

EE241B : Advanced Digital Circuits

Lecture 23 – Sleep Modes

Borivoje Nikolić



Wave Computing and MIPS Wave Goodbye

by Mike Gianfagna on 04-19-2020 at 8:00 Word on the virtual street is that Wave Computing is closing down. The company has reportedly let all employees go and filed for Chapter 11. As one of the many promising new companies in the field of AI, Wave Computing was founded in 2008 with the mission “to revolutionize deep learning with real-time AI solutions that scale from the edge to the datacenter.”



Announcements

- Assignment 4 due on Friday.
- Reading
 - Rabaey, LPDE, Chapter 8

Outline

- **Module 5**
 - Sleep modes
 - Optimal thresholds and supplies

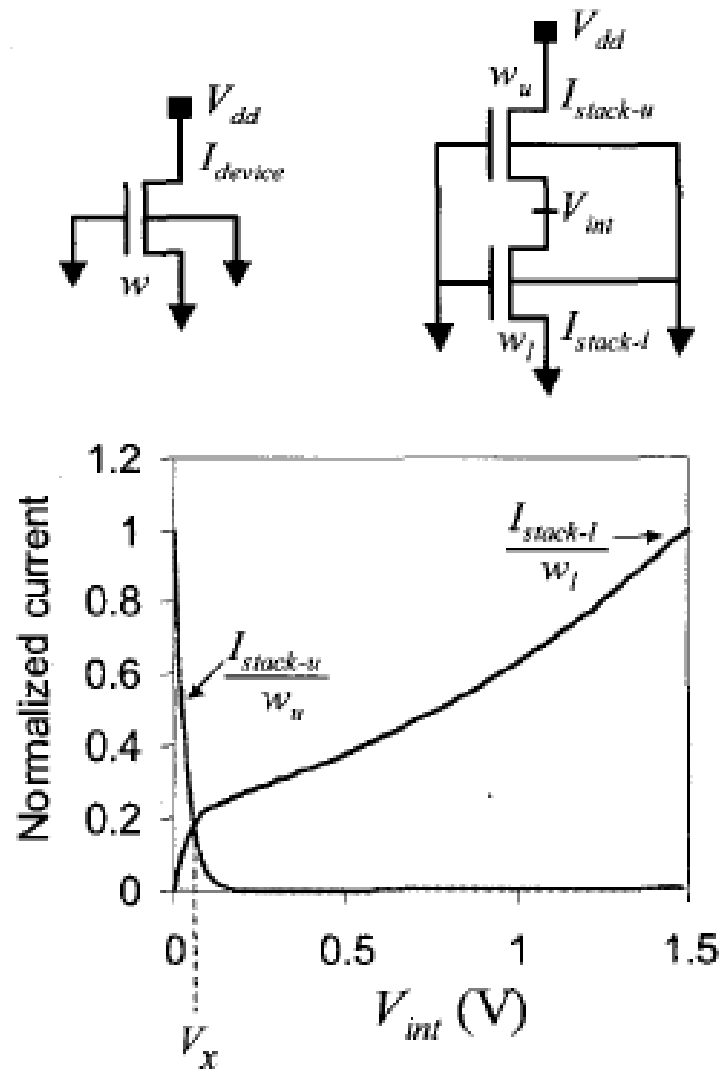


5.J Lowering Leakage During Design: Transistor Stacking

Power /Energy Optimization Space

	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Sleep Mode	Run Time
Active	Logic design Scaled V_{DD} Trans. sizing Multi- V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi- V_{Th}	Sleep T's Multi- V_{DD} Variable V_{Th} + Input control	+ Variable V_{Th}

Stack Effect

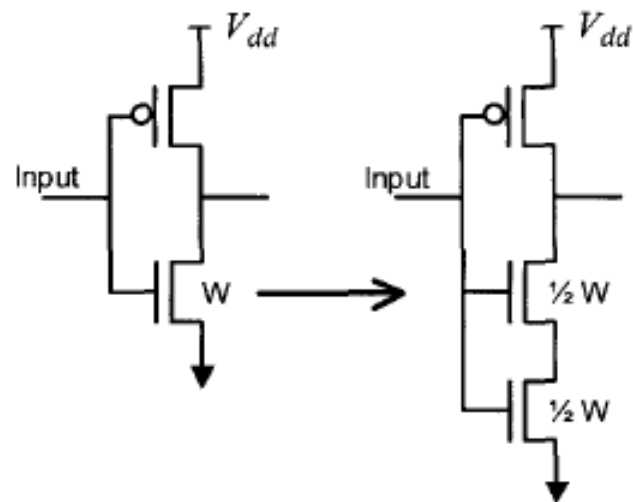


Reduction (in 0.13 μ):

	High V_t	Low V_t
2 NMOS	10.7X	9.96X
3 NMOS	21.1X	18.8X
4 NMOS	31.5X	26.7X
2 PMOS	8.6X	7.9X
3 PMOS	16.1X	13.7X
4 PMOS	23.1X	18.7X

Narendra, ISLPED'01

Stack Forcing



Tradeoffs:

- $W/2$ – $1/3$ of drive current, same loading
- $1.5W$ – $3x$ loading, same drive current

Narendra, ISLPED'01



5.L Power Gating

Power /Energy Optimization Space

	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Sleep Mode	Run Time
Active	Logic design Scaled V_{DD} Trans. sizing Multi- V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi- V_{Th}	Sleep T's Multi- V_{DD} Variable V_{Th} + Input control	+ Variable V_{Th}

Power Gating with Sleep Transistors

- Key components:
 - Power gates (& controller)
 - Leakage vs size
 - Switched cap
 - Slew-rate/rush current
 - State preservation
 - Energy overhead of sleep/wake-up transitions

How to Size the Sleep Transistor?

- Don't need both header and footer
- Circuits in active mode see the sleep transistor as extra power line resistance
 - The wider the sleep transistor, the better
- Wide sleep transistors cost area and are slow to turn on/off
 - Minimize the size of the sleep transistor for given ripple (e.g. 5%)
- Need to find the worst case vector
- Sleep transistor is not for free – it will degrade the performance in active mode
- Charging and discharging the virtual rails costs energy
- Need to sequentially wake up

Sleep Transistor

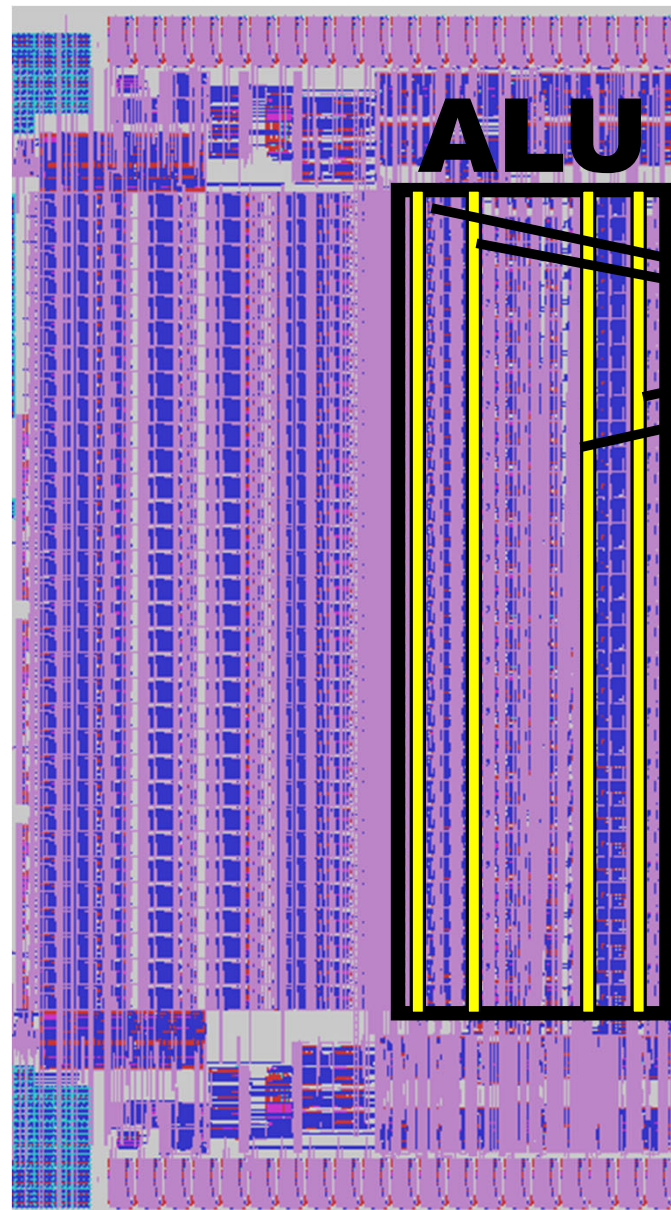
High- V_{TH} transistor (many in parallel) has to be very large for low resistance in linear region.

Low- V_{TH} transistor needs much less area for the same resistance.

	MTCMOS	Boosted Sleep	Non-Boosted Sleep
Sleep-TR size	5.1%	2.3%	3.2%
Leakage power reduction	1450X	3130X	11.5X
Virtual supply bounce	60 mV	59 mV	58 mV

Courtesy: R. Krishnamurthy, Intel

Sleep Transistor Layout



ALU

**Sleep
transistor
cells**

Area overhead	
PMOS	6%
NMOS	3%

Tschanz, ISSCC'03

Sleep in Standard Cells

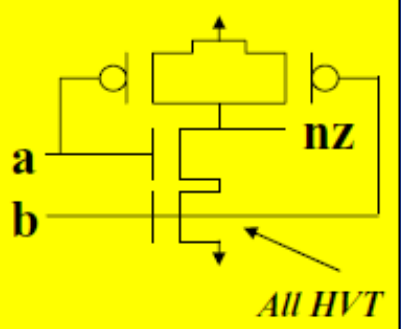
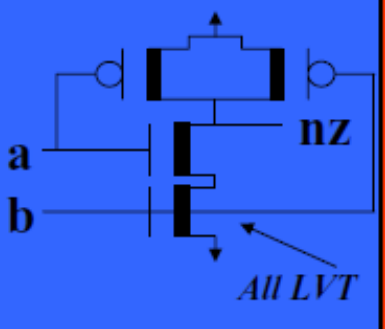
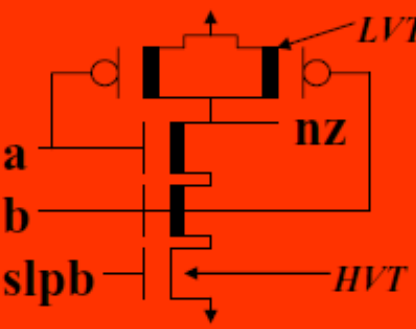
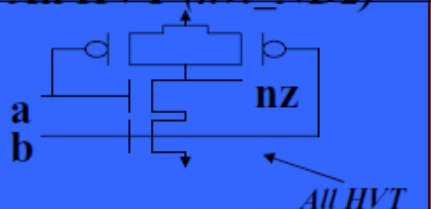
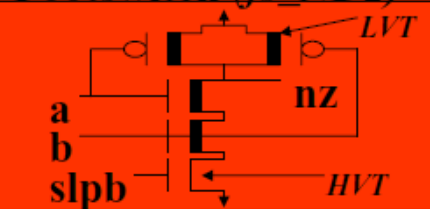
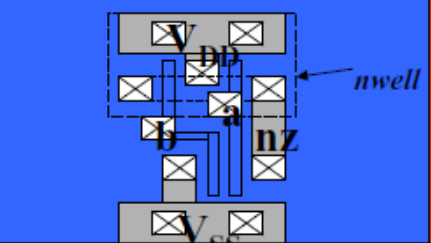
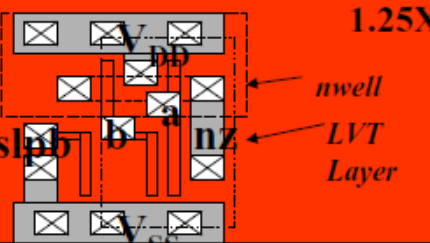
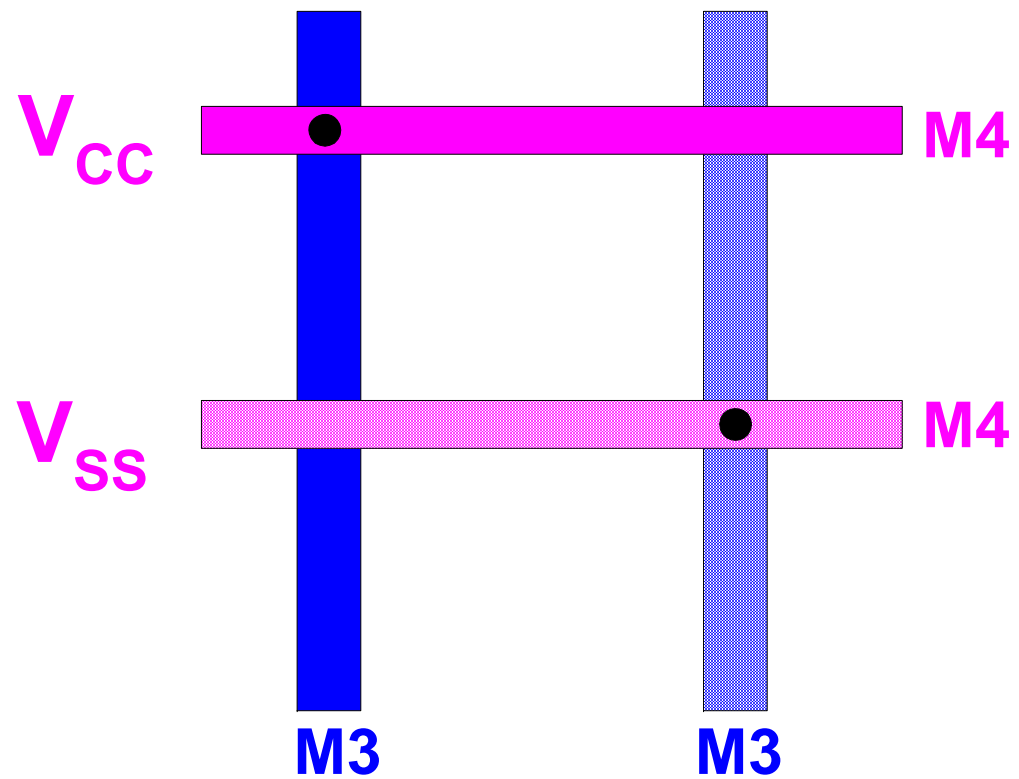
	All HVT (<i>hvt_ND2</i>)	All LVT (<i>lvt_ND2</i>)	Footswitch (<i>fs_ND2</i>)
Schematics			
Perf.	1X	1.5X - 2X	1.4X - 1.8X
Leakage	1X	70X - 100X	≈ 1X
Area	1X	1X	1.25X

