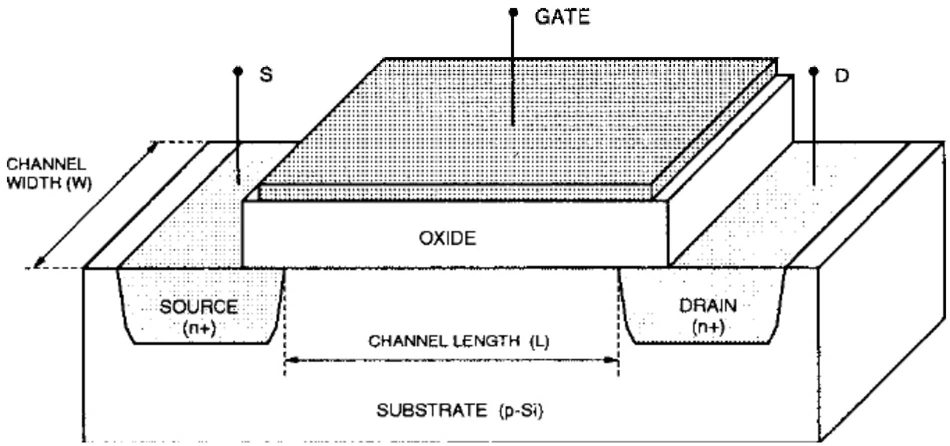
Minimizing Leakage Power in CMOS: Technology Issues

Centre SI Summer School on
Nanoelectronic Circuits and Tools

Massoud Pedram
Dept. of Electrical Engineering
University of Southern California

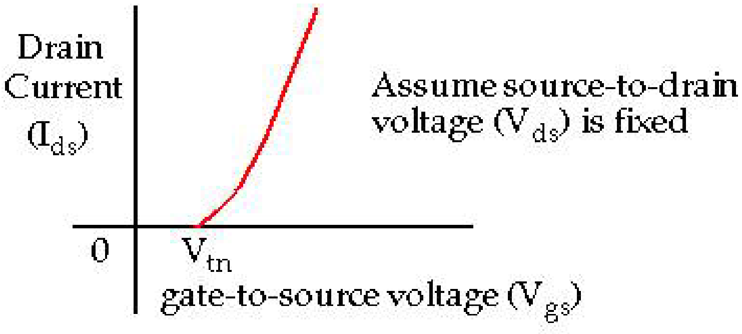
July 15, 2008

Physical Structure of a Long N-Channel Enhancement-Type MOSFET



- Threshold voltage, V_T, is defined as the voltage at which an MOS transistor begins to conduct. For voltages less than V_T , the channel is cut off.

n-channel enhancement MOS



Assume source-to-drain voltage (V_{ds}) is fixed

Threshold Voltage of an nMOS Transistor

For V_{SB}=0, the threshold voltage,V_{T0}, is defined as the gate potential V_G at which the surface potential ϕ_s changes by $2\phi_F$, i.e., the surface becomes strongly inverted.

Flat band voltage   Ideal threshold voltage

$$\boxed{V_{T0} = \left(\Phi_{GC} - \frac{qN_{OX}}{C_{OX}}\right) + \left(-2\phi_F - \frac{Q_{B0}}{C_{OX}}\right) + \frac{qN_I}{C_{OX}}}$$

Channel-implant induced shift

Φ_{GC} ⇨ The work function difference between the gate and the channel

qN_{OX} ⇨ Positive charge density at the gate Si-oxide interface due to impurities and lattice imperfections at the interface (Sign is always positive)

ϕ_F ⇨ The substrate Fermi potential

Q_{BO} ⇨ Depletion charge density at surface inversion

qN_I ⇨ Additional channel implant density (Sign is positive for p-type and negative for n-type implant)

nMOS transistor: $Q_{BO} = -\sqrt{2qN_A\varepsilon_{Si}\,|-2\phi_F\,|}$

pMOS transistor: $Q_{BO} = \sqrt{2qN_D\varepsilon_{Si}\,|-2\phi_F\,|}$

Threshold Voltage (Cont'd)

$$\Phi_{GC} = \phi_F - \phi_{F(gate)}$$

$$\phi_{F(gate)} = \begin{cases} 0.55V & \text{for heavily doped n-type polysilicon gate (edge of conduction band)} \\ -0.55V & \text{for heavily doped p-type polysilicon gate (edge of valence band)} \\ \phi_M & \text{for metal gate} \end{cases}$$

$$\varepsilon_{ox} = 0.34 \times 10^{-12}\, Fcm^{-1},\ \varepsilon_{si} = 1.06 \times 10^{-12}\, Fcm^{-1}$$

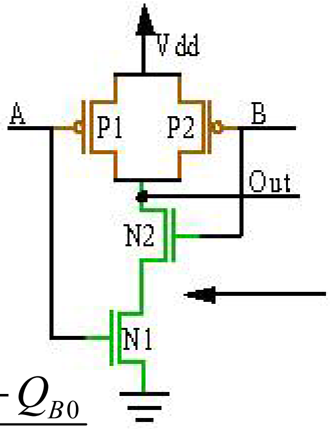
$$\boxed{C_{OX} = \frac{\varepsilon_{OX}}{t_{OX}}}$$

Threshold voltage determinants:

- Gate conductor materials
- Gate oxide material & thickness
- Substrate doping
- Channel Ion Implantation
    - p-type (n-type) impurities, V_T is made more positive (negative)
- Impurities in Si-oxide interface, Q_{ox}
- Source-bulk voltage, V_{SB}
- Temperature, T

Threshold Voltage (Cont'd)

For $V_{SB} \neq 0$, threshold voltage is denoted as V_T



Drain of N1 is source of N2

$$Q_B = -\sqrt{2qN_A\varepsilon_{Si}\,|-2\phi_F + V_{SB}|}$$

$$V_T = \Phi_{GC} - 2\phi_F - \frac{Q_B}{C_{OX}} - \frac{Q_{OX}}{C_{OX}}$$

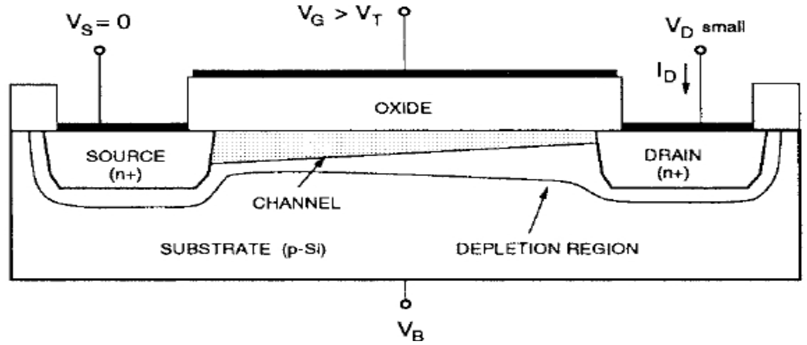
$$= \Phi_{GC} - 2\phi_F - \frac{Q_{B0}}{C_{OX}} - \frac{Q_{OX}}{C_{OX}} - \frac{Q_B - Q_{B0}}{C_{OX}} = V_{T0} - \frac{Q_B - Q_{B0}}{C_{OX}}$$

where $\frac{Q_B - Q_{B0}}{C_{OX}} = -\frac{\sqrt{2qN_A\varepsilon_{Si}}}{C_{OX}}\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right)$

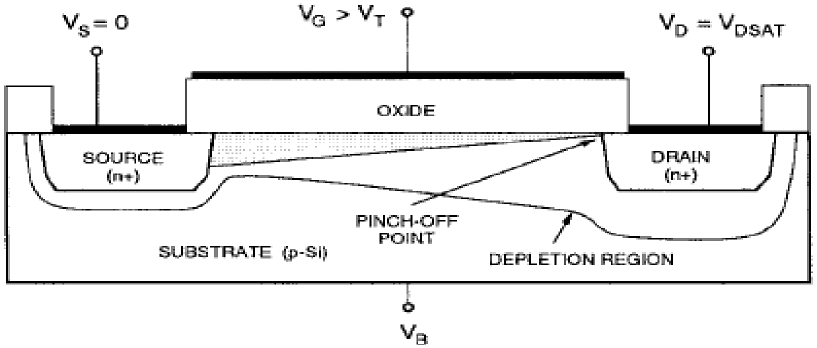
$$\boxed{V_T = V_{T0} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right)}$$

where $\gamma = \text{body effect coefficient} = \frac{\sqrt{2qN_A\varepsilon_{Si}}}{C_{OX}}$

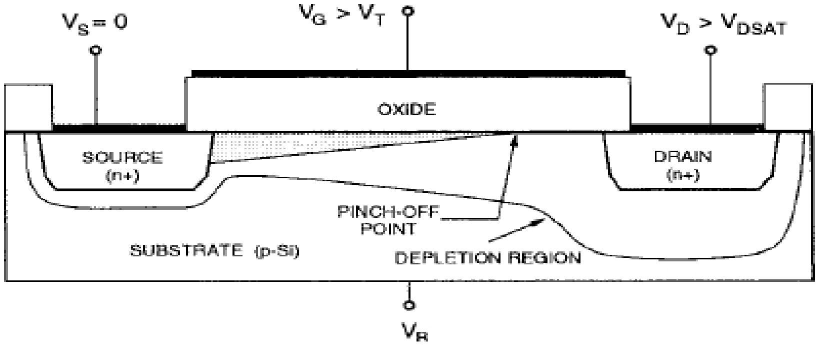
Cross-sectional View of an nMOS Transistor when $V_G > V_T$



Operating in the linear region

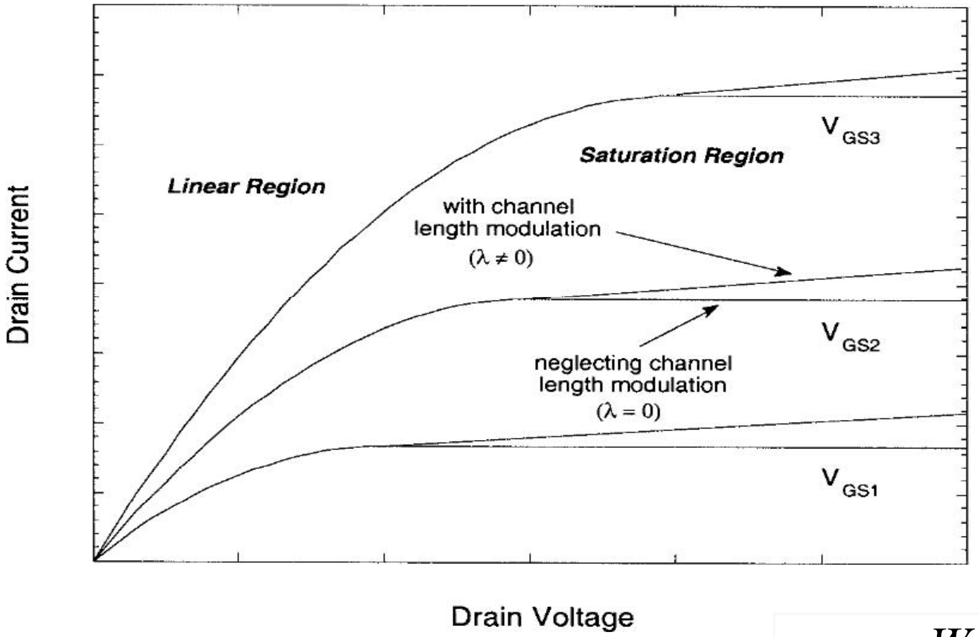


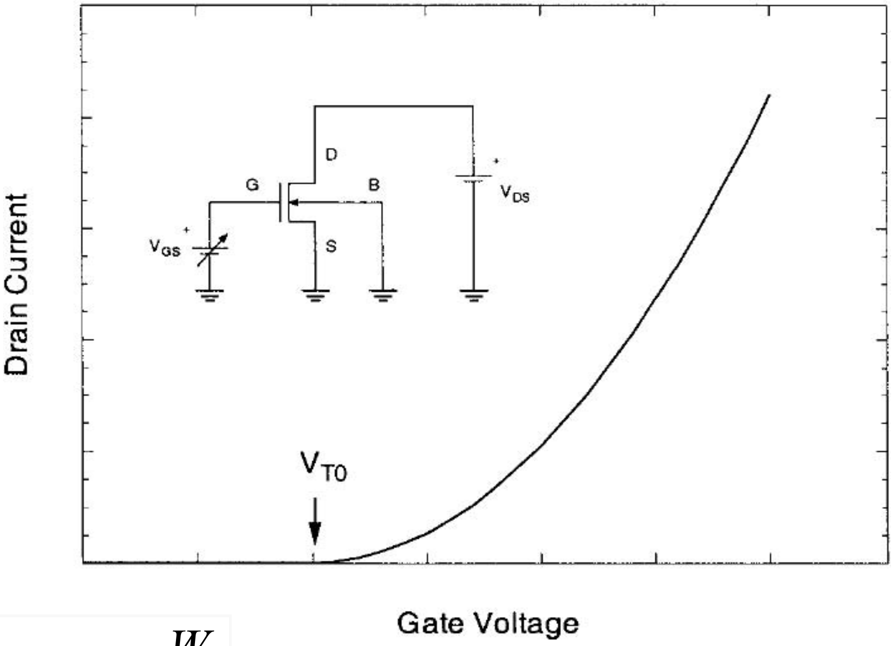
Operating at the edge of saturation



Operating beyond saturation

NMOS I_D-V_{DS} and I_D-V_{GS} Curves



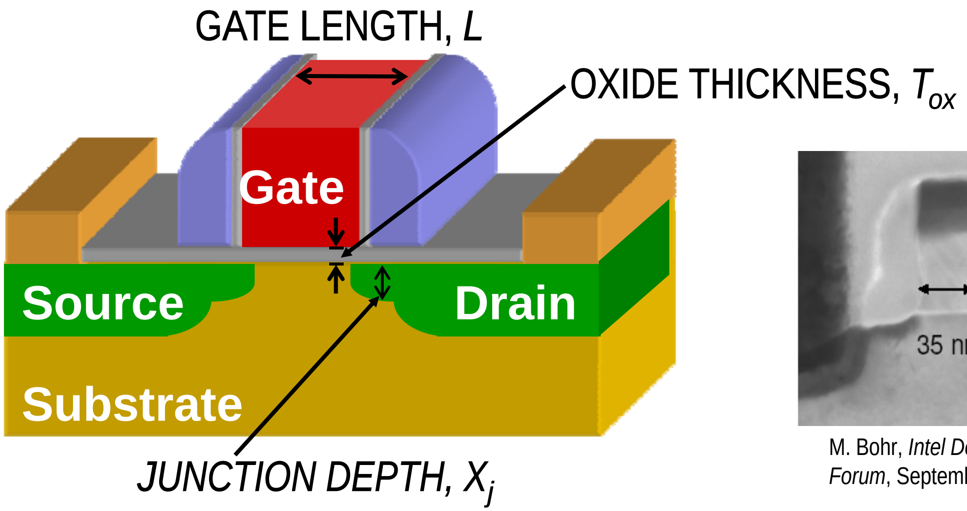


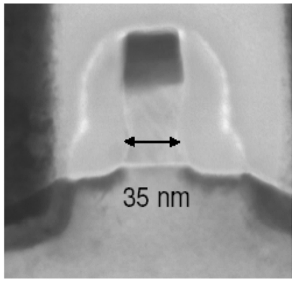
- nMOS transistor, with $k_n = k_n^{'} \dfrac{W}{L} = \mu_n C_{ox} \dfrac{W}{L}$

$$\boxed{\begin{aligned}
&I_D(cutoff) = 0 \quad V_{GS} < V_T \\
&I_D(lin) = \frac{k_n}{2}\left(2\left(V_{GS} - V_T(V_{SB})\right)V_{DS} - V_{DS}^2\right) \quad V_{GS} \geq V_T\,, V_{DS} < V_{GS} - V_T \\
&I_D(sat) = \frac{k_n}{2}\left(V_{GS} - V_T(V_{SB})\right)^2\left(1 + \lambda V_{DS}\right) \quad V_{GS} \geq V_T\,, V_{DS} \geq V_{GS} - V_T
\end{aligned}}$$

Short-Channel Effects

- A MOS transistor is called a short-channel device if its channel length is on the same order of magnitude as the depletion region thickness of the source and drain junctions
- The short-channel effects are attributed to two physical phenomena:
  1. Limitation on the electron drift characteristics in the channel
  2. Reduction of the threshold voltage due to shortening of the channel length





M. Bohr, *Intel Developer Forum*, September 2004

Short-Channel Effect on Electron Drift Characteristics

- In short-channel MOS transistor, the carrier velocity in the channel is also a function of the vertical component of the electric field, E_x
- Since the vertical field influences the scattering of the carriers in the surface, the surface mobility is reduced with respect to the bulk mobility
- The surface electron mobility can be expressed as follows:

$$\mu_n(eff) = \frac{\mu_{n0}}{1+\zeta\left(V_{GS}-V_T\right)}$$

where μ_{n0} is the low-field surface mobility and ζ is an empirical factor

Alpha-Power Current Equation for Short-Channel Devices

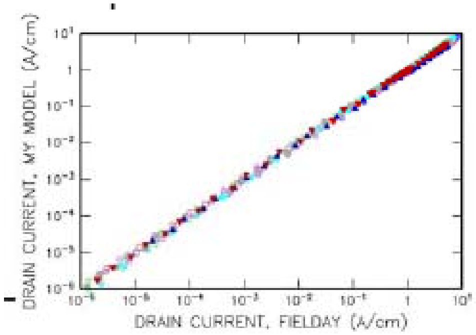
- In some textbooks, we see the following simplified equation for short-channel MOSFET current in saturation:

$$I_D(sat) = W v_{d,sat} C_{ox} \left(V_{GS} - V_T - V_{DSAT} \right)$$

- More often, we adopt the alpha-power current equation for short-channel MOSFETs, which is as follows:

$$\boxed{I_D(sat) = \frac{k_{n(p)}}{2} \left(V_{GS} - V_T \right)^{\alpha} \quad where \quad 1 < \alpha \le 2}$$

For a 60nm bulk CMOS process,$\alpha = 1.45$.



Plot of actual drain current vs. alpha-power current predictions for a 60nm bulk CMOS process

Example

- Suppose you were to design an NMOS transistor in a 0.18mm CMOS process. The transistor width is 0.72µm, and length is 0.18µm. The manufacturing process could result in a 25% variation in the threshold voltage, a 20% variation in the oxide thickness, and a 0.1µm variation in the width and in the length for the actual device that is fabricated. Assume that $V_{GS} = V_{DS} = 1.8V$ and that the threshold voltage is 0.5V. What is the ratio of the maximum value of the drain current to the minimum value of the drain current that could flow through the fabricated device when it is in saturation?
- **Solution:** We use the alpha power saturated current equation with α=1.4.

For max drain current, t_{ox1}=0.8t_{ox} , W_1=W+0.1µm=0.82µm, L_1=L-0.1µm=0.08µm

V_{th1}=0.75V_{th}=0.375, V_{th2}=1.25V_{th}=0.625

For min drain current, t_{ox2}=1.2t_{ox} , W_2=W-0.1µm =0.62µm, L_2=L+0.1µm=0.28µm

$$\frac{I_{d\max}}{I_{d\min}} = \frac{\frac{W_1}{t_{ox1}L_1}(V_{DD}-V_{th1})^{1.4}}{\frac{W_2}{t_{ox2}L_2}(V_{DD}-V_{th2})^{1.4}} = \frac{\frac{0.82}{0.8\times 0.08}(1.8-0.375)^{1.4}}{\frac{0.62}{1.2\times 0.28}(1.8-0.625)^{1.4}} = 9.1$$

Short Channel Effect (SCE) and Drain Induced Barrier Lowering (DIBL)

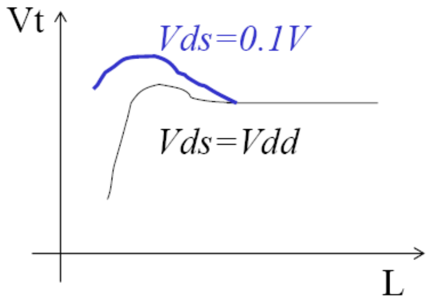
- In short-channel MOS, there is a significant amount of depletion charge around the source and drain, and therefore, the long channel model overestimates the depletion charge that must be supported by the gate voltage

$$\boxed{V_{T0}(short\ channel) = V_{T0} - \Delta V_{T0}}$$

- ΔV_{T0} is the threshold voltage reduction due to the short-channel effect:

$$\Delta V_{T0} = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{Si}N_A|2\phi_F|}\frac{x_j}{L}\left[\frac{1}{2}\left(\sqrt{1+\frac{2x_{dS}}{x_j}}+\sqrt{1+\frac{2x_{dD}}{x_j}}\right)-1\right]$$

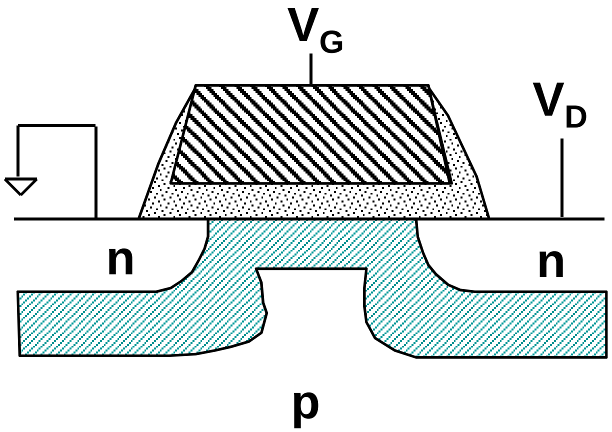
- V_T is also reduced due to the Drain Induced Barrier Lowering (DIBL)

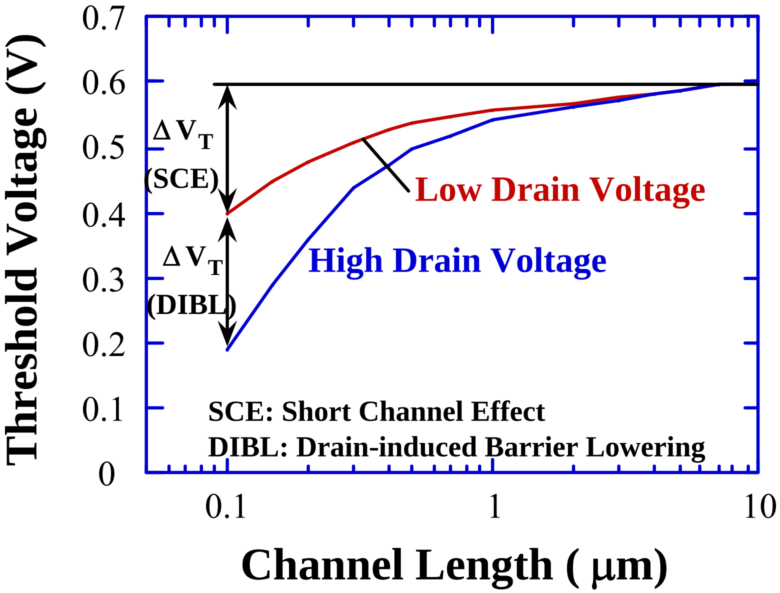


DIBL effect on barrier height (higher V_{DS} causes V_T of a short channel transistor to decrease)

Combined Effect of SCE and DIBL on Threshold Voltage

- Most noticeable in short-channel devices
  - Especially important in the subthreshold regime





Subthreshold Current (I_{sub})

- If the drain-source voltage is above 0V, the potential barrier for the electrons in the channel decreases and we have current even though$V_{GS}<V_{T0}$
- The channel current that flows under these conditions ($V_{GS}<V_{T0}$) is called the *sub-threshold current*:

$$I_D(sub) \equiv I_{sub} = \frac{W}{L}\mu_e(n-1)C_{ox}\vartheta_T^2\,\mathrm{e}^{\frac{V_{GS}-V_T+\eta V_{DS}}{n\vartheta_T}}\left(1-\mathrm{e}^{\frac{-V_{DS}}{\vartheta_T}}\right) \propto e^{\frac{V_{GS}-V_T+\eta V_{DS}}{n\vartheta_T}} = 10^{\frac{V_{GS}-V_T+\eta V_{DS}}{S}}$$

- The inverse subthreshold slope, S, is equal to the voltage required to increase I_D by 10X, i.e.,

$$S = \left(\frac{\partial(\log_{10} I_{sub})}{\partial V_{GS}}\right)^{-1} = n\vartheta_T \ln 10 = 2.3n\frac{kT}{q}$$

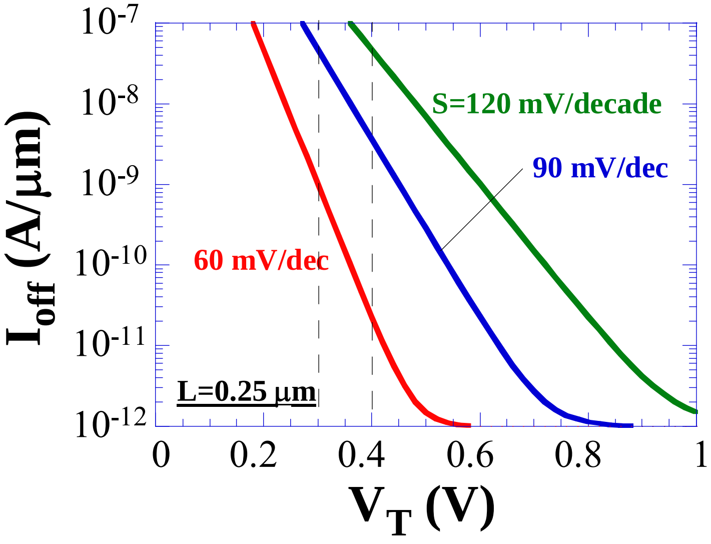
  - If n = 1, S = 60 mV/dec at 300 K
  - We want S to be small to shut off the MOSFET quickly
  - In well designed devices, S is 70 - 90 mV/dec at 300 K.

