuación (2.1), repetida aquí por conveniencia:

$$J = AE^2 \exp\left(-\frac{B}{E}\right) \quad . \tag{2.35}$$

No existe mucho acuerdo en cuanto a los valores que se utilizan para A y B, pero se pueden tomar los reportados en [7]:

$$A = 1.88 \times 10^{-6}\ A/V^2$$

$$B = 2.55 \times 10^8\ V/cm.$$

2.3. Memoria de compuerta flotante de doble inyector.

Ya se ha establecido que un dispositivo adecuado para ser utilizado como sinapsis es el FGMOS, por la propiedad de lograr la multiplicación escalar y el almacenamiento del peso al mismo tiempo. La forma de programación y borrado es otro criterio que influye en la elección de la estructura apropiada para las aplicaciones que se tengan como objetivo. Se mencionó que es posible aplicar inyección por electrones calientes o por tunelamiento FN durante la programación e invariablemente, tunelamiento FN para el borrado. Los voltajes necesarios para la inyección por cualquiera de los dos fenómenos, deben de ser considerados ya que también es deseable que sean lo más bajos posible, para poder tener una mayor integración de las fuentes dentro del circuito, además de evitar alcanzar valores críticos de ruptura de las uniones y del óxido.

Una estructura reportada en la literatura y que tiene gran potencial como elemento de memoria analógica en diferentes aplicaciones, aparte de las RNA's, es la que se indica en la referencia [20]. Esta estructura se puede fabricar mediante los servicios ofrecidos por compañías como Orbit y consta de un transistor MOS, una compuerta flotante y una compuerta de control. Tiene además, dos inyectores con los que se puede tener un mejor control de la carga inyectada o extraída de la compuerta flotante, mediante el fenómeno de tunelamiento FN. La Fig. 2.18 muestra el diseño de este tipo de estructura.

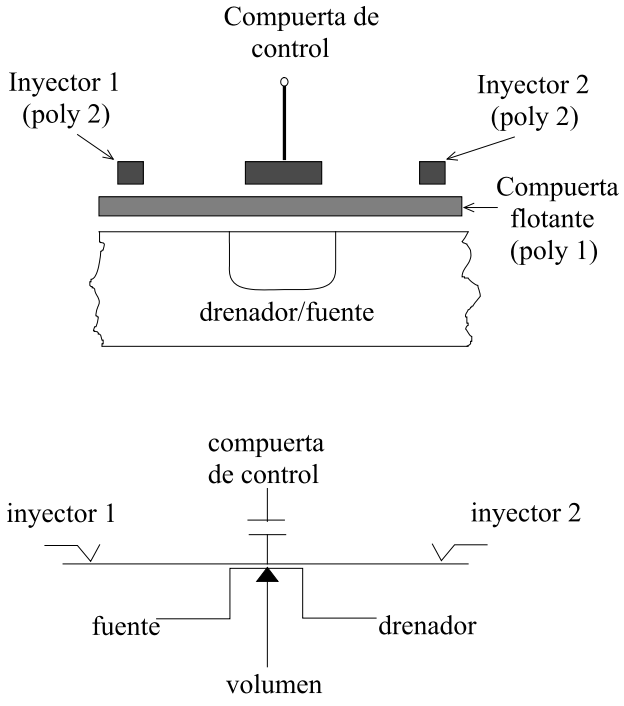


Fig. 2.18. Memoria de compuerta flotante con dos inyectores.

Como se puede observar, las terminales con las que cuenta este dispositivo son: drenador, fuente, substrato, compuerta de control, inyector 1 e inyector 2. Basándose en los parámetros de Orbit [21], el polisilicio 1 está separado del substrato en la región de compuerta por un óxido de 40 nm de espesor, el polisicio 1 y el polisilicio 2 están separados por un óxido de 70 nm tanto en la región de compuerta, como en la región de los inyectores y será el que se considere como el óxido de tunelamiento. El área de traslape

del polisicio de los inyectores sobre el polisilicio de la compuerta flotante puede ser tan pequeño como 4 m^2, pero el área de traslape del polisilico de la compuerta de control sobre la compuerta flotante, requiere que sea lo suficientemente grande, para lograr un acople capacitivo cercano a uno, para poder inducir un voltaje lo más parecido al voltaje aplicado a la compuerta de control, lo que será tratado en la sección 2.3.1.

Los inyectores tienen un papel importante en la modificación del voltaje de umbral, ya que al aplicar voltaje simultáneamente en la compuerta de control y en uno de los inyectores, es posible establecer un campo eléctrico a través del óxido de tunelamiento, de tal forma que exista una corriente fluyendo hacia o desde la compuerta flotante. El sentido de la corriente dependerá de la polaridad del voltaje aplicado en ambas terminales, existiendo un flujo de electrones hacia la compuerta flotante cuando el inyector y la compuerta de control tengan polarización negativa, y en sentido inverso, cuando ambas tengan polarización positiva. Si se mantienen fijas las polarizaciones de los inyectores (uno con voltaje positivo y otro con voltaje negativo), cambiando únicamente la de la compuerta de control, el sentido del flujo de corriente de electrones será controlada por la señal aplicada en ésta, sin que se presente simultáneamente un flujo en ambas direcciones. Esto se explicará más claramente en la sección 2.3.2.

2.3.1. Coeficiente de acoplamiento capacitivo.

El transistor MOS de enriquecimiento incluido en las memorias de compuerta flotante, requiere que le sean aplicados voltajes en sus terminales para que exista flujo de corriente entre drenador y fuente. En estructuras MOS clásicas, se necesita aplicar un voltaje en la compuerta para poder formar el canal entre drenador y fuente, lo cual se puede hacer ya que la terminal de compuerta está accesible. Sin embargo, cuando el dispositivo consta de una compuerta flotante a la cual no se puede acceder directamente para aplicar un voltaje, con la finalidad de poder conservar la carga presente, se debe inducir un voltaje sobre la compuerta flotante por medios indirectos. Una forma apropiada de lograrlo, es mediante un divisor de voltaje formado con capacitores, para aprovechar la retención de carga propia de estos dispositivos. En otros términos, esto correspondería a acoplar capacitivamente el voltaje aplicado en la compuerta de control, sobre la compuerta flotante. En la estructura de la Fig. 2.18, se tienen varias terminales separadas por un dieléctrico, en este caso SiO_2, que tienen como punto común la compuerta flotante, lo que da origen a un circuito eléctrico equivalente, como el mostrado en la Fig. 2.19.

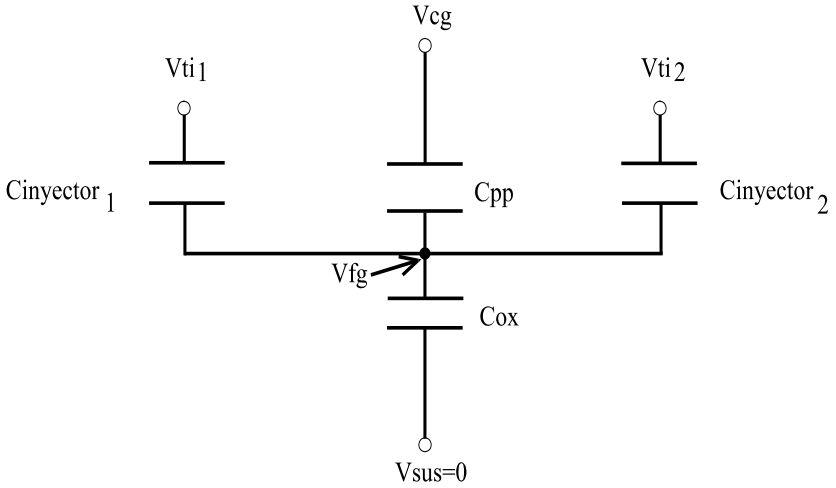


Fig. 2.19. Circuito equivalente de la estructura de compuerta flotante, para el cálculo del coeficiente de acoplamiento.

En este caso, se pueden distinguir cuatro capacitancias formadas con el diseño anterior:

Cpp: capacidad formada entre compuerta de control y compuerta flotante.
Cox: capacidad formada entre compuerta flotante y canal.
$Cinyector_1$: capacidad formada entre compuerta flotante e inyector 1.
$Cinyector_2$: capacidad formada entre compuerta flotante e inyector 2.

El voltaje en la compuerta flotante, Vfg, se puede encontrar por superposición considerando cada uno de los voltajes Vcg, Vti1 y Vti2. Esto se podría expresar de la siguiente forma:

$$Vfg = \sum_i KiVi + KcgVcg \quad , \tag{2.36}$$

donde:

$$Ki = \frac{Ci}{Ctot} \quad , \tag{2.37}$$

$$Kcg = \frac{Cpp}{Ctot} \quad , \tag{2.38}$$

$$Ctot = \sum Ci + Cpp \quad . \tag{2.39}$$

Vi corresponde a los voltajes aplicados en terminales diferentes a la compuerta de control y Vcg es el voltaje aplicado en la compuerta de control; Ci es la capacidad asociada con la terminal correspondiente. Las ecuaciones (2.37) y (2.38) expresan a los coeficientes de acoplamiento, dependientes de las capacidades, que a su vez son función directa del área y función inversa de la separación entre electrodos, es decir, entre mayor área tengan, mayor será la capacidad y entre menor separación tengan los electrodos, mayor será su capacidad. Lo que interesa, es que el voltaje sobre la compuerta flotante sea lo más parecido al voltaje aplicado a la compuerta de control y en consecuencia, se debe buscar que la capacidad Cpp sea lo más alta posible, pero cumpliendo con un compromiso debido a que Cpp también se encuentra en el denominador de (2.37). Ya que la separación del óxido entre polisilicio 1 y polisilicio 2 ya está fijo por la tecnología, el grado de libertad se encuentra por el lado del área, conduciendo esto a diseñar un traslape grande entre la compuerta de control y la compuerta flotante. Como la función de los inyectores es crear únicamente un campo eléctrico y no un acople de voltaje, el área de traslape entre inyectores y compuerta flotante, puede se tan pequeña como (2x2) m^2, lo que reduce enormemente el coeficiente de acoplamiento Ki y por lo tanto, cuando no existe carga presente en la compuerta flotante (Qfg = 0), el término dominante en la ecuación (2.36) resulta ser KcgVcg. Una desventaja que esto trae, es que el área total del dispositivo se ve incrementada, reduciendo por tanto la capacidad de integración.

La Fig. 2.20 muestra la tendencia del coeficiente de acoplamiento con respecto a un cambio en la capacidad Cpp por aumento del área de traslape. Al graficar la ecuación (2.38), se consideraron los parámetros tecnológicos de Orbit para las capacidades $Cinyector_1$, $Cinyector_2$ y Cox.

Como se puede observar, al aumentar la capacidad, con el consecuente aumento de área, la pendiente de la curva disminuye, siendo más difícil que Kcg alcance el deseado valor de 1; en este caso, la máxima capacidad graficada, corresponde a un área de 100 m^2, con lo que se alcanza como extremo para Kcg, el valor de 0.666.

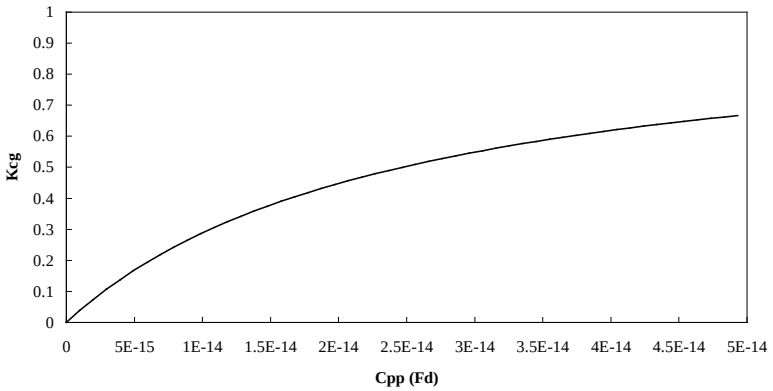


Fig. 2.20. Coeficiente de acoplamiento capacitivo Kcg en función de la capacidad entre polisilicios, Cpp.

2.3.2. Inyección y extracción de carga.

La programación y borrado, para alterar el voltaje de umbral de la estructura mostrada en la Fig. 2.18, se realiza a través del óxido de tunelamiento que se encuentra entre los inyectores y la compuerta flotante. Para esto, se necesita establecer un campo eléctrico con el que se tenga un flujo de corriente de electrones, aplicando una polarización entre estos electrodos. Como existe un voltaje inducido Vfg, debido al coeficiente de acoplamiento, es necesario calcular el voltaje que se debe aplicar a los inyectores para tener la diferencia de potencial mínima para tener la condición de tunelamiento. De ésta manera, es posible alterar la cantidad de carga en la compuerta flotante, con el consecuente cambio del voltaje de umbral del dispositivo, lo cual se puede aprovechar para cambiar su conductancia y aprovechar tal hecho para aplicarlo en sistemas que requieran modificar la resistencia, como es el caso de las redes neuronales.

Para conocer con certeza el campo eléctrico aplicado al óxido de silicio que se encuentra entre los inyectores de polisilicio 1 y polisilicio 2, se requiere conocer el voltaje presente en la compuerta flotante (Vfg). El voltaje Vfg, cuando existe carga presente, se puede calcular con la siguiente ecuación [20]:

$$Vfg = Vo + KcgVcg + \frac{Qfg}{Ctot} \quad , \tag{2.40}$$

donde Qfg es la carga presente en la compuerta flotante y Vo representa la suma de los voltajes constantes en otras terminales.

En la ecuación (2.40), los tres términos de la derecha dependen de factores geométricos dados en el diseño por las áreas de las capacitancias y también de los voltajes aplicados. El parámetro que permite cambiar el voltaje de umbral, es la carga Qfg y su relación es de la siguiente forma:

$$Qfg = \Delta Vth * Cpp \quad , \tag{2.41}$$

donde ΔVth es la magnitud del cambio del voltaje de umbral y Cpp es la capacitancia de la compuerta de control. Por lo tanto, si se conoce Cpp y el cambio deseado para el voltaje de umbral, se puede calcular la carga en la compuerta flotante.

En la Fig. 2.21, se grafica la ecuación (2.40) para tres casos distintos: 1) ΔVth = 0, 2) ΔVth = ±1V y 3) ΔVth = ±3 V, con un Kcg = 0.4. Se puede ver que para el primer caso, la recta pasa por el origen y corresponde a una carga Qfg = 0; el segundo caso, pasa desplazada por encima del origen para un

desplazamiento ΔVth = -1 V, correspondiendo a una carga positiva y, por debajo para ΔVth = +1 V para una carga negativa. Lo mismo sucede para el tercer caso, donde la recta cruza por magnitudes mayores en el eje de Vfg, respecto al caso (2). Con esta gráfica se puede observar el cambio de Vfg en función del voltaje de la compuerta de control, Vcg, teniéndose la carga Qfg como parámetro.

Por otro lado, también se requiere conocer la caída de potencial mínima en el óxido de tunelamiento (inyectores), necesaria para comenzar la corriente a través del óxido. Con los campos indicados anteriormente, junto con los datos tecnológicos de Orbit (separación entre poly1 y poly2), esta caída de potencial se puede evaluar como sigue:

$$V_{p2 \to 1} = E_{c2 \to 1} * e_{ox} \quad , \tag{2.42}$$

$$V_{p1 \to 2} = E_{c1 \to 2} * e_{ox} \quad , \tag{2.43}$$

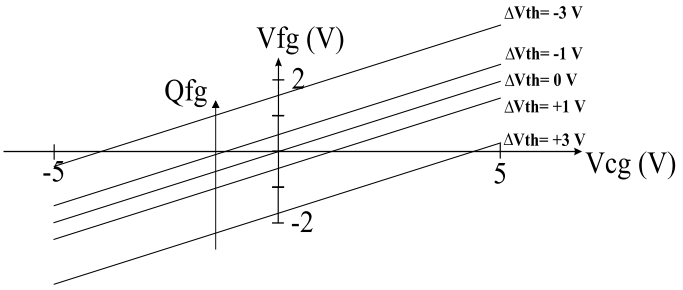


Fig. 2.21. Vfg en función del voltaje de la compuerta de control Vcg.

donde: $Vp_{2\to1}$: potencial del inyector 1 (V)
$Vp_{1\to2}$: potencial del inyector 2 (V)
$Ec_{2\to1}$:$1.37x10^8$ V/m
$Ec_{1\to2}$:$1.01x10^8$ V/m
e_{ox}: 78 nm (se considera este espesor como el peor de los casos tomando en cuenta la dispersión en el proceso).

Sustituyendo los valores en las ecuaciones (2.42) y (2.43), se obtienen dos valores importantes para la inyección de carga:

$$Vp_{2\to1} = -10.7 \text{ V}$$

$$Vp_{1\to2} = 7.9 \text{ V}.$$

El signo, en este caso, indicará hacia dónde es la inyección de los electrones: -10.7 V es con referencia al voltaje aplicado al inyector 1, lo que significa que se inyectarán electrones hacia la compuerta flotante; 7.9 V es con referencia al inyector 2, indicando extracción de electrones de la compuerta flotante. Así entonces, existirá corriente de tunelamiento cuando la diferencia de potencial entre el inyector 1 y la compuerta flotante, esté por debajo de -10.7 V o cuando la diferencia de potencial entre el inyector 2 y la compuerta flotante esté por encima de 7.9 V. Esto se puede expresar como sigue:

$$I_{tunel} = 0 \text{ para } V_{p1 \to 2} \rangle Vti - Vfg \rangle V_{p2 \to 1} \quad . \tag{2.44}$$

Esta desigualdad se puede graficar por separado para cada inyector, en función del voltaje aplicado en la compuerta de control, sustituyendo las siguientes igualdades en el límite de tunelamiento, dentro de la ecuación (2.40):

$$Vti_2 - Vfg = Vp_{1 \to 2} \quad , \tag{2.45}$$

$$Vti_1 - Vfg = Vp_{2 \to 1} \quad . \tag{2.46}$$

En la Fig. 2.22 se muestra el cambio del voltaje de los inyectores en función del voltaje de la compuerta de control, para el inyector 1 (Fig. 2.22a) y para el inyector 2 (Fig. 2.22b), para $\Delta Vth = \pm 3$ V y $\Delta Vth = \pm 10$ V.

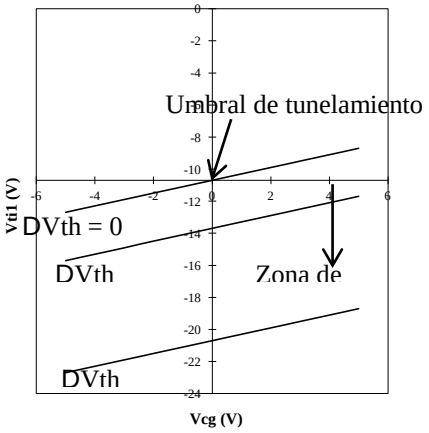
Para un funcionamiento apropiado de la estructura, se debe cumplir que no debe haber tunelamiento para Vcg = 0, sin embargo, los voltajes Vti en Vcg = 0 en 2.22(a) y 2.22(b), junto con Vfg, indican que están justo en el inicio del tunelamiento, esto es:

ΔVth = -10 V $\rightarrow$ Vt i= 17.9 V y Vfg = 10 V
Vti - Vfg = 7.9 V = $Vp_{1\rightarrow2}$

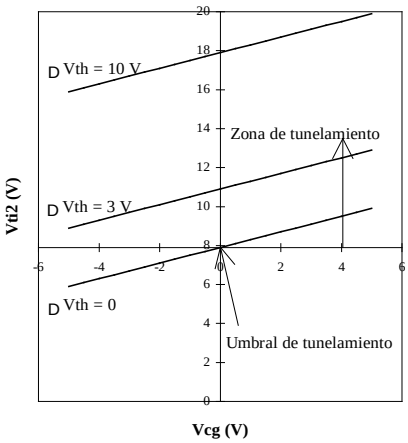
ΔVth = -3 V $\rightarrow$ Vti = 10.9 V y Vfg = 3 V
Vti - Vfg = 7.9 V = $Vp_{1\rightarrow2}$

ΔVth = +10 V $\rightarrow$ Vti = -20.7 V y Vfg = -10 V
Vti - Vfg = -10.7 V = $Vp_{2\rightarrow1}$

ΔVth = +3 V $\rightarrow$ Vti = -13.7 V y Vfg = -3 V
Vti - Vfg = -10.7 V = $Vp_{2\rightarrow1}$



(a)



(b)

Fig. 2.22. Gráfica del cambio de Vti con respecto a Vcg para ΔVth = ±3 V y ± 10 V.
a) inyector 1.
b) inyector 2.

Por lo tanto, el voltaje en los inyectores debe cambiarse para cumplir la condición anterior, junto con el requisito establecido en la ecuación (2.44). Un criterio utilizado en [20] es mantener 1 Volt de diferencia entre el voltaje Vp y la diferencia de potencial Vti - Vfg, para controlar la inyección con Vcg. Esto equivale entonces a tener los voltajes indicados en la Tabla 2.2, para los casos en que se quiera cambiar el voltaje de umbral del FGMOS en el intervalo de ΔVth = ±10 V.

Es importante hacer notar también que nunca habrá inyección simultánea en los dos inyectores, ya que el voltaje que controla esto es Vcg y, como se puede ver en la Fig. 2.22, cuando Vcg toma valores negativos favorece la inyección de electrones por el inyector 1 y desfavorece la misma en el inyector 2, y viceversa cuando Vcg toma valores positivos, esto, independiente de si se trata de un FGPMOS o un FGNMOS, ya que lo único que se hace es cambiar la cantidad de carga en la compuerta flotante, sin tomar en cuenta si el substrato es tipo N o tipo P. El comportamiento de los transistores en este sentido se verá en la sección 4.1.2.

Tabla 2.2. Voltaje aplicado en el inyector 1 e inyector 2,
para diferentes variaciones de voltajes de umbral.

ΔVth	Vti_1	ΔVth	Vti_2
+10 V	-16.9 V	-10 V	19.7 V
+9 V	-15.9 V	-9 V	18.7 V
+8 V	-14.9 V	-8 V	17.7 V
+7 V	-13.9 V	-7 V	16.7 V
+6 V	-12.9 V	-6 V	15.7 V
+5 V	-11.9 V	-5 V	14.7 V
+4 V	-10.9 V	-4 V	13.7 V
+3 V	-9.9 V	-3 V	12.7 V
+2 V	-8.9 V	-2 V	11.7 V
+1 V	-7.9 V	-1 V	10.7 V

Nota: esta tabla es válida para la tecnología Orbit,
2 μm, doble polisilicio, doble metal y pozo N.

Cuando se quiere inyectar carga negativa, se requieren voltajes Vcg < -2.5 V y, cuando se quiere extraer carga, se requieren voltajes Vcg > 2.5 V. Esto último se encuentra al usar el criterio de la diferencia de 1 Volt entre el Vp y Vti - Vfg. De esta manera, si se sabe anticipadamente el cambio deseado del voltaje de umbral de la estructura, manteniendo fijos los voltajes Vti en la magnitud correspondiente, se puede controlar este parámetro haciendo cambiar a Vcg.

2.4. Sumario.

Este capítulo comprendió la comparación de diferentes estructuras de compuerta flotante que se han reportado en la literatura, donde se resumen las diferentes características con las que cuentan cada una de ellas, tanto desde el punto de vista de factibilidad tecnológica, como desde el punto de vista de propiedades de programación y borrado. En este último aspecto, se comentaron los diferentes mecanismos que existen para la inyección de carga, concluyéndose que los más adecuados son los de electrones calientes por avalancha de drenador y el tunelamiento Fowler-Nordheim. El primero requiere de la creación de portadores por avalancha; el segundo, requiere de un campo eléctrico alto para provocar tunelamiento a través del óxido.

Dada la sencillez de diseño y control sobre la carga inyectada, la estructura reportada por Thomsen y Brooke aparece como una de las estructuras mas apropiadas para su aplicación en el diseño de circuitos para RNA's. La programación y borrado se hace en base al tunelamiento FN con ayuda de dos inyectores con los que se tiene un control adecuado de la carga almacenada y se explicó la forma como se presenta este fenómeno para un sistema metal-óxido-semiconductor, pero se ha encontrado que el campo eléctrico crítico para un sistema polisilicio-óxido-polisilicio es diferente al del primero.

Un elemento importante en el diseño topológico de la estructura, corresponde al factor de acoplamiento Kcg, con el que se puede inducir un voltaje sobre la compuerta flotante para que pueda funcionar el transistor MOS incluido en la estructura y para poder polarizar esta terminal con la finalidad de poder crear un campo eléctrico sobre el óxido de tunelamiento. Se requiere que Kcg sea lo más cercano posible a 1, lo que involucra aumentar el área de traslape de la compuerta de control sobre la compuerta flotante, pero sin comprometer el área a instancias de mejorar el parámetro Kcg.

Finalmente, en base a los parámetros tecnológicos planteados por la tecnología de Orbit, en este trabajo se dedujeron los voltajes que se han de aplicar en los inyectores para modificar el voltaje de umbral de una estructura de compuerta flotante con dos inyectores, equivalente a las fases de escritura y borrado.

Referencias.

1.- J. M. Zurada. "Analog implementation of neural networks", *Circuits and Devices Magazine*, Sep. 1992, pp. 36-41.

2.- S. Y. Fuo, L. R. Anderson and Y. Takefuji. "Analog components for the VLSI of neural networks", *Circuits and Devices Magazine*, July 1990, pp. 18-26.

3.- A. A. Amin. "Design, selection and implementation of flash erase EEPROM memory cells", *IEE Proceedings-G*, Vol.139, No. 3, June 1992, pp. 370-376.

4.- A. Concannon, S. Keeney, A. Mathewson, R. Bez and C. Lombardi. "Two-dimensional numerical analysis of floating-gate EEPROM devices", *IEEE Trans. on Elect. Dev.*, Vol. 40, No. 7, July 1993, pp. 1258-1262.

5.- S. Keeney, A. Mathewson, L. Ravazzi and C. Lombardi. "Simulation of EPROM device programming using the hydrodynamic model", *Microelectronic Engineering* 19 (1992) pp. 261-264.

6.- A. J. Montalvo and J. J. Paulos. "Improved Floating-gate devices using standard CMOS technology", *IEEE Elec. Dev. Lett.*, Vol. 14, No. 8, August 1993, pp. 372-374.

7.- A. Kolodny, S. T. K. Nieh, B. Eitan and J. Shappir. "Analysis and modeling of floating-gate EEPROM cells", *IEEE Trans. Elec. Dev.*, Vol. ED-33, No. 6, June 1986, pp. 835-844.

8.- G. Samachisa et. al., "A 128K Flash EEPROM using double polisilicon technology", *IEEE Journal of Solid State Circuits*, Vol. SC-22, No. 5, October 1987, pp. 676-683.

9.- D. A. Durfee and F. S. Shoucair. " Comparison of floating-gate neural network memory cells in standard VLSI CMOS technology", *IEEE Trans. on Neural Networks*, Vol. 3, No. 3, May 1992, pp. 347-352.

10.-A. Thomsen and M. A. Brooke. "A floating-gate MOSFET with tunneling injector fabricated using a standard double-polisilicon CMOS process", *IEEE Elec. Dev. Lett.*, Vol. 12, No. 3, March 1991, pp. 111-113.

11.-L. R. Carley. "Trimming analog circuit using floating-gate analog MOS memory", *IEEE Jour. of Solid State Circ.*, Vol. 24, No. 6, Dec. 1989, pp. 1569-1575.

12.-L. C. Liong and P. Liu. "A theoretical model for the current-voltage characteristics of a floating-gate EEPROM cell", *IEEE Trans. Elec. Dev.*, Vol. 40, No. 1, January 1993, pp. 146-151.

13.-C. K. Sin, U. H. Robert and P. K. Ko. "EEPROM as an analog storage device, with particular applications in neural networks", *IEEE Trans. Elec. Dev.*, Vol. 39, No. 6, June 1992, pp. 1410-1419.

14.-J. J. Sánchez and T. A. DeMassa, "Review of carrier injection in the silicon/silicon dioxide system", *IEE Proceedings-G*, Vol. 138, No. 3, June 1991, pp. 377-389.

15.-M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO2", *J. Appl. Phys.*, Vol. 40, 1969, pp. 278-283.

16.-K. Yang and A. G. Andreou, "Subthreshold analysis of floating-gate MOSFET's", *Proc. Tenth Biennial VGIM Symposium*, Research Triangle Park, NC, May 1993, pp. 141-144.

17.-E. Säckinger and W. Guggenbühl, "An analog trimming circuit based on a floating-gate device", *IEEE Jour. Of Solid State Circuits*, Vol. 23, No. 6, Dec. 1988, pp. 1437-1440.

18.-A. A. M. Amin, "Speed optimised array architecture for flash EEPROM's", *IEE Proceedings-G*, Vol. 140, No. 3, June 1993, pp. 177-181.

19.-A. E. El-Hennawy, G. G. Al-Barakati and T. L. Al-Harbi, "Threshold-voltage instability by the hot-carrier substrate current in MOSFET's", *Int. J. Electronics*, Vol. 75, No. 1, 1993, pp. 49-55.

20.-A. Thomsen and M. A. Brooke, "Low control voltage programming of floating-gate MOSFET's and applications", *IEEE Trans. on Circuits and Systems*, Vol. 41, No. 6, June 1994, pp. 443-451.

21.-Orbit, *Orbit Electrical Parameter Set* (1993).

Capítulo 3.

Diseño de una RNA tipo BAM.

En el Capítulo 1 se mencionaron algunas de las arquitecturas más comunes en el estudio de las Redes Neuronales Artificiales. De entre ellas, la Memoria Asociativa Bidireccional (BAM) se presenta como una de las arquitecturas con mayor atractivo para su implementación en circuitos integrados, dada la alta simetría de su arquitectura, lo que permite la integración de una RNA en un circuito pequeño, y lo ilustrativo de su comportamiento como memoria asociativa representada por una matriz sencilla, capaz de aprender una serie de patrones en el marco de las redes neuronales.

Este tipo de arquitectura, se presta también para ensayar la aplicación del elemento estudiado como sinapsis, en base a la memoria analógica utilizando transistores MOS de compuerta flotante. Mediante la aplicación de circuitos sencillos CMOS, es posible crear una configuración básica que incluya a la sinapsis y a la neurona, interconectados de tal forma que, en conjunto, realice la función buscada. Como vehículo de prueba acerca de la aplicación de estos elementos con almacenamiento tipo analógico, se propone el diseño, simulación y fabricación de las componentes estructurales de una arquitectura neuronal, por ejemplo una del tipo Memoria Asociativa Bidireccional. Aunque el objetivo fundamental de esta tesis es el diseño, simulación y caracterización de estructuras MOS de compuerta flotante, el tipo de estructuras de prueba y sus respectivos voltajes, corrientes y dimensiones, corresponden a elementos prácticos que se pueden aplicar en un diseño dado. Así, a continuación, se propone el diseño de una estructura de memoria asociativa, mencionando algo de su teoría de funcionamiento, su simulación, e inclusive algunas de las limitaciones que posee este tipo de arquitecturas.

3.1. Fundamentos teóricos.

Desde el punto de vista funcional, una memoria digital almacena un dato en una localidad definida dentro de un arreglo matricial y para extraer ese dato, se requiere indicar la dirección exacta donde se encuentra almacenado, es decir, es puntual. Una memoria asociativa, como la BAM, es capaz de almacenar datos de manera distribuida, es decir, no se tiene una dirección específica en donde localizar al dato, sino que se aprovecha todo el arreglo en conjunto, para correlacionar los datos que han sido almacenados y basta con indicar incluso solo parte de los datos que se quieren extraer, para obtenerlos, mediante una asociación. Esto trae como ventaja, que el dato correcto puede ser extraído cuando los requisitos presentados a la memoria, están incompletos, con errores o incluso si el arreglo de la memoria tiene algunos elementos defectuosos.

Como se mencionó anteriormente, una memoria asociativa puede clasificarse como heteroasociativa, cuando la dimensión del dato o patrón de entrada es diferente al de salida, o autoasociativa, cuando ambas dimensiones son iguales. El primer caso asocia eventos diferentes y el segundo caso asocia a un evento consigo mismo.

La BAM fue introducida por Kosko en 1987 [1, 2], y a partir de ahí, ha sido extensamente estudiada y utilizada para diversas aplicaciones en control [3], procesamiento de imágenes [4] o en reconocimiento de patrones [5]. Esta arquitectura ha causado tal interés, que se siguen proponiendo modificaciones al procedimiento de codificación de la red propuesta por Kosko [5, 6, 7, 8]. Esto se debe a que la BAM de Kosko tiene una capacidad limitada para reconocer patrones con errores y mediante modificaciones tanto a la arquitectura como al método de codificación, se logra mejorar su capacidad de reconocimiento.

Una Memoria Asociativa Bidireccional, consiste de dos capas de neuronas, donde cada una de las neuronas de una capa, está interconectada con todas las demás que forman la otra capa; no se interconectan neuronas de la misma capa ni cada neurona consigo misma. Esto conduce a tener un matriz de dimensión *nxm*, donde n es el número de neuronas en la capa A y m es el número de neuronas en la capa B, como se indica en la Fig. 3.1.

A este tipo de redes, se le puede presentar un vector A_k (a(k)1, a(k)2,...a(k)n) a la capa A y obtener una respuesta en la capa B, o bien, se le puede presentar un vector B_k (b(k)1, b(k)2,...,b(k)m) a la capa B y obtener una respuesta en la capa A, lo que corresponde a una asociación de parejas de vectores (A_k, B_k), donde k=1,....p, es decir, se pueden tener p parejas diferentes. Los vectores se pueden representar en forma binaria (ai, bi [0, 1]) o en forma bipolar (ai, bi [-1, 1]), existiendo mayor eficiencia de reconocimiento cuando se representan en forma bipolar [2]. Las conexiones entre capas corresponden a los pesos o sinapsis, formándose matrices de interconexión de dimensión *n*x*m*.

Una BAM opera relacionando la pareja de vectores (A_k, B_k) mediante una matriz de correlación, formada al multiplicar ambos vectores como se indica en la siguiente ecuación:

$$M_k = A_k^T B_k \quad , \tag{3.1}$$

donde para obtener la matriz resultante, se requiere multiplicar un vector, por la transpuesta del otro. Si se desea almacenar p patrones en una misma red, entonces es suficiente con hacer la suma de todas las matrices de correlación de cada una de las parejas:

$$M = M_1 + M_2 + ... + M_p \quad . \tag{3.2}$$

Obtenida la matriz M (con elementos m_{ij}) mediante la ecuación (3.2), uno de los vectores asociados se puede encontrar multiplicando al vector de entrada por M, o bien por la transpuesta de M (M^T), según sea el caso. Por ejemplo, cuando se usa como entrada al vector A_k, se puede encontrar B_k, de la siguiente manera:

$$B_k = A_k M \quad , \tag{3.3}$$

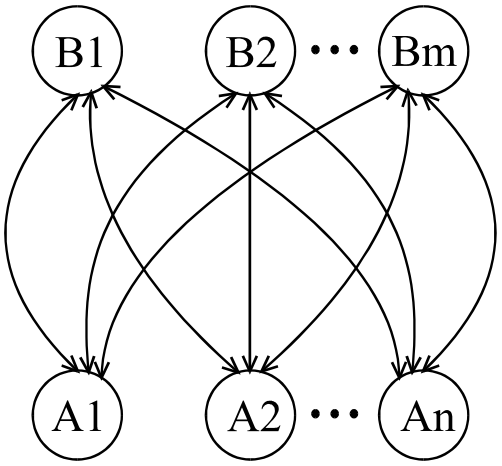


Fig. 3.1. Memoria Asociativa Bidireccional.

y cuando se usa B_k como entrada, se multiplica este vector por M^T, para encontrar al vector A_k:

$$A_k = B_k M^T \quad . \tag{3.4}$$

Por otro lado, la función de activación de las neuronas en cada una de las capas puede ser una función discreta, resultando en un conjunto bivalente; o una función contínua, donde se representa a un conjunto multivaluado, lo que distingue a una BAM discreta y a una BAM contínua, respectivamente. Estas funciones son como se expresa a continuación:

$$f(x) = \begin{cases} 1 & \text{si } x \geq \theta \\ 0 & \text{si } x \leq \theta \end{cases} \quad , \quad \text{para patrones binarios} \tag{3.5}$$

$$f(x) = \begin{cases} 1 & \text{si } x \geq \theta \\ -1 & \text{si } x \leq \theta \end{cases} \quad , \quad \text{para patrones bipolares} \tag{3.6}$$

$$f(x) = \frac{1}{1+\exp(-\sigma x)} \quad , \quad \text{para patrones binarios} \tag{3.7}$$

$$f(x) = \frac{2}{1+\exp(-\sigma x)} - 1 \quad . \quad \text{para patrones bipolares} \tag{3.8}$$

Las ecuaciones (3.5) y (3.6) se aplican a una BAM discreta y las ecuaciones (3.7) y (3.8) a una BAM continua. es un umbral de la función y es el parámetro de ganancia con un valor entre 0 y 1. En las cuatro expresiones anteriores, x es la suma de todas las señales que llegan a cada neurona y f(x) es la salida de la neurona. Expresado de otra manera, esto se puede ver como sigue:

$$x = a_i n_i = \sum b_j m_{ji} \quad , \tag{3.9}$$

$$a_i = f(x) \quad , \tag{3.10}$$

o bien:

$$x = b_i n_j = \sum a_i m_{ij} \quad , \tag{3.11}$$

$$b_i = f(x) \quad , \tag{3.12}$$

dependiendo de si se quiere obtener el vector A o el vector B, respectivamente.

En un momento dado, una BAM continua puede tender hacia la BAM discreta si se hace que el parámetro de ganancia sea cercano a uno, lo que correspondería a una sigmoide muy parecida a un escalón y esto simplifica el análisis de la red por la simplicidad de la función. En el mismo sentido, se dice que la BAM aprende la asociación de los patrones que se hayan elegido, sin embargo, mediante el uso de las ecuaciones (3.1) y (3.2) se infiere el valor de los pesos de interconexión entre las dos capas para k pares asociados y ya no se requiere cambiar el valor. Esto implica estrictamente que no existe un proceso de aprendizaje para la BAM, ya que para cumplir con este concepto, se debe cumplir que los pesos deben adaptarse conforme se presentan los patrones de entrenamiento y el par (A_k, B_k) forma por sí mismo su propio pozo de mínima energía.

Lo anterior se ve como una resonancia adaptativa entre los estados de ambas capas como se muestra en el siguiente esquema:

A M B,

$$
\begin{array}{lll}
A' & M^T & B, \\
A' & M & B', \\
A'' & M^T & B', \\
 & . & \\
 & . & \\
 & . & \\
A_f & M & B_f, \\
A_f & M^T & B_f, \\
 & . & \\
 & . & \\
 & . &
\end{array}
$$

donde la red entra rápidamente en equilibrio hacia los estados A_f y B_f, y se espera que este par sea igual al par almacenado (A_k, B_k) o por lo menos cercano a este y en consecuencia, la matriz de pesos es bidireccionalmente estable. Los valores m_{ij} cambiarán durante el proceso de resonancia, pero como se considera que los cambios en el tiempo de las sinapsis son considerablemente más lentos que los cambios en las activaciones de las neuronas, se pueden tomar los pesos constantes, como una aproximación válida.

En realidad, se puede pensar que conforme cambian las activaciones, estas influyen en el valor de las sinapsis para que vayan modificando su valor conforme se presentan los patrones de entrenamiento, lo que lleva a una expresión para los pesos de la siguiente forma:

$$\dot{m}_{ij} = -m_{ij} + S(a_i)\,s(b_j) \quad , \tag{3.13}$$

la cual es un algoritmo de aprendizaje de la forma de la regla de Hebb, con solución:

$$m_{ij}(t) = e^{-t} m_{ij}(0) + \left(1 - e^{-t}\right) \quad , \tag{3.14}$$

donde $\dot{m}_{ij}$ es la derivada de m_{ij} con respecto al tiempo, $S(a_i)$ y $S(b_j)$ son las funciones de activación de las neuronas y $m_{ij}(0)$ es el peso de inicialización; esto sucede cuando $S(a_i)S(b_j) = 1$. El valor de m_{ij} en la ecuación (3.14) cuyo valor límite tiende exponencialmente hacia 1 de una manera rápida, independientemente de las condiciones iniciales, y la regla de Hebb de la ecuación (3.13) tiende asintóticamente a la ecuación (3.2), pudiéndose utilizar esta última para la programación de la red prototipo a diseñar. El algoritmo de aprendizaje es la base de la llamada BAM adaptativa [1].

La regla de Hebb es el método mas común y simple para determinar los pesos de una memoria asociativa y se puede expresar como sigue:

$$w_{ij} = w'_{ij} + x_i y_j \quad , \tag{3.15}$$

donde w_{ij} es el nuevo peso calculado a partir del peso anterior, w'_{ij}, del vector de entrada, **x** y del vector de salida, **y**. El producto del segundo término a la derecha de la ecuación (3.15) se encuentra justamente con la expresión de la ecuación (3.1) y se puede ver la justificación de porqué los pesos se encuentran rápidamente.