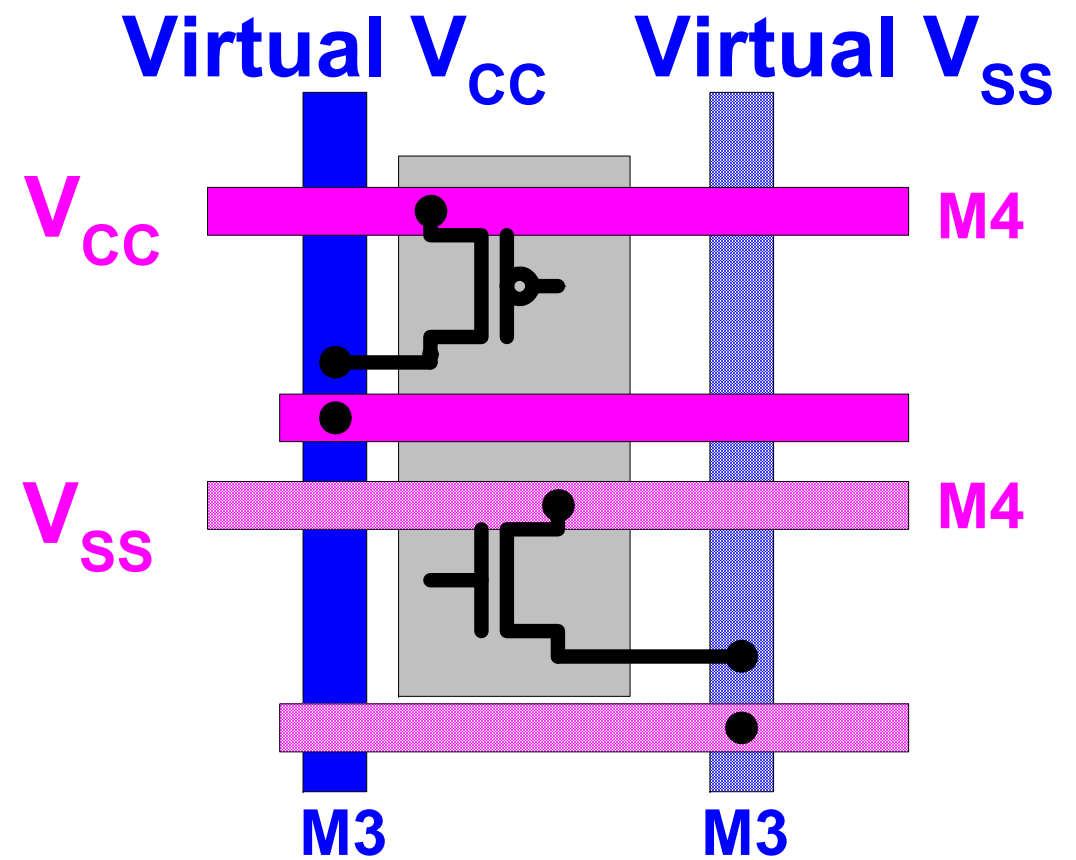
Diagram	All HVT (<i>hvt_ND2</i>)	Footswitch (<i>fs_ND2</i>)
SCHEMATICS		
LAYOUT		

Sleep Transistor Grid

No sleep transistor



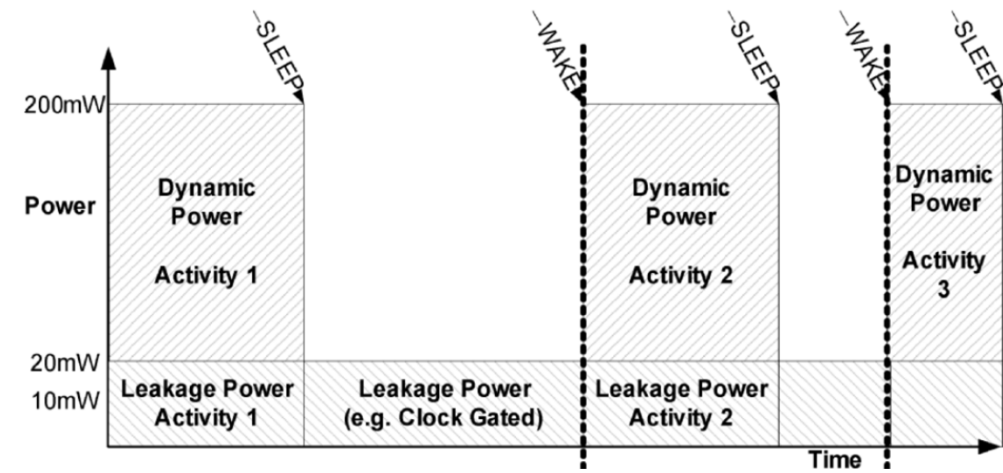
PMOS & NMOS
sleep transistors



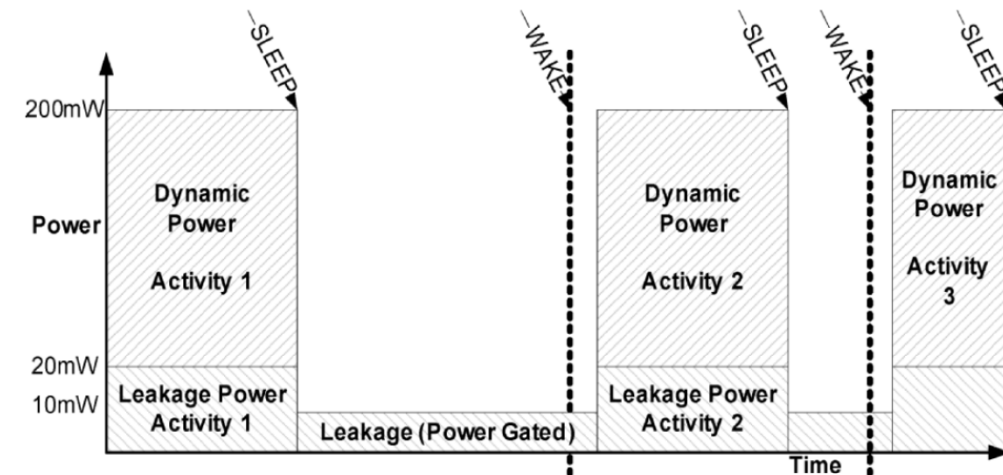
Tschanz, ISSCC'03

Power Gating

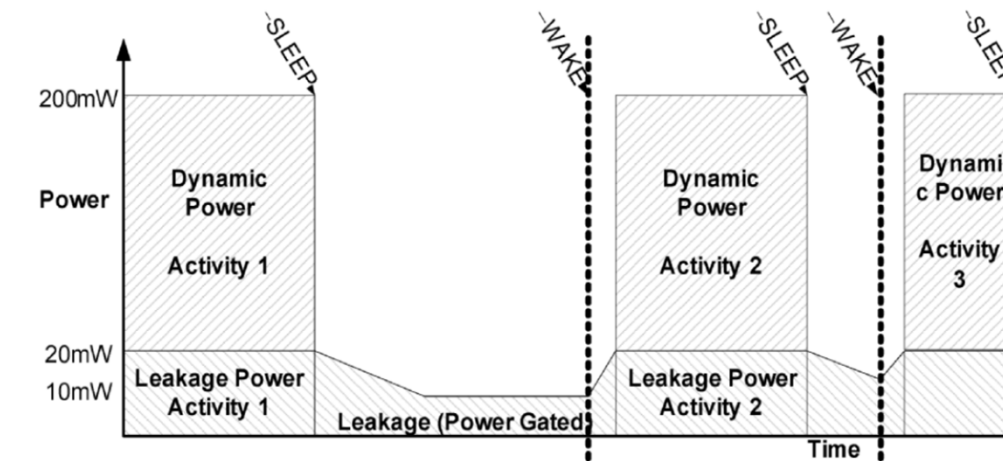
➤ No power gating



➤ “Ideal” power gating



➤ Realistic profile

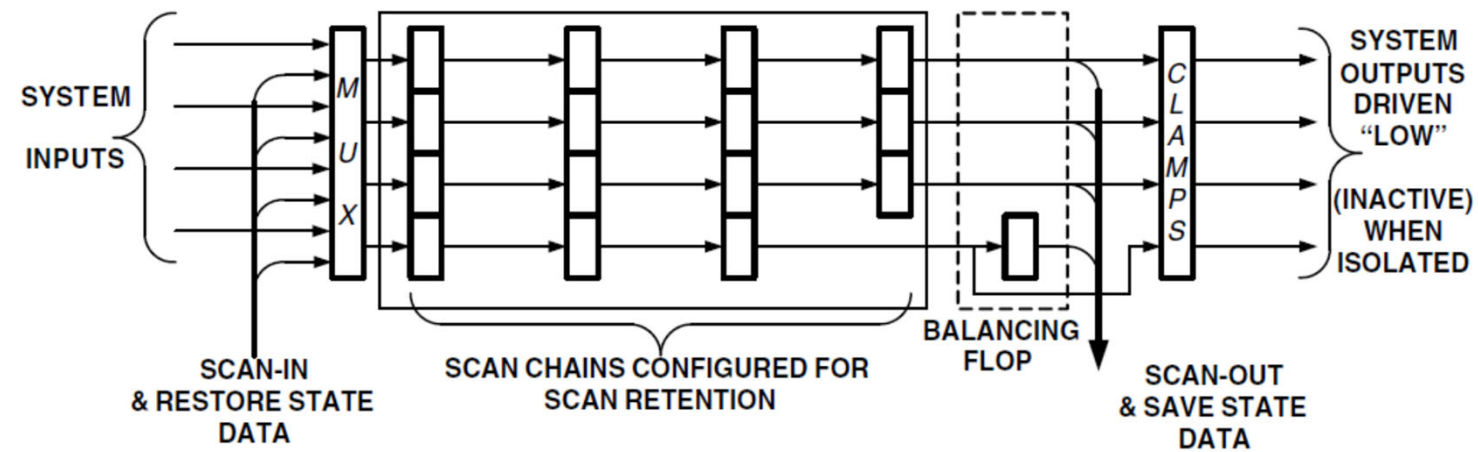


Preserving State

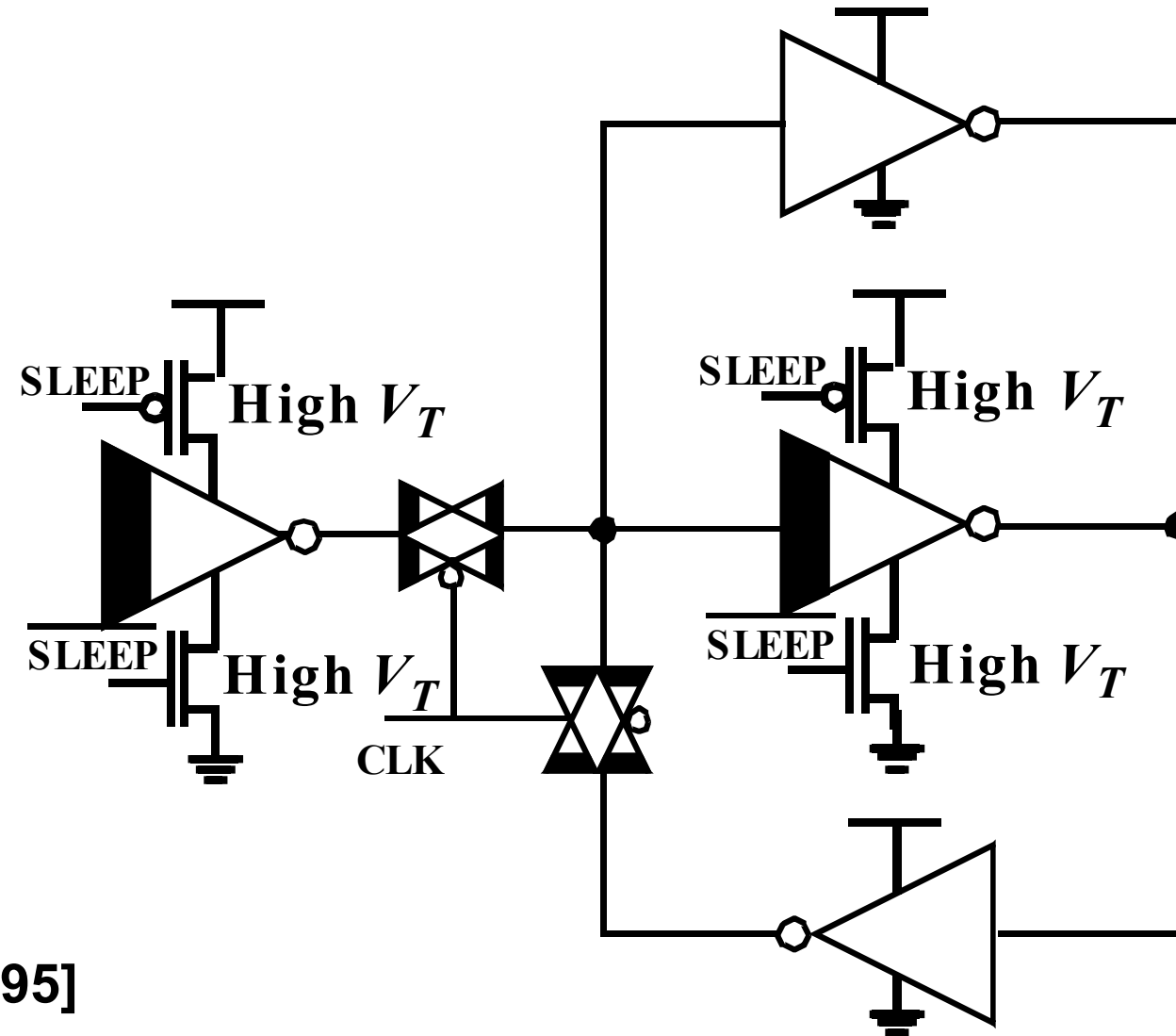
- Virtual supply collapse in sleep mode will cause the loss of state in registers
- Putting the registers at nominal VDD would preserve the state
 - These registers leak
 - The second supply needs to be routed as well
- Can lower VDD in sleep
 - Some impact on robustness, noise and SEU immunity
- State preservation and recovery

Scan-Based Retention

- Scan-out/scan-in state to preserve/restore state

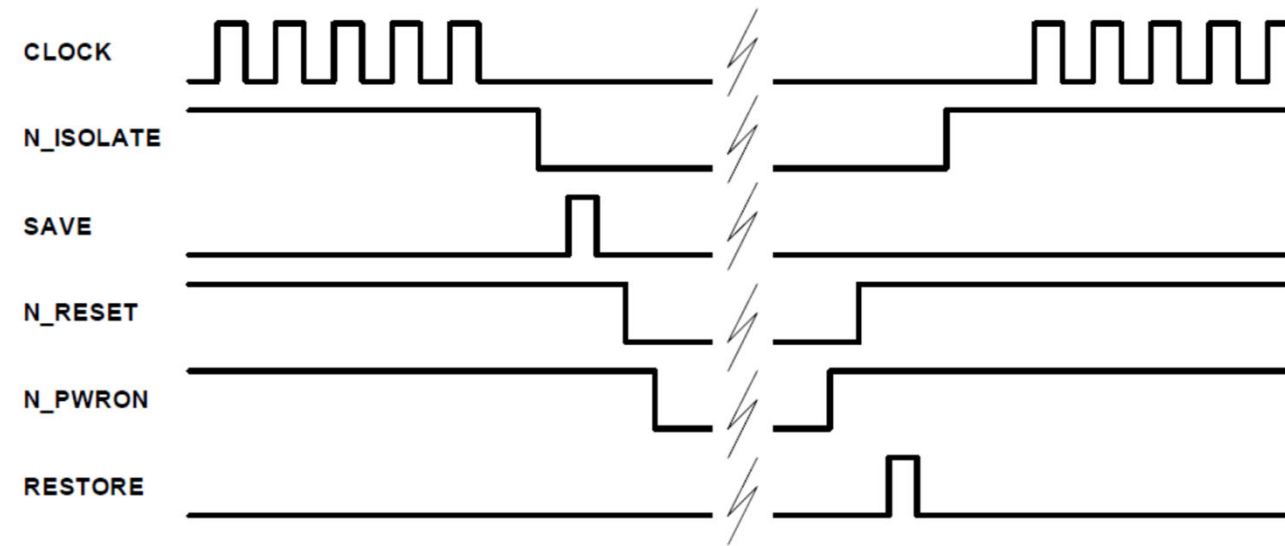


Retention Register Design



[Mutoh95]

