USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)

Tesis doctoral

DEVELOPMENT OF TOOLS FOR THE SIMULATION OF NANOMETRIC TRANSISTORS USING ADVANCED COMPUTATIONAL ARCHITECTURES

Presentada por:

Guillermo Indalecio Fernández

Dirigida por:

Antonio Jesús García Loureiro
Natalia Seoane Iglesias

Santiago de Compostela, junio de 2016

Antonio Jesús García Loureiro, Profesor Titular del Área de Electrónica de la Universidad de Santiago de Compostela

Natalia Seoane Iglesias, Investigadora Postdoctoral del Centro Singular de Investigación en Tecnoloxías da Información de la Universidad de Santiago de Compostela

HACEN CONSTAR:

Que la memoria titulada **Development of tools for the simulation of nanometric transistors using advanced computational architectures** ha sido realizada por **Guillermo Indalecio Fernández** bajo nuestra dirección en el Centro Singular de Investigación en Tecnoloxías da Información de la Universidade de Santiago de Compostela, y constituye la Tesis que presenta para optar al título de Doctor.

Santiago de Compostela, junio de 2016

Antonio Jesús García Loureiro
Director/a de la tesis

Natalia Seoane Iglesias
Director/a de la tesis

Guillermo Indalecio Fernández
Autor de la tesis

Antonio Jesús García Loureiro, Profesor Titular del Área de Electrónica de la Universidad de Santiago de Compostela

Natalia Seoane Iglesias, Investigadora Postdoctoral del Centro Singular de Investigación en Tecnoloxías da Información de la Universidad de Santiago de Compostela

como directores de la tesis titulada:

Development of tools for the simulation of nanometric transistors using advanced computational architectures

Por la presente DECLARAN:

Que la tesis presentada por Don **Guillermo Indalecio Fernández** es idónea para ser presentada, de acuerdo con el artículo 41 del *Regulamento de Estudos de Doutoramento*, por la modalidad de compendio de ARTÍCULOS, en los que el doctorando ha tenido participación en el peso de la investigación y su contribución fue decisiva para llevar a cabo este trabajo. Y que está en conocimiento de los coautores, tanto doctores como no doctores, participantes en los artículos, que ninguno de los trabajos reunidos en esta tesis serán presentados por ninguno de ellos en otras tesis de Doctorado, lo que firmamos bajo nuestra responsabilidad.

Santiago de Compostela, junio de 2016

Antonio Jesús García Loureiro	**Natalia Seoane Iglesias**
Director/a de la tesis	Director/a de la tesis

A mind needs books like a sword needs a whetstone.

Tyrion Lannister

Agradecimientos

Jamás habría llegado a ser doctor si no fuese por mi madre, porque ella me enseñó a leer y a sumar. A partir de ahí ya es fácil.

Santiago de Compostela, junio de 2016

Resumen

La tecnología electrónica tiene un profundo impacto en la sociedad y en la ciencia, aportando cada día nuevas soluciones tanto a nivel personal como profesional. En el caso particular de la ciencia, estas mejoras tecnológicas ofrecen la posibilidad de avanzar en nuevos campos y además a un ritmo mas rápido, mediante herramientas de todo tipo. La mayor parte de las mejoras están relacionadas con los transistores, que son el componente principal de cualquier dispositivo electrónico, como por ejemplo los procesadores (CPU), los procesadores gráficos (GPU) o la memoria volátil (RAM). Estos elementos se diseñan, fabrican y venden utilizando transistores cada vez más avanzados, lo que permite ofrecer en general un producto más rápido, con menos consumo de energía, más pequeño, o más barato. Los expertos de esta industria publican periódicamente el ITRS (International Technology Roadmap of Semiconductor), una hoja de ruta que trata de caracterizar la evolución que debe realizarse en los materiales y procesos para poder mantener el ritmo de avance de la industria de transistores. El ITRS también analiza los problemas que surgen de la miniaturización de los mismos. Utilizando este documento, los investigadores deben hacer frente a los problemas de manera anticipada, para que estos no obstaculicen el avance de las soluciones tecnológicas. Una herramienta poderosa para afrontar estos problemas son las simulaciones, que permiten ahorrar mucho tiempo y dinero, al proporcionar una estimación de cómo se comportará un dispositivo sin necesidad de crearlo en la cadena de producción.

Para analizar correctamente un dispositivo mediante técnicas de simulación, éstas tienen que ser lo más precisas posible. El modelo de arrastre-difusión, que calcula las corrientes de arrastre y la de difusión usando diversas aproximaciones, es una solución rápida y simple. Si se acopla con correcciones para el confinamiento cuántico, como el modelo de gradiente de densidad, puede simular correctamente las características sub-umbral del dispositivo, incluso con tamaños de puerta del orden de nanómetros. Existen otros modelos más precisos como el

método Monte Carlo que considera las partículas de manera individual o como meta-partículas, y tiene en cuenta los procesos de dispersión que sufren a través del dispositivo. Con este modelo, se obtiene buena precisión especialmente en el régimen on, a costa de ser bastante más costosa computacionalmente que la solución de arrastre-difusión. Finalmente, utilizar funciones de Green fuera de equilibrio para resolver el transporte cuántico con la ecuación de Schrödinger, da lugar a uno de los métodos con más precisión de los simuladores disponibles. Como era de esperar, este método es todavía más costoso computacionalmente que los anteriores.

En nuestro caso particular, mediante la simulación de transistores queremos analizar el problema de las fuentes de variabilidad que surgen en el proceso de fabricación de los mismos, porque tienen un gran impacto en el rendimiento del dispositivo, dando lugar incluso a fallos de funcionamiento. Para realizar un análisis fiable necesitamos seleccionar una técnica de simulación que nos permita desplegar tantas simulaciones como sea posible, pero que por otra parte sea lo suficientemente precisa como para extraer información significativa. Seleccionamos el simulador basado en el modelo de arrastre-difusión con correcciones cuánticas como el candidato adecuado para empezar este análisis.

Teniendo en cuenta lo anterior, vamos a centrar nuestro trabajo en dos frentes diferentes: por un lado, estudiar las fuentes de variabilidad que se presentan en las arquitecturas modernas de dispositivos electrónicos y caracterizar su efecto. Por otra parte, desarrollar las herramientas computacionales que necesitamos con el fin de poder gestionar miles de simulaciones y procesar los resultados.

Las fuentes de variabilidad surgen como diferencias respecto de la definición del dispositivo que se quiere fabricar y el resultado final. Estas desviaciones aleatorias son de dos tipos: inherentes al material, o relacionadas con etapas del proceso de fabricación. Es prioritario comprender el efecto que tienen estas desviaciones en el comportamiento del dispositivo, porque normalmente su efecto se agrava con la miniaturización del mismo. Puesto que estas fuentes de variabilidad son diferencias respecto de la definición del dispositivo ideal, se ha decidido que las modificaciones que se realicen del simulador no afecten al núcleo del mismo, sino que sólo alteren la estructura del dispositivo. De esta manera, se han podido aplicar las fuentes de variabilidad tanto a un simulador de Monte Carlo como a uno de arrastre-difusión. Por otro lado nuestro enfoque es modelar de la manera más realista posible las fuentes de variabilidad, para que estas alteraciones de la estructura del dispositivo sean fiables. Debido a la naturaleza aleatoria de las fuentes de variabilidad, es necesario dar soporte a la realización de cientos o miles de simulaciones para tener unos resultados estadísticamente sólidos, y por tanto una buena caracterización de los parámetros en juego.

La metodología desarrollada utiliza un proceso de perturbación que consta de tres componentes:

- El perfil de perturbación es cualquier fichero o recurso que indica cómo se debe modificar el dispositivo. Este fichero permite abstraer la fuente de variabilidad del simulador y representa una perturbación única del dispositivo. Para analizar una fuente de variabilidad, se generan tantos perfiles como simulaciones se deseen.

- El generador de perfiles es un código externo que se encarga de crear los perfiles atendiendo al tipo de variabilidad que se quiera estudiar, y también a los parámetros que la caracterizan. En nuestro caso, este generador suele estar programado en Matlab.

- El lector de perfiles es una modificación en el código del simulador que se encarga de cargar y aplicar el perfil de perturbación, independientemente de la naturaleza del mismo. Esta modificación del código del simulador es muy simple dado que solamente debe encargarse de leer un único perfil de perturbación y modificar el dispositivo como sea necesario.

Hemos aplicado esta metodología basada en perturbaciones a dos fuentes de variabilidad diferentes: Line Edge Roughness (LER) y Metal Gate Granularity (MGG). En los artículos presentados hemos aplicado estas fuentes de variabilidad exitosamente en una amplia variedad de escenarios: distintas arquitecturas como nanohilos y FinFETs, distintas aleaciones como InGaAs o Silicio, varios materiales de puerta como TiN, TaN o WN, y dos métodos de simulación, arrastre-difusión con correcciones cuánticas, y Monte Carlo.

La naturaleza de LER son las irregularidades que aparecen en el líneas de un dispositivo respecto a la forma recta ideal. En general, cualquier interfaz entre los materiales del dispositivo es un candidato a padecer este tipo de variabilidad, debido a que su origen es el propio proceso litográfico. Este efecto aumenta según se reducen las dimensiones del dispositivo si no se mejora el proceso litográfico, por tanto es muy importante caracterizarlo adecuadamente.

Nuestra aproximación fue utilizar una transformada inversa de Fourier con un espectro de ruido con distribución gausiana o exponencial. El espectro de ruido caracteriza las deformaciones que sufre la línea original del dispositivo, pero en el espacio de frecuencias. Medidas experimentales sobre imágenes TEM avalan las dos distribuciones seleccionadas. Esta transformada inversa recupera la información del espacio de frecuencias al espacio real, y por tanto genera un perfil de deformación que indica en qué cantidad se va a deformar una línea concreta del dispositivo. El lector de perfiles debe encargarse de la modificación de la malla que define al dispositivo de manera que no se generen tetraedros degenerados, y el resto de la simulación puede realizarse como si no hubiese fuente de variabilidad alguna.

Hemos analizado el efecto que tiene sobre el comportamiento del dispositivo los parámetros que definen el espectro de ruido, que son la altura cuadrática media (Δ), y la longitud de correlación (Λ). En todos los casos se ha aplicado en la dirección de transporte de carga, puesto que es la contribución más importante que genera esta fuente de variabilidad. Usando esta técnica se ha estudiado el efecto del LER en varios dispositivos, y se ha comparado el efecto cruzado de cambiar la aleación del semiconductor y el tamaño del mismo.

Además de LER, también hemos aplicado nuestra metodología a MGG. En este caso, la naturaleza de la variabilidad son los dominios, o granos, que surgen en el metal con el que se fabrica el contacto de la puerta del dispositivo. Entre otras tecnologías que se han desarrollado para aumentar la capacitancia del contacto de puerta, se encuentra el conjunto de dieléctrico con high-κ y puerta metálica. Esta solución está siendo aplicada ampliamente, pero tiene la contrapartida de que en el contacto metálico surgen dominios que tienen distinta orientación cristalográfica. Estos dominios, que tienen formas y orientaciones aleatorias, dependen del material depositado, y además presentan distintos valores de función de trabajo, lo que tiene un efecto perjudicial sobre la variabilidad del dispositivo.

Para modelar esta fuente de variabilidad, una de las opciones es dividir la puerta del dispositivo como si estuviera compuesta por varias puertas en paralelo, y aplicar un modelo analítico para tener en cuenta el efecto de esta partición. Este método es sólo aplicable para los MOSFETs, y es una primera aproximación, pero carece de la precisión necesaria para abordar el problema cuando el tamaño del dispositivo se reduce por debajo de un cierto umbral, que es precisamente el rango que nos interesa estudiar. Otro enfoque es modelar la puerta mediante granos cuadrados que cubran el área de la puerta, y aplicarle a cada uno de estos granos un valor distinto de función de trabajo, para luego simular el dispositivo. Estos cuadrados pueden tener diferentes tamaños, y orientaciones, según el material que se quiera simular. El principal inconveniente de esta técnica es que los granos reales tienen una forma artificiosa, no cuadrada, y aunque hay otros enfoques donde se intenta ajustar la distribución de granos para contrarrestar esta carencia, unos granos de forma cuadrada no van a representar adecuadamente los resultados experimentales. El enfoque más costoso y preciso es el uso de imágenes de TEM del material con el fin de tener un patrón que pueda ser aplicado a la simulación. Este enfoque requiere imágenes TEM como datos de entrada, por lo que se ve limitado por la disponibilidad de los mismos.

Nuestra aportación es el algoritmo de Voronoi. Esta técnica se ha diseñado para imitar el proceso de deposición de metal, en la que puntos de nucleación se definen por los primeros átomos que llegan a la superficie, y los siguientes átomos se concentran alrededor de ellos. Los dominios surgen de la concentración de átomos alrededor de puntos de nucleación, y un diagrama de Voronoi reproduce exactamente esa estructura. La ubicación aleatoria de los puntos de nucleación, junto con la orientación aleatoria que cada dominio recibe acorde con el material en estudio, permite crear varios perfiles de perturbación para cada conjunto de parámetros. Para el caso de MGG, la parámetros involucrados son el tamaño medio de los granos, que es controlado en nuestro caso a través del número de puntos de nucleación, las posibles orientaciones, su probabilidades y la función de trabajo que tiene cada orientación.

Utilizando este método, es decir simulando la partición del contacto de puerta en dominios, la distribución de tamaños de los mismos sigue de manera natural una distribución Gamma. Hemos demostrado esta afirmación por medio de datos experimentales, comparando la distribución de tamaños visible en imágenes TEM de distintos materiales con la distribución que surge de nuestro modelo, Gamma. Los resultados apoyan nuestra aproximación sobre otras soluciones como el modelo de Rayleigh propuesto por otros investigadores, que también analizamos con el mismo mecanismo y resultados experimentales, pero que resultó ser inadecuado para representar esta fuente de variabilidad. Este enfoque ha sido probado con diferentes materiales de compuerta, como el TiN, TaN y WN. También ha sido verificado en dispositivos y materiales semiconductores diferentes, y los resultados publicados en diversas revistas.

Con el fin de tener más información sobre el comportamiento intrínseco del dispositivo en virtud de las fuentes de variabilidad, hemos desarrollado una herramienta matemática, el mapa de sensibilidad de fluctuaciones (FSM). Utilizando el FSM es posible determinar qué partes del dispositivo son más sensibles a una cierta fuente de variabilidad, pudiendo saber de qué manera se ve afectada una figura de mérito ante un perfil de perturbación concreto. Esta sensibilidad espacial se puede calcular para diferentes figuras de mérito, como tensión umbral o corriente en las zonas on y off, y también para diversas fuentes de variabilidad. El FSM es una característica única de cada dispositivo una vez fijada la figura de mérito y la fuente de variabilidad, de tal manera que comparando el FSM de varios dispositivos obtenemos una relación entre los propios dispositivos. Finalmente, es posible utilizar el FSM para realizar predicciones sobre el comportamiento del dispositivo ante un conjunto de perfiles de perturbación. Esto permite obtener una estimación de los parámetros del dispositivo sin tener que llegar a simularlo, lo que la convierte en una primera aproximación de muy bajo coste computacional y con una precisión adecuada.

Cuando se estudia la variabilidad de dispositivos semiconductores a través de la simulación numérica, nos introducimos en el campo de los estudios estadísticos, en el sentido de que tendremos una mayor precisión en los resultados a medida que aumentemos el número de simulaciones , es decir la carga computacional de trabajo que estamos utilizando. Esta situación se da en otros campos de investigación, como oceanografía, biología, ingeniería civil, y normalmente se resuelve creando una infraestructura adaptada al problema concreto, lo que conlleva que la solución esté ligada al problema resuelto, no siendo así aplicable en otros campos, y generalmente tampoco se puede adaptar a recursos computacionales distintos. Se han desarrollado también soluciones genéricas que actúan como un middleware o como una plataforma científica, pero igualmente presentan dificultades para abordar problemas nuevos, o para ser adaptadas a recursos computaciones distintos de los inicialmente previstos.

Nuestro objetivo es reducir el tiempo de simulación, con el fin de obtener los resultados tan pronto como sea posible, pudiendo así realizar más simulaciones. En nuestro caso de análisis de variabilidad, aumentar el número de simulaciones nos va a permitir caracterizar más adecuadamente el efecto de la misma en el dispositivo. La principal dificultad es que, normalmente, los recursos computacionales disponibles son incompatibles entre sí, y por tanto no se pueden lanzar simulaciones en todos ellos de una forma totalmente inmediata. Para resolver este problema, hemos creado cuatro herramientas que permiten procesar eficientemente cientos o miles de simulaciones: el TaskManager as a Service, el General Workload Manager, el Auto-calibrador, y la reescritura del núcleo del simulador para utilizar OpenCL.

Para caracterizar el TaskManager as a Service, hemos utilizado el enfoque que se adopta en computación en la nube, es decir, una taxonomía de modelos de computación que comúnmente consiste en la Infraestructura como Servicio (IaaS), Plataforma como servicio (PaaS) y Software como Servicio (SaaS). En todos estos modelos de computación en la nube se presenta una interfaz al usuario, y se abstrae el contenido de las capas inferiores, definiendo así un servicio nuevo. Por ejemplo, el IaaS abstrae el hardware de varios equipos a través de las máquinas virtuales, y le ofrece al usuario la posibilidad de poner en marcha y administrar máquinas virtuales. Hemos presentado por tanto un modelo de computación que se adapta a esta taxonomía para mantener un lenguaje común con otros investigadores.

La idea detrás de la TMaaS es aislar el acceso a los recursos informáticos, y ofrecer al usuario la posibilidad de definir y gestionar tareas computacionales. En cada tarea computacional hay que definir un conjunto de componentes: el entorno de ejecución, la aplicación que se desea lanzar y el conjunto de recursos de entrada y de salida. Estos componentes deben ser proporcionados por el usuario para que el TMaaS pueda gestionar la tarea de manera transparente en los recursos computacionales disponibles, sean estos o no homogéneos. Por un lado el TMaaS se encarga de la comunicación con el sistema de colas o sistema operativo que esté instalado en cada recurso computacional, al igual que del despliegue de máquinas si se trata de un recurso de computación en la nube, y de la gestión y monitorización de la tarea concreta. Por otro lado, el TMaaS ofrece al usuario el control de las tareas, para que pueda gestionarlas, independientemente de la naturaleza de las mismas. De esta manera resolvemos el problema de que la solución quede ligada a un campo concreto.

Para implementar y probar el TMaaS hemos desarrollado el General Workload Manager (GWM). Esta herramienta cumple con los requisitos antes mencionados, y permite al usuario utilizar los recursos informáticos heterogéneos de una manera transparente. El GWM tiene una arquitectura cliente-servidor, y utiliza REST para comunicar ambos actores, lo cual permite descubrir las características de la herramienta con facilidad. Como cliente, hemos desarrollado dos versiones: un cliente de línea de comandos que permite gestionar el sistema completo desde un terminal UNIX, y un cliente habilitado para web que permite al usuario controlar el comportamiento del servidor desde un navegador web. Esta aplicación web se ha construido con tecnologías modernas para que la comunicación con el servidor sea mínima, proporcionando una experiencia sólida y rápida para el usuario.

La estructura del GWM ha sido diseñada para que sea expansible, de tal modo que pueda proporcionar soporte a distintos recursos computacionales de manera transparente. Mediante esta estructura, se han implementado módulos para el GWM de comunicación con varios shells, como bash, sh o ksh, y para comunicarse con varios sistemas de colas, como PBS/Torque o SGE. Para aprovechar las soluciones modernas de cloud computing de IaaS, también hemos implementado el soporte con varios proveedores de cloud computing, incluyendo CloudStack, OpenStack, y Amazon EC2, de tal manera que un usuario puede solicitar la instanciación de nuevos recursos computacionales en cualquiera de estas plataformas, y el GWM los muestra de manera transparente para la ejecución de las tareas definidas.

Utilizando el GWM hemos sido capaces de realizar la mayoría de las simulaciones que se presentan en esta tesis en tres clústeres de HPC, que tienen tanto el hardware como el sistema de colas incompatible entre si. En cualquier caso, el usuario sólo tuvo que definir la tarea que quería que se ejecutase, y el GWM se encargó del lanzamiento y monitorización de la tarea en los recursos computacionales disponibles.

Otra de las soluciones desarrolladas para abordar el problema de cálculo es un auto-calibrador. Todas las simulaciones de dispositivos electrónicos presentados en esta tesis necesitan ser calibradas con alguna fuente externa. Por lo general, se utilizan datos experimentales cuando están disponibles, pero también se puede calibrar contra datos de simulaciones más precisas, como NEGF o Monte Carlo. En ambos casos, la calibración requiere que el usuario averigüe los parámetros de entrada del simulador mediante ensayo y error. Este proceso es costoso y lento. Para mejorarlo hemos desarrollado un auto-calibrador que utiliza un algoritmo genético para encontrar los valores de los parámetros que ajustan el comportamiento del dispositivo a la curva de calibración deseada. Esta herramienta utiliza el GWM como infraestructura para desplegar los cientos o miles de tareas que serán necesarios hasta alcanzar un calibrado suficientemente preciso. Los resultados obtenidos con este auto-calibrador han sido muy satisfactorios, con curvas de calibración más ajustadas que cuando se calibra manualmente, y sin interacción del usuario alguna, más allá de definir el dispositivo, la curva de calibración deseada y los valores iniciales de los parámetros.

El simulador que estamos utilizando está implementado en C, utilizando MPI para comunicar los nodos de computación de memoria distribuida que se quieren utilizar. Esta implementación está muy bien probada y optimizada, así que no hay mucho margen de mejora posible. No obstante, nuevas arquitecturas como unidades de procesamiento gráfico de propósito general (GPGPU) o aceleradores como el Intel Xeon Phi, están surgiendo como una buena alternativa para alcanzar rendimientos muy elevados. Estas arquitecturas están más orientadas a sistemas con matrices densas, puesto que el modelo de computación de hilos que presentan favorece una carga de trabajo homogénea entre ellos. En nuestro caso, dado que utilizamos elementos finitos en los simuladores que ejecutamos, nuestras matrices son dispersas, lo que da lugar a un problema más complicado y no tan explorado. Para utilizar estas nuevas arquitecturas, hemos implementado las operaciones del núcleo de los simuladores, que es la parte más costosa computacionalmente, en OpenCL, un lenguaje que permite ejecutar código en paralelo en arquitecturas GPGPU o Xeon Phi, entre otras. Este trabajo es preliminar, pero ya hemos realizado algunas publicaciones con los resultados obtenidos y se presentan en la bibliografía.

En conclusión, el autor empezó esta tesis con el objetivo de avanzar el conocimiento existente en dispositivos semiconductores nanométricos. Concretamente seleccionó el análisis de variabilidad como un problema que exige una combinación interesante de diversas habilidades. Por una parte, requiere conocimiento de los mecanismos físicos que afectan al comportamiento de los semiconductores, y también de los procesos de fabricación, debido a su impacto en la variabilidad bajo estudio. Por otra parte, requiere herramientas potentes para simular miles de simulaciones y así comprender el efecto de las fuentes de variabilidad. Durante el desarrollo de esta tesis se han estudiado dos fuentes de variabilidad distintas, utilizando un simulador de arrastre-difusión y otro de tipo Monte Carlo. Estas fuentes de variabilidad se han estudiado en distintos tipos de dispositivos electrónicos, con distintas aleaciones y con varios tamaños de puerta diferentes. Finalmente, se han desarrollado herramientas novedosas con las que poder desplegar las simulaciones en recursos computacionales heterogéneos y optimizar el tiempo de simulación.

Contents

Chapter 1

Introduction

1.1 Motivation

Electronic technology has a deep impact in today's society, as well as in science. Society has introduced new several solutions in both personal and professional environments. Similarly, scientific research of all kinds take advantage of the possibilities that technology provides. Modern improvements had provided science the tools it needs to advance at a faster pace. A representation of how important this factor is in modern society and science, is the high economical impact that several technological corporations have in the worldwide market.

Most of these improvements are backed up by transistors, which are the main component of any digital electronic device, specifically of central processing units (CPUs), graphic processing units (GPUs), and volatile memory (RAM). Foundries design, manufacture and sell transistors as a component for digital devices. These foundries rely on cutting edge knowledge to provide faster, less power consuming, smaller or cheaper solutions. To achieve these improvements, there has to be advance in the many steps of the fabrication process [1].

In order to foresee the evolution of transistors, hence technology, a group of semiconductor industry experts publish the ITRS [2, 3], a road map that characterizes the evolution that transistors have to follow in order to maintain the desired rhythm of advance. Problems that may arise due to the continuous miniaturization of the transistors are also explained in this document. Using the ITRS, researchers can try to tackle the foreseen problems before they actually occur, so they do not hinder the advance of technology.

Semiconductor device simulations are a powerful tool that allow scientists to save time and money, by being able to predict how a device will behave without the need to create the manufacturing pipeline [4–6]. In order to understand the behavior of the real device, the simulation process has to be as precise as possible. The drift-diffusion approach, which calculates only the current and moment conservation of the carriers, is a simple but fast solution. When coupled with corrections for the quantum confinement like density gradient [7], this method, once calibrated, is able to accurately simulate the subthreshold characteristics of state-of-the-art semiconductor devices in the nanometre regime. The next step in complexity could be the hydrodynamic approximation. This model is similar to the previous one, but includes out of equilibrium effects that improve the simulation in certain situations. A more complex simulation methodology is to use Monte Carlo, which considers the particles individually or as meta-particles, and the scattering processes along the device, to obtain a very good precision, specially in the on regime [8–10]. The downside of this approach is that each simulation is very costly in comparison with drift-diffusion. An even more precise simulation method is based on Non-Equilibrium Green Functions and it solves the quantum transport with the Schrödinger equation [11]. As expected, this simulator is the most costly of the ones presented.

One of the problems that we want to simulate, and hence give information back to the scientific community and foundries, is the variability sources that appear in the process of manufacturing the nanodevice [12]. This has a very big impact on the devices behavior, decreasing their performance or some times generating operational failures [12–15].

