DISS. ETH NO. 27345

# *AB INITIO* QUANTUM TRANSPORT IN CONDUCTIVE BRIDGING RANDOM ACCESS MEMORY

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

FABIAN DUCRY
MSc EEIT, ETH Zurich
born on 6 November 1988
citizen of Zurich, Switzerland

accepted on the recommendation of

Prof. Dr. Mathieu Luisier, examiner
Prof. Dr. Alexander Shluger, co-examiner

2021

# Contents

# Acknowledgments

Foremost I would like to express my gratitude to Prof. Dr. Mathieu Luisier for supervising my PhD project and giving me the opportunity to work on this fascinating research topic. I highly value his support and advice and the freedom I was given in pursuing my work. I would also like to thank Prof. Dr. Alexander Shluger for accepting to be the co-examiner of my thesis.

I thank Prof. Dr. Jürg Leuthold and Prof. Dr. Thomas Schimmel for the productive collaboration and discussions. This is extended to all members of this research effort as part of the Werner Siemens Stiftung Center for Single Atom Electronics and Photonics, namely Dr. Alexandros Emboras, Dr. Ueli Koch, Dr. Roland Gröger, Dr. Stefan Walheim, Dr. Florian Wertz, Dr. Fangqing Xie, Jan Äschlimann, Bojun Cheng, Mila Lewerenz, Elias Passerini, Torben Staiger, and Samuel Zumtaugwald.

I am grateful to the people of the Nano-TCAD group for the pleasant atmosphere, inspiring discussions, and encouragement, in particular Cedric Klinkert, Dr. Dominik Bauer and Dr. Tarun Agarwal. I greatly enjoyed working with all of you.

Furthermore, I would like to thank Christine Haller for handling administrative matters as well as Christoph Wicki, Frank K. Gürkaynak, and Adam Feigin for keeping the IT-infrastructure running and dealing with all my problems and requests.

Finally, I am deeply grateful to my family, to my parents, Esther and Andre, for all their support throughout my life, to my sister Jasmin, and to Theresa. Without her encouragement and support this thesis would not have been possible.

# Abstract

In line with the growing popularity of data-driven IT applications, the importance of data storage and memory has not stopped increasing during the last decade. However, the complementary metal-oxide-semiconductor (CMOS)-based storage and memory technology is reaching its scaling limits and has become ill-suited for low power and energy efficient operations. Therefore, new nonvolatile memory technologies are being developed and are slowly entering the mass market. Among those, conductive bridging random access memory (CBRAM) cells promise great advances in terms of power consumption, integration density, and cointegration with CMOS-based logic circuits. CBRAMs are metal-insulator-metal heterojunctions through which a metallic filament grows and dissolves, which creates two distinctive resistance states corresponding to logical 1 or 0. Nevertheless, several challenges remain to be addressed before they can compete with traditional technologies, particularly with respect to variability and reliability. Solving these issues is complicated by the fact that the operating mechanisms of CBRAMs are not fully understood.

The aim of this thesis was to use quantum transport simulations to elucidate the switching characteristics of CBRAM cells. We used density-functional theory to compute the electronic structure of nanoscale components, from which structural, electrical, thermal and electro-thermal properties of memory cells were derived in a parameter-free manner.

First, the theoretical foundation of first principles-based quantum transport is presented. Such simulation techniques induce an immense computational burden when applied to large models such as CBRAM. This challenge is assessed and addressed in the subsequent chapters. For that purpose, the so-called mode-space approximation originally derived within the effective mass context was generalized and applied to CBRAM systems. Moreover, the process of obtaining the required mode-space

transformation matrix was largely automatized which drastically simplifies the usage of this approach. Its strength and reliability is demonstrated with the help of a CBRAM configuration whose electrical current was computed both with and without the mode-space approximation. We found that the discrepancy between both models can be kept below 2%. Furthermore, the reduction of the required computational resources was evaluated and we observed that it can be as large as two to three orders of magnitude, depending on the targeted accuracy.

Next, a model for a $Cu/SiO_2/Cu$ CBRAM cell is presented. Its ON-OFF switching behavior was investigated and an estimate of the number of atoms that contribute to the process is provided. Multiple intermediate states were simulated to present the electrical conductance as a function of the number of "dissolved" atoms. The results qualitatively agree with experimental data. Moreover, the electro-thermal properties for CBRAM cells of varying size were determined, leading to an explanation for the experimentally discovered improved performance of ultra-scaled devices. Bringing the memory cell closer to their ballistic transport limit by decreasing their thickness minimizes the interaction between electrons and atomic lattice vibrations. This in turn lowers the lattice temperature within the cell, thereby improving its thermal stability.

In the last chapter of this thesis, the impact of the CBRAM material stack on the electrical current is examined. We observed that for a given atomic configuration of the filamentary structure in the ON-state, the metal of the counter electrode has little influence on the conductance. However, the current density is affected by the choice of the electrode material. This fact is likely due to the asymmetric configuration of typical CBRAM cells, which is challenging to account for in *ab initio* quantum transport. However, to assess the resistance state of a filament, a we found that assuming symmetric metallic electrodes is sufficient. Yet, calculating further properties of the system such as the interaction with the surrounding oxide, requires representing the full asymmetric nature of CBRAM stacks.

# Zusammenfassung

Die zunehmende Beliebtheit von datengestützten Programmen hat die Bedeutung von Datenspeicherung im letzten Jahrzehnt weiter gesteigert. Die Optimierung von Speicherechnologien welche auf komplementären Metall-Oxid-Halbleitern (CMOS) basieren, nähert sich hingegen fundamentalen physikalischen Grenzen. Dadurch eignet sich diese Technologie immer weniger für energiesparende elektronische Speicherkomponenten. Aus diesen Gründen etablieren sich langsam Alternativen zu den herkömmlichen Speicherelementen. Unter diesen neuen, nichtflüchtigen Datenspeichern (NVRAM) verspricht vor allem conductive bridging random access memory (CBRAM) grosse Fortschritte in Bezug auf z.B. Energieverbrauch und Integrationsdichte. CBRAMs sind Metall/Isolator/Metall Heteroübergänge, in welchen metallische Filamente wachsen und sich auflösen. Die An- oder Abwesenheit einer leitenden, metallischen Verbindung zwischen den Elektroden erzeugt unterschiedliche Widerstandszustände. Damit können die logischen Zustände 1 oder 0 kodieren werden. Um jedoch mit herkömmlichen CMOS-Technologien konkurrenzfähig zu werden, müssen mehrere verbleibendende Schwierigkeiten in Bezug auf Variabilität und Zuverlässigkeit von CBRAM beseitigt werden. Das Ausmerzen dieser Probleme wird jedoch dadurch erschwert, dass die Funktionsweise von CBRAM nicht im Detail geklärt ist.

Das Ziel dieser Arbeit war es, mit Quantentransport-Simulationen Erkenntnisse über die Funktionsweise von CBRAM zu gewinnen. Dichtefunktionaltheorie ist ein verbreitetes Verfahren um die elektronische Struktur von Nanokomponenten parameterfrei zu bestimmen. Basierend auf diesen Resultaten konnten strukturelle, elektrische und thermische Eigenschaften von CBRAM-Zellen abgeleitet werden.

Zuerst werden die theoretischen Grundlagen von ab initio Quantentransport präsentiert. Solche Simulationen verursachen einen erheblichen

Rechenaufwand, wenn sie auf grosse Modelle wie zum Beispiel CBRAM angewendet werden. Dies wird im darauffolgenden Kapitel untersucht und angegangen. Dazu wird die mode-space (MS) Methode verallgemeinert um sie auf CBRAM-Systeme anwenden zu können. Zudem wurde der Prozess um die dazu benötigte Transformationsmatrix zu berechnen weitgehend automatisiert, was die Anwendung der Methode erheblich vereinfacht. Die Vorteile und Zuverlässigkeit der Methode wurde anhand des Modells einer CBRAM-Zelle untersucht, indem der elektrische Strom mit und ohne MS-Methode berechnet wurde. Wir sahen, dass die Abweichungen im Resultat unter 2% gehalten werden konnten. Des Weiteren eruierten wir, wie viele Ressourcen mit der MS-Methode eingespart werden konnten und ermittelten eine Reduktion um 2 bis 3 Grössenordnungen, je nach angestrebter Genauigkeit.

Danach wird das Modell einer $Cu/SiO_2/Cu$ CBRAM-Zelle präsentiert. Deren Abschaltvorgang wurde untersucht und eine Abschätzung zur Anzahl Atome die dazu bewegt werden müssen wird gegeben. Mehrere Widerstandszustände zwischen dem Ein- und dem Auszustand wurden simuliert um die Leitfähigkeit als Funktion der Anzahl der diffundierten Atome zu berechnen. Die Resultate stimmen qualitativ mit experimentellen Untersuchungen überein. Zudem wurden elektrothermische Eigenschaften von CBRAM-Zellen verschiedener Grössen bestimmt. Diese Resultate zeigten auf, warum das Verkleinern der Speicherzellen deren Zuverlässigkeit verbessern kann. Je dünner das Speicherelement ist, je weniger interagieren die Elektronen mit den Atomen im Kristallgitter im Herzen des Speichers und geben entsprechen weniger Energie in Form von Wärme an dieses ab. Dies verringert die Betriebstemperatur des Speichers und erhöht damit die thermische Stabilität.

Im letzten Kapitel dieser Arbeit wird der Einfluss der Materialzusammensetzung auf den elektrischen Strom untersucht. Bei einer vorgegebenen Form der atomaren Struktur des Metallfilaments hatte die Wahl des Materials der Elektrode einen vernachlässigbaren Einfluss auf die Leitfähigkeit der Speicherzelle im eingeschalteten Zustand. Deren Stromdichte

hingegen wurde stark vom Metall beeinflusst. Diese asymmetrische Konfiguration von CBRAM-Zellen ist jedoch schwierig in ab initio Simulationen aufzunehmen. Um nur den Widerstand eines Filaments zu ermitteln, eignen sich daher vereinfachte symmetrische Modelle. Für weitergehende Untersuchungen hingegen, muss diese Vereinfachung aufgegeben werden und der asymmetrische Charakter vom CBRAM-Zellen berücksichtigt werden.

Chapter 1

# Introduction

## 1.1 Current Memory Landscape

Data storage is a fundamental component of any information processing system. Present-day memories are based on the complementary metal-oxide-semiconductor (CMOS) technology, e.g. static random-access memory (SRAM), dynamic random-access memory (DRAM), or Flash cells. Because of the different capabilities of these technologies with respect to speed, latency, storage density, and cost, computing architectures rely on hierarchical storage systems. High-density and low-cost Flash drives provide permanent long-term storage, but they are slow to access. DRAM and SRAM units implement temporary, volatile memory, which is faster to access, but at the expense of storage density and cost. The access latency and memory bandwidth limit the data transfer between the storage hierarchy levels and create a speed gap between the storage and computing units. This is known as the so-called "memory wall," which represents a major bottleneck in the current computing architectures [1, 2]. Moreover, the increasing demand for portable electronic devices powered by low-capacity batteries imposes additional hardware requirements. Altogether, these issues have led to a shift of the focus of the semiconductor industry towards low power computing solutions, although they are challenging to implement with CMOS-based memories. Finally, downscaling the feature size of CMOS components has brought this technology close to its fundamental physical limits, hampering further performance

improvements and its cost effectiveness [3].

The limitations encountered by traditional memory implementations have sparked intense research interest for alternative technologies that could complement or replace CMOS-based components. These alternative technologies are collectively termed emerging nonvolatile memories (NVM) [4, 5]. Of particular interest is the concept of storage class memory (SCM), which combines the advantages of low-cost and high density Flash memories with those of fast and low access latency SRAM cells, while offering low power operations. Promising candidates as emerging NVMs for SCM applications include both types of resistive random-access memories (ReRAMs) [6]. To store data, ReRAM relies on the reversible resistance changes of a soft breakdown in a dielectric layer induced by electrochemical effects [7]. This type of NVM cells can be further divided into two categories: valence change memories (VCM) and conductive bridging random access memories (CBRAM), which are at the core of this thesis. ReRAM memory cells can be scaled down to sub-10 nm sizes while keeping long retention times [8], and can be densely integrated [9]. Such memory cells feature other desirable properties, e.g. high erase-write endurances and window margins [10], low operating energies [11], and sub-nanosecond switching speeds [12, 13]. It is, however, important to note that some of these properties have conflicting design requirements such that not all of them can be simultaneously implemented within the same storage unit [14, 15]. Thus, engineering a ReRAM array satisfying all requirements of SCM and being competitive with CMOS-based solutions remains to be demonstrated [6].

A resistive switching behavior has been shown in a large number of material stacks [16]. Each individual configuration comes with its own characteristics such as switching speed, retention time, or turn-on voltage. A huge design space must be explored to obtain a memory cell that optimally fulfills specific application requirements. Moreover, not all aspects of the operating mechanisms in ReRAM are quantitatively understood, the origin of certain effects still being debated [17, 18]. An in-

depth comprehension of the underlying physics is one of the key challenges to address in ReRAM research as it could critically enhance the reliability of these memory cells [19]. Device modeling and computer-aided design can provide valuable insights into the processes governing the ReRAM operation and thereby support on-going experimental activities [20].

An accurate modeling of ReRAMs must cover an enormous time scale and spatial extent. The ON-OFF switching typically occurs within nanoseconds or less and relies on the relocations of single atoms, which can happen within picoseconds. At the same time, the retention capabilities of NVM is measured in years and a memory array is composed of billions of individual cells extending over mm square surfaces. Since no modeling technique can span the enormous time and spatial ranges needed to fully describe ReRAM memory arrays, multiscale approaches are required [20]. *Ab initio* calculations form an integral part of such approaches due to their ability to model atomistic systems in a parameter-free manner. The majority of *ab initio* studies dedicated to ReRAMs focuses on the extraction of relevant physical quantities such as formation energies or diffusion barriers that are used to parameterize macroscopic simulation approaches. On the other hand, the electrical properties of the nanostructures at the core of ReRAMs have attracted little attention [21].

Modeling ReRAM memory cells faces several challenges because they feature complex filamentary structures and material interfaces. For example, the discrete properties of the switching layer play a prominent role. Moreover, there is ample evidence that the atomic configuration of ReRAM memory cells changes from cycle-to-cycle, which creates atomic-scale variability and manifests itself in fluctuations in the resistance measured across the memory cell [22]. Controlling these variations is of paramount importance to implement reliable memories. Therefore, atomistic methods that are able to capture these characteristics should be employed. Another consequence of the atomic scale operation of ReRAM is that electron transport has entered the mesoscopic regime where classical models such as the drift-diffusion equations [23] fail to capture the full

range of the physics at play: quantum mechanical phenomena, e.g. energy quantization, confinement, or tunneling must be taken into account [24]. The presence of quantum mechanical and atomic-scale effects call for an *ab initio* treatment of ReRAMs. Such methods are directly derived from physical principles and thus do not require any input parameters.

The ground-state electronic structure of CBRAM cells can be obtained from density-functional theory (DFT) [25, 26], which is a widespread tool to perform *ab initio* calculations. Out-of-equilibrium properties, e.g. electrical and thermal currents, can be computed using the Non-equilibrium Green's Function (NEGF) formalism [27, 28]. The latter allows for the extraction of transport properties under an external applied bias. Even though ultra-scaled devices operate close to their ballistic limit, many experimental features can only be explained by simultaneously accounting for electrical and thermal effects. This is the case of self-heating, which may be responsible for CBRAM failures at high current densities [29]. Coupled together, DFT and NEGF compose a versatile *ab initio* quantum transport (QT) simulation framework [30, 31].

## 1.2 Conductive Bridging RAM

CBRAM is a (non)volatile memory technology that stores data through a reversible change in the resistance of a dielectric switching layer (SL). The SL is embedded between an oxidizable metal and an electrochemically inert electrode. The lowering of the resistance is induced by the growth of one or several metallic filament(s) through the SL. Resistive switching was first shown with a VCM-type device using titanium oxide as SL in 1968 [32]. In 1976, Y. Hirose et al. demonstrated a CBRAM-like memory effect in a metal-insulator-metal (MIM) stack by growing Ag dendrites through a layer of $As_2S_3$ [33]. The first use of this effect in integrated circuits occurred twenty years later when Swaroop et al. replaced

analogue CMOS components with CBRAM cells to fabricate artificial solid-state synapses [34]. The first implementation of a vertical CBRAM for memory applications was demonstrated in 1999 by Kozicki et al. using Ag and Cu dendrite growth in $As_2S_3$ and $GeSe_2$, respectively [35]. This finding made the CBRAM technology an appealing candidate for data storage in integrated circuits. In 2008, ReRAM cells in general were linked to the theoretical concept of memristor [36], which is an electrical two-terminal component whose resistance depends on the history of the current that flew through it. The memristor is a mathematical construct first hypothesized in 1971 by L. Chua [37].

CBRAMs possess numerous advantages over CMOS-based memory technologies such as Flash, DRAM, and SRAM. First, the resistance state of CBRAM cells can be switched on sub-nanosecond time scales, enabling high speed SET and RESET operations [13]. Second, the ratio between the ON- and OFF-state can typically reach several orders of magnitudes and may reach more than $10^7$ [38]. Third, the size of CBRAM can be scaled down to 10 $nm^2$ areas [39]. Fourth, the relatively simple MIM structure of CBRAM paves the way for high density arrays that can potentially be stacked on top of each other to form 3D structures [40, 41]. Fifth, CBRAM can be operated at low voltages, which is favorable for low power applications [29]. Sixth, the memory state can be retained for long times [42]. Last, CBRAM can be constructed solely of CMOS-compatible materials and directly integrated into standard CMOS processes during the back end of line (BEOL) [43]. Despite these benefits, CBRAM remains an experimental technology and a niche product from an industrial perspective. This is mainly due to the fact that low power and high speed are usually in competition with reliability and retention times, forcing design engineers to make compromises regarding the aspects that should be favored [14, 15]. Nevertheless, in 2014, a chip relying on the CBRAM technology with 16 GB storage capacity was built [44] and the feasibility of automotive grade CBRAM was suggested in 2018 [45].

In addition to electrical operation, the resistance state of CBRAM can

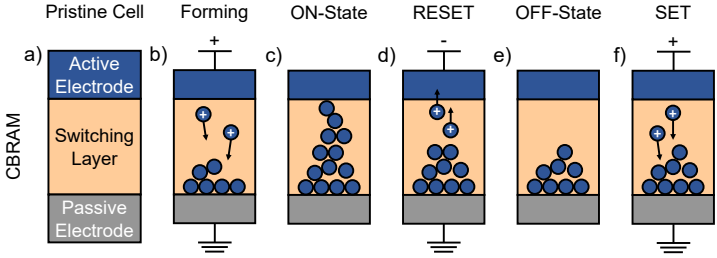


**Figure 1.1:** Illustration of the bipolar filamentary type switching process in a CBRAM cell. The underlying structure is built of an active electrode (often Ag or Cu), an insulating switching layer (e.g. a-$SiO_2$), and an inert metal electrode (a). The application of a forward voltage oxidizes metal atoms at the interface of the active electrode. These ions migrate towards the passive electrode where they are reduced (b). Eventually, a stable metallic filament is formed that bridges the switching layer (c). By reversing the voltage polarity, the filament can be disrupted (d). Finally, the OFF-state with a partial filament left in the oxide layer is reached (e). The subsequent SET process (f) brings the CBRAM cell back to its ON-state (c). Adapted from [50].

also be read or set by other physical stimuli. For example, a CBRAM cell can be coupled to optical signals by inserting it into a plasmonic cavity, enabling the optical readout of the memory state of the cell [46]. Furthermore, CBRAM cells can function as an electro-optical modulator [47] or detector [48]. Besides memory applications, CBRAMs are investigated as active building blocks of neuromorphic circuits, a potential application currently receiving a lot of attention [49]. While the focus of this thesis lies on memory applications, the presented models and simulations apply to CBRAM cells irrespective of their purpose.

### 1.2.1 Working Principle

The memory effect in CBRAM is the consequence of a reversible change in the resistance of the SL [7]. In the simplest picture, CBRAM cells are

two-terminal MIM stacks, with the central dielectric acting as the SL, as shown in Fig. 1.1(a). In contrast to capacitors, which share the same MIM configuration, the dielectric medium changes its resistance upon application of an appropriate voltage between the two electrodes. In the pristine high-resistance state, called OFF-state, the dielectric behaves as an insulator separating the metallic contacts. The low-resistance ON-state is characterized by a soft breakdown of the insulator through the presence of a metallic filament that connects and thereby short-circuits the two electrodes. Multilevel cells can be realized by carefully controlling the length and thickness of the filament [51]. Consequently, the resistance state of the SL, and thus the memory cell, can exhibit different values.

The reversible breakdown of the SL is illustrated in Fig. 1.1(b-f). It is caused by the growth and dissolution of a metallic filament composed of few atoms and extending from one of the electrodes towards the other [52]. Reliable operations require an asymmetric MIM configuration in which the active electrode (AE) is composed of an electrochemically active metal. This chemical species can diffuse into and out of the SL, thus creating the desired ON-OFF resistance switching. The counter electrode (CE) is built from an inert material that does not actively contribute to the filament formation.

During the SET process a positive voltage is applied to the AE causing three principal actions. First, the surface atoms (of type M) of the AE are oxidized

$$\mathrm{M} \longrightarrow \mathrm{M}^{z+} + z\mathrm{e}^{-}, \tag{1.1}$$

where $z$ is the number of electrons with elementary charge $e^-$ that are exchanged during the oxidation and $\mathrm{M}^{z+}$ is an ion dissolved in the SL. Next, under the high electric field induced by the applied voltage, the ions drift through the SL towards the CE. Finally, the $\mathrm{M}^{z+}$ ions are

reduced back to their elementary state

$$M^{z+} + ze^{-} \longrightarrow M \tag{1.2}$$

on the surface of the CE, eventually forming a metallic filament. Upon bridging the SL, the filament short-circuits the memory cell, drastically reduces the voltage drop across the dielectric, and inhibits the oxidation (Eq. 1.1). Thus, the growth of the filament stops.

During the RESET operation, the voltage polarity is reversed such that (i) the oxidation takes place on the surface of the filament instead of the surface of the AE, (ii) the ions drift back towards the AE instead of towards the CE, and (iii) they are reduced on the surface of the AE. The dissolution of the filament into the dielectric eventually disrupts the inter-electrode contact and resets the memory cell into its OFF-state. Because the switching mechanism relies on the oxidation and reduction of metallic species, CBRAMs are also known as electrochemical metallization (ECM) or programmable metallization cells (PMC) in literature [53].

The direction of the growth of the filament during the SET operation, as described above, remains controversial and up for discussion. There is experimental evidence for growth in either direction, towards the CE as well as towards the AE [16]. When growing towards the CE, the reduction reaction of Eq. (1.2) takes place within the dielectric when electrons tunnel from the CE to the $M^{n+}$ ion. The direction of growth supposedly depends on the cation mobility in the SL and the tunneling probability of electrons through the insulating barrier [16]. Other factors such as the device layout likely impact the growth direction as well.

The kinetics of the aforementioned oxidation-drift-reduction processes largely determine the switching characteristics of the memory cell, e.g. operation speed and SET threshold voltage. The stability of the filament and thus the retention time of the ON-state are given by the oxidation rate together with the diffusivity of the $M^{n+}$ ions in the dielectric, which in turn, is closely related to the drift and diffusion properties. Because all

these effects depend on the electrochemical and physical characteristics of the AE, the SL, and the CE, the choice of the material for each of these three regions critically impact the performance of the memory cell.

The **Active Electrode** supplies the ions that build the filament. For low voltage operations, both the oxidation and reductions of its atoms, i.e. Eqs. (1.1-1.2), need to occur with low energy barriers. Therefore, the most popular choices as active materials are Ag and Cu because of their small standard electrode potential [53, 54]. Beside these two elements a number of other materials including Te and Ru have been shown to support CBRAM-type switching. For example, Te improves the filament stability and retention time when compared to Ag or Cu [44, 45, 55, 56]. Ru electrodes are claimed to exhibit more linear switching characteristics than other materials due to the conductive nature of its oxide which makes more intermediate resistance states available between the lowest and highest resistance states [57]. However, both Te and Ru filaments require larger external voltages to be grown and dissolved. Other electrochemically active metals that enable resistive switching include Ni [58], Al [59], Ti [60], Zn [61], Nb [62], or even Au [63]. To favorably affect the chemical environment at the interface between the AE and the SL, the active material in CBRAM cells can be combined with a different metal to form alloys such as AgTe [64] and AgCu [65]. A similar approach consists of creating multi-layer electrodes, e.g. by alternating W and Cu layers [66]. This leads to similar effects as alloying.

The filament grows on top of the **Counter Electrode**. Thus the active $M^{z+}$ ions have to be easily reducible and oxidizable on its surface. The atoms constituting the CE should neither react electrochemically nor dissolve into the SL. Commonly used CE materials are Pt, TiN or W. Moreover, M should alloy with the CE as little as possible to avoid the phenomenon of negative SET. Alternatively, an ion blocking layer on top of the CE can be inserted to prevent a negative SET [67, 68].

Filaments grow through the **Switching Layer** and thereby enable the reversible nonvolatile switching. A wide range of material classes can be

employed as SL, from liquid to solid electrolytes, oxides, nitrides or carbides as well as organic materials [16]. The first CBRAM-type memories featured SL composed of materials known to be solid electrolytes for Ag and Cu such as chalcogenides glasses [69]. Both Ag and Cu have high diffusion in GeSe, for example, which also supports high concentrations of metal ions and favors high switching speeds. The major drawback of such material stacks is their lack of retention times and their low thermal budget during fabrication [70]. As a result, research has shifted its focus towards oxides and their improved insulating properties and better stability of the metallic filament. Longer data retentions are observed in oxide-based CBRAMs. Oxide SLs are the most prevalent ones because they offer the best switching characteristics in terms of ON/OFF ratios, switching speed and endurance [16]. Typical oxides include $SiO_2$, $HfO_2$, $Al_2O_3$, and $Ta_2O_5$. As alternatives a-Si [71], a-SiCN [72], and a-BN [73] have been tested, but with less success. Double-layer SL were proven beneficial at enhancing some of the CBRAM characteristics, in particular the switching window and cycling endurance [72, 74]. These achievements were attributed to a reduced defect generation in the SL as well as a reduced stack degradation.

### 1.2.2 Modeling of CBRAM[1]

Although experimental studies have revealed the switching principle of CBRAMs, the precise mechanisms that control the transition from the OFF- to the ON-state as well as the nature of the conducting path are still under intense investigations [18]. Despite their simple MIM structure, the modeling of CBRAM is a complex endeavor that requires a multiscale approach to capture all of its aspects [75, 76]. Continuum models [77], in which partial differential equations describe the atomic motions (drift and diffusion), can very accurately reproduce and explain experimental data such as the "current vs. voltage" ($I$-$V$) characteristics during a

[1] This subsection is based on [50]

switching cycle [78, 79] or the conductive filament life time [80], at low computational cost. However, their efficiency depends on the availability of a large set of material parameters that must be determined in one way, e.g. from higher-order simulations, or the other, e.g. through fitting. In addition, any information about the actual atomic configuration is lost, which might become an issue when the stochastic relocation of few atoms can change the electronic current by several orders of magnitude. Therefore, atomistic models are needed to highlight the mechanisms underlying the switching behavior of CBRAM cells.

One such example is kinetic Monte Carlo [81] (KMC), a simulation approach that allows to generate atomistic filament structures and to link them to continuum methods [76]. The KMC simulation box is typically discretized into a grid with quadratic tiles representing the atomic positions. The edge length of a square (2-D) or cube (3-D) corresponds to the hopping distance of the filament forming species. In a KMC model, all relevant processes occurring in a CBRAM cell, e.g. oxidation and reduction, adsorption and desorption, nucleation as well as ionic migration within the insulating layer or along interfaces, are described by rate equations obeying an Arrhenius-type behavior [77]. Each rate equation depends on the energy barrier that the specific reaction has to overcome, for instance the activation energy of ionic diffusion or of oxidation. Since the activation energy can be lowered by an external voltage, the formation and dissolution of filaments can be exponentially accelerated by increasing either the applied voltage or the temperature. In a typical KMC simulation, the rate of each individual process is first calculated and stored in a table. At each step of the KMC algorithm, the event to be executed is randomly chosen based on the occurrence probability of the various processes. After each event, the atomic configuration is potentially modified until a stationary state is reached.

KMC has been successfully applied to grow and dissolve filaments with an atomic resolution [82, 83], in excellent agreement with experimental data. Despite valuable insights into the filament dynamics, structures

generated with this method suffer from multiple limitations. First of all, most KMC models are two-dimensional, although three-dimensional implementations have been recently demonstrated [83, 84]. Second, cubic grids have difficulties treating amorphous structures and materials with a non-cubic lattice are only approximately represented. Lastly, the SL, contrary to the filament, is described as a continuum rather than atomic medium. Thus, more advanced models are needed that can enhance the spatial resolution of KMC and better account for the broad range of material properties encountered in CBRAMs.

Classical molecular dynamics (MD) based on force-field (FF) approaches [85] meet these requirements and can capture the detailed atomic structure of both the filament and the insulating layer as well as their dynamics. In such simulations, a parameter set describes the different types of atoms and their interactions. The parameters are fitted to reproduce reference data from experiments, quantum mechanical calculations, or both [86]. The obtained forces are then used to determine the trajectories of the atoms based on Newton's equations of motion. To model the growth and dissolution of a filament through the SL of a CBRAM cell, simulation domains containing thousands of atoms must be constructed [87]. Additionally, time spans of several nanoseconds must be considered to model a full switching cycle [29]. FF-based molecular dynamics achieves that at reasonable computational cost.

Elaborate schemes are needed to construct suitable amorphous structures and interface them with metallic electrodes. For example, a melt-and-quench approach can be used for that purpose [88]. Starting with a chunk of crystalline oxide or randomly placed atoms, MD must be performed for several hundreds of picoseconds at a temperature above the melting point of the oxide. Then, the melt is quenched to 300 K with cooling rates in the order of 15 K/ps [87]. Post-quench annealing at room or slightly elevated temperatures can be beneficial to eliminate coordination defects and reduce the stress inside the amorphous structure.

