VLSI DESIGN
1998, Vol. 6, Nos. (1–4), pp. 147–153
Reprints available directly from the publisher
Photocopying permitted by license only

Hierarchy of Full Band Structure Models for Monte Carlo Simulation

U. RAVAIOLI, A. DUNCAN, A. PACELLI, C. WORDELMAN and K. HESS

Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801

This paper discusses the various hierarchy levels that are possible when the full band structure is considered. At the highest level, the scatterings are treated using complete k-k' transition rates, which entail extremely memory intensive computational applications. At the lowest level, the scattering anisotropy is neglected and the scattering rate is considered to be a constant average value on energy isosurfaces of the bandstructure. This model is more practical for device simulation. In between the two extremes, it is possible·to design intermediate models which preserve some essential features of both. At all levels of the band structure hierarchy of models, there are similar issues of numerical noise, related to the sampling of real and momentum space that the Monte Carlo method necessarily performs with a relatively small number of particles. We discuss here computationally efficient approaches based on the assignment of variable weights to the simulated particles, in conjunction with careful gather-scatter procedures to split particles of large weight and combine particles of small weight.

Keywords: Monte Carlo, Device Simulalation, Semiconductors, Scattering, Bandstructure, Variance Reduction

1. INTRODUCTION

The increasing miniaturization of integrated devices requires simulation models which are more accurate than standard drift-diffusion techniques. Many approaches have been pursued to arrive at a practical solution of the Boltzmann transport equation (BTE) to improve the description of hot carriers in simulation models. The hydrodynamics approach, for instance, is based on moments of the BTE and adds to the drift-diffusion model two equations for the momentum and energy balance [1]. Direct solution of the BTE has been implemented in various forms, including spherical harmonics expansion [2, 3], scattering matrix approach [4], and cellular automata technique [5]. The main advantage of these direct BTE solution approaches is computational efficiency.

Most of the work in the area of hot electron simulation has been carried out with statistical Monte Carlo techniques, where a large sample of carriers is tracked in the phase space, according to free flight and scattering statistics produced with computer generated pseudo-random sequences [6]. Monte Carlo techniques have computational disadvantages, because they requires large memory resources and need large amounts of CPU time, in order to gather sufficiently accurate statistical results and resolve the intrinsic noise of the random processes. However, the Monte Carlo approach has been very popular because the implementation of complete scattering models is relatively simple. The availability of affordable superscalar workstations and vector processors, and advances in optimization of the algorithms [7], have contributed to keep Monte Carlo a competitive technique for physical investigation of semiconductor transport and device properties.

An increasing number of applications require an accurate knowledge of transport at very high energies. The analysis of reliability in deep submicron MOSFETs must resolve impact ionization effects and carrier injection into the oxide in order to arrive at device designs and scaling rules that reduce to a minimum such indesirable phenomena that cause device degradation. The design of flash memories, in contrast, calls for enhancement of hot carrier injection into floating gates beyond the oxide interface barrier. Very high energies, up to several eV's, are involved with such hot carrier phenomena, and for these investigations the Monte Carlo approach has a distinct advantage over other methods, since a complete description of the bandstructure can be included in the model. The simplified bandstructures based on non-parabolic approximations are not realistic at high electron energies, where the density of states, the effective mass, and the carrier velocity deviate enormously from the simple pictures which are calibrated in an energy range close to the bottom of the conduction band.

Here we will review the hierarchy of physical models which have been developed to include the full details of the bandstructure into Monte Carlo models, and we will describe recent work which has the goal to arrive at practical simulation of semiconductor materials and devices with full band Monte Carlo.

2. FULL BAND MONTE CARLO

Early development of Full Band Monte Carlo was pioneered in the early '80s at the University of Illinois [8]. Because of the limited computing resources available at the time, applications were confined to single particle bulk calculations. The first large scale application to device simulation came with the DAMOCLES project developed at IBM, Yorktown Heights [9]. A comprehensive review of the foundations of Full Band Monte Carlo can be found in [7].