In order to characterize the variability as well as possible, we have to run thousands of simulations, to obtain a more reliable statistical insight on the nature and effect of the variability sources [16]. Therefore, the selected simulation technique has to be simple enough to allow us to deploy as many simulations as possible while keeping an accuracy level that grants us meaningful information. In our case that will be the drift-diffusion simulator with quantum corrections, calibrated against experimental data when possible.

Another problem that we also want to tackle is the lack of general solutions that allow a scientist to easily manipulate the computing capabilities needed in order to launch thousands of simulations, or any other large workload. The existing solutions are too complex, or tailored to certain problems and limited by their infrastructure.

In summary, we want to focus our work in two different fronts: i) to study the variability sources that arise in modern nanodevice architectures, characterizing them and their effect on the devices, and ii) to develop the computational tools that we need in order to be able to manage thousands of simulations and post process the results.

1.2 Variability sources

Once the semiconductor nanodevice is defined and ready to be produced, certain deviations from the blueprints are to be expected. These deviations are random, and can be of two types: related to different stages of the building process, or inherent to the semiconductor material and physics. The effect of these deviations on the behavior of the device is called variability, and the nature of the deviation is the variability source. These intrinsic fluctuations [17], increase when the device is scaled down, which aggravates its importance.

We want to study different variability sources, and how are they related to the scaling of the device. Each variability source under study will have an impact on the device characteristics, that will depend on the parameters that characterize the variability source. Studying the relation between those parameters and the impact on the device characteristics, we can conclude which steps had to be taken in order to minimize the negative effect of the variability source on the device behavior. Similarly, this allows us to compare the variability sources between themselves.

To apply the variability source, considering that their nature is the deviation from the ideal device, we modified the source code of the numerical simulator to account for the difference. Our approach has to be as much realistic as possible, without modifying the simulator more than necessary. All the modifications in the code have to be possible to deactivate, in order to restore the original behavior. Also, because the variability is a statistical process, we need more than one simulation to account for the effect of the variability source. More concretely, considering that some parameters that characterize the variability are not fixed but also are variables, we may want to deploy hundreds or thousands of simulations to have good statistics and a proper characterization of the variability source.

The methodology chosen is common to all the variability sources under study: we analyze the effect of the variability via a perturbation process. This perturbation methodology is composed of:

1. The **perturbation profile** is any kind of file or set of files that represent how the device has to be perturbed. This allows to take the actual variability source out of the simulator, so a single compilation of the simulation can deal with different instances of perturbations. This perturbation profile is generated offset, and deployed with the simulator and the corresponding device characteristics, like the mesh, in order to have a full simulation of the source under variability.

2. The **profile generator** is an external code, that using the variability parameters is able to generate a profile that represents how the device has to be perturbed. This profile generator usually creates not one, but hundreds or thousands of profiles. The variability parameters and the nature of this specific source of variability is treated in this stage, so the simulator does not have to account for the details of the variability that is being studied. In our case, this profile generator has been programmed in Matlab.

3. The **profile loader** in the simulator is an addition to the code base of the simulator that will load the perturbation profile and modify the device accordingly. This profile loader is oblivious to the characteristics of the perturbation that is being applied. Also, even if the user wants to simulate hundreds of perturbations in order to get statistics, the profile loader only has to deal with one at the time. This allows the modification in the code base to be as small as possible, to be of little intrusion to the other developers that work with the same code.

We have applied this methodology to two different variability sources: Metal Gate Granularity (MGG) and Line Edge Roughness (LER). The same methodology is valid for different device structures. For instance, it has been applied to InGaAs and Silicon nanowires [18, 19], and InGaAs and Silicon FinFETs [20, 21]. Since this perturbation is not an integral part of the simulation, the application to different simulation engines is straightforward, like drift-diffusion [19] or Monte Carlo [22]. Detaching the profile from the simulator allows for a single compilation of the code, less maintenance of the source, and also allows for the combination of variability sources. Next we present these variability sources and their main characteristics.

1.2.1 Line Edge Roughness

The nature of the Line Edge Roughness is the irregularities that appear in the lines of a device from the ideal straight shape. In general, any interface between materials created via spacers in the lithography process is a candidate to suffer this variability. If the patterning is resist-defined, the result is a random uncorrelated deformation in the line, and for spacer-defined patterning, the shape of the deformation get transferred first to a dummy spacer, and from there to the Fin, generating a correlated deformation [15, 23, 24]. This variability is found in the several lines of FinFET devices [25], in MOSFET devices [26], and in other devices [27].

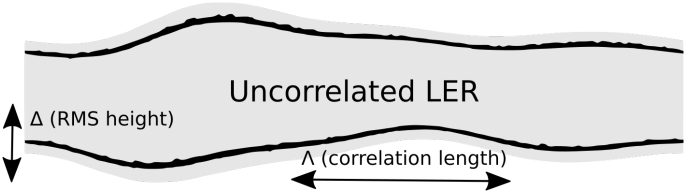

Figure 1.1: Representation of uncorrelated LER applied for a FinFET device. A cross section of the device body is shown.

LER is a source of variability that will worsen as the device is scaled down, so it has to be studied and mitigated [28]. This shape has been observed via TEM images and can be characterized with an inverse Fourier transformation of a noise profile. This characterization of the TEM images also allows to generate [26, 29] the required deformation profiles to be used in our simulator.

Considering a power spectra $S(k)$, the deformation height can be calculated from a set of random phases $\phi(k)$, such that:

$$H(x) = F^{-1}S(k)\phi(k),$$

being F^{-1} the inverse Fourier transformation from the wavelength space to the real space. This transformation will depend on the random phases, which will give us different possible perturbations for a given power spectra. Also, the power spectra will depend on some parameters, and will also have a certain functional dependency.

We have analyzed two different power spectra: Gaussian and exponential, as suggested by [26]. In both cases, we are using two parameters to account for the variability. The root mean square height of the deformation, Δ, represents how much the line is deformed in average. The correlation length of the spectra, Λ, represents the spatial frequency of the deformation. Small values of Λ represent elongated deformations, where big values of Λ will correspond to shorter ones. In Figure 1.1, an example a LER deformation applied to a device is shown to clarify this parameters.

The expressions for the Gaussian and Exponential spectra are:

$$S_G(k) = \sqrt{\pi}\Delta^2\Lambda e^{-(k^2\Lambda^2/4)}$$

and

$$S_E(k) = \frac{2\Delta^2\Lambda}{1+k^2\Lambda^2}.$$

We applied the LER deformation along the body of the device, which is called Fin Edge Roughness (FER), because is the most important contribution to the variability. Another applications of LER are to be explored in future work, especially when changing the shape of the device, which could unbalance the relative effect of each LER option.

The perturbation profile for this variability source is a file representing how much the device has to be deformed. Fixing the Δ and Λ values, we can generate several perturbation devices by introducing different random phases $\phi(k)$. The profile loader has to be able to deform the device and keep the mesh quality, which means no degenerated tetrahedra or close to degeneration should be created. This is achieved by doing a gradual deformation of the device and monitoring the tetrahedra, so if the deformation is not possible, the user is warned.

1.2.2 Metal Gate Granularity

A technology that has been used in production and is still projected to smaller device sizes, is the metal/high-κ gate stack. This metal contact in the gate exhibits a problem that gains importance in deca nanometre devices: the metal has domains with different orientations [30]. These domains will depend on the material, and each domain has a different work function. The difference in work function implies that the behavior of the device will depend on the grains that compose the gate and their orientation. The impact of this variability in SRAM cells was studied [12], and it was confirmed that it is comparable or worse than the effect of LER. Similar studies for single transistors [31–33] present the same conclusion.

This metal grain pattern and its effect on the device behavior is the nature of the MGG variability source. Several approaches model this variability source. One of the options is to partition the gate of the device as if it was composed of several gates in parallel, and apply an analytic model to account for the effect of this partition. This approach is only applicable for MOSFETs, and it is a first approach to this variability, but lacks the precision necessary to tackle the problem for smaller devices [12, 33, 34].

Another widely used approach is to model the grains of the gate as squares that span the area of the gate. These squares can have different sizes, and so they can take into account the fact that the metal grains have not only random placement and orientation, but also random sizes around a given mean value [31]. The main downside of this technique is that the grains are always presented as squares, and this is not the observed behavior in nature. Other approaches [35] try to use an artificial distribution of grain sizes to better describe the behavior of the device.

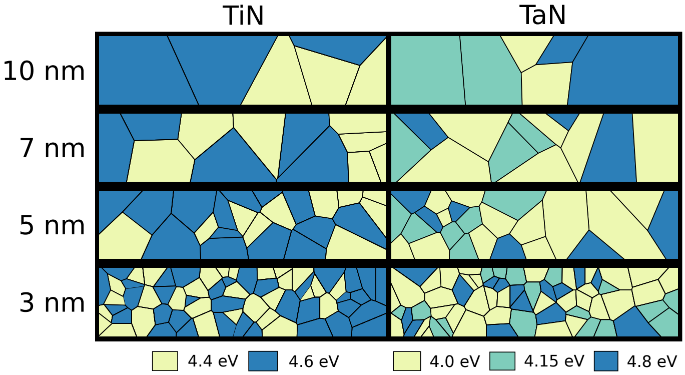

Figure 1.2: Example of Metal Gate Granularity perturbation profiles for different materials and grain sizes.

The most costly approach is to use TEM images of the material in order to have a pattern that can be applied to the simulation. This approach requires TEM images as input data, so it is limited by the availability of that data [32].

We base our approach in trying to model the experimental data in the most realistic possible way, like the TEM images, but allowing for thousands of simulations without much overhead. Because of that, we have developed the Voronoi model [36,37] of perturbations for the gate. This algorithm consists on the definition of a random set of points in the surface of the gate contact, randomly placed, r_i. Once the points have been located, we define the grains as the regions of the gate surface r such that:

$$G_i = \left\{ r | d(r, r_i) < d(r, r_j) \forall j \right\},$$

with d being the distance between two points measured along the gate surface. This is the definition of a Voronoi diagram, which divides the surface in regions such that the points in each region are closer to the related randomly placed point that to the other points.

We show several perturbation profiles in Figure 1.2, with different mean grain sizes and two different materials, using our Voronoi approach. Once the material is chosen, the number of orientations, their relative probability and the work function of each one changes.

This algorithm mimics the behavior of the metal deposition stage, in which nucleation points are defined by the first atoms that reach the surface, and the next atoms gather around them and define a single orientation. The random location of the nucleation points, along with the random orientation that each grain receives after the grain boundary is defined, allows to generate several perturbation profiles from a single set of parameters. For the case of MGG, the parameters involved are the mean grain size, that is controlled in our case with the number of nucleation points, the possible orientations, their probabilities and the work function that each orientation has.

Using this method to generate the grains, their area distribution arises naturally as a Gamma distribution. We have checked with experimental data to compare the actual grain area distribution visible in TEM images with the grain area distribution that arises from our model [38]. The results support our model over other solutions like the Rayleigh model [39,40].

This approach has been tested with different gate materials, like TiN, TaN and WN. Also, with different devices and semiconductor materials, and several publications present the obtained data [19–21,36,38,41,42].

1.2.3 FSM, a tool for variability analysis

In order to have more information about the intrinsic behavior of the device under variability sources, we have developed a mathematical tool which creates a fluctuation sensitivity map (FSM) that registers how sensitive certain parts of the device are under the perturbation that they suffer when a given variability source is being applied. The sensitivity can be calculated for different figures of merit, like threshold voltage or off current. For a given figure of merit, and a variability source, the FSM will be unique to the device under study, so comparisons between FSMs of different devices provide interesting information about how they react to the variability source. In certain cases, because the FSM represents the sensitivity of the device, a prediction can also be carried out, in which the variability of the figure of merit can be calculated by using the FSM and the perturbation profiles that are going to be used.

We have applied the FSM to analyze the MGG variability. In this case, the FSM takes the shape of a matrix that represents each of the points of the discrete gate contact. After simulating an ensemble of perturbation profiles, we can calculate the FSM with the following procedure, which we present particularized to the MGG variability and its effect in the threshold voltage:

Let V_i be the threshold voltage that results for each of the perturbation profile. Let $f : (u,v) \to (x,y,z)$ a continuous function that maps the elements from the FSM matrix to the points in the gate surface, and let $WF_i(x,y,z)$ be the work function that is present in the given coordinates of the gate. For each point of the matrix, (u,v), we can do the following least

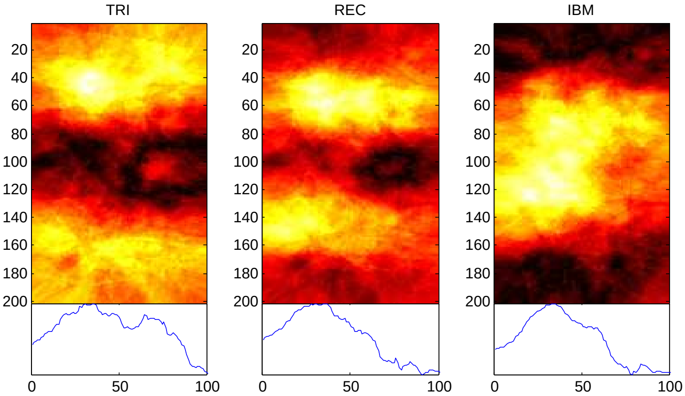

Figure 1.3: FSM applied to three different devices over the threshold voltage figure of merit.

squares linear fit:

$$V_i \sim WF_i(f(u,v)),$$

which will return a different slope $m(u,v,V,WF)$ for each of the matrix elements, so we define $FSM_{u,v}(V,WF) = m(u,v,V,WF)$.

We present in Figure 1.3 the result of applying this algorithm to the threshold voltage in three similar devices, all of them a representing a 10.4 nm gate length InGaAs FinFET transistor. The image from the left corresponds to a triangular body shape, the center image is a rectangular body shape with a big buffer of oxide at the top of the gate, under the contact, and the last image is a rounded Fin. The figures represent the gate sensitivity, such that the center of the figure corresponds to the top of the gate, and the extremes of the figure with those of the gate. Usually the most sensitive part of the gate (light color in the figure) is in the sides close to the top of the gate. Both in the TRI and REC devices, this sensitivity is reduced in the apex of the contact. In the first case, due to the narrowing of the body, and in the second one, because of the buffer of oxide. More details are shown in the published article [43].

1.3 Computational problem

When studying the variability of semiconductor devices via numerical simulation, we are stepping in the field of the statistical studies, in the sense that we are going to have more precision in our results as we increase the computational workload that we are deploying. This kind of problem is also present in other areas of science, in which upgrading the computational capabilities available will return a better solution to their problem. Similar problems raised in other fields like oceanography or biology, has been solved via creating solutions tailored to a particular problem [44–46]. Because of this, the solutions are only valid for the correspondent field of study. Another solution based on science gateways is close to solving that problem [47], but it only provides a community-specific set of tools, and does not allow a scientist to deploy his code independently.

Our objective is the optimization of the simulation time, in order to have the results as soon as possible or to have more simulations that allows for a better result. The problem is having to use computational resources that are incompatible between themselves. In our case, deploying a big amount of simulations is a key point in order to properly analyze the effect of the variability source on the device behavior. Therefore, we have developed four tools to efficiently process hundreds or thousands of simulations, and we briefly describe them in the following subsections: the Task Manage as a Service, the General Workload Manager, the Self-Calibrator, and the OpenCL implementation of the simulator engine.

1.3.1 Task Manager as a Service

The cloud computing environment has defined an approach that we can adopt in order to tackle the presented computational problem. The taxonomy of cloud computing services [48] is commonly represented via the Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). In all these cloud computing models, there is an abstraction of certain layers of computation, and an interface is offered to the user so he can deal with them without knowing their internal details. For example, the IaaS abstracts the hardware of several machines via virtual machines, that can be launched and managed by the user.

We present the Task Manager as a Service, which solves the aforementioned computational problem. This computing model has also been implemented in the form of the General Workload Manager, explained in the next subsection.

The idea behind the TMaaS is to isolate the access to the computing resources, and to present the user with the ability to define and manage tasks. We define a computational task as a set of components: the environment, the executable that is to be launched, the possible set of input and output resources. The TMaaS is a layer that allows a user to define and manage the life cycle of tasks using the available computing resources transparently.

Once the TMaaS is up and running, the only interaction of the user with the computational resources is the task. With this unit, it is very easy to monitor the tasks in several ways. It also allows to schedule the tasks following different scheduling mechanisms that will adapt to the time deadlines, the status of the computing resources, or the scientist needs. This computing model does not depend on the field that the scientist is working on, so its applicable to the aforementioned cases, and of course to our nanodevice simulator.

1.3.2 General Workload Manager

To implement and test the TMaaS we have developed the General Workload Manager (GWM). This tool complies with the requirements mentioned before, and allows the user to use heterogeneous computing resources in a transparent way.

The tool was developed following a client-server architecture. A server is installed that monitors some ports for REST petitions. By using REST, the application is easy to extend and to discover from the user point of view. The client that communicates with the server via the REST architecture is controlled by the user. We have implemented two different clients with the same capabilities: one command line client which allows to manage the full system from a UNIX terminal, and one web enabled client that allows the user to control the behavior of the server from a web browser. This web browser application is developed using modern technologies for communicating with the server, and displaying the state, to provide a easy, fast and modern experience to the user. Using a Model View Controller paradigm, with AJAX in order to maintain the state of the application in the client, and REST to communicate with the server, the result is that the management of thousands of tasks is not more difficult for the user than that of an online mail client.

The GWM is expansible because it has been conceived as a plugin-based architecture. This allowed up to implement modules for the GWM to communicate with several shells, like bash, sh, or ksh. The same plugin-based architecture is used to facilitate the access to queuing engines, like PBS/Torque or SGE, so the user does not have to deal with the differences between them. Also, the GWM is capable of communicating with several cloud computing providers, like CloudStack, OpenStack, Amazon EC2, and more. So the instantiation of new computing resources is done transparently. One of the developed schedulers, called intelligent scheduler, allows the user to define a stopping metric that can be calculated from the simulation results, and the GWM will deploy only the required simulations to obtain that metric. This is done by calculating the value of the metric after each simulation and using that information as feedback.

Using the GWM we were able to deploy most of the simulations that are presented in this thesis. In most cases, the simulations were run in three different high performance clusters, with incompatible hardware and different task management enqueuing. In any case, the user only had to define the computing task and the GWM would take care of the task management.

1.3.3 Self-calibrator

Another of the solutions developed to tackle the computational problem is a self-calibrator. All the nanodevice simulations presented in this thesis need to be calibrated to some external source. Usually the source is either experimental data, when available, or results from more precise simulations, like NEGF or Monte Carlo. In both cases, the calibration requires the user to guess the right values for the parameters that characterize our drift-diffusion simulator and that fit the behavior of the device as close as possible. To find these parameters, the original procedure is to change their values, simulate the device, compare the behavior and repeat. We developed a self-calibrator that uses the device specifications and the desired behavior to obtain the values for the parameters that closely match that desired behavior. This self-calibrator uses a genetic algorithm to decide the values of the parameters for each iteration, and the GWM to manage the tasks.

1.3.4 OpenCL implementation

The simulator that we are using is implemented in C with MPI to take account of the communication between nodes. This implementation is very well tested and optimized, so no much margin of improvement is possible. New architectures like General Purpose Graphics Processing Units (GPGPUs) or accelerators, like the Intel Xeon Phi, are being used nowadays to obtain faster running times [49], even if they are tailored to dense systems instead of the sparse we are working with. We have implemented the required operations to transfer the engine of our simulator from the MPI-enabled to a OpenCL implementation, which can be run in several different architectures without changing the source code. This is still a work in progress, but the preliminary articles already published in the topic are listed in section 1.5.

1.4 Outline

In the following chapters we provide the key articles that represent the main body of work for this thesis. In all these articles, the author of the thesis has been the main contributor, or a coauthor that highly contributed to the paper. These articles have been either published in JCR journals or in high quality international conferences: IEEE Transactions on Electron Devices, Semiconductor Science and Technology, International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), and IEEE International Conference on Communication (CORE A). This selection of articles has been made to delve into the main points mentioned in the introduction, and to have a more complete representation of the work carried out doing this thesis, the full reference list in section 1.5 should be considered. In that section we list a full compendium of the journal publications and conference presentations related to this thesis, which include journals like IEEE Electron Device Letters and IEEE Internet Computing.

In chapter 2, we explain the Voronoi method introduced in section 1.2.2, to model the Metal Gate Granularity. We also analyze the effect of changing the device body shape from a complete square to a rounded corner shape. The first measures of MGG variability were presented for a 25 nm gate length Silicon SOI FinFET device. This presentation of the Voronoi method was well received by the scientific community and the findings of this article where cited several times. The Voronoi method is being used today by several researchers to model the MGG variability.

Following a recently published approach to calculate the MGG variability via the Rayleigh distribution [39], in chapter 3, we compare our Voronoi model with the Rayleigh approach, using the equivalent Gamma distribution that arises naturally from the grain area distribution of a Voronoi diagram. We also compare both algorithms with TEM images. We found that our approach is way more suitable to match the experimental results, and that the Rayleigh distribution overstates the value of the variability. The analysis was done with experimental data of different materials, provided by Dr. Kenji Ohmori, from the Nanotechnology Laboratory of Waseda University, Tokyo [30].

Using both the Line Edge Roughness, explained in section 1.2.1, and the Metal Gate Granularity, we present in chapter 4 an analysis of the effect of both these variability sources in a 25 nm Silicon SOI FinFET device, the same device that was used in chapter 2. This is the first article in which we present our methodology to generate LER profiles, as an application of the same perturbation pipeline. We have found that the MGG has a negative effect in the power consumption and the switching speed, decreasing the quality of the device, as the grain size grows. Similarly for LER, we have found that both the correlation length and the rms height have a negative effect in the variability of all figures of merit, but more pronounced in the case of the rms height for the studied parameters. In general, this device shows more sensitivity for LER than for MGG.

In order to expand the knowledge of both variability sources and device fabrication, we simulated the same variability sources as in chapter 4, but for two state-of-the-art devices: a Silicon SOI FinFET, and an InGaAs III-V-OI FinFET with a similar shape. In both cases, we have also reduced the size of the device from 25 nm to 10.7 and 10.4 nm, respectively. We used data from Monte Carlo simulations to calibrate the simulator, because there was no experimental data available at the moment. The results of this comparison are shown in chapter 5, where we found that in the sub-threshold region, the InGaAs device is more resilient to MGG variability than the Silicon device, specially for the subthreshold swing, and produces similar results for the LER variability. Nevertheless, the results for on-current present the opposite trend.

To obtain the previous results, we have to run several thousands of simulations, to account for the different devices, variability sources and parameters. The proposed Task Manager as a Service infrastructure was used to test its validity in real world situations. In chapter 6 we present the General Workload Manager, our implementation of the TMaaS computing model. We have applied the GWM to different scenarios to show how it can handle workloads independently of the nature of them, and we also present how it can deal with three incompatible clusters and a cloud provider in order to deploy and manage the computational tasks.

Finally, in chapter 7, we present the conclusions of the thesis and of the articles reproduced in the following chapters, along with the future work that naturally arises from the articles written in this thesis.

1.5 List of publications

This is the list of publications written by the author throughout the development of the thesis.

