EECS 507: Introduction to Embedded Systems Research
Power, Energy, and Temperature

Robert Dick

University of Michigan

Outline

Deadlines and Announcements I

28 Sep.: J. Polastre, R. Szewczyk, A. Mainwaring, D. Culler, and J. Anderson, "Analysis of wireless sensor networks for habitat monitoring," in *Wireless Sensor Networks*, C. S. Raghavendra, K. M. Sivalingam, and T. Znati, Eds. Springer US, 2004, ch. 18, pp. 399–423.

4 Oct.: U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: An overview," *IEEE Communications Surveys and Tutorials*, vol. 19, no. 2, 2017.

6 Oct: "Research challenges for energy-efficient computing in automated vehicles," *IEEE Computer*, 2022, to appear. No student presentation for this one, but come prepared to discuss.

11 Oct.: E. Ronen, A. Shamir, A.-O. Weingarten, and C. O'Flynn, "IoT goes nuclear: Creating a ZigBee chain reaction," in *Proc. Symp. on Security and Privacy*, May 2017.

Deadlines and Announcements II

13 Oct.: Project checkpoint 1.

25 Oct.: Midterm exam.

Early December: Project presentations.

9 Dec.: Project deadline.

8am–10am 15 Dec.: Final exam.

Context

We just finished scheduling and RTOSs.

Will discuss L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. P. Dick, Z. M. Mao, and L. Yang, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones," in *Proc. Int. Conf. Hardware/Software Codesign and System Synthesis*, Oct. 2010, pp. 105–114.

Lecture on power, energy, and temperature.

Outline

General case

Many components.

Each may have many power management/activity states.

System-wide power consumption depends on the specific combination of component states.

How many samples?

- 10 components.
- 5 states, each.
- $5^{10} \simeq$ 10-million system-wide states.

How to get enough samples to characterize?

Independence assumption for embedded system power modeling

What if a component's power consumption were mostly independent of the power management/activity states of other components?

How many samples?

- 10 components.
- 5 states, each.
- 50 samples of interest.

The assumption is often correct.

When it is not, can treat the two interdependent components as a single component.

Practical embedded system power estimation

1. For each component.
   1. Put all other components in lowest power state.
   2. Measure component power consumption in each state.
2. Can manually use measurements to build expression for system-wide power consumption.
3. Also works for incomplete sampling by using linear regression to find the relationship between each state variable and the system-wide power consumption.

Applying the power model

Estimate/measure the proportion of time each component spends in each state.

Sum the products of time proportions and component–state power consumptions to get system-wide average power consumption.

This is often inaccurate for instantaneous power consumption.

Not good for power supply provisioning or thermal design.

Often very accurate over timescales of minutes.

Good for battery lifespan estimation.

Outline

Definitions

Temperature: Average kinetic energy of particle.

Heat flow: Transfer of this energy.

Heat always flows from regions of higher temperature to regions of lower temperature.

Particles move.

What happens to a moving particle in a lattice?

Acoustic phonons

- Lattice structure.
- Transverse and longitudinal waves.
- Electron–phonon interactions.
- Effect of carrier energy increasing beyond optic phonon energy?

Optic phonons

- Only occur in lattices with more than one atom per unit cell.
- Optic phonons out of phase from primitive cell to primitive cell.
- Positive and negative ions swing against each other.
- Low group velocity.
- Interact with electrons.

Nanostructure heat transfer

Boundary scattering.

Quantum effects when phonon spectra of materials do not match.

Why do wires get hot?

Scattering of electrons due to destructive interference with waves in the lattice.

What are these waves?

What happens to the energy of these electrons?

What happens when wires start very, very cool?

What is electrical resistance?

What is thermal resistance?

Why do transistors get hot?

Scattering of electrons due to destructive interference with waves in the lattice.

Where do these waves come from?

Where do the electrons come from?

- Intrinsic carriers.
- Dopants.

What happens as the semiconductor heats up?

- Carrier concentration increases.
- Carrier mobility decreases.
- Threshold voltage decreases.

Power consumption trends

- Initial optimization at transistor level.
- Further research-driven gains at this level difficult.
- Research moved to higher levels, e.g., RTL.
- Trade area for performance and performance for power.
- Clock frequency gains linear.
- Voltage scaling ${V_{DD}}^2$ – important.

Power consumption in synchronous CMOS

$$P = P_{SWITCH} + P_{SHORT} + P_{LEAK}$$

$$P_{SWITCH} = C \cdot {V_{DD}}^2 \cdot f \cdot A$$

$$\dagger\ P_{SHORT} = \frac{b}{12}(V_{DD} - 2 \cdot V_T)^3 \cdot f \cdot A \cdot t$$

$$P_{LEAK} = V_{DD} \cdot (I_{SUB} + I_{GATE} + I_{JUNCTION} + I_{GIDL})$$

C : total switched capacitance

V_{DD} : high voltage

f : switching frequency

A : switching activity

b : MOS transistor gain

V_T : threshold voltage

t : rise/fall time of inputs

† P_{SHORT} usually $\leq 10\%$ of P_{SWITCH}

Smaller as $V_{DD} \rightarrow V_T$

Adiabatic charging

Voltage step function implies $E = C{V_{CAP}}^2/2$.

Instead, vary voltage to hold current constant: $E = C{V_{CAP}}^2 \cdot RC/t$.

Lower energy if $T > 2RC$.

Impractical when leakage significant.

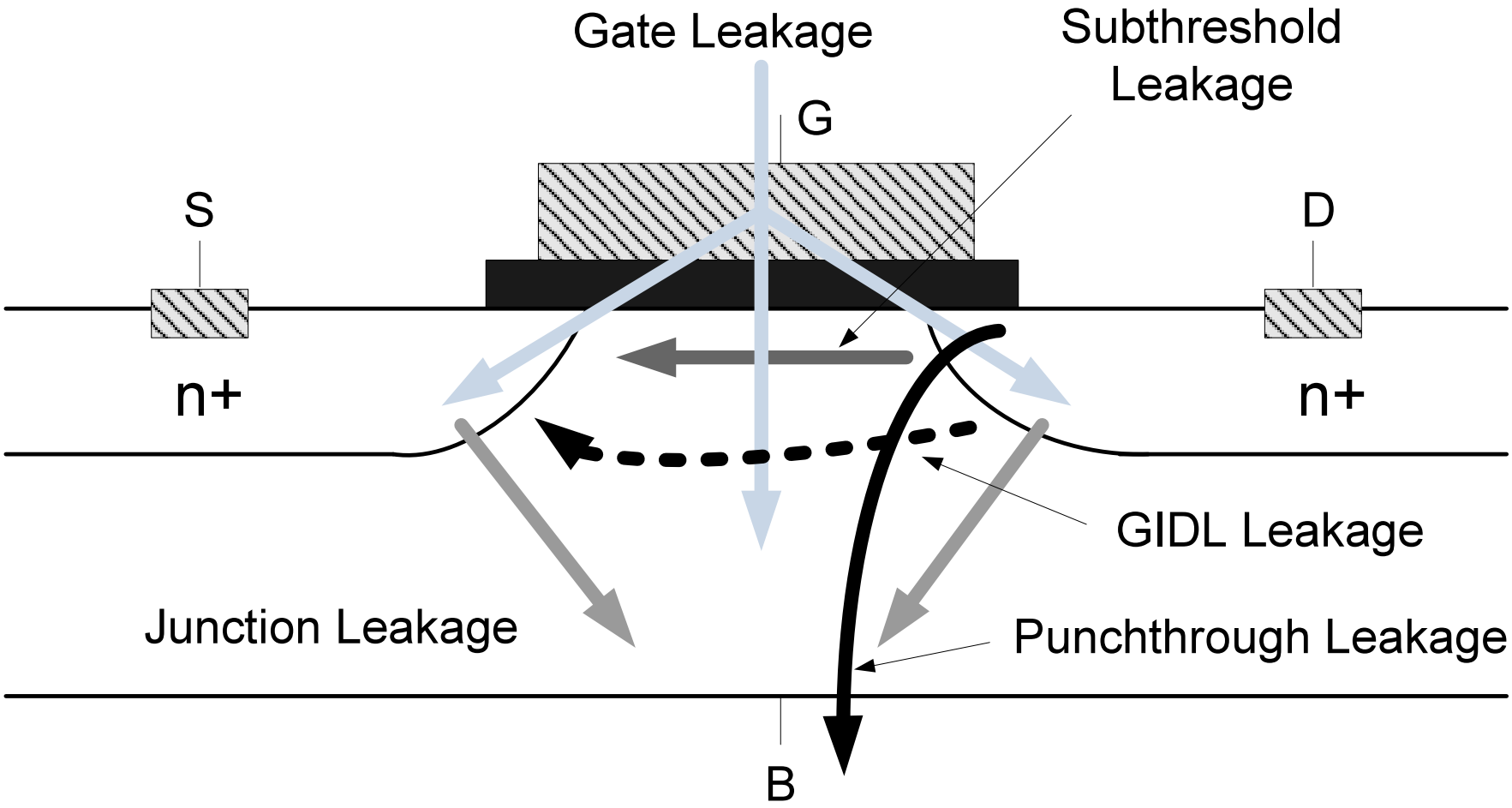
Wiring power consumption

In the past, transistor power $\gg$ wiring power.

Process scaling $\Rightarrow$ ratio changing.

Conventional CAD tools neglect wiring power.

Leakage


Gate Leakage
Subthreshold Leakage
G
S
D
n+
n+
GIDL Leakage
Junction Leakage
Punchthrough Leakage
B

Subthreshold leakage current

$$I_{subthreshold} = A_s \frac{W}{L} {v_T}^2 \left(1 - e^{\frac{-V_{DS}}{v_T}}\right) e^{\frac{(V_{GS} - V_{th})}{n v_T}},$$

where A_s is a technology-dependent constant,

V_{th} is the threshold voltage,

L and W are the device effective channel length and width,

V_{GS} is the gate-to-source voltage,

n is the subthreshold swing coefficient for the transistor,

V_{DS} is the drain-to-source voltage, and

v_T is the thermal voltage.

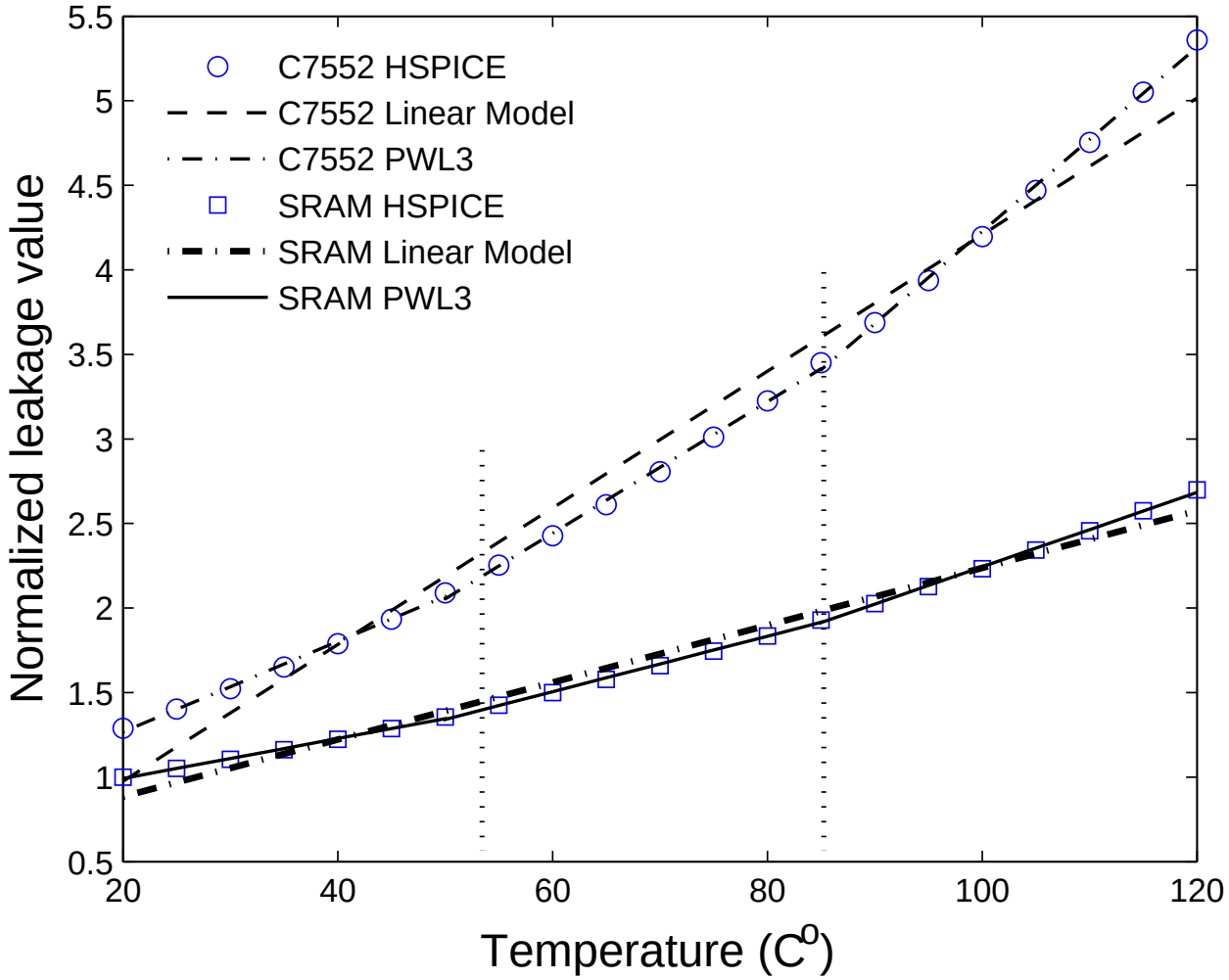
A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. IEEE Press, 2001

Simplified subthreshold leakage current

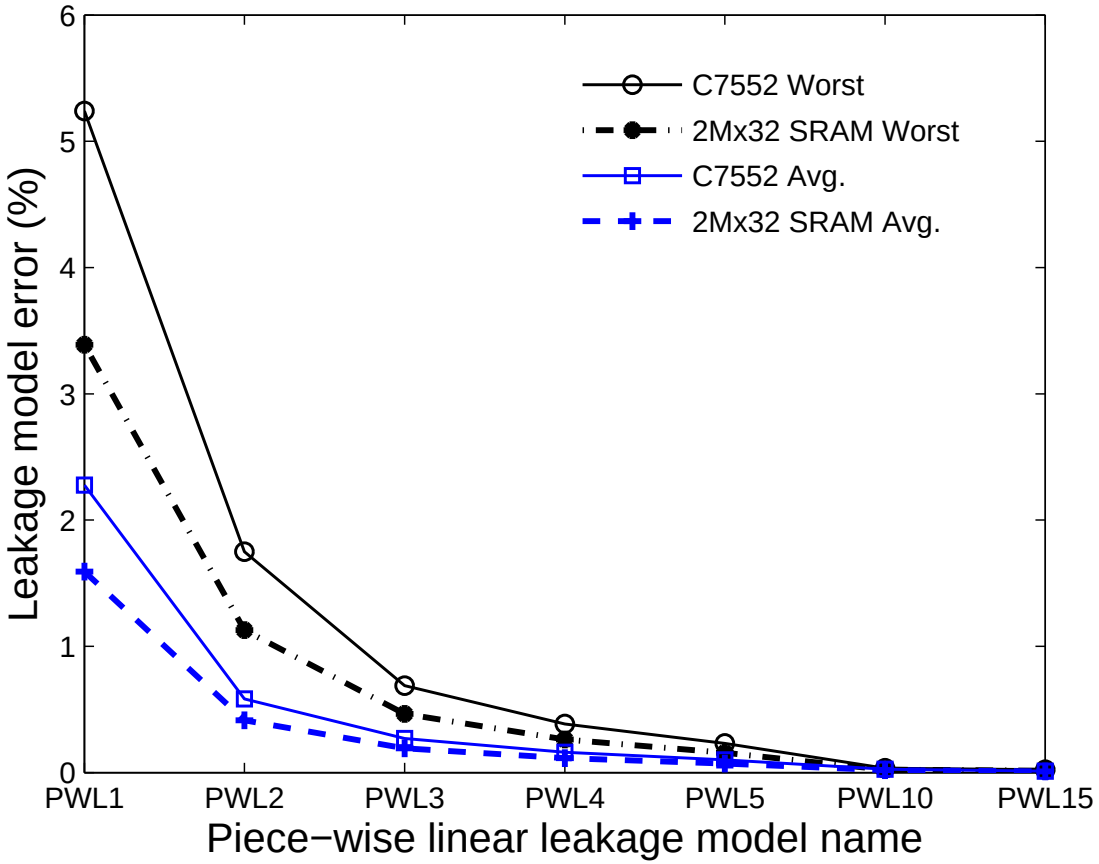
$V_{DS} \gg v_T$ and $v_T = \frac{kT}{q}$. q is the charge of an electron. Therefore, equation can be simplified to

$$I_{subthreshold} = A_s \frac{W}{L} \left(\frac{kT}{q} \right)^2 e^{\frac{q(V_{GS} - V_{th})}{nkT}} \quad (1)$$

Exponential?



Piece-wise linear error



Y. Liu, R. P. Dick, L. Shang, and H. Yang, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *Proc. Design, Automation & Test in Europe Conf.*, Mar. 2007, pp. 1526–1531

Gate leakage

Caused by tunneling between gate and other terminals.

$$I_{gate} = WLA_J \left(\frac{T_{oxr}}{T_{ox}} \right)^{nt} \frac{V_g V_{aux}}{T_{ox}^2} e^{-BT_{ox}(a-b|V_{ox}|)(1+c|V_{ox}|)}$$

where A_J, B, a, b, and c are technology-dependent constants,

nt is a fitting parameter with a default value of one,

V_{ox} is the voltage across gate dielectric,

T_{ox} is gate dielectric thickness,

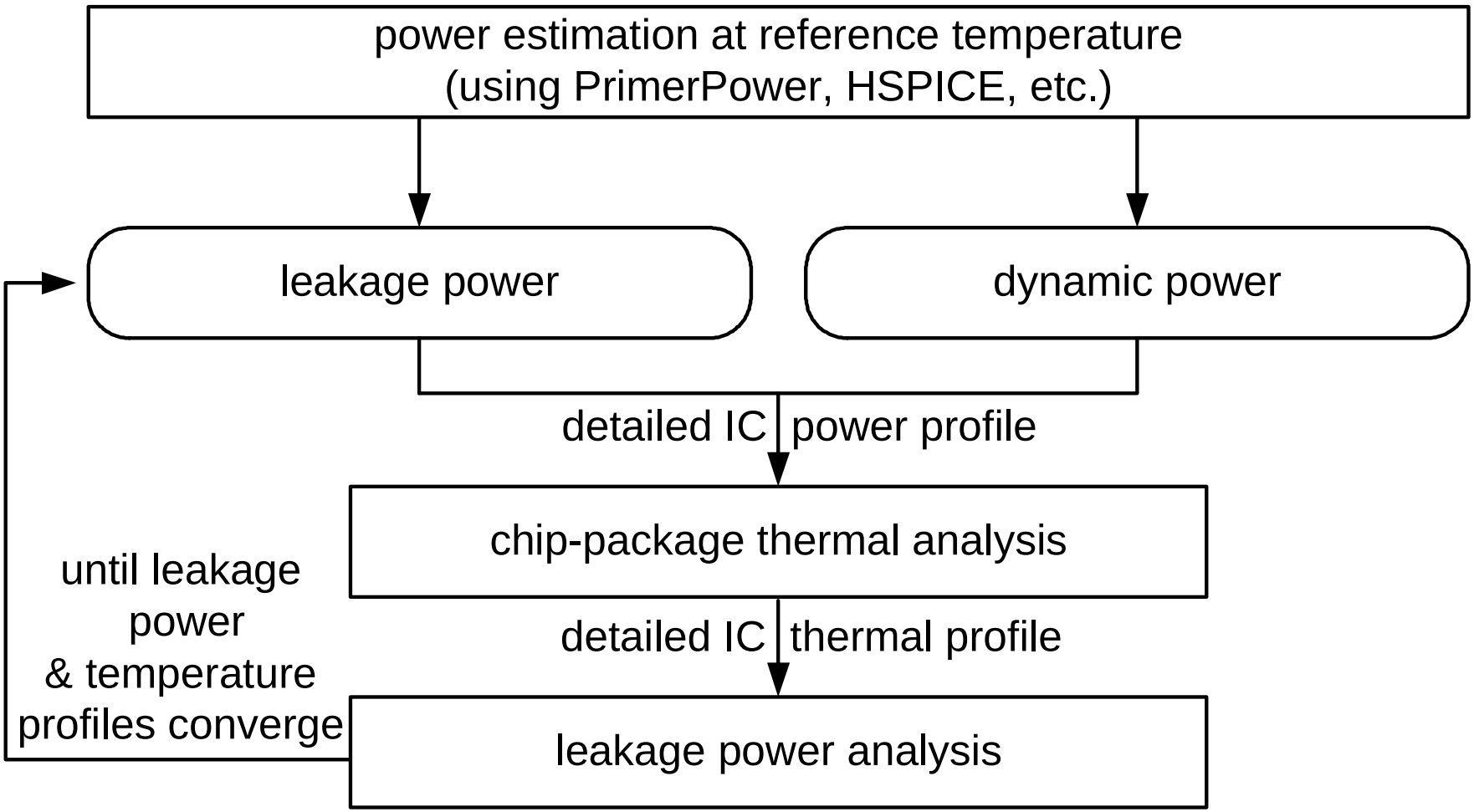
T_{oxr} is the reference oxide thickness,

V_{aux} is an auxiliary function that approximates the density of tunneling carriers and available states, and

V_g is the gate voltage.

K. M. Cao, W. C. Lee, W. Liu, X. Jin, P. Su, S. K. H. Fung, J. X. An, B. Yu, and C. Hu, "BSIM4 gate leakage model including source-drain partition," in *IEDM Technology Dig.*, Dec. 2000, pp. 815–818

Temperature-aware leakage estimation



Power consumption conclusions

Voltage scaling is currently the most promising low-level power-reduction method: V^2 dependence.

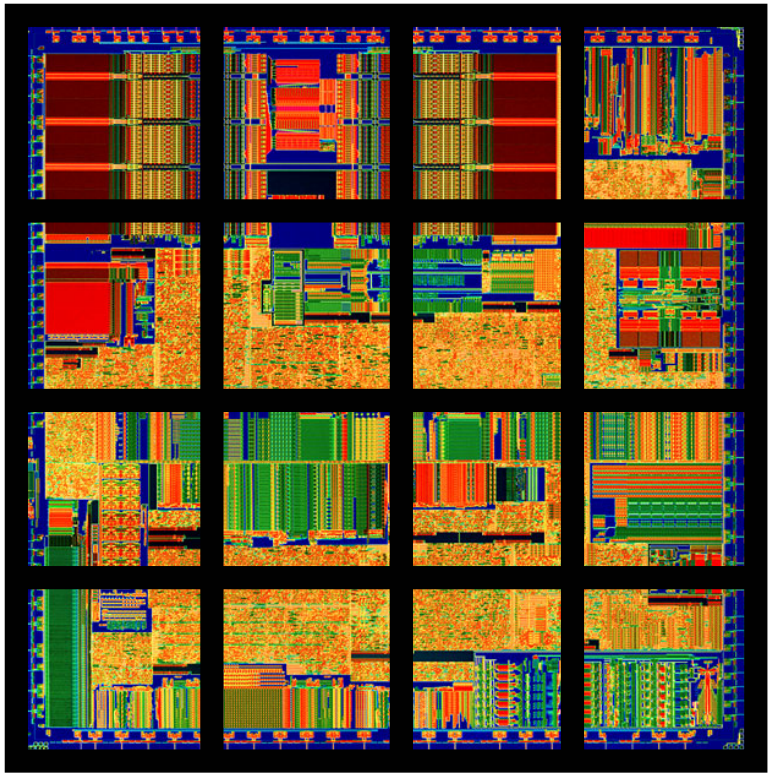
As V_{DD} reduced, V_T must also be reduced.

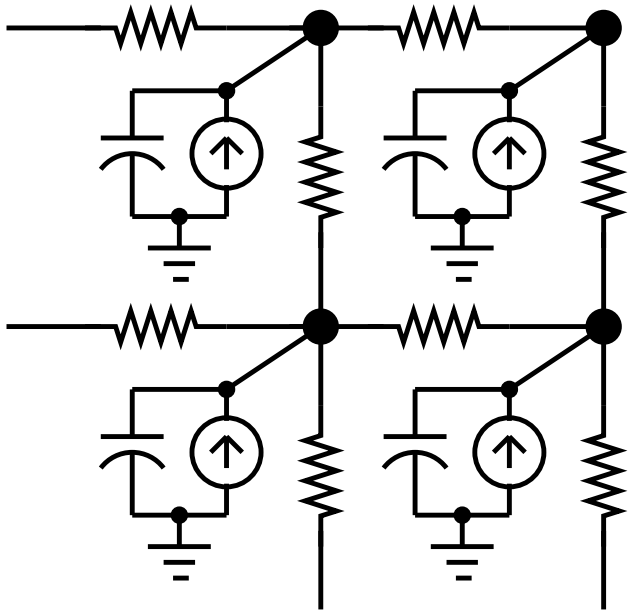
Sub-threshold leakage becomes significant.

What happens if $P_{LEAK} > P_{SWITCH}$?

Outline

R(C) model





Partition into 3-D elements (diagram 2-D for simplicity)

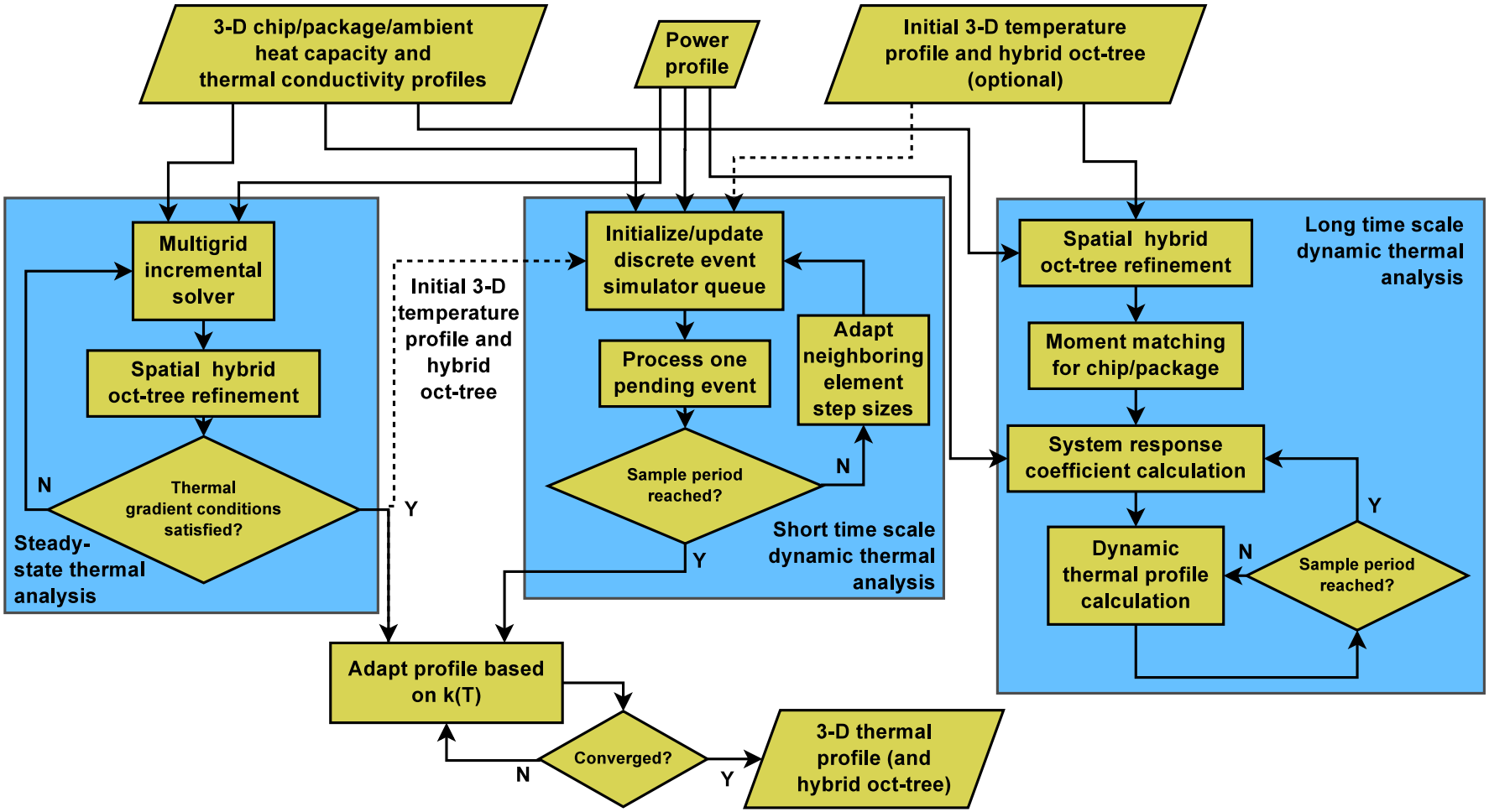
Thermal resistance	$\leftrightarrow$	Resistance
Heat flow	$\leftrightarrow$	Current
For dynamic: Heat capacity	$\leftrightarrow$	Capacitance

Problem definition

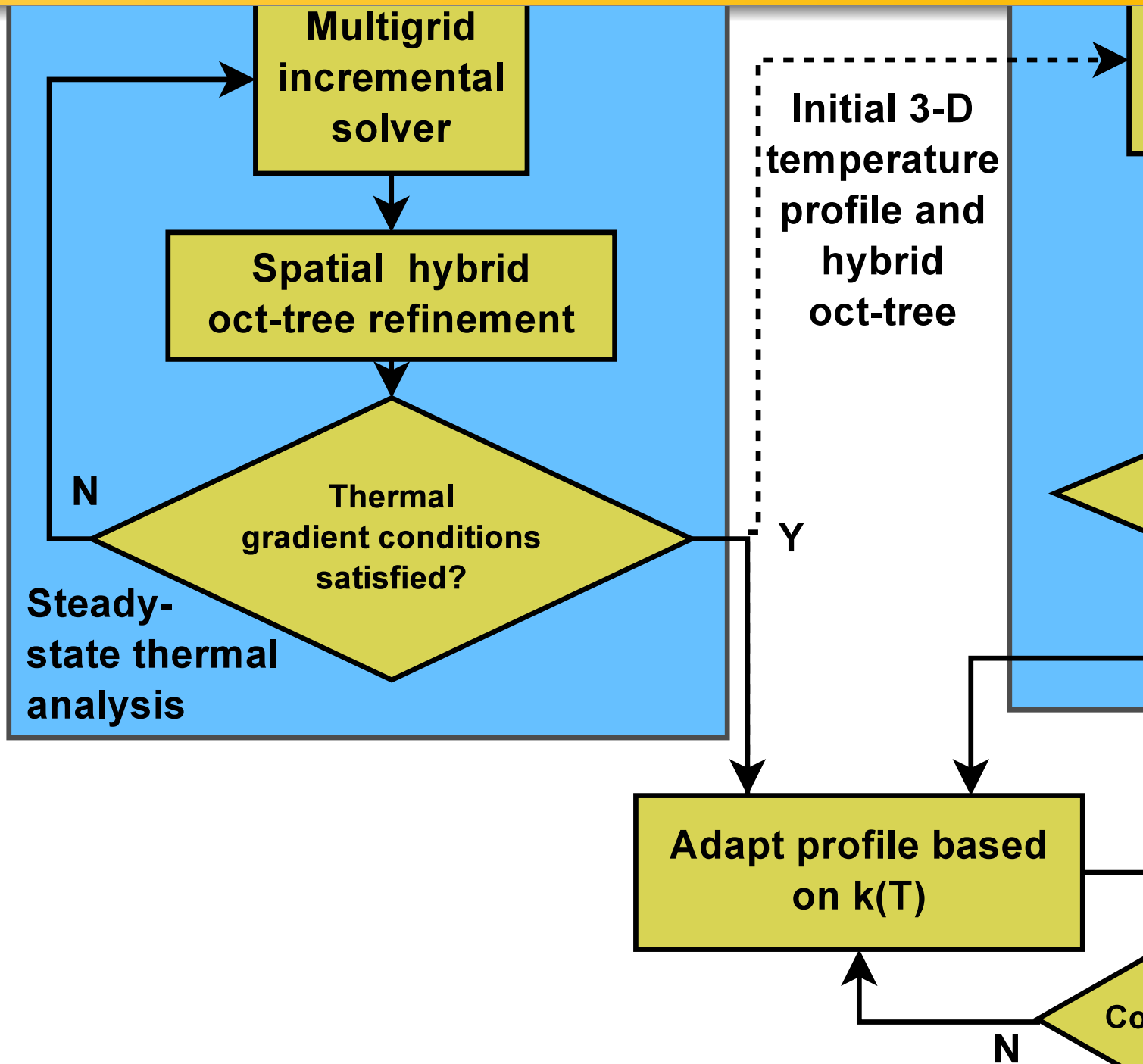
$$\mathbf{C}\frac{d\mathbf{T}(t)}{dt} = \mathbf{A}\mathbf{T}(t) + \mathbf{P}U(t)$$

- **A** is the thermal conductivity matrix
- Steady-state: Initial temperature and **C** unnecessary
- Dynamic: Transient temperature analysis, must also consider heat capacity

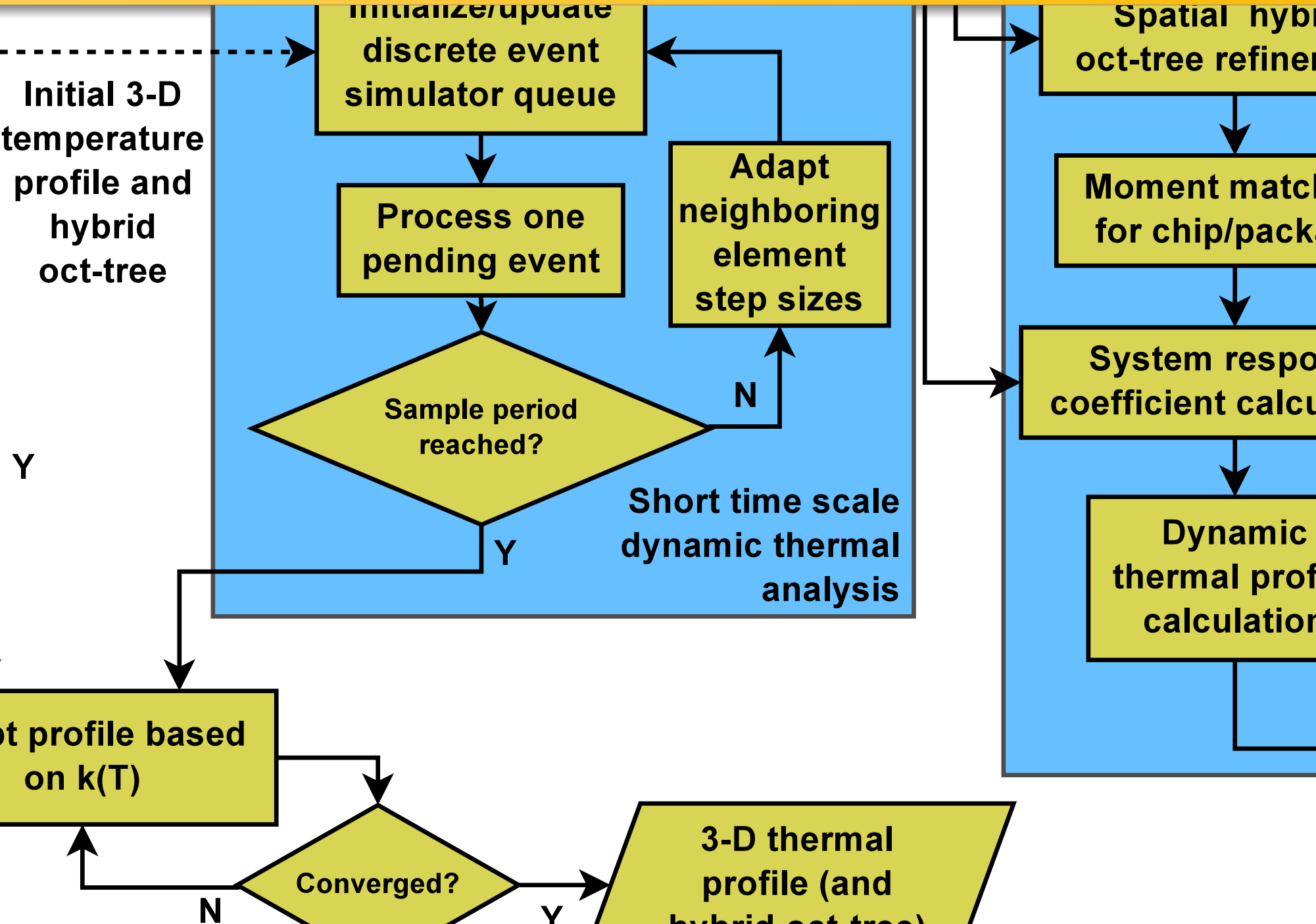
Thermal analysis infrastructure overview



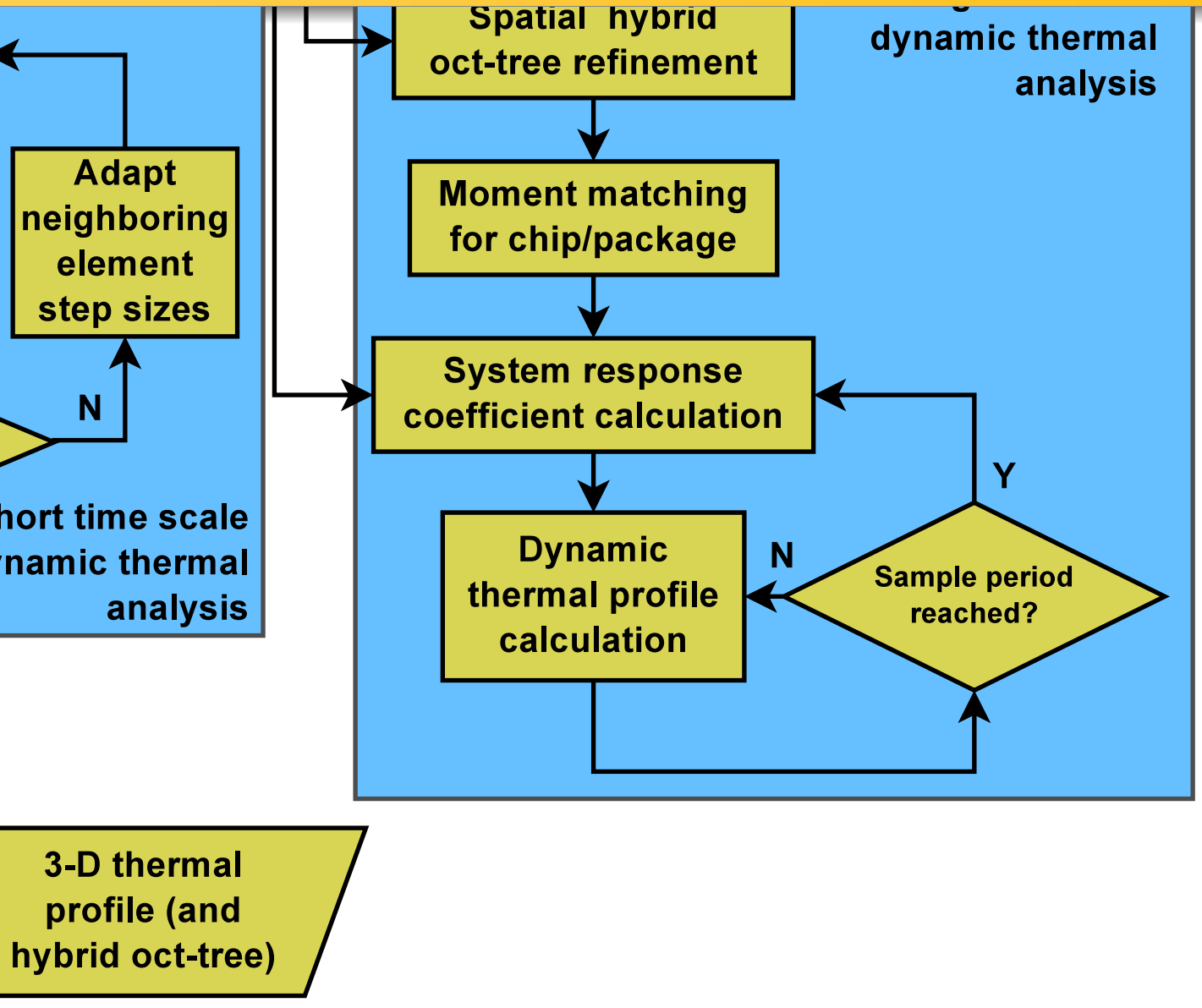
Thermal analysis infrastructure overview



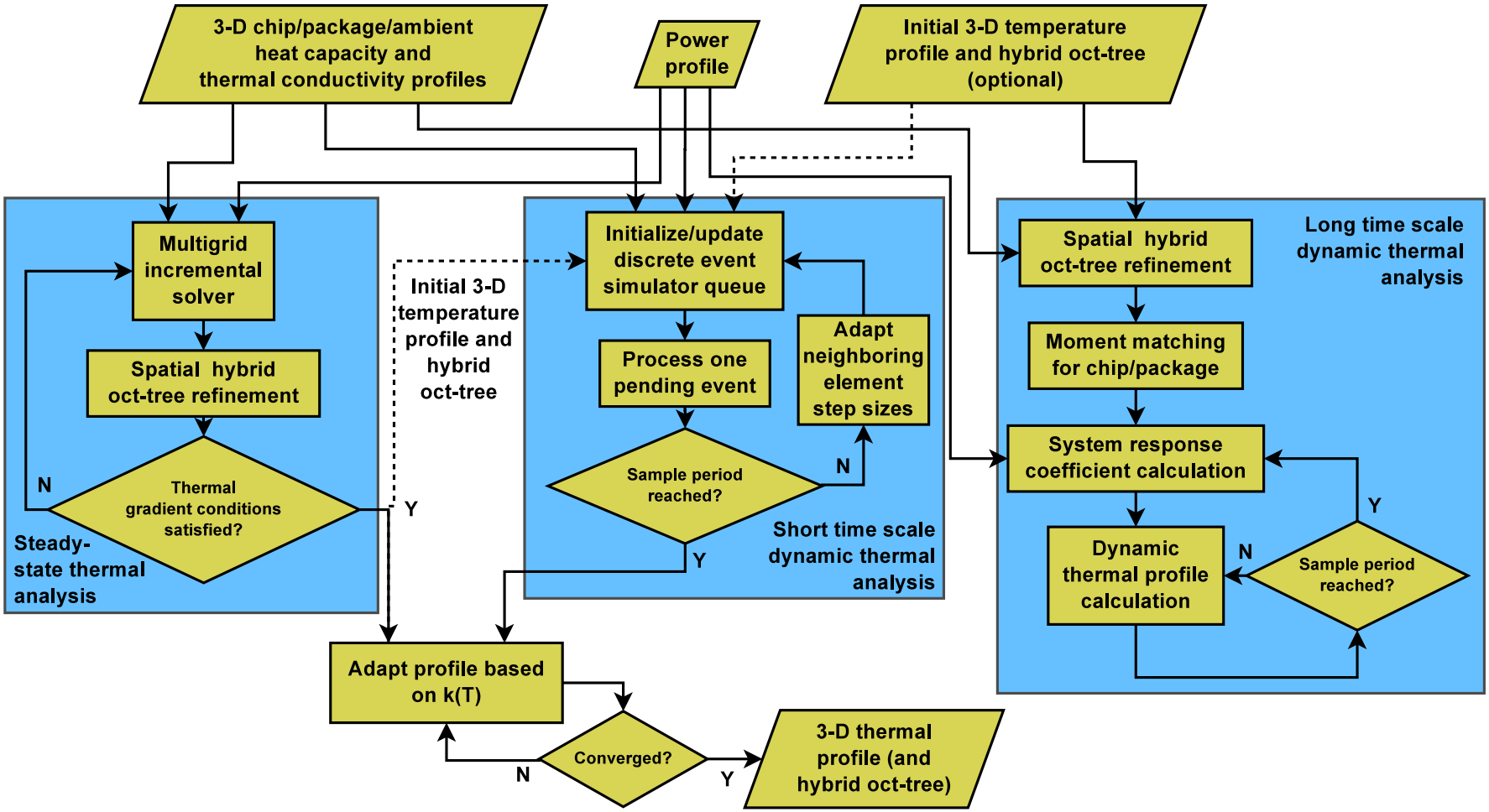
Thermal analysis infrastructure overview



Thermal analysis infrastructure overview



Thermal analysis infrastructure overview



Steady-state thermal analysis

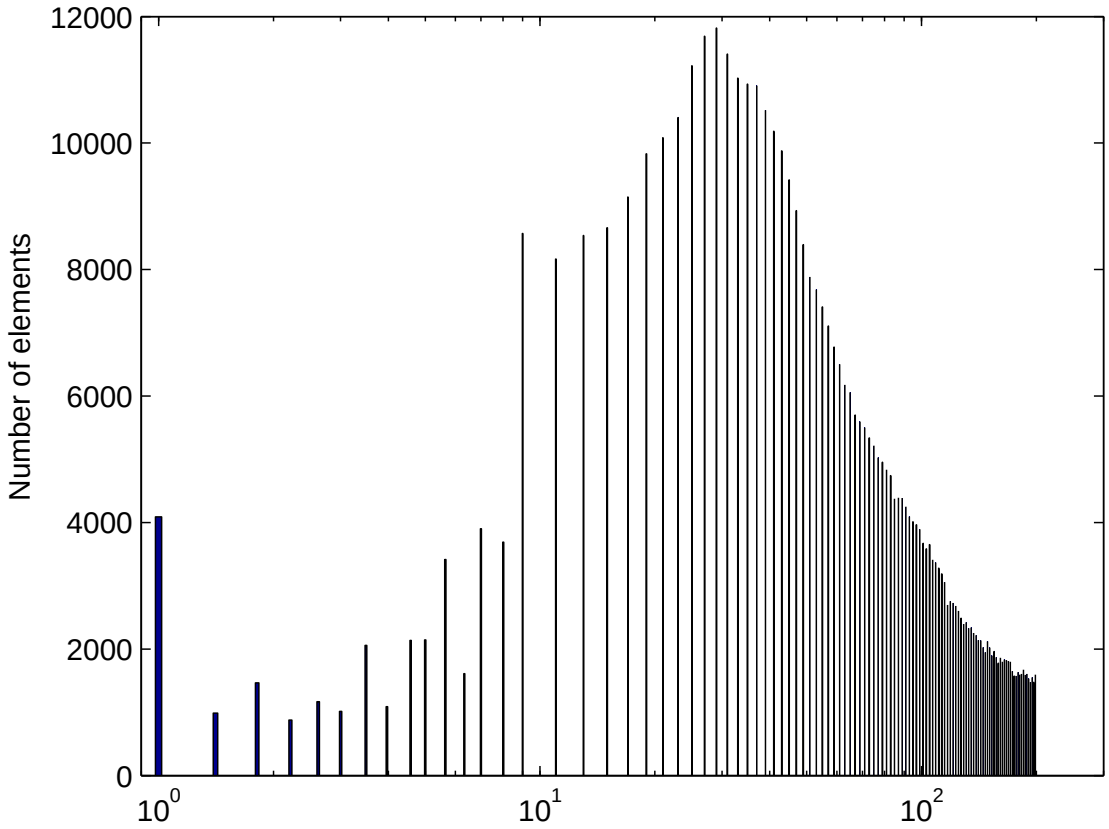
Basis: Multigrid analysis

Fast, multi-resolution relaxation method for matrix solving.

1. Iterative solver (relaxation) on fine grid.
2. Coarsen and propagate residual upward.
3. Iterative solver for error at coarser level.
4. Correct fine-grained solution based on coarse-grained error.
5. Iterative solver for error at fine level.

- Main challenge: Too slow for repeated use on large structures, especially 3-D chip-package modeling.
- Observation: Steepness of thermal gradients vary across IC.

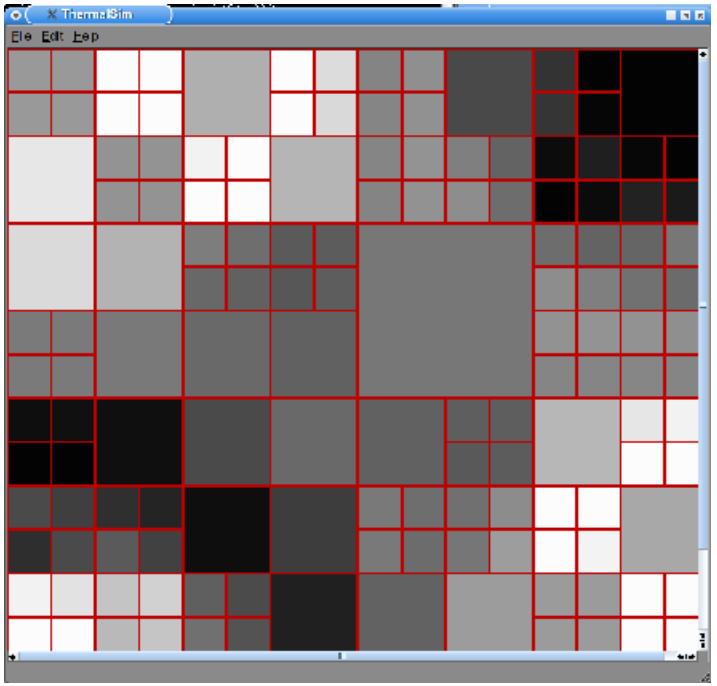
Neighbor temperature difference histogram



Spatial adaptation can improve performance w.o. loss of accuracy.

Hybrid oct-tree

- Reduce element count by merging when $\Delta_T < \epsilon$
- Conventional oct-tree inefficient for chip-package model
- Anisotropic thermal gradients
- We generalize to hybrid oct-tree
- Arbitrary partitioning on each axis



Hybrid oct-tree