Multi-scale modelling for devices and circuits

F. Ortmann[1], S. Roche[2], J. C. Greer[3], G. Huhs[4], X. Oriols[5], T. Shulthess[6], T. Deutsch[7], P. Weinberger[8], M. Payne[9], J. M. Sellier[10], J. Sprekels[11], J. Weinbub[12], K. Rupp[12,13], M. Nedjalkov[12], D. Vasileska[14], E. Alfinito[15], L. Reggiani[15], D. Guerra[16], D. K. Ferry[16], M. Saraniti[16], S. M. Goodnick[16], A. Kloes[17], L. Colombo[18], K. Lilja[19], J. Mateos[20], T. Gonzalez[20], E. Velazquez[20], P. Palestri[21], A. Schenk[22], M. Macucci[23]

[1]CEA, INAC, SPRAM, GT - 17 rue des Martyrs, 38054 Grenoble Cedex 9, France.
[2]CIN2 (ICN-CSIC) and Universitat Autónoma de Barcelona, Catalan Institute of Nanotechnology, Campus de la UAB - 08193 Bellaterra (Barcelona), Spain.
[3]Tyndall National Institute, Lee Maltings, Prospect Row, Cork, Ireland.
[4]BSC - Centro Nacional de Supercomputación, Nexus I - Planta 2 C/ Gran Capita, 2-4, 08034 Barcelon, Spain.
[5]Departament d'Enginyeria Electronica, Universitat Autonoma de Barcelona, 08193, Bellaterra, Spain.
[6]Institute für Theoretische Physik, ETH Zürich, 8093 Zürich, Switzerland.
[7]Laboratoire de simulation atomistique (L_Sim) INAC/SP2M, C.E.A. Grenoble 17, rue des Martyrs, 38054 Grenoble Cedex 9 France.
[8]Center for Computational Nanoscience, Seilerstätte 10/21, A-1010 Wien, Austria.
[9]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom, and Department of Chemistry.
[10]Network for Computational Nanotechnology, 207 S. Martin Jischke Drive, Purdue University, West Lafayette, IN, USA.
[11]Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39 D-10117 Berlin, Germany.
[12]Institute for Microelectronics, TU Wien, Gußhausstraße 27-29/E360, A-1040 Wien, Austria.
[13]Institute for Analysis and Scientific Computing, TU Wien, Wiedner Hauptstraße 8-10/E101, A-1040 Wien, Austria.
[14]School of Electrical, Computer and Energy Engineering, Ira A. Fulton School of Engineering, Arizona State University Tempe, AZ 85287-5706 USA.
[15]Innovation Energy Department, Lecce University, Campus universitario,Corpo O 73100 Lecce.
[16]Center for Computational Nanoscience, Arizona State University, Tempe, AZ, USA.
[17]Technische Hochschule Mittelhessen, Competence Center for Nanoelectronics and Photonics, Wiesenstrasse 14, Giessen 35390, Germany.
[18]Department of Physics - University of Cagliari, Cittadella Universitaria - 09042 Monserrato, Italy.
[19]Robust Chip Inc, 7901 Stoneridge Drive, Pleasanton, CA 94588, USA.
[20]Departamento de Física Aplicada, Universidad de Salamanca, Plaza de la Merced, s/n, E-37008 Salamanca, Spain.
[21]DIEGM, Via delle Scienze 208, 33100 Udine, Italy.
[22]Integrated Systems Laboratory, Gloriastrasse 35, ETH Zurich, 8092 Zurich, Switzerland.
[23]Dipartimento di Ingegneria dell'Informazione, Università di Pisa, via Girolamo Caruso 16, 56122 Pisa, Italy.

Introduction

In this document we try to address the role of modelling, and in particular of multi-scale modelling of devices and circuits, in the development of nanoelectronics in the next future. We focus on the contribution that modelling can provide to developers of new technologies, device and circuit designers, and other researchers in the field. Several issues need to be addressed, starting from the approaches required to achieve a reasonable computational efficiency, to the more challenging requirements in the simulation of nanoscale devices and circuits with respect to what happened in traditional microelectronics, to the specifically computational aspects, and to the needs of industry at this time of transition between the final downscaling of silicon CMOS technology and the rise of new material and device concepts.

This document has been widely circulated in the circuit and device modelling community and is the result of a wide consensus, as testified by the number of authors and of contributors that have been acknowledged.

We have sought to synthesize a shared view of the whole community about the future of modelling and about the needs that are seen as most urgent to support European industry.

Material and device level

Indeed, as feature sizes are scaled down, approaching the few-atom limit, device modelling becomes more and more interwoven with material modelling, and the distinction between these two fields starts to fade. In the simulation of gate stacks for MOS devices, for example, the direct connection with atomistic material modelling has been apparent for a while [1], but, as conventional devices are being shrunk further and emerging devices are based on few-nanometer size structures, a seamless connection is being created between modelling of the functional unit and material modelling, which need to be integrated into a single multi-scale hierarchy. Novel technologies are often based on new functional materials put into praxis, such as in smart mobile devices (starting from the very basic battery), as similarly pointed out by the European Commission, stating that technology developments are largely driven by such advancements: "Alternative paths to components and systems development - including nano-electronics, more integration of functionalities on chips, the use of new materials and progress in photonics - will drive a large part of technology developments" [2].

Many fundamental questions are still open in the field of nano-electronics, which cannot be answered by means of conventional simplified approaches. This type of research is highly interdisciplinary, covering the domains of chemistry, material science, physics, mathematics, and engineering with their methods and the extremely wide range of length and complexity scales. Advanced knowledge of such fields has necessarily to be combined. In addition, the complexity of quantum laws, which start playing a dominant role in nano-electronics, makes attempts to treat complete devices (and even more so complete circuits) practically impossible, if approaches at the level of a full quantum description are used. Thus, at the cross-road of new materials and nanoelectronics, multi-scale modelling approaches appear as the only option allowing fast progress in the near future. This is true not only for emerging devices, but also for aggressively scaled CMOS and other advanced traditional devices. In particular self-heating and thermodynamic issues are expected to become relevant for all devices at the nanoscale, which requires the coupled solution of the electrical and heat problems, as discussed also in the example on the development of thermodynamically consistent hybrid models.

A manifold of flavors and codes for ab initio simulation exist, such as VASP, AB-INIT, ONETEP, or SIESTA, to mention a few, each of which is adapted to the specific needs of their users. These software tools provide an atomistic viewpoint by simulating the atomic structure from monomers to larger clusters of atoms and molecules, with a trend to treat larger systems with lower accuracy (a commonly accepted practice). Ab initio methods were very successful in recent years, although LDA (Local Density Appoximation) approaches, which represent the bulk of currently existing ab initio efforts, have well-known problems with the evaluation of the bandgap. The ab initio community has developed more reliable, although much more expensive from a computational point of view, methods, such as GW, which still represent an open field of research and that can in principle provide more accurate quantitative estimates of the relevant quantities. Often experimental findings are supported by means of simulations of at least basic properties. There is in general a great need for modelling research devices accurately over a range of length-scales, taking into account the appropriate physics at each scale. For example, there is a particular need to model the effect of surface gates on low-dimensional systems, taking proper account of exposed surfaces between the gates.

The computational cost of performing large scale DFT (Density Functional Theory) calculations is unfortunately very heavy, but it can be significantly reduced by adopting linear scaling approaches as implemented in codes such as CONQUEST, ONETEP, SIESTA (though SIESTA also does conventional cubic scaling and is mostly used in that mode), BigDFT, OpenMX and MondoSCF. However, while these codes significantly reduce the computational cost for large systems, there is still the problem that relative timescales increase rapidly with system size and hence a huge computational challenge results if it is necessary to sample a relatively large region of parameter space.

The requirement of self-consistency further slows down such calculations. Although many concepts exist to weaken this impact, there is a practical limitation to the size of systems that can be described directly and in full detail with ab initio methods, at present a few nanometers. This, however, is not the length scale on which the functionality of even the simplest devices can be described, which is at least one order of magnitude above. Even the basic characteristics of functional materials are defined on a length scale beyond the reach of direct ab initio methods. This is because their properties are defined by the presence of low concentrations of dopants, impurities or structural defects.

Multi-scale modelling is capable of bridging across the different length scales. The concept is based on the observation that not all interactions must necessarily be treated within the first principles framework. In an ensemble of small systems, certain variables average out and can be replaced by mean characteristic values which enter as parameters into the next level. This allows one to introduce a hierarchy of interactions, which might be founded either on very general considerations or just adapted and valid for the specifically studied properties. Based on this, a hierarchy of levels of treatment may be introduced. The lowest (at the nanoscale or, better, at the atomic scale) level deals with the smallest objects at the highest accuracy. It can be identified with the full ab initio level. Multi-scale modelling defines first the models on each level and then the interfaces for transferring relevant information to the respective upper (or even lower, to create a feedback loop) level where they are further processed. The advantage is that not all information available on the computationally heavy lower level enters the upper-level modelling but only relevant condensed information, which is precisely where multi-scale modelling is benefiting from. In addition, the modelling of interactions at the upper macroscopic level replaces the full analysis of couplings at the more refined lower level. This allows one to reduce the work performed at the lower level by limiting it to the treatment of smaller parts (non-interacting subsystems).

For instance, a finite-range impact on electrostatics and on electronic properties can be expected from impurities or dopants, depending on the local surrounding of a host crystal. Additional long-range parts such as the Coulomb interaction might be separable, and can be treated on the upper level. The information on local electronic properties can still be obtained with massively parallel ab initio methods using large supercells. On the other hand, the evolution of a system as a whole, consisting of millions of atoms, with the inclusion of a given distribution of such dopants, cannot be predicted by means of ab initio methods with the entire system included at the same level of detail.

Taking advantage of the above observation and of simplified, but still realistic, modelling of a macroscopic part of the system at the upper level, a solution can be found by treating fully ab initio only a subsystem, in order to extract relevant information. Such information, together with separate long-range interactions, completes the upper-level model for the full system. In the case of transport properties of novel materials, the interconnections between subsystems are analyzed by means of traveling quantum particles. The interactions may be nontrivial

and in such situation, a time-consuming accurate numerical analysis needs to be performed.

At the macroscopic level it may be possible to reduce the many-electron problem to the study of single quasi-particles moving into averaged fields and weakly interacting with the environment. A proper description can be provided by alternative formulations of quantum mechanics in terms of Green's functions, density matrix, Wigner and Bohm representations. However the boundary conditions in large systems weakly affect the phonons, which become one of the major factors of the environment to influence the carrier transport process.

To design a realistic scenario of the influence of disorder on transport properties of materials and devices one has to consider different length scales and modelling strategies simultaneously. First, a microscopic picture of the atomic structure is necessary to access electronic properties. This can only be provided with state-of-the-art first principles simulations. These calculations can be carried out using simple unit cells in the case of clean systems. When considering crystal imperfections or dopants, larger supercells with few impurity atoms or defect sites are necessary for the simulation. For the analysis of larger structures, for which first-principle simulations would be computationally prohibitive, simpler approaches, such as tight-binding and envelope function methods [3] (optimized, if possible, for the specific application [4,5]), are preferable. In a multi-scale framework, the parameters on which these models operate are extracted by means of more detailed ab initio studies performed on smaller structures.

Another interesting approach to multi-scale simulation is "coarse graining," a method whereby one attempts to obtain a coarse model from ab initio techniques. An example can be found in Ref. [6], where the authors propose the modelization of graphene, from the point of view of its mechanical properties, in terms of "blobs" consisting of hexagonal clusters of carbon atoms and described by means of center of mass variables.

The interface part is an essential ingredient of modelling. It defines which information is exchanged, i.e. which features of the ab initio simulation are relevant enough to affect the macroscopic length scale. From these simulations one extracts electronic structure parameters which represent the interactions at this level. For the efficiency of the multi-scale approach it is very advantageous if the extracted parameters are generic. This should be considered when setting up the modelling strategy to reduce or, at best, avoid feedback effects, whenever possible.

Once the macroscopic model is well defined and its parameters are provided through the interface, a variety of situations can be investigated keeping the interface parameter fixed but changing the arrangement or interconnection of such subsystems and/or environmental conditions. In order to make usage of the codes accessible to a wide audience and to device and circuit designers, a special effort will have to be spent on the automation of some of the most challenging aspects, such as adaptive grid generation or the proper choice of underrelaxation and convergence parameters. We believe that flexibility will be one of the central advantages of multi-scale approaches. This will also help in achieving a better understanding of complex systems, which at present are otherwise not accessible.