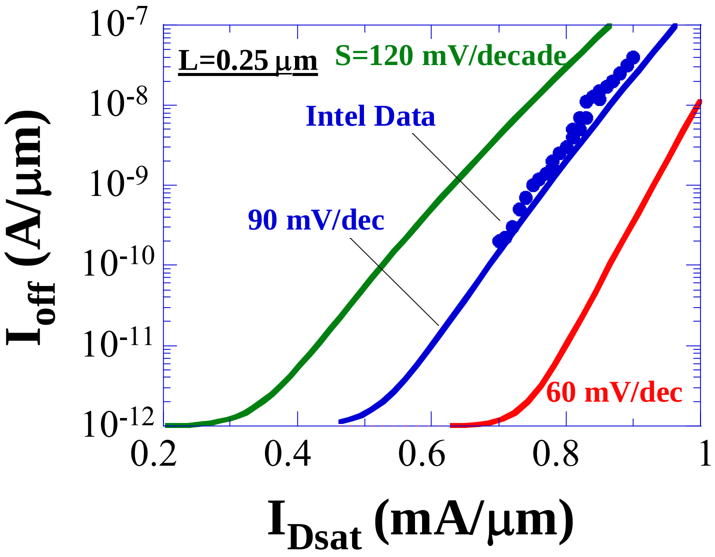
Modeling the Off Current (I_{off})

- Note that $n = 1 + \frac{C_{dep} + C_{it}}{C_{ox}}$
- Modulation of V_T in a short channel transistor
  - L ↓ ⇒ V_T ↓: "V_T Rolloff"
  - V_{DS} ↑ ⇒ V_T ↓:"Drain Induced Barrier Lowering"
  - V_{SB} ↑ ⇒ V_T ↑: "Body Effect"
- If $V_{DS} = 0 \Rightarrow I_{sub} = 0$
- long-channel device with $V_{DS} > 3n\vartheta_T \Rightarrow I_{sub} = \frac{W}{L}\mu_e(n-1)C_{ox}\vartheta_T^2 e^{\frac{V_{GS}-V_T}{n\vartheta_T}}$
- Now, we have: $\boxed{I_{off} \equiv I_{sub}(V_{GS} = 0) = \frac{W}{L}\mu_e(n-1)C_{ox}\vartheta_T^2 e^{\frac{-V_T}{n\vartheta_T}}}$
- Key dependencies of the subthreshold slope:
  - $t_{ox} \downarrow \Rightarrow C_{ox} \uparrow \Rightarrow n \downarrow \Rightarrow$ sharper subthreshold
  - $N_A \uparrow \Rightarrow C_{sth} \uparrow \Rightarrow n \uparrow \Rightarrow$ softer subthreshold
  - $V_{SB} \uparrow \Rightarrow C_{sth} \downarrow \Rightarrow n \downarrow \Rightarrow$ sharper subthreshold
  - $T \uparrow \Rightarrow$ softer subthreshold

Subthreshold Swing

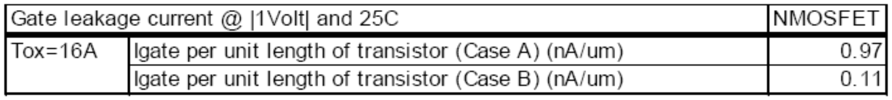
- V_T ⇩, I_{off} ⇧
- Subthreshold swing (S) ⇧ , I_{off} ⇧
- S ⇧ with increased doping density, reduced gate length (drain-induced barrier lowering)
- SOI is able to hold S = 60 mV/decade



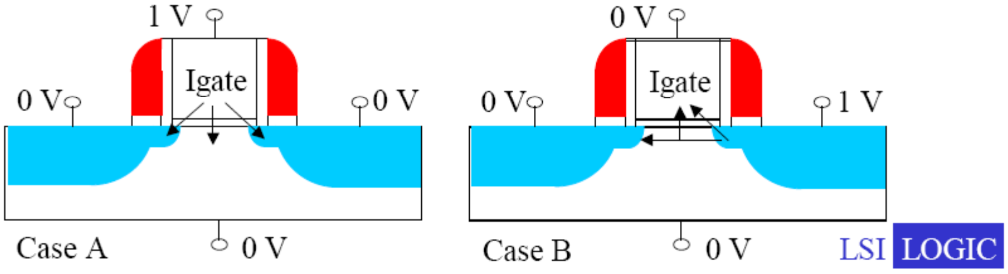


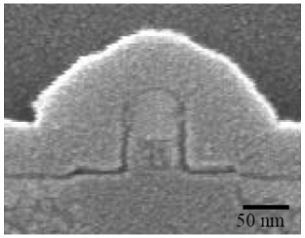
Gate Oxide Leakage

- At 15 Å the dielectric material will only have a bulk thickness of three atomic layers of silicon
  - Around this thickness, electrical leakage current through the dielectric becomes excessive and is expected to cause problems due to either high power dissipation or circuit reliability
- Gate oxide leakage current per unit length of the 65 nm nMOS FET (with t_{ox} of 15˚A) is below 1 nA/um
  - This gate leakage current is well below the transistor I_{off} of about 30nA/um at a V_{DD} of 1V and the nMOS transistor I_D(sat) of 775uA/um (this value is 270uA/um for the pMOS transistor)



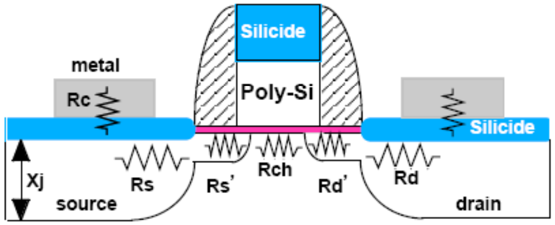
Gate leakage current @ |1Volt| and 25C		NMOSFET
Tox=16A	Igate per unit length of transistor (Case A) (nA/um)	0.97
	Igate per unit length of transistor (Case B) (nA/um)	0.11

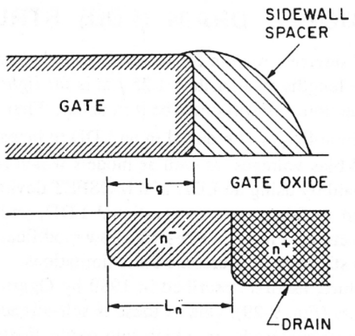


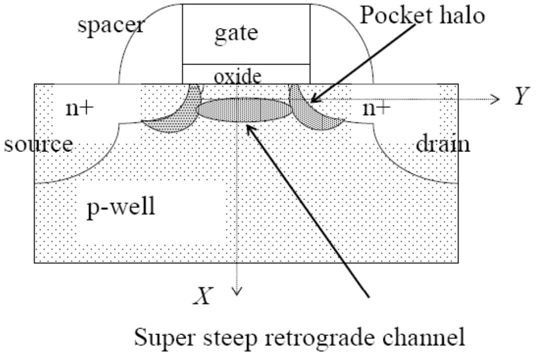


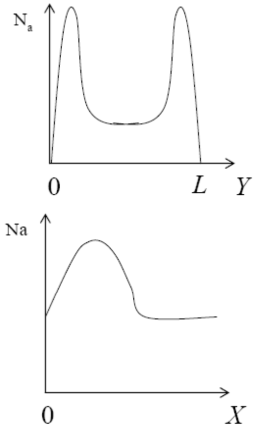
Modern Si MOSFET Structure

- Silicided junctions
  - Minimizes junction parasitic resistances
- Lightly doped drain (LDD)
  - A.k.a. shallow junction extension
  - Reduces hot carrier effect by lowering horizontal electric field
  - Need to watch for performance degradation
- Packet Halo implant and super steep retrograde channel implant with LDD
  - Control short channel effects
  - Suppresses punch-through





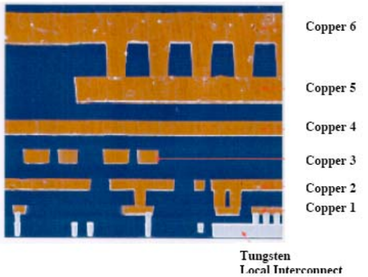




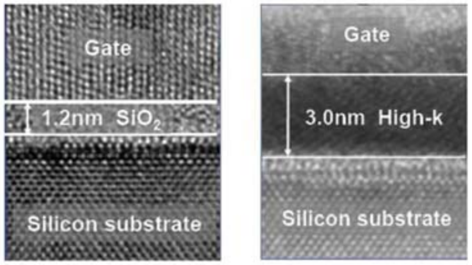
Source/Drain and Channel Engineering

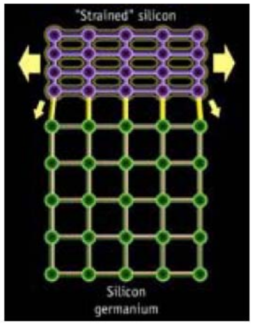
Modern Si MOSFET (Cont’d)

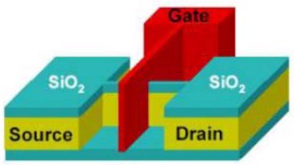
- Cu interconnect and low-k ILD
  - Improves VLSI interconnect performance
- High-k gate dielectric and metal gate
  - Example: replace SiO_2 (k=4) with Si_3N_4 (k=8) or HfO_2 (k ≈ 25)
  - Reduces tunneling gate leakage current by increasing the effective oxide thickness
- Strained Si or strained SiGe
  - Enhances electron/hole mobility and increases saturation velocity
- Multi-gate devices
  - Example: Double-gate FinFET
  - Provides better gate control over the channel
  - Improves the subthreshold swing

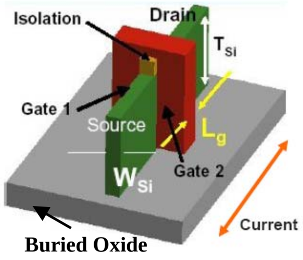


Current Cu technology





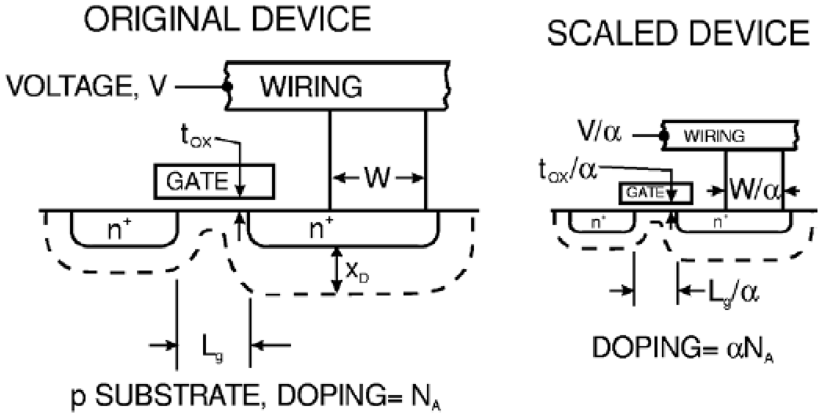




Constant-Field Scaling

- This scaling option attempts to preserve the magnitude of internal electric fields in the MOSFET, while the dimensions are scaled down by a factor of *S*
- The scaling factor for different parameters in this case is as follows:
  1. All dimensions, including those vertical to the surface, 1/S
  2. Device voltages, 1/S
  3. Concentration densities, S

Quantity	**Before Scaling**	**After Scaling**
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power supply voltage	V_{DD}	$V_{DD}' = V_{DD}/S$
Threshold voltage	V_{T0}	$V_{T0}' = V_{T0}/S$
Doping densities	N_A & N_D	$N_A' = S.N_A$ $N_D' = S.N_D$



Full Scaling (Cont’d)

- For linear mode and saturation mode drain current we have:

$$I_D^{'}(lin) = \frac{k_n^{'}}{2}\left[2\left(V_{GS}^{'} - V_T^{'}\right)V_{DS}^{'} - V_{DS}^{'2}\right]$$

$$= \frac{Sk_n}{2}\frac{1}{S^2}\left[2\left(V_{GS} - V_T\right)V_{DS} - V_{DS}^{2}\right] = \frac{I_D(lin)}{S}$$

$$I_D^{'}(sat) = \frac{k_n^{'}}{2}\left(V_{GS}^{'} - V_T^{'}\right)^2 = \frac{Sk_n}{2}\frac{1}{S^2}\left(V_{GS} - V_T\right)^2 = \frac{I_D(sat)}{S}$$

- For power dissipation of the transistor, we obtain:

$$P^{'} = I_D^{'}V_{DS}^{'} = \frac{1}{S^2}I_D V_{DS} = \frac{P}{S^2}$$

- With the device area reduction by S^2, we find that the *power density* (W/cm^2) remains unchanged for the scaled device

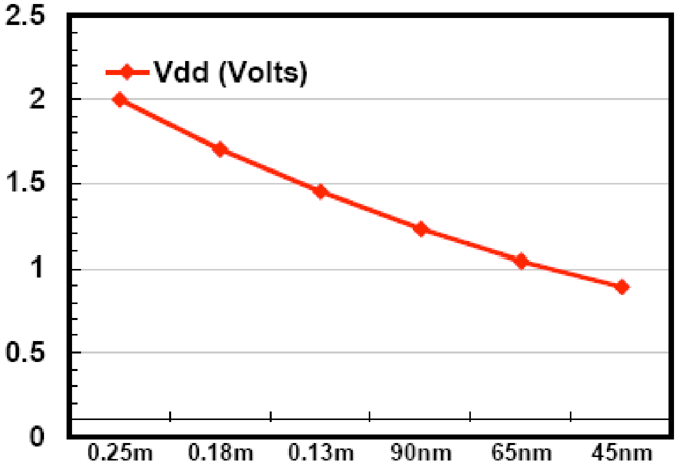
Example of Typical CMOS Scaling

- Consider a more realistic scaling scenario where voltage is scaled down by S while all dimensions and doping densities are scaled up by M:

$$V_{new} = \frac{1}{S} V_{old} \quad , \quad I_{new} = \frac{M}{S^2} I_{old} \quad , \quad (C_L)_{new} = \frac{1}{M} (C_L)_{old}$$

$$freq_{new} = \frac{M^2}{S} freq_{old} \quad , \quad power_{new} = \frac{M}{S^3} power_{old}$$

$$energy_{new} = \frac{1}{MS^2} energy_{old} \quad , \quad pow\_dens_{new} = \frac{M^3}{S^3} pow\_dens_{old}$$



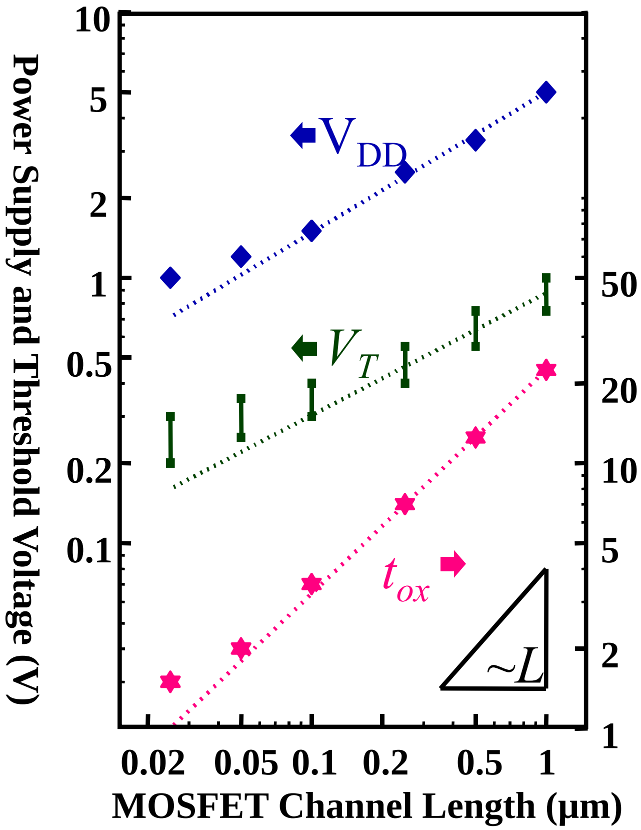
- With S^{-1}=0.85 and M^{-1}=0.7, we obtain:

$$energy_{new} = 0.506\, energy_{old} \quad , \quad pow\_dens_{new} = 1.79\, pow\_dens_{old}$$

- With each generation, voltage has decreased to 0.85x, not 0.7x for constant field scaling. Thus, energy dissipation per logic gate decreases by (1-0.85^2*0.7)=50% rather than by the ideal (1-$0.7^{3)}$=66% per generation
- However, the number of logic gates in a chip has been increasing by 3x per generation (since the die size is increasing correspondingly), thus a net increase in the energy consumption per chip
- The power density is increasing by about 80% per generation

Reality of CMOS Scaling

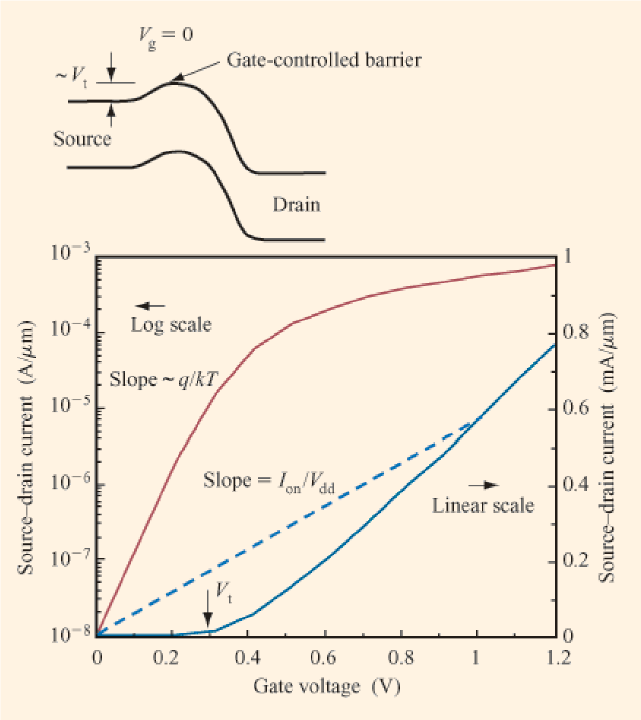
- Scaling increases:
  - Transistor density and functionality
  - Speed of operations
- V_{DD} scaling is needed to maintain device reliability and reduce power dissipation
  - V_T scaling needed to maintain switching speeds
  - t_{ox} scaling needed to maintain the current drive and keep V_T variations under control when dealing with short-channel effects
- V_T /V_{DD} is increasing with scaling
  - Effect of process variations on delay becomes higher
  - Delay sensitivity becomes intolerable when the V_T /V_{DD} ratio is at 0.5 or higher



Source: Taur

Fundamental Limits to CMOS Scaling

- Fundamental non-scaling effects are caused by the fact that neither the thermal voltage kT/q nor the silicon band gap changes with scaling
  - The first results in non-scaling of the subthreshold swing parameter
  - The latter results in non-scalability of built-in junction potential, depletion layer width, and short channel effects
- Maximum integration density is limited by the power density while maximum circuit speed is limited by the *parametric variability*
  - Because of the field dependence of the carrier mobility, the gate speed will not improve linearly with scaling
  - There is adverse impact on device reliability due to high electric field stress



Source: Taur

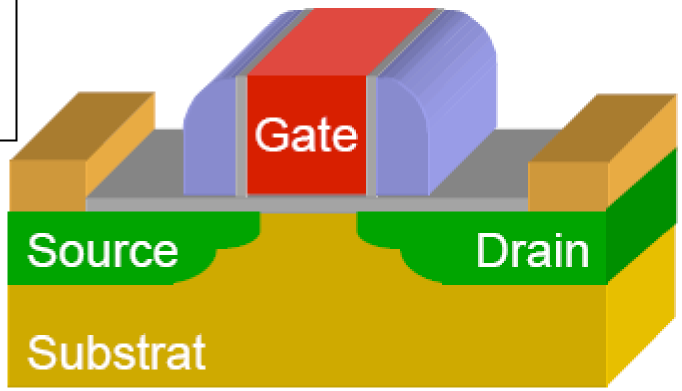
Physical Limits in Scaling Si MOSFET

Gate stack

- Tunneling current
- Gate depletion, resistance

Source/Drain

- Contact resistance
- Band-to-band tunneling
- Doping level, abruptness



High E-Field

- Mobility degradation
- Reliability

Channel

- Surface scattering - the “universal mobility” tyranny
- Subthreshold slope limited to 60mV/decade (kT/q)
- V_G - V_T decrease
- DIBL ⇒ leakage

Power Dissipation and Temperature

- Power consumption and heat removal are limited by practical considerations:
  - Low power applications must be battery powered
  - Many must be light-weight → power < ~few watts
  - Disposable batteries can cost > $500/watt over the life of device
  - Rechargeables can cost > $50/watt over the life of device
- Home electronics is limited to < ~1000W due to heat generation in the rooms and the cost of electricity
- High performance is limited by difficulty of heat removal from chip (~100 W/chip) (Cost of electricity is ~$5/watt over the life of device)
- Every 10°C increase on operating temperature double failure rate for the electronic components

Components of Power Consumption

- The power consumption in CMOS digital circuits has three main components:
  - Capacitive (switching) power consumption
  - Short-circuit (rush-thru) power consumption

  } dynamic power consumption

  - Leakage power consumption
- Chips with circuits other than conventional CMOS gates that have continuous current paths between the power supply and the ground, have an extra power component:
  - Static (DC) power consumption

Switching Power Consumption

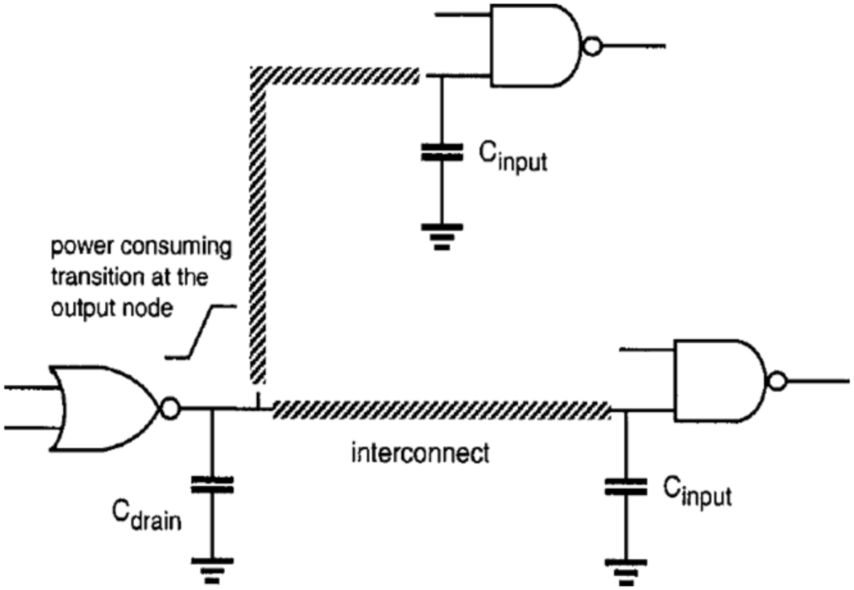
- In digital CMOS circuits, switching power is dissipated when energy is drawn from the power supply to charge up the output node capacitance

- Total capacitive load at the output of a NOR gate consists of

  i) the output node cap. of the gate itself

  ii) the total interconnect cap.

  iii) the input cap. of the driven gates



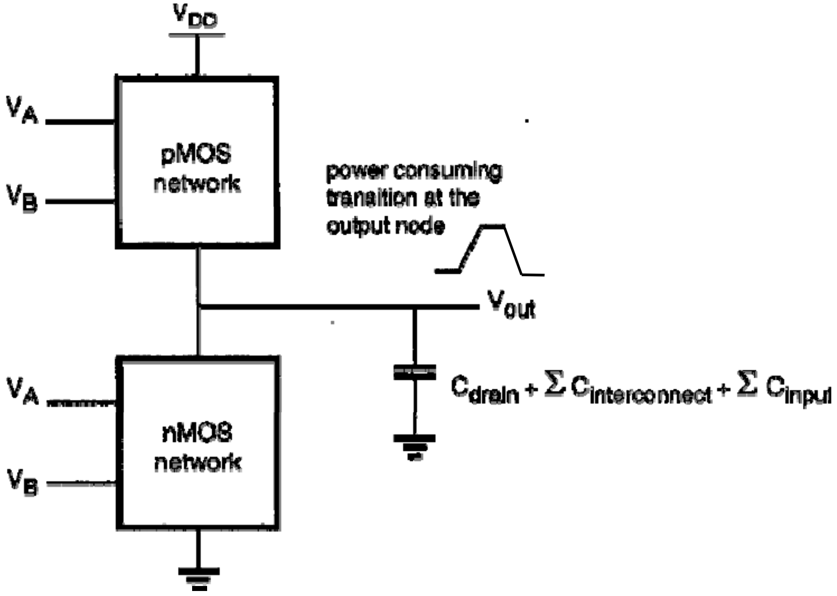
A NOR gate driving two NAND gates through interconnection lines

Derivation of Switching Power Consumption

- The average power dissipation can be calculated from the energy required to charge up the output node to V_{DD} and charge down the total output load capacitance to ground level.

