Full Band Monte Carlo Simulation

Umberto Ravaioli

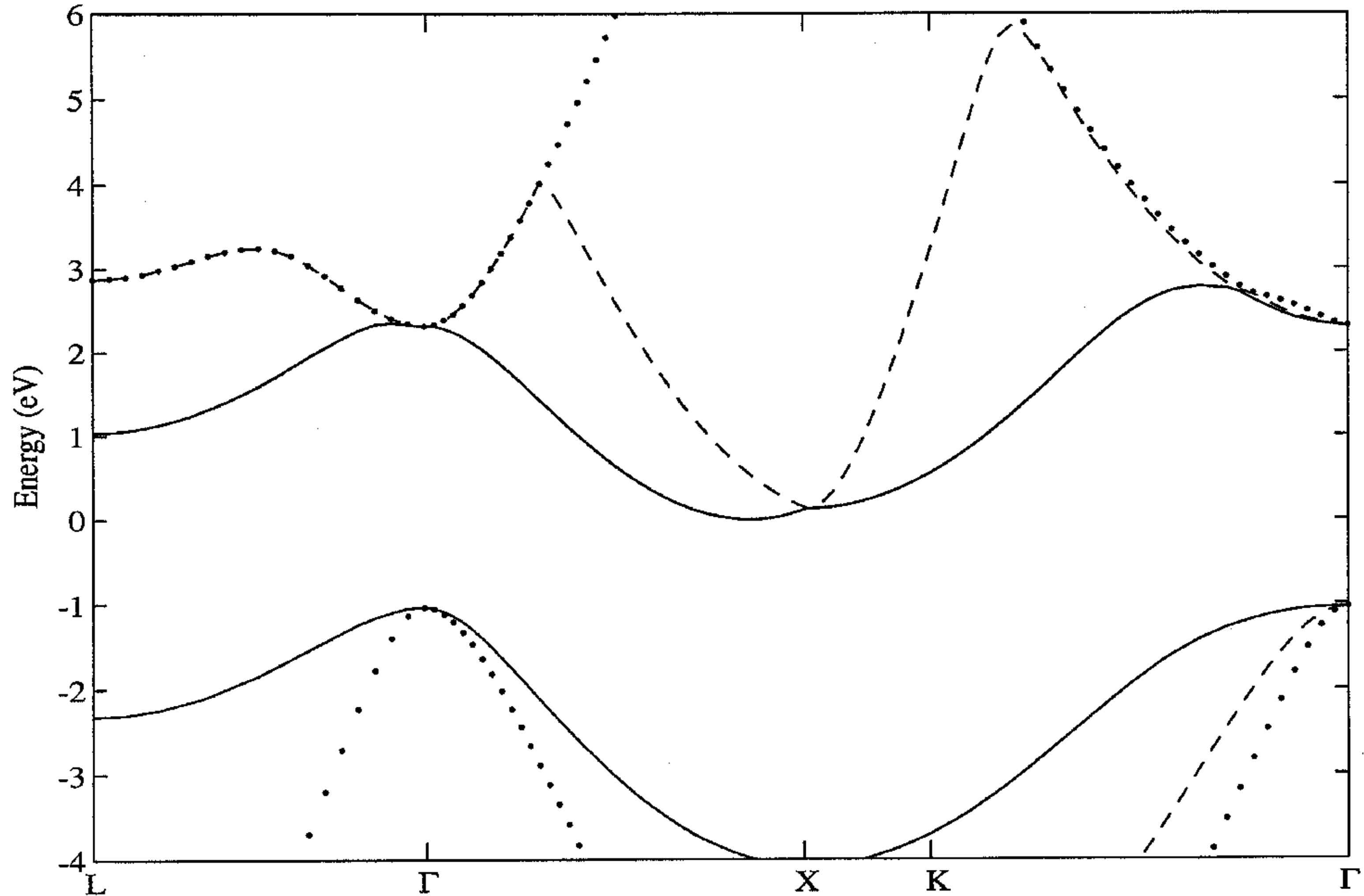
Beckman Institute and
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Issues

- **The conduction (valence) band is represented by a parabolic dispersion relation, only close to the energy minima.**
- **Non-parabolicity may extend analytical representations of the bands but may also have limited validity. In silicon, for instance, typically above 1.0 eV, the density of states in the conduction band may not be approximated by a non-parabolic dispersion relation.**
- **Valence bands may have strong warping, which is difficult to represent analytically.**
- **At higher energies there may be satellite minima and other band braches start to appear.**

