ECE 225
High-Speed Digital IC Design
Lecture 11

CMOS Devices: Power Dissipation, Scaling Scheme, and Power-Performance Trade-offs,

Future Device perspectives…

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

Power Dissipation

Where Does Power Go in CMOS?

- **Dynamic Power**

  Due to charging/discharging of capacitors….

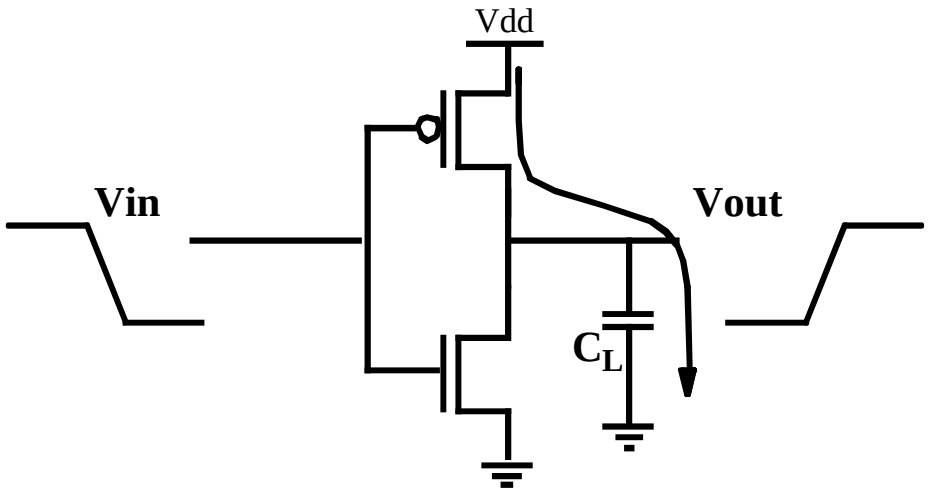
- **Leakage Power**

  - *Subthreshold*
  - *Gate*
  - *Junction*

- **Short-Circuit Power**

  When NMOS and PMOS are both turned ON….

Dynamic Power Dissipation



Energy/transition = C_L * V_{dd}^2

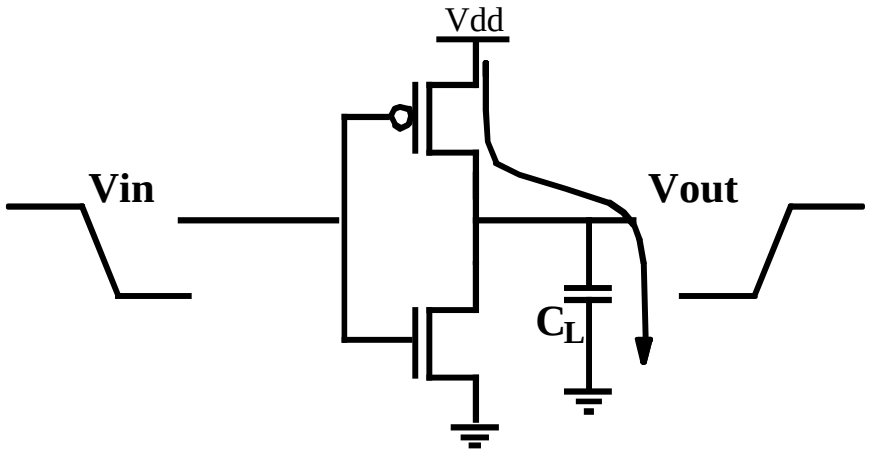
Power = Energy/transition * f = C_L * V_{dd}^2 * f

- **Not a function of transistor sizes!**
- **Need to reduce C_L, V_{dd}, and f to reduce power.**

Energy taken from supply during transition:

$$E_{V_{DD}} = \int_0^\infty i_{V_{DD}}(t) V_{DD}\, dt = V_{DD} \int_0^\infty C_L \frac{dv_{out}}{dt}\, dt = C_L V_{DD} \int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2$$

Dynamic Power Dissipation



Note: during the discharge phase, charge is removed from C_L and its energy is dissipated in the NMOS

Energy taken from supply during transition:

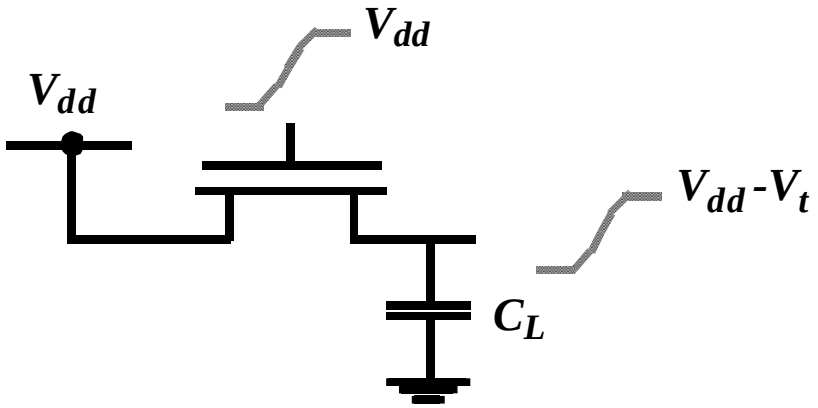
$$E_{V_{DD}} = \int_0^\infty i_{V_{DD}}(t) V_{DD}\, dt = V_{DD} \int_0^\infty C_L \frac{dv_{out}}{dt}\, dt = C_L V_{DD} \int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2$$

Energy stored in the capacitor:

$$E_C = \int_0^\infty i_{V_{DD}}(t) v_{out}\, dt = \int_0^\infty C_L \frac{dv_{out}}{dt} v_{out}\, dt = C_L \int_0^{V_{DD}} v_{out}\, dv_{out} = \frac{C_L V_{DD}^2}{2}$$

Where is the other half of the energy?….Dissipated by the PMOS

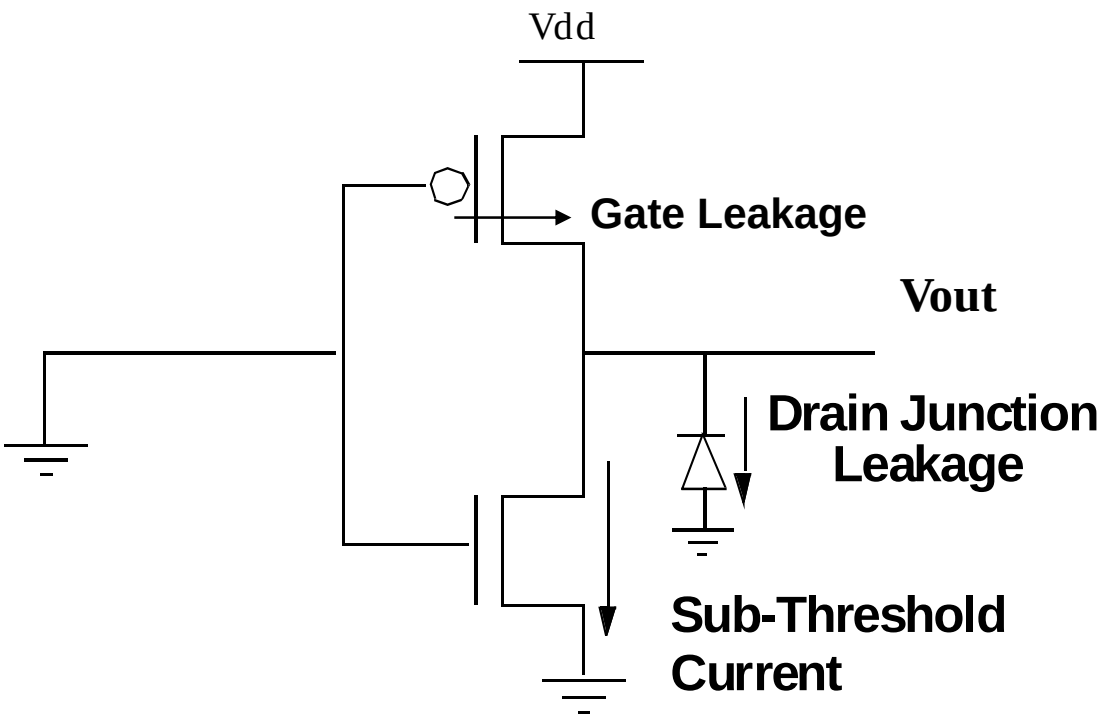
Modification for Circuits with Reduced Swing