$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2} V_{out}\left(-C_{load}\frac{dV_{out}}{dt}\right)dt + \int_{T/2}^{T}(V_{DD}-V_{out})\left(C_{load}\frac{dV_{out}}{dt}\right)dt\right]$$

$$P_{avg} = \frac{1}{T}C_{load}V_{DD}{}^2 \quad \text{or} \quad P_{avg} = C_{load}V_{DD}{}^2 f_{CLK}$$


V_DD
V_A
V_B
pMOS network
power consuming transition at the output node
V_out
V_A
V_B
nMOS network
C_drain + Σ C_interconnect + Σ C_input

Switching Power Consumption (Cont'd)

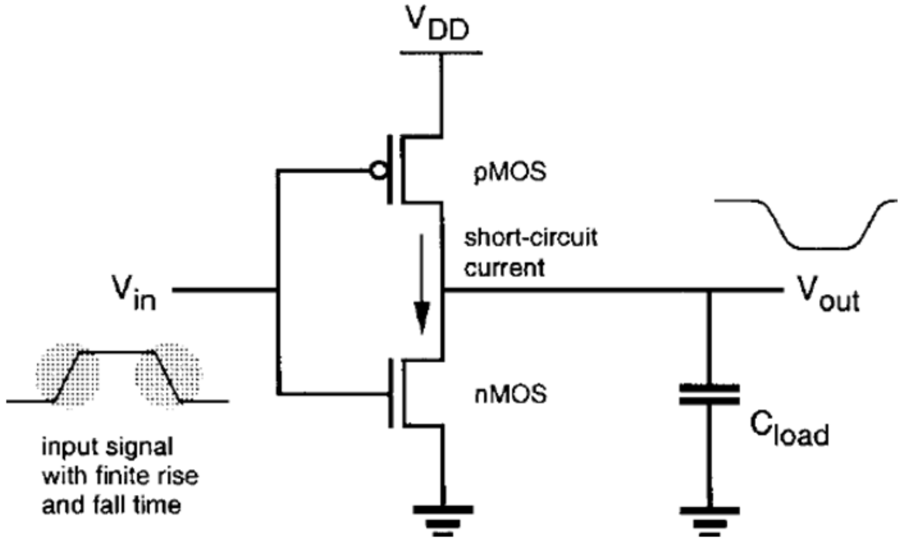
- Internal node voltage transitions can be partial, i.e., the node voltage swing may be only *ΔV*$_i$, which is in general smaller than the full voltage swing of V_{DD}

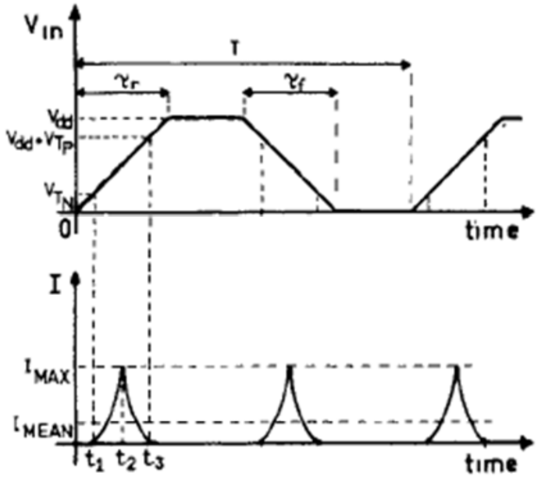
$$\boxed{P_{avg} = \frac{1}{2} V_{DD} f_{CLK} \left(\sum_{i=1}^{\#of\ nodes} \beta_i \, C_i \, \Delta V_i \right)}$$

where C_i = the parasitic capacitance associated with each node in the circuit
β_i = the corresponding activity factor associated with the node

Short-Circuit Power Dissipation

- Let $\tau_r = \tau_f = \tau$ denote the transition time of the input voltage, V_{in}
- Now t_1 is the time when the input voltage reaches the threshold voltage of nMOS while t_3 is the time when the input voltage reaches the threshold voltage of pMOS
- The short-circuit current flows between t_1 and t_3 and reaches it maximum at t_2 when V_{out}=V_{dd}/2





Short Circuit Power Calculation

- Turgis et al model, which is based on the concept of an equivalent short circuit capacitance calculated under the assumption that the input and the output waveforms are linear, is as follows:

$$I_{sc}(\text{rising input}) = \frac{1}{6} k_p \tau_{in,r} V_{DD}^{2} \left(1 - \frac{V_{T,n} + |V_{T,p}|}{V_{DD}}\right)^2 \left(\frac{\tau_{in,r}}{\tau_{in,r} + \tau_{out,f}}\right) f_{CLK}$$

$$C_{sc} = \frac{I_{sc} T}{V_{DD}} = \frac{1}{6} k_p \tau_{in,r} V_{DD} \left(1 - \frac{V_{T,n} + |V_{T,p}|}{V_{DD}}\right)^2 \left(\frac{\tau_{in,r}}{\tau_{in,r} + \tau_{out,f}}\right)$$

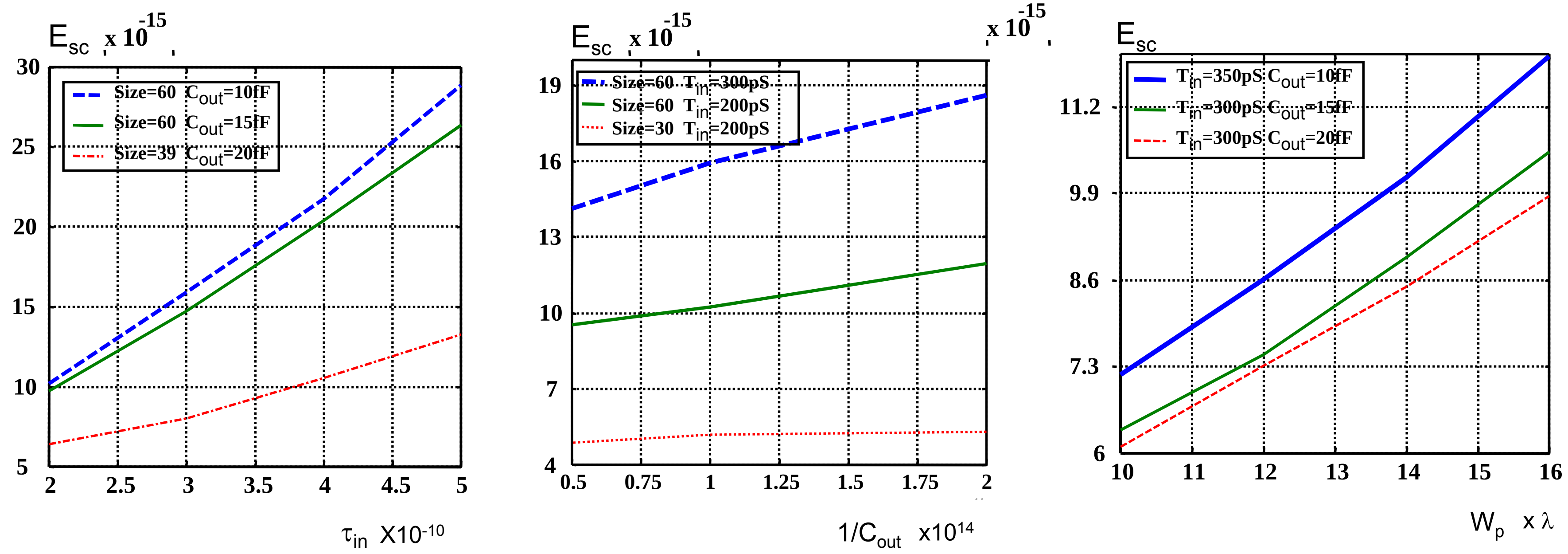
$$P_{sc}(\text{rising input}) = \frac{1}{2} C_{sc} V_{DD}^{2} f_{CLK} = \frac{1}{12} k_p \tau_{in,r} V_{DD}^{3} \left(1 - \frac{V_{T,n} + |V_{T,p}|}{V_{DD}}\right)^2 \left(\frac{\tau_{in,r}}{\tau_{in,r} + \tau_{out,f}}\right) f_{CLK}$$

- For a symmetric CMOS inverter with $k_n = k_p = k$, $V_{T,n} = |V_{T,p}| = V_T$, and equal input rise and fall times, the above equation becomes:

$$\boxed{P_{sc} = \frac{1}{12} k \, \tau_{in} V_{DD} \left(V_{DD} - 2V_T\right)^2 \left(\frac{1}{1 + \tau_{out}/\tau_{in}}\right) f_{CLK} \beta}$$

which almost reduces to Veendrick's result for $\tau_{out}/\tau_{in} = 1$

Regression-based Short Circuit Power Equation



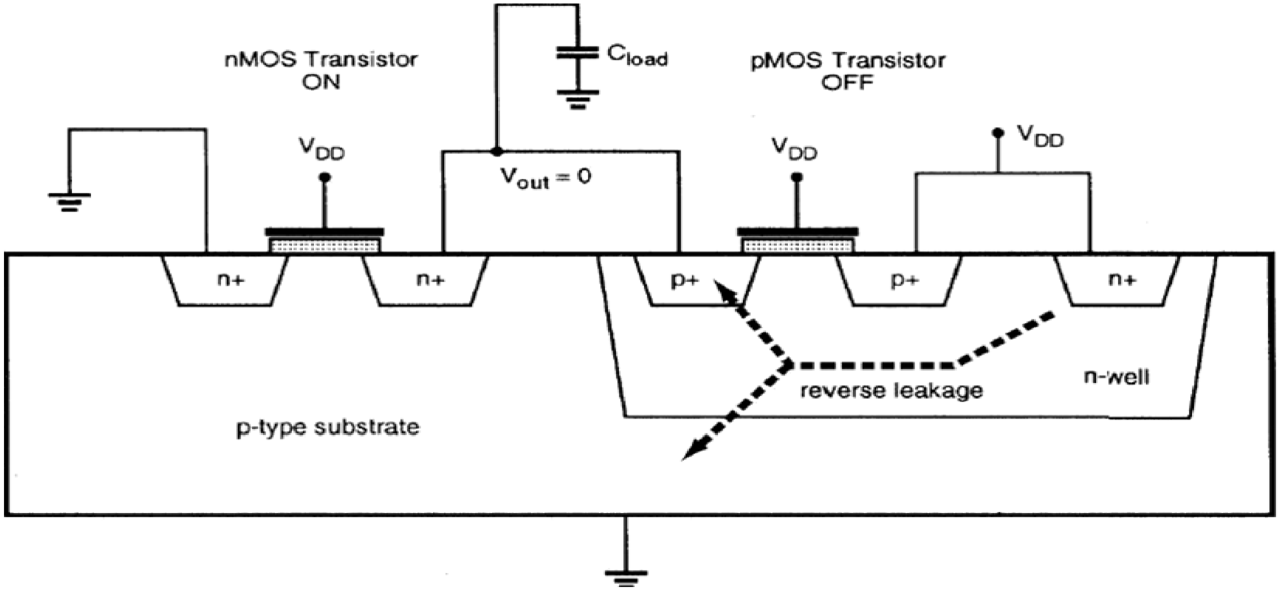
$$P_{sc}\left(\tau_{in},k,C_{out}\right)\propto\frac{k\,\tau_{in}}{C_{out}}V_{DD}f_{CLK}\beta$$

Dynamic Power Minimization Techniques

- Power management
  - Dynamic voltage and frequency scaling
  - Multiple voltage islands
- Trading area or latency for power
  - Pipelining
  - Parallelization
- Glitch suppression
- Clock gating
- Driving buses
  - Bus encoding
  - Low Swing buses and split buses
- Adiabatic circuits, stepwise charging, charge recycling

Reverse-Biased Junction Leakage

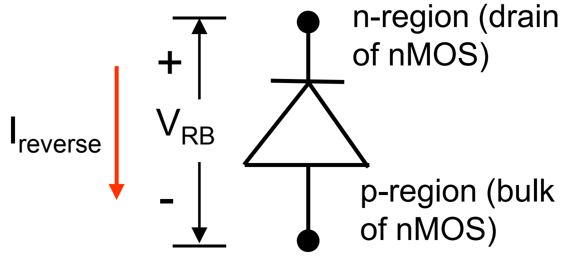
- Consider a CMOS inverter with a high input voltage
  - Although pMOS transistor is turned off, there will be a reverse potential difference of V_{DD} between its drain and the n-well



Reverse-Biased Junction Leakage (Cont'd)

- The reverse leakage current of a pn-junction is expressed by

$$\boxed{I_{reverse} = A \cdot J_s \cdot (1 - e^{-\frac{V_{RB}}{n\vartheta_T}})}$$

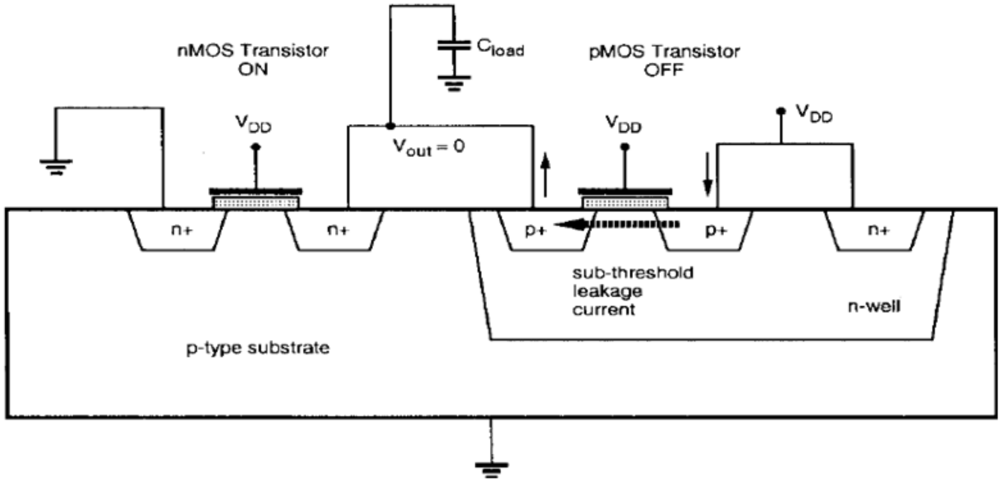


  - A : the junction area
  - J_s : the *maximum reverse saturation current density* (typically 1-5 pA/μm^2)
  - n : the *emission coefficient*, usually set to 1, although can be larger depending on the type of junction
  - V_{RB} : the *reverse bias voltage* across the junction, i.e., the voltage of drain diffusion with respect to the bulk or well
  - $\upsilon_T = kT/q$ denotes the thermal voltage at absolute junction temperature, T

- $I_{reverse}$ is maximum when V_{RB} is largest, that is why we focus on the drain side and not the source side of the nMOS transistor

Subthreshold Conduction Leakage

- Another component of leakage current is the subthreshold current, which is due to carrier diffusion between the source and the drain regions of the transistor in weak inversion



- The subthreshold leakage current can occur even when there is no switching activity in the circuit

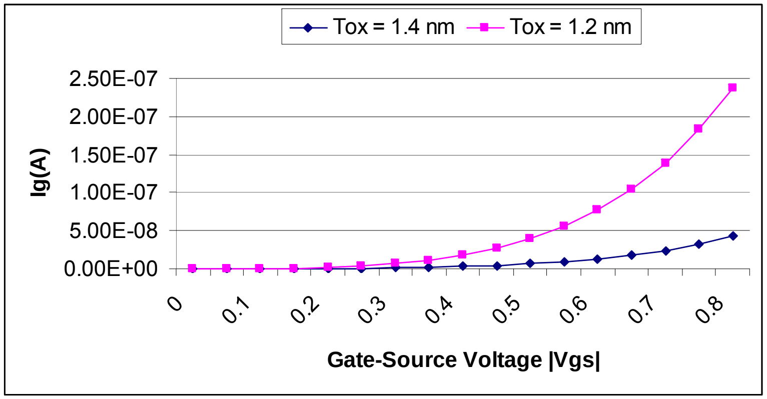
$$\boxed{I_{subthreshold} \cong \mu_0 C_{ox} \frac{W}{L} \vartheta_T^2 (n-1)\, \mathrm{e}^{\frac{V_{GS} - V_T + \eta V_{DS}}{n\vartheta_T}}}$$

Gate Leakage

- With the advent of deep-submicron devices comes the reduction of the gate-oxide thickness. This reduction leads to a higher electric field across the oxide
  - The tunneling of electrons through the gate oxide into the substrate and from substrate to the gate becomes possible. This current is referred to as gate leakage
  - Being quantum mechanical in nature, the gate leakage current is virtually temperature-independent

$$\boxed{I_{gate} \cong \kappa WL \left(\frac{V_{GB}}{t_{ox}} \right)^2 e^{-\alpha \frac{t_{ox}}{V_{GB}}}}$$

where κ and α are fitting parameters, W is the transistor width, V_{GB} denotes the gate to bulk voltage, and t_{ox} denotes the gate oxide thickness



Source and bulk are tied together, i.e., $V_{GB}=V_{GS}$

Total Power Dissipation in CMOS VLSI Circuits

- The total power dissipation is the sum of two components: dynamic (switching plus short-circuit) and leakage (reverse biased junction, subthreshold and gate currents)

$$P_{total} = \frac{1}{2}\left(C_{load} V_{DD} + \frac{k\,\tau_{in}}{6\left(1+\frac{\tau_{out}}{\tau_{in}}\right)} \left(V_{DD} - 2V_T\right)^2 \right) V_{DD} f_{CLK} \beta + V_{DD} I_{leakage}$$

$$I_{leakage} = I_{reverse} + I_{subthreshold} + I_{gate}$$

Example

Calculate the capacitive, short circuit, and leakage components of power dissipation of a CMOS inverter with $W_p/L=2W_n/L=8$, driving an identical inverter with the following parameters: $\mu_n = 2\mu_p = 600\text{cm}^2/\text{V-sec}$, $C_{ox}=2\times10^{-7}\ \text{F/cm}^2$, $V_{T,n}=-V_{T,p}=0.6\text{V}$, $V_{DD}=2.8\text{V}$, $t_{in}=100\text{ps}$, $t_{out}=300\text{ps}$, activity factor $\beta=0.2$, $f_{CLK}=500\text{MHz}$, die Temperature $T=85$ °C, the subthreshold shape parameter, $n=1.5$, Boltzmann constant, $k=1.38\times10^{-23}\text{J/K}$, electron/hole charge, $q=1.6\times10^{-19}$ C, and $L=0.25\mu\text{m}$.

$$k'_n = \mu_n C_{ox} = 2k'_p = 2\left(\mu_p C_{ox}\right) = \left(600\frac{cm^2}{V\cdot s}\right)\left(2\times10^{-7}\frac{F}{cm^2}\right) = 120\frac{\mu A}{V^2}$$

$$k_n = k_p = \mu_n C_{ox}\frac{W_n}{L} = \mu_p C_{ox}\frac{W_p}{L} = \left(120\frac{\mu A}{V^2}\right)(4) = \left(60\frac{\mu A}{V^2}\right)(8) = 480\frac{\mu A}{V^2}$$

$$C_{load} = \left(W_n + W_p\right)LC_{ox} = 3W_nLC_{ox} = \frac{3}{2}W_pLC_{ox} = (12\times10^{-4}cm)(2\times10^{-4}cm)\left(2\times10^{-7}\frac{F}{cm^2}\right) = 48fF$$

$$\vartheta_T = 1.38\times10^{-23}\frac{273+85}{1.6\times10^{-19}} = 308.8\times10^{-4}V = 30.9mV$$

Note that since inverter is driving identical load: $\mu_n C_{ox}\frac{W_n}{L} = \mu_n\frac{C_{load}}{3L^2} = \left(2\mu_p\right)\frac{C_{load}}{3L^2} = \mu_p C_{ox}\frac{W_p}{L}$

$$P_{total} = \frac{1}{2}C_{load}V_{DD}{}^2 f_{CLK}\beta + \frac{k_n\tau_{in}}{12\left(1+\frac{\tau_{out}}{\tau_{in}}\right)}\left(V_{DD}-2V_T\right)^2 V_{DD}f_{CLK}\beta + \frac{\mu_n C_{load}}{3L^2}\vartheta_T^2(n-1)\,\mathrm{e}^{\frac{-V_T}{n\vartheta_T}}V_{DD}$$

$$P_{total} = \frac{1}{2}(48f)(2.8V)^2(500MHz)(0.2)$$

$$+\frac{\left(480\frac{\mu A}{V^2}\right)(100ps)}{12\left(1+\frac{300}{100}\right)}\left(2.8V-2\times0.6V\right)^2(2.8V)(500MHz)(0.2)$$

$$+\frac{\left(600\frac{cm^2}{V\cdot s}\right)(48fF)}{3\left(0.25\times10^{-4}cm\right)^2}\left(0.0309V\right)^2(0.5)e^{\frac{-0.6}{1.5\times0.0309}}(2.8V)$$

$$=18.816\mu W + 0.717\mu W + 0.863nW$$

Example (Cont'd)

Consider a case in which the circuit is in busy state consuming capacitive and short-circuit power for time T_{busy} and then remains idle for a time T_{idle} during which only leakage power is dissipated. Let's define the *duty factor* as ψ = T_{busy} / (T_{busy} + T_{idle}). Calculate the minimum value of ψ such that the idle state energy dissipation is no more than 10^{-3} times the total energy dissipation of the CMOS inverter.

$$\psi = \frac{T_{busy}}{T_{busy} + T_{idle}} = \frac{1}{1 + \frac{T_{idle}}{T_{busy}}}$$

$$E_{idle} \le 10^{-3} \times E_{tot}$$

$$10^3 \times P_{leak} \times T_{idle} \le (P_{cap} + P_{sc}) \times T_{busy} + P_{leak} \times T_{idle}$$

$$\frac{T_{idle}}{T_{busy}} \le \frac{P_{cap} + P_{sc}}{999 \times P_{leak}} = \frac{18.816\mu W + 0.717\mu W}{999 \times 0.863 nW} = \frac{19.533\mu W}{862.137 nW} = 22.66$$

$$\psi \ge \frac{1}{1 + 22.66} \approx 0.0423$$

Effects due to High Die Temperatures

- Thermal effects are an inseparable aspect of electrical power generation and signal transmission
  - They arise from the substrate power generation and self-heating in the interconnects
- High temperature reduces the interconnect performance due to increase in electrical resistance and lowers the mean time to failure (MTTF) of VLSI interconnections due to more severe Electro-migration effect
  - Every 10 degrees Celsius increase in the die temperature increases wire delays by 5% and reduces the MTTF by 50%
- They are expected to become more severe due to CMOS technology scaling

Electrical-Thermal Analogy

- Analogous quantities
  - Electrical potential, *V (Volt)* ⇔ Temperature, *T* (Kelvin)
  - Charge, *Q* (Coulomb) ⇔ Heat, *q* (Joule)
  - Current, *I* (Ampere) ⇔ Heat flux = power, *P=dq/dt,* (*Watt*)
  - Electrical resistance, *R* (V/I=Ω) ⇔ Thermal resistance, R_T (K/W)
  - Electrical capacitance, *C* (C/V=F) ⇔ Thermal capacitance, C_T (J/K)

$$\boxed{\frac{\partial^2 V}{\partial z^2} = RC\frac{\partial V}{\partial t} \leftrightarrow \frac{\partial^2 T}{\partial z^2} = R_T C_T \frac{\partial T}{\partial t}}$$

