

ECE 225

High-Speed Digital IC Design

Lecture 9

Nanoscale Power and Thermal Management:

Design-Specific Metric Optimization

Self-consistent Thermal Profile Estimation

IC Cooling Analysis

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



Design-Specific Metric Optimization

S-C. Lin and K. Banerjee, "A Design-Specific and Thermally-Aware Methodology for Trading-Off Power and Performance in Leakage-Dominant CMOS Technologies," TVLSI Vol. 16, No. 11, pp. 1488-1498, 2008.

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

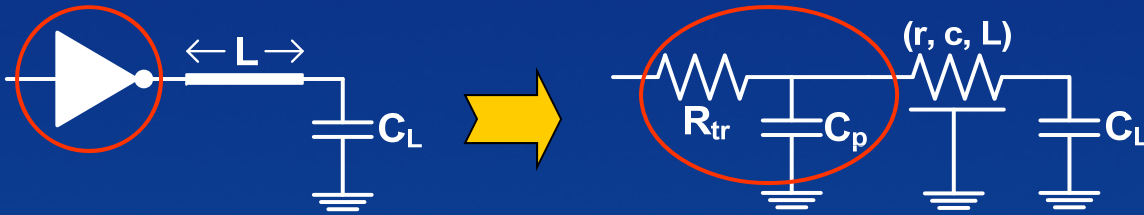
T. Sakurai and A. R. Newton, 1990

Energy

Can be calculated by total power dissipation



Incorporate Interconnect Effects in EDP Optimization



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

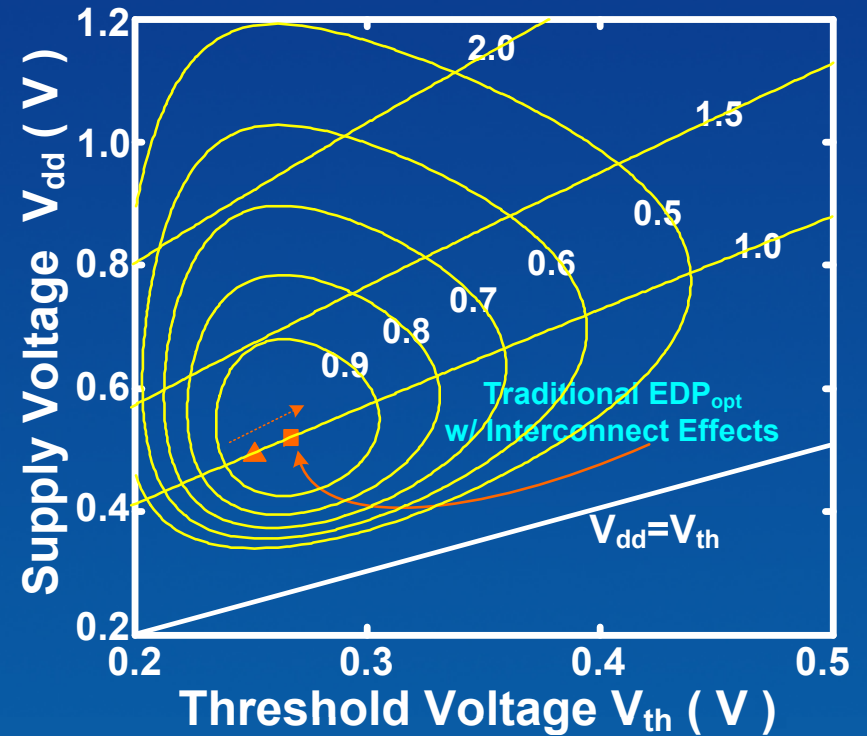
$$T_g \cong 0.69R_{tr}(C_p + C_L + cL) + 0.69rLC_L + 0.38rcL^2$$

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

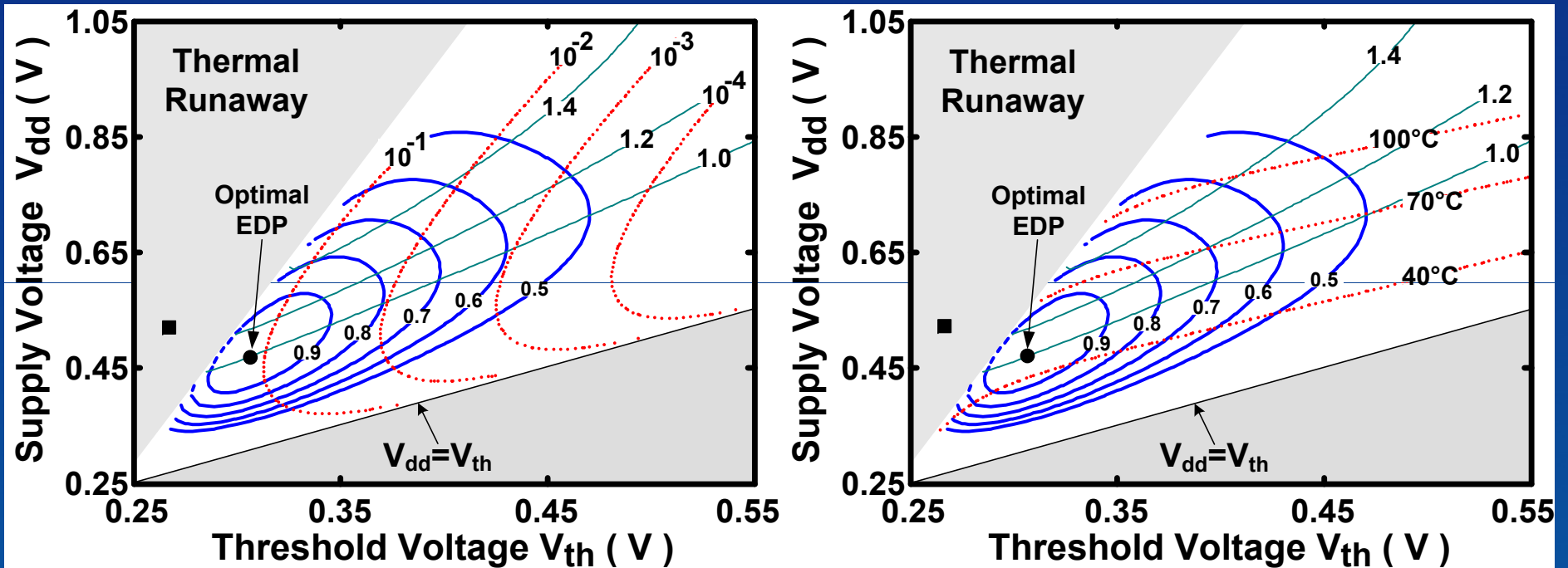
$$P_{dynamic} = aC_{eff}V_{dd}^2F$$

$$P_{static} = I_s e^{-V_{th}/\gamma V_0} (1 - e^{-V_{ds}/\gamma V_0}) W_{eff} V_{dd}$$