$$E_{0 \rightarrow 1} = C_L \bullet V_{dd} \bullet (V_{dd} - V_t)$$

- **Can exploit reduced swing to lower power (e.g., reduced bit-line swing in memory)**

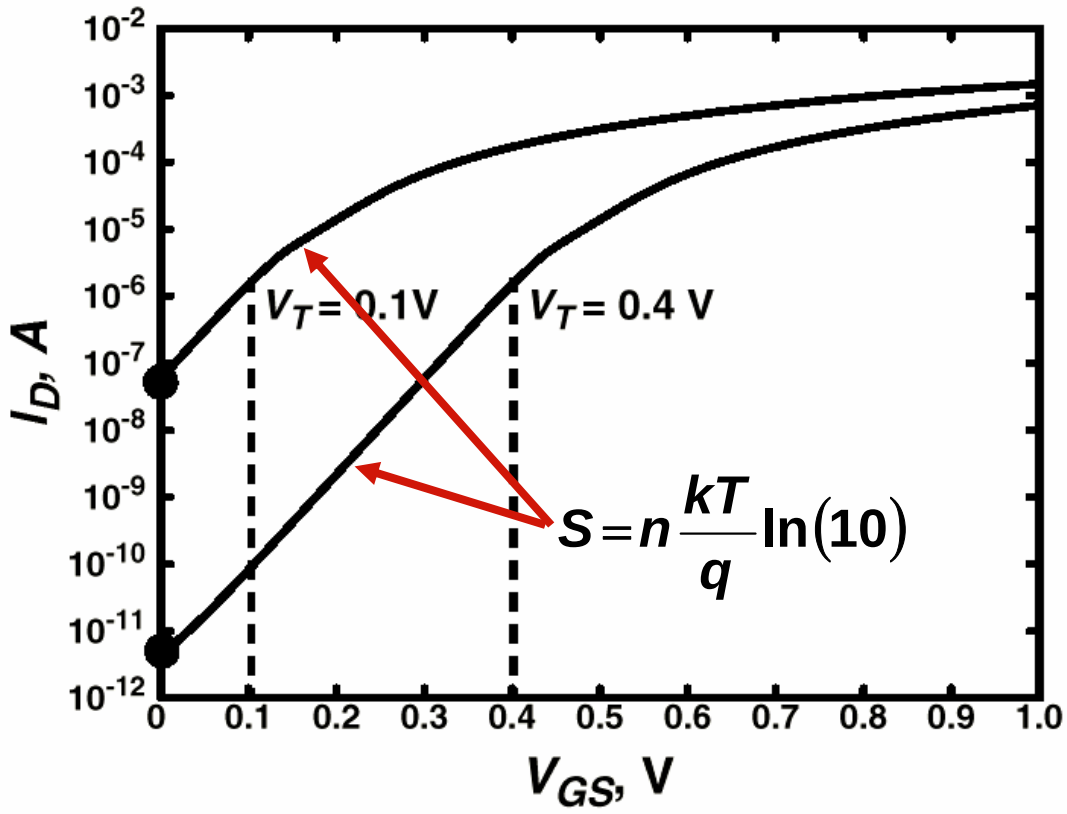
Primary Leakage Mechanisms

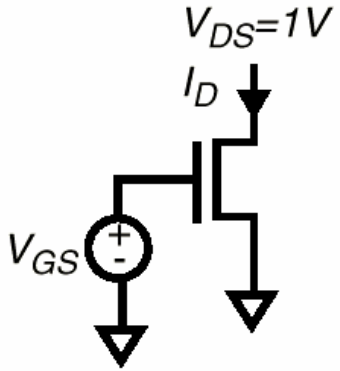


Sub-threshold current one of most compelling issues in low-energy circuit design!

Subthreshold Leakage Component

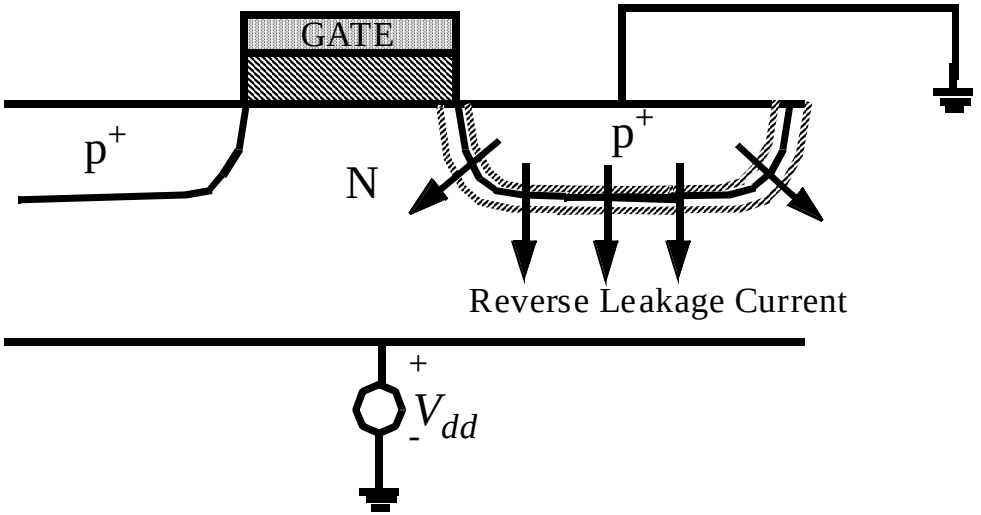
Subthreshold Slope = S = 60 mv/decade (for ideal transistor with n=1)





- **Leakage control is critical for low-voltage operation**

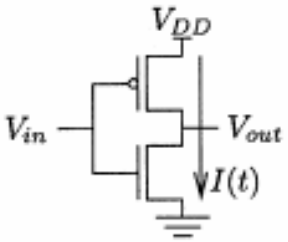
Reverse-Biased Diode Leakage



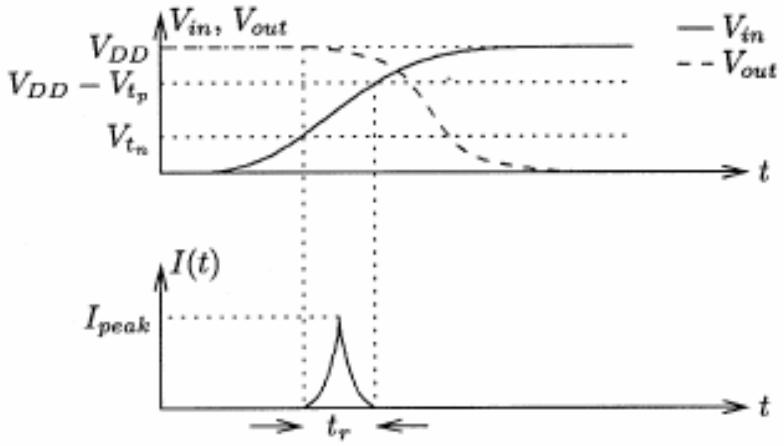
$$I_{DL} = J_S \times A$$

***JS* = 10-100 pA/μm2 at 25 deg C for 0.25μm CMOS**
***JS* doubles for every 9 deg C!**

Short Circuit Currents



(a)
CMOS
inverter.