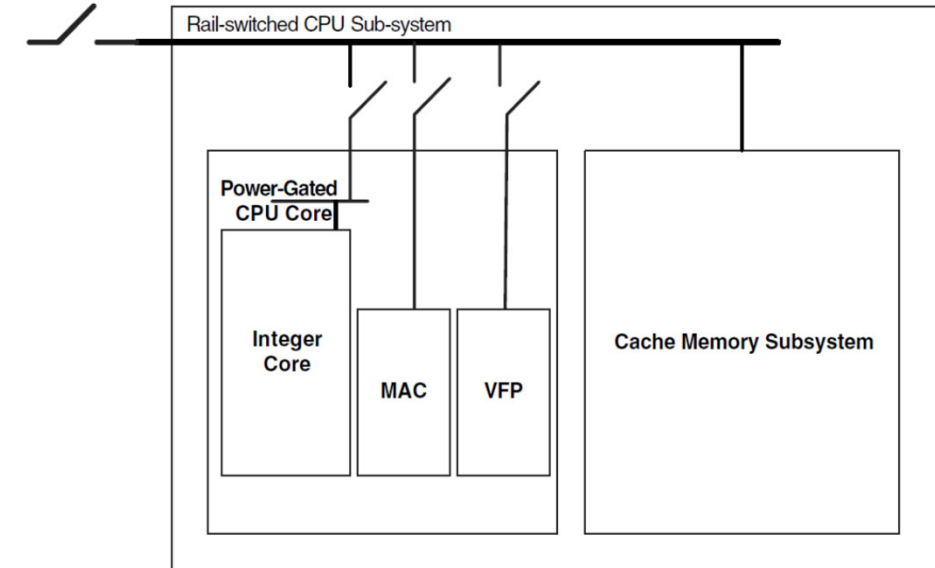
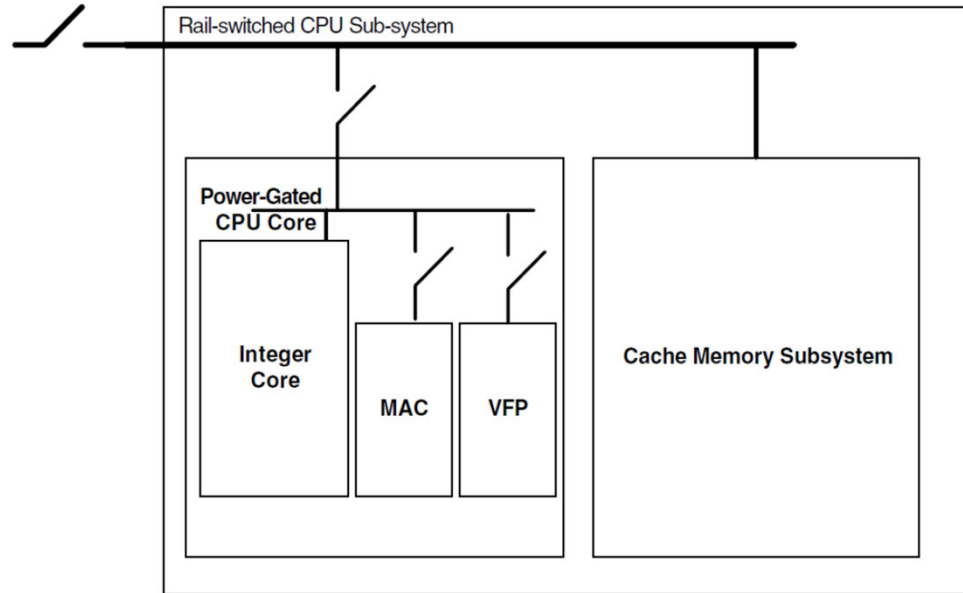
Gating Sequences



- **Sequence of steps:**

- Gate clock
- Isolate inputs
- Save (scan out)
- Reset
- Gate power

Hierarchical Power Gating



Cache	CPU	MAC	VFP	Power State
(OFF)	(OFF)	-	-	Shutdown (Cache cleaned, VDDCPU off)
ON	OFF	-	-	Deep Sleep (Cache preserved)
ON	ON	OFF	OFF	Normal Operation
ON	ON	ON	OFF	DSP workload
ON	ON	OFF	ON	Graphics workload
ON	ON	ON	ON	Intensive multimedia mode

Cache	CPU	MAC	VFP	Power State
(OFF)	(OFF)	(OFF)	(OFF)	Shutdown (Cache cleaned, VDDCPU off)
ON	OFF	OFF	OFF	Deep Sleep (Cache preserved)
ON	ON	OFF	OFF	Normal Operation
ON	ON	ON	OFF	DSP workload
ON	ON	OFF	ON	Graphics workload
ON	ON	ON	ON	Intensive multimedia mode

Project reports

- Due *May 4*, up to 6 pages
- Presentations on *May 4* in the afternoon
 - 12min + 3 min Q&A
 - 15min for 3-person teams



5.M Dynamic Threshold Scaling

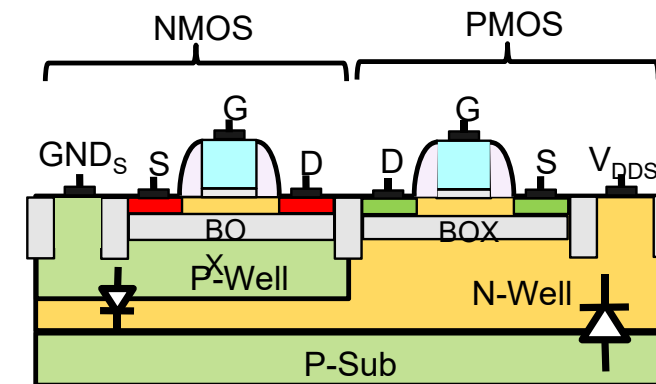
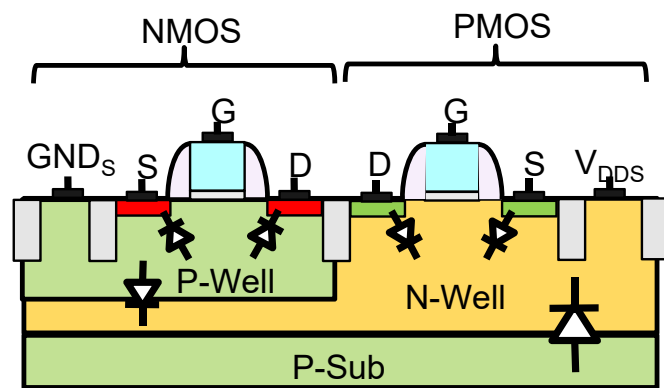
Power /Energy Optimization Space

	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Sleep Mode	Run Time
Active	Logic design Scaled V_{DD} Trans. sizing Multi- V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi- V_{Th}	Sleep T's Multi- V_{DD} Variable V_{Th} + Input control	DVS< Variable V_{Th}

Dynamic Body Bias

- Similar concept to dynamic voltage scaling
- Control loop adjusts the substrate bias to meet the timing/leakage goal
 - Can be used just as runtime/sleep
- Limited range of threshold adjustments in bulk ($<100\text{mV}$)
 - Limited leakage reduction ($<10\text{x}$)
- Works well in FDSOI ($80\text{-}85\text{mV/V}$, with $\sim 1.8\text{V}$ range)
- No delay penalty
 - Can increase speed by forward bias
- Energy cost of charging/discharging the substrate capacitance
 - but doesn't need a regulator

FDSOI and Bulk



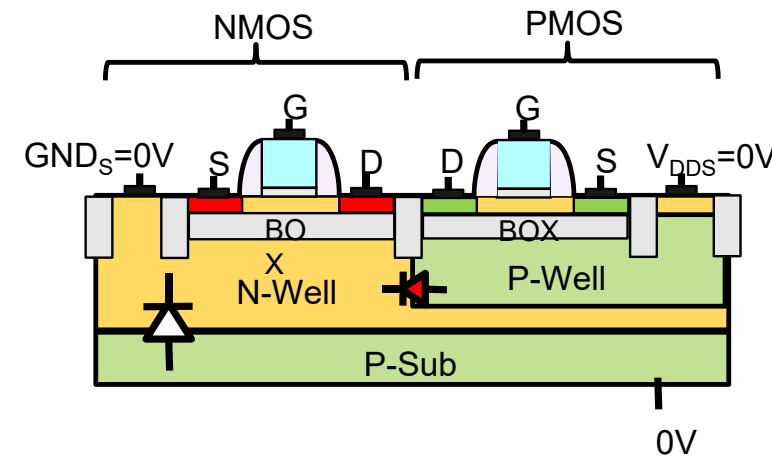
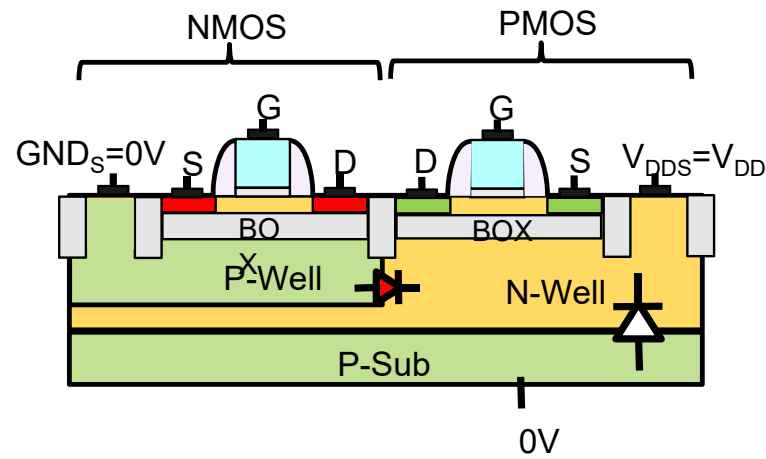
- Bulk CMOS

- Leakage paths through bulk
- RDF dominates local variability
- Diodes and B2B tunneling limit back-bias range

- UTBB FD-SOI

- Thin body for short-channel control
- No doping – less RDF
- Extended back-bias range

FDSOI Wells and Back Bias



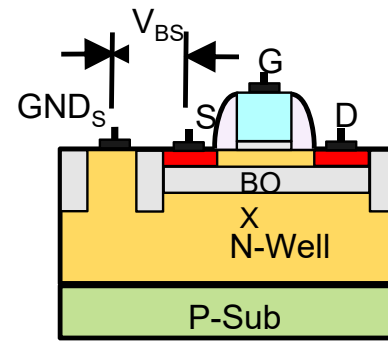
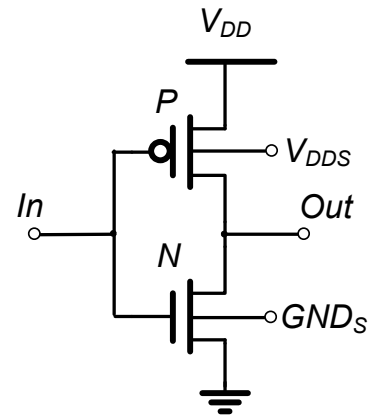
Typical (RVT)

- ▶ $GND_{S,nom} = 0V, V_{DDS,nom} = V_{DD}$
- ▶ Reverse body bias, $V_{BSN} < 0V$
- ▶ $(-3V) < GND_S < V_{DD}/2 + 0.3V$
 - ▶ Limit due to diodes, BOX
- ▶ Can reverse bias 2-3V each

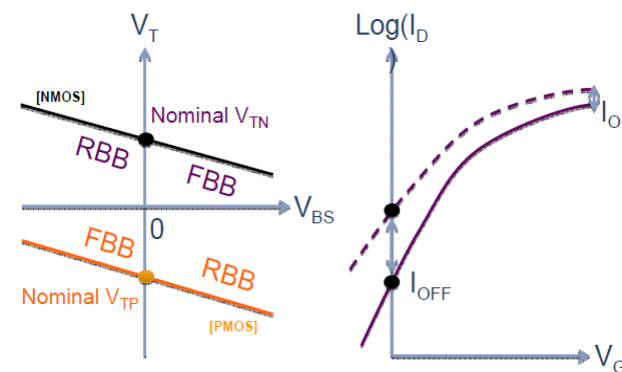
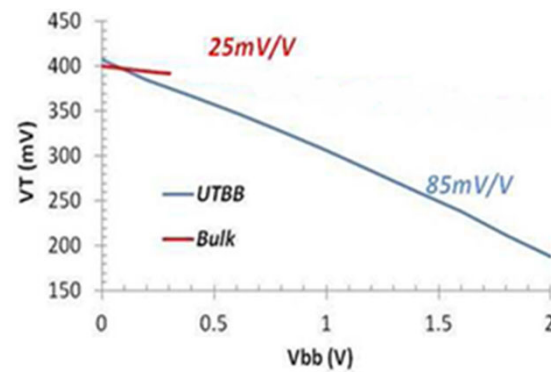
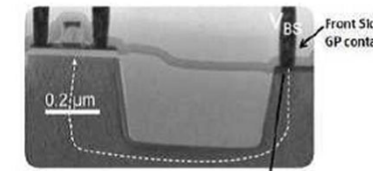
Flip-well (LVT)

- $V_{DDS,nom} = GND_{S,nom} = 0V$
- Forward body bias $V_{BSN} > 0V$
- $0.3V < GND_S < (3V)$
 - Limit due to diodes, BOX
- Can forward bias 2-3V each