Silicon band structure

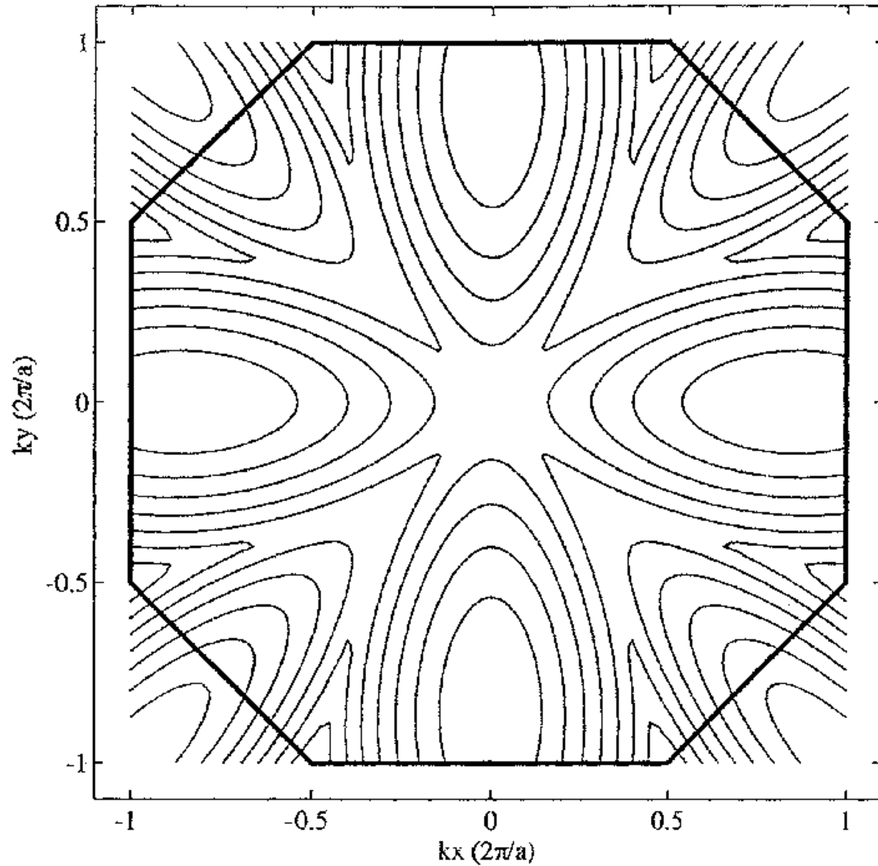
- Textbook picture of band structure



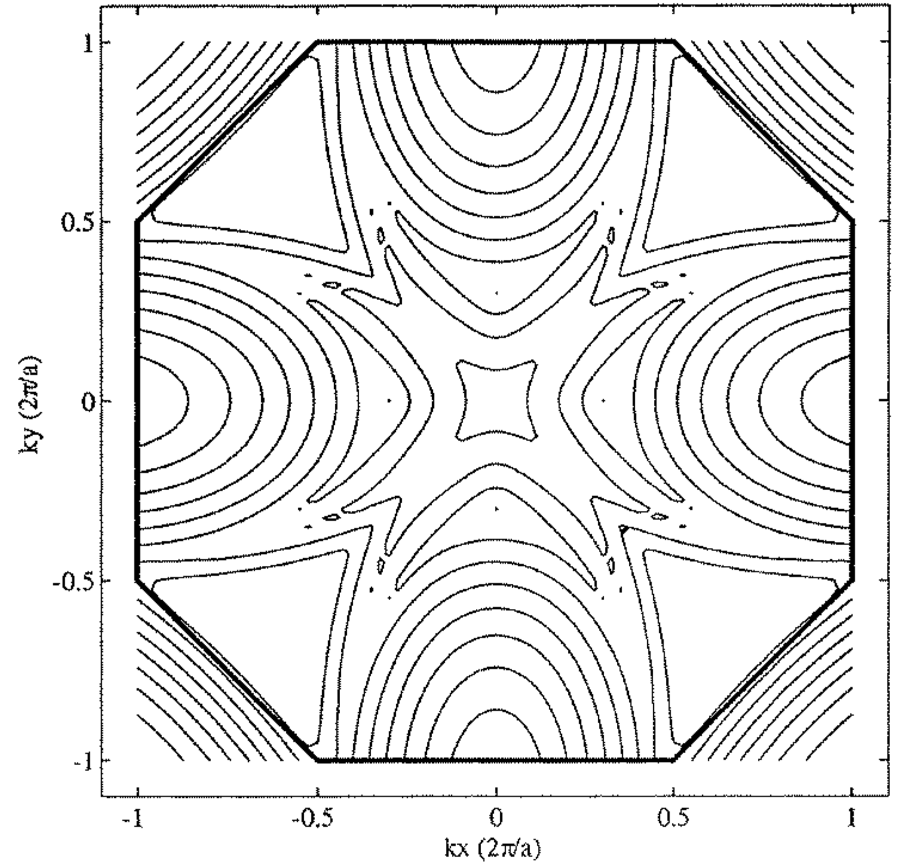
Silicon band structure

- **Contour plots**

First band

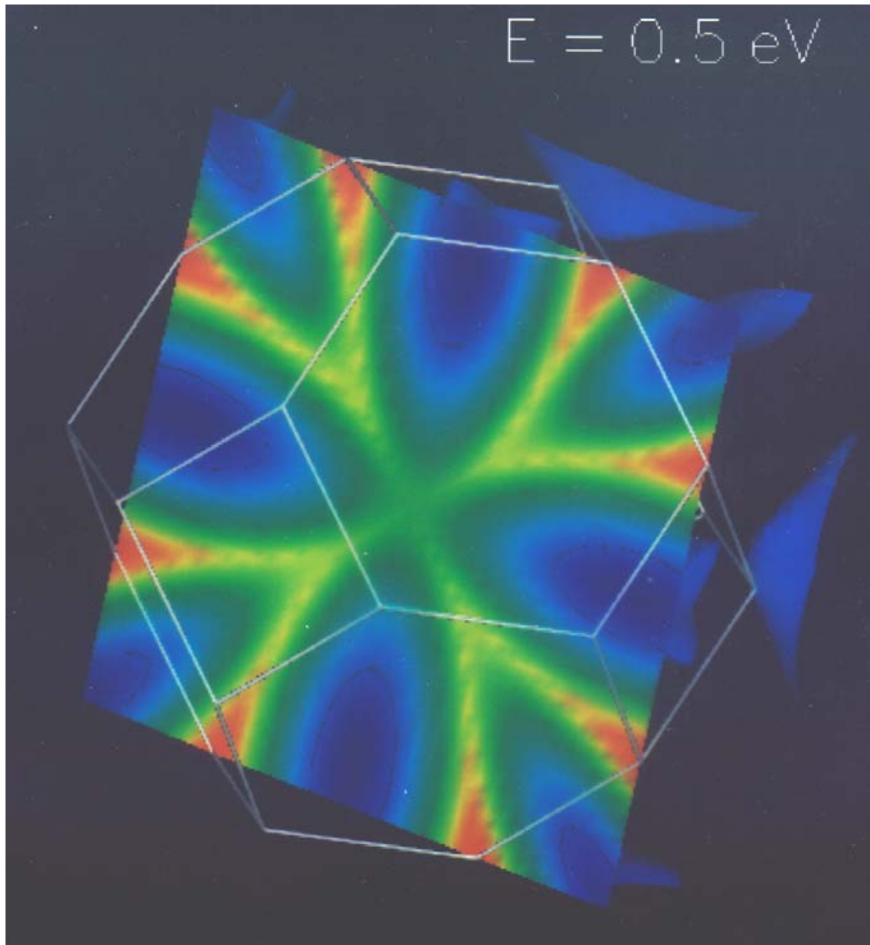


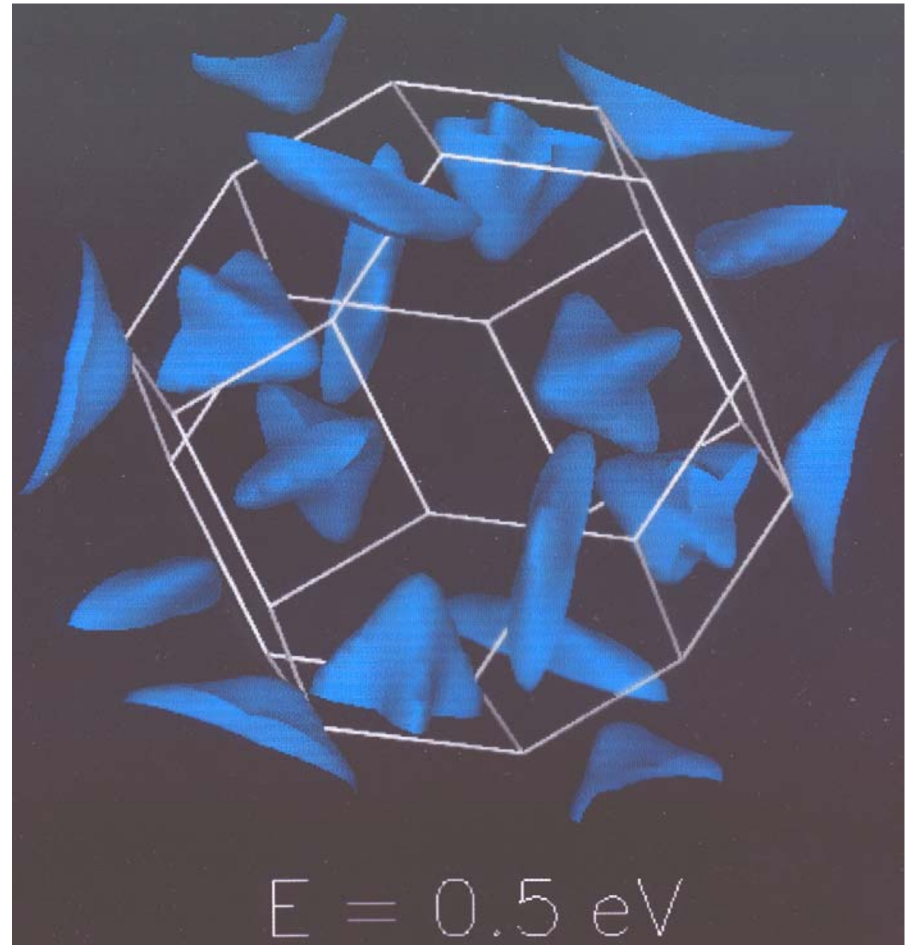
Second band



Silicon band structure

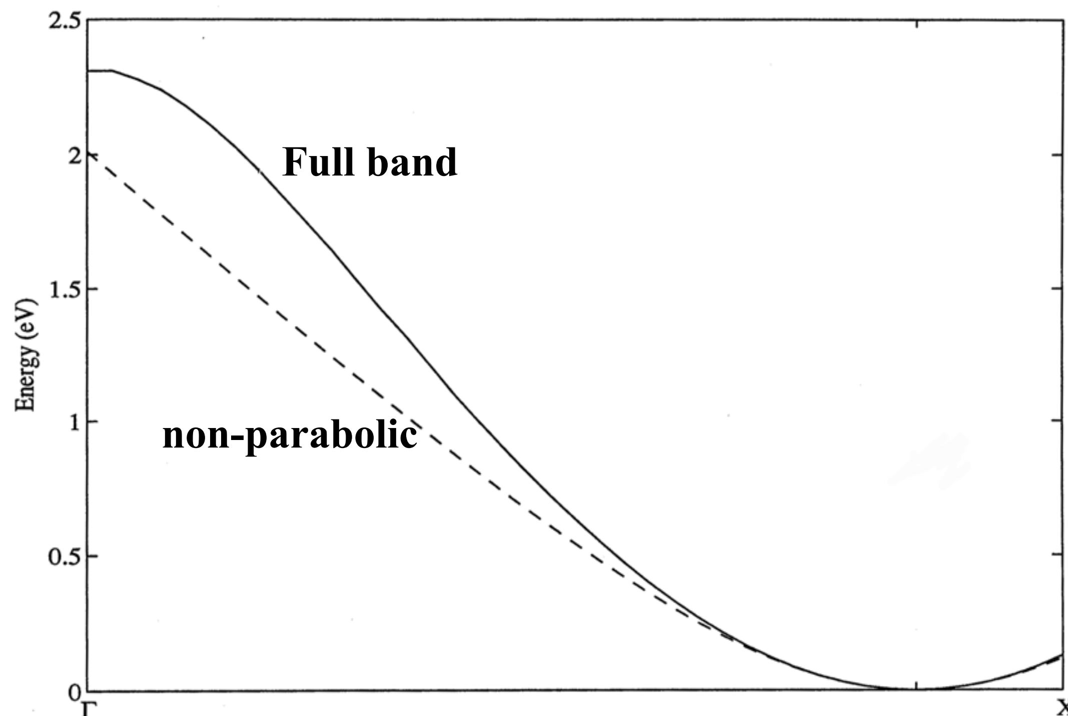
- **3D isosurfaces - Video animation**



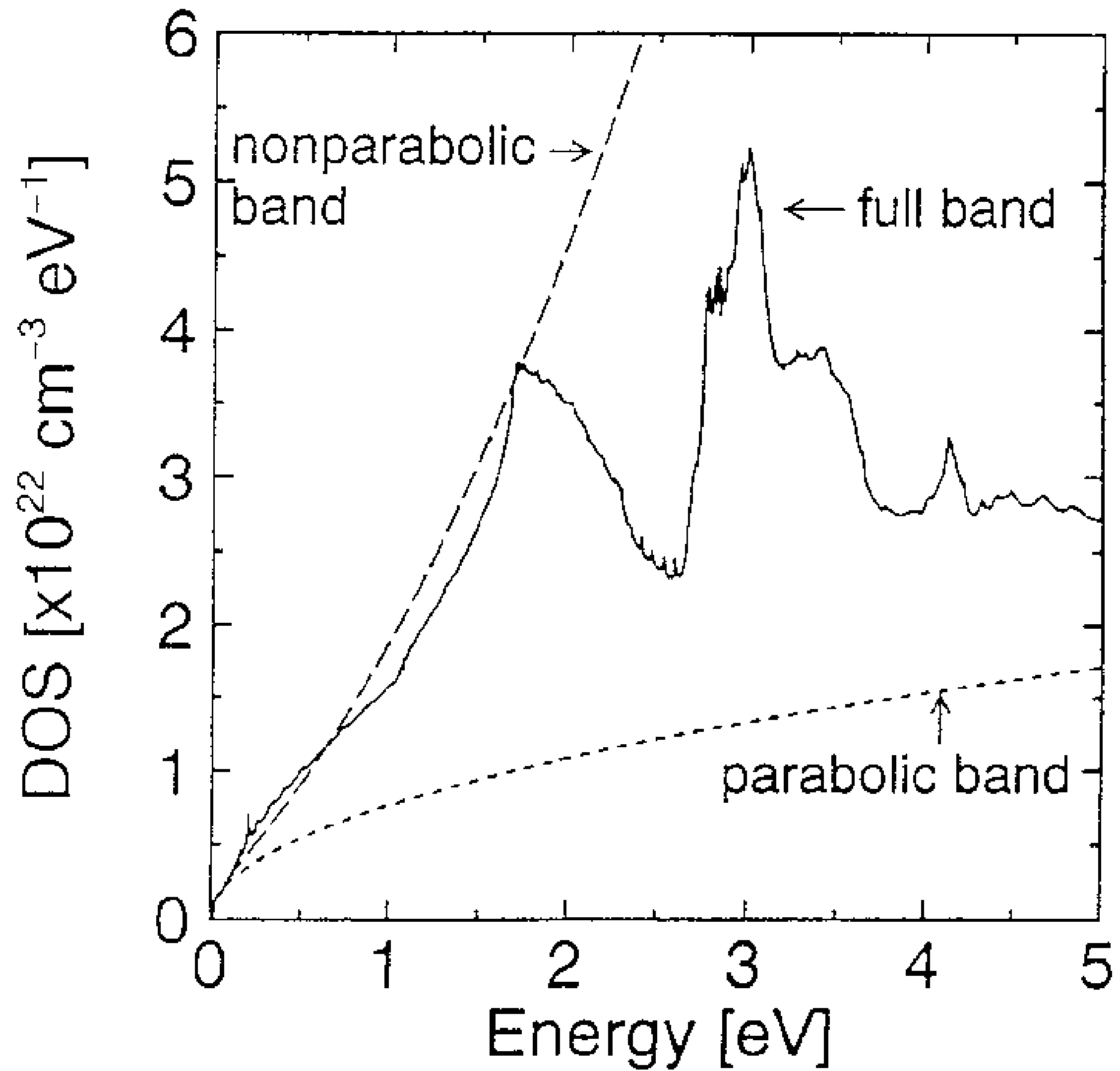


Silicon band structure

- **First conduction band near minimum**



Silicon density of states



Band structure calculations

- The band structure of semiconductors may be obtained by using a variety of methods, known as tight-binding, $k \cdot p$, local and non-local pseudopotentials, etc.

- For semiconductor bandstructure, we find it useful to adopt pseudopotential methods. We review in the following the steps to solve a bandstructure problem with the local empirical pseudopotential approach.

- The pseudopotential method assume a "reasonable form" of the potential due to the atom from nucleus up to core electrons. The states for valence electrons are solved, and as outcome of the program we obtain the eigenvalues corresponding to valence and conduction bands.

Empirical pseudopotentials

- Our goal is to solve the Schrödinger equation

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi(\vec{r})+V(\vec{r})\psi(\vec{r})=E\psi(\vec{r})$$

to obtain the allowed values of energy E for a given value of k for the valence electrons.

- From Block's theorem

$$\psi\left(\vec{k},\vec{r}\right)=e^{\hat{j}\vec{k}\cdot\vec{r}}u_k\left(\vec{r}\right)$$

$$V(\vec{r}) \quad \text{and} \quad u_k\left(\vec{r}\right) \quad \text{are periodic with the lattice}$$

- **We can expand in terms of reciprocal lattice vectors and substitute in Schrödinger equation**

$$u_k(\vec{r}) = \sum_h A_{\vec{K}_h} e^{\hat{j}\vec{K}_h \cdot \vec{r}}$$

$$V(\vec{r}) = \sum_m V_{\vec{K}_m} e^{\hat{j}\vec{K}_m \cdot \vec{r}}$$

$$\Rightarrow \frac{\hbar^2}{2m_0} \sum_h \left| \vec{k} + \vec{K}_h \right|^2 A_{\vec{K}_h} e^{\hat{j}(\vec{k}+\vec{K}_h)\cdot\vec{r}}$$

$$+ \sum_m \sum_h V_{\vec{K}_m} e^{\hat{j}(\vec{k}+\vec{K}_h+\vec{K}_m)\cdot\vec{r}} = E(\vec{k}) \sum_h A_{\vec{K}_h} e^{\hat{j}(\vec{k}+\vec{K}_h)\cdot\vec{r}}$$

- **Now multiply by**

$$e^{-\hat{j}(\vec{k}+\vec{K}_l)\cdot\vec{r}} \qquad l = 0, 1, \ldots, n$$

and integrate over the volume of the crystal (Fourier transform), with result

$$\frac{\hbar^2}{2m_0} A_{\vec{K}_h} \left(\vec{k}+\vec{K}_l\right)^2 + \sum_h V_{\vec{K}_l-\vec{K}_h} A_{\vec{K}_h} = E(\vec{k}) A_{\vec{K}_l}$$

- **In matrix form**

$$\begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & \cdots & a_{0,n} \\ a_{1,0} & a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,0} & a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ a_{n,0} & a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} A_{\vec{K}_0} \\ A_{\vec{K}_1} \\ A_{\vec{K}_2} \\ \vdots \\ A_{\vec{K}_n} \end{bmatrix} = E\left(\vec{K}\right) \begin{bmatrix} A_{\vec{K}_0} \\ A_{\vec{K}_1} \\ A_{\vec{K}_2} \\ \vdots \\ A_{\vec{K}_n} \end{bmatrix}$$

$$a_{ij} = \begin{cases} \dfrac{\hbar^2}{2m_0} \left|\vec{k} + \vec{K}_i\right|^2 & i = j \\ V_{\vec{K}_i - \vec{K}_j} & i \neq j \end{cases}$$

- **For a given wave vector k, we can determine the band structure values E(k) by solving this eigenvalue solver.**

- **Technical issue: how big should n be? Band structure experts indicate that $n \approx 120$ should give satisfactory results.**
- **The coefficients expressing the potential can be decomposed into a symmetric and an antisymmetric part**

$$V(\vec{r}) = \sum_m V_{\vec{K}_m} e^{\hat{j}\vec{K}_m \cdot \vec{r}}$$

$$V_{\vec{K}_m} = \overbrace{\underbrace{S^s(\vec{K}_m)}_{\text{structure factor}} \underbrace{F^s_{\vec{K}_m}}_{\text{pseudopotential form factor}}}^{\text{symmetric}} + \hat{j}\,\overbrace{\underbrace{S^a(\vec{K}_m)}_{\text{structure factor}} \underbrace{F^a_{\vec{K}_m}}_{\text{pseudopotential form factor}}}^{\text{antisymmetric}}$$

The diamond and zinc-blende type semiconductor have fcc structure with 2 atoms per cell. The origin is taken halfway between these two atoms with position

$$\vec{r}_1 = (1/8, 1/8, 1/8) \;\; ; \qquad \vec{r}_2 = (-1/8, -1/8, -1/8)$$

and

$$S^s\left(\vec{K}_m\right) = \cos\left(\vec{K}_m \cdot \tau\right) \;\; ; \quad S^a\left(\vec{K}_m\right) = \sin\left(\vec{K}_m \cdot \tau\right)$$

$$F^s_{\vec{K}_m} = \frac{V_1\left(\vec{K}_m\right) + V_2\left(\vec{K}_m\right)}{2} \;\; ; \quad F^a_{\vec{K}_m} = \frac{V_1\left(\vec{K}_m\right) - V_2\left(\vec{K}_m\right)}{2}$$

$$V_1\left(\vec{K}_m\right) = \frac{2}{\Omega}\int V_1(\vec{r})\, e^{-\hat{j}\vec{K}_m \cdot \vec{r}}\, d^3\vec{r}$$

$$V_2\left(\vec{K}_m\right) = \frac{2}{\Omega}\int V_2(\vec{r})\, e^{-\hat{j}\vec{K}_m \cdot \vec{r}}\, d^3\vec{r}$$

$V_1(\vec{r})$ and $V_2(\vec{r})$ are the pseudopotentials due to single atoms in the lattice and Ω is the volume of the unit cell.