Example: Electron-Phonon Interaction Models

The phase space description of electron transport allows a convenient formulation of a hierarchy of electron-phonon interaction models, which range from the most comprehensive quantum description to the classical Boltzmann picture, as sketched in Fig. 1. The generalized Wigner equation provides the pure state for a system with a single electron moving in a potential and coupled to a phonon environment. If the information about the phonon subsystem is

obsolete, the phonon variables may be traced out under certain assumptions, and with approximations including a weak scattering limit and an equilibrium environment [7]. The obtained model is an equation for the electron Wigner function, coupled to two auxiliary equations. Further assumptions give rise to the Levinson and the Barker-Ferry equations for bulk material or quantum wires, where the interaction with phonons is rich in quantum phenomena such as collisional broadening, collision retardation, non-Markovian evolution and lack of energy conservation of the individual processes. These are related to the finite duration of the interaction process, which typically occurs at the femtosecond scale. The establishment of the energy conserving delta function occurs over longer time scales, in the picosecond range, given by the classical limit in the phonon term and characterized by instantaneous, local in position, Boltzmann collisions. Finally, slowly varying potentials can be represented by their local gradients, thereby leading to the classical Boltzmann equation at the very bottom of the figure.

Overall, the more physically refined approaches in the upper part of the hierarchy are extremely demanding from the computational point of view. For instance, for the Barker-Ferry model of evolution of an initial electron distribution in a quantum wire, a simulation of a time evolution above 250 femtoseconds is hardly reachable, even with massive GRID simulations.

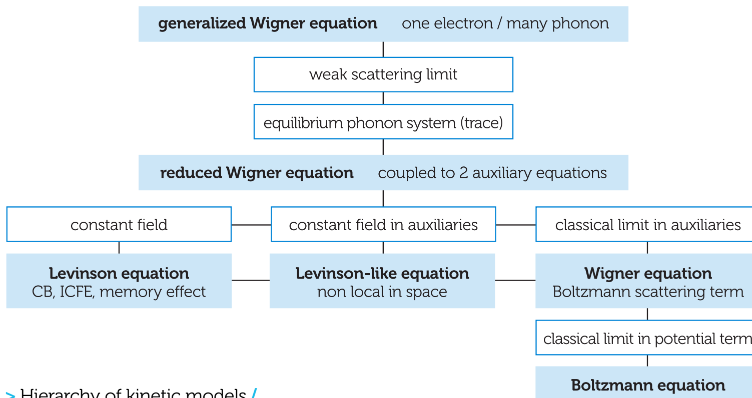

Fig. 1 > Hierarchy of kinetic models./

WF MODEL	SYSTEM DESCRIPTION	PHYSICAL FEATURES	NUMERICAL TREATMENT
Generalized	detailed quantum electron + phonons	pure state for the electron + phonons	not possible
Reduced (electron)	quantum electron, quantum el.-phon. interaction	interference between coherent and phonon processes	not available
Wigner-Boltzmann	quantum electron, Boltzmann el.-phon. interaction	coherent and phase-breaking processes	Monte Carlo 1D
Barker-Ferry or Levinson model	classical electron, quantum el.-phon	intra-collisional field effect, collisional broadening and retardation effects	Monte Carlo; deterministic; homogeneous or quantum wire
Boltzmann	classical limit for electron and el.-phon. interaction	classical transport	Monte Carlo: 3D devices deterministic

Table 1 > Electron-phonon interaction models./

Example: Organic Semiconductors

Recent advances of multi-scale modelling include the simulation of polaron transport in organic matter using first-principles and a Kubo approach for quantum transport [8]. This methodology evaluates the macroscopic current response to an applied electric field, as probed in standard measurements on carrier mobility in organic semiconductors. The method is based on microscopic parameters which are computed in the ab initio framework but is an extension with respect to previous conventional approaches [9]. Figure 2 shows the theoretical carrier mobility of holes in naphthalene compared to experimental measurements on highly purified material. The effect of disorder is assumed to be minor in this study, while the strong temperature-dependence is governed by phonon scattering of charge carriers.

A significant step forward in our understanding of transport in organic matter requires further clarification of the impact of disorder. In particular, the interplay of impurity scattering with phonon scattering is poorly understood at present. One of the fundamental open questions is still about the temperature dependence of the transport regime (coherent, or phonon-assisted hopping) as well as its dependence on the direction and on dimensionality. Consequently, besides the study of intrinsic properties, the impact of structural disorder and of dopants on transport is equally important. The way disorder effects are described in the novel approach is essentially with a multi-scale technique involving the simulation of a macroscopic 3D sample with a size of few hundred nanometers. It is based on material parameters known from previous studies, complemented with the inclusion of realistic disorder in real space.

Example: Mobility gaps

Undoped single layer graphene behaves as a zero-gap semiconductor, and thus it turns out to be an unsuitable material for achieving efficient field effect transistor functionality for digital applications. Indeed, experimental measurements reported ratios between the current in the ON state and the current in the OFF state not larger than one order of magnitude. A possibility to increase the (zero) gap of graphene single layers is to shrink their lateral dimensions [3]. Using e-beam lithographic techniques and oxygen plasma etching, graphene nanoribbons can be fabricated with ribbon widths of a few tens of nanometers, down to about 10 nm. This lateral confinement creates an electronic

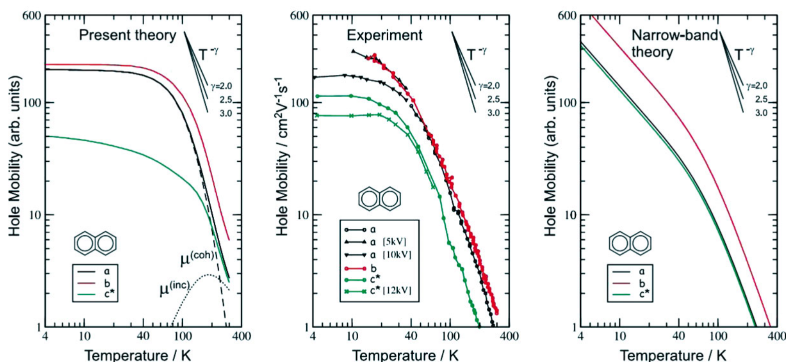

Fig. 2 > Charge carrier mobility in ultrapure naphthalene organic single crystals. Comparison between experimental results (middle), new theoretical prediction (left), as well as previous theory (right) for mobility anisotropy and temperature dependence. Adapted from [9]./

bandgap [10] with a magnitude decreasing with increasing nanoribbon width. However, theoretical predictions and experimental results have reported energy bandgaps far too small or very unstable as a result of edge reconstruction and defects, thus preventing, at the present stage, from envisioning graphene-based devices that outperform ultimate CMOS-FETs.

To circumvent such an effect [11], one could instead resort to larger-width graphene nanoribbons (above 10 nm in lateral size), seeking to compensate bandgap shrinking by triggering mobility gaps through chemical doping (such as substitutional boron or nitrogen doping). These mobility gaps are unique consequences of a wide distribution of quasibound states over the entire valence band (for acceptor-type impurities), resulting from dopants that are randomly distributed across the ribbon width.

In Fig. 3 the conductance (computed with the Landauer-Büttiker method) [12] of a 10 nm wide armchair nanoribbon with low boron doping is reported. For a doping density of about 0.2%, the system presents a mobility gap of the order of 1 eV. When lowering the doping level to 0.05%, the mobility gap reduces to about 0.5 eV and finally becomes less than 0.1 eV for lower doping. The 0.2% case is obtained for a given nanoribbon width and length, so that adjustments need to be performed if upscaling either the lateral or longitudinal sizes, but the recipe is straightforward, once the transport length scales (mean free paths, localization length) have been computed.

Transport asymmetries can be induced in graphene also as a result of adsorbates, as shown in [13], where conductivity is computed with an analytical approach that yields results in good agreement with numerical simulations. The analytical approach in [13] is based on the kinetic theory for lower adsorbate concentrations and on a renormalization group analysis for higher concentrations, always starting from tight-binding parameters derived with DFT techniques.

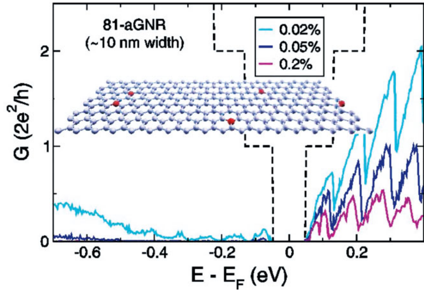

Fig. 3 > Main panel: average conductance as a function of energy for the semiconducting 81-aGNR and three selected doping rates (0.02%, 0.05% and 0.2%, from top to bottom). Inset: schematic plot of a randomly doped 34-aGNR./

One notes however that the existence of mobility gaps (with conductance several orders magnitude lower than the conductance quantum) cannot yield a straightforward quantitative estimation of resulting ON/OFF current ratio, since this will require computing the charge flow in a self-consistent manner (using a Schrödinger-Poisson solver). This will be essential since accumulated charges inside the ribbon channel will screen the impurity potential in an unpredictable way. This evaluation definitely deserves specific consideration and stand as an important challenge of multi-scale modelling in the Information and Communication Technologies (ICT) domain.

Example: Nanowire transistors

Short-channel effects begin to dominate behavior in classical, planar MOSFETs when transistor gate lengths are below 10 nm. In this regime, new device structures that maintain electrostatic control of the channel region by multiple-gates (“MuGFETs” or “finFETS”) and gate-all-around geometries (GAA) are being pursued, as well as being currently implemented in commercial digital integrated electronics fabrication. These devices, the so-called multi-gate MOSFETs, are effectively a silicon nanowire with the gate electrode wrapped around the nanowire. This simple

structure permits effective gate control of the channel, enabling switching between the high current "ON" state and a low current "OFF" state, with the current in the "ON" state meeting the stringent requirements for current drive and the current in the "OFF" state small enough to ensure low power consumption. Atomic scale simulation is playing a key role in designing and assessing device performance over lengths of the order of a few nanometers. For these dimensional scales, quantum transport effects can be dominant and device simulators need to account for the quantum mechanical nature of the charge carriers and the atomic structure of the materials in the transistor (a hierarchical simulation approach has been followed, for example, in the numerical analysis of Ref. [14]). Fully functional gate-all-around transistors based on a silicon nanowire with a diameter of only 3 nm, that is, roughly the same diameter as typical carbon nanotubes, were reported in 2006 [15]. Recently, a device that circumvents the need to make abrupt p-n junctions at the atomic scale has been proposed and fabricated [16]. Almost synonymous with transistors is the need to form p-n junctions, but on the length scale of a few nanometers the fabrication and the physics of the p-n junction become challenging. The creation of p-n junctions requires atomic precision, and the behavior of dopants in small nanowires is significantly different, due to quantum confinement and polarization effects reflected in the electronic band structure of nanowires. Extremely large doping concentration gradients become necessary; the laws of diffusion and the statistical nature of the distribution of the doping atoms work against this requirement. Hence the concept of a junctionless transistor is attractive as an "end-of-the-roadmap" device for silicon technologies. Following the experimental realizations of gated nanowire transistors, proof-of-concept simulations for junctionless gated Si nanowire transistors have been performed, based on quantum mechanical calculations that take into account an explicit description of the atomic scale structure of the nanowires [17]. This has made possible to explore transistor action in devices with gate lengths below 3 nm. Simulations reveal that Si-based transistors are physically possible without major changes to the design philosophy at scales of approximately 1 nm wire diameter and 3 nm gate length. Metal or silicide source/drain regions are very interesting for the reduction of parasitic access resistances [18]. Quantum-mechanical tunneling of both electrons and holes at the source/drain barriers plays a major role in the calculation of the device current both in the ON and in the OFF state. Traditional TCAD tools do not provide a satisfactory treatment of such phenomena, in particular when the channel length is below 10 nm. Therefore atomistic approaches are definitely needed for a proper estimation of device characteristics.

This is also true for the accurate simulation of the device current in "steep slope" devices, where a steep switching behavior (with a subthreshold slope less than 60 mv/dec) is achieved as a result of the modulation of the tunneling barrier at the source by the gate bias [19]. When several devices will be connected to obtain specific functionalities, multi-scale modelling will be necessary to handle at the same time the active areas of the devices and their interfaces to contacts (treated atomistically), the bulk region of the contacts and the interconnections, as well as the overall circuit.