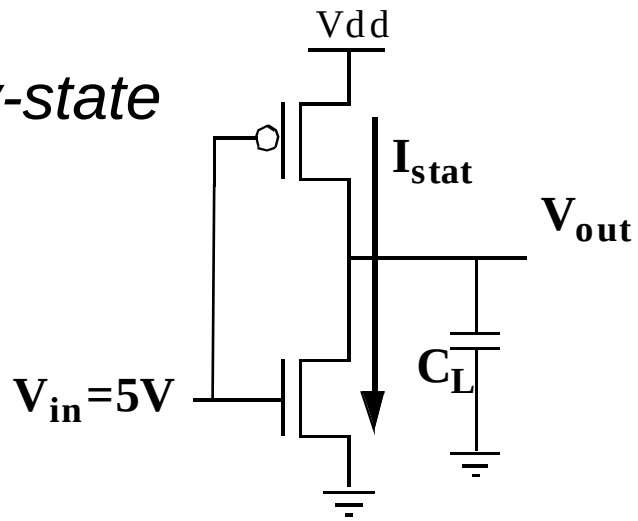
(b) Voltage and current waveforms.

Fig. 5. Voltage and current waveforms of a CMOS inverter.

K. Banerjee and A. Mehrotra, IEEE Transactions on Electron Devices, Vol. 49, No. 11, 2002.

Static Power Consumption

In the absence of switching….steady-state operation



Sources:

Due to all previously mentioned leakage currents……

$$P_{stat} = P_{(In=1)}.V_{dd} \cdot I_{stat}$$

Wasted energy …
Should be avoided in almost all cases…..
but could help reducing energy in others (e.g. sense amps)

Principles for Power Reduction

- Prime choice: Reduce voltage!
  - Recent years have seen an acceleration in supply voltage reduction
  - Design at very low voltages still open question (0.6 … 0.9 V by 2010!)
- Reduce switching activity
- Reduce physical capacitance
  - Device Sizing

A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, pp. 498–523, Apr. 1995.

Leakage Power Reduction

- Process scaling
  - V_T reduces with each new process (historically)
  - Leakage increases ~10X!
- Leakage vs. performance tradeoff:
  - For high-speed, need small V_T and L
  - For low leakage, need high V_T and large L
- One solution: dual-V_T process
  - Low-V_T transistors: use in critical paths for high speed
  - High-V_T transistors: use to reduce power

Scaling Theory:

Parameters		Constant Vdd scaling	Constant E scaling
Dimentions	Width = w	0.7	
	Length = L	0.7	
	Oxide thickness t_{ox}	0.7	
	Junction depth X_j	0.7	
Die Area		$(0.7)^2$	
Gate capacitance per unit area ($C_{gox} = \varepsilon_{ox}/t_{ox}$)		$\frac{1}{0.7}$	
Gate capacitance ($C_g = wLC_{gox}$)		0.7	
Total Capacitance (C)		0.7	
Supply Voltage (V_{DD})		1	0.7
Current per device ($I_{DS} \propto \frac{w}{L}\frac{\varepsilon_{ox}}{t_{ox}}(V_{gs} - V_{th})V_{DD}$)		1	0.7
Intrinsic Gate Delay ($\tau = \frac{C_g \Delta V}{I_{AV}}$)		0.7	0.7
Frequency ($f \propto \frac{1}{\tau}$)		$\frac{1}{0.7}$	
Active Power Dissipation ($P_{active} = CV_{DD}^2 f$)		1	$(0.7)^2$
Energy-Delay Product ($CV_{DD}^2 \tau$)		$(0.7)^2 = 0.49$	$(0.7)^4 = 0.2401$
Power Dissipation density ($P_{active}/Area$)		$\frac{1}{(0.7)^2} \approx 2$	$\frac{(0.7)^2}{(0.7)^2} \approx 1$

Design Optimization…

High Performance Design

Low Power Design

- Lower supply voltage
- Higher threshold voltage

- Higher supply voltage
- Lower threshold voltage

Need a design metric for optimization

Energy-Delay Product (EDP)….

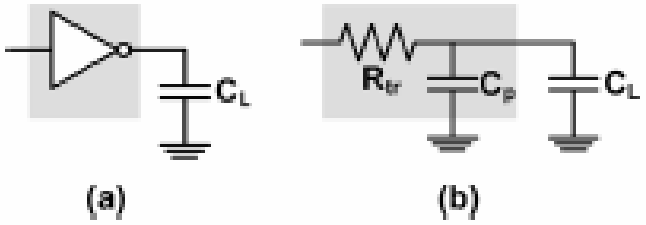


Fig. 1. (a) Schematic diagram of an inverter gate with a capacitive load (C_L). (b) Equivalent *RC* model of an inverter gate with a capacitive load (C_L). The inverter gate [shaded region in (a)] is represented by an effective on-resistance (R_{tr}) and an output parasitic capacitance (C_p).

$$T_g = \frac{C\Delta V}{I_{avg}}$$

$$T_g = \frac{KV_{dd}}{(V_{dd} - V_{th})^{\alpha}}.$$

$$F = \frac{1}{T_g L_d}.$$

From Alpha-Power model…

K lumps the capacitance and other process parameters….

Energy-Delay Product (EDP)….

$$P_{\text{dynamic}} = aC_{\text{eff}}V_{\text{dd}}^2F$$
$$P_{\text{static}} = I_s e^{-V_{\text{th}}/\gamma V_o}(1 - e^{-V_{\text{th}}/\gamma V_o})W_{\text{eff}}V_{\text{dd}}$$

$$\text{Energy} = (P_{\text{dynamic}} + P_{\text{static}}) \bullet \text{Delay}.$$

If P = power and T= delay

EDP = Energy. Delay = PT^2

R. Gonzalez, B. M. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, Aug. 1997.

EDP Optimization…traditional

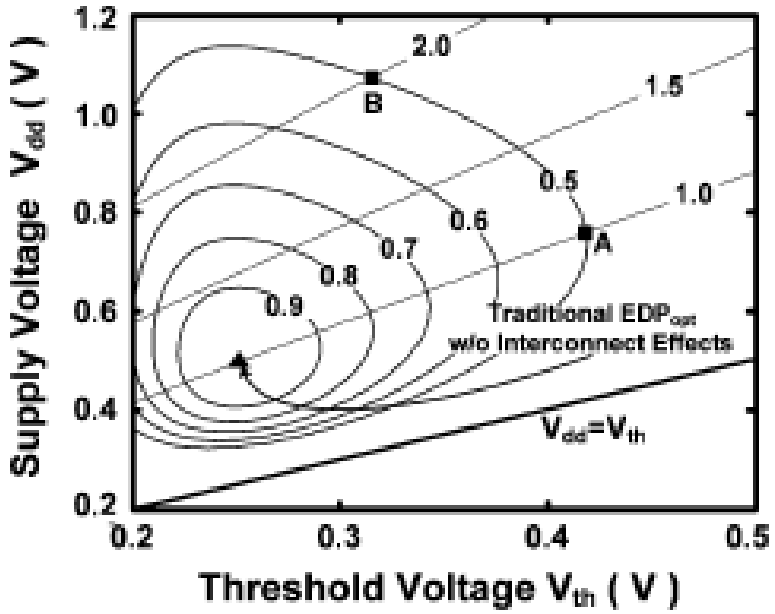


Fig. 2. Traditional V_{dd}/V_{th} optimization uses EDP as a design metric (without considering the effect of interconnects). The EDP contours and iso-performance curves are obtained by simple numerical calculation without considering electrothermal couplings between temperature and static power dissipation at 100-nm technology node. Note that although five EDP contours (0.9, 0.8, 0.7, 0.6, 0.5) and three iso-performance curves (1.0, 1.5, 2.0) are drawn in this figure, other values of EDP contour and iso-performance curve can be easily employed.

