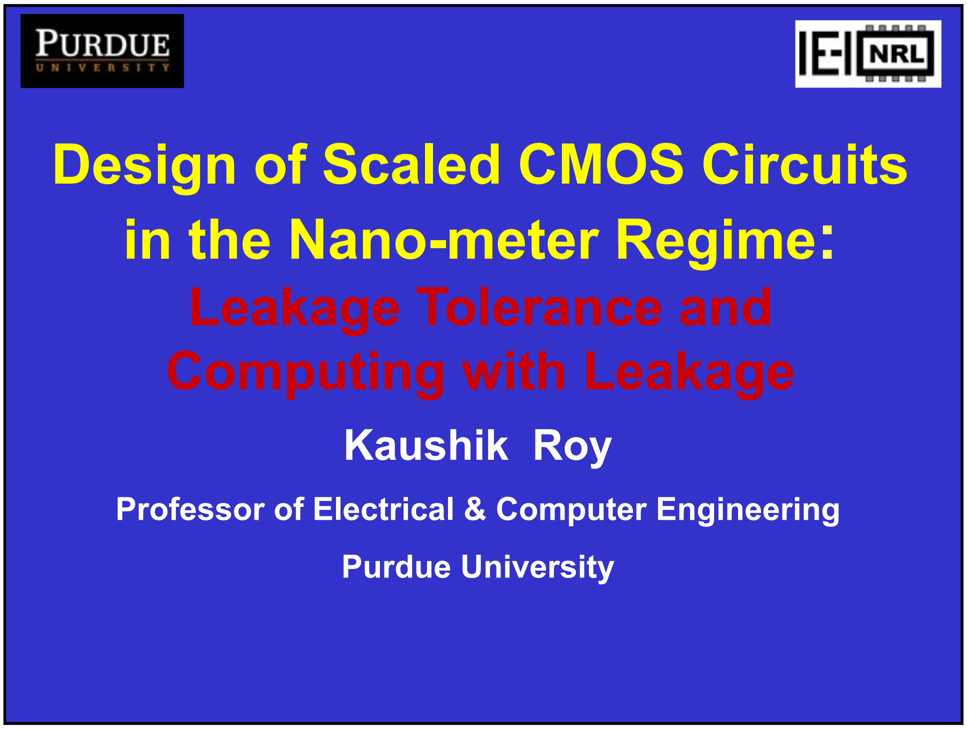
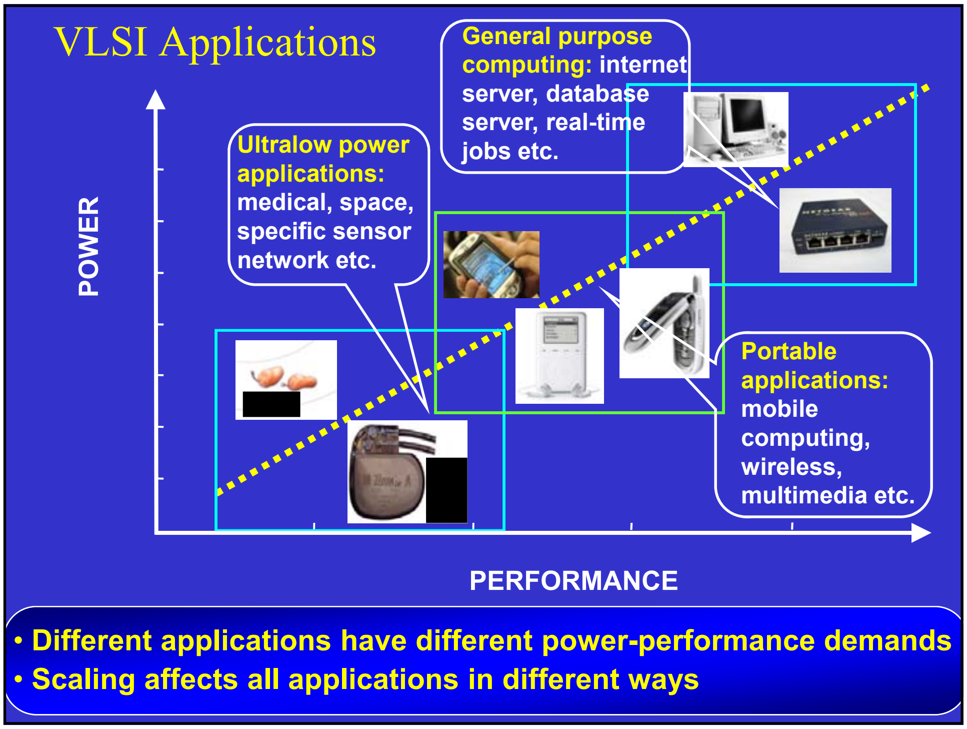


Design of Scaled CMOS Circuits in the Nano-meter Regime: Leakage Tolerance and Computing with Leakage

Kaushik Roy

Professor of Electrical & Computer Engineering

Purdue University



VLSI Applications

[redacted]

[redacted]

General purpose computing: internet server, database server, real-time jobs etc.

Ultralow power applications: medical, space, specific sensor network etc.

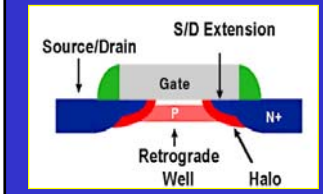
Portable applications: mobile computing, wireless, multimedia etc.

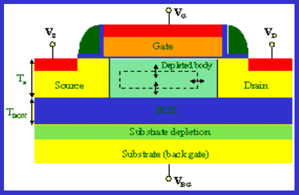
- Different applications have different power-performance demands
- Scaling affects all applications in different ways

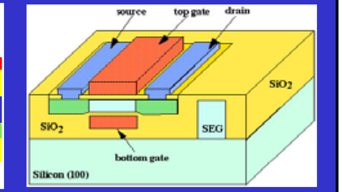
Challenges ahead …

in Si nanometer regime

Challenge No. 1: Device Scaling







Bulk MOS

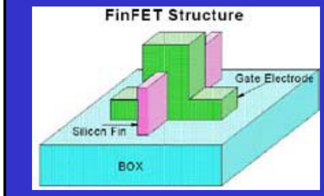
Retograde Well, Halo
Strained Channel

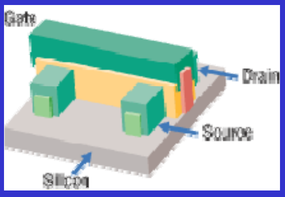
UTB-SOI MOS

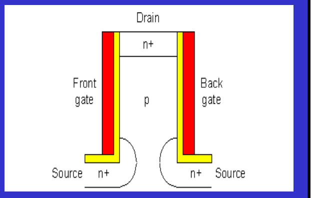
Fully-depleted ultra-thin body
Ground-plane

DG-SOI MOS

Planar double-gate structure
Independent gate control







FinFET

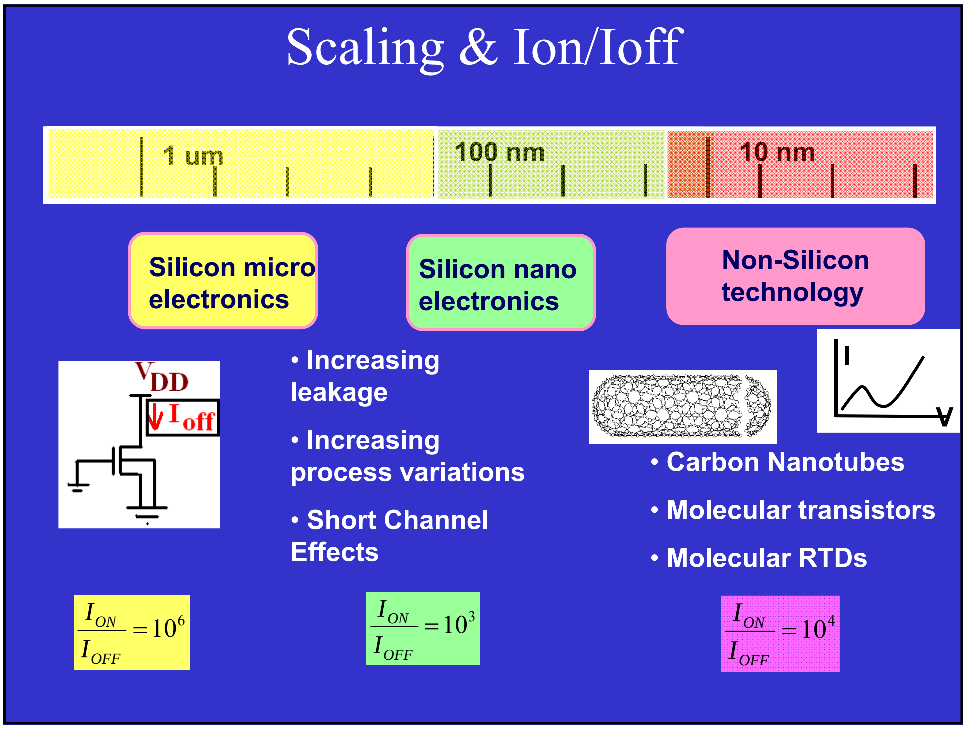
Quasi-planar DG structure
Most promising device

Tri-gate

Quasi-planar with 3 gates
Better area efficiency

Vertical MOS

Conduction normal to plane
Difficult to fabricate



Scaling & Ion/Ioff

Silicon micro electronics

Silicon nano electronics

Non-Silicon technology

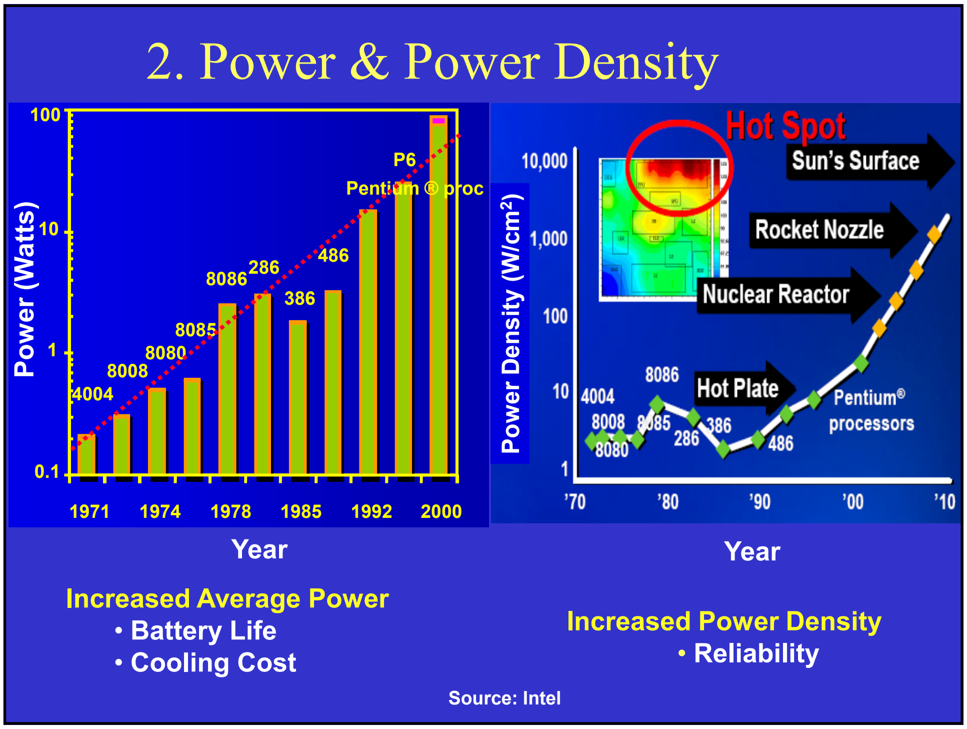
- Increasing leakage
- Increasing process variations
- Short Channel Effects

- Carbon Nanotubes
- Molecular transistors
- Molecular RTDs

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

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

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



2. Power & Power Density

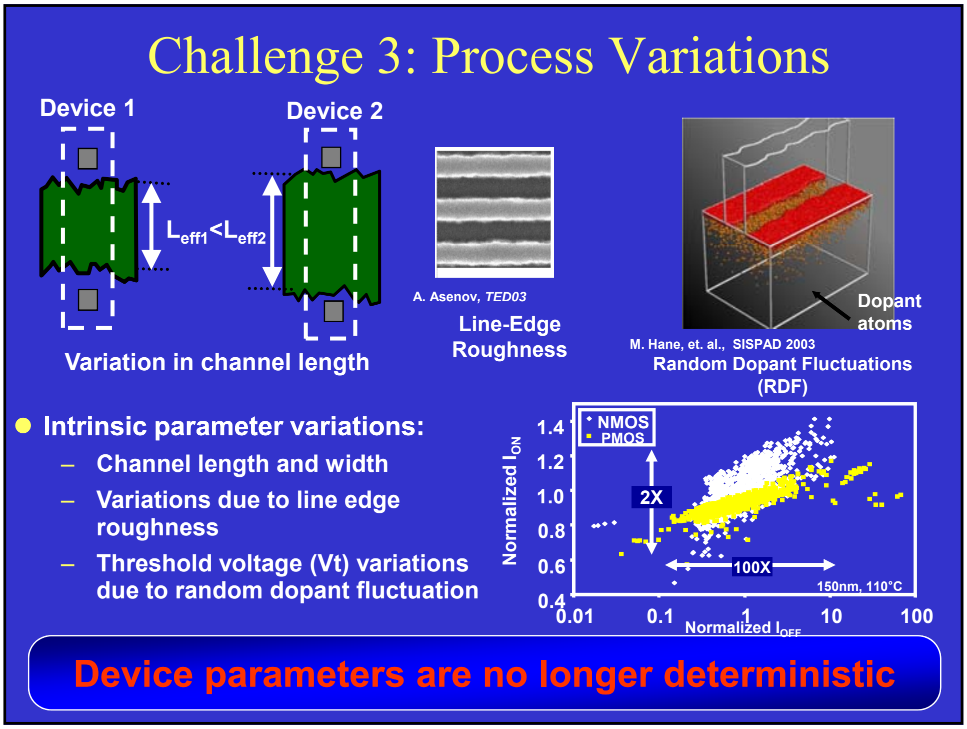
Year

Year

Increased Average Power
- Battery Life
- Cooling Cost

Increased Power Density
- Reliability

Source: Intel



Challenge 3: Process Variations

Variation in channel length

A. Asenov, *TED03*

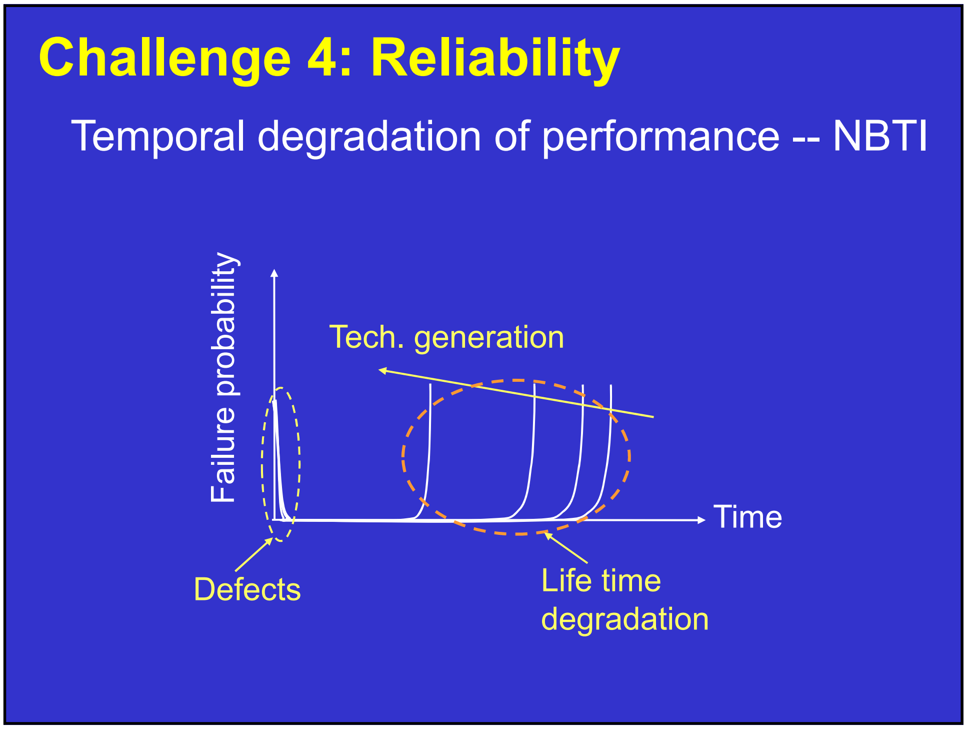
Line-Edge Roughness

M. Hane, et. al., SISPAD 2003

Random Dopant Fluctuations (RDF)

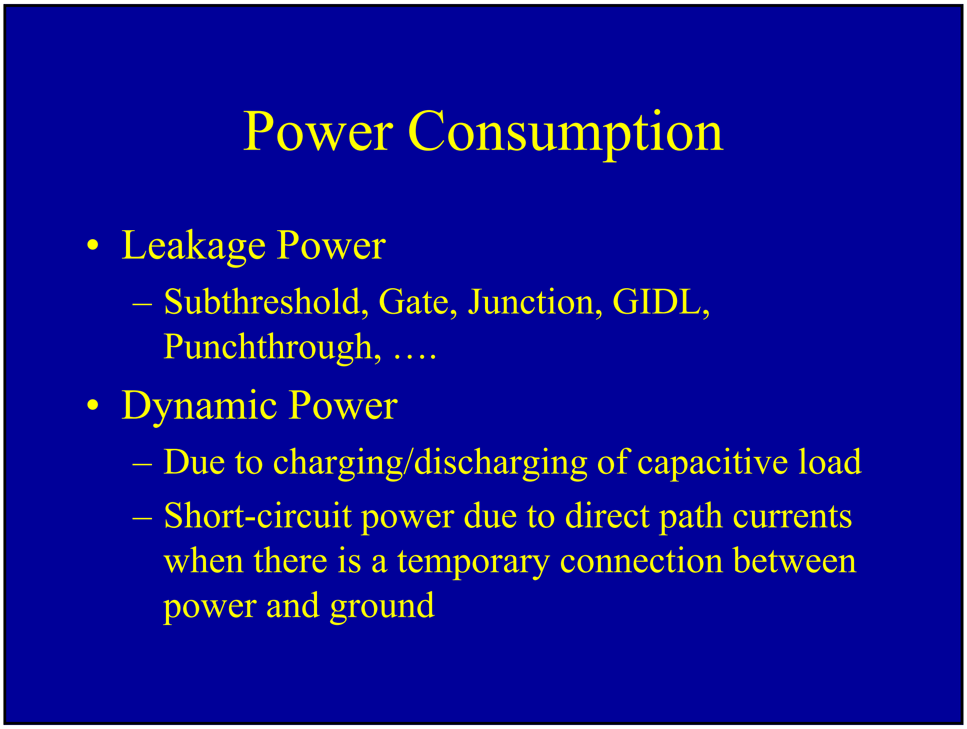
- Intrinsic parameter variations:
  - Channel length and width
  - Variations due to line edge roughness
  - Threshold voltage (Vt) variations due to random dopant fluctuation

Device parameters are no longer deterministic



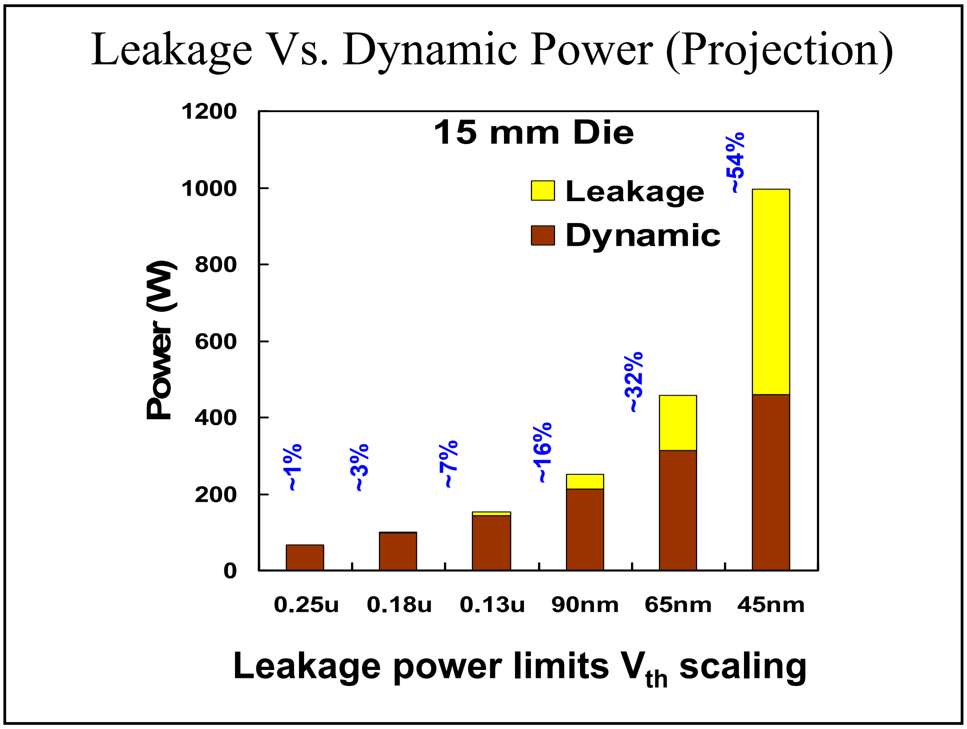
Challenge 4: Reliability

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



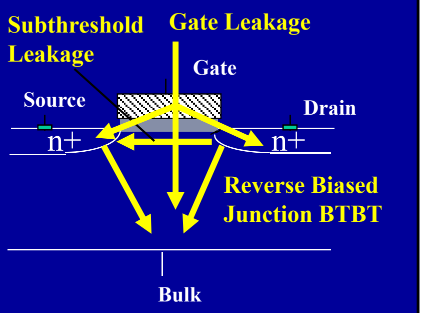
Leakage Vs. Dynamic Power (Projection)

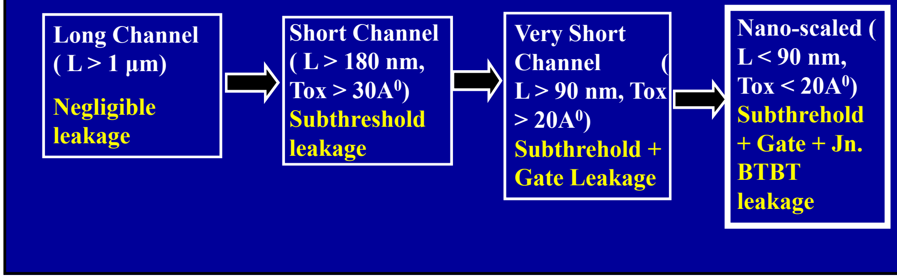
Leakage power limits V_{th} scaling

Leakage Power

Scaling and Other Leakage Components

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





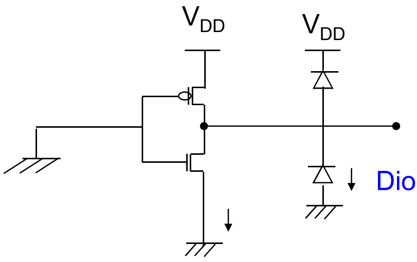
Long Channel (L > 1 µm)
Negligible leakage

Short Channel (L > 180 nm, Tox > $30A^0$)
Subthreshold leakage

Very Short Channel (L > 90 nm, Tox > $20A^0$)
Subthrehold + Gate Leakage

Nano-scaled (L < 90 nm, Tox < $20A^0$)
Subthrehold + Gate + Jn. BTBT leakage

Leakage Power Consumption



$V_{out} = V_{DD}$

Diode leakage

$$I_O = i_s(e^{Vq/kT} - 1)$$

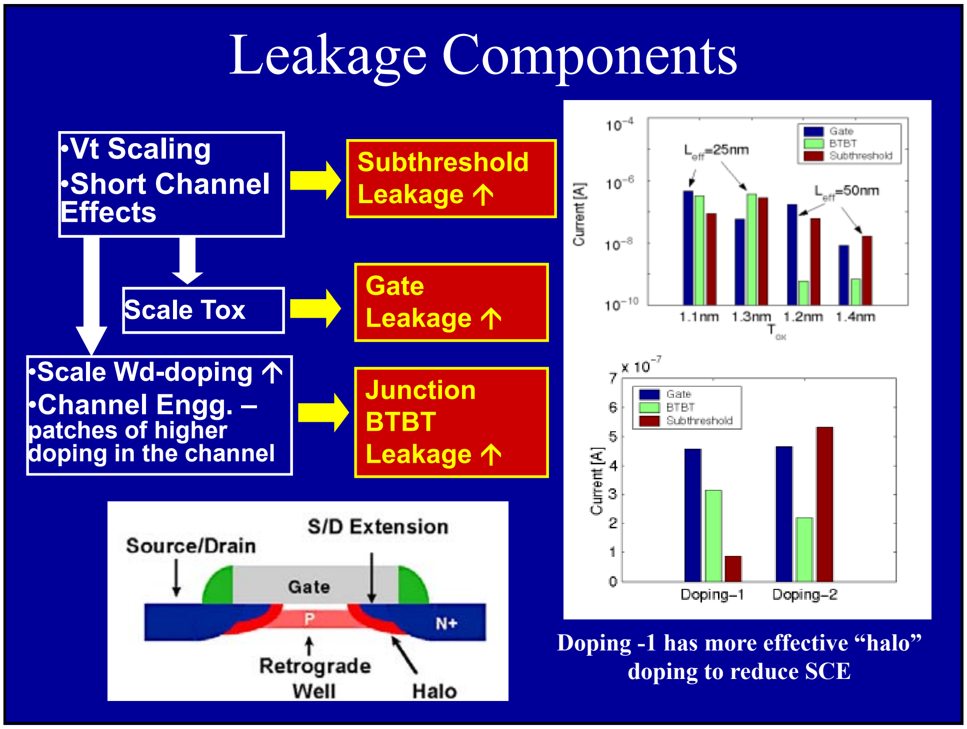
Sub-threshold leakage

$$I_D = K \cdot e^{(V_{gs} - V_t)q/nkT}(1 - e^{V_{ds}q/kT})$$

$P_{static} = I_{leakage} \cdot V_{DD}$

Diode Leakage

- Leakage current through the reverse biased diode junctions
- For typical devices it is between 0.1nA - 0.5nA at room temperature
- For a die with 1 million devices operated at 5 V, this results in 0.5mW power consumption → not much
- Junction leakage current is caused by thermally generated carriers -> therefore is a strong function of temperature
- More important is sub-threshold leakage, gate leakage, and Junction BTBT leakage



Leakage Components

- Vt Scaling
- Short Channel Effects

Subthreshold Leakage ↑

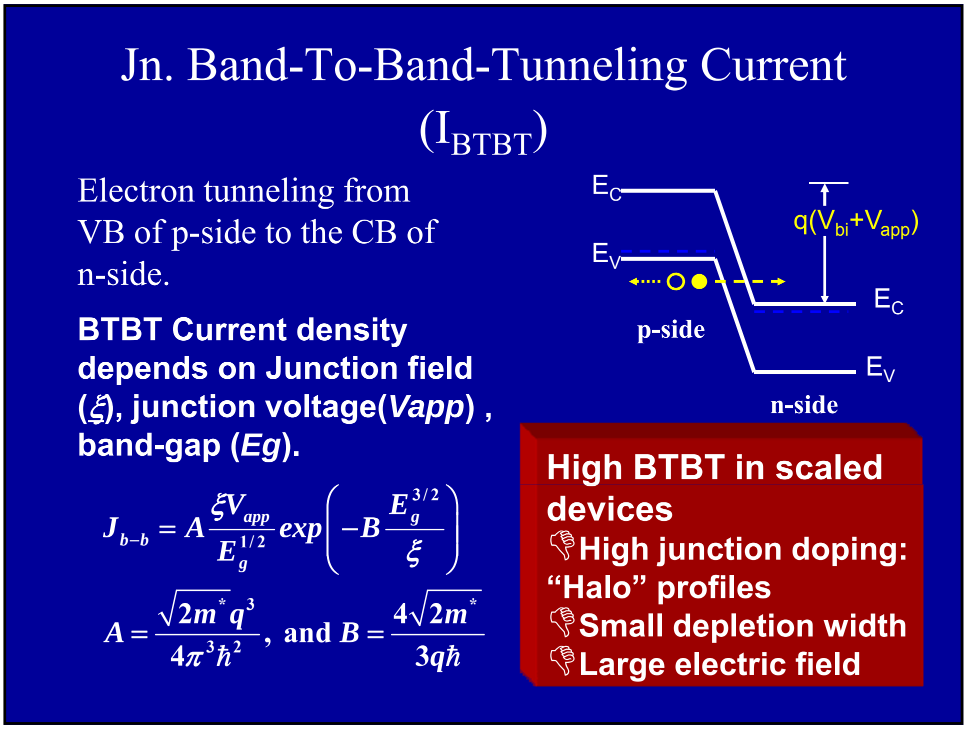
Scale Tox

Gate Leakage ↑

- Scale Wd-doping ↑
- Channel Engg. – patches of higher doping in the channel

Junction BTBT Leakage ↑

Doping -1 has more effective "halo" doping to reduce SCE



Jn. Band-To-Band-Tunneling Current (I_{BTBT})

Electron tunneling from VB of p-side to the CB of n-side.