The first atomistic simulation of a complete CBRAM switching cy-

cle was demonstrated by Onofrio et al. [87] using a so-called reactive FF method. In contrast to traditional FFs, reactive force-fields such as ReaxFF [89] are able to describe the formation and breaking of bonds and therefore to model chemical reactions. A much simpler model of a conductive filament can be obtained by manually inserting metal atoms into the amorphous insulating layer instead of explicitly growing a structure [90]. A shape must be defined and all atoms within it are replaced by metal. The result can be used as starting point for reactive MD under an electric field. However, models relying on continuous rather than localized electric fields have not lead to realistic filament morphologies so far, at least not for complete ON-OFF switching cycles [91].

The parameterization of force-fields is often tailored such that the processes of interest are accurately described, whereas less relevant phenomena are not well accounted for. Therefore, the usage of force-fields to perform MD in complex systems such as CBRAM cells, where many different sub-processes are encountered, can result in misleading behavior. A higher level of accuracy can be achieved by using *ab initio* molecular dynamics (AIMD) where the forces acting on each atom are derived from DFT. The latter is a quantum mechanical modeling method that can describe the electronic structure of any given atomic configuration without the need of fitting parameters. However, the high computational demand of AIMD limits the time range it can access to a few picoseconds and the system size to a few thousand atoms [92]. To benefit from the advantages of FFs and DFT, both methods can be combined. First, atomistic filamentary type CBRAM structures are created by using FF approaches. Then, the structures are relaxed and optimized using AIMD before a variety of physical properties such as the evolution of the electronic density of states (DOS) [92], the activation energy of ion diffusion [93] as well as the nucleus formation energy [94] in CBRAM cells are extracted with the help of DFT. Due to the disordered nature of the structures, calculating meaningful physical properties can only be achieved by averaging over an ensemble of independent measurements [92].

First principles studies have succeeded in deducing a multitude of characteristics and properties of CBRAM and have thereby enabled the parameterization of high-level modeling techniques. The majority of *ab initio* modeling works is concerned with evaluating microscopic quantities such as the charge states of metal interstitials in oxide layers, their diffusion barriers and energy of formation, e.g. Ag in $SiO_2$ [95] or $Ta_2O_5$ [96], which affects the kinetics of the CBRAM ON-OFF switching as well as data retention. Recent studies have also been concerned with the atomic composition of filamentary structures and their properties [96, 97]. Typically, the conductance value of emerging NVM is estimated based on the charge density only. However, only few studies leveraged the power of NEGF in combination with DFT to shed light on the electron transport mechanism in CBRAM [98, 99], VCM [21, 100–103], or phase change memory (PCM) [104, 105].

## 1.3 Motivation and Objective of the Thesis

Detailed understanding of electron transport in CBRAM is crucial to improve their implementation, enhance their performance, and ultimately meet the required specifications for mass production. In light of these challenges, gaining insight into the processes taking place within individual memory cells is vital. The purpose of this thesis is to investigate quantum transport phenomena at the atomic scale in realistically sized, ultra-scaled CBRAM cells. The emphasis is set on the extraction of electrical and electro-thermal properties of structures featuring nanoscale filaments as found in the ON-state of such memories. To reach this goal we combined DFT and NEGF, an equally versatile as efficient modeling toolkit. DFT is the most widespread *ab initio* approach to electronic structure calculations. It is capable of predicting ground-state properties of any material or nanostructure, taking only the initial atomic coordinates as input. This is particularly attractive to describe amorphous

or disordered materials and their interface with metallic electrodes, as encountered in CBRAM. All of which are notoriously difficult to parameterize. The second component of this framework, NEGF, is a popular statistical approach to out-of-equilibrium quantum transport phenomena. It can be interfaced with DFT, thereby benefiting from the parameter-free nature of DFT. However, these first principle based calculations are computational intensive. Approximate schemes must be devised to address this immense challenge and efficiently evaluate the transport properties of nanostructures with an atomic resolution.

The treatment of CBRAM cells with DFT and NEGF provide insight into the properties of nanoscale filaments that cannot be probed experimentally due to their tiny dimensions and their encapsulations within other materials. In modeling, however, the effect of the relocation of individual atoms on electron transport can be traced back and used to elucidate the conductance behavior during the ON-OFF switching process. Moreover, electro-thermal simulations can help identify critical current paths in nanoscale extrusions. The latter can be linked to the reliability of CBRAM cells.

While this work is dedicated to the modeling of CBRAM structures, the developed methodologies can be equally applied to other emerging NVM technologies, as mentioned in the Conclusion chapter.

## 1.4 Outline of the Thesis

After this introduction about CBRAM cells, their functionality, and their modeling, in Chapter 2 the theoretical background of DFT is presented and various atomistic modeling techniques that are employed throughout this thesis are described. Electrical and thermal quantum transport within the NEGF framework is also reviewed and an approximate method to reduce the intensity of quantum transport calculations is outlined.

Chapter 3 is dedicated to the mode-space approximation, which is a

numerical procedure to enable quantum transport calculations of large devices at moderate computational cost. The mode-space approximation is then generalized to inhomogeneous simulation domains and applied to CBRAM systems. The scheme is benchmarked against the original real-space description, with a focus on speed and accuracy.

In Chapter 4 an ultra-scaled Cu/a-$SiO_2$/Cu CBRAM cell is modeled and the results discussed. A technique to implement the dissolution of a filament at the QT level is introduced and the ON-OFF switching of a memory cell is examined with it by extracting the current density at various intermediate stages. The ON-state is further investigated with electro-thermal simulations. Results in the ballistic limit of transport are compared to fully coupled electron and phonon calculations. Finally, the impact of the thickness of the oxide layer on the device stability is revealed.

Chapter 5 focuses on the CE metal, in particular on its influence on the current densities in an Ag filament embedded in $SiO_2$. Multiple metals are used with a fixed filamentary configuration. The characteristics of all created structures are extracted and their implications on the computational model requirements are discussed.

Ultimately, conclusions are drawn and an outlook is given in Chapter 6. The developed simulation environment is now read to explore alternative material stacks, device geometries, and memory types, e.g. VCM or breaking junctions.

Chapter 2

# Quantum Transport from First Principles

## 2.1 Introduction

From "first principles" or "*ab initio*" describes a type of calculation that is directly based on established laws of physics and does not require any empirical inputs nor fitting parameters. When applied to electronic structure investigations, such methods imply that the equations are directly derived from quantum mechanics. This has the obvious advantage no prior knowledge about the properties of the described system is needed, thus precluding any input bias. Furthermore, any kind of system can be treated, whether to reproduce and explain experimental results or to predict a set of properties. The principal challenge posed by *ab initio* electronic structure calculations is their heavy computational burden, which limits the size of ensembles that can be studied. Density-functional theory (DFT) [25, 26] has become the "de facto" standard in electronic structure calculations for solid-state physics due to the computational efficiency of the method. DFT indeed enables the study of larger systems than other approaches.

Electronic structure calculations are predominantly concerned with determining the ground-state properties of either bulk or isolated systems. Quantum transport (QT) simulations, on the other hand, couple the simulation domain to its environment and enable the treatment of

out-of-equilibrium situations. Popular approaches to QT are the Non-equilibrium Green's Function (NEGF) [27, 28] formalism and the quantum transmitting boundary method (QTBM) [106]. Both are statistical approaches to solve the Schrödinger equation in the presence of external stimuli that are treated perturbatively. These techniques require a precise description of the electronic structure of the system of interest in the form of a Hamiltonian matrix. A range of methods have been applied to construct this quantity such as the effective mass approximation (EMA) [107], the k·p [108], or the tight-binding method (TB) [109]. Their restriction to a single point in the Brillouin Zone and/or their reliance on empirical parameterizations limits their predictive capabilities and restricts their use to materials with known properties. A more flexible approach to QT is achieved by combining it with *ab initio* electronic structure calculations [30].

The theory behind these simulations is presented in this chapter as follows: first the electronic structure procedure behind density-functional theory is summarized. Subsequently, the NEGF formalism is outlined and its connection to DFT described. Lastly, the mode-space (MS) approximation for use together with NEGF, is introduced. It can help tackle the computational burden induced by combining NEGF and DFT.

## 2.2 Density-functional Theory

The properties of a stationary quantum mechanical system can be entirely determined by its many-body wave function $\Psi$. The latter obeys the time-independent Schrödinger equation

$$\hat{H}\Psi = E\Psi, \tag{2.1}$$

which is the fundamental quantum mechanical problem that needs to be solved. Here, $E$ is the total energy of the system and the corresponding

Hamiltonian operator $\hat{H}$ can be separated into a nuclear and electronic kinetic energy term $\hat{T}_c$ and $\hat{T}_e$, a nuclear-nuclear and electron-electron interaction term $\hat{U}_c$ and $\hat{U}_n$, and lastly a nuclear-electron interaction term $\hat{U}_{ce}$. In an atomic structure with $N_c$ cores and $N_e$ electrons, the many-body $\hat{H}$ is given by

$$\begin{aligned}\hat{H} = &-\sum_{n=1}^{N_c}\frac{\hbar^2}{2m_n}\nabla^2 - \sum_{i=1}^{N_e}\frac{\hbar^2}{2m_0}\nabla^2 + \\ &\frac{1}{4\pi\epsilon_0}\sum_{n<m}\frac{Z_n Z_m e^2}{|R_n - R_m|} + \frac{1}{4\pi\epsilon_0}\sum_{i<j}\frac{e^2}{|r_i - r_j|} - \\ &\frac{1}{4\pi\epsilon_0}\sum_{n,i}\frac{Z_n e^2}{|R_n - r_i|},\end{aligned} \tag{2.2}$$

where $m_n$ and $m_0$ are the masses of the cores and the rest mass of electrons, respectively, $Z_n$ is the number of protons in atom $n$, $R_n$ and $r_i$ are the positions of atom $n$ and electron $i$, respectively, $\hbar$, $e$, and $\epsilon_0$ are the reduced Planck's constant, the electron charge, and the permittivity of free space. With this Hamiltonian the wave function $\Psi$ becomes a function that depends on $3(N_c + N_e)$ variables, thus resulting in a problem that can only be solved analytically for the simplest systems such as harmonic oscillators or the hydrogen atom. For realistic applications Eq. (2.1) has to be solved numerically.

Even numerically, Eqs. (2.1)-(2.2) remain extremely challenging to handle without further approximations. The most fundamental one is the Born-Oppenheimer approximation, which separates the motion of the atomic nuclei and of the electrons. The difference in masses justifies the assumption that the particles, the core and the electrons, move on different time scales, which implies that the electrons can always relax to their ground state within the time scale the atomic cores need to move. This effectively decouples the wave function of the core and electrons and allows one to describe the total wave function $\Psi$ as the product of the

two separate components, the nuclei $\Psi_c$ and electron $\Psi_e$. In this picture the electrons move freely within the external potential induced by the nuclei. The Hamiltonian for the electrons is then given by

$$\hat{H}_e = \hat{T}_e + \hat{U}_e + \hat{V}, \tag{2.3}$$

where $\hat{T}_e$ and $\hat{U}_e$ keep their definition from Eq. (2.2), whereas the last term $\hat{V}$ describes the external potential induced by the nuclei. Owing to their large mass, the atomic cores are typically treated in a classical manner, as explained in Section 2.2.1. This restricts the quantum mechanical part to the description of the electrons, thereby significantly reducing the computational complexity. Nevertheless, $\Psi_e$ still is a function of $3N_e$ spatial variables.

Multiple methodologies to approximate Eq. (2.3) have been developed with a wide range of computational requirements and accuracy. The gold standard in computational chemistry is set by the coupled-cluster (CC) technique [110], which uses the exponential cluster operator to express the electron correlation. By approximating the exponential with a power series CC offers a systematic approach to converge the wave function. Typically, CCSD(T) which includes first- and second-order terms and treats the third-order term perturbatively, delivers results which agree exceptionally well with experiments [110]. CC, however, scales with $(O^6)$ to $(O^8)$ with respect to the number of electrons, depending on the number of terms included. This fact renders CC impractical for systems beyond a few dozen atoms. Post-Hartree-Fock methods introduce further approximations and reduce the polynomial scaling down to $(O^5)$, though it still limits them to fairly small atomic configurations or molecules. The high polynomial scaling can be avoided altogether by reformulating Eq. (2.1), such that $\Psi_e$ does not have to be calculated. A powerful approach to this endeavor is given by the Hohenberg-Kohn (H-K) theorems, which lay out the foundation of DFT. The latter relies on the electron density $\rho(r)$, which is a function of only three spatial variables. For this reason, DFT

is the most widely used method for electronic structure calculations.

The first H-K theorem states that two systems of electrons, which have the same ground-state density $\rho(r)$, necessarily reside in the same external potential $V(r)$, up to a constant. That is, the external potential and the total energy are unique functionals of the ground-state density. Not only is the electron density determined by the wave function

$$\rho(r) = \int dr_2^3 ... \int dr_{N_e}^3 \Psi_e^*(r, r_2, ..., r_{N_e}) \Psi_e(r, r_2, ..., r_{N_e}), \qquad (2.4)$$

but the reverse is true as well

$$\Psi_0(r_1, ..., r_{N_e}) = \Psi_e[\rho_0(r)] \qquad (2.5)$$

for the ground-state wave function. Consequently, $\rho_0(r)$ determines all observable properties of a system of electrons and it is not necessary to compute the true many-body wave function. This is particularly attractive because $\rho(r)$ only depends on the 3 spatial coordinates $r = x, y, z$ instead of $3N_e$ variables, $r_1, ..., r_{N_e}$, which greatly reduces the computational impact. The second H-K theorem states that the functional $F[\rho(r)]$, which delivers the ground-state energy $E$, returns the lowest possible energy if and only if the electron density is the true ground-state density. This implies that a functional exists such that

$$E[\rho(r)] = F[\rho(r)] + \int dr^3 V(r)\rho(r) \qquad (2.6)$$

where $E[\rho(r)]$ reaches its minimum at the true ground state $\rho_0(r)$. Therefore, a variational principle can be deduced to compute the ground-state density $\rho_0(r)$ of any system where the only input is $V(r)$, which in turn is solely determined by the presence and position of the atomic nuclei. While the H-K theorems prove that the many-body wave function does not need to be computed, they do not provide the functional $F[\rho(r)]$ to obtain $E[\rho(r)]$. The many-body interactions do not disappear, but they

are merely hidden in the functional.

Instead of attempting to solve the many-body Schrödinger equation directly, Kohn and Sham introduced an auxiliary system of non-interacting particles [26]. Each of them obeys the single-particle Schrödinger equation

$$[-\frac{\hbar^2}{2m} + V_s(r)]\psi_i(r) = E_i\psi_i(r). \tag{2.7}$$

While the wave functions $\psi_i(r)$ solving Eq. (2.7) have no physical meaning themselves, they give the correct ground-state density though

$$\rho_0(r) = \sum_{i=1}^{N} \psi_i^*(r)\psi_i(r). \tag{2.8}$$

The non-interacting particles reside in the single-electron effective potential $V_s(r)$, which contains the external potential $V(r)$ as well as the electron-electron interactions:

$$V_s(r) = \int \frac{e^2\rho(r')}{|r-r'|} d^3r' + V_{xc}[\rho(r)] + V(r). \tag{2.9}$$

The first term of Eq. (2.9) is called Hartree potential and contains the Coulomb repulsion between the electrons in a mean-field approach. The computationally challenging many-body interactions are cast into the second term called exchange-correlation potential $V_{xc}$ (xc-functional). To date the exact form of this functional has not been universally determined, but a multitude of approximations exist and the search for improved forms is still ongoing [111, 112]. The two H-K theorems together with the Kohn-Sham equations define the KS-DFT scheme, subsequently referred to as DFT. It should be noted that KS-DFT is not the only approach, but by far the most widespread one [113]. Furthermore, the theory behind DFT is exact. The inaccuracies of modern implementations of DFT originate from the approximations made to render the equations computationally tractable, predominantly those applied to $V_{xc}$, sometimes termed

density-functional approximation (DFA). Since the entire computational procedure is solely derived from theory and does not require empirical inputs, DFT is a first principle or *ab initio* method. For historical reasons the second denomination (*ab initio*) is often reserved to wave function based methods. The two terms are used interchangeably in this work.

Numerous forms of xc-functionals have been proposed. They are typically categorized according to the level of approximation in the so-called Jacob's ladder [114]. On the lowest rung of the ladder one finds local-density approximation (LDA) functionals. These are derived from the homogeneous electron gas model and assume the charge density to be locally constant to model the exchange and Coulomb correlations. Despite its crude level of approximation, the LDA functional is rather successful at describing many properties of metallic compounds. Its success can be attributed to the fact that errors in the exchange and correlation energies tend to cancel out. For semiconductors or insulators as well as molecular systems the accuracy of LDA greatly varies. The second rung of Jacob's ladder is made of the large family of generalized-gradient approximation (GGA) functionals. In addition to the magnitude of the local density GGA also includes its gradient to estimate the correlation effects. Among the GGA-implementations the formulation of Perdew-Burke-Ernzerhof (PBE) [115] is consistently the most popular. While not the most accurate functional it produces reasonable results for most systems. Among the shortcomings of LDA and GGA, the most significant one for *ab initio* QT is that the band gap of semiconductors and insulators is severely underestimated. Meta-GGA functionals implement the next level of theory over GGA by adding a kinetic term to the functional, related to the second derivative of the density [116]. Hybrid functionals constitute the fourth rung of Jacob's ladder. They mark a huge step upwards, both in terms of accuracy as well as computational complexity. Most hybrid functionals rely on GGA where the exchange term is improved by adding a fraction of the exact Hartree-Fock (HF) exchange [117]. Many properties such as the band gap and binding energies are

much more accurately predicted by hybrid functionals as compared to GGA or LDA. This enhancement comes at the prize of worse polynomial scaling and far larger memory requirements, rendering the use of hybrid functionals impractical for systems containing many atoms. In addition to mixing HF and GGA exchanges into the functional, second-order perturbative Coulomb correlation can be introduced [118]. These double hybrid functionals can provide high accuracy at the cost of even higher computation cost. Because of this they are restricted to niche application, but are slowly gaining in popularity.

Independent of Jacob's ladder, functionals can also be grouped by their composition. One such group is purely or mostly derived from theory with zero or few fitting parameters such as the PBE functional. The latter is often chosen to minimize the error of various properties over a large range of different compounds or molecules. The second group consists of functionals that are based on a number of parameters, which are fitted to accurately reproduce specific properties of certain materials such as B3LYP [119], which is tailored for organic molecules. Both forms have their respective advantages: fitted functionals are often more accurate when correctly employed, i.e. on compounds that are similar to the ones for which the xc-functional was parameterized. PBE-type functionals, on the other hand, are typically less accurate than the fitted ones, but have the advantage of being applicable to a wide range of compounds, i.e. they are more versatile [120].

### 2.2.1 Molecular Dynamics

Molecular dynamics (MD) is a simulation approach to investigate the movements of atoms and probe the dynamic evolution of an ensemble. The considered particles are allowed to interact for a certain period of time by numerically integrating Newton's equations of motion. MD can be combined with DFT as a consequence of the Born-Oppenheimer approximation, which separates the treatment of electrons and nuclei.

During the electronic structure calculation, the atoms are frozen in place and are only accounted for through the external potential in Eq. (2.3). Once the ground-state electron density is obtained, the forces acting on the atoms can be derived as the negative gradient of the potential energy with respect to the atom positions. Because the potential energy surface is obtained from first principles this type of simulation is called *ab initio* molecular dynamics (AIMD) or alternatively Born-Oppenheimer molecular dynamics (BOMD).

Different ensembles can be defined within MD. The simplest form is the microcanonical ensemble (NVE), where the number of particles (N), the volume (V), and the total energy (E) of the system are conserved. The particles can freely exchange potential for kinetic energy and vice-versa. The instantaneous temperature T of the ensemble is defined by equating the kinetic energy of the atoms to $3/2k_BT$, where $k_B$ is Boltzmann's constant. The estimated temperature can change over time when the system is undergoing exo- or endothermic reactions. Alternatively, a canonical ensemble (NVT) can be defined, where the temperature instead of the total energy is conserved. A thermostat couples the finite simulation domain to a bath which absorbs (injects) kinetic energy from (into) the system, thereby regulating the instantaneous temperature. Additionally, the pressure tensor can be derived from the electron density. By applying a barostat to the AIMD simulation, the pressure (P) within the ensemble can be controlled and kept constant. Deploying both a thermostat and a barostat to an AIMD simulation results in an isothermal-isobaric (NPT) ensemble.

The time step employed to integrate the equations of motions should be kept small enough to capture all atomic movements The shortest time scale to consider typically is the phonon oscillation period. Therefore, time steps in the order of femtoseconds are required. Because of the large computational burden induced by DFT and the short time steps required to capture all relevant motions, AIMD simulations are typically restricted to dozens of picoseconds and several 100's to few 1000's of atoms.

As an alternative, instead of integrating the forces over time, the atom positions can be optimized such that the force on each atom vanishes. This process of driving the ensemble towards a local minimum in the potential energy surface is referred to as *geometry optimization* throughout this thesis. By including the pressure tensor in the optimization process, the lattice parameter of the unit cell under consideration can be optimized as well. Together with the optimization of the forces this process is termed *cell optimization*.

## 2.3 Quantum Transport[1]

The NEGF formalism offers a powerful framework to calculate the non-equilibrium properties of quantum mechanical systems [121]. It is widely used to perform quantum transport (QT) simulations [122, 123]. This approach to non-equilibrium statistical mechanics is based on the works of Kadanoff and Baym [27] and Keldysh [28]. It requires a description of the electronic structure of the system under study in the form of a Hamiltonian matrix. Approaches relying on first principles concepts, where the Hamiltonian matrix is obtained from DFT calculations, are known as *ab initio* QT [30, 124, 125]. The coupling of NEGF and DFT was introduced by Lang [124], where the representation of the device was based on DFT and the electrodes modeled using a jellium approximation. Fully atomistic simulations were proposed by Taylor et al. [30] featuring an atomistic representation based on DFT for both the device region and the contacts. Since these pioneering works, several packages capable of treating quantum transport from first principles have been developed [126–131]. Some of them are freely, others commercially available.

The majority of DFT+NEGF calculations are performed in the ballistic limit of transport, where the energy of each particle is conserved throughout the simulation domain. The effect of inelastic interactions

[1] This section is based on [50]

can naturally be incorporated in NEGF through the use of scattering self-energies [132]. Besides pure electrical or thermal transport, coupled electro-thermal simulations can be perturbatively carried out through the self-consistent Born-approximation (SCBA) [31, 133]. Owing to the large computational burden induced by the SCBA, such simulations are typically restricted to small systems or require large computational resources. Furthermore, calculating the phonon properties and the electron-phonon coupling from first principles is a challenging task [134, 135]. Nevertheless, such calculations have been applied to a wide range of nanoscale devices going from 2-D field-effect transistors (FETs) [134], FinFETs [135], or CBRAM cells [136] to the modeling of inelastic electron-tunneling spectroscopy (IETS) in molecular junctions [137, 138].

It should be emphasized that the SCBA is not the only possibility to account for electro-thermal effects. Lowest order-expansion techniques have been used as well. They have a lower computational burden, but at the cost of additional approximations [137, 139]. Due to its perturbative nature, SCBA may fail to converge in the presence of strong electron-phonon coupling or other scattering mechanisms representing strong interactions. An exact but also computationally more expensive techniques capable of treating such systems is the hierarchical equations of motion [140].

This section is dedicated to the description of electro-thermal QT calculations with a focus on DFT-based electronic structure representations within the SCBA. The following subsections describe the coupling of DFT with NEGF for the case of electrical and thermal transport. Lastly, the coupling of electrons and phonons via scattering self-energies is explained.

### 2.3.1 Electron Transport

As described in the previous section, DFT is a powerful method to calculate the electronic structure of atomic systems. Thus, the most advanced

electron transport frameworks typically rely on the ground-state Kohn-Sham Hamiltonians [26] calculated with DFT [30, 125]. To derive the equations governing QT the stationary Schrödinger equation Eq. (2.1) is a convenient starting point. Inserting the KS-Hamiltonian and left-multiplying Eq. (2.1) with $< \Psi|$ gives rise to the Hamiltonian and overlap matrices

$$H =< \Psi|\hat{H}_{KS}|\Psi > \text{ and } S =< \Psi|\Psi > \tag{2.10}$$

that can be used to rewrite Eq. (2.1) in matrix form as

$$H \cdot \psi(E) = ES \cdot \psi(E), \tag{2.11}$$

where $\psi(E)$ is a vector related to the wave function that still needs to be determined. Bloch's theorem [141] implies that the wave function $|\Psi >$ and the matrices $H$ and $S$ depend on an additional quantity $k$ called wave vector. As the atomic configurations considered in this thesis are typically large, measuring multiple nanometer along each direction, the $k$-dependency of $|\Psi >$ can be safely neglected. This simplifies the notation and reduces our analysis to $\Gamma$-point calculations where $k = 0$.

The computational efficiency of solving matrix equations can be directly related to their number of nonzero elements and their sparsity pattern. To maximize the sparsity of $H$ and $S$ a suitable basis must be selected to expand $|\Psi >$. Localized basis sets such as Gaussian-type orbitals (GTO) [142] are ideal for that purpose as they produce sparsely populated, banded matrices. In contrast, plane-waves, which are popular for electronic structure calculations, lead to dense matrices. While not impossible [143, 144], the use of plane-waves is relatively seldom in quantum transport calculations, in particular for large systems. If necessary, plane-waves can still be localized, with the help of, e.g., Wannier functions [145]. When employing a localized basis set, $|\Psi >$ becomes

$$|\Psi > = \sum_n \sum_{l(n)} c_n^l(E)\phi_n^l(r - r_n). \tag{2.12}$$

The valence electrons of atom $n$ are expanded in $l(n)$ basis functions $\phi_n^l$ centered at position $r_n$. The number of basis functions per atom $l$ may vary for different chemical species. The $c_n^l(E)$'s are the occupation coefficients of their respective basis function. The wave function is then the sum over all individual basis functions of each atom weighted by the occupation factor. With this choice of basis expansion $\psi(E)$ becomes a vector containing all coefficients $c_n^l(E)$ and the size of the matrices $H$ and $S$ is the sum of $l(n)$ over all atoms. If the $\phi_n^l$ are orthogonal to each other, $S$ is the identity matrix and Eq. (2.11) becomes a regular eigenvalue problem (EVP). In most cases the overlap between localized basis functions is nonzero and a generalized EVP must be solved.

It should be noted that the $H$ and $S$ matrices are not unique to DFT. They can also be created based on other methods such as tight-binding (TB) [109], where Löwdin orbitals are parameterized to reproduce experimentally measured or DFT band structures. The Hamiltonian is not necessarily an atomistic quantity either, it could be expressed in the effective mass approximation (EMA) [107] on a finite difference or finite element discretized grid. The transport equations presented in the next paragraphs apply to Hamiltonian matrices obtained with all these methods. The continuum nature of EMA and the required TB parameterization, however, make both methods ill-suited to deal with disordered structures. The major difficulty in TB models lies in the derivation of a parameter set that accurately captures both amorphous phases or defects such as vacancies and interstitials as well as interfaces between different materials.

In contrast to electronic structure calculations, which are typically restricted to the ground-state of a system, electrons in device simulations must be able to enter and leave an open domain so that a non-equilibrium current can flow. External potentials are applied to contact regions to drive a device out of equilibrium. As a consequence, the boundary conditions applied to Eq. (2.11) require special attention. Whereas periodic (PBC) or closed boundary conditions (CBC) are typically used in DFT,

open boundary conditions (OBC) [146] are at the core of quantum transport investigations. In OBCs the contacts and the device region are first treated separately, as illustrated in Fig. 2.1(a). Each contact is modeled as a semi-infinite lead in thermal equilibrium, with a flat electrostatic potential, as schematized in Fig. 2.1(b). It should be represented by at least two identical blocks of atoms. These blocks correspond to the first and last unit cell of the central (device) region. The leads serve as launching pads or collectors for electrons, phonons or other particles. They are connected to the device through so-called retarded boundary self-energies $\Sigma^{R,B}(E)$ that must be introduced into Eq. (2.11). Additionally, an injection vector $I_{nj}(E)$ acts as a source term to model the incoming electrons. OBC are not limited to two-terminal systems, but can be readily generalized to structures with multiple leads [146]. Techniques to compute $\Sigma^{R,B}(E)$ and $I_{nj}(E)$ can be found in Ref. [147].

In the presence of OBC Eq. (2.11) takes the following form

$$(E\ S - H - \Sigma^{R,B}(E)) \cdot \psi(E) = I_{nj}(E). \tag{2.13}$$

This equation must be solved for all discrete energies belonging to the interval of interest, which extends over the Fermi energy of both contacts. In ballistic and coherent simulations, i.e. in the absence of scattering with energy relaxation, all $E$'s are independent of each other and Eq. (2.13) directly yields $\psi(E)$. Such an approach is known as the QTBM [106]. It is computationally attractive as it involves the solution of sparse linear systems of equations with multiple right-hand-sides, but it does not lend itself naturally to the simulation of dissipative transport.