S-C. Lin and K. Banerjee, "**A Design-Specific and Thermally-Aware Methodology for Trading-off Power and Performance in Leakage-Dominant CMOS Technologies,"** *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 16, No. 11, pp. 1488-1498, Nov. 2008.

Other Metrics?

$$\text{EDP} = \text{Energy} \cdot \text{Delay} = \text{PT}_g^2$$
$$\text{PDP} = \text{Power} \cdot \text{Delay} = \text{Energy} = \text{PT}_g$$
$$\text{PEP} = \text{Power} \cdot \text{Energy} = \text{P}^2\text{T}_g.$$

…Which metric is appropriate?

Design-Specific and Thermally-Aware Power-Performance Trade-offs….

S-C. Lin and K. Banerjee, "**A Design-Specific and Thermally-Aware Methodology for Trading-off Power and Performance in Leakage-Dominant CMOS Technologies,"** *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 16, No. 11, pp. 1488-1498, Nov. 2008.

- Optimization metric should be determined by design-specific requirements
- A systematic methodology has been proposed
  - allows circuit designers to comprehend reliability and packaging constraints
  - incorporates electrothermal couplings between temperature, power dissipation, and performance
  - chooses design-specific metrics for power and performance optimization
- A more meaningful comparison basis is proposed
- The design-specific metric is sensitive to technology scaling

Incorporating Interconnect Effects…

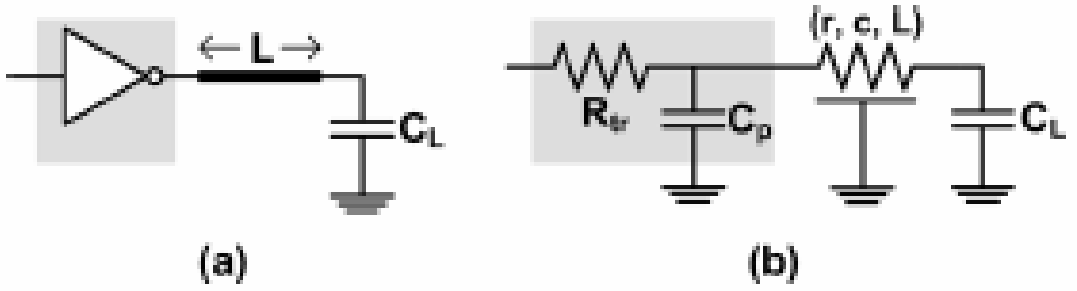


Fig. 3. (a) Schematic diagram of an inverter gate driving a capacitive load (C_L) through a wire of length L. (b) Equivalent *RC* model for (a). The inverter gate [shaded region in (a)] is represented by an effective on-resistance (R_{tr}) and an output parasitic capacitance (C_p). Note that C_L is the input capacitance of the gate in the next stage.

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

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

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

EDP Optimization Incorporating Interconnect Effect

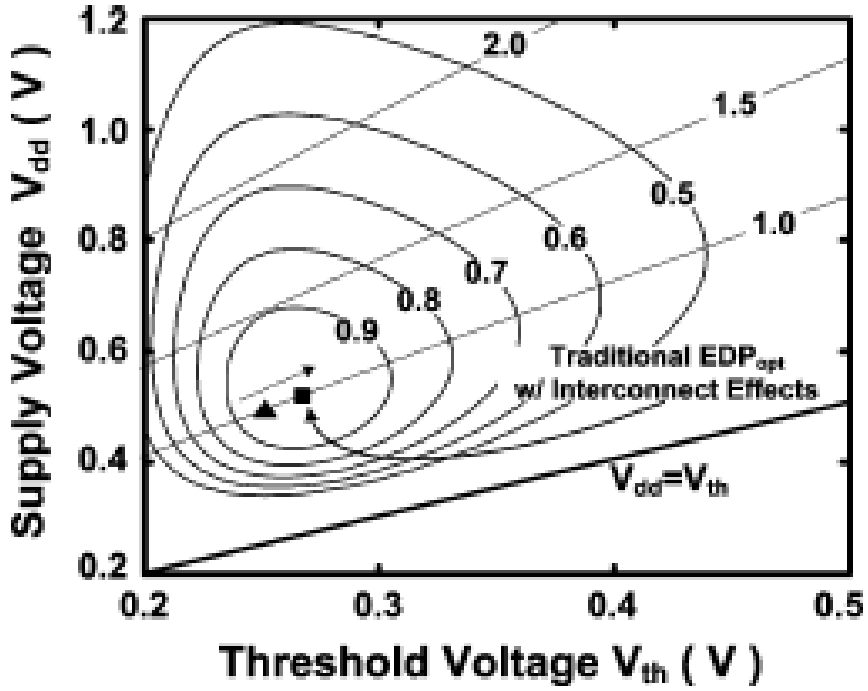
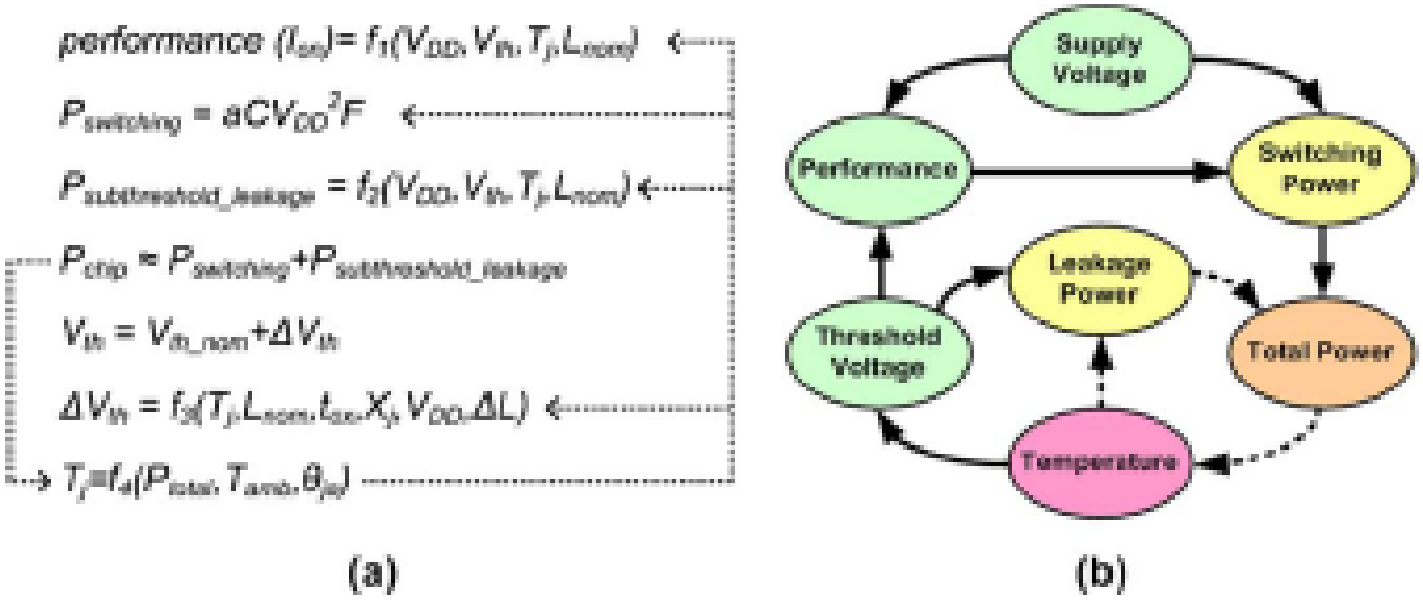


