Leakage Tolerant Circuits, Sub-threshold Logic

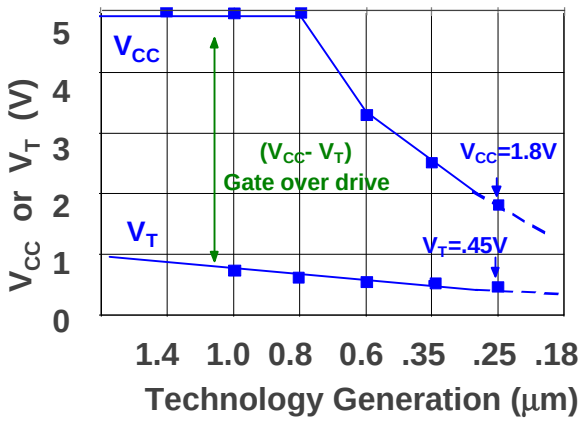
Kaushik Roy
Electrical & Computer Engineering
Purdue University

Outline

- High Performance and Low-Power Circuits
  - Leakage control (CAD and circuit techniques)
    - Stacked CMOS with Gated-Vdd (Application: DRI-cache)
    - Multiple VT
    - Dynamic VT
  - Circuit techniques:
    - MTCMOS, VTCMOS, DTMOS, SCCMOS, etc.
  - SOI implementation
- Ultra low voltage digital sub-threshold logic
  - Medical applications, bursty versus non-bursty mode

Constant Voltage vs Field Scaling

- Recently: constant e-field scaling, aka voltage scaling
- V_{CC} → 1V
- V_{CC} & modest V_T scaling
- Loss in gate overdrive (V_{CC}-V_T)



Voltage scaling is good for controlling IC's active power, but it requires aggressive V_T scaling for high performance

Delay

$$t_d = \frac{C_L V_{DD}}{I_D}$$

$$t_d = \frac{C_L}{(\frac{W}{2L})\mu C_{ox} V_{DD}(1-\frac{V_T}{V_{DD}})^2}$$ Long Channel MOSFET

$$t_d = \frac{C_L}{WC_{ox}u_{SAT}(1-\frac{V_T}{V_{DD}})}$$ Short Channel MOSFET

$$t = \frac{CL^{0.5}T_{ox}^{0.5}}{V_{DD}^{0.3}(0.9-V_T/V_{DD})^{1.3}}(\frac{1}{W_n}+\frac{2.2}{W_p})$$ [1]

[1] C. Hu, "Low Power Design Methodologies," Kluwer Academic Publishers, p. 25.

Performance significantly degrades when V_{DD} approaches $3V_T$.

V_T Scaling: V_T and I_{OFF} Trade-off

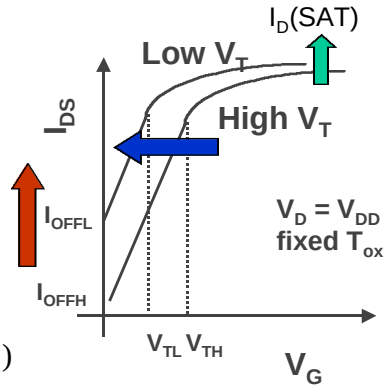
Performance vs

$V_T \downarrow I_{OFF} \uparrow I_D(SAT) \uparrow$

$$I_{OFF} \propto I_{subth} \propto \frac{W_{eff}}{L_{eff}} K_1 e^{(V_{GS}-V_T)}$$

$$I_D(SAT) \propto \frac{W_{eff}}{L_{eff}} K_2 (V_{GS}-V_T)^2$$

$$I_D(SAT) \propto K_3 W_{eff} C_{ox} u_{SAT} (V_{GS}-V_T)$$

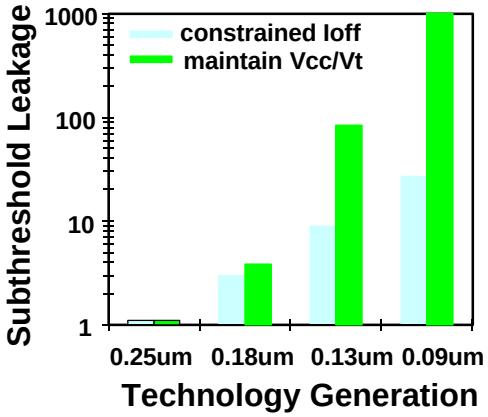


⇨ As V_T decreases, sub-threshold leakage increases

⇨ is a barrier to voltage scaling

5

Barriers to Voltage Scaling

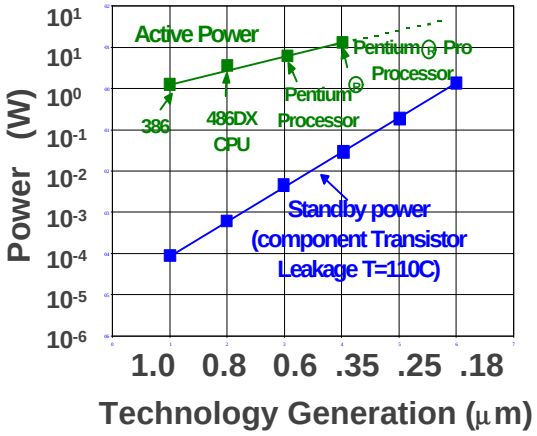


- Leakage power
- Short-channel effects
- Soft error
- Special circuit functionality

6

Why Excessive Leakage an Issue?

- Leakage component to active power becomes significant % of total power
- Approaching ~10% in 0.18 μm technology
- Acceptable limit less than ~10%, implies serious challenge in V_T scaling!



7

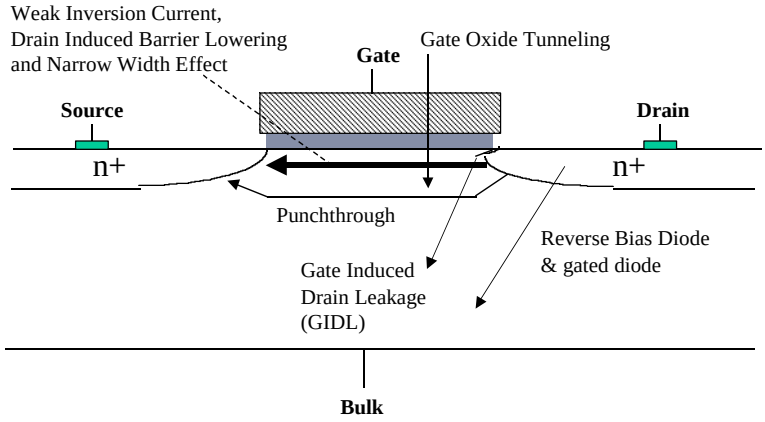
Low-Vdd Low-Vt Design

- Stacked CMOS
- Dual-threshold CMOS
- Dynamic-threshold CMOS

Leakage control techniques

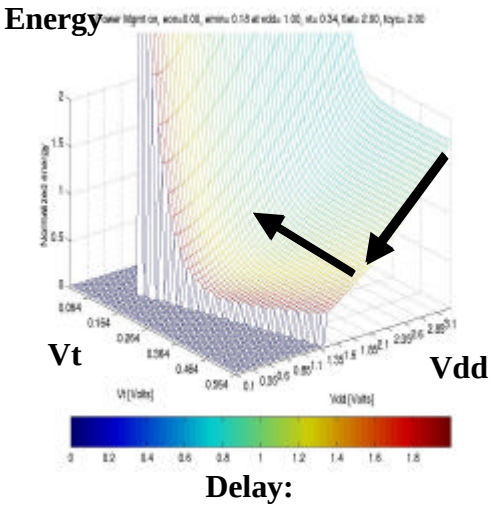
Sources of Leakage

- From Keshavarzi, Roy, & Hawkins (ITC 1997)



Motivation

- **Reduce Vdd**
  - Switching energy α Vdd^2
  - Increase delay
- **Reduce Vt**
  - Decrease delay
  - Leakage energy α $e^{(-Vt)}$
- **Leakage Control permits**
  - lower voltage
  - lower Vt



Leakage Control

- Needed most when circuit is idle
  - inputs latched & clocking removed
  - supply voltage is still applied
- Can exploit input dependence
  - turn off stacks of transistors
  - intrinsic self-reverse biasing
- Multiple V_t useful
  - High V_{th} : suppress sub-threshold leakage
  - Low V_{th}: achieve high performance

