ECE 225
High-Speed Digital IC Design
Lecture 12

Nanoscale Power and Thermal Management:
Electrothermal Couplings in Nanoscale CMOS, Design-Package Co-Optimization, Self-Consistent Temperature Profile Estimation, Cooling Analysis

Prof. Kaustav Banerjee
Electrical and Computer Engineering
E-mail: kaustav@ece.ucsb.edu
Outline

- Introduction
  - Technology Scaling Challenges
  - Power and Thermal Problems
  - Integrated Approach

- Integrated Power-Thermal Management
  - Electrothermal Couplings in Nanoscale CMOS
  - Design-Package Co-Optimization
  - Self-Consistent Temperature Profile Estimation
In 1965, Gordon Moore sketched out his prediction of the pace of silicon technology. Decades later, Moore’s Law remains true, driven largely by Intel’s unparalleled silicon expertise.
Process-Induced Random Variations in Devices
- Critical Dimension Variation
- Random Dopant Fluctuation

Random variations cause negative shift of threshold voltage, which increases leakage substantially.
Technology Scaling Challenges — 2

Leakage

Say goodbye to “happy scaling”…….
Technology Scaling Challenges — 3

- **Leakage:**
  - Gate Stack Processes and Materials (Reduce gate leakage)

1. Silicon Dioxide (1.2 nm @ 90 nm node)
2. Polysilicon

- High-K gate dielectric
- Metal gate electrode with proper work function

1. Silicon Dioxide is too thin to be a good insulator.
2. Alternative gate electrode materials are needed.
   (compatible with high-K gate dielectric)
Subthreshold leakage increases with scaling and temperature...
Increasing Power Demand...

- Higher performance
- More Functionalities
- Higher interconnect power
- Increasing leakage

Power Density (W/cm²)

- Rocket Nozzle
- Nuclear Reactor
- Hot Plate

International Technology Roadmap for Semiconductors

Max Power (Watt)

- ITRS 2006 High-Performance
- ITRS 2006 Cost-Performance
- IBM Power4
- Intel Itanium 2
- Intel Xeon

Technology node

- Pentium® 4
  ~ 1.31 cm²

- Core™2 Duo
  ~ 1.43 cm²

Courtesy Intel
Power has been a technology driver....

Power Wall: Module Heat Flux Trend

Year of Announcement

Integrated Circuit 1957

Power is the only limiter

R. Schmidt, T-C. Chen, IBM
Integrated Power & Thermal Management

Temperature distribution plays a key role...

Heat Transfer Mechanisms
- Active Power
  - Device
  - Interconnect
- Leakage Power
  - Device
- Layout geometry
- Packaging Cooling

Full chip power dissipation and temperature profile estimation

System level
- Packaging & Cooling
- System power dissipation
- System performance

Circuit level
- Timing
- Placement / Routing
- Buffer insertion

Device / Interconnect level
- Self-Heating
- Leakage mechanisms
- Reliability (EM, ESD)

Temperature-Aware Design

Wider implications for emerging technologies

3D ICs
NEMS
CNFET
CNT
SET
Outline

- Introduction
  - Technology Scaling Challenges
  - Power and Thermal Problems
  - Integrated Approach

- Integrated Power-Thermal Management
  - Electrothermal Couplings in Nanoscale CMOS
  - Design-Package Co-Optimization
  - Self-Consistent Temperature Profile Estimation
Design/ Packaging Conflicts!

- Packaging
- Cooling Solutions
- System Power Performance Optimization
- Placement Routing Schemes
- Accurate Temperature Profile
- Clock Timing Buffer Insertion
- Device Reliability
- Interconnect Reliability

Key piece of the design puzzle

Design

Packaging

Thermal Management
- Placement Routing Schemes
- Packaging Cooling Solutions
- System Power Performance Optimization
- Power Ground Integrity
- Leakage
- Interconnect Reliability
- Device Reliability
- Interconnect Reliability

Accurate Temperature Profile

Socket
- PCB
- Substrate
- Core (die)
- TIM 1
- TIM 2
Positive feedback between temperature, leakage and total power leads to electrothermal couplings, which had been inconspicuous for earlier generation of ICs.

Banerjee et al., IEDM 2003
Self-Consistent Methodology

Lin et al., TED Dec 2007

Initial $V_{dd}$ & $T_j$

Frequency Estimation

$V_{dd}$ & Temperature dependent $I_{on}$ & $I_{off}$ from Circuit Simulation

Full-Chip Power Estimation

Packaging/Cooling Models

Chip Temperature Estimation

Obtain a self-consistent average junction temperature

Check $T_j$ Convergence

Reliability Check

Update $T_j$

No

Yes

Self-Consistent $T_j$

1. Logic Depth (critical path)
   Capacitance (driver / load)
   Ion ($V_{dd}$, $T_j$)

2. $P_{switching}$ (logic / memory)
   $P_{leakage}$ (logic / memory)
   $P_{short}$-circuit

3. $P_{switching}$ (logic / memory)
   $P_{leakage}$ (logic / memory)
   $P_{short}$-circuit
1. Leakage power becomes a major contributor to total power when junction temperature is high.
2. For leakage dominant technologies, lowering supply voltage may even improve performance.
Outline

- Introduction
  - Technology Scaling Challenges
  - Power and Thermal Problems
  - Integrated Approach

- Integrated Power-Thermal Management
  - Electrothermal Couplings in Nanoscale CMOS
  - Design-Package Co-Optimization
  - Self-Consistent Temperature Profile Estimation
  - IC Cooling for Power/Thermal Management

- Conclusions
Bring Design/ Packaging Closer....how?

Lin and Banerjee, TVLSI 2008 (to appear)

**Target:**
Find a minimal value of Energy-Delay-Product for operation.
M. Horowitz, 1997

**Delay**
Delay of various gates in the critical path remains proportional to the delay of an equivalent inverter.
Can be evaluated by Alpha-Power model
T. Sakurai and A. R. Newton, 1990

**Energy**
Can be calculated by total power dissipation
Incorporate Interconnect Effects in EDP Optimization

\[ \tau = R_{tr} (C_p + C_L + cL) + rL C_L + \left( \frac{1}{2} \right) r c L^2 \]

\[ T_g \approx 0.69 R_{tr} (C_p + C_L + cL) + 0.69 rL C_L + 0.38 r c L^2 \]

\[ T_g \approx \frac{K V_{dd}}{(V_{dd} - V_{th})^\alpha} \left[ 1 + \frac{cL}{C_p + C_L} \right] + 0.69 rL C_L + 0.38 r c L^2 \]

\[ P_{dynamic} = a C_{eff} V_{dd}^2 F \]

\[ P_{static} = I_s e^{-V_{th}/V_o} (1 - e^{-V_{ds}/V_o}) W_{eff} V_{dd} \]

\[ Energy = (P_{dynamic} + P_{static}) \cdot T_g L_d \]
Electrothermally-Coupled EDP

- A shift in optimal point, EDP and performance contours
- An overall change in shape of EDP contours
- Iso-temperature curves generated using self-consistent methodology
- Operation region restricted by electrothermal constraints

Lin et al., ICCD 2005
Generalized Metric for Optimization

- EDP is not the only metric

- Other Metrics...

  EDP is Energy X Delay = PT^2
  PDP is Power X Delay = PT
  PEP is Power X Energy = P^2T

Generalized Metric: PT^μ

μ is the ratio of the exponent of delay over that of power (μ = 2 for EDP)

How to choose a design-specific metric?
Optimal Operation Locus

\[ V_{dd} = V_{th} \]

- Restricted by temperature limit
- Restricted by \( V_{dd} > V_{th} \) constraint or required performance constraint

- \( \mu \) value

- Thermal runaway

\[ \mu \text{ is determined by thermal and performance requirements} \]
Choose an optimization metric to meet thermal/packaging requirements (For T = 70°C, µ = 3.3)

Impact of scaling should be taken into account (For T = 70°C, µ = 3)
Outline

- Introduction
  - Technology Scaling Challenges ✓
  - Power and Thermal Problems ✓
  - Integrated Approach ✓

- Integrated Power-Thermal Management
  - Electrothermal Couplings in Nanoscale CMOS ✓
  - Design-Package Co-Optimization ✓
  - Self-Consistent Temperature Profile Estimation
Non-Uniform Power Density

Intel's first dual-core Itanium-Montecito 90nm

2.5W

Cache

Core 1

Core 2

AMD's first dual-core Athlon 64 X2 90nm SOI

40W

Cache

Core 1

Core 2

Multi-Core

Multi-HotSpots

S. Borkar, DAC 2003
C. Poirier, ISSCC 2005

Non-uniform power density generates on-chip temperature gradient
Methodology Overview

**Packaging / Cooling Model**
- Heatsink
- Heat Spreader
- TIM 2
- TIM 1
- Core (die)
- Substrate
- Socket
- PCB

**Layout Geometry & Power Dissipation**

**Electrothermal Couplings**

**Boundary Condition**

**Parabolic Heat Partial Differential Equations**

**3-D Electrothermally-Aware Spatial Temperature Estimation**

\[
\frac{\partial T}{\partial n_j} = \frac{h}{K} [T - T_{amb}]
\]

\[
\frac{\partial T}{\partial t} = \left( \frac{K}{\rho C_p} \right) \left( \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) + \frac{p}{\rho C_p}
\]
Numerical Approach

\[
\frac{\partial T}{\partial t} = \frac{k}{\rho C_p} \left( \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) + \frac{\rho}{\rho C_p} \Delta T
\]

\[
\beta_x = \frac{k \Delta t}{2 \Delta x^2 \rho C_p}
\]

\[
\frac{T^{n+1} - T^n}{\Delta t} = \left( \frac{k}{\rho C_p} \right) \left[ \left( \frac{M_x}{2 \Delta x^2} + \frac{M_y}{2 \Delta y^2} + \frac{M_z}{2 \Delta z^2} \right) \left( T^{n+1} + T^n \right) \right] + \frac{\rho}{\rho C_p}
\]

\[
(1 - \beta_x M_x - \beta_y M_y - \beta_z M_z) (T^{n+1} - T^n) = 2(\beta_x M_x + \beta_y M_y + \beta_z M_z) T^n + \frac{\Delta t}{\rho C_p} \rho
\]

\[
(1 - \beta_x M_x)(1 - \beta_y M_y)(1 - \beta_z M_z) (T^{n+1} - T^n) = 2(\beta_x M_x + \beta_y M_y + \beta_z M_z) T^n + \frac{\Delta t}{\rho C_p} \rho
\]

Linear Equations:

\[
(1 - \beta_x M_x) U_{x,y,z} = V_{x,y,z}
\]
Self-Consistent Evaluation

From $P(t_0)$, solve $T(t_0+\Delta t)$

From $P(t_0+\Delta t)$, solve $T(t_0+2\Delta t)$

Steady-State Temperature profile

Steady-State $P$
Methodology Verification

For a case with negligible leakage

Commercial Tool (IcePak)  Proposed Method

Equally select 400 data points from identical locations for comparison (total = 10000 pts)

Proposed method is validated against an industrial-quality software.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Max. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>h=1000</td>
<td>2.13 %</td>
</tr>
<tr>
<td>h=2000</td>
<td>2.48%</td>
</tr>
<tr>
<td>h=3000</td>
<td>3.26 %</td>
</tr>
<tr>
<td>h=4000</td>
<td>3.28 %</td>
</tr>
<tr>
<td>h=5000</td>
<td>3.06 %</td>
</tr>
</tbody>
</table>
Implementation & Setup

High leakage power region

High power density region

<table>
<thead>
<tr>
<th>Layer</th>
<th>Area (mm²)</th>
<th>Thickness (mm)</th>
<th>Specific Heat (J/kg°C)</th>
<th>Density (kg/m³)</th>
<th>K (W/m°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>10 X 10</td>
<td>0.8</td>
<td>712</td>
<td>2330</td>
<td>120</td>
</tr>
<tr>
<td>TIM 1</td>
<td>10 X 10</td>
<td>0.2</td>
<td>230</td>
<td>7310</td>
<td>30</td>
</tr>
<tr>
<td>IHS</td>
<td>30 X 30</td>
<td>1.8</td>
<td>385</td>
<td>8930</td>
<td>390</td>
</tr>
<tr>
<td>TIM 2</td>
<td>30 X 30</td>
<td>0.2</td>
<td>2890</td>
<td>900</td>
<td>6.4</td>
</tr>
<tr>
<td>Heatsink</td>
<td>60 X 60</td>
<td>6.4</td>
<td>385</td>
<td>8930</td>
<td>360</td>
</tr>
</tbody>
</table>
Impact of Considering ET-couplings

Traditional evaluation neglects electrothermal couplings

Considering electrothermal couplings leads to an additional hot-spot at the highest leakage power region
Impact of Package Models

Cubic Package Thermal Model

Realistic Package Thermal Model

- Simple cubic package thermal model ignores physical dimensions of different packaging layers and neglects lateral heat spreading.
- Simple cubic model over-estimates the temperature profile.
Implications for Power Estimation

- Power is not consistent with temperature in traditional evaluation
- Considering cubic package model overestimates leakage power
Efforts on Low-Power.....without hurting performance ...

- **Device Engineering**
  - Enhanced Channel Mobility - Strained Silicon
  - Reduced Gate Leakage - High-K Gate
  - Reduced Junction Leakage - Nitrogen Doped

- **Circuit Level**
  - Adaptive body-biasing, Dual $V_{DD}$-CMOS, Dual-$V_{th}$
  - Sleep Transistor, Clock/Power Gating

- **Micro-Architecture Level**
  - Multi-Core
Does IC Cooling Make Sense?

- **Range of IC Cooling**
  - Sub-Ambient Range
  - Practical Commercial Range

- **Prior Research on Cooling**
  - **Analysis under Cryogenic Condition**
    - Operating near liquid nitrogen temperature (77 K) requires complex cooling facilities
    - NOT for commercial purpose
  - **Analysis of Performance under Cooling**
    - Does NOT consider electrothermal couplings
    - Does NOT consider the system level power dissipation
Higher drive current can be achieved by cooled operation across all technology nodes due to higher carrier mobility.
Cooling---Benefit at the Circuit Level

- Enhance $I_{on}$ to $I_{off}$ ratio
- Reduce propagation delay and variance
- Benefit back-end performance and reliability

S. Borkar, Intel

- 9-stage Inverter Chain
- $\mu = 1.41e-9$, $\sigma^2 = 2.0e-21$, $T = 25^\circ C$
- $\mu = 1.55e-9$, $\sigma^2 = 2.3e-21$, $T = 125^\circ C$
Further Exploiting the Benefit of Cooling

N-MOSFET (90 nm)

Relative Performance Factor

Operating Temperature (°C)

Performance can be further improved by different design strategies

I. Aller et al., ISSCC 2000
Cooling---System Level Considerations

Beyond this point, further cooling does not lead to any power saving.

The limit occurs at a lower temperature as technology scales.
Practical limit can be further extended towards a lower temperature and higher performance is achieved by enhancing cooling efficiency.
Hot-Spot Management

Global Cooling
Thermal resistance reduces 20% 

Localized Cooling 

Localized cooling will be more effective for hot-spot management
For power/thermal management of nanoscale CMOS ICs, **electrothermal couplings** must be incorporated for trading-off chip-level power, performance, reliability, and cooling cost.

An accurate leakage-aware **self-consistent** power and silicon-substrate **temperature-profile-estimation** technique has been developed for nanoscale CMOS technologies.

A systematic methodology for power-performance optimization has been developed to **bring design and packaging closer**.

**Chip cooling** combining device, circuit, and system level considerations can extend CMOS scaling.
Additional Issues....

- Extend CMOS scaling
  - Time per node can be expanded via cooling

- Application Based
  - High-performance servers
  - Hand-held applications

- Implications for Ultra Low-Voltage Applications
  - Positive Temperature Dependence