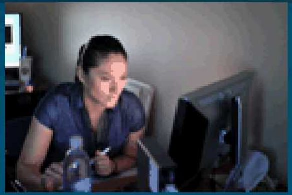
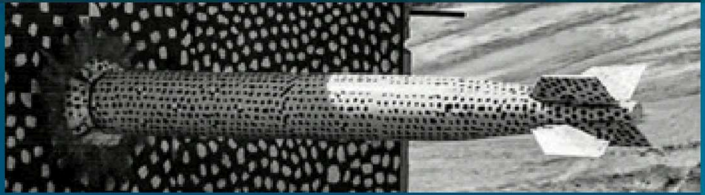
SAND2020-1754C

Technology Computer Aided Design Modeling of Semiconductor Devices in Parallel Computing Architectures





Lawrence C Musson, Gary Hennigan, Jason Gates, Xujiao Gao, Mihai Negoita, and Andy Huang

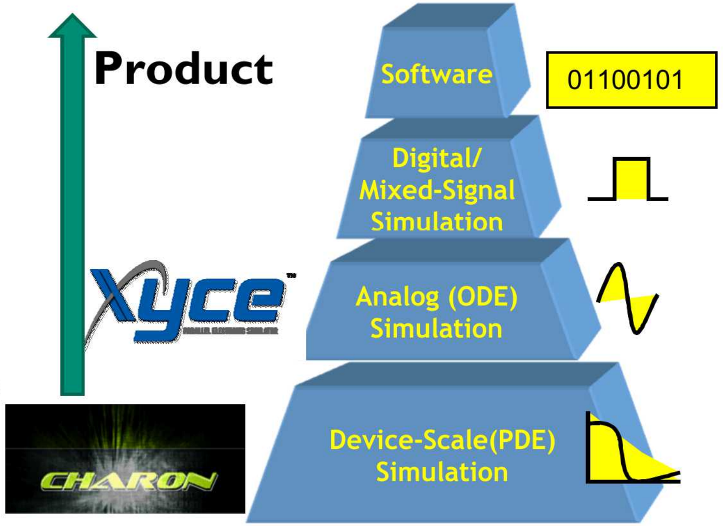
Sandia National Laboratories

SIAM PP20

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Charon Basics & Mission

- **Charon is a technology computer aided design (TCAD) code**
  - Solve partial differential equations to predict carrier transport of semiconductor devices
  - Include the effects of radiation on carrier transport and device performance
- **Charon strategy is to engage the national defense community for existing and future technologies**
  - Advocate Charon for sensitive radiation effects modeling and massively parallel over commercial alternatives
  - Develop capabilities according to customers' modeling needs
  - Target capability development for future technologies: what will be important in 5 years, 10 years, etc.?
- **Support all mission applications at Sandia**
  - Nuclear Deterrence, Satellites, TRUST, Beyond-Moore Computing
- **Support other national interests through other defense contractors**
  - Atomic Weapons Establishment
  - Air Force Research Laboratory
  - Air Force Institute of Technology
  - Naval Surface Warfare Center
  - Draper labs
- **Open source release**
  - Spring 2020
- **Engage larger community through technical interchange meetings & conferences**
  - HEART, SISPAD, NSREC, RADECS
- **Charon is the starting point for developing Strategically Radiation Hardened (SRH) electronic products**



What does Charon do?

- Drift-Diffusion PDE solver for modeling charge carrier flow

Electric Potential
$$\begin{cases} \nabla \cdot \left(\epsilon \vec{\mathbf{E}}\right) = q\,(p - n + C) \\ \vec{\mathbf{E}} = -\nabla V \end{cases}$$

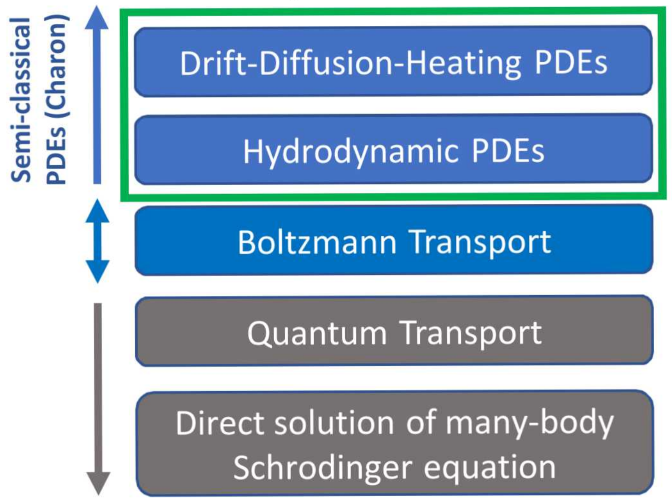
Constitutive Relations
$$\begin{cases} \vec{\mathbf{J}}_n = q\left(n\mu_n\vec{\mathbf{E}} + D_n\nabla n\right) \\ \vec{\mathbf{J}}_p = q\left(p\mu_p\vec{\mathbf{E}} - D_p\nabla p\right) \end{cases}$$

Conservation
$$\begin{cases} \nabla \cdot \vec{\mathbf{J}}_n - qR = q\frac{\partial n}{\partial t} \\ -\nabla \cdot \vec{\mathbf{J}}_p - qR = q\frac{\partial p}{\partial t} \end{cases}$$

Lattice Heating
$$\nabla \cdot (\kappa \nabla T_L) + H = \rho c \frac{\partial T_L}{\partial T}$$

TCAD code for modeling semiconductor performance including ionizing radiation and displacement damage as a result of radiation

Hierarchy of transport models

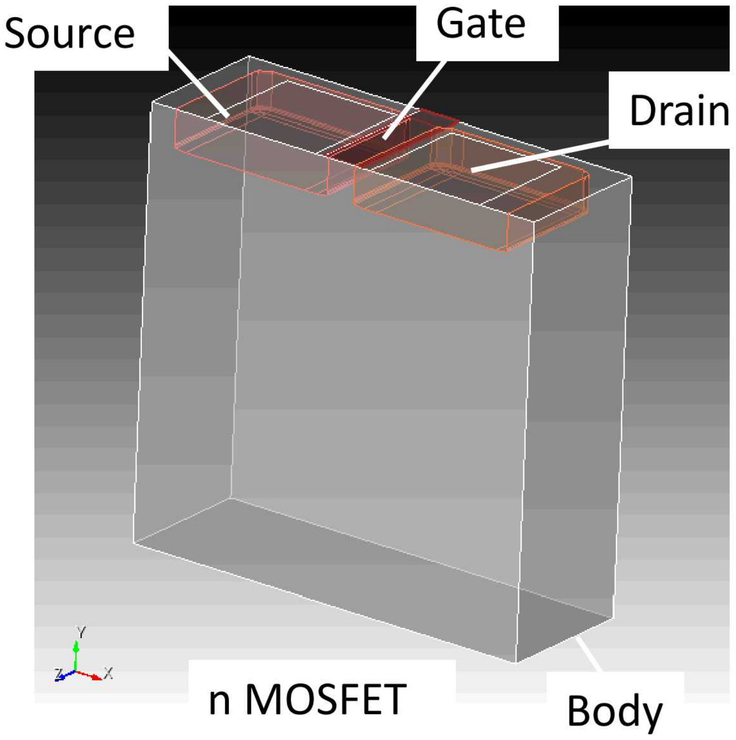


Capabilities Provided by Charon

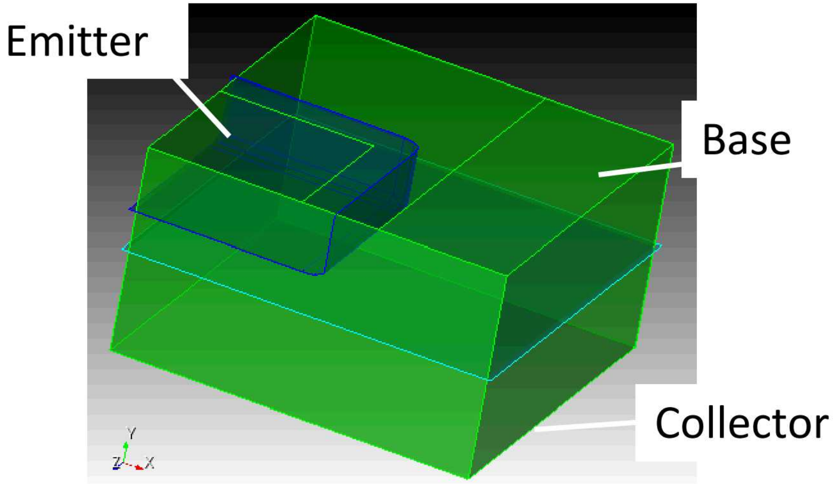
- Two & three dimensional + parallel capability
- General code models most common devices
  - Diodes
  - Bipolar junction transistors
  - Field effect transistors
- Models the effects of common radiation environments
  - Ionizing radiation (X-ray,γ-ray)
  - Total ionizing dose radiation
  - Neutron irradiation
- Some capability to model emerging devices and materials
  - APAM devbices
  - Memristors
  - III-V materials
    - Gallium Nitride
    - Gallium Arsende
    - Indium Gallium Phosphide
- Production quality code using current best practices for software development
  - Adheres to formal SQE practices
  - Incorporating agile development methods (scrum-ban)
- Utilizes latest computational technology
  - Solvers in Sandia’s Trilinos toolkit
  - Galerkin and Scharfetter-Gummel discretizations
  - Steady-state, time and frequency domain calculations
  - Next Generation Platforms (in process)