- Analogous laws

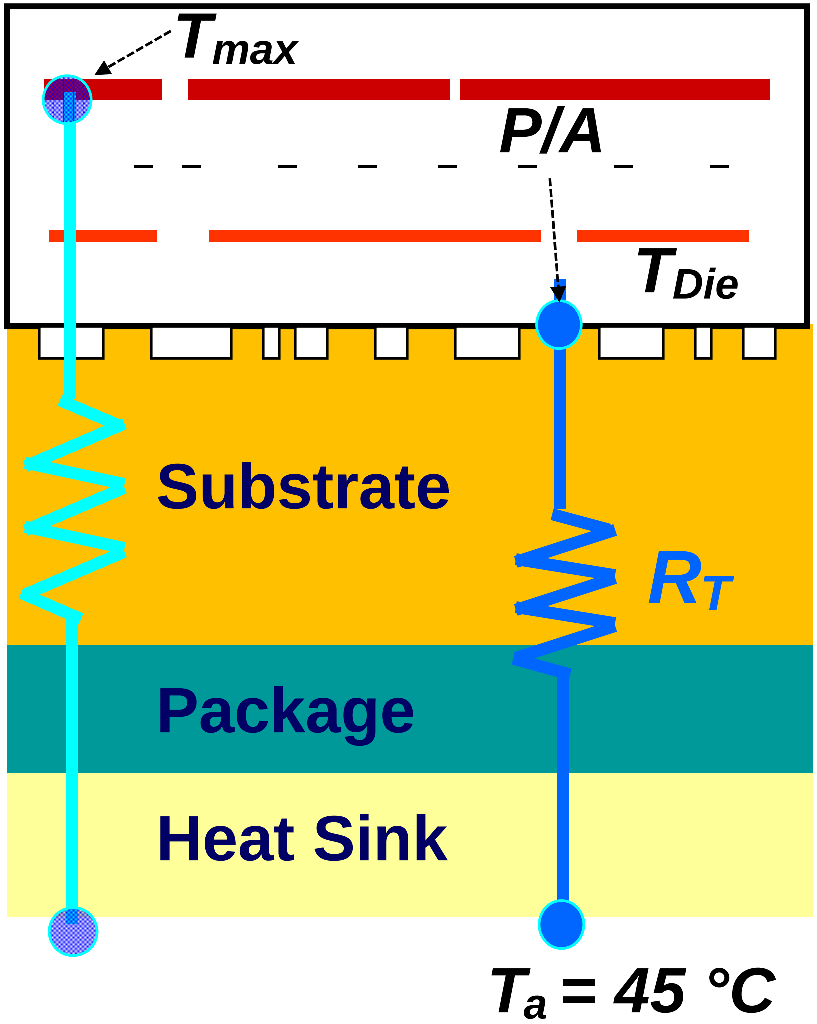
$$\boxed{\begin{aligned} &V = RI \leftrightarrow T = R_T P \\ &I = C\frac{\partial V}{\partial t} \leftrightarrow P = C_T \frac{\partial T}{\partial t} \end{aligned}}$$

Die Temperature Calculation

- 1-D heat conduction model

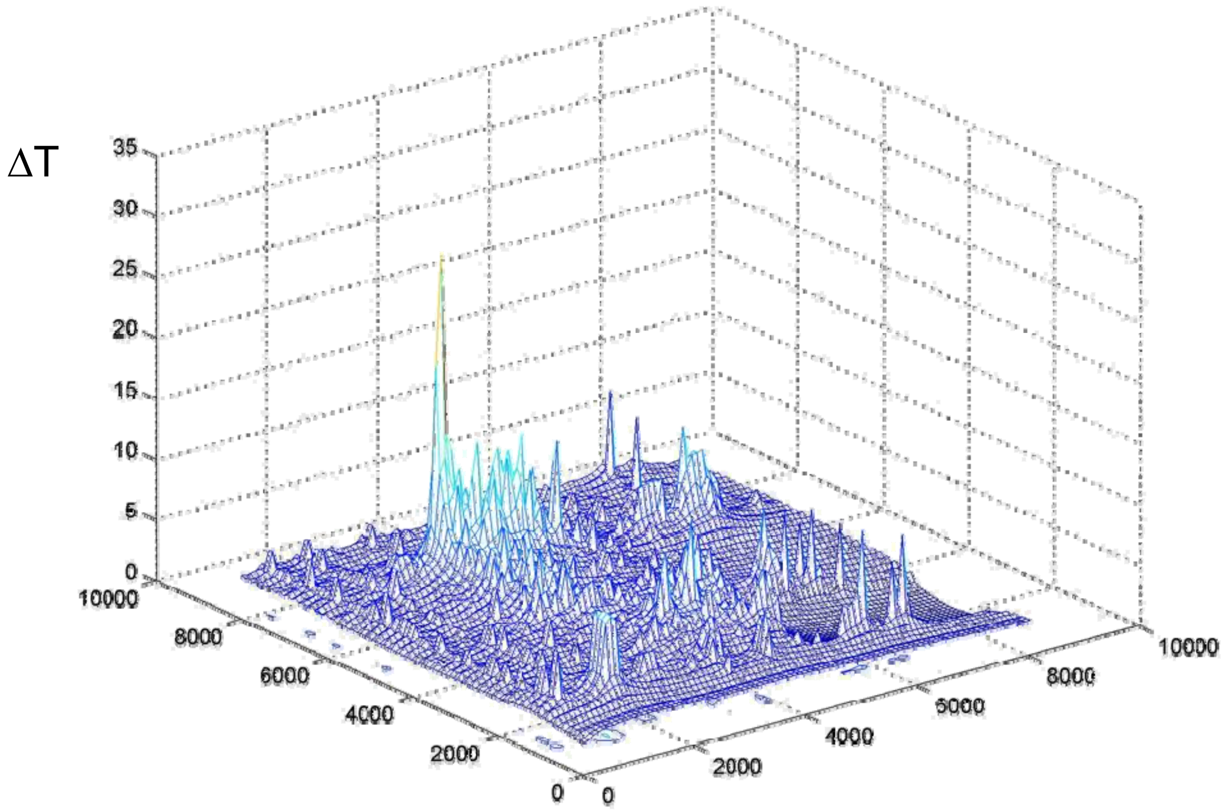
$$\boxed{T_{Die} = T_a + R_T \left(\frac{P}{A} \right)}$$

- T_{Die} = 120 °C (180 nm)
- R_T = 1.07 cm^2 °C/W
- For given packaging and cooling technologies (R_T), the die temperature (T_{die}) can be calculated for any ambient temperature (T_a) and any technology node (P and A)
- Note that maximum temperature occurs in uppermost metal lines



Example of Die Temperature Profile

- Thermal map of a 9mm by 9mm ASIC chip – Su, ISLPED 2003



V_T Dependence on Temperature

- Assuming that Φ_{GC} and qN_{ox} remains unchanged with temperature and that $n_i \propto T^{1.5}$, we have:

$$V_T = \left(\Phi_{GC} - \frac{qN_{OX}}{C_{OX}} + \frac{qN_I}{C_{OX}} \right) + \left(2\frac{kT}{q}\ln(\frac{N_A}{n_i}) + \frac{\sqrt{2qN_A\varepsilon_{Si}\left(\frac{kT}{q}\ln(\frac{N_A}{n_i}) + V_{SB}\right)}}{C_{OX}} \right)$$

$$= A + BT\ln(\frac{\kappa}{T^{1.5}}) + C\sqrt{T\ln(\frac{\kappa}{T^{1.5}})} \qquad A,B,C,D > 0, \text{Ignoring } V_{SB}$$

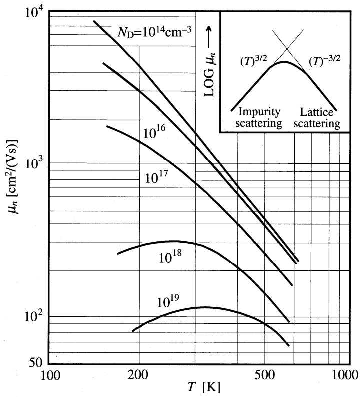
$$\frac{\partial V_T}{\partial T} = B\left(\ln(\frac{\kappa_1}{T^{1.5}}) - \frac{3}{2} \right) + \frac{C\left(\ln(\frac{\kappa_1}{T^{1.5}}) - \frac{3}{2} \right)}{2\sqrt{T\ln(\frac{\kappa_1}{T^{1.5}})}}$$

Hard to say, but we expect: $\frac{\partial V_T}{\partial T} < 0$

- V_T decreases with Temperature, i.e., the gate overdrive voltage, V_{GS}–V_T, goes up at higher temperatures

Mobility Dependence on Temperature

- For short channel devices, the surface electron mobility is expressed as follows (V_{SB}=0V):



$$\mu_n(eff) = \frac{\mu_{n0}}{1+\zeta\left(V_{GS}-V_T\right)} \qquad \zeta \geq 0$$

$$\frac{\alpha T^{-1.5}}{1+\zeta\left(V_{GS}-A-BT-C\sqrt{T}\right)} = \frac{\alpha}{(1+\zeta V_{GS}-\zeta A)T^{1.5}-\zeta CT^2-\zeta BT^{2.5}} \qquad \alpha \geq 0$$

$$\frac{\partial\mu_n}{\partial T} = \frac{-\alpha\left(\frac{3}{2}(1+\zeta V_{GS}-\zeta A)T^{0.5}-2\zeta CT-\frac{5}{2}\zeta BT^{1.5}\right)}{\left((1+\zeta V_{GS}-\zeta A)T^{1.5}-\zeta CT^2-\zeta BT^{2.5}\right)^2}$$

Hard to say, but we expect: $\frac{\partial\mu_n}{\partial T} < 0$

- Carrier mobility degrades at higher temperatures

Temperature Effect on the ON current (I_{on})

- We consider the I_D(sat) equation here:

$$I_D(sat) = \frac{\mu_n C_{ox}}{2}\frac{W}{L}\left(V_{GS} - V_T\right)^2\left(1 + \lambda V_{DS}\right) \qquad \lambda \geq 0$$

$$\frac{\partial I_D}{\partial T} = A\left(\left(V_{GS} - V_T\right)^2 \frac{\partial \mu_n}{\partial T} - 2\mu_n\left(V_{GS} - V_T\right)\frac{\partial V_T}{\partial T}\right)\left(1 + \lambda V_{DS}\right)$$

Hard to say, but we expect : $\frac{\partial I_D}{\partial T} < 0$

- I_{on} decreases with temperature
- Increase in gate overdrive is smaller compared to carrier mobility degradation when the temperature goes up. That is why the MOSFET drain current degrades when the temperature is increased from 25°C to 125°C
  - For a 65 nm node, the V_T of an nMOS decreases by about 40mV for this temperature rise range; the carrier mobility is cut in nearly half

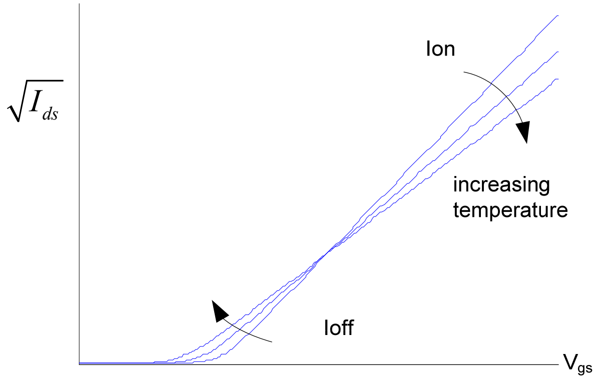
Effect on the Off Current (I_{off})

$$I_{off} = \frac{W}{L}\mu_e (n-1) C_{ox} \left(\frac{kT}{q}\right)^2 e^{\frac{q\left(-A-BT-C\sqrt{T}\right)}{nkT}} = \rho T^2 e^{-\beta T^{-1}-\chi-\eta T^{-0.5}} \qquad \rho,\beta,\chi,\eta>0$$

$$\frac{\partial I_{sub}}{\partial T} = 2\rho T e^{-\beta T^{-1}-\chi-\eta T^{-0.5}} + \rho T^2\left(\beta T^{-2} + \frac{\eta}{2} T^{-1.5}\right) e^{-\beta T^{-1}-\chi-\eta T^{-0.5}}$$

$$= \rho\left(2T + \beta + \frac{\eta}{2}\sqrt{T}\right) e^{-\beta T^{-1}-\chi-\eta T^{-0.5}} > 0$$

- I_{off} increases at higher temperatures
- The I_{on} to I_{off} ratio is significantly reduced with higher temperatures



Summary

- CMOS scaling trends
- Power dissipation in CMOS logic gates and circuits
- Dynamic power minimization techniques
- Effect of temperature on I_{on}/I_{off} ratio

- Next we shall consider leakage power minimization techniques

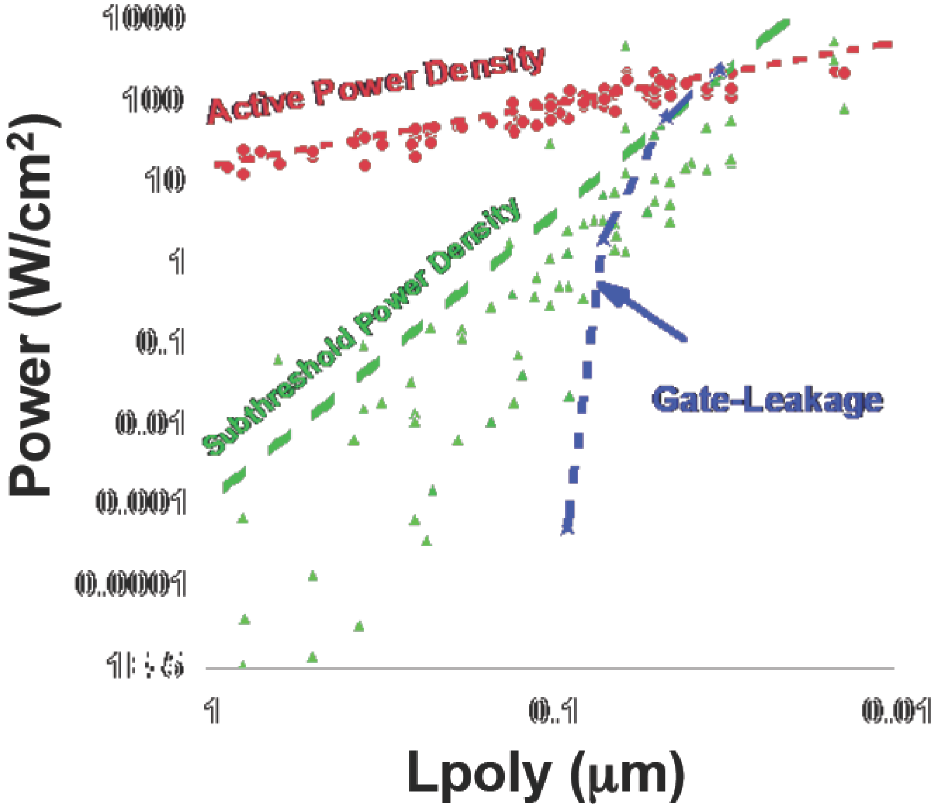
Minimizing Leakage Power in CMOS: Design Issues

Centre SI Summer School on
Nanoelectronic Circuits and Tools

Massoud Pedram
Dept. of Electrical Engineering
University of Southern California

July 15, 2008

Power Density Trends

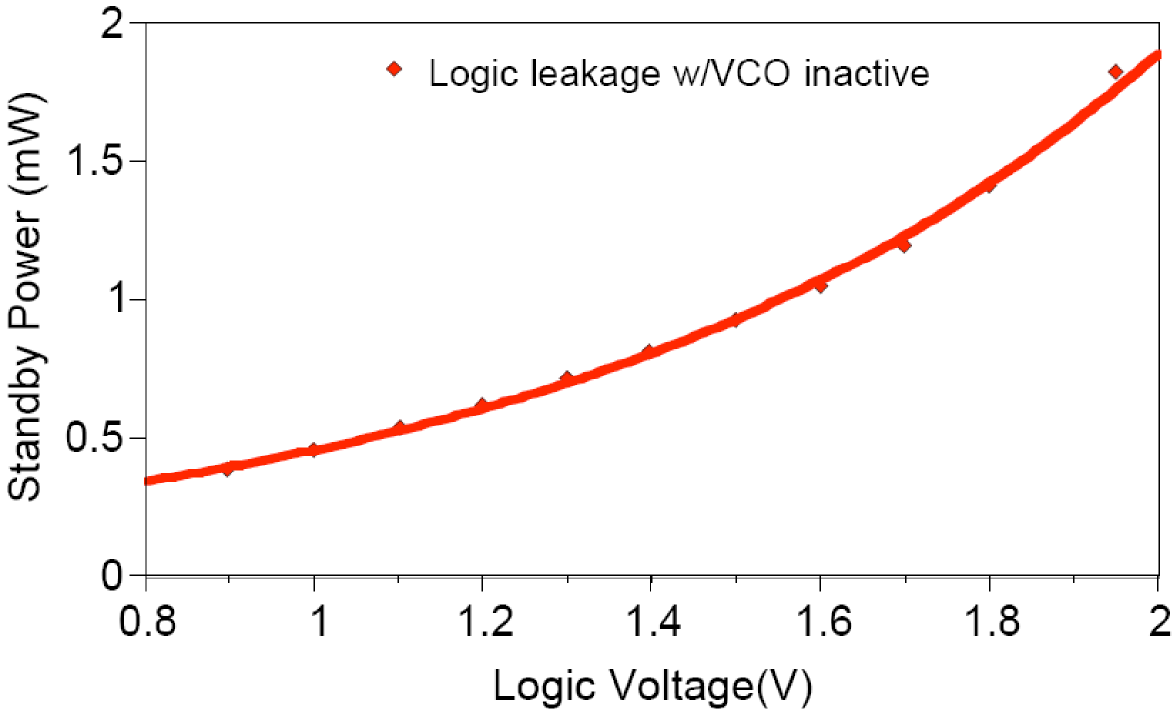


After Nowak, et al.

Leakage Power Minimization Techniques

- Lowering and/or turning off V_{DD}
- Gate length biasing (V_{th} roll-off effect)
- Transistor stacking
- Applying minimum leakage input vector in sleep mode
- Utilizing the dual-Vth devices (possibly combined with V_{th} roll-off effect)
  - Static approach: assigns low-V_{th} to timing-critical logic cells, high-V_{th} to other cells
  - Dynamic approach (a.k.a. power gating): requires a control signal (SLEEP signal) to turn off devices in the standby mode
- Body-biasing
  - Bias the body of NMOS (PMOS) device V_b < GND ($V_b > V_{DD}$) in sleep mode

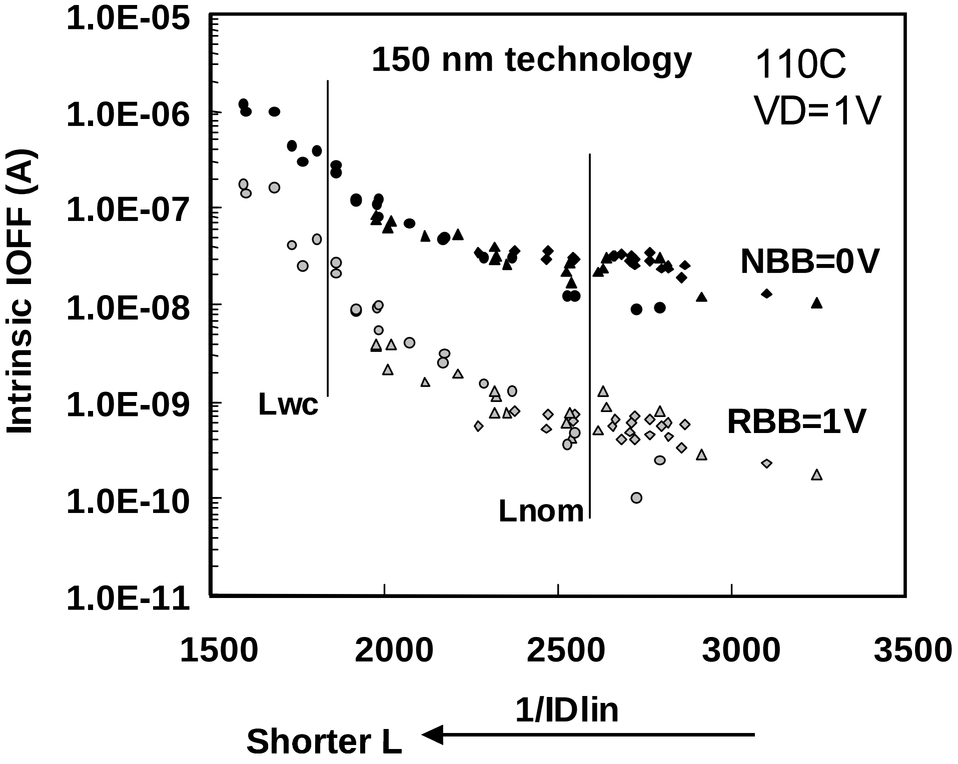
Effect of Supply Voltage Scaling



Subthreshold dominated technology

Source: Nowka, ISSCC-02

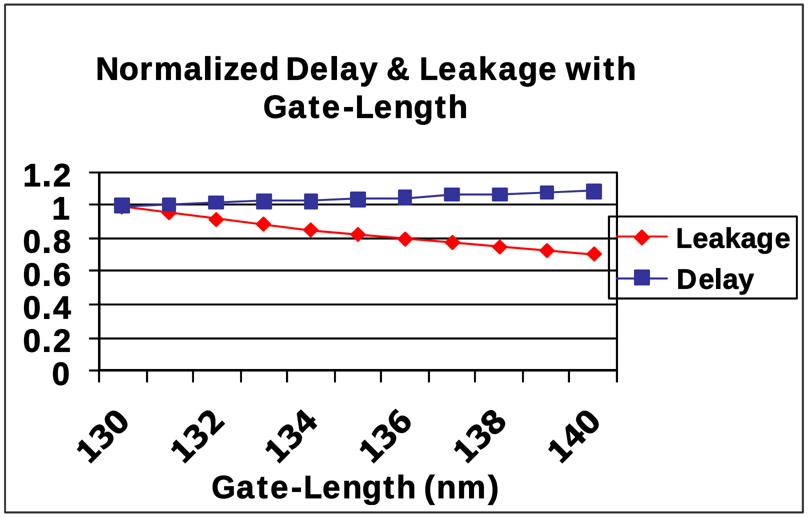
Impact of Gate Length Variation



NBB: No Body Biasing
RBB: Reverse Body Biasing

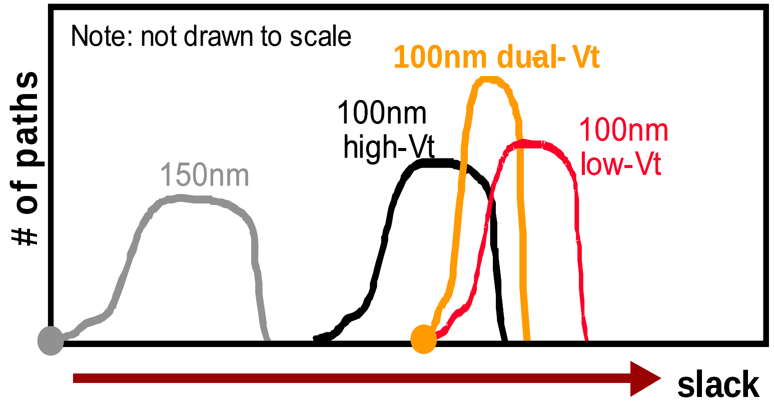
Source: De, 2004

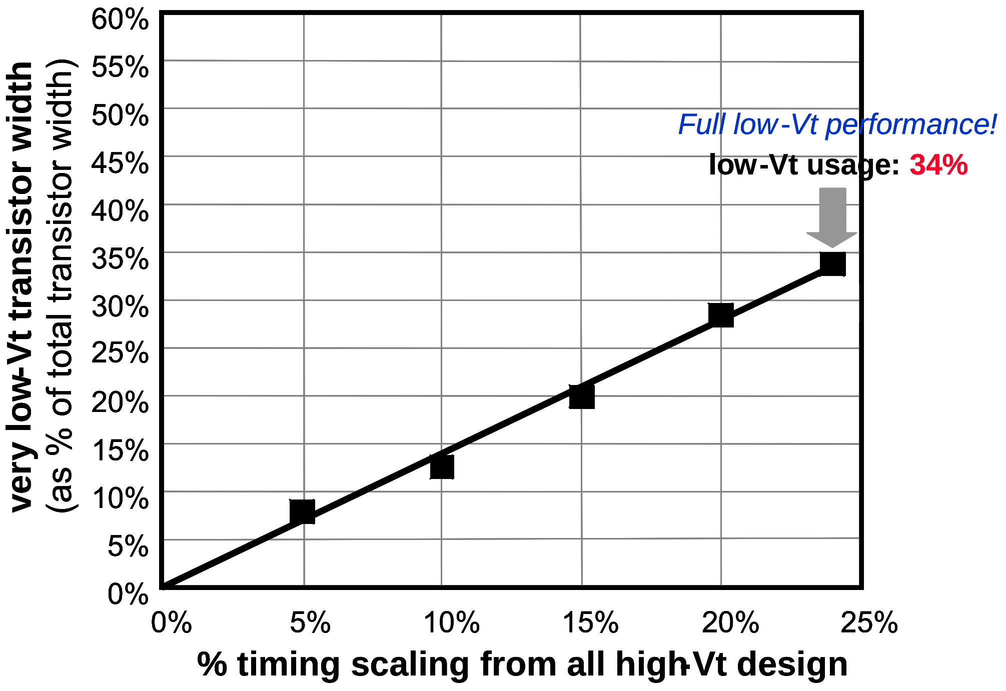
Gate Length Biasing



- Slightly increase (bias) the gate-length (line width) of devices
  - Slightly increases delay
  - Significantly reduces leakage
  - Bias only the non-critical devices
- Advantages:
  - Reduces runtime leakage and leakage variability
  - Can work in conjunction w/ V_{th} assignment → Gives finer control over delay-leakage tradeoff
  - Post-layout technique, no additional masks required
- 15-40% leakage and 30-60% leakage variability reduction for 90nm with dual-V_{th} assignment [Source: Gupta et al]