Alternatively, Eq. (2.13) can be rewritten in terms of Green's Functions (GF) as

$$(E\ S - H - \Sigma^{R,B}(E)) \cdot G^{R}(E) = I, \tag{2.14}$$

$$G^{\lessgtr}(E) = G^{R}(E) \cdot \Sigma^{\lessgtr,B}(E) \cdot G^{A}(E), \tag{2.15}$$

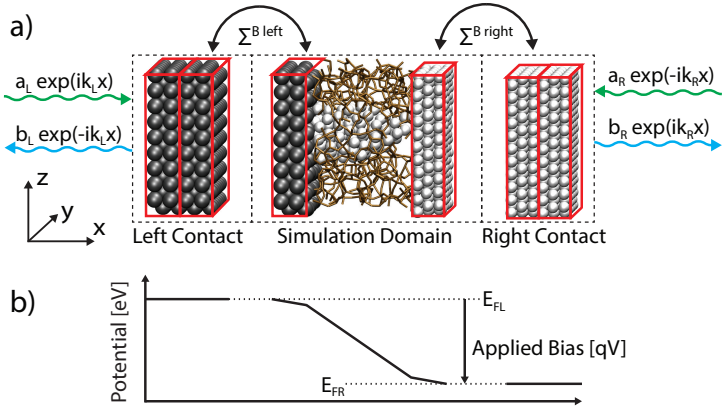


**Figure 2.1:** (a) Illustration of a CBRAM atomic simulation domain and its division into three regions. The open boundary conditions are calculated in the leads based on a plane-wave ansatz. Incoming (green) and outgoing (blue) waves are considered. The incoming waves inject electrons into the simulation domain with a probability $a_L$ (from left) and $a_R$ (from right). The outgoing waves encompass the transmitted and reflected electrons with amplitudes $b_R$ and $b_L$, respectively. The leads are coupled to the simulation domain through the boundary self-energies $\Sigma^B$. (b) Average electrostatic potential of a typical CBRAM along the x-axis. The potential in the leads is kept constant so that plane-waves are the exact solution of the Schrödinger equation in these regions. The potential difference between the left ($E_{FL}$) and right ($E_{FR}$) Fermi energy is equal to the externally applied voltage $V$ times the elementary charge $q$.

with $I$ being the identity matrix of appropriate size. In Eqs. (2.14-2.15) the GFs are of different flavors, namely retarded ($G^R$), advanced ($G^A$), lesser ($G^<$), and greater ($G^>$) and the influence of $I_{nj}$ is indirectly accounted for in the lesser and greater boundary self-energies $\Sigma^{\lessgtr,B}(E)$. The advanced GF is the hermitian transposed of the retarded GF, i.e. $G^A = G^{R\dagger}$. The wave function $\psi(E)$ can be related to $G^R(E)$ through

$$\psi(E) = G^R(E) \cdot I_{nj}(E). \tag{2.16}$$

Eqs. (2.14-2.15) are known as the NEGF formalism [27]. Intuitively, the lesser (greater) boundary self-energy, $\Sigma^{B,\lessgtr}(E)$, indicates the probability that a state gets filled ($\Sigma^{B<}(E)$) or emptied ($\Sigma^{B>}(E)$) through interactions with the contact. The off-diagonal elements of the lesser and greater GF, $G^{\lessgtr}(E)$, describe the correlation between the involved basis functions, whereas the diagonal entries contain the probability that a state $n$ is occupied ($G^<_{nn}(E)$) or unoccupied ($G^>_{nn}(E)$). Therefore, it is not necessary to convert back the results of Eqs. (2.14-2.15) to a wave function $\psi(E)$. All observable quantities such as the charge density $\rho(r)$ at position $r$ and the electrical current $I_d$ can be directly derived from selected entries of $G^{\lessgtr}(E)$. The latter can be computed efficiently with an iterative procedure called recursive GFs (RGF) algorithm [148]. Nevertheless, for ballistic simulations this approach is computationally much more expensive than QTBM. The strength of NEGF comes from its natural integration of scattering mechanisms through self-energies, as will be introduced in Section 2.3.3.

Regardless of the transport type the charge density $\rho(r)$ can be computed as

$$\rho(r) = -i \sum_{m,n} \sum_{k,l} \int \frac{dE}{2\pi} \, \phi_m^{k\,*}(r - r_m) G^{<kl}_{mn}(E) \phi_n^l(r - r_n), \tag{2.17}$$

where the $(k, l)$ indices refer to orbital types and the $(m, n)$ ones to

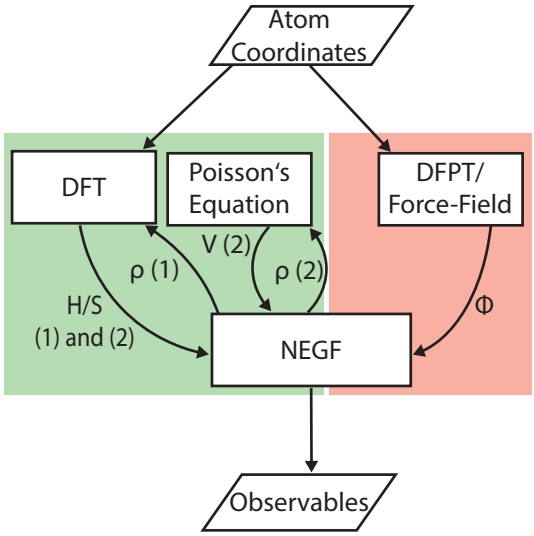


**Figure 2.2:** Flow chart illustrating the interactions between the different methods used to perform DFT+NEGF simulations. The data flow for the section covering electrons is shown in green, the one for thermal transport in red. Electron transport does not require any input beside the coordinates of the atoms. For a given atomic system relaxed with DFT the Hamiltonian $H$ and overlap $S$ matrices are passed to NEGF. Subsequently, either the NEGF-DFT (1) or the NEGF-Poisson (2) loop is executed until the charge density $\rho$ is converged. Lastly, observables such as the electrical and energy currents are computed. The path for thermal transport has no feedback loop because the dynamical matrix $\Phi$ is assumed not to depend on any out-of-equilibrium quantity. In addition to the atomic coordinates a parameterization of the forces is needed in classical approaches to obtain the dynamical matrix. This requirement is superfluous if the dynamical matrix is calculated at the *ab inito* level.

atomic position. If the spread of the basis functions is very narrow and the orbitals orthogonal, the expression for $\rho(r)$ can be simplified to

$$\rho(r) = -i \sum_l \sum_n \int \frac{dE}{2\pi} G^{<ll}_{nn}(E)\delta(r - r_n). \tag{2.18}$$

Injecting electrons into the device domain drives it out of equilibrium, which changes the distribution of electrons and modifies the charge density $\rho(r)$. This in turn affects the KS-Hamiltonian $H_{KS}$ through the first two terms in the effective potential $V_{eff}(r)$ in Eq. (2.9), giving rise to a mutual dependence of Eq. (2.7) and Eqs. (2.14-2.15). It must be resolved in a self-consistent manner until $\rho(r)$ is converged [30].

Fully coupled NEGF+DFT simulations come with a heavy computational burden, even in the ballistic case. This can be somewhat alleviated by assuming that the charge density only affects the electron-electron repulsions and by neglecting the change of exchange and correlation. By solving Poisson's equation

$$\nabla^2 V_{pot}(r) = -\frac{\rho(r)}{\epsilon(r)}, \tag{2.19}$$

where $\epsilon(r)$ represents the position-dependent dielectric function, the electrostatic potential $V_{pot}(r)$ is obtained. Instead of recomputing the $H$ matrix with Eqs. (2.7-2.10), the influence of $V_{pot}$ can be directly incorporated in Eq. (2.14) by assuming that

$$V^{kl}_{mn} = \int d^3r\, \phi^k_m(r-r_m) V_{pot}(r) \phi^l_n(r-r_n) \approx S^{kl}_{mn} \frac{V_{pot}(r_m) + V_{pot}(r_n)}{2}, \tag{2.20}$$

and hence, $H^{kl}_{mn} \to H^{kl}_{mn} + V^{kl}_{mn}$ [147]. As in the original NEGF+DFT scheme, Eqs. (2.14), (2.15) and (2.19) must be solved self-consistently in an NEGF-Poisson loop, but this second approach is computationally advantageous. The organization of the original and simplified method is illustrated in the left part of Fig. 2.2. Both procedures rely on the

DFT calculation of the Hamiltonian $H_{KS}$, after which either the DFT or Poisson feedback loop is executed. If QTBM is deployed instead of NEGF the same dependence arises between $H$, $S$, $V_{pot}$, and $\rho(r)$.

After convergence of the selected self-consistent loop, physical observables can be extracted. In ballistic simulations the function $T(E)$ describes the transmission probability of an electron from the left to the right side of an open system (or vice-versa) at energy $E$ [132]. It can be calculated according to

$$T(E) = \Gamma_L(E)\ G^A(E)\ \Gamma_R(E)\ G^R(E). \tag{2.21}$$

The broadening function $\Gamma_C(E)$ of contact $C$ depends on the retarded boundary self-energy $\Sigma_C^{B,R}(E)$ and is defined as

$$\Gamma_C(E) = i(\Sigma_C^{B,R}(E) - \Sigma_C^{B,R\dagger}(E)), \text{ with } C = R \text{ or } L. \tag{2.22}$$

In Eq. (2.22) $i$ is the imaginary unit and $^\dagger$ denotes the hermitian transpose operator. The current is conveniently obtained from $T(E)$ through the Landauer-Büttiker formula [149]

$$I_d = -\frac{e}{\hbar}\int \frac{dE}{2\pi} T(E)\ (f_L(E) - f_R(E)), \tag{2.23}$$

where $f_L(E)$ ($f_R(E)$) is the Fermi distribution function of the left (right) contact, $e$ the electron charge, and $\hbar$ Planck's reduced constant. The electrical current can also be directly calculated from the GF with [132]

$$I_d = \frac{e}{\hbar}\sum_{m,n}\sum_{k,l}\int \frac{dE}{2\pi}(H_{mn}^{kl}G_{nm}^{<lk}(E) - G_{mn}^{<kl}(E)H_{nm}^{lk}) \tag{2.24}$$

where the subscripts $m$ and $n$ denote two atoms situated in two consecutive slabs (unit cells) of the simulated structure and $(k, l)$ refers to the corresponding basis indices. The $H_{mn}^{kl}$ off-diagonal entries connect the orbital $k$ on atom $m$ with the orbital $l$ on atom $n$. In analogy with

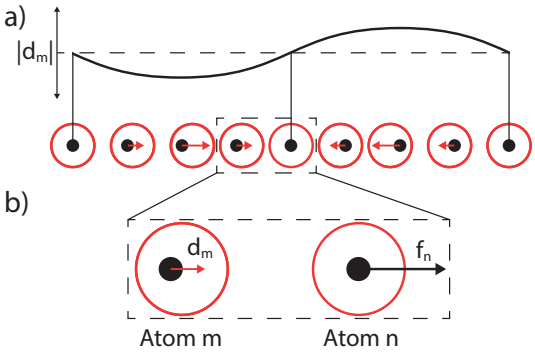


**Figure 2.3:** (a) Visualization of a phonon wave. The black dots represent a 1-D chain of atoms in their equilibrium position. The larger red circles around the atoms help visualize their current displacement, which is marked by the horizontal arrow. The first, middle, and last atom remain at their equilibrium position. The curve above the atoms illustrates the envelope of the phonon wave function. The magnitude of the atomic displacement is proportional to it. (b) Illustration of the interplay between the displacement of an atom and the force it induces on a neighbor atom. Atom $m$ is moved by the vector $d_m$ towards atom $n$. Consequently, the latter feels a repulsive force $f_n$ even though it remains at its lattice site.

the electrical current, the energy current carried by electrons is given by [150]

$$I_{dE,el} = \frac{1}{\hbar} \sum_{m,n} \sum_{k,l} \int \frac{dE}{2\pi} E \, (H^{kl}_{mn} G^{<lk}_{nm}(E) - G^{<kl}_{mn}(E) H^{lk}_{nm}). \quad (2.25)$$

This formulation of the electrical and energy currents, $I_d$ and $I_{dE,el}$ is more general and holds even when no transmission function can be defined, i.e. in the presence of a dissipative scattering mechanism.

### 2.3.2 Thermal Transport

Thermal transport at the nanoscale is conveniently modeled through the propagation of phonons [123]. Ballistic phonon transport can be formulated in the QTBM and NEGF formalisms, as demonstrated for electrons in the previous section. While QTBM is more efficient to solve ballistic problems, NEGF is required to model electro-thermal interactions and account for self-heating effects. Therefore, only the NEGF equations are shown here, for the sake of brevity.

Phonons are quasi-particles that arise from the coupled motion of atoms around their equilibrium position. [107]. An illustration of such a wave of coupled atomic motions is given in Fig. 2.3(a) for a 1-D wire. What is represented is an excited state of the lattice whose amplitude is related to the crystal temperature. To mathematically describe phonons, the total energy $E_{Tot}$ of a perturbed atomic system with equilibrium energy $E_0$ should be considered. The displacement of atom $m$ along the cartesian direction $i$ is labeled $d_m^i$. In the harmonic approximation $E_{Tot}$ is expanded in a Taylor series up to the second order of the displacement

$$\begin{aligned} E_{Tot} = E_0 + \sum_m \sum_i \frac{\partial E_{Tot}}{\partial d_m^i} d_m^i + \\ \frac{1}{2} \sum_{m,n} \sum_{i,j} d_m^i F_{mn}^{ij} d_n^j + O^3(d_m^i),\ i,j = x,y,z. \end{aligned} \tag{2.26}$$

In an equilibrium configuration the system resides in a (local) minimum so that $dE_{Tot}/d_m^i$, the force acting on atom $m$ in direction $i$, vanishes from Eq. (2.26). The second-order term

$$F_{mn}^{ij} = \frac{\partial^2 E_{Tot}}{\partial d_m^i\ \partial d_n^j} \tag{2.27}$$

is the force constant matrix and is made of the second derivative of the total energy with respect to the displacements of atoms $m$ and $n$. With

the knowledge of the force constant matrix and by applying Newton's classical equation of motions, the displacement of each atom from its equilibrium position, $\mu(r,t)$, can be computed. Applying PBC or CBC and assuming that the atoms oscillate with a frequency $\omega$, we end up in the stationary regime with the following eigenvalue problem to solve for $\mu(\omega)$ [151]

$$\Phi \cdot \mu(\omega) = \omega^2 \mu(\omega). \tag{2.28}$$

Here, $\Phi$ is the dynamical matrix with entries $\Phi^{ij}_{mn} = -F^{ij}_{mn}/\sqrt{M_m M_n}$, $M_{m/n}$ being the mass of atom $m/n$, and $\mu(\omega)$ the phonon wave function or polarization vector. The energy of a phonon is related to its frequency through $E = \hbar\omega$. As Hamiltonian matrices may be $k$-dependent quantities, $\Phi$ may also depend on the phonon wave vector $q$. Here, this dependence is neglected for the same reasons given for electrons and only the $\Gamma$-point is considered.

Eq. (2.28) is the phonon equivalent to Eq. (2.11) for electrons. To derive the thermal NEGF equations OBC are introduced into Eq. (2.28). They can be constructed following the same prescriptions as for electrons and illustrated in Fig. 2.1(a). Their calculation takes either the form of an eigenvalue problem [152] or of a complex contour integral [153]. By incorporating the resulting phonon boundary self-energies $\Pi^B$ into Eq. (2.28) and by transforming the wave function expressions into GFs, we obtain the following system of equations to solve:

$$(\omega^2 I - \Phi - \Pi^{R,B}(\omega)) \cdot D^R(\omega) = I \tag{2.29}$$

$$D^{\lessgtr}(\omega) = D^R(\omega) \cdot \Pi^{\lessgtr,B}(\omega) \cdot D^A(\omega). \tag{2.30}$$

The quantities in the equations above are the phonon GFs ($D^R$, $D^A$ and $D^{\lessgtr}$) and the boundary self-energies ($\Pi^{R,B}$ and $\Pi^{\lessgtr,B}$). The labeling conventions for retarded, advanced, lesser, and advanced remain the same as in the previous section. Eqs. (2.29-2.30) must be solved for all phonon frequencies of interest. The phonon density and current are then derived

from $D^<$ using similar expressions as for the electrons. Besides, the energy current carried by phonons can be computed as

$$I_{dE,ph} = \sum_{mn}\sum_{ij}\int \frac{d\omega}{2\pi}\hbar\omega\ (\Phi^{ij}_{mn}D^{<ji}_{nm}(\omega) - D^{<ij}_{mn}(\omega)\Phi^{ji}_{nm}), \quad (2.31)$$

where $m$ and $n$ refer to atoms situated in two adjacent slabs and $(i, j)$ to the cartesian coordinates x,y, or z [154].

While Eqs. (2.29-2.30) have the same form as the equations for electrons, there is an important difference. Namely, Eq. (2.28) does not depend on the phonon population, therefore eliminating the need for an iterative solution process and reducing the computational cost as compared to electrons. The procedure involved in atomistic thermal transport simulations is depicted in the right-hand side of Fig. 2.2.

Whereas DFT has become the most widely used method for electronic structure calculations, even in large systems, competing approaches exist to generate the dynamical matrix $\Phi$ in Eq. (2.28) [155]. It can be directly produced from the total energy of a given system with density-functional perturbation theory (DFPT) [156]. Alternatively, it can be observed that the $F^{ij}_{mn}$'s in Eq. (2.27) correspond to the first derivative of the forces acting on each atom. These elements can therefore be calculated from finite differences through the frozen phonon scheme [151]. Both DFPT and frozen phonons induce a large computational burden that makes them impractical for large atomic systems, if performed at the *ab initio* level.

The frozen phonon approach is the method of choice for large, disordered systems because of its relatively low computational complexity as compared to DFPT. It relies on the evaluation of the first derivative of the force $f^i_m = dE_{Tot}/dr^i_m$, which is typically evaluated using forward or central differences [137, 151]

$$F^{ij}_{mn} \approx -\frac{f^{+i}_m}{d^j_n} \approx -\frac{f^{+i}_m - f^{-i}_m}{2d^j_n}. \quad (2.32)$$

In Eq. (2.32) $f_m^{\pm i}$ is the force acting on atom $m$ along the cartesian coordinate $i$. In equilibrium it is zero, but upon displacing atom $n$ along $\pm j$ by a distance $d_n^j$ it becomes finite, as illustrated in Fig. 2.3(b). To compute the entire dynamical matrix each atom must be displaced individually three (six) times when using forward (central) differences. This results in $3N_{atom}$ ($6N_{atom}$) configurations to simulate and for each of them a force evaluation must be performed. Finally, the $3N_{atom} \times 3N_{atom}$ dynamical matrix $\Phi$ is obtained, which is computationally very expensive if evaluated at the *ab initio* level. This fact is particularly relevant when modeling large systems such as realistic CBRAMs where thousands of atoms are involved. As force-field parameterizations are available that produce reasonably accurate forces and lattice dynamics for a large number of atom combinations, employing such a classical approach is an attractive compromise between accuracy adn computational cost.

### 2.3.3 Electro-thermal Coupling

Ballistic electron and phonon transport simulations, as introduced in the two previous subsections, provide valuable insights into a large variety of device operation regimes. To offer a comprehensive picture under any bias condition and to investigate certain failure mechanisms such as temperature-induced breakdowns, coupled electrical and thermal simulations are required. The electron-phonon (e-ph) interactions can take different forms, e.g. electron scattering on deformation potentials [157] or scattering on polar-optical phonons through the Fröhlich interaction [158]. The computational framework to couple electron and phonon transport is the same in all cases, the difference coming from the electron-phonon coupling elements. Subsequently, a description of scattering on deformation potentials is given.

To couple the electron and phonon populations, the energy of the fermionic and bosonic system must be considered. It can be described by

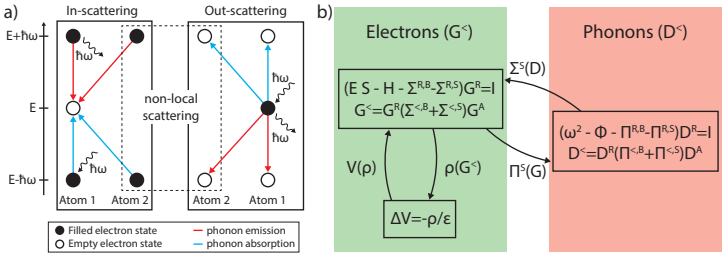


**Figure 2.4:** (a) Scattering events leading to a change in the energy of an electron when it interacts with a phonon. These events can be divided into two categories, in- and out-scattering. In the in-scattering case an empty state at energy $E$ is filled by an electron at energy $E \pm \hbar\omega$ by emission or absorption of a phonon with energy $\hbar\omega$. Out-scattering describes the situation where an occupied state is emptied to a state at $E \mp \hbar\omega$ by emitting or absorbing a phonon with energy $\hbar\omega$. Both in- and out-scattering can be local or nonlocal events. In the former case the involved electron remains located on the same atom, whereas in the latter it may relocate to a different place. (b) Coupled electron-phonon system of equations within NEGF. The coupling is included perturbatively through the scattering self-energies $\Sigma^S$ and $\Pi^S$ that must be solved self-consistently with the GFs. This process is known as the self-consistent Born approximation (SCBA).

the total Hamiltonian

$$H_{tot} = H + H_{ph-kinetic} + H_{ph-harmonic} + H_{e-ph}. \tag{2.33}$$

The first term corresponds to the electron Hamiltonian from Eq. (2.11), while the second and third ones are captured by the dynamical matrix $\Phi$. The last Hamiltonian contains the interaction between electrons and phonons. It is treated perturbatively and cast into the scattering self-energies $\Sigma^{TS}$ and $\Pi^{TS}$ of type $T \in \{R, A, <, >\}$ for electrons and phonons, respectively. The lesser and greater components can be written as [150]

$$\Sigma^{\lessgtr,S}(E) = i\hbar \sum_{i,j} \int \frac{d\omega}{2\pi} \nabla^i H \; G^{\lessgtr}(E - \hbar\omega) \; \nabla^j H \; D^{\lessgtr ij}(\omega) \tag{2.34}$$

and

$$\Pi^{\lessgtr,S \; ij}(\omega) = -i \int \frac{dE}{2\pi} tr\{\nabla^i H \; G^{\lessgtr}(E + \hbar\omega) \; \nabla^j H \; G^{\gtrless}(E)\}. \tag{2.35}$$

In Eqs. (2.34-2.35), all $G^{\lessgtr}$, $\Sigma^{\lessgtr}$, and $\nabla H$ blocks are matrices of size $N_{orb} \times N_{orb}$ and the summation over neighbor atoms are omitted for brevity. The superscripts $i$ and $j$ denote the entries in $\nabla H$, the phonon GFs, and self-energies corresponding to the cartesian coordinates $i$ and $j \in \{x, y, z\}$. The strength of the electron-phonon coupling is determined by $\nabla^i H$, which represents the derivative of the electron Hamiltonian with respect to the displacement of the atoms along the direction $i$. It thus couples the lattice dynamics created by the phonons to its electronic response. The retarded scattering self-energies, $\Sigma^{R,S}$ and $\Pi^{R,S}$, can be derived from $\Sigma^{\lessgtr,S}$ and $\Pi^{\lessgtr,S}$. Very often, their real part is neglected for simplicity [132].

To give an intuitive interpretation of Eqs. (2.34-2.35), we first recall that the diagonal elements of the lesser GFs, $G^{<}(E)$ and $D^{<}(\omega)$, indicate whether a state at energy $E$ is occupied by an electron and the number

of phonons that occupy a state at energy $\hbar\omega$. The same elements of the greater GFs, $G^{>}(E)$ and $D^{>}(\omega)$, determine whether an electronic state at energy $E$ is unoccupied or the number of free phonon states at energy $\hbar\omega$. A specific transition is only possible if an (un-) occupied electron state is available at an energy $\hbar\omega$ above or below the state of interest, as illustrated in Fig. 2.4(a). If a scattering event is allowed, the likelihood of in-scattering, i.e. an empty state at energy $E$, $G^{>}(E)$, gets filled is proportional to the lesser scattering self-energy $\Sigma^{<}(E)$. Such a process can happen through either phonon emission or absorption. An electron at energy $E \pm \hbar\omega$ ($G^{<}(E \pm \hbar\omega)$) emits (+) or absorbs (-) a phonon with energy $\hbar\omega$ ($D^{>}(\omega)$ for emission, $D^{<}(\omega)$ for absorption) and changes its energy to $E$. The out-scattering probability is given by $\Sigma^{>}(E)$. An occupied state at energy $E$, $G^{<}(E)$, gets emptied to $E \mp \hbar\omega$ ($G^{>}(E \mp \hbar\omega)$) by the emission (-) or absorption (+) of a phonon ($D^{>}(\omega)$ and $D^{<}(\omega)$). A similar interpretation can be made for the phonon scattering self-energies $\Pi^{\lessgtr}(\omega)$. They refer to the probabilities that an unoccupied ($D^{>}(\omega)$) or free ($D^{<}(\omega)$) state gets filled ($\Pi^{<}(\omega)$) or emptied ($\Pi^{>}(\omega)$) when an electron transitions from one state to the other through phonon emission or absorption.

The diagonal entries of the scattering self-energies describe local interactions, that is, the electron remains on the same atom during the process. The off-diagonal elements, on the other hand, account for nonlocal transitions where the electron does not only change its energy, but also its position in real-space. Nonlocal scattering events are numerically difficult to handle. In our approach they are neglected for electrons and only close-neighbor interactions within a predefined cutoff radius are taken into account for phonons. This is necessary to ensure energy conservation in our NEGF calculations. By scaling the strength of the local entries of $\Sigma^{\lessgtr}$, the influence of the nonlocal events can be indirectly modeled [159].

The derivative of the Hamiltonian matrix in Eq. (2.10) with respect

to a perturbation in real-space, $\partial r$, is given by

$$\nabla H = \frac{\partial < \Psi|\hat{H}_{KS}|\Psi >}{\partial r} - < \frac{\partial \Psi}{\partial r}|\hat{H}_{KS}|\Psi > - < \Psi|\hat{H}_{KS}|\frac{\Psi}{\partial r} > . \tag{2.36}$$

The last two terms appear because the basis changes with perturbation [137]. The expression in Eq. (2.36) can be computed at various levels of accuracy. The simplest picture considers the derivative of the Hamiltonian with respect to bond stretching between two neighboring atoms [160]. This scheme can be expanded to include more harmonic terms, e.g. bond angles. An alternative approach, similar to the calculation of the dynamical matrix, helps to determine $\nabla H$ with respect to the displacement of each individual atom. This is the most accurate method as it does not rely on any assumption regarding the nature of the bonds or angles connecting two atoms, but it comes at the price of increased computational cost. While the bond stretching technique only requires to perform one additional ground-state DFT calculation under hydrostatic strain, typically 0.01% to 0.1%, the second method is even more expensive than computing the dynamical matrix from first principle. Because of this, the bond stretching approach is the preferred one for large systems.

The chosen hydrostatic stress is applied to the atomic configuration by increasing the simulation box, thus stretching the bond $r_{mn}$ between atoms $m$ and $n$ without affecting the angle between them. If the change in basis functions induced by the stress is small the last two terms in Eq. (2.36) vanish. The strained Hamiltonian $H^s$ is then computed and the derivative with respect to a change in the bond length evaluated as

$$\frac{\partial H_{mn}}{\partial r_{mn}} \approx \frac{H^s_{mn} - H_{mn}}{\Delta r_{mn}}, \tag{2.37}$$

where $\Delta r_{mn}$ is the bond length variation due to the applied strain [160]. The $\nabla^i H_{mn}$ entries are obtained by projecting the derivative of $H_{mn}$

onto the cartesian coordinate $i$ with

$$\nabla^i H_{mn} = \frac{\partial H_{mn}}{\partial r_{mn}} \frac{r^i_{mn}}{|r_{mn}|}. \tag{2.38}$$

Here, $|r_{mn}|$ is the length of the bond between atoms $m$ and $n$, $r^i_{mn}$ its signed component along $i$.

When the scattering self-energies are included in NEGF the equations for electrons become [150]

$$(E\ S - H - [\Sigma^{R,B} + \Sigma^{R,S}]) \cdot G^R = I \tag{2.39}$$

$$G^{\lessgtr} = G^R \cdot [\Sigma^{\lessgtr,B} + \Sigma^{\lessgtr,S}(G^{\lessgtr}, D^{\lessgtr})] \cdot G^A, \tag{2.40}$$

with

$$\Sigma^R \approx \frac{i}{2}(\Sigma^> - \Sigma^<). \tag{2.41}$$

For phonons we have

$$(\omega^2 I - \Phi - [\Pi^{R,B} + \Pi^{R,S}]) \cdot D^R = I \tag{2.42}$$

$$D^{\lessgtr} = D^R(\omega) \cdot [\Pi^{\lessgtr,B} + \Pi^{\lessgtr,S}(G^{\lessgtr}, G^{\gtrless})] \cdot D^A, \tag{2.43}$$

and

$$\Pi^R \approx \frac{i}{2}(\Pi^> - \Pi^<). \tag{2.44}$$

The dependence of the GFs and self-energies on the energy $E$ and frequency $\omega$ has been dropped out in the above equations and substituted by $D^{\lessgtr}$ and $G^{\lessgtr}$ for the scattering self-energies to emphasize the interplay between the electron and phonon populations. Eqs. (2.39-2.40) now depend on Eqs. (2.42-2.43) and vice versa. These two sets of equations must be solved iteratively until convergence is reached. The fulfillment of this property can be verified by looking at the electrical current, Eq. (2.24), and the sum of the electronic and thermal energy currents, Eqs. (2.25) and (2.31). Both quantities have to be conserved along the transport

axis of the investigated device, when the GFs do not vary anymore. This process is known as the self-consistent Born-approximation (SCBA). The system of equations to be tackled is depicted in Fig. 2.4(b).

After converging the electron and phonon densities, physical quantities can be extracted. In addition to the currents that are given by Eqs. (2.24) and (2.31), the lattice temperature is of particular interest to quantify the effect of self-heating. Different possibilities exist to assign a local temperature to individual atoms [150]. In the so-called population approach the effective temperature $T_n^{eff}$ of atom $n$ is adjusted such that the Bose-Einstein distribution reproduces the phonon population of each individual atom, $N_n^{eff}$. The phonon density is derived from the GF

$$N_n^{eff} = i \int \frac{d\omega}{\pi} \omega \; tr[D_{nn}^<], \tag{2.45}$$

where $D_{nn}^<$ is the matrix block of size $3 \times 3$ corresponding to atom $n$. The temperature $T_n^{eff}$ is a fitting parameter that is adjusted so that the value of $N_n^{eff}$ can be reproduced with

$$N_n^{eff} = \int \frac{d\omega}{\pi} \omega \; N_{Bose}(\hbar\omega, T_n^{eff}) DOS_n(\omega). \tag{2.46}$$

Here $N_{Bose}$ is the Bose-Einstein distribution and $DOS_n(\omega)$ the frequency-resolved phonon density-of-states of atom $n$, which is proportional to the difference between the diagonal elements of $D^<$ and $D^>$.

## 2.4 Mode-Space Approximation[2]

The computational intensity of NEGF calculations can be decreased by reducing the size of the Hamiltonian matrix in Eqs. (2.14)-(2.15) or (2.13) through the use of the mode-space (MS) approximation [162–164]. This technique decomposes the real-space (RS) domain into two directions,

[2] This section is based on [161]

the transport and the transverse one, the two being orthogonal to each other. The MS approximation exploits the fact that only few transverse modes are required to accurately reproduce the electronic structure in a limited energy window [162]. In the case of transport, only states around the Fermi energy contribute to the current, which makes the MS approximation ideally suited for quantum transport. Based on the transverse modes of a single block of the device structure a transformation matrix $U$ can be computed. Subsequently, the Hamiltonian and overlap matrices are converted into MS with $H_{MS} = U^\dagger H_{RS} U$ and $S_{MS} = U^\dagger S_{RS} U$. On the down side, such approaches introduce unphyiscal energies into the band structure. A procedure to obtain an accurate MS representation and free of unphysical bands and called equivalent model, has recently been developed [163] for orthogonal basis sets. It has been extended to the nonorthogonal case [164], which is relevant for *ab initio* QT and outlined in this section.