La correlación entre los vectores de entrada a la red, puede causar diferencia en la respuesta, dependiendo de si son ortogonales o no. Esto puede provocar que la respuesta de la red, después de la presentación de un vector de entrenamiento en la entrada, sea exactamente su vector asociado o una mezcla de todos vectores asociados, respectivamente. Para la explicación de este concepto, es adecuado introducir la definición de ortogonalidad. Se dice que dos vectores son ortogonales si su producto punto es igual a cero, lo que se puede expresar de la siguiente manera:

$$A(k)A^T(p) = 0 \quad , \tag{3.16}$$

$$\sum_{i=1}^{n} A_i(k)A_i(p) = 0 \quad , \tag{3.17}$$

donde A(k) y A(p) son dos vectores de entrada diferentes.

Repitiendo por conveniencia las ecuaciones (3.1) y (3.3) y expandiendo para generalizar, se tiene que:

$$M = \sum_{p=1}^{P} A^T(p)\,B(p) \quad , \tag{3.18}$$

$$B(k) = A(k)\,M \quad , \tag{3.19}$$

sustituyendo (3.18) en (3.19) y expandiendo:

$$\begin{aligned} A(k)\,M &= \sum_{p=1}^{P} A(k)\,A^T(p)\,B(p) \\ A(k)\,M &= A(k)\,A^T(k)\,B(k) + \sum_{p\neq k} A(k)\,A^T(p)\,B(p) \end{aligned} \quad , \tag{3.20}$$

de donde se puede ver que si se cumple la condición dada en (3.16) para dos vectores ortogonales, el término dentro de la sumatoria de (3.20) se elimina y el resultado será el vector B(k) multiplicado por un factor igual al cuadrado de la norma del vector de entrada A(k); a este resultado se le aplica la función de transferencia de la neurona y resulta finalmente en B(k). Sin embargo, cuando no se cumple (3.16), es decir, los vectores de entrada no son ortogonales entre sí, la respuesta incluirá valores de cada uno de los vectores con los que A(k) no es ortogonal.

Respecto al pozo de mínima energía mencionado anteriormente, este se puede comprender si se supone a la energía como una superficie plana, donde cada par asociado forma su propio pozo, llamado mínimo local. De esta manera, una vez entrenada la red y formados los pozos, cuando se presenta un vector similar a alguno de los vectores almacenados, análogamente correspondería a colocar una esfera cerca del pozo formado por este último, con una alta probabilidad de deslizarse hacia el pozo. Si por el contrario, el vector presentado no es similar, sería como si la esfera se estuviera colocando muy alejada del pozo, con una mínima probabilidad de caer en él.

Una restricción importante que se tiene con el método anterior de codificación o deducción de la matriz de correlación (M) para *p* vectores, es que no se pueden almacenar más vectores que lo indicado por la siguiente expresión [2]:

$$s \quad \min(n,m) \quad , \tag{3.21}$$

es decir, la BAM será capaz de almacenar y reconocer tantos vectores como la mínima dimensión de cualquiera de los vectores de los pares almacenados, por ejemplo, si n < m, se almacenarán y reconocerán correctamente s = n vectores. También, Kosko menciona que la BAM puede ser confundida si se asocian entradas similares con salidas diferentes, por lo que se debe cumplir la siguiente condición de continuidad:

$$\frac{1}{n}H(A_i, A_j) \approx \frac{1}{p}H(B_i, B_j) \quad , \tag{3.22}$$

siendo $H(A_i, A_j)$ la distancia de Hamming entre el vector A_i y el vector A_j y lo mismo para $H(B_i, B_j)$.

Una vez presentados los antecedentes teóricos de una BAM, se pueden proponer los vectores que se pretende almacenar en la red que se quiere integrar. Se propone, de manera arbitraria y tratando de ajustarse al área que se impone para la integración de un circuito en un encapsulado de cuarenta terminales (Tiny Chip) como los ofrecidos por MOSIS, una red de seis neuronas de entrada por tres neuronas de salida (6x3). En este caso, n = 6, m = 3 y s = 3. A continuación, se indican tres parejas de vectores cuyas entradas y salidas son ortogonales entre sí, expresadas en forma binaria:

A1: (1 0 0 1 0 0)	B1: (0 1 0)
A2: (0 1 0 0 1 0)	B2: (1 0 0)
A3: (0 0 1 0 0 1)	B3: (0 0 1)

Se mencionó anteriormente que, para tener mejor eficiencia de reconocimiento, es preferible la notación en bipolar, por lo que las parejas anteriores se pueden expresar de esa manera, cambiando los 0's por -1's, teniéndose de esta manera lo siguiente:

A1: (1 -1 -1 1 -1 -1)	B1: (-1 1 -1)
A2: (-1 1 -1 -1 1 -1)	B2: (1 -1 -1)
A3: (-1 -1 1 -1 -1 1)	B3: (-1 -1 1)

Usando las ecuaciones (3.1) y (3.2), se pueden obtener las matrices de correlación M1, M2, M3 y M:

$$M1 = \begin{bmatrix} -1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \end{bmatrix}$$

$$M2 = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \end{bmatrix}$$

$$M3 = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$$

$$M = \begin{bmatrix} -1 & 3 & -1 \\ 3 & -1 & -1 \\ -1 & -1 & 3 \\ -1 & 3 & -1 \\ 3 & -1 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

La transpuesta de M es:

$$M^T = \begin{bmatrix} -1 & 3 & -1 & -1 & 3 & -1 \\ 3 & -1 & -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 & -1 & 3 \end{bmatrix}$$

Con las especificaciones anteriores, la arquitectura de la BAM queda como se muestra en la Fig. 3.2(a) para el diagrama esquemático y 3.2(b) para la configuración usada para el circuito electrónico. Este último se precisará en función de los elementos básicos empleados: *la neurona* y *la sinapsis*.

Analíticamente, se puede comprobar mediante las ecuaciones (3.3) y (3.4), que usando las matrices de correlación M y M^T, se obtienen los vectores almacenados A y B. Cada elemento de estas matrices son las sinapsis y para su implementación en el circuito, se requerirán dos etapas, ya que desde el punto de vista del circuito utilizado para la sinapsis, no existe bidireccionalidad. Los valores de cada elemento (peso) de estas matrices (m_{ij}), tienen sentido cuando el análisis es matemático, pero su equivalencia dentro de la dinámica del circuito electrónico debe ser representada por el voltaje de umbral para que tengan sentido en la funcionalidad del circuito. En este caso, esto indica que cada sinapsis expresada con (-1) tendrá un voltaje de umbral diferente a aquellas que están expresadas por (3), y es donde se aprovecha la característica de los elementos de compuerta flotante para cambiar este parámetro. La determinación del voltaje de umbral que se ha de utilizar, se hace en el siguiente capítulo.

Los elementos anteriores son suficientes para tener una idea de la concepción del circuito básico que conformará a la BAM y a continuación se explicarán con detalle los dispositivos y circuitos que deberán incluirse para el diseño de una red neuronal que almacene tres parejas de patrones.

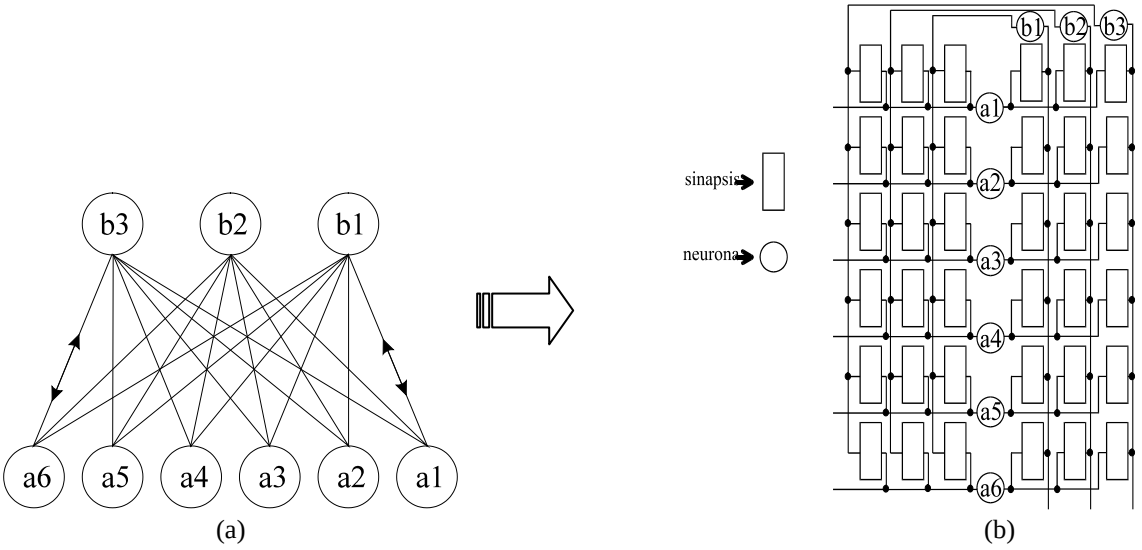


Fig. 3.2. Arquitectura propuesta para la BAM. a) Diagrama esquemático; b) Diagrama a bloques.

3.2. Aplicación de la memoria de compuerta flotante en la RNA-BAM.

En el campo de las redes neuronales artificiales, se están realizando muchos esfuerzos para lograr que lo que se ha desarrollado teóricamente en cuanto a las diferentes arquitecturas y algoritmos, se puedan desarrollar físicamente mediante un circuito. Las limitaciones tecnológicas han frenado el desarrollo de esto último, comparado con el desarrollo computacional, el cual ha sido muy avanzado. El prototipo que se diseña en esta tesis, se puede ver como una aportación en el desarrollo de circuitos neuronales, con los que

se pueda comprender mejor tanto el concepto de redes neuronales artificiales, como el funcionamiento de elementos de memoria analógicos como dispositivos potencialmente adecuados para la implementación física de las redes.

Anteriormente, se mencionó que el peso o interconexión tiene su analogía en un elemento de resistencia variable. En consecuencia, cambiar el peso es equivalente a cambiar la resistencia y si se utiliza la característica de salida de un inversor CMOS con Vth variable, en la parte de transición, como se explicó en el Capítulo 1, se tiene la posibilidad de armar una arquitectura como la que se desea, utilizando al inversor como un elemento de la matriz de la BAM. Las funciones de transferencia que son aplicadas a la suma de las señales provenientes de otras neuronas, se pueden realizar también mediante circuitos CMOS, con lo que se puede tener una BAM básica. El diseño de los elementos que componen a esta red, se explican a continuación.

3.2.1. Diseño de los elementos electrónicos básicos.

La red mostrada en la Fig. 3.2, se puede realizar usando los circuitos mostrados en las Figs. 1.17 y 1.20, que corresponden a la sinapsis y a la neurona, respectivamente. Uno de los criterios de partida, es la alimentación que se aplicará a la red y es conveniente que sea de un valor bajo y estándar para cumplir con requerimientos como bajo consumo de potencia y compatibilidad con alimentaciones convencionales. Se eligió para este diseño, un valor de 5 V con lo que se pueden cumplir las condiciones anteriores y además, de esta manera se pueden manejar patrones bipolares para poder tener la característica de inhibición y excitación.

La base de cómo se programarán las sinapsis, es la matriz M que se encontró en la sección anterior y en el siguiente capítulo se determinará el cambio del voltaje de umbral que equivale proponer a esta matriz en función de este parámetro. Una vez determinado el voltaje de umbral correspondiente, se puede utilizar la tabla 2.2 para elegir el voltaje que se ha de aplicar en los inyectores de la sinapsis. En este caso, la programación de cada una de ellas se hará independientemente, por lo que se requiere un circuito codificador para seleccionar cada punto de la matriz durante la programación y que una vez hecho esto, se puedan habilitar todas para permitir el funcionamiento de la red en su conjunto.

El circuito total se puede dividir en tres etapas: neurona, sinpasis y circuitos periféricos, los cuales se diseñaron en base a los datos y reglas de diseño de Orbit y se explican a continuación. Las especificaciones del diseño se resumen en la Tabla 3.1.

Tabla 3.1. Especificaciones para el diseño del circuito de la BAM.

Fuentes de polarización	V_{DD} = 5 V; V_{SS} = -5 V
Arqutectura	BAM en arreglo 6 x 3
Cantidad de patrones almacenados	3
Tipo de vector de los patrones	Bipolar
Función de transferencia de las neuronas	Tipo escalón
Programación de las sinapsis	Una a la vez (sin algoritmo)

3.2.1.1. Neurona.

La función sigmoidal es una función de activación muy utilizada en redes neuronales y una de sus ventajas es debida a la simplificación del proceso computacional que se deriva de poder tener una relación entre el valor de la función en un cierto punto y el valor de la derivada de la función en el mismo punto, durante el entrenamiento de la red. Un circuito sencillo con el que se logra tener esta función, es mediante dos inversores conectados en serie, como se mostró en la Fig. 1.20(a). La parte de transición de la función (cuando pasa de un valor bajo a un valor alto) representa la ganancia del circuito y las dimensiones de los

transistores empleados en él, determinan la pendiente o ganancia. Como se había mencionado anteriormente, con la idea de simplificar el análisis de la respuesta de la red, se puede aproximar a una función escalón o de alta ganancia procurando tener una sigmoide cuya pendiente en la transición sea muy pronunciada. Las dimensiones se deben restringir para evitar que el tamaño global del circuito se incremente demasiado y en consecuencia se reduzca el área de integración.

Las dimensiones pueden ser obtenidas analíticamente [10], proponiendo la ganancia (alta para el caso de la neurona), pero a partir de una solución no trivial de la expresión de ganancia para el inversor empleado.Una manera mas simple de deducir las dimensiones W y L es usando el programa PSpice [9]; se simularon varias combinaciones de las distintas dimensiones de los transistores canal N y canal P para obtener la respuesta apropiada y finalmente, las dimensiones con las que quedó la etapa de la neurona fueron:

Transistores canal P:	W = 29 μm	L = 10 μm
Transistores canal N:	W = 9 μm	L = 10 μm .

En la Fig. 3.3 se muestran los detalles del circuito utilizado como neurona. Para el caso del circuito prototipo, deberán existir nueve circuitos similares: seis para la capa A y tres para la capa B.

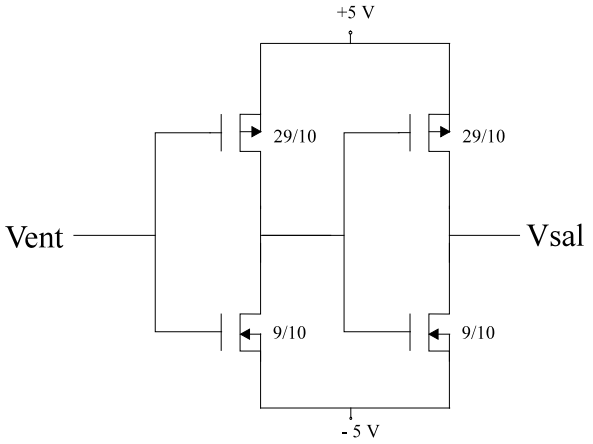


Fig. 3.3. Doble inversor en serie para obtener una sigmoide.

3.2.1.2. Sinapsis.

La misma función de transferencia de un inversor CMOS se puede aprovechar para implementar un circuito cuya característica de salida I-V sea similar a la de una resistencia, pero procurando que se tenga la mayor linealidad posible en el intervalo de voltajes de entrada a la sinapsis, que será de ± 5 V. Esto conlleva a que la pendiente de la función de transferencia no sea tan pronunciada como en el caso anterior, ya que se requiere que la corriente cambie monotónicamente a lo largo de los 10 V de cambio de voltaje. Por lo tanto, las dimensiones son diferentes y es suficiente con un solo inversor.

En este caso, los transistores que componen al inversor, son los que tienen incluidos los inyectores y la compuerta flotante, para poder cambiar el voltaje de umbral (la conductancia, equivalentemente) y cada voltaje de umbral responde con una curva de pendiente distinta, como se verá en el siguiente capítulo, donde se reportan las simulaciones hechas con PSpice. La Fig. 3.4 muestra las dimensiones con las que se realizaron las simulaciones. También se pueden obtener mediante la expresión de ganancia del inversor push-pull, considerando en este caso, una ganancia baja [10].

Las dimensiones que resultaron con mejores resultados, cumpliendo un compromiso entre área y funcionamiento, fueron las siguientes:

Transistor canal P: W = 10 μm L = 9 μm
Transistor canal N: W = 4 μm L = 10 μm .

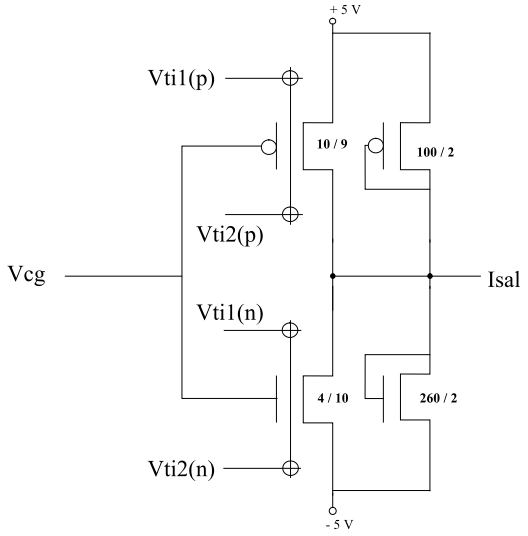


Fig. 3.4. Inversor CMOS para la sinapsis con resistencias de carga para la simulación.

Con el propósito de simular correctamente la característica de salida I-V de la sinapsis, se utilizó un circuito de carga, para mantener el voltaje de salida a un nivel de $(V_{DD} - |V_{SS}|)/2$, mediante un transistor canal P y otro canal N, con las dimensiones y configuración mostradas en la Fig. 3.4. También en este caso, la polarización que se aplica al inversor es de $V_{DD} = 5$ V y $V_{SS} = -5$ V.

El voltaje de entrada Vcg estará dentro del intervalo de ± 5 V y el voltaje que se aplica en los inyectores, denotados por Vti1(n, p) y Vti2(n, p), dependerá del cambio del voltaje de umbral que se desee alcanzar.

Un solo circuito de estos (sin los transistores de carga) ocupará un lugar de la matriz M y M^T y como ya se mencionó que es necesario implementar las dos matrices, deberán incluirse 36 sinapsis CMOS en el circuito.

3.2.1.3. Circuitos periféricos.

Para poder hacer funcional al circuito de la BAM, es necesario complementarlo con algunos circuitos adicionales que permitan hacer la programación de las sinapsis, por un lado, y en general, hacer funcionar globalmente a toda la red. En el caso del circuito prototipo, al no incluirse algún algoritmo de aprendizaje, se programarán las sinapsis una vez calculado el valor de voltaje de umbral correspondiente a cada una de ellas. La estrategia a seguir, será programar dos sinapsis a la vez, aprovechando que las matrices son iguales dado que el elemento de una será igual al elemento transpuesto de la otra. De esta manera, serán necesarios 18 pasos de programación en lugar de 36, si se programaran independientemente cada una de ellas. Esto ayuda a reducir además, el número de terminales utilizadas en el circuito integrado.

Sin embargo, existe un conflicto que se debe resolver para programar ambos transistores de la sinapsis. Según la referencia [10], para tener una respuesta adecuada de la sinapsis integrada, el desplazamiento del voltaje de umbral del transistor canal P deberá ser en sentido contrario al desplazamiento del voltaje de umbral del transistor canal N, es decir, si uno se desplaza hacia voltajes más positivos, el otro deberá ser desplazado hacia voltajes más negativos y viceversa. De la revisión del mecanismo de inyección y extracción de carga que se vió en la sección 2.3.2, se puede ver que para que exista un desplazamiento del voltaje de umbral hacia valores negativos, es preciso aplicar voltajes positivos

tanto en el inyector como en la compuerta y si es hacia valores positivos, los voltajes deberán ser ambos negativos. Con la configuración mostrada en la Fig. 3.4, donde la compuerta de ambos transistores corresponde a un mismo nodo, no sería posible cumplir con la condición reportada en [10], ya que ambos transistores tendrían una misma polaridad aplicada en su compuerta y por lo tanto, el desplazamiento del voltaje de umbral sería en el mismo sentido para ambos. Esto lleva a la necesidad de separar las compuertas para poder polarizar opuestamente ambas compuertas; los inyectores Vti1 de todos los transistores de la red pueden estar conectados a un mismo nodo, así como los inyectores Vti2, con lo que se utilizan únicamente dos terminales para todo el circuito.

Con lo anterior se resuelve parte del problema, ya que al separar las compuertas se puede programar independientemente cada transistor de la sinapsis; una vez terminado el proceso de programación, se pueden conectar ambas compuertas al mismo nodo y obtener la respuesta esperada. Esto último (cortocircuitar las compuertas) deberá ser posible para habilitar el modo de funcionamiento de la BAM. En resumen, debe existir la posibilidad de abordar al circuito en dos fases independientes entre sí: 1) programación de los transistores y 2) funcionamiento del circuito. El arreglo mostrado en la Fig. 3.5, muestra la forma cómo se puede elegir la operación en una fase o en otra, mediante el uso de interruptores o compuertas de transmisión CMOS.

Con esta configuración, al habilitar los interruptores S3 y S4 e inhabilitar S1 y S2, se tendría la fase de *programación*; cuando se habilitan S1 y S2 mientras S3 y S4 se encuentran inhabilitados, el circuito se encuentra en la fase de *funcionamiento*. Las terminales Vcg(n) y Vcg(p) son dos terminales comunes a todas las sinapsis del circuito, por lo que se ocupan únicamente dos terminales; El hecho de que el voltaje llegue a ser aplicado a la compuerta de cada transistor, dependerá de si el interruptor correspondiente esté habilitado o no y la elección de cada punto de la matriz, se puede llevar a cabo mediante un decodificador conectado a las terminales de habilitación/inhabilitación del interruptor CMOS. Este decodificador puede ser construido a base de compuertas NAND, como se muestra en la Fig. 3.6.

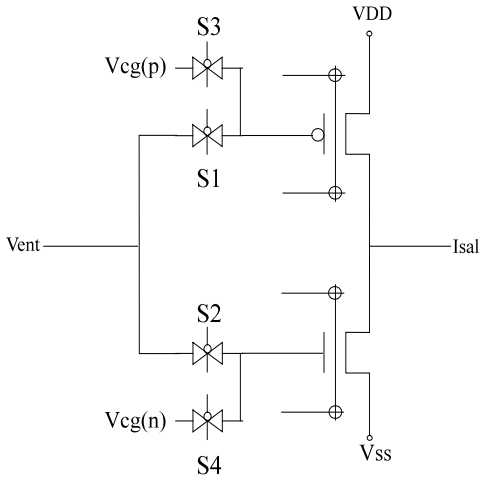


Fig. 3.5. Separación de las compuertas del NMOS y del PMOS de la sinapsis.

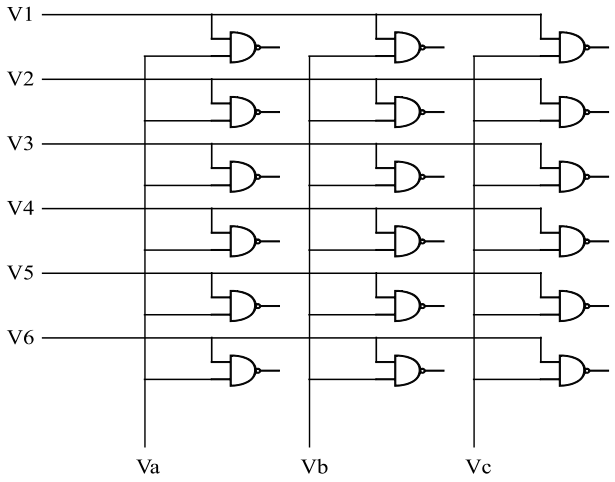


Fig. 3.6. Decodificador para seleccionar independientemente la sinapsis a programar.

Con este arreglo y con las compuertas NAND, se debe procurar que solo aquella compuerta que corresponda a la posición de la sinapsis que se quiere programar, tenga presente en sus dos entradas un uno (1) lógico con el que se obtiene un cero (0) lógico a la salida, y será este estado (junto con su complemento) el que habilitará la compuerta de transmisión correspondiente. En la tabla 3.2 se indica la notación que se le da a cada posición de la matriz y en la tabla 3.3, se indican los valores de voltaje que deberán tener cada una de las nueve terminales del decodificador para seleccionar exclusivamente a una posición.

Una vez programadas y listas las sinapsis para funcionar como una BAM, basta con establecer simultáneamente un voltaje de + 5 V en las terminales Va, Vb y Vc, para inhabilitar todos los interruptores S3 y S4. La habilitación de los interruptores S1 y S2 se puede hacer con una sola compuerta NAND con sus dos entradas conectadas a un mismo nodo y cuyas salidas se conectan a estos últimos interruptores. Esta compuerta deberá tener sus entradas en un cero lógico cuando se está en la fase de *programación* y un uno lógico cuando se esté en la fase de *funcionamiento* (se considera a + 5 V como uno lógico y – 5 V como cero lógico).

El decodificador se diseñó en base a la biblioteca disponible en el paquete L-Edit de Tanner, donde se usó la compuerta NAND2C, formada por 6 transistores, usándose en total, 108 transistores debido a los 18 elementos del decodificador. Para la compuerta habilitadora de la fase de funcionamiento, se usó la biblioteca de la compuerta NAND2CH, también compuesta de 6 transistores. Cada sinapsis consta de 2 transistores, lo que tomando en cuenta las dos matrices de 18 elementos cada una, hace un total de 72 transistores; cada compuerta de transmisión se forma con dos transistores, multiplicada por 4 interruptores usados por sinapsis, hace un total de 288 transistores para los interruptores. Finalmente, cada neurona está formada por 4 transistores y siendo 9 neuronas, hace un total de 36 transistores.

Haciendo un recuento de los transistores mencionados anteriormente, se puede ver la cantidad de transistores utilizados en total, dentro del circuito:

Decodificador:	108 transistores.
Sinapsis	72 transistores.
Interruptores	288 transistores.
Habilitador	4 transistores.
Neuronas	36 transistores.
Total	**508 Transistores.**

También se puede determinar la cantidad de terminales que se utilizarán en la BAM:

Patrón de entrada 6 terminales.

Patrón de salida	3 terminales.
Decodificador	9 terminales.
Inyectores	2 terminales.
V_{DD}	1 terminal.
V_{SS}	1 terminal.
Habilitador	1 terminal.
Total	**23 terminales.**

El encapsulado que se eligió para la integración de este circuito, consta de 40 terminales (Tiny Chip), por lo que se pueden usar otras terminales para conectar circuitos de prueba, si el espacio del dado lo permite.

El diagrama esquemático de la red neuronal queda como se muestra en la Fig. 3.7. No se indican los circuitos periféricos por simplicidad del diagrama, las neuronas se ilustran con el símbolo de un amplificador diferencial y solo se indican las configuraciones de las sinapsis de la primera fila de cada matriz.

Tabla 3.2.
Notación de la posición de la sinapsis.

11	21	31
12	22	32
13	23	33
14	24	34
15	25	35
16	26	36

Tabla 3.3. Voltajes en las entradas del decodificador para seleccionar solo una sinapsis.

Elemento	**Va**	**Vb**	**Vc**	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**
11	**5**	-5	-5	**5**	-5	-5	-5	-5	-5
21	-5	**5**	-5	**5**	-5	-5	-5	-5	-5
31	-5	-5	**5**	**5**	-5	-5	-5	-5	-5
12	**5**	-5	-5	-5	**5**	-5	-5	-5	-5
22	-5	**5**	-5	-5	**5**	-5	-5	-5	-5
32	-5	-5	**5**	-5	**5**	-5	-5	-5	-5
13	**5**	-5	-5	-5	-5	**5**	-5	-5	-5
23	-5	**5**	-5	-5	-5	**5**	-5	-5	-5
33	-5	-5	**5**	-5	-5	**5**	-5	-5	-5
14	**5**	-5	-5	-5	-5	-5	**5**	-5	-5
24	-5	**5**	-5	-5	-5	-5	**5**	-5	-5
34	-5	-5	**5**	-5	-5	-5	**5**	-5	-5
15	**5**	-5	-5	-5	-5	-5	-5	**5**	-5
25	-5	**5**	-5	-5	-5	-5	-5	**5**	-5
35	-5	-5	**5**	-5	-5	-5	-5	**5**	-5
16	**5**	-5	-5	-5	-5	-5	-5	-5	**5**
26	-5	**5**	-5	-5	-5	-5	-5	-5	**5**
36	-5	-5	**5**	-5	-5	-5	-5	-5	**5**

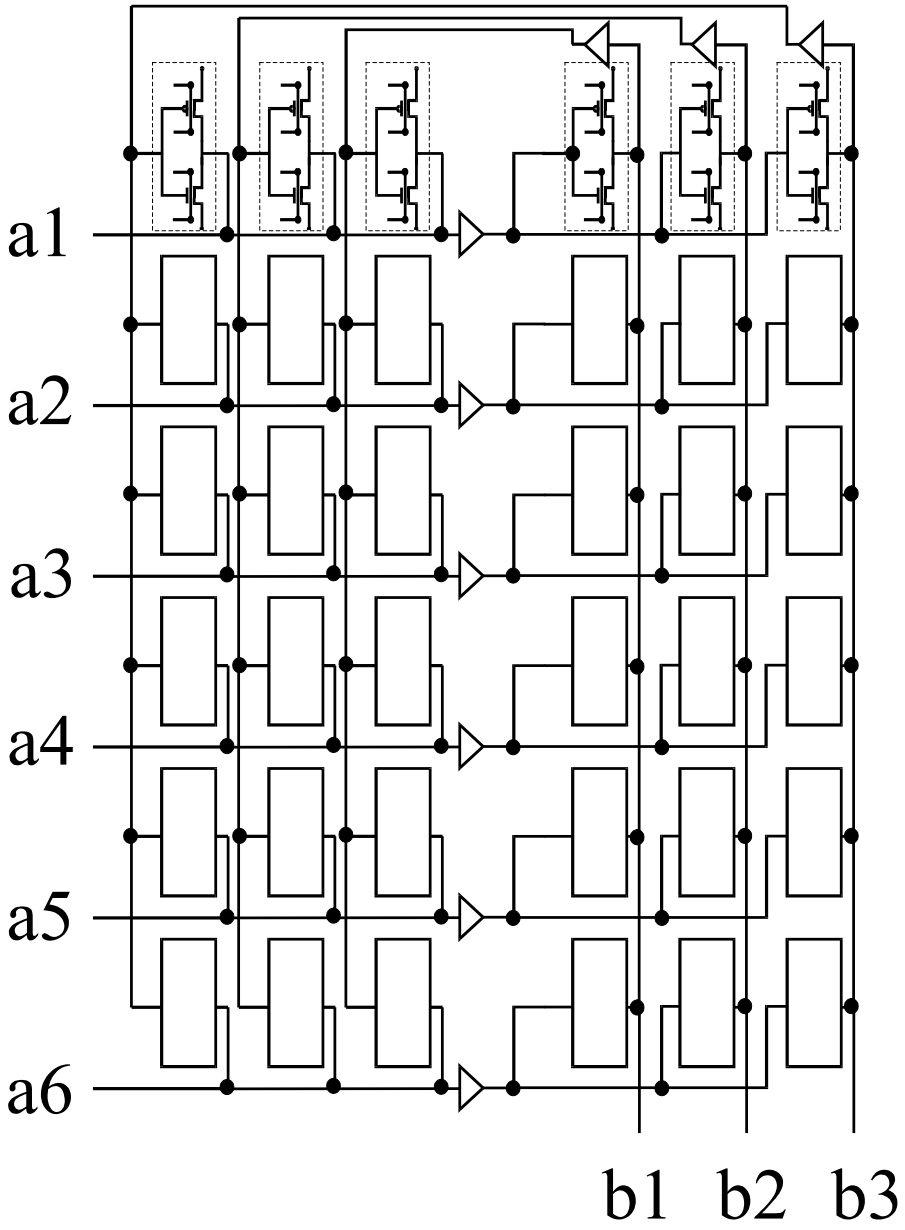


Fig. 3.7. Diagrama esquemático de la BAM de dimensión 6x3.

3.3. Sumario.

En este capítulo se establecieron los circuitos apropiados para implementar los elementos básicos de una red neuronal, conectada en configuración BAM. Tanto para la neurona como para la sinapsis, se puede utilizar un inversor CMOS, pero con diferentes dimensiones cada uno. La neurona consta de dos inversores en serie, además que sus dimensiones fueron ajustadas para tener una función sigmoidal de alta ganancia para aproximarse a la función escalón y simplificar el análisis. La sinapsis consta de un solo inversor con dimensiones más pequeñas para tener menor ganancia en la función de transferencia y poder tener una aproximación lineal de la característica I-V de una resistencia. Ambos circuitos son alimentados con ± 5 V con el propósito de considerar patrones bipolares.

Para lograr una funcionalidad práctica del circuito, se requirió agregar circuitos periféricos con el propósito de poder programar independientemente cada sinapsis de la red, para posteriormente hacer funcionar al circuito completo en las condiciones adecuadas. Esto es posible con un decodificador que habilita a las compuertas de transmisión del NMOS y del PMOS de una sola sinapsis, para permitir el paso de voltaje hacia las compuertas de control de cada transistor.

El número total de transistores MOS utilizados en la BAM prototipo, propuesta en esta tesis, es de 508 transistores y la cantidad de terminales usadas es de 23. El encapsulado elegido para la integración del circuito consta de 40 terminales y la diferencia puede ser utilizada para conectar circuitos de prueba, si el espacio de silicio para este encapsulado lo permite.

Referencias.

1.- B. Kosko, "Adaptive bidirectional associative memories", *Applied Optics*, Vol. 26, No. 23, Dec. 1987, pp. 4947-4960.

2.- B. Kosko, "Bidirectional associative memories", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 18, No. 1, Jan/Feb 1988, pp. 49-60.

3.- B. Bavarian, "Introduction to neural networks for intelligent control", *IEEE Control Systems Magazine*, Apr. 1988, pp. 3-7.

4.- G. Dunning, E. Marom, Y. Owechko and B. Soffer, "Optical holographic associative memory using a phase conjugate resonator", *Proc. of the SPIE*, Vol. 625, 1986, pp. 205-213.

5.- Y. F. Wang, J. B. Cruz and H. Mulligan, "Two coding strategies for bidirectional associative memory", *IEEE Trans. Neural Networks*, Vol. 1, No. 1, 1990, pp. 81-92.

6.- W. Wang and D. Lee, "A modified bidirectional decoding strategy based on the BAM structure", *IEEE Trans. Neural Networks*, Vol. 4, No. 4, Jul. 1993, pp. 710-717.

7.- B. Zhang, B. Xu and C. Kwong, "Performance analysis of the bidirectional associative memory and an improved model from the matched-filtering viewpoint", *IEEE Trans. Neural Nerworks*, Vol. 4, No. 5, Sept. 1993, pp. 864-881.

8.- T. Wang, X. Zhuang and X. Xing, "Designing bidirectional associative memories with optimal stability", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 24, No. 5, May 1994, pp. 778-790.

9.- Programa PSpice. Producto de MicroSim Corporation. Versión 5.1. *The Design Center. Analysis-reference manual*. Irvine, California, January 1992.

10.- S. Kim, Y. Shin, N. C. R. Bogineni and R. Shridhar, "A programmable analog CMOS synapse for neural networks", *Analog Integrated Circuits and Signal Processing*, Vol. 2, 1992, pp. 345-352.

11.- P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*, Saunders College Publishing, 1987.

Capítulo 4.

Simulación circuital de la BAM.

Cuando ya se tienen definidos los circuitos básicos y la estructura necesarios para configurar una red neuronal tipo BAM, se puede realizar el diseño del circuito prototipo, a partir del cual se pueden hacer las simulaciones pertinentes para comprobar inicialmente el comportamiento y funcionamiento del circuito, antes de pasar al diseño topológico. Estas simulaciones se pueden efectuar mediante la utilización del programa PSpice, el cual permite cambiar tanto los parámetros de los dispositivos, así como los estímulos, de tal forma que se abarcan diferentes condiciones de funcionamiento del circuito.

Las simulaciones contempladas abarcan a los circuitos básicos de manera independiente, con el objetivo de encontrar las características de funcionamiento con las que, una vez integradas a la red en su conjunto, sea posible obtener el funcionamiento esperado, así como a la BAM completa, para determinar su comportamiento a la presentación de los patrones elegidos en el capítulo anterior junto con algunas variantes que permitan observar la eficiencia al reconocimiento de patrones con errores. Esto se refiere a probar a la red cuando se le introducen vectores con diferentes distancias de Hamming respecto al vector almacenado.

La simulación también aporta un apoyo en la definición del voltaje de umbral que se debe utilizar en las sinapsis, ya que la respuesta de la red es dependiente de este parámetro. Mediante la simulación se pueden conocer los voltajes de umbral tanto para reconocimiento como para bidireccionalidad. El procedimiento de asociar la matriz de correlación, obtenida mediante el método de Kosko, con una matriz a base de voltajes de umbral, no se ha reportado en la literatura y es una aportación de la presente tesis para el diseño de una BAM que almacene patrones. A continuación, se presentan las simulaciones que se realizaron en el estudio del diseño de la BAM.

4.1. Células básicas.

Como se explicó en el capítulo anterior, la BAM se puede formar en base a dos elementos básicos: la sinapsis, que se puede construir mediante un inversor CMOS, con elementos de compuerta flotante que permitan el cambio de la conductancia mediante la programación del voltaje de umbral, y la neurona, que se puede realizar usando dos inversores CMOS en serie, para entregar una función de transferencia tipo sigmoidal.

Estas simulaciones de la sinapsis y de la neurona se realizaron de manera independiente, procurando ajustar las respuestas que cumplieran con un compromiso de funcionalidad y área, para después conjuntar los elementos en la configuración de la red deseada.

4.1.1. Neurona.

Una característica importante de la neurona es su función de transferencia, la cual permite acotar la suma de las señales presentes en su entrada, dentro de un intervalo binario o bipolar, según se desee. Ya se había justificado con anterioridad la conveniencia de utilizar vectores bipolares y de utilizar una sigmoide que tendiera a una función de transferencia tipo escalón. Las dimensiones para los transistores MOS con las que se obtuvieron mejores resultados dentro del compromiso requerido fueron (ver Fig. 3.3.):

W = 29 μm	L = 10 μm	para canal P
W = 9 μm	L = 10 μm	para canal N.

La polarización del circuito es de V_{DD} = 5 V y V_{SS} = -5 V.

El listado utilizado para la simulación del inversor usado como neurona se presenta a continuación:

Listado para la simulación de la neurona

```
neurona
m1 3 2 1 1 mp w = 29u   l = 10u
m2 3 2 4 4 mn w = 9u    l = 10u
m3 5 3 1 1 mp w = 29u   l = 10u
m4 5 3 4 4 mn w = 9u    l = 10u
vdd 1 0 5
vss 4 0 -5
vin 2 0 5
.lib C:\alfredo\rlevel2.lib
.dc vin -5 5 .2
.print dc V(5)
.probe
.end
```

El resultado de la simulación usando el listado anterior se presenta en la Fig. 4.1.

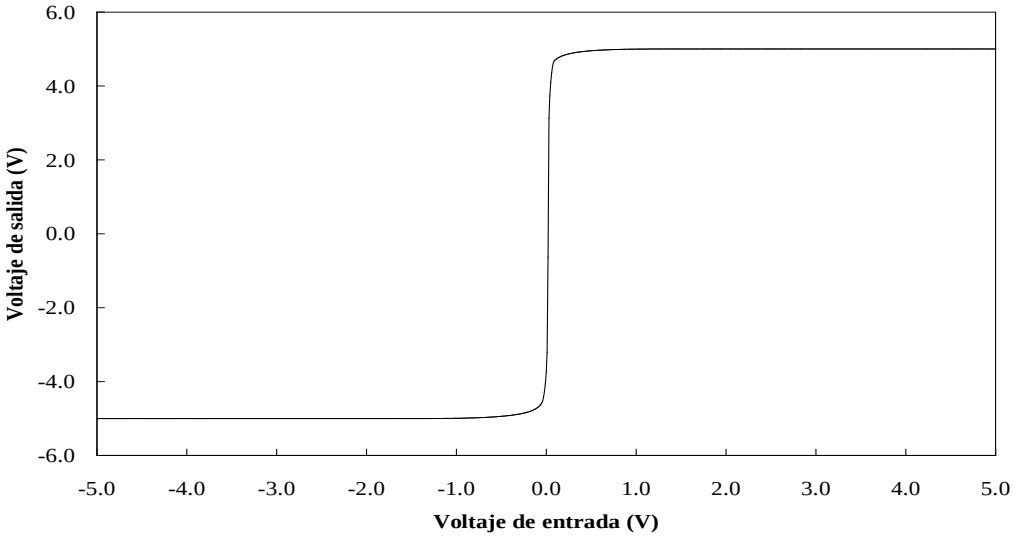


Fig. 4.1. Simulación de la neurona con dos inversores CMOS en serie.

