ECE 225
High-Speed Digital IC Design
Lecture 9

Nanoscale Power and Thermal Management:

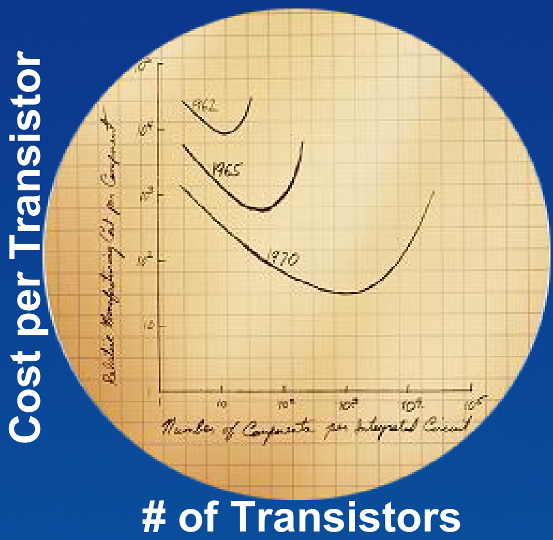
Electrothermal Couplings in Nanoscale CMO, Design-Package Co-Optimization, Self-Consistent Temperature Profile Estimation

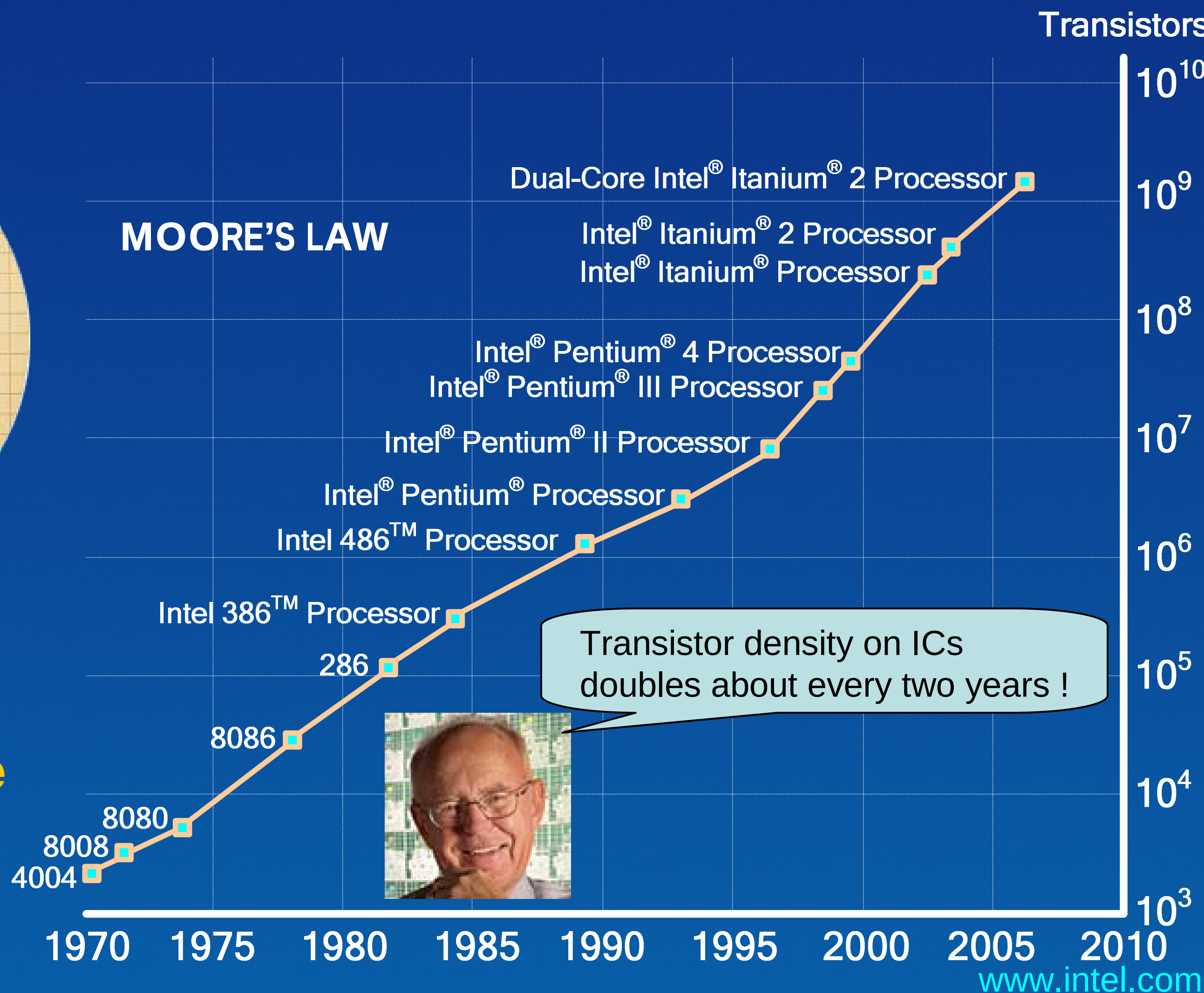
Prof. Kaustav Banerjee
Electrical and Computer Engineering
E-mail: kaustav@ece.ucsb.edu

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

Technology Scaling — Moore’s Law





Better Performance

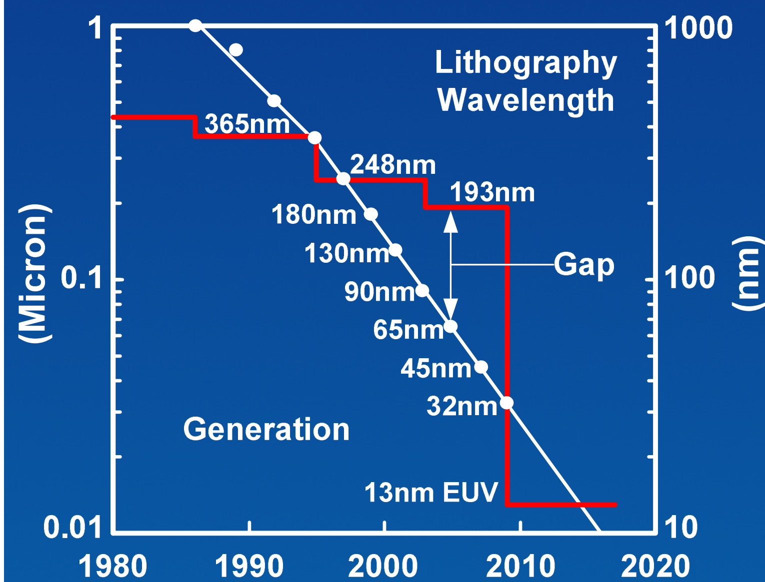
Lower Cost

More Functionality

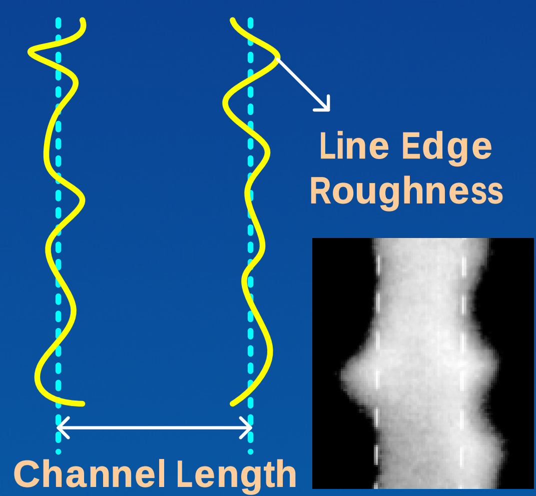
In 1965, Gordon Moore sketched out his prediction of the pace of silicon technology. Decades later, Moore’s Law remains true, driven largely by Intel’s unparalleled silicon expertise.

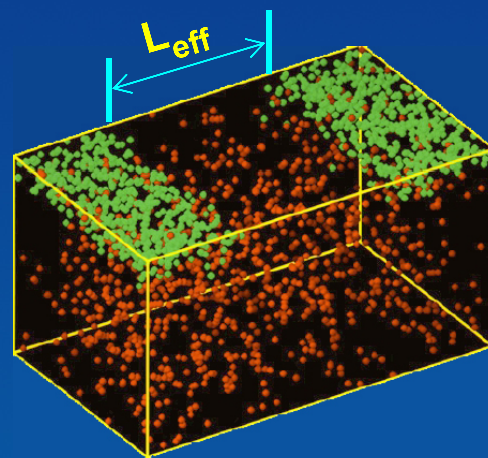
Technology Scaling Challenges — 1

- Process-Induced Random Variations in Devices
  - Critical Dimension Variation
  - Random Dopant Fluctuation



