

Design of CMOS Circuits in the Nano-meter Regime: **Leakage Tolerance**

Kaushik Roy

Professor of Electrical & Computer Engineering

Purdue University

Challenges ahead ...

in Si nanometer regime

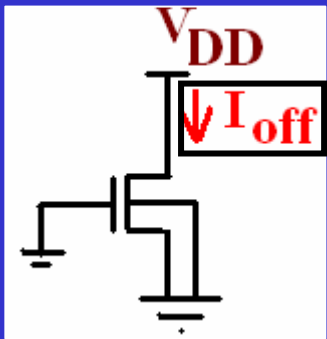
Scaling & Ion/Ioff



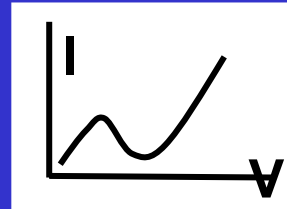
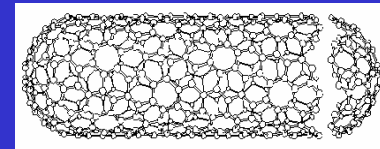
**Silicon micro
electronics**

**Silicon nano
electronics**

**Non-Silicon
technology**



- Increasing leakage
- Increasing process variations
- Short Channel Effects



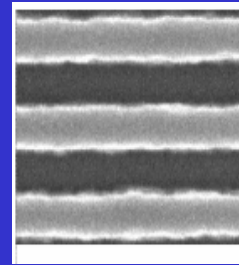
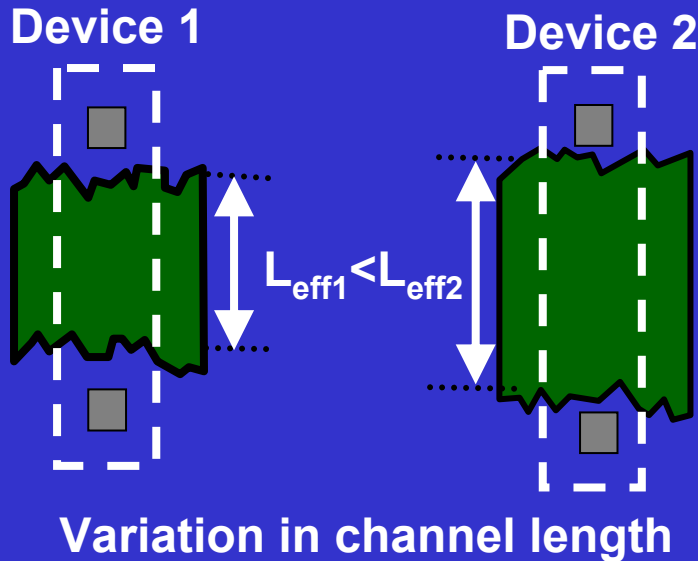
- Carbon Nanotubes
- Molecular transistors
- Molecular RTDs

$$\frac{I_{ON}}{I_{OFF}} = 10^6$$

$$\frac{I_{ON}}{I_{OFF}} = 10^3$$

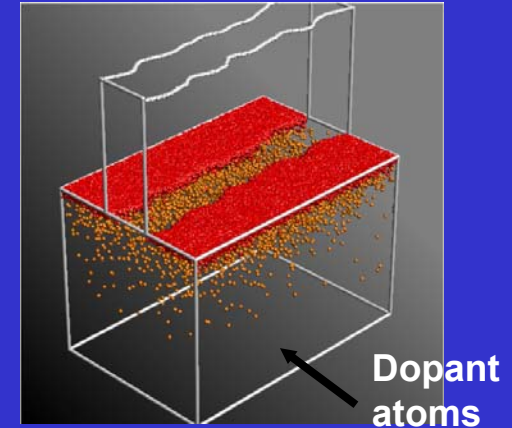
$$\frac{I_{ON}}{I_{OFF}} = 10^4$$

Process Variations



A. Asenov, TED03

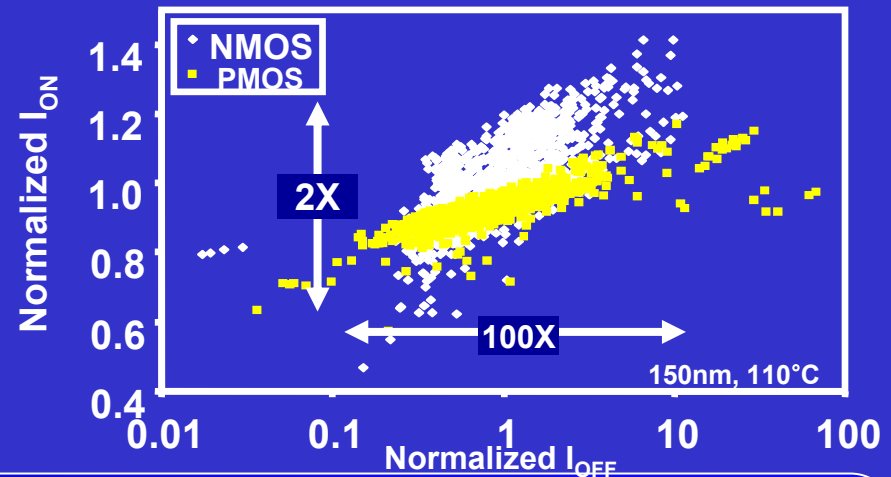
Line-Edge
Roughness



M. Hane, et. al., SISPAD 2003

Random Dopant Fluctuations
(RDF)

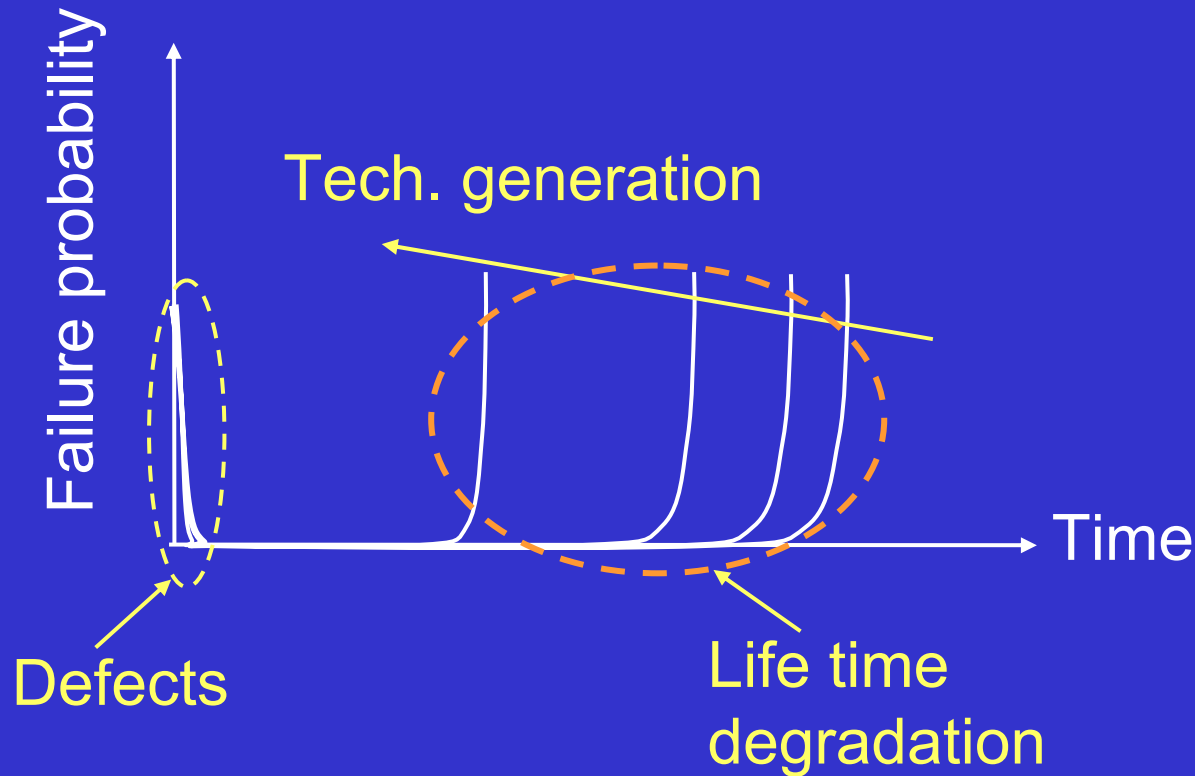
- Intrinsic parameter variations:
 - Channel length and width
 - Variations due to line edge roughness
 - Threshold voltage (V_t) variations due to random dopant fluctuation



Device parameters are no longer deterministic

Reliability

Temporal degradation of performance -- NBTI



Power Consumption

- Leakage Power
 - Subthreshold, Gate, Junction, GIDL, Punchthrough,
- Dynamic Power
 - Due to charging/discharging of capacitive load
 - Short-circuit power due to direct path currents when there is a temporary connection between power and ground

Switching/Dynamic Power

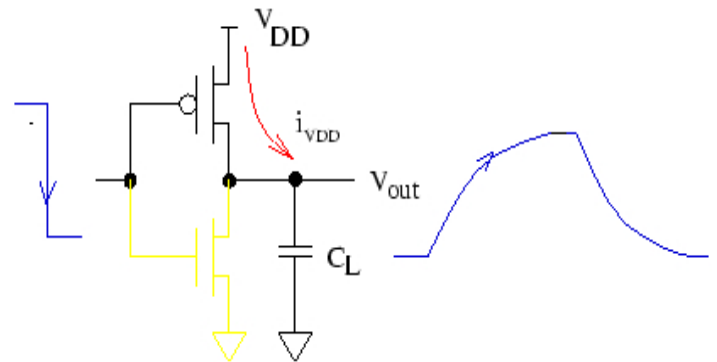
Switching Power

- Signal properties
 - Signal probability, P_i , - probability of a signal being logic ONE
 - Signal activity, a_i , - probability of signal switching(0->1, or 1->0)
- Energy dissipated per transition

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

$$= C_L V_{DD} \int_0^{V_{DD}} dv_{out} = \boxed{C_L V_{DD}^2}$$

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

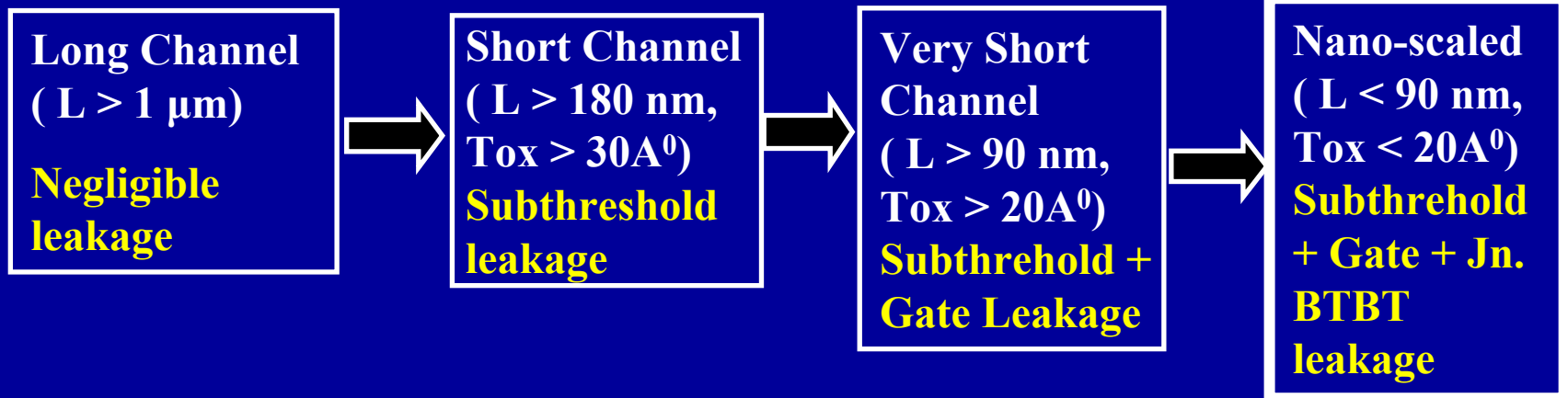
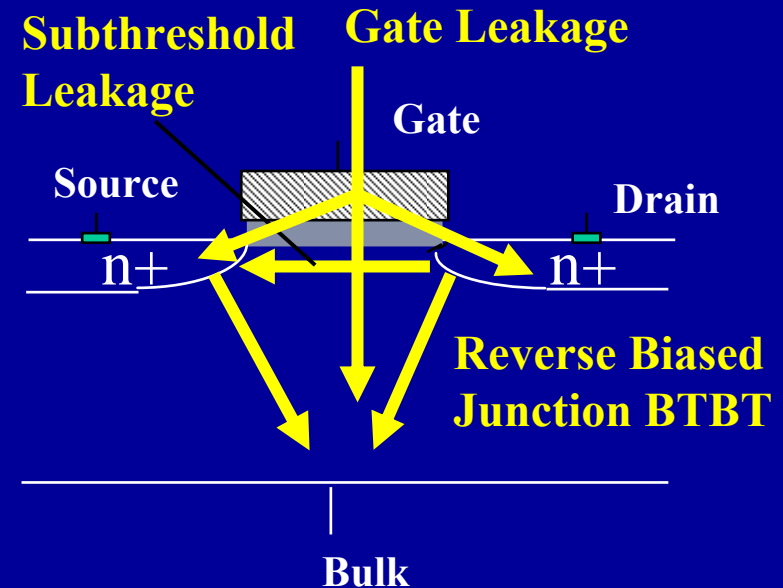


Energy dissipated for 1->0 or 0->1 transition: $C_L V_{DD}^2 / 2$

Leakage Power

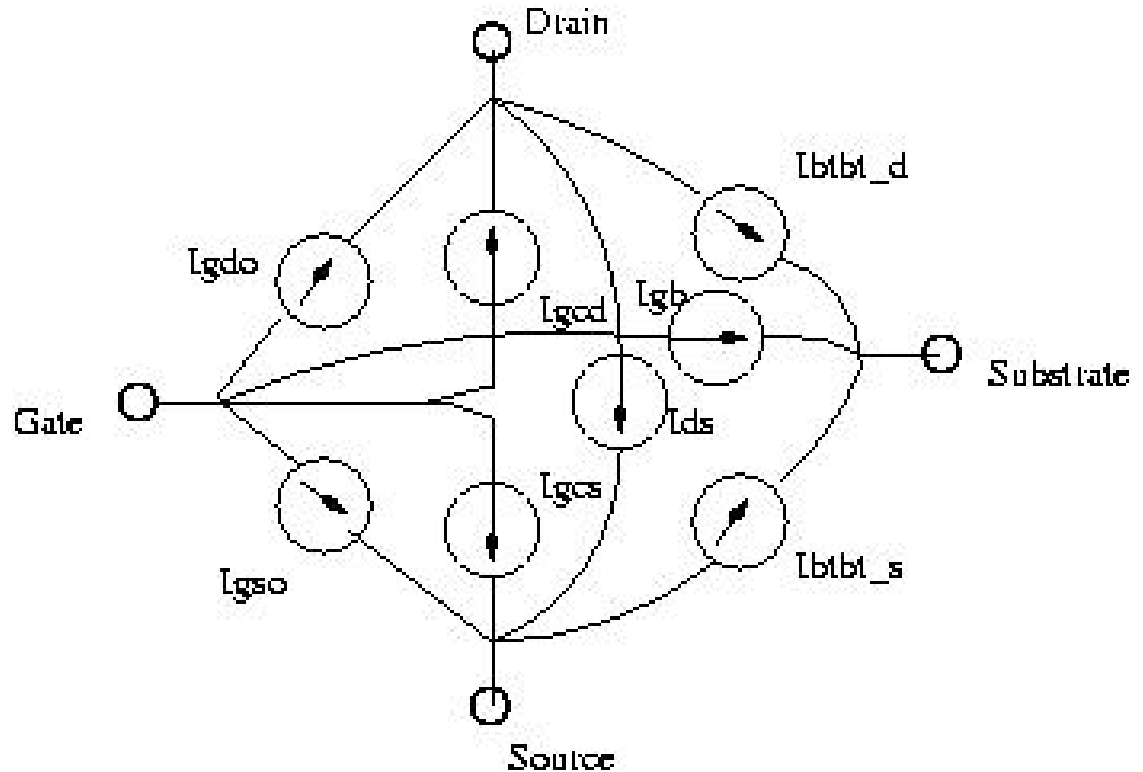
Scaling and Other Leakage Components

- Leakage Components
 - Subthreshold Leakage
 - Gate Leakage
 - Reverse-biased Junction Band-To-Band-Tunneling (BTBT) Leakage.
 - Others



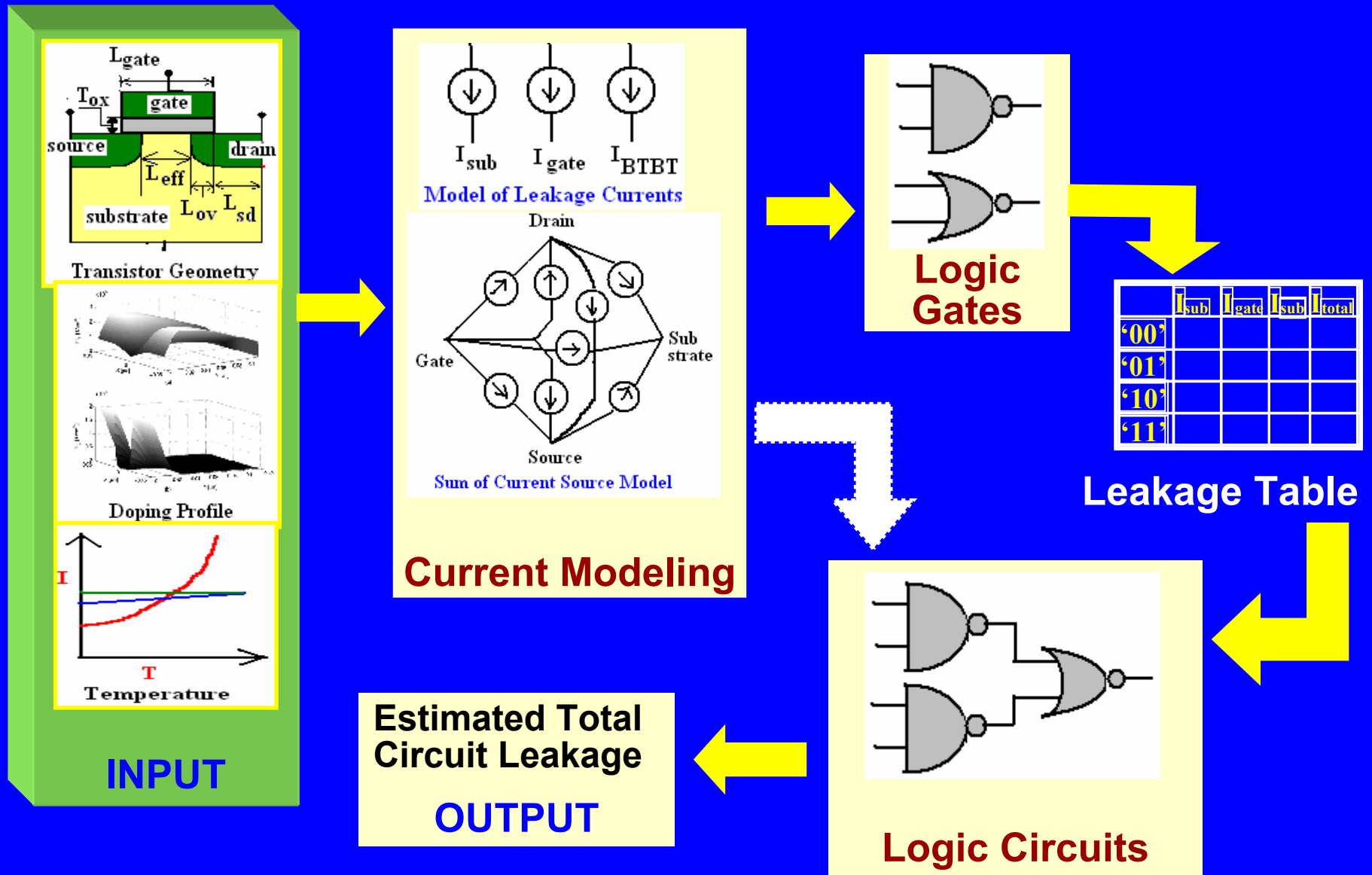
Total Leakage

“Sum of Current Source Model”
Voltage Controlled Current Sources describing each leakage comp.



$$\text{Total Transistor Leakage} = I_{\text{overall}} = I_{BTBT} + I_{\text{sub}} + I_{\text{gate}}$$

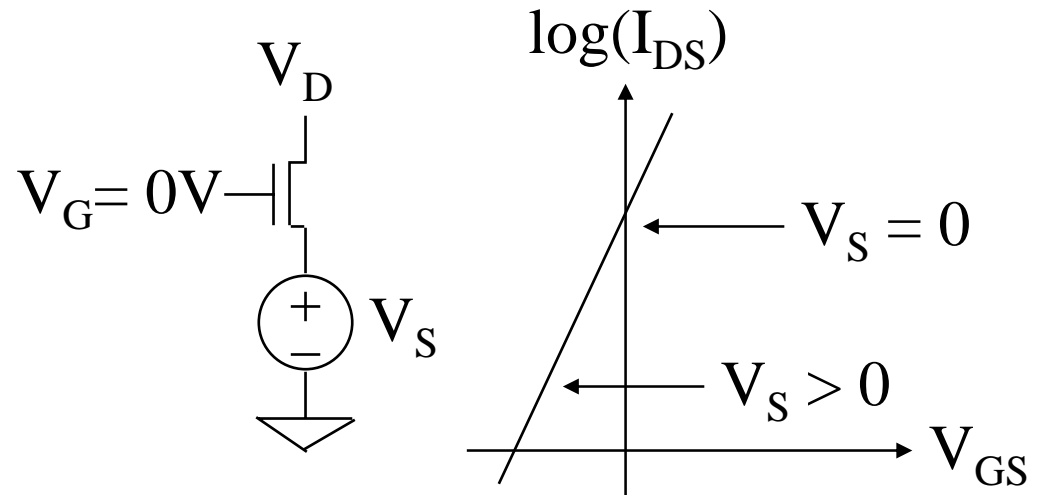
Leakage Estimation Method



Leakage Reduction: Logic & Memory

Self-Reverse Bias (Source-Biasing, Supply-Gating, Stacking)