Como se puede ver en la figura anterior, la ganancia del inversor permite acercarse de manera muy aproximada a la función escalón, conservando dimensiones relativamente pequeñas de los transistores, que pueden ser fabricados con la tecnología de Orbit.

En el listado se puede observar que la terminal de substrato de los transistores se conecta al mismo nodo que la fuente. Por esto, se debe tener cuidado de que cuando se haga el diseño topográfico del

circuito, los transistores canal P deben tener su fuente conectada al pozo y los transistores canal N lo deberán tener conectado al substrato y así tener las mismas condiciones de funcionamiento y de simulación.

4.1.2. Sinapsis.

El siguiente circuito básico de interés para simular, es la sinapsis implementada en base a un solo inversor CMOS, pero con inyectores de carga hacia la compuerta flotante. Hay que recordar que la función de la sinapsis, es proporcionar un elemento de resistencia variable para tener la capacidad de aprendizaje. Esta condición implica diseñar un inversor con baja ganancia, de tal forma que la transición de la función de transferencia sea más gradual y menos abrupta, comparada con el inversor usado como neurona.

Un aspecto importante que se tiene que considerar en la simulación de la respuesta de la sinapsis, es cómo cambiar el voltaje de umbral de los transistores, ya que los modelos contenidos en la biblioteca de PSpice sólo consideran al transistor MOS normal sin compuerta flotante, con ninguna posibilidad de simular la programación. Sin embargo, es posible utilizar el nivel 2 del modelo del transistor MOS, donde dentro de los parámetros considerados se encuentra el voltaje de umbral dependiente de la polarización de substrato, VTO. Cambiando este parámetro a conveniencia (con $V_{SB} = 0$), se puede hacer la simulación de un inversor con dispositivos con compuerta flotante a diferentes voltajes de umbral, dentro de un intervalo práctico.

Partiendo de [1] es posible realizar incrementos de voltajes de umbral de la misma magnitud pero en sentido inverso para el transistor canal P, con respecto del transistor canal N, es decir, simétricos, con lo que la respuesta I-V de la sinapsis es a su vez más simétrica. Cuando no se hace de esta manera, no existe una buena aproximación a la linealidad deseada en el comportamiento de la sinapsis. Siguiendo con los lineamientos necesarios para proceder con la simulación de este elemento, se ha reportado [2] que el cambio del voltaje de umbral en un transistor con compuerta flotante puede alcanzar un intervalo de variación del voltaje de umbral de $\Delta Vth = \pm 10$ V sin degradación del dispositivo.

El problema que se presenta con los modelos de PSpice disponibles para los transistores MOS, en cuanto a que no se consideran compuertas flotantes, afortunadamente se puede resolver tomando en cuenta el factor de acoplamiento mediante un voltaje multiplicado por una ganancia, que para este caso debe ser menor a uno, donde el voltaje resultante será el aplicado a la compuerta flotante, en otras palabras, se aplica un voltaje en el nodo de la compuerta de control y este voltaje multiplicado por el factor de acoplamiento, Kcg (descrito en la ecuación 2.38), será el aplicado al nodo de la compuerta flotante. Debido a esto, es necesario calcular de antemano este factor para aplicarlo en ambos transistores.

De los datos tecnológicos proporcionados por la fabrica elegida para la realización del diseño, que en nuestro caso es Orbit, se pueden determinar los valores de las capacitancias y encontrar finalmente el factor de acoplamiento teórico de los transistores NMOS y PMOS. Los parámetros de importancia para esto son el espesor de óxido entre los dos polisilicios y el espesor de óxido de compuerta, reportados a continuación:

Espesor del óxido entre Poly1 y Poly2: 70 nm ± 8 nm.
Espesor del óxido de compuerta: 40 nm ± 3 nm.

Con estos valores, se puede calcular la capacitancia por unidad de área para el óxido entre los dos polisilicios, de la siguiente manera:

$$C_{pp} = \frac{\varepsilon_0 \varepsilon_{ox}}{t_{ox}} \quad , \tag{4.1}$$

$$C_{pp} = \frac{8.85x10^{-14}\left(\frac{F}{cm}\right)3.9}{700x10^{-8}(cm)} = 4.93x10^{-8}\frac{F}{cm^2} \quad .$$

Con el valor anterior, se puede proponer el área del capacitor de acoplamiento para conocer la capacitancia. Con el objeto de asegurar tentativamente un factor de acoplamiento cercano a uno (según se procura en otro tipo de estructuras), se propone un capacitor cuadrado de 36 μm por lado. Por lo tanto, el área a considerar será de:

$$A = 1296\ \mu m^2 \quad .$$

Tomando esta área, la capacitancia resultante del capacitor de acoplamiento puede ser calculada como sigue:

$$C_{pp} = \frac{4.93x10^{-16}(F)x1296(\mu m^2)}{1(\mu m^2)} \quad ,$$

$$C_{pp} = 639\ fF \quad ,$$

por lo tanto, la capacitancia Cpp mostrada en la Fig. 2.19 del Capítulo 2, será de 639 x 10^{-15} F. Tanto el transistor NMOS como el PMOS tendrán su propio capacitor de acoplamiento de las mismas dimensiones pero el cálculo del factor de acoplamiento se verá afectado por las diferentes áreas en la región de compuerta.

Esta región forma la capacitancia designada como Cox en la Fig. 2.19 y recordando las dimensiones encontradas en el Capítulo 3 para los transistores MOS:

NMOS: W x L = 4 μm x 10 μm = 40 μm^2
PMOS: W x L = 10 μm x 9 μm = 90 μm^2

se pueden conocer las capacitancias Cox para cada uno de ellos, como a continuación se muestra:

$$C_{ox} = \frac{8.85x10^{-14}\left(\frac{F}{cm}\right)x3.9}{400x10^{-8}(cm)} \quad ,$$

$$C_{ox} = 8.63x10^{-8}\frac{F}{cm^2} = 8.63x10^{-16}\frac{F}{\mu m^2} \quad .$$

Multiplicando este resultado por el área del NMOS, se obtiene la capacitancia correspondiente:

$$C_{ox}(N) = 8.63x10^{-16}\left(\frac{F}{\mu m^2}\right)x40\mu m^2 \quad ,$$

$$C_{ox}(N) = 34.5\ fF \quad .$$

Lo mismo se puede hacer para el PMOS, obteniéndose lo siguiente:

$$C_{ox}(P) = 8.63x10^{-16}\left(\frac{F}{\mu m^2}\right)x90\mu m^2$$

$$C_{ox}(P) = 77.67 \text{ fF} \quad .$$

Cuando se considera al transistor canal N, la capacitancia total necesaria para aplicar la ecuación 2.38 está formada por la suma de Cpp y Cox(N) y cuando se considera al transistor canal P, esta se forma por la suma de Cpp y Cox(P). Entonces, a partir de la ecuación 2.38 se tiene que:

$$Ctot(N) = 639x10^{-15} + 34.5x10^{-15} = 673.5x10^{-15}\,F \quad ,$$

$$Kcg(N) = \frac{639x10^{-15}}{673.5x10^{-15}} = 0.94 \quad ,$$

$$Ctot(P) = 639x10^{-15} + 77.67x10^{-15} = 716.6x10^{-15}\,F \quad ,$$

$$Kcg(P) = \frac{639x10^{-15}}{716.6x10^{-15}} = 0.89 \quad .$$

Siendo estos los valores que se deberán tomar en cuenta durante la simulación del circuito sináptico, mediante una fuente de voltaje controlada por voltaje, como se mencionó anteriormente. Estos valores se deben considerar por el momento como un punto de partida, ya que existen otras capacidades parásitas que se deben incluir en la suma de las capacidades (C_{tot}), con lo cual se reduce el factor de acoplamiento.

Tomando en cuenta todas las consideraciones anteriores, la simulación se puede hacer con el listado de PSpice que se muestra mas adelante. En dicho listado, se establece a M1 como el transistor PMOS con una relación W/L= 10 μm/9 μm y a M2 como el transistor NMOS, con W/L= 4 μm/10 μm. La polarización del circuito inversor, tipo push-pull, se eligió como V_{DD}= 5 V y V_{SS}= -5 V. A continuación de la declaración de las fuentes de polarización, se establece la fuente de control para el voltaje de entrada al inversor, denominada VG y que es la que controla las fuentes de voltaje controladas por voltaje EVGP y EVGN, siendo estas últimas, las que aplican el voltaje a la compuerta flotante de cada transistor. Utilizando el nivel 2 del modelo MOS, se puede cambiar el parámetro VTO (que considera la carga total simulada en el TMOS), para obtener la respuesta del inversor o sinapsis a diferentes voltajes de umbral. La Fig. 4.2 muestra el circuito considerado para la simulación de la sinapsis, donde el circuito de la Fig. 4.2(a) corresponde al circuito equivalente reflejado en el listado de Pspice y al modelo propuesto para resolver el problema de la carencia de modelos compatibles con PSpice.

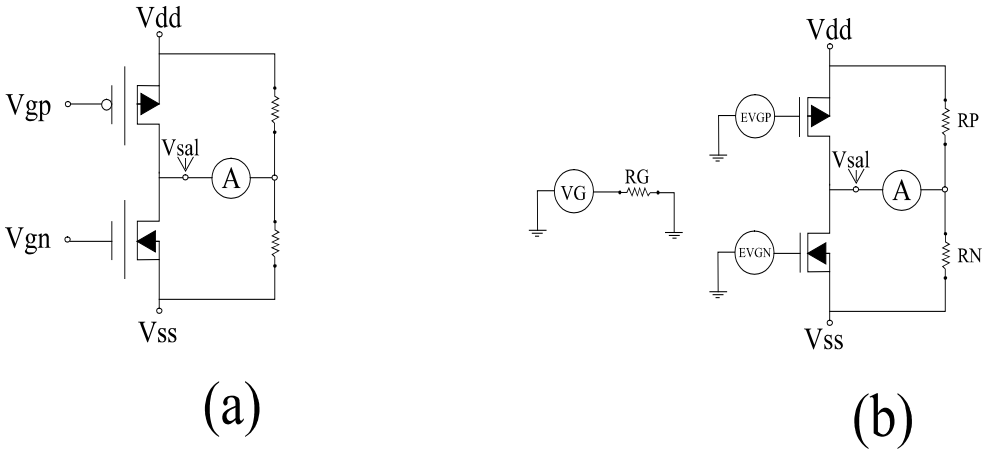


Fig. 4.2. a) Inversor con transistores de compuerta flotante usado como sinapsis; b) circuito equivalente usado para la simulación.

Listado para la simulación de la sinapsis

```
INVERSOR CON VTO'S VARIABLES
M1 3 2  1 1 MP1 W=10U L=9U    ;TRANSISTOR PMOS DE COMPUERTA FLOTANTE
M2 3 2G 4 4 MN1 W=4U  L=10U   ;TRANSISTOR NMOS DE COMPUERTA FLOTANTE
VDD 1 0  5              ;POLARIZACION POSITIVA
VSS 4 0 -5              ;POLARIZACION NEGATIVA
VG VG 0  5              ;VOLTAJE APLICADO A LA COMPUERTA DE CONTROL
RG VG 0 1k              ;RESISTENCIA ASOCIADA A VG
EVGP  2 0 VG 0 0.89     ;FUENTE DE VOLTAJE CONTROLADA POR VOLTAJE (PMOS)
EVGN 2G 0 VG 0 0.94     ;FUENTE DE VOLTAJE CONTROLADA POR VOLTAJE (NMOS)
RP 1 5 1k               ;RESISTENCIA DE CARGA
RN 4 5 1k               ;RESISTENCIA DE CARGA
V0 3 5 0                ;AMPERIMETRO PARA MEDIR LA CORRIENTE DE SALIDA
*NIVEL=2
*
.MODEL MN1 NMOS Level=2.0 UO=608.3 VTO=825.3E-3 NSS=0.000 NFS=0.105E+12
+ TPG=+1.000 TOX=40.0E-9 NSUB=7.755E+15 UCRIT=50E+3 UEXP=78.26E-3
+ UTRA=0.000 VMAX=49.89E+3 RSH=50.15 XJ=450.0E-9 LD=112.1E-9 DELTA=3.714
+ PB=0.44 JS=10.0E-6 NEFF=3.358 WD=46.34E-9 CJ=323.1E-6 MJ=461.5E-3
+ CJSW=929.9E-12 MJSW=268.3E-3 CGSO=96.77E-12 CGDO=96.77E-12
+ CGBO=40.00E-12 FC=500.0E-3 XQC=1.000
*
.MODEL MP1 PMOS Level=2.0 UO=205.1 VTO=-703.0E-3 NSS=0.000 NFS=.010E+12
+ TPG=-1.000 TOX=40.0E-9 NSUB=1.486E+16 UCRIT=70E+3 UEXP=184.2E-3
+ UTRA=0.000 VMAX=40.76E+3 RSH=69.46 XJ=450E-9 LD=230.5E-9 DELTA=1.843
+ PB=0.958 JS=10.0E-6 NEFF=0.688 WD=117.6E-9 CJ=804.9E-6 MJ=525.0E-3
+ CJSW=749.1E-12 MJSW=495.4E-3 CGSO=199.0E-12 CGDO=199.0E-12
+ CGBO=101.5E-12 FC=500.0E-3 XQC=1.000
*
.DC VG 5 -5 -.1   ;VARIACION DEL VOLTAJE DE ENTRADA AL INVERSOR
.PRINT DC I(V0)   ;IMPRIME LA CORRIENTE DE SALIDA DEL INVERSOR CON CARGA
.PRINT DC V(3)    ;IMPRIME EL VOLTAJE DE SALIDA DEL INVERSOR
.PROBE
.END
```

El valor dado a XQC, correspondiente a la fracción de carga de canal atribuido al drenador, corresponde al análisis del modelo en una dimensión. Para el análisis bidimensional y dispositivos de longitud de canal comparabale con el espesor de la región de deserción en drenador, existe una dependencia del voltaje de umbral con la carga adicional en esta zona desértica. Se modifica el factor de cuerpo λ para el cálculo del voltaje de umbral y en consecuencia, el valor de XQC debe ser menor a 1. Sin embargo, para los parámetros establecidos en los modelos del listado anterior, aún considerando el análisis dimensional, la variación de λ es despreciable, por lo que es válido en este caso el valor unitario para XQC.

De la figura anterior, se tiene que las resistencias RP y RN son resistencias de carga para forzar al voltaje de salida del inversor a un voltaje igual a:

$$V_{sal} = \frac{V_{DD} + V_{SS}}{2} \quad , \tag{4.2}$$

además, estas resistencias permiten el flujo de corriente a la salida de la sinapsis, de tal forma que se pueda obtener la característica I-V de la celda cuando se modifican los voltajes de umbral en los transistores de compuerta flotante. El resultado de la simulación de la sinapsis, usando el listado anterior y considerando cambios del voltaje de umbral según la tabla 4.1, se muestra en la Fig. 4.3, donde se observa una respuesta

no lineal, pero aún así, apropiada para el funcionamiento del circuito, ya que presenta cambio de la resistencia en base a los distintos voltajes de umbral empleados, la cual es la característica deseada para la sinapsis.

Tabla 4.1.

ΔVth(NMOS)	ΔVth(PMOS)
0 V	0 V
+ 1 V	- 1 V
+ 2 V	- 2 V
+ 3 V	- 3 V
+ 4 V	- 4 V
+ 5 V	- 5 V
+ 6 V	- 6 V

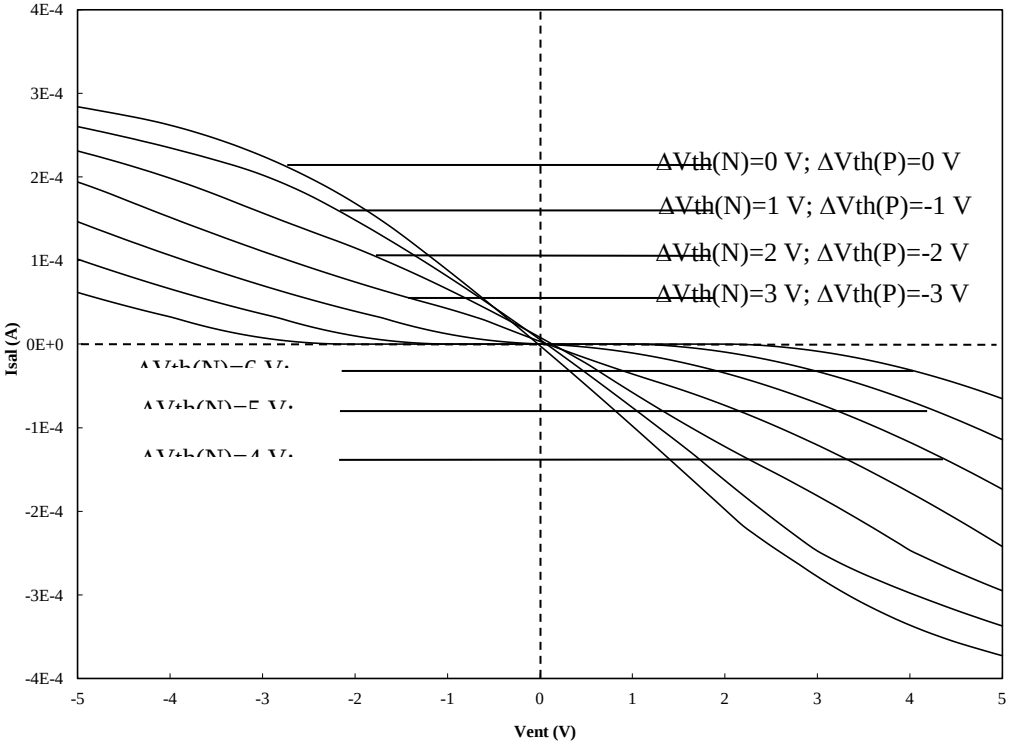


Fig. 4.3. Respuesta I-V simulada de la sinapsis diseñada.

Aproximando las pendientes de las curvas de la Fig. 4.3, se puede tener una idea cercana de las resistencias que se cubren con la sinapsis diseñada, siendo en este caso un intervalo que va desde 8.5 kΩ hasta 43 kΩ aproximadamente.

Con objeto de comprender el comportamiento de los dispositivos de compuerta flotante cuando se cambia el voltaje de umbral, es oportuno explicar el mecanismo que se sigue para lograr dicho objetivo y el tipo de carga que causa tal variación.

Como se mencionó anteriormente, se recomienda que la variación del voltaje de umbral en los transistores del inversor sea simétrica y en sentido inverso, es decir, cuando el sentido de la variación en uno de los transistores sea hacia valores positivos, el otro transistor deberá ser cambiado hacia valores

negativos, como se observa en la Fig. 4.3. En la Fig. 4.4 se ilustra la característica Id-Vg de los transistores NMOS y PMOS de donde se puede extrapolar el valor del voltaje de umbral y el sentido en el cual se considera que estos voltajes aumentan o disminuyen, según sea la polaridad de la carga que se inyecte en la compuerta flotante. Esto es, cuando la carga inyectada en la compuerta flotante sea negativa (Qfg < 0) el voltaje de umbral del NMOS y el del PMOS aumenta; cuando la carga inyectada sea positiva (Qfg > 0), el voltaje de umbral del NMOS y el del PMOS disminuye.

Esto se puede ver más claro en la Fig. 4.5, donde se ilustran ambos tipos de transistores con el canal formado debido a la compensación de carga en el canal y en la compuerta. Se puede ver, en el caso del NMOS, que si se aumenta la carga positiva en la compuerta flotante, se atrae más carga negativa sobre el canal y por lo tanto, se requerirá menos voltaje positivo en la compuerta de control para inducir el canal, lo que es equivalente a disminuir el voltaje de umbral. Lo contrario sucede si se disminuye la carga positiva en la compuerta flotante. El caso del PMOS es similar, ya que cuando se aumenta la carga negativa en la compuerta flotante, se requerirá menos voltaje negativo en la compuerta de control para inducir el canal y por lo tanto, aumenta el voltaje de umbral (se recorre hacia valores positivos de Vth).

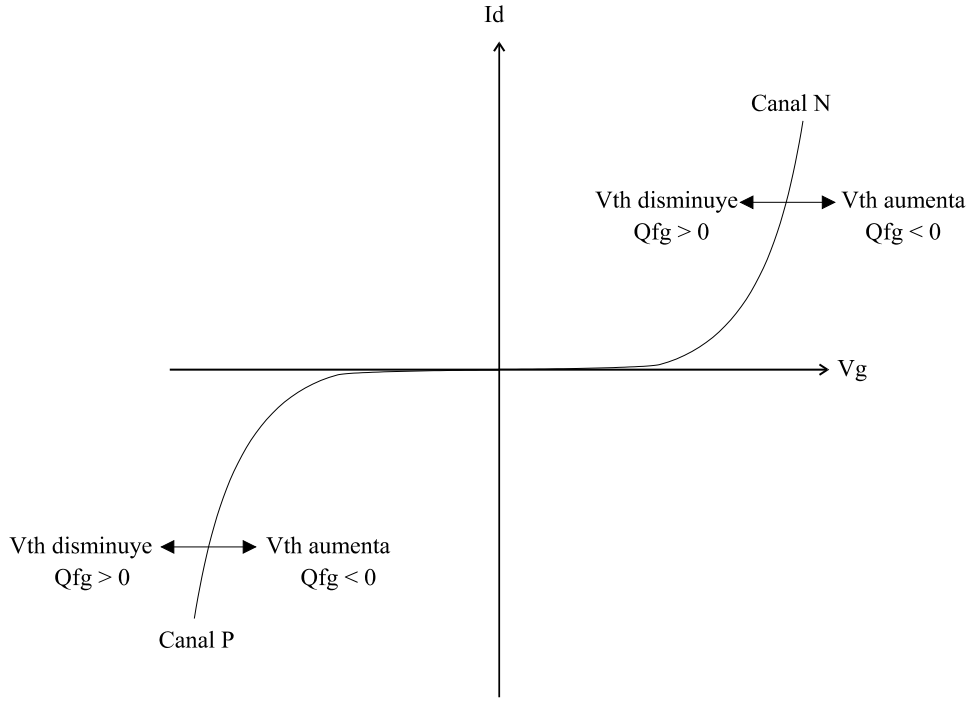


Fig. 4.4. Cambio del voltaje de umbral de transistores PMOS y NMOS según la polaridad de la carga inyectada en la compuerta flotante.

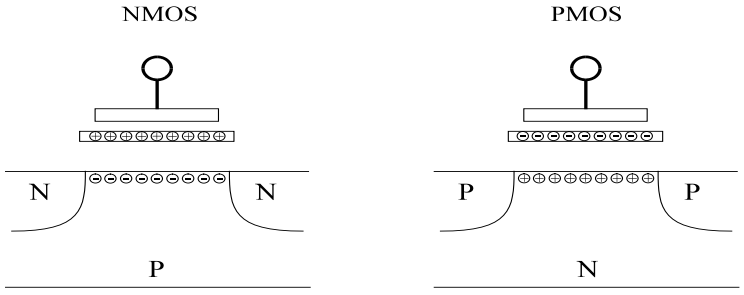


Fig. 4.5. Formación del canal en los transistores, según la carga presente en la compuerta flotante.

Con el análisis anterior se justifica la afirmación que se hizo en la sección 3.2.1.3 del capítulo anterior, en el sentido de que se requieren incluir interruptores MOS en las compuertas de los transistores del inversor para seleccionar y programar independientemente a cada uno. Como ilustración de esto, se puede tomar la curva $\Delta Vth(N) = -4$ V, $\Delta Vth(P) = +4$ V, de la Fig. 4.3, donde para lograr esta respuesta, se deberá inyectar carga positiva en la compuerta flotante del NMOS y carga negativa en la compuerta flotante del PMOS, según la Fig. 4.4. Una vez definido el cambio deseado de voltaje de umbral para cada transistor, se requiere conocer tanto las magnitudes de los voltajes que se deberán aplicar a los inyectores y a la compuerta de control, como los anchos del pulso, en base al mecanismo de programación y borrado, que en este caso es el de tunelamiento Fowler-Nordheim.

4.2. Programación y borrado del dispositivo de compuerta flotante.

Ya se mencionó que la inyección de carga sobre la compuerta flotante provoca un cambio del voltaje de umbral en el dispositivo MOS. Cuando se quiere hacer esto de manera controlada, es preciso aplicar los voltajes adecuados a través del óxido de tunelamiento, de tal manera que provoquen la inyección de la carga necesaria. El cambio de voltaje de umbral en función de la carga, se encuentra con la siguiente ecuación:

$$\Delta Vth = Vth_{final} - Vth_{inicial} = -\frac{Q_{fg}}{C_{pp}} \quad , \tag{4.3}$$

donde Q_{fg} es la carga inyectada y C_{pp} es la capacidad que existe entre la compuerta de control y la compuerta flotante, es decir, entre los dos polisilicios (Poly1 y Poly2). Por otro lado, la cantidad de carga dependerá de la corriente de tunelamiento, que a su vez es función de la diferencia de potencial aplicada entre las terminales del inyector, como se establece en la siguiente ecuación:

$$J_{tun} = \alpha E_{tun}^{\,2} \exp\left(-\beta / E_{tun}\right) \quad , \tag{4.4}$$

$$E_{tun} = \frac{V_{tun}}{X_{tun}} \quad , \tag{4.5}$$

donde α y β son constantes características del fenómeno de tunelamiento FN, V_{tun} es la diferencia de potencial a través del óxido de tunelamiento y X_{tun} es el espesor de este óxido. Dado que Vtun es el voltaje aplicado entre el polisilicio del inyector y el de la compuerta flotante, se requiere conocer el coeficiente de acoplamiento del sistema MOS para saber la magnitud del voltaje inducido sobre esta última y, junto con el campo eléctrico crítico o umbral (ver sección 2.3.2) para el comienzo de la inyección de carga mediante el tunelamiento F-N, poder conocer la magnitud del voltaje que se ha de aplicar al inyector correspondiente, para inyectar la carga positiva o negativa apropiada para desplazar el voltaje de umbral. Utilizando el modelo ilustrado en la Fig. 2.19, el coeficiente de acoplamiento se deduce como sigue:

$$K_{cg} = \frac{C_{pp}}{C_{pp} + C_{ox} + C_{tun} + C_{par}} \quad , \tag{4.6}$$

donde C_{pp} es la capacidad entre la compuerta de control y la compuerta flotante, C_{ox} es la capacidad de compuerta, C_{tun} es la capacidad en el óxido de tunelamiento y C_{par} incluye a las capacidades parásitas que pudieran estar presentes en el sistema. Por lo tanto, del divisor de voltaje formado por estas capacidades (ver Fig. 2.19), se encuentra que el voltaje inducido en la compuerta flotante es el siguiente:

$$V_{fg} = K_{cg} V_{cg} \quad , \tag{4.7}$$

en el caso de que no exista alguna carga presente en la compuerta flotante; sin embargo, cuando se tiene la condición de carga presente, la ecuación anterior deberá ser:

$$V_{fg} = K_{cg}V_{cg} + \frac{Q_{fg}}{C_{pp} + C_{ox} + C_{tun} + C_{par}} \quad . \qquad (4.8)$$

Si se establece al voltaje aplicado al inyector como Vi, se tiene entonces, que la diferencia de potencial a través del óxido de tunelamiento, V_{tun}, se puede expresar como a continuación se indica:

$$V_{tun} = V_i - V_{fg} \quad , \qquad (4.9)$$

$$V_{tun} = V_i - K_{cg}V_{cg} - \frac{Q_{fg}}{C_{pp} + C_{ox} + C_{tun} + C_{par}} \quad . \qquad (4.10)$$

Las ecuaciones (4.3) - (4.10), se desarrollaron adaptando el modelo reportado en [3], donde se utiliza una memoria EEPROM de compuerta flotante, pero cuyo diseño permite la programación mediante la polarización de la compuerta de control, aterrizando al drenador, y el borrado mediante la aplicación de voltaje en el drenador y aterrizando la compuerta de control, es decir, por un lado se tiene corriente de tunelamiento cruzando el óxido presente entre los dos polisilicios (escritura), y por el otro, la corriente de tunelamiento de borrado fluye a través de un óxido ultrafino presente entre el drenador y la compuerta flotante. Para el caso que se está tratando, hay que recordar que la estructura diseñada cuenta con dos inyectores fuera de la zona activa del dispositivo, uno de los cuales, con las condiciones adecuadas, inyecta carga negativa y el otro inyecta carga positiva; el óxido de tunelamiento es únicamente el presente entre los dos polisilicios. El aporte de esta modificación al modelo de Kolodny [3], consiste en incluir tanto el voltaje inducido a la compuerta flotante, como el voltaje aplicado a los inyectores. El uso de las ecuaciones anteriores, permite por lo tanto el cálculo del cambio del voltaje de umbral en función del tiempo durante el cual se aplica dicho voltaje, con ayuda de la siguiente ecuación diferencial:

$$\frac{dQ_{fg}}{dt} = A_{tun}J_{tun} \quad , \qquad (4.11)$$

donde A_{tun} corresponde al área del óxido de tunelamiento. Sustituyendo en la ecuación anterior las ecuaciones (4.3), (4.4), (4.5), (4.6) y (4.10) y resolviendo para el voltaje de umbral final, Vth_{final}, se puede graficar este cambio en función del tiempo, para conocer el ancho de pulso que se ha de aplicar en las terminales para alcanzar un voltaje de umbral deseado. Como se podrá observar de las ecuaciones anteriores, se involucran parámetros geométricos del diseño, considerados en las capacidades y en el área de tunelamiento, por lo que dichos valores se deben deducir del diseño geométrico realizado de los dispositivos del inversor. De la ecuación (4.11) se pueden encontrar dos soluciones dependiendo del tipo de carga que se esté inyectando.

$$\int_0^t dt = \int_{Vth_{inicial}}^{Vth_{final}} \frac{\exp\left[\dfrac{B}{K_{cg}\left(\dfrac{V_i}{K_{cg}} - V_{cg} + V_{th_{nom}} - V\right)}\right]}{K_{cg}A\left(\dfrac{V_i}{K_{cg}} - V_{cg} + V_{th_{nom}} - V\right)} \text{ para } Q_{fg} < 0 \, , \qquad (4.12)$$

de donde, resolviendo para Vth_{final}, se obtiene:

$$V_{thfinal} = \frac{V_i}{K_{cg}} - V_{cg} + V_{thnom} - \frac{B}{K_{cg}\left[\ln\left(BAt + \frac{1}{\exp\left(\frac{B}{-V_i + K_{cg}V_{cg} - K_{cg}V_{thnom} + K_{cg}V_t(0)}\right)}\right)\right]}, \quad (4.13)$$

La otra solución, parte de la siguiente ecuación:

$$\int_0^t dt = \int_{V_{thinicial}}^{V_{thfinal}} \frac{\exp\left[\frac{B}{K_{cg}\left(\frac{V_i}{K_{cg}} - V_{cg} - V_{thnom} + V\right)}\right]}{K_{cg}A\left(\frac{V_i}{K_{cg}} - V_{cg} - V_{thnom} + V\right)} \text{ para } Q_{fg} > 0\,, \quad (4.14)$$

de donde, resolviendo para $V_{thfinal}$, se obtiene:

$$V_{thfinal} = -\frac{V_i}{K_{cg}} + V_{cg} + V_{thnom} + \frac{B}{K_{cg}\left[\ln\left(BAt + \exp\left(\frac{B}{V_i - K_{cg}V_{cg} - K_{cg}V_{thnom} + K_{cg}V_t(0)}\right)\right)\right]}, \quad (4.15)$$

A continuación se definen las constantes A y B usadas en (4.12) - (4.15):

$$A = \frac{A_{tun}\alpha}{X_{tun}^2\left(C_{pp} + C_{ox} + C_{tun} + C_{par}\right)}\,, \quad (4.16)$$

$$B = \beta X_{tun}\,, \quad (4.17)$$

A su vez, V_{thnom} es el voltaje de umbral nominal del dispositivo, tal y como sale de la fabricación y $V_{th}(0)$ corresponde al voltaje de umbral en t=0 ($V_{thinicial}$). Según lo ilustrado en la Fig. 4.4, se puede prever que el voltaje de umbral calculado con la ecuación (4.13) tenderá hacia valores más positivos y el calculado mediante la ecuación (4.15) tenderá hacia valores más negativos.

4.2.1. Cálculo de las constantes α y β.

En la sección 4.1.2, se obtuvieron los parámetros geométricos relacionados con los transistores NMOS y PMOS utilizados en el inversor de la sinapsis, los cuales serán los que se incluyan en las ecuaciones (4.13) y (4.15); sin embargo, falta aún por definir las constantes α y β, las cuales cambian según el intervalo de espesores de óxido considerado [4], [5], [6], [7]. La diferencia en los valores reportados para estas constantes, puede ser crítica al momento de derivar el ancho de pulso del voltaje aplicado para alcanzar el voltaje de umbral deseado, por lo tanto, es deseable considerar aquellas constantes derivadas de la tecnología con la cual se fabricará al circuito, Orbit en este caso. De un estudio hecho a estructuras de inyección de carga sobre compuerta flotante [8], fabricadas mediante Orbit, cuyo espesor de óxido entre Poly1 y Poly2 era de 84 nm, donde únicamente reportan la cantidad de carga FN inyectada en función del voltaje aplicado entre estas dos terminales, se dedujeron los valores de las constantes α y β, las cuales se consideran como apropiadas para los cálculos del presente trabajo, dado que la fábrica es la misma y el espesor del óxido para la opción tecnológica elegida (70 nm), es cercano al utilizado en [7] (opción tecnológica alternativa). Por lo tanto, dada la importancia de la derivación de estas constantes, se incluye a continuación el procedimiento analítico, para su posible utilización en caso de que se haga un estudio posterior de caracterización de la inyección de carga mediante el tunelamiento Fowler-Nordheim, lo cual puede quedar como un trabajo futuro.

En la referencia [7], se pueden encontrar las gráficas de carga inyectada desde Poly2 hacia Poly1 (escritura $\Rightarrow$ Qfg<0) y desde Poly1 hacia Poly2 (borrado $\Rightarrow$ Qfg>0), en función de la diferencia de potencial. Si se grafican estos valores como 1/V vs C/V^2, se tiene que, dada la dependencia exponencial de la corriente de tunelamiento en función del voltaje aplicado (ver ecuación 4.4), en una gráfica semilogarítmica el cruce cuando V=0 corresponderá al valor de α y la pendiente corresponderá al valor de β. Las Figs. 4.6 y 4.7 muestran las gráficas de escritura y borrado, respectivamente, donde se indican las ecuaciones de ajuste a los valores experimentales. De aquí, se deducen dos pares de constantes:

$$\left.\begin{array}{l} \alpha = 5.1 \times 10^{-3}\ A/V^2 \\ \beta = 373\ V \end{array}\right\} \quad \text{cuando } Q_{fg} < 0\ ,$$

y también:

$$\left.\begin{array}{l} \alpha = 13.94 \times 10^{-1}\ A/V^2 \\ \beta = 337\ V \end{array}\right\} \quad \text{cuando } Q_{fg} > 0\ .$$

Esta diferencia en las constantes, puede ser debida a la posible diferencia en la interfaz de las superficies de ambos polisilicios, ya que el óxido que los separa es depositado y puede causar algunos defectos en las superficies emisoras en diferente magnitud, y se refleja también en la diferencia del campo eléctrico umbral en un sentido y en otro de inyección, considerado en la sección 2.3.2. Así entonces, la primer pareja de constantes se deberá sustituir en la ecuación (4.13) y la segunda pareja se deberá sustituir en la ecuación (4.15), pero para que las unidades sean consistentes, el valor de β tiene que ser dividido por el valor del espesor del óxido, 70 nm en este caso. En la Fig. 4.6, las barras de error comprenden un 20 % de error en la graficación y la Fig. 4.7 las barras de error comprenden un 10 % de error.

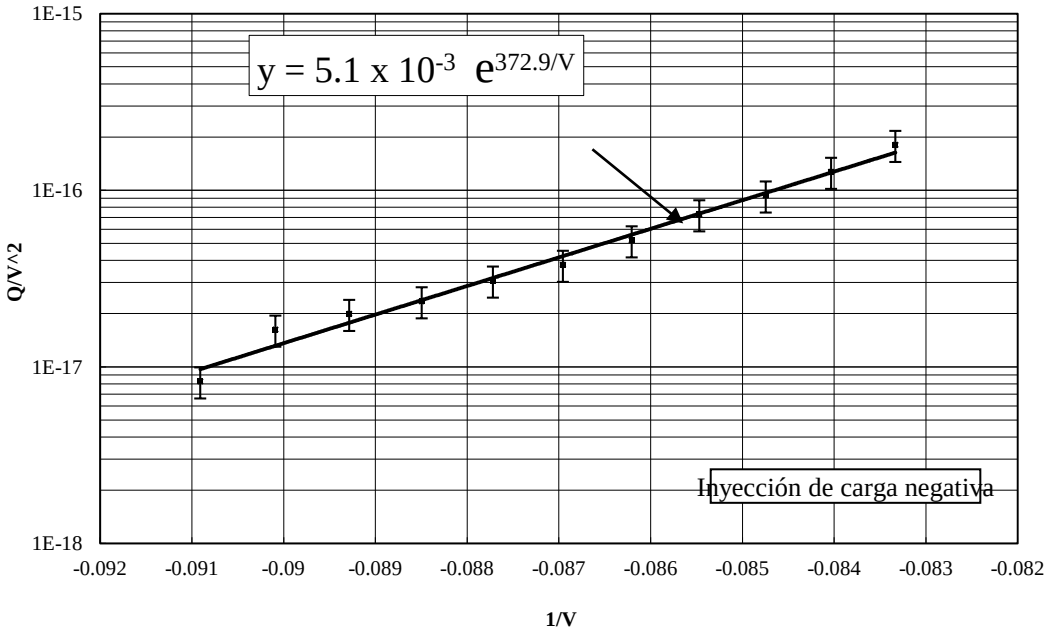


Fig. 4.6. Carga negativa inyectada en función del voltaje aplicado entre Poly1 y Poly2. Deducción de las constantes α y β.

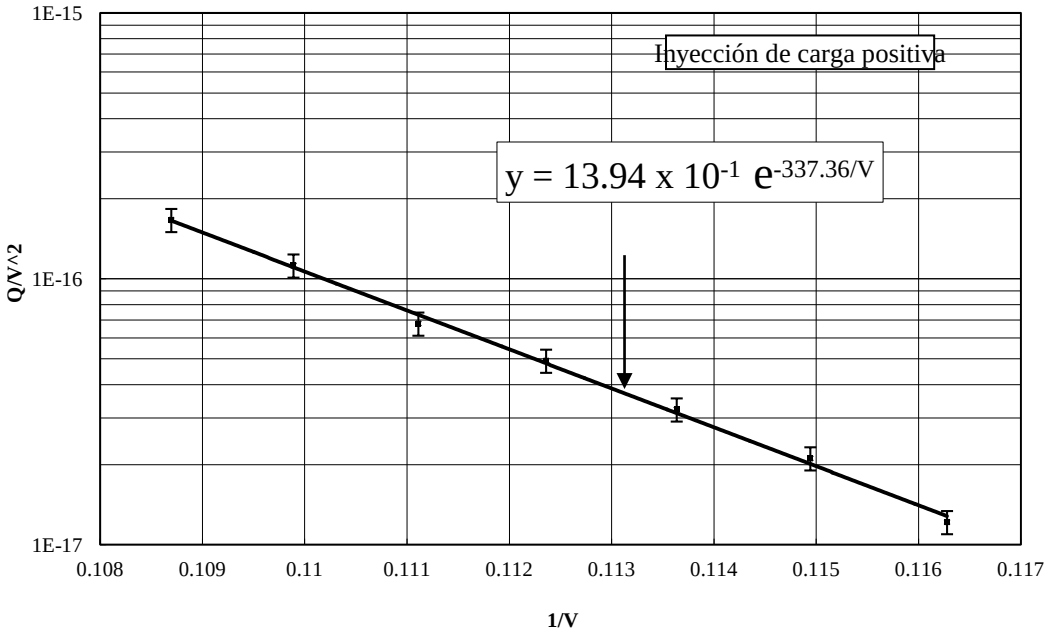


Fig. 4.7. Carga positiva inyectada en función del voltaje aplicado entre Poly1 y Poly2. Deducción de las constantes α y β.