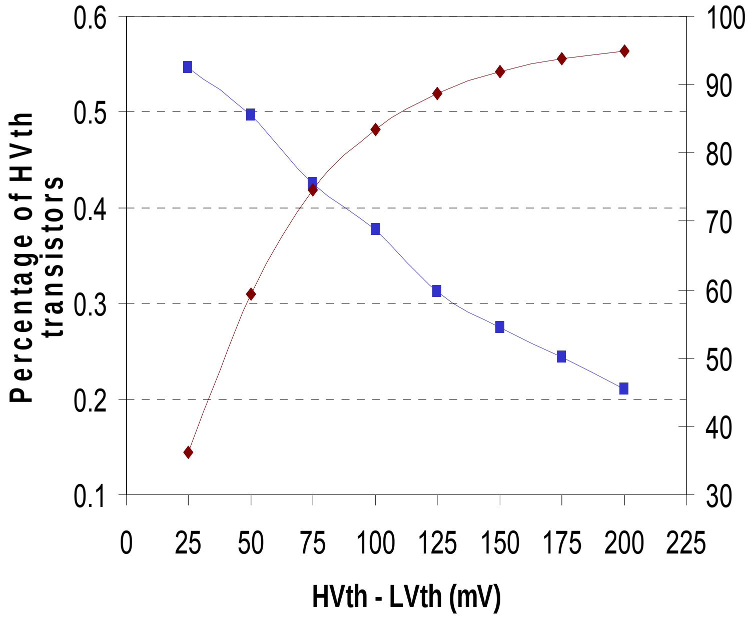
Dual-Vt Design for Leakage Control

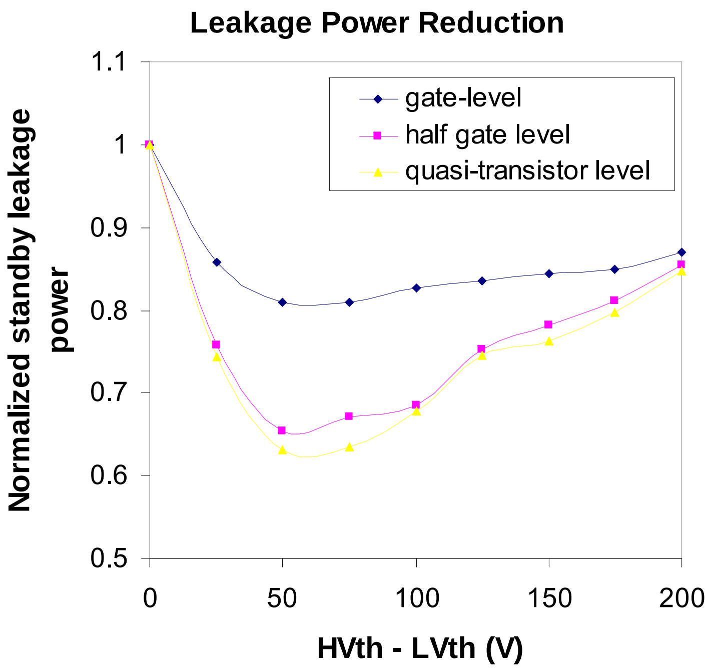




Source: De et al

Optimal Vt Choices



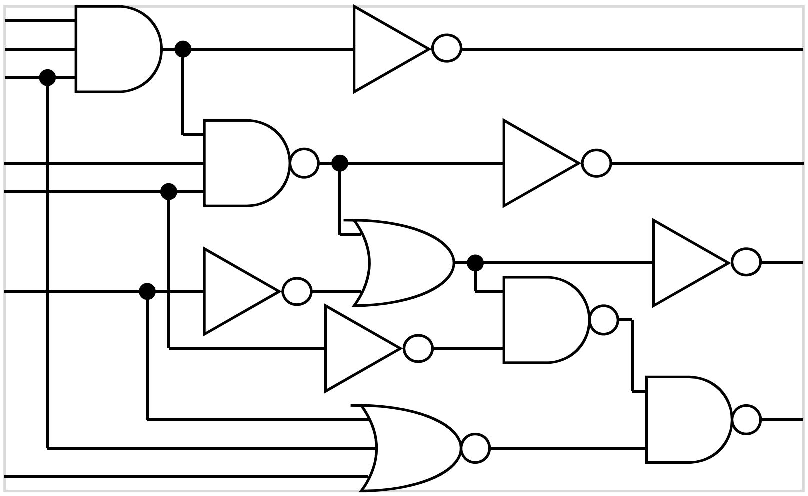


Example Dual Vt Optimization

- The following circuit is designed in 65nm CMOS technology using low threshold transistors. Each gate has a delay of 5ps and a leakage current of 10nA. Given that a gate with high threshold transistors has a delay of 12ps and leakage of 1nA, optimally design the circuit with dual-threshold gates to minimize the leakage current without increasing the critical path delay.

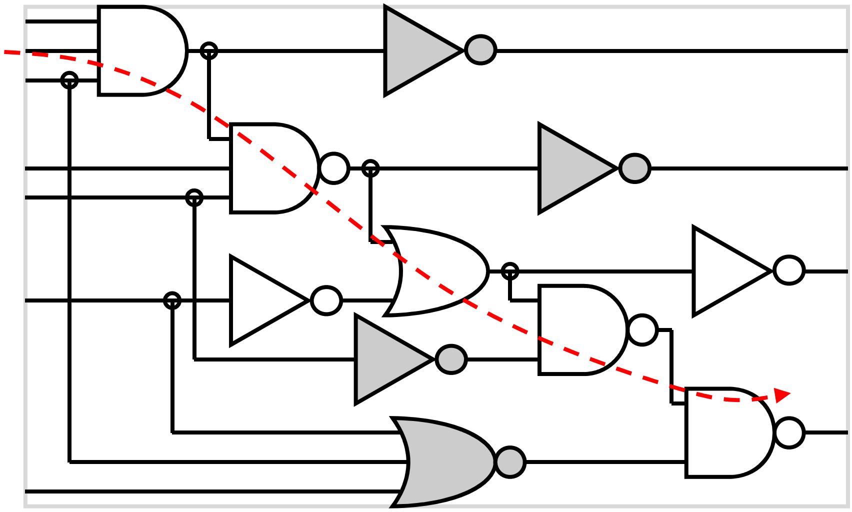
  (a) What is the percentage reduction in leakage power?

  (b) What will the leakage power reduction be if a 30% increase in the critical path delay is allowed?



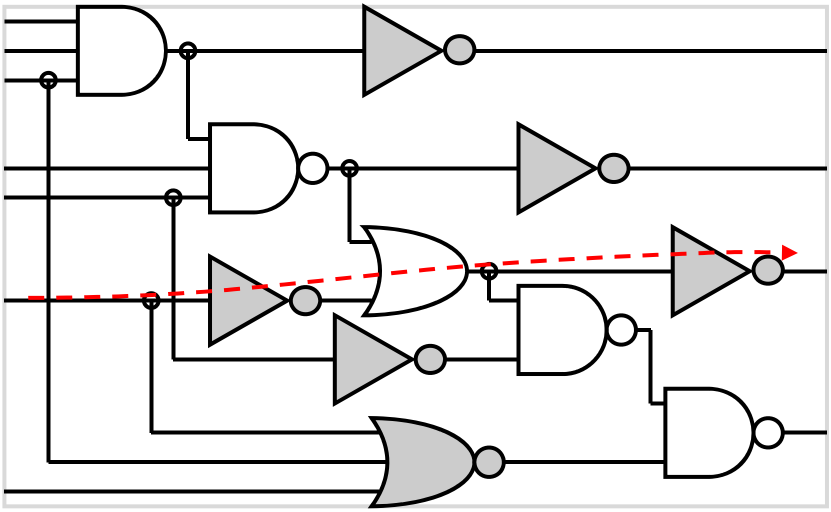
Dual Vt Example (Cont'd)

- Part (a): Three critical paths are from the first, second and third inputs to the last output, shown by a dashed line arrow. Each has five gates and a delay of 25ps. None of the five gates on the critical path (red arrow) can be assigned a high threshold. Also, the two inverters that are on four-gate long paths cannot be assigned high threshold because then the delay of those paths will become 27ps. The remaining three inverters and the NOR gate can be assigned high threshold. These gates are shaded grey in the circuit. The reduction in leakage power = 1 – (4×1+7×10)/(11×10) = 32.73%.
- **Critical path delay = 25ps**

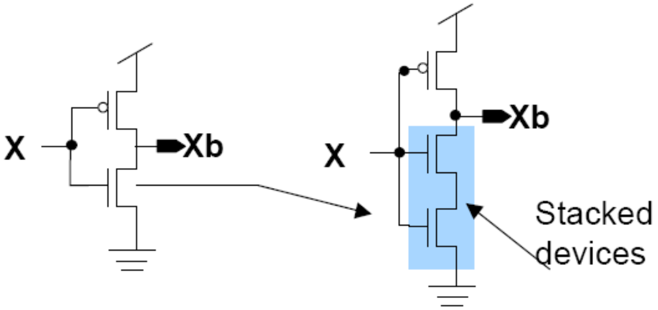


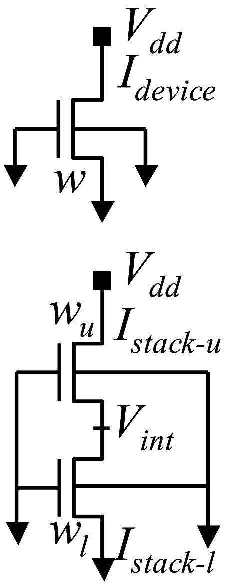
Dual Vt Example (Cont’d)

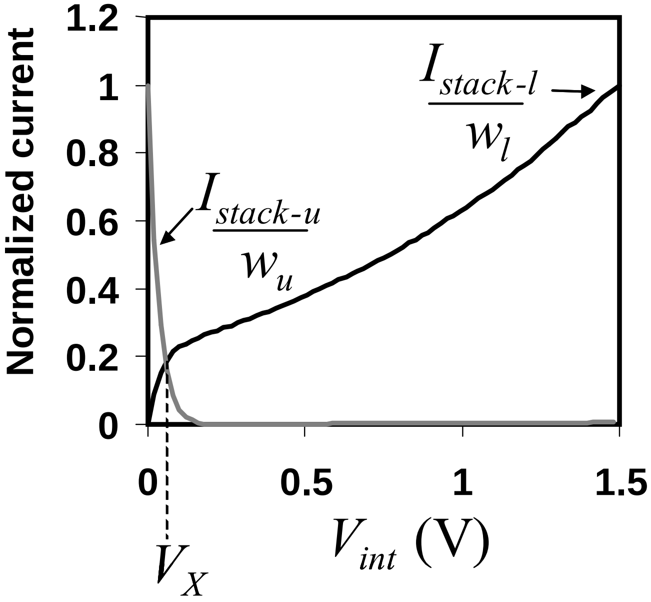
- Part (b): Several solutions are possible. Notice that any 3-gate path can have 2 high threshold gates. Four and five gate paths can have only one high threshold gate. One solution is shown in the figure below where six high threshold gates are shown with shading and the critical path is shown by a dashed red line arrow. The reduction in leakage power = 1 – (6×1+5×10)/(11×10) = 49.09%.
- **Critical path delay = 29ps**

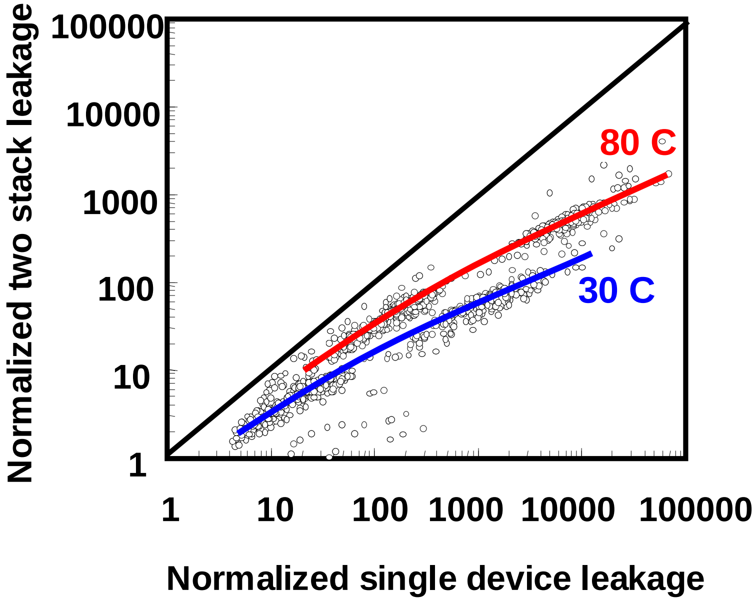


Leakage Current of Transistor Stacks



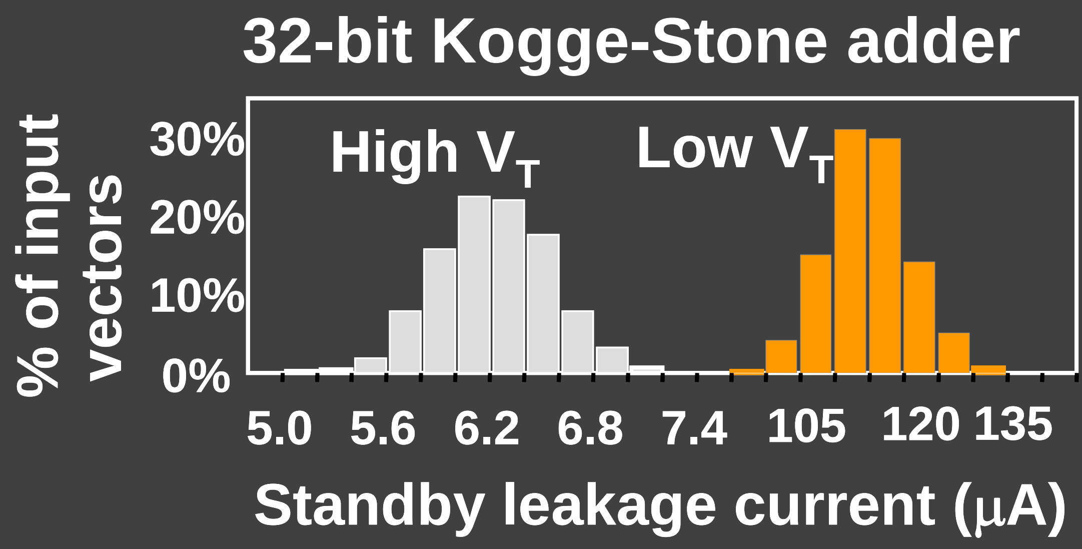






Source: De-2004

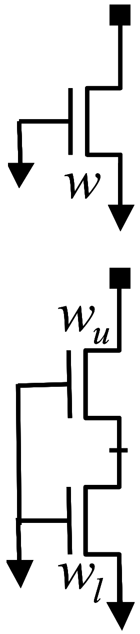
Exploiting Natural Stacks

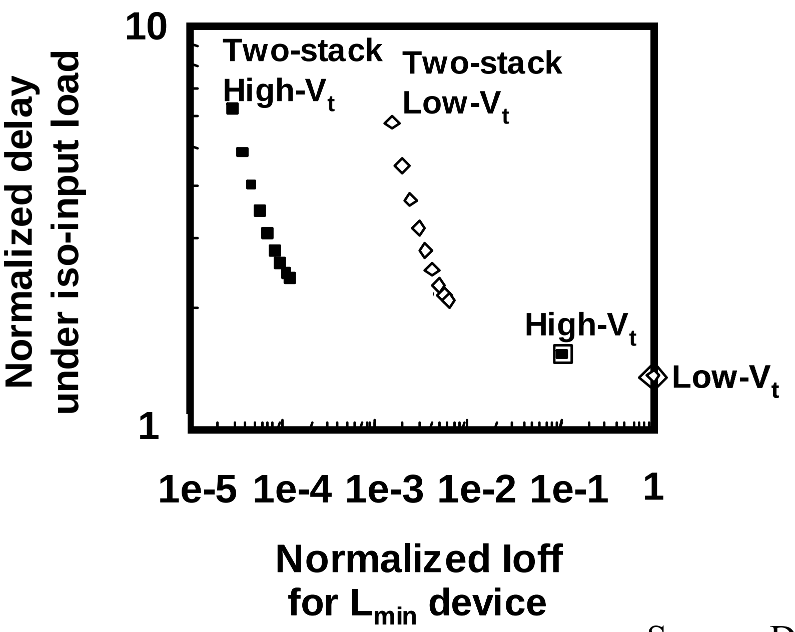


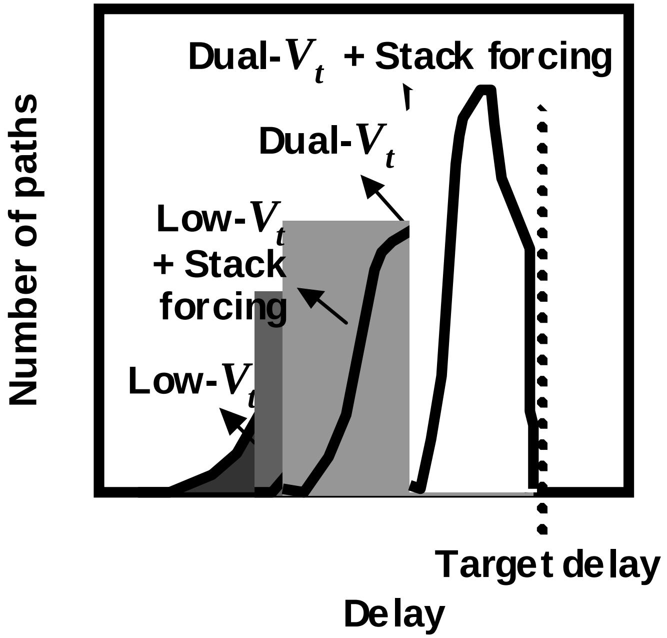
Reduction	Avg	Worst
High V_T	1.5X	2.5X
Low V_T	1.5X	2X