The semiconductor bandstructure utilized for the transport model is usually obtained from pseudopotential calculations [10]. The main difference between models of the hierarchy is in the implementation of scattering rates.

2.1. Fúll k-k' scattering model

The most complete scattering model is based on transition rates between any initial state **k** and any possible final state **k'** in the Brillouin zone. Even after taking advantage of the symmetry of the Brillouin zone, a very large set of tables must be precalculated to account for all the possible scattering rates in the various conduction band branches considered. One of difficulties of the approach is to formulate an appropriate physical model for the calculation of the transition rates. Since the bandstructure must be calculated, the same parameters could be used for the evaluation of transition rates, thus eliminating the need to find the deformation potentials which have to be determined empirically for most of the scattering rates used in standard Monte Carlo approaches. Therefore, the unified theory for bandstructure and transport parameters can be based only on several material constants, reducing the most considerable source of uncertainty in the Monte Carlo scattering models that are based on a fit of deformation potentials.

Work in this area was done by using empirical pseudopotentials [11] [12] and *ab initio* pseudopotentials [13]. The electron-phonon scattering model is based on a rigid pseudo-ion approach. The empirical pseudopotentials are assumed to be sums of spherical potentials moving rigidly with the crystal atoms, in electron-phonon calculations. In the *ab initio* density functional approach, ionic pseudopotentials are assumed to be spherical and rigidly transferable from the atom. The total Kohn-Sham equation contains contributions from the self-consistent crystalline charge density under distortion.

The deformation potentials are obtained directly from the theory, and depend on the pseudopotentials used to calculate the bandstructure. Therefore, scattering rates and bandstructure are consistent. The unfortunate aspect of this approach is that some tunable degrees of freedom are lost, leaving less room for calibration. The quality of the transition rates is therefore dependent on the complexity of the theory, and there is still work to be done to achieve a precise quantitative match. The results in [13] show that the deformation potentials, usually taken to be constant for each scattering rate in the standard approaches, may vary quite strongly as a function of momentum and energy.

2.2. Valley-dependent scattering formulation

Models that directly use the **k-k'** transition rates for the scattering tend to be too costly, in terms of computer time and memory, to be suitable for efficient device simulation. Therefore, it is useful to formulate an approximation of this model to improve efficiency while retaining the main physical features. A possible way is to examine the variation of the deformation potentials in the Brillouin zone and define regions (valleys) where it can be reasonably taken to be a constant. This procedure preserves the essential anisotropy of the scattering rate, but within each valley an energy-dependent rate is evaluated, thus avoiding the burden involved with large **k-k'** transition rate tables and reducing the work necessary for final state evaluation.

A preliminary version of this model has been implemented [14], starting from the results of the *ab initio* pseudopotential calculations [13]. Two conduction band branches have been considered, each subdivided into four valleys centered around the Γ, X, L, and K points of the Brillouin zone. The decomposition into valleys requires the definition of suitable intervalley scattering rates to map correctly the **k-k'** transition rates. Figure 1 shows the total scattering rates obtained by averaging the **k-k'** rates over energy shells of the momentum space, and the corresponding rates obtained by performing the relatively simple decomposition into four valleys, as compared to the total integrated rates from the *ab initio* **k-k'** approach. Numerical tests indicate a large improvement in computational efficiency, and work is in progress to refine the overall behavior of the model, in the complete range of electric fields of interest for device simulation.