Back-Bias in FDSOI

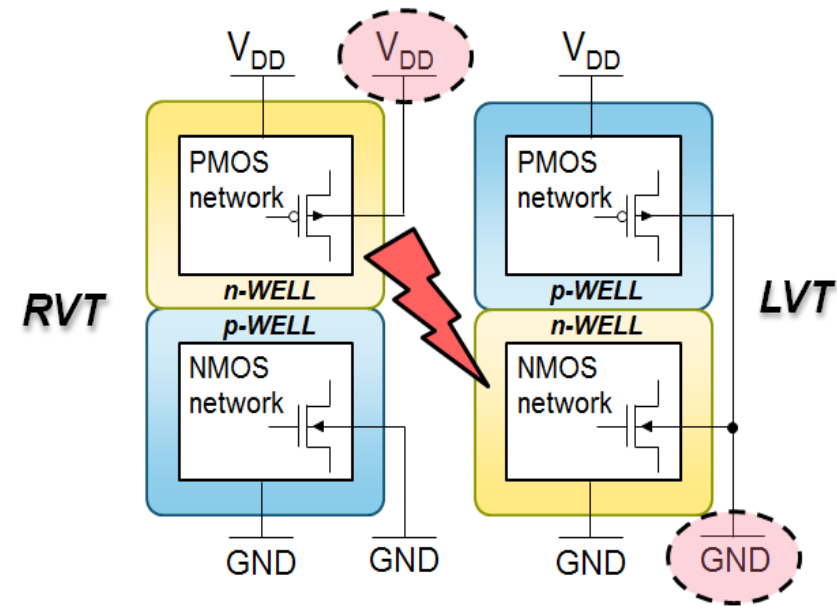


Back-gate contact



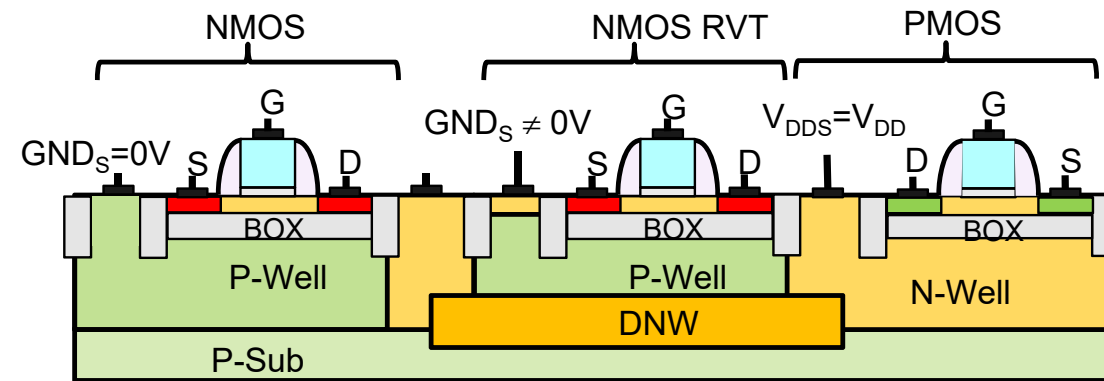
- $\gamma = 85\text{mV/V}$ body coefficient, and extended voltage range
- Lower coefficient and voltage range in bulk, finFET

Multi V_{Th}



- No channel implant in 28FDSOI
 - No multi V_{Th}
- Can't abut wells
 - RVT and LVT require different well biases

Back Bias in FDSOI



- Triple well (deep N-Well, DNW) allows for separate back bias
- Layout penalty; capacitance to drive

Digital Logic: UPF

- Supply, back-bias defined in Universal Power Format (UPF)
 - Or Common Power Format (CPF)
- Handled by synthesis, place and route tools

```
UPF description of PT_TOP with GND, VDD, GNDS and VDDS supplies.
create_power_domain PD_TOP

create_supply_port GND
create_supply_port VDD
create_supply_net GND -domain PD_TOP
connect_supply_net GND -ports {GND}
create_supply_net VDD -domain PD_TOP
connect_supply_net VDD -ports {VDD}

set_domain_supply_net PD_TOP -primary_power_net VDD -primary_ground_net GND

# Body-bias specification
create_supply_port VDDS
create_supply_port GNDS
create_supply_net VDDS -domain PD_TOP
connect_supply_net VDDS -ports {VDDS vddgndvdds*/VDDSCORE}
create_supply_net GNDS -domain PD_TOP
connect_supply_net GNDS -ports {GNDS gnds*/VDDCORE1V8}

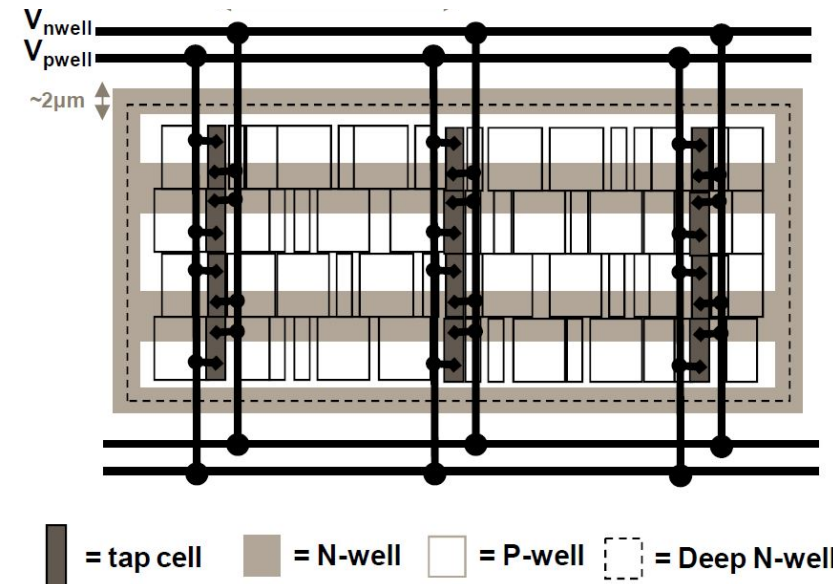
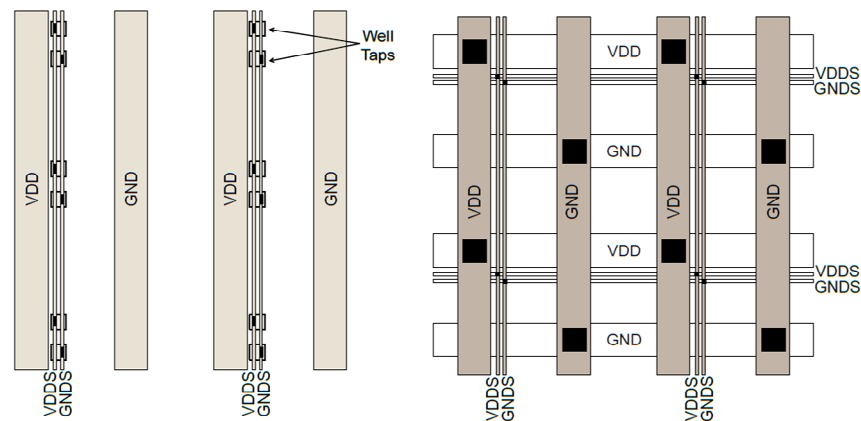
create_supply_set back_bias_set \
-function {nwell VDDS} \
-function {pwell GNDS} \
-reference_gnd {GND} \

create_power_domain PD_TOP -update -supply bias
associate_supply_set back_bias_set -handle PD_TOP.bias
```

M.Blagojevic, Ph.D. Dissertation, ISEP 2017

Digital Logic - Implementation

- Well taps added explicitly
 - Difference from bulk

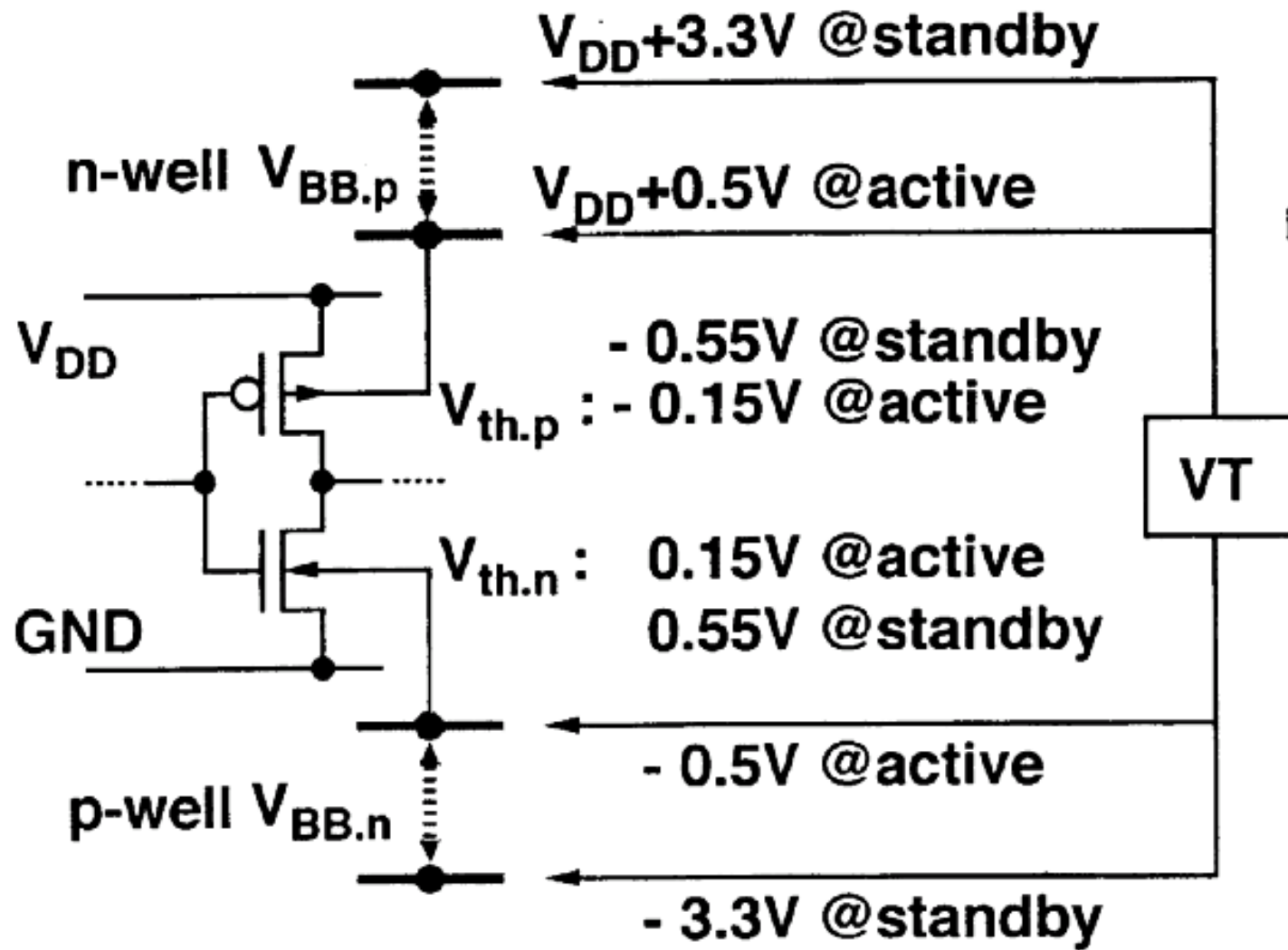


➤ Back bias straps

- Low DC current
- Except for very fast transitions

Dynamic Body Bias (Bulk)

ISSCC'96 pp.166-167



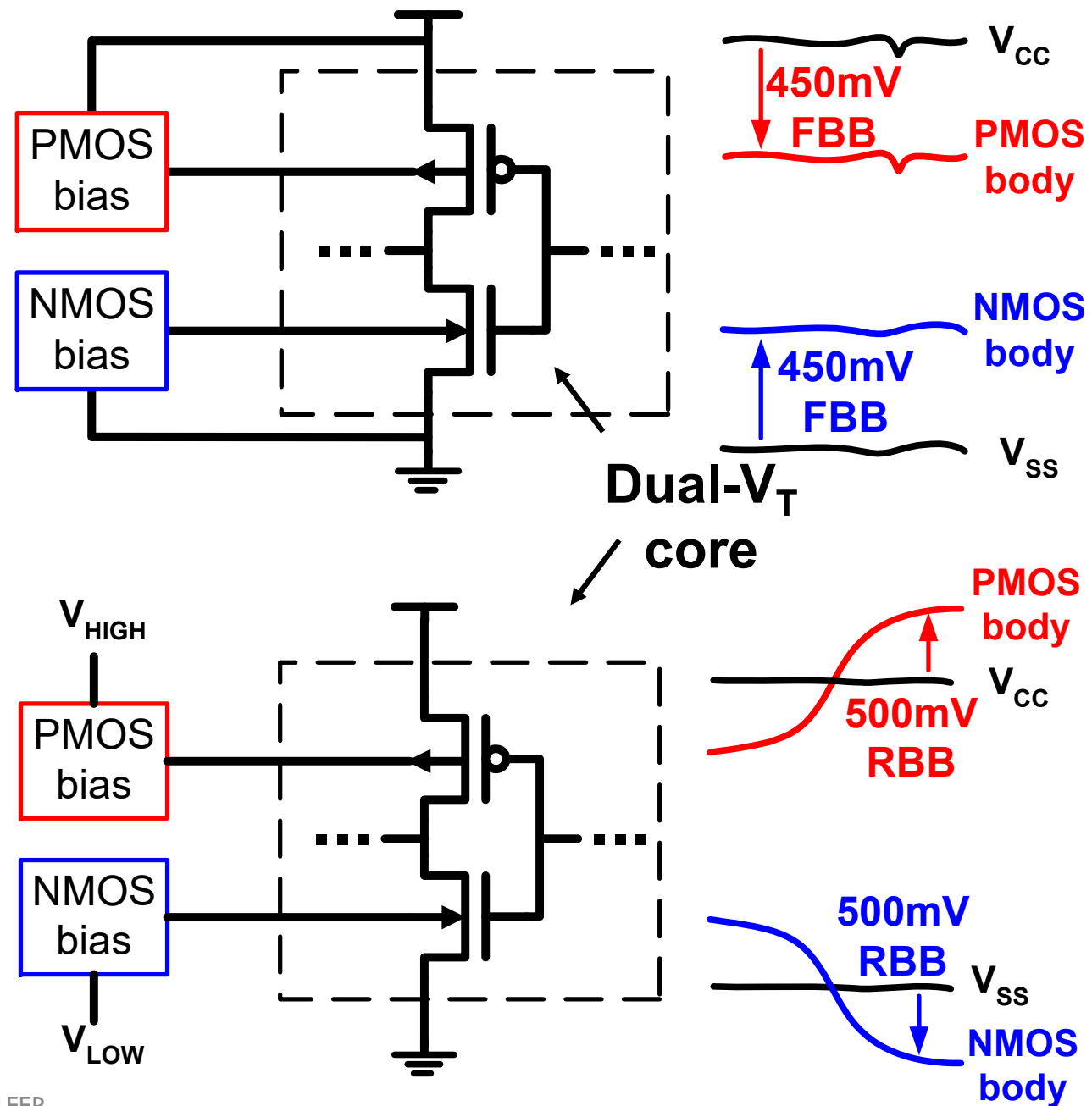
VTCMOS :

Dynamic V_{th} control for low power through backgate bias

example:

(SATS) or (SPR) or (SATS + SPR)

Dynamic Body Bias (Bulk)