For diamond structure

$$V^s_{\vec{K}_m} = V_1(\vec{K}_m) = V_2(\vec{K}_m)$$

$$V^a_{\vec{K}_m} = 0$$

The first five of the reciprocal lattice vectors for fcc structure have squared magnitude of 0,3,4,8,11 in the units of $(2\pi/a)^2$.

Only these vectors are allowed to have a non-zero potential.

Form factors are taken in a local and static approximation, to be independent of momentum and energy, so that they can produce average potentials for the whole Brillouin zone and the entire range of energy of the valence and conduction bands.

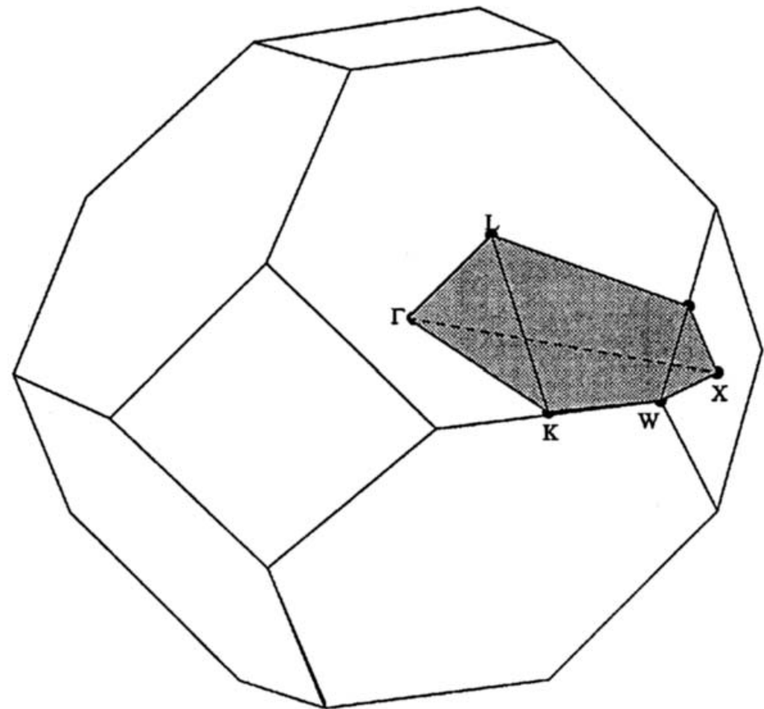
Below is a table of form factors for Si and Ga. The local pseudopotential approximation is better for Si. For a polar material like GaAs, anti-symmetric form factors are to be considered, leading to a complex eigenvalue problem. Some details of the bandstructure are more accurately reproduced if a more complicated non-local approach is taken for GaAs.

Element	$F^s_{\sqrt{3}}$	$F^s_{\sqrt{4}}$	$F^s_{\sqrt{8}}$	$F^s_{\sqrt{11}}$	$F^a_{\sqrt{3}}$	$F^a_{\sqrt{4}}$	$F^a_{\sqrt{8}}$	$F^a_{\sqrt{11}}$	a
Si	–0.224	0.0	0.055	0.072	0.0	0.0	0.0	0.0	5.431
GaAs	–0.23	0.0	0.01	0.06	0.07	0.12	0.07	0.02	5.65

The computational cost of band structure calculations can be significantly reduced if the symmetry of the Brillouin zone is exploited. Only 1/48 (irreducible wedge) of the fcc Brillouin zone is necessary. The irreducible wedge is determined by the following equations (wave vectors normalized with 2π/a)

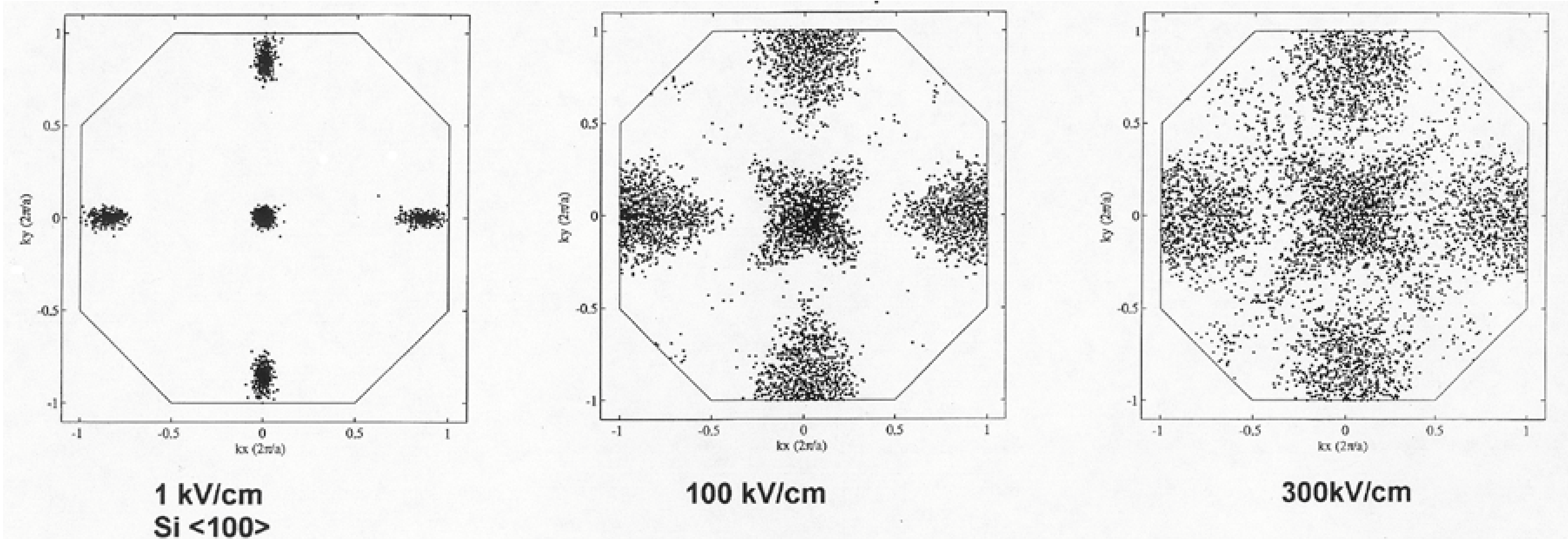
$$0 \le k_z^* \le k_x^* \le k_y^* \le 1$$

$$k_x^* + k_y^* + k_z^* \le \frac{3}{2}$$



Bulk Monte Carlo simulation of Silicon

Effect of band structure on high-field transport



Full band Monte Carlo implementations

Complete model:

k-dependent scattering rates
- Requires storage of very large scattering rate tables
- Take fully into account material anisotropy
- Suitable for material parameter calculations

Simplified model:

Energy-dependent scattering rates
- Same storage requirements as conventional MC
- Neglects anisotropy of scattering
- Suitable for device simulation

Two main objections can be raised against full band structure approaches:

(i) Details of the band structure at high energy are not well known, therefore the full band approach may be unreliable.

There has been great progress in band structure calculations using pseudopotential and density functional approaches. Not only the accuracy of the band structure has improved, but it is also possible to develop a unified theory which treats band structure and scatterings consistently, minimizing the number of input parameters.

(ii) Full band calculations require extremely large computational resources. CPU run times are extremely long, making the collection of results too slow to be of practical use.

The power of supercomputers, workstations and PCs has been increasing tremendously. Self-consistent calculations for realistic structures may still require hours or even days on the most powerful machines, but particularly for 2-D simulation the turn-around time is acceptable even on inexpensive computing platforms.

Monte Carlo based on energy-dependent scattering rates is the approach of choice for practical device simulation.

While scattering rate calculations may be very time consuming if done using first principle approaches from the band structure parameters, in a simplified approach rates may be obtained with the standard non-parabolic band approach. A scaling based on the knowledge of the full band density of state is applied to get the proper behavior. This procedure is acceptable for most device simulation applications and and it adds negligible cost with respect to the standard Monte Carlo approach.

It is then possible to update scattering table on-the-fly if one wants to change temperature, material composition, etc.

Specific algorithms for full band Monte Carlo

Particle trajectories

The Brilluoin zone is discretized, and trajectories are tracked in momentum space to obtain the appropriate wave group velocity, which in turn is used for the real space trajectories.

Some approaches apply a tetrahedra discretization to momentum space, so that one can use linear functions to determine directly the evolution in momentum space.

If a simple cubic discretization of momemtum space is used, then an interpolation scheme is applied.

Specific algorithms for full band Monte Carlo

Final state determination after scattering

Final state must be consistent with the band structure description. The band structure table is sorted by energy, so that, given the energy at the end of the scattering, a new state can be selected from the table.

Band structure

k_x	k_y	k_z	E	band index
0	0	0	0.0	1
1	0	0	0.01	1
⋮	⋮	⋮	⋮	⋮

Sorted list

E	band index	k_x	k_y	k_z
0.0	1	0	0	0
0.01	1	1	0	0
⋮	⋮	⋮	⋮	⋮

Energy broadening may be applied, by selecting a certain broadening range about the final energy, and a state in that range is selected randomly as the final state after scattering, to restart the new flight .

Recent code developments

3-D Parallel Monte Carlo Simulation of sub-0.1 Micron MOSFETs on a Cluster Based Supercomputer

Asim Kepkep and Umberto Ravaioli

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

Recent code developments

3-D Parallel Monte Carlo Simulation of sub-0.1 Micron MOSFETs on a Cluster Based Supercomputer

Asim Kepkep and Umberto Ravaioli

- **Performance issues**
- **Charge-charge interaction**
- **Code validation - Well tempered MOSFET tests**

3-D Code Development

- **MOCA 3-D version for MOSFET has been completed.**
- **Detailed charge-charge interaction implemented coupling to Poisson equation (long range) a local molecular dynamics (short range) using the particle-particle-particle-mesh (P3M) method.**
- **Test calculations and validation performed for bulk mobility (ionized impurity scattering with granular doping charge-charge interaction) and well-tempered MOSFET.**

3-D Parallel Monte Carlo

- **Parallel (MPI) 3-D version addresses computation on modern clusters.**
- **Parallelization of the Monte Carlo portion is accomplished by using an optimal domain decomposition in space for MOSFET structures.**
- **3-D Poisson equation is based on Conjugate Gradients. Matrix operations are parallelized.**
- **Production runs conducted on several NCSA clusters.**

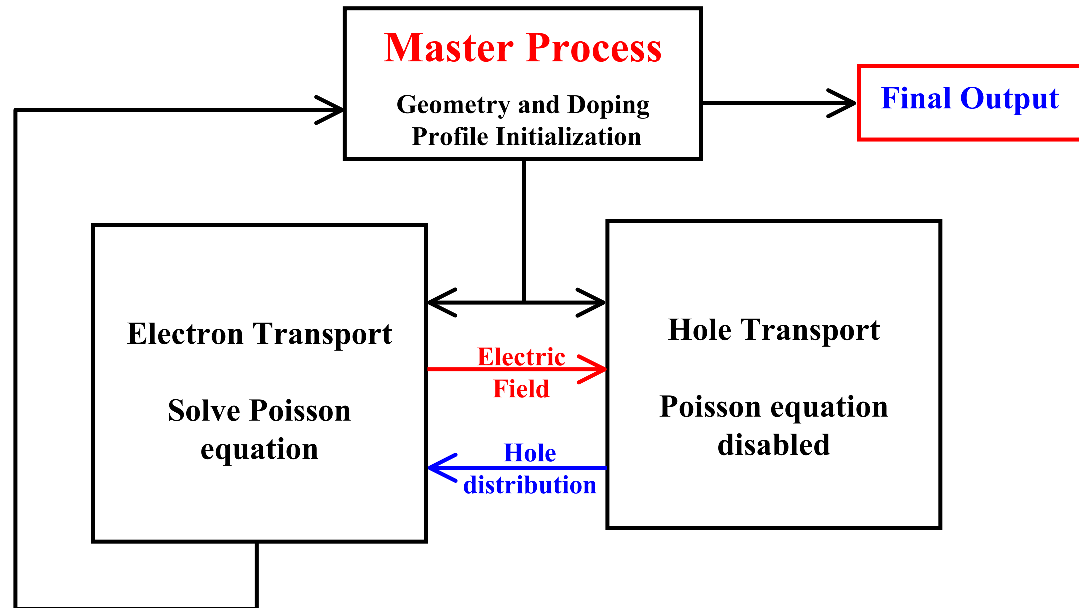
Parallel Monte Carlo implementation

- **Developed using MPI, mainly in FORTRAN and C enabling, future migration to other parallel platforms.**
- **The transport core has been kept meshless to maintain flexibility.**
- **Each "node instance" of the code handles particles in a 3-D domain with rectangular cross-section.**
- **To minimize node to node interaction, structural boundaries of the device are duplicated on all processors, allowing particles to be processed locally for the entire time-step.**

Sources of Parallelism

- **Dynamics of distant particles are independent of each other during sub-flights (lattice scattering, boundary crossings, etc.). Particles interact with each other only during force evaluation steps.**
- **Particles within a sub-domain may be viewed as 'ensembles within the ensemble' giving another mode of parallelism. (This mode is not utilized in this work)**
- **In bipolar mode, electrons and holes interact with each other only through forces obtained from the solution of Poisson's equation therefore a two-fold parallelism exists between force evaluation steps.**

Sources of Parallelism



Two separate "synchronized" executables implement electron and hole transport in our implementation.