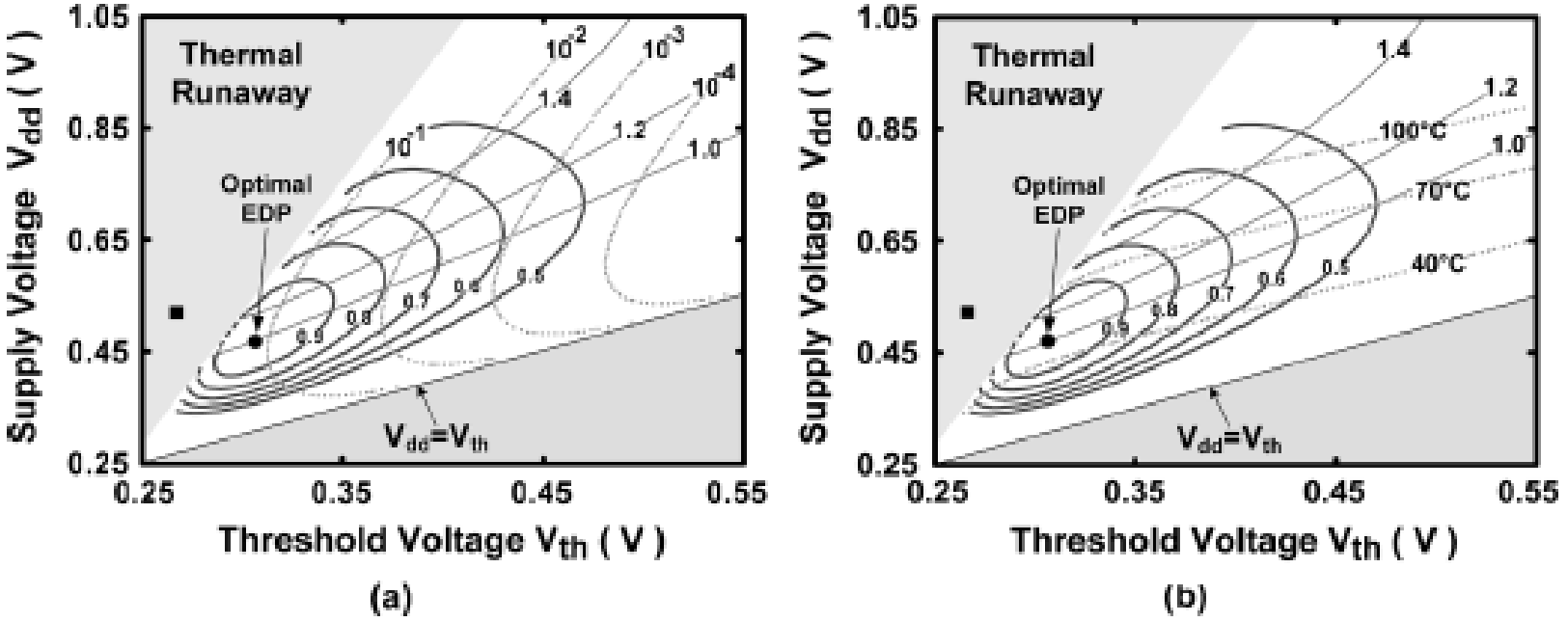
Fig. 4. V_{dd}/V_{th} optimization using EDP as a design metric (with the effect of interconnects included). The EDP contours and iso-performance curves are obtained by simple numerical calculation without considering electrothermal couplings between temperature and static power dissipation at 100-nm technology node. EDP_{opt} without considering the effect of interconnects is also shown ("▲") for comparison.

Electrothermal Couplings….



*S-C. 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,**" IEEE Transactions on Electron Devices, Vol. 54, No. 12, pp. 3342-3350, 2007*

Thermal Aware EDP Optimization…



Comparisons between different Metrics….

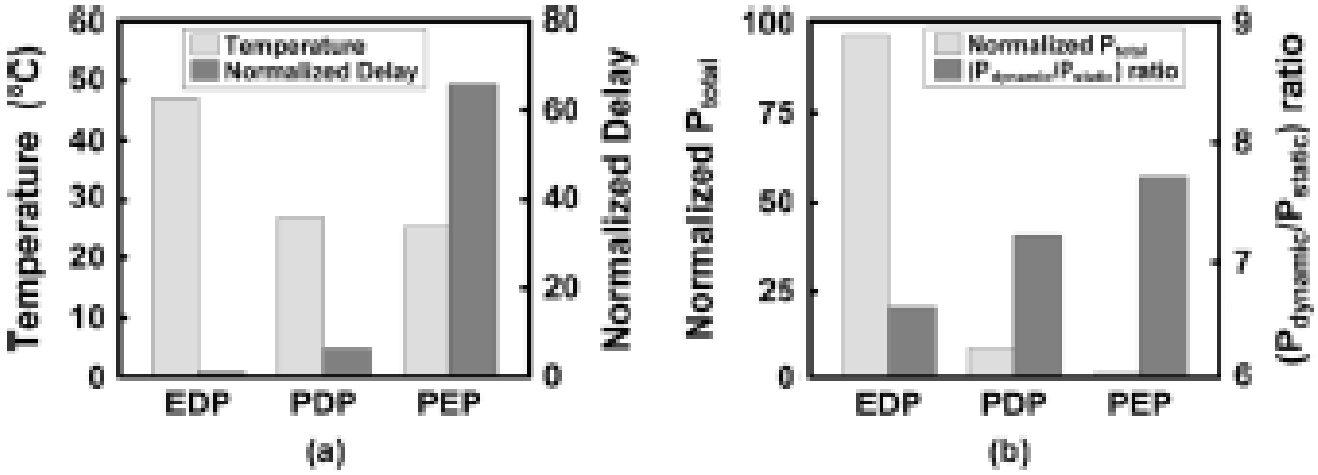


Fig. 11. Comparison of delay, temperature, and power between different optimization metrics. (a) Normalized delay (w.r.t. the EDP case) and average junction temperature corresponding to the optimal operating points obtained using three different optimization metrics (EDP, PDP, and PEP). (b) Normalized total power dissipation (w.r.t. the PEP case) and the $(P_{dynamic}/P_{static})$ ratio corresponding to optimal operating points of different optimization metrics.

