Leakage Power Reduction

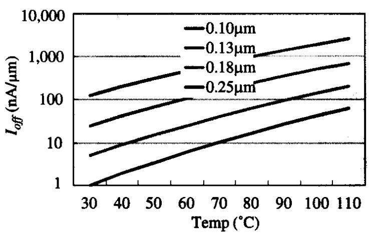
Ken Stevens

Leakage and Scaling

- Leakage and scaling are directly related
  - Constant field scaling would result in continued exponential increase in leakage
    - V_t would continue to be reduced, increasing I_{off}
    - $I_{off} = I_0^{(-\frac{qV_t}{mkT})}$
      $I_0 =$ current at threshold: $\approx$ 1–10μA/μm for 100nm devices
      $m =$ body coefficient (≈ 1.2)
    - $I_0 \propto$ inversion charge density at threshold
      $Q_i \approx (1\ldots2)\frac{kT}{q}C_{ox}$

Leakage and Scaling

- $P_{off} = W_{tot}V_{dd}I_{off}$
- W_{tot} can increase by $\approx 50\%$ per generation
- Overall leakage can increase by $\approx 7.5\times$ per generation



Scaling Throttled by Leakage

- Low V_t adversely effects:
  - Short channel effects due to V_t roll-off
  - DIBL (Drain Induced Barrier Lowering)
  - Within-die process variation

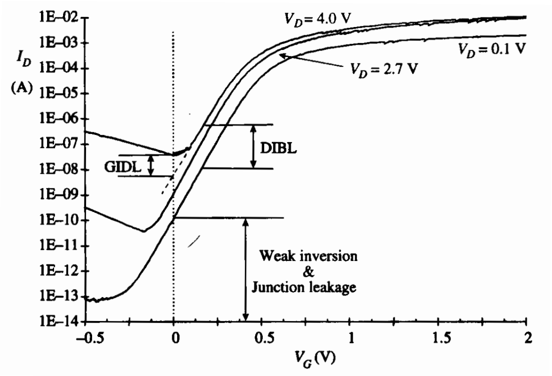
Primary Leakage Problems

1. Overall power of a chip
2. Stability of 6T SRAM cells
3. Noise immunity of dynamic logic gates

Sources of Leakage

1. p-n junction reverse bias current
2. Weak inversion
3. DIBL (Drain Induced Barrier Lowering)
4. GIDL (Gate Induced Drain Leakage)
5. Punchthrough
6. Narrow width effect
7. Gate oxide tunneling
8. Hot carrier injection

Leakage Sources



P-N Junction Reverse Bias Current

- Current can leak from source and drain into the well or substrate
- Relatively minor contributor
- A function of junction area and doping concentration
- Heavy doping can cause Zener and band-to-band tunneling
- Two main mechanisms
  - Minority carrier diffusion or drift near edge of depletion region
  - Electron-hole pair generation in the depletion region of the reverse biased junctions.

Weak Inversion or Junction Leakage

- Current leaking between source and drain
- Major contributor
- $0 < V_{gs} < V_t$
- Exponential current in semi-log plot
- Carriers diffuse along channel surface
- Caused by cross coupling on gate wires, power supply noise, etc.
- Exacerbated by low V_t (slope of exponential)

DIBL - Drain Induced Barrier Lowering

- Current leaking between source and drain
- Primary contributor
- Source potential barrier lowered by high voltages on drain
  - Depletion region of drain interacts with source
  - Occurs near channel surface
- Source injects carriers without gate playing a significant role
- Enhanced by short L_{eff}
- Effectively increases linear region current

DIBL Model

Accurate leakage model that includes DIBL
Similar to our standard leakage model

$$I_{off} = I_0 e^{\frac{q(V_g - V_s - V_t - \gamma' V_s + \eta V_{ds})}{mkT}} \left(1 - e^{-\frac{qV_{ds}}{kT}}\right)$$

$$I_0 = \mu_0 C_{ox} \frac{W}{L_{eff}} \left(\frac{kT}{q}\right)^2 e^{1.8} e^{-\frac{q\delta V_t}{\eta kT}}$$