Source: Mark Bohr, Intel



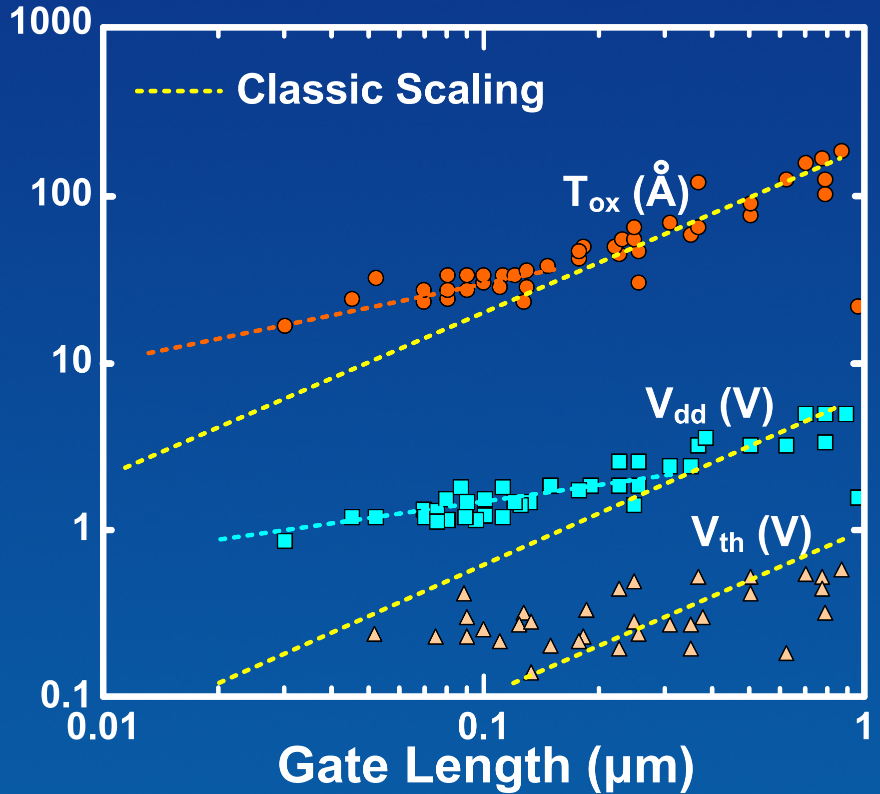


L_{eff}=30nm W_{eff}=50nm
Source: A. Asenov

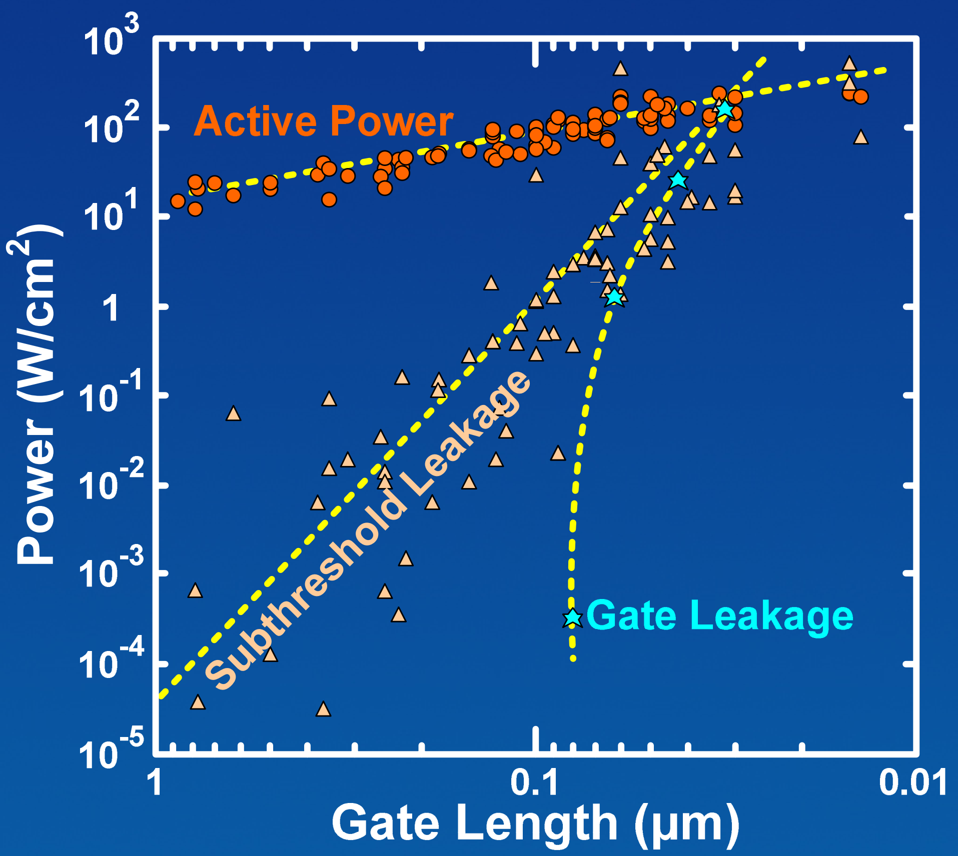
Random variations cause negative shift of threshold voltage, which increases leakage substantially.

Technology Scaling Challenges — 2

- Leakage



E.J. Nowak, 2002



E.J. Nowak, 2002

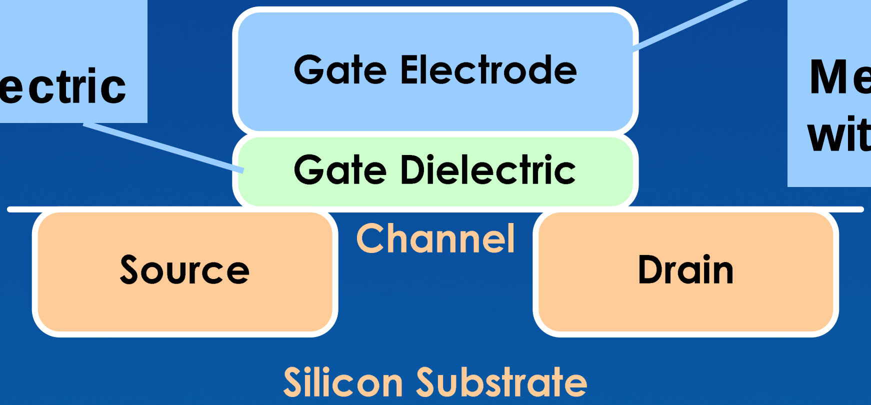
Say goodbye to "happy scaling" ......

Technology Scaling Challenges — 3

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

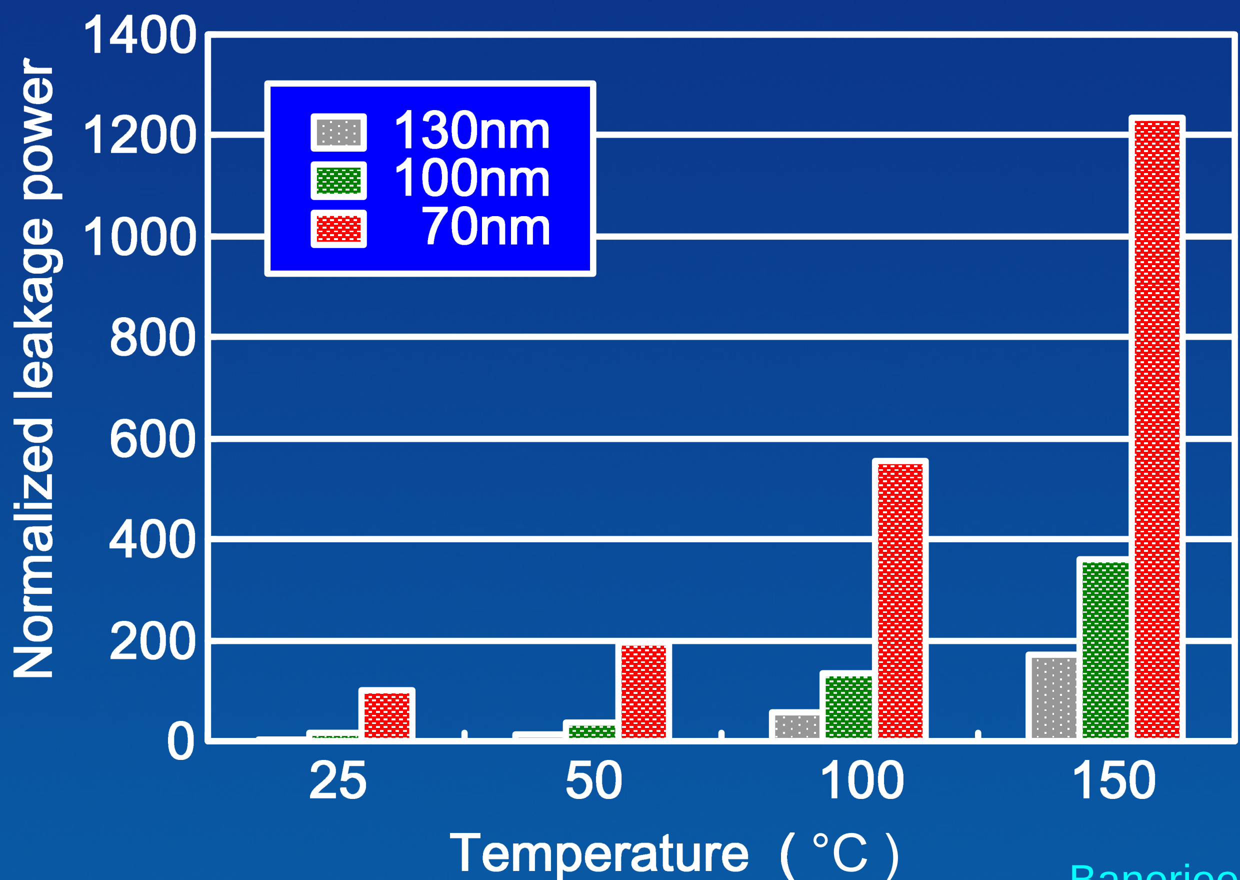
Silicon Dioxide
(1.2 nm @ 90 nm node)
↓
High-K gate dielectric

Polysilicon
↓
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)

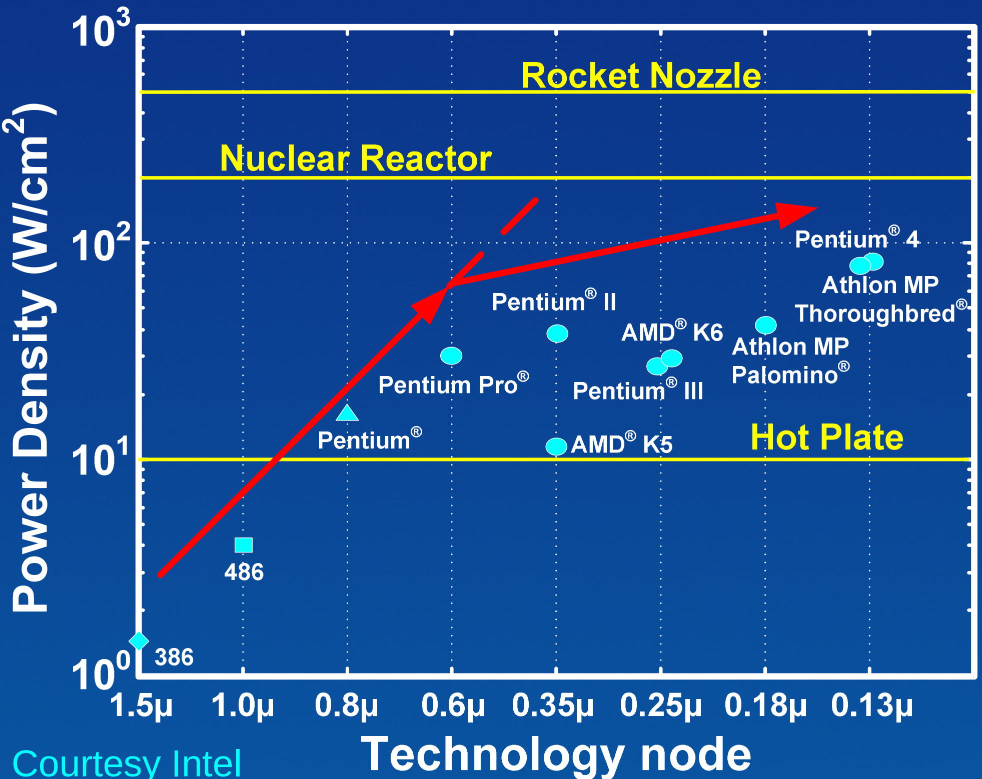
Technology Scaling Challenges — 4



Banerjee et al. IEDM 2003

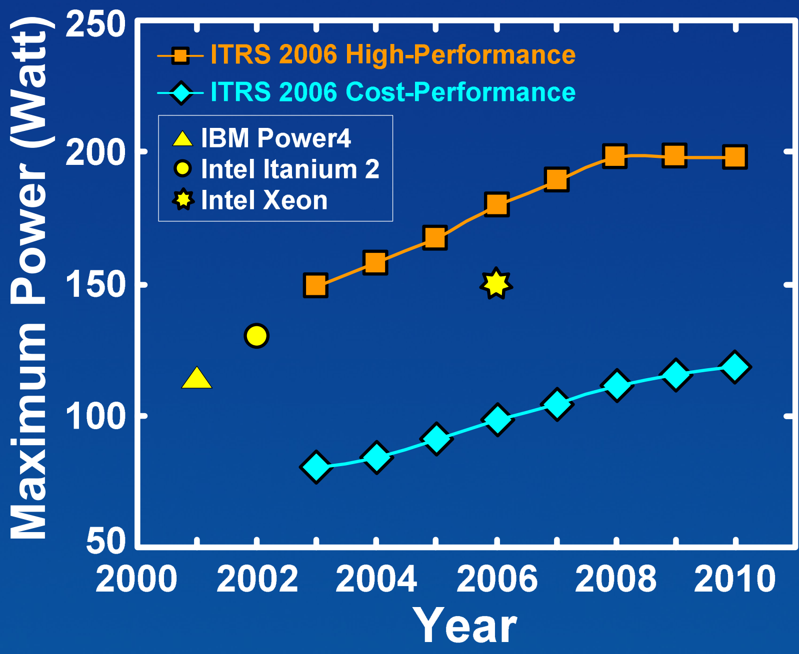
Subthreshold leakage increases with scaling and temperature……

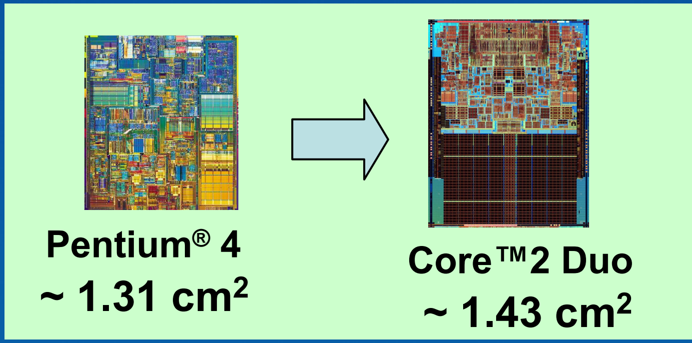
Increasing Power Demand.....