Source: De-2004

Stack Forcing for Leakage Control



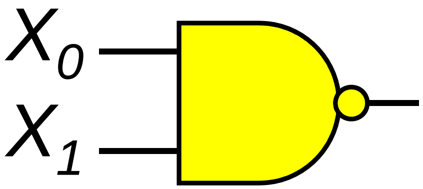




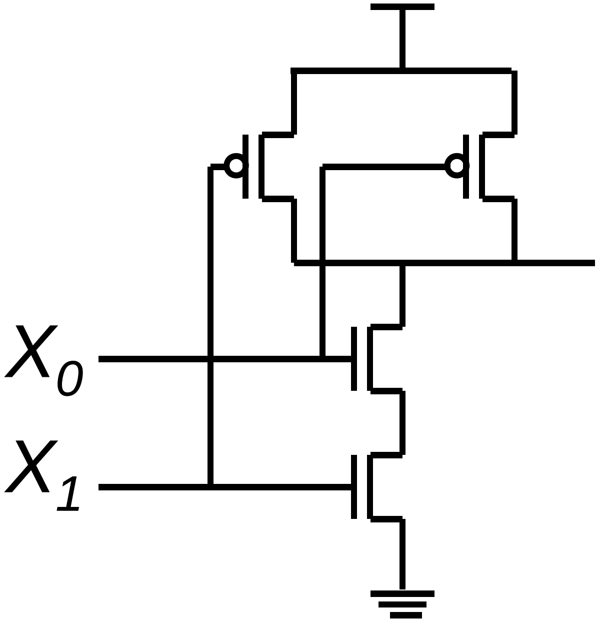
Source: De-2004

Low-Vt + stack-forcing reduces leakage power by 3X

Input Dependence of the Leakage Current

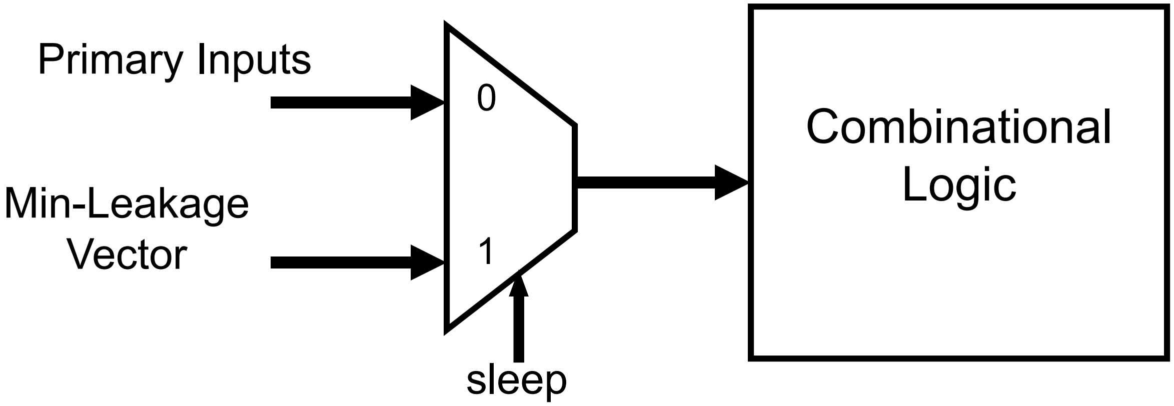


Technology: 0.18 μm
Supply Voltage = 1.5V
Threshold Voltage = 0.2V

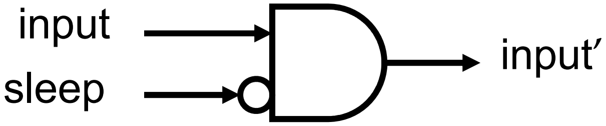


X_0	X_1	Leakage
0	0	23.60 nA
0	1	47.15 nA
1	0	51.42 nA
1	1	82.94 nA

Input Vector Control During Sleep Mode

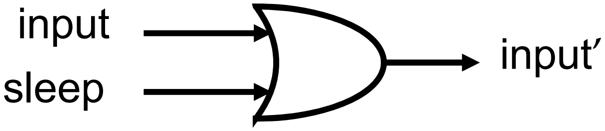


Min-Leakage Input = 0



sleep	input′
0	input
1	0

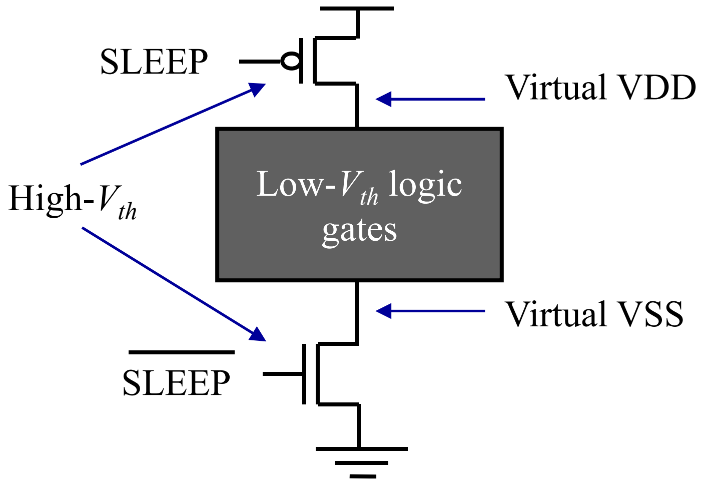
Min-Leakage Input = 1

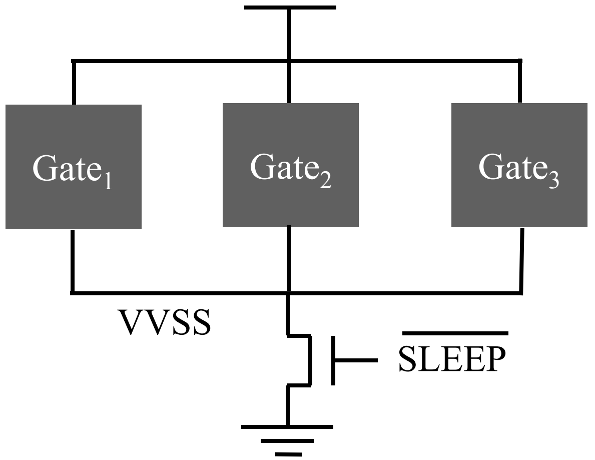


sleep	input′
0	input
1	1

Multi-Threshold CMOS

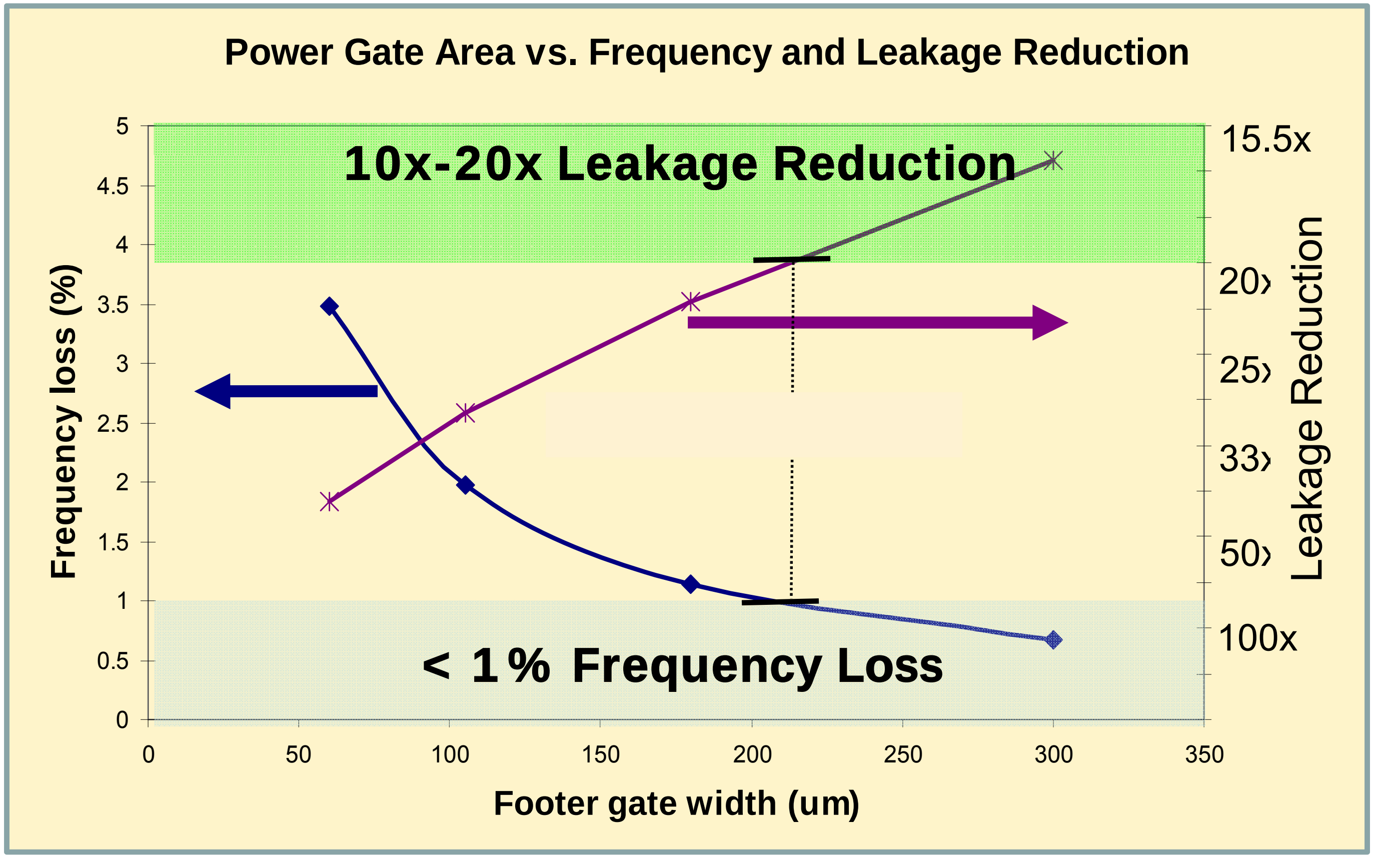
- High-V_{th} power switches are connected to low-V_{th} logic gates
  - Achieves high performance due to low-V_{th} logic gates
  - Reduces leakage power dramatically due to the series-connected high-V_{th} power switch
- Typically only a header or a footer sleep transistor is used, not both
- A single sleep transistor may be shared among several logic gates





Footer Gate Width Selection

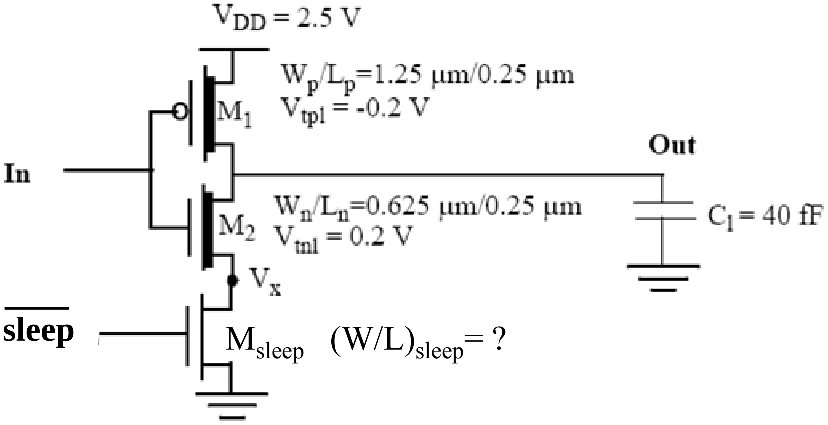
A 2-stage pipelined 40-bit ALU (IBM)



Sleep Transistor Sizing Example

- Consider the following MTCMOS inverter. Assuming zero effective resistance, $R_{DS,ON}$, for the sleep transistor, M_{sleep}, calculate, I_{active}, the peak value of the current that discharges the load capacitance in a high to low transition at the output?

  Solution: This is the value of the current through transistor M2 immediately after *In* switches from 0 to V_{DD}. M2 is in the saturation region:



$$I_{active} = I_{DS,M2} = 57.5\mu \times 2.5 \times (2.3)^2 = 760\mu A$$

Sleep Transistor Sizing

- Suppose that the maximum delay penalty of the MTCMOS circuit compared to the original CMOS inverter is 5%. Calculate the max value of V_x to ensure this timing requirement.

$$\frac{\tau_{d,SLEEP}}{\tau_d} = \frac{(V_{DD}-V_{tnl})^2}{(V_{DD}-V_X-V_{tnl})^2} = 1.05, \quad \frac{V_{DD}-V_{tnl}}{V_{DD}-V_X-V_{tnl}} = 1.025$$

$$V_X = \frac{0.025\times(V_{DD}-V_{tnl,0})}{1.025} = \frac{0.025\times 2.3}{1.025} = 0.0561V$$

$$V_{tnl}\left(V_{SB}=0.056\right) = V_{T0} + \gamma\left(\sqrt{\left|2\phi_F\right| + V_{SB}} - \sqrt{\left|2\phi_F\right|}\right) = 0.2 + 0.8\left(\sqrt{0.6+0.0561} - \sqrt{0.6}\right) = 0.228$$

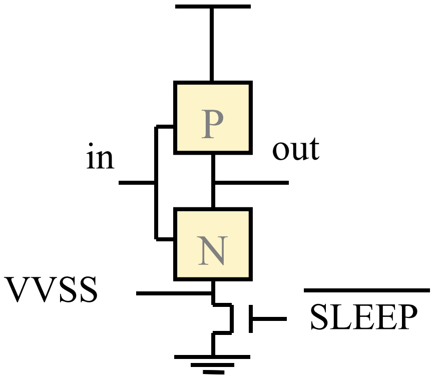
$$V_X = \frac{0.025\times(V_{DD}-V_{tnl,0})}{1.025} = \frac{0.025\times 2.272}{1.025} = 0.0554V$$

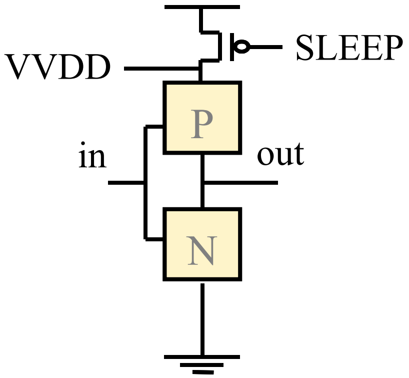
- Using I_{active} and V_x, find the minimum size of the sleep transistor, $(W/L)_{sleep}$. We write the current equation through the sleep transistor when its V_{DS} is equal to V_x obtained above and set this current equal to I_{active}.

$$I_{sleep} = 115\frac{\mu A}{V^2}\times\left(\frac{W}{L}\right)_{sleep}\times\left((V_{DD}-V_{th,sleep})V_x - \frac{V_x^2}{2}\right) = I_{active}$$

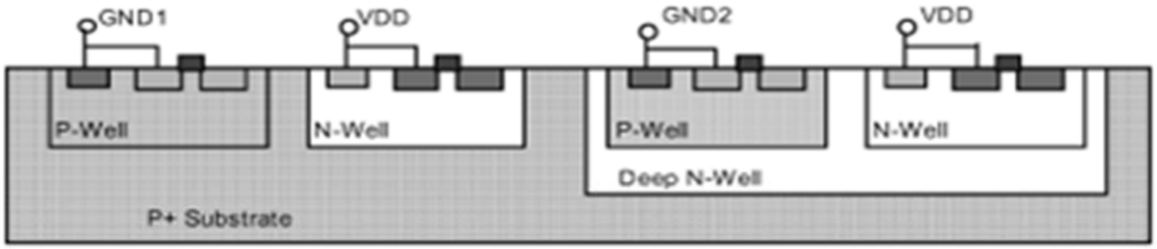
$$115\times\left(\frac{W}{L}\right)_{sleep}\times\left(2\times 0.056 - \frac{(0.056)^2}{2}\right) = 760, \quad \left(\frac{W}{L}\right)_{sleep} = 60$$

Header vs. Footer Switches





- Area and power dissipation overhead of NMOS footer transistors are lower due to higher mobility of electrons
- PMOS header transistors are more compatible with two-well bulk CMOS process where a high-performance NMOS transistor realized in the substrate is desirable

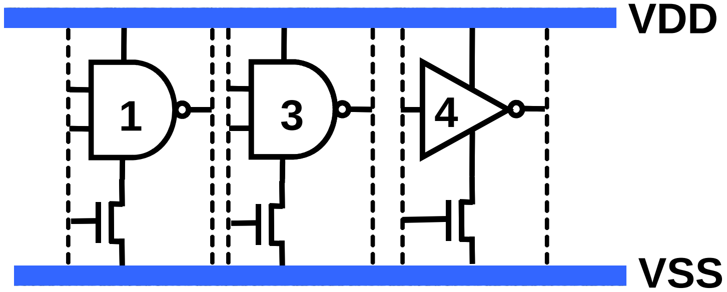


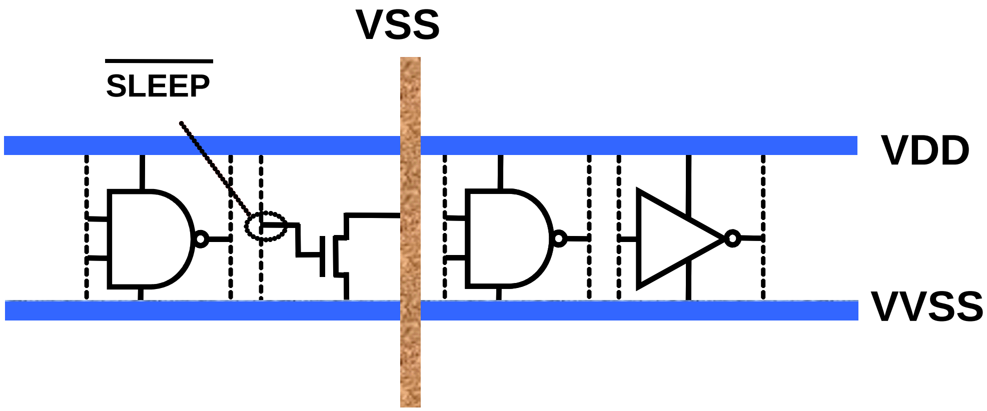
Triple well structure provides an isolating N layer between the local P-well and the P-substrate.

Coarse-Grain vs. Fine-grain MTCMOS

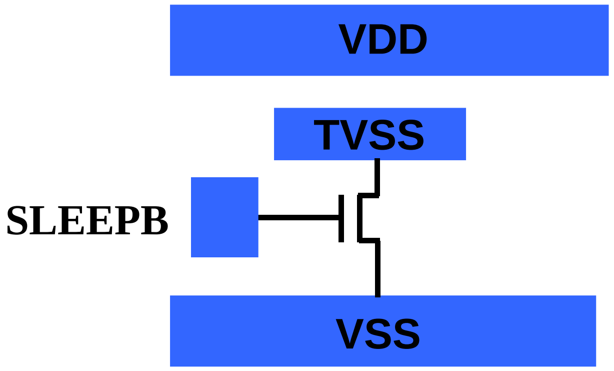
- Merits of fine-grain MTCMOS
  - Easier to incorporate into existing EDA flows and tools
  - Less parasitics on the virtual node and less EM problems

- Merits of coarse-grain MTCMOS
  - Smaller sleep transistor area and mode transition power overheads
  - Lower leakage due to smaller sleep transistor width

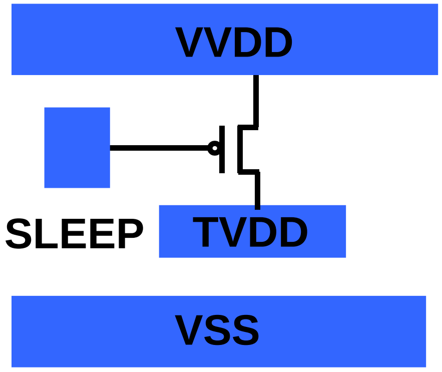




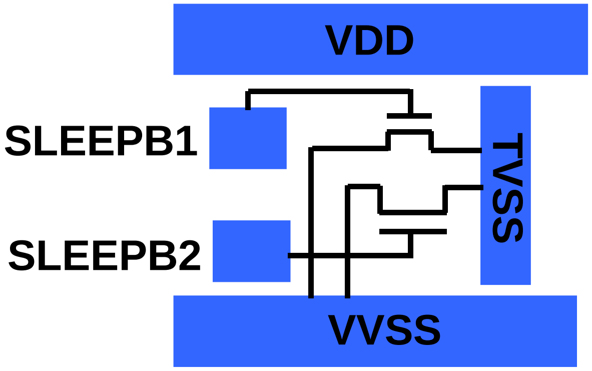
Sleep Transistor Layouts for Internal Switches



Single transistor footer switch

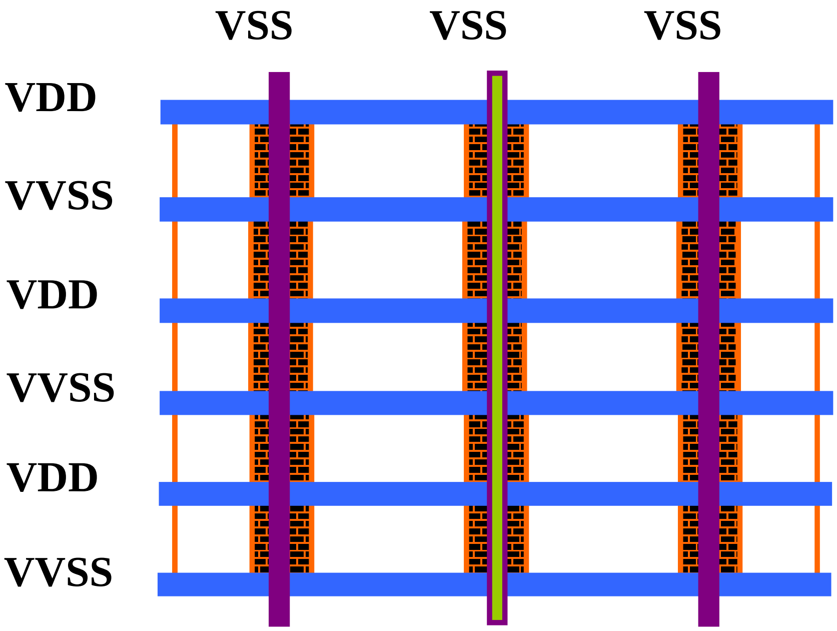


Single transistor header switch

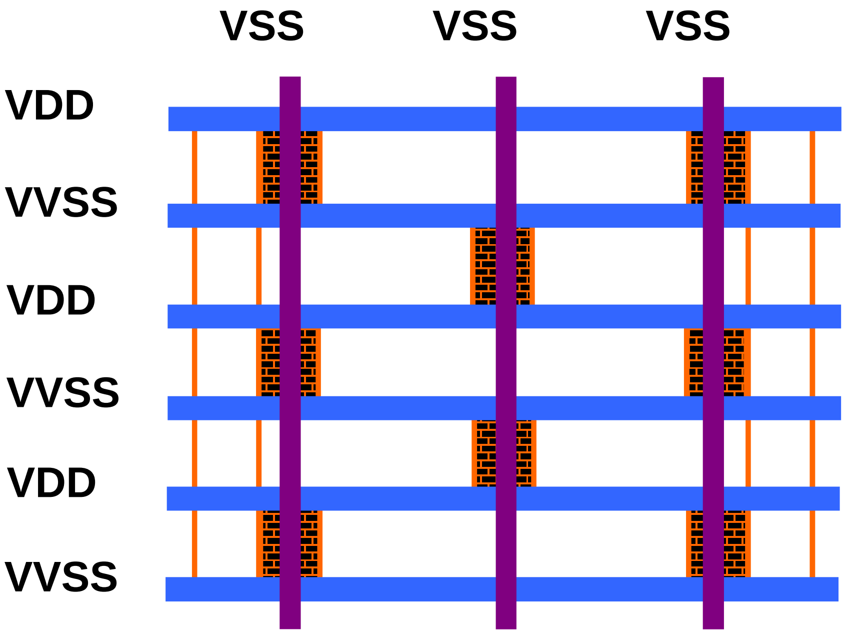


Double-transistor
(mother/daughter) footer switch

Layout Styles for Internal Footer Switch



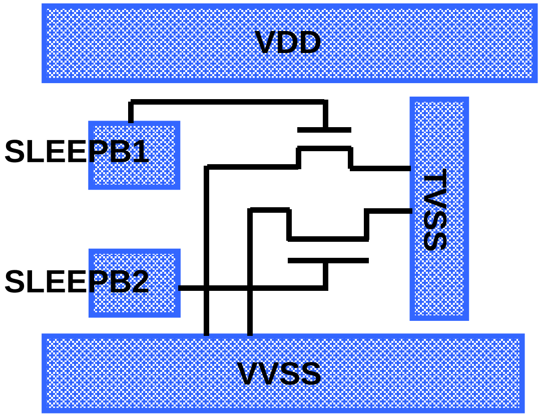
Column-based



Staggered

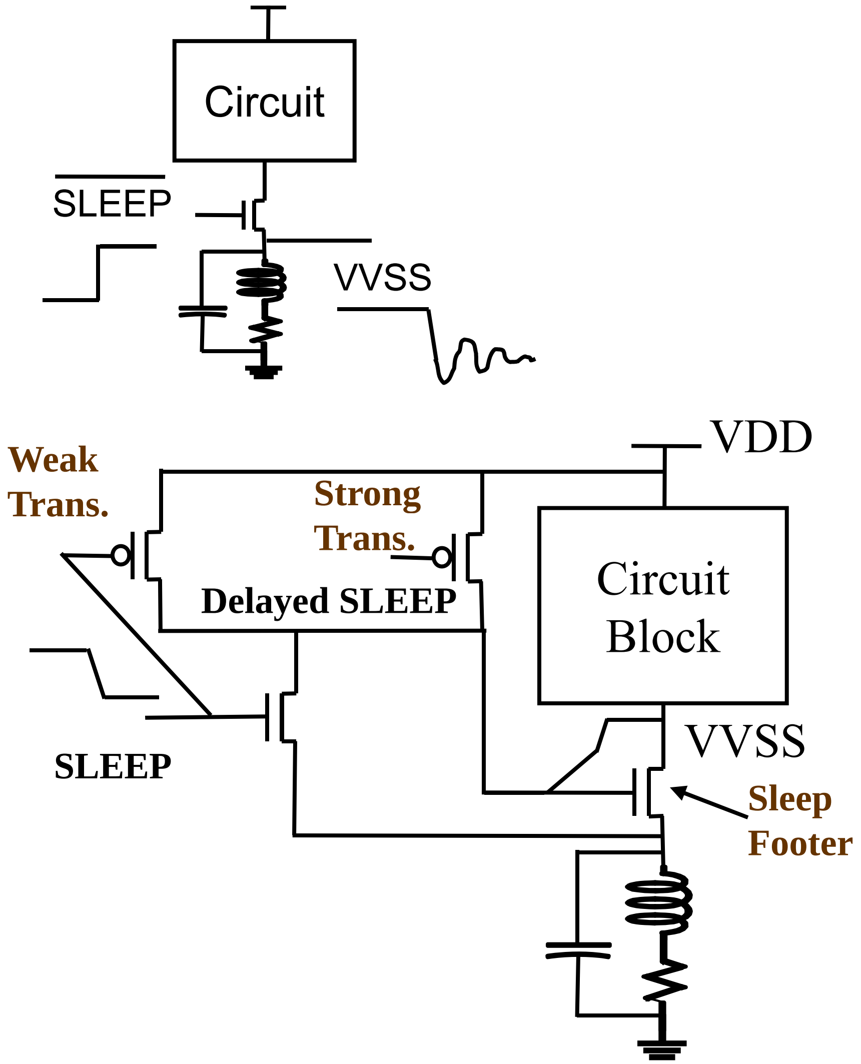
Sleep Signal Scheduling for Two Parallel Sleep Transistors

- There is a large current rush in sleep to active transition which can cause EM and IR-drop issues
  - Rush current can be reduced by using parallel (mother-daughter) sleep transistors
  - Total sleep transistor width, the summation of the mother and daughter switch widths, is determined by using a sizing algorithm
- The peak rush current and IR drop across the switch can be controlled by optimizing the ratio of the daughter and mother transistor widths and scheduling the turn-on times of the two switches so as to minimize the wakeup delay



Staircase Sleep Scheduling

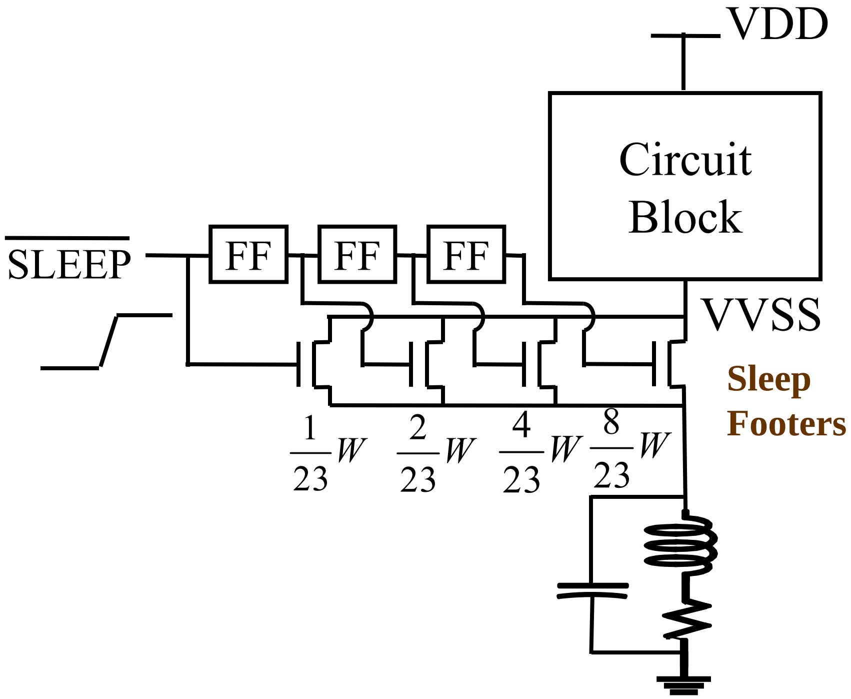
- Reduce ground bounce by turning on the sleep transistor in two steps:
  - First use a weak PMOS: $V_{gs} < V_{dd}$ for the sleep transistor. Originally, V_{ds} is high. So, the peak current is controlled
  - Next use a strong PMOS: $V_{gs} = V_{dd}$ for the sleep transistor. V_{ds} is however low. Therefore, the peak current is reduced