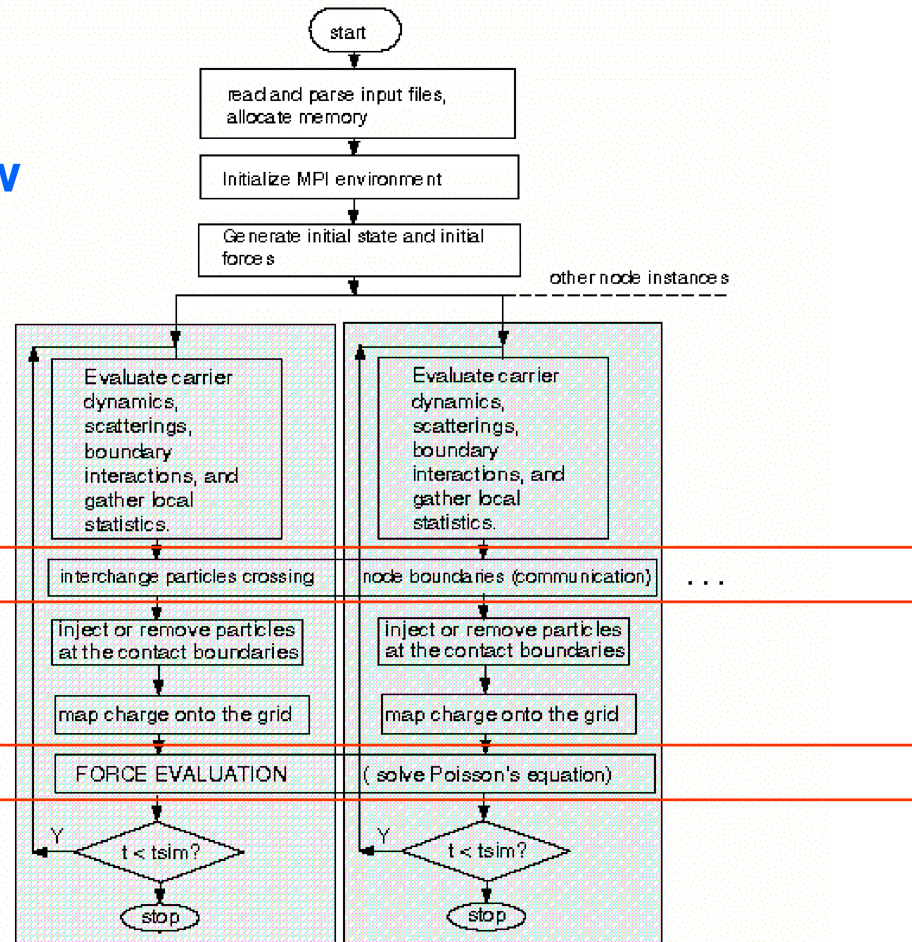
Parallel Monte Carlo implementation

- **Developed using MPI, mainly in FORTRAN and C enabling, future migration to other parallel platforms.**
- **The transport core has been kept meshless to maintain flexibility.**
- **Each “node instance” of the code handles particles in a 3-D domain with rectangular cross-section.**
- **To minimize node to node interaction, structural boundaries of the device are duplicated on all processors, allowing particles to be processed locally for the entire time-step.**

3-D Parallel Monte Carlo

Program flow

Intensive data communication



Computational Environment at NCSA

- **The systems are evolving in terms of processor, system software and operating systems.**
- **Initial work on NT cluster, with 96 double Pentium III nodes.**
- **New higher performance cluster with high-performance Intel chip and Linux operating system has been available over the last year. NT cluster converted to Linux over time. Security issues in Linux are being improved, but still of some concern for a public access facility.**
- **New system announced, with Intel Itanium chips, part of a national distributed facility.**
- **Issue: Faster and higher performance system may be available *before* efficient compilers are.**

Charge-charge interaction

Application of the particle-particle-particle-mesh approach (P^{3}M)

- The total force acting on a generic particle is

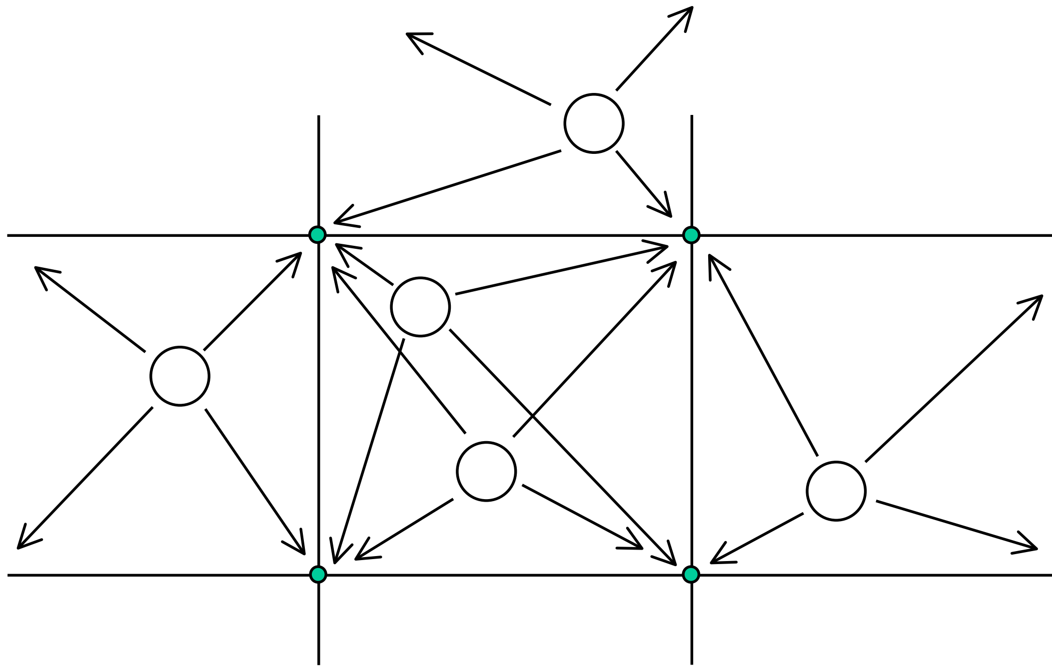
$$F_i = \sum_{j \neq i} F_{ij}^{coul} + F_i^{ext}$$

$$F_{ij}^{coul} = \frac{q_i q_j (\vec{r}_i - \vec{r}_j)}{4\pi\varepsilon \left|\vec{r}_i - \vec{r}_j\right|^3}$$

$$F_i^{ext} = \text{ due to bias}$$

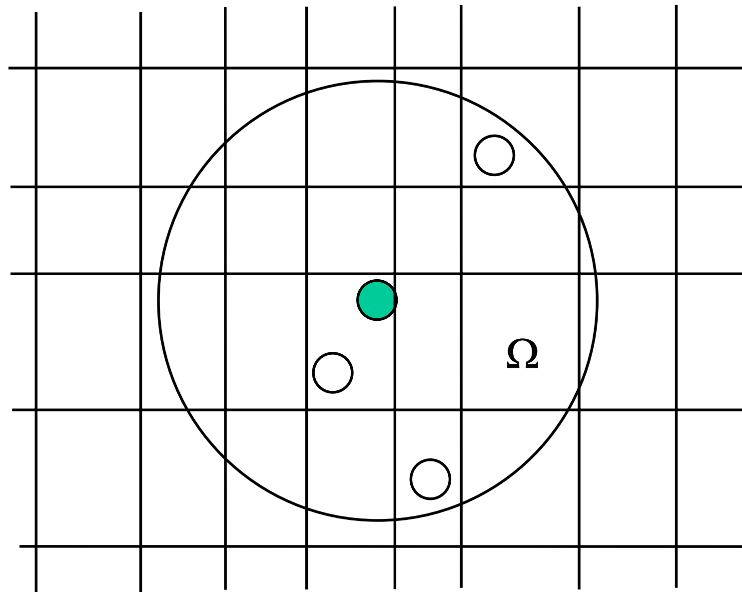
Charge-charge interaction

- **The particle charge is mapped to a grid where Poisson equation is solved. The "long-range" (plasmon) is well resolved, the "short-range" (e-e scattering) not completely due to the finite size of the mesh.**



Charge-charge interaction

- **To resolve completely the short-range, we have adopted a technique that adapts well to the non-uniform rectangular grid used in MOCA. The full coulomb forces are evaluated in a neighborhood Ω of the particle**



Charge-charge interaction

- **But Poisson equation also solves on mesh points inside the short-range domain, so there is an overlap and a double-counting. A correction (reference force is evaluated to remove the double-counting**

$$F_i = \sum_{\substack{i \neq j \\ \text{in } \Omega}} F_{ij}^{sr} + \underbrace{\sum_{\substack{i \neq j \\ \text{outside } \Omega}} F_{ij}^{mesh}}_{\text{from Poisson equation}}$$

$$F_{ij}^{sr} = F_{ij}^{coul} - \underbrace{R_{ij}}_{\text{reference force}}$$

- **For details on the reference force calculation, a good source is the book by Hockney and Eastwood, Computer Simulation Using Particles, 1981.**

Charge-charge interaction

- **The reference force is neede to avoid double counting of short-range force due to overlap of particle-mesh and short-range domains.**
- **It could be calculated by associating particle j with the particle-mesh and calculating the resulting force on particle i with external forces equal to zero. This procedure is costly.**
- **A possible approximation is to associate a finite size to particle i to achieve a smoothing of the total interparticle force between long- and short-range domains. We adopted the choice indicated by Hockney, of a radial uniformly decreasing density profile**

$$S(r) = \begin{cases} \dfrac{48}{\pi r_{sr}^4}\left(r_{sr}/2 - r\right) & r \le r_{sr}/2 \\ 0 & r > r_{sr}/2 \end{cases}$$

Charge-charge interaction

- **The reference force is then defined as**

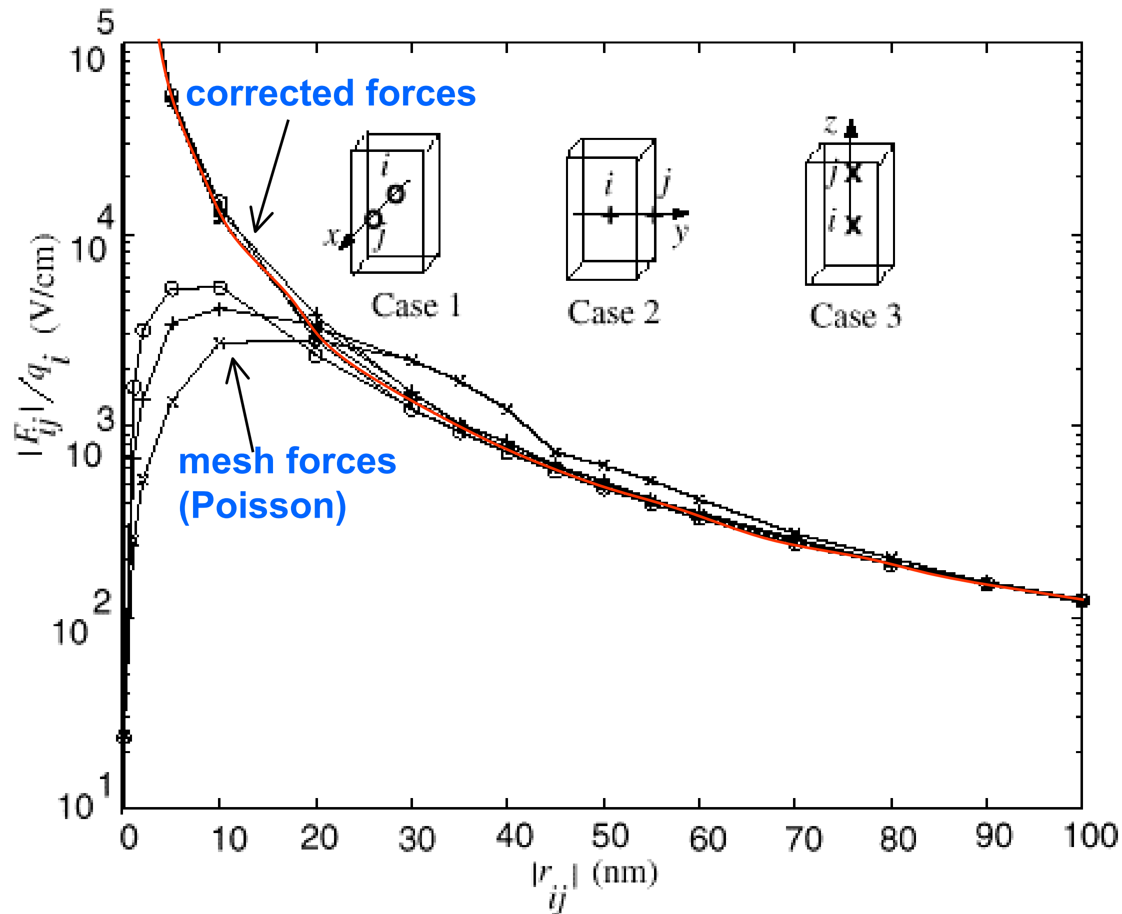
$$\vec{R}(r) = \frac{\hat{r}}{4\pi\varepsilon}\int d\,\vec{x}'\int d\,\vec{x}''\; S(x'')S(x'-x'')\frac{(\vec{x}'-\vec{x}'')}{|\vec{x}'-\vec{x}''|^3}$$