Courtesy Intel

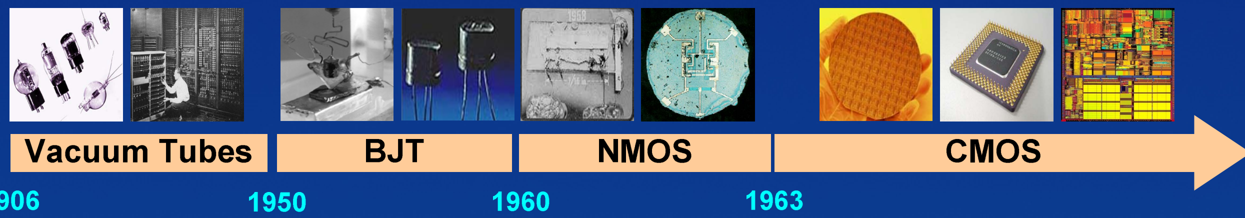
International Technology Roadmap for Semiconductors

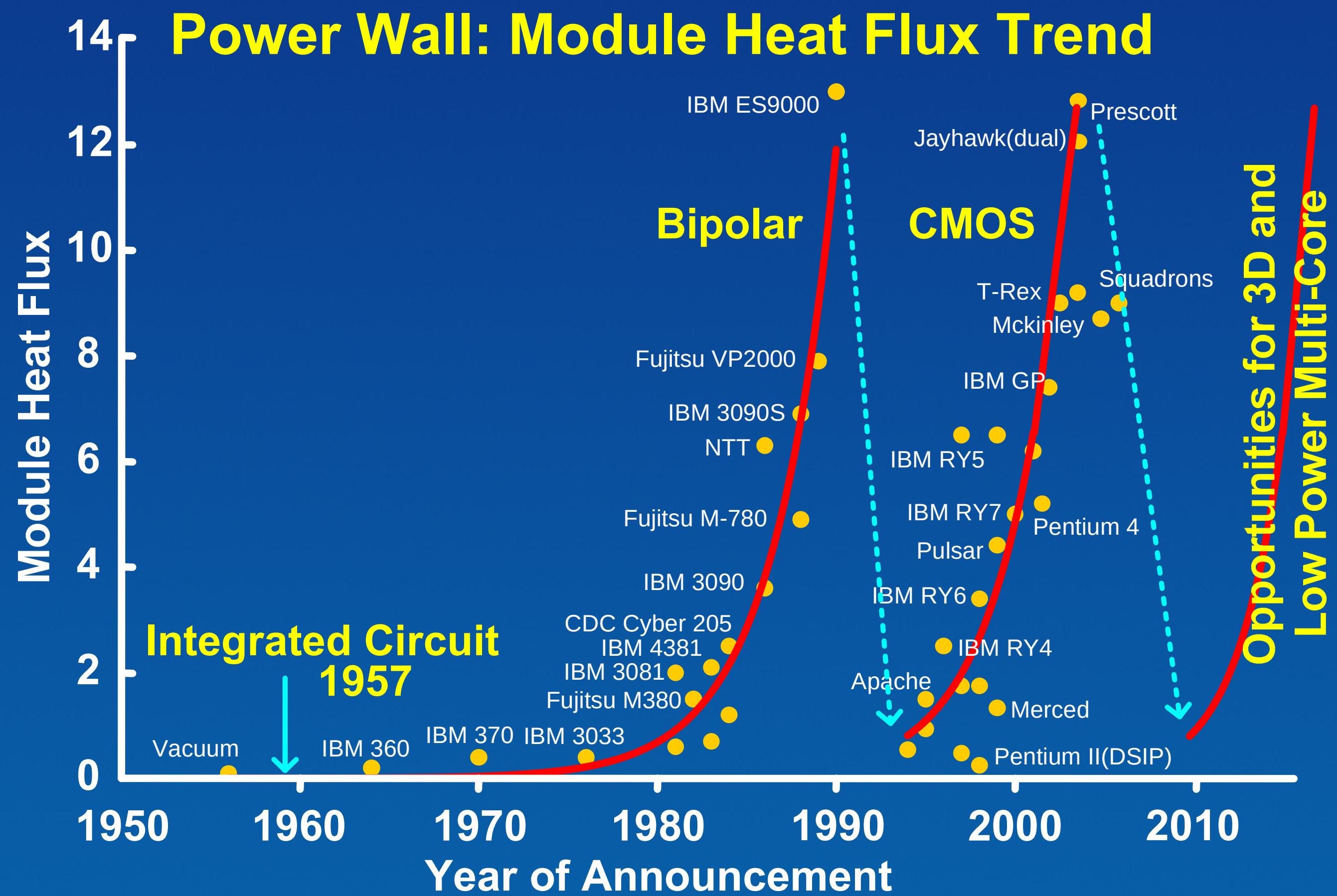




- Higher performance
- More Functionalities
- Higher interconnect power
- Increasing leakage

Power has been a technology driver....



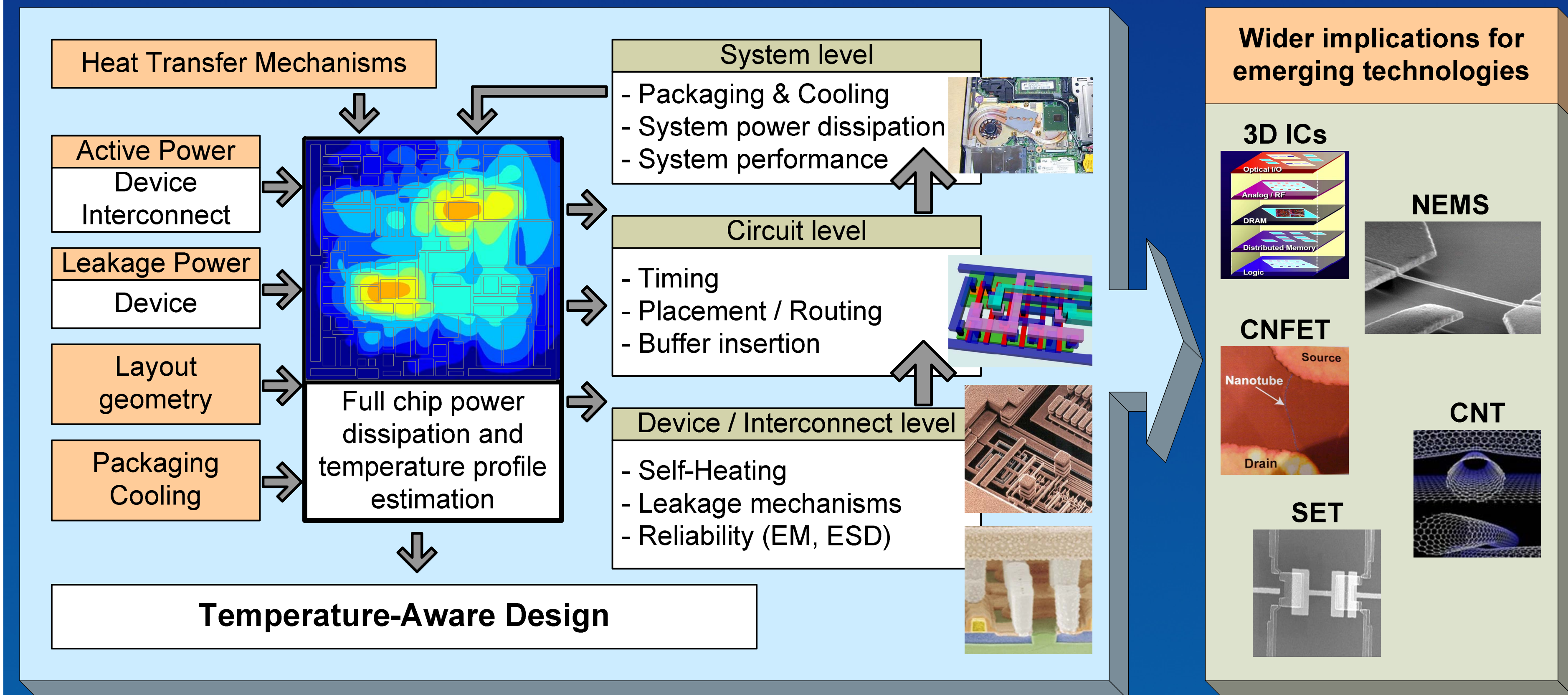


P. Gelsinger, Intel

R. Schmidt, T-C. Chen, IBM

Integrated Power & Thermal Management

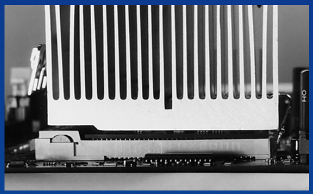
Temperature distribution plays a key role…

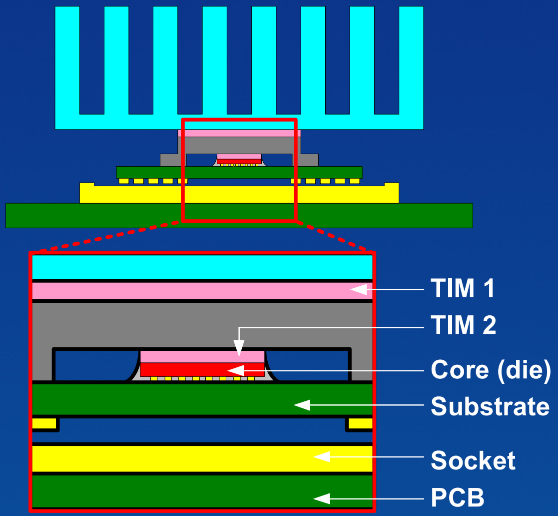


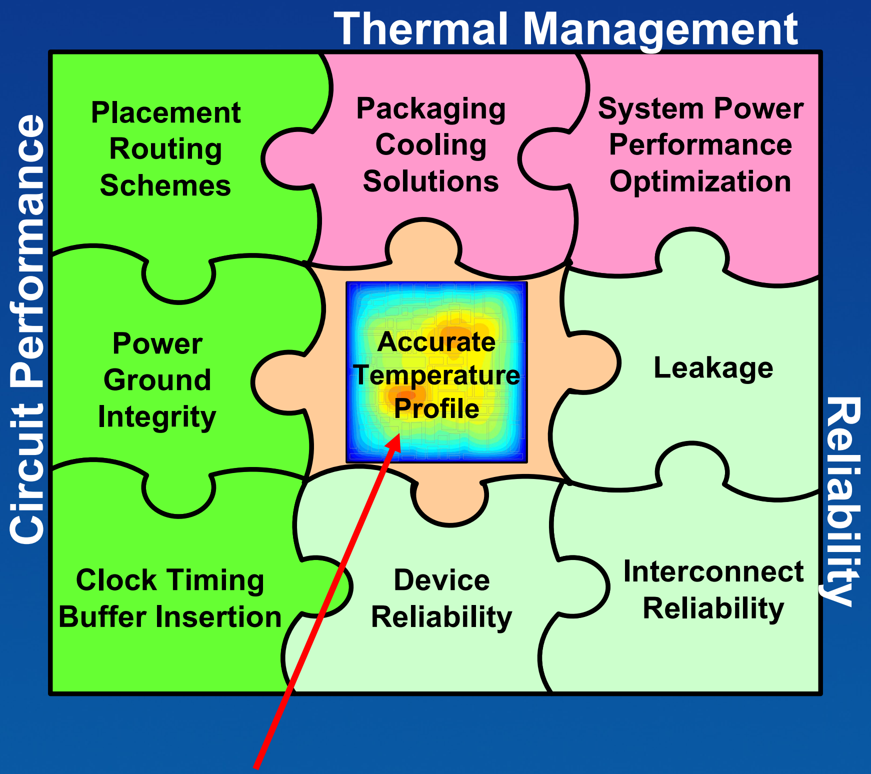
Outline

- 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!