Nominal Devices

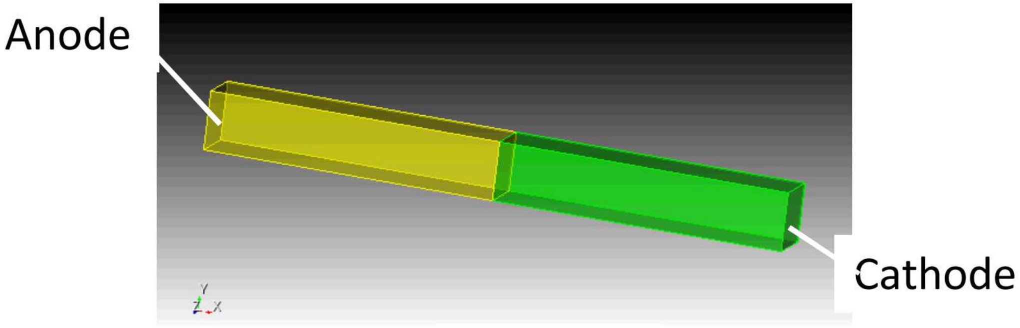
- Three devices selected for this presentation
- Each is nominal
  - Not based on any real commercial or Sandia device
- Each is commonly modeled by Sandia analysts
- Each is designed to be less numerically complicated than devices often are



n MOSFET



npn Bipolar Junction Transistor

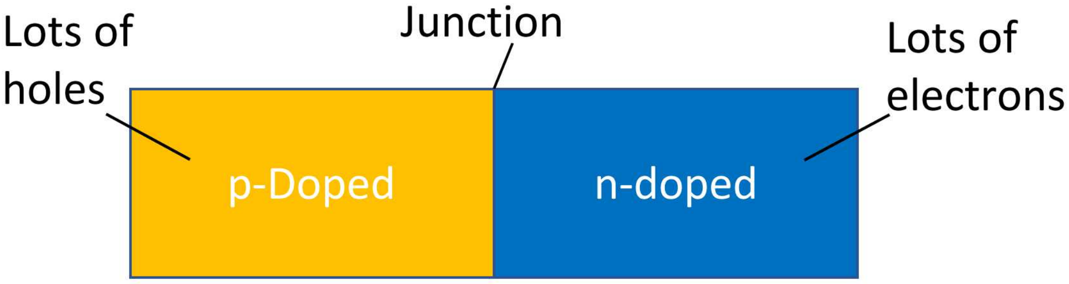


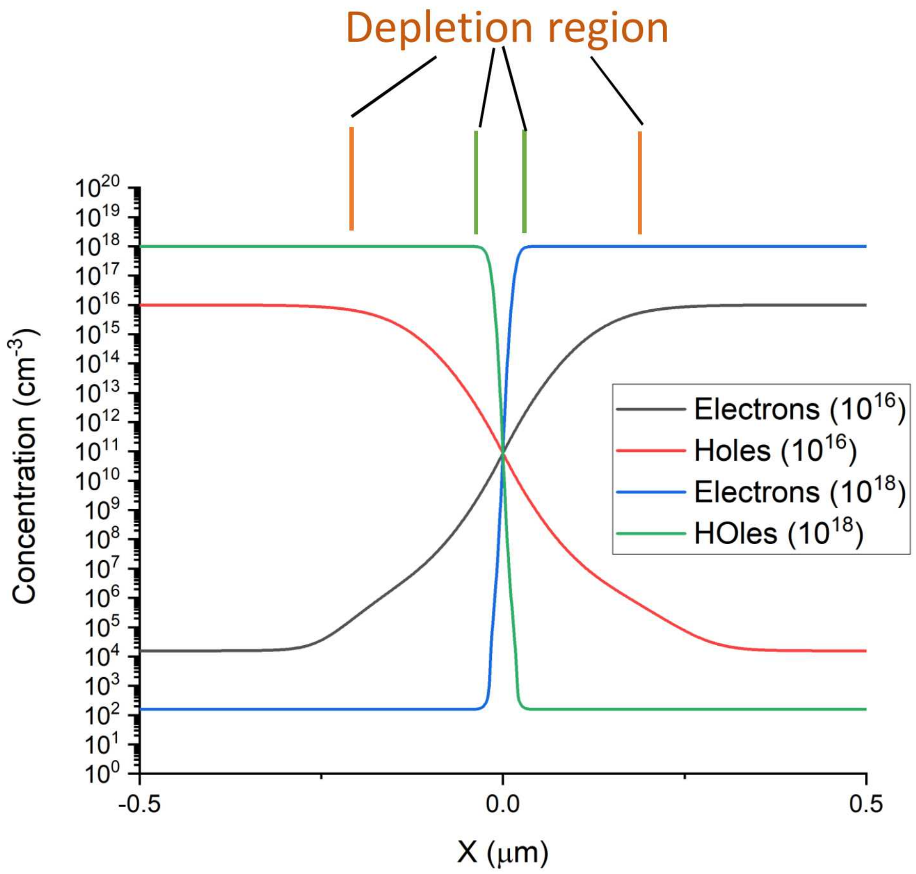
pn diode

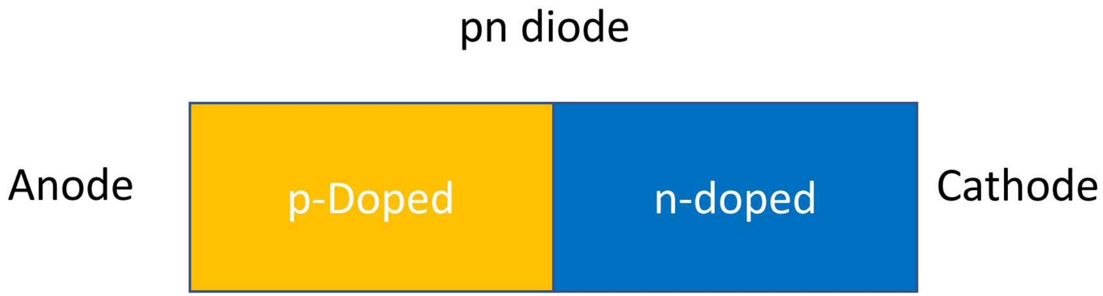
In these simulations…

- Discretization
  - SUPG-stabilized Galerkin finite element method
  - EFFPG finite element method
  - Scharfetter-Gummel finite volume method
- Matrix solution method
  - AztecOO GMRES with Ifpack ILU preconditioner
  - Belos GMRES with ML preconditioner
  - Currently in Trilinos/Epetra
  - Transitioning to Trilinos/Tpetra/Kokkos
- Timings
  - Averaged to the cost of a single Newton iteration
    - Regardless of discretization type, drift-diffusion equations exhibit strong mesh dependency
    - Linear solver iteration count can vary widely even during a single simulation
- All calculations done on SNL's Skybridge capacity cluster
  - 1,848 nodes
  - 16 cores per node
  - 2.6 GHz Intel Sandy Bridge