Optimal Operation Locus…

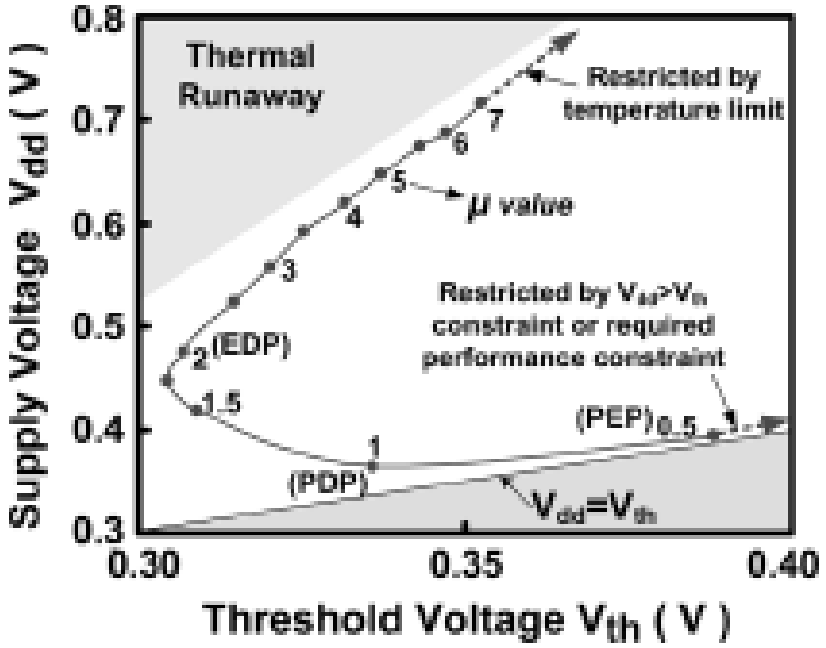


Fig. 12. Optimal operation locus for 100-nm technology node. The optimal locus is formed by the optimal points (denoted by "•") with respect to different values of μ, i.e., the minimal values of the product $P(T_d)^{\mu}$. Although values of μ between 0.5 and 7 are shown in the plot, μ can be chosen to be any positive rational number depending on design space and constraints. The shaded regions in the two corners correspond to thermal runaway region and the region where the supply voltage is less than the threshold voltage.

Choosing the right metric….

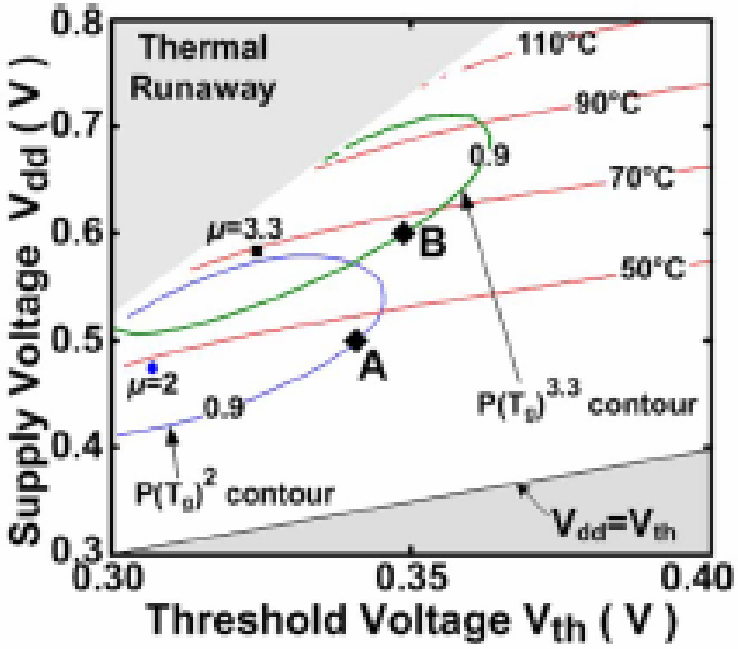


Fig. 14. Example comparing the use of the proposed metric $\mathrm{P}(\mathrm{T}_g)^{3.3}$ in choosing between two design options (A and B) against that of conventional EDP evaluation. "•" indicates EDP ($\mu = 2$) optimal operating point and "■" indicates the optimal point that meets all design-specific requirements. Since design optimization is preferably to find an operation set closest to the optimal point, it is clear that the choice between the two points A and B changes depending on the metric of optimization chosen.

Understanding choice of packaging/cooling solutions at the circuit design level…

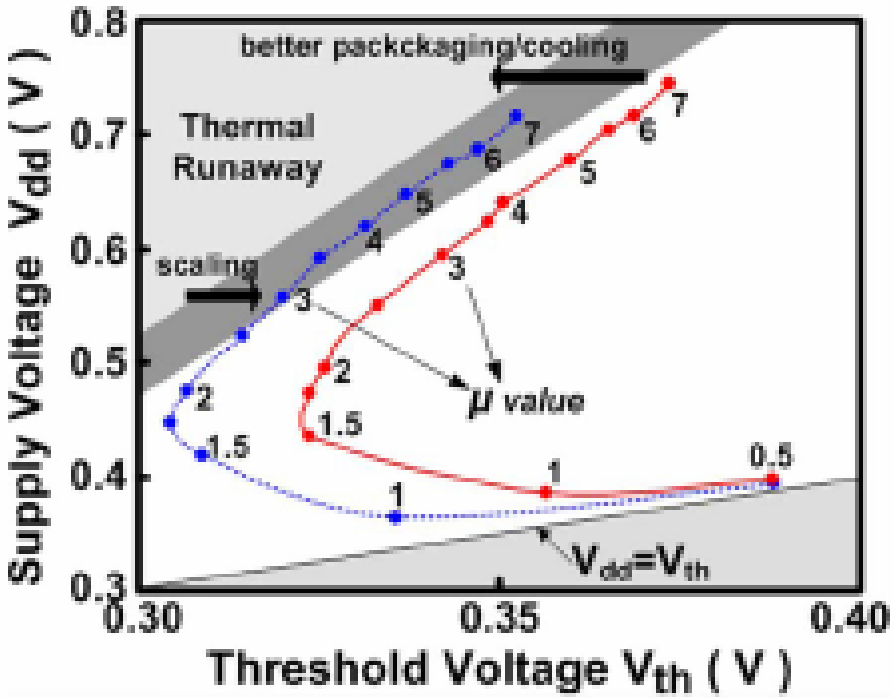


Fig. 15. Scaling analysis of optimal operating points obtained by applying different optimization metrics (shown for 100- and 70-nm technology nodes). Note that the thermal runaway region expands to the right due to technology scaling.

Understanding impact of scaling and variations…

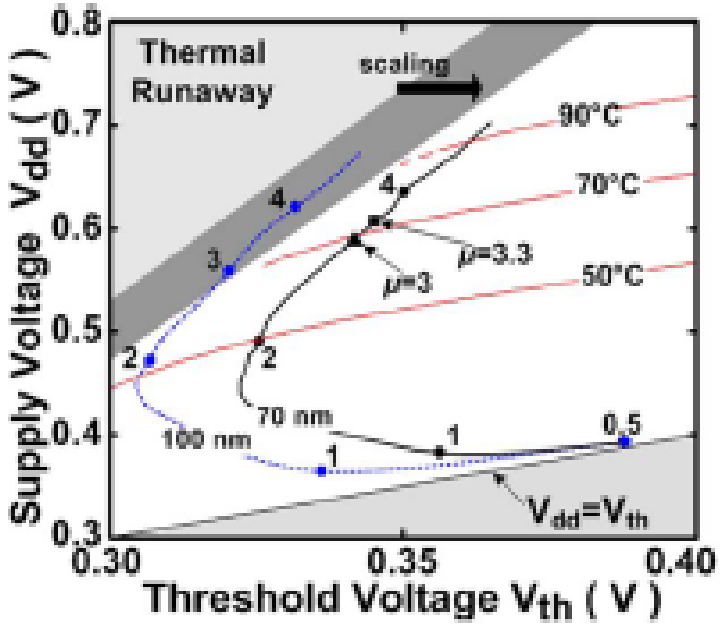


Fig. 16. Impact of technology scaling on the proposed optimization methodology. The optimal operation locus of 100-nm technology node is shown by a dotted curve (the same locus shown in Fig. 13) while the locus of 70-nm technology node with the same design is shown by a solid curve. Three iso-temperature curves (50 °C, 70 °C, 90 °C) are superimposed. "•" indicates the optimal operating points with different optimization metrics. "■" indicates the optimal operating point when $P(T_g)^{3.3}$ is chosen for 70-nm technology node. "▲" indicates the optimal operating point when $P(T_g)^3$ is chosen. Note that the metric ($\mu = 3$) meets all design-specific requirements at 70-nm technology node.