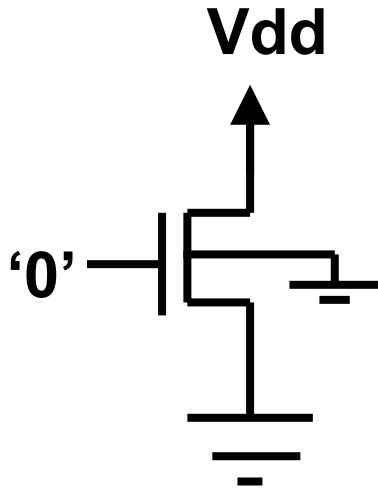
- Primary effect:
 - $V_{GS} < 0$
 - move down subthreshold slope
- Secondary effects:
 - Drain Induced Barrier Lowering
 - Body effect



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

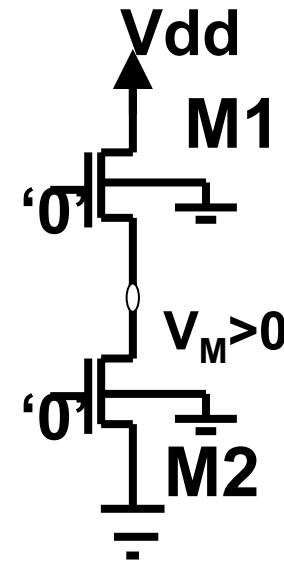
$$V_S \uparrow \Rightarrow V_T \uparrow$$

Leakage Control: Stacking



$V_{gs}=0, V_{bs}=0, V_{ds}=V_{dd}$

- ✓ Negative V_{gs} ,
 - ✓ Negative V_{bs} - More Body effect,
 - ✓ Reduced V_{ds} -Less DIBL
- 2-T stack has lower subthreshold leakage



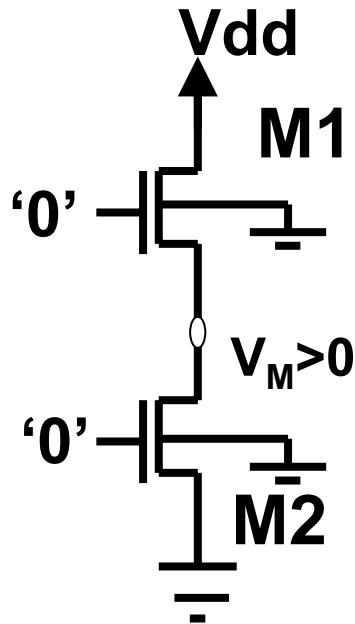
For M1:

$V_{gs} = -V_M < 0, V_{bs} = -V_M < 0,$
 $V_{ds} = V_{dd} - V_M < V_{dd}$

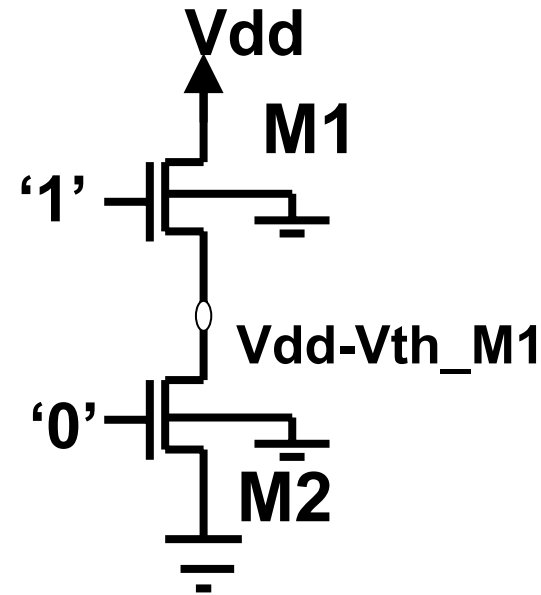
For M2:

$V_{gs} = 0, V_{bs} = 0,$
 $V_{ds} = V_M < V_{dd}$

Input Vector Control - Subthreshold



Minimum V_{gs} is For M1:
 $V_{gs_M1} < 0$,
 $V_{ds_M1} = V_{dd} - V_M$



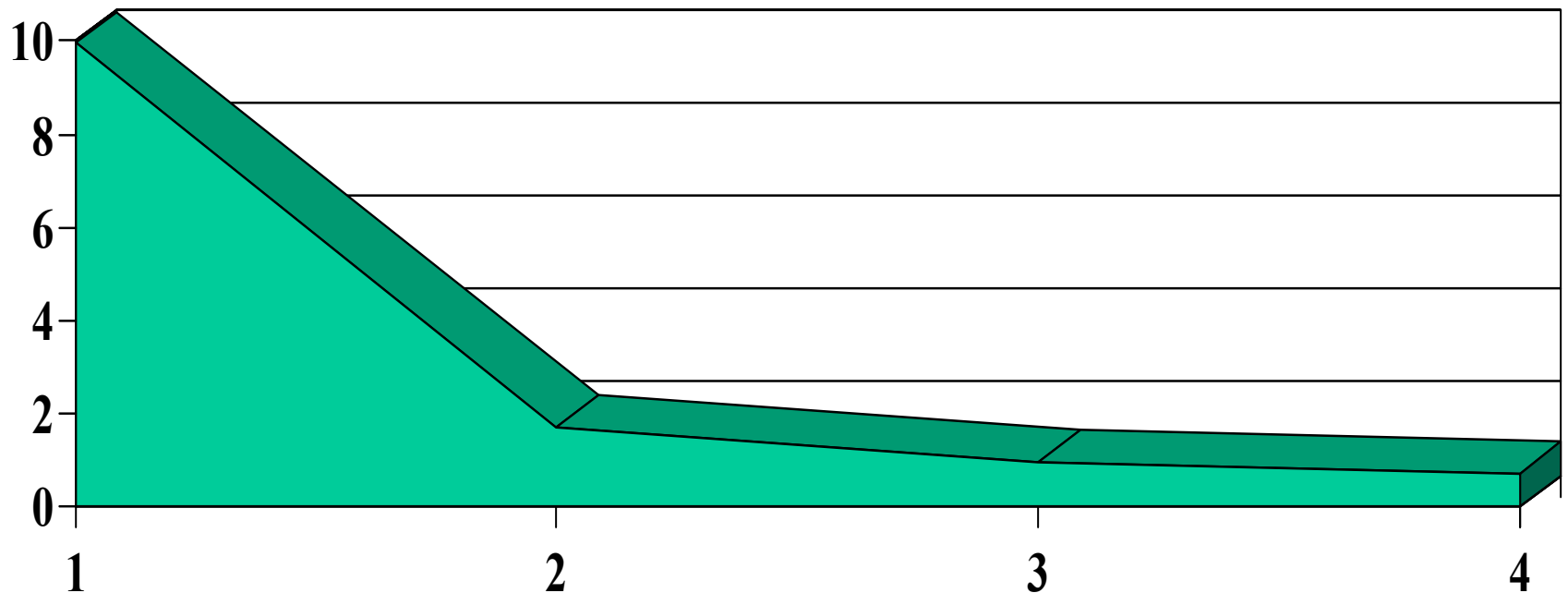
Minimum V_{gs} is For M2:
 $V_{gs_M2} = 0$,
 $V_{ds_M2} = V_{dd} - V_{th_M1}$

'00' gives minimum subthreshold leakage.

Turn 'off' maximum number of transistors in a stack to reduce subthreshold leakage

Leakage vs. Transistors Off

Leakage [nA]



Number of transistors off in stack

Input Vector Control – Gate Leakage

✓ $V_g = '0'$ – EDT dominates

➤ $I_g = I_{gdo} + I_{gso}$

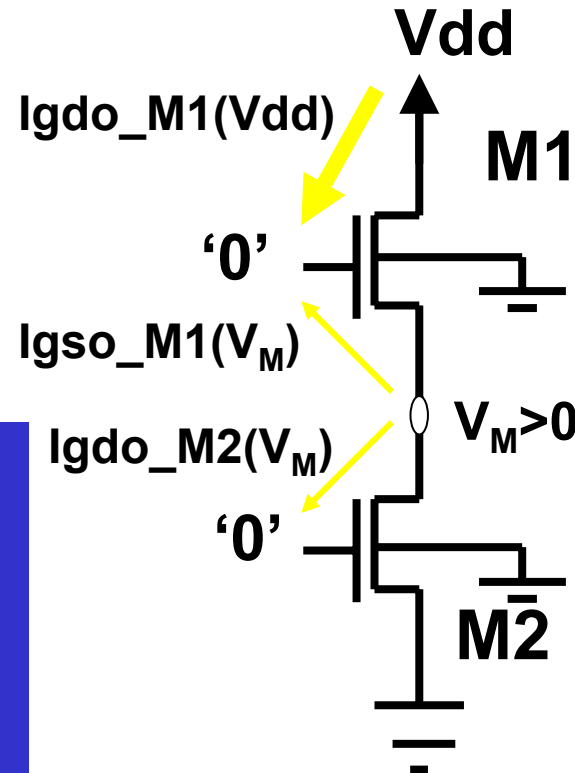
✓ $V_g = '1'$ – Gate to Channel tunneling is significant

➤ $I_g = I_{gdo} + I_{gso} + I_{gc}$

With '00' –

$I_{gdo_M1}(V_{dd}) \gg I_{gso_M1}(V_M) + I_{gdo_M2}(V_M)$

I_{gdo} of M1 dominates the total gate current



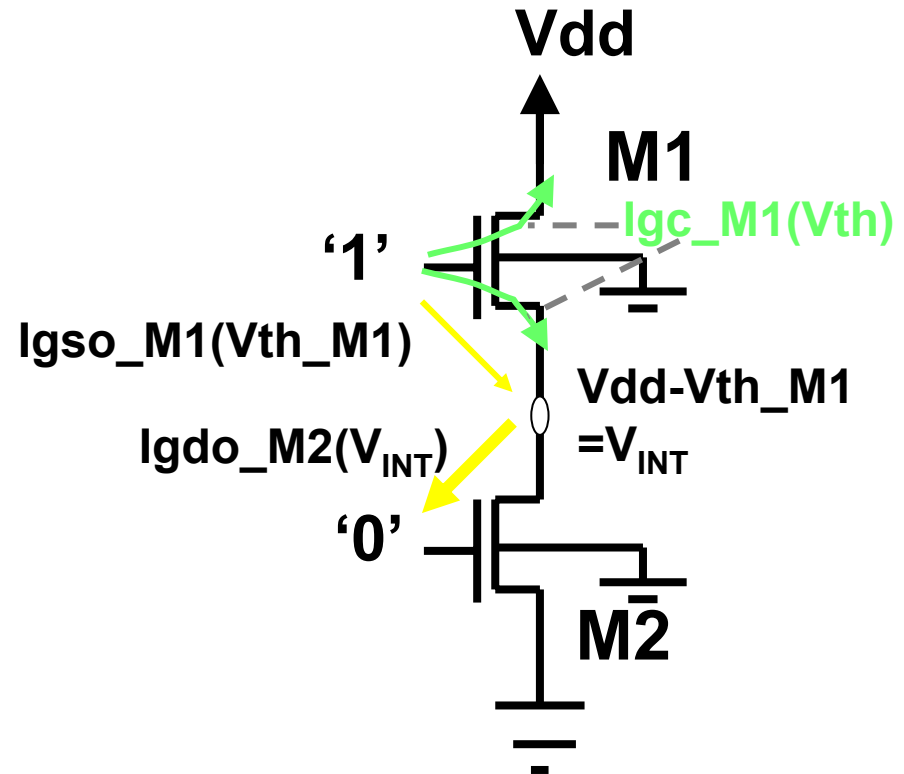
$$I_{gstack} = WL_{SDE} A (V_{dd} / T_{ox})^2 \exp \left(\frac{-B \left(1 - (1 - V_{dd} / \phi_{ox})^{3/2} \right)}{V_{dd} / T_{ox}} \right)$$

Input Vector Control – Gate Leakage

With '10' the major gate currents are:

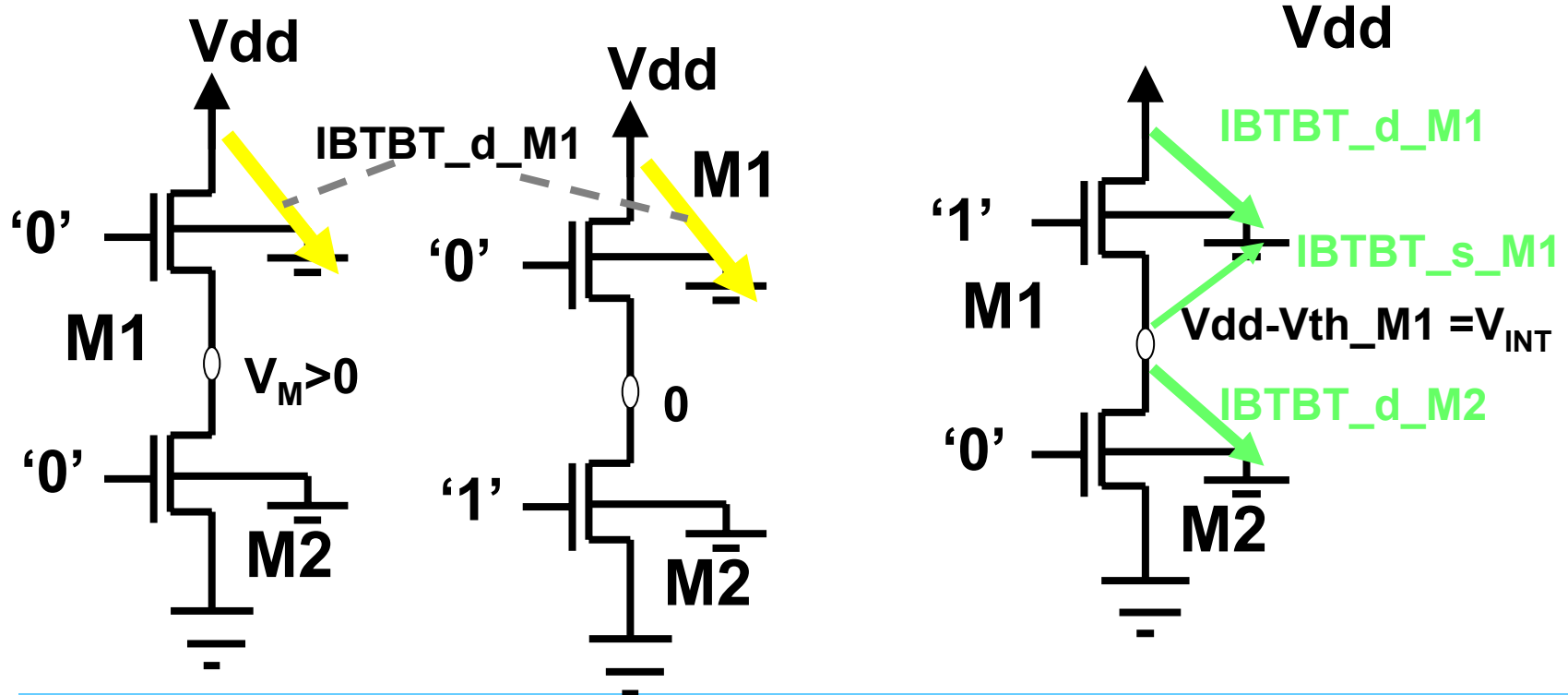
- ✓ $I_{gso_M1}(V_{th})$
- ✓ $I_{gdo_M2}(V_{dd} - V_{th_M1})$
- ✓ $I_{gc_M1}(V_{gs} = V_{th})$

I_{gdo_M2} dominates the total current.



$$I_{gstack} = WL_{SDE}A \left(\frac{(V_{dd} - V_{th_M1})}{T_{ox}} \right)^2 \exp \left(\frac{-B \left(1 - \left(1 - (V_{dd} - V_{th_M1}) / \phi_{ox} \right)^{3/2} \right)}{(V_{dd} - V_{th_M1}) / T_{ox}} \right)$$

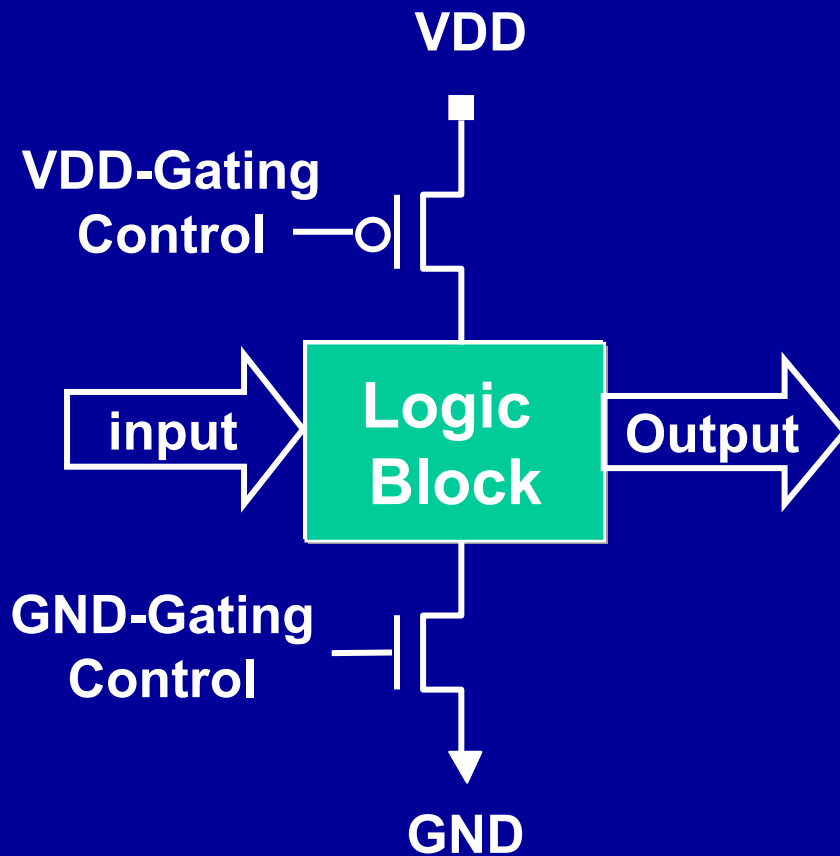
Input Vector Control – BTBT



'00' and '01' –drain-substrate BTBT of M1 dominates.
'10' – additional BTBT components drain-substrate of M2 and source-substrate of M1.

'10' gives maximum BTBT. However, BTBT is not very sensitive to stacking.