BTBT Current density depends on Junction field (ξ), junction voltage(*Vapp*) , band-gap (*Eg*).

$$J_{b-b} = A\frac{\xi V_{app}}{E_g^{1/2}} exp\left(-B\frac{E_g^{3/2}}{\xi}\right)$$

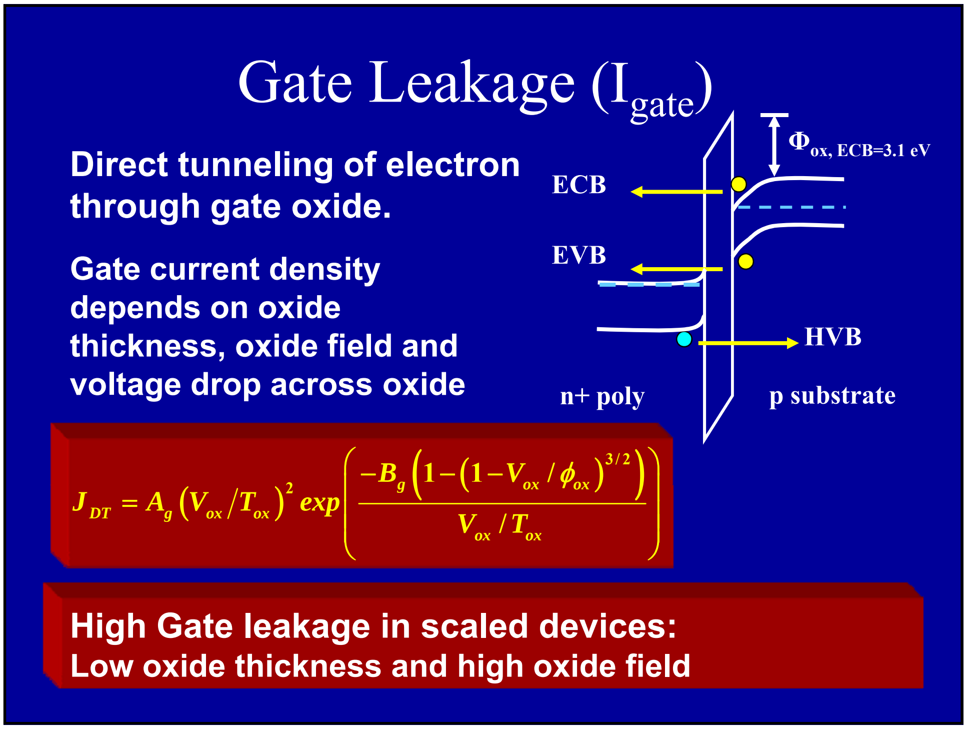
$$A = \frac{\sqrt{2m^* q^3}}{4\pi^3\hbar^2}, \text{ and } B = \frac{4\sqrt{2m^*}}{3q\hbar}$$

High BTBT in scaled devices

👎 **High junction doping: "Halo" profiles**

👎 **Small depletion width**

👎 **Large electric field**



Gate Leakage (I_{gate})

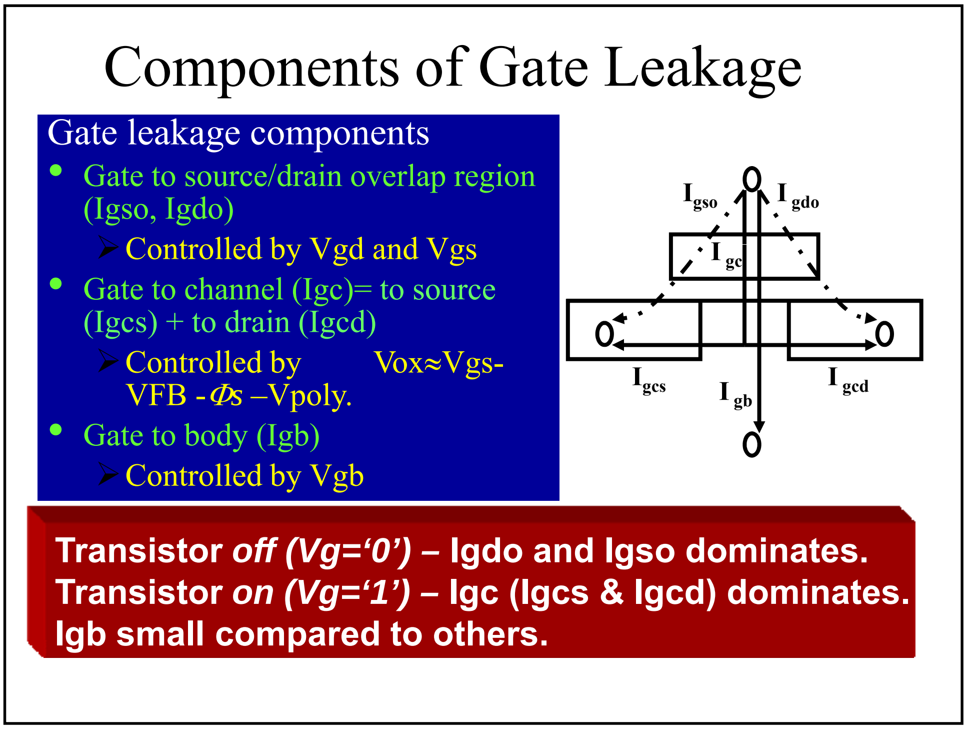
Direct tunneling of electron through gate oxide.

Gate current density depends on oxide thickness, oxide field and voltage drop across oxide

$\Phi_{ox, ECB=3.1 eV}$
ECB
EVB
HVB
n+ poly
p substrate

$$J_{DT} = A_g (V_{ox}/T_{ox})^2 \, exp\left(\frac{-B_g\left(1-(1-V_{ox}/\phi_{ox})^{3/2}\right)}{V_{ox}/T_{ox}}\right)$$

High Gate leakage in scaled devices:
Low oxide thickness and high oxide field



Components of Gate Leakage

Gate leakage components

- Gate to source/drain overlap region (Igso, Igdo)
  - ➢ Controlled by Vgd and Vgs
- Gate to channel (Igc)= to source (Igcs) + to drain (Igcd)
  - ➢ Controlled by Vox≈Vgs-VFB -Φs –Vpoly.
- Gate to body (Igb)
  - ➢ Controlled by Vgb

I_{gso}
I_{gdo}
I_{gc}
I_{gcs}
I_{gb}
I_{gcd}

Transistor *off (Vg='0')* – Igdo and Igso dominates.
Transistor *on (Vg='1')* – Igc (Igcs & Igcd) dominates.
Igb small compared to others.

Subthreshold leakage (I_{sub})

Exponentially dependence on Vgs and Vth.

$$I_{sub} = \frac{w_{eff}}{L_{eff}} \mu \sqrt{\frac{q \varepsilon_{si} N_{cheff}}{2\Phi_s}} v_T{}^2 exp\left(\frac{V_{gs} - V_{th}}{n v_T}\right)\left(1 - exp\left(\frac{-V_{ds}}{v_T}\right)\right)$$

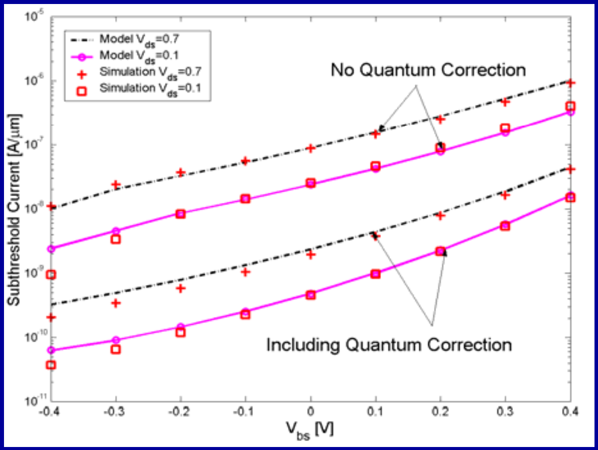
Vth modulation

- ✓ **Short channel Effect – Vth reduction due to**
  - ➢ **Increase in Vds (DIBL),**
  - ➢ **Reduction in Channel Length (Vth roll off).**
- ✓ **Body effect – negative Vbs increases Vth.**
- ✓ **Quantum confinement effect – increases Vth**

$$V_{th} = V_{FB} + \left(\Phi_{s0} - \Delta\Phi_s\right) + \gamma\sqrt{\Phi_{s0} - V_{bs}}\left(1 - \lambda \frac{W_{dm}}{L_{eff}}\right) + V_{nce} + V_{QM}$$

$$\Delta\Phi_s = \left[2(V_{bi} - \phi_{S0}) + V_{ds}\right] \times \left[e^{-L/2l_c} + 2e^{-L/l_c}\right] \quad and \quad l_C = \sqrt{\left(\varepsilon_{si}/\varepsilon_{ox}\eta\right) T_{ox} W_{dm}}$$

Subthreshold Leakage

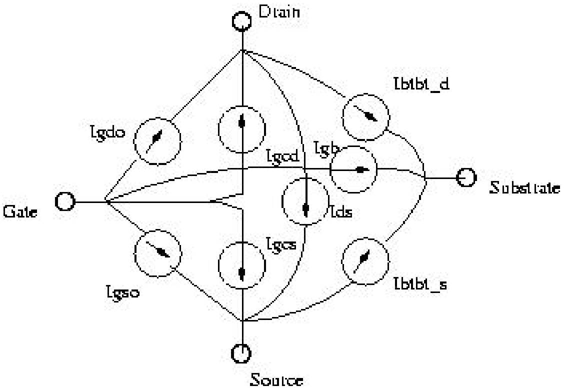


Subthreshold leakage reduces with
✓Negative Vbs, Reduction of Vds
✓Application of Quantum Correction

Total Leakage

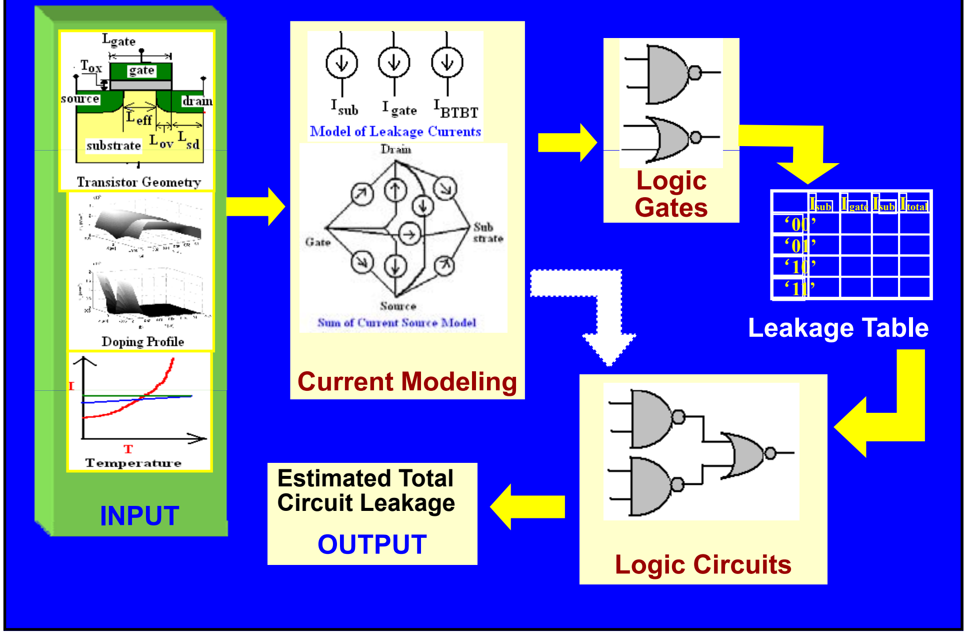
“Sum of Current Source Model”

Voltage Controlled Current Sources describing each leakage comp.



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

Leakage Estimation Method



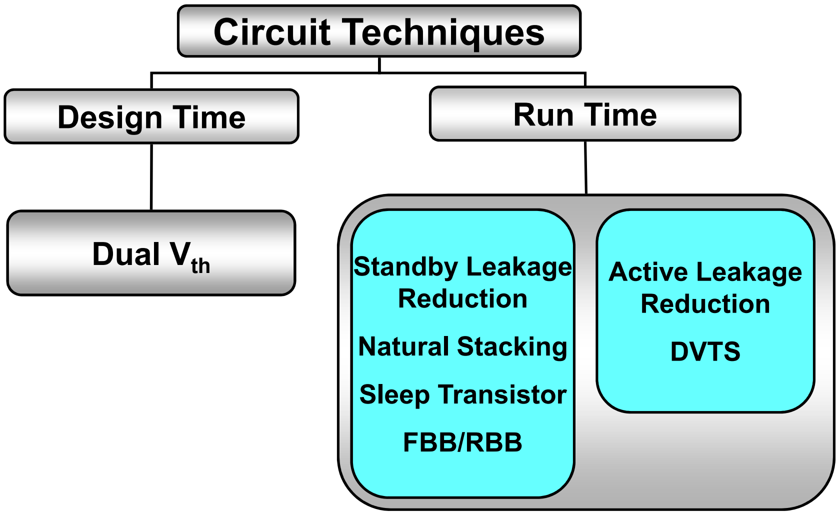
	I_{sub}	I_{gate}	I_{sub}	I_{total}
‘00’				
‘01’				
‘10’				
‘11’				

Low-Vdd Low-Vt Design

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

Leakage control techniques

Leakage Reduction (Logic & Memory)



Circuit Techniques

Design Time

Dual V_{th}

Run Time

Standby Leakage Reduction

Natural Stacking

Sleep Transistor

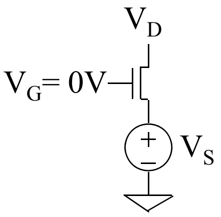
FBB/RBB

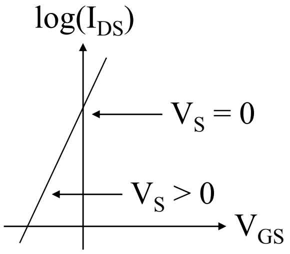
Active Leakage Reduction

DVTS

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

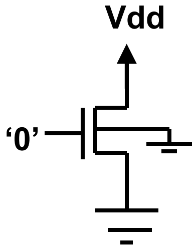


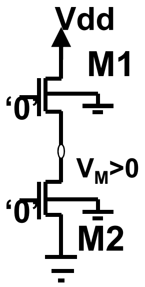


$V_{DS} \downarrow \Rightarrow V_T \uparrow$

$V_S \uparrow \Rightarrow V_T \uparrow$

Leakage Control: Stacking

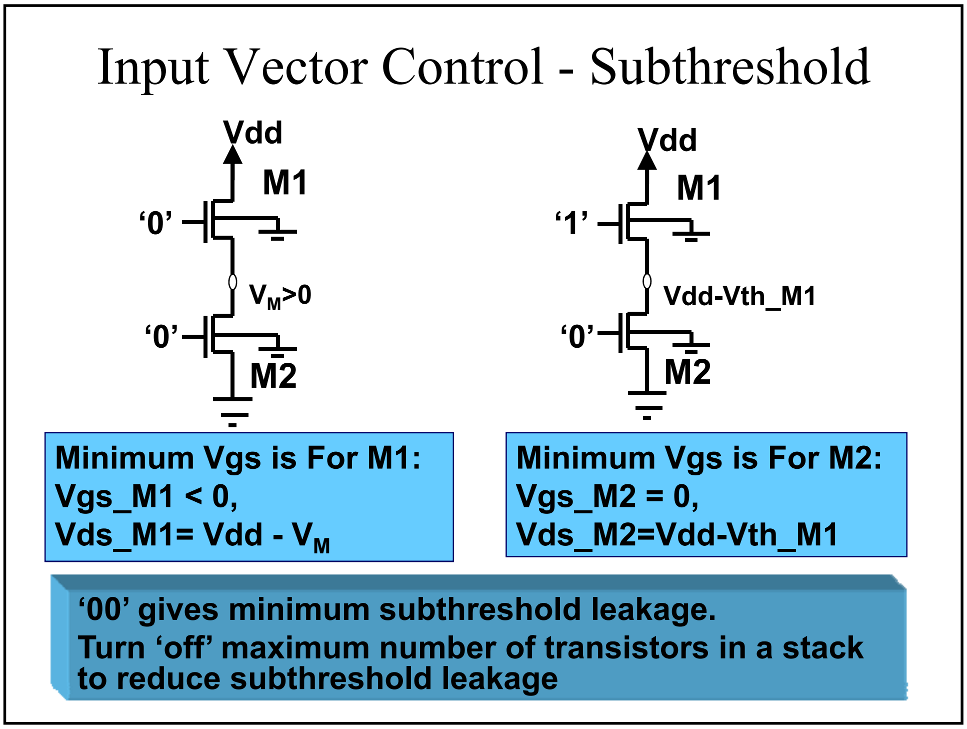




Vgs=0,Vbs=0,Vds=Vdd

✓Negative Vgs,
✓Negative Vbs- More Body effect,
✓Reduced Vds-Less DIBL
2-T stack has lower subthreshold leakage

For M1:
Vgs =-V_M< 0,Vbs =-V_M<0,
Vds = Vdd-V_M<Vdd
For M2:
Vgs =0,Vbs =0,
Vds = V_M < Vdd

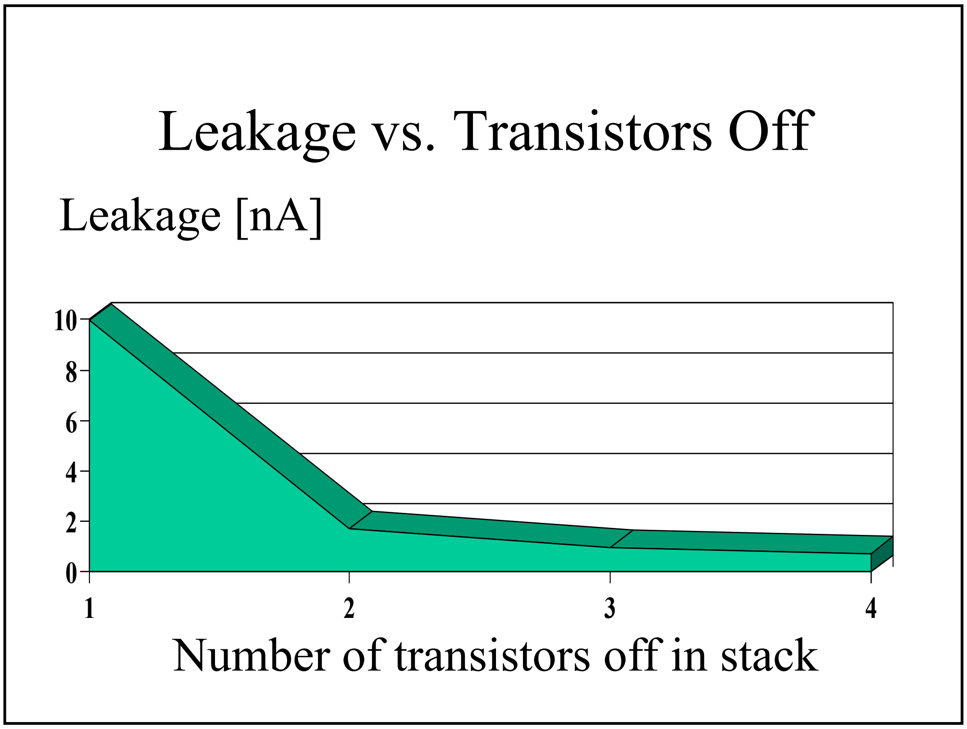


Input Vector Control - Subthreshold

Minimum Vgs is For M1:
Vgs_M1 < 0,
Vds_M1= Vdd - V_M

Minimum Vgs is For M2:
Vgs_M2 = 0,
Vds_M2=Vdd-Vth_M1

'00' gives minimum subthreshold leakage.
Turn 'off' maximum number of transistors in a stack to reduce subthreshold leakage



Leakage vs. Transistors Off

Input Vector Control – Gate Leakage

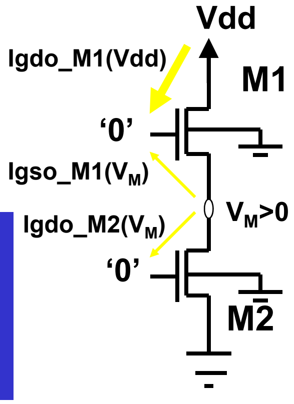
✓Vg=‘0’ – EDT dominates

- ➢Ig = Igdo + Igso

✓Vg=‘1’ – Gate to Channel tunneling is significant

- ➢Ig = Igdo + Igso + Igc

With ‘00’ –
Igdo_M1(Vdd) >>
Igso_M1(V_M) + Igdo_M2(V_M)
Igdo of M1 dominates the total gate current

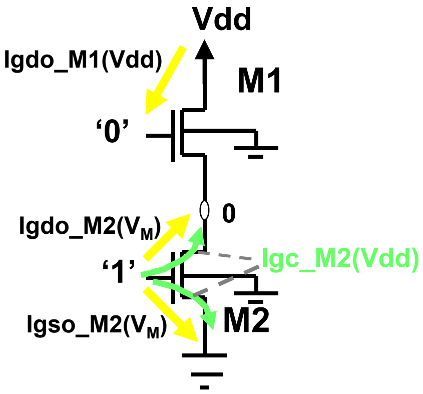


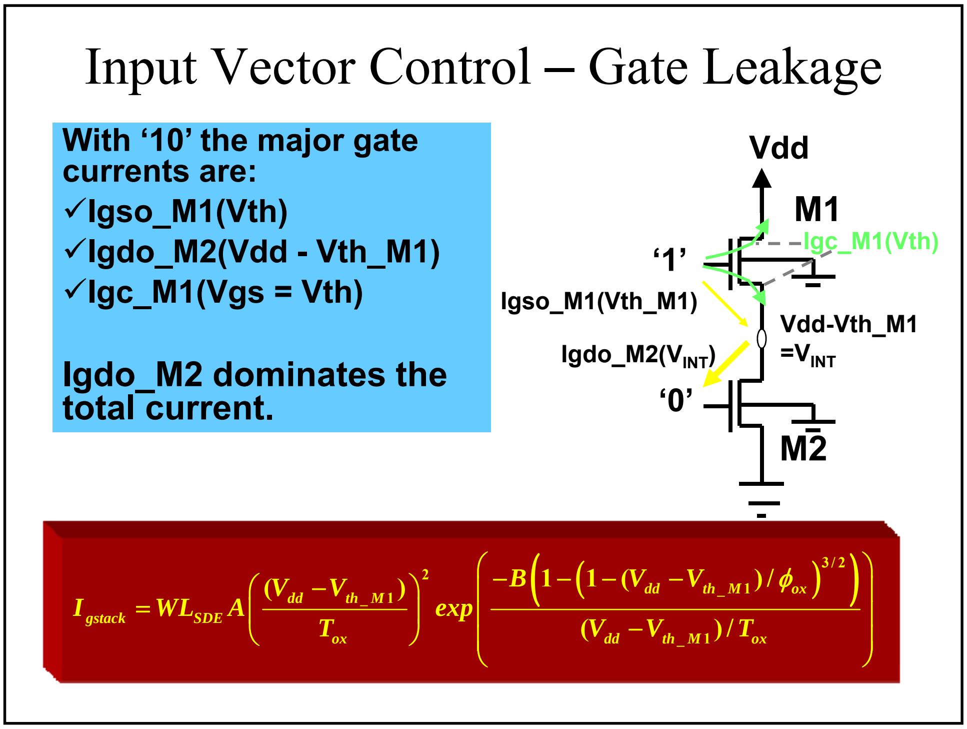
$$I_{gstack} = WL_{SDE} A\left(V_{dd}/T_{ox}\right)^2 exp\left(\frac{-B\left(1-\left(1-V_{dd}/\phi_{ox}\right)^{3/2}\right)}{V_{dd}/T_{ox}}\right)$$

Input Vector Control – Gate Leakage

With ‘01’ –
Igdo_M1(Vdd) is high.
Igdo_M2(Vdd) is high.
Igso_M2(Vdd) is high.
Gate to channel leakage of M2 is controlled by :
Vox_M2 = Vdd - VFB -Φs – Vpoly.

Total gate current is high.