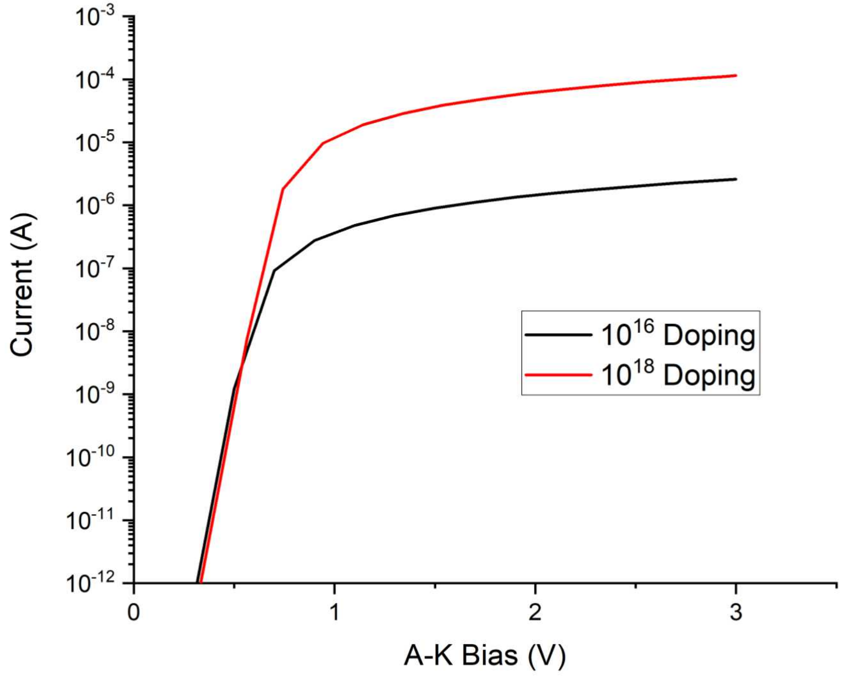
TCAD Junctions & Diode





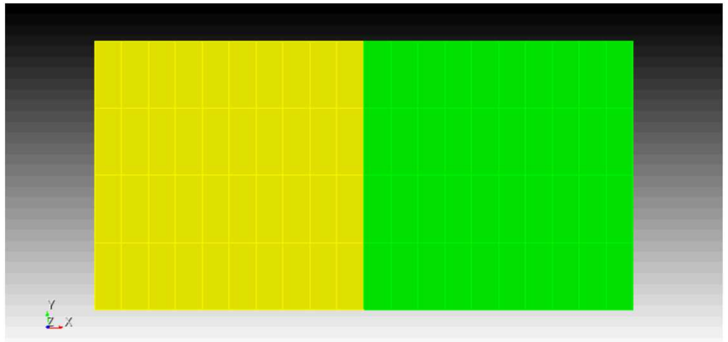


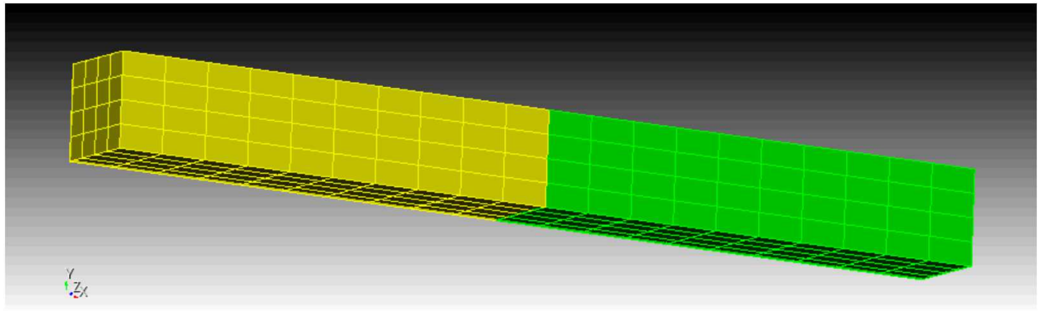
Elevating potential on the anode causes current to flow



Diode Mesh Setup


2D/1D

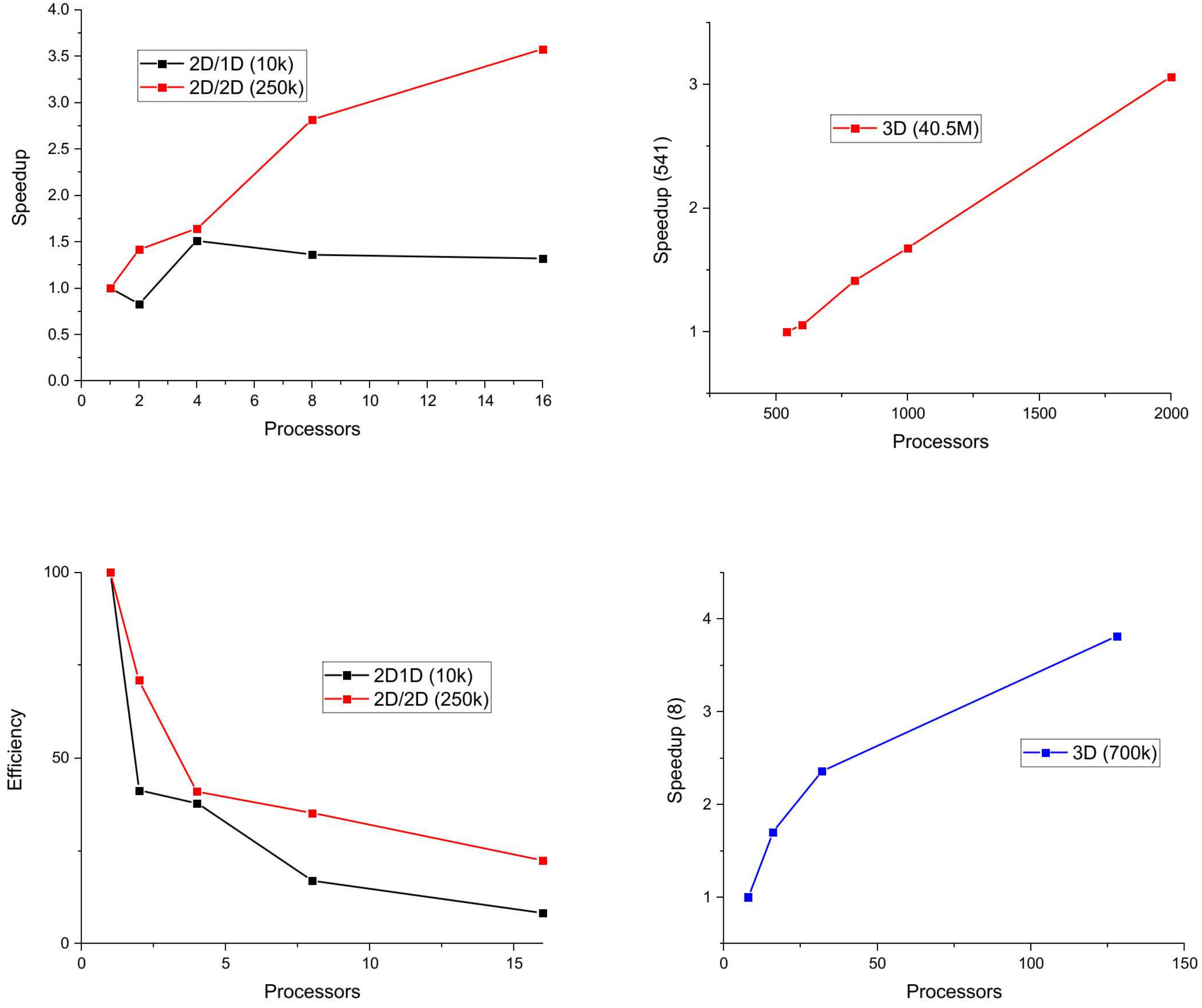

2D/2D


3D

- pn diode is a 1D device
  - Typically solved as quasi-1D
- Three types of meshes used for diode
  - All hex or quad meshing
  - 2D/1D-refined in flow direction only
  - 2D/2D-refined in flow and lateral directions
  - 3D-uniformly refined in 3D
- In this study, model is over-resolved
  - Wanted to see scaling in simplest possible configuration