An interesting tool for the simulation of transport in nanowire devices is represented by OMEN Nanowire, which is based on a bandstructure calculation performed with a semi-empirical tight-binding model [20], followed by a transport calculation with a wave-function approach.

Often elastic models are not sufficient for the proper description of transport, since charge carriers face incoherent scattering. This can be neglected only at temperatures of a few Kelvin and below, while at room temperature phonons disrupt the coherent propagation. Therefore, a realistic model for transport in semiconductor nanodevices has to treat not only carrier confinement, tunneling, quantum interference, but also incoherent scattering with a unitary approach.

It is well established that the nonequilibrium Green's function theory (NEGF) is a most general scheme for the analysis of coherent and incoherent quantum transport. Since its introduction in the 1960's this formalism has been successfully been applied to a great variety of transport problems. The non-equilibrium Green's function (NEGF) formalism provides a sound conceptual basis for the development of the atomic-level quantum mechanical simulators that are needed for nanoscale devices of the future. The approach has been already successfully applied to a variety of technologically relevant devices like RTDs and MOSFETs, but also to emerging devices such as those based on graphene and carbon nanotubes.

Example: Nanowires and interactions with biological molecules

The comparable length scale of some biomolecules such as DNA, enzymes, and antibodies with respect to nanowire cross sections or nanoparticle dimensions creates the possibility of developing nanofabrication technologies that take advantage of the interaction between inorganic and organic systems. These technologies exploit patterning at the nanoscale with the specific recognition and biocatalytic properties of biomolecules. Biomolecule-nanowire hybrid systems are being explored for medical diagnostics with emphasis on low-cost "point of care" technologies and for designing biomolecular assays directly interfaced to electronic systems. Key to understanding and designing the systems is a detailed knowledge of the molecule-surface interactions, and the translation of these interactions into electrical signals. In this context, biological sensors are being explored that base their operation on a threshold voltage shift in nanowire transistors due to "gate dielectric" change. In this configuration, the nanowire is treated with a receptor that permits binding of specific biomolecules to the surface. Biomolecules are also used as chromphores for light absorption and subsequent charge transfer to inorganic electrodes in the form of a nanowire or nanoparticle. These processes are being tailored to optimize organic photovoltaic cells. Biocatalysis using hybrid nanowire/ biomolecular systems is being explored by monitoring molecular transformation of surface adsorbed species. In this scheme the photoabsorption of the nanowire is modified following catalysis of the biomolecule and used as a marker for the catalytic reactions. It is thus essential to develop a multi-scale simulation framework that can handle the large-scale calculations needed to atomistically describe the combined nanowire/biomolecule system, and, at the same time, be able to treat the electronic structure of these systems including the interaction with light and electronic transport phenomena. These approaches could be extended to achieve further functionalities, such as a description of the information coded in chemical or biological material, with which Information and Communication Technology will have sooner or later to deal.

Example: Modelling of transport properties in biological materials

In general, the electrical conductivity of proteins is close to that of insulators, but it can increase significantly in particular environmental conditions, such as in organic semiconductors. Furthermore, the microscopic mechanism of charge transport in proteins is very articulated and includes electron, ion, and proton transport in a not yet completely understood sequence.

The difficulty to discriminate among these types of transport is further enhanced by the lack of experiments, mainly related to the difficulty of working with biological materials at the nano-scale. The most qualified mechanisms to describe electron transport in proteins are the hopping/tunneling processes. In this framework, a transport model which assumes the existence of privileged pathways of sequential tunneling has gained a wide consensus.

By embracing this picture, we describe the protein as a network and the electron transport as due to sequential tunneling

along preferential channels. In doing so, we can construct a map of the protein by taking its tertiary structure from the protein data base (PDB) , and setting up a graph whose nodes correspond to the amino-acids (identified by the α-carbons). Two nodes are then connected with a link only if they can interact within a given interaction radius R_c. At the current stage, thermal fluctuations are neglected because experiments do not suggest a relevant contribution from them.

In order to investigate the conduction properties of the protein, the graph is transformed into a resistor network, with each link replaced by an elemental resistance, which mimics the different charge transfer property between neighboring amino-acids associated with their mutual distance [21,22].

To detect the differences in conduction properties between two protein states, such as those produced by a conformational change, we adopt the simplest parameter model by choosing the same resistance for each amino-acid. In this way, the tertiary structure and the interaction radius are the only relevant input parameters of the theory.

As elemental resistance of the network we take

$$R_{i,j}=l_{i,j}\,\rho/A_{i,j}$$

where $A_{i,j}=\pi(R_c^2-l_{i,j}^2/4)$ is the cross-sectional area between the spheres of radius R_c centered on the i-th and j-th node, respectively, $l_{i,j}$ is the distance between these centers, ρ is the resistivity.

By construction, each elemental resistance depends upon the distance between nodes. Therefore, following up a conformational change, such as the one occurring in bacteriorhodopsin as a result of exposure to light [23], the variation of this distance implies a variation of each elemental resistance, which eventually leads to a variation of the network resistance (and thus of the protein resistance). As a consequence, a topological transformation can be monitored by means of resistance measurements.

By taking the first and last amino-acids of the protein primary structure as ideal electrical contacts, the protein resistance for a given R_c is calculated by solving the corresponding linear resistance network within a standard Kirchhoff framework. Being interested into a change of the resistance, rather than into its absolute value, this choice for the contacts does not modify the results of the model. Numerical calculations show that for R_c below a threshold value (typically around 4 Å) the network is disconnected and thus not conducting at all. By contrast, for R_c well above the average distance between amino-acids (typically above 15 Å) each amino-acid is connected with most of the others, thus conduction becomes insensitive to modification of the structure. Between these two critical values of R_c, the change of resistance exhibits a smooth behavior. Accordingly, one can select the value of R_c which maximizes the variation of resistance due to conformational change, which is found to be around 6 Å, a well accepted value for the interacting radius between amino-acids. To account for the strong measured super-linear I-V characteristic, the model implements a barrier-limited current mechanism as follows. For increasing voltage, each elemental resistance is associated with a second value, which, playing the role of a small series resistance, is several order of magnitude lower than the first one. This choice mimics a barrier-limited mechanism, in analogy with the case of an organic molecular layer. In this way, the initial linear increase of current with applied voltage turns into a super-linear increase, with the value of the barrier energy to be fitted by comparison with experiments. Accordingly, the stochastic selection is taken to be ruled by the direct tunneling probability

$$P_{i,j}=\exp\{-2\,l_{i,j}/\hbar\}[2m(\Phi-eV_{i,j})]^{(1/2)}\}$$

where $V_{i,j}$ is the potential drop between the i-th and j-th node, m is an effective electron mass, here taken equal to that of the free electron, and Φ is the barrier height.

We notice that the model does not take into account the protein environment, mainly for two reasons. One reason is that the focus is only on the variation of protein response, and this should be independent of the environment, since it remains unchanged during the conformational change. The other reason is a consequence of experimental results, showing that transmembrane electron transport occurs essentially only via the protein and not the lipid bilayer.

Example: Current Collapse in GaN

As the technology of crystal growth has improved over the past 10 years, research on electronic devices based on GaN has also made significant progress. GaN, with a wider band gap with respect to most other semiconductors, possesses some unique physical properties that are of fundamental interest. The first attractive property is that GaN devices have high breakdown fields and can operate at much higher voltages and temperatures. GaN has also excellent electron transport properties, including good mobility, and higher saturated drift velocity. Compared to GaAs, electrons in GaN have three times higher effective mass and also 30% lower dielectric constant. Even though the electron mobility is much lower than that in GaAs, big strides have been made to obtain higher mobility in GaN by improving the quality of GaN templates used in molecular beam epitaxy (MBE) growth. Other attractive properties of III-V nitrides include high mechanical and thermal stability and large piezoelectric constants [24]. GaN also has a larger g-factor and has been predicted to have a longer spin coherence lifetime than most other commonly used material systems. The strongest feature of the III-V nitrides is the heterostructure technology it can support - Quantum welsl, modulation-doped heterointerfaces, and heterojunction structures can all be made in this system. Table 2 shows a comparison of the important material properties of GaN and other conventional semiconductors.

GaN based metal-semiconductor field-effect transistors (MESFETs), GaN/AlGaN modulation doped field-effect transistors (MODFETs) or high electron mobility transistors (HEMTs), and GaN/AlGaN superlattice structures have been demonstrated by many groups in the past decade [25,26,27]. In particular, HEMTs fabricated from wurtzite based AlGaN/GaN heterostructures have shown great potential for microwave power devices. In these heterostructures, there is a difference in the polarization field between the top layer (AlGaN) and that in the bottom layer (GaN). This polarization in wurtzite crystals is due to the bulk properties with asymmetric lattice structure and ionicity of the bonds. In addition, strain in one or both layers leads to additional built-in fields due to the polarization effect. As a result, the discontinuity of this polarization

PROPERTY	Si	GaAs	SiC	GaN
Band gap E_g (eV)	1.12	1.42	3.25	3.40
Breakdown field (MV/cm)	0.25	0.4	3.0	4.0
Electron Mobility μ (cm^2/Vs)	1350	6000	800	1300
Maximum velocity υ_d (10^7cm/s)	1.0	2.0	2.0	3.0
Thermal Conductivity κ (W/cmK)	1.5	0.5	4.9	1.3
Dielectric constant ε	11.8	12.8	9.7	9.0

Table 2 > Comparison of semiconductor material properties at 300K./

field at the interface leads to much higher carrier densities than conventional GaAs/AlGaAs heterostructures. The typical charge density is as high as 2×10^{13} cm^{-2}, which is about ten times higher than in AlGaAs/GaAs HEMTs [28].

Various groups have demonstrated large output power densities in their GaN HEMTs, but significant developmental work remains for GaN HEMTs to become viable. One of the key remaining issues is that of device reliability, and it has been reported that the output power from these devices can permanently degrade, to varying degrees, over relatively short periods of time (~10 hours) [29,30,31]. It has been observed that a current collapse can be induced in the device drain characteristics as a result of short-term DC bias stress. The term "current collapse" is taken here as a reduction in DC drain current after the application of a high drain–source voltage.

Current collapse reduces the maximum available drain current and increases the knee voltage, thereby limiting the drain current and voltage excursions and resulting in a reduced microwave output power. Thus, induced current collapse can be a significant factor in the observed degradation of HEMT output power. Trapping/Detrapping of electrons at either surface or buffer trap sites has been suggested as a possible mechanism for the above current collapse and RF dispersion effects. Trapping can compensate part of the surface component of spontaneous polarization and reduce the 2DEG electron concentration. Experimental correlation between deteriorated RF performance and surface traps has been demonstrated, however the exact mechanism, as well as the trap dynamics, are still not clearly understood.

The modelling of the current collapse phenomena, as shown in Figure 4, requires multi-time and multi-length-scale models. In the modelling strategy proposed in Figure 4, the outputs of the device simulator are Joule heating terms and the charge density, which enter as inputs into the finite volume strain calculation module, which in turn delivers the local strains and stresses of the structure to the atomistic band structure calculation module. This module later calculates shifts in the band-edges which are then treated within the deformation potential theory as scattering potentials. This information is successively used in the creation of the temperature and position dependent scattering tables that are part of the Monte Carlo module. This is a multi-length-scale problem because as information is passed between different modules we essentially move from device level to purely atomistic level.

The different time-scales that are involved in this scheme come about from the coupling of the Monte Carlo device simulator with Synopsys Simulation software. Namely, after full calibration of the Synopsys software using the thermal Monte Carlo

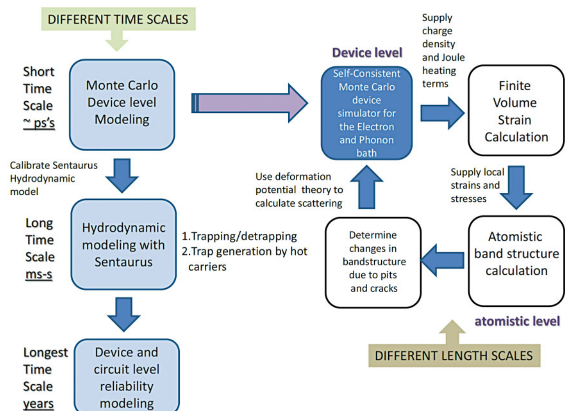

Fig. 4 > Multi-length and multi-time scale modelling of the current collapse effect in GaN HEMTs./

device solvers, it is possible to solve, using Sentaurus, the rate equations for trapping and detrapping of charge at the gate-drain interface. This represents a multi-time-scale problem because Monte Carlo calculations are typically performed on a ps time-frame, whereas the trapping-detrapping of charge happens on a millisecond to a second time scale. Finally, when these solvers are coupled to circuit level simulations, another dimension is added to the problem, on a much longer time scale.