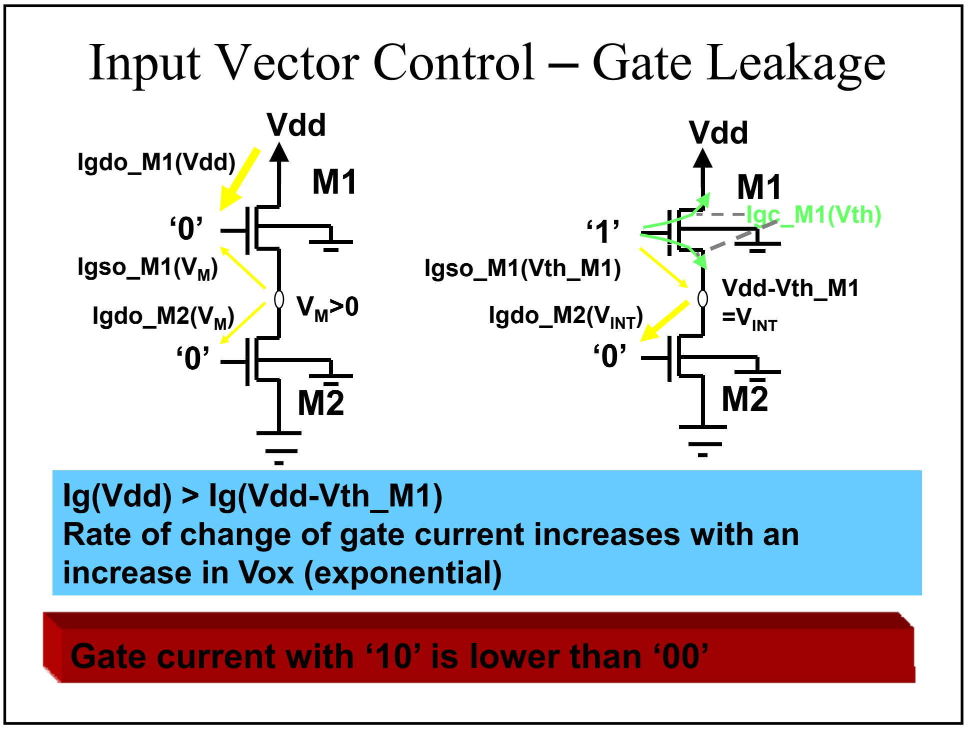
Input Vector Control – Gate Leakage

With ‘10’ the major gate currents are:

- ✓Igso_M1(Vth)
- ✓Igdo_M2(Vdd - Vth_M1)
- ✓Igc_M1(Vgs = Vth)

Igdo_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 – Gate Leakage

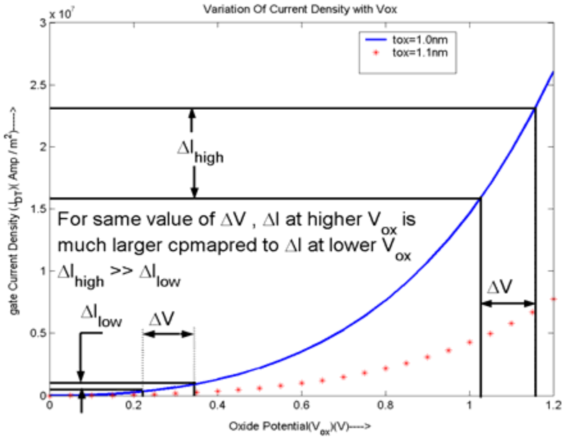
Ig(Vdd) > Ig(Vdd-Vth_M1)
Rate of change of gate current increases with an increase in Vox (exponential)

Gate current with ‘10’ is lower than ‘00’

Gate Leakage

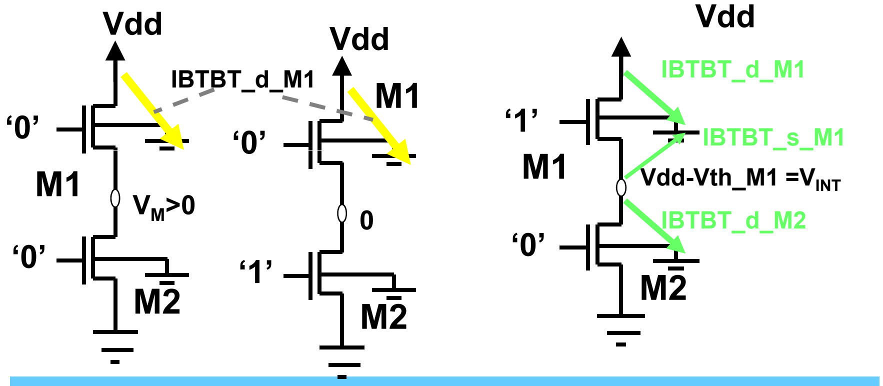
Gate leakage increases with

- ✓ **Increase in Vox**
- ✓ **Reduction in Tox.**



Rate of change of current is higher at higher Vox

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.

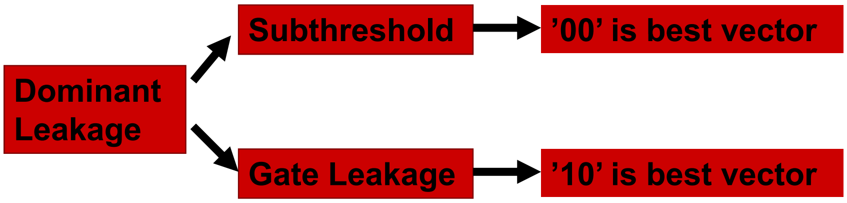
Input Vector Control – Total Leakage

Leakage difference between '10' and '00' =

$$\Delta I_{leakage} = I_{'10'} - I_{'00'}$$

$$= (I_{sub-10} - I_{sub-00}) + (I_{gdo2-10} - I_{gdo1-00}) + (I_{btbt-10} - I_{btbt-00})$$

- ✓ **Isub-10 > Isub-00**
- ✓ **Igdo2-10 < Igdo1-00**
- ✓ **Ibtbt-10 ≥ Ibtbt-00**

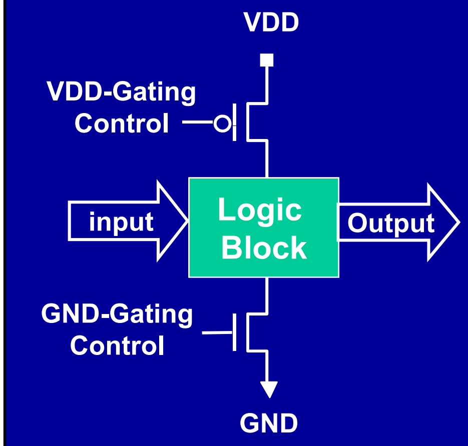


Dominant Leakage

Subthreshold → **'00' is best vector**

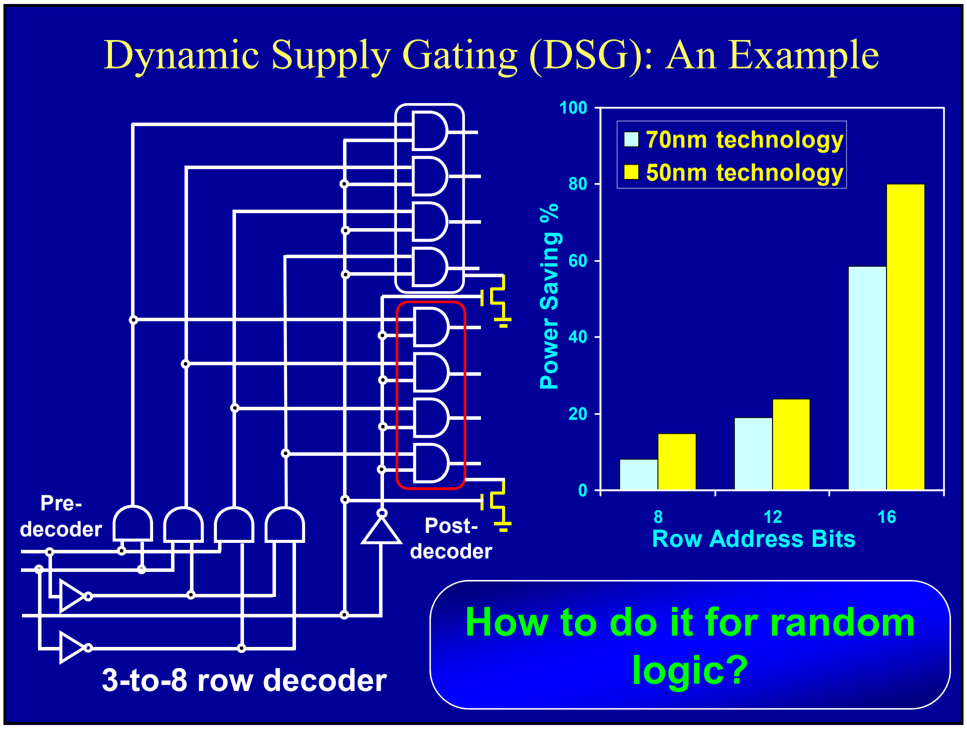
Gate Leakage → **'10' is best vector**

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

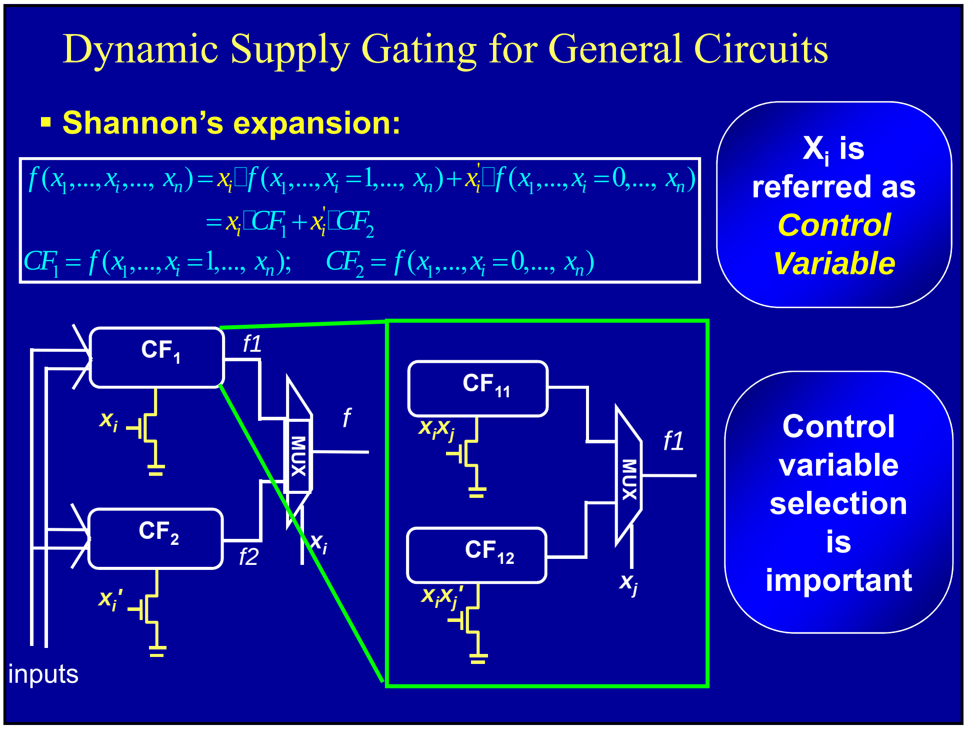
How to use supply gating dynamically in active mode?



Dynamic Supply Gating (DSG): An Example

3-to-8 row decoder

How to do it for random logic?



Dynamic Supply Gating for General Circuits

- **Shannon's expansion:**

$$f(x_1,\ldots,x_i,\ldots,x_n) = x_i \cdot f(x_1,\ldots,x_i=1,\ldots,x_n) + x_i' \cdot f(x_1,\ldots,x_i=0,\ldots,x_n)$$

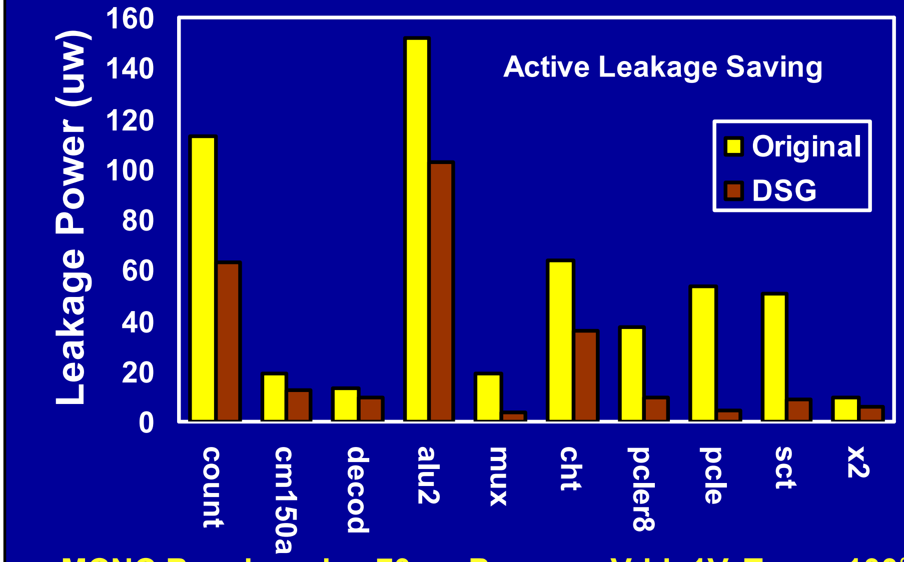
$$= x_i \cdot CF_1 + x_i' \cdot CF_2$$

$$CF_1 = f(x_1,\ldots,x_i=1,\ldots,x_n); \quad CF_2 = f(x_1,\ldots,x_i=0,\ldots,x_n)$$

X_i is referred as *Control Variable*

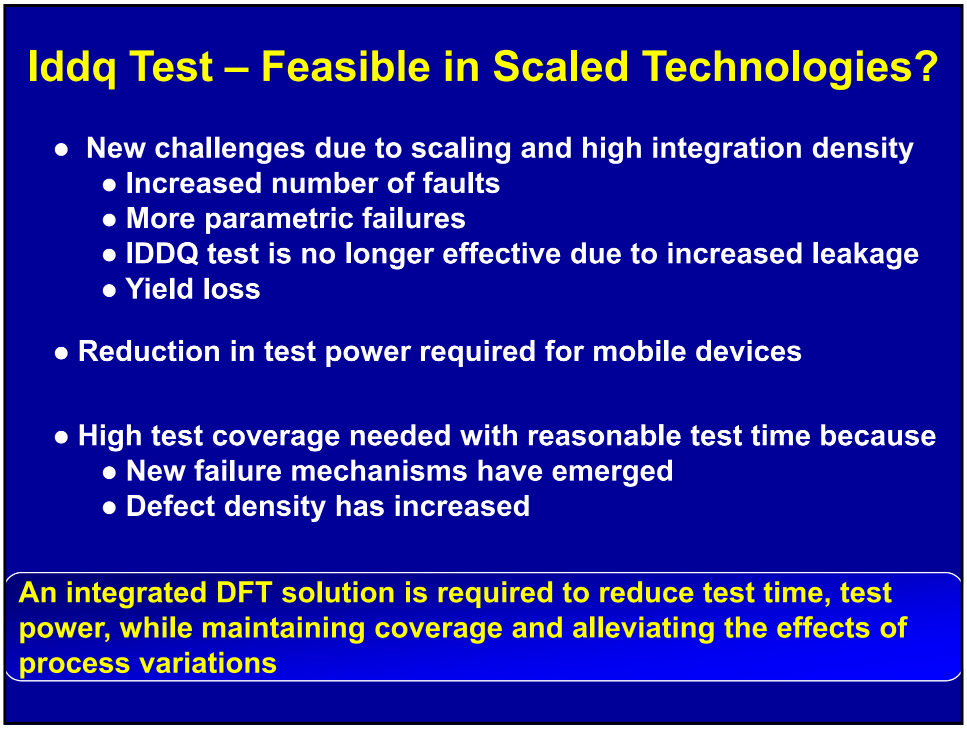
Control variable selection is important

Simulation Results



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

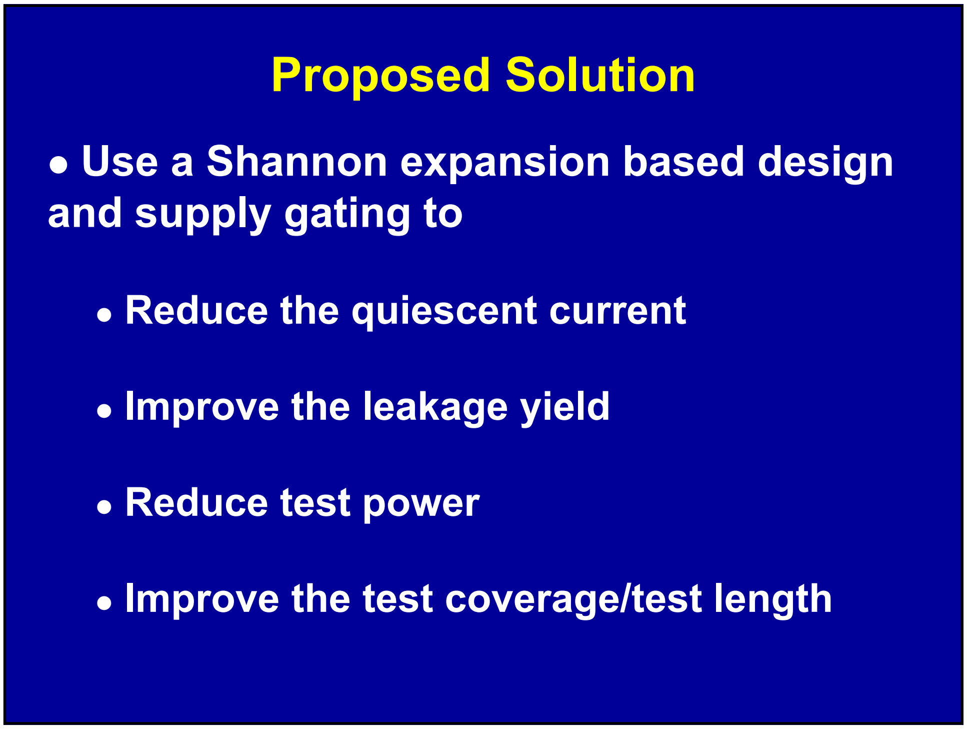
Supply-Gating & Test



Iddq Test – Feasible in Scaled Technologies?

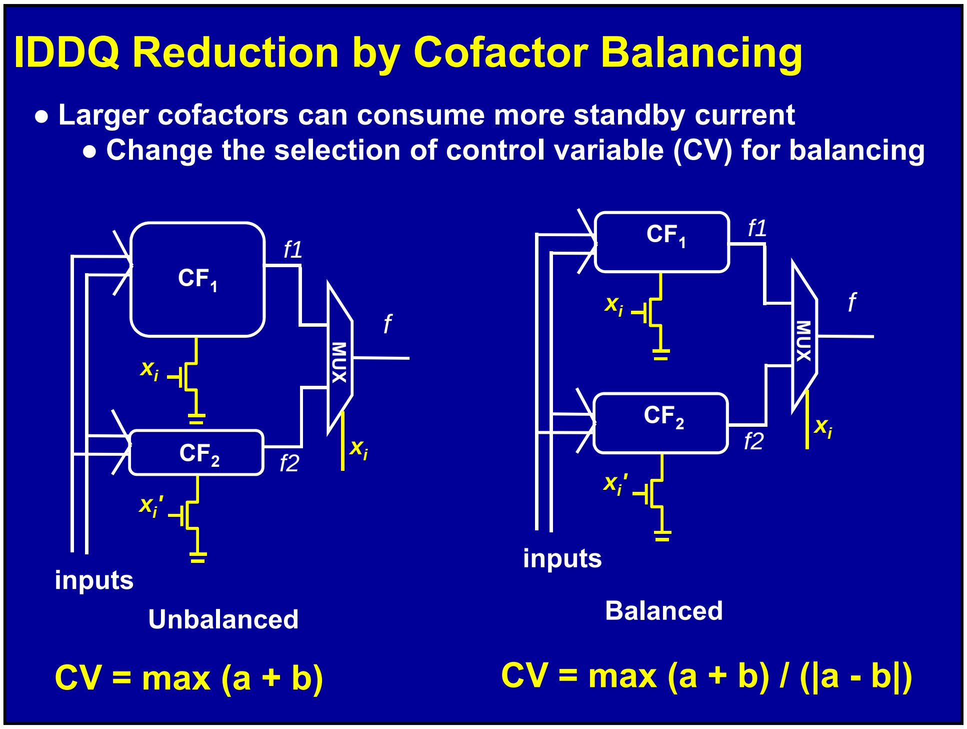
- New challenges due to scaling and high integration density
  - Increased number of faults
  - More parametric failures
  - IDDQ test is no longer effective due to increased leakage
  - Yield loss
- Reduction in test power required for mobile devices
- High test coverage needed with reasonable test time because
  - New failure mechanisms have emerged
  - Defect density has increased

An integrated DFT solution is required to reduce test time, test power, while maintaining coverage and alleviating the effects of process variations



Proposed Solution

- Use a Shannon expansion based design and supply gating to
  - Reduce the quiescent current
  - Improve the leakage yield
  - Reduce test power
  - Improve the test coverage/test length

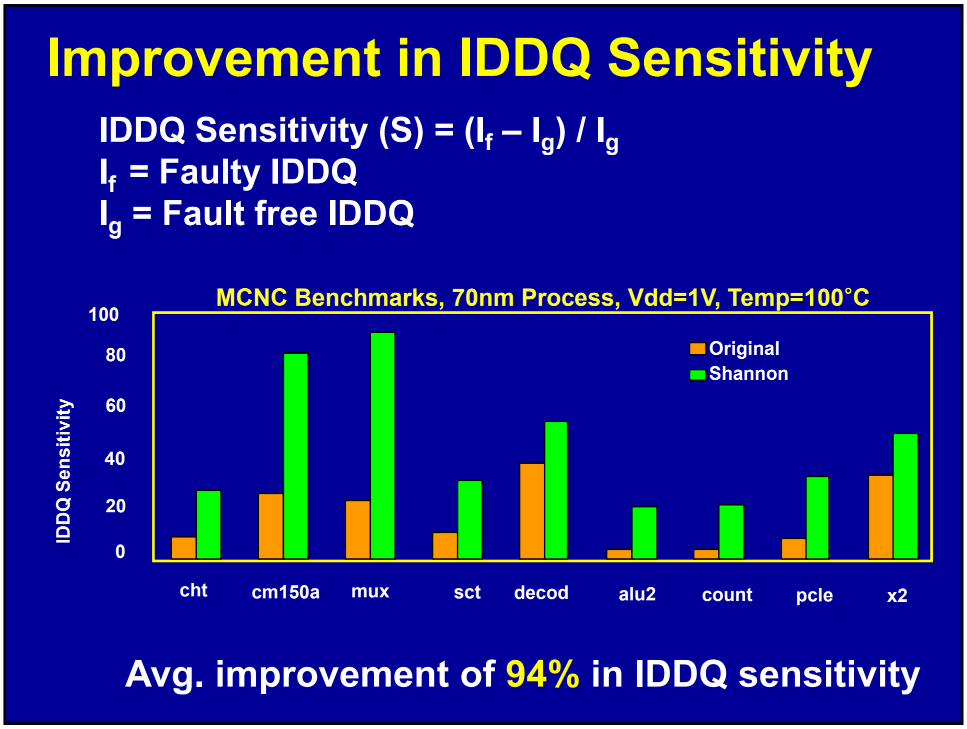


IDDQ Reduction by Cofactor Balancing

- Larger cofactors can consume more standby current
  - Change the selection of control variable (CV) for balancing

CV = max (a + b)

CV = max (a + b) / (|a - b|)



Improvement in IDDQ Sensitivity

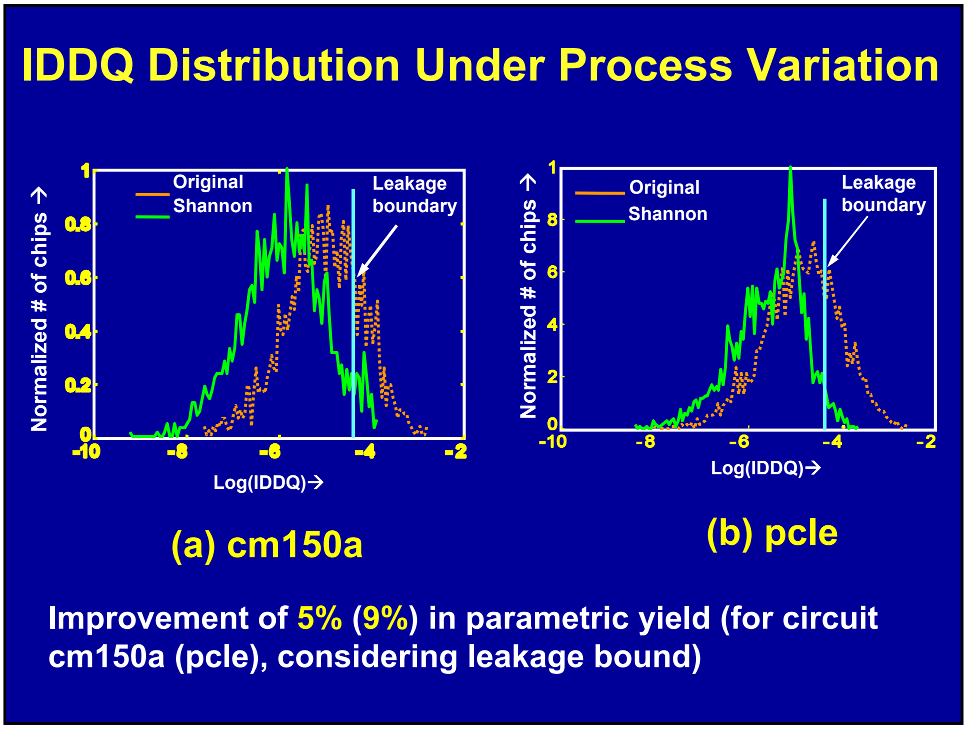
IDDQ Sensitivity (S) = $(I_f - I_g) / I_g$

I_f = Faulty IDDQ

I_g = Fault free IDDQ

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

Avg. improvement of 94% in IDDQ sensitivity

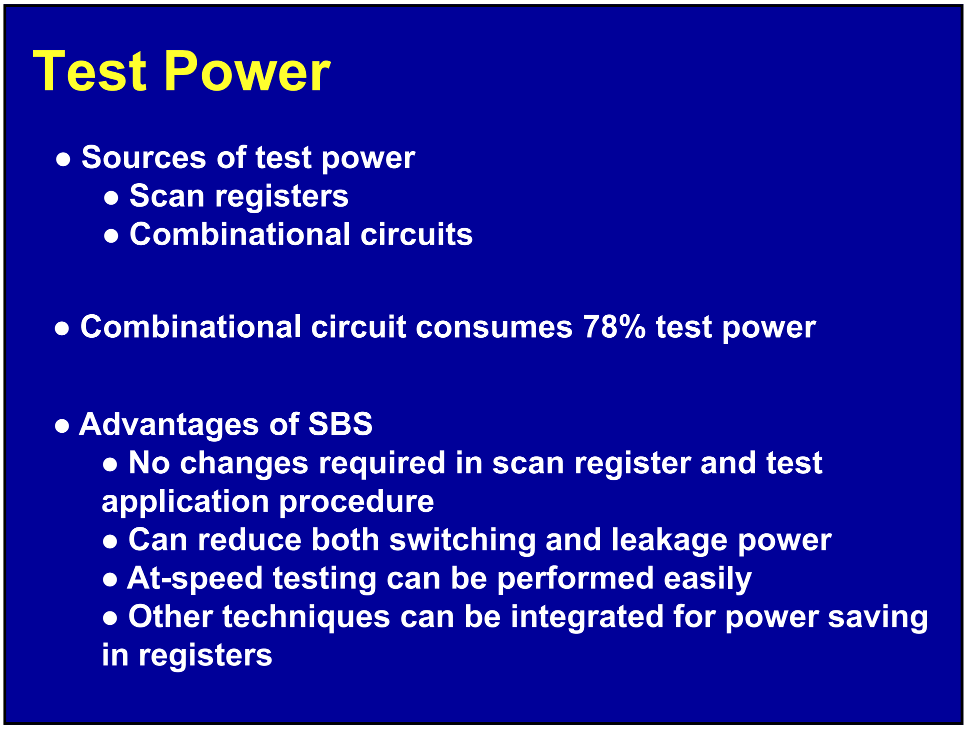


IDDQ Distribution Under Process Variation

(a) cm150a

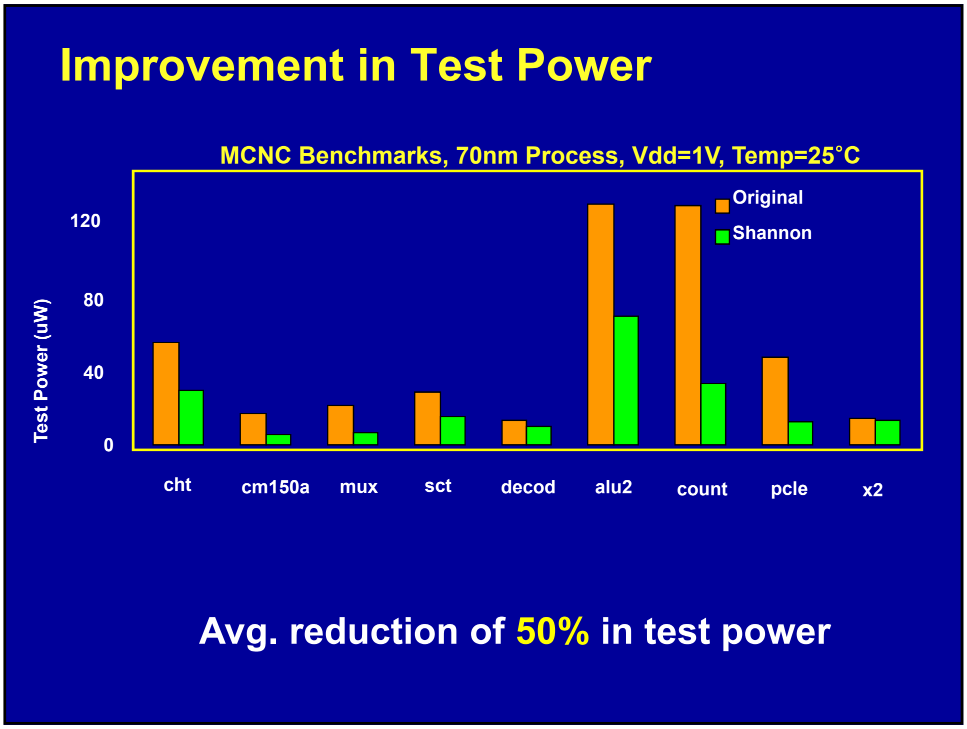
(b) pcle

Improvement of 5% (9%) in parametric yield (for circuit cm150a (pcle), considering leakage bound)



Test Power

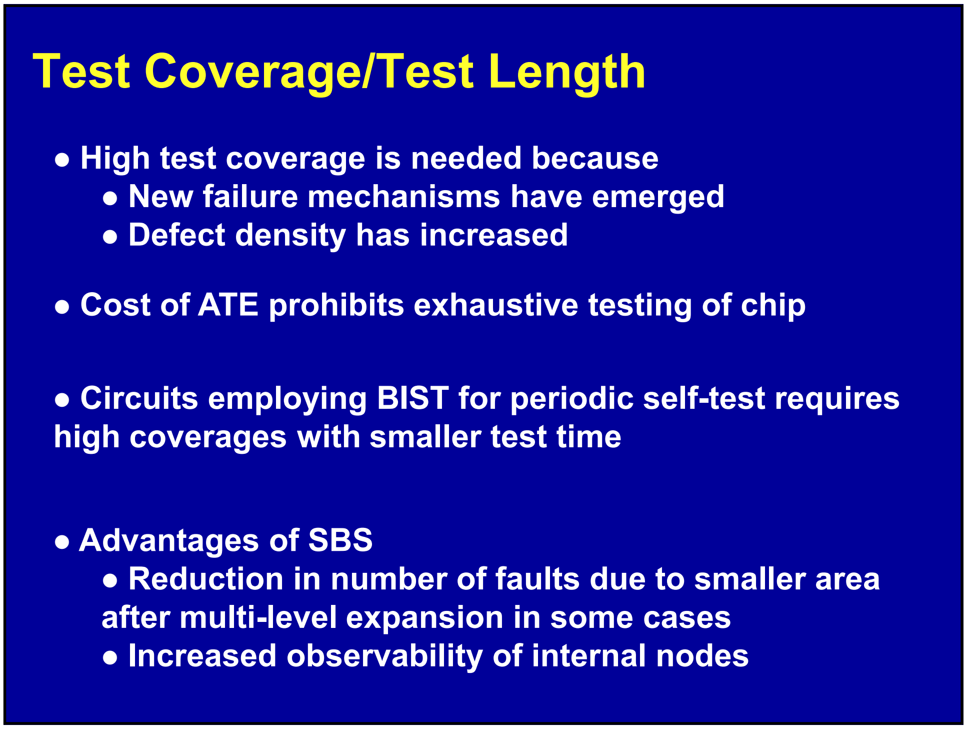
- Sources of test power
  - Scan registers
  - Combinational circuits
- Combinational circuit consumes 78% test power
- Advantages of SBS
  - No changes required in scan register and test application procedure
  - Can reduce both switching and leakage power
  - At-speed testing can be performed easily
  - Other techniques can be integrated for power saving in registers



Improvement in Test Power

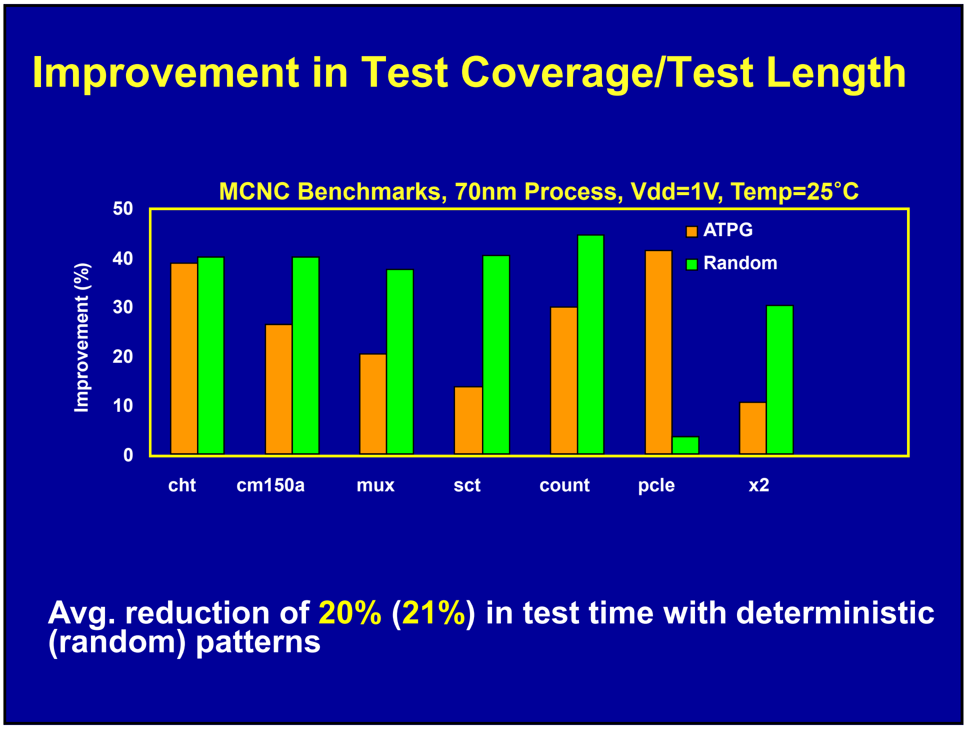
MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=25˚C

Avg. reduction of 50% in test power



Test Coverage/Test Length

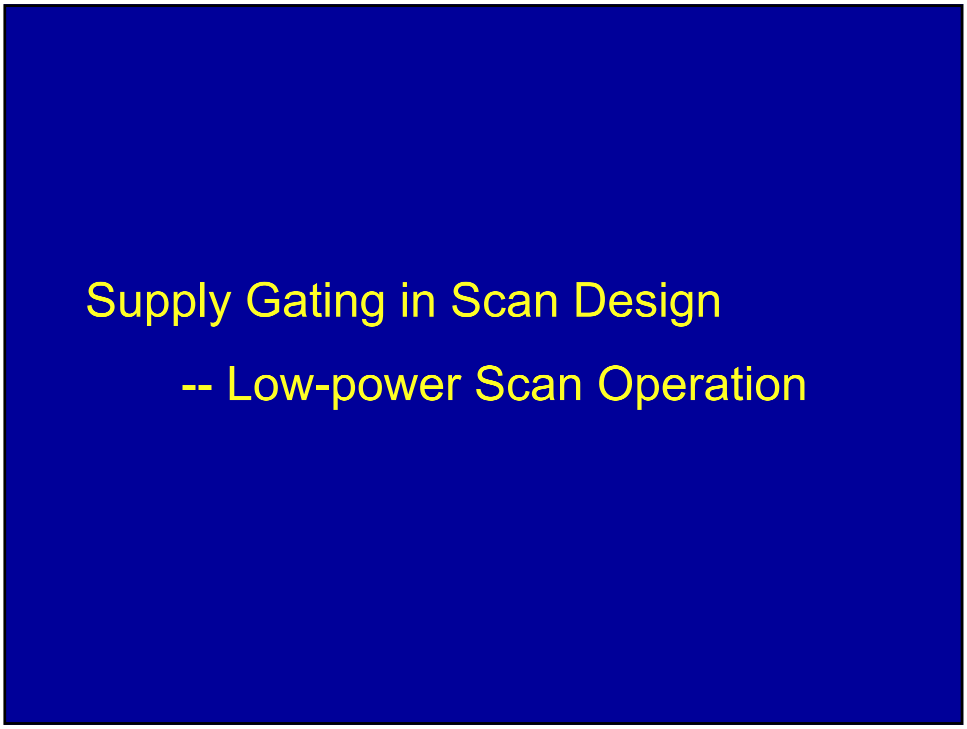
- High test coverage is needed because
  - New failure mechanisms have emerged
  - Defect density has increased
- Cost of ATE prohibits exhaustive testing of chip
- Circuits employing BIST for periodic self-test requires high coverages with smaller test time
- Advantages of SBS
  - Reduction in number of faults due to smaller area after multi-level expansion in some cases
  - Increased observability of internal nodes



Improvement in Test Coverage/Test Length

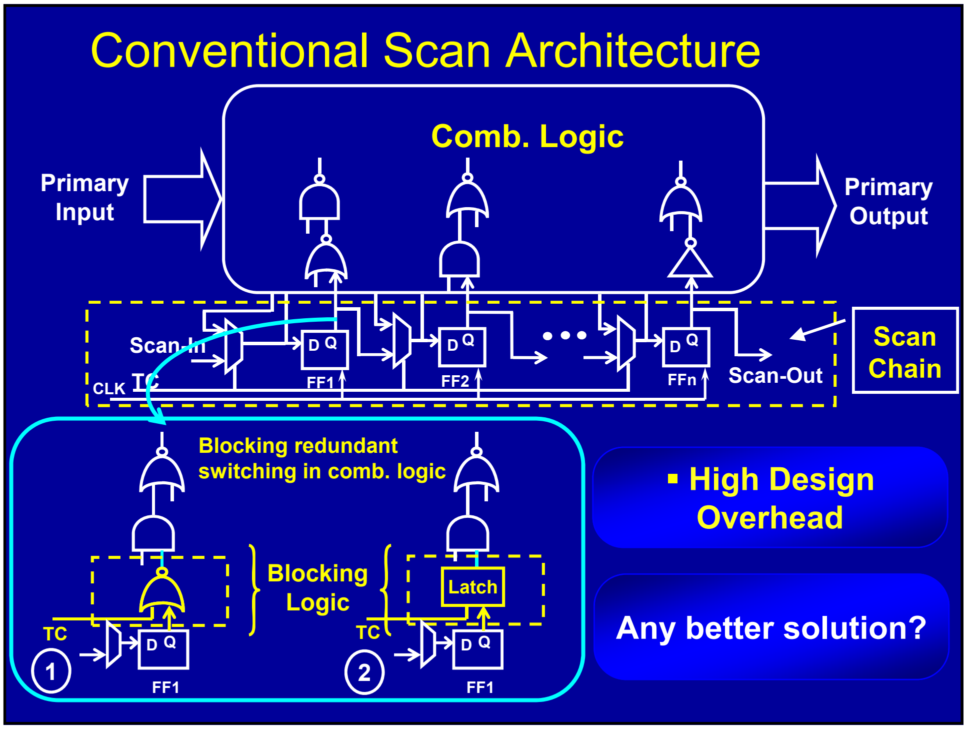
MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=25°C

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



Supply Gating in Scan Design

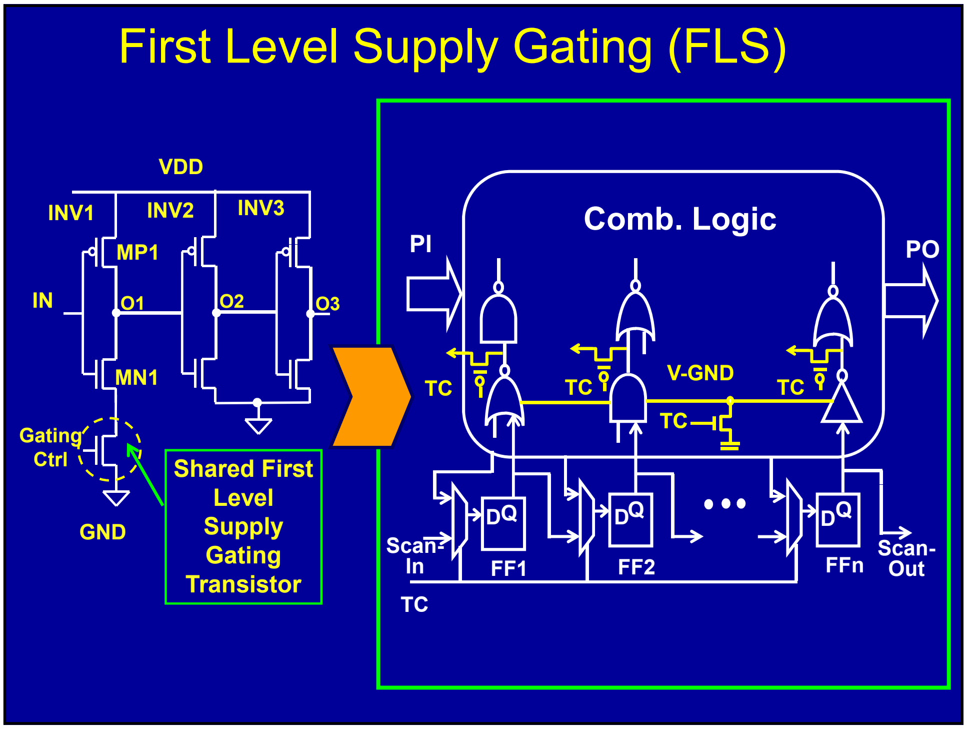
-- Low-power Scan Operation



Conventional Scan Architecture

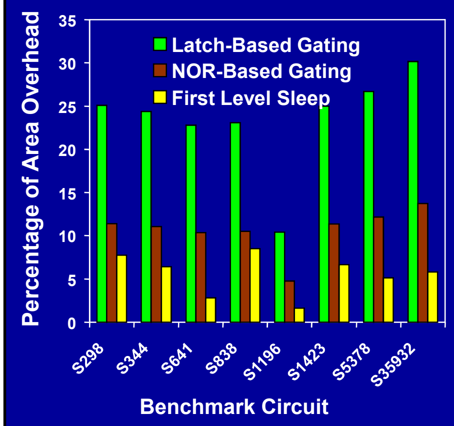
- High Design Overhead

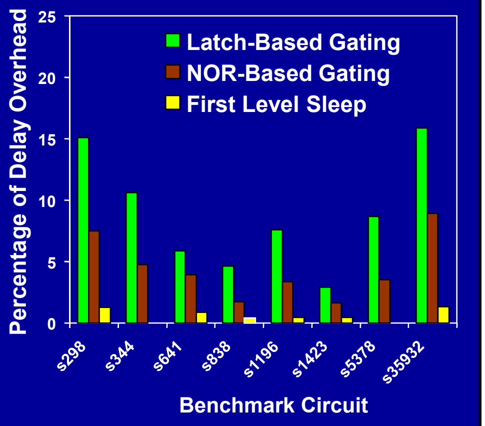
Any better solution?



First Level Supply Gating (FLS)

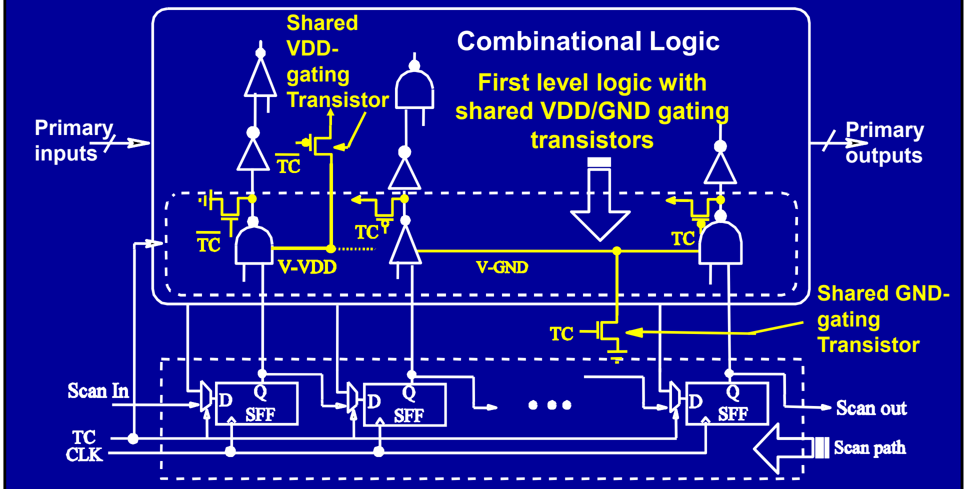
Results and Comparisons for FLS





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

Input Vector Control for Leakage Reduction

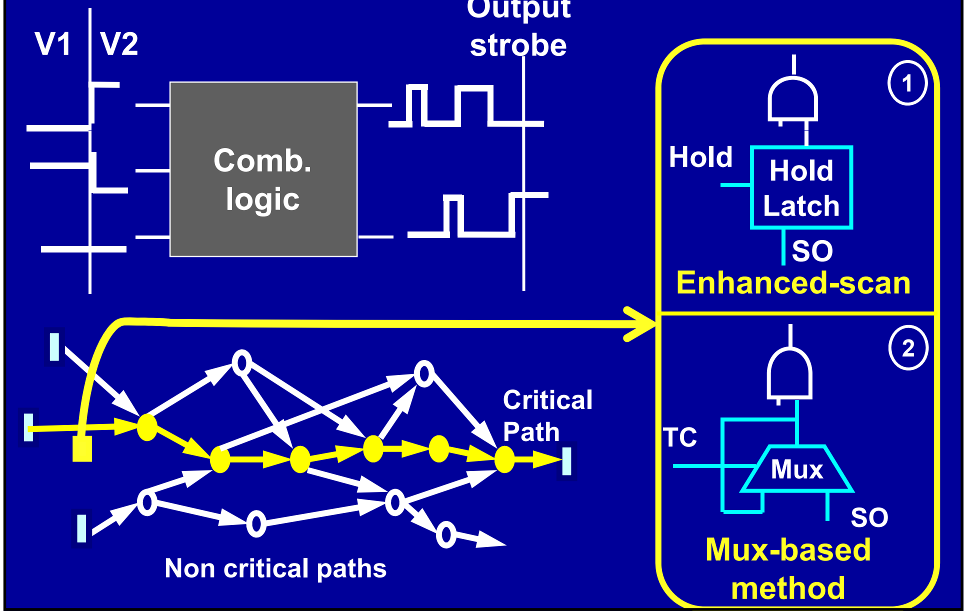

Combinational Logic
First level logic with shared VDD/GND gating transistors
Shared VDD-gating Transistor
Primary inputs
Primary outputs
TC
TC
TC
TC
V-VDD
V-GND
Shared GND-gating Transistor
TC
Scan In
D
Q
SFF
D
Q
SFF
D
Q
SFF
Scan out
Scan path
TC
CLK

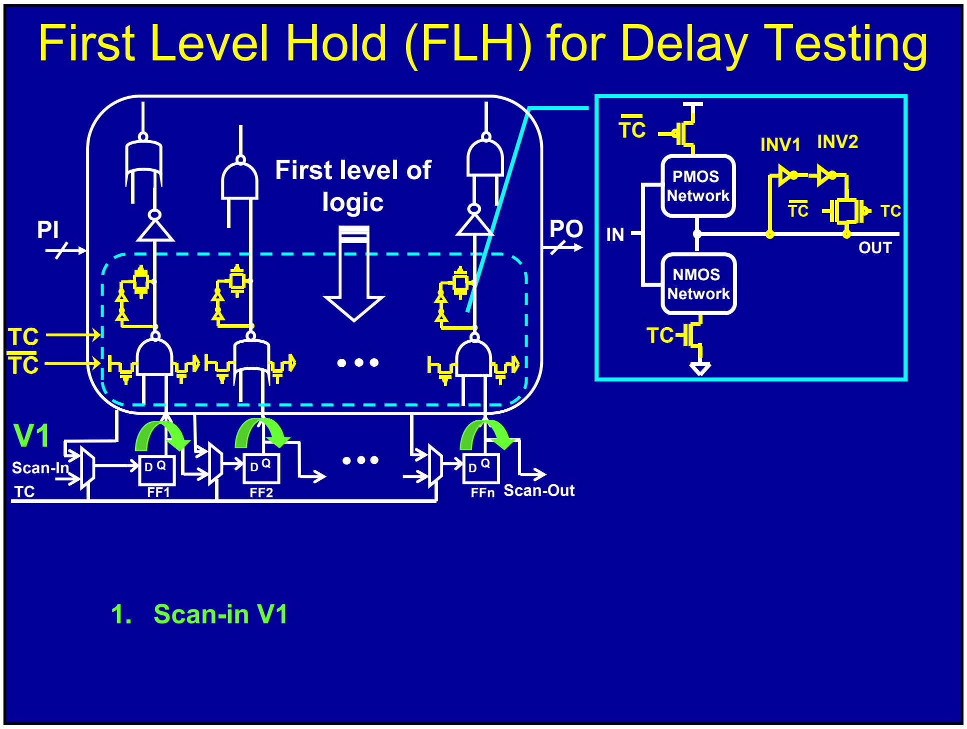
- **Application of best input during scan shifting can save leakage power**
- **About 38% leakage saving with Mixed VDD/GND FLS over NOR gating**

Low-Overhead Delay Fault Testing With Supply Gating

A Delay Test (V1,V2)

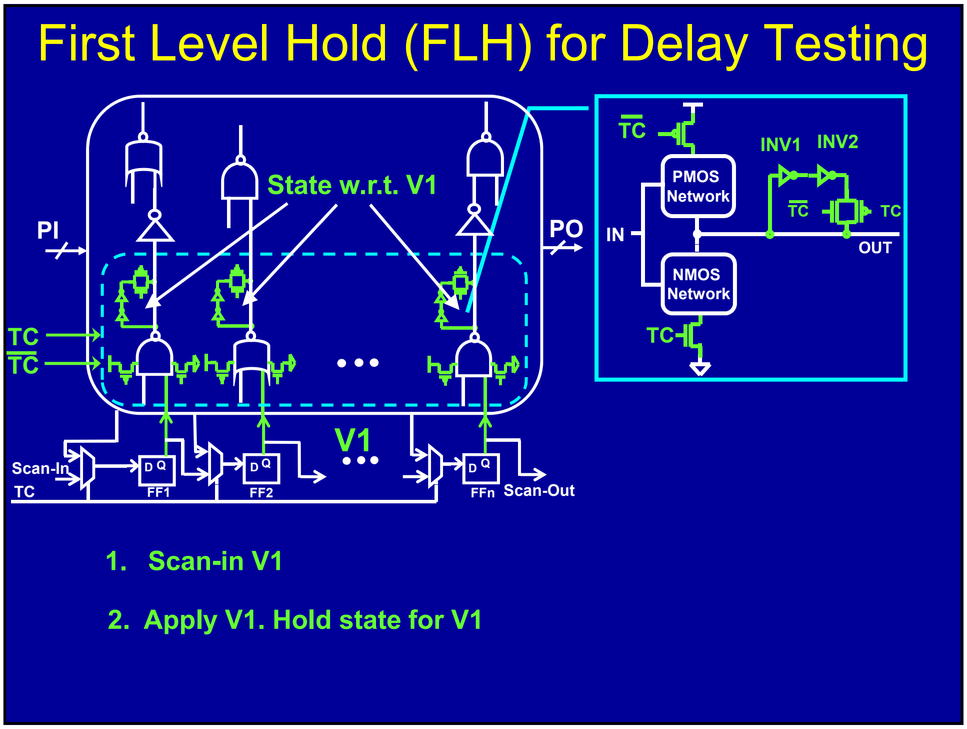
Existing Delay Testing schemes





First Level Hold (FLH) for Delay Testing

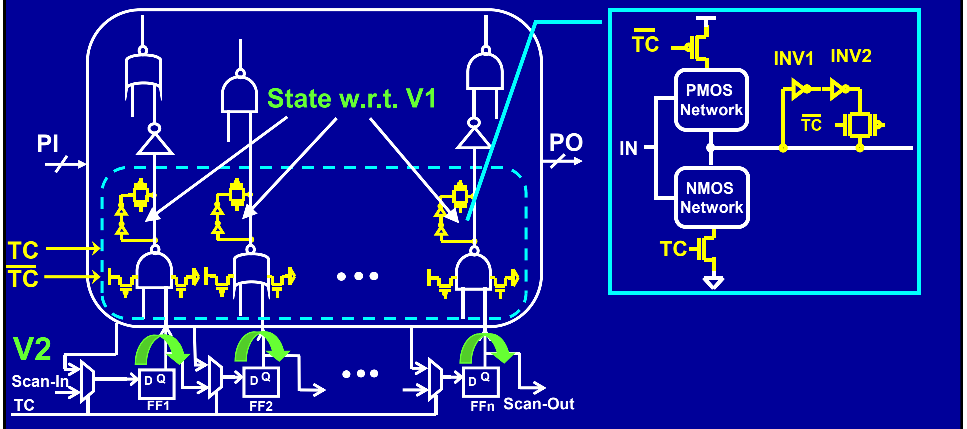
1. Scan-in V1



First Level Hold (FLH) for Delay Testing

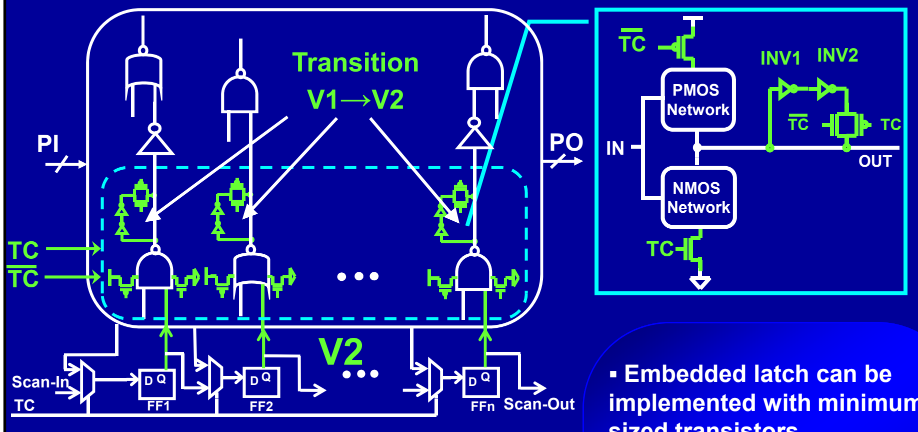
1. Scan-in V1
2. Apply V1. Hold state for V1

First Level Hold (FLH) for Delay Testing



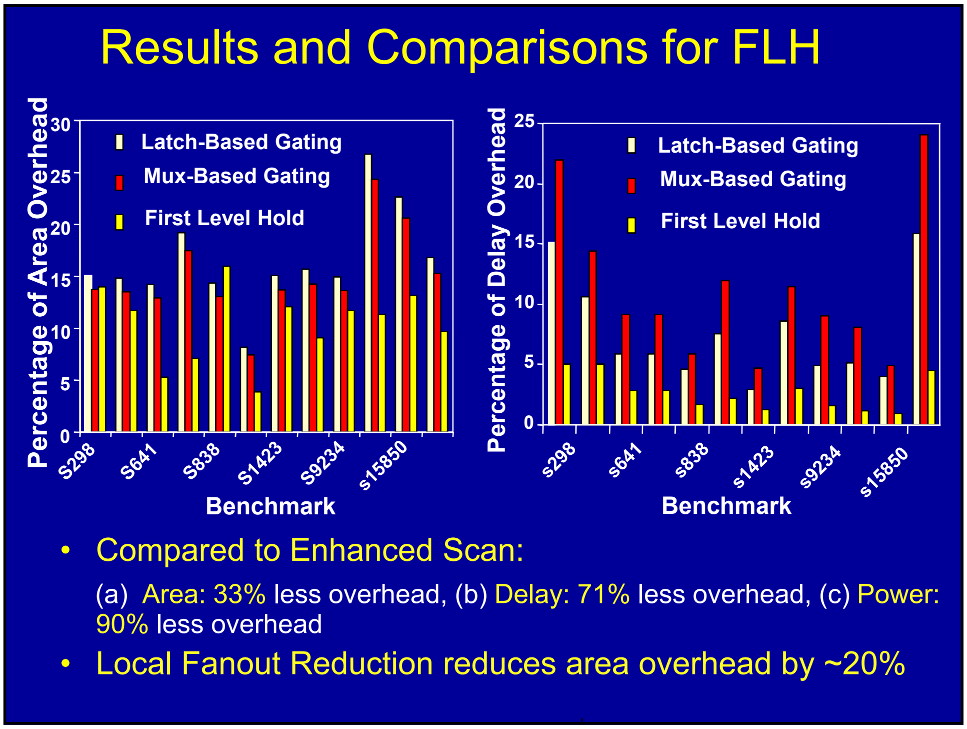
1. Scan-in V1
2. Apply V1. Hold state for V1
3. Scan-in V2

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

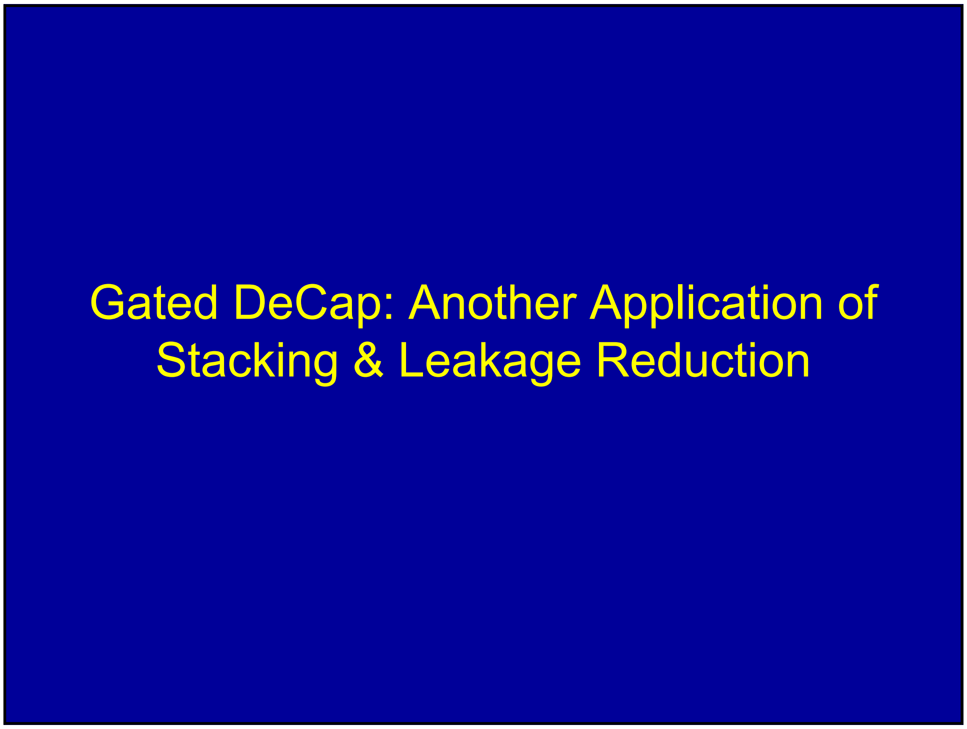


Results and Comparisons for FLH

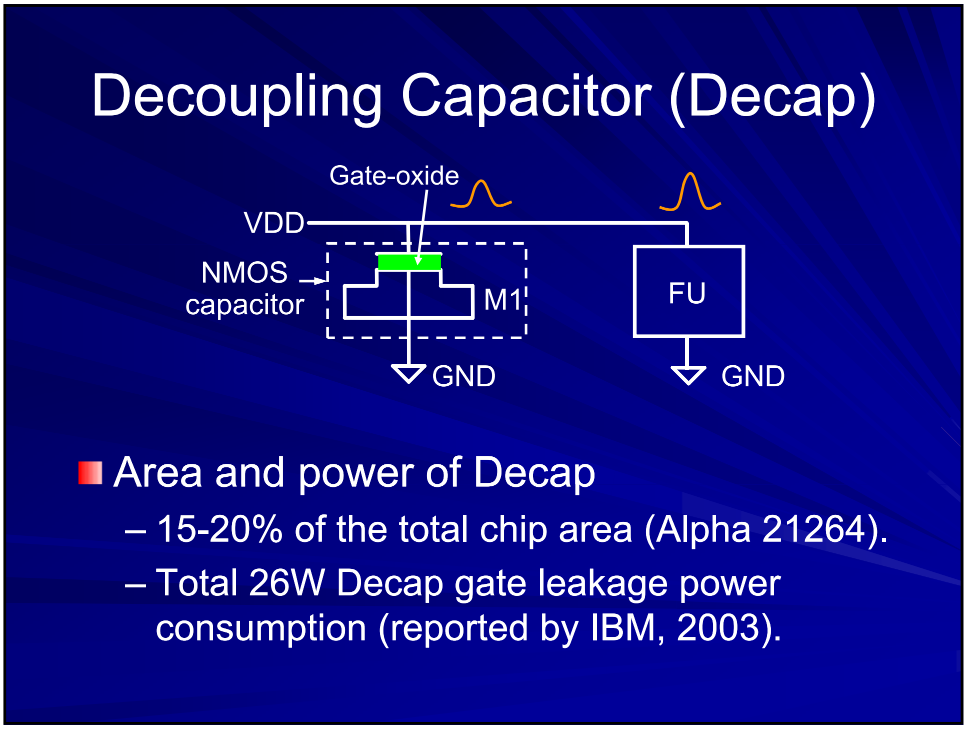
- Compared to Enhanced Scan:

  (a) Area: 33% less overhead, (b) Delay: 71% less overhead, (c) Power: 90% less overhead

- Local Fanout Reduction reduces area overhead by ~20%

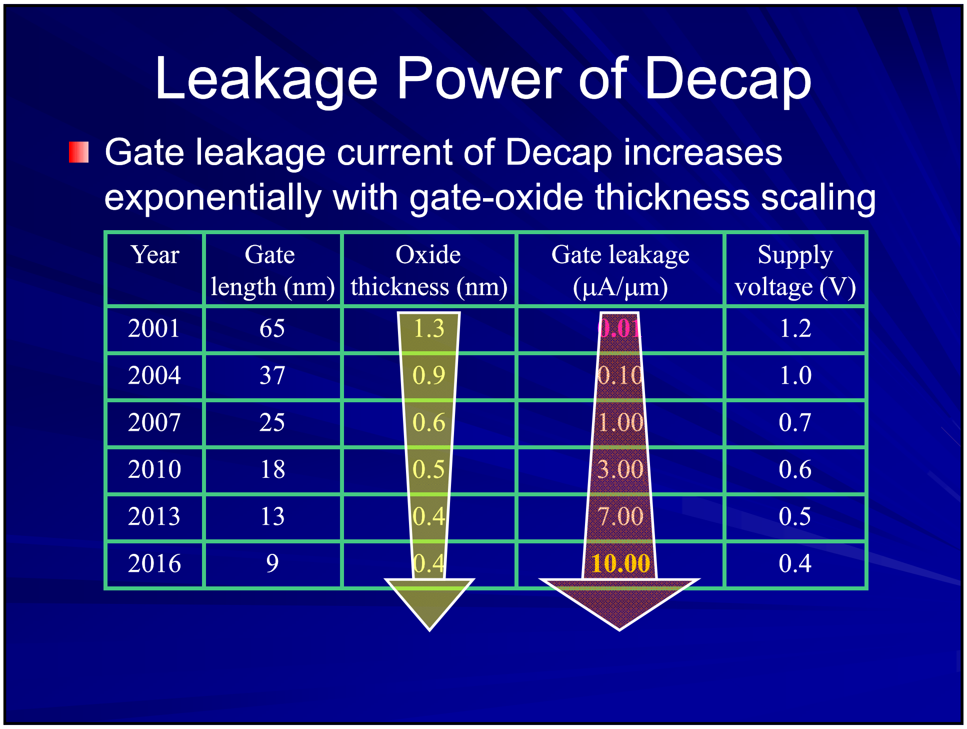


Gated DeCap: Another Application of Stacking & Leakage Reduction



Decoupling Capacitor (Decap)

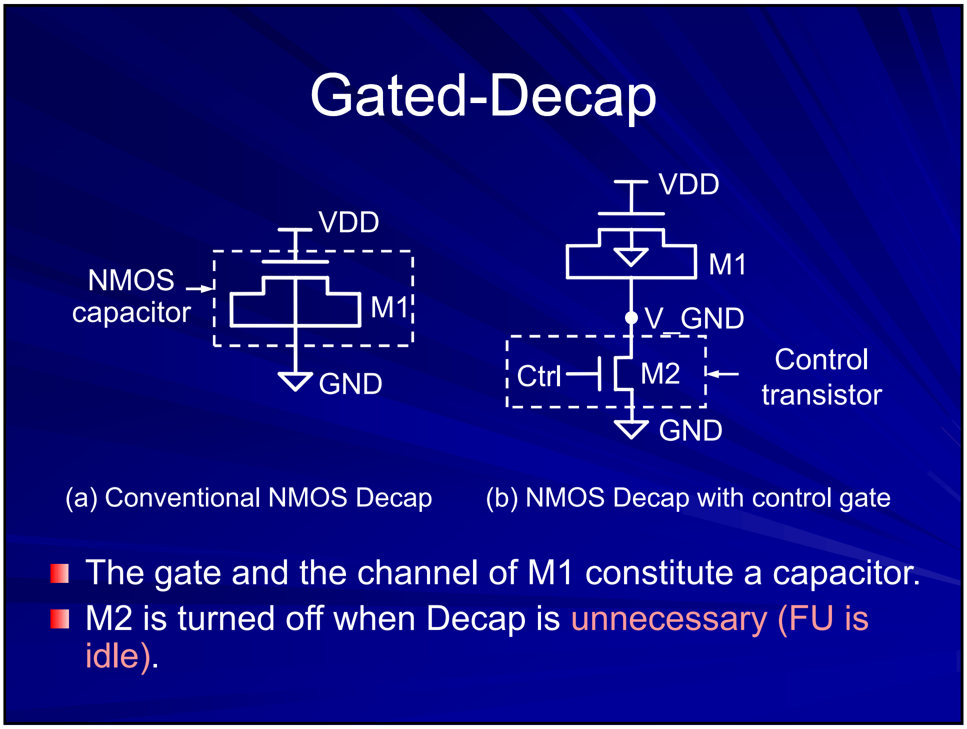
- Area and power of Decap
  - 15-20% of the total chip area (Alpha 21264).
  - Total 26W Decap gate leakage power consumption (reported by IBM, 2003).



Leakage Power of Decap

- Gate leakage current of Decap increases exponentially with gate-oxide thickness scaling

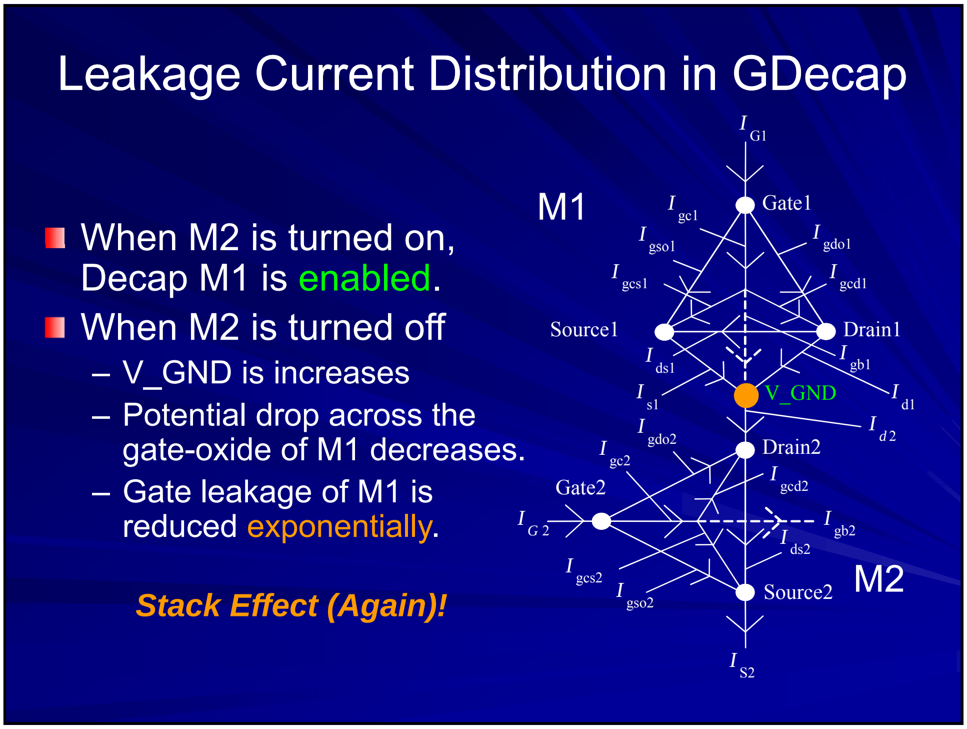
Year	Gate length (nm)	Oxide thickness (nm)	Gate leakage (µA/µm)	Supply voltage (V)
2001	65	1.3	0.01	1.2
2004	37	0.9	0.10	1.0
2007	25	0.6	1.00	0.7
2010	18	0.5	3.00	0.6
2013	13	0.4	7.00	0.5
2016	9	0.4	**10.00**	0.4



Gated-Decap

(a) Conventional NMOS Decap (b) NMOS Decap with control gate

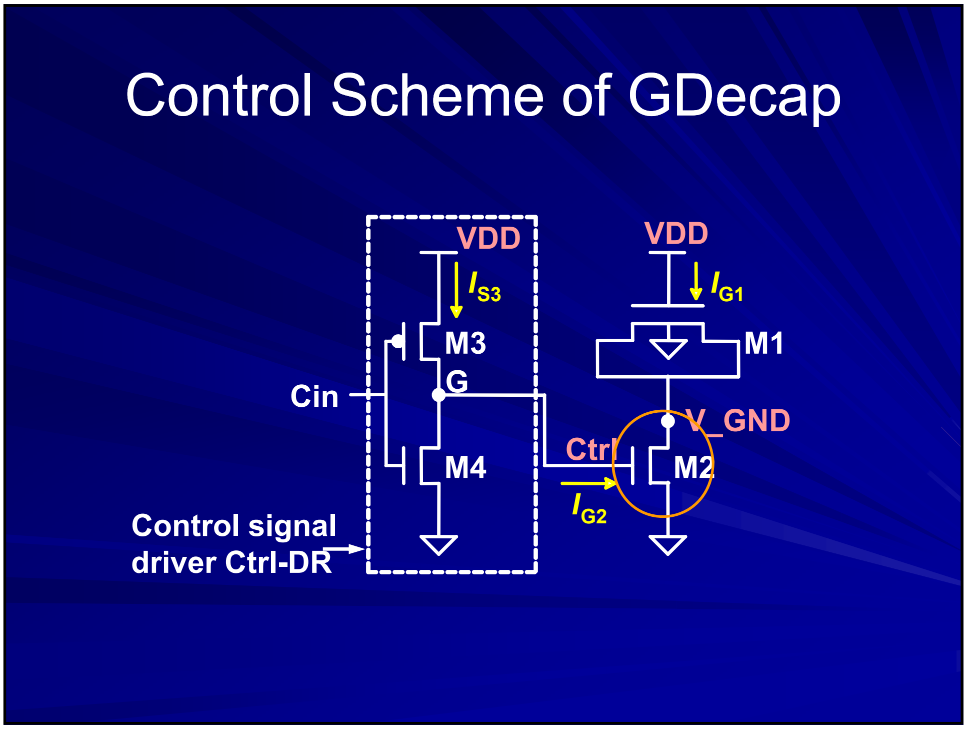
- The gate and the channel of M1 constitute a capacitor.
- M2 is turned off when Decap is unnecessary (FU is idle).



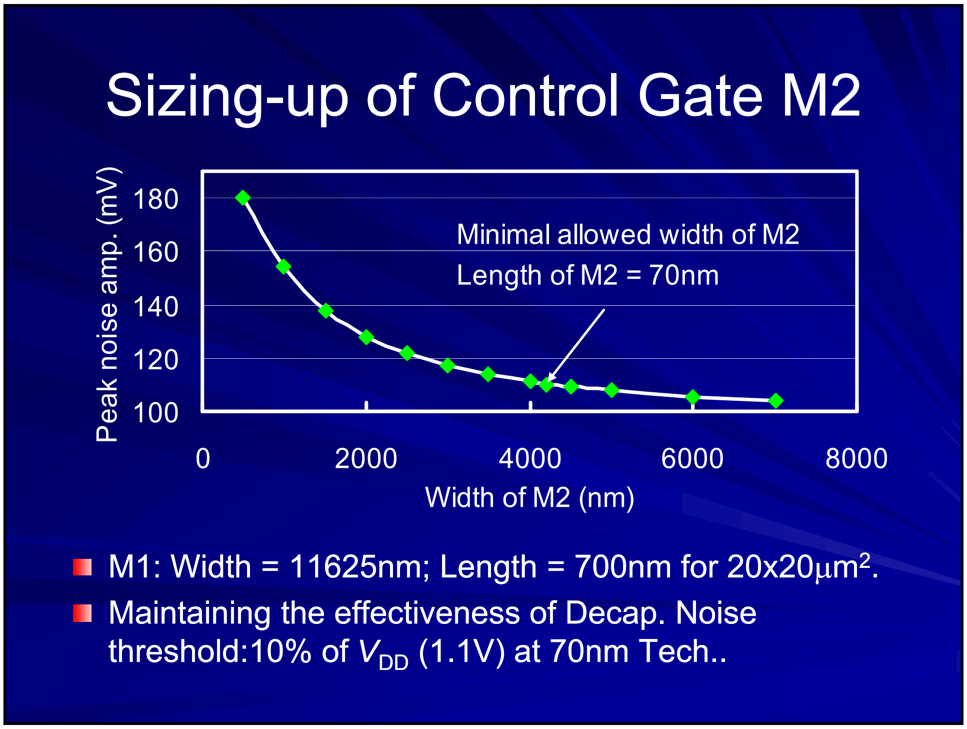
Leakage Current Distribution in GDecap

- When M2 is turned on, Decap M1 is enabled.
- When M2 is turned off
  - V_GND is increases
  - Potential drop across the gate-oxide of M1 decreases.
  - Gate leakage of M1 is reduced exponentially.

Stack Effect (Again)!

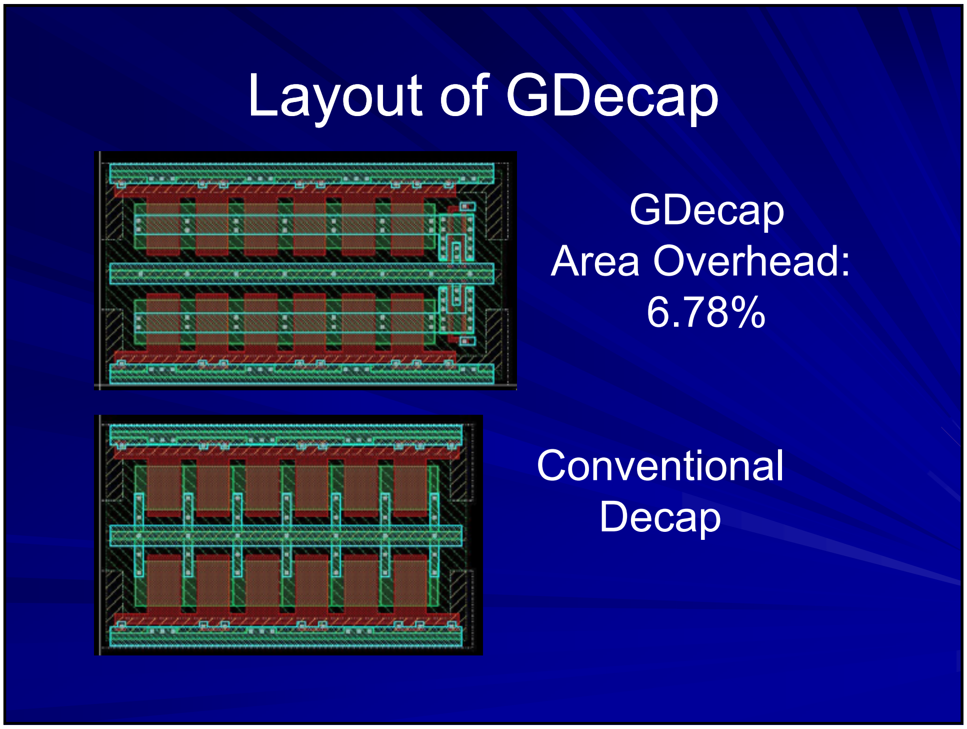


Control Scheme of GDecap



Sizing-up of Control Gate M2

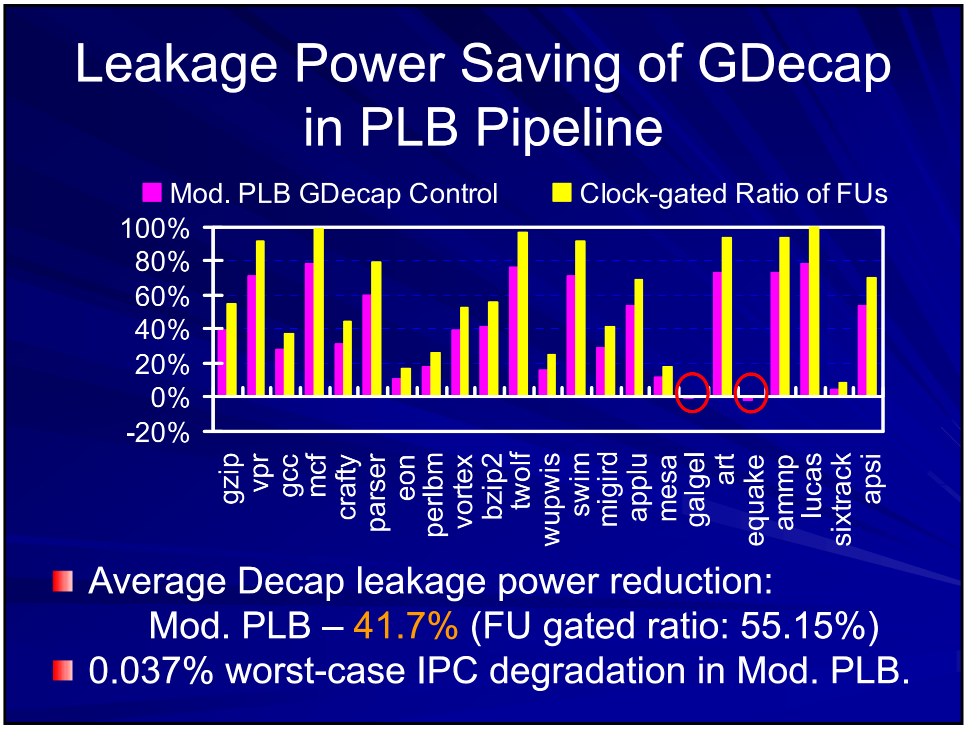
- M1: Width = 11625nm; Length = 700nm for 20x20μm^2.
- Maintaining the effectiveness of Decap. Noise threshold:10% of V_{DD} (1.1V) at 70nm Tech..



Layout of GDecap

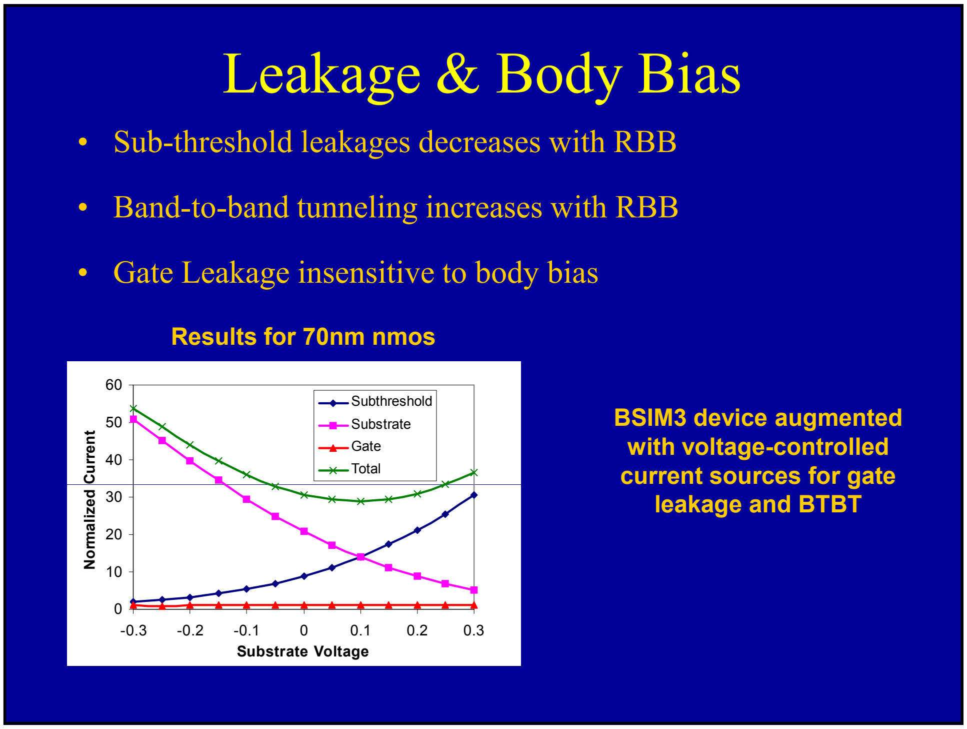
GDecap
Area Overhead:
6.78%

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.

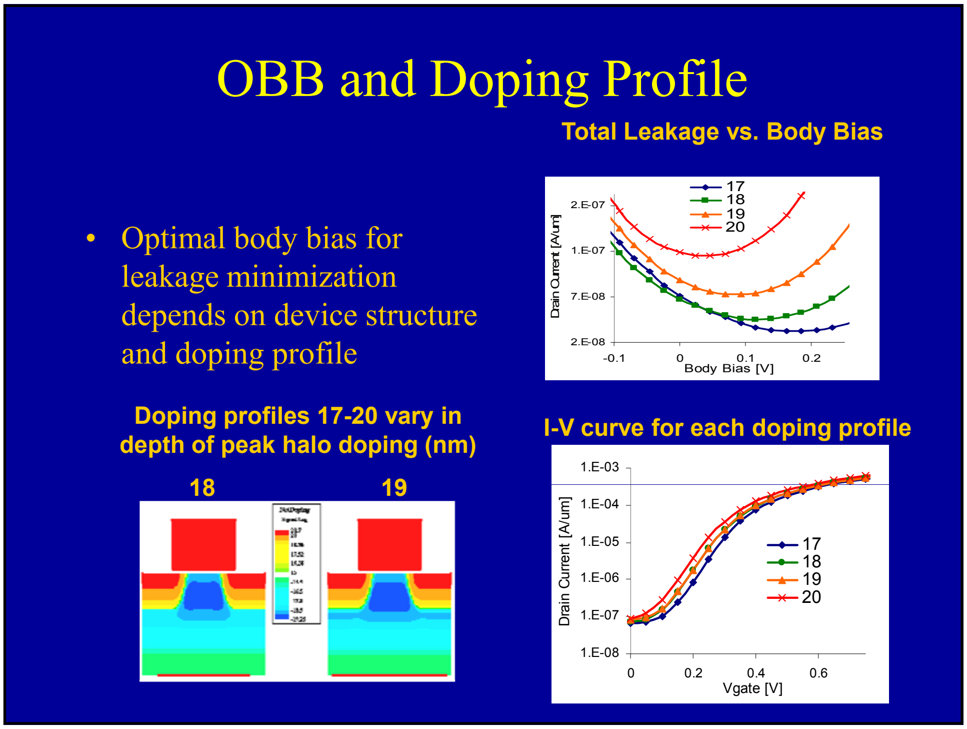


Leakage & Body Bias

- 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



OBB and Doping Profile

- Optimal body bias for leakage minimization depends on device structure and doping profile

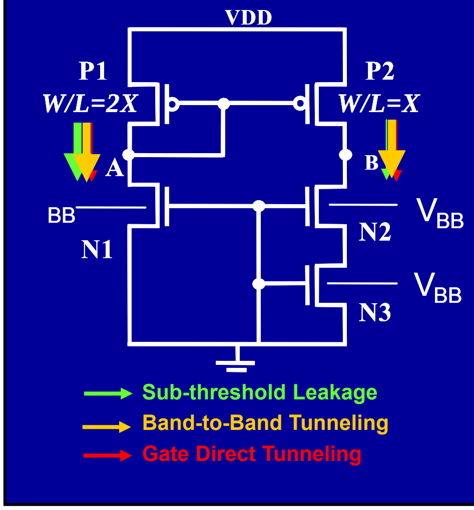
Total Leakage vs. Body Bias

Doping profiles 17-20 vary in depth of peak halo doping (nm)

I-V curve for each doping profile

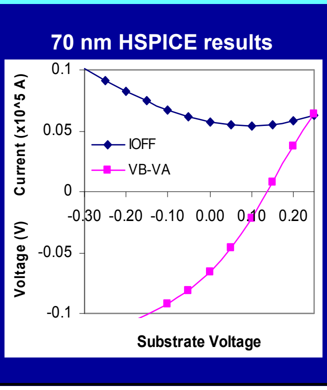
OBB Selection Circuit

- Body bias minimizes leakage when BTBT leakage is approximately equal to the sub-threshold leakage.



Adjust body bias until V(A) = V(B). Leakage current on the left side of the current mirror is twice the leakage current on the right side.

70 nm HSPICE results



Leakage Reduction with OBB

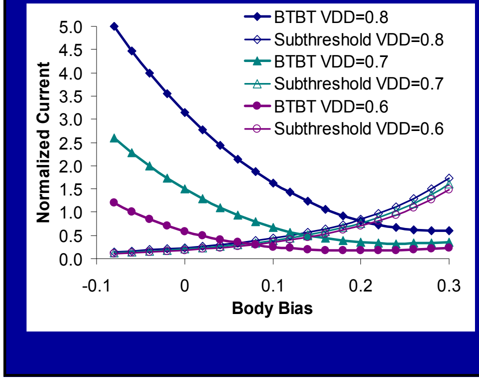
- Leakage savings ranged from 14-55% compared to zero body bias case for nominal 70nm and 50nm transistors in Taurus device simulations.

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	

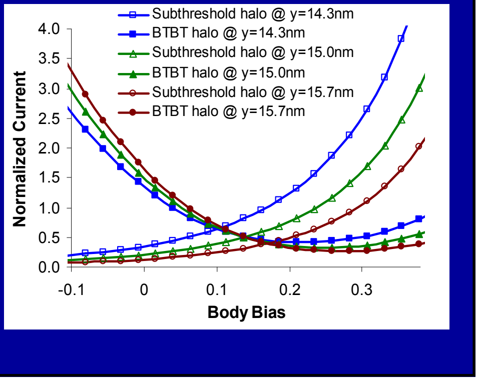
Variation Effects on OBB

- Optimal Body Bias is affected by variations in:
  - Supply voltage
  - Gate length
  - Doping Profile
  - Temperature

Variation in Supply Voltage



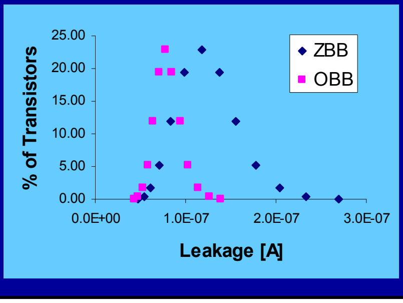
Variation in Halo Doping Location



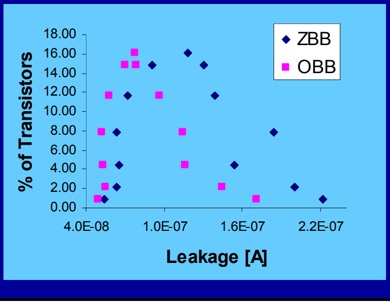
Variation Reduction with OBB

- OBB selector circuit automatically adjusts to process and operating conditions to reduce variation in leakage
  - Leakage values determined for 50nm transistors with Gaussian distributed parameter variations. Spread of leakage values reduced with OBB compared to ZBB

Variation in Supply Voltage
Gaussian (μ = 0.7V, σ = 0.035V)



Variation in Channel Length
Gaussian (μ = 50nm, σ = 2.5nm)



Variation Reduction Results

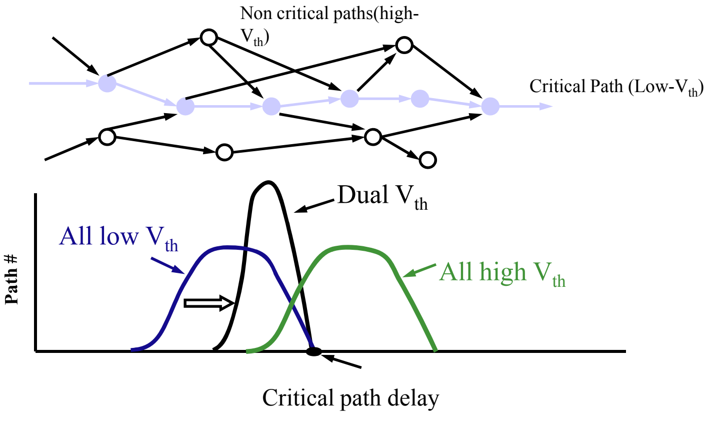
- OBB reduces mean leakage by 30-37%
- OBB reduces the spread of leakage values by 40-71%

Taurus Device simulation results for 50nm nmos with Gaussian distributed parameter variations

Device Variation	Leakage Variation			
	μ @ ZBB [A]	μ @ OBB [A]	σ @ ZBB	σ @ OBB
Length	1.14e-7	7.97e-8	3.89e-8	2.32e-8
VDD	1.20e-7	7.87e-8	3.19e-8	1.33e-8
Peak Halo Doping X	1.27e-7	7.96e-8	1.96e-8	5.70e-9

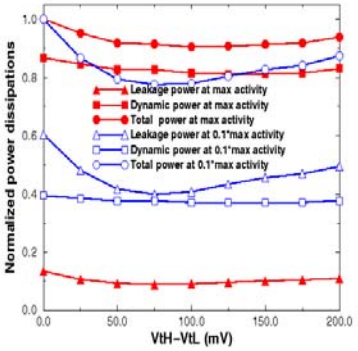
Dual Threshold CMOS

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



Total Power of 32-bit Adder

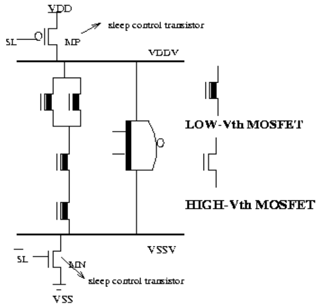
- Total power can be reduced by 9% for high activity
- Total power can be reduced by 22% at low activity



Process Variation & Dual-Vt

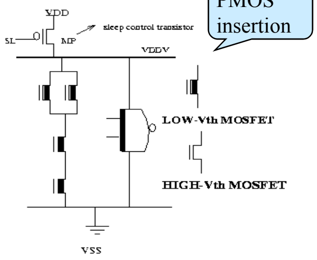
MTCMOS

- 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 “of leakage is suppressed.

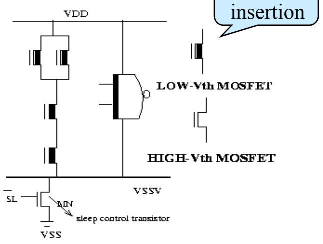


MTCMOS (cont’d)

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



PMOS insertion



NMOS insertion

- NMOS size smaller

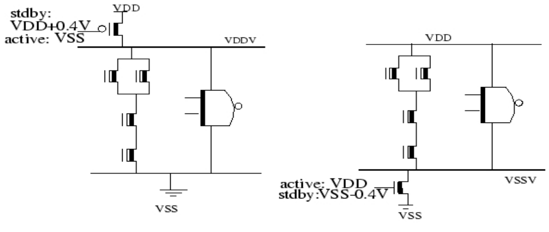
⇨ NMOS insertion is preferable

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

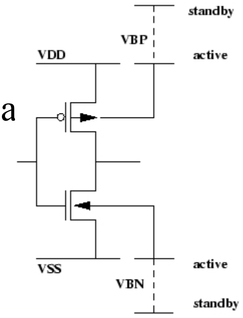


(a) PMOS insertion

(b) NMOS insertion

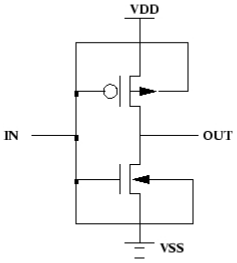
VTCMOS

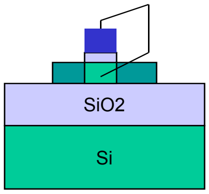
- Variable Threshold CMOS (from T. Kuroda, ISSCC, 1996)
- In active mode:
  - Zero or slightly forward body bia for high speed
- In standby mode:
  - Deep reverse body bias for low leakage
- Triple well technology required

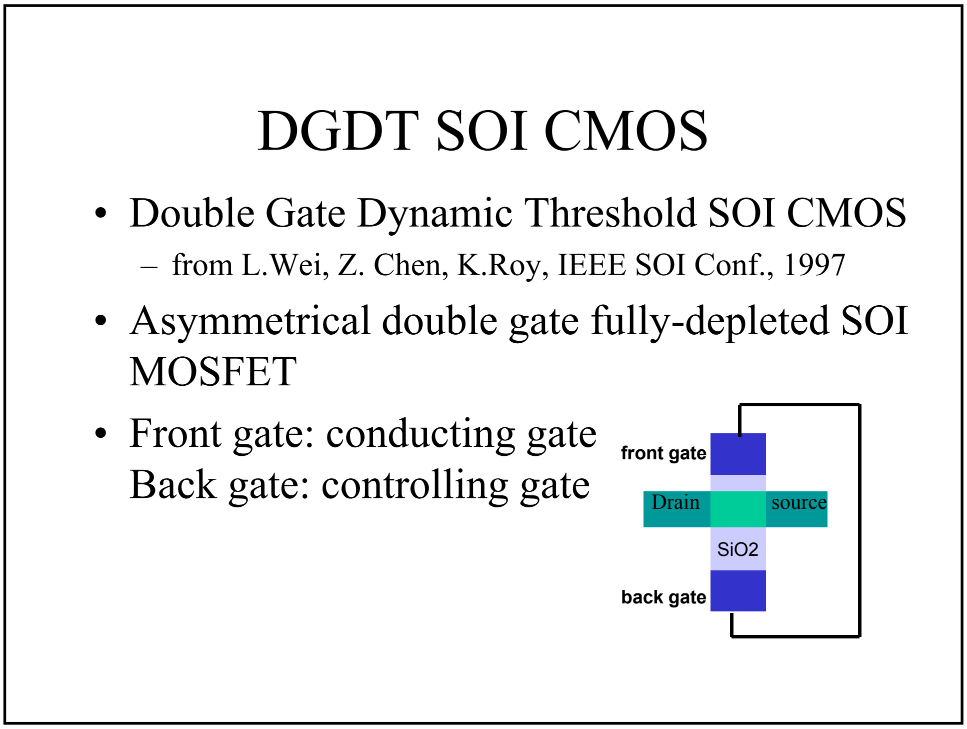


DTMOS

- 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



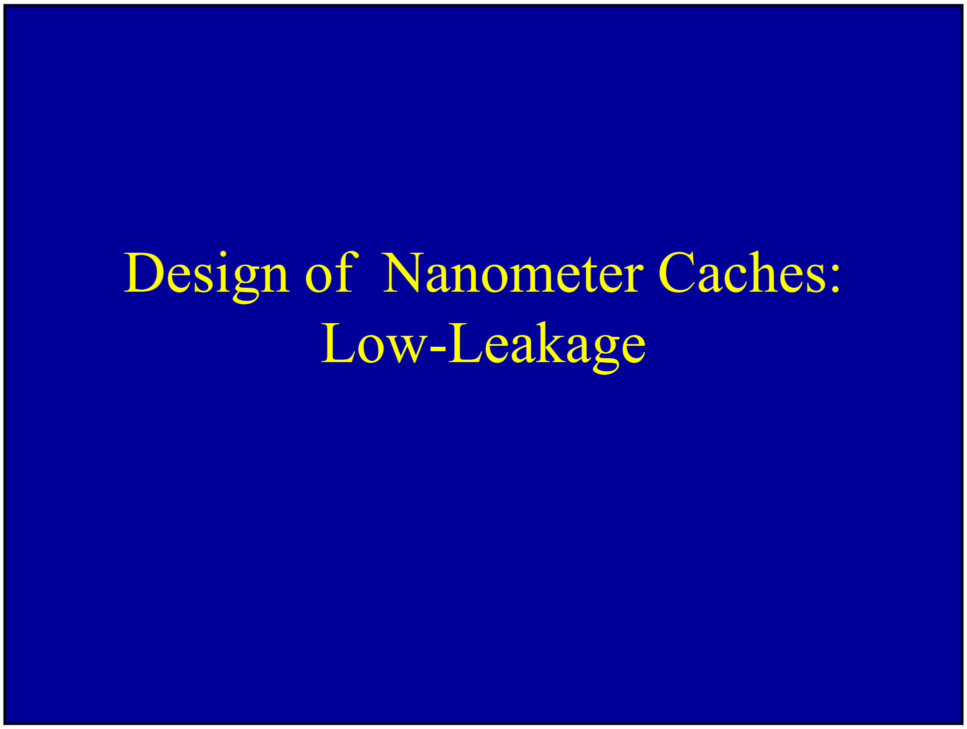




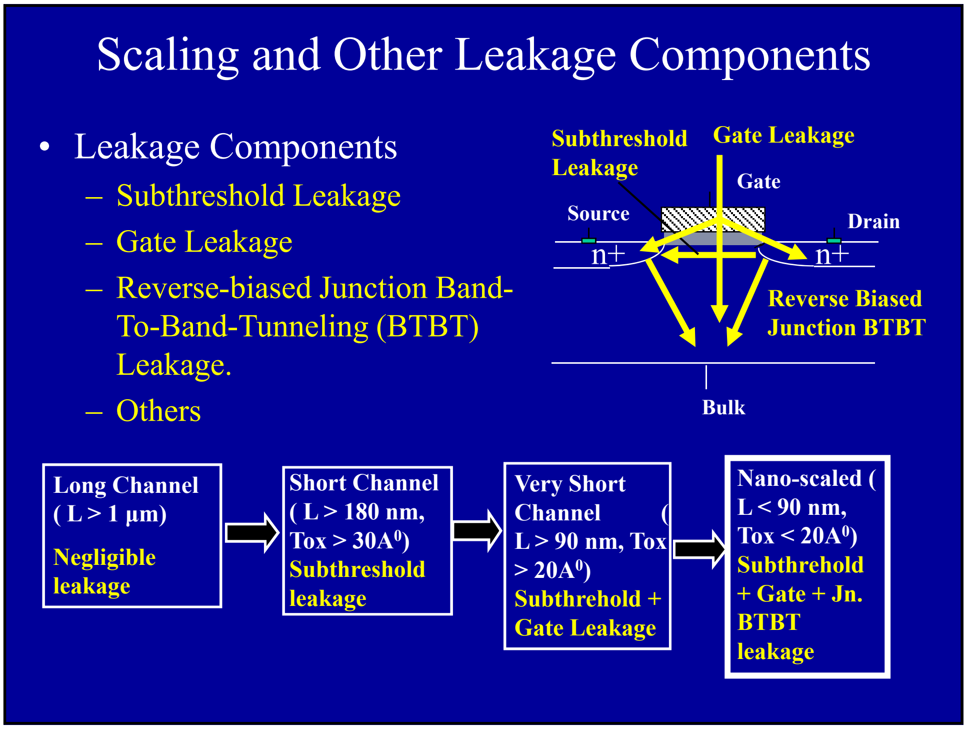
DGDT SOI CMOS

- 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

front gate
Drain
source
SiO2
back gate



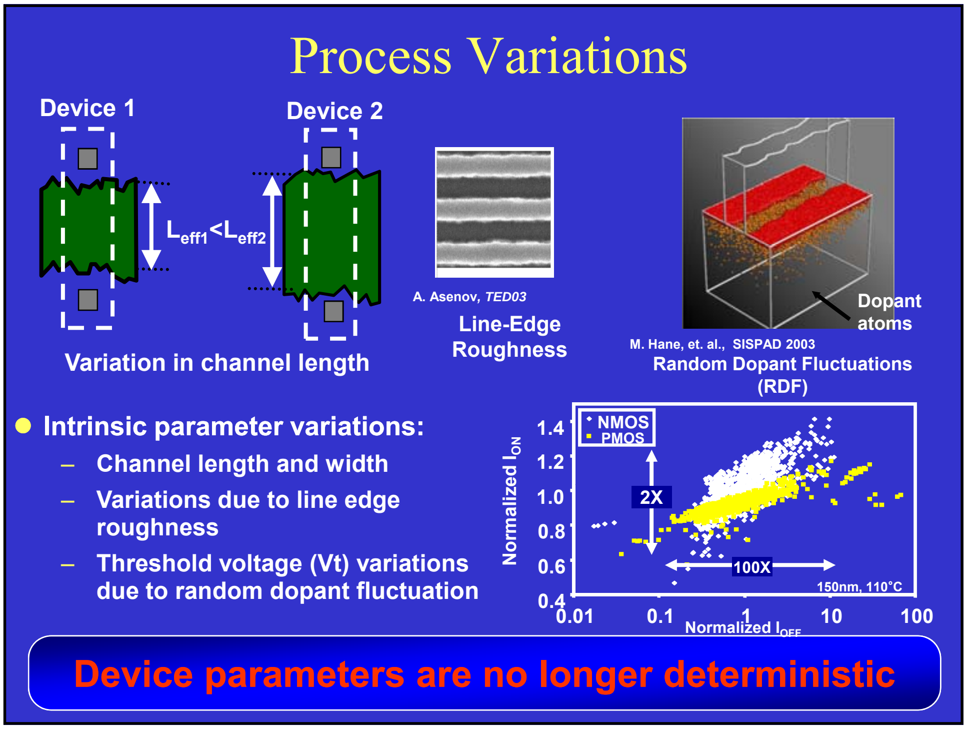
Design of Nanometer Caches: Low-Leakage



Scaling and Other Leakage Components

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

Long Channel (L > 1 µm) **Negligible leakage** → **Short Channel (L > 180 nm, Tox > 30A^0) Subthreshold leakage** → **Very Short Channel (L > 90 nm, Tox > 20A^0) Subthreshold + Gate Leakage** → **Nano-scaled (L < 90 nm, Tox < 20A^0) Subthrehold + Gate + Jn. BTBT leakage**



Process Variations

Variation in channel length

A. Asenov, *TED03*

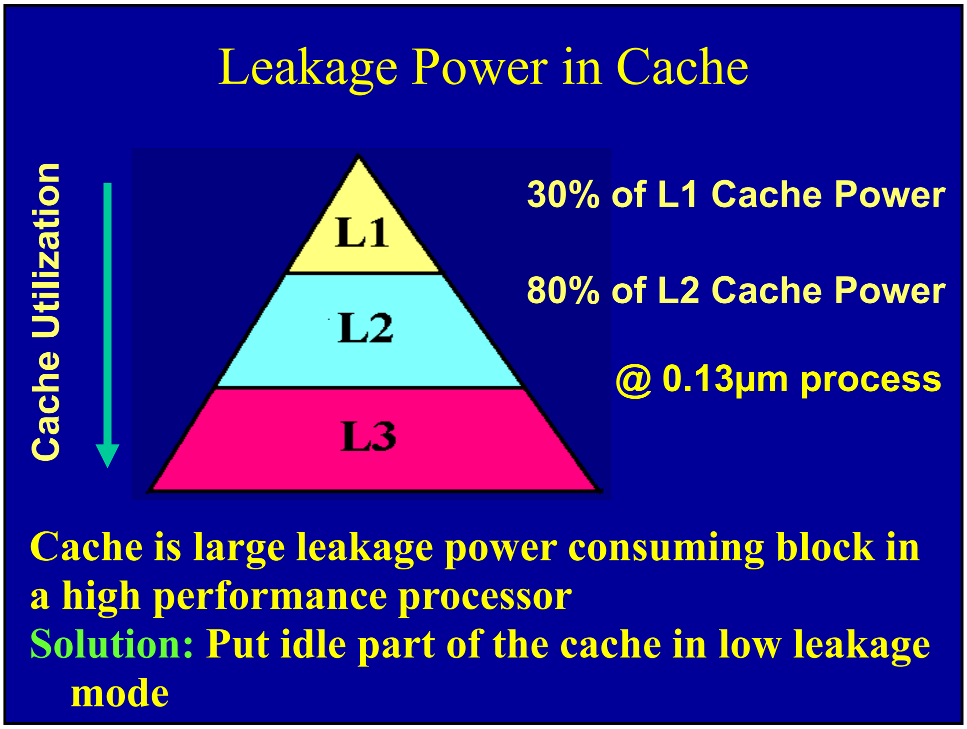
Line-Edge Roughness

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 (Vt) variations due to random dopant fluctuation**

Device parameters are no longer deterministic



Leakage Power in Cache

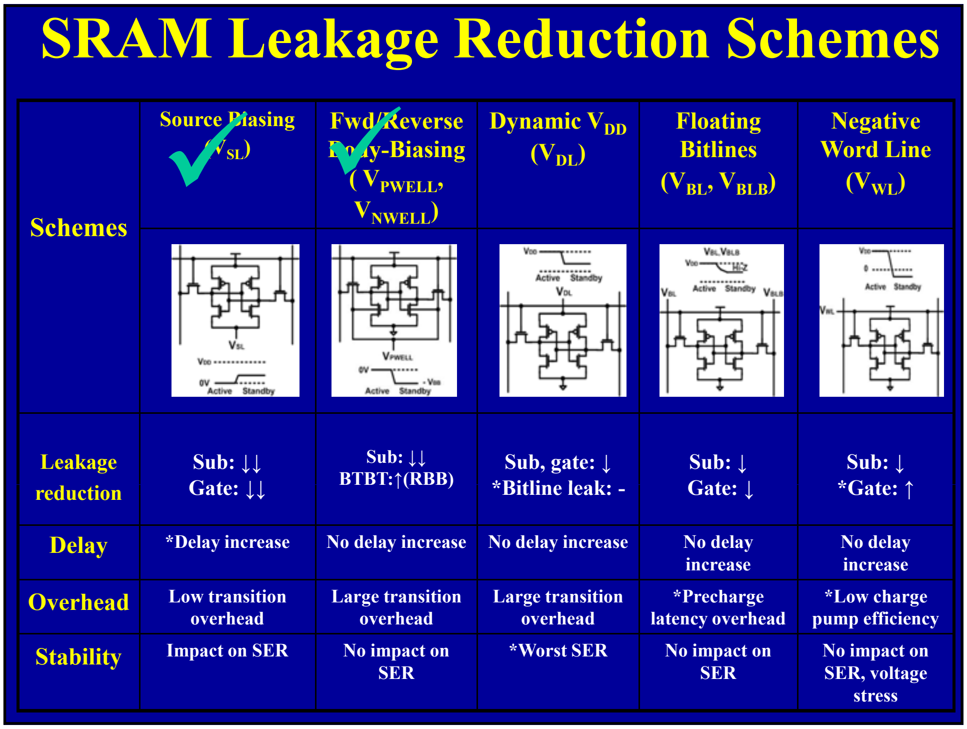
30% of L1 Cache Power

80% of L2 Cache Power

@ 0.13μm process

Cache is large leakage power consuming block in a high performance processor

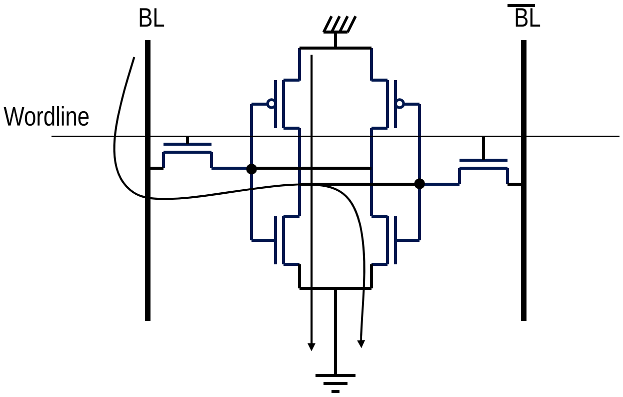
Solution: Put idle part of the cache in low leakage mode



SRAM Leakage Reduction Schemes

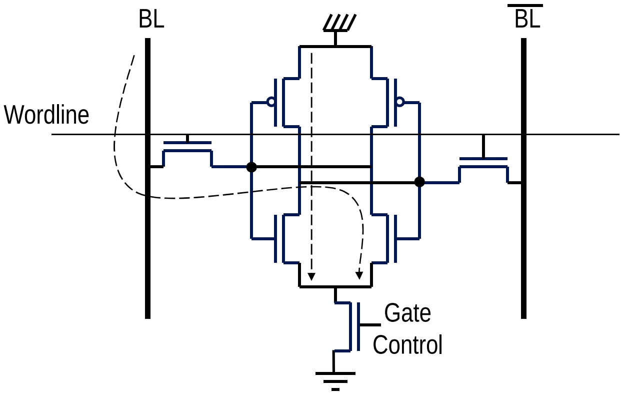
Schemes	Source Biasing ✓ (V_{SL})	Fwd/Reverse Body-Biasing ✓ (V_{PWELL}, V_{NWELL})	Dynamic V_{DD} (V_{DL})	Floating Bitlines (V_{BL}, V_{BLB})	Negative Word Line (V_{WL})
Leakage reduction	Sub: ↓↓ Gate: ↓↓	Sub: ↓↓ BTBT:↑(RBB)	Sub, gate: ↓ *Bitline leak: -	Sub: ↓ Gate: ↓	Sub: ↓ *Gate: ↑
Delay	*Delay 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

Conventional Cell Leakage Paths



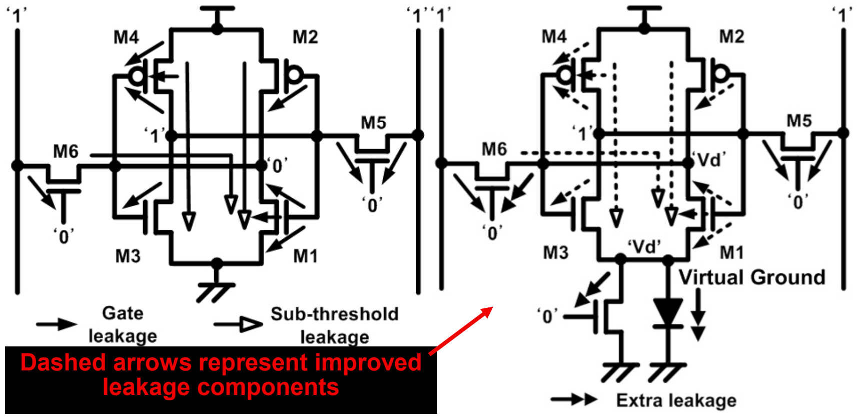
- V_{dd} to ground path
- Bitline to ground path

Gated-Ground (Source-Biased) SRAM



- Gating options: NMOS, Dual-V_t, PMOS

Leakage Reduction in Diode Footed Cache

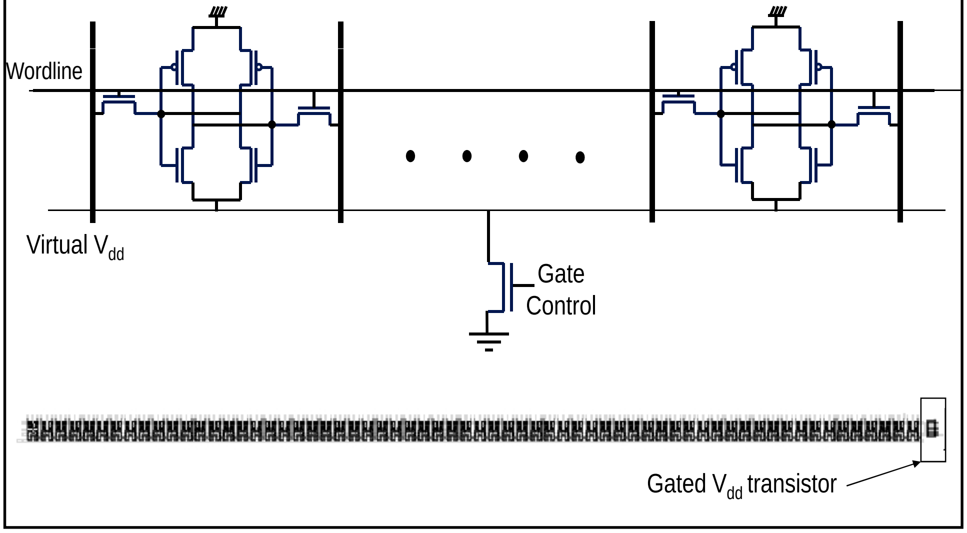


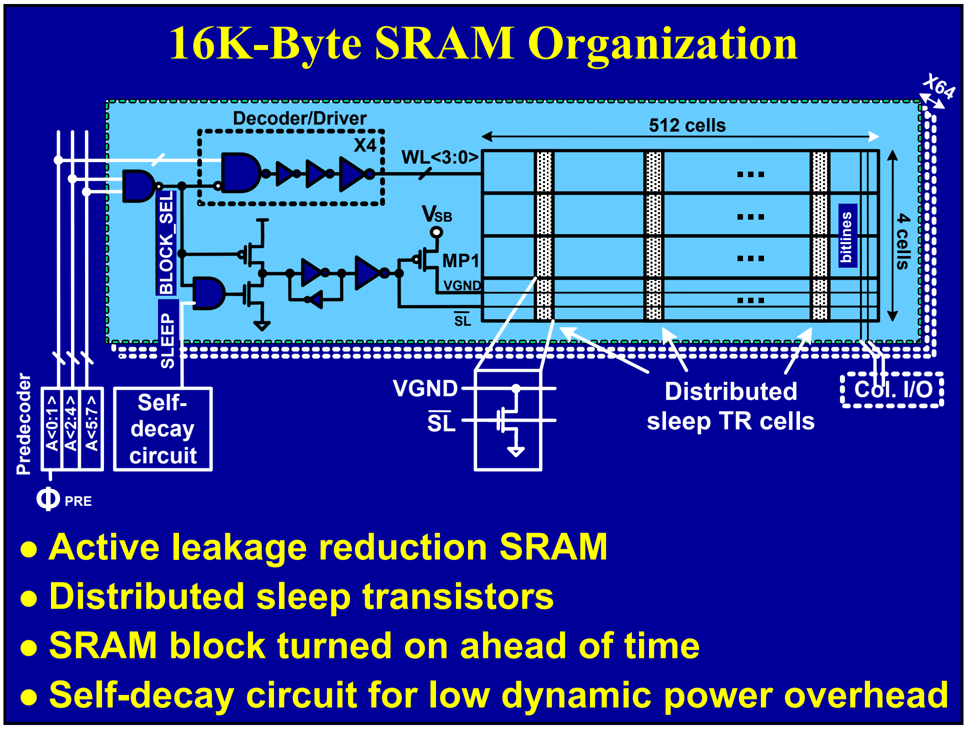
Dashed arrows represent improved leakage components

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

Reduces gate and subthreshold leakage

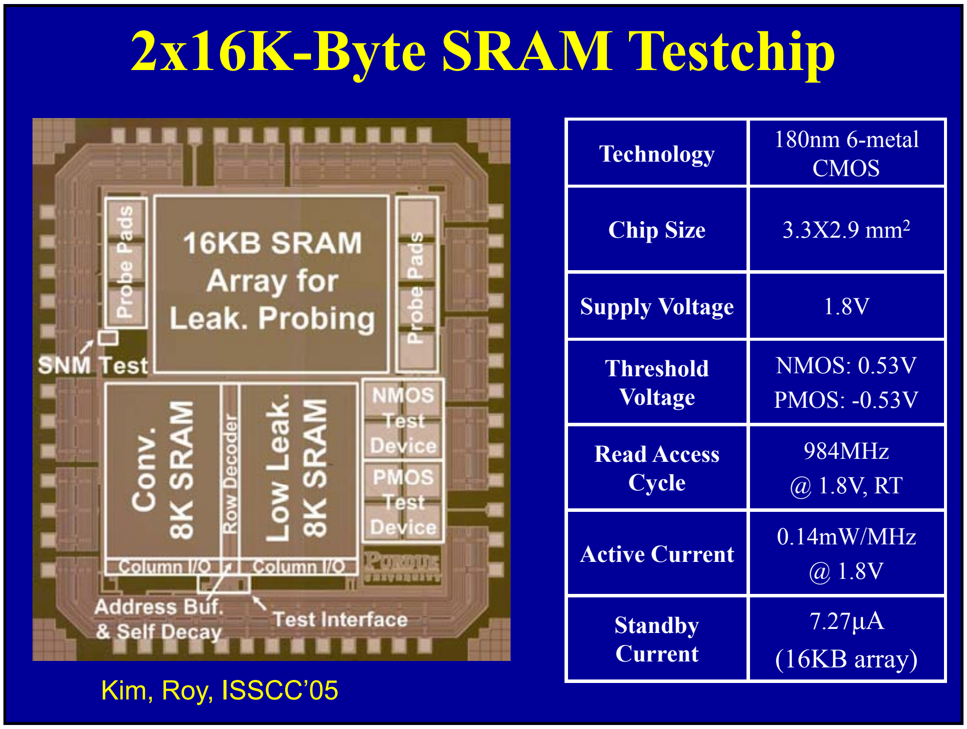
Gated-Ground Transistor Sharing





16K-Byte SRAM Organization

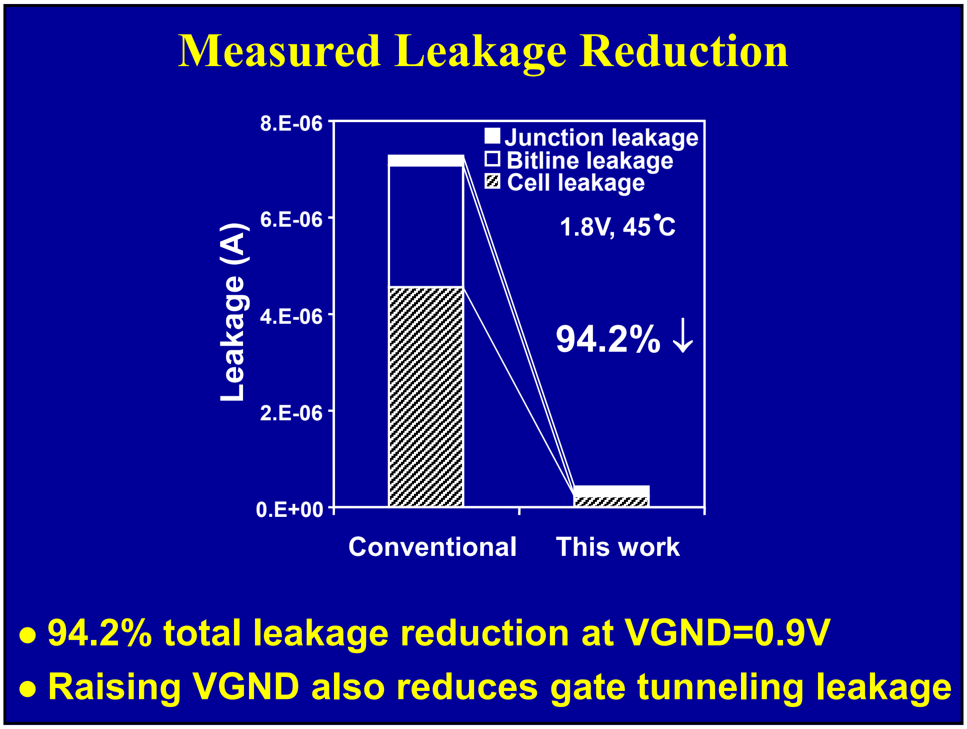
- Active leakage reduction SRAM
- Distributed sleep transistors
- SRAM block turned on ahead of time
- Self-decay circuit for low dynamic power overhead



2x16K-Byte SRAM Testchip

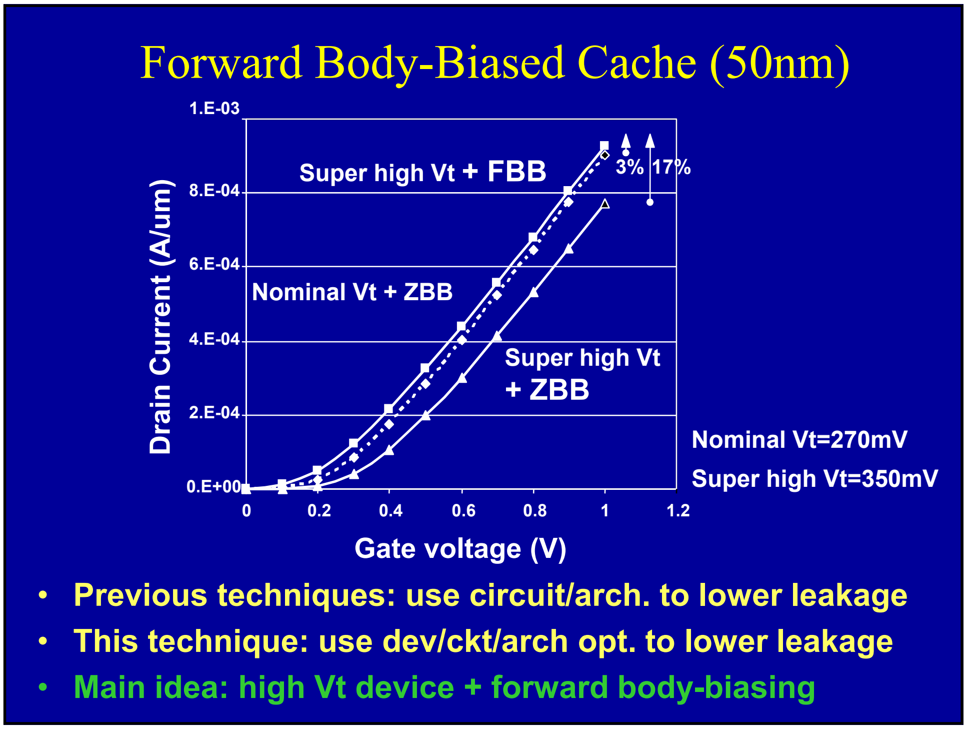
Technology	180nm 6-metal CMOS
Chip Size	3.3X2.9 mm^2
Supply Voltage	1.8V
Threshold Voltage	NMOS: 0.53V PMOS: -0.53V
Read Access Cycle	984MHz @ 1.8V, RT
Active Current	0.14mW/MHz @ 1.8V
Standby Current	7.27μA (16KB array)

Kim, Roy, ISSCC'05



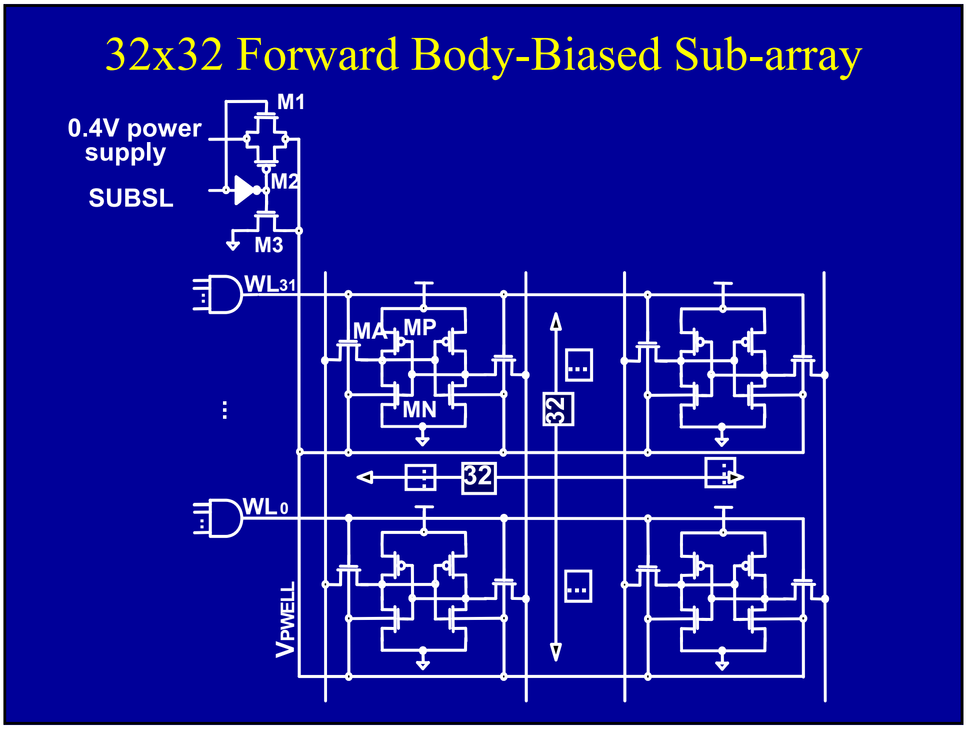
Measured Leakage Reduction

- 94.2% total leakage reduction at VGND=0.9V
- Raising VGND also reduces gate tunneling leakage

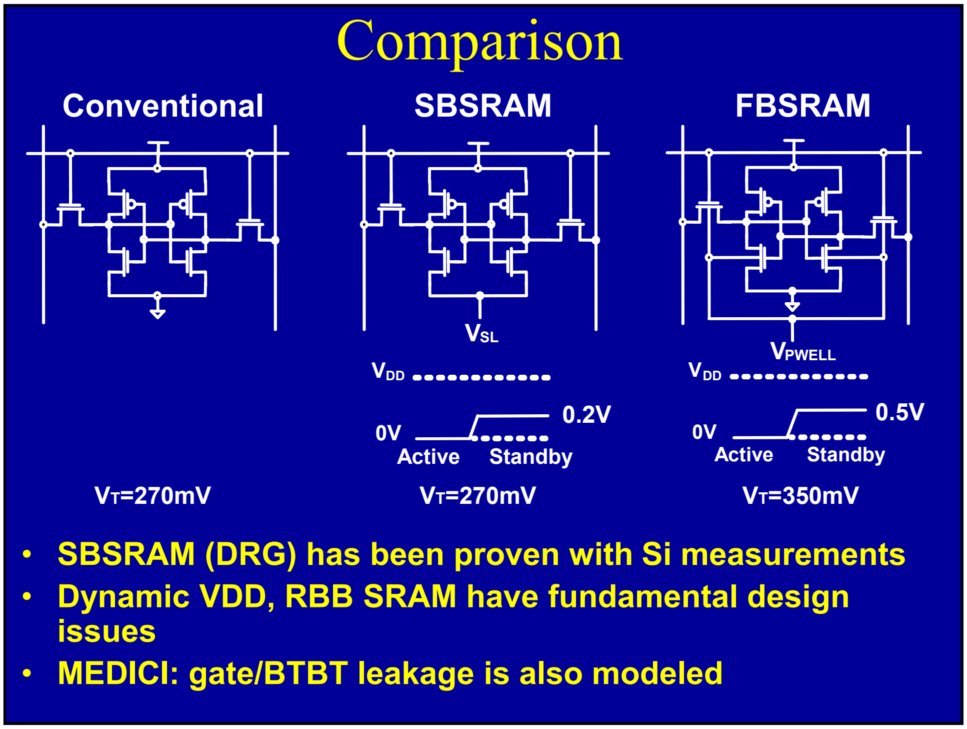


Forward Body-Biased Cache (50nm)

- 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



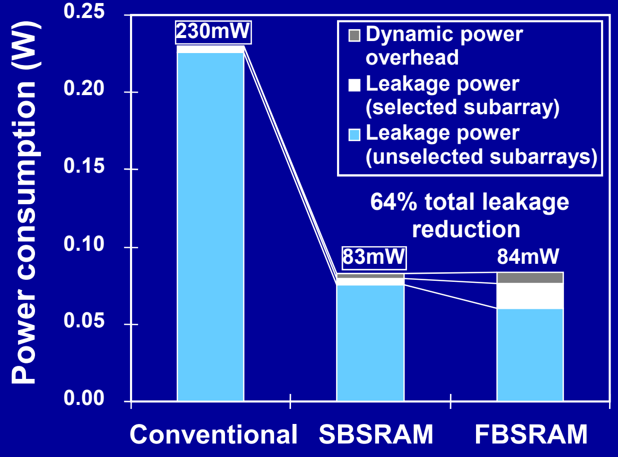
32x32 Forward Body-Biased Sub-array



Comparison

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

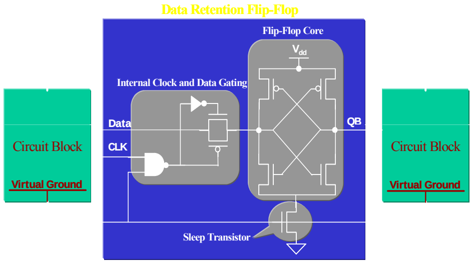
32KB Cache Total Leakage Reduction

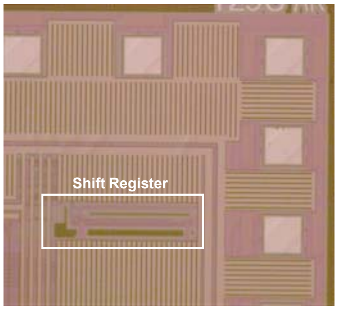


- SBSRAM and FBSRAM are designed to give iso-leakage savings
- 64% total leakage reduction including overhead

Another Application: Data Retention Flip-Flop

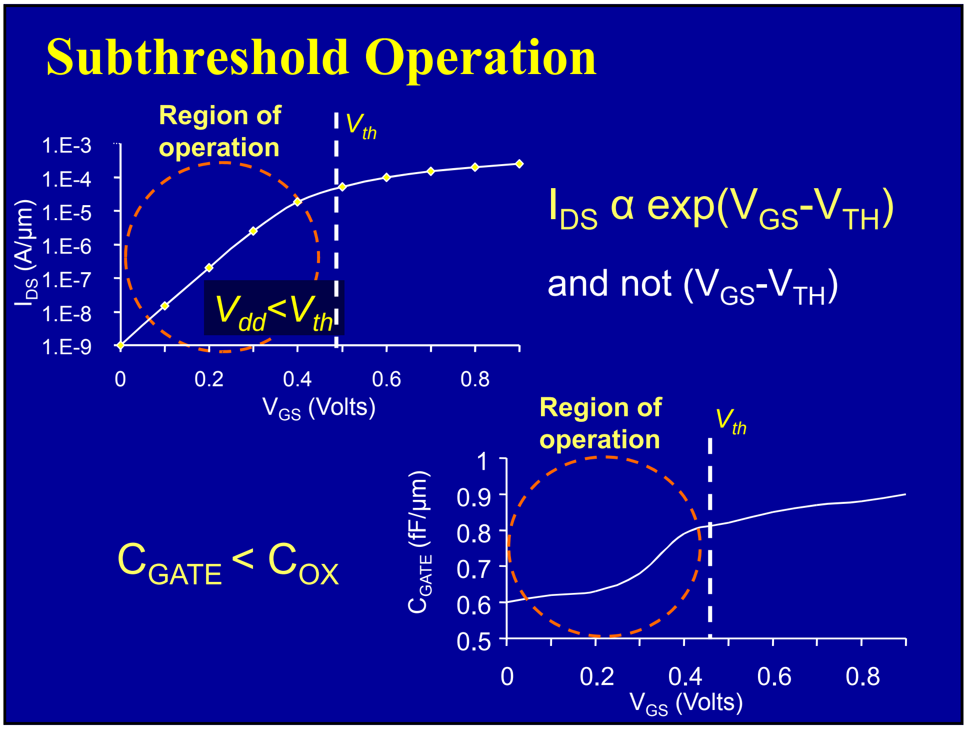
- Cross-coupled inverters are cores of any flip-flops
- Cross-coupled inverters retain data under gated ground
- Data and clock gating is required to preserve data
- Successful fabrication and test:
  - 16-bit shift-register based on our data-retention FF





40% power reduction by enabling power-down mode

Computing with Leakage for Ultralow Power: Digital Subthreshold Logic

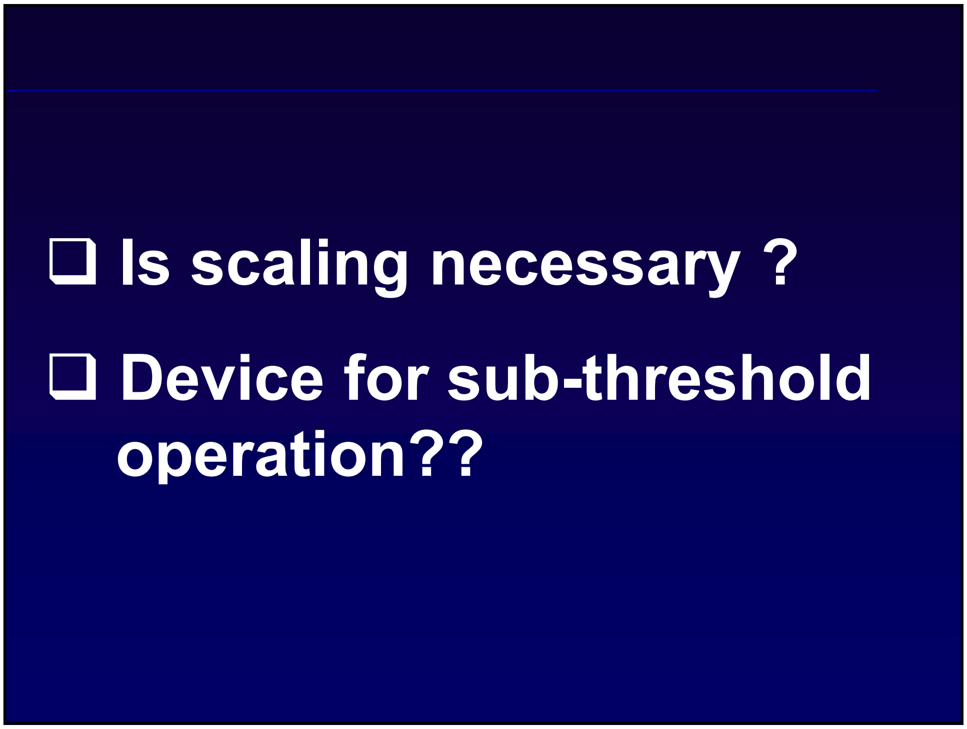


Subthreshold Operation

I_{DS} α $\exp(V_{GS}-V_{TH})$

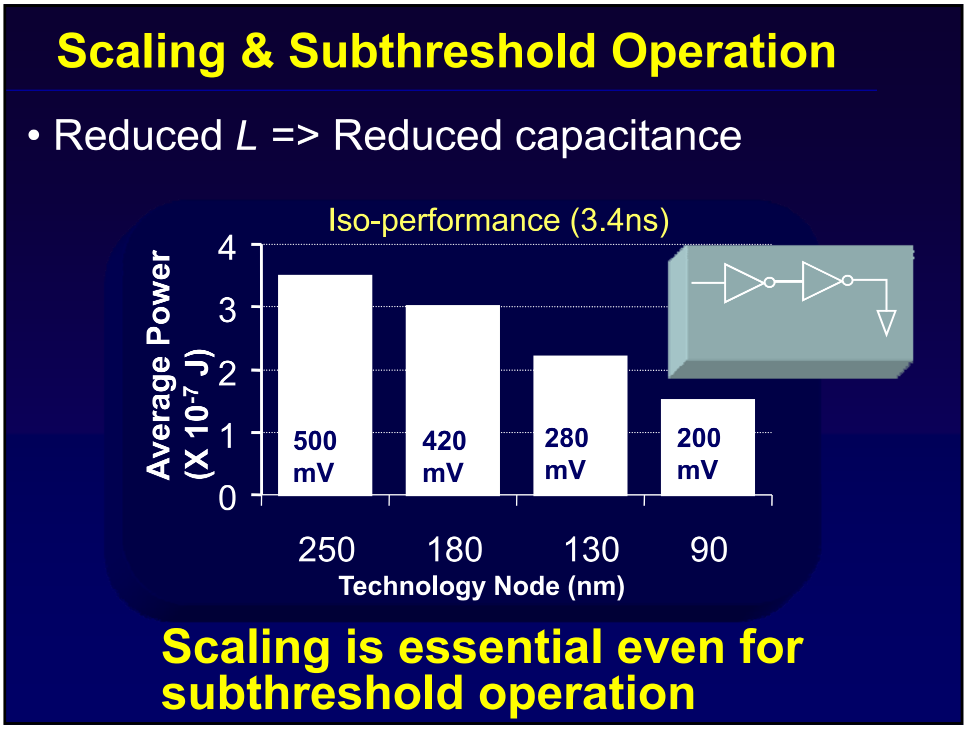
and not $(V_{GS}-V_{TH})$

$C_{GATE} < C_{OX}$



❑ Is scaling necessary ?

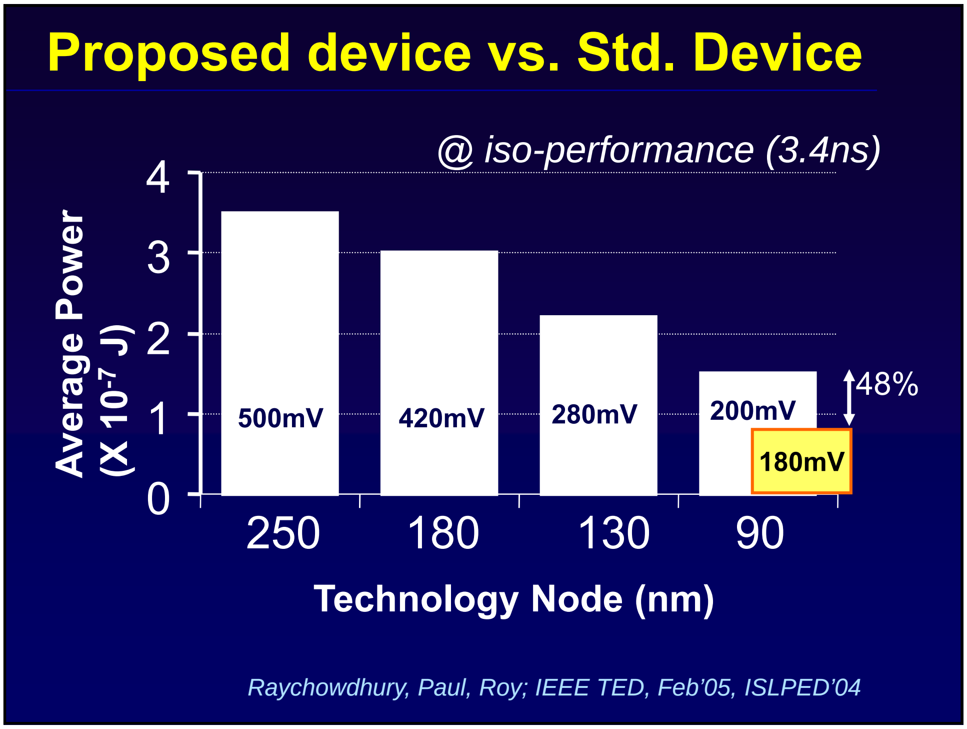
❑ Device for sub-threshold operation??