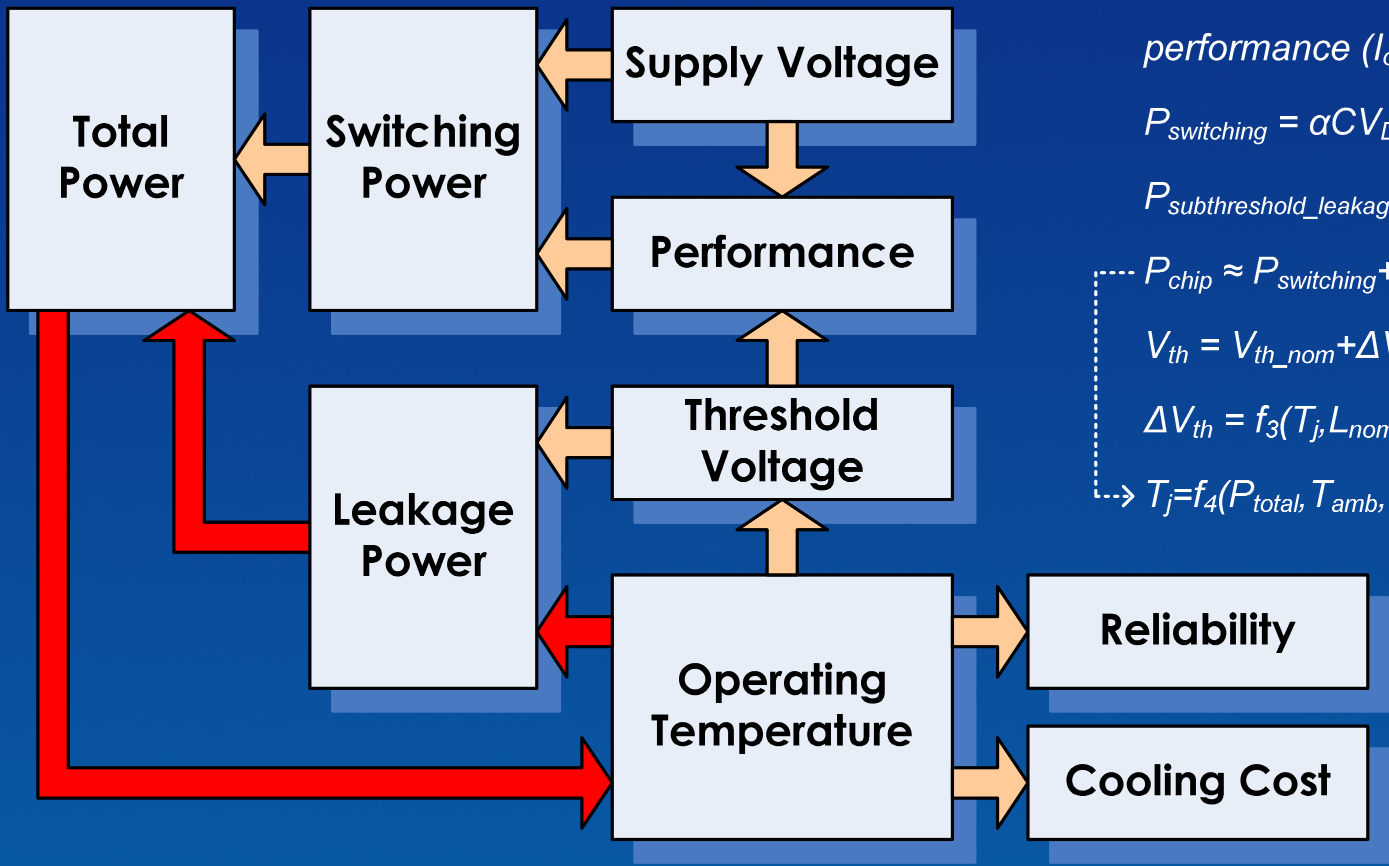






Key piece of the design puzzle

Electrothermal (ET) Couplings



$performance\ (I_{on}) = f_1(V_{DD}, V_{th}, T_j, L_{nom})$

$P_{switching} = \alpha C V_{DD}^2 F$

$P_{subthreshold\_leakage} = f_2(V_{DD}, V_{th}, T_j, L_{nom})$

$P_{chip} \approx P_{switching} + P_{subthreshold\_leakage}$

$V_{th} = V_{th\_nom} + \Delta V_{th}$

$\Delta V_{th} = f_3(T_j, L_{nom}, t_{ox}, X_j, V_{DD}, \Delta L)$

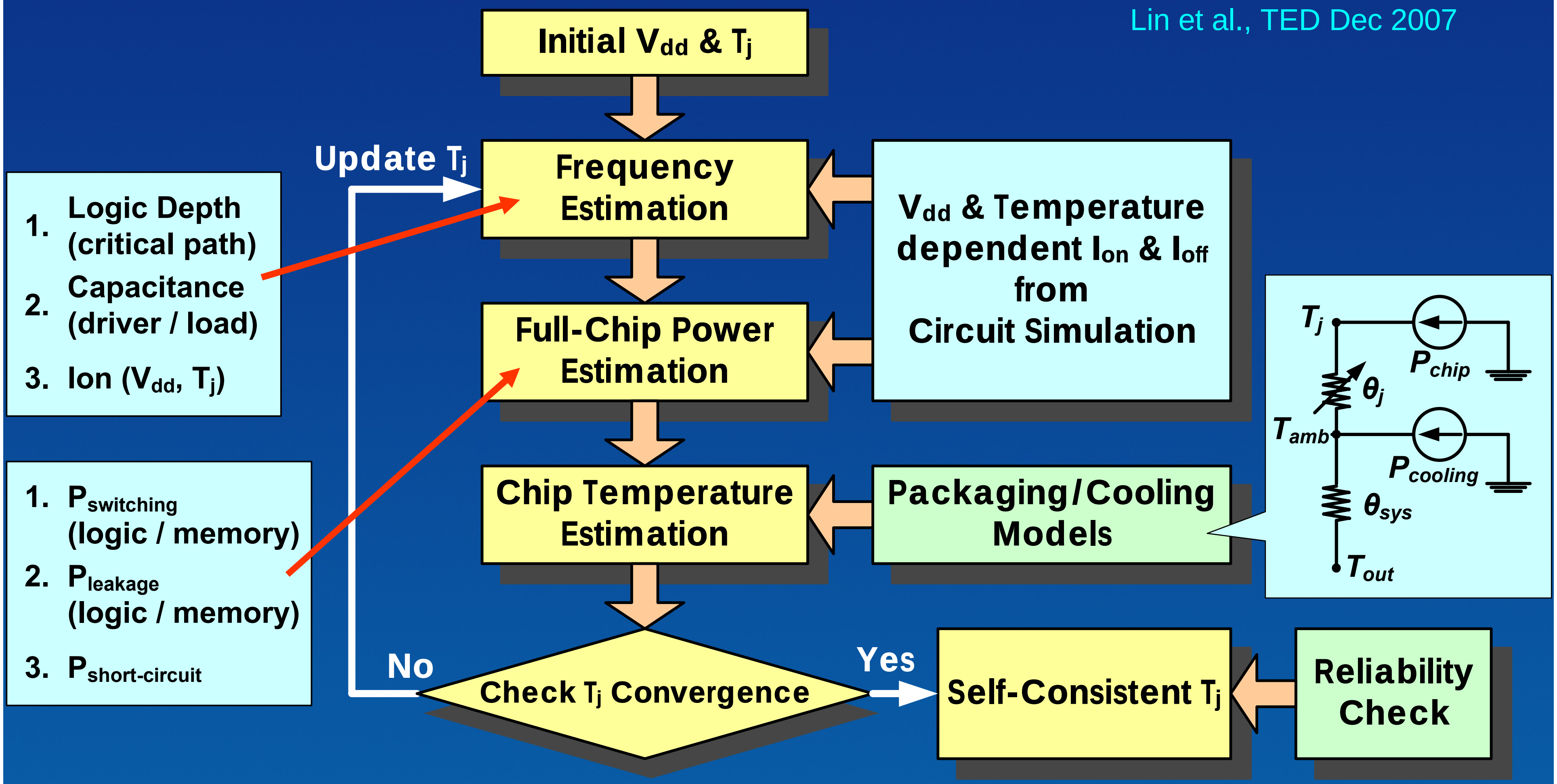
$T_j = f_4(P_{total}, T_{amb}, \theta_{ja})$

Banerjee et al., IEDM 2003

Positive feedback between temperature, leakage and total power leads to electrothermal couplings, which had been inconspicuous for earlier generation of ICs.

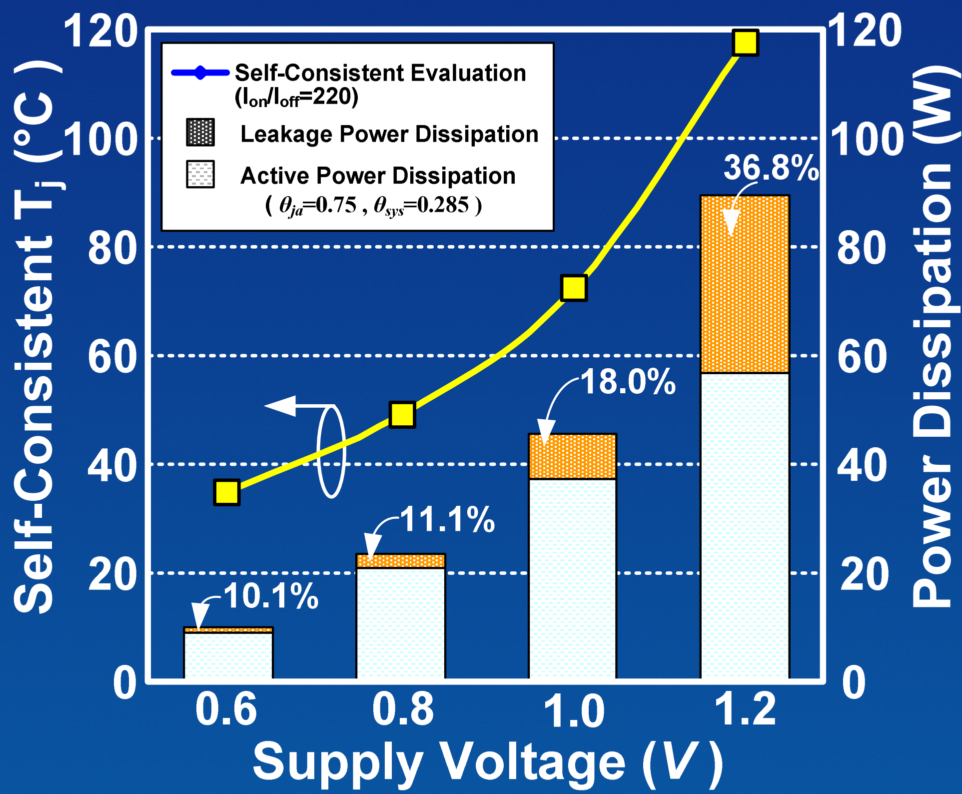
Self-Consistent Methodology

Lin et al., TED Dec 2007



Obtain a self-consistent average junction temperature

Self-Consistent Evaluation



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.