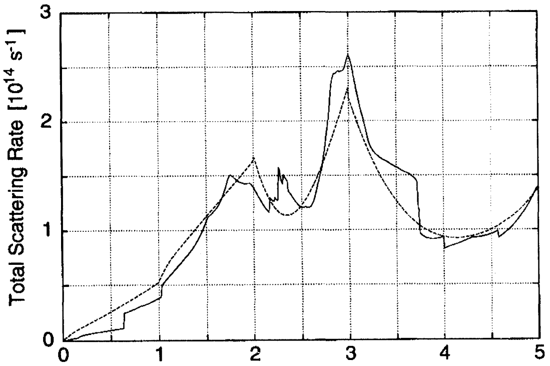


FIGURE 1 Total phonon scattering rate obtained in the valley-dependent formulation by averaging the **k-k'** transition rates over energy shells of momentum space

2.3. Energy-dependent scattering rates

The lower level in the hierarchy of full band models includes energy-dependent scattering rates, as implemented in analytical bands Monte Carlo, but the rates are scaled with the density of states from the actual bandstructure data. The anisotropy of the scattering rates is neglected, but it is preserved within an energy shell for the final state after scattering, because the **k** dependent density of state is considered in the bandstructure tables. This approach preserves the computational efficiency that has been achieved for the traditional analytical band Monte Carlo, and is therefore suitable for large scale device simulation. The goal of our work in this area has been to demonstrate high efficiency on standard workstations, by seeking an optimal trade-off between model accuracy and memory usage/CPU requirements. We present here several examples which are based on a full band implementation with two conduction band branches. An ensemble Monte Carlo approach with constant time technique optimization for the time of flight, which minimizes self-scattering, was used to develop MOS structure simulators. The model has been used as post-processor, to investigate large structures where the potential was calculated with a drift-diffusion approach, and in a self-consistent implementation, where the substrate hole distribution is evaluated within a 2-D non-linear Poisson solution.

The post-processor mode was used to investigate large MOS structures with a single floating gate, intended for long term memory elements. The 2-D potential in the structure, shown in the Fig. 2, was obtained with the PISCES-IIB drift-diffusion simulator, yielding typical potential profiles along the channel as shown also in Fig. 2. The full band Monte Carlo model was then used to map the electron energy distribution throughout the device, by injecting a cold distribution from the source. The energy distributions of electrons in Fig. 3, for several locations under the floating gate, show that only near the left edge of the gate there is a sufficiently high energy tail for injection across the oxide barrier which is around 3 eV. Gate currents were estimated by running a second simulation only for a high energy tail, to achieve statistical enhancement [15]. This work was carried out on HP 735 workstations. While the first type of simulation to obtain energy distributions as in Fig. 3 is fairly rapid, the gathering of statistics for gate currents with the high energy tail simulation is feasible on workstations but requires many hours per data point, particularly at low voltages.

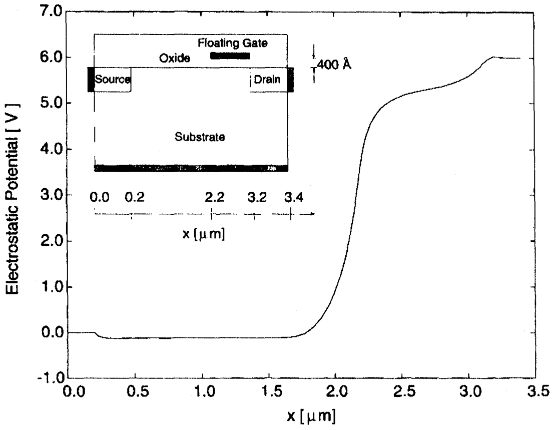


FIGURE 2 Potential profile along the channel of a single floating gate MOS device, shown in the inset. Bias voltages are $V_d = 6.0V$, $V_{fg}= 7.0V$ and $V_B = 0.7V$

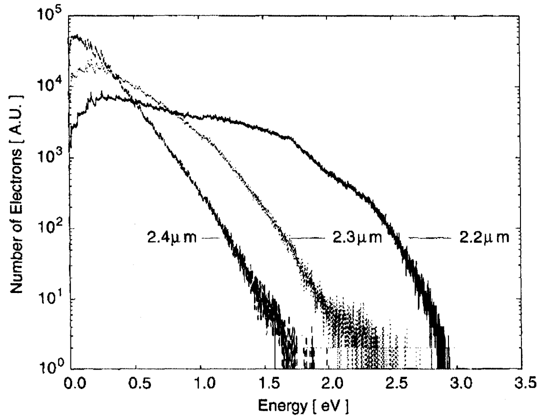


FIGURE 3 Energy distribution of electrons for the structure and biases of Fig. 2, for several locations along the channel

