TESIS DOCTORAL

DOCTORADO EN INGENIERÍA
MENCIÓN PROCESAMIENTO DE SEÑALES E IMÁGENES

Título:

“Degradación de estructuras MOS (metal-óxido-semiconductor) y sus aplicaciones”

Autor: Fernando Leonel Aguirre

Director de Tesis: Dr. Félix Roberto Mario Palumbo

Co-Director de Tesis: Dr. Pedro Julián

Buenos Aires - 2021

Resumen

HISTÓRICAMENTE, la mejora de rendimiento de los circuitos microelectrónicos ha ido asociada a un aumento en la cantidad de transistores por unidad de área. Esto ha sido posible debido a una reducción sostenida en el tamaño y costo de cada dispositivo. Sin embargo, el ritmo de miniaturización se ha ralentizado sensiblemente en los últimos años, por limitaciones tanto económicas como tecnológicas, con lo que la industria nano-electrónica se ha visto forzada a buscar alternativas considerando nuevos materiales así como también el cambio de los paradigmas computacionales. En ambos casos, la confiabilidad es un apartado clave en el desarrollo de las tecnologías de integración venideras. Pero así como la introducción de materiales novedosos siembra interrogantes a responder en el ámbito de la fiabilidad, también da lugar a nuevas aplicaciones y al surgimiento de tecnologías disruptivas. En esta tesis, se discuten los principales fenómenos de degradación observados en estructuras MOS (Metal-Óxido-Semiconductor) implementadas con materiales alternativos, así como las oportunidades que estas brindan para la aparición de alternativas superadoras frente a la microelectrónica convencional.

Para ello, se plantea en primera instancia el estudio de la degradación y ruptura en estructuras basadas en semiconductores de alta movilidad y óxidos delgados no nativos como reemplazo de la combinación silicio-dióxido de silicio (Si/SiO_2). En segundo lugar, el fenómeno de ruptura dieléctrica reversible en aislantes de alta constante dieléctrica ha dado lugar a su aplicación como dispositivo de memoria y sinapsis artificial en circuitos neuromórficos, los cuales son un cambio disruptivo respecto a los actuales sistemas de cómputo y tienen múltiples aplicaciones en el paradigma del Internet del Todo (*Internet of Everything*, IoE), el reconocimiento de patrones y la inteligencia artificial, solo por citar algunos ejemplos.

Con respecto al estudio de la confiabilidad a nivel de los dispositivos nano-electrónicos, se analiza el origen e impacto de los defectos del óxido en la dinámica de degradación de estructuras MOS sobre sustratos de alta movilidad, llamados a ser el reemplazo de la tecnología Si/SiO_2. Se abordan los efectos del proceso de fabricación sobre la distribución de defectos y el atrapamiento de carga, reportando las diferencias entre los principales candidatos.

A su vez, se estudia la mecánica de generación de defectos en óxidos delgados de alta constante dieléctrica, como origen de la ruptura dieléctrica. Se reporta la influencia de las propiedades intrínsecas de óxidos de alta constante dieléctrica en la estadística de ruptura de dispositivos MOS, así como de la densidad de defectos. Para ello, se han llevado adelante complejos experimentos de irradiación altamente localizada a fin de al-

terar artificialmente la concentración de defectos. Como resultado, se reporta evidencia experimental que deja entrever la naturaleza correlacionada de la generación de defectos en materiales de alta constante dieléctrica, ayudando a comprender mejor el fenómeno de ruptura.

Si bien la ruptura dieléctrica representa un desafío de confiabilidad, el mecanismo físico subyacente es al mismo tiempo responsable de posibilitar el funcionamiento de las memorias de conmutación resistiva, de gran utilidad en aplicaciones de almacenamiento de información y circuitos neuromórficos. En este trabajo de tesis, se discuten los aspectos temporales del evento de conmutación tanto en memorias no volátiles como volátiles, utilizando aislantes de alta constante dieléctrica así como nitruro de boro hexagonal (h-BN), un aislante de dos dimensiones (2D) que representa una alternativa con considerable potencial tecnológico. Se contribuye con una interpretación para la velocidad de la conmutación centrada en el papel de la temperatura y las características intrínsecas y geométricas del medio de conmutación.

Justamente la aplicación de memorias de conmutación resistiva en circuitos neuromórficos es el foco de estudio en la segunda parte de esta tesis. Los mismos son de gran interés dada su capacidad para procesar grandes volúmenes de información con baja latencia y bajo consumo de energía, sin mencionar la gran densidad de dispositivos integrables por unidad de área. Sin embargo, las múltiples no-idealidades que dominan el funcionamiento de estos dispositivos suponen un problema para su desarrollo futuro. En este contexto, se presenta la relación entre confiabilidad y rendimiento en redes neuronales implementadas en *hardware* mediante *crossbars* de memorias de conmutación resistiva y destinadas al reconocimiento de patrones.

Para ello se propone un flujo integral que contempla el modelado eléctrico de la física de conmutación resistiva, el entrenamiento de redes neuronales para la clasificación de imágenes, su representación a nivel circuital y simulación eléctrica contemplando la posible existencia de fallas de enclavamiento distribuidas aleatoriamente. En relación al último punto, se plantean también métodos para mitigar las consecuencias de las mismas. Puntualmente, se reportan resultados considerando perceptrones mono y multi capa, destacándose el impacto de la resistencia parásita de las interconexiones, las características eléctricas de las dispositivos de conmutación resistiva y el impacto de las mencionadas fallas de enclavamiento.

Abstract

HISTORICALLY, the improvement of the microelectronic circuit's performance has been associated with an increase in the number of transistors per area. This has been possible due to the sustained reduction of the device's size and cost. Nonetheless, the pace of miniaturisation has sensibly slowed down during in the last decades, because of both technological and economic reasons, forcing the nano-electronics industry to look for alternatives considering novel materials as well as disruptive computational paradigms. In both cases, reliability is a key in the development of the upcoming integration technologies, raising challenges to the introduction of new materials. However, the very same physical mechanisms that threatens the device reliability is responsible for allowing novel applications which are expected to be a breakthrough in computing technologies. In this thesis, the main wear-out and breakdown phenomena observed in MOS (Metal-Oxide-Semiconductor) structures implemented with alternative materials are studied, as well as the opportunities they create against the conventional nano-electronic devices and techniques.

To comply with this objective, the first part considers the study of the wear-out and breakdown of structures based on high-mobility substrates and thin, non-native oxides proposed as replacement of the well-known Si-SiO_2 combination. Then, the second part talks about the phenomenon of reversible dielectric breakdown in high constant dielectrics (*high–*κ dielectrics), which has allowed their applicability as memory devices and artificial synapses in neuromorphic circuits, which supposes a change of paradigm with respect to the current computing systems and has a pletora of applications in the field of *Internet of Everything* (IoE), pattern recognition, and artificial intelligence, just to mention the most common examples.

Regarding the first part, the study of the reliability at the nano-electronic device level is organised around the origin and impact of the oxide defects on the wear-out dynamics of MOS structures fabricated on high-mobility substrates, proposed as replacement for the Si-SiO_2 technology. In this context, the effects of the fabrication process on the defect distributions and charge trapping phenomena are discussed, reporting the difference between the main candidates to succeed the Si-SiO_2 technology.

The generation dynamics of defects in thin *high–*κ dielectrics is also studied given their key role in the breakdown phenomenon. The influence of the intrinsic properties of the *high–*κ dielectrics, as well as the defect densities, in the breakdown statistics of MOS devices is reported. To do so, complex experiments involving highly localised irradiation with high-energy ions were carried out in order to artificially and precisely tune the defect concentration. As a result, this thesis reports experimental evidence suggesting the

correlated nature of the generation of defects in *high–κ* insulators, paving the way to a better understanding of t he breakdown phenomenon.

Although the dielectric breakdown supposes a reliability challenge, the underlying physical mechanism is at the same time the enabler of the resistive switching memories, of great utility in storage applications and neuromophic circuits. In this work, the temporal aspects of the switching event in volatile and non-volatile memories are discussed, considering both *high–κ* and hexagonal boron-nitride (h-BN, a two-dimensional dielectric with outstanding technological potential), respectively. In this regard, this thesis contributes with an interpretation of the switching speed based on the role of the temperature and the intrinsic and geometrical characteristics of the switching medium.

The application of resistive switching memories in neuromophic circuits is precisely the core of the study reported in the second part of this thesis. These are of great interest given their capacity to process large volumes of information with low latency and energy consumption, not to mention the high density of integration per unit area. However, the multiple non-idealities (inherent to this devices) pose a threat to their further development of these devices. In this context, the trade-off between reliability and performance in neural networks implemented in hardware with resistive-switching memories and intended for pattern recognition tasks.

To do so, this thesis proposes an integral workflow considering the electrical modelling of the resistive switching devices, the training of neural networks for the classification of images, its representation as a circuit and its electrical simulation considering parasitic devices as well as randomly distributed stuck-at-faults. With regard to the last point, methods intended to mitigate the impact of stuck-at-faults are proposed and tested. Among other aspects, the reported result include single and multi-layer perceptrons and the impact of the interconnections resistance, the electrical characteristics of the resistive switching devices and the aforementioned stuck-at-faults.

Dedicatoria y Agradecimientos

ESTA tesis está dedicada a todas a todas aquellas personas que han contribuido a su desarrollo ya sea desde el aporte técnico/científico o bien acompañándome durante estos años de trabajo. Sin su aporte, el cumplimiento de los objetivos propuestos al inicio de este camino no podrían haber sido alcanzados. A todos y cada uno de ellos, mi más sincero y profundo agradecimiento.

A mis directores, Dr. Félix Palumbo y Dr. Pedro Julián, no sólo por el valor de su vasto conocimiento en las temáticas de esta tesis sino también por su predisposición para conmigo y este trabajo y por su atención constante en proveerme las facilidades necesarias para la ejecución de tareas, ya sea financiamiento, disponibilidad de equipamiento o simplemente sus valiosos consejos. El apoyo que me han brindado a lo largo de estos años es difícil de transmitir en pocas palabras, y es por eso mismo que estoy profundamente agradecido no solo por su aporte académico sino por haber podido compartir este período con ellos.

A todas las autoridades y personal de la UTN-FRBA, pero particularmente al Ing. Guillermo Oliveto, decano de la facultad, al Departamento de Ingeniería Electrónica (Ing. Alejandro Furfaro, Ing. Marcelo Doallo y Ing. Carlos Navarro), a la Secretaría de Ciencia, Tecnología e Innovación Productiva (Lic. Patricia Cibeira y Lic. Florencia Counyo) y por último a las autoridades del Doctorado en Ingeniería de la UTN-FRBA (entre ellos Dr. Ricardo Armentano, Dr. Walter Legnani y Dr. Leandro Cymberknop) que me han acompañado en este camino desde mucho antes de comenzar el doctorado y que han puesto todo su esfuerzo y predisposición en garantizar las mejores condiciones de trabajo posibles. A su vez, al Dr. Enrique Miranda y al Dr. Jordi Suñé, por haberme recibido en la Universidad Autónoma de Barcelona, España, y no solo hacerme sentir como en mi casa, sino también por su invaluable aporte a mi trabajo de tesis. Asimismo, agradezco a las autoridades del Grupo de Materia Condensada y del laboratorio TANDAR, por recibirme durante los primeros años de beca doctoral. En particular al grupo de la linea del micro-haz de iones pesados (Dr. Mario Debray, Dr. Nahuel Vega), por permitirme acceder a sus instalaciones para el desarrollo de esta tesis.

Al Laboratorio de Nanoelectrónica de la UIDI/UTN-FRBA, por su apoyo constante, discusiones enriquecedoras y sobre todo, por su increíble valor humano: Dr. Félix Palumbo, Dr. Hernán Giannetta, Ing. Andrés Fontana, Lic. Santiago Boyeras e Ing. Gabriel Maroli. Particularmente, a mi colega devenido en amigo, Ing. (y más recientemente Dr.) Sebastián Pazos, con quien he compartido interminables horas de trabajo, desafíos, momentos de alegría y desazón, y todos los condimentos de un doctorado. El trabajo desarrollado en esta tesis no hubiera sido posible sin el aporte del equipo de trabajo, cuya participación va más allá de lo que el listado de autores asociados a las publicaciones enmarcadas en este trabajo pueda atestiguar.

Un especial agradecimiento va dirigido a todos los colaboradores externos, nacionales y extranjeros, que contribuyeron a estos años de trabajo. A los investigadores y autoridades de la Universidad Tecnológica Nacional Facultad Regional Villa María por su guía y predisposición. Asimismo, a los investigadores del Technion Israel Insitute of Technology, Israel (Dr. Moshe Eizenberg, Dr. Igor Krylov, Dr. Sivan Fadida), el IMM-CNR en Catania, Italia (Dr. Salvatore Lombardo), de la Universidad de Soochow, China (Dr. Mario Lanza), de la Universidad de Singapur de Tecnología y Diseño, Singapur (Dr. Nagarajan Raghavan, Dr. Alok Ranjan y Dr. Kin Leong Pey), de Applied Materials, EE.UU. (Dr. Andrea Padovani) Todos y cada uno de ellos contribuyeron a distintos aspectos de esta tesis, ya sea proveyendo muestras para experimentos, discusiones valiosas sobre los resultados o incluso su compañía y sus consejos durante conferencias en el extranjero.

A Cintia por su amor y apoyo incondicional. A Pura y Jesús por su guía y por haberme legado sus virtudes y su ética. A Paula por acompañarme en esta incursión al mundo académico. A la memoria de mis abuelos, por que toda historia tiene un principio. A Visita y Felisa, a día de hoy las matriarcas de una familia amplia. A Luis y Betty, por estar siempre a mi lado. A todos y cada uno de integrantes de una familia grande y unida. Y por su puesto, a mis amigos, esa parte de la familia que se elige. A todos, infinitas gracias por interesarse en mi trabajo, por ponerse a mi disposición o sencillamente compartir los momentos buenos y/o malos.

A la comunidad de microelectrónica de la Argentina, con colegas que recorrieron el doctorado a la par e investigadores que siempre se pusieron a disposición para conversar, compartir experiencias y trabajar en conjunto, entre ellos no puedo dejar de mencionar a los Doctores Gabriel Sança, Fabricio Alcalde y Nicolás Calarco. Por último pero no por ello menos relevante, a los Ingenieros Federico Di Vruno y Emilio Álvarez, que me recibieron como estudiante de grado para dar mis primeros pasos en el área de la investigación. Estas lineas difícilmente hubieran sido escritas de no ser por ellos.

Al CONICET por su apoyo financiero a través de la beca doctoral que me permitió dedicarme exclusivamente a este trabajo y a los mecanismos de financiamiento del MINCYT y la UTN-FRBA, sin los cuales muchos resultados no hubieran sido posibles.

Finalmente, un profundo agradecimiento a los destacados miembros del tribunal de evaluación, tanto titulares como suplentes, por su enorme predisposición y por sus

valiosas contribuciones a la versión final de esta tesis.

Declaración del autor

DECLARO que el trabajo en esta disertación fue llevado a cabo de acuerdo con los requerimientos fijados por la Escuela de Posgrado de la UTN-FRBA para el Doctorado en Ingeniería Mención en Procesamiento de Señales e Imágenes, y no fue presentada en ningún otro ámbito académico. Excepto en donde se aclara por referencias, este trabajo ha sido desarrollado por el alumno que lo remite. Trabajos en colaboración, o con asistencia de terceros, están indicados como tal. Cualquier opinión o conclusión que se presenta en este trabajo, es desde el punto de vista del autor.

FIRMADO: ..... FECHA: .....

Índice general

Página

Lista de tablas **XVII**

Lista de figuras **XIX**

1 Introducción **1**
1.1 Organización del trabajo de tesis 8

2 Conceptos básicos **11**
2.1 La estructura MOS 11
2.1.1 Curvas de Capacidad-Tensión 14
2.1.2 Curvas de Corriente-Tensión (I-V) 21
2.1.3 Degradación y ruptura de dispositivos MOS 22
2.2 Fundamentos de Redes Neuronales 26
2.2.1 Estructura básica 26
2.2.2 Entrenamiento supervisado y retro-propagación (*backpropagation*) . . 29

3 Degradación en estructuras MOS **35**
3.1 Materiales de alta movilidad 35
3.2 Sustratos de compuestos III-V 37
3.2.1 Dispersión de capacidad — impacto de las Trampas de Borde 38
3.2.2 Histéresis de capacidad — impacto de la capa interfacial 40
3.3 Sustratos de Germanio 43
3.3.1 Rol de los defectos en el atrapamiento de carga 44
3.3.2 Influencia del proceso de fabricación 51
3.4 Conclusiones 53

4 Dinámica de ruptura en dieléctricos **55**
4.1 Diferencias en la estadística de ruptura: SiO_2 y *high–*κ 55
4.2 Daño inducido por radiación en estructuras MOS 57
4.2.1 Impacto en las curvas de C–V 62
4.2.2 Impacto en las curvas I–V 63
4.3 Generación de defectos: Efecto sobre TDDB 64
4.3.1 Modelo de *Clustering* 67

4.3.2 Distribución de Weibull vs. modelo de *Clustering* 69
4.3.3 Plataforma de simulación multi-física 71
4.3.4 Identificación de la dinámica espacio-temporal de ruptura en dieléctricos *high–*κ 72
4.4 Conclusiones 74

5 Conmutación Resistiva en dieléctricos **75**
5.1 Similitudes con el mecanismo de ruptura progresiva 76
5.2 Conmutación Volatil y No-Volatil 78
5.2.1 Conmutación No-Volátil, o *Memory Resistive Switching* 79
5.2.2 Conmutación Volatil, o *Threshold Resistive-Switching* 82
5.3 Modelo compacto para la transición del estado de alta a baja resistividad . . 85
5.3.1 Dependencia con el espesor del dieléctrico 87
5.3.2 Rol de las especies migrantes 89
5.3.3 Ajuste de los datos experimentales 91
5.4 Conclusiones 93

6 Redes Neuronales **95**
6.1 *Cross-bar arrays* de memristores en redes neuronales 96
6.2 Modelo Cuasi-Estático del *Memdiodo* (QMM) 100
6.3 Simulación eficiente en SPICE de redes neuronales 104
6.3.1 Circuitos neuromórficos basados en RRAM 107
6.3.2 Complejidad computacional 110
6.4 Impacto de los elementos circuitales parásitos 111
6.4.1 Perceptrón Mono-Capa (*Single Layer Perceptron*, SLP) 112
6.4.2 Perceptrón Multi-Capa (*Multi Layer Perceptron*, MLP) 118
6.5 Aspectos de confiabilidad 120
6.5.1 Consideraciones de diseño 120
6.5.2 Estrategias para la mitigación de fallas de enclavamiento 124
6.6 Conclusiones 130

7 Conclusiones y próximos pasos **131**
7.1 Contribuciones 131
7.2 Perspectivas a futuro 133

A Lista de publicaciones centrales **135**
A.1 Artículos en revistas con referato 135
A.2 Artículos en *proceedings* de conferencias indexadas 136
A.3 Participación en publicaciones relacionadas a la línea de trabajo 137

B Modelos, *datasets* y resultados de sim. SPICE de ANN **139**
B.1 Modelo lineal de *Cross-bar* de memristores 139

B.2 Modelo SPICE del Memdiodo Cuasi-Estático (*Quasi-Static Memdiode*, QMM) 141
B.3 Base de datos utilizadas en los análisis 142
B.4 Validación de los algoritmos de entrenamiento 143
B.5 Métricas adicionales para el rendimiento de inferencia 146
B.6 Algoritmos de re-mapeo para la minimización de SAFs 151

Bibliografía **153**

Lista de tablas

TABLA **Página**

4.1 Códigos asignados a las diferentes muestras considerando diferentes fluencias y áreas de irradiación 60

4.2 Valores de R^2 (Coeficiente de determinación) para el ajuste con el modelo de *Clustering* tanto para las irradiaciones de área pequeña y área grande. Las celdas sombreadas indican el caso de mayor R^2 para cada escenario . . 70

6.1 Comparación entre diferentes tipos de redes neuronales implementadas con RRAM y los algoritmos de entrenamiento empleados. Nótese que en todos los casos las capas sinápticas están implementadas con CPAs y las simulaciones no modelan correctamente los efectos de resistencia de linea ni contemplan modelos realistas de RRAM. Dado que el CPA es un bloque constructivo fundamental en estas redes neuronales realizadas en hardware, la simulación realista en SPICE de los CPA es de suma importancia. . . . 106

6.2 Rangos de conductancia utilizados en la bibliografía 112

6.3 Estructura de los MLPs abordados en las simulaciones consideradas en esta sección. En todos los casos se utilizó como patrón de entrada a las imágenes del MNIST re-escaladas a una resolución de 8×8 px. 118

6.4 Cociente $V_{celda}/V_{lectura}$ calculado con el equivalente serie simplificado 123

6.5 Combinaciones de fallas recuperables y no-recuperables 125

B.1 Modelo SPICE del *memdiodo*: `S(x)` y `R(x)` son las funciones logísticas Γ^+(V) y Γ^-(V). `A(x)` y `RS(x)` representan a los parámetros α y R del memdiodo, los cuales son una función del estado memorizado. El estado memorizado λ es indicado por `V(H)`, con `H0` el estado inicial. Los parámetros que modelizan la transición HRS-LRS son `etaset`, `etares`, `vset` y `vres` para η^+, η^-, V^+ y $V-$, respectivamente. Se utilizan fuentes de corriente controladas por tensión para implementar la Ec. 6.1. (`GD` y el resistor `RS`) y la Ec. 6.2 (`GH` y el capacitor `CH`). Los diodos en anti-paralelo se modelan mediante la fuente de corriente controlada `GD` en el sub-circuito. `beta` define si la conducción es simétrica (igual para tensiones positivas y negativas, `beta`=0.5) o no (`beta`≠0.5). La Ec. 6.1 corresponde al caso de `beta`=1 pero es simetrizada utilizando la función de valor absoluto. El modelo está escrito utilizando la sintaxis de H-SPICE 141

B.2 $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST, codificada en 8×8 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %) 144
B.3 $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-F, codificada en 8×8 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %) 144
B.4 $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-K, codificada en 8×8 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %) 144
B.5 $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST, codificada en 28×28 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %) 145
B.6 $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-F, codificada en 28×28 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %) 145
B.7 $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-K, codificada en 28×28 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %) 145

Lista de figuras

FIGURA **Página**

1.1 (a) Evolución de la cantidad de transistores por *chip* desde la introducción del primer circuito integrado hasta la actualidad. Reproducido de [2]. (b) Fin de la era del escalamiento pronosticado hacia el inicio de la década de los 2020. Reproducido de [3]. (c) Integración de los enfoques *More Moore* y *More than Moore* como medio para el desarrollo futuro de la microelectrónica, con la integración tri-dimensional como gran ventaja. (d) Ejemplo de integración monolítica multicapa de un sistema *In-Memory Computing*. Reproducido de [4] 2

1.2 (a) Movilidad de portadores en distintos sustratos considerados para la estructura MOS. Los símbolos de color indican movilidad de electrones (cargas negativas) y los símbolos vacíos movilidad de huevos (cargas positivas). Reproducido de [20]. La combinación de transistores P-MOS (Ge) con transistores N-MOS (III-V) sobre una misma oblea de Si se propone como una opción para optimizar el rendimiento de futuros sistemas CMOS. Reproducido de [21] 4

1.3 Dieléctricos *high–*κ: Aparte de la constante dieléctrica, otros aspectos del material aislante son relevantes en su funcionamiento como dieléctrico de compuerta, como el ancho de la zona prohibida y el desfase entre las bandas de conducción. Estos se muestra en (a) Constante dieléctrica vs. Ancho de la zona prohibida, y (b) Constante dieléctrica vs. Desfase entre bandas de conducción. Reproducido de [32] 5

1.4 (Der.) Curvas de Corriente–Tensión (I–V) típicas mostrando las distintas etapas de la formación del CF. En los *inset* (a) y (b) se representa esquemáticamente el CF. En (c) y (d) se indica la migración de iones/defectos desde y hacia el CF, produciendo variaciones aleatorias en la corriente. En (e) se representa el punto en el cual el CF se completa, produciendo un aumento abrupto de la corriente a través de la estructura. Finalmente en (f) se muestra el caso de un CF completamente formado. Reproducido de [33]. (Izq.) Principales mecanismos explotados en memorias emergentes. Además de las ya mencionadas, se pueden citar las memorias RAM de cambio de fase (*Phase Change Random Acces Memory*, PCRAM) y las RAM magnéticas (*Magnetic Random Access memory*, MRAM). Reproducido de [34] 6

1.5 (a) Ejemplo de un patrón de entrada a clasificar, en este caso la letra 'S'. (b) Representación esquemática de una red neuronal de una sola capa. En este caso contiene 49 neuronas de entrada, y 26 neuronas de salida, con un total de 49×26=1274 sinapsis. Las lineas rojas indican las conexiones activas para el patrón de entrada asumido. (c) Implementación en *hardware* de la red neuronal utilizando un *crossbar* de RRAM de 2 dimensiones. (d) Una opción más avanzada sería la construcción de un *crossbar* vertical de memorias RRAM (VRRAM), el cual permita el mismo rendimiento con un mejor aprovechamiento del área disponible en silicio. Reproducido de [46] . 7

2.1 (Arriba) Estructura básica de un capacitor MOS. (Abajo) (a)-(d) Diagramas de bandas de energía en cada condición de operación, según potencial aplicado V_G tanto para sustrato P (fila superior) como N (fila inferior). Adaptado de [51]. 12

2.2 (a) Circuito equivalente de capacidades y (b) curvas capacidad-tensión para una estructura MOS ideal a baja (lineas a trazos) y alta (lineas llenas) frecuencia. (c) Impacto de defectos en el óxido y la interfaz sobre la curva C-V real (líneas rojas) respecto a la ideal (líneas negras). Reproducido de [52] 15

2.3 (a) Diagrama esquemático de la distribución de defectos defectos en la estructura MOS, dando lugar a distintos tipos dependiendo de su posicionamiento. (b) Diagrama de bandas considerando los distintos tipos de trampas. Reproducido de [52].(c) Curvas C-V típicas para distintas frecuencias de una estructura MOS fabricada sobre un sustrato de alta movilidad (InGaAs), mostrando los efectos de estados de interfaz y las trampas de frontera. Reproducido de [54] 16

2.4 Comparación entre los cambios inducidos por la frecuencia de medición, en las curvas C–v y G–V. Reproducido de [51] 18

2.5 Mecanismos de conducción a través de dieléctricos: (a) Emisión Termoiónica o Emisión de Schottky, (b) Emisión de Poole-Frenkel, (c) Túnel de Fowler-Nordheim, (d) Túnel Directo y (e) Túnel Asistido por Trampas. La esfera verde opaca representa la localización inicial del portador y la esfera traslucida la final. Se ha supuesto SiO_2 como aislante, pero los conceptos se hacen extensivos a otros dieléctricos. 21

2.6 *Tests* de degradación y ruptura para las estructuras MOS. Reproducido de [67] 23

2.7 (a) Múltiples eventos (le ruptura del tipo *HARD*) en una misma muestra y una interpretación en términos de la formación de contactos localizados. (b) Características I-V típicas correspondientes a los modos de conducción *SOFT* y *HARD*. Reproducido de [67] 24

2.8 (a) Modelo esquemático de la teoría de percolación: las esferas (defectos) se generan aleatoriamente hasta formar un camino percolativo entre los electrodos, atravesando el dieléctrico. Reproducido de [27]. (b) Confiabilidad de óxidos delgados representada en un diagrama de Weibit en términos de la teoría de percolación. Reproducido de [24] 25
2.9 (Arr.) Estructura típica de una DNN. A modo de ejemplo se presenta un red con 4 entradas, 3 capas ocultas (con 6, 6 y 4 neuronas, respectivamente) y 3 salidas. Cada neurona se vincula con todas las neuronas de la capa anterior y posterior, a través de sinapsis artificiales. (Aba. Izq.) Representación gráfica de una neurona y sinapsis biológicas, donde el axón es responsable de posibilitar la sinapsis con otra neurona. (Aba. Der.) Representación de una neurona artificial. La misma realiza una suma ponderada de los impulsos recibidos de las neuronas de la capa anterior, y emite una señal proporcional a las neuronas de la capa siguiente. 27

3.1 Estructura de las muestras utilizadas en este estudio. El espesor relativo de los films que forman la bi-capa aislante son (de izquierda a derecha): 0 %-100 %, 30 %-70 %, 40 %-60 % y 100$-0 %. Nótese que las figuras no están a escala. 38
3.2 Curvas MFCV medidas entre 200 Hz y 1 MHz para muestras de InGaAs (arriba) e InP (abajo) con una proporción de Al_2O_3-HfO_2 de: 0 %-100 % (a) y (f), 60 %-40 % (b) y (g), 70 %-30 % (c) y (h), 90 %-10 % (d) y (i), 100 %-0 % (e) y (j). Las lineas de trazo discontinuo azul y negra indican en cada caso, la capacidad mínima en inversión y la tensión V_{FB}, respectivamente. 39
3.3 Curvas de estrés C–V a una única frecuencia (SFCV) resultantes de variar (a) V_{start} y (b) V_{stress}. 40
3.4 ΔV_{FB} y ΔV_{Hys} medidos para las muestras bi-capa, tanto para el estrés negativo como positivo. ΔV_{FB} para InGaAs bajo estrés negativo (a) y positivo (b). ΔV_{Hys} para InGaAs bajo estrés negativo (c) y positivo (d). ΔV_{FB} para InP bajo estrés negativo (e) y positivo (f). ΔV_{Hys} para InP bajo estrés negativo (g) y positivo (h) 41

3.5 Curvas C–V para las muestras bajo estudio, medidas a 200 kHz. (a) Muestras sin tratamiento de FGA (no-FGA, control) y (b) muestras con tratamiento de FGA (FGA). En cada gráfico, la linea de trazos indica la tensión de *flat-band*, calculada de acuerdo con la técnica del punto de inflexión. En el *inset* de (a) se muestra el D_{it} para las muestras de HfO_2 calculado mediante *i)* el método de Alta Frecuencia-Baja Frecuencia y *ii)* el método de la conductancia paralelo. En ambos casos, el D_{it} es mayor para las muestras de control (no-FGA). (c) y (d) muestran las curvas C–V medidas a múltiples frecuencias (200Hz–300 kHz) (MFCV) para las estructuras de HfO_2, antes y después del tratamiento de FGA. Nótese que hay una clara reducción tanto de la dispersión del valor de capacidad medido en acumulación así como del llamado "*weak inversion hump*", lo cual es consistente con la reducción observada del D_{it} 45

3.6 Mediciones de histéresis en curvas C–V para una única frecuencia de prueba (200 kHz). Las muestras de control (no-FGA) estresadas a tensión negativa y positiva se muestran en (a) y (b), respectivamente. De la misma forma en (c) y (d) se muestran las muestras sometidas al tratamiento de FGA estresadas a tensiones negativa y positiva, respectivamente. En el *inset* de (d) se muestra la evolución de ΔD_{it} calculado en el *mid-gap* en función de las tensiones de estrés V_{stress} y V_{start} respecto de la muestra sin estresar. Las tensiones de V_{stress} y V_{start} se indican en (a) y (b) para los experimentos de estrés a tensión negativa y positiva, respectivamente. . . 47

3.7 La evolución de la tensión de estrés para los experimentos ilustrados en la Fig. 3.6 se muestra en (a) para el caso de estrés a tensión negativa. Se indica el periodo de estrés y los barridos C-V de inversión a acumulación y *vice versa*, así como las tensiones de V_{stress} y V_{start}. Procedimiento de medición utilizado para el estrés de larga duración cuyos resultados se reportan en la Fig. 3.10. $V_G - V_{FB}$ se mantiene constante por un periodo de $\sim$30 seg. de forma de polarizar el dispositivo en acumulación, siendo periódicamente interrumpido para evaluar V_{FB} y V_{Hys}. 48

3.8 ΔV_{FB} y ΔV_{hys} para diferentes muestras multi-laminadas para tensiones de estrés negativas y positivas (se reduce V_{stress} y se aumenta V_{start}, respectivamente). El corrimiento de V_{FB} (ΔV_{FB}) para muestras con y sin tratamiento de FGA estresadas a tensión (a) negativa y (b) positiva. La variación de la histéresis de C–V (ΔV_{hys}) para muestras con y sin tratamiento de FGA estresadas a tensión (c) negativa y (d) positiva. Las lineas de trazos roja y azul indican en (a) la máxima variación de V_{FB} en el rango de -3.5 a -1.5 V (zona sombreada) para las muestras no-FGA y FGA, respectivamente. 49

3.9 Mediciones I–V para las distintas muestras consideradas: (a) muestras no-FGA (control), (b) muestras FGA. Se observa una nivel de corriente apenas mayor para el caso de las muestras no-FGA 49

3.10 Impacto del tratamiento de FGA en el corrimiento de (a) V_{FB} y (b) la histéresis de C–V (V_{hys}) para las diferentes estructuras MOS de Ge consideradas, bajo condiciones de estrés prolongadas (30 seg.) Se puede ver como el tratamiento de FGA reduce tanto ΔV_{FB} como V_{hys} en todos los casos. 50

4.1 (Arriba) Sección final del micro-haz de iones pesados del acelerador lineal TANDAR de la CNEA-CAC. (Abajo). Comparación entre el tamaño del *spot* del micro-haz y un circuito CMOS analógico Full-Custom fabricado en un proceso de longitud nominal de 180 nm. Cada punto indica una posición en la que el haz fue enfocado. Reproducido de [149] 59

4.2 Simulación en SRIM. Se muestra el número de vacancias promedio creado por cada ion incidente en el dieléctrico por unidad de material atravesado (Å). Nótese que se genera aproximadamente el mismo número de vacancias de Hafnio que de Oxígeno. 60

4.3 Diagrama del dispositivo bajo estudio, irradiado con iones de Carbono (C^{+4}), con una energía de 40 MeV. (1) Defectos generados en la capa dieléctrica, (2) caminos filamentarios parcialmente formados, generados en la capa dieléctrica, (3) estados de interfaz y (4) iones implantados en el sustrato. . . 61

4.4 Mediciones C–V a múltiples frecuencias: (a) Muestra de control (no irradiada) (b) muestras sometida a la máxima fluencia de irradiación. La resistencia serie aumenta en función de la fluencia debido a los iones implantados en el sustrato (véase la Fig. 4.3). Las mediciones fueron corregidas para descartar el efecto de la misma. A pesar del pequeño incremento en el "*weak-inversion hump*", el D_{it} calculado mediante el método de la conductancia paralelo no muestra variaciones significativas en función de la fluencia. Los *inset* mostrados en (a)-(c) muestran el detalle de la conductancia paralela normalizada en cada caso. Nótese que el pequeño incremento de G_P está directamente relacionado al leve aumento de D_{it}. 62

4.5 Mediciones I–V para diferentes fluencias de radiación. La corriente de fuga aumenta en función de la fluencia en el régimen de bajo campo eléctrico, y se mantiene aproximadamente constante para campos más elevados (Indicando que no existe SBD para las muestras bajo estudio). 63

4.6 Mediciones CVS realizadas sobre las muestras (a) F#, (b) S#3 y (c) L#3. Independientemente de la fluencia de irradiación utilizada, todas las muestras exhiben un una dinámica de ruptura progresiva. (d) Representación de I_{Init}, I_{SBD} y I_{HBD} durante la evolución de la corriente de fuga. También se indica la tasa de degradación (DR) = $\Delta I/\Delta t$, donde ΔI=$I_{HBD} - I_{SBD}$ y $\Delta t = t_{HBD} - t_{SBD}$. (e) Los valores calculados de DR se grafican para todas las fluencias de radiación, tanto para el área pequeña como para el área grande (Los datos de las muestras S#2, S#3, L#2 y L#3 están levemente desplazados horizontalmente por claridad). La corriente en el momento previo al HBD y la corriente inicial se grafican para cada combinación de área y fluencia. El valor de I_{Init} concuerda con las mediciones $I - V$ en todos los casos. I_{HBD} está en el rango de 1-10 μA. (f) muestra de control (sin irradiar) -F#-, (g) muestras L#1, (h) L#2 y S#2 e (i) muestras L#3 y S#3. 65

4.7 Distribuciones de TDDB medido para la ruptura abrupta (HBD) para cada fluencia de irradiación (25 dispositivos en cada caso). Los datos fueron obtenidos mediante experimentos de CVS a $V_G - V_{FB} = 2{,}4V$. (a) Irradiación de área grande y (b) irradiación de área pequeña. Los datos en (a) y (b) fueron ajustados utilizando el modelo de *Clustering*. La comparación entre el ajuste usando el modelo de Weibull y el modelo de *Clustering* se muestra en cada caso en las figuras (c) a (f), donde la concavidad "hacia abajo" se vuelve evidente. Los parámetros de ajuste, es decir el factor de *Clustering* (α_C círculos azules), pendiente de Weibull (β, diamantes rojos) y $t_{63\,\%}$ (tiempo de vida medio, cuadrados azules) se muestran en (g) para el caso del modelo de *Clustering* y (h) para la distribución de Weibull. Las lineas de trazo indican las irradiaciones de área pequeña y las continuas la irradiación de área grande. 66

4.8 (a) Variaciones típicas del espesor en lineas metálicas de interconexión en el BEOL (*Back-End Of Line*). (b) Incremento localizado del campo eléctrico entre lineas metálicas en el BEOL debido a los cambios en el espesor de las lineas metálicas. Reproducido de [162] 67

4.9 Representación esquemática de los diferentes escenarios de simulación. A tiempo $t = 0$, hay solamente 2 distribuciones diferentes: *i)* defectos distribuidos uniformemente y *ii)* defectos agrupados eh caminos filamentarios parcialmente formados. Luego de un tiempo de estrés t_{stress}, y considerando dos mecanismos diferentes para la generación de nuevos defectos (espacialmente correlacionado y no-correlacionado), surgen 4 escenarios posibles. Caso A: distribución de defectos uniforme con generación no-correlacionada, Caso B: distribución de defectos uniforme con generación correlacionada, Caso C: caminos filamentarios parcialmente formados con generación no-correlacionada y Caso D: caminos filamentarios parcialmente formados con generación correlacionada. 72

4.10 Tendencias observadas de la pendiente de Weibull (β) obtenidas mediante simulaciones realizadas utilizando la plataforma Ginestra^{TM}, en función de la densidad inicial de defectos (la cual es ajustada artificialmente mediante la fluencia de irradiación en el estudio experimental), correspondientes a los Casos (a) A, (b) B y (c) D definidos en la Fig. 4.9. Para cada caso, se incluye una representación gráfica de los defectos distribuidos al momento de HBD, generada a partir de una simulación tomada al azar. Nótense las significativas variaciones en el factor de *clustering* (generación no-correlacionada o correlacionada) y del número de defectos necesarios para alcanzar el estado de HBD. 73

5.1 Comparación entre (a) *Memory Type RS* y (b) *Threshold Type RS*. La parte superior de cada sub-figura muestra una representación esquemática del fenómeno mientras que en la parte inferior se presentan ejemplos experimentales consistentes en una estructura de (TiO_X amorfo)/(nano-partículas de Ag)/(TiO_X poli-cristalino). Adaptado de [190] y [191]. 78

5.2 Mediciones I–V para las muestras bajo prueba (50 curvas en total). La línea continua roja indica la curva promedio, mientras que la azul de trazos representa el 1er ciclo. En el *inset* de la derecha se muestra el arreglo 1T1R, formada por el dispositivo RRAM y el correspondiente transistor N-MOS de control. 79

5.3 Evolución de la corriente en función del tiempo para mediciones CVS. Por claridad solo se muestran los casos correspondientes a la (a) mínima —450 mV— y máxima —650 mV— tensión. Los marcadores 1, 2 y 3 indican la corriente inicial (I_{init}), el inicio del aumento progresivo de la corriente (I_{on}) y el momento del salto a la corriente límite (I_{end}), para las mediciones CVS realizadas a (c) 450 mV, (d) 500 mV, (e) 550 mV, (f) 600 mV y (g) 650 mV. En las figuras (c)-(g), C_i indica el número de ciclo. (h) Tasa de transición (*Transition Rate*, TR=dI_{Tr}/dt) de las muestras bajo estudio. El valor medio de TR junto con ∼100 valores medidos se reporta para cada tensión de estrés (450, 500, 550, 600 y 650 mV). Nótese que el valor medio se incrementa en aproximadamente 1 orden de magnitud cada 50 mV. 80

5.4 Estructura y características eléctricas de los dispositivos Ag/h-BN/Au bajo estudio. (a) Imagen TEM de sección transversal de la multi-capa de h-BN crecido mediante CVD sobre el sustrato de Cu, el cual está cubierto por un film delgado de Au/Ti, demostrando la existencia de defectos nativos en el h-BN. (b) Imagen SEM de una estructura Ag/h-BN/Au. Nótese la disposición perpendicular de los electrodos, dando lugar a una estructura tipo *cross-point*. (c) Ciclos de RS volátil obtenidos sobre un dispositivo Ag/h-BN/Au con una limitación de corriente de 1 μA. (d) Ciclos de RS no-volátil obtenidos sobre el mismo tipo de dispositivos pero con una limitación de corriente de 10 μA. Las regiones sombreadas en (c) indican el SET (aprox. entre 0.5 V y 0.6 V) y RESET (aprox. entre 0.2 V y 0.3 V) durante las rampas de tensión creciente y decreciente, respectivamente. En forma análoga, se reporta en (d) el caso de la tensión de SET entre 0.5 V y 0.7 V aprox. (nótese que en este caso no se observa el RESET) 83

5.5 Mediciones PVS de la estructura Ag/h-BN/Au realizadas a diferentes tensiones V_E. (a) Secuencia de pulsos con V_E=2 V mostrando la transición del nivel de corriente entre HRS a LRS cuando cada pulso es aplicado y la relajación una vez removida la tensión de estrés. Para cada valor de V_E se han recogido entre 20 y 30 pulsos. Nótese que durante los pulsos de lectura, la corriente se mantiene en el piso de ruido, indicando una relajación completa. (b) Detalle de un pulso (trazo rojo) para el caso de V_E=2,5 V (trazo azul). Se indican 3 puntos, correspondientes a (1) el inicio de la tensión de estrés, (2) el momento en que la corriente a través del dispositivo comienza a aumentar y (3) el punto en el cual se estabiliza. TR se define entre los puntos (2) y (3). (c) Ídem (b) para el caso de V_E=4 V. (d) Los valores obtenidos de TR para cada uno de los valores de V_E (aprox. 20-30 puntos para cada tensión) muestran una clara dependencia con la tensión V_E. 84

5.6 Representación esquemática de la reducción de t_{ox} debido a la aparición de una discontinuidad en el CF. (a) estructura MIM antes del electro-formado, (b) estado LRS y (c) estado HRS. Las esferas azules representan las especies atómicas de los electrodos, mientras que las rojas las vacancias de oxigeno (VOs). (d) Modelo de I_{SET} descrito en la Ec. 5.2 comparado con las mediciones experimentales, para el caso de HfO_2. (e) Cálculo de la temperatura en función de la tensión aplicada y de la energía perdida por los portadores en la constricción del CF para el caso de HfO_2. Tanto en (d) como en (e) la zona sombreada en rojo indica el rango de tensiones utilizadas en los experimentos de CVS (0.45 a 0.65 V). (f) y (g) presentan los resultados correspondientes a los dispositivos de h-BN. 86

5.7 Distribución del tiempo al SET obtenida para los experimentos de CVS realizados, considerando ∼100 mediciones para cada escenario. Los símbolos indican datos experimentales y las lineas el ajuste utilizando la distribución de Weibull. (b) Temperatura en la constricción del CF vs. la tensión de estés, el t_{ox} efectivo y la pérdida de energía en la constricción (f_2). La zona sombreada en rojo indica el rango de tensiones utilizadas en los experimentos de CVS (0.45 a 0.65 V). 88

5.8 Difusividad de las especies atómicas presentes en la estructura vs. el recíproco de la temperatura. (a) La difusivdad de las VOs [210] es $\sim 10^4$ veces más alta que aquella observada en iones metálicos [18]. E_{act} está en el rango de 0.3 a 0.7 eV. Se puede apreciar que la difusividad D requerida para ajustar los datos experimentales de TR está en el mismo rango de la difusividad de las VOs. (b) Difusividad de las especies involucradas en el SET en memorias volátiles (h-BN). Nótese que el mejor ajuste obtenido para los datos presentados en este capítulo presenta una energía de activación sustancialmente menor a otros reportados en la literatura, lo cual podría explicarse por los diferentes medios considerados para la difusión. 90

5.9 Ajuste de los datos experimentales de TR para (a) HfO_2 (los círculos indican el valor medio mientras que los cuadrados distintos valores medidos) asumiendo la difusión de VOs y un t_{ox} efectivo igual a t_{gap} (linea de trazo color cian —curva n°1—). Adicionalmente se muestra TR vs. tensión aplicada utilizando combinaciones alternativas de D y t_{ox} reportadas en la literatura —curvas n° $2,\ 3$ y 4—. TR aumenta casi 1 orden de magnitud cada 50 mV, con un valor medio de 2x10^{-3} A/seg, $1{,}4$x10^{-2} A/seg, $2{,}1$x10^{-1} A/seg, $1{,}8$x10^{0} A/seg y 1x10^{1} A/seg para tensiones aplicadas de 450 mV, 500 mV, 550 mV, 600 mV y 650 mV, respectivamente. (b) muestra los resultados correspondientes a los dispositivos de h-BN. Nótese que se observa un ajuste igualmente aceptable donde se han considerado 3 curvas I-V diferentes, resultando en un $TR(V)$ máximo, típico y mínimo. 92

6.1 (a) Esquema de la estructura del CPA. Las flechas rojas y azules indican el flujo de corriente a través de los memdiodos que conectan las lineas superiores (*Word Lines*, WL) e inferiores (*Bit Lines*, BL). Diferentes estados de conducción son representados esquemáticamente (HRS y LRS). La linea de azul de trazo discontinuo ejemplifica el denominado problema de *sneak-path*. La resistencia serie parásita de las lineas de conexión se indica tanto para WL_i y BL_i. (b) Modelo del histerón con las funciones logísticas Γ^+ (Ec. 6.3) y Γ^- (Ec. 6.4). Ω es el espacio de estados posibles S. Las lineas negras de trazo discontinuo superpuestas al modelo del histerón indican la trayectoria de la variable de estado λ dentro de Ω desde un estado inicial S_1 hasta un estado final S_2. El *inset* izquierdo muestra la representación circuital de la ecuación de transporte (Ec. 6.1) incluyendo la resistencia serie. Las propiedades de conducción de cada diodo están determinadas por el estado de memoria del dispositivo y solo un diodo se encuentra activo en un instante t. La característica I–V típica del memdiodo obtenida mediante simulación del modelo propuesto se muestra superpuesta al histerón. La evolución de la corriente se indica mediante las flechas azules. El *inset* de la derecha muestra la transición de una característica exponencial en HRS a lineal en LRS, mediante la variación de λ. La región sombreada en rojo indica el rango de posibles tensiones aplicadas al dispositivo. Las corrientes I_{HRS} e I_{LRS} a la tensión de ajuste se indican mediante los símbolos gris y blanco, respectivamente. Nótese que puede existir una sobre-estimación de la corriente I_{HRS} cuando se considera un modelo lineal para el régimen de HRS y se utilizan tensiones mucho más bajas que la utilizada para el ajuste, como indican los símbolos cían, azul y negro. 99

6.2 Curvas I–V experimentales para diferentes materiales reportados en la literaturas, ajustadas con el modelo QMM: (a) HfO_2 [271], (b) Al_2O_3 [272], (c) MnO_3 [273], (d) CuO_2 [274], (e) $La_{1-x}Ca_xMnO_3$ [275] y (f) TaO_X [269]. Los parámetros de ajuste del modelo QMM se muestran para cada caso. Como referencia, las curvas HRS y LRS se indican en (a) y los eventos de SET y RESET en (b). Nótese que en (a) se impuso una limitación de corriente de 200 μA para prevenir la ruptura dieléctrica permanente, la cual es adecuadamente representada por el QMM. 103

6.3 (a) Representación esquemática del proceso de Escritura–Verificación utilizado para programar los dispositivos del CPA. La forma de onda superior representa los pulsos alternados de escritura y verificación, mientras que la inferior da cuenta de los cambios de conductancia asociados. Una representación simplificada del circuito utilizado para este proceso se muestra en el *inset* de la derecha. (b) Curvas experimentales de RESET ajustadas mediante la utilización del modelo QMM para un dispositivo de SiO_X (Datos experimentales reportados por el *University College London* (UCL) en [276]). Nótese el control sobre los estados intermedios. En el *inset* se muestra la señal de tensión aplicada. 104
6.4 Diagrama de flujo del procedimiento de entrenamiento, modelado circuital y simulación. Partiendo del tamaño de las imágenes de la base de datos, R_L, V_{read} y el esquema de conexionado, el set de rutinas de MATLAB escala la base de datos, entrena la ANN (SLP o MLP, en el diagrama se indica SLP solo por simplicidad), la traduce a nivel circuital, agrega la electrónica de control necesaria, realiza las simulaciones y procesa los resultados. Las tareas realizadas en MATLAB están agrupadas por el recuadro azul y las operaciones de SPICE por el recuadro verde. 105
6.5 (a) Circuito equivalente simplificado de un MLP. Cada uno de los dos CPA (positivo y negativo) de la 1^{er} capa sináptica está dividido en N particiones idénticas para minimizar las caídas de tensión en las resistencias parásitas de las lineas de interconexión. En la salida de cada partición, se indica el resultado parcial de la multiplicación vector-matriz efectuada en cada bloque. (b) Circuito esquemático equivalente de un CPA. Las flechas rojas y azules ejemplifican el flujo de corriente a través de los memristores conectando *wordlines* y *bitlines*. (c) Celda RRAM individual con la correspondiente resistencia R_L. 109
6.6 Costo computacional (tiempo de simulación y uso de memoria RAM) de la simulación de *crossbars* realizados en base al modelo QMM y comparado con los modelos propuestos por Yakopcic [252], Laiho [302] y la universidad de Michigan [303]. (a) Tiempo de simulación y (b) Uso total de RRAM en función del tamaño del CPA (medido en términos del número de dispositivos). El tiempo de simulación y el uso de memoria RAM se reportan también normalizados respecto del caso completamente lineal en (c) y (d) respectivamente, indicando que la simulación mediante el modelo QMM permite modelar con gran precisión las características I–V sin causar un aumento de la complejidad computacional. 110

6.7 Ajustes (a) $A1$–$A4$, (b) $B1$–$B4$ y (c) $C1$–$C4$ del modelo QMM. (d) Pérdida de legibilidad de las imágenes de la base de datos del MNIST al ser reescaladas. Precisión de inferencia en función de la característica (e) R_{ON} y (f) R_{OFF} del ajuste del modelo QMM. Se observa un mayor impacto de R_{ON} en la pérdida de precisión. (g) El cociente R_L/R_{ON} pone de manifiesto la dependencia de la precisión de inferencia con la resistencia de linea R_L, y como esta empeora con el tamaño o resolución de la imagen. Asimismo se puede observar que una relación R_{OFF}/R_{ON} mayor a 100 es necesaria para minimizar la sensibilidad a las variaciones del tipo dispositivo a dispositivo. (i) No obstante, deben evitarse resistencias R_{OFF} o R_{ON} de alto valor ya que las mismas implican corrientes sumamente bajas que comprometen el SNR y por consiguiente la precisión. Por último, se muestra la dependencia de la precisión con el tamaño de la imagen, de donde se ve como la relación de compromiso entre legibilidad y caída parásita de tensión, resulta en un tamaño óptimo de CPA 113

6.8 Precisión de inferencia en función de R_L, normalizada respecto de la precisión obtenida para el caso de $R_L \longrightarrow 0\Omega$. Se consideran dos grupos diferentes de MLPs (#a y #b, véase la Tabla 6.3) así como el caso del SLP. El *inset* indica la precisión de inferencia obtenida para $R_L \longrightarrow 0\Omega$ para cada uno de los casos considerados. Nótese que la dependencia con R_L está determinada por el tamaño de la capa sináptica más grande en el MLP, y presenta muy poca sensibilidad al número de capas ocultas. De hecho, agregando más capas ocultas es posible incrementar notoriamente la precisión, sin que esto implique una mayor degradación producida por R_L. 119

6.9 El cambio en (a) la distribución de elementos de W_M según las distintos métodos de normalización se muestra en (b)-(f). La precisión de inferencia en función del porcentaje de dispositivos defectuosos considerando distintos métodos de normalización se presenta para fallas del tipo (g) SA1, (h) SA0 y (i) SA0_nE. La potencia disipada en el SLP durante la fase de inferencia se indica en el *inset* de (h) en función del tamaño del SLP. Similarmente, la precisión de inferencia obtenida en el SLP libre de fallas se muestra para cada método de normalización en el *inset* de (i). Por otro lado, la precisión de inferencia en función del porcentaje de dispositivos defectuosos se presenta en (j)-(l) para las fallas de tipo SA1, SA0 y SA0_nE, teniendo en cuenta cuatro valores distintos de R_L. (m) Precisión de inferencia obtenida con el SLP en función del porcentaje de dispositivos defectuosos polarizados. Cada símbolo corresponde a una ejecución de las simulaciones de Monte-Carlo. Los datos se codifican en términos del porcentaje total de dispositivos defectuosos (tipo de símbolo) y tamaño del SLP (color del símbolo). Por ejemplo, los círculos azules corresponden a los resultados de inferencia para simulaciones de SLPs de 64×10 con un 1 % de dispositivos defectuosos. El caso puntual de SLPs con un 30 % de dispositivos defectuosos (pentágonos) han sido señalados con el fin de ejemplificar la reducción del porcentaje de dispositivos defectuosos polarizados a medida que aumenta el tamaño del SLP (∼5 % en el SLP de 1280 dispositivos (imágenes de 8×8), ∼3.7 % con 5120 dispositivos (imágenes de 16×16) y finalmente ∼2.3 % con 15680 dispositivos (imágenes de 28×28). 122

6.10 (a) Representación esquemática del algoritmo de re-mapeo número 1, describiendo la compensación de conductancias que permite tolerar fallas en la primer y última filas (celdas coloreadas en verde) pero que es incapaz de manejar otras SAFs (celdas coloreadas en gris, las cuales son irrecuperables). Para solucionar este inconveniente, se propone la permutación de filas (abajo) para transformar fallas irrecuperables en recuperables (véase la Tabla 6.5. (b) La permutación de filas también es usada en los Algoritmos 2 y 3. En el último esta se utiliza para re-mapear las filas con más dispositivos defectuosos a los píxeles menos activos. 126

6.11 (a) Muestra de las imágenes de la base de datos de rostros de la Universidad de Yale, mostrando 3 clases diferentes con resoluciones de 32×32 px. (arriba) y 16×16 px. (abajo). En ambos casos los ejes x e y en las imágenes del extremo izquierdo indican la posición de cada píxel. Los algoritmos de re-mapeo 1–3 fueron puestos a prueba tanto con las imágenes de la base de datos del MNIST como con las presentadas en (a). Las tendencias correspondientes a la inyección de fallas del tipo SA1 y SA0 se muestran para las primeras (MNIST) en las figuras (b) y (c) respectivamente, mientras que para el caso de las segundas (Universidad de Yale) se muestra en las figuras (d) y (e). En ambos casos el Algoritmo 1 muestra los mejores resultados al considerar fallas del tipo SA1, mientras que para fallas del tipo SA0, es conveniente el Algoritmo 2. 129

B.1 Imágenes de ejemplo de las tres bases de datos estilo MNIST consideradas en este trabajo. Para todos los casos, las imágenes tienen una resolución de 28×28 píxeles. El brillo de cada píxel está codificado en 256 niveles entre 0 (completamente apagado, y por ende negro) y 1 (completamente prendido y por ende blanco). (a) Base de datos MNIST de dígitos manuscritos. (b) Base de datos MNIST-F [278] de artículos de vestir. (c) Base de datos MNIST-K [279] de ideogramas Kanji japoneses manuscritos. 142

B.2 Validación cruzada de 5 grupos (*5-fold cross-validation*) y 10 repeticiones, para los 11 algoritmos de aprendizaje considerados [289]-[296]. La precisión de inferencia obtenida se grafica en función del tiempo de CPU requerido por el algoritmo, para las 3 bases de datos (MNIST, MNIST-F y MNIST-K) y dos resoluciones diferentes (imágenes de 8×8 y 28×28 px.). La precisión promedio de cada algoritmo se reporta en el encabezado de las Tablas B.2-B.7 para cada par Base de Datos - Resolución: imágenes de 8×8 px. de las bases (a) MNIST, (b) MNIST-F y (c) MNIST-K, e imágenes de 28×28 px. de las bases (d) MNIST, (e) MNIST-F y (f) MNIST-K. Si bien LM muestra la mayor precisión, es también el más lento, especialmente para redes de gran tamaño, como las requeridas para imágenes de 28×28 px. Los test de significación estadística reportados en las Tablas B.2-B.7 indican con un 95 % de confianza que para de $|z|$ >1,96 los valores de precisión obtenidos son estadísticamente diferentes (celdas sombreadas en gris). Asumiendo una relación de compromiso entre precisión y tiempo de aprendizaje, se elige SCG por sobre los demás, ya que la diferencia de precisión con LM no es estadísticamente relevante (las diferencias podrían deberse a fluctuaciones en los datos de entrada). 143

B.3 Métricas de inferencia complementarias para el reconocimiento de imágenes de la base de datos del MNIST, en función de R_L y tamaño del CPA (# dispositivos): (a) y (f) Sensibilidad (*Recall*), (b) y (g) Especificidad, (c) y (h) Precisión, (d) y (i) F1-*Score* y (e) y (j) coeficiente-κ. 146
B.4 (a) Impacto de la resistencia de linea (R_L) en la precisión de inferencia para la base de datos MNIST-F, considerando los ajustes $C1$-$C4$ del QMM. (b) La precisión de inferencia se grafica en función del cociente R_L/R_{ON} mostrando una tendencia unificada entre todos los ajustes. Matrices de confusión para el ajuste $C2$ con R_L igual a (c) 0,1 Ω, (d) 4,53 Ω (16 nm [233]) y (e) 81,3 Ω (10 nm [306]). En todos los casos, el CPA es conectado por ambos lados (DSC) y las imagenes no se re-escalan (resolución de 28×28 px.). Otras metricas de inferencia incluyen: (f) Sensibilidad, (g) Especificidad, (h) Precisión, (i), F1-*Score* y (j) coeficiente–κ. 147
B.5 (a) Pérdida de legibilidad de las imágenes de la base de datos MNIST-F para resolución decreciente de 28×28 px. (caso I) a 8×8 px. (caso IV). (b) Impacto del tamaño del CPA (cantidad de dispositivos) sobre la precisión de inferencia para el ajuste $C2$, y R_L barrido entre 1 y 100 Ω. Dos esquemas de particionado diferentes fueron utilizados: P1 indica *arrays* no particionados y P4 que cada CPA fue dividido en 4 sub-*arrays*. Los símbolos triangulares indican los puntos de tamaño máximo y precisión máxima, mostrando que los CPA particionados permiten mayor precisión en *arrays* más grandes. (c) Precisión de inferencia en función del cociente R_L/R_{ON} para imágenes de 28×28 (dos *arrays* de 784×10 arrays, ∼15.6k disp., símbolos vacíos) y 8×8 (dos 64×10 *arrays*, ∼1.2k disp., símbolos llenos). Matrices de confusión para el ajuste $C2$ y R_L=10 Ω para imágenes de (d) 28×28 px., (e) 20×20 px. y (f) 8×8 px. Otras métricas para el el reconocimiento de imágenes incluyen (g) sensibilidad, (h) especificidad, (i) precisión, (j) F1-Score y (k) coeficiente-κ.148
B.6 (a) Impacto de la resistencia de linea (R_L) en la precisión de inferencia para la base de datos MNIST-K, considerando los ajustes $C1$-$C4$ del QMM. (b) La precisión de inferencia se grafica en función del cociente R_L/R_{ON} mostrando una tendencia unificada entre todos los ajustes. Matrices de confusión para el ajuste $C2$ con R_L igual a (c) 0,1 Ω, (d) 4,53 Ω (16 nm [233]) y (e) 81,3 Ω (10 nm [306]). En todos los casos, el CPA es conectado por ambos lados (DSC) y las imagenes no se re-escalan (resolución de 28×28 px.). Otras metricas de inferencia incluyen: (f) Sensibilidad, (g) Especificidad, (h) Precisión, (i), F1-*Score* y (j) coeficiente–κ. 149

B.7 (a) Pérdida de legibilidad de las imágenes de la base de datos MNIST-K para resolución decreciente de 28×28 px. (caso I) a 8×8 px. (caso IV). (b) Impacto del tamaño del CPA (cantidad de dispositivos) sobre la precisión de inferencia para el ajuste $C2$, y R_L barrido entre 1 y 100 Ω. Dos esquemas de particionado diferentes fueron utilizados: P1 indica *arrays* no particionados y P4 que cada CPA fue dividido en 4 sub-*arrays*. Los símbolos triangulares indican los puntos de tamaño máximo y precisión máxima, mostrando que los CPA particionados permiten mayor precisión en *arrays* más grandes. (c) Precisión de inferencia en función del cociente R_L/R_{ON} para imágenes de 28×28 (dos *arrays* de 784×10 arrays, ∼15.6k disp., símbolos vacíos) y 8×8 (dos 64×10 *arrays*, ∼1.2k disp., símbolos llenos). Matrices de confusión para el ajuste $C2$ y R_L=10 Ω para imágenes de (d) 28×28 px., (e) 20×20 px. y (f) 8×8 px. Otras métricas para el el reconocimiento de imágenes incluyen (g) sensibilidad, (h) especificidad, (i) precisión, (j) F1-Score y (k) coeficiente-κ. 150

1 Introducción

En la actualidad, los sistemas electrónicos integrados son parte fundamental de una gran variedad de aplicaciones, representando por lo tanto el corazón de las tecnologías de la información y la comunicación, tal como lo estipula el plan Argentina Innovadora 2020 [1] al proponer a ésta como un área de interés estratégico. Dicho paradigma es una consecuencia de la agresiva reducción en las dimensiones del bloque constructivo fundamental de los circuitos electrónicos, el transistor de efecto de campo (*Field Effect Transistor*, FET), lo cual ha permitido incluir un número creciente de dispositivos en cada circuito. Para ello, el desarrollo y maduración de la tecnología MOS (Metal-Óxido-Semiconductor) ha sido clave, y en esto ha tenido un rol preponderante el silicio, dado su reducido costo, abundancia y la existencia de un óxido nativo estable tal como el dióxido de Silicio (SiO_2). Tal ha sido el avance de este enfoque, que durante las últimas 4 décadas el número transistores (*Metal-Oxide-Semiconductor Field Effect Transistor*, MOSFET) integrables por unidad de área se ha duplicado aproximadamente cada 18 meses, lo cual ha sido descrito por la célebre Ley de Moore.

Si bien ésta no es una ley física, desde que fue introducida por Gordon Moore en 1975 [5] ha establecido una dirección para la industria microelectrónica e incentivado los esfuerzos que permitieron reducir el costo y potencia consumida en cada componente. De esta forma, se ha pasado de unos 50 transistores integrados mediante un proceso con transistores de 50 μm en 1965, a miles de millones de dispositivos integrados en los procesos actuales de una longitud de 10 nm, tal como se indica en la Fig. 1.1a. Durante dicha evolución, se pueden distinguir dos fases con marcadas diferencias. En primera instancia, el escalamiento "clásico" o "tradicional" fue utilizado exitosamente hasta fines de los años 90, siendo uno de los últimos nodos tecnológicos el de 130 nm. Esta metodología de escalamiento del transistor MOS fue descrita por Dennard *et al.* [6] en 1974 y permitió producir transistores de menor área, capaces de conducir mayor corriente (mayor rendimiento) y menor capacidad parásita (menor potencia dinámica) mediante el escalamiento conjunto (factor de escala constante) de todas las dimensiones del dispositivo MOS, pero sin modificaciones sustanciales en la morfología ni en los materiales involucrados en el proceso. Como consecuencia de la madurez alcanzada por la tecnología MOS en este pun-

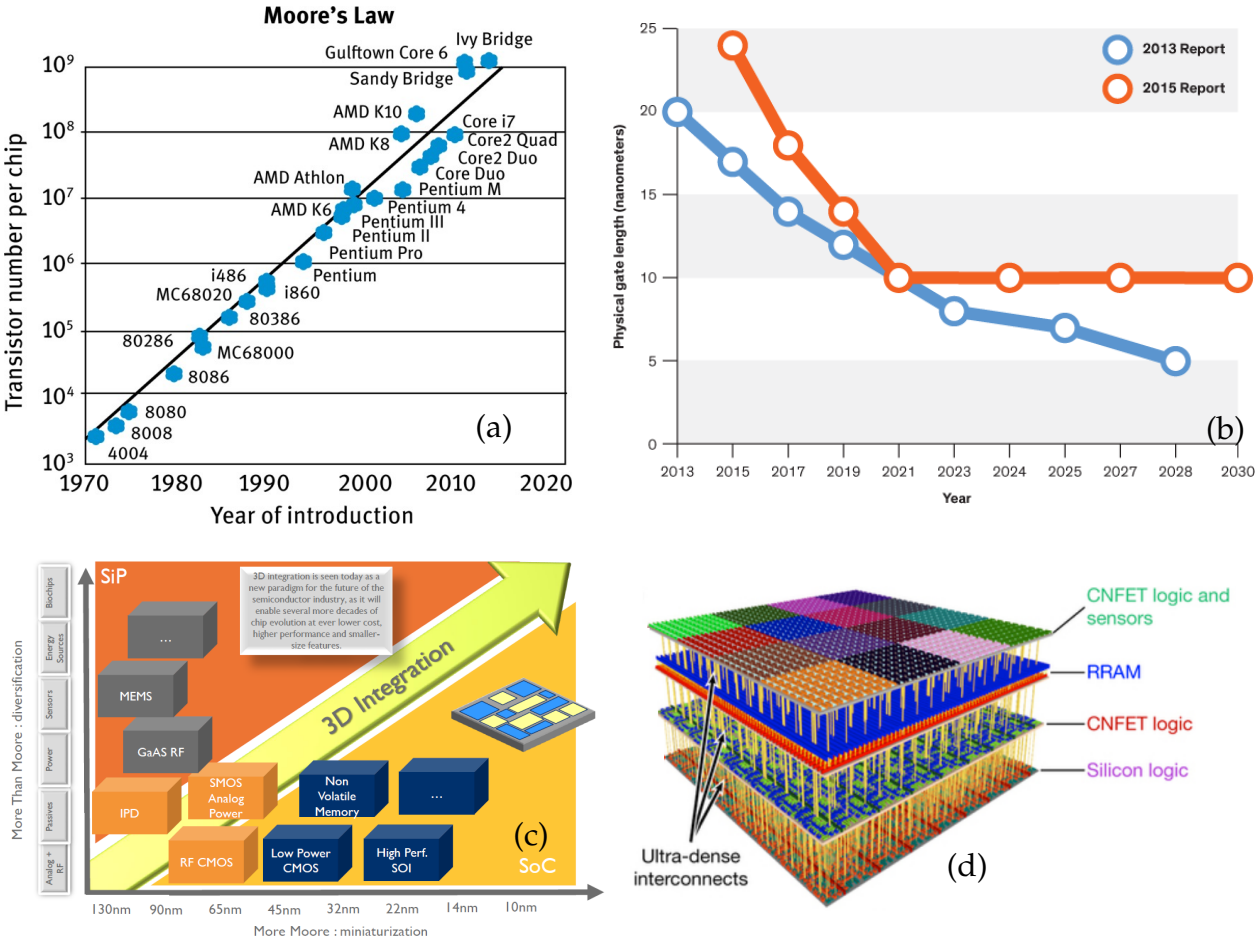


Figura 1.1: (a) Evolución de la cantidad de transistores por *chip* desde la introducción del primer circuito integrado hasta la actualidad. Reproducido de [2]. (b) Fin de la era del escalamiento pronosticado hacia el inicio de la década de los 2020. Reproducido de [3]. (c) Integración de los enfoques *More Moore* y *More than Moore* como medio para el desarrollo futuro de la microelectrónica, con la integración tri-dimensional como gran ventaja. (d) Ejemplo de integración monolítica multicapa de un sistema *In-Memory Computing*. Reproducido de [4]

to, en el año 2001 llegaron a existir 19 actores de peso abocados a la fabricación de circuitos integrados en el estado del arte [3]. A partir de ese entonces la complejidad de sortear las limitaciones impuestas por la física del estado sólido para continuar el escalamiento por debajo de los 100 nm comenzó a afectar sensiblemente la tasa de miniaturización. Es en este punto que se comienzan a introducir las primeras modificaciones morfológicas a la estructura del transistor MOS y a experimentar con nuevos materiales. En relación a las primeras, la tecnología de fabricación planar de transistores MOSFET de silicio a dado paso al desarrollo de dispositivos tri-dimensionales, como los denominados *FinFETs* o a los sustratos SOI (*Silicon-On-Insulator*, o Silicio Sobre Aislante), mientras que en el segundo caso, se han introducido nuevos materiales (por ejemplo el reemplazo del dióxido de silicio como aislante por dióxido de hafnio a partir del nodo de 45 nm) o la utilización de compresión mecánica en el sustrato de silicio para mejorar la movilidad de los portadores. Sin embargo, esta diversificación y aun mayor complejidad de los procesos asociados dispararon los costos asociados de la industria, y como resultado, en el corriente año (2021)

solo 3 firmas continúan produciendo circuitos integrados con longitudes de canal en el estado del arte. A su vez, las perspectivas del escalamiento indican que este terminará a principios de la década de los 2020, con el nodo de 7 nm [3] (véase la Fig. 1.1b).

Ante este panorama, mediante la colaboración entre industria y academia se están realizando importantes avances tecnológicos dentro de dos grandes lineas denominadas "Más Moore"(*More Moore*) y "Más que Moore"(*More than Moore*), representadas esquemáticamente en la Fig. 1.1c. Por un lado, la primera alternativa se enfoca en mantener el escalamiento de los dispositivos MOS previsto por la Ley de Moore, aumentando la cantidad de dispositivos integrados en cada *chip* (de allí que también se la denomine como enfoque *System on Chip*, SoC) mediante la combinación inteligente de nuevos materiales e interfaces. Por ejemplo, así como los aislantes de alta constante dieléctrica (denominados dieléctricos *high–*κ) tales como HfO_2 y Al_2O_3 reemplazaron al SiO_2 como dieléctrico de compuerta en FET ultra-escalados por su capacidad para reducir la corriente de fuga [7], se espera que el Germanio (Ge) y los semiconductores de los grupos III y V de la tabla periódica (Semiconductores III-V) puedan reemplazar al silicio como material de sustrato, dada su mayor movilidad de portadores [8]. Asimismo, la aparición de nuevos materiales bidimensionales (2D) con propiedades eléctricas, físicas, químicas, térmicas y ópticas superiores sugiere ahora una transición tecnológica similar [9]. Por otro lado, la segunda alternativa (también referida en la literatura como *System in Package*, SiP) aboga por empaquetar un número mayor de *chips* en un mismo encapsulado en lugar de aumentar la cantidad de transistores por *chip*. Como casos paradigmáticos de este enfoque se pueden citar la integración monolítica tri-dimensional de transistores MOS en múltiples capas, o los sistemas de "Computación en memoria" (*In-Memory-Computing*) los cuales aprovechan las nuevas tecnologías en materia de memorias no-volátiles desarrolladas recientemente y de gran interés en el ámbito de las redes neuronales o circuitos neuromórficos (véase la Fig. 1.1d).

No obstante el camino por recorrer para alcanzar la madurez tecnológica en estos dos enfoques es largo y está plagado de interrogantes. Considérese en primera instancia el enfoque SoC y la introducción de nuevos materiales a la estructura MOS. Aunque se conoce ampliamente mayor movilidad de electrones y huecos en semiconductores III-V y Germanio, respectivamente (véase la Fig. 1.2a), e incluso se ha llegado a reportar en forma exitosa la integración conjunta de transistores MOSFET tipo P y N con dieléctricos *high–*κ en un mismo *chip* utilizando Ge y materiales III-V (como InGaAs, InP, y GaN [10], [11]), respectivamente [12] (véase la Fig. 1.2b), existe un sin número de desafíos que ha postergado su introducción a nivel industrial, y al día de hoy continúan sin estar completamente solucionados. Particularmente, la interfaz semiconductor/óxido (Ge/*high–*κ o III-V/*high–*κ)[13] es un campo de intensa investigación, debido a su importante rol en los aspectos de confiabilidad. El atrapamiento de carga en el gran número de defectos existentes tanto en el óxido como en la interfaz y la gran variabilidad entre dispositivos [14], [15], particularmente si se consideran dieléctricos *high–*κ (HfO_2, Al_2O_3, nitruros, etc.,

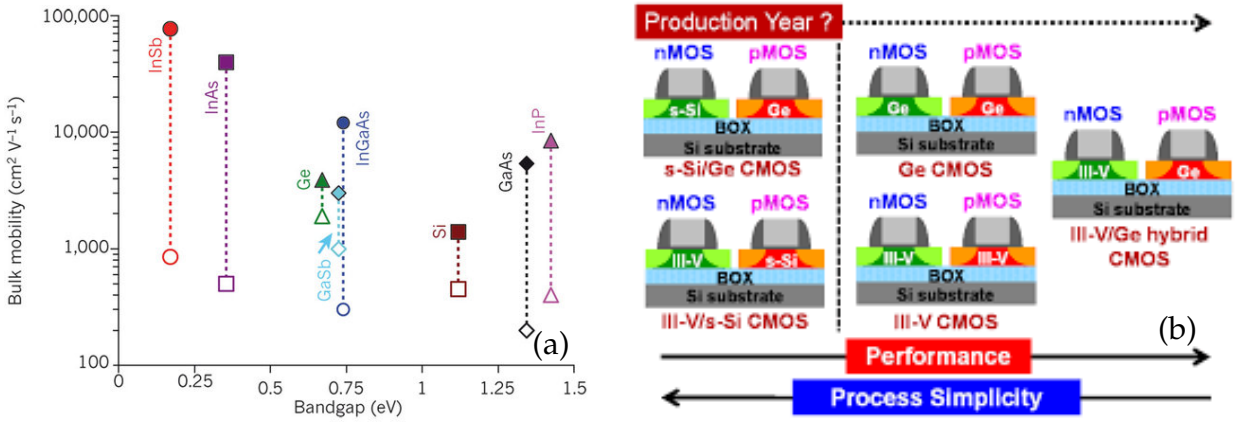


Figura 1.2: (a) Movilidad de portadores en distintos sustratos considerados para la estructura MOS. Los símbolos de color indican movilidad de electrones (cargas negativas) y los símbolos vacíos movilidad de huevos (cargas positivas). Reproducido de [20]. La combinación de transistores P-MOS (Ge) con transistores N-MOS (III-V) sobre una misma oblea de Si se propone como una opción para optimizar el rendimiento de futuros sistemas CMOS. Reproducido de [21]

véase la Fig. 1.3) atenta contra la tradicional operación fiable durante 10 años, dado que induce una degradación de las características eléctricas del dispositivo (tales como la tensión de umbral, el rango de sub-umbral, la transconductancia, y finalmente la corriente de conducción). En estas, el gran número de defectos en la interfaz, las dificultades de su cuantificación respecto al sistema Si/SiO_2 [7] y sus efectos sobre los mecanismos de túnel [16]-[19] comprometen su fiabilidad y vida útil.

Como se ha expuesto previamente, la introducción de los dieléctricos *high–*κ permitió reducir las corrientes a través del aislante de compuerta en los transistores MOS. Esto se debe a que al poseer una constante dieléctrica mayor al SiO_2, se pueden utilizar óxidos mas gruesos y mantener la capacidad de compuerta necesaria para garantizar el funcionamiento de los dispositivos. No obstante, esto no ha eliminado los problemas de ruptura dieléctrica (un fenómeno general en aislantes) que comprometen la fiabilidad, sino que por el contrario ha disparado nuevos interrogantes en este apartado. Al respecto del fenómeno de ruptura, debe diferenciarse entre ruptura extrínseca e intrínseca: En relación a la primera, la ruptura a bajo campo está relacionada con defectos macroscópicos tales como partículas, variaciones en el espesor del óxido, impurezas metálicas, etc., y ha sido minimizada en tecnologías Si/SiO_2 dado su avanzado estado de desarrollo [22]. Por el contrario, la ruptura intrínseca, especialmente producida por campos eléctricos altos, es inevitable. Esta ha sido ampliamente estudiada en SiO_2 [23]-[25], y actualmente se sabe que la degradación del óxido de compuerta es consecuencia del transporte de portadores (corriente de túnel) a través del dieléctrico de compuerta [24]: Esta promueve la migración de átomos de los electrodos (compuerta y/o sustrato en una estructura MOS) hacia el óxido, así como la generación de defectos eléctricamente activos [26], [27]. Alcanzada una densidad crítica de defectos, se produce un filamento conductivo (*Conductive Filament*, CF) entre los electrodos, formado por defectos nanométricos [23], [28] y se dispara una dinámica de realimentación positiva, en la cual la corriente a través del fi-

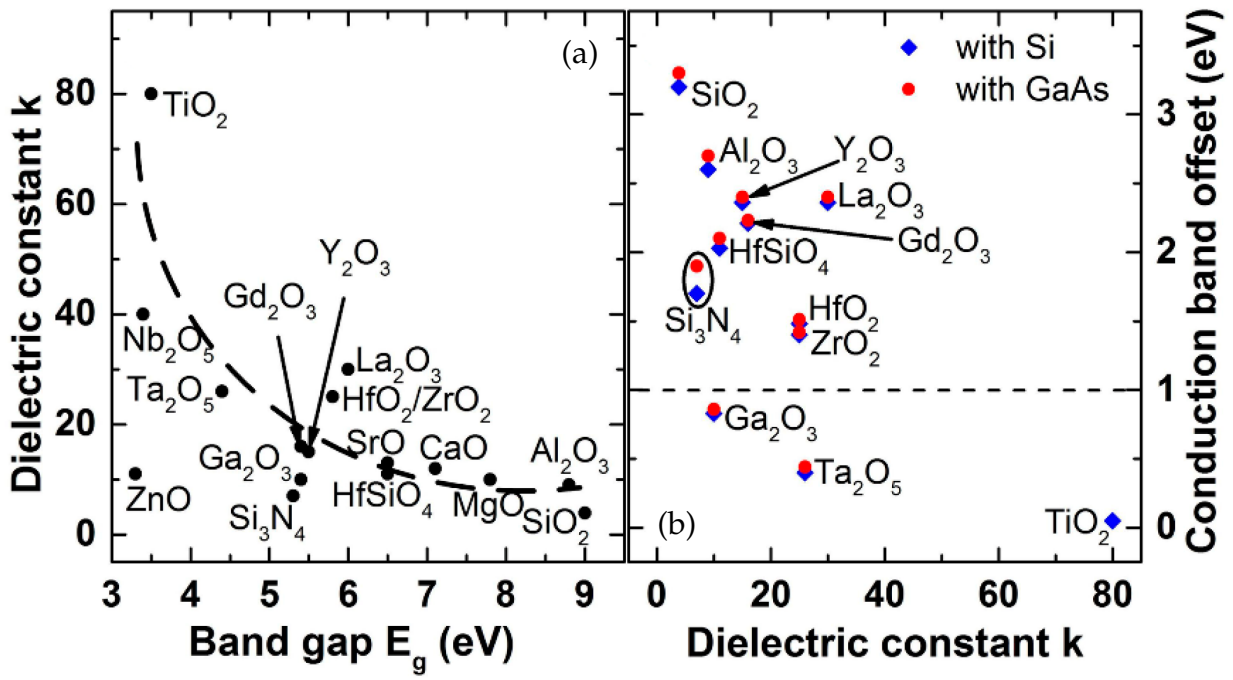


Figura 1.3: Dieléctricos *high–κ*: Aparte de la constante dieléctrica, otros aspectos del material aislante son relevantes en su funcionamiento como dieléctrico de compuerta, como el ancho de la zona prohibida y el desfase entre las bandas de conducción. Estos se muestra en (a) Constante dieléctrica vs. Ancho de la zona prohibida, y (b) Constante dieléctrica vs. Desfase entre bandas de conducción. Reproducido de [32]

lamento aumenta abruptamente, produciendo una gran disipación de calor que conduce a la destrucción del medio aislante. Por lo tanto, la ruptura dieléctrica intrínseca puede considerarse como un proceso de 3 etapas [29], [30]: *i)* La fase de acumulación de defectos, *ii)* el evento de ruptura y *iii)* de degradación abrupta por temperatura o progresiva [18], [31]. Este proceso tiene una gran variabilidad, por lo que es estudiado con herramientas estadísticas para poder hacer estimaciones de la vida útil de los dispositivos en condiciones de trabajo nominales. Sin embargo, al cambiar las propiedades térmicas y de generación de defectos entre el SiO_2 y los dieléctricos *high–κ* también cambia la estadística de la ruptura dieléctrica, representando un problema abierto para la comunidad científica.

Si bien propone un desafío de confiabilidad, este fenómeno ha dado lugar al mecanismo de conmutación resistiva (*Resistive Switching*, RS), el cual ha sido explotado en la operación de dispositivos tales como las recientemente sugeridas memorias resistivas de acceso aleatorio (*Resistive Random Access Memory*, RRAM) [35]-[37]. El mecanismo de RS tiene lugar en ciertos materiales aislantes tales como óxidos de metales de transición (TMO) (TiO_2, TaO_X) entre los cuales se encuentran los dieléctricos *high–κ* como el HfO_2 y Al_2O_3, así como también en dieléctricos 2D (h-BN, MoS_2), y consiste a grandes rasgos en una ruptura dieléctrica "reversible". Mediante la modulación y eventual disolución del filamento conductivo formado durante un evento de ruptura dieléctrica controlada, los dispositivos RRAM pueden oscilar entre dos estados de baja y alta resistencia, de forma que pueden almacenar información codificándola como un valor de resistencia. Al

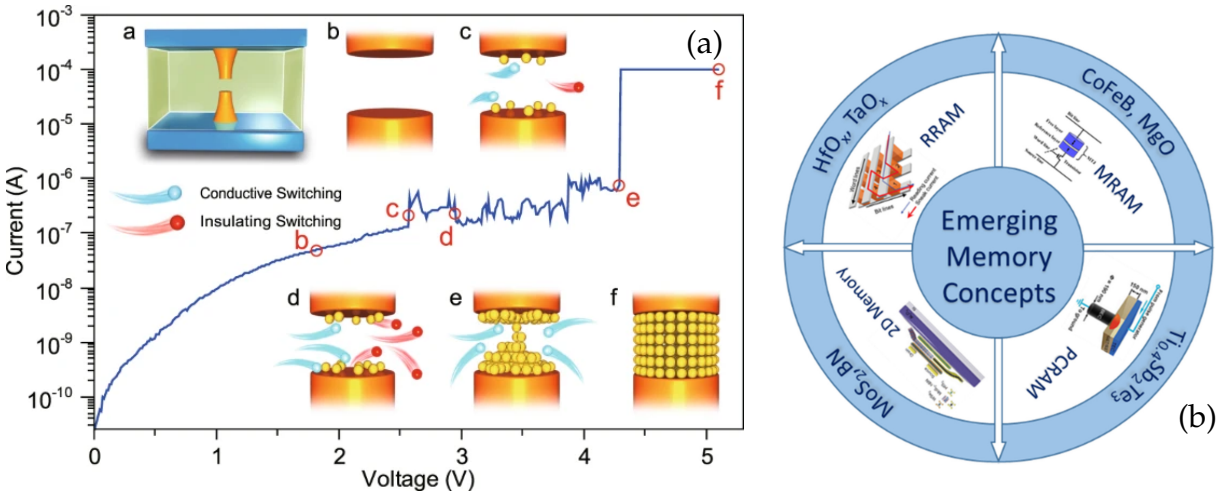


Figura 1.4: (Der.) Curvas de Corriente–Tensión (I–V) típicas mostrando las distintas etapas de la formación del CF. En los *inset* (a) y (b) se representa esquemáticamente el CF. En (c) y (d) se indica la migración de iones/defectos desde y hacia el CF, produciendo variaciones aleatorias en la corriente. En (e) se representa el punto en el cual el CF se completa, produciendo un aumento abrupto de la corriente a través de la estructura. Finalmente en (f) se muestra el caso de un CF completamente formado. Reproducido de [33]. (Izq.) Principales mecanismos explotados en memorias emergentes. Además de las ya mencionadas, se pueden citar las memorias RAM de cambio de fase (*Phase Change Random Acces Memory*, PCRAM) y las RAM magnéticas (*Magnetic Random Access memory*, MRAM). Reproducido de [34]

mismo tiempo, el almacenamiento puede ser Volátil o No-Volátil, dependiendo de si el estado de resistencia almacenado se pierde una vez removido el estímulo eléctrico, o no. Independientemente del caso, las características estáticas de los dispositivos RRAM han sido ampliamente investigadas. Sin embargo, las evolución temporal del fenómeno de conmutación ha sido poco abordada en la literatura, y resulta de sumo interés para poder estimar la máxima velocidad de conmutación y su dependencia con las propiedades intrínsecas del medio de conmutación. Esto es de vital importancia dadas las características de las principales aplicaciones para la tecnología de RRAM. Cabe destacar, tal como se indica en la Fig. 1.4b, que si bien las RRAM son las más difundidas, existe un gran número de memorias no-volátiles actualmente bajo estudio, utilizando principios resistivos, magnéticos o de cambio de fase, entre otros.