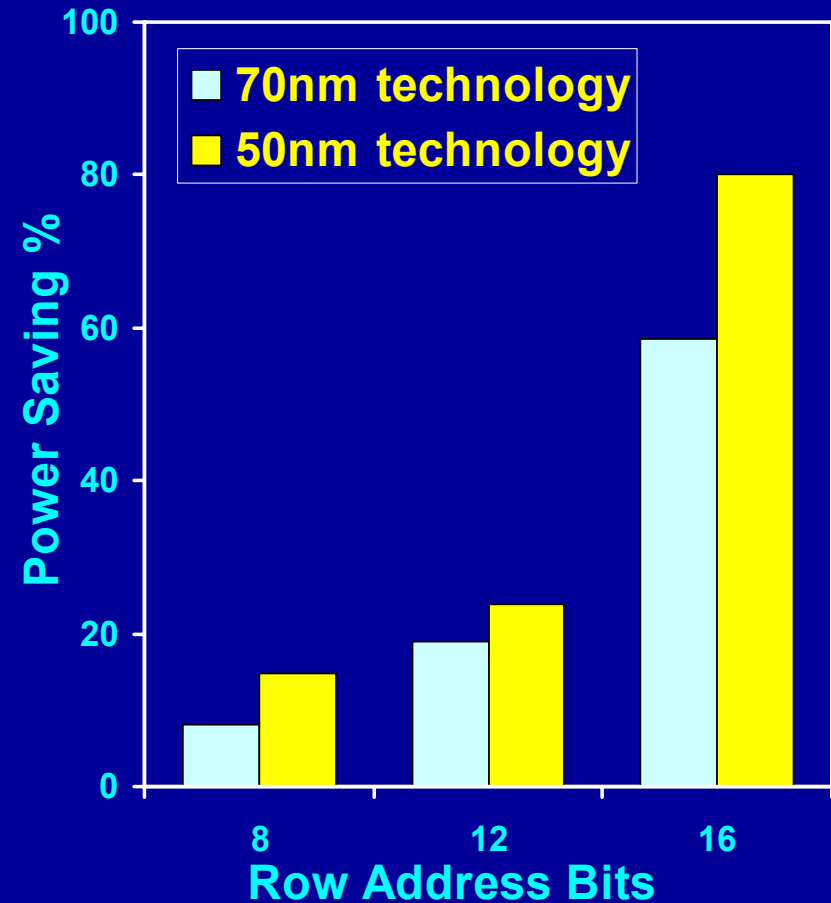
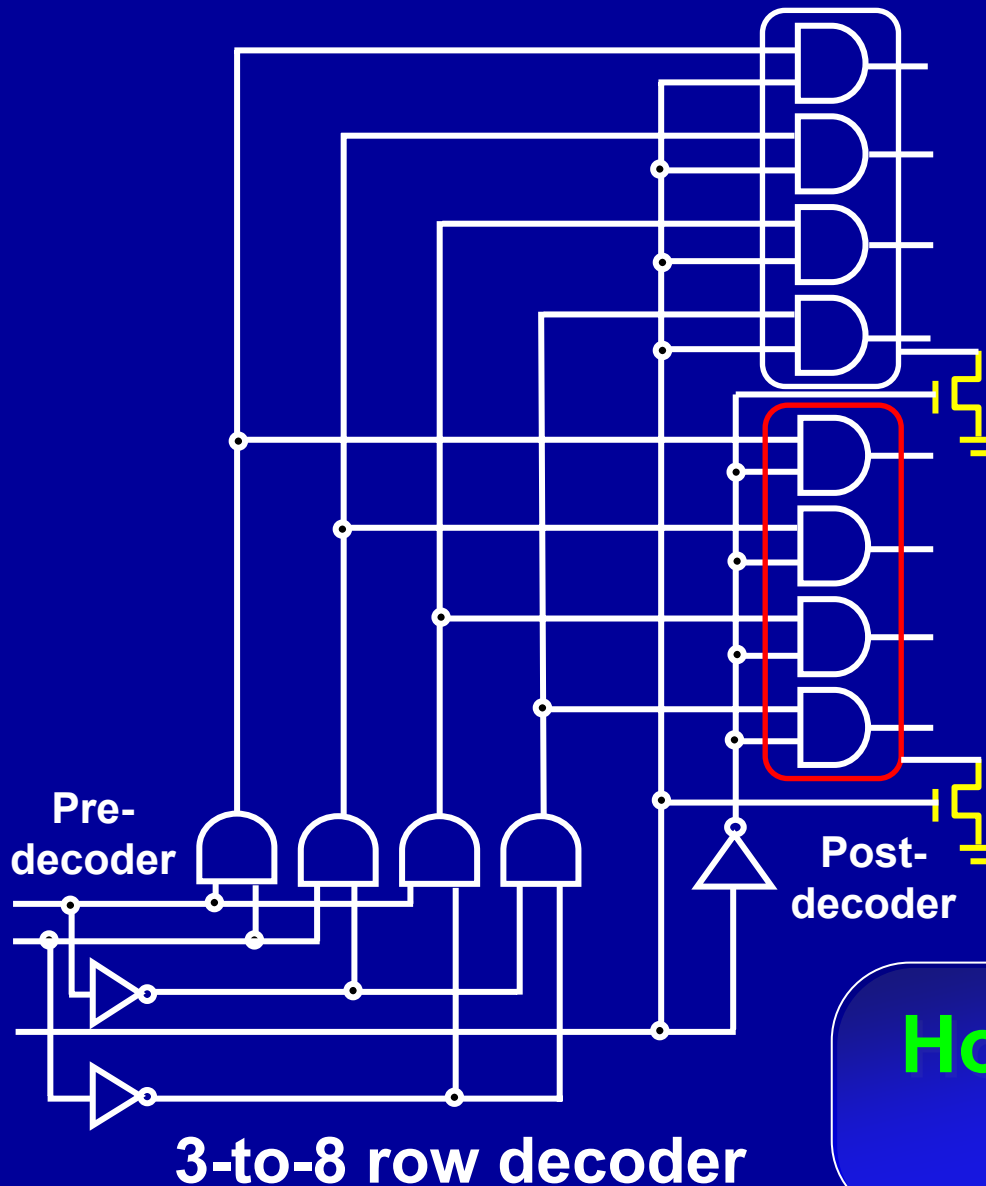
Supply Gating for Logic



Pros	Cons
5-20X Leakage Reduction	Delay/Area Overhead
Scalable	Floated Output
Design ease	Can be applied to idle sections only

How to use supply gating dynamically in active mode?

Dynamic Supply Gating (DSG): An Example



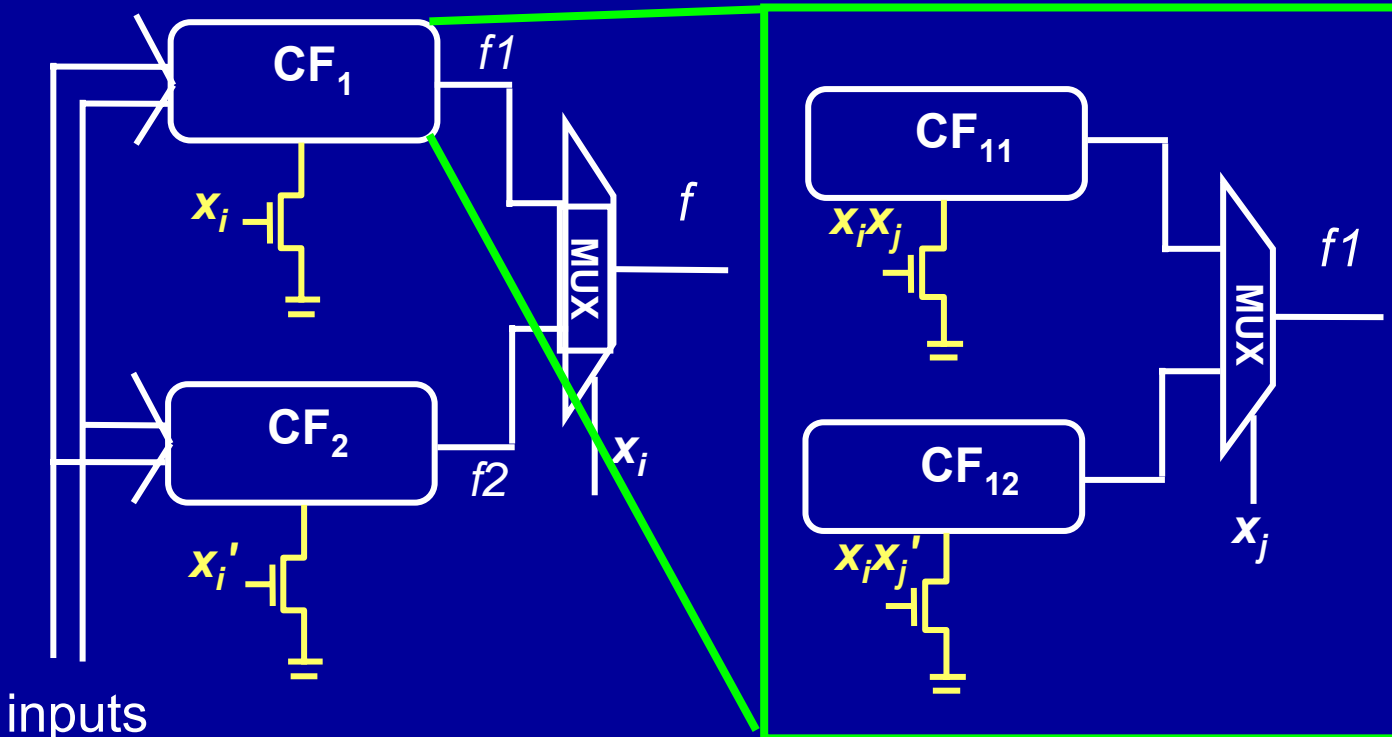
How to do it for random logic?

Dynamic Supply Gating for General Circuits

■ Shannon's expansion:

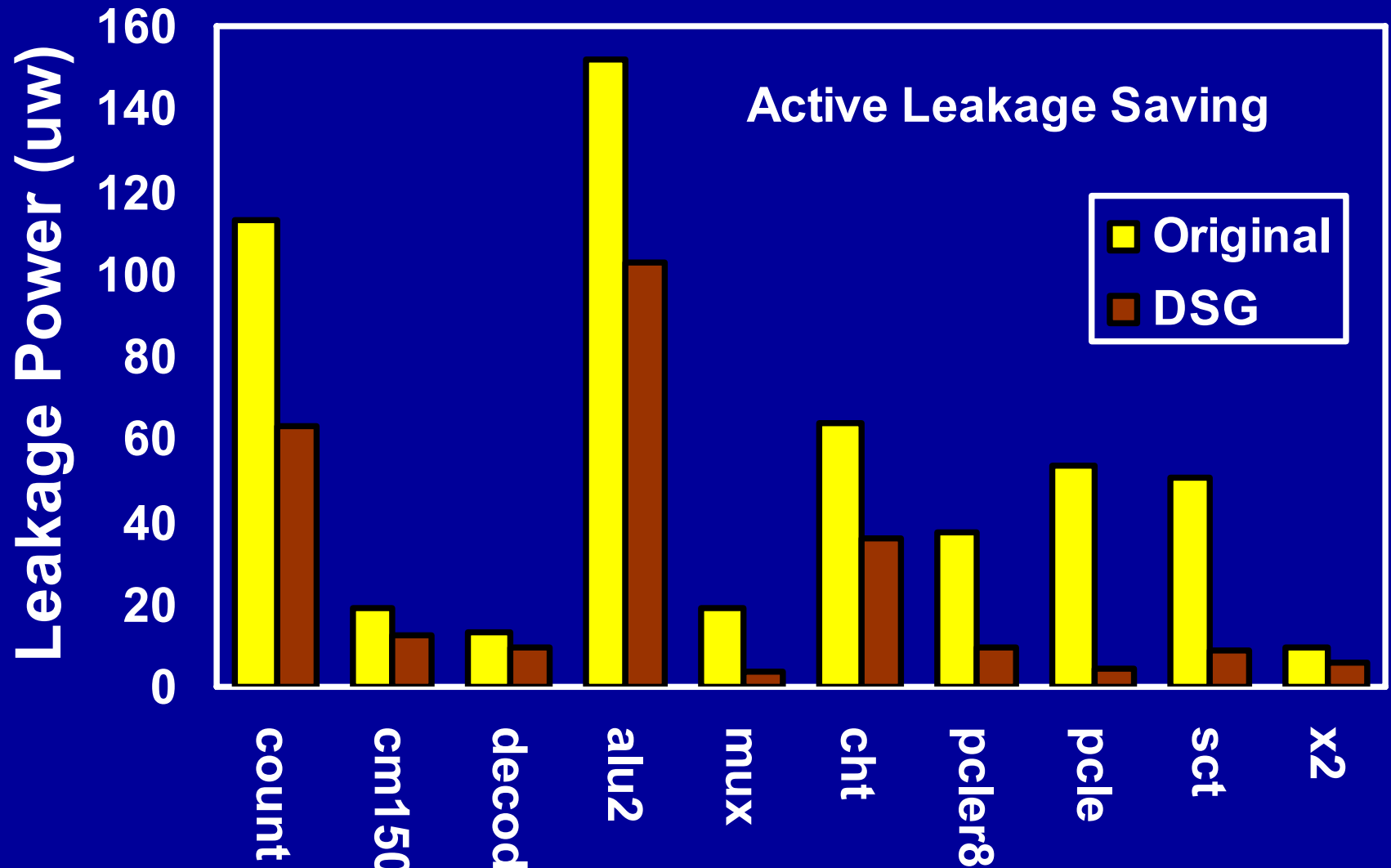
$$\begin{aligned} f(x_1, \dots, x_i, \dots, x_n) &= x_i \cdot f(x_1, \dots, x_i = 1, \dots, x_n) + x_i' \cdot f(x_1, \dots, x_i = 0, \dots, x_n) \\ &= x_i \cdot CF_1 + x_i' \cdot CF_2 \\ CF_1 &= f(x_1, \dots, x_i = 1, \dots, x_n); \quad CF_2 = f(x_1, \dots, x_i = 0, \dots, x_n) \end{aligned}$$

x_i is referred as **Control Variable**



Control variable selection is important

Simulation Results



MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=100°C

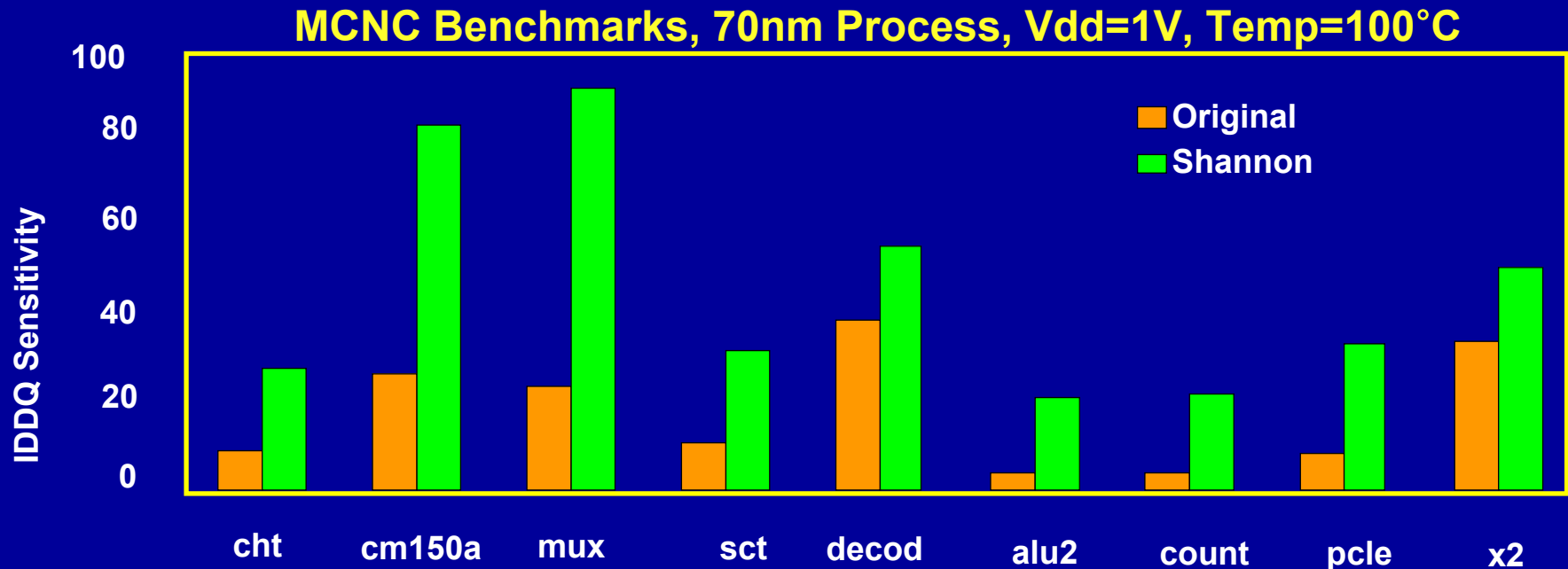
Supply-Gating & Test

Improvement in IDDQ Sensitivity

$$\text{IDDQ Sensitivity (S)} = (I_f - I_g) / I_g$$

I_f = Faulty IDDQ

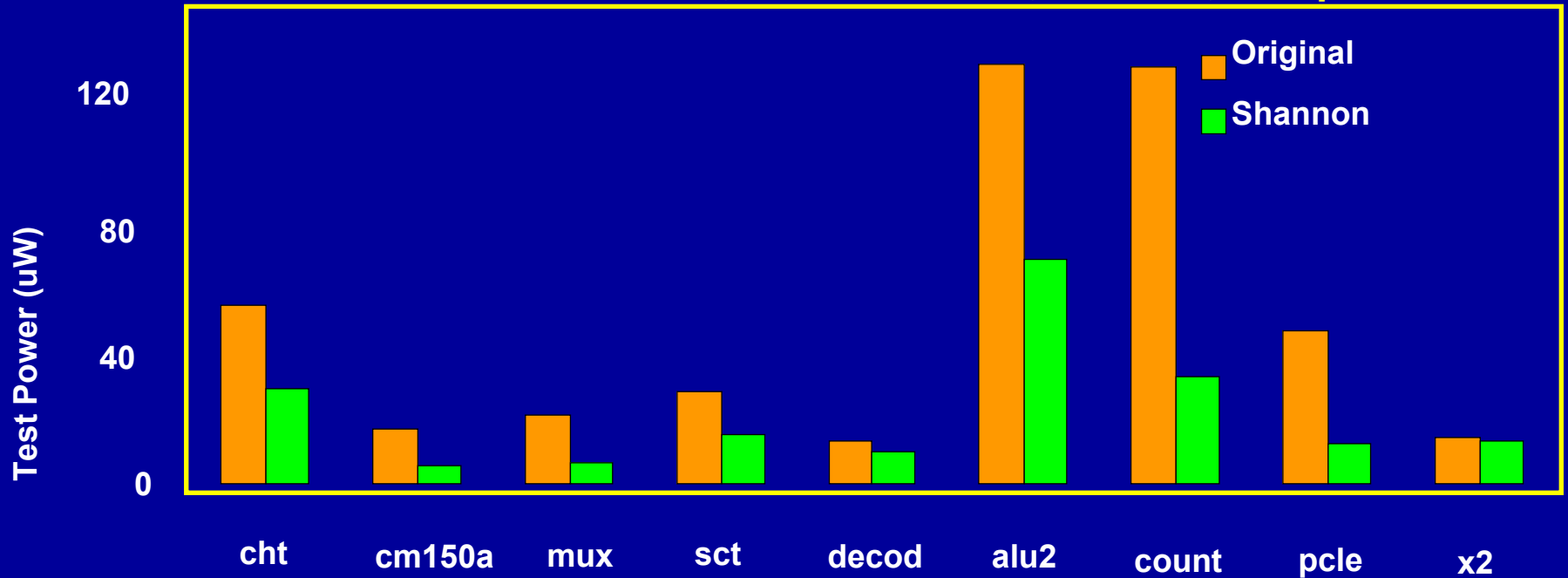
I_g = Fault free IDDQ



Avg. improvement of **94%** in IDDQ sensitivity

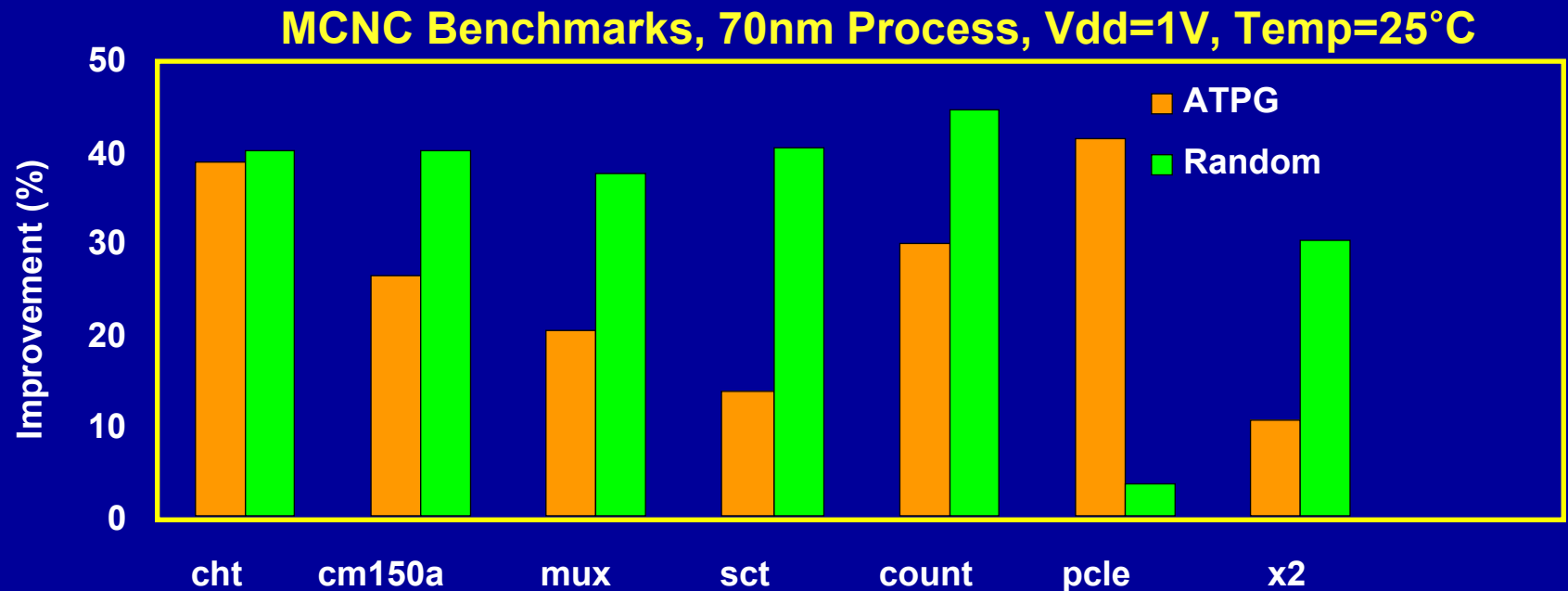
Improvement in Test Power

MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=25°C



Avg. reduction of **50%** in test power

Improvement in Test Coverage/Test Length

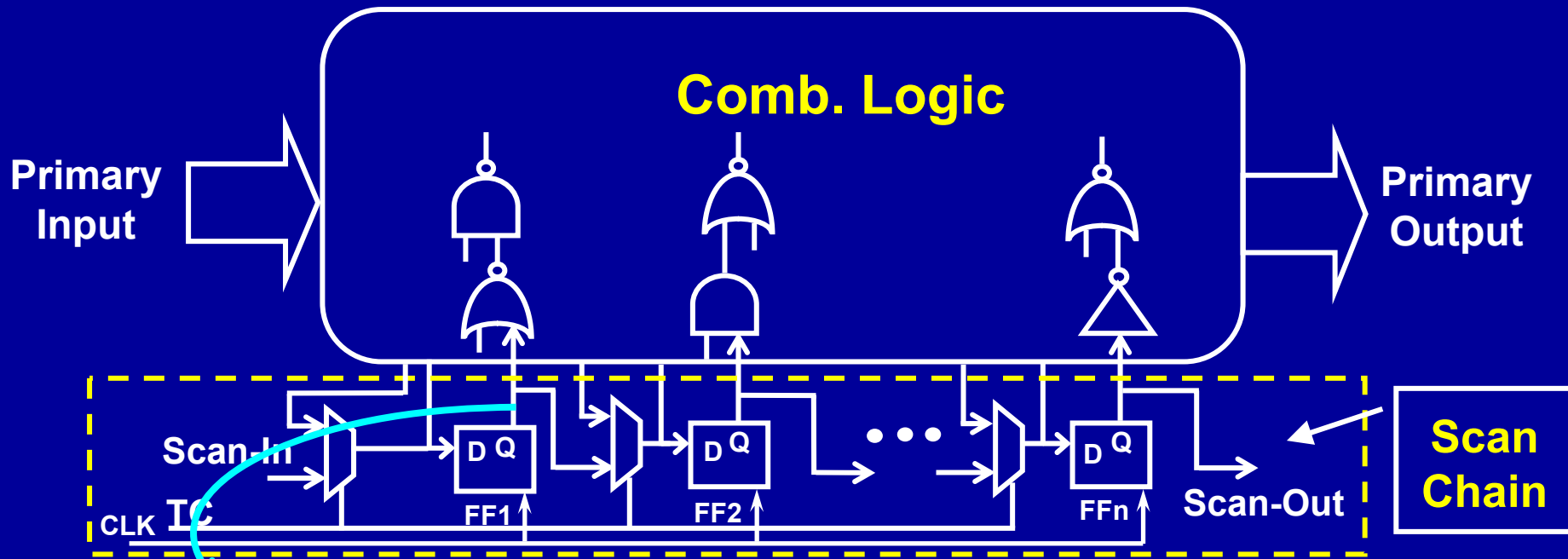


Avg. reduction of **20%** (**21%**) in test time with deterministic (random) patterns