Selfconsistent simulations are now actively pursued to understand the scaling properties of MOS structures, down to 0.1 μm and below. Conventional, LDD, and SOI MOSFETs are considered in our work. We present here a sample of representative results. Figure 4 shows the electric field profile for three conventional MOSFET structures, with 0.3 μm (continuous line), 0.15 μm (broken line) and 0.075 μm (dotted line) channel length, with drain voltage $V_d = 3.0V$ and gate voltage $V_g= 3.0V$. The corresponding energy distribution plots shown in Fig. 5, obtained at the channel-drain junction, indicate that, although higher fields are present, the 0.075 μm structure has a tail which extends less into the high energy region. The fine detail plotted for this case was obtained with the

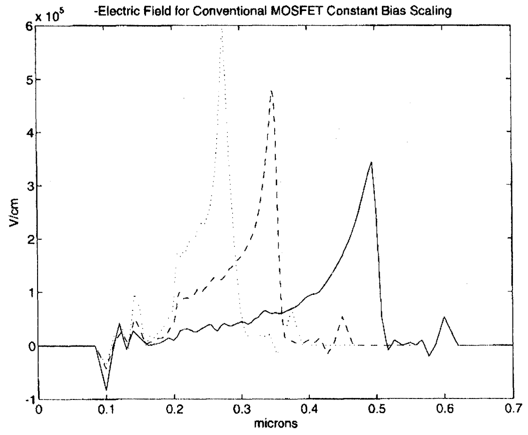


FIGURE 4 Electric field profiles for three scaled conventional MOSFET structures, with 0.3 μm (continuous line), 0.15 μm (broken line) and 0.075 μm (dotted line) channel length, with identical bias $Vd = Vg = 3.0V$

statistical enhancement technique described later, in order to have a verification of this effect. The result suggests that at such a short channel length the electron population does not have enough room to build up a tail. Such effects then must be carefully examined when scaling into such small scale is attempted. The energy distribution, again at the channel-drain junction, is shown in Fig. 6 for the case of bias scaled to $V_g = V_d = 2.12V$ and $V_g = V_d = 1.5V$, for the 0.15 and 0.075 μm devices, respectively. Because of the lower fields at scaled bias, the energy distributions are even cooler.

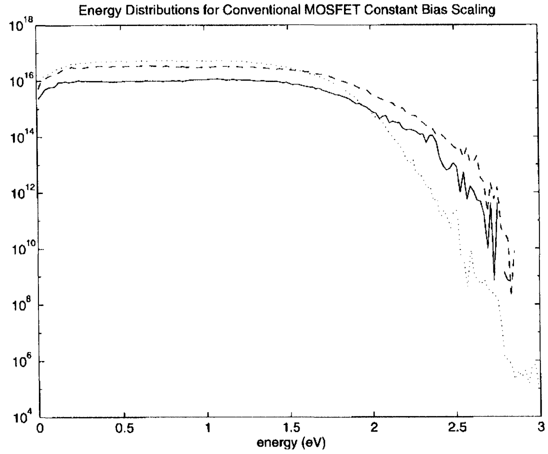


FIGURE 5 Energy distributions of electrons at the channel-drain junction for the structures in Fig. 4

These simulations were obtained with a sample of nearly 20,000 electrons, with a timestep of 1 *fs* for the particle flights and nonlinear solution of Poisson equation at 2 *fs* intervals. In double precision, a maximum of 50MB RAM is necessary for up to 30,000 particles and 200 × 200 nonuniform Poisson grids, while meaningful computational experiments may require quite less. A batch of simulations to characterize a device may take up to several hours on a multiuser environment and on medium power workstations. We find therefore that this full band approach is relatively practical to answer key questions in the design of new structures.