Articles in peer reviewed journals:

- G. Indalecio, A.J. Garcia-Loureiro, N. Seoane, and K. Kalna, *Study of Metal-Gate Work-Function Variation Using Voronoi Cells: Comparison of Rayleigh and Gamma Distributions*, IEEE Transactions on Electron Devices, **63**, pp. 2625-2628, 2016

- G. Indalecio, F. Gomez-Folgar, and A.J. Garcia-Loureiro, *GWMEP: Task-Manager-as-a-Service in Apache CloudStack*, IEEE Internet Computing, **20**, pp. 42-49, 2016

- G. Indalecio, N. Seoane, M. Aldegunde, K. Kalna, and A. J. Garcia-Loureiro, *Variability Characterisation of Nanoscale Si and InGaAs Fin Field-Effect-Transistors at Subthreshold.*, Journal of Low Power Electronics, **11**, pp. 256-263, 2015

- G. Indalecio, M. Aldegunde, N. Seoane, K.Kalna and A. J. Garcia-Loureiro, *Statistical study of the influence of LER and MGG in SOI MOSFET*, Semiconductor Science and Technology, **29**, 045005, 2014

- N. Seoane, M. Aldegunde, D. Nagy, M.A. Elmessary, G. Indalecio, A.J. Garcia-Loureiro and K. Kalna *Simulation study of scaled $In_{0.53}Ga_{0.47}As$ and Si FinFETs for sub-16 nm technology nodes*, Semiconductor Science and Technology, **31**, 075005, 2016

- N. Seoane, G. Indalecio, M. Aldegunde, D. Nagy, M.A. Elmessary, A.J. Garcia-Loureiro, K. Kalna, *Comparison of Fin-Edge Roughness and Metal Grain Work Function Variability in InGaAs and Si FinFETs*, IEEE Transactions on Electron Devices, **63**, pp. 1209-1215, 2016

- E. Coronado-Barrientos, G. Indalecio and A. Garcia-Loureiro, *Study of basic vector operations on Intel Xeon Phi and NVIDIA Tesla using OpenCL*, Annals of Multicore and GPU Programming, **2**, 15, 2015

- N. Seoane, G. Indalecio, E. Comesana, M. Aldegunde, A. J. Garcia-Loureiro and K. Kalna, *Random Dopant, Line-Edge Roughness, and Gate Workfunction Variability in a Nano InGaAs FinFET*, IEEE Transactions on Electron Devices, **61**, pp. 466-472, 2014

- N. Seoane, G. Indalecio, E. Comesaña, A. J. Garcia-Loureiro, M. Aldegunde, and K. Kalna, *Three-Dimensional Simulations of Random Dopant and Metal-Gate Workfunction Variability in an $In_{0.53}Ga_{0.47}As$ GAA MOSFET*, IEEE Electron Device Letters, **34**, pp. 205-207, 2013

Articles published in international conferences:

- G. Indalecio, F. Gomez-Folgar and A.J. Garcia-Loureiro, *General Workload Manager: a Task Manager as a Service*, IEEE International Conference on Communications, pp. 1859-1864, 2015

- G. Indalecio, N. Seoane, M. Aldegunde, K. Kalna and A. J. Garcia-Loureiro, *Variability characterisation of nanoscale Si and InGaAs FinFETs at subthreshold*, 5th European Workshop on CMOS Variability, 2014

- G. Indalecio, N. Seoane, M. Aldegunde, K. Kalna, A. J. Garcia-Loureiro, *Scaling of Metal Gate Workfunction Variability in nanometer SOI-FinFETs*, 15th International Conference on Ultimate Integration on Silicon, pp. 105-108, 2014

- G. Indalecio, M. Aldegunde, A.J. Garcia-Loureiro, *Static Multipole Method Applied to Boundary Conditions for Semiconductor Device Simulations* The 2012 International Conference on High Performance Computing & Simulation, pp. 654-659, 2012

- G. Indalecio, A.J. Garcia-Loureiro, M. Aldegunde, and K. Kalna, *3D Simulation Study of Work-Function Variability in a 25 nm Metal-Gate FinFET with Curved Geometry using Voronoi Grains*, 2012 International Conference on Simulation of Semiconductor Processes and Devices, pp. 149-152, 2012

- M.A. Elmessary, D. Nagy, M. Aldegunde, N. Seoane, G. Indalecio, J. Lindberg, W. Dettmer, D. Peri, A.J. Garcia-Loureiro and K. Kalna, *Scaling/LER Study of Si GAA Nanowire FET using 3D Finite Element Monte Carlo Simulations*, International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon, pp. 52-55, 2016

- F. Gomez-Folgar, G. Indalecio, N. Seoane, A. J. Garcia-Loureiro, and T. F. Pena, *Study of Point-to-Point Communication Latency for MPI Implementations in Cloud*, The 22nd International Conference on Parallel and Distributed Processing Techniques and Applications, ACCEPTED, 2016

- F. Gomez-Folgar, G. Indalecio, A.J. Garcia-Loureiro and T.F. Pena, *A Flexible Cluster System for the Management of Virtual Clusters in the Cloud*, IEEE 17th International Conference on High Performance Computing and Communications, pp. 1693-1698, 2015

- M. Fortes, E. Comesaña, G. Indalecio, J. Rodriguez, P. Otero, A. Garcia-Loureiro, M. Vetter, *Design and Monte Carlo Simulation of a LED-based Optic Coupler*, 17th International Conference on Computer Modelling and Simulation, pp. 577-581, 2015

- A. Abdikarimov, G. Indalecio, E. Comesaña, N. Seoane, K. Kalna, A.J. Garcia-Loureiro, A.E. Atamuratov, *Influence of device geometry on electrical characteristics of a 10.7 nm SOI-FinFET*, 17th International Workshop on Computational Electronics, pp. 247-248, 2014

- N. Seoane, G. Indalecio and A.J. García-Loureiro, K. Kalna, *Impact of cross-section of 10.4 nm gate length $In_{0.53}Ga_{0.47}As$ FinFETs on metal grain variability*, 2016 International Conference on Simulation of Semiconductor Processes and Devices, ACCEPTED, 2016

- N. Seoane, G. Indalecio, E. Comesaña, M. Aldegunde, A. J. Garcia-Loureiro and K. Kalna, *WN and TiN metal gate workfunction variability in a 10.4 nm gate length InGaAs FinFET*, 17th International Workshop on Computational Electronics, pp. 239-240, 2014

- N. Seoane, A. Garcia–Loureiro, E. Comesaña, R. Valin, G. Indalecio, M. Aldegunde and K. Kalna, *3D simulations of random dopant induced threshold voltage variability in inversion–mode $In_{0.53}Ga_{0.47}As$ GAA MOSFETs*, 2012 International Conference on Simulation of Semiconductor Processes and Devices, pp. 392-395, 2012

Articles published in national conferences:

- G. Indalecio, F. Gomez-Folgar and A. J. Garcia-Loureiro, *Comparison of state-of-the-art distributed computing frameworks with the GWM*, 10th Spanish Conference on Electron Devices, 2015

- G. Indalecio, M. Aldegunde, K. Kalna, A. Garcia-Loureiro, *Study of statistical variability in nanoscale transistors introduced by LER, RDF and MGG*, 2013 Spanish Conference on Electron Devices, pp. 95-98, 2013

- E. Coronado-Barrientos, G. Indalecio and A.J Garcia-Loureiro, *Implementation and performance analysis of the AXPY, DOT, and SpMV functions on Intel Xeon Phi and NVIDIA Tesla using OpenCL*, Segundas Jornadas de Programacion Paralela Multicore y GPU, 2015

- E. Coronado-Barrientos, A. Garcia-Loureiro, G. Indalecio N. Seoane, *Implementation of numerical methods for nanoscaled semiconductor device simulation using OpenCL*, 10th Spanish Conference on Electron Devices, 2015

- F. Gomez-Folgar, G. Indalecio, E. Comesana, A. J. Garcia-Loureiro, T. F. Pena, *A tool to deploy nanodevice simulations on Cloud*, 10th Spanish Conference on Electron Devices, 2015

Chapter 2

3D Simulation Study of Work-Function Variability in a 25 nm Metal-Gate FinFET with Curved Geometry using Voronoi Grains

Following is a reproduction of an article of which the author of this thesis is a main contributor. This is a verbatim reproduction, and the original can be found online with the following information:

International Conference on Simulation of Semiconductor Processes and
Devices, 2012, pp. 149-152

G. Indalecio, A.J. García-Loureiro, M. Aldegunde and K.Kalna

2.1 Abstract

A full-scale 3D simulation study of the impact of metal gate granularity (MGG) on the off-state of a 25 nm length gate SOI FinFET is carried out. The 3D simulations are performed using a parallel finite-element simulator within the drift-diffusion approximation using density gradient quantum corrections. The shapes in the device are described by using splines, and metal grains are modelled using Voronoi diagrams. We study two different grain sizes and silicon fin corner geometries. While the impact of the geometry is found to be negligible in

our simulations due to a relatively large size of the device, the grain size has a large impact on the variability of subthreshold characteristics.

2.2 Introduction

As the ITRS imposes new requirements over the device characteristics for the next generation of digital circuits, FinFET architectures have became the leading solution to continue the scaling beyond the 32 nm node [50]. These devices are manufactured using high-κ/metal gate stack technology, which brings about a reduced EOT without compromising on leakage current. However, grains of different orientations in the metal gate of MOSFETs with sub-32 nm metal gate/high-K gate stacks induce a work-function variability whose impact is comparable with random dopant fluctuations (RDF) and line edge roughness (LER) [12, 51, 52]. The variability induced by the different metal grain orientation will become a dominating source of variability over RDF and LER at the 22 nm technology node [12]. Therefore, the work-function variability is expected to affect also significantly FinFET MOSFETs which use the similar metal gate/high-K gate stacks. Other sources of fluctuations (e.g., RDF and LER) have been studied extensively in the past because they were a main concern for scalability for several technological nodes. However, the metal gate granularity has become a source of concern in the last technology nodes with the substitution of polysilicon by a metal in the transistor gate stack.

In this work, we use an in-house developed parallel code to carry out the study of metal grains induced variability. This code implements a full-scale finite element simulation under the drift-diffusion approximation with quantum corrections [53], needed to account for the quantum confinement effects occurring in FinFET architectures [7].

2.3 Device Description and Simulation Approach

The device under study is a 25 nm length silicon FinFET with a 30 nm tall and 12 nm wide silicon body [54]. The dimensions and shape of the oxide layer are modelled following the data of the actual high-κ based dielectric shown in [54, 55]. The geometry of the device is described by using spline functions instead of the more common block-shapes [56, 56–58], achieving a more realistic simulation. The doping profile was defined analytically to reproduce experimental sub-threshold characteristics at low and high drain biases.

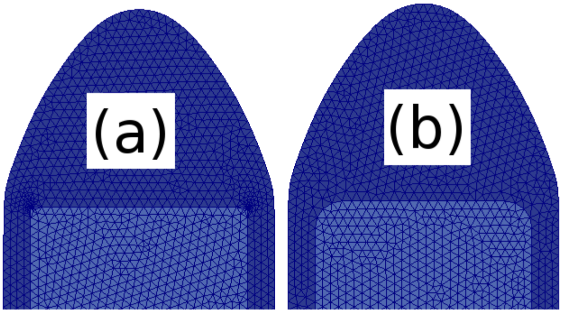

Figure 2.1: Details of ’block’ (a) and ’curved’ (b) geometry of the Si Fin, note the difference at the inner top corners.

The device is modelled using a tetrahedral mesh suitable for finite element simulations. This modelling was made with Gmsh. This is a generic CAD tool, which allows us to define the whole device and tag each surface and volume, with arbitrary geometry. In this way, we can change the geometry of the device and reflect possible differences, like curvatures. An example is shown in Fig. 2.1, where we changed the geometry of the silicon fin from a squared to a rounded one, with the objective of analyse the possible effect of the fin curvature.

The simulation code has been parallelised using the MPI communication library. One of the advantages is lower running time, which allows us to run more samples of the work-function or geometries. Another is to have more memory because the distributed scheme have access to available memory at all the nodes, which allows us to run finer meshes and evaluate fine details, like the curvature on Fig. 2.1 or small patches on the voronoi diagram. This MPI communication requires to divide the mesh in as many parts as nodes of computation. This division is made not using a CAD tool, but using Metis, a well known graph partitioner very suitable for this purpose.

Fig. 2.2 shows the geometry of the device and an example of the distribution of work-function values on the gate. The work-function granularity is modelled using a grain approximation in which we define two possible orientations for the grains in the metal gate with work-functions of 4.4 and 4.6 eV as shown in Table 2.1 together with their respective probability of occurrence [56, 57]. Instead of the usual approximation which employs squares for the shape of the grains [57, 58], we calculated a Voronoi diagram of a set of randomly generated points, which results in variations in size and shape as in realistic crystalline structures such as those used in [56]. Then, each polygon is assigned a certain orientation based on the

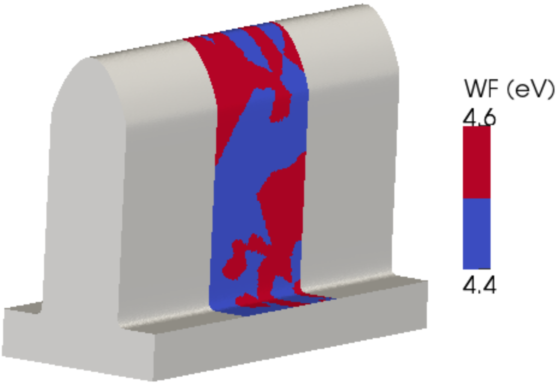

Figure 2.2: Render of a simulated FinFET device, with a gate coloured according to work-function values for 5 nm metal grains.

experimental probabilities shown in Table 2.1. Examples of such diagrams for two different grain sizes are shown in Fig. 2.3.

These stripes with randomly assigned work-functions are then mapped onto the metal gate of the simulation domain using the inverse of the spline-functions which define its shape. The result of this process for the diagrams in Fig. 2.3 can be seen in Fig. 2.4 for the two different grain sizes.

The use of Poisson-Voronoi diagrams to simulate the polycrystalline grain structure of materials, although common in other fields [59–61], has not been used previously for the metal gates in semiconductor device simulations, where a block-based approach is more common (probably because of its simplicity of implementation). The use of Poisson-Voronoi diagrams to simulate polycrystalline materials has some problems associated to its ability to reproduce certain statistical properties of the grains observed experimentally. This has lead to few studies of alternative approaches which try to obtain a closer match to experimentally observed properties [62, 63]. However, the inclusion of these methods can be computationally very de-

Table 2.1: Probability and work-function of the different orientations of the grains for the metal gate used in the simulations.

Orientation	Prob.	WF (eV)
<200>	60 %	4.6
<111>	40 %	4.4

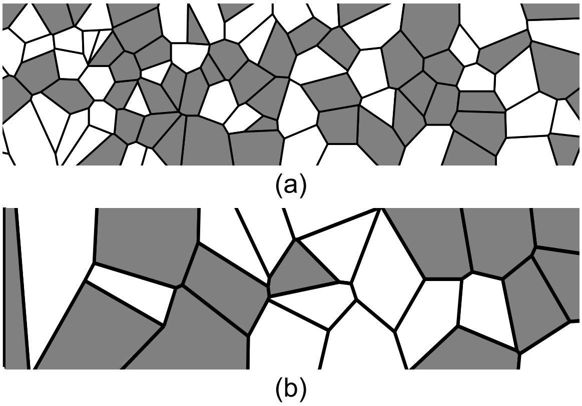

Figure 2.3: Example of two Voronoi diagrams used to randomise metal grains, with a mean diameter of (a) 5 nm and (b) 10 nm. Different colour represents different work-function values, 4.4 eV (white) and 4.6 eV (grey).

manding and as such is beyond the scope of this work. We feel that the approach selected in this work represents good compromise and significantly improves on block based generation methods. Thus more realistic simulations of polycrystalline structures following [62, 63] will bring small quantitative changes to the results obtained in the present work.

2.4 Grain Induced Variability

For the present study of variability, we have generated 200 different work-function patterns with two different average grain sizes using the methodology described in the previous section. Half of patterns have an average grain size of 10 nm, and the other half of 5 nm. We have also changed the shape of the Si body, as shown in Fig. 2.1, from a simple block ('block') to one with corners of 1.5 nm of radii ('curved'), following a similar approximation as in [64].

All the simulation were carried out on FinisTerrae supercomputer from CESGA supercomputing facilities at Galicia, Spain. The FinisTerrae supercomputer is an integrated system with shared memory nodes with a NUMA SMP architecture. It is composed of 143 computing nodes (142 HP Integrity rx7640 nodes with 16 Itanium Montvale cores and 128 GB of memory each, 1 HP Integrity Superdome node with 128 Itanium Montvale cores and 1,024 GB of memory). This is a total of 2528 processing cores and 19.670 TB of memory, interconnected with an INFINIBAND 4xDDR at 20 Gbps.

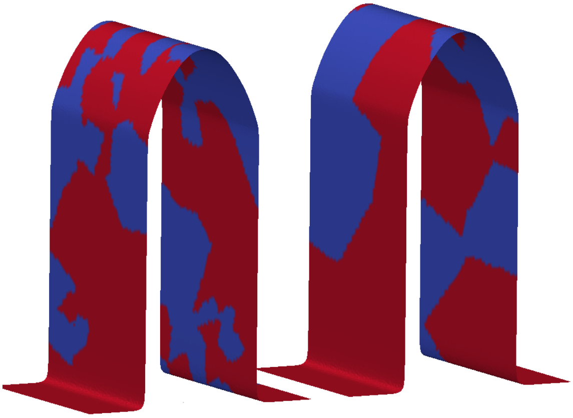

Figure 2.4: Example of work-functions for 5 nm (a) and 10 nm (b) grain sizes. These are two of the 200 profiles used in our simulations showing different colour for different work-function values of 4.4 eV (blue) and 4.6 eV (red).

The simulation results for the four situations, 5 nm and 10 nm average grain sizes and two geometries, are summarised in Fig. 2.5. This figure shows the distribution of threshold voltage, sub-threshold swing and off-current for all the simulated cases. The figure also indicates the mean for each case (thick line) as a reference.

Comparing the simulations using 'block' and 'curved' geometry, only small differences can be found, which are mainly statistical fluctuations. This is not the case for the average

Table 2.2: Threshold voltage (V_t), sub-threshold swing (SS), and off-current (I_{off}) for indicated metal grain sizes and their respective standard deviations.

'Block' geometry						
Grain [nm]	V_t [V]	$\sigma(V_t)$	SS [mv/dec]	$\sigma(SS)$	I_{off} [A]	$\sigma(I_{off})$
5 nm	0.356	0.020	72.752	1.211	-10.734	0.317
10 nm	0.349	0.028	73.220	1.259	-10.610	0.417

'Curved' geometry						
Grain [nm]	V_t [V]	$\sigma(V_t)$	SS [mv/dec]	$\sigma(SS)$	I_{off} [A]	$\sigma(I_{off})$
5 nm	0.355	0.024	72.981	1.199	-10.712	0.366
10 nm	0.356	0.026	73.216	1.362	-10.703	0.390

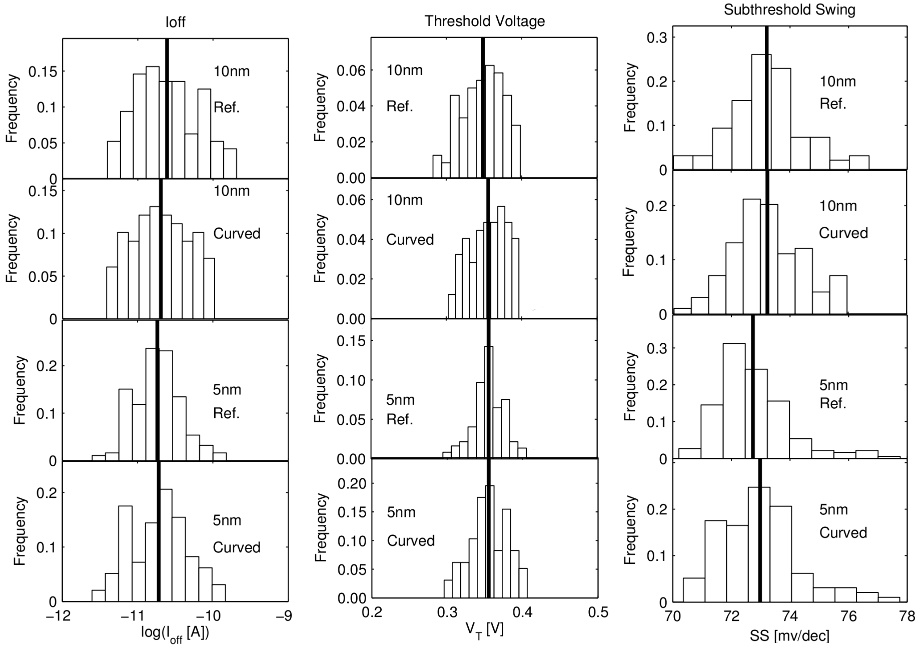

Figure 2.5: Histograms for the simulated I_{off} (off-current), V_T (threshold voltage) and *SS* (subthreshold swing), showing the four possible parameter combinations of the curvature ('block' and 'curved') and the grain diameter (10 nm and 5 nm), vertically stacked to have a common *x*-axis. The mean of the data is shown by a thick black line in every plot.

diameter of the grains. For every set on 2.5, the mean and variability does not depend on the gran size, but the shape of the histogram shows some change. Looking at the *y*-axis range on the threshold voltage, the data are more centered on the 5 nm grain case, and more spread on the 10 nm grain case.

2.5 Conclusions

In this work, we have studied the metal grain induced variability in the sub-threshold characteristics of a 25 nm gate length Si SOI FinFET using quantum corrected 3D FE DD simulations. We have introduced a new approach for modelling the metal grain granularity using Voronoi diagrams instead of the simple square-based approach. This method allows for a more physical representation of the grain shape and size distribution at very little extra computational cost. Furthermore, as the size of the gate decreases, this method will demonstrate its advantages over simpler physical representation of grain shapes. We also conclude that he study on grain size for the work-function variability produced a similar behaviour to that found on bulk MOSFETs.

When the grain size is comparable to the gate size, the distribution of V_{th} becomes wider. The mean values for the three parameters (I_{off}, V_{th} and SS) are similar, but the dispersion of V_{th} is 40 % larger for the 10 nm grain size than for the 5 nm one. Also, the dispersion of I_{off} is about 30 % larger for the bigger size of grains of 10 nm. This increase in the dispersion is not desired for device characteristics because not only the device figures-of-merit have to be matched but they have to exhibot also small dispersions.

On the other hand, the impact of corner effects on the metal grain induced variability are found to be negligible for the simulated device. The mean values for the parameters are similar in the both 'block' and 'curved' geometries deep inside the uncertainty margin. The dispersion σ shows that the larger grains make the distribution wider thus σ increases. These differences would call for a more extensive study taking a larger variety in work-function granularities into consideration.

CHAPTER 3

STUDY OF METAL-GATE WORK-FUNCTION VARIATION USING VORONOI CELLS: COMPARISON OF RAYLEIGH AND GAMMA DISTRIBUTIONS

Following is a reproduction of an article of which the author of this thesis is a main contributor. This is a verbatim reproduction, and the original can be found online under the following DOI: 10.1109/TED.2016.2556749, or with this information:

IEEE Transactions on Electron Devices, vol. 63, no. 6, pp. 2625-2628, 2016

G. Indalecio, A. J. Garcia-Loureiro, N. Seoane and K. Kalna

3.1 Abstract

We have demonstrated, via validation to experimental data for TiN and Ru, that the grains which appear in the metal gate stacks of nanoscale CMOS devices can be characterized via a two-parameter Gamma distribution (p-values 0.17 and 0.42 for TiN and Ru). Conversely, a previously presented fit which used Rayleigh distribution does not reproduce the experimental data (p-values 3×10^{-14} and 0.0029 for TiN and Ru). Poisson Voronoi Diagrams (PVDs) are shown as a suitable algorithm to generate grains with Gamma distribution, via fitting of the distribution of 10000 grains. We have also compared the PVD variability against the

Rayleigh model. for both TiN and TaN metal gates, and concluded that Rayleigh approach overestimates the device variability (by 11.9% for the TiN and by 7.14% for the TaN).

3.2 Introduction

The metal gate granularity (MGG) [65] is one of the most important sources of variability affecting nano-scaled devices studied both experimentally [30] and in simulations [32, 41]. The metal grains that appear in the gate contact will have different sizes and orientations depending on the material and the annealing temperature [66, 67]. The orientation of the metal lattice in each grain will change a work-function (WF) of the metal contact affecting the channel formation and inducing variability into device characteristics [68]. A physically based modelling of the MGG variability requires a realistic characterization of the distribution of the metal grains that accurately reproduces the behaviour found in the experimental devices. Most of the current simulation approaches use square grains which are unrealistic and lack the flexibility to correctly represent gates with very large grains or on the nanoscale regime [31]. Recently, a grain size distribution governed by the Rayleigh distribution was proposed [35] which represented closely simulation results [32]. However, no physical basis has been argued for choosing that particular distribution.

In this paper, we initially establish, via comparison to experimental data, that the random grains arising from the metal gate contacts are characterised via a Gamma distribution. This distribution, unlike the previously adopted Rayleigh fit [35], has a physical justification and will provide a correct description of the metal grain induced device variability. We have demonstrated that our algorithm generates Voronoi cells that follow the expected theoretical Gamma distribution. Finally, we have compared these two distributions, Gamma and Rayleigh, and their predicted MGG variability.

3.3 Poisson Voronoi Diagrams

The realistic growth of the metal grains over amorphous substrates is determined by the nature of a deposition process [69]. The first metal atoms that reach the oxide will deposit at random positions and serve as nucleation points. Next deposited atoms will drift towards their closest nucleation points creating a domain with a specific lattice orientation. A Poisson Voronoi Diagram (PVD) [37] reproduces this behaviour being able to generate realistic grains that account for the shape of domains growing from these randomly placed nucleation points. The

PVD approach has been previously used [21,36,70] to simulate the impact of the metal grain WF variability in nanoscale FinFETs. The physical meaning of the PVD makes it a suitable tool to model the grains of the metal contact. We will demonstrate that the area distribution of the grains generated in a PVD profile is a Gamma distribution. Analysing the area distribution of experimental gates, we are able to validate the Gamma distribution. This process provides the physical basis to use the Gamma distribution that the Rayleigh one lacks.

The PVD is a mathematical structure that consists in seeds of points randomly placed in any n-dimensional space. In our case, we are working with a surface that represents a metal gate contact, to generate a profile that can be applied to the device. Once the seeds are located, all the points from the space are classified taking into consideration the nearest seed, defining a PVD. In our case, the domains represent the grains of the metal contact. This profile is generated from material parameters and device dimensions to particularize it for the device under study.

3.4 Simulation results

3.4.1 Experimental validation

Both Gamma and Raileigh distributions try to account for the grain distribution of metal gates, so in order to prove which one is more suitable, the best approach is to compare experimental data [30] against Rayleigh and Gamma distributions. To do this comparison, we are going to fit the areas from experimental TEM images to the following density functions:

$$Rayleigh(x;a) = \frac{x}{a^2}\exp\left(-x^2/2a^2\right) \tag{3.1}$$

$$Gamma(x;a,b) = \frac{1}{b^a\Gamma(a)}x^{(a-1)}\exp\left(-x/b\right)) \tag{3.2}$$

where x is the normalized grain area and a, and b are fitting parameters.

Two metal poly-crystalline films have been compared: TiN, which produces nano-sized grains with mean diameter of 4.3 nm, and Ru, with larger-sized grains with mean diameter of 18 nm [30]. Figs. 3.1 and 3.2 show the experimental histograms of the distribution of grain areas (normalised by the mean grain area) for the TiN and Ru metals and their comparison to Rayleigh and Gamma distributions. For both metals, the Gamma distribution accurately reproduces the shape of the experimentally observed metal grains. However, Rayleigh distribution underestimates the number of small grains and overestimates the number of large ones.

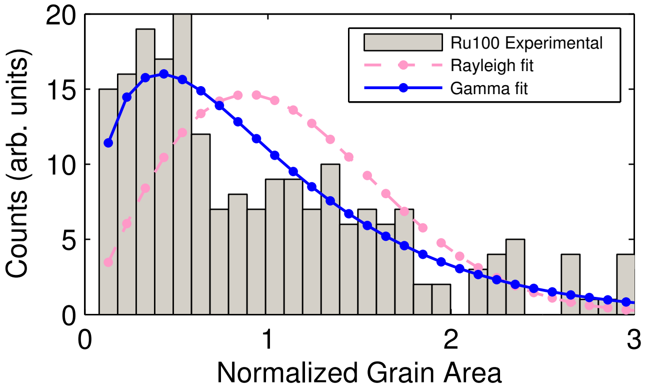

Figure 3.1: Experimental distribution of the normalized Ru grain area fitted to Rayleigh and Gamma functions.