Supply Gating in Scan Design

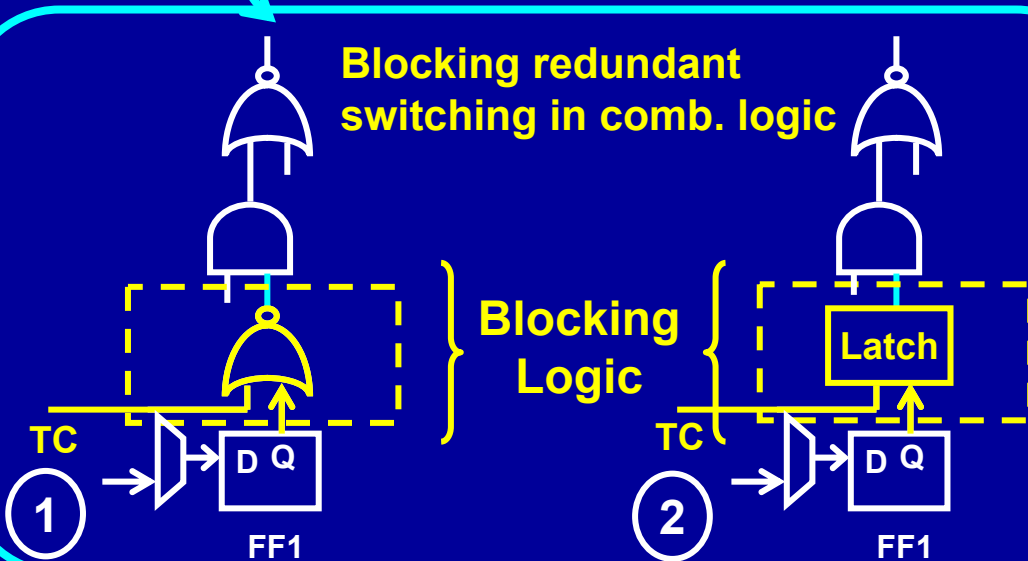
- Low-power Scan Operation

Conventional Scan Architecture



Blocking redundant switching in comb. logic

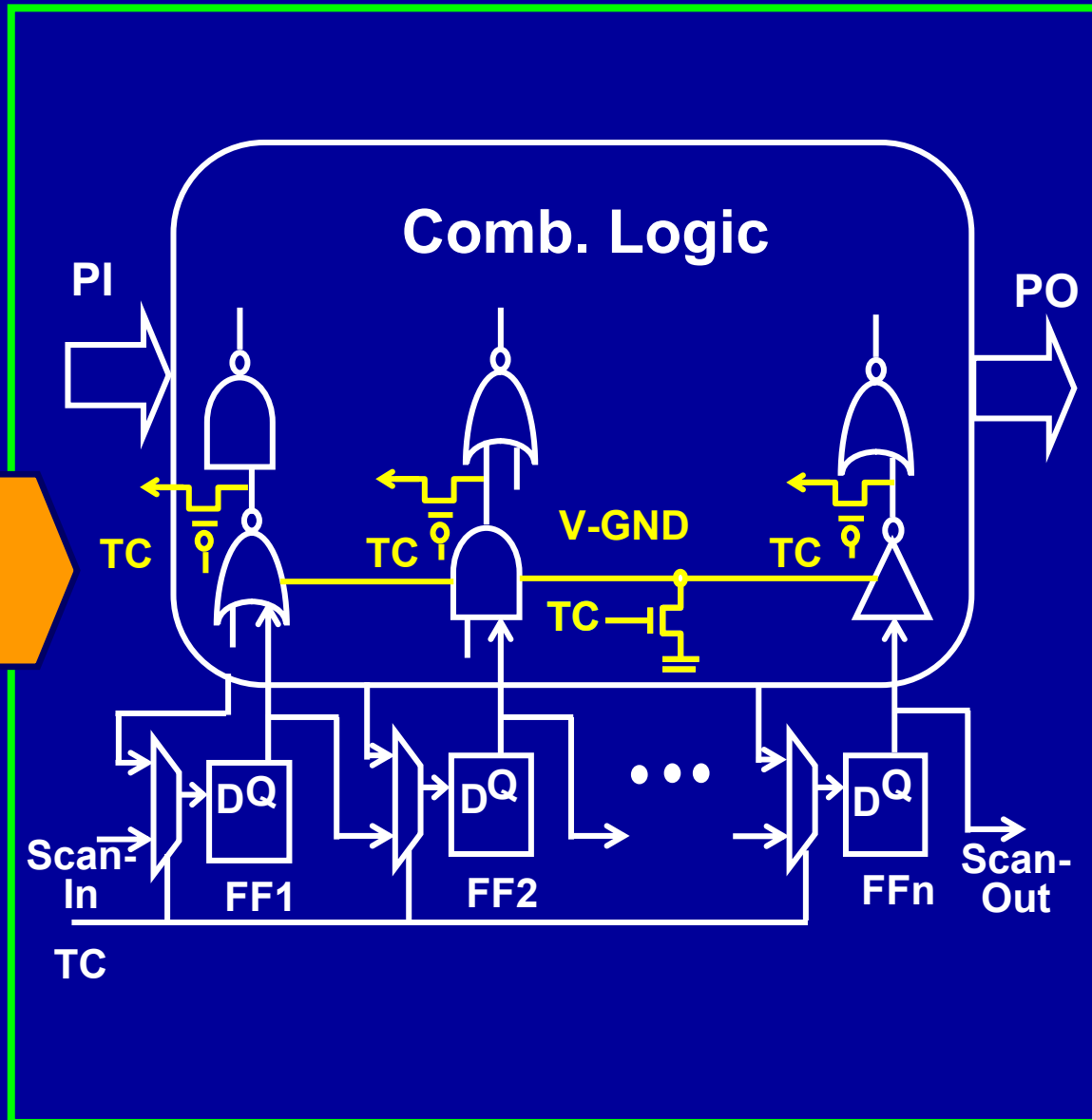
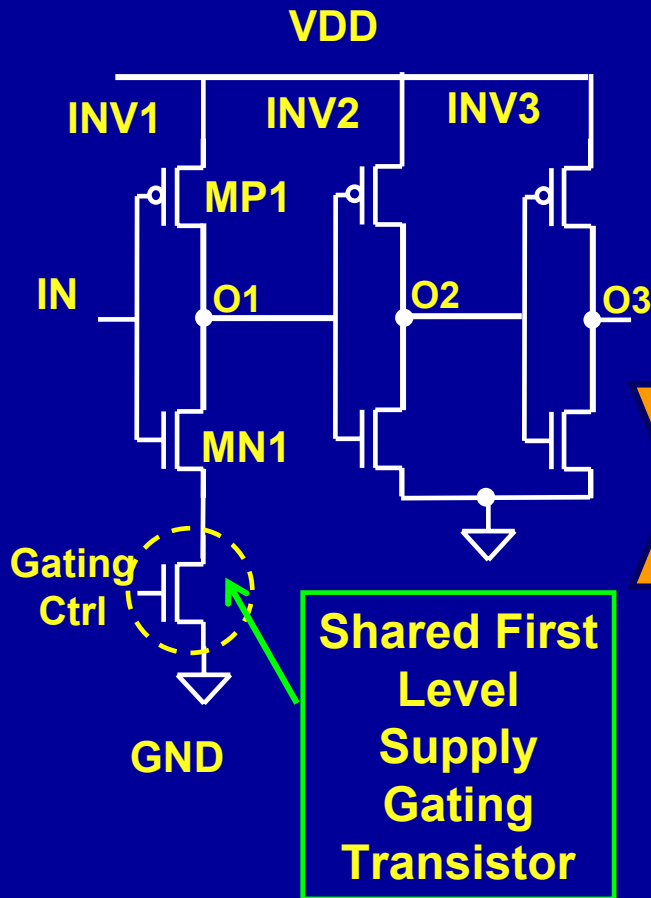
Blocking Logic



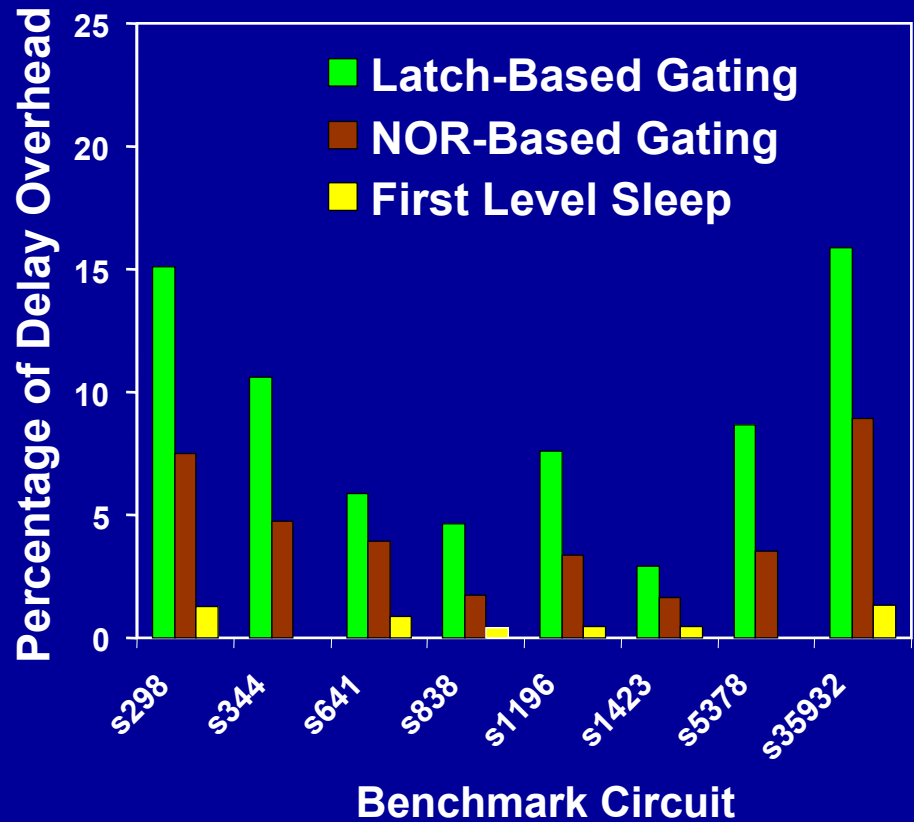
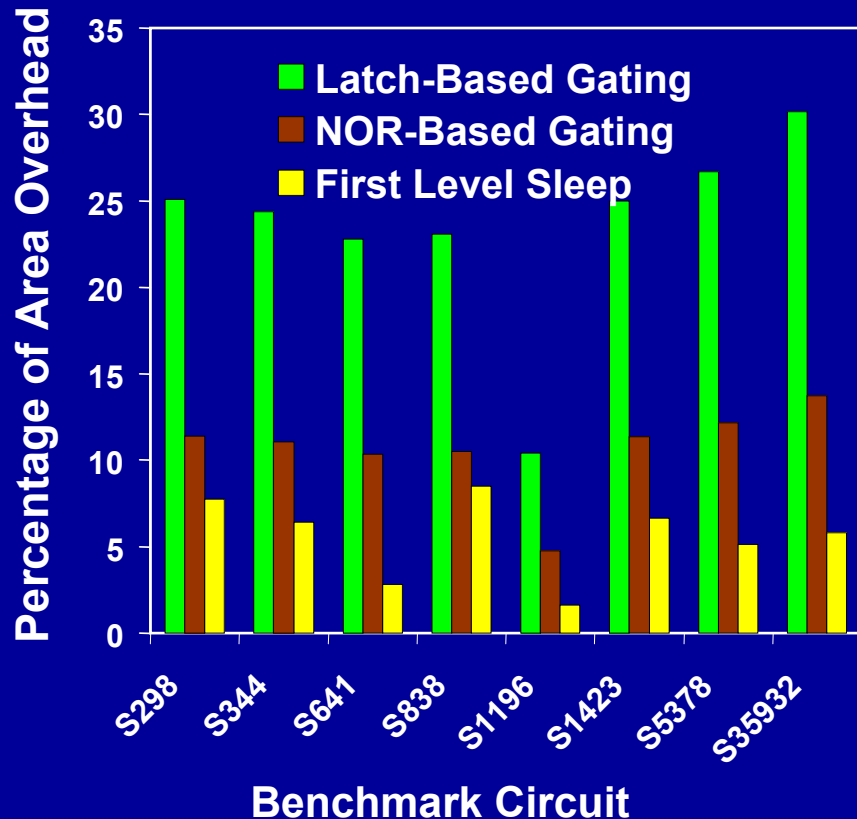
■ **High Design Overhead**

Any better solution?

First Level Supply Gating (FLS)



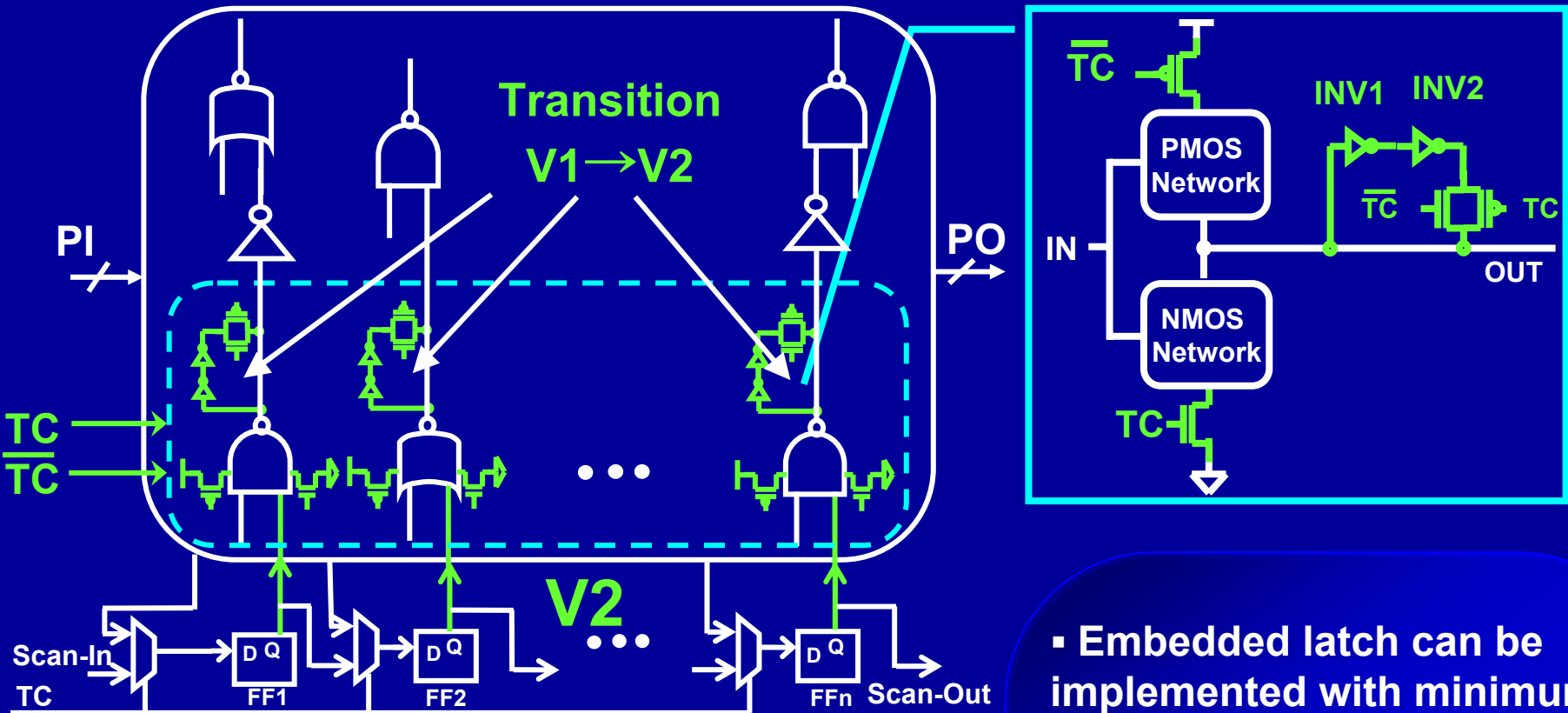
Results and Comparisons for FLS



- Compared to Nor-based Gating:
 - **Area: 62% less overhead**
 - **Delay: 94% less**

Low-Overhead Delay Fault Testing With Supply Gating

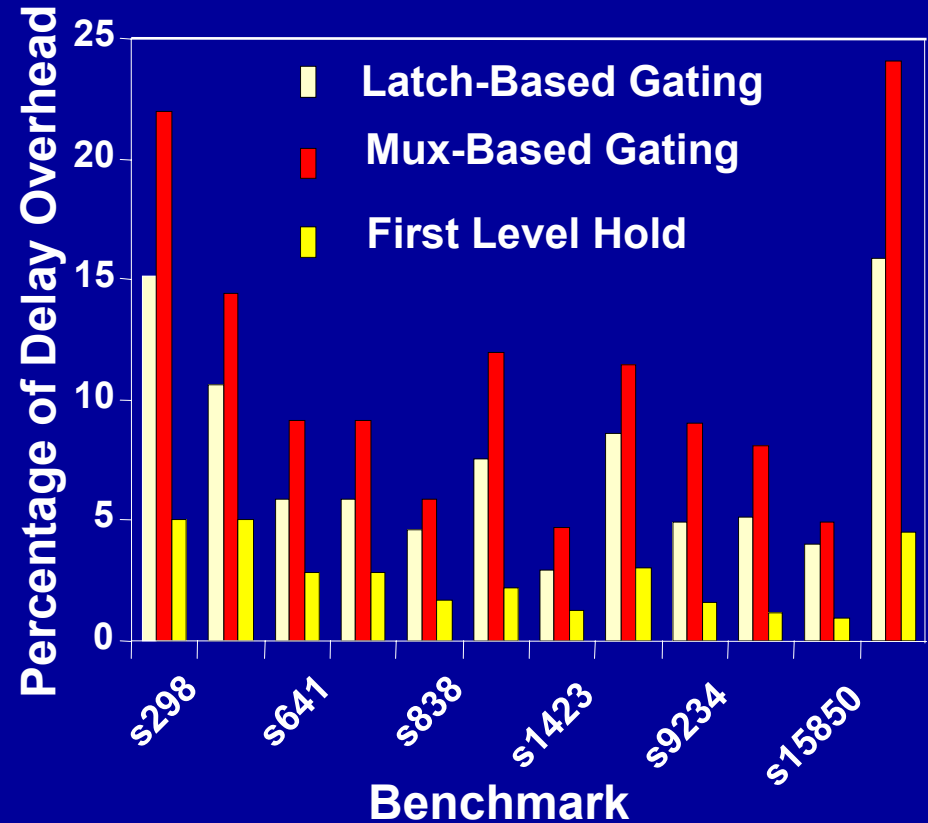
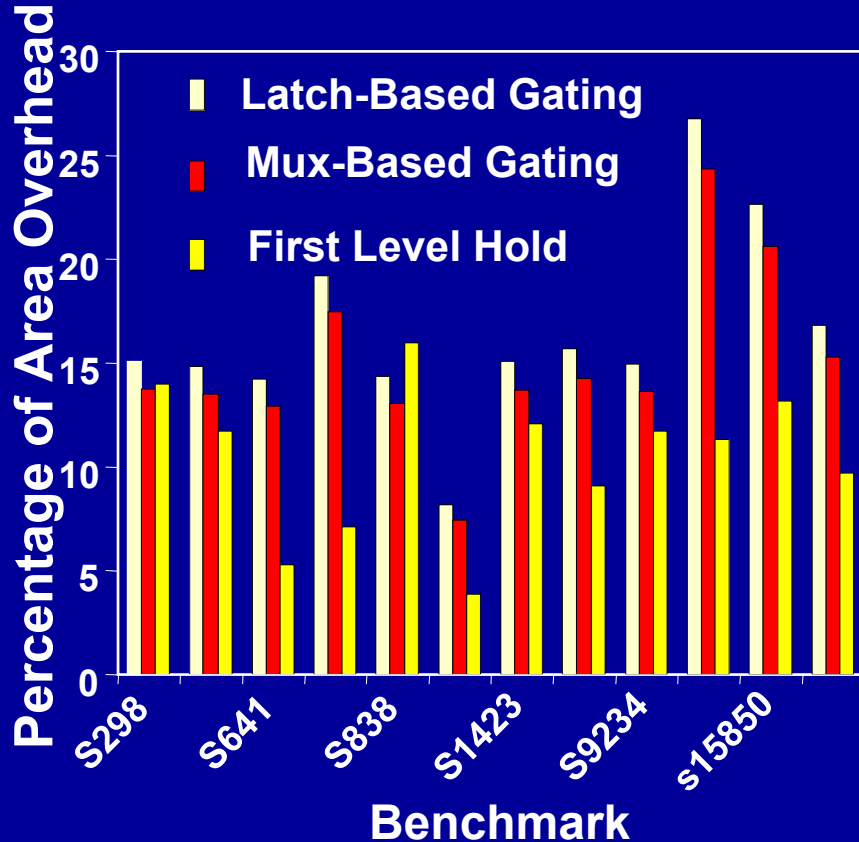
First Level Hold (FLH) for Delay Testing