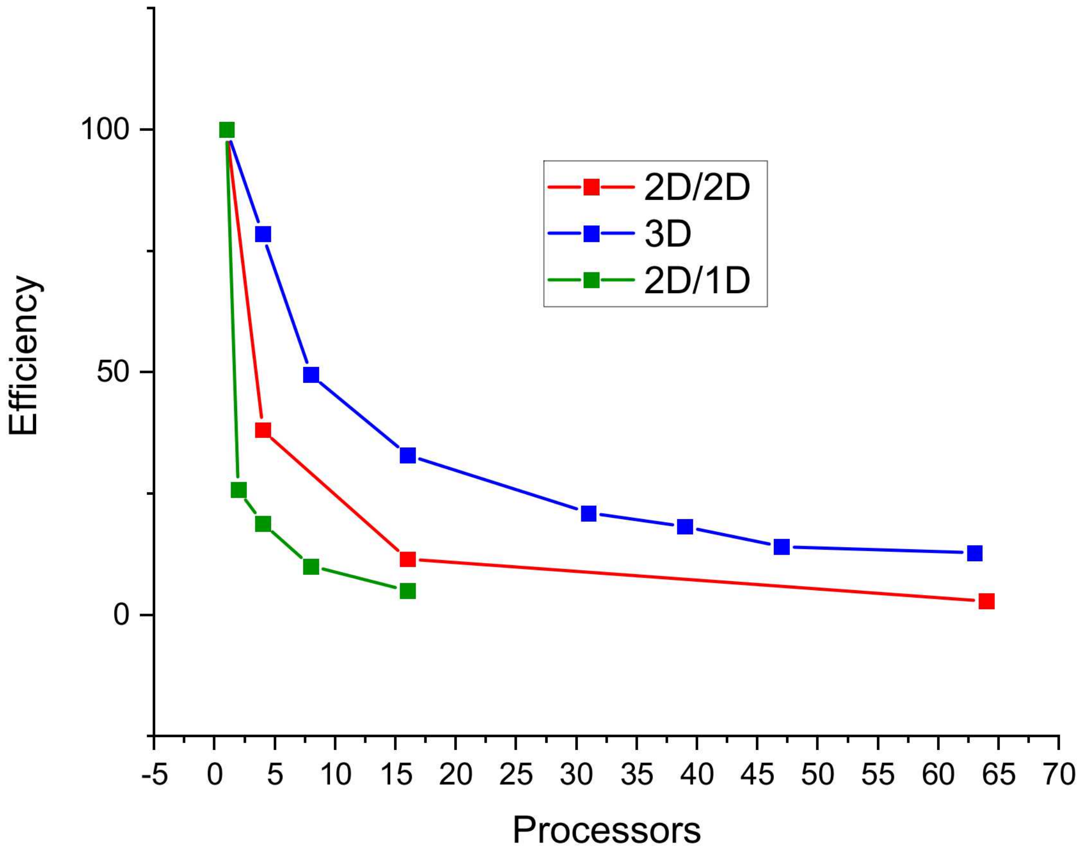
Diode Strong Scaling

- In General, 2D simulations scaled poorly
  - On node, no switch
  - 1D shares only 2 nodes at processor boundaries
- 3D simulations scale well
  - Resolved well beyond necessary

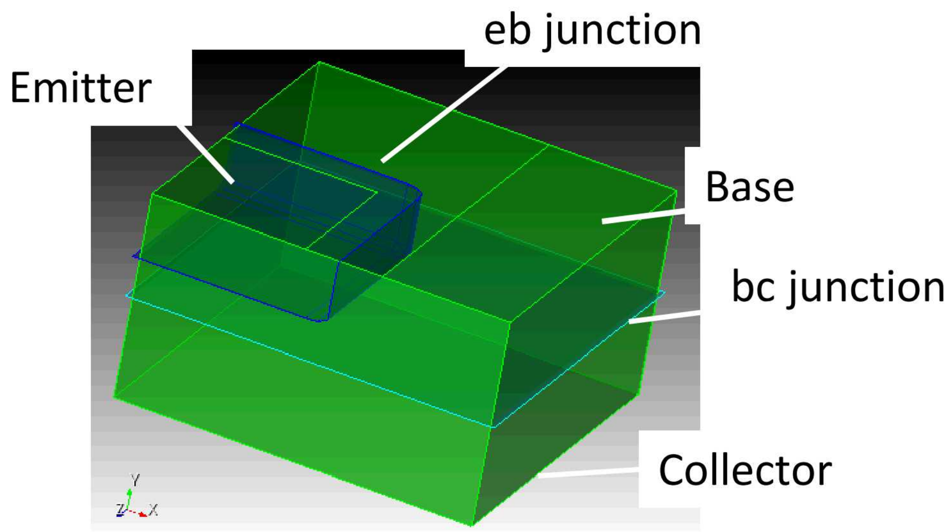


Diode Weak Scaling

- Weak scaling is poor for 2D or 3D
- 3D scaling starts from memory limit of a node—80k DoFs/processor
- Starts a theme that over-resolution in regions away from junctions may hinder weak scaling



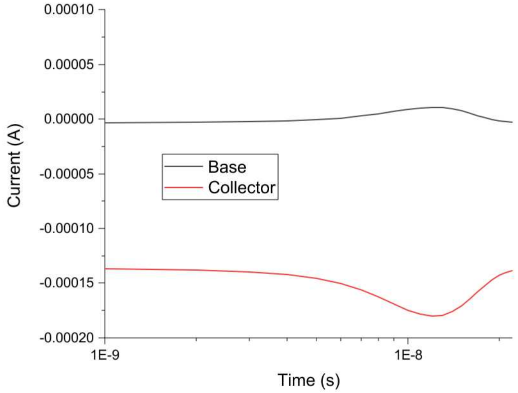
Bipolar Junction Transistor Setup

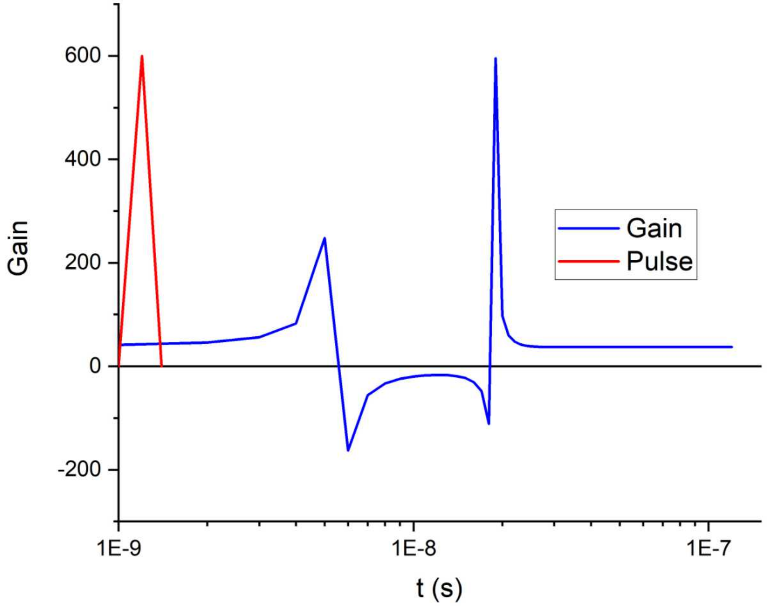


npn Bipolar Junction Transistor

- Three terminal device
  - Emitter, base, collector
- Common component of circuits
- Used for switching or amplification
- Historically, most studied with Charon
  - Neutron irradiation
  - Some dose rate (x-ray, γ-ray) radiation