4.2.2. Cálculo del cambio del voltaje de umbral para los transistores NMOS y PMOS, en función del tiempo.

Hasta este momento, ya se cuenta con todos los parámetros necesarios para graficar el cambio del voltaje de umbral en función del ancho de pulso aplicado, sin embargo, es oportuno precisar los valores de un par de parámetros, como lo son Vg y Cpar. El primero se elige como 5 V para hacerlo compatible con las magnitudes de voltaje de los patrones bipolares y de polarización y el segundo, se definirá arbitrariamente como 0.5 pF, dado que con este valor y las otras tres capacidades consideradas a partir del diseño, se obtiene un factor de acoplamiento ligeramente superior a 0.5, el cual podría llegar a ser un valor real. Finalmente, los valores de Vth_{nom}(NMOS) = 0.825 V y Vth_{nom}(PMOS) = -0.703 V, se obtienen de los datos proporcionados por el fabricante, con lo que se pueden sustituir valores en las ecuaciones (4.13) y (4.15) para hacer la estimación de la variación del voltaje de umbral, tanto del transistor NMOS como del PMOS. Para tener la certeza de que los cálculos realizados son los apropiados, se requiere medir experimentalmente el coeficiente de acoplamiento y el voltaje de umbral nominal de las estructuras y en una próxima sección, se propone un método sencillo para poder determinarlos.

Las Figs. 4.8 y 4.9 muestran los resultados de sustituir los valores correspondientes a cada tipo de transistor, considerando que Vcg es de +5 o -5 V con Vi como parámetro, además, se consideró a Vth_{nom} igual a Vth(0), es decir, se ilustra el primer paso de programación o borrado de la estructura. La razón de tener diferentes gráficas, se debe a la diferencia de las dimensiones de los transistores considerados.

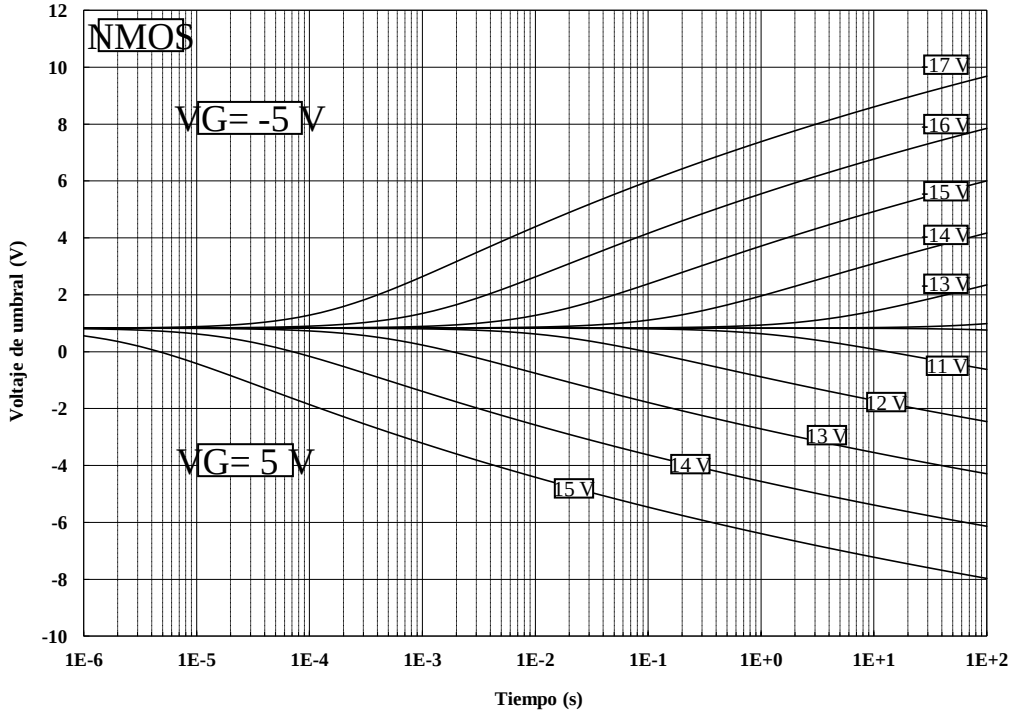


Fig. 4.8. Tiempo de programación para el NMOS. Se considera la aplicación de un voltaje de compuerta de Vcg = 5 V y Vi = 11 V, 12 V, 13 V, 14 V y 15 V para la inyección de carga positiva y con Vcg = - 5 V y Vi = -13 V, -14 V, -15 V, -16 V y -17 V para la inyección de carga negativa.

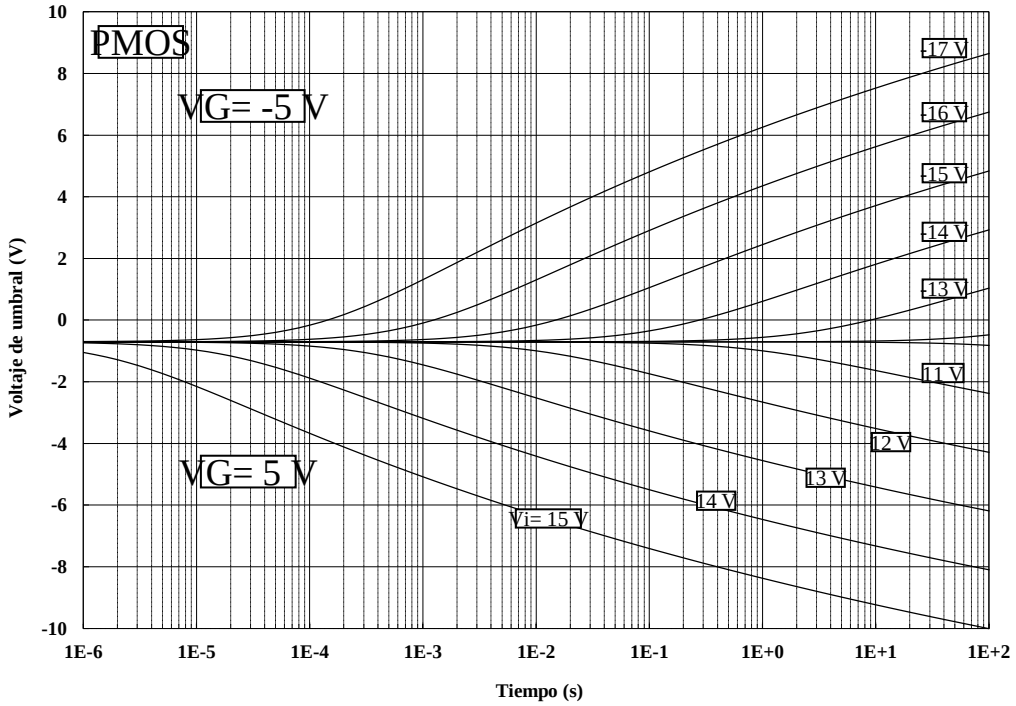


Fig. 4.9. Tiempo de programación para el PMOS. Se considera la aplicación de un voltaje de compuerta de Vcg = 5 V y Vi = 11 V, 12 V, 13 V, 14 V y 15 V para la inyección de carga positiva y con Vcg = - 5 V y Vi = -13 V, -14 V, -15 V, -16 V y -17 V para la inyección de carga negativa.

De lo anterior, se puede ver que entre mayor sea el voltaje aplicado al inyector, se necesitará menor tiempo de programación para desplazar al voltaje de umbral en una misma magnitud. Estas gráficas sirven entonces para determinar las características del pulso que se ha de usar en la red para programar las sinapsis de la BAM.

4.3. Procedimiento para el cálculo del coeficiente de acoplamiento.

Debido a la importancia que tiene el conocer el valor del coeficiente de acoplamiento para el cálculo del cambio del voltaje de umbral en función del tiempo, es conveniente proponer un método con el cual se pueda determinar dicho parámetro. En esta sección, se establece un método sencillo que hace uso del inversor con dispositivos de compuerta flotante, polarizándolo de tal forma que se pueda caracterizar independientemente cada transistor. Esto se hace así por motivo de utilización de área de silicio, ya que un inversor como motivo de prueba incluido en el circuito integrado, puede ser utilizado en diferentes caracterizaciones si se polariza adecuadamente. El método que se propone puede ser realizado sobre dispositivos aislados, con los puntos de operación apropiados aplicados a éste.

El arreglo utilizado para la caracterización del coeficiente de acoplamiento se muestra en la Fig. 4.10, el cual es muy similar al ilustrado en la Fig. 4.2, pero en este caso, se requiere aplicar un voltaje en la terminal de salida del inversor. Cinco son los voltajes involucrados en este arreglo:

1. Voltaje de compuerta del transistor PMOS.
2. Voltaje de compuerta del transistor NMOS.
3. Voltaje de polarización positiva V_{DD}.
4. Voltaje de polarización negativa V_{SS}.
5. Voltaje de drenador V0.

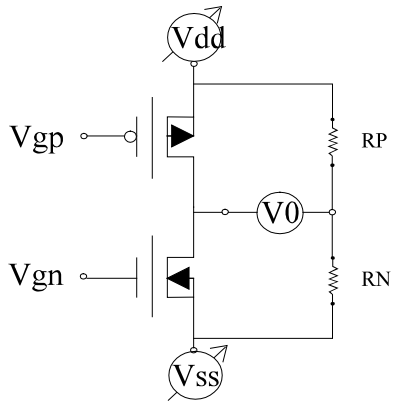


Fig. 4.10. Montaje para la caracterización del coeficiente de acoplamiento, usando un inversor con dispositivos de compuerta flotante.

Para poder considerar sólo uno de los transistores, es necesario llevar al otro a un régimen de corte para evitar que influya en la caracterización del primero. Mediante simulaciones hechas con PSpice y suponiendo valores arbitrarios tanto de voltaje de umbral nominal como de coeficiente de acoplamiento, para tener una idea de las magnitudes de voltajes que se habrán de utilizar para llevar a corte al transistor que no se desea que influya, se deducen las ecuaciones que permiten obtener experimentalmente el valor de Kcg, con el mismo arreglo y para cualquier transistor de compuerta flotante.

Considerando, por ejemplo, los siguientes parámetros para los transistores del inversor:

NMOS: Kcg = 0.1
Vth = 0.825 V

PMOS: Kcg = 0.4
Vth = -0.703 V

y polarizando con V_{DD} = 5 V, V_{SS} = - 5 V y Vgn = - 40 V, se puede obtener la Fig. 4.11.

De la Fig. 4.11 se extrapola una lectura del voltaje de umbral para el transistor PMOS (10.78 V) , el cual no corresponde al Vth nominal especificado, debido a la acción del coeficiente de acoplamiento. Sin embargo, se puede aprovechar esta gráfica para encontrar la siguiente ecuación:

$$V_{DD} = K_{cg}V_{gps} - Vth(PMOS) \quad , \qquad (4.18)$$

donde Vgps corresponde a la lectura extrapolada de voltaje de umbral en la gráfica de $Id^{1/2}$ vs Vgp. Sustituyendo valores se puede ver que se cumple lo anterior:

$$5\ V = (0.4)\ x\ (10.78\ V) - (-0.703)\ .$$

Lo mismo se puede hacer para el transistor NMOS, cuando se polariza la compuerta del transistor PMOS con Vgp = 15 V y se obtiene la ecuación correspondiente para el NMOS:

$$V_{SS} = K_{cg}V_{gns} - Vth(NMOS)\ . \qquad (4.19)$$

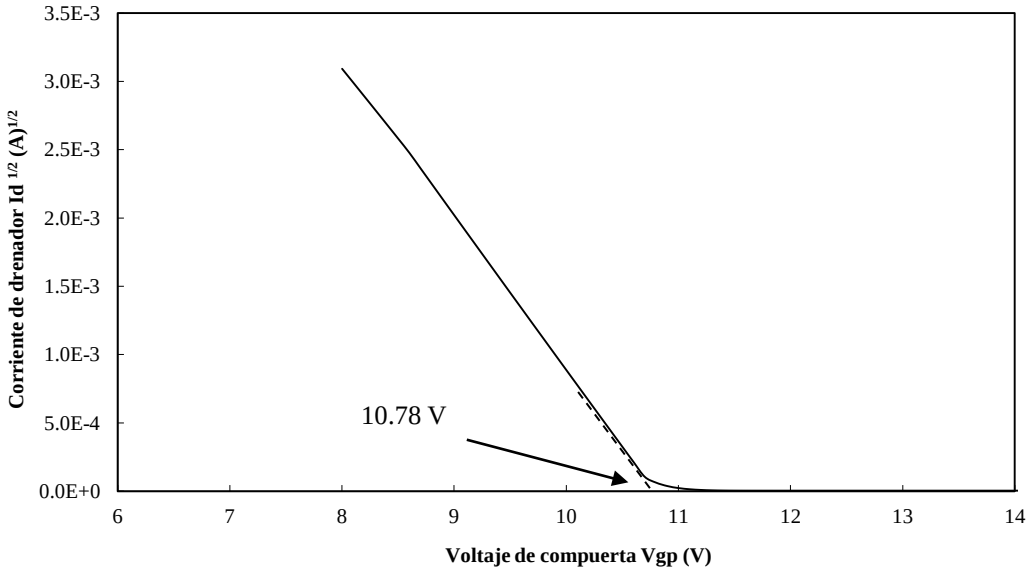
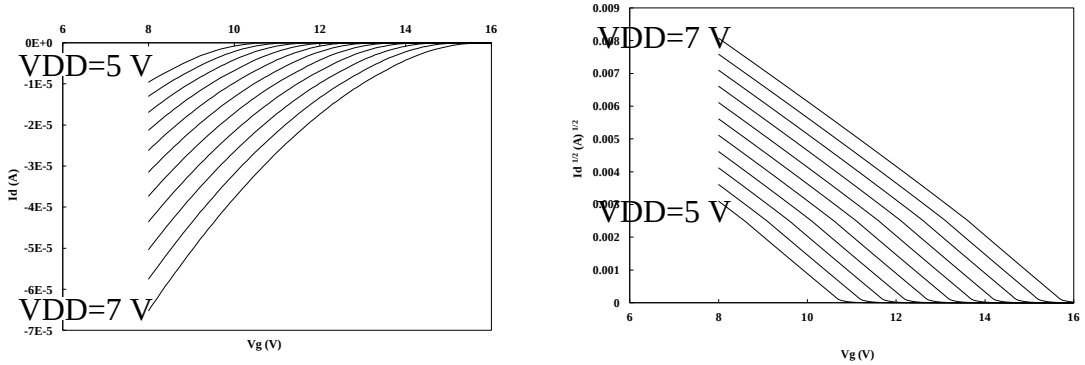


Fig. 4.11. Gráfica de $Id^{1/2}$ vs Vgp del transistor PMOS, usando el arreglo de la Fig. 4.10.

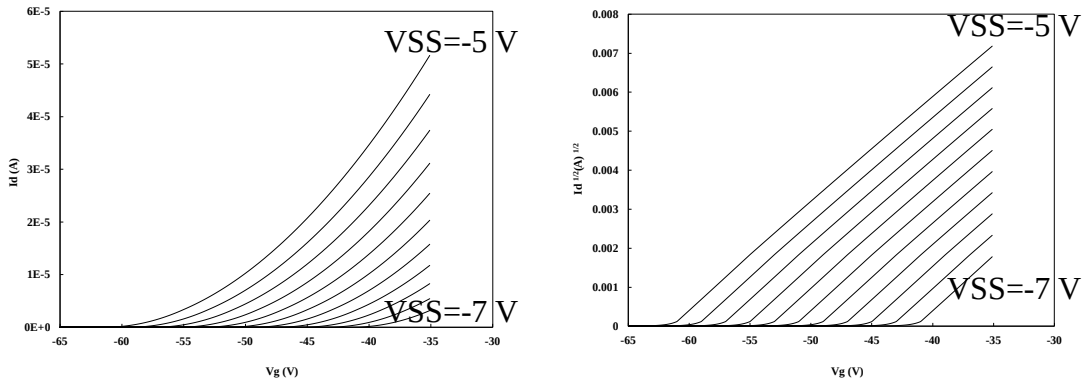
Las ecuaciones (4.18) y (4.19) se pueden utilizar entonces para el cálculo del coeficiente de acoplamiento de cada transistor, siempre y cuando se conozca el voltaje de umbral del dispositivo. Dado que en el caso de la sinapsis, se requiere programar a los transistores, el voltaje de umbral puede tomar un valor aleatorio y ser desconocido en un momento dado, lo que presenta el problema de tener una ecuación con dos incógnitas. Lo anterior se puede resolver polarizando con diferentes pares de voltajes V_{DD} y V_{SS}, de tal forma que se tenga al menos un sistema de dos ecuaciones con dos incógnitas, cuya solución analítica es sencilla. La fuente de voltaje V0 de la Fig. 4.10 se declara con una salida igual a cero volts y dado que $V_{DS} = V_D - V_S$, esta diferencia de potencial estará dada por V_{SS} para el NMOS y por V_{DD} para el PMOS. En el caso del presente trabajo, se utilizó el Sistema I-V, modelo 90 de Keithley, con el cual, se puede declarar una fuente de voltaje de cero volts para que funcione como amperímetro y por consiguiente, no se requieren las resistencias de carga Rp y Rn.

De esta manera, se pueden obtener gráficas como las mostradas en las Fig. 4.12 y 4.13, donde se usaron diferentes parejas de V_{DD} y V_{SS} para obtener diferentes lecturas tanto de Vgps como de Vgns.



(a) (b)

Fig. 4.12. Deducción del factor de acoplamiento para el PMOS.
a) Gráfica Id vs Vgp
b) Gráfica de $Id^{1/2}$ vs Vgp



(a) (b)

Fig. 4.13. Deducción del factor de acoplamiento para el NMOS.
a) Gráfica Id vs Vgn
b) Gráfica de $Id^{1/2}$ vs Vgn

Los valores cubiertos de V_{DD} y V_{SS}, van desde 5 V hasta 7 V y desde -5 V hasta -7 V, respectivamente. Las lecturas correspondientes a estos valores extremos, son las siguientes:

PMOS:	Vgp(5 V) = 10.761 V	Vgp(7 V) = 15.789 V
	V_{DD} = 5 V	V_{DD} = 7 V
	V_{SS} = -5 V	V_{SS} = -7 V
	Kcg = 0.4	Kcg = 0.4

con lo que ya se puede resolver el sistema de dos ecuaciones con Vth(PMOS) y Kcg como incógnitas, teniéndose como resultado que:

Kcg = 0.4
Vth(PMOS) = -0.681 ,

donde se puede ver que el cálculo de Kcg arrojó el valor exacto considerado y el valor del voltaje de umbral presente un error de tan solo el 3 %, que se puede considerar como aceptable y quizá debido al método de extrapolación del voltaje de umbral.

NMOS:	Vgn(5 V) = -41.386 V	Vgn(7 V) = -61.386 V
	V_{DD} = 5 V	V_{DD} = 7 V
	V_{SS} = -5 V	V_{SS} = -7 V
	Kcg = 0.1	Kcg = 0.1 ,

cuya solución para las mismas incógnitas es de:

Kcg = 0.1
Vth(NMOS) = 0.861 ,

donde una vez más, el valor del coeficiente de acoplamiento calculado corresponde al mismo considerado para la simulación, mientras que el voltaje de umbral presenta tan solo un 4 % de error con respecto al propuesto.

Con esto se demuestra la utilidad del método propuesto, además de la sencillez del montaje experimental necesario y del álgebra involucrada.

4.4. Funcionamiento de la BAM.

Una vez desarrollado el modelo para determinar el comportamiento de las celdas que conforman a la BAM, se puede continuar con la simulación del circuito completo, para el estudio del desempeño de la red en su conjunto. Esto se lleva a cabo conectando las sinapsis y los inversores según la configuración establecida en la Fig. 3.7 del capítulo anterior. Utilizando el nivel 2 para el modelo del transistor MOS para poder considerar diferentes valores de voltaje de umbral al cambiar el parámetro VTO de acuerdo a la matriz propia del diseño de la BAM (ver sección 3.1), se puede simular la respuesta del circuito para encontrar el valor más apropiado de Vth en los transistores NMOS y PMOS de la sinapsis. Los transistores de la neurona, al no ser de compuerta flotante, conservan el Vth nominal. El listado utilizado para dicho propósito se puede consultar en el apéndice al final del capítulo y aquí se discutirán los resultados obtenidos de las simulaciones realizadas.

De la matriz deducida en el capítulo anterior, es inmediato pensar que sólo existen dos valores de voltaje de umbral asociados con el (-1) y (3) presentes en la matriz de la BAM de interés. Por otro lado, de la Fig. 4.3 es evidente que la respuesta I-V de la sinapsis no reportaría ninguna ventaja si se amplía el intervalo de voltajes de umbral, dado que como se recordará, se tiene la restricción de simetría en los Vth de los transistores y la linealidad de las curvas por encima de $\Delta Vth = |4|$ V se reduce, además de implicar la aplicación de voltajes de programación mayores, lo cual se desea evitar.

Entonces, la duda de cuál de las curvas de la Fig. 4.3 permite un mejor desempeño de la BAM, se puede despejar cuando se sustituyen las sinapsis programables por simples resistencias y se analiza el comportamiento. El resultado de hacer lo anterior indica que entre mayor sea la diferencia entre las resistencias asociadas con el (-1) y el (3), el desempeño de la red es mejor y esto conduce a elegir las siguientes curvas: [$\Delta Vth(N) = +4$ V, $\Delta Vth(P) = -4$ V] y [$\Delta Vth(N) = -0$ V, $\Delta Vth(P) = 0$ V]; para este caso, la primer pareja tiene asociada una mayor resistencia que la segunda pareja. En el mismo sentido asociado a las resistencias, fue necesario definir con cuál de las dos parejas anteriores, se debería asociar al (-1) y cuál al (3) de la matriz. Escribiendo de nuevo la matriz de la BAM, se tiene que:

$$\begin{bmatrix} -1 & 3 & -1 \\ 3 & -1 & -1 \\ -1 & -1 & 3 \\ -1 & 3 & -1 \\ 3 & -1 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

si se usan las resistencias de la siguiente forma:

$$\begin{bmatrix} 5000G\Omega & 10\Omega & 5000G\Omega \\ 10\Omega & 5000G\Omega & 5000G\Omega \\ 5000G\Omega & 5000G\Omega & 10\Omega \\ 5000G\Omega & 10\Omega & 5000G\Omega \\ 10\Omega & 5000G\Omega & 5000G\Omega \\ 5000G\Omega & 5000G\Omega & 10\Omega \end{bmatrix}$$

la respuesta entregada por la red a los tres patrones almacenados fue la siguiente:

Tabla 4.2.

Patrón de entrada	Patrón de salida	respuesta simulada
1 -1 -1 1 -1 -1	-1 1 -1	0 1 0
-1 1 -1 -1 1 -1	1 -1 -1	1 0 0
-1 -1 1 -1 -1 1	-1 -1 1	0 0 1

y cuando las resistencias se arreglan de la siguiente forma:

$$\begin{bmatrix} 10\Omega & 5000G\Omega & 10\Omega \\ 5000G\Omega & 10\Omega & 10\Omega \\ 10\Omega & 10\Omega & 5000G\Omega \\ 10\Omega & 5000G\Omega & 10\Omega \\ 5000G\Omega & 10\Omega & 10\Omega \\ 10\Omega & 10\Omega & 5000G\Omega \end{bmatrix}$$

la respuesta entregada en este caso para los tres patrones almacenados, fue la siguiente:

Tabla 4.3.

Patrón de entrada	Patrón de salida	respuesta simulada
1 -1 -1 1 -1 -1	-1 1 -1	1 -1 1
-1 1 -1 -1 1 -1	1 -1 -1	-1 1 1
-1 -1 1 -1 -1 1	-1 -1 1	1 1 -1

Analizando la Tabla 4.2, se puede ver que la respuesta es similar a la esperada, con la diferencia de que se recibe un cero en lugar de -1, pero se puede interpretar como un estado sin cambio, es decir, permanece el estado anterior a la transición. La respuesta ilustrada en la Tabla 4.3, corresponde sin embargo al complemento de la respuesta, situación que no se desea, pero que pudiera llegar a ser útil en un momento dado. De lo anterior, se puede decir que la mayor resistencia debe estar asociada al (-1) y la menor resistencia al (3), sin embargo, al trasladar esto a las sinapsis CMOS se obtuvo lo contrario y esto se debe al tipo de respuesta que entrega la sinapsis programable, que como se puede ver de la Fig. 4.3, la pendiente de la respuesta I-V, es negativa. En conclusión, la mayor resistencia debe estar asociada al (3) y la menor al (-1) y esto corresponde a lo mostrado en la Fig. 4.14:

$$\begin{bmatrix} \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = +4\\ \Delta P = -4\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} \\ \begin{pmatrix}\Delta N = +4\\ \Delta P = -4\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = -4\\ \Delta P = +4\end{pmatrix} \\ \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = +4\\ \Delta P = -4\end{pmatrix} \\ \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = +4\\ \Delta P = -4\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} \\ \begin{pmatrix}\Delta N = +4\\ \Delta P = -4\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} \\ \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = 0\\ \Delta P = 0\end{pmatrix} & \begin{pmatrix}\Delta N = +4\\ \Delta P = -4\end{pmatrix} \end{bmatrix}$$

Fig. 4.14. Matriz resultante de la BAM, en función de voltajes de umbral.

quedando de esta manera, definida la matriz de la BAM en función de voltajes de umbral, por lo que se deberán elegir los parámetros de programación que conduzcan a estos valores. Una prueba adicional hecha a la red con estos últimos valores sustituidos en el parámetro VTO de las sinapsis correspondientes, comprobó la bidireccionalidad de la red, es decir, se introdujeron los patrones bipolares de tres elementos a la entrada y se obtuvieron los correctos a la salida, lo mismo que al hacerlo en sentido inverso, en ambos casos, sí se obtuvo -1 en lugar de cero, a diferencia de lo obtenido en la tabla 4.2, al simular con resistencias.

4.4.1. Simulación del funcionamiento en régimen dinámico.

Haciendo una simulación de la BAM en condiciones estáticas, es decir, únicamente se presentaba un patrón a la entrada y analizando la respuesta, se obtenían los patrones deseados representados por voltajes de + 5 V o – 5 V en las terminales correspondientes. Dado que se supone un comportamiento dinámico de la red, la siguiente simulación que corresponde, tiene que ver con el desempeño de la red cuando los vectores son introducidos secuencialmente, lo que relaciona al tiempo y en consecuencia, a las capacitancias. Para fines de claridad de las gráficas que se presentan a continuación, se definen los patrones almacenados como sigue:

Patrón 1: 1 -1 -1 1 -1 -1 **Respuesta: -1 1 -1**
Patrón 2: -1 1 -1 -1 1 -1 **Respuesta: 1 -1 -1**
Patrón 3: -1 -1 1 -1 -1 1 **Respuesta: -1 -1 1**

donde cada elemento del patrón y de la respuesta se representa por:

Patrón: A1 A2 A3 A4 A5 A6 **Respuesta: B1 B2 B3**

En este tipo de caracterización, las capacidades toman un papel importante en el funcionamiento de los circuitos. Una de las capacitancias que mayor influencia puede tener, es la debida al capacitor de acoplamiento de la compuerta de control con la compuerta flotante, lo que involucra al coeficiente de acoplamiento. La simulación con varios valores de este parámetro, puede dar una idea del comportamiento de la red en función del coeficiente de acoplamiento. Los coeficientes utilizados en la simulación, fueron los siguientes:

Kcg(N) = 0.8
Kcg(P) = 0.5

Kcg(N) = 0.8
Kcg(P) = 0.7

Kcg(N) = 0.95
Kcg(P) = 0.9

donde a todos los transistores de compuerta flotante dentro de la red, se les asigna el valor correspondiente, según sean NMOS o PMOS. Los resultados de dicha simulación, se pueden observar en las Figs. 4.15, 4.16 y 4.17.

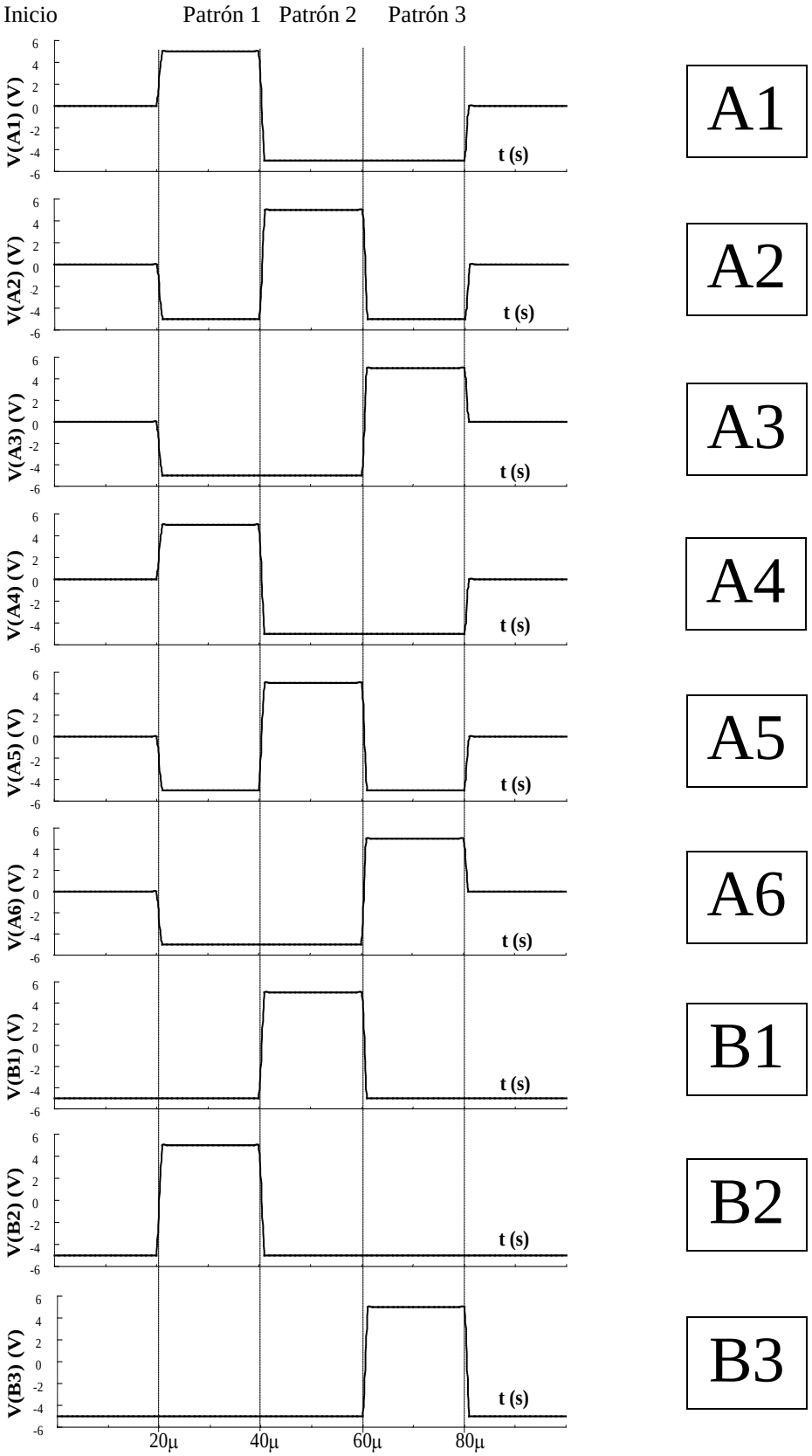


Fig. 4.15. Respuesta de la BAM con Kcg(N) = 0.8 y Kcg(P) = 0.7.

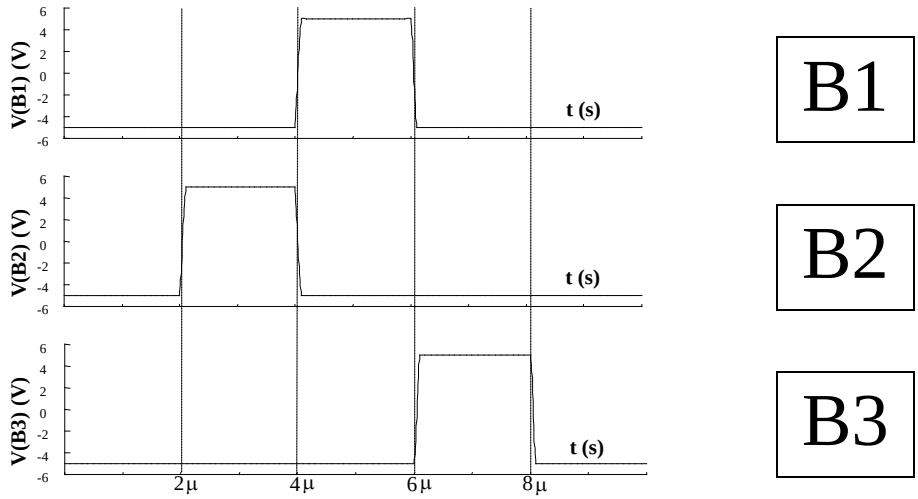


Fig. 4. 16. Respuesta de la BAM con Kcg(N) = 0.8 y Kcg(P) = 0.5.

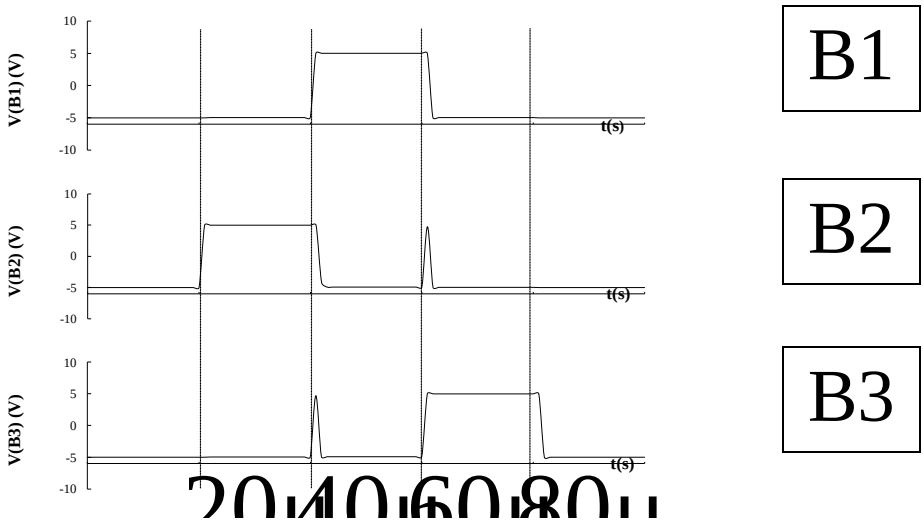
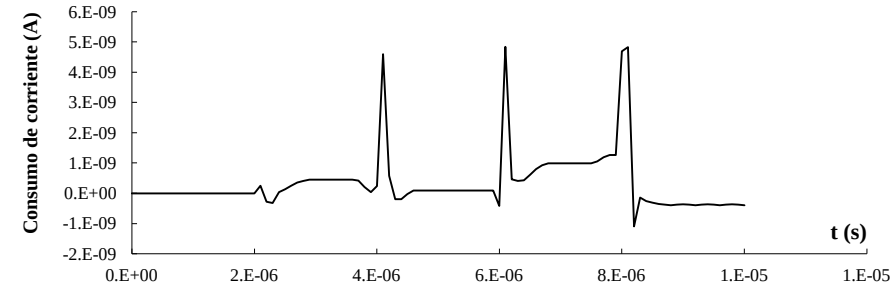


Fig. 4.17. Respuesta de la BAM con Kcg(N) = 0.95 y Kcg(P) = 0.9.

En las figuras anteriores, se presenta la respuesta de la BAM cuando se consideran las tres parejas diferentes de coeficientes de acoplamiento planteados anteriormente. En todas ellas, el reconocimiento de los tres patrones almacenados se logra; sin embargo, la influencia de las capacidades se hace evidente en el procedimiento de simulación, ya que con las siguientes parejas: Kcg(N) = 0.8 y Kcg(P) = 0.7 o bien con Kcg(N) = 0.95 y Kcg(P) = 0.9, el proceso de simulación no convergía mientras el ancho de pulso fuera menor a 20 μs, mientras que cuando se usaba Kcg(N) = 0.8 y Kcg(P) = 0.5, el límite inferior del ancho de pulso para tener convergencia en la simulación fue de 2 μs, causado esto por el algoritmo de procesamiento de PSpice. Por esta razón, en las Fig. 4.15 y 4.17, el eje de tiempo es diferente al presentado en la Fig. 4.16. Otra diferencia evidente se observa en la Fig. 4.17, donde debido a los coeficientes usados, se presentan unos picos no deseados en las salidas B2 y B3.

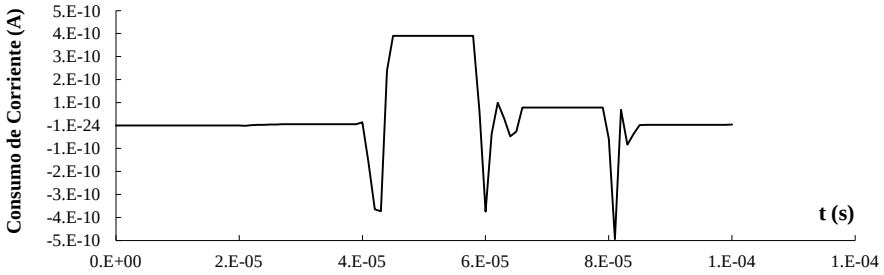
Otra manera de comparar la influencia de Kcg, se puede ver en la Fig. 4.18, donde se grafica el consumo de corriente por una sola neurona para cada pareja de Kcg. Algo que se tiene en común, es el bajo

consumo de corriente, a pesar de que la tercer pareja provoca corriente del orden de microamperes en tan solo dos transitorios; el consumo con las otras dos es mínimo, entre 10^{-9} y 10^{-10} amperes.



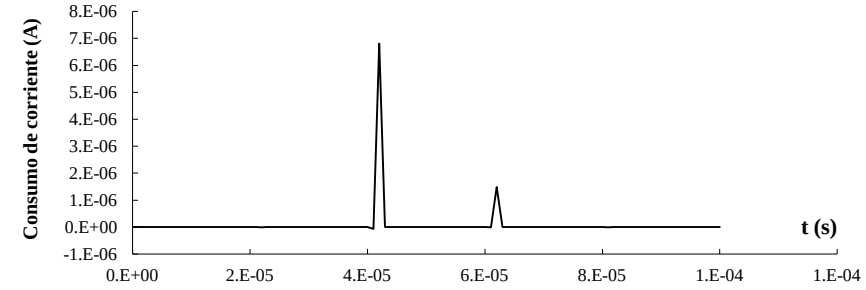
Kcg(N) = 0.8

Kcg(P) = 0.7



Kcg(N) = 0.8

Kcg(P) = 0.5



Kcg(N) = 0.95

Kcg(P) = 0.9

Fig. 4.18. Consumo de corriente de una neurona de la BAM, aplicando diferentes coeficientes de acoplamiento a los transistores de las sinapsis.

Lo anterior permite confirmar la importancia que tiene el diseño del capacitor de acoplamiento en los dispositivos de compuerta flotante, cualquiera que sea su aplicación. También es importante mencionar que los mismos resultados fueron obtenidos cuando se hicieron las simulaciones considerando temperaturas de operación de 0 °C y 150 °C, el cual es un intervalo apropiado de operación para la red.

4.5. Sumario.

En este capítulo se abordó el cálculo y simulación tanto de las células básicas utilizadas, como de una red completa. Se presenta el tipo de función de transferencia elegida para el acotamiento de la señal de entrada a cada neurona, que fue tipo sigmoidal pero muy aproximada a la función escalón, la cual simplifica el análisis teórico de una BAM. Respecto a la sinapsis, se dan los criterios de diseño y simulación de dispositivos de compuerta flotante, así como de la determinación de las capacidades necesarias para el cálculo del coeficiente de acoplamiento, factor de particular importancia en el funcionamiento del transistor de compuerta flotante y su aplicación en circuitos. En este mismo inciso, se aporta un método analítico para la determinación de las características del pulso de programación o borrado del transistor con compuerta flotante, a partir del mecanismo de tunelamiento Fowler-Nordheim, aplicado a los inyectores del dispositivo. Siguiendo con el diseño de la sinapsis, con el objetivo de aplicarla a una red asociativa, también se analiza la forma de determinar la magnitud de la variación del voltaje de umbral de los transistores, de tal forma que se tenga un desempeño eficiente de la BAM prototipo, así como la relación de este último parámetro con la matriz de correlación de la red. Dada la importancia tanto del voltaje de umbral como del coeficiente de acoplamiento, se propone un método sencillo para extraer el valor de estos parámetros, a partir de un sistema de ecuaciones con dos incógnitas. Finalmente, se presentan los resultados de la simulación de la red en conjunto, en condición estática y en condición dinámica, así como en función de las capacitancias presentes en el dispositivo, en la simulación relacionada con el tiempo.

APENDICE.