1. Scan-in V1
2. Apply V1. Hold state for V1
3. Scan-in V2
4. Launch V2

- Embedded latch can be implemented with minimum-sized transistors
- No extra signal; simple control
- Eliminates redundant test power in comb. logic

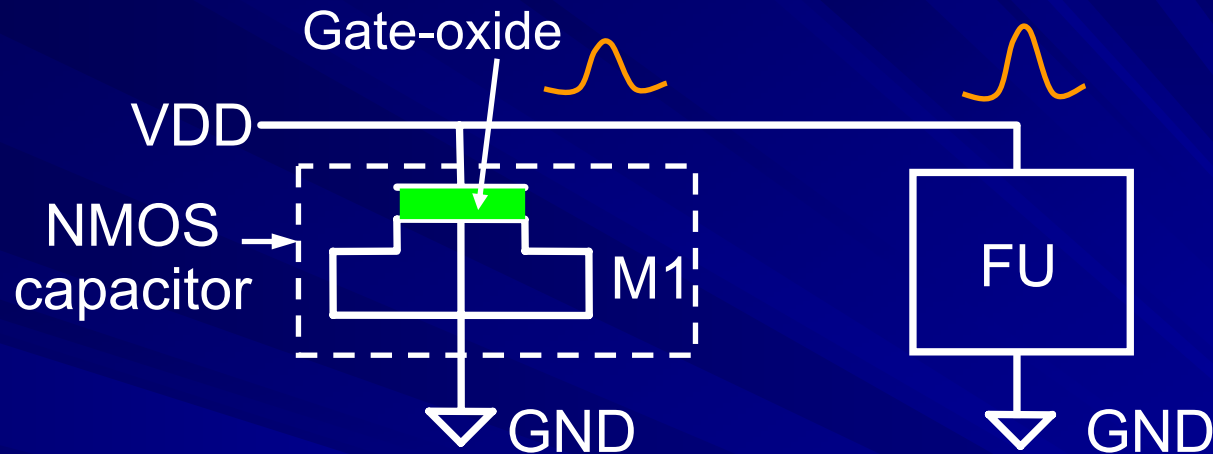
Results and Comparisons for FLH



- Compared to Enhanced Scan:
 - (a) Area: 33% less overhead, (b) Delay: 71% less overhead, (c) Power: 90% less overhead
- Local Fanout Reduction reduces area overhead by ~20%

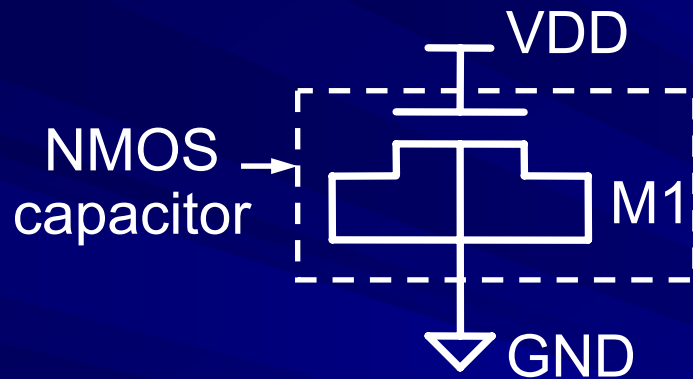
Gated DeCap: Another Application of Stacking & Leakage Reduction

Decoupling Capacitor (Decap)

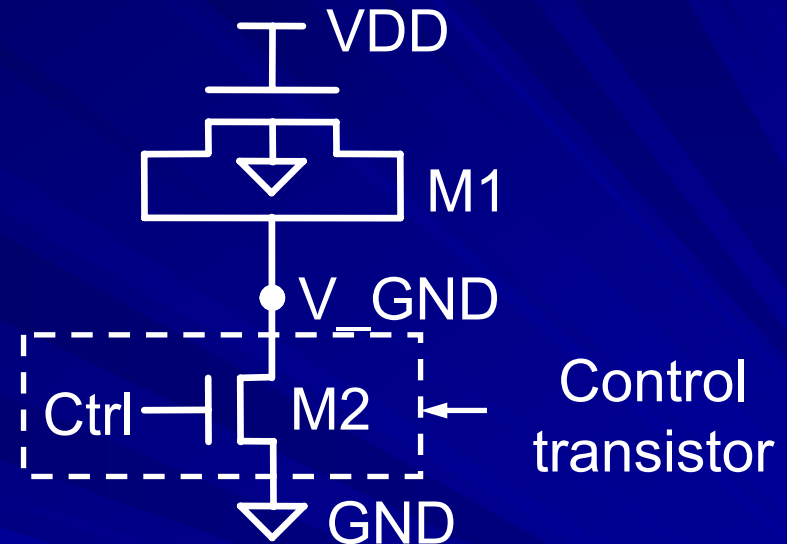


- Area and power of Decap
 - 15-20% of the total chip area (Alpha 21264).
 - Large Decap gate leakage power consumption (reported by IBM, 2003).

Gated-Decap



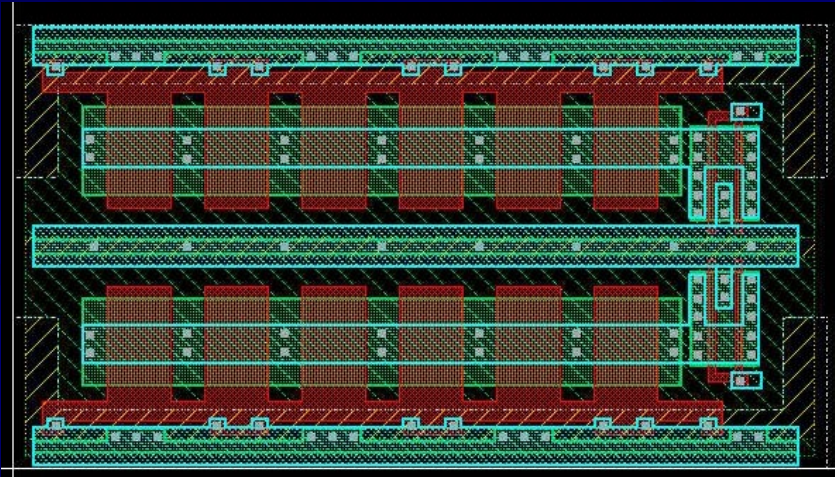
(a) Conventional NMOS Decap



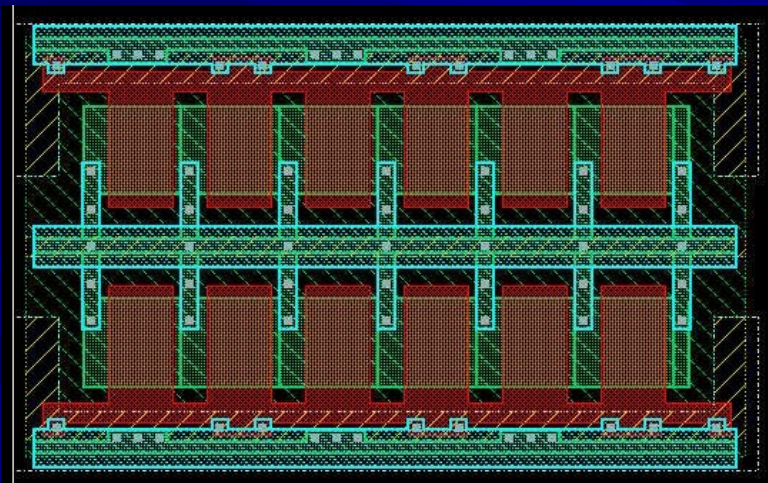
(b) NMOS Decap with control gate

- The gate and the channel of M1 constitute a capacitor.
- M2 is turned off when Decap is unnecessary (FU is idle).

Layout of GDecap

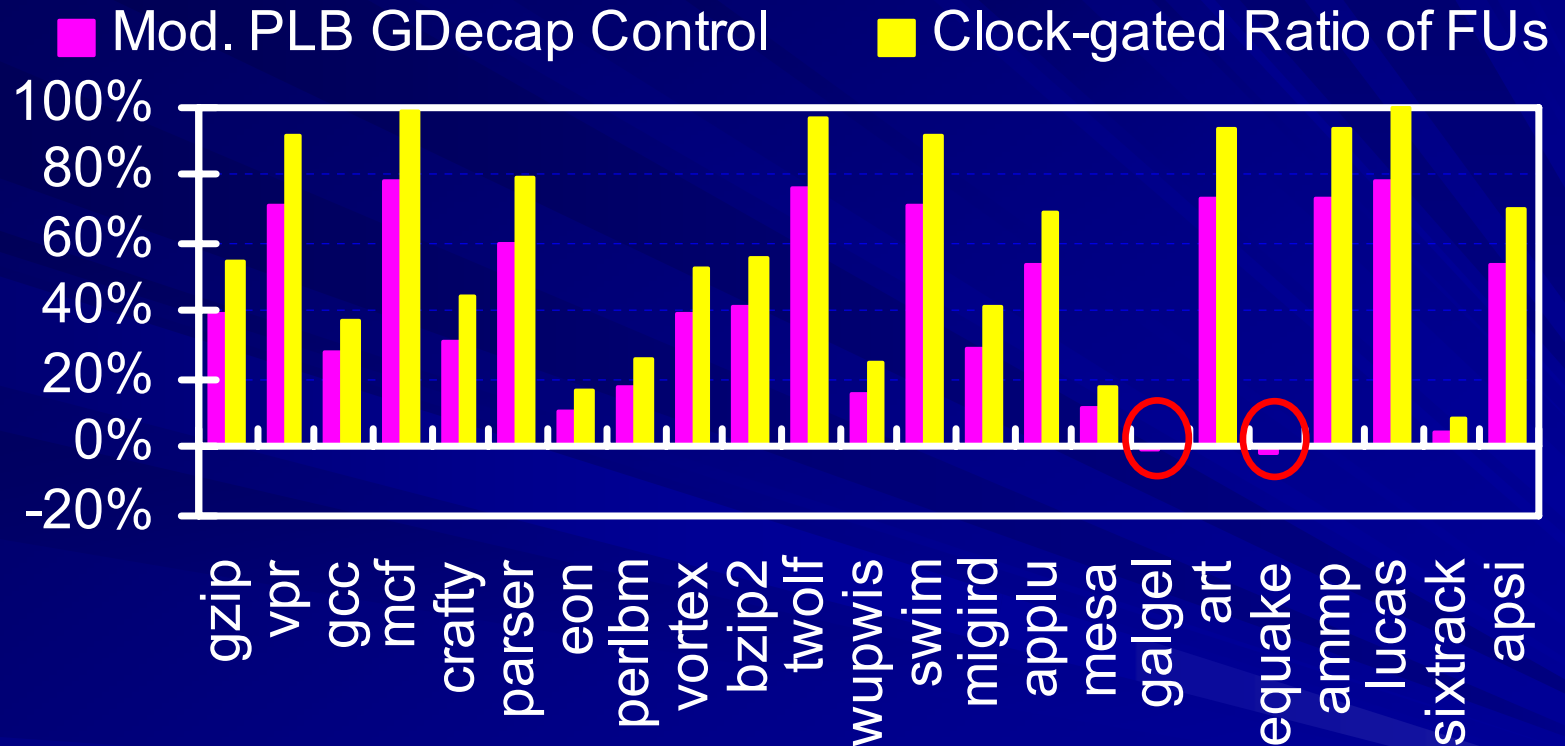


GDecap
Area Overhead:
6.78%



Conventional
Decap

Leakage Power Saving of GDecap in PLB Pipeline

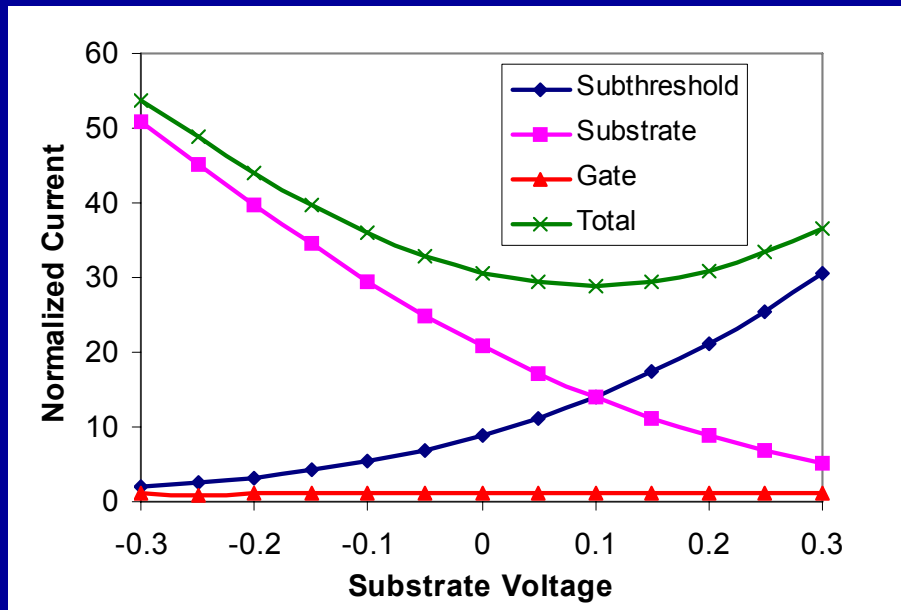


- Average Decap leakage power reduction:
Mod. PLB – **41.7%** (FU gated ratio: 55.15%)
- 0.037% worst-case IPC degradation in Mod. PLB.

Leakage & Body Bias

- Sub-threshold leakages decreases with RBB
- Band-to-band tunneling increases with RBB
- Gate Leakage insensitive to body bias

Results for 70nm nmos



**BSIM3 device augmented
with voltage-controlled
current sources for gate
leakage and BTBT**

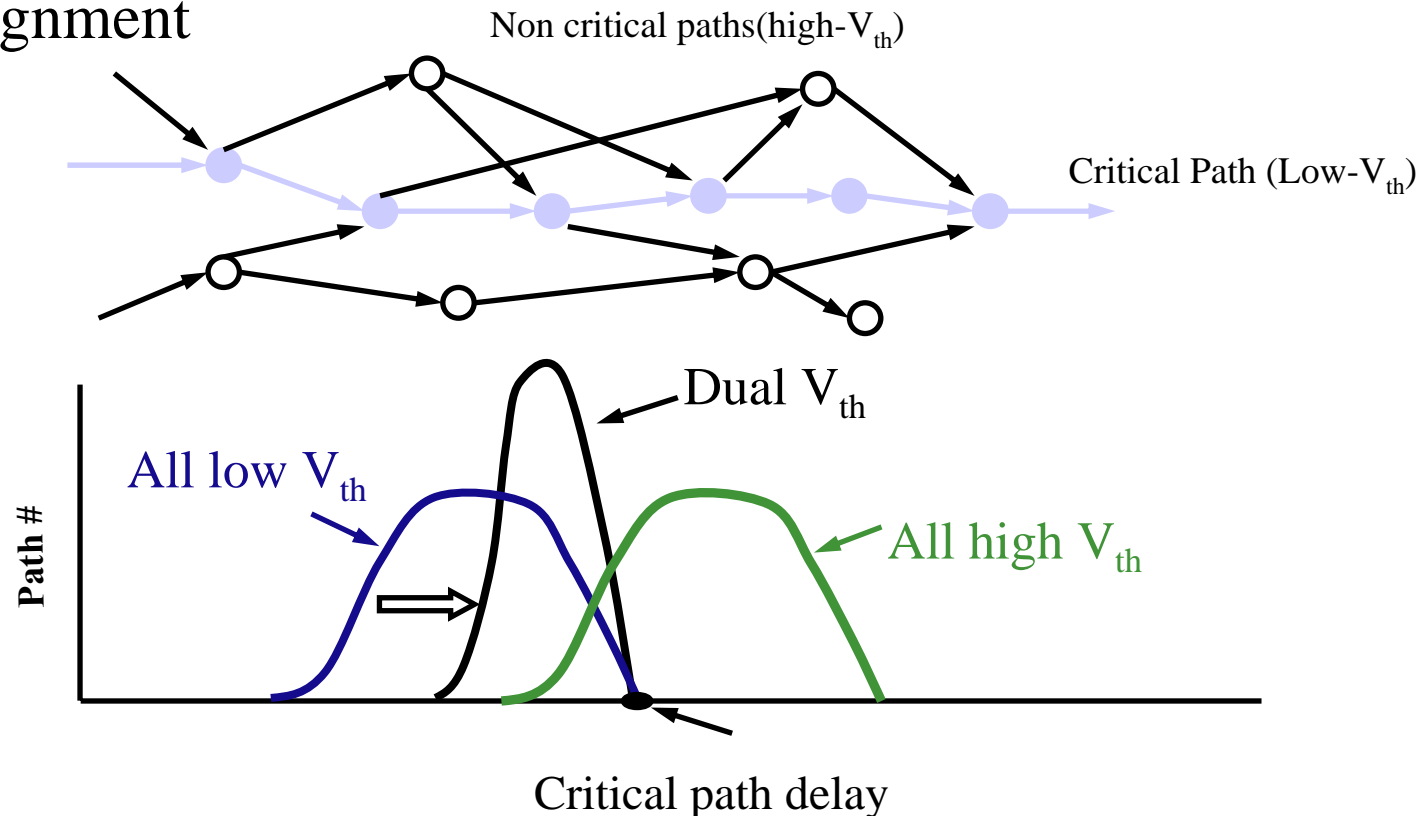
Leakage Reduction with OBB

- Leakage savings ranged from 14-55% compared to zero body bias case for nominal 70nm and 50nm transistors in Taurus device simulations.

Tech.	Temp (°C)	V _B (V)	I _{OFF} (normalized)	I _{ON} (normalized)	I _{ON} /I _{OFF}	Leakage Reduction
70nm	25	0	1	97115	97115	43%
	25	-0.16	0.57	91005	159657	
	70	0	5.14	120673	23477	55%
	70	-0.20	2.30	118269	51421	
50nm	25	0	1	3478	3478	45%
	25	0.15	0.55	3992	7258	
	70	0	2.51	4044	1611	14%
	70	0.09	2.15	4286	1993	

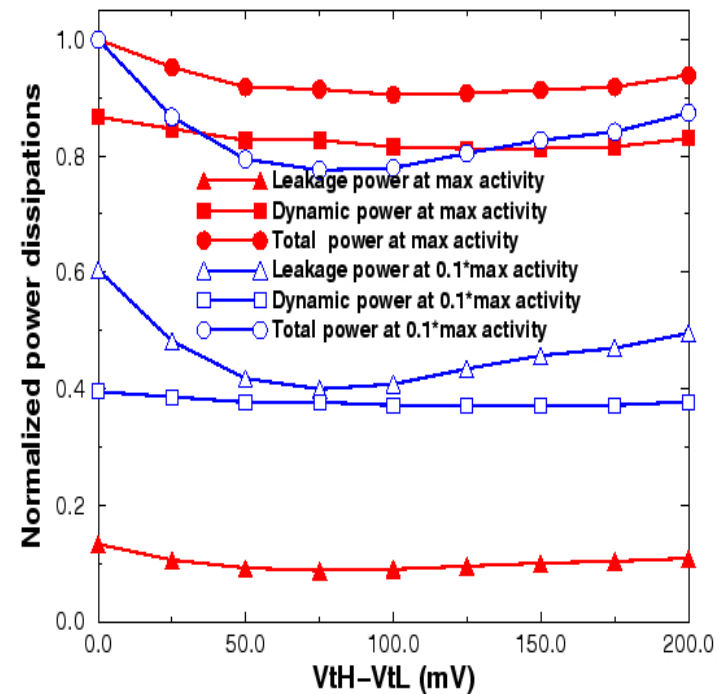
Dual Threshold CMOS

- Low- V_{th} transistors in critical path for high performance
- Some high- V_{th} transistors in non-critical paths to reduce leakage
- Impact on yield – need to consider variations and V_t -assignment