Volviendo a la Figura 1.1c y 1.1d la maduración tecnológica de las memorias RRAM (tanto volátiles como no volátiles) se posiciona como un paso clave en el desarrollo del enfoque SiP, dado su enorme potencial de aplicación en circuitos neuromórficos y redes neuronales. Específicamente, las redes Neuronales Profundas (*Deep Neural Networks*, DNN, véanse las Figs. 1.5a y 1.5b) [38] han demostrado un éxito comercial significativo en los últimos años, con rendimientos que sobrepasan alternativas previas mucho más sofisticadas para el reconocimiento de voz [39] e imágenes [40]-[42] y se aproximan o incluso superan niveles humanos. Sin embargo, estas implementaciones requieren de una fase de entrenamiento que es extremadamente costosa en términos computacionales y por ende reduce su aplicabilidad en ciertos escenarios (en algunos casos se pueden llegar a requerir días o semanas para entrenar una DNN frente a problemas realistas, in-

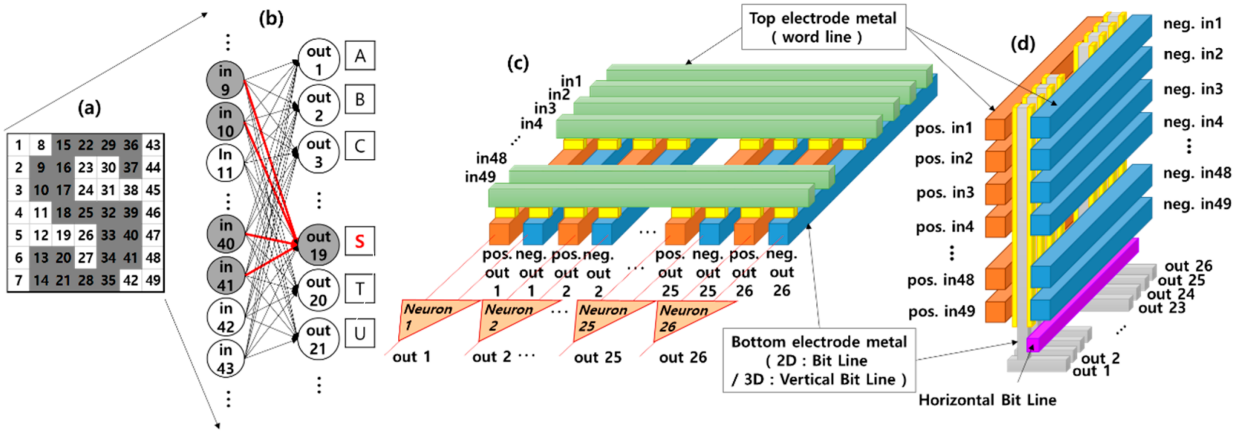


Figura 1.5: (a) Ejemplo de un patrón de entrada a clasificar, en este caso la letra 'S'. (b) Representación esquemática de una red neuronal de una sola capa. En este caso contiene 49 neuronas de entrada, y 26 neuronas de salida, con un total de 49×26=1274 sinapsis. Las lineas rojas indican las conexiones activas para el patrón de entrada asumido. (c) Implementación en *hardware* de la red neuronal utilizando un *crossbar* de RRAM de 2 dimensiones. (d) Una opción más avanzada sería la construcción de un *crossbar* vertical de memorias RRAM (VRRAM), el cual permita el mismo rendimiento con un mejor aprovechamiento del área disponible en silicio. Reproducido de [46]

cluso utilizando *hardware* del estado del arte). Esto se debe a la cantidad de parámetros a ajustar durante el entrenamiento y la propia naturaleza de la arquitectura de Von Neumann: cada parámetro debe ser enviado a las unidades de cómputo (CPU y/o GPU) para procesarlo, y luego retornado a memoria para su almacenamiento. Más aún, este cuello de botella también afecta la fase de funcionamiento normal (es decir, con la DNN ya entrenada). Con el objetivo de minimizar el costo computacional se han explorado múltiples alternativas, con distintos grados de éxito. En todas ellas (ya sea involucrando GPUs [43], FPGAs [44] o ASICs), la premisa ha sido aumentar el paralelismo y reducir el movimiento de datos entre bloques del sistema (cómputo y memoria). Eventualmente, el caso ideal sería poder realizar el cómputo y el almacenamiento en el mismo dispositivo [45], con lo que la complejidad temporal se podría reducir de $O(n^2)$ a $O(1)$.

Una posibilidad para alcanzar este ideal, particularmente para el caso de redes completamente conectadas (*Fully Connected*, FC) es la utilización de *cross-bars* de memorias no volátiles tales como RRAM, como los indicados en las Figs. 1.5c y 1.5d. En estos, existencia de estados intermedios permite un ajuste gradual de cada parámetro. Dado que en este escenario la operación de Multiplicación Vector Matriz (*Matrix-Vector-Multiplication*, MVM), fundamental en el funcionamiento de DNNs, puede realizarse directamente en el *crossbar* en forma altamente paralelizada en cada nodo, se pueden obtener mejoras en el rendimiento y eficiencia energética de hasta 3 a 5 órdenes de magnitud en comparación a arquitecturas convencionales tipo "Von Neumann". Pero a pesar de la prometedora potencialidad de estas implementaciones, existen ciertas características intrínsecas de los dispositivos RRAM (usualmente ignoradas o incluso beneficiosas cuando se lo utiliza en aplicaciones de memoria) tales como la relación entre los estados alta y baja resistencia, la carencia de estados intermedios, y la asimetría entre la escritura y el borrado, que suponen problemas mayores para su aplicación en DNNs [47].

En esta primera sección del presente capítulo introductorio, se ha presentado una breve síntesis del estado del arte y de las problemáticas que deberá afrontar la nanoeléctrónica en los próximos años. Continuando la filosofía ascendente adoptada en esta introducción, esta tesis aborda algunos de los mecanismos de degradación y ruptura en los sustratos y dieléctricos propuestos como protagonistas centrales en los procesos de integración venideros, y los desafíos de confiabilidad que esto implica. Acto seguido, la ruptura dieléctrica es abordada desde el punto de vista del fenómeno de conmutación resistiva, como base de los dispositivos de memoria no volátil llamados a jugar un rol fundamental en las implementaciones de *hardware* de inteligencia artificial del futuro cercano, las cuales se abordan en detalle en el final del trabajo. La organización del contenido se detalla en la siguiente sección.

1.1. Organización del trabajo de tesis

En esta tesis se reportan aspectos asociados a la dinámica de degradación y ruptura en dispositivos semiconductores del estado del arte, dada su gran injerencia sobre la fiabilidad de los mismos. Se pone particular énfasis sobre el fenómeno de ruptura, abordándolo con un enfoque ascendente que parte desde el nivel fenomenológico (física de ruptura en dispositivos aislados) y culmina con el estudio de su aplicación como elemento de memoria no-volátil en circuitos neuromórficos y los aspectos de confiabilidad asociados. En este punto es imprescindible señalar que todos los dispositivos estudiados experimentalmente fueron provistos por colaboradores internacionales del grupo de trabajo, debidamente mencionados en cada sección, y es gracias a su predisposición y la calidad de sus dispositivos que las actividades desarrolladas fueron posibles. De la misma forma, el acceso a la instalaciones del laboratorio TANDAR en el Centro Atómico Constituyentes (Buenos Aires) y el Departamento de Ingeniería Electrónica de la UAB (Barcelona, España) ha sido factores clave en el desarrollo de este trabajo. De este modo, el trabajo de tesis se separa en capítulos alrededor de estos dos ejes troncales de la investigación, de la siguiente manera:

Capítulo 2: Conceptos básicos

En este apartado se revisan los conceptos básicos de la física de dispositivos MOS y conducción en óxidos, centrando la atención en las características eléctricas de los mismos, los métodos de caracterización fundamentales y la pertinencia tecnológica de la información obtenible de los mismos. Asimismo, se introducen los conceptos de degradación y ruptura del óxido de compuerta, centrales para los resultados posteriores. Por último, se plantean los fundamentos de redes neuronales y los procedimientos asociados.

Capítulo 3: Degradación en estructuras MOS

A pesar de los prometedores resultados iniciales, los semiconductores de alta movilidad propuestos para reemplazar al silicio como material de sustrato en transistores MOS aún presentan múltiples interrogantes y desafíos a resolver que frenan su implementación a escala industrial. Es por ello que en este capítulo se estudian mediante mediciones de estrés eléctrico y de Capacidad–Tensión, los fenómenos de atrapamiento de carga y generación de estados de interfaz en materiales III–V (InGaAs e InP, con alta movilidad de portadores negativos) y Germanio (con alta movilidad de portadores positivos). Se contribuye con una explicación fundamentada de los mismos, indicando su dependencia con el proceso de fabricación y las características de los materiales aislantes involucrados.

Capítulo 4: Dinámica de ruptura en dieléctricos

Mediante experimentos sistemáticos de ruptura por estrés eléctrico e irradiación controlada con iones de alta energía de estructuras MOS, se analiza la estadística de ruptura de dieléctricos *high–*κ, y su relación con las propiedades geométricas e intrínsecas del medio, así como con la naturaleza de los defectos existentes en el mismo. Con el agregado de simulaciones multi-físicas se demuestra que en aislantes *high–*κ los defectos agregados durante el estrés eléctrico siguen una dinámica de generación con una marcada correlación espacial, lo que permite explicar la escasa dependencia entre la dispersión en el tiempo de ruptura y el espesor de la capa aislante en algunos materiales *high–*κ.

Capítulo 5: Conmutación Resistiva en dieléctricos

Si bien la generación de defectos en la capa aislante producida por la aplicación de una tensión eléctrica es responsable del evento de ruptura dieléctrica y por lo tanto supone un problema de fiabilidad, el mismo fenómeno también da lugar al mecanismo de conmutación resistiva. Este ha demostrado un gran potencial en aplicaciones de almacenamiento de memoria (memorias resistivas de acceso aleatorio) y como sinapsis artificiales en circuitos neuromórficos. En este capítulo se investiga la evolución temporal de la resistencia filamentaria en memorias de conmutación resistiva, tanto para el caso no-volátil (utilizando un óxido *high–*κ) como medio de conmutación, como el caso volátil (utilizando *cross-bars* de h-BN como material aislante). Teniendo en cuenta las similitudes de los cambios eléctricos y micro-estructurales observados entre el evento de ruptura dieléctrica y el SET en conmutación resistiva, se propone un modelo compacto que captu-

re la dinámica de transición del SET en términos de la conductividad térmica del material, la difusividad de las especies migrantes y la geometría de la estructura.

Capítulo 6: Redes Neuronales

La implementación en *hardware* de redes neuronales mediante *cross-bars* de memorias resistiva es en la actualidad un tópico de gran interés en la comunidad científica dada su capacidad para procesar grandes volúmenes de información con baja latencia, consumo de potencia y área requerida. Existe sin embargo, un sin número de desafíos a resolver antes de poder llevar esta tecnología a la escala industrial. En este capítulo se demuestra la aplicabilidad del modelo compacto de memdiodo en la simulación eléctrica realista de perceptrones mono y multi-capa. Para ello, se tiene en cuenta el impacto de no idealidades tales como la resistencia de linea, la reducción de la ventana resistiva de los dispositivos, la degradación de la relación señal a ruido y la variabilidad dispositivo a dispositivo, entre otras. También se hacen contribuciones enfocadas en los aspectos de fiabilidad, proponiendo estrategias para la mitigación de fallas de enclavamiento en las memorias resistivas.

Capítulo 7: Conclusiones y próximos pasos

Se resumen los resultados más importantes del trabajo de tesis, enumerando sus contribuciones a fin de poner de manifiesto su novedad. Asimismo, se apuntan algunos interrogantes que pueden desprenderse de los resultados, considerando las perspectivas de esta línea de investigación.

2 Conceptos básicos

La primer propuesta documentada de un dispositivo de efecto de campo construida en un semiconductor se remonta al año 1928. Desde entonces, numerosos intentos infructuosos de implementarlo en un dispositivo superficial se sucedieron hasta su introducción masiva en la electrónica, que no se produjo sino hasta casi 40 años después, limitado por la posibilidad de fabricar en forma consistente una estructura MOS de la calidad suficiente para obtener un desempeño aceptable de los transistores [48]. De allí en más, el escalamiento y la evolución de estos dispositivos ha impulsado el crecimiento asombroso de la industria de los semiconductores. Para lograrlo, el estudio de los fenómenos físicos detrás su funcionamiento han sido, y continúan siendo al día de hoy, fundamentales para comprender y modelar el comportamiento eléctrico esperado de los dispositivos. La profundidad de este estudio posibilita la obtención de información de gran relevancia tecnológica a partir de la caracterización de variables eléctricas tanto en dispositivos MOS. Por su rol central en este trabajo, este capítulo se dedica a revisar las nociones básicas de la física de la estructura MOS, haciendo particular énfasis en la relación entre la caracterización eléctrica de los dispositivos y los aspectos tecnológicos asociados a las variables medidas. El objetivo de las secciones subsiguientes será definir conceptos, nomenclaturas y parámetros que se utilizarán a lo largo de este documento. Para mayor profundidad en los desarrollos y teoría asociada, se refiere al lector a la literatura disponible en el tema [49]-[51].

2.1. La estructura MOS

Las propiedades físicas y eléctricas de la estructura Metal-Óxido-Semiconductor (MOS) han sido ampliamente estudiadas en las últimas décadas, por ser estos dispositivos fundamentales en la mayoría de los dispositivos planares en tecnologías de circuitos integrados. El capacitor MOS tiene una estructura como la que se observa en la Fig. 2.1a, donde mediante el terminal metálico (compuerta), se controla la densidad de carga en

el semiconductor (sustrato) a partir del efecto capacitivo provisto por el óxido (aislante) delgado que separa ambos terminales. Un punto departida apropiado para entender el comportamiento eléctrico de la estructura MOS consiste en considerar su diagrama energético de bandas, el cual se presentan en las Figs. 2.1ba-2.1bd para dopaje P (arriba) y N (abajo). Según la polaridad del potencial aplicado y el tipo de dopaje, el sustrato (en la región próxima a la interfaz con el óxido) puede presentar una condición de carga nula o "bandas planas"(a), acumulación de portadores mayoritarios (b), vaciamiento de por-

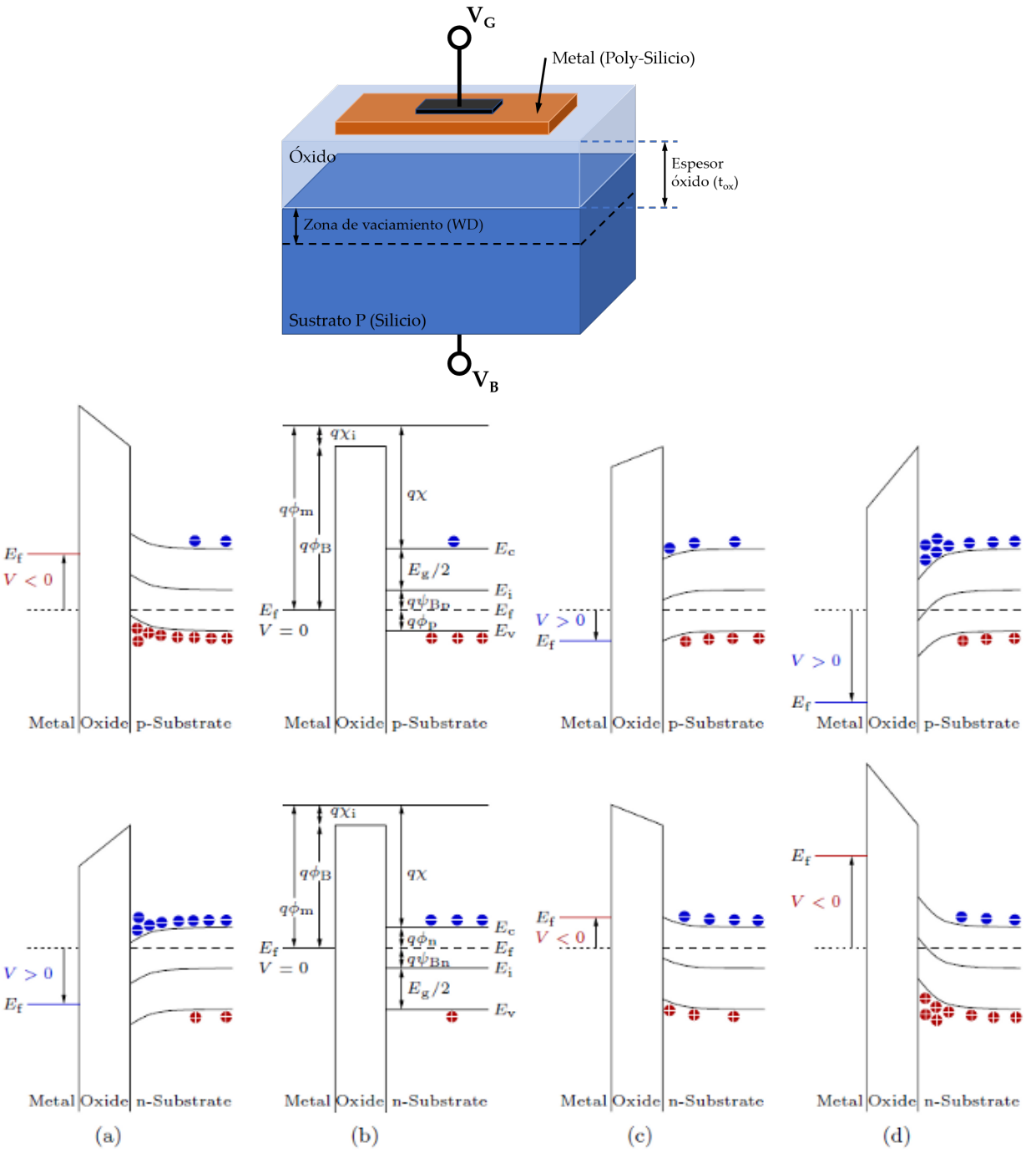


Figura 2.1: (Arriba) Estructura básica de un capacitor MOS. (Abajo) (a)-(d) Diagramas de bandas de energía en cada condición de operación, según potencial aplicado V_G tanto para sustrato P (fila superior) como N (fila inferior). Adaptado de [51].

tadores (c) o inversión del sustrato (acumulación de portadores minoritarios) (d). Estas condiciones son fundamentales en el análisis de las propiedades físicas de los dispositivos MOS, ya que son fuertemente dependientes de los materiales involucrados y de la distribución de defectos y cargas. En dichas figuras, ϕ_m es la función trabajo del metal, χ la afinidad electrónica del semiconductor, χ_i afinidad electrónica del aislante, ϕ_B la barrera de potencial entre el metal y el aislante, y Ψ_B la diferencia de potencial entre el nivel de Fermi E_F y el nivel de Fermi intrínseco E_i. E_C y E_V corresponden, respectivamente, al fondo de la banda de conducción y al tope de la banda de valencia en el semiconductor. Para una descripción completa de las propiedades de las estructuras MOS y el origen de dicho diagrama energético, se pueden consultar las referencias [50], [51].

La existencia de una curvatura de bandas en el silicio es sinónima de una caída de potencial, con las consecuentes variaciones de campo eléctrico ($\mathcal{E}$) y carga (Q) a lo largo del dispositivo. Poniendo el foco en las variaciones de carga (dado que la magnitud a controlar es la carga en el semiconductor), la relación entre la carga en el semiconductor y el potencial de superficie $\Psi_S(V)$ puede hallarse mediante el planteamiento de la ecuación unidimensional de Poisson (Ec. (2.1)):

$$\frac{d^2\Psi}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{q}{\epsilon_S}[p(x) - n(x) + N_d^+(x) - N_a^-(x)] \tag{2.1}$$

donde Ψ denota el potencial en cada punto de la estructura MOS, q la carga del electrón, ϵ_S la permitividad eléctrica del semiconductor y $p(x)$, $n(x)$, $N_d^+(x)$ y $N_a^-(x)$ son las distribuciones de portadores e impurezas ionizadas en el semiconductor, respectivamente. $\Psi_S(V)$ da cuenta de la curvatura total de bandas sobre la interfaz óxido-semiconductor. La solución a la ecuación de Poisson resulta en la Ec. 2.2 ,

$$Q_S = -\epsilon_S\mathcal{E}_S = \pm\sqrt{2\epsilon_S kTN_a}\left[\left(e^{-q\Psi_s/kT} + \frac{q\Psi_s}{kT} - 1\right) + \frac{n_i^2}{N_a}\left(e^{q\Psi_s/kT} - \frac{q\Psi_s}{kT} - 1\right)\right]^{1/2} \tag{2.2}$$

(véase la Ref. [49]). Aquí, k es la constante de Boltzmann, T es la temperatura absoluta y n_i la densidad intrínseca de portadores del semiconductor. Aplicando ley de Gauss, $Q_S = \epsilon_S\mathcal{E}_S$, con $\mathcal{E}_S$ el campo eléctrico en el semiconductor, podemos relacionar directamente la carga integrada por unidad de área en el semiconductor Q_S con el potencial aplicado en el terminal de compuerta V_G según la Ec. (2.3), donde V_G es el potencial aplicado a la compuerta, V_{FB} es el potencial de bandas planas (*flat-band*), V_{ox} es la caída de potencial en el óxido y $C_{ox} = \epsilon_{ox}/t_{ox}$ es la capacidad por unidad de área del óxido, inversamente proporcional a su espesor t_{ox}.

$$V_G - V_{FB} = V_{ox} + \Psi_s = \frac{-Q_S}{C_{ox}} + \Psi_s. \tag{2.3}$$

No obstante, a la hora de contrastar los valores obtenidos mediante la Ec. 2.2 no es posible determinar experimentalmente el nivel de carga en el semiconductor. Un método

indirecto para obtener Q_S experimentalmente es analizar la relación Capacidad–Tensión del dispositivo MOS, la cual puede obtenerse a partir de la caracterización de la impedancia de pequeña señal de la estructura.

2.1.1. Curvas de Capacidad-Tensión

La impedancia (o admitancia) de pequeña señal medida en un dispositivo MOS puede interpretarse en términos de un circuito Capacitor-Conductancia (C–G) paralelo (existen también otras variantes). En este contexto, la parte imaginaria de la admitancia medida en función de la tensión aplicada, corresponde la capacidad total de la estructura MOS. La capacidad total entre terminales se reparte entre la capacidad del óxido C_{ox} y la capacidad del semiconductor. Esta puede separarse a su vez en la capacidad de vaciamiento C_d, debido a la distribución de carga espacial en el sustrato, y la capacidad de inversión C_i, debido a la alta concentración de portadores minoritarios en la interfaz en condición de inversión. Existe también una tercer componente debida a estados de interfaz C_{it} (los cuales se discutirán más adelante en el texto). El circuito equivalente resultante de las capacidades en una estructura MOS se muestra en la Fig. 2.2a de donde la capacidad total equivalente puede escribirse como se muestra en la Ec. (2.4)

$$\frac{1}{C_G} = \frac{1}{C_{ox}} + \frac{d\Psi_s}{d(-Q_S)} = \frac{1}{C_{ox}} + \frac{1}{C_D + C_i + C_{it}} \tag{2.4}$$

en donde la capacidad del semiconductor vendrá dada por la curvatura de bandas, representada en Ψ_s. Estas curvas son generalmente conocidas como curvas Capacidad–Tensión o C–V. En la Fig. 2.2b se puede apreciar la curva de capacidad en función de la tensión aplicada al terminal de compuerta V_G para un sistema MOS ideal, computada mediante la derivada de la carga en el semiconductor respecto al potencial de superficie de la ecuación (2.2) y resolviendo de forma acoplada la ecuación (2.3).

De derivar la Ec. 2.2 para obtener la curva de capacidad tensión, se puede observar que existe una relación entre ciertos parámetros tecnológicos fundamentales del dispositivo MOS, y los valores de capacidad. Por ejemplo, la permitividad y el espesor del óxido se relacionan directamente con la capacidad en acumulación $C_{acc} = C_{ox} = Area\ \epsilon_{ox}/t_{ox}$, o bien la capacidad en inversión a baja frecuencia (línea a trazos). Por otro lado, la condición de bandas planas, caracterizada por la tensión V_{FB}, es de vital importancia por su relación directa con la tensión de umbral (V_{TH}) a la cual se asume la inversión del canal. Un estimador recientemente introducido para V_{FB} y de amplia aplicación para estructuras MOS sobre semiconductores de alta movilidad se conoce como el método del punto de inflexión [53]. El mismo consiste en inferir este valor característico a partir de la tensión a la cual se produce el máximo de la derivada primera (o cero de la derivada segunda) de la característica C–V. Este método no es influenciado por las posibles

no idealidades de la estructura, que se discuten brevemente en la sección 2.1.1.1, ni por la frecuencia de caracterización, y solo presenta el error introducido por el instrumento.

Una importante característica de la curva C–V es su dependencia con la frecuencia de la señal de AC utilizada en la medición de impedancia. Mientras en acumulación, los portadores mayoritarios pueden responder idealmente a las frecuencias típicas de caracterización, en el orden de 2 Hz - 2 MHz, los minoritarios necesitan ser generados y recombinados al ritmo de la señal de AC en la región de inversión. Para valores típicos de tiempo de vida medio en sustratos ligeramente dopados, frecuencias superiores a los 100 Hz suelen ser suficientes para que el canal de inversión no pueda responder a esta señal, no existiendo entonces el incremento abrupto de capacidad debido a la inversión del canal y observando un valor mínimo constante con la tensión (línea llena). Este valor mínimo está determinado por la máxima profundidad de vaciamiento en el semiconductor, antes de ser apantallado su efecto por el canal de inversión [49], calculándose como $C_{min} = \epsilon_S / W_{Dmax}$, donde W_{Dmax} es el espesor de la zona de vaciamiento en el inicio de la inversión fuerte. La máxima profundidad de vaciamiento puede estimarse según $W_{Dmax} = \sqrt{4\epsilon_S kT \ln(N_a/n_i)/q^2 N_a}$. En condiciones ideales, la capacidad mínima es representativa del dopaje del sustrato. Sin embargo, algunos defectos presentes en la estructura, particularmente en condiciones de vaciamiento y hacia la inversión, pueden responder a frecuencias más altas que los portadores minoritarios en sustrato, alterando la curva C–V. Esta característica es fundamental para cuantificar la contribución de defectos cerca de la interfaz, como se resume en la siguiente sección.

2.1.1.1. Carga atrapada: Trampas de borde y defectos de interfaz

Las estructuras MOS reales presentan no idealidades respecto al caso teórico, como diferencias de funciones trabajo, efectos de la compuerta, cargas fijas y móviles en el óxido (Q_{ox}), trampas de frontera (BT, *"Border Traps"* distribuidas desde la interfaz óxido-semiconductor hacia el cuerpo del óxido y la contribución de defectos de interfaz

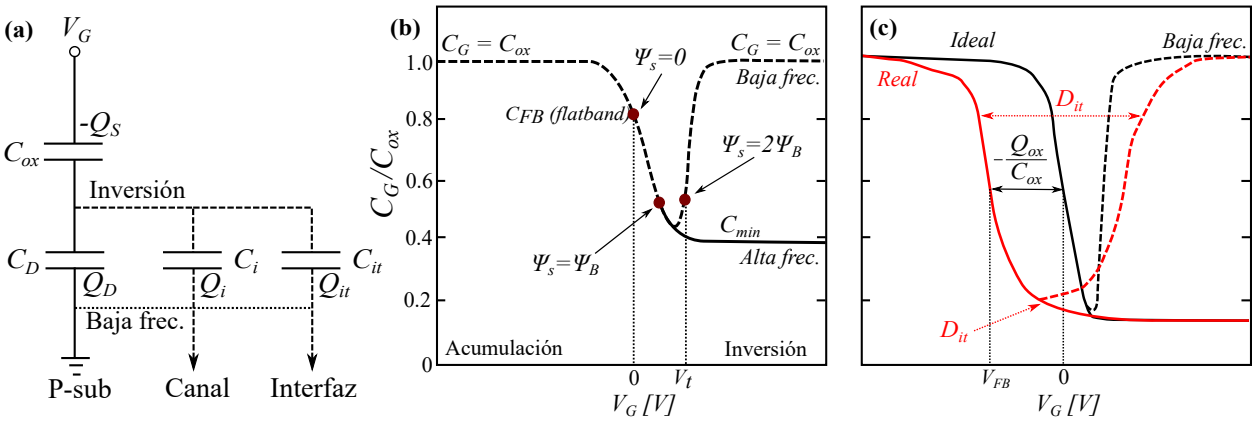


Figura 2.2: (a) Circuito equivalente de capacidades y (b) curvas capacidad-tensión para una estructura MOS ideal a baja (lineas a trazos) y alta (lineas llenas) frecuencia. (c) Impacto de defectos en el óxido y la interfaz sobre la curva C-V real (líneas rojas) respecto a la ideal (líneas negras). Reproducido de [52]

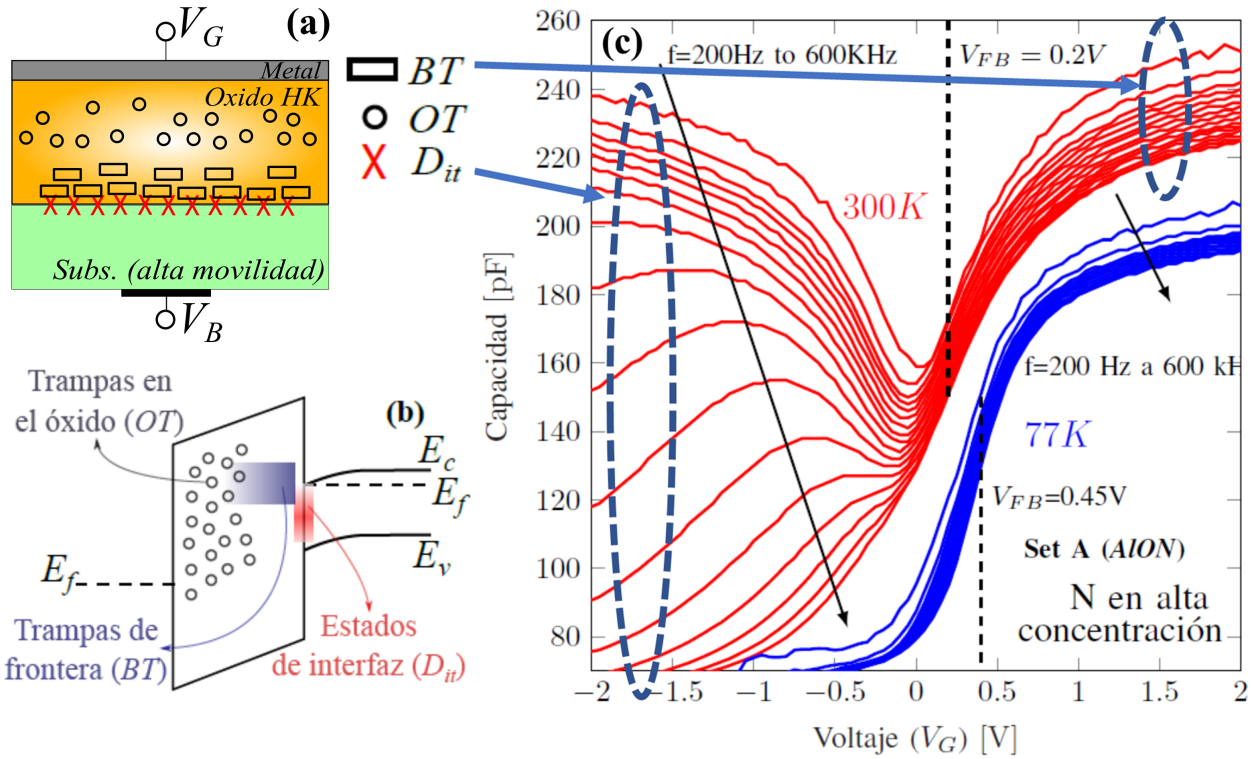


Figura 2.3: (a) Diagrama esquemático de la distribución de defectos defectos en la estructura MOS, dando lugar a distintos tipos dependiendo de su posicionamiento. (b) Diagrama de bandas considerando los distintos tipos de trampas. Reproducido de [52].(c) Curvas C-V típicas para distintas frecuencias de una estructura MOS fabricada sobre un sustrato de alta movilidad (InGaAs), mostrando los efectos de estados de interfaz y las trampas de frontera. Reproducido de [54]

($D_{it}(\Psi_s)$) [49]. Una representación esquemática de la presencia de estos defectos un una estructura MOS se observa en la Fig. 2.3a, mientras que su análisis a nivel de bandas de energía se detalla en la Fig. 2.3b.

La medición de curvas C–V experimentales es una técnica ampliamente estudiada para caracterizar estas no idealidades [50]. La comparación entre una curva C–V teórica (línea negra) y su equivalente experimental (línea roja) se muestra cualitativamente en la Fig. 2.2c. En primer lugar, el desplazamiento de la tensión V_{FB} puede asociarse a carga fija o carga de interfaz que es atrapada a medida que las bandas se curvan en el barrido de tensión, como se abordará en detalle en el capítulo 3. De este modo, la carga total en el óxido Q_{ox} tendrá una contribución que puede ser fija debido a trampas en el óxido (Q_{ot}) y una dependiente de la curvatura, es decir de una distribución energética de estados a lo largo de la banda prohibida, generalmente localizados en la cercanía de la interfaz ($Q_{it}(\Psi_s)$). Podemos entonces escribir $Q_{ox} = Q_{it}(\Psi_s) + Q_{ot}$, resultando su influencia en un desplazamiento horizontal de la curva $\Delta V_{FB} = -Q_{ox}/C_{ox}$. A su vez, la interacción de los estados de interfaz D_{it} pueden resultar en un aumento de la capacidad observada en la zona de inversión débil y también hacia inversión fuerte (si la contribución de C_{it} en la Fig. 2.2a es considerable), además de un ensanchamiento (*"stretch-out"*) de la curva en la zona de inversión, debido a defectos con energías desde el *"midgap"* hacia la banda de conducción, mostrando dependencia con la frecuencia de medición.

La cuantificación de Q_{ot} y D_{it} se encuentran ampliamente revisados en la lite-

ratura [49], [50] y son indicadores de la calidad de la estructura MOS. De hecho, con el cambio de escenario impuesto por la renovación de los materiales en la estructura MOS, nuevas consideraciones han sido necesarias para que estas técnicas puedan ser satisfactoriamente utilizadas en, por ejemplo, dispositivos con sustratos de alta movilidad [10], [55]. Algunos aspectos de interés discutidos en la literatura asociada serán abordados en detalle en el capítulo 3, pero a modo introductorio resulta conveniente discutir brevemente en este punto algunos de los métodos de cuantificación de D_{it} más difundidos en la literatura así como el impacto de la frecuencia de la señal de AC de excitación en las curvas experimentales C–V debido a la presencia de no idealidades en la estructura MOS. De este modo, será posible separar la contribución de los defectos y facilitar el análisis de los resultados experimentales de los capítulos 3 y 4.

2.1.1.2. Método de Baja-Alta Frecuencia (Castagné-Vapaille)

Este método, que combina mediciones C–V a alta y baja frecuencia fue desarrollado originalmente por Castagné y Vapaille [56]. Inicialmente este método no requiere de cálculos teóricos para la comparación del D_{it} entre dos estructuras diferentes, lo cual es una ventaja significativa debido a la complejidad que tienen dichos cálculos cuando el perfil de dopaje es no uniforme, si es que acaso se pueda conocer. Partiendo de las Ecs. 2.5 y 2.6 para baja y alta frecuencia

$$C_{LF} = \frac{C_{ox}\left(C_D + C_{it}\right)}{C_{ox} + C_D + C_{it}} \tag{2.5}$$

$$C_{HF} = \frac{C_{ox}C_D}{C_{ox} + C_D} \tag{2.6}$$

donde C_{LF} y C_{HF} son las capacidades a baja y alta frecuencia, respectivamente, se puede expresar C_{it} como:

$$C_{it} = \left(\frac{1}{C_{LF}} - \frac{1}{C_{ox}}\right)^{-1} - C_D = \left(\frac{1}{C_{LF}} - \frac{1}{C_{ox}}\right)^{-1} - \left(\frac{1}{C_{HF}} - \frac{1}{C_{ox}}\right)^{-1} \tag{2.7}$$

definiendo la diferencia de capacidades como $\Delta C \equiv C_{LF} - C_{HF}$ y considerando la relación $D_{it} = C_{it}/q^2$, se puede obtener directamente la densidad de estados de interfaz como se indica en la Ec. 2.8.

$$\begin{aligned} D_{it} &= \frac{C_{ox}}{q^2}\left[\left(\frac{1}{\Delta C/C_{ox} + C_{HF}/C_{ox}} - 1\right)^{-1} - \left(\frac{1}{C_{HF}/C_{ox}} - 1\right)^{-1}\right] \\ &= \frac{\Delta C}{q^2}\left(1 - \frac{C_{HF} + \Delta C}{C_{ox}}\right)^{-1}\left(1 - \frac{C_{HF}}{C_{ox}}\right)^{-1} \end{aligned} \tag{2.8}$$

para cada valor de tensión aplicada a la estructura (D_{it}-V). Como se puede apreciar en la Ec. 2.8 en primer orden, la densidad de trampas es proporcional a la diferencia de capacidades ΔC. En el caso de requerirse la distribución energética de dichas trampas

(D_{it}-Ψ), debe de considerarse ya sea el método de alta frecuencia (método de Terman [51], [57]) o la integración de la capacidad de baja frecuencia para poder determinar Ψ–V. En este trabajo de tesis, se ha optado el segundo método, siendo este originalmente propuesto por Berglund [51], [58] para obtener la relación Ψ–V. Partiendo de la Ec. 2.9

$$\begin{aligned}\frac{d\psi_s}{dV} &= \frac{C_{ox}}{C_{ox}+C_D+C_{it}} = 1-\frac{C_D+C_{it}}{C_{ox}+C_D+C_{it}} \\ &= 1-\frac{C_{LF}}{C_{ox}}\end{aligned} \tag{2.9}$$

e integrando se obtiene

$$\psi_s(V_2)-\psi_s(V_1)=\int_{V_1}^{V_2}\left(1-\frac{C_{LF}}{C_{ox}}\right)dV+\text{ constante} \tag{2.10}$$

donde el término constante puede ser el punto inicial en la región de acumulación, ya que ψ_S es conocido y tiene una débil dependencia con la tensión aplicada.

2.1.1.3. Método de la Conductancia

Uno de los problemas que presentan los métodos basados en mediciones de capacidad es que la capacidad medida incluye no solo la capacidad debida a los estados de interfaz sino que también la capacidad del óxido y la capacidad asociada a la capa de

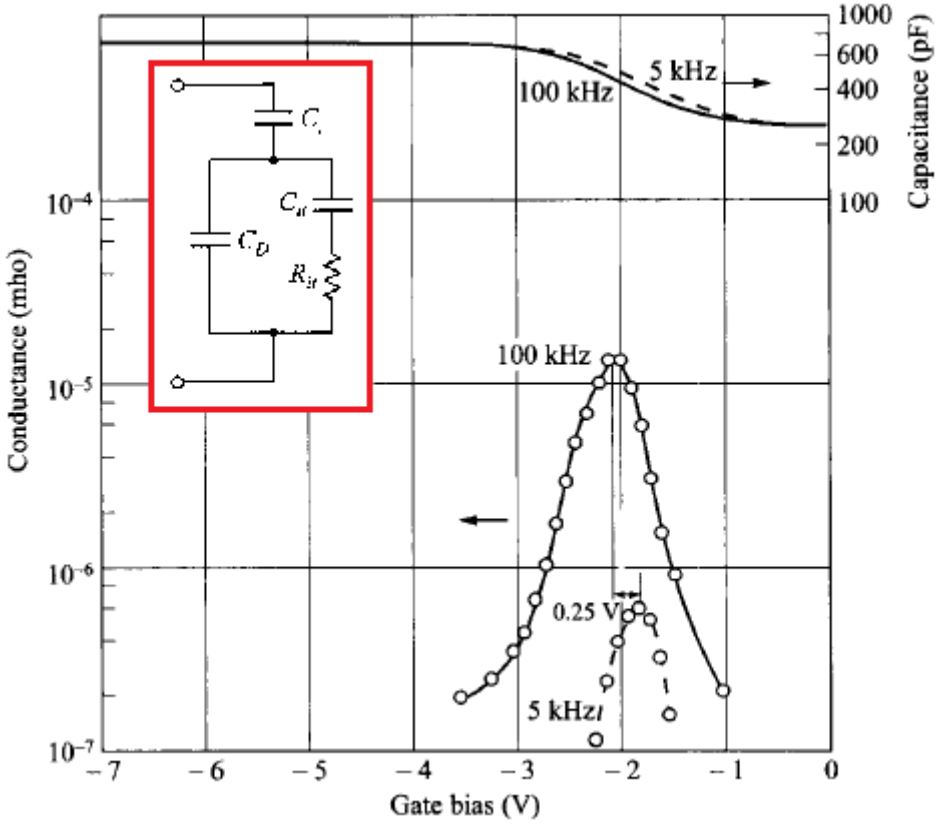


Figura 2.4: Comparación entre los cambios inducidos por la frecuencia de medición, en las curvas C–v y G–V. Reproducido de [51]

vaciamiento. A su vez, la inexactitud se ve aún más agravada si se considera que se debe calcular la diferencia entre dos mediciones de capacidad. Afortunadamente y tal como se mencionara previamente, tanto la capacidad como la conductancia en función de la tensión y la frecuencia contienen la misma información al respecto de los estados de interfaz. Una de las ventajas del segundo caso (método de la conductancia) es que la conductancia medida es directamente proporcional al D_{it}, por lo que este método permite resultados más precisos y confiables, especialmente para niveles bajos de D_{it} como sucede en sistemas Si/SiO_2 crecidos mediante oxidación térmica. Otra ventaja de este método es una mayor sensibilidad de la conductancia frente a la capacidad, como se puede observar en la Fig. 2.4 para frecuencias de prueba de 5 y 100 kHz. Para el cálculo del D_{it} se utiliza el circuito que se muestra en el inset de la Fig. 2.4. De esta forma, el primer paso es sustraer la impedancia debida a la capacidad (reactancia capacitiva) del dieléctrico C_i de la impedancia medida. Luego, la impedancia resultante se convierte a admitancia, lo que deja un paralelo entre C_D y un circuito serie R_{it}–C_{it} que modela los efectos de las trampas de interfaz. Operando algebraicamente y renombrando ciertos términos, se puede expresar la conductancia G_P normalizada a partir de la conductancia y capacidad medidas, como se indica en el primer y segundo miembro de la Ec. 2.11.

$$\frac{G_p}{\omega} = \frac{\omega C_{ox}^2 G_{in}}{G_{in}^2 + \omega^2 \left(C_{ox} - C_{in}\right)^2} = \frac{C_{it}\omega\tau_{it}}{1 + \omega^2\tau_{it}^2} \tag{2.11}$$

donde C_{in} es la capacidad medida, G_{in} es la conductancia medida, ω es la frecuencia angular y τ_{it} es el tiempo de vida medio de los defectos. Finalmente, el tercer término de la Ec. 2.11 surge de convertir el circuito paralelo del *inset* de la Fig. 2.4 en un circuito paralelo G_P-C_P. Entonces, mediante el segundo y tercer término es posible obtener C_{it} y de allí D_{it} considerando la relación $D_{it} = C_{it}/q^2$. τ_{it} es una función de ψ_S, pudiendo aproximarse su dependencia mediante las Ecs. 2.12 y 2.13

$$\tau_{it} = \frac{1}{\bar{v}\sigma_p n_i} \exp\left[-\frac{q\left(\psi_{Bp} - \bar{\psi}_s\right)}{kT}\right] \quad \text{para tipo } P \tag{2.12}$$

$$\tau_{it} = \frac{1}{\bar{v}\sigma_n n_i} \exp\left[-\frac{q\left(\psi_{Bn} - \bar{\psi}_s\right)}{kT}\right] \quad \text{para tipo } N \tag{2.13}$$

donde σ_p y σ_n son la sección eficaz de captura de huecos y electrones, respectivamente y $\bar{\nu}$ es la velocidad térmica promedio. ψ_{Bp} y ψ_{Bn} indican la diferencia entre el nivel de fermi y el nivel intrínseco del semiconductor, para semiconductores tipo P y N, respectivamente. Los resultados típicos obtenidos mediante este método sugieren, para sistemas Si/SiO_2 convencionales, un D_{it} relativamente constante en la región media de la zona prohibida, que aumenta sensiblemente en las cercanías de la banda de conducción y valencia, observándose una clara influencia de la orientación del cristal de silicio utilizado como sustrato [51].

2.1.1.4. Curvas C–V multifrecuencia de MOS sobre III-V

Tal como se discutiera en el capítulo 1, el reemplazo del silicio como material semiconductor para la tecnología CMOS está cobrando mucha importancia frente a las dificultades impuestas por el escalamiento de la tecnología estándar. No obstante, existe aún un sin número de desafíos a resolver para permitir la irrupción a escala industrial de los semiconductores de alta movilidad, como el Germanio o los semiconductores III-V (InAs, InGaAs, InP, entre otros). Para resolver estos problemas, el análisis de capacitores MOS de puerta metálica (*Metal Gate*, MG), óxidos *high–*κ y sustratos de Germanio o III-V (MG/*high–*κ/Ge o III-V) es una herramienta indispensable. En este contexto, y aunque el principio de funcionamiento básico se mantiene alrededor de lo discutido hasta el momento para estructuras MOS de Si/SiO_2, para poder obtener indicadores confiables del desempeño a partir de mediciones eléctricas, se deben realizar numerosas consideraciones en cuanto a la caracterización y modelado de estos dispositivos novedosos.

Entre otras no idealidades [55], uno de los aspectos a tener en cuenta es la mayor densidad de defectos, tanto en el volumen (Trampas de borde) como en la interfaz con el semiconductor (Estados de interfaz), ya que puede alterar las conclusiones obtenidas alrededor del comportamiento de los capacitores MOS MG/*high–*κ/III-V. El estudio de los segundos es de particular interés dado su influencia sobre el desplazamiento del nivel de Fermi, y la distribución asimétrica de estados en la banda de conducción de semiconductores III-V.

Por esta razón, la medición de la admitancia de pequeña señal a diversas frecuencias (medición "multi-frecuencia") representa una herramienta de muchísima utilidad, ya que permite obtener las curvas de Conductancia-Tensión (G–V) y C–V. Si a su vez, se combina con una variación controlada de la temperatura de la muestra, estas curvas permiten cuantificar el impacto de las no idealidades en estructuras MOS y han sido ampliamente utilizadas sobre dispositivos basados en sustratos de alta movilidad. Suponiendo una estructura MOS de compuerta metálica, óxido de alta constante dieléctrica y sustrato semiconductor de alta movilidad tipo N, una curva C–V multi-frecuencia típica, a frecuencias entre 200 Hz y 1 MHz, se asemeja a la presentada en la Fig. 2.3c. En esta, una fuerte respuesta de D_{it} (común para estos dispositivos) resultan en el incremento de la capacidad en la zona de inversión débil y el corrimiento de la curva para distintas frecuencias en la zona de vaciamiento, lo cual ha sido objeto de profundas investigaciones para su correcta cuantificación [55]. Por otro lado, la interacción de trampas de borde con los portadores del semiconductor produce un efecto conocido como dispersión con la frecuencia, caracterizado por una marcada reducción de la capacidad en la zona de acumulación C_{acc} (capacidad a tensiones positivas elevadas, y que se espera constante con la frecuencia y asociada al espesor y permitividad dieléctrica del óxido, véase Fig. 2.2), la cual se agrava con la frecuencia de medición (debido a que la interacción entre portadores y trampas está fuertemente condicionada por la frecuencia [59]-[61]).

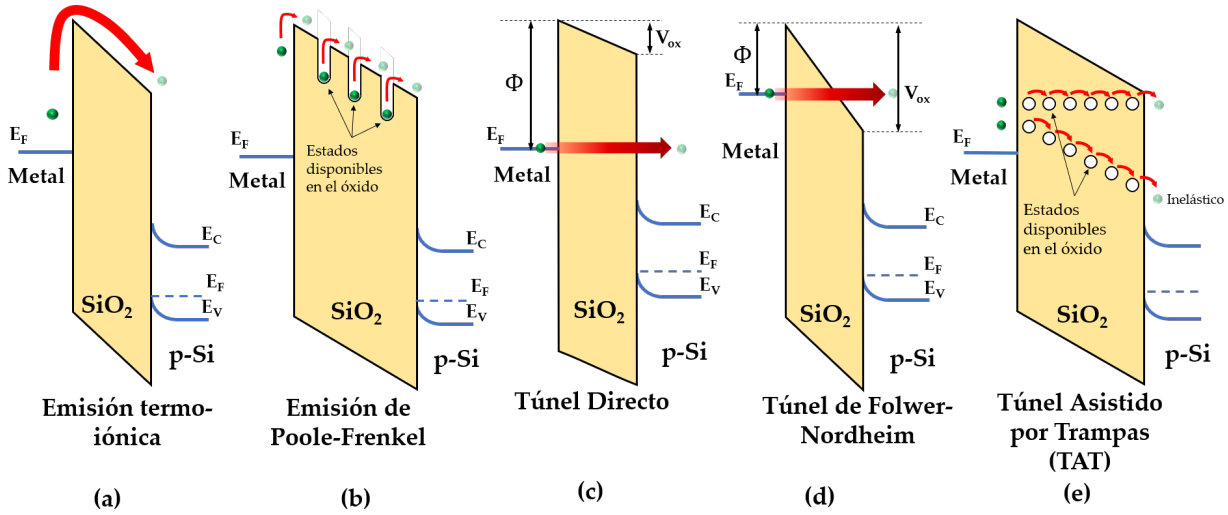


Figura 2.5: Mecanismos de conducción a través de dieléctricos: (a) Emisión Termo-iónica o Emisión de Schottky, (b) Emisión de Poole-Frenkel, (c) Túnel de Fowler-Nordheim, (d) Túnel Directo y (e) Túnel Asistido por Trampas. La esfera verde opaca representa la localización inicial del portador y la esfera traslucida la final. Se ha supuesto SiO_2 como aislante, pero los conceptos se hacen extensivos a otros dieléctricos.

2.1.2. Curvas de Corriente-Tensión (I-V)

Si bien, idealmente, un aislante es un material que no permite la conducción eléctrica, está claro que esta definición esta lejos de ser realista y es la resistividad propia del material conjuntamente con la de los electrodos y los contactos interfaciales los que determinan las características conductoras de la estructura MOS. Los mecanismos de conducción resultantes, pueden clasificarse dentro de dos grandes grupos [62]: *i)* los mecanismos de conducción asociados a los electrodos (normalmente descritos en la literatura como *"electrode-limited conduction mechanisms"*) y *ii)* los mecanismos de conducción asociados al dieléctrico en si (normalmente descritos en la literatura como *"bulk-limited conduction mechanisms"*). En el caso de los primeros, la conducción depende exclusivamente de las propiedades del contacto entre los electrodos y el dieléctrico, siendo el parámetro más importante la altura de la barrera de potencial en la interfaz electrodo-dieléctrico. Entre otros, estos mecanismos incluyen por ejemplo a la Emisión de Schottky, y el Túnel Directo y de Fowler-Nordheim. Por otro lado, los mecanismos incluidos en el segundo grupo dependen de las propiedades eléctricas del dieléctrico, especialmente del nivel de energía de las defectos (trampas) distribuidas en el volumen del óxido. En este grupo se incluyen por ejemplo, la emisión de Poole-Frenkel y el Túnel Asistido por Trampas. A continuación se describen brevemente los mecanismos mencionados en este párrafo, siendo el lector referido a la literatura especializada para una recopilación más exhaustiva de los mecanismos existentes [49], [51], [62].

La emisión Schottky corresponde a la conducción termoiónica, es decir a la inyección de portadores por encima de la barrera de potencial como se muestra en la Fig. 2.5a, mientras que la emisión Poole-Frenkel está asociada a la excitación térmica de los

electrones atrapados en la banda de conducción (véase la Fig. 2.5b). Si la película aislante es suficientemente delgada resulta imperioso considerar la corriente que atraviesa la estructura debida al mecanismo de efecto túnel. Experiencias basadas en la separación de portadores en estructuras MOS parecen indicar que los portadores responsables de la corriente son mayoritariamente los electrones [63], [64], aunque el tema continúa siendo materia de debate. Si bien los principios del mecanismo de túnel se conocen desde los albores de la mecánica cuántica, su aplicación a la física de sólidos se desarrollo recién en los años cincuenta con el advenimiento de la tecnología de materiales y en particular a las estructuras MOS a partir de los años sesenta.

Si bien se suele diferenciar la corriente en túnel en estructuras MOS en Túnel Directo (TD, esquematizado en la Fig. 2.5c) y Túnel Fowler-Nordheim (TFN, representado en la Fig. 2.5d), el mecanismo físico involucrado es el mismo. El primero de tales modos de conducción está asociado a las bajas tensiones, $V_{ox} < \Phi$, siendo V_{ox} la caída de potencial en el óxido y $\Phi \sim 3{,}2$ eV la altura de la barrera catódica (En estructuras basadas en SiO_2). Tal como muestra la figura, el túnel tiene lugar esencialmente a través de una barrera de potencial trapezoidal formada por la banda prohibida del SiO_2. La longitud de esta barrera medida en el nivel de Fermi es igual al espesor del óxido. El segundo de los regímenes nombrados tiene lugar a campos mas altos, $V_{ox} > \Phi$), y está esencialmente asociado a una barrera de potencial triangular y por lo tanto con una barrera de túnel de longitud menor que el espesor del óxido. La transición de un régimen a otro está gobernada por la tensión aplicada a la puerta, teniendo, el cambio en la forma de la barrera, un efecto notorio sobre el comportamiento general de la corriente.

Por último, el mecanismo de Túnel Asistido por Trampas (Fig. 2.5e) es más reciente y es de gran difusión en la actualidad. La idea subyacente a este tipo de conducción es que la presencia de trampas en el seno del aislante aumenta notablemente la probabilidad de transmisión a través de la estructura. Se trata del tránsito de electrones entre electrodos asistido por estados intermediarios localizados en la banda prohibida del aislante y gobernado por un proceso secuencial de captura-liberación de los portadores [65].

2.1.3. Degradación y ruptura de dispositivos MOS

Cuando la estructura MOS es sometida a un estrés eléctrico el óxido de puerta se degrada paulatinamente, con lo cual pierde sus características aislantes iniciales. Si el estrés se prolonga en el tiempo, se termina produciendo la ruptura dieléctrica del material, la cual se manifiesta como un cambio abrupto en la magnitud de la corriente que atraviesa la estructura. En aplicaciones reales, la degradación puede provenir del mismo mecanismo de conducción involucrado en el funcionamiento del dispositivo (carga y descarga en memorias de puerta flotante) o por la aparición de electrones altamente energéticos inyectados en el aislante desde el canal (electrones calientes en transistores).

A pesar de ser fenómenos muy estudiados, como consecuencia de la renovación de los materiales involucrados [66] y el profundo escalamiento de las dimensiones físicas, mecanismos de degradación conocidos con anterioridad pero que eran considerados de menor impacto, se han convertido en serias amenazas a la confiabilidad de un circuito completo. Para estudiar de manera controlada el proceso de degradación y ruptura se utilizan *tests* consistentes en aplicar una tensión o una corriente constante (CVS, por sus siglas en inglés *"Constant Voltage Stress"*, o CCS *"Constant Current Stress"*) en la puerta del dispositivo hasta detectar un cambio abrupto en la característica de conducción (ver Fig. 2.6). Alternativamente, también se utiliza una rampa de tensión (RVS por sus siglas en inglés, *Ramped Voltage Stress"*) o corriente aunque los resultados que provee este tipo de test resultan más complejos de analizar. Los test de fiabilidad para el óxido de puerta consisten en analizar un conjunto suficientemente grande de muestras y registrar el tiempo (t_{BD}) en el que se produce tal cambio abrupto. Otras descripciones posibles se basan en el registro de la carga inyectada hasta la ruptura o, para una rampa I-V, de la tensión o campo eléctrico en el que se produce el cambio en la característica.

Por lo general, la degradación del aislante se manifiesta como una deriva en alguno de sus parámetros característicos: tensión de bandas planas, corriente de túnel, etc. Sin embargo, si la degradación persiste por un tiempo prolongado, finalmente se produce la ruptura dieléctrica del material, la cual se manifiesta, tal como se ha mencionado, por un cambio abrupto en el mecanismo de conducción. Según han mostrado resultados recientes,la ruptura del aislante consiste en la aparición de "spots" o sitios conductores que actúan como cortocircuitos entre puerta y substrato [68]. Se pueden tener múltiples eventos de ruptura sobre una misma muestra tal como lo muestra la Fig. 2.7a. En dicha figura, se observa que la resistencia de la estructura varía abruptamente con la aparición de cada *spot*. A la ruptura dieléctrica del aislante de puerta se la suele clasificar en *SOFT* (SBD) o *HARD* (HBD) de acuerdo a la magnitud del evento. Las características de conducción típicas para ambos modos son las que se muestran en la Fig. 2.7b.

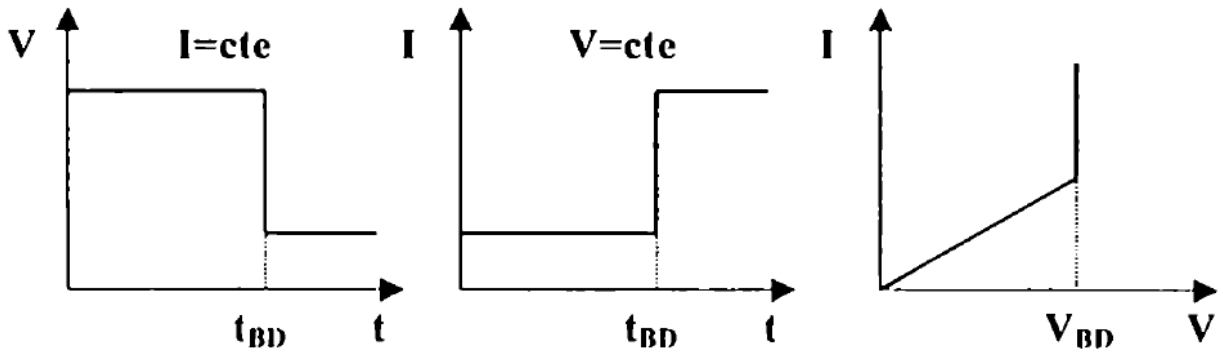
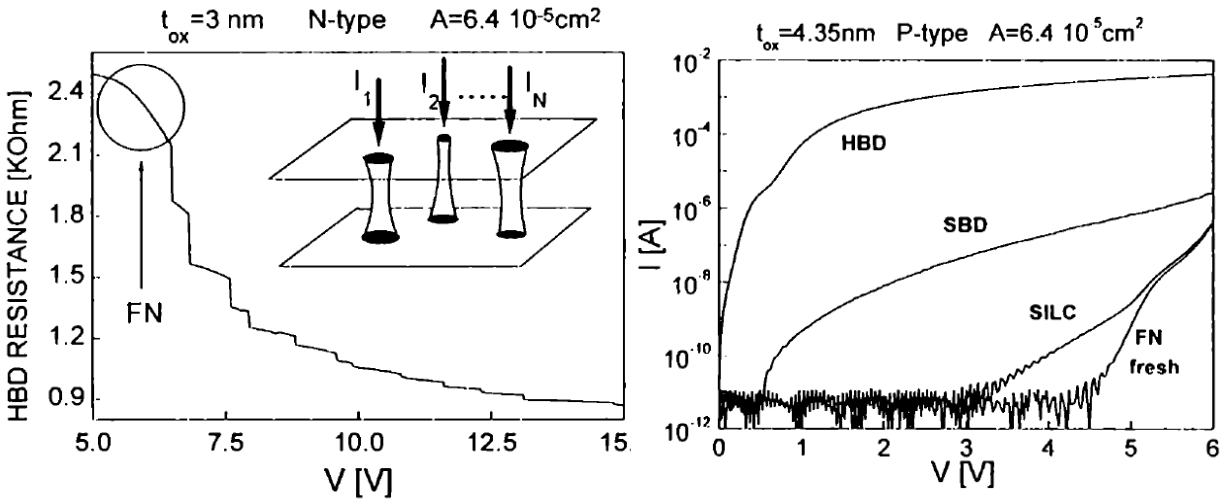


Figura 2.6: *Tests* de degradación y ruptura para las estructuras MOS. Reproducido de [67]



(a) (b)

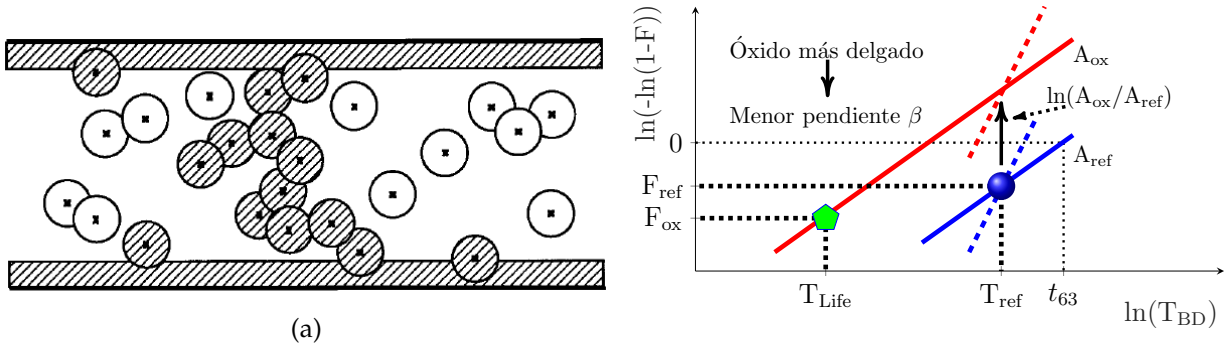
Figura 2.7: (a) Múltiples eventos (le ruptura del tipo *HARD*) en una misma muestra y una interpretación en términos de la formación de contactos localizados. (b) Características I-V típicas correspondientes a los modos de conducción *SOFT* y *HARD*. Reproducido de [67]

2.1.3.1. Ruptura dieléctrica dependiente del tiempo (TDDB)

La ruptura dieléctrica dependiente del tiempo es el proceso por el cual un material pierde sus propiedades aislantes cuando es estresado bajo un campo eléctrico de cierta intensidad por un período de tiempo suficientemente extenso. Si bien el fenómeno es ampliamente conocido, en los últimos 15 años se desarrollaron enormes avances en la comprensión del mecanismo físico en la ruptura de dispositivos nano-electrónicos [31], [69], [70]. La introducción de técnicas experimentales avanzadas de microscopía electrónica y de fuerza atómica, sumada a las técnicas tradicionales y a la enorme diversidad de dispositivos y materiales estudiados a la fecha, permiten continuar expandiendo el conocimiento en esta temática [31], [69], [70].

Desde el punto de vista experimental, el fenómeno de ruptura en dispositivos MOS o MIM (*Metal-Insulator-Metal*, Metal Aislante Metal) suele estudiarse a partir de una muestra suficientemente grande de dispositivos idénticamente fabricados, sometiéndolos a condiciones de estrés aceleradas. Tal como se explicara anteriormente, esto consiste en aplicar un campo eléctrico o potencial elevado constante pero por debajo de sus valores críticos (medición CVS) y adquirir la corriente en función del tiempo hasta el momento en que se produce la ruptura conocida como ruptura dieléctrica dependiente del tiempo (*Time Dependent Dielectric Breakdown*, TDDB). Otra alternativa es aplicar un estrés incremental (como en el caso de RVS) hasta un valor crítico de potencial V_{BD} o campo eléctrico E_{BD} al cual se produce la ruptura. Más allá del método utilizado, la ruptura es un fenómeno estocástico y por ende es representado utilizando herramientas estadísticas. En el caso del estudio por CVS, el resultado será una distribución de tiempos de ruptura, mientras que para RVS, una distribución te tensiones.

Durante muchos años, el modelo percolativo [24], [71], [72] ha representado el comportamiento del TDDB con el avance tecnológico de la nano-electrónica. Este mo-



(a) (b)

Figura 2.8: (a) Modelo esquemático de la teoría de percolación: las esferas (defectos) se generan aleatoriamente hasta formar un camino percolativo entre los electrodos, atravesando el dieléctrico. Reproducido de [27]. (b) Confiabilidad de óxidos delgados representada en un diagrama de Weibit en términos de la teoría de percolación. Reproducido de [24]

delo, de los conocidos como geométricos, se apoya en el concepto esquematizado en la Fig. 2.8a: a lo largo del tiempo, el campo eléctrico aplicado induce defectos (esferas en el diagrama) aleatoriamente distribuidos en el aislante hasta que un filamento conductivo conecta los dos terminales que el aislante separa. Este camino, al ser de elevada conductividad, resulta en la pérdida parcial o total de las propiedades dieléctricas, impactando negativamente en el desempeño o quitando de funcionamiento al dispositivo en cuestión. Bajo esta teoría, la distribución de los tiempos de ruptura para una muestra de dispositivos idénticos sigue una distribución de Weibull [24], caracterizada por una pendiente β. Por el escalamiento directo con el área del dispositivo bajo estudio, la industria encontró en este modelo una herramienta poderosa para extrapolar las condiciones de trabajo (es decir, la tensión nominal de alimentación) para garantizar una tasa de falla determinada en el tiempo de vida esperado del producto, como lo representa la gráfica de Weibit de la Fig. 2.8b. Los detalles analíticos de este modelo serán de utilidad en el desarrollo de los resultados del capítulo 4.

Analíticamente, la función densidad de probabilidad (PDF) de Weibull se caracteriza por la Ec. (2.14), donde β es el parámetro de forma o pendiente y $t_{63\,\%}$ es el tiempo de vida característico o parámetro de escala. La Fig. 2.8b representa el denominado *weibit* que, calculado como $ln(-ln(1-F))$, permite representar la distribución de Weibull en función de $ln(T_{BD})$ como una recta de pendiente β, en donde T_{BD} es el tiempo t al cual se observa la ruptura dieléctrica para cada elemento de la distribución, es decir para cada dispositivo caracterizado. El cruce por el Weibit "0"$ln(-ln(1-F)) = 0$ se corresponde con el tiempo característico $t_{63\,\%}$ de la distribución.

$$F = \frac{\beta}{t_{63\,\%}} \left(\frac{t}{t_l}\right)^{\beta-1} e^{-\left(\frac{t}{t_{63\,\%}}\right)^{\beta}} \tag{2.14}$$

Tecnológicamente, la pendiente β en la distribución de TDDB se volvió más baja a medida que los óxidos de compuerta de los dispositivos se hicieron más delgados. Considerando que los percentiles más bajos de la distribución son los de interés para garantizar una

baja tasa de fallas, una pendiente más baja (o distribución más ancha) supone un tiempo de vida más corto para una tasa determinada, como muestra la Fig. 2.8b. Otra propiedad tecnológicamente importante es el escalamiento de la distribución con el área del óxido bajo estudio [72]. Esta propiedad fundamental permite extrapolar desde un área de dispositivos de referencia A_{ref} utilizados para el estudio estadístico, al área eficaz de óxidos de compuerta en un producto final A_{ox} (en electrónica de consumo, ésta área puede alcanzar algunos milímetros cuadrados). De este modo, el tiempo de vida T_{Life} esperado para el producto considerando una tasa de fallas F_{ox} puede ser extrapolado a partir de la información disponible. Finalmente, TDDB se caracteriza por una fuerte aceleración con la tensión aplicada, que puede modelarse como una dependencia tipo ley de potencias (de la forma V^m) o bien con una función exponencial exponencial [73].

Es importante mencionar en este punto que, si bien la estadística de Weibull es una importante herramienta de estudio y toma de decisiones en lo pertinente a TDDB, el proceso de ruptura no necesariamente es catastrófico, sino que la condición de falla depende de la corriente a través del óxido que puede ser soportada por un dispositivo para que el sistema en el que se desempeña pueda seguir cumpliendo su función. De este modo, y en la búsqueda de una representación eléctrica válida del impacto de la ruptura dieléctrica sobre el correcto funcionamiento de un transistor, diversos modelos fueron propuestos, algunos mucho más elaborados que otros [74].

2.2. Fundamentos de Redes Neuronales

2.2.1. Estructura básica

Las Redes Neuronales Artificiales (*Artificial Neural Networks*, ANN) están fuertemente inspiradas en sus equivalentes biológicos, tales como el cerebro humano. Por lo tanto, las ANN no solo tienen una naturaleza morfológica similar al cerebro, sino que también replican una serie de características fundamentales, como el aprendizaje basado en la experiencia, la generalización a partir de ejemplos previos y la abstracción de las características principales de una serie de datos. Desde el punto de vista constitutivo, dos elementos sobresalen: Por un lado las neuronas, que cumplen el rol de elemento procesador, y por otro las sinapsis, que son las responsables de propagar las señales producidas por una neurona hacia las adyacentes.

De los dos elementos mencionados, la neurona (véase la Fig. 2.9 inferior izquierda) representa el bloque fundamental del sistema nervioso y en particular del cerebro. Cada neurona es una simple unidad de procesamiento que recibe y combina señales desde y hacia otras neuronas. Mientras algunas de las entradas tienden a excitar a la célula, otras inhibirla. Cuando la excitación acumulada supera un valor umbral, la neurona envía

una señal a otras neuronas de su entorno. Dicho proceso de producción y transmisión de impulsos entre neuronas biológicas fue descrito por primera vez en 1952 por Hodgkin y Huxley [75] en términos de un conjunto de cuatro ecuaciones diferenciales ordinarias no lineales, que aproximan las características eléctricas de células excitables donde canales iónicos de K y Na permiten al núcleo de la neurona recibir impulsos de potenciación y depresión. Dicho impulsos son conducidos por el axón (salida) de la neurona, cuyo extremo se ramifica y conecta a las dendritas (entradas) de otras neuronas a través de uniones llamadas sinapsis. Esta conexión es fundamental en el sistema nervioso, ya que el proceso de aprendizaje consiste a grandes rasgos en el fortalecimiento selectivo de ciertas

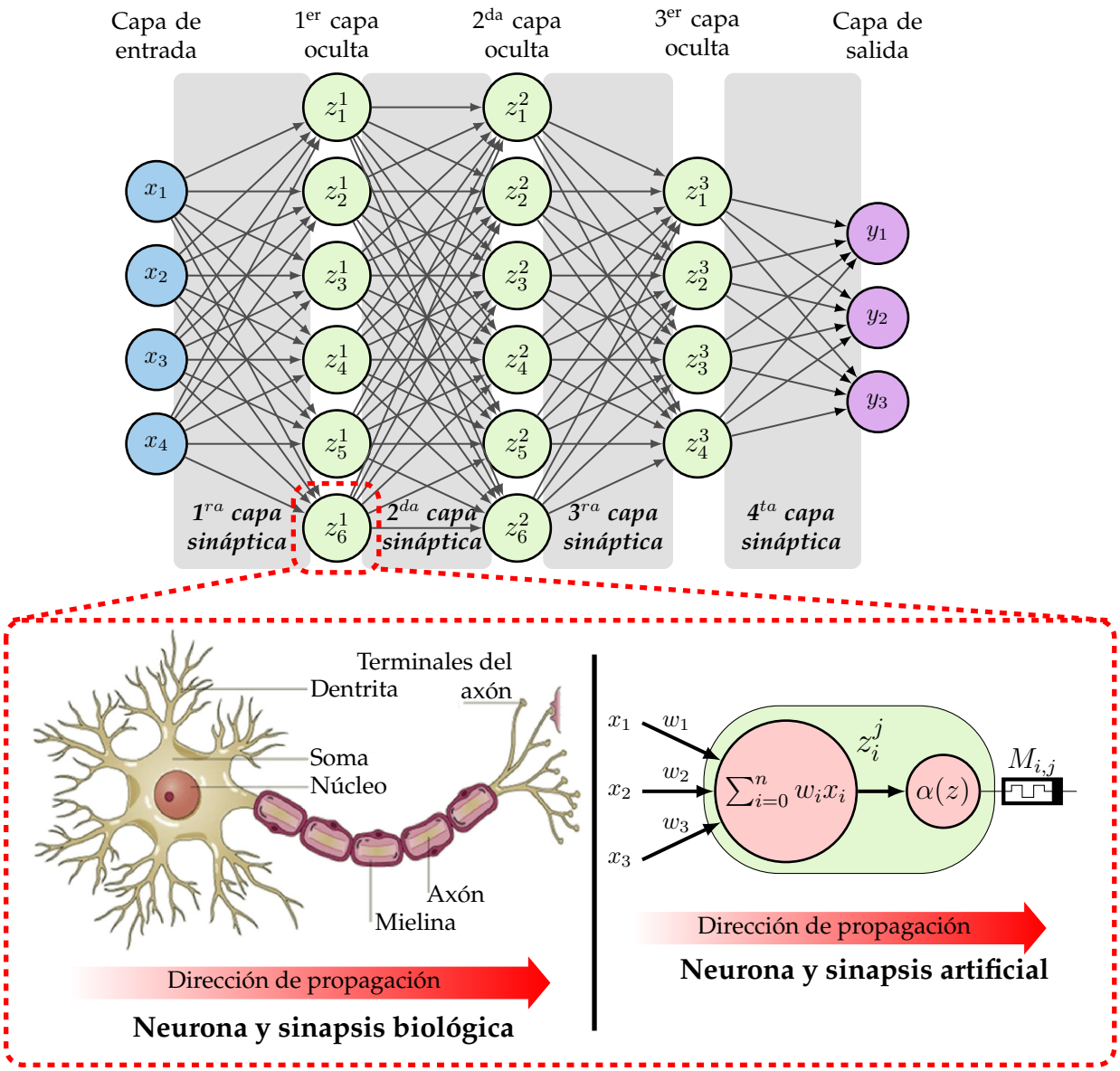


Figura 2.9: (Arr.) Estructura típica de una DNN. A modo de ejemplo se presenta un red con 4 entradas, 3 capas ocultas (con 6, 6 y 4 neuronas, respectivamente) y 3 salidas. Cada neurona se vincula con todas las neuronas de la capa anterior y posterior, a través de sinapsis artificiales. (Aba. Izq.) Representación gráfica de una neurona y sinapsis biológicas, donde el axón es responsable de posibilitar la sinapsis con otra neurona. (Aba. Der.) Representación de una neurona artificial. La misma realiza una suma ponderada de los impulsos recibidos de las neuronas de la capa anterior, y emite una señal proporcional a las neuronas de la capa siguiente.

conexiones sinápticas y el debilitamiento de otras.

En cuanto a la contra-parte artificial de la neurona, su formulación es más sencilla, siendo en su versión más elemental una suma ponderada afectada por una función de transferencia usualmente no lineal (véase la Fig. 2.9 inferior derecha). No obstante, su implementación cuando se consideran ANNs implementadas en *hardware* mediante tecnología CMOS, requiere de múltiples transistores para la realización de la suma ponderada y la función de activación, por no mencionar las limitaciones al rango de funcionamiento que esto acarrea. Si bien una célula nerviosa (neurona) tiene muchas complejidades y excepciones, este modelo básico es suficiente para el desarrollo de diversos tipos de ANNs. En dichas redes, las neuronas están organizados en grupos llamados niveles o capas, donde el tipo de red determina el conexionado particular. Una red típica consiste en una secuencia de capas con conexiones (sinapsis) entre capas adyacentes consecutivas. Existen dos capas con conexiones con el mundo exterior: Una capa de entrada, o *buffer* de entrada, donde se presentan los datos a la red, y una capa (*buffer*) de salida que mantiene la respuesta de la red a una entrada. El resto de las capas reciben el nombre de capas ocultas. La sub-figura superior de la Fig. 2.9 muestra el aspecto de una ANN. A nivel de la implementación en *hardware* la realización eficiente de las conexiones sinápticas constituye un desafío no menor, ya que deben poder mantener invariante la fuerza de la conexión sináptica (memoria no-volátil) pero a su vez permitir su actualización durante la etapa de aprendizaje o entrenamiento.

Justamente la capacidad de aprender mediante un proceso de entrenamiento es una de las características sobresalientes de las ANN. Dicha fase de entrenamiento de las ANN muestra algunos paralelismos con el desarrollo intelectual de los seres humanos, siendo el objetivo del entrenamiento de una ANN conseguir que dado un conjunto de entradas, una aplicación determinada produzca el conjunto de salidas deseadas o mínimamente consistentes. El proceso de entrenamiento consiste en la aplicación secuencial de diferentes conjuntos o vectores de entrada para que se ajuste la fuerza de las conexiones sinápticas (también llamados “pesos”) de las interconexiones según un procedimiento predeterminado. Durante la sesión de entrenamiento los pesos convergen gradualmente hacia los valores que hacen que cada entrada produzca el vector de salida deseado. Los algoritmos de entrenamiento o los procedimientos de ajuste de los valores de las conexiones de las ANN se pueden clasificar en dos grupos: Supervisado y No Supervisado. En este trabajo de tesis se considera exclusivamente el caso de entrenamiento Supervisado, motivo por el cual el mismo es detallado para el caso del algoritmo de retro-propagación en la siguiente Sub-Sección.

2.2.2. Entrenamiento supervisado y retro-propagación (*backpropagation*)

En aprendizaje automático y minería de datos, el aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento, capaz de predecir el valor correspondiente a cualquier objeto de entrada válido después de haber visto una serie de ejemplos (los datos de entrenamiento). Para ello, la red tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente. Los datos de entrenamiento consisten en pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (problemas de regresión) o una etiqueta de clase (problemas de clasificación). El proceso de entrenamiento consiste en presentar un vector de entrada a la red, calcular la salida y contrastarla contra el resultado deseado para definir el error de la red. Tal error es una función (función de costo) $\mathbb{R}^n \longrightarrow \mathbb{R}$ de los n pesos sinápticos de la red, y por lo tanto el aprendizaje puede pensarse como el proceso de minimización de la función de coste. Para ello las parejas de vectores del conjunto de entrenamiento se aplican secuencialmente y de forma cíclica (un "*Epoch*" implica la presentación de todas las imágenes de entrenamiento 1 vez). Para cada *epoch* se re-ajustan los pesos de la red hasta que el error para el conjunto de entrenamiento entero sea un valor pequeño y aceptable.

Para llegar al mínimo de la función de coste, usualmente se utiliza su gradiente. Sin embargo dada la complejidad de la alta dimensionalidad de la función de coste su cómputo ha representado un obstáculo significativo para el proceso de aprendizaje y por lo tanto para el desarrollo de las redes neuronales hasta la introducción del algoritmo de retro-propagación (*backpropagation*) en 1986 por Rumelhart *et al* [76]. La retro-propagación o propagación hacia atrás de errores es un método de cálculo del gradiente que emplea un ciclo propagación–adaptación de dos fases. En la fase de propagación, se calcula la señal de error para cada salida en función de los patrones de entrada. Luego en la fase de de adaptación, el error se propagan hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida. Sin embargo las neuronas de la capa oculta solo reciben una fracción de la señal total del error, basándose aproximadamente en la contribución relativa de cada una a la salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido una señal de error que describa su contribución relativa al error total. A medida que se entrena la red, las neuronas de las capas ocultas se organizan a sí mismas de tal modo que distintas neuronas aprenden a reconocer distintas características del espacio total de entrada. Después del entrenamiento, cuando se les presente un patrón arbitrario de entrada que contenga ruido o que esté incompleto, las neuronas de la capa oculta de la red responderán con una salida activa si la nueva entrada contiene un patrón que se asemeje a aquella característica que las neuronas individuales hayan aprendido a reconocer durante su entrenamiento.

Para explicar brevemente el algoritmo de retro-propagación se parte de una red

como la expuesta en la sub-figura superior de la Fig. 2.9, en la que en cada conexión tiene un peso $w_{l,i,j}$ (l indica la capa sináptica, i es la neurona en la capa l y j la neurona en la capa $l-1$). La función de activación de cada neurona es f e y, la función de base, es lineal. Se dispone de un conjunto de K pares de entrenamiento que se denota $\{X_K, d_K\}$, donde X es la entrada y d la salida deseada. Se define la señal de error para el nodo j de la capa de salida y el par de entrenamiento k de la siguiente manera:

$$e_j(k) = d_j(k) - y_j(k) \tag{2.15}$$

Se define el valor del error cuadrático de la neurona j como $1/2e_j^2(k)$. Consecuentemente se define el valor instantáneo $\xi(k)$ según la Ec. 2.16 como la suma de los errores cuadráticos $1/2e_j^2(k)$ en todas las neuronas en la capa de salida:

$$\xi(k) = \frac{1}{2}\sum_{j=1}^{N_L} e_j^2(k) \tag{2.16}$$

donde N_L es el número de neuronas en la capa de salida. El error cuadrático medio sobre el total K de pares de entrenamiento se obtiene entonces como se muestra en la ecuación 2.17:

$$E = \frac{1}{K}\sum_{k=1}^{K} \xi(k) \tag{2.17}$$

Tanto $\xi(k)$ como E son función de todos los parámetros libres de la red, es decir pesos sinápticos y sesgos. Para un conjunto de pares de entrenamiento, E representa la función de coste como medida del proceso de aprendizaje. Dado que la función de activación es derivable, es posible el uso de gradientes. Los ajustes de los pesos se realizan de acuerdo con los respectivos errores calculados para cada par de entrenamiento introducido a la red. El promedio de esos cambios individuales de los pesos sobre todos los pares de entrenamiento es una estimación del cambio que deben sufrir los pesos para minimizar la función de coste E sobre el conjunto de entrenamiento entero. Para detallar el proceso, se parte de una neurona como por ejemplo la indicada en la sub-figura inferior derecha de la Fig. 2.9. La corrección de los pesos sinápticos se realiza mediante un incremento en estos como se indica en la Ec. 2.18.

$$w'_{l,i,j} = w_{l,i,j} + \Delta w_{l,i,j} \tag{2.18}$$

El algoritmo de retro-propagación aplica una corrección $\Delta w_{l,i,j}(k)$ a los pesos sinápticos $w_{l,i,j}(k)$ la cual es proporcional al gradiente $\partial\xi(k)/\partial w_{l,i,j}(k)$, el cual representa la dirección de búsqueda en el espacio de los pesos para la conexión sináptica $w_{l,i,j}$. De acuerdo a la regla de la cadena se puede expresar el gradiente de la siguiente forma:

$$\frac{\partial \xi(k)}{\partial w_{l,i,j}(k)} = \frac{\partial \xi(k)}{\partial e_j(k)}\frac{\partial e_j(k)}{\partial y_j(k)}\frac{\partial y_j(k)}{\partial v_j(k)}\frac{\partial v_j(k)}{\partial w_{l,i,j}(k)} \tag{2.19}$$

Donde se sabe que

$$\frac{\partial \xi(k)}{\partial e_j(k)} = e_j(k) \tag{2.20}$$

es la derivada del error cuadrático total (Ec. 2.16) respecto del error e_j de cada salida,

$$\frac{\partial e_j(k)}{\partial y_j(k)} = -1 \tag{2.21}$$

es la derivada del error de la salida j (Ec. 2.15) respecto de la suma ponderada $y_j(k)$ de cada salida afectada por la función de activación,

$$\frac{\partial y_j(k)}{\partial v_j(k)} = f_j'\left(v_j(k)\right) \tag{2.22}$$

es la derivada de la salida y_j de cada neurona (afectada por la función de activación) respecto de la suma ponderada v_j,

$$\frac{\partial v_j(k)}{\partial w_{i,j}(k)} = y_i(k) \tag{2.23}$$

es la derivada de la suma ponderada v_j respecto de los pesos sinápticos que conectan a la neurona j de la capa l con las de la capa $l-1$.

Remplazando las ecuaciones 2.20, 2.21, 2.22 y 2.23 en la ecuación 2.19 se llega a:

$$\frac{\partial \xi(k)}{\partial w_{i,j}(k)} = -e_j(k) f_j'\left(v_j(k)\right) y_i(k) \tag{2.24}$$

La corrección $\Delta w_{i,j}(k)$ aplicada a $w_{i,j}(k)$ es definida por la regla delta:

$$\Delta w_{i,j}(k) = -\eta \frac{\partial \xi(k)}{\partial w_{i,j}(k)} \tag{2.25}$$

donde η es una constante que determina la velocidad de aprendizaje del algoritmo de retro-propagación.

Para simplificar los cálculos posteriores, la Ec. 2.25 se re-escribe como:

$$\Delta w_{i,j}(k) = \eta \delta_j(k) y_i(k) \tag{2.26}$$

donde $\delta_j(k)$ es el denominado "error imputado a la neurona" y se define como indica la Ec. 2.27

$$\delta_j(k) = -\frac{\partial \xi(k)}{\partial e_j(k)} \frac{\partial e_j(k)}{\partial y_j(k)} \frac{\partial y_j(k)}{\partial v_j(k)} = e_j(k) f_j'\left(v_j(k)\right) \tag{2.27}$$

De la ecuación 2.27 se deduce que un elemento clave envuelto en el cálculo del ajuste en los pesos $\Delta w_{ij}(k)$ es la señal de error $e_j(k)$ en la neurona de salida j. En este contexto se pueden identificar dos casos distintos dependiendo de a que capa neuronal pertenece la neurona j. Caso I: La neurona j está en la capa de salida, con lo que para

calcular la señal de error en cada nodo simplemente hay que restar de la salida deseada la señal producida por la red, tal y como indica la Ec. 2.15. Caso II: La neurona j está en una capa oculta, con lo cual las neuronas no son directamente accesibles, y comparten la "culpabilidad" en el error que se produzca en los nodos de salida.

2.2.2.1. Caso I: La Neurona j está en la capa de salida

Cuando la neurona j está ubicada en la capa de salida de la red, debe ser contrastada la salida natural con la respuesta deseada (parte del par de entrenamiento). De ahí que se use la Ec. 2.26 para calcular la señal de error $e_j(k)$ asociada a la neurona. Con $e_j(k)$ es un problema trivial el cómputo del gradiente local $\delta(k)$ usando la ecuación Ec. 2.27.

