Electrothermal Analysis and Optimization Techniques for Nanoscale Integrated Circuits

Yong Zhan, Brent Goplen, and Sachin S. Sapatnekar
ECE Department
University of Minnesota
sachin@ece.umn.edu
Motivation: Temperature problems
The problem of heat removal

[Joshi, Georgia Tech]  [Viswanath et al., Intel]
Interesting (certainly novel) approaches to cooling

Cooking oil as a coolant

Multitasking

Strip Out The Fans:
The DIY Oil PC

[ http://www.tomshardware.com/2006/01/09/strip_out_the_fans ]

By Trubador, available at
http://www.phys.ncku.edu.tw/~htsu/humor/fry_egg.html
Chip cooling technologies
(“Cooling a 200W Light Bulb that is the Size of a Postage Stamp”)

Air cooling (passive or active)
- Heat sink
- Thermal interface materials (TIMs), heat spreaders
  - Next generation TIMs shows much better thermal conductivity
- Thermal vias

Exotic cooling techniques
- Microchannels
- Peltier elements

[Viswanath et al., Intel]

[www.cooligy.com]
Cost of cooling a microprocessor

[Intel, via Hannemann]
Physical limitations on heat sinks

[Joshi, Georgia Tech]  [Goodson, Stanford]
Motivation: Electrothermal effects
Heat flux maps vs. temperature maps

[Viswanath et al., Intel]
On-chip temperature variations

Heat Flux (W/cm²) Results in $V_{cc}$ variation

Temperature Variation (°C)

[Borkar, Intel]
Temperature contours: Core vs. cache

Core

120ºC

Cache

70ºC

Temp (ºC)

108
105.2
102.4
99.53
96.7
93.86
91.03
88.2
85.36
82.53
79.7
76.87
74.03
71.2
68.37
65.53
Power as a function of application

Average and Peak Power as a % of Max Peak

[McGowen, Intel]
Leakage current effects

- Leakage current varies exponentially with temperature
- Self-consistent solutions

Thermal runaway

Variability effects

Ref: M Miller, NGBI, 2001

[Vassighi, Intel] [Borkar, Intel]
Reliability impact

- Electromigration
  - Black’s equation: increased temperature reduces mean time to failure
    \[ \text{MTTF} = A_0 (J - J_{\text{crit}})^{-n} e^{-\frac{E_A}{kT}} \]
- Hot carrier injection
- Negative bias temperature instability (NBTI)

[Schroder]

[Alam]
Electrothermal design

- Simple approach
- Integrated approach
  - Include thermal effects during analysis, optimization
  - Tightly coupled analysis/optimization
- Temperature affects
  - Leakage power
  - Timing
    - Higher temperatures reduce $V_T$, reduce mobility
- Temperature is affected by
  - Leakage power
  - Timing
Thermal analysis
Thermal analysis

- **Heat generation**
  - Switching gates/blocks act as heat sources
  - Time constants for heat of the order of ms or more
- **Temperature alters device behavior, switching speeds**
- **Strong local spatial characteristics**
Thermal analysis

- Thermal equation: partial differential equation

\[ K_x \frac{\partial^2 T}{\partial x^2} + K_y \frac{\partial^2 T}{\partial y^2} + K_z \frac{\partial^2 T}{\partial z^2} + Q(x, y, z) = 0 \]

- Boundary conditions corresponding to the ambient, heat sink, etc.

- Self-consistency
  - Power is a function of temperature, which is a function of power!
  - Often handled using iterations

- Some solution techniques
  - Numerical: solve large, sparse systems of linear equations
    - Finite difference method
    - Finite element method
  - System structure is similar to power grid systems
  - Semi-analytical
    - Green functions
The finite difference approach

\[ K_x \frac{\partial^2 T}{\partial x^2} + K_y \frac{\partial^2 T}{\partial y^2} + K_z \frac{\partial^2 T}{\partial z^2} + Q(x, y, z) = 0 \]

- Finite difference method
  - Discretize into elements; assume element temp. constant
  - Thermal-electrical analogy
    - Can find “thermal resistance” values between element centers
- Eliminate internal mesh nodes to get
  \[ G T = P \]
  - G is the thermal conductance matrix
  - T and P are the temperature and power density vector over mesh nodes on the top surface of the wafer
The finite element approach

\[ K_x \frac{\partial^2 T}{\partial x^2} + K_y \frac{\partial^2 T}{\partial y^2} + K_z \frac{\partial^2 T}{\partial z^2} + Q(x, y, z) = 0 \]

- Also a discretization methods
  - Discretize into elements; use polynomial interpolation based on values at nodes
  - Use “element stamps” and assemble these into a larger matrix
  - Apply boundary conditions to get
    \[ G \mathbf{T} = \mathbf{P} \]
  - \(G\) here is denser and smaller than for FDM
**Element stiffness matrix**

Stamp for a hexahedral element

- Rows and columns correspond to nodes 1 - 8

\[
[k] = \begin{bmatrix}
\text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F} & \text{G} & \text{H} \\
\text{B} & \text{A} & \text{D} & \text{C} & \text{F} & \text{E} & \text{H} & \text{G} \\
\text{C} & \text{D} & \text{A} & \text{B} & \text{G} & \text{H} & \text{F} & \text{E} \\
\text{D} & \text{C} & \text{B} & \text{A} & \text{H} & \text{F} & \text{E} & \text{G} \\
\text{E} & \text{F} & \text{H} & \text{G} & \text{A} & \text{B} & \text{C} & \text{D} \\
\text{F} & \text{E} & \text{G} & \text{H} & \text{B} & \text{A} & \text{D} & \text{C} \\
\text{G} & \text{H} & \text{E} & \text{F} & \text{C} & \text{D} & \text{A} & \text{B} \\
\text{H} & \text{G} & \text{F} & \text{E} & \text{D} & \text{C} & \text{B} & \text{A} \\
\end{bmatrix}
\]

where

\[
A = \frac{K_xh_d}{9w} + \frac{K_yw_d}{9h} + \frac{K_zw_h}{9d}, \quad B = -\frac{K_xh_d}{9w} + \frac{K_yw_d}{18h} + \frac{K_zw_h}{18d} \\
C = -\frac{K_xh_d}{18w} - \frac{K_yw_d}{18h} + \frac{K_zw_h}{36d}, \quad D = \frac{K_xh_d}{18w} - \frac{K_yw_d}{9h} + \frac{K_zw_h}{18d} \\
F = \frac{K_xh_d}{18w} + \frac{K_yw_d}{18h} - \frac{K_zw_h}{9d}, \quad E = -\frac{K_xh_d}{18w} + \frac{K_yw_d}{36h} - \frac{K_zw_h}{18d} \\
G = -\frac{K_xh_d}{36w} - \frac{K_yw_d}{36h} - \frac{K_zw_h}{36d}, \quad H = \frac{K_xh_d}{36w} - \frac{K_yw_d}{18h} - \frac{K_zw_h}{18d}
\]
Element and global matrices

- Elements are aligned in a grid pattern
- Element matrices, $k$, are calculated for each element
- Similar to the Modified Nodal Formulation:
  - These stamps, $K$, are added to the global matrix, $K_{\text{global}}$
- Now solve
  
  $$K_{\text{global}} \mathbf{T} = \mathbf{P}$$

- $\mathbf{P}$ = power vector, $\mathbf{T}$ = temperature vector
Reducing the global matrices using fixed temperatures ("ground nodes")

- Starting with a global system of equations
  - $X_1$ are the unknown values
  - $X_2$ are fixed values

$$
\begin{bmatrix}
K_{11} & K_{12} \\
K_{21} & K_{22}
\end{bmatrix}
\begin{bmatrix}
X_1 \\
X_2
\end{bmatrix} =
\begin{bmatrix}
F_1 \\
F_2
\end{bmatrix}
$$

- Eliminate rows and columns corresponding to fixed values

$$
K_{11} \{X_1\} = \{F_1\} - K_{12} \{X_2\}
$$

- Results in a reduced system of equation
- Applicable to both FEA and force-directed methods
The Green function method

- Problem definition

\[ \nabla^2 T(x, y, z) = 0 \]

\[ \frac{\partial T(x, y, z)}{\partial x} \bigg|_{x=0,a} = \frac{\partial T(x, y, z)}{\partial y} \bigg|_{y=0,b} = 0 \]

\[ k \frac{\partial T(x, y, z)}{\partial z} \bigg|_{z=0} = P_d(x, y) \]

\[ k \frac{\partial T(x, y, z)}{\partial z} \bigg|_{z=-d} = hT(x, y, z) \bigg|_{z=-d} \]