Using a χ^2 test [71], we can quantify how well these two distributions represent the experimental data. The χ^2 test compares the observed histogram measures (O_i) and the expected statistical distribution (E_i) using the normalized difference for the n measured points:

$$\chi^2 = \sum_i^n \left(\frac{E_i - O_i}{O_i} \right)^2 \tag{3.3}$$

Large values of χ^2 represent a mismatch between the observed and the expected data. For any χ^2 exists a corresponding p-value (tabulated in standard distribution tables [71]) that represents the probability that the set of data follows the proposed distribution. If the p-value is over a lower-bound previously set (typically 0.05) it is considered that the distribution matches properly the data; if the p-value is below the lower-bound the distribution will excessively differ from the observed data.

A χ^2 analysis of the data presented in Figs. 3.1 and 3.2 shows that the Gamma distribution

Table 3.1: Physical properties of the TiN and TaN metals.

Material	Orientation	Probability	WF (eV)
TiN	$< 200 >$	60 %	4.6
[69]	$< 111 >$	40 %	4.4
TaN	$< 100 >$	50 %	4.0
	$< 200 >$	30 %	4.15
	$< 220 >$	20 %	4.8

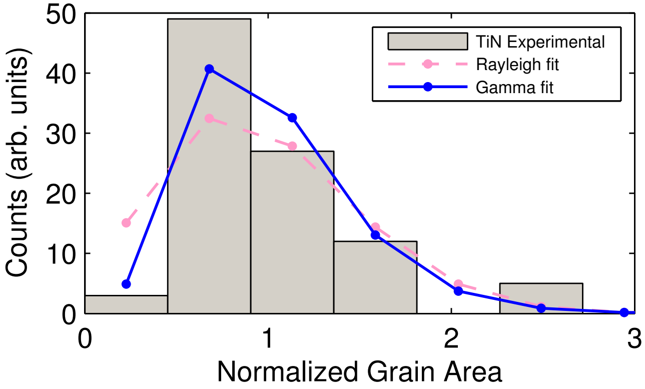

Figure 3.2: Experimental distribution of the normalized TiN grain area fitted to Rayleigh and Gamma functions.

of the grain areas for both metals (p-values 0.17 and 0.42 for TiN and Ru, respectively) fits much better the experimental data than the results produced by Rayleigh distribution (p-values 3×10^{-14} and 0.0029 for TiN and Ru, respectively). Therefore, p-values show that only Gamma distribution is above the lower-bound of 0.05, reproducing the experimental data.

3.4.2 PVD simulations: comparison with Rayleigh and Gamma

In this section, we model the distribution of metal grains via Voronoi based simulations. This model will be independent of the metal employed in the gate, since it only depends on the grain distribution and not on the orientation of the grains. As an example, we show TiN and TaN as possible metals for the gate. Their physical properties are collected in Table 3.1. Fig. 3.3 shows an example of Voronoi WF distributions for these two metal gates for four different grain sizes (10, 7, 5 and 3 nm).

The distribution of grains created via PVD will be now analised to show that Voronoi approach inherently follows the Gamma distribution. We have generated several hundreds of metal grain mappings on the gate with an average grain size of 4.3 nm. Fig. 3.4 shows a histogram of the normalized grain area distribution for metal grains when Voronoi approach is used together with its fit to Rayleigh and Gamma distributions. The Voronoi distribution fits accurately to a Gamma distribution with a p-value of 0.38. On the other hand, Rayleigh distribution is not only ill-fitted (with a p-value of 0.033) but it also does not represent correctly a position of the mode of the distribution, which is the grain area that has the largest

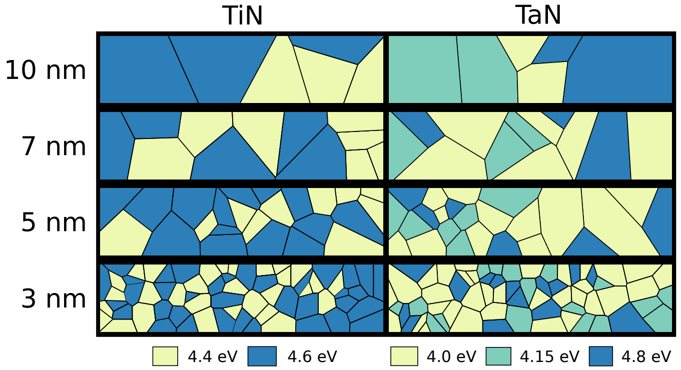

Figure 3.3: TiN and TaN gates for different grain sizes. Each colour represent one of the possible WF values for the given material.

frequency. This can be seen in Fig. 3.1, in which modes for both distributions are shifted. Only the Gamma mode matches the experimental data. The Gamma distribution parameters a and b (see Eq. (2)) were fitted to a=b^{-1}=3.47 via the least squares method. Those values are very close to those predicted by [37] (a=b^{-1}=3.50). The Rayleigh distribution parameter a fitted to 0.92 is giving the best possible fit to experimental data.

3.4.3 Impact on the estimation of the WF variability

Having demonstrated that the experimental grains follow a Gamma distribution, we aim to analyse the impact of using Rayleigh distribution instead of Gamma distribution to generate the grains for MGG variability studies. To estimate the impact of the gate length and the grain size, we initially define the RGG (average grain size divided by the total gate area) as previously done in Ref. [35]. To obtain an average WF value for all the gates generated via the Voronoi approach, we use the following expression:

$$WF(eV) = \sum_{i=0}^{N} \frac{A_i \cdot WF_i}{A}, \tag{3.4}$$

where N is the number of grains present in the gate, A_i (nm^2) the area of the grain i, A (nm^2) the total area of the gate, and WF_i (eV) the WF value assigned to the grain i. This is a simplification done in order to compare our results with the Rayleigh approach, because it uses an average of grain areas instead of simulating the full device. The downside of this

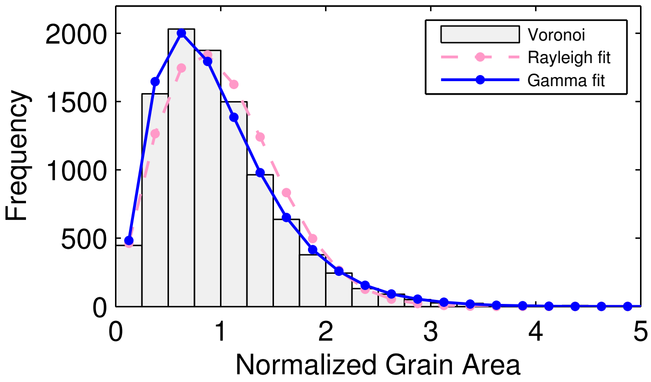

Figure 3.4: Distribution of metal grains obtained via Voronoi based simulations. Results have been fitted to both Rayleigh and Gamma distributions.

simplification is an underestimation of the variability that will affect all scenarios, as noted in [32].

Fig. 3.5 shows the TiN and TaN metal gate WF variability extracted from Voronoi based simulations and compares it with: our proposed Gamma fit, and the Rayleigh fit (data extracted from [35]). The number of metal gates which is used to obtain an accurate grain distribution while minimise statistical error depends on the grain size and ranges between 500 to 1000.

The Rayleigh linear fit consistently overestimates the gate WF variability with respect to the Gamma fit by 11.9% for the TiN and by 7.14% for the TaN. This overstimation is based on the fact that Rayleigh distribution is unable to correctly capture the grain size distribution, as seen in Figures 3.1, 3.2 and 3.4. This inaccuracy does not play a significant role when the number of grains present in a gate is very large but when there are only a few grains in the gate as in typical nano-scale multi-gate FETs [41], it may lead to a significant overestimation of a variability of the threshold voltage (V$_T$) of devices. As a rough estimation, if we take into account that the V$_T$ of a MOS device depends linearly on its gate WF, the correlation between the metal gate WF and the V$_T$ variabilities [72] is:

$$\sigma(WF)/eV = \sigma(V_T)/mV \tag{3.5}$$

As an example, the overestimation in $\sigma(V_T)$ is 5 mV when RGG=0.2 if the Rayleigh linear fit is used and it increases to 12.5 mV when RGG=0.5 for TaN.

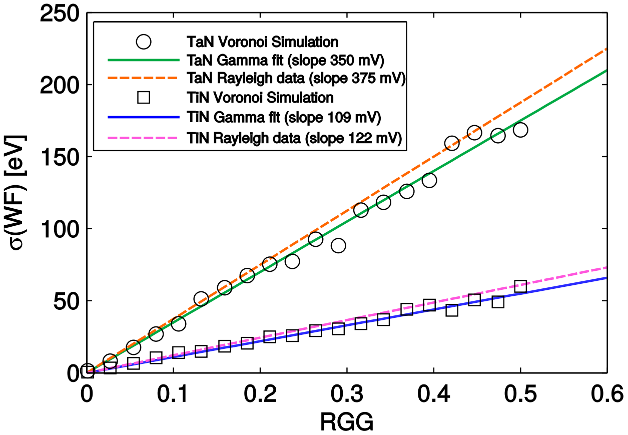

Figure 3.5: TiN and TaN gate WF variability Voronoi simulations compared to [32] and to linear Rayleigh and Gamma fits.

3.5 Conclusion

We have demonstrated, via validation to experimental data [30], that the metal grains which appear in the metal gate stacks of state-of-the-art nano-scaled devices can be characterized via a two-parameter Gamma distribution. We have shown that a previously presented fit which used Rayleigh distribution [35] is not accurately reproducing the experimental data. However, the two-parameter Gamma distribution of the grain areas is well fitted (p-values 0.17 and 0.42 for TiN and Ru, respectively) while the Rayleigh distribution of the grains is unsatisfactory (p-values 3×10^{-14} and 0.0029 for TiN and Ru, respectively).

Finally, we have compared the Poisson Voronoi Diagram (PVD) variability against the Rayleigh model for both TiN and TaN metal gates. The PVD is an optimum method [19, 73] to generate metal grains since this approach represents the shape of domains that grow from randomly placed nucleation points as observed in a real fabrication [69], and the grain distribution generated matches the experimental results. We have shown that the Rayleigh approach overestimates the device variability (by 11.9% for the TiN and by 7.14% for the TaN), whereas the variability provided by the Gamma distribution is much closer to the realistic metal gate induced device variability.

Chapter 4

Statistical study of the influence of LER and MGG in SOI MOSFET

Following is a reproduction of an article of which the author of this thesis is a main contributor. This is a verbatim reproduction, and the original can be found online under the following DOI: 10.1088/0268-1242/29/4/045005, or with this information:

Semiconductor Science and Technology, vol. 29, no. 4, pp. 045005, 2014

G. Indalecio, M. Aldegunde, N. Seoane, K. Kalna and A.J. García-Loureiro

4.1 Abstract

A 3D drift-diffusion device simulation tool with quantum corrections has been applied to study the off-current, threshold voltage and sub-threshold slope variability induced by the metal gate granularity, using a Voronoi approach, and line edge roughness, using Fourier synthesis, in a 25 nm Si FinFET. The discretisation based on the finite element method allows for an accurate description of the 3D geometry. We have simulated 4000 variations of the device to study the metal gate granularity using four different metal grain sizes. The results for the threshold voltage variability ranged from 8.6 mV, for a 3 nm grain size, to 25.9 mV, for a 10 nm grain size. The effect of the grain size was studied and found an inverse square root dependence of the variability for the three figures of merit. The mean threshold voltage and sub-threshold slope have monotonous decrease with the decrease in metal grain size suggesting that the device power consumption and switching speed can be improved by reducing

the grain size. The corresponding threshold voltage variability can reach up to 8.2 mV when RMS=3 nm and the correlation length is 50 nm.

4.2 Introduction

The next generations of digital technology impose new challenges on device solutions which has to satisfy requirements defined in the ITRS [3]. To meet these requirements, fin field-effect-transistor (FinFET) architectures arise as the strongest candidates to continue the scaling at and beyond the 22 nm node [50]. These non-planar devices have to use a high-κ/metal gate stack in order reduce a gate leakage thanks to a thicker physical oxide thickness. The FinFET architecture also do not require high doped source/drain (S/D) regions resulting in a decrease in random discrete dopant variability [74]. However, the metal gate technology brings about a new source of variability, the random orientation of the grains which compose a gate. The actual size of the metal grains depends on the processing technology and their associated variability. As it has been shown in other works [58, 75], this size has a serious impact in the sub-threshold region of the device, causing fluctuations in the threshold voltage and in other device parameters like the sub-threshold swing and the off current.

In this work, we explore the polycrystalline TiN metal gate work function variability in the sub-threshold region of a 25 nm gate length SOI FinFET. We use a finite element framework to properly account for the geometry of the device and we develop a novel Voronoi diagram approach to provide a realistic representation of the metal grain granularity (MGG), following the design of an actual FinFET device [54]. We also explore the line edge roughness (LER) using a method that allows to control the shape of the roughness with two parameters, a correlation length and a root mean square (RMS) deviation [26]. Due to the existence of these variability sources, the ITRS [3] suggest that the accurate modelling of the devices is fundamental for the prediction and optimisation of the device performance. We have used a distributed-memory 3D finite-element (FE) simulations, that use the drift-diffusion (DD) transport model and includes quantum corrections via the density gradient (DG) approach adapted for the FE method [7].

4.3 Model and simulation

The study is carried out with our in-house parallel 3D FE DD simulator [7] which includes quantum corrections using the DG approach. The device under study is a fully depleted 25 nm

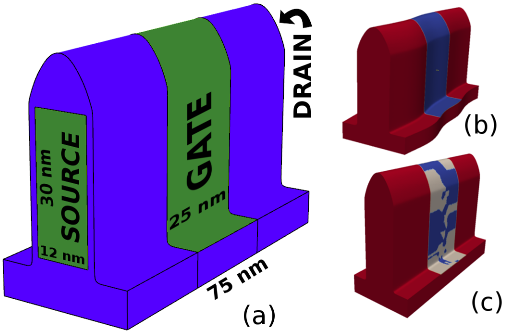

Figure 4.1: a) A schematic of the simulated 25 nm gate length SOI FinFET showing the S/D and gate contacts. Apart from the dimensions shown, the oxide has a width of 1.5 nm, the buried-oxide (BOX) has a width of 30 nm and a height of 7.75 nm. The total height of the device is 51 nm. b) Example of a LER-modified device, with RMS of 2 and correlation of 20 nm. c) Example of a WFV-modified device, showing in two colors the possible orientation of the grains.

gate length SOI FinFET with a silicon body height of 30 nm and a width of 12 nm [54, 55]. Thanks to the use of the FE method, we can describe arbitrary geometries and reproduce a real device design. The silicon body, oxide layer dimensions, and the shape of channel were accurately modelled following a TEM visualisation of the real devices [54, 55]. In order to reproduce as closely as possible the experimental data, a spline was used to model the shapes of the oxide and gate structure, see Figure 4.1.

The doping profile was carefully calibrated to reproduce the experimental sub-threshold characteristics at low/high drain biases while taking into account all experimentally known limitations. S/D n-type regions were doped to 10^{20} cm^{-3} with a Gaussian decay of $\sigma =$ 8.07 nm towards the channel, which is nominally undoped with a p-type doping concentration of 10^{15} cm^{-3}. The result of this process can be seen in Figure 4.2 which shows the experimental measured I_D-V_G characteristics at a low drain bias of 0.05 V and a high drain bias of 1.0 V which are the operating polarizations of this device when used in digital applications.

This calibration has been done by changing the parameters that define the DD simulation and the QC. Macroscopic parameters like saturation velocity and ECN for the high field dependency of the mobility (we are using the Caughey and Thomas expression for high field mobility corrections), and relative masses for the quantum corrections are used to reproduce the behaviour of the experimental data. The saturation velocity was finally set at 1.04×10^7 cm/s, the normal electric field has a value of 6.49×10^4 V/cm. The relative electron mass is

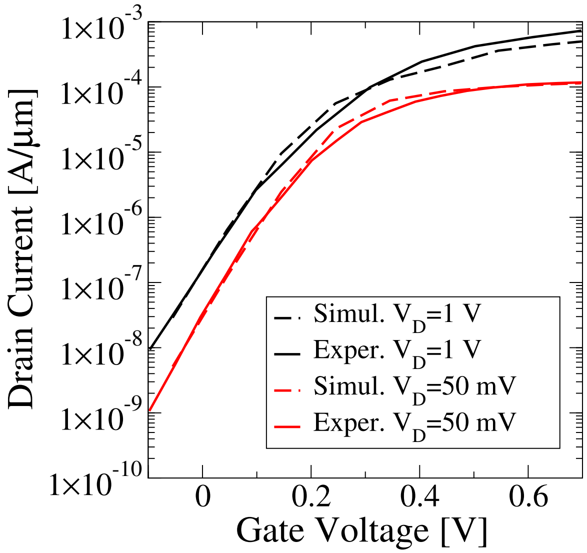

Figure 4.2: I_D-V_G characteristics for the 25 nm SOI FinFET at V_D=0.05 V and V_D=1 V comparing the results from our 3D FE DD simulation with quantum corrections against the experimental results (full lines) [54].

to calculate the quantum corrections is set to 0.32, for the electrons in the silicon, and 0.4 for the oxide.

We have obtained a good agreement in the sub-threshold region which is essential for a study of the device sub-threshold figures of merit. We observe an underestimation of the on-current at $V_D = 1.0\ V$ since the mobility model does not reproduce the drain current in this region. This is consistent with the presence of highly non-equilibrium carrier transport at these large drain as well as gate voltages [9]. Fortunately, the on-current variability is largely suppressed in FinFETs thanks to their electrostatic integrity [76].

4.3.1 Voronoi Grains

For the analysis of the MGG, the gate is modified by changing its work-function to match the values and shapes of different grain orientations observed experimentally [34]. These patterns were generated using a Voronoi diagram over a plain surface for a fixed amount of random points [70, 77] providing a set of patches with a given average surface, which represent the grains. The next step is to assign different work-function values to every patch, following the parameters shown in Table 4.1. This aims to provide a physical representation of the

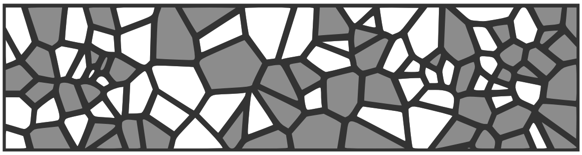

Figure 4.3: Example of Voronoi grains used in the MGG simulation. The edges of the grains are shown with a thick black line, and the patch colour represents different values of the work-function as defined in Table 4.1 (white: 4.4 eV, grey: 4.6 eV)

grains [56] rather than the usual approximation of using square grains without any differential shape [57, 75] which is very artificial. Figure 4.3 represents an example of a 5 nm grain distribution, as used in one of our simulations. These patterns are then mapped onto the metal gate using the mathematical definition of the shape for the gate. In this way, we can adjust any pattern to different device simulations, keeping the advantages of the FE approach. This contrasts with previous FinFET simulations where the shape of the gate was left as a single rectangle [56] instead of modelling the real shape of the device.

4.3.2 Line Edge Roughness

LER is the geometrical difference between a straight theoretical ideal mask and a real device shape produced in a lithographic fabrication process which shows roughness in its defining lines [78], [25]. We can define different planes in a 3D device geometry in which a line roughness is present. The most significant are Fin Edge Roughness and Gate Edge Roughness. Since the Fin Edge Roughness has a paramount impact on the behaviour of the device characteristics as reported in [25], we will be focusing on that phenomenon.

The LER is implemented as previously done in other works [15], deforming the mesh according to the shape of the roughness. To represent the roughness, we generate a set of profiles, which will define how much each part of the device will be deformed. We use a Fourier synthesis method, in which we define and adjust the amplitude of the roughness and also the spatial correlation, using the root mean square (RMS) of the deviations, Δ, and

Table 4.1: Probability and work-functions for a TiN metal gate.

Orientation	Prob.	WF (eV)
<200>	60 %	4.6
<111>	40 %	4.4

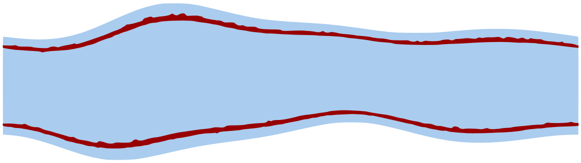

Figure 4.4: Example of a LER profile used in our simulations applied to a transversal cut of the device, parallel to the ground, with a 2 nm RMS and a 20 nm correlation length. The interfaces between the silicon body (inside) and the oxide (outside) are in red colour.

the correlation length Λ [26]. The first represents the average deviation for each line, and depends on the process used to fabricate the device [25]. We employ in our simulator values found in experiments ranging from 1 nm to 3 nm [25]. Gaussian spectra (S_G) were used for the generation of the roughness via the equation:

$$S_G = \sqrt{\pi}\Delta^2\Lambda e^{-k^2\Lambda^2/4}, \tag{4.1}$$

where k represents the frequency values that are defined by the discretisation in real space. The spectrum S_G is multiplied by an array of complex random numbers, and transformed back to real space by using an inverse fast Fourier transform. Figure 4.4 shows one example of a generated profile for $\Lambda = 20$ nm and $\Delta = 2$ nm, which is then used to modify the structure of the device.

4.4 Simulation results

All the simulations were deployed in a computation cluster, using our in-house developed scheduler which allows for an easy deployment, monitoring and retrieval of tasks. Using our scheduler, the computing capabilities of the CESGA SVGD supercomputer [79] and a local cluster, we were able to deploy 8500 simulations, distributed between the two variability sources that we want to analyse.

4.4.1 Metal gate granularity

We have generated 1000 different profiles for each average grain size, namely: 3 nm, 5 nm, 8 nm and 10 nm. In each simulation, all the other parameters and variables are kept constant, and no other variability sources are studied.

For each grain size, we have obtained three figures of merit: off-current (I_{off}), threshold voltage (V_{th}), and sub-threshold swing (SS) at low drain bias, $V_D = 0.05V$. For each figure of merit, we have studied the three main statistical parameters: mean, standard deviation and skewness. The statistical values obtained for the three figures of merits as a function of the grain size are shown in Table 4.2.

The threshold voltage variability can be compared with previous studies done on SOI FinFETs [56, 75] with slightly different gate lengths. For a device with a gate length and a fin width of 16 nm [75], the observed standard deviation for a threshold voltage distribution is 19.2 mV and the mean value is 250 mV, for a grain size of 8 nm. In our case, for the same grain size, the standard deviation is 21.8 mV and the mean threshold voltage is 350 mV. For a SOI FinFET with a gate length of 20 nm and a fin width of 10 nm [56], the standard deviation for the threshold voltage around 26 mV and 15 mV for the studied grain sizes of 10 nm and

Table 4.2: Statistical parameters of the studied variables, grouped by figure of merit and sorted by average grain size.

Sub-threshold Swing (SS [mV/dec])			
Grain	Mean	Stdev	Skewness
10 nm	65.939	0.951	-0.657
8 nm	65.881	0.831	-0.784
5 nm	65.816	0.591	-0.720
3 nm	65.809	0.357	-0.460
Threshold Voltage (V_{th} [V])			
Grain	Mean	Stdev	Skewness
10 nm	0.3547	0.0259	-0.276
8 nm	0.3532	0.0218	-0.354
5 nm	0.3496	0.0137	-0.024
3 nm	0.3466	0.0086	0.102
Off-current ($\log_{10}(I_{\text{off}}$ [A]))			
Grain	Mean	Stdev	Skewness
10 nm	-10.655	0.410	0.166
8 nm	-10.660	0.342	0.236
5 nm	-10.654	0.216	0.052
3 nm	-10.648	0.137	-0.059

5 nm, respectively. In our device, the variability results are almost identical, observing V_{th} standard deviations of 25.9 mV and 13.7 mV for the grain sizes of 10 nm and 5 nm. The threshold voltage variability decreases in a factor 3 when the grain size is reduced from 10 nm to 3 nm. It is important to note that in the simulation study presented in [56, 75], the shape of the gate was just modelled as a simple rectangle, without considering its real shape.

Figure 4.5 shows a panel of histograms, with each figure of merit in a different column, and each grain size in a different row. As expected, the standard deviation of all figures of merit decreases with the reduction of the average grain size due to a better self-averaging of the work-function when there are more grains in the gate. This behaviour is common to all the *SS*, V_{th} and $\log(I_{\mathrm{off}})$, but it is more noticeable in the case of V_{th} as shown earlier. This decrease is very strong, with a standard deviation ranging from 0.951 mV/dec to 0.357 mV/dec for the SS, from 0.0259 V to 0.0086 V for the V_{th} and from 0.410 to 0.137 $log(A)$ for the off-current.

The mean of figures of merit shows a small monotonous decrease for the SS and V_{th} as the grain size decreases. For the $\log(I_{\mathrm{off}})$, it remains almost constant. This is an important result, because it means an improvement of power consumption and speed with the decrease of grain size. This can be seen in Table 4.2, but it is almost imperceptible in the histograms.

For large grain sizes, the SS shows a negative skewness, and I_{off} a positive one. In both cases the skewness approaches zero when the grain size decreases. This can be seen in both the histograms and the numerical table. However, this statistical parameter is very sensitive to extremal data, so using 1000 simulations is on the limit where this parameter can be analysed [16]. The skewness has two consequences: on one side, if the skewness is different from zero, the normality assumption doesn't hold, which limits the analysis that can be done in the data; on other side, a non-zero skewness has an impact on the precission of the mean when only a few samples of the population are used, positive skewness implies that the mean will be overestimated, negative skewness implies that the mean will be understimated. Populations with skewness near to zero will be more predictable and easier to analyse.