γ' is the body effect coefficient

η is the DIBL coefficient r

GIDL - Gate Induced Drain Leakage

- Current leaking between drain and well or substrate
- Minor contributor
- Gate to drain overlap (bird's beak) can cause deep depletion
  - Mechanisms: band-to-band tunneling, trap-assisted tunneling, thermal emission and tunneling.
- Enhanced by reducing t_{ox} and increasing V_g
- Worst for moderate doping
  - Low doping won't create needed high electric fields
  - High doping limits depletion volume
- Mostly an issue at “burn-in” voltages and for FLASH

Punchthrough

- Current deep in well or substrate between source and drain
- Minor contributor
- Depletion regions of drain and source become close enough deep in the channel to conduct
- Current varies quadratically with V_d
- This is considered a subsurface version of DIBL

Narrow Width Effect

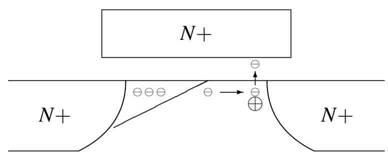
- Current between source and drain
- Minor negative contributor
- Thresholds *increase* for very short gate widths

Gate Oxide Tunneling

- Current between Gate and well or substrate
- Minor contributor
- Electric field across oxide can cause tunneling through oxide bands
- Largely controlled in current processes

Hot Carrier Injection

- Increased or decreased leakage current (DIBL, weak inversion, etc.)
- Significant contributor, larger for new nodes
- V_t offset caused by charge trapped in gate oxide



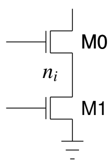
Controlling Leakage

- Leakage can be controlled in four ways:

1. **Topologically**
2. Reducing total transistor width $P_{off} = W_{tot} V_{dd} I_{off}$
3. Removing power (sleep transistors) – "dark silicon"
4. Modifying transistor thresholds V_t

Transistor Stack Effect

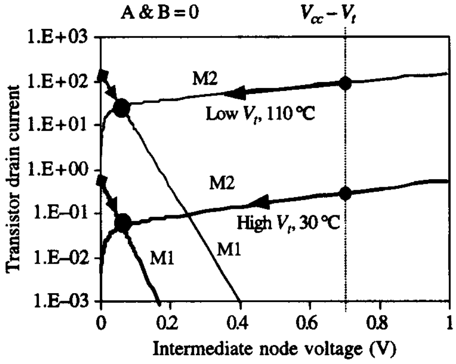
- Leakage is effected by data dependencies
  - Sensitive to vectors
- If series transistors are turned off, leakage is greatly reduced
  - Leakage through M0 will create an intermediate parasitic voltage on node n_i
  - Creates $V_{gs} < 0$ for M1
  - Exponentially reduces leakage in M1
  - DIBL vastly decreases in M1 due to voltage on n_i
  - Body effect increases V_t further reducing leakage



Transistor Stack Effect

- Internal node voltage determined by transistor crosspoints
- Voltage stabilizes at $\approx$ 50-100mV at all corners
- Note: Substantial delay to reach stable point (standby mode)
- For NAND gate:
  - When $n_i = 0$, voltage determined by current through M0
  - When $n_i = V_{dd} - V_t$, voltage determined by M1

Transistor Stack Effect



Transistor Stack Effect

- Reduction mainly dependent on
  - Temperature (current $\propto kT$)
  - Threshold V_t
  - Number of series transistors

	High V_t	Low V_t
2 NMOS	$10.7\times$	$10.0\times$
3 NMOS	$21.1\times$	$18.8\times$
4 NMOS	$31.5\times$	$26.7\times$
2 PMOS	$8.6\times$	$7.9\times$
3 PMOS	$16.1\times$	$13.7\times$
4 PMOS	$23.1\times$	$18.7\times$

Stack Effect Time Constants

- Convergence depends on:
  - Drain-body junction and gate-overlap (bird's beak) caps
  - Initial voltage condition
  - Subthreshold leakage current (with T and V_t dependence)

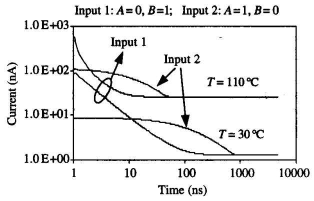


Figure 3.10 Temporal behavior of leakage current in transistor stacks for different temperatures and initial input conditions.

Input A closest to ground, so slowest when internal node charged low.

Stacked Transient Model

- Time constant determines leakages and utility of low leakage vectors
- Stabilization delays vary from microseconds to milliseconds, and
- For low V_t in deep sub-μ technology, this could be 5–50ns
- Largest delays when:
  - Low temperature
  - High V_t
  - Internal nodes charged high
- Approximate delay calculated is sum of delay of each transistor, starting from transistor closest to rail, discharges its parasitic node through the series stack.