El siguiente listado corresponde a la simulación de la BAM en condición de CD.

```
*Memoria asociativa bidireccional*
m1          3    2S  1  1        mp w=29u l=10u     ;NEURONAS DE ENTRADA
m2          3     S  4  4        mn w=9u  l=10u     ;                    :
mB1        33     3  1  1        mp w=29u l=10u     ;                    :
m22        33     3  4  4        mn w=9u  l=10u     ;                    :
m3         13    8S  1  1        mp w=29u l=10u     ;                    :
m4         13    8S  4  4        mn w=9u  l=10u     ;                    :
m33       133    13  1  1        mp w=29u l=10u     ;                    :
m44       133    13  4  4        mn w=9u  l=10u     ;                    :
m5         14    9S  1  1        mp w=29u l=10u     ;                    :
m6         14    9S  4  4        mn w=9u  l=10u     ;                    :
m55       144    14  1  1        mp w=29u l=10u     ;                    :
m66       144    14  4  4        mn w=9u  l=10u     ;                    :
m7         15   10S  1  1        mp w=29u l=10u     ;                    :
m8         15   10S  4  4        mn w=9u  l=10u     ;                    :
m77       155    15  1  1        mp w=29u l=10u     ;                    :
m88       155    15  4  4        mn w=9u  l=10u     ;                    :
m9         16   11S  1  1        mp w=29u l=10u     ;                    :
m10        16   11S  4  4        mn w=9u  l=10u     ;                    :
m99       166    16  1  1        mp w=29u l=10u     ;                    :
m100      166    16  4  4        mn w=9u  l=10u     ;                    :
m11        17   12S  1  1        mp w=29u l=10u     ;                    :
m12        17   12S  4  4        mn w=9u  l=10u     ;                    :
m111      177    17  1  1        mp w=29u l=10u     ;                    :
m122      177    17  4  4        mn w=9u  l=10u     ;                    :
m13         5    18  1  1        mp w=29u l=10u     ;NEURONAS DE SALIDA
m14         5    18  4  4        mn w=9u  l=10u     ;                    :
m133       18  188S  1  1        mp w=29u l=10u     ;                    :
m144       18  188S  4  4        mn w=9u  l=10u     ;                    :
m15         6    19  1  1        mp w=29u l=10u     ;                    :
m16         6    19  4  4        mn w=9u  l=10u     ;                    :
m155       19  199S  1  1        mp w=29u l=10u     ;                    :
m166       19  199S  4  4        mn w=9u  l=10u     ;                    :
m17         7    20  1  1        mp w=29u l=10u     ;                    :
m18         7    20  4  4        mn w=9u  l=10u     ;                    :
m177       20  200S  1  1        mp w=29u l=10u     ;                    :
m188       20  200S  4  4        mn w=9u  l=10u     ;                    :
*
.lib c:\alfredo\NIVEL4.lib
*ELEMENTOS SALIDA-ENTRADA.
Mpr11       2 5G 1 1        MP2 W=10U L=9U
MNR11       2 5GN 4 4       MN2 W=4U  L=10U
Mpr21       2 6G 1 1        MP1 W=10U L=9U
MNR21       2 6GN 4 4       MN1 W=4U  L=10U
Mpr31       2 7G 1 1        MP2 W=10U L=9U
MNR31       2 7GN 4 4       MN2 W=4U  L=10U
Mpr12       8 5G 1 1        MP1 W=10U L=9U
MNR12       8 5GN 4 4       MN1 W=4U  L=10U
Mpr22       8 6G 1 1        MP2 W=10U L=9U
MNR22       8 6GN 4 4       MN2 W=4U  L=10U
Mpr32       8 7G 1 1        MP2 W=10U L=9U
MNR32       8 7GN 4 4       MN2 W=4U  L=10U
Mpr13       9 5G 1 1        MP2 W=10U L=9U
MNR13       9 5GN 4 4       MN2 W=4U  L=10U
Mpr23       9 6G 1 1        MP2 W=10U L=9U
MNR23       9 6GN 4 4       MN2 W=4U  L=10U
Mpr33       9 7G 1 1        MP1 W=10U L=9U
MNR33       9 7GN 4 4       MN1 W=4U  L=10U
Mpr14       10 5G 1 1       MP2 W=10U L=9U
MNR14       10 5GN 4 4      MN2 W=4U  L=10U
Mpr24       10 6G 1 1       MP1 W=10U L=9U
MNR24       10 6GN 4 4      MN1 W=4U  L=10U
Mpr34       10 7G 1 1       MP2 W=10U L=9U
MNR34       10 7GN 4 4      MN2 W=4U  L=10U
```

```
Mpr15       11 5G 1 1       MP1 W=10U L=9U
MNR15       11 5GN 4 4      MN1 W=4U  L=10U
Mpr25       11 6G 1 1       MP2 W=10U L=9U
MNR25       11 6GN 4 4      MN2 W=4U  L=10U
Mpr35       11 7G 1 1       MP2 W=10U L=9U
MNR35       11 7GN 4 4      MN2 W=4U  L=10U
Mpr16       12 5G 1 1       MP2 W=10U L=9U
MNR16       12 5GN 4 4      MN2 W=4U  L=10U
Mpr26       12 6G 1 1       MP2 W=10U L=9U
MNR26       12 6GN 4 4      MN2 W=4U  L=10U
Mpr36       12 7G 1 1       MP1 W=10U L=9U
MNR36       12 7GN 4 4      MN1 W=4U  L=10U
*
*ELEMENTOS ENTRADA-SALIDA.
Mprr11      188 33G 1 1     MP2 W=10U L=9U
MNRR11      188 33GN 4 4    MN2 W=4U  L=10U
Mprr21      199 33G 1 1     MP1 W=10U L=9U
MNRR21      199 33GN 4 4    MN1 W=4U  L=10U
Mprr31      200 33G 1 1     MP2 W=10U L=9U
MNRR31      200 33GN 4 4    MN2 W=4U  L=10U
Mprr12      188 133G 1 1    MP1 W=10U L=9U
MNRR12      188 133GN 4 4   MN1 W=4U  L=10U
Mprr22      199 133G 1 1    MP2 W=10U L=9U
MNRR22      199 133GN 4 4   MN2 W=4U  L=10U
Mprr32      200 133G 1 1    MP2 W=10U L=9U
MNRR32      200 133GN 4 4   MN2 W=4U  L=10U
Mprr13      188 144G 1 1    MP2 W=10U L=9U
MNRR13      188 144GN 4 4   MN2 W=4U  L=10U
Mprr23      199 144G 1 1    MP2 W=10U L=9U
MNRR23      199 144GN 4 4   MN2 W=4U  L=10U
Mprr33      200 144G 1 1    MP1 W=10U L=9U
MNRR33      200 144GN 4 4   MN1 W=4U  L=10U
Mprr14      188 155G 1 1    MP2 W=10U L=9U
MNRR14      188 155GN 4 4   MN2 W=4U  L=10U
Mprr24      199 155G 1 1    MP1 W=10U L=9U
MNRR24      199 155GN 4 4   MN1 W=4U  L=10U
Mprr34      200 155G 1 1    MP2 W=10U L=9U
MNRR34      200 155GN 4 4   MN2 W=4U  L=10U
Mprr15      188 166G 1 1    MP1 W=10U L=9U
MNRR15      188 166GN 4 4   MN1 W=4U  L=10U
Mprr25      199 166G 1 1    MP2 W=10U L=9U
MNRR25      199 166GN 4 4   MN2 W=4U  L=10U
Mprr35      200 166G 1 1    MP2 W=10U L=9U
MNRR35      200 166GN 4 4   MN2 W=4U  L=10U
Mprr16      188 177G 1 1    MP2 W=10U L=9U
MNRR16      188 177GN 4 4   MN2 W=4U  L=10U
Mprr26      199 177G 1 1    MP2 W=10U L=9U
MNRR26      199 177GN 4 4   MN2 W=4U  L=10U
Mprr36      200 177G 1 1    MP1 W=10U L=9U
MNRR36      200 177GN 4 4   MN1 W=4U  L=10U
*
C1S 2 0 .01P        ;CAPACIDAD DE ENTRADA A NEURONAS DE ENTRADA
C2S 8 0 .01P
C3S 9 0 .01P
C4S 10 0 .01P
C5S 11 0 .01P
C6S 12 0 .01P
C1SS 33 0 .01P      ;CAPACIDAD DE SALIDA A NEURONAS DE ENTRADA
C2SS 133 0 .01P
C3SS 144 0 .01P
C4SS 155 0 .01P
C5SS 166 0 .01P
C6SS 177 0 .01P
C1E 188 0 .01P      ;CAPACIDAD DE ENTRADA A NEURONAS DE SALIDA
C2E 199 0 .01P
C3E 200 0 .01P
CES1 5 0 .01P       ;CAPACIDAD DE SALIDA A NEURONAS DE SALIDA
CES2 6 0 .01P
CES3 7 0 .01P
*
```

```
*
*
*A CONTINUACION SE ESTABLECEN LOS MODELOS DE LOS TRANSISTORES
*QUE SERVIRAN COMO MEMORIA (FGMOSFET) USANDO EL NIVEL 2,
*CAMBIANDO EL VOLTAJE DE UMBRAL SEGUN LA RESISTENCIA REQUERIDA
*PARA TENER UNA MATRIZ QUE APRENDA TRES PATRONES DE ENTRADA Y
*LOS RECONOZCA.
*
*DELTA VTH(N)=0 V
.MODEL MN1 NMOS Level=2.000 UO=608.3 VTO=825.3E-3 NSS=0.000
+ NFS=0.105E+12 TPG=+1.000 TOX=40.0E-9 NSUB=7.755E+15 UCRIT=50E+3
+ UEXP=78.26E-3 UTRA=0.000 VMAX=49.89E+3 RSH=50.15 XJ=450.0E-9
+ LD=112.1E-9 DELTA=3.714 PB=0.44 JS=10.0E-6 NEFF=3.358 WD=46.34E-9
+ CJ=323.1E-6 MJ=461.5E-3 CJSW=929.9E-12 MJSW=268.3E-3 CGSO=96.77E-12
+ CGDO=96.77E-12 CGBO=40.00E-12FC=500.0E-3 XQC=1.000

*DELTA VTH(P)=0 V
.MODEL MP1 PMOS Level=2.000 UO=205.1 VTO=703.0E-3 NSS=0.000
+ NFS=0.010E+12 TPG=-1.000 TOX=40.0E-9 NSUB=1.486E+16 UCRIT=70E+3
+ UEXP=184.2E-3 UTRA=0.00 VMAX=40.76E+3 RSH=69.46 XJ=450E-9 LD=230.5E-9
+ DELTA=1.843 PB=0.958 JS=10.0E-6 NEFF=0.688 WD=117.6E-9 CJ=804.9E-6
+ MJ=525.0E-3 CJSW=749.1E-12 MJSW=495.4E-3 CGSO=199.0E-12 CGDO=199.0E-12
+ CGBO=101.5E-12 FC=500.0E-3 XQC=1.000
*
*DELTA VTH(N)=+4 V
.MODEL MN2 NMOS Level=2.000 UO=608.3 VTO=4825.3E-3 NSS=0.000
+ NFS=0.105E+12 TPG=+1.000 TOX=40.0E-9 NSUB=7.755E+15 UCRIT=50E+3
+ UEXP=78.26E-3 UTRA=0.000 VMAX=49.89E+3 RSH=50.15 XJ=450.0E-9
+ LD=112.1E-9 DELTA=3.714 PB=0.44 JS=10.0E-6 NEFF=3.358 WD=46.34E-9
+ CJ=323.1E-6 MJ=461.5E-3 CJSW=929.9E-12 MJSW=268.3E-3 CGSO=96.77E-12
+ CGDO=96.77E-12 CGBO=40.00E-12 FC=500.0E-3 XQC=1.000

*DELTA VTH(P)=-4 V
.MODEL MP2 PMOS Level=2.000 UO=205.1 VTO=-4703.0E-3 NSS=0.000
+ NFS=0.010E+12 TPG=-1.000 TOX=40.0E-9 NSUB=1.486E+16 UCRIT=70E+3
+ UEXP=184.2E-3 UTRA=0.00 VMAX=40.76E+3 RSH=69.46 XJ=450E-9 LD=230.5E-9
+ DELTA=1.843 PB=0.958 JS=10.0E-6 NEFF=0.688 WD=117.6E-9 CJ=804.9E-6
+ MJ=525.0E-3 CJSW=749.1E-12 MJSW=495.4E-3 CGSO=199.0E-12 CGDO=199.0E-12
+ CGBO=101.5E-12 FC=500.0E-3 XQC=1.000
*
*V0 2A 2 0  ;AMPERIMETRO PARA LA CORRIENTE DE ENTRADA A NEURONA 1
vdd 1 0  5
vss 4 0 -5
*v1  5 0  -5  ; |
*v2  6 0  -5  ; > PATRON DE SALIDA B1, B2, B3
*v3  7 0   5  ; |
*
EV5  5G 0 5 0 0.5     ;          |
EV5N  5GN 0 5 0 0.8 ;            |
EV6  6G 0 6 0 0.5     ;          \ FUENTES DE VOLTAJE CONTROLADAS POR
EV6N  6GN 0 6 0 0.8 ;            / VOLTAJE CON KCG DE 0.5 Y 0.8. CAPA B
EV7  7G 0 7 0 0.5     ;          |
EV7N 7GN 0 7 0 0.8  ;            |
*
EV33 33G 0 33 0 0.5              ;          |
EV33N 33GN 0 33 0 0.8            ;          |
EV133 133G 0 133 0 0.5           ;          |
EV133N 133GN 0 133 0 0.8         ;          |
EV144 144G 0 144 0 0.5           ;          |
EV144N 144GN 0 144 0 0.8         ;          \ FUENTES DE VOLTAJE CONTROLADAS POR
EV155 155G 0 155 0 0.5           ;          / VOLTAJE CON KCG DE 0.5 Y 0.8. CAPA A.
EV155N 155GN 0 155 0 0.8         ;          |
EV166 166G 0 166 0 0.5           ;          |
EV166N 166GN 0 166 0 0.8         ;          |
EV177 177G 0 177 0 0.5           ;          |
EV177N 177GN 0 177 0 0.8         ;          |

V33         33          0  5         ; |
V133        133         0 -5         ; |
V144        144         0 -5         ; \ PATRON DE ENTRADA A1, A2 ,A3, A4, A5, A6
```

```
V155      155       0  5       ; /
V166      166       0 -5       ; |
V177      177       0 -5       ; |
*
.OP
.OPTION NOPAGE
.print DC v(33) v(133) v(144) v(155) v(166) v(177)
.print DC v(5) v(6) v(7)
.end
```

El siguiente listado corresponde a la simulación de la BAM en función del tiempo, con los patrones introducidos secuencialmente.

```
*Memoria asociativa bidireccional*
m1          3    2S  1  1       mp w=29u l=10u     ;NEURONAS DE ENTRADA
m2          3     S  4  4       mn w=9u  l=10u     ;                    :
mB1        33     3  1  1       mp w=29u l=10u     ;                    :
m22        33     3  4  4       mn w=9u  l=10u     ;                    :
m3         13    8S  1  1       mp w=29u l=10u     ;                    :
m4         13    8S  4  4       mn w=9u  l=10u     ;                    :
m33       133    13  1  1       mp w=29u l=10u     ;                    :
m44       133    13  4  4       mn w=9u  l=10u     ;                    :
m5         14    9S  1  1       mp w=29u l=10u     ;                    :
m6         14    9S  4  4       mn w=9u  l=10u     ;                    :
m55       144    14  1  1       mp w=29u l=10u     ;                    :
m66       144    14  4  4       mn w=9u  l=10u     ;                    :
m7         15   10S  1  1       mp w=29u l=10u     ;                    :
m8         15   10S  4  4       mn w=9u  l=10u     ;                    :
m77       155    15  1  1       mp w=29u l=10u     ;                    :
m88       155    15  4  4       mn w=9u  l=10u     ;                    :
m9         16   11S  1  1       mp w=29u l=10u     ;                    :
m10        16   11S  4  4       mn w=9u  l=10u     ;                    :
m99       166    16  1  1       mp w=29u l=10u     ;                    :
m100      166    16  4  4       mn w=9u  l=10u     ;                    :
m11        17   12S  1  1       mp w=29u l=10u     ;                    :
m12        17   12S  4  4       mn w=9u  l=10u     ;                    :
m111      177    17  1  1       mp w=29u l=10u     ;                    :
m122      177    17  4  4       mn w=9u  l=10u     ;                    :
m13         5    18  1  1       mp w=29u l=10u     ;NEURONAS DE SALIDA
m14         5    18  4  4       mn w=9u  l=10u     ;                    :
m133       18  188S  1  1       mp w=29u l=10u     ;                    :
m144       18  188S  4  4       mn w=9u  l=10u     ;                    :
m15         6    19  1  1       mp w=29u l=10u     ;                    :
m16         6    19  4  4       mn w=9u  l=10u     ;                    :
m155       19  199S  1  1       mp w=29u l=10u     ;                    :
m166       19  199S  4  4       mn w=9u  l=10u     ;                    :
m17         7    20  1  1       mp w=29u l=10u     ;                    :
m18         7    20  4  4       mn w=9u  l=10u     ;                    :
m177       20  200S  1  1       mp w=29u l=10u     ;                    :
m188       20  200S  4  4       mn w=9u  l=10u     ;                    :
*
.lib c:\alfredo\NIVEL4.lib
*ELEMENTOS SALIDA-ENTRADA.
Mpr11       2 5G 1 1        MP2 W=10U L=9U
MNR11       2 5GN 4 4       MN2 W=4U  L=10U
Mpr21       2 6G 1 1        MP1 W=10U L=9U
MNR21       2 6GN 4 4       MN1 W=4U  L=10U
Mpr31       2 7G 1 1        MP2 W=10U L=9U
MNR31       2 7GN 4 4       MN2 W=4U  L=10U
Mpr12       8 5G 1 1        MP1 W=10U L=9U
MNR12       8 5GN 4 4       MN1 W=4U  L=10U
Mpr22       8 6G 1 1        MP2 W=10U L=9U
MNR22       8 6GN 4 4       MN2 W=4U  L=10U
Mpr32       8 7G 1 1        MP2 W=10U L=9U
MNR32       8 7GN 4 4       MN2 W=4U  L=10U
Mpr13       9 5G 1 1        MP2 W=10U L=9U
MNR13       9 5GN 4 4       MN2 W=4U  L=10U
Mpr23       9 6G 1 1        MP2 W=10U L=9U
MNR23       9 6GN 4 4       MN2 W=4U  L=10U
```

```
Mpr33       9 7G 1 1          MP1 W=10U L=9U
MNR33       9 7GN 4 4         MN1 W=4U  L=10U
Mpr14       10 5G 1 1         MP2 W=10U L=9U
MNR14       10 5GN 4 4        MN2 W=4U  L=10U
Mpr24       10 6G 1 1         MP1 W=10U L=9U
MNR24       10 6GN 4 4        MN1 W=4U  L=10U
Mpr34       10 7G 1 1         MP2 W=10U L=9U
MNR34       10 7GN 4 4        MN2 W=4U  L=10U
Mpr15       11 5G 1 1         MP1 W=10U L=9U
MNR15       11 5GN 4 4        MN1 W=4U  L=10U
Mpr25       11 6G 1 1         MP2 W=10U L=9U
MNR25       11 6GN 4 4        MN2 W=4U  L=10U
Mpr35       11 7G 1 1         MP2 W=10U L=9U
MNR35       11 7GN 4 4        MN2 W=4U  L=10U
Mpr16       12 5G 1 1         MP2 W=10U L=9U
MNR16       12 5GN 4 4        MN2 W=4U  L=10U
Mpr26       12 6G 1 1         MP2 W=10U L=9U
MNR26       12 6GN 4 4        MN2 W=4U  L=10U
Mpr36       12 7G 1 1         MP1 W=10U L=9U
MNR36       12 7GN 4 4        MN1 W=4U  L=10U
*
*ELEMENTOS ENTRADA-SALIDA.
Mprr11      188 33G 1 1               MP2 W=10U L=9U
MNRR11      188 33GN 4 4              MN2 W=4U  L=10U
Mprr21      199 33G 1 1               MP1 W=10U L=9U
MNRR21      199 33GN 4 4              MN1 W=4U  L=10U
Mprr31      200 33G 1 1               MP2 W=10U L=9U
MNRR31      200 33GN 4 4              MN2 W=4U  L=10U
Mprr12      188 133G 1 1              MP1 W=10U L=9U
MNRR12      188 133GN 4 4             MN1 W=4U  L=10U
Mprr22      199 133G 1 1              MP2 W=10U L=9U
MNRR22      199 133GN 4 4             MN2 W=4U  L=10U
Mprr32      200 133G 1 1              MP2 W=10U L=9U
MNRR32      200 133GN 4 4             MN2 W=4U  L=10U
Mprr13      188 144G 1 1              MP2 W=10U L=9U
MNRR13      188 144GN 4 4             MN2 W=4U  L=10U
Mprr23      199 144G 1 1              MP2 W=10U L=9U
MNRR23      199 144GN 4 4             MN2 W=4U  L=10U
Mprr33      200 144G 1 1              MP1 W=10U L=9U
MNRR33      200 144GN 4 4             MN1 W=4U  L=10U
Mprr14      188 155G 1 1              MP2 W=10U L=9U
MNRR14      188 155GN 4 4             MN2 W=4U  L=10U
Mprr24      199 155G 1 1              MP1 W=10U L=9U
MNRR24      199 155GN 4 4             MN1 W=4U  L=10U
Mprr34      200 155G 1 1              MP2 W=10U L=9U
MNRR34      200 155GN 4 4             MN2 W=4U  L=10U
Mprr15      188 166G 1 1              MP1 W=10U L=9U
MNRR15      188 166GN 4 4             MN1 W=4U  L=10U
Mprr25      199 166G 1 1              MP2 W=10U L=9U
MNRR25      199 166GN 4 4             MN2 W=4U  L=10U
Mprr35      200 166G 1 1              MP2 W=10U L=9U
MNRR35      200 166GN 4 4             MN2 W=4U  L=10U
Mprr16      188 177G 1 1              MP2 W=10U L=9U
MNRR16      188 177GN 4 4             MN2 W=4U  L=10U
Mprr26      199 177G 1 1              MP2 W=10U L=9U
MNRR26      199 177GN 4 4             MN2 W=4U  L=10U
Mprr36      200 177G 1 1              MP1 W=10U L=9U
MNRR36      200 177GN 4 4             MN1 W=4U  L=10U
*
C1S 2 0 .01P          ;CAPACIDAD DE ENTRADA A NEURONAS DE ENTRADA
C2S 8 0 .01P
C3S 9 0 .01P
C4S 10 0 .01P
C5S 11 0 .01P
C6S 12 0 .01P
C1SS 33 0 .01P        ;CAPACIDAD DE SALIDA A NEURONAS DE ENTRADA
C2SS 133 0 .01P
C3SS 144 0 .01P
C4SS 155 0 .01P
C5SS 166 0 .01P
```

```
C6SS 177 0 .01P
C1E 188 0 .01P          ;CAPACIDAD DE ENTRADA A NEURONAS DE SALIDA
C2E 199 0 .01P
C3E 200 0 .01P
CES1 5 0 .01P           ;CAPACIDAD DE SALIDA A NEURONAS DE SALIDA
CES2 6 0 .01P
CES3 7 0 .01P
*
*A CONTINUACION SE ESTABLECEN LOS MODELOS DE LOS TRANSISTORES
*QUE SERVIRAN COMO MEMORIA (FGMOSFET) USANDO EL NIVEL 2,
*CAMBIANDO EL VOLTAJE DE UMBRAL SEGUN LA RESISTENCIA REQUERIDA
*PARA TENER UNA MATRIZ QUE APRENDA TRES PATRONES DE ENTRADA Y
*LOS RECONOZCA.
*
*DELTA VTH(N)=0 V
.MODEL MN1 NMOS Level=2.000 UO=608.3 VTO=825.3E-3 NSS=0.000
+ NFS=0.105E+12 TPG=+1.000 TOX=40.0E-9 NSUB=7.755E+15 UCRIT=50E+3
+ UEXP=78.26E-3 UTRA=0.000 VMAX=49.89E+3 RSH=50.15 XJ=450.0E-9
+ LD=112.1E-9 DELTA=3.714 PB=0.44 JS=10.0E-6 NEFF=3.358 WD=46.34E-9
+ CJ=323.1E-6 MJ=461.5E-3 CJSW=929.9E-12 MJSW=268.3E-3 CGSO=96.77E-12
+ CGDO=96.77E-12 CGBO=40.00E-12FC=500.0E-3 XQC=1.000

*DELTA VTH(P)=0 V
.MODEL MP1 PMOS Level=2.000 UO=205.1 VTO=703.0E-3 NSS=0.000
+ NFS=0.010E+12 TPG=-1.000 TOX=40.0E-9 NSUB=1.486E+16 UCRIT=70E+3
+ UEXP=184.2E-3 UTRA=0.00 VMAX=40.76E+3 RSH=69.46 XJ=450E-9 LD=230.5E-9
+ DELTA=1.843 PB=0.958 JS=10.0E-6 NEFF=0.688 WD=117.6E-9 CJ=804.9E-6
+ MJ=525.0E-3 CJSW=749.1E-12 MJSW=495.4E-3 CGSO=199.0E-12 CGDO=199.0E-12
+ CGBO=101.5E-12 FC=500.0E-3 XQC=1.000
*
*DELTA VTH(N)=+4 V
.MODEL MN2 NMOS Level=2.000 UO=608.3 VTO=4825.3E-3 NSS=0.000
+ NFS=0.105E+12 TPG=+1.000 TOX=40.0E-9 NSUB=7.755E+15 UCRIT=50E+3
+ UEXP=78.26E-3 UTRA=0.000 VMAX=49.89E+3 RSH=50.15 XJ=450.0E-9
+ LD=112.1E-9 DELTA=3.714 PB=0.44 JS=10.0E-6 NEFF=3.358 WD=46.34E-9
+ CJ=323.1E-6 MJ=461.5E-3 CJSW=929.9E-12 MJSW=268.3E-3 CGSO=96.77E-12
+ CGDO=96.77E-12 CGBO=40.00E-12 FC=500.0E-3 XQC=1.000

*DELTA VTH(P)=-4 V
.MODEL MP2 PMOS Level=2.000 UO=205.1 VTO=-4703.0E-3 NSS=0.000
+ NFS=0.010E+12 TPG=-1.000 TOX=40.0E-9 NSUB=1.486E+16 UCRIT=70E+3
+ UEXP=184.2E-3 UTRA=0.00 VMAX=40.76E+3 RSH=69.46 XJ=450E-9 LD=230.5E-9
+ DELTA=1.843 PB=0.958 JS=10.0E-6 NEFF=0.688 WD=117.6E-9 CJ=804.9E-6
+ MJ=525.0E-3 CJSW=749.1E-12 MJSW=495.4E-3 CGSO=199.0E-12 CGDO=199.0E-12
+ CGBO=101.5E-12 FC=500.0E-3 XQC=1.000
*
*V0 2A 2 0  ;AMPERIMETRO PARA LA CORRIENTE DE ENTRADA A NEURONA 1
vdd 1 0  5
vss 4 0 -5
*v1  5 0  -5  ; |
*v2  6 0  -5  ; > PATRON DE SALIDA B1, B2, B3
*v3  7 0   5  ; |
*
EV5  5G 0 5 0 0.5     ;          |
EV5N  5GN 0 5 0 0.8 ;            |
EV6  6G 0 6 0 0.5     ;          \ FUENTES DE VOLTAJE CONTROLADAS POR
EV6N  6GN 0 6 0 0.8 ;            / VOLTAJE CON KCG DE 0.5 Y 0.8. CAPA B
EV7  7G 0 7 0 0.5     ;          |
EV7N 7GN 0 7 0 0.8  ;            |
*
EV33 33G 0 33 0 0.5          ;          |
EV33N 33GN 0 33 0 0.8        ;          |
EV133 133G 0 133 0 0.5       ;          |
EV133N 133GN 0 133 0 0.8     ;          |
EV144 144G 0 144 0 0.5       ;          |
EV144N 144GN 0 144 0 0.8     ;          \ FUENTES DE VOLTAJE CONTROLADAS POR
EV155 155G 0 155 0 0.5       ;          / VOLTAJE CON KCG DE 0.5 Y 0.8. CAPA A.
EV155N 155GN 0 155 0 0.8     ;          |
EV166 166G 0 166 0 0.5       ;          |
EV166N 166GN 0 166 0 0.8     ;          |
```

```
EV177 177G 0 177 0 0.5          ;           |
EV177N 177GN 0 177 0 0.8        ;           |
*
*

V33         33          0  5        ;           |
V133        133         0 -5        ;           |
V144        144         0 -5        ;           \ PATRON DE ENTRADA A1, A2 ,A3, A4, A5, A6
V155        155         0  5        ;           /
V166        166         0 -5        ;           |
V177        177         0 -5        ;           |
*
.OP
.OPTION NOPAGE
*.print TRAN v(33) v(133) v(144) v(155) v(166) v(177)
*.print TRAN v(5) v(6) v(7)
.PRINT TRAN I(V0)
*
V33D 33 0
+ PWL(
+  0.000000E+00  0.000000E+00
+  2.000000E-06  0.000000E+00
+  2.020000E-06  5.000000E+00
+  4.000000E-06  5.000000E+00
+  4.020000E-06 -5.000000E+00
+  6.000000E-06 -5.000000E+00
+  6.020000E-06 -5.000000E+00
+  8.000000E-06 -5.000000E+00
+  8.020000E-06  0.000000E+00
+  1.000000E-05  0.000000E+00
+)
V133D 133 0
+ PWL(
+  0.000000E+00  0.000000E+00
+  2.000000E-06  0.000000E+00
+  2.020000E-06 -5.000000E+00
+  4.000000E-06 -5.000000E+00
+  4.020000E-06  5.000000E+00
+  6.000000E-06  5.000000E+00
+  6.020000E-06 -5.000000E+00
+  8.000000E-06 -5.000000E+00
+  8.020000E-06  0.000000E+00
+  1.000000E-05  0.000000E+00
+)
V144D 144 0
+ PWL(
+  0.000000E+00  0.000000E+00
+  2.000000E-06  0.000000E+00
+  2.020000E-06 -5.000000E+00
+  4.000000E-06 -5.000000E+00
+  4.020000E-06 -5.000000E+00
+  6.000000E-06 -5.000000E+00
+  6.020000E-06  5.000000E+00
+  8.000000E-06  5.000000E+00
+  8.020000E-06  0.000000E+00
+  1.000000E-05  0.000000E+00
+)
V155D 155 0
+ PWL(
+  0.000000E+00  0.000000E+00
+  2.000000E-06  0.000000E+00
+  2.020000E-06  5.000000E+00
+  4.000000E-06  5.000000E+00
+  4.020000E-06 -5.000000E+00
+  6.000000E-06 -5.000000E+00
+  6.020000E-06 -5.000000E+00
+  8.000000E-06 -5.000000E+00
+  8.020000E-06  0.000000E+00
+  1.000000E-05  0.000000E+00
+)
```

```
V166D 166 0
+ PWL(
+  0.000000E+00  0.000000E+00
+  2.000000E-06  0.000000E+00
+  2.020000E-06 -5.000000E+00
+  4.000000E-06 -5.000000E+00
+  4.020000E-06  5.000000E+00
+  6.000000E-06  5.000000E+00
+  6.020000E-06 -5.000000E+00
+  8.000000E-06 -5.000000E+00
+  8.020000E-06  0.000000E+00
+  1.000000E-05  0.000000E+00
+)
V177D 177 0
+ PWL(
+  0.000000E+00  0.000000E+00
+  2.000000E-06  0.000000E+00
+  2.020000E-06 -5.000000E+00
+  4.000000E-06 -5.000000E+00
+  4.020000E-06 -5.000000E+00
+  6.000000E-06 -5.000000E+00
+  6.020000E-06  5.000000E+00
+  8.000000E-06  5.000000E+00
+  8.020000E-06  0.000000E+00
+  1.000000E-05  0.000000E+00
+)
.TRAN .1U 10U
.END
```

Referencias.

1.- S. Kim, Y. Shin, N. C. R. Bogineni and R. Shridhar, "A programmable analog CMOS synapse for neural networks", *Analog Integrated Circuits and Signal Processing*, Vol. 2, 1992, pp. 345-352.

2.- A. Thomsen and M. A. Brooke, "A floating-gate MOSFET with tunneling inyector fabricated using a standard double-polysilicon CMOS process", *IEEE Electron Devices Letters*, Vol. 12, No. 3, March 1991, pp. 111-113.

3.- A. Kolodny, S. T. K. Nieh, B. Eitan and J. Shappir, "Analysis and modeling of floating-gate EEPROM cells", *IEEE Transactions on Electron Devices*, Vol. ED-33, No. 6, June 1986, pp. 835-844.

4.- D. A. Durfee and F. S. Shoucair, "Comparison of floating-gate neural network memory cells in standard VLSI CMOS technology", *IEEE Transactions on Neural Networks*, Vol. 3, No. 3, May 1992, pp. 347-352.

5.- H. C. Card and W. R. Moore, "EEPROM synapse exhibiting pseudo-hebbian plasticity", *Electronics Letters*, Vol. 25, No. 12, June 1989, pp. 805-806.

6.- S. Keeney, R. Bez, D. Cantarelli, F. Piccini, A. Mathewson, L. Ravazzi and C. Lombardi, "Complete transient simulation of flash EEPROM devices", *IEEE Transactions on Electron Devices*, Vol. 39, No. 12, Dec. 1992, pp. 2750-2757.

7.- A. Concannon, S. Keeney, A. Mathewson, R. Bez and C. Lombardi, "Two dimensional numerical analysis of floating-gate EEPROM devices", *IEEE Transactions on Electron Devices*, Vol. 40, No. 7, July 1993, pp. 1258-1262.

8.- K. Chao and M. Chen, "Fowler-Nordheim limited band-to-band tunneling (FNBB) for p-MOSFET gate current in a floating bulk condition", *Solid State Electronics*, Vol. 38, No. 1, 1995, pp. 135-137.

9.- Y. Wang, J. B. Cruz and J. H. Mulligan, "Two coding strategies for bidirectional associative memory", *IEEE Transactions on Neural Networks*, Vol. 1, No. 1, March 1990, pp. 81-92.

10.- W. Wang and D. Lee, "A modified bidirectional decoding strategy based on the BAM structure", *IEEE Transactions on Neural Networks*, Vol. 4, No. 4, July 1993, pp. 710-717.

11.- B. Zhang, B. Xu and C. Kwong, "Performance analysis of the bidirectional associative memory and an improved model from the matched-filtering viewpoint", *IEEE Transactions on Neural Networks*, Vol. 4 No. 5, Sept. 1993, pp, 864-881.

12.- T. Wang, X. Zhuang and X. Xing, "Designing bidirectional associative memories with optimal stability", *IEEE Transaction on Systems, Man and Cybernetics*, Vol. 24, No. 5, May 1994, pp. 778-790.

Capítulo 5.

Diseño topológico.

En este capítulo se presenta el diseño topológico de las celdas presentadas en el capítulo anterior, utilizando las reglas de diseño de la tecnología Orbit, de 2 μm, doble polisilicio, doble metal y pozo N. El programa de edición gráfica que se utiliza, es el proporcionado por la compañía **Tanner Research, Inc**, llamado **L-Edit**, que contiene las reglas de diseño de diferentes tecnologías ofrecidas por la fábrica de silicio; en particular, se cuenta con las de **Orbit**.

Se llaman reglas de diseño, al conjunto de directrices necesarias para dibujar geométricamente a los dispositivos que forman un circuito integrado, y que aseguren el buen funcionamiento de éste. Las reglas toman en cuenta dos factores para las limitaciones del diseño topológico: la resolución fotolitográfica y los parámetros eléctricos de los modelos de los dispositivos. Conforme la tecnología avanza hacia dimensiones más pequeñas, se hace necesario el desarrollo de nuevos modelos que ayuden a la simulación de los circuitos y que tomen en cuenta las dimensiones topológicas.

El programa L-Edit con el que se cuenta tiene bibliotecas pre-diseñadas de circuitos comunes y que pueden ser utilizadas para configurar un circuito en particular. Se tiene la libertad de hacer las modificaciones pertinentes, como por ejemplo la longitud de canal (L) o el ancho de canal (W) de los MOSFET, para ajustarlos según el diseño realizado. En caso de que no existan las bibliotecas, es posible hacer el diseño siempre y cuando se sigan las reglas, como es el caso de los FGMOSFET, los cuales no cuentan con bibliotecas.

Como referencia para la posible utilización del paquete empleado en este trabajo, se transcribe la ficha de los archivos utilizados, tal y como los presenta L-Edit:

Vendor: MOSIS: Orbit Semiconductor.
Technology: 2.0U N-Well (Lambda = 1.0 μm, Technology = SCNA).
Technology Setup File: mORBn20. tdb.

Estos datos son importantes para el envío a fabricación del circuito a través de MOSIS, medio que fue utilizado para tal fin. Cabe mencionar que el significado de esta palabra corresponde a: **MOS I**mplementation **S**ystem, la cual es una institución que proporciona el servicio de fabricación del circuito integrado a sus abonados, mediante su envío a fabricas de silicio establecidas en los Estados Unidos. La información referente a este servicio puede ser consultada por Internet en la siguiente dirección:

http://www.isi.edu/mosis/

donde se proporcionan los pasos a seguir para todas aquellas personas interesadas en este servicio. Existen otras opciones europeas como Alcatel Mietec, AMS, ATMEL-ES2 y CNM-España que también cuentan con sus propias reglas de diseño. Si se piensa en alguna de estas últimas como opción, se debe contar con los programas correspondientes.

El diseño de la BAM prototipo se ajustó a una de las variedades de encapsulado ofrecidas por Orbit, llamada **Tiny-Chip**, que consta de 40 terminales con un área de silicio de 2220 μm x 2250 μm con un encapsulado del tipo **40PC22X22 TINY CHIP**. Esta área permitió dar cabida al arreglo de 6X3, junto con el decodificador y tres motivos de prueba (dos inversores con compuerta flotante y uno sin compuerta flotante). De esta manera, todas las terminales fueron ocupadas a excepción de las cuatro esquinas, que no se utilizan, es decir, el número de terminales realmente usadas es de 36.

Con base en el diseño expuesto en el capítulo 3, el número total de transistores MOS usados en el circuito, se puede ver en la tabla 5.1. A este número de transistores, se deben agregar los de los tres motivos de prueba, que en total suman 6 transistores usados en los tres inversores.

Tabla 5.1. Transistores utilizados en la BAM.

Tipo de Celda	No. de transistores	Función
Celda NAND2C: 6 transistores x 18	108 transistores	Decodificador
Sinapsis: 2 transistores x 18 x 2	72 transistores	Elementos de aprendizaje
Interruptores: 2 transistores x 4 x 18 x 2	288 transistores	Selecc. de programación. o RNA.
Celda INV2: 4 transistores	4 transistores	Función Habilitar
Neuronas: 4 transistores x 9	36 transistores	Función de transferencia
	Total: 508 transistores	

Por lo tanto, se cuenta con 514 transistores MOS dentro del circuito integrado. Otro aspecto importante en el diseño topológico, se refiere al tipo de terminales empleadas (PADS), ya que existen varias opciones y la elección depende de la función que se desea que realicen. En este caso, se utilizaron cuatro tipos de terminales:

1. BlankPad.
2. BareIO Pad.
3. VDD Pad.
4. GND Pad.

La primera (BlankPad) cuenta con diodos de protección para las compuertas de los transistores MOS, cuyo óxido es muy sensible a descargas electrostáticas y se requiere proteger contra éstas. Las terminales BareIO, se usaron para alimentar a los inyectores de carga en los FGMOSFET, dado que existe la posibilidad de utilizar voltajes por encima del voltaje de ruptura de los diodos de protección; estas terminales no cuentan con diodos de protección y alimentan directamente el voltaje aplicado. Por último, las terminales VDD y GND se usan para conectar las fuentes de alimentación del circuito y polarizar a los diodos de protección.

5.1. Células básicas.

Las Figs. 5.1 y 5.2 muestran a la neurona y a la sinapsis, respectivamente, como quedaron finalmente en el diseño topológico.

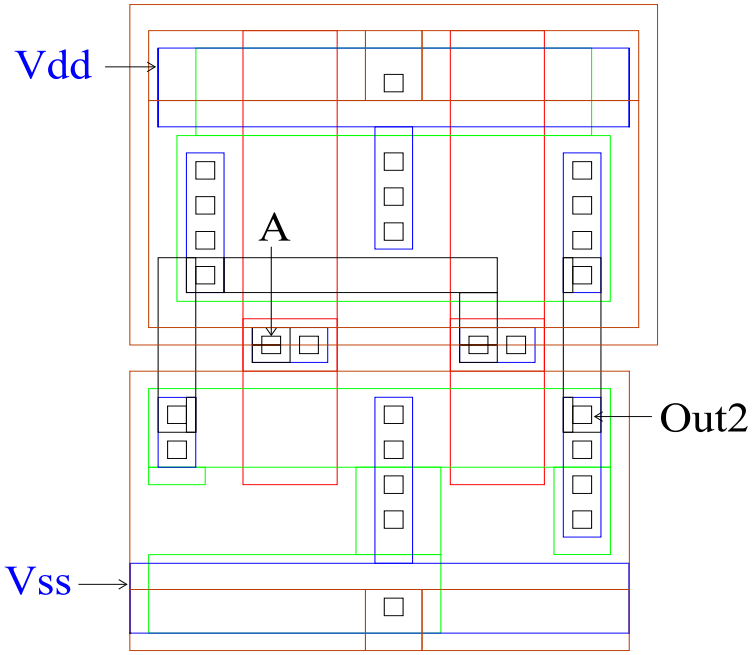


Fig. 5.1. Diseño topológico de la neurona.