3. VARIANCE REDUCTION

An important problem in Monte Carlo simulations is the presence of statistical noise. In order to make full band approaches more practical for realistic application, it is important to develop strategies for the statistical enhancement of rare events and high energy tails, through variance reduction schemes. This is a key factor which plays as important a role as improvement of physical models and efficient implementation of Monte Carlo algorithms. We have developed a selfconsistent variance reduction scheme based on the assignment to the i-th particle in the simulation a weight w_i, where we assume that each Monte carlo particle represents a number of electrons and holes (with associated mass and charge) proportional to its weight. In conventional simulation particles have the same weight, and the variance reduction technique consists of manipulating the particle weight to equalize the sampling of phase space. A larger number of particles with smaller weights are placed in sparsely populated regions of the phase space. To do so, a large particle can be split into many identical particles with a fraction of the original weight. Gathering of particles is performed to avoid an indefinite increase of particles in the actual simulation. Gathering reduces a set of N particles into one larger particle

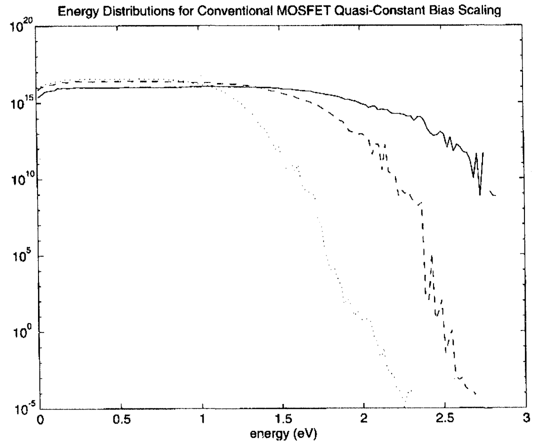


FIGURE 6 Energy distribution of electrons for the structures in Fig. 4 with scaled bias $V_g = V_d = 2.12V$ (0.15 μm channel) and $V_g = V_d = 1.5V$(0.075 μm channel)

without affecting the simulation results appreciably. The method used to perform gathering is fairly crucial, because if the states of N particles are simply averaged into one larger particle, unphysical effects might take place. The averaging of two particles with opposite wave vectors (in a symmetric band) would result in a particle with zero momentum, violating energy conservation well beyond any approximation tolerance of the Monte Carlo method.

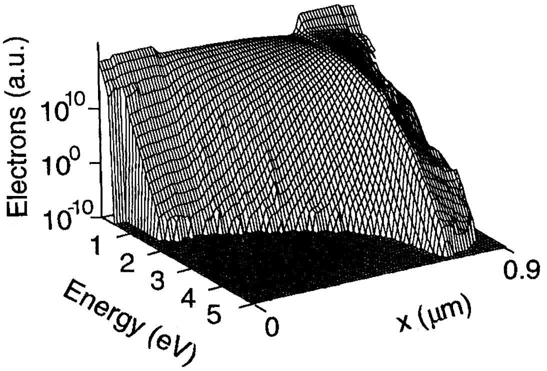


FIGURE 7 Energy distribution of electrons inside a conventional MOSFET with 0.5 μm channel, obtained with variance-reduction applied to the full band Monte Carlo simulation

We perform gathering by selecting one particle at random from the chosen set, removing the remaining ones from the simulation. To ensure charge conservation, the cumulative weight is assigned to the particle which is retained. While momentum and energy are not conserved for a particular random choice, the gathering can be performed so that the expected values of the observables are unchanged. The possible gathering strategies can be *local* and *non - local.* A local gathering set includes particles from a single bin in phase space, while a non-local set includes particles from many bins. A simple local strategy applies gathering to bins with many particles, considering them *over sampled,* but in this way only number, not weight, of particles would be considered. As an alternative, one could look for bins with particles that do not appreciably contribute to the estimators, because large and small particles coexist. However, if small particles in such situations are eliminated systematically, the variance of the distribution close to the boundaries between bins may actually increase.