Stack Leakage Current Calculation

- Assume most leakage is from drain to source in n-FET stacks
- Current will be identical in each transistor
- Use Kirchoff's Current Law to calculate the current
- Recursively calculate V_{ds}
- For transistor i:

$$V_{ds_i} = \frac{nkT}{q(1+\gamma')} \ln\left(1 + \frac{I_{0_{i-1}}}{I_{0_i}}\left(1 - e^{-\frac{qV_{ds_{i-1}}}{kT}}\right)\right)$$

where γ' is the linearized body effect coefficient

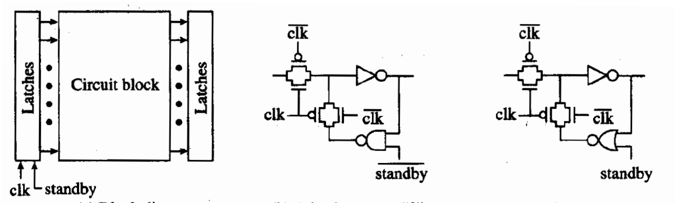
Standby Leakage Input Vector Activation

- Most gates in function blocks contain stacks
- Three methods used to calculate low leakage vectors

1. Topological algorithms
   - Good for regular data structures (adders, multipliers)
2. Genetic/testability search algorithms
   - Can work well for random function logic
3. Random vector simulations

Standby Leakage Input Vector Activation

- Find the vector for lowest leakage
- Generate standby logic that drives low leakage vector



(a) Block diagram (b) A latch to store "0" (c) A latch to store "1"

Standby Leakage Input Vector Activation

- Example:
  - 32-bit Kogge-Stone adder
  - Static gates
  - One high V_t and one low V_t design
- 1,000 random vectors generated

Threshold	vector	% reduction
high V_t	best	0.00%
high V_t	average	35.4%
high V_t	worst	60.7%
low V_t	best	0.00%
low V_t	average	33.3%
low V_t	worst	56.5%

Controlling Leakage

- Leakage can be controlled in three ways:

1. Topologically
2. **Reducing total transistor width** $P_{off} = W_{tot}V_{dd}I_{off}$
3. Removing power (sleep transistors) – "dark silicon"
4. Modifying transistor thresholds V_t

Reducing total transistor width

- Leakage can be significantly reduced by efficient designs:
  - Efficient transistor structures (domino, pass gate, etc.)
  - Less concurrency in designs
  - Optimizing power delay points (design compiler algorithms, etc.)
  - Asynchronous design
    - Fewer inverters (especially big clock drivers) – maybe not!
    - Sequential logic design

Controlling Leakage

- Leakage can be controlled in three ways:

1. Topologically
2. Reducing total transistor width $P_{off} = W_{tot}V_{dd}I_{off}$
3. **Removing power (sleep transistors) – "dark silicon"**
4. Modifying transistor thresholds V_t

Sleep Transistors

- Gate power supply to blocks with *sleep transistor*
- Can be applied to dual-V_t design:
  - High V_t for sleep transistor
  - Low V_t for majority of logic

Sleep Transistors

- Drawbacks:
  - Negative performance impact
  - Large area for sleep transistors
  - Power-up latency
  - Calculating optimal sizing of sleep transistor can be difficult
  - Potential for power supply noise when turning on and off

Sleep Transistors

- Extra area required due to real and virtual V_{DD} to cell
- Sizing of sleep transistor
  - Too large: waste area and excess energy
  - Too small: degrade circuit speed, power supply noise
  - Data dependence
  - Different delays than critical path

Sleep Transistors

Delay of 8 bit multiplier with sleep transistor

	Static	Dual-V_t sleep tran	
	Delay	$\frac{W}{L}=60$	$\frac{W}{L}=170$
Vector		rel. perf (%)	rel. perf (%)
A: X=00→FF, Y=00→81	8.96ns	81.9%	95.0%
A: X=7F→FF, Y=81→81	8.93ns	95.2%	98.3%

Power Down Schemes

- Can save power but has challenges
  - 100μs or more settling time
  - Partitioning due to periodic access requirements
  - Loss of data due to power-down
    - Data backup, *drowsy latches*, . . .
  - Interface between blocks powered off and on
    - Gate keeper or pull-down resistor at interface
    - Chip inputs low before power-off due to ESD leakage
    - Chip inputs reset before power on