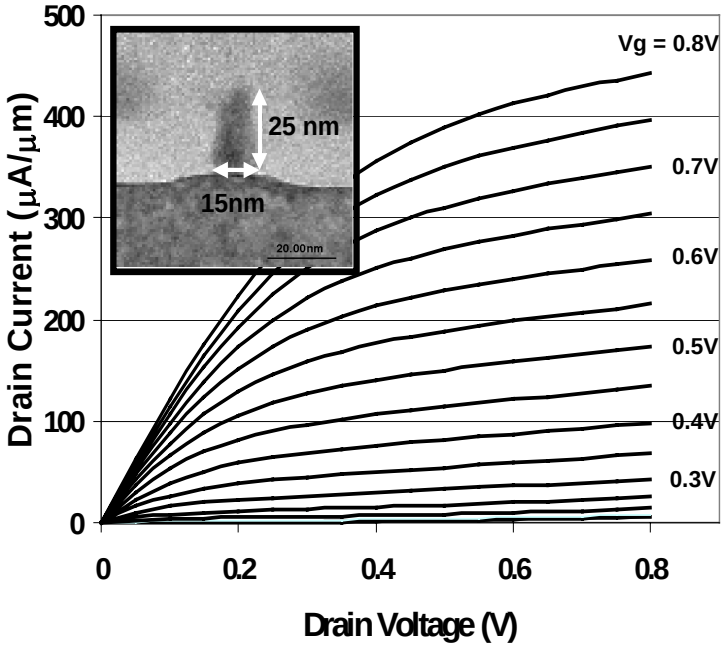
Pushing CMOS: the 10 nm wall

- **Smart engineering, new materials and evolutionary architectures have continuously pushed the CMOS limits…**

  I_{on} > 700 μA/μm
  I_{off} < 1000 pA/μm
  @ room temperature
  low operating power logic

- **Key problems of sub-10 nm MOSFET**

  (i) electrostatic limits
  (ii) source-to-drain tunnelling
  (iii) carrier mobility
  (iv) process variations
  (v) static leakage
  (vi) power density

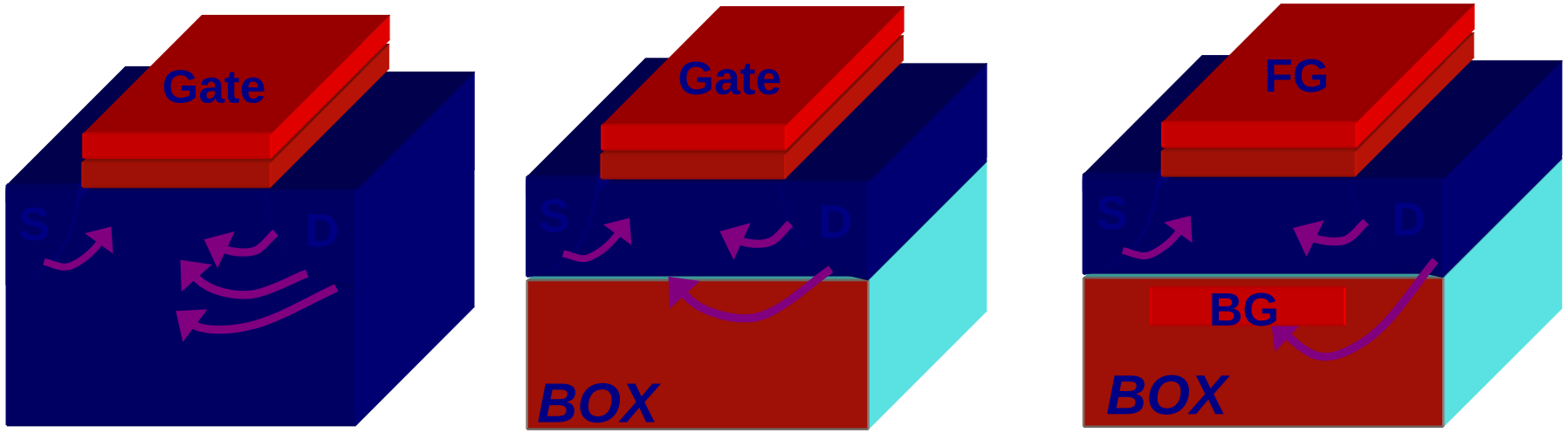


Intel's 15 nm channel length MOSFET

Nanoscale CMOS Problems and Solutions

- High E-field: mobility degradation
  - Use Ge or SiGe (introduce strain) and increase mobility
- Gate: tunneling and depletion
  - Use high-k dielectric, metal gate (work function engineering)
- Source/Drain: parasitic contact resistance
  - Use Schottky S/D
- Channel: reduced Vg-Vt causes reduction in Ion (electrostatics)
  - Use DG, FinFET or Tri-gate to retain gate control over channel, minimize Ioff

Device Topologies….

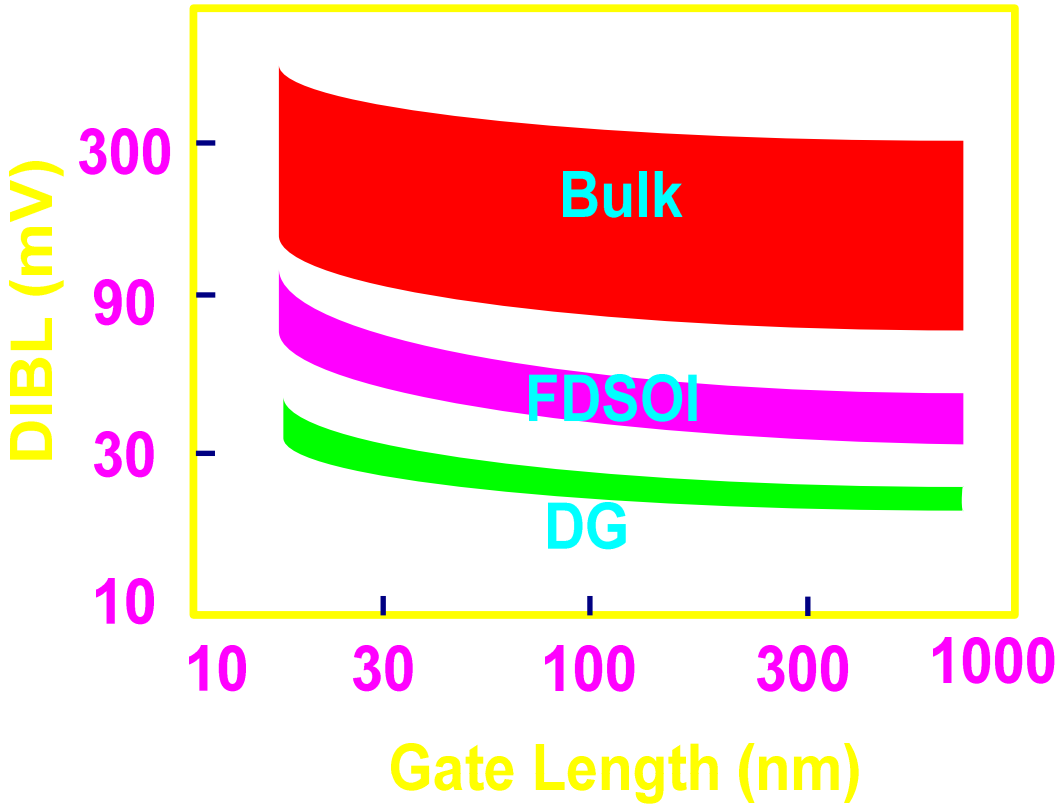


Bulk ***FDSOI*** ***Double Gate***

- Proximity of the BG to the channel
- Electric field lines from Drain are effectively shielded by back-gate in DG devices