The energy-momentum dispersion of a RS Hamiltonian with periodicity along the $x$ direction can be computed from the following generalized eigenvalue problem:

$$H_{k_x} \Psi_{k_x} = E(k_x) S_{k_x} \Psi_{k_x}, \tag{2.47}$$

where $H_{k_x}$ and $S_{k_x}$ are the $k_x$-dependent Hamiltonian and overlap matrices, respectively, while $E(k_x)$ is the energy of the transverse mode $\Psi_{k_x}$ at momentum $k_x$. The matrix $H_{k_x}$ is given by

$$H_{k_x} = H_0 + \sum_{n=1}^{N_N} \left( H_n e^{ink_x} + H_n^\dagger e^{-ink_x} \right). \tag{2.48}$$

Here, $H_0$ is the on-site and diagonal RS Hamiltonian block corresponding to a unit cell of width $L_x$. It is connected to $2N_N$ neighboring cells situated at a distance $\pm nL_x$ from it, through the off-diagonal RS Hamiltonian blocks $H_n$ and $H_n^\dagger$. Note that the overlap matrix $S_{k_x}$, which is

equal to the identity matrix in case of an orthogonal basis set, has the same form as $H_{k_x}$.

The initial guess $U_0$ for the transformation matrix $U$ is constructed from selected $\Psi_{k_x}$'s. These are obtained from solving Eq. (2.47) for all energies within a predefined window with lower and upper limits $E_1$ and $E_2$, and for a set of $n_k$ $k_x$-points selected according to

$$k_{x_j} = 2\pi j/(n_k - 1) - \pi,\ j = 0, ..., n_k - 1. \tag{2.49}$$

Overall, for each $k_x$-point, a total of $m(k_x)$ modes $\Psi(k_x)$ is calculated, which satisfy $E_1 \leq E(k_x) \leq E_2$. A basis

$$\begin{aligned}\Psi(n_k) = [&\Psi_{k_{x1},1}(E),\ ...,\ \Psi_{k_{x1},m(k_{x1})}(E),\\ &\Psi_{k_{x2},1}(E),\ ...,\ \Psi_{k_{xn_k},m(k_{n_k})}(E)]\end{aligned} \tag{2.50}$$

is formed by grouping all these modes. It is orthonormalized using singular value decomposition (SVD). To reduce the size of the MS basis when building $U_0$, only those left-singular vectors corresponding to the largest singular values (SVs) are retained. All the others are discarded.

With this initial guess for $U$ Eq. (2.47) is transformed into MS according to

$$[U^\dagger H_{k_x} U]\Psi_{k_x} = E(k_x)[U^\dagger S_{k_x} U]\Psi_{k_x}. \tag{2.51}$$

By applying $U_0$ to Eq. (2.51), the energies obtained from Eq. (2.47) are reproduced accurately between the bounds $E_1$ and $E_2$. Unfortunately, apart from the real energies, spurious unphysical states are introduced into the band structure. They originate from the compression of $U$ through the SVD. To eliminate these spurious energies and obtain the final, physically valid transformation matrix $U$, a refinement procedure must be applied. It consists of iteratively adding carefully selected modes $\psi$ to $U_0$. The latter shift the unphysical energies below $E_1$ or above $E_2$, until no unphysical energies remain between $E_1$ and $E_2$. To ensure that the real energies are preserved, the additional $\psi$'s are required to be orthogonal

to the modes responsible for the real energies. Finding the necessary $\psi$'s is done by minimizing the functional $\mathcal{F}(C)$ [163, 164] of a vector $C$:

$$\mathcal{F}(C) = \frac{1}{n_z}\sum_{q=1}^{n_q}\sum_{j=1}^{n_z}\frac{C^T A(k_q, z_j)C}{C^T B(k_q, z_j)C} + (C^T C - 1)^2, \tag{2.52}$$

where the $k_q$ is a set of trial k-points and the $z_j$ points are situated on a complex contour around the energy $(E_1 + E_2)/2$ [163]. The matrices $A$ and $B$ are defined as [164]

$$\begin{aligned} A(k_q, z_j) = (z_j - \epsilon_c)\Xi^\dagger(\Lambda U\widetilde{\Lambda}^{-1}U^\dagger S_{k_q} U\widetilde{\Lambda}^{-1}U^\dagger\Lambda - \\ S_{k_q}^\dagger U\widetilde{\Lambda}^{-1}U^\dagger\Lambda - \Lambda U\widetilde{\Lambda}^{-1}U^\dagger S_{k_q} + S_{k_q})\Xi, \end{aligned} \tag{2.53}$$

$$B(k_q, z_j) = \Xi^\dagger(\Lambda - \Lambda U\widetilde{\Lambda}^{-1}U^\dagger\Lambda)\Xi. \tag{2.54}$$

Here, $\Xi$ is a trial basis to be chosen later, while $\epsilon_c$ and $z_j$ as well as the matrices $\Lambda$ and $\widetilde{\Lambda}$ obey the following equations:

$$\epsilon_c = \frac{E_1 + E_2}{2}, \tag{2.55}$$

$$z_j = (\epsilon_c + \frac{E_2 - E_1}{2})e^{2\pi i(j-1/2)/n_z}, \tag{2.56}$$

$$\Lambda(k_q, z_j) = z_j S_{k_q} - H_{k_q}, \tag{2.57}$$

$$\widetilde{\Lambda}(k_q, z_j) = U^\dagger\Lambda U. \tag{2.58}$$

From the $C$ that minimizes Eq. (2.52) the new modes $\psi$ can be computed as $\psi = \Xi C$ and added to the transformation matrix $U$. This operation typically shifts one unphysical band out of the energy window of interest. Solving Eq. (2.52) has to be repeated until no spurious energies remain.

As the matrices $A$ and $B$ appear only in quadratic form, i.e. $C^T AC$ and $C^T BC$, in Eq. (2.52), they can be symmetrized without loss of generality. This trick is found to improve the stability of the minimization

and it simplifies the form of the analytical derivative of $\mathcal{F}(C)$ to

$$\frac{\partial \mathcal{F}(C)}{\partial C} = \sum_q \sum_z \left[ \frac{C^T A(q,z)}{C^T B(q,z) C} - \frac{C^T A(q,z) C}{(C^T B(q,z) C)^2} C^T B(q,z) \right] + 4(C^T C - 1)C. \tag{2.59}$$

With this form of $\frac{\partial \mathcal{F}(C)}{\partial C}$ and the use of a derivative-based minimizer such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [165–168], the calculation of $C$ that minimizes $\mathcal{F}$ can be considerably accelerated.

Once a transformation matrix that returns a clean MS band structure is obtained, each block of the RS Hamiltonian and overlap matrices can be transformed individually according to

$$\tilde{H}_i = U^\dagger H_i U, \;\; \tilde{S}_i = U^\dagger S_i U, \;\; i = 0, 1, \tag{2.60}$$

where the tilde ($\sim$) denotes quantities represented in MS. The matrices $\tilde{H}$ and $\tilde{S}$ can finally be used to solve Eqs. (2.14-2.15) or (2.13).

The NEGF-Poisson loop requires an expression for the evaluation of the charge in RS. The charge density in MS is calculated from

$$\tilde{\rho}_{nn} = \tilde{G}_{i,i-1} \tilde{S}_1 + \tilde{G}_{i,i} \tilde{S}_0 + \tilde{G}_{i,i+1} \tilde{S}_1^\dagger. \tag{2.61}$$

The MS density matrix is then transformation from MS to RS

$$\rho = U \tilde{\rho}_i U^\dagger. \tag{2.62}$$

Remaining observables such as the device current or the transmission can be computed directly in MS using the equations given in Section 2.3. Analogous to the back transformation of the density matrix, expressions for the scattering self-energies in Eqs. (2.34-2.35) could potentially be

derived. This, however, has not been demonstrated to date and goes beyond the scope of this thesis.

Chapter 3

# Application of the Mode-Space Approximation to CBRAM Cells[1]

## 3.1 Introduction

As the thickness of CBRAM switching layers is shrunk and the active switching volume approaches the nanometer scale, it becomes essential to model the memory cells with an atomistic resolution. Modeling CBRAM at the atomic scale requires a description of many elements in a wide range of chemical environments, rendering the use of methods relying on empirically fitted parameters extremely challenging. The use of parameter-free tools elegantly circumvents this problem. Therefore, DFT-based NEGF calculations are the method of choice for QT from first principles, as explained in Chapter 2. Owing to their immense computational burden such calculations are typically limited to a small number of atoms, in the range of few hundreds to thousands [122, 169]. One approach to solve this challenge are massively parallel calculations [170]. The required computational resources, however, render this approach unfeasible for large-scale investigations, particularly for a technology that relies on the relocation of atoms and thus needs different Hamiltonian representations throughout the calculation of its $I$-$V$ characteristics.

Alternatively, the high computational intensity can be mitigated by methods suchs low-rank approximation [171], model order reduction [172],

[1] This chapter is based on [161]

or MS approximation [162], which was introduced in Section 2.4. They are all similar in that they construct a much smaller basis to represent the electronic structure of the device. Hence, these techniques are appealing because of their ability to reduce the computational burden by several orders of magnitude while keeping the errors of the computed device current and charge density within a few percent. Originally developed for orthogonal basis sets such as tight binding (TB) [162, 163], the MS scheme has recently been expanded to nonorthogonal basis sets [164], which are often employed in DFT. A major disadvantage of the MS approximation is that the band structure of the investigated device must be available. In effect, to apply an MS transformation it is necessary that all unit cells building the device domain are periodic replica of a representative cell so that the overall band structure is uniquely defined. As a consequence, interfaces and defective, disordered, or amorphous materials cannot be readily transformed into MS. Such features and layers are, however, at the core of CBRAM cells as well as many other realistic applications, e.g. structures with surface roughness [173], grain boundaries and interfaces in interconnects [174, 175], and tunnel and break junctions [176, 177].

The seemingly incompatible modeling of CBRAMs and application of the MS transformation can be reconciled by dividing the device structure into multiple regions. The OBC, used in the NEGF formalism to inject or collect particles, rely on the same principles as the MS transformation, periodicity along the transport direction. The semi-infinite leads that couple the device domain to its surroundings are made of replica of the same cell. Moreover, specific parts of the device could also consist of repeated unit cells. This fact can be capitalized on to apply a MS transformation to the periodic regions of the CBRAM and to the OBC. Here, we propose a hybrid MS-RS scheme which locally transforms suitable regions into MS, whereas the nonperiodic features along the transport direction remain expressed in RS [178]. Such an approach combines the advantages of the parameter-free first principles description of the CBRAM cell via the DFT framework with those of the computational efficiency of the MS

transformation.

## 3.2 Atomic Device Structure

To explain the RS-to-MS transformation and its hybrid coupling and to demonstrate the applicability of the scheme, the CBRAM cell illustrated in Fig. 3.1 is used. The cell is composed of two Cu contacts surrounding a layer of amorphous silicon dioxide (a-$SiO_2$). A Cu filament bridges the switching layer (SL) and short-circuits the electrodes, putting the CBRAM cell into its ON-state. The length of the oxide $L_{ox}$ is 1.6 nm and the size of the cross-section 2.1×2.3 $nm^2$. The resulting device is schematically illustrated in Fig. 3.1(a). The contacts are in fact longer than shown in Fig. 3.1(a) and comprise seven unit cells of Cu each. The total number of atoms in the contacts and oxide is 3390. The CBRAM cell is examined in detail in Chapter 4. Here, it only serves as a benchmarking structure for the hybrid MS-RS scheme.

The unit cells delimited by the black boxes in Fig. 3.1(b) are perfectly periodic. The corresponding Hamiltonian and overlap matrices have a block structure. Taking advantage of it, an MS basis can be created for the contact extensions of the device. The RS-to-MS transformation is performed on one of these Cu unit cell, which is composed of 240 atoms with 25 orbitals each. The Hamiltonian and overlap matrices were computed with the CP2K package [179] using the Perdew-Burke-Ernzerhof (PBE) [115] exchange-correlation functional, a double zeta-valence polarized (DZVP) Gaussian-type orbital (GTO) basis set [180], and Goedecker-Teter-Hutter (GTH) pseudopotentials [181].

## 3.3 Parameterization of the Equivalent Model

The efficiency and accuracy of the algorithm presented in Chapter 2.4 sensitively rely on a certain number of parameters. Selecting them re-

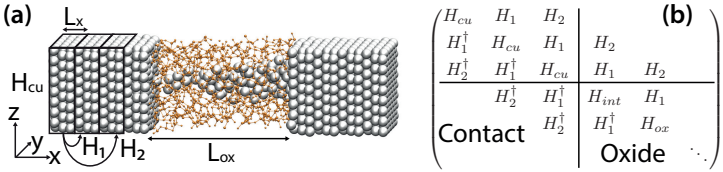


**Figure 3.1:** (a) Atomistic representation of the Cu/a-$SiO_2$/Cu CBRAM cell structure studied in this chapter with the help of the proposed hybrid mode-space – real-space approach. The structure is composed of 3390 atoms, with the large white spheres representing Cu atoms (25 orbitals per atom), the orange lines representing bonds between Si and O atoms (both with 13 orbitals per atom). The black rectangles mark the periodic Cu contact blocks. (b) Block pentadiagonal structure of the real-space Hamiltonian matrix $H$ corresponding to the structure in (a). Each block represents a unit cell (three Cu layers). Because of long-range interactions, each unit cell is connected to $N_N$=2 neighboring cells.

quires extensive tests and good intuition, particularly for large metallic cells as encountered in the contacts of CBRAM. To reduce the need for expensive parameter sweeps we developed a scheme to select most parameters automatically and provide simple guidelines to choose the reminder of them. This automatization has been achieved through the development of metrics to measure the convergence of several characteristics. The RS-to-MS transformation then boils down to setting appropriate thresholds for the metrics. The automation of the process the is key to make the use of the MS approximation worthwhile for large unit cells. The guidelines to select the parameters for transforming RS Hamiltonian blocks corresponding to metallic cells to their MS equivalents are outlined here. It is important to note that the strategy for the selection of the parameters does not necessarily return the most optimal parameter. Possibly, there might exist a combination that leads to a solution with lower computational effort. The ones given here, however, reliably lead to a clean MS band structure without unphysical states in a entirely

automatized manner.

Eqs. (2.47)-(2.60) describe the mathematical procedure to obtain the MS transformation matrix. Six parameters must be carefully selected for the process to be successful and to complete within a reasonable computational time. Namely, (1) the value of $n_k$ in Eq (2.49) to calculate the RS band structure, (2) the number of modes $N_M$ that compose the initial guess for the transformation matrix $U_0$, (3) the trial k-point set $k_q$, which is used during the refinement procedure, (4) the energy integration points $z_j$, (5) the initial guess for $C$ in Eq. (2.52), and (6) the trial basis $\Xi$. Parameters (1) and (2) determine both the accuracy of the MS band structure and the number of unphysical bands present in the initial guess. Parameters (3)-(6) on the other hand are critical to control the refinement procedure and to ensure fast convergence.

The k-point sampling is chosen based on the SVD of the basis of transverse modes $\Psi(n_k)$ in Eq. (2.50). The grid should be fine enough to ensure that the singular values (SV) are converged with respect to $n_k$. This convergence is quantified by the quotient of the SV of two $\Psi(n_k)$. Let $s_{1,i}$ and $s_{2,i}$ be the sorted SV of two bases $\Psi(n_{k,1})$ and $\Psi(n_{k,2})$, where $n_{k,2} > n_{k,2}$. The normalized ratio of each pair of SV has to be greater than a predefined threshold $q_{min}$

$$(\min_i\{\frac{s_{1,i}}{s_{2,i}}\})\frac{s_{2,1}}{s_{1,1}} > q_{min}. \tag{3.1}$$

The second ratio on the left-hand-side, $\frac{s_{2,1}}{s_{1,1}}$, is needed to normalize the SV's because those originating from two different bases can be offset by a constant value. This concept of convergence is illustrated in Fig. 3.2(c), where the $\frac{s_{1,i}}{s_{2,i}}$ are plotted for several pairs of $\{n_{k,1}, n_{k,2}\}$ for a rectangular Cu cell with 240 atoms. The horizontal dashed line markes $q_{min} = 0.6$ the designated threshold, indicating that $n_k \geq 96$ are a robust choice, whereas any lower value might cause convergence issues. To minimize the computational effort, the lowest value satisfying Eq. (3.1) should be selected. Please note that the cutoff of $q_{min} = 0.6$ is an empirical value

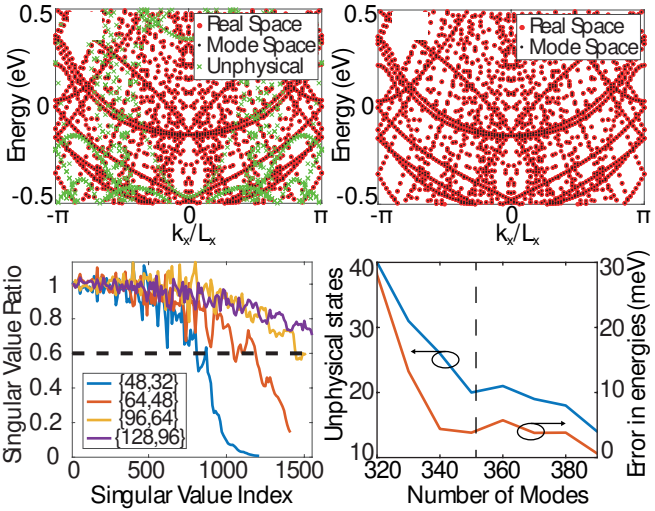


**Figure 3.2:** (a) Band structure of a Cu cell containing 240 atoms as computed in RS (large red dots) and MS (small black dots). The unphysical energies coming from the initial MS transformation are shown as green crosses. (b) Same as (a), but after applying the basis refinement of Eq. (2.52). No spurious, unphysical energy remains. (c) Ratio of the singular values resulting from the orthonormalization of several pairs of bases $\Psi(n_k)$ corresponding to a Cu cell made of 240 atoms. (d) Maximum number of unphysical energy levels at a single k-point (left axis) when the RS and MS bands are compared as a function of the number of initial modes $N_M$ composing $U_0$. Reprinted from [161]

that served well to transform the metallic band structure of Cu. It might have to be readjusted for different materials. Moreover, $\frac{s_{1,i}}{s_{1,1}} < 10^{-11}$ are discarded due to numerical noise.

To find the number of left-singular vectors $N_M$ to include in $U_0$, the MS band structure must be computed as a function of $N_M$ from Eq. (2.51) using $U_0(N_M)$ as the transformation matrix $U$ and compared against the RS results obtained from Eq. (2.47). The initial basis $U_0(N_M)$ must be large enough to reproduce the real energies. Further increasing $N_M$ reduces the error in the physical energies computed in MS. Hence, $N_M$ should be large enough such that the maximum error $\Delta_E$, measured as the difference between the RS and physical MS energies, decreases to below a threshold value which was empirically determined to be 5 meV. To identify the physical MS energies, each RS energy level is mapped to the closest MS energy value. Consequently, the MS energies without a corresponding RS level are deemed unphysical. The number of unphysical bands, $N_s$, i.e. the number of spurious energies that are introduced by the MS transformation, follows the same trend as $\Delta_E$ and decreases as $N_M$ increases. Thereby, the number of refinement iterations needed to clean the MS band structure is reduced. The dependence of $\Delta_E$ and $N_s$ on $N_M$ is illustrated in Fig. 3.2(d). The smallest $N_M$ enabling a reliable refinement procedure is marked by the vertical dashed black line. Increasing $N_M$ beyond the minimum value satisfying the aforementioned conditions improves the accuracy of the final $U$, but at the cost of its size and ultimately computational time. The impact of $N_M$ on the result of the refinement procedure is summarized in Table 3.1. A suitable initial guess for $N_M$ corresponds to the number of normalized SVs, $\frac{s_{1,i}}{s_{1,1}}$, that are larger than $10^{-2}$.

Selecting the proper set of trial k-points $k_q$ has a major influence on the refinement procedure, regarding both the convergence and computational resources. A large $n_q$ is needed for a comprehensive sampling of the k-space, when $k_q$ is selected according to Eq. (2.49), ensuring that all unphysical bands are sampled. To overcome this limitation, Huang

**Table 3.1:** Comparison of the efficiency of the real-space-to-mode-space transformation as a function of the size of $U_0$ for a rectangular Cu cell with 240 atoms. The energy window covers a range of $\pm 0.5$ eV around the Fermi energy. A total of $n_k = 96$ k-points is used to sample the band structure of the Cu cell. The original RS Hamiltonian block size is equal to 6000, made up of 240 atoms times 25 orbitals per atom. Using a larger initial guess leads to a reduction in both the number of refinement iterations and the error in the resulting MS band structure, but it also increases the MS block size.

| Size of initial guess ($N_M$) | 350 | 380 | 390 |
|---|---|---|---|
| Refinement iterations | 38 | 34 | 26 |
| MS block size | 388 | 414 | 416 |
| Size reduction (%) | 93.5 | 93.1 | 93.1 |
| Max. error (meV) | 0.89 | 0.62 | 0.34 |

et al. [182] suggested to choose $k_q$ according to the position of the unphysical energies. Such a scheme results in a nonhomogeneous k-point grid. In the absence of equidistant k-points, however, instabilities can occur during the minimization, which may lead to a divergence of the refinement procedure, when the RS basis is composed of nonorthogonal functions. We opted for an intermediate scheme where the $k_q$ are selected based on Eq. (2.49) with $n_q = 5$. Such a grid is too coarse to sample all unphysical bands and does not guarantee convergence on its own. This shortcoming is eliminated by augmenting $k_q$ with points based on an analysis of Eq. (2.51). The $\pm k_x$ that correspond to the positions with the largest number of unphysical energy states ($N_s$) are added to $k_q$. At every iteration of the refinement procedure, the MS band structure is recomputed with the updated $U$, and the $\pm k_x$ are adjusted accordingly. This scheme ensures the same stability as that obtained with a regular k-point grid, it converges through sampling of at least one unphysical band at the time, and it keeps $n_q$ small for computational efficiency.

The number of integration points in Eq. (2.52), $n_z$, is not a critical parameter. With $n_z = 6$ convergence can be readily achieved. Increasing $n_z$ could avoid the addition of few modes to the final transformation matrix, but it rapidly increases the computational cost too.

Minimizing Eq. (2.52) is a challenging task as the functional has many local minima. This is one of the biggest issues encountered in the refinement procedure because there is no algorithm available that guarantees that the global minimum is identified if the sum in Eq. (2.52) has more than one term [183]. Having a good initial guess is thus essential to determine modes that improve the MS band structure. From numerical testing, it appears that the rows of $U$ form a suitable initial guess for the $C$'s. To enable a fast convergence of the refinement with the lowest possible number of iterations, we minimize Eq. (2.52) with multiple initial guesses and always select the mode that produces the smallest value of $\mathcal{F}(C)$. To reduce the computational burden, we perform the minimization with the first 10% of the rows of $U$.

The last parameter to be selected is the trial basis $\Xi$. The matrix proposed in literature [163, 164] is

$$\Xi(U) = (\mathcal{I} - UU^\dagger) \cdot [S_{k_1}^{-1} H_{k_1} \oplus S_{k_2}^{-1} H_{k_2}] \cdot U, \tag{3.2}$$

where $\mathcal{I}$ is the identity matrix of appropriate size and $[\oplus]$ the concatenation operator. We find that by replacing $[\oplus]$ with the commutator $[,]$ a smaller variational space is obtained

$$\Xi(U) = (\mathcal{I} - UU^\dagger) \cdot [S_{k_1}^{-1} H_{k_1} \ , \ S_{k_2}^{-1} H_{k_2}] \cdot U, \tag{3.3}$$

which is sufficient and often outperforms the result of Eq. (3.2) in terms of accuracy, besides offering lower computational cost. It is possible, however, that towards the end of the refinement, the $\Xi$ obtained from Eq. (3.3) does not provide enough degrees of freedom anymore. Thus, when Eq. (2.52) stops providing modes satisfying $\mathcal{F}(C) < -1$, we switch to Eq. (3.2)

**Table 3.2:** Timing data for the MS transformation procedure. The profiling is performed on a workstation featuring two Intel Xeon E5-2680 v4 CPUs with 28 cores in total, and the procedure is implemented in MATLAB 2018b. The refinement iteration is broken down into three parts: (1) the preparation of the matrices A and B, (2) the global minimization, and (3) the calculation of the band structure.

| | Runtime |
|---|---|
| RS band structure (s) | 1096.7 |
| Initial guess $U_0$ (s) | 19.4 |
| Refinement iteration ($N_M = 350$) (s) | 65.4 |
| (1) Matrices $A$ and $B$ (s): | 28.3 |
| (2) Minimization of $\mathcal{F}(C)$ (s) | 34.1 |
| (3) MS band structure (s) | 3.0 |

to determine $\Xi$. To fully define $\Xi$, two wave vectors $k_1$ and $k_2$ must be selected; the first one corresponds to the position with the highest $N_s$ in the initial MS band structure, whereas the second one is the $\Gamma$ point.

The computational resources required to obtain the MS transformation matrix and to eliminate spurious states from the band structure depend on the size of the matrices $H_{RS}$ and $S_{RS}$, on the width of the energy window [164], and on the choice of the parameters discussed in this section. The times needed to acquire the MS transformation matrix for the band structure shown in Fig. 3.1(a) are listed in Table 3.2. The time per refinement iteration, 65.4 s, is roughly split in half between solving Eqs. (2.53-2.58) and minimizing Eq. (2.52), both taking about 30 s. Calculating the MS band structure, on the other hand, hardly contributes to the computational effort. The 38 refinement iterations needed to clean up the band structure add up to 3644 s. Together with the computation of the RS band structure, the entire procedure takes 4740 s.

## 3.4 Hybrid mode-space – real-space device Hamiltonian

The Hamiltonian and overlap matrices $H$ and $S$ in Eqs. (2.13) and (2.14-2.15) are first computed with DFT a using localized GTO DZVP [180] basis set. Such a basis has the advantage of producing a banded RS Hamiltonian $H_{RS}$ that is directly suitable for QT calculations without further processing [30, 125]. To reduce the size of the matrices $H$ and $S$, the MS transformation is employed. Because the original matrices are not homogeneous throughout the device structure due to the MIM stack of CBRAMs, the transformation is applied locally. In the contact extensions the diagonal and off-diagonal blocks of $H_{MS}$ are given by

$$\mathcal{H}_n = U^\dagger H_n U. \tag{3.4}$$

The $H_n$'s in Eq. (3.4) correspond to the on-site term ($n = 0$) and to the coupling of one Cu unit cell and its 240 atoms, to its $n$th nearest neighbor ($n > 0$). The Cu blocks at the $SiO_2$ interface and the oxide regions embedded between the metallic electrodes remain expressed in RS. At the transition between the periodic area of the device, which is transformed into MS, and the RS active region, the off-diagonal blocks must be treated carefully as they exhibit a hybrid character. They undergo a one-sided transformation

$$\mathcal{H}_{RM,n} = U^\dagger H_n, \ \mathcal{H}_{MR,n} = H_n U. \tag{3.5}$$

The subscript RM indicates that $\mathcal{H}_{RM,n}$ connects an RS to an MS region and vice-versa. The MS transformation is not optimized for the surface blocks of the Cu contacts, which are in direct contact with the $SiO_2$ layer and whose atomic structure is different from that of the other Cu cells due to the additional relaxation imposed by the neighboring oxide. Still, the MS transformation can theoretically be applied to them using the

same matrix $U$ as for the other Cu cells because they are made of the same number of atoms. In the next section, we present two strategies, one with all metal blocks transformed into MS and one with one metal block at the oxide interface remaining in RS. The sparsity patterns of the RS and the two hybrid Hamiltonian matrices are depicted in Fig. 3.3. The large Cu blocks at the interface between the RS and MS domains disappear and become rectangular. Only the active region, through which a metallic filament grows and then dissolves, remains in RS. As this region is not affected by the MS transformation, the matrix $U$ does not need to be recomputed when structural changes occur within the oxide, e.g. atomic relocations. This feature is especially appealing when one considers components whose atomic geometry evolves with time such as CBRAM cells.

An alternative, more general approach is to treat the entire simulation domain in RS and compute only the OBCs in MS. The coupling between MS and RS is achieved by transforming the boundary self-energies from MS to RS with $\Sigma^R_{RS}(E) = U \cdot \Sigma^R_{MS}(E)U^\dagger$ or by constructing hybrid blocks connecting the MS boundary self-energy to the RS central part of the device. The benefit from the evaluation of the OBCs in MS is equal to that offered by the first approach. The solution of the transport equations with QTBM or NEGF does not benefit from the MS transformation. While providing a lower gain in computational efficiency, this method is generally applicable to any quantum transport simulation, regardless of the length of the contacts. As such, the proposed MS-RS scheme finds applications beyond the CBRAM cells considered here.

It should be noted that the two contacts are independent of each other. While the device illustrated in Fig. 3.2(a) features two identical electrodes, the proposed hybrid scheme does not enforce such a constraint. The MS transformation matrix can be computed and applied to the two contacts independently by transforming the respective blocks obtained from the Hamiltonian and overlap matrices with Eqs. (3.4-3.5).

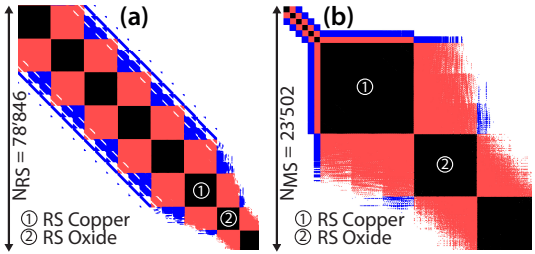


**Figure 3.3:** Sparsity patterns of the (a) RS and (b) hybrid MS-RS Hamiltonian matrices corresponding to the left half of the CBRAM cell in Fig. 3.1(a). The presence of second-nearest-neighbor block connections translates into a pentadiagonal matrix structure, with on-site blocks (black), first-nearest-neighbor blocks (red), and second-nearest-neighbor blocks (blue). The electrode region of the hybrid Hamiltonian is composed of six MS and one RS block. The overall matrix size is reduced from 78 846 to 23 502. Although the electrode region is transformed into MS, the active device area remains in RS, which is recognizable from the much larger block size in subplot (b). The hybrid off-diagonal blocks connecting the RS and MS take the form of long thin rectangular blocks. Note that the blocks labeled ① in subplots (a) and (b) are identical in both representations. The same holds true for the blocks labeled and ②.