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



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

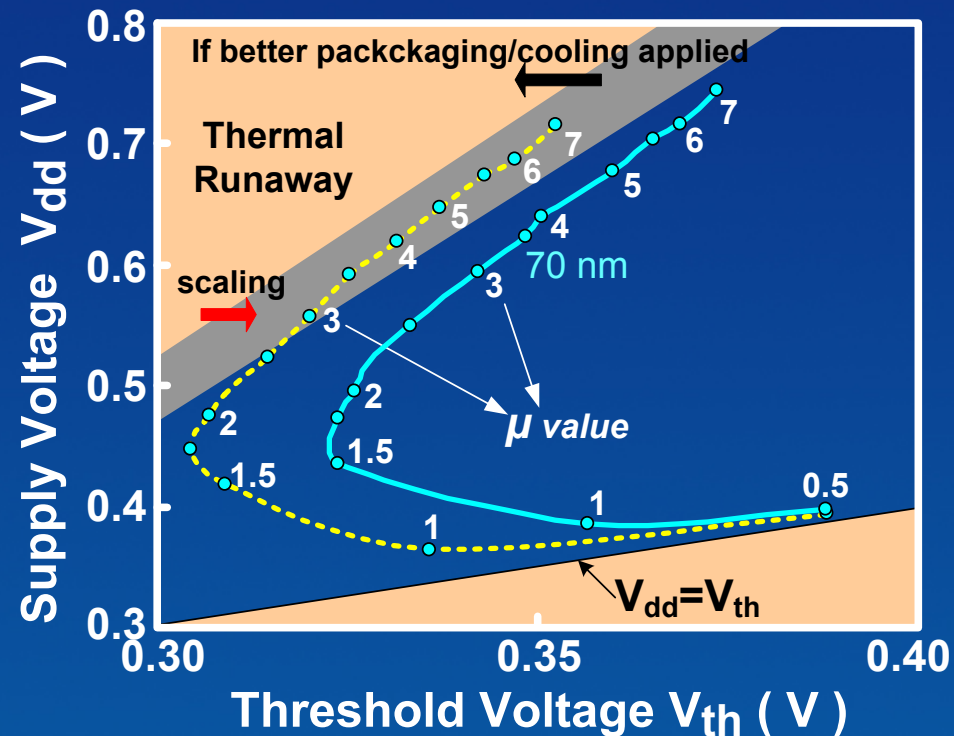
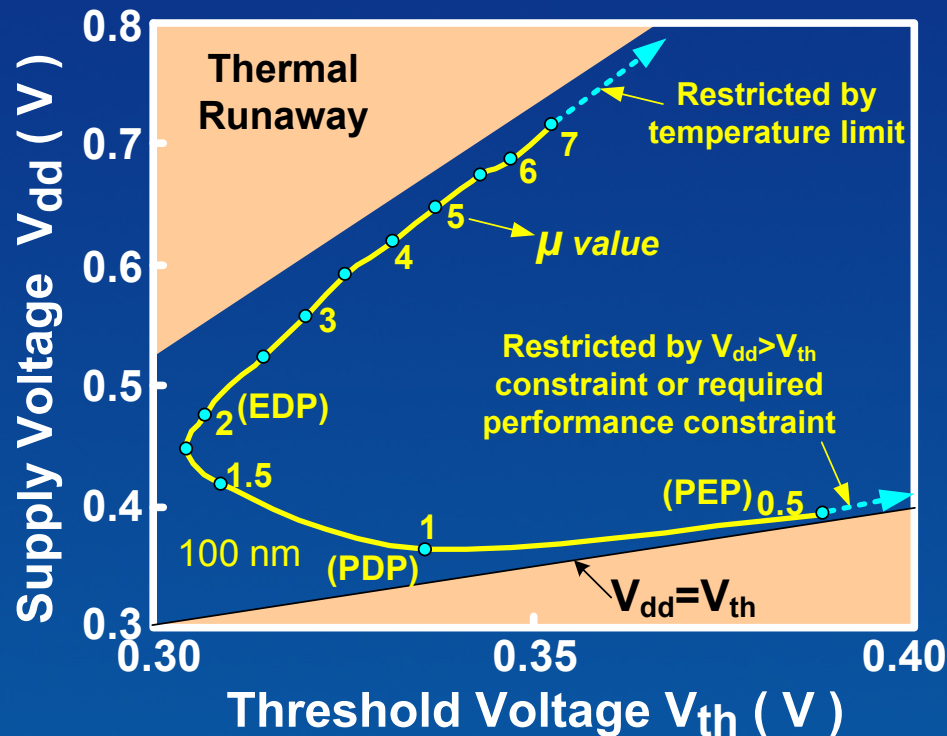


Generalized Metric: PT^μ

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

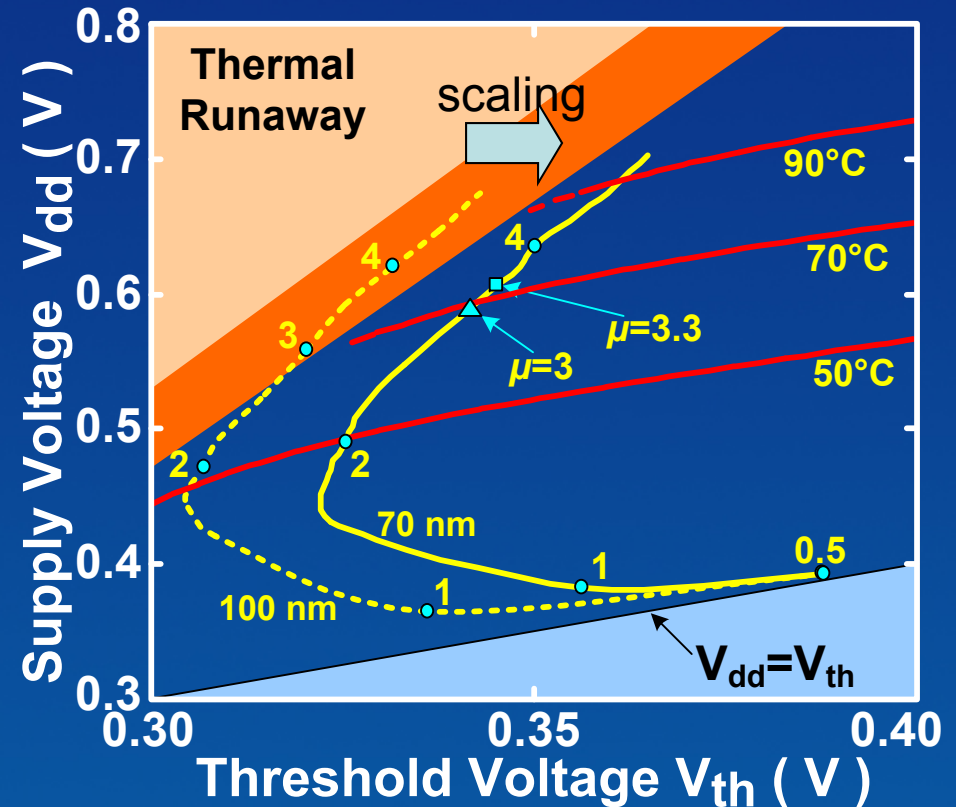
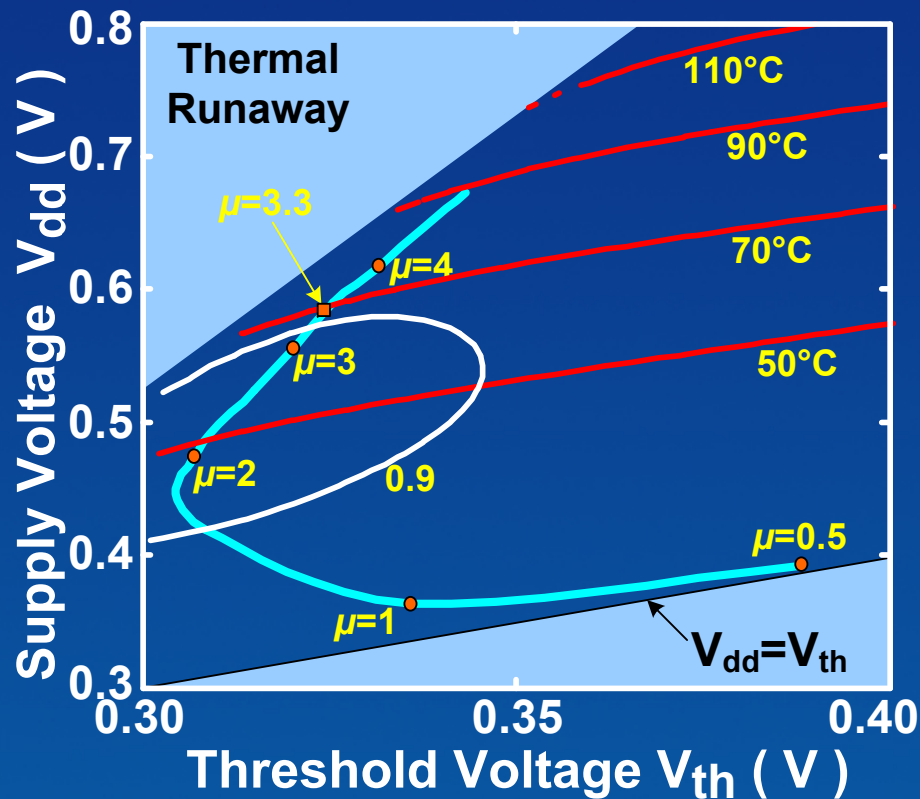
How to choose a design-specific metric ?

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, $\mu = 3.3$)
- ❖ Impact of scaling should be taken into account (For $T = 70$ C, $\mu = 3$)

Self-Consistent Temperature Profile Estimation

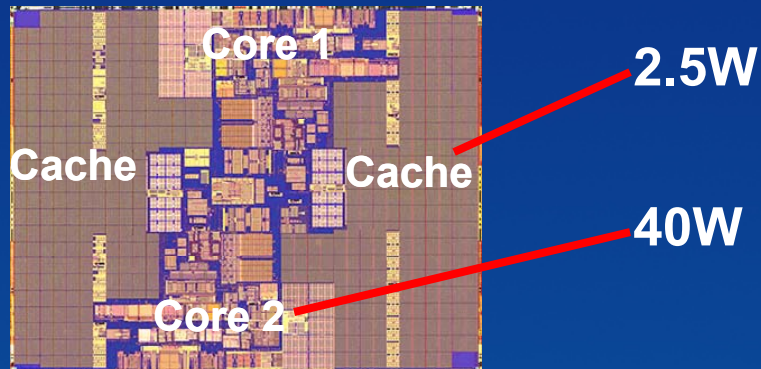
S-C. Lin Lin, G. Chrysler, R. Mahajan, V. De and K. Banerjee, "A Self-Consistent Substrate Thermal Profile Estimation Technique for Nanoscale ICs—Part I: Electrothermal Couplings and Full-Chip Package Thermal Model," TED Vol. 54, No. 12, pp. 3342-3350, 2007.