Example: Development of thermodynamically consistent hybrid models

The self-consistent simulation of charge transport, heat transport, and optical properties in semiconductor lasers based on semi-classical models such as drift-diffusion has been proven to be a powerful tool for the understanding and optimization of their DC and AC properties [32,33]. For devices based on semiconductor nanostructures, effects on many scales have to be covered, leading to multi-scale problems [33,34,35,36]. This ranges from a quantum mechanical description for processes like quantum confinement, scattering and optical transitions (on the microscopic scale of the semiconductor nanostructure forming the functional part of the device) to a semi-classical description of the charge and heat transport through the bulk part of the device, on the coarsest scale.

Suitable embedding of microscopic models for the nanostructure into semi-classical models for the whole device results in comprehensive multi-scale description for such devices [33,34,35,36].

This hybrid modelling approach goes beyond pure microscopic and pure semi-classical ones in terms of the following aspects. On the one hand, pure microscopic models can only be applied to regions of a few-nanometer width and the influence of spatial inhomogeneities at the macroscopic level (e.g. the spatial inhomogeneous current injection), as well as that of the long-range electrostatic interaction on the scale the device, is neglected. On the other hand, in pure spatially resolved semi-classical approaches for the whole device, the modifications or adjustments for the models necessary to cover the impact of novel effects arising at the microscopic level of the nanostructure material are often “a priori” unknown. For example, the functional dependence of the electronic mobility on the carrier density and the electric field strength may be unknown (or at least the proper parameters to be inserted into existing mobility models may not be available).

Multi-scale simulations based on microscopic-macroscopic hybrid models [33,34,35] can help to explore the potential of novel promising nanostructures at the device level. In general, feedback effects from the macroscopic level to the microscopic level should be minimized within the multi-scale modelling strategy. Whenever the feedback from the coarser levels to the finer levels is negligible, multi-scale modelling can be used to extract parameters or characteristics for coarser scale models from the simulation of finer ones [33,34,35]. The simulation of the whole device can then be carried out at the macroscopic level alone, at the end of this upscaling process. If there are feedback effects (e.g. by long-range electrostatic interaction) from the macroscopic level to the microscopic level, fully self-consistent multi-scale simulations [36] of microscopic-macroscopic hybrid models have to be performed.

Device and circuit level

Device-circuit interaction

It is important to point out that in nanocircuits the distinction between the different levels is becoming more and more blurred, thereby also making the synthesis of parameters at one level to be transferred to the level above more difficult. A good case study is the one of single-electron circuits, based on the Coulomb blockade phenomenon. Here the separation between simulation at

the device level or at the circuit level is very difficult to make: the behavior of the circuit depends on the charge state of all the islands (nodes that are connected to tunnel junctions), including those represented by the interconnections among devices. As a result, a SPICE-like approach, in which each device is represented simply by way of the mathematical relationships between the currents and voltages at its terminals, is not applicable, except when special conditions occur, i.e. when the capacitances associated with the interconnections are much larger than the internal capacitances of the devices (i.e those of the tunnel junctions), as recognized by Yu and coworkers [37], the authors of the first successful SPICE model of a single-electron device. Such conditions are often verified in circuits set up for research purposes, with a limited number of components and no constraint on the miniaturization, which explains the relative success of a few existing SPICE models for single electron devices. The conditions for the validity of SPICE models are even more restrictive when an analysis of the time-dependent behavior of circuits is desired, as shown again by Yu and coworkers [38]. If in the future single electronics will overcome the practical problems represented by the random nature of stray charges and the associated unpredictability of device characteristics, it is conceivable that developers will want to achieve extremely high device densities, thereby minimizing the length of interconnects, which will make SPICE models fail, and require approaches based on Monte Carlo techniques or the master equation. Such approaches already exist, but they are computationally expensive and therefore applicable only to small circuits: it will be an important challenge within multi-scale modelling to find sustainable methods to handle circuits made up of a large number of single-electron devices.

This difficulty in making a clear distinction between simulation at the device level and at the circuit level is not exclusive to the relatively limited field of single-electron systems, but extends to many other emerging technologies, such as, for example, QCA (Quantum Cellular Automaton) circuits, where the gates controlling one cell may have a nonnegligible effect also on other cells [39] or graphene-based circuits, where, according to some proposed implementations [40], devices and interconnects will all be made with graphene, with a differentiation only in term of orientation.

Along with the development of the multi-scale hierarchy and as an additional result of it, compact models will have to be set up, in order to be able to quickly compare different device solutions and to assist in the early stage of circuit design. Not to be neglected is the interaction with the experimental work consisting in the fabrication of test structures and in their characterization, which will be essential for the validation of the multi-scale results and for the calibration of the parameters contained in the compact models.

Also for advanced versions of more "traditional" devices there are difficulties in the definition of effective and standardized compact models: well-established compact models are not yet available, for example, for multi-gate field-effect transistors or for tunnel FETs. It is likely that for many novel devices it will not be possible to have compact models that can be used in a way that is fully independent from the lower-level situation. Indeed even for MOSFETs existing compact models are still based on a drift-diffusion view of carrier transport and are hardly applicable to the current technological nodes. New ways of defining compact models will have to be devised, starting from the output of refined device simulations. It is expected that some degree of interaction will have to be introduced between levels, trying to limit it to what is really necessary to obtain a reliable simulation, and keeping the computational complexity under control.

Multi-scale aspects of device-circuit simulation in advanced traditional devices

A large majority of current multi-scale device and circuit simulations (mainly for traditional devices) are performed with available commercial simulators. A simulator like Sentaurus-Device [41] provides different transport models in one package (drift-diffusion, energy-balance, full-band Monte Carlo, electro-thermal, electro-optical, mixed-mode, transient). The same device can be simulated with models for different scales, including their own boundary conditions, which has some advantages (see below). Furthermore, multi-scale features in the form of combinations of different transport models exist, at least with an initial version. One example is the "mixed-mode" where devices can be connected to circuits, and passive elements can be included. Here, the device transport equations are coupled with the Kirchhoff laws. Another example is represented by integrated Schrödinger-Poisson solvers [42] where the quantum-mechanical solution on a quantum domain is coupled with the solution on the drift-diffusion domain. Using the terminology of Ref. [43], an overlap method for the electrostatic potential is combined with a bridge method for the densities, and a smoothing technique is applied at the boundary of the quantum window. Such hybrid methods have their own problems, both numerical and conceptual. For example, a proper geometrical definition of the sub-domain and its placement is difficult when the nature of the electron gas changes from a 3DEG (source) to eventually 2DEG (channel) in a MOSFET. Therefore, an elegant method to include the effect of confinement is to extend the system of equations by an equation for the quantum potential (also known as density-gradient method) which is solved on the whole drift-diffusion domain. The necessary fit parameter is obtained from calibrations based on Schrödinger-Poisson solutions (parametric coupling). Hybrid techniques for the coupling between Monte Carlo and drift-diffusion [44], which can be seen as combining different energy scales, have been abandoned in favor of global Monte Carlo simulations, because the additional expense is minor compared to the required iterative hybrid solution. The latter involves particle injection at the interior boundaries with (unphysical) hemi-Maxwellian statistics and again poses the problem of a suitable size and placement of the Monte Carlo window. In the time domain, global transient simulations are sometimes advised. A striking example is the industrial measurement of dark currents and breakdowns in CMOS imagers, where the voltage of typically 10 V is ramped within 10-100 ms, which defines a time scale between "static" and "RF". Depending on the temperature, *I-V* curves can be completely dominated by Maxwell's displacement currents.

The multi-scale simulator TiberCAD [45] couples semiclassical transport equations with an atomistic treatment of active regions. By means of continuum models also strain and heat transport are taken into consideration.

It is common language to state that nanoscale and microscale models have to be solved concurrently, as they are seldom decoupled [43]. This requires a clarification: in many cases the feedback from the nanoscale domain to the microscale domain is minor, compared to the strong impact that the solution on the microscale domain has on the nanoscale solution. In such cases, a "frozen-field like" approach is sufficient, which avoids expensive iterations. Examples are the weak electrostatic feedback of hot carriers in Monte Carlo simulations or of electron-hole pairs generated by band-to-band tunneling in steep pn-junctions. Another example is given by the "single-electron" memory, where single charges on the nanoscale floating gate (quantum dot) notably change the threshold voltage of the transistor in the memory cell. Here, a second correction of the quantum-mechanical charge distribution on the floating gate based on the first change of the global potential results in a further shift of the threshold voltage by less than just 1 mV.

Hence, a self-consistent simulation is not necessary.

Often a device is "mesoscopic," i.e. the scale is in between two different physical scales. A well-known example is "quasi-ballistic" transport which features both dissipative scattering and ballistic effects. In such cases it is advised and admissible to use different transport models for the whole device and to compare the outputs. This can help preventing wrong conclusions.

Another interesting example of application of multi-scale simulation is in the field of reliability and error correction, specifically the treatment of single-event generated soft errors (due to cosmic rays or other sources of ionizing radiation). The physical origin of errors occurs at a low level in the simulation hierarchy (3D charge transport simulation) and the coupling with the simulation at the compact model level is too strong to allow for de-coupled transfer of information between levels. This issue has been solved by Robust Chip Inc. by reducing the simulation at the lower level to only the long-range effect that requires a self-consistent, coupled, description, and coupling the two modelling levels self-consistently at appropriate interfaces in the simulated structure (the contact regions). This approach has led to the development of a new layout methodology to contain soft errors, and it allows an accurate prediction of error rates for a given circuit design.

Circuit-device interactions in millimeter-wave power amplifiers

Monte Carlo (MC) particle-based device simulations allow a detailed description of the nanoscale phenomena occurring inside a device. However, in the experimental characterization of an actual high-frequency device circuit packaging and the extrinsic parasitic elements due to interconnections can play a role as important as the intrinsic device behavior when the frequency increases. In particular, simulating only the intrinsic device may lead to an overestimation of DC and RF performance. Specifically, the reactive parasitic elements related to interconnections can dominate the ultra-fast device operation at high frequency, and their effect may no longer be neglected. Furthermore, the interaction between the device and the external circuits, such as matching networks, load impedance, and transmission lines, must be included, with either an analytical or non-analytical approach, in order to properly reproduce the device operating conditions as well as to obtain a more global RF characterization.

Small-signal AC analysis fails to predict large-signal device performance, which must be assessed by simulating the device within the full range of actual operating conditions. In particular, when simulating a transistor used as power amplifier, the highly frequency selective high-Q matching network (required to suppress undesired harmonics generated due to non-linearity typical of mm-wave band power amplifier classes such as -AB, -B, -F, and hard-driven Class-A) must be included in order to properly simulate the large voltage swing, due to the load-line, at the device output contacts. A time domain solution, through a Time Domain (TD) circuit solver, of these matching networks, which include large reactive elements, implies long transient times (i.e. RC, L/C time constants). On the other hand, a frequency domain circuit solver such as Harmonic Balance (HB) allows to directly obtaining the steady-state behavior of an external passive reactive network without the need for simulating the transient. The HB frequency domain solution of the circuit can be self-consistently coupled to the MC time domain solution of the device. At the end of each simulation, the iterative HB solver evaluates in the frequency-domain whether a solution of the coupled circuit-device structure has been obtained. In the negative case, a new circuit-device simulation is run where the voltage waveforms applied to the device contacts are calculated from the inverse Fourier transform of the HB frequency domain solution of the circuit.

