Numerical Parallel Algorithms for Large-Scale Nanoelectronics Simulations using NESSIE

Eric Polizzi, Ahmed Sameh

Department of Computer Sciences,

Purdue University

NESSIE

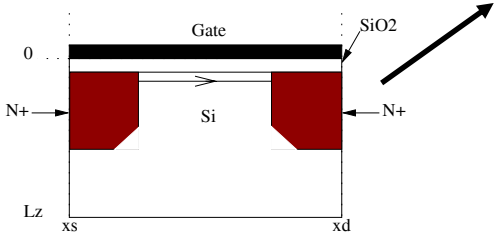
- NESSIE: a ‘top-down” multidimensional (1D, 2D, 3D) nanoelectronics simulator including:
  - Full quantum ballistic transport within NEGF/Poisson (transport Schrodinger/Poisson)
  - PDE-based model within effective mass or “multi-band” approach and FEM discretization
  - Non-equilibrium transport in 3-D structures using exact 3-D open boundary conditions
  - A Gummel iteration technique to handle the non-linear coupled transport /electrostatics problem
  - Semi-classical and/or hybrid approximations to obtain a good initial guess at equilibrium
  - General multidimensional subband decomposition approach (mode approach)
  - Asymptotic treatment of the mode approach: quasi-full dimensional model
  - The most ``suitable'' parallel numerical algorithms for the target high-end computing platforms

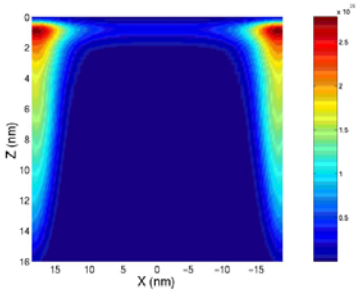
- NESSIE (1998-2004) has been used to simulate
  - 3D electron waveguide devices- III-V heterostructures: E. Polizzi, N. Ben Abdallah, PRB 66, (2002)
  - 2D MOSFET and DGMOSFET: E. Polizzi, N. Ben Abdallah, JCP in press (2004)
  - 3D Silicon Nanowire Transistors, see J. Wang, E. Polizzi, M. Lundstrom, JAP, 96, (2004)

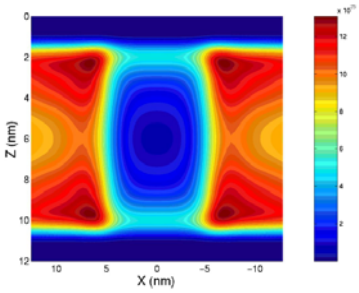
- NESSIE can be used to study a wide range of characteristics (current-voltage, etc…) of many other multidimensional realistic quantum structures.

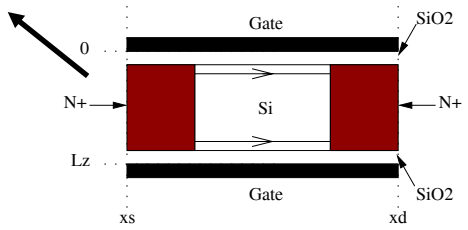
➔ By allowing the integration of different physical models, new discretization schemes, robust mathematical methods, and new numerical parallel techniques, NESSIE is becoming an extremely robust simulation environment

Simulation Results using NESSIE

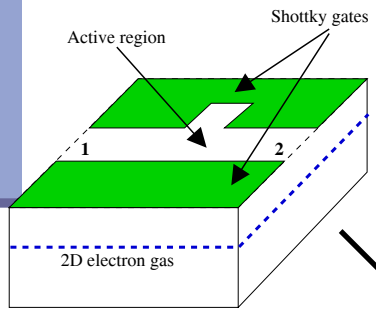




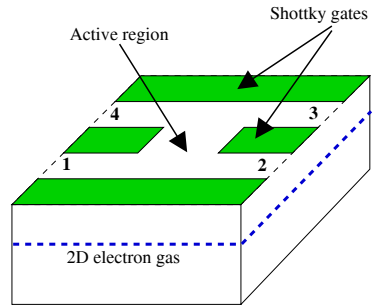


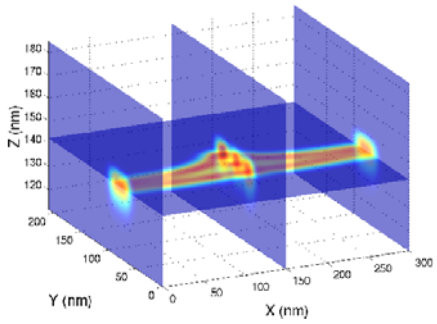


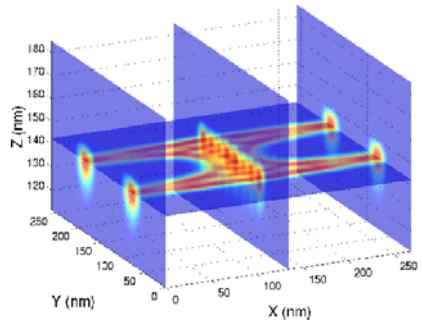
nanoMOSFETS: Full 2D simulation



III-V heterostructures: Full 3D simulation







Numerical Techniques

- linear systems on the Green's function or wave function:

$$\left(E[S] - [H(V)] - [\Sigma_E]\right)\ \mathbf{X_E} = \mathbf{F_E}, \quad \forall E$$

  - (E[S]-[H]) is large, sparse, real symmetric (hermitian in general case)
  - $[\Sigma_E]=[\Sigma_1(E)]+\ldots+[\Sigma_p(E)]$, and $[\Sigma_j]$ is "small", dense, complex symmetric

  ➔ Parallel MPI procedure on the energy where each processor handles many linear systems

  ➔ Krylov subspace iterative method uses on one processor

- Linear system on the potential (modified Poisson equation)

AX=F

  - A is large, sparse, s.p.d

Simulation Results using NESSIE

For only one point in the I-V curve	Full 2D	Full 3D
Matrix size	$O(10^4)$	$O(10^6)$
linear systems to solve by iteration	$O(10^3)$	$O(10^3)$
Number of Gummel iterations	O(10)	O(10)
Simulation time (uniprocessor)	O(**hours**)	O(**days**)

➔**Current algorithms for obtaining I-V curves are in need of improvement**

- Remark: for particular devices, the dimension of the transport problem can be reduced using a subband decomposition approach (mode approach)-

Poster session

- Silicon Nanowire Transistors: J. Wang, E. Polizzi, A. Ghosh, S. Datta, M. Lundstrom
- A WKB based method: N. Ben Abdallah, N. Negulescu, M. Mouis, E. Polizzi

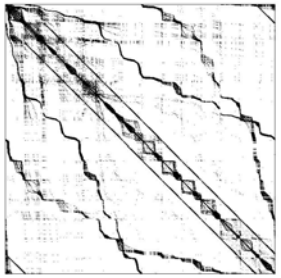
The need of high-performance parallel numerical algorithms

- Problem for large-scale computation:
  - Each processor handles many linear systems
  - The size N_j of $[\Sigma_j]$ (dense matrix) will increase significantly
  - Integration over the energy on a non-uniform grid (quasi-bound states)

- New proposed strategy:
  - Each linear system is solved in parallel
  - Strategy of preconditioning to address all these problems

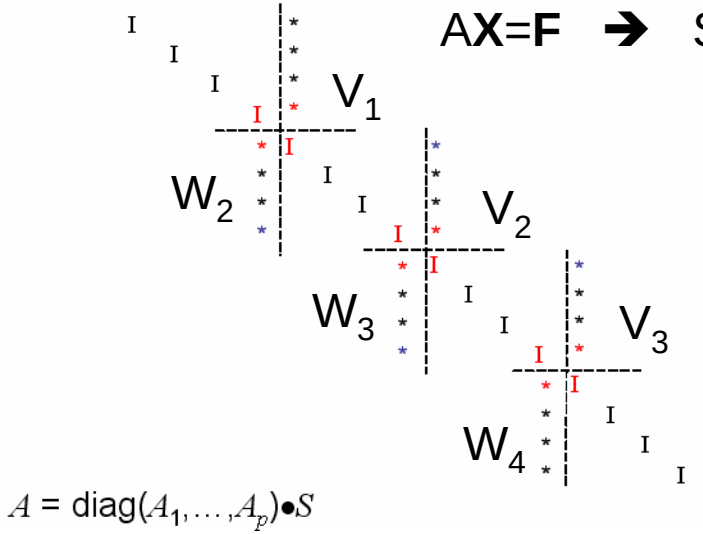
SPIKE: A parallel hybrid banded linear solver

- Engineering problems usually produce large sparse matrices
- Banded structure is often obtained after reordering
- SPIKE partitions the banded matrix into a block tridiagonal form
- Each partition is associated with one node or one CPU ➔ multilevel of parallelism



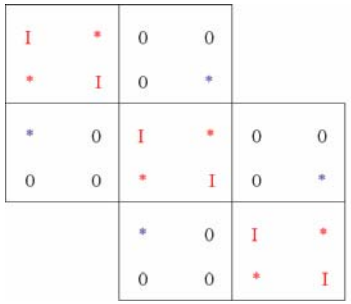
"NESSIE matrix"

After RCM reordering



$A = \text{diag}(A_1,\ldots,A_p)\bullet S$

$AX=F \rightarrow SX=\text{diag}(A_1^{-1},\ldots,A_p^{-1})\ F$



Reduced system

Retrieve solution

SPIKE: improvement over ScaLAPACK

N=480, 000; RHS=1; #procs= 32, dense within the band

IBM-SP

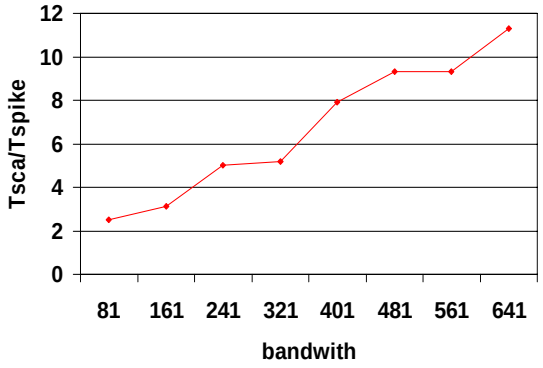
SPIKE as Preconditioner

SPIKE Preprocessing on A'

ITERATIVE METHOD

- SPIKE SOLVER A'z=r
- MATRIX-VECTOR MULTI. Ax=r

If "zero-pivot" detected in preprocessing



Spike w/o pivoting

Time (s) and Tscal/Tspike	Preprocess.		Solver		Total	
	Tscal	Tspike	Tscal	Tspike	Tscal	Tspike
bandwith b=81	0.49	0.21 **2.4**	0.090	0.022 **4.1**	0.58	0.23 **2.5**
b=161	1.63	0.53 **3.1**	0.130	0.044 **2.9**	1.75	0.57 **3.1**
b=241	5.24	1.03 **5.1**	0.20	0.064 **3.1**	5.44	1.10 **5.0**
b=321	8.83	1.65 **5.3**	0.25	0.078 **3.2**	9.08	1.73 **5.2**
b=401	20.61	2.56 **8.1**	0.31	0.099 **3.1**	20.61	2.66 **7.9**
b=481	34.75	3.68 **9.5**	0.37	0.12 **3.1**	35.12	3.79 **9.3**
b=561	47.99	5.05 **9.5**	0.48	0.14 **3.6**	48.47	5.19 **9.3**
b=641	75.69	6.56 **11.5**	0.66	0.17 **3.9**	76.36	6.74 **11.3**

SPIKE: Scalability

b=161; RHS=1;

IBM-SP

Spike (RL0)

N=480,000; b=161; RHS=1

# procs.	4	8	16	32	64	128	256	512
Tscal.(s)	13.06	6.60	3.4	1.78	0.95	0.56	0.38	0.40
Tspike (s)	4.17	2.22	1.12	0.58	0.3	0.18	0.17	0.22
Tscal/Tspike	**3.1**	**3.0**	**3.0**	**3.1**	**3.2**	**3.1**	**2.2**	**1.8**

N=960,000; b=161; RHS=1

# procs.	4	8	16	32	64	128	256	512
Tscal. (s)	26.21	12.98	6.76	3.42	1.83	0.98	0.60	0.39
Tspike (s)	8.4	4.42	2.23	1.13	0.62	0.32	0.22	0.17
Tsca/Tspike	**3.1**	**2.9**	**3.0**	**3.0**	**2.9**	**3.1**	**2.8**	**2.3**

N=1,920,000; b=161; RHS=1

# procs.	4	8	16	32	64	128	256	512
Tscal. (s)		26.23	13.35	6.74	3.44	1.89	1.00	0.70
Tspike (s)	17.20	8.68	4.42	2.25	1.14	0.63	0.34	0.27
Tsca/Tspike		**3.0**	**3.0**	**3.0**	**3.0**	**3.0**	**3.0**	**2.6**

SPIKE inside NESSIE

- Problem for large-scale computation in NESSIE:

1) Each processor handles many linear systems
2) The size N_j of $[\Sigma_j]$ (dense matrix) will increase significantly
3) Integration over the energy on a non-uniform grid (quasi-bound states)

- SPIKE inside NESSIE

1) Each linear system is solved in parallel using SPIKE
2) $(E_1[S]-[H])$ is a good preconditioner for $(E_1[S]-[H]-[\Sigma_{E1}])$
   - Neumann B.C. for the preconditioner
   - ~2-3 outer-iterations of BiCG-stab
   - $[\Sigma_{E1}]$ is now requiring only in mat-vec multiplications that can be done on the fly for very large system
3) We use $(E_1[S]-[H])$ as preconditioner for $(E_2[S]-[H]-[\Sigma_{E2}])$
   - $(E_2-E_1) < \delta E$, the preconditioner is updated if # of iteration > N_{max}
   - Solver time of SPIKE<< preprocessing time
     ➔ Fast algorithm
     ➔ Refinement of the energy grid

Conclusion and Prospect

- NESSIE: A robust 2D/3D simulator and a nanoelectronics simulation environment
- SPIKE: An efficient parallel banded linear solver
  - Significant improvement vs ScaLapack
  - A version of Spike for matrices that are sparse within the band is under development
- SPIKE inside NESSIE: strategy to address large-scale nanoelectronics simulations