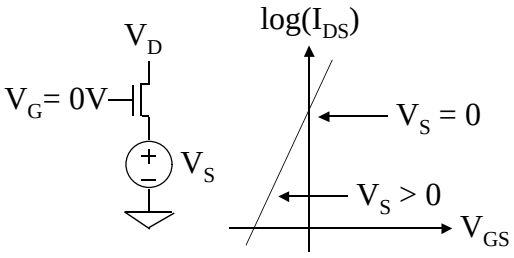
Leakage Control using Self Reverse Bias

- Subthreshold current dominant in sub-μ
  - this is the component we concentrate on
- Drain induced barrier lowering (DIBL) and body effect modeled as V_t shift
- Gate induced drain leakage (GIDL) and gate oxide tunneling may be problem in future
- Subthreshold current model based on BSIM
  - body effect linear for small V_S

$$I_{subth} = A \times e^{\frac{1}{\eta u_T}(V_G - V_S - V_{TH0} - \gamma' V_S + \eta V_{DS})} \times \left(1 - e^{\frac{-V_{DS}}{u_T}}\right)$$

Self-reverse bias

- Primary effect:
  - $V_{GS} < 0$
  - move down subthreshold slope
- Secondary effects:
  - Drain Induced Barrier Lowering
  - Body effect

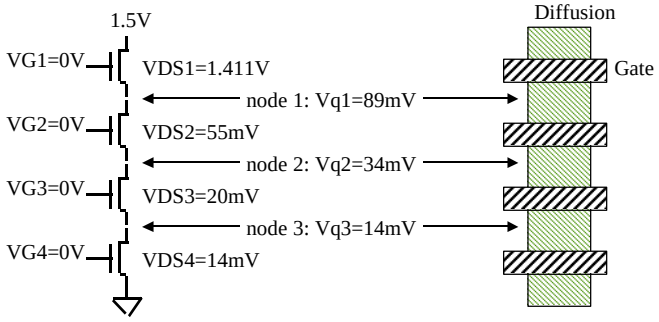


$V_{DS} \downarrow \Rightarrow V_T \uparrow$

$V_S \uparrow \Rightarrow V_T \uparrow$

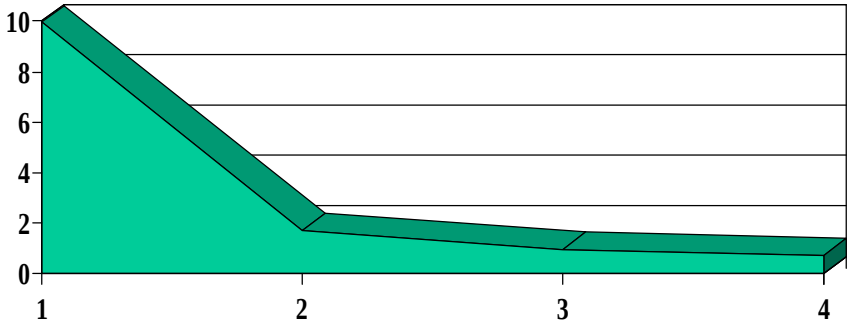
"Stacking Effect"

- Intrinsic self-reverse biasing of V_{GS} in stack



Leakage vs. Transistors Off

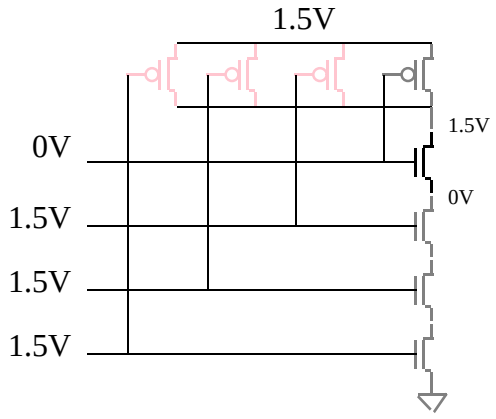
Leakage [nA]



Number of transistors off in stack

Input dependence of leakage

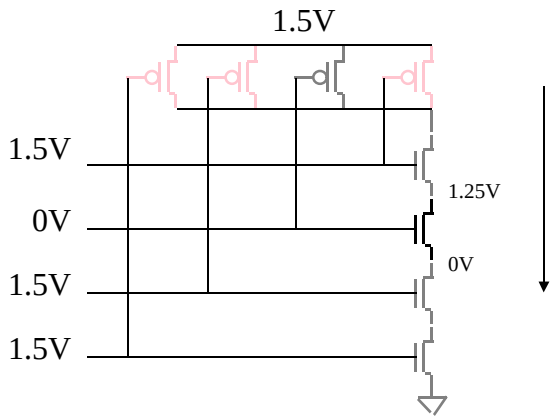
- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



I_{ddq} = 9.96nA

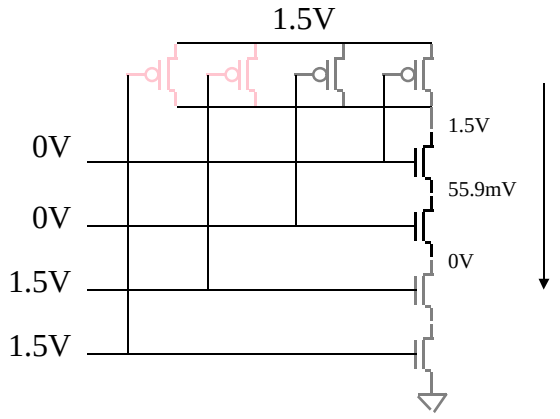
Input dependence of leakage

- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



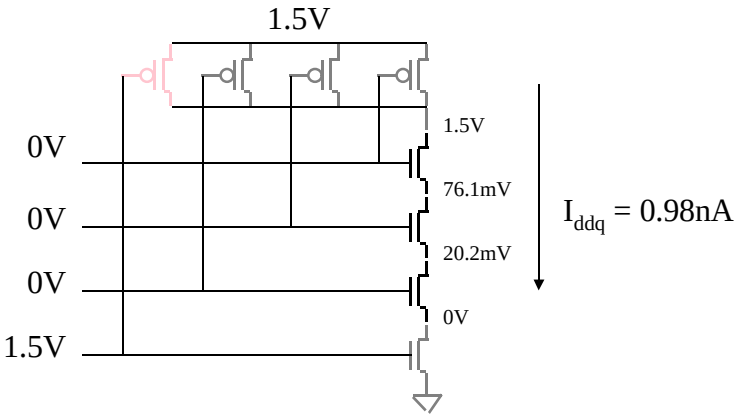
Input dependence of leakage

- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



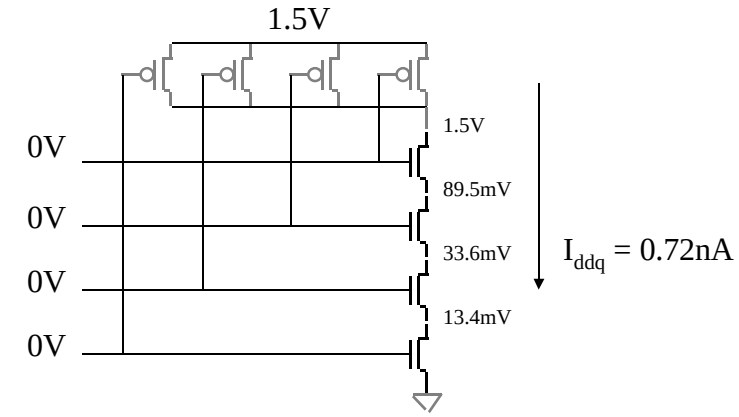
Input dependence of leakage

- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



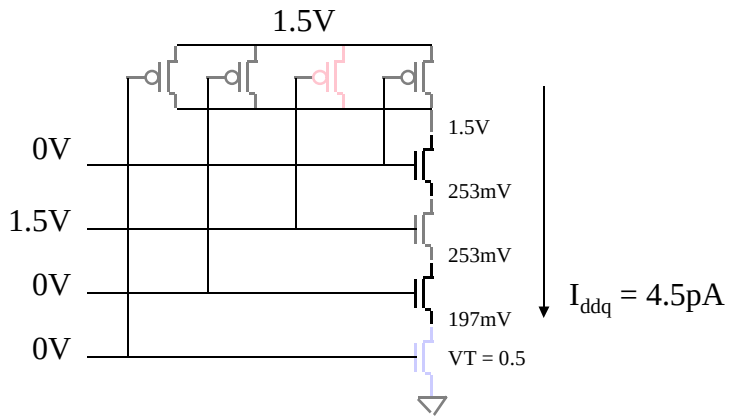
Input dependence of leakage

- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



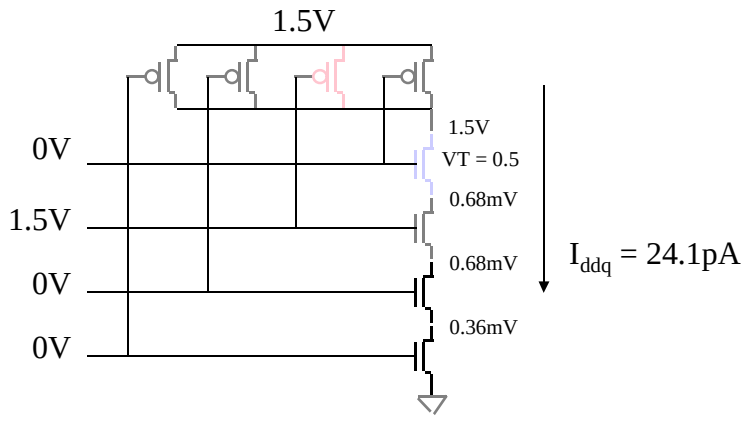
Input dependence of leakage

- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



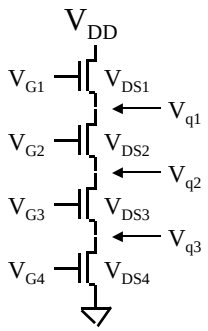
Input dependence of leakage

- Consider 4 input NAND (V_{DD} = 1.5V, V_T = 0.25V)