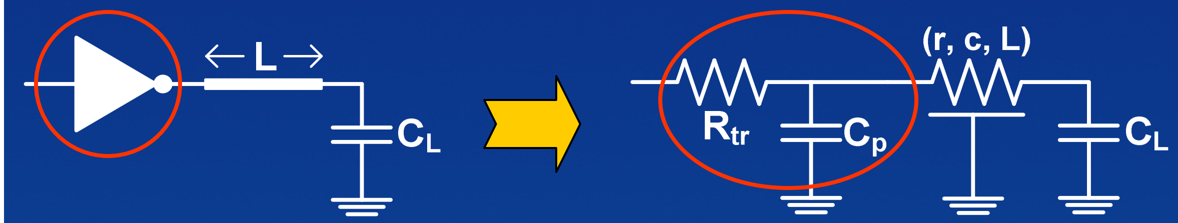
Can be evaluated by Alpha-Power model

T. Sakurai and A. R. Newton, 1990

Energy

Can be calculated by total power dissipation

Incorporate Interconnect Effects in EDP Optimization



$$\tau = R_{tr}(C_p + C_L + cL) + rLC_L + (1/2)rcL^2$$

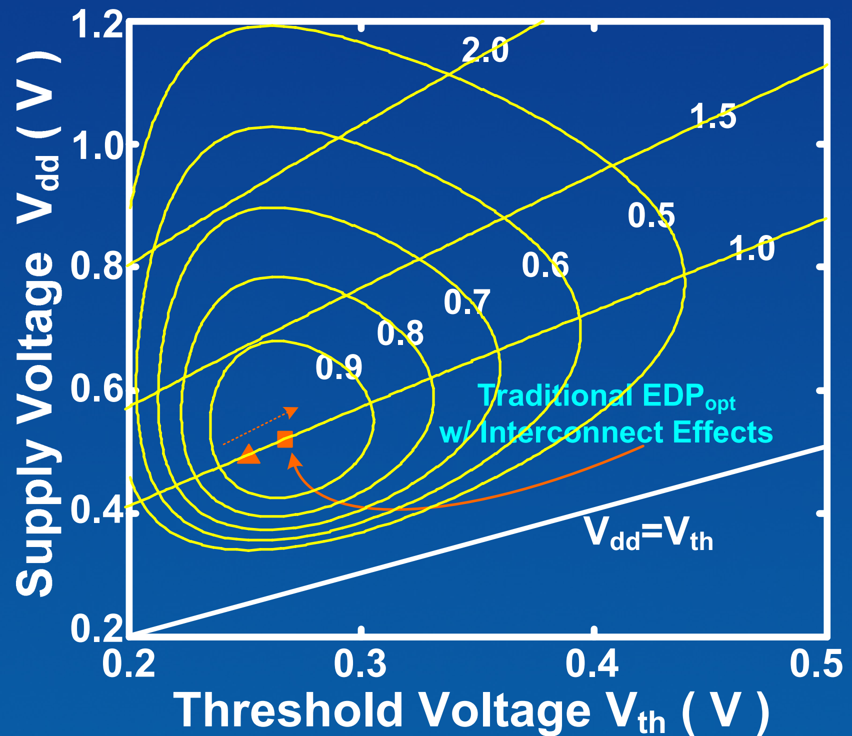
$$T_g \cong 0.69R_{tr}(C_p + C_L + cL) + 0.69rLC_L + 0.38rcL^2$$

$$T_g \cong \frac{KV_{dd}}{(V_{dd} - V_{th})^{\alpha}}\left[1 + \frac{cL}{C_p + C_L}\right] + 0.69rLC_L + 0.38rcL^2$$

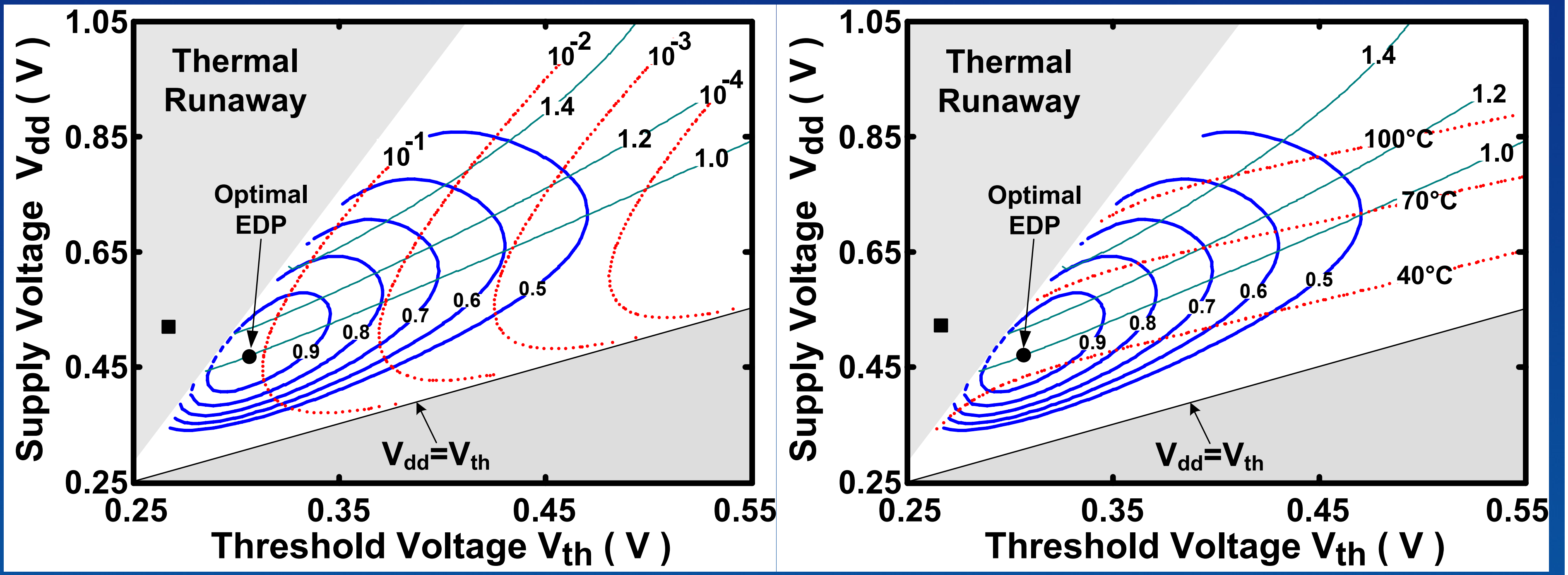
$$P_{dynamic} = aC_{eff}V_{dd}^2 F$$

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

$$Energy = (P_{dynamic} + P_{static}) \cdot T_g L_d$$



Electrothermally-Coupled EDP



Lin et al., ICCD 2005

- 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

Generalized Metric for Optimization

- EDP is not the only metric
- Other Metrics......

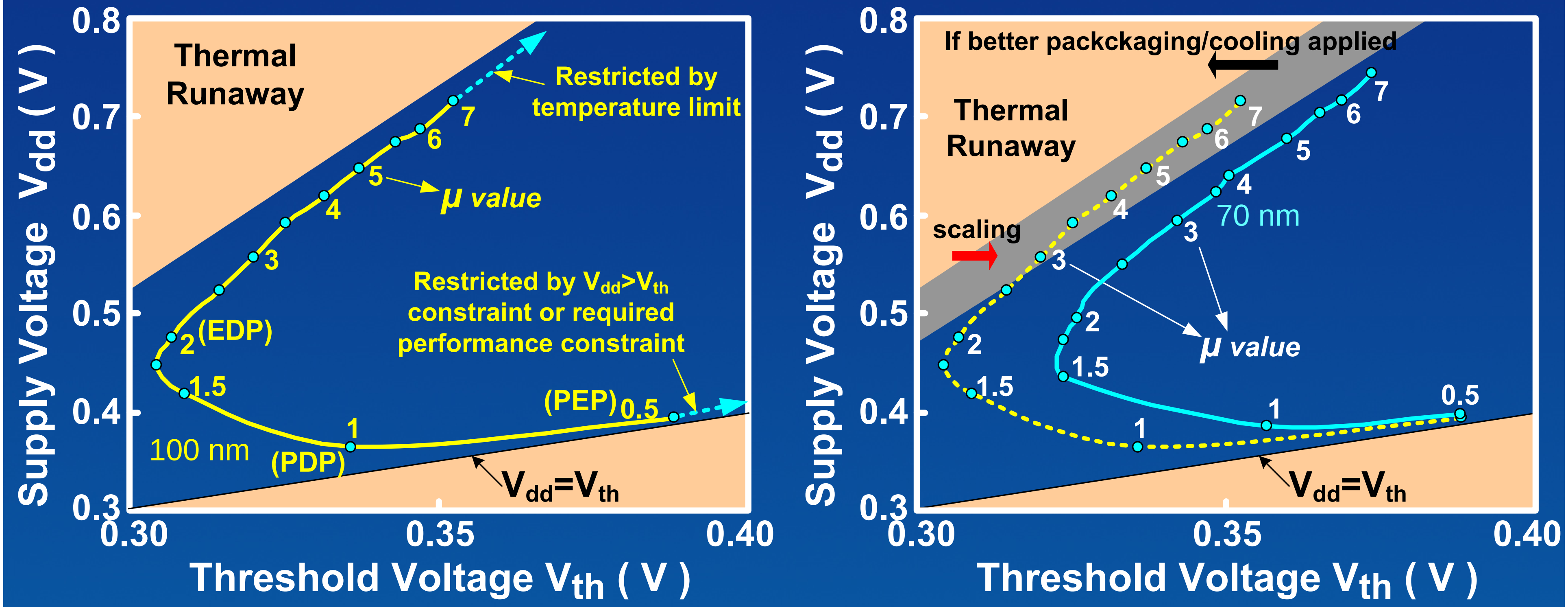
EDP is Energy X Delay = PT^2
PDP is Power X Delay = PT
PEP is Power X Energy = P^2T