Figure 4.6 shows the dependence of the normalised standard deviation with the number of grains in the gate, which is directly related to their average diameter. The normalisation is obtained by dividing the values by the biggest standard deviation for each figure of merit. The behaviour is proportional to the inverse square root of the number of grains as predicted in [74]:

$$\sigma_{Vt} = \frac{\sigma_{WF}}{\sqrt{N_{grain}}} \tag{4.2}$$

In equation (4.2), σ_{WF} is the standard deviation of the metal grain work function and N_{grain}

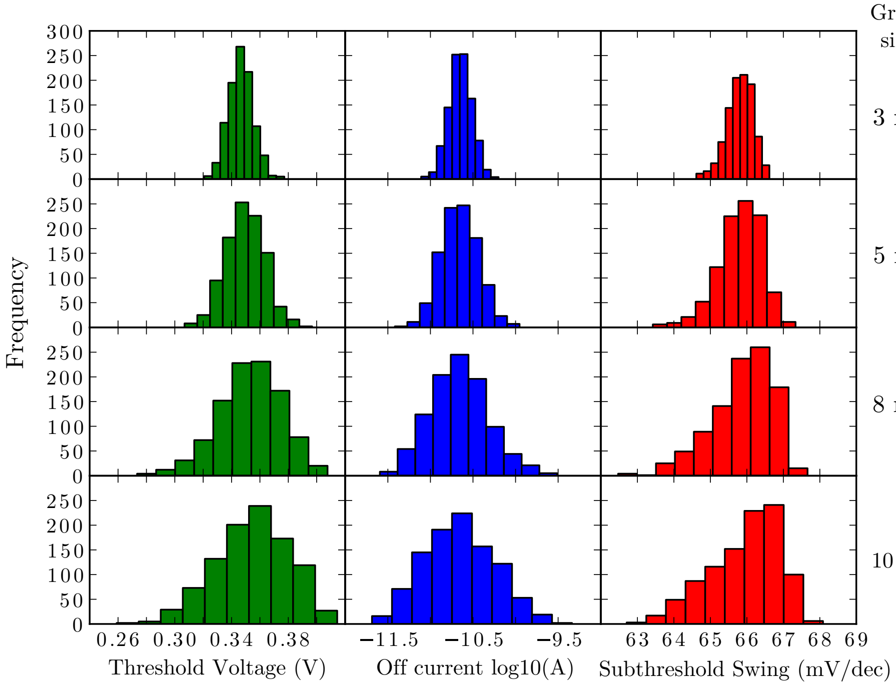

Figure 4.5: Panel showing the frequency histograms for the metal gate granularity (MGG). Each column is one of the three figures of merit studied, namely the threshold voltage (left), the $\log_{10}(I_{\mathrm{off}})$ (middle) and the subthreshold-swing (right).

is the number of grains covering the gate. This behaviour is due to the metal grain size becoming smaller compared to the gate size. The metal gate becomes more uniform and the work function deviation starts to self-average. In the limiting case of infinite number of grains, the work function is constant, without any variability, and we can reach a uniformly distributed work function with $\sigma = 0$.

4.4.2 Line Edge Roughness

To analyse the effect of the line edge roughness, we have generated 4500 random profiles, distributed amongst the values of two parameters: RMS and correlation length. Each parameter is given three different values: 1 nm, 2 nm and 3 nm for the RMS, and 10 nm, 20 nm

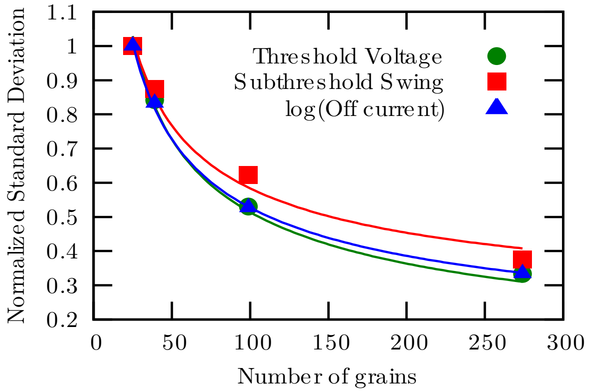

Figure 4.6: Dependence of the normalised standard deviation of the figures of merit with the average of the number of grains. A fit to the inverse of the square root (continuous lines) with offset is shown for comparison.

and 50 nm for the correlation length. The cross product of the values defines nine possible situations.

Figure 4.8 shows the threshold voltage variability for each RMS height and correlation length combination. The threshold voltage variability increases from 1.47 mV in the best case ($\Lambda = 10$ nm, $\Delta = 1$ nm) to 8.01 mV in the worst case ($\Lambda = 50$ nm, $\Delta = 3$ nm), this is more than a five-fold increase. Figure 4.7 shows that the most important contributor to the variability is the RMS value, for every figure of merit. This behaviour is also found in similar devices and simulations as in [25].

An increase in the RMS from 1 nm to 3 nm has a huge impact on the deviation figures which is even more noticeable if the correlation length is large. In the case of the V_{th}, with a correlation length of 10 nm, the variability goes from 1.471 mV to 6.500 mV, and with a correlation of 50 nm, from 2.680 mV to 8.088 mV. In the first case the difference is 5.029 mV and in the second 5.408 mV. The same behaviour is found for the I_{off} and SS. Numerically, a slightly larger results were found in a 22 nm gate length Si inverse mode (IM) FinFET [80], in which the threshold voltage variability is around 4 mV for a RMS of 1 nm and a correlation length of 15 nm.

Figure 4.8 shows the threshold voltage histogram for each combination of RMS and correlation length. The threshold voltage variability becomes smaller when both the RMS and correlation length are reduced, but clearly the effect of RMS plays a more important role than the correlation length (on the scales of their respective realistic values), the same behaviour

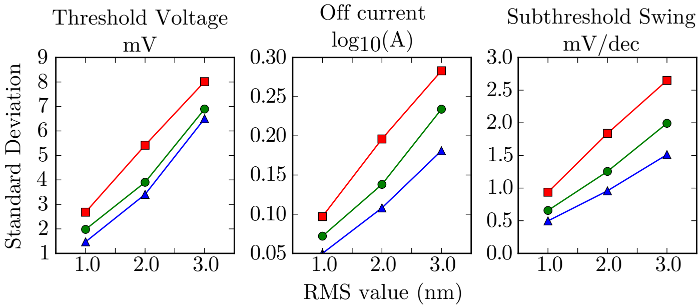

Figure 4.7: Standard deviation for the threshold voltage (left), off-current (centre) and sub-threshold swing (right), showing the change with the RMS value and the correlation length. Results for correlation lengths of 10 nm (blue triangles), 20 nm (green circles) and 50 nm(red squares) are shown. The numerical values are presented in Table 4.2.

is found in other studies [15, 26]. Both the sub-threshold swing and the off-current show a similar trend when changing the RMS or correlation length. In the case of the sub-threshold swing, the variability change is almost sixfold from the worst to the best case (see Figure 4.7).

We are not making an numerical analysis for the skewness values for this variability source, because the number of samples used (500) doesn't allow to calculate a valid skewness value.

4.5 Conclusion

Using our 3-D FE DD simulator, quantum corrected via the density gradient approach, we have studied both the impact of the metal gate granularity and line edge roughness in the off-current, subthreshold slope and threshold voltage in the 25 nm gate length Si SOI FinFET. The effect of the grain size is analysed using a Voronoi diagram approach to accurately describe the real physical shape of the metal grains. We have then run one thousand simulations per grain size to obtain highly significant statistical results. The statistical distribution of the subthreshold figures of merit for the analysed parameters would be incorrectly analysed only if a small set of samples are taken from the population, specially when the grain size of the metal gate is bigger, because of the skewness being different from zero, which implies that the distribution is not Gaussian.

The mean of SS and V_{th} shows a monotonous decrease as the grain size decreases, and it remains almost constant for I_{off}. These results translate to important message: the power

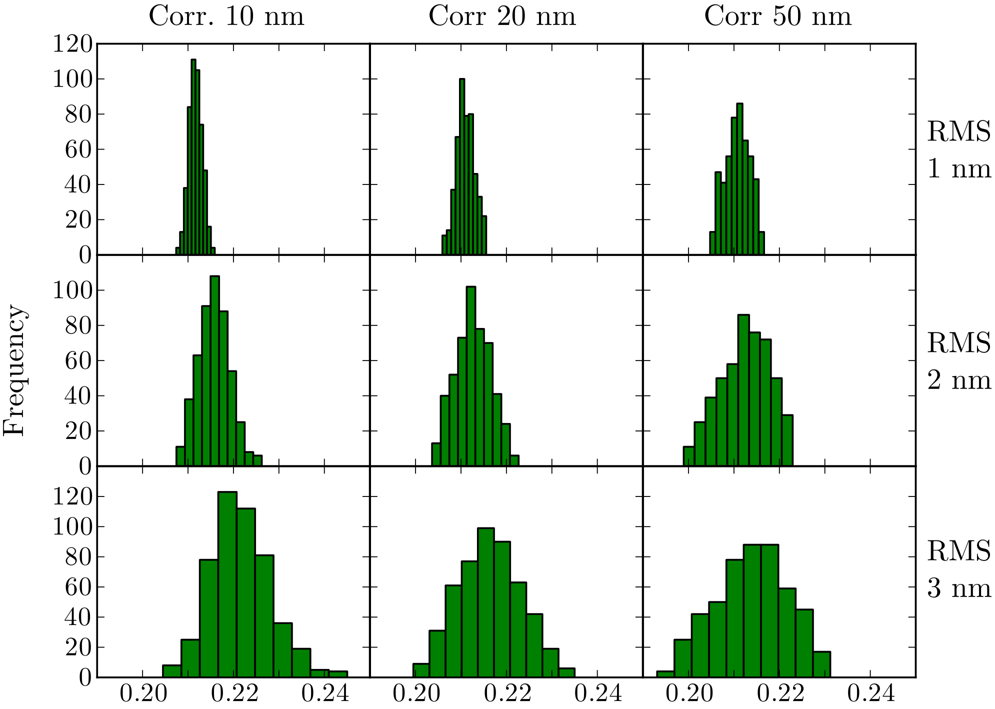

Figure 4.8: Threshold voltage variability for each considered RMS height (marked as RMS in the figure) and correlation length (marked as Corr in the figure).

consumption and the switching speed of FinFETs improve with the decrease in the metal gate grain size. The standard deviation also decreases, from 0.951 mV/dec to 0.357 mV/dec for the SS, from 0.0259 V to 0.0086 V for the V_{th} and from 0.410 to 0.137 $log(A)$ for the off-current, in a monotonous and smooth way, as the mean grain size decreases. In the case of the threshold voltage, the ratio from the worst case to the best is 3.24.

The effect of LER is analysed generating a set of deformation profiles with a given RMS deviation amplitude and correlation length. Changing the RMS from 1 nm to 3 nm has a huge impact on the deviation figures. This impact is even larger if the correlation length is high. In the case of the V_{th}, with a correlation length of 10 nm, the variability goes from 1.471 mV to 6.500 mV, and with a correlation of 50 nm, from 2.680 mV to 8.088 mV. In the first case, the difference is 5.029 mV and in the second 5.408 mV. The same behaviour is found for the I_{off} and SS. This results also translate in a message: the power consumption and speed of this device will degrade if the LER RMS is not keep to a minimum value, being more important that the correlation.

The LER variability, when the RMS is 3 nm, can be similar to the MGG variability for the smallest grain size of 3 nm. When the grain size is increased, the LER is considerably lower than the MGG.

CHAPTER 5

COMPARISON OF FIN EDGE ROUGHNESS AND METAL GRAIN WORK FUNCTION VARIABILITY IN INGAAS AND SI FINFETS

Following is a reproduction of an article of which the author of this thesis is a main contributor. This is a verbatim reproduction, and the original can be found online under the following DOI: 10.1109/TED.2016.2516921, or with this information:

IEEE Transactions on Electron Devices, vol. 63, no. 3, pp. 1209-1216, 2016

N. Seoane, G. Indalecio, M. Aldegunde, D. Nagy, M.A. Elmessary,
A.J. García-Loureiro and K. Kalna

5.1 Abstract

The fin edge roughness (FER) and the TiN metal grain work-function (MGW) induced variability affecting off and on device characteristics is studied and compared between a 10.4 nm gate length $In_{0.53}Ga_{0.47}As$ FinFET and a 10.7 nm gate length Si FinFET. We have analysed the impact of variability by assessing five figures of merit (V_T, SS, I_{OFF}, DIBL and I_{ON}) using two state-of-the-art in-house-build 3-D simulation tools based on the finite element method. Quantum-corrected 3-D drift-diffusion simulations are employed for variability studies in the sub-threshold region while, in the on-region, we use quantum-corrected 3-D ensemble Monte

Carlo simulations. The $In_{0.53}Ga_{0.47}As$ FinFET is more resilient to the FER and MGW variability in the sub-threshold compared to the Si FinFET due to a stronger quantum carrier confinement present in the $In_{0.53}Ga_{0.47}As$ channel. However, the on-current variability is between 1.1-2.2 times larger for the $In_{0.53}Ga_{0.47}As$ FinFET than for the Si counterpart, respectively.

5.2 Introduction

Non-planar multi-gate transistors like FinFETs are leading solutions for the future sub-14 nm digital technology [81]. To meet the ITRS requirements, the future multi-gate transistors may use III-V channel materials which are intensively researched as a possible replacement for *n*-type Si channels because of their higher electron mobility and saturation velocity [82]. These further scaled solutions require not only a realistic assessment of their performance, which is strongly affected by the exact device geometry and design, but also the determination of how different sources of device variability can affect characteristics and reliability.

Variability of transistor characteristics is not only a problem that mainly affects the device fabrication process but it has become an universal concern affecting CMOS and SRAM [83] scaling and perturbing digital logic circuits [84]. New design processes are require to incorporate this phenomena at every level [85]. Nowadays, variability is the main factor restricting the scaling of the supply voltage which, in turn, can lead to unacceptable power dissipation. Random sources of variability such as random dopant fluctuations, line-edge roughness, and metal gate work-function variations, have become dominant in both Si and III-V channel-based nano-MOSFETs [20, 80, 86].

In this work, we have studied and compared the uncorrelated fin-edge roughness (FER) and TiN metal grain work-function (MGW) induced variability in $In_{0.53}Ga_{0.47}As$ and Si FinFETs (designed following the ITRS specifications [87]) using state-of-the-art in-house-build 3-D simulation tools. We simulate the variability for device threshold voltage, off-current, sub-threshold slope, drain-induced-barrier-lowering and on-current at both low and high drain biases.

5.3 FinFET Modelling

The variability study has been performed for a 10.4 nm gate length $In_{0.53}Ga_{0.47}As$ FinFET and a 10.7 nm gate length Si FinFET. These devices have been designed following the 2013 ITRS targets for high-performance logic multi-gate devices [87] assuming a *n*-type Gaussian-

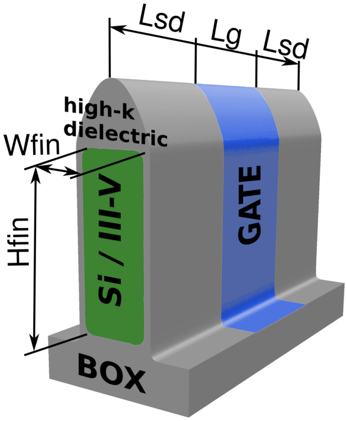

Figure 5.1: Schematic structure of the simulated $In_{0.53}Ga_{0.47}As$ and Si FinFETs.

like doping profile in the source/drain regions (with a N_{SD} peak value) and a *p*-type uniform doping in the channel (N_{ch}) [9]. The geometry of the simulated devices is shown in Fig. 5.1 and their dimensions, doping concentrations and applied drain biases are listed in Table 5.1. The work-function of the TiN metal was set to be 4.52 eV. Table 5.2 shows the nominal performance values yield by both FinFET devices. On the one hand, the $In_{0.53}Ga_{0.47}As$ FinFET delivers a larger on-current than the Si device but for the price of increase in leakage current when compared to than observed in the Si FinFET. Therefore, the (I_{ON}/I_{OFF}) ratio, close to $6x10^4$, is similar for the both devices.

This variability study uses three different simulation tools in a hierarchical workflow from a quantum-transport through a semi-classical to a classical technique. First, we use a 3-D parallel finite-element (FE) drift-diffusion (DD) device simulator [19, 88] with integrated FE density gradient (DG) quantum corrections [7] and Fermi-Dirac statistics [89]. We have calibrated quantum corrections through the effective masses that characterise the DG solution, which mimic the source-to-drain tunnelling and quantum confinement effects [20]. After that, this simulator has been validated at both low and high drain biases against 3-D Non-Equilibrium Green's Functions (NEGF) simulations [11, 90] with an excellent agreement as seen in Figs 5.2 and 5.3.

Finally, for studies in the on-region of $In_{0.53}Ga_{0.47}As$ and Si devices, we use a 3-D quantum-corrected FE ensemble Monte Carlo (MC) simulation tool. In the MC simulator, the quantum corrections have been included via the solution of the DG equation for the

Table 5.1: Dimensions, doping concentrations, and applied drain biases for the simulated $In_{0.53}Ga_{0.47}As$ and Si FinFETs.

Symbol	$In_{0.53}Ga_{0.47}As$	Si
L_G(nm)	10.4	10.7
EOT(nm)	0.59	0.62
W_{fin}(nm)	6.10	5.80
H_{fin}(nm)	15.2	15.0
L_{SD}(nm)	10.4	10.7
N_{ch}(cm^{-3})	10^{17}	10^{15}
N_{SD}(cm^{-3})	5x10^{19}	10^{20}
V_{Dlin}(V)	0.05	0.05
V_{Dsat}(V)	0.60	0.70
WF(eV)	4.52	4.52

Table 5.2: Nominal parameters for the simulated $In_{0.53}Ga_{0.47}As$ and Si FinFETs.

Symbol	$In_{0.53}Ga_{0.47}As$	Si
V_{Tlin}(V)	0.227	0.227
V_{Tsat}(V)	0.183	0.178
SS(mV/dec)	78	76
DIBL(mV/V)	80	75
I_{OFF}(μA/μm)	0.031	0.027
I_{ON}(mA/μm)	1.77	1.56
(I_{ON}/I_{OFF}) ratio	5.7x10^4	5.8x10^4

$In_{0.53}Ga_{0.47}As$ device [9], and of the 2-D Schrödinger equation for the Si device [8], respectively. The MC simulation tool uses an analytic non-parabolic anisotropic model [91] and includes the interface roughness via Ando's model [92]. Note that the 3-D quantum-corrected FE Monte Carlo simulations were verified against experimental I_D-V_G characteristics of a 25 nm gate Si FinFET [9]. The MC considers the following scattering mechanisms: acoustic

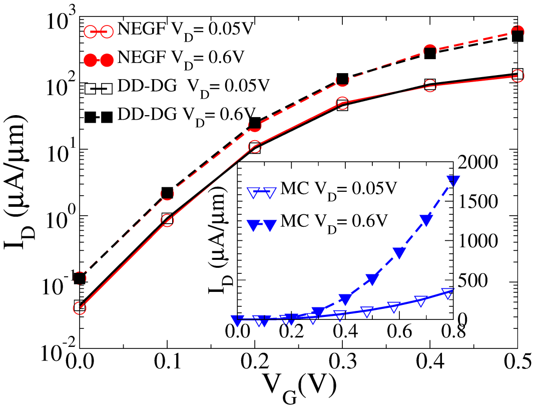

Figure 5.2: I_D-V_G characteristics of the 10.4 nm gate length $In_{0.53}Ga_{0.47}As$ FinFET comparing 3-D DD-DG to ballistic NEGF simulations [90] in the sub-threshold region. Inset: Monte Carlo (MC) simulations of on-current shown on a linear scale.

phonon, non-polar optical intra-valley, non-polar optical inter-valley and ionized impurities (using Ridley's third-body exclusion model [93]). Polar optical phonon, piezoelectric and alloy scattering have also been included for the $In_{0.53}Ga_{0.47}As$ device simulations [94]. The I_D-V_G characteristics of Si and $In_{0.53}Ga_{0.47}As$ FinFETs obtained from the 3-D MC are shown in the insets of Figs 5.2 and 5.3 on a linear scale. The channel orientation is $\langle 100 \rangle$. The Γ valley confinement effective mass of $0.083m_0$, deduced from tight-binding calculations for III-V ultra-thin body SOI MOSFETs [95], has been used in the InGaAs FinFETs while the effective masses in the *L* and *X* valleys are assumed to be bulk.

5.4 Variability Comparison Between Si and $In_{0.53}Ga_{0.47}As$ FinFETs

We have employed the meticulously calibrated DD-DG simulations to analyse the variability affecting the sub-threshold region of the device comparing four figures of merit: threshold voltage (V_T), sub-threshold slope (SS), off-current (I_{OFF}) and drain-induced-barrier-lowering (DIBL). In the on-region, we have studied the on-current (I_{ON}) variability with the quantum-corrected 3-D FE MC simulation tool. Ensembles of 300 and 100 devices have been used in the analysis of the sub-threshold and on-regions, respectively. To extract the threshold voltage, we have used the same constant current criterion for both devices ($I_T = 17.56$ A/m). The on-current has been calculated as the drain current when V_G=V_{Dsat}+V_{Tsat} and the off-current has

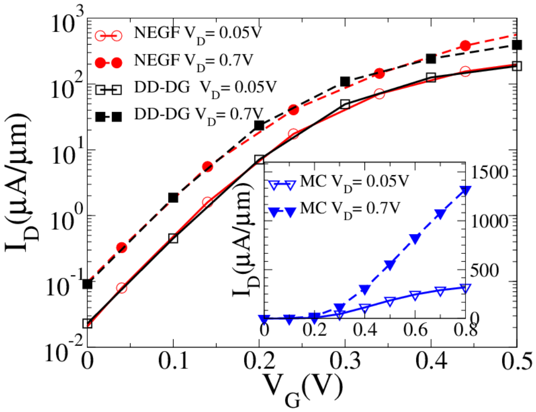

Figure 5.3: I_D-V_G characteristics of the 10.7 nm gate length Si FinFET comparing 3-D DD-DG to NEGF simulations with scattering [11] in the sub-threshold region at low and high drain biases. Inset: 3-D Monte Carlo (MC) simulations of on-current shown on a linear scale.

been extracted at $V_G = 0.0$ V.

5.4.1 FER and MGW Variability Models

The effect of uncorrelated FER is studied using Fourier synthesis with Gaussian autocorrelation [26]. The FER is implemented as previously described in [20, 42]. DD simulations that include DG quantum confinement corrections have been widely used for the analysis of line-edge-roughness variability [51, 56, 86]. The DG requires calibration which is carried out for ideal nominal device. Once accurately calibrated, the DG quantum corrections will mimic very well the position of the lowest bound state [53]. However, the FER will induce a shift in the ground state, particularly for low mass materials such as InGaAs, which would require small adjustments of DG fitting parameters for each simulated sample [8]. These adjustments, computationally prohibitive in variability studies, would introduce small changes in the carrier density distributions [96].

During the simulations, two values of the correlation length (CL=10 and 20 nm) and three root mean square values (RMS=1, 0.8 and 0.6 nm) are analysed. These values have been chosen in order to represent foreseeable trends required by industry and observed in experiments [80]. Fig. 5.4 (left) shows an example of the quantum potential inside the Si FinFET body for a particular FER profile (CL=20 nm, RMS=1 nm) at V_G=0.92 V and V_D=0.7 V.

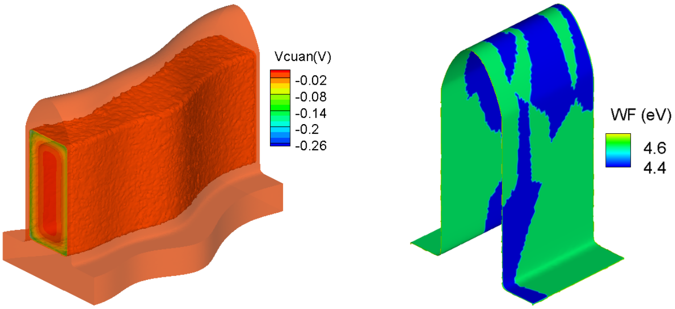

Figure 5.4: (Left) Iso-surfaces showing the quantum potential inside a Si FinFET body for a particular fin-edge roughness (FER) profile (CL=20 nm, RMS=1 nm) at V_G=0.92 V and V_D=0.7 V. (Right) Example of a work function distribution in the TiN metal gate due to MGWV for a 5 nm average GS.

The TiN MGW variability (MGWV) [33] is obtained via the calculation of Voronoi diagrams for a set of randomly generated points which modify the size and shape of the grains. A full description of the followed methodology can be found in [20, 42]. In this work, we analyse four different grain sizes (GSs) (10, 7, 5 and 3 nm) and assume that TiN has two possible grain orientations with MGWs of 4.6 and 4.4 eV and probabilities 60% and 40%, respectively [33]. Fig. 5.4 (right) shows an example of a particular work-function distribution in the TiN metal gate due to MGWV for a 5 nm average GS.

5.4.2 FER Impact on FinFET Variability

Fig. 5.5 shows a comparison of the $\log_{10}(I_{OFF})$, SS, DIBL, V_T and I_{ON} variability due to FER for the 10.4 nm gate length In$_{0.53}$Ga$_{0.47}$As and 10.7 nm Si FinFETs as a function of the drain bias, the correlation length, and the RMS height.

In the presence of FER, the observed variations for the three figures of merit related to the off-region of the device (I_{OFF}, SS and DIBL) are smaller in the In$_{0.53}$Ga$_{0.47}$As FinFET than in the Si FinFET. Note here that the standard deviations for all the figures of merit are strongly affected by the drain bias and the correlation length values in the Si FinFET whereas their impact on the In$_{0.53}$Ga$_{0.47}$As FinFET is smaller as previously seen in [20].