Active mode

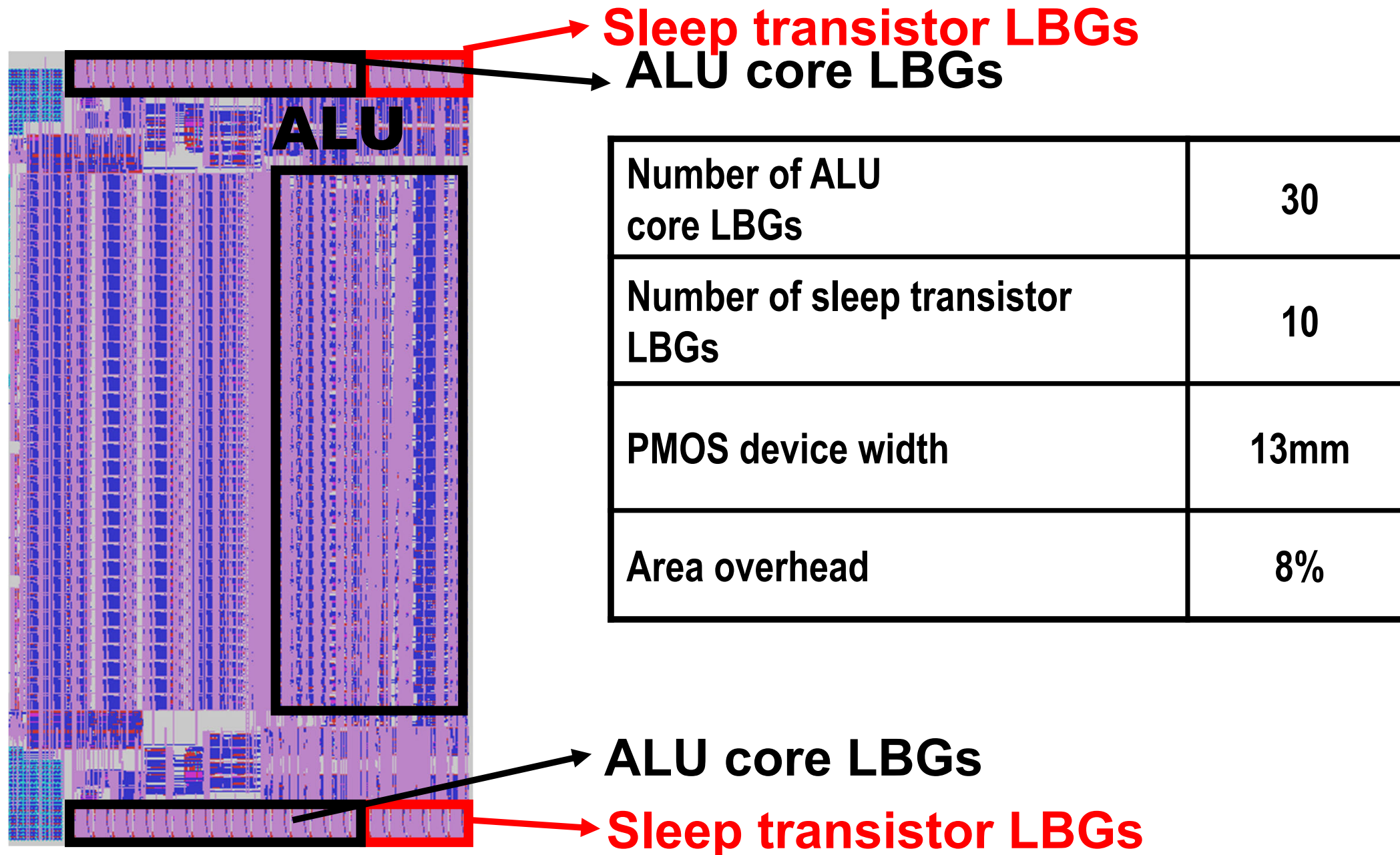
Forward body bias (FBB)
Local V_{CC} tracking

Idle mode

Reverse body bias (RBB)
Triple well needed

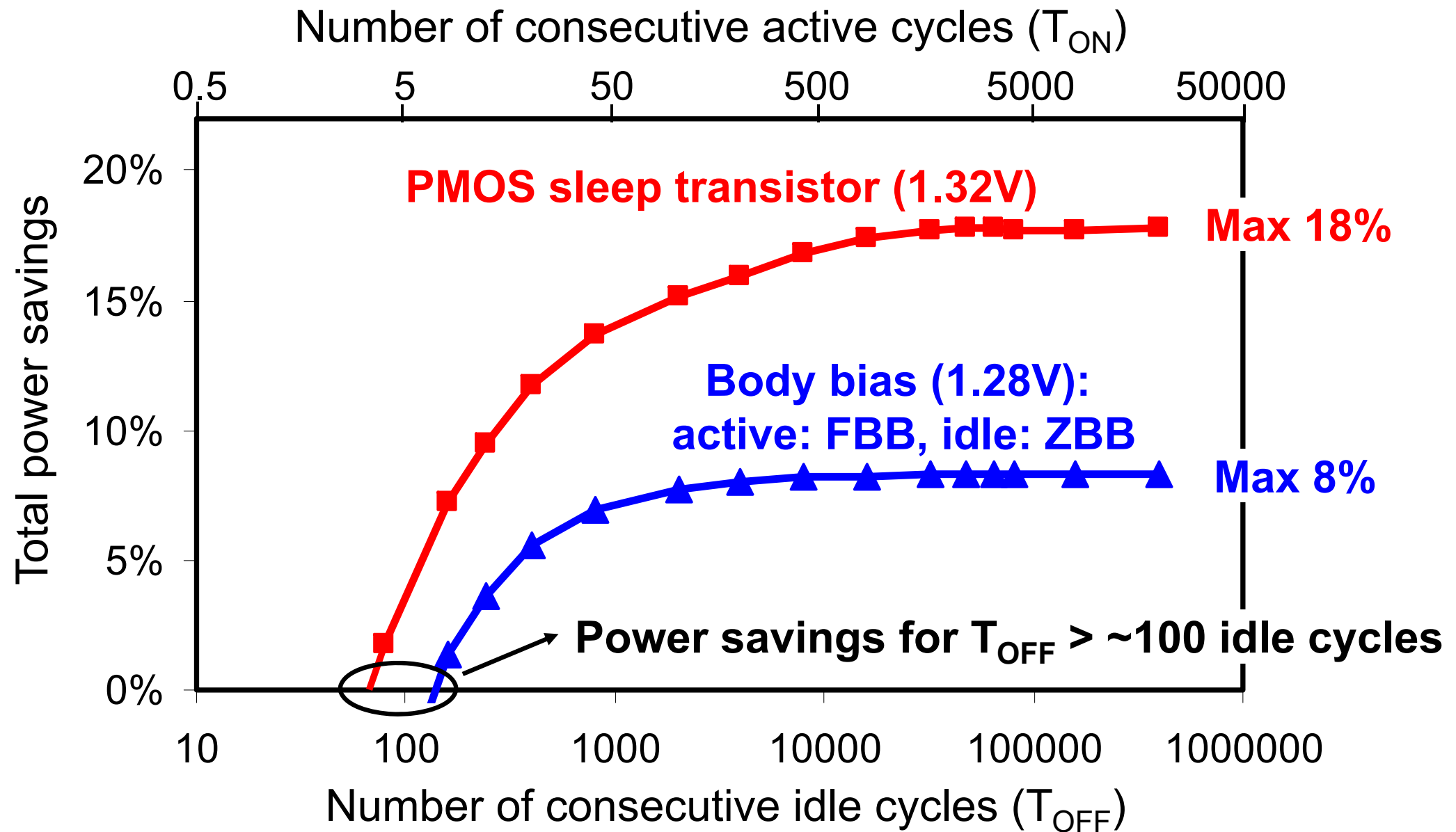
Tschanz, ISSCC'03

Body Bias Layout



Total Active Power Savings

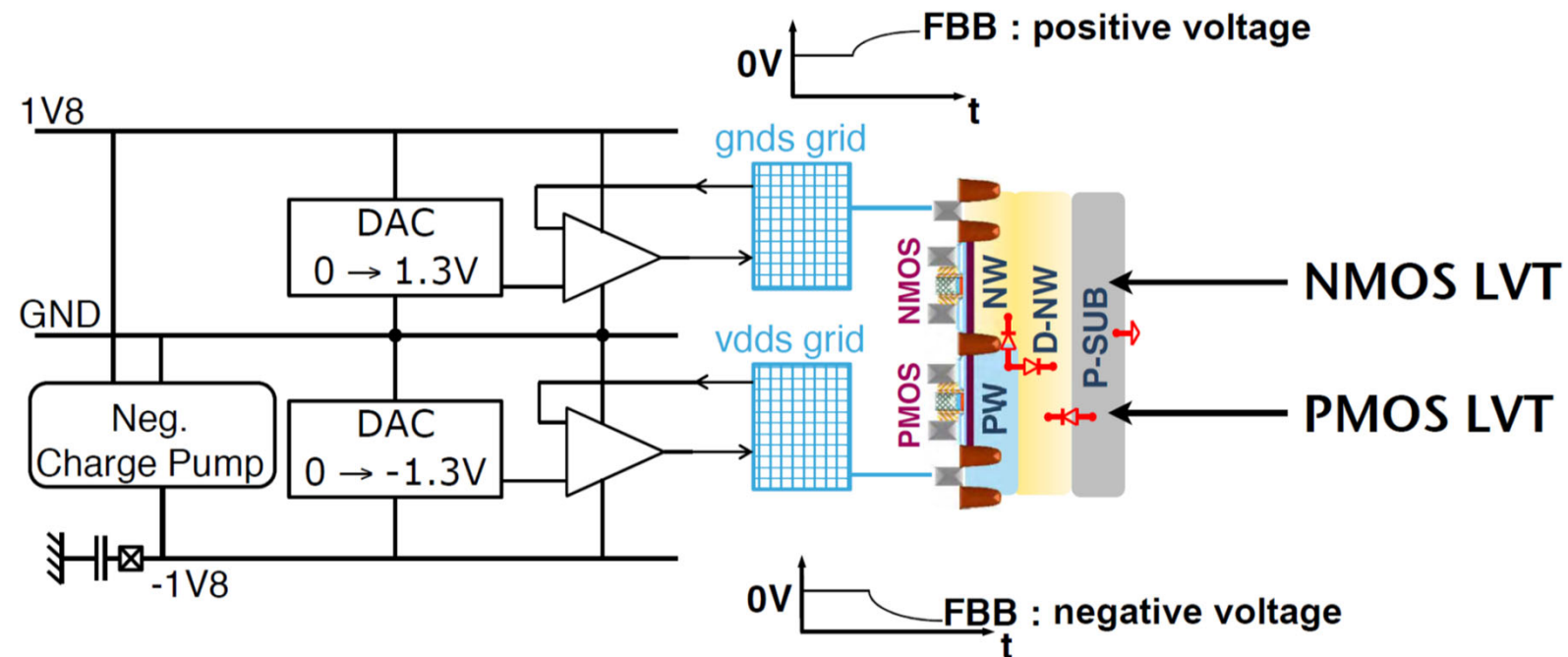
(Fixed activity: $\alpha = 0.05$)



Reference: 450mV FBB to core with clock gating, 1.28V, 4.05GHz, 75°C

Generating Back-Bias

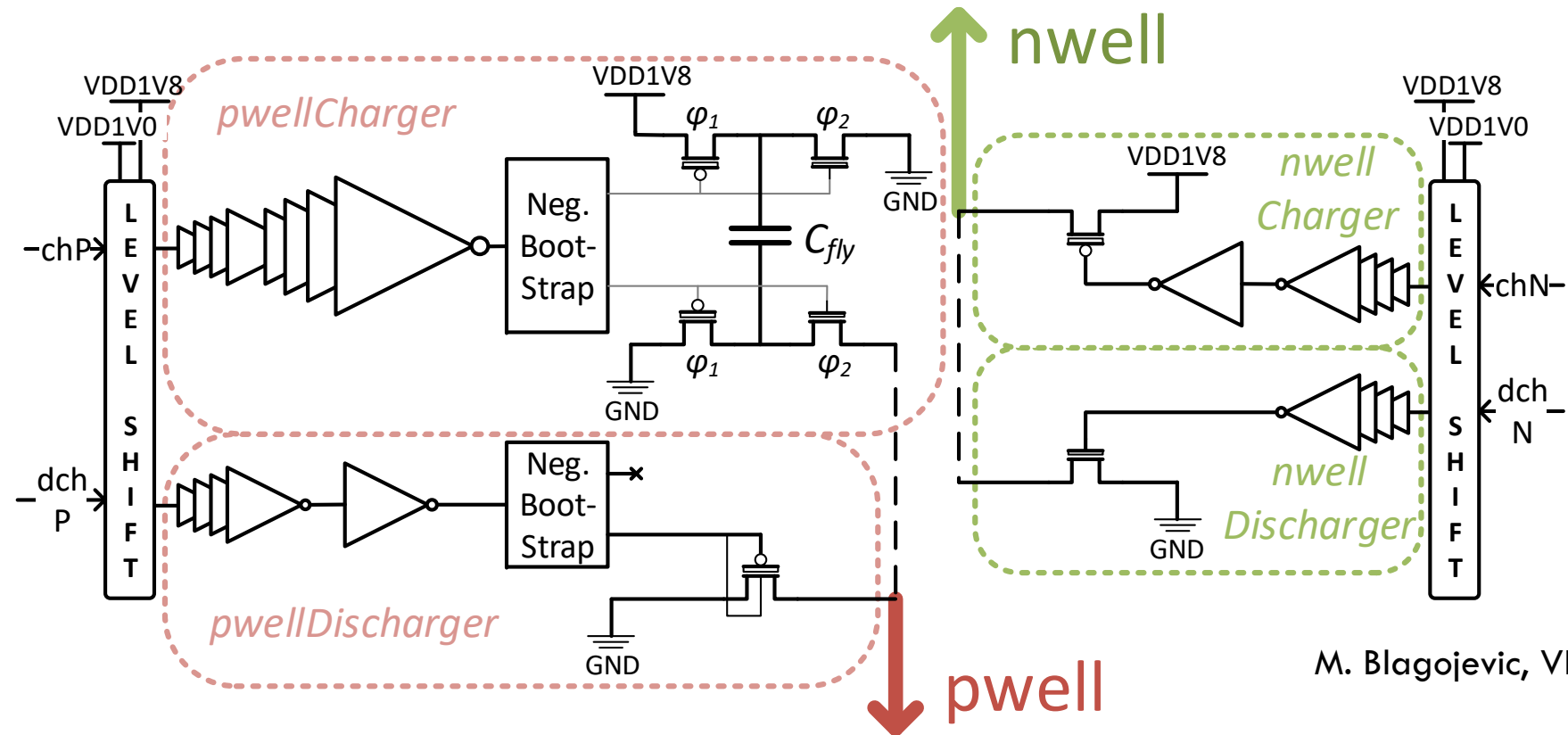
- Tradeoff – speed of charging and discharging well caps
- Often measure V_{BB} indirectly (leakage)
- Challenge: Generating $-V_{SS}$
- 28nm FDSOI implementation