The complete large-signal characterization of an intrinsic device coupled with an external

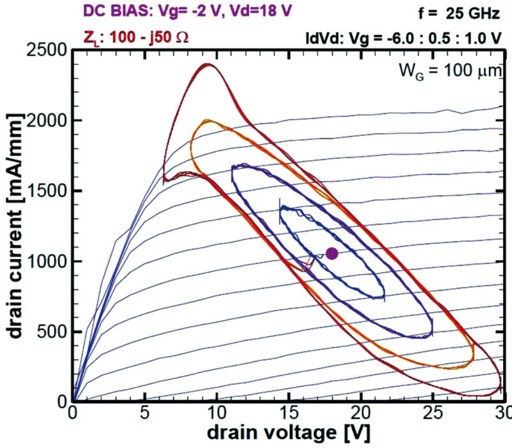

Fig. 5 > Example of the simulated dynamic load-lines (i.e.drain current swing vs. drain voltage swing) for increasing output power in AlGaN/GaN HEMT Class-A amplifier./

circuit can be performed by combining the HB circuit solver (to emulate the high-Q matching network and the load line voltage swing) with the TD circuit solver (to solve the extrinsic parasitic network that usually involves small reactive elements). In such a way, two different simulation options are available to include parasitic elements within large-signal operations: a fully HB circuit solver and a hybrid HB/TD circuit solver. Both options are self-consistently coupled to the MC particle-based device simulations that provide the time-domain solution of the device. On the other hand, the frequency domain circuit solution provided by HB allows including the effect of those passive structures that are more easily characterized in the frequency domain than in the time domain. For instance, the HB algorithm can include the effect of a transmission line by using the S-parameters characterization, which still holds even under large-signal operations for a passive and bias independent structure. In general, HB can be used to couple time-domain device simulations with circuits that are generally difficult to analyze in the time-domain such as high-Q matching networks, transmission lines, RF filters, and multi-port sub-networks, all of them described by the Y-,Z-, or S-parameters. The frequency-domain approach of Harmonic Balance can be easily applied to these structures. Furthermore, the Harmonic Balance analysis is applicable to a wide variety of transistors under large-signal operations occurring in microwave circuits such as power amplifiers, frequency multipliers, and mixers.

The typical characterization flow, in the case that preliminary experimental DC and small-signal RF characterization is available for the device under investigation, initially starts with a fit of the experimental DC characteristics. Then the RF small-signal performance are also assessed through small-signal MC simulations (by applying small-step perturbations in order to extract a two-port network characterization of the device through Y-parameters), and the agreement with the experimental measurements can be verified. At this point, once the agreement

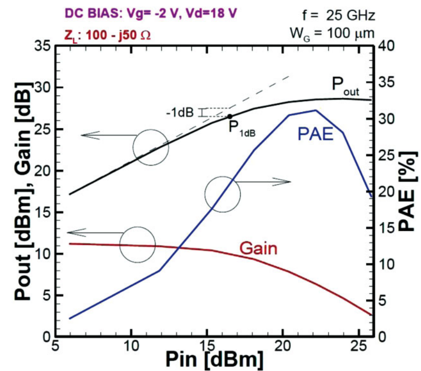

Fig. 6 > Example of the P_{out} vs. P_{in} large-signal characterization of an AlGaN/GaN HEMT Class-A amplifier./

between simulated and experimental DC/RF small-signal performance is assessed as starting baseline, one can finally proceed with the Power Amplifier characterization through HB-MC self-consistently coupled circuit-device simulations. In particular, this approach allows, in an early stage of the design, the optimization of the intrinsic device layout, as well as real device issues, such as parasitic elements (i.e. gate resistance, source/drain inductors), material defects/reliability (i.e. threading dislocations, surface traps, interface roughness), and self-heating/thermal management.

Multi-physics aspects of device and circuit simulation

The integration into multi-scale simulation codes of physical phenomena beyond charge transport, such as heat transport, optical properties, electromagnetic radiation, mechanical degrees of freedom, is a relevant issue. Future nanosystems are expected to conjugate different degrees of freedom: electrical and mechanical in nanomechanical memories, electrical and thermal in nanowire based thermoelectric generators, etc. The problem of the thermal budget for aggressively scaled devices is boosting the importance and the urgency of a fundamental understanding of thermal transport at the nanoscale, in order to devise better ways to effectively remove heat. All of this requires multiphysics approaches and makes it even more challenging to transfer parameters between different levels. Furthermore, it is expected that the yield, in terms of working devices, will be reduced and this will force acceptance of circuits that contain one or more defective devices. Thus, results will be probabilistic, and simulation tools will have to take this into account, up to circuit level.

In terms of general needs for multi-scale modelling we should mention the definition of standard benchmarks useful to verify the consistency between different levels of multi-scale simulations, in such a way that a consensus could easily be built about the validation of a new approach. Another issue, of fundamental nature, is represented by the integration of approaches such as Density Functional Theory (DFT) (which is an equilibrium theory) into Non Equilibrium Green's Function (NEGF), which are by definition nonequilibrium theories, as well as the limits of validity of standard exchange and correlation terms in the presence of Dirichlet boundary conditions close to the system being considered (which leads to significant shielding of the electrostatic interaction).

There are fields where work still has to be done to define reliable models, such as that of wide-bandgap semiconductors, which present particular difficulties because of the large energy ranges and the presence of extended defects that make the comparison between simulations and measurements rather difficult, that of high-temperature superconductors, for which a lot of approximations are still in use, and that of single-molecule transport, in which treatment of the leads is not yet well-established and differences of orders of magnitude in the computed current exist between different methods. For molecular calculations full configuration-interaction is considered to be the reference, but its application to practical transport problems is often not possible with the currently available computational resources.

Example: Development of plasmonic devices for Terahertz (THz) and photovoltaic applications

The electromagnetic (EM) spectral range 0.1-10 THz has many applications from the investigation of fundamental excitations in matter to medical and security imaging. Nevertheless the lack of suitable and reliable sources in the THz range (THz gap) has severely limited the launch of terahertz technology into the public domain. Up to now the most successful technique to generate and detect T-rays (THz-rays) uses two photoconductive metal-antennas fabricated on a low-temperature growth (LTG) GaAs substrate, the first antenna is excited by a

femto-second laser and the second one is used to detect the THz electromagnetic response to the pulse of carriers photo-generated in the first antenna by the laser beam. Of course, the optical components in the described time-domain system need to be integrated by means of semiconductor technology in order to gain in reliability, to increase its usability, and to reduce both initial price and maintenance costs. Semiconductor technology will be key to fuel the development and deployment of new THz applications [46]. The maximum cut-off frequency of conventional transistors (i.e. based on carrier transit time) lies in the sub-THz range. Nevertheless, in deep submicron/nanometer FET (field-effect transistors) devices plasma wave frequencies lie in the THz range and its frequency can be tuned adjusting the gate bias [47]. Since the plasma wave frequency is much larger than the inverse of the electron transit time in the device, it becomes easier to build THz detectors and emitters based on the transport of "ballistic" plasma waves (plasmonic devices) than on that of ballistic carriers. Moreover, the channel of the FET may be designed with dimensions such that it acts as a resonant cavity at THz frequencies, in order to enhance resonant detection. Additionally, it has been recently shown [48] that the scattering from metal nanoparticles near their localized plasmon resonance is a promising way to increase light absorption in thin-film solar cells. Enhancements in photocurrent have been observed for a wide range of semiconductors and solar cell configurations.

In the first decade of this century research in plasmonics has been developing at a breathtaking pace. Nevertheless, to keep this pace a better understanding of the coupling between the incoming/outgoing electromagnetic radiation and the device is needed. Semiconductor and EM modelling and simulation have been two almost completely separated worlds up to now and, although commercial tools exist in both fields, none of them successfully incorporates a fully coupled reliable simulation. Electromagnetic solvers, both commercial and academic, are available and are routinely used up to millimeter-wave range to design RF devices such as antennas, amplifier coupling, etc. but at this level the description of the transistors is rather simplistic. In strong contrast to this, the carrier transport in transistors used in high-frequency circuits is extremely complex and, therefore, advanced models (NEGF, Monte Carlo, hydrodynamic) are necessary to simulate such devices. A considerable effort will be needed in order to couple EM solvers to, for instance, 2D/3D Monte Carlo codes, but this effort will allow to fully understand the coupling of surface plasmons to EM radiation [49]. For example, it has been demonstrated that voltage variations on a short gate can effectively tune higher-order ungated plasmon resonances in FETs, nevertheless gated plasmons are more efficient to radiate THz waves.

New tools are necessary to design plasmonic devices. Major difficulties may be expected when coupling the Maxwell equations solver to the carrier transport one, since both time and space scales may be very different (for instance, a grating of the FET gate used to couple incoming THz radiation in a sensor has an area of hundreds of square-microns), the interaction with phonons needs to be carefully taken into account (the energy of the THz radiation, 1-10 meV, is in the range of acoustic semiconductor phonons), etc.

Thermal transport in nanostructured thermoelectrics

Thermoelectrics represent a potentially important energy conversion technology due to their ability to convert heat into electricity. A significant example is the automotive industry where heat extracted from high-temperature exhaust gases could be efficiently converted into electric power. A similar approach could be as well used in heavy industry on an even larger scale.

Despite this potential, thermoelectric devices are at present only used in niche markets

(where reliability and durability are more important than the overall performance) because of their low efficiency.

The key issue is how to maximize the thermoelectric figure of merit $Z=(S^2s)/K$ (where S is the Seebeck coefficient, while s and K are respectively the electrical and thermal conductivity), which basically dictates the heat-to-electric power conversion efficiency. In other words, we need a material with high electrical conductivity, large Seebeck coefficient, and low thermal conductivity. Such a material is unfortunately not provided by Nature.

A way to bypass this limitation and straightforwardly increase Z involves reducing K (still preserving a good s), which involves designing new materials for which the lattice thermal conductivity is more significantly affected by some additional structural feature than in pristine systems.

It has been proposed that a primary approach to obtain this reduction is to generate suitable nano-composite semiconductor materials, where the typical grain size is smaller than the phonon mean-free-path, but it is still larger than the electron (or hole) one. This typically occurs for nanometer-sized grains and warrants that, while charge carriers are basically unaffected by the underlying nano-composite structure, phonons (i.e. heat carriers) are strongly scattered at the grain boundaries. This ultimately causes a dramatic increase in the Z figure of merit [50,51].

Si-Ge systems appear to be the most promising as nano-composite thermoelectrics since they are easily and cheaply fabricated in large quantities and they are fully compatible with present-day technology.

While the first generation of Si-Ge-based applications is close to commercialization, any further improvement of this energy production technology indeed requires a better fundamental understanding of heat and charge transport in such complex materials [52].

Multi-scale modelling - based on the concurrent and/or hierarchic combination of statistical mechanics methods, materials physics and atomistic simulations - is believed to provide a powerful and valuable tool for improving our basic understanding of thermoelectric power production in such systems.

In order to understand thermal transport in Si-Ge nanostructured materials, a first important concern is represented by the generation of reliable models of Si-Ge nano-composites with tailored structural characteristics.

There are two possible approaches, both of which should be explored, in order to establish the relative advantages and disadvantages. One possibility is to start from the insertion of a pre-assigned ensemble of Ge nano-grains into a Si crystalline host, followed by a simulated annealing procedure to achieve the final nano-composite configuration. Alternatively, the Ge nano-grains can be at first randomly inserted into a Si melt and then, upon cooling down to room temperature, the final solid-state sample can be obtained.

In any case, the protocol for the sample generation must have the following features under control: (i) the grain size distribution and the average grain size; (ii) the crystallographic orientation of grains with respect to the host; (iii) the grain density and spatial distribution, (iv) the ratio of the amount of crystalline matter to that of the amorphous one (as discussed below), (v) the overall stoichiometry of the composite.

Another issue of great importance is the full characterization of Si-Ge boundaries, in terms of their effective area, orientation, and thickness. In particular, it is well known that large-angle grain boundaries in semiconductor materials are likely to be disordered (because of the release of a large amount of locally accumulated strain energy), so that the inner crystalline core of the grain is sheathed in an amorphous layer.

Once that Si-Ge nano-composite samples are defined, the investigation of their thermal

transport properties can be performed with non-equilibrium molecular dynamics techniques.

In particular, it is important to investigate how and to what extent the thermal conductivity is affected by the overall structural features listed above, as well as by the population of amorphous grain boundaries. In this respect, special care must be taken in understanding what happens at a single Si-Ge interface, as far as the flux of heat carriers is concerned. It is important to establish how such a flux is adsorbed, or transmitted, or reflected, depending on the boundary crystallographic orientation, roughness, thickness, and state of aggregation.

Another relevant issue is the difference, if any, between a homogeneous vs. graded distribution of grains. While the first configuration refers to an ordinary bulk nano-composite, the second one is also referred to as nano-graded interface. In this graded case, grains are distributed with decreasing average density along a given direction, for instance from left to right. In this way, it is possible to generate a structure where a pure Ge composition on the left is gradually changed into a pure Si composition on the right. Such a graded interface is inherently anisotropic and, therefore, new features are expected in thermal transport (as well as in any other physical property).