2.2.2.2. Caso II: La Neurona j está en una capa oculta

Cuando la neurona j está ubicada en una capa oculta de la red no hay una respuesta deseada especificada para esta neurona. Por lo tanto la señal de error de esta neurona se determina recursivamente en términos de las señales de error de todas las neuronas a las que está conectada la neurona en cuestión. De acuerdo con la Ec. 2.27 se puede redefinir el gradiente local $\delta_j(k)$ para la neurona oculta j como:

$$\delta_j(k) = -\frac{\partial \xi(k)}{\partial y_j(k)} \frac{\partial y_j}{\partial v_j} = -\frac{\partial \xi(k)}{\partial y_j(k)} f_j' \cdot \big(v_j(k)\big) \tag{2.28}$$

donde en el segundo miembro se ha hecho uso de la ecuación Ec. 2.22. Para calcular la derivada parcial $\partial \xi(k)/\partial y_j(k)$, se procede como sigue. Se toma la Ec. 2.16 y se reescribe cambiando el índice j por m. Esto se hace para evitar confusión con el índice j que se usa ahora para la neurona oculta.

$$\xi(k) = \frac{1}{2} \sum_{m \in N_L} e_m^2(k) \tag{2.29}$$

Si se deriva esta ecuación respecto la señal $y_j(k)$ se tiene:

$$\frac{\partial \xi(k)}{\partial y_j(k)} = \sum_m e_m \frac{\partial e_m(k)}{\partial y_j(k)} \tag{2.30}$$

Ahora se puede calcular la derivada parcial $\partial e_m(k)/\partial y_j(k)$ utilizando la regla de la cadena y reescribir la Ec. 2.30 en la forma equivalente:

$$\frac{\partial \xi(k)}{\partial y_j(k)} = \sum_m e_m(k) \frac{\partial e_m(k)}{\partial v_m(k)} \frac{\partial v_m(k)}{\partial y_j(k)} \tag{2.31}$$

$$e_m(k) = d_m(k) - y_m(k) = d_m(k) - f_m\big(v_m(k)\big) \tag{2.32}$$

siendo la neurona m una neurona de la capa de salida. De esto se sigue:

$$\frac{\partial e_m(k)}{\partial v_m(k)} = -f'_m\left(v_m(k)\right) \tag{2.33}$$

El nivel de actividad interna de la red para la neurona k es:

$$v_m(k) = \sum_{j=0}^{q} w_{m,j}(k) y_j(k) \tag{2.34}$$

donde q es el número total de entradas (sin contar el offset) aplicadas a la neurona k. Aquí de nuevo, el peso sináptico $w_{m,0}(k)$ es igual al offset $\theta_m(k)$ aplicado a la neurona k y la correspondiente entrada y_0 se fija al valor 1. Para cualquier caso, derivando la Ec. 2.34 con respecto a $y_j(k)$ conduce a:

$$\frac{\partial v_m(k)}{\partial y_j(k)} = w_{m,j}(k) \tag{2.35}$$

De esto, usando las Ecs. 2.31,2.33 y 2.35 se tiene que:

$$\frac{\partial \xi(k)}{\partial y_j(k)} = -\sum_m e_m(k) f'_m\left(v_m(k)\right) \times w_{m,j}(k) = -\sum_m \delta_m(k) \times w_{m,j}(k) \tag{2.36}$$

donde en el tercer miembro, se ha usado la definición del gradiente local $\delta_m(k)$ dada en la Ec. 2.27 con el índice m sustituido por j. Finalmente, usando la Ec. 2.36 en 2.28, se obtiene el gradiente local $\delta_m(k)$ para una neurona oculta j, tras la reorganización de términos como sigue:

$$\delta_j(k) = f'_j\left(v_j(k)\right) \sum_m \delta_m(k) w_{m,j}(k) \tag{2.37}$$

siendo la neurona j oculta. El factor $f'_j\left(v_j(k)\right)$ envuelto en la computación del gradiente local $\delta_m(k)$ en la Ec. 2.28 depende solamente de la función de activación asociada a la neurona oculta j. El otro factor, el sumatorio depende de dos conjuntos de términos. El primer conjunto de términos, $\delta_m(k)$, requiere del conocimiento de todas las señales de error $e_m(k)$, de todas las neuronas de la capa siguiente a la neurona oculta j, que están directamente conectadas a esta. El segundo conjunto términos, $w_{m,j}(k)$, consiste en los pesos sinápticos asociados a las conexiones.

3 Degradación en estructuras MOS

El reemplazo del silicio como material de sustrato en la estructura MOS por otros materiales semiconductores de mayor movilidad se ha posicionado como un paso clave en el desarrollo de nuevas tecnologías de integración que puedan sostener el escalamiento previsto por la Ley de Moore. Puntualmente el Germanio [77], [78] y los materiales compuestos formados con elementos de los grupos III y V [10], [11] de la tabla periódica han demostrado resultados muy prometedores en la realización de transistores MOSFET tipo P y N, respectivamente [12]. No obstante, la ausencia de un óxido nativo estable tal como lo es el SiO_2 a los sustratos de Silicio, la baja calidad de la interfaz resultante cuando se consideran dieléctricos *high–*κ (Ge/*high–*κ o III-V/*high–*κ) causada por el gran número de defectos (tambien llamados "estados") de interfaz y trampas de borde y las dificultades de su cuantificación respecto a Si/SiO_2 [7], continúan demorando su irrupción a escala industrial. En pos de superar estas dificultades se han planteado distintas variantes, tales como modificaciones en los procesos de deposición/crecimiento de la estructura de capas (por ejemplo, *annealing* bajo diversas atmósferas) y la introducción de dieléctricos compuestos, utilizando una capa de interfaz (*Interfacial Layer*, IL) entre el sustrato y el dieléctrico *high–*κ. En este capítulo se estudian los fenómenos de atrapamiento de carga y generación de estados de interfaz en Germanio y semiconductores III–V. Se contribuye con una explicación fundamentada de los mismos, indicando su dependencia con el proceso de fabricación y las características de los materiales aislantes involucrados.

3.1. Materiales de alta movilidad

A medida que las dimensiones físicas de los transistores se reducen, la saturación de velocidad de los portadores en el canal de los dispositivos MOSFET (principalmente causado por la emisión de fonones ópticos [79]) se ha vuelto un factor limitante en el rendimiento de los mismos [10], [77], [80]. La introducción de tensón mecánica en el canal

de Si ha sido la alternativa más exitosa hasta el momento [77] para superar este problema y mejorar la movilidad en los procesos CMOS basados en Si, dado que de esta forma se altera la estructura de bandas del canal e incrementa la ocupación del valle de baja masa efectiva [81]. Sin embargo, para poder mantener el ritmo de incremento en la cantidad de dispositivos MOSFET integrados en un *chip* de Si, se requiere un cambio de paradigma que considere los nuevos materiales propuestos para la realización del canal de los MOSFET, tal como se fija en los objetivos del *International Technology Roadmap of Semiconductors* (ITRS) [82]. Con el objetivo de cumplir con este requisito, se ha propuesto la utilización de materiales con alta movilidad de portadores y por lo tanto una mayor velocidad en el canal tales como los compuestos III-V (como el Arseniuro de Galio-Indio, InGaAs, o el fosfato de Indio, InP) —hasta 2 veces mayor movilidad de electrones que el Si [77], [78]— y el Germanio (Ge) —hasta 4 veces mayor movilidad de *huecos* que el Si [83]— . Más aún, recientemente se ha demostrado que es posible lograr una alta densidad de integración de ambos materiales (Ge e InGaAs) sobre un sustrato convencional de Si (Tecnología Ge–III-V híbrida), utilizando procesos CMOS estándar: Czornomaz *et al.* han demostrado en [12] la posibilidad de fabricar transistores NMOS de InGaAs y PMOS de SiGe con una separación de apenas 25 nm junto con celdas inversoras y memorias SRAM implementadas con dispositivos finFET. En este contexto, es de suma importancia entender los factores limitantes en los dispositivos MOS de Ge e InGaAs, dados los distintos tipos de defectos que comprometen su rendimiento y fiabilidad en el largo plazo [8], [84], [85].

En primer lugar, los compuestos III-V como el InGaAs presentan una interfaz con dieléctricos *high–*κ muy pobre (es decir, con una gran densidad de defectos [86]-[89]), escenario que se agrava en el caso de sustratos InP [55], [90]. Esto ha representado un escollo significativo para el desarrollo de las tecnologías III-V, motivo por el cual se ha puesto especial atención a la búsqueda de posibles soluciones. Como resultado se han identificado dos estrategias con resultados prometedores en términos de la producción de interfaces III-V/high–κ de alta calidad. Estas son, en primer lugar, el uso de técnicas de deposición por capas atómicas (*Atomic Layer Deposition*, ALD) para el crecimiento del óxido de compuerta y en segundo lugar, la utilización de una capa interfacial de pasivación (*Interfacial Pasivation Layer*, IPL). Con relación a la segunda, el uso de una capa delgada de Al_2O_3 como IPL entre el dieléctrico de HfO_2 y el sustrato III-V ha sido intensamente investigado [91]-[93], ya que de esta forma se combinan tanto la mayor calidad de la interfaz Al_2O_3/III-V con la alta constante dieléctrica del HfO_2.

No obstante, se ha descubierto que la densidad de defectos en el dieléctrico no solo depende de la composición de la estructura aislante, sino también del material de sustrato sobre el que se deposita [93]. Esta dependencia ha sido reportada en la bibliografía [61], [91], [93], [94] mediante el estudio de la dispersión de capacidad en función de la frecuencia para capacitores MOS polarizados en la región de acumulación. Sin embargo, no existe consenso sobre su influencia en términos del corrimiento de la tensión de banda planas (V_{FB}) e histéresis (V_{Hys}) en las curva de Capacidad–Tensión (C–V). Por

último pero no menos importante, debe mencionarse que otra alternativa para mejorar la interfaz es el tratamiento de la superficie del sustrato III-V mediante una atmósfera controlada de NH_4OH y/o *Forming Gas* (H_2/N_2) *Annealling* (FGA) ya que suprime las uniones Ga–O. Desafortunadamente se ha observado [95], [96] que esta alternativa contribuye a aumentar el atrapamiento de carga en el dieléctrico al someter el dispositivo a tensiones de estrés [95], produciendo un incremento significativo de la tasa de degradación medida mediante la histéresis de las curvas de C–V.

En segundo lugar, recientemente se ha demostrado que es posible depositar aislantes *high–*κ sobre sustratos de Ge [81], [97]. Entre los dieléctricos *high–*κ considerados, el HfO_2 emerge como un candidato atractivo dada su conocida integración con sustratos de Si. Sin embargo, y tal como sucede en sustratos III-V, el HfO_2 ha demostrado ser inestable al ser depositado sobre GeO_2 [98]. Por otro lado, Al_2O_3 forma una mejor interfaz con GeO_2 [99] pero tiene una constante dieléctrica relativamente baja. Por lo tanto, al igual que para el caso anterior de sustratos III-V, en la literatura usualmente se considera una estructura multi-capa ($Ge/GeO_2/Al_2O_3/$*high–*κ). No obstante, la baja constante dieléctrica de la bi-capa GeO_2/Al_2O_3 reduce sensiblemente la constante dieléctrica equivalente de la estructura compuesta. Con el fin de compensar esta reducción, se ha optado por incluir aleaciones entre HfO_2 y otros óxidos metálicos para obtener un óxido ternario metálico con una mayor constante dieléctrica.

Pero en cualquier caso, la interfaz Ge/aislante no está exenta de defectos, con lo que se han ensayado distintos métodos de pasivación. Entre los procedimiento estudiados se encuentran la nitridación, la pasivación mediante Fluor y los tratamientos mediante sulfuro [97], [100]-[102]. En particular, se ha reportado una reducción del D_{it} en capacitores MOS fabricados sobre (100)-Ge [101] (100 indica la dirección cristalográfica) mediante el tratamiento con FGA posterior a la deposición del electrodo de compuerta, tanto para los casos de utilizar HfO_2 y Al_2O_3 como material de aislante [84], [101], [103]. Sin embargo, a pesar de la vasta literatura disponible sobre la buena pasivación de la interfaz y el excelente rendimiento obtenido con este método, muy pocos estudios de fiabilidad en estructuras MOS de Ge se han reportado en relación a la pasivación de los estados de interfaz mediante FGA. A pesar de que se ha mostrado que los efectos de Inestabilidad por Tensión Negativa (*Negative Bias Temperature Instability*, NBTI) se pueden reducir en estructuras SiO_2/Ge [104]-[106], se requiere más información para describir y modelar los mecanismos responsables de la degradación en estructuras más novedosas tales como aquellas del tipo *high-*κ/Ge.

3.2. Sustratos de compuestos III-V

Con el objetivo de estudiar la distribución de defectos en el óxido en estructuras MOS bi-capa realizadas en sustratos III-V se han realizado mediciones C–V sobre dispo-

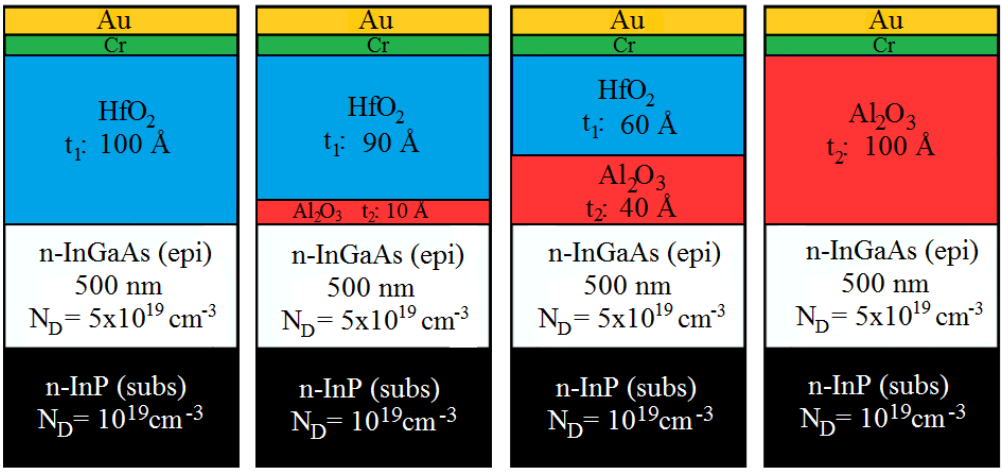


Figura 3.1: Estructura de las muestras utilizadas en este estudio. El espesor relativo de los films que forman la bi-capa aislante son (de izquierda a derecha): 0 %-100 %, 30 %-70 %, 40 %-60 % y 100$-0 %. Nótese que las figuras no están a escala.

sitivos MOS con diferentes proporciones de dieléctricos de HfO_2 y Al_2O_3 pero con un espesor constante e igual a 10 nm del dieléctrico compuesto[1]. Ambos dieléctricos fueron depositados por *Atomic Layer Deposition* (ALD), mientras que el electrodo de compuerta es un bi-capa de Cr/Au depositado mediante la técnica de *lift-off*. En todos los casos considerados, la capa de Al_2O_3 cumple la función de IPL, con espesores en el rango de 0,5 nm a 4 nm (con lo que la capa de HfO_2 restante tiene espesores de entre 9,5 nm y 6 nm). Adicionalmente se han considerado estructuras de control conteniendo una única capa, tanto de Al_2O_3 como HfO_2, a modo de referencia. Dos tipos de sustrato fueron considerado, siendo estos $In_{0,53}Ga_{0,47}As$ (5×10^{16} cm^{-3}) e InP (4×10^{16} cm^{-3}) tipo N, ambos con un espesor de 500 nm y fabricados mediante crecimiento epitaxial por haces moleculares metal-orgánicos (*Metal-Organic Molecular Beam Epitaxy*, MOMBE) sobre un sustrato tipo N de InP altamente dopado (10^{19} cm^{-3}). En la Fig. 3.1 se presenta un resumen de algunas de las estructuras estudiadas. Para prevenir un aumento de la corriente de fuga a través del dieléctrico, se omitió cualquier tipo de recocido (*annealing*) post-deposición. Detalles relacionados al proceso de fabricación pueden ser encontrados en la bibliografía [61][91], [94]. El análisis de la histéresis de la curva C–V y el corrimiento de la tensión de bandas planas (V_{FB}) permite inferir la distribución energética de los defectos, su contribución al funcionamiento del dispositivo y el impacto de los diferentes materiales de sustrato en la calidad de la estructura MOS [59].

3.2.1. Dispersión de capacidad — impacto de las Trampas de Borde

La Fig.3.2 muestra los resultados de la caracterización mediante mediciones C–V a múltiples frecuencias (*Multi-Frequency* C–V, MFCV) de los dispositivos fabricados con

[1]Dispositivos provistos por el Departamento de Ciencia de Materiales e Ingeniería del Instituto Technion de Israel

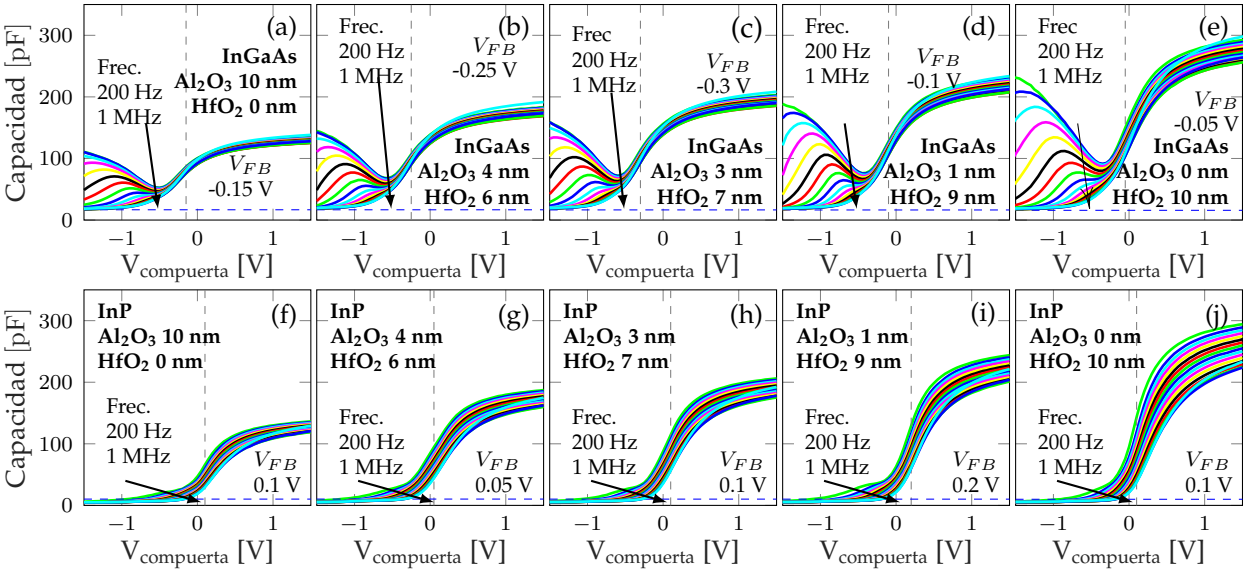


Figura 3.2: Curvas MFCV medidas entre 200 Hz y 1 MHz para muestras de InGaAs (arriba) e InP (abajo) con una proporción de Al_2O_3-HfO_2 de: 0 %-100 % (a) y (f), 60 %-40 % (b) y (g), 70 %-30 % (c) y (h), 90 %-10 % (d) y (i), 100 %-0 % (e) y (j). Las lineas de trazo discontinuo azul y negra indican en cada caso, la capacidad mínima en inversión y la tensión V_{FB}, respectivamente.

una composición variable del bi-capa de Al_2O_3/HfO_2 (0 %-100 %, 40 %-60 %, 30 %-70 %, 10 %-90 % y 100 %-0 %). Se puede apreciar un incremento de la capacidad en acumulación a medida que aumenta el espesor de la capa de HfO_2 en las Figs. 3.2a a 3.2e (para InGaAs) y las Figs. 3.2f a 3.2j (para InP). Esto puede ser explicado teniendo en cuenta la alta constante dieléctrica del HfO_2 ($e_{r_{HfO_2}}$~17.5) en comparación al Al_2O_3 ($e_{r_{Al_2O_3}}$~7). Para cada muestra, la tensión V_{FB} (indicada en cada sub-figura de la Fig. 3.2 mediante las lineas grises de trazo discontinuo) ha sido extraída mediante la técnica del punto de inflexión en la curva C–V medida a 500 kHz [53]. En acumulación, la capacidad se aproxima al valor teórico de la capacidad del óxido C_{Ox}. Asimismo, las lineas de trazo azul discontinuo en cada sub-figura representan la capacidad mínima teórica, mostrando concordancia entre los valores medidos y calculados lo cual permite descargar el enclavamiento del nivel de Fermi (*Fermi-Level Pinning*) en la interfaz óxido/semiconductor [55].

Con el propósito de poder realizar una comparación cuantitativa, el D_{it} para cada caso fue calculado utilizando el método de Alta-Frecuencia Baja-Frecuencia (*High-Frequency Low-Frequency*, HF-LF) [55]. Para las estructuras de InGaAs, el D_{it} calculado varía entre $1{,}2 \times 10^{12}$ $cm^{-3}eV^{-1}$ y 2×10^{12} $cm^{-3}eV^{-1}$, mientras que para InP la variación se da entre 5×10^{11} $cm^{-3}eV^{-1}$ y 1×10^{12} $cm^{-3}eV^{-1}$, lo cual indica que no existe una influencia significativa del espesor del IPL sobre la densidad de estados de interfaz. No obstante, a pesar de los valores similares de D_{it} para InGaAs e InP, el llamado "*weak inversion hump*" claramente visible en las muestras de InGaAs no se observa en sus contrapartes de InP. Dicho "*weak inversion hump*" usualmente se atribuye a la interacción entre los estados de interfaz y los portadores minoritarios en la región de inversión débil. Pero para el caso de las muestras de InP, la tasa de generación-recombinación es mucho más lenta que la frecuencia de medición y por lo tanto su interacción se puede considerar

despreciable [90]. Otra importante diferencia a señalar entre los sustratos de InGaAs e InP es la mayor dispersión de la capacidad en acumulación en función de la frecuencia en las segundas, lo cual puede ser relacionado con una mayor densidad de trampas de borde en las estructuras de InP [61], [93]. A su vez, las diferentes constantes de tiempo asociadas al intercambio de carga por efecto túnel entre las trampas de borde y el semiconductor [60] contribuyen a producir un aumento de la dispersión de frecuencia a medida que se incrementa la porción de HfO_2 de la bi-capa aislante. Por último vale mencionar que aunque mucho menor, tal dispersión de frecuencia puede observarse para tensiones por debajo del régimen de acumulación.

3.2.2. Histéresis de capacidad — impacto de la capa interfacial

La Figura 3.3a muestra una sucesión de curvas C–V medidas considerando un barrido de tensión que parte de una tensión inicial (V_{start}) variable en inversión, y aumenta hasta llegar a una tensión de estrés fija en acumulación V_{stress} (curva *forward*), para luego retornar a la tensión V_{start} (curva *backward*). Para el caso del estrés a tensiones negativas, en cada ciclo V_{start} se reduce gradualmente, mientras que se mide la histéresis para la condición de banda planas (calculada como la diferencia entre la tensión V_{FB} medida en la curva *forward* y la tensión V_{ret} de la curva *backward* que produce la misma capacidad C_{FB} que se obtiene para V_{FB}). Análogamente, la Fig. 3.3b presenta los resultados de las curvas correspondientes al estrés positivo. En este caso la tensión V_{start} es se mantiene constante mientras que V_{stress} se incrementa gradualmente en cada ciclo. Tales mediciones se repitieron para cada uno de las estructuras consideradas, y los resultados en términos ΔV_{FB} (calculado como $V_{FB_{i-cycle}} - V_{FB_0}$) y ΔV_{Hys} (calculado como $V_{Hys_n} - V_{Hys_o}$, es decir,

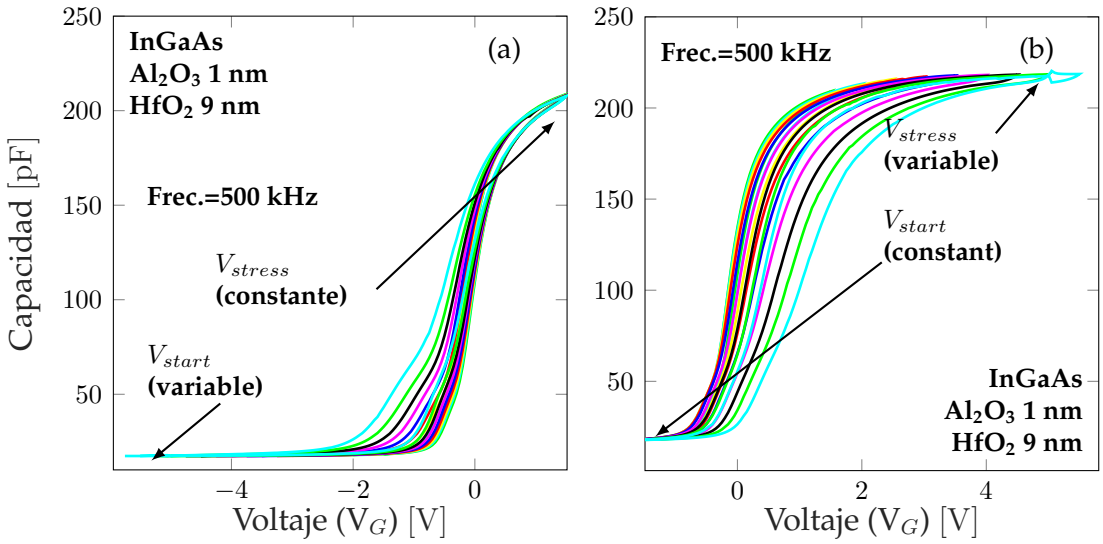


Figura 3.3: Curvas de estrés C–V a una única frecuencia (SFCV) resultantes de variar (a) V_{start} y (b) V_{stress}.

el incremento de la histéresis con respecto al primer lazo) se presentan en la Fig. 3.4 en función de las tensiones V_{start} y V_{stress}, tanto para InGaAs como para InP.

La tendencia general indica que, independientemente del sustrato utilizado, tanto el estrés a tensión positiva como negativa causan un corrimiento de V_{FB} tal como se puede apreciar en las Figs. 3.4a, 3.4b, 3.4e y 3.4f. Esto se debe a que durante la curva *backward* no toda la carga atrapada durante la curva *forward* es liberada [59], [107]. Con relación a la dependencia del ancho del lazo de histéresis con las tensiones de estrés positivo y negativo (V_{stress} y V_{start}, respectivamente) presentada en las Figs. 3.4c, 3.4d, 3.4g y 3.4h, se puede observar en ambos casos, un aumento en ΔV_{Hys}. Para el caso del estrés a tensión negativa (Figs. 3.4c y 3.4g) el incremento de ΔV_{Hys} se puede explicar al considerar la existencia de una amplia distribución de defectos en las estructuras MOS III-V/*high*–κ [59]. En este escenario una fracción significativa de los defectos del óxido con energías debajo de la banda de conducción del semiconductor permanecen cargados con electrones en la condición de bandas planas [59], [108]. Suponiendo entonces una tensión V_{start} para el inicio de la curva *forward* por debajo de V_{FB} (i.e. $V_{start} < V_{FB}$ en las Figs. 3.4c y 3.4g), una fracción de estos defectos en el óxido comienza descargado y por lo tanto contribuye a atrapar carga durante el doble barrido C–V. A medida que V_{start} decrece, la fracción inicial de defectos descargados aumenta, y por lo tanto aumenta también ΔV_{Hys} [107]. Finalmente, el incremento lineal de ΔV_{Hys} para el caso de estrés positivo (Figs. 3.4d y 3.4h) puede explicarse análogamente, considerando en este caso una distribución amplia de defectos cerca del nivel de Fermi [59].

Habiendo presentado ya las tendencias generales de las mediciones reportadas, es importante realizar una disquisición entre las estructuras de InGaAs e InP. Si bien en

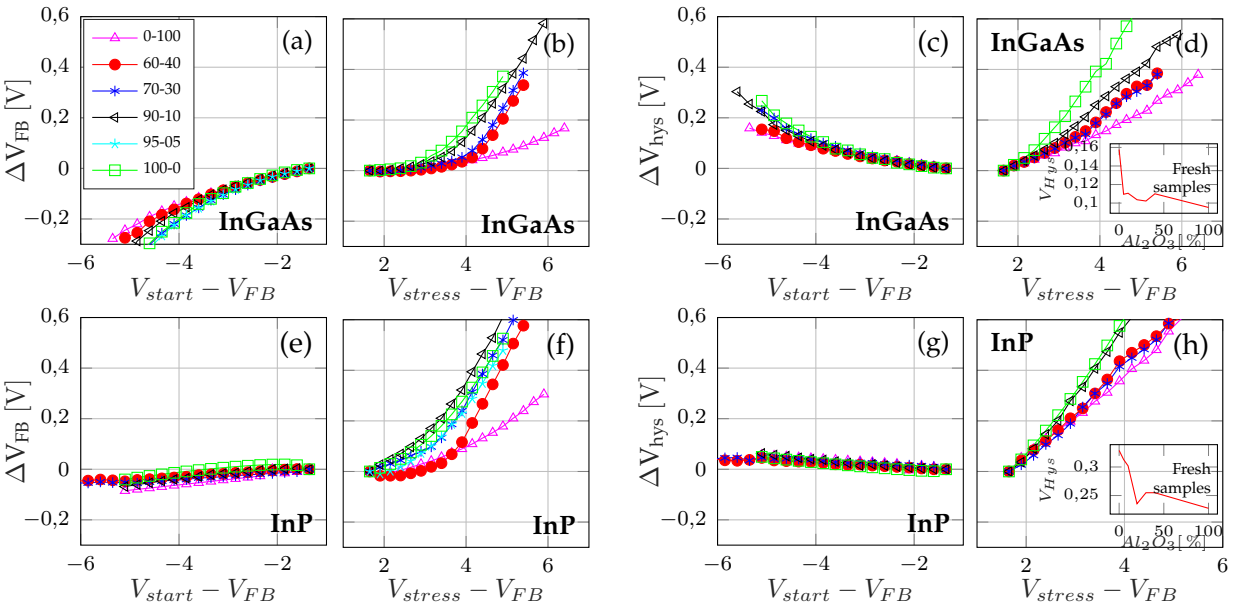


Figura 3.4: ΔV_{FB} y ΔV_{Hys} medidos para las muestras bi-capa, tanto para el estrés negativo como positivo. ΔV_{FB} para InGaAs bajo estrés negativo (a) y positivo (b). ΔV_{Hys} para InGaAs bajo estrés negativo (c) y positivo (d). ΔV_{FB} para InP bajo estrés negativo (e) y positivo (f). ΔV_{Hys} para InP bajo estrés negativo (g) y positivo (h)

ambos casos ΔV_{FB} se mueve hacia tensiones más positivas o negativas según sea el tipo de estrés, la magnitud de dicho corrimiento es mayor para el estrés negativo que en el positivo para InGaAs, mientras que en InP existe una descarga casi total de la carga atrapada durante el estrés negativo, con lo que V_{FB} prácticamente permanece invariante. Dicho comportamiento indica que la densidad de defectos presentes en el óxido con energías por encima del nivel de Fermi resulta mayor cuando el óxido es depositado sobre un sustrato de InP, independientemente de que el D_{it} sea similar para ambas estructuras (InGaAs/*high–*κ y InP/*high–*κ).

Esto puede justificarse por la naturaleza inerte de la superficie del sustrato InP en comparación al InGaAs, debido a la mayor estabilidad de las uniones In–P (198 kJ/mol) frente a las uniones In–As (94 kJ/mol). Como resultado, la reactividad de los precursores metal-orgánicos con la superficie del sustrato InP puede ser menor que en el de InGaAs, y las primeras mono-capas dieléctricas depositadas sobre el InP pueden incluir una mayor concentración de defectos [93], [94]. Dado que el D_{it} resulta similar entre ambos sustratos, se puede concluir que dichos defectos se concentran en el volumen del dieléctrico (Trampas de Borde), pero lejos de la interfaz. Con respecto a las diferencias encontradas para el caso de ΔV_{Hys}, se observa una fuerte dependencia con V_{stress} para los dispositivos InP (Fig. 3.4h), mientras que la dependencia con V_{start} es prácticamente despreciable (Fig. 3.4g). Estas diferencias sugieren que la distribución de defectos en el óxido se localiza completamente alrededor de la banda de conducción del semiconductor. De esta forma, ya para condición de bandas planas, todos los defectos del óxido se encuentran descargados y por lo tanto pueden contribuir en su totalidad a la histéresis. De esta forma, no importa que tanto decrezca V_{start} la cantidad de defectos descargados no variará (ya que todos se encuentran descargados) y por lo tanto tampoco lo hace la histéresis medida ($\Delta V_{Hys} \sim 0$, véase la Fig. 3.4g).

Finalmente, la dependencia observada en las mediciones en la composición de la estructura para un determinado sustrato debe ser mencionada. Las mediciones del ΔV_{Hys} para los dispositivos de 100 % HfO_2 y 100 % Al_2O_3 estresados positivamente (Figs. 3.4d y 3.4h), revelan un incremento más pronunciado para los dispositivos con un dieléctrico exclusivamente de HfO_2 (más carga atrapada), lo que está en concordancia con los resultados reportados en [109]. Este fenómeno puede explicarse teniendo en cuenta la mayor densidad de defectos presente en las primeras mono-capas depositadas de HfO_2. Adicionalmente, los dispositivos bi-capa conteniendo una porción creciente de Al_2O_3 (60 %-40 %,70 %-30 % y 90 %-10 %) fueron también analizados. A pesar de que se ha propuesto que el fenómeno de atrapamiento de carga sucede en una región planar cerca de la interfaz, y que es independiente del espesor del óxido (se indica en [109] para $t_{ox} > 5$ nm), los datos experimentales aquí presentados para distintos espesores de IPL muestran distintos comportamientos en términos de la histéresis medida para IPLs delgadas, que se asemejan al caso de 100 % HfO_2 a medida que la porción de HfO_2 en el bi-capa se incrementa.

Dentro de este marco, V_{Hys} puede calcularse de acuerdo a la Eq. 3.1:

$$V_{Hys} = \frac{Q_{trapped}\left(\epsilon_{\mathrm{Al_2O_3}} t_{\mathrm{HfO_2}} + \epsilon_{\mathrm{HfO_2}} t_{\mathrm{Al_2O_3}}\right)}{\epsilon_o \epsilon_{\mathrm{Al_2O_3}} \epsilon_{\mathrm{HfO_2}}} \tag{3.1}$$

donde $Q_{trapped}$ es la carga neta atrapada, ϵ_o, $\epsilon_{Al_2O_3}$ y ϵ_{HfO_2} son las constantes dieléctricas del vacío y relativas del Al_2O_3 y HfO_2, mientras que $t_{Al_2O_3}$ y t_{HfO_2} son los espesores de las capas de Al_2O_3 y HfO_2, respectivamente. Asumiendo que $Q_{trapped}$ se mantiene constante para distintos espesores de IPL, se podría esperar que la histéresis aumente a medida que aumenta la porción de Al_2O_3 (La capacidad del dispositivo aumenta). Sin embargo en el *inset* de las Figs. 3.4d y 3.4h se observa un comportamiento diametralmente opuesto, lo cual sugiere que $Q_{trapped}$ lejos de mantenerse constante, aumenta a medida que se reduce el espesor de la IPL para espesores por debajo de los 5 nm. Nótese que $Q_{trapped}$ es una representación en la forma de una "hoja de carga" de la carga atrapada a distintas distancias de la interfaz semiconductor/óxido. En este contexto, y para el caso de dieléctricos delgados, la densidad de defectos equivalente en la interfaz y su distribución energética serán determinadas en forma conjunta mediante el IPL (material y espesor) y el dieléctrico depositado sobre el mismo. Por lo tanto, muestras con una menor proporción de Al_2O_3 exhiben un comportamiento marcadamente similar a las muestras 100 % HfO_2, mientras que las muestras 60 % HfO_2-40 % Al_2O_3 tienden a comportarse como las 100 % Al_2O_3.

3.3. Sustratos de Germanio

Habiendo abordado en primera instancia los dispositivos fabricados en sustratos III-V, en esta segunda sección se pondrá el foco sobre los sustratos de Germanio. Para ello, se realizaron experimentos con dispositivos[2] fabricados sobre una oblea de Ge sobre Si, donde la capa de Ge fue crecida epitaxialmente hasta lograr un espesor de 1 μm. Posteriormente, la superficie de la oblea fue limpiada por 30 segundos en una solución de HF (2 wt. %) para remover el óxido nativo. Acto seguido, la superficie de Ge fue re-oxidada de forma controlada para obtener un film de GeO_2 de $\sim$0.7 nm [99], [110]. La re-oxidación se produce como un efecto secundario de la deposición por ALD de una capa de $\sim$2 nm de Al_2O_3 utilizando trimethylaluminum (TMA)/O_3 como precursor. Durante dicho proceso de ALD se ha reportado también la formación de GeO_x [111], [112] (Germanio en estados de oxidación menores a +4). Sobre la estructura de $Ge/GeO_2/Al_2O_3$ resultante se depositó una capa de $\sim$4 nm de un dieléctrico *high–*κ a base de Hf. Tales dieléctricos consisten en HfO_2, $HfAlO_x$ y $HfGdO_x$ y en todos los casos la deposición se realizó mediante ALD. Los precursores utilizados para la deposición de la capa de *high–*κ fueron $HfCl_4$ y H_2O para la deposición de HfO_2 y TMA, $Gd(iPrCp)_3$, $HfCl_4$ y H_2O para las deposiciones

[2]Dispositivos provistos por el Departamento de Ciencia de Materiales e Ingeniería del Instituto Technion de Israel

de $HfAlO_x$ y $HfGdO_x$, respectivamente. Por último, se depositó una capa de 40 nm de Pt como electrodo de compuerta. Parte de las muestras fabricadas fueron sometidas a un *annaeling* en FG durante 30 minutos a una temperatura de 400 °C (muestras FGA), mientras que el resto fue resguardado como muestras de control (muestras no-FGA). Estas condiciones de *annealing* fueron elegidas dado que permiten obtener mejores resultados en términos de pasivación de defectos que el caso del *annealing* en vacío (10^{-7}Torr, 30 minutos a 400 °C). En resumen las muestras consideradas en este estudio tienen unos 7 nm de espesor. Más detalles al respecto del proceso de fabricación puede encontrarse en la literatura [99], [110], [113].

Los dispositivos previamente descritos fueron analizados mediante mediciones C–V realizadas con un analizador de impedancias Agilent 4285A y mediciones de I–V e I–t —también denominadas mediciones bajo estrés a tensión constante (*Constant Voltage Stress*, CVS)— utilizando una unidad de medición Keithley 2636B. Durante las mediciones CVS, la tensión de estrés fue interrumpida periódicamente a fin de realizar mediciones C–V para cuantificar la degradación de los parámetros del dispositivos, tales como la tensión de bandas planas (V_{FB}) y la tensión de histéresis (V_{hys}). Para cada condición de estrés, 10 dispositivos fueron medidos. Asimismo, para evitar errores de medición asociados a la fase de relajamiento de carga, se mantuvo un pequeño tiempo de espera (constante en 100 mSeg.) entre la remoción de la tensión de estrés y el inicio de la medición C–V. Por último el calculo de la tensión de *flat-band* fue realizado mediante la técnica del punto de inflexión [53].

3.3.1. Rol de los defectos en el atrapamiento de carga

En las Figuras 3.5a y 3.5b se pueden apreciar las curvas C–V medidas a 200 kHz para ambos *sets* de muestras (FGA y no-FGA) respectivamente. En ambos casos, las diferencias entre los distintos materiales *high–*κ basados en Hf son significativas. Teniendo en cuenta que la estructura multi-capa dieléctrica tiene un espesor total de $\sim$7 nm (corroborado por Microscopía Electrónica de Transmisión, TEM [99]) en todos los casos, la variación de la capacidad medida en la región de acumulación indica cambios en la constante dieléctrica de la estructura. En este sentido, la constante dieléctrica efectiva para una estructura multi-capa de $GeO_2/Al_2O_3/$*high–*κ se ha reportado en la literatura [99] y corresponde a 8.5, 7.7 y 9.5 para HfO_2, $HfAlO_x$ y $HfGdO_x$, respectivamente. Nótese que debido a la capa de interfacial de GeO_2 (ubicada entre el sustrato de Ge y la bi-capa $Al_2O_3/$*high–*κ) los valores mencionados de constante dieléctrica son más bajos de lo que se esperaría. Dado que la única diferencia en las muestras radica en la capa de *high–*κ, puede observarse que la adición de Gd a la capa de Hf permite aumentar la constante dieléctrica efectiva mientras que el agregado de Al la reduce [99].

Por otro lado, las figuras 3.5c y 3.5d presentan las curvas C–V multi–frecuencia

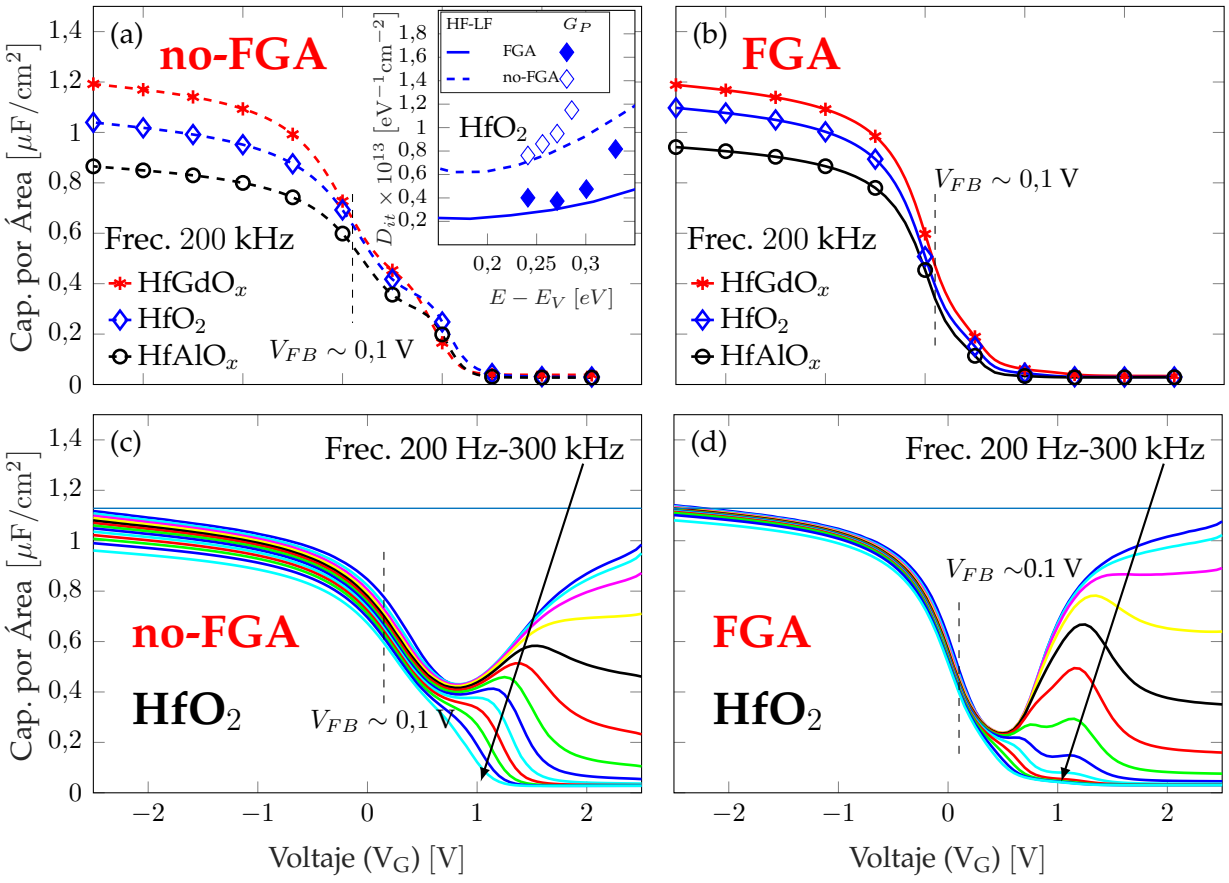


Figura 3.5: Curvas C–V para las muestras bajo estudio, medidas a 200 kHz. (a) Muestras sin tratamiento de FGA (no-FGA, control) y (b) muestras con tratamiento de FGA (FGA). En cada gráfico, la linea de trazos indica la tensión de *flat-band*, calculada de acuerdo con la técnica del punto de inflexión. En el *inset* de (a) se muestra el D_{it} para las muestras de HfO_2 calculado mediante *i)* el método de Alta Frecuencia-Baja Frecuencia y *ii)* el método de la conductancia paralelo. En ambos casos, el D_{it} es mayor para las muestras de control (no-FGA). (c) y (d) muestran las curvas C–V medidas a múltiples frecuencias (200Hz–300 kHz) (MFCV) para las estructuras de HfO_2, antes y después del tratamiento de FGA. Nótese que hay una clara reducción tanto de la dispersión del valor de capacidad medido en acumulación así como del llamado "*weak inversion hump*", lo cual es consistente con la reducción observada del D_{it}

(200 Hz–300 kHz) medidas a temperatura ambiente (*Room Temperature*, RT) para las muestras FGA y no-FGA de HfO_2, respectivamente. La primera diferencia notoria a señalar es la mayor dispersión en frecuencia observada en las muestras no-FGA (Fig. 3.5c) para la región comprendida de vaciamiento a acumulación (Voltajes de compuerta menores a V_{FB}). Esto indica que el proceso de FGA logra efectivamente pasivar los defectos cuyas energías se encuentran cerca de la banda de valencia (*Valence Band*, VB) del Ge [101]. Los resultados obtenidos para $HfGdO_x$ y $HfAlO_x$ no se muestran dada su similitud a los de HfO_2. Con respecto a la característica de dispersión (también llamada *"Weak Inversion Hump"*) observada en la región restante (vaciamiento a inversión, es decir, tensiones de compuerta mayores a V_{FB}) esta se debe al atrapamiento y des-atrapamiento de electrones y/o huecos en defectos cuyas energías van desde el *mid-gap* hasta la banda de conducción (*Conductance Band*, CB) y se encuentran cerca de la interfaz con el GeO_2 [101].

Dado que el área bajo el denominado *"weak inversion hump"* y el "estiramiento" de la curva C–V son proporcionales a la cantidad de estados en el *mid-gap* [99], se puede asumir una reducción de los mismos para todas las muestras luego del FGA, independientemente de la capa de *high–*κ. Esto está en concordancia con resultados previos presentados tanto por Fadida *et al.* [99] para muestras experimentales con una composición muy similar, como también reportes experimentales indicando la reducción en la densidad de estados de interfaz para dieléctricos de HfO_2 y Al_2O_3 depositados por ALD [101], [114], [115].

A fin de cuantificar este último item, el D_{it} tanto para las muestras FGA como no-FGA fue calculado en la región de *mid-gap* utilizando dos métodos diferentes: mediciones C–V de Alta Frecuencia y Baja Frecuencia (Castagné-Vapaille) y mediciones de conducantacia paralela [50], [55], [56]. En el inset de la figura 3.5a se muestra el perfil de D_{it} alrededor de *mid-gap* para el caso de HfO_2 como dieléctrico *high–*κ. Para el caso del método de Castagné-Vapaille el potencial de superficie ψ_S fue calculado mediante la integral de Berglund [50], mientras que para el método de la conductancia mediante el tiempo de respuesta de los defectos (estimado mediante la estadística de Shockley-Read-Hall para la tasa de captura y emisión [55]. Para ambos métodos de cálculo se puede observar que las muestras no-FGA son las más defectuosas en términos de D_{it}. Al igual que en el caso de las curvas MFCV, las gráficas de D_{it} vs. ψ_S para $HfGdO_x$ y $HfAlO_x$ no se incluyen dada su notoria similitud con el caso de HfO_2.

Es importante señalar que para ambos métodos de extracción de D_{it} considerados, no se observaron diferencias relevantes entre las estructuras con distintas capas de *high–*κ. Esto sugiere que en las muestras aquí consideradas, la capa interfacial (GeO_x/ GeO_2/Al_2O_3 con un espesor total de $\sim$3 nm [99]) apantalla el posible impacto de los defectos del *high–*κ en la calidad de la estructura MOS en términos del D_{it} [116], [117]. Por lo tanto, la discusión se centrará en los efectos del FGA sobre la característica de degradación de las muestras al ser sometidas a un estrés eléctrico, y no en la influencia del film de *high–*κ. Finalmente, debe mencionarse que a pesar de que tanto el método de Castagné-Vapaille (Alta Frecuencia-Baja Frecuencia) [56] como el de conductancia paralela [50] pierden precisión cuando se evalúa el D_{it} a temperatura ambiente cerca de las bandas de conducción y valencia, estos son ampliamente utilizados en la literatura para estudiar el D_{it} alrededor del *mid-gap* [99], [118]-[122]. En este estudio, ambas técnicas se utilizaron para hacer una comparación relativa del D_{it} en la región de *mid-gap* [55], [99], [118] antes y después del tratamiento de FGA y no con el objetivo de determinar fehacientemente el D_{it} exacto en cada caso. Por otro lado, los resultados obtenidos son consistentes con otros reportados recientemente [55], [101], [103], [118].

Para clarificar el rol del tratamiento de FGA posterior a la deposición del dieléctrico en la dinámica de degradación, se realizaron mediciones de estrés eléctrico sobre los dispositivos multicapa de Al_2O_3/$HfMO_x$, similares a las reportadas en la Fig. 3.3. En ellas se estudia el corrimiento relativo de V_{FB} (ΔV_{FB}) en función de la tensión de estrés, tanto

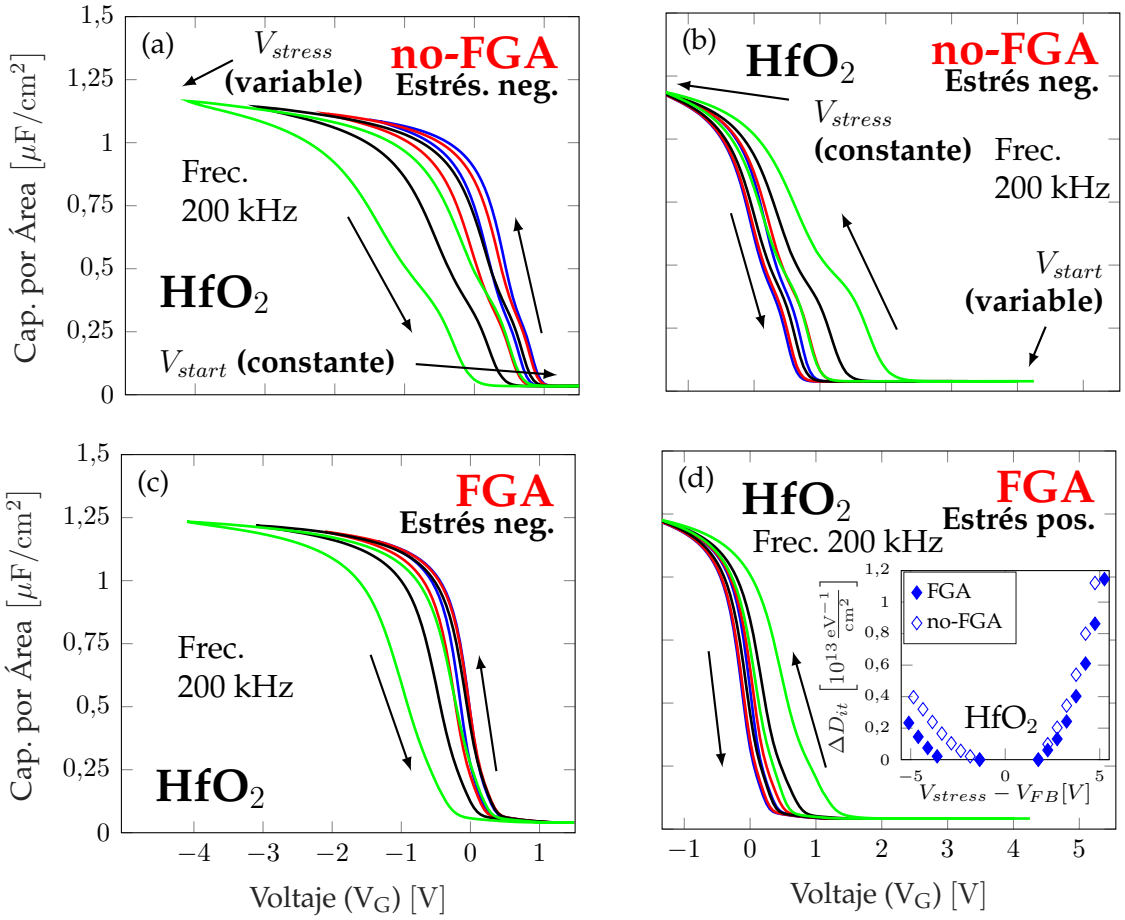


Figura 3.6: Mediciones de histéresis en curvas C–V para una única frecuencia de prueba (200 kHz). Las muestras de control (no-FGA) estresadas a tensión negativa y positiva se muestran en (a) y (b), respectivamente. De la misma forma en (c) y (d) se muestran las muestras sometidas al tratamiento de FGA estresadas a tensiones negativa y positiva, respectivamente. En el *inset* de (d) se muestra la evolución de ΔD_{it} calculado en el *mid-gap* en función de las tensiones de estrés V_{stress} y V_{start} respecto de la muestra sin estresar. Las tensiones de V_{stress} y V_{start} se indican en (a) y (b) para los experimentos de estrés a tensión negativa y positiva, respectivamente.

para tensiones negativas como positivas. Para las primeras, V_{FB} fue medido a partir de curvas C–V sucesivas obtenidas a 200 kHz, en las cuales la tensión mínima en acumulación (V_{stress} como se indica en la Fig. 3.6a) se reduce gradualmente. La misma es mantenida constante durante periodos de corta duración (∼1 seg.). Por el contrario la tensión máxima en inversión (V_{start}, como se indica en la Fig. 3.6a) se mantiene constante. Las curvas C–V resultantes para el caso de las muestras con HfO_2 se pueden observar en las Figs. 3.6a–3.6d, mientras que la metodología de medición es representada esquemáticamente en la Fig. 3.7a. A partir de dichas gráficas, se calculó ΔV_{FB} como $\Delta V_{FBN} = V_{FBN} - V_{FB0}$, donde V_{FBN} es la tensión de *flat-band* extraída de la enésima curva C–V medida en el barrido que va desde la región de inversión a acumulación y V_{FB0} es la tensión de *flat-band* extraída para el primer barrido. En este contexto, un valor positivo de ΔV_{FB} representa un corrimiento de V_{FB} hacia tensiones positivas mientras que un ΔV_{FB} negativo, lo contrario. Dado que el des-atrapamiento de la carga atrapada puede ser rápido en los materiales aquí considerados [104], [108], el *delay* entre dos mediciones C–V sucesivas se

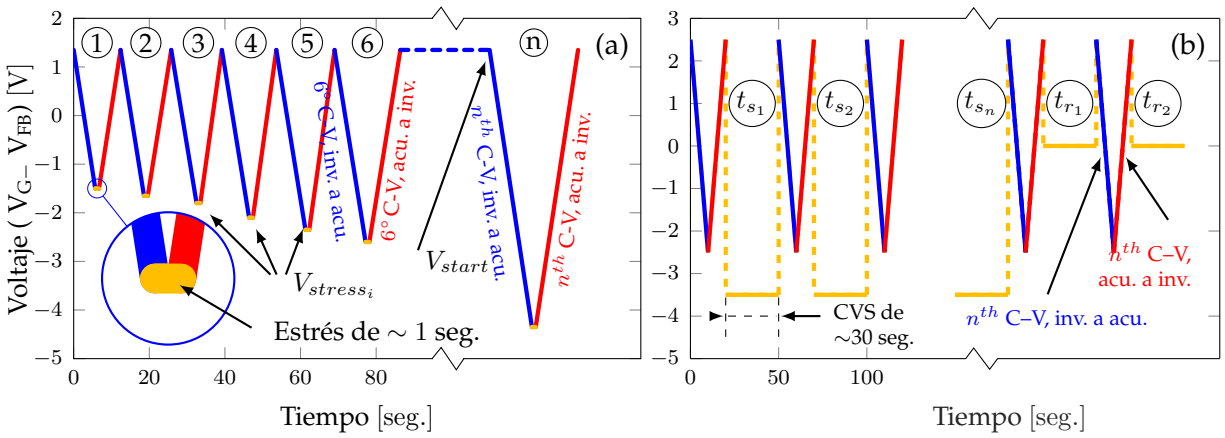


Figura 3.7: La evolución de la tensión de estrés para los experimentos ilustrados en la Fig. 3.6 se muestra en (a) para el caso de estrés a tensión negativa. Se indica el periodo de estrés y los barridos C-V de inversión a acumulación y *vice versa*, así como las tensiones de V_{stress} y V_{start}. Procedimiento de medición utilizado para el estrés de larga duración cuyos resultados se reportan en la Fig. 3.10. $V_G - V_{FB}$ se mantiene constante por un periodo de ∼30 seg. de forma de polarizar el dispositivo en acumulación, siendo periódicamente interrumpido para evaluar V_{FB} y V_{Hys}.

mantuvo por debajo de los 100 mseg. Esta metodología, ampliamente utilizada para estresar estructuras MOS [19], [106], [108], [109] permite aumentar la velocidad del barrido aprovechando el equipamiento disponible y por lo tanto reducir el porcentaje de carga que relaja durante el fin de un ciclo y el comienzo del siguiente [106], [108].

La figura 3.8a muestra el efecto de V_{stress} en el corrimiento de V_{FB} para las distintas combinaciones consideradas (HfO_2, $HfGdO_x$ y $HfAlO_x$, considerando FGA y no-FGA). V_{FB} se corre hacia tensiones más negativas (ΔV_{FB} <0), indicando atrapamiento de carga positiva (huecos). Las muestras sujetas a FGA muestran un ΔV_{FB} en función de V_{stress} mucho menor que aquellas que no fueron sometidas a dicho tratamiento (no-FGA). Siguiendo una metodología análoga a la discutida en el párrafo anterior, en la figura 3.8b se presenta la evolución de ΔV_{FB} en función de V_{start}: En cada barrido C–V se aumenta progresivamente V_{start} mientras V_{stress} se mantiene constante (véase las figuras 3.6b y 3.6d). Al contrario del estrés a tensiones negativas, y aparte del lógico atrapamiento de carga negativa (ΔV_{FB} >0), en este caso no se observan diferencias en el ΔV_{FB} medido para muestras tratadas con FGA y aquellas sin ser sometidas a dicho tratamiento.

Para cumplimentar los resultados presentados en las Figuras 3.8a y 3.8b, se analizó también la histéresis de la característica C–V (V_{hys}). La misma se define como la diferencia entre V_{FB} (medida en el barrido de inversión a acumulación) y una tensión $V_{FB'}$, definida como aquella para la cual la capacidad medida durante el barrido de acumulación a inversión, coincide con la medida en para la tensión de *flat-band*. Los resultados obtenidos se muestran en las figuras 3.8c y 3.8d. En la primera se puede apreciar que el estrés a tensiones negativas (mediante la disminución progresiva de V_{stress}) produce un aumento levemente mayor de ΔV_{hys} en las muestras no tratadas (no-FGA). Esta diferencia es similar a la observada para ΔV_{FB} (véase la figura 3.8a), pero menos pronunciada.

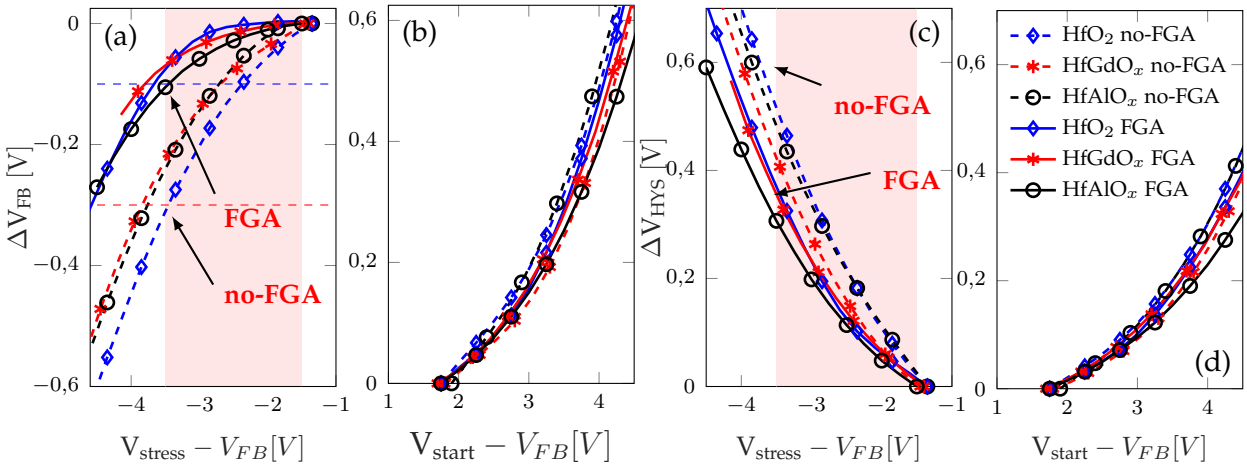


Figura 3.8: ΔV_{FB} y ΔV_{hys} para diferentes muestras multi-laminadas para tensiones de estrés negativas y positivas (se reduce V_{stress} y se aumenta V_{start}, respectivamente). El corrimiento de V_{FB} (ΔV_{FB}) para muestras con y sin tratamiento de FGA estresadas a tensión (a) negativa y (b) positiva. La variación de la histéresis de C–V (ΔV_{hys}) para muestras con y sin tratamiento de FGA estresadas a tensión (c) negativa y (d) positiva. Las lineas de trazos roja y azul indican en (a) la máxima variación de V_{FB} en el rango de -3.5 a -1.5 V (zona sombreada) para las muestras no-FGA y FGA, respectivamente.

Por el contrario, para el estrés a tensión positiva (Figura 3.8d) no se observan diferencias en las variaciones de ΔV_{hys} entre las muestras tratadas y no tratadas (FGA y no-FGA). Por lo tanto del análisis de las métricas presentadas en las Figuras 3.8a-3.8d se concluye que el principal efecto del tratamiento por FGA se evidencia a tensiones negativas de estrés. Finalmente debe mencionarse que el bajo nivel de corriente (por debajo de 1 nA) para el rango de tensiones aplicadas (de -4 a 0 V, aproximadamente, véanse las curvas I–V en las Figuras 3.9a y 3.9b, para las muestras no-FGA y FGA, respectivamente), permite descartar que las variaciones observadas sean producidas por el mecanismo de *hot-carrier-injection*.

Para analizar la influencia del tiempo de estrés y respaldar los resultados obtenidos bajo estrés a tensiones negativas en la Figura 3.8a, se estudió la evolución de la tensión

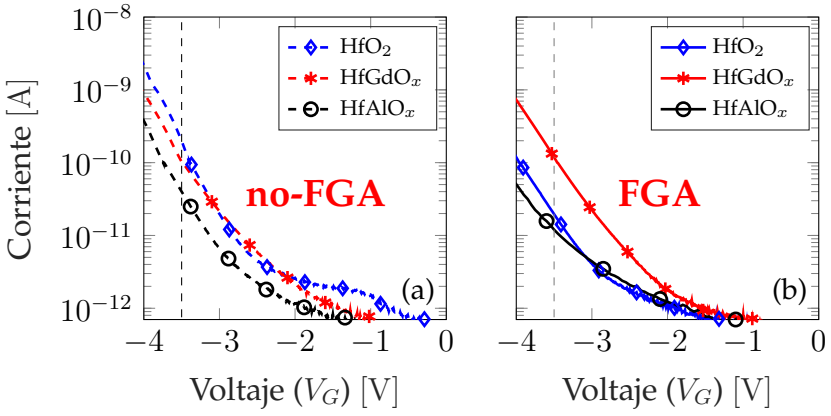


Figura 3.9: Mediciones I–V para las distintas muestras consideradas: (a) muestras no-FGA (control), (b) muestras FGA. Se observa una nivel de corriente apenas mayor para el caso de las muestras no-FGA

de *flat-band* en función del tiempo de estrés a tensión constante (CVS) y a temperatura ambiente. Para ello se aplicó una tensión V_G tal que $V_G - V_{FB_0}$ sea constante para todas las estructuras consideradas. Para obtener las variaciones de V_{FB} y V_{hys}, a intervalos regulares de ∼30 seg. se quitó el estrés para realizar una medición C–V sobre la muestra [19]. Luego de un periodo de estrés total de unos ∼ 750 segundos, la muestra se dejó en reposo (sin tensión aplicada) y el proceso de des-atrapamiento fue medido utilizando curvas C–V a periodos regulares de ∼30 seg. El proceso completo se presenta esquemáticamente en la Fig. 3.7b. Los resultados de este procedimiento de medición se presentan en las Figuras 3.10a y 3.10b. En primera instancia, se puede ver que el corrimiento hacia tensiones negativas de V_{FB} es coherente con lo expuesto en la Fig. 3.8. En ambos casos los corrimientos son más pronunciados en el caso de las muestras sin tratamiento de FGA, siendo ΔV_{FB} la que mayores variaciones presenta. La medición a tensión constante permite también evaluar las características temporales de los procesos de atrapamiento y des-atrapamiento de carga. Al ajustar las curvas de ΔV_{FB} (Fig.3.10a) mediante una función de ley exponencial $\Delta V_{FB} = Ae^{-t/\tau}$ tanto para el inicio del estrés como el inicio de la relajación, se obtuvieron constantes de tiempo de 20-100 seg. para el primero, y de 200-600 seg. para el segundo.

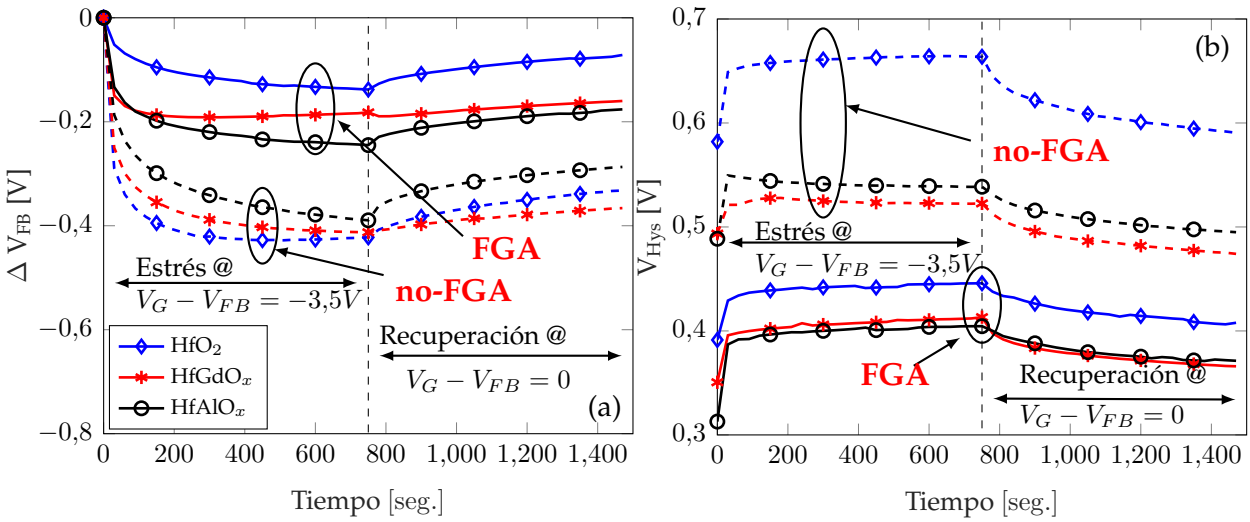


Figura 3.10: Impacto del tratamiento de FGA en el corrimiento de (a) V_{FB} y (b) la histéresis de C–V (V_{hys}) para las diferentes estructuras MOS de Ge consideradas, bajo condiciones de estrés prolongadas (30 seg.) Se puede ver como el tratamiento de FGA reduce tanto ΔV_{FB} como V_{hys} en todos los casos.

De esta forma no solo el proceso de relajación (cuya magnitud está en linea con trabajos previamente reportados [19]) es incompleto (ΔV_{FB} no retorna a 0, ni V_{hys} a su valor inicial) sino que también es más lento que la dinámica de degradación (medida durante el estrés). Es importante resaltar que el nivel de corriente resultante de la tensión aplicada durante el estrés eléctrico utilizado para obtener los resultados presentados en la Fig. 3.10 ($V_G - V_{FB}$= -3.5V) está en el orden de los 100pA para ambos *sets* de muestras (véanse las características I–V expuestas en la Figura 3.9). Se evita así la creación de carga inducida por una alta corriente de fuga a través del dieléctrico. Por otro lado, las diferen-

cias existentes entre las dinámicas de degradación para los distintos dieléctricos *high–*κ están en linea con aquellas reportadas en al literatura: La histéresis en la curva C–V para los dispositivos de HfO_2 es la mayor entre los otros óxidos ternarios considerados. Asimismo, aunque la mejora producida por el tratamiento de FGA es observable para todos los materiales, esta es mucho más pronunciada en el caso de HfO_2 [123]. El análisis del origen de dichas discrepancias escapa a los alcances de esta tesis.

3.3.2. Influencia del proceso de fabricación

De la comparación de la dinámica de ΔV_{FB} y ΔV_{hys} causada por el estrés eléctrico se pueden extraer conclusiones muy importantes con relación al rol que juega el tratamiento de FGA en las características de degradación. A tensiones negativas (esto es, variando V_{stress}-V_{FB}), ΔV_{FB} y ΔV_{Hys} (véanse las figuras 3.8a y 3.8c) muestran comportamientos disimiles entre los *sets* de muestras FGA y no-FGA. En las primeras (FGA), ΔV_{Hys} muestra un crecimiento monótono hasta $\sim$+0,4 V en el rango de tensiones de -3.5 a -1.5 V, mientras que ΔV_{FB} no varía en más de $\sim$0.1 V para el mismo rango de tensiones de estrés. Esto es coherente con el hecho de que el las curvas C–V de inversión a acumulación (de tensión positiva a negativa) se superponen, e indica que la mayoría de la carga atrapada se libera durante el barrido de acumulación a inversión (tensión negativa a positiva). Por el contrario, las muestras sin tratamiento de FGA muestran una reducción sostenida de ΔV_{FB} junto con un aumento de ΔV_{hys}, indicando que la carga atrapada por el estrés a tensión negativa no se libera durante el barrido de acumulación a inversión.

La característica permanente del atrapamiento de carga para las muestras sin tratamiento de FGA puede relacionarse con una mayor barrera de energía para remover la carga positiva atrapada en los defectos [113]. Por otro lado, el mayor atrapamiento de carga en las muestras no-FGA con respecto a las FGA, puede relacionarse con una mayor densidad de defectos de atrapamiento carga positiva, con niveles de energía alineados con la banda de valencia del semiconductor (i.e., tensión de compuerta negativa). En una investigación reciente, Zhang *et al.* [101] mostraron, mediante *X-rays Synchrotron Photoelectron Spectroscopy*, evidencia experimental de los estrados químicos del Ge en la región de interfaz en estructuras Ge/ALD-Al_2O_3, durante el tratamiento de FGA. En este estudio, se mostró que el óxido de Germanio (GeO_X, Germanio con estados de oxidación menores que +4) formado en la interfaz *high–*κ/Ge puede degradar significativamente la calidad de la estructura, al introducir defectos en niveles de energía cerca de las Bandas de Conducción (trampas de electrones) y de Valencia (trampas de huecos). Asimismo, también se reporta que las mismas pueden ser pasivadas mediante el tratamiento de FGA.

En este contexto, la evidencia experimental reportada en [101] permite una interpretación en la cual, durante el tratamiento de FGA, átomos de hidrógeno (presente en la atmósfera del FG) se pueden difundir hacia la interfaz Al_2O_3/Ge y reaccionar con los

hidroxilos residuales incorporados durante el crecimiento de la capa de Al_2O_3 por ALD, produciendo de esta forma moléculas de H_2O. Estás posteriormente reacciona con el Ge y GeO_X para formar GeO_2 (una variante del óxido nativo menos defectuoso), lo cual reduce el número de defectos en la interfaz. Vale la pena mencionar que las reacciones propuestas son consistentes con experimentos previos realizados en estructuras *high–*κ/Ge [84], [103] y que la existencia de GeO_X en sustratos tratados con HF durante la fabricación ha sido ampliamente documentada en la literatura [111], [124], [125]. Por otro lado, el Pt en el electrodo de compuerta ayuda a disociar el H_2 en H atómico, lo que puede contribuir a la reacción con el grupo -OH en la capa de Al_2O_3 y por lo tanto a la formación de GeO_2.

Adicionalmente, las diferencias en los enlaces existentes en la interfaz Ge/GeO_X y Ge/GeO_2 han sido analizadas mediante modelado de primeros principios por Zhang *et al.* [101]. Los resultados muestran que la interfaz Ge/GeO_2 está libre de defectos en el rango de energías del *gap* de Ge, mientras que la interfaz $Ge/GeO_X/GeO_2$ puede tener una gran densidad de defectos (vacancias de oxígeno) con niveles de energía cerca de la Banda de conducción (trampas de electrones) o de valencia (trampas de huecos) [114], [115]. En este escenario, y considerando la interpretación previa, el hidrógeno reacciona con las trampas de electrones y huecos, removiendo aquellos en el rango de energías del *gap* [101].

Al contrario de las tendencias observadas para el caso de estrés a tensión negativa, para el estrés a tensión positiva existe un crecimiento monótono tanto de ΔV_{FB} como de ΔV_{hys} en función de V_{start}, tal como se puede observar en las Figuras 3.8b y 3.8d respectivamente. Un aspecto clave a discutir es la gran similitud entre las tendencias exhibidas por las muestras tratadas con FGA y las no tratadas (no-FGA). El corrimiento hacia tensiones positivas de V_{FB} (atrapamiento de electrones) es indicativo de que el daño causado por el estrés positivo es permanente, y la característica C–V original no puede ser recuperada. En otras palabras, los defectos responsables por el atrapamiento de electrones durante el barrido de inversión a acumulación son creados durante propio estrés a tensión positiva. Esta interpretación se sustenta por el incremento en el ΔD_{it} (véase el *inset* de la figura 3.6d) computado para cada ciclo de medición C–V alrededor del *mid-gap* (esto es, el incremento en el D_{it} de las muestras sometidas a estrés con respecto al D_{it} para las muestras no estresadas). Un detalle no menor del análisis de D_{it} en función de las tensiones de estrés (V_{stress} y V_{start}) es que el aumento de ΔD_{it} es significativamente mayor para el estrés a tensión positiva (V_{start}) lo cual siguiere que cuando la carga es inyectada desde el sustrato, el daño en la interfaz resulta más importante.

Como se mencionó previamente, el hidrógeno (H) presente en la atmósfera durante el FGA reacciona con los defectos de la interfaz $Ge/GeO_X/GeO_2$, pasivando tanto trampas de atrapamiento de electrones como de huecos con energías en el rango del *gap* del material. Sin embargo, la barrera de energía necesaria para remover cada uno de estos dos tipos de trampas es diferente, siendo más grande la energía necesaria para pasivar los defectos que permiten el atrapamiento de electrones [101], [126]. Esto ayuda a expli-

car las diferencias observadas entre las muestras FGA y no-FGA cuando se estresan con tensiones positivas.

En resumen, en base al corrimiento de V_{FB} se ha observado que el estrés a tensiones negativa y positiva contribuye con el atrapamiento de huecos y electrones respectivamente en estructuras MOS fabricadas sobre un sustrato de Ge. El tratamiento de FGA tiene mayor impacto en el estrés a tensiones negativas, dado que la energía necesaria para pasivar los centros de atrapamiento de huecos es sustancialmente menor que para los centros de atrapamiento de electrones. Dado que las mismas tendencias se reportaron para diferentes materiales dieléctricos ($HfAlO_X$, $HfAGdO_X$, HfO_2) depositados sobre un sustrato de Ge con una bi-capa interfacial de GeO_2/Al_2O_3, se concluye que las características de dicha bi-capa han de jugar un rol preponderante en la dinámica de degradación de la estructura completa, y la susceptibilidad al tratamiento de FGA. Sin embargo, vale la pena mencionar que el espesor de la capa multilaminada de $GeO_X/GeO_2/Al_2O_3$ de $\sim$3 nm, podría estar apantallando la influencia de la capa de *high–*κ sobre los mecanismos de atrapamiento de carga [116], [117], los cuales usualmente tienen lugar en la vecindad de la región de interfaz. Por lo tanto, este es un desafío que requerirá futuras investigaciones considerando capas interfaciales más delgadas.

3.4. Conclusiones

En este capítulo se discutieron aspectos del atrapamiento de carga en dieléctricos *high–*κ depositados sobre sustratos de alta movilidad, propuestos como reemplazo del silicio en tecnologías CMOS futuras. Dado el rol preponderante que tiene el atrapamiento de carga en las variaciones de la tensión de bandas planas de la estructuras MOS (y por consiguiente en la tensión de umbral en el transistor MOSFET), su análisis requiere particular atención. Mediante mediciones experimentales realizadas sobre estructuras MOS utilizando sustratos de Germanio y semiconductores III-V (InP e InGaAs), se cuantificó el atrapamiento de cargas positiva y negativa en ambos casos, estableciendo una relación del mismo con la densidad de trampas de borde y estados de interfaz presentes en la capa aislante y la interfaz semiconductor/óxido, respectivamente. Asimismo, se analizan alternativas para su minimización: En primer lugar, para el caso de los sustratos de Germanio, se investiga la posibilidad de utilizar un tratamiento de recocido térmico en una atmósfera de de H_2/N_2 (*Forming Gas Annealing*). En este escenario, el proceso de recocido produce una pasivación selectiva de los centros de atrapamiento de carga positiva, debido a una menor energía necesaria en comparación a las centros de atrapamiento de carga positiva. En segundo lugar, para el caso de sustratos III-V, se considera la utilización de una capa de interfaz, depositada entre el sustrato y el dieléctrico *high–*κ. Esto se debe a que la distribución energética de dichos defectos (responsables del atrapamiento) está fuertemente condicionada tanto por la capa aislante como por el sustrato. En última ins-

tancia, y teniendo en cuenta los resultados experimentales aquí reportados en conjunto con la literatura especializada, es importante señalar que el tratamiento de FGA posterior a la metalización del electrodo de compuerta afectan la dinámica de degradación en Ge y semiconductores III-V en formas diferentes: Mientras que el atrapamiento de carga se reduce en una estructura del tipo Ge/GeO_2/*high*–κ luego del tratamiento de FGA, el efecto opuesto se observa en estructuras III-V/*high*–κ. Por lo tanto, la correcta elección de las condiciones del FGA capaces de mejorar la interfaz semiconductor/*high*–κ sin afectar la fiabilidad de la estructura, son clave para permitir el futuro desarrollo de las tecnologías Ge/III–V híbridas.

4 Dinámica de ruptura en dieléctricos

El evento de ruptura implica la pérdida de las propiedades dieléctricas del aislante de compuerta. Como se discutiera previamente, esta capa es crucial para le correcto funcionamiento de los dispositivos MOS, parte fundamental de la nanoelectrónica actual. Por otro lado, la gran variabilidad de este fenómeno implica que su estudio debe ser abordado mediante herramientas estadísticas. En consecuencia, la estadística de ruptura dieléctrica ha sido un tema de estudio dado su enorme importancia tecnológica, y sus grandes implicaciones en las estimaciones de fiabilidad tanto para aislantes convencionales basados en silicio (SiO_2, SiON, SiO_X) y dieléctricos *high–*κ (HfO_2 y Al_2O_3, entre otros). En este capítulo se analiza la estadística de ruptura de dielectricos *high–*κ, y su relación con las propiedades geométricas e intrínsecas del medio, así como con la distribución y dinámica de generación de defectos en el mismo. Cabe mencionar que para ello se han combinado satisfactoriamente herramientas de simulación, experimentos de irradiación controlada con iones de alta energía y diversas técnicas de caracterización eléctrica.

4.1. Diferencias en la estadística de ruptura: SiO_2 y *high–*κ

Los aislantes de alta constante dieléctrica (aislantes *high–*κ) fueron introducidos en los procesos de fabricación CMOS para reducir las crecientes corrientes de fuga existentes en los nodos de fabricación de 45nm y otros más avanzados. Sin embargo, a pesar de sus beneficios, los aislantes *high–*κ han agravado ciertos problemas de fiabilidad tales como la ruptura dieléctrica dependiente del tiempo (*Time Dependent Dielectric Breakdown*, TDDB) [127], [31]. Tal como se ha expuesto brevemente en la sección 2.1.3.1, en los materiales dieléctricos, la ruptura dieléctrica se produce cuando se alcanza una densidad de defectos (de un tamaño en el orden del nm [128]) crítica en el volumen del material [24], [71] y [129] lo que da lugar a la formación de un filamento conductivo entre el ánodo y el cátodo. Tal interpretación ha sido clave para entender la fiabilidad de los óxidos de

compuerta y es independiente de la física de generación de defectos [31] y [24]. Desde un punto de vista estadístico, la ruptura del óxido de compuerta sigue una distribución de *Weibull*. En este contexto, la pendiente de *Weibull* (β) se usa para escalar la distribución del tiempo de vida a distintas áreas de óxido y bajos percentiles [127], [130], por lo que es un parámetro de suma importancia (véase la Fig. 2.8).

Para un film de SiO_2 crecido térmicamente y de 7nm de espesor, se obtiene un $\beta \sim$6 para el caso de la ruptura intrínseca [127]. No obstante, y en línea con la teoría percolativa [128], este valor decrece a medida que se reduce el espesor de la capa dieléctrica, lo cual compromete sensiblemente la fiabilidad en óxidos ultra-delgados [130]. Por el contrario, tal comportamiento no se observa en todos los dieléctricos *high–*κ. Por ejemplo, la pendiente de Weibull extraída de distribuciones de TDDB en dieléctricos *high–*κ ultra-delgados tales como HfO_2 y Al_2O_3 es mucho menor ($\beta \sim$2) para el mismo espesor del dieléctrico y no escala con el espesor del material [128], [88]. Esto es de una enorme importancia, dado que bajos valores de β no solamente implican una alta dispersión en los tiempos de ruptura, sino que la leve dependencia de β con el espesor del óxido, contradice la teoría percolativa sobre la cual se basa la gran mayoría de los modelos. Por lo tanto, un mejor entendimiento de la dinámica espacio-temporal de la generación de defectos responsables de la ruptura en dieléctricos *high–*κ es necesaria.

Se ha demostrado que aparte de una mayor densidad de defectos [127] en dieléctricos *high–*κ policristalinos, existen concentraciones de defectos altamente localizados en la proximidad de los bordes de grano, producidos por el proceso de fabricación [131]-[133]. En este escenario, los defectos no se distribuyen uniformemente en el volumen del óxido como en SiO_2 amorfo, sino que forman "*clusters*" de defectos alrededor de los bordes de grano. Por otro lado, mediante simulaciones DFT (*Density Functional Theory*) se sabe que a diferencia de lo que sucede en SiO_2, la generación de defectos en HfO_2 no sigue un proceso de Poisson (aleatorio) sino que se ve favorecida en las inmediaciones de los defectos ya existentes [134], lo que supone un fenómeno de generación correlacionado [135]. Tal comportamiento podría ser descrito estadísticamente mediante el denominado "modelo de *clustering*" introducido por Wu y colaboradores [136], que si bien puede ajustar los datos estadísticos obtenidos experimentalmente de TDDB para materiales *high–*κ, no es suficiente para esclarecer los pormenores de la dinámica de generación de defectos, y por lo tanto tampoco la muy leve dependencia entre β y t_{ox} observada en dieléctricos *high–*κ.

En este contexto, un avance significativo ha sido logrado mediante simulaciones físicas del proceso de TDDB publicadas por Padovani *et al.* [137] considerando tanto una alta densidad de defectos localizados y una generación espacialmente correlacionada de nuevos defectos. Sin embargo, y a pesar de la extensa investigación llevada a cabo en este campo, no existe suficiente evidencia experimental para sustanciar los resultados computacionales. Esto se debe a que la fabricación de dispositivos con una densidad controlable de defectos localizados sobre las cuales extraer la estadística de TDDB es sumamente problemática.

Ante este desafío, en este capítulo se propone la utilización de irradiación localizada con iones de alta energía como herramienta para controlar la densidad de defectos localizados. Sobre esta base, se analiza experimentalmente (y también por medio de simulaciones físicas multi-escala que contemplen la generación de defectos en dieléctricos [138]) el impacto de la densidad inicial de defectos sobre la pendiente de Weibull, mediante mediciones de TDDB en capacitores MOS con dieléctrico de *high–*κ (HfO_2).

4.2. Daño inducido por radiación en estructuras MOS

Una herramienta sumamente interesante para introducir *clusters* de defectos en óxidos prístinos surge de considerar la interacción de partículas de alta energía con la estructura del material. El daño inducido por radiación (*Radiation Induced Damage*) en dispositivos CMOS ha sido ampliamente estudiado en la literatura [139] e incluye fenómenos tales como el aumento de la corriente de fuga por efectos de radiación (*Radiation Induced Leakage Current*, RILC) y la ruptura dieléctrica inducida por radiación (*Radiation Induced TDDB*, RITDDB) [140]. A su vez, el daño inducido por iones pesados ha sido extensamente abordado para una gran variedad de materiales amorfos y cristalinos, incluyendo aislantes, semiconductores y metales [141]-[143]. En dieléctricos, la interacción entre dichas partículas de alta energía y la estructura cristalina del óxido resulta en una alta disipación de energía a lo largo de la trayectoria del ion, dando lugar incluso a la fundición localizada del material [144]. Otros fenómenos reportados causados por la interacción ion-materia incluyen la difusión de oxígeno en el corto rango y la re-cristalización localizada del Silicio en las cercanías a la trayectorias del haz [145], [146].

Además, los iones incidentes causan daño por desplazamiento que se correlaciona con la la creación de defectos de vacancias intersticiales en la vecindad de la trayectoria del ion, cuya vida media es dependiente de la energía de la partícula [147]. Asimismo vale la pena notar que la creación de dichos defectos no es continua a lo largo de la trayectoria seguida por el ion, y la magnitud de las modificaciones estructurales causada aumenta junto con la energía y masa de la partícula incidente. Mientras que para bajas energías, las partículas incidentes producen daño disperso y discontinuo, a altas energía el daño puede ser continuo. En el primer caso, la naturaleza dispersa de los defectos intersticiales a lo largo de la trayectoria del ion [148] evita la formación de un camino conductivo a lo largo del óxido que implique la ruptura dieléctrica del aislante.

Estos conceptos son entonces utilizados para controlar la densidad de defectos localizados en capacitores MOS (MOSCAP) de HfO_2[1]. Los mismos fueron fabricados sobre un sustrato de Silicio tipo N altamente dopado de 15 μm de espesor, crecido epita-

[1]Dispositivos provistos por el Prof. Joel Molina del Instituto Nacional de Astrofísica, Óptica y Electrónica de México (INAOE) a través de una colaboración internacional involucrando al grupo del Prof. Kin Leong Pey de la Universidad de Tecnología y Diseño de Singapur (SUTD)

xialmente sobre una oblea de silicio de 600 μm. La capa aislante fue depositada capa por ALD hasta lograr un espesor de $\sim$7 nm, el cual fue corroborado por elipsometria. Como electrodo superior (*Top Electrode*, TE) se utilizó una capa de Al de entre 400 y 500 nm. El área de los dispositivos resultantes es de $\sim$3600 μm^2. Para crear diferentes densidades de defectos, los dispositivos fueron irradiados con diferentes fluencias de una misma especie iónica en el micro-haz de iones pesados del Laboratorio TANDAR, perteneciente al Centro Atómico Constituyentes de la Comisión Nacional de Energía Atómica (CNEA-CAC). Dicha facilidad comprende un acelerador en tándem, modelo *National Electrostatic Corporation 20UD* con una fuente de iones SNICS acoplada a un micro haz *Oxford Microbeams, Ltd.* OM55 y un quadrupolo magnético tríple, capaz de enfocar los iones con energías de hasta $\sim$160 MeV amu/q2, en un haz de no más de 5 μm de diámetro. Adicionalmente, el agregado de una cámara de irradiación con 3 grados de libertad (X, Y y Z) permite el preciso posicionamiento de la muestra y el monitoreo en tiempo-real de la fluencia de iones. En la mitad superior de la Figura 4.1 se puede ver una fotografía del *set-up* experimental. De esta forma, fue posible en este estudio definir fehacientemente un área de irradiación que contenga solamente la muestra a ser irradiada (60 μm $\times$ 60 μm e incluso menor) y ajustar precisamente la fluencia de radiación para cada muestra. La capacidad de este *set-up* para enfocar con precisión el haz de iones ha sido ampliamente probada en [149], donde el haz es enfocado con precisión en regiones específicas de un circuito analógico *Full-Custom* fabricado en un proceso de 180 nm de longitud nominal de canal (véase la sub-figura inferior de la Fig. 4.1), a fin de evaluar la sensibilidad del circuito a la eventos espurios. Para mayores detalles en relación a este *set-up* el lector es referido a publicaciones previas del grupo de trabajo [150]-[152].