Model of stacking

- In a simple transistor stack
  - ignore transistors which are ON
  - calculate V_{DS} of each transistor
  - use I_{subth} equation to calculate leakage



Selection of Standby Mode Inputs

- At gate level
  - evaluate all possible inputs
  - for each input vector
    - replace ON transistors by shorts
    - decompose remaining circuit into disjoint leakage paths
    - apply stack leakage model to each path
- For more complex circuits
  - build look-up table for each sub-circuit
  - use ATPG or GA approach to select minimum leakage input vectors

Results

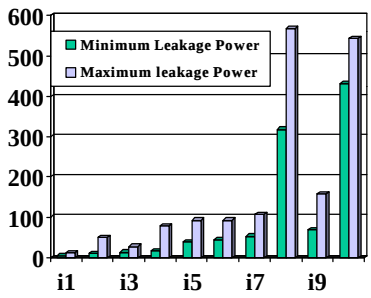
Circuit & Input vector	Model Iddq [nA]	HSPICE Iddq [nA]	Comments
4 input NAND			
ABCD=0000	0.72	0.60	Best
ABCD=1111	23.2	24.1	Worst
3 input NOR			
ABC=111	0.13	0.13	Best
ABC=000	29.9	29.5	Worst
Full Adder			
A,B,Ci=111	7.5	7.8	Best
A,B,Ci=001	56.0	62.3	Worst
4 bit ripple Add			
A=B=0000, Ci=0	102.6	91.3	Best
A=B=1111, Ci=1	102.6	94.0	Best
A=B=0101, Ci=1	258.9	282.9	Worst

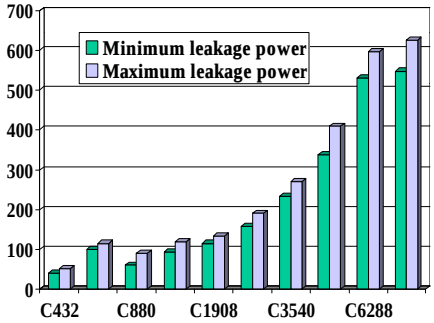
Results

Circuit & Input vector	Model Iddq [nA]	HSPICE Iddq [nA]	Comments
8 Bit Carry Select			Uses 4 bit stages
A=B=11111111,Ci=1	259.0	246.2	Best
A=B=01010101,Ci=1	690.4	759.6	Worst
4 Bit Manchester Carry Chain			Dynamic
All Gi, Pi=1, CLK=1	16.8	13.5	Best
All Gi, Pi=0, CLK=0	15.6	15.9	Best
All Gi, Pi=1, CLK=0	49.7	55.3	Worst
4 Bit MCC based Adder			
CLK=1, others=1	154.4	126.6	Best
CLK=0, others=0	144.4	134.4	Best
CLK=0, others=1	198.8	190.4	Worst

Results on Benchmark Circuits

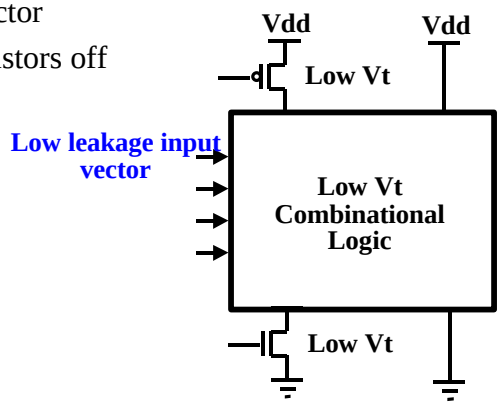
Vdd=1.0V, Vth0=0.2V
NMOSFET: L=0.5U, W=1.8U
PMOSFET: L=0.5U, W=3.6U





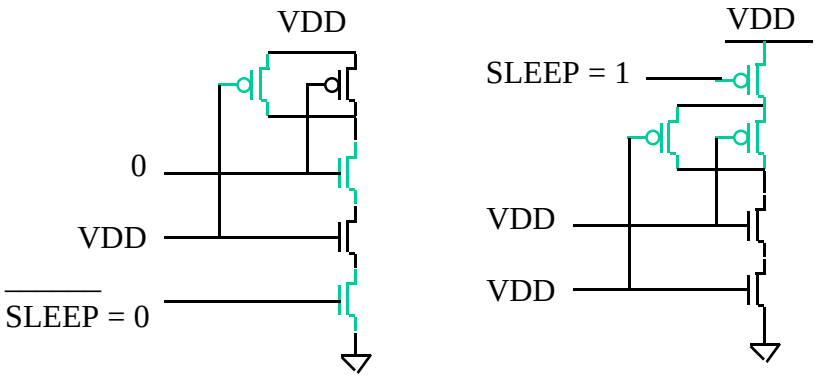
Leakage Control - Single Threshold

- Sleep mode:
  - low leakage input vector
  - leakage control transistors off
- Sub-circuits in low leakage state or in critical path
  - Normal power & ground
- Sub-circuits in high leakage state
  - Power & ground via leakage control transistors

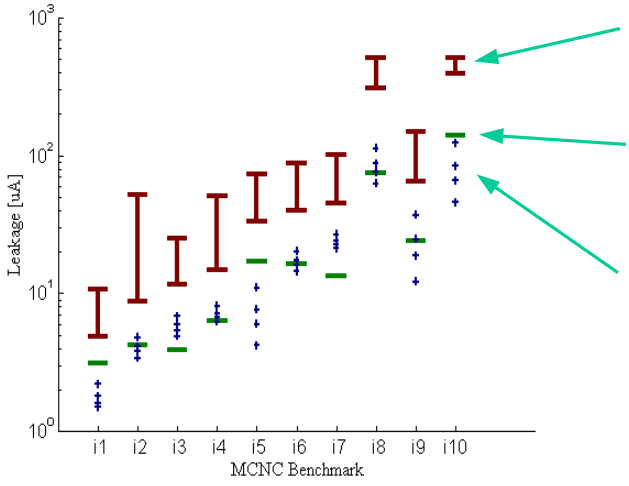


Insertion of leakage control

Create stacks of transistors which are off
(control transistors can be shared)



Leakage Control Results



Leakage range by choice of inputs

Sum of minimum gate leakages

Leakage using low Vt control transistor for 100%, 50%, 30%, and 10% sizing

Application: ICALP

- Integrated Circuit/Architecture Approach to Low Power – application of trans. stacking
  - Leakage power (DRI cache) – using transistor stacking (gated-Vdd) and simple hardware monitors
  - Dynamic power (L1 & L2 caches)

ICALP: An Integrated Approach

- Use both circuit & architecture techniques
  - aggressive low power circuit techniques
  - architecture techniques to configure hardware
- Customize hardware to fit app demand
  - e.g., caches, functional units, etc.
  - simple hardware monitors
  - compiler estimates
- Use circuit to reconfigure hardware

ICALP Goals

- Minimize both leakage and dynamic power
- Redefine architecture from power perspective
  - both architectural & compiler techniques
  - propose & evaluate power-aware systems
  - e.g., our design for a power-aware I-Cache
- Develop the first integrated evaluator
  - cycle-driven performance estimator
  - accurate power estimators

Power Management Trigger

- ICALP ISA
  - augment ISA with power mgt instructions
- Compiler + ICALP ISA
  - e.g., loop size => required I-cache size
  - e.g., estimate ILP => required issue width
- Simple hardware monitors
  - e.g., monitor miss rate and compare to threshold

ICALP I-Cache: An Example

- RAM cells large fraction of #transistors
- Potentially large fraction of leakage
- Illustrate usage of Gated-Vdd
- Integrate circuit/architecture schemes

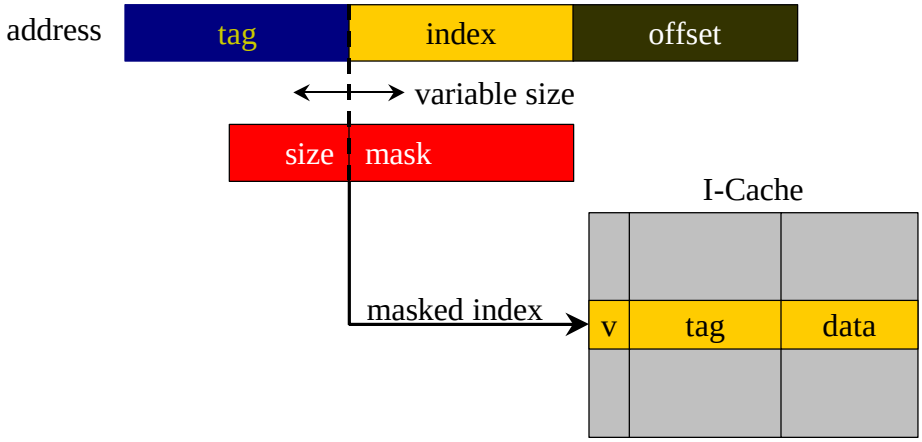
DRI-Cache: Overview

Dynamically Resizable I-Cache (DRI-Cache)

- Monitor dynamic miss rate
- Upsize if miss rate > threshold
- Downsize if miss rate < threshold
- Turn-off power to unused cache blocks
  - using Gated-Vdd
- Simple hardware implementation

DRI-Cache Resizing: How?