- **Using the previous definition of** $S(r)$ **and** $\xi = 2r/r_{sr}$

$$R_{ij} = \frac{q_i q_j}{4\pi\varepsilon}\begin{cases} \frac{1}{35 r_{sr}^2}\left(224\xi - 224\xi^3 + 70\xi^4 + 48\xi^5 - 21\xi^6\right) & 0 \le r \le r_{sr}/2 \\ \frac{1}{35 r_{sr}^2}\left(\frac{12}{\xi^2} - 224 + 896\xi - 840\xi^2 + 224\xi^3 + 70\xi^4 - 48\xi^5 + 7\xi^6\right) & r_{sr}/2 \le r \le r_{sr} \\ \frac{1}{r^2} & r > r_{sr} \end{cases}$$

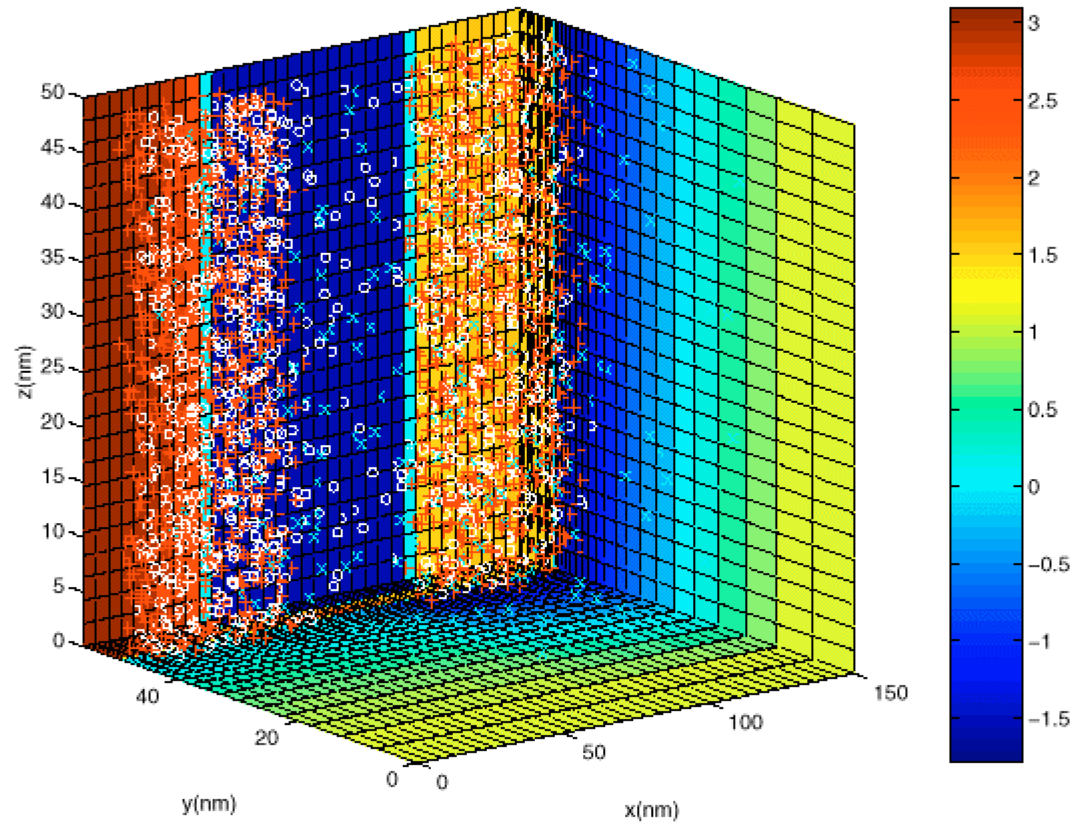
Charge-charge interaction

The test of the P^3M scheme is done by considering two particles in a mesh and the exact Coulomb law (in red).



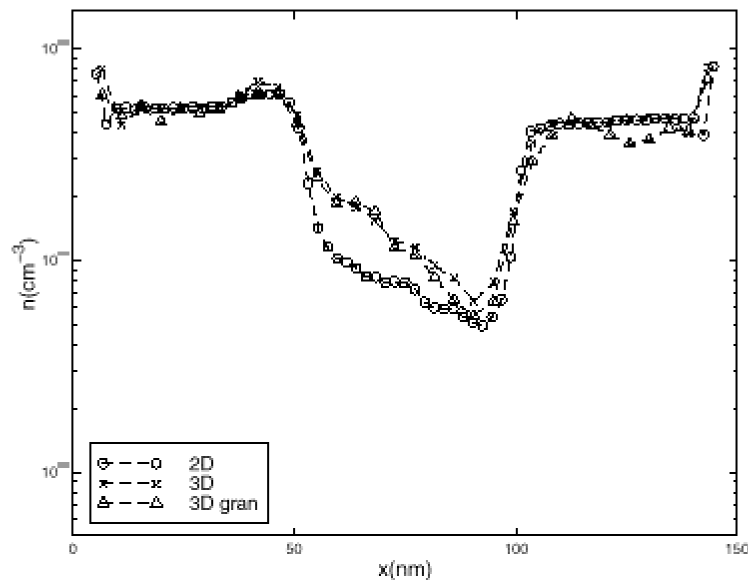
Charge-charge interaction

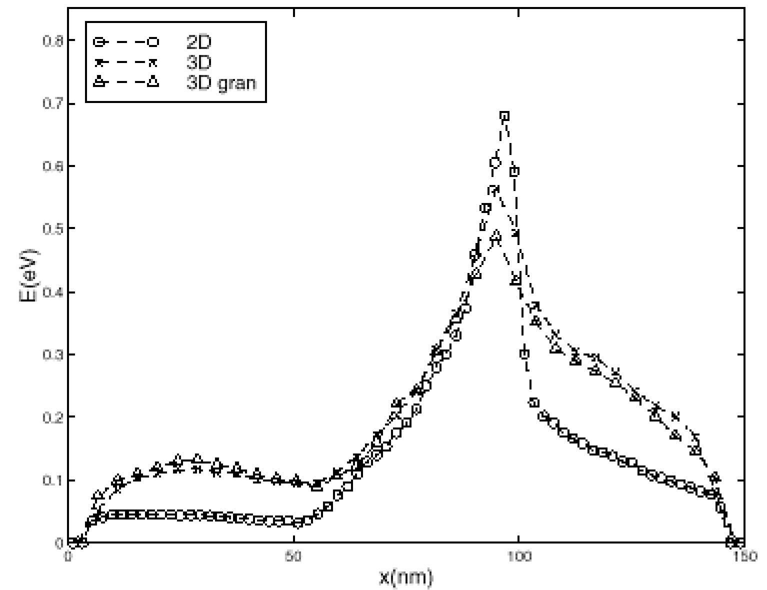
- **An example of MOSFET implementation (50nm gate) (granular dopant distribution)**



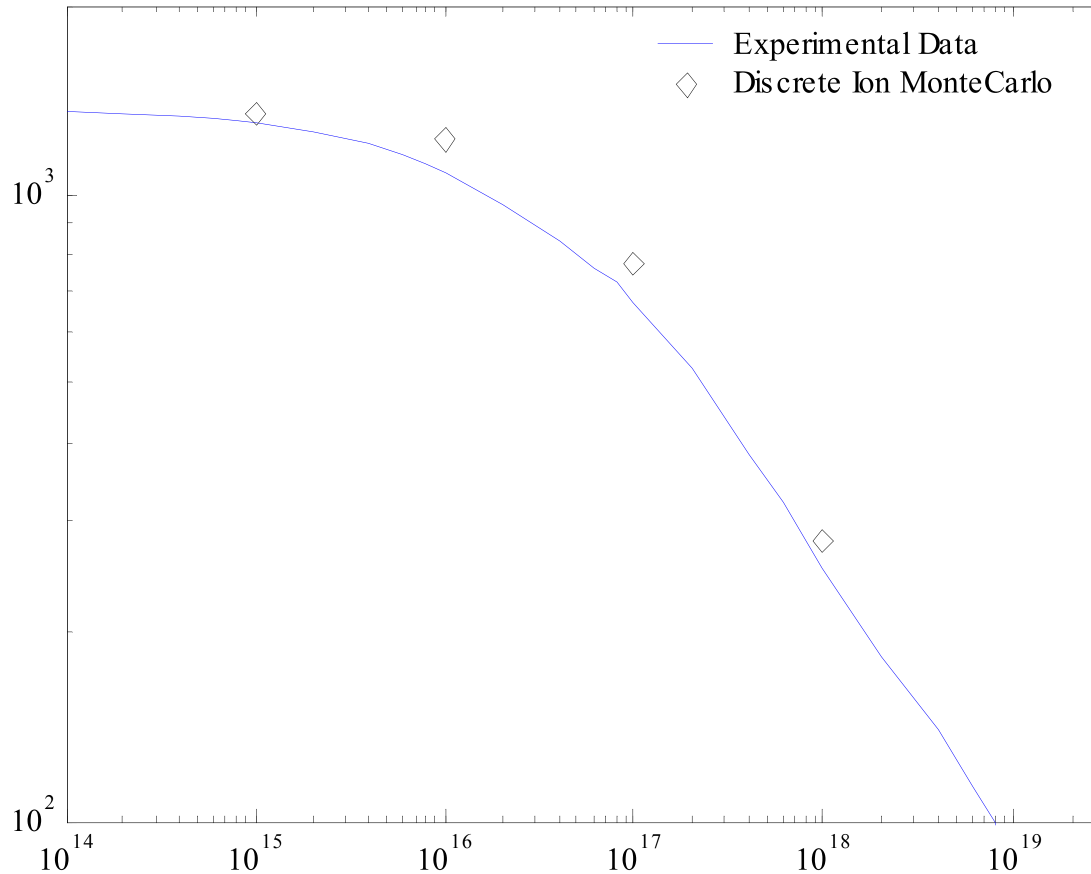
Charge-charge interaction

- **An example of n-MOSFET implementation (50nm gate) ($V_G = V_D = 1.5$ V)**



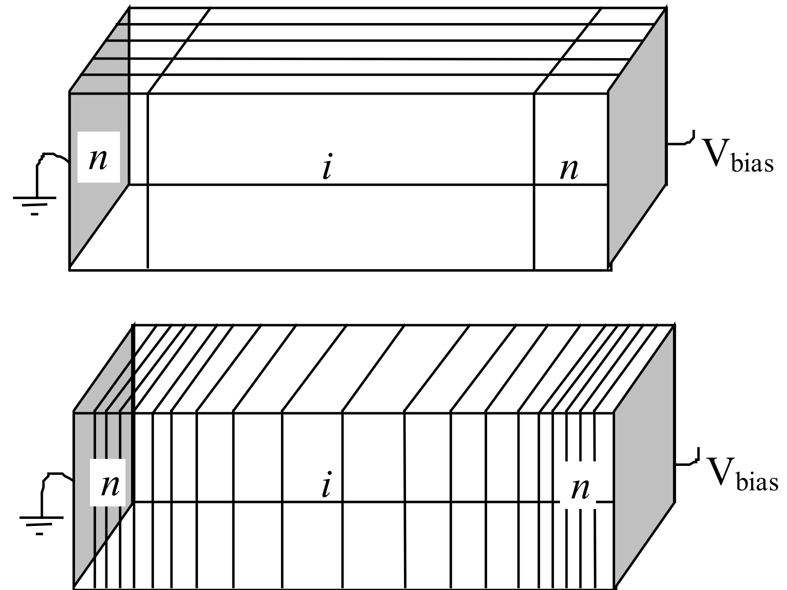


Code Validation: Electron mobility as a function of doping concentration



Prototype test structure

- An n-i-n structure have been simulated. Length of the slab is 0.5 μm.
- A velocity-field curve for electrons produced to verify functionality.
- Two partitioning schemes are compared to see impact on communication costs.
- The grid is 100x20x20.