D. Jacquet, VLSI 2013

Generating Back Bias

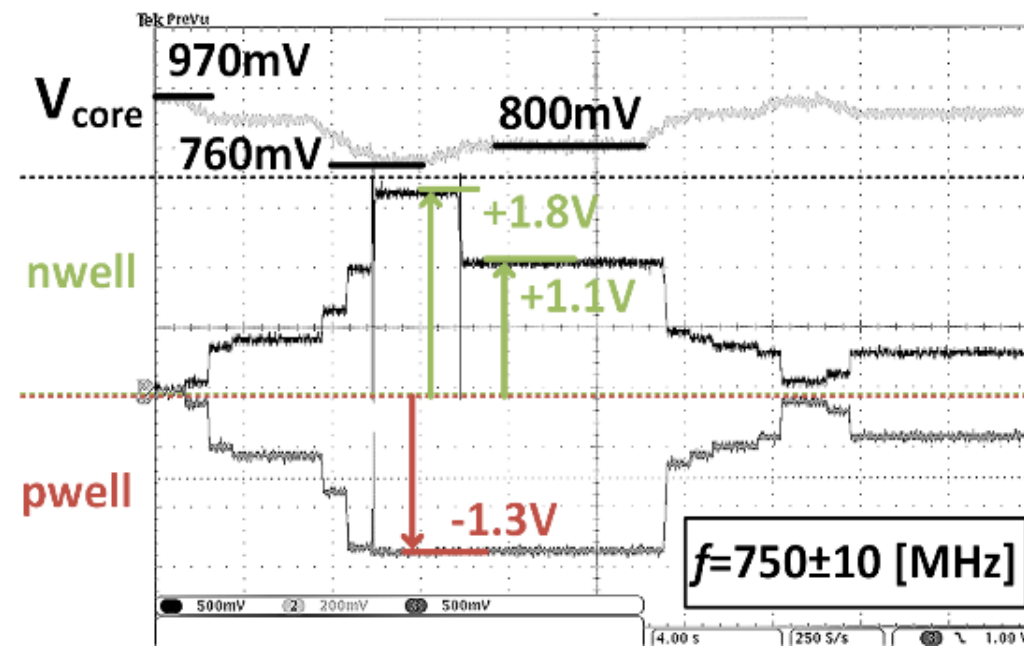
- Fast and wide voltage range back-bias in FDSOI



M. Blagojevic, VLSI 2016

Switched capacitors generate negative bias and pump substrate

Supply/Process Compensation



- Able to track $\sim 200\text{mV}$ supply droops and maintain constant frequency (measured by a replica) by back-bias adjustments



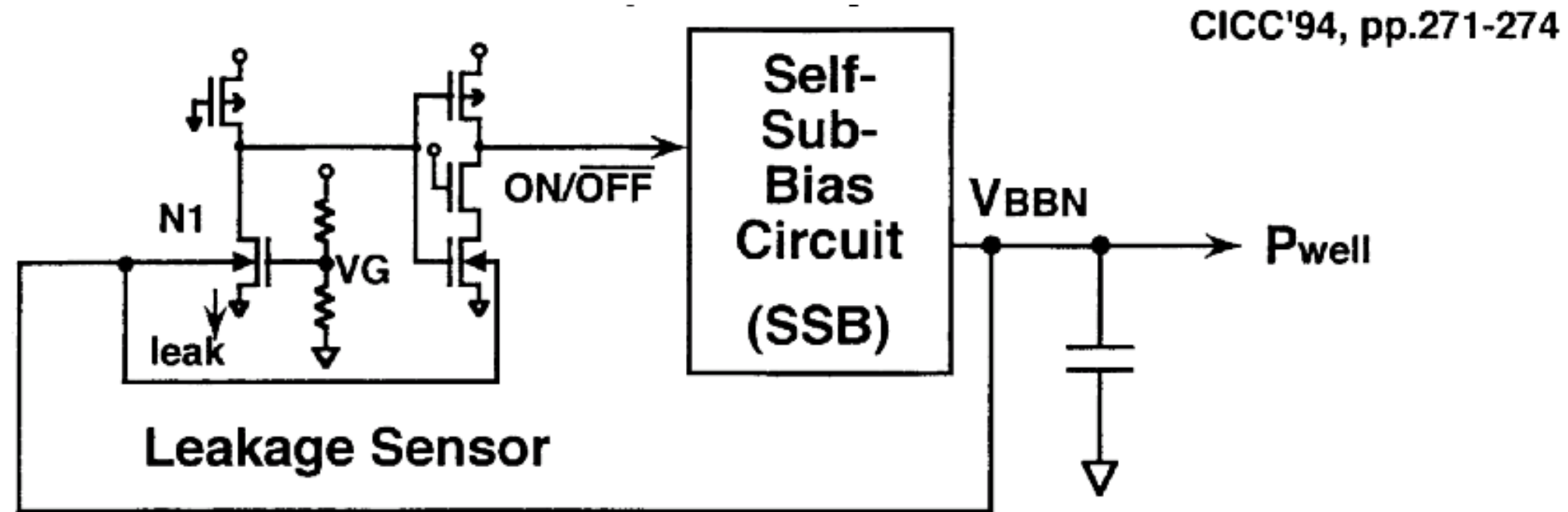
5.N Dynamic Threshold Scaling and Variations

Body Biasing and Variations

- Body biasing with a local control loop can be used to lower the impact of process variations
- Used to limit die-to-die and within-die variations

Self-Adjusting Threshold-Voltage Scheme (SATS)

- Older bulk technologies had stronger body effect



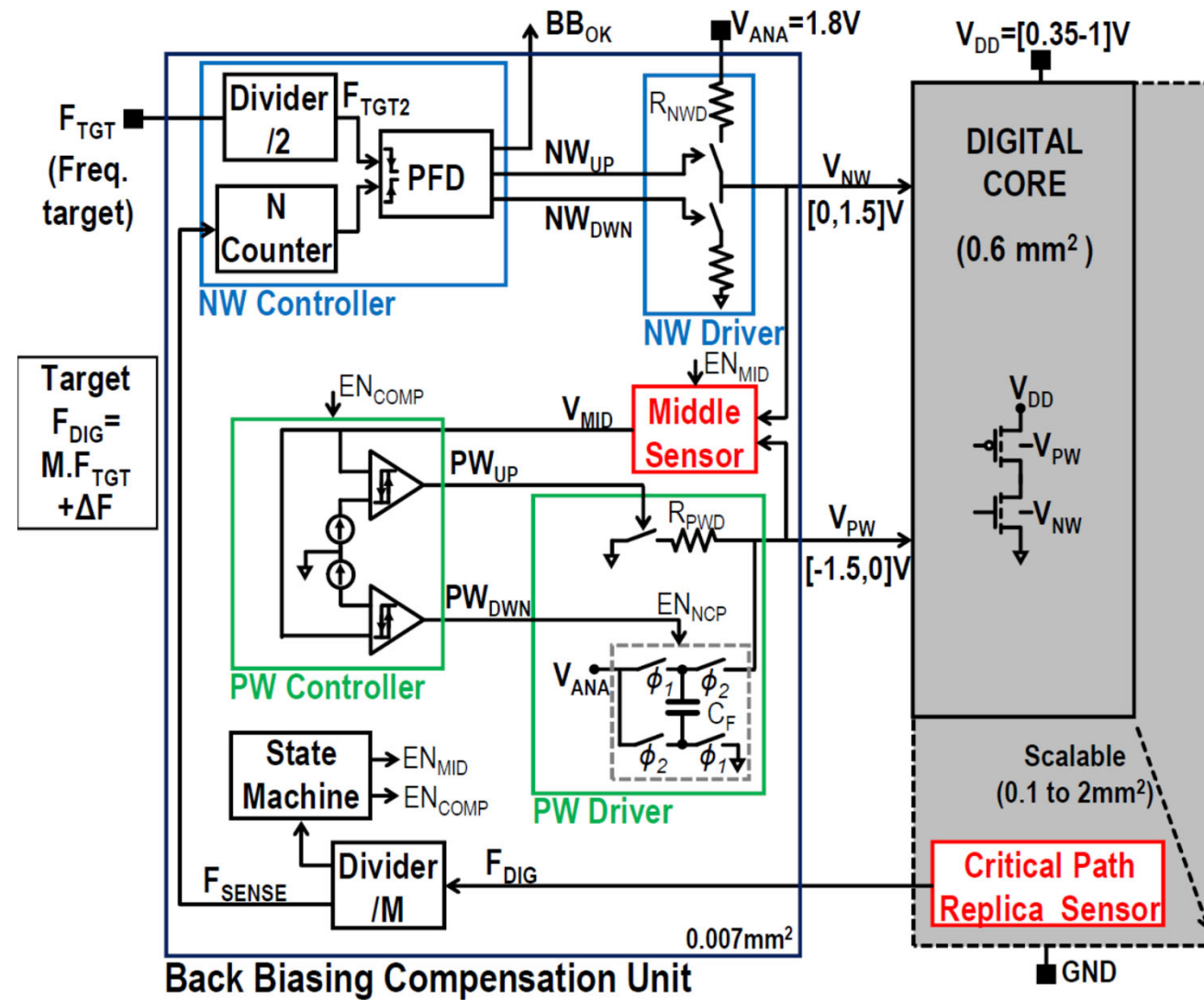
low V_{th} → large leakage → SSB ON → deep V_{BB} → high V_{th}

high V_{th} → little leakage → SSB OFF → shallow V_{BB} → low V_{th}



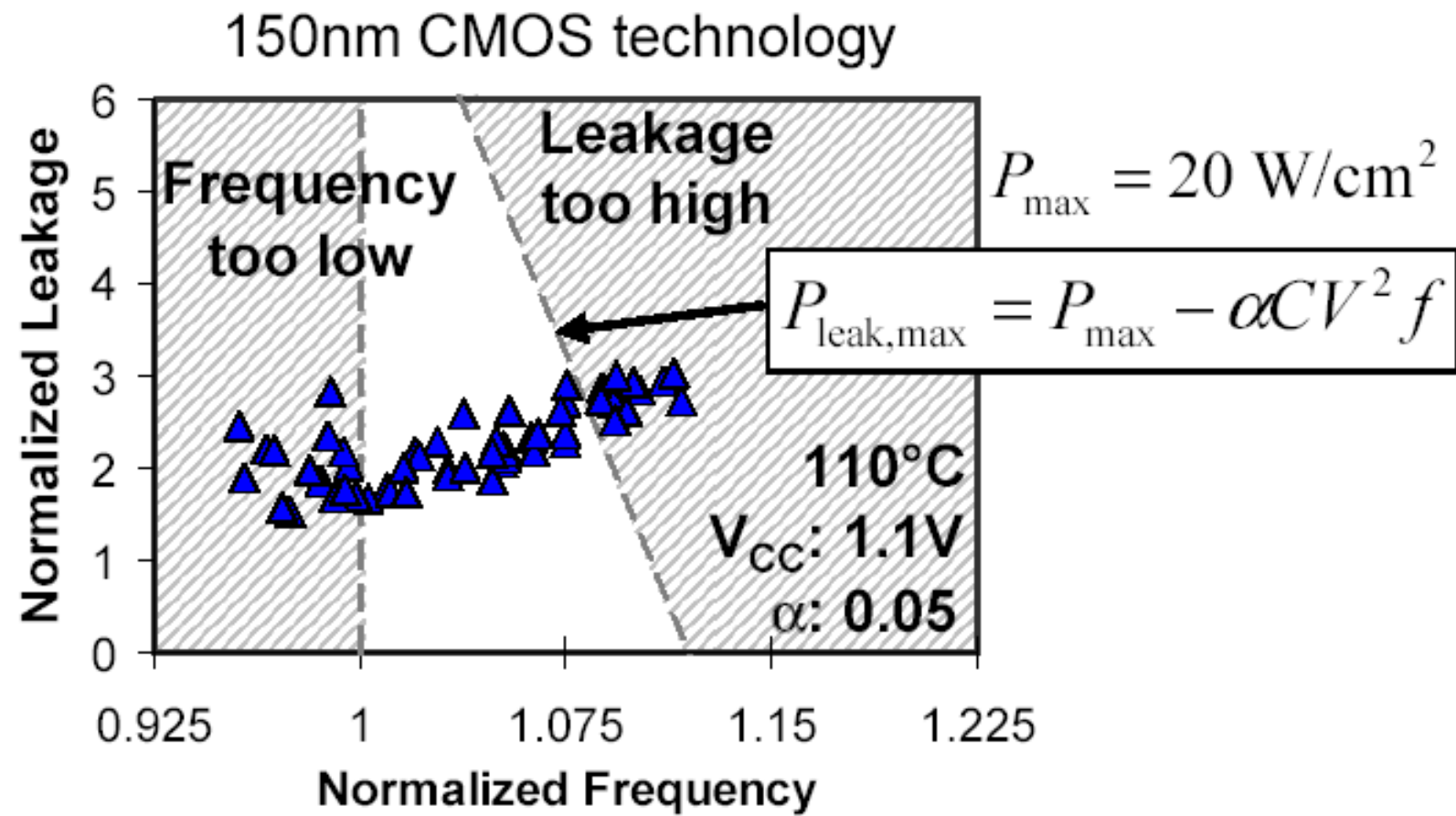
- control V_{th} to adjust leakage current
- compensate V_{th} fluctuation