Generalized Metric: PT^{μ}

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

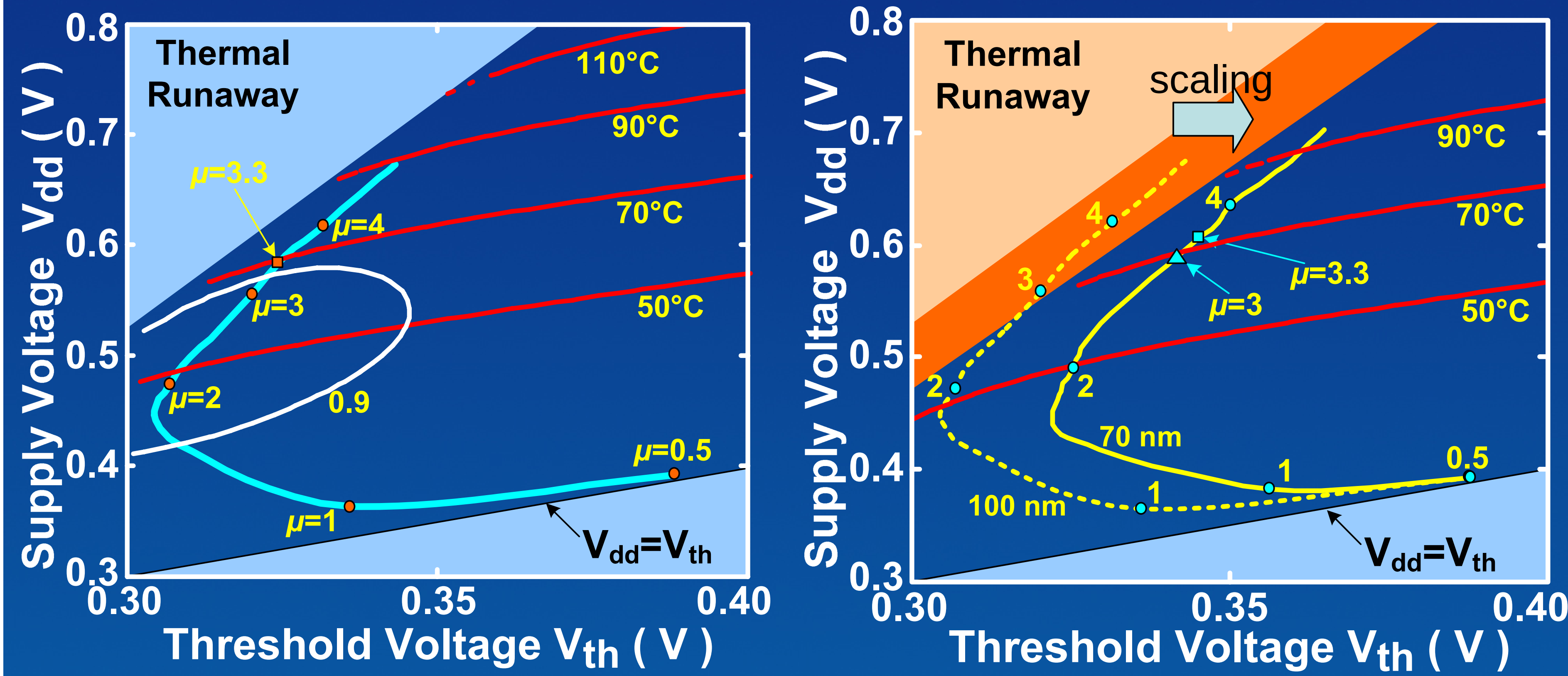
How to choose a design-specific metric ?

Optimal Operation Locus



μ is determined by thermal and performance requirements

Thermal/Packaging Aware Design Optimization



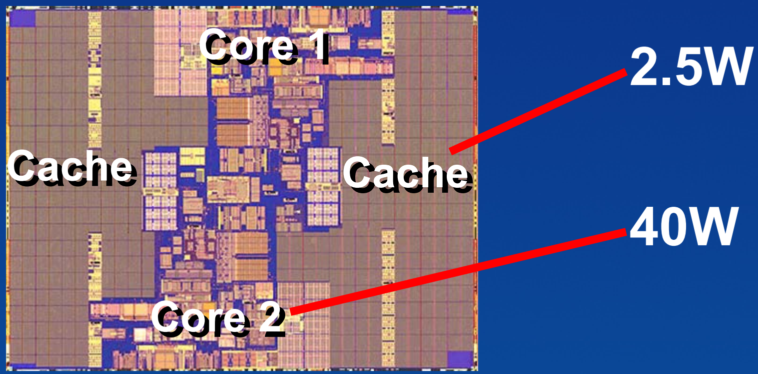
- ❖ Choose an optimization metric to meet thermal/packaging requirements (For T = 70 C, μ = 3.3)
- ❖ Impact of scaling should be taken into account (For T = 70 C, μ =3)

Outline

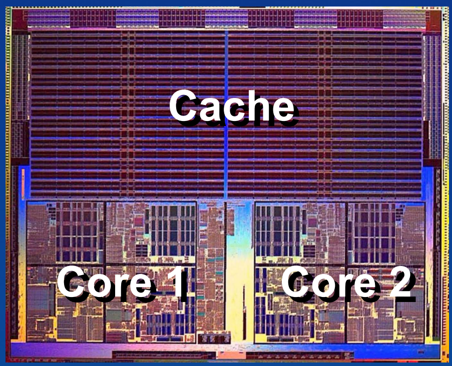
- Introduction
  - Technology Scaling Challenges ✓
  - Power and Thermal Problems ✓
  - Integrated Approach ✓
- Integrated Power-Thermal Management
  - Electrothermal Couplings in Nanoscale CMOS ✓
  - Design-Package Co-Optimization ✓
  - Self-Consistent Temperature Profile Estimation

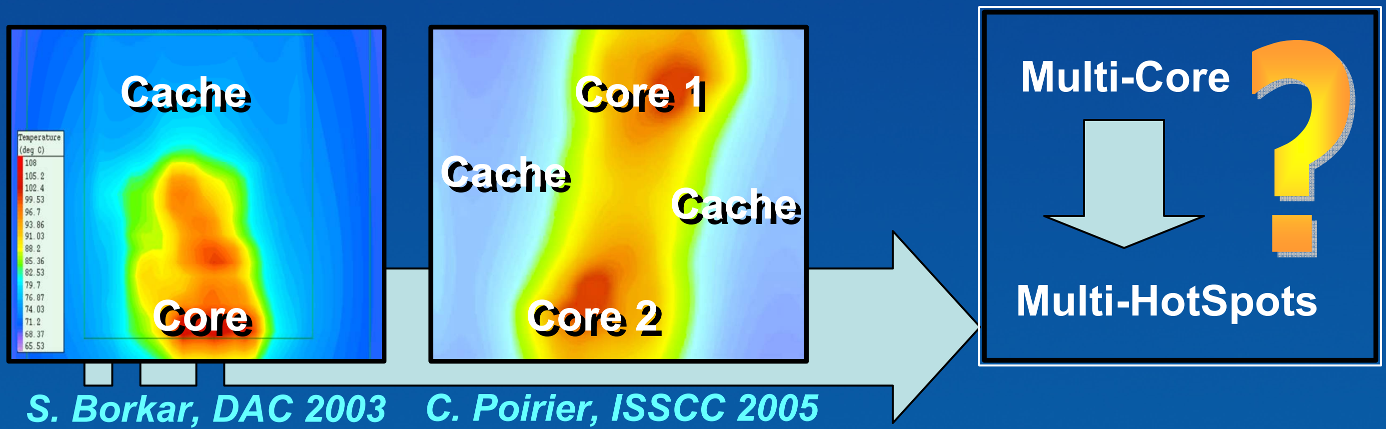
Non-Uniform Power Density

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



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



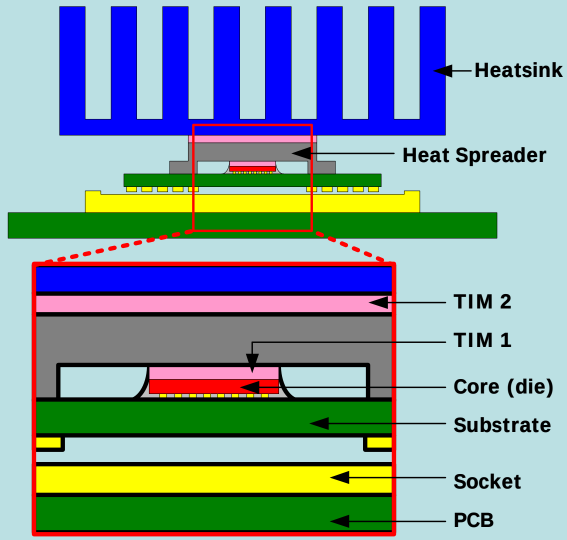


S. Borkar, DAC 2003 *C. Poirier, ISSCC 2005*

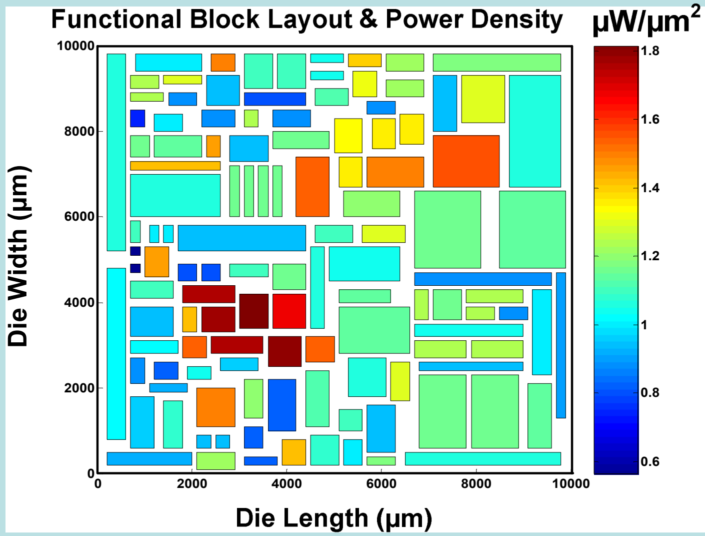
Non-uniform power density generates on-chip temperature gradient

Methodology Overview

Packaging / Cooling Model



Layout Geometry & Power Dissipation



$$\frac{\partial T}{\partial n_i} = \frac{h}{K}\left[T - T_{amb}\right]$$

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

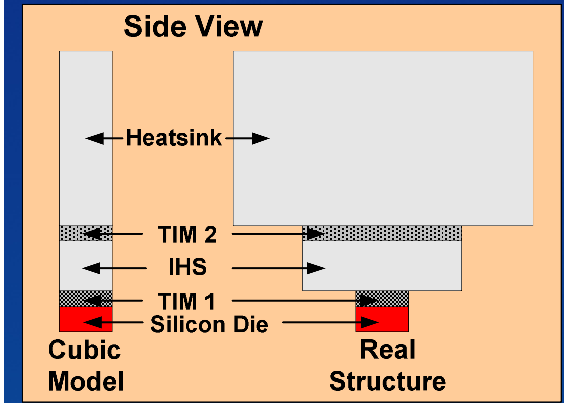
Boundary Condition → Parabolic Heat Partial Differential Equations

Parabolic Heat Partial Differential Equations → 3-D Electrothermally-Aware Spatial Temperature Estimation

Electrothermal Couplings ↔ 3-D Electrothermally-Aware Spatial Temperature Estimation

Numerical Approach

Lin et al., TED Dec 2007



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

$$\frac{M_x T}{\Delta x^2} = \frac{T_{i+1,j,k} - 2T_{i,j,k} + T_{i-1,j,k}}{\Delta x^2}$$

$$\frac{\partial^2 T}{\partial x^2} = \frac{1}{2}\left(\frac{M_x T^{n+1}}{\Delta x^2} + \frac{M_x T^n}{\Delta x^2}\right)$$