Performance of parallel code

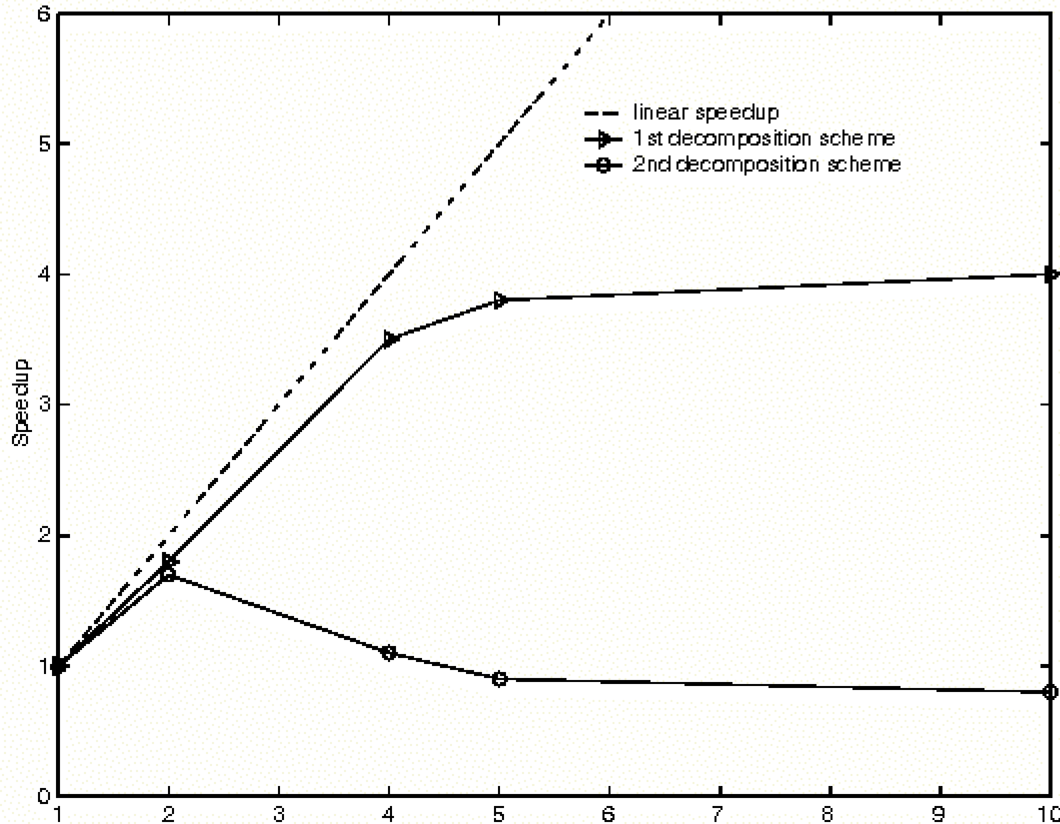
- **Performs well for the decomposition scheme with long elements along current flow direction. In this case inter-domain particle crossings are less frequent. Communication load is balance with computation load.**

- **Performance eventually saturates due to limited size of the test problem and scalar Poisson's equation solver.**

- **For the second decomposition scheme, performance degradation occurs with increase in number of processors used, after an initial improvement, due to large number of particles crossing domain boundaries.**

3-D Parallel Monte Carlo

Performance of parallel code



V_{bias} = 0.25 V

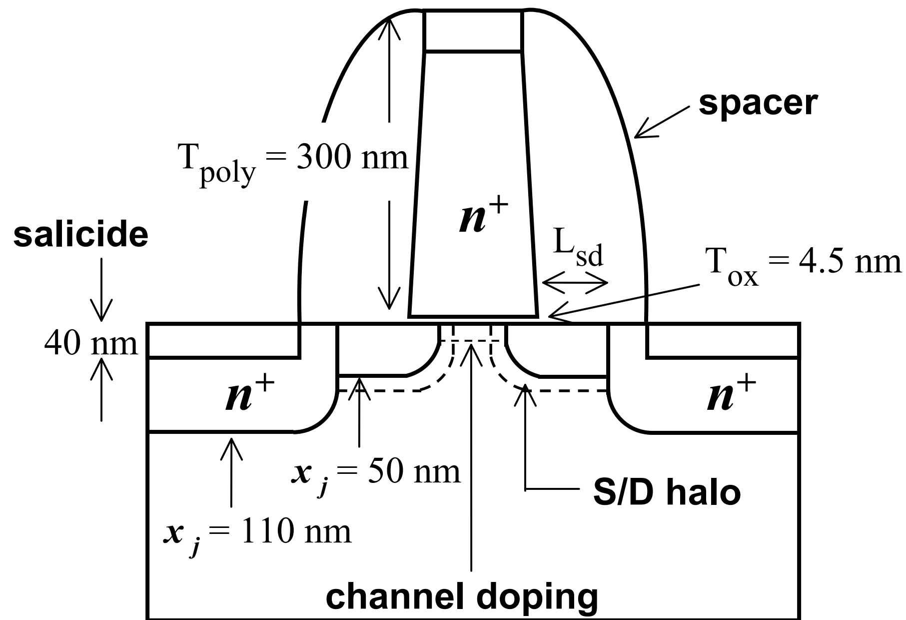
Some Conclusions (1)

- **One needs to have some intuition about current flow patterns inside the device to be simulated. This helps design of a decomposition scheme, which minimizes communication costs.**
- **If device structure is complex and has asymmetries in the current flow, adaptive decomposition schemes could help.**
- **Two guidelines one could use for decomposition scheme design: i) volume to surface ratio of sub-domains should be maximized, ii) the topology of the sub-domain should minimize boundary crossings in the direction of current flow.**

Conclusions (2)

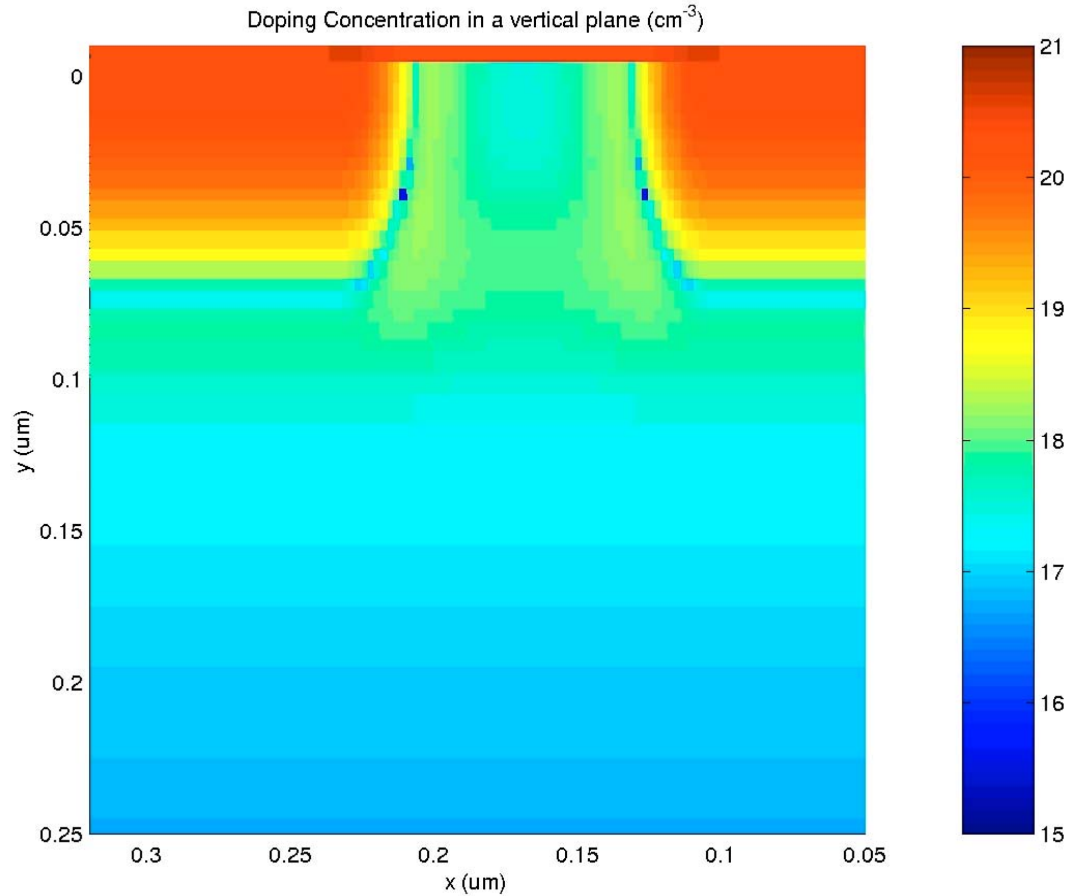
- **It has been demonstrated, even for small problem sizes, that achieving good speed-up is possible.**
- **Speed-up obtainable strongly depends on the quality of the decomposition. An adaptive decomposition scheme generator would help, achieving high performance levels for unfamiliar problems.**
- **It has been demonstrated that an MPI based parallel Monte Carlo simulator running over a cluster of commodity computers can yield performance comparable to much more expensive custom parallel computers.**

Well-Tempered MOSFET prototype 1



Sample structures at: http://www-mtl.mit.edu/Well/

90 nm Well-Tempered MOSFET doping profile



90 nm Well-Tempered MOSFET simulation conditions

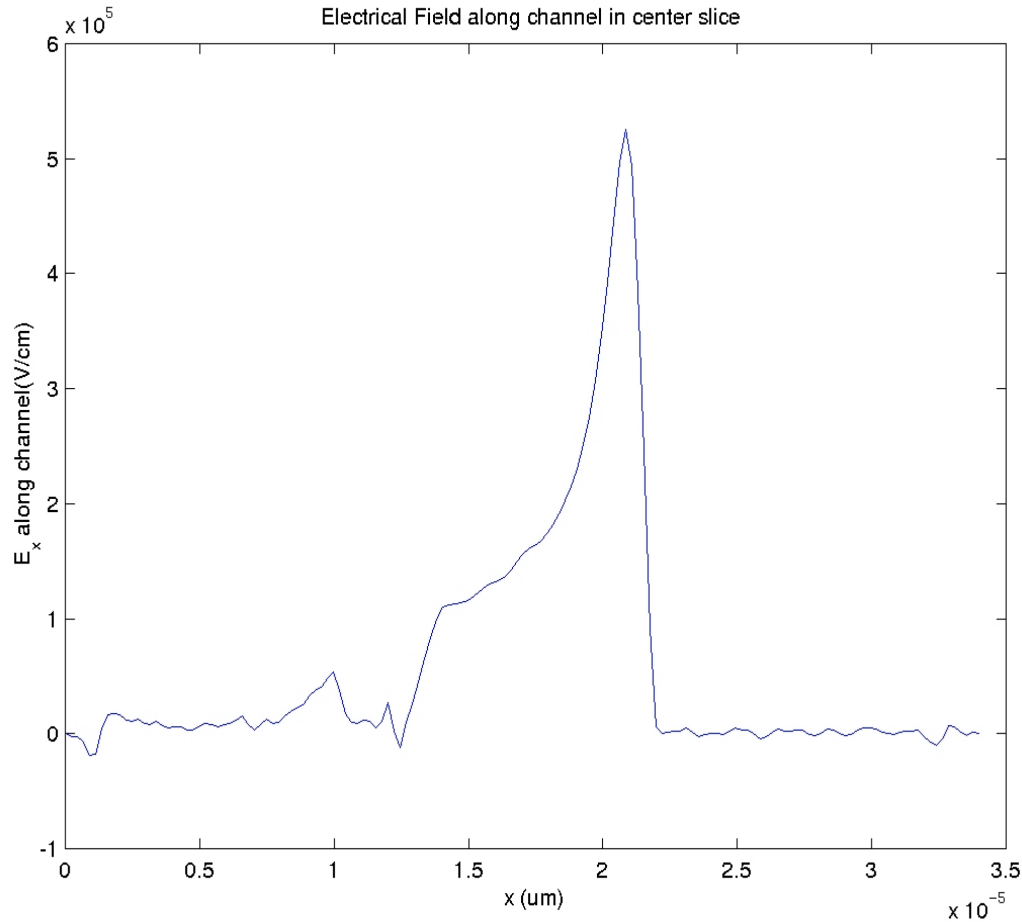
A total of about 50,000 electrons are used in typical simulations.

Transport time-step is the same as the interval between consecutive Poisson solutions, and was 0.1 fs.