We believe that a smaller variability of the In$_{0.53}$Ga$_{0.47}$As FinFET compared to the Si device can be understood as follows. In the sub-threshold region, where electrostatics dominates, the variability is governed by the strength of the quantum carrier confinement in the nanoscale channel which is related to the separation of energy levels. The In$_{0.53}$Ga$_{0.47}$As,

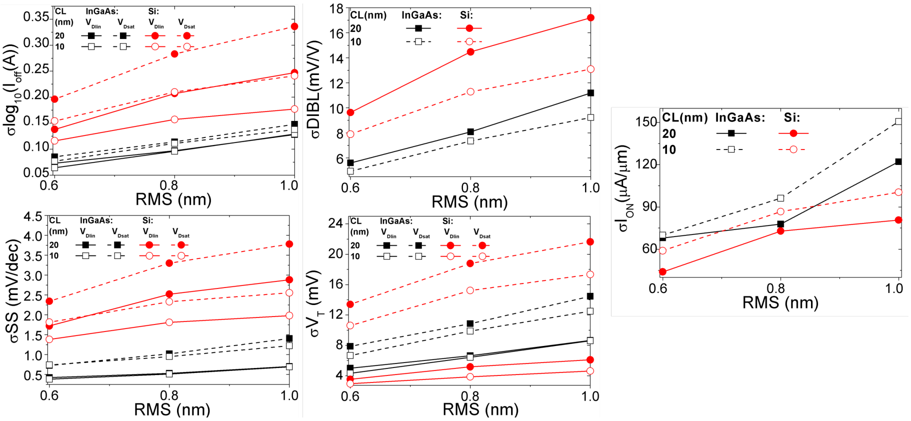

Figure 5.5: Comparison of the $\log_{10}(I_{OFF})$, SS, DIBL, V_T and I_{ON} variability due to FER for the studied $In_{0.53}Ga_{0.47}As$ and Si FinFETs at low (V_{Dlin}=0.05 V) and high drain biases ($V_{Dsat} = 0.6$ V and $V_{Dsat} = 0.7$ V, respectively) as a function of the correlation length, and the RMS height.

being a III-V material, provides a stronger confinement (thanks to a smaller electron effective mass) of electron density in the channel than that in the Si channel [97]. This stronger confinement keeps a large number of carriers in the middle of the channel (a strong body inversion). Thus the carriers in the $In_{0.53}Ga_{0.47}As$ channel will be less affected by disruptions of electrostatics induced by FER leading to a lower variability than the observed in the Si channel. The carriers in the Si channel can spread closer to the FER profile because of the weaker confinement thus interacting with the profile more strongly leading to a larger variability.

On the other hand, we observe that at high drain bias, the FER induced V_T variability is lower for the $In_{0.53}Ga_{0.47}As$ device than for the Si one (same behaviour as in the sub-threshold region magnitudes), while the V_T variability is larger for the $In_{0.53}Ga_{0.47}As$ device than that for the Si one at low drain bias (same behaviour as we will see in the on-region). This opposite behaviour at low and high drain biases is due to the change in the transport regime (V_T is a figure of merit measured at the transition between the off- and on-regions of a device).

In the on-current region, the non-equilibrium carrier transport dominates. We think that the variability will be governed by the efficiency of carrier transport through the channel from the source to the drain and by the gate control of those carriers during the transport process. The $In_{0.53}Ga_{0.47}As$ device has a smaller average effective transport mass when compared to the Si channel providing a faster transport. However, in the on-current regime, the effect of the strong confinement will be less important than in the sub-threshold region. Therefore,

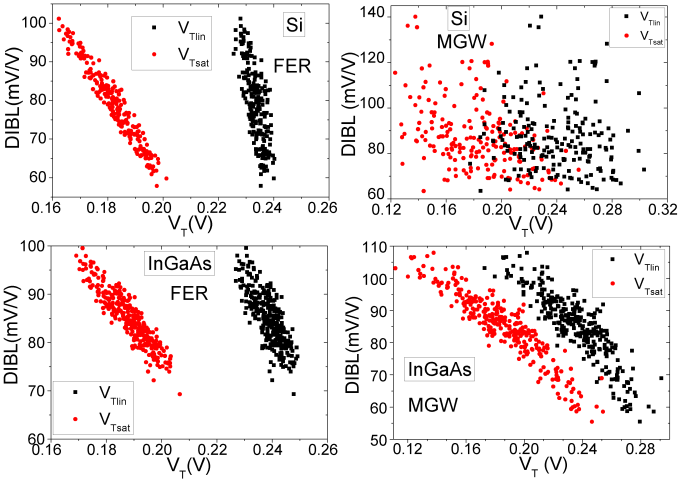

Figure 5.6: Scatter plots showing the DIBL variation as a function of the V_T, at both low (V_{Dlin}=0.05 V) and high (V_{Dsat}=0.6 V and V_{Dsat}=0.7 V, respectively) drain biases, due to MGW (GS=5 nm) and FER variations (CL=20 nm and RMS=0.6 nm) for the 10.4 nm $In_{0.53}Ga_{0.47}As$ and the 10.7 nm Si FinFETs.

these faster III-V carriers with a smaller effective mass interact more strongly with any FER induced electrostatic potential disruptions than the slower carriers with a larger effective mass in Si which leads to a larger variability of the $In_{0.53}Ga_{0.47}As$ transistor.

The I_{ON} variability (Fig. 5.5) is between 1.1-1.5 times larger for the $In_{0.53}Ga_{0.47}As$ FinFET than for the Si device. Note that for both devices the variability is larger for a smaller correlation length of 10 nm which is opposite to the behaviour observed in the sub-threshold region. In the on-region, where the conductivity is large, the device shape variations create effective paths for electrons to pass through the channel more easily. This behaviour will happen less frequently for a smaller correlation length since there will be a higher probability of having an uncorrelated variation (both sides of the device will deform towards opposite directions) followed by a correlated one (both sides will deform towards the same direction). For the Si FinFET, when the correlation length is 10 nm, σI_{ON} ranges from 58 μA/μm when RMS=0.6 nm to 105 μA/μm when RMS=1.0 nm. For the same correlation length, the on-current variability $In_{0.53}Ga_{0.47}As$ FinFET is ranging from 69 to 151 μA/μm when the RMS increases from 0.6 to 1.0 nm.

Fig. 5.6 shows the DIBL variability as a function of V_T at low and high drain biases due to FER (CL=20 nm and RMS=0.6 nm) for the Si (top left figure) and the $In_{0.53}Ga_{0.47}As$ (bottom left figure) FinFETs. For both devices, the DIBL shows strong negative correlations with V_{Tlin} (correlation coefficient (CC) around -0.7) and V_{Tsat} (CC larger than -0.9). In general, the overall narrowing of the channel of the devices due to FER leads to a higher V_T at both low and high drain biases and a better immunity against the short channel effects. We define the threshold voltage shift ($V_{T-shift}$) as the difference between the mean value of statistical sample, $\langle V_T \rangle$ and the threshold voltage for nominal device (see Table 5.2). Thus, the FER-induced $V_{T-shift}$ at both low and high drain biases are around 10 mV for the $In_{0.53}Ga_{0.47}As$ FinFET and increase to around 45 mV for the Si device. The larger DIBL variability observed in the Si FinFET indicates a larger penetration of the electric field into the channel region and therefore a larger loss of gate control at a high drain bias than in the $In_{0.53}Ga_{0.47}As$ device.

5.4.3 MGW Impact on FinFET Variability

Fig. 5.7 shows (from top to bottom) a comparison of the $\log_{10}(I_{OFF})$, SS, DIBL and V_T variability due to MGW for the studied 10.4 nm $In_{0.53}Ga_{0.47}As$ and 10.7 nm Si FinFETs as a function of the drain bias and the average number of grains present in the gate. Note here that we have opted to not represent these magnitudes as a function of the grain size because the results could be misleading since the TiN metal gate area is slightly different for both devices.

The V_T and $\log_{10}(I_{OFF})$ MGWV is very similar for the $In_{0.53}Ga_{0.47}As$ and Si FinFETs when the average number of metal grains in the gate is large (GS small) and slightly smaller for the $In_{0.53}Ga_{0.47}As$ FinFET when the gate is composed by a few grains. The $In_{0.53}Ga_{0.47}As$ FinFET is noticeably less resilient to the SS MGWV than the Si device. A major transport process affecting the drain current in the sub-threshold region is the S/D tunnelling, which influences the SS. The S/D tunnelling is much larger in the $In_{0.53}Ga_{0.47}As$ device than in the Si counterpart, mostly due to a smaller average effective transport mass, leading to the observed larger variability. However, the variability in the DIBL is smaller for the $In_{0.53}Ga_{0.47}As$ device than that for the Si one because, for this figure of merit, the strength of the quantum carrier confinement becomes the major factor while the impact of the S/D tunnelling decreases at the threshold. The carriers are more weakly confined in the Si device leading to a worse electrostatic integrity and thus to a larger variability.

In the sub-threshold region, the MGWV is the dominant source of V_T and $\log_{10}(I_{OFF})$ fluctuations in both Si and III-V FinFETs when compared to the FER. The FER variability (for a RMS=1 nm) is only comparable to MGWV when the number of grains present in the gate is very large (GS 3 nm). The impact of the MGW and the FER (when CL=20 nm) on the

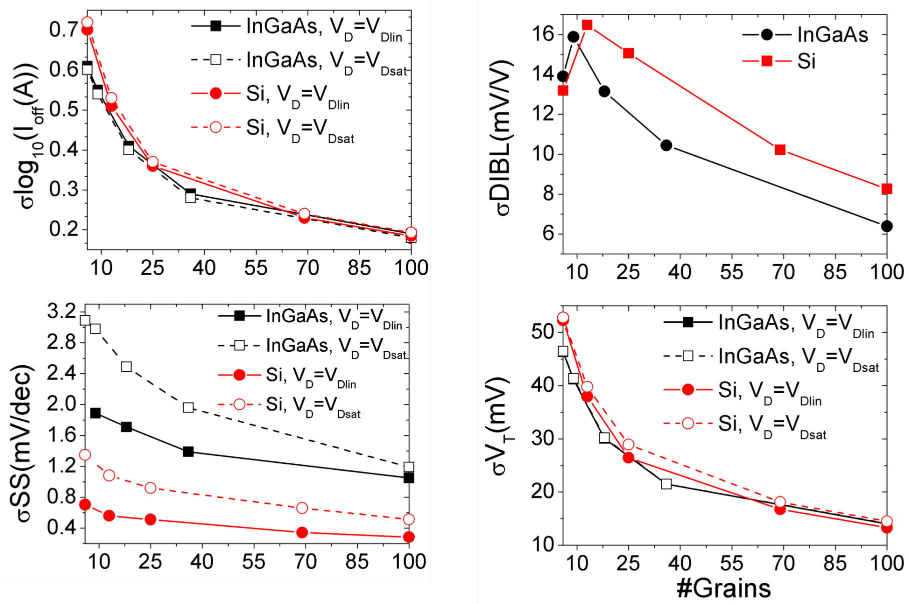

Figure 5.7: Comparison of the log$_{10}$(I$_{OFF}$), SS, DIBL and V$_T$ variability due to MGW for the studied In$_{0.53}$Ga$_{0.47}$As and Si FinFETs as a function of the drain bias (at low V$_{Dlin}$=0.05 V, at high V$_{Dsat}$=0.6 V and V$_{Dsat}$=0.7 V, respectively) and the average number of grains present in the gate.

DIBL variability of both devices is similar. However, the FER becomes the largest source of variability affecting the SS of Si FinFETs while, conversely, the MGW is the dominant source influencing the SS of In$_{0.53}$Ga$_{0.47}$As FinFETs.

Fig.5.6 also shows scatter plots of the DIBL variability as a function of V$_T$ at low and high drain biases due to MGWV (GS=5 nm) for the Si (top right figure) and the In$_{0.53}$Ga$_{0.47}$As (bottom right figure) FinFETs. For the In$_{0.53}$Ga$_{0.47}$As FinFET, the DIBL shows very strong negative correlations with V$_T$ at both low (CC=−0.88) and high drain biases (CC=−0.92). However, for the Si FinFET, the DIBL is practically uncorrelated with V$_T$ at both low (CC is −0.09) and high drain biases (CC=−0.42). The different behaviour observed in the DIBL for both devices can be explained through an analysis of the relation between the threshold voltages at low and high drain biases. Fig. 5.8 shows the scatter plots of V$_{Tlin}$ versus V$_{Tsat}$ for the Si (left figure) and In$_{0.53}$Ga$_{0.47}$As (right figure) FinFETs. The device with an uniform gate has been added for comparative purposes (red line). For the In$_{0.53}$Ga$_{0.47}$As FinFET, the threshold voltage at low and high drain biases are very strongly correlated (CC=0.99). On the other hand, for the Si FinFET, the correlation between V_{Tlin} and V_{Tsat} is weaker (CC=0.94). The larger the CC value, the less sensitive the variability is to a change in the drain bias. As previously seen in Fig. 5.7, the MGW induced V_{T} variability is independent of V_{D} for the $\mathrm{In_{0.53}Ga_{0.47}As}$ FinFET whereas, for the Si device, it slightly increases with the applied drain bias.

Fig. 5.9 shows a comparison of the on-current variability due to MGW both for the $\mathrm{In_{0.53}Ga_{0.47}As}$ and Si FinFETs as a function of the grain size. As expected, the standard deviation of the $\mathrm{I_{ON}}$ decreases with a reduction in the grain size. For the Si FinFET, $\sigma\mathrm{I_{ON}}$ ranges from 59 μA/μm when GS=5 nm to 107 μA/μm when GS=10 nm. For the $\mathrm{In_{0.53}Ga_{0.47}As}$ FinFET, the on-current variability is around 2.2 times larger than that observed for the Si device, with $\sigma\mathrm{I_{ON}}$ ranging from 132 to 237 μA/μm when GS increases from 5 to 10 nm. This very large on-current variability is related to a lower average electron effective transport mass and thus a higher mobility of III-V materials resulting in a faster carrier transport. The MGWV is recognised as a major source of variability in multi-gate transistors with high-K/metal gate stacks [33, 42] due to much stronger disruptions of electrostatic potential in the channel region controlled by the gate (as compared to the FER or random dopants). The disruptions of electrostatic potential in the $\mathrm{In_{0.53}Ga_{0.47}As}$ channel will affect much more its faster non-equilibrium transport thus resulting in a larger difference between the MGWV of the $\mathrm{In_{0.53}Ga_{0.47}As}$ and Si devices than that observed for the FER variability. For the Si FinFET, the impact of the FER and MGW variabilities on the on-current is similar. However, for the $\mathrm{In_{0.53}Ga_{0.47}As}$ FinFET, the MGW and FER induced on-current standard deviations are only similar when GS=5 nm and RMS=1.0 nm. Any other combination of parameters will lead to a larger MGWV than the variability observed due to the FER.

5.5 Conclusion

A 3-D quantum-corrected FE DD and MC simulation study of two sources of statistical variability induced by the fin-edge roughness (FER) and the metal gate work-function (MGW) is performed for the $\mathrm{In_{0.53}Ga_{0.47}As}$ with a gate length of 10.4 nm and Si FinFET with a gate length of 10.7 nm. We have analysed the influence of these two sources of variability on five figures of merit: 1) threshold voltage, 2) sub-threshold slope, 3) off-current, 4) drain-induced-barrier-lowering and 5) on-current. This study is done at both low (0.05 V) and high drain biases (0.6 V for the $\mathrm{In_{0.53}Ga_{0.47}As}$ FinFET and 0.7 V for the Si device). The main

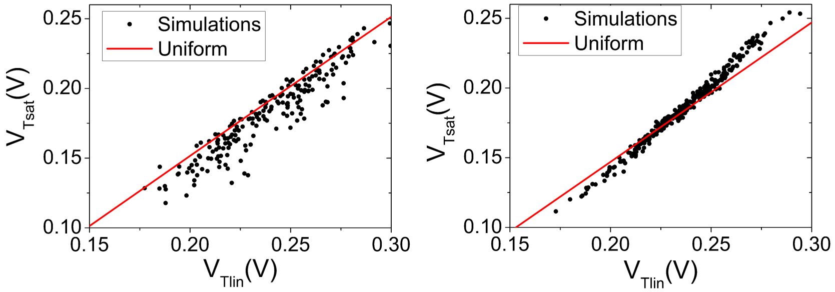

Figure 5.8: Scatter plots showing the threshold voltage at a low drain bias (V_{Tlin}) versus the threshold voltage at a high drain bias (V_{Tsat}) due to MGWV for the Si (left) and $In_{0.53}Ga_{0.47}As$ (right) FinFETs. The device with an uniform gate has been added for comparison (red line).

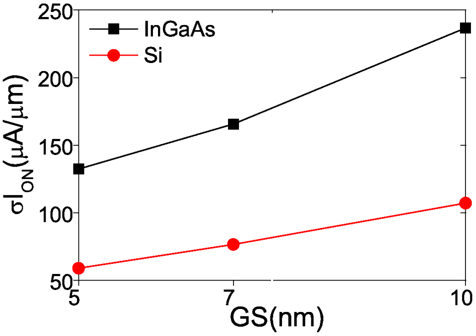

Figure 5.9: Comparison of the I_{ON} variability due to MGW for the $In_{0.53}Ga_{0.47}As$ and Si FinFETs as a function of the grain size (GS).

conclusions can be summarised as follows; in the sub-threshold region:

- The V_T and $\log_{10}(I_{OFF})$ MGWV is very similar for the InGaAs and Si FinFETs when the GS is small, and slightly smaller for the InGaAs FinFET when the GS is large.
- The InGaAs FinFET is less resilient to the SS MGWV than the Si device but there is a smaller variability in the DIBL for the InGaAs device than that for the Si counterpart because of a stronger quantum electron confinement in the III-V channel.
- In the presence of FER, the V_T, $\log_{10}(I_{OFF})$, SS and DIBL variations in the InGaAs FinFET are generally smaller at both low and high drain biases than the ones observed in the Si FinFET.

- The MGW variability is the dominant source of V_T and $\log_{10}(I_{OFF})$ fluctuations in both Si and III-V FinFETs when compared to the FER.

In the on-region:

- The on-current variability due to FER is between 1.1-1.5 times larger for the InGaAs FinFET than for the Si device.
- The on-current variability due to MGW is around 2.2 times larger for the InGaAs FinFET than for the Si device.
- For the Si FinFET, the impact of the FER and MGW variabilities on the on-current is similar.
- For the InGaAs FinFET, the on-current MGW variability is generally larger than that observed due to the FER.

Chapter 6

General Workload Manager: a Task Manager as a Service

Following is a reproduction of an article of which the author of this thesis is a main contributor. This is a verbatim reproduction, and the original can be found online under the following DOI: 10.1109/ICCW.2015.7247451, or with this information:

IEEE International Conference on Communications - Workshop on Cloud Computing Systems, Networks, and Applications, pp. 1859-1864, 2015

G. Indalecio, F. Gómez-Folgar and A.J. García-Loureiro

6.1 Abstract

During the recent past, the demand on High Throughput Computing has been increasing because of the new scientific challenges. Since the access to several computational resources to manage thousands of simulations can be difficult for scientists, different initiatives have tried to provide the scientific community with interfaces that are user-friendly for several computational resources. Usually, these are designed for some specific codes and for a given research field, such as oceanographic, climate modeling and physics, among others. To overcome this situation, we have developed the General Workload Manager (GWM), a universal-purpose very light management system, capable of working with different computing resources with the least configuration as possible, such as HPC and HTC clusters, standalone worker nodes, hypervisor-enabled servers, and cloud platforms. The suggested system is able to deploy

thousands of different simulation tasks using several computing resources, and collecting the results in an easy way.

6.2 Introduction

During the past years, the demands on High Throughput Computing (HTC) has been increasing, due to the fact that the new scientific challenges require more computational power, usually to improve the scope or the precision of their results. To provide the scientific community with the resources they need, several solutions that implement in one way or another a clusterization of capabilities like commodity-clusters, grid, and Cloud, were created. One of the most common solution to HTC is the grid technology, which allows the transparent deployment of tasks in a heterogeneous and decoupled pool of workers. In order to serve grid capabilities to users, several infrastructures have been built such as EGI for Europe [98], SEE-GRID for south eastern Europe [99], and EELA in collaboration between Europe and Latin America [100]. In these infrastructures, gLite is widely used. This middleware has way more capabilities than just managing the workload between the computing elements, which is reserved to a component called the Workload Management System.

Over the grid technology, several solutions have been developed tailored to specific research topics. For example: oceanography [101] in which the computational resources are provided by a grid infrastructure and a virtual database focused on metadata was designed specifically to store the geospatial information, climate modeling [44] in which the addressing challenge is associated with the management, discovery, access and analysis of important datasets distributed in heterogeneous computational environment that is basically a grid problem, chemistry [45] in which the harmonization of the scientists work is done by means of using a web service that can be run on the grid, bioinformatics [46] in which bioinformatics alignment tools are wrapped as web services in grid, physics simulations [102] providing distributed computing resources to the LHCb, ATLAS, CMS or ALICE experiments, the largest at CERN. Also generic job submission tools such as science gateways [47] in which a community-specific set of tools, applications, and data collection are integrated providing access to grid resources.

This is promising, but there are several problems to the day-to-day use of these technologies: mostly the huge overhead, not in performance but in deployment time that the users have to work with. From the authentication to task specification, it also includes the various middleware solutions that convert the use of these tools in a very difficult task if the user is not a computer scientist.

The approach followed in the General Workload Manager (GWM), the tool we have developed is a simplistic one: a general-purpose very light management system, high efficient, capable of working with different computing resources with the least configuration as possible. In our case, only a ssh connection is required. There is no configuration needed in the workers, providing that the worker is a server capable of developing the task that the user wants to distribute. As a rule of thumb, if the user can run a task in a given worker, manually, our system is capable of deploying, monitoring and retrieving the same task. The GWM can manage and deploy tasks in several computation services such as worker nodes, HPC clusters, virtualized nodes, and Cloud platforms, and collecting the obtained results.

6.3 Description of the General Workload Manager

We have designed our system structure to be as simple as possible. It does not have any footprint, this meaning there is no necessity for any configuration in the server in order to be able to launch and monitor tasks. To do this, it is only necessary that the servers are able to run the task that the user wants to deploy. This is also true for the user machine: only a python installation is needed. Next we are going to summarize the main characteristics of the GWM.

- **Only ssh connection needed** to deploy and manage tasks. Only a valid ssh connection is required between the user and the worker, which nowadays is a must-have in any computer. All the communication is carried in that service, taking advantage of the inherent encryption. Any form of identification that enables the user to log into the server (password, private key or private key with encryption password) can be assigned, and our system will login in the same way.

- **Multiple computation services supported** when interacting with the workers. In our system, we define a computation service as the logic object used to manage workers, for example a CloudStack infrastructure, Oracle VM enabled nodes, KVM enabled nodes, instances of Amazon EC2, or a cluster.

- **Multiple engines supported** in the workers. In our system we define an engine as the software used in the worker to manage tasks. For example, the SGE engine in a cluster is treated as an engine in the GWM.

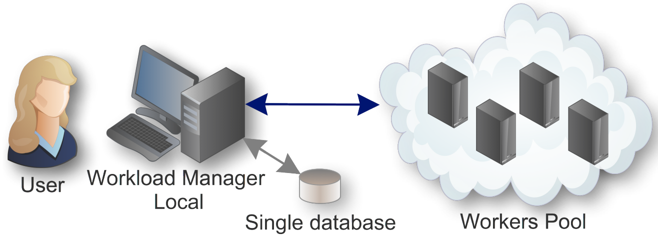

Figure 6.1: Local mode of operation. Note the database is stored in the user computer.

- **The system run in two modes** of operation, which are local and server. The former (see Fig. 6.1) is a distributed mode, in which each user has the full-control of his own database and runs the manager locally, deploying tasks and retrieving them to his local computer. The later (see Fig. 6.2) is a centralized mode, in which the manager runs in a server which allows users to log-in via a web browser, command line client or just use the API to deploy tasks. The server manages the tasks between the computers of every user and the pool of workers, transferring data from one to other.

- **Very easy login**, as in opposition to grid solutions. The authentication is just a pair user/password, as used in every UNIX system, storing only a shadow of the password and using SSL encryption to move every piece of information. This is a very important point, because it allows a new user to deploy a working system easily.

- **Fully featured API**, which allows a developer to extend the functionality of the system from outside. The API is modeled to be as simple and easy to understand as possible, but giving access to all the features of the workload manager.

- **Command line** access, which allows the user to define workers, or deploy, update and retrieve tasks. The command line access is great for advanced users, and for writing automatization scripts.

- **Web interface** allows users to access from a web browser, manage both tasks and workers and change settings. All this is made as user friendly as possible.

From the point of view of the user, two different ways to interact with the system were provided: via the use of a web service or the command line interface.

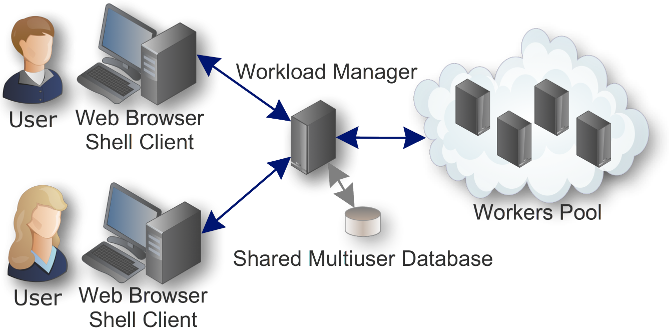

Figure 6.2: Server mode of operation. Note the database is stored in a remote machine.

The web interface is designed as a web application, using AJAX to provide a quick and smooth experiencie when dealing with the system, enabling encrypted connection, and providing a user-friendly modern interface. The web service can be deployed in both a running apache server or in a provided python standalone server that can be run with a non-privileged account in the port 8080. This standalone server allows each user to deploy it's own web service to have easier access to the data and simulations. No configuration is needed by the user, and is intended to be used when the operation mode is local.

On the other hand, a command line interface is also provided, which allows the user to login with the same credentials as in the web browser, and works in a similar way, but using the terminal as an output device instead of a browser. The output is easy to parse, and the commands are short and comprehensive, so it can be used in an interactive mode and it can also be used from any script or in remote mode.