Dynamic Frequency Loop in FDSOI



Quelen, ISSCC'18

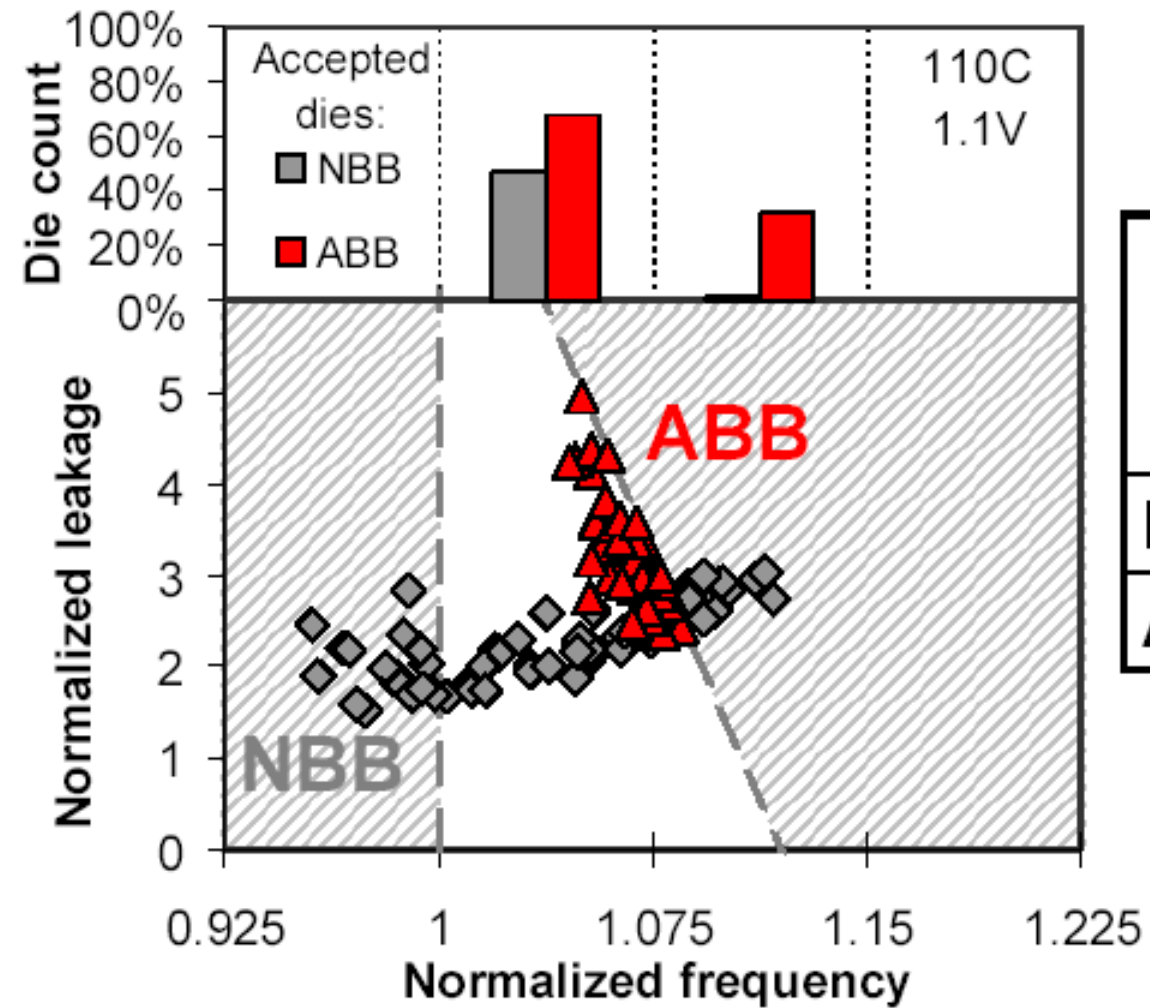
Substrate Biasing



Tschanz, JSSC 11/02

Effectiveness of Substrate Bias

Die-to-die variations

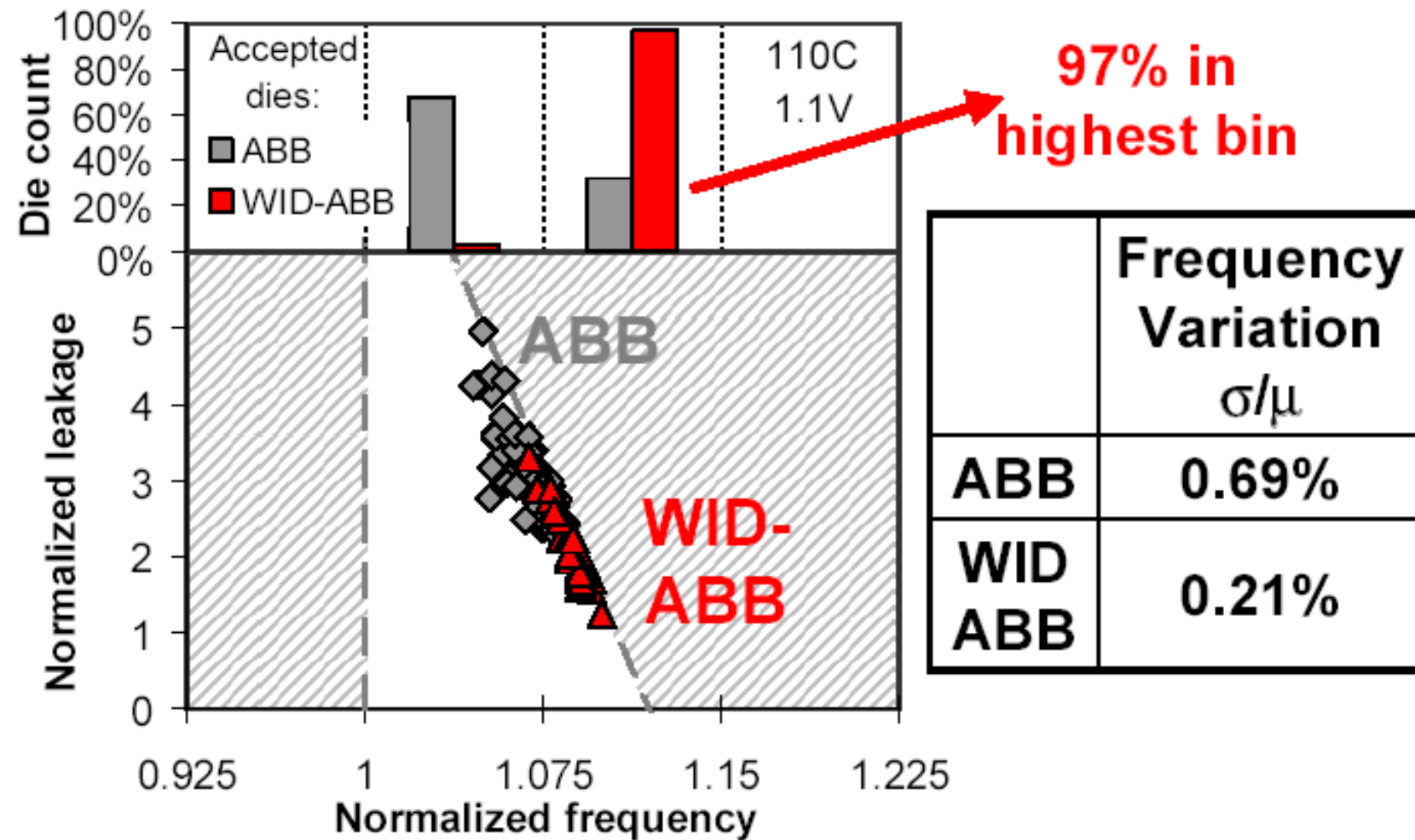


	Frequency Variation σ/μ
NBB	4.1%
ABB	0.69%

- NBB: No body bias
- ABB: Adaptive body bias

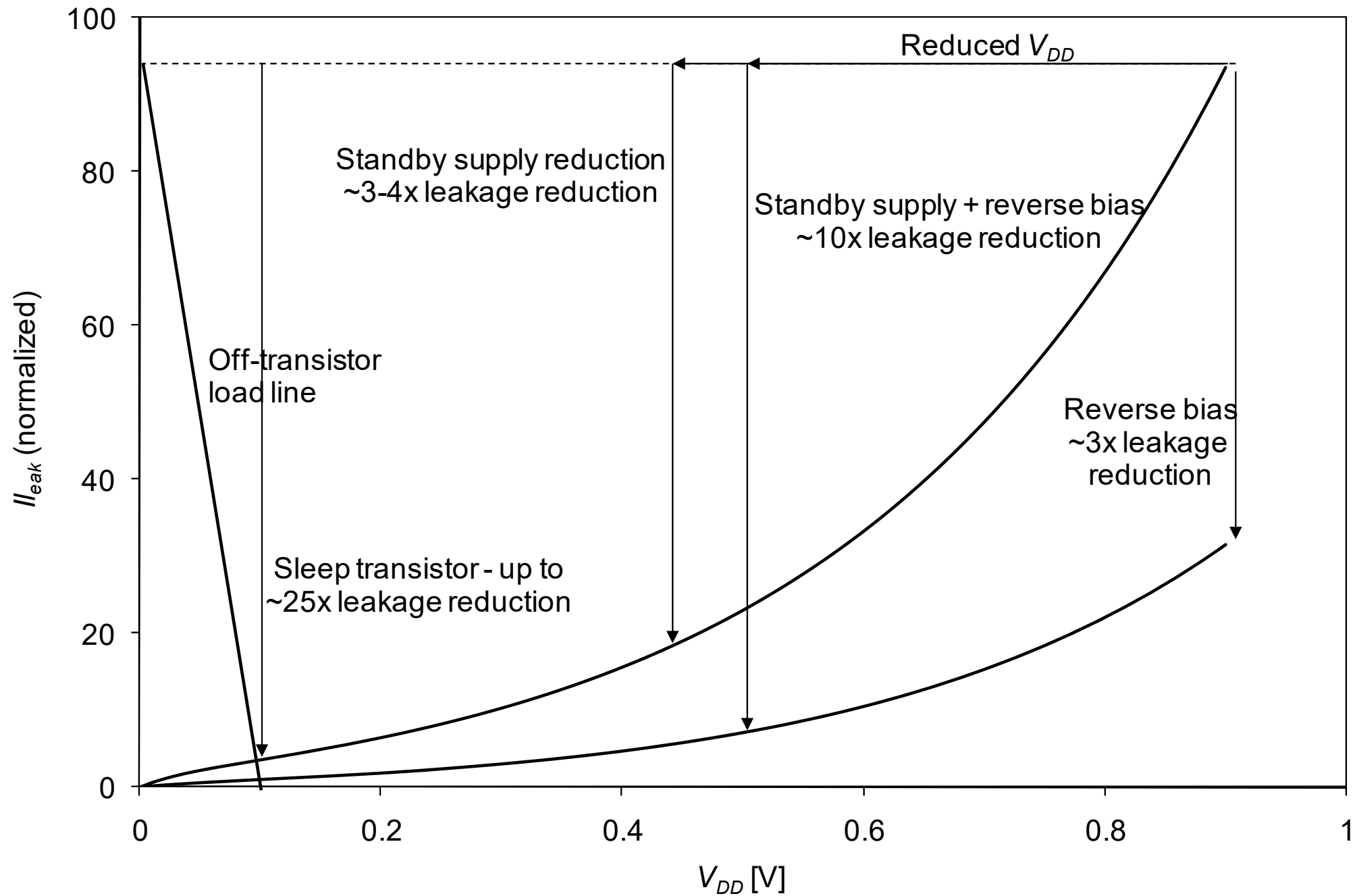
Effectiveness of Substrate Bias

Within-die variations



- ABB with multiple within die (WID) sensors

Techniques Summary (around 130nm node)



Power /Energy Optimization Space

	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Sleep Mode	Run Time
Active	Logic design Scaled V_{DD} Trans. sizing Multi- V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi- V_{Th}	Sleep T's Multi- V_{DD} Variable V_{Th} + Input control	+ Variable V_{Th}

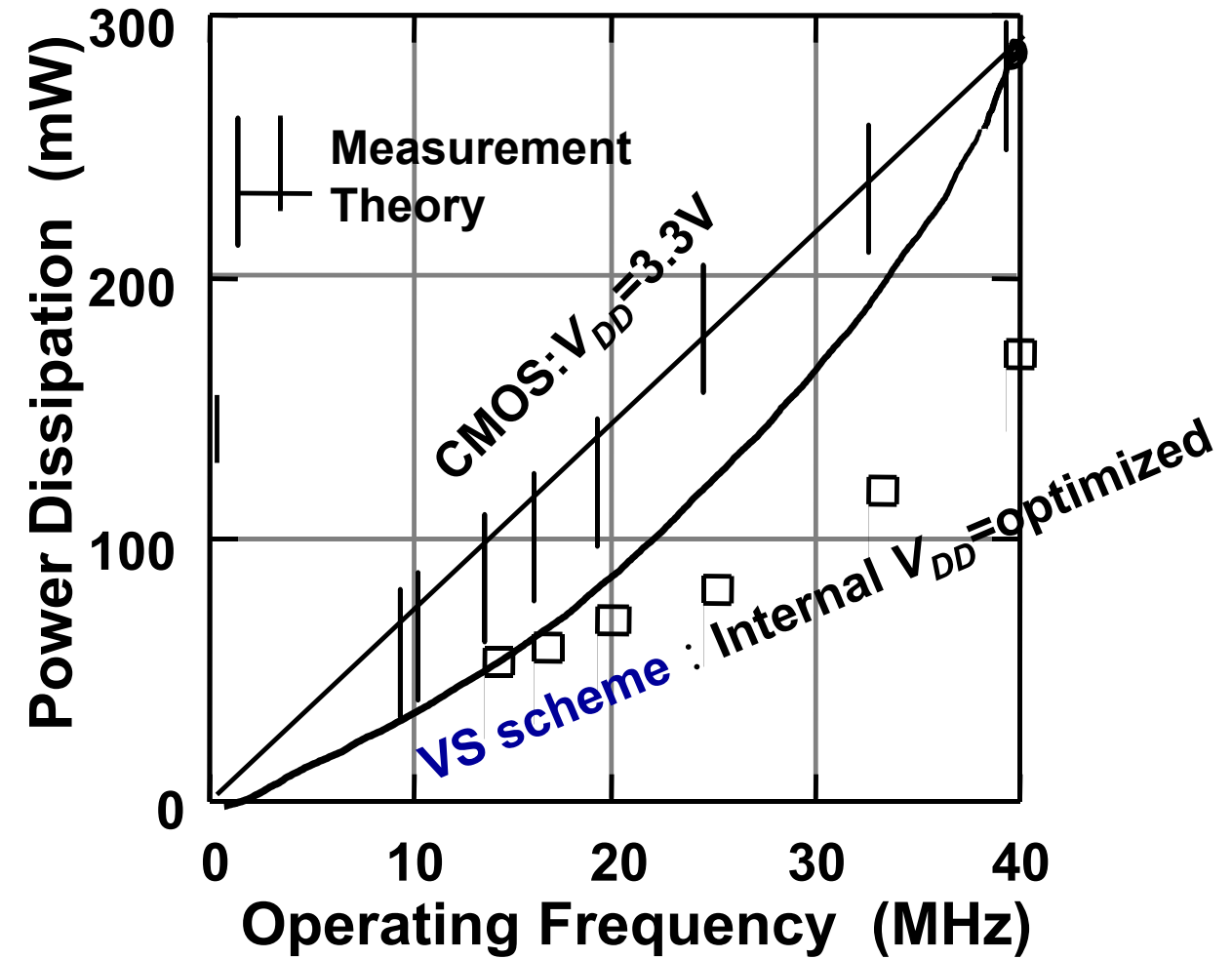
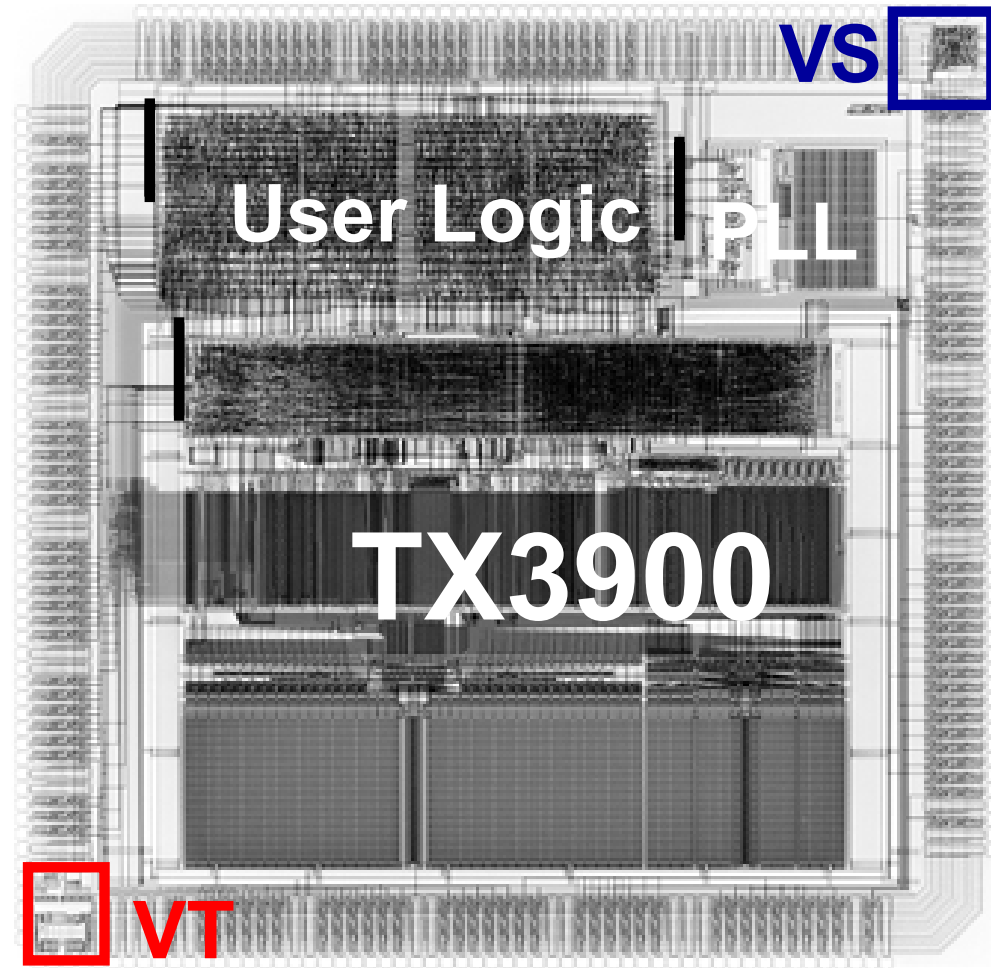