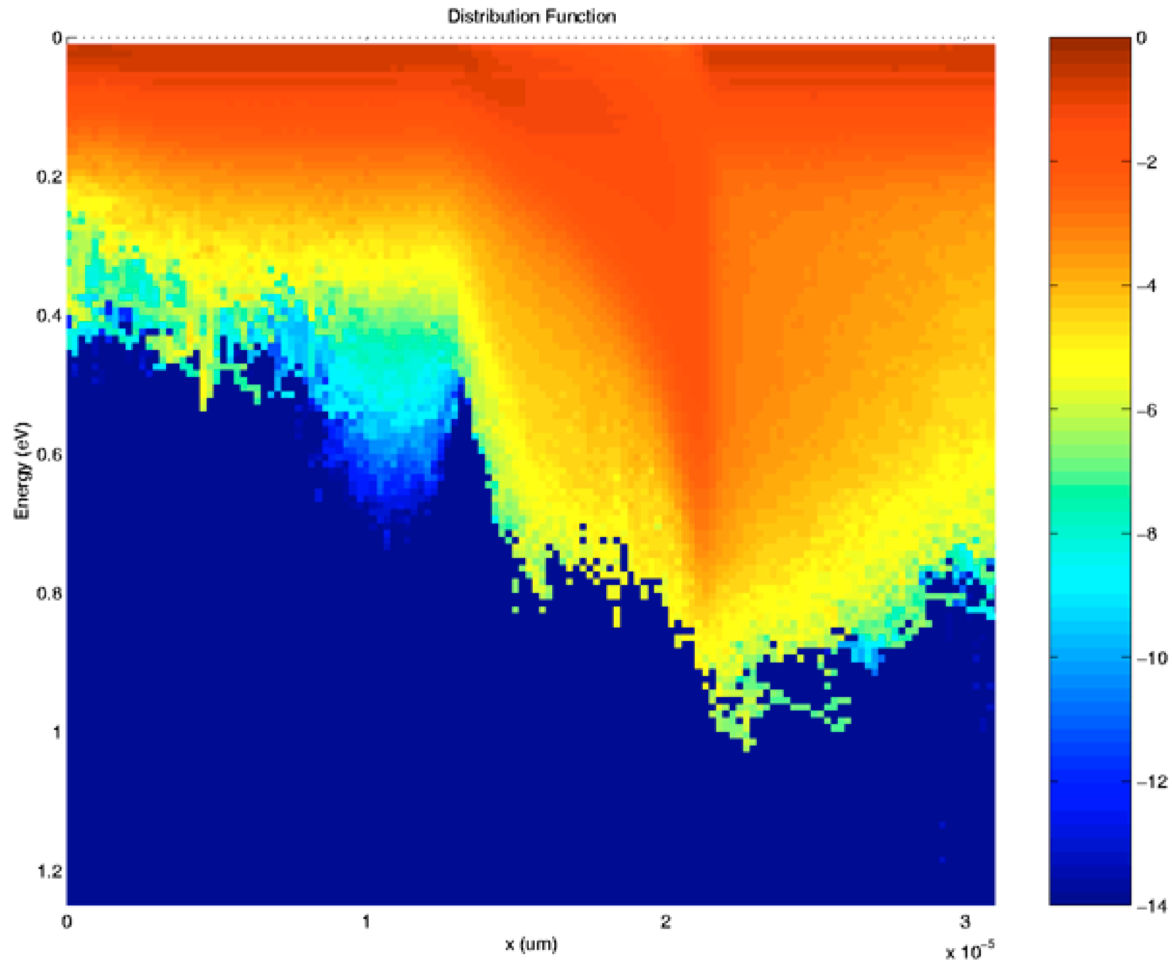
120,000 time steps simulated = 12 ps simulation time. Data was collected after a transient of 50,000 time step.

Wall clock time of simulation 8 hours on 16 CPUs.

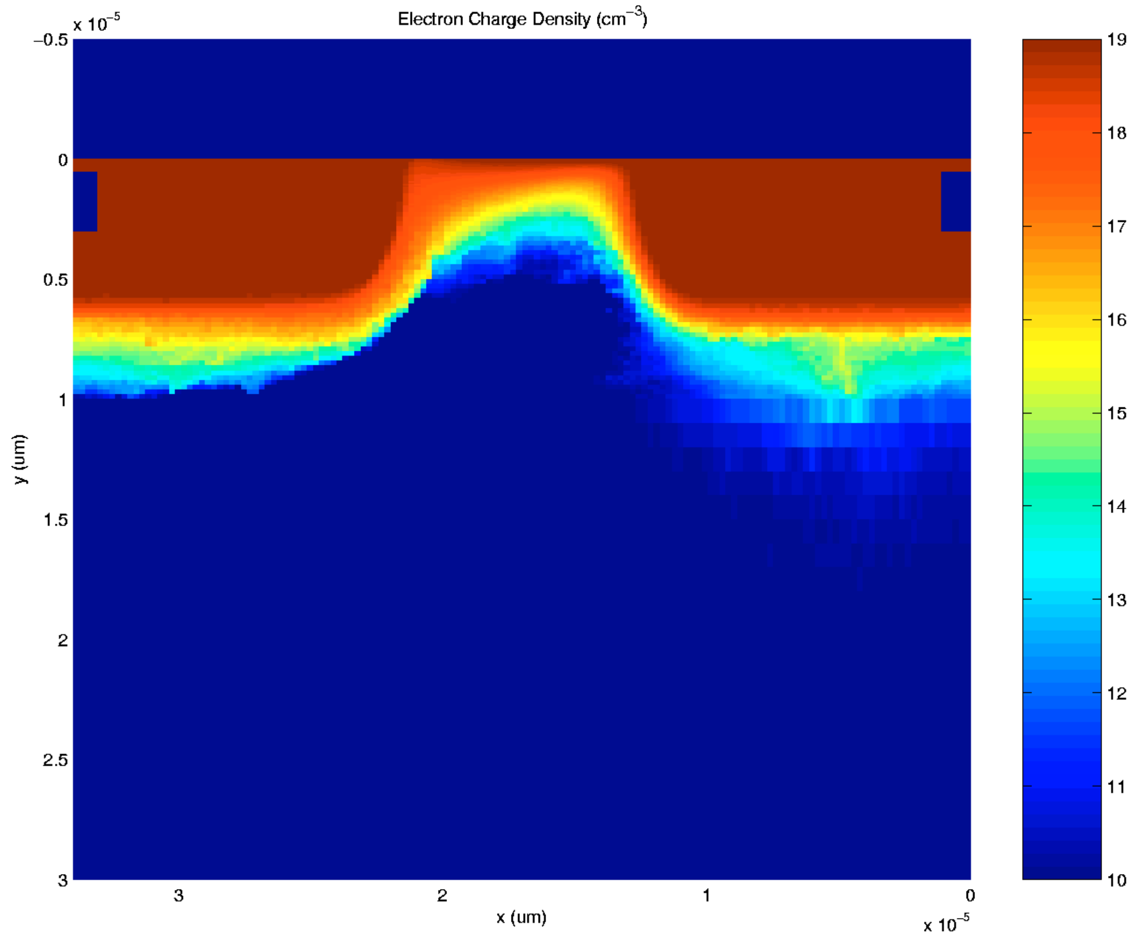
90 nm Well-Tempered MOSFET – longitudinal field result



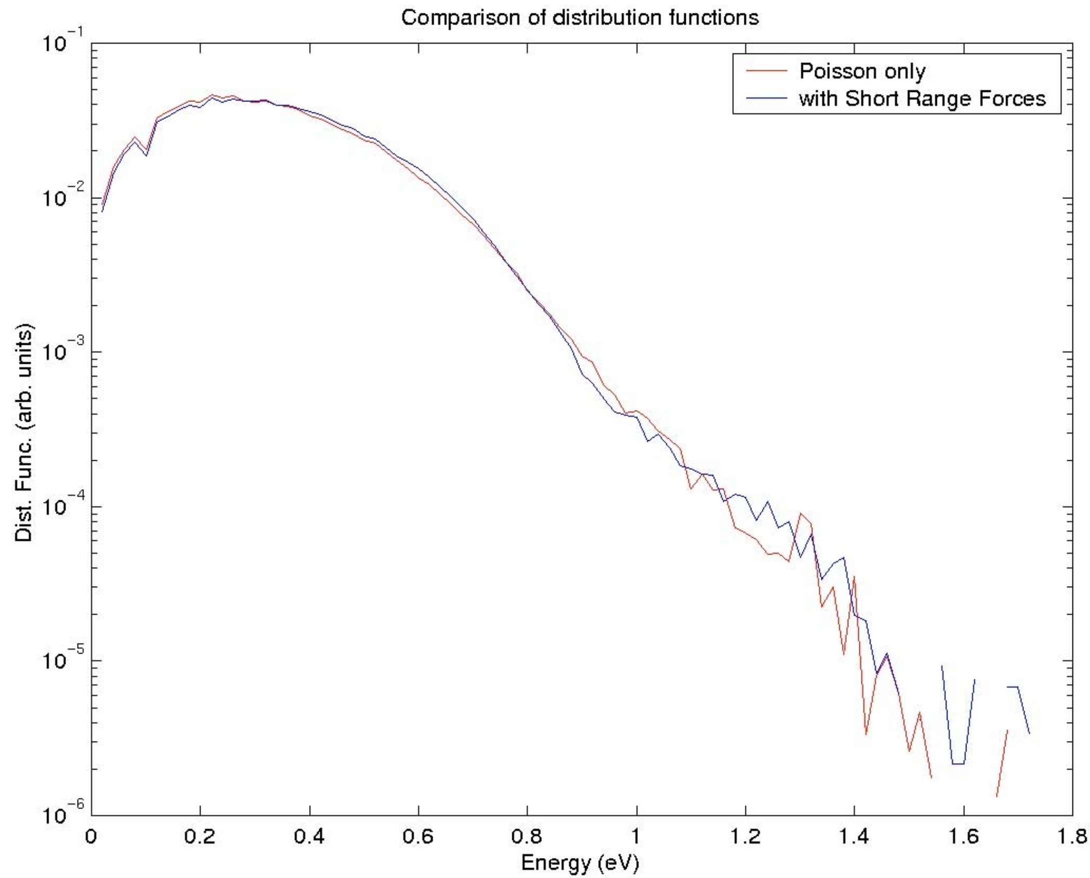
90 nm Well-Tempered MOSFET – distribution function



90 nm Well-Tempered MOSFET – electron density



90 nm Well-Tempered MOSFET – Test of P^3M short range calculations



90 nm Well-Tempered MOSFET – Test of P^3M short range calculations

- **Communication pattern between nodes changes significantly because of the different coupling, when charge-charge interaction is included. This affects the run time of the simulation. Additional domain decomposition strategies may be necessary.**

- **Results with and without P^3M are similar, except for a small amount of heating caused by the additional short range forces. This means that the fine Poisson mesh used is very effective for this channel length and regime of operation.**

- **With P^3M the Poisson mesh can be coarsened to reduce overall simulation cost.**

90 nm Well-Tempered MOSFET – Test of P^3M short range calculations

- **The particle-particle interaction introduces additional data exchanges among nodes, reducing the overall performance of the code. The short range interaction radius is chosen to be minimum distance which spans two mesh cells in any direction.**
- **If the slice thickness in the domain decomposition is set to four meshes, information on all the particles in the domain must be reported to the neighboring nodes on either side of the slice, as all particles are within the short-range interaction radius from the domain interfaces.**
- **As a result the runtime performance of the code suffers significantly in parallel mode, with 2 to 3 times increase in execution cost.**

90 nm Well-Tempered MOSFET – Test of P^{3}M short range calculations

- In the device simulation presented, dopants are still treated as smooth varying functions of positions. Charge-ion interaction is included using scattering rates.
- Granular dopant can be included simply, since the P^{3}M method treats electron-ion interaction in the same way as electron-electron interaction.
- Granularity will be important in very small devices. For instance, if the channel density is 10^{19} cm^{-3} average ion distance is about 5 nm. Only a few channel ions would be on the path of a particle.

Monte Carlo simulation of Schottky barrier MOSFET

(Winstead and Ravaioli, IEEE TED, vol. 47, p. 1241, 2001)

Quantum injection through the contacts is coupled to semi-classical Monte Carlo.

