Enabling cutting-edge semiconductor simulation through grid technology

DAVE REID 1,*, CAMPBELL MILLAR 1, SCOTT ROY 1, GARETH ROY 1, RICHARD SINNOTT 2, GORDON STEWART 2, GRAEME STEWART 3 AND ASEN ASENOV 1

1 Device Modelling Group, 2 National e-Science Centre, and 3 Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, UK

The progressive scaling of complementary metal oxide semiconductor (CMOS) transistors drives the success of the global semiconductor industry. Detailed knowledge of transistor behaviour is necessary to overcome the many fundamental challenges faced by chip and systems designers. Grid technology has enabled the unavoidable statistical variations introduced by scaling to be examined in unprecedented detail. Over 200,000 transistors have been simulated, the results of which provide detailed insight into underlying physical processes. This paper outlines recent scientific results of the nanoCMOS project and describes the way in which the scientific goals have been reflected in the grid-based e-Infrastructure.

Keywords: nano complementary metal oxide semiconductor electronics; virtual organization; security; variability

1. Introduction to nanoCMOS challenges

The progressive scaling of complementary metal oxide semiconductor (CMOS) transistors drives the success of the global semiconductor industry. This is often described by the widely known Moore’s Law (ITRS 2005). As device dimensions approach the nanometre scale, however, chip and systems designers must overcome many fundamental challenges. The Engineering and Physical Sciences Research Council (EPSRC)-funded project meeting the design challenges of nanoCMOS electronics (nanoCMOS) has been formed to explore and tackle the problems caused by these atomistic scale effects throughout the semiconductor electronics design process.

In future technology generations, the industry is primarily concerned with unavoidable intrinsic parameter fluctuations, where the behaviour of individual solid-state components varies due to effects caused by the inherent discreteness of charge and matter. The EPSRC-funded, nanoCMOS e-Science pilot project (Sinnott et al. 2006) aims to apply e-Science techniques in such a way as to support the computationally intensive simulation methodologies required to gain a deeper understanding of the various sources of statistical variability and
their impact on circuit and system design. These technologies will also enable the management and manipulation of the large amounts of resultant simulation output data. The large numbers of simulation are necessitated by variations in the atomic structure of nanoscale devices, which require three-dimensional simulation of ensembles of devices to be performed, rather than a single idealized device (Frank & Taur 2002), as has been the norm. The increasing number of transistors in modern chips also necessitates the simulation of very large statistical samples to allow the study of statistically rare devices, with potentially detrimental effects on circuit performance, to be examined. Previously, the computational complexity of three-dimensional device simulation has restricted studies of variability to small ensembles of approximately 200 devices (Roy et al. 2006; Brown et al. 2007); however, as we show, this results in inaccurate predictions of the statistical distribution of transistor behaviour (Millar et al. 2008). Therefore, it is necessary to employ grid technologies to simulate the large statistical samples required and to apply statistical and data mining techniques to process the large amount of output data generated.

The increasingly small dimension of modern transistors means that both the number and position of individual dopant atoms inside are beginning to affect the behaviour of these devices. A metal oxide semiconductor field effect transistor (MOSFET) is essentially a gate-controlled switch. The switching gate voltage is called threshold voltage ($V_T$), and variation of this parameter between the transistors within circuits is the primary source of timing and power variability in modern chips. In the current 45 nm technology generation (ITRS 2005), the main source of threshold voltage fluctuation in bulk MOSFETs comes from random discrete dopants (RDDs). The billions of transistors that constitute modern chips also mean that statistically rare devices (beyond $5–6\sigma$ from the mean) are beginning to have a significant effect on designs and yield. In this paper, we present groundbreaking results in which ensembles in excess of 100 000 three-dimensional devices have been simulated using a grid-based methodology, for 35 nm (Inaba et al. 2002) and 13 nm gate-length devices. In total, these simulations required approximately 20 CPU years on a 2.4 GHz Advanced Micro Devices (AMD) Opteron system.

2. nanoCMOS e-Infrastructure and simulation methodology

The nanoCMOS project has adopted a hierarchical simulation methodology, which is shown in figure 1. In this paper, we are primarily concerned with the bottom level of this hierarchy, where individual devices are simulated ‘atomistically’. Results from this stage of the methodology will be passed up to higher levels of abstraction in the design flow in the form of transistor compact models, which will be used in circuit-level simulation program with integrated circuit emphasis (SPICE)-like (Berkeley SPICE—http://bwrc.eecs.berkeley.edu/Classes/icbook/SPICE/) simulations.