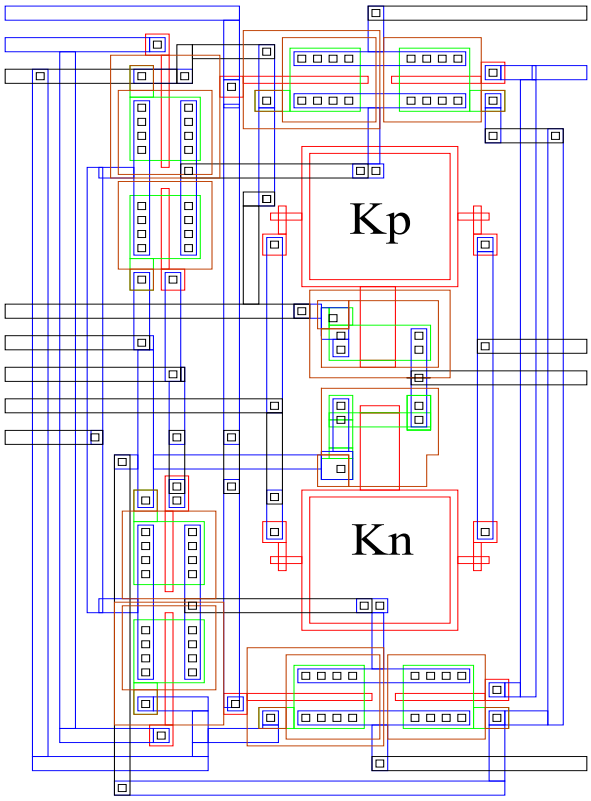


Fig. 5.2. Diseño topológico de la sinapsis.

En la Fig. 5.1, el nodo etiquetado con A corresponde a la entrada a la neurona, proveniente del arreglo matricial de las sinapsis y el nodo Out2 corresponde a la salida de la neurona. En la Fig. 5.2, se puede apreciar la forma como están conectados los inyectores del NMOS y del PMOS. Las compuertas etiquetadas como Kp y Kn se conectan a las compuertas de transmisión que rodean a la sinapsis y que se habilitarán según se desee programar a cada transistor, o se cortocircuiten para habilitar el funcionamiento de la red. Estas compuertas de transmisión o interruptores CMOS, son controlados por el decodificador.

5.2. Matrices sinapticas.

Las celdas básicas de la sinapsis y de la neurona, se pueden interconectar para configurar las matrices según se estableció en el capítulo 3 y se ilustra en la Fig. 3.7. Se muestra en la Fig. 5.3 una de las matrices que se requieren para lograr la bidireccionalidad típica de este tipo de redes. La otra matriz es idéntica pero sus salidas se conectan a la entrada de la otra capa de neuronas.

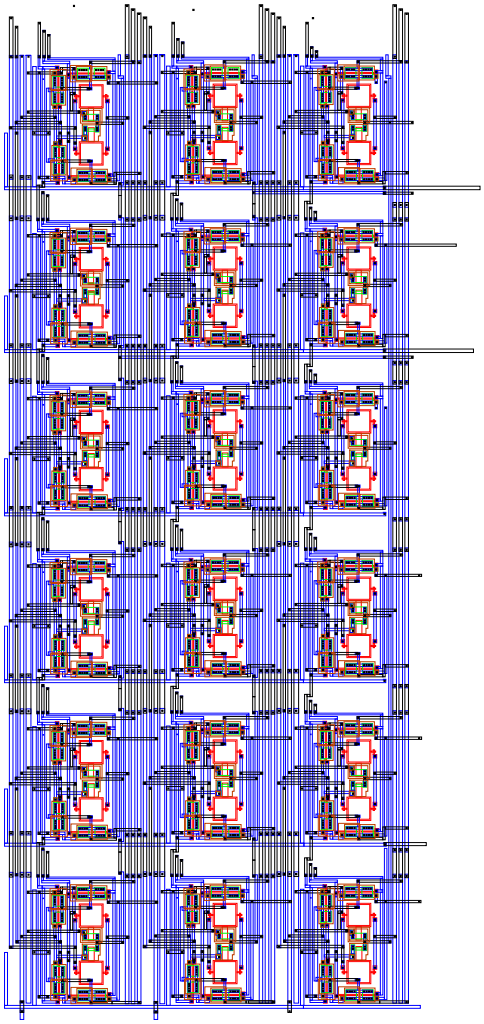


Fig. 5.3. Diseño topológico de la matriz de sinapsis.

5.3. Decodificador para la programación de la sinapsis.

El decodificador necesario para seleccionar cada transistor de las sinapsis de la BAM, de tal forma que se puedan programar independientemente, se presenta en la Fig. 5.4 en una sola hilera formada por tres celdas NAND2C y, en la Fig. 5.5, en todo su arreglo. La compuerta NAND2C, es una compuerta NAND de dos entradas con salida complementaria, y las tres líneas verticales en la parte inferior van directamente a los PADS de entrada, junto con la línea horizontal que sale hacia la izquierda. Las seis líneas que se dirigen hacia la derecha son las señales y sus complementos, que se dirigen a las compuertas de transmisión. Dependiendo de la polarización aplicada en los PADS, alguna de estas tres celdas cerrará al interruptor correspondiente al transistor que se desea programar.

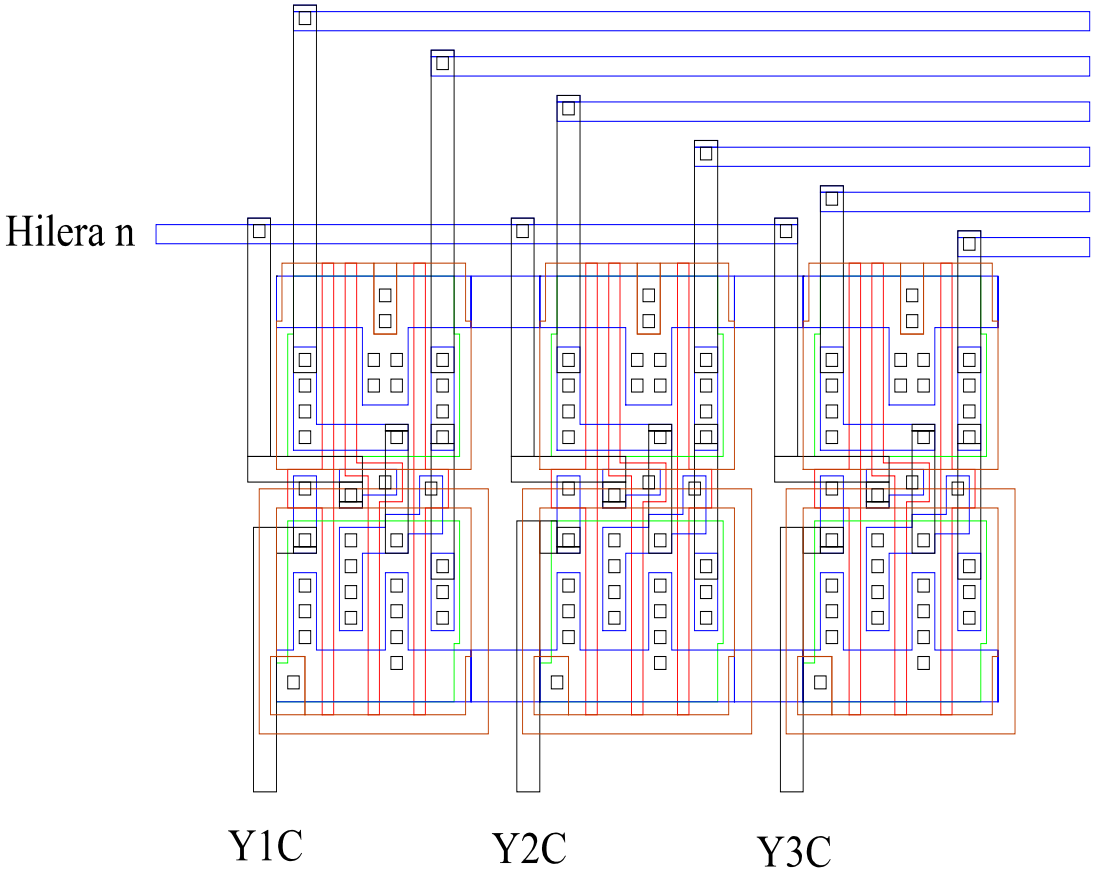


Fig. 5.4. Una hilera del decodificador.

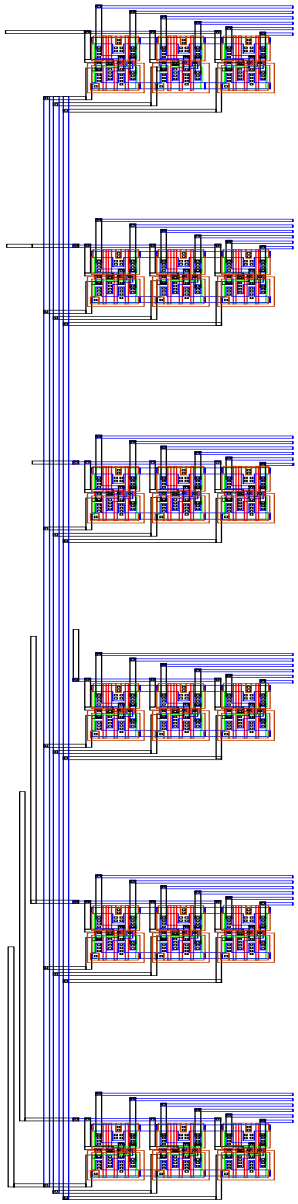


Fig. 5.5. Arreglo completo del decodificador.

Los voltajes que se deben aplicar a las entradas del decodificador, se indican en la tabla 3.2 del capítulo 3. Una vez que se hayan programado todos los transistores de la BAM, se deben poner las entradas Y1C, Y2C y Y3C a una polarización de + 5 V para inhabilitar los interruptores correspondientes.

5.4. Circuito de habilitación de la red.

Cuando se desea hacer funcionar a la red, se debe poner al decodificador como el mostrado en la Fig. 5.5, con las entradas a la polarización indicada anteriormente y habilitar los interruptores que conectan las

compuertas de los FGMOSFET a un mismo nodo. Esto se hace con un circuito habilitador, como el mostrado en la Fig. 5.6, que consiste simplemente en un doble inversor, con la salida del primer inversor conectada al mismo tiempo a la entrada del segundo inversor y a la compuerta complementaria de los interruptores CMOS; la salida del segundo inversor se conecta a la compuerta no complementaria del interruptor CMOS. La tabla 5.2 muestra la tabla de verdad de dicho doble inversor.

Tabla 5.2. Tabla de verdad del doble inversor.

A	B	Sal1	Sal2
0	0	1	1
0	1	1	0
1	0	0	1
1	1	0	0

De esta manera, cuando se están programando las sinapsis, la entrada A debe estar a - 5 V y cuando se quiera hacer funcionar la red, A debe estar a + 5 V.

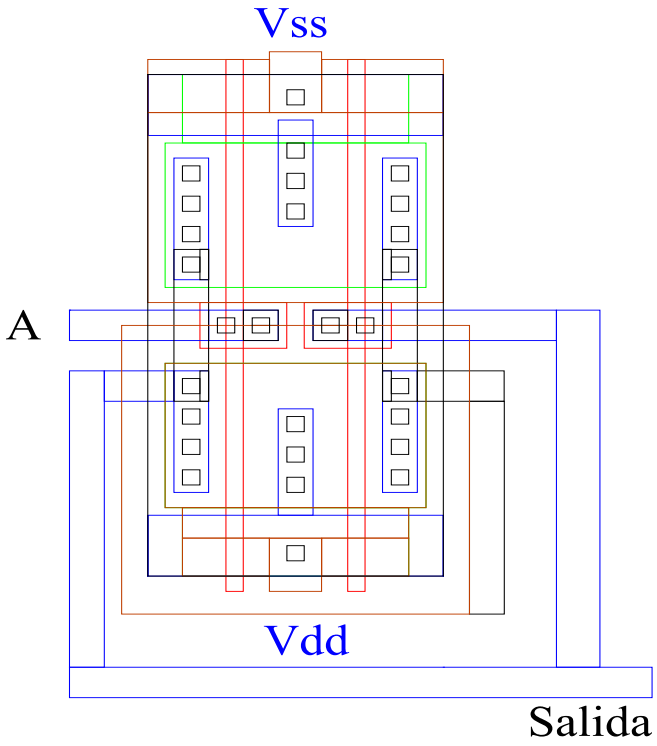


Fig. 5.6. Circuito habilitador de la BAM.

5.5. Circuitos de prueba.

Con el fin de tener la posibilidad de caracterizar los dispositivos de compuerta flotante, se incluyeron tres estructuras de prueba, como los mostrados en las Figs. 5.7 y 5.8. Dos de ellos consisten en inversores iguales con transistores de compuerta flotante y el tercero un inversor con las mismas dimensiones pero sin compuerta flotante. De esta manera, se puede comparar el comportamiento de uno con respecto al otro. Estos circuitos de prueba permiten experimentar en el proceso de programación y determinación del coeficiente de acoplamiento.

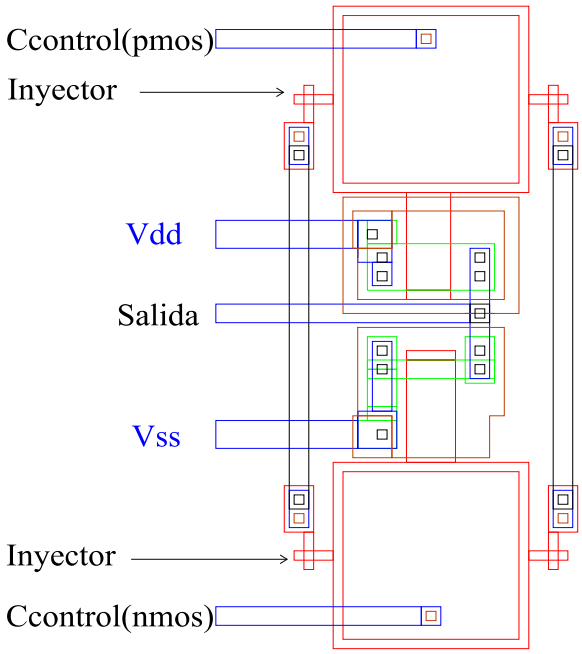


Fig. 5.7. Circuito de prueba. Inversor con compuerta flotante.

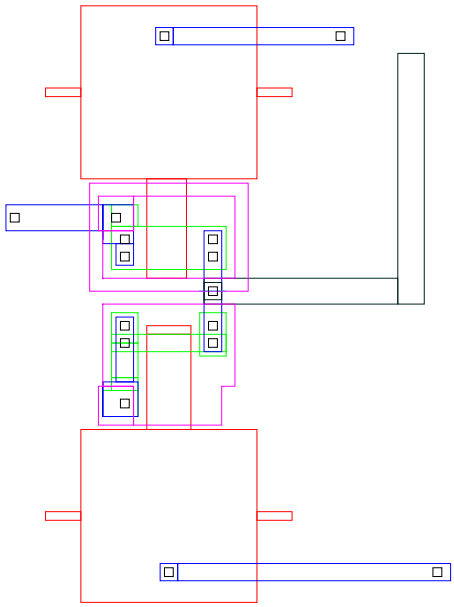


Fig. 5.8. Circuito de prueba. Inversor sin compuerta flotante.

Los circuitos de prueba anteriores, corresponden a una de las corridas que se fabricaron, con los que se hicieron las mediciones inhabilitando uno de los transistores del inversor para caracterizar al complementario, razón por la cual, como se verá en el siguiente capítulo, las corrientes de la curva de salida y de transconductancia de los transistores comienzan a incrementarse desde un voltaje de – 5 V y tienen su

máxima corriente en cero volts, para los PMOS y desde + 5 V hasta cero volts, para los NMOS (tanto normales como de compuerta flotante). Otra de las corridas que se fabricaron, incluyó únicamente transistores sencillos, con lo que se podía caracterizar a cada transistor independientemente. En la Fig. 5.9 se muestra un arreglo de transistores FGPMOS con dos inyectores y la fuente y substrato de todos los transistores, conectadas a un mismo pad, esto con el fin de disminuir el número de terminales ocupadas. Un arreglo similar, pero con FGNMOS, se muestra en la Fig. 5.10. Con estos transistores, las curvas que se obtienen, tanto de salida como de transconductancia, se presentan en ejes como normalmente se encuentran en la literatura, es decir, parten desde cero para tener su magnitud máxima de corriente en 5 V dependiendo si es un PMOS o un NMOS. (Esta observación es importante para poder interpretar correctamente las gráficas que se presentarán en el siguiente capítulo.)

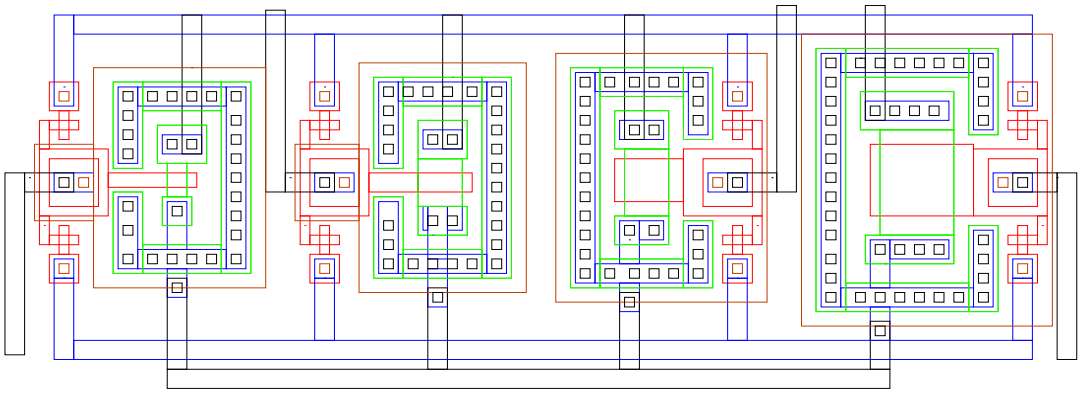


Fig. 5.9. Transistores FGPMOS independientes.

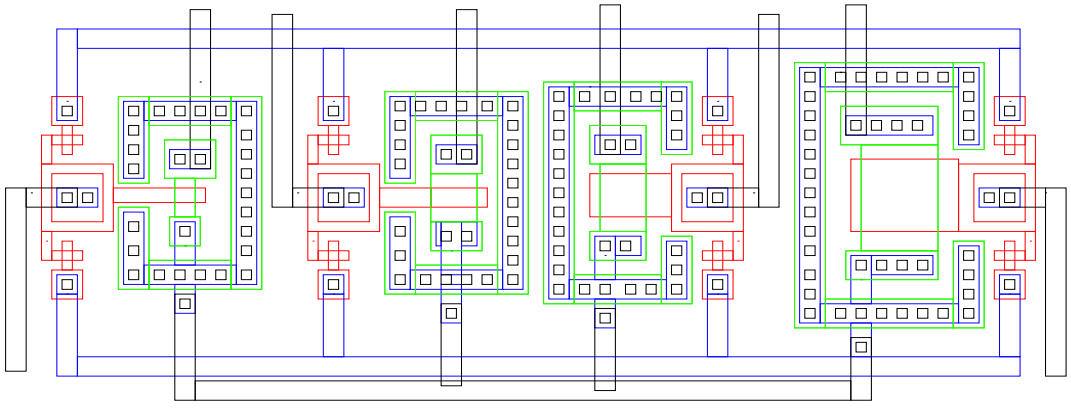


Fig. 5.10. Transistores FGNMOS independientes

5.6. Circuito integrado completo.

Al conjuntar todos los circuitos anteriores, se tiene finalmente el diseño topológico completo del circuito integrado, resultado del diseño presentado en la presente tesis y el cual se muestra en la Fig. 5.11.

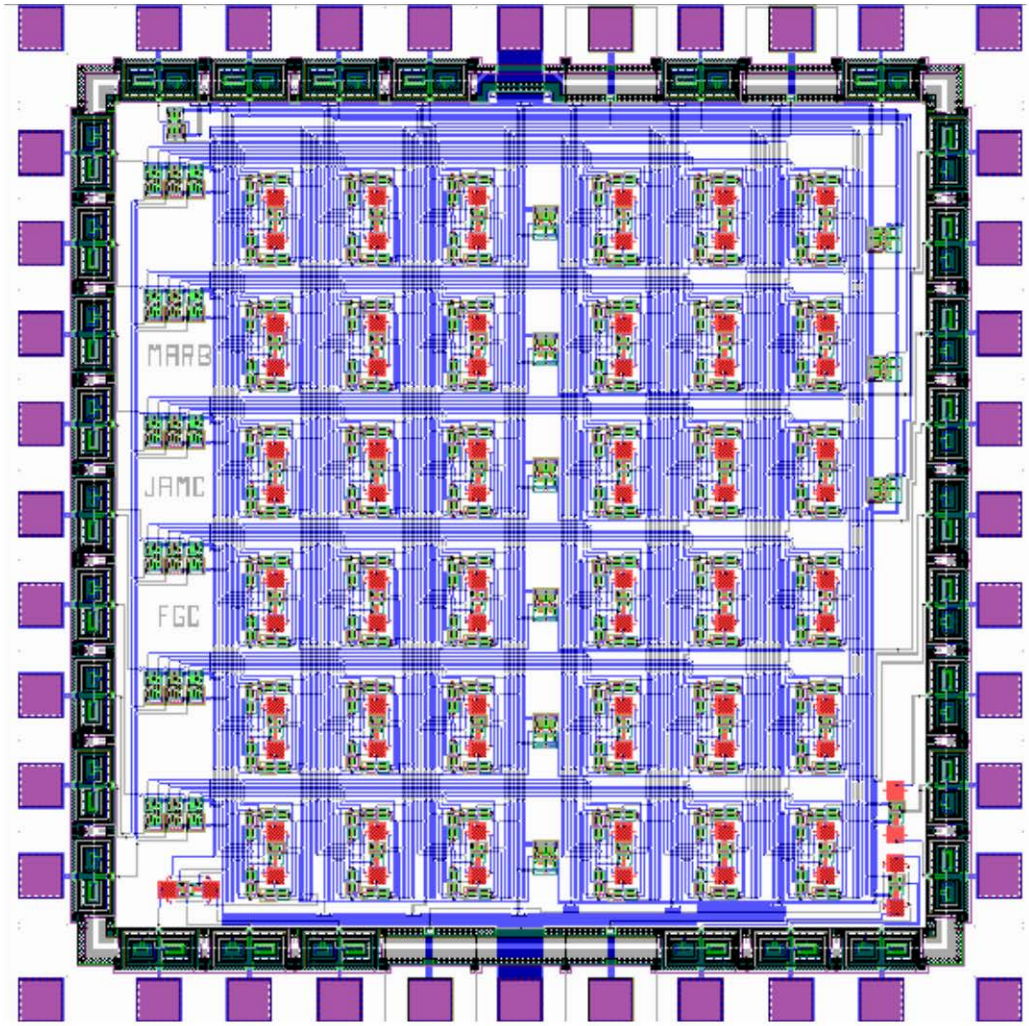


Fig. 5.11. Diseño topológico completo del circuito integrado.

En el apéndice se incluyen las hojas requeridas por MOSIS cuando se solicita la fabricación de un circuito. Estas hojas, llamadas “Forma de envío del proyecto” deben incluir el número de cuenta otorgado por MOSIS al cliente, además de los datos referentes al tipo de tecnología con la que se realizó el diseño, el tipo de encapsulado solicitado y datos referentes a la longitud del archivo (Byte Count y Checksum). Existen dos alternativas para el envío de esta forma junto con el diseño:

1. Mediante correo electrónico.
2. Correo normal.

Existen varias fechas a lo largo del año para el inicio de la corrida y el tiempo de fabricación puede cambiar entre las seis semanas y las catorce semanas, dependiendo de la época del año, de la tecnología solicitada y otros factores que no dependen del fabricante.

5.7. Sumario.

En este capítulo se presenta el diseño topológico de las celdas básicas como las neuronas, las sinapsis, el decodificador, el circuito habilitador y los circuitos de prueba, utilizando las reglas de diseño de la fábrica Orbit, para la tecnología de 2 μm, pozo N, doble polisilicio y doble metal. Esto se puede hacer por medio del servicio de MOSIS, institución que recaba la información de diferentes circuitos que se desean fabricar y los incluye en una sola corrida para disminuir el costo de fabricación.

PROJECT SUBMISSION FORM

1

Submitted by:

Name: ALFREDO REYES BARRANCA

Phone Number: (525)747-7000 Fax Number: (525)747-7114

Mailing Address: CINVESTAV-IPN, ELECTRICAL ENGINEERING DEPT.

P. O. BOX 14-740

MEXICO CITY 07000 D.F. MEXICO

Account Number: XXXXXXXXX

(must be included — call MOSIS at (310) 822–1511 if you're unsure of your account)

Purchase Order Number XXX/XX

(must be included if your project is going to be charged to a commercial account)

2

Indicate only **one** of the following submission technologies (not one from each category):

SCMOS	CMOSN	VENDOR
Lambda: 1.0	Lambda:	
SCP SCE SCNLC SCN3M (SCNA) SCN SCPE SCELC SCE3M SCEA SCNE SCEE	NSCN NSCP NSCE	ORBIT_CP HP_CMOS26B ORBIT_CN HP_CMOS34 HP_AMOSI VITESSE_HGAAS3 AMI_ABN

3

Project Specifications:

MOSIS Project ID (if known): Project Name: BAM

Project Description: PROTOTYPE FLOATING GATE CMOS CIRCUIT
(optional if commercial)

Number of Pads: 40

Project Size (in Microns): 2220 X 2250

Is this a TinyChip project? [X] Yes [ ] No

Before indicating the quantity of parts desired, refer to the MOSIS price schedule for price and quantity guidelines.

Number of packaged parts desired: 4

Number of unpackaged parts desired: 0

Total number of parts desired (indicate total quantity): 4

If any packaged parts are desired, indicate choice below:

- [ ] **Customer–Supplied Bonding Diagram (p. 3)** (not available on TinyChips)
- [ ] **Best Fit (calculated by MOSIS; diagrams shipped with packages)**
- [X] **MOSIS Standard Frame** (specify frame name): 40PX22X22

Check desired options:
- [ ] Substrate (not applicable to TinyChips)
- [ ] Hermetically Sealed Lids (not applicable to TinyChips)
- [ ] Foundry (specify:)

695wt

Call Sam Reynolds at (310) 822–1511 with any questions about Options.

4 **Tape Information:**

Record Length (default 2048 bytes/record): _____

Density (default 1600 BPI): _____

Data (Please check one):

☐ CIF ANSI ☐ CIF TAR ☐ CIF VMS_BACKUP ☐ GDSII STREAM ☐ MEBES PAM

5 **CIF Checksum:** 64429975 Byte Count: 1331251

6 **Top Structure / Library**

Top Structure Name (GDSII only): _____

Name of Referenced MOSIS Library (e.g., CMOSN_30A): _____

7 **Layer Map**

☐ MOSIS default ☐ Other *(Please note: No layer merging or Boolean operations available.)*

Layer Name	GDS #	CIF Name	Layer Name	GDS #	CIF Name
N_WELL	42	CWN	ELECTRODE_CONTACT	55	CCE
ACTIVE	43	CAA	ELECTRODE	56	CEL
P_PLUS_SELECT	44	CSP	METAL_1	49	CMF
N_PLUS_SELECT	45	CSN	VIA	50	CVA
POLY	46	CPG	METAL_2	51	CMS
POLY_CONTACT	47	CCP			
ACTIVE_CONTACT	48	CCA			

8 **Authorization:**

Authorized Signature: _____

Name (Please Print): _____

Date: _____

MAIL THIS FORM WITH YOUR OFFLINE SUBMISSION TO: **SAM REYNOLDS**
THE MOSIS SERVICE
4676 ADMIRALTY WAY
MARINA DEL REY, CA 90292–6695

Bonding Diagram Preparation Guidelines

MOSIS–SUPPLIED PACKAGES:

1. When selecting packages, you must allow for a cavity size that will accomodate the chip; 350 extra microns per side is usually adequate. However, if there is to be a downbond, allow an extra 200 microns on that side.

2. Don't 'fudge' in trying to make a layout fit; be certain of the pad placement. Packagers are confused by bonding diagrams that do not match the actual chip and package.

3. Cut and paste a (to scale) plot of the passivation layer of your chip in the space provided on the next page; be sure to indicate orientation if pad placement is symmetrical.

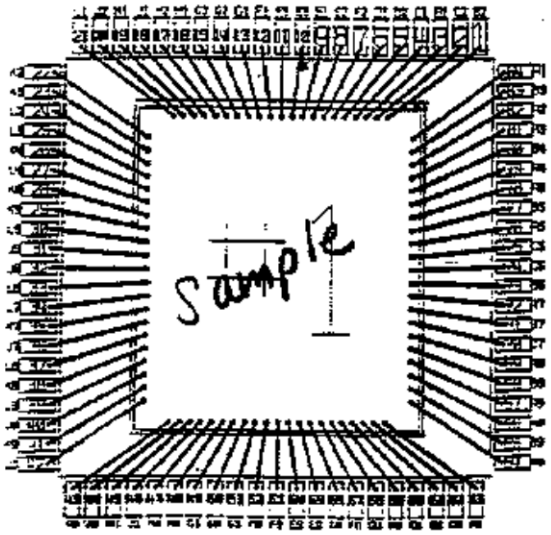
You MUST draw in your bonding wires with a straight edge. Curved lines are NOT allowed. Be sure to draw your chip to the scale of the package cavity.

4. If you have pads which are NOT bonded, please

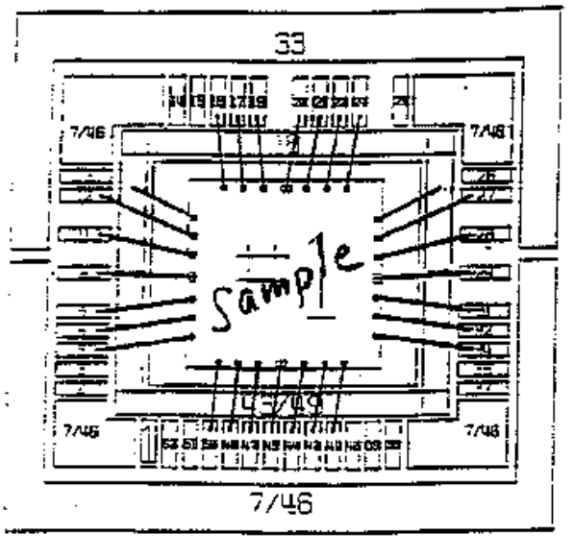
highlight them in some way

5. If you are FAXing your diagram, make sure the copy is EXTREMELY clean and well drawn. Detail can be easily lost, especially in packages with many pads.

6. MOSIS provides blank bonding diagrams for each package and cavity size. If you would like a template for any MOSIS package, contact Terry Dosek at (310) 822–1511.



PGA 84



LDCC 52

CUSTOMER–SUPPLIED PACKAGES:

1. If you need packages that we do not provide, you have the option of supplying your own. We cannot assist in the purchase of these packages, but we can provide the names of the suppliers used by MOSIS.

2. When supplying packages to MOSIS, you must also include package lids, bonding diagrams and package drawings. All of these items must be received by MOSIS one week prior to the scheduled run–closing date. Please ship them to the attention of Terry Dosek.

Customer–Supplied Bonding Diagram

Organization Name CINVESTAV-IPN Project Name BAM

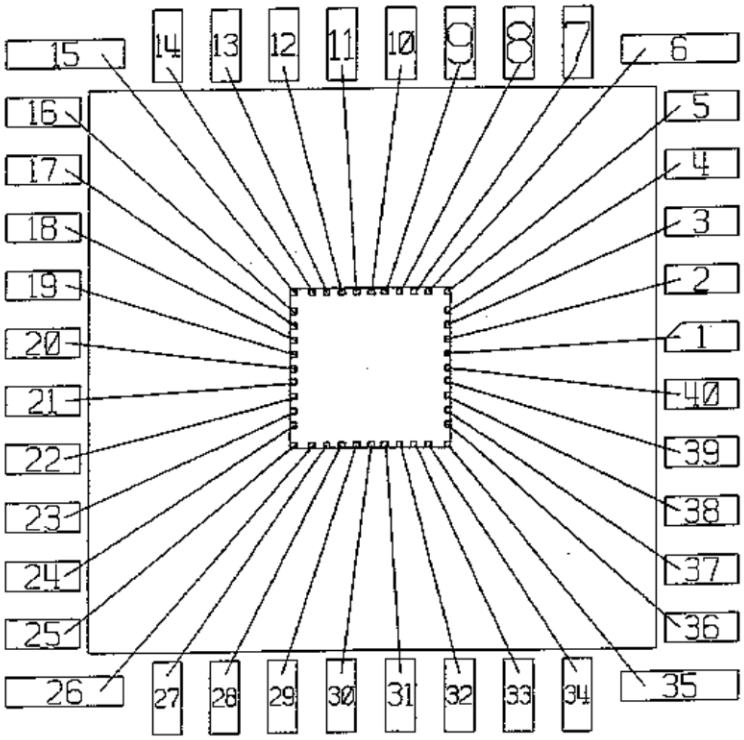
Designer Name ALFREDO REYES BARRANCA X _____ Y _____

MOSIS 2 MICRON (40pc22x22) TINY CHIP FRAME

P WELL AND N WELL RUNS

SUBMITTED PROJECT SIZE = 2220 BY 2250 MICRONS

PIN#	X	Y
1	2170	1325
2	2170	1525
3	2170	1725
4	2170	1925
5	2170	2200
6	1910	2200
7	1710	2200
8	1510	2200
9	1310	2200
10	1110	2200
11	910	2200
12	710	2200
13	510	2200
14	310	2200
15	50	2200
16	50	1925
17	50	1725
18	50	1525
19	50	1325
20	50	1125
21	50	925
22	50	725
23	50	525
24	50	325
25	50	50
26	310	50
27	510	50
28	710	50
29	910	50
30	1110	50
31	1310	50
32	1510	50
33	1710	50
34	1910	50
35	2170	50
36	2170	325
37	2170	525
38	2170	725
39	2170	925
40	2170	1125



<-- NOTE: X Y REFERS TO CENTERS OF 100 BY 100 MICRON PADS

Capítulo 6.

Resultados experimentales.

En el presente capítulo, se presenta la caracterización que se hizo a las estructuras de compuerta flotante en el régimen de Corriente Directa y la caracterización de la programación de las mismas, junto con la determinación del parámetro del coeficiente de acoplamiento, para finalmente presentar los resultados del funcionamiento de la estructura implementada como una sinapsis. Los montajes para la medición de las características, se trataron en el capítulo 4 y aquí se reportan los resultados de las mediciones realizadas.

Las mediciones se realizaron en un sistema de adquisición de datos maraca Keithley ® modelo 90, apoyado en el programa de adquisición y manejo de datos llamado Metrics ©. Este sistema consta de una unidad controladora de disparo modelo 2361, una fuente de voltaje Quad, modelo 213 y cuatro unidades de fuente/medición (SMU), modelo 236 conectadas a una computadora controladora, mediante una interfaz GPIB-488, desde donde con el uso del paquete Metrics, se configura el montaje de medición según las necesidades y se inicia la secuencia de adquisición de datos. Este programa incluye facilidades mediante las cuales se pueden extrapolar datos importantes a partir de las curvas de los dispositivos medidos. Para el caso del presente trabajo, el parámetro mas usado, es el voltaje de umbral (Vth), obtenido con el método tradicional de extrapolación al cruce con el eje de las ordenadas, de la recta de $Ids^{1/2}$ vs Vgs.

6.1. Transistores MOS sin compuerta flotante.

Como ya se había mencionado con anterioridad, la neurona se implementa con un inversor CMOS utilizando transistores sin compuerta flotante, dado que su tarea es la de incluir la función de transferencia de la red neuronal –una sigmoide en este caso- para lo cual no se requiere modificar el voltaje de umbral nativo de los transistores. Por lo tanto, estos transistores se pueden aprovechar para tener datos acerca de las características de este tipo de dispositivos, sobre todo de su voltaje de umbral. Este representará el voltaje a partir del cual se introduce o extrae carga a la compuerta flotante, como se analizó en el diseño de la sinapsis.

La Fig. 6.1(a) presenta las curvas de salida de un transistor PMOS y la Fig. 6.1(b) las de un transistor NMOS, de donde se pueden ver características típicas de corriente-voltaje presentadas por los dispositivos fabricados por ORBIT. En la Fig. 6.2(a), se presenta la gráfica de transconductancia del transistor PMOS. Como se puede observar, se tienen varias curvas de transconductancia ya que se utilizó el mismo montaje mencionado en la sección 4.3 del capítulo 4, para la obtención del coeficiente de

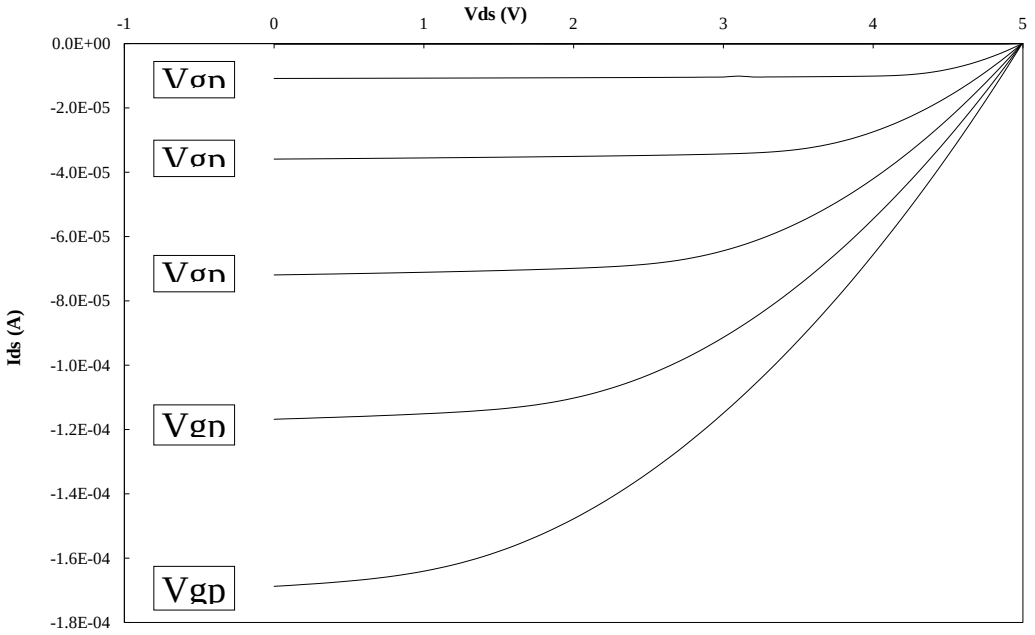
acoplamiento (Kcg), sin menoscabo del dispositivo que se va a medir (inhabilitando al transistor NMOS). Esto implica aplicar diferentes voltajes de Vdd (recordar que fuente/substrato del PMOS se conectan a Vdd, ver Fig. 4.10), en diferentes pasos y hacer un barrido en el voltaje de compuerta. La ventaja de hacer esto, radica en que se pueden encontrar tantos valores de voltaje de umbral, mediante la extrapolación de la recta de la gráfica de $Ids^{1/2}$ vs Vg, como curvas se tengan, para de esta manera, obtener un promedio. Recordando la fórmula (4.18), repetida aquí por conveniencia, pero considerando que en este caso el coeficiente de acoplamiento es unitario, dado que no existe compuerta flotante, se tiene que el voltaje de umbral para cada curva, se puede obtener de la siguiente manera:

$$Vth = Vgp - Vdd \quad , \tag{6.1}$$

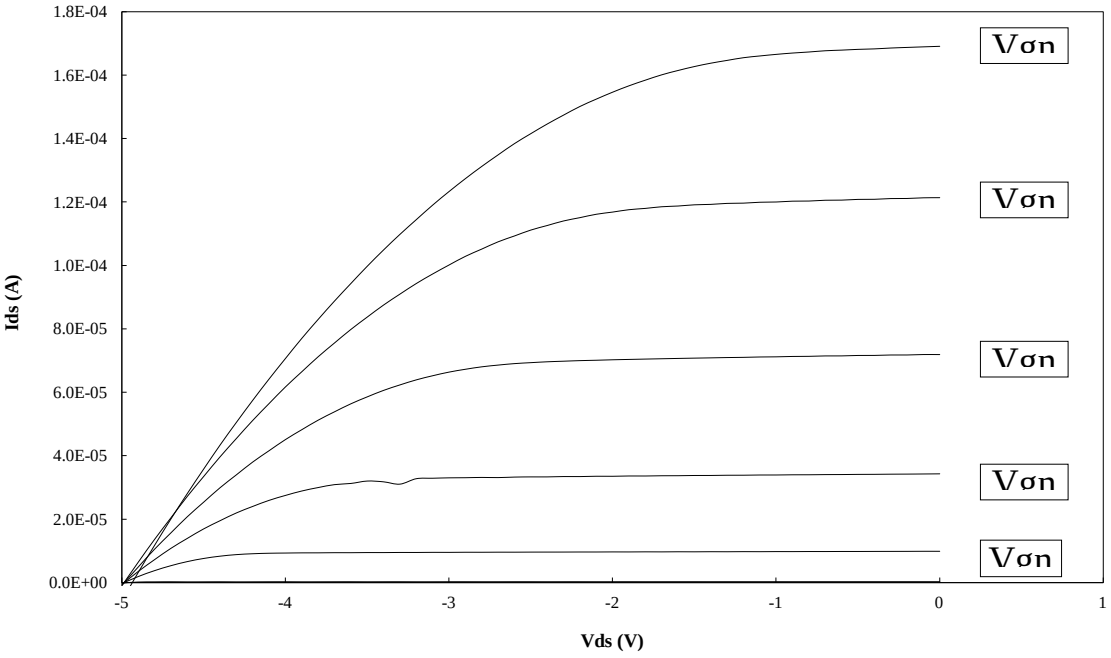
donde Vth es el voltaje de umbral que se desea conocer, Vgp es el voltaje extrapolado de la recta obtenida al graficar $Ids^{½}$ vs Vgs y Vdd son los diferentes pasos de voltaje aplicados a la fuente/substrato del arreglo. También se puede utilizar el montaje de la Fig. 4.10 para medir el transistor NMOS, pero ahora inhabilitando al transistor PMOS, con los resultados mostrados en la Fig. 6.2(b), encontrándose el respectivo voltaje de umbral con una fórmula similar, pero ahora cambiando los pasos de voltaje al Vss aplicado a la fuente/substrato del NMOS:

$$Vth = Vgn - Vss \quad , \tag{6.2}$$

y de la misma manera, se pueden obtener diferentes valores para promediar el voltaje de umbral.