Parallel Sleep Transistors

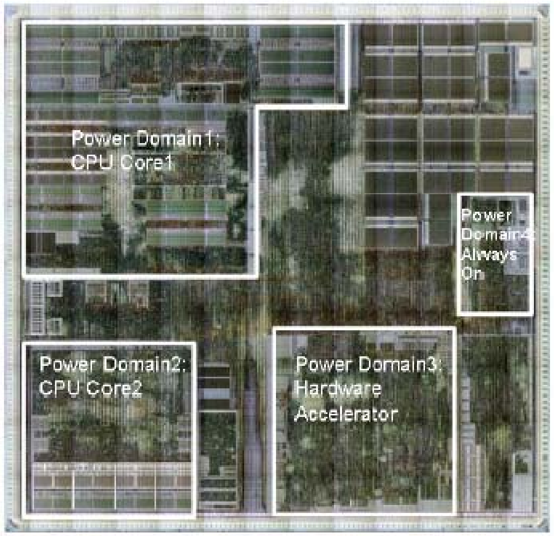
- Alternatively, one may use several sleep transistors

  Successively turn them on with cycle delays
- The resistance between the virtual ground and the ground is reduced as the V_{ds} of the sleep transistor is lowered. This reduces the peak current

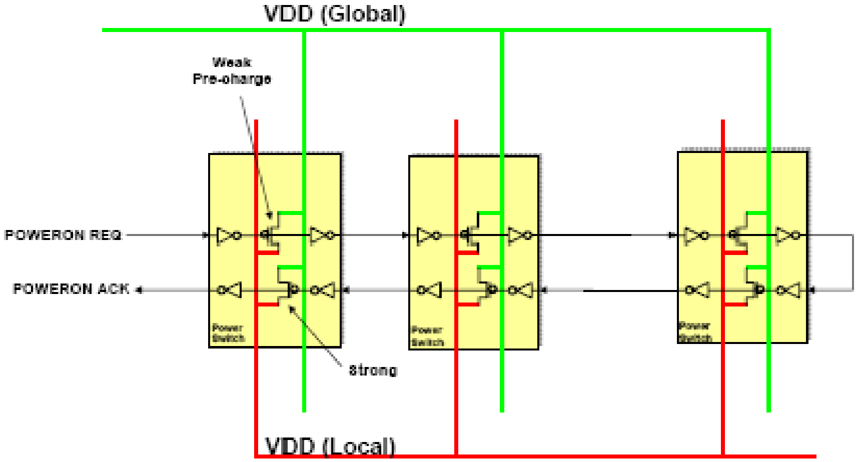


Sleep Footers

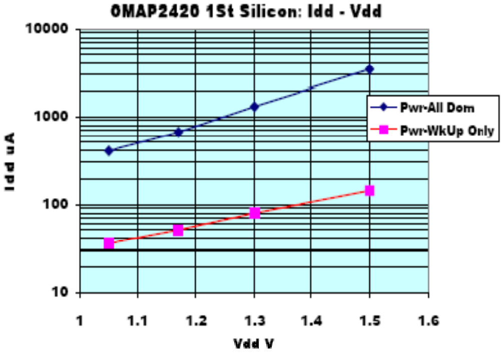
Power Gating Example



90nm OMAP2420 SoC



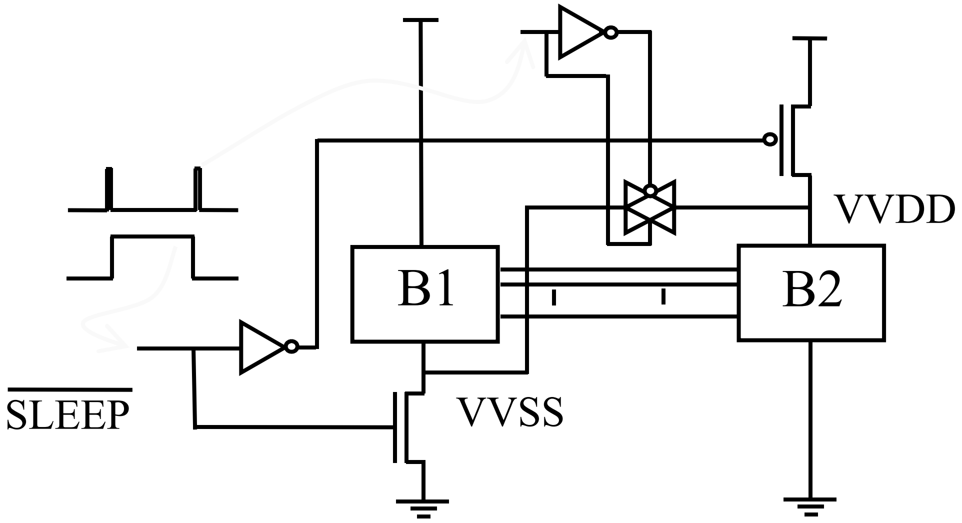
Power Switch used in OMAP



- Five power domains in OMAP SoC enabled by power gating
- Power switches gate V_{DD}, consists of
  - Weak PMOS: Sinks low current for power restoration
  - Strong PMOS: Deliver current for normal operation
- 2-pass power turn-on mechanism to prevent current surges
  - Weak switches turned on first to almost fully restore V_{DD} (local), and then the strong switches are turned on to support normal operation

Mode Transition Energy Minimization

- Energy consumption needed for mode transitions can be significant for power-gated circuits
  - A charge-recycling technique can be used to minimize the power consumption during the mode transition in an MTCMOS circuit while maintaining, or sometimes even improving, the wake up time
  - The charge recycling switch cell is turned on right before going from sleep to active and right after going from active to sleep

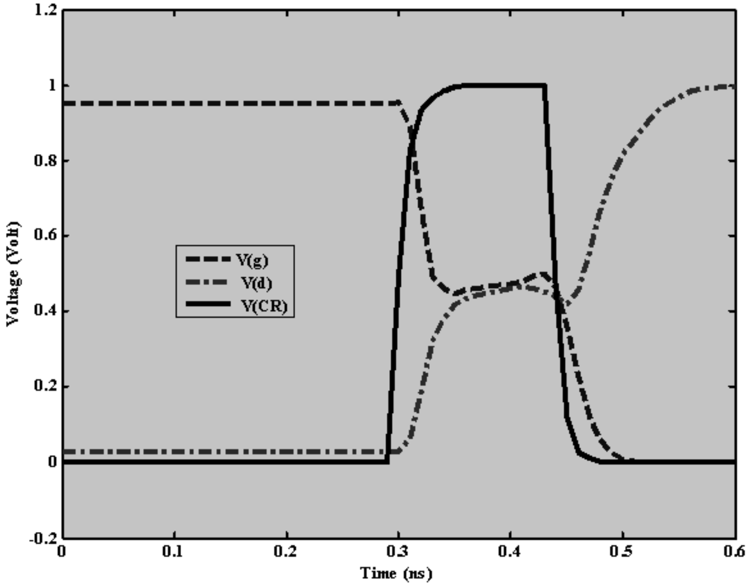


Mode Transition Energy (cont'd)

- During the sleep mode, voltage values for VVSS and VVDD reach V_{DD} and 0, respectively
- The circuit is put to a half-wakeup state by turning the charge recycling circuitry on at the sleep-to-active transition edge and right before turning on the sleep transistors
- After charge recycling is complete, the charge recycling circuitry is turned off and the sleep transistors are turned on to completely wake up the circuit. A similar strategy is used during the active-sleep transition
- Total energy saving is:

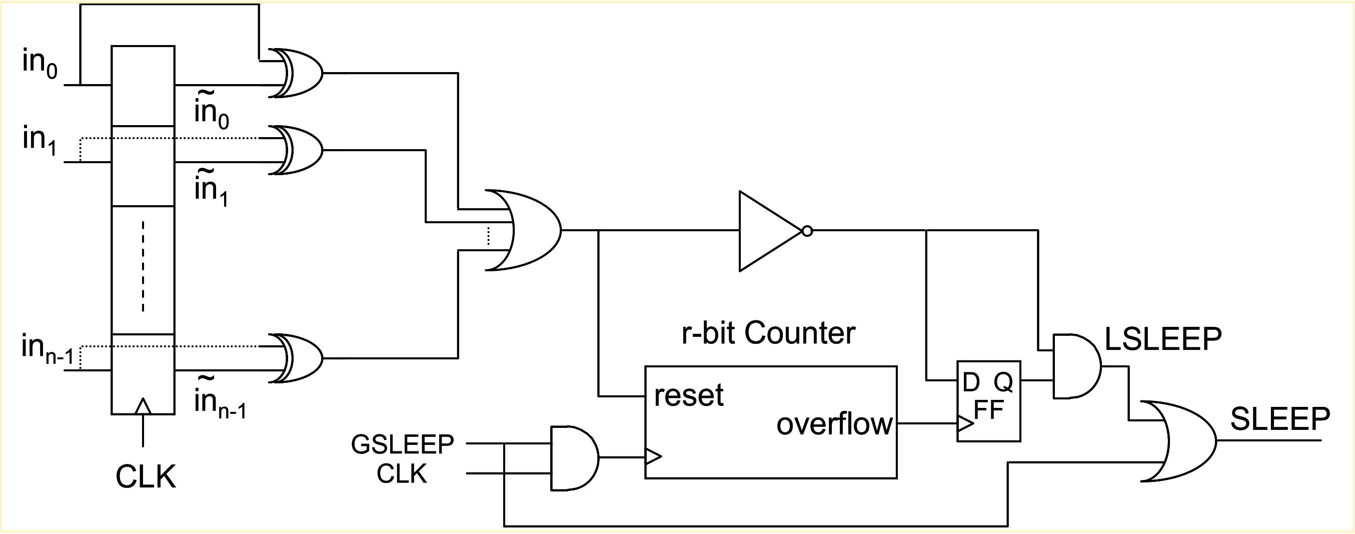
$$ESR = \frac{2C_{VVSS}C_{VVDD}}{\left(C_{VVSS} + C_{VVDD}\right)^2}$$

- The maximum ESR of 50% is obtained when $C_{VVSS} = C_{VVDD}$



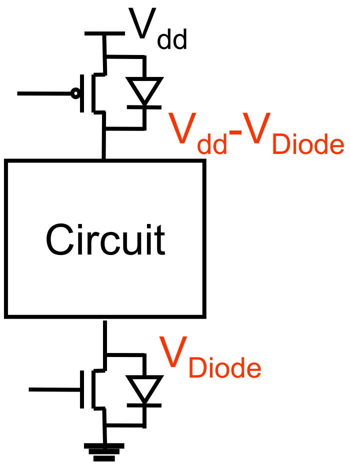
Local Sleep Signal Generation for Autonomous Power Gating

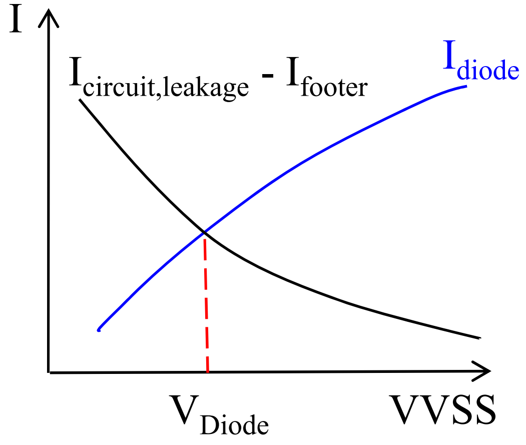
- A local sleep signal for each block can be generated which can automatically put the block into sleep independent of the global sleep signal
- A small circuitry may be added to compare the current input signals to the block with the input signals of the previous clock, and generate a local sleep signal when there is no change in the input signals for a given number of the clock cycles



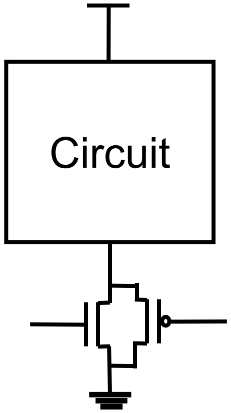
Voltage Rail Clamp (VRC) and Park Mode

- Voltage Rail Clamp: Reduce the virtual supply and ground voltages using two diodes
  - This allows state retention
  - It reduces noise during transition to active mode
  - However, the leakage saving in sleep mode is reduced



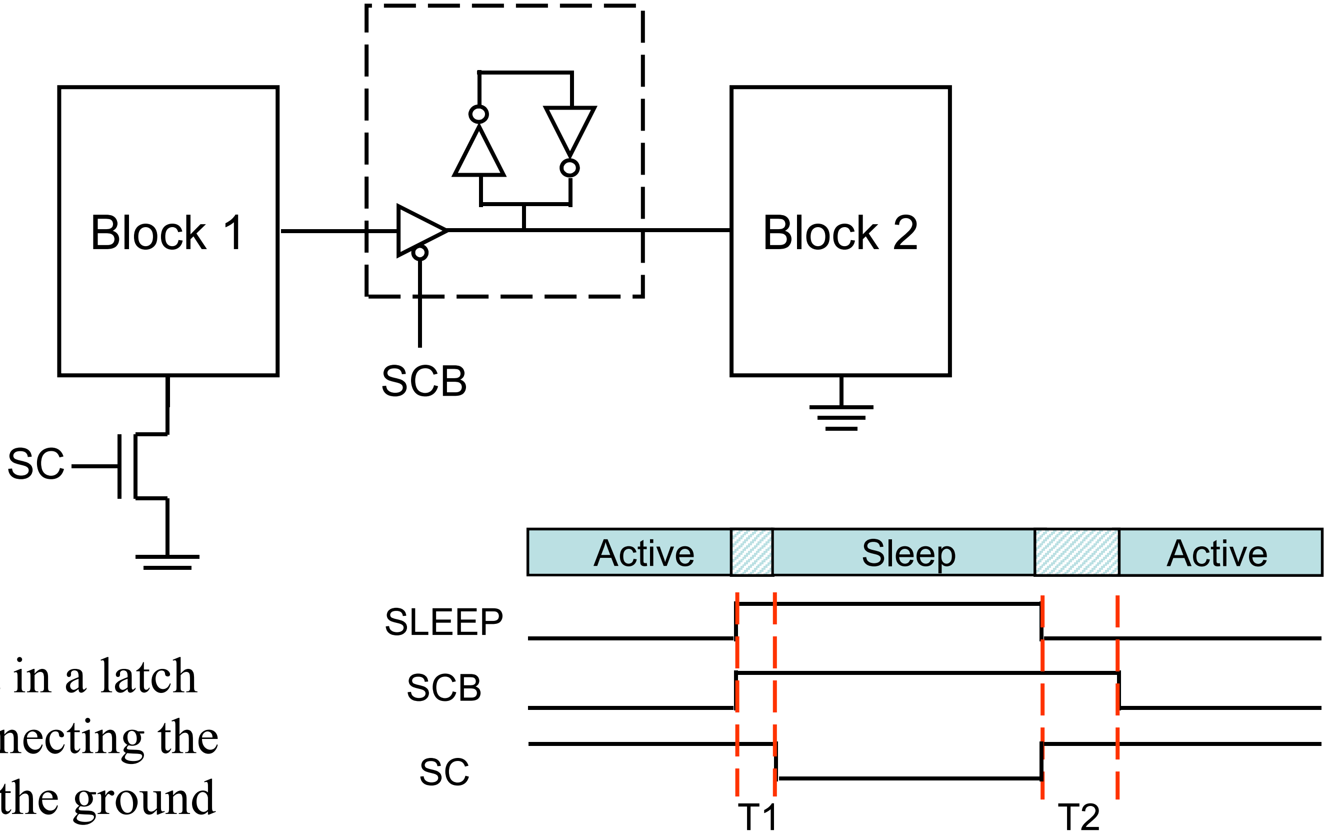


- Park Mode: Use a normally-on PMOS transistor to clamp virtual ground
  - Reduces leakage and bounce noise during wakeup
  - Keeps the internal state
  - Can turn off the PMOS transistor when in the sleep mode to achieve higher leakage saving. However, internal state will be lost



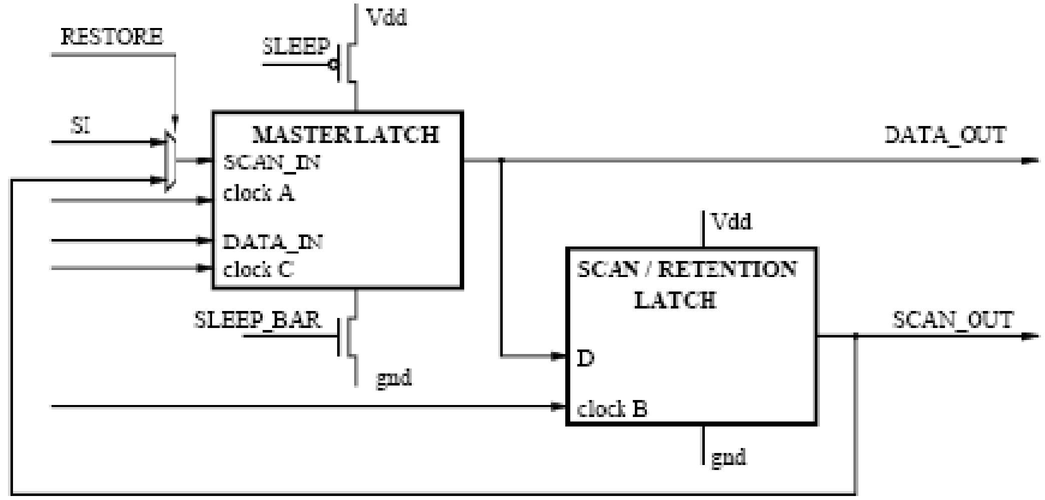
MTCMOS Fencing

Floating Prevention Circuit (FPC)



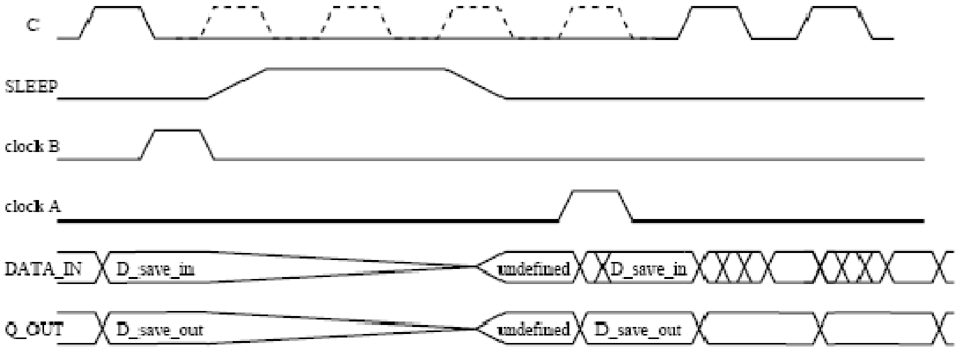
- Store the data in a latch before disconnecting the module from the ground

State Retention



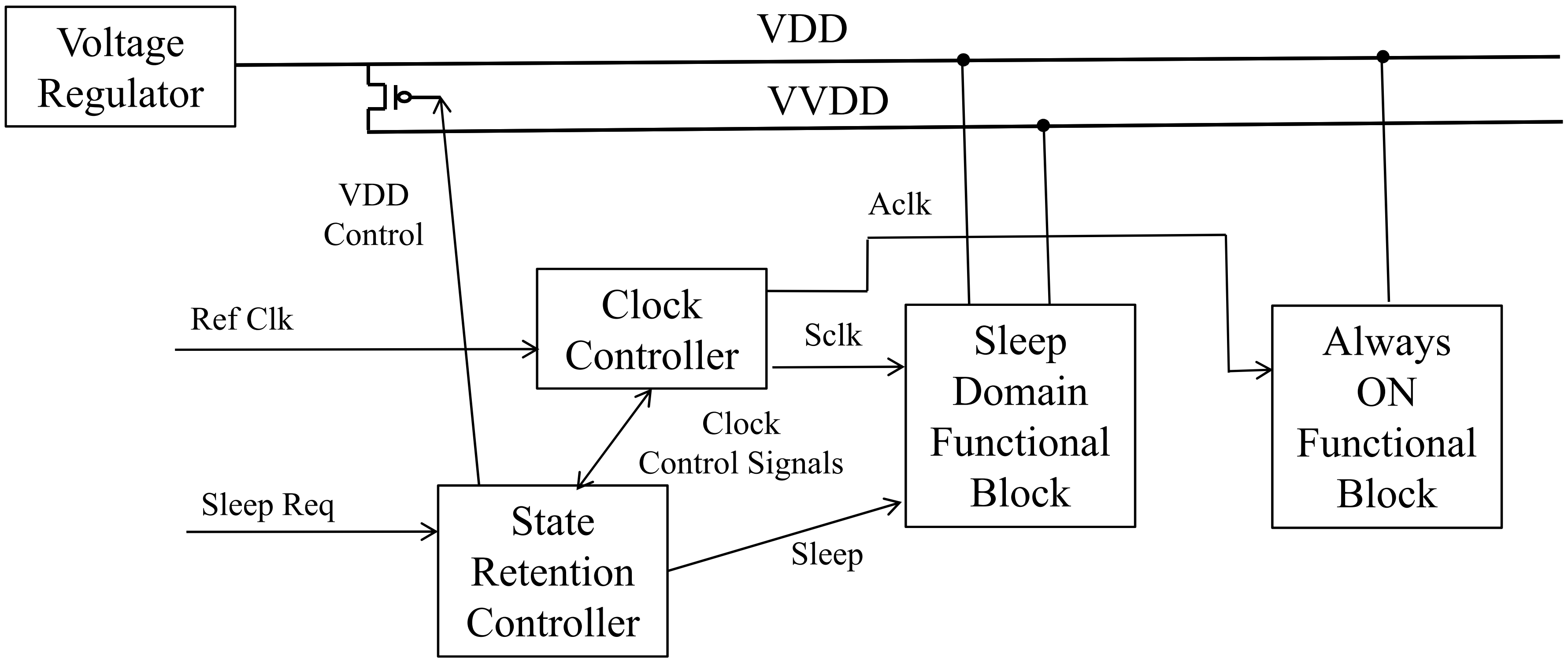
Integrated Scan Retention

Source: Zyuban-ISLPED02



Save and Restore Operations

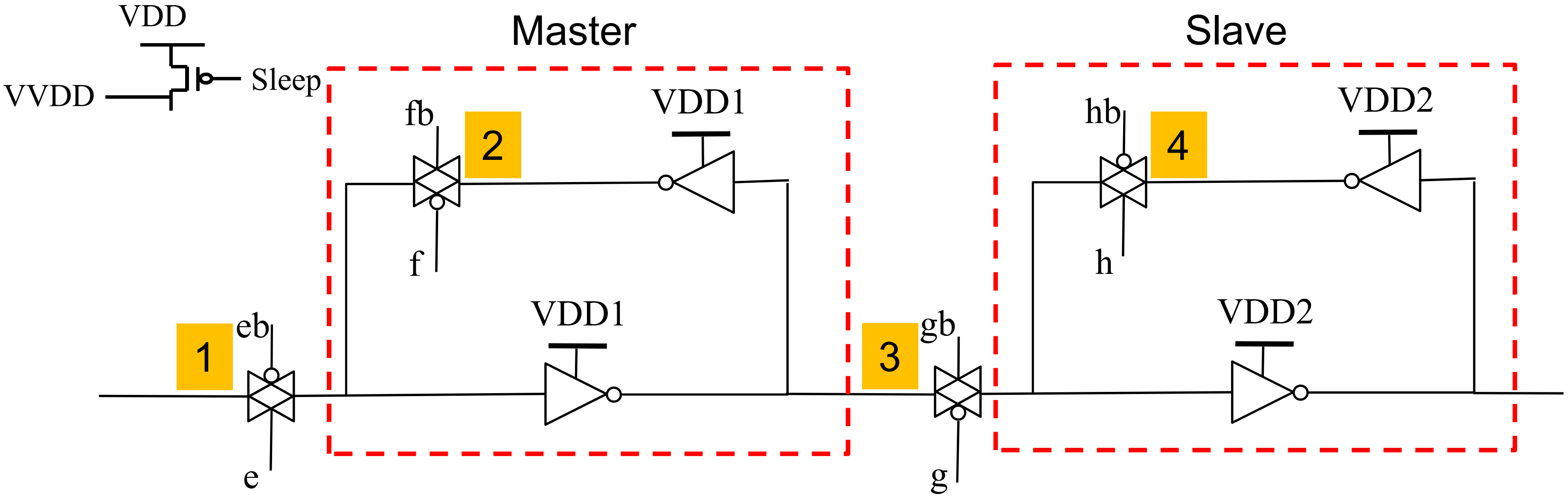
Data Processor with Power Gating Support



Sleep Req=1→ Stop Clk=1 →Sclk disabled→Sleep=1 →VDD Control=1→VVDD floats

State-Retentive Master-Slave Flip-Flop

(Freescale)



Note that at the point that the sleep domain clock (Sclk) stops, the slave (master) portion of the pos (neg)- edge FF contains the state information to be retained

Pos-Edge FF (w.r.t. Sclk)

Sleep	Sclk	Switch 1	Switch 2	Switch 3	Switch 4
0	0	ON	OFF	OFF	ON
0	1	OFF	ON	ON	OFF
1	x	ON	OFF	OFF	ON

Neg-Edge FF (w.r.t. Sclk)

Sleep	Sclk	Switch 1	Switch 2	Switch 3	Switch 4
0	0	OFF	ON	ON	OFF
0	1	ON	OFF	OFF	ON
1	x	OFF	ON	ON	OFF

	VDD1	VDD2
Pos-Edge FF	VVDD	VDD
Neg-Edge FF	VDD	VVDD

Power Management Example

- Consider a logic circuit that can be in one of two modes of operation: *Busy state* where it is doing useful work and *Idle state* where it is sitting idle waiting for workload to arrive. The circuit can be placed in one of two states: *Active mode* where the full V_{DD} level is applied to the circuit and a *Sleep mode* where the circuit is power gated using a high V_T sleep transistor. We assume the power dissipations in active and sleep modes are 1mW and 50μW, respectively.
- A power management controller counts the number of cycles that the circuit has been idle, and after 100 idle cycles, it will generate a control signal to a power-gating sleep transistor in order to transition the circuit into the sleep mode. The transitions into and out of the sleep mode take 10 cycles each, during which the circuit consumes ¼ of the active power dissipation on average. For each of the two mode transitions, the energy dissipation by the driver of the sleep transistor is 1nJ.
- We calculate the minimum duration of the *sleep mode* (in number of cycles) for the transition from the active to sleep mode and back to active mode to result in energy saving compared to the case that the circuit is never put to sleep and stays in the active mode all the time (clock frequency is100MHz).