To support these simulations, the project has worked with a variety of grid middleware and middleware providers, including many of the technologies provided by the Open Middleware Infrastructure Institute (OMII-UK—http://www.omii.ac.uk) and the enabling grids for e-Science (EGEE—http://public.eu-egee.org) project. The early phase of work focused upon development of a family of
OMII-UK services, which supported the device modelling and compact model generation phases of electronics design. These services were developed to exploit the OMII-UK GridSAM job submission system (GridSAM—Grid Job Submission and Monitoring Web Service, http://gridsam.sourceforge.net).

The aim of GridSAM is to provide a web service for submitting and monitoring jobs managed by a variety of distributed resource managers (DRMs). This web service interface allows jobs to be submitted from a client in a job submission

Figure 1. Hierarchical simulation methodology adopted by the nanoCMOS project.
description language (Anjomshoaa et al. 2005) document and supports retrieval of their status as a chronological list of events detailing the state of the job. GridSAM translates the submission instruction into a set of resource-specific actions—file staging, launching and monitoring—using DRM connectors for each stage. Proof-of-concept nanoCMOS job submission solutions were implemented with GridSAM that showed how access to resources such as the National Grid Service (NGS), Sun grid engine clusters and Condor pools at Glasgow could be supported. However, several limitations were identified with GridSAM, which were fed back to OMII-UK. Among other things, these included issues with GridSAM failing if all of the files specified for staging were not present, and problems with staging binary files. Both of these have been acknowledged as issues by OMII-UK, and are currently being resolved for future technology releases.

However, the main computational resource available to nanoCMOS was the ScotGrid cluster (http://www.scotgrid.ac.uk/) at the University of Glasgow. This infrastructure has been established as a shared resource across campuses and is fully integrated into the EGEE project. Since November 2007, ScotGrid has also been made available to the NGS. The ScotGrid cluster supports the EGEE middleware stack, including its own particular flavour of job submission and management software, all based around the gLite middleware. One of the front-end software environments used for EGEE job submission is GANGA (http://ganga.web.cern.ch/ganga/). GANGA supports bulk job submission, where the maximum number of concurrent jobs per user is limited to 1000, but since multiple devices can be simulated in a single job this was sufficient for our methodology. While this method of submission alleviates many of the issues with large parallel job submissions, GANGA is not without problems. GANGA automates the process of job submission and monitoring, but, since it is a front-end to the existing system, some of the deficiencies of grid middleware such as Globus are still evident; an example is job submission, which is extremely slow (approx. 1–2 hours to submit 500 jobs). Additionally, it was also found that GANGA sometimes fails to properly track jobs (due to both bugs in GANGA and resource broker problems in our experience), resulting in it becoming impossible for the user to control them, and requires administrator intervention to cancel the execution of such rogue jobs.

Despite these issues, the nanoCMOS researchers have become the primary end users of the ScotGrid cluster with a combined CPU usage of 23 per cent of the total resource since the project started. We note that prior to the nanoCMOS project, the electronics community at Glasgow had less than 1 per cent usage of ScotGrid, despite it being a shared campus resource. Thus far, the simulations undertaken in nanoCMOS have resulted in over 175 000 CPU hours of device simulations run on ScotGrid. More details on the implementation of the simulation framework for nanoCMOS are available in Han et al. (2007).

In running such large-scale simulations for nanoCMOS researchers, one of the primary challenges that have been faced is in dealing with the data and metadata associated with the simulations. This is particularly complicated given the commercial sensitivity of some of the datasets that are being dealt with. The project has explored a variety of data management and security solutions in this space. Given that the vast majority of the simulation data are file-based, a performance and security evaluation of the storage resource broker (http://www.sdsc.edu/srb/index.php) and the Andrew file system (AFS) (Zayas 1991) was
undertaken. The results of this are described in Sinnott et al. (2008d) along with our justification for the adoption of AFS. A variety of metadata associated with simulations are captured and made available to targeted metadata services, which are in turn coupled with the predominantly file-based AFS data. A range of query interfaces to these metadata systems are also under production.

Key to both the data management and simulation frameworks of nanoCMOS is support for fine-grained security. The need to protect access to commercial resources is paramount for many of the industrial partners involved in the project. We have identified that no single security solution fulfils the needs of nanoCMOS research. Instead, it has been necessary to integrate a range of security solutions, including Kerberos (for AFS; Kohl et al. 1994), the virtual organization membership service (Alfieri et al. 2003), MyProxy (Basney et al. 2005), PERMIS (Chadwick et al. 2003), Globus Security Infrastructure (http://www.globus.org/security/) and Internet2 Shibboleth architecture and protocols (http://shibboleth.internet2.edu). The justification for integrating these technologies and the way in which they have been integrated are described in more detail in Sinnott et al. (2008a–c).