Scaling & Subthreshold Operation

- Reduced *L* => Reduced capacitance

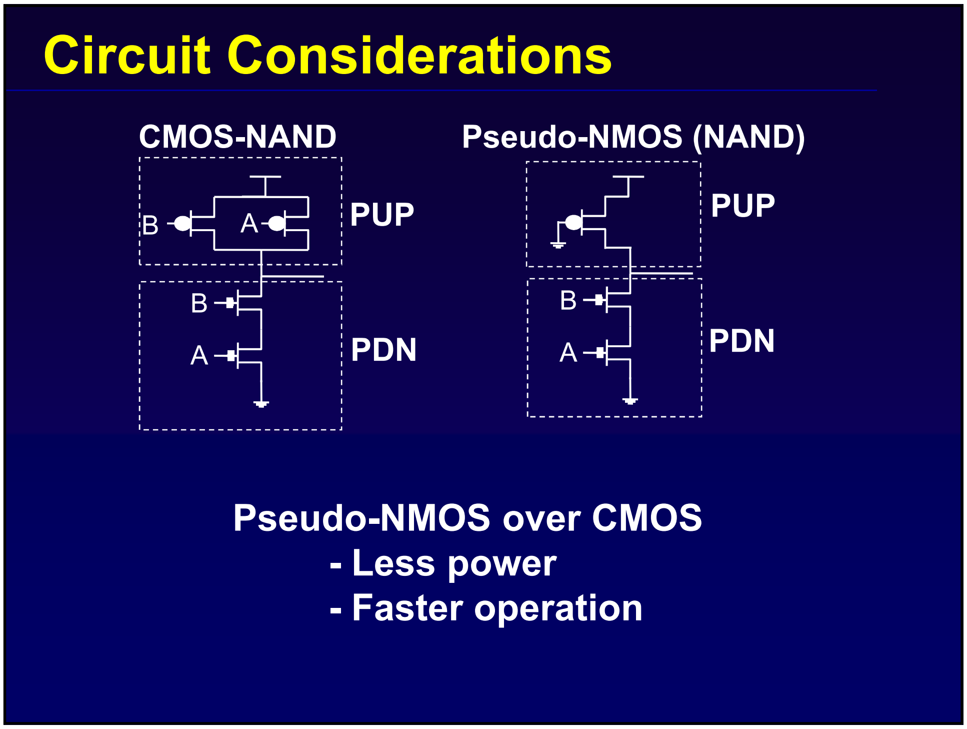
Scaling is essential even for subthreshold operation



Proposed device vs. Std. Device

@ iso-performance (3.4ns)

Raychowdhury, Paul, Roy; IEEE TED, Feb'05, ISLPED'04

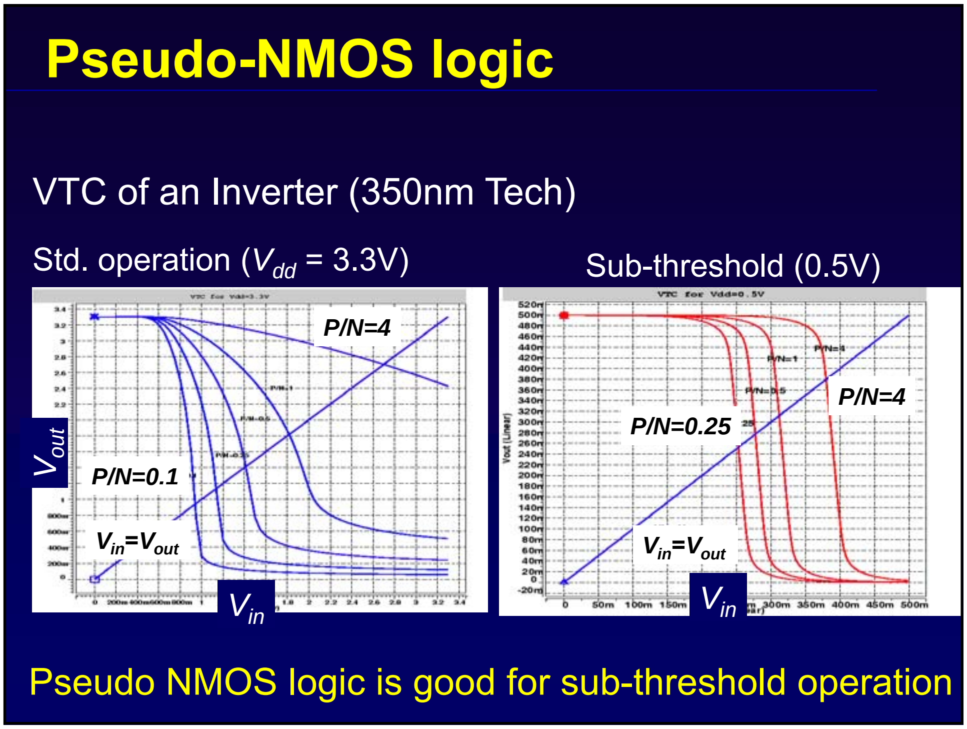


Circuit Considerations

Pseudo-NMOS over CMOS

- Less power

- Faster operation



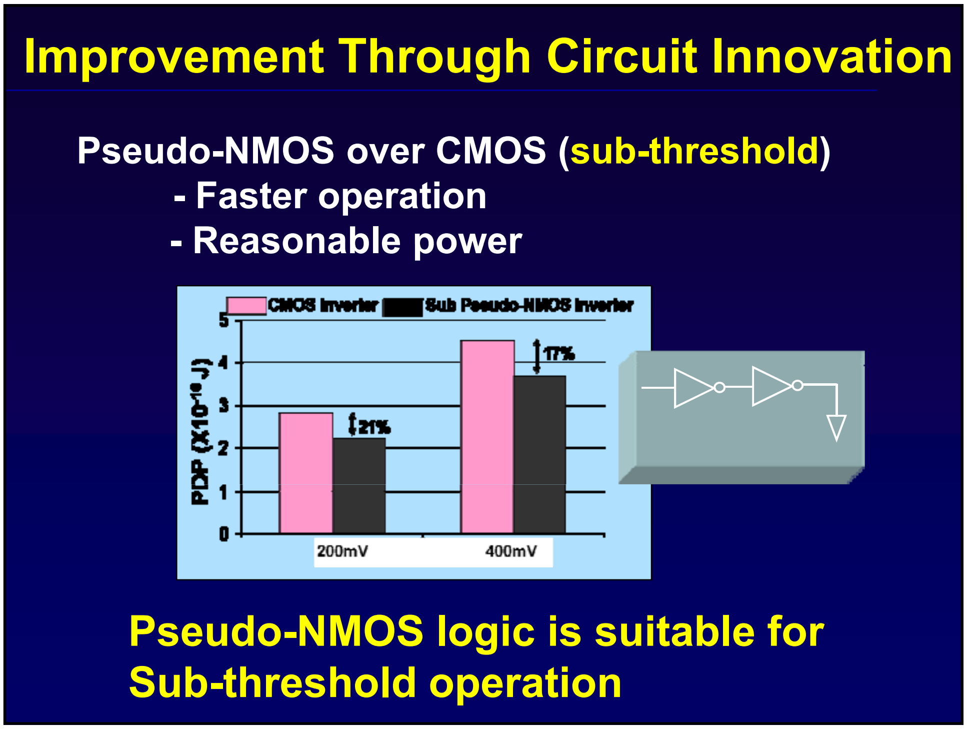
Pseudo-NMOS logic

VTC of an Inverter (350nm Tech)

Std. operation (V_{dd} = 3.3V)

Sub-threshold (0.5V)

Pseudo NMOS logic is good for sub-threshold operation



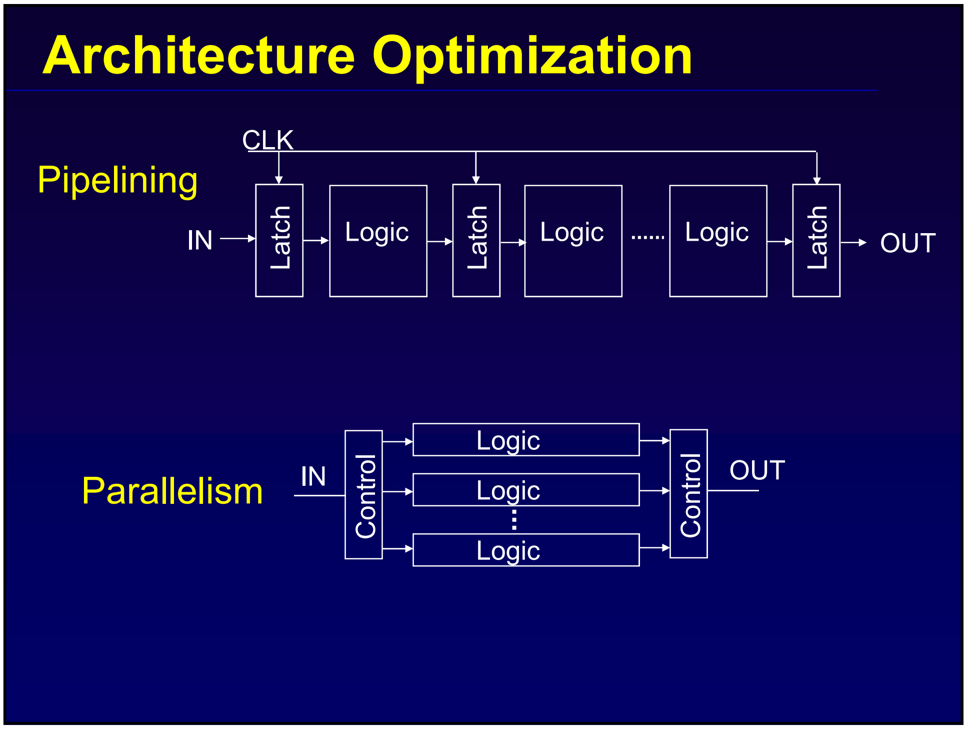
Improvement Through Circuit Innovation

Pseudo-NMOS over CMOS (sub-threshold)

- Faster operation

- Reasonable power

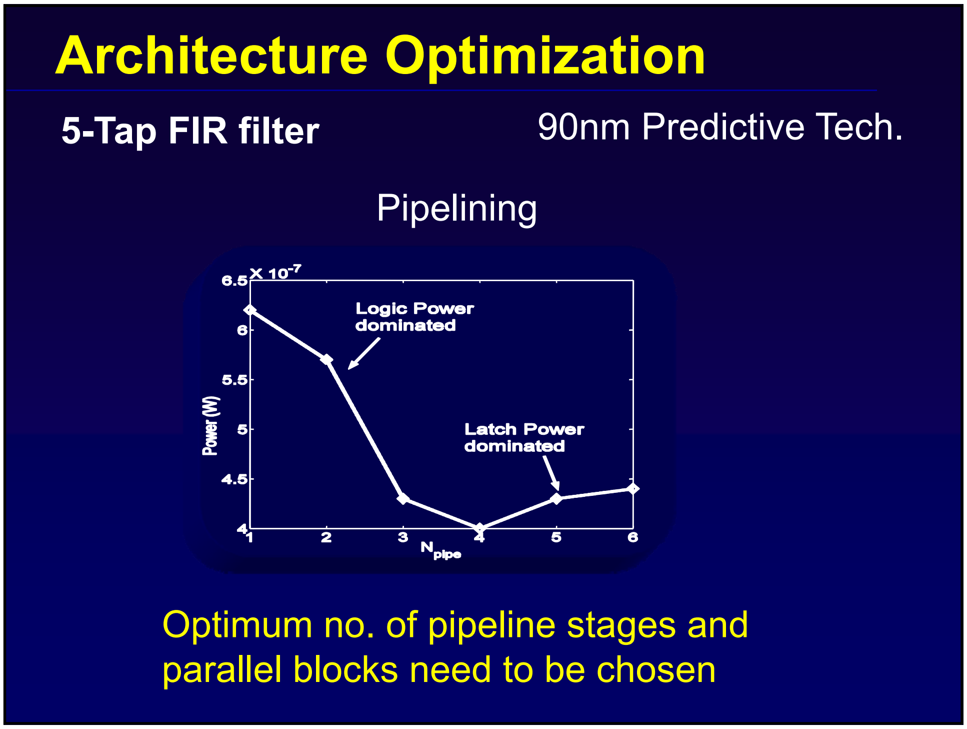
Pseudo-NMOS logic is suitable for Sub-threshold operation



Architecture Optimization

Pipelining

Parallelism



Architecture Optimization

5-Tap FIR filter

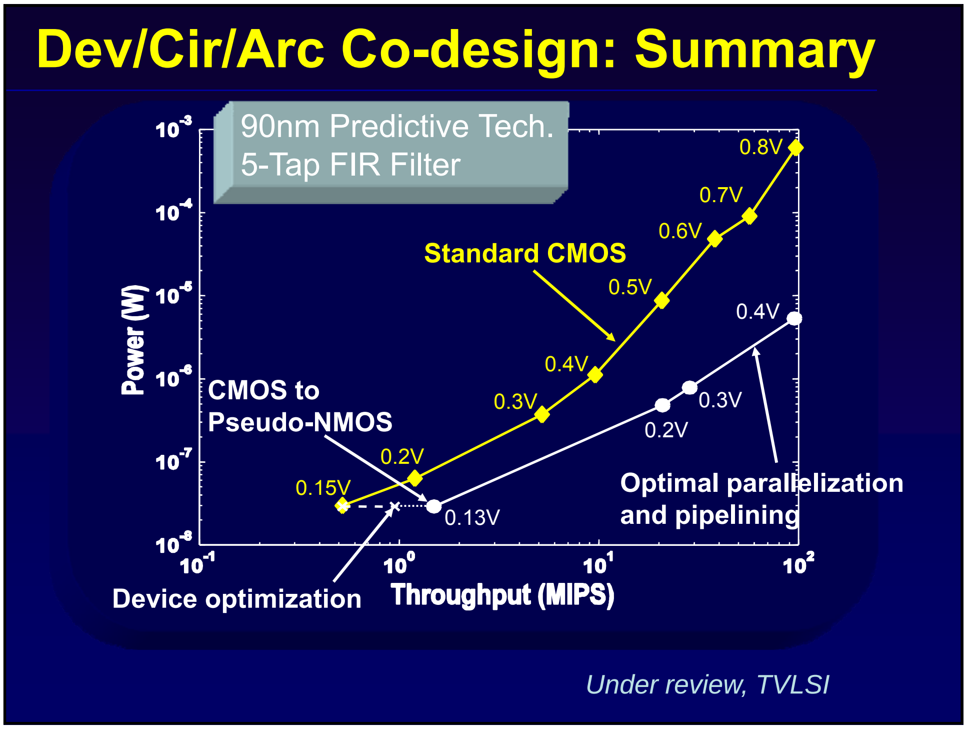
90nm Predictive Tech.

Pipelining

Logic Power dominated

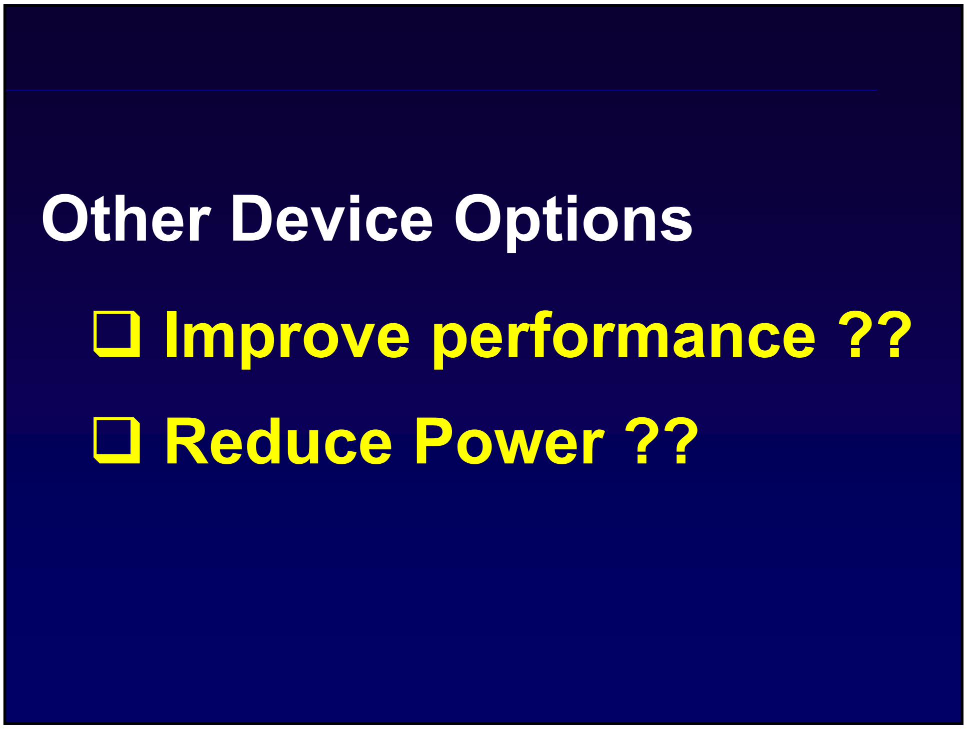
Latch Power dominated

Optimum no. of pipeline stages and parallel blocks need to be chosen



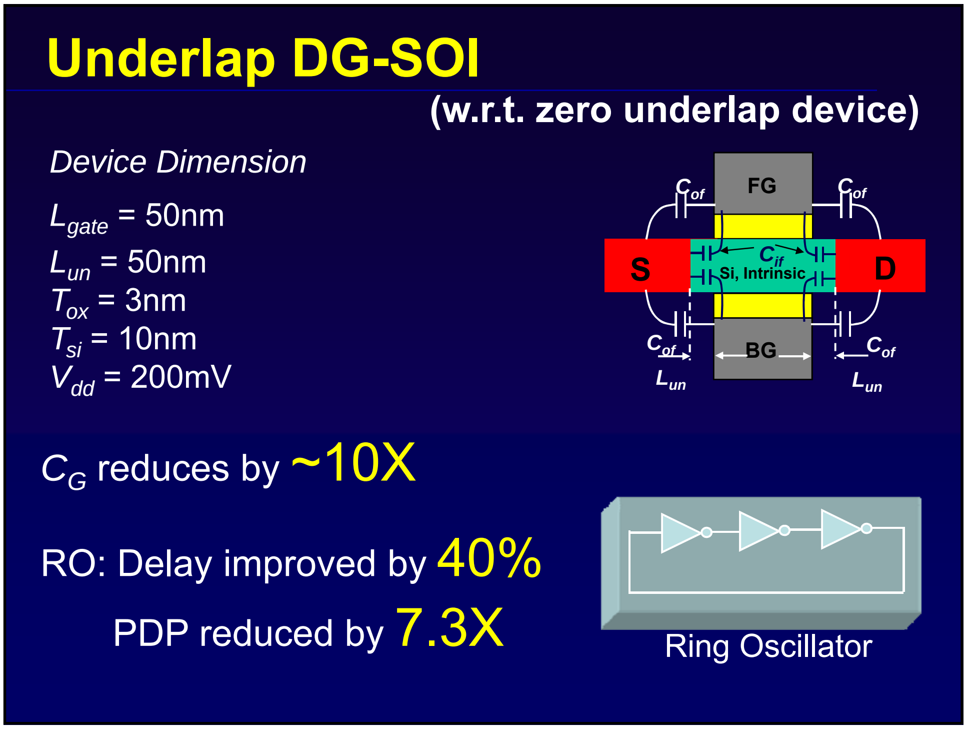
Dev/Cir/Arc Co-design: Summary

Under review, TVLSI



Other Device Options

- ❑ Improve performance ??
- ❑ Reduce Power ??



Underlap DG-SOI

(w.r.t. zero underlap device)

Device Dimension

L_{gate} = 50nm

L_{un} = 50nm

T_{ox} = 3nm

T_{si} = 10nm

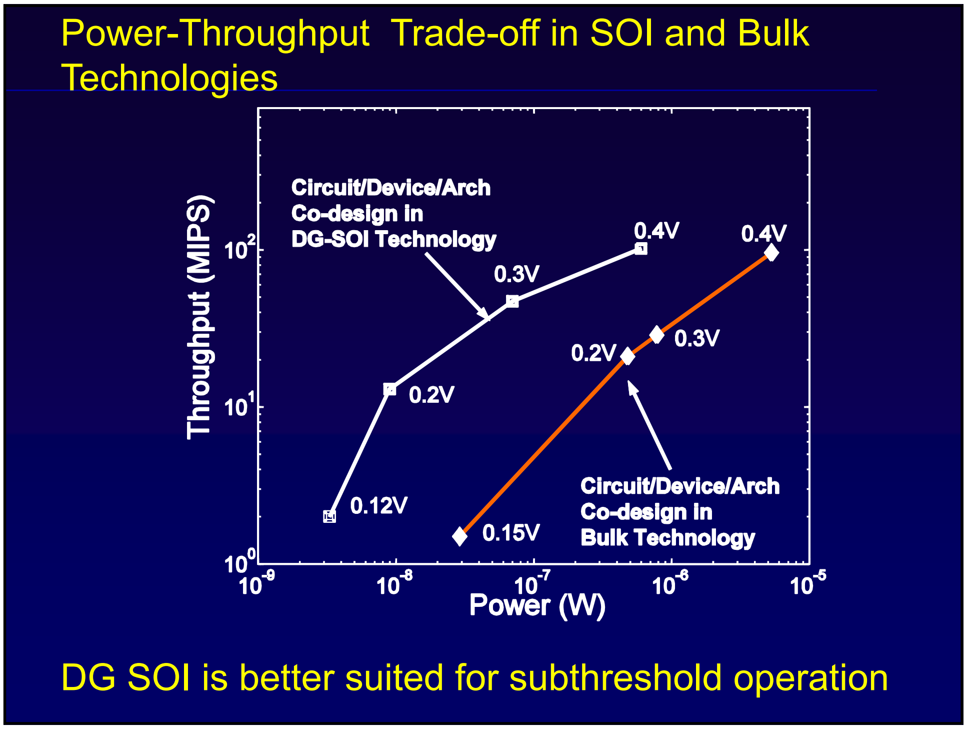
V_{dd} = 200mV

C_G reduces by ~10X

RO: Delay improved by 40%

PDP reduced by 7.3X

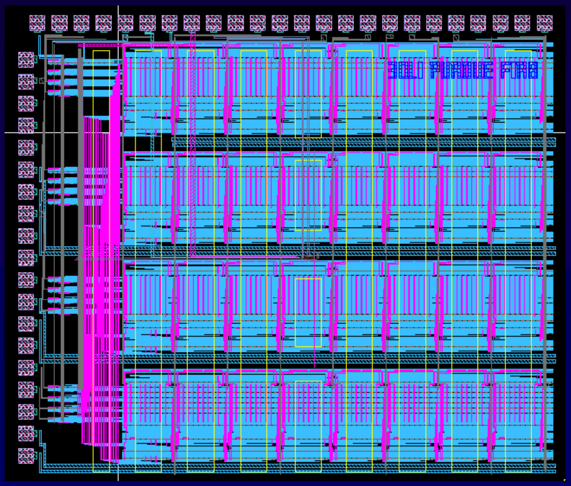
Ring Oscillator



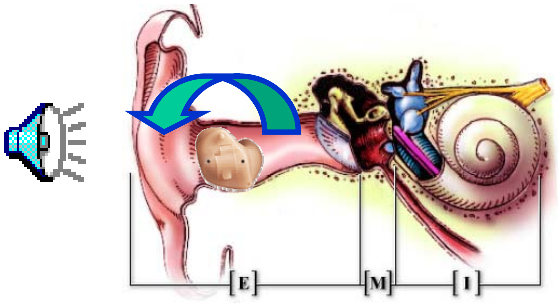
Power-Throughput Trade-off in SOI and Bulk Technologies

DG SOI is better suited for subthreshold operation

8 tap FIR in MITLL 3D FDSOI Process

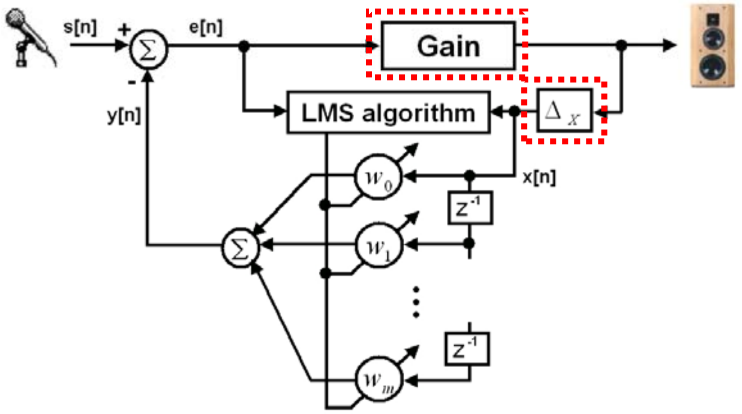


Example Application: Adaptive Filters in Digital Hearing Aid Devices



- Adaptive filters are used to cancel out the annoying high intensity oscillation
  - Acoustic feedback through the human body
  - Hearing aid output leaking into the input again

Prototype Adaptive Filter For Hearing Aid Devices



- Subtracts the unwanted acoustic feedback noise
- Reference signal : delayed error output

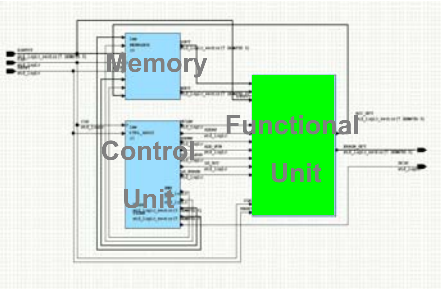
Filter Architecture With Single Functional Unit

	LMS (Least Mean Square)
# of FU	Single
Algorith m	$W(n+1) = W(n) + \mu e(n)U(n)$ $e(n) = d(n) - W^T(n)U(n)$

$W(n) = [\omega_0(n)\ \omega_1(n)\ \ \dots\ \ \omega_{N-1}(n)]^T$: *Filter coefficients*

$U(n) = [u(n)\ u(n-1)\ \ \dots\ \ u(n-N+1)]^T$: *Data input*

t_m : *Multiplier delay*, t_a : *Adder delay*, N : *Filter length*



LMS filter with a single FU

CLK = 22kHz*34cycle/sample

= 748 kHz

- Not suitable for ultra-low voltage operation
- LMS algorithm cannot be implemented in a parallel architecture

Filter Architecture With Multiple Functional Units

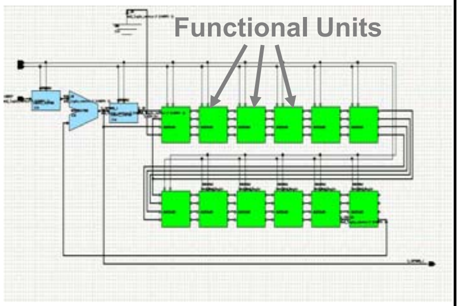
	DLMS (Delayed Least Mean Square)
# of FU	Multiple
Algorithm	$W(n+1) = W(n) + \boxed{\mu e(n-N)U(n-N)}$ $e(n-N) =$ $d(n-N) - W^T(n-N)U(n-N)$

$W(n) = [\omega_0(n)\ \omega_1(n)\ \ \ldots\ \ \omega_{N-1}(n)]^T$: *Filter coefficients*

$U(n) = [u(n)\ u(n-1)\ \ \ldots\ \ u(n-N+1)]^T$: *Data input*

t_m : *Multiplier delay*, t_a : *Adder delay*, N : *Filter length*

M. Meyer, et al., IEEE Trans. Circuits Syst. II, 1993



DLMS filter with multiple FU

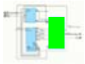
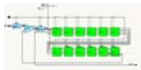
CLK = 22kHz*1cycle/sample

= 22 kHz

- **DLMS algorithm enables parallel architecture**
- **Trading off area for power**

Power Consumption

- Architecture & Logic Styles -

Implementation	Clock frequency	Vdd	Energy /Operation	# of Transistors
 + Sub-CMOS	748 kHz	650 mV	19.1 nJ	31k
+ Sub-CMOS	22 kHz	450 mV	2.47 nJ	111k
 + Sub-Pseudo NMOS	22 kHz	400 mV	1.77 nJ	86k

- **Parallel architecture lowers the clock rate, reduces power dissipation by 87%**
- **Pseudo NMOS logic styles provides another 28%reduction**