Tanto la especie iónica como la energía y fluencia necesarias han sido cuidadosamente elegidas en base a simulaciones SRIM. SRIM (*Stoping Range of Ions in Matter*) es un programa de simulación computacional ampliamente aceptado para simular la interacción ion-materia [153], [154]. En base a los resultados obtenidos por este medio y reportados en la figura 4.2, se optó por iones de carbono (C) dado que estos son relativamente livianos y con una energía de 40 MeV (C^{+4} 40 MeV) causan menos de un (1) desplazamiento atómico por Armstrong. Esto minimiza las posibilidades de crear un filamento conductivo completo (*Soft Breakdown*) durante la irradiación, produciendo por el contrario caminos filamentarios parcialmente formados [139], [153]. Una vez conocido el número de vacancias creadas por ion por Armstrong (V_i [vacancies/(ion $\times$ Å)]) mediante simulaciones SRIM, la densidad de defectos (D_D [cm^{-3}]) creada para una determinada fluencia (D [ions/cm^{-2}]) se puede estimar como $D_D = V_i \times D$ [139].

Se han considerado densidades de defectos en el rango de $\sim 10^{16}$ cm^{-3} a $\sim 10^{18}$ cm^{-3}, dado que se ha demostrado en el estudio realizado por Padovani *et al.* [137] mediante simulaciones multi-escala que tales densidades son capaces de causar variaciones claras en la pendiente de Weibull. Las fluencias requeridas para inducir tales densidades de defectos fueron calculadas mediante simulaciones SRIM [153], [154] y van de 10^{11} a

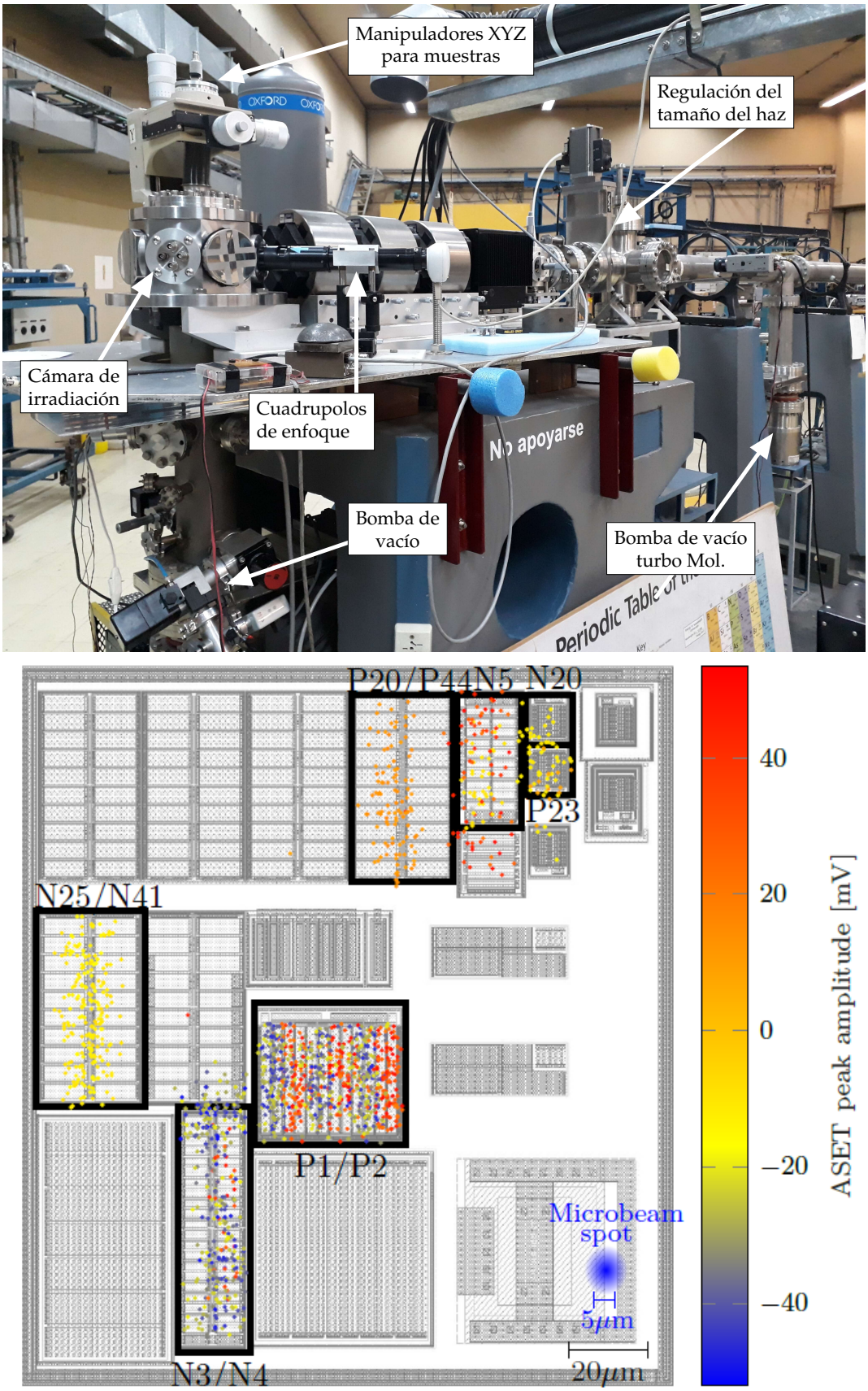


Figura 4.1: (Arriba) Sección final del micro-haz de iones pesados del acelerador lineal TANDAR de la CNEA-CAC. (Abajo). Comparación entre el tamaño del *spot* del micro-haz y un circuito CMOS analógico Full-Custom fabricado en un proceso de longitud nominal de 180 nm. Cada punto indica una posición en la que el haz fue enfocado. Reproducido de [149]

Tabla 4.1: Códigos asignados a las diferentes muestras considerando diferentes fluencias y áreas de irradiación

Fluencia	**Área o patrón de irradiación**	
	Área grande (60μm×60μm)	**Área pequeña (10μm×10μm)**
Sin irradiar	F#	
$10^{11}\ \frac{iones}{cm^2}$	L#1	——
$10^{12}\ \frac{iones}{cm^2}$	L#2	S#2
$10^{13}\ \frac{iones}{cm^2}$	L#3	S#3

10^{13} ion/cm^2. Vale la pena mencionar que fluencias mas bajas (y por lo tanto densidades de defectos menores) no logran agregar suficientes vacancias en la capa de óxido como para modificar sustancialmente la estadística de TDDB. Por otro lado, fluencias más altas producen la degradación de la movilidad de los portadores del canal y el atrapamiento de carga en defectos asociados al daño por radiación [139], [155]. Esto resulta en un incremento inaceptable de la resistencia serie (R_S), la cual se vuelve no despreciable y puede alterar los resultados de TDDB.

Como se ha mencionado, la interacción ion-materia en el volumen del dieléctrico conduce a la formación de caminos discontinuos de daño a lo largo de la trayectoria seguida por el ion, replicando a grandes rasgos los *clusters* de defectos en dieléctricos *high–κ*. Mientras la fluencia de radiación permanezca baja, la distribución de tales caminos de daño será heterogénea sobre el área de la muestra. No obstante, se podría discutir que al aumentar la fluencia, los defectos ya no formarán *clusters*, sino que dado su gran número formarán una distribución uniforme en el volumen del óxido. Por lo tanto, para asegurar la característica altamente localizada de la generación de defectos inducidos por la radiación, los experimentos de irradiación fueron repetidos sobre dos áreas diferen-

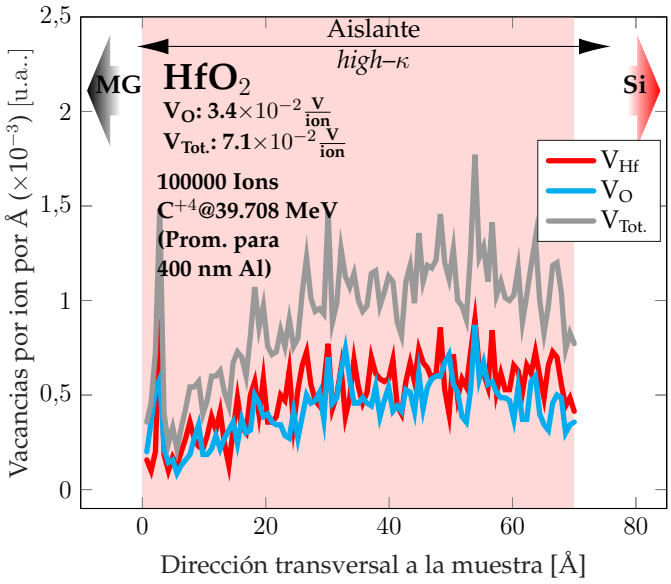


Figura 4.2: Simulación en SRIM. Se muestra el número de vacancias promedio creado por cada ion incidente en el dieléctrico por unidad de material atravesado (Å). Nótese que se genera aproximadamente el mismo número de vacancias de Hafnio que de Oxígeno.

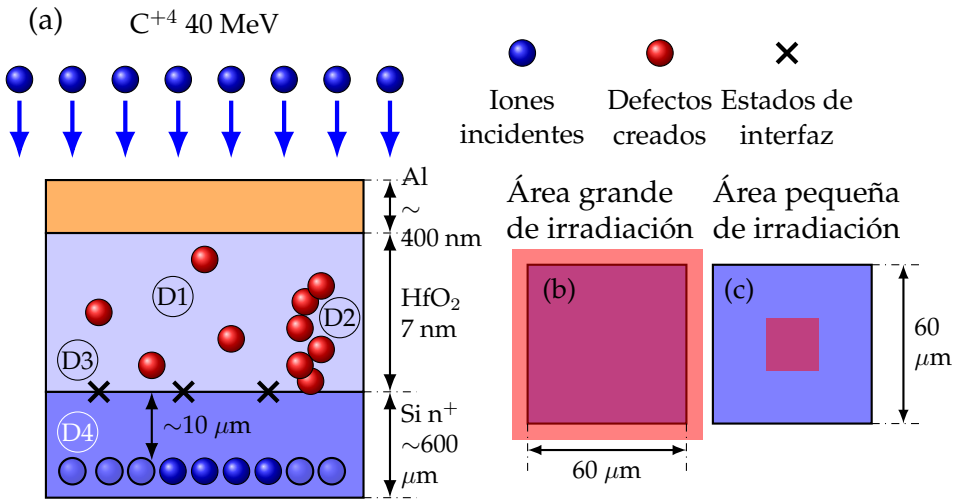


Figura 4.3: Diagrama del dispositivo bajo estudio, irradiado con iones de Carbono (C^{+4}), con una energía de 40 MeV. (1) Defectos generados en la capa dieléctrica, (2) caminos filamentarios parcialmente formados, generados en la capa dieléctrica, (3) estados de interfaz y (4) iones implantados en el sustrato.

tes, utilizando la misma fluencia y especie atómica. Un primer grupo de dispositivos fue irradiado sobre un área coincidente con el área del electrodo superior (60 μm $\times$ 60 μm) utilizando 3 fluencias diferentes y crecientes (*Sets* L#1, L#2, y L#3). El segundo grupo de dispositivos fue sometido a irradiaciones con las mismas fluencias, pero confinada a un área menor (10 μm $\times$ 10 μm) al área del electrodo superior (*Sets* S#2, y S#3). Para ello, el haz de iones fue enfocado solamente en una pequeña región de los capacitores MOS. Nótese que no se ha considerado el *set* S#1 (la misma fluencia que para L#1 pero con menor área) dado que el número de iones requeridos para este caso es demasiado bajo como para poder ser controlado con precisión. La nomenclatura y fluencias utilizadas para cada caso pueden verse en la Tabla 4.1.

Una representación tanto de la estructura MOS considerada en este análisis, como del daño causado por la radiación se muestra en la Figura 4.3a, independientemente del área de irradiación (grande -Fig. 4.3b- o pequeña -4.3c-). Iones de C^{+4} con una energía de 40 MeV impactan homogeneamente toda la superficie de la estructura MOS y con una dirección normal a la misma, tal como indican las flechas azules. Aparte de los defectos intrínsecos iniciales distribuidos en forma aleatoria (1), cada ion crea un *cluster* discontinuo de defectos a lo largo de la trayectoria trazada a través del dieléctrico (2). Asimismo, los iones podrían crear defectos en la interfaz óxido / silicio (estados de interfaz) (3) antes de finalmente detenerse algunos micrómetros debajo de la interfaz, luego de haber transferido toda su energía a la estructura cristalina del silicio (4).

Con la finalidad de comprobar el estado de degradación de las muestras utilizadas para este análisis antes y después del procedimiento de irradiación, se realizaron (en condiciones de oscuridad para evitar la generación de portadores por estimulo lumínico) mediciones C–V para múltiples frecuencias (*Multi-Frequency Capacitance Voltage*, MFCV) e I–V, utilizando un analizador de impedancias Agilent 4285A y un SMU Keithley 2636B, respectivamente, ambos conectados a una estación de prueba con un *chuck* con conexio-

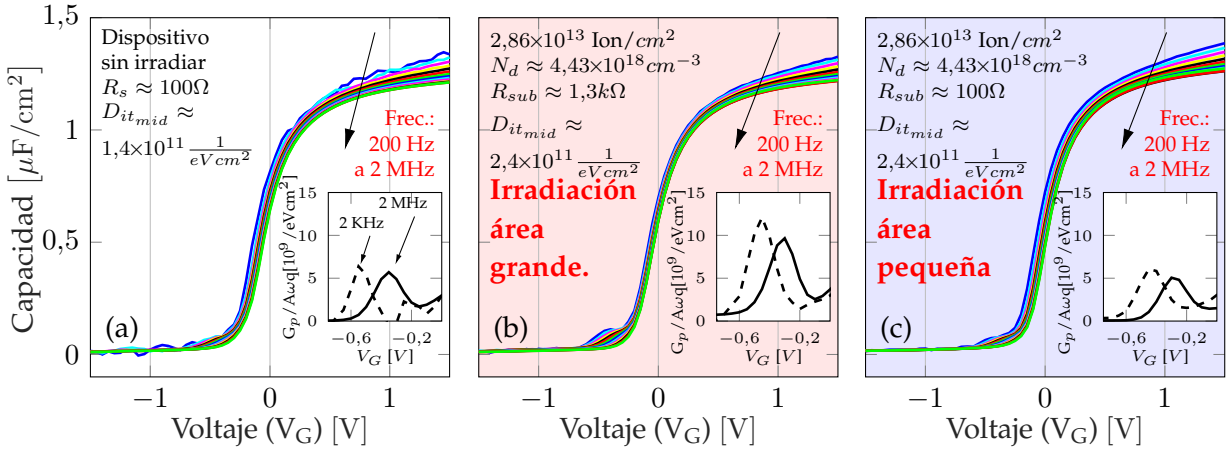


Figura 4.4: Mediciones C–V a múltiples frecuencias: (a) Muestra de control (no irradiada) (b) muestras sometida a la máxima fluencia de irradiación. La resistencia serie aumenta en función de la fluencia debido a los iones implantados en el sustrato (véase la Fig. 4.3). Las mediciones fueron corregidas para descartar el efecto de la misma. A pesar del pequeño incremento en el "*weak-inversion hump*", el D_{it} calculado mediante el método de la conductancia paralelo no muestra variaciones significativas en función de la fluencia. Los *inset* mostrados en (a)-(c) muestran el detalle de la conductancia paralela normalizada en cada caso. Nótese que el pequeño incremento de G_P está directamente relacionado al leve aumento de D_{it}.

nes tri-axiales[2]. Acto seguido, los dispositivos fueron utilizados para realizar experimentos de CVS con una tensión aplicada de 2.4V, y una limitación de corriente $I_{comp} \sim 1$ mA, utilizando un *set-up* de bajo ancho de banda implementado con un SMU Keithley 2636B. Esta configuración permite una resistencia serie despreciable y un piso de ruido de aproximadamente 100 fA, aunque una resolución temporal de algunos milisegundos.

4.2.1. Impacto en las curvas de C–V

Previamente a los experimentos de TDDB, los dispositivos irradiados fueron caracterizados mediante curvas I-V y MFCV (caracterización eléctrica) para corroborar que no hayan sufrido daño durante el proceso de irradiación, aparte del aumento requerido en la densidad de defectos. En la Figura 4.4 se comparan las curvas MFCV (corregidas en función de la resistencia serie, R_S, y para frecuencias de 200 Hz hasta 2 MHz) para los casos sin irradiar y con la máxima fluencia de radiación considerada (10^{13} ion/cm^2), tanto para el área de irradiación pequeña como la grande. En todos los casos, la capacidad máxima en acumulación por unidad de área ($\sim$1.2 μF/cm^2 en linea con el valor teórico esperado) y la tensión de *flat-band* (V_{FB} $\sim$-0.1 V, calculado con la técnica del punto de inflexión [53]) se mantienen constantes y la densidad de estados de interfaz (D_{it}, calculada alrededor del *mid-gap* y a temperatura ambiente, considerando el método de la conductancia paralelo [50]) varía muy poco desde $\sim$1.4x10^{11} eV^{-1}cm^{-2} hasta $\sim$2.4x10^{11}

[2]Equipamiento del Laboratorio de Nano-electrónica de la UTN-FRBA

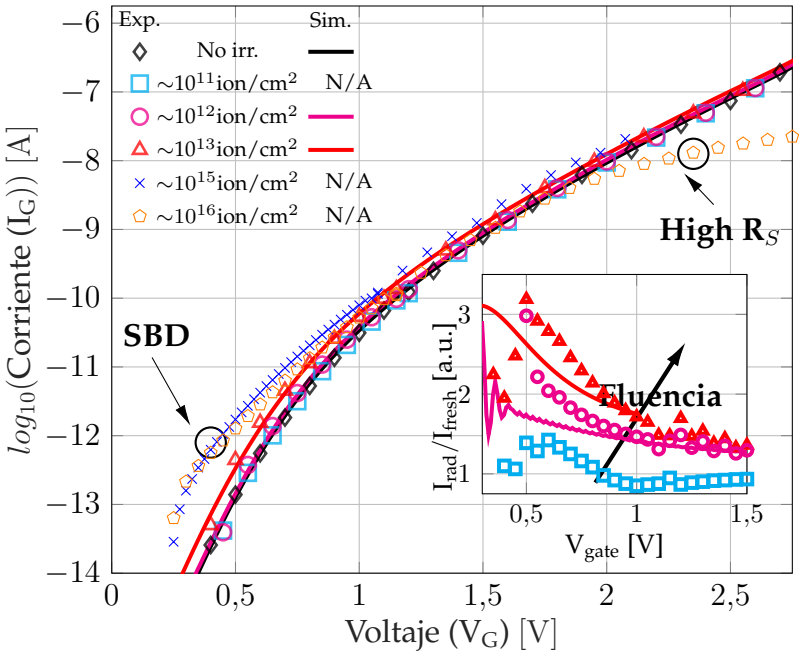


Figura 4.5: Mediciones I–V para diferentes fluencias de radiación. La corriente de fuga aumenta en función de la fluencia en el régimen de bajo campo eléctrico, y se mantiene aproximadamente constante para campos más elevados (Indicando que no existe SBD para las muestras bajo estudio).

$eV^{-1}cm^{-2}$. El pequeño incremento en el denominado "*weak inversion hump*" apreciable en las Figuras 4.4a, 4.4c y 4.4b constituye una clara evidencia experimental de esto [55], ya que es proporcional al incremento de la conductancia paralelo normalizada (véase el *inset* de las Figuras 4.4a a 4.4b), calculada como $Gp/(Aq\omega)$, donde G_p es la conductancia paralelo, A es el área del dispositivo, q es la carga elemental y ω es la frecuencia de medición medida en rad/sec. Tales características indican, respectivamente, que los defectos inducidos por la radiación deberían tener carga neutra dado que no se observan cambios significativos en V_{FB} en función de la fluencia y que los defectos son creados principalmente en el volumen del óxido, con poco o ningún impacto en la interfaz óxido/semiconductor (el D_{it} permanece casi constante).

4.2.2. Impacto en las curvas I–V

Con relación a la característica I-V de los dispositivos bajo estudio, la Figura 4.5 muestra tanto los resultados experimentales como así también las curvas simuladas. Las mediciones fueron realizadas sobre dispositivos irradiados con fluencias de 10^{12} y 10^{13} ion/cm^2, mientras que para las simulaciones se utilizó la densidad de defectos esperadas para dichas fluencias (las cuales son 10^{16} y 10^{17}, respectivamente). Los resultados muestran que no hay diferencias apreciables cuando se aplican campos eléctricos elevados, y solamente se observa un pequeño incremento en función de la fluencia cuando el campo eléctrico aplicado es bajo (véase el *inset* de la Figura 4.5), lo cual se ajusta a la literatura pre-

via sobre el tema [140]. Esto sugiere que los dispositivos no sufren *Soft-Breakdown* (SBD) durante el procedimiento de irradiación, lo cual es una posibilidad especialmente cuando se consideran altas fluencias de iones pesados [156], [157]. Las simulaciones fueron realizadas utilizando Ginestra^{TM} [137], [158], una plataforma de simulación multi-escala que tiene en cuenta los mecanismos físicos que rigen su funcionamiento, y teniendo en cuenta el pequeño incremento en la densidad de defectos causado por la irradiación, los cuales sustentan el mecanismo de tunel asistido por trampas (*Trap-Assisted-Tunneling*, TAT).

4.3. Generación de defectos: Efecto sobre TDDB

La estadística de TDDB se obtuvo para las muestras L#1, L#2 y L#3 (irradiación de área grande) y S#2 y S#3 (irradiación de área pequeña) utilizando experimentos de CVS. Los dispositivos fueron sometidos a estrés eléctrico hasta la ruptura abrupta ($I_{comp} \sim 1$ mA). La evolución de la corriente de compuerta durante el estrés eléctrico para los *sets* de muestras F# (sin irradiar), S#3 y L#3 se pueden observar en las figuras 4.6a-4.6c, respectivamente. Independientemente de la fluencia y el área de irradiación, las curvas de corriente en función del tiempo exhiben las características generales de la ruptura progresiva en óxidos delgados en estructuras MOS [18], [24], [31], como se ve en la Fig. 4.6d. Aquí, los puntos de inicio, SBD y HBD representan la corriente en el instante $t = 0$ (I_{init}), la corriente en el instante del primer evento de SBD (I_{SBD}) y en el instante previo al salto abrupto a la corriente límite (I_{HBD}), respectivamente. Para todos los casos considerados, se observa un pequeño incremento en la tasa de degradación promedio (*Degradation Rate*[18], [24], [31], estimada mediante dI/dt) en función de la fluencia (véase la Figura 4.6e), la cual puede ser despreciada dado que cae dentro del rango de dispersión. La corriente de fuga inicial para todas las fluencias de irradiación consideradas es ∼30 nA, como se muestra en las Figs.4.6f-i. Asimismo, hay una pequeña reducción con respecto a la corriente inicial, tal como se espera debido a la alta concentración de defectos en el óxido. Antes del primer evento de SBD, la corriente de fuga a través de la estructura MOS decrece debido al atrapamiento de cargas negativas, de acuerdo con la ley de Curie-von Schweidler [159]. Una vez que se produce la formación del primer camino de SBD, la corriente de fuga muestra un aumento progresivo estocástico, posiblemente debido a la competencia entre la progresiva degradación del camino ya formado, y la formación simultanea de caminos adicionales [160]. Esta dinámica se mantiene hasta que la corriente alcanza $\sim 5\mu$A y salta abruptamente al nivel máximo de corriente permitido (HBD). Los valores de la corriente de post-HBD se muestran también en las Figs. 4.6f-i.

El tiempo transcurrido hasta la ruptura abrupta (t_{HBD}) se grafica en escala de Weibull en las Figuras 4.7a y 4.7b para las áreas de irradiación grande y pequeña, junto con el caso de control (sin irradiación). Para cuantificar el impacto de la fluencia de impacto de la fluencia de la radiación, los datos de t_{HBD} fueron ajustados considerando dos

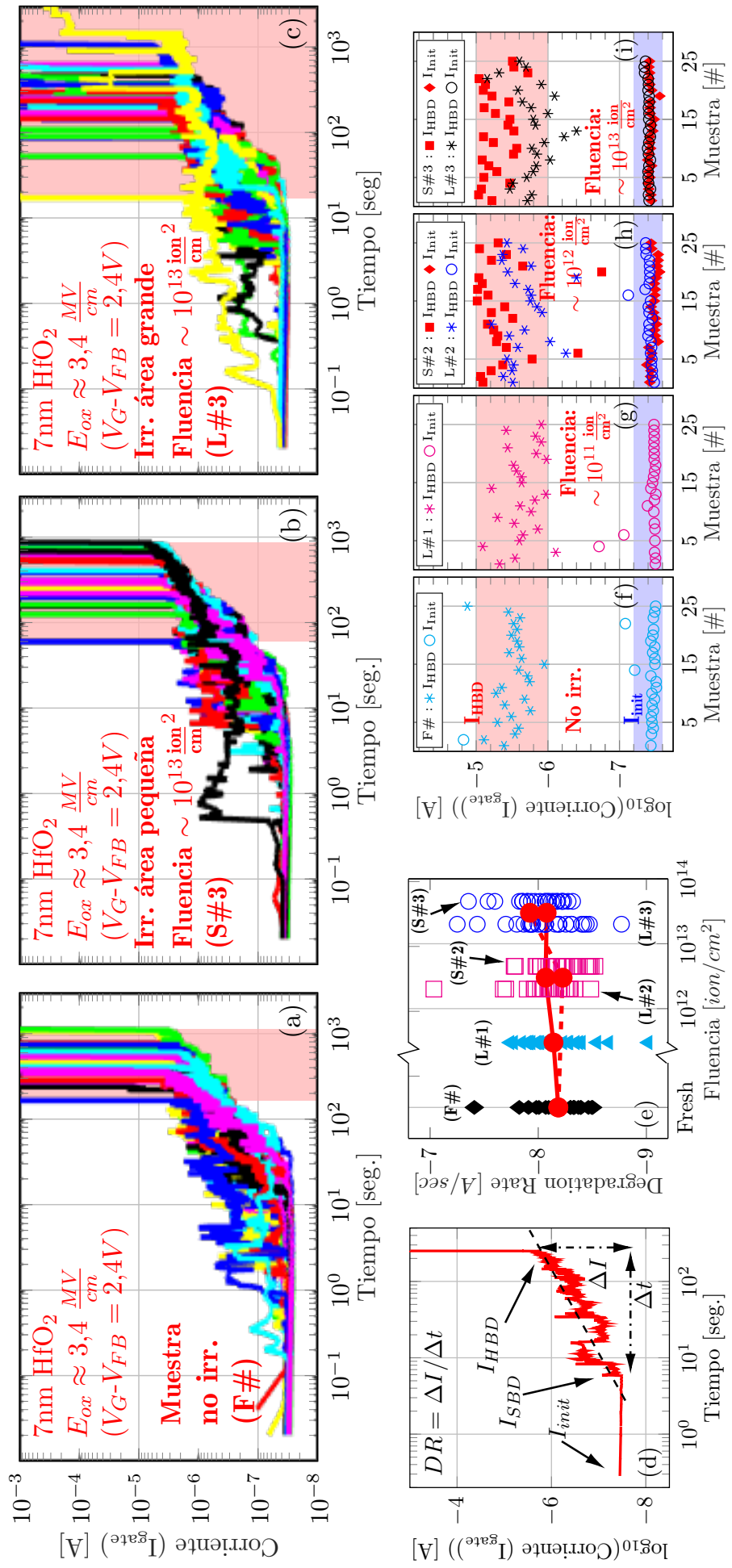


Figura 4.6: Mediciones CVS realizadas sobre las muestras (a) F#, (b) S#3 y (c) L#3. Independientemente de la fluencia de irradiación utilizada, todas las muestras exhiben un una dinámica de ruptura progresiva. (d) Representación de I_{Init}, I_{SBD} y I_{HBD} durante la evolución de la corriente de fuga. También se indica la tasa de degradación (DR) = $\Delta I/\Delta t$, donde $\Delta I = I_{HBD} - I_{SBD}$ y $\Delta t = t_{HBD} - t_{SBD}$. (e) Los valores calculados de DR se grafican para todas las fluencias de radiación, tanto para el área pequeña como para el área grande (Los datos de las muestras S#2, S#3, L#2 y L#3 están levemente desplazados horizontalmente por claridad). La corriente en el momento previo al HBD y la corriente inicial se grafican para cada combinación de área y fluencia. El valor de I_{Init} concuerda con las mediciones $I - V$ en todos los casos. I_{HBD} está en el rango de 1-10 μA. (f) muestra de control (sin irradiar) -F#-, (g) muestras L#1, (h) L#2 y S#2 e (i) muestras L#3 y S#3.

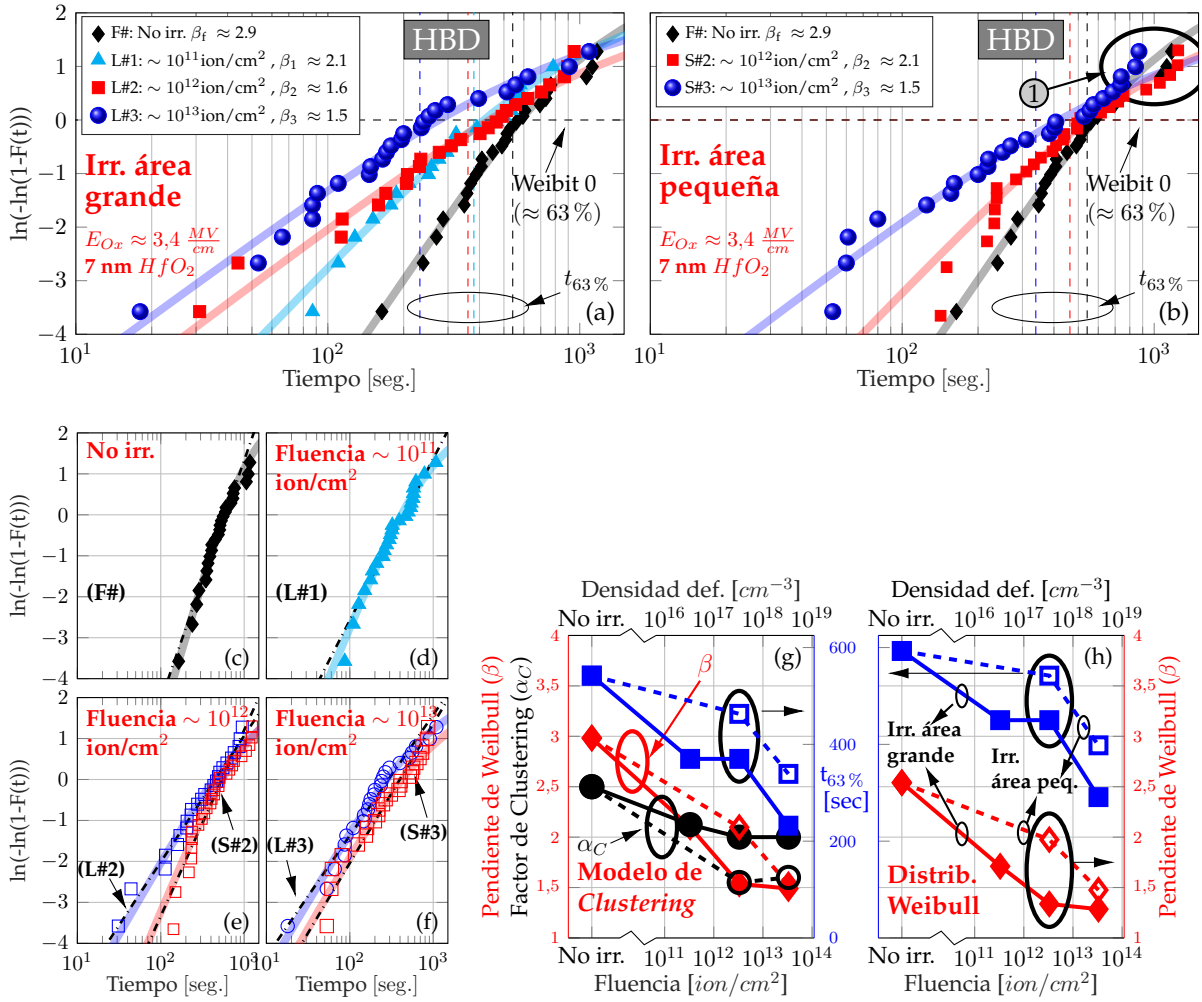


Figura 4.7: Distribuciones de TDDB medido para la ruptura abrupta (HBD) para cada fluencia de irradiación (25 dispositivos en cada caso). Los datos fueron obtenidos mediante experimentos de CVS a $V_G - V_{FB} = 2{,}4V$. (a) Irradiación de área grande y (b) irradiación de área pequeña. Los datos en (a) y (b) fueron ajustados utilizando el modelo de *Clustering*. La comparación entre el ajuste usando el modelo de Weibull y el modelo de *Clustering* se muestra en cada caso en las figuras (c) a (f), donde la concavidad "hacia abajo" se vuelve evidente. Los parámetros de ajuste, es decir el factor de *Clustering* (α_C círculos azules), pendiente de Weibull (β, diamantes rojos) y $t_{63\,\%}$ (tiempo de vida medio, cuadrados azules) se muestran en (g) para el caso del modelo de *Clustering* y (h) para la distribución de Weibull. Las lineas de trazo indican las irradiaciones de área pequeña y las continuas la irradiación de área grande.

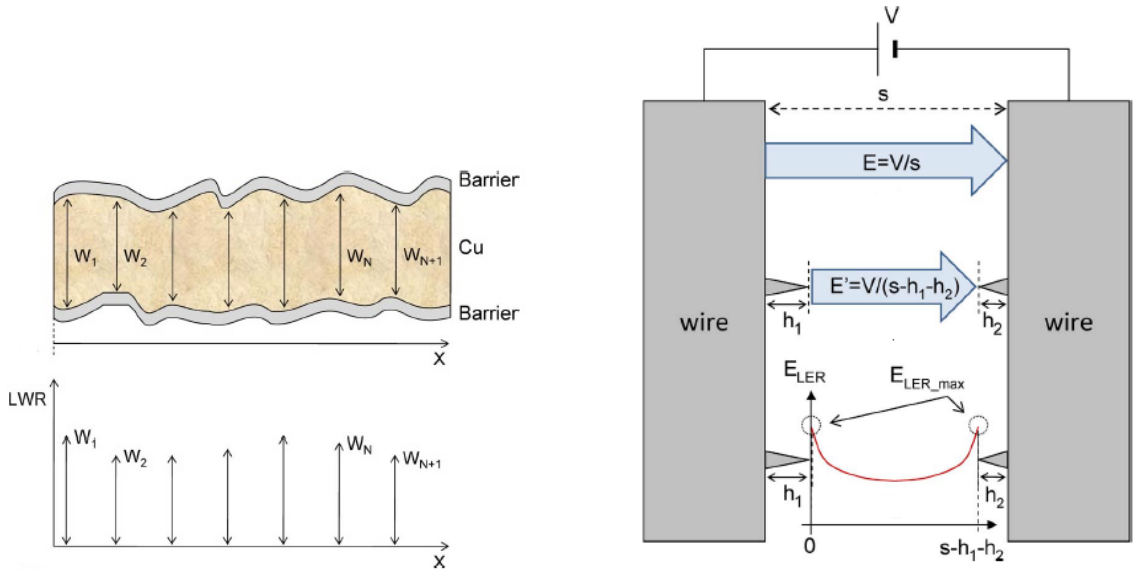
distribuciones de probabilidad comúnmente adoptadas para describir el tiempo de vida de los dieléctricos de compuerta. Estas son la distribución de Weibull (descrita en la subsección 2.1.3.1, reproduciéndose para comodidad del lector su función de probabilidad acumulada en la Ec. 4.1) y el modelo de *Clustering* [136], [161] (descrito por la Ec. 4.2). En ambas ecuaciones β refiere a la pendiente de Weibull y η es el factor de forma o tiempo de vida característico. En la ecuación 4.2 el parámetro α_C es el llamado factor de *Clustering*. Cuando α_C tiende a infinito, el modelo de *Clustering* se aproxima a la distribución de Weibull [136], [161]. Dada su novedad, en la siguiente Sub-Sección se provee una breve descripción del modelo de clustering.

$$F_{Weibull} = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{4.1}$$

$$F_{Clustering} = 1 - \left(1 + \frac{1}{\alpha_C}\left(\frac{t}{\eta}\right)^{\beta}\right)^{-\alpha_C} \tag{4.2}$$

4.3.1. Modelo de *Clustering*

Tal como se indicara en la Sub-sección 2.1.3.1 en forma introductoria, la conocida distribución de Weibull ha sido ampliamente utilizada para describir el tiempo de ruptura dieléctrica durante varias décadas, siendo muy adecuado para los dieléctricos de SiO_2. En estos (siempre que el espesor del óxido sea perfectamente uniforme) la generación de defectos tiene una probabilidad uniforme en todo el volumen de la capa aislante. No obstante, a medida que la tecnología microelectrónica sigue avanzando con el advenimiento de tecnologías como los FinFETs, la complejidad de las estructuras involucradas y los mayores desafíos impuestos a los procesos fotolitográficos, han resultado en variaciones localizadas del espesor de las diversas capas de óxido involucradas en el proceso de fabricación, por ejemplo el óxido que separa las lineas metálicas en el *Back-End Of Line* (BEOL). En los puntos donde la separación entre lineas es menor (y el óxido más delgado) el campo eléctrico entre lineas aumenta (véase la Fig. 4.8) y por ende en dicha región del óxido la tasa de generación de defectos también [162]. Estas son entones regiones de



(a) (b)

Figura 4.8: (a) Variaciones típicas del espesor en lineas metálicas de interconexión en el BEOL (*Back-End Of Line*). (b) Incremento localizado del campo eléctrico entre lineas metálicas en el BEOL debido a los cambios en el espesor de las lineas metálicas. Reproducido de [162]

.

mayor vulnerabilidad y en consecuencia modifican la distribución del tiempo de ruptura, la cual se vuelve no-Weibull dada la no uniformidad en la generación de defectos.

En este caso, existen dos variables aleatorias, siendo ellas el espesor del óxido, y el tiempo de ruptura, con lo que su estudio ha sido originalmente planteado mediante la convolución de dos distribuciones con un total de 6 parámetros. De aquí se puede inferir que este método de análisis es poco práctico en el caso de abordar un problema con múltiples variables aleatorias (>2). Una alternativa sumamente atractiva para el análisis de este problema fue propuesta por Wu *et al.* [136] mediante el denominado modelo de *Clustering*. Este modelo, es capaz de describir una distribución no-Weibull con solamente 3 parámetros, siendo sumamente útil para proyecciones de fiabilidad en fenómenos de ruptura que no sigan una distribución de Weibull.

Para derivar la formulación del modelo de *Clustering* se debe partir de un escenario en el cual los defectos no se distribuyan uniformemente. En este contexto, se tiene que la densidad de defectos es una variable aleatoria en lugar de una constante, con lo que la confiabilidad (Y) en función de la densidad de defectos puede escribirse como:

$$Y = 1 - F = \int_0^\infty e^{-AD'} f\left(D'\right) dD' \tag{4.3}$$

donde es F la probabilidad de falla acumulada y A el área del dispositivo. $f(D)$ es la función de densidad de probabilidad para la densidad de defectos, la cual sigue una distribución Γ

$$f(D) = \frac{1}{\Gamma(\alpha)\theta^\alpha} D^{\alpha-1} \exp\left(-\frac{D}{\theta}\right) \tag{4.4}$$

donde θ y α son los parámetros de escala y forma de la distribución Γ, respectivamente. Mediante la integración de la Ec. 4.3 se obtiene el modelo de confiabilidad binomial negativo

$$Y = 1 - F = \left(1 + \frac{A\bar{D}}{\alpha}\right)^{-\alpha} = \left(1 + \frac{\lambda}{\alpha}\right)^{-\alpha} \tag{4.5}$$

donde $\lambda = A\bar{D}$ es el número promedio de defectos para un área A y $\bar{D}$ es la densidad de defectos promedio. α se denomina "factor de *clustering*" y describe cuan agrupados en *clusters* están los defectos en el óxido (es decir, que tan no-uniforme es la distribución). Cuando $\alpha \longrightarrow \infty$ se puede demostrar matemáticamente que la Ec. 4.5 tiende a una distribución de Poisson, es decir $Y = e^{-AD} = e^{-\alpha}$. La Ec. 4.5 es conocida como modelo de *clustering* para confiabilidad. En dicho caso, solo se considera el contexto espacial (es decir, invariante en el tiempo). Al considerar el escenario de la ruptura dieléctrica, debe introducirse una variable dependiente del tiempo, lo cual se hace re-formulando $\lambda = AD$ como $\lambda = \left(\left(t/\tau\right)^\beta\right.$, en la Ec. 4.5, con lo que se llega a la expresión del modelo de *Clustering* dependiente del tiempo indicada en la Ec. 4.2.

4.3.2. Distribución de Weibull vs. modelo de *Clustering*

Suponiendo que la generación de defectos en el dieléctrico durante el estrés eléctrico siguiera una distribución uniforme en el volumen del dieléctrico (estadística de Poisson, independiente de las trampas inducidas por el proceso de fabricación), la estadística de BD estaría descrita por la distribución de Weibull [24]. Este ha sido el caso de los *films* de SiO_2/SiON usados típicamente hasta la introducción del nodo de fabricación de 45 nm ($t_{ox} \geq 2$nm). Sin embargo, en las tecnologías modernas de ULSI (*Ultra Large Scale Integration*) que comprenden materiales high-κ tales como el HfO_2, los defectos tienden a generase cerca de los ya existentes (siendo un caso extremo el del HfO_2 policristalino, donde los defectos se agrupan en la cercanía de los bordes de grano [135], [163]) y consecuentemente la estadística de TDDB resultante no siempre sigue un escalamiento de área en los altos percentiles [136].

Este escenario de generación heterogenea de defectos en el óxido de compuerta (producido por las características intrínsecas del material) se asemeja al observado en el dieléctrico de las capas BEOL a raíz las variaciones de espesor. Por lo tanto, si bien fue originalmente planteado alrededor de la variabilidad aleatoria del espesor del óxido de las capas BEOL, el modelo de *Clustering* es apropiado para otros procesos que no sigan una estadística de Poisson [136], [161]) y por ende también puede ser utilizado para el caso donde existan variaciones localizadas en la probabilidad de generación de defectos, tal como sucede en la proximidad de los bordes de grano en dieléctricos *high–*κ. Tal no-uniformidad en la generación de defectos se espera en las muestras de HfO_2 analizadas en este capítulo en la cercanía de los *clusters* de defectos inducidos por los iones incidentes [137]. Para comparar las bondades de ambos modelos, en esta sección se analizan las características del ajuste que cada uno permite sobre los datos experimentales obtenidos en la Sec. 4.3.

Los detalles del los ajustes de la estadística de TDDB se muestran en las figuras 4.7c-4.7f para el caso de control y tres fluencias diferentes de irradiación. Se puede observar que el modelo de *Clustering* reproduce mejor la leve concavidad "hacia abajo" observable en gráfico en escala Weibull de las mediciones experimentales, lo cual resulta en un ajuste con un mayor coeficiente de determinación (R^2, véase la Tabla 4.2). Los valores extraídos de β, $t_{63\,\%}$ y α_C en función de la fluencia de irradiación se muestran en las Figuras 4.7g-4.7h. Independientemente del área de irradiación y el modelo considerado, todos los parámetros de ajuste (β, $t_{63\,\%}$ y α_C) exhiben una tendencia decreciente con la fluencia de irradiación. Para el caso del ajuste utilizando el modelo de *Clustering* y considerando el incremento en la densidad inicial de defectos causada por la radiación (véase la figura 4.7g), β decrece de $\sim$2.9 a $\sim$1.5 para las dos áreas de irradiación consideradas. De la misma forma, α_C muestra una tendencia a la baja, decreciendo de $\sim$2.5 a $\sim$2.0 (para el área de irradiación más grade) y desde $\sim$2.5 a $\sim$1.6 (área más pequeña). $t_{63\,\%}$ también decrece desde $\sim$550 hasta $\sim$230 sec para la irradiación de mayor área y de $\sim$550 a $\sim$330

Tabla 4.2: Valores de R^2 (Coeficiente de determinación) para el ajuste con el modelo de *Clustering* tanto para las irradiaciones de área pequeña y área grande. Las celdas sombreadas indican el caso de mayor R^2 para cada escenario

Fluencia	**Área pequeña**			**Área grande**		
	Código	***Clustering***	**Weibull**	**Código**	***Clustering***	**Weibull**
Sin irradiar	F#	0.984	0.954	F#	0.984	0.954
$10^{11}\ \frac{iones}{cm^2}$	L#1	0.981	0.967	—	—	—
$10^{12}\ \frac{iones}{cm^2}$	L#2	0.993	0.984	S#2	0.990	0.909
$10^{13}\ \frac{iones}{cm^2}$	L#3	0.978	0.961	S#3	0.989	0.950

sec para la de menor. El valor de β ($\sim$2.9) extraído de la estadística de TDDB para las muestras de control, coincide con el valor estimado para dieléctricos high-κ mediante la teoría percolativa (asumiendo un tamaño de defectos, $a_0 \sim 1$nm y un tiempo de vida medio de los defectos de $\sim$0.35) [128], [164]. Vale la pena notar que los bajos valores de R_S (100 Ω-1 kΩ) en los dispositivos sujetos a la mayor área de irradiación y el máximo valor alcanzado por I_{HBD} (muy por debajo de los 10μA) resultan en una caída de tensión despreciable (menor a $\sim$10 mV), la cual no afecta la dinámica de TDDB.

Teniendo en cuenta estudios previos de radiación en estructuras MOS [148], [165], se sabe que los defectos generados por el proceso de irradiación antes del estrés eléctrico son creadas a lo largo de la trayectoria de los iones incidentes. Esto se debe a la interacción de los mismos con la estructura cristalina del material, lo cual causa daño por desplazamiento, la fundición localizada y la difusión de oxígeno [144], [147], [148]. Dado que este proceso implica una transferencia de energía el daño por desplazamiento no es estrictamente continuo continuo a lo largo de la trayectoria [148] sino discontinuo, dando lugar a *clusters* de defectos en la forma de caminos filamentarios parcialmente formados, con un diámetro de hasta $\sim$10 nm [145]. Dado que fluencias más altas inducen mas defectos, menos defectos adicionales deben ser agregados para completar alguno de los caminos filamentarios, lo que causa una mayor dispersión en el tiempo medio de ruptura (t_{HBD}). Por último, debe mencionarse que la estadística de TDDB obtenida en las muestras irradiadas es independiente del área de irradiación, y altamente dependiente de la fluencia utilizada, puesto que a iguales fluencias para distintas áreas se obtienen resultados similares en términos de la estadística de TDDB. Esto sugiere que los cambios inducidos por la irradiación son localizados y se producen solamente en las proximidades de la trayectoria del haz de iones.

Debido a que los datos experimentales son mejor reproducidos por el modelo de *Clustering*, la generación correlacionada de defectos emerge como la mejor explicación para las tendencias estadísticas observadas. Sin Embargo, la dinámica espacio-temporal de la generación de nuevos defectos debe ser estudiada en detalle para poder descartar el rol de defectos extrínsecos que enmascaren las tendencias estadísticas. Por lo tanto, el mecanismo de ruptura es estudiado en las próximas secciones mediante simulaciones físicas multi-escala.

4.3.3. Plataforma de simulación multi-física

Para complementar los resultados experimentales reportados en este capítulo, en este trabajo de tesis se ha considerado la utilización del Software Ginestra™. Ginestra™ es una herramienta de simulación física multi - escala que describe de manera auto - consistente los principales mecanismos físicos que tienen lugar en un dispositivo MOS sujeto a estrés eléctrico: Disipación de potencia e incremento de la temperatura, distorsión, ruptura de las uniones atómicas promovidas por el campo eléctrico y atrapamiento y transporte de carga, este último usualmente dominado por la conducción por efecto de túnel inelástico asistido por trampas (debido al relajamiento estructural y al acople de fonones).

El transporte de portadores se modela mediante varios mecanismos de conducción, tales como corriente de arrastre, Túnel Directo o de Fowler-Nordheim, tunel asistido por trampas (TAT) y emisión termo-iónica (emisión de Schottky) (véase la Sección 2.1.2 en el Capítulo 2). Entre todos ellos, el mecanismo de túnel asistido por trampas se posiciona como el mecanismo de transporte de carga dominante en los momentos previos al inicio de la ruptura dieléctrica [166]. La potencia disipada por los portadores en las proximidades de los defectos se utiliza para calcular el perfil de temperatura a través del dispositivo utilizando la ecuación de flujo calórico de Fourier [167]. Dicho perfil es utilizado para calcular por ejemplo, el campo eléctrico y la tasa de generación de nuevos defectos en el óxido (por ejemplo vacancias de oxigeno y defectos intersticiales, V0s), tomando en cuenta la correlación espacial del proceso de generación de defectos. Para ello, se considera una reducción en la barrera de energía necesaria para la formación de V0s en la vecindad (dentro de una rango equivalente a la distancia de una unión atómica, $\sim$3 Å) de un defecto existente. El incremento en la densidad de defectos causado por el estrés eléctrico y el incremento localizado de la temperatura conduce a una mayor de-localización de electrones y formación de defectos, los cuales a su vez propician un aumento de la corriente de arrastre, haciéndola la componente dominante de la corriente que circula entre los electrodos del dispositivo. También se consideran los cambios en las propiedades de los materiales (constante dieléctrica y conductividad térmica) en las regiones del aislante altamente degradadas, tales como el filamento conductivo que propicia la ruptura, lo que permite considerar también su naturaleza cuasi-metálica.

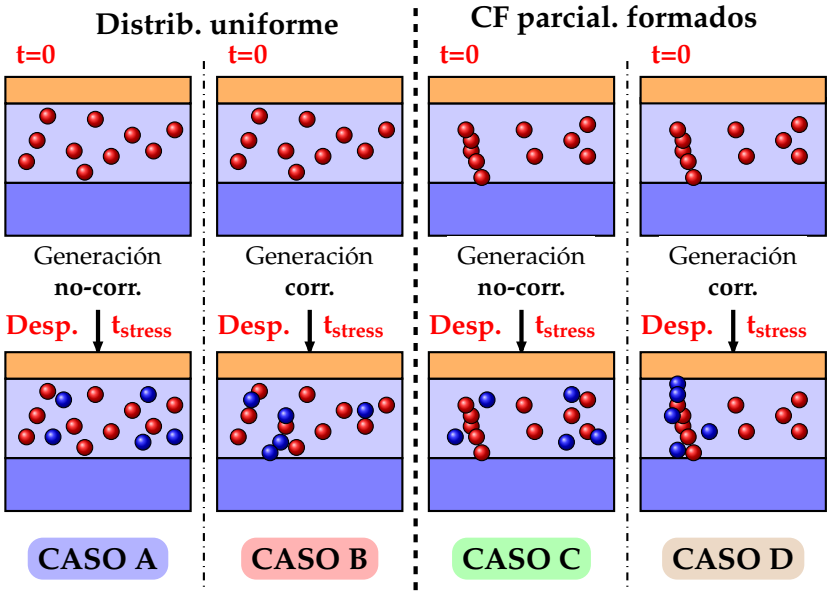


Figura 4.9: Representación esquemática de los diferentes escenarios de simulación. A tiempo $t = 0$, hay solamente 2 distribuciones diferentes: *i)* defectos distribuidos uniformemente y *ii)* defectos agrupados eh caminos filamentarios parcialmente formados. Luego de un tiempo de estrés t_{stress}, y considerando dos mecanismos diferentes para la generación de nuevos defectos (espacialmente correlacionado y no-correlacionado), surgen 4 escenarios posibles. Caso A: distribución de defectos uniforme con generación no-correlacionada, Caso B: distribución de defectos uniforme con generación correlacionada, Caso C: caminos filamentarios parcialmente formados con generación no-correlacionada y Caso D: caminos filamentarios parcialmente formados con generación correlacionada.

4.3.4. Identificación de la dinámica espacio-temporal de ruptura en dieléctricos *high–*κ

La evolución de la corriente de compuerta y las estadística de ruptura resultante fue simulada para varios dispositivos utilizando Ginestra™[137], [158][3]. El método cinético de Monte-Carlo fue utilizado en las simulaciones para describir la naturaleza estocástica del proceso de generación de defectos[132], considerando la misma estructura MOS que para los experimentos de irradiación. Aproximadamente 200 simulaciones fueron realizadas para cada uno de los cuatro escenarios representados esquemáticamente en la Figura 4.9. Dos de los escenarios considerados corresponden a una distribución uniforme de defectos y los dos restantes a *clusters* de defectos espacialmente localizados, i.e. caminos filamentarios parcialmente formados. Para cada uno de estos casos, la generación de nuevos defectos durante el estrés eléctrico fue modelada asumiendo tanto un mecanismo correlacionado como no-correlacionado. Esto resulta en cuatro escenarios diferentes (A-D) donde la dependencia de β con la fluencia de irradiación es estudiada para el caso de la ruptura abrupta (HBD). Las figuras 4.10a-4.10c muestran las gráficas de β vs. fluencia junto con la evolución temporal de los caminos filamentarios en cada caso. Nótese que el caso C se omite por brevedad, dado que es muy similar el caso A pero con

[3]Simulaciones realizadas en estrecha colaboración con Andrea Padovani, de *Applied Materials*

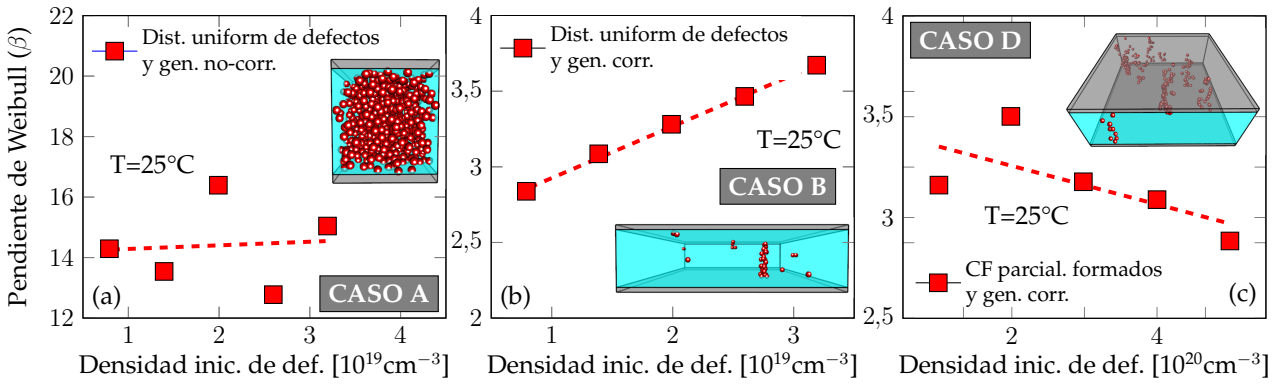


Figura 4.10: Tendencias observadas de la pendiente de Weibull (β) obtenidas mediante simulaciones realizadas utilizando la plataforma Ginestra^{TM}, en función de la densidad inicial de defectos (la cual es ajustada artificialmente mediante la fluencia de irradiación en el estudio experimental), correspondientes a los Casos (a) A, (b) B y (c) D definidos en la Fig. 4.9. Para cada caso, se incluye una representación gráfica de los defectos distribuidos al momento de HBD, generada a partir de una simulación tomada al azar. Nótense las significativas variaciones en el factor de *clustering* (generación no-correlacionada o correlacionada) y del número de defectos necesarios para alcanzar el estado de HBD.

valores de β levemente menores. Para simular la generación correlacionada de nuevos defectos, se asumió una reducción de la energía de activación (E_a) de 0.5 eV, lo cual está en linea con los resultados obtenidos por DFT [132], [163].

En primer lugar, a partir de la Fig. 4.10a se puede inferir que si los defectos existentes antes del estrés eléctrico están distribuidos uniformemente en el volumen del dieléctrico y la generación de nuevos defectos por estrés eléctrico es no-correlacionada (Caso A), entonces β es independiente de la fluencia de radiación, a diferencia de los resultados observados experimentalmente. En segundo lugar, para el caso B donde se considera una distribución inicial de defectos uniforme y un mecanismo de generación correlacionado (véase la Figura 4.10b) se puede apreciar un incremento de β en función de la fluencia de irradiación, también contrario a los datos experimentales expuestos en la Fig. 4.7. Es importante resaltar que los bajos valores de β obtenidos mediante el procedimiento de simulación de la ruptura dieléctrica coinciden con los valores hallados experimentalmente solo si se considera la correlación espacial en la generación de defectos (Casos B y D). Por el contrario, para los casos donde se asume una generación de defectos no-correlacionada (Casos A y C), β es significativamente más alto. Interesantemente, la combinación de bajos valores de β con una tendencia a la baja es solo observable para el caso D (Fig. 4.10c) [137]. Estos resultados sostienen la hipótesis de que mediante experimentos de irradiación controlada se pueden crear caminos filamentarios parcialmente formados y que el subsiguiente estrés eléctrico produce defectos en la proximidad de los defectos pre-existentes. Para las condiciones de irradiación consideradas hay una reducción en la densidad de defectos requerida para producir la ruptura dieléctrica, lo cual explica la leve dependencia entre β y t_{ox} [88], [127].

4.4. Conclusiones

En este capítulo, se investigó la evolución espacio–temporal de los defectos que gobiernan la ruptura dieléctrica dependiente del tiempo (TDDB) en dieléctricos *high–*κ, considerando el fenómeno de *Clustering*, mediante experimentos de radiación localizada con iones de alta energía para inducir una densidad controlada de defectos en la capa dieléctrica, y complementado mediante simulaciones multi-físicas. La misma estadística de TDDB se obtuvo considerando distintas áreas de irradiación para diferentes fluencias, lo cual indica una clara dependencia con la fluencia, sugiriendo un fenómeno localizado. Interesantemente, la estadística experimental de TDDB solo puede ser representada apropiadamente mediante el modelo de *Clustering* en lugar de la distribución de Weibull. Los experimentos de CVS fueron repetidos mediante simulaciones, considerando tanto (I) defectos distribuidos aleatoriamente o formando caminos filamentarios parciales (*Clusters* de defectos) y (II) correlación o no-correlación espacial en la generación de nuevos defectos. Se ha descubierto que solo al considerar caminos filamentarios parcialmente formados y generación espacialmente correlacionada de nuevos defectos en las simulaciones, la tendencia observada de la pendiente de Weibull en función del espesor del óxido reproduce los datos experimentales. Esto da sustento experimental a las teorías estadísticas que sugieren la necesidad del modelo de *Clustering* para el estudio de fiabilidad en dieléctricos *high–*κ. En este escenario, dado que los nuevos defectos se generan preferentemente en la cercanía de los ya existentes, el número crítico de defectos necesarios para desencadenar la ruptura se reduce notablemente, y por consecuencia se reduce también la dependencia con el espesor del óxido, lo cual explica la relativa insensibilidad de la pendiente de Weibull con respecto al espesor del óxido de compuerta. No obstante, debe señalarse como trabajo futuro una análisis por seperado de la estadística asociada al *Soft-Breakdown*, para identificar el impacto del efecto de *Clustering* en estas etapas. Asimismo, esto podría ser de utilidad para interpretar el rol de los puntos de disipasión térmica a lo largo del filamento conductivo. Por otro lado, el caso de estructuras bi-capa debe ser considerado dada su altísima importancia tecnológica y posible aplicación en memorias no volátiles.

5 Conmutación Resistiva en dieléctricos

COMO se ha discutido en el capítulo anterior, la ruptura dieléctrica es un fenómeno de gran interés para la comunidad de fiabilidad, dado que limita la vida útil de los dispositivos CMOS. Sin embargo, ciertos materiales aislantes tales como los óxidos de metales de transición (TMO) o el nitruro de boro hexagonal (h-BN) presentan una ruptura dieléctrica reversible, denominada conmutación resistiva (*Resistive Switching*, RS). En este escenario, la capa aislante es utilizada como un medio para almacenar información, la cual se codifica en términos del estado de conducción del material, dando lugar a las denominadas Memorias Resistivas de Acceso Aleatorio (*Resistive Random Access Memory, RRAM*). Estas son actualmente un interesante tópico de estudio debido a su prometedora capacidad para la integración en gran escala de celdas de memoria [168] y su aplicación en sistemas de computación neuromórfica [169]. En particular, las características estáticas de la relación I-V en dispositivos RRAM han sido extensamente investigados. Sin embargo, la evolución temporal del fenómeno de RS no ha sido completamente desvelada aún. En este contexto, es claro que a pesar de los recientes avances en relación al modelado de la dependencia I-V de los dispositivos RRAM [170], [171], se requiere más información sobre el evento de SET (es decir, el evento de conmutación). Un conocimiento detallado de la evolución morfológica del medio aislante durante el proceso de electro-formado y el SET así como la naturaleza de las curvas I-V asociadas son necesarios para modelar precisamente el fenómeno de RS y realizar estimaciones de confiabilidad. En última instancia, el estudio de la dependencia temporal del RS es imprescindible para estimar el tiempo mínimo requerido para escribir un bit de información en una celda RRAM [172]. En este capítulo, se estudia la dinámica temporal de la conmutación resistiva y su dependencia con la tensión de *switch* sobre un gran número de mediciones de corriente en función del tiempo. En base a una descripción física de de la dinámica de ruptura en dieléctricos de uso común en procesos de fabricación CMOS planares y su similitud con el evento de SET en memorias RRAM, se propone un modelo para la evolución del estado resistivo en RRAMs en función de la tensión de *switch* aplicada.

5.1. Similitudes con el mecanismo de ruptura progresiva

Los dispositivos RRAM basan su funcionamiento en el mecanismo de Conmutación Resistiva (*Resistive Switching*, RS), el cual consiste en la formación y disolución de un filamento conductivo (CF, *Conductive Filament*) nanoscópico a través de la capa aislante en una estructura Metal-Aislante-Metal (*Metal-Insulator-Metal*, MIM), típicamente el óxido de un metal de transición tal como NiO [173], TiO_x [174], HfO_x [175], TaO_x[176], entre otros, o tambien en dieléctricos 2D, tales como el hBN [169]. El filamento conductivo se genera inicialmente por una operación de electro-formado (*electro-forming*), la cual consiste en la ruptura dieléctrica (*Breakdown*, BD) controlada de la capa aislante de la estructura MIM, mediante la apropiada limitación de la corriente que circula por el dispositivo [177]. En este punto, el dispositivo presenta un estado de baja resistencia (*Low Resistance State, LRS*). Acto seguido, la operación de RESET produce la desconexión de dicho filamento al generar una pequeña discontinuidad (*gap*) en el mismo, lo cual produce una transición de LRS a un estado de alta resistencia (*High Resistance State*, HRS). La transición complementaria, es decir de HRS a LRS, se produce por el contrario al completar el *gap* del filamento en HRS mediante una operación de SET. Tanto la operación de SET como de RESET se logran mediante la aplicación de pulsos de tensión, de polaridades iguales (conmutación unipolar) u opuestas (conmutación bipolar). El mecanismo de RS ha sido ampliamente investigado en los últimos años [168], [169], [171], [173]-[181], y un gran número de autores concuerdan en que los procesos de SET y RESET se deben a la migración de iones inducida por el aumento localizado tanto del campo eléctrico como de la temperatura (debido al efecto Joule) [179]. Dicha migración, incluye componentes de arrastre y difusión, en las cuales el mecanismo fundamental es el salto de iones (*ion-hopping*), donde los mismos se mueven entre pozos de potencial, los cuales proveen estados para la localización de iones [180], [181].

Por otro lado, recientemente se ha identificado el proceso físico detrás de la dinámica de ruptura en dieléctricos ultra-delgados Al_2O_3, HfO_2, SiO_2, Si_2N_4) comúnmente usados en los procesos de integración CMOS planares[18], [182]. En pocas palabras, la transferencia de energía desde el CF hacia su entorno, promueve la difusión de las especies atómicas presentes en la estructura, las cuales contribuyen a un engrosamiento gradual del CF que une los electrodos del dispositivo MIM, causando un aumento progresivo de la corriente de fuga. Interesantemente, los eventos de SET y BD muestran el mismo comportamiento estadístico[171], y cambios micro-estructurales similares en el dieléctrico [31], [183]-[186]. Varios autores [31], [183], [184] han mostrado que la región de ruptura (*BD spot*) está caracterizada por la formación de una región rica en Silicio (en dispositivos poly-Si/SiO_xN_y/Si) o rica en metal (en estructuras *Metal Gate*/high-K/Si) en el dieléctrico de compuerta. En este sentido, Privitera *et al.*[185] han reportado la presencia de especies metálicas en el oxido de compuerta (HfO_2) para el caso de dispositivos MIM RRAM luego del proceso de electro-formado.

Se ha demostrado experimentalmente que tanto el evento de SET como el de *forming* en dispositivos RRAM [171], [187] y la ruptura dieléctrica del óxido de compuerta [18], [182] tienen ciertos aspectos en común. Aparte del incremento ruidoso y progresivo de la corriente de fuga, cuya tasa de crecimiento depende de la tensión de estrés, un análisis más detallado de estos eventos revela otros puntos de contacto. Mediante el estudio del impacto de la limitación de corriente se ha observado para ambos fenómenos (evento de SET [36] y BD [177]), una clara dependencia entre las características del CF y la máxima corriente que fluye a través del dispositivo. Adicionalmente, imágenes TEM (Microscopía de Transmisión Electrónica) de capacitores MOS de Si capturadas antes y después del evento de BD [31], [183], [184] y celdas RRAM de HfO_2 tomadas durante el ciclado [185], [186] muestran cambios micro-estructurales comparables en el óxido, sugiriendo la difusión de las especies atómicas del ánodo hacia la capa dieléctrica en ambos escenarios.

Por lo tanto, no solamente existen similitudes a nivel de las características eléctricas ente los eventos de SET en RRAM y BD en óxidos de compuerta sino también en cuanto a los cambios micro-estructurales que se producen, lo cual sugiere que ambos fenómenos podrían estar dominados por el mismo mecanismo físico. En este contexto, se propone modelar los resultados obtenidos para el evento de SET en RRAM en el marco del modelo propuesto por Palumbo *et al.* [18]. El mismo ha sido exitosamente utilizado para analizar el fenómeno de BD en óxidos de compuerta tanto en óxidos ultra delgados de SiO_2 sobre sustratos de Silicio, como también diversos materiales *high-k* con distintas conductividades térmicas sobre sustratos III-V así como también materiales 2D [70]. Dicho modelo tiene en cuenta la naturaleza progresiva del evento de BD y la cuantifica en términos de dI_{BD}/dt (denominado *Degradation Rate*, DR), donde I_{BD} es la corriente a través del dispositivo durante la ruptura progresiva (PBD).

Con respecto a la física detrás de este fenómeno, el modelo asume que el proceso de BD está relacionado con la transferencia de energía desde el CF hacia la red cristalina que lo rodea. De acuerdo con esta idea, el incremento localizado de la temperatura, asociado a la alta densidad de corriente fluyendo por el CF (en el orden de los MA/cm^2 a través de un CF con sección de 1-50 nm^2 [18], [31], [183]) promueve la electro-migración de ciertas especies atómicas presentes en la estructura, lo cual contribuye al crecimiento del CF que conecta los electrodos. La presencia de tal mecanismo de electro-migración ha sido documentado por Tang *et al.* [188]. Esta interpretación del fenómeno de BD tiene en cuenta la dependencia de las características del CF con la máxima corriente que circula por el CF y la presencia de las especies atómicas del ánodo en la capa dieléctrica después del evento de BD, por lo que puede ser usado para modelar el evento de SET en RRAMs.

5.2. Conmutación Volatil y No-Volatil

En la mayoría de las estructuras RRAM del tipo MIM, el evento de conmutación resistiva es un fenómeno no-volátil, es decir que el cambio en la conductancia se mantiene estable incluso luego de quitar la polarización [189]. Sin embargo, cuando se utilizan determinadas combinaciones de materiales para los electrodos y la capa aislante, el cambio de conductancia es volátil, de forma que luego del evento de SET, el estado inicial de conducción puede ser recuperado al remover la tensión de excitación [189]. Por otro lado, este fenómeno, comúnmente denominado Conmutación Resistiva de tipo Umbral (*Threshold Type RS*), puede producirse tanto antes como después del proceso de electroformado. La diferencia entre *Memory Type RS* y *Threshold Type RS*, puede verse en la Fig. 5.1.

Con relación al caso de RS volátil, recientemente se ha reportado que dispositivos MIM con electrodos de Ag, muestran un excelentes características de *Threshold RS* debido a la alta difusividad de los iones de Ag^{+} en diversos aislantes (incluyendo SiO_X, HfO_X y MgO_X) [192]. Más aún, se ha descubierto que la combinación de electrodos de Ag y un dieléctrico de nitruro de boro hexagonal (hBN) multicapa fabricado por (*Chemical Vapour Deposition*, CVD) resulta en dispositivos de conmutación volátil con un consumo de energía ultra bajo (∼8.8 zJ)[193] – Nótese que este es el consumo de energía más bajo reportado a la fecha para dispositivos de conmutación resistiva de cualquier tipo, y

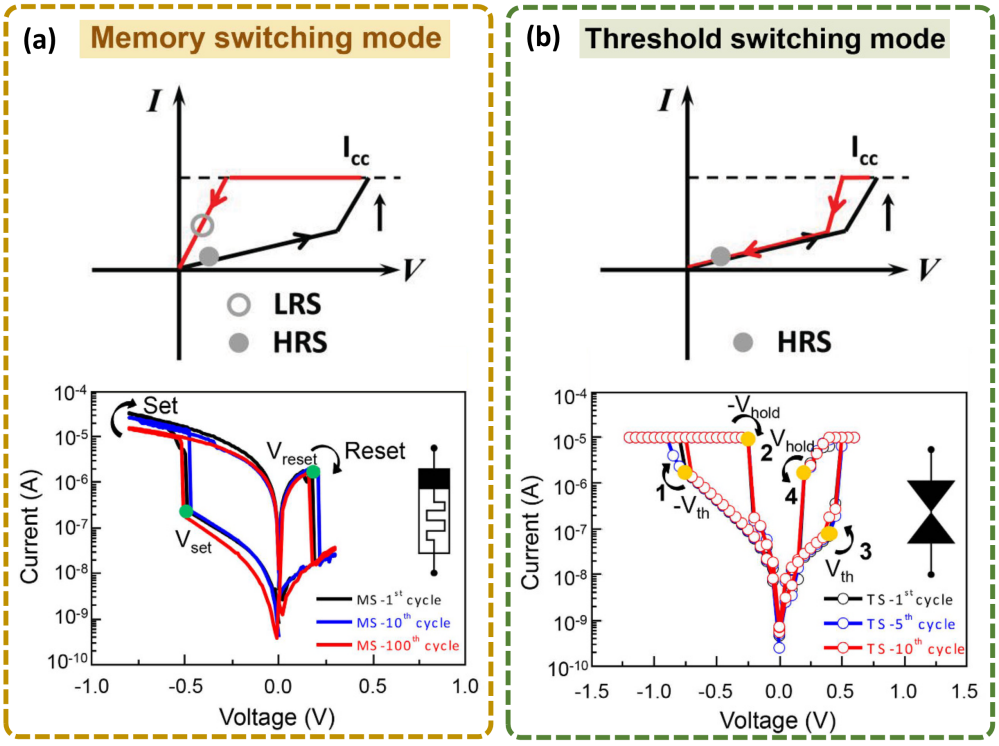


Figura 5.1: Comparación entre (a) *Memory Type RS* y (b) *Threshold Type RS*. La parte superior de cada sub-figura muestra una representación esquemática del fenómeno mientras que en la parte inferior se presentan ejemplos experimentales consistentes en una estructura de (TiO_X amorfo)/(nano-partículas de Ag)/(TiO_X poli-cristalino). Adaptado de [190] y [191].

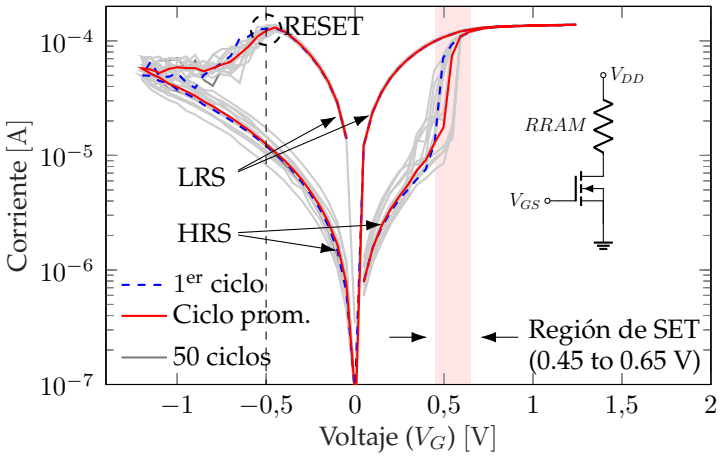


Figura 5.2: Mediciones I–V para las muestras bajo prueba (50 curvas en total). La línea continua roja indica la curva promedio, mientras que la azul de trazos representa el 1er ciclo. En el *inset* de la derecha se muestra el arreglo 1T1R, formada por el dispositivo RRAM y el correspondiente transistor N-MOS de control.

se encuentra próximo al límite fundamental impuesto por el ruido térmico (4.1 zJ) [194]. Además, otra característica sobresaliente de los dispositivos metal/hBN/metal es que exhiben la co-existencia de conmutación volátil y no volátil [195], lo que permite usarlos como sinápsis electrónicas que emulen mecanismos de plasticidad tanto de largo (*Long Term Plasticity*, LTP) como de corto plazo (*Short Term Plasticity*, STP) [169]. Se cree que esto depende de la sección del CF creado durante el electro-formado, el cual puede ser controlado limitando la energía disipada durante dicho evento [169], ya sea utilizando una limitación de corriente en mediciones con rampas de tensión (*Ramped Voltage Stress*, RVS) o bajas tensiones o tiempos cortos durante mediciones de estrés pulsado (*Pulsed Voltage Stress*, PVS). No obstante, los factores clave que permiten este comportamiento dual no están completamente desvelados y se requiere más estudio de este fenómeno.

En las siguientes Sub-Secciones 5.2.1 y 5.2.2 se presentan, mediante experimentos de caracterización eléctrica, el comportamiento estático (Corriente–Tensión) y dinámico (Corriente–Tiempo) para memorias volátiles y no volátiles, respectivamente.

5.2.1. Conmutación No-Volátil, o *Memory Resistive Switching*

Para abordar el estudio de mas memorias RRAM de conmutación resistiva no-volátil (o *Memory Type Resistive Switching*) se consideraron dispositivos RRAM de HfO_2[1]. La estructura MIM correspondiente consiste en un *film* de 10 nm de HfO_2 depositado por ALD entre dos electrodos de Ti y TiN. La celda resultante se conecta en serie con un transistor de efecto de campo (MOSFET) de canal N, formando un arreglo del tipo "1T1R"(1 Transistor-1 Resistor), como se muestra en el *inset* de la Figura 5.2. Dicho transistor controla la corriente máxima que puede fluir a través de la celda de memoria, la

[1]Mediciones realizadas en la Universidad Autónoma de Barcelona

cual es responsable de determinar la ventana resistiva (resistencia en HRS, R_{HRS}, y resistencia en LRS, R_{LRS}) del dispositivo. [36], [177]. Todas las mediciones fueron realizadas aplicando una tensión entre compuerta y fuente (V_{GS}) de 1.2 V.

Las mediciones I-V quasi-estáticas se realizaron utilizando un analizador paramétrico para semiconductores (*Semiconductor Parametric Analyzer*, SPA) Keithley 4200-SCS equipado con una unidad de medición rápida y generador de pulsos (*Fast Measurement and Pulse Generator Unit*, FM-PGU) 4225-RPM, capaz de proveer la resolución temporal adecuada (hasta 200 ns para la tensión más alta, como se ve en la Figura 5.3b). En linea con trabajos previamente publicados considerando muestras similares [170], el evento de electro-formado se produce a $\sim 3{,}8V$ y se puede apreciar la bi-polaridad del evento de RS. Nótese además que las tensiones promedio a las cuales se producen los eventos de SET y RESET son aproximadamente simétricas ($\sim \pm 0{,}5$V). Esta característica sugiere que en HfO_2 ambos son procesos controlados por la tensión aplicada, lo cual se ajusta a los resultados reportados en la literatura [196]. Más detalles al respecto de los dispositivos y el *set-up* experimental se pueden encontrar en [171].

Con el objetivo de estudiar el evento de SET, el proceso completo (incluyendo la fase de degradación progresiva y transición HRS a LRS) debe ser explorado. Por lo

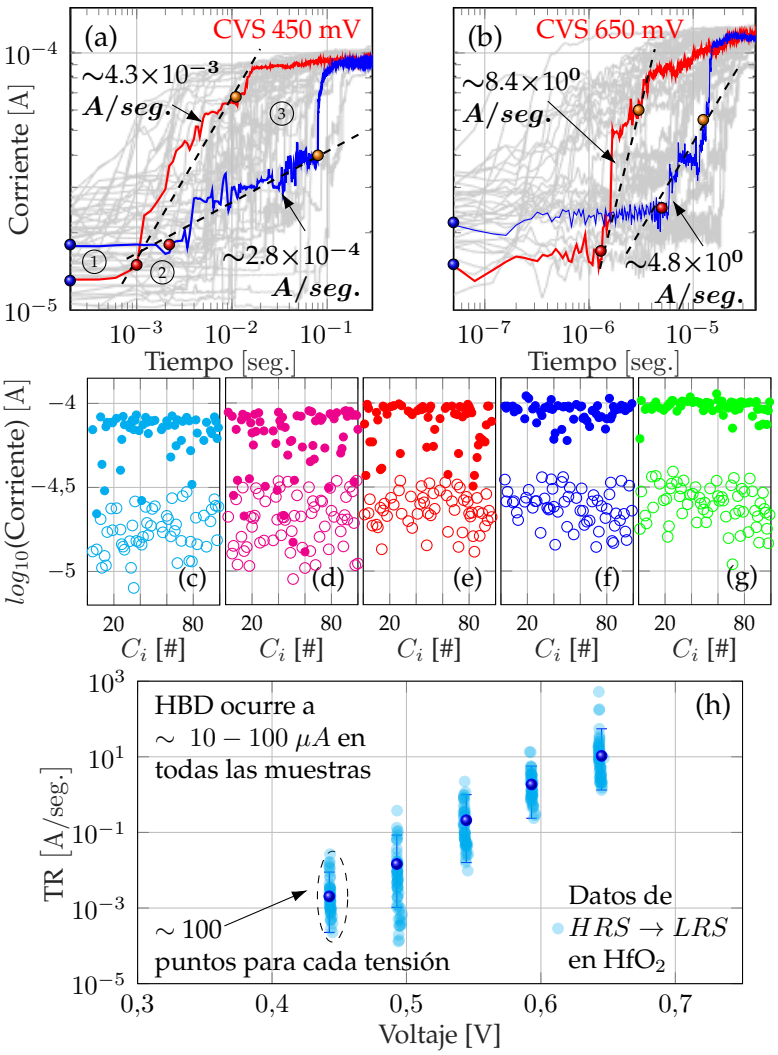


Figura 5.3: Evolución de la corriente en función del tiempo para mediciones CVS. Por claridad solo se muestran los casos correspondientes a la (a) mínima —450 mV— y máxima —650 mV— tensión. Los marcadores 1, 2 y 3 indican la corriente inicial (I_{init}), el inicio del aumento progresivo de la corriente (I_{on}) y el momento del salto a la corriente límite (I_{end}), para las mediciones CVS realizadas a (c) 450 mV, (d) 500 mV, (e) 550 mV, (f) 600 mV y (g) 650 mV. En las figuras (c)-(g), C_i indica el número de ciclo. (h) Tasa de transición (*Transition Rate*, TR=dI_{Tr}/dt) de las muestras bajo estudio. El valor medio de TR junto con ~100 valores medidos se reporta para cada tensión de estrés (450, 500, 550, 600 y 650 mV). Nótese que el valor medio se incrementa en aproximadamente 1 orden de magnitud cada 50 mV.

tanto, se realizaron mediciones de corriente en función del tiempo para tensión constante (*Constant Voltage Stress*, CVS) utilizando el *set-up* de gran ancho de banda mencionado en el párrafo anterior, hasta que la transición rápida de HRS a LRS ocurre o, en otras palabras, hasta que la corriente a través del dispositivo alcance la limitación de corriente ($\sim 100\ \mu$A). Las mediciones CVS fueron efectuadas a tensiones de 450 mV, 500 mV, 550 mV, 600 mV y 650 mV. Por brevedad, solo aquellas transiciones correspondientes a tensiones de estrés de 450 mV y 650 mV se muestran en las Figs. 5.3a y 5.3b. Dichas tensiones de estrés fueron aplicadas en el estado HRS y fueron seleccionadas en base a las curvas I-V representadas en la Fig. 5.2, donde la transición de HRS a LRS toma lugar para voltajes en un rango de 0.45 a 0.65 V.

La corriente a través de la estructura MIM durante la transición HRS a LRS (I_{Tr}) se incrementa gradualmente a lo largo del tiempo[197], dando cuenta de la naturaleza progresiva del evento de SET (ver Figs. 5.3a y 5.3b). Es un proceso ruidoso y progresivo que coincide con lo expresado en la literatura [171], [177], [179], [187] y cuya duración muestra una fuerte dependencia y dispersión de voltaje. Dicha dispersión será abordada en la Sec. 5.3. La tasa de crecimiento de la corriente máxima está limitada por el ancho de banda del equipo. Las Figuras 5.3c a 5.3g muestran las corrientes iniciales (I_{init}, símbolos vacíos) y finales (I_{end}, símbolos completos) de los transitorios adquiridos, incluyendo aquellos mostrados en la Fig. 5.3a y 5.3b. Cabe destacar que la corriente inicial coincide con los datos I-V ilustrados en Fig. 5.2, donde la región SET muestra corrientes iniciales cercanas a $10\ \mu$A. La corriente I_{end} representa el nivel desde el cual el transitorio adquirido presenta un salto a la corriente límite (ver Fig. 5.3), dejando a la muestra en LRS.

La evolución a lo largo del tiempo de la transición de HRS a LRS es cuantificada por la pendiente dI_{Tr}/dt, tal cual está definida en las Refs.[18], [198], [199]. Dicha métrica será subsecuentemente re-definida como Tasa de Transición (*Transition Rate*, TR) [A/s]. Los valores TR fueron experimentalmente evaluados a través de mediciones como aquellas de la Fig. 5.3a y 5.3b y se reportan para aproximadamente 100 mediciones para cada valor de voltaje (véase la Fig. 5.3h). La comparación entre TR en las Figs. 5.3a y 5.3b y la tendencia en la Fig. 5.3h sugieren una alta dependencia de voltaje, dado que TR se incrementa casi cuatro órdenes de magnitud entre los dos casos, es decir en un rango de 200 mV. Mediciones similares de la transición HRS-LRS han sido previamente reportadas [187], mostrando una dependencia de voltaje comparable (TR se incrementa a medida que lo hace el voltaje aplicado).

En este punto es crucial resaltar que la dependencia con la tensión observada en la métrica de TR es mucho más fuerte que la observada para el DR durante el evento de BD (dI_{BD}/dt, 3-5 órdenes de magnitud por volt)[18]. Para entender el origen de tal diferencia, en las siguientes secciones se planteará un análisis pormenorizado del modelo propuesto así como también de los cambios en los parámetros involucrados entre el caso del BD del óxido de compuerta y el SET en memorias RRAM

5.2.2. Conmutación Volatil, o *Threshold Resistive-Switching*

Para estudiar el fenómeno de conmutación resistiva volátil, se han utilizado estructuras Ag/h-BN/Au con un tamaño de 150 nm×150 nm[2]. Las mismas han sido fabricadas siguiendo un proceso que podría ser divido en dos etapas: En primera instancia, se sintetizó un sistema multi-capa de h-BN (aproximadamente 4 capas) mediante CVD, utilizando borano de amoníaco (H_3NBH_3) como precursor y un sustrato de cobre (Cu) de ∼20 μm. La estructura multi-capa resultante, así como su espesor efectivo (∼1.3 nm) fue posteriormente verificado mediante microscopía TEM de sección transversal (JEOL JEM-2100), presentándose la misma en la Fig. 5.4a. Nótese que para el proceso de obtención de las imágenes TEM, se ha depositado una bi-capa de protección —Ti(20nm)/Au(40nm)— sobre el h-BN. En segundo lugar, el h-BN multi-capa fue transferido utilizando la técnica reportada en [169] sobre electrodos de Au (150 nm×100 μm×50 nm —ancho, largo, alto—), los cuales fueron previamente depositados sobre una oblea de SiO_2(300 nm)/Si utilizando litografía de haz de electrones (*Electron Beam Litography*). Finalmente, sobre el multi-capa de h-BN transferido se depositó un electrodo de Ag de las mismas dimensiones que el inferior, pero rotado 90°. Dada la rotación de los electrodos superior e inferior la estructura MIM resultante tiene un área de 150 nm×150 nm. En todos los casos las capas metálicas fueron depositadas mediante evaporación por haz de electrones (*Electron Beam Evaporation*). En la Fig. 5.4b se presenta una imagen obtenida por microscopía electrónica de barrido (*Scanning Electron Microscopy*, SEM) de la estructura Ag/h-BN/Au completa.

La caracterización eléctrica de los dispositivos fabricados fue realizada utilizando una estación de prueba Cascade modelo M150, conectada a un SPA Keysight modelo B1500A, aplicando la tensión de prueba sobre el electrodo superior (Ag) mientras el electrodo inferior (Au) se conecta al potencial de referencia (0V). La existencia del fenómeno de conmutación resistiva en estos dispositivos fue verificada aplicando secuencias de mediciones RVS con diferentes limitaciones de corriente. Las gráficas de I–V se presentan en las Figuras 5.4c y 5.4d, demostrando que estos dispositivos presentan conmutación volátil cuando la corriente límite se establece en valores iguales o menores a 1 μA, así como también conmutación no-volátil cuando dicha corriente límite se fija en valores superiores. En este último caso, la potencia disipada en la vecindad del CF aumenta, lo que incrementa localmente la temperatura causando la expansión lateral del CF debido a la transferencia de momento desde los portadores que circulan por el CF hacia los átomos cercanos [200].