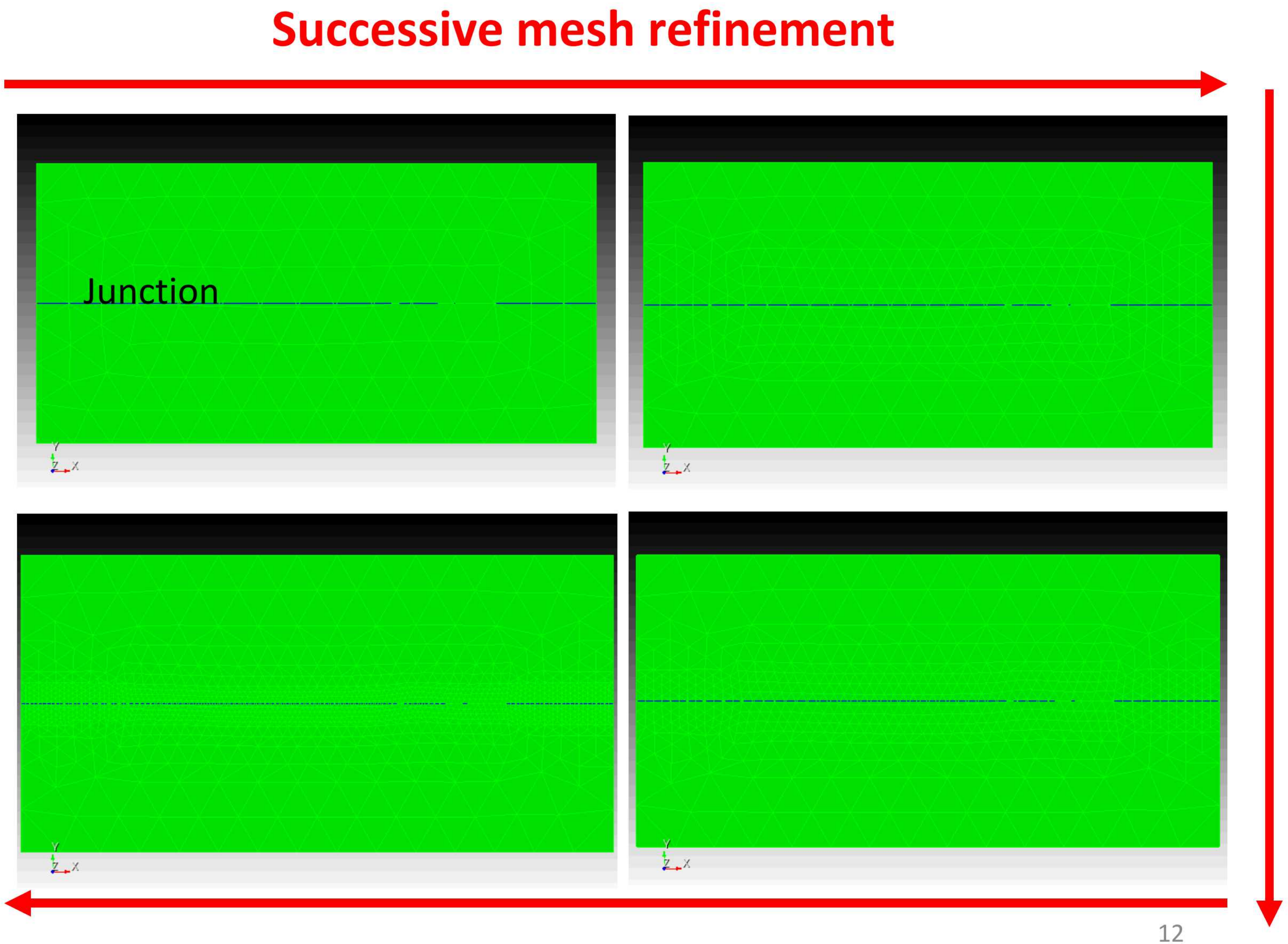
- Simple example of "made up" BJT under dose rate radiation
  - Electron-hole pairs produced in large quantities
  - Gain (I_c/I_b) evolves with carrier transport
- Gain changes dramatically and returns to normal after excess carriers dissipate





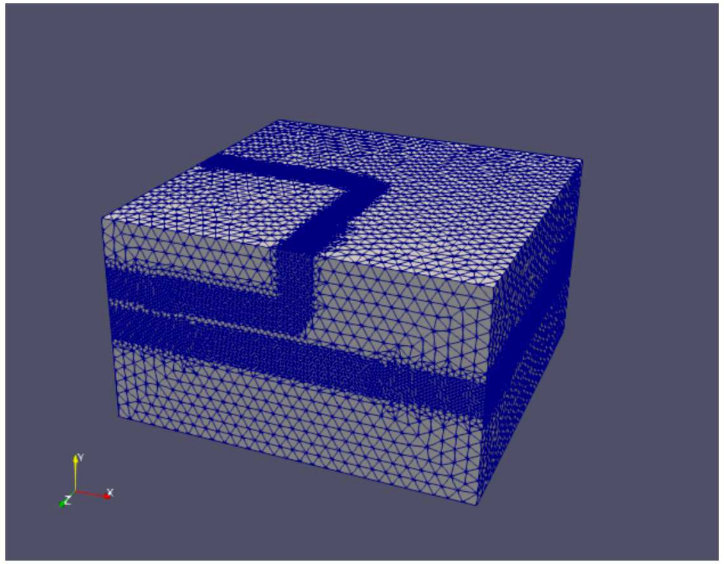
Automated Mesh Refinement

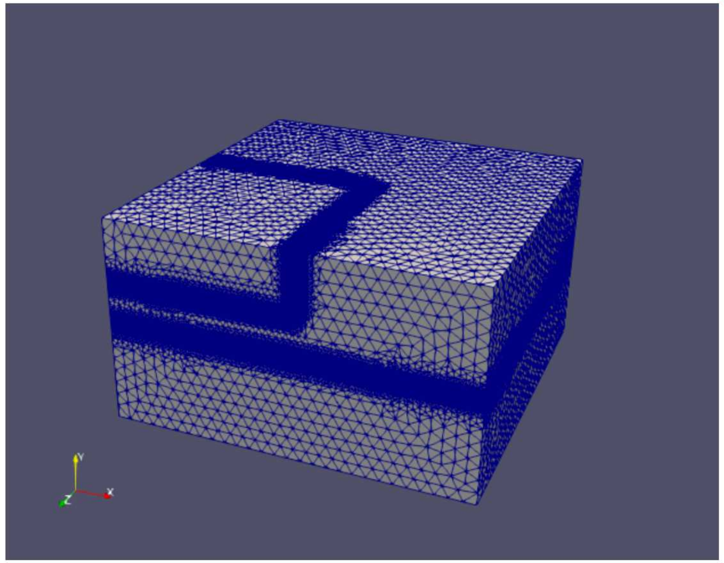
- Meshes for doping more complex than a diode problems must be refined around junctions
- Native Cubit was unable to produce meshes for complex-shaped junctions
- Charon pyMesh was created to address TCAD meshing needs
- Python based tool reads standard Cubit journal files plus special refinement directives.
- Tool creates a base mesh from Pythonized Cubit and then instructs Cubit which cells to refine.

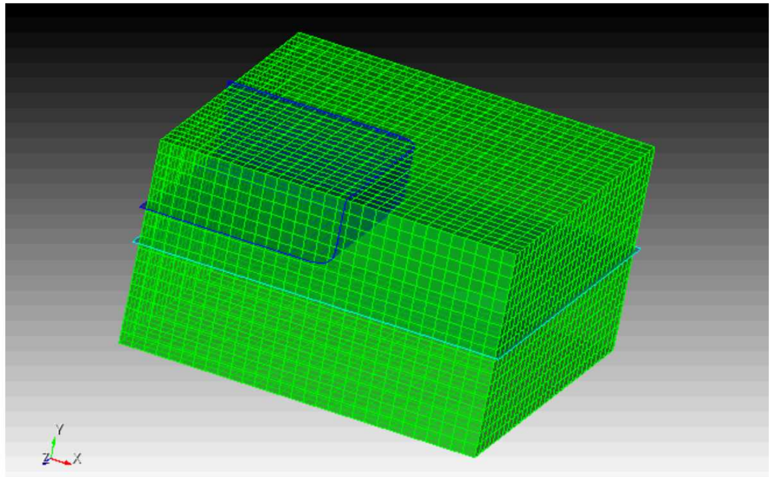


BJT Meshes

- Hexahedral meshes are uniform (almost)
  - Conform to boundaries and contacts
  - Must be excessively refined to resolve of junctions