The actual simulations carried out at the base of figure 1 primarily involve using the Glasgow three-dimensional ‘atomistic’ drift/diffusion (DD) simulator (Roy et al. 2006). The DD approach accurately models transistor characteristics in the sub-threshold regime, making it well suited for the study of $V_T$ fluctuations. The simulator includes density gradient quantum corrections (Ancona & Tiersten 1987), which accurately capture quantum confinement effects and are essential for preventing artificial charge trapping in the sharply resolved Coulomb potential of discrete impurities. Each device is also fully independent, allowing the problem to be easily parallelized using a task-farming approach. The 35 nm gate length transistor used in the simulation studies was published by Toshiba (Inaba et al. 2002) and the simulator was calibrated to the experimental characteristics of this device. Structural data for the device were obtained through commercial technology computer-aided design (TCAD) process simulation. The doping profile structure and characteristics of the Toshiba device are shown in figure 2.

In order to accurately model $V_T$ variations, it is necessary to simulate very large numbers of devices. This is necessary both to reduce statistical noise in the resulting parameter distributions and to allow statistically rare devices to be studied in detail. Performing these large numbers of simulations required significant technical challenges to be overcome. It is extremely important to be able to track failed and/or numerically unstable simulations, since it is of vital importance to ensure that duplicate devices are not included in the output ensemble in order to preserve the integrity of statistical calculations. It is generally beyond the scope of the grid software to track specific conditions within an individual job, so various tools had to be developed to manage the output data during the lifetime of active simulations.

3. nanoCMOS scientific results and discussion

Having performed 100 000 three-dimensional simulations of the 35 nm device, we have been able to investigate its properties on a statistical level, and the distribution of random dopant-induced threshold voltage fluctuations in the
simulated ensemble can be seen in figure 3. These fluctuations arise from the fact that each macroscopically similar device will have a different number of and microscopically different configurations of dopant atoms. It is important to have

Figure 2. $I_D$ $V_G$ characteristics of the 35 nm Toshiba device as produced by the Glasgow simulator at a low drain voltage of 50 mV and a high drain voltage of 850 mV, calibrated against results obtained from both TCAD process simulation and experiment. The continuous doping profile is shown as inset. Black solid lines, 35 nm Toshiba MOSFET; grey solid line, Taurus Simulation; dashed line, three-dimensional DD simulator.

Figure 3. (a) The distribution of simulated $V_T$ data compared with Gaussian and Pearson IV distributions. (b) Tails of the $V_T$ distribution. The inaccuracy of the Gaussian fit to the distribution in the tails of the distribution can clearly be seen. Filled circles, data; dashed line, Pearson IV; solid line, Gaussian.

simulated ensemble can be seen in figure 3. These fluctuations arise from the fact that each macroscopically similar device will have a different number of and microscopically different configurations of dopant atoms. It is important to have

Phil. Trans. R. Soc. A (2009)
accurate models of the tails of the distribution (at 6σ and greater), so that the impact of statistically rare devices on circuit performance may be properly assessed. This is presently possible only through a ‘brute force’ approach, necessitating very large numbers of simulations. In order to assess the accuracy of the simulation ensemble, we have calculated the $\chi^2$ errors of the statistical moments of the threshold voltage distribution as a function of the ensemble size. These are shown in figure 4, where it can be seen that enabling larger scale simulations using grid technology has greatly improved the accuracy of our description of random dopant fluctuations.

It can be seen in figure 3 that RDD-induced variations are asymmetric in nature. This is confirmed by the non-zero skew value, and such asymmetry has been reported in experimental measurements (Nassif et al. 2007), which provide a great confidence in the accuracy of the three-dimensional simulation methodology. Currently, it is generally assumed that RDD-induced fluctuations are Gaussian in nature. However, if a Gaussian with the data mean and standard deviation is compared with the experimental distribution, we find a relatively large $\chi^2$ error of 2.42. Fitting the data using a distribution that has skew and kurtosis, such as a Pearson type IV (Heinrich 2004), results in a much better fit, with a $\chi^2$ error of 0.38. The analytical fits are shown in figure 3a,b for comparison.

Table 1. Statistical moments of the $V_T$ distributions. (Note that the mean value has been normalized to the experimental mean threshold.)

<table>
<thead>
<tr>
<th>distribution</th>
<th>mean (mV)</th>
<th>s.d. (mV)</th>
<th>skew</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>225.9</td>
<td>30.28</td>
<td>0.1597</td>
<td>0.0486</td>
</tr>
<tr>
<td>calculated Gaussian</td>
<td>225.9</td>
<td>30.28</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Pearson IV</td>
<td>225.894</td>
<td>30.37</td>
<td>0.16168</td>
<td>0.0811</td>
</tr>
</tbody>
</table>

Figure 4. $\chi^2$ error of the statistical moments as a function of sample size. It is assumed that the final values are the population values. Black solid line, mean; grey solid lines, s.d.; dark grey lines, skew; black dashed lines, kurtosis.
and the moments of the distribution and analytical fits are given in table 1. A more detailed view of the tails is shown in figure 3b, where the systematic error in the Gaussian is more apparent. Most semiconductor designers, and design tools, assume a Gaussian distribution of $V_T$ in order to make decisions about the safety margins and robustness of a given design. However, as we have seen, this variation is not Gaussian in nature, which can lead to the incorrect estimation of design margins at higher levels in the semiconductor design-chain and this can have significant adverse effects on production yield optimization.