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

Crank-Nicholson Implicit Method

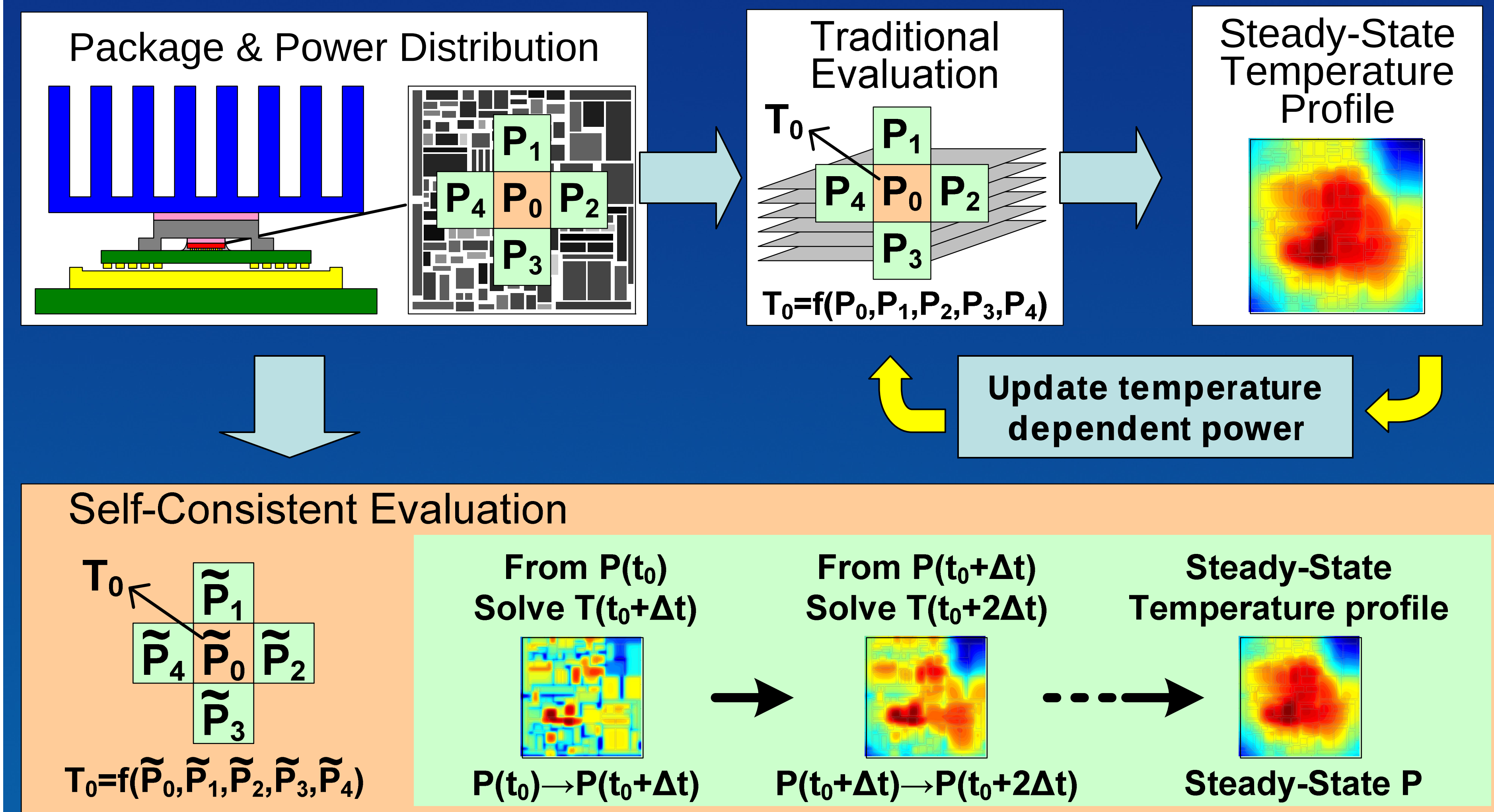
$$\frac{T^{n+1} - T^n}{\Delta t} = \left(\frac{k}{\rho C_p}\right)\left[\left(\frac{M_x}{2\Delta x^2} + \frac{M_y}{2\Delta y^2} + \frac{M_z}{2\Delta z^2}\right)\left(T^{n+1} + T^n\right)\right] + \frac{p}{\rho C_p}$$

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

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

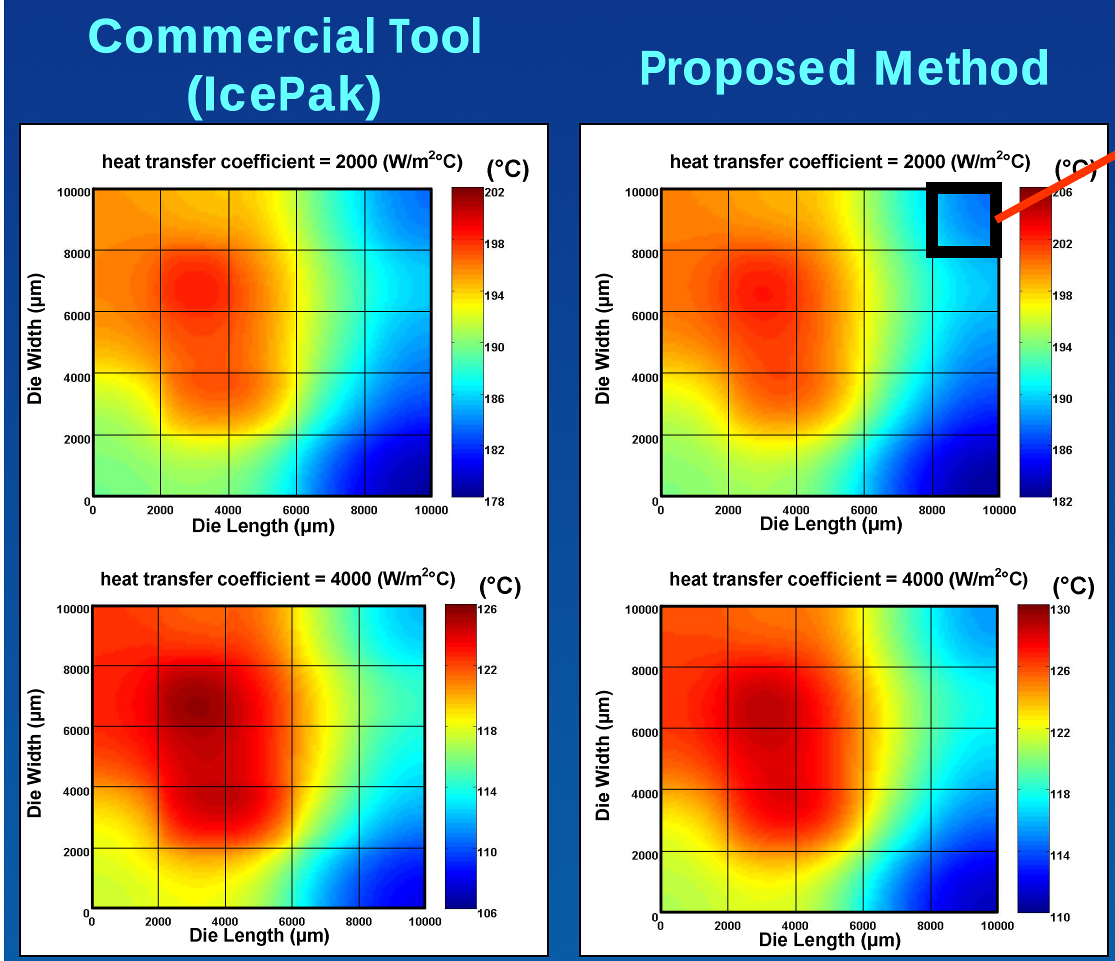
Linear Equations: $(1 - \beta_x M_x) U_{x,y,z} = V_{x,y,z}$

Self-Consistent Evaluation



Methodology Verification

For a case with negligible leakage

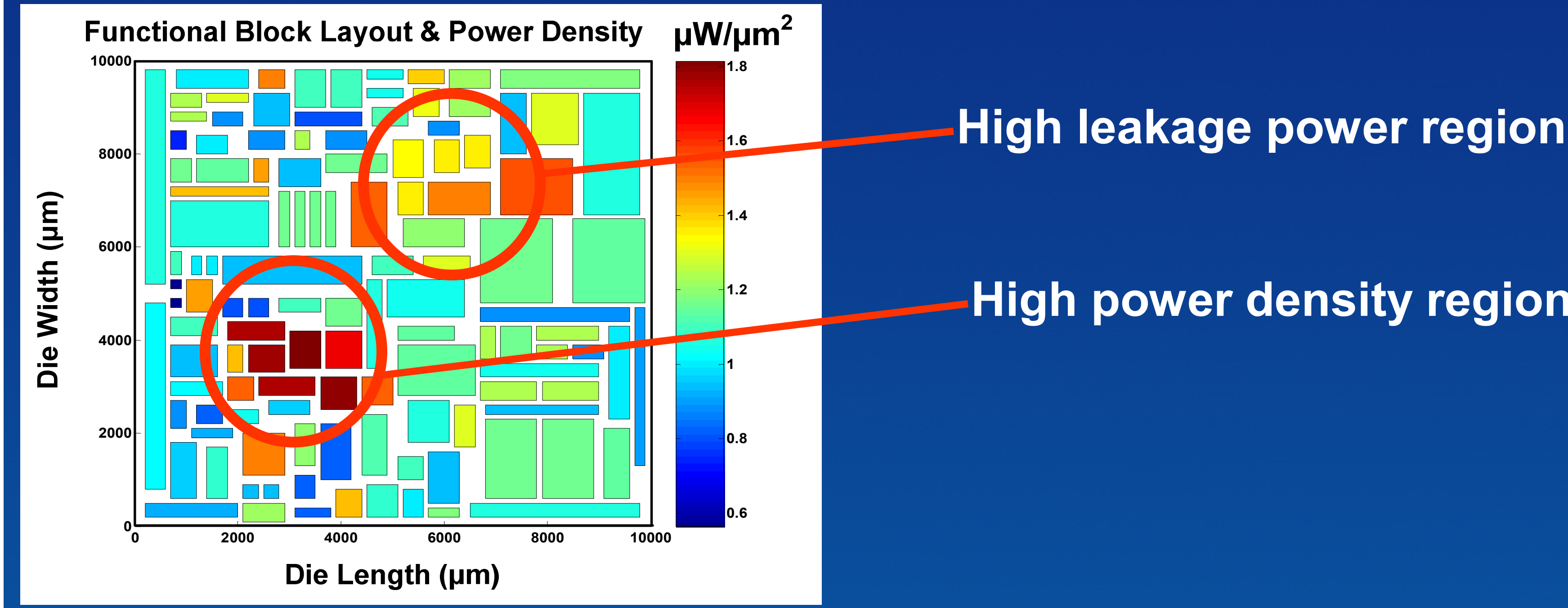


Equally select 400 data points from identical locations for comparison (total = 10000 pts)

Coefficient	Max. Deviation
h=1000	2.13 %
h=2000	2.48%
h=3000	3.26 %
h=4000	3.28 %
h=5000	3.06 %

Proposed method is validated against an industrial-quality software.