We apply a non-local gathering strategy, where we keep a linked list of small particles lying in bins that contain much larger particles. Gathering is then performed by ensuring that variance of the estimators is properly reduced. We present in Fig. 7 an example obtained for a 0.5 μm MOSFET structure, with $V_d = 3.0V$, where the electron energy distribution is smoothly resolved with a dynamic range of 20 orders of magnitude throughout the device. Such a variance reduction technique was also applied in the MOSFET simulations described above, with considerable contributions to the overall efficiency.

4. CONCLUSIONS

Full band Monte Carlo approaches have progressed considerably in the last decade, and are now becoming practical physical investigation tools on commonly available workstations. The simplest approach with energy-dependent scattering, coupled with a careful variance reduction scheme, is already fairly practical for selfconsistent simulation of deep submicron structures and can be a valuable tool to help the designer of new integrated device families.

Acknowledgements

This work is supported by the Joint Services Electronics program, grant n. N00014-90-J-1270, and the Semiconductor Research Corporation, contract n. 95-SJ-406.

References

[1] K. Blotekjaer, "Transport Equations for Electrons in Two-Valley Semiconductors," *IEEE Trans. Electron Devices,* vol. ED-17, p. 38, 1970.

[2] D. Ventura, A. Gnudi, G. Baccarani and F. Odeh, "Multidimensional Spherical Harmonics Expansion of Boltzmann Equation for Transport in Semiconductors," *Appl. Math. Lett.,* vol. 5, n. 4, p. 85, 1992

[3] K.A. Hennacy, Y.-j. Wu, N. Goldsman and I.D. Mayergoyz, "Deterministic MOSFET Simulation Using a Generalized Spherical Harmonic Expansion of the Boltzmann Equation," *Solid-State Electronics,* vol. 38, p. 1485, 1995.

[4] A. Das and M.S. Lundstrom, "A Scattering Matrix Approach to Device Analysis," *Solid-State Electron.,* vol. 33, p. 1299, 1990.

[5] K. Kometer, G. Zandler and P. Vogl, "Lattice-Gas Cellular-Automaton Method for Semiclassical Transport in Semiconductors," Phys. Rev. B, vol. 46, p. 1382, 1992.

[6] R.W. Hockney and J.W. Eastwood, *Computer Simulation Using Particles,* New York: Mc Graw Hill, 1981.

[7] K. Hess, ed., *Monte Carlo Device Simulation: Full Band and Beyond,* Kluwer Academic Publishers, Norwell, Massachusetts, 1991.

[8] H. Shichijo and K. Hess, "Band-Structure-Dependent Transport and Impact Ionization in GaAs," *Phys. Rev. B.,* vol. 23, p. 4197, 1981.

[9] M.V. Fischetti and S.E. Laux, "Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices Including Band-Structure and Space-Charge Effects," *Phys. Rev. B,* vol. 38, p.9721, 1988.

[10] M.L. Cohen and J.R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors,* Springer Series in Solid-State Sciences, vol. 75, Springer-Verlag, Berlin, 1988.

[11] M.V. Fischetti and J.M. Higman, Chapter 5 in *Monte Carlo Device Simulation: Full Band and Beyond,* K. Hess, ed., Kluwer Academic Publishers, Norwell, Massachusetts, 1991.

[12] T. Kunikiyo, M. Takenaka, Y. Kamakura, M, Yamaji, H. Mizuno, M. Morifuji, K. Taniguchi and C. Hamaguchi, "A Monte Carlo Simulation of Anisotropic Electron Transport in Silicon Including Full Band Structure and Anisotropic Impact Ionization Model," *J. of Appl. Physics,* vol. 75, p. 297, 1994.

[13] P.D. Yoder, V.D. Natoli and R.M. Martin, "*Ab Initio* Analysis of the Electron-Phonon Interaction in Silicon," *J. of Appl. Physics,* vol. 73, p. 4378, 1993.