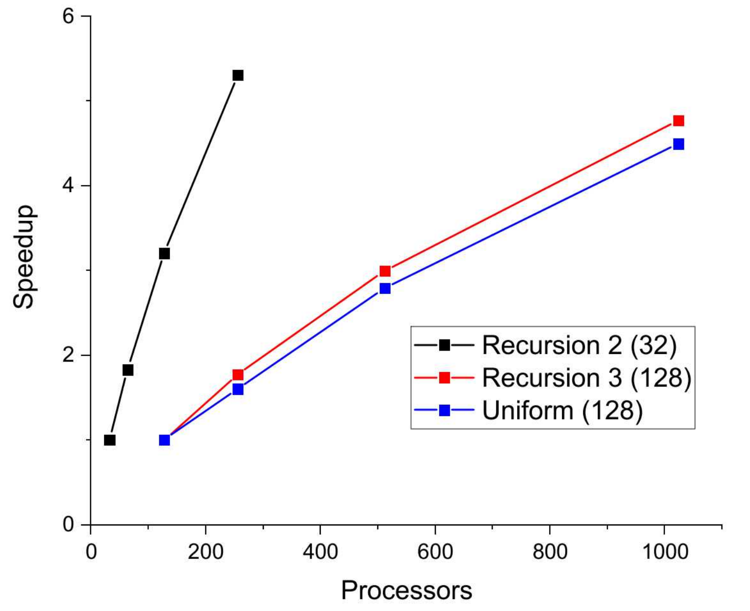


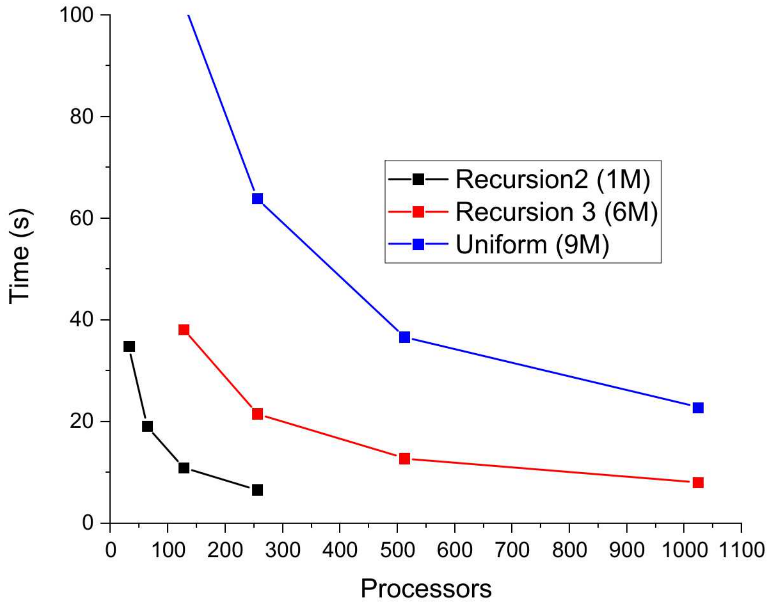


- Tetrahedral meshes are recursively refined
  - Conformal to boundaries and contacts
  - Base mesh of 1.5nm
  - Recursed 3 times to 1M DoFs
  - Recursed 4 times to 6M DoFs

BJT Strong Scaling

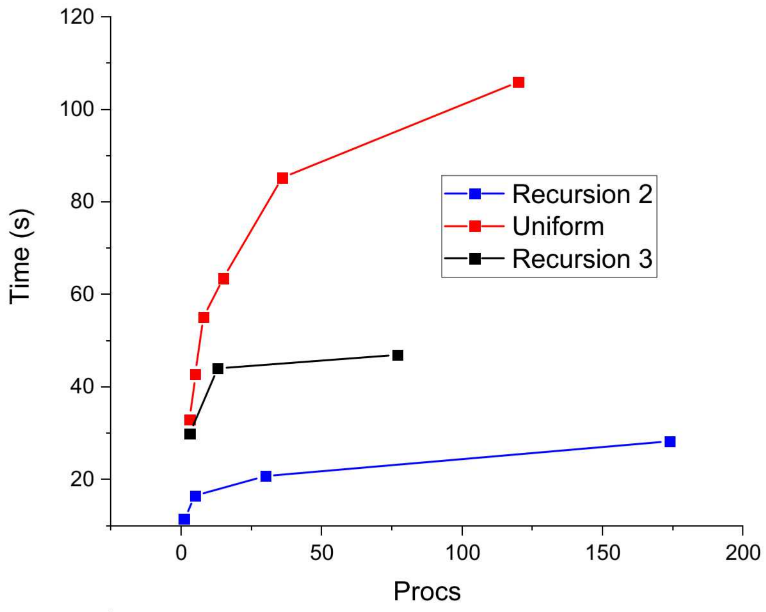
- Strong scaling of three different meshes—uniform, two recursively refined
- Starting point is near node memory limit
- All three performed about the same in terms of speedup relative to the fewest processors
  - ~5x speedup with an ideal of 8x

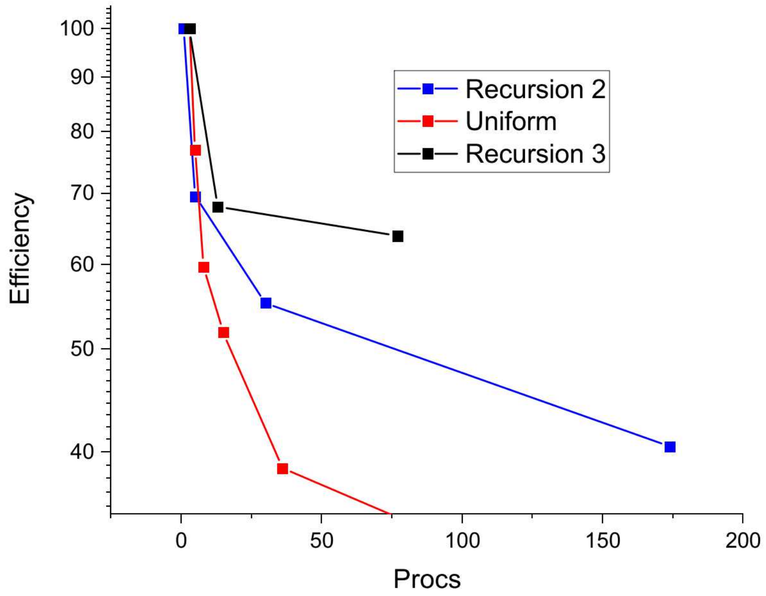




BJT Weak Scaling

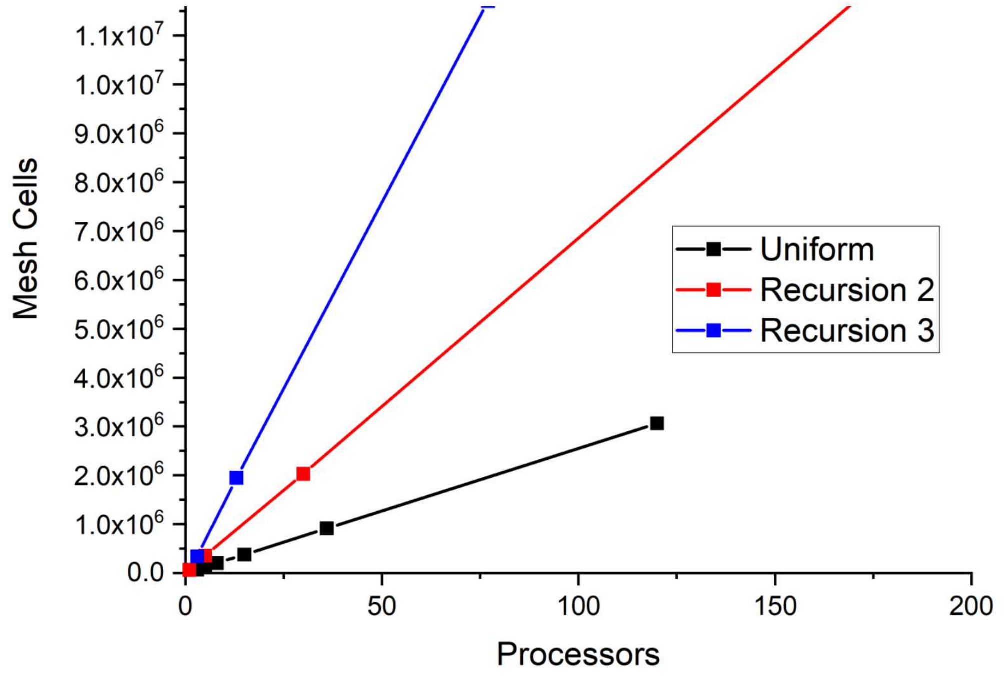
- Weak scaling performed across two meshing strategies
- Uniform meshes were uniformly refined
- Automated meshes start from different base meshes
  - Recursive refinement was held fixed across scaling
- Uniform meshes weak scale poorly
- Recursive meshes preform better