5.0 Optimal V_{DD} , V_{Th}

Dynamic Voltage Scaled Microprocessor

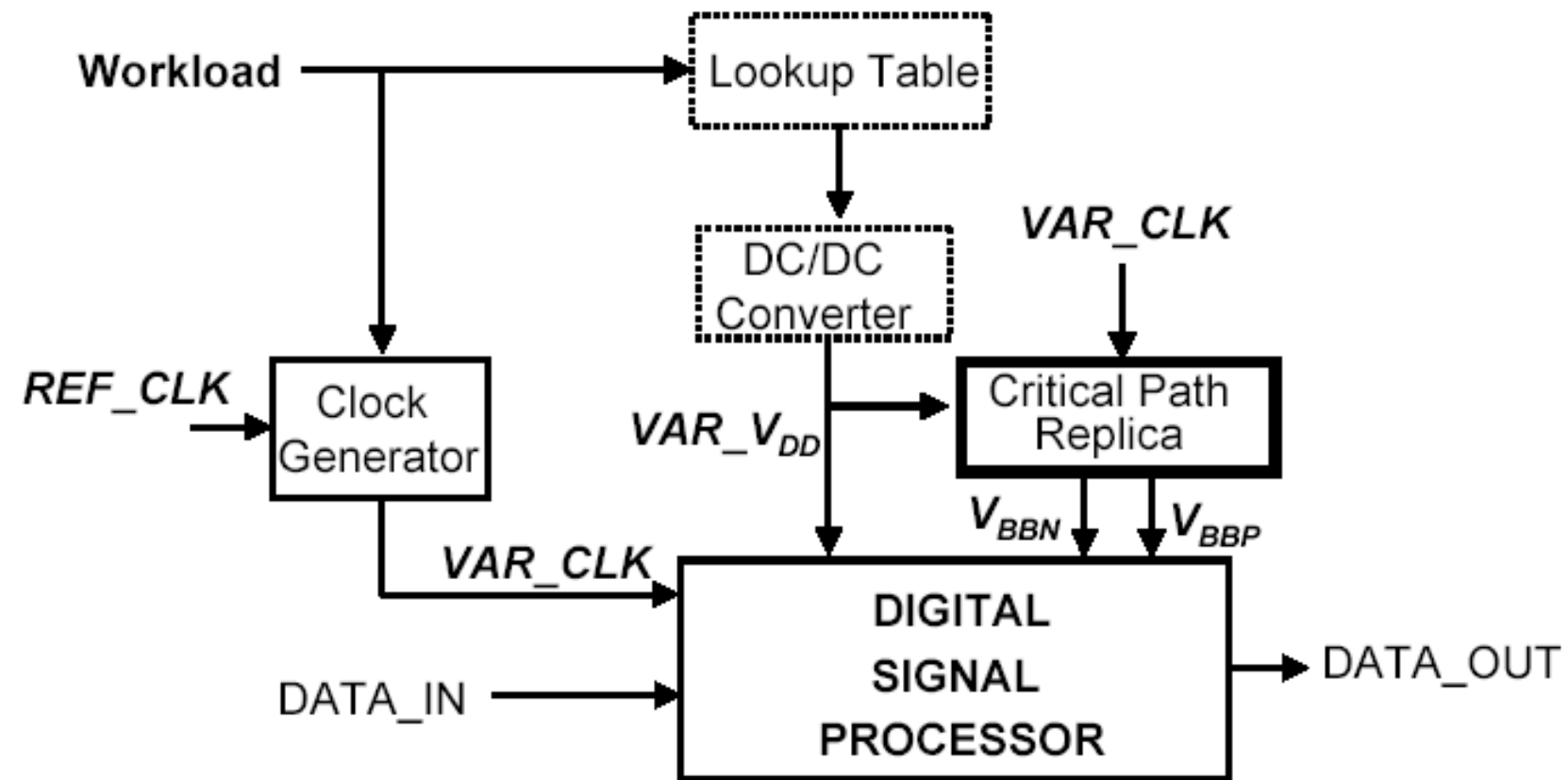
External V_{DD} $3.3V \pm 10\%$

Internal V_{DDL} $0.8V \sim 2.9V \pm 5\%$



Courtesy: Prof. Kuroda

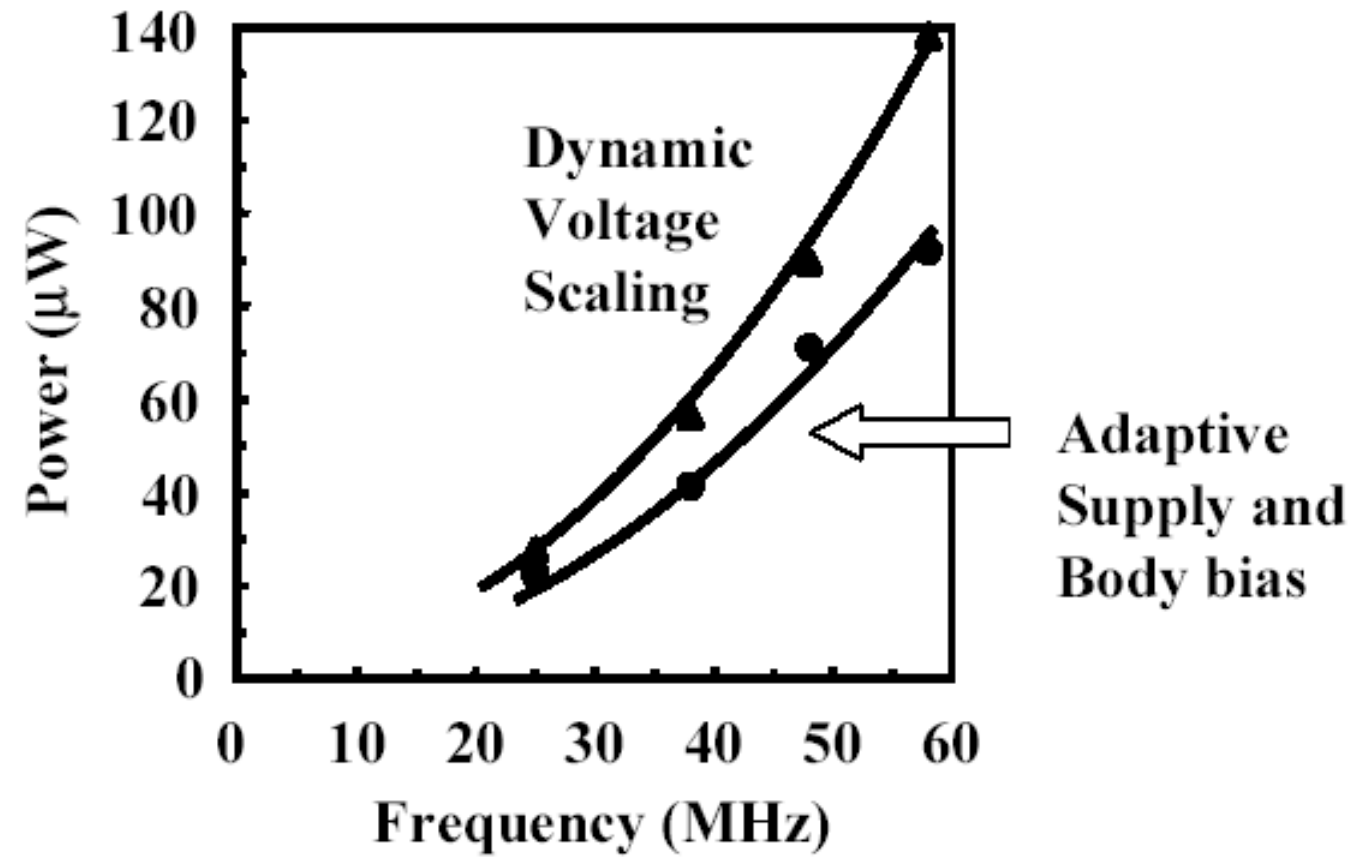
Adapting V_{DD} and V_{TH}



- Adapting both V_{DD} and V_{Th} during runtime
 - V_{Th} is much less sensitive

Miyazaki, ISSCC'02

Adapting V_{DD} and V_{TH}



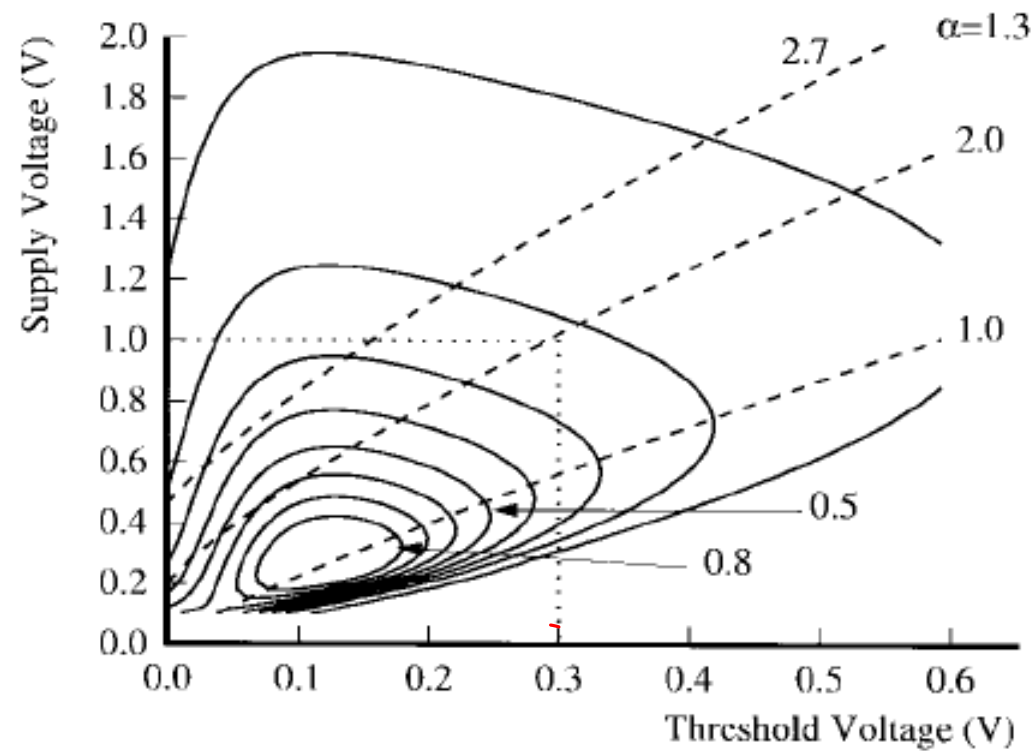
Miyazaki, ISSCC'02

Optimal V_{DD} , V_{Th}

- Adjusting V_{DD} , V_{Th} trades of energy and delay
- We studied energy-limited design
 - And alternate ways for optimizing energy and delay together
 - E.g. energy-delay product (EDP)
 - Or $E^n D^m$, $n, m > 1$

Optimal EDP Contours

- Plot of EDP curves in V_{DD} , V_{Th} plane



Gonzalez, JSSC 8/97

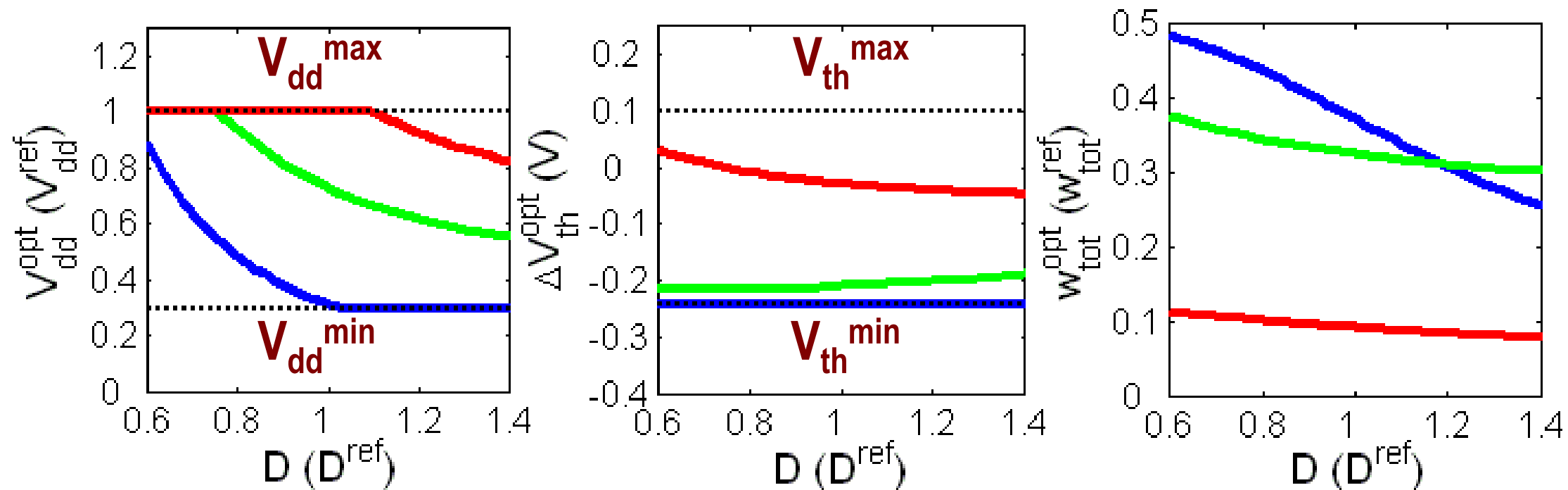
Sizing, Supply, Threshold Optimization

Reference Design:

$D^{\text{ref}} (V_{\text{dd}}^{\text{max}}, V_{\text{th}}^{\text{ref}})$