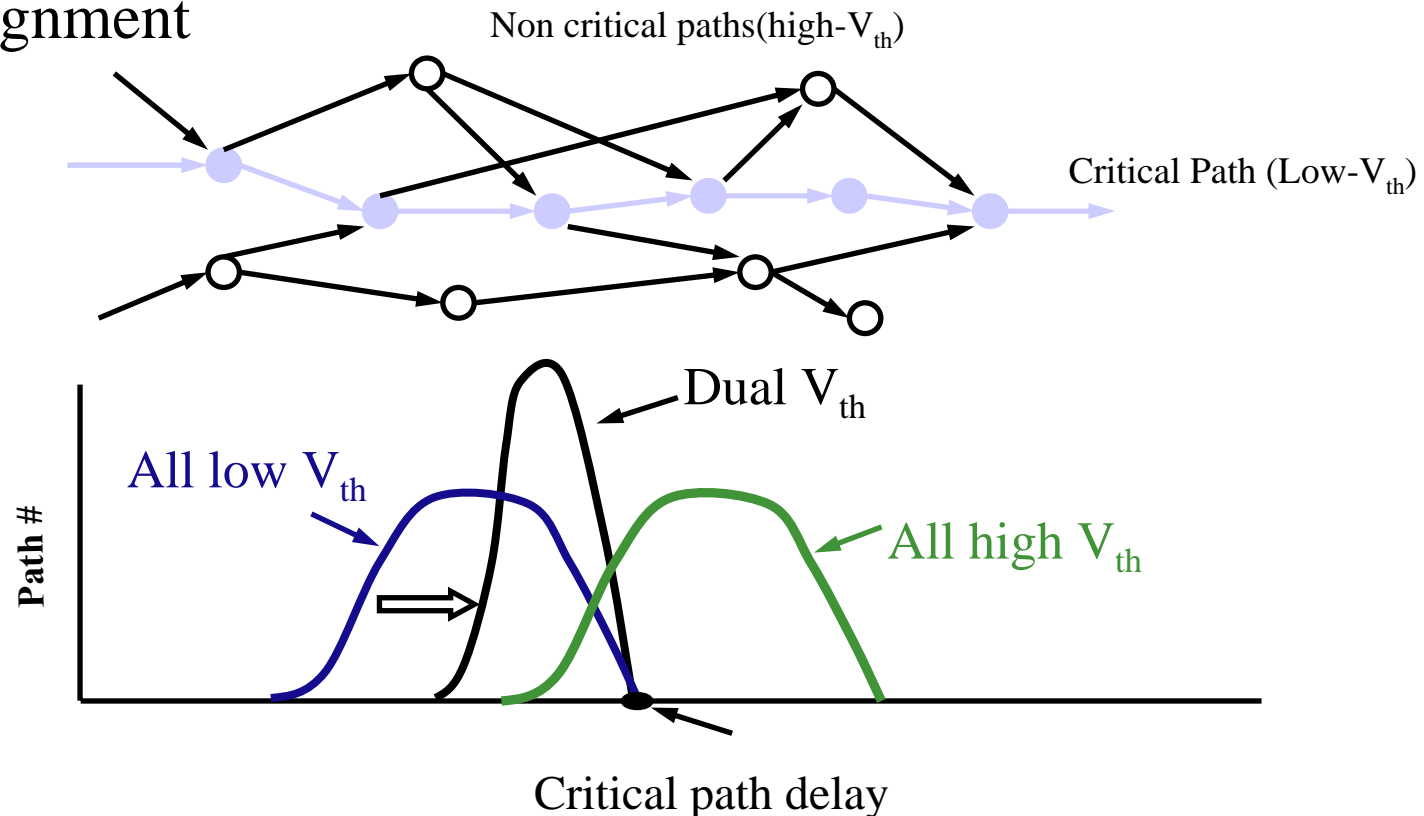
Total Power of 32-bit Adder

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



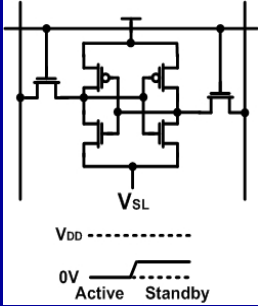
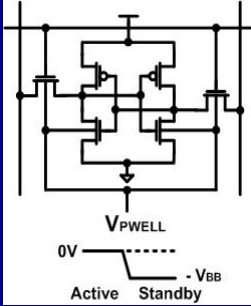
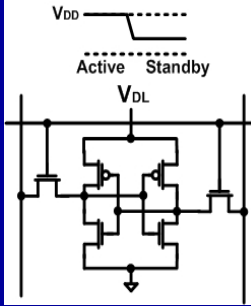
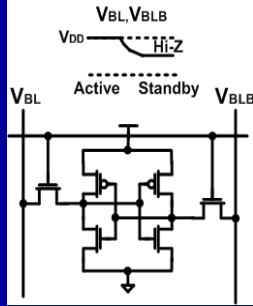
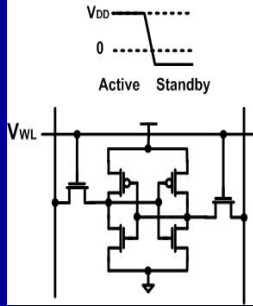
Dual Threshold CMOS

- Low- V_{th} transistors in critical path for high performance
- Some high- V_{th} transistors in non-critical paths to reduce leakage
- Impact on yield – need to consider variations and V_t -assignment

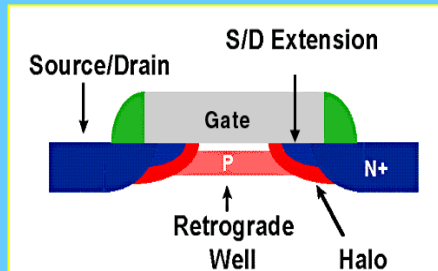


Design of Nanometer Caches: Low-Leakage

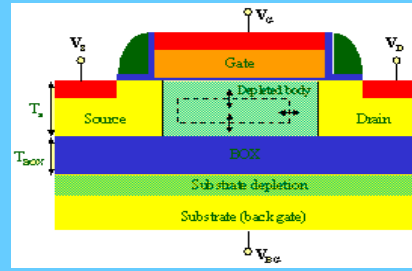
SRAM Leakage Reduction Schemes

Schemes	Source Biasing (V_{SL})	Fwd/Reverse Body-Biasing (V_{PWELL} , V_{NWELL})	Dynamic V_{DD} (V_{DL})	Floating Bitlines (V_{BL} , V_{BLB})	Negative Word Line (V_{WL})
					
Leakage reduction	Sub: ↓↓ Gate: ↓↓	Sub: ↓↓ BTBT: ↑ (RBB)	Sub, gate: ↓ *Bitline leak: -	Sub: ↓ Gate: ↓	Sub: ↓ *Gate: ↑
Delay	*Delay increase	No delay increase	No delay increase	No delay increase	No delay increase
Overhead	Low transition overhead	Large transition overhead	Large transition overhead	*Precharge latency overhead	*Low charge pump efficiency
Stability	Impact on SER	No impact on SER	*Worst SER	No impact on SER	No impact on SER, voltage stress

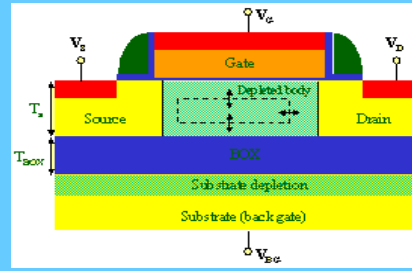
Device-aware Circuit/Microarch: Cache



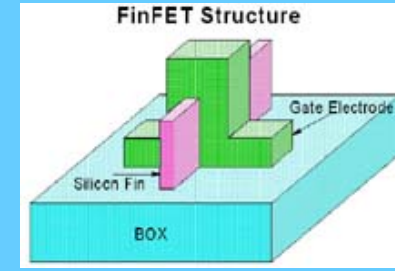
Bulk Ultra-high V_t



Nominal V_t



Ground-plane SOI



FinFET

Circuit Design Issues

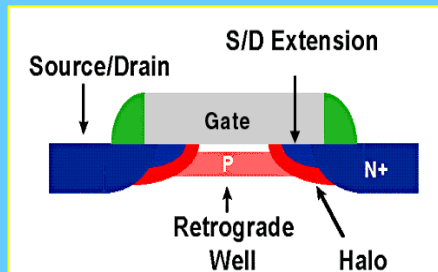
Leakage – Sub-threshold, Gate, Junction, BTBT
Stability – Read noise margin, Writability, Soft errors
Delay – Decoder, Wordline, Bitline, MUX, Sense-amp, Driver
Transition between active and standby modes
Variations – Process, V_{dd} , Temperature

Microarch Design Issues

Array aspect ratio – # cells WL/BL
Sub-array structure and selection strategy
Active-Standby transition frequency, delay, energy

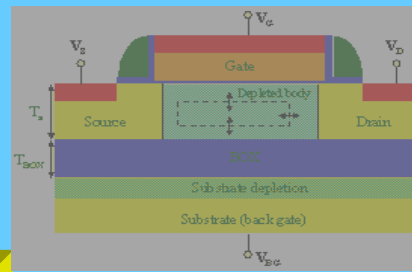
How do you co-design?

Bulk Nominal V_t Source-biased Cache

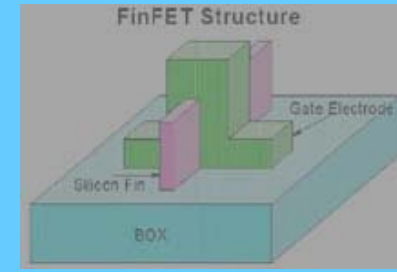


Bulk Ultra-high V_t

Nominal V_t

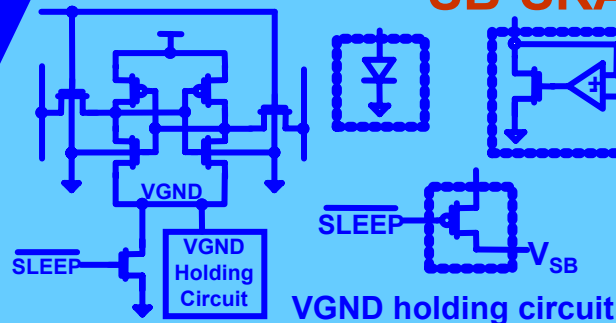


Ground-plane SOI

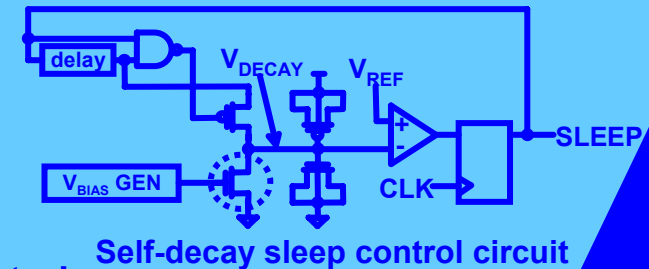


FinFET

SB-SRAM Circuit Design Issues

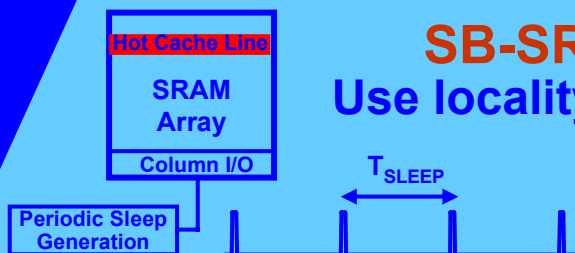


- Data retention (VGND should be strapped)
- Noise issue
- Process variation tracking sleep control



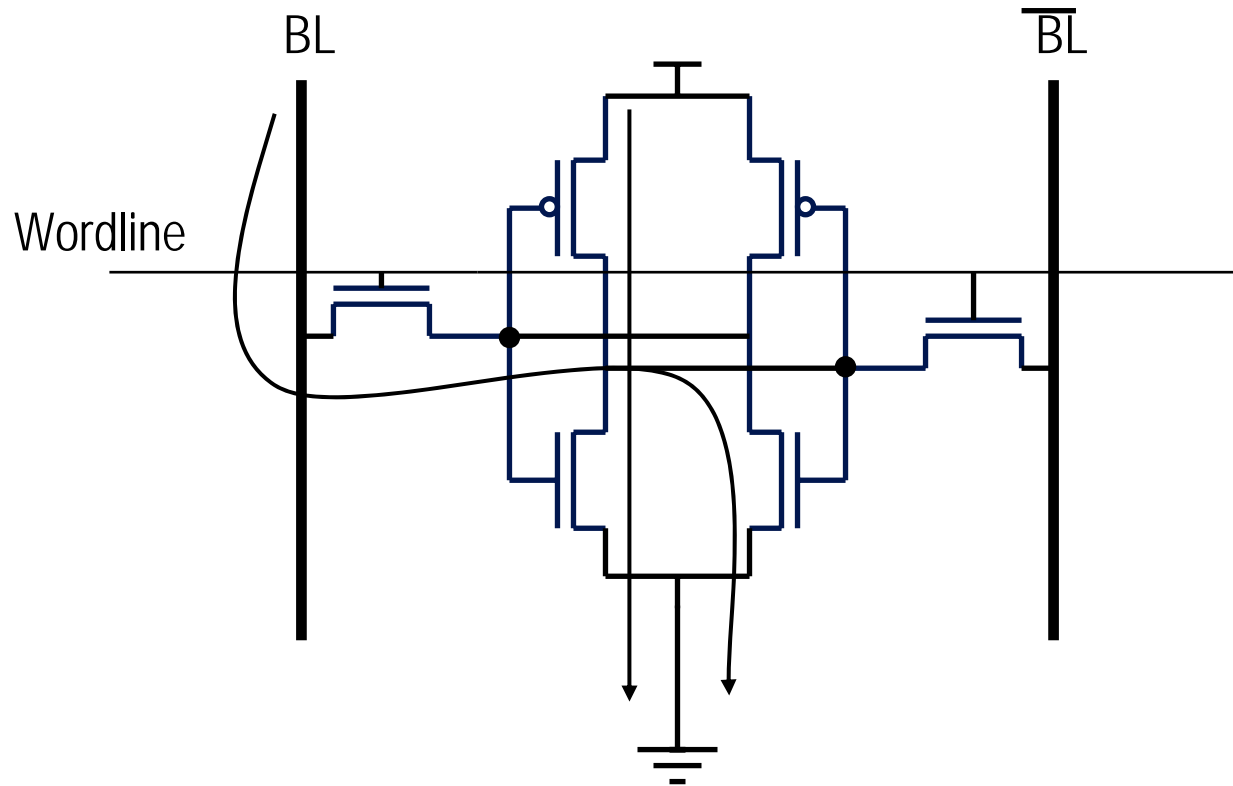
SB-SRAM Microarch Design Issues

Use locality of reference in cache to reduce transition energy
Optimum memory sub-array size selection
Sleep time T_{sleep} selection



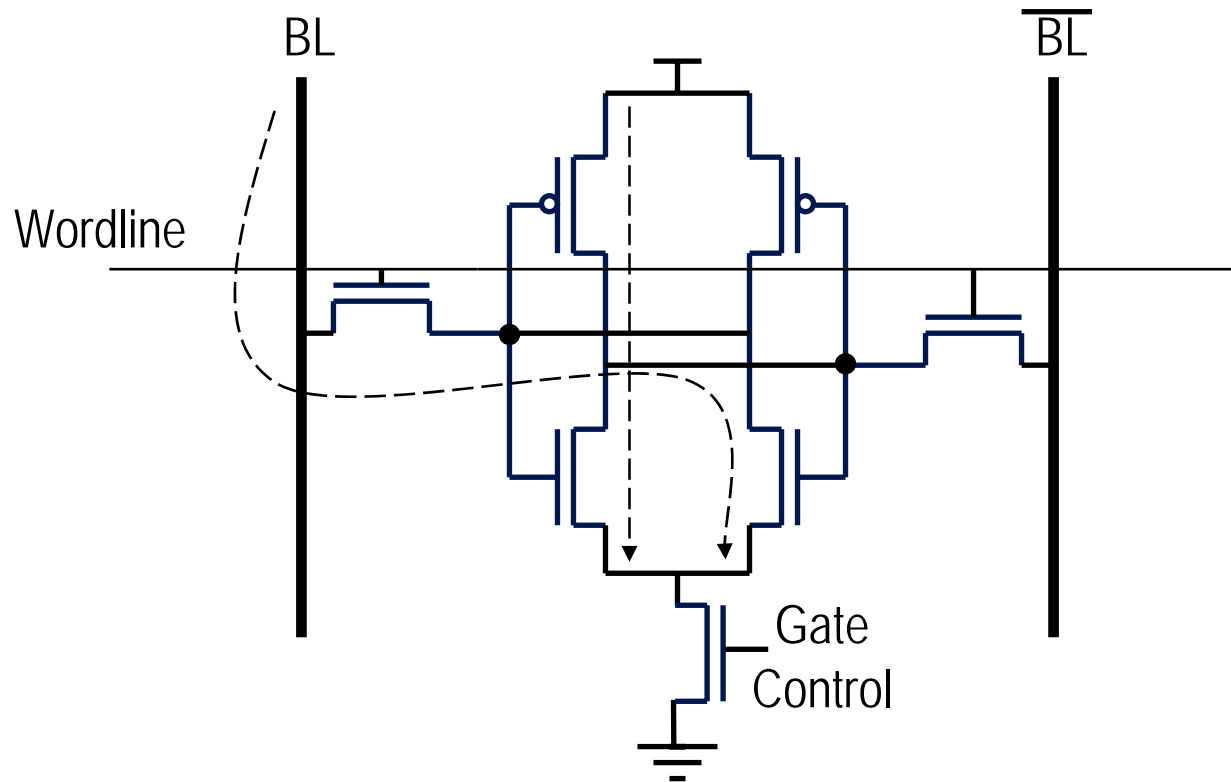
Co-design approach leads to higher payoffs and more opportunities

Conventional Cell Leakage Paths



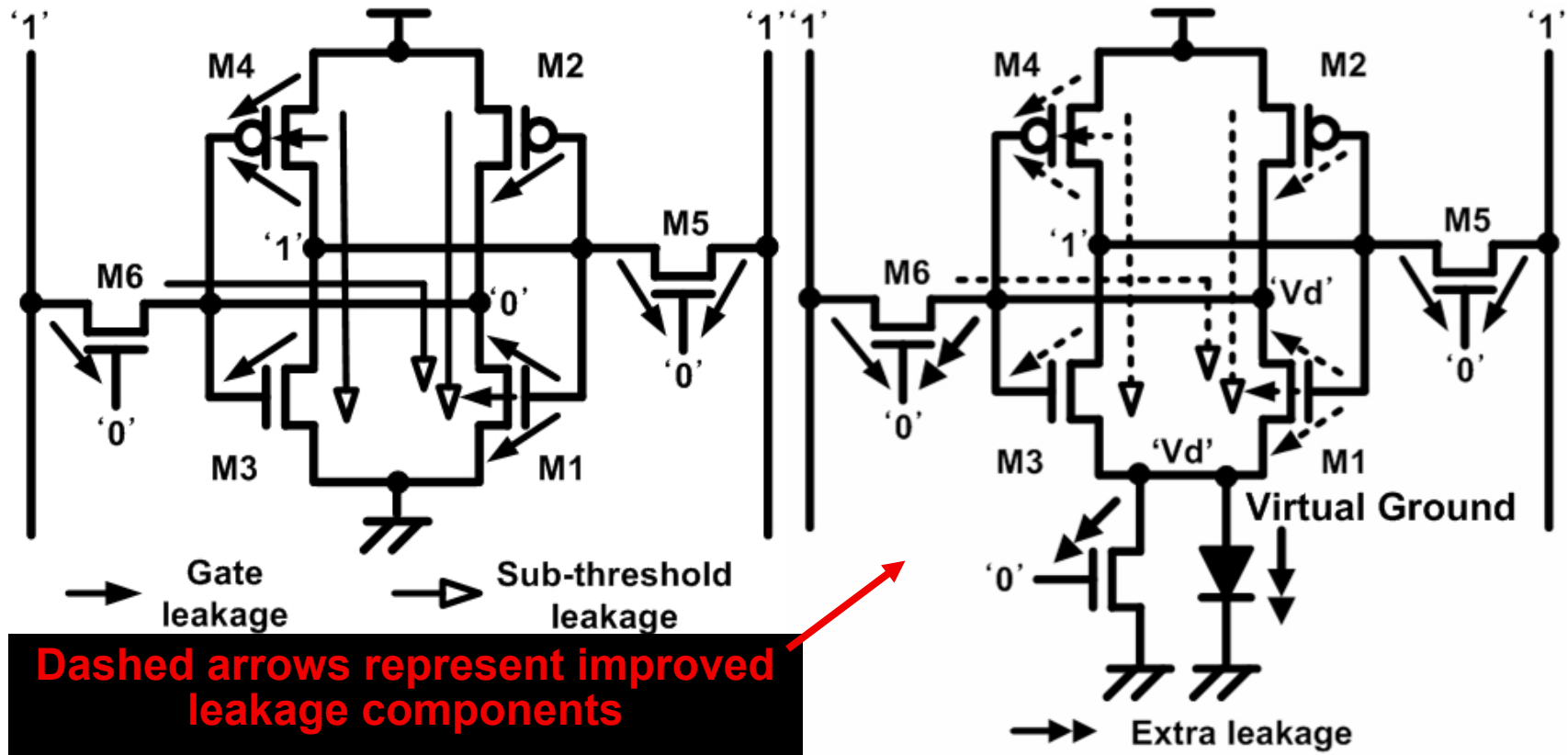
- V_{dd} to ground path
- Bitline to ground path

Gated-Ground (Source-Biased) SRAM



- Gating options: NMOS, Dual- V_t , PMOS

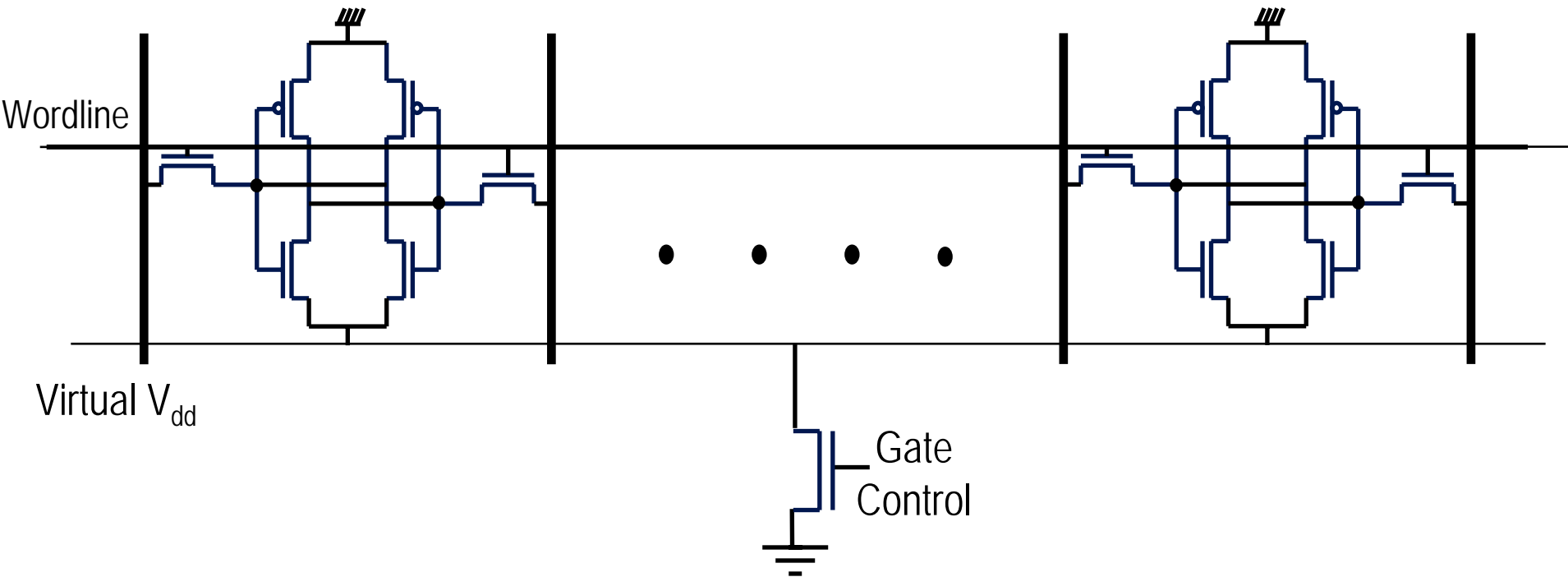
Leakage Reduction in Diode Footed Cache



Voltages across terminals get reduced by V_d (diode intrinsic voltage)