6.4 Details of the architecture of the General Workload Manager

The GWM is designed to adapt to the requirements of the user by being as general as possible with regard to the workload being deployed. The architecture of the system is presented in the Fig. 6.3 which shows the main components of the system. Most of them can be easily extended by a developer so as to implement new services or engines. For normal operation this is not required. In the following subsections we are going to explore an overview of the GWM architecture, and several details about the corresponding components.

6.4.1 Architecture Overview

The *Web Client* provides a user-friendly interface for the API of the system, in both local and server mode. The same interface can be accessed by the *Command Line*. Both applications use a *REST API* created to comply with the REST requirements, providing JSON objects via a state-less connection.

The *Secure Connection Manager* is responsible for establishing connections to workers or user computers. The *Task Engine Manager* is capable of managing the tasks at each step: matching the tasks with workers, deploying them in the matched worker, updating the state of the tasks, and retrieving them. The *Worker Service Manager* is similar to the Task Engine Manager, but applied to workers. This component interacts with the computation services, managing virtual machines and requesting computational resources. The *Configurable Scheduler* uses the three managers to control tasks and workers in a transparent way, this is, independent on the task or worker type. This scheduler operates like a finite state machine, taking the states of the system as input, and taking decisions. To be able to manage large sets of tasks instead of individual ones, they are aggregated in labels, which group tasks and allow the scheduler to take different actions depending on the label and its statistics.

Finally, a *SQL Database* using SQLite stores all the information, and also the state of the system. Using this implementation it is very easy to do backups and deploy the GWM in any user computer, because it doesn't need a running SQL server.

6.4.2 Schedulers

Using the capabilities provided by the other components of the system, a scheduler is a fairly simple component of the system. Only the logic of the state machine has to be implemented, using the information provided by the database. In our case we implemented the following schedulers:

First-Come-First-Served (FCFS), consists on the deployment of the available tasks at the order of definition in the system, using the available workers until there are no tasks to be deployed.

Round-Robin (RR), applied to the labels that group tasks. This distributes the computational power between the labels, instead of deploying the tasks in definition order as in FCFS. This allows the user to define a set of labels, and the system will deploy the tasks from

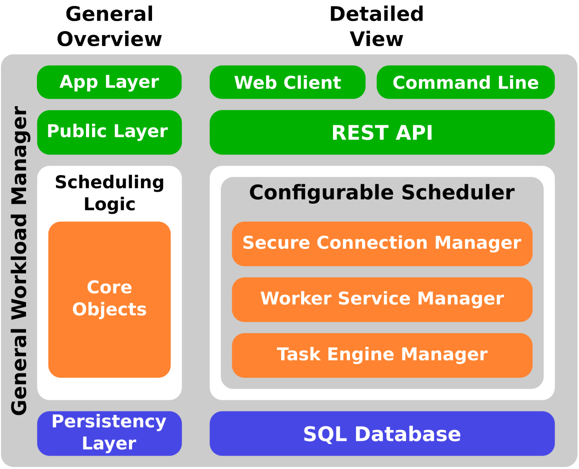

Figure 6.3: Internal architecture of the GWM. We present an overview (left), and a more detailed view (right).

the labels in a balanced round-robin fashion. The effect is that the tasks assigned to the labels will end roughly at the same time.

Fixed-Priority (FP), useful in giving high priority values to important last-time executions. This scheduler behaves more like a service to the user than a resource balancer.

6.4.3 Computation Services

A computation service is an external system, which is able to manage workers. This system will be employed by the GWM, using any kind of connection. An example of computation service is the CloudStack infrastructure, which allows the user to deploy new virtual machine instances, destroy them or pool their state. In any case, the GWM will use its API to access the interface and manage the workers. Amazon EC2 is another example that also has a web API so the implementation is similar.

It is not necessary to have a web API to access a computational service. For example, just a pool of KVM instances in a server can be used as a computational service. In that case, GWM will log into the server and manage the workers by invoking libvirt.

6.4.4 Engines

In the GWM, an engine is defined as any software available for the worker that is capable to schedule the execution of the task. The engine is able to start tasks, pool the state of the task, and stop tasks. There are plenty of engines deployed in clusters and they are not compatible between themselves, so a tool like GWM has to be aware and able to interact with them in order to manage tasks in that worker.

Most of the engines are accessed through the ssh connection of the worker. For example, the SGE running in a computer cluster is an engine. The user has to have authorization to schedule jobs in the SGE queue, and communicate the credentials to the GWM. Usually, any kind of engine returns to the user the identification for the job deployed, so the user can query information using that identification. The GWM calls that a ticket, stores it in the database for future queries of the condition of the task, and also exposes this ticket to the user, so he can manually login into the server and ask information about the task.

6.4.5 Database

The database is implemented using SQLite. This is a server-free database, that implements a subset of the SQL language, but allows very fast access, and the single-user database is perfect for deploying the GWM as a distributed system, in which each user has his database.

All the important information about workers and tasks is stored in the database, along with the encrypted credentials to access different workers or machines. We choose the SQLite database because it is ACID (Atomicity, Consistency, Isolation, Durability) compliant.

Every operation in the system is logged into the database. For each user that has an active account in the system the database manages its workers and tasks. There is no hierarchy of users or privileges, so the system is maintained in a simple manner. The only different user is the administrator, who has privileges to manage all the state of the system.

6.4.6 User Authentication

The user authentication is as easy as possible. Each user has an account in the system, identified with the username and password. The password is stored with a shadow and each login attempt is checked using the sha1 encryption. It works similarly to a UNIX account, and completely different to the GRID authentication schema which is a huge overhead to the deployment and usage of the system.

6.4.7 Task and Worker Definitions

To set a new task, a task definition file has to be written. This task definition file is inspired on the Job Description Language of gLite. To define a job, only the executable path, input and output files, output path for the results and engine is needed. The result of this is a small file with less than 10 lines for most of the tasks. With this task definition written, it can be submitted to the server by using the command line interface.

To define a new worker a similar file is needed. In this case, the fields required will depend on the computation service. For the simplest one, only an IP, username, authentication and also information about the engine that is configured in the server is necessary. If the worker belongs to a given computation service, a new set of options is needed, preceded by an identifier like Oracle VM, KVM or CloudStack. After this file is loaded into the system, the GWM will be able to communicate with the engine that is running in the worker, using the private key provided in the configuration file.

The tasks will evolve following a state machine approach controlled by the scheduler, and it can also be controlled from the command line interface. When the tasks are done, the results will appear in the folder defined by the user in the task configuration file. The results of each task will be stored in a unique folder, which its name is the identification number of the task.

6.5 Test production runs

Our full testing suite consists in deploying the system, adding a set of tasks and workers and iterating the tasks to solve them in the pool of workers. In order to test the main capabilities of the proposed GWM, we have used it in three different cases: Monte Carlo tasks in heterogeneous systems, image processing and nanoelectronic numerical simulation. So here we will describe the nature of the tasks and the experience of the execution using different computational resources.

The section 6.5.1 shows the capability of the system to manage several heterogeneous infrastructures. We have selected a Monte Carlo [47] code to deploy 100 simulations in several incompatible architectures, operating systems and queuing mechanisms.

In the section 6.5.2 we will show the adaptability of the system, running a standard linux tool to do image processing. We have used 100 high resolution images to study the performance of the system. This was used as an example of the fact that not only simulations but any task can be deployed, as long it is compatible with the workers.

The section 6.5.3 shows an in-house code developed to simulate nano semiconductor devices. Se have created 4000 tasks that represent variability study of the characteristics of a given semiconductor device. We are using this code in a daily basis and deploying thousands of simulations, using a big ammount of disk space for the IO, and MPI paralelism. This is representative of a general simulation code, with specific paralelism requirements, a big amount of IO and a particular code that needs to be compiled and deployed.

6.5.1 Monte Carlo on a heterogeneous system

In order to demonstrate that GWM is able to deal with heterogeneous infrastructures and workers, we have deployed a Monte Carlo simulation in four systems, with different queue engines, processors with several architectures that force us to deploy different binaries and different Operating Systems (OS). The computing facilities are provided by systems like cloud platforms and HPC and HTC clusters, as follows:

CLOUD In this case, the simulations were executed in a Virtual Machine (VM) deployed by means of Apache CloudStack 4.0.1 under KVM hypervisor. The physical Computer node have Intel Core i7-2600 at 3.4GHz with 8 GB of memory. This processor has 4 cores and 8 threads. The VM has a 1 Virtual QEMU CPU, 1 GB of memory and 5 GB of disk, employing CentOS 5.5 64 bits as OS. This VM does not has a queuing system installed, therefore the simulations will be automatically executed in a shell, and the queuing mechanism of the Linux kernel will be used to retrieve the state of the task.

FT The Finisterrae supercomputer in the CESGA instalations. This supercomputer is composed of 142 HP Integrity rx7640 nodes, with Itanium Montvale processors and 128 GB of memory per core. This supercomputer use SGE as the queuing system. The OS is SUSE Linux Enterprise Server 10 (ia64), in both the head and the computing nodes.

SVG The Super Ordenador Virtual Gallego in the CESGA instalations. This supercomputer is composed of 46 nodes with twin AMD Opteron 6174 processors at 2.2 Ghz with 12 cores, the memory per node varies between 32 and 64 GB. This supercomputer also uses SGE as queuing mechanism. The OS is Red Hat Enterprise Linux release 4 for the head, and Scientific Linux 6.4 for the computing nodes.

CTCOMP The CITIUS HPC Cluster, a local cluster used in our organization. This has a variety of computing nodes, but we are using the Dell PowerEdge M910, that has twin

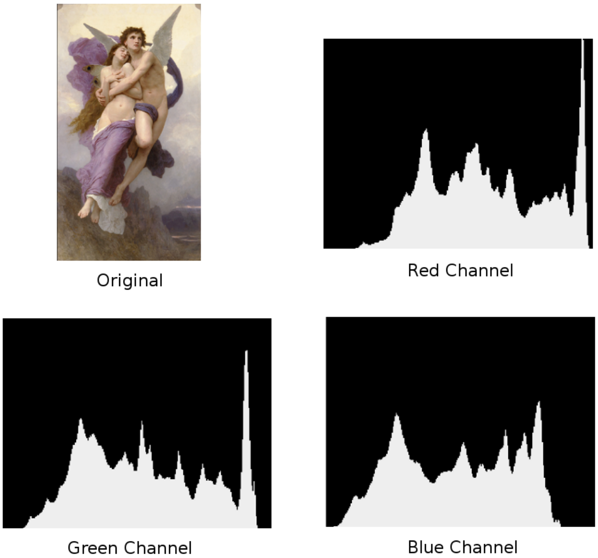

Figure 6.4: Original image and calculated histogram, for one of the 100 William Bouguereau artworks processed.

Intel Xeon L7555 with 8 nodes, and 64 GB of memory per node. Notice that this worker is using PBS as a queuing mechanism. The OS in all the nodes is Debian 7.1.

We have defined and deployed 100 Monte Carlo tasks in four different computing systems, using all of them to run the simulations. The distribution of the tasks between the workers is done dinamically by the GWM Schedulers. As a result, more tasks are deployed in the computing facilities that finish the simulations in less time. In our case the distribution of tasks is shown in Table 6.1, along with the time required to deploy the 100 tasks in each worker alone.

When using all the workers to deploy the simulations, a total time of 911 seconds was used to run all the 100 simulations. Compare this with the invidivual times that would be neccesary in the case of using only one of the available computing resources, in Table 6.1. Time measurements have been taken from the start of the first task deployment, to the end of the retrieval of the last task, including file transfers and the execution of the Monte Carlo simulation.

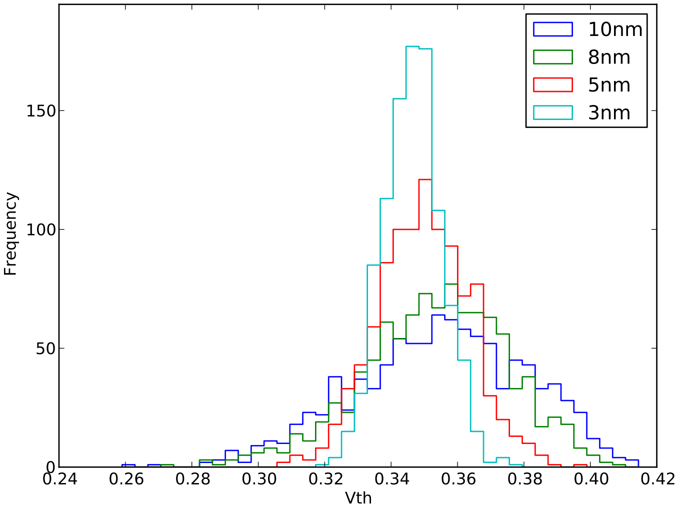

Figure 6.5: Histogram of the 4000 nanodevice simulations carried out in a cluster, with four different input parameter configurations.

6.5.2 Image handling

This example is a batch of image manipulation. This is a very useful task that can be needed in several areas of work. In our case, we are calculating a histogram of a high-resolution picture of an art-work. We collected 100 open-domain pictures of William Bouguereau to use as a benchmark, to represent a possible realistic workload. We are using the common Linux software Image Magick, which allows us to apply a gaussian blur and calculate the histogram of the result.

In this case, we are using the shell engine. Also, the input files are the script and all the pictures. Providing that we have about 100 pictures, this will spawn 100 tasks, and we only

Table 6.1: Distribution of the tasks among the workers when running together, and total time when running the worker alone.

Worker	Shared Task Count	Alone time (s)
FT	20	3285
SVGD	35	1746
CTCOMP	37	1469
CLOUD	8	6319

want the output file with the histograms of each color channel (represented in Fig. 6.4).

This code has been executed in a pool of 8 workers provided by CloudStack. This is independent of the task definition; any other computational service could be used, provided that it has the required software. Furthermore, the workers are KVM instances of CentOS with 512 MB of RAM each, with Image Magick installed.

Using the pool of workers, all the histograms were calculated from the high quality images in 60 minutes. If instead of the cloud pool, one of the hosts is used in a sequential execution the wall time needed is around 420 minutes, which is 7 times larger. This gives us a fairly good efficiency.

6.5.3 Nanodevice simulations

In order to represent a workload of computational intensive tasks, we have defined a big set of nanodevice simulations, and we used a local cluster to run the tasks and retrieve the data. The cluster is running SGE, and the user has access to the cluster in order to launch tasks.

The first step was to compile the executable in the cluster. The simulator needs some input files common to all the simulations except one, which we want to change in every simulation. What we want in this case is to take all the files in a given directory (work-function), and spawn a new task for each file in an automatic way.

The important point here is that each simulation will use about 100 MB of storage. The 4000 tasks are going to dry the available storage on the cluster, which is a shared and spare resource. Our system sends the data to each simulation, ran the program and then retrieved the data to the local computer of the scientist. The queue of the cluster only allowed 30 tasks to be run at the same time, which means that only about 3 GB of storage is used in the cluster during the execution, saving space for the other users of the computer system. If we just submit the 4000 tasks to the system without the general workload manager, the space needed would be of 40 GB.

The results are shown on a histogram in the Fig. 6.5. The 4000 simulations are grouped in 4 sets of 1000 simulations each with four different values to an internal parameter (10 to 3 nm). The histogram shows four histograms overlaped, one for each set.

6.6 Related Work

Several other software solutions exist that cope with the same problem, but in a different approach and with other limitations and strong points. For example, the Grid powered applications homogeneize the deployment of tasks in Grid infrastructures so that the user does not need to know about the underlying computational resources. It is presented with an easy to use interface and the scheduling and task deployment is automatic. This is very similar to our approach, but in our case the application that can be deployed is way generic that the field specific Grid applications.

Other solution is the Grid middleware. This is probably the most similar, as already explained in the introduction. The problem with this middleware is that is heavy, difficult to install and use, and tailored to the existing Grid infrastructure. From this point of view, our solution can be explained as a Grid middleware but capable of easily work with existing cloud providers, clusters, even local workstations.

Finally, several cloud computing solutions provide Platform as a Service. This is similar, but again tailored to programmers and limited to cloud frameworks. This Platform as a Service presents a set of libraries and storage capabilities so the developer can create an application, and it will use the underlying cloud infrastructure to run it. In our case, the task orientation can define our tool as a Task Manager as a Service.

6.7 Conclusion

We have developed a general-purpose very light and high efficient workload manager system allowing users to manage thousands of simulations in an easy way. It provides API, command-line, web interface and several deployment schedulers. Compared with the existing solutions like gLite, GWM provides: an integrated web interface, a simple configuration, and a modular and expandable codebase, while mantaining a syntax similar to the Job Description Language. This system allows scientists to work with different computing resources including HPC and HTC clusters, standalone worker nodes, hypervisor-enabled servers, and cloud platforms, by means of a ssh connection, without requiring any ad-hoc configuration in these systems.

Three different scenarios have been tested with GWM. The first one is intended to test the heterogeneous capabilities of our system. In this scenario, we have run the same Monte Carlo code in several processor architectures with incompatible scheduling engines, and also in a cloud provided virtual machine. The second one has the purpose of showing the ability of this tool to run applications that are not compiled scientific codes. In this case, we have employed an imaging processing tool to manipulate a collection of high resolution pictures to get their histogram data. In the last one, we present the ability of GWM to deal with heavy workloads in a production environment, deploying a set of 4000 heavyweight semiconductor device simulations.

The suggested general workload manager is capable of deploying, running, monitoring and retrieving sets of tasks in an automatic way using different infrastructures. It would help scientists to run hundreds or thousands of simulations using different computational resources in an easy way. GWM reduces the time necessary to distribute the tasks and collect the results improving the usage of computer resources.

CHAPTER 7

CONCLUSION

The author started this thesis with the objective of advancing the existing knowledge of semiconductor devices in the nanoscale regime. In order to do that, the analysis of variability sources was selected as an interesting combination that involves several abilities. On the one hand, it requires knowledge of the physical mechanisms that affect the semiconductors behavior, and also of the manufacturing process, because of its impact on the variability to be studied. On the other hand, it requires powerful tools to be able to simulate thousands of devices to understand the effect of small changes on the device characteristics.

As a starting point we developed a pipeline based in a perturbation model that allows to modify the simulation to account for different variability sources, without many changes in the simulator code. Using this pipeline, we have implemented two variability sources: the Metal Gate Granularity (MGG) and the Line Edge Roughness (LER). These variability sources have been applied to several devices: Silicon and InGaAs FinFETs and gate-all-around Nanowires. These tools are currently being used by other authors in the Universities of Santiago de Compostela and in Swansea University, to further study the effect of that variability sources.

The simulator that was used and modified is a drift-diffusion 3D simulator. It uses density gradient corrections to account for the quantum effects that arise when shrinking the device under certain sizes. The device is modeled with a tetrahedral mesh, because the simulator uses finite elements to discretize the problem. Several meshes where generated for this simulator, with different shape, size or density, to manage the associated convergence problems that can happen if the density is too low and to explore different architectures.

The Metal Gate Granularity was studied using our own approach which is based on the mathematical structure of the Voronoi diagram. To implement the Line Edge Roughness, we have developed a inverse Fourier transform of a spectra. To obtain comprehensive data of the effect of this variability sources in semiconductor devices, we need to change the parameters that define the sources of variability, and also use different devices. We deployed several thousands of simulations in several computing resources thanks to the General Workload Manager, which was also developed during all the period of this thesis.

The following bullet list summarizes some of the findings presented in the previous chapters that were achieved throughout this thesis:

- We have developed a pipeline based in a perturbation model, that allows to implement several variability sources in our semiconductor device simulators. This pipeline introduces the variability source as a perturbation, without many changes in the original source of the simulator. This is currently being used by several scientists from two different research institutions.

- One of the most important applications of this pipeline, the Voronoi approach for the Metal Gate Granularity variability, was presented and validated against experimental data. These values have been provided by Dr. Kenji Ohmori [30], and consisted on TEM images of different materials: TiN and Ru. In both cases, our Voronoi approach generates a grain distribution that fits properly the experimental grain distribution, with p-values of 0.17 and 0.42, for TiN and Ru, respectively. We have also checked with the same experimental data an option developed by another authors: the Rayleigh approach, and concluded that is not adequate to account for the grain distribution of MGG simulations. The same fit to the same experimental data resulted in p-values of 3×10^{-14} and 0.0029 for TiN and Ru. We have also demonstrated that the variability calculated with Rayleigh overestimates the real variability by 11.9% and 7.14% for TiN and TaN materials, which our approach does not.

- Using the presented pipeline, we have analyzed the impact of the MGG and LER sources of variability in the performance of several state-of-the-art semiconductor devices. This is a key point to understand the process of device fabrication and how it has an impact on the device characteristics. We have simulated 10.7 and 10.4 nm gate length Silicon and InGaAs devices modeled according to the ITRS predictions. Those simulations were calibrated using the data from a more precise but slower simulator, based on 3-D Non-Equilibrium Green's Functions, because no experimental data was available at the moment. From this comparison we have found that the InGaAs device is more resilient to the variability sources in the subthreshold regime. The behavior for the on-current variability is the opposite, having more sensitivity in the InGaAs device.

- Independently of the device, for the MGG we have found and characterized a dependency of the variability on threshold voltage, off current and subthreshold swing with the inverse of the root square of the grain size. Also, we have found that the device power consumption and switching speed diminish when the grain size if large. This means that not only a variation of the parameters is to be expected, but also a net reduction of the quality of the device.

- Regarding the LER, we have found that the effect of the correlation length is smaller than the effect of the root mean square of the height, for the parameters that are usually studied. This result is found to be applicable for both Silicon and InGaAs FinFETs. We have also studied the impact of correlated versus uncorrelated LER, and we have concluded that the uncorrelated LER has more impact on the variability because it changes the device width along the current flow direction.

- In order to further understand the effect of the variability sources, we have implemented and presented a Fluctuation Sensitivity Map (FSM) to study the MGG variability. The FSM shows us that we can detect the position in the device where the oxide is wider, because it reduces the sensitivity of the device to the grain orientation. Also, we have found that a reduction of the width of the device body near the top of the Fin has a similar effect of a oxide buffer: it reduces the sensitivity.

- Finally, regarding the infrastructure to manage tasks, the GWM, we have tested it using heterogeneous work loads, computing resources incompatible between themselves, different queuing engines for the tasks, and cloud infrastructures. We have also benchmarked the system with a 16 nodes cloud machine, and found that the GWM is capable of keeping a mean usage of 14.98 nodes during the simulations, leveraging the available resources. Almost all the simulations of this work have been carried with this tool, and the results are positive.

7.1 Future work

We present here a comprehensive list of future tasks that can be carried in order to continue the work started in this thesis.

- The MGG variability can be further improved by taking into consideration the effect of the gate-first and gate-last techniques. Doing this, we could generate Voronoi grains that represent the gate in the two possible implantation techniques and compare them directly.

- The LER variability source can be applied in different lines of the device. We have only used the most important, the FER, which is applied in the body of the device in the direction of the current flow. Applying this variability along the gate, transverse to the device, may prove useful.

- The FSM can be applied to another variability source other than MGG, and also for more devices geometry in order to improve our knowledge of the sensitivity of the device.

- GWM is being expanded right now to implement new mechanisms, like dependency between tasks that allows the user to define not only a task but a pipeline of data between tasks. This would allow complex interactions to be carried on automatically.

Bibliography

[1] Kelin J. Kuhn, Martin D. Giles, David Becher, Pramod Kolar, Avner Kornfeld, Roza Kotlyar, Sean T. Ma, Atul Maheshwari, and Sivakumar Mudanai. Process Technology Variation. *IEEE Transactions on Electron Devices*, 58(8):2197–2208, aug 2011.

[2] The International Technology Roadmap for Semiconductors (ITRS), 2009, http://www.itrs2.net/. 2009.

[3] The International Technology Roadmap for Semiconductors (ITRS), 2011, http://www.itrs2.net/. 2011.

[4] T Matsukawa, S O, K Endo, Y Ishikawa, H Yamauchi, Y X Liu, J Tsukada, K Sakamoto, and M Masahara. Comprehensive Analysis of Variability Sources of FinFET Characteristics. In *Symposium on VLSI Technology*, pages 159–160, 2009.

[5] R.E. Shannon. Introduction to the art and science of simulation. In *1998 Winter Simulation Conference. Proceedings (Cat. No.98CH36274)*, volume 1, pages 7–14. IEEE, 1998.

[6] A.B. Fortes, J. Figueiredo, and M.S. Lundstrom. Virtual Computing Infrastructures for Nanoelectronics Simulation. *Proceedings of the IEEE*, 93(10):1839–1847, oct 2005.

[7] Antonio Jesus Garcia-Loureiro, Natalia Seoane, Manuel Aldegunde, Raúl Valin, Asen Asenov, Antonio Martinez, and Karol Kalna. Implementation of the Density Gradient Quantum Corrections for 3-D Simulations of Multigate Nanoscaled Transistors. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30(6):841–851, jun 2011.

[8] Jari Lindberg, Manuel Aldegunde, Daniel Nagy, Wulf G Dettmer, Karol Kalna, Antonio Jesus Garcia-Loureiro, and Djordje Peric. Quantum Corrections Based on the 2-D Schrödinger Equation for 3-D Finite Element Monte Carlo Simulations of Nanoscaled FinFETs. *IEEE Transactions on Electron Devices*, 61(2):423–429, feb 2014.

[9] Manuel Aldegunde, Antonio Jesus Garcia-Loureiro, and Karol Kalna. 3D Finite Element Monte Carlo Simulations of Multigate Nanoscale Transistors. *IEEE Transactions on Electron Devices*, 60(5):1561–1567, may 2013.

[10] Manuel Aldegunde and K. Kalna. Energy conserving, self-force free Monte Carlo simulations of semiconductor devices on unstructured meshes. *Computer Physics Communications*, 189:31–36, apr 2015.

[11] Antonio Martinez, Manuel Aldegunde, Natalia Seoane, Andrew R. Brown, John R. Barker, and Asen Asenov. Quantum-Transport Study on the Impact of Channel Length and Cross Sections on Variability Induced by Random Discrete Dopants in Narrow Gate-All-Around Silicon Nanowire Transistors. *IEEE Transactions on Electron Devices*, 58(8):2209–2217, aug 2011.