Importance of High Performance Computing

In the longer term, having a strong and realistic simulation capability will provide a strategic tool to support product development in all fields of applications including ICT and beyond. In that perspective, the use of High Performance Computing is a convenient and necessary strategy for all scientific fields, particularly for making advances in frontier developments in the fields of first-principles calculations and multi-scale methodologies.

The desire for an increased numerical resolution of additional physical details results in an insatiable need for more computing resources. Such a need is further exacerbated by the growing importance of statistical effects, whose treatment effectively increases the dimensionality of the problem by one: the size of the sample adds up to the spatial dimensions. In previous years, Moore's law has provided a doubling of components per chip every 18 to 24 months, which has also resulted in respective performance gains of processor cores. Consequently, a once implemented algorithm could naturally benefit from faster hardware without continuous adoptions. As transistor dimensions have reached scales where power dissipation inhibits further increases of the clock frequency, multi- and many-core processors are being introduced in the main-stream market. They allow for an efficient utilization of the still increasing number of transistors per chip, but they may only provide their full computational power to parallel algorithms and implementations. Consequently, we are now facing the fact that existing implementations no longer lead to higher performance on novel computing hardware, unless suitable parallelization measures are applied.

While additional parallelism is employed in general purpose central processing units (CPUs), graphics processing units (GPUs) have become more versatile over the last years and are now being used for general purpose computations, which is often abbreviated GPGPU. Unlike CPUs, GPUs are able to execute hundreds of lightweight threads simultaneously. This leads to a peak performance which is up to one order of magnitude higher than for multi-core CPUs. Similarly, the memory bandwidth of GPUs is also a factor of about ten larger than for CPUs, provided that regular memory access patterns across the individual threads can be ensured. A drawback of GPUs for general-purpose use is the limited amount of memory, which can be a considerable constraint for development. The acceptance of GPUs for high performance computing is reflected by the fact that leading-edge machines in the TOP 500 list of supercomputers

are now equipped with GPUs. It is worth mentioning that a single modern GPU board in 2012 offers the performance of the fastest supercomputers back in 1997. This enables numerical experiments on average desktop machines, which have been impossible even on supercomputers 15 years ago – provided that the underlying algorithm allows for a high degree of parallelism.

The increased degree of parallelism poses an additional challenge for algorithm design on the one hand, and implementation on the other hand. While a parallelization of an inherently parallel algorithm is often straightforward, a parallelization of a conceptually serial algorithm may be even impossible. Consequently, it can be beneficial to substitute an established serial algorithm by a parallel variant. While the parallel algorithm may be less efficient in a serial processing environment, it is able to utilize all computational resources available on modern architectures, ultimately leading to shorter execution times. In the context of multi-scale methodologies, the plethora of different algorithms employed requires a careful replacement of bottlenecks caused by the use of serial algorithms. In this regard it is crucial to first rely on suitable parallel algorithms, since inherently serial algorithms cannot be implemented in a massively parallel manner. In addition to the increased physical details in multi-scale simulations, close inter-disciplinary collaboration of engineers and physicists with numerical mathematicians on the one hand, and computer scientists on the other hand is the key to access the full potential of modern computing architectures in high performance computing, in order to carry out frontier developments and research.

A multitude of different approaches for programming parallel hardware has emerged. The message passing interface (MPI) has become a de-facto standard for use on distributed memory machines such as computing clusters. On shared memory machines such as current workstations, the use of OpenMP allows for adding parallelism to existing code using suitable compiler directives. Good performance enhancements can be obtained by OpenMP in certain cases, while the probability of breaking working code is minimized. Full control over the individual threads is provided by libraries such as the POSIX threads (pthreads), which is available on a wide range of different platforms. Additional vendor-specific compiler extensions are also available, most noteworthy is Cilk for C and C++. A different approach is automated code generation, which tries to automatically detect parallelism in existing code and may aid in migrating existing code. Examples of such an approach are HMPP by CAPS Enterprise or PGI Accelerators compilers. In this last case, as OpenMP for the multicore programming in shared memory systems, only some new directives need to be added into the original code as comments, which are then understood only by the specific compilers. With this approach the programming effort is less but performances are sometimes poorer. Indeed, they strongly depend on the characteristics of the algorithms. With HMPP it is possible to have a version of the code that is automatically generated but amenable to manual optimization afterwards.

In contrast to the vast number of programming approaches for CPUs, only two mainstream approaches for programming GPUs are in use. On the one hand, CUDA is a proprietary language introduced by NVIDIA. On the other hand, OpenCL is a standardized framework for general multi- and many-core architectures. It is maintained by the Khronos group and implementations from AMD, Intel and NVIDIA are available. Unlike CUDA, OpenCL can be employed on GPUs and multi-core CPUs for different vendors. Although the OpenCL standard addresses multiple target devices, full hybrid support of OpenCL implementations for devices from different vendors still needs to evolve. Despite the unified programming approach, hardware-specific implementations are required in order to obtain good performance. OpenCL can in principle also be used with special-purpose hardware such as field-programmable gate

arrays (FPGAs). A non-optimal situation in the exploitation of GPUs may derive from the fact that there is a fast development of new hardware, which is not guaranteed to be compatible with the software tools created for the previously existing hardware. This is a result of the fact that the market force driving the development of GPU hardware is computer graphics and not high-performance computation.

In a large-scale hybrid computing environment, additional challenges arise due to the inhomogeneous computing resources. A common approach for harvesting the processing power in such a setting is to combine the available approaches. Typically, an MPI layer is used for inter-node communication, whereas shared memory techniques such as OpenMP, CUDA or OpenCL are employed on the respective computing node.

Very recently a new parallel programming standard, OpenACC, has been announced by NVIDIA, Cray Inc, PGI and CAPS enterprise. This is a new open parallel programming standard allowing simplified access to hybrid CPU/GPU systems, which appear to be the trend for the future in supercomputing applications, in particular after the announcement of the new Cray hybrid supercomputer, the XK6, which will be based on a combination of multi-core AMD CPUs and NVIDIA GPUs, scalable up to 500,000 cores. OpenACC, which is being implemented into several compilers, will be operable on multiple platforms and should allow a very straightforward acceleration of legacy codes.

All of these options have in common that performance is always obtained through a massive exploitation of parallelism at different levels. It is therefore important to fully assess the possible impact of these new technologies and, if the preliminary evaluations are confirmed, to invest heavily in the development of optimized codes, implementing advanced simulation strategies for the device structures of main current interest. While some device concepts will finally move from the research field to that of actual development and industrial exploitation, it will be essential that engineers have advanced simulation codes available and that they are able to run them for realistic device structures, with the possibility of exploring a vast parameter space. This is not likely to happen if simulation codes will require state of the art supercomputers, because the availability of such facilities is going to be certainly limited also in the future. However, since the next generation of computing facilities affordable even for small companies will be massively parallel, too, investing in appropriate code now is certainly advisable. An effort should be made also in terms of engineering curricula at universities, in order to make students familiar with high performance computing and with the development of parallel codes.

For simulating complex devices, more physical insight in models and suitable use of first principles will be required in addition to high performance computing resources. For the case that access to supercomputers is nevertheless required, an infrastructure of European supercomputers named PRACE, gathering the leading platforms, already exists. Access to supercomputers facilities can be particularly beneficial, especially for those developing material simulation to serve as guidance for applications. Another way of accessing significant computational resources for the solution of a class of simulation problems is represented by grid computing. Grid computing relies on a loosely connected, geographically distributed collection of generally inhomogeneous computational resources, which can be exploited for parallel batch processing via a dedicated middleware. The main difference with respect to a traditional supercomputer or even with respect with a cluster is represented by the lack of a high-speed connection among the different nodes (which usually communicate via the Internet), but there is a significant advantage in terms of costs, because large-scale production hardware can be used, and it is possible to

exploit the spare CPU time of systems that have a specific main purpose and that would otherwise be idle.

A similar internet-based large-scale high performance computing environment is available via cloud services such as Amazon's Web Services (AWS). Cloud instances can be configured to provide heterogeneous computing instances, providing the ability to select application specific facilities. Compared to grid-based solutions, additional computing resources can be allocated on demand at short notice.

With the presence of parallel computing hardware in average desktop machines, the same programming model can be used on small- and large-scale hardware, thus simplifying the migration.

Role of physics-based modelling in predicting new phenomena

There are several examples, in the recent history of electron devices, in which physics-based modelling has played an essential role in pointing out new avenues for research or in explaining unexpected phenomena. One interesting example, which has had also important consequences in terms of applications, is that of the prediction of much improved tunneling magnetoresistance (TMR) of magnetic tunnel junctions. Up to a decade ago, the achieved values of TMR in structures with ferromagnet - dielectric - ferromagnet were at most around 70%, because of the amorphous nature of the dielectric layer. Accurate first-principles calculations [53] made it possible to predict that with crystalline dielectric layers the tunneling process was much more complicated than the previously used simple barrier model, because of the relevance of the symmetry of Bloch states at the Fermi energy and the effect on wave function matching of interfacial resonance states, and much higher TMR values were to be expected. These results triggered experimental activity aimed at improving the quality of the growth process for the dielectrics and to the actual measurement of the predicted high TMR values [54,55]. The resulting magnetic tunnel junctions have found a very important field of application in the readout heads for hard disks, allowing an unprecedented increase in the information density on a magnetic substrate, and represent the building block of magnetic memories, which are likely to be one of the main technologies of the future in the market of non-volatile storage. This further demonstrates the importance of simulation codes that can provide quantitatively reliable estimates of the parameters of interest, instead of approximate models.

Another example is represented by the application of physics-based models based on the solution of the Boltzmann transport equation to the analysis of the high-frequency performance of heterojunction bipolar transistors (HBTs) [56,57]. Significantly higher values of f_T (and in particular of the product $f_T V_{BC,br}$, where $V_{BC,br}$ is the breakdown voltage of the collector-base junction) than expected on the basis of previous theories have been found with the application of advanced transport models. This can be explained on the basis of the device structure (HBT instead of BJT), of the consequences of velocity overshoot, and of the fact that, as a consequence of the particular doping profile, the region in the collector with high field is only one fifth of the ionization length, therefore curbing impact ionization and limiting the avalanche current.

Variability, which is acquiring an increasing relevance in CMOS technology as device sizes are scaled down, can be modeled effectively only with physics-based approaches. The impact of impurity distributions, for example, has to be calculated on the basis of the position of individual atoms, which requires an atomistic approach, yielding results for a single device that must be then passed in an adequate form to higher levels of the multi-scale simulation hierarchy.

In general, from the point of view of the industry real progress is made by means of simulation tools when they can tell when

things are expected not to go according to intuition. This would be far more important than having a very sophisticated code showing that just a small correction has to be made with respect to the behavior of the device that could be predicted just by intuition or by means of standard simulators. If for the industry the real added value in a model is not the extremely precise evaluation of a quantity, the question may arise whether drift-diffusion is all they need, because with small changes it has been adapted to successive generations of devices, always yielding reasonable results. Indeed, this is not the case, because most of the adjustments have been made "a posteriori," once the behavior of a device was know from measurements or from more sophisticated models. Therefore simple approaches, such as drift-diffusion, cannot be of great help when trying to determine novel effects in devices based on new concepts.

Simulation in the time domain

The semiclassical Monte Carlo (MC) solution of the Boltzmann equation (sometimes called "numerical experiment") took a prominent role in the simulation of microelectronic devices because of its accuracy and versatility [58,59]. This numerical technique can be applied to any type of electron device with different materials, geometries and bias conditions. It provides information on the behavior of electron devices under DC, transient and AC conditions. It is not only able to provide average values, but also current and voltage fluctuations. However, there exist approximations that allow introducing some of the most significant quantum effects into MC simulators. For example the effect of the Pauli Exclusion Principle (both in the injection statistics [60] and electron dynamics inside the device [61,62]) or the occurrence of tunneling processes [63] can be semiclassically accounted for by means of numerically generating complex distributions using a simple rejection technique [64]. This could be already considered a rudimentary implementation of multi-scale modelling in which the low level quantum effects are transferred to the higher level MC semiclassical model by means of an undemanding approximation. Even if this approach is only valid under certain conditions (and therefore improvements of such approximations are necessary in order to refine the quantum description of the system at low temperature or high degeneracy), the agreement with the experimental results of real devices is remarkable.