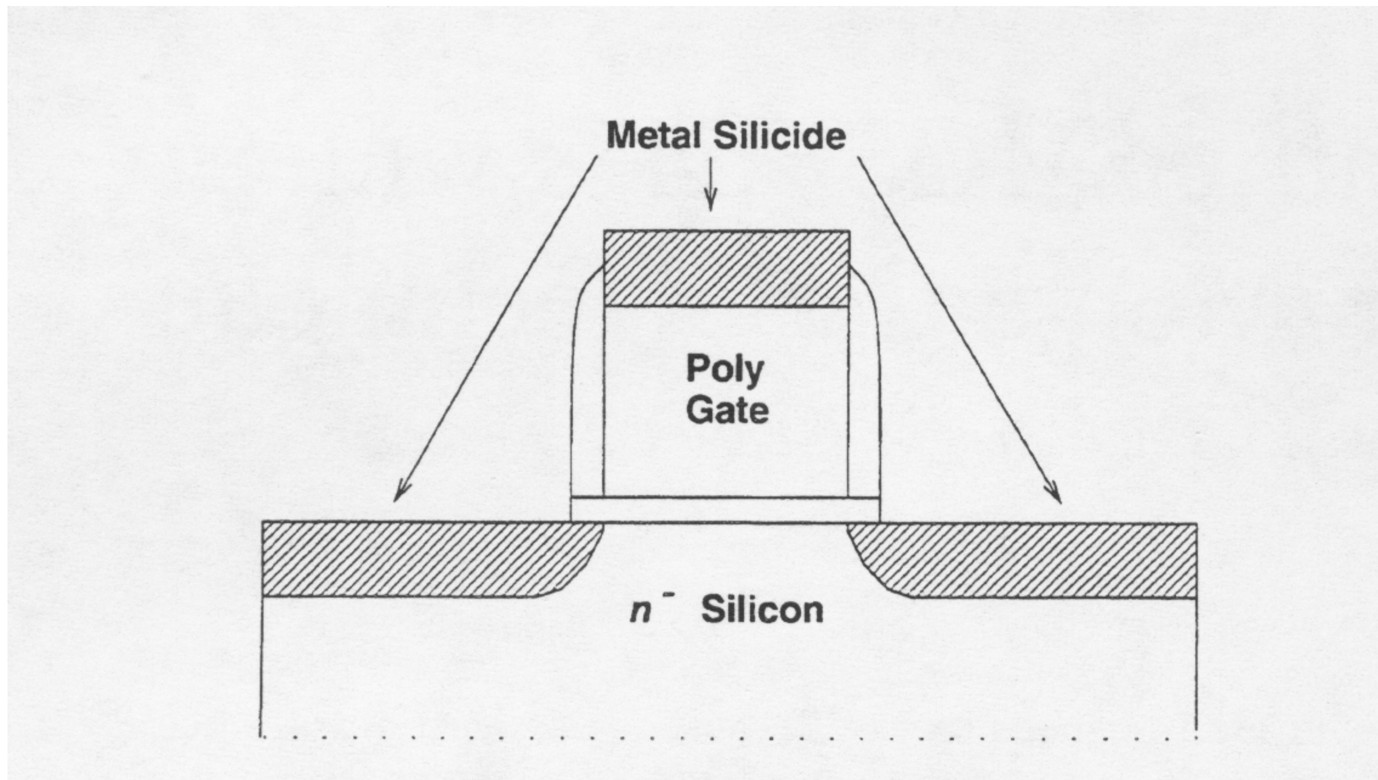


Fig. 1. Schematic structure of a Schottky barrier MOSFET. Only light doping exists in the channel, and the source and drain are metal silicide.

Monte Carlo simulation of Schottky barrier MOSFET

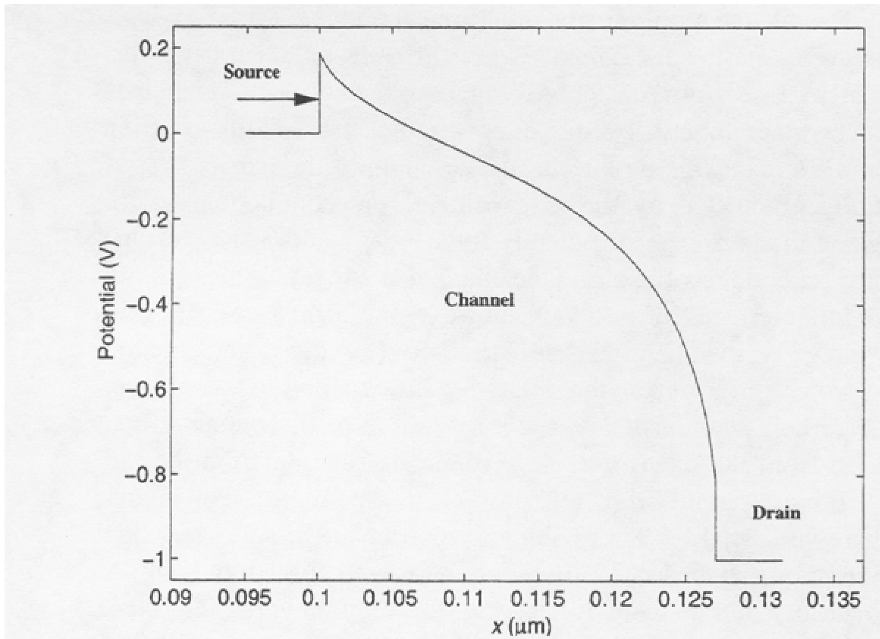


Fig. 2. Potential energy profile along the interface of a Schottky barrier MOSFET under subthreshold conditions. The thickness of the barrier significantly reduces the tunneling current.

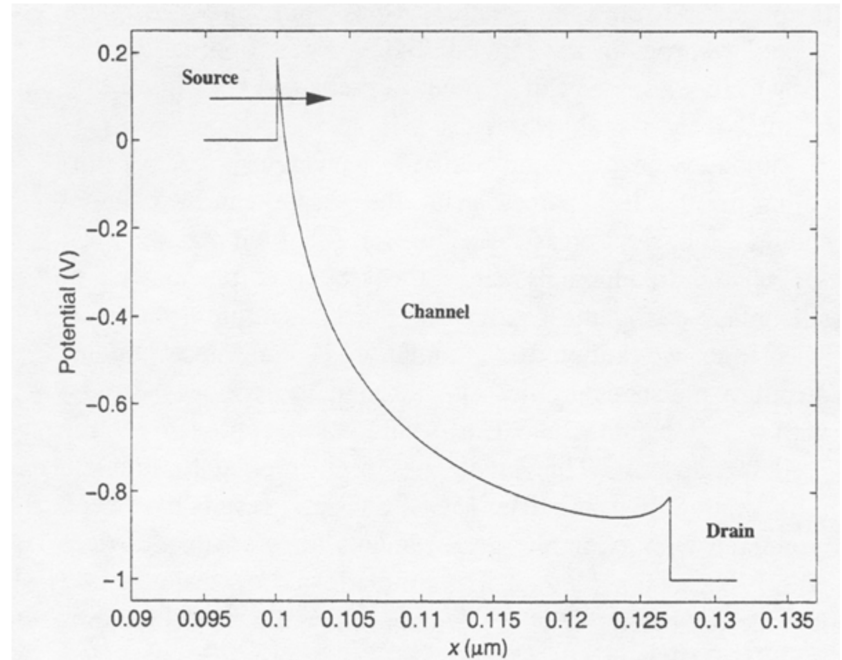


Fig. 3. Potential energy profile along the interface of a Schottky barrier MOSFET under high-field conditions. The barrier becomes transparent to current flow as the barrier becomes thin enough for quantum mechanical tunneling.

Monte Carlo simulation of Schottky barrier MOSFET

The silicide contacts are modeled as ideal metals. A distribution of injected particles is obtained statistically by using a transmission probability across the Schottky barrier, calculated a set of 1-D Schrödinger equations using the Airy function approach.

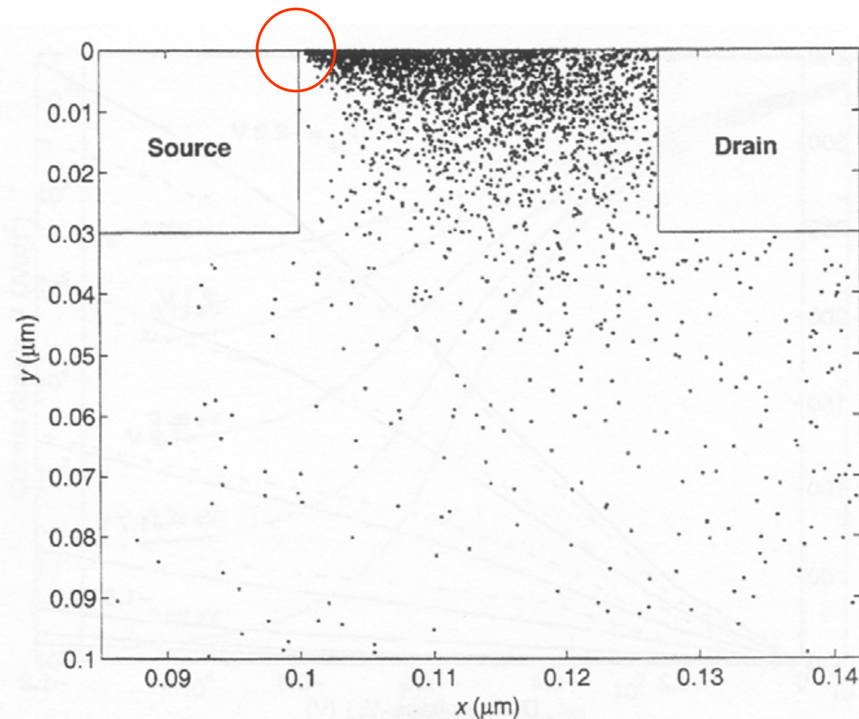


Fig. 4. Typical particle snapshot in time for the Monte Carlo simulation of a 270-Å p-channel Schottky barrier MOSFET in saturation bias. Approximately 90% of the current is injected within 50 Å of the Si-SiO_2 interface.

Monte Carlo simulation of Schottky barrier MOSFET

The particles in the contacts are not simulated explicitly, therefore the channel resolution is very good and subthreshold characteristics can be resolved clearly.

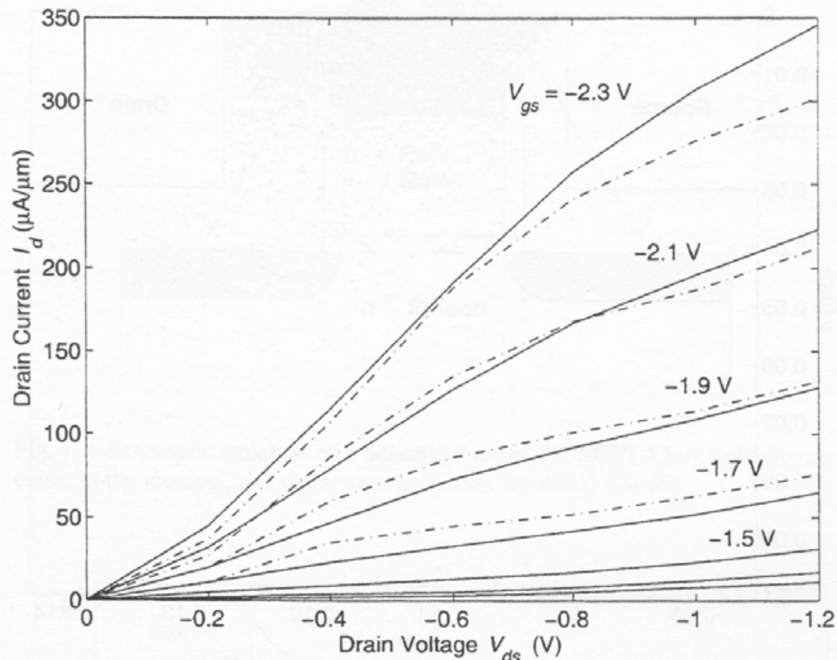


Fig. 5. I–V characteristics of a 270 Å p-channel Schottky barrier MOSFET. The oxide thickness is 19 Å. The solid lines are experimental results from [1] and the dash-dot lines are from Monte Carlo simulation.

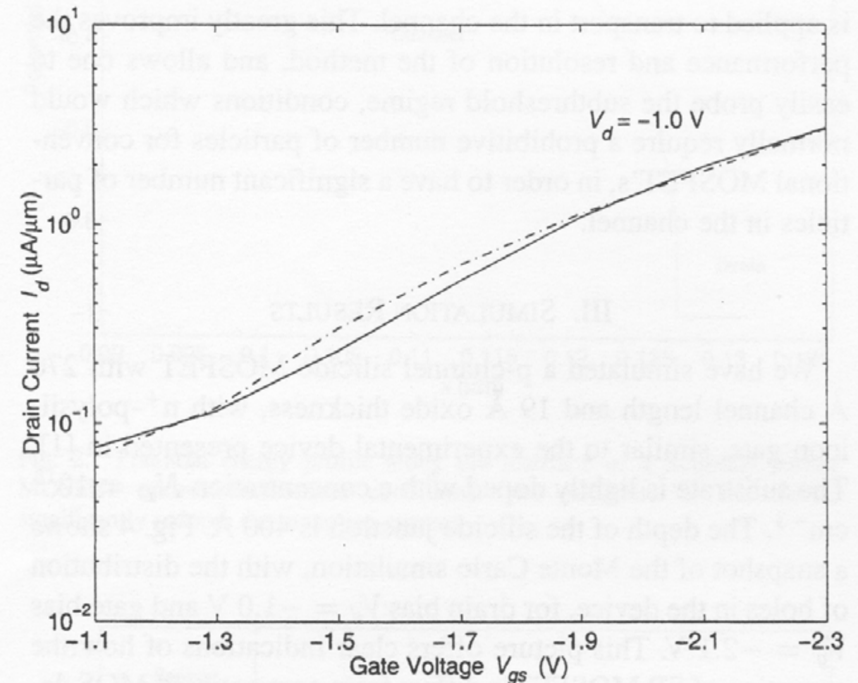


Fig. 6. Subthreshold characteristics of the same device in Fig. 5. The solid line is an experimental result from [1] and the dash-dot line is from Monte Carlo simulation.

[1] C. Wang, J.P. Snyder and J.R. Tucker, Appl. Phys. Lett., vol. 74, pp. 1174-1176, 1999.