[12] Xiao Zhang, Jing Li, M Grubbs, M Deal, B Magyari-Kope, B M Clemens, and Y Nishi. Physical model of the impact of metal grain work function variability on emerging dual metal gate MOSFETs and its implication for SRAM reliability. In *IEEE Electron Device Letters*, pages 1–4, 2009.

[13] K. Sivasankaran, P S Mallick, and T R K Kumar Chitroju. Impact of device geometry and doping concentration variation on electrical characteristics of 22nm FinFET. In *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, number Iceccn, pages 528–531. IEEE, mar 2013.

[14] Nattapol Damrongplasit, Sung Hwan Kim, Changhwan Shin, and Tsu-Jae King Liu. Impact of Gate Line-Edge Roughness (LER) Versus Random Dopant Fluctuations (RDF) on Germanium-Source Tunnel FET Performance. *IEEE Transactions on Nanotechnology*, 12(6):1061–1067, nov 2013.

[15] E Baravelli, A Dixit, R Rooyackers, M Jurczak, N Speciale, and K De Meyer. Impact of Line-Edge Roughness on FinFET Matching Performance. *IEEE Transactions on Electron Devices*, 54(9):2466–2474, sep 2007.

[16] D Reid, C Millar, G Roy, S Roy, and Asen Asenov. Analysis of threshold voltage distribution due to random dopants: A 100 000-sample 3-D simulation study. *IEEE Transactions on Electron Devices*, 56(10):2255–2263, 2009.

[17] Natalia Seoane, Antonio Jesus Garcia-Loureiro, K. Kalna, and Asen Asenov. Impact of intrinsic parameter fluctuations on the performance of HEMTs studied with a 3D parallel drift-diffusion simulator. *Solid-State Electronics*, 51(3):481–488, mar 2007.

[18] Muhammad A. Elmessary, Daniel Nagy, Manuel Aldegunde, Natalia Seoane, Guillermo Indalecio, Jari Lindberg, Wulf Dettmer, Djordje Peric, Antonio J. Garcia-Loureiro, and Karol Kalna. Scaling/LER study of Si GAA nanowire FET using 3D Finite Element Monte Carlo simulations. *2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, pages 52–55, 2016.

[19] Natalia Seoane, Guillermo Indalecio, E. Comesana, Antonio Jesus Garcia-Loureiro, Manuel Aldegunde, and K. Kalna. Three-dimensional simulations of random dopant and metal-gate workfunction variability in an In0.53Ga0.47As GAA MOSFET. *IEEE Electron Device Letters*, 34(2):205–207, feb 2013.

[20] Natalia Seoane, Guillermo Indalecio, Enrique Comesana, Manuel Aldegunde, Antonio Jesus Garcia-Loureiro, and Karol Kalna. Random Dopant, Line-Edge Roughness, and Gate Workfunction Variability in a Nano InGaAs FinFET. *IEEE Transactions on Electron Devices*, 61(2):466–472, feb 2014.

[21] Guillermo Indalecio, Natalia Seoane, Manuel Aldegunde, K Kalna, and Antonio Jesus Garcia-Loureiro. Scaling of metal gate workfunction variability in nanometer SOI-FinFETs. In *2014 15th International Conference on Ultimate Integration on Silicon (ULIS)*, number Dd, pages 105–108. IEEE, apr 2014.

[22] Natalia Seoane, Guillermo Indalecio, Manuel Aldegunde, Daniel Nagy, Muhammad A. Elmessary, Antonio J. García-Loureiro, and Karol Kalna. Comparison of Fin-Edge Roughness and Metal Grain Work Function Variability in InGaAs and Si FinFETs. *IEEE Transactions on Electron Devices*, 63(3):1209–1216, 2016.

[23] A Dixit, K. G. Anil, E Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, and K. De Meyer.

Impact of Stochastic Mismatch on Measured SRAM Performance of FinFETs with Resist/Spacer-Defined Fins: Role of Line-Edge-Roughness. In *2006 International Electron Devices Meeting*, volume 3001, pages 1–4. IEEE, 2006.

[24] G Kokkoris, V Constantoudis, and E Gogolides. Nanoscale roughness effects at the interface of lithography and plasma etching: Modeling of line-edge-roughness transfer during plasma etching. *Plasma Science, IEEE Transactions on*, 37(9):1705–1714, 2009.

[25] K Patel, Tsu-Jae King Liu, and C J Spanos. Gate Line Edge Roughness Model for Estimation of FinFET Performance Variability. *IEEE Transactions on Electron Devices*, 56(12):3055–3063, 2009.

[26] Asen Asenov, S. Kaya, and A.R. Brown. Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness. *IEEE Transactions on Electron Devices*, 50(5):1254–1260, may 2003.

[27] S Yu, Y Zhao, L Zeng, G Du, J Kang, R Han, and X Liu. Impact of line-edge roughness on double-gate Schottky-barrier field-effect transistors. *IEEE Transactions on Electron Devices*, 56(6):1211–1219, 2009.

[28] P Oldiges, Q Lin, K Petrillo, M Sanchez, M Ieong, and M Hargrove. Modeling line edge roughness effects in sub 100 nanometer gate length devices. In *Simulation of Semiconductor Processes and Devices, 2000. SISPAD 2000. 2000 International Conference on*, pages 131–134. IEEE, 2000.

[29] Shimeng Yu, Yuning Zhao, Yuncheng Song, Gang Du, Jinfeng Kang, Ruqi Han, and Xiaoyan Liu. Full 3-D simulation of gate line edge roughness impact on sub-30nm FinFETs. In *2008 IEEE Silicon Nanoelectronics Workshop*, pages 1–2. IEEE, jun 2008.

[30] Kenji Ohmori, T Matsuki, D Ishikawa, T Morooka, T Aminaka, Y Sugita, T Chikyow, K Shiraishi, Y Nara, and K Yamada. Impact of additional factors in threshold voltage variability of metal/high-k gate stacks and its reduction by controlling crystalline structure and grain size in the metal gates. In *2008 IEEE International Electron Devices Meeting*, number 110, pages 1–4. IEEE, dec 2008.

[31] Yiming Li, Hui-Wen Cheng, Chun-Yen Yiu, and Hsin-Wen Su. Nanosized metal grains induced electrical characteristic fluctuation in 16-nm-gate high-κ/metal gate bulk FinFET devices. *Microelectronic Engineering*, 88(7):1240–1242, jul 2011.

[32] Xingsheng Wang, Andrew R. Brown, Niza Idris, Stanislav Markov, Gareth Roy, and Asen Asenov. Statistical Threshold-Voltage Variability in Scaled Decananometer Bulk HKMG MOSFETs: A Full-Scale 3-D Simulation Scaling Study. *IEEE Transactions on Electron Devices*, 58(8):2293–2301, aug 2011.

[33] Hamed F Dadgour, Kazuhiko Endo, Vivek K De, and Kaustav Banerjee. Grain-Orientation Induced Work Function Variation in Nanoscale Metal-Gate Transistors—Part I: Modeling, Analysis, and Experimental Validation. *IEEE Transactions on Electron Devices*, 57(10):2504–2514, oct 2010.

[34] Hamed F Dadgour, Vivek De, and Kaustav Banerjee. Modeling and analysis of grain-orientation effects in emerging metal-gate devices and implications for SRAM reliability. *2008 IEEE International Electron Devices Meeting*, 3:1–4, dec 2008.

[35] Hyohyun Nam and Changhwan Shin. Study of High-k/Metal-Gate Work-Function Variation Using Rayleigh Distribution. *IEEE Electron Device Letters*, 34(4):532–534, apr 2013.

[36] Guillermo Indalecio, Antonio Jesus Garcia-Loureiro, Manuel Aldegunde, and Karol Kalna. 3D Simulation Study of Work-Function Variability in a 25 nm Metal-Gate FinFET with Curved Geometry using Voronoi Grains. In *Simulation of Semiconductor Processes and Devices (SISPAD), 2012 International Conference on*, pages 149–152, 2012.

[37] Járai-Szabó Ferenc and Zoltán Néda. On the size distribution of Poisson Voronoi cells. *Physica A: Statistical Mechanics and its Applications*, 385(2):518–526, nov 2007.

[38] Guillermo Indalecio, Antonio J. Garcia-Loureiro, Natalia Seoane, and Karol Kalna. Study of Metal-Gate Work-Function Variation Using Voronoi Cells: Comparison of Rayleigh and Gamma Distributions. *IEEE Transactions on Electron Devices*, 63(6):2625–2628, 2016.

[39] Hyohyun Nam and Changhwan Shin. Comparative study in work-function variation: Gaussian vs. Rayleigh distribution for grain size. *IEICE Electronics Express*, 10(9):20130109–20130109, 2013.

[40] Hyohyun Nam and Changhwan Shin. Study of High-k/Metal-Gate Work Function Variation in FinFET: The Modified RGG Concept. *IEEE Electron Device Letters*, 34(12):1560–1562, dec 2013.

[41] Guillermo Indalecio, Natalia Seoane, Manuel Aldegunde, K Kalna, and Antonio Jesus Garcia-Loureiro. Variability characterisation of nanoscale Si and InGaAs FinFETs at subthreshold. In *2014 5th European Workshop on CMOS Variability (VARI)*, pages 1–6. IEEE, sep 2014.

[42] Guillermo Indalecio, Manuel Aldegunde, Natalia Seoane, K Kalna, and Antonio Jesus Garcia-Loureiro. Statistical study of the influence of LER and MGG in SOI MOSFET. *Semiconductor Science and Technology*, 29(4):045005, apr 2014.

[43] N. Seoane, Guillermo Indalecio, K. Kalna, and A.J. García-Loureiro. Impact of cross-section of 10.4 nm gate length $In_{0.53}Ga_{0.47}$ FinFETs on metal grain variability. In *International Conference on Simulation of Semiconductor Processes and Devices SISPAD*, 2016.

[44] David Bernholdt, Shishir Bharathi, David Brown, Kasidit Chanchio, Meili Chen, A N N Chervenak, Luca Cinquini, B O B Drach, I A N Foster, Peter Fox, Jose Garcia, Carl Kesselman, R O B Markel, D O N Middleton, Veronika Nefedova, Line Pouchard, Arie Shoshani, Alex Sim, and Gary Strand. The Earth System Grid : Supporting the Next Generation of Climate Modeling Research. 93(3):485–495, 2005.

[45] C. Manuali, a. Laganà, and S. Rampino. GriF: A Grid framework for a Web Service approach to reactive scattering. *Computer Physics Communications*, 181(7):1179–1185, jul 2010.

[46] Maria Mirto, Sandro Fiore, Italo Epicoco, Massimo Cafaro, Silvia Mocavero, Euro Blasi, and Giovanni Aloisio. A Bioinfomatics Grid Alignment Toolkit. *Future Generation Computer Systems*, 24(7):752–762, jul 2008.

[47] Nancy Wilkins-Diehr. Special Issue: Science Gateways—Common Community Interfaces to Grid Resources. *Concurrency and Computation: Practice and Experience*, 19(6):743–749, apr 2007.

[48] Bhaskar Prasad Rimal, Eunmi Choi, and Ian Lumb. A Taxonomy and Survey of Cloud Computing Systems. *2009 Fifth International Joint Conference on INC, IMS and IDC*, pages 44–51, 2009.

[49] Hartwig Anzt, Jack Dongarra, and Enrique S. Quintana-Ortí. Adaptive precision solvers for sparse linear systems. *Proceedings of the 3rd International Workshop on Energy Efficient Supercomputing - E2SC '15*, (April):1–10, 2015.

[50] J Kavalieros, B Doyle, S Datta, G Dewey, M Doczy, B Jin, D Lionberger, M Metz, W Rachmady, M Radosavljevic, U Shah, N Zelick, and R Chau. Tri-Gate Transistor Architecture with High-k Gate Dielectrics, Metal Gates and Strain Engineering. In *VLSI Technology, 2006. Digest of Technical Papers. 2006 Symposium on*, pages 50–51, 2006.

[51] D Reid, C Millar, S Roy, and Asen Asenov. Understanding LER-Induced MOSFET V_ {T} Variability—Part I: Three-Dimensional Simulation of Large Statistical Samples. *IEEE Transactions on Electron Devices*, 57(11):2801–2807, 2010.

[52] D Reid, C Millar, S Roy, and Asen Asenov. Understanding LER-Induced MOSFET V_ {T} Variability—Part II: Reconstructing the Distribution. *IEEE Transactions on Electron Devices*, 57:2808–2813, 2010.

[53] M. Ancona and G. Iafrate. Quantum correction to the equation of state of an electron gas in a semiconductor. *Physical Review B*, 39(13):9536–9540, may 1989.

[54] V S Basker, T Standaert, H Kawasaki, C.-C. Yeh, K Maitra, T Yamashita, J Faltermeier, H. Adhikari, H. Jagannathan, J Wang, H. Sunamura, S. Kanakasabapathy, S. Schmitz, J. Cummings, A. Inada, C. H. Lin, P. Kulkarni, Y. Zhu, J. Kuss, T. Yamamoto, A. Kumar, J. Wahl, A. Yagishita, L. F. Edge, R. H. Kim, E. Mclellan, S. J. Holmes, R. C. Johnson, T. Levin, J. Demarest, M. Hane, M. Takayanagi, M. Colburn, V K Paruchuri, R J Miller, H Bu, B Doris, D. McHerron, E Leobandung, and J. O'Neill. A 0.063 um2 FinFET SRAM cell demonstration with conventional lithography using a novel

integration scheme with aggressively scaled fin and gate pitch. In *2010 Symposium on VLSI Technology*, volume 50, pages 19–20. IEEE, jun 2010.

[55] Sujata Paul, Frank Yeh, Kingsuk Maitra, Andreas Kerber, Pranita Kulkarni, Hemanth Jagannathan, Veeraraghavan S Basker, and Robert J Miller. Extraction of Effective Oxide Thickness for SOI FINFETs With High-k Metal Gates Using the Body Effect. In *IEEE Electron Device Letters*, volume 31, pages 650–652, jul 2010.

[56] Xingsheng Wang, Andrew R Brown, and Asen Asenov. Statistical variability and reliability in nanoscale FinFETs. In *2011 International Electron Devices Meeting*, volume 58, pages 5.4.1–5.4.4. IEEE, dec 2011.

[57] Yiming Li, Hui-Wen Cheng, and Ming-Hung Han. Quantum hydrodynamic simulation of discrete-dopant fluctuated physical quantities in nanoscale FinFET. *Computer Physics Communications*, 182(1):96–98, jan 2011.

[58] Hui-Wen Cheng, Fu-Hai Li, Ming-Hung Han, Chun-Yen Yiu, Chia-Hui Yu, Kuo-Fu Lee, and Yiming Li. 3D device simulation of work function and interface trap fluctuations on high-k / metal gate devices. In *IEEE Electron Device Letters*, pages 15.6.1 –15.6.4, 2010.

[59] Kai Kadau, Timothy C Germann, Peter S Lomdahl, Robert C Albers, Justin S Wark, Andrew Higginbotham, and Brad Lee Holian. Shock Waves in Polycrystalline Iron. *Phys. Rev. Lett.*, 98(13):135701, mar 2007.

[60] H Van Swygenhoven, D Farkas, and A Caro. Grain-boundary structures in polycrystalline metals at the nanoscale. *Phys. Rev. B*, 62(2):831–838, jul 2000.

[61] Horacio D. Espinosa and Pablo D. Zavattieri. *A grain level model for the study of failure initiation and evolution in polycrystalline brittle materials. Part I: Theory and numerical implementation*, volume 35. mar 2003.

[62] Alberto Leonardi, Paolo Scardi, and Matteo Leoni. Realistic nano-polycrystalline microstructures: beyond the classical Voronoi tessellation. *Philosophical Magazine*, 92(8):986–1005, 2012.

[63] Zhigang Fan, Yugong Wu, Xuanhe Zhao, and Yuzhu Lu. Simulation of polycrystalline structure with Voronoi diagram in Laguerre geometry based on random closed packing of spheres. *Computational Materials Science*, 29(3):301–308, mar 2004.

[64] F J Garcia Ruiz, A Godoy, F Gamiz, C Sampedro, and L Donetti. A Comprehensive Study of the Corner Effects in Pi-Gate MOSFETs Including Quantum Effects. *IEEE Transactions on Electron Devices*, 54(12):3369–3377, 2007.

[65] Hamed F Dadgour and Kaustav Banerjee. Statistical modeling of metal-gate Work-Function Variability in emerging device technologies and implications for circuit design. In *2008 IEEE/ACM International Conference on Computer-Aided Design*, pages 270–277. IEEE, nov 2008.

[66] Qintao Zhang, Cindy Wang, Hailing Wang, Christopher Schnabel, Dae-gyu Park, Scott K Springer, and Effendi Leobandung. Experimental Study of Gate-First FinFET Threshold-Voltage Mismatch. *IEEE Transactions on Electron Devices*, 61(2):643–646, feb 2014.

[67] S Deora, G Bersuker, T. W. Kim, D H Kim, C Hobbs, P D Kirsch, K. C. Sahoo, and A. S. Oates. Positive bias instability in gate-first and gate-last InGaAs channel n-MOSFETs. In *2014 IEEE International Reliability Physics Symposium*, pages 3C.5.1–3C.5.4. IEEE, jun 2014.

[68] Seid Hadi Rasouli, Kazuhiko Endo, Jone F. Chen, Navab Singh, and Kaustav Banerjee. Grain-Orientation Induced Quantum Confinement Variation in FinFETs and Multi-Gate Ultra-Thin Body CMOS Devices and Implications for Digital Design. *IEEE Transactions on Electron Devices*, 58(8):2282–2292, aug 2011.

[69] S. DiCenzo and G. Wertheim. Monte Carlo calculation of the size distribution of supported clusters. *Physical Review B*, 39(10):6792–6796, apr 1989.

[70] Shao-Heng Chou, Ming-Long Fan, and Pin Su. Investigation and Comparison of Work Function Variation for FinFET and UTB SOI Devices Using a Voronoi Approach. *IEEE Transactions on Electron Devices*, 60(4):1485–1489, 2013.

[71] P R Bevington and D K Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill Higher Education. McGraw-Hill, 2003.

[72] Hamed F Dadgour, Kazuhiko Endo, Vivek K De, and Kaustav Banerjee. Grain-Orientation Induced Work Function Variation in Nanoscale Metal-Gate Transistors—Part II: Implications for Process, Device, and Circuit Design. *IEEE Transactions on Electron Devices*, 57(10):2515–2525, oct 2010.

[73] Guillermo Indalecio, Natalia Seoane, Manuel Aldegunde, K. Kalna, and Antonio Jesus Garcia-Loureiro. Variability Characterisation of Nanoscale Si and InGaAs Fin Field-Effect-Transistors at Subthreshold. *Journal of Low Power Electronics*, 11(2):256–262, jun 2015.

[74] P Oldiges, R Muralidhar, P Kulkarni, C-H. Lin, K Xiu, D Guo, M Bajaj, and N Sathaye. Critical analysis of 14nm device options. In *2011 International Conference on Simulation of Semiconductor Processes and Devices*, pages 5–8. IEEE, sep 2011.

[75] Hui-Wen Cheng, Yiming Li, Chun-Yen Yiu, and Hsin-Wen Su. Nanosized metal grains induced electrical characteristic fluctuation in 16 nm bulk and SOI FinFET devices with TiN/HfO<inf>2</inf> gate stack. *2011 International Conference on Simulation of Semiconductor Processes and Devices*, pages 287–290, sep 2011.

[76] J Colinge. The New Generation of SOI MOSFETs. *Romanian Journal of Information Science and Technology*, 11(1):3–15, 2008.

[77] Guillermo Indalecio, Antonio Jesus Garcia-Loureiro, and Manuel Aldegunde. Static multipole method applied to boundary conditions for semiconductor device simulations. In *2012 International Conference on High Performance Computing & Simulation (HPCS)*, number 1, pages 654–659. IEEE, jul 2012.

[78] J.A. Croon, G Storms, S Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, and H.E. Maes. Line edge roughness: characterization, modeling and impact on device behavior. In *Digest. International Electron Devices Meeting,*, pages 307–310. IEEE, 2002.

[79] Galician Centre for Supecomputation, http://www.cesga.es/. 2016.

[80] Greg Leung and Chi On Chui. Variability of Inversion-Mode and Junctionless FinFETs due to Line Edge Roughness. *IEEE Electron Device Letters*, 32(11):1489–1491, nov 2011.

[81] Terence B. Hook. Fully depleted devices for designers: FDSOI and FinFETs. In *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, number Figure 3, pages 1–7. IEEE, sep 2012.

[82] Karol Kalna, Natalia Seoane, Antonio Jesus Garcia-Loureiro, I.G. Thayne, and Asen Asenov. Benchmarking of Scaled InGaAs Implant-Free NanoMOSFETs. *IEEE Transactions on Electron Devices*, 55(9):2297–2306, sep 2008.

[83] Zheng Guo, Andrew Carlson, Liang-Teck Pang, Kenneth T. Duong, Tsu-Jae King Liu, and Borivoje Nikolic. Large-Scale SRAM Variability Characterization in 45 nm CMOS. *IEEE Journal of Solid-State Circuits*, 44(11):3174–3192, nov 2009.

[84] Yu Cao, Jyothi Velamala, Ketul Sutaria, Mike Shuo-Wei Chen, Jonathan Ahlbin, Ivan Sanchez Esqueda, Michael Bajura, and Michael Fritze. Cross-Layer Modeling and Simulation of Circuit Reliability. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(1):8–23, jan 2014.

[85] Cheng Zhuo, Kaviraj Chopra, Dennis Sylvester, and David Blaauw. Process Variation and Temperature-Aware Full Chip Oxide Breakdown Reliability Analysis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30(9):1321–1334, sep 2011.

[86] Nidhi Agrawal, Yoshie Kimura, Reza Arghavani, and Suman Datta. Impact of Transistor Architecture (Bulk Planar, Trigate on Bulk, Ultrathin-Body Planar SOI) and Material (Silicon or III-V Semiconductor) on Variation for Logic and SRAM Applications. *IEEE Transactions on Electron Devices*, 60(10):3298–3304, oct 2013.

[87] The International Technology Roadmap for Semiconductors (ITRS), 2013, http://www.itrs2.net/. 2013.

[88] Natalia Seoane, Manuel Aldegunde, Antonio Jesus Garcia-Loureiro, R Valin, and K Kalna. 3D ‘atomistic’ simulations of dopant induced variability in nanoscale implant free $In_{0.75}Ga_{0.25}As$ MOSFETs. *Solid-State Electronics*, 69(0):43–49, 2012.

[89] A Islam and K Kalna. Monte Carlo simulations of mobility in doped GaAs using self-consistent Fermi–Dirac statistics. *Semiconductor Science and Technology*, 26(5):055007, may 2011.

[90] Silvaco. *ATLAS User’s Manual.* Santa Clara, CA, USA, 2012.

[91] Carlo Jacoboni and Lino Reggiani. The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials. *Reviews of Modern Physics*, 55(3), 1983.

[92] D. Ferry. *Semiconductor Transport.* Taylor & Francis, New York, USA, 2000.

[93] B K Ridley. Reconciliation of the Conwell-Weisskopf and Brooks-Herring formulae for charged-impurity scattering in semiconductors: Third-body interference. *Journal of Physics C: Solid State Physics*, 10(10):1589–1593, may 1977.

[94] Aynul Islam, Brahim Benbakhti, and Karol Kalna. Monte Carlo Study of Ultimate Channel Scaling in Si and $In_{0.3}Ga_{0.7}As$ Bulk MOSFETs. *IEEE Transactions on Nanotechnology*, 10(6):1424–1432, nov 2011.

[95] Yang Liu, Neophytos Neophytou, Gerhard Klimeck, and Mark S. Lundstrom. Band-Structure Effects on the Performance of III-V Ultrathin-Body SOI MOSFETs. *IEEE Transactions on Electron Devices*, 55(5):1116–1122, may 2008.

[96] M. G. Ancona. Density-gradient theory: a macroscopic approach to quantum confinement and tunneling in semiconductor devices. *Journal of Computational Electronics*, 10(1-2):65–97, jun 2011.

[97] Jiangjiang J. Gu, Heng Wu, Yiqun Liu, Adam T. Neal, Roy G. Gordon, and Peide D. Ye. Size-Dependent-Transport Study of $In_{0.53}Ga_{0.47}As$ Gate-All-Around Nanowire MOSFETs: Impact of Quantum Confinement and Volume Inversion. *IEEE Electron Device Letters*, 33(7):967–969, jul 2012.

[98] Tiziana Ferrari and Luciano Gaido. Resources and Services of the EGEE Production Infrastructure. *Journal of Grid Computing*, 9(2):119–133, mar 2011.

[99] A. Bala. Development of Grid e-Infrastructure in South-Eastern Europe. *Journal of Grid Computing*, 9:135–154, 2011.

[100] Francisco Brasileiro, Matheus Gaudencio, Rafael Silva, Alexandre Duarte, Diego Carvalho, Diego Scardaci, Leandro Ciuffo, Rafael Mayo, Herbert Hoeger, Michael Stanton, Raul Ramos, Roberto Barbera, Bernard Marechal, and Philippe Gavillet. Using a Simple Prioritisation Mechanism to Effectively Interoperate Service and Opportunistic Grids in the EELA-2 e-Infrastructure. *Journal of Grid Computing*, 9(2):241–257, jan 2011.

[101] Carmen Cotelo, Andrés Gómez, J. Ignacio López, David Mera, José M. Cotos, J. Pérez Marrero, and Constantino Vázquez. Retelab: A geospatial grid web laboratory

for the oceanographic research community. *Future Generation Computer Systems*, 26(8):1157–1164, oct 2010.

[102] Andrei Tsaregorodtsev, Vincent Garonne, and Ian Stokes-rees. DIRAC : A Scalable Lightweight Architecture for High Throughput Computing.

List of Figures

List of Tables