There are indeed many further flavors of the Monte Carlo technique which feature approaches allowing the inclusion of quantum effects. Among them, the Multi-Subband Monte Carlo (MSMC) [65,66] accounts for quantization normal to transport. Subband Smoothening applied to MSMC allows taking into consideration also quantum effects in the transport direction [67] and it was verified to reproduce NEGF results for devices with a channel length down to 6 nm. Solvers for the Boltzmann transport equation (deterministic or Monte Carlo) can be coupled with the 2D Schrödinger equation and the 3D Poisson equation to model nanowires [68]. Such approaches are very heavy from the computational point of view.

The Wigner Monte Carlo [69] is a quantum transport framework able to handle quite "large" device including also scattering (although not as accurately as in NEGF) .

It seems natural to wonder whether the predicting capabilities of the present quantum electron transport simulators available for the electronic industry are comparable to those of the semiclassical Monte Carlo technique. Unfortunately, the answer is definitely no. Although most of the present quantum simulators provide very accurate information about the DC behavior, they are unable to reliably predict the transient, AC or noise properties of quantum devices in most practical operating conditions. There are mainly two reasons that explain why such predictions are so difficult. First, at high-frequency, one has to deal with the role of the "displacement current" that

requires a reasonable approximation for the (time-dependent or frequency-dependent) Coulomb interaction among electrons [70,71]. Second, in order to compute correlations (the second or higher moments of the electrical current for instance) one has to measure at least at two different times. One has to take into account that a first measurement "collapses" the evolution of the quantum system and, consequently, it will affect the outcome of a second (or further) measurement [72] (See schematic representation in Fig. 7).

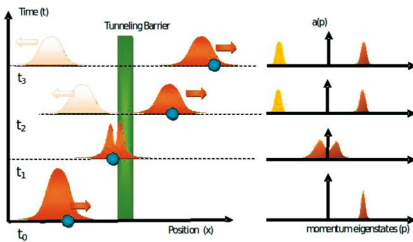

Fig. 7 > (Left) Time-evolution of a (single-particle) wave-packet impinging upon a barrier. The transmitted and reflected wave packets are spatially separated at times t_2 and t_3. (Right) representation of the probability of the momentum eigenstates, *a(p)*, at different times. At time t_2, if a positive value of the current is measured, then, only the evolution of the right-moving (dark) wavepacket will be relevant to describe the quantum system at a later time t_3. High frequency measurement of the current (AC) in quantum system has to deal with such multi-time measurement phenomena./

There are some elegant theoretical proposals in the literature that show the path to include time-dependence and correlations into practical quantum simulators. Among others, we cite the work of Büttiker and coworkers, who generalized the successful (DC) Landauer [73,74] model towards current fluctuations and AC conductances [75,76], as well as the work of Levitov, Lesovik and others in the application of Full Counting Statistics in mesoscopic devices for the computation of higher moments of the electrical current [77,78,79]. Recently, a novel proposal that treats quantum transport with quantum (Bohmian) trajectories has also demonstrated its ability to provide accurate approximations for the many body problem and the computation of higher moments in time-dependent scenarios [80,81,82] (see Fig. 7). Unfortunately, because of their computational burden, these proposals have only been used to predict noise and time-dependent behavior of very simple and idealized quantum devices.

As happened for the microelectronic industry some time ago, in the next future the nanoelectronic industry will ask for realistic predictions about the time-dependent and noise behavior of these novel quantum devices. Thus, an important effort must be made by the scientific community to develop accurate and versatile quantum simulators providing information beyond the DC predictions and interesting enough for the nanoelectronic industry, in terms of accuracy, versatility and required computational resources.

In this respect, multi-scale modelling will be essential, because the time-dependent behavior of a device or circuit involves the interaction of a large number of particles (charge carriers), but at the same time is influenced by detailed quantum interactions. Therefore it requires a treatment based on a hierarchy of models, with a proper microscopic description of particle correlations, a semiclassical transport model within the potential landscape of the device, and a higher-level description for a circuit made up of more that one device. In addition, to be appealing for the industry, simulation codes must be really simple to use, with a very intuitive user interface and automated setup of the grids, convergence parameters, etc.

Moreover, when going into the nanoscale in order to reach higher operation frequencies (arriving to the THz range), it may also be necessary to solve the Maxwell equations instead of using the static Poisson approximation [83]. Also, slow mechanisms such as carrier generation/recombination

and the associated time evolution of the charge state of traps and surface states can be crucial for the device operation, but unaffordable simulation times would be necessary to account for such processes. Another example is the treatment of self-heating of the devices, which can be critical at the nanoscale [84,85]. The combined solution of Bolztmann (and/or Schrödinger) equation together with those accounting for the heat transfer should be performed with a different time step (heat transfer is much slower than electron transport) and in a much larger simulation domain (possibly including contacts, wiring and heat sinks). As a consequence, the implementation of the multi-scale modelling approach within the MC simulations (or other time domain techniques) may be useful not only for the treatment of the problem at different physical levels, but also at different time and spatial scales. Such a complete multi-scale physical model based on the MC method would be of high interest for the time domain simulation at the nanoscale.

Temporal aspects of switching: spin valves, tunnelling junctions and domain wall motions

It is a well-known fact that most physical phenomena encountered in spintronics are caused by spin-orbit interactions. Although there are ab initio methods to include relativistic effects correctly even for layered systems and to calculate on this level not only the electronic and magnetic structure but also the corresponding electric and optical transport properties [86], the problem of the time-dependence of such phenomena is far more complicated [87].

In order to bridge formal difficulties, multi-scale approaches can be applied. For example by making use of the Landau-Lifshitz-Gilbert equation in terms of an internal energy, derived from the derivative of the relativistically calculated ab initio free energy with respect to the orientation of the magnetization, switching times and switching probabilities of spin valves [88] and tunnelling junctions [89] can be calculated that agree rather well with their experimental counterparts [90,91]. Furthermore, by mapping such free energies and corresponding electric resistivities onto Ohm's law, even critical switching currents can be evaluated [88] that are in good agreement with experiment.

In the case of race track memories [89] relativistically obtained ab initio-like data in terms of the Ginzburg-Landau expansion, can not only be used to evaluate the equilibrium width and the electric properties of domain walls [92], but also, in combination with the Landau-Lifshitz-Gilbert equation as a second step, their speed of motion. Thus even a domain wall motion [93] as underlying principle for a race track memory is accessible.

It is important to note that in describing switching times, switching probabilities or critical currents of spintronic devices the computationally expensive step in all multi-scale procedures is the application of an appropriate ab initio scheme, those then based on phenomenological or macroscopic equations are usually very fast indeed. This ratio of computing efforts applies, e.g., also to the multi-scale procedure to evaluate (time-independent) magneto-optical properties, since the microscopic optical conductivity tensor at a given photo frequency has to be combined with a classical optics package in order to yield macroscopic Kerr rotation and ellipticity angles.

Conclusions

From the analysis that we have performed, multi-scale, multi-physics modelling stands out as the way to go for the simulation of nanoscale devices and circuits, due to the wide range of involved dimensional scales and to the variety of physical phenomena acting at the same time. It is also clear that the structure of the new hierarchies of simulation tools will exhibit conceptual differences with respect to those typical of traditional microelectronics: for example the possibility of a SPICE-like description of circuits is not warranted any

longer, because of the more complex nature of the interaction among devices and the fact that, at least for some material systems, the sharp separation between devices and interconnects seems to be fading. Furthermore, some feedback between levels cannot be avoided, and its handling must be optimized in order to keep the computational effort within reasonable limits.

An effort will then be needed to reach out to wider communities: the language of specialists in many narrow fields will have to be translated into something accessible to device and circuit designers.

There is also a clear need for coordination, benchmarking and validation of the simulation tools, in order to create modelling suites that are supported by the consensus of a vast community and can provide useful information for novel approaches, too, for which intuition alone is not sufficient to predict the main behavior. For the tools to be accessible, computational issues will have to be solved, since it is not expected that in the near future most of the interested players will have access to supercomputing facilities: GPUs may be a possible solution, due to the low cost and the extremely high degree of parallelism. GPUs, however, pose a few difficulties from the point of view of programming, requiring specific skills and quite a bit of effort, in exchange for an order of magnitude gain in computational power. Thus, automated code optimization for GPUs or in general for highly parallel machines will have to be pursued.

In part we need to assemble into multi-scale hierarchies tools that have already been developed, but there is also a lot of work that still needs to be done: for example the quantum mechanical treatment of the time-dependent behavior of nanoscale devices and circuits is still in its infancy, and the same can be said for the treatment of transport through single molecules. Novel functionalities, novel physical aspects, and novel materials are rising in importance, such as thermoelectric devices based on the properties of nanoscale structures, wide-bandgap materials, biomolecules: they require new simulation approaches that need to be integrated into general-purpose simulators.

Accurate, physics-based modelling has been an engine of development in the past, motivating research that has led to breakthroughs with relevant industrial impact and will certainly play an even more important role in the future, for technologies based on the quantum properties of matter. Europe should leverage on the vast expertise and on the excellence that has available, and develop initiatives to stay at the forefront of this field, which, for traditional microelectronics, has so far been dominated by non-European companies.

Acknowledgments

We would like to acknowledge, for useful discussion, suggestions, and support, Chris Ford, Joe Jerome, Jing Guo, Karl Hess, Joe Cychosz, Aldo Di Carlo, Michael Rieth, Mark Ratner, Doris Schmitt-Landsiedel, Jürgen Wilde, Alessandro Toscano, Mihail Nedjalkov, Siegfried Selberherr, Peter Vogl, Asen Asenov, Abel Garcia Barrientos, Robert Baptist, Arnaud Beaumont, Marek Turowski, Anna Perri, Jürgen Sprekels, Ivan Dimov, Fausto Rossi, Christoph Jungemann, Andreas Greiner, David Esseni, Rosana Rodríguez Martínez, Henning Schomerus, Francis Calmon,

This work was financially supported by the European Commission through the Project Mult.Eu,Sim (270062).

References

[1] M. Auf der Maur, M. Povolotskyi, F. Sacconi, A. Pecchia, and A. Di Carlo, J. Comput. Electron. 7, 398 (2008).

[2] CT-INFORMATION AND COMMUNICATION TECHNOLOGIES, WP 2011-12 *ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/ict-wp-2011-12_en.pdf*

[3] P. Marconcini and M. Macucci, La Rivista del Nuovo Cimento 34, 489 (2011).

[4] P. Marconcini and M. Macucci, Carbon 45, 1018 (2007).