[14] C. Wordelman, *Valley-Dependent Monte Carlo Simulation,* M.S. Thesis, University of Illinois at Urbana-Champaign, 1995.

[15] C.H. Lee, U. Ravaioli, K. Hess, C. Mead and P. Hasler, "Simulation of a Long Term Memory Device with a Full Bandstructure Monte Carlo Approach," *IEEE Electron Device Lett.,* vol. 16, p. 360, 1995.

Biographies

Umberto Ravaioli is a Professor in the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign. His main research interests are in the areas of semiconductor device physics and simulation, numerical methods, and high performance computing.

Amanda Duncan is a Ph.D. student in the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign. Her thesis work is in the area of full band Monte Carlo simulation of ultra small silicon devices.

Andrea Pacelli was a visiting scientist with the Beckman Institute of the University of Illinois at Urbana-Champaign, and he is now a Ph.D. student at the Politecnico di Milano, Italy, working on physics and simulation of silicon devices.

Carl Wordelman is a Ph.D. student in the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign. His research work is in the area of full band Monte Carlo simulation of transport in silicon.

Karl Hess is a Professor in the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign. His research interests are in the areas of high energy transport in semiconductors, laser simulation, and quantum nanostructures.

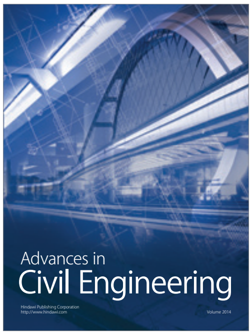
Journal of
Engineering

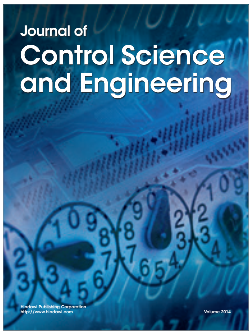
The Scientific
World Journal

International Journal of
Rotating
Machinery

Journal of
Sensors

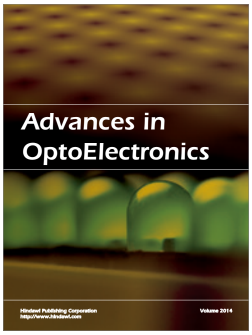
International Journal of
Distributed
Sensor Networks


Advances in
Civil Engineering


Journal of
Control Science
and Engineering

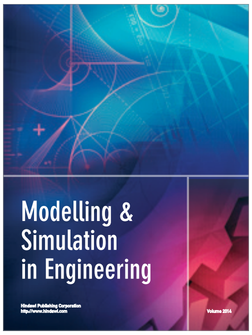
Journal of
Robotics

Journal of
Electrical and Computer
Engineering

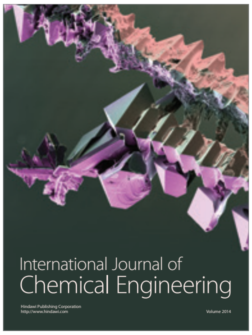

Advances in
OptoElectronics

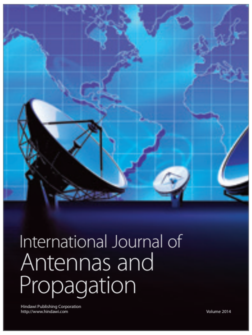
VLSI Design

International Journal of
Navigation and
Observation

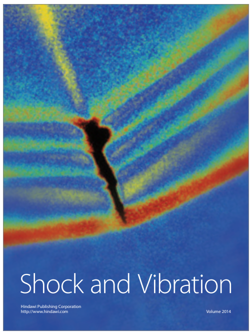

Modelling &
Simulation
in Engineering

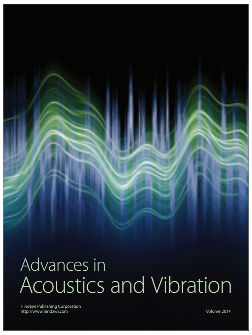
International Journal of
Aerospace
Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and
Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration