Scaling Deterministic Numerical Solutions of the Boltzmann Transport Equation on Heterogeneous Computing Platforms

Karl Rupp[1,2], Andreas Morhammer[1],
Tibor Grasser[1], Ansgar Jüngel[2]

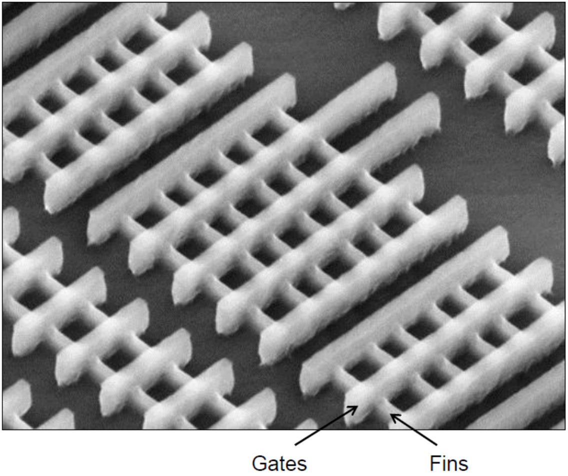
[1] Institute for Microelectronics
[2] Institute for Analysis and Scientific Computing
TU Wien, Austria

Scalable Methods for Kinetic Equations
Oak Ridge National Laboratory
October 20th, 2015

Contents

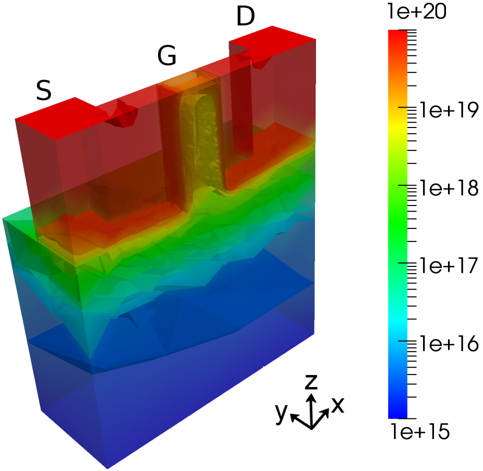
Semiconductor Devices in 3D: FinFET

Intel Trigate transistors



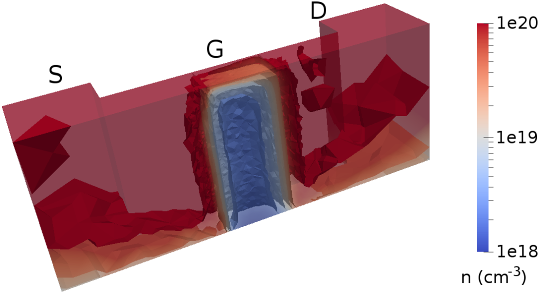
http://download.intel.com/newsroom/kits/22nm/gallery/images/Intel-22nm_Transistor_2.jpg

Semiconductor Devices in 3D: FinFET



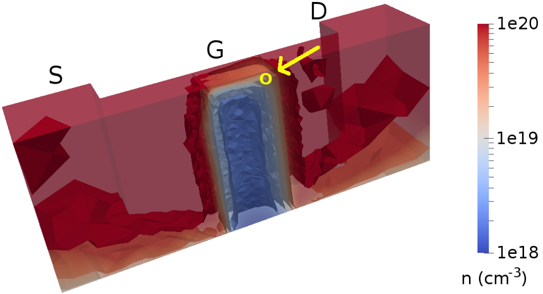
Electron Density in a FinFET

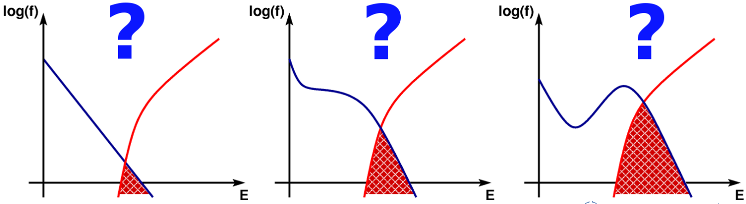
Density of electrons at each point $\mathbf{x}$



Electron Energy Distribution?

Distribution of electrons with respect to energy at $\mathbf{x}$?





Electron Energy Distribution?

Macroscopic Transport Models

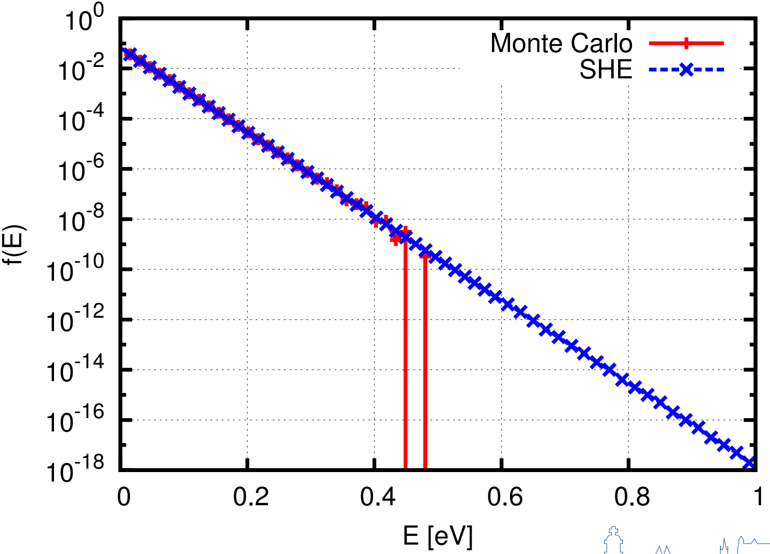
- Invalid in deca-nanometer regime
- “Fitting” only treats the symptoms, not the cause
- Only averaged quantities of the carrier ensemble modeled

Boltzmann Transport Equation (BTE)

$$\frac{\partial f}{\partial t} + \mathbf{v}(\mathbf{k}) \cdot \nabla_{\mathbf{x}} f + \mathbf{F}(\mathbf{x}) \cdot \nabla_{\mathbf{k}} f = Q\{f\}$$

- Best semi-classical description of carrier transport
- Posed in a seven-dimensional $(\mathbf{x}, \mathbf{k}, t)$ space
- Most popular solution method: Monte Carlo

Electron Energy Distribution?



Spherical Harmonics Expansion Method

Spherical Symmetries

Maxwell distribution of carriers at equilibrium

Dispersion relation (Herring-Vogt transform, approx.)

Spherical Harmonics Expansion (SHE)

$$f(\mathbf{x}, \mathbf{k}, t) \simeq \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{l,m}(\mathbf{x}, E, t) Y_{l,m}(\theta, \varphi)$$

New unknowns: $f_{l,m}(\mathbf{x}, E, t)$

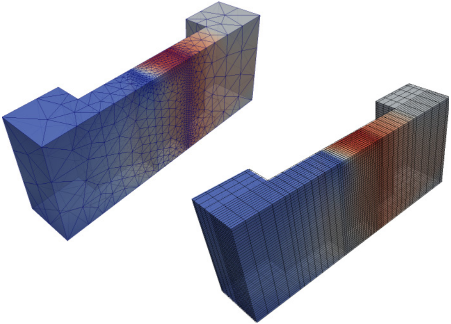
Solution in five-dimensional $(\mathbf{x}, E, t)$-space

S.-M. Hong and C. Jungemann, *J Comput Electron* (2009):
Fifth-order, three-dim. $(\mathbf{x}, E)$-space, 26 GB memory, 12 hours

Unstructured Grids

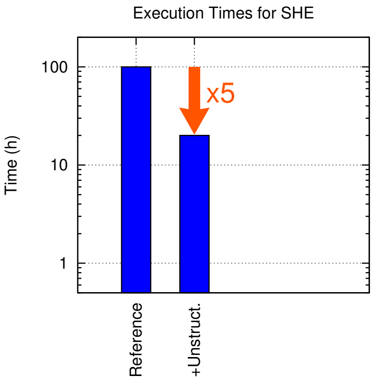
- State-of-the-art in modern TCAD
- Only structured grids in publications on higher-order SHE in 2D
  - [S.-M. Hong and C. Jungemann (2008), S.-M. Hong and C. Jungemann (2009)]
- Extension of discretization proposed by Hong and Jungemann

4838 nodes

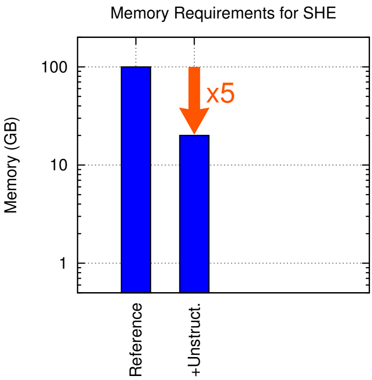


27456 nodes

Summary



Execution Times for SHE
Time (h)
100
10
1
x5
Reference
+Unstruct.



Memory Requirements for SHE
Memory (GB)
100
10
1
x5
Reference
+Unstruct.

Spherical Harmonics Expansion

$$f(\mathbf{x}, \mathbf{k}, t) \simeq \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{l,m}(\mathbf{x}, E, t) Y_{l,m}(\theta, \varphi)$$

$(L+1)^2$ unknown functions $f_{l,m}(\mathbf{x}, E, t)$

$L = 0$ sufficient in equilibrium

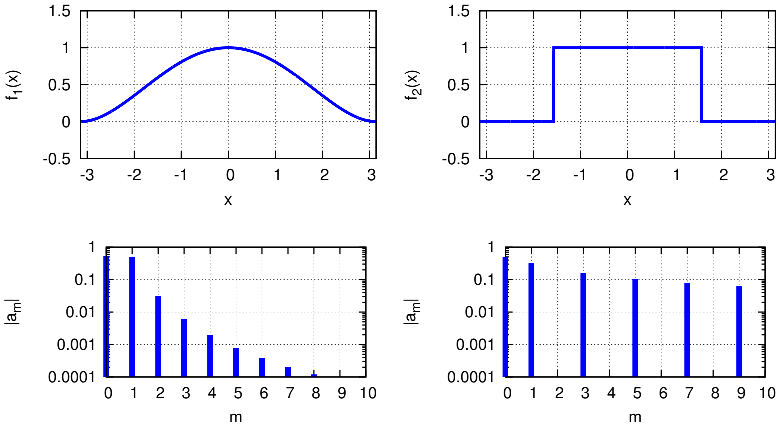
Higher-order expansions in active regions

Therefore: Variable-order SHE:

$$f(\mathbf{x}_i, \mathbf{k}_n, t) \simeq \sum_{l=0}^{L(\mathbf{x}_i, E_n)} \sum_{m=-l}^{l} f_{l,m}(\mathbf{x}_i, E_n, t) Y_{l,m}(\theta, \varphi)$$

How to choose $L(\mathbf{x}_i, E_n)$ in the simulation domain?

Motivation from Fourier series

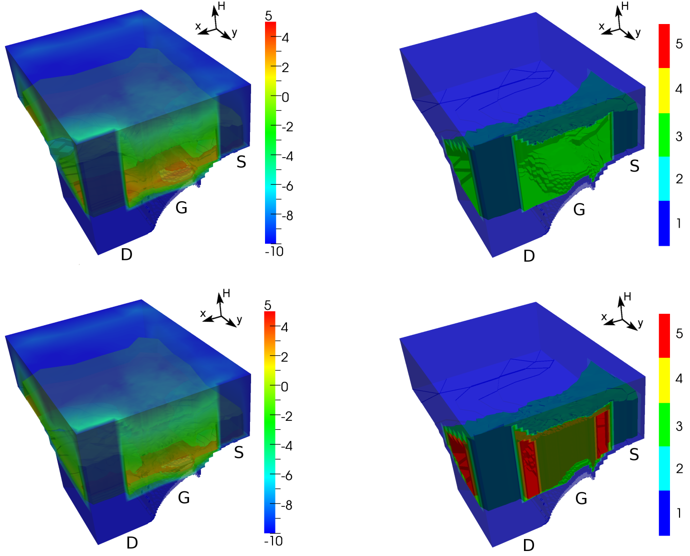


Adaptive Variable-Order SHE

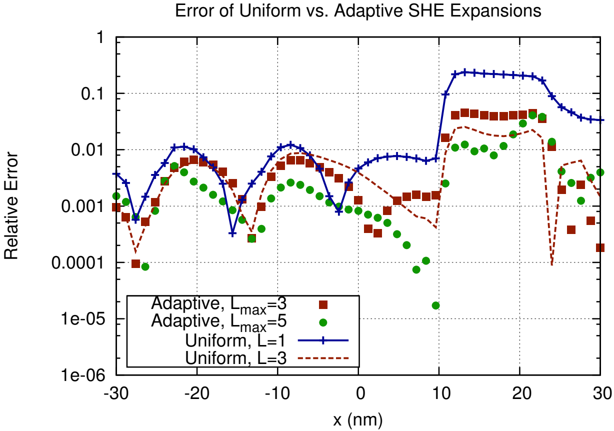
Error indicator: | Expansion order:

$L = 1$

$L = 3$



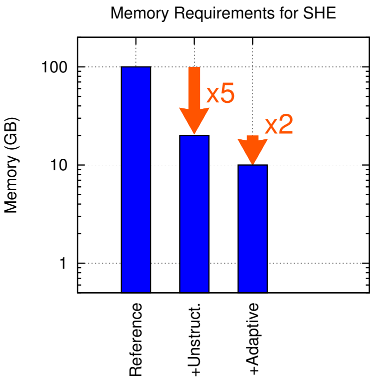
Adaptive Variable-Order SHE



$L = 3$: **306 261** instead of **476 061** unknowns (factor **1.5**)
$L = 5$: **606 671** instead of **1 146 120** unknowns (factor **1.9**)

Summary

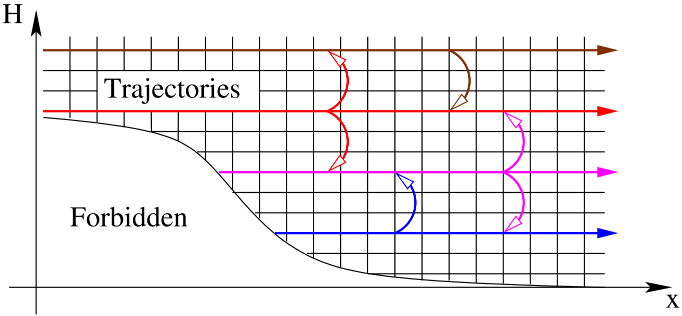

Execution Times for SHE
Time (h)
100
10
1
x5
x2
Reference
+Unstruct.
+Adaptive


Memory Requirements for SHE
Memory (GB)
100
10
1
x5
x2
Reference
+Unstruct.
+Adaptive

Parallelization

Preconditioner for Iterative Linear Solvers

- No fast general-purpose parallel preconditioner available
- Physics-based parallel block preconditioner developed



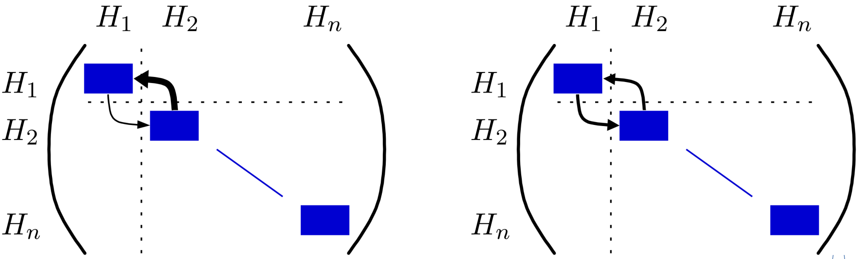
Scaling of Solution Variables

Exponential decay with energy: $f(E_i) \sim \exp(-\frac{E_i}{k_B T})$

Rescale unknowns: $\tilde{f}(E_i) = \exp(\frac{E_i}{k_B T}) f(E_i)$

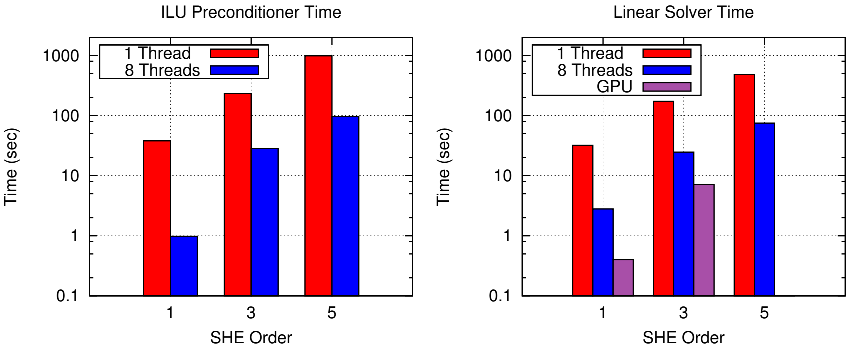
New system: $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \mathbf{ADD}^{-1}\mathbf{x} = \mathbf{b}$

Row normalization: $\hat{\mathbf{A}}\tilde{\mathbf{x}} = \mathbf{P}\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \mathbf{Pb}$



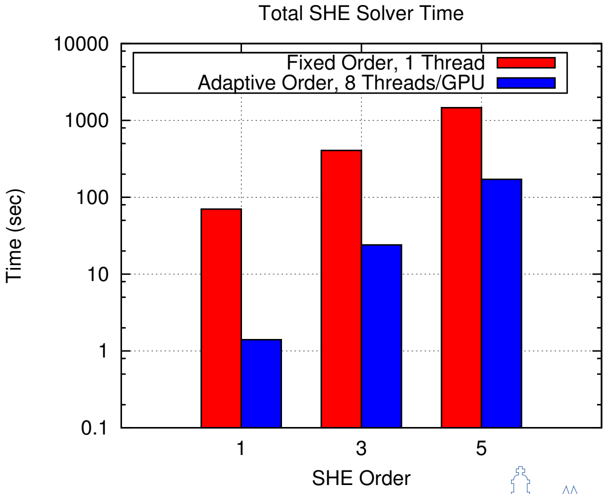
Parallelization

Benchmark results for a FinFET (INTEL Core i7 960, NVIDIA GTX 580)



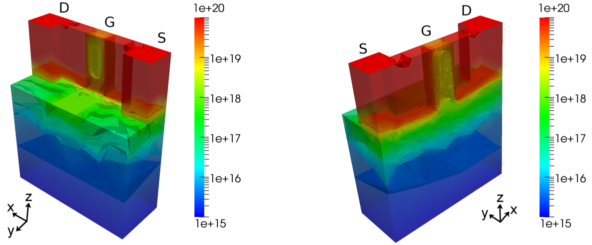
Parallelization

Benchmark results for a FinFET (INTEL Core i7 960, NVIDIA GTX 580)

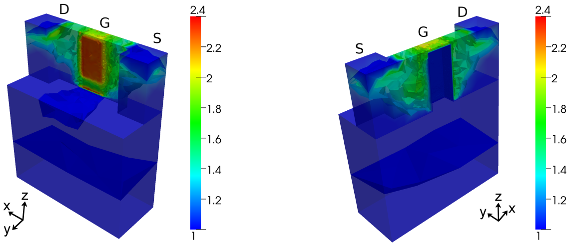


Results

Electron Concentration (cm^{-3})

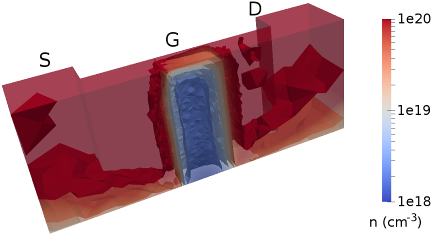


Avg. Expansion Order

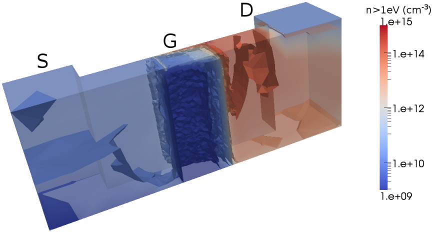


Results

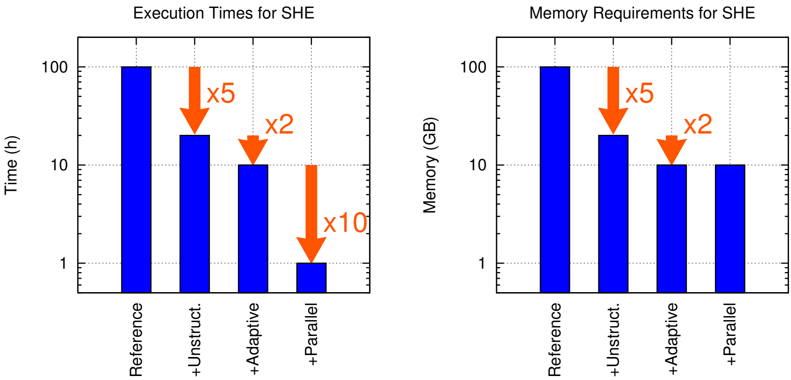
Electron Concentration (cm^{-3})



Electron Density with Energy above 1eV (cm^{-3})



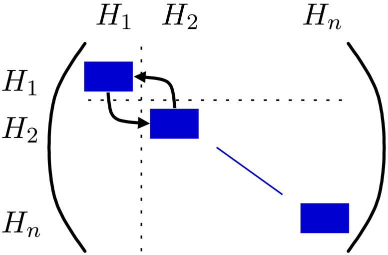
Summary for Shared Memory Machines



Part 2

Current Work:
Development of a Parallel Preconditioner
for (Heterogeneous) Distributed Memory Machines

Distributed SHE



Blueprint

Keep block-Jacobi based on total energies

Map total energies to MPI ranks

Requires fine-grained parallel preconditioner per block

Parallel ILU

General

Approximate factorization $\mathbf{A} \approx \mathbf{LU}$

Proposed by Chow and Patel (SISC, vol. 37(2)) for CPUs and MICs

Available in ViennaCL for CUDA, OpenCL, OpenMP

Preconditioner Setup

Nonlinear parallel sweeps to obtain l_{ij} and u_{ij}

Massively parallel (one thread per row)

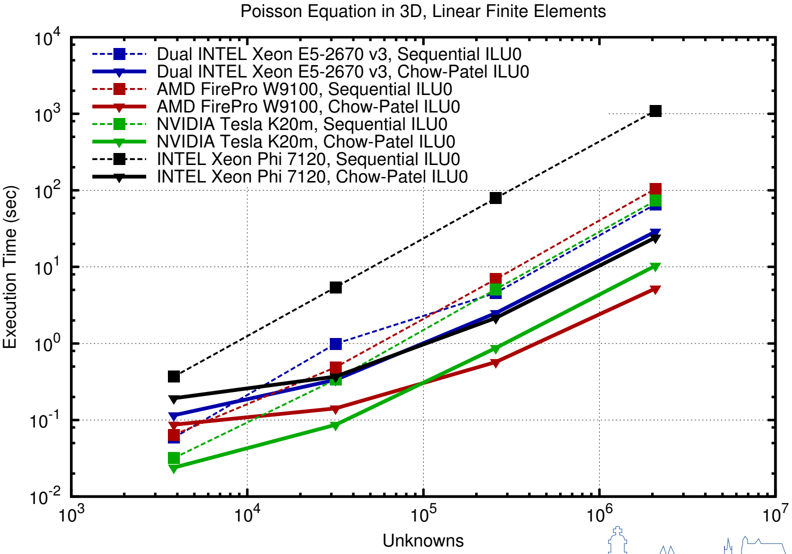
Preconditioner Application

Truncated Neumann series:

$$\mathbf{L}^{-1} \approx \sum_{k=0}^{K} (\mathbf{I} - \mathbf{L})^k, \quad \mathbf{U}^{-1} \approx \sum_{k=0}^{K} (\mathbf{I} - \mathbf{U})^k$$

Exact triangular solves not necessary

Parallel ILU



Preconditioner Evaluation

Resolution

Total energy spacing: approx. 10 meV

Thus: Hundred MPI ranks per eV (slight load imbalance)

Typical minimum range: 3-5 eV

Granularity

Fine-grained: One thread per matrix row on each MPI rank

Coarse-grained: One run per voltage bias

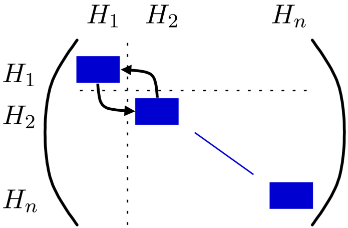
Evaluation

Preconditioner sufficient for typical TCAD workloads

Ready for upcoming HPC hardware

Spatial domain decomposition required for strong scaling limit

Preconditioner Alternatives



Other Shared Memory Preconditioners

Algebraic multigrid?

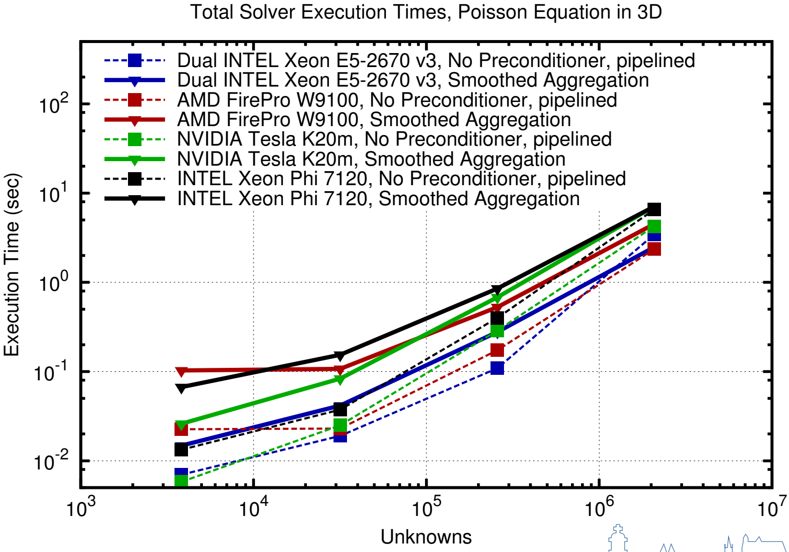
Polynomial preconditioners?

Block-Diagonal Inversion

Additive Schwarz (no overlap)

Sparse direct solver for each block?

Parallel AMG



Conclusion

SHE Method

- Viable alternative to Monte Carlo
- Full 3D device simulations possible
- Convergence behavior similar to drift-diffusion model
- Free open-source simulator: ViennaSHE

Large-Scale Solution

- Physics-based block-Jacobi preconditioner
- Replication of spatial mesh on all MPI ranks
- Fine-grained parallel ILU
- Combine functionality in PETSc and ViennaCL libraries