Al igual que en la Sub-Sección 5.2.1, el estudio de la dinámica de SET implica la medición de la evolución de la corriente a través del dispositivo de RS en función del tiempo. No obstante en este caso se ha optado por una metodología de PVS, es decir, aplicando pulsos de tensión constante a fin de replicar las condiciones de funcionamiento de estos dispositivos cuando se utilizan como sinapsis artificiales en redes neuronales de impulsos (*Spiking Neural Networks*, SNN) [201]. De esta forma, se realizaron sucesivos

[2]Dispositivos fabricados por la Universidad de Soochow, China

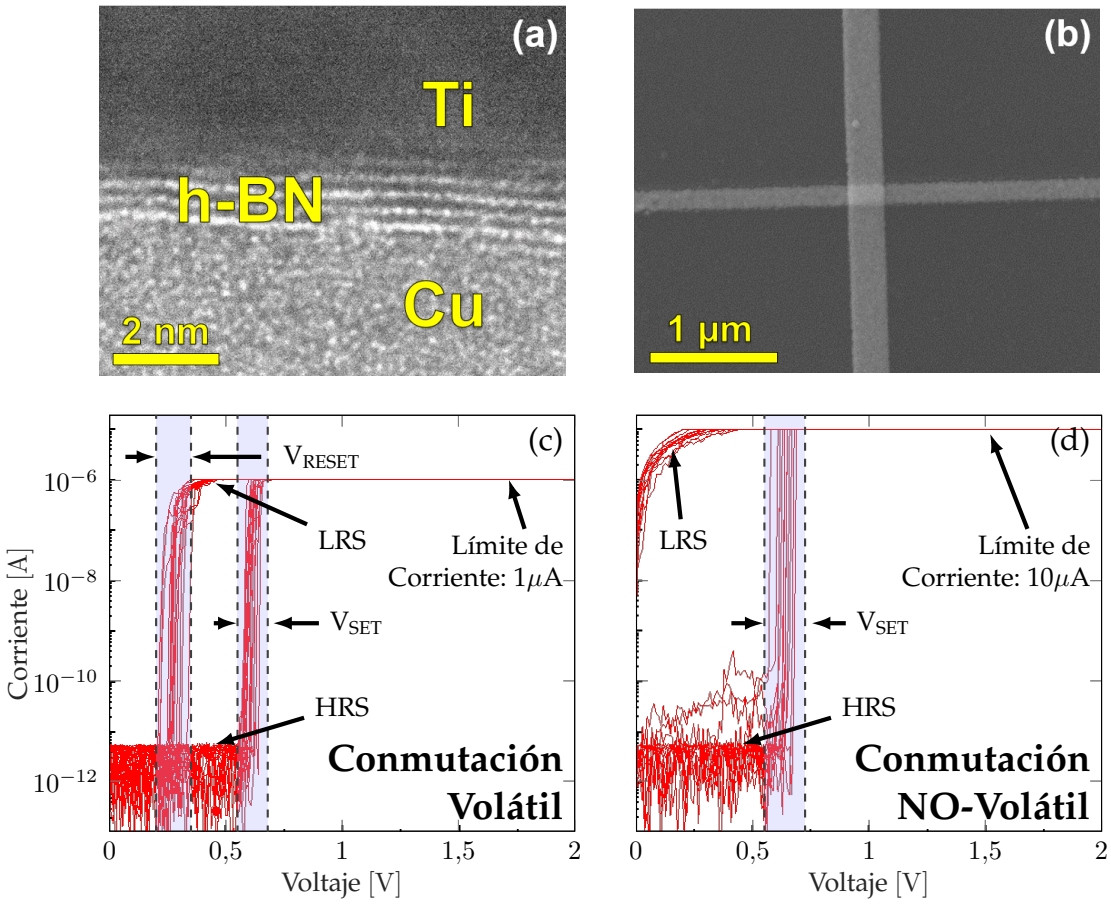


Figura 5.4: Estructura y características eléctricas de los dispositivos Ag/h-BN/Au bajo estudio. (a) Imagen TEM de sección transversal de la multi-capa de h-BN crecido mediante CVD sobre el sustrato de Cu, el cual está cubierto por un film delgado de Au/Ti, demostrando la existencia de defectos nativos en el h-BN. (b) Imagen SEM de una estructura Ag/h-BN/Au. Nótese la disposición perpendicular de los electrodos, dando lugar a una estructura tipo *cross-point*. (c) Ciclos de RS volátil obtenidos sobre un dispositivo Ag/h-BN/Au con una limitación de corriente de 1 μA. (d) Ciclos de RS no-volátil obtenidos sobre el mismo tipo de dispositivos pero con una limitación de corriente de 10 μA. Las regiones sombreadas en (c) indican el SET (aprox. entre 0.5 V y 0.6 V) y RESET (aprox. entre 0.2 V y 0.3 V) durante las rampas de tensión creciente y decreciente, respectivamente. En forma análoga, se reporta en (d) el caso de la tensión de SET entre 0.5 V y 0.7 V aprox. (nótese que en este caso no se observa el RESET)

.

experimentos de PVS utilizando varias tensiones para los pulsos de escritura (V_E, fijada en 2 V, 2,5 V, 3 V, 3,5 V y 4 V) a fin de observar los cambios en la conductancia de los dispositivos MIM, e intercalando entre cada uno de ellos, un pulso de lectura de una amplitud V_L fija y siempre igual a 0.1 V. La duración de cada uno de estos pulsos (T_E y T_L) se mantuvo siempre en 2 mseg., así como la separación entre ellos (T_{Bajo}), periodo en que la tensión aplicada es de 0 V (V_{Bajo}). La forma de onda resultante se puede apreciar en la Fig. 5.5a (linea azul) así como la corriente conducida a través de la estructura Ag/h-BN/Au, para el caso de V_E=2 V. Durante los periodos T_{Bajo}, la corriente fluctúa alrededor de $\sim$1 nA, nivel coincidente con el piso de ruido del SPA para el rango de medición utilizado en este experimento y coincide también con otros casos reportados en la literatura [202]. Al igual que en el caso no-volátil presentado en la Sub-Sección 5.2.1, la corriente de

transición (I_{Tr}) a través del dispositivo muestra un aumento progresivo y ruidoso hasta alcanzar un nivel casi constante (I_{end}) que depende de V_E (véase el punto 3 en la Fig. 5.5b). Para cada valor de V_E, I_{end} presenta cierta variabilidad, pero la tendencia general con V_E es creciente. Posteriormente a la aplicación de cada pulso de escritura, la corriente vuelve a ∼1 nA, incluso durante la aplicación de los pulsos de lectura, lo que indica que el estado de conducción de la estructura Ag/h-BN/Au previo al pulso de escritura fue totalmente recuperado.

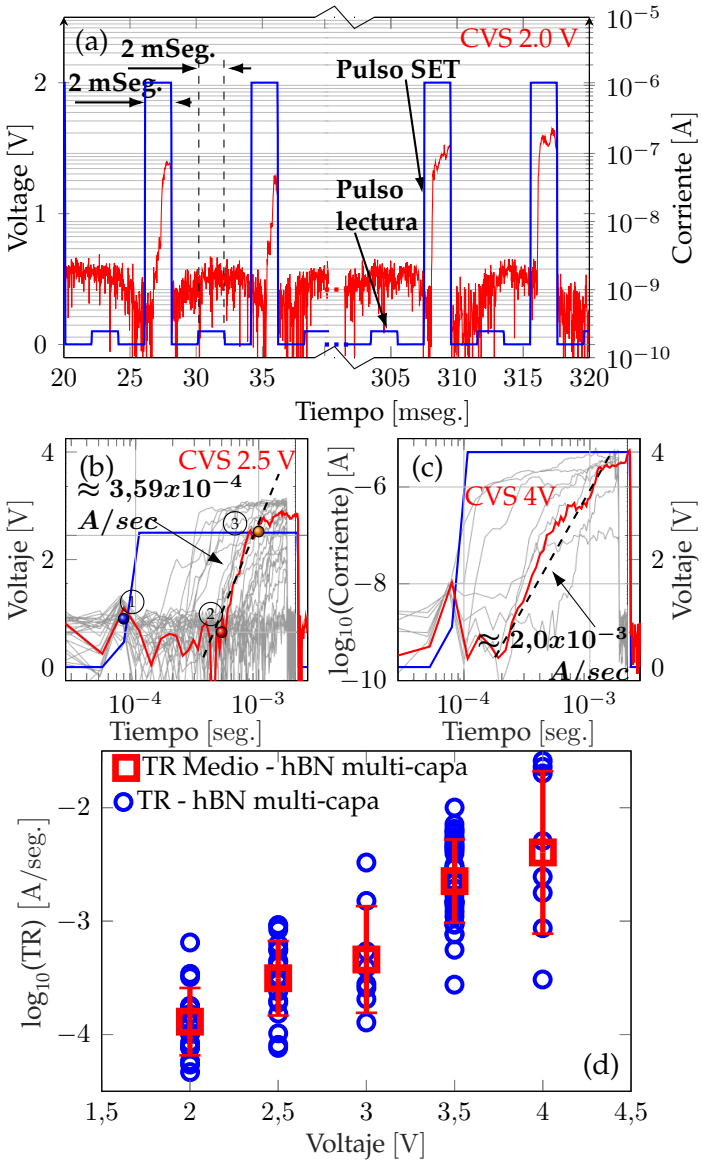


Figura 5.5: Mediciones PVS de la estructura Ag/h-BN/Au realizadas a diferentes tensiones V_E. (a) Secuencia de pulsos con V_E=2 V mostrando la transición del nivel de corriente entre HRS a LRS cuando cada pulso es aplicado y la relajación una vez removida la tensión de estrés. Para cada valor de V_E se han recogido entre 20 y 30 pulsos. Nótese que durante los pulsos de lectura, la corriente se mantiene en el piso de ruido, indicando una relajación completa. (b) Detalle de un pulso (trazo rojo) para el caso de V_E=2,5 V (trazo azul). Se indican 3 puntos, correspondientes a (1) el inicio de la tensión de estrés, (2) el momento en que la corriente a través del dispositivo comienza a aumentar y (3) el punto en el cual se estabiliza. TR se define entre los puntos (2) y (3). (c) Ídem (b) para el caso de V_E=4 V. (d) Los valores obtenidos de TR para cada uno de los valores de V_E (aprox. 20-30 puntos para cada tensión) muestran una clara dependencia con la tensión V_E.

5.3. Modelo compacto para la transición del estado de alta a baja resistividad

En base en los resultados experimentales y el mecanismo físico subyacente común a los eventos de SET en RRAM y BD, se propone un fenómeno de difusión propiciado por la transferencia de energía desde el CF hacia su entorno, teniendo en cuenta la dependencia con el espesor de la capa dieléctrica y la tensión aplicada. En este contexto, se puede expresar el TR de la transición HRS a LRS como en la Ec. 5.1

$$TR = \frac{dI_{Tr}}{dt} = \frac{qVf_1}{k_B T t_{ox}^2} D I_{SET} \tag{5.1}$$

donde t_{ox} es el espesor de la capa aislante, T es la temperatura en la discontinuidad (*gap*) del CF (ver 5.6c), k_B es la constante de Boltzmann, D es la constante de difusión de las especies atómicas responsables de la transición HRS a LRS, V es la tensión aplicada, I_{SET} es el nivel de corriente al momento del inicio de la transición HRS a LRS (evento de SET) tal como se expresa en la Ec. 5.2 y $f_1 = n_e \lambda_e \sigma_e$, con n_e siendo la densidad de electrones en el CF, λ_e el camino libre medio para los electrones y σ_e la sección transversal para la colisión entre átomos y electrones (responsables de la transferencia de momento). f_1 es aproximadamente 1 ya que la concentración de defectos en el CF es muy alta [185].

A partir de la Ec. 5.1 se puede ver que dI_{Tr}/dt es proporcional a $D \times I_{SET}$. Esto significa que la tasa de crecimiento del CF aumenta ya sea por un incremento de I_{SET} (la corriente a través del CF) o la difusividad D de las especies atómicas de los electrodos. También se pone de manifiesto la dependencia con t_{ox}, V y T. Por lo tanto, para poder utilizar la Ec. 5.1 es necesario establecer un modelo que describa I_{SET} y D en función de la tensión aplicada. Adicionalmente, el valor real de t_{ox} así como la naturaleza de las especies atómicas involucradas en la transición HRS a LRS se analiza en detalle en las secciones 5.3.1 y 5.3.2, respectivamente.

Se ha mostrado que el proceso de SET está controlado por la característica I-V del dispositivo, la cual dictamina la temperatura localizada [18], [36] que finalmente impulsa la formación y ruptura del CF a través de la capa aislante[179] (véanse las Figs. 5.6b y 5.6c, en estas se ha considerado un aislante *high–*κ simplemente a modo de ejemplo, y aplica también al caso de h-BN). Esto permite controlar la resistencia en LRS mediante la limitación de la corriente que circula por el dispositivo [179]. La corriente I_{SET} puede ser descrita mediante el modelado de la discontinuidad en el CF (véase Fig. 5.6c) como un filamento de tamaño cuántico con una constricción, de acuerdo con el formalismo de Landauer-Buttiker [178] como ha sido previamente demostrado [177]. Sin embargo, por simplicidad en esta tesis se ha optado por modelar la característica $I_{SET} - V$ del CF en HRS mediante una ecuación de interpolación (Ec. 5.2) [18], [70], minimizando así el número de parámetros requeridos. El ajuste del modelo empírico propuesto se muestra

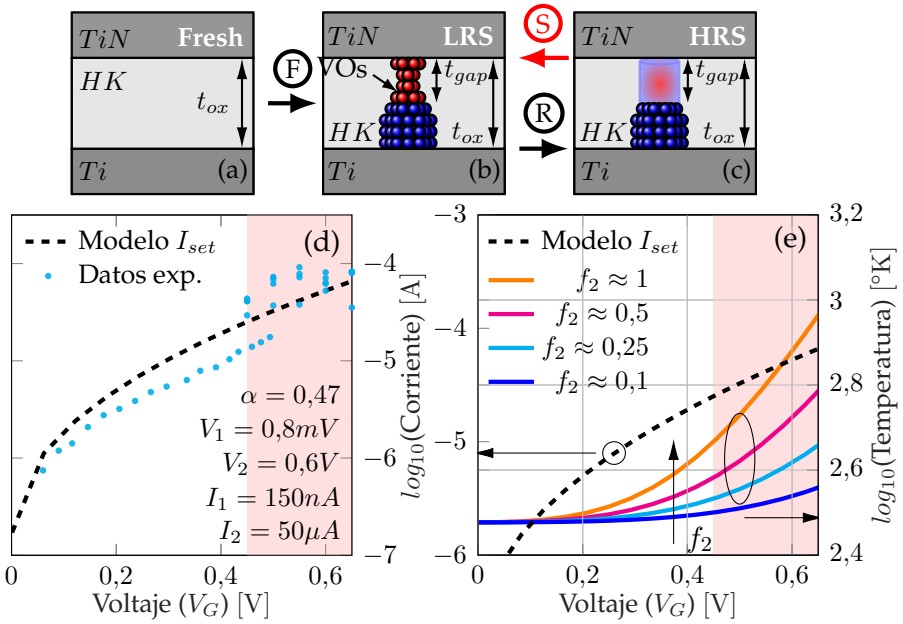


Figura 5.6: Representación esquemática de la reducción de t_{ox} debido a la aparición de una discontinuidad en el CF. (a) estructura MIM antes del electro-formado, (b) estado LRS y (c) estado HRS. Las esferas azules representan las especies atómicas de los electrodos, mientras que las rojas las vacancias de oxigeno (VOs). (d) Modelo de I_{SET} descrito en la Ec. 5.2 comparado con las mediciones experimentales, para el caso de HfO_2. (e) Cálculo de la temperatura en función de la tensión aplicada y de la energía perdida por los portadores en la constricción del CF para el caso de HfO_2. Tanto en (d) como en (e) la zona sombreada en rojo indica el rango de tensiones utilizadas en los experimentos de CVS (0.45 a 0.65 V). (f) y (g) presentan los resultados correspondientes a los dispositivos de h-BN.

en la Figura 5.6d (por brevedad, se muestran solo las curvas ajustadas para el caso de HfO_2, siendo el realizado para las muestras de h-BN un proceso idéntico). I_1,V_1 y I_2,V_2 representan dos puntos en la curva experimental $I_{SET} - V$ (véanse las curvas en HRS en la Fig. 5.2) tomados a baja y alta tensión aplicada, respectivamente, mientras que α es una constante de ajuste

$$I_{SET} = I_1 exp\left(log\left(\frac{I_2}{I_1}\right)\left(\frac{V - V_1}{V_2 - V_1}\right)^{\alpha}\right) \tag{5.2}$$

Luego, para obtener un modelo de D primero es necesario modelar la temperatura T en la constricción/discontinuidad del CF. Para ello, se ha resuelto la ecuación de transferencia de calor para dicho sistema. Para hacerlo, se ha asumido una geometría esférica simplificada alrededor de la constricción del CF y considerado que la densidad de potencia emitida por el *BD spot* sobre la superficie interna de la esfera. El resultado es el expresado por la Ec. 5.3, donde f_2 es la fracción de energía qV que cada electrón pierde al atravesar la constricción, la potencia disipada en el CF es proporcional a $I_{SET} \times V$, T_{amb} es la temperatura ambiente y k es la conductividad térmica del dieléctrico.

$$T = \frac{f_2 V I_{SET}}{2\pi t_{ox} k} + T_{amb} \tag{5.3}$$

Vale aclarar que es perfectamente razonable asumir que la disipación de potencia tiene

lugar dentro de la capa aislante. Takagi *et al.*[203] han mostrado para el caso de la ruptura dieléctrica de los óxidos de compuerta en MOSFETs, que los electrones que atraviesan dicha capa mediante efecto túnel a través de los defectos responsables del aumento de corriente inducida por estrés eléctrico (*Stress Induced Leakage Current*, SILC) pierden una energía considerable en el óxido, sugiriendo un proceso inelástico. Este comportamiento se ha explicado como una consecuencia de la relajación de los defectos [204] y fue recientemente confirmado por Lombardo *et al.*[200] para dispositivos MOS ultra-delgados.

f_2 representa la fracción de energía qV perdida por los portadores inyectados en el dieléctrico, y su valor oscila entre 0 y 1. Mientras que para tensiones aplicadas elevadas f_2 tiende a 1 reflejando la alta porción de energía perdida por los electrones, para bajas tensiones f_2 decrece, tendiendo a 0 para V=0. f_2 también depende de la temperatura, principalmente debido a la dispersión fonón-electrón [200]. Por lo tanto, f_2 es una función de la tensión aplicada y de la temperatura, cuyo comportamiento es determinado por un procedimiento de ajuste. La influencia de f_2 en la temperatura del filamento se muestra en la Figura 5.6e para diferentes valores de f_2 (nuevamente, se considera a modo de ejemplo, el caso de los dispositivos de HfO_2).

Una vez que la temperatura (T) en la constricción del CF es estimada, la difusividad D puede ser descrita mediante la ley exponencial expuesta en la Ec. 5.4, donde D_0 es un término pre-exponencial y E_{act} es la energía de activación para el proceso de difusión. Cabe mencionar que tal como ha sido reportado por Lombardo *et al.*[200], al sustituir las ecuaciones 5.4 y 5.3 en la Ec.5.1 aparece una dependencia exponencial de TR en f_2. De esta forma, se puede explicar la significativa dispersión observada en las Figs. 5.3a y 5.3b. Si por ejemplo f_2 tiene una dispersión del 20 %, el TR puede presentar un desvío estándar de hasta dos órdenes de magnitud.

$$D = D_0 e^{\frac{-E_{act}}{k_B T}} \tag{5.4}$$

5.3.1. Dependencia con el espesor del dieléctrico

Para extender el modelo descrito en la sección anterior al caso del evento de SET, se debe analizar en detalle el espesor del dieléctrico en la región donde sucede el fenómeno de RS. En este contexto debe hacerse una disquisición entre los dispositivos de RS que no requieren un paso previo de electro-formado (como en el caso de los dispositivos de conmutación volátil en base a h-BN abordados en la sección 5.2.2), y aquellos que si (como por ejemplo en las memorias no-volátiles de HfO_2 descritas en la sección 5.2.1). En el primer caso, y despreciando las posibles variaciones locales, el espesor en la región de conmutación coincide con el espesor nominal del óxido. Sin embargo, en el segundo caso el evento de RS se produce sobre el CF electro-formado (proceso asemejado a un evento de ruptura dieléctrica controlada, véase la Fig. 5.6b) a través de una capa

dieléctrica (HfO_2 en la Fig. 5.6a) prístina. En este contexto, el mecanismo de RS consiste en la creación de una discontinuidad (RESET) en el CF y la completitud de la misma (SET). El espesor (t_{gap}) de dicha discontinuidad es independiente del espesor de la capa de óxido (t_{ox}) (véase Fig. 5.6c) pero dependiente de la corriente máxima que circula durante el evento de SET [31], [180], [181]. Por lo tanto para el caso de dispositivos que requieren un paso previo de electro-formado, conocer el espesor t_{gap} es imprescindible para poder ajustar el modelo propuesto.

Como se explicara en el Capítulo 4, en una capa dieléctrica, se atribuye el evento de BD a la formación de un CF entre el ánodo y el cátodo debido a la generación de defectos con tamaños del orden de ∼1 nm [31]. También se sabe, que para el caso de SiO_2 dicho evento se puede modelar siguiendo una distribución estadística de Weibull (si bien en el capítulo 4, se han discutido las falencias de la distribución de Weibull para óxidos *high–κ*, en este capítulo se la utilizará por simplicidad). Dado que la probabilidad para la formación de tal CF depende del espesor del óxido [205], es posible usar la estadística del tiempo al SET (t_s, es decir el tiempo necesario para completar la transición entre HRS y LRS) bajo tensión constante para estimar el espesor efectivo remanente t_{gap} del film dieléctrico luego del evento de electro-formado y RESET).

La figura 5.7a muestra la estadística de t_s en escala de Weibull, donde la pendiente de las rectas de ajuste (indicada como β y constante para todos las tensiones de estrés consideradas) es ∼1.2. Esto sugiere, a partir de la comparación con la literatura [205], [206], un espesor de ∼2 nm en el dieléctrico de HfO_2 remanente en la discontinuidad. Desde el punto de vista del modelo, esto puede ser interpretado en la ecuación 5.1 como una reducción localizada del espesor del óxido (t_{ox}), como se ilustra en las figuras 5.6a y 5.6c. De la Ec. 5.3 es claro que tal reducción de t_{ox} causa un aumento significativo de la temperatura en la constricción del CF, como se indica en la Fig. 5.7b mediante las lineas de trazo y continuo para los casos de $t_{ox} \sim 10$ nm (dispositivo sin electro-formar) y

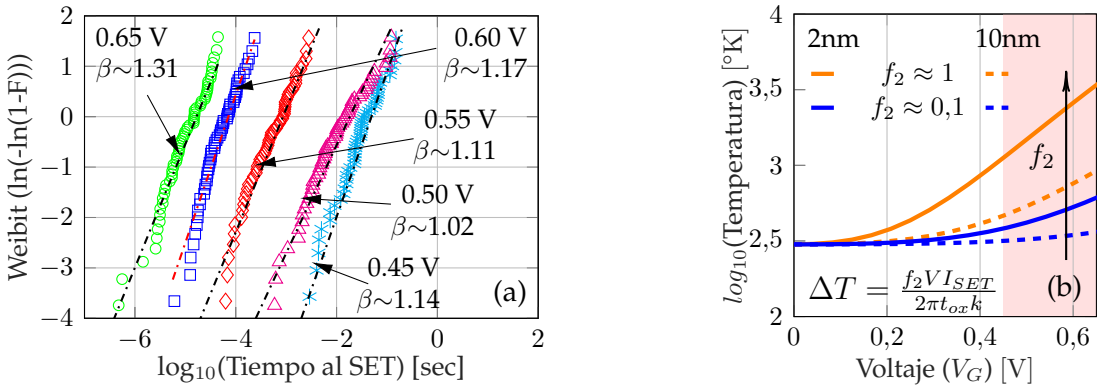


Figura 5.7: Distribución del tiempo al SET obtenida para los experimentos de CVS realizados, considerando ∼100 mediciones para cada escenario. Los símbolos indican datos experimentales y las lineas el ajuste utilizando la distribución de Weibull. (b) Temperatura en la constricción del CF vs. la tensión de estés, el t_{ox} efectivo y la pérdida de energía en la constricción (f_2). La zona sombreada en rojo indica el rango de tensiones utilizadas en los experimentos de CVS (0.45 a 0.65 V).

$t_{ox} \sim t_{gap} \sim 2$ nm. Es importante resaltar que la reducción localizada del aislante de compuerta causada por la creación de una región rica en iones metálicos en el óxido ha sido observada utilizando microscopía HRTEM (TEM de alta resolución) [184], [185], [207].

5.3.2. Rol de las especies migrantes

En el marco del modelo propuesto por Palumbo *et al.* en la Referencia [18], la creación y crecimiento del CF metálico que conecta el ánodo y cátodo (BD progresivo) son debidas a la difusión de las especies atómicas provenientes de dichos electrodos hacia el dieléctrico de compuerta. Este fenómeno ha sido reportado [18] mediante el ajuste de los datos de DR observados para estructuras MOS tanto con dieléctrico de SiO_2, *H+high-K* y h-BN, mediante la Ec. 5.1, y evaluando la difusividad D requerida en función de la temperatura con la Ec. 5.4, tal como se muestra en la figura 5.8 (véanse las curvas A, B y C). A pesar de los altos valores de difusividad (en el orden de 10^{-13} cm^2/seg. a 1000 °K, con bajas energías de activación en el rango de 0.3 a 0.7 eV), los mismos se encuentran en un rango compatible con la difusividad de iones metálicos en dieléctricos bajo condiciones similares a las aquí reportadas (por ejemplo, la difusión de Cu en *films* de SiO_2[208])

De la misma manera, las difusividades requeridas para ajustar los resultados de TR en función de la tensión aplicada fueron calculadas para los dos casos considerados (no-volátil –HfO_2– y volátil –h-BN–) con las ecuaciones 5.1-5.4. Para el primer caso (dispositivos MIM de HfO_2) se asumieron los mismos valores para las constantes de ajuste que en [18] (E_{act} en el rango de 0.3 a 0.7 eV, $f_2 \sim 0{,}1$ y $f_1 \sim 1$) y se grafica asimismo en función del recíproco de la temperatura en la Fig. 5.8a (Curva D). Los valores obtenidos de D son significativamente mayores ($\sim 10^{-6}$ cm^2/seg. a 1000 °K) que las difusividades obtenidas para el caso del BD del óxido de compuerta. En este sentido, vale la pena mencionar que incluso en el escenario poco probable de que los portadores que fluyen por CF pierdan el 100 % de su energía en la constricción (i.e.: $f_2 = 1$), la difusividad necesaria para ajustar los datos de TR sería significativamente mas alta que la difusividad obtenida al ajustar los datos de ruptura progresiva en óxidos de compuerta (Por claridad, se omiten dichos resultados). De la misma forma, para el caso de la conmutación volátil, y asumiendo valores bajos de $E_{act} \sim 0,1$eV, $f_2 \sim 0{,}05$ y $f_1 \sim 1$, la difusividad D a 100K °K es de aprox. 10^{-13}cm^2/seg., tal como se indica en la Fig. 5.8b, y en linea con [209]. Aquí es importante señalar que los datos experimentales de TR solamente pueden ser ajustados mediante el modelo de la Ec. 5.1 si se consideran valores bajos para la energía de activación, siendo estos son mucho menores que los valores requeridos para ajustar los valores de DR medidos en ciertos casos de ruptura dieléctrica observada en dispositivos MIM metal/h-BN/metal reportados en [209] (E_{act} = 1.3 eV).

En este punto, es valioso recapitular sobre como se produce la transición de HRS a LRS. Por un lado, para el caso de memorias no-volátiles que requieran un proceso pre-

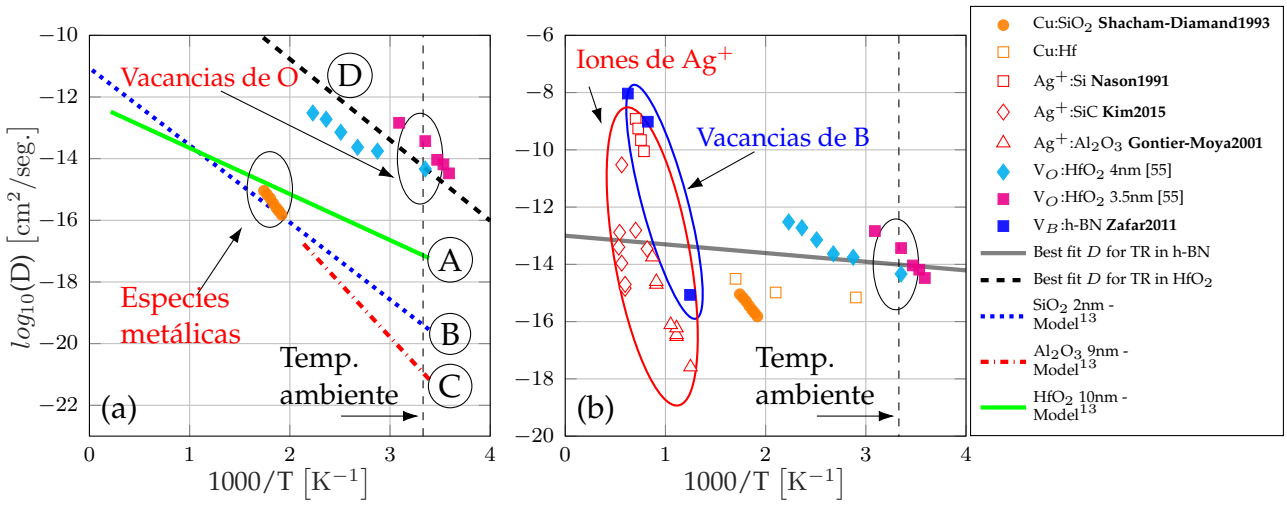


Figura 5.8: Difusividad de las especies atómicas presentes en la estructura vs. el recíproco de la temperatura. (a) La difusivdad de las VOs [210] es $\sim 10^4$ veces más alta que aquella observada en iones metálicos [18]. E_{act} está en el rango de 0.3 a 0.7 eV. Se puede apreciar que la difusividad D requerida para ajustar los datos experimentales de TR está en el mismo rango de la difusividad de las VOs. (b) Difusividad de las especies involucradas en el SET en memorias volátiles (h-BN). Nótese que el mejor ajuste obtenido para los datos presentados en este capítulo presenta una energía de activación sustancialmente menor a otros reportados en la literatura, lo cual podría explicarse por los diferentes medios considerados para la difusión.

vio de electro-formado (dispositivos de HfO_2 en este capítulo), el SET se ha explicado como el desvanecimiento de una discontinuidad en el CF. Durante la operación de *forming*, el CF se crea mediante la migración de las especies atómicas conductivas nativas de los electrodos (esferas azules) hacia el volumen del dieléctrico, y la disociación de iones de oxígeno(O^{2-}), los cuales son arrastrados hacia el electrodo superior (*Top Electrode*, TE), con lo que se generan vacancias de oxígeno cargadas positivamente (esferas rojas) en el volumen del óxido y un reservorio de iones O^{2-} en el TE (véanse las figuras 5.6a y 5.6b) [181], [211]. Acto seguido, el proceso de RESET re-abre la discontinuidad en el CF debido a la re-combinación de las vacancias de oxigeno del CF con los iones de oxígeno O^{2-} que retornan desde el TE hacia el CF por un proceso de difusión (Fig. 5.6c). En este punto, el dispositivo se encuentra en HRS, con una discontinuidad en el CF el cual consiste principalmente de especies anódicas (esferas azules) que se difundieron durante el *forming*.

Dentro de este marco, el evento de SET se explica como la unión del *gap* debido a la migración de iones de O^{2-} en dirección del TE mediante mecanismos dependientes del campo eléctrico y con activación térmica. Al desplazarse los iones de O^{2-}, estos dejan tras si vacancias de oxigeno que completan la discontinuidad del CF (véase la figura. 5.6b) [170], [211], [212]. Este punto es de suma importancia, dado que la difusividad de las vacancias de oxigeno (VOs), o equivalentemente de los iones de O^{2-} en un film de HfO_2 de espesor similar al *gap* en el CF se encuentran en un rango similar al de las difusividad requerida para ajustar los datos experimentales de TR [210] (véase la curva D en la Fig. 5.8). Por lo tanto, esto podría, en conjunción con la mencionada reducción del espesor

efectivo de t_{ox}, explicar la los altos valores de TR para estas muestras.

otro lado, en el caso del SET en memorias volátiles sin electro-formado (dispositivos de h-BN en este capítulo), se debe considerar un fenómeno físico similar pero con ciertas diferencias. Shi *et al.* han mostrado en [169] que los cambios de conductancia volátiles pueden explicarse considerando que los iones de Ag^+ provenientes del TE que se difunden hacia el interior del dieléctrico de h-BN durante el SET no logran formar uniones estables y por ende retornan al electrodo superior durante T_{Bajo} (e incluso T_L) debido a su alta difusividad [192]. Esta interpretación es consistente con el hecho de que la aplicación de PVS de polaridad opuesta no muestra conmutación volátil. En este escenario, los iones de Au^+ que se difunden desde el electrodo inferior no tienen la difusividad suficiente y permanecen en la capa aislante cuando se remueve la tensión aplicada. Es importante mencionar que la completa recuperación del estado de alta resistencia después de largas secuencias de PVS es una propiedad sobresaliente que la amplia mayoría de los dispositivos de RS no exhiben [192], [213], [214] y puede explicarse mediante la excelente estabilidad mecánica y química de la estructura de múltiples capas de h-BN [215]-[217], así como también su alta conductividad térmica [218].

5.3.3. Ajuste de los datos experimentales

Los resultados del ajuste de los datos experimentales obtenidos de TR en función de la tensión aplicada, en términos del modelo propuesto por las ecuaciones 5.1-5.4 se presentan en la figura 5.9. Considérese en primer lugar el caso de la conmutación no-volátil (HfO_2), particularmente la curva n°1 de la Fig. 5.9a. Estos resultados han sido superpuestos al gráfico de dispersión de los datos experimentales de TR (cuadrados de color cían), mostrando un alto grado de coincidencia con el valor medio obtenido para cada tensión (indicado por las esferas azules). El modelo propuesto tiene en cuenta tanto la reducción efectiva del óxido (t_{ox} se considera igual al espesor de la discontinuidad, t_{gap} ∼2 nm) y el incremento en la difusividad (D_0 en el orden de $\sim 10^{-6}$ cm^2/seg dado el cambio en las especies que completan el CF, i.e. vacancias de oxigeno). El resto de los parámetros involucrados se mantienen sin alteraciones ($E_{act} \sim$ 0.3-0.7 eV, f_2∼0.1 y f_1∼1).

Con el objetivo de clarificar el impacto que tanto la reducción de t_{ox} como el incremento de la difusividad (D_0) tienen TR, distintos ajustes fueron realizados (curvas n°2-n°4). La curva n°2 muestra a TR en función de la tensión aplicada sumiendo un valor de t_{ox} fijo e igual al espesor nominal del óxido ($t_{ox} \sim$ 10 nm), pero un incremento en la difusividad de las especies anódicas ($D_0 \sim 10^{-6}$ cm^2/sec). La curva n°3 presenta el ajuste resultante cuando solo se tiene en cuenta la reducción el t_{ox} (i.e. $t_{ox} \sim t_{gap} \sim$ 2 nm y D_0 corresponde a la difusión de iones metálicos en el óxido de compuerta $\sim 10^{-13}$ cm^2/seg). Finalmente, la curva n°4 presenta la dependencia de TR con la tensión de compuerta cuando se consideran los mismos parámetros que para el caso de la ruptura del óxido de

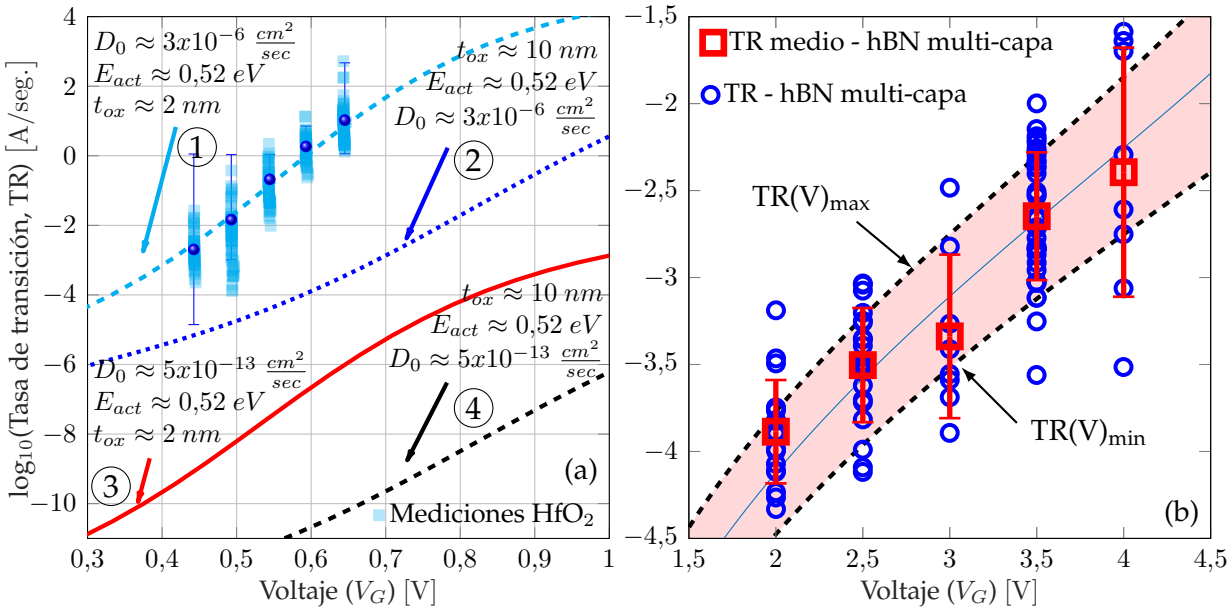


Figura 5.9: Ajuste de los datos experimentales de TR para (a) HfO_2 (los círculos indican el valor medio mientras que los cuadrados distintos valores medidos) asumiendo la difusión de VOs y un t_{ox} efectivo igual a t_{gap} (linea de trazo color cian —curva n°1—). Adicionalmente se muestra TR vs. tensión aplicada utilizando combinaciones alternativas de D y t_{ox} reportadas en la literatura —curvas n° 2, 3 y 4—. TR aumenta casi 1 orden de magnitud cada 50 mV, con un valor medio de 2x10^{-3} A/seg, 1,4x10^{-2} A/seg, 2,1x10^{-1} A/seg, 1,8x10^{0} A/seg y 1x10^{1} A/seg para tensiones aplicadas de 450 mV, 500 mV, 550 mV, 600 mV y 650 mV, respectivamente. (b) muestra los resultados correspondientes a los dispositivos de h-BN. Nótese que se observa un ajuste igualmente aceptable donde se han considerado 3 curvas I-V diferentes, resultando en un $TR(V)$ máximo, típico y mínimo.

compuerta, es decir la difusividad de iones metálicos en láminas de oxido ($D_0 \sim 10^{-13}$ cm^2/seg) y sin reducción del espesor del óxido ($t_{ox} \approx 10$ nm). Este último caso es coincidente con el DR durante el BD de óxidos de compuerta.

A partir de la comparación de las curvas n°1 y n°4 se torna evidente que el TR es significativamente más alto que el DR esperable para el rango de tensiones ensayadas. Cabe también notar que el TR calculado considerando solamente una reducción de t_{ox} o un incremento de la difusividad (D_0) no puede replicar los valores medidos experimentalmente (véase la comparación entre las curvas n°1 vs. n°3 y n°1 vs. n°2, respectivamente). Esto sugiere que el TR como función de la tensión aplicada está conjuntamente determinado por una combinación de ambos efectos, dado que ninguno de ellos puede replicar los resultados experimentales por si solo.

Con relación al ajuste de los datos obtenidos mediante los experimentos realizados sobre los dispositivos de conmutación volátil (h-BN), se puede observar un ajuste igualmente aceptable mediante el modelo de TR (véase la Fig. 5.9b). En este caso se ha probado el caso de mantener constante el espesor del dieléctrico (T_{ox} ~1,3 nm, equivalente a 4 capas de h-BN) y la difusividad de las especies migrantes ($D_0 \sim 10^{-13}$$cm^2/seg.$), pero evaluando diferentes relaciones I-V. Ante la inconsistencia de los datos reportados en la literatura para la conductividad térmica (k, los cuales varían entre 10-300 mW/°K dado que depende de si la disipación térmica se considera en el mismo plano de la capa

o a través de varias capas), el ajuste de los datos experimentales mediante el TR, sirve como una medida para estimar k, siempre que se puedan hacer ciertas consideraciones en el resto de los parámetros. Entre ellas el valor de f_2 y la energía de activación E_{act}. En el caso de la primera, ha sido mostrado por Lombardo *et al.* [200] que el valor de f_2 para un espesor de aprox. 1.3 nm y una tensión de conmutación de entre 2 y 4 V oscila entre 1×10^{-2} y 1×10^{-1}, con lo cual el valor de f_2 sugerido es una asunción lógica. Por otro lado, asumir una baja energía de activación para los iones de Ag^+ es coherente con los resultados reportados en la literatura para diversos dieléctricos (SiO_XN_Y, SiO_X, HfO_X y MgO_X [192]) donde se reportan energías de activación tan bajas como ∼0,2 eV, pudiendo ser aún más baja en los desarreglos de la red cristalina. Sobre la base de estas consideraciones, el valor de k puede estimarse en ∼100.

5.4. Conclusiones

En este capítulo, se ha mostrado que la transición (evento de SET) entre los estados de alta resistencia (*High Resistance State*, HRS) y baja resistencia (*Low Resistance State*, LRS) en estructuras MIM utilizadas como memorias de conmutación resistiva, tanto no–volátiles (en este caso de con un aislante de HfO_2 y denominadas *Memory Resistive Switching*) como volátiles (en base a hBN y denominadas *Threshold Resistive Switching*) es un fenómeno progresivo cuya dependencia con la tensión aplicada puede ser modelada de forma similar a la ruptura progresiva de óxidos de compuerta [18] si se realiza un ajuste cuidadoso de los parámetros involucrados. En este marco, el tiempo requerido para completar una transición de SET fue estudiado como una función de la tensión aplicada mediante experimentos de estrés a tensión constante (*Constant Voltage Stress*, CVS). Tal como para el caso de la ruptura progresiva, se propone que el evento de SET se debe a la transferencia de energía desde el filamento conductivo (*Conductive Filament*, CF) a su entorno circundante, lo cual promueve la electro-migración de las especias atómicas más difusivas presentes en la estructura, contribuyendo al crecimiento del CF. A pesar de que la magnitud y aceleración por tensión de la Tasa de Degradación (*Degradation Rate*, DR) durante la ruptura dieléctrica progresiva (*Progressive Breakdown*, PBD) y la tasa de transición (*Transition Rate*, TR) durante el evento de SET son diferentes, en el contexto del modelo presentado tales diferencias son explicadas en términos de la conductividad térmica del material, la naturaleza de las especies atómicas migrantes y una eventual reducción localizada en la capa de óxido. Si bien la conductividad térmica del medio aislante se considera invariante entre PBD y SET y dependiente solo del material, se debe contemplar un cambio de especies migrantes entre PBD y SET.

Si bien en el primero, se considera que existe migración de especies atómicas de los electrodos hacia el dieléctrico, en el segundo caso se debe considerar también el movimiento de vacancias (de oxígeno –VOs– en el caso de HfO_2) aunque también de iones

metálicos –Ag^+– en las muestras consideradas de hBN) como responsables del evento de conmutación. En segundo lugar, una reducción localizada del espesor del óxido puede ocurrir debido a que el evento de SET se origina en una capa de óxido previamente degradada, lo cual aumenta la velocidad de la transición. El modelo aquí presentado tiene en cuenta tales efectos y está en concordancia con resultados previamente reportados en la bibliografía [169], [170], [185], [211], los cuales presentan evidencia experimental tanto de la reducción localizada del espesor del óxido como del rol jugado por las vacancias. Finalmente, se debe resaltar que mediante el análisis aquí presentado, es posible estimar a partir de las características intrínsecas de los materiales y su geometría, los tiempos de transición (o escritura) en memorias resistivas, un dato no menor teniendo en cuenta su potencial aplicación en sistemas neuromórficos o de almacenamiento. Al mismo tiempo, permite una estimación indirecta de características intrínsecas del material, tal como la conductividad térmica.

6 Redes Neuronales

Las Redes Neuronales Profundas [38] (*Deep Neural Networks*, DNNs) han demostrado un éxito comercial significativo en los últimos años, con rendimientos que sobrepasan alternativas previas mucho más sofisticadas en el reconocimiento de voz [39] e imágenes [40]-[42] y se aproximan o incluso superan niveles humanos. Sin embargo, estas implementaciones tienen un alto costo computacional que reduce su aplicabilidad en ciertos escenarios. Esto se debe en parte a la arquitectura de Von Neumann, por lo que se han explorado múltiples alternativas considerando GPUs [43], FPGAs [44] o ASICs, con distintos grados de éxito. En este contexto, la implementación mediante *cross-bars* de memorias resistivas (RRAMs) se presenta como la más prometedora dada su capacidad para procesar grandes volúmenes de información con baja latencia, consumo de potencia y área requerida. Pero a pesar de la potencialidad de estas implementaciones, existen ciertas características intrínsecas de los dispositivos RRAM (usualmente ignoradas o incluso beneficiosas cuando se lo utiliza en aplicaciones de memoria) tales como la relación entre los estados alta y baja resistencia, la carencia de estados intermedios, la asimetría entre la escritura y el borrado, la variabilidad entre dispositivos y confiabilidad de los mismos, que se transforman en problemas mayores para su aplicación en DNNs [47]. En este capítulo se demuestra la aplicabilidad del modelo compacto de memdiodo en la simulación eléctrica realista de perceptrones mono y multicapa destinados al reconocimiento de patrones. Para ello, se tiene en cuenta el impacto de no idealidades tales como resistencia de linea, la reducción de la ventana resistiva de los dispositivos, la degradación de la relación señal a ruido y la variabilidad dispositivo a dispositivo. También se hacen contribuciones enfocadas en los aspectos de fiabilidad, proponiendo estrategias para la mitigación de fallas de enclavamiento en las memorias resistivas.

6.1. *Cross-bar arrays* de memristores en redes neuronales

Los denominados *Cross-bar Arrays* o *Cross-Point Arrays* (CPA) de RRAM o memristores (véase la Fig. 6.1a) son en este momento objeto de intensa investigación, dado que sus propiedades como memorias no-volátiles de bajo consumo tienen un enorme potencial en campos de la inteligencia artificial (*Artificial Intelligence*, AI) [219], [220] y el almacenamiento de información [221]. Más aún, el desarrollo del nuevo paradigma del Internet del Todo (*Internet of Everything*, IoE) requerirá el procesamiento de grandes cantidades de datos con muy poco consumo de energía. En este contexto, el método de Multiplicación-Vector-Matriz (MVM) usado en muchas de estas aplicaciones es muy adecuado para ser implementado mediante CPAs [222]. Adicionalmente, la estructura del CPA puede ser escalada hasta $4F^2$, siendo F la longitud nominal de un determinado nodo tecnológico de nano fabricación [223], lo cual hace posible una alta densidad de integración de unidades de memoria. Esta serie de características permiten a los CPA implementar funciones de AI tales como la clasificación de diversos patrones (sonido, imágenes, electrocardiogramas, etc.) con un menor consumo de energía que los sistemas basados en la arquitectura convencional de Von Neumann [222].

Tales aplicaciones han sido extensamente estudiadas en la literatura [219], [224]-[229] considerando distintas arquitecturas basadas en CPAs, así como diversos modelos para los dispositivos RRAM. Hu *et al.* mostraron en [219] un caso de estudio basado en simulaciones donde se demuestra la factibilidad de su aplicación en el reconocimiento de caracteres. Para ello, se utilizan dos CPA de 256×26 (i.e. 256 filas por 26 columnas, en total ∼13k dispositivos RRAM) para poder representar conexiones sinápticas positivas y negativas, y los dispositivos RRAM se modelan mediante un modelo no-lineal descrito en Verilog-A [230]. Posteriormente, y con el objetivo de reducir tanto el área como el consumo de potencia resultantes de tener dos CPA, Truong *et al.* presentaron en [224] una arquitectura involucrando un CPA de 64×26 (∼1.6k dispositivos), utilizando el mismo modelo de memristor, pero con un solo CPA. Este esquema también se ha demostrado aplicable para el reconocimiento de voz [226] aunque escalado hasta un número de ∼2.5k dispositivos RRAM.

Sin embargo, a pesar de los prometedores resultados reportados, los CPA no están libres de limitaciones prácticas tales como la resistencia de linea (R_L), la ventana resistiva de los dispositivos (R_{ON} y R_{OFF}), la degradación de la relación señal a ruido (SNR), la latencia de la inferencia, la variabilidad dispositivo a dispositivo (D2D), así como también las características conductivas inherentes de los CPA, como el denominado efecto de *sneak-path* (véase la Fig. 6.1a). Mientras que los primeros son fundamentalmente una consecuencia del incremento de R_L a medida que la longitud mínima del proceso de fabricación decrece [228], [231], la cual en combinación con una ventana resistiva reducida causa una significativa caída de tensión en las lineas de interconexión, el último punto refiere a la corriente no nula que circula a través de los dispositivos no seleccionados.

Estas no idealidades dan origen a errores tanto en el proceso de inferencia (clasificación) como de escritura de los pesos sinápticos en la red [231].

Métodos tanto de *Hardware* [232] como de *Software* [219], [224]-[229], [231], [233], [234] se han propuesto para estudiar estas problemáticas. A pesar de que los primeros permitan efectivamente mejorar el rendimiento de los sistemas, estos suelen ser por lo general costosos de desarrollar [232]. Por el contrario, las soluciones de *Software* son mucho más versátiles y permiten un estudio sistemático. Los mismos pueden ser divididos en tres grupos: En primer lugar, un gran número de autores [231], [233]-[238] ha propuesto resolver el sistema de ecuaciones diferenciales acopladas que surge de considerar la ley de Kirchoff de corrientes en cada nodo del CPA, asumiendo que cada dispositivo RRAM es un resistor de valor fijo. A pesar de haberse probado útil, este método no permite tener en cuenta la electrónica digital de control, ni considera las características de conducción no-lineal de los *memristores*, con lo que su aplicabilidad es limitada. En segundo lugar, se han propuesto métodos basados en Python [239], [240], capaces de incorporar modelos realistas de los dispositivos RRAM. Sin embargo, en estos casos se suele omitir los efectos parásitos de los CPA, o ignorar la lógica electrónica de control. Finalmente, la simulación SPICE emerge como el método más adecuado, dado que permite modelar el sistema completo (CPA con no idealidades así como la electrónica de control) [219], [224]-[229]. Este enfoque, sin embargo, está limitado por las peculiaridades de los modelos de memristor considerados y por el tamaño del CPA, dado los recursos computacionales que suelen demandar [241]. Como ejemplo de ello, el simulador NVM-SPICE [242] es capaz de simular CPAs de hasta 32×32 en un tiempo acotado, pero tiene dificultades para lidiar con variaciones abruptas en las señales de excitación o cláusulas condicionales en el modelo de memristor utilizado [243].

Dada su importancia para el desarrollo de simulaciones fiables en SPICE, se ha puesto gran atención en el perfeccionamiento del modelo compacto del memristor. Como consecuencia, distintas alternativas han sido reportadas en la literatura, existiendo entre ellas grandes discrepancias al respecto de que mecanismo de conducción y ecuación de memoria (*Memory Equation*, ME, ecuación diferencial de primer orden que vincula la corriente fluyendo por el dispositivo y la tensión aplicada, con el estado de memoria del dispositivo) se ajusta mejor al amplio rango de comportamientos memristivos [244]-[246]. Como resultado de esto, ha florecido una gran variedad de modelos tanto comportamentales como físico-fenomenológicos, posibilitando una solución de compromiso entre simplicidad de cómputo y precisión.

En líneas generales, los modelos disponibles del memristor pueden ser clasificados en tres grupos. En primer lugar, los modelos comportamentales más simples [230], [247], [248] son útiles en las etapas tempranas del diseño de circuitos, es decir, cuando se necesita una prueba de concepto a la brevedad. En segundo lugar, los modelos físico-fenomenológicos tales como los de Pickett [249] y Bayat [250] para dispositivos MIM de TiO_2 son los que permiten la mayor precisión, pero a un elevado costo computacional.

Esto último los hace poco prácticos para escenarios donde se involucra un gran número de dispositivos [251]. En último lugar pero no menos importante, los modelos fenomenológicos generalizados tales como el de Yakopcic [252], TEAM [253], VTEAM [254] o Eshraghian [255] pueden ajustar satisfactoriamente los datos experimentales de ciertos dispositivos. Sin embargo, y al igual que el segundo grupo, estos basan su funcionamiento en varias ecuaciones internas o en la introducción de funciones ventana en la ME (normalmente para modelar las transiciones de SET/RESET), lo cual supone serios problemas matemáticos que atentan contra la convergencia del modelo [251], [256]. Una alternativa prometedora a estos, es el modelo compacto propuesto por Miranda *et al.* en [257], [258], capaz de proveer alta precisión con un costo computacional reducido. Su expresión continua y derivable para la relación I-V y la naturaleza recursiva del cómputo de la variable de estado, lo hace apropiado para lidiar con señales de entrada arbitrarias (continuas y discontinuas, derivables y no derivables). Este modelo es llamado "Memdiodo Cuasi-estático" (*Quasi-static Memdiode Model*, QMM) y es una pieza central en el desarrollo de este capítulo. Cabe resaltar que hasta el momento ha sido ampliamente explorado como dispositivo aislado o en una configuración serie/anti-serie/paralelo/anti-paralelo de dos dispositivos [257]-[259], pero su aplicabilidad en tareas de reconocimiento de patrones sigue inexplorada.

Adicionalmente, la falta de madurez de la tecnología de fabricación todavía supone una limitación mayor para el sucesivo desarrollo de la computación neuromórfica basada en CPA. Diferentes tipos de fallas pueden ocurrir en estructuras CPA y las mismas pueden ser divididas en dos grupos: *soft-faults* y *hard-faults*. Las primeras pueden ser corregidas dado que la conductancia del memristor es modificable, por lo tanto representan un problema menor [260], [261]. Por el contrario, las denominadas *hard-faults* tales como las fallas de enclavamiento (*Stack-at-Fault*, SAF) suponen una amenaza cada vez mayor para el correcto funcionamiento del CPA. Una SAF en un memristor implica que la conductancia del dispositivo queda fijada a un valor alto (*Stuck-at-ON*, SA1) o bajo (*Stuck-at-OFF*, SA0). Estas fallas de enclavamiento son consecuencia tanto del proceso de fabricación como de la utilización intensiva del dispositivo, y aunque que las redes neuronales tienen una robustez intrínseca frente a variaciones en los pesos sinápticos [238], estas fallas terminan por degradar severamente la precisión en la etapa de inferencia. Dado que las propiedades de conducción del óxido metálico de un dispositivo RRAM dependen del espesor de dicha capa y el proceso de electro-formado [262], es difícil prevenir la ocurrencia de las fallas de enclavamiento [263]. Por ejemplo, un bloque RRAM de 4 Mb puede tener hasta un 10 % de dispositivos defectuosos [264].

Con el objetivo de minimizar el impacto de las fallas de enclavamiento en los CPA se han propuestos diversos métodos, tales como esquemas de redundancia [265], el re-entrenamiento de la red neuronal [228], [238] y un mapeo alternativo de los pesos sinápticos en el CPA de memristores [265], todos ellos con ventajas y desventajas. Por ejemplo, la primera opción incurre inevitablemente en un mayor costo de *hardware* y di-

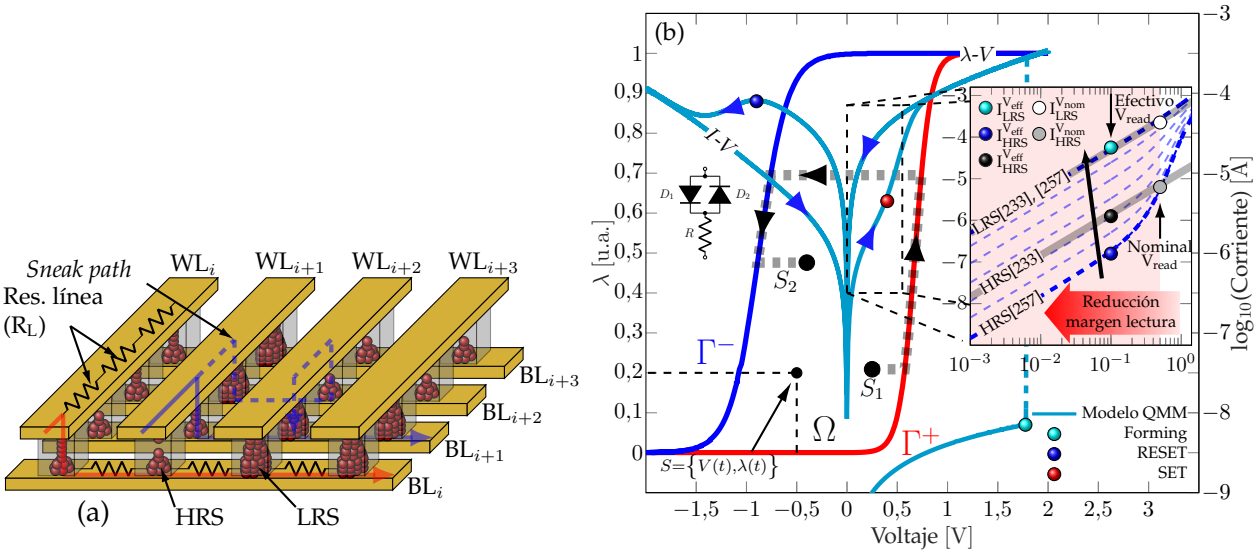


Figura 6.1: (a) Esquema de la estructura del CPA. Las flechas rojas y azules indican el flujo de corriente a través de los memdiodos que conectan las lineas superiores (*Word Lines*, WL) e inferiores (*Bit Lines*, BL). Diferentes estados de conducción son representados esquemáticamente (HRS y LRS). La linea de azul de trazo discontinuo ejemplifica el denominado problema de *sneak-path*. La resistencia serie parásita de las lineas de conexión se indica tanto para WL_i y BL_i. (b) Modelo del histerón con las funciones logísticas Γ^+ (Ec. 6.3) y Γ^- (Ec. 6.4). Ω es el espacio de estados posibles S. Las lineas negras de trazo discontinuo superpuestas al modelo del histerón indican la trayectoria de la variable de estado λ dentro de Ω desde un estado inicial S_1 hasta un estado final S_2. El *inset* izquierdo muestra la representación circuital de la ecuación de transporte (Ec. 6.1) incluyendo la resistencia serie. Las propiedades de conducción de cada diodo están determinadas por el estado de memoria del dispositivo y solo un diodo se encuentra activo en un instante t. La característica I–V típica del memdiodo obtenida mediante simulación del modelo propuesto se muestra superpuesta al histerón. La evolución de la corriente se indica mediante las flechas azules. El *inset* de la derecha muestra la transición de una característica exponencial en HRS a lineal en LRS, mediante la variación de λ. La región sombreada en rojo indica el rango de posibles tensiones aplicadas al dispositivo. Las corrientes I_{HRS} e I_{LRS} a la tensión de ajuste se indican mediante los símbolos gris y blanco, respectivamente. Nótese que puede existir una sobre-estimación de la corriente I_{HRS} cuando se considera un modelo lineal para el régimen de HRS y se utilizan tensiones mucho más bajas que la utilizada para el ajuste, como indican los símbolos cían, azul y negro.

sipación de potencia, dada la electrónica extra requerida. Esto al mismo tiempo limita su aplicación solamente a redes de pequeño tamaño. En el segundo caso, el re-entrenamiento de la red no es la opción más eficiente ya que es computacionalmente muy costoso, por no mencionar que los sucesivos ciclos de escritura pueden conducir a un incremento en el número de dispositivos RRAM con fallas de enclavamiento [229]. Por último, la utilización de algoritmos de re-mapeo es una alternativa prometedora, ya que al contrario de las anteriores, esta implica muy poco *hardware* adicional y evita el costo computacional añadido del re-entrenamiento. Ejemplos de estos algoritmos son la inversión de filas (*Row-Flip*), la permutación de filas (*Row-Permutaiton*) y la modificación del rango dinámico de las conductancias (*Value Range Transformation*) [237], [238]. Sin embargo, tales métodos han sido estudiados en escenarios idealizados y desde un punto de vista funcional y su análisis teniendo en cuenta las no-idealidades antes mencionadas, no ha sido abordado en la literatura hasta el momento.

En las siguientes secciones, se demuestra no solo que el modelo QMM puede ajustar adecuadamente los datos experimentales de I-V de un amplio rango de dispositivos memristivos, sino que también puede ser utilizado en la simulación SPICE de redes neuronales basadas en memristores y destinadas al reconocimiento de patrones, sin incrementar el costo computacional de dicha simulación. Considerando entrenamiento *ex-situ* y la clasificación de imágenes en escala de grises (entre ellas las imágenes de caracteres numéricos manuscritos de la base de datos del MNIST [266]) como métrica del rendimiento, se investigan el perceptrón mono (SLP) y multi (MLP) capa. Alrededor de estos casos de estudio, se presenta un análisis exploratorio de la dependencia de la precisión de inferencia con diversas características del CPA y el modelo QMM, tales como: $i)$ el cociente R_{ON}/R_{OFF}, $ii)$ R_L, $iii)$ el cociente R_L/R_{ON}, $iv)$ el tamaño del CPA y la resolución de las imágenes, $v)$ el esquema de particionado de la estructura del CPA y $vi)$ la variabilidad dispositivo-a-dispositivo (D2D). También se estudia $vii)$ la potencia disipada y $Viii)$ la relación señal a ruido (SNR) en función del cociente R_{ON}/R_{OFF}, R_L y el tamaño del CPA. Adicionalmente, se reporta $ix)$ una comparación entre el modelo QMM y un equivalente completamente lineal en términos de la precisión de inferencia para diferentes tensiones aplicadas y $x)$ una comparación de la complejidad computacional de la simulación eléctrica de CPAs implementados con diferentes modelos de memristores. Se evalúan también $xi)$ los aspectos estructurales de las redes, analizando el impacto del número de capas y neuronas en la inferencia, y $xii)$ el impacto de las SAFs para diferentes tamaños de CPA, valores de R_L y mapeo de pesos sinápticos. Finalmente, también se analiza $vii)$ la latencia de la inferencia en función del nodo tecnológico de fabricación, aunque por brevedad de este trabajo de tesis, este resultado se reporta por separado en las publicaciones relacionadas [267].

Debe mencionarse, que un análisis con el nivel de detalle aquí presentado, no ha sido reportado en la literatura hasta el momento. El resto del presente capítulo se organiza de la siguiente manera: La Sección 6.2 describe los fundamentos del modelo QMM: las características I-V y la ecuación de memoria. La Sección 6.3 explica el entrenamiento y simulación de las redes neuronales basadas en CPA. La Sección 6.4 discute los resultados de simulación obtenidos en términos de las variables previamente enumeradas, proveyendo las relaciones de compromiso sumamente útiles para el diseño de redes neuronales basadas en *hardware*. En la Sección 6.5 se abordan los aspectos de fiabilidad, analizando el impacto de la ruptura de los dispositivos RRAM del CPA sobre la precisión de inferencia. Finalmente, en la Sección 6.6 finalmente se presentan las conclusiones del capítulo.

6.2. Modelo Cuasi-Estático del *Memdiodo* (QMM)

Desde el punto de vista de la modelización, el modelo compacto originalmente propuesto por Miranda en [257] y luego extendido por Patterson *et al.* en [258] es capaz

de describir la relación I–V en los estados HRS y LRS, y la transición gradual entre ellos en memorias de conmutación resistiva bipolares. Esto es posible gracias a considerar una ecuación de transporte no lineal basada en dos diodos idénticos conectados en paralelo, y en serie con un resistor, como se muestra en el *inset* de la Fig. 6.1b. La relación I–V resultante se asemeja a la de un diodo con memoria, y por ello es que este dispositivo fue bautizado como memdiodo. Por completitud, el modelo QMM será brevemente resumido en los siguientes párrafos.

Físicamente, el memdiodo se asocia con una barrera de potencial que controla el flujo de electrones en el filamento conductivo (CF) de las memorias RRAM. Las propiedades de conducción de este dispositivo no lineal cambian de acuerdo con la variación de dicha barrera. Sin embargo, en la modelización se ha optado por utilizar el nivel corriente del diodo en lugar de la altura de la barrera de potencial, principalmente debido a la gran incertidumbre en el área del CF. En este contexto y siguiendo el enfoque propuesto por Leon Chua, el modelo propuesto involucra dos ecuaciones, una para el transporte de portadores (*Transport Equation*, TE) y otra para el estado de memoria del dispositivo (*Memory Equation*, ME) la cual está basada en un operador de histéresis. En relación a la primera, la característica I–V del memdiodo está dada por la expresión:

$$I = sgn(V)\left\{\frac{W\left(\alpha R I_0(\lambda) e^{\alpha\left(abs(V)+RI_0(\lambda)\right)}\right)}{\alpha R} - I_0(\lambda)\right\} \tag{6.1}$$

donde $I_0(\lambda) = I_{min}(1-\lambda) + I_{max}\lambda$ es la corriente del diodo, α es una constante de ajuste y R es una resistencia serie. La Ec. 6.1 es la solución al problema de la corriente a través de un diodo con resistencia serie y $W()$ es la función de Lambert. I_{min} e I_{max} son los valores mínimo y máximo de la corriente, respectivamente. $abs()$ indica el valor absoluto de la tensión aplicada y $sgn()$ es la función signo. A medida que I_0 aumenta en la Ec. 6.1, la curva I–V cambia su forma de exponencial a lineal, a través de un continuo de estados, como se ha observado experimentalmente en este tipo de dispositivos. λ es un parámetro de control que oscila entre 0 (HRS) y 1 (LRS) y está dado por el operador recursivo indicado en la Ec. 6.2:

$$\lambda(V) = min\left\{\Gamma^-(V), max\left[\lambda(\overleftarrow{V}), \Gamma^+(V)\right]\right\} \tag{6.2}$$

donde $min()$ y $max()$ son las funciones de mínimo y máximo, respectivamente, y $\overleftarrow{V}$ es la tensión aplicada al dispositivo en el instante previo a V. Las funciones logísticas positiva $\Gamma^+(V)$ y negativa $\Gamma^-(V)$ en la Ec. 6.2 representan las transiciones de HRS a LRS (SET) y *vice versa* (RESET), respectivamente, y físicamente pueden ser asociadas a la formación completa y disolución del CF [268], [269]. Las mismas están definidas por las Ecs. 6.3 y 6.4:

$$\Gamma^+(V) = \left\{1 + e^{-\eta^+\left(V-V^+\right)}\right\}^{-1} \tag{6.3}$$

$$\Gamma^{-}(V) = \left\{1 + e^{-\eta^{-}\left(V - V^{-}\right)}\right\}^{-1} \tag{6.4}$$

donde η^{+} y η^{-} son las tasas de transición y V^{+} y V^{-} las tensiones de umbral para el SET y RESET, respectivamente. $\lambda(V)$ define el así llamado histerón logístico o mapa de memoria del dispositivo y determina el estado de conducción del dispositivo a partir de la tensión aplicada (véase la Fig. 6.1b). λ calculado a partir de la Ec. 6.2 da cuenta de la transición entre HRS y LRS a través de un cambio en las propiedades de los diodos indicados en el *inset* de la Fig. 6.1b. La combinación de las Ecs. 6.1 y 6.2 resulta en un lazo I–V tal como el indicado en la Fig. 6.1b, el cual comienza en HRS (λ=0) y evoluciona según indican las flechas azules superpuestas al trazo I–V. El nombre Cuasi–estático proviene del hecho de que el tiempo característico de conmutación se asume infinito para cualquier estado definido dentro del histerón. No obstante, el modelo QMM puede ser transformado en un modelo dinámico fácilmente mediante la incorporación de un módulo temporal como se describe en [258].

La transición de HRS (exponencial) a LRS (Lineal) se detalla en el *inset* derecho de la Fig. 6.1b (linea azul continua) junto con algunos estados intermedios (lineas azules discontinuas). Con el propósito de comparar la capacidad de distintos modelos, se incluye un modelo lineal [231], [233], [234] (lineas grises). Nótese que los modelos basados en $sinh()$ se han omitido dado que requieren el ajuste simultaneo de múltiples parámetros para replicar la transición gradual de exponencial a lineal, o incluso expresiones independientes para cada régimen [234], [270]. A pesar de que ambos modelos coinciden para bajas tensiones donde exhiben un comportamiento claramente lineal, aparecen grandes discrepancias a medida que la tensión aumenta. Como se puede apreciar, el modelo lineal no es capaz de capturar el alejamiento de de las curvas de HRS de la característica lineal para tensiones intermedias. Como tal, de usarse para ajustar I_{LRS} e I_{HRS} a una tensión nominal V_{read}, se puede incurrir en una sobre estimación de la corriente a través del dispositivo cuando se utilicen tensiones más bajas. Por el contrario, el modelo QMM puede describir con precisión tanto las curvas I–V de HRS como de LRS con solamente cambiar un único parámetro en la ecuación de transporte. A medida que λ es barrido desde λ_{min} (por ejemplo 10^{-5}) hasta 1, I_0 en la Ec. 6.1 varía entre I_{min} e I_{max}, causando que la curva I–V cambie su forma gradualmente de exponencial (régimen HRS) a lineal (régimen LRS), como consecuencia de la caída de tensión en la resistencia serie, lo cual linealiza la ecuación de transporte.

Para demostrar la capacidad de ajuste del modelo, el mismo fue utilizado para modelizar datos experimentales extraídos de la bibliografía. En particular, la Fig. 6.2 muestra los resultados obtenidos para estructuras de HfO_2 [271], Al_2O_3 [272], MnO_3 [273], CuO_2 [274], $La_{1\text{-}x}Ca_xMnO_3$ [275] y TaO_x [269], medidas a temperatura ambiente y con barridos de tensión DC. Los datos experimentales fueron replicados con el modelo de SPICE presentado en la Tabla B.1 presente en el apéndice B.2, el cual está basado en las Ecs. 6.1 y 6.2, y considerando la aplicación de estímulos descritos en las referencias correspondientes. Los parámetros de ajuste se muestran en cada una de las sub-figuras de la figura 6.2

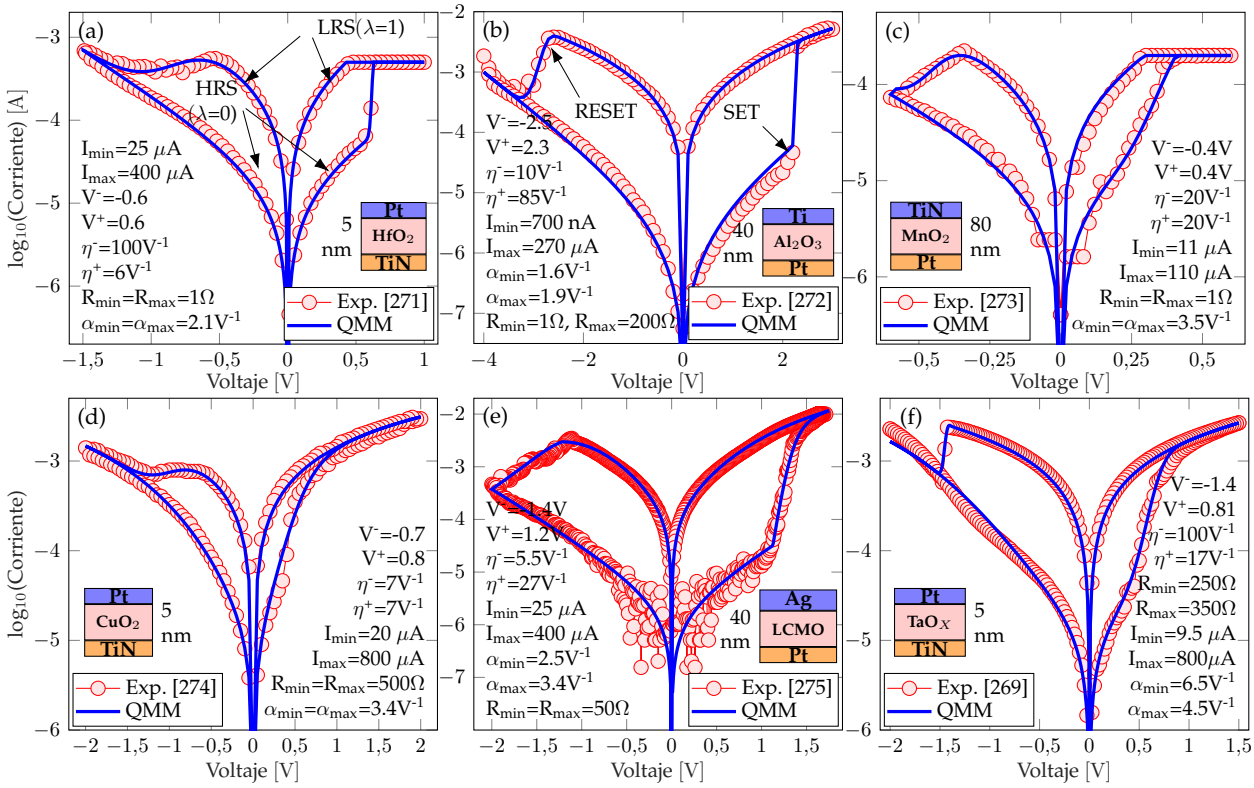


Figura 6.2: Curvas I–V experimentales para diferentes materiales reportados en la literaturas, ajustadas con el modelo QMM: (a) HfO_2 [271], (b) Al_2O_3 [272], (c) MnO_3 [273], (d) CuO_2 [274], (e) $La_{1-x}Ca_xMnO_3$ [275] y (f) TaO_X [269]. Los parámetros de ajuste del modelo QMM se muestran para cada caso. Como referencia, las curvas HRS y LRS se indican en (a) y los eventos de SET y RESET en (b). Nótese que en (a) se impuso una limitación de corriente de 200 μA para prevenir la ruptura dieléctrica permanente, la cual es adecuadamente representada por el QMM.

como referencias, así como los detalles de la estructura RRAM. Debe mencionarse que el modelo QMM propuesto no solo provee una implementación compatible con SPICE para los dispositivos de memoria resistiva, sino que también una gran versatilidad, dado que puede ajustar correctamente los lazos I–V medidos en una gran variedad de dispositivos RRAM. Nótese que mediante la apropiada selección de parámetros, el modelo QMM es capaz de modelar transiciones abruptas o suaves tanto en el evento de SET (véase el SET en las Figs. 6.2a y 6.2c) o RESET (véase el RESET en las Figs. 6.2d y 6.2f).

Por último, el modelo QMM también permite también simular la programación del dispositivo a un determinado nivel de conductancia (resistencia) mediante un procedimiento iterativo de Escritura–Verificación (*Write–Verify*), tal como se muestra esquemáticamente en la Fig. 6.3a. Suponiendo tal método, se aplican pulsos de amplitud creciente (Escritura) al dispositivo hasta que la conductancia deseada es alcanzada (Verificación) [277]. En caso de que se exceda el valor de conductancia, se aplican pulsos de la polaridad opuesta con el objetivo de reducir gradualmente la conductancia hasta obtener la conductancia deseada (dentro de un margen de error). Esta metodología de programación implica una transición como la que se muestra en la Fig. 6.1b mediante la linea roja de trazo discontinuo, donde los pulsos incrementales causan que el sistema evolucione desde un estado inicial S_1 hasta un estado final S_2 siguiendo la función logística

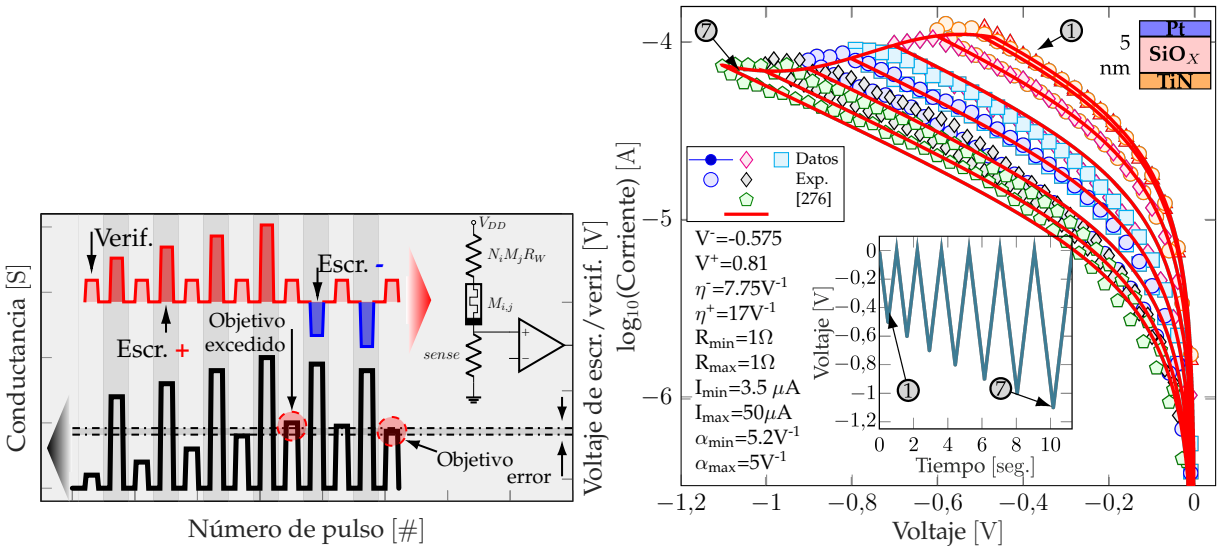


Figura 6.3: (a) Representación esquemática del proceso de Escritura–Verificación utilizado para programar los dispositivos del CPA. La forma de onda superior representa los pulsos alternados de escritura y verificación, mientras que la inferior da cuenta de los cambios de conductancia asociados. Una representación simplificada del circuito utilizado para este proceso se muestra en el *inset* de la derecha. (b) Curvas experimentales de RESET ajustadas mediante la utilización del modelo QMM para un dispositivo de SiO_X (Datos experimentales reportados por el *University College London* (UCL) en [276]). Nótese el control sobre los estados intermedios. En el *inset* se muestra la señal de tensión aplicada.

Γ^+. En el caso de exceder la conductancia deseada, el sistema continua su evolución a lo largo de la función logística Γ^- producto de la aplicación de pulsos de tensión de la polaridad apropiada. Esta última parte del procedimiento es presentada experimentalmente en la Fig. 6.3b para un dispositivo RRAM construido con un dieléctrico de SiO_x [276] y adecuadamente modelado mediante el modelo QMM. Con el objetivo de explorar los estados intermedios entre HRS y LRS, se aplican 7 rampas de tensión omitiendo el paso de verificación, tal como se muestra en el *inset* de la Fig. 6.3b

6.3. Simulación eficiente en SPICE de redes neuronales

Con el objetivo de poder evaluar sistemáticamente la aplicabilidad del modelo QMM para la simulación realista de ANNs basadas en CPAs destinadas al reconocimiento de patrones, se propone un procedimiento que contempla la creación, entrenamiento y simulación tanto de SLPs como MLPs. Las bases de datos de evaluación consideradas incluyen las denominadas MNIST [266], MINST-F [278], MNIST-K [279], CIFAR-10 [280], SVHN [281] (ambas dos re-escaladas y convertidas a escala de grises) y la *Yale Face Database* [282]. Todo este proceso se resume mediante el diagrama de flujo presentado en la Fig. 6.4a, de donde se observa que las tareas realizadas pueden dividirse en dos grupos: El primero de ellos comprende un set de sub-rutinas de MATLAB para crear, entrenar y escribir el código (*netlist*) de SPICE asociado a la ANN bajo estudio, mientras que la

segunda parte involucra la simulación SPICE del circuito propuesto durante la fase de inferencia (clasificación de patrones). Nótese que por simplicidad, en la Fig. 6.4a se considera el caso del SLP, pero el mismo proceso aplica para la creación de MLPs. Con relación al tipo de redes considerados, vale mencionar que si bien el caso del SLP o MLP [47], [277], [283] es una variante más sencilla respecto de otras implementaciones más complejas de redes neuronales con RRAM reportadas en la literatura (tales como las redes neuronales convolucionales [284], o de impulsos [285], etc., véase la Tabla 6.1) estas redes permiten estudiar las limitaciones impuestas a las ANNs por los efectos parásitos y no-idealidades propios de los CPAs de memristores, así como también estimar el costo computacional de las simulaciones implementadas mediante el modelo QMM.

Con relación a las bases de datos de evaluación utilizadas, se ha hecho hincapié en dos de ellas: La base de datos del MNIST (*Modified National Institute of Standards and Technology*) de dígitos manuscritos, y la base de datos de rostros de la Universidad de Yale *Yale Face Database*. Cada una de ellas comprende una serie de m vectores de entrada $x(m)$ y un vector de etiquetas $t(m)$ con 10 dimensiones para el caso del MNIST (una por dígito) y 38 para la base de datos de Yale (una por cada persona considerada). Cada una de las posiciones t_c de $t(m)$ será $t_c(m) = 1$ si m pertenece a la clase c o 0 en el caso contrario. Por otro lado, el brillo de cada píxel esta codificado en una escala de grises de 256 niveles entre 0 (pixel negro o apagado) y 1 (píxel blanco o encendido). Con relación a la base de datos del MNIST, esta contiene 60.000 imágenes de entrenamiento junto con

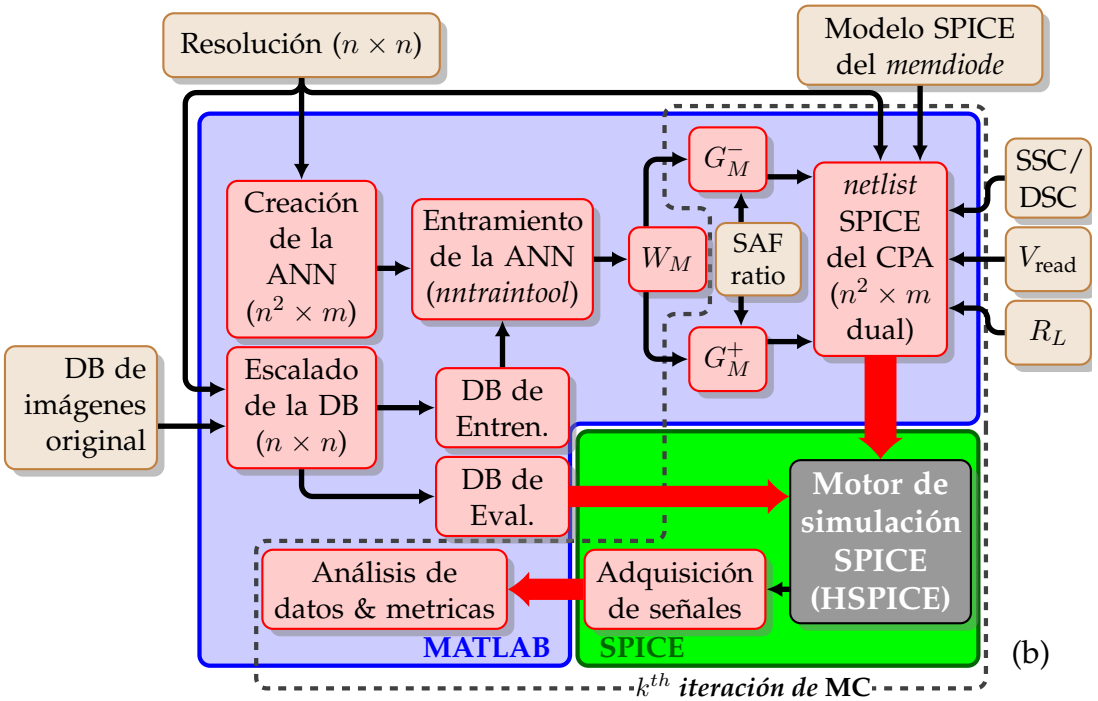


Figura 6.4: Diagrama de flujo del procedimiento de entrenamiento, modelado circuital y simulación. Partiendo del tamaño de las imágenes de la base de datos, R_L, V_{read} y el esquema de conexionado, el set de rutinas de MATLAB escala la base de datos, entrena la ANN (SLP o MLP, en el diagrama se indica SLP solo por simplicidad), la traduce a nivel circuital, agrega la electrónica de control necesaria, realiza las simulaciones y procesa los resultados. Las tareas realizadas en MATLAB están agrupadas por el recuadro azul y las operaciones de SPICE por el recuadro verde.

10.000 imágenes de prueba, todas con una resolución de 28×28 px. Algunos ejemplos de estas imágenes pueden verse en la Fig. B.1 en el apéndice B.3, donde x e y indican el número de píxel. Por su parte, la base de datos de rostros de la Universidad de Yale, contiene imágenes de 38 personas diferentes con aproximadamente 64 condiciones de iluminación diferentes y 9 poses distintas. Para los análisis propuestos en este capítulo se han considerado solamente las imágenes frontales, las cuales fueron luego centradas y recortadas a una resolución de 32×32 px. No obstante, y al igual que para el caso de las imágenes del MNIST, se ha utilizado el método de interpolación bicúbica para poder evaluar ANNs de diversos tamaños.

Tabla 6.1: Comparación entre diferentes tipos de redes neuronales implementadas con RRAM y los algoritmos de entrenamiento empleados. Nótese que en todos los casos las capas sinápticas están implementadas con CPAs y las simulaciones no modelan correctamente los efectos de resistencia de linea ni contemplan modelos realistas de RRAM. Dado que el CPA es un bloque constructivo fundamental en estas redes neuronales realizadas en hardware, la simulación realista en SPICE de los CPA es de suma importancia.