Reduces gate and subthreshold leakage

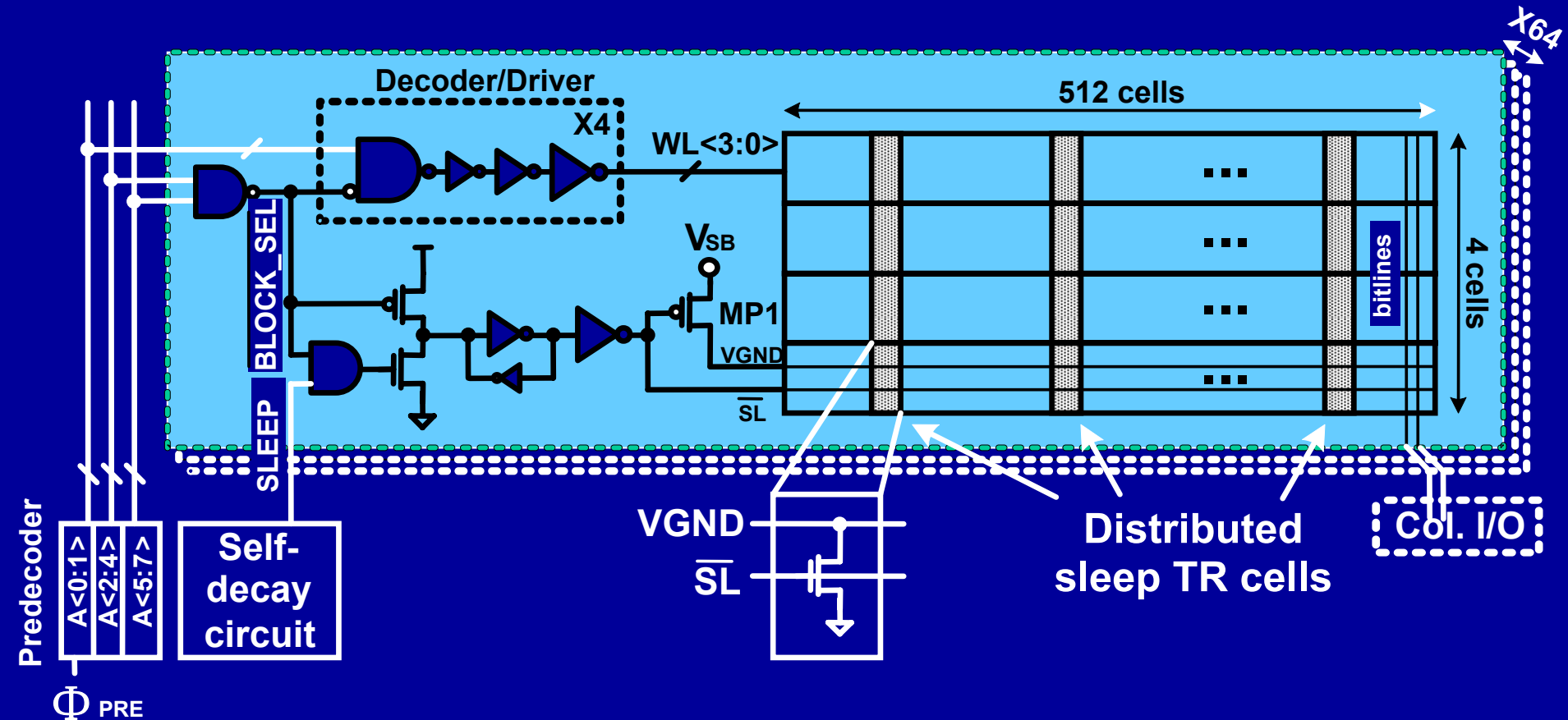
Gated-Ground Transistor Sharing



Gated V_{dd} transistor

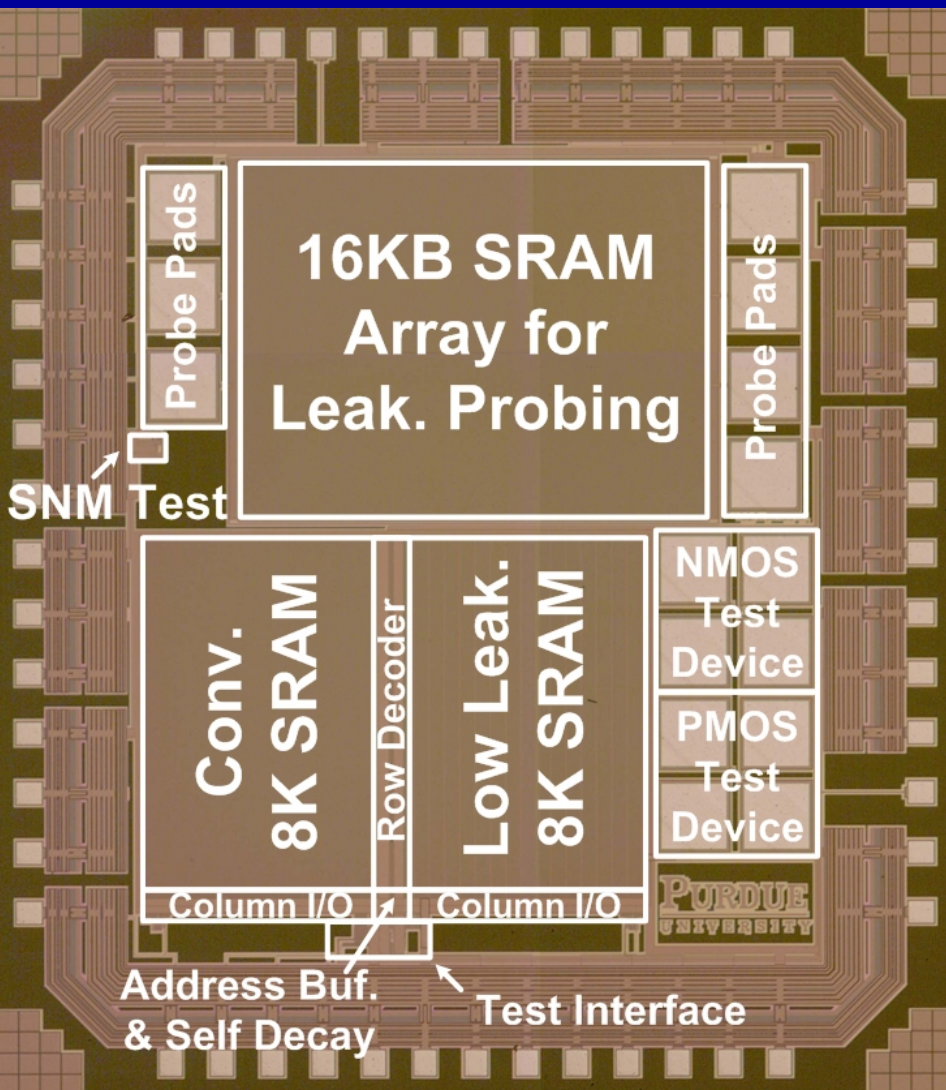


16K-Byte SRAM Organization



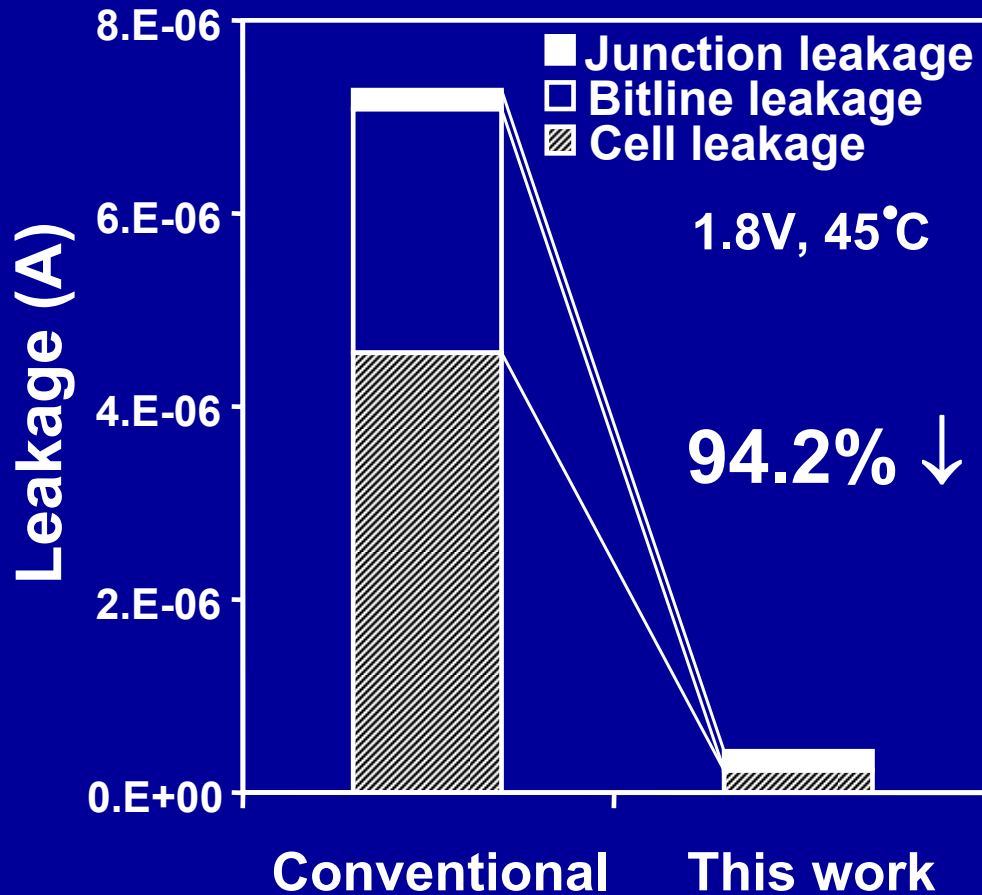
- Active leakage reduction SRAM
- Distributed sleep transistors
- SRAM block turned on ahead of time
- Self-decay circuit for low dynamic power overhead

2x16K-Byte SRAM Testchip



Technology	180nm 6-metal CMOS
Chip Size	3.3X2.9 mm ²
Supply Voltage	1.8V
Threshold Voltage	NMOS: 0.53V PMOS: -0.53V
Read Access Cycle	984MHz @ 1.8V, RT
Active Current	0.14mW/MHz @ 1.8V
Standby Current	7.27μA (16KB array)

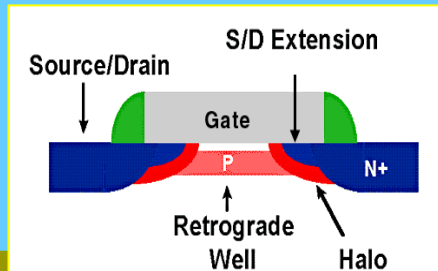
Measured Leakage Reduction



- 94.2% total leakage reduction at $V_{GND}=0.9V$
- Raising V_{GND} also reduces gate tunneling leakage

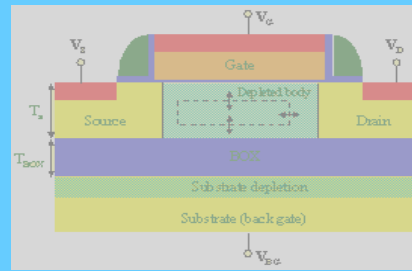
Forward-Body Biased Cache

Bulk Ultra-High V_t Forward-biased Cache

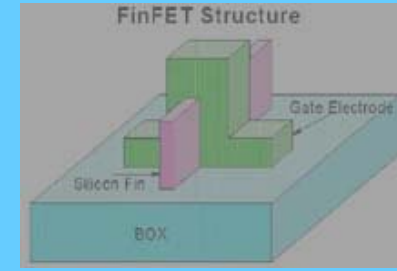


Bulk Ultra-high V_t
Strong halo, Low I_{SUB}
FBB to $\uparrow I_{ON}$

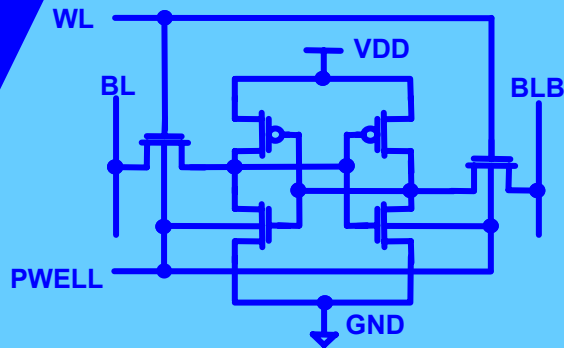
Nominal V_t



Ground-plane SOI



FinFET



FB-SRAM Circuit Design Issues

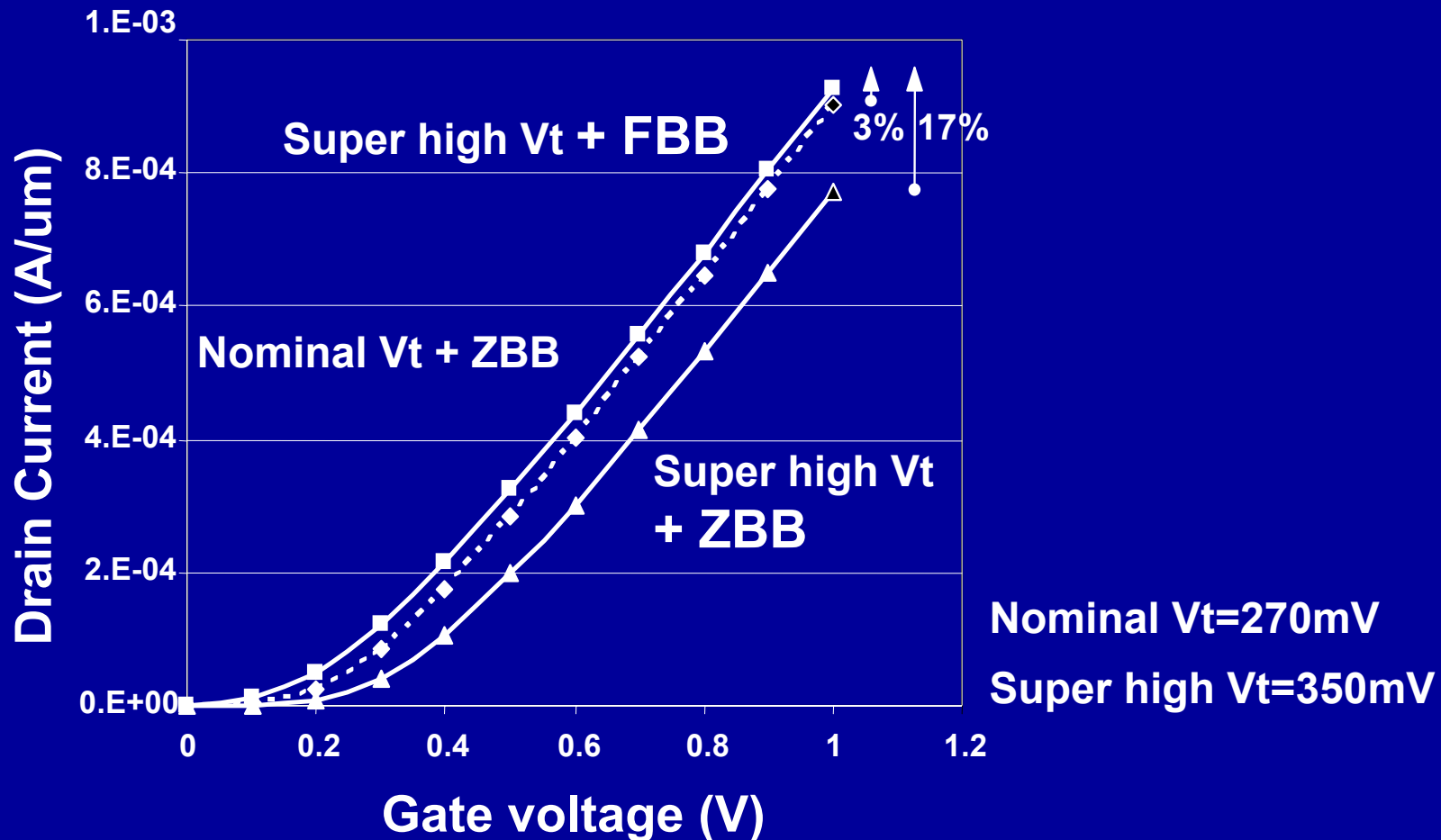
- Zero body bias in standby to reduce leakage
- FBB in active-mode to improve speed
- Early sub-array selection to hide body-bias transition latency

FB-SRAM Microarch Design Issues

Use MSB of memory address for early selection of memory sub-array
Use locality of reference in cache to reduce transition energy

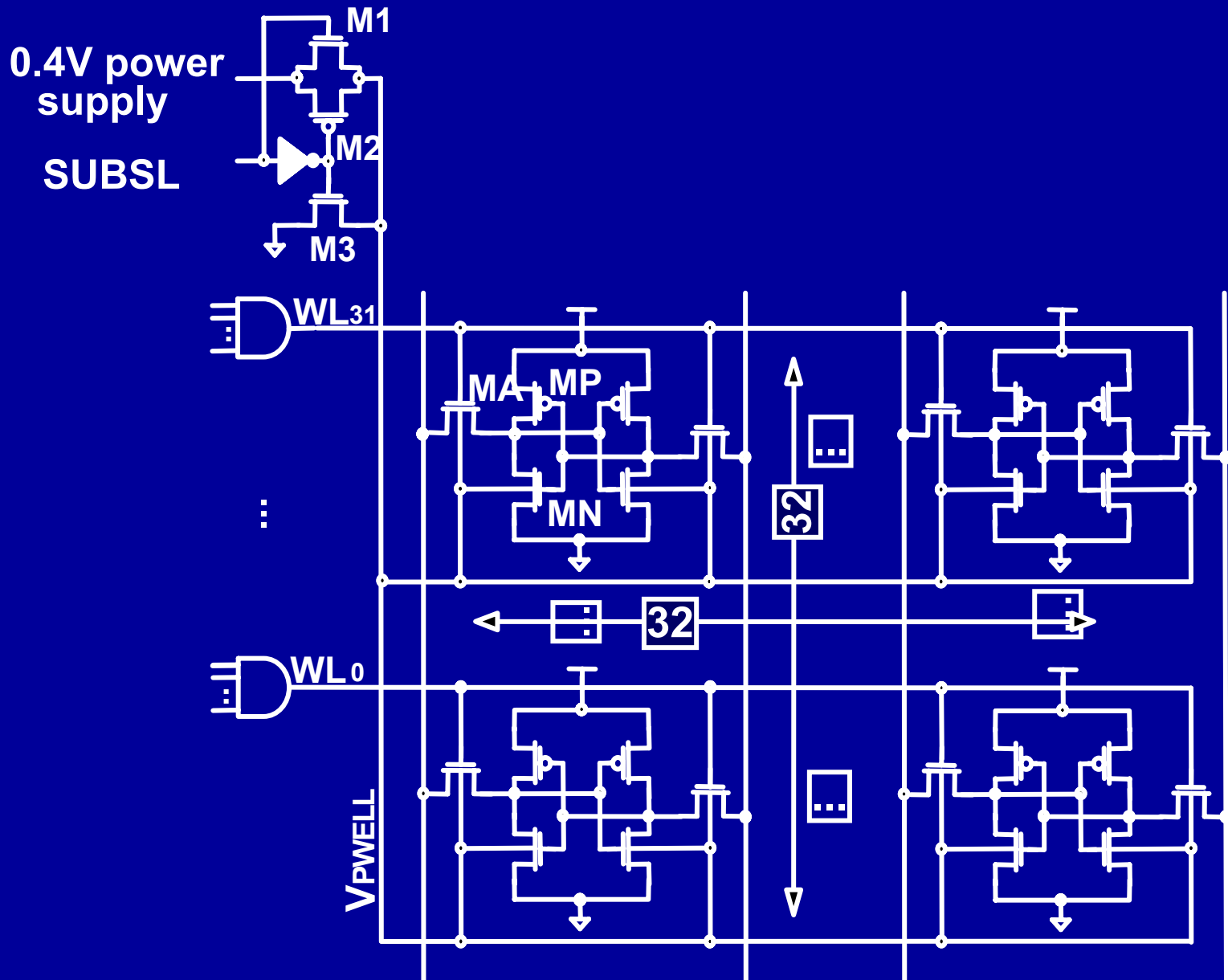
Co-design approach gives 64% leakage savings

Forward Body-Biased Cache (50nm)