Controlling Leakage

- Leakage can be controlled in three ways:

1. Topologically
2. Reducing total transistor width $P_{off} = W_{tot} V_{dd} I_{off}$
3. Removing power (sleep transistors) – “dark silicon”
4. **Modifying transistor thresholds V_t**

Dual-V_t Design

- Individual transistors or gates get different dopings
- Doping based on circuit timing requirements
- Critical paths get low V_t for increased performance
- Non-critical paths get high V_t for decreased leakage
- Challenging CAD algorithms to determine partitioning
  - optimizations include sizing, logic gate complexity (logical effort), . . .
- Diminishing return as number of critical paths increases
  - unfortunately this is the best design target for power/performance

Dual-V_t Domino

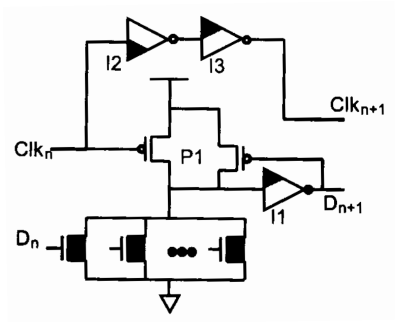
- Individual domino gates can have both high and low V_t
- Significant standby leakage improvements
- No reduction in performance!
- Set/reset functions assigned low V_t transistors
- Keeper function mapped to high V_t transistors
- In precharged (clocked) design, can make precharge transistors high-V_t

Dual-V_t Domino

- Precharged design naturally precharges to near optimal standby leakage vector
  - all low-V_t devices will automatically be turned off
    - domino stage will have all inputs low
    - static stages will have all inputs high (turns off static transition logic stack)
  - only need to explicitly set inputs to domino pipeline from other logic families

Dual-V_t Domino

What other means can you use to improve performance?



Adaptive Body Biasing (ABB)

- Modulate V_t through body biasing
- Reverse body bias transistors to increase V_t
- Can also be applied to
  - speed up slow paths
  - mitigate process variation
- Normally want a lower nominal V_t when applying ABB

Adaptive Body Biasing (ABB)

- Can be applied across wide range
  - all n- or p-FETs
  - entire function block (FUB)
  - critical paths
  - gate
  - single transistor

Body Biasing Effects

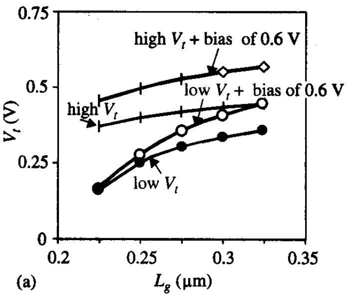
- Source and drain voltages in deep submicron devices effect channel
- Source/drain-body reverse biased diode junction depletion regions contribute significantly to channel charge
- Reduces control of gate and body over channel
  - produces V_t roll-off and body effect reduction
- DIBL and V_t **roll-off** are affected by body biasing
  - transistors degrade due to widening diode depletions

Scaling & Effectiveness of Body Biasing

- Body Biasing requires lower V_t devices for positive bias
- But low V_t devices show vastly diminished affects
- Lower body effects caused by:
  1. reduced channel doping for V_t reduction
  2. low V_t has more diode depletion charge
  3. body biasing further increases diode depletion

Effectiveness of Body Biasing

- Diminishes in deeper submicron devices
- Diminishes with lower thresholds



Low Voltage Technologies

- Tradeoff between voltage and throughput:

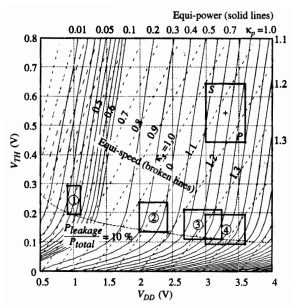
$$P = C_{sw} V_{DD}^2 f + I_0^{\left(-\frac{qV_t}{mkT}\right)}$$
$$T_{pd} = \frac{\beta C_L V_{DD}}{(V_{DD}-V_t)^{\alpha}}$$

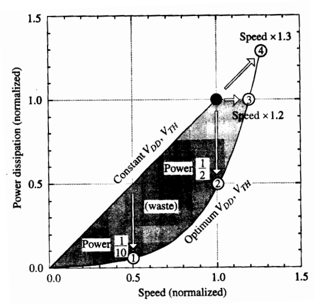
- How does one maintain performance with low V_{DD}?
  1. **lower** V_t
  2. dual-V_t
  3. multiple supply voltages (RAMs)
  4. parallelism to compensate for slower devices
- What are the costs for the performance?