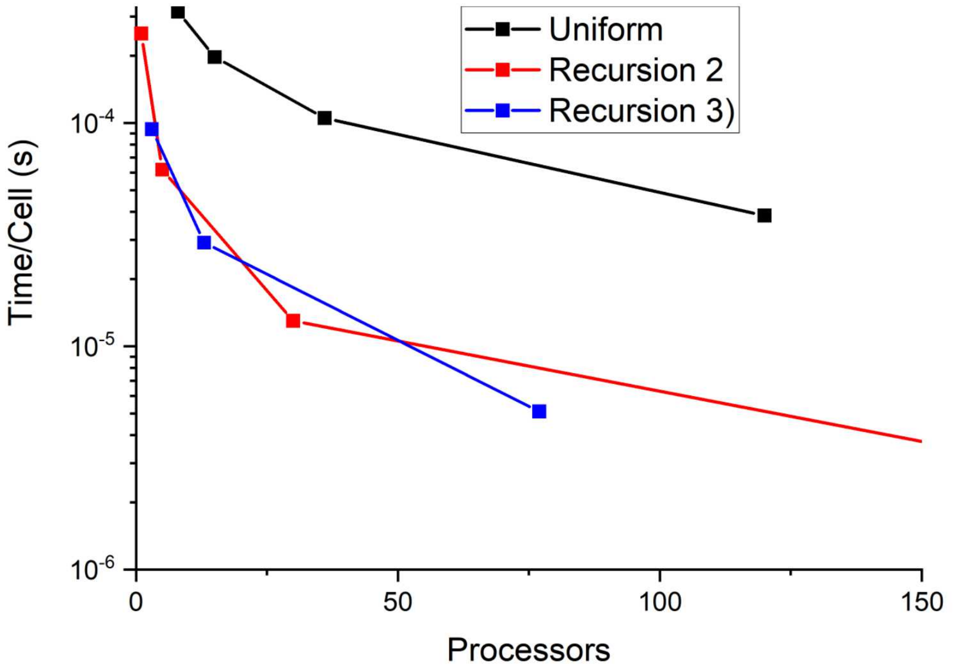




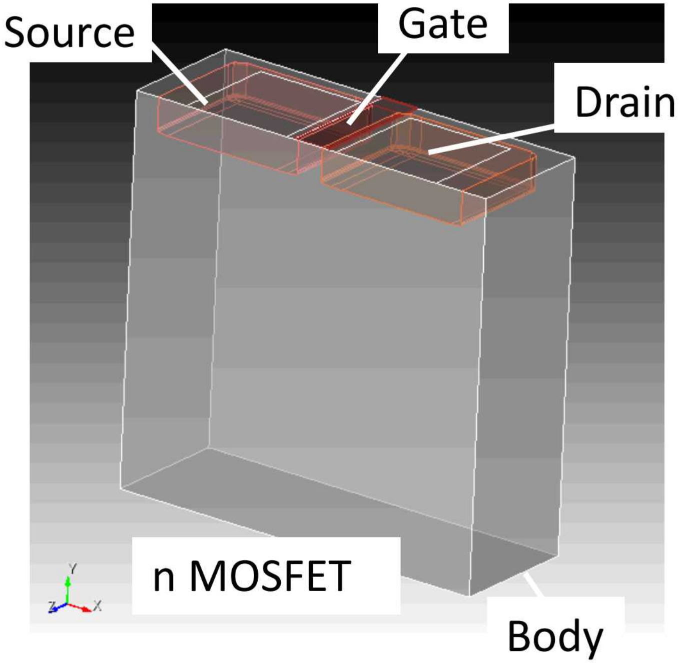
BJT Grind Time

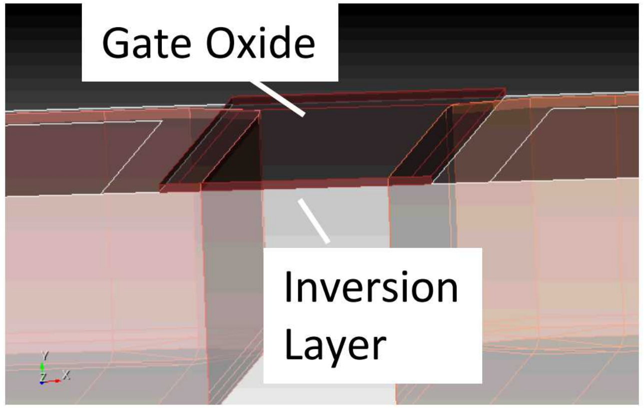
- Grind time examined across same weak scaling range
- Assembly time for hex meshes about 10% of overall solution time
  - Node/cell ratio nearly 1:1
  - Matrix condition number $O(10^8)$
- Assembly time for tet meshes 20%-25% of overall solution
  - Node/cell ratio 1:5
  - Matrix condition number $O(10^6)$