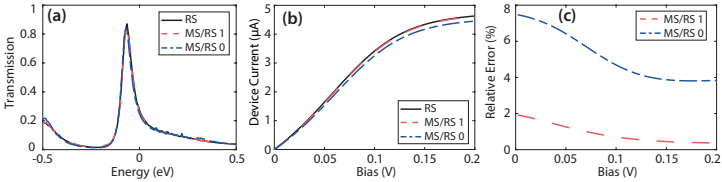


**Figure 3.4:** (a) Ballistic transmission through the device structure shown schematically in Fig. 3.1 with different representations of the Hamiltonian. The RS reference curve is plotted in black, while the red and blue dashed curves correspond to the hybrid model MS-RS$_1$ and MS-RS$_2$, respectively. (b) Low-field current versus applied voltage between the Cu electrodes for the same structure, Hamiltonian matrices, and plotting conventions as in (a). (c) Relative error between the hybrid model 1 (dashed red curve) and 2 (blue curve with small dashes) and the reference RS results.

## 3.5 Ballistic Simulations: Benchmark

In the present section, we benchmark the MS-RS hybrid scheme and compare two different versions of it with the RS reference, in terms of both accuracy and computational efficiency. The first hybrid version, MS-RS$_1$, has all contact blocks transformed into MS using Eqs. (3.4-3.5), except for the surface Cu blocks, which form the interface with the insulating layer. In the second version, MS-RS$_2$, all metallic blocks, including those in contact with $SiO_2$, are transformed into MS. A single transformation matrix $U$ is used in all cases.

The quantities of interest are the ballistic transmission function and the current flowing through the device. The transmission function is shown in Fig. 3.4(a) within the energy window where the MS transformation is valid. The MS-RS$_1$ model agrees almost perfectly with the RS reference; MS-RS$_2$, on the other hand, shows some discrepancies. The latter are attributed to the fact that projecting the relaxed surface block onto the same MS basis as the perfectly crystalline Cu blocks does not fully work. The same behavior is reflected in the ballistic current in Fig. 3.4(b).

The latter is computed with Eq. (2.23), keeping the transmission function constant and varying the Fermi distribution function of the right contact as a function of the applied voltage. It can be seen that MS-RS$_1$ follows almost exactly the RS reference, while MS-RS$_2$ slightly underestimates the current. The observed error is quantified in Fig. 3.4(c), where the relative difference between the currents obtained from the hybrid and the RS simulations is reported. For MS-RS$_1$ this error does not exceed 2% at very low bias, and decreases at larger voltages. MS-RS$_2$ displays deviations of up to 8%, reaching its maximum at low voltages.

All calculations are performed on CPU-GPU hybrid nodes with 64 GB of memory each, using 20 CPUs and 2 GPUs to treat each energy point. These computational resources are distributed over as few nodes as possible, while simultaneously accommodating the memory constraints. Detailed timings are recorded in Table 3.3. The size of the matrices is reduced by factors of 3 and 7 for MS-RS$_1$ and MS-RS$_2$, respectively. The number of nonzero elements is cut down by even more, by factors of 6 and 25, reducing the memory requirement accordingly. The time required to compute the OBCs is roughly equal for both hybrid schemes, as it only depends on the size of the boundary blocks, which is the same in MS-RS$_1$ and MS-RS$_2$. A speedup of more than 150 is achieved for the OBCs. Solving the linear system of equations resulting from Eq. (2.13) is shortened by factors of 48 and 191. The gain overall in computational time is less significant than could be expected from the timings of the OBCs and linear system calculations, with speedups of 27 and 95 for MS-RS$_1$ and MS-RS$_2$, respectively. In effect, in the current implementation of the algorithm, the postprocessing stage, where $T(E)$ and the DOS are derived from the wave function, does not offer a substantial speedup in MS, as it mainly depends on the number of states injected into the simulation domain. This number is the same in RS and MS. The postprocessing accounts for less than 10% of the run time in the RS simulation, but it increases to about 50% of the time in both MS approaches. Addressing this bottleneck and optimizing the underlying

algorithms for the hybrid scheme could therefore lead to an even larger speedup. Other than the timing, the memory footprint is a large cost factor. While the RS simulations need to be performed on 20 nodes, the MS-RS simulations can run on 4 and 2 nodes, depending on the model used (1 or 2), adding to the cost reduction. Combining these two aspects into a cost function "run time $\times$ number of nodes" we report a reduction factor of 136 and 955 for MS-RS$_1$ and MS-RS$_2$, respectively.

## 3.6 Conclusion

We developed a robust scheme to select parameters for successful MS transformations of metallic layers. Based on a nonorthogonal DFT RS Hamiltonian, a size reduction of over 90% can be achieved. We also introduced a scheme to locally transform the Hamiltonian matrix of a device into MS. This hybrid MS-RS approach extends the use of MS to structures with inhomogeneities such as metal-insulator interfaces or amorphous layers along the electron-transport direction. Numerical benchmarks against an RS reference reveal that the accuracy of the proposed method is excellent, with relative errors of the electrical current below 2%. The gain in computational efficiency together with the reduction in the memory footprint results in performance improvements by 2 to 3 orders of magnitude.

More generally, if the region transformed to MS and the energy window of interest remain the same, as in most resistive switching memory cells, the same transformation matrix can be reused each time the active layer undergoes a structural change. This feature renders our approach very powerful to investigate dynamically evolving devices. Here, this was demonstrated with the ON-state configuration of a CBRAM cell.

Apart from CBRAMs, the hybrid MS-RS approach is especially appealing to study devices with metallic contacts, where their treatment dominates the overall computational time. This could accelerate, for in-

**Table 3.3:** Summary of the computing times for the RS and the different RS-MS hybrid schemes. The calculation of the OBCs and the solution of the resulting linear system of equations are done in parallel, the former on CPUs, the latter on GPUs. The total time includes the aforementioned components plus postprocessing such as the extraction of the density of states and the transmission function. The numbers are averaged over the calculation of 20 energy points. The number of nodes required to run the simulation is directly related to the memory needed to store and process the data. It shrinks with the number of nonzero elements in the Hamiltonian matrix. The total cost of the simulation, defined as run time $\times$ number of nodes, is reduced by 2 and almost 3 orders of magnitude with the help of the MS-RS$_1$ and MS-RS$_2$ schemes, respectively.

| | RS | MS-RS$_1$ | Gain vs RS | MS-RS$_2$ | Gain vs RS |
|---|---|---|---|---|---|
| Total matrix size | 78 846 | 22 726 | 3.47 | 11 502 | 6.86 |
| Nonzero elements | 673.8e6 | 104.9e6 | 6.43 | 27.5e5 | 24.51 |
| Time for OBCs (s) | 308.8 | 1.61 | 191.5 | 1.84 | 167.5 |
| Time for linear system (s) | 273.9 | 5.72 | 47.9 | 1.44 | 190.7 |
| Time for $T(E)$ and DOS (s) | 27.2 | 6.58 | 4.14 | 1.66 | 16.35 |
| Total time (s) | 336.0 | 12.3 | 27.3 | 3.52 | 95.5 |
| Nodes | 20 | 4 | 5 | 2 | 10 |
| Cost (time$\times$nodes): | 6720 | 49.3 | 136.4 | 7.19 | 954.9 |

stance, the characterization of the contact resistance for a multitude of materials and devices. While the influence of contact resistances is often ignored in quantum transport simulations, it can severely limit the performance of nanoscale devices in reality.

To further improve the hybrid MS-RS approach, the inclusion of scattering through self-energies should be considered. Mil'nikov et al. [163] demonstrated the feasibility of incorporating scattering in pure MS simulations, but the proposed technique needs careful adjustments and verifications in the hybrid approach. Moreover, the Hamiltonian is not treated self-consistently in the present work. Shin et al. [164] performed self-consistent MS-only calculations. Errors in the charge density arising from the partial MS treatment of the simulation domain, even though small, could affect the electrostatic potential. Fully self-consistent hybrid simulations should be performed and compared to real-space calculations to verify the correctness of the approach. Solving the Poisson equation would enable the simulation of larger bias voltages than those considered here.

Chapter 4

# Filament in a Cu/a-$SiO_2$/Cu Cell[1]

## 4.1 Introduction

The localized nature of the filamentary switching mechanism in combination with the simple metal-insulator-metal (MIM) structure of CBRAM are keys to scale such devices to extremely small feature sizes [29]. This constitutes a major advantage of CBRAM over competing nonvolatile memory (NVM) technologies [4] because a high-density integration is possible [184]. Shrinking the cross-section of such a memory cell can also improve the endurance of CBRAM [185]. Moreover, reducing the active switching volume through geometry engineering can improve the memory characteristics, e.g. reduced operating voltages and lower cycle-to-cycle variability [184]. These phenomena have been attributed to the fact that the stochasticity of the filament growth process can be decreased by constraining the volume through which it grows and by minimizing the amount of metal ions that is dissolved in the solid electrolyte.

As the dimensions of CBRAM cells decrease, unique physical effects emerge that are masked or absent in larger structures. A peculiarity found experimentally is that the stability with respect to device failure improves if the thickness of the SL is shrunk [29]. Specifically, the maximum current that is supported by a single cell drastically increases. The thinnest ever reported CBRAMs with SL of 1 nm of $SiO_2$ only endured the largest operating currents before permanently breaking down. The maximum

[1] This chapter is based on [50], [136] and [161]

current versus the insulator thickness is illustrated in Fig. 4.1(a). Another feature observed in experiments is that during the ON-OFF switching cycle the conductance of the CBRAM does not change gradually, but in a steplike fashion with multiple discrete intermediate stages, as shown in Fig. 4.1(b).

To investigate these effects, we performed QT simulations on structures containing already formed metallic filaments. The switching characteristics of CBRAM cells arise due to the relocation of atoms, a process that is computationally expensive to account for at the *ab initio* level [186]. Typically, QT simulations assume therefore that atoms are located at static positions. To investigate the observed behavior, we implemented a straightforward scheme to study the influence of the filament dissolution on the CBRAM transport properties. Furthermore, we examined the electro-thermal properties of CBRAM cells with respect to the thickness of the SL. Due to the extremely narrow dimensions of these filaments, high current densities are expected, with potentially significant self-heating effects. To design better performing CBRAM cells it is therefore critical to precisely understand the interplay between electron transport and atom positions as well as their influence on the lattice temperature. In particular, we focused on the impact of self-heating and thermal conductance.

After briefly summarizing few computational details, the results presented in this chapter will be divided into three separate subsections. In the first one, the intermediate conductance states originating from the dissolution of the Cu filament are investigated. The second subsection presents the electro-thermal properties of a filamentary structure in the ON-state with a fixed oxide thickness of 3.5 nm. The third subsection is concerned with the dependence of these electro-thermal properties on the thickness of the SL. Three devices in the ON-state with oxide thicknesses ranging from 1.6 nm to 3.5 nm are compared to each other.

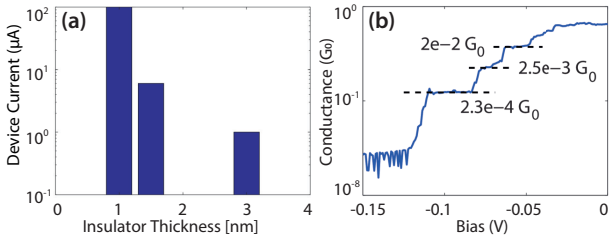


**Figure 4.1:** (a) Thickness dependence of the current supported by $Ag/SiO_2/Pt$ CBRAM cells. The bars represent the maximum current achieved before device breakdown. Values are reported for different thicknesses of the switching layer. Adapted from [29]. (b) Extracted conductance $G$ vs. voltage $V$ of a single ON-OFF switching event for a $Cu/SiO_2/Pt$ CBRAM cell. The dashed lines mark three intermediate conductance plateaus obtained during the Cu filament dissolution.

## 4.2 Computational Details

The cross-section of the considered $Cu/SiO_2/Cu$ atomic structure is $2.34\times2.38$ nm$^2$. Both electrodes are composed of the same metal and measure 4.1 nm along the MIM axis. They are made of 7 slabs containing 240 atoms each, for a total of 3360 Cu atoms. The surface blocks of both electrodes, i.e. those in contact with the insulator, are optimized, whereas the atoms composing the reminder of the contact are fixed to their bulk positions. Three filamentary structures are studied: the thickness of their SL measure 3.5 nm, 2.4 nm, and 1.6 nm, respectively. In the largest configuration 96 Cu atoms build the metallic filament. The oxide matrix is composed of 331 Si and 662 O atoms. The DFT calculations are fully periodic in all three cartesian directions. This ensures that no vacuum interface is present that could distort the electronic structure. Cross-sections of more than $2\times2$ nm$^2$ are sufficiently large to avoid cross-talks between the filament of interest and its periodic images. For simplicity, both metal contacts are made of the same metal. This approximation greatly reduces the computational complexity, and is necessary to per-

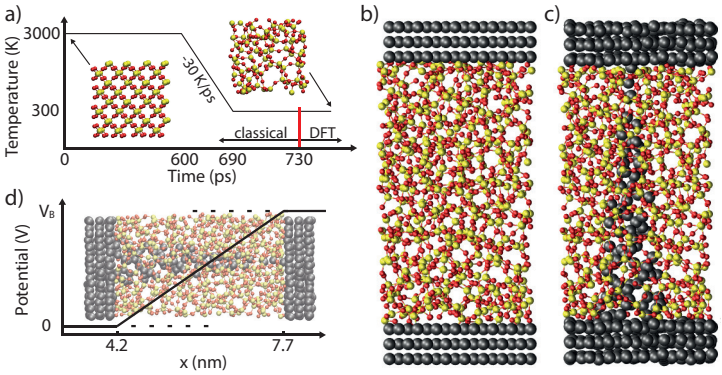


**Figure 4.2:** (a) Melt-and-quench procedure to generate samples of a-$SiO_2$. A box with crystalline $SiO_2$ or randomly placed Si and O atoms is melted at 3000 K. Subsequently, it is cooled at a rate of 30 K/ps and annealed at 300 K. The melt-and-quench is performed classically, the post-annealing and optimization with DFT. (b) Metal-insulator-metal structure of a pristine CBRAM cell. Metal electrodes are attached to a slab of a-$SiO_2$ obtained with the procedure from (a). (c) Filamentary configuration of a CBRAM cell in the ON-state. A metal filament is inserted into the a-$SiO_2$ from (b) by replacing all Si and O atoms within a cone and annealing the result with DFT. (d) Shape of the potential applied to the CBRAM cell in the quantum transport calculations. The potential increases linearly over the switching layer from 0 to $V_B$ and is constant within the Cu electrodes.

form the calculations at hand. Its validity, however, is challenged and examined in Chapter 5.

Assembling the largest CBRAM cell with 3.5 nm of amorphous $SiO_2$ as SL requires several steps. First, amorphous $SiO_2$ (a-$SiO_2$) with a density of 2.20 g/cm$^3$ is obtained by using a melt-and-quench procedure [88]. A sample of cristobalite $SiO_2$ is melted and quickly cooled to room temperature to freeze in an amorphous configuration, as illustrated in Fig. 4.2(a). Melting happens at 3000 K for 600 ps. It is followed by a quenching phase with a cooling rate of 30 K/ps to reach 300 K. Lastly, the amorphous structure is annealed for 40 ps at 300 K. The MD sim-

ulations are performed with a ReaxFF force-field [87], as implemented in the QuantumATK 2017.1 software [131, 187]. Next, the Cu electrodes are attached to the slab of the a-$SiO_2$ layer and a conical Cu filament is inserted such that it spans the entire $SiO_2$ and connects the electrodes. The resulting structure is shown in Fig. 4.2(b-c). The filamentary geometry is obtained by converting all Si and O atoms within a cone to Cu ones. The assembled CBRAM model is then optimized within DFT and annealed with AIMD at 800 K for 3-4 ps to relax the stress induced by attachment of electrodes to the a-$SiO_2$ and the insertion of the metallic filament into the SL.

The two other configurations with a-$SiO_2$ layer thicknesses of 2.4 nm and 1.6 nm are created by shortening the SL of the original CBRAM model with 3.5 nm of $SiO_2$. The conductance of the Cu filament is postulated to be largely determined by its thinnest part, while the length is much less important. This assumption will be justified in Section 4.3.3. Therefore, to ensure an identical atomic configuration of the filament tip in all three structures, the two shortest ones are obtained from the 3.5 nm cell by cutting out the base, i.e. the thicker end of the conical filament. The shortened cells with 2.4 nm (1.6 nm) a-$SiO_2$ feature 44 (18) filament atoms as well as 241 (164) Si and 482 (328) O atoms.

The DFT simulations are performed with the CP2K code [179], which employs a GTO basis set with the linear combinations of atomic orbitals (LCAO)[141] method. Such a localized basis is suitable for subsequent NEGF quantum transport simulations. The exchange-correlation energy is approximated by the PBE functional [115], and the atomic cores are described by Goedecker-Teter-Hutter (GTH) pseudopotentials [181]. All metal atoms are represented by a double zeta-valence polarized (DZVP)[180] basis set to construct the atomic structure. To calculate the $H$ and $S$ matrices required for Eq. (2.11) a single zeta-valence (SZV) basis [180] is used to represent the Cu atoms. It is combined with the 3SP parameterization of Zijlstra et al. [188] to describe the Si and O atoms. Both basis sets, although small, are sufficient to represent the

electronic structure of the conductive states around the Fermi energy. The energy-resolved transmission function is shown in Fig. 4.3(a) and compared to the one obtained with the larger DZVP basis set. Only minor differences can be observed, thus justifying the usage of the lighter basis. The OFF-state, however, is not as accurately captured by the SZV + 3SP combination, as can be deduced from Fig 4.3(b). Therefore, the ON-OFF switching behavior is investigated based on $H$ and $S$ matrices obtained with DZVP basis sets for all atoms. The electron-phonon coupling elements are computed with the bond stretching scheme introduced in Section 2.3.3 that relies on a hydrostatically strained Hamiltonian. This approach is computationally less demanding than more elaborate techniques, which is crucial when dealing with large systems, as in this study.

The dynamical matrix accounting for the thermal properties of the CBRAM cells is determined with the frozen-phonon approach discussed in Section 2.3.2. Computing the forces required to construct $\Phi$ from first principles is prohibitive for systems that comprise more than 3000 atoms. Therefore, these calculations are performed with a ReaxFF [89] force-field specifically parameterized [87] to capture the switching behavior of CBRAM. To minimize the number of negative eigenvalues in the dynamical matrix the atomic structure is reoptimized with the force-field. This step is followed by the calculation of the dynamical matrix. As a consequence, the atomic structure is not identical to the DFT case. Because the configuration does not change considerably, the impact of the optimization is expected to remain small.

The Hamiltonian, overlap, and dynamical matrices as well as the derivatives of the Hamiltonian are imported into the OMEN quantum transport simulator [150, 154]. The latter implements the NEGF-Poisson scheme for electrons, phonons, and their coupling, as detailed in Section 2.3. The ballistic NEGF simulations for electrons are performed in the low field limit, which assumes that the applied bias does not significantly alter the charge density, and thus the electronic structure, within the considered

domain. Hence, the Hamiltonian and overlap matrices from CP2K are not updated based on the non-equilibrium charge derived from NEGF so that no solution of Poisson's equation is needed. Such a scheme dramatically reduces the computational burden. The calculations including electron-phonon interactions are done with a potential difference applied between the metallic electrodes. As a realistic approximation a linear drop of the potential energy is imposed over the SL, as illustrated in Fig. 4.2(d). Poisson's equation is not solved in this case either.

Evaluating the intermediate resistance states of the CBRAM during the ON-OFF switching necessitates to perform multiple QT simulations. The latter only differ in the number of filament atoms, whereas the contacts and the oxide remain identical. This numerical experiment can ideally leverage the efficiency of the hybrid MS-RS approach presented in Chapter 3. The contact regions of the Hamiltonian matrix are transformed into MS while the filament remains in real-space. In accordance with the results of the previous chapter, the MS-$RS_1$ model is employed, with one RS metal block at the interface between the MS region and the amorphous oxide.

## 4.3 Results

Using the model described in Section 4.2 above the switching behavior of Cu/a-$SiO_2$/Cu CBRAM cells is examined and their electro-thermal properties are calculated in the ON-state. The results from these investigations are presented in the three next subsections.

### 4.3.1 ON-Off Switching

In the first CBRAM structure, a metallic filament extends through the entire oxide layer, creating a conductive path between the two metallic electrodes, as shown in Fig. 3.1(a). A rendering of the lines of the current field in RS is plotted in Fig. 4.4(a). It can be observed that the current

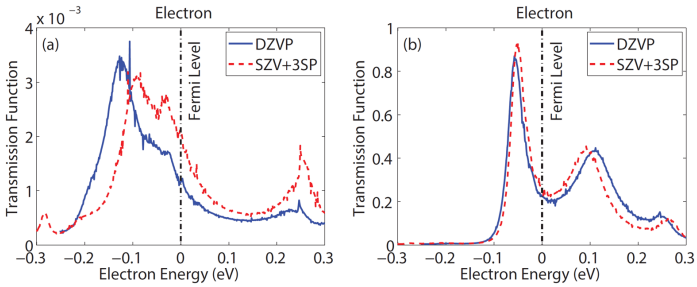


**Figure 4.3:** (a) Energy-resolved electron transmission function $T(E)$, as calculated with Eq. (2.23), through an incomplete nanofilament formed in a Cu/a-SiO$_2$/Cu CBRAM (OFF-state). A gap length of 1 nm between the tip of the filament and the active electrode is left without Cu atoms. It corresponds to 9 missing atoms (87 instead of 96 Cu atoms in the filament). The reported data were computed with a Hamiltonian matrix created by CP2K and expressed either in a double-zeta valence polarized (DZVP, solid blue line) basis set or a combination of 3SP for the Si and O atoms and single-zeta valence (SZV) for Cu (dashed red line). Note that the Fermi level energy was shifted to E=0. (b) Same as (a), but for the complete nanofilament (ON-state).

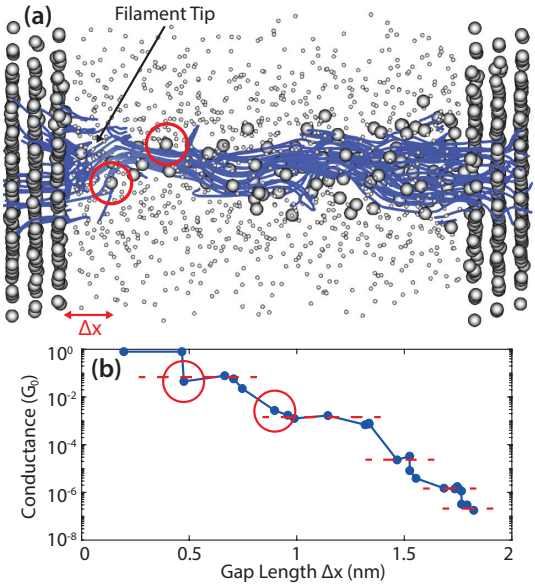


**Figure 4.4:** (a) Rendering of the electrical current filed lines for the CBRAM device shown in Fig. 3.1(a) under an applied bias of 1 mV. Large gray spheres represent Cu atoms, small spheres Si and O ones. The current follows a curly path through the filament and focuses at the tip before spreading again through the right contact. Not all Cu atoms carry current, especially surface atoms and certain atoms at the tip are avoided. Some of them are indicated by a red circle (b) Conductance of the CBRAM given in units of $G_0$ vs. $\Delta x$, the gap created at the filament tip when atoms are removed. Each blue dot corresponds to a different filament configuration. The first point on the left represents the conductance of the device in sub-plot (a), the second point the same structure, but with one atom less at the tip of the filament. This procedure has been repeated until a gap length $\Delta x = 1.8$ nm was obtained. Red circles refer to the atoms marked equally in (a), i.e. ones that do not carry electrical current. Red dashed lines indicate the discrete conductance levels.

is confined within the filament. The field lines are focused onto the tip of the filament, where the highest current density is found. These lines also reveal that a substantial part of the current flows through the oxide around the tip, so that the narrowest current distribution happens slightly before the filament tip. The current is rather "curly." Not all atoms appear to contribute to the current.

Starting at the tip, the atoms that build the filament are removed one-by-one. The conductance is reported in Fig. 4.4(b) as a function of the length of the gap that is formed between the filament tip and the closest metallic electrode. The conductance decreases exponentially as the gap increases, showing the same dependence as the current produced by the tunneling of an electron through a potential barrier. The decrease is, however, not uniform. It resembles a steplike process, as indicated by the presence of multiple plateaus in Fig. 4.4(b). The removal of certain atoms does not reduce the conductance, thus leading to the observed behavior. The atoms whose absence does not affect the device conductance are those that do not carry any current in Fig. 4.4(a), i.e., those that are not crossed by field lines and that are circled in red, for example. The results in Fig. 4.4(a) reveal that a gap of 1.5 nm is sufficient to achieve an ON-OFF ratio of six orders of magnitude. For the case of the investigated filament 20 atoms need to be displaced to open this gap.

Close investigations of experimental ON→OFF switching characteristics in Ag/a-$SiO_2$/Pt CBRAM cells [48] show the same qualitative behavior although the material system is different. The conductance in a voltage-sweep experiment does not decrease continuously, but instead exhibits a steplike descent, just as observed in our simulations. In experiments, this characteristic was attributed to discrete relocations of filament atoms that change the electrical properties of the filament. In the modeling, the elimination of a filament atom corresponds to the relocation of a Cu atom out of the simulation domain, thus mimicking the experimental diffusion. It therefore appears that these plateaus can to geometrical effects.

### 4.3.2 Electro-thermal Properties of the CBRAM in ON-state

Here, we examine electro-thermal effects in the ON-state of the CBRAM cell introduced above with a $SiO_2$ thickness of 3.5 nm. The ballistic current and the current with electron-phonon interactions that flow through the filament configuration were computed. The $I$-$V$ characteristics are shown in Fig. 4.5(b). An ohmic behavior is revealed at low biases, i.e. a linear increase of the current vs. voltage. The electron-phonon limited current reaches about 69% of the value of the ballistic one at room temperature. The impact of scattering on the magnitude of the electrical current is rather low, which is to be expected from such a short device whose length barely exceeds its mean free path for scattering. Hence, it operates near its ballistic limit if only the electrical current is considered. The resistance, extracted as a linear fit to the $I$-$V$ curve, is 48.1 k$\Omega$ in the ballistic case and 57.6 k$\Omega$ in the dissipative one. These values lie in the range of typical CBRAM ON-state resistances [7]. The current field lines with electron-phonon scattering are rendered in Fig. 4.5(a) at an applied voltage of 0.2 V. As in the ballistic case the current is confined to the filament but the field lines are less "curly," though some atoms of the filament still do not carry any current. The influence of electron-phonon scattering on the current can be best visualized in the form of a spectral plot representing its energy and spatial distribution, as given in Fig. 4.5(c). Electrons lose part of their energy when propagating through the device. Such an energy dissipation is not possible in ballistic electron transport. Most of the energy loss occurs close to the tip of the filament or in the metallic contact attached to it. This fact agrees well with the observation that the device operates close to its ballistic limit. The propagation of most electrons through the oxide layer is too fast to allow them to interact with phonons. Therefore, they cannot relax their energy within the filament. It should also be noted that a large portion of the electron population enters (leaves) the simulation domain with an energy below (above) the Fermi energy of the respective contact.

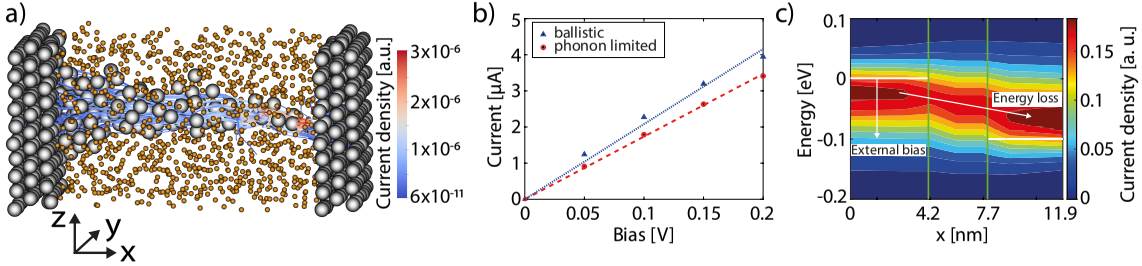


**Figure 4.5:** (a) Atomic structure of a CBRAM cell containing a metallic filament. The copper atoms are represented by gray spheres, silicon and oxygen by orange ones. The current flowing through this configuration upon an external bias of 0.2 V is plotted as the blue-red lines. The red color at the tip of the filament (right-hand-side of the oxide layer) indicates a large current density which decreases (dark blue) towards the left. (b) Ballistic (blue triangles) and electron-phonon limited (red diamonds) $I$-$V$ characteristics of the device in (a). A linear fit to the data points is given by the blue dotted (ballistic) and red dashed (e-ph) lines. (c) Energy- and position-resolved current flowing through the cell in (a) in the presence of electron-phonon scattering. The largest current density is observed just below (above) the equilibrium potential in the left (right) contact, which is indicated by a white line.

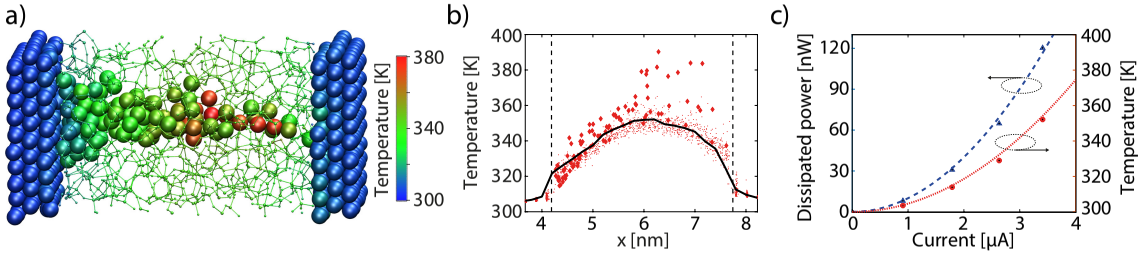


**Figure 4.6:** (a) Atomically-resolved temperature in the device of Fig. 4.5(a) at an applied bias V=0.2 V. The heating is more prominent in the copper atoms in the right half of the filament, marked by the red color. (b) Atomic temperature projected onto the x-axis (transport). The filament atoms are marked with the red diamonds, the oxide by the red dots. The vertical dashed lines indicate the electrode-device interfaces. The black line refers to the average temperature over the cell cross-section (yz-plane). The maximum temperature is given by the top of this black curve. (c) Maximum temperature and dissipated power as a function of the CBRAM device current. The simulated values are given as symbols. A quadratic fit of the data is superimposed as dashed (power dissipation) and dotted (maximum temperature) lines.