Mask index bits to resize cache



DRI-Cache Resizing: When?

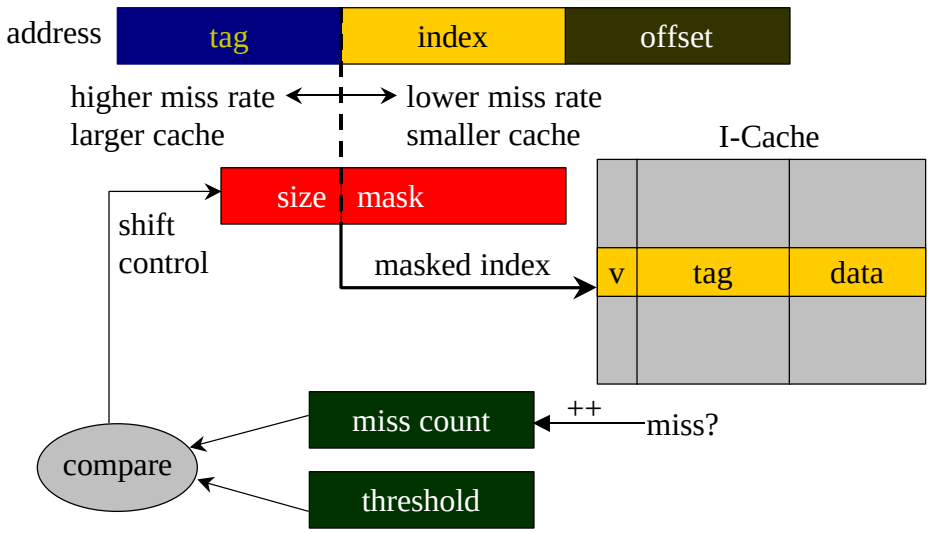
Monitor miss rate for an interval

- Hardware register tracks miss rate

At end of interval

- Compare with threshold
- Shift size mask right if lower miss rate
- Shift size mask left if higher miss rate

DRI-Cache Resizing: When?



DRI-Cache: Architectural Issues

Resizing may cause aliasing

As many tag bits as minimum size

- Larger tag RAM
- Slower tag compares

Size mask in address path

- May affect access time

Extending to D-Cache has problems

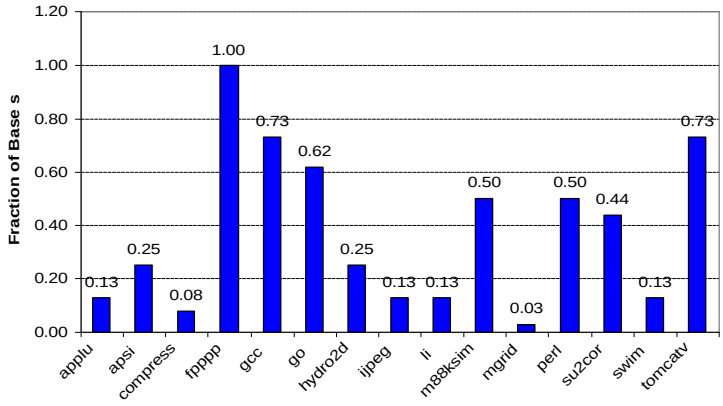
- Dirty data needs to be written back

DRI-Cache: Layout Issues

Extra transistor for Vdd or Gnd

- Less than 3% area increase
- Amortize over one/many cache blocks

DRI-Cache: Average Size



DRI-Cache: Preliminary Results

Simulation Parameters

- Simplescalar simulator
- SPEC95 benchmarks
- 2-way, 64K L1 I- and D-Cache
- 4-way, out-of-order issue
- Sense interval : 256K I-Cache accesses
- Threshold set to base case miss rate
- Measure size - estimate power

DRI-Cache: Conclusions

DRI-Cache results are encouraging

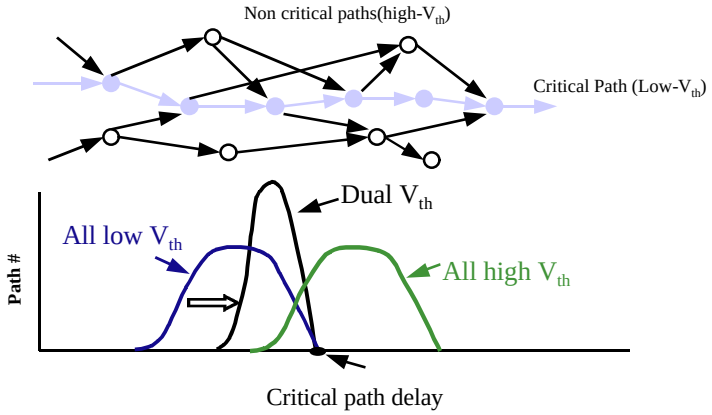
- Within 3% of base performance on average
  - size reduction by 11% - 96%
- Relatively simple hardware

Actual power reduction

- Using spice simulations on our layout
  - average static power reduction of 62%

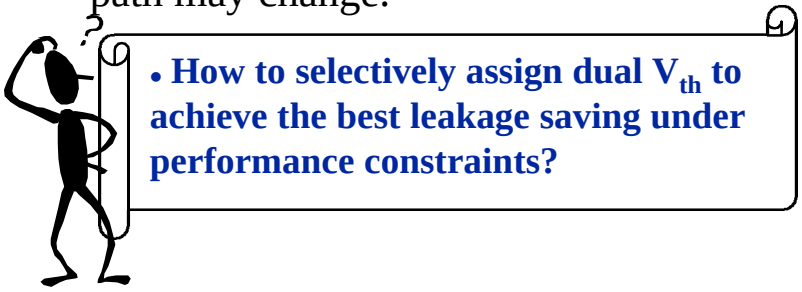
Dual Threshold CMOS

- Low-V_{th} transistors in critical path for high performance
- Some high-V_{th} transistors in non-critical paths to reduce leakage



Dual Threshold CMOS (cont'd)

- Due to the complexity of a circuit, not all the transistors in non-critical paths can be assigned a high-V_{th}, otherwise, the critical path may change.



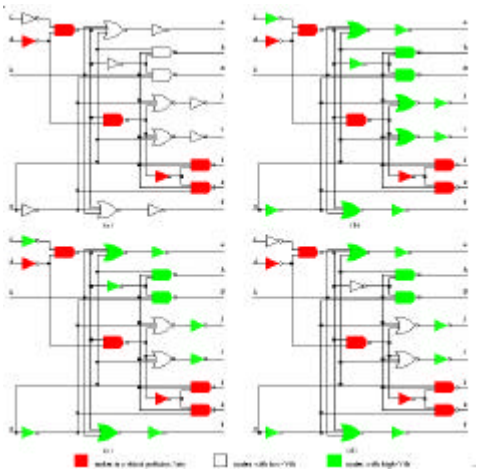
- **How to selectively assign dual V_{th} to achieve the best leakage saving under performance constraints?**

Dual Vth Results

- Example

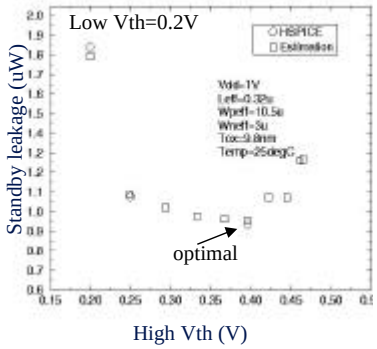
(a) original single low Vth circuit
(Vth=0.2V, Vdd=1V)
(b) Dual Vth (VtL=0.2V, VtH=0.25V)
(c) Dual Vth (VtL=0.2V, VtH=0.395V)
(d) Dual Vth (VtL=0.2V, VtH=0.46V)