The effect of RDDs on an individual device can be most readily understood by examining the electrostatic potential within the devices. Figure 5 shows the potential profiles extracted from devices taken from the upper and lower tails, along with a ‘mean’ device chosen from the centre of the distribution for comparison. This clearly illustrates the physical causes of variation in device characteristics, since the behaviour of a MOSFET is determined by the height of the potential barrier in the channel. It can be seen that, even at the nominal threshold voltage, the device from the lower tail has already been switched on and the device from the upper tail is off. While statistically rare devices can be generated artificially, it is only through the large-scale simulation undertaken here that we can obtain the correct statistical description that governs the occurrence of such devices and through this can be achieved a better understanding of the underlying physical causes of variability.

In sub-50 nm devices, there are typically fewer than 100 dopant atoms which are significant in determining the behaviour of the device. In order to determine which dopants are important, data mining techniques were applied to the generated data. Each device within the ensemble was analysed by sectioning into 1 nm thick volumes in the $x$ and $z$ directions, as demonstrated in figure 6a. Having done this, the correlation between the number of dopants in each volume and the measured $V_T$ of the device could be calculated. This provides a two-dimensional description showing the correlation between dopant position and threshold voltage, as seen in figure 6b. This correlation plot could then be used to estimate the relative effect of a dopant at a given position on the threshold voltage of the devices. This defines the statistically significant region (SSR) where individual dopants could affect the bulk device behaviour, which, as expected, corresponds approximately to the device channel but also includes some small portions of the source and drain regions as well.

For a fixed number of dopants in the SSR, there will also be a distribution of threshold voltage arising from the different positional configurations of dopants, which can occur in the silicon lattice. The $V_T$ distributions for $n=35$, 45 and 55 dopants (corresponding to lower tail, mean and upper tail devices, respectively)
in the SSR can be seen in figure 7. From this it is clear that both mean and standard deviation of $V_T$ increase with the number of dopants. By further examining this relationship using all the available data, it can be determined that the mean and standard deviation depend linearly on the number of dopants (Reid et al. 2008). Therefore, this relationship can be extrapolated to arbitrary numbers within the SSR as necessary.

From the knowledge of the physical processes governing dopant implantation, the number of dopants within the SSR must be governed by a Poisson distribution. Therefore, a $V_T$ distribution can be constructed from the convolution of this Poisson distribution with the Gaussian distributions from the random position of dopants (Reid et al. 2008), which extends to very large values of $\sigma$. This calculation is illustrated graphically in figure 8. The linear
relationship between the number of dopants and the mean and standard deviation of the positional Gaussians allows the distribution to be extrapolated to an arbitrary value of $\sigma$, resulting in a significantly more accurate prediction of the probability of finding devices in tails of the distribution, where real gains can be made in billion transistor count chips. The accuracy of the distribution resulting from this convolution is shown in figure 9a, along with the extrapolated distribution; and the $\chi^2$ errors calculated from the comparison of this distribution with simulation data can be seen in figure 9b. The total calculated values for the $\chi^2$ error are 0.94 for the convolution using simulation data and 0.55 for the extrapolated distribution. These values are comparable to the fitting error of the Pearson IV, demonstrating the accuracy of this analytical description.
4. Conclusions and future work

It is clear that small statistical ensembles do not provide sufficient accuracy and information to design when variations at $6\sigma$ and beyond are important. Large datasets are necessary as some effects, such as the asymmetry of the $V_T$ distribution observed here, are not visible in small datasets and will have seriously detrimental effects if not incorporated into design strategies. We have shown that, by applying grid and e-Science technology to the problem of intrinsic parameter fluctuations, we have gained a fundamental insight into device behaviour in the presence of random dopants. The large datasets obtained have provided sufficient data to develop and verify an accurate analytical model of the underlying physical processes that affect $V_T$ fluctuations in nano-scale semiconductor MOSFETs. Future work will be to explore the impact of these atomic variations throughout the design process. We are also exploring the atomistic variability of a variety of novel device architectures.

This work was funded by a grant from the UK Engineering and Physical Sciences Research Council. We gratefully acknowledge their support.

References


Sinnott, R. O. *et al.* 2006 Meeting the design challenges of nanoCMOS electronics: an introduction to an EPSRC pilot project. In *Proc. UK e-Science All Hands Meeting*.