Example (Cont'd)

- **Solution:**

  Each clock cycle=10ns

  Number of cycles in the transition from Active to sleep mode and back to active mode = 100+10×2=120

  Number of cycles in the sleep mode = x

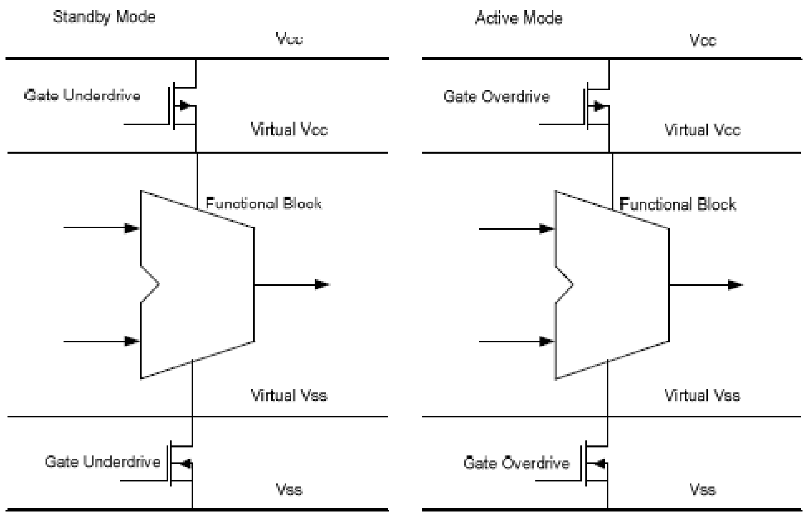
  Energy dissipation if the circuit was always in the active mode = $(120+x)\times 10\text{ns}\times 1\text{mW}=(1.2+0.01\,x)\times 10^{-9}\text{nJ}$

  Energy dissipation if the circuit is put to sleep and awakened = $100\times 10\text{ns}\times 1\text{mW}+20\times 10\text{ns}\times 0.25\text{mW}+1\times 2\text{nJ}+x\times 10\text{ns}\times 0.050\text{mW}=(3.05+0.0005x)\times 10^{-9}\text{ nJ} \rightarrow 1.2+0.01x=3.05+0.0005x \rightarrow 0.0095x=1.85 \rightarrow x=194.7$

  The minimum number of cycles to be in the sleep mode=195.

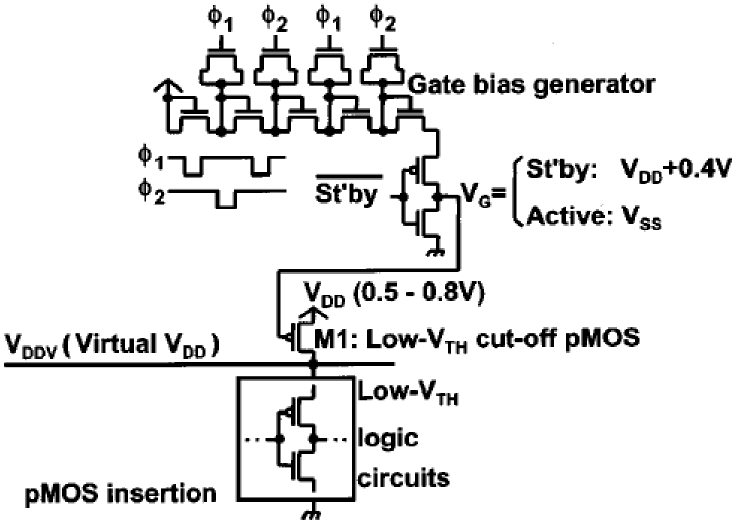
Supper Cut-Off CMOS (SCCMOS)

- MTCMOS uses high V_{th} as a cut-off MOSFET in series with low-V_{th} logic circuits to cut-off leakage current in standby mode
- MTCMOS does not work below 0.6V supply voltage because the high-V_{th} MOSFET does not turn on
  - MTCMOS cannot be used in sub-1-V V_{DD}
- Super cut-off CMOS has been proposed to solve this problem
  - The cut-off device in SCCMOS is low-V_{th} MOSFET, thus no need for high-V_{th} MOSFET
  - The low-V_{th} assures high speed operation

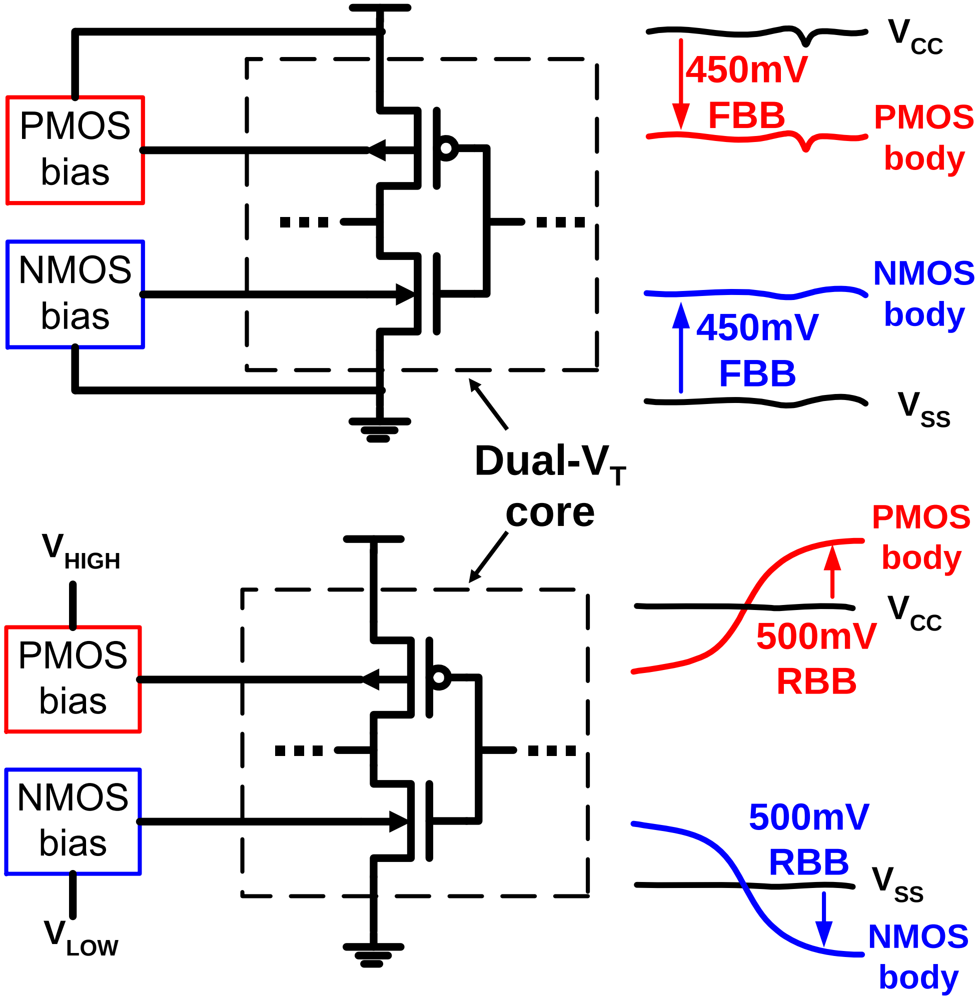


SCCMOS

- Instead of increasing V_{th}, SCCMOS increase the $|V_{GS}|$ value in the off region of the cut-off device
  - SCCMOS with a pMOS insertion case is shown below
- The low-V_{th} cut-off pMOS, M1, is inserted in series to the logic circuit consisting of low-V_{th} MOSFETs
- The gate voltage of M1, V_G, is grounded in active mode
- When the logic circuits enter standby operation, V+ is overdriven to V_{DD}+0.4 V to completely cut off the leakage current
- This is because the low-V_{th} of 0.1–0.2 V is lower by 0.4 V than conventional high-V_{th} (0.5–0.6 V), and thus this overdriven mechanism can sustain the standby current level



Dynamic Body Biasing for Active Mode Leakage Control



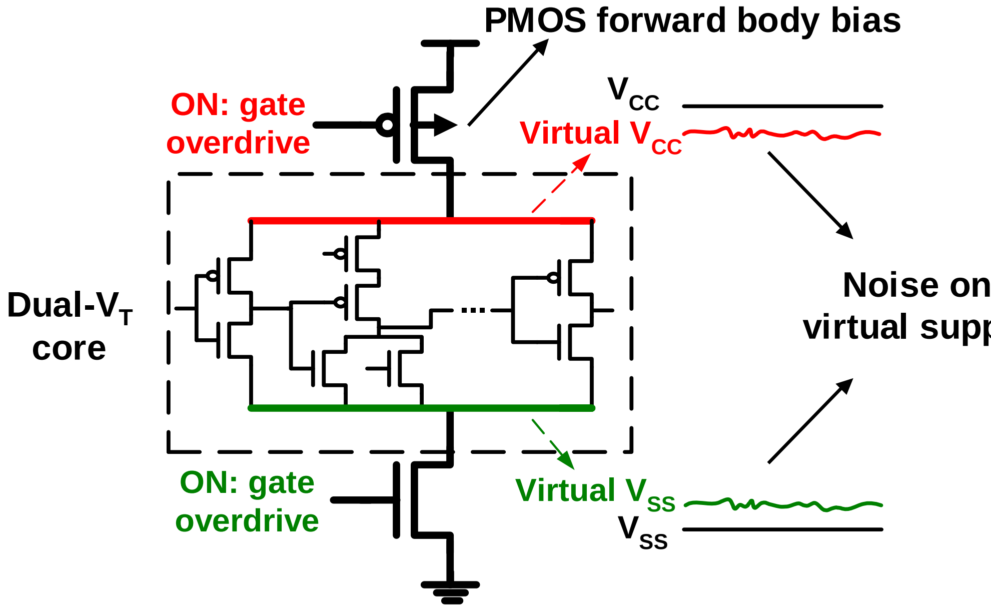
Active mode:

Forward body bias (FBB)

Idle mode:

Reverse body bias (RBB)

Dynamic Sleep Transistor

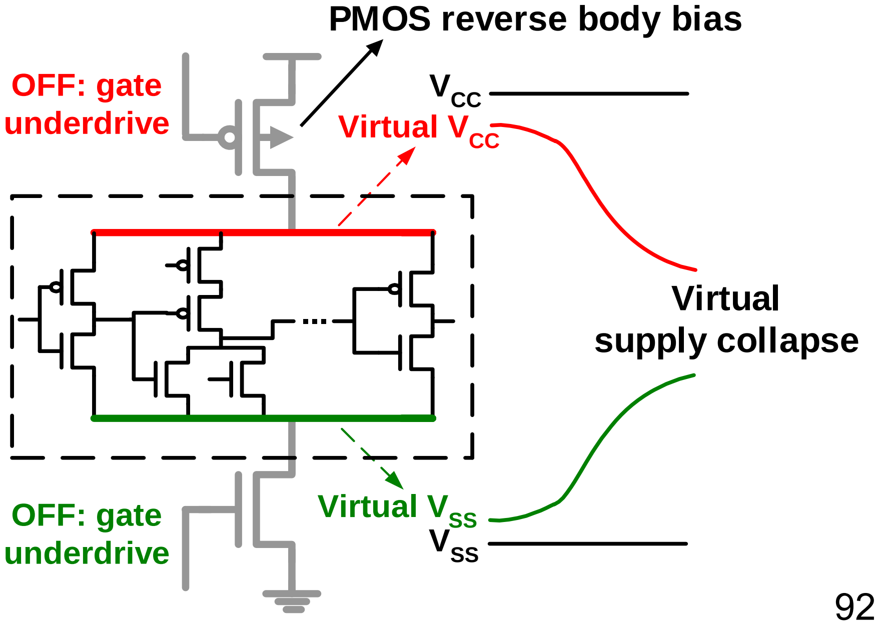


Idle mode:

Sleep transistor OFF

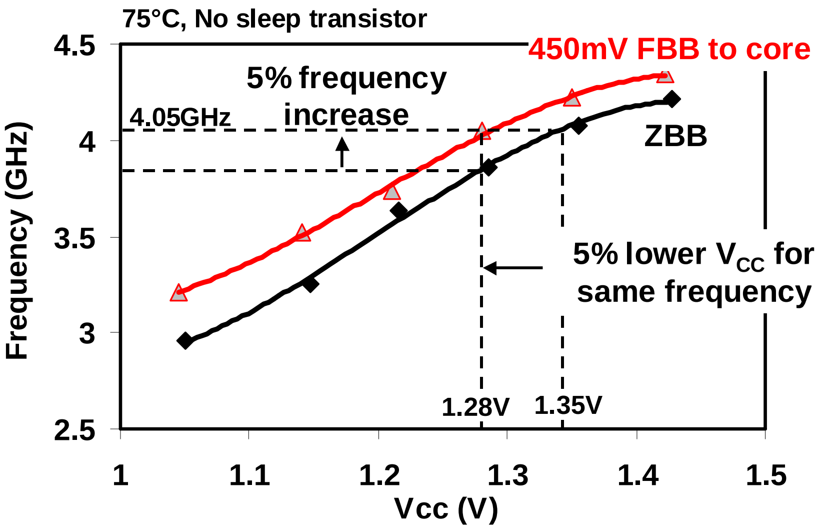
Active mode:

Sleep transistor ON

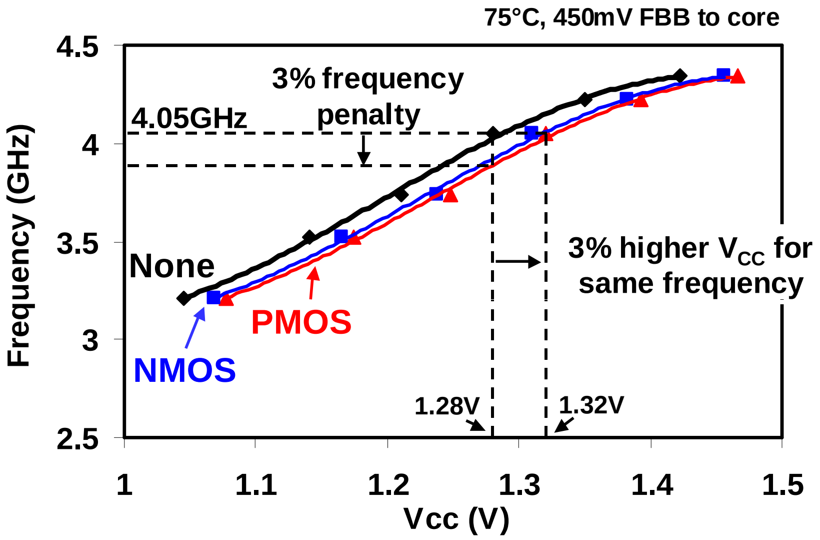


Performance Impact

Body bias



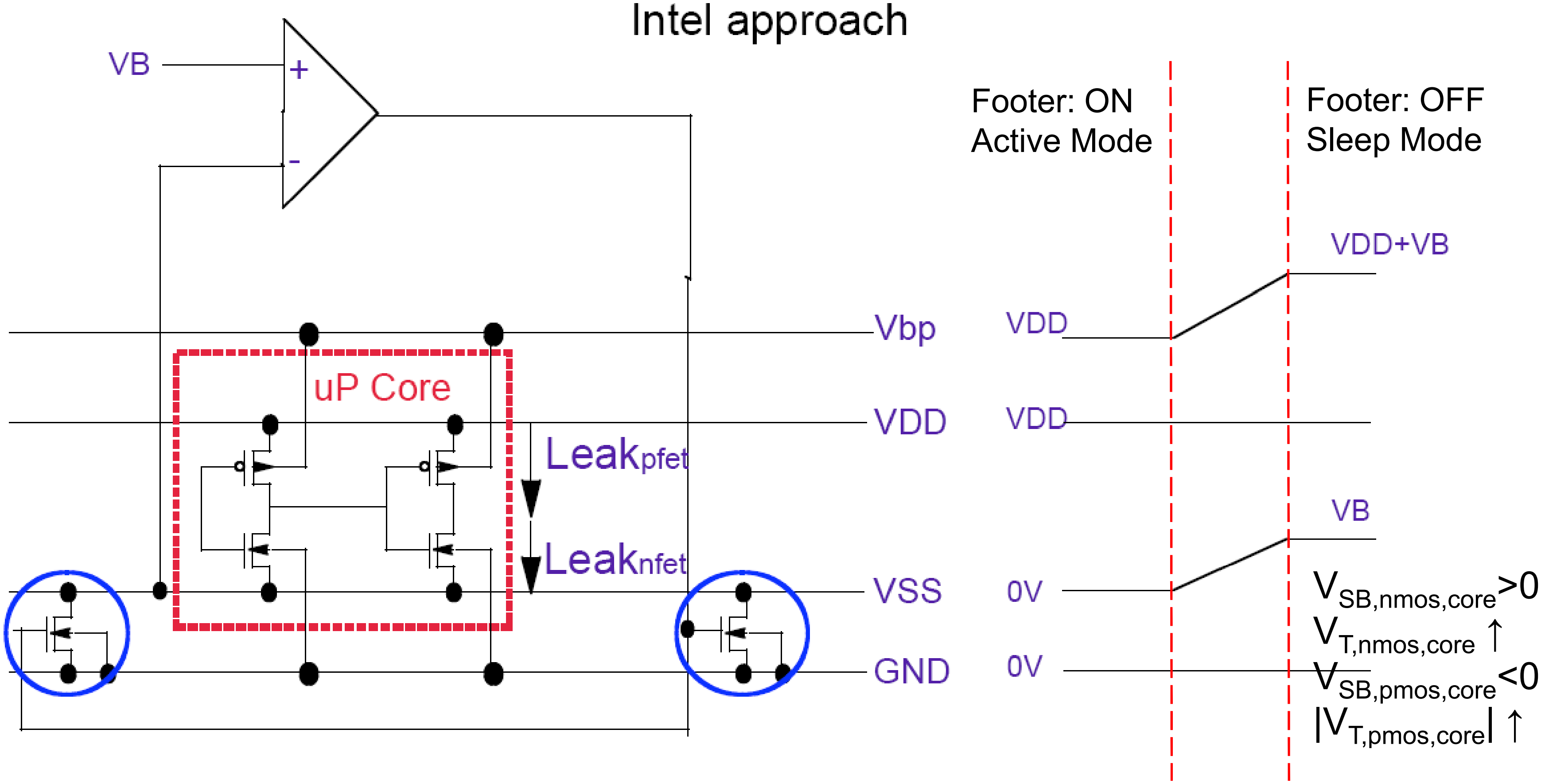
Sleep transistor



Source: De-2004

Nwell Biasing in Two-Well Process

Intel approach



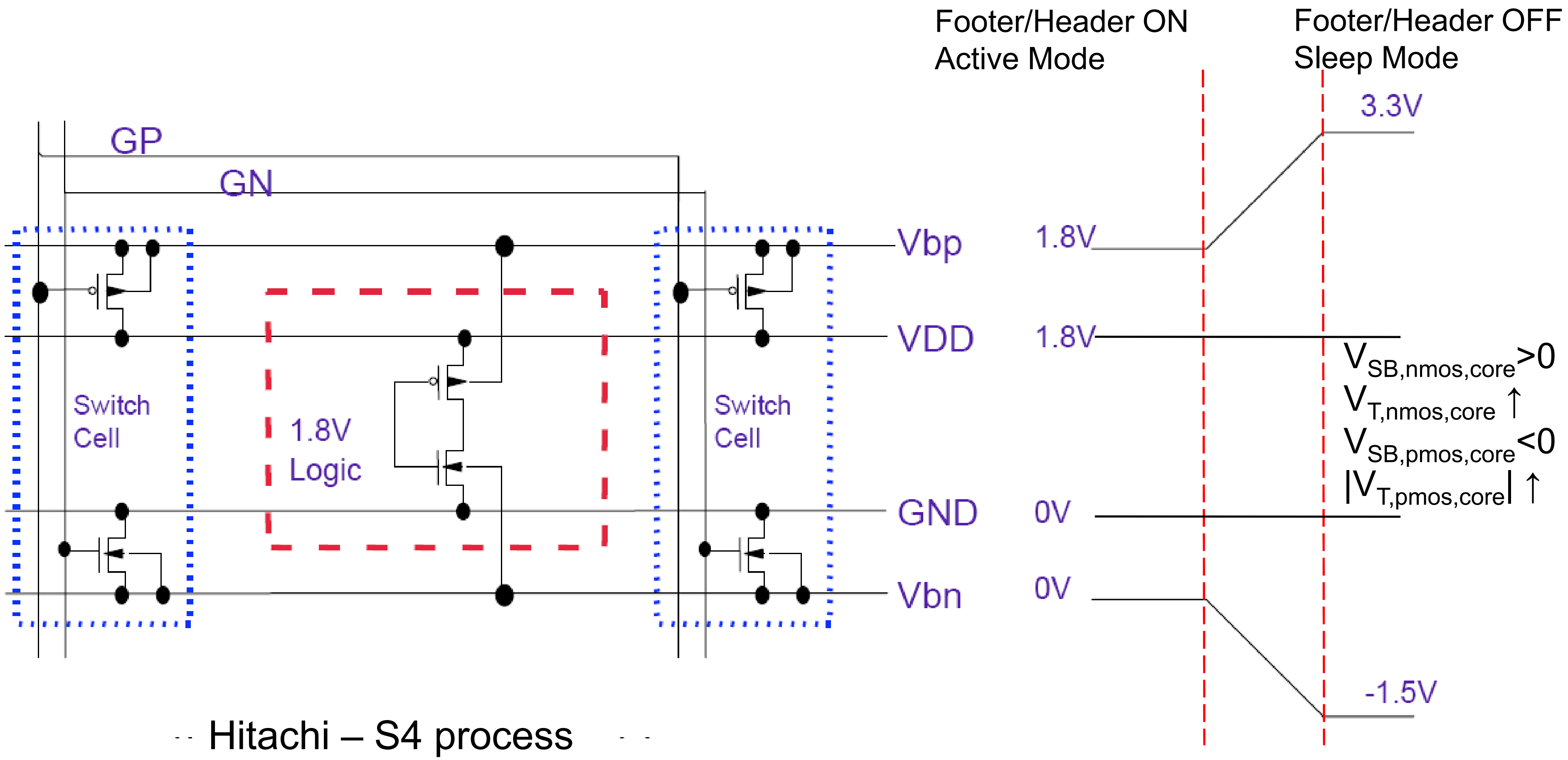
$V_{SB,nmos,core} > 0$

$V_{T,nmos,core}$ ↑

$V_{SB,pmos,core} < 0$

$|V_{T,pmos,core}|$ ↑

Nwell and Pwell Biasing in Triple-Well Process



- - Hitachi – S4 process - -

Basic Guidelines for Power Minimization

- Do not do more than necessary
  - avoid wasteful power dissipation: clock gating
  - do not optimize for ‘worst case’ but for the ‘current case’: DVFS
  - react to the environment: DPM
  - use bus encoding, reduced swing signaling, etc.
- Use Locality of reference
  - store results locally
  - avoid communication over long distances
  - avoid off-chip communications (1000 times more expensive)
- Be energy aware at all levels of your system: technological, system architecture, operating system, applications
  - do the tasks at the most energy-efficient platform/way
  - match algorithm with architecture