Performance constraints satisfied

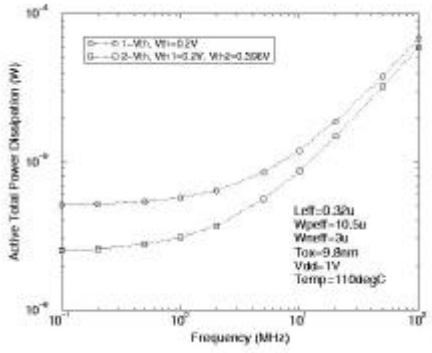


Dual Vth Results (cont'd)

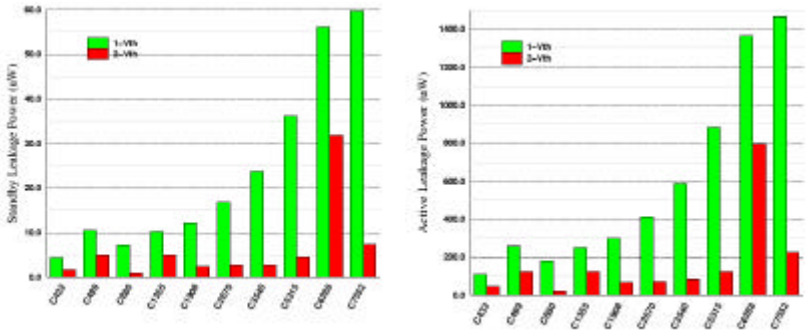
Standby Leakage



Active Power



ISCAS Benchmark Results

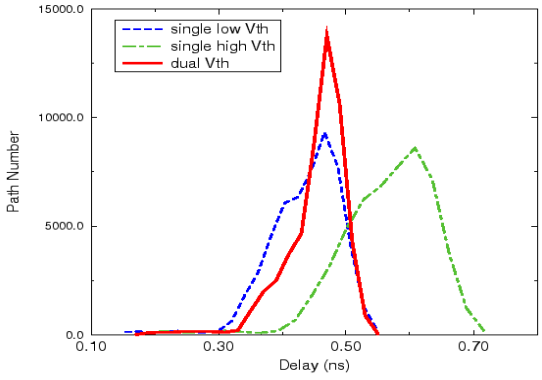


Leakage power saving can be more than 80% for some benchmark circuits

Implementation and Results

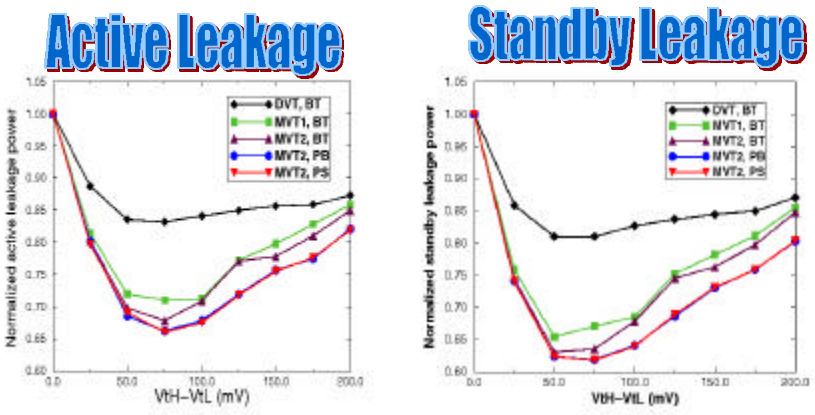
Path Distribution

- 32-bit Adder



(PathMill simulation)

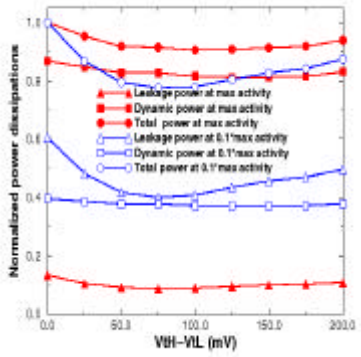
Leakage of 32-bit Adder



- **Mixed-V_{th} design technique can provide 20% more leakage savings than gate-level dual-V_{th} technique**

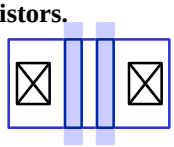
Total Power of 32-bit Adder

- Total power can be reduced by 9% for high activity
- Total power can be reduced by 22% at low activity



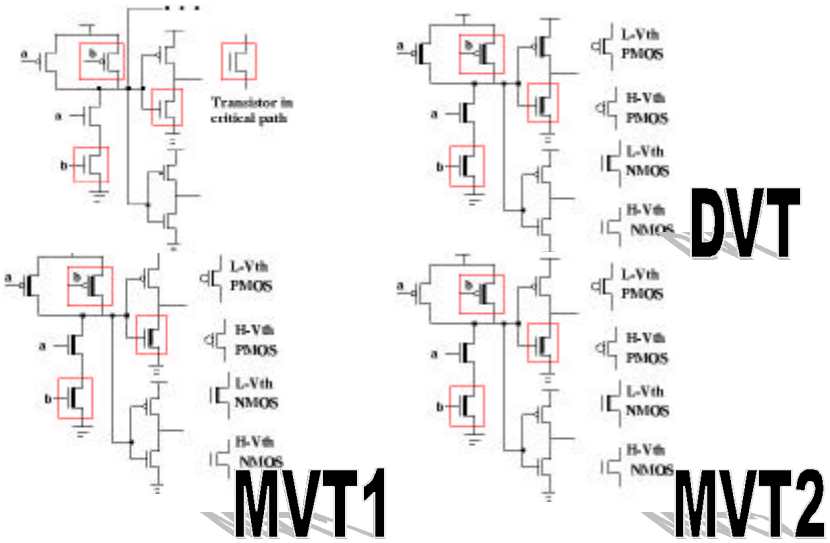
Mixed-V_{th} (MVT) CMOS Schemes

- Mixed-V_{th} (MVT) CMOS Schemes
  - **Scheme I (MVT1)**
    - **There is no mixed V_{th} in p pull-up or n pull-down trees.**
  - **Scheme II (MVT2)**
    - **Mixed-V_{th} is allowed anywhere except for the series connected transistors.**

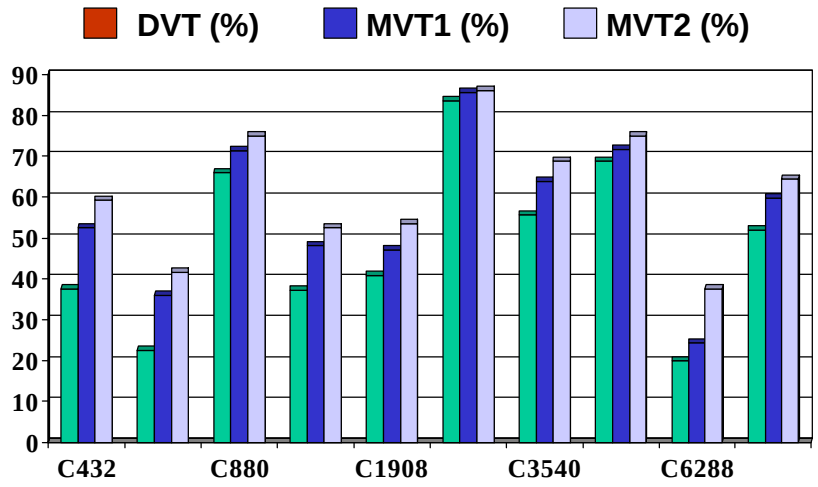


- **Due to the process variation, the worst case corner should be used in the analysis.**

Mixed-V_{th} (MVT) CMOS Schemes

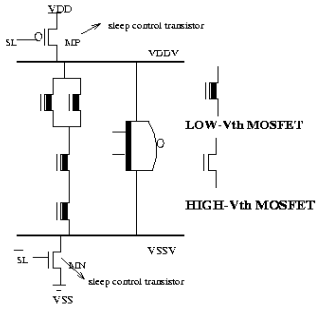


Leakage Power Saving for Different Circuit Schemes



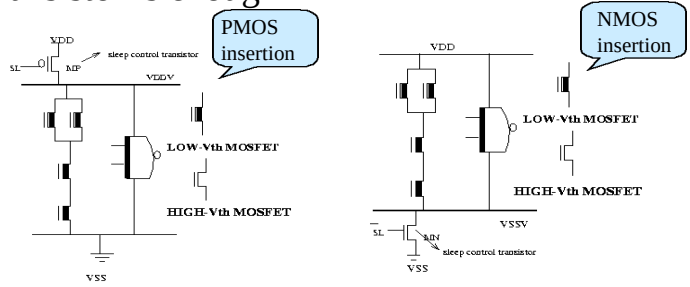
MTCMOS

- Multi-Threshold CMOS (From S. Mutoh, etc. JSSC 1995)
- In active mode:
  - SL=0, MP and MN are “on VDDV and VSSV almost function as VDD and VSS.
- In standby mode:
  - SL=1, MP and MN are “off leakage is suppressed.