This indicates that a lot of the power dissipation happens outside the simulated device, namely in the leads. The fraction of the power that is consumed within the simulation domain is called the internal dissipation fraction and labeled $\alpha$. It can be computed as

$$\alpha = \frac{P_{Dev}}{VI} = \frac{|I_{dE,ph,R}| - |I_{dE,ph,L}|}{VI} \tag{4.1}$$

where $P_{Dev}$ is the power dissipated inside the device. It corresponds to the difference between the energy current carried by the phonons evaluated at the right ($I_{dE,ph,R}$) and left ($I_{dE,ph,L}$) contact boundaries. Note that due to energy conservation $I_{dE,ph,R} - I_{dE,ph,L} = I_{dE,el,L} - I_{dE,el,R}$, the difference is the energy current carried by electrons between the left and right electrodes. In other words, $P_{Dev}$ is the amount of electrical power that is converted into heat inside the simulation domain. In Eq. (4.1), $V$ is the applied bias and $I$ the resulting electrical current. For the considered CBRAM cell we find that $\alpha \approx 0.18$, which signifies that more than 80% of the power is dissipated outside the simulation domain, inside the metallic electrodes.

Due to energy conservation, a reduction in electron energy must be compensated by the creation of additional phonons corresponding to an increase of the lattice temperature. Generally, the temperature is an average quantity characterizing an ensemble of particles embedded within a reservoir. It is therefore a macroscopic property. Nevertheless, by relating the excess phonon population on each atom to an equilibrium temperature through the Bose-Einstein distribution, as in Eqs. (2.45-2.46), self-heating effects can be mapped to the familiar scale of temperature. The heat map of the device from Fig. 4.5(a) at 0.2 V is provided in Figs. 4.6(a-b) with a 3-D atomic resolution and after projection onto the transport axis. The highest temperature of the device lies in the middle of the SL, as indicated by the red atoms. This does not correlate with the highest electron-phonon scattering rate, which can be found in the electrodes or at the oxide-metal interface. Evidently, there has to be a

second mechanism involved in the heating process. This is identified as the heat extraction rate from the oxide layer, which can be cast into the thermal resistance $R_{th}$. We define the latter as

$$T_{max} = T_0 + R_{th}\alpha R I^2 \tag{4.2}$$

and use the simulated temperature data to determine its value. In Eq. (4.2) $T_0$ is the ambient temperature, $R$ the electrical resistance, and $T_{max}$ the maximum calculated temperature averaged over the cross-section of the CBRAM cell. This quantity is shown in Fig. 4.6(b). The power dissipation and maximum temperature are plotted as a function of the device current in Fig. 4.6(c). Both behave as predicted by simpler physical models [17], i.e. they increase quadratically with the current. Up to a current $I = 1\mu A$, $P_{dev}$ and $T_{max}$ do not exceed 10 nW and 305 K, respectively, so that self-heating is almost negligible. Past this point, the situation rapidly deteriorates as $I$ increases.

Using Eq. (4.2) a thermal resistance of $R_{th} \approx 0.4K/nW$ can be extracted from the simulation data. In the fitting, it is crucial to consider the average temperature, and not individual values because large variations may occur between the coldest and hottest atoms across the yz plane, as can be observed in Fig. 4.6(b). In particular, the lattice temperature of the Cu atoms forming the nanofilament tends to be much higher than that of the left and right contacts and of the $SiO_2$ matrix. With this respect, Fig. 4.6(a) distinctly shows that a local hot spot is situated in the middle and second half of the filament, where the hottest atoms have temperatures up to 40 K larger than the average of their immediate surrounding. At high current densities these particles might acquire enough kinetic energy to change site, alter the filament geometry, and destroy its conductivity in the process. As a side note, it should be emphasized that classical simulation approaches using the same relation as above to model $T_{max}$ cannot properly describe the large, atomic-scale temperature variations of Fig. 4.6(a) and might underestimate the influence

of self-heating. Still, coupling molecular dynamics and QT calculations might be key to reveal the thermal stability of nanofilaments.

### 4.3.3 Influence of the Oxide Thickness on Self-Heating

The case study in the previous subsection provided an overview on the operation of a CBRAM cell in the ON-state on the basis of a single filament. In this third study, the impact of the oxide thickness on the same properties is presented. In addition to the Cu/a-$SiO_2$/Cu system with an $SiO_2$ SL of 3.5 nm, two additional structures with oxide thicknesses of 2.4 nm and 1.6 nm have been constructed [29]. All three memory cells share the same atomic configuration at the tip of the filament. The original and the two shorter cells are depicted in Figs. 4.7(a-c). As postulated, the electrical resistance only slightly depends on the oxide thickness with a 10% reduction when going from the 3.5 nm to the 1.6 nm a-$SiO_2$ layer, as can be seen in Fig. 4.7(d). This confirms the assumption that the current is mostly limited by the filament extremity, and not by its length. Consequently, self-heating can be compared between the three CBRAM cells using Eq. (4.2), which depends on the current and resistance. The temperature averaged over the device cross-section is displayed in Fig. 4.7(e) for all structures at an applied bias of 0.2 V. It is apparent that self-heating diminishes with the oxide thickness. Apart from a higher temperature profile in the thicker oxide, it can be noticed that the peak temperature moves further away from the contact towards the middle of the a-$SiO_2$ layer.

The magnitude of the self-heating depends on two factors according to Eq. (4.2): (1) the amount of power that is converted to heat ($\alpha R I^2$) and (2) the efficiency at which the excess phonon population can be extracted from the active region into the leads ($R_{th}$). To determine the origin of the increased temperature in thicker devices both effects are examined separately. The internal power dissipation factor $\alpha$ versus the oxide thickness is plotted in Fig. 4.7(f). It can be observed that this

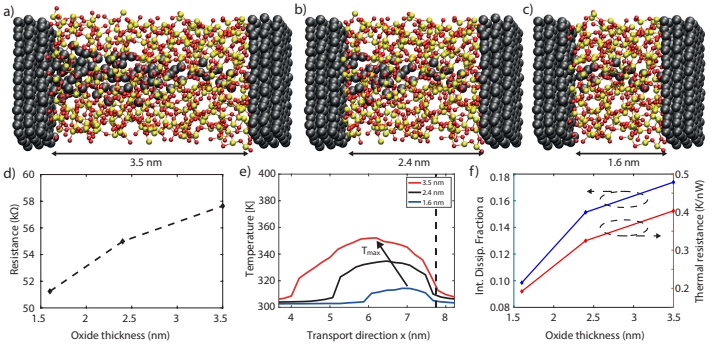


**Figure 4.7:** (a) CBRAM cell with an oxide thickness of 3.5 nm. (b) Same as (a), but with an oxide shortened from the left to 2.4 nm. The atomic configuration of the right-hand extremity of the filament is preserved. (c) CBRAM cell with an oxide layer shortened to 1.6 nm in the same fashion as (b). (d) Filament resistance as a function of the $SiO_2$ layer thickness for an external bias of 0.2 V. The diamonds mark the simulated values, the dashed line serves as a guide to the eye. (e) Average temperature along the transport axis of the three CBRAM cells with different $SiO_2$ layer thickness. For convenience, the filament tips are aligned at 7.6 nm. The vertical dashed line indicates the right device-electrode boundary. (f) Internal dissipation fraction $\alpha$ (left axis) and thermal resistance $R_{th}$ (right axis) as a function of the $SiO_2$ layer thickness.

quantity decreases with the oxide thickness. Hence, fewer phonons are emitted within the shorter cell. This behavior can be related to the time electrons spend propagating through the filament. The shorter the time, the lower the probability that an electron can interact with the surrounding lattice and dissipate energy.

To examine the thermal resistance of the device, $R_{th}$, this quantity is drawn as a function of the oxide thickness, also in Fig. 4.7(f). We find that $R_{th}$ exhibits the same characteristics as the power dissipation, i.e. $R_{th}$ is smaller in thinner oxides: the phonons emitted by electrons remain longer in thicker oxides because they need more time to escape. The fact that copper is an excellent heat conductor, whereas $SiO_2$ is not explains this behavior. As a result, the heat distribution is different in the three structures. The maximum temperature, which is located at the tip extremity in the shortest device, moves to the middle of the structure when the oxide thickness increases and requires therefore more time to escape.

As both $\alpha$ and $R_{th}$ increase with the oxide thickness, a strong dependence of self-heating on the dimensions of the SL is observed in CBRAM cells. If we assume that excessive joule heating is the main reason for device failure, which is commonly believed [189], the experimental findings of Ref. [29] can be readily explained: devices with a shorter oxide layer can endure larger currents before reaching the breakdown temperature. Stated differently, at a given current magnitude, self-heating is more pronounced in CBRAM cells with longer filaments because more phonons are emitted and they have more difficulties to leave the amorphous oxide region and attain the metallic contacts.

## 4.4 Conclusion

In this chapter, the QT characteristics of a Cu filament embedded within a Cu/a-$SiO_2$/Cu CBRAM cell have been examined. The transport prop-

erties have been studied both in the ballistic limit and in the presence of electron-phonon scattering. The electronic structure and the electron-phonon coupling strength were determined with the help of DFT, while the thermal properties were assessed based on a force-field approach. By coupling electron and phonon transport we could shed light on the mechanisms that limit the performance of such nanodevices.

By expelling filament atoms one-by-one we obtained discrete conductance levels during the ON-OFF switching of a CBRAM. These results are in qualitative agreement with experimental data. We could show that not all metal atoms of the filament contribute to the current flow of the structure. As a consequence, the presence or absence of atoms carrying little current does not affect the overall device conductance. Furthermore, we were able to demonstrate that memory cells with thin SL see a twofold benefit in terms of thermal stability. Firstly, the time needed for electrons to traverse the filament is shorter, which results in less scattering events and consequently less resistive heating. Secondly, the thermal resistance of shorter filaments is lower because the well conducting electrodes are closer to the heat source, thereby extracting excess phonons faster. We have also shown that the point of the highest temperature shifts further away from the tip of the filament, towards the middle of the oxide for thicker SLs.

Considering only a single metal for the active electrode (AE) and the counter electrode (CE) supposedly affects the thermal device resistance. Cu is an excellent heat conductor, therefore, a CE composed of Pt, W, or TiN should increase the overall device temperature and move $T_{max}$ further towards the CE. The diameter of the filament likely impacts the thermal resistance as well. The filaments considered in this work are extremely thin, while experimental imaging indicates that its diameter could be larger [190]. A wider filament would increase the thermal conductance of the base of the filament, shifting the $T_{max}$ towards the tip. Because of the competing effect of the approximations made in this work, parts of the induced errors might cancel each other, at least to a certain

degree. Which error dominates should be further investigated in future works.

Chapter 5

# Impact of the Counter Electrode Metal in Ag/$SiO_2$/M Cells[1]

## 5.1 Introduction

The CE plays a crucial role in the switching of CBRAM because it is inert and its atoms do not move during the SET and RESET processes. This "not-moving" is of utmost importance as a memory cell with two AEs could not be properly reset into a pristine state. Reversing the voltage polarity across the SL would simply supply $M^{z+}$ ions from the second AE, regrowing the filament in the reverse direction. Even though the CE is inert and does not undergo any reaction the choice of metal of this region affects the chemical environment at the CE-SL interface and thus the redox reactions on its surface. The direct impact of the electrode metal on the current, on the other hand, is unknown.

Constructing a CBRAM cell in which one electrode is different from the other requires a special treatment of the DFT domain because the symmetry is broken along the transport axis. Periodicity in the transverse directions requires that the cross-section is identical throughout the device. Two metals typically have different lattice constants, which breaks this requirement and creates a mismatch between electrodes. Three options are available to resolve this mismatch: First, to break the periodicity in the transverse directions, which would relief the need for matching

[1] This chapter is based on [50] and [191]

cells. This introduces material-vacuum interfaces all along the edges of the simulation domain, which greatly distorts the electronic structure. These surface effects could mask the properties of the filamentary structure we would like to study. Second, to strain one or both electrodes in a way that their cross-sections become identical. However, due to large mismatches in the lattice constants and crystal symmetries this typically distorts the band structure and renders this approach futile. Third, supercells can be constructed in a way that the dimensions of both materials agree. Large supercells are typically needed in this case. While this is a physically valid approach the computational cost of considering large structures quickly grows intractable. In practice, a combination of the second and third approach proves to be optimal. It is often possible to find a common cross-section with both a reasonable number of atoms as well as a bearable strain. Evidently, this becomes more and more difficult if several different crystals should be fitted within the same model to allow for direct comparison. Amorphous materials, on the other hand, are inherently disordered on a long range scale and have no fixed cross-section dimensions. The periodicity imposed by the boundary condition directly contradicts this notion of disorder. If the cell is sufficiently large, however, the electronic structure of such "quasi"-amorphous materials approaches the one of an amorphous state [192]. If grown carefully, the quasi-amorphous lattice constant can be tailored to any length and be made to match the dimensions of the contacts.

In the Cu/a-$SiO_2$/Cu CBRAM cells studied in Chapter 4 both metal contacts are composed of the same chemical species as the filament. The approximation to use the same active metal for both electrodes facilitates the structure construction because no cell matching is needed. It further reduces the computational burden as smaller cross-sections can be employed. One the one hand, no strain needs to be induced as the oxide is grown to match the size of the contact. On the other hand, the chemical environment is changed because of the absence of a third material and a symmetric configuration is used. Therefore, the impact

of this approximation needs to be carefully analyzed.

The aim of this chapter is to examine different metals as the counter CBRAM electrode and study their impact on the current distribution through the filamentary structure. To facilitate the distinction between the effect of a change of atomic configurations and the replacement of the CE metal, the atomic configuration in the oxide layer should not change. In reality, the choice of the electrode is likely to modify the shape of the filament, but this is not the subject of the present study.

## 5.2 Computational Details

The AE and the filament are composed of Ag atoms, while three different metals are used as CE, namely Ag, Pt, and W. To minimize the impact of strain on the results, the metal supercells that build the left and right contacts have been chosen such that their dimensions closely match. To do that, a cross-section of 2.5 x 2.4 $nm^2$ was selected. It ensures that the strain level does not exceed 1% for each electrode. The amorphous $SiO_2$ does not have a lattice constant and hence does not suffer from strain.

Multiple steps of AIMD and geometry optimization within DFT are needed to construct the modeled structures, as shown in Fig. 5.1(a-c). Prior to the DFT stage, the a-$SiO_2$ is obtained by a melt-and-quench approach using force-field molecular dynamics. A $SiO_2$ $\beta$-cristobalite crystal is melted at 5000 K for 550 ps. To speed up the melting process the volume is increased by a factor of 2.75 during the first 300 ps. Subsequently, the sample is rescaled to the target density of 2.2 $g/cm^3$ and annealed for another 250 ps. Then, the $SiO_2$ is cooled down to room temperature with a cooling rate of 40 K/ps and constant volume and lastly further annealed for 40 ps at 300 K. To start the DFT process, the Ag contacts are attached on two sides of the slab of 2 nm of a-$SiO_2$ and the latter optimized. During this process the contact atoms remain immobile. In

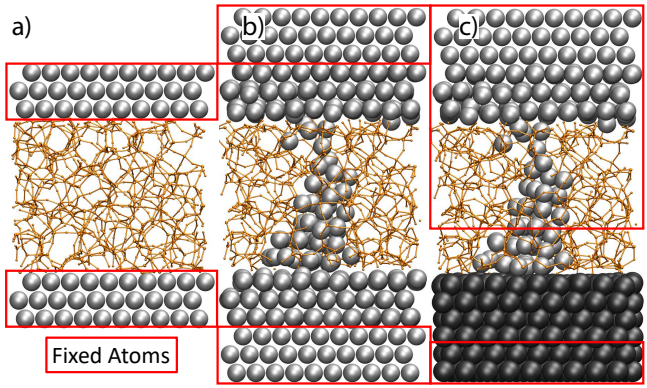


**Figure 5.1:** Assembly steps of the Ag/a-$SiO_2$/M CBRAM cell. The top contact is the active electrode made of Ag, the bottom contact the counter electrode made of Ag, W, or Pt. The a-$SiO_2$ matrix is represented by its bonds only. (a) An Ag contact is attached to the a-$SiO_2$, produced by a melt-and-quench approach, and the MIM structure is optimized with DFT. (b) Next, the filament is inserted and the contacts extended. The filament and the top metal layers are relaxed and annealed with AIMD. (c) Finally, the CE is replaced with an electrochemically inert metal, Pt or W. The CE surface and the bottom of the filament are further annealed.

order to reduce the computational cost, the same material is used for both contacts, eliminating the AE-CE metal interface periodic boundary conditions (PBC) would induce. Next, the filament is inserted by replacing all Si and O atoms within a cone by Ag ones. The system is again relaxed using optimization and AIMD for 1.5 ps, now including the top three atom layers of the electrode. After annealing, the filament is no longer perfectly conical, but assumes a shape that minimizes its potential energy. It displays an hourglass shape with a large cone on the CE and a small one on the AE side, as illustrated in Fig. 5.3(a). Finally, the bottom contact is replaced by an inert metal (Pt or W). To preserve the filament-AE interface, only the filaments basis near the CE and the top CE layers are annealed for 0.5 ps.

The force-field molecular dynamics simulations are performed with an empirical pairwise potential [193], as implemented in the QuantumATK 2017.1 package [131, 187], while the DFT calculations are done with CP2K [179] based on GTH pseudopotentials [181] and the PBE exchange-correlation functional [115]. For metals, DZVP basis sets [180] are employed, 3SP [188] for Si and O. To compute the Hamiltonian, the metal atoms are expressed in SZV basis sets instead of DZVP, as they provide accurate results at much lower computational cost, as illustrated in Chapter 4. Owing to the large cell size, only Γ-point sampling was used.

Based on the Hamiltonian and overlap matrices obtained from DFT, ballistic electron transport simulations are carried out via the OMEN quantum transport simulator [150, 154] relying on the NEGF formalism. Phonons and electron-phonon scattering are not considered here. The NEGF simulations for electrons are performed in the low field limit, i.e. Poisson's equation is not solved. Due to the localized nature of the Gaussian basis functions, the Hamiltonian has a sparse and banded form, as illustrated in Fig. 5.2: it resembles a tight-binding Hamiltonian. As PBC are used throughout the DFT calculations, special care is required to treat the boundaries due to the AE-CE metal-metal interface across

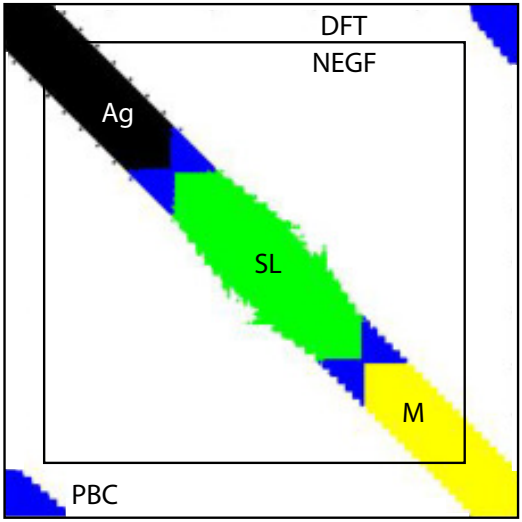


**Figure 5.2:** Visualization of the sparsity pattern of the Hamiltonian matrix corresponding to a CBRAM cell with two different electrodes. The black entries represent the Ag active electrode, the green ones the switching layer, which includes the Ag filament, and the yellow colored area depicts the counter electrode. The off-diagonal blocks are blue-colored and connect the materials. In the DFT simulation the two electrodes are linked by the two blocks in the bottom left and top right corner, while they are removed for the NEGF calculations.

the cell boundary. The nonzero entries in the Hamiltonian corresponding to the PBC are removed. As the metal-metal interface also affects the diagonal elements of the Hamiltonian close to the interface, the entries corresponding to the first few layers of metal atoms from the most left and rightest parts of the CBRAM cell are deleted, shrinking the size of the Hamiltonian. The extent of the interface is determined from the PBC blocks and is typically six layers. After the cut, both ends of the Hamiltonian correspond to bulk AE and CE materials, respectively. The overlap matrix is treated analogously.

## 5.3 Results

The energy-resolved transmission function around the Fermi energy is very similar in all CBRAM cells. It is reported in Fig. 5.3(b). This implies that the probability of electrons to be transmitted through the oxide layer is independent of the contact material and determined solely by the filamentary geometry. In ballistic simulations the current is directly related to the transmission through Eq. (2.23). Therefore, the low-field $I$-$V$ characteristics also resemble each other, Fig. 5.3(c), and confirm that the device current is mostly unaffected by the choice of the electrode metal. The conductance values extracted from linear fits to the current range from 0.13 $G_0$ for the Ag CE to 0.18 $G_0$ with the CE made of W, which is well in the range of experimental ON-state values [7]. The current behavior is largely ohmic, indicating that there is no open tunneling gap. Provided the same filamentary structure can be grown on different substrates, this demonstrates that the ON-state conductance is independent of the CE.

From a memory perspective the resistance state of the CBRAM cell is all that matters. The SET and RESET operations as well as stability, on the other hand, could well be affected by the distribution of the current within the cell. While experiments cannot directly probe this

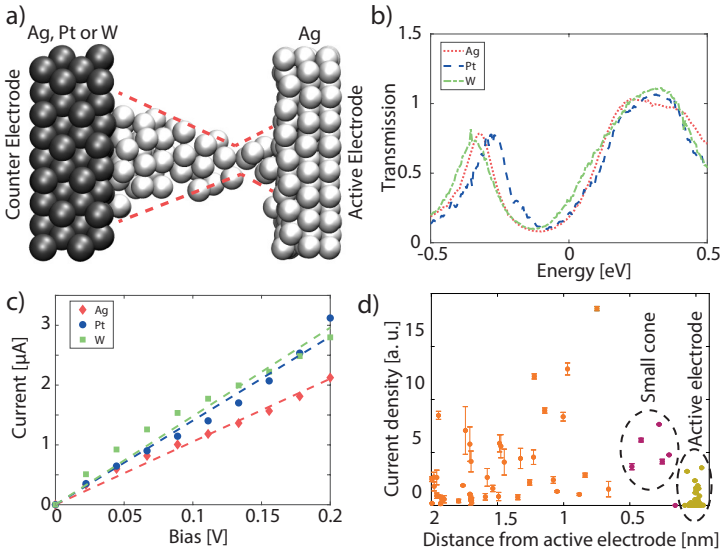


**Figure 5.3:** (a) Atomic configuration of a CBRAM cell with two different electrodes and an hourglass filament. White spheres represent silver atoms, the black ones tungsten. The red dashed line delimits the shape of the filament with a large cone on the left and a small one on the silver electrode. (b) Energy-resolved transmission function for the structures illustrated in (a) with Ag (red dotted), Pt (blue dashed), and W (green dash-dotted) as counter electrodes. (c) *I-V* characteristics of the three devices in (a) with Ag (red diamond), Pt (blue circles), and W (green squares) as counter electrode. Linear fits are provided for convenience (dashed lines). (d) Current magnitude on the individual silver atoms in the filament (orange and magenta squares) and on the surface of the active electrode (yellow squares). The error bars give a measure of the variance of the current between the three structures. The atoms belonging to the active electrode and the small cone (magenta colored) attached to it are indicated by the red circles.

property within a nanoscale structure, this can be done relatively easily in atomistic simulations. To assess the current distribution, its component along the (x,y,z) directions is computed for each atom. The expression for that can be inferred from Eq. (2.24): instead of summing all flows, each individual contribution is recorded and stored in the form of a current vector field. To allow for direct comparisons between the three device structures, the current vector fields are extracted at a constant current magnitude instead of a constant voltage.

The current magnitude passing through each individual filamentary Ag and electrode surface atom is plotted in Fig. 5.3(d). The dots refer to the mean value between the three simulations and the error bar measures the standard deviation. Three distinct regimes can be discerned: (1) in the AE there is very little spread in the current; (2) in the small cone situated on the AE, there is no variation in the current between the different structures as well; (3) some Ag atoms situated inside the large cone attached to the CE (Ag, Pt, W) show a large spread in current. From these results, it can be deduced that the nature of the interface between the {Ag, Pt, W}-electrode and the Ag filament influences the current distribution. The current in the Ag cone residing on the AE, on the other hand, is not affected. Because the conductance of the three cells is roughly constant, we infer that it is determined solely by the atomic configuration of the thinnest region of the filament. In other words, the filament resistance depends on its atomic morphology. The current density, on the other hand, is affected by the interface connecting the CE and the filament. Therefore, simulations considering two identical metal electrodes correctly predict the device resistance, but not the spatial distribution of the current. Electro-thermal effects, for instance, strongly depend on the current density, not only on its magnitude. Further studies will be needed to shed light on these effects and refine the results from Chapter 4.

## 5.4 Conclusion

In this chapter the influence of the material composition of the CE on the transport properties of Ag/a-$SiO_2$/M CBRAM cells in the ON-state has been studied. To this end, three models with a similar filamentary configuration, but different inert electrodes have been assembled and their ballistic QT characteristics have been investigated with the help of a DFT-based NEGF solver.

The choice of the CE material is found to exert little influence on the current magnitude, but affects its spatial distribution, in particular at the CE-SL interface. Therefore, simulations considering two identical metal electrodes correctly predict the device resistance, but not the current density within the filament of CBRAM cells. Electro-thermal effects, for instance, which depend on the current density, not only on its magnitude, could be affected by the choice of the CE. This study highlights the challenges associated with *ab initio* studies of entire nanoscale devices and emphasizes the limitations of simplified approaches.

The present analysis is limited to the ballistic regime. As future work, electro-thermal effects should be included and the atomically resolved temperature of the three investigated CBRAM cells calculated to identify possible failure mechanisms. Moreover, the atomic configuration of the filament, which was kept identical in all electrode configurations, is in reality likely affected by the choice of the CE. The filament growth process could be the subject of future studies that go beyond the scope of this thesis.

Chapter 6

# Conclusion and Outlook

## 6.1 Summary

In this thesis, out-of-equilibrium properties of CBRAM cells were calculated in a parameter-free manner by using *ab initio* quantum transport simulations. The combination of DFT and NEGF enabled us to perform these calculations on atomistic models of CBRAM cells in various resistance states and material compositions.

First, to mitigate the large computational burden of *ab initio* QT simulations an approach for matrix size reduction, the mode-space (MS) transformation, was generalized and benchmarked versus full sized real-space (RS) calculations. The local periodicity within the CBRAM models was exploited to successfully apply the transformation to nonhomogeneous device structures. This hybrid MS-RS scheme allowed for a reduction of the computational burden of QT calculations by two to three orders of magnitude, as compared to the traditional RS approach. Moreover, the procedure to obtain the MS transformation matrix was automatized, which both simplified the process and made it less time consuming.

Next, the DFT-based approach for calculating the ground-state electronic structure and the electron-phonon coupling strength was combined with a force-field baesed method to determine the dynamical matrix. Thereby, we obtained the required quantities for fully coupled electro-thermal QT calculations at the NEGF level of theory. When performing these simulations on a Cu filament embedded within a Cu/a-$SiO_2$/Cu

structure, an explanation for the improved stability of extremely thin CBRAM cells could be devised: First, thin switching layers (SL) only require a short filament to connect both electrodes. Consequently, electrons quickly traverse the SL, which results in fewer opportunities for electrons to interact with lattice vibrations, as compared to devices with thicker oxide layers. Secondly, the low thermal conductance of the filament and of the $SiO_2$ region is restricted to a tiny volume, enabling an efficient removal of excess phonons from the switching area. Combined, these two effects result in a strong reduction of the maximum temperature of ultra-scaled CBRAM cells at high current densities and thus improve their thermal stability.

Lastly, the impact of the contact material was investigated by assembling CBRAM models in which electrodes could be readily replaced by a different metal. We calculated the current density through the Ag filamentary structure embedded within a $SiO_2$ matrix with multiple CEs. The current magnitude was found to be independent of the contact material, suggesting that the atomic configuration of the filament is solely responsible for the device resistance state. The current density, on the other hand, was found to vary when changing the electrode metal. Consequently, a change in electrode metal may have important ramifications on processes that rely on the current density such as self-heating and redox reactions. While simplified device models with two equal electrodes can deliver informative qualitative results, asymmetric configurations will be needed for thorough investigations.

## 6.2 Outlook

In this thesis we have demonstrated the capability of currently available computational methods and resources to simulate physical processes and properties of entire memory cells comprising several 1000s of atoms from first principles. The developed methodology opens the possibility to pre-

dict and compare the transport characteristics of CBRAM cells without the need for experiments. However, the extreme computational burden inherent to such calculations limits them to the analysis of static cases, that is, the ON- or OFF-state. Switching processes can, in any case, only be approximated by simple models. Thus, to maximize the benefit of *ab initio* QT device calculations, they should be combined with other levels of modeling such as the finite element method for device investigation or compact models for circuit simulations. In both cases, our approach can supply material parameters to reduce the need for fitted values. Furthermore, the current implementation of the hybrid MS-RS technique has only been applied to ballistic simulations. Omitting scattering effects, however, might hide important effects, as demonstrated in this thesis. Therefore, the hybrid scheme should be extended to include electro-thermal effects that could benefit from an immense reduction of computational requirements.

It should finally be emphasized that the proposed *ab initio* simulation environment is not limited to CBRAM cells and was already successfully applied to several other applications, as indicated in Fig. 6.1. First, it was used to investigate the switching behavior of a nano-electromechanical relay. Such devices can be implemented by combining CBRAM-type switching with a mechanically controllable break junction (MCBJ) in a single environment [194]. Experimentally, both the active and counter electrodes are deposited on a flexible substrate. A metallic filament is grown between the two electrodes through a liquid electrolyte. A third input terminal is realized by a piezo element, which can bend the substrate and thereby rupture the metallic filament. The filament formation and fracturing has been simulated by implementing a pull and push scheme within DFT. An exemplary filament of the MCBJ in the ON-state is shown in Fig. 6.1(a). The device conductance, recorded at multiple intermediate steps during the switching process in Fig. 6.1(b), displays a mechanically-induced hysteresis that qualitatively agrees with experimental data [194].