[5] P. Marconcini, D. Logoteta, M. Fagotti, and M. Macucci, Proceedings of the 14th International Workshop on Computational Electronics, Pisa, Italia, 26-29 October 2010, edited by G. Basso and M. Macucci, IEEE Conference Proceedings, IEEE Catalog Number: CFP10462-PRT, Ed. Plus, p. 53 (2010).
[6] D. Kauzlaric, J. T. Meier, P. Espanol, S. Succi, A. Greiner, and J. G. Korvink, J. Chem. Phys. 134, 064106 (2011).
[7] "Nano-Electronic Devices: Semiclassical and Quantum Transport Modelling," edited by D. Vasileska and S. Goodnick, (Springer, 2011), and references therein.
[8] F. Ortmann and S. Roche, Phys. Rev. B 84, 180302(R) (2011).
[9] F. Ortmann, F. Bechstedt, and K. Hannewald, New J. Phys. 12, 023011 (2010).
[10] A. Cresti, N. Nemec, B. Biel, G. Niebler, F. Triozon, G. Cuniberti, and S. Roche, Nano Research 1, 361 (2008).
[11] G. Iannaccone, G. Fiori, M. Macucci, P. Michetti, M. Cheli, A. Betti, and P. Marconcini, Proceedings of IEDM 2009, Baltimora, USA, 7-9 December 2009, IEEE Conference Proceedings, IEEE Catalog Number: CFP09IED-CDR, p. 245 (2009).
[12] B. Biel, X. Blase, F. Triozon, and S. Roche, Phys. Rev. Lett. 102, 096803 (2009).
[13] P. Robinson, H. Schomerus, L. Oroszlány, and V. I. Fal'ko, Phys. Rev. Lett. 101, 196803 (2008).
[14] P. Marconcini, G. Fiori, M. Macucci, and G. Iannaccone, J. Comput. Electron. 7, 415 (2008).
[15] N. Singh, A. Agarwal, L. K. Bera, T. Y. Liow, R. Yang, S. C. Rustagi, C. H. Tung, R. Kumar, G. Q. Lo, N. Balasubramanian, and D. L. Kwong, IEEE Electron Device Letters 27. 383 (2006).
[16] J.-P. Colinge, C. Lee, A. Afzalian, N. Dehdashti Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A. M. Kelleher, B. McCarthy, and R. Murphy, Nature Nanotechnology 5, 225 (2010).
[17] L Ansari, B. Feldman, G. Fagas, J.-P. Colinge, and J. C. Greer, Appl. Phys. Lett. 97, 062105 (2010).
[18] Bing-Yue Tsui and Chi-Pei Lu, "Current Transport Mechanisms of Schottky Barrier and Modified Schottky Barrier", in the Proceedings of ESSDERC 2007.
[19] C. Sandow, J. Knoch, C. Urban. Q.-T. Zhao, and S. Mantl, Solid-state Electronics 53, 1126 (2009).
[20] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, Phys. Rev. B 74, 205323 (2006).
[21] E. Alfinito, C. Pennetta, and L. Reggiani, Nanotechnology 19, 065202 (2008).
[22] E. Alfinito, C. Pennetta, and L. Reggiani, J. Appl. Phys. 105, 084703 (2009).
[23] E. Alfinito, and L. Reggiani, Europhysics Letters 85, 68002 (2009).
[24] S. J. Pearton, F. Ren, A. P. Zhang, K.P. Lee, Materials Science and Engineering R 30, 55 212 (2000).
[25] M. A. Khan, J. N. Kuznia, D. T. Olson, W. Schatt, J. Burm, and M. S. Shur, Appl. Phys. Lett., 65, 1121 (1994).
[26] Y. F. Wu, P. M. Chavarkar, M. Moore, P. Parikh, B. P. Keller, and U. K. Mishra, Int. Electron Devices Meeting, CA, 2000.
[27] T. H. Yu and K. F. Brennan, J. Appl. Phys. 89, 3827 (2001).
[28] Y. F. Wu, B. P. Keller, P. Fini, S. Keller, T. J. Jenkins, L. T. Kehias, S. P. DenBaars, and U. K. Mishra, J. Appl. Phys. 19, 50 (1998).
[29] H. Kim, V. Tilak, B. M. Green, H. Cha, J. A. Smart, J. R. Shealy, and L. F. Eastman, in Proceedings of the Thirty-Ninth International Reliability Physics Symposium (IEEE, Piscataway, NJ) 214, 2001.
[30] S. C. Binari, P. B. Klein, and T. E. Kazior, Proc. IEEE 90, 1048 (2002).
[31] K. V. Smith, Proceedings of the GaAs Reliability Workshop, Monterey, CA, 2002 (IEEE, Piscataway, NJ) 2002.
[32] U. Bandelow, R. Hünlich, and T. Koprucki, IEEE Journal of selected topics in quantum electronics 9, 798 (2003).
[33] U. Bandelow, H. Gajewski, and R. Hünlich, Chapter 3: "Fabry--Perot Lasers: Thermodynamics-based Modelling, in: Optoelectronic Devices --- Advanced Simulation and Analysis," J. Piprek, ed., Springer, (New York, 2005) p. 63.

[34] Th. Koprucki, H.-C. Kaiser, and J. Fuhrmann, "Electronic states in semiconductor nanostructures and upscaling to semi-classical models," in A. Mielke, Editor, Analysis, Modelling and Simulation of Multiscale Problems, Springer Verlag, (Heidelberg, 2006), p. 365.

[35] Th. Koprucki, A. Wilms, A. Knorr, and U. Bandelow, Optical and Quantum Electronics 42, 777 (2011).

[36] S. Steiger, R.G. Veprek, and B. Witzigmann, J. Comput. Electron. 7, 509 (2008).

[37] Y. S. Yu, S. W. Hwang, and D. Ahn, IEEE Trans. Elec. Dev. 46, 1667 (1999).

[38] Y. S. Yu, S. W. Hwang, and D. Ahn, IEE Proc.- Circuits, Devices, Syst. 152, 691 (2005).

[39] M. Girlanda and M. Macucci, J. Appl. Phys. 92, 536 (2002).

[40] A. Naeemi and J. D. Meindl, IEEE Electron Device Letters 29, 497 (2008).

[41] Synopsys Inc., Sentaurus Device User Guide, Version 2010.06, Mountain View, California (2009).

[42] A. Schenk and A. Wettstein, "Simulation of DGSOI MOSFETs with a Schrödinger-Poisson Based Mobility Model", in Proc. SISPAD, p. 21, Kobe (Japan), Sept. 4-6, 2002.

[43] M. Auf der Maur, G. Romano, G. Penazzi, F. Sacconi, A. Pecchia, and A. Di Carlo, IEEE Transactions on Electron Devices 58, 1425 (2011).

[44] U. Krumbein, D. Yoder, A. Benvenuti, A. Schenk, and W. Fichtner , "Full-Band Monte Carlo Transport Calculation in an Integrated Simulation Platform," Proceedings of SISDEP'95, vol. 6, p. 400 , Erlangen, Germany, September 6-8, 1995.

[45] M. Auf der Maur, M. Povolotskyi, F. Sacconi, A. Pecchia, G. Romano, G. Penazzi, and A. Di Carlo, Opt. Quant. Electron. 40, 1077 (2008).

[46] M. Shur, "Plasma wave terahertz electronics", 33rd International Conference on Infrared, Millimeter and Terahertz Waves, 2008. IRMMW-THz 2008, 15-19 Sept. 2008, Pasadena, CA, p. 1.

[47] M. Dyakonov and M. Shur, Phys. Rev. Lett. 71, 2465 (1993).

[48] Harry A. Atwater and Albert Polman, Nature Materials 9, 205 (2010).

[49] V. V. Popov, A. N. Koudymov, M. Shur, and O. V. Polischuk, J. Appl. Phys. 104, 024508 (2008).

[50] G. J. Snyder and E. S. Toberer, Nature Materials 7, 105 (2008).

[51] M. S. Dresselhaus, G. Chen, M. Y. Tang, R. G. Yang, H. Lee, D. Z. Wang, Z. F. Ren, J.-P. Fleurial, and P. Gogna, Advanced Materials 19, 1043 (2007).

[52] A. J. Minnich, M. S. Dresselhaus, Z. F. Ren, and G. Chen, Energy Environ. Sci. 2, 466 (2009).

[53] W. H. Butler, X.-G. Zhang, T. C. Schulthess, and J. M. MacLaren, Phys. Rev. B 63, 054416 (2001).

[54] S. S. P. Parkin, C. Kaiser, A. Panchula, P. M. Rice, B. Hughes, M. Samant, and S. Yang, Nature Materials 3, 862 (2004).

[55] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, Nature Materials 3, 868 (2004).

[56] M. Schröter, G. Wedel, B. Heinemann, C. Jungemann, J. Krause, P. Chevalier, and A. Chantre, IEEE Transactions on Electron Devices 58, 3687 (2011).

[57] M. Schröter, J. Krause, N. Rinaldi, G. Wedel, B. Heinemann, P. Chevalier, and A. Chantre, IEEE Transactions on Electron Devices 58, 3697 (2011).

[58] C. Jacoboni and P. Lugli, "The Monte Carlo Method for Semiconductor Device Simulation" (Springer-Verlag Wien, New York, 1989).

[59] K. Hess, "Advanced Theory of Semiconductor Devices," (IEEE Press, Piscataway, 2000), p. 129.

[60] T. Gonzalez, J. Mateos, D. Pardo, and L. Reggiani, Semicond. Sci. Tech. 14, L37 (1999).

[61] K. Kalna, L. Yang, and A. Asenov, Proc. of the 35th European Solid-State Device Conference, 169 (2005).

[62] J. Mateos, T. González, D. Pardo, V. Hoel, H. Happy, and A. Cappy, IEEE Trans. Electron Devices, 47, 250 (2000).

[63] E. Pascual, R. Rengel, and M. J. Martin, Semicond. Sci. and Tech. 22, 1003 (2007).
[64] P. Lugli and D. K. Ferry, IEEE Trans. Electron Devices 32, 2431 (1985).
[65] M. V. Fischetti and S. E. Laux, Phys. Rev. B 48, 2244 (1993).
[66] L. Lucci, P. Palestri, D. Esseni, L. Bergagnini, and L. Selmi, IEEE Trans. Electron Devices 54, 1156 (2007).
[67] P. Palestri, L. Lucci, S. Dei Tos, D. Esseni, and L. Selmi, Semiconductor Science and Technology, 25, 055011 (2010).
[68] M. Lenzi, P. Palestri, E. Gnani, S. Reggiani, A. Gnudi, D. Esseni, L. Selmi, and G. Baccarani,. 55, 2086, IEEE Transactions on Electron Devices (2008).
[69] D. Querlioz, J. Saint-Martin, K. Huet, A. Bournel, V. Aubry-Fortuna, S. Galdin-Retailleau, and P. Dollfus, IEEE Trans. Electron Devices 54 2232 (2007).
[70] Y. M. Blanter and M. Büttiker, Phys. Rep. Rev. 336, 2 (2000).
[71] R. Landauer, Physica Scripta T42, 110 (1992).
[72] M. Di Ventra, "Electrical Transport in Nanoscale systems" (Cambridge University Press, 2008).
[73] R. Landauer, Philos. Mag. 21, 863 (1970).
[74] R. Landauer, IBM J. Res. Dev. 1, 223 (1957).
[75] M. Büttiker, A. Prêtre, and H. Thomas, Phys. Rev. Lett. 70, 4114 (1993).
[76] M. Büttiker, Phys.Rev.Lett. 65, 2901 (1990).
[77] L. S. Levitov and G. B. Lesovik, Pis'ma Zh. Eksp. Teor. Fiz 55, 555 (1992).
[78] L. S. Levitov, H. Lee, and G. B. Lesovik, J. Math. Phys. 37, 4845 (1996).
[79] Yu. Bomze, G. Gershon, D. Shovkun, L. S. Levitov, and M. Reznikov, Phys. Rev. Lett. 95, 176601 (2005).
[80] X. Oriols, Phys. Rev. Lett. 98, 066803 (2007).
[81] G. Albareda, J. Suñé and X. Oriols, Phys. Rev. B 79, 075315 (2009).
[82] G. Albareda, H. López, X. Cartoixà, J. Suñé, and X. Oriols, Phys. Rev. B 82, 085301 (2010).
[83] J. S. Ayubi-Moak, S. M. Goodnick, S. J. Aboud, M. Saraniti, and S. El-Ghazaly, J. Comput. Electron. 2, 183 (2003).
[84] E. Pop, R. W. Dutton, and K. E. Goodson, Appl. Phys. Lett. 86, 082101 (2005).
[85] S. Vitanov, V. Palankovski, S. Maroldt, and R. Quay, Solid-State Electronics 54, 1105 (2010).
[86] J. Zabloudil, R. Hammerling, L. Szunyogh, and P. Weinberger, "*Electron Scattering in Solid Matter: a theoretical and computational treatise,*" Springer Berlin Heidelberg New York, 2005, ISBN 3-540-22524-2; P. Weinberger, "*Magnetic Anisotropies in Nanostructured Matter,*" CRC Press, 2009, ISBN 978-1-4200-7265-5.
[87] A. Vernes, B. L. Gyorffy, and P. Weinberger, Phys. Rev. B 76, 012408 (2007); A. Vernes, L. Szunyogh, and P. Weinberger, Phys. Rev. B 78, 155129 (2008).
[88] P. Weinberger, Phys. Rev. B 83, 134413 (2011).
[89] D. Bedau, H. Liu, J.-J. Bouzaglou, A. D. Kent, J. Z. Sun, J. A. Kantine, E. E. Fullerton, and S. Mangin, Appl. Phys. Lett. 96, 022514 (2010).
[90] Y.-T. Cui, G. Finocchio, C. Wang, J. A. Katine, R. A. Buhrman, and D. C. Ralph, Phys. Rev. Lett. 104, 097201 (2010).
[91] L. Thomas, M. Hayashi, X. Jiang, R. Moriya, C. Rettner, and S. S. P. Parkin, Science 315, 1553 (2007).
[92] P. Weinberger, Phys. Rev. Lett. 100, 017201 (2008).
[93] P. Weinberger, E. Y. Vedmedenko, R. Wieser, and R. Wiesendanger, Phil. Mag. 91, 2248 (2011).