S-C. Lin Lin, G. Chrysler, R. Mahajan, V. De and K. Banerjee, "A Self-Consistent Substrate Thermal Profile Estimation Technique for Nanoscale Part II: Implementation and Implications for Power Estimation and Thermal Management," TED Vol. 54, No. 12, pp. 3351-3360, 2007 .

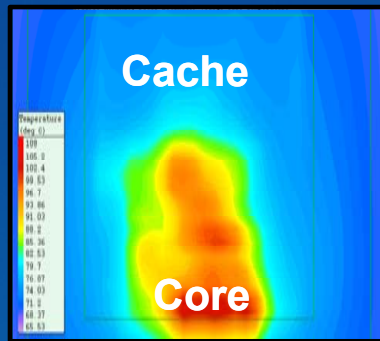
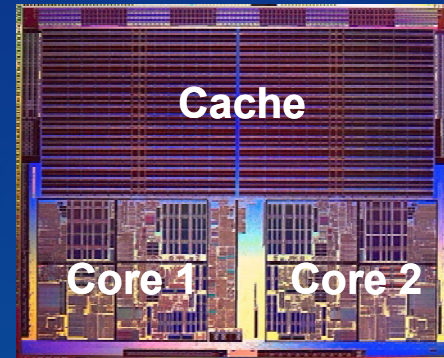


Non-Uniform Power Density

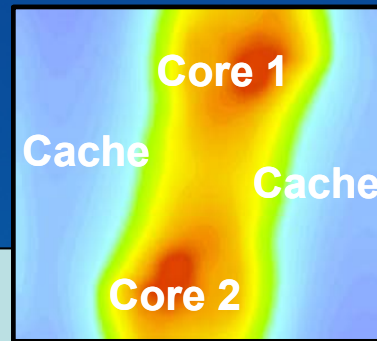
*Intel's first dual-core
Itanium-Montecito 90nm*



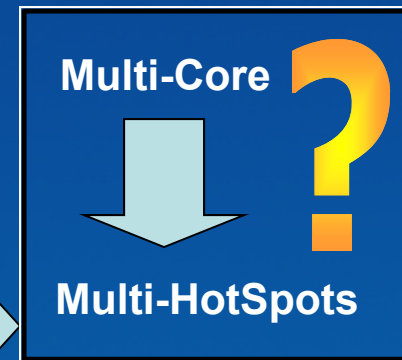
*AMD's first dual-core
Athlon 64 X2 90nm SOI*



S. Borkar, DAC 2003



C. Poirier, ISSCC 2005

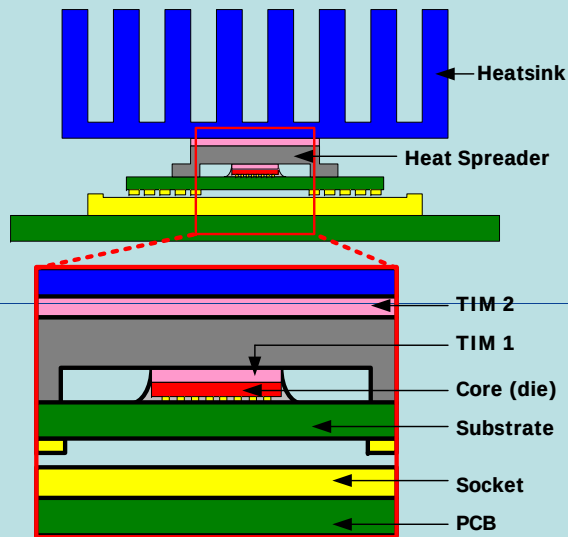


Non-uniform power density generates on-chip temperature gradient

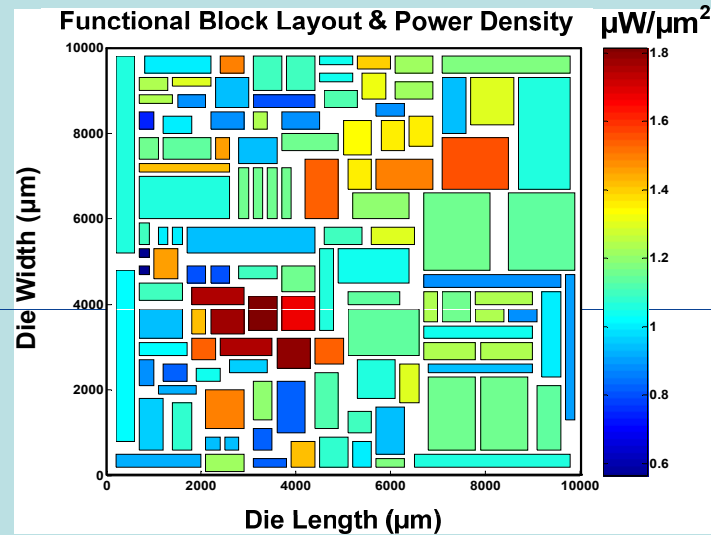


Methodology Overview

Packaging / Cooling Model



Layout Geometry & Power Dissipation



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

Boundary Condition

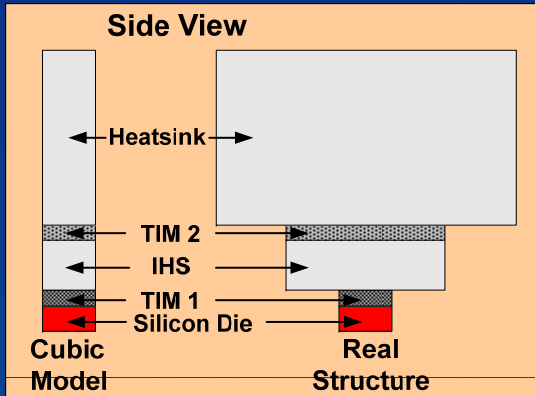
$$\frac{\partial T}{\partial t} = \left(\frac{K}{\rho C_p} \right) \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) + \frac{p}{\rho C_p}$$

Parabolic Heat
Partial Differential Equations

Electrothermal
Couplings

3-D Electrothermally-Aware
Spatial Temperature Estimation

Numerical Approach



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

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

Crank-Nicholson
Implicit Method

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

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

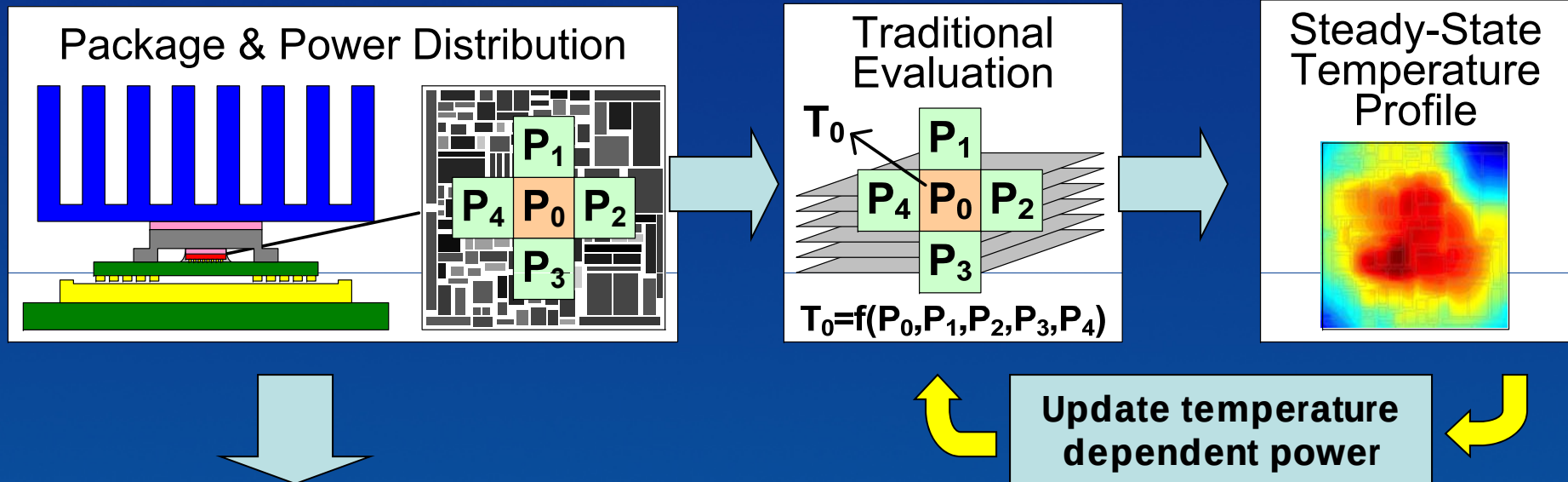
$$\frac{T^{n+1} - T^n}{\Delta t} = \left(\frac{k}{\rho C_p} \right) \left[\left(\frac{M_x}{2 \Delta x^2} + \frac{M_y}{2 \Delta y^2} + \frac{M_z}{2 \Delta z^2} \right) (T^{n+1} + T^n) \right] + \frac{p}{\rho C_p}$$