Tipo de red neuronal	**Algoritmo de entrenamiento**	**Base de datos**	**Tamaño**	**Prec. (Sim.)**	**Prec. (Exp.)**	**Platforma**	**Ref.**
Single-Layer Perceptron (SLP)	Scaled Conjugate Gradient	MNIST ($n \times n$ px.)	1 capa (n^2×10)	See paper		SPICE sim. QMM model	[267]
	Manhattan update rule	Custom-pattern	1 capa (10×3)	ND		Exp. (TaO_X / Al_2O_3)	[286]
		Yale-Face	1 capa (320×3)	∼91.7 %		Exp. (TaO_X)	[287]
Multi-Layer Perceptron (MLP)	Stochastic Gradient Descent	MNIST (8×8 px)	2 capas (64×54×10)	∼91.7 %	∼91.7 %	Exp. (HfO_2)	[283]
	Backprop.	MNIST (14×14 px)	2 capas (196×20×10)	∼92 %	∼82.3 %	Software / Exp. (HfO_2)	[277]
		MNIST (22×24 px.)	3 capas (528×250... ×125×10)	∼83 %	∼97 %	Software / Exp. (PCM)	[47]
		MNIST (28×28 px)	2 capas (784×100×10)	∼97 %		Software (Python)	[239]
	Sign-Backprop.	MNIST (28×28 px)	2 capas (784×300×10)	∼94.5 %		Software (MATLAB)	[288]
Convol. Neural Network (CNN)	Backprop.	MNIST (28×28 px)	2-layer (1^{st} Conv., 2^{nd} FC)	∼94 %		Software	[284]
Spike Neural Network (SNN)	Spike Timing Dependent Plasticity (Unsup.)	MNIST (28×28 px)	2 capas (784×300×10)	∼93.5 %		Software (C++ "Xnet")	[285]

6.3.1. Circuitos neuromórficos basados en RRAM

Con respecto al set de rutinas implementadas en MATLAB, el primer paso consiste en crear la base de datos de imágenes, re-escalando una base de datos determinada a $n \times n$ pixels. Acto seguido, se crea y entrena un SLP o MLP con n^2 entradas, m salidas y un número N de capas neuronales ocultas (cada una de ellas con m_i neuronas), utilizando la base de datos de imágenes de entrenamiento previamente re-escalada. El SLP (o MLP) es entrenado *ex-situ* (entrenamiento fuera de linea o *off-line*) considerando como algoritmo de entrenamiento el Gradiente Conjugado Escalado (*Scaled Conjugated Gradient*, SCG) [289], tal como se ha propuesto en [267], dado que el mismo supone una interesante relación de compromiso entre precisión y tiempo de entrenamiento para los distintas bases de datos de evaluación consideradas. Más aún, a pesar de que el método de Levenberg-Marquardt (ML) [290] provee la mayor precisión con el máximo costo computacional entre los métodos considerados [289]-[296], la diferencia observada en la precisión de inferencia obtenida con este método y el SGC no es estadísticamente significativa, como se muestra en los estudios de validación cruzada con 10 iteraciones presentado en la Fig. B.2 y las Tablas B.2-B.7 de los apéndices B.4.1 y B.4.2. Esto produce $N+1$ matrices de pesos sinápticos $W_{M_k} \in \mathbb{R}$ con $k \in 1, 2, \ldots, N+1$ (por ejemplo, para dos capas neuronales ocultas, con m_1 y m_2 neuronas por capa, se obtienen tres matrices W_{M_1} , W_{M_2} y W_{M_3}, con tamaños $n^2 \times m_1$, $m_1 \times m_2$ y $m^2 \times 10$, respectivamente). Para poder representar tanto los elementos negativos como positivos de W_{M_k} con las conductancias siempre positivas de los memristores, cada peso sináptico es implementado utilizando dos memdiodos tal como se sugiere en [286], [297], resultando en 2 CPA por cada capa sináptica. Por lo tanto, cada matriz W_{M_K} es dividida en dos matrices $W^+_{M_k}$ y $W^-_{M_k}$ como:

$$w^+_{M_{i,j}} \begin{cases} w_{M_{i,j}}, & w_{M_{i,j}} > 0 \\ 0, & w_{M_{i,j}} \leq 0 \end{cases} \tag{6.5}$$

$$w^-_{M_{i,j}} \begin{cases} 0, & w_{M_{i,j}} \geq 0 \\ -w_{M_{i,j}}, & w_{M_{i,j}} < 0 \end{cases} \tag{6.6}$$

cada una de ellas conteniendo solamente pesos positivos, de tal forma que $W_M = W^+_M - W^-_M$. En el siguiente paso, las matrices de conductancia G^+_M y G^-_M a ser mapeadas a los CPAs se calculan mediante la transformación lineal propuesta en [238], [298]:

$$G^{+,-}_M = \frac{G_{max} - G_{min}}{max\{W_M\} - min\{W_M\}} W^{+,-}_M + \left[G_{max} - \frac{(G_{max} - G_{min})\ max\{W_M\}}{max\{W_M\} - min\{W_M\}} \right] \tag{6.7}$$

donde $[G_{min}, G_{max}]$ es el rango de conductancia para la multiplicación vector matriz. Por simplicidad, se ha considerado $G_{max} = G_{LRS} = 1/R_{ON}$ and $G_{min} = G_{HRS} = 1/R_{OFF}$, donde $max\{W_M\}$ y $min\{W_M\}$ son los pesos máximo y mínimo en la matriz W_M obtenida

por *software*. De esta forma, los pesos sinápticos en las matrices W_M^+ y W_M^- son convertidos a una conductancia en el rango [G_{HRS},G_{LRS}].

Las sub-rutinas siguientes generan el *netlist* del circuito del MLP (SLP) $n^2 \times m_i$, $m_i \times m_{i+1}, \ldots, m_N \times m$ utilizando para ello dos CPAs para cada capa sináptica y añadiendo la resistencia parásita de interconexión (R_L), el esquema de conexionado y la lógica de control necesaria para realizar la clasificación de patrones. Como se ha reportado en la Ref. [267], la utilización de un único CPA para implementar matrices de grandes dimensiones es altamente ineficiente. Dado que tanto R_L y el cociente R_{ON}/R_{OFF} están fijados por el nodo de fabricación y el mecanismo de RS respectivamente, una alternativa de diseño ampliamente aceptada [299], [300] consiste en dividir dichas matrices en particiones más pequeñas, cuyo tamaño reducido maximiza la porción de tensión efectivamente entregada a los dispositivos memristivos. La Figura 6.5a muestra una representación simplificada del circuito correspondiente al MLP implementado con CPAs particionados y las interconexiones necesarias para realizar la Multiplicación Vector-Matriz completa en la primera capa sináptica. Explotando la integrabilidad de los CPA de memristores con la fabricación CMOS convencional, es posible emplazar las conexiones y electrónica analógica de censado debajo de la estructura CPA particionada, manteniendo casi inalterada el área ocupada.

Cada memdiodo en los CPAs particionados es programado al valor de conductancia correspondiente de las matrices $G_{M_k}^+$ y $G_{M_k}^-$ mediante el ajuste del parámetro de control λ. El valor requerido de λ se obtiene resolviendo la Ec. 6.1 $I = g_{k_{1,j}}^{+(-)} V$, siendo $g_{k_{1,j}}^{+(-)}$ cada uno de los elementos en $G_{M_k}^+(G_{M_k}^-)$. Dado que en este capítulo se hace foco en el modelado de las sinapsis artificiales utilizando el modelo QMM, las neuronas ocultas en la $k - esima$ capa neuronal oculta conectando las capas de sinapsis adyacentes $k-1$ y $k+1$, se implementan mediante un modelo comportamental de SPICE. En este contexto, el modelo de cada neurona involucra un amplificador de Trans-Impedancia (*Trans-Impedance Amplifier*, TIA) que traduce la corriente de salida en la *bitline i* correspondiente a la capa sináptica $k-1$ a una tensión que es proporcionada a una función de activación que luego la propaga a la *wordline j* a la capa siguiente ($k+1$). En este capítulo se considera una función de activación log-sigmoidal $(1/(1+e^{-x}))$ aunque la metodología es extensible a otras funciones de activación.

La tensión asignada a cada sinapsis es aplicada siguiendo una esquema de conexionado dual, tal como se muestra en la Figura 6.5a. A pesar de una mayor complejidad de los circuitos periféricos, este esquema mejora la tensión entregada a cada sinapsis al conectar ambos terminales de las *wordlines* a la señal de entrada [301]. El patrón de señales de entrada a la 1^{er} capa sináptica se obtiene al re-organizar cada una de las imágenes reescaladas a $n \times n$ de la base de datos de evaluación en un vector de $n^2 \times 1$ y escalando el valor (adimensional) de cada pixel por una tensión V_{read}. V_{read} se elige de forma tal de evitar que su aplicación altere el estado de conducción de los memdiodos durante el proceso de inferencia. De esta forma, durante el dicha fase, cada una de las imágenes de

evaluación es presentada al MLP (o SLP) como un vector de tensiones analógicas en el rango [0, V_{read}]. Una vez que el *netlist* del circuito es generado, es transferido al simulador SPICE el cual evalúa las tensiones y corrientes en el MLP (o SLP) mientras este clasifica las imágenes de entrada y retorna a MATLAB las formas de onda para la extracción de métricas.

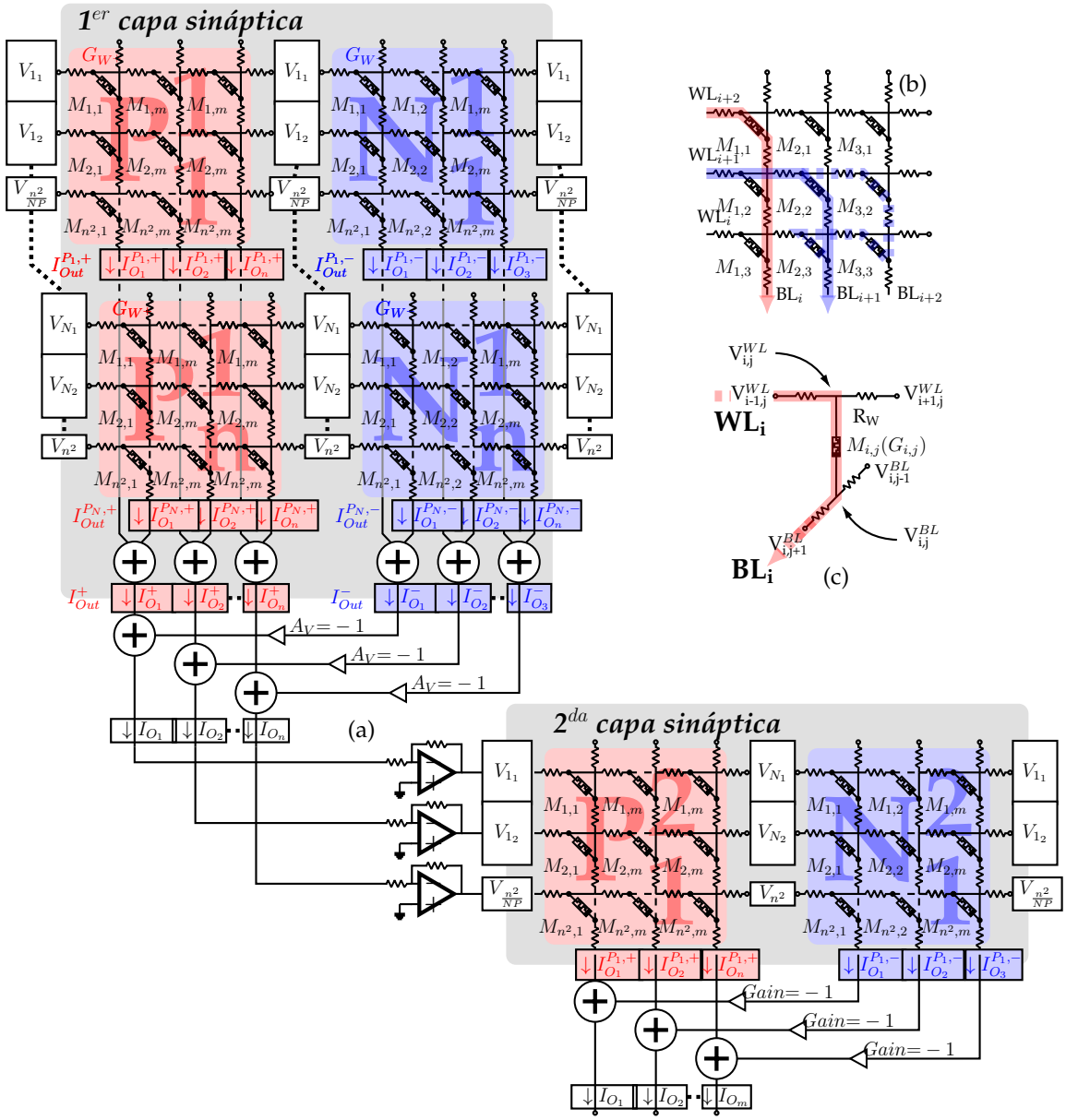


Figura 6.5: (a) Circuito equivalente simplificado de un MLP. Cada uno de los dos CPA (positivo y negativo) de la 1^{er} capa sináptica está dividido en N particiones idénticas para minimizar las caídas de tensión en las resistencias parásitas de las lineas de interconexión. En la salida de cada partición, se indica el resultado parcial de la multiplicación vector-matriz efectuada en cada bloque. (b) Circuito esquemático equivalente de un CPA. Las flechas rojas y azules ejemplifican el flujo de corriente a través de los memristores conectando *wordlines* y *bitlines*. (c) Celda RRAM individual con la correspondiente resistencia R_L.

6.3.2. Complejidad computacional

Durante la fase de inferencia, la realización en *hardware* de la multiplicación Vector-Matriz mediante CPAs permite reducir la complejidad temporal de dicha operación a $O(1)$ (definida en la notación denominada *big-O* [304]), independientemente del tamaño del vector de entrada, la matriz de pesos sinápticos o el algoritmo de entrenamiento utilizado. Sin embargo, el estudio de la complejidad computacional (esto es, la complejidad tanto temporal –tiempo o número de ciclos necesarios para completar el algoritmo en función del tamaño de la entrada a procesar– como espacial –memoria consumida por el algoritmo—) es una métrica interesante para comprar el rendimiento de distintos modelos RRAM utilizados en la simulación eléctrica de CPAs. Para ello, se ha

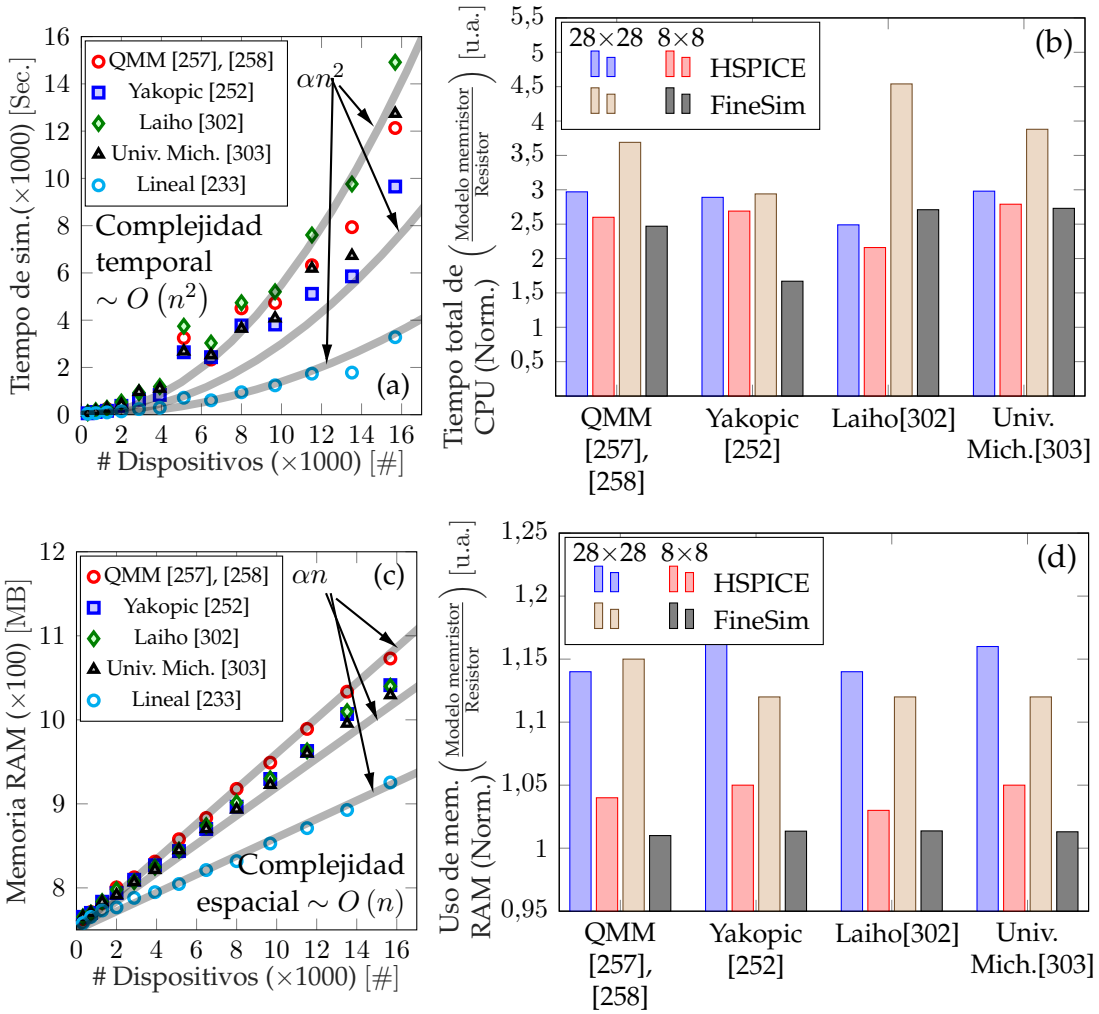


Figura 6.6: Costo computacional (tiempo de simulación y uso de memoria RAM) de la simulación de *crossbars* realizados en base al modelo QMM y comparado con los modelos propuestos por Yakopcic [252], Laiho [302] y la universidad de Michigan [303]. (a) Tiempo de simulación y (b) Uso total de RRAM en función del tamaño del CPA (medido en términos del número de dispositivos). El tiempo de simulación y el uso de memoria RAM se reportan también normalizados respecto del caso completamente lineal en (c) y (d) respectivamente, indicando que la simulación mediante el modelo QMM permite modelar con gran precisión las características I–V sin causar un aumento de la complejidad computacional.

optado por medir empíricamente el tiempo de ejecución y utilización de memoria RAM durante la simulación de circuitos CPA conteniendo entre 320 y 15680 dispositivos, ya que la estimación analítica de la complejidad temporal no es posible. Para minimizar posibles errores inducidos por el *hardware* utilizado, para cada circuito se realizaron múltiples simulaciones, reportándose el valor medio como resultado. Esto se repitió para 5 modelos diferentes utilizados para representar a los dispositivos RRAM: El modelo QMM utilizado en esta tesis, y los modelos propuestos por Yakopcic [252], Laiho-Biolek [302], la Universidad de Michigan [303] y un modelo completamente lineal [233], (un resistor de valor fijo y equivalente al peso sináptico a representar).

En primer lugar, el tiempo total de simulación en función del tamaño de CPA se presenta en la Fig. 6.6a. Puede observarse que sin importar el modelo RRAM utilizado, el tiempo de simulación demandado aumenta siguiendo una relación cuadrática con el número de dispositivos, lo que sugiere una complejidad temporal del tipo $O(n^2)$. El análisis es luego complementado con la comparativa entre distintas herramientas de simulación (H–SPICE y FineSim, este último un simulador del tipo FastSPICE) para dos tamaños diferentes (1280 y 15680 dispositivos), la cual se presenta en la Fig. 6.6b. Dado que la mínima complejidad temporal posible se obtiene utilizando un modelo completamente linea, los resultados se reportan normalizados respecto del caso lineal. De aquí se observa una mayor sensibilidad del tiempo de CPU con el tamaño del CPA al utilizar la herramienta de FastSPICE. En segundo lugar, se realiza un análisis similar de la memoria utilizada a fin de estimar la complejidad espacial. Como se puede ver en la Fig. 6.6c la utilización de memoria RAM durante la simulación crece linealmente con el número de dispositivos en el CPA independientemente del modelo RRAM utilizado, lo que sugiere una complejidad espacial del tipo $O(n)$ para todos ellos. Interesantemente, al comparar la memoria RAM utilizada para los casos de 1280 y 15680 dispositivos con los simuladores antes mencionados (véase la Fig. 6.6d), se observa una dependencia idéntica sin importar la herramienta utilizada. De la comparativa presentada entre el QMM y otros modelos reportados en la bibliografía en términos del tiempo de simulación y uso de memoria RAM normalizadas, se concluye que el modelo QMM permite una replicar las características I–V de los dispositivos RRAM (tal como se indica en la Sección 6.2) sin incurrir en un alto costo computacional (la complejidad temporal y espacial son similares).

6.4. Impacto de los elementos circuitales parásitos

En base al flujo de entrenamiento y simulación descrito en la Sección 6.3 se propone un análisis exploratorio del espacio de diseño con el objetivo de analizar el impacto de tres variables diferentes sobre la precisión de la inferencia. Estás son: *i*) la ventana resistiva de los memristores (R_{ON} y R_{OFF}) [47], [232], [277], [283], [305], *ii*) la resistencia de las lineas de interconexión (R_L) [233], [306] y *iii*) la resolución de las imágenes de las bases

de datos ($n \times n$ pixels) [47], [277], [283]. Adicionalmente también se discute el efecto de *iv*) la variabilidad entre dispositivos y de *iv*) la relación señal a ruido (SNR).

6.4.1. Perceptrón Mono-Capa (*Single Layer Perceptron*, SLP)

6.4.1.1. Dependencia con la relación R_{ON}/R_{OFF}

Para poder operar con un bajo consumo de potencia, el SLP basado en CPAs de memristores requiere de dispositivos RRAM con corrientes I_{HRS} y I_{LRS} reducidas, o equivalentemente, grandes valores de resistencia R_{OFF} y R_{ON}, respectivamente. Esto se puede lograr minimizando el tamaño del CF [36], al costo de incrementar la variabilidad de la resistencia del dispositivo, especialmente R_{OFF} [36], [277]. Esta relación de compromiso entre variabilidad y baja corriente de operación puede ser parcialmente resuelta mediante la utilización de celdas de memoria con grandes ventanas resistivas (es decir, la relación entre R_{ON} y R_{OFF}). Bajo esta premisa, se han estudiado diferentes materiales y mecanismos de RS (RRAM, CBRAM, PCM) con distintos valores de R_{ON} y R_{OFF} (véase la Tabla 6.2) [47], [232], [277], [283], [305]. Con el objetivo de estudiar como la ventana resistiva del memristor afecta la precisión de la clasificación, 12 ajustes diferentes del modelo QMM (descrito en la Sec. 6.2) han sido considerados. Estos se pueden dividir en 3 grupos y han sido definidos mediante *i*) el escalado conjunto de las curvas de HRS y LRS por un factor de 10: A1 (R_{OFF} ∼1 MΩ y R_{ON} ∼100 kΩ), A2 (∼100 kΩ y ∼10 kΩ), A3 (∼10 kΩ y ∼1 kΩ), y A4 (∼1 kΩ y ∼100 Ω), *ii*) escalando la curva de HRS por un factor de 10 mientras se mantiene fija la de LRS: B1 (∼1 MΩ y ∼100 Ω), B2 (∼100 kΩ y ∼100 Ω), B3 (∼10 kΩ y ∼100 Ω), y B4 (∼1 kΩ y ∼100 Ω) y *iii*) escalando la curva de LRS por un factor de 10 mientras que se mantiene fija la curva de HRS: C1 (∼1 MΩ y 100 ∼kΩ), C2 (∼1 MΩ y ∼10 kΩ), C3 (∼1 MΩ y ∼1 kΩ), y C4 (∼1 MΩ y ∼100 Ω). Las curvas I–V correspondientes se muestran en las Figs. 6.7a-6.7c. La simulación eléctrica del proceso de clasificación fue repetida para cada ajuste del modelo QMM ($A1$-$A4$, $B1$-$B4$ y $C1$-$C4$) considerando DSC, R_L=0.1-10 Ω, V_{read}=300 mV, con un ratio de dispositivos en SAF igual a 0 y asumiendo imágenes en la máxima resolución disponible (28×28 px., i.e. por lo que el SLP resultante contiene 15680 sinapsis). Por brevedad, se muestran solo los resultados obtenidos para la base de datos del MNIST.

Tabla 6.2: Rangos de conductancia utilizados en la bibliografía

Ref.	**Dispositivo RRAM**	R_{OFF} $(1/G_{min})$	R_{ON} $(1/G_{max})$	*ratio*	Ajuste
[277]	TiN/Ti/HfAlO/ TiN RRAM	∼1 MΩ	∼5 kΩ	200	∼*C2*
[232]	Ta/HfO_2/Pt RRAM	∼100 kΩ	∼2.5 kΩ	40	∼*A2*
[47]	GST-PCM	∼1 MΩ	∼50 kΩ	20	∼*A1*
[305]	Ta/Al_2O_3/ZrTe CBRAM	∼1 MΩ	∼5 kΩ	200	∼*C2*
[283]	Ta/HfO_2/Pd RRAM	∼10 kΩ	∼1 kΩ	10	∼*A3*

La precisión de la inferencia obtenida se presenta en las Figs. 6.7e y 6.7f en función de la resistencia R_{ON} y R_{OFF} de cada ajuste del modelo QMM, respectivamente. La figura 6.7e muestra una clara pérdida de precisión a medida que R_{OFF} decrece (la ventana resistiva se mueve hacia arriba) para los ajustes $A1$ A $A4$, con la precisión obtenida para el ajuste $A1$ cerca de los resultados obtenidos para un SLP implementado en *software* (90.9 %) [307]. Por el contrario, no se observa una dependencia clara entre la precisión de inferencia y R_{OFF} cuando se consideran los ajustes $B1$-$B4$ (la ventana resistiva se amplía para un R_{ON} fijo). Por último, la precisión de inferencia se evalúa en función de R_{ON} y

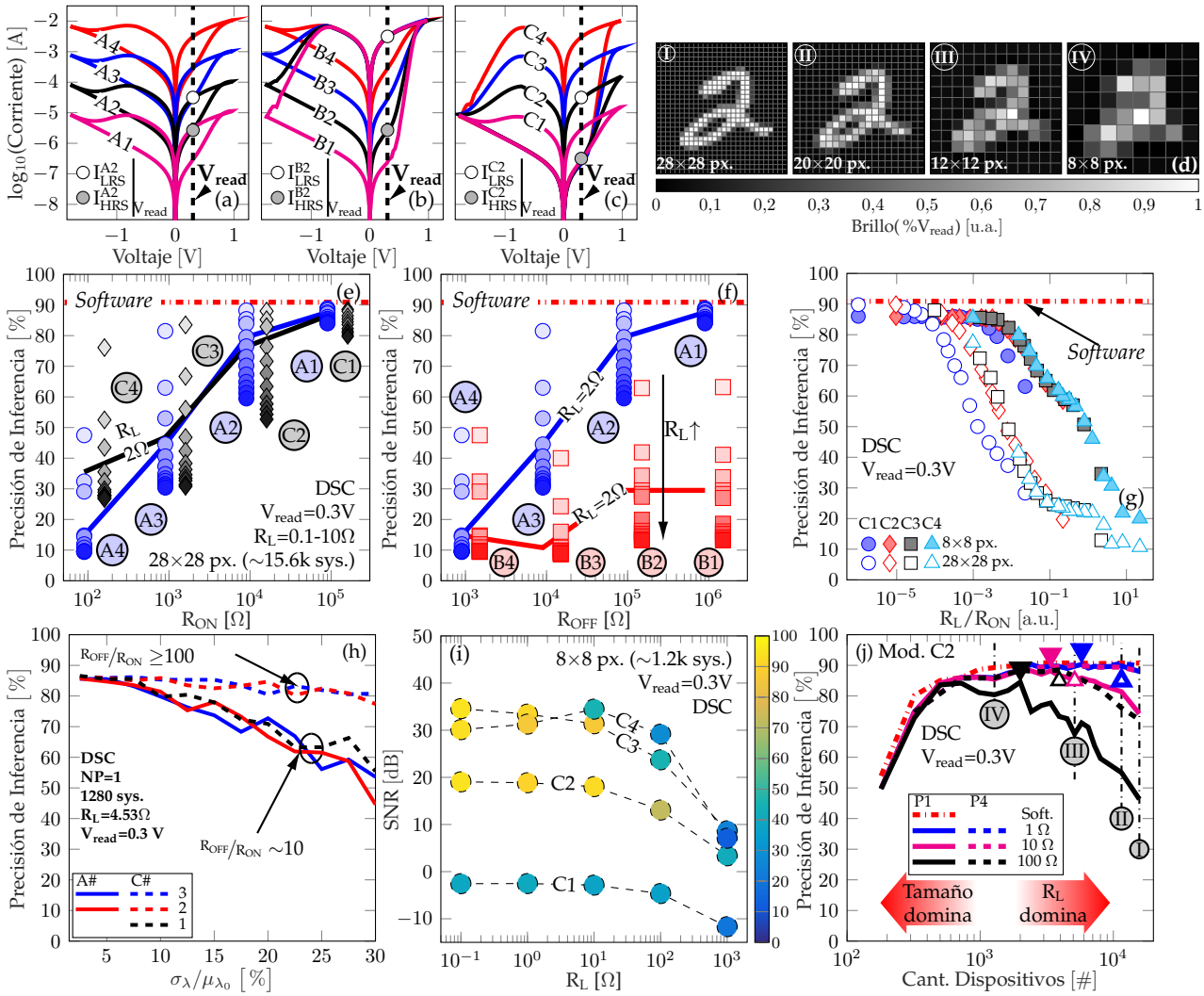


Figura 6.7: Ajustes (a) $A1$–$A4$, (b) $B1$–$B4$ y (c) $C1$–$C4$ del modelo QMM. (d) Pérdida de legibilidad de las imágenes de la base de datos del MNIST al ser re-escaladas. Precisión de inferencia en función de la característica (e) R_{ON} y (f) R_{OFF} del ajuste del modelo QMM. Se observa un mayor impacto de R_{ON} en la pérdida de precisión. (g) El cociente R_L/R_{ON} pone de manifiesto la dependencia de la precisión de inferencia con la resistencia de linea R_L, y como esta empeora con el tamaño o resolución de la imagen. Asimismo se puede observar que una relación R_{OFF}/R_{ON} mayor a 100 es necesaria para minimizar la sensibilidad a las variaciones del tipo dispositivo a dispositivo. (i) No obstante, deben evitarse resistencias R_{OFF} o R_{ON} de alto valor ya que las mismas implican corrientes sumamente bajas que comprometen el SNR y por consiguiente la precisión. Por último, se muestra la dependencia de la precisión con el tamaño de la imagen, de donde se ve como la relación de compromiso entre legibilidad y caída parásita de tensión, resulta en un tamaño óptimo de CPA

compara frente al rendimiento de los ajustes $A1$-$A4$ en la Fig. 6.7f. Tanto los ajustes $A1$-$A4$ como $C1$-$C4$ presentan una dependencia casi idéntica con R_{ON} a pesar de las significativas diferencias en las ventanas resistivas.

En su conjunto, estas observaciones indican que a pesar de que se esperaría que el ensanchamiento de la ventana resistiva mejore la precisión de inferencia, los resultados de simulación muestran que esta métrica está altamente limitada por el valor de la resistencia R_{ON}, dada la gran similitud entre los resultados obtenidos para los ajustes $A1$-$A4$ y $C1$-$C4$. Esto se puede explicar si se considera que la caída de tensión en las lineas de interconexión aumenta significativamente con la reducción de R_{ON}. Como resultado, cada uno de los dispositivos RRAM del CPA "ve" una tensión de de lectura menor a la aplicada a las entradas del CPA. La relación entre estas tensiones se denomina margen de lectura y se define como $V_{celda}/V_{lectura}$) [306]. Esto está concordancia con resultados previamente reportados en la bibliografía por Liang Liang *et al.* [306], donde se muestra que el margen de lectura está gobernado principalmente por R_{ON} con una muy leve dependencia con la ventana resistiva. Dado que la caída de tensión en las líneas de interconexión está determinada tanto por la resistencia de los dispositivos RRAM y la resistencia de linea, la dependencia de la precisión de inferencia con el valor de la resistencia de línea también debe ser analizada.

6.4.1.2. Dependencia con la resistencia de linea

En un escenario real, las líneas metálicas de conexión que forman las WL y BL del CPA están caracterizadas por una resistencia parásita (R_L) que degrada severamente los márgenes de lectura del CPA. R_L, definida como la resistencia entre dos celdas de memoria adyacentes, puede calcularse como $R_L = \rho L/(WT)$, donde L, W y T son el largo, ancho y espesor de la linea de interconexión entre dos celdas adyacentes, y se toman igual a la longitud nominal del proceso (F). R_L oscila entre 1 y 10 Ω cuando se asume $d > 10$ nm, dado que la resistividad convencional de las lineas de metal (ρ) varía entre 10^{-8} y $10^{-7}\Omega$·m. Por lo tanto, para una estructura *crossbar*, R_L puede ser estimada en $\sim$4.53, 2.97, y 1.55 Ω para los nodos tecnológicos de 16, 22 and 32 nm, respectivamente [233]. Sin embargo, para el nodo de 10 nm y posteriores, el efecto de dispersión en la superficie y en los bordes de grano causa un aumento significativo de la resistividad [308]-[310] dado que el camino libre medio de los electrones se vuelve comparable a las dimensiones de los conductores. Estos dos efectos han sido ampliamente investigados y pueden ser cuantificados por medio de los modelos de Fuchs-Sondheimer (FS) [311] y Mayadas-Shatzkes (MS) [312], revelando que para nodos altamente escalados ($\sim$5 nm) R_L puede ser tan grande como $\sim$100 kΩ [306]. El incremento de ρ puede ser descrito mediante la Ec. 6.8:

$$\frac{\rho}{\rho_{Cu}} = \frac{3}{4}(1-p)\frac{l_0}{w} + 3\left[\frac{1}{3} - \frac{\alpha}{2} + \alpha^2 - \alpha^3 ln\left(1+\frac{1}{\alpha}\right)\right]^{-1}, \alpha = \frac{l_0}{d}\frac{R}{1-R} \tag{6.8}$$

donde ρ_{Cu} es la resistividad del Cu (1.9 $\mu\Omega$·cm), p es la porción de dispersión especular, l_0 es el camino libre medio en Cu (39 nm a temperatura ambiente), W es el ancho de los conductores, R e la probabilidad de los electrones de reflejarse en los bordes de grano y d es el tamaño promedio de los granos. $p = 0,25$ y $R = 0{,}3$ se toman en base a los valores reportados en al literatura y d se asumen igual al ancho de las lineas de interconexión [308], [310]. Por último, se toma una relación de aspecto igual a 1 (W igual a la longitud nominal del proceso) y se considera una reducción de 2 nm a cada lado del conductor para representar en ancho de la barrera [306].

Como se ha mencionado en la sub-sección 6.4.1.1, la precisión de inferencia del SLP implementado mediante CPAs se ve severamente afectada por la degradación de los margenes de lectura del CPA. Dado que cada memristor se encuentra conectado en serie con un número de resistores de interconexión de valor R_L (siendo estos impuestos por las características de las *wordlines* y *bitlines*), se puede demostrar que el margen de lectura es proporcional al cociente R_L/R_{ON}. Por lo tanto, la precisión de inferencia es estudiada en función del cociente R_L/R_{ON} en la Fig. 6.7h para dos tamaños de imagen diferentes: 28×28 px. (SLPs de 784×10) y 8×8 px. (SLPs de 64×10). Interesantemente, los datos obtenidos considerando diferentes ajustes del modelo QMM ($C1$-$C4$) y diversos valores de R_L exhiben una tendencia unificada para cada uno de los tamaños de imagen considerados. En el caso de las imágenes de 28×28 px., para valores de R_L/R_{ON} inferiores a 10^{-4} no existe influencia de la resistencia de linea, dado que esta resulta despreciable frente al valor de la resistencia en LRS (R_{ON}) de los memristores, y por lo tanto casi la totalidad de la tensión aplicada en los terminales de las *wordlines* del CPA es aplicada a los dispositivos RRAM. Por el contrario, cuando el valor del cociente R_L/R_{ON} sobrepasa el umbral de 10^{-1}, la caída de tensión en las lineas de interconexión domina la distribución de voltajes en el CPA, resultando en en errores significativos de reconocimiento. Para valores entre estos dos extremos, hay un aumento gradual de la fracción de tensión que se pierde en las resistencias de linea, y por lo tanto, una reducción sostenida de la precisión de inferencia.

Al considerar las imágenes más pequeñas (de 8×8 px.) se puede observar un comportamiento muy similar pero desplazado hacia la derecha. Esto se puede explicar al considerar que en un *crossbar* más pequeño, habrá una menor degradación de los márgenes de lectura para el mismo valor del cociente R_L/R_{ON}. En conclusión, la resistencia de linea tolerable depende fuertemente del valor de la resistencia en LRS de los memristores utilizados. Dado que las proyecciones indican un aumento significativo de la resistencia de linea para los nodos de fabricación más avanzados, se requerirán importantes esfuerzos desde el área de ingeniería de materiales para desarrollar dispositivos RRAM con altas resistencias de HRS y LRS. Esto sin embargo, trae aparejados una mayor sensibilidad a problemas de variabilidad y ruido, los cuales se discutirán en las próximas secciones. -

6.4.1.3. Variabilidad Dispositivo a Dispositivo (D2D) y relación Señal a Ruido (SNR)

El mecanismo de conmutación resistiva ha sido mostrado tanto para materiales amorfos como policristalinos. En ambos casos, la posibilidad de depositar el medio dieléctrico a baja temperatura permite la fabricación de sistemas con múltiples capas sin producir perturbaciones en la electrónica CMOS ubicada debajo. Sin embargo, la poca controlabilidad sobre la densidad de defectos en el medio dieléctrico en combinación con la naturaleza estocástica del fenómeno de conmutación puede inducir un alto grado de variabilidad [313]. Tal variabilidad en las estructuras memristivas aumenta la posibilidad de errores en el resultado de la multiplicación vector matriz [314]. Más aún, si cada dispositivo tiene características levemente diferentes y las mismas varían en el tiempo, programar cada peso sináptico se prevé como una tarea desafiante. La variabilidad puede ser analizada de dos formas: ciclo a ciclo para un mismo dispositivo (*Cycle-to-Cycle*, C2C) o entre distintos dispositivos (*Device-to-Device*, D2D). En ambos casos, convencionalmente la variabilidad se informa normalizada como σ/μ, donde σ es el desvío estándar y μ el valor medio de las distribuciones de las resistencias R_{ON} y R_{OFF}. La misma está influenciada en gran medida por los materiales utilizados para la estructura (por ejemplo, un dieléctrico mono-capa como HfO_2, o una estructura bi-capa como HfO_X+TaO_X) [315], [316] así como también el tamaño de los dispositivos. En relación al segundo punto, el escalamiento extremo parece reducir la variabilidad, probablemente debido a una reducción del área donde ocurre la conmutación [317].

En la Fig. 6.7g se estudia el impacto que tiene sobre la precisión de la inferencia, la variabilidad de la resistencia programada en los dispositivos RRAM (parámetro λ, cuya variabilidad normalizada $\sigma_\lambda/\mu_\lambda$ se varia entre 0 y 30 %). Este análisis se repite para los ajustes $A1$-$A3$y $C1$-$C3$), una resistencia de linea correspondiente un nodo tecnológico de fabricación de 16 nm (R_L=4.5 Ω), y sin variabilidad en las curvas de HRS y LRS ($\sigma_{R_{OFF}}$ = $\sigma_{R_{ON}}$=0). De aquí se observan claramente dos tendencias marcadamente diferentes. Por un lado, los ajustes con un ratio R_{OFF}/R_{ON} igual o mayor que 100 ($C2$ y $C3$) exhiben una sensibilidad muy reducida a las variaciones de λ (la degradación de la precisión de inferencia es menor al 5 %). Por otro lado, existe una evidente reducción de la precisión para los ajustes $A1$-$A3$ ($A1$ y $C1$ son equivalentes) para el mismo rango de $\sigma_\lambda/\mu_\lambda$. De aquí podría entonces sugerirse la existencia de una relación de compromiso entre la minimización de la caída de tensión las resistencias de linea (utilizando dispositivos con altos valores de R_{ON}) y la mitigación del impacto de la variabilidad D2D (utilizando dispositivos con un ratio R_{OFF}/R_{ON} igual o mayor que 100. No obstante, existe otra arista a considerar.

En caso de tener en cuenta las fuentes de ruido térmico y *flicker* en el circuito del CPA, la relación señal a ruido (*Signal-to-Noise Ratio*, SNR) se degrada sensiblemente al considerar altos valores de R_{ON}, dado que las corrientes resultantes resultan comparables al piso de ruido del sistema. Esto se puede ver la Fig. 6.7j, donde se gráfica el SNR extraído de las simulaciones para diferentes valores de R_L utilizando los ajustes $C1$-$C4$.

El color de los símbolos indica la precisión de inferencia para cada simulación en base a la escala de la derecha. Como se puede observar, para el ajuste $C1$ la precisión de la clasificación se reduce significativamente al considerar el ruido añadido, por lo que el ajuste $C2$ se presenta como el más indicado en términos de mayor SNR, menor impacto de la variabilidad $\sigma_\lambda/\mu_\lambda$ y menor impacto de la resistencia de linea.

6.4.1.4. Dependencia con el tamaño de imagen

La base de datos del MNIST ha sido ampliamente utilizada en la bibliografía para evaluar la precisión de inferencia en ANNs implementadas con CPA. Para sortear las limitaciones impuestas por el tamaño de los CPA disponibles, una práctica usual es re-escalar las imágenes de la base de datos. Por ejemplo, las imágenes de entrenamiento y prueba se re-escalan a 8×8 px. en [283], 14×14 px. en [277] y 22×24 px. en [47] utilizando el método de interpolación bicúbica. Sin embargo, en la Fig. 6.7h puede observarse que para un mismo ratio R_L/R_{ON} hay una sensible diferencia en la precisión de la clasificación para dos resoluciones diferentes (8×8 px. y 28×28 px.). Esto puede explicarse fácilmente si se tiene en cuenta que CPAs más grandes implican una mayor caída de tensión en las resistencias de linea. Por lo tanto sería razonable esperar que al reducir el tamaño del CPA, se aumente progresivamente la precisión de la inferencia. No obstante, como se puede apreciar en la Fig. 6.7d las imágenes se vuelven apenas reconocibles para el ojo humano cuando la resolución cae por debajo de 12×12 px. Por lo tanto se deja entrever que existe una relación de compromiso entre minimizar la caída de tensión en las resistencias de linea (reduciendo el tamaño del CPA) y mejorar la legibilidad de las imágenes (aumentando el tamaño del CPA).

Con el fin de investigar como la resolución de las imágenes afecta el rendimiento general de la clasificación en ANNs basadas en CPAs, se han evaluado los casos de imágenes de 28×28 px. (imágenes originales) hasta 3×3 px. En cada caso, las imágenes de la base de datos fueron re-escaladas mediante el método de interpolación bi-cúbica y la resistencia de linea variada paramétricamente entre 1 y 100 Ω. Asimismo se consideraron dos implementaciones diferentes: Por un lado CPAs sin particionar, y por otro CPAs divididos en 4 particiones. A modo de referencia, los casos de 28×28 (I), 20×20 (II), 12×12 (III) y 8×8 px. (IV) usados como ejemplo en la Fig. 6.7d han sido señalados en la Fig. 6.7i.

Para el caso sin particionar (NP=1), los resultados de las simulaciones muestran en primera instancia, que R_L tiene un impacto mayor en CPAs de gran tamaño, con el caso I exhibiendo una reducción de la inferencia desde 88.21 % a 46.43 % cuando R_L pasa de 1 a 100 Ω. Por el contrario, el caso IV muestra una variación mucho menor de la precisión para el mismo rango de R_L (baja de 88.91 % a 84 %). En segunda instancia, es posible observar que para cada valor de R_L existe un tamaño de imagen (y por ende de CPA) que maximiza la precisión de la inferencia. Tales valores se indican en la Fig. 6.7i, siendo 2000, 3380 y 5780 dispositivos R_L igual a 100, 10 y 1 Ω respectivamente. Interesantemente al analizar el caso particionado (NP=4), se puede observar como se minimiza el impacto de R_L, lo

cual explica la mejora en la métrica de inferencia para el caso I y $R_L = 100\Omega$, que pasa de 46.33 % a 72.63 %. En la misma linea, los puntos de máxima precisión - máximo tamaño, pasan a 3920, 5120 y 11520 dispositivos para R_L igual a 100, 10 y 1 Ω, respectivamente.

6.4.2. Perceptrón Multi-Capa (*Multi Layer Perceptron*, MLP)

A diferencia del caso del SLP donde el tamaño de la red (cantidad de dispositivos/sinapsis) queda determinado por el tamaño de los patrones de entrada y el número de clases diferentes, en el caso de MLPs la introducción de capas neuronales ocultas resulta en una infinidad de posibles redes para la clasificación de una única base de datos [296]. En este contexto, se sabe que al agregar más capas neuronales ocultas es posible incrementar la precisión de la red. Sin embargo, al considerar una implementación en *hardware* de los mismos utilizando memristores y teniendo en cuenta que cada capa sináptica puede considerarse como un SLP, existe una degradación de las señales propagadas a través de cada capa sináptica debido a la resistencia de linea y el denominado efecto de *sneak-path*, tal como se discutió en la Sección 6.4.1. Consecuentemente es posible que las capas neuronales ocultas propagan señales erróneas y por ende atenten contra la precisión del MLP. Con el objetivo de clarificar este punto, 5 MLPs con diferentes números de capas neuronales ocultas y de neuronas por capa fueron analizados. Los resultados obtenidos se muestran en la Tabla 6.3, considerando para todos los casos las imágenes del MNIST re-escaladas a una resolución de 8×8 px, DSC, una única partición (NP=1), V_{read}=300 mV y R_L entre 100 mΩ y 1 kΩ.

Los resultados de las simulaciones se presentan en función de R_L en la Fig. 6.8, donde la precisión de inferencia se muestra normalizada con respecto al caso de $R_L \longrightarrow 0\Omega$. Un punto central a resaltar es la mayor sensibilidad de la precisión frente a variaciones de R_L para el MLP, independientemente del número de capas ocultas y neuronas por capa. Esto puede explicarse teniendo en cuenta el mayor tamaño de las capas sinápticas involucradas en los MLPs (en el caso del SLP la capa sináptica tiene un tamaño de 64×10 mientras que para los MLP –MLP-#a– en la primer capa tiene como mínimo de

Tabla 6.3: Estructura de los MLPs abordados en las simulaciones consideradas en esta sección. En todos los casos se utilizó como patrón de entrada a las imágenes del MNIST re-escaladas a una resolución de 8×8 px.

Capas ocultas	**Código**	**Estructura de la red**	**Num. de memristores**	**Precisión** $R_L \longrightarrow 0\Omega$	**Precisión (Software)**
0	SLP	64×10	1.280	89.6 %	91.14 %
1	MLP-2a	64×54×10	7.992	92.3 %	95.95 %
	MLP-2b	64×100×10	14.800	92.7 %	96.89 %
2	MLP-3a	64×54×34×10	11.263	95.2 %	96.30 %
	MLP-3b	64×100×50×10	23.800	96 %	96.92 %
3	MLP-4	64×54×34×24×10	12.696	94.3 %	95.81 %

64×54). En este escenario, la utilización de CPAs sin particionar para implementar dichas capas sinápticas degrada la tensión efectiva aplicada a las sinapsis localizadas lejos de los terminales de entrada y salida del CPA. Esta interpretación encuentra sustento además en los resultados obtenidos para los MLPs etiquetados como MLP-#b. Tal como podría esperarse de considerar un MLP con capas sinápticas más grandes (64×100 para el MLP-#b frente a 64×54 para MLP-#a), en estas simulaciones se observa una sensibilidad aún mayor con R_L. Por otro lado también vale la pena notar que tanto MLP-#a y MLP-#b siguen tendencias unificadas decrecientes en función de R_L e independientes del número de capas sinápticas. Por lo tanto aumentar el número de capas neuronales ocultas no compromete significativamente la sensibilidad de la precisión de inferencia frente a la resistencia de linea R_L pero si resulta en un aumento no despreciable de la precisión cuando se considera un escenario de mínima R_L, como se muestra en el *inset* de la Fig. 6.8. Por el contrario, el aumento del número de neuronas por cada capa neuronal causa una degradación sustancial de la precisión de inferencia en función de R_L, dado que involucra CPAs sustancialmente más grandes.

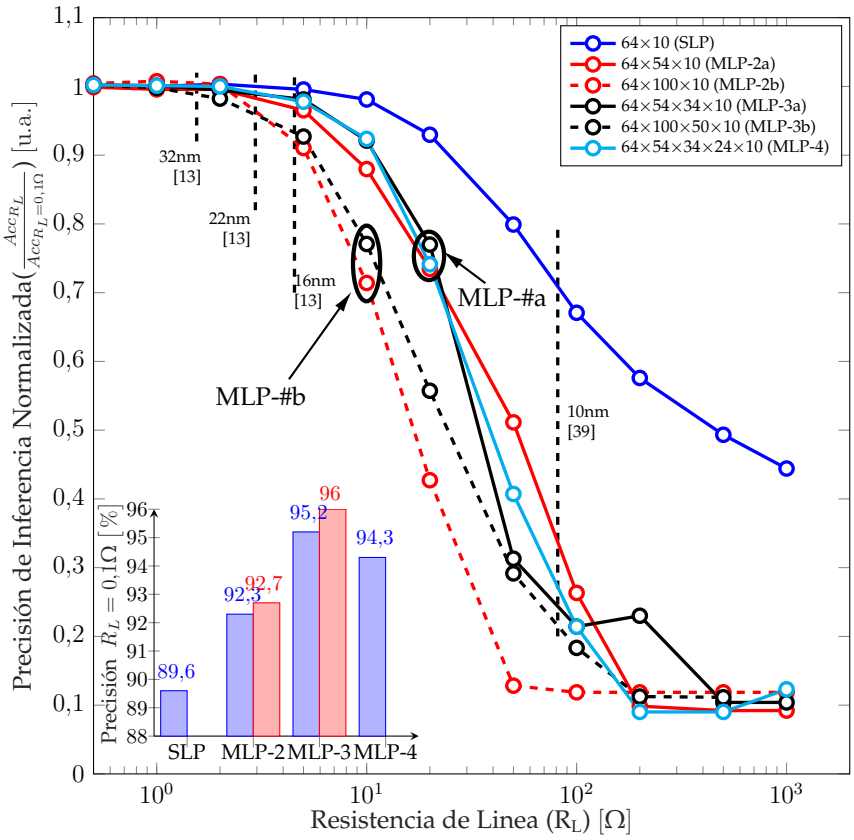


Figura 6.8: Precisión de inferencia en función de R_L, normalizada respecto de la precisión obtenida para el caso de $R_L \longrightarrow 0\Omega$. Se consideran dos grupos diferentes de MLPs (#a y #b, véase la Tabla 6.3) así como el caso del SLP. El *inset* indica la precisión de inferencia obtenida para $R_L \longrightarrow 0\Omega$ para cada uno de los casos considerados. Nótese que la dependencia con R_L está determinada por el tamaño de la capa sináptica más grande en el MLP, y presenta muy poca sensibilidad al número de capas ocultas. De hecho, agregando más capas ocultas es posible incrementar notoriamente la precisión, sin que esto implique una mayor degradación producida por R_L.

6.5. Aspectos de confiabilidad

Junto a los elementos parásitos y fuentes de variabilidad descritos en la sección 6.4, la falta de maduración en los procesos de fabricación asociados a la producción de CPAs de memristores resulta en la existencia de dispositivos defectuosos. Entre estos, las fallas de enclavamiento (*Stuck-at-Faults*, SAFs) son un gran desafío a resolver, dado que causan la potenciación (dispositivo enclavado en el estado de LRS o *Stuck-at-ON*, SA1) o depreciación (dispositivo enclavado en el estado de HRS o *Stuck-at-OFF*, SA0) no deseada de las conexiones sinápticas del CPA [236], [265]. En la presente sección, la precisión de inferencia se estudia en función de la porción de SAFs en el CPA, además de tener en cuenta la posibilidad de tener dispositivos que no hayan sido electro-formados (SA0_nE). En este contexto, el modelo QMM considerado es particularmente útil para inyectar defectos en el CPA ya que esto se puede hacer mediante la modificación de un solo parámetro: λ ($\lambda = 1$ corresponde a fallas del tipo SA1, $\lambda = \lambda_{min}$ a SA0, y $\lambda = 0$ a fallas SA0_nE). Dada la naturaleza estocástica de la distribución espacial de fallas SAFs en el CPA [264], [318], se realizaron simulaciones de tipo Monte Carlo (MC) asumiendo distintas porciones de dispositivos en SAF. En cada simulación, los dispositivos defectuosos son distribuidos aleatoriamente en el CPA siguiendo una distribución uniforme [265], [318] y el CPA defectuoso obtenido es utilizado para clasificar las imágenes de una determinada base de datos. Dichas SAFs se inyectan en las matrices de conductancia G_M^+ y G_M^- (véase el diagrama de flujo de la Fig. 6.4a). La precisión de clasificación registrada es entonces promediada entre todas las simulaciones de MC para un determinado ratio de SAFs y presentada en la Fig. 6.9. Se considera el ajuste $C2$ y dos resoluciones de imágenes diferentes (8×8 px. y 16×16 px.) y R_L entre 1 Ω y 100 Ω. Dado el tamaño relativamente pequeño de los SLP considerados, no se ha considerado el particionado de los mismos. No obstante se evalúan distintos métodos de normalización (*Normalization Method*, NM) para mapear W_M en G_M^+ y G_M^- a fin de determinar su robustez frente a SAFs y su impacto en la precisión de inferencia. Por último, unas 10 simulaciones de Monte-Carlo se consideran para cada combinación de R_L, NM, tamaño de imagen y ratio de SAFs, resultando en un total de ∼4.3k simulaciones.

6.5.1. Consideraciones de diseño

Los elementos de W_M se encuentran en el rango $\left[min\{W_M\}, max\{W_M\}\right]$ y siguen la distribución que se indica en la Fig. 6.9a. Con el objetivo de traducir dichos valores al rango $[G_{HRS}, G_{LRS}]$, estos deben ser primeros normalizados al rango $[-1, 1]$. Usualmente [238], [298] tal normalización se realiza dividiendo W_M por el elemento de W_M con el máximo valor absoluto (Método de Normalización 1, MM-1) o por la diferencia máxima entre elementos de W_M (Método de Normalización 2, NM-2). Como cabe esperar, las ma-

trices normalizadas $W_{M_{N1}}$ y $W_{M_{N2}}$ preservan la misma distribución que W_M y la misma relación $max\{W_M\}/min\{W_M\}$, como se puede apreciar en las Figs. 6.9c y 6.9b, respectivamente. Interesantemente, para el caso de las imágenes de la base de datos del MNIST re-escaladas a 8×8 px., 95 % de los elementos de $W_{M_{N1}}$ cae dentro del rango [-0.5,0.5], mientras que el porcentaje asciende al 99 % cuando se consideran los elementos de $W_{M_{N2}}$. De aquí se desprenden dos grandes problemas de estos métodos de normalización.

En primer lugar, ninguno de los dos explota adecuadamente el rango dinámico de los memristores dado que la amplia mayoría de ellos estarán programados en un nivel de conductancia en el rango $\left[G_{HRS}, \left(\frac{(G_{LRS}+G_{HRS})}{2}\right)\right]$. En segundo lugar, la concentración de pesos sinápticos cercanos a 0 (G_{HRS}) exacerba el impacto de las fallas tipo SA1 [319]: Dado que una fracción significativa de dispositivos están mapeados cerca de G_{HRS} (y por lo tanto $\lambda \longrightarrow \lambda_{min}$) una falla del tipo SA1 (λ=1) causa un corrimiento importante respecto de la conductancia esperada y por lo tanto degrada la inferencia. Para mitigar estos problemas, se propone un método alternativo de normalización (NMM-3) basado en la distribución de tipo normal que siguen los elementos de W_M. De esta forma, el elemento $w_{i,j} \in W_M$ tiene una probabilidad P_i de estar en el rango $\mu_{W_M} \pm i\sigma_{W_M}$, donde μ_{W_M} y σ_{W_M} son la media y desvío estándar de los valores de W_M. Para i entre 1 y 4, el ∼68.3 %, ∼95.5 %, ∼99.7 % y ∼99.9 % de los pesos sinápticos estarán contenidos dentro de dicho rango, respectivamente [320]. Por lo tanto, los valores que excedan dichos límites son fijados a $\mu_{W_M} \pm i\sigma_{W_M}$ y normalizados para obtener $W_{M_{N3}}$. Los histogramas mostrando la distribución de elementos resultantes para cada caso, se muestran en las Figs. 6.9d-6.9f para 2, 3 y $4\sigma_{W_M}$, respectivamente.

El impacto del método de normalización (NM) en la precisión de la inferencia se presenta en función del ratio de dispositivos defectuosos en las Figs. 6.9g-6.9i, considerando fallas del tipo SA1, SA0 y SA0_nE. R_L se mantuvo fija en 10 Ω para los 3 casos. Tal como se ha reportado en la literatura [239], [283], las fallas del tipo SA1 tienen un impacto mucho mayor sobre la precisión de clasificación que las fallas de tipo SA0, mientras que poca o ninguna diferencia existe entre los casos SA0 y SA0_nE. A su vez, para las fallas tipo SA1, se observa una importante dependencia con el método de normalización utilizado, con el NM-3 mostrando la mayor robustez frente a SA1. A diferencia de la abrupta merma de la precisión observada para los métodos NM-1 y NM-2, el caso NM-3 tiene una mayor tolerancia a los dispositivos defectuosos. De hecho, cuanto mas se aleja la distribución de los elementos en $W_{M_{N3}}$ de una distribución normal, menor es el impacto de los dispositivos defectuosos en la precisión de inferencia (Véase la diferencia del método NM-3 para $2\sigma_{W_M}$ y $4\sigma_{W_M}$).

No obstante, esta mejora tiene asociado un mayor consumo de potencia (alcanzando ∼10 mW en un SLP con ∼15.6k sinapsis) durante la fase de inferencia (véase el *inset* de la Fig. 6.9h) y una pequeña pérdida de precisión para el caso de considerar un CPA libre de fallas (véase el *inset* de la Fig. 6.9i). Tal incremento en la potencia consumida es una consecuencia esperable de mapear una fracción mayor de elementos de W_M cerca

de G_{LRS}, lo que inevitablemente incrementa las corrientes a través del CPA. Esto también juega un rol clave en la reducción de precisión observada para altos valores de R_L en el SLP libre de fallas (véase el *inset* de la Fig. 6.9i), dado que podría interpretarse como un aumento en el ratio R_L/R_{ON}. Más aún, inclusive para valores reducidos de R_L hay una

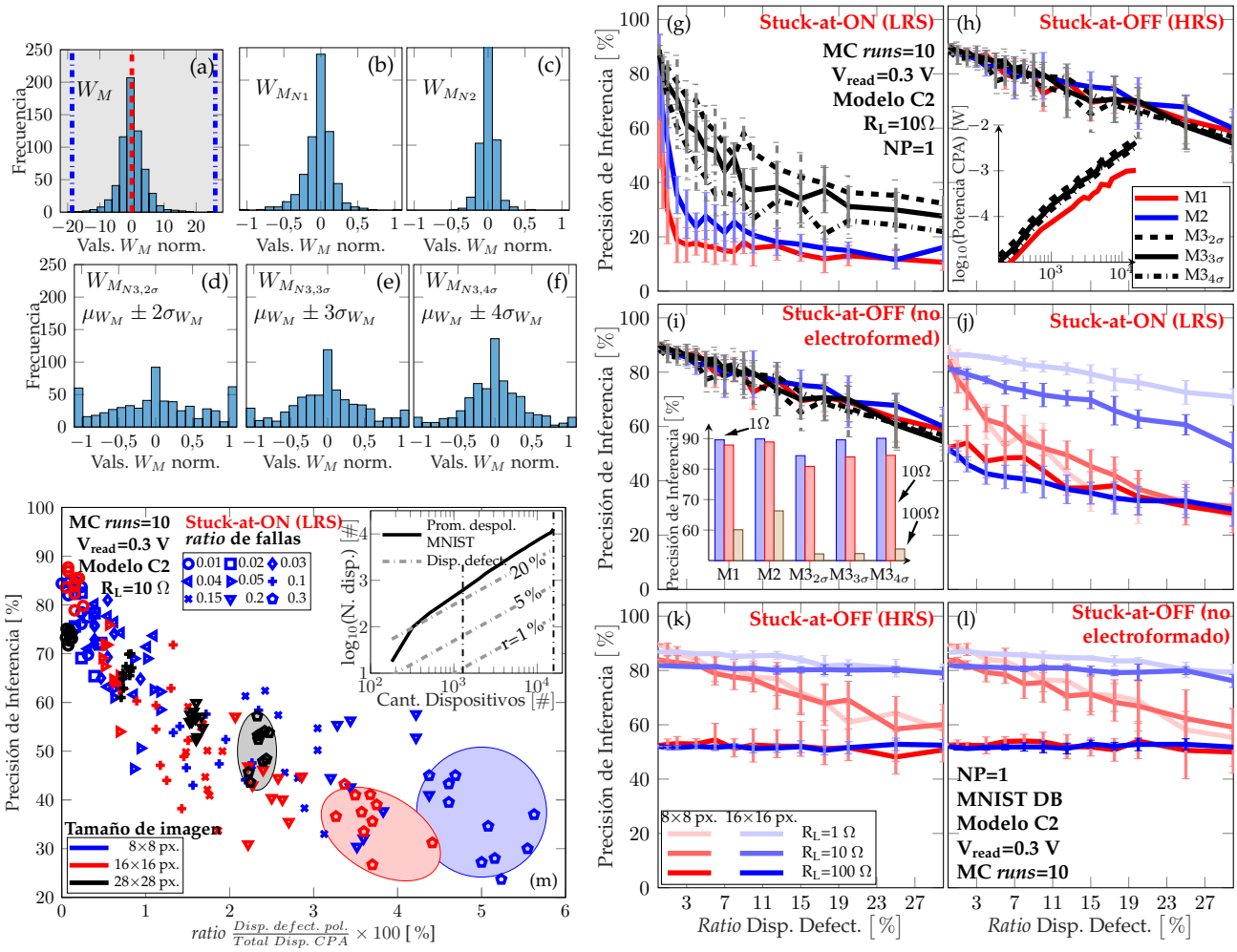


Figura 6.9: El cambio en (a) la distribución de elementos de W_M según las distintos métodos de normalización se muestra en (b)-(f). La precisión de inferencia en función del porcentaje de dispositivos defectuosos considerando distintos métodos de normalización se presenta para fallas del tipo (g) SA1, (h) SA0 y (i) SA0_nE. La potencia disipada en el SLP durante la fase de inferencia se indica en el *inset* de (h) en función del tamaño del SLP. Similarmente, la precisión de inferencia obtenida en el SLP libre de fallas se muestra para cada método de normalización en el *inset* de (i). Por otro lado, la precisión de inferencia en función del porcentaje de dispositivos defectuosos se presenta en (j)-(l) para las fallas de tipo SA1, SA0 y SA0_nE, teniendo en cuenta cuatro valores distintos de R_L. (m) Precisión de inferencia obtenida con el SLP en función del porcentaje de dispositivos defectuosos polarizados. Cada símbolo corresponde a una ejecución de las simulaciones de Monte-Carlo. Los datos se codifican en términos del porcentaje total de dispositivos defectuosos (tipo de símbolo) y tamaño del SLP (color del símbolo). Por ejemplo, los círculos azules corresponden a los resultados de inferencia para simulaciones de SLPs de 64×10 con un 1 % de dispositivos defectuosos. El caso puntual de SLPs con un 30 % de dispositivos defectuosos (pentágonos) han sido señalados con el fin de ejemplificar la reducción del porcentaje de dispositivos defectuosos polarizados a medida que aumenta el tamaño del SLP (∼5 % en el SLP de 1280 dispositivos (imágenes de 8×8), ∼3.7 % con 5120 dispositivos (imágenes de 16×16) y finalmente ∼2.3 % con 15680 dispositivos (imágenes de 28×28).

sutil reducción de la precisión causada por la re-distribución de los pesos sinápticos entre W_M y $W_{M_{N3}}$. Por lo tanto, NM-3 se considerará a lo largo de esta sección dado que provee la mayor robustez frente a SAFs y la menor pérdida de precisión en el escenario libre de fallas.

Por otro lado, la precisión de clasificación en CPA con SAFs fue estudiada para diferentes valores de R_L (1 Ω, 10 Ω y 100 Ω) y resoluciones de las imágenes de la base de datos del MNIST (8×8 px. y 16×16 px.). En ambos casos, la precisión de inferencia para un ratio de dispositivos defectuosos tendiente a cero muestra un corrimiento hacia abajo a medida que R_L aumenta de 1 Ω a 100 Ω, en linea con los resultados presentados en la Fig. 6.7h. Para las imágenes más pequeñas (8×8 px., SLP de 64×10) e independientemente del tipo de SAF, la sensibilidad de la precisión de inferencia a los dispositivos defectuosos crece sensiblemente a medida que se reduce R_L, siendo el caso más notorio el de $R_L = 1\Omega$, en las Figs. 6.9j-6.9l. Nótese que en la clasificación de las imágenes de 16×16 px. la precisión de inferencia se vuelve insensible a los dispositivos defectuosos para valores de R_L mayores a 10 Ω cuando se consideran fallas del tipo SA0.

Este comportamiento puede explicarse teniendo en cuenta dos factores. Por un lado, se ha mostrado en la sección 6.4 [267], [301], [306] que el margen de lectura de los dispositivos del CPA está determinado en forma conjunta por la resistencia de los dispositivos (R_{memd}, que varía entre R_{OFF} y R_{ON}) y la resistencia R_L. Mediante un análisis de primer orden, se puede decir que cada memristor forma parte de un camino conductivo entre la entrada de la WL *i* y la salida de la bitline *j*. En un SLP de $N \times M$, la resistencia parásita promedio asociada a tal camino es $R_L[(N + M)/2 + 1]$ [267], [283]. En este escenario simplificado, el cociente V_{cell}/V_{read} puede ser obtenido planteando un divisor de tensión entre R_{memd} y $R_L[(N + M)/2 + 1]$. Los resultados se muestran en la Tabla 6.4 para las dos resoluciones de imagen consideradas y los distintos valores de R_L, considerando tanto fallas del tipo SA0 y SA1. A pesar de ser una limitación para mejorar el rendimiento en CPAs libres de fallas, la reducción del margen de lectura como consecuencia del aumento de R_L tiene un efecto positivo cuando se consideran SAFs, dado que se reduce la tensión aplicada a los dispositivos defectuosos. Esto es particularmente notorio en el caso de los SLP de 256×10 sujetos a fallas del tipo SA1 y evaluados con las imágenes de 16×16 px. del MNIST. En estos, solo el ∼49 % de la tensión de entrada es aplicada a los dispositivos, lo que en consecuencia reduce la contribución de los mismos a la corriente de salida total de la BL asociada.

Por otro lado, las imágenes de la base de datos del MNIST incluyen una fracción de píxeles inactivos (por ejemplo, aquellos próximos a los bordes). Interesantemente, tal

Tabla 6.4: Cociente $V_{celda}/V_{lectura}$ calculado con el equivalente serie simplificado

	8×8 px. MNIST (64×10 SLP)			16×16 px. MNIST (256×38 SLP)		
	$R_L = 1\Omega$	$R_L = 10\Omega$	$R_L = 100\Omega$	$R_L = 1\Omega$	$R_L = 10\Omega$	$R_L = 100\Omega$
SA1: R_{Memd}=R_{ON} (10 kΩ)	∼0.99	∼0.96	∼0.72	∼0.99	∼0.90	∼0.49
SA0: R_{Memd}=R_{OFF} (1 MΩ)	∼1	∼0.99	∼0.99	∼1	∼0.99	∼0.98

fracción no se mantiene constante cuando se reduce la resolución de las imágenes, como se ejemplifica en la Fig. 6.7d. Por el contrario, imágenes más pequeñas muestran un porcentaje menor de píxeles inactivos. Por lo tanto, cuando las entradas de un SLP de tamaño $n^2 \times m$ son conectadas a los píxeles de una imagen de prueba de n×n, el porcentaje de dispositivos del SLP que sean polarizados a 0 voltios aumenta con la resolución de las imágenes (n), como se muestra en el *inset* de la Fig. 6.9m. En esta se puede ver que el número de dispositivos no polarizados en el CPA tiene un crecimiento más abrupto que el número de dispositivos defectuosos, para ratios de SAFs del 1 %, 5 % y 20 %. Dado que los dispositivos en SAFs se distribuyen uniformemente en todo el CPA, es razonable esperar que parte de dichos dispositivos con falla de enclavamiento se encuentren polarizados con 0 voltios, y por lo tanto no jueguen un rol en el proceso de clasificación. Para probar esta interpretación, el valor obtenido de precisión de inferencia para cada simulación ejecutada con $R_L = 10\Omega$ en la Fig. 6.9j se presenta en el gráfico de dispersión de la Fig. 6.9m en función del ratio de dispositivos defectuosos polarizados. Por completitud, se agrega también el caso de las imágenes de 28×28. De aquí sobresalen dos características. En primer lugar, a pesar de las claras diferencias observadas para las imágenes de 8×8 y 16×16 píxeles en la Fig. 6.9j se puede observar una tendencia unificada cuando se considera la precisión de inferencia en función del ratio de dispositivos defectuosos polarizados para SLPs de diferentes tamaños. En segundo lugar, solo una fracción de los dispositivos defectuosos son polarizados, y dicha porción decrece a medida que aumenta el tamaño del CPA (A modo de referencia, se señala el caso de CPAs con 30 % de dispositivos defectuosos).

6.5.2. Estrategias para la mitigación de fallas de enclavamiento

A partir de los resultados presentados en las Secciones 6.4 y 6.5, es evidente que las SAFs (tanto SA1 como SA0) tienen un impacto no despreciable en al precisión de clasificación obtenida en SLPs, independientemente del método de normalización elegido, resolución de imágenes y relación R_L/R_{ON}. En este contexto, es necesaria la utilización de técnicas que permitan mitigar tales defectos para permitir la operación confiable de redes neuronales basadas en CPAs de memristores. Esto implica alterar los pesos sinápticos obtenidos durante el entrenamiento. Para ello, en esta sub-sección se proponen 3 métodos diferentes, los cuales parten de dos premisas: *i*) que es posible conocer la localización de las celdas RRAM defectuosas y *ii*) que las filas del CPA pueden ser permutadas. Con respecto al primer punto, en [321] se ha presentado un método simple de evaluación de la actividad de conmutación en CPAs mediante el cual es posible obtener fácilmente la localización espacial de los dispositivos defectuosos. En segundo lugar, el orden de las filas y columnas en el CPA puede ser permutado sin cambiar el resultado final de la multiplicación vector matriz si las entradas y salidas son permutadas simultáneamente y siguiendo

el mismo orden [236]. Los algoritmos en cuestión se discuten a continuación y basan su funcionamiento en *i*) la compensación de las celdas RRAM defectuosas en un esquema de doble CPA, *ii*) la minimización de la variación de la suma ponderada y *iii*) la permutación basada en la actividad promedio de cada pixel.

6.5.2.1. Algoritmo 1

Como en múltiples estudios reportados en la bibliografía [219], [265], [286], [299] se utilizan dos memristores para representar cada uno de los elementos de la matriz W_M normalizada (pesos sinápticos). A su vez, y tal como se explica en la Sub-Sección 6.3.1, W_{Norm} se computa como $W^+_{Norm} - W^-_{Norm}$, con W^+_{Norm} y W^-_{Norm} siendo los elementos positivos y y negativos de W_{Norm}, respectivamente. Tal técnica permita implementar un método de remapeo simple pero efectivo para minimizar el efecto de las SAFs. Este implica que para cada celda RRAM $g^{-(+)}_{i,j}$ defectuosa en el CPA positivo (o negativo), la celda RRAM $g^{-(+)}_{i,j}$ correspondiente en el CPA negativo (o positivo) es re-ajustada para compensar el error en $g^{-(+)}_{i,j}$. Esto puede resumirse mediante la Ec. 6.9:

$$g^{+(-)}_{i,j} = \begin{cases} g^{+(-)}_{i,j}, & w^+_{M_{Norm}} \wedge w^-_{M_{Norm}} \text{está libre de defectos} \\ g^{+(-)}_{i,j} - g^{-(+)}_{i,j} w^+_{M_{Norm}} & \vee\ w^-_{M_{Norm}} \text{está libre de defectos} \end{cases} \tag{6.9}$$

Por ejemplo, si el peso sináptico ($w_{M_{Norm}}$) es positivo pero la celda RRAM correspondiente en el CPA positivo ($w^+_{M_{Norm_{eff}}}$) contiene una falla de tipo SA1, el mapeo ideal incurre en un error ya que la celda RRAM positiva no puede ser ajustada al valor requerido por estar enclavada en LRS. Para corregir este problema, se aumenta el valor de la conductancia en la celda RRAM correspondiente del CPA negativo. No obstante, es importante notar que la aplicación directa de la Ec. 6.9 para mitigar el efecto de los dispositivos defectuosos puede resultar en pesos sinápticos no realizables. Supóngase por ejemplo que un peso sináptico positivo debe ser realizado con un memristor libre de fallas en el CPA positivo, y un memristor enclavado en LRS en el CPA negativo. En este caso, la Ec. 6.9 indicará que el peso sináptico necesario en el CPA positivo para la corrección es mayor que 1 y por lo tanto fuera de los límites. Por lo tanto, ciertas fallas pueden ser toleradas por la aplicación

Tabla 6.5: Combinaciones de fallas recuperables y no-recuperables

Target $(W_{i,j})$	**Estado de la celda RRAM en el CPA +**	**Estado de la celda RRAM en el CPA -**	**¿Recuperable?**
Positivo	*Stuck-at-ON*	Libre de falla	**Si**
Positivo	*Stuck-at-OFF*	Libre de falla	**No**
Positivo	Libre de falla	*Stuck-at-ON*	**No**
Positivo	Libre de falla	*Stuck-at-OFF*	**Si**
Negativo	*Stuck-at-ON*	Libre de falla	**No**
Negativo	*Stuck-at-OFF*	Libre de falla	**Si**
Negativo	Libre de falla	*Stuck-at-ON*	**Si**
Negativo	Libre de falla	*Stuck-at-OFF*	**No**

directa de este método (fallas recuperables) y otras requieren un procesamiento previo (fallas no recuperables), como se muestra en la Tabla 6.5

Mediante un algoritmo iterativo de permutación, las fallas no-recuperables pueden ser convertidas en fallas recuperables, como se indica esquemáticamente en la sección inferior de la Fig. 6.10a. Nótese, a modo de ejemplo, que los pares de filas $\{i, j\}$ y $\{k, l\}$ han sido permutados de forma de eliminar las fallas no-recuperables en $\left\{g^{+}_{i,1}, g^{+}_{j,1}, g^{-}_{k,1}, g^{-}_{l,3}\right\}$. El algoritmo completo de re-mapeo incluyendo la permutación de filas y compensación de conductancia se presenta en términos de pseudo-código en el Algoritmo 1 en el apéndice B.6.

6.5.2.2. Algoritmo 2

Llámese $WV = \left|w^{SAF}_{i,j} - w_{i,j}\right|$ a la diferencia existente entre un peso sináptico ($w_{i,j}$) de la matriz $W_{M_{Norm}}$ obtenida por simulación y libre de fallas, y el peso sináptico realizable ($w^{SAF}_{i,j}$) en un CPA con dispositivos defectuosos ($W^{SAF}_{M_{Norm}}$). Esta puede extenderse a todo el CPA para definir la métrica de Variación Total de Pesos (*Sum Weight Variation*, SWV) [319] con el fin de cuantificar cuanto se aleja la matriz de pesos efectivamente

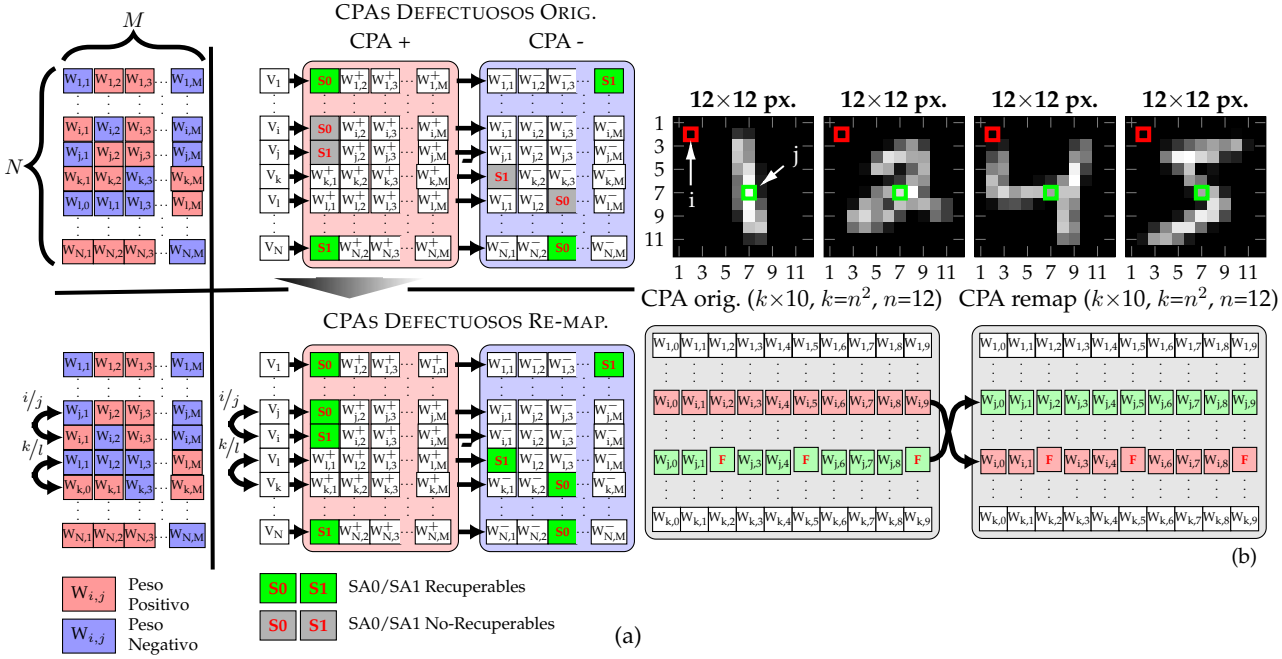


Figura 6.10: (a) Representación esquemática del algoritmo de re-mapeo número 1, describiendo la compensación de conductancias que permite tolerar fallas en la primer y última filas (celdas coloreadas en verde) pero que es incapaz de manejar otras SAFs (celdas coloreadas en gris, las cuales son irrecuperables). Para solucionar este inconveniente, se propone la permutación de filas (abajo) para transformar fallas irrecuperables en recuperables (véase la Tabla 6.5. (b) La permutación de filas también es usada en los Algoritmos 2 y 3. En el último esta se utiliza para re-mapear las filas con más dispositivos defectuosos a los píxeles menos activos.

mapeada al CPA defectuoso, de la matriz de pesos ideal, según se indica en la Ec. 6.10:

$$\text{SWV} = \sum_{i=1}^{M} \sum_{j=1}^{N} \mid w_{i,j}^{eff} - w_{i,j} \mid \tag{6.10}$$

donde M y N indican el número de filas y columnas de la matriz $W_{M_{Norm}}$. De la Ec. 6.10, se desprende que cuanto menor sea SWV, menor será el impacto de las SAFs en el mapeo de la matriz de pesos. El algoritmo propuesto entonces consiste en reducir progresivamente SWV mediante la permutación secuencial de filas hasta hallar el mínimo valor posible. El procedimiento se indica en pseudo-código en el Algoritmo 2 en el apéndice B.6.

6.5.2.3. Algoritmo 3

Como se muestra en la parte superior de la Fig. 6.10b para el caso de la base de datos del MNIST re-escalada a un tamaño de 12×12 px., un número considerable de píxeles permanece en negro (inactivos) mientras algunos otros son normalmente blancos (activos), independientemente del dígito que representan. Teniendo en cuenta que el brillo de cada píxel se codifica como un voltaje en el rango de 0 a V_{read}, se puede obtener un brillo promedio para cada píxel de la base de datos. Ambos casos se ilustran en la Fig. 6.10b: el píxel i indica un píxel normalmente inactivo (por ejemplo aquellos cerca de los bordes de la imagen), mientras que el píxel j indica un píxel normalmente activo (por ejemplo aquellos en el centro de la imagen). Para el caso de las imágenes de 12×12 px. cada uno de los 144 px. resultantes es usado para polarizar cada una de las filas de dos CPA (positivo y negativo) de tamaño 144×10. En el escenario más simple, el 1^{er} píxel (esquina superior izquierda de la imagen) se asocia a la 1^{er} fila del CPA. A continuación, el i^{th} pixel se conecta a la i^{th} fila, el j^{th} píxel a la j^{th} fila y finalmente el 144^{th} píxel (esquina inferior derecha de las imagen) con la 144^{th} fila. Sin embargo, tal mapeo no tiene en cuenta la distribución de dispositivos defectuosos en el CPA. De conocerla es posible determinar cuantos dispositivos defectuosos están conectados a cada fila del CPA y por consiguiente re-mapear los píxeles más activos a las filas "menos defectuosas" (filas con menos cantidad de celdas RRAM defectuosas asociadas) y los píxeles menos activos a las celdas con más dispositivos defectuosos. Este procedimiento se representa esquemáticamente en la Fig. 6.10b en términos de los píxeles i y j y en el Algoritmo 3 en el apéndice B.6.

6.5.2.4. Resultados

La capacidad de los algoritmos de re-mapeo presentados en las secciones 6.5.2.1–6.5.2.3 para mitigar el efecto de las fallas de enclavamiento se discute en la presente sección. Para ello se consideran dos escenarios. En primer lugar se considera la clasificación de las imágenes del MNIST en una resolución de 8×8 px. mediante un SLP de 64×10 sin particionar (NP=1), dado que este caso se observa como el más sensible en la Fig. 6.9. En segundo lugar, se considera una base de datos alternativa, la base de datos de rostros

de la Universidad de Yale (*Yale Face Dataset*). Dichas imágenes son re-escaladas a un tamaño de 16×16 y clasificadas con un SLP de 256×38 con 4 particiones (NP=4). Algunas imágenes de este dataset se muestran en la Fig. 6.11a. Al igual que en las simulaciones previas, las características I–V de los memristores fueron representadas mediante el ajuste $C2$, R_L fijado en 10Ω, G_M^+ y G_M^- obtenidos mediante el método de normalización 3 (NM-3) y 10 simulaciones se realizaron para cada ratio de dispositivos defectuosos. Nótese que solamente los casos SA1 y SA0 fueron considerados, dada la gran similitud entre los resultados obtenidos para SA0 y SA0_nE presentados en la Fig. 6.9. Adicionalmente es importante resaltar que este estudio se presenta para los casos de tener exclusivamente fallas del tipo SA1 o SA0, dado que considerar su efecto combinado requeriría de múltiples escenarios con diferentes ratios de fallas SA1 y SA0, aumentando notablemente el tiempo y recursos necesario para su realización. En este escenario, se realizan 4 simulaciones diferentes para un dado k-esimo CPA: *i*) se simula el mapeo original, *ii*)-*iv*) las celdas RRAM no defectuosas son alteradas siguiendo los algoritmos 1-3 mientras se mantienen las celdas defectuosas. De esta forma se realizaron en total unas ∼1700 simulaciones.

La precisión de inferencia promedio se presenta en función del ratio de dispositivos defectuosos, para las imágenes en 8×8 px. de la base de datos del MNIST considerando fallas de tipo SA1 y SA0, en las Figs. 6.11b y 6.11c, respectivamente. La precisión del caso libre de fallas (∼90.1 %) se indica a modo de referencia en ambos casos. A pesar de que los tres algoritmos considerados muestran una mejora de la precisión independientemente del tipo de SAF considerada, existen diferencias sustanciales que ameritan un discusión detallada. Por un lado, para las fallas del tipo SA1 (véase la Fig. 6.11b) el Algoritmo 1 ofrece los mejores resultados, permitiendo una precisión por encima del ∼75 % con hasta un 10 % de dispositivos defectuosos. A partir de este punto, el rendimiento es superado por el re-mapeo propuesto por los Algoritmos 2 y 3 para CPAs con hasta un 20 % de dispositivos defectuosos, no obstante la mejora que permiten no excede el ∼10 % en un escenario de muy baja precisión (debajo del ∼30 %). Más aún, los Algoritmos 2 y 3 muestran una mejora estadísticamente idéntica sobre el rango evaluado. Por el contrario, el Algoritmo 2 muestra una mejora sobresaliente de la precisión de inferencia cuando se consideran fallas del tipo SA0 (véase la Fig. 6.11c), alcanzando una precisión de hasta ∼80 % con un porcentaje de dispositivos defectuosos del 30 %. Una vez más, la mejora provista por el Algoritmo 1 es insuficiente cuando el porcentaje de dispositivos en SAF excede el ∼20 %. Esto podría explicarse teniendo en cuenta que cuando una fracción significativa de los dispositivos del CPA presenta una falla de enclavamiento, no hay suficientes filas libres de fallas en el CPA para convertir fallas irrecuperables en recuperables (véase la Tabla 6.5) mediante la permutación de filas. Por lo tanto el número de SAFs que no puede ser compensado aumenta y en consecuencia limita la funcionalidad del algoritmo de mitigación.