Lowering V_t and V_{DD}

- For a given process we can pick V_{DD} and V_t points
- Following two graphs show tradeoff for:
  - aggressively lowering V_t
  - leakage limits for lowering V_t
  - lowering power supply

Lowering V_t and V_{DD}





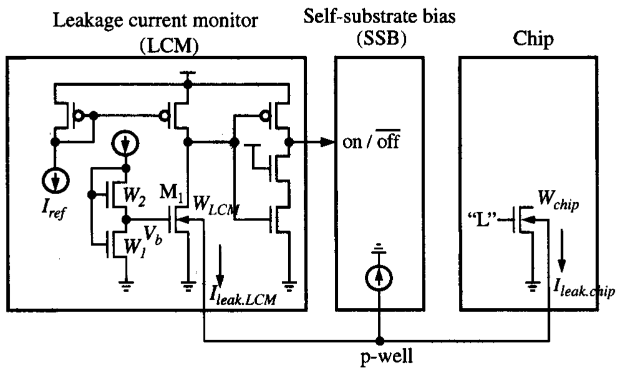
Lowering V_t and V_{DD}

- Observations:
  1. up to 30% increase in power and performance for aggressive V_t scaling
  2. joint scaling V_t and V_{dd} allow
     - 20% speed improvement at same power
     - same speed at 50% power reduction
     - up to 90% reduction in power for 50% loss in performance
  3. today's processes don't afford these ranges. . .
     - starting at 0.9 – 1.2 V
     - V_t at 0.2 – 0.35 V
  4. this largely ignores the effect of leakage that limits these benefits

Variable Threshold Voltage CMOS

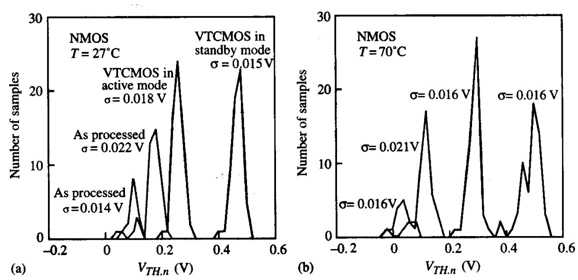
- V_t control by substrate bias control with feedback
  1. feedback by dynamically changing effective size/ratio of mirrored current.
  2. if leakage $I_{leak.LCM}$ is larger than target current I_{ref} self substrate bias turns on due to body bias.
  3. self-substrate bias lowers V_{BB} which raises V_t, reduces $I_{leak.LCM}$.
  4. when $I_{leak.LCM}$ becomes smaller than I_{ref}, self-substrate bias is turned off.
- Current mirrors set to control bias of substrate, pumping out

Variable Threshold Voltage CMOS



Variable Threshold Voltage CMOS

- In 300nm CMOS, 3.3V power supply:
- How effective will this be in 32nm process?



Leakage Review

1. major components of leakage (temp, threshold, body coefficient, etc.)
2. how leakage has scaled in the past
3. how leakage will scale in the future
4. how leakage effects Moore's Law
5. gate structures with highest leakage (series or parallel)
6. how leakage effects noise immunity of domino gates

Leakage Review

1. primary sources of leakage
   a. DIBL (cause & effect of L_{eff})
   b. weak inversion or junction leakage (cause & V_t effect?)
   c. hot carrier and gate oxide tunneling
2. four *design* methods of controlling leakage
3. main source of reduced leakage in series stack (DIBL, body effect)
4. dc characteristics of stacked transistors
5. approximate scaling of reduction in leakage of transistor stacks
6. stack effect time constants and pattern dependencies

Leakage Review

1. three main effects for time constants: temp, V_t, parasitics
2. algorithms for stack vector generation
3. circuits for standby mode
4. transistor width reduction methods
5. sleep transistor design, sizing
6. sleep transistor drawbacks

Leakage Review

1. threshold modifications
   a. dual-V_t concept
   b. best design targets for dual-V_t
   c. dual-V_t domino design
   d. dual-V_t domino benefits
   e. initial threshold target using body biasing