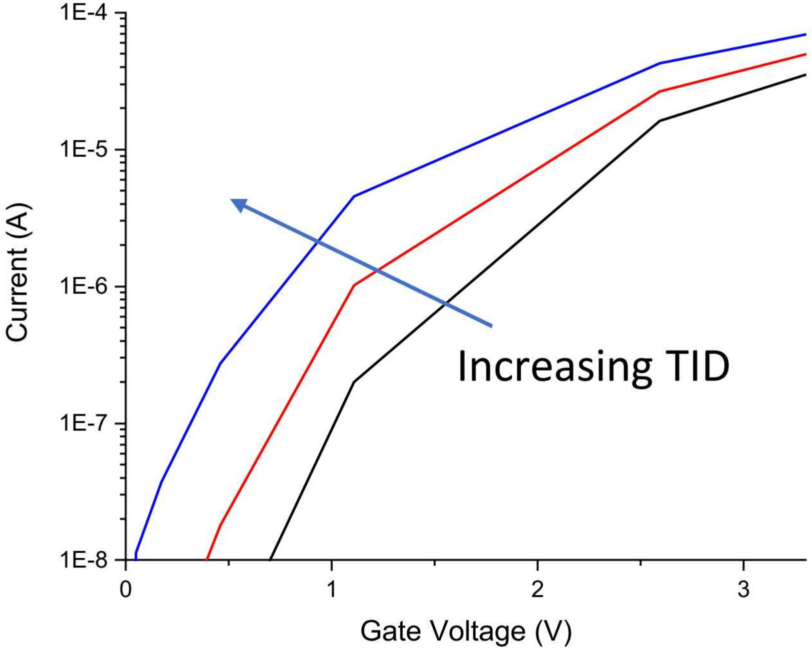


MOSFET Setup

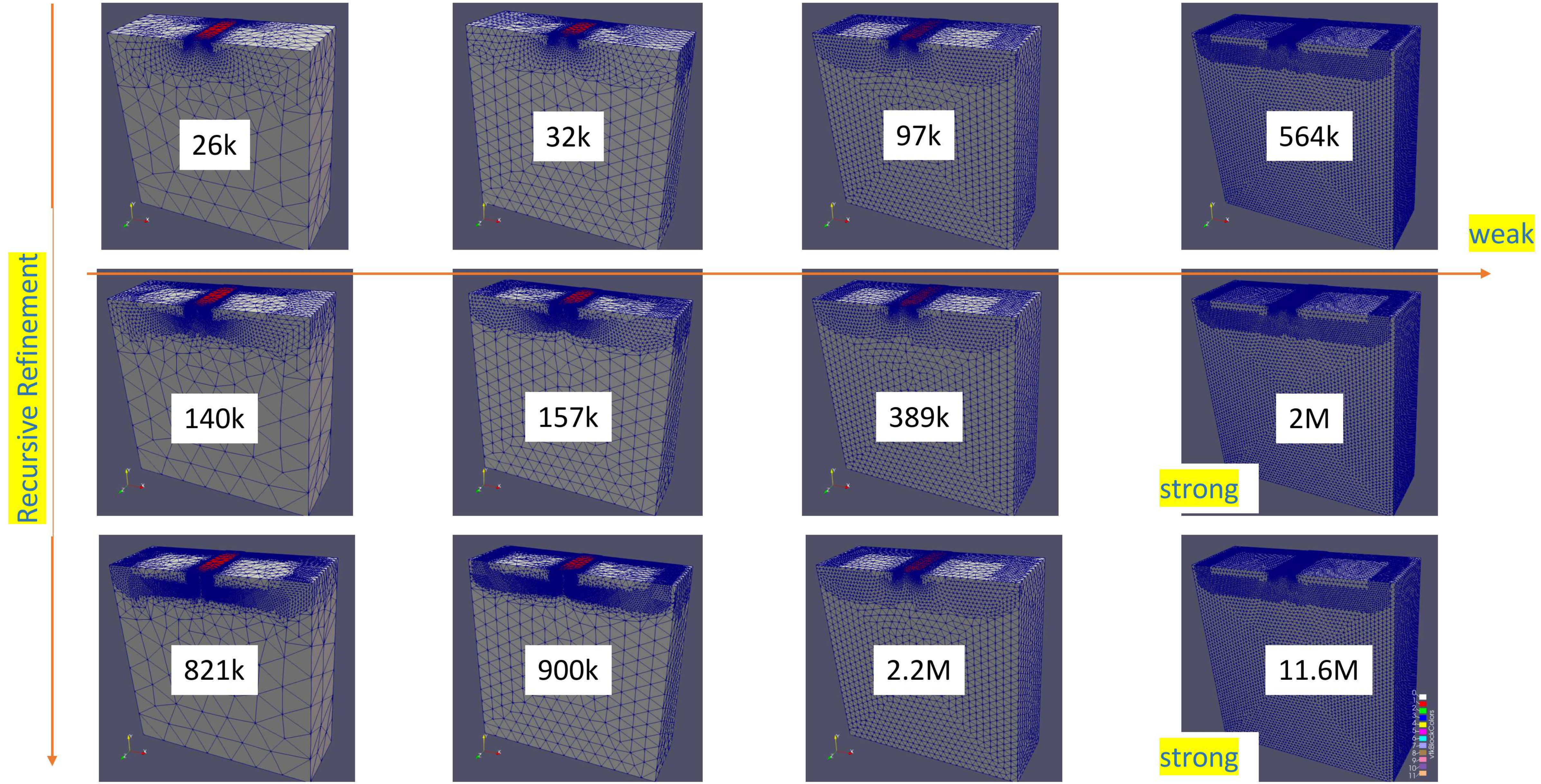




- Four terminal field effect transistor
- Bias is applied across source & drain
- No current flows until gate voltage increases above threshold—switching
- Total Ionizing Dose (TID) radiation is a chief concern
  - Causes charge buildup between gate oxide & semiconductor
  - Modifies threshold voltage
  - Enough radiation can leave device on permanently



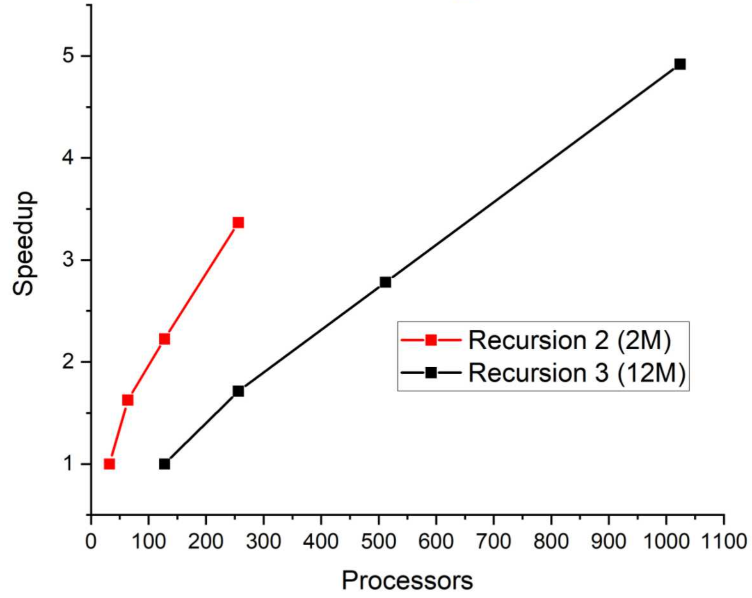
MOSFET Meshes

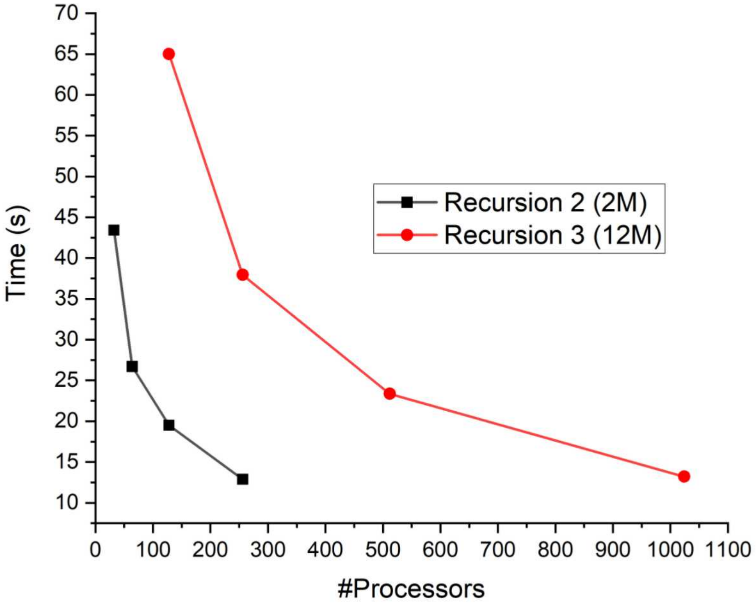


MOSFET Scaling

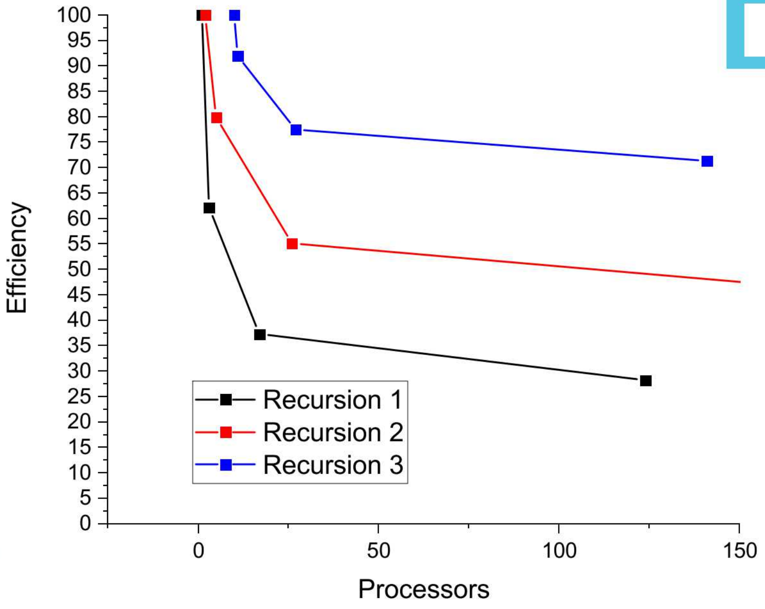
- Strong scaling across 2 & 3 recursions
    - Speedup relative to starting point 32 & 128 respectively
    - Recursion 2 achieved speedup of 3.5 with ideal of 8
    - Recursion 3 achieved speedup of 5 with ideal of 8
- Weak scaling across 1,2 & 3 recursions
    - Recursions held fixed with different base meshes
    - Scaling improves with higher levels of recursive refinement
    - Lower end performance poor—probably started too coarse

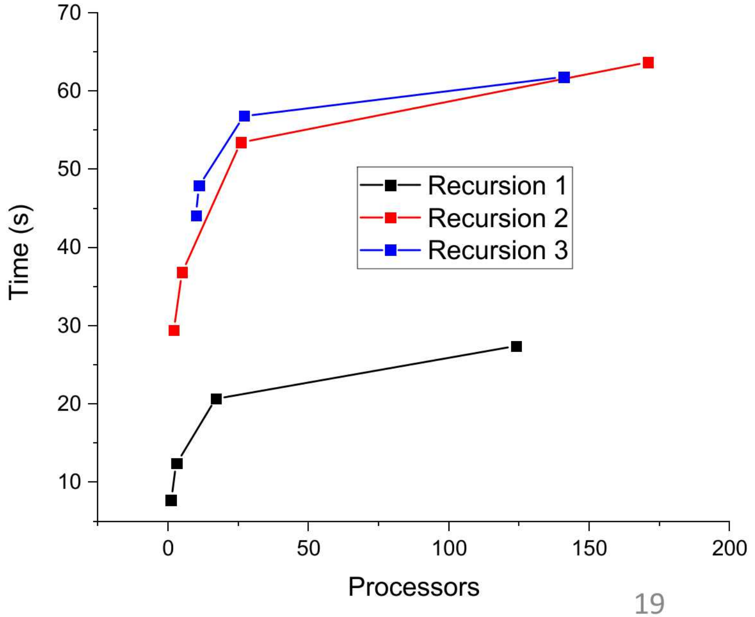
Strong





Weak





Wrap up

- Strong & weak scaling studies were done for three common devices Charon simulates
- Grind time examined across the week scaling spectrum for bipolar junction transistor
- Over-refinement in “quiescent” regions appears to hinder weak scaling—matrix condition numbers are consistently higher
- These will serve as benchmark as Charon transitions from Epetra to Tpetra/Kokkos