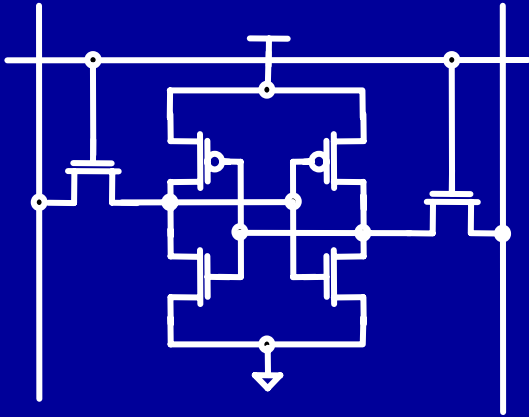
- Previous techniques: use circuit/arch. to lower leakage
- This technique: use dev/ckt/arch opt. to lower leakage
- Main idea: high Vt device + forward body-biasing

32x32 Forward Body-Biased Sub-array



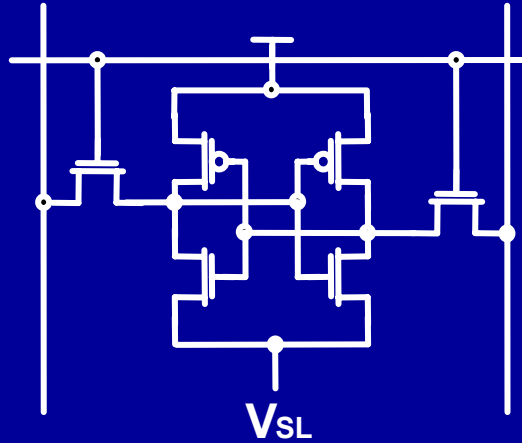
Comparison

Conventional



$V_T=270\text{mV}$

SBSRAM

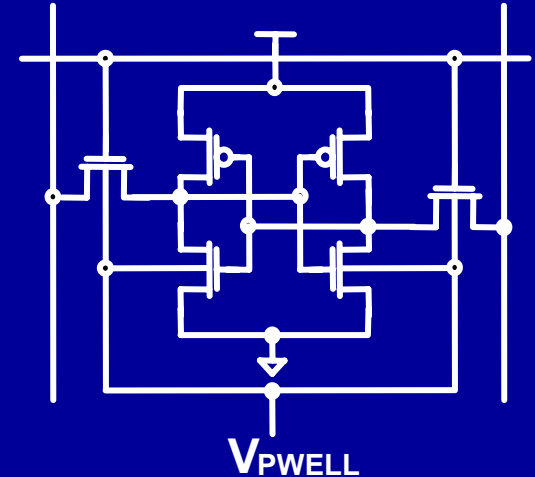


V_{DD}

0V _____ 0.2V
Active Standby

$V_T=270\text{mV}$

FBSRAM



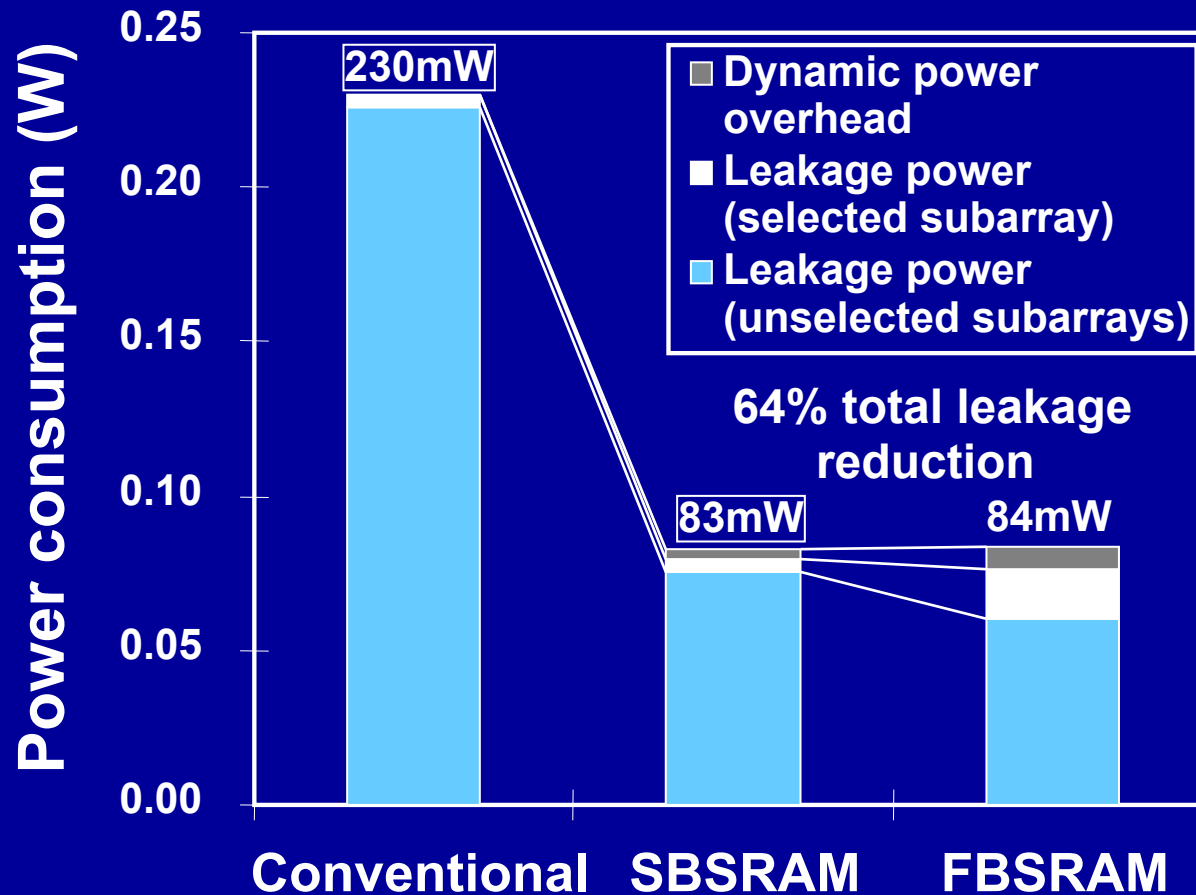
V_{DD}

0V _____ 0.5V
Active Standby

$V_T=350\text{mV}$

- **SBSRAM (DRG) has been proven with Si measurements**
- **Dynamic VDD, RBB SRAM have fundamental design issues**
- **MEDICI: gate/BTBT leakage is also modeled**

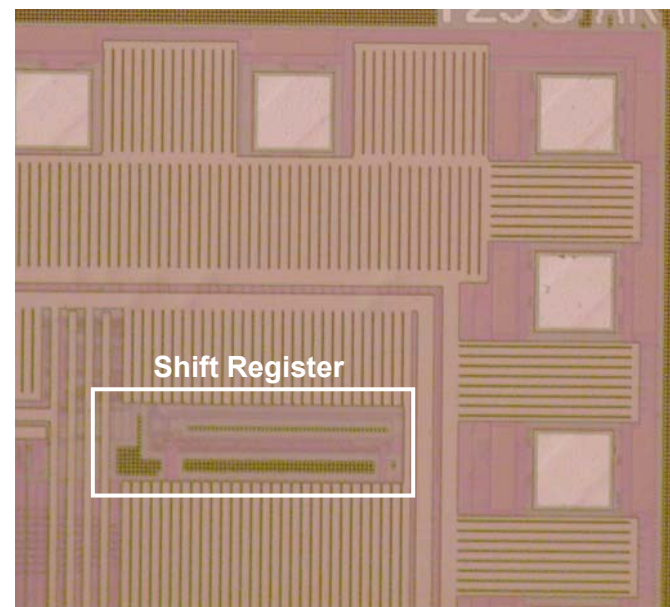
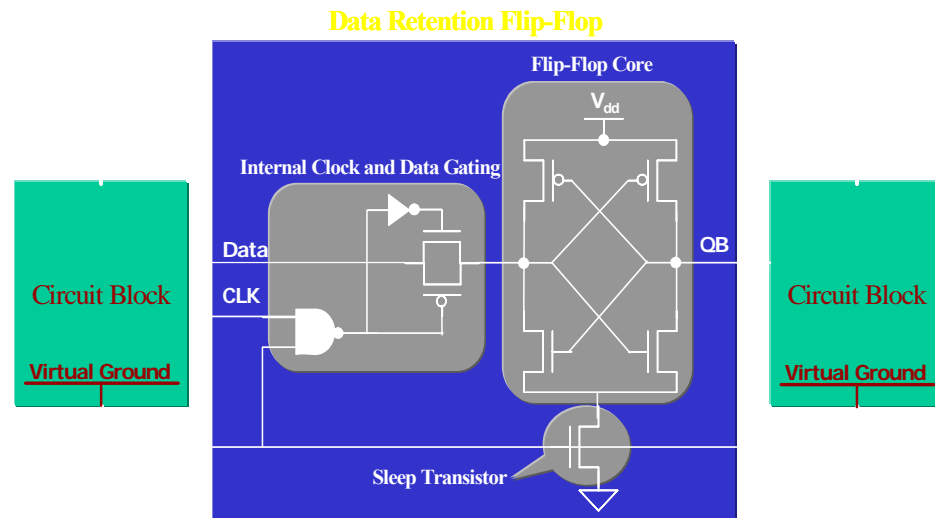
32KB Cache Total Leakage Reduction



- **SBSRAM and FBSRAM are designed to give iso-leakage savings**
- **64% total leakage reduction including overhead**

Another Application: Data Retention Flip-Flop

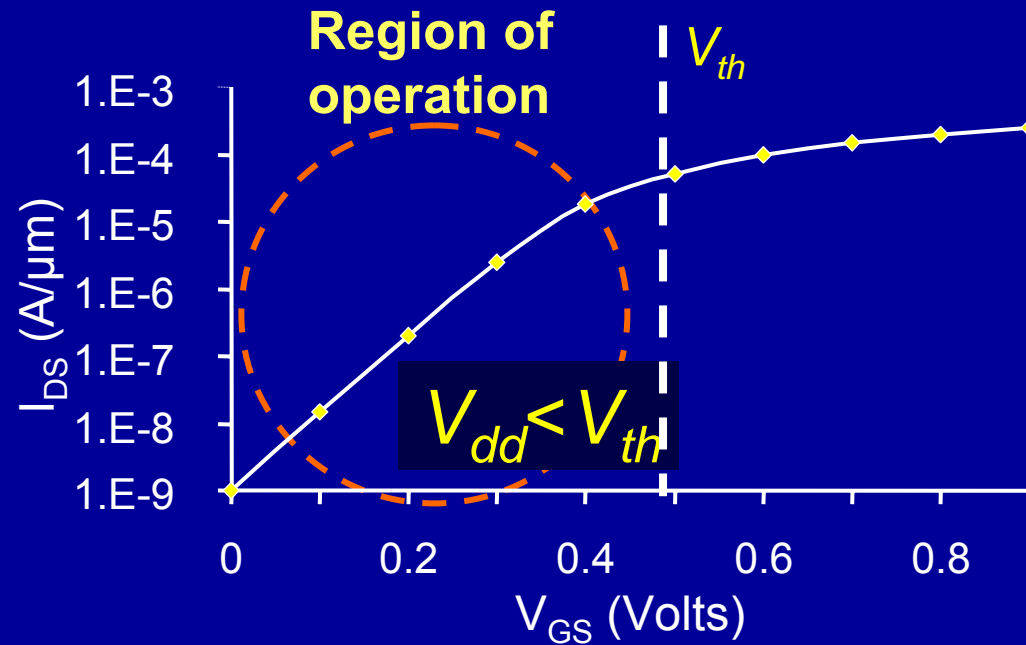
- Cross-coupled inverters are cores of any flip-flops
- Cross-coupled inverters retain data under gated ground
- Data and clock gating is required to preserve data
- Successful fabrication and test:
 - 16-bit shift-register based on our data-retention FF



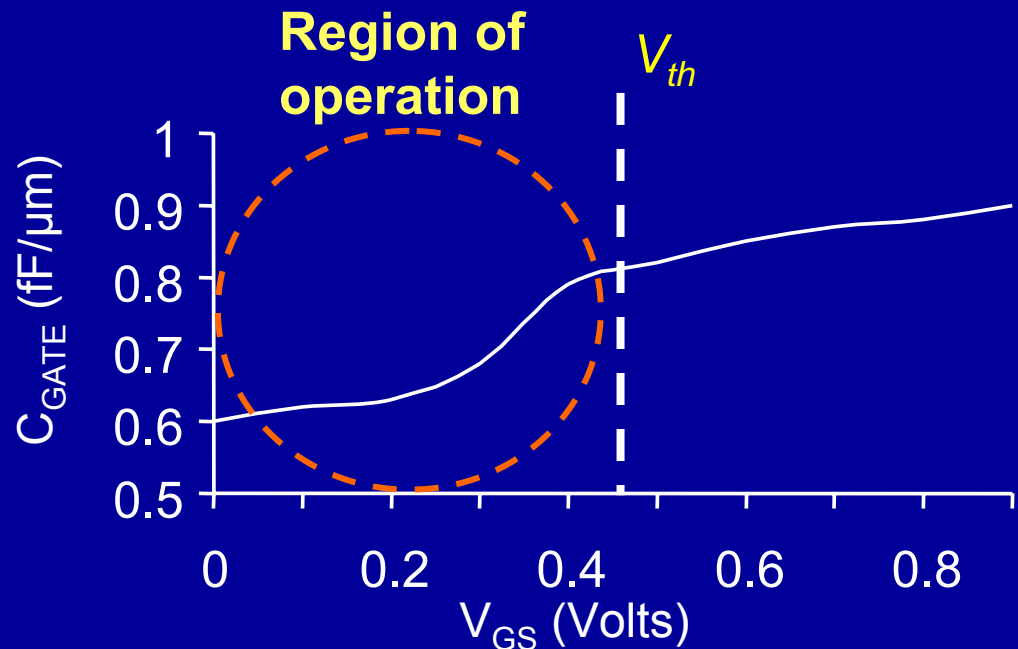
**40% power reduction by
enabling power-down mode**

Computing with Leakage for Ultralow Power: Digital Subthreshold Logic

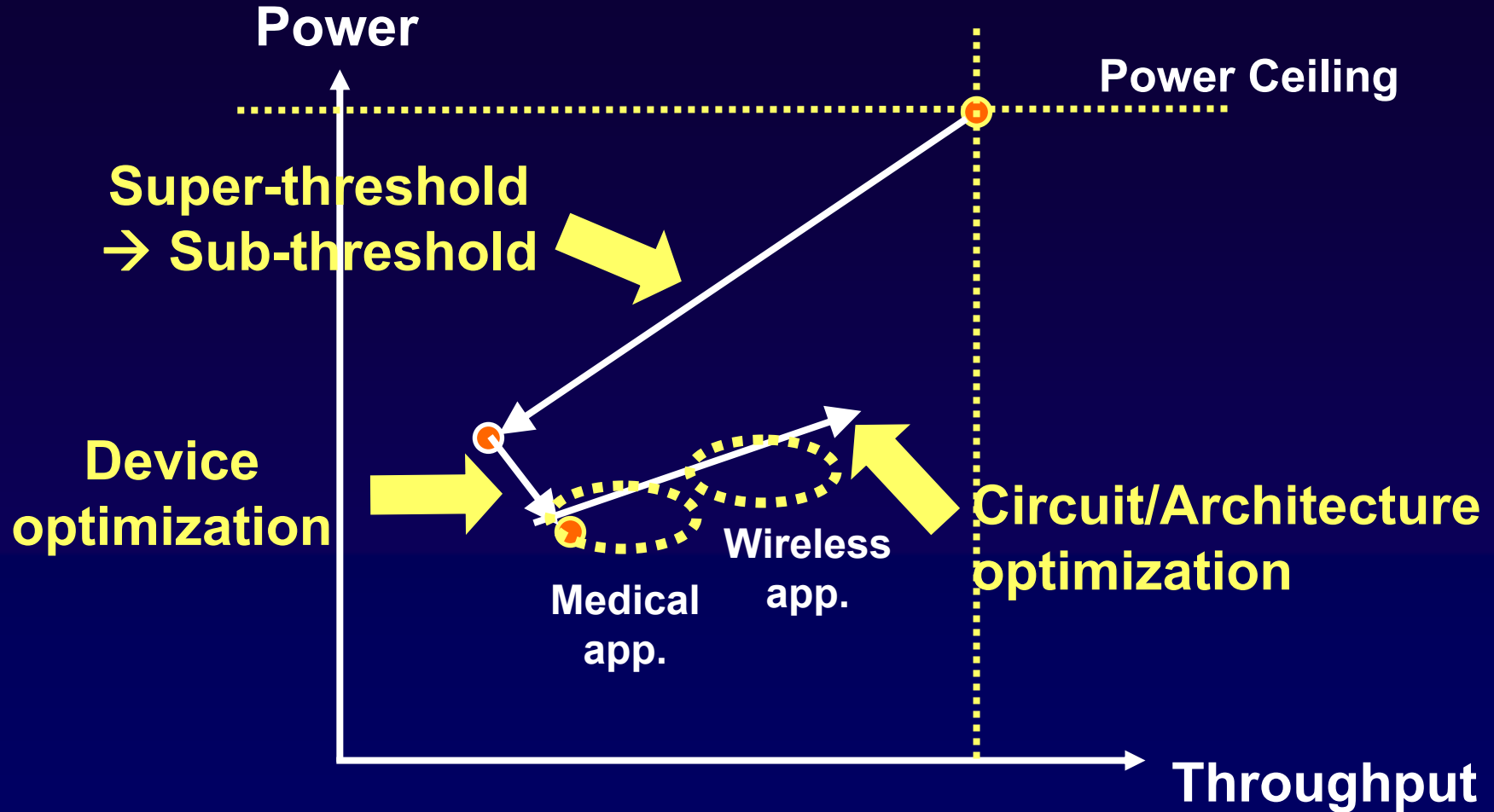
Subthreshold Operation



$$C_{GATE} < C_{OX}$$



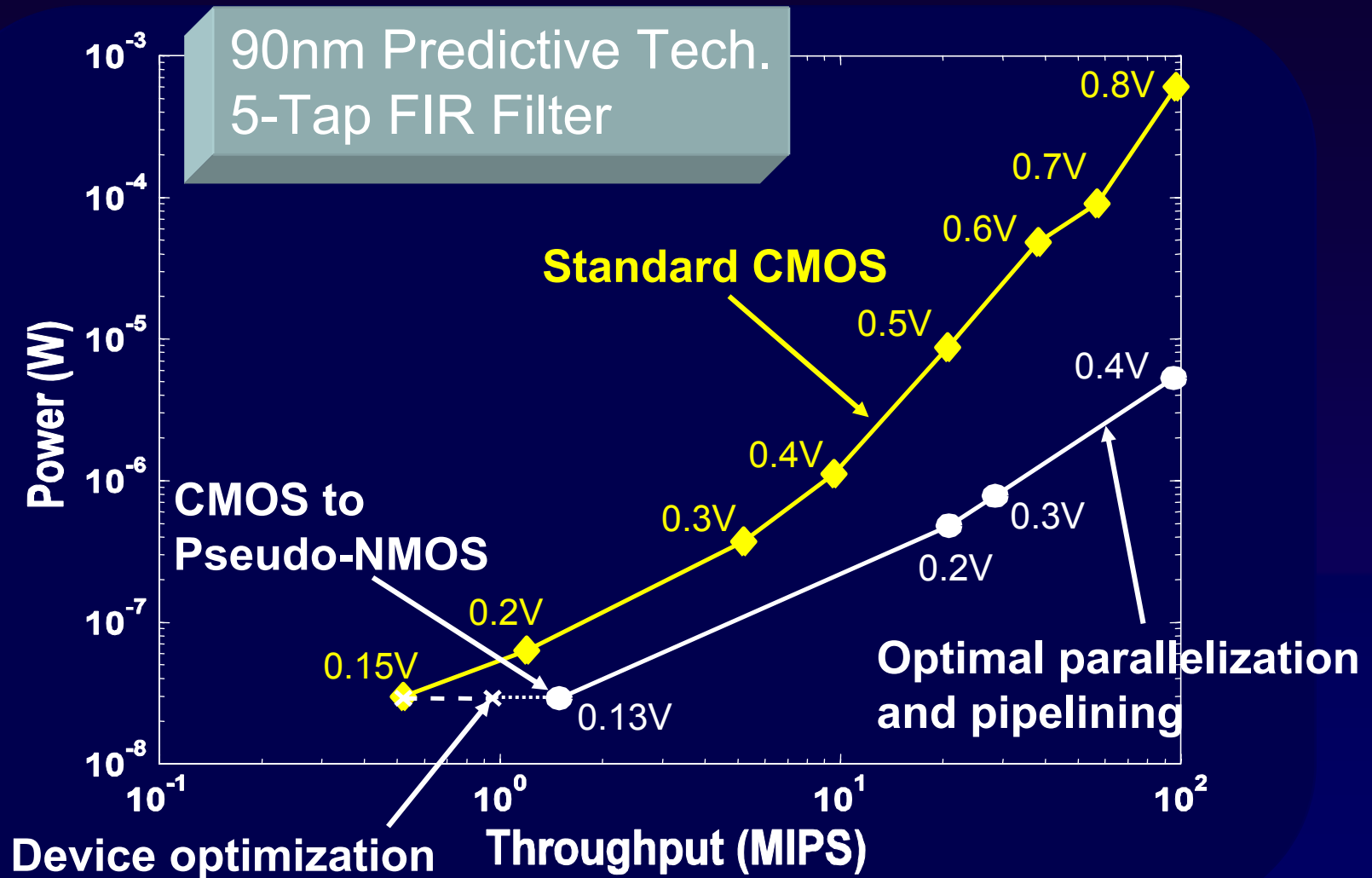
Computing Using Leakage Current



Dev/Cir/Arch co-optimization is necessary

Switching back-and-forth between sup. and sub. operations

Dev/Cir/Arc Co-design: Summary



Under review, TVLSI

Conclusions

- Power considerations (both dynamic and leakage) are very important for scaled technologies
 - Leakage control techniques are becoming essential!
 - Leakage problem is expected in other variations of Si technologies
 - One can effectively use some of the leakage control circuits for testability enhancement
- An integrated approach to design – device/circuit/arch. – is essential for an optimized design
- Subthreshold leakage for computing – ultralow power

Questions and Answers