MTCMOS (cont'd)

- Only one type of high-Vth sleep control transistor is enough



- NMOS size smaller

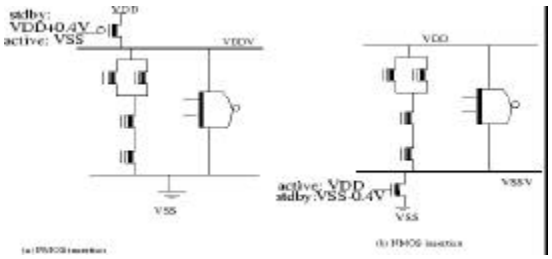
  ⇒ NMOS insertion is preferable

MTCMOS (cont'd)

- Advantage:
  - Effective for standby leakage reduction
  - Easily implemented based on existing circuits
  - 1-V MTCMOS DSP chip for mobile phone application (1996)
- Disadvantage:
  - Increase area and delay
  - If data retention is required in standby mode, an extra high-V_{th} memory circuit is needed

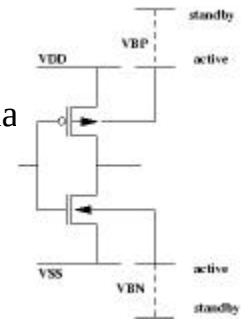
SCCMOS

- Super Cut-off CMOS (From H. Kawaguchi, ISSCC, 1998)
- Single-low-V_{th} circuit
  - Low-V_{th} sleep control transistor with smaller size
  - Minimal V_{dd} is lower than that of MTCMOS
- A gate bias generator is required



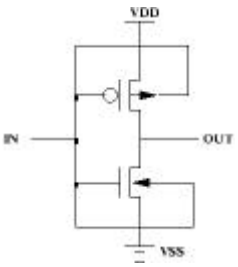
VTCMOS

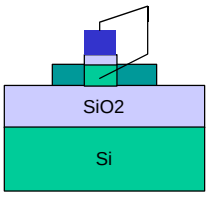
- Variable Threshold CMOS (from T. Kuroda, ISSCC, 1996)
- In active mode:
  - Zero or slightly forward body bia for high speed
- In standby mode:
  - Deep reverse body bias for low leakage
- Triple well technology required



DTMOS

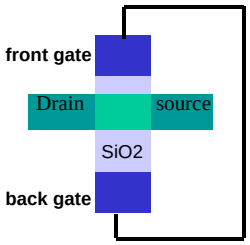
- Dynamic Threshold CMOS
  - from F. Assaderaghi, IEDM, 1994
- Vth altered dynamically to suit the operation state of the circuit
- Vdd<0.6V
- Triple well required for BULK silicon technology
- DTMOS in partially-depleted SOI





DGDT SOI CMOS

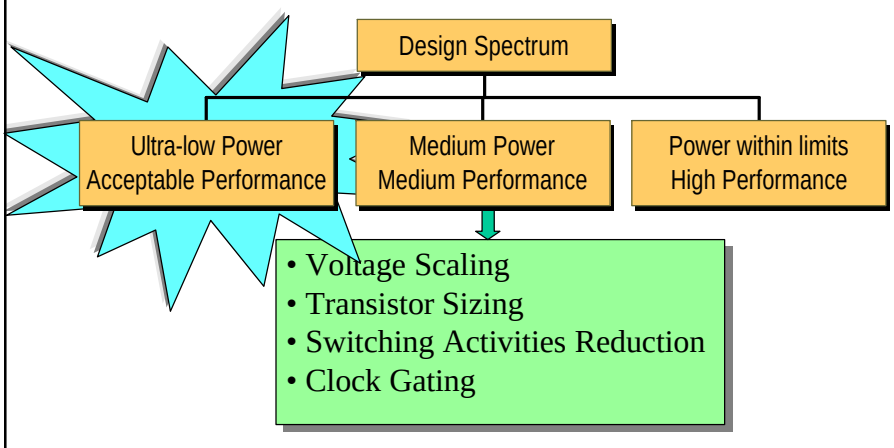
- Double Gate Dynamic Threshold SOI CMOS
  - from L.Wei, Z. Chen, K.Roy, IEEE SOI Conf., 1997
- Asymmetrical double gate fully-depleted SOI MOSFET
- Front gate: conducting gate
  Back gate: controlling gate



Conclusions

- Efficient leakage control technique required to maintain high performance in future designs
- For some niche applications (where performance is of secondary concern) ultra low power sub-threshold digital circuits can be used

Digital Sub-Threshold Logic

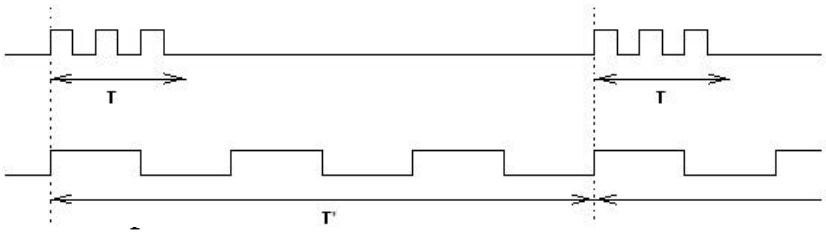


Digital Sub-threshold Logic

- Operate in sub-threshold region ($V_{gs} < V_{th}$)
- Uses leakage current as the operating switching current
- Suitable for ultra-low power applications where performance is of secondary importance
- Utilizes sub-threshold characteristics of MOS transistors

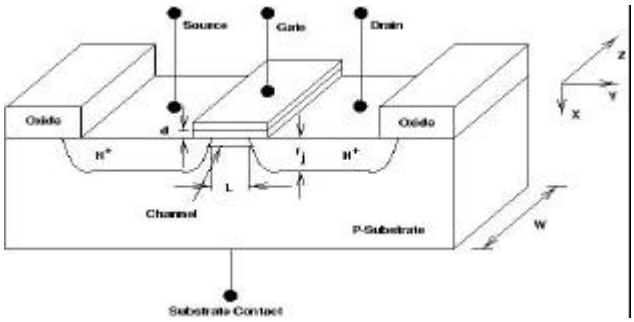
Application Areas

- Primary requirement: *Ultra-low Power*



- Bursty applications

Sub-threshold Characteristics of MOS Transistor



$$I_{ds} = m_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{\frac{(V_g - V_{th})}{mkT/q}}\left(1 - e^{-\frac{V_{ds}}{kT/q}}\right)$$

Sub-threshold Characteristics of MOS Transistor

$$I_{ds} = m_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{\frac{(V_g - V_{th})}{mkT/q}}\left(1 - e^{-\frac{V_{ds}}{kT/q}}\right)$$

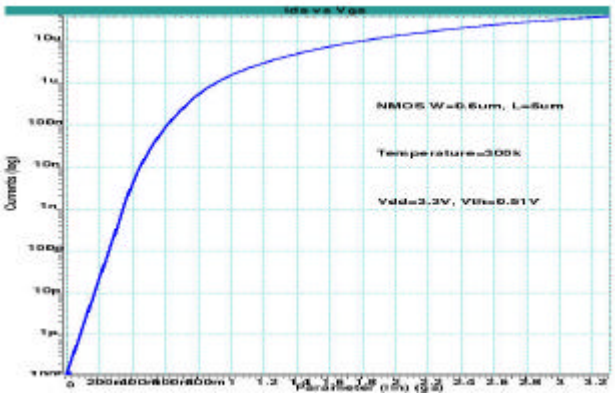
- m is body-effect coefficient i.e. $m = 1 + \frac{C_{depl}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{depl}}$
- C_{depl} is bulk depletion capacitance i.e. $C_{depl} = \frac{e_{Si}}{W_{depl}}$
- C_{ox} is oxide capacitance per unit area i.e. $C_{ox} = \frac{e_{ox}}{t_{ox}}$

Sub-threshold Characteristics of MOS Transistor

$$I_{ds} = m_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{\frac{(V_g - V_{th})}{mkT/q}}\left(1 - e^{-\frac{V_{ds}}{kT/q}}\right)$$

$$I_{ds} \propto e^{V_{gs}}$$

$$g_m \equiv \frac{\partial I_{ds}}{\partial V_{gs}}\Big|_{V_{ds}=const}$$



Threshold Voltage

$$V_{th} = V_{fb} + 2y_b + \frac{\sqrt{2e_{Si}qN_A(2y_b - V_{sb})}}{C_{ox}}$$

$$\Delta V_{th} = V_{th}(V_{sb}) - V_{th}(V_{sb}=0) = \frac{\sqrt{2e_{Si}qN_A}}{C_{ox}}\left(\sqrt{2y_b - V_{sb}} - \sqrt{2y_b}\right)$$

$$= ab\left(\sqrt{\frac{2y_b - V_{sb}}{b}} - \sqrt{\frac{2y_b}{b}}\right)$$

$$a \equiv \sqrt{2}\left(\frac{e_{Si}}{L_D}\right)\Big/ C_{ox}$$

- V_{fb} is the flat-band voltage
- $y_b = V_{fermi} - V_{intrinsic\ of\ bulk\ Si}$
- e_{Si} is the permittivity of Silicon
- $C_{ox} = e_{ox}/t_{ox}$ i.e. oxide capacitance/unit area
- b = kT/q i.e. thermal voltage
- $L_D = (be_{Si}/(qN_A))^{1/2}$ i.e. extrinsic Debye length

Temperature Effect on V_{th}

$$V_{th} = V_{fb} + 2y_b + \frac{\sqrt{2e_{Si}qN_A(2y_b - V_{sb})}}{C_{ox}}$$

$$\frac{dV_{th}}{dT} = \frac{dy_b}{dT}\left(2 + \frac{1}{C_{ox}}\sqrt{\frac{e_{Si}qN_A}{y_b}}\right)$$

$$\frac{dy_b}{dT} = \pm\frac{1}{T}\left(\frac{E_g(T=0)}{2q} - |y_b(T)|\right)$$

E_g is the band-gap energy

Sub-threshold Slope (S)

Gate voltage swing needed to change the drain current by one decade

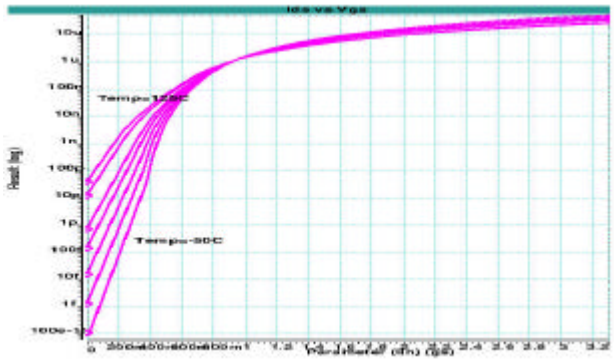
$$S \equiv \frac{dV_{gs}}{d(\ln I_{ds})}\ln 10 = b\,\frac{d\left(V_{gs}/b\right)}{d(\ln I_{ds})}\ln 10 = b\left[1 + \frac{C_{depl}}{C_{ox}}\right]\left[1 - \left(\frac{2}{a^2}\right)\left(\frac{C_{depl}}{C_{ox}}\right)^2\right]\ln 10$$

For $a >> \left(\frac{C_{depl}}{C_{ox}}\right)$: $\quad S \approx \frac{kT}{q}\left[1 + \frac{C_{depl}}{C_{ox}}\right]\ln 10$

C_{depl} is the depletion layer capacitance

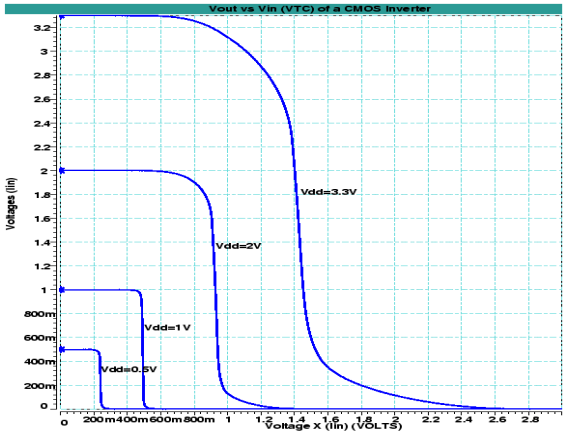
Temperature Variation

$$I_{ds} = m_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{\frac{(V_g - V_{th})}{mkT/q}}\left(1 - e^{-\frac{V_{ds}}{kT/q}}\right)$$

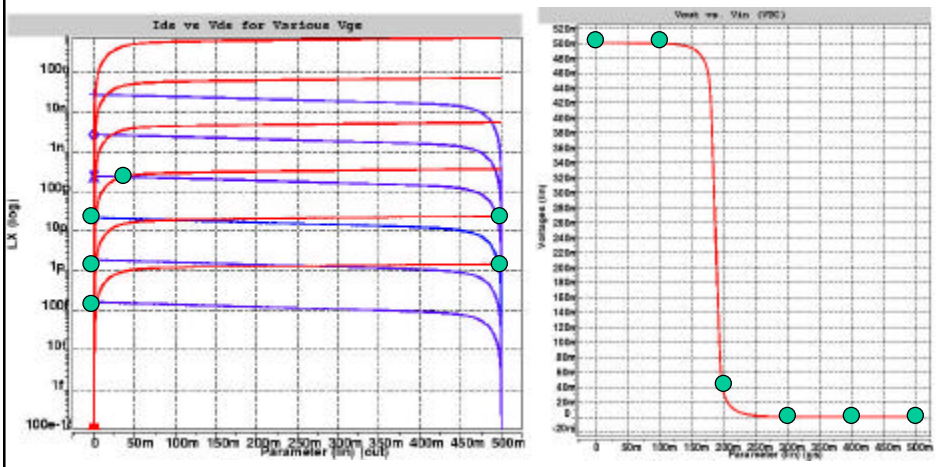


Sub-threshold Static Logic: Sub-CMOS Logic

- CMOS logic operates in sub-threshold region
- $V_{dd} < V_{th}$
- Near ideal VTC
- High gain
- High g_m
- Better NM

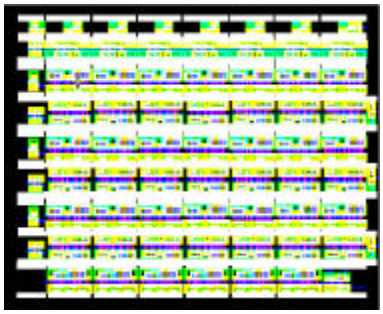


Sub-CMOS Logic



Simulation Results

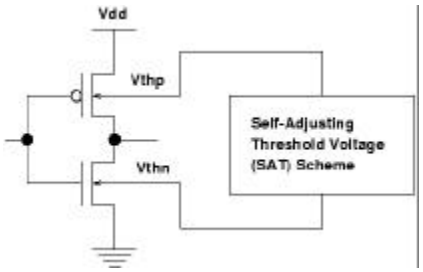
- Logic Gates (SPICE using 0.35m process)
- 8x8 Carry-Save Array Multiplier (0.35um)



377.4um x 279.1um

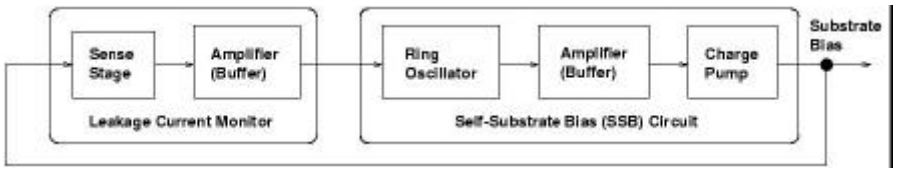
Inversion	Power(W)	Delay(s)	PDP(J)
Strong	9.27e-3	1.883e-9	17.46e-12
Weak	30.1e-9	21.38e-6	0.644e-12

Robust Sub-Static Logic: Sub-VT-CMOS logic



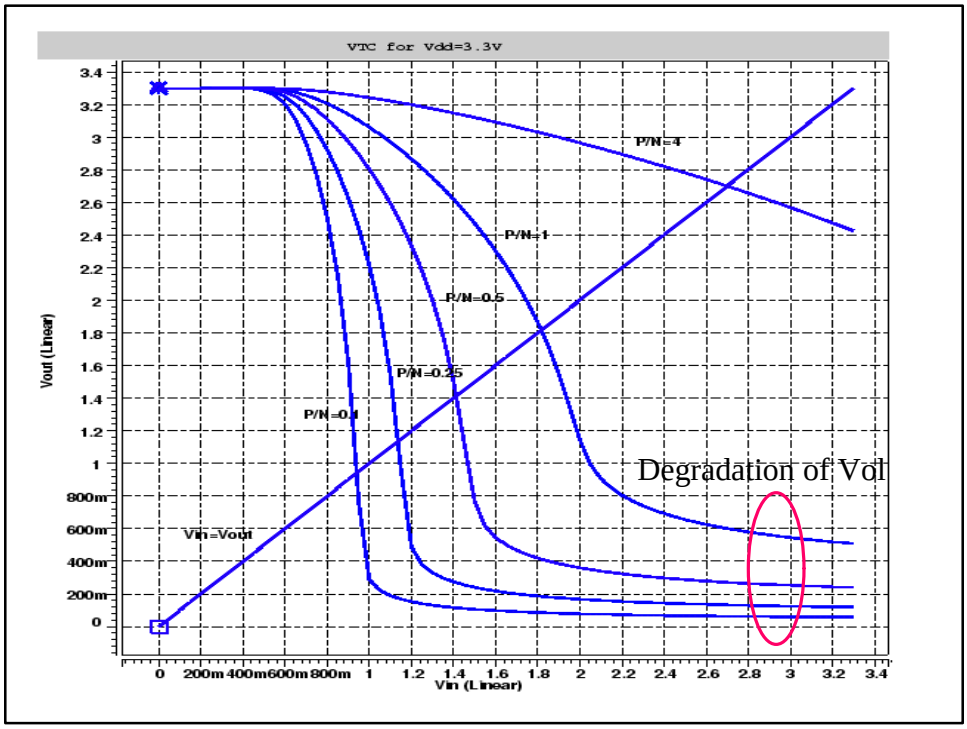
- Negative feedback principle:
  - Monitors any change in leakage current (due to process and temperature variations)
  - Stabilizes the circuit by applying appropriate substrate bias

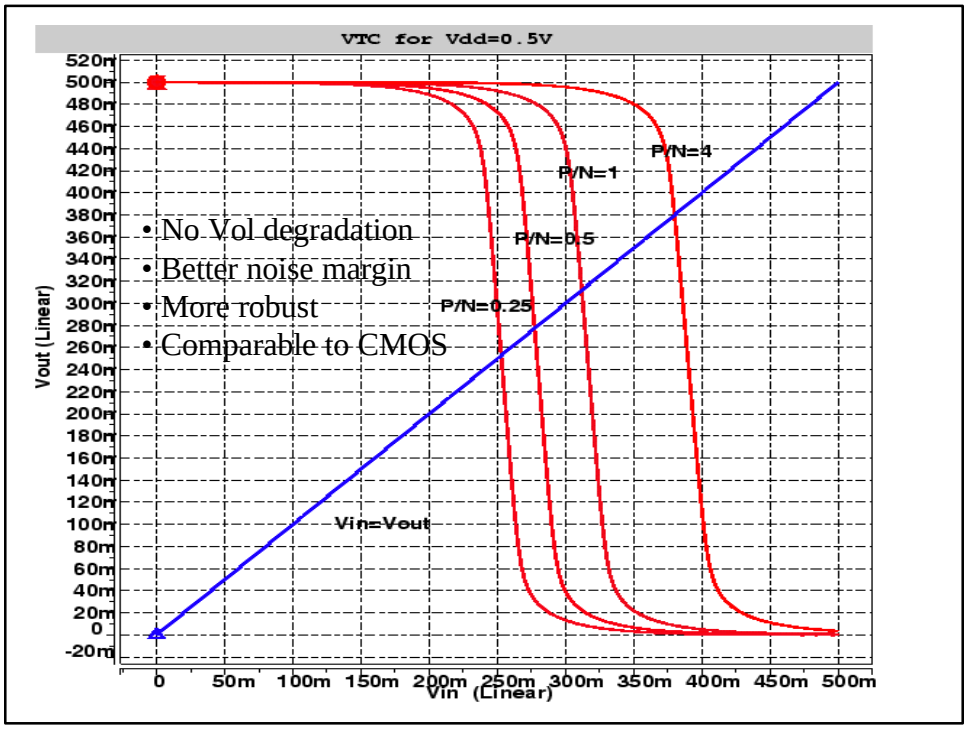
Sub-VT-CMOS Logic



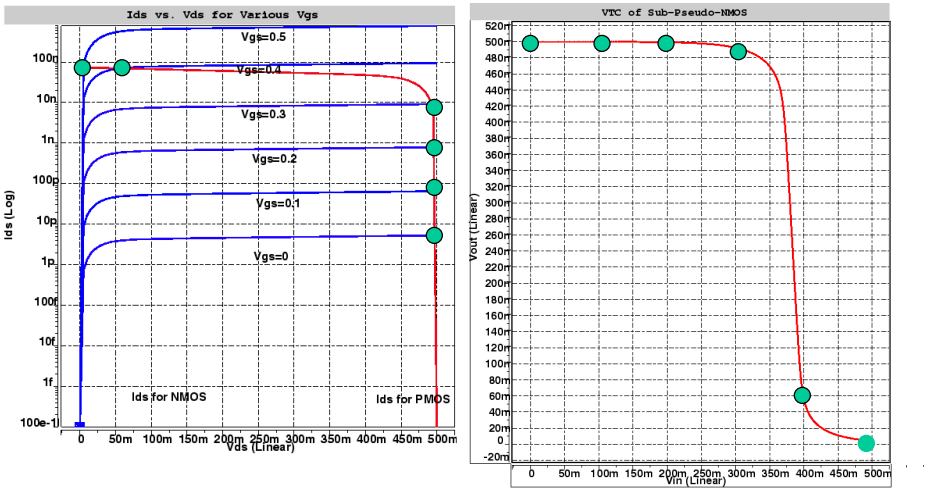
Two components of stabilization scheme:

- Leakage Current Monitor (LCM):
  - Leakage current sensor
  - Activates/De-activates the SSB circuit
- Self-Substrate Bias (SSB) circuit:
  - Uses charge-pump to accumulate charge
  - Applies bias to the substrate


VTC for Vdd=3.3V
Vout (Linear)
Vin (Linear)
P/N=4
P/N=1
P/N=0.5
P/N=0.25
P/N=0.1
Vin=Vout
Degradation of Vol


VTC for Vdd=0.5V
Vout (Linear)
Vin (Linear)
P/N=4
P/N=1
P/N=0.5
P/N=0.25
Vin=Vout
• No Vol degradation
• Better noise margin
• More robust
• Comparable to CMOS

VTC of Sub-Pseudo-NMOS



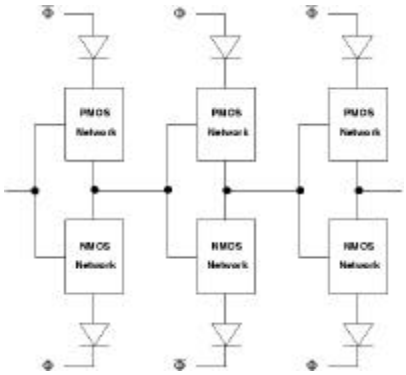
Simulation Results

- 8x8 carry-save array multipliers
- TSMC 0.35um process technology
- 50kHz with Vdd=0.5V and temp=55ºC

	Area(um^2)	Power(uW)
CMOS	$228.6x10^3$	11.59
Sub-CMOS	$228.6x10^3$	0.163
Sub-Pseudo-NMOS	$210.9x10^3$	1.056
Sub-True-NMOS	$202.7x10^3$	1.497

Comparison with other known ultra-low power logic: Energy-Recovery Logic

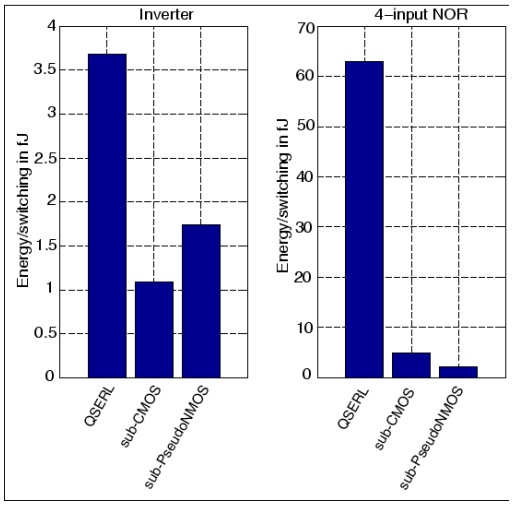
- To validate the power saving effectiveness of sub-threshold logic
- QSERL (Quasi-Static Energy-Recovery Logic): close resemblance to CMOS logic



• Sinusoidal power supply generator is assumed to have 100% efficiency (In practice, only 80-90% efficiency is achieved)

• Use ideal Schottky diodes instead of diode-connected, low Vth MOS transistors (leaky transistors make it difficult for QSERL to hold the output voltage properly during long hold time)

Comparison Results



Conclusions

- Digital sub-threshold logic is used to satisfy the ultra-low power requirement
- Sub-threshold logic is readily implemented and derived from the traditional existing circuits by lowering the V_{dd} to be less than V_{th}
- Improved characteristics including higher gain, better noise margin, and more energy efficient
- Ratio-ed logic (Pseudo/True-NMOS) is comparable to CMOS logic in sub-threshold logic in terms of robustness, noise margin and power consumption
- Sub-dynamic logic is faster, smaller and has better noise margin than sub-CMOS
- Sub-threshold logic is easier to design and more efficient as compared to other known ultra-low power logic, such as energy-recovery logic