Al considerar la clasificación de las imágenes de la base datos de rostros de la Universidad de Yale codificadas en una resolución de 16×16 px. se observan resultados

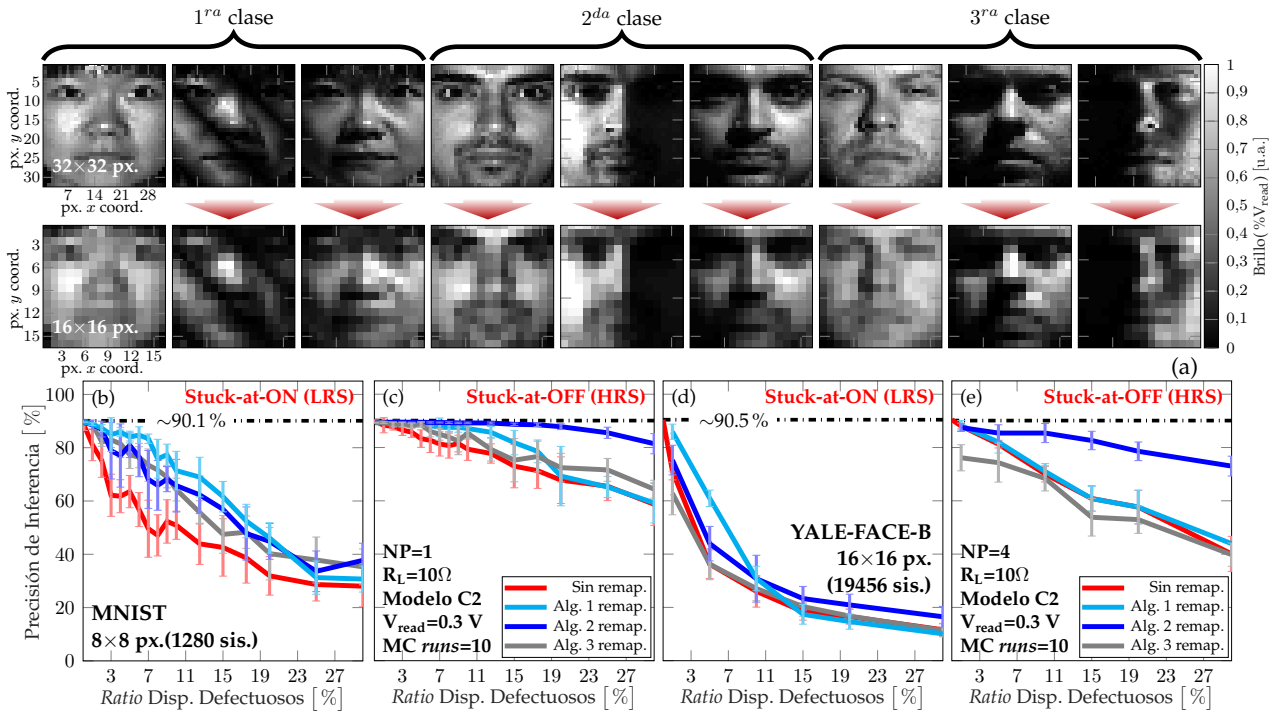


Figura 6.11: (a) Muestra de las imágenes de la base de datos de rostros de la Universidad de Yale, mostrando 3 clases diferentes con resoluciones de 32×32 px. (arriba) y 16×16 px. (abajo). En ambos casos los ejes x e y en las imágenes del extremo izquierdo indican la posición de cada píxel. Los algoritmos de re-mapeo 1–3 fueron puestos a prueba tanto con las imágenes de la base de datos del MNIST como con las presentadas en (a). Las tendencias correspondientes a la inyección de fallas del tipo SA1 y SA0 se muestran para las primeras (MNIST) en las figuras (b) y (c) respectivamente, mientras que para el caso de las segundas (Universidad de Yale) se muestra en las figuras (d) y (e). En ambos casos el Algoritmo 1 muestra los mejores resultados al considerar fallas del tipo SA1, mientras que para fallas del tipo SA0, es conveniente el Algoritmo 2.

similares tanto para el caso de fallas del tipo SA1 como SA0, como se presenta en las Figuras 6.11d y 6.11e, respectivamente. En este caso, los Algoritmos 1 y 2 emergen como las mejores opciones para tolerar fallas del tipo SA1 y SA0. Sin embargo, el ratio máximo de dispositivos defectuosos donde estos algoritmos son capaces de producir una mejora se reduce sensiblemente. Para fallas del tipo SA1 (véase la Fig. 6.11d) el Algoritmo 1 no puede producir mejoras si la fracción de dispositivos con fallas excede el 10 %. De hecho, por encima del 15 % el Algoritmo 1 reduce la precisión del SLP obtenida con el mapeo original. En cambio, el Algoritmo 2 provee una mejora de la precisión constante aunque no mayor al ∼5 % sobre todo el rango evaluado. Al considerar las fallas del tipo SA0 (véase la Fig. 6.11e) el Algoritmo 2 muestra los mejores resultados. Finalmente vale mencionar que el menor rendimiento obtenido por el Algoritmo 3 para las imágenes de Yale puede explicarse en términos las propias características del dataset. Dado que este algoritmo basa su funcionamiento en mapear las filas del CPA a los píxeles con la menor actividad promedio, la no existencia de píxeles normalmente inactivos en esta base de datos reduce su eficacia.

6.6. Conclusiones

En este capítulo se ha explorado la implementación de redes neuronales en *hardware* considerando memorias de conmutación resistiva (RRAM). Las mismas son de gran interés dado su gran rendimiento en tareas de clasificación de patrones con una alta eficiencia energética. El análisis ha sido abordado mediante simulaciones eléctricas realistas realizadas en SPICE, utilizando para ello el modelo Cuasi-Estático del memristor. De esta forma se ha demostrado su aplicabilidad a la simulación de redes neuromórficas con miles de dispositivos destinadas al reconocimiento de imágenes. Dicho modelo es considerado no solo por su alta precisión para modelar las características eléctricas de las memorias resistivas, sino también por su versatilidad y reducido costo computacional. Considerando tanto redes mono-capa (sin neuronas ocultas) como multi-capa (con neuronas ocultas), se ha indagado sobre el efecto de las no idealidades presentes en estructuras tipo *crossbar* de dispositivos RRAM, tales como la resistencia de linea (R_L) y la relación entre resistencia mínima y máxima (R_{OFF}/R_{ON}), mostrando que la relación R_L/R_{ON} juega un rol fundamental en los márgenes de lectura y por lo tanto en la precisión de la inferencia. Al mismo tiempo el impacto de la variabilidad dispositivo a dispositivo (D2D) y la relación señal a ruido (SNR) sobre la capacidad de la red de clasificar correctamente los patrones de entrada también fueron analizadas. Teniendo en cuenta estas dependencias, se analiza el particionado del *crossbar* como herramienta para minimizar la degradación de la inferencia. En cuanto a la estructura de la red, se ha identificado un mayor impacto de la resistencia de linea en redes multi-capa que involucran un gran número de neuronas en las capas ocultas, mientras que el número de capas ocultas no daña significativamente el rendimiento de la red. Finalmente, se ha indagado sobre las consecuencias de la escasa madurez en los procesos de fabricación de dispositivos RRAM sobre la inferencia. Suponiendo fallas de enclavamiento distribuidas aleatoriamente en la estructura *crossbar* y simulaciones del tipo Monte Carlo, se cuantificó su impacto y la dependencia del mismo con la resistencia de linea, tamaño del *crossbar* y el método de representación utilizado para traducir pesos sinápticos a conductancia. Finalmente, se proponen diversos métodos de mitigación de fallas de enclavamiento haciendo uso de técnicas de permutación y re-ajuste, mostrando resultados alentadores.

7 Conclusiones y próximos pasos

7.1. Contribuciones

En el presente trabajo de tesis, se discuten los principales desafíos de confiabilidad que enfrenta la introducción de nuevos materiales que permitan mantener el desarrollo sostenido de la industria microelectrónica, así como las oportunidades que brindan para el el surgimiento de alternativas superadoras frente a la microelectrónica convencional, los cuales podrían encuadrarse en los enfoques de *System on Chip* y *System in Package*. En relación al primer bloque conceptual, se estudió la dinámica de degradación paramétrica en estructuras MOS con sustratos de alta movilidad y dieléctricos *high–*κ. Acto seguido, se estudió la estadística de ruptura de dichos aislantes desde el punto de vista de la dinámica espacio-temporal de la generación de defectos. El mecanismo de ruptura fue entonces abordado como el fenómeno subyacente en las memorias de conmutación resistiva (RRAM), a fin de mejorar el entendimiento de la evolución temporal de dichos dispositivos. En cuanto al segundo bloque conceptual, se puso foco en el modelado compacto de las memorias RRAM, su integración en estructuras de *crossbar* y simulación eléctrica realista, cuantificando el impacto de fallas de enclavamiento y proponiendo técnicas de mitigación de las mismas. Sobre estas áreas generales, se realizaron las siguientes contribuciones específicas:

- Para el caso de estructuras Metal-Óxido-Semiconductor (MOS) fabricadas sobre sustratos de Germanio (llamado a reemplazar al Silicio como semiconductor en los transistores P-MOS de las futuras tecnologías CMOS) se identificó la capacidad del tratamiento de recocido en una atmósfera de H_2/N_2 (*Forming Gas Annealing*) para la pasivación selectiva de los centros de atrapamiento de carga positiva (con energías cercanas a la banda de valencia) localizados en la proximidad de la interfaz óxido/-semiconductor

- En el caso de los transistores N-MOS en las tecnologías venideras, son los semiconductores III-V los que se consideran como reemplazo del silicio. Por este motivo,

se ha estudiado el atrapamiento de carga en estructuras con dieléctricos bi-capa sobre III-V, identificando una influencia del material de sustrato en la distribución energética de las trampas presentes en el dieléctrico. Al mismo tiempo se ha identificado la dependencia de estas con el material dieléctrico utilizado para la capa interfacial.

- En estructuras con dieléctricos *high–*κ se contribuyó con evidencia experimental de la generación correlacionada de defectos, la cual ayuda a explicar las variaciones estadísticas observadas en el tiempo de ruptura dieléctrica. Para ello se realizaron experimentos de estrés a tensión constante sobre dispositivos sometidos a fluencias precisamente controladas de irradiación altamente localizada, a fin de crear densidades variables de defectos altamente controladas, y se complementó con simulaciones multi-físicas.

- Dada su potencial aplicación como dispositivo de memoria o sinapsis artificial, las dinámica temporal de la transición entre baja y alta resistencia (SET) en memorias RRAM no-volátiles basadas en dieléctricos *high–*κ (HfO_2) fue analizada mediante numerosas mediciones experimentales a diferentes tensiones. En base a su simulitud fenomenológica con el proceso de ruptura dieléctrica, se contribuye con un modelo para cuantificar la velocidad de la transición en términos de las propiedades de los materiales.

- El caso de memorias RRAM volátiles basas en materiales 2D (h-BN) es igualmente considerando. Mediante el modelado físico de la transición de SET, se contribuye a identificar a la especie atómica responsable de dicho evento, pudiendo ser estas iones de plata (Ag^+), al menos en los dispositivos considerados en esta tesis.

- Se demostró la aplicabilidad del modelo compacto denominado Cuasi-Estático de memorias RRAM para la simulación eléctrica realista de estructuras tipo *crossbar* con decenas de miles de dispositivos, destinadas a la clasificación de patrones. Se verifica que el mismo aporta un modelado preciso del dispositivo sin incurrir en altos costos computacionales.

- En base al modelo Cuasi-Estático de la memoria RRAM, se realizó un estudio exploratorio del impacto de las no idealidades de las estructuras *crossbar* (resistencia de linea, ventana resistiva, capacidad de linea, relación señal a ruido, variabilidad dispositivo a dispositivo, entre otras) sobre la precisión de clasificación lograda por un perceptrón mono-capa. Como resultado, se proponen como herramienta para el diseño de redes neuronales en *hardware* ciertos lineamientos para el dimensionamiento de *crossbars* (cantidad de dispositivos), características de los dispositivos RRAM que optimicen el funcionamiento y particionado del *crossbar* que permita minimizar el impacto de la resistencia de las lineas de interconexión. Asimismo, estos resultados son luego extendidos al caso de perceptrones multi-capa, estableciendo

la influencia del número y tamaño de las capas de neuronas ocultas en el funcionamiento de la red.

- Se evaluó el impacto de las denominadas fallas de enclavamiento sobre circuitos neuromórficos basados en RRAM. Las mismas son una consecuencia de la falta de madurez tecnológica y su influencia sobre la precisión de la clasificación de patrones depende de su número, pero también de las características constructivas del *crossbar* (resistencia de linea) y mapeo de los pesos sinápticos mediante el rango de conductancia disponible en los dispositivos RRAM. De esta forma, se proponen ciertos lineamientos para el correcto mapeo de pesos sinápticos a dispositivos RRAM a fin de mitigar el impacto de las fallas de enclavamiento. Por otro lado, se proponen 3 posibles técnicas para corregir las mismas, utilizando una permutación inteligente de las conexiones sinápticas del *crossbar*.

En síntesis, siguiendo un enfoque ascendente que va desde desde la física de degradación hacia la implementación de circuitos neurmórficos, se realizaron contribuciones relevantes en el campo de la fiabilidad de dispositivos CMOS novedosos y sus aplicaciones. A nivel de dispositivo se indagó en el atrapamiento de carga en sustratos de alta movilidad, un paso crítico para permitir el desarrollo futuro de la integración CMOS híbrida entre Ge y III-V. A su vez, se mejoró el entendimiento de la estadística de ruptura en dieléctricos *high–*κ y su similitud con la transición de SET en memorias RRAM tanto volátiles como no volátiles. Estás son de gran utilidad para la implementación de redes neuronales en *hardware*. Sobre esta linea se analizó tanto el impacto de no idealidades en el rendimiento de las mismas, y los aspectos de confiabilidad asociado a la falta de madurez tecnológica en los procesos de fabricación.

7.2. Perspectivas a futuro

En base a las problemáticas presentadas en la introducción de este trabajo de tesis y las contribuciones realizadas en el mismo, se desprenden los siguientes interrogantes abiertos, dando un panorama de los próximos problemas a abordar.

- Así como se ha tratado estadística de ruptura dieléctrica, el evento de *Soft-Breakdown* merece un tratamiento similar. Puntualmente, se espera que mediante el mismo se pueda obtener una mejor comprensión del inicio del fenómeno de ruptura. Por otro lado, podría ser de utilidad para identificar la secuencia de ruptura en dieléctricos bi-capa.

- Dada la clara influencia de la densidad de defectos en la ruptura de dieléctricos, la cual en este trabajo fue controlada mediante experimentos de irradiacion controlada, se plantea como interrogante el impacto del proceso de fabricación en la

estadística de ruptura. En otras palabras, si variables asociadas al proceso de fabricación tales como la temperatura de deposición del dieléctrico modifican la densidad de defectos en aislante, se espera encontrar una dependencia de la estadística de ruptura con dichas variables.

- Actualmente se sabe que el fenómeno físico detrás del fenómeno de ruptura dieléctrica y el SET en memorias RRAM es el mismo. Habiendo demostrado que la evolución temporal de la corriente en ambos casos puede ser descrita por el mismo modelo, mediante un reajuste cuidadoso de ciertos parámetros, se plantea como un paso siguiente el análisis del proceso de electro formado en memorias RRAM.

- La resistencia de linea en redes neuronales de *hardware* basadas en memorias RRAM es un importante limitante de su rendimiento. El desarrollo de un método de entrenamiento que tenga en cuenta este efecto de forma tal de minimizar su impacto se posiciona entonces como un problema relevante. No obstante es un desafío no menor dado su alto costo computacional y la necesidad de optar por métodos numéricos no basados en el algoritmo de *back-propagation*.

- Al igual que las memorias RRAM, las memorias de cambio de fase (*Phase Change Memories*, PCM) se posicionan como otro candidato para aplicaciones neuromórficas. En este contexto, su modelado compacto y utilización en simulaciones eléctricas realistas constituye otro desafío abierto para el desarrollo de estos sistemas. Sin embargo, la variabilidad de las mismas, y desafíos de su caracterización eléctrica, suponen serios obstáculos para su implementación.

A

Lista de publicaciones centrales

A.1. Artículos en revistas con referato

- F. L. Aguirre, S. M. Pazos, F. Palumbo, M. Antoni, J. Suñé y E. A. Miranda, "Assessment and Improvement of the Pattern Recognition Performance of Memdiode-Based Cross-Point Arrays with Randomly Distributed Stuck-at-Faults," *Electronics*, vol. 10, n.º 19, pág. 2427, oct. de 2021. DOI: 10.3390/electronics10192427
- F. L. Aguirre, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "SPICE Simulation of RRAM-Based Crosspoint Arrays Using the Dynamic Memdiode Model," *Frontiers in Physics*, vol. 9, pág. 548, 2021. DOI: 10.3389/FPHY.2021.735021
- F. L. Aguirre, N. M. Gomez, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "Minimization of the Line Resistance Impact on Memdiode-Based Simulations of Multilayer Perceptron Arrays Applied to Pattern Recognition," *Journal of Low Power Electronics and Applications 2021*, vol. 11, n.º 1, pág. 9, feb. de 2021. DOI: 10.3390/JLPEA11010009
- F. L. Aguirre, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "Application of the Quasi-Static Memdiode Model in Cross-Point Arrays for Large Dataset Pattern Recognition," *IEEE Access*, vol. 8, págs. 1-1, 2020. DOI: 10.1109/ACCESS.2020.3035638
- F. L. Aguirre, A. Rodriguez-Fernandez, S. M. Pazos, J. Sune, E. Miranda y F. Palumbo, "Study on the Connection Between the Set Transient in RRAMs and the Progressive Breakdown of Thin Oxides," *IEEE Transactions on Electron Devices*, vol. 66, n.º 8, págs. 1-7, 2019. DOI: 10.1109/ted.2019.2922555
- F. Palumbo, C. Wen, S. Lombardo, S. Pazos, F. Aguirre y col., "A Review on Dielectric Breakdown in Thin Dielectrics: Silicon Dioxide, High- k , and Layered Dielectrics," *Advanced Functional Materials*, pág. 1 900 657, abr. de 2019. DOI: 10.1002/adfm.201900657

- F. L. Aguirre, S. M. Pazos, F. Palumbo, S. Fadida, R. Winter y M. Eizenberg, "Effect of forming gas annealing on the degradation properties of Ge-based MOS stacks," *Journal of Applied Physics*, vol. 123, n.º 13, pág. 134 103, abr. de 2018. DOI: 10.1063/1.5018193

A.2. Artículos en *proceedings* de conferencias indexadas

- F. L. Aguirre, N. Gomez, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "Line Resistance Impact in Memristor-based Multi Layer Perceptron for Pattern Recognition," en *2021 5th Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA), Aceptado para publicación*, 2021

- F. L. Aguirre, F. Palumbo y P. Julian, "Piecewise-linear Modelling of CMOS Gates Propagation Delay as a Function of PVT Variations and Aging," Institute of Electrical y Electronics Engineers (IEEE), abr. de 2021, págs. 25-31. DOI: 10.1109/cae51562.2021.9397560

- F. L. Aguirre, A. Padovani, A. Ranjan, N. Raghavan, N. Vega y col., "Spatio-Temporal Defect Generation Process in Irradiated HfO_2 MOS Stacks: Correlated versus Uncorrelated Mechanisms," en *2019 IEEE International Reliability Physics Symposium (IRPS)*, Monterey: IEEE, mar. de 2019, págs. 13-14. DOI: 10.1109/IRPS.2019.8720539

- F. L. Aguirre, S. M. Pazos, F. Palumbo, S. Fadida, R. Winter y M. Eizenberg, "Impact of forming gas annealing on the degradation dynamics of Ge-Based MOS stacks," en *2018 IEEE International Reliability Physics Symposium*, Burlingame, EE.UU.: IEEE, mar. de 2018, págs. 21-25. DOI: 10.1109/CAMTA.2017.8058136, accepted for publication

- F. L. Aguirre, S. M. Pazos, F. Palumbo, I. Krylov y M. Eizenberg, "Substrate influence on the behavior of capacitance hysteresis of III-V bilayered MOS stacks," en *2017 32nd Symposium on Microelectronics Technology and Devices (SBMicro)*, Fortaleza, Brasil: IEEE, ago. de 2017, págs. 1-4. DOI: 10.1109/SBMicro.2017.8112972

- F. L. Aguirre, S. M. Pazos y F. Palumbo, "Experimental study of progressive breakdown in different conductance states of resistive switching structures," en *2017 1st Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA)*, Bariloche, Argentina: IEEE, feb. de 2017, págs. 1-4. DOI: 10.1109/PRIME-LA.2017.7899167

A.3. Participación en publicaciones relacionadas a la línea de trabajo

- S. M. Pazos, F. L. Aguirre, F. Palumbo y F. Silveira, "Hot-carrier-injection resilient RF power amplifier using adaptive bias," *Microelectronics Reliability*, vol. 114, pág. 113 912, oct. de 2020. DOI: 10.1016/j.microrel.2020.113912
- S. M. Pazos, S. Boyeras Baldomá, F. L. Aguirre, I. Krylov, M. Eizenberg y F. Palumbo, "Impact of bilayered oxide stacks on the breakdown transients of MOS devices: An experimental study," *Journal of Applied Physics*, vol. 127, n.º 17, pág. 174 101, mayo de 2020. DOI: 10.1063/1.5138922
- S. Boyeras Baldomá, S. M. Pazos, F. L. Aguirre y F. R. Palumbo, "Breakdown transients in high-k multilayered MOS stacks: Role of the oxide-oxide thermal boundary resistance," *Journal of Applied Physics*, vol. 128, n.º 3, pág. 034 103, jul. de 2020. DOI: 10.1063/5.0012918
- S. Pazos, F. Aguirre, F. Palumbo y F. Silveira, "Reliability-aware design space exploration for fully integrated RF CMOS PA," *IEEE Transactions on Device and Materials Reliability*, 2019. DOI: 10.1109/TDMR.2019.2957489
- S. Boyeras, S. M. Pazos, F. L. Aguirre, H. Giannetta, C. Delgado y F. Palumbo, "Progressive breakdown on bi-layered gate oxide stacks," en *SBMicro 2019 - 34th Symposium on Microelectronics Technology and Devices*, Institute of Electrical y Electronics Engineers Inc., ago. de 2019. DOI: 10.1109/SBMicro.2019.8919480
- A. Fontana, S. Pazos, F. Aguirre, N. Vega, N. Muller y col., "Pulse quenching and charge sharing effects on heavy-ion microbeam induced ASET in a full-custom CMOS OpAmp," *IEEE Transactions on Nuclear Science*, págs. 1-1, 2019. DOI: 10.1109/TNS.2019.2908174
- S. M. Pazos, F. L. Aguirre, K. Tang, P. McIntyre y F. Palumbo, "Lack of correlation between C-V hysteresis and capacitance frequency dispersion in accumulation of metal gate/high-k/n-InGaAs MOS stacks," *Journal of Applied Physics*, vol. 124, n.º 22, pág. 224 102, 2018. DOI: 10.1063/1.5031025
- S. M. Pazos, F. L. Aguirre, F. Palumbo y F. Silveira, "Performance-reliability tradeoffs in short range RF power amplifier design," *Microelectronics Reliability*, sep. de 2018. DOI: 10.1016/J.MICROREL.2018.06.089
- F. Palumbo, F. L. Aguirre, S. M. Pazos, I. Krylov, R. Winter y M. Eizenberg, "Influence of the spatial distribution of border traps in the capacitance frequency dispersion of Al_2O_3/InGaAs," *Solid-State Electronics*, 2018. DOI: 10.1016/J.SSE.2018.07.006

- A. Fontana, S. M. Pazos, F. L. Aguirre, F. Palumbo, N. Vega y col., "Heavy Ion Microbeam Experimental Study of ASET on a Full-Custom CMOS OpAmp," en *2018 31st Symposium on Integrated Circuits and Systems Design (SBCCI)*, IEEE, ago. de 2018, págs. 1-5. DOI: 10.1109/SBCCI.2018.8533232

- S. M. Pazos, F. L. Aguirre, S. Lombardo, E. Miranda y F. Palumbo, "Experimental Study of Progressive Breakdown in Different Conductance States of Resistive Switching Structures," en *China RRAM International Workshop 2017*, Soochow University, China, jun. de 2017

- A. Fontana, S. M. Pazos, F. L. Aguirre y F. Palumbo, "Automatic ASET sensitivity evaluation of a custom-designed 180nm CMOS technology operational amplifier," en *2017 Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA)*, Buenos Aires, Argentina: IEEE, jul. de 2017, págs. 21-25. DOI: 10.1109/CAMTA.2017.8058136

- S. M. Pazos, F. L. Aguirre y F. Palumbo, "Charge trapping effects on Metal-Gate/High-k/III-V MOS devices assessed through C-V hysteresis," en *2017 Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA)*, Buenos Aires, Argentina: IEEE, jul. de 2017, págs. 21-25. DOI: 10.1109/CAMTA.2017.8058135

- S. M. Pazos, F. Palumbo y F. L. Aguirre, "Analysis and comparison of the CV-Dispersion of high-k, bi-layered MOS InGaAs/InP stacks," en *2017 1st Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA)*, Bariloche, Argentina: IEEE, feb. de 2017, págs. 1-4. DOI: 10.1109/PRIME-LA.2017.7899166

- F. Palumbo, S. M. Pazos, F. L. Aguirre, R. Winter, I. Krylov y M. Eizenberg, "Temperature dependence of trapping effects in metal gates/Al_2O_3/InGaAs stacks," *Solid-State Electronics*, vol. 132, págs. 12-18, 2017. DOI: http://dx.doi.org/10.1016/j.sse.2017.03.009

- S. M. Pazos, F. Palumbo y F. L. Aguirre, "Analysis and comparison of the CV-Dispersion of high-k, bi-layered MOS InGaAs/InP stacks," en *2017 1st Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA)*, Bariloche, Argentina: IEEE, feb. de 2017, págs. 1-4. DOI: 10.1109/PRIME-LA.2017.7899166

B

Modelos, bases de datos e información complementaria a la simulación SPICE de redes neuronales

B.1. Modelo lineal de *Cross-bar* de memristores

Si se considera que no se modifica el estado de memoria almacenado en los memristores, se puede derivar un equivalente eléctrico lineal del CPA considerando el circuito esquemático asociado a cada celda de memristor presentado en la Fig. 6.5a y teniendo en cuenta la 2^{nd} ley de Kirchoff (Ley de Kirchoff de Corrientes) en los terminales de cada uno de los memristores del CPA (véase la Fig. 6.5b), las cuales tendrán la forma de alguna de las Ec. B.1)-(B.6, dependiendo de la localización del dispositivo en el CPA. En las Ecs. B.1-B.6, G_L es la conductancia de linea ($1/R_L$), $G^{BL}_{i,in}$=$G^{WL}_{i,in}$ son las conductancias de entrada a las *wordline* y *bitline* (resistencia en los terminales de las WL y BL) ($G^{WL}_{i,in}$=$1/R^{WL}_{i,in}$ y $G^{BL}_{i,in}$=$1/R^{BL}_{i,in}$), $V^{WL}_{i,app}$ son las tensiones aplicadas en los terminales de las WL (correspondiente a las imágenes de test) y $V^{BL}_{j,app}$ está puesto a tierra a través de un resistor de censado.

$$(WL,(i,j)): \;\; G_L\left(V^{WL}_{i,j}-V^{WL}_{i,j-1}\right)-G_{i,j}\left(V^{BL}_{i,j}-V^{WL}_{i,j}\right)-G_L\left(V^{WL}_{i,j+1}-V^{WL}_{i,j}\right)=0 \quad \text{(B.1)}$$

$$(WL,j=1): \;\; G^{WL}_{i,in}\left(V^{WL}_{i,1}-V^{WL}_{i,app}\right)-G_{i,j}\left(V^{BL}_{i,1}-V^{WL}_{i,1}\right)-G_L\left(V^{WL}_{i,2}-V^{WL}_{i,1}\right)=0 \quad \text{(B.2)}$$

$$(WL,j=n): \;\; G_W\left(V^{WL}_{i,n}-V^{WL}_{i,n-1}\right)-G_{i,n}\left(V^{BL}_{i,n}-V^{WL}_{i,n}\right)=0 \quad \text{(B.3)}$$

$$(BL,(i,j)): \;\; G_L\left(V^{BL}_{i+1,j}-V^{BL}_{i,j}\right)-G_{i,j}\left(V^{BL}_{i,j}-V^{WL}_{i,j}\right)-G_L\left(V^{BL}_{i,j}-V^{BL}_{i-1,j}\right)=0 \quad \text{(B.4)}$$

$$(BL,i=m): \;\; G^{BL}_{j,in}\left(V^{BL}_{j,app}-V^{BL}_{m,j}\right)-G_{m,j}\left(V^{WL}_{i,1}-V^{BL}_{i,1}\right)-G_L\left(V^{BL}_{i,2}-V^{BL}_{i,1}\right)=0 \quad \text{(B.5)}$$

$$(BL, i = 1): \;\; G_W \left(V_{2,j}^{BL} - V_{1,j}^{BL} \right) - G_{i,j} \left(V_{1,j}^{BL} - V_{1,j}^{WL} \right) = 0 \tag{B.6}$$

Esto resulta en un sistema de $2mn$ ecuaciones acopladas, con $2mn$ tensiones desconocidas correspondientes a las tensiones en los nodos sobre las WL ($V_{WL} = \left[V_{1,1}^{WL}, V_{1,2}^{WL}, ..., V_{1,n}^{WL}, V_{2,1}^{WL}, ..., V_{n,m}^{WL}\right]^T$) y BL ($V_{BL} = \left[V_{1,1}^{BL}, V_{1,2}^{BL}, ..., V_{1,n}^{BL}, V_{2,1}^{BL}, ..., V_{n,m}^{BL}\right]^T$). Definiendo los vectores columna E_{WL} y E_{BL} como $\left[G_{1,in}^{WL} V_{1,in}^{WL}, 0, ..., G_{2,in}^{WL} V_{2,in}^{WL}, 0, ..., G_{m,in}^{WL} V_{m,in}^{WL}\right]$ y $\left[G_{1,in}^{BL} V_{1,in}^{BL}, 0, ..., G_{2,in}^{BL} V_{2,in}^{BL}, 0, ..., G_{n,in}^{BL} V_{n,in}^{BL}\right]$ respectivamente, las Ecs. B.1-B.6 pueden ser representadas siguiendo una notación matricial como se muestra en la Ec. B.7:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_{WL} \\ V_{BL} \end{bmatrix} = \begin{bmatrix} E_{WL} \\ E_{BL} \end{bmatrix} \tag{B.7}$$

donde todos los términos (A, B, C y D) son matrices de $m \times n$. Para una descripción detallada de la estructura de estas matrices el lector es referido a la Ref. [301]. Bajo esta formulación, la solución a este sistema da como resultado un vector fila ($1 \times n$) de corrientes definido como $I_{Out} = V_{n,j}^{BL} G_L$, con $1 \leq j \leq m$.

B.2. Modelo SPICE del Memdiodo Cuasi-Estático (*Quasi-Static Memdiode*, QMM)

Tabla B.1: Modelo SPICE del *memdiodo*: `S(x)` y `R(x)` son las funciones logísticas Γ^+(V) y Γ^-(V). `A(x)` y `RS(x)` representan a los parámetros α y R del memdiodo, los cuales son una función del estado memorizado. El estado memorizado λ es indicado por `V(H)`, con `H0` el estado inicial. Los parámetros que modelizan la transición HRS-LRS son `etaset`, `etares`, `vset` y `vres` para η^+, η^-, V^+ y $V-$, respectivamente. Se utilizan fuentes de corriente controladas por tensión para implementar la Ec. 6.1. (`GD` y el resistor `RS`) y la Ec. 6.2 (`GH` y el capacitor `CH`). Los diodos en anti-paralelo se modelan mediante la fuente de corriente controlada `GD` en el sub-circuito. `beta` define si la conducción es simétrica (igual para tensiones positivas y negativas, `beta`=0.5) o no (`beta`≠0.5). La Ec. 6.1 corresponde al caso de `beta`=1 pero es simetrizada utilizando la función de valor absoluto. El modelo está escrito utilizando la sintaxis de H-SPICE

```
.subckt memdiode p n
   .param H0=0 CH0=1e-4 beta=0.5
   *Transition parameters
   .param etaset=10.5 vset=0.78 etares=7.2 vres=-0.79
   *I-V parameters
   .param imax=6.06e-3 imin=1.16e-4 alphamax=3.5 alphamin=5.6 rsmax=47
   + rsmin=47
   *Auxiliary functions
   .param I0(x)='imax*x+imin*(1-x)'
   .param A(x)='alphamax*x+alphamin*(1-x)'
   .param RS(x)='rsmax*x+rsmin*(1-x)'
   .param R(x)='1/(1+exp(-etares*(x-vres)))'
   .param S(x)='1/(1+exp(-etaset*(x-vset)))'
   *********H-V*********
   GH gnd! H cur='min(R(V(p,n)),max(S(V(p,n)),V(H)))'
   Rpar gnd! H R=1
   CH H gnd! C='CH0' IC='H0'
   *********I-V*********
   RS p D R='RS(V(H))'
   GD D n cur='I0(V(H))*(exp(beta*A(V(H))*V(D,n))-
   + exp(-(1-beta)*A(V(H))*V(D,n)))'
.ends memdiode
```

B.3. Base de datos utilizadas en los análisis

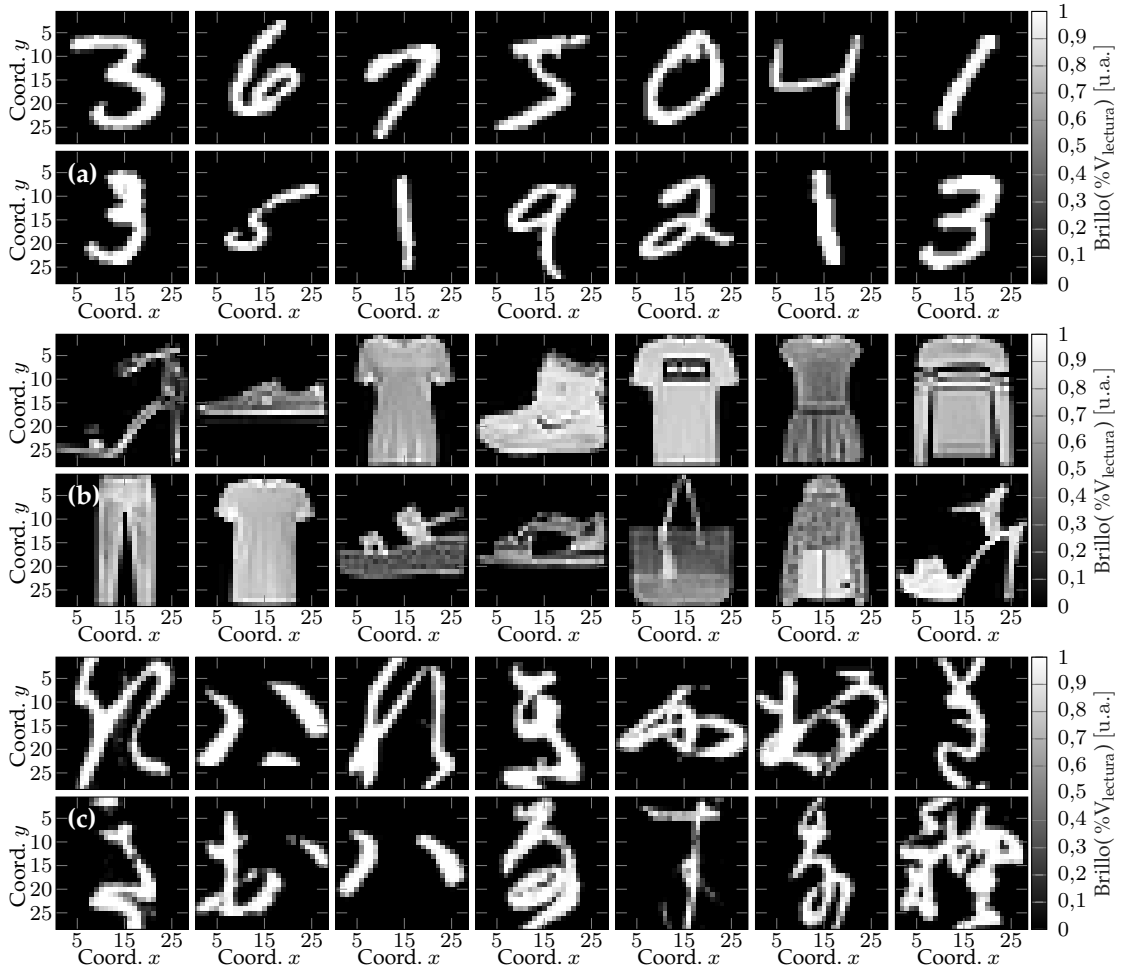


Figura B.1: Imágenes de ejemplo de las tres bases de datos estilo MNIST consideradas en este trabajo. Para todos los casos, las imágenes tienen una resolución de 28×28 píxeles. El brillo de cada píxel está codificado en 256 niveles entre 0 (completamente apagado, y por ende negro) y 1 (completamente prendido y por ende blanco). (a) Base de datos MNIST de dígitos manuscritos. (b) Base de datos MNIST-F [278] de artículos de vestir. (c) Base de datos MNIST-K [279] de ideogramas Kanji japoneses manuscritos.

B.4. Validación de los algoritmos de entrenamiento

B.4.1. Validación cruzada de k-iteraciones

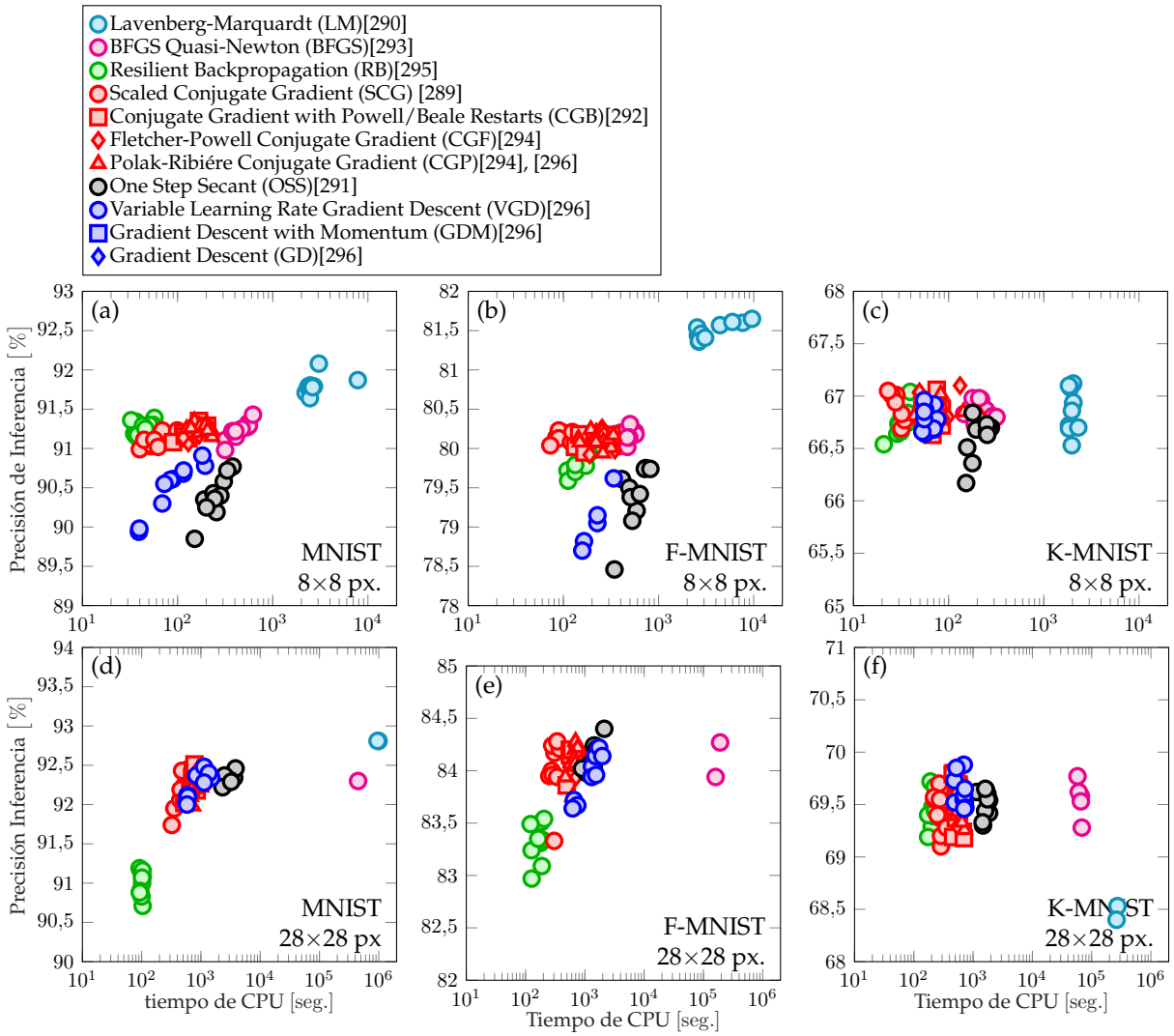


Figura B.2: Validación cruzada de 5 grupos (*5-fold cross-validation*) y 10 repeticiones, para los 11 algoritmos de aprendizaje considerados [289]-[296]. La precisión de inferencia obtenida se grafica en función del tiempo de CPU requerido por el algoritmo, para las 3 bases de datos (MNIST, MNIST-F y MNIST-K) y dos resoluciones diferentes (imágenes de 8×8 y 28×28 px.). La precisión promedio de cada algoritmo se reporta en el encabezado de las Tablas B.2-B.7 para cada par Base de Datos - Resolución: imágenes de 8×8 px. de las bases (a) MNIST, (b) MNIST-F y (c) MNIST-K, e imágenes de 28×28 px. de las bases (d) MNIST, (e) MNIST-F y (f) MNIST-K. Si bien LM muestra la mayor precisión, es también el más lento, especialmente para redes de gran tamaño, como las requeridas para imágenes de 28×28 px. Los test de significación estadística reportados en las Tablas B.2-B.7 indican con un 95 % de confianza que para de $|z|$ >1,96 los valores de precisión obtenidos son estadísticamente diferentes (celdas sombreadas en gris). Asumiendo una relación de compromiso entre precisión y tiempo de aprendizaje, se elige SCG por sobre los demás, ya que la diferencia de precisión con LM no es estadísticamente relevante (las diferencias podrían deberse a fluctuaciones en los datos de entrada).

B.4.2. *Tests* de significación estadística

Tabla B.2: $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST, codificada en 8×8 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %)

	LM	**BFGS**	**RS**	**SCG**	**CGB**	**CGF**	**CGP**	**OSS**	**VGD**	**GDM**	**GD**
	91.80 %	91.22 %	91.27 %	91.09 %	91.20 %	91.22 %	91.24 %	90.39 %	90.51 %	70.41 %	70.14 %
LM											
BFGS	1.47										
RS	1.34	0.13									
SCG	1.80	0.33	0.46								
CGB	1.52	0.05	0.18	0.28							
CGF	1.48	0.01	0.14	0.32	0.04						
CGP	1.42	0.05	0.08	0.38	0.10	0.06					
OSS	3.51	2.04	2.17	1.71	1.99	2.03	2.09				
VGD	3.23	1.76	1.89	1.43	1.71	1.75	1.81	0.28			
GDM	38.65	37.38	37.49	37.10	37.34	37.37	37.43	35.59	35.84		
GD	39.03	37.76	37.87	37.48	37.72	37.75	37.81	35.98	36.23	0.41	

Tabla B.3: $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-F, codificada en 8×8 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %)

	LM	**BFGS**	**RS**	**SCG**	**CGB**	**CGF**	**CGP**	**OSS**	**VGD**	**GDM**	**GD**
	81.50 %	80.14 %	79.80 %	80.15 %	80.10 %	80.11 %	80.10 %	79.39 %	78.82 %	62.64 %	63.05 %
LM											
BFGS	2.44										
RS	3.04	0.60									
SCG	2.42	0.02	0.62								
CGB	2.51	0.07	0.54	0.09							
CGF	2.50	0.05	0.55	0.08	0.01						
CGP	2.52	0.08	0.52	0.10	0.01	0.02					
OSS	3.77	1.32	0.72	1.34	1.26	1.27	1.25				
VGD	4.75	2.31	1.71	2.33	2.24	2.25	2.23	0.98			
GDM	29.72	27.38	26.80	27.40	27.32	27.33	27.30	26.10	25.15		
GD	29.14	26.79	26.21	26.81	26.73	26.74	26.72	25.51	24.56	0.60	

Tabla B.4: $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-K, codificada en 8×8 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %)

	LM	**BFGS**	**RS**	**SCG**	**CGB**	**CGF**	**CGP**	**OSS**	**VGD**	**GDM**	**GD**
	66.87 %	66.87 %	66.75 %	66.90 %	66.83 %	66.92 %	66.83 %	66.60 %	66.78 %	41.71 %	41.74 %
LM											
BFGS	0.00										
RS	0.17	0.17									
SCG	0.05	0.05	0.22								
CGB	0.05	0.05	0.12	0.09							
CGF	0.08	0.08	0.25	0.03	0.12						
CGP	0.06	0.06	0.11	0.11	0.01	0.14					
OSS	0.41	0.41	0.23	0.45	0.36	0.48	0.34				
VGD	0.13	0.13	0.05	0.17	0.08	0.20	0.06	0.28			
GDM	35.71	35.71	35.55	35.76	35.67	35.79	35.65	35.32	35.59		
GD	35.67	35.67	35.50	35.71	35.62	35.74	35.61	35.28	35.55	0.05	

Tabla B.5: $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST, codificada en 28×28 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %)

	LM	**BFGS**	**RS**	**SCG**	**CGB**	**CGF**	**CGP**	**OSS**	**VGD**	**GDM**	**GD**
	92.81 %	92.30 %	91.00 %	92.11 %	92.30 %	92.31 %	92.23 %	92.32 %	92.22 %	84.97 %	84.97 %
LM											
BFGS	1.37										
RS	4.70	3.33									
SCG	1.87	0.50	2.84								
CGB	1.36	0.01	3.34	0.51							
CGF	1.34	0.03	3.36	0.53	0.02						
CGP	1.55	0.18	3.16	0.32	0.19	0.21					
OSS	1.32	0.06	3.39	0.55	0.05	0.02	0.23				
VGD	1.58	0.20	3.13	0.29	0.21	0.24	0.03	0.26			
GDM	17.63	16.32	13.10	15.84	16.33	16.35	16.15	16.38	16.13		
GD	17.64	16.32	13.10	15.85	16.34	16.36	16.15	16.38	16.13	0.00	

Tabla B.6: $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-F, codificada en 28×28 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %)

	LM	**BFGS**	**RS**	**SCG**	**CGB**	**CGF**	**CGP**	**OSS**	**VGD**	**GDM**	**GD**
	83.99 %	84.10 %	75.98 %	84.00 %	84.10 %	84.12 %	84.08 %	84.13 %	83.94 %	72.03 %	72.01 %
LM											
BFGS	0.22										
RS	14.15	14.37									
SCG	0.03	0.20	14.18								
CGB	0.21	0.01	14.36	0.19							
CGF	0.25	0.03	14.40	0.23	0.04						
CGP	0.18	0.04	14.33	0.15	0.03	0.07					
OSS	0.28	0.05	14.42	0.25	0.06	0.02	0.09				
VGD	0.09	0.31	14.06	0.12	0.30	0.35	0.27	0.37			
GDM	20.41	20.63	6.37	20.44	20.62	20.66	20.59	20.68	20.32		
GD	20.44	20.66	6.40	20.47	20.65	20.69	20.62	20.71	20.36	0.03	

Tabla B.7: $|z|$-scores para el test de dos proporciones [343] (Véase la Ec. B.8), para la base de datos MNIST-K, codificada en 28×28 px. H_0 : Las proporciones son las mismas, H_1 Las proporciones son diferentes. H_0 es rechazada para $|z|$ >1.96 (intervalo de confianza del 95 %)

	LM	**BFGS**	**RS**	**SCG**	**CGB**	**CGF**	**CGP**	**OSS**	**VGD**	**GDM**	**GD**
	68.47 %	69.56 %	69.43 %	69.41 %	69.48 %	69.54 %	69.56 %	69.51 %	69.62 %	56.64 %	56.66 %
LM											
BFGS	1.67										
RS	1.47	0.19									
SCG	1.45	0.22	0.03								
CGB	1.55	0.12	0.07	0.10							
CGF	1.64	0.03	0.16	0.19	0.09						
CGP	1.67	0.00	0.20	0.23	0.12	0.03					
OSS	1.60	0.07	0.12	0.15	0.05	0.04	0.08				
VGD	1.77	0.10	0.29	0.32	0.22	0.13	0.10	0.17			
GDM	17.28	18.93	18.74	18.71	18.81	18.90	18.94	18.86	19.03		
GD	17.24	18.89	18.70	18.68	18.78	18.87	18.90	18.82	18.99	0.04	

$$z = \frac{p_1 - p_2}{\sqrt{\frac{2p(1-p))}{n}}}, \quad p = \frac{p_1 + p_2}{2} \tag{B.8}$$

B.5. Métricas adicionales para el rendimiento de inferencia

B.5.1. Base de datos MNIST

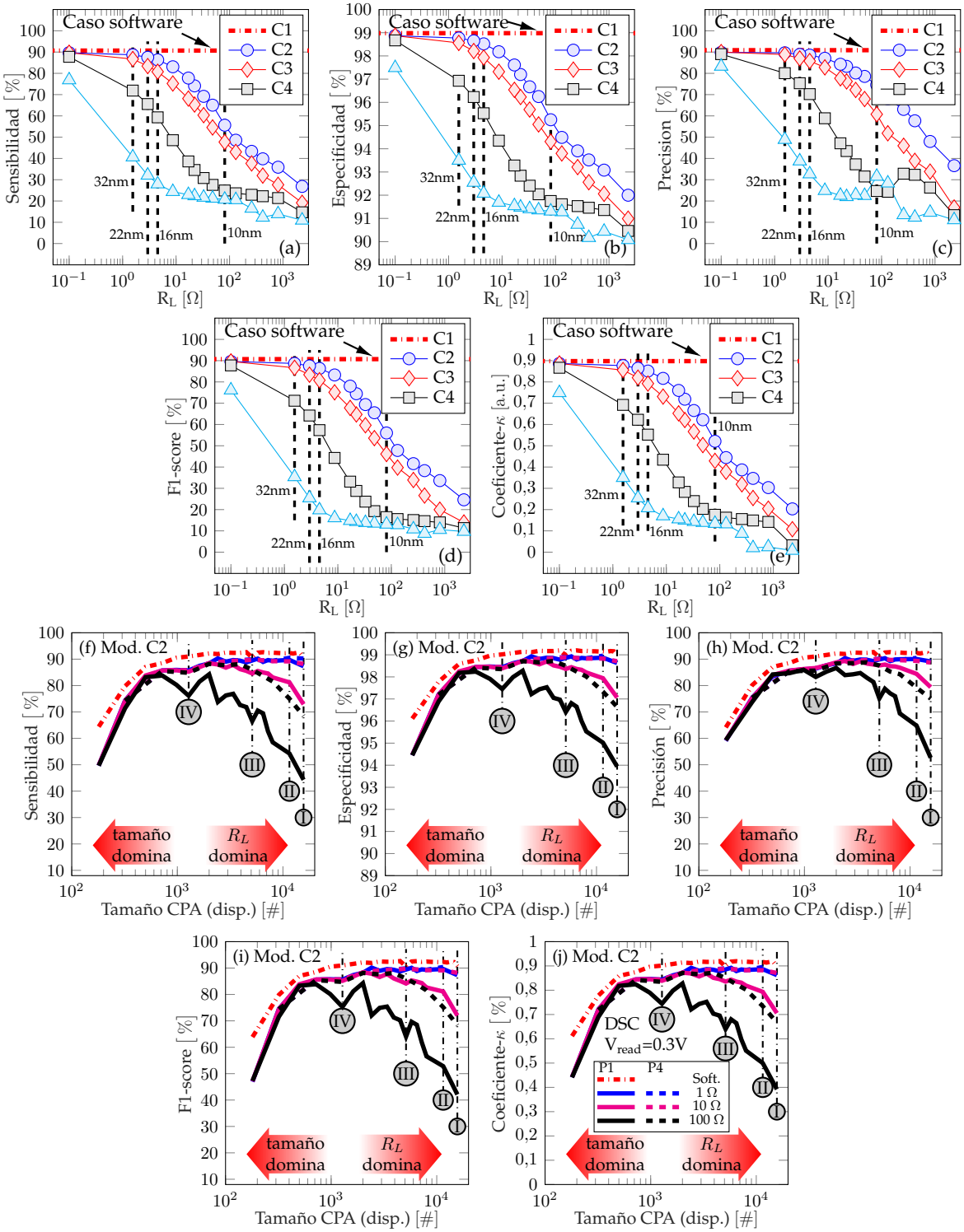


Figura B.3: Métricas de inferencia complementarias para el reconocimiento de imágenes de la base de datos del MNIST, en función de R_L y tamaño del CPA (# dispositivos): (a) y (f) Sensibilidad (*Recall*), (b) y (g) Especificidad, (c) y (h) Precisión, (d) y (i) F1-*Score* y (e) y (j) coeficiente-κ.

B.5.2. Base de datos F-MNIST

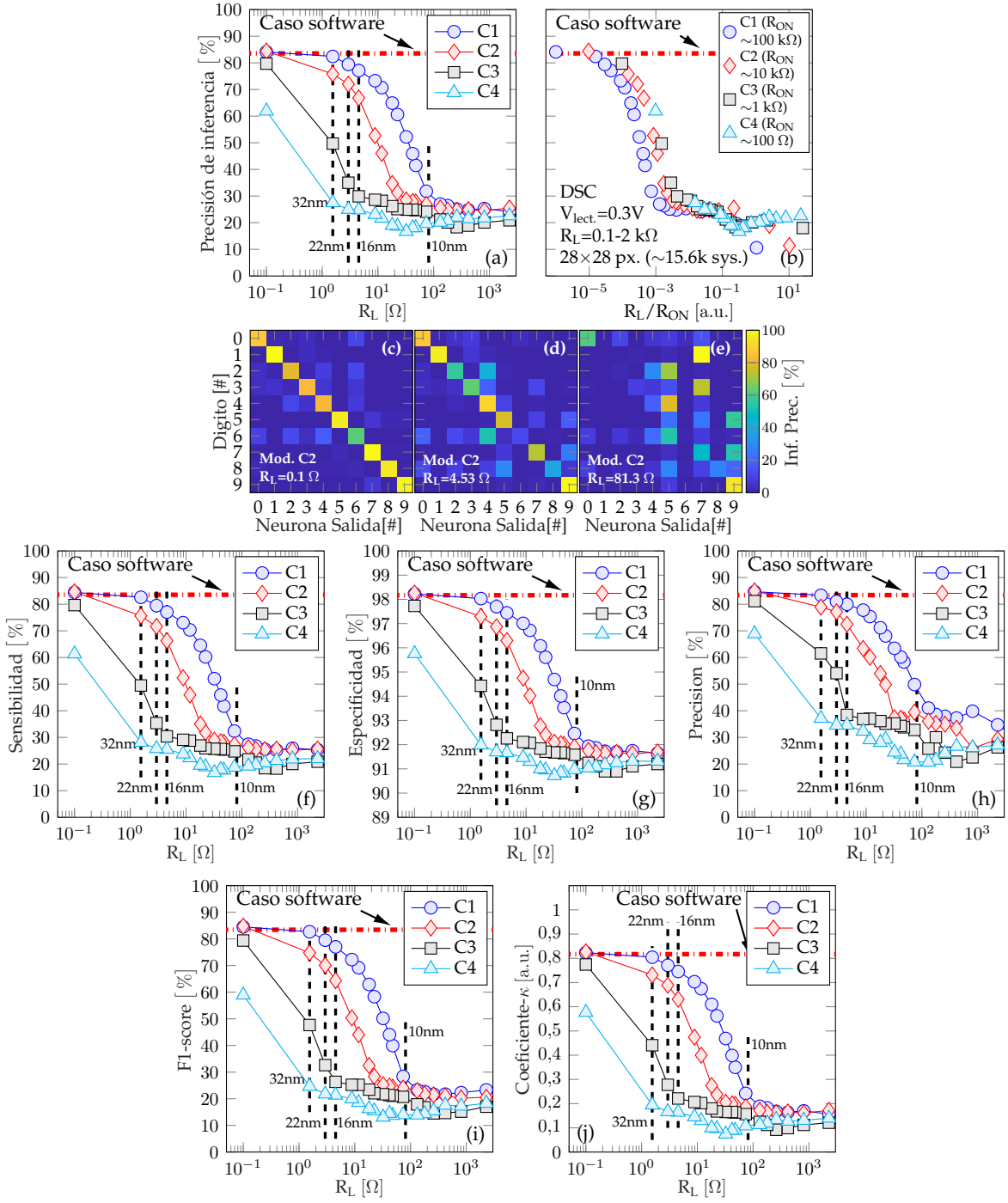


Figura B.4: (a) Impacto de la resistencia de linea (R_L) en la precisión de inferencia para la base de datos MNIST-F, considerando los ajustes $C1$-$C4$ del QMM. (b) La precisión de inferencia se grafica en función del cociente R_L/R_{ON} mostrando una tendencia unificada entre todos los ajustes. Matrices de confusión para el ajuste $C2$ con R_L igual a (c) 0,1 Ω, (d) 4,53 Ω (16 nm [233]) y (e) 81,3 Ω (10 nm [306]). En todos los casos, el CPA es conectado por ambos lados (DSC) y las imagenes no se re-escalan (resolución de 28×28 px.). Otras metricas de inferencia incluyen: (f) Sensibilidad, (g) Especificidad, (h) Precisión, (i), F1-*Score* y (j) coeficiente–κ.

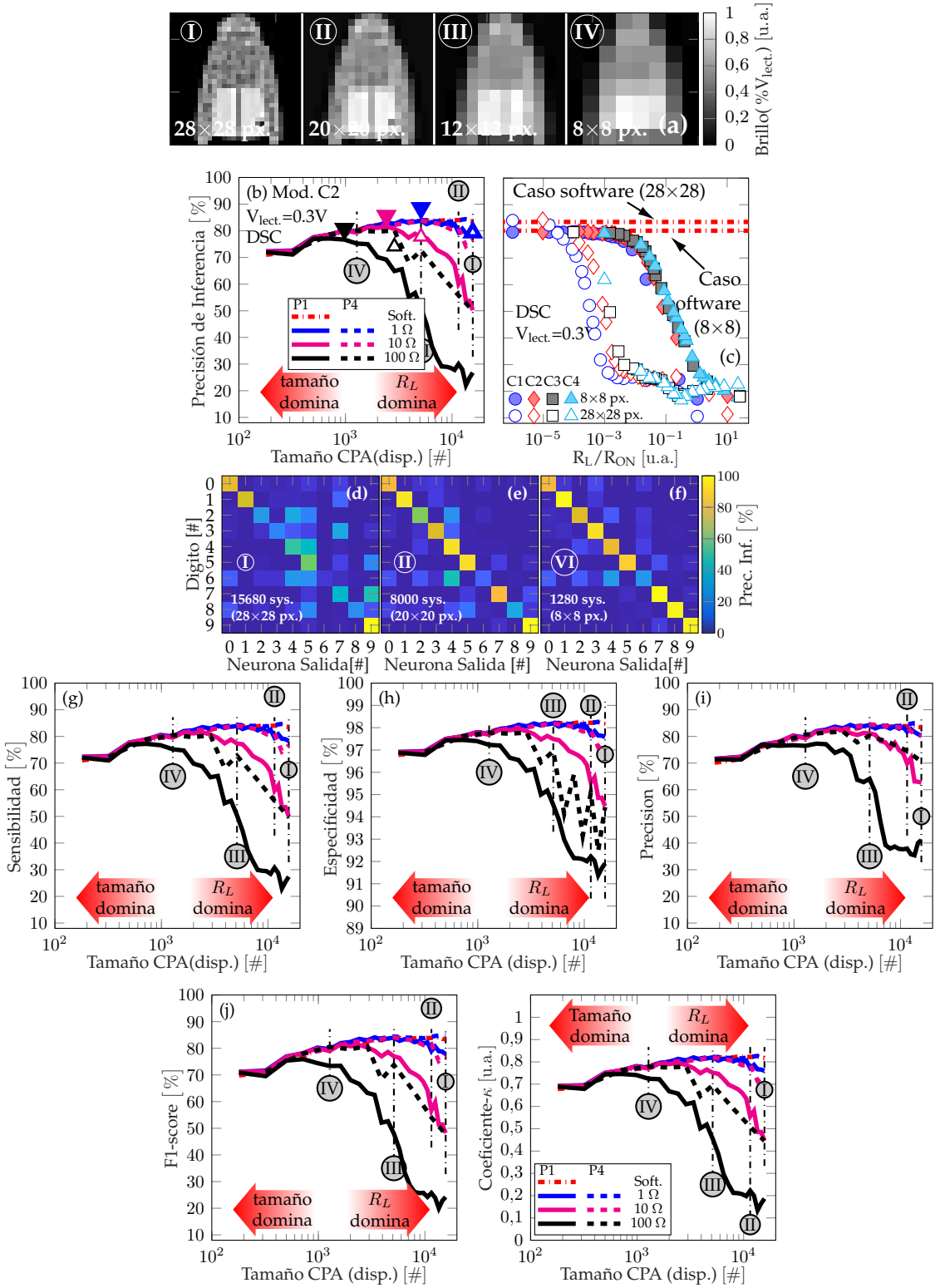


Figura B.5: (a) Pérdida de legibilidad de las imágenes de la base de datos MNIST-F para resolución decreciente de 28×28 px. (caso I) a 8×8 px. (caso IV). (b) Impacto del tamaño del CPA (cantidad de dispositivos) sobre la precisión de inferencia para el ajuste $C2$, y R_L barrido entre 1 y 100 Ω. Dos esquemas de particionado diferentes fueron utilizados: P1 indica *arrays* no particionados y P4 que cada CPA fue dividido en 4 sub-*arrays*. Los símbolos triangulares indican los puntos de tamaño máximo y precisión máxima, mostrando que los CPA particionados permiten mayor precisión en *arrays* más grandes. (c) Precisión de inferencia en función del cociente R_L/R_{ON} para imágenes de 28×28 (dos *arrays* de 784×10 arrays, ∼15.6k disp., símbolos vacíos) y 8×8 (dos 64×10 *arrays*, ∼1.2k disp., símbolos llenos). Matrices de confusión para el ajuste $C2$ y R_L=10 Ω para imágenes de (d) 28×28 px., (e) 20×20 px. y (f) 8×8 px. Otras métricas para el el reconocimiento de imágenes incluyen (g) sensibilidad, (h) especificidad, (i) precisión, (j) F1-Score y (k) coeficiente-κ.

B.5.3. Base de datos K-MNIST

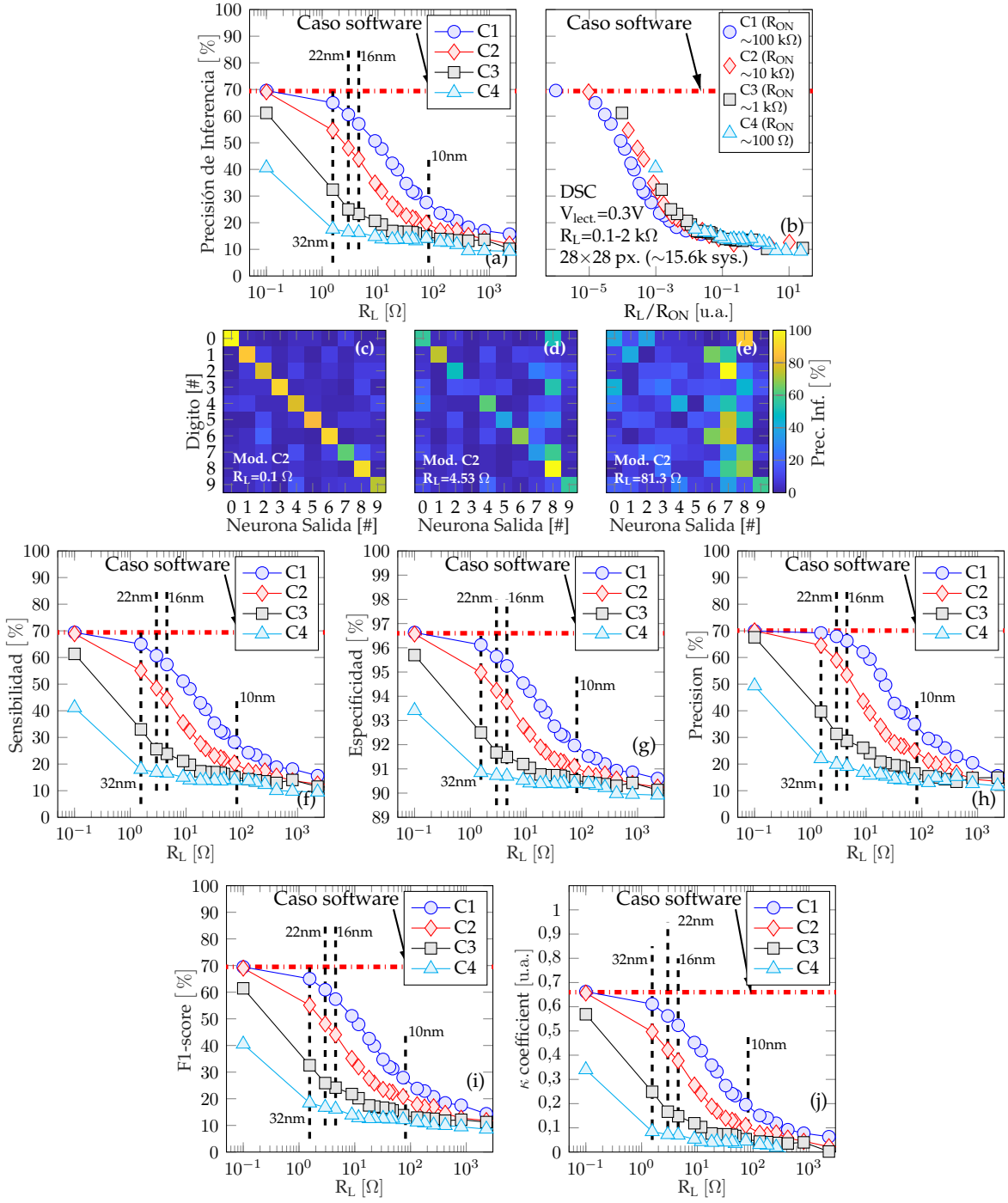


Figura B.6: (a) Impacto de la resistencia de linea (R_L) en la precisión de inferencia para la base de datos MNIST-K, considerando los ajustes $C1$-$C4$ del QMM. (b) La precisión de inferencia se grafica en función del cociente R_L/R_{ON} mostrando una tendencia unificada entre todos los ajustes. Matrices de confusión para el ajuste $C2$ con R_L igual a (c) 0,1 Ω, (d) 4,53 Ω (16 nm [233]) y (e) 81,3 Ω (10 nm [306]). En todos los casos, el CPA es conectado por ambos lados (DSC) y las imagenes no se re-escalan (resolución de 28×28 px.). Otras metricas de inferencia incluyen: (f) Sensibilidad, (g) Especificidad, (h) Precisión, (i), F1-*Score* y (j) coeficiente–κ.

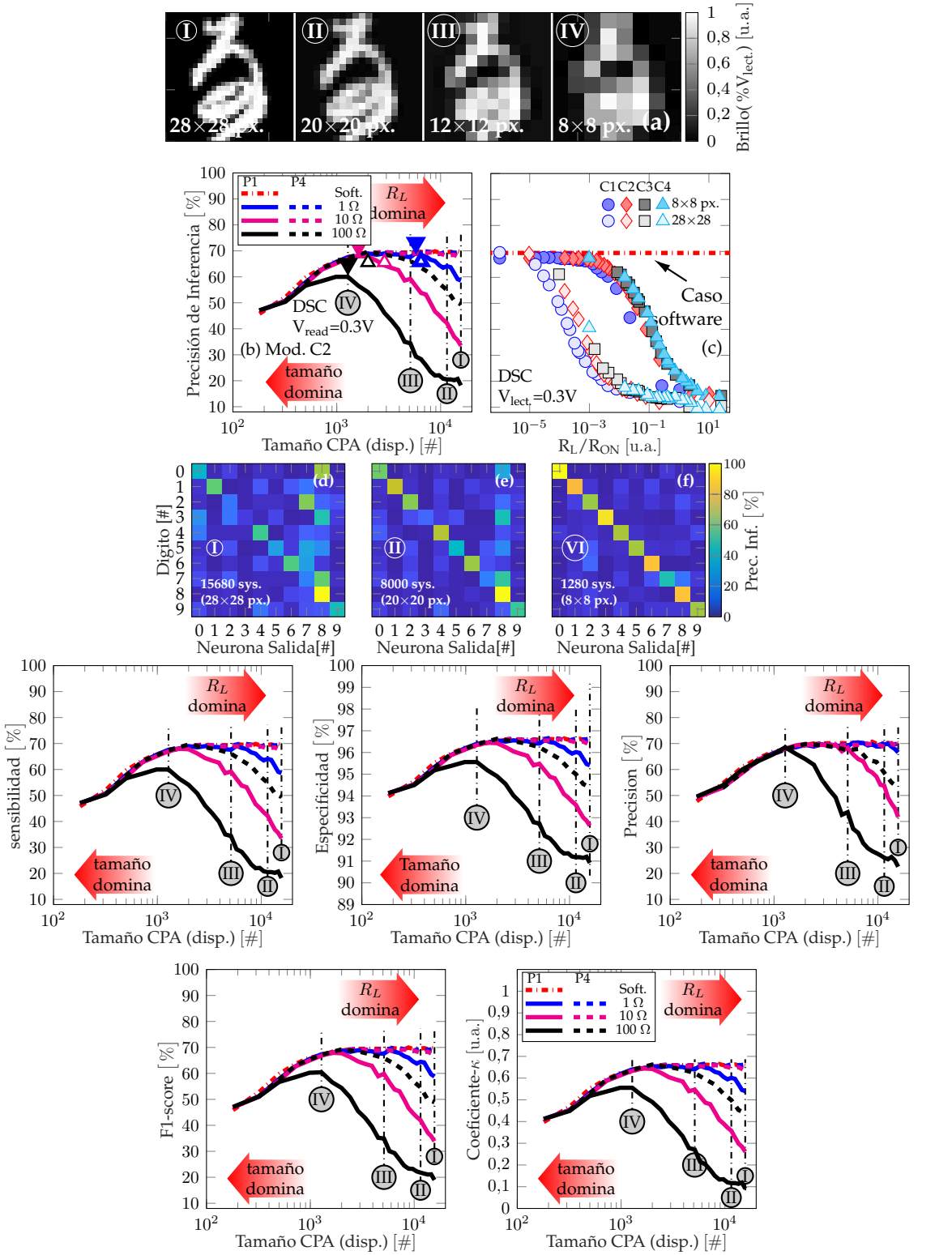


Figura B.7: (a) Pérdida de legibilidad de las imágenes de la base de datos MNIST-K para resolución decreciente de 28×28 px. (caso I) a 8×8 px. (caso IV). (b) Impacto del tamaño del CPA (cantidad de dispositivos) sobre la precisión de inferencia para el ajuste $C2$, y R_L barrido entre 1 y 100 Ω. Dos esquemas de particionado diferentes fueron utilizados: P1 indica *arrays* no particionados y P4 que cada CPA fue dividido en 4 sub-*arrays*. Los símbolos triangulares indican los puntos de tamaño máximo y precisión máxima, mostrando que los CPA particionados permiten mayor precisión en *arrays* más grandes. (c) Precisión de inferencia en función del cociente R_L/R_{ON} para imágenes de 28×28 (dos *arrays* de 784×10 arrays, ∼15.6k disp., símbolos vacíos) y 8×8 (dos 64×10 *arrays*, ∼1.2k disp., símbolos llenos). Matrices de confusión para el ajuste $C2$ y R_L=10 Ω para imágenes de (d) 28×28 px., (e) 20×20 px. y (f) 8×8 px. Otras métricas para el el reconocimiento de imágenes incluyen (g) sensibilidad, (h) especificidad, (i) precisión, (j) F1-Score y (k) coeficiente-κ.

B.6. Algoritmos de re-mapeo para la minimización de SAFs

Algoritmo 1: Mapeo adaptativo con tolerancia a pérdidas

Entrada: ${G_{M0}}^{+}(i,j)$, ${G_{M0}}^{-}(i,j)$ *(matrices de conductancia sin pérdidas),* ${G_M}^{+}(i,j)$ y ${G_M}^{-}(i,j)$ *(matrices de conductancia con pérdidas), con* $i = \{1, ..., n^2\}$ *y* $j = \{1, ..., m\}$;
Salida: $G_{M_{remap}}^{+}(i,j)$, $G_{M_{remap}}^{-}(i,j)$ *(matrices de conductancia sin pérdidas),* ${G_M}^{+}(i,j)$ y ${G_M}^{-}(i,j)$ *(matrices de conductancia con pérdidas), con* $i = \{1, ..., n^2\}$ *y* $j = \{1, ..., m\}$;
Asignar al número de filas con fallas irrecuperables a la variable unrec_faults;
while *iteration_i* $<$ *max_iterations* $\vee$ *unrec_faults* $>$ *0* **do**
    for $i = 1 : n^2$ **do**
        if *Row(i) tiene fallas irrecuperables* **then**
            for $j = 1 : n^2$ **do**
                Permutar pesos del CPA en Row(i) por Row(j);
            end
        end
    end
    Recalcular unrec_rows;
end
for $i = 1 : n^2$ **do**
    for $j = 1 : m$ **do**
        if ${G_M}^{+}(i,j) = SAI \wedge {G_M}^{-}(i,j) = OK \wedge W(i,j) > 0$ **then**
            $G_{M_{remap}}^{-}(i,j) = {G_{M0}}^{-}(i,j) + ({G_M}^{+}(i,j) - {G_{M0}}^{+}(i,j))$;
        end
        if ${G_M}^{+}(i,j) = OK \wedge {G_M}^{-}(i,j) = SAI \wedge W(i,j) < 0$ **then**
            $G_{M_{remap}}^{+}(i,j) = {G_{M0}}^{+}(i,j) + ({G_M}^{-}(i,j) - {G_{M0}}^{-}(i,j))$;
        end
    end
end

Algoritmo 2: Permutación de filas basada en la minimización del SWV

Entrada: $G_M{}^+(i,j)$, $G_M{}^-(i,j)$ *with* $i = \{1,...,n^2\}$ *and* $j = \{1,...,m\}$, *las imágenes están codificadas en nxn pixeles*;
Salida: $G_{M^+_{remap}}(i,j)$, $G_{M^-_{remap}}(i,j)$;
Calcular SWV según Ec. 6.10;
for $i = 1 : n^2$ **do**
  if *Row(i) contiene SAFs* **then**
    for $j = 1 : n^2$ **do**
      Permutar filas i y j: $G_{M^+_{remap}}(i,:) = G_M{}^+(j,:)$;
      Calcular SWV según Ec. 6.10;
      if *new_SWV<SWV* **then**
        SVN=new_SWV;
        break;
      else
        Deshacer permutación de filas;
      end
    end
  end
end

Algoritmo 3:

Entrada: $G_M{}^+(i,j)$, $G_M{}^-(i,j)$ *con* $i = \{1,...,n^2\}$ *y* $j = \{1,...,m\}$, *las imágenes están codificadas en n×n pixeles*;
Salida: $G_{M^+_{remap}}(i,j)$, $G_{M^-_{remap}}(i,j)$;
Clasificar los índices de los píxeles según su nivel de brillo medio en orden decreciente y almacenarlos en mean_br(k), con $k = \{1,...,n^2\}$;
Clasificar las filas de los CPAs según el número de celdas SAF en orden creciente y almacenarlas en Rows(k);
for $k = 1 : n^2$ **do**
  if *Row(i) contiene SAFs* **then**
    for $j = 1 : n^2$ **do**
      Asignar $G_M{}^+(Rows(k),j)$ *a* $G_{M^+_{remap}}(mean\_br(k),j)$;
      Asignar $G_M{}^-(Rows(k),j)$ *a* $G_{M^-_{remap}}(mean\_br(k),j)$;
      Permutar los datos de entrada entre las filas indicadas por Rows(k) y mean_br(k);
    end
  end
end

Bibliografía

[1] Ministerio de Ciencia Tecnología e Innovación Productiva, "Plan Argentina Innovadora 2020: Plan Nacional de Ciencia, Tecnología e Innovación: Lineamientos estratégicos 2012-2015," 2015.

[2] A. Ostendorf y K. König, "Tutorial laser in material nanoprocessing," en *Optically Induced Nanostructures: Biomedical and Technical Applications*, Walter de Gruyter GmbH, mayo de 2015, págs. xxiii-xl.

[3] R. Courtland, "Transistors could stop shrinking in 2021," *IEEE Spectrum*, vol. 53, n.º 9, págs. 9-11, 2016. DOI: 10.1109/MSPEC.2016.7551335.

[4] M. M. Shulaker, G. Hills, R. S. Park, R. T. Howe, K. Saraswat y col., "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, vol. 547, n.º 7661, págs. 74-78, 2017. DOI: 10.1038/nature22994.

[5] M. T. Bohr e I. A. Young, "CMOS Scaling Trends and Beyond," *IEEE Micro*, vol. 37, n.º 6, págs. 20-29, nov. de 2017. DOI: 10.1109/MM.2017.4241347.

[6] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous y A. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, n.º 5, págs. 256-268, oct. de 1974. DOI: 10.1109/JSSC.1974.1050511.

[7] J. Robertson y R. M. Wallace, "High-K materials and metal gates for CMOS applications," *Materials Science and Engineering: R: Reports*, vol. 88, págs. 1-41, feb. de 2015. DOI: 10.1016/j.mser.2014.11.001.

[8] M. Heyns, A. Alian, G. Brammertz, M. Caymax, G. Eneman y col., "Challenges for introducing Ge and III/V devices into CMOS technologies," en *2012 IEEE International Reliability Physics Symposium (IRPS)*, IEEE, abr. de 2012, págs. 5D.1.1-5D.1.10.

[9] F. Hui, C. Pan, Y. Shi, Y. Ji, E. Grustan-Gutierrez y M. Lanza, "On the use of two dimensional hexagonal boron nitride as dielectric," *Microelectronic Engineering*, vol. 163, págs. 119-133, sep. de 2016. DOI: 10.1016/J.MEE.2016.06.015.

[10] J. A. del Alamo, "Nanometre-scale electronics with III–V compound semiconductors," *Nature*, vol. 479, n.º 7373, págs. 317-323, nov. de 2011. DOI: 10.1038/nature10677.

[11] K. J. Kuhn, "Considerations for Ultimate CMOS Scaling," *IEEE Transactions on Electron Devices*, vol. 59, n.º 7, págs. 1813-1828, jul. de 2012. DOI: 10.1109/TED.2012.2193129.

[12] L. Czornomaz, V. Djara, V. Deshpande, E. O'Connor, M. Sousa y col., "First demonstration of InGaAs/SiGe CMOS inverters and dense SRAM arrays on Si using selective epitaxy and standard FEOL processes," en *2016 IEEE Symposium on VLSI Technology*, IEEE, jun. de 2016, págs. 1-2. DOI: 10.1109/VLSIT.2016.7573391.

[13] G. F. Jiao, W. Cao, Y. Xuan, D. M. Huang, P. D. Ye y M. F. Li, "Positive bias temperature instability degradation of InGaAs n-MOSFETs with Al2O3 gate dielectric," en *2011 International Electron Devices Meeting*, IEEE, dic. de 2011, págs. 27.1.1-27.1.4.

[14] B. Kaczer, T. Grasser, P. J. Roussel, J. Franco, R. Degraeve y col., "Origin of NBTI variability in deeply scaled pFETs," en *2010 IEEE International Reliability Physics Symposium*, IEEE, 2010, págs. 26-32. DOI: 10.1109/IRPS.2010.5488856.

[15] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger y col., "The Paradigm Shift in Understanding the Bias Temperature Instability: From Reaction–Diffusion to Switching Oxide Traps," English, *IEEE Transactions on Electron Devices*, vol. 58, n.º 11, págs. 3652-3666, nov. de 2011. DOI: 10.1109/TED.2011.2164543.

[16] G. Ribes, P. Mora, F. Monsieur, M. Rafik, F. Guarin y col., "High-K gate stack breakdown statistics modeled by correlated interfacial layer and high-k breakdown path," en *2010 IEEE International Reliability Physics Symposium*, IEEE, 2010, págs. 364-368.

[17] K. Okada, H. Ota, A. Hirano, A. Ogawa, T. Nabatame y A. Toriumi, "Roles of high-k and interfacial layers on TDDB reliability studied with HfAlOX/SiO2 stacked gate dielectrics," en *2008 IEEE International Reliability Physics Symposium*, IEEE, abr. de 2008, págs. 661-662.

[18] F. Palumbo, S. Lombardo y M. Eizenberg, "Physical mechanism of progressive breakdown in gate oxides," *Journal of Applied Physics*, vol. 115, n.º 22, 2014. DOI: 10.1063/1.4882116.

[19] F. Palumbo, M. Eizenberg y S. Lombardo, "General features of progressive breakdown in gate oxides: A compact model," en *IEEE International Reliability Physics Symposium Proceedings*, vol. 2015-May, 2015, 5A11-5A16. DOI: 10.1109/IRPS.2015.7112737.

[20] F. Schwierz, J. Pezoldt y R. Granzner, "Two-dimensional materials and their prospects in transistor electronics," *Nanoscale*, vol. 7, n.º 18, págs. 8261-8283, mayo de 2015. DOI: 10.1039/c5nr01052g.

[21] S. Takagi y M. Takenaka, "(Invited) III-V/Ge MOS Transistor Technologies for Future ULSI," *ECS Transactions*, vol. 54, n.º 1, págs. 39-54, jun. de 2013. DOI: 10.1149/05401.0039ecst.

[22] T. Hori, *Gate dielectrics and MOS ULSIs: principles, technologies and applications*. 2012.

[23] D. Dimaria, "Defect production, degradation, and breakdown of silicon dioxide films," *Solid-State Electronics*, vol. 41, n.º 7, págs. 957-965, jul. de 1997. DOI: 10.1016/S0038-1101(97)00006-3.

[24] J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology," *IBM Journal of Research and Development*, vol. 46, n.º 2.3, págs. 265-286, mar. de 2002. DOI: 10.1147/rd.462.0265.

[25] M. Alam, B. Weir, J. Bude, P. Silverman y A. Ghetti, "A computational model for oxide breakdown: theory and experiments," *Microelectronic Engineering*, vol. 59, n.º 1-4, págs. 137-147, nov. de 2001. DOI: 10.1016/S0167-9317(01)00657-8.

[26] D. J. DiMaria y J. W. Stasiak, "Trap creation in silicon dioxide produced by hot electrons," *Journal of Applied Physics*, vol. 65, n.º 6, págs. 2342-2356, mar. de 1989. DOI: 10.1063/1.342824.

[27] R. Degraeve, J. Ogier, R. Bellens, P. Roussel, G. Groeseneken y H. Maes, "A new model for the field dependence of intrinsic and extrinsic time-dependent dielectric breakdown," *IEEE Transactions on Electron Devices*, vol. 45, n.º 2, págs. 472-481, 1998. DOI: 10.1109/16.658683.

[28] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas y H. Maes, "A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides," en *Proceedings of International Electron Devices Meeting*, IEEE, 1995, págs. 863-866. DOI: 10.1109/IEDM.1995.499353.

[29] M. Depas, T. Nigam y M. Heyns, "Soft breakdown of ultra-thin gate oxide layers," *IEEE Transactions on Electron Devices*, vol. 43, n.º 9, págs. 1499-1504, 1996. DOI: 10.1109/16.535341.

[30] M. Nafría, J. Suñé y X. Aymerich, "Exploratory observations of post-breakdown conduction in polycrystalline-silicon and metal-gated thin-oxide metal-oxide-semiconductor capacitors," *Journal of Applied Physics*, vol. 73, n.º 1, págs. 205-215, ene. de 1993. DOI: 10.1063/1.353884.

[31] S. Lombardo, J. H. Stathis, B. P. Linder, K. L. Pey, F. Palumbo y C. H. Tung, "Dielectric breakdown mechanisms in gate oxides," *Journal of Applied Physics*, vol. 98, n.º 12, pág. 121 301, dic. de 2005. DOI: 10.1063/1.2147714.

[32] L. N. Liu, W. M. Tang y P. T. Lai, *Advances in la-based high-k dielectrics for MOS applications*, abr. de 2019. DOI: 10.3390/coatings9040217.

[33] Y. Li, M. Zhang, S. Long, J. Teng, Q. Liu y col., "Investigation on the Conductive Filament Growth Dynamics in Resistive Switching Memory via a Universal Monte Carlo Simulator," *Scientific Reports*, vol. 7, n.º 1, págs. 1-11, dic. de 2017. DOI: 10.1038/s41598-017-11165-5.

[34] Z. Zhang, Z. Wang, T. Shi, C. Bi, F. Rao y col., "Memory materials and devices: From concept to application," *InfoMat*, vol. 2, n.º 2, págs. 261-290, mar. de 2020. DOI: 10.1002/inf2.12077.

[35] D. Ielmini, R. Bruchhaus y R. Waser, "Thermochemical resistive switching: materials, mechanisms, and scaling projections," *Phase Transitions*, vol. 84, n.º 7, págs. 570-602, jul. de 2011. DOI: 10.1080/01411594.2011.561478.

[36] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," *Semiconductor Science and Technology*, vol. 31, n.º 6, pág. 063 002, jun. de 2016. DOI: 10.1088/0268-1242/31/6/063002.

[37] F. Pan, C. Chen, Z.-s. Wang, Y.-c. Yang, J. Yang y F. Zeng, "Nonvolatile resistive switching memories-characteristics, mechanisms and challenges," *Progress in Natural Science: Materials International*, vol. 20, págs. 1-15, nov. de 2010. DOI: 10.1016/S1002-0071(12)60001-X.

[38] Y. Lecun, Y. Bengio y G. Hinton, *Deep learning*, mayo de 2015. DOI: 10.1038/nature14539.

[39] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed y col., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, n.º 6, págs. 82-97, 2012. DOI: 10.1109/MSP.2012.2205597.

[40] A. Krizhevsky, I. Sutskever y G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, ene. de 2012. DOI: 10.1145/3065386.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed y col., "Going deeper with convolutions," en *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, IEEE Computer Society, oct. de 2015, págs. 1-9. DOI: 10.1109/CVPR.2015.7298594.

[42] K. Simonyan y A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. abs/1409.1556, 2015.

[43] A. Coates, B. Huval, T. Wang, D. J. Wu, A. Y. Ng y B. Catanzaro, "Deep Learning with COTS HPC Systems," en *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ép. ICML'13, Atlanta, GA, USA: JMLR.org, 2013, III–1337–III-1345.

[44] S. Gupta, A. Agrawal, K. Gopalakrishnan y P. Narayanan, "Deep Learning with Limited Numerical Precision," feb. de 2015.

[45] C. Lehmann, M. Viredaz y F. Blayo, "A Generic Systolic Array Building Block For Neural Networks with On-Chip Learning," *IEEE Transactions on Neural Networks*, vol. 4, n.º 3, págs. 400-407, 1993. DOI: 10.1109/72.217181.