**Chip**

**Packaging**

**Adiabatic**

**Convective**

\( P_d \) – power density, \( k \) – thermal conductivity, \( h \) – heat transfer coefficient
Advantages:
- no 3D meshing necessary
- can do localized solve efficiently
The Green function method (contd.)

$$\nabla^2 G(x, y, z, x', y') = 0$$

$$\frac{\partial G(x, y, z, x', y')}{\partial x} \bigg|_{x=0, a} = \frac{\partial G(x, y, z, x', y')}{\partial y} \bigg|_{y=0, b} = 0$$

$$k \frac{\partial G(x, y, z, x', y')}{\partial z} \bigg|_{z=0} = \delta(x - x') \delta(y - y')$$

$$k \frac{\partial G(x, y, z, x', y')}{\partial z} \bigg|_{z=-d} = hG(x, y, z, x', y') \bigg|_{z=-d}$$

$$G(x, y, x', y') = G(x, y, z = 0, x', y') = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} c_{mn} \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) \cos\left(\frac{m\pi x'}{a}\right) \cos\left(\frac{n\pi y'}{b}\right)$$
Fast computation techniques

- Algorithm I: Table look-up approach [Zhan, ASPDAC05]
  - Solve the double infinite summation issue
  - Suitable for sensitivity analysis and incremental calculation

- Algorithm II: Frequency domain computation approach [Zhan, ICCAD05]
  - Solve the pair-wise calculation issue
  - Suitable for full-chip temperature profiling

- Algorithm III: Precorrected FFT approach
  - Solve problems with local high accuracy requirements
Sample results

Runtime comparison
Algorithm I: 30msec
Algorithm II: 10msec
Another example

1024x1024 grid cells
Electrothermal Optimization/Mitigation
Electrothermal optimization

- Various techniques at all levels of design

- Some examples
  - Architectural optimizations
  - Thermal mitigation
    - Placement
    - Application of body biases
Recall: Power as a function of application

Average and Peak Power as a % of Max Peak

[McGowen, Intel]
Architectural mitigation

- Work by Skadron, Stan et al. at Virginia
- Coarse FDM model (HotSpot)
- Coupled with microarchitectural simulator
- Can be used for analyzing and optimizing microarchitectures
- Integrate clock gating/dynamic voltage scaling optimizations
Placement

- Spatial distribution of cells can affect temperature distributions
- 3D circuits: thermal issues are much stronger
- Force-directed approach

Spring force attracts connected cells together

Thermal gradients force cells away from hot spots
Heat removal through thermal vias (3D)

- Rows of Standard Cells
- Thermal via regions
- Inter-row region
- Inter-layer elements
- Layer elements
- Standard cells (heat sources)
- Bulk substrate elements

Thermal Via
Substrate
Heat removal through thermal vias (3D)

Before

After
Measuring temperature

- Place thermal sensors (diodes) at various points

[McGowen, Intel]
Adaptive body bias

Circuit Block

Bias Gen.

PD

PD = Phase detector and critical path

Apply $F_{\text{target}}$

Apply NMOS bias

PMOS bias adapts

Measure $P_{\text{leak}}$ of circuit block

Pick best NMOS/PMOS bias (minimize $P_{\text{leak}}$)

Circuit block $P_{\text{leak}} < P_{\text{leak,max}}$?

NO

YES

Reduce $F_{\text{target}}$

[J. Tschanz, ISSCC02]
Within-die adaptive body bias

Compensating for within-die variation

Area overhead:
Similar to ABB

Apply F_{target}  Reduce F_{target}

Apply NMOS bias to all circuit blocks

Circuit Block 1:
Adapt PMOS bias

Measure P_{leak} of block

Pick best NMOS/PMOS bias

Measure total die leakage P_{leak}

P_{leak} < P_{leak,max}?  NO

YES
Conclusion

- Temperature issues are vital for nanometer-scale designs
- Old metrics (power, etc.) aren’t good enough
- A coordinated electrothermal design strategy is essential