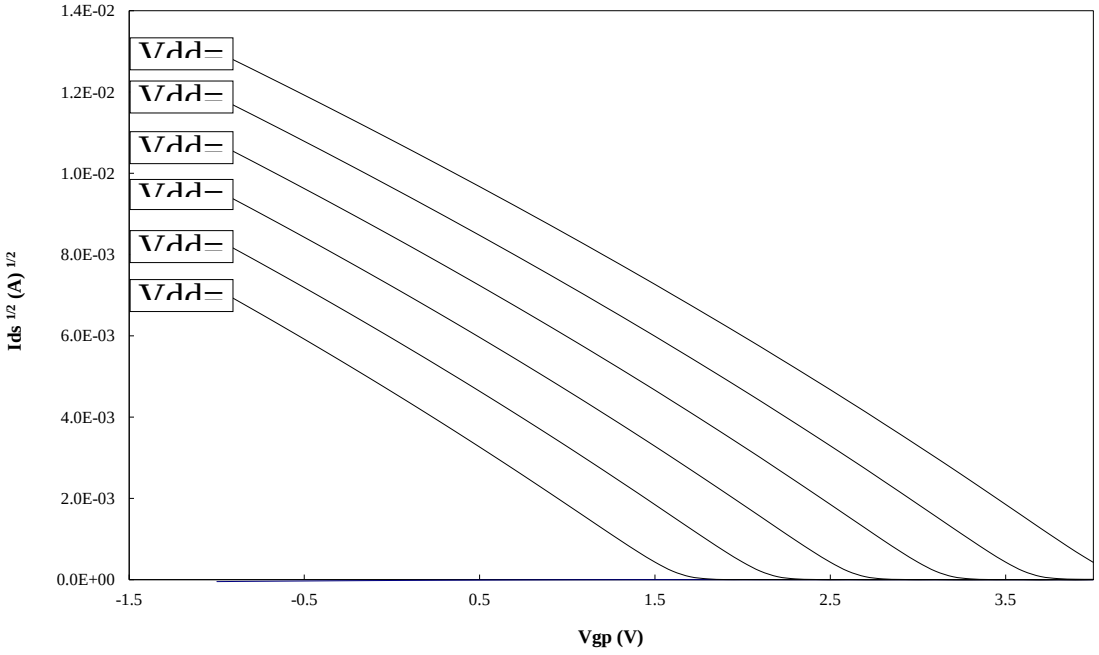


(a)

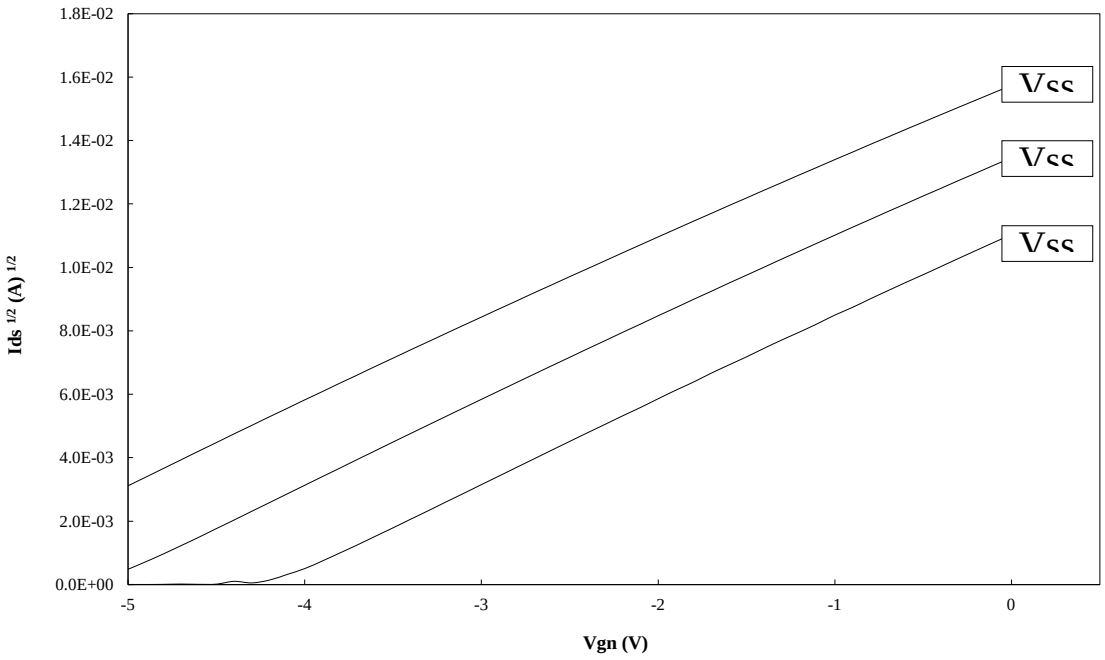


(b)

Fig. 6.1. Características de salida, a) de un transistor PMOS; b) de un transistor NMOS. Nota: los voltajes del eje X, se deben al montaje de la Fig. 4.10.



(a)



(b)

Fig. 6.2. Curva de transconductancia, a) de un transistor PMOS; b) de un transistor NMOS.

Para el transistor PMOS, se tiene la siguiente tabla, donde se muestra el voltaje de umbral nativo medido:

Vdd (V)	Vgp (V)	Vth (V)
5.0	4.0701	-0.9299
4.5	3.5766	-0.9234
4.0	3.0884	-0.9116
3.5	2.6202	-0.8798
3.0	2.1244	-0.8756
2.5	1.6268	-0.8732

$$Vth(PMOS) = -0.8989 \pm 0.023 \text{ V}$$

Y para el transistor NMOS, la tabla correspondiente, es la siguiente:

Vss (V)	Vgn (V)	Vth (V)
-5.0	-4.1800	0.8200
-6.0	-5.1718	0.8282
-7.0	-6.1710	0.8290
-5.0	-4.1567	0.8433
-6.0	-5.1471	0.8529
-7.0	-6.1447	0.8553
-5.0	-4.1292	0.8708
-6.0	-5.1510	0.8490
-7.0	-6.1503	0.8497

$$Vth(NMOS) = 0.8442 \pm 0.015 \text{ V}$$

Estos serán los voltajes de umbral que se tomarán como partida para la determinación de la carga inyectada o extraída en la compuerta flotante de los dispositivos que la contienen.

6.2. Transistores MOS con compuerta flotante.

También es conveniente una caracterización similar para los transistores que tienen compuerta flotante, tanto para ver su estado de funcionamiento, como para determinar si existe alguna carga presente en la compuerta flotante (polisilicio 2, para la tecnología usada). Esto último se debe a que se ha reportado en la literatura [1], [2], [3], que es normal que, de fábrica estos dispositivos tengan presente una carga, a pesar de que no hayan sido programados intencionalmente. Esto puede tener una consecuencia para el procedimiento de escritura/borrado de los FGMOSFET, ya que sería erróneo considerar que la compuerta flotante está libre de carga. De hecho, algunos investigadores para hacer su caracterización, comienzan con un borrado con luz ultravioleta para eliminar esta carga y comenzar sus mediciones en ese estado.

6.2.1. Estado inicial de los transistores fabricados en ORBIT.

Dado que esta carga no es introducida intencionalmente, su valor es relativamente aleatorio en cuanto a su valor, pero lo que si es consistente, es en cuanto a que se trata de una carga positiva. Esto se puede ver en la Fig. 6.3, donde se presenta la característica de salida de un transistor FGNMOS. En este caso, las líneas continuas corresponden a la respuesta del transistor medido por primera vez, es decir, con

las características que presenta de fábrica. Se puede notar que a pesar de que los voltajes de compuerta aplicados cubren un intervalo desde cero hasta 5 volts, su corriente de salida es muy alta y aún cuando el voltaje de drenador-fuente llega hasta los 5 volts, el dispositivo se encuentra en la región ohmica de operación; esto es un claro indicativo de que su voltaje de umbral se encuentra desplazado hacia voltajes negativos, lo que corresponde a contener una carga positiva en la compuerta flotante. En la misma figura se puede ver con línea punteada, la respuesta del transistor una vez inyectada carga negativa en la compuerta flotante, en tal magnitud, que en el mismo intervalo de voltaje de drenador-fuente, el dispositivo ya alcanza la saturación.

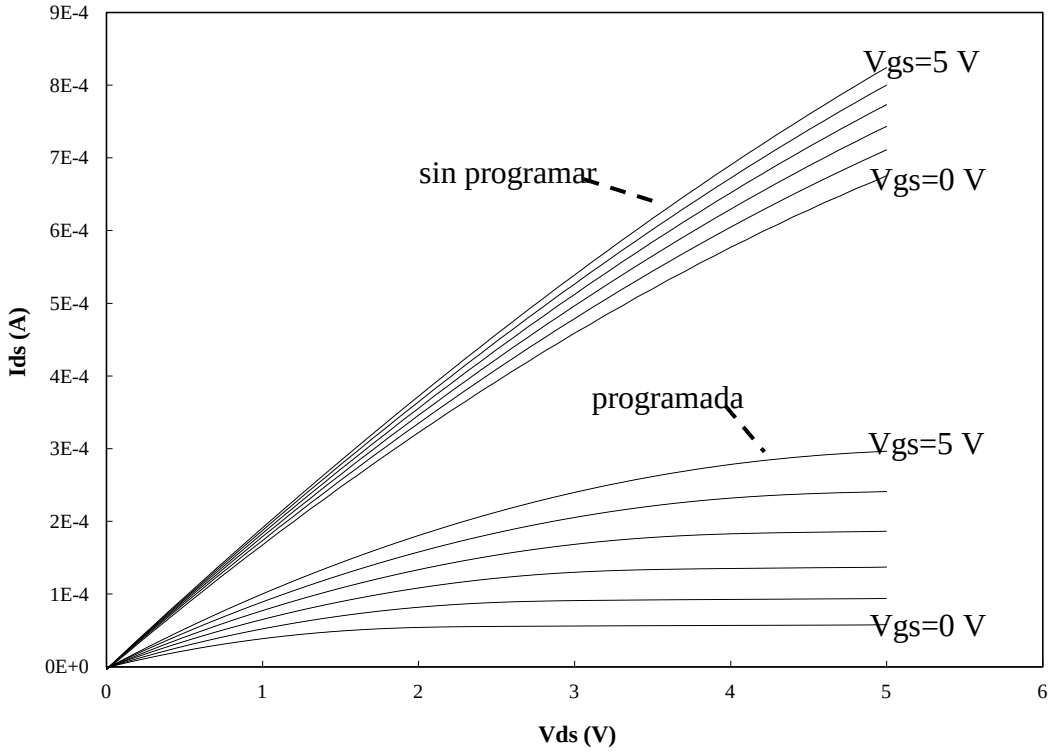


Fig. 6.3. Curva Ids vs Vds de un transistor FGNMOS, con las características presentadas de fábrica.

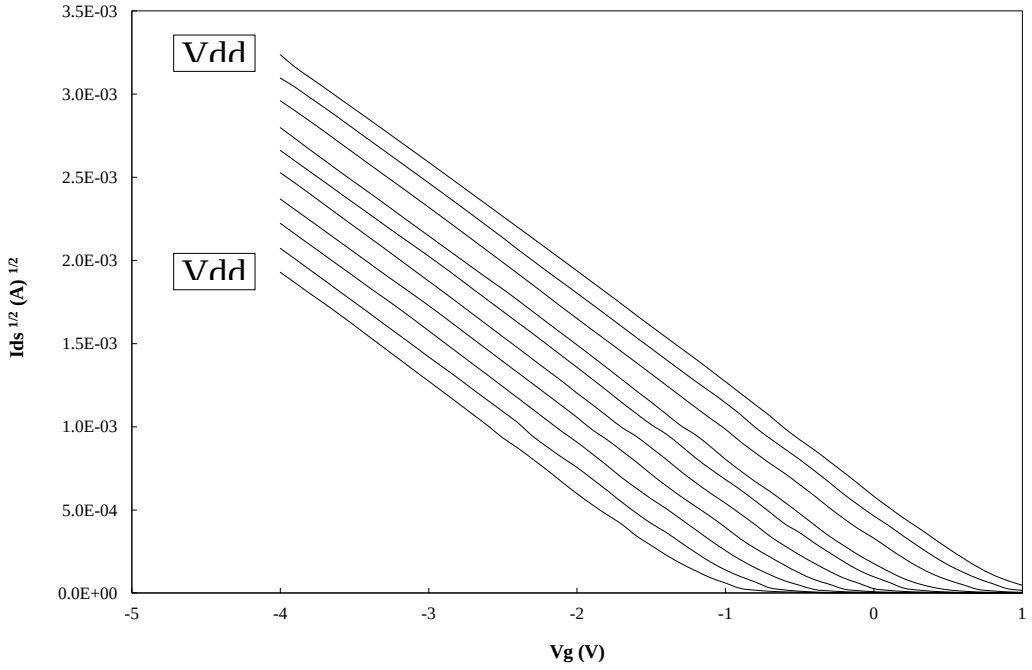
En mediciones realizadas a diferentes dispositivos, el valor de voltaje de umbral estaba por debajo de –3 volts, alcanzando en ocasiones valores cercanos a los –8 volts, valores muy alejados de los voltajes nativos de los transistores sin compuerta flotante determinados con anterioridad. Para lo presentado en la Fig. 6.3, sólo se intentó determinar el estado inicial de los transistores sin necesariamente obtener el voltaje de umbral. De la misma manera, con la única finalidad de observar su comportamiento en función de la programación (inyección o extracción de carga), se disminuyó el voltaje de umbral del FGNMOS aplicando un voltaje negativo en la compuerta de control de –5 V y un voltaje negativo arbitrario, en varios pasos, de tal forma que se obtuviera la característica marcada como *programada* en la Fig. 6.3.

6.2.2. Medición del coeficiente de acoplamiento y del voltaje del umbral.

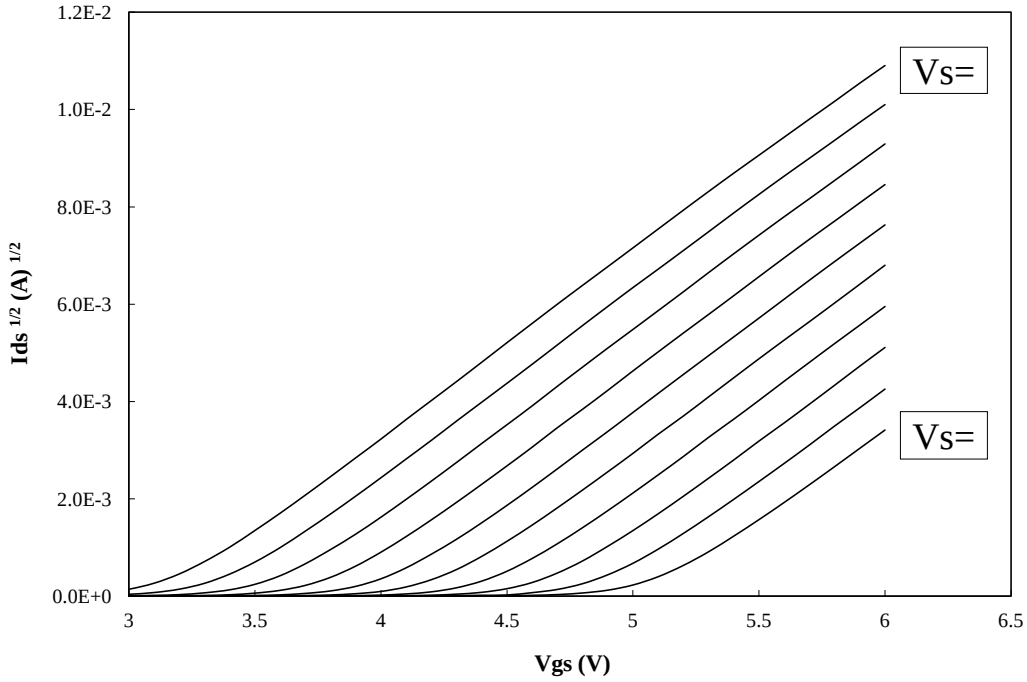
Mediante la técnica ilustrada en la sección 4.3 del capítulo 4, se midieron los coeficientes de acoplamiento de los transistores de la sinapsis y al mismo tiempo se pudo determinar el voltaje de umbral real en los dispositivos. Según se mencionó en la sección 4.1.2, el diseño realizado a los transistores arrojaba como resultado los siguientes valores para Kcg:

$$Kcg(NMOS) = 0.94$$
$$Kcg(PMOS) = 0.89$$

En la Fig. 6.4(a) se pueden apreciar las curvas típicas de este tipo de caracterización, para un transistor FGPMOS y en la Fig. 6.4(b) las correspondientes al transistor FGNMOS, lo cual confirma lo deducido en el capítulo 4, siendo este un método no reportado en la literatura y por lo mismo, original, que además facilita la medición de este parámetro, de bastante importancia en el funcionamiento de este tipo de dispositivos.



(a)



(b)

Fig. 6.4. Curvas experimentales para la determinación del coeficiente de acoplamiento, Kcg, y el voltaje de umbral, Vth, para a) un transistor FGPMOS y b) un transistor FGNMOS.

Tomando la lectura de la extrapolación de la recta hacia el eje de Vg, para curvas similares a las presentadas en las figuras anteriores, se pueden evaluar los valores de Kcg y Vth mediante el uso de las ecuaciones (4.18) y (4.19), para ambos transistores. Las tablas VI.1, VI.2, VI.3 y VI.4 muestran un ejemplo de valores obtenidos, a partir de los cuales se deducen los valores buscados.

Tablas VI.1, VI.2, VI.3 y VI.4. Cálculo de Kcg a partir de la gráfica de Id vs Vg.

Tabla VI.1.

Transistor FGNMOS	
Vs (V)	Vg (V)
-3.0	6.60
-2.5	6.08
-2.0	5.56
-1.5	5.04
-1.0	4.52
-0.5	4.00
Kcg(real)=0.962	
Kcg(diseñado)=0.94	
Vth=-3.346 V	

Tabla VI.2.

Transistor FGPMOS	
Vs (V)	Vg (V)
5.0	6.17
4.5	5.58
4.0	5.00
3.5	4.41
3.0	3.82
2.5	3.24
Kcg(real)=0.853	
Kcg(diseñado)=0.89	
Vth=-0.262 V	

Tabla VI.3.

Transistor FGPMOS	
Vs (V)	Vg (V)
5.0	2.8
4.5	2.2
4.0	1.6
3.5	1.0
3.0	0.4
2.5	-0.1
Kcg(real)=0.82	
Kcg(diseñado)=0.89	
Vth=-2.649 V	

Tabla VI.4.

Transistor FGPMOS	
Vs (V)	Vg (V)
5.0	2.72
4.5	2.14
4.0	1.55
3.5	0.97
3.0	0.38
2.5	-0.21
Kcg(real)=0.853	
Kcg(diseñado)=0.89	
Vth=-2.675 V	

En general, para las mediciones hechas del coeficiente de acoplamiento, Kcg, en diferentes estructuras y circuitos integrados, el promedio de este parámetro es: Kcg(N) = 0.8585 ± 0.016 para el transistor FGNMOS, y para el FGPMOS es de Kcg(P) = 0.9562 ± 0.024. De estos valores, se tiene que el error para Kcg(N) es de 4.7 % y para el Kcg(P) es de 1.7 %. Las capacitancias parásitas pueden ser las responsables de la desviación que se tiene, dado que es de esperarse que los traslapamientos y las pistas de conexión contribuyan con capacitancias no incluidas en el modelo. Sin embargo, el método demuestra ser práctico y de gran utilidad para ser usado convenientemente en la caracterización de los dispositivos.

Cabe mencionar en este momento, que otra forma de calcular el coeficiente de acoplamiento, aparte de las ecuaciones (4.18) y (4.19), se tiene con las siguientes fórmulas:

$$\mathrm{Kcg(NMOS)} = \frac{\left|\mathrm{Vss}_2 - \mathrm{Vss}_1\right|}{\left|\mathrm{Vgn}_2 - \mathrm{Vgn}_1\right|} \quad , \tag{6.3}$$

$$\mathrm{Kcg(PMOS)} = \frac{\left|\mathrm{Vdd}_2 - \mathrm{Vdd}_1\right|}{\left|\mathrm{Vgp}_2 - \mathrm{Vgp}_1\right|} \quad . \tag{6.4}$$

Mediante estas fórmulas, el coeficiente de acoplamiento se puede calcular directamente de la gráfica, trazando una recta horizontal a un nivel arbitrario de corriente, que cruce todas las curvas en su parte lineal, como se observa en la Fig. 6.5. Se eligen dos curvas cualquiera, cuyo valor de Vdd o Vss se conoce, y se lee verticalmente el voltaje Vg correspondiente al cruce entre la recta trazada y la curva medida. Se puede elegir el par de curvas que se desee y de esa manera se obtienen las parejas (Vdd, Vgp) para el transistor FGPMOS, o (Vss,Vgn) para el transistor FGNMOS. Cuando no se requiere conocer el voltaje de umbral, este método es bastante práctico a su vez.

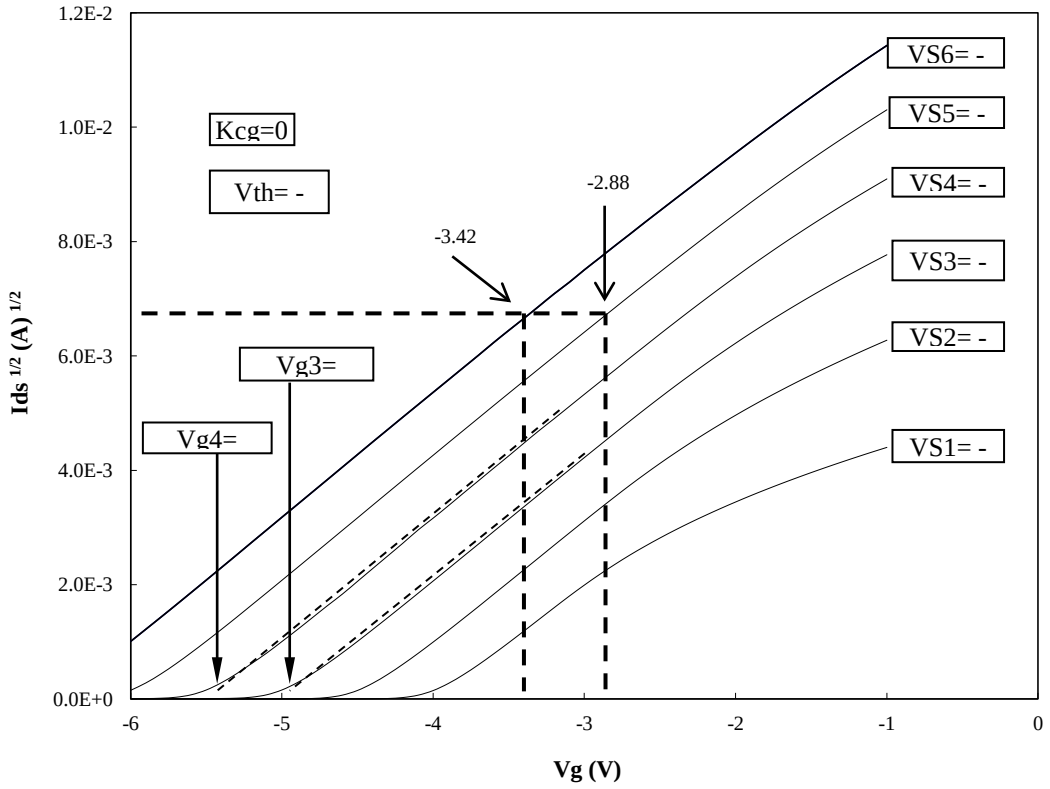


Fig. 6.5. Método alternativo para la determinación de Kcg.

6.3. Programación de los transistores de compuerta flotante.

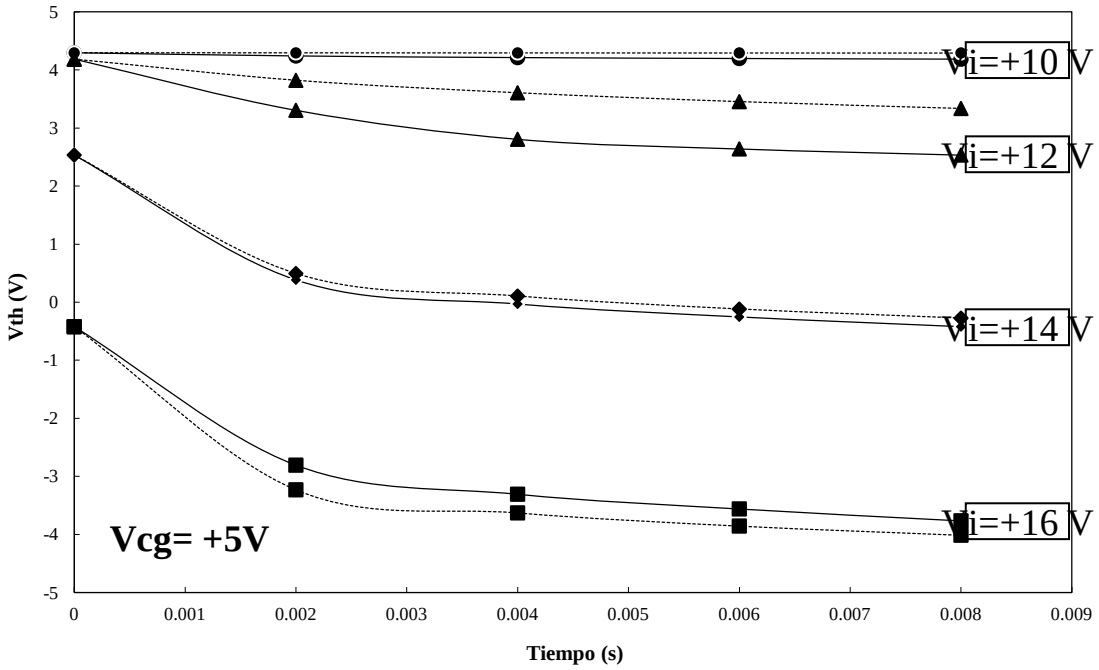
El cambio del voltaje de umbral se tiene que hacer de manera controlada, por lo que es indispensable conocer el comportamiento de los transistores de compuerta flotante cuando se les aplica un voltaje entre la compuerta de control y el inyector. La inyección o extracción de carga en la compuerta flotante se hace mediante el fenómeno de tunelamiento Fowler-Nordheim, como se explicó anteriormente. Dependiendo de la diferencia de potencial existente entre la compuerta flotante y el inyector, se tendrá cierta magnitud de carga fluyendo a través del óxido de tunelamiento: entre mayor sea la diferencia de potencial, mayor será la carga que atraviesa el óxido. Como se demostró con la ecuación (2.40) de la sección 2.3.2, puede existir un voltaje en la compuerta flotante aún cuando no se esté aplicando un voltaje en la compuerta de control, consecuencia de alguna carga presente en la compuerta flotante (debido al término Qfg/Ctot). Esto puede influir al intentar inyectar o extraer carga, ya que para un mismo voltaje de inyector, la diferencia de potencial entre el óxido de tunelamiento dependerá de esta carga presente de antemano en la compuerta flotante.

Se recuerda aquí, que en el análisis hecho en la sección 2.3.2 para el tunelamiento, se consideró que voltajes de la misma polaridad se aplicaban tanto al inyector como a la compuerta de control. Por lo tanto, la estrategia seguida fue la misma durante la caracterización de las estructuras. Además, durante el procedimiento, se aplicó un voltaje constante a la compuerta de control, mientras que al inyector se le aplicaban pulsos de magnitud y ancho específicos.

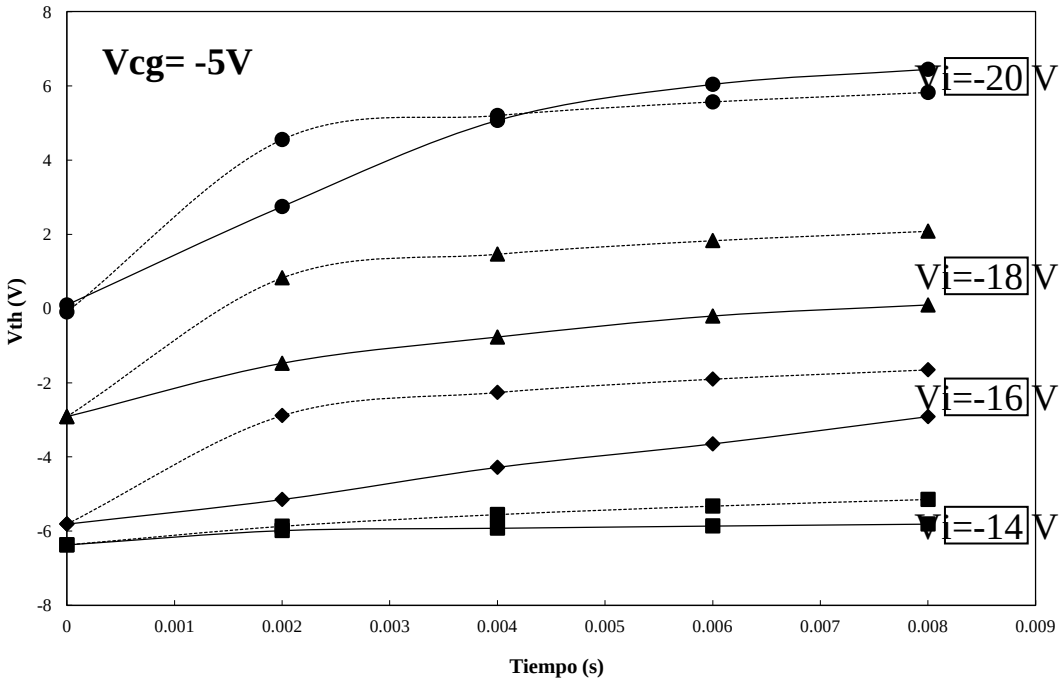
En la Fig. 6.6(a) se presentan las características de borrado (extracción de carga negativa) sobre la estructura. Partiendo de un voltaje de umbral de 4.23 V en la estructura (el comportamiento es igual tanto en transistores PMOS como para NMOS), se comenzó aplicando un voltaje constante de +5 V en la compuerta de control y posteriormente se aplicó un pulso de +10 V de magnitud y 2 ms de ancho, midiendo el voltaje de umbral después de cada pulso, hasta completar cuatro pulsos, es decir, 8 ms en total. Después de cada 8 ms, se mantenía el voltaje de compuerta y se aumentaba el voltaje del inyector, considerando el voltaje de umbral en t=0, a aquel que se medía al terminar el ciclo de cuatro pulsos con un voltaje de inyector dado. Estos voltajes fueron 10, 12, 14 y 16 V, como se indica en la Fig. 6.6(a), con sus respectivas respuestas.

Se puede apreciar de esta gráfica, que la diferencia de potencial a través del óxido de tunelamiento para el voltaje de umbral inicial usado (4.23 V) y con un voltaje de inyector de +10 V, era insuficiente para provocar extracción de carga de la compuerta flotante, pero a medida que se aumentaba este voltaje, la extracción era más pronunciada, llegándose a tener un cambio de voltaje de umbral de cerca de 8 volts, ya que con +16 V aplicados en el inyector, se llegó a tener un voltaje de umbral de –4 V. Esto permite tener un intervalo bastante amplio de voltaje de umbral con voltajes de inyector relativamente bajos.

El mismo procedimiento se aplicó a la escritura de la estructura (inyección de carga negativa), con los resultados mostrados en la Fig. 6.6(b). En este caso, los voltajes aplicados tanto al inyector como a la compuerta flotante, fueron negativos. El voltaje de umbral de partida fue de –6.4 V y se puede apreciar que los voltajes requeridos para poder inyectar carga son mayores que los necesarios para extraer carga, ya que con –14 V se logra una ligera variación del voltaje de umbral inicial y se requieren alrededor de –19 V en el inyector para tener una variación de cerca de 8 volts en el voltaje de umbral. Esto es consistente con lo analizado en la sección 4.2.1, donde se dedujeron los coeficientes característicos del tunelamiento Fowler-Nordheim (α y β), a partir de los datos reportados en el uso de una tecnología similar a la que aquí se emplea [4] (Orbit, pozo n, 2 μm, doble polisilicio, doble metal). De este análisis resultaba que era necesario aplicar un voltaje mayor para inyectar carga negativa, que para extraerla.



(a)



(b)

Fig. 6.6. Programación de estructuras de compuerta flotante, a) borrado y b) escritura.

En la Fig. 6.6(a), se muestran con línea continua los resultados experimentales y con línea punteada los resultados de una modelado de la programación. Este modelo se muestra en el apéndice del capítulo como un listado para usarse en el programa MathCAD, simulando la excitación constante para la compuerta de control y pulsos en el inyector, donde se utiliza la ecuación de densidad de corriente de tunelamiento Fowler-Nordheim con las siguientes constantes:

$$\alpha = 1.394 \times 10^{-1} \ \frac{A}{V^2}$$

$$\beta = 392 \ V$$

que fueron con las que se ajustaron las curvas teóricas, correspondiendo por lo tanto, a las constantes características de la corrida fabricada para este trabajo, cuando se extrae carga de la compuerta flotante (Qfg > 0). Se puede ver que el ajuste teórico se aproxima al experimental para +14 V y +16 V, pero existe cierta desviación con +12 V, lo que se puede justificar al recordar que el modelo FN ajusta a un comportamiento perfectamente exponencial y sin embargo, en la práctica esto tiene sus limitaciones. A pesar de eso, el ajuste se puede considerar como aceptable.

El mismo ajuste se realizó para la inyección de carga negativa (Qfg < 0) -línea punteada en la Fig. 6.6(b)- donde se aprecia que el ajuste no es lo suficientemente bueno. Las constantes empleadas en el modelo fueron:

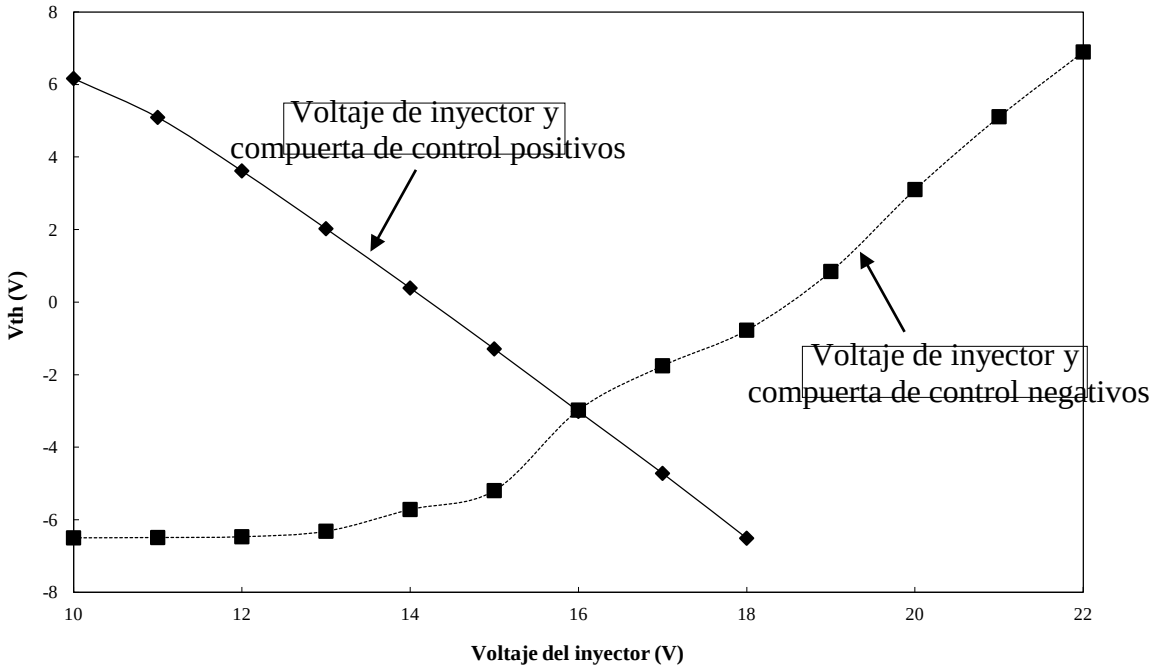
$$\alpha = 5 \times 10^{-4} \ \frac{A}{V^2}$$

$$\beta = 457 \ V$$

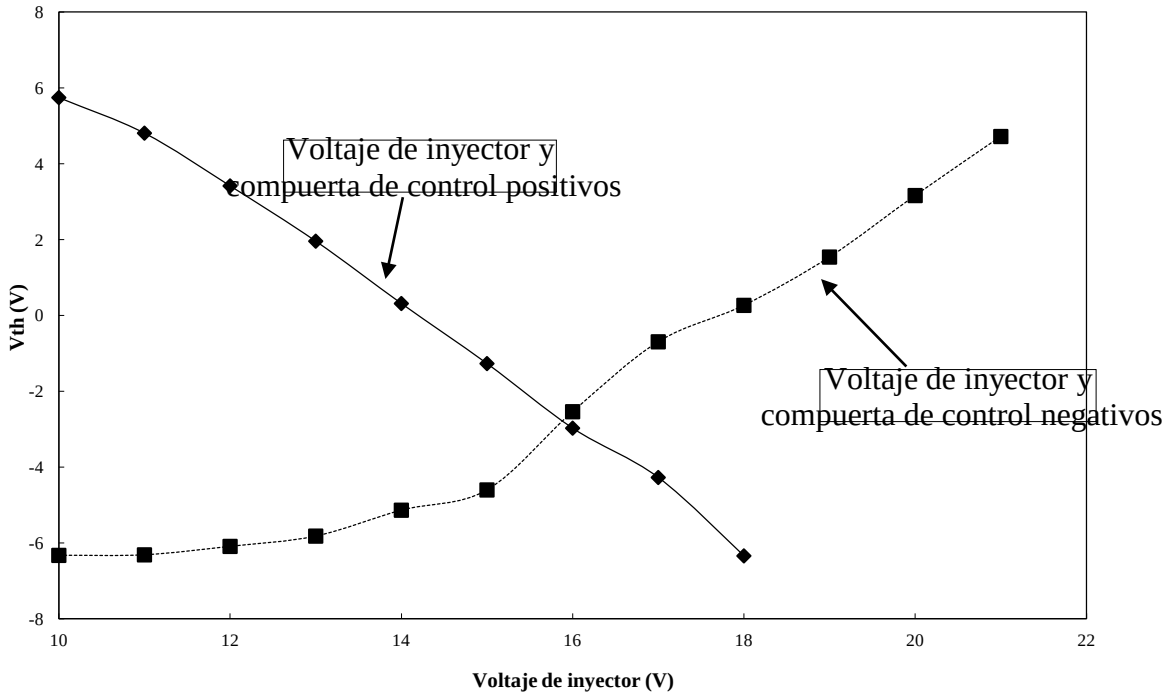
La razón de esta diferencia, sin embargo, se puede considerar como normal dado que las superficies emisoras de carga son relativamente de diferente naturaleza, como se mencionó anteriormente, y dado el hecho de que durante la extracción de carga, el polisilicio 1 es el que funciona como emisor de carga, mientras que durante la inyección de carga, el polisilicio 2 es el que lo hace, una superficie emitirá en base a unas constantes diferentes a las que modelan la inyección de la otra superficie. Esto explica en la práctica la relativa falta de simetría entre la escritura y el borrado y que se refleja en la diferencia entre las constantes α y β, las cuales en teoría se asumen iguales, independientemente del sentido de inyección, ya que se consideran superficies iguales.