[46] W. Sun, S. Choi, B. Kim y J. Park, "Three-Dimensional (3D) Vertical Resistive Random-Access Memory (VRRAM) Synapses for Neural Network Systems," *Materials*, vol. 12, n.º 20, pág. 3451, oct. de 2019. DOI: 10.3390/ma12203451.

[47] G. W. Burr, R. M. Shelby, C. Di Nolfo, J. W. Jang, R. S. Shenoy y col., "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2015-Febru, n.º February, págs. 29.5.1-29.5.4, 2015. DOI: 10.1109/IEDM.2014.7047135.

[48] I. M. Ross, "The invention of the transistor," *Proceedings of the IEEE*, vol. 86, n.º 1, págs. 7-28, 1998. DOI: 10.1109/5.658752.

[49] Y. Taur y T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998, pág. 469.

[50] E. H. Nicollian y J. R. Brews, *MOS (metal oxide semiconductor) physics and technology*. Wiley-Interscience, 2003, pág. 906.

[51] S. M. Sze y K. K. Ng, *Physics of Semiconductor Devices*. John Wiley & Sons, 2006, vol. 3, pág. 832.

[52] Sebastían Matías Pazos, "Desafíos de Confiabilidad en dispositivos Metal-Óxido-Semiconductor y circuitos integrados de radiofrecuencia.," Doctorado, Universidad Tecnológica Nacional, mar. de 2021, pág. 202.

[53] R. Winter, J. Ahn, P. C. McIntyre y M. Eizenberg, "New method for determining flat-band voltage in high mobility semiconductors," *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, vol. 31, n.º 3, pág. 030 604, mayo de 2013. DOI: 10.1116/1.4802478.

[54] F. Palumbo, F. L. Aguirre, S. M. Pazos, I. Krylov, R. Winter y M. Eizenberg, "Influence of the spatial distribution of border traps in the capacitance frequency dispersion of Al_2O_3/InGaAs," *Solid-State Electronics*, 2018. DOI: 10.1016/J.SSE.2018.07.006.

[55] R. Engel-Herbert, Y. Hwang y S. Stemmer, "Comparison of methods to quantify interface trap densities at dielectric/III-V semiconductor interfaces," *Journal of Applied Physics*, vol. 108, n.º 12, pág. 124 101, dic. de 2010. DOI: 10.1063/1.3520431.

[56] R. Castagné y A. Vapaille, "Description of the SiO_2/Si interface properties by means of very low frequency MOS capacitance measurements," *Surface Science*, vol. 28, n.º 1, págs. 157-193, nov. de 1971. DOI: 10.1016/0039-6028(71)90092-6.

[57] L. M. Terman, "An investigation of surface states at a silicon/silicon oxide interface employing metal-oxide-silicon diodes," *Solid State Electronics*, vol. 5, n.º 5, págs. 285-299, 1962. DOI: 10.1016/0038-1101(62)90111-9.

[58] C. N. Berglund, "Surface States at Steam-Grown Silicon-Silicon Dioxide Interfaces," *IEEE Transactions on Electron Devices*, vol. ED-13, n.º 10, págs. 701-705, 1966. DOI: 10.1109/T-ED.1966.15827.

[59] A. Vais, H.-C. Lin, C. Dou, K. Martens, T. Ivanov y col., "Temperature dependence of frequency dispersion in III–V metal-oxide-semiconductor C-V and the capture/emission process of border traps," *Applied Physics Letters*, vol. 107, n.º 5, pág. 053 504, ago. de 2015. DOI: 10.1063/1.4928332.

[60] C. Dou, D. Lin, A. Vais, T. Ivanov, H.-P. Chen y col., "Determination of energy and spatial distribution of oxide border traps in In0.53Ga0.47As MOS capacitors from capacitance–voltage characteristics measured at various temperatures," *Microelectronics Reliability*, vol. 54, n.º 4, págs. 746-754, abr. de 2014. DOI: 10.1016/j.microrel.2013.12.023.

[61] I. Krylov, D. Ritter y M. Eizenberg, "The physical origin of dispersion in accumulation in InGaAs based metal oxide semiconductor gate stacks," *Journal of Applied Physics*, vol. 117, n.º 17, pág. 174 501, mayo de 2015. DOI: 10.1063/1.4919600.

[62] F. C. Chiu, *A review on conduction mechanisms in dielectric films*, 2014. DOI: 10.1155/2014/578168.

[63] Z. A. Weinberg, W. C. Johnson y M. A. Lampert, "Determination of the sign of carrier transported across SiO2 films on Si," *Applied Physics Letters*, vol. 25, n.º 1, págs. 42-43, jul. de 1974. DOI: 10.1063/1.1655271.

[64] A. G. O'Neill, "An explanation of the asymmetry in electron and hole tunnel currents through ultra-thin SiO2 films," *Solid State Electronics*, vol. 29, n.º 3, págs. 305-310, 1986. DOI: 10.1016/0038-1101(86)90208-X.

[65] D. J. Dumin, J. R. Cooper, J. R. Maddux, R. S. Scott y D. P. Wong, "Low-level leakage currents in thin silicon oxide films," *Journal of Applied Physics*, vol. 76, n.º 1, págs. 319-327, jul. de 1994. DOI: 10.1063/1.357147.

[66] A. E. Islam, "Current Status of Reliability in Extended and Beyond CMOS Devices," *IEEE Transactions on Device and Materials Reliability*, vol. 16, n.º 4, págs. 647-666, 2016. DOI: 10.1109/TDMR.2014.2348940.

[67] E. Miranda, "Mecanismos de conducción a través del aislante de puerta en estructuras MOS (Metal-Oxido-Semiconductor)," Doctorado, Universidad de Buenos Aires, 2002.

[68] S. Lombardo, F. Crupi, A. La Magna, C. Spinella, A. Terrasi y col., "Electrical and thermal transient during dielectric breakdown of thin oxides in metal-SiO2-silicon capacitors," *Journal of Applied Physics*, vol. 84, n.º 1, págs. 472-479, jul. de 1998. DOI: 10.1063/1.368050.

[69] N. Raghavan, K. L. Pey y K. Shubhakar, "High-κ dielectric breakdown in nanoscale logic devices – Scientific insight and technology impact," *Microelectronics Reliability*, vol. 54, n.º 5, págs. 847-860, 2014. DOI: 10.1016/J.MICROREL.2014.02.013.

[70] F. Palumbo, C. Wen, S. Lombardo, S. Pazos, F. Aguirre y col., "A Review on Dielectric Breakdown in Thin Dielectrics: Silicon Dioxide, High- k , and Layered Dielectrics," *Advanced Functional Materials*, pág. 1 900 657, abr. de 2019. DOI: 10.1002/adfm.201900657.

[71] R. Degraeve, B. Kaczer y G. Groeseneken, "Degradation and breakdown in thin oxide layers: mechanisms, models and reliability prediction," *Microelectronics Reliability*, vol. 39, n.º 10, págs. 1445-1460, 1999. DOI: 10.1016/S0026-2714(99)00051-7.

[72] J. H. Stathis, "Percolation models for gate oxide breakdown," *Journal of Applied Physics*, vol. 86, n.º 10, págs. 5757-5766, nov. de 1999. DOI: 10.1063/1.371590.

[73] E. Y. Wu, "Facts and Myths of Dielectric Breakdown Processes-Part I: Statistics, Experimental, and Physical Acceleration Models," *IEEE Transactions on Electron Devices*, vol. 66, n.º 11, págs. 4523-4534, nov. de 2019. DOI: 10.1109/TED.2019.2933612.

[74] P. E. Nicollian, R. T. Cakici, A. T. Krishnan, V. K. Reddy y A. Seshadri, "Device characteristics and equivalent circuits for NMOS gate-to-drain soft and hard breakdown in polysilicon/SiON gate stacks," *IEEE Transactions on Electron Devices*, vol. 58, n.º 4, págs. 1170-1175, abr. de 2011. DOI: 10.1109/TED.2011.2105878.

[75] A. L. Hodgkin y A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, n.º 4, págs. 500-544, ago. de 1952. DOI: 10.1113/jphysiol.1952.sp004764.

[76] D. E. Rumelhart, G. E. Hinton y R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, n.º 6088, págs. 533-536, 1986. DOI: 10.1038/323533a0.

[77] S. Takagi y M. Takenaka, "Advanced non-Si channel CMOS technologies on Si platform," en *2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, IEEE, nov. de 2010, págs. 50-53. DOI: 10.1109/ICSICT.2010.5667857.

[78] M. Heyns, A. Alian, G. Brammertz, M. Caymax, Y. Chang y col., "Advancing CMOS beyond the Si roadmap with Ge and III/V devices," en *2011 International Electron Devices Meeting*, IEEE, dic. de 2011, págs. 13.1.1-13.1.4. DOI: 10.1109/IEDM.2011.6131543.

[79] A. W. Strong, E. Y. Wu, R.-P. Vollertsen, J. Sune, G. L. Rosa y col., *Reliability Wearout Mechanisms in Advanced CMOS Technologies*, 1.ª ed. Wiley-IEEE Press, 2009, pág. 624.

[80] S. Takagi, R. Zhang, T. Hoshii, N. Taoka y M. Takenaka, "MOS Interface Control Technologies for III-V/Ge Channel MOSFETs," en *ECS Transactions*, vol. 41, The Electrochemical Society, 2011, págs. 3-20. DOI: 10.1149/1.3633015.

[81] Chi On Chui, Hyoungsub Kim, P. McIntyre y K. Saraswat, "A germanium NMOSFET process integrating metal gate and improved high-K dielectrics," en *2003 IEEE IEDM*, IEEE, 2003, págs. 18.3.1-18.3.4. DOI: 10.1109/IEDM.2003.1269316.

[82] "ITRS 2.0 Executive Report," inf. téc., 2015.

[83] D. P. Brunco, B. De Jaeger, G. Eneman, J. Mitard, G. Hellings y col., "Germanium MOSFET Devices: Advances in Materials Understanding, Process Development, and Electrical Performance," *Journal of The Electrochemical Society*, vol. 155, n.º 7, H552, 2008. DOI: 10.1149/1.2919115.

[84] Y. Oshima, Y. Sun, D. Kuzum, T. Sugawara, K. C. Saraswat y col., "Chemical Bonding, Interfaces, and Defects in Hafnium Oxide - Germanium Oxynitride Gate Stacks on Ge(100)," *Journal of The Electrochemical Society*, vol. 155, n.º 12, G304, dic. de 2008. DOI: 10.1149/1.2995832.

[85] P. Batude, X. Garros, L. Clavelier, C. Le Royer, J. M. Hartmann y col., "Insights on fundamental mechanisms impacting Ge metal oxide semiconductor capacitors with high-k/metal gate stacks," *Journal of Applied Physics*, vol. 102, n.º 3, pág. 034 514, ago. de 2007. DOI: 10.1063/1.2767381.

[86] T. D. Lin, Y. H. Chang, C. A. Lin, M. L. Huang, W. C. Lee y col., "Realization of high-quality HfO 2 on In 0.53 Ga 0.47 As by in-situ atomic-layer-deposition," *Applied Physics Letters*, vol. 100, n.º 17, pág. 172 110, abr. de 2012. DOI: 10.1063/1.4706261.

[87] H. D. Trinh, E. Y. Chang, P. W. Wu, Y. Y. Wong, C. T. Chang y col., "The influences of surface treatment and gas annealing conditions on the inversion behaviors of the atomic-layer-deposition Al_2O_3/n-$In_{0,53}Ga_{0,47}$ metal-oxide-semiconductor capacitor," *Applied Physics Letters*, vol. 97, n.º 4, pág. 042 903, jul. de 2010. DOI: 10.1063/1.3467813.

[88] É. O'Connor, S. Monaghan, K. Cherkaoui, I. M. Povey y P. K. Hurley, "Analysis of the minority carrier response of n -type and p -type Au/Ni/Al_2O_3/$In_{0,53}Ga_{0,47}$/InP capacitors following an optimized (NH 4) 2 S treatment," *Applied Physics Letters*, vol. 99, n.º 21, pág. 212 901, nov. de 2011. DOI: 10.1063/1.3663535.

[89] I. Krylov, A. Gavrilov, M. Eizenberg y D. Ritter, "Correlation between Ga-O signature and midgap states at the $Al_2O_3In_{0,53}Ga_{0,47}$ interface," *Applied Physics Letters*, vol. 101, n.º 6, pág. 063 504, 2012. DOI: 10.1063/1.4745012.

[90] G. Brammertz, A. Alian, D. H.-C. Lin, M. Meuris, M. Caymax y W.-E. Wang, "A Combined Interface and Border Trap Model for High-Mobility Substrate Metal–Oxide–Semiconductor Devices Applied to In0.53Ga0.47As and InP Capacitors," *IEEE Transactions on Electron Devices*, vol. 58, n.º 11, págs. 3890-3897, nov. de 2011. DOI: 10.1109/TED.2011.2165725.

[91] I. Krylov, D. Ritter y M. Eizenberg, "The dispersion in accumulation at InGaAs-based metal/oxide/semiconductor gate stacks with a bi-layered dielectric structure," *Journal of Applied Physics*, vol. 118, n.º 8, pág. 084 502, ago. de 2015. DOI: 10.1063/1.4928960.

[92] C. Mahata, Y. An, S. Choi, Y.-C. Byun, D.-K. Kim y col., "Electrical properties of the HfO2–Al2O3 nanolaminates with homogeneous and graded compositions on InP," *Current Applied Physics*, vol. 16, n.º 3, págs. 294-299, 2016. DOI: 10.1016/j.cap.2015.11.022.

[93] A. Vais, J. Franco, H.-C. Lin, N. Collaert, A. Mocuta y col., "Impact of starting measurement voltage relative to flat-band voltage position on the capacitance-voltage hysteresis and on the defect characterization of InGaAs/high-k metal-oxide-semiconductor stacks," *Applied Physics Letters*, vol. 107, n.º 22, pág. 223 504, nov. de 2015. DOI: 10.1063/1.4936991.

[94] I. Krylov, D. Ritter y M. Eizenberg, "The role of the substrate on the dispersion in accumulation in III-V compound semiconductor based metal-oxide-semiconductor gate stacks," *Applied Physics Letters*, vol. 107, n.º 10, pág. 103 503, sep. de 2015. DOI: 10.1063/1.4930202.

[95] F. Palumbo y M. Eizenberg, "Degradation characteristics of metal/Al 2 O 3 /n-InGaAs capacitors," *Journal of Applied Physics*, vol. 115, n.º 1, pág. 014 106, ene. de 2014. DOI: 10.1063/1.4861033.

[96] F. Palumbo, I. Krylov y M. Eizenberg, "Comparison of the degradation characteristics of AlON/InGaAs and Al2O3/InGaAs stacks," *Journal of Applied Physics*, vol. 117, n.º 10, 2015. DOI: 10.1063/1.4914492.

[97] Chi On Chui, Hyoungsub Kim, D. Chi, B. Triplett, P. McIntyre y K. Saraswat, "A sub-400 C germanium MOSFET technology with high-κ dielectric and metal gate," en *Digest. International Electron Devices Meeting,*, IEEE, 2002, págs. 437-440. DOI: 10.1109/IEDM.2002.1175872.

[98] E. Golias, L. Tsetseris, A. Chroneos y A. Dimoulas, "Interaction of metal impurities with native oxygen defects in GeO2," *Microelectronic Engineering*, vol. 104, págs. 37-41, abr. de 2013. DOI: 10.1016/J.MEE.2012.11.012.

[99] S. Fadida, F. Palumbo, L. Nyns, D. Lin, S. Van Elshocht y col., "Hf-based high-k dielectrics for p-Ge MOS gate stacks," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 32, n.º 3, pág. 03D105, mayo de 2014. DOI: 10.1116/1.4837295.

[100] D. Bodlaki, H. Yamamoto, D. H. Waldeck y E. Borguet, "Ambient stability of chemically passivated germanium interfaces," *Surface Science*, vol. 543, n.º 1-3, págs. 63-74, oct. de 2003. DOI: 10.1016/S0039-6028(03)00958-0.

[101] L. Zhang, H. Li, Y. Guo, K. Tang, J. Woicik y col., "Selective Passivation of GeO 2 /Ge Interface Defects in Atomic Layer Deposited High- k MOS Structures," *ACS Applied Materials & Interfaces*, vol. 7, n.º 37, págs. 20 499-20 506, sep. de 2015. DOI: 10.1021/acsami.5b06087.

[102] Q. Xie, S. Deng, M. Schaekers, D. Lin, M. Caymax y col., "Germanium surface passivation and atomic layer deposition of high- k dielectrics—a tutorial review on Ge-based MOS capacitors," *Semiconductor Science and Technology*, vol. 27, n.º 7, pág. 074 012, jul. de 2012. DOI: 10.1088/0268-1242/27/7/074012.

[103] S. Swaminathan, Y. Sun, P. Pianetta y P. C. McIntyre, "Ultrathin ALD-Al_2O_3 layers for Ge(001) gate stacks: Local composition evolution and dielectric properties," *Journal of Applied Physics*, vol. 110, n.º 9, pág. 094 105, nov. de 2011. DOI: 10.1063/1.3647761.

[104] J. Franco, B. Kaczer, P. J. Roussel, J. Mitard, M. Cho y col., "SiGe Channel Technology: Superior Reliability Toward Ultrathin EOT Devices—Part I: NBTI," *IEEE Transactions on Electron Devices*, vol. 60, n.º 1, págs. 396-404, ene. de 2013. DOI: 10.1109/TED.2012.2225625.

[105] J. Franco, B. Kaczer, J. Mitard, M. Toledano-Luque, P. J. Roussel y col., "NBTI Reliability of SiGe and Ge Channel pMOSFETs With SiO2 / HfO2 Dielectric Stack," *IEEE Transactions on Device and Materials Reliability*, vol. 13, n.º 4, págs. 497-506, dic. de 2013. DOI: 10.1109/TDMR.2013.2281731.

[106] J. Franco, B. Kaczer, P. J. Roussel, J. Mitard, S. Sioncke y col., "Understanding the suppressed charge trapping in relaxed- and strained-Ge/SiO_2/HfO_2 pMOSFETs and implications for the screening of alternative high-mobility substrate/dielectric CMOS gate stacks," en *2013 IEEE International Electron Devices Meeting*, IEEE, dic. de 2013, págs. 15.2.1-15.2.4. DOI: 10.1109/IEDM.2013.6724634.

[107] F. Palumbo, S. M. Pazos, F. L. Aguirre, R. Winter, I. Krylov y M. Eizenberg, "Temperature dependence of trapping effects in metal gates/Al_2O_3/InGaAs stacks," *Solid-State Electronics*, vol. 132, págs. 12-18, 2017. DOI: http://dx.doi.org/10.1016/j.sse.2017.03.009.

[108] J. Franco, A. Alian, B. Kaczer, D. Lin, T. Ivanov y col., "Suitability of high-k gate oxides for III-V devices: A PBTI study in $In_{0,53}Ga_{0,47}$ devices with Al_2O_3," en *2014 IEEE International Reliability Physics Symposium*, IEEE, jun. de 2014, 6A.2.1-6A.2.6. DOI: 10.1109/IRPS.2014.6861098.

[109] J. Lin, Y. Y. Gomeniuk, S. Monaghan, I. M. Povey, K. Cherkaoui y col., "An investigation of capacitance-voltage hysteresis in metal/high- k/$In_{0,53}Ga_{0,47}$ metal-oxide-semiconductor capacitors," *Journal of Applied Physics*, vol. 114, n.º 14, pág. 144 105, oct. de 2013. DOI: 10.1063/1.4824066.

[110] S. Fadida, L. Nyns, S. Van Elshocht y M. Eizenberg, "Effect of Remote Oxygen Scavenging on Electrical Properties of Ge-Based Metal–Oxide–Semiconductor Capacitors," *Journal of Electronic Materials*, págs. 1-7, ago. de 2016. DOI: 10.1007/s11664-016-4841-6.

[111] A. Delabie, A. Alian, F. Bellenger, G. Brammertz, D. P. Brunco y col., "Atomic Layer Deposition of High-k Dielectric Layers on Ge and III-V MOS Channels," *ECS Transactions*, vol. 16, n.º 10, págs. 671-685, oct. de 2008. DOI: 10.1149/1.2986824.

[112] L. Nyns, D. Lin, G. Brammertz, F. Bellenger, X. Shi y col., "Interface and Border Traps in Ge-Based Gate Stacks," en *ECS Transactions*, vol. 35, The Electrochemical Society, 2011, págs. 465-480. DOI: 10.1149/1.3569938.

[113] S. Fadida, M. Eizenberg, L. Nyns, S. Van Elshocht y M. Caymax, "Band alignment of Hf–Zr oxides on Al2O3/GeO2/Ge stacks," *Microelectronic Engineering*, vol. 88, n.º 7, págs. 1557-1559, 2011. DOI: 10.1016/j.mee.2011.03.075.

[114] H. Matsubara, T. Sasada, M. Takenaka y S. Takagi, "Evidence of low interface trap density in GeO_2-Ge metal-oxide-semiconductor structures fabricated by thermal oxidation," *Applied Physics Letters*, vol. 93, n.º 3, pág. 032 104, jul. de 2008. DOI: 10.1063/1.2959731.

[115] D. Kuzum, T. Krishnamohan, A. J. Pethe, A. K. Okyay, Y. Oshima y col., "Ge-Interface Engineering With Ozone Oxidation for Low Interface-State Density," *IEEE Electron Device Letters*, vol. 29, n.º 4, págs. 328-330, abr. de 2008. DOI: 10.1109/LED.2008.918272.

[116] J. Franco, B. Kaczer, J. Roussel, M. Cho, T. Grasser y col., "BTI reliability of high-mobility channel devices: SiGe, Ge and InGaAs," en *2014 IEEE International Integrated Reliability Workshop Final Report (IIRW)*, IEEE, oct. de 2014, págs. 53-57. DOI: 10.1109/IIRW.2014.7049510.

[117] R. Zhang, N. Taoka, Po-Chin Huang, M. Takenaka y S. Takagi, "1-nm-thick EOT high mobility Ge n- and p-MOSFETs with ultrathin GeOx/Ge MOS interfaces fabricated by plasma post oxidation," en *2011 International Electron Devices Meeting*, IEEE, dic. de 2011, págs. 28.3.1-28.3.4. DOI: 10.1109/IEDM.2011.6131630.

[118] Chi On Chui, F. Ito y K. Saraswat, "Nanoscale germanium MOS Dielectrics-part I: germanium oxynitrides," *IEEE Transactions on Electron Devices*, vol. 53, n.º 7, págs. 1501-1508, jul. de 2006. DOI: 10.1109/TED.2006.875808.

[119] G. Dushaq, A. Nayfeh y M. Rasras, "Passivation of Ge/high- ¡i¿κ¡/i¿interface using RF Plasma nitridation," *Semiconductor Science and Technology*, vol. 33, n.º 1, pág. 015 003, ene. de 2018. DOI: 10.1088/1361-6641/aa98cd.

[120] A. Ghosh, M. B. Clavel, P. D. Nguyen, M. A. Meeker, G. A. Khodaparast y col., "Growth, structural, and electrical properties of germanium- on -silicon heterostructure by molecular beam epitaxy," *AIP Advances*, vol. 7, n.º 9, pág. 095 214, sep. de 2017. DOI: 10.1063/1.4993446.

[121] A. Delabie, F. Bellenger, M. Houssa, T. Conard, S. Van Elshocht y col., "Effective electrical passivation of Ge(100) for high-k gate dielectric layers using germanium oxide," *Applied Physics Letters*, vol. 91, n.º 8, pág. 082 904, ago. de 2007. DOI: 10.1063/1.2773759.

[122] S. Kar, *High Permittivity Gate Dielectric Materials*. 2013, vol. 43, págs. 425-457. DOI: 10.1007/978-3-642-36535-5.

[123] I. Krylov, D. Ritter y M. Eizenberg, "Hf x Al y O ternary dielectrics for InGaAs based metal-oxide-semiconductor capacitors," *Journal of Applied Physics*, vol. 122, n.º 3, pág. 034 505, jul. de 2017. DOI: 10.1063/1.4993905.

[124] P. Ponath, A. B. Posadas y A. A. Demkov, "Ge(001) surface cleaning methods for device integration," *Applied Physics Reviews*, vol. 4, n.º 2, pág. 021 308, jun. de 2017. DOI: 10.1063/1.4984975.

[125] B. Onsia, M. Caymax, T. Conard, S. De Gendt, F. De Smedt y col., "On the Application of a Thin Ozone Based Wet Chemical Oxide as an Interface for ALD High-k Deposition," *Solid State Phenomena*, vol. 103-104, págs. 19-22, 2005. DOI: 10.4028/www.scientific.net/SSP.103-104.19.

[126] K. M. Wong, W. K. Chim, J. Q. Huang y L. Zhu, "Scanning capacitance microscopy detection of charge trapping in free-standing germanium nanodots and the passivation of hole trap sites," *Journal of Applied Physics*, vol. 103, n.º 5, pág. 054 505, mar. de 2008. DOI: 10.1063/1.2875776.

[127] T. Kauerauf, R. Degraeve, E. Cartier, C. Soens y G. Groeseneken, "Low Weibull slope of breakdown distributions in high-κ layers," *IEEE Electron Device Letters*, vol. 23, n.º 4, págs. 215-217, abr. de 2002. DOI: 10.1109/55.992843.

[128] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas y col., "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Transactions on Electron Devices*, vol. 45, n.º 4, págs. 904-911, abr. de 1998. DOI: 10.1109/16.662800.

[129] J. Suñé, I. Placencia, N. Barniol, E. Farrés, F. Martín y X. Aymerich, "On the breakdown statistics of very thin SiO_2 films," *Thin Solid Films*, vol. 185, n.º 2, págs. 347-362, mar. de 1990. DOI: 10.1016/0040-6090(90)90098-X.

[130] E. Y. Wu, J. Suñé y W. Lai, "On the Weibull shape factor of intrinsic breakdown of dielectric films and its accurate experimental determination-Part II: Experimental results and the effects of stress conditions," *IEEE Transactions on Electron Devices*, vol. 49, n.º 12, págs. 2141-2150, dic. de 2002. DOI: 10.1109/TED.2002.805603.

[131] N. Raghavan, K. L. Pey, K. Shubhakar y M. Bosman, "Modified percolation model for polycrystalline high-κ gate stack with grain boundary defects," en *IEEE Electron Device Letters*, vol. 32, ene. de 2011, págs. 78-80. DOI: 10.1109/LED.2010.2085074.

[132] G. Bersuker, J. Yum, L. Vandelli, A. Padovani, L. Larcher y col., "Grain boundary-driven leakage path formation in HfO_2 dielectrics," *Solid-State Electronics*, vol. 65-66, n.º 1, págs. 146-150, 2011. DOI: 10.1016/j.sse.2011.06.031.

[133] K. Shubhakar, K. L. Pey, N. Raghavan, S. S. Kushvaha, M. Bosman y col., "Study of preferential localized degradation and breakdown of HfO_2/SiO_X dielectric stacks at grain boundary sites of polycrystalline HfO_2 dielectrics," *Microelectronic Engineering*, vol. 109, págs. 364-369, sep. de 2013. DOI: 10.1016/j.mee.2013.03.021.

[134] A. M. El-Sayed, M. B. Watkins, A. L. Shluger y V. V. Afanas'Ev, "Identification of intrinsic electron trapping sites in bulk amorphous silica from ab initio calculations," *Microelectronic Engineering*, vol. 109, págs. 68-71, sep. de 2013. DOI: 10.1016/j.mee.2013.03.027.

[135] S. R. Bradley, A. L. Shluger y G. Bersuker, "Electron-Injection-Assisted generation of oxygen vacancies in monoclinic HfO_2," *Physical Review Applied*, vol. 4, n.º 6, pág. 064 008, dic. de 2015. DOI: 10.1103/PhysRevApplied.4.064008.

[136] E. Y. Wu, B. Li y J. H. Stathis, "Modeling of time-dependent non-uniform dielectric breakdown using a clustering statistical approach," *Applied Physics Letters*, vol. 103, n.º 15, pág. 152 907, oct. de 2013. DOI: 10.1063/1.4824035.

[137] A. Padovani y L. Larcher, "Time-dependent dielectric breakdown statistics in SiO_2 and HfO_2 dielectrics: Insights from a multi-scale modeling approach," en *IEEE International Reliability Physics Symposium Proceedings*, vol. 1, IEEE, mar. de 2018, págs. 86-93. DOI: 10.1109/IRPS.2018.8353552.

[138] L. Vandelli, A. Padovani, L. Larcher y G. Bersuker, "Microscopic Modeling of Electrical Stress-Induced Breakdown in Poly-Crystalline Hafnium Oxide Dielectrics," *IEEE Transactions on Electron Devices*, vol. 60, n.º 5, págs. 1754-1762, mayo de 2013. DOI: 10.1109/TED.2013.2255104.

[139] Glenn Knoll, *Radiation Detection and Measurement*. John Wiley, 2013, págs. 489-522.

[140] J. S. Suehle, E. M. Vogel, P. Roitman, J. F. Conley, A. H. Johnston y col., "Observation of latent reliability degradation in ultrathin oxides after heavy-ion irradiation," *Applied Physics Letters*, vol. 80, n.º 7, págs. 1282-1284, feb. de 2002. DOI: 10.1063/1.1448859.

[141] W. Wesch, A. Kamarou y E. Wendler, "Effect of high electronic energy deposition in semiconductors," en *Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms*, vol. 225, North-Holland, ago. de 2004, págs. 111-128. DOI: 10.1016/j.nimb.2004.04.188.

[142] K. Awazu, S. Ishii y K. Shima, "Structure of latent tracks created by swift heavy-ion bombardment of amorphous," *Physical Review B - Condensed Matter and Materials Physics*, vol. 62, n.º 6, págs. 3689-3698, ago. de 2000. DOI: 10.1103/PhysRevB.62.3689.

[143] A. Meftah, F. Brisard, J. M. Costantini, E. Dooryhee, M. Hage-Ali y col., "Track formation in SiO_2 quartz and the thermal-spike mechanism," *Physical Review B*, vol. 49, n.º 18, págs. 12 457-12 463, mayo de 1994. DOI: 10.1103/PhysRevB.49.12457.

[144] J. F. Carlotti, A. D. Touboul, M. Ramonda, M. Caussanel, C. Guasch y col., "Growth of silicon bump induced by swift heavy ion at the silicon oxide-silicon interface," *Applied Physics Letters*, vol. 88, n.º 4, págs. 1-3, ene. de 2006. DOI: 10.1063/1.2166476.

[145] M. Toulemonde, C. Trautmann, E. Balanzat, K. Hjort y A. Weidinger, "Track formation and fabrication of nanostructures with MeV-ion beams," en *Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms*, vol. 216, North-Holland, feb. de 2004, págs. 1-8. DOI: 10.1016/j.nimb.2003.11.013.

[146] P. S. Chaudhari, T. M. Bhave, D. Kanjilal y S. V. Bhoraskar, "Swift heavy ion induced growth of nanocrystalline silicon in silicon oxide," *Journal of Applied Physics*, vol. 93, n.º 6, págs. 3486-3489, mar. de 2003. DOI: 10.1063/1.1542913.

[147] C. J. Dale, P. W. Marshall, G. P. Summers, E. A. Wolicki y E. A. Burke, "Displacement damage equivalent to dose in silicon devices," *Applied Physics Letters*, vol. 54, n.º 5, págs. 451-453, ene. de 1989. DOI: 10.1063/1.100949.

[148] A. M. Carvalho, M. Marinoni, A. D. Touboul, C. Guasch, H. Lebius y col., "Discontinuous ion tracks on silicon oxide on silicon surfaces after grazing-angle heavy ion irradiation," *Applied Physics Letters*, vol. 90, n.º 7, pág. 073 116, feb. de 2007. DOI: 10.1063/1.2591255.

[149] A. Fontana, S. Pazos, F. Aguirre, N. Vega, N. Muller y col., "Pulse quenching and charge sharing effects on heavy-ion microbeam induced ASET in a full-custom CMOS OpAmp," *IEEE Transactions on Nuclear Science*, págs. 1-1, 2019. DOI: 10.1109/TNS.2019.2908174.

[150] F. Palumbo, M. Debray, N. Vega, C. Quinteros, A. Kalstein y F. Guarin, "Evolution of the gate current in 32nm MOSFETs under irradiation," *Solid-State Electronics*, vol. 119, págs. 19-24, 2016. DOI: 10.1016/j.sse.2016.02.004.

[151] S. Sondon, A. Falcon, P. Mandolesi, P. Julian, N. Vega y col., "Diagnose of radiation induced single event effects in a PLL using a heavy ion microbeam," en *2013 14th Latin American Test Workshop - LATW*, IEEE, abr. de 2013, págs. 1-5. DOI: 10.1109/LATW.2013.6562682.

[152] A. Fontana, S. M. Pazos, F. L. Aguirre, F. Palumbo, N. Vega y col., "Heavy Ion Microbeam Experimental Study of ASET on a Full-Custom CMOS OpAmp," en *2018 31st Symposium on Integrated Circuits and Systems Design (SBCCI)*, IEEE, ago. de 2018, págs. 1-5. DOI: 10.1109/SBCCI.2018.8533232.

[153] J. F. Ziegler, M. D. Ziegler y J. P. Biersack, "SRIM - The stopping and range of ions in matter (2010)," *Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms*, vol. 268, n.º 11-12, págs. 1818-1823, jun. de 2010. DOI: 10.1016/j.nimb.2010.02.091.

[154] R. Stoller, M. Toloczko, G. Was, A. Certain, S. Dwaraknath y F. Garner, "On the use of SRIM for computing radiation damage exposure," *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 310, págs. 75-80, jul. de 2013. DOI: 10.1016/J.NIMB.2013.05.008.

[155] J. R. Srour, C. J. Marshall y P. W. Marshall, "Review of displacement damage effects in silicon devices," *IEEE Transactions on Nuclear Science*, vol. 50 III, n.º 3, págs. 653-670, jun. de 2003. DOI: 10.1109/TNS.2003.813197.

[156] F. W. Sexton, D. M. Fleetwood, M. R. Shaneyfelt, P. E. Dodd, G. L. Hash y col., "Precursor ion damage and angular dependence of single event gate rupture in thin oxides," *IEEE Transactions on Nuclear Science*, vol. 45, n.º 6 PART 1, págs. 2509-2518, 1998. DOI: 10.1109/23.736492.

[157] M. Ceschia, A. Paccagnella, M. Turrini, A. Candelori, G. Ghidini y J. Wyss, "Heavy ion irradiation of thin gate oxides," en *IEEE Transactions on Nuclear Science*, vol. 47, 2000, págs. 2648-2655. DOI: 10.1109/23.903821.

[158] Applied Materials, *Ginestra*^{TM}.

[159] E. Miranda, C. Mahata, T. Das y C. K. Maiti, "An extension of the Curie-von Schweidler law for the leakage current decay in MIS structures including progressive breakdown," en *Microelectronics Reliability*, vol. 51, Pergamon, sep. de 2011, págs. 1535-1539. DOI: 10.1016/j.microrel.2011.06.035.

[160] S. Sahhaf, R. Degraeve, P. J. Roussel, B. Kaczer, T. Kauerauf y G. Groeseneken, "A new TDDB reliability prediction methodology accounting for multiple SBD and wear out," *IEEE Transactions on Electron Devices*, vol. 56, n.º 7, págs. 1424-1432, jul. de 2009. DOI: 10.1109/TED.2009.2021810.

[161] N. Raghavan, "Failure of Weibull distribution to represent switching statistics in OxRAM," *Microelectronic Engineering*, vol. 178, págs. 230-234, jun. de 2017. DOI: 10.1016/J.MEE.2017.05.007.

[162] M. Stucchi, P. J. Roussel, Z. Tokei, S. Demuynck y G. Groeseneken, "A Comprehensive LER-Aware TDDB Lifetime Model for Advanced Cu Interconnects," *IEEE Transactions on Device and Materials Reliability*, vol. 11, n.º 2, págs. 278-289, jun. de 2011. DOI: 10.1109/TDMR.2011.2121909.

[163] S. R. Bradley, G. Bersuker y A. L. Shluger, "Modelling of oxygen vacancy aggregates in monoclinic HfO_2: can they contribute to conductive filament formation?" *Journal of Physics Condensed Matter*, vol. 27, n.º 41, pág. 415 401, sep. de 2015. DOI: 10.1088/0953-8984/27/41/415401.

[164] N. Raghavan, K. L. Pey, W. H. Liu y X. Li, "New statistical model to decode the reliability and Weibull slope of high-κ and interfacial layer in a dual layer dielectric stack," en *IEEE International Reliability Physics Symposium Proceedings*, IEEE, 2010, págs. 778-786. DOI: 10.1109/IRPS.2010.5488735.

[165] D. Fink, A. V. Petrov, K. Hoppe, W. R. Fahrner, R. M. Papaleo y col., "Etched ion tracks in silicon oxide and silicon oxynitride as charge injection or extraction channels for novel electronic structures," en *Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms*, vol. 218, North-Holland, jun. de 2004, págs. 355-361. DOI: 10.1016/j.nimb.2003.12.083.

[166] L. Vandelli, A. Padovani, L. Larcher, R. G. Southwick, W. B. Knowlton y G. Bersuker, "A physical model of the temperature dependence of the current through SiO¡sub¿2¡/sub¿/HfO¡sub¿2¡/sub¿stacks," *IEEE Transactions on Electron Devices*, vol. 58, n.º 9, págs. 2878-2887, sep. de 2011. DOI: 10.1109/TED.2011.2158825.

[167] A. Padovani, D. Z. Gao, A. L. Shluger y L. Larcher, "A microscopic mechanism of dielectric breakdown in SiO¡sub¿2¡/sub¿films: An insight from multi-scale modeling," *Journal of Applied Physics*, vol. 121, n.º 15, pág. 155 101, abr. de 2017. DOI: 10.1063/1.4979915.

[168] R. Waser, R. Dittmann, G. Staikov y K. Szot, "Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges," *Advanced Materials*, vol. 21, n.º 25-26, págs. 2632-2663, jul. de 2009. DOI: 10.1002/adma.200900375.

[169] Y. Shi, X. Liang, B. Yuan, V. Chen, H. Li y col., "Electronic synapses made of layered two-dimensional materials," *Nature Electronics*, vol. 1, n.º 8, págs. 458-465, ago. de 2018. DOI: 10.1038/s41928-018-0118-9.

[170] A. Rodriguez-Fernandez, C. Cagli, J. Sune y E. Miranda, "Switching Voltage and Time Statistics of Filamentary Conductive Paths in HfO_2-based ReRAM Devices," *IEEE Electron Device Letters*, págs. 1-1, 2018. DOI: 10.1109/LED.2018.2822047.

[171] A. Rodriguez-Fernandez, C. Cagli, L. Perniola, J. Suñé y E. Miranda, "Identification of the generation/rupture mechanism of filamentary conductive paths in ReRAM devices using oxide failure analysis," *Microelectronics Reliability*, vol. 76-77, págs. 178-183, sep. de 2017. DOI: 10.1016/J.MICROREL.2017.06.088.

[172] S. S. Sheu, K. H. Cheng, M. F. Chang, P. C. Chiang, W. P. Lin y col., "Fast-write resistive RAM (RRAM) for embedded applications," *IEEE Design and Test of Computers*, vol. 28, n.º 1, págs. 64-71, ene. de 2011. DOI: 10.1109/MDT.2010.96.

[173] D. Ielmini, R. Bruchhaus y R. Waser, "Thermochemical resistive switching: materials, mechanisms, and scaling projections," *Phase Transitions*, vol. 84, n.º 7, págs. 570-602, jul. de 2011. DOI: 10.1080/01411594.2011.561478.

[174] J. Park, S. Jung, J. Lee, W. Lee, S. Kim y col., "Resistive switching characteristics of ultra-thin TiO_x," *Microelectronic Engineering*, vol. 88, n.º 7, págs. 1136-1139, jul. de 2011. DOI: 10.1016/J.MEE.2011.03.050.

[175] G. Bersuker, D. C. Gilmer, D. Veksler, P. Kirsch, L. Vandelli y col., "Metal oxide resistive memory switching mechanism based on conductive filament properties," *Journal of Applied Physics*, vol. 110, n.º 12, pág. 124 518, dic. de 2011. DOI: 10.1063/1.3671565.

[176] L. Zhang, R. Huang, M. Zhu, S. Qin, Y. Kuang y col., "Unipolar TaO-Based Resistive Change Memory Realized With Electrode Engineering," *IEEE Electron Device Letters*, vol. 31, n.º 9, págs. 966-968, sep. de 2010. DOI: 10.1109/LED.2010.2052091.

[177] F. Palumbo, E. Miranda, G. Ghibaudo y V. Jousseaume, "Formation and Characterization of Filamentary Current Paths in HfO_2-Based Resistive Switching Structures," *Ieee Electron Device Letters*, vol. 33, n.º 7, págs. 1057-1059, 2012. DOI: 10.1109/led.2012.2194689.

[178] E. Miranda y J. Suñé, "Electron transport through broken down ultra-thin SiO_2 layers in MOS devices," *Microelectronics Reliability*, vol. 44, n.º 1, págs. 1-23, 2004. DOI: 10.1016/j.microrel.2003.08.005.

[179] D. Ielmini, "Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature-Driven Filament Growth," *IEEE Transactions on Electron Devices*, vol. 58, n.º 12, págs. 4309-4317, dic. de 2011. DOI: 10.1109/TED.2011.2167513.

[180] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer y D. Ielmini, "Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling," *IEEE Transactions on Electron Devices*, vol. 59, n.º 9, págs. 2468-2475, sep. de 2012. DOI: 10.1109/TED.2012.2202320.

[181] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer y D. Ielmini, "Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part I: Experimental Study," *IEEE Transactions on Electron Devices*, vol. 59, n.º 9, págs. 2461-2467, sep. de 2012. DOI: 10.1109/TED.2012.2202319.

[182] F. Palumbo, P. Shekhter, K. Cohen Weinfeld y M. Eizenberg, "Characteristics of the dynamics of breakdown filaments in Al_2O_3/InGaAs stacks," *Applied Physics Letters*, vol. 107, n.º 12, pág. 122 901, sep. de 2015. DOI: 10.1063/1.4931496.

[183] C. H. Tung, K. L. Pey, L. J. Tang, M. K. Radhakrishnan, W. H. Lin y col., "Percolation path and dielectric-breakdown-induced-epitaxy evolution during ultrathin gate dielectric breakdown transient," *Applied Physics Letters*, vol. 83, n.º 11, págs. 2223-2225, sep. de 2003. DOI: 10.1063/1.1611649.

[184] F. Palumbo, G. Condorelli, S. Lombardo, K. Pey, C. Tung y L. Tang, "Structure of the oxide damage under progressive breakdown," *Microelectronics Reliability*, vol. 45, n.º 5-6, págs. 845-848, mayo de 2005. DOI: 10.1016/J.MICROREL.2004.11.034.

[185] S. Privitera, G. Bersuker, B. Butcher, A. Kalantarian, S. Lombardo y col., "Microscopy study of the conductive filament in HfO_2 resistive switching memory devices," *Microelectronic Engineering*, vol. 109, págs. 75-78, sep. de 2013. DOI: 10.1016/J.MEE.2013.03.145.

[186] H. Du, C. L. Jia, A. Koehl, J. Barthel, R. Dittmann y col., "Nanosized Conducting Filaments Formed by Atomic-Scale Defects in Redox-Based Resistive Switching Memories," *Chemistry of Materials*, vol. 29, n.º 7, págs. 3164-3173, abr. de 2017. DOI: 10.1021/acs.chemmater.7b00220.

[187] Y. Nishi, K. Fleck, U. Bottger, R. Waser y S. Menzel, "Effect of RESET Voltage on Distribution of SET Switching Time of Bipolar Resistive Switching in a Tantalum Oxide Thin Film," *IEEE Transactions on Electron Devices*, vol. 62, n.º 5, págs. 1561-1567, 2015. DOI: 10.1109/TED.2015.2411748.

[188] K. Tang, A. C. Meng, F. Hui, Y. Shi, T. Petach y col., "Distinguishing Oxygen Vacancy Electromigration and Conductive Filament Formation in TiO_2 Resistance Switching Using Liquid Electrolyte Contacts," *Nano Letters*, vol. 17, n.º 7, págs. 4390-4399, jul. de 2017. DOI: 10.1021/acs.nanolett.7b01460.

[189] *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications* D. Ielmini y R. Waser. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, jun. de 2016. DOI: 10.1002/9783527680870.

[190] H. Wang, Y. Du, Y. Li, B. Zhu, W. R. Leow y col., "Configurable Resistive Switching between Memory and Threshold Characteristics for Protein-Based Devices," *Advanced Functional Materials*, vol. 25, n.º 25, págs. 3825-3831, 2015. DOI: 10.1002/adfm.201501389.

[191] Y. J. Huang, S. C. Chao, D. H. Lien, C. Y. Wen, J. H. He y S. C. Lee, "Dual-functional memory and threshold resistive switching based on the push-pull mechanism of oxygen ions," *Scientific Reports*, vol. 6, n.º March, págs. 1-10, 2016. DOI: 10.1038/srep23945.

[192] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya y col., "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," *Nature Materials*, vol. 16, n.º 1, págs. 101-108, ene. de 2017. DOI: 10.1038/nmat4756.

[193] S. Chen, M. R. Mahmoodi, Y. Shi, C. Mahata, B. Yuan y col., "Wafer-scale integration of two-dimensional materials in high-density memristive crossbar arrays for artificial neural networks," *Nature Electronics*, vol. 3, n.º 10, págs. 638-645, oct. de 2020. DOI: 10.1038/s41928-020-00473-w.

[194] B. E. A. Saleh y M. C. Teich, *Fundamentals of photonics*. Wiley-Interscience, 2007, pág. 1177.

[195] C. Pan, Y. Ji, N. Xiao, F. Hui, K. Tang y col., "Coexistence of Grain-Boundaries-Assisted Bipolar and Threshold Resistive Switching in Multilayer Hexagonal Boron Nitride," *Advanced Functional Materials*, vol. 27, n.º 10, pág. 1 604 811, mar. de 2017. DOI: 10.1002/adfm.201604811.

[196] A. Fantini, D. J. Wouters, R. Degraeve, L. Goux, L. Pantisano y col., "Intrinsic Switching Behavior in HfO_2 RRAM by Fast Electrical Measurements on Novel 2R Test Structures," en *2012 4th IEEE International Memory Workshop*, IEEE, mayo de 2012, págs. 1-4. DOI: 0.1109/IMW.2012.6213646.

[197] B. P. Linder, S. Lombardo, J. H. Stathis, A. Vayshenker y D. Frank, "Voltage dependence of hard breakdown growth and the reliability implication in thin dielectrics," *IEEE Electron Device Letters*, vol. 23, n.º 11, págs. 661-663, nov. de 2002. DOI: 10.1109/LED.2002.805010.

[198] S. Pazos, F. L. Aguirre, E. Miranda, S. Lombardo y F. Palumbo, "Comparative study of the breakdown transients of thin Al_2O_3 and HfO_2 films in MIM structures and their connection with the thermal properties of materials," *Journal of Applied Physics*, vol. 121, n.º 9, pág. 094 102, mar. de 2017. DOI: 10.1063/1.4977851.

[199] R. Pagano, S. Lombardo, F. Palumbo, P. Kirsch, S. Krishnan y col., "A novel approach to characterization of progressive breakdown in high-k/metal gate stacks," *Microelectronics Reliability*, vol. 48, n.º 11-12, págs. 1759-1764, nov. de 2008. DOI: 10.1016/J.MICROREL.2008.07.071.

[200] S. Lombardo, E. Y. Wu y J. H. Stathis, "Electron energy dissipation model of gate dielectric progressive breakdown in n- and p-channel field effect transistors," *Journal of Applied Physics*, vol. 122, n.º 8, pág. 085 701, ago. de 2017. DOI: 10.1063/1.4985794.

[201] T. Gokmen e Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," *Frontiers in Neuroscience*, vol. 10, n.º JUL, jul. de 2016. DOI: 10.3389/fnins.2016.00333.

[202] R. Midya, Z. Wang, J. Zhang, S. E. Savel'ev, C. Li y col., "Anatomy of Ag/Hafnia-Based Selectors with 10^{10} Nonlinearity," *Advanced Materials*, vol. 29, n.º 12, pág. 1 604 457, mar. de 2017. DOI: 10.1002/adma.201604457.

[203] S. Takagi, N. Yasuda y A. Toriumi, "Experimental evidence of inelastic tunneling in stress-induced leakage current," *IEEE Transactions on Electron Devices*, vol. 46, n.º 2, págs. 335-341, 1999. DOI: 10.1109/16.740899.

[204] P. E. Blöchl y J. H. Stathis, "Hydrogen Electrochemistry and Stress-Induced Leakage Current in Silica," *Physical Review Letters*, vol. 83, n.º 2, págs. 372-375, jul. de 1999. DOI: 10.1103/PhysRevLett.83.372.

[205] E. Y. Wu, J. H. Stathis y L.-K. Han, "Ultra-thin oxide reliability for ULSI applications," *Semiconductor Science and Technology*, vol. 15, n.º 5, págs. 425-435, mayo de 2000. DOI: 10.1088/0268-1242/15/5/301.

[206] Young Hee Kim, K. Onishi, Chang Seok Kang, Hag-Ju Cho, Rino Choi y col., "Thickness dependence of Weibull slopes of HfO_2 gate dielectrics," *IEEE Electron Device Letters*, vol. 24, n.º 1, págs. 40-42, ene. de 2003. DOI: 10.1109/LED.2002.807314.

[207] S. Slesazeck y T. Mikolajick, "Nanoscale resistive switching memory devices: a review," *Nanotechnology*, mayo de 2019. DOI: 10.1088/1361-6528/ab2084.

[208] Y. Shacham-Diamand, A. Dedhia, D. Hoffstetter y W. G. Oldham, "Copper Transport in Thermal SiO_2," *Journal of The Electrochemical Society*, vol. 140, n.º 8, pág. 2427, ago. de 1993. DOI: 10.1149/1.2220837.

[209] F. Palumbo, X. Liang, B. Yuan, Y. Shi, F. Hui y col., "Bimodal Dielectric Breakdown in Electronic Devices Using Chemical Vapor Deposited Hexagonal Boron Nitride as Dielectric," *Advanced Electronic Materials*, vol. 4, n.º 3, pág. 1 700 506, mar. de 2018. DOI: 10.1002/aelm.201700506.

[210] S. Zafar, H. Jagannathan, L. F. Edge y D. Gupta, "Measurement of oxygen diffusion in nanometer scale HfO_2 gate dielectric films," *Applied Physics Letters*, vol. 98, n.º 15, pág. 152 903, abr. de 2011. DOI: 10.1063/1.3579256.

[211] P. Huang, X. Y. Liu, B. Chen, H. T. Li, Y. J. Wang y col., "A Physics-Based Compact Model of Metal-Oxide-Based RRAM DC and AC Operations," *IEEE Transactions on Electron Devices*, vol. 60, n.º 12, págs. 4090-4097, dic. de 2013. DOI: 10.1109/TED.2013.2287755.

[212] M. Bocquet, D. Deleruyelle, H. Aziza, C. Muller, J.-M. Portal y col., "Robust Compact Model for Bipolar Oxide-Based Resistive Switching Memories," *IEEE Transactions on Electron Devices*, vol. 61, n.º 3, págs. 674-681, mar. de 2014. DOI: 10.1109/TED.2013.2296793.

[213] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski y M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," *Nature Materials*, vol. 10, n.º 8, págs. 591-595, jun. de 2011. DOI: 10.1038/nmat3054.

[214] G. Milano, M. Luebben, Z. Ma, R. Dunin-Borkowski, L. Boarino y col., "Self-limited single nanowire systems combining all-in-one memristive and neuromorphic functionalities," *Nature Communications*, vol. 9, n.º 1, págs. 1-10, dic. de 2018. DOI: 10.1038/s41467-018-07330-7.

[215] A. Falin, Q. Cai, E. J. Santos, D. Scullion, D. Qian y col., "Mechanical properties of atomically thin boron nitride and the role of interlayer interactions," *Nature Communications*, vol. 8, n.º 1, págs. 1-9, jun. de 2017. DOI: 10.1038/ncomms15815.

[216] L. H. Li, J. Cervenka, K. Watanabe, T. Taniguchi e Y. Chen, "Strong oxidation resistance of atomically thin boron nitride nanosheets," *ACS Nano*, vol. 8, n.º 2, págs. 1457-1462, feb. de 2014. DOI: 10.1021/nn500059s.

[217] Z. Liu, Y. Gong, W. Zhou, L. Ma, J. Yu y col., "Ultrathin higherature oxidation-resistant coatings of hexagonal boron nitride," *Nature Communications*, vol. 4, n.º 1, págs. 1-8, oct. de 2013. DOI: 10.1038/ncomms3541.

[218] Q. Cai, D. Scullion, W. Gan, A. Falin, S. Zhang y col., "High thermal conductivity of high-quality monolayer boron nitride and its thermal expansion," *Science Advances*, vol. 5, n.º 6, eaav0129, jun. de 2019. DOI: 10.1126/sciadv.aav0129.

[219] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose y R. W. Linderman, "Memristor crossbar-based neuromorphic computing system: A case study," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, n.º 10, págs. 1864-1878, 2014. DOI: 10.1109/TNNLS.2013.2296777.

[220] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang y H.-S. P. Wong, "A Low Energy Oxide-Based Electronic Synaptic Device for Neuromorphic Visual Systems with Tolerance to Device Variation," *Advanced Materials*, vol. 25, n.º 12, págs. 1774-1779, mar. de 2013. DOI: 10.1002/adma.201203680.

[221] R. F. Freitas y W. W. Wilcke, "Storage-class memory: The next storage system technology," *IBM Journal of Research and Development*, vol. 52, n.º 4.5, págs. 439-447, jul. de 2008. DOI: 10.1147/rd.524.0439.

[222] N. K. Upadhyay, S. Joshi y J. J. Yang, "Synaptic electronics and neuromorphic computing," *Science China Information Sciences*, vol. 59, n.º 6, pág. 061 404, jun. de 2016. DOI: 10.1007/s11432-016-5565-1.

[223] Y. Sasago, M. Kinoshita, T. Morikawa, K. Kurotsuchi, S. Hanzawa y col., "Cross-point phase change memory with $4F^2$ cell size driven by low-contact-resistivity poly-Si diode," en *Digest of Technical Papers - Symposium on VLSI Technology*, jul. de 2009, págs. 24-25.

[224] S. N. Truong y K. S. Min, "New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing," *Journal of Semiconductor Technology and Science*, vol. 14, n.º 3, págs. 356-363, jun. de 2014. DOI: 10.5573/JSTS.2014.14.3.356.

[225] S. Truong, S.-J. Ham y K.-S. Min, "Neuromorphic crossbar circuit with nanoscale filamentary-switching binary memristors for speech recognition," *Nanoscale Research Letters*, vol. 9, n.º 1, pág. 629, 2014. DOI: 10.1186/1556-276X-9-629.

[226] S. N. Truong, S. H. Shin, S. D. Byeon, J. S. Song y K. S. Min, "New Twin Crossbar Architecture of Binary Memristors for Low-Power Image Recognition With Discrete Cosine Transform," *IEEE Transactions on Nanotechnology*, vol. 14, n.º 6, págs. 1104-1111, nov. de 2015. DOI: 10.1109/TNANO.2015.2473666.

[227] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri y col., "RRAM-based synapse for neuromorphic system with pattern recognition function," en *Technical Digest - International Electron Devices Meeting, IEDM*, 2012. DOI: 10.1109/IEDM.2012.6479016.

[228] B. Liu, H. Li, Y. Chen, X. Li, T. Huang y col., "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD*, vol. 2015-Janua, n.º January, págs. 63-70, 2015. DOI: 10.1109/ICCAD.2014.7001330.

[229] S. J. Ham, H. S. Mo y K. S. Min, "Low-Power VDD/3 write scheme with inversion coding circuit for complementary memristor array," *IEEE Transactions on Nanotechnology*, vol. 12, n.º 5, págs. 851-857, 2013. DOI: 10.1109/TNANO.2013.2274529.

[230] D. B. Strukov, G. S. Snider, D. R. Stewart y R. S. Williams, "The missing memristor found," *Nature*, vol. 453, n.º 7191, págs. 80-83, mayo de 2008. DOI: 10.1038/nature06932.

[231] C. Chen, S. Gao, G. Tang, H. Fu, G. Wang y col., "Effect of Electrode Materials on AlN-Based Bipolar and Complementary Resistive Switching," *ACS Applied Materials & Interfaces*, vol. 5, n.º 5, págs. 1793-1799, mar. de 2013. DOI: 10.1021/am303128h.

[232] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge y col., "Analogue signal and image processing with large memristor crossbars," *Nature Electronics*, vol. 1, n.º 1, págs. 52-59, ene. de 2018. DOI: 10.1038/s41928-017-0002-z.

[233] Y. K. Lee, J. W. Jeon, E.-S. Park, C. Yoo, W. Kim y col., "Matrix Mapping on Crossbar Memory Arrays with Resistive Interconnects and Its Use in In-Memory Compression of Biosignals," *Micromachines*, vol. 10, n.º 5, pág. 306, 2019. DOI: 10.3390/mi10050306.

[234] R. Han, P. Huang, Y. Zhao, X. Cui, X. Liu y J. Kang, "Efficient evaluation model including interconnect resistance effect for large scale RRAM crossbar array matrix computing," *Science China Information Sciences*, vol. 62, n.º 2, págs. 1-11, 2019. DOI: 10.1007/s11432-018-9555-8.

[235] B. Zhang, N. Uysal, D. Fan y R. Ewetz, "Handling Stuck-at-faults in Memristor Crossbar Arrays using Matrix Transformations," en *Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC*, New York, NY, USA: Institute of Electrical y Electronics Engineers Inc., ene. de 2019, págs. 474-479. DOI: 10.1145/3287624.3287707.

[236] ——, "Handling Stuck-at-fault Defects using Matrix Transformation for Robust Inference of DNNs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, págs. 1-1, oct. de 2019. DOI: 10.1109/tcad.2019.2944582.

[237] L. Xia, M. Liu, X. Ning, K. Chakrabarty e Y. Wang, "Fault-Tolerant Training with On-Line Fault Detection for RRAM-Based Neural Computing Systems," en *Proceedings - Design Automation Conference*, vol. Part 12828, Institute of Electrical y Electronics Engineers Inc., jun. de 2017. DOI: 10.1145/3061639.3062248.

[238] C. Liu, M. Hu, J. P. Strachan y H. H. Li, "Rescuing Memristor-based Neuromorphic Design with High Defects," en *Proceedings - Design Automation Conference*, vol. Part 12828, Institute of Electrical y Electronics Engineers Inc., jun. de 2017. DOI: 10.1145/3061639.3062310.

[239] A. Mehonic, D. Joksas, W. H. Ng, M. Buckwell y A. J. Kenyon, "Simulation of inference accuracy using realistic RRAM devices," *Frontiers in Neuroscience*, vol. 13, n.º JUN, págs. 1-15, 2019. DOI: 10.3389/fnins.2019.00593.

[240] C. Lammie, S. Member, W. Xiang y S. Member, "MemTorch : An Open-source Simulation Framework for Memristive Deep Learning Systems," págs. 1-14,

[241] C. Yakopcic, R. Hasan, T. M. Taha, M. R. McLean y D. Palmer, "Efficacy of memristive crossbars for neuromorphic processors," *Proceedings of the International Joint Conference on Neural Networks*, págs. 15-20, 2014. DOI: 10.1109/IJCNN.2014.6889807.

[242] H. Yu, Y. Wang, W. Fei e Y. Shang. (). "http://www.nvmspice.org/." available online.

[243] C. Yakopcic, T. M. Taha, G. Subramanyam y R. E. Pino, "Memristor SPICE modeling," en *Advances in Neuromorphic Memristor Science and Applications*, Springer Netherlands, ene. de 2012, págs. 211-244. DOI: 10.1007/978-94-007-4491-2-12.

[244] D. Panda, P. P. Sahu y T. Y. Tseng, "A Collective Study on Modeling and Simulation of Resistive Random Access Memory," *Nanoscale Research Letters*, vol. 13, 2018. DOI: 10.1186/s11671-017-2419-8.

[245] J. Zha, H. Huang, T. Huang, J. Cao, A. Alsaedi y F. E. Alsaadi, "A general memristor model and its applications in programmable analog circuits," *Neurocomputing*, vol. 267, págs. 134-140, 2017. DOI: 10.1016/j.neucom.2017.04.057.

[246] P. Sheridan, K. H. Kim, S. Gaba, T. Chang, L. Chen y W. Lu, "Device and SPICE modeling of RRAM devices," *Nanoscale*, vol. 3, n.º 9, págs. 3833-3840, 2011. DOI: 10.1039/c1nr10557d.

[247] Y. N. Joglekar y S. J. Wolf, "The elusive memristor: Properties of basic electrical circuits," *European Journal of Physics*, vol. 30, n.º 4, págs. 661-675, 2009. DOI: 10.1088/0143-0807/30/4/001.

[248] T. Prodromakis, B. P. Peh, C. Papavassiliou y C. Toumazou, "A versatile memristor model with nonlinear dopant kinetics," *IEEE Transactions on Electron Devices*, vol. 58, n.º 9, págs. 3099-3105, sep. de 2011. DOI: 10.1109/TED.2011.2158004.

[249] M. D. Pickett, D. B. Strukov, J. L. Borghetti, J. J. Yang, G. S. Snider y col., "Switching dynamics in titanium dioxide memristive devices," *Journal of Applied Physics*, vol. 106, n.º 7, 2009. DOI: 10.1063/1.3236506.

[250] F. Merrikh Bayat, B. Hoskins y D. B. Strukov, "Phenomenological modeling of memristive devices," *Applied Physics A: Materials Science and Processing*, vol. 118, n.º 3, págs. 779-786, 2015. DOI: 10.1007/s00339-015-8993-7.

[251] D. Biolek, Z. Biolek, V. Biolkova y Z. Kolka, "Modeling of TiO_2 memristor: From analytic to numerical analyses," *Semiconductor Science and Technology*, vol. 29, n.º 12, págs. 2-7, 2014. DOI: 10.1088/0268-1242/29/12/125008.

[252] C. Yakopcic, T. M. Taha, G. Subramanyam y R. E. Pino, "Generalized memristive device SPICE model and its application in circuit design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, n.º 8, págs. 1201-1214, 2013. DOI: 10.1109/TCAD.2013.2252057.

[253] S. Kvatinsky, E. G. Friedman, A. Kolodny y U. C. Weiser, "TEAM: Threshold adaptive memristor model," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, n.º 1, págs. 211-221, 2013. DOI: 10.1109/TCSI.2012.2215714.

[254] S. Kvatinsky, M. Ramadan, E. G. Friedman y A. Kolodny, "VTEAM: A General Model for Voltage-Controlled Memristors," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, n.º 8, págs. 786-790, ago. de 2015. DOI: 10.1109/TCSII.2015.2433536.

[255] K. Eshraghian, O. Kavehei, K. R. Cho, J. M. Chappell, A. Iqbal y col., "Memristive device fundamentals and modeling: Applications to circuits and systems simulation," en *Proceedings of the IEEE*, vol. 100, jun. de 2012, págs. 1991-2007. DOI: 10.1109/JPROC.2012.2188770.

[256] D. Biolek, Z. Biolek, V. Biolkova y Z. Kolka, "Reliable modeling of ideal generic memristors via state-space transformation," *Radioengineering*, vol. 24, n.º 2, págs. 393-407, 2015. DOI: 10.13164/re.2015.0393.

[257] E. Miranda, "Compact Model for the Major and Minor Hysteretic I-V Loops in Nonlinear Memristive Devices," *IEEE Transactions on Nanotechnology*, vol. 14, n.º 5, págs. 787-789, sep. de 2015. DOI: 10.1109/TNANO.2015.2455235.

[258] G. A. Patterson, J. Sune y E. Miranda, "Voltage-Driven Hysteresis Model for Resistive Switching: SPICE Modeling and Circuit Applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, n.º 12, págs. 2044-2051, dic. de 2017. DOI: 10.1109/TCAD.2017.2756561.

[259] S. Petzold, E. Miranda, S. U. Sharath, J. Muñoz-Gorriz, T. Vogel y col., "Analysis and simulation of the multiple resistive switching modes occurring in HfO_X -based resistive random access memories using memdiodes," *Journal of Applied Physics*, vol. 125, n.º 23, 2019. DOI: 10.1063/1.5094864.

[260] L. Xia, P. Gu, B. Li, T. Tang, X. Yin y col., "Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication," *Journal of Computer Science and Technology*, vol. 31, n.º 1, págs. 3-19, 2016. DOI: 10.1007/s11390-016-1608-8.

[261] B. Li, Y. Wang, Y. Chen, H. H. Li y H. Yang, "ICE: Inline calibration for memristor crossbar-based computing engine," EDAA, abr. de 2014, págs. 1-4. DOI: 10.7873/date.2014.197.

[262] R. Degraeve, A. Fantini, N. Raghavan, L. Goux, S. Clima y col., "Causes and consequences of the stochastic aspect of filamentary RRAM," *Microelectronic Engineering*, vol. 147, págs. 171-175, nov. de 2015. DOI: 10.1016/j.mee.2015.04.025.

[263] Y. Y. Chen, R. Degraeve, S. Clima, B. Govoreanu, L. Goux y col., "Understanding of the endurance failure in scaled HfO2-based 1T1R RRAM through vacancy mobility degradation," en *Technical Digest - International Electron Devices Meeting, IEDM*, 2012. DOI: 10.1109/IEDM.2012.6479079.

[264] C. Y. Chen, H. C. Shih, C. W. Wu, C. H. Lin, P. F. Chiu y col., "RRAM defect modeling and failure analysis based on march test and a novel squeeze-search scheme," *IEEE Transactions on Computers*, vol. 64, n.º 1, págs. 180-190, ene. de 2015. DOI: 10.1109/TC.2014.12.

[265] L. Xia, W. Huangfu, T. Tang, X. Yin, K. Chakrabarty y col., "Stuck-at Fault Tolerance in RRAM Computing Systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, n.º 1, págs. 102-115, mar. de 2018. DOI: 10.1109/JETCAS.2017.2776980.

[266] Y. LeCun, C. Cortes y C. J. Burges. (1998). "MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges," dirección: http://yann.lecun.com/exdb/mnist/ (visitado 21-11-2019).

[267] F. L. Aguirre, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "Application of the Quasi-Static Memdiode Model in Cross-Point Arrays for Large Dataset Pattern Recognition," *IEEE Access*, vol. 8, págs. 1-1, 2020. DOI: 10.1109/ACCESS.2020.3035638.

[268] F. L. Aguirre, A. Rodriguez-Fernandez, S. M. Pazos, J. Sune, E. Miranda y F. Palumbo, "Study on the Connection Between the Set Transient in RRAMs and the Progressive Breakdown of Thin Oxides," *IEEE Transactions on Electron Devices*, vol. 66, n.º 8, págs. 1-7, 2019. DOI: 10.1109/ted.2019.2922555.

[269] K. Fröhlich, I. Kundrata, M. Blaho, M. Precner, M. Tapajna y col., "Hafnium oxide and tantalum oxide based resistive switching structures for realization of minimum and maximum functions," *Journal of Applied Physics*, vol. 124, n.º 15, oct. de 2018. DOI: 10.1063/1.5025802.

[270] W. Choi, K. Moon, M. Kwak, C. Sung, J. Lee y col., "Hardware implementation of neural network using pre-programmed resistive device for pattern recognition," *Solid-State Electronics*, vol. 153, págs. 79-83, mar. de 2019. DOI: 10.1016/j.sse.2018.12.018.

[271] J. Blasco, P. Jančovič, K. Fröhlich, J. Suñé y E. Miranda, "Modeling of the switching I-V characteristics in ultrathin (5 nm) atomic layer deposited HfO_2 films using the logistic hysteron," *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, vol. 33, n.º 1, 01A102, ene. de 2015. DOI: 10.1116/1.4900599.

[272] C.-Y. Lin, C.-Y. Wu, C.-Y. Wu, C. Hu y T.-Y. Tseng, "Bistable Resistive Switching in Al_2O_3 Memory Thin Films," *Journal of The Electrochemical Society*, vol. 154, n.º 9, G189, 2007. DOI: 10.1149/1.2750450.

[273] M. K. Yang, J. W. Park, T. K. Ko y J. kook Lee, "Resistive switching characteristics of TiN/MnO_2/Pt memory devices," *Physica Status Solidi - Rapid Research Letters*, vol. 4, n.º 8-9, págs. 233-235, sep. de 2010. DOI: 10.1002/pssr.201004213.

[274] *Resistance Random Access Memory (RRAM)*, http://archive.today/c6PS, Accessed: 2010-09-30.

[275] E. Miranda, W. Román Acevedo, D. Rubi, U. Lüders, P. Granell y col., "Modeling of the multilevel conduction characteristics and fatigue profile of Ag/$La_{1/3}Ca_{2/3}MnO_3$/Pt structures using a compact memristive approach," *Journal of Applied Physics*, vol. 121, n.º 20, pág. 205 302, mayo de 2017. DOI: 10.1063/1.4984051.

[276] A. Mehonic, A. L. Shluger, D. Gao, I. Valov, E. Miranda y col., "Silicon Oxide (SiO_X): A Promising Material for Resistance Switching?" *Advanced Materials*, vol. 30, n.º 43, págs. 1-21, 2018. DOI: 10.1002/adma.201801187.

[277] V. Milo, C. Zambelli, P. Olivo, E. Pérez, M. K. Mahadevaiah y col., "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *APL Materials*, vol. 7, n.º 8, 2019. DOI: 10.1063/1.5108650.

[278] H. Xiao, K. Rasul y R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," ago. de 2017.

[279] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto y D. Ha, "Deep Learning for Classical Japanese Literature," *ArXiv*, vol. abs/1812.01718, 2018.

[280] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," inf. téc., 2009.

[281] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu y A. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning," *NIPS*, 2011.

[282] A. S. Georghiades, P. N. Belhumeur y D. J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, n.º 6, págs. 643-660, 2001.

[283] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu y col., "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature Communications*, vol. 9, n.º 1, págs. 1-8, dic. de 2018. DOI: 10.1038/s41467-018-04484-2.

[284] Z. Dong, Z. Zhou, Z. Li, C. Liu, P. Huang y col., "Convolutional Neural Networks Based on RRAM Devices for Image Recognition and Online Learning Tasks," *IEEE Transactions on Electron Devices*, vol. 66, n.º 1, págs. 793-801, 2019. DOI: 10.1109/TED.2018.2882779.

[285] D. Querlioz, O. Bichler, P. Dollfus y C. Gamrat, "Immunity to device variations in a spiking neural network with memristive nanodevices," *IEEE Transactions on Nanotechnology*, vol. 12, n.º 3, págs. 288-295, 2013. DOI: 10.1109/TNANO.2013.2250995.

[286] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev y D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, n.º 7550, págs. 61-64, mayo de 2015. DOI: 10.1038/nature14441.

[287] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang y col., "Face classification using electronic synapses," *Nature Communications*, vol. 8, n.º May, págs. 1-8, 2017. DOI: 10.1038/ncomms15199.

[288] Q. Zhang, H. Wu, P. Yao, W. Zhang, B. Gao y col., "Sign backpropagation: An on-chip learning algorithm for analog RRAM neuromorphic computing systems," *Neural Networks*, vol. 108, págs. 217-223, 2018. DOI: 10.1016/j.neunet.2018.08.012.

[289] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, n.º 4, págs. 525-533, ene. de 1993. DOI: 10.1016/S0893-6080(05)80056-5.

[290] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, n.º 2, págs. 431-441, jun. de 1963. DOI: 10.1137/0111030.

[291] R. Battiti, "First- and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method," *Neural Computation*, vol. 4, n.º 2, págs. 141-166, mar. de 1992. DOI: 10.1162/neco.1992.4.2.141.

[292] M. J. Powell, "Restart procedures for the conjugate gradient method," *Mathematical Programming*, vol. 12, n.º 1, págs. 241-254, dic. de 1977. DOI: 10.1007/BF01593790.

[293] J. E. Dennis y R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial y Applied Mathematics, ene. de 1996. DOI: 10.1137/1.9781611971200.

[294] R. Fletcher, "Function minimization by conjugate gradients," *The Computer Journal*, vol. 7, n.º 2, págs. 149-154, feb. de 1964. DOI: 10.1093/comjnl/7.2.149.

[295] M. Riedmiller y H. Braun, "Direct adaptive method for faster backpropagation learning: The RPROP algorithm," en *1993 IEEE International Conference on Neural Networks*, Publ by IEEE, 1993, págs. 586-591. DOI: 10.1109/icnn.1993.298623.

[296] M. Hagan y H. Demuth, "Neural Network Design," *Neural Networks in a Softcomputing Framework*, págs. 1-1012, 2014.

[297] M. Hu, H. Li, Q. Wu, G. S. Rose e Y. Chen, "Memristor crossbar based hardware realization of BSB recall function," en *The 2012 International Joint Conference on Neural Networks (IJCNN)*, IEEE, jun. de 2012, págs. 1-7. DOI: 10.1109/IJCNN.2012.6252563.

[298] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila y col., "Dot-product engine for neuromorphic computing," en *DAC '16: Proceedings of the 53rd Annual Design Automation Conference*, New York, NY, USA: Association for Computing Machinery, 2016, págs. 1-6. DOI: 10.1145/2897937.2898010.

[299] M. E. Fouda, S. Lee, J. Lee, A. Eltawil y F. Kurdahi, "Mask Technique for Fast and Efficient Training of Binary Resistive Crossbar Arrays," *IEEE Transactions on Nanotechnology*, vol. 18, págs. 704-716, 2019. DOI: 10.1109/tnano.2019.2927493.

[300] J. Liang y H. S. Wong, "Cross-point memory array without cell selectors-device characteristics and data storage pattern dependencies," *IEEE Transactions on Electron Devices*, vol. 57, n.º 10, págs. 2531-2538, 2010. DOI: 10.1109/TED.2010.2062187.

[301] A. Chen, "A comprehensive crossbar array model with solutions for line resistance and nonlinear device characteristics," *IEEE Transactions on Electron Devices*, vol. 60, n.º 4, págs. 1318-1326, 2013. DOI: 10.1109/TED.2013.2246791.

[302] M. Laiho, E. Lehtonen, A. Russell y P. Dudek, "Memristive synapses are becoming reality," *The Neuromorphic Engineer*, págs. 10-12, 2010. DOI: 10.2417/1201011.003396.

[303] T. Chang, S. H. Jo, K. H. Kim, P. Sheridan, S. Gaba y W. Lu, "Synaptic behaviors and modeling of a metal oxide memristive device," *Applied Physics A: Materials Science and Processing*, vol. 102, n.º 4, págs. 857-863, mar. de 2011. DOI: 10.1007/s00339-011-6296-1.

[304] D. Z. Du y K. I. Ko, *Theory of Computational Complexity: Second Edition*, 2nd. Wiley, 2014, vol. 9781118306, págs. 1-494. DOI: 10.1002/9781118595091.

[305] Y. Shi, L. Nguyen, S. Oh, X. Liu, F. Koushan y col., "Neuroinspired unsupervised learning and pruning with subquantum CBRAM arrays," *Nature Communications*, vol. 9, n.º 1, págs. 1-11, 2018. DOI: 10.1038/s41467-018-07682-0.

[306] J. Liang, S. Yeh, S. Simon Wong y H. S. Philip Wong, "Effect of wordline/bitline scaling on the performance, energy consumption, and reliability of cross-point memory array," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 9, n.º 1, págs. 1-14, 2013. DOI: 10.1145/2422094.2422103.

[307] Y. LeCun, L. Bottou, Y. Bengio y P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, n.º 11, págs. 2278-2323, 1998. DOI: 10.1109/5.726791.

[308] S. M. Rossnagel y T. S. Kuan, "Alteration of Cu conductivity in the size effect regime," en *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 22, American Vacuum Society AVS, ene. de 2004, págs. 240-247. DOI: 10.1116/1.1642639.

[309] D. Josell, S. H. Brongersma y Z. Tőkei, "Size-Dependent Resistivity in Nanoscale Interconnects," *Annual Review of Materials Research*, vol. 39, n.º 1, págs. 231-254, ago. de 2009. DOI: 10.1146/annurev-matsci-082908-145415.

[310] W. Steinhögl, G. Schindler, G. Steinlesberger, M. Traving y M. Engelhardt, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100 nm and smaller," *Journal of Applied Physics*, vol. 97, n.º 2, pág. 023706, ene. de 2005. DOI: 10.1063/1.1834982.

[311] K. Fuchs, "The conductivity of thin metallic films according to the electron theory of metals," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 34, n.º 1, págs. 100-108, 1938. DOI: 10.1017/S0305004100019952.

[312] A. F. Mayadas y M. Shatzkes, "Electrical-resistivity model for polycrystalline films: The case of arbitrary reflection at external surfaces," *Physical Review B*, vol. 1, n.º 4, págs. 1382-1389, feb. de 1970. DOI: 10.1103/PhysRevB.1.1382.

[313] G. C. Adam, A. Khiat y T. Prodromakis, *Challenges hindering memristive neuromorphic hardware from going mainstream*, dic. de 2018. DOI: 10.1038/s41467-018-07565-4.

[314] W. Yi, Y. Kim y J. J. Kim, "Effect of Device Variation on Mapping Binary Neural Network to Memristor Crossbar Array," *Proceedings of the 2019 Design, Automation and Test in Europe Conference and Exhibition, DATE 2019*, págs. 320-323, 2019. DOI: 10.23919/DATE.2019.8714817.

[315] A. Chen y M. R. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," en *IEEE International Reliability Physics Symposium Proceedings*, 2011. DOI: 10.1109/IRPS.2011.5784590.

[316] Q. Luo, X. Xu, T. Gong, H. Lv, D. Dong y col., "8-Layers 3D vertical RRAM with excellent scalability towards storage class memory applications," en *Technical Digest - International Electron Devices Meeting, IEDM*, Institute of Electrical y Electronics Engineers Inc., ene. de 2018, págs. 2.7.1-2.7.4. DOI: 10.1109/IEDM.2017.8268315.

[317] S. Pi, P. Lin y Q. Xia, "Cross point arrays of 8 nm × 8 nm memristive devices fabricated with nanoimprint lithography," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 31, n.º 6, 06FA02, nov. de 2013. DOI: 10.1116/1.4827021.

[318] J. Wang, X. Dong, Y. Xie y N. P. Jouppi, "I2WAP: Improving non-volatile cache lifetime by reducing inter- and intra-set write variations," en *Proceedings - International Symposium on High-Performance Computer Architecture*, 2013, págs. 234-245. DOI: 10.1109/HPCA.2013.6522322.

[319] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu y T. Huang, "Vortex: Variation-aware training for memristor X-bar," *Proceedings - Design Automation Conference*, vol. 2015-July, n.º c, págs. 1-6, 2015. DOI: 10.1145/2744769.2744930.

[320] D. C. Montgomery y G. C. Runger, *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 2010, pág. 784.

[321] E. Miranda, A. Morell, J. Muñoz-Gorriz y J. Suñé, "Simple method for monitoring the switching activity in memristive cross-point arrays with line resistance effects," *Microelectronics Reliability*, vol. 100-101, sep. de 2019. DOI: 10.1016/j.microrel.2019.06.019.

[322] F. L. Aguirre, S. M. Pazos, F. Palumbo, M. Antoni, J. Suñé y E. A. Miranda, "Assessment and Improvement of the Pattern Recognition Performance of Memdiode-Based Cross-Point Arrays with Randomly Distributed Stuck-at-Faults," *Electronics*, vol. 10, n.º 19, pág. 2427, oct. de 2021. DOI: 10.3390/electronics10192427.

[323] F. L. Aguirre, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "SPICE Simulation of RRAM-Based Crosspoint Arrays Using the Dynamic Memdiode Model," *Frontiers in Physics*, vol. 9, pág. 548, 2021. DOI: 10.3389/FPHY.2021.735021.

[324] F. L. Aguirre, N. M. Gomez, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "Minimization of the Line Resistance Impact on Memdiode-Based Simulations of Multilayer Perceptron Arrays Applied to Pattern Recognition," *Journal of Low Power Electronics and Applications 2021*, vol. 11, n.º 1, pág. 9, feb. de 2021. DOI: 10.3390/JLPEA11010009.

[325] F. L. Aguirre, S. M. Pazos, F. Palumbo, S. Fadida, R. Winter y M. Eizenberg, "Effect of forming gas annealing on the degradation properties of Ge-based MOS stacks," *Journal of Applied Physics*, vol. 123, n.º 13, pág. 134 103, abr. de 2018. DOI: 10.1063/1.5018193.

[326] F. L. Aguirre, N. Gomez, S. M. Pazos, F. Palumbo, J. Suñé y E. Miranda, "Line Resistance Impact in Memristor-based Multi Layer Perceptron for Pattern Recognition," en *2021 5th Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA), Aceptado para publicación*, 2021.

[327] F. L. Aguirre, F. Palumbo y P. Julian, "Piecewise-linear Modelling of CMOS Gates Propagation Delay as a Function of PVT Variations and Aging," Institute of Electrical y Electronics Engineers (IEEE), abr. de 2021, págs. 25-31. DOI: 10.1109/cae51562.2021.9397560.

[328] F. L. Aguirre, A. Padovani, A. Ranjan, N. Raghavan, N. Vega y col., "Spatio-Temporal Defect Generation Process in Irradiated HfO_2 MOS Stacks: Correlated versus Uncorrelated Mechanisms," en *2019 IEEE International Reliability Physics Symposium (IRPS)*, Monterey: IEEE, mar. de 2019, págs. 13-14. DOI: 10.1109/IRPS.2019.8720539.

[329] F. L. Aguirre, S. M. Pazos, F. Palumbo, S. Fadida, R. Winter y M. Eizenberg, "Impact of forming gas annealing on the degradation dynamics of Ge-Based MOS stacks," en *2018 IEEE International Reliability Physics Symposium*, Burlingame, EE.UU.: IEEE, mar. de 2018, págs. 21-25. DOI: 10.1109/CAMTA.2017.8058136, accepted for publication.

[330] F. L. Aguirre, S. M. Pazos, F. Palumbo, I. Krylov y M. Eizenberg, "Substrate influence on the behavior of capacitance hysteresis of III-V bilayered MOS stacks," en *2017 32nd Symposium on Microelectronics Technology and Devices (SBMicro)*, Fortaleza, Brasil: IEEE, ago. de 2017, págs. 1-4. DOI: 10.1109/SBMicro.2017.8112972.

[331] F. L. Aguirre, S. M. Pazos y F. Palumbo, "Experimental study of progressive breakdown in different conductance states of resistive switching structures," en *2017 1st Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA)*, Bariloche, Argentina: IEEE, feb. de 2017, págs. 1-4. DOI: 10.1109/PRIME-LA.2017.7899167.

[332] S. M. Pazos, F. L. Aguirre, F. Palumbo y F. Silveira, "Hot-carrier-injection resilient RF power amplifier using adaptive bias," *Microelectronics Reliability*, vol. 114, pág. 113 912, oct. de 2020. DOI: 10.1016/j.microrel.2020.113912.

[333] S. M. Pazos, S. Boyeras Baldomá, F. L. Aguirre, I. Krylov, M. Eizenberg y F. Palumbo, "Impact of bilayered oxide stacks on the breakdown transients of MOS devices: An experimental study," *Journal of Applied Physics*, vol. 127, n.º 17, pág. 174 101, mayo de 2020. DOI: 10.1063/1.5138922.

[334] S. Boyeras Baldomá, S. M. Pazos, F. L. Aguirre y F. R. Palumbo, "Breakdown transients in high-k multilayered MOS stacks: Role of the oxide-oxide thermal boundary resistance," *Journal of Applied Physics*, vol. 128, n.º 3, pág. 034 103, jul. de 2020. DOI: 10.1063/5.0012918.

[335] S. Pazos, F. Aguirre, F. Palumbo y F. Silveira, "Reliability-aware design space exploration for fully integrated RF CMOS PA," *IEEE Transactions on Device and Materials Reliability*, 2019. DOI: 10.1109/TDMR.2019.2957489.

[336] S. Boyeras, S. M. Pazos, F. L. Aguirre, H. Giannetta, C. Delgado y F. Palumbo, "Progressive breakdown on bi-layered gate oxide stacks," en *SBMicro 2019 - 34th Symposium on Microelectronics Technology and Devices*, Institute of Electrical y Electronics Engineers Inc., ago. de 2019. DOI: 10.1109/SBMicro.2019.8919480.

[337] S. M. Pazos, F. L. Aguirre, K. Tang, P. McIntyre y F. Palumbo, "Lack of correlation between C-V hysteresis and capacitance frequency dispersion in accumulation of metal gate/high-k/n-InGaAs MOS stacks," *Journal of Applied Physics*, vol. 124, n.º 22, pág. 224 102, 2018. DOI: 10.1063/1.5031025.

[338] S. M. Pazos, F. L. Aguirre, F. Palumbo y F. Silveira, "Performance-reliability trade-offs in short range RF power amplifier design," *Microelectronics Reliability*, sep. de 2018. DOI: 10.1016/J.MICROREL.2018.06.089.

[339] S. M. Pazos, F. L. Aguirre, S. Lombardo, E. Miranda y F. Palumbo, "Experimental Study of Progressive Breakdown in Different Conductance States of Resistive Switching Structures," en *China RRAM International Workshop 2017*, Soochow University, China, jun. de 2017.

[340] A. Fontana, S. M. Pazos, F. L. Aguirre y F. Palumbo, "Automatic ASET sensitivity evaluation of a custom-designed 180nm CMOS technology operational amplifier," en *2017 Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA)*, Buenos Aires, Argentina: IEEE, jul. de 2017, págs. 21-25. DOI: 10.1109/CAMTA.2017.8058136.

[341] S. M. Pazos, F. L. Aguirre y F. Palumbo, "Charge trapping effects on Metal-Gate/High-k/III-V MOS devices assessed through C-V hysteresis," en *2017 Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA)*, Buenos Aires, Argentina: IEEE, jul. de 2017, págs. 21-25. DOI: 10.1109/CAMTA.2017.8058135.

[342] S. M. Pazos, F. Palumbo y F. L. Aguirre, "Analysis and comparison of the CV-Dispersion of high-k, bi-layered MOS InGaAs/InP stacks," en *2017 1st Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA)*, Bariloche, Argentina: IEEE, feb. de 2017, págs. 1-4. DOI: 10.1109/PRIME-LA.2017.7899166.

[343] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, n.º 7, págs. 1895-1923, oct. de 1998. DOI: 10.1162/089976698300017197.