→ **improved SCE**

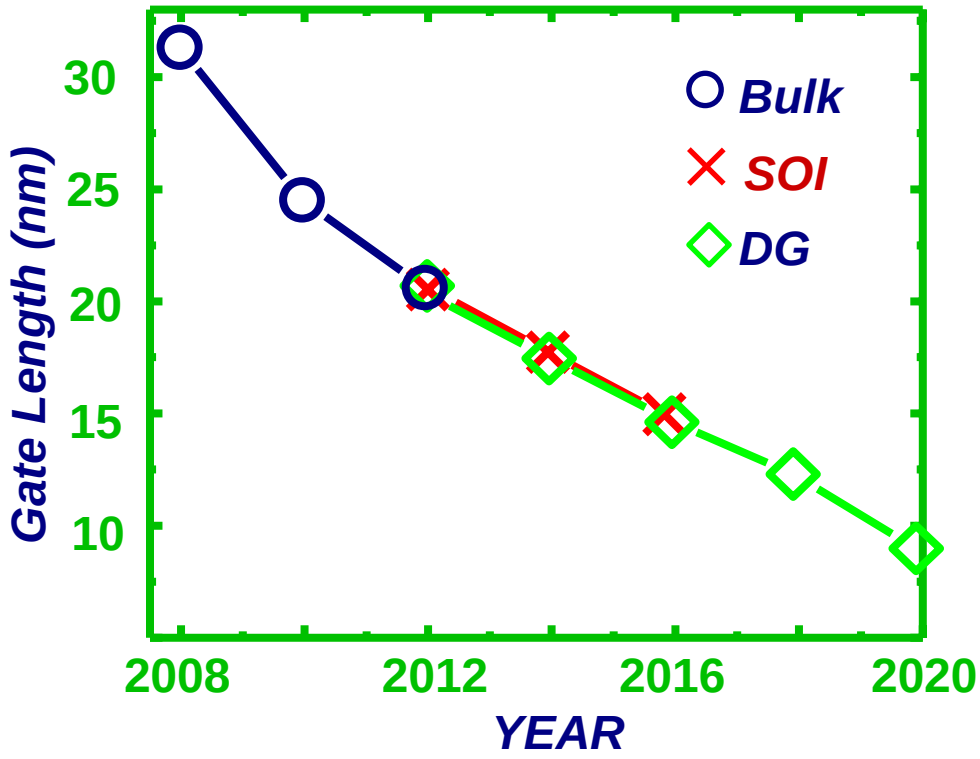
DIBL Control.....



- Better control of the channel by two gates reduces Drain-Induced Barrier-Lowering (DIBL) in DG devices

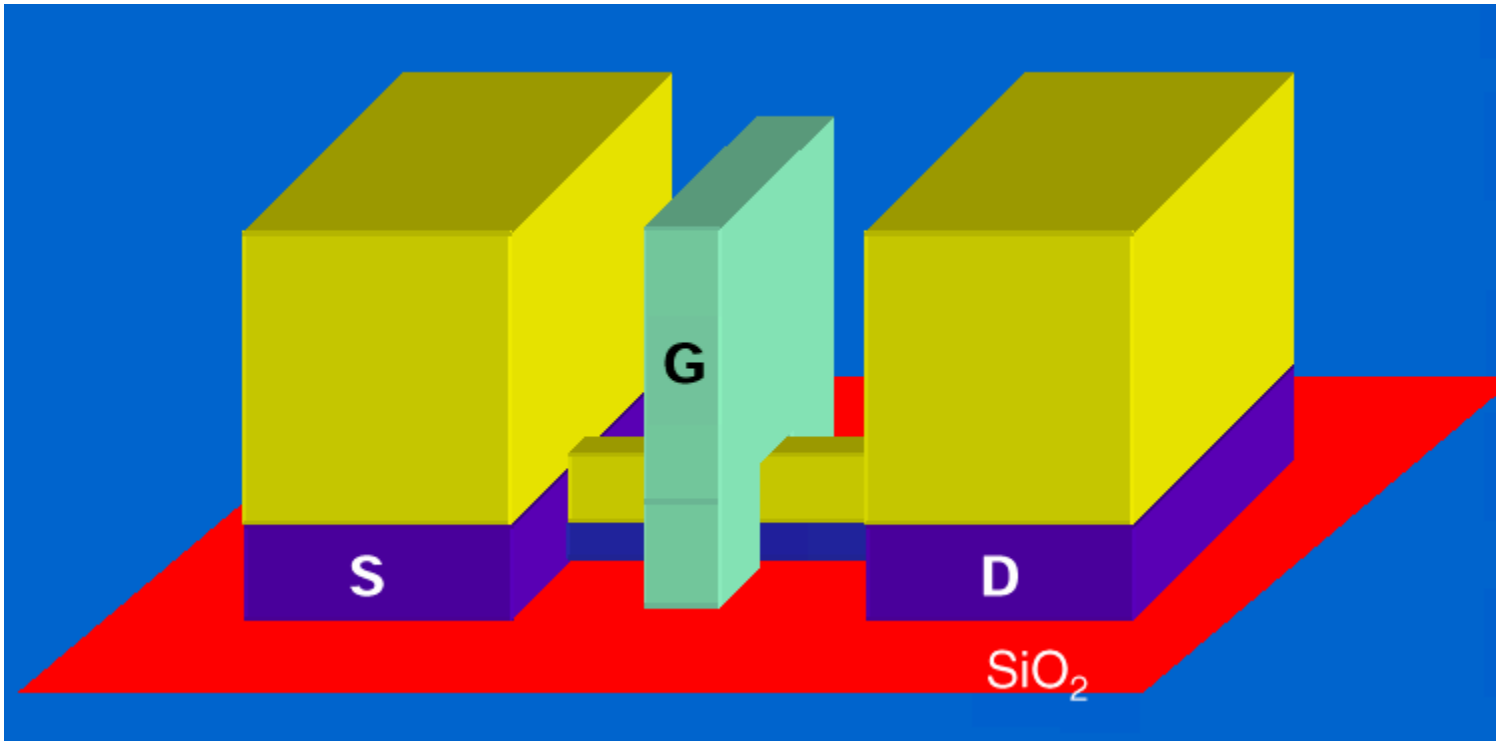
→ **improved leakage**

Device Technology Scaling Trend....



- Double-Gate devices needed for ultimate scaling!!

Future Perspectives……



25 nm FINFET MOS transistor (Berkeley)