The characteristics of conducting paths through the switching layer of valence change memory (VCM) cells has also been studied. The electrical conduction through amorphous $HfO_2$ (a-$HfO_2$) structures has so far only been evaluated in the ballistic limit of transport [101]. To investigate the impact of electron-phonon scattering on the electrical current flowing through amorphous switching layers we have assembled Pt/a-$HfO_2$/Pt VCM cells, as illustrated in Fig. 6.1(c). Oxygen vacancies were introduced by randomly removing 10% to 50% of the oxygen atoms from the $HfO_2$ layer. The current flowing through a system with 25% of the oxygen removed can be visualized in Fig. 6.1(d). We found that accounting for electron-phonon scattering can increase the device current at all considered concentrations of oxygen vacancies. This indicates that trap-assisted tunneling (TAT) is an important conduction mechanism, even at high concentrations of vacancies.

Two-dimensional (2D) semiconductors are promising candidates to implement field-effect transistors at the ultimate scaling limit [195]. The most common 2D insulator is hexagonal boron nitride (hBN) [196]. Its suitability as gate insulator, however, is questioned [197]. We computed the leakage current through crystalline hBN at the *ab initio* level in an Au/hBN/Si gate stack. The structure and the $I$-$V$ characteristics are both shown in Fig. 6.1(e-f). Defects in the hBN can induce molecular bridges between the individual layers of the 2D material [198], which could severely degrade the insulating properties of the material. As defects can be readily introduced into the atomistic model of the gate stack, their impact on the current will be examined in a future work.

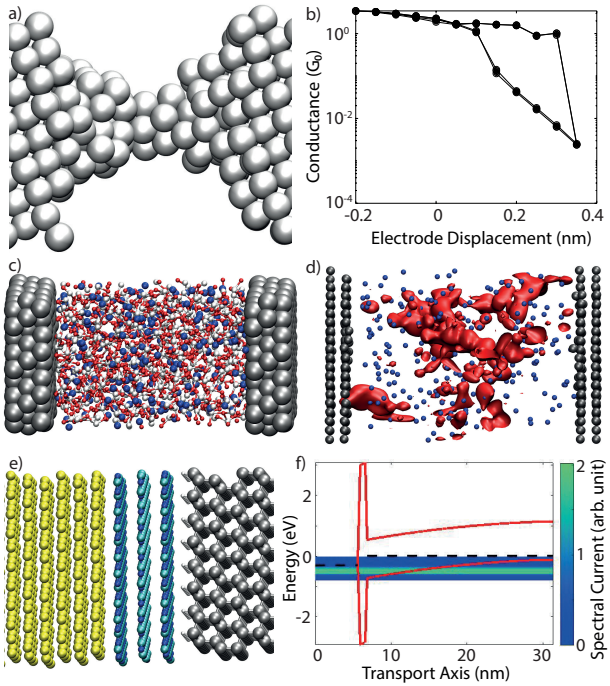


**Figure 6.1:** (a) Schematic view of a Sn-filament inserted in the middle of a mechanically controllable break junction. (b) The conductance of the Sn structure in (a) during two consecutive ON-OFF switching events where the distance between the electrode (horizontal axis) is first increased and then decreased. A hysteretic behavior can be observed. (c) Atomistic representation of a $Pt/HfO_2/Pt$ valence change memory cell. The gray spheres represent Pt, the white ones Hf, and the red ones O. Oxygen vacancies (blue spheres) have been randomly distributed. (d) Visualization of the electrical current flowing through the structure in (c). The Pt and oxygen vacancies are shown together with the current density (red). (e) Atomic structure of an Au/hBN/Si gate stack where the Au, B, N, Si, and H atoms are represented as yellow, cyan, blue, gray, and white spheres, respectively. (f) Energy- and position-resolved tunneling current flowing through the stack in (e) at an applied voltage of 0.3 V between the Si and Au electrodes. The band diagram is superimposed to the current as red lines, whereas the Fermi energies of the Au and Si contacts are marked by the black dashed line.

# References


[1] W. A. Wulf and S. A. McKee, "Hitting the memory wall", *ACM SIGARCH Computer Architecture News*, vol. 23, no. 1, pp. 20–24, 1995 (see p. 1).

[2] M. M. Aly, M. Gao, G. Hills, C. S. Lee, G. Pitner, M. M. Shulaker, T. F. Wu, M. Asheghi, J. Bokor, F. Franchetti, K. E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. Ré, H. S. Wong, and S. Mitra, "Energy-efficient abundant-data computing: The N3XT 1,000", *Computer*, vol. 48, no. 12, pp. 24–33, 2015 (see p. 1).

[3] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics", *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018 (see p. 2).

[4] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications", *Solid-State Electronics*, vol. 125, pp. 25–38, 2016 (see pp. 2, 71).

[5] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems", *Nature Electronics*, vol. 1, no. 1, pp. 22–29, 2018 (see p. 2).

[6] B. Hudec, C. W. Hsu, I. T. Wang, W. L. Lai, C. C. Chang, T. Wang, K. Fröhlich, C. H. Ho, C. H. Lin, and T. H. Hou, "3D resistive RAM cell design for high-density storage class memory—a review", *Science China Information Sciences*, vol. 59, no. 6, pp. 1–21, 2016 (see p. 2).

[7] R. Waser, R. Dittmann, C. Staikov, and K. Szot, *Redox-based resistive switching memories nanoionic mechanisms, prospects, and challenges*, Jul. 2009 (see pp. 2, 6, 81, 97).

[8] B. Govoreanu, G. S. Kar, Y. Y. Chen, *et al.*, "10x10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation", *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 729–732, 2011 (see p. 2).

[9] M. A. Zidan, H. Omran, R. Naous, A. Sultan, H. A. Fahmy, W. D. Lu, and K. N. Salama, "Single-Readout High-Density Memristor Crossbar", *Scientific Reports*, vol. 6, pp. 2–10, 2016 (see p. 2).

[10] K. H. Kim, S. Hyun Jo, S. Gaba, and W. Lu, "Nanoscale resistive memory with intrinsic diode characteristics and long endurance", vol. 96, no. 5, pp. 2–5, 2010 (see p. 2).

[11] J. Zhou, F. Cai, Q. Wang, B. Chen, S. Gaba, and W. D. Lu, "Very low-programming-current RRAM with self-rectifying characteristics", *IEEE Electron Device Letters*, vol. 37, no. 4, pp. 404–407, 2016 (see p. 2).

[12] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, "Sub-nanosecond switching of a tantalum oxide memristor", *Nanotechnology*, vol. 22, no. 48, p. 485 203, 2011 (see p. 2).

[13] L. Goux, K. Sankaran, G. Kar, N. Jossart, K. Opsomer, R. Degraeve, G. Pourtois, G. M. Rignanese, C. Detavernier, S. Clima, Y. Y. Chen, A. Fantini, B. Govoreanu, D. J. Wouters, M. Jurczak, L. Altimime, and J. A. Kittl, "Field-driven ultrafast sub-ns programming in W\Al 2O 3\Ti\CuTe-based 1T1R CBRAM system", *Digest of Technical Papers - Symposium on VLSI Technology*, 2012, pp. 69–70 (see pp. 2, 5).

[14] C. Nail, G. Molas, P. Blaise, *et al.*, "Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations", *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 1–4, 2017 (see pp. 2, 5).

[15] D. Alfaro Robayo, G. Sassine, Q. Rafhay, G. Ghibaudo, G. Molas, and E. Nowak, "Endurance Statistical Behavior of Resistive Memories Based on Experimental and Theoretical Investigation", *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3318–3325, 2019 (see pp. 2, 5).

[16] F. Pan, S. Gao, C. Chen, C. Song, and F. Zeng, "Recent progress in resistive random access memories: Materials, switching mechanisms, and performance", *Materials Science and Engineering R: Reports*, vol. 83, no. 1, pp. 1–59, 2014 (see pp. 2, 8, 10).

[17] D. Ielmini, "Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling", *Semiconductor Science and Technology*, vol. 31, no. 6, pp. 063 002–063 027, 2016 (see pp. 2, 85).

[18] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang, and H. Qian, "Conduction mechanisms, dynamics and stability in ReRAMs", *Microelectronic Engineering*, vol. 187-188, pp. 121–133, Feb. 2018 (see pp. 2, 10).

[19] E. Ambrosi, P. Bartlett, A. I. Berg, *et al.*, *Electrochemical metallization ReRAMs (ECM) - Experiments and modelling: General discussion*, 2019 (see p. 3).

[20] M. Lanza, H. S. Wong, E. Pop, *et al.*, "Recommended Methods to Study Resistive Switching Devices", *Advanced Electronic Materials*, vol. 5, no. 1, pp. 1–28, 2019 (see p. 3).

[21] C. Funck, A. Marchewka, C. Bäumer, P. C. Schmidt, P. Müller, R. Dittmann, M. Martin, R. Waser, and S. Menzel, "A Theoretical and Experimental View on the Temperature Dependence of the Electronic Conduction through a Schottky Barrier in a Resistively Switching SrTiO3-Based Memory Cell", *Advanced Electronic Materials*, vol. 4, no. 7, pp. 1–12, Jul. 2018 (see pp. 3, 14).

[22] J. Guy, G. Molas, C. Cagli, M. Bernard, A. Roule, C. Carabasse, A. Toffoli, F. Clermidy, B. De Salvo, and L. Perniola, "Guidance to reliability improvement in CBRAM using advanced KMC modelling", *IEEE International Reliability Physics Symposium Proceedings*, PM2.1–PM2.5, 2017 (see p. 3).

[23] D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator", *IEEE Transactions on Electron Devices*, vol. 16, no. 1, pp. 64–77, 1969 (see p. 3).

[24] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, "A review of hydrodynamic and energy-transport models for semiconductor device simulation", *Proceedings of the IEEE*, vol. 91, no. 2, p. 249, 2003 (see p. 4).

[25] P. Hohenberg, "Inhomogeneous Electron Gas", *Physical Review*, vol. 136, no. 3B, B864–B871, 1964 (see pp. 4, 17).

[26] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects", *Physical Review*, vol. 140, no. 4A, A1133–A1138, 1965 (see pp. 4, 17, 22, 28).

[27] L. P. Kadanoff and G. Baym, *Quantum statistical mechanics: Green's function methods in equilibrium and nonequilibrium problems*. New York: Benjamin, 2018, pp. 1–203 (see pp. 4, 18, 26, 32).

[28] L. V. Keldysh, "DIAGRAM TECHNIQUE FOR NONEQUILIBRIUM PROCESSES", *SOVIET PHYSICS JETP*, vol. 20, no. 4, pp. 1018–1026, 1965 (see pp. 4, 18, 26).

[29] B. Cheng, A. Emboras, Y. Salamin, F. Ducry, P. Ma, Y. Fedoryshyn, S. Andermatt, M. Luisier, and J. Leuthold, "Ultra compact electrochemical metallization cells offering reproducible atomic scale memristive switching", *Communications Physics*, vol. 2, no. 1, pp. 1–9, 2019 (see pp. 4, 5, 12, 71, 73, 86, 88).

[30] J. Taylor, H. Guo, and J. Wang, "Ab initio modeling of quantum transport properties of molecular electronic devices", *Physical Review B - Condensed Matter and Materials Physics*, vol. 63, no. 24, pp. 1–13, 2001 (see pp. 4, 18, 26, 28, 34, 63).

[31] T. Frederiksen, M. Brandbyge, N. Lorente, and A. P. Jauho, "Inelastic scattering and local heating in atomic gold wires", *Physical Review Letters*, vol. 93, no. 25, 2004 (see pp. 4, 27).

[32] F. Argall, "Switching phenomena in titanium oxide thin films", *Solid State Electronics*, vol. 11, no. 5, pp. 535–541, 1968 (see p. 4).

[33] Y. Hirose and H. Hirose, "Polarity-dependent memory switching and behavior of Ag dendrite in Ag-photodoped amorphous As2S3 films", *Journal of Applied Physics*, vol. 47, no. 6, pp. 2767–2772, 1976 (see p. 4).

[34] B. Swaroop, W. C. West, G. Martinez, M. N. Kozicki, and L. A. Akers, "Programmable current mode Hebbian learning neural network using Programmable Metallization Cell", *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 3, pp. 33–36, 1998 (see p. 5).

[35] M. N. Kozicki, M. Yun, and A. Singh, "Applications of programmable resistance changes in metal-doped chalcogenides", *Solid-State Ionic Devices*, E. D. Wachsman, M. Liu, J. R. Akridge, and N. Yamazoe, Eds., Seattle, USA: The Electrochemical Society, 1999, pp. 298–309 (see p. 5).

[36] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found", *Nature*, vol. 453, no. 7191, pp. 80–83, 2008 (see p. 5).

[37] L. Chua, "Memristor - The Missing Circuit Element", *IEEE Transactions on Circuit Theory*, vol. 18, no. 5, pp. 507–519, 1971 (see p. 5).

[38] A. Bricalli, E. Ambrosi, M. Laudato, M. Maestro, R. Rodriguez, and D. Ielmini, "SiOx-based resistive switching memory (RRAM) for crossbar storage/select elements with high on/off ratio", *Technical Digest - International Electron Devices Meeting, IEDM*, IEEE, 2017, pp. 1–4 (see p. 5).

[39] J. Guy, G. Molas, E. Vianello, *et al.*, "Investigation of the physical mechanisms governing data-retention in down to 10nm nanotrench Al2O3/CuTeGe conductive bridge RAM (CBRAM)", *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 742–745, 2013 (see p. 5).

[40] Y. Wang, C. Zhang, H. Yu, and W. Zhang, "Design of low power 3D hybrid memory by non-volatile CBRAM-crossbar with block-level data-retention", *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 197–202, 2012 (see p. 5).

[41] D. Jana, S. Roy, R. Panja, M. Dutta, S. Z. Rahaman, R. Mahapatra, and S. Maikap, "Conductive-bridging random access memory: challenges and opportunity for 3D architecture", *Nanoscale Research Letters*, vol. 10, no. 1, Dec. 2015 (see p. 5).

[42] F. G. Aga, J. Woo, S. Lee, J. Song, J. Park, J. Park, S. Lim, C. Sung, and H. Hwang, "Retention modeling for ultra-thin density of Cu-based conductive bridge random access memory (CBRAM)", *AIP Advances*, vol. 6, no. 2, 2016 (see p. 5).

[43] C. Gopalan, Y. Ma, T. Gallo, J. Wang, E. Runnion, J. Saenz, F. Koushan, P. Blanchard, and S. Hollmer, "Demonstration of Conductive Bridging Random Access Memory (CBRAM) in logic CMOS process", *Solid-State Electronics*, vol. 58, no. 1, pp. 54–61, 2011 (see p. 5).

[44] R. Fackenthal, M. Kitagawa, W. Otsuka, K. Prall, D. Mills, K. Tsutsui, J. Javanifard, K. Tedrow, T. Tsushima, Y. Shibahara,

and G. Hush, "A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology", *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 57, IEEE, 2014, pp. 338–339 (see pp. 5, 9).

[45] J. R. Jameson, J. Dinh, N. Gonzales, S. Hollmer, S. Hsu, D. Kim, F. Koushan, D. Lewis, E. Runnion, J. Shields, A. Tysdal, D. Wang, and V. Gopinath, "Towards automotive grade embedded RRAM", *European Solid-State Device Research Conference*, vol. 2018-Septe, pp. 58–61, 2018 (see pp. 5, 9).

[46] A. Emboras, I. Goykhman, B. Desiatov, N. Mazurski, L. Stern, J. Shappir, and U. Levy, "Nanoscale plasmonic memristor with optical readout functionality", *Nano Letters*, vol. 13, no. 12, pp. 6151–6155, 2013 (see p. 6).

[47] A. Emboras, J. Niegemann, P. Ma, C. Haffner, A. Pedersen, M. Luisier, C. Hafner, T. Schimmel, and J. Leuthold, "Atomic Scale Plasmonic Switch", *Nano Letters*, vol. 16, no. 1, pp. 709–714, 2016 (see p. 6).

[48] A. Emboras, A. Alabastri, F. Ducry, B. Cheng, Y. Salamin, P. Ma, S. Andermatt, B. Baeuerle, A. Josten, C. Hafner, M. Luisier, P. Nordlander, and J. Leuthold, "Atomic Scale Photodetection Enabled by a Memristive Junction", *ACS Nano*, vol. 12, no. 7, pp. 6706–6713, 2018 (see pp. 6, 80).

[49] J. H. Cha, S. Y. Yang, J. Oh, S. Choi, S. Park, B. C. Jang, W. Ahn, and S. Y. Choi, "Conductive-bridging random-access memories for emerging neuromorphic computing", *Nanoscale*, vol. 12, no. 27, pp. 14 339–14 368, 2020 (see p. 6).

[50] F. Ducry, J. Aeschlimann, and M. Luisier, "Electro-thermal transport in disordered nanostructures: A modeling perspective", *Nanoscale Advances*, vol. 2, no. 7, pp. 2648–2667, 2020 (see pp. 6, 10, 26, 71, 91).

[51] U. Russo, D. Kamalanathan, D. Ielmini, A. L. Lacaita, and M. N. Kozicki, "Study of multilevel programming in Programmable Metallization Cell (PMC) memory", *IEEE Transactions on Electron Devices*, vol. 56, no. 5, pp. 1040–1047, 2009 (see p. 7).

[52] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories - Fundamentals, applications, prospects", *Nanotechnology*, vol. 22, no. 25, p. 254 003, 2011 (see p. 7).

[53] W. Banerjee, "Challenges and Applications of Emerging Nonvolatile Memory Devices", *Electronics*, vol. 9, no. 6, p. 1029, 2020 (see pp. 8, 9).

[54] M. N. Kozicki and H. J. Barnaby, "Conductive bridging random access memory - Materials, devices and applications", *Semiconductor Science and Technology*, vol. 31, no. 11, Oct. 2016 (see p. 9).

[55] J. R. Jameson, P. Blanchard, J. Dinh, *et al.*, "Conductive Bridging RAM (CBRAM): Then, Now, and Tomorrow", *ECS Transactions*, vol. 75, no. 5, pp. 41–54, 2016 (see p. 9).

[56] S. Lee, J. Song, S. Lim, S. A. Chekol, and H. Hwang, "Excellent data retention characteristic of Te-based conductive-bridge RAM using semiconducting Te filament for storage class memory", *Solid-State Electronics*, vol. 153, no. December 2018, pp. 8–11, 2019 (see p. 9).

[57] J. H. Yoon, J. Zhang, P. Lin, N. Upadhyay, P. Yan, Y. Liu, Q. Xia, and J. J. Yang, "A Low-Current and Analog Memristor with Ru as Mobile Species", *Advanced Materials*, vol. 32, no. 9, pp. 1–9, 2020 (see p. 9).

[58] J. Sun, Q. Liu, H. Xie, X. Wu, F. Xu, T. Xu, S. Long, H. Lv, Y. Li, L. Sun, and M. Liu, "In situ observation of nickel as an oxidizable electrode material for the solid-electrolyte-based resistive random

access memory", *Applied Physics Letters*, vol. 102, no. 5, 2013 (see p. 9).

[59] C. Pearson, L. Bowen, M. W. Lee, A. L. Fisher, K. E. Linton, M. R. Bryce, and M. C. Petty, "Focused ion beam and field-emission microscopy of metallic filaments in memory devices based on thin films of an ambipolar organic compound consisting of oxadiazole, carbazole, and fluorene units", *Applied Physics Letters*, vol. 102, no. 21, 2013 (see p. 9).

[60] P. Peng, D. Xie, Y. Yang, Y. Zang, X. Gao, C. Zhou, T. Feng, H. Tian, T. Ren, and X. Zhang, "Resistive switching behavior in diamond-like carbon films grown by pulsed laser deposition for resistance switching random access memory application", *Journal of Applied Physics*, vol. 111, no. 8, 2012 (see p. 9).

[61] Z. Wang, P. B. Griffin, J. McVittie, S. Wong, P. C. McIntyre, and Y. Nishi, "Resistive switching mechanism in ZNxCd1-xS nonvolatile memory devices", *IEEE Electron Device Letters*, vol. 28, no. 1, pp. 14–16, 2007 (see p. 9).

[62] X. Liu, S. M. Sadaf, S. Park, S. Kim, E. Cha, D. Lee, G. Y. Jung, and H. Hwang, "Complementary resistive switching in niobium oxide-based resistive memory devices", *IEEE Electron Device Letters*, vol. 34, no. 2, pp. 235–237, 2013 (see p. 9).

[63] C. N. Peng, C. W. Wang, T. C. Chan, W. Y. Chang, Y. C. Wang, H. W. Tsai, W. W. Wu, L. J. Chen, and Y. L. Chueh, "Resistive switching of Au/ZnO/Au resistive memory: An in situ observation of conductive bridge formation", *Nanoscale Research Letters*, vol. 7, pp. 1–6, 2012 (see p. 9).

[64] W. Devulder, K. Opsomer, J. Meersschaut, D. Deduytsche, M. Jurczak, L. Goux, and C. Detavernier, "Combinatorial study of ag-te thin films and their application as cation supply layer in CBRAM

cells", *ACS Combinatorial Science*, vol. 17, no. 5, pp. 334–340, 2015 (see p. 9).

[65] Y. T. Tseng, I. C. Chen, T. C. Chang, J. C. Huang, C. C. Shih, H. X. Zheng, W. C. Chen, M. H. Wang, W. C. Huang, M. C. Chen, X. H. Ma, Y. Hao, and S. M. Sze, "Enhanced electrical behavior from the galvanic effect in Ag-Cu alloy electrode conductive bridging resistive switching memory", *Applied Physics Letters*, vol. 113, no. 5, pp. 1–6, 2018 (see p. 9).

[66] L. Gao, Z. Song, Y. Li, G. Qian, and F. Ma, "Controllable growth of conductive filaments in sandwiched CBRAM cells using self-assembled Cu/W multilayer thin films as the electrodes", *Journal of Alloys and Compounds*, vol. 803, pp. 601–609, 2019 (see p. 9).

[67] F. Yuan, Z. Zhang, C. Liu, F. Zhou, H. M. Yau, W. Lu, X. Qiu, H. S. Wong, J. Dai, and Y. Chai, "Real-Time Observation of the Electrode-Size-Dependent Evolution Dynamics of the Conducting Filaments in a SiO2 Layer", *ACS Nano*, vol. 11, no. 4, pp. 4097–4104, 2017 (see p. 9).

[68] R. Cao, S. Liu, Q. Liu, X. Zhao, W. Wang, X. Zhang, F. Wu, Q. Wu, Y. Wang, H. Lv, S. Long, and M. Liu, "Improvement of Device Reliability by Introducing a BEOL-Compatible TiN Barrier Layer in CBRAM", *IEEE Electron Device Letters*, vol. 38, no. 10, pp. 1371–1374, 2017 (see p. 9).

[69] M. N. Kozicki and M. Mitkova, "Mass transport in chalcogenide electrolyte films - materials and applications", *Journal of Non-Crystalline Solids*, vol. 352, no. 6-7 SPEC. ISS. Pp. 567–577, 2006 (see p. 10).

[70] V. Sousa, "Chalcogenide materials and their application to Non-Volatile Memories", *Microelectronic Engineering*, vol. 88, no. 5, pp. 807–813, 2011 (see p. 10).

[71] K. H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications", *Nano Letters*, vol. 12, no. 1, pp. 389–395, 2012 (see p. 10).

[72] D. Kumar, R. Aluguri, U. Chand, and T. Y. Tseng, "Enhancement of resistive switching properties in nitride based CBRAM device by inserting an Al2O3 thin layer", *Applied Physics Letters*, vol. 110, no. 20, pp. 1–6, 2017 (see p. 10).

[73] Y. R. Jeon, Y. Abbas, A. S. Sokolov, S. Kim, B. Ku, and C. Choi, "Study of in Situ Silver Migration in Amorphous Boron Nitride CBRAM Device", *ACS Applied Materials and Interfaces*, 2019 (see p. 10).

[74] M. Barci, G. Molas, C. Cagli, E. Vianello, M. Bernard, A. Roule, A. Toffoli, J. Cluzel, B. De Salvo, and L. Perniola, "Bilayer metal-oxide conductive bridge memory technology for improved window margin and reliability", *IEEE Journal of the Electron Devices Society*, vol. 4, no. 5, pp. 314–320, 2016 (see p. 10).

[75] S. Ambrogio, B. Magyari-Köpe, N. Onofrio, M. Mahbubul Islam, D. Duncan, Y. Nishi, and A. Strachan, "Modeling resistive switching materials and devices across scales", *Journal of Electroceramics*, vol. 39, no. 1-4, pp. 39–60, 2017 (see p. 10).

[76] L. Larcher and A. Padovani, "Multiscale modeling of oxide RRAM devices for memory applications: from material properties to device performance", *Journal of Computational Electronics*, vol. 16, no. 4, pp. 1077–1084, Dec. 2017 (see pp. 10, 11).

[77] S. Menzel, "Comprehensive modeling of electrochemical metallization memory cells", *Journal of Computational Electronics*, vol. 16, no. 4, pp. 1017–1037, Dec. 2017 (see pp. 10, 11).

[78] W. Wang, M. Laudato, E. Ambrosi, A. Bricalli, E. Covi, Y. H. Lin, and D. Ielmini, "Volatile Resistive Switching Memory Based on Ag Ion Drift/Diffusion Part I: Numerical Modeling", *IEEE Transactions on Electron Devices*, vol. 66, no. 9, pp. 3795–3801, Sep. 2019 (see p. 11).

[79] W. Wang, M. Laudato, E. Ambrosi, A. Bricalli, E. Covi, Y. H. Lin, and D. Ielmini, "Volatile Resistive Switching Memory Based on Ag Ion Drift/Diffusion - Part II: Compact Modeling", *IEEE Transactions on Electron Devices*, vol. 66, no. 9, pp. 3802–3808, Sep. 2019 (see p. 11).

[80] W. Wang, M. Wang, E. Ambrosi, A. Bricalli, M. Laudato, Z. Sun, X. Chen, and D. Ielmini, "Surface diffusion-limited lifetime of silver and copper nanofilaments in resistive switching devices", *Nature Communications*, vol. 10, no. 1, p. 81, Dec. 2019 (see p. 11).

[81] W. M. Young and E. W. Elcock, "Monte Carlo studies of vacancy migration in binary ordered alloys: I", *Proceedings of the Physical Society*, vol. 89, no. 3, pp. 735–746, 1966 (see p. 11).

[82] S. Menzel, P. Kaupmann, and R. Waser, "Understanding filamentary growth in electrochemical metallization memory cells using kinetic Monte Carlo simulations", *Nanoscale*, vol. 7, no. 29, pp. 12 673–12 681, Aug. 2015 (see p. 11).

[83] S. Dirkmann and T. Mussenbrock, "Resistive switching in memristive electrochemical metallization devices", *AIP Advances*, vol. 7, no. 6, Jun. 2017 (see pp. 11, 12).

[84] S. Dirkmann, J. Kaiser, C. Wenger, and T. Mussenbrock, "Filament Growth and Resistive Switching in Hafnium Oxide Memristive Devices", *ACS Applied Materials and Interfaces*, vol. 10, no. 17, pp. 14 857–14 868, May 2018 (see p. 12).

[85] M. P. Allen, "An Introduction to Molecular Dynamics Simulation", *Computer Simulations of Surfaces and Interfaces*, vol. 23, Forschungszentrum Jülich, 2004, pp. 1–28 (see p. 12).

[86] L. P. Wang, T. J. Martinez, and V. S. Pande, "Building force fields: An automatic, systematic, and reproducible approach", *Journal of Physical Chemistry Letters*, vol. 5, no. 11, pp. 1885–1891, Jun. 2014 (see p. 12).

[87] N. Onofrio, D. Guzman, and A. Strachan, "Atomic origin of ultrafast resistance switching in nanoscale electrometallization cells", *Nature Materials*, vol. 14, no. 4, pp. 440–446, 2015 (see pp. 12, 13, 75, 76).

[88] E. A. Chagarov and A. C. Kummel, "Ab initio molecular dynamics simulations of properties of a- Al 2O3 /vacuum and a-Zr O2 /vacuum vs a- Al 2O3 Ge (100) (2x1) and a-Zr O2 Ge (100) (2x1) interfaces", *Journal of Chemical Physics*, vol. 130, no. 12, p. 124 717, 2009 (see pp. 12, 74).

[89] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, "ReaxFF: A reactive force field for hydrocarbons", *Journal of Physical Chemistry A*, vol. 105, no. 41, pp. 9396–9409, Oct. 2001 (see pp. 13, 76).

[90] S. Watanabe and B. Xiao, "Atomistic Simulations for Understanding Microscopic Mechanism of Resistive Switches", *Atomic Switch*, Springer, 2020, pp. 95–125 (see p. 13).

[91] S. C. Pandey, R. Meade, and G. S. Sandhu, "Cu impurity in insulators and in metal-insulator-metal structures: Implications for resistance-switching random access memories", *Journal of Applied Physics*, vol. 117, no. 5, Feb. 2015 (see p. 13).

[92] N. Onofrio, D. Guzman, and A. Strachan, "Atomistic simulations of electrochemical metallization cells: Mechanisms of ultra-fast

resistance switching in nanoscale devices", *Nanoscale*, vol. 8, no. 29, pp. 14 037–14 047, 2016 (see p. 13).

[93] K. Sankaran, L. Goux, S. Clima, M. Mees, J. A. Kittl, M. Jurczak, L. Altimime, G.-M. Rignanese, and G. Pourtois, "Modeling of Copper Diffusion in Amorphous Aluminum Oxide in CBRAM Memory Stack", *ECS Transactions*, vol. 45, no. 3, pp. 317–330, 2012 (see p. 13).

[94] B. Xiao, X. F. Yu, and J. B. Cheng, "Atomic insight into the origin of various operation voltages of cation-based resistance switches", *ACS Applied Materials and Interfaces*, vol. 8, no. 46, pp. 31 978–31 985, 2016 (see p. 13).

[95] K. Patel, J. Cottom, M. Bosman, A. J. Kenyon, and A. L. Shluger, "An oxygen vacancy mediated Ag reduction and nucleation mechanism in SiO2 RRAM devices", *Microelectronics Reliability*, vol. 98, no. September 2018, pp. 144–152, 2019 (see p. 14).

[96] Q. Hu, Z. Fan, A. Huang, X. Zhang, R. Zhao, Q. Gao, Y. Ji, W. Dou, M. Wang, H. Shi, Z. Xiao, X. Jiang, and P. K. Chu, "Competitive conductive mechanism of interstitial Ag and oxygen vacancies in Ag/Ta2O5/Pt stack", *Journal of Applied Physics*, vol. 126, no. 6, p. 065 104, 2019 (see p. 14).

[97] C. Nail, P. Blaise, G. Molas, M. Bernard, A. Roule, A. Toffoli, L. Perniola, and C. Vallée, "Atomistic mechanisms of copper filament formation and composition in Al2O3-based conductive bridge random access memory", *Journal of Applied Physics*, vol. 122, no. 2, 2017 (see p. 14).

[98] B. Xiao, T. Gu, T. Tada, and S. Watanabe, "Conduction paths in Cu/amorphous-Ta2O5/Pt atomic switch: First-principles studies", *Journal of Applied Physics*, vol. 115, no. 3, 2014 (see p. 14).

[99] H. Nakamura and Y. Asai, "Competitive effects of oxygen vacancy formation and interfacial oxidation on an ultra-thin HfO2-based resistive switching memory: Beyond filament and charge hopping models", *Physical Chemistry Chemical Physics*, vol. 18, no. 13, pp. 8820–8826, 2016 (see p. 14).

[100] Z. Wang, S. Tsukimoto, M. Saito, and Y. Ikuhara, "Quantum electron transport through SrTiO3: Effects of dopants on conductance channel", *Applied Physics Letters*, vol. 94, no. 25, pp. 1–4, 2009 (see p. 14).

[101] X. Cartoixa, R. Rurali, and J. Sune, "Transport properties of oxygen vacancy filaments in metal/crystalline or amorphous HfO2/metal structures", *Physical Review B - Condensed Matter and Materials Physics*, vol. 86, no. 16, Oct. 2012 (see pp. 14, 104).

[102] O. Pirrotta, A. Padovani, L. Larcher, L. Zhao, B. Magyari-Kope, and Y. Nishi, "Multi-scale modeling of oxygen vacancies assisted charge transport in sub-stoichiometric TiOx for RRAM application", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, pp. 37–40, 2014 (see p. 14).

[103] M. Andrä, C. Funck, N. Raab, M. A. Rose, M. Vorokhta, F. Dvorák, B. Šmíd, V. Matolín, D. N. Mueller, R. Dittmann, R. Waser, S. Menzel, and F. Gunkel, "Effect of Cationic Interface Defects on Band Alignment and Contact Resistance in Metal/Oxide Heterojunctions", *Advanced Electronic Materials*, vol. 6, no. 1, 2020 (see p. 14).

[104] H. Nakamura, I. Rungger, S. Sanvito, N. Inoue, J. Tominaga, and Y. Asai, "Resistive switching mechanism of GeTe-Sb2Te3 interfacial phase change memory and topological properties of embedded two-dimensional states", *Nanoscale*, vol. 9, no. 27, pp. 9386–9395, 2017 (see p. 14).

[105] N. Inoue and H. Nakamura, "Structural transition pathway and bipolar switching of the GeTe-Sb2Te3 superlattice as interfacial phase-change memory", *Faraday Discussions*, vol. 213, pp. 303–319, 2019 (see p. 14).

[106] C. S. Lent and D. J. Kirkner, "The quantum transmitting boundary method", *Journal of Applied Physics*, vol. 67, no. 10, pp. 6353–6359, 1990 (see pp. 18, 30).

[107] C. Kittel, *Introduction to Solid State Physics*, 8th Editio. Wiley, 2004 (see pp. 18, 29, 37).

[108] J. M. Luttinger and W. Kohn, "Motion of Electrons and Holes in Perturbed Periodic Fields", *Physical Review*, vol. 97, no. 4, pp. 869–883, 1955 (see p. 18).

[109] J. C. Slater and G. F. Koster, "Simplified LCAO method for the periodic potential problem", *Physical Review*, vol. 94, no. 6, pp. 1498–1524, 1954 (see pp. 18, 29).

[110] R. J. Bartlett and M. Musiał, "Coupled-cluster theory in quantum chemistry", *Reviews of Modern Physics*, vol. 79, no. 1, pp. 291–352, 2007 (see p. 20).

[111] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals", *Molecular Physics*, vol. 115, no. 19, pp. 2315–2372, 2017 (see p. 22).

[112] B. Kanungo, P. M. Zimmerman, and V. Gavini, "Exact exchange-correlation potentials from ground-state electron densities", *Nature Communications*, vol. 10, no. 1, 2019 (see p. 22).

[113] D. Chakraborty, S. Kar, and P. K. Chattaraj, "Orbital free DFT versus single density equation: A perspective through quantum domain behavior of a classically chaotic system", *Physical Chemistry Chemical Physics*, vol. 17, no. 47, pp. 31 516–31 529, 2015 (see p. 22).

[114] J. P. Perdew, "Jacob's ladder of density functional approximations for the exchange-correlation energy", *AIP Conf. Proc.*, vol. 577, no. August 2001, pp. 1–20, 2001 (see p. 23).

[115] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple", *Physical Review Letters*, Phys. Rev. Lett. (USA), vol. 77, no. 18, pp. 3865–3868, 1996 (see pp. 23, 55, 75, 95).

[116] Y. Zhang, D. A. Kitchaev, J. Yang, T. Chen, S. T. Dacek, R. A. Sarmiento-Pérez, M. A. Marques, H. Peng, G. Ceder, J. P. Perdew, and J. Sun, "Efficient first-principles prediction of solid stability: Towards chemical accuracy", *npj Computational Materials*, vol. 4, no. 1, 2018 (see p. 23).

[117] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model", *Journal of Chemical Physics*, vol. 110, no. 13, pp. 6158–6170, 1999 (see p. 23).

[118] L. Goerigk and S. Grimme, "Double-hybrid density functionals", *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 6, pp. 576–600, 2014 (see p. 24).

[119] J. Tirado-Rives and W. L. Jorgensen, "Performance of B3LYP density functional methods for a large set of organic molecules", *Journal of Chemical Theory and Computation*, vol. 4, no. 2, pp. 297–306, 2008 (see p. 24).

[120] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, "Density functional theory is straying from the path toward the exact functional", *Science*, vol. 356, no. 6337, p. 496c, 2017 (see p. 24).

[121] A. Kamenev, *Field theory of non-equilibrium systems*, eng. Cambridge: Cambridge University Press, 2011, pp. 1–341 (see p. 26).

[122] J. Maassen, M. Harb, V. Michaud-Rioux, Y. Zhu, and H. Guo, "Quantum transport modeling from first principles", *Proceedings of the IEEE*, vol. 101, no. 2, pp. 518–530, 2013 (see pp. 26, 53).

[123] J. S. Wang, B. K. Agarwalla, H. Li, and J. Thingna, "Nonequilibrium Green's function method for quantum thermal transport", *Frontiers of Physics*, vol. 9, no. 6, pp. 673–697, 2014 (see pp. 26, 37).

[124] N. D. Lang, "Resistance of atomic wires", *Physical Review B*, vol. 52, no. 7, pp. 5335–5342, 1995 (see p. 26).

[125] M. Brandbyge, J. L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, "Density-functional method for nonequilibrium electron transport", *Physical Review B - Condensed Matter and Materials Physics*, vol. 65, no. 16, pp. 1 654 011–16 540 117, 2002 (see pp. 26, 28, 63).

[126] J. Ferrer, C. J. Lambert, V. M. García-Suárez, D. Z. Manrique, D. Visontai, L. Oroszlany, R. Rodríguez-Ferradás, I. Grace, S. W. Bailey, K. Gillemot, H. Sadeghi, and L. A. Algharagholy, "GOLLUM: A next-generation simulation tool for electron, thermal and spin transport", *New Journal of Physics*, vol. 16, 2014 (see p. 26).

[127] *TB_Sim, http://www.mem-lab.fr/en/Pages/L_SIM/Softwares/TB_Sim.aspx, last access 24.02.2021* (see p. 26).

[128] *NEMO5, https://engineering.purdue.edu/gekcogrp/software-projects/nemo5/, last access 24.02.2021* (see p. 26).

[129] G. Fiori and G. Iannaccone, *NanoTCAD ViDES*, 2008 (see p. 26).

[130] *TranSiesta, https://departments.icmab.es/leem/siesta/, last access 24.02.2021* (see p. 26).

[131] S. Smidstrup, T. Markussen, P. Vancraeyveld, *et al.*, "QuantumATK: An integrated platform of electronic and atomic-scale modelling tools", *Journal of Physics Condensed Matter*, vol. 32, no. 1, p. 015 901, 2020 (see pp. 26, 75, 95).

[132] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices", *Journal of Applied Physics*, vol. 81, no. 12, pp. 7845–7869, 1997 (see pp. 27, 35, 42).

[133] W. Lee, N. Jean, and S. Sanvito, "Exploring the limits of the self-consistent Born approximation for inelastic electronic transport", *Physical Review B - Condensed Matter and Materials Physics*, vol. 79, no. 8, p. 085 120, 2009 (see p. 27).

[134] A. Afzalian and G. Pourtois, "Atomos: An atomistic modelling solver for dissipative DFT transport in ultra-scaled hfs2 and black phosphorus mosfets", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2019-Septe, IEEE, 2019, pp. 1–4 (see p. 27).

[135] A. N. Ziogas, T. Ben-Nun, G. I. Fernández, T. Schneider, M. Luisier, and T. Hoefler, "A data-centric approach to extreme-scale ab initio dissipative quantum transport simulations", *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2019, pp. 1–13 (see p. 27).

[136] F. Ducry, A. Emboras, S. Andermatt, M. H. Bani-Hashemian, B. Cheng, J. Leuthold, and M. Luisier, "Ab-initio modeling of CBRAM cells: From ballistic transport properties to electro-thermal effects", *Technical Digest - International Electron Devices Meeting, IEDM*, IEEE, 2018, pp. 1–4 (see pp. 27, 71).

[137] T. Frederiksen, M. Paulsson, M. Brandbyge, and A. P. Jauho, "Inelastic transport theory from first principles: Methodology and application to nanoscale devices", *Physical Review B - Condensed Matter and Materials Physics*, vol. 75, no. 20, pp. 1–22, 2007 (see pp. 27, 39, 44).

[138] A. Gagliardi, G. Romano, A. Pecchia, A. Di Carlo, T. Frauenheim, and T. A. Niehaus, "Electron-phonon scattering in molecular elec-

tronics: From inelastic electron tunnelling spectroscopy to heating effects", *New Journal of Physics*, vol. 10, 2008 (see p. 27).

[139] N. Cavassilas, M. Bescond, H. Mera, and M. Lannoo, "One-shot current conserving quantum transport modeling of phonon scattering in n-type double-gate field-effect-transistors", *Applied Physics Letters*, vol. 102, no. 1, pp. 10–13, 2013 (see p. 27).

[140] H. Tian and G. H. Chen, "Application of hierarchical equations of motion (HEOM) to time dependent quantum transport at zero and finite temperatures", *European Physical Journal B*, vol. 86, no. 10, 2013 (see p. 27).

[141] F. Bloch, "Über die Quantenmechanik der Elektronen in Kristallgittern", *Zeitschrift für Physik*, vol. 52, no. 7-8, pp. 555–600, 1929 (see pp. 28, 75).

[142] P. M. Gill, "Molecular integrals Over Gaussian Basis Functions", *Advances in Quantum Chemistry*, vol. 25, no. C, pp. 141–205, 1994 (see p. 28).

[143] L. W. Wang, "Elastic quantum transport calculations using auxiliary periodic boundary conditions", *Physical Review B - Condensed Matter and Materials Physics*, vol. 72, no. 4, pp. 1–10, 2005 (see p. 28).

[144] A. Garcia-Lekue, M. G. Vergniory, X. W. Jiang, and L. W. Wang, "Ab initio quantum transport calculations using plane waves", *Progress in Surface Science*, vol. 90, no. 3, pp. 292–318, 2015 (see p. 28).

[145] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, "Maximally localized Wannier functions: Theory and applications", *Reviews of Modern Physics*, vol. 84, no. 4, pp. 1419–1475, 2012 (see p. 28).

[146] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors", *Journal of Applied Physics*, vol. 91, no. 3, pp. 2343 – 2354, 2002 (see p. 30).

[147] S. Brück, "Ab-Initio Quantum Transport Simulations for Nanoelectronic Devices", PhD thesis, ETH Zurich, 2017 (see pp. 30, 34).

[148] R. Haydock, "The recursive solution of the schroedinger equation", *Computer Physics Communications*, vol. 20, pp. 11 – 16, 1980 (see p. 32).

[149] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, "Generalized many-channel conductance formula with application to small rings", *Physical Review B*, vol. 31, no. 10, pp. 6207 – 6215, 1985 (see p. 35).

[150] R. Rhyner and M. Luisier, "Atomistic modeling of coupled electron-phonon transport in nanowire transistors", *Physical Review B - Condensed Matter and Materials Physics*, vol. 89, no. 23, pp. 1 – 12, 2014 (see pp. 36, 42, 45, 46, 76, 95).

[151] A. Togo, L. Chaput, and I. Tanaka, "Distributions of phonon lifetimes in Brillouin zones", *Physical Review B - Condensed Matter and Materials Physics*, vol. 91, no. 9, 2015 (see pp. 38, 39).

[152] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the s p3 d5 s* tight-binding formalism: From boundary conditions to strain calculations", *Physical Review B - Condensed Matter and Materials Physics*, vol. 74, no. 20, pp. 205 323 – 205 335, 2006 (see p. 38).

[153] S. Brück, M. Calderara, M. H. Bani-Hashemian, J. VandeVondele, and M. Luisier, "Efficient algorithms for large-scale quantum transport calculations", *Journal of Chemical Physics*, vol. 147, no. 7, 2017 (see p. 38).

[154] M. Luisier and A. Schenk, "Atomistic simulation of nanowire transistors", *Journal of Computational and Theoretical Nanoscience*, vol. 5, no. 6, pp. 1031–1045, 2008 (see pp. 39, 76, 95).

[155] P. Carbonniere, A. Dargelos, and C. Pouchan, "Vibrational analysis from Quantum Mechanic molecular dynamics trajectories", *AIP Conference Proceedings*, vol. 963, no. 1, pp. 329–336, 2007 (see p. 39).

[156] X. Gonze, "Perturbation expansion of variational principles at arbitrary order", *Physical Review A*, vol. 52, no. 2, pp. 1086–1095, 1995 (see p. 39).

[157] C. Herring and E. Vogt, "Transport and deformation-potential theory for many-valley semiconductors with anisotropic scattering", *Physical Review*, vol. 101, no. 3, pp. 944–961, 1956 (see p. 40).

[158] H. Fröhlich, "Electrons in lattice fields", *Advances in Physics*, vol. 3, no. 11, pp. 325–361, 1954 (see p. 40).

[159] M. Luisier and G. Klimeck, "Atomistic full-band simulations of silicon nanowire transistors: Effects of electron-phonon scattering", *Physical Review B - Condensed Matter and Materials Physics*, vol. 80, no. 15, pp. 1–11, 2009 (see p. 43).

[160] Á. Szabó, R. Rhyner, and M. Luisier, "Ab initio simulation of single- and few-layer MoS2 transistors: Effect of electron-phonon scattering", *Physical Review B - Condensed Matter and Materials Physics*, vol. 92, no. 3, pp. 1–10, 2015 (see p. 44).

[161] F. Ducry, M. H. Bani-Hashemian, and M. Luisier, "Hybrid Mode-Space-Real-Space Approximation for First-Principles Quantum Transport Simulation of Inhomogeneous Devices", *Physical Review Applied*, vol. 13, no. 4, p. 44 067, 2020 (see pp. 46, 53, 58, 71).

[162] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, "Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches", *Journal of Applied Physics*, vol. 92, no. 7, pp. 3730–3739, 2002 (see pp. 46, 47, 54).

[163] G. Mil'nikov, N. Mori, and Y. Kamakura, "Equivalent transport models in atomistic quantum wires", *Physical Review B - Condensed Matter and Materials Physics*, vol. 85, no. 3, p. 035 317, 2012 (see pp. 46, 47, 49, 54, 61, 70).

[164] M. Shin, W. J. Jeong, and J. Lee, "Density functional theory based simulations of silicon nanowire field effect transistors", *Journal of Applied Physics*, vol. 119, no. 15, p. 154 505, 2016 (see pp. 46, 47, 49, 54, 61, 62, 70).

[165] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms 1. General considerations", *IMA Journal of Applied Mathematics (Institute of Mathematics and Its Applications)*, vol. 6, no. 1, pp. 76–90, 1970 (see p. 50).

[166] R. Fletcher, "A new approach to variable metric algorithms", *The computer journal*, vol. 13, no. 3, pp. 317–322, 1970 (see p. 50).

[167] D. Goldfarb, "A Family of Variable-Metric Methods Derived by Variational Means", *Mathematics of Computation*, vol. 24, no. 109, p. 23, 1970 (see p. 50).

[168] D. F. Shanno, "Conditioning of Quasi-Newton Methods for Function Minimization", *Mathematics of Computation*, vol. 24, no. 111, p. 647, 1970 (see p. 50).

[169] Y. Lv, H. Wang, S. Chang, J. He, and Q. Huang, "Band Structure Effects in Extremely Scaled Silicon Nanowire MOSFETs With Different Cross Section Shapes", *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3547–3553, 2015 (see p. 53).

[170] M. Calderara, S. Brück, A. Pedersen, M. H. Bani-Hashemian, J. Vandevondele, and M. Luisier, "Pushing back the limit of Ab-initio quantum transport simulations on hybrid supercomputers", *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, vol. 15-20, ACM, 2015, pp. 1–12 (see p. 53).

[171] L. Zeng, Y. He, M. Povolotskyi, X. Liu, G. Klimeck, and T. Kubis, "Low rank approximation method for efficient Green's function calculation of dissipative quantum transport", *Journal of Applied Physics*, vol. 113, no. 21, 2013 (see p. 53).

[172] J. Z. Huang, W. C. Chew, J. Peng, C. Y. Yam, L. J. Jiang, and G. H. Chen, "Model order reduction for multiband quantum transport simulations and its application to p-type junctionless transistors", *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2111–2119, 2013 (see p. 53).

[173] S. Kim, M. Luisier, A. Paul, T. B. Boykin, and G. Klimeck, "Full three-dimensional quantum transport simulation of atomistic interface roughness in silicon nanowire FETs", *IEEE Transactions on Electron Devices*, vol. 58, no. 5, pp. 1371–1380, 2011 (see p. 54).

[174] M. Ye, X. Jiang, S. S. Li, and L. W. Wang, "Large-scale ab initio quantum transport simulation of nanosized copper interconnects: The effects of defects and quantum interferences", *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2019-Decem, 2019, pp. 1–24 (see p. 54).

[175] N. A. Lanzillo, B. D. Briggs, R. R. Robison, T. Standaert, and C. Lavoie, "Electron Transport Across Cu/Ta(O)/Ru(O)/Cu Interfaces in Advanced Vertical Interconnects", *Computational Materials Science*, vol. 158, no. December 2018, pp. 398–405, 2019 (see p. 54).

[176] M. Qiu, S. Ye, W. Wang, J. He, S. Chang, H. Wang, and Q. Huang, "Spin transport properties of magnetic tunnel junction based on zinc blende CrS", *Superlattices and Microstructures*, vol. 133, no. July, p. 106 199, 2019 (see p. 54).

[177] M. Li and M. Smeu, "Atomistic simulation of the structural and conductance evolution of Au break junctions", *Computational Materials Science*, vol. 164, no. December 2018, pp. 147–152, 2019 (see p. 54).

[178] F. Ducry, M. H. Bani-Hashemian, and M. Luisier, "A hybrid mode-spacereal-space scheme for DFT+negf device simulations", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2019-Septe, Udine, Italy: IEEE, 2019, pp. 1–4 (see p. 54).

[179] T. D. Kühne, M. Iannuzzi, M. Del Ben, *et al.*, "CP2K: An electronic structure and molecular dynamics software package -Quickstep: Efficient and accurate electronic structure calculations", *Journal of Chemical Physics*, vol. 152, no. 19, 2020 (see pp. 55, 75, 95).

[180] J. VandeVondele and J. Hutter, "Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases", *Journal of Chemical Physics*, vol. 127, no. 11, p. 114 105, 2007 (see pp. 55, 63, 75, 95).

[181] S. Goedecker, M. Teter, and J. Hutter, "Separable dual-space Gaussian pseudopotentials", *Physical Review B*, vol. 54, no. 3, p. 1703, 1996 (see pp. 55, 75, 95).

[182] J. Z. Huang, H. Ilatikhameneh, M. Povolotskyi, and G. Klimeck, "Robust mode space approach for atomistic modeling of realistically large nanowire transistors", *Journal of Applied Physics*, vol. 123, no. 4, p. 044 303, 2018 (see p. 60).

[183] L. F. Wang and Y. Xia, "A linear-time algorithm for globally maximizing the sum of a generalized Rayleigh quotient and a quadratic form on the unit sphere", *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 1844–1869, 2019 (see p. 61).

[184] B. K. You, J. M. Kim, D. J. Joe, K. Yang, Y. Shin, Y. S. Jung, and K. J. Lee, "Reliable Memristive Switching Memory Devices Enabled by Densely Packed Silver Nanocone Arrays as Electric-Field Concentrators", *ACS Nano*, vol. 10, no. 10, pp. 9478–9488, 2016 (see p. 71).

[185] S. Fujii, J. A. C. Incorvia, F. Yuan, S. Qin, F. Hui, Y. Shi, Y. Chai, M. Lanza, and H. S. Wong, "Scaling the CBRAM Switching Layer Diameter to 30 nm Improves Cycling Endurance", *IEEE Electron Device Letters*, vol. 39, no. 1, pp. 23–26, 2018 (see p. 71).

[186] D. Stradi, U. G. Vej-Hansen, P. A. Khomyakov, M. E. Lee, G. Penazzi, A. Blom, J. Wellendorff, S. Smidstrup, and K. Stokbro, "Atomistic modeling of nanoscale ferroelectric capacitors using a density functional theory and non-equilibrium green's-function method", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2019-Septe, 2019 (see p. 72).

[187] *QuantumATK version 2017.1, https://www.synopsys.com/silicon/quantumatk.html, last access 24.02.2021* (see pp. 75, 95).

[188] E. S. Zijlstra, N. Huntemann, A. Kalitsov, M. E. Garcia, and U. Von Barth, "Optimized Gaussian basis sets for Goedecker-Teter-Hutter pseudopotentials", *Modelling and Simulation in Materials Science and Engineering*, vol. 17, no. 1, pp. 015 009–015 019, 2009 (see pp. 75, 95).

[189] A. Mehonic, S. Cueff, M. Wojdak, S. Hudziak, C. Labbé, R. Rizk, and A. J. Kenyon, "Electrically tailored resistance switching in silicon oxide", *Nanotechnology*, vol. 23, no. 45, 2012 (see p. 88).

[190] U. Celano, G. Giammaria, L. Goux, A. Belmonte, M. Jurczak, and W. Vandervorst, "Nanoscopic structural rearrangements of the Cu-filament in conductive-bridge memories", *Nanoscale*, vol. 8, no. 29, pp. 13 915–13 923, 2016 (see p. 89).

[191] F. Ducry, K. Portner, S. Andermatt, and M. Luisier, "Investigation of the Electrode Materials in Conductive Bridging RAM from First-Principle", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2018-Septe, Austin, Texas: IEEE, 2018, pp. 107–110 (see p. 91).

[192] J. Strand, M. Kaviani, D. Gao, A. M. El-Sayed, V. V. Afanas'Ev, and A. L. Shluger, "Intrinsic charge trapping in amorphous oxide films: Status and challenges", *Journal of Physics Condensed Matter*, vol. 30, no. 23, 2018 (see p. 92).

[193] A. Pedone, G. Malavasi, M. C. Menziani, A. N. Cormack, and U. Segre, "A new self-consistent empirical interatomic potential model for oxides, silicates, and silicas-based glasses", *Journal of Physical Chemistry B*, vol. 110, no. 24, pp. 11 780–11 795, Jun. 2006 (see p. 95).

[194] T. Staiger, F. Wertz, F. Xie, M. Heinze, P. Schmieder, C. Lutzweiler, and T. Schimmel, "Macro-mechanics controls quantum mechanics: Mechanically controllable quantum conductance switching of an electrochemically fabricated atomic-scale point contact", *Nanotechnology*, vol. 29, no. 2, 2018 (see p. 103).

[195] D. Akinwande, C. Huyghebaert, C.-H. Wang, M. I. Serna, S. Goossens, L. J. Li, H. S. Wong, and F. H. Koppens, "Graphene and two-dimensional materials for silicon technology", *Nature*, vol. 573, no. 7775, pp. 507–518, 2019 (see p. 104).

[196] Y. Liu, N. O. Weiss, X. Duan, H. C. Cheng, Y. Huang, and X. Duan, "Van der Waals heterostructures and devices", *Nature Reviews Materials*, vol. 1, no. 9, 2016 (see p. 104).

[197] T. Knobloch, Y. Y. Illarionov, F. Ducry, C. Schleich, S. Wachter, K. Watanabe, T. Taniguchi, T. Mueller, M. Waltl, M. Lanza, M. I. Vexler, M. Luisier, and T. Grasser, "The performance limits of hexagonal boron nitride as an insulator for scaled CMOS devices based on two-dimensional materials", *Nature Electronics*, vol. 4, no. February, 2021 (see p. 104).

[198] J. Strand, L. Larcher, and A. L. Shluger, "Properties of intrinsic point defects and dimers in hexagonal boron nitride", *Journal of Physics Condensed Matter*, vol. 32, no. 5, 2020 (see p. 104).

# List of Publications

### Journal papers


[1] A. Emboras, A. Alabastri, F. Ducry, B. Cheng, Y. Salamin, P. Ma, S. Andermatt, B. Baeuerle, A. Josten, C. Hafner, M. Luisier, P. Nordlander, and J. Leuthold, "Atomic Scale Photodetection Enabled by a Memristive Junction", *ACS Nano*, vol. 12, no. 7, pp. 6706 – 6713, 2018.

[2] S. Andermatt, M. H. Bani-Hashemian, F. Ducry, S. Brück, S. Clima, G. Pourtois, J. Vandevondele, and M. Luisier, "Microcanonical RT-TDDFT simulations of realistically extended devices", *Journal of Chemical Physics*, vol. 149, no. 12, 2018.

[3] B. Cheng, A. Emboras, Y. Salamin, F. Ducry, P. Ma, Y. Fedoryshyn, S. Andermatt, M. Luisier, and J. Leuthold, "Ultra compact electrochemical metallization cells offering reproducible atomic scale memristive switching", *Communications Physics*, vol. 2, no. 1, pp. 1 – 9, 2019.

[4] F. Ducry, M. H. Bani-Hashemian, and M. Luisier, "Hybrid Mode-Space-Real-Space Approximation for First-Principles Quantum Transport Simulation of Inhomogeneous Devices", *Physical Review Applied*, vol. 13, no. 4, p. 44 067, 2020.

[5] F. Ducry, J. Aeschlimann, and M. Luisier, "Electro-thermal transport in disordered nanostructures: A modeling perspective", *Nanoscale Advances*, vol. 2, no. 7, pp. 2648 – 2667, 2020.

[6] A. Emboras, A. Alabastri, P. Lehmann, K. Portner, C. Weilenmann, P. Ma, B. Cheng, M. Lewerenz, E. Passerini, U. Koch, J. Aeschlimann, F. Ducry, J. Leuthold, and M. Luisier, "Opto-

electronic memristors: Prospects and challenges in neuromorphic computing", *Applied Physics Letters*, vol. 117, no. 23, 2020.

[7] T. Knobloch, Y. Y. Illarionov, F. Ducry, C. Schleich, S. Wachter, K. Watanabe, T. Taniguchi, T. Mueller, M. Waltl, M. Lanza, M. I. Vexler, M. Luisier, and T. Grasser, "The performance limits of hexagonal boron nitride as an insulator for scaled CMOS devices based on two-dimensional materials", *Nature Electronics*, vol. 4, no. February, 2021.

### Conferences

[8] F. Ducry, A. Emboras, S. Andermatt, M. H. Bani-Hashemian, B. Cheng, J. Leuthold, and M. Luisier, "Ab-initio modeling of CBRAM cells: From ballistic transport properties to electro-thermal effects", *Technical Digest - International Electron Devices Meeting, IEDM*, IEEE, 2018, pp. 1–4.

[9] J. Leuthold, A. Emboras, B. Cheng, M. Luisier, S. Andermatt, F. Ducry, and T. Schimmel, "Single atom electronics and photonics (Conference Presentation)", *SPIE Photonics Europe*, Strasbourg, France, 2018, p. 4.

[10] M. Luisier, F. Ducry, M. Hossein, H. Bani, S. Bruck, M. Calderara, and O. Schenk, "Advanced Algorithms for Ab-initio Device Simulations", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2018-Septe, pp. 62–66, 2018.

[11] F. Ducry, K. Portner, S. Andermatt, and M. Luisier, "Investigation of the Electrode Materials in Conductive Bridging RAM from First-Principle", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2018-Septe, Austin, Texas: IEEE, 2018, pp. 107–110.

[12] C. Weilenmann, F. Ducry, S. Andermatt, B. Cheng, M. Lewerenz, P. Ma, J. Leuthold, A. Emboras, and M. Luisier, "Investigation of light-controlled filament dynamics in an electro-optical memristive photodetector (Conference Presentation)", *Proc. SPIE 11031, Integrated Optics: Design, Devices, Systems, and Applications V, 110310C (13 May 2019)*, 2019, p. 11.

[13] F. Ducry, M. H. Bani-Hashemian, and M. Luisier, "A hybrid mode-spacereal-space scheme for DFT+negf device simulations", *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2019-Septe, Udine, Italy: IEEE, 2019, pp. 1–4.

# Curriculum Vitae

*Name* Fabian Ducry
*Date of birth* 6 November 1988
*Place of birth* Lenzburg, Switzerland
*Nationality* Swiss

## Education

2016 – 2021 **PhD ETH in Electrical Engineering and Information Technology**, ETH Zurich, Switzerland
- *Ab initio* Quantum Transport in Conductive Bridging Random Access Memory

2014 – 2016 **MSc ETH in Electrical Engineering and Information Technology**, ETH Zurich, Switzerland
- *Master's Thesis*: Modeling of a Metal-Insulator-Metal Light Emitting Tunnel Junction

2011 – 2014 **BSc ETH in Electrical Engineering and Information Technology**, ETH Zurich, Switzerland

2009 **Matura**, Kantonsschule Enge, Zurich, Switzerland