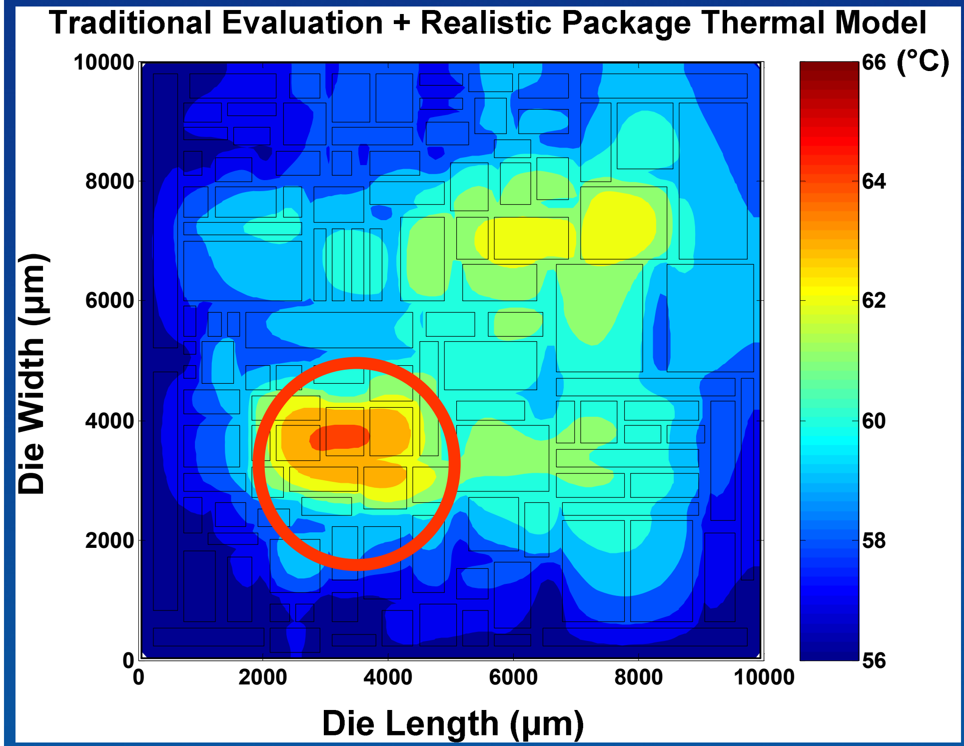
Implementation & Setup



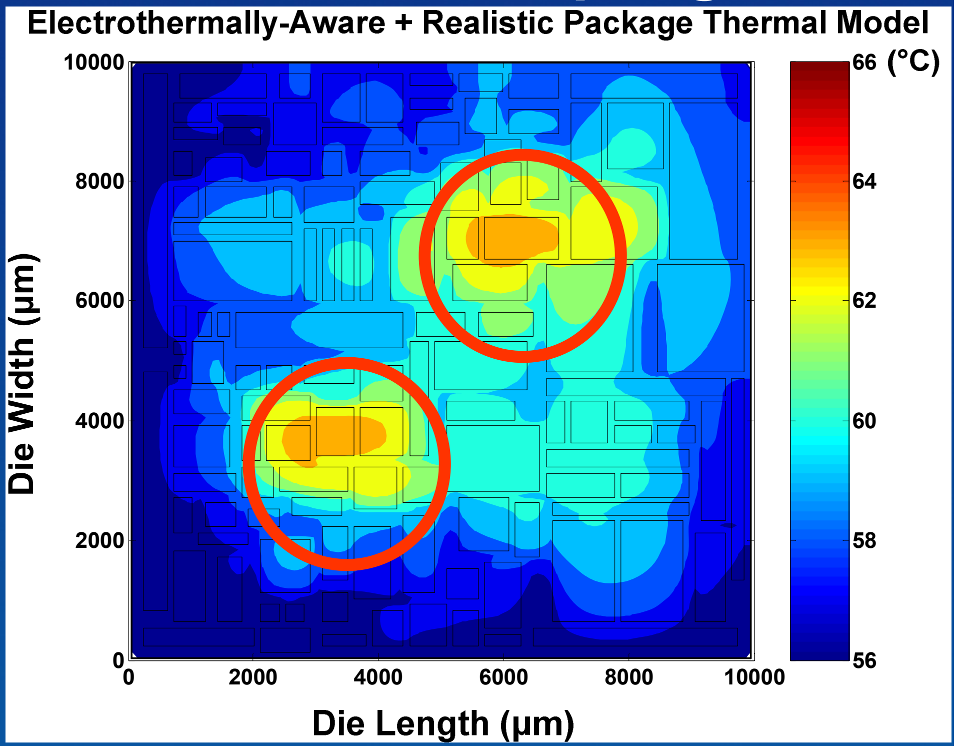
Layer	Area (mm^2)	Thickness (mm)	Specific Heat (J/kg°C)	Density (kg/m^3)	K (W/m°C)
Die	10 X 10	0.8	712	2330	120
TIM 1	10 X 10	0.2	230	7310	30
IHS	30 X 30	1.8	385	8930	390
TIM 2	30 X 30	0.2	2890	900	6.4
Heatsink	60 X 60	6.4	385	8930	360

Impact of Considering ET-couplings

Without ET-Couplings



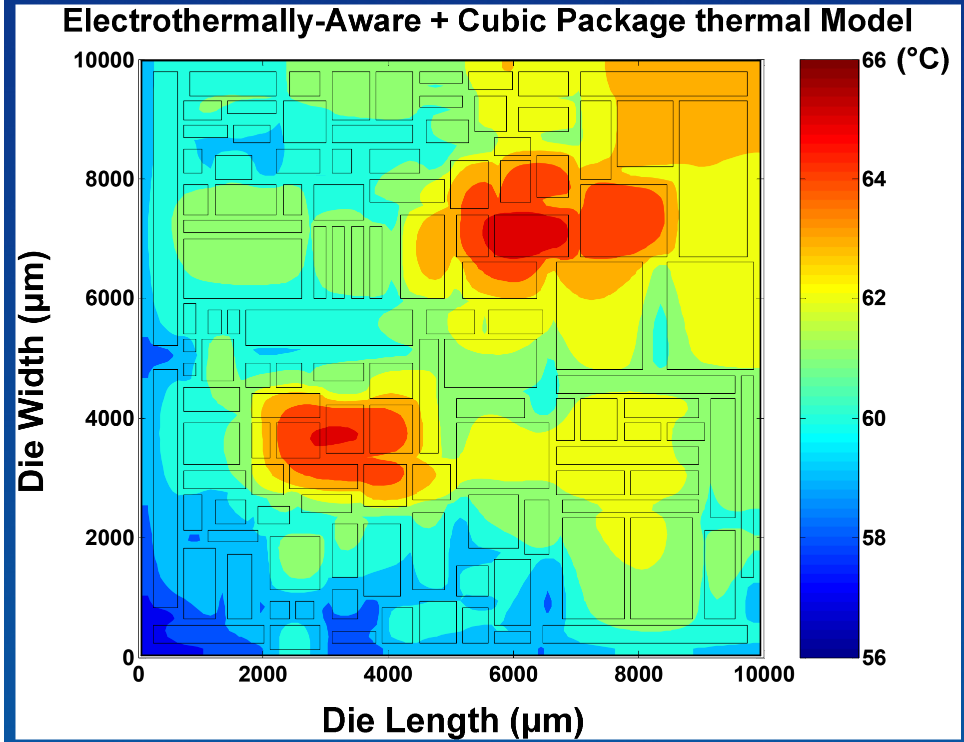
With ET-Couplings



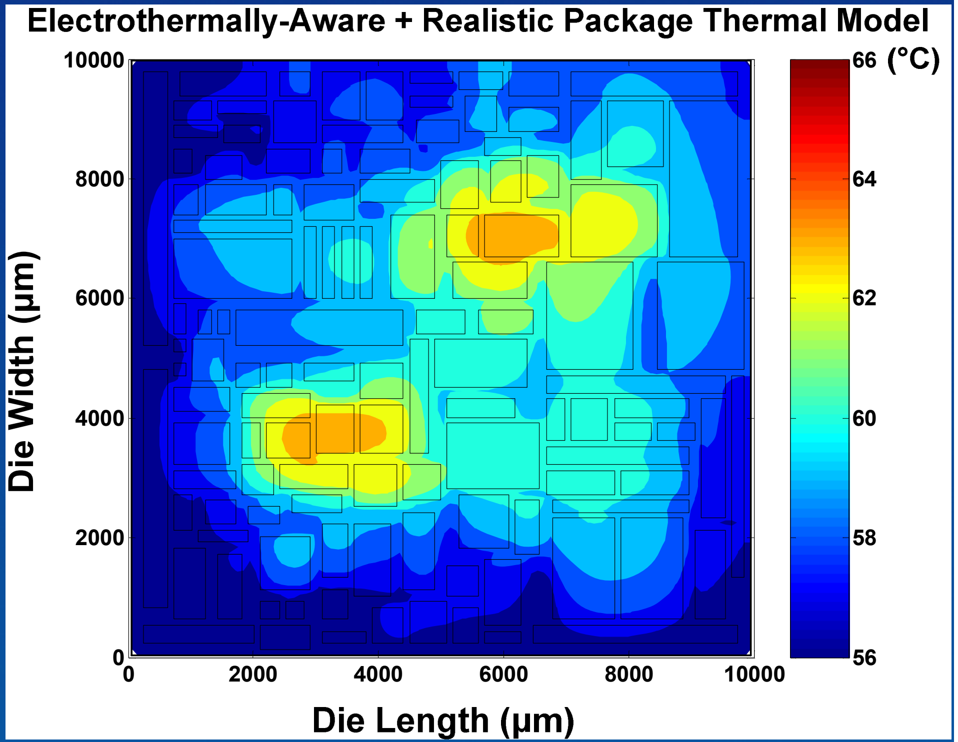
- Traditional evaluation neglects electrothermal couplings
- Considering electrothermal couplings leads to an additional hot-spot at the highest leakage power region

Impact of Package Models

Cubic Package Thermal Model

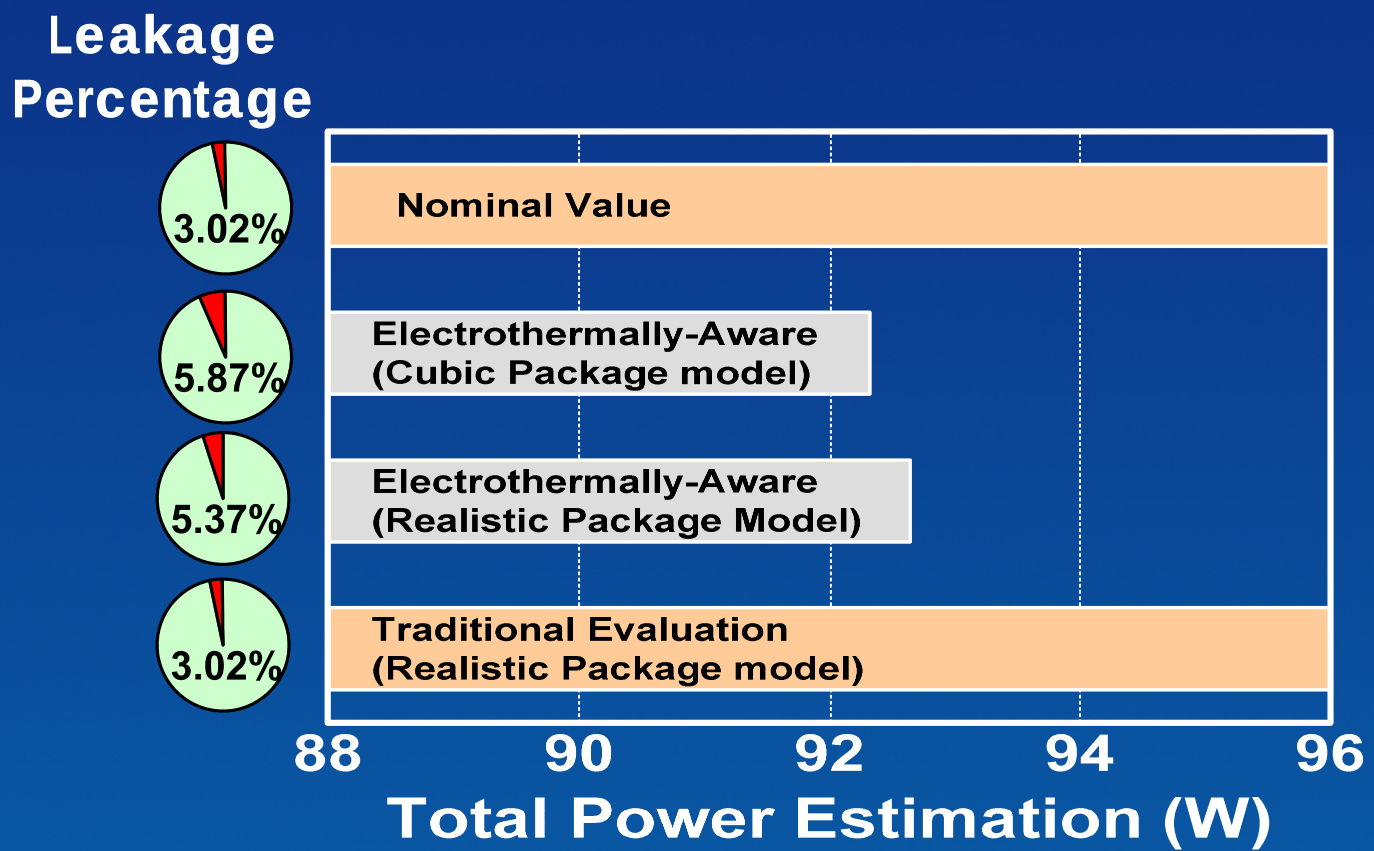


Realistic Package Thermal Model



- 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