Una comprobación de lo anterior, se puede ver en la Fig. 6.7(a) y (b), donde se grafican dos ventanas de programación (Vth en función del voltaje en el inyector) para estructuras distintas, al aplicar un pulso de un mismo ancho (2 ms) y voltaje de inyector variable (el voltaje de umbral se mide después de cada pulso). Se puede ver de estas gráficas, cómo al extraer carga negativa (disminución del Vth), se tiene una reducción monotónica y uniforme, mientras que para la inyección de carga negativa (aumento del Vth), ésta se presenta de una manera irregular. Esto quiere decir, que la superficie del polisilicio 1 es mas regular que la superficie del polisilicio 2.

Finalmente, vale la pena comentar que se logra tener un intervalo de voltaje de umbral de alrededor de 12 V, sin que la estructura se dañe, con voltajes de inyector de +18 V y –22 V, los cuales están por debajo de los aplicados a otras estructuras, que llegan a ser de ± 30 V, aproximadamente.



(a)



(b)

Fig. 6.7. Ventana de programación de las estructuras de compuerta flotante.

6.4. Mediciones sobre la sinapsis.

Como se había mencionado anteriormente, se requiere una configuración con la cual se pudiera cambiar la conductancia a la entrada de la neurona, para almacenar los pesos requeridos de tal forma que se pudiera ejecutar una tarea en particular con la red neuronal. La configuración propuesta es la que se mostró en la Fig. 3.4 del capítulo3, con la que mediante la aplicación de voltaje entre la compuerta de control y el inyector, se inyecta o extrae carga para cambiar el voltaje de umbral de los dispositivos. Las curvas esperadas como respuesta de la sinapsis, se mostraron en la Fig. 4.3 del capítulo 4.

La Fig. 6.8 muestra la caracterización de la sinapsis con transistores de compuerta flotante, a los cuales se les hizo cambiar el voltaje de umbral, según los valores mostrados en la inserción de la figura. En este caso, se puede apreciar la diferencia entre la pendiente presentada en esta figura con respecto a la Fig. 4.3, lo cual es debido al sentido de la lectura de la corriente, es decir, en PSpice la lectura de la corriente se hizo en sentido contrario a como se hizo con el sistema Keithley. Sin embargo, la respuesta es la deseada para la sinapsis que se pretende emplear. Cabe comentar que los voltajes de umbral se deben ajustar apropiadamente, para que todas las curvas de la sinapsis crucen por el origen (0,0), ya que de no ser así, se tendría un desplazamiento en las curvas (offset) que podría inducir a una interpretación errónea de la

corriente en la sinapsis, por ejemplo, comportarse como una corriente excitatoria cuando en realidad es inhibitoria, o viceversa. El procedimiento para lograr las curvas de la Fig. 6.8, fue de manera iterativa, de tal manera que se encontrara un voltaje de umbral para cada uno de los transistores con los que la curva fuera lo suficientemente simétrica y cruzara por (0,0), siendo esta la curva de partida para cambiar los voltajes de umbral con los que se generaran las demás curvas. La programación de los dispositivos se hizo independientemente, aplicando voltajes positivos o negativos tanto al inyector como a la compuerta de control, según se requiriera disminuir o aumentar el voltaje de umbral, respectivamente.

En la Fig. 6.9 se tiene otro conjunto de curvas, donde se tienen otros voltajes de umbral en los dispositivos, pero aún así se logra tener el comportamiento requerido. Los niveles de corriente obtenidos (alrededor de 15 a 20 μA), se deben a las resistencias de carga empleadas en la medición, cuyo valor era de 10 kΩ , pero la corriente puede ser mayor o menor, dependiendo de la carga que se tenga. Con esto se demuestra la funcionalidad de la configuración de la sinapsis, mediante el inversor CMOS con dispositivos de compuerta flotante.

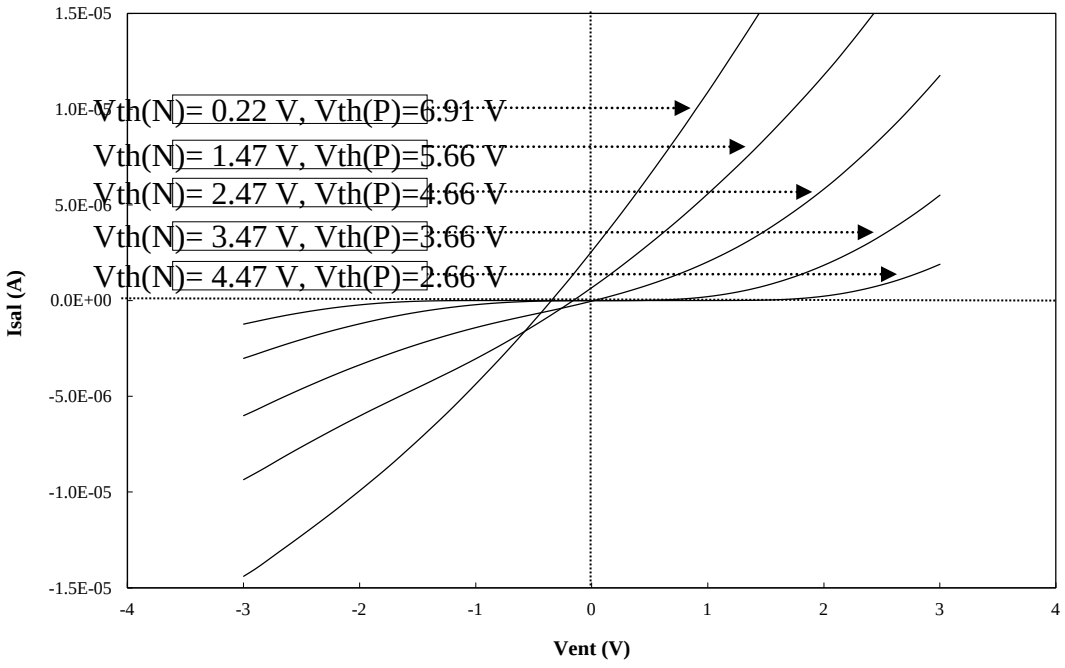


Fig. 6.8. Característica I-V de una sinapsis, con voltajes de umbral de partida de Vth(N)=0.22 V y Vth(P)=6.91 V.

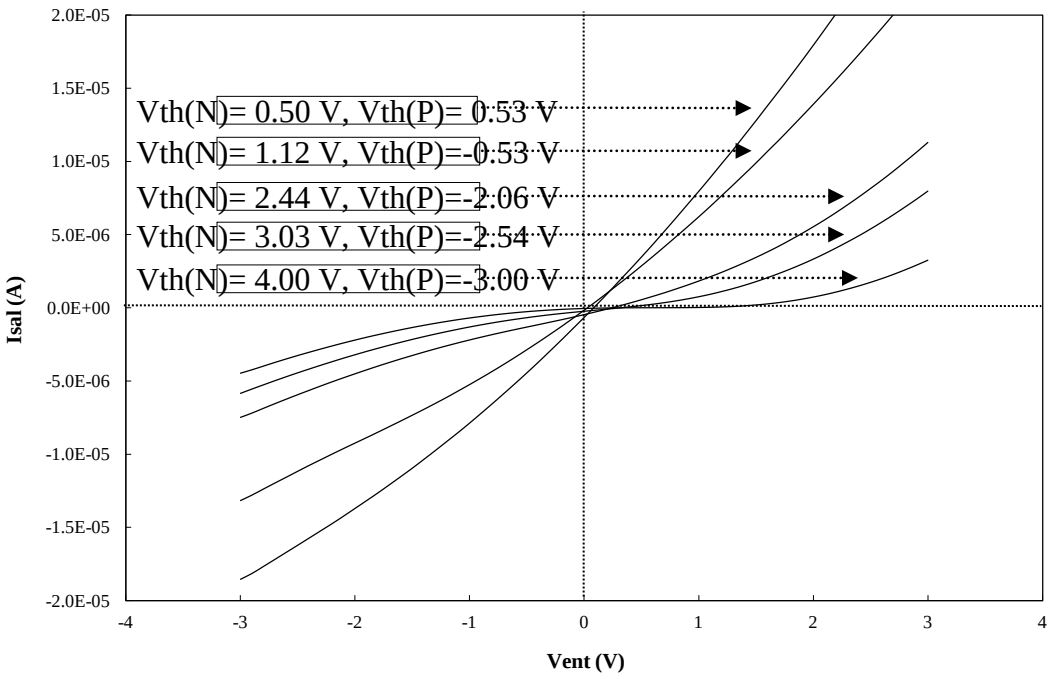


Fig. 6.9. Característica I-V de una sinapsis, con voltajes de umbral de partida de Vth(N)=0.53 V y Vth(P)=0.50 V.

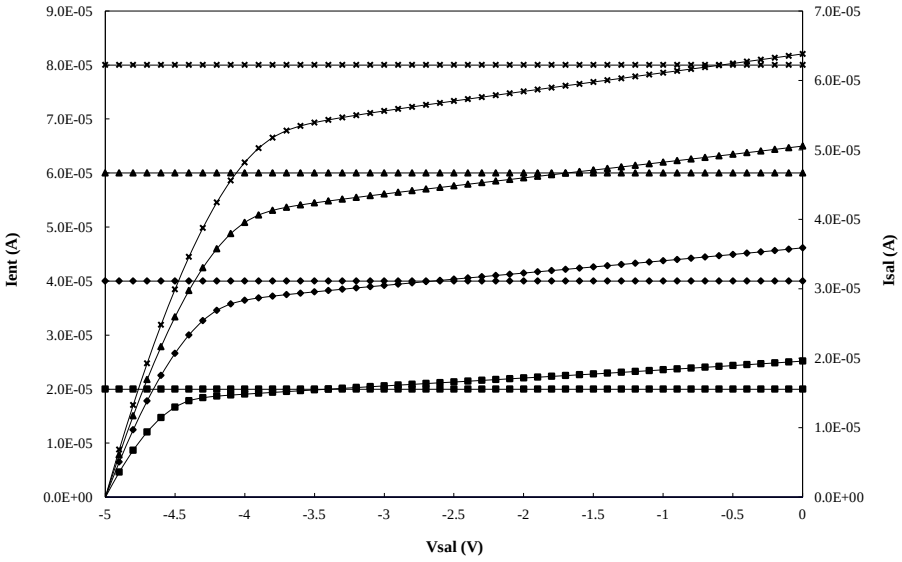
6.5. Espejo de corriente programable.

Como comprobación de la utilidad de las estructuras de compuerta flotante dentro de los circuitos analógicos, se pensó en configurar un espejo de corriente simple a partir de dos transistores programables. En este caso, se tomaron dos de los transistores FGNMOS con las siguientes características:

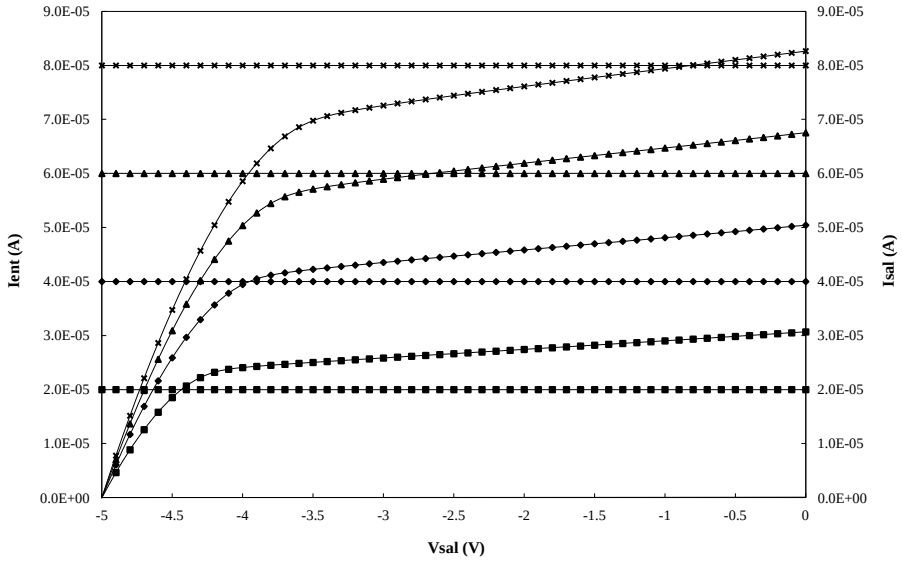
$$W = 4\ \mu m \qquad L = 3\ \mu m$$

y con estos se propuso un espejo de corriente simple, el cual se alimentó con pasos de corriente de 20 μA, empezando con 20 μA y terminando con 100 μA. En teoría, la corriente de salida del espejo debe ser igual a la de entrada, sin embargo, dadas las características de modulación de la longitud de canal de los transistores MOS, existe una diferencia dependiendo del voltaje de salida con que esté polarizado el espejo (pendiente de las características de salida debido a la λ del transistor). Inicialmente, se estableció un voltaje de umbral para ambos transistores de 1 volt y se obtuvieron las características de salida del espejo (Fig. 6.10(a), lo que corresponde al comportamiento de un circuito convencional. Posteriormente, únicamente al transistor de salida, se le programó el voltaje de umbral, llevándolo a un valor de 0 volts de la manera

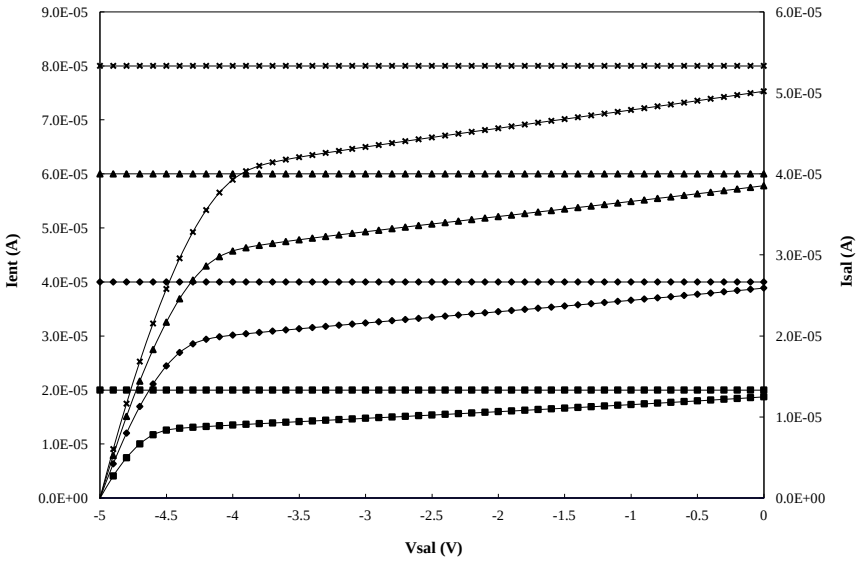
explicada en la sección 6.3, para después medir las características de salida del espejo, mostradas en la Fig. 6.10(b) y de donde se puede ver que la corriente de salida del espejo es mayor que cuando ambos transistores tenían el mismo voltaje de umbral. Finalmente, se llevó al transistor de salida del espejo a un voltaje de umbral de 1.95 V (~ 1 volt por encima del Vth del transistor de entrada), obteniéndose a la salida, una corriente menor a la de entrada (Fig. 6.10(c)).



(a)



(b)



(c)

Fig. 6.10. Comportamiento de un espejo programable, con transistores de compuerta flotante; a) Voltaje de umbral de ambos transistores, igual; b) Vth del transistor de salida 1 volt por debajo del Vth del transistor de entrada; c) Vth del transistor de salida 1 volt por encima del Vth del transistor de entrada.

De lo anterior, se puede ver que a pesar de que los transistores tienen las mismas dimensiones, la corriente de salida se puede escalar, a diferencia de la aproximación usual, en la que para hacer el escalamiento de las corrientes se requiere usar diferentes razones W/L en los transistores de salida. Esto permite también su uso en sistemas neuronales donde se puede diseñar una sinapsis mediante diferencia de corrientes [5], y de lógica difusa para implementar circuitos analógicos con universos de discurso variables y fuentes de corriente programables [6].

6.6. Comentarios acerca del diseño de la BAM.

La simulación presentada en el capítulo 4, con respecto a la Memoria Asociativa Bidireccional (BAM), fue posible gracias a que se lograron determinar los voltajes de umbral con los que la red funciona en el reconocimiento de los patrones escogidos y de esa manera establecerlos dentro del modelo de los transistores MOS en el programa PSpice, para los transistores de compuerta flotante. Esta simulación pasa por alto los procedimientos de programación, que sería el complemento de la implementación de la BAM, dado que no es necesario incluirlos en la simulación. Aún así, al hacer el diseño topológico se procuró complementar la red con circuitos periféricos que permitieran la programación independiente de cada una de las sinapsis, correspondiente a la matriz de correlación, como el decodificador y los interruptores CMOS, según se vio en el capítulo 3.

Este prototipo permitiría ajustar los procedimientos de almacenamiento de los patrones para optimizar el arreglo, es decir, el diseño presentado es una red de evaluación cuyo objetivo era obtener información a partir de la cual se establecieran criterios que permitieran la aplicación de las estructuras estudiadas, en algún sistema neuronal.

Vale la pena mencionar, por lo tanto, algunas de las dificultades que se presentaron al abordar la BAM incluida en el circuito integrado. Una de estas dificultades, se basó en el hecho de que se tiene un voltaje de umbral diferente en los transistores de compuerta flotante, como se había comentado anteriormente, y al detectarse que su distribución es en cierto modo aleatoria, se requiere ya sea un modo

de sensado con el cual se establezca de manera puntual el valor del voltaje de umbral, o bien, diseñar la red de tal forma que se pueda eliminar la carga de la compuerta flotante de todos los transistores FGMOS, mediante un borrado con luz ultravioleta, pero sin afectar los transistores sin compuerta flotante. Esto sugiere que se diseñe la red en forma modular, incluyendo únicamente las sinapsis en un circuito integrado y conectar externamente el decodificador (en caso de ser necesario) y las neuronas. Esto daría la ventaja de ampliar la capacidad de la BAM, ya que en este caso por falta de espacio, se ajustó a una red de 6x3 neuronas, dado que se empleó el encapsulado llamado Tiny-Chip, con una área de 2 mm x 2 mm con 40 terminales, lo que limita el diseño, o bien implementar una arquitectura diferente.

Por otro lado, como se había visto, la configuración de las estructuras de compuerta flotante mencionadas en este trabajo, tienen dos inyectores que se pueden manejar independientemente, uno para inyectar carga negativa y el otro para extraer carga negativa, siendo la magnitud y la polaridad del voltaje de la compuerta de control la que controlara el sentido de flujo de corriente de tunelamiento (ver sección 2.3.2). Este procedimiento implica fijar cada inyector a un voltaje de polarización, sin embargo, esto limita a su vez el intervalo del voltaje de umbral, como se ve en las Figs. 6.6 y 6.7. En el procedimiento presentado en esta tesis, los voltajes de umbral para la BAM fueron determinados de antemano por lo que no se requería de un proceso iterativo que ajustara este parámetro, como cuando se emplea un algoritmo de aprendizaje. Por lo tanto, en este caso en particular, no se necesitan los dos inyectores, que fueron incluidos en el diseño actual y, en consecuencia, se puede reducir el número de terminales ocupadas por los inyectores para ser aprovechados con otros fines.

Con respecto a lo comentado en el párrafo anterior, se podría pensar que para aumentar el intervalo del voltaje de umbral, se puede optar por aumentar la magnitud de los voltajes fijos aplicados a los inyectores, pero esto implica un problema: al elevar la magnitud, se aumenta la diferencia de potencial entre el inyector y la compuerta flotante lo que lleva, en teoría, a alcanzar un cambio mayor de voltaje de umbral, pero llevado esto al procedimiento práctico de programación, dado que se requiere aplicar el voltaje de compuerta de control en forma de pulsos, en un momento dado esta última tendría cero volts, pudiendo ser aún mayor la diferencia de potencial requerida, saliéndose del intervalo, por consecuencia, la inyección o extracción de carga. El hecho de trabajar con pulsos lo impone el arreglo capacitivo del modelo equivalente de la compuerta flotante, y de no hacerse así, es decir con DC, provoca un daño irreversible en los inyectores dejando inoperante a la red. Por lo tanto, se recomienda que la inyección o extracción de carga se controle mediante los inyectores, fijando el voltaje de la compuerta de control a ± 5 V y se apliquen los pulsos al (los) inyector(es). Este último procedimiento también fue examinado, dando excelentes resultados ya que se controlaba bien el cambio del voltaje de umbral aún cuando los dispositivos se encontraran polarizados. Una vez programados los transistores, los inyectores se aterrizan y se procura no aplicar voltajes mayores al umbral de tunelamiento en la compuerta de control, durante el funcionamiento del circuito.

6.7. Sumario.

En este capítulo se presentaron las mediciones realizadas a las estructuras de compuerta flotante, desde la determinación de las características de salida tanto de los transistores MOS normales, así como de los transistores de compuerta flotante, las curvas de transconductancia para la determinación práctica del coeficiente de acoplamiento, hasta la generación de la gráfica de variación del voltaje de umbral en función del tiempo, comparado con el modelo teórico. Se discuten los resultados en cada una de las caracaterizaciones, haciendo observaciones útiles y prácticas para la aplicación de los métodos propuestos en trabajos futuros relacionados con estructuras de compuerta flotante. Los resultados son consistentes con la teoría presentada en capítulos anteriores y termina de fundamentar los elementos teóricos y prácticos necesarios para operar este tipo de dispositivos, cuyo panorama en la aplicación de los circuitos analógicos y en particular de las RNA’s, parece ser promisorio.

APENDICE.

Modelo realizado en MathCAD para el cálculo del cambio del voltaje de umbral en función del tiempo.

Archivo: Pulso4.MCD
Inyección de carga negativa.
Alfa y Beta extraídos del artículo de Durfee y Shoucair, IEEE Trans. on Neural Networks
Vol. 3, No. 3, May 1992, pp. 347-352.
Las capacitancias corresponden al diseño topológico de las estructuras.

$\alpha := 4 \cdot 10^{-4}$ A/V**2 $\beta := 4.896 \cdot 10^{7}$ V/cm

$Kcg := 0.85$ Coeficiente de acoplamiento

$\varepsilon := 690 \cdot 10^{-8}$ Espesor del óxido de tunelamiento (cm)

$Ctun := 1.972 \cdot 10^{-15}$ Capacitancia del óxido de tunelamiento (F)

$Cgox := 3.452 \cdot 10^{-14}$ Capacitancia del óxido de compuerta (F)

$Cpp := 6.389 \cdot 10^{-13}$ Capacitancia del capacitor de acoplamiento (F)

$Vg := 5$ Voltaje en la compuerta de control (V)

$I :=$ $Qfg_0 \leftarrow -4.755 \cdot 10^{-13}$

for $i \in 0..100$

$t_i \leftarrow 1 \cdot 10^{-4} \cdot i$

$V_i \leftarrow \frac{15}{4 \cdot 10^{-3}} \cdot t_i$

V

$G_i \leftarrow \frac{5}{4 \cdot 10^{-3}} \cdot t_i$

G

$Vfg_i \leftarrow Kcg \cdot \left(if\left(G_i > 5, 5, G_i\right)\right) + \frac{Qfg_i}{Cpp + Cgox + Ctun}$

$J_i \leftarrow \alpha \cdot \left(\frac{if\left(V_i > 15, 15, V_i\right) - Vfg_i}{690 \cdot 10^{-8}}\right)^2 \cdot e^{\frac{-\beta \cdot \varepsilon}{if\left(V_i > 15, 15, V_i\right) - Vfg_i}}$

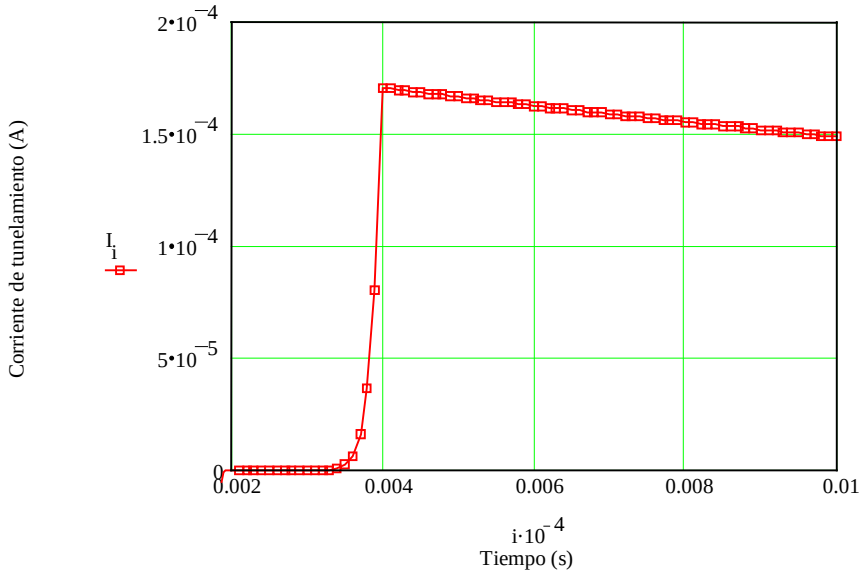
$q_i \leftarrow \int_{t_i}^{t_i + 1 \cdot 10^{-4}} J_i \cdot 3.5 \cdot 10^{-8}\, dt$

$Qfg_{i+1} \leftarrow Qfg_i + q_i$

Qfg

J

$i := 0..100$

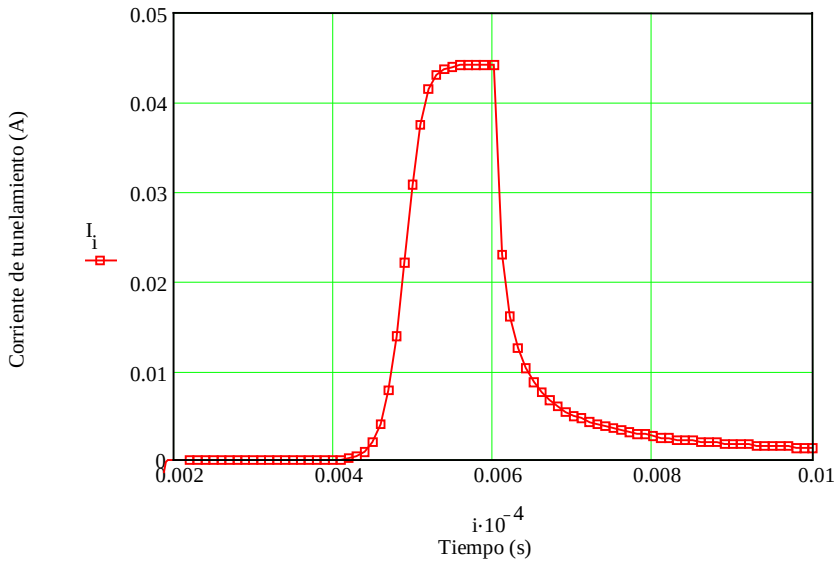


Inyección de carga positiva.
Alfa y Beta extraídos del artículo de Durfee y Shoucair, IEEE Trans. on Neural Networks Vol. 3, No. 3, May 1992, pp. 347-352.
Las capacitancias corresponden al diseño topológico de las estructuras.

$\alpha := 1.394 \cdot 10^{-1}$ A/V**2	$\beta := 4.819 \cdot 10^{7}$ V/cm
$Kcg := 0.85$	Coeficiente de acoplamiento
$\varepsilon := 700 \cdot 10^{-8}$	Espesor del óxido de tunelamiento (cm)
$Ctun := 1.972 \cdot 10^{-15}$	Capacitancia del óxido de tunelamiento (F)
$Cgox := 3.452 \cdot 10^{-14}$	Capacitancia del óxido de compuerta (F)
$Cpp := 6.389 \cdot 10^{-13}$	Capacitancia del capacitor de acoplamiento (F)
$Vg := 5$	Voltaje en la compuerta de control (V)

$$
\begin{aligned}
I := \;& Qfg_0 \leftarrow 0 \cdot 10^{-14} \\
& \text{for } i \in 0..100 \\
& \quad t_i \leftarrow 1 \cdot 10^{-4} \cdot i \\
& \quad V_i \leftarrow \frac{18}{6 \cdot 10^{-3}} \cdot t_i \\
& \quad V \\
& \quad G_i \leftarrow \frac{5}{6 \cdot 10^{-3}} \cdot t_i \\
& \quad Vfg_i \leftarrow Kcg \cdot if\left(G_i > 5, 5, G_i\right) + \frac{Qfg_i}{Cpp + Cgox + Ctun} \\
& \quad J_i \leftarrow \alpha \cdot \left(\frac{if\left(V_i > 18, 18, V_i\right) - Vfg_i}{680 \cdot 10^{-8}} \right)^2 \cdot e^{\frac{-\beta \cdot \varepsilon}{\left| if\left(V_i > 18, 18, V_i\right) - Vfg_i \right|}} \\
& \quad q_i \leftarrow \int_{t_i}^{t_i + 1 \cdot 10^{-4}} J_i \cdot 3.5 \cdot 10^{-8} \, dt \\
& \quad Qfg_{i+1} \leftarrow Qfg_i + q_i \\
& \quad J
\end{aligned}
$$

$i := 0..100$



$\alpha := 5.0 \cdot 10^{-4}$ A/V**2 $\beta := 6.527 \cdot 10^{7}$ V/cm

$Kcg := 0.85$ Coeficiente de acoplamiento

$\varepsilon := 700 \cdot 10^{-8}$ Espesor del óxido de tunelamiento (cm)

$Ctun := 1.972 \cdot 10^{-15}$ Capacitancia del óxido de tunelamiento (F)

$Cgox := 4.096 \cdot 10^{-14}$ Capacitancia del óxido de compuerta (F)

$Cpp := 4.929 \cdot 10^{-14}$ Capacitancia del capacitor de acoplamiento (F)

$Vga := 5$ Voltaje en la compuerta de control (V)

$Vi := 14$ Voltaje en el inyector (V)

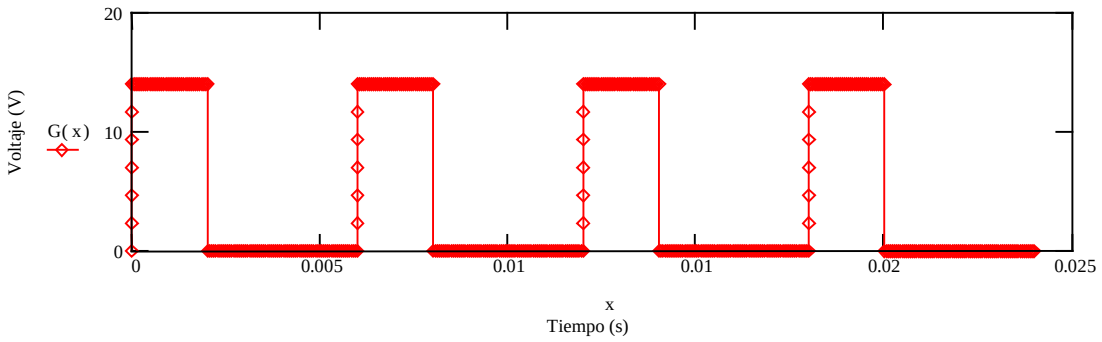
$V(x) := \frac{Vi}{6 \cdot 10^{-6}} \cdot x$ $Vg(x) := \frac{Vg}{6 \cdot 10^{-6}} \cdot x$ $period := 6 \cdot 10^{-3}$

$F(x) := if(V(x) > Vi, Vi, V(x))$

$H(x) := if\left(x - period > -4 \cdot 10^{-3}, 0, F(x)\right)$

$G(x) := if(x < period, H(x), G(x - period))$

$x := 0, 1 \cdot 10^{-6} .. 24 \cdot 10^{-3}$



$$\frac{Cpp}{Cpp + Cgox + Ctun} = 0.53447$$

$$
\begin{aligned}
Vth := \;& Qfg_0 \leftarrow 3.573 \cdot 10^{-13} \\
& \text{for } i \in 0..24000 \\
& \quad t_i \leftarrow 1 \cdot 10^{-6} \cdot i \\
& \quad Vfg_i \leftarrow \frac{-Cpp}{Cpp + Cgox + Ctun} \cdot Vga + \frac{Qfg_i}{Cpp + Cgox + Ctun} \\
& \quad J_i \leftarrow \alpha \cdot \left(\frac{-G\left(i \cdot 10^{-6}\right) - Vfg_i}{\varepsilon} \right)^2 \cdot e^{\frac{-\beta \cdot \varepsilon}{\left| -G\left(i \cdot 10^{-6}\right) - Vfg_i \right|}} \\
& \quad q_i \leftarrow J_i \cdot \int_{t_i}^{t_i + 1 \cdot 10^{-6}} 8 \cdot 10^{-8} \, dt \\
& \quad Qfg_{i+1} \leftarrow Qfg_i - q_i \\
& \quad Vth_i \leftarrow 0.88 - \frac{Qfg_{i+1}}{Cpp} \\
& \quad Vth
\end{aligned}
$$

$i := 0..24000$

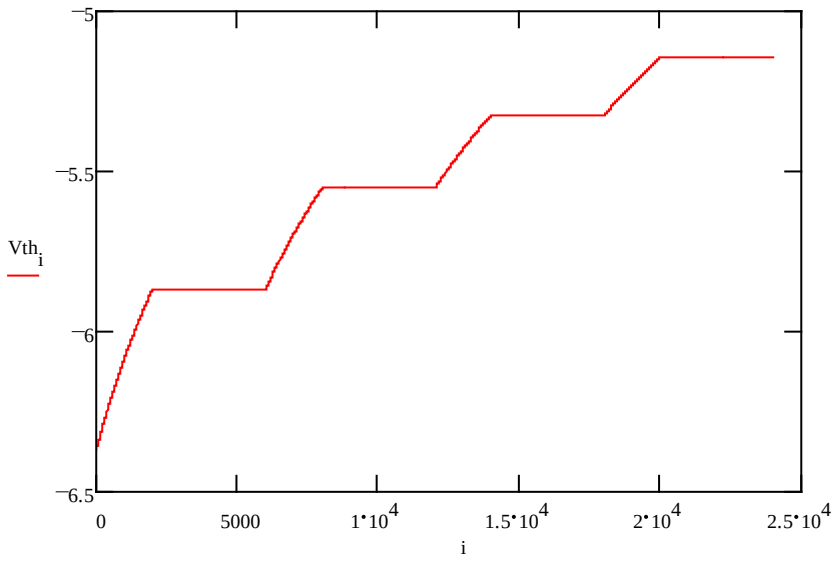
$Vth_0 = -6.36893$

$Vth_{6000} = -5.86746$

$Vth_{12000} = -5.55279$

$Vth_{18000} = -5.32443$

$Vth_{24000} = -5.14572$



Referencias.

1.- K. Yang and A. G. Andreou, "Multiple input floating-gate MOS differential amplifiers and applications for analog computation", Proc. 36th Midwest Symposium on Circuits and Systems, pp. 1212-1216, New York, NY, Aug. 1993.

2.- K. Yang and A. G. Andreou, "Subthreshold analysis of floating-gate MOSFET's", *Proc. Tenth Biennial UGIM symposium*, pp. 141-144, Research Triangle Park, NC, May 1993.

3.- A. Kolodny, S. T. K. Nieh, B. Eitan and J. Shappir, "Analysis and modeling of floating-gate EEPROM cells", *IEEE Trans. On Electron Devices*, Vol. ED-33, No. 6,pp.835-844, June, 1986.

4.- D. A. Durfee and F. S. Shoucair, "Comparison of floating-gate neural network memory cells in standard VLSI CMOS technology", *IEEE Trans. on Neural Networks*, Vol. 3, No. 3, pp. 347-352, May, 1992.

5.- T. H. Borgstrom, M. Ismail and S. B. Bibyk, "Programmable corrent-mode neural network for implementation in analogue MOS VLSI", *IEE Proc.*, Vol. 137, Pt. G, No. 2, pp.175-184, Apr 1990.

6.- Juan Jesús Ocampo Hidalgo, "Diseño de circuitos CMOS de entrada para un sistema de inferencias difusas, *Tesis de Maestría*, Julio 1997.

Capítulo 7

Conclusiones.

Dada la importancia que se está dando al desarrollo de electrónica para la aplicación práctica de las redes neuronales artificiales, es básico el estudio de estructuras que permitan el desarrollo de diseños prototipo para su aplicación en esta rama y para contribuir a su vez en el desarrollo de sistemas neuronales. Es así, que en este trabajo se estudiaron elementos de memoria analógicos no volátiles, con los que se pueden tener, por ejemplo, configuraciones aplicadas como sinapsis, con capacidad de variación de conductancia, almacenamiento de información y con la posibilidad de procesar fácilmente la señal de entrada.

Como conclusiones importantes, derivadas del desarrollo de esta tesis, se pueden enumerar las siguientes:

1) Se diseñó una estructura CMOS, en base a un inversor, para operar como elemento de resistencia variable, cuya programación se hace por medio de tuneleo Fowler-Nordheim. Su fabricación se hizo aprovechando una tecnología estándar y de bajo costo

2) Se profundizó en el estudio de las características físicas y operacionales de la estructura elegida, de tal forma que se pudo proponer un modelo simple que se pudiera aplicar en el programa de simulación de circuitos, PSpice que incluye: el coeficiente de acoplamiento derivado del diseño físico y eléctrico y se aprovecha el nivel 2 del modelo de los transistores MOS. Los resultados obtenidos, muestran que el modelo es válido y funcional.

3) Uno de los parámetros de mas importancia que hay que considerar en el diseño de las estructuras propuestas es el coeficiente de acoplamiento. Se vio la importancia que tiene en el desempeño de un circuito, ya que un valor no óptimo de éste puede provocar una respuesta no deseada para la BAM, por lo que se requiere su optimización tanto desde el punto de vista de área como de funcionamiento.

4) Desde el punto de vista físico, se propuso un método sencillo y novedoso que permite caracterizar el coeficiente de acoplamiento de estas estructuras, apoyado en el conocido método de extrapolación del voltaje de umbral usando la gráfica de transconductancia del TMOS. Con el método propuesto no se requieren cálculos ni montajes de medición complicados. Esta es una aportación importante de este trabajo.

5) Haciendo uso del modelo de simulación de operación de la estructura de compuerta flotante, se propuso el diseño de un prototipo para una Memoria Asociativa Bidireccional (BAM), con lo que se demuestra que los elementos estudiados pueden ser aplicados de manera práctica a redes neuronales artificiales. Esto es una aportación en el desarrollo de la electrónica dentro de este campo, ya que el desarrollo de dispositivos y circuitos ha sido mas lento que el de programas de cómputo.

6) De manera práctica, se comprobó la validez del modelo propuesto al medir las características de la sinapsis diseñada, siendo estas como las previstas por la simulación. En el mismo sentido, se obtuvieron las características de salida de un espejo de corriente programable, con el que se puede escalar la corriente de entrada sin necesidad de escalar la razón geométrica del transistor de salida, sino variando el voltaje de umbral de este último.

Trabajo futuro.

El trabajo mostrado y las conclusiones expuestas, se pueden considerar como la demostración de que los conceptos básicos del funcionamiento de la estructura de compuerta flotante elegida, fueron comprendidos y abren una buena posibilidad de seguir ampliando su estudio. Dados los resultados exitosos presentados en este trabajo, con características de estructuras que pueden ser aplicables a las RNA's, uno de los trabajos futuros en forma natural sería el desarrollo práctico de una arquitectura como la BAM, o un poco mas compleja, incluyendo adaptabilidad o aprendizaje en línea.

También existe un estudio derivado del análisis hecho en este trabajo, con respecto al coeficiente de acoplamiento, en el cual se intenta disminuir el área obteniéndose beneficios en cuanto a la densidad de integración y la velocidad de programación.