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

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

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

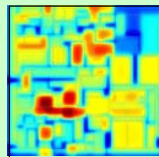
Self-Consistent Evaluation



Self-Consistent Evaluation

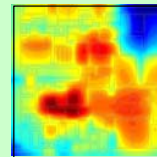
$T_0 = f(\tilde{P}_0, \tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4)$

From $P(t_0)$
Solve $T(t_0 + \Delta t)$



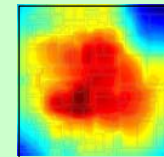
$P(t_0) \rightarrow P(t_0 + \Delta t)$

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



$P(t_0 + \Delta t) \rightarrow P(t_0 + 2\Delta t)$

Steady-State
Temperature profile

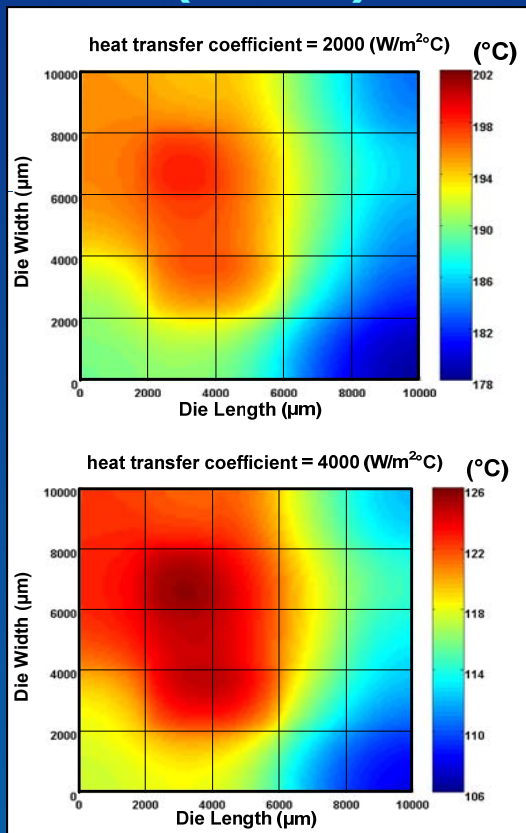


Steady-State P

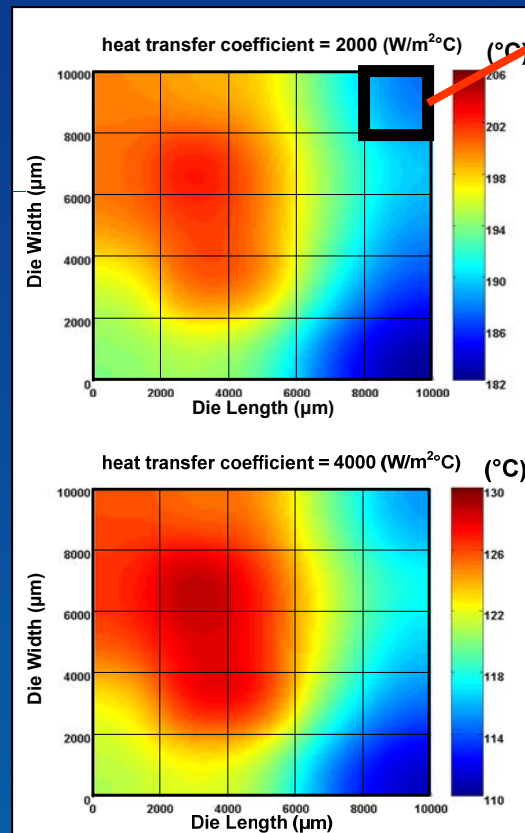
Methodology Verification

For a case with negligible leakage

Commercial Tool
(IcePak)



Proposed Method

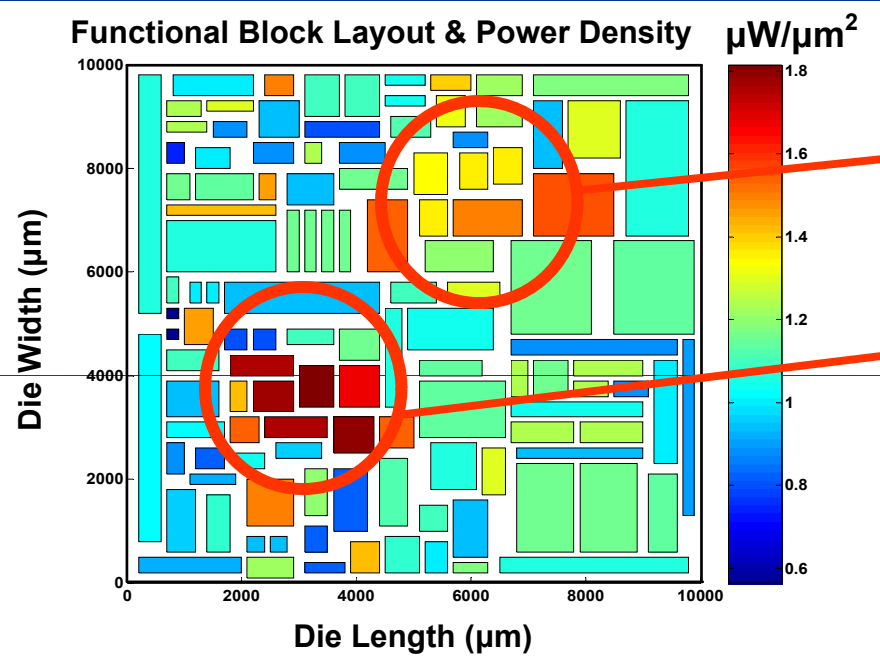


Equally select 400 data points
from identical locations for
comparison (total = 10000 pts)

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



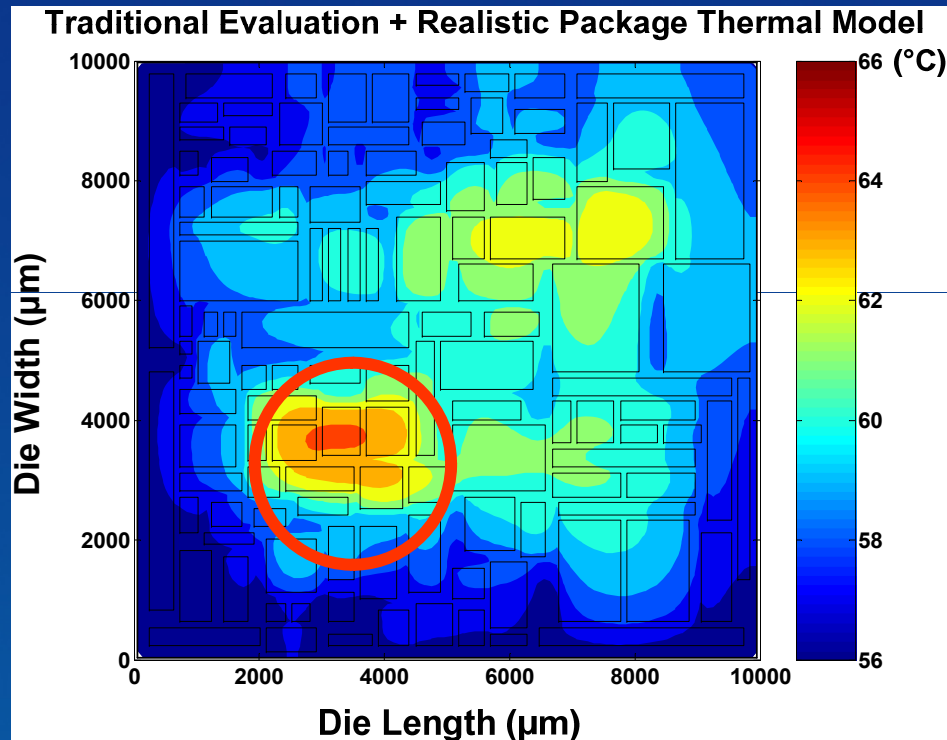
High leakage power region

High power density region

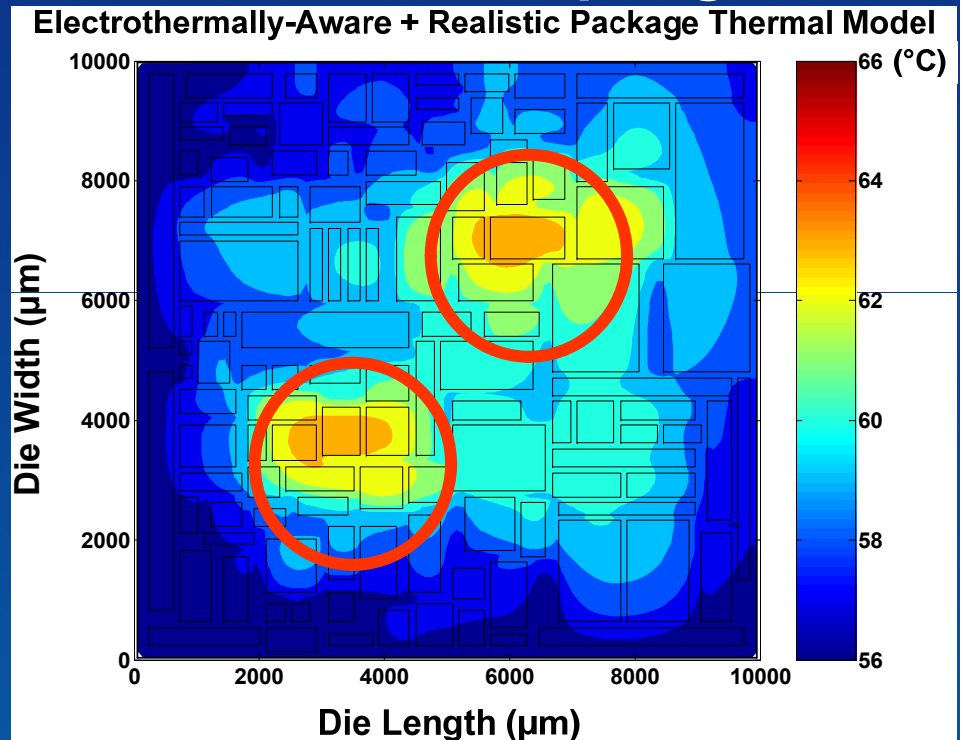
Layer	Area (mm^2)	Thickness (mm)	Specific Heat ($\text{J}/\text{kg}^\circ\text{C}$)	Density (kg/m^3)	K ($\text{W}/\text{m}^\circ\text{C}$)
Die	10 X 10	0.8	712	2330	120
TIM 1	10 X 10	0.2	230	7310	30
IHS	30 X 30	1.8	385	8930	390
TIM 2	30 X 30	0.2	2890	900	6.4
Heatsink	60 X 60	6.4	385	8930	360

Impact of Considering ET-couplings

Without ET-Couplings



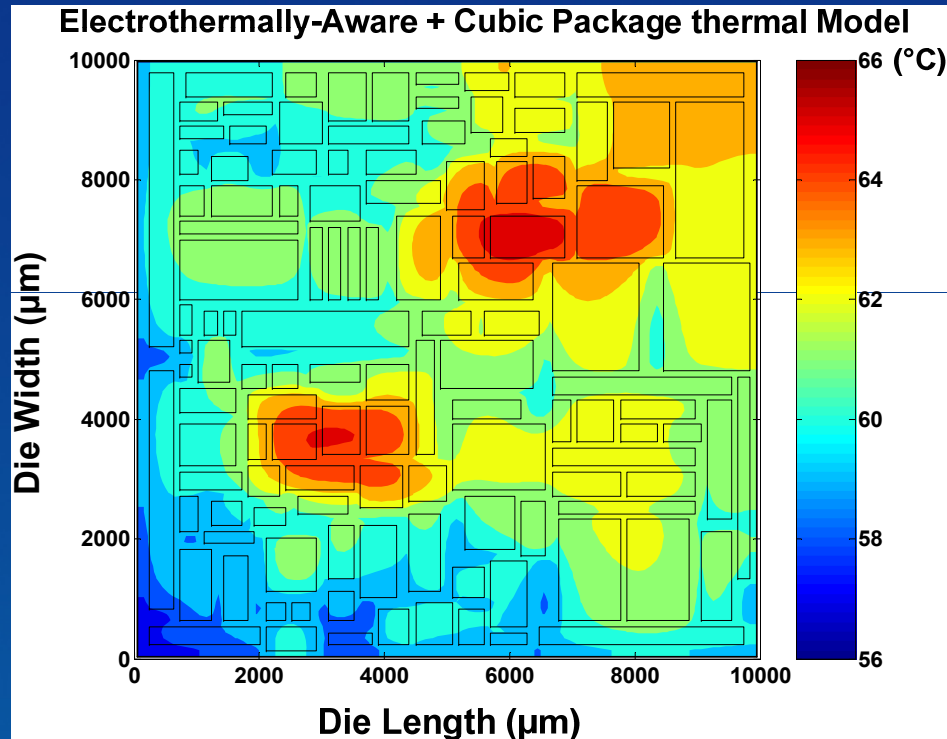
With ET-Couplings



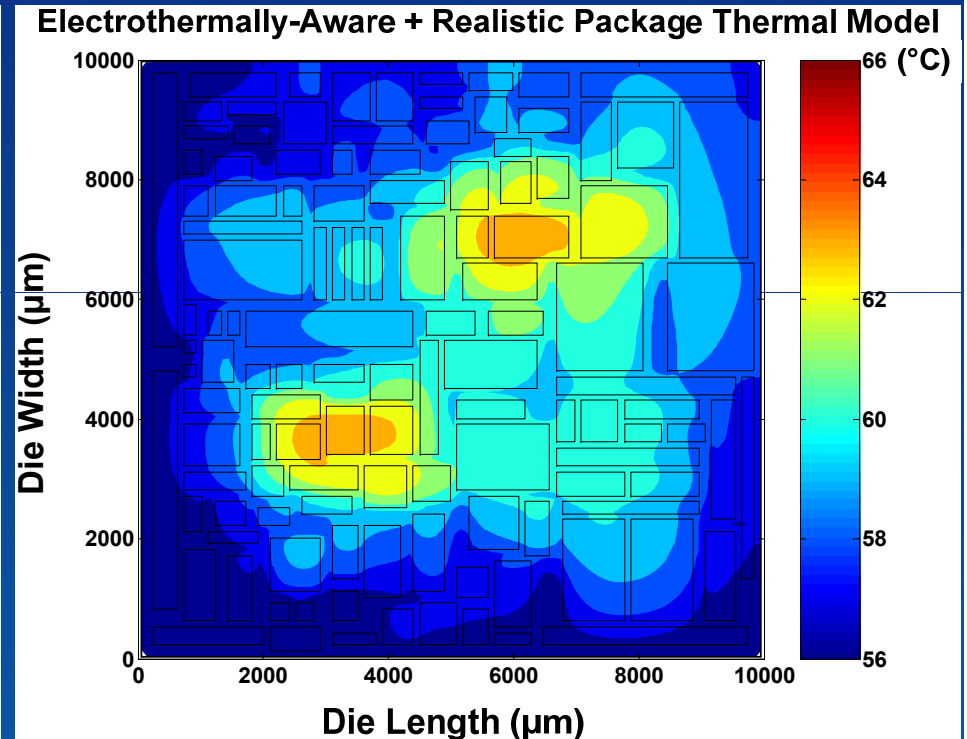
- ❖ Traditional evaluation neglects electrothermal couplings
- ❖ Considering electrothermal couplings leads to an additional hot-spot at the highest leakage power region

Impact of Package Models

Cubic Package Thermal Model



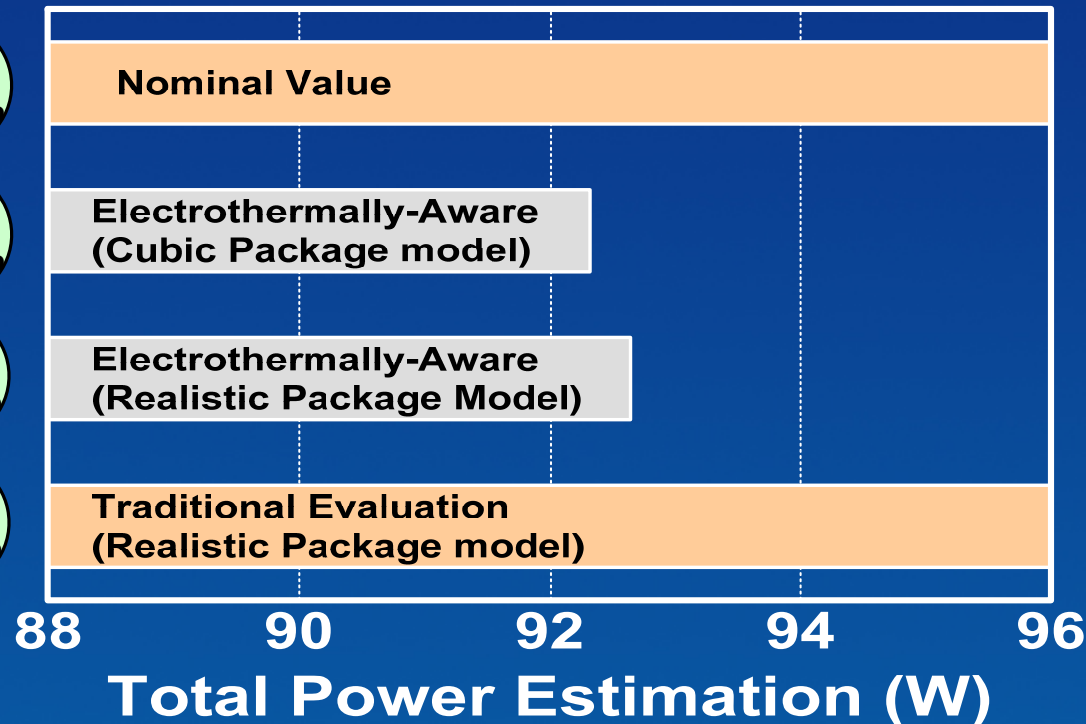
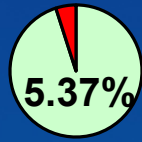
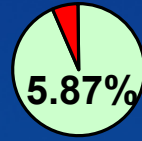
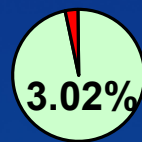
Realistic Package Thermal Model



- ❖ 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

Leakage
Percentage



- ❖ Power is not consistent with temperature in traditional evaluation
- ❖ Considering cubic package model overestimates leakage power

IC Cooling for Power/Thermal Management?

S-C. Lin and K. Banerjee, "Cool Chips: Opportunities and Implications for Power and Thermal Management," TED Vol. 55, No. 1, pp. 245-255, Jan 2008.

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**

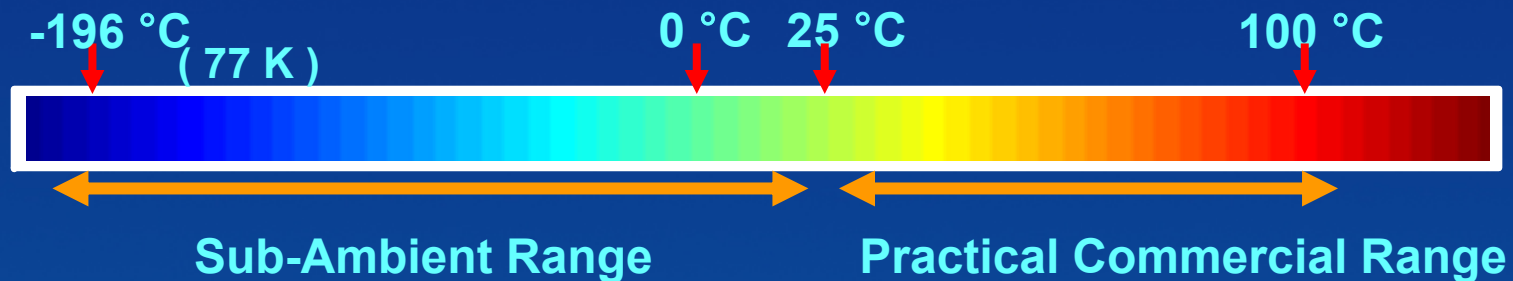
❖ Micro-Architecture Level

- ❖ **Multi-Core**



Does IC Cooling Make Sense?

❖ Range of IC Cooling



❖ Prior Research on Cooling

❖ Analysis under Cryogenic Condition

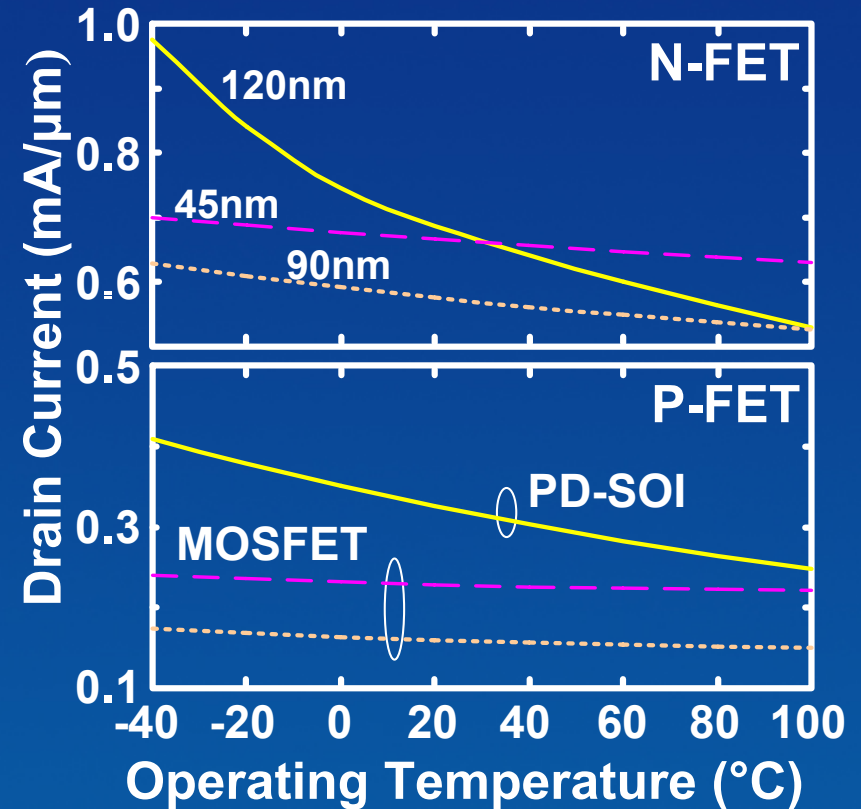
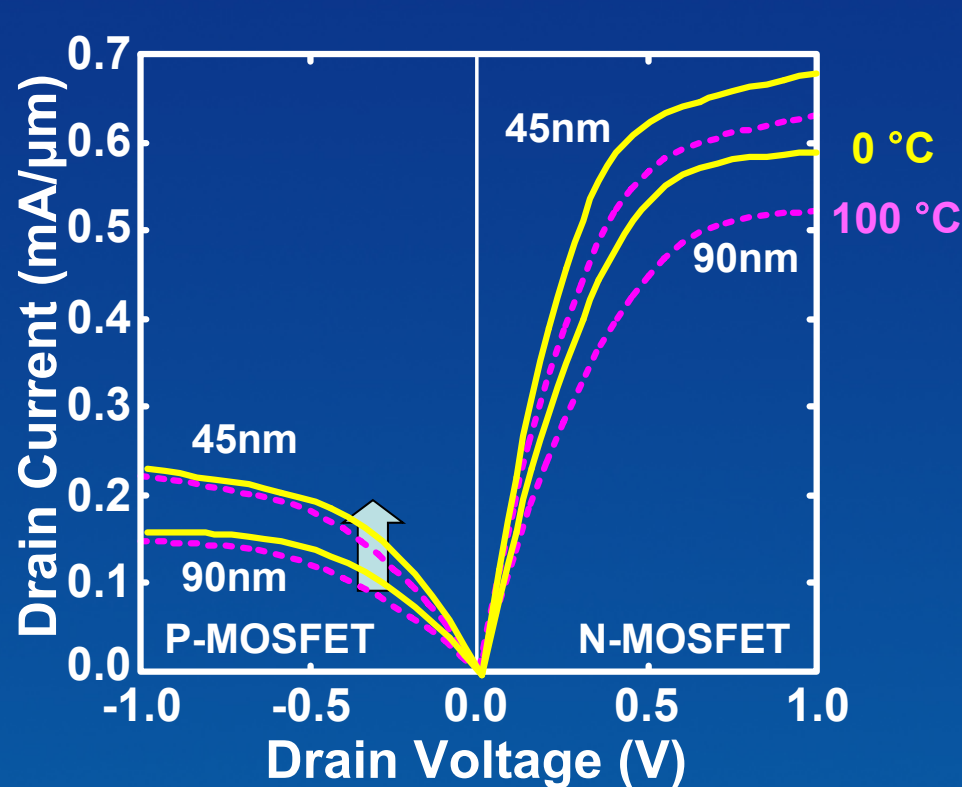
- Operating near liquid nitrogen temperature (77 K) requires complex cooling facilities
- NOT for commercial purpose

❖ Analysis of Performance under Cooling

- Does NOT consider electrothermal couplings
- Does NOT consider the system level power dissipation

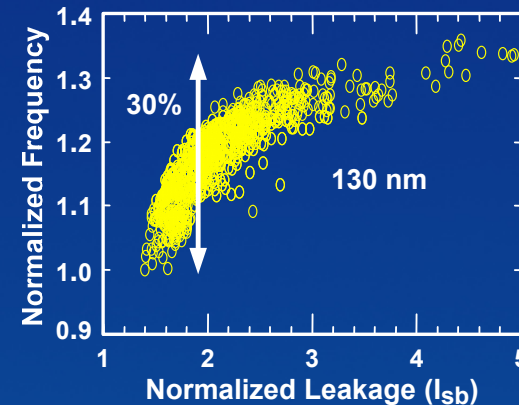
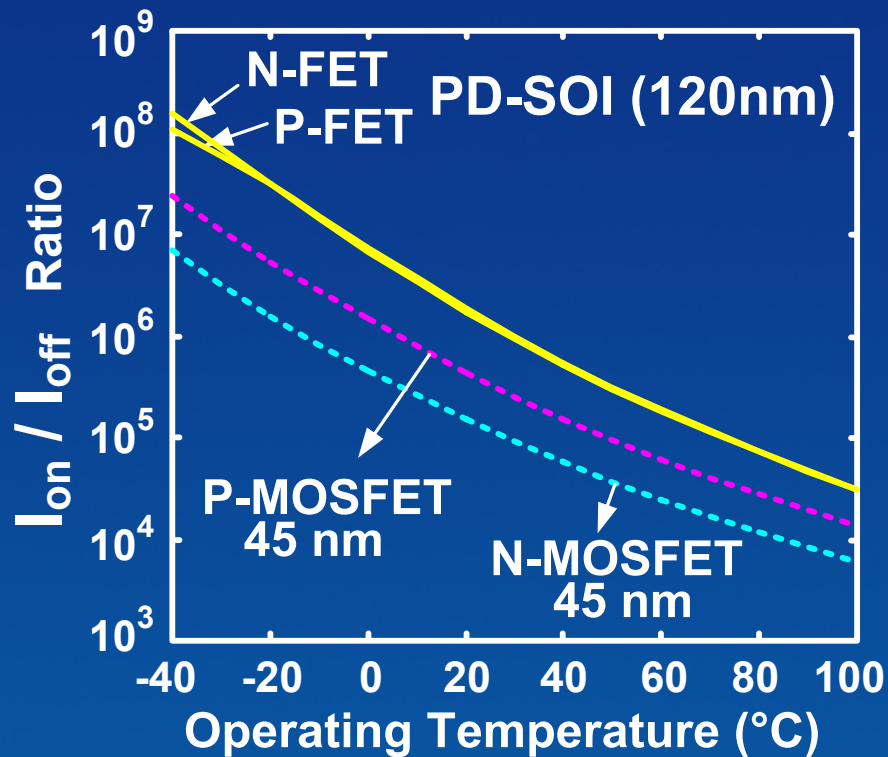
Cooling---Benefit at the Device Level

Transistor Drive Current Increases



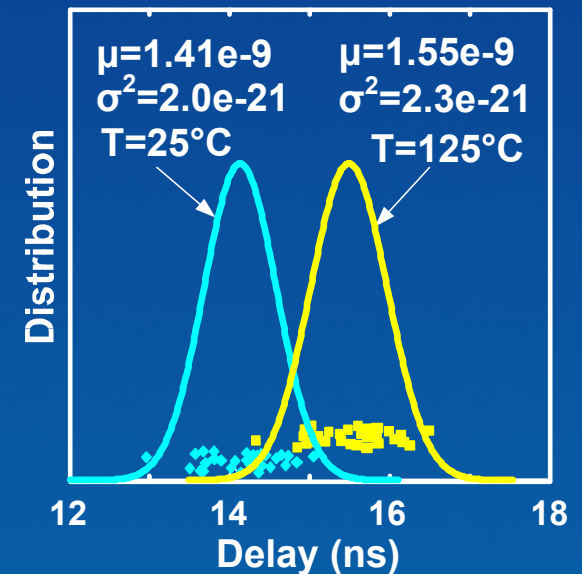
Higher drive current can be achieved by cooled operation across all technology nodes due to higher carrier mobility

Cooling---Benefit at the Circuit Level



S. Borkar, Intel

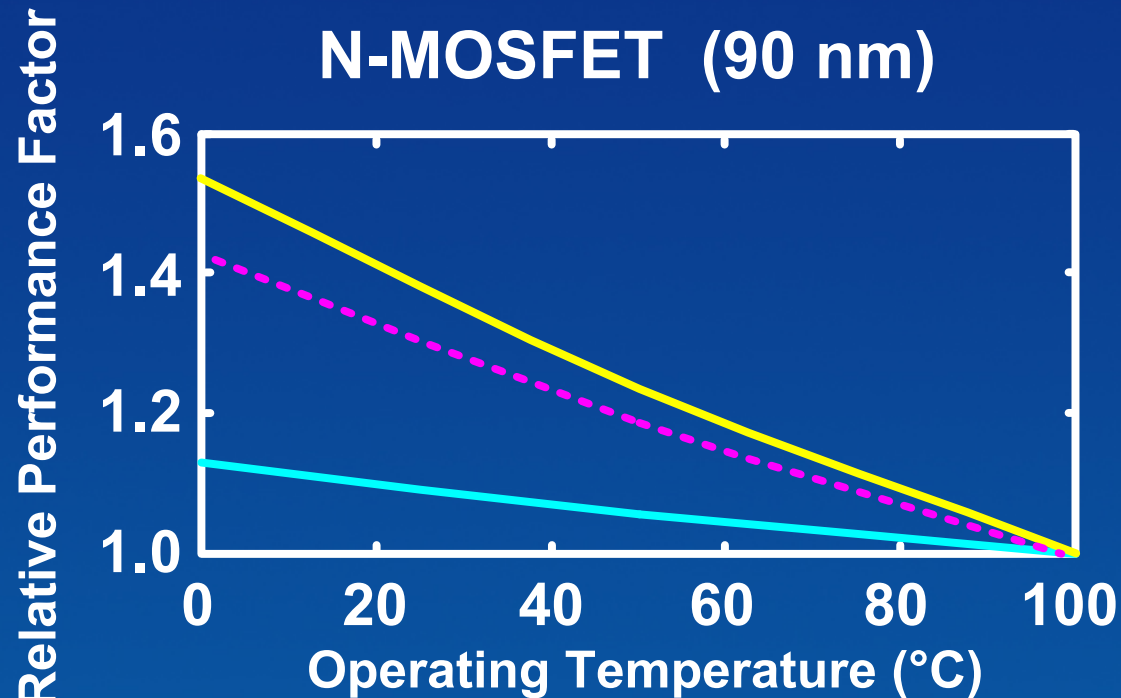
9-stage Inverter Chain



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



Further Exploiting the Benefit of Cooling



I. Aller et al., ISSCC 2000

— Same Device



... Same V_{th}

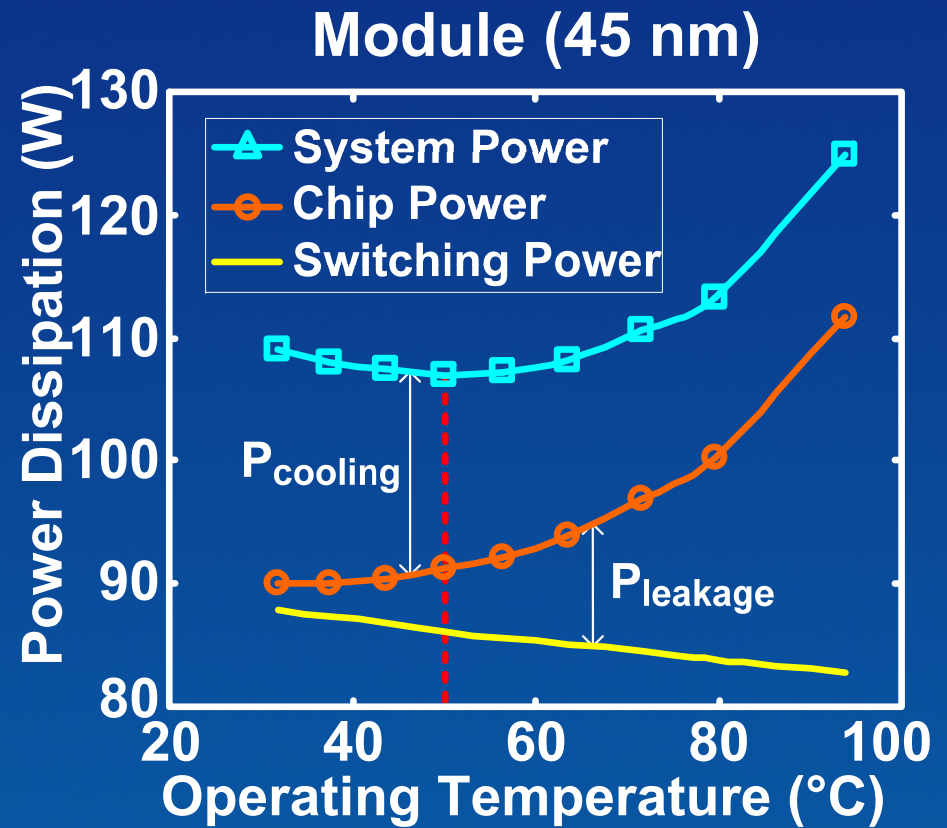
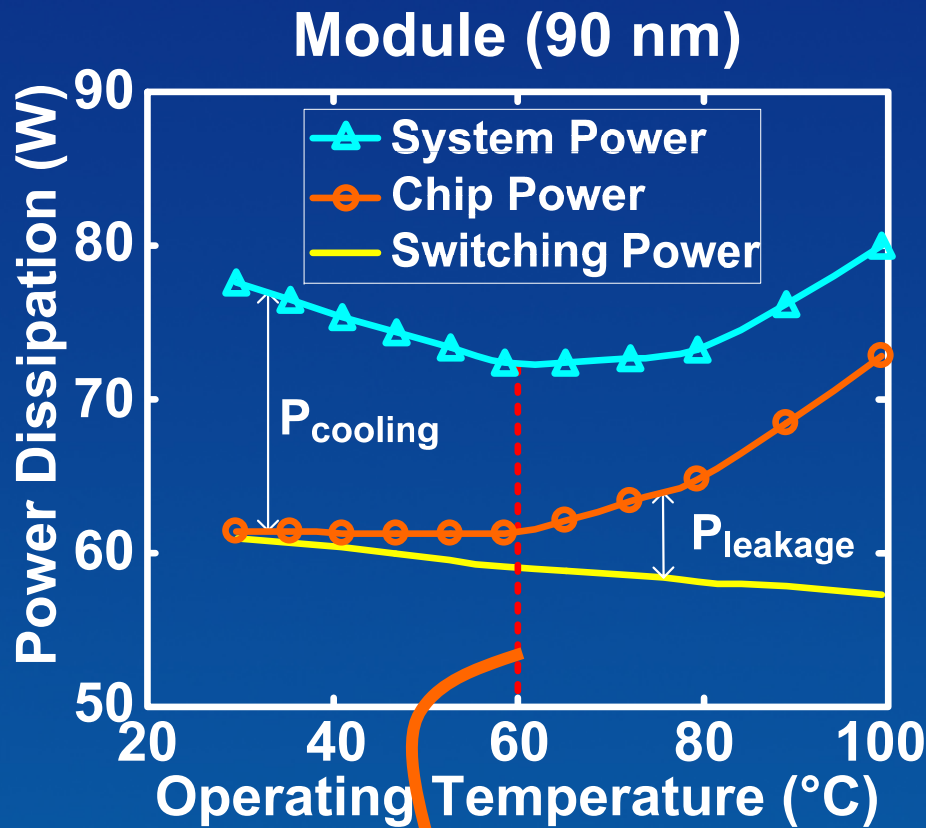
Applying Forward Body Bias (FBB)



— Same Leakage
Increase FBB Further

Performance can be further improved by different design strategies

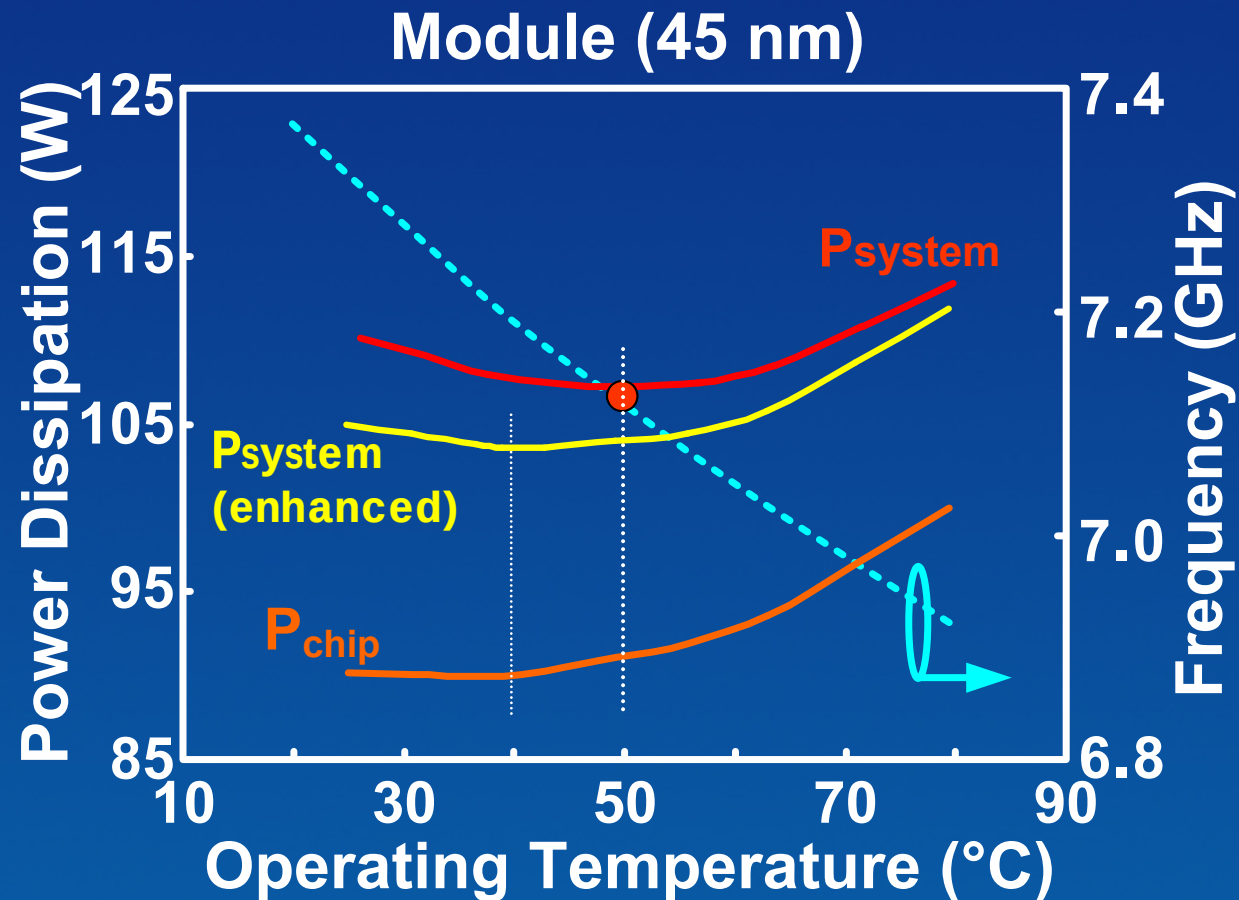
Cooling---System Level Considerations



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

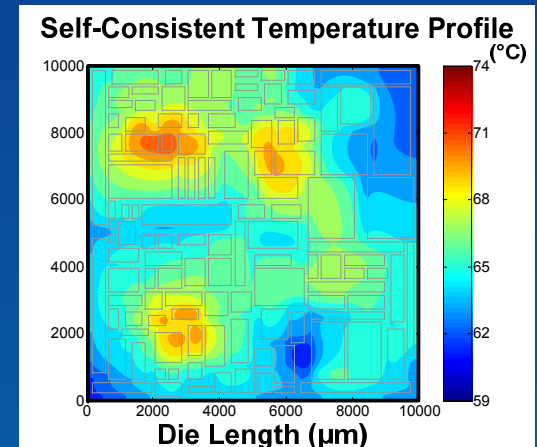
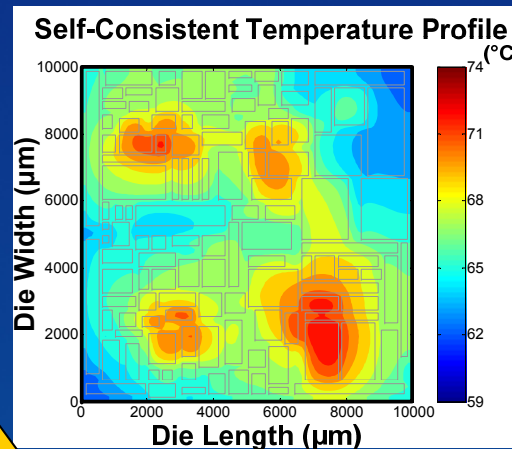
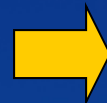
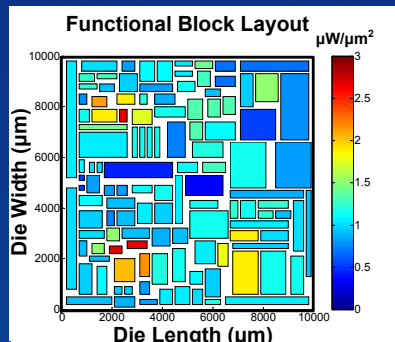
The limit occurs at a lower temperature as technology scales.

Extending the Practical Limit of Cooling



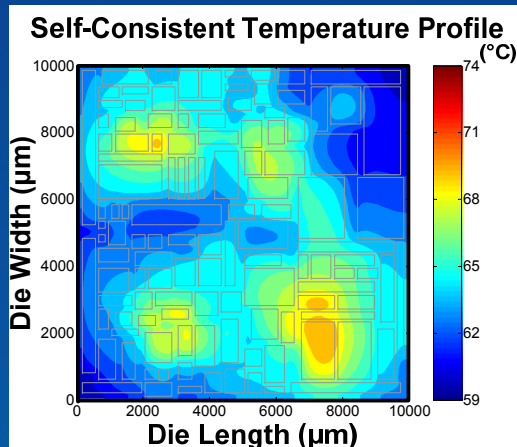
Practical limit can be further extended towards a lower temperature and higher performance is achieved by enhancing cooling efficiency

Hot-Spot Management



Localized Cooling

T_{MAX} decreases
but hot-spots remain



Global Cooling

Thermal resistance reduces 20%

Localized cooling will be more effective for hot-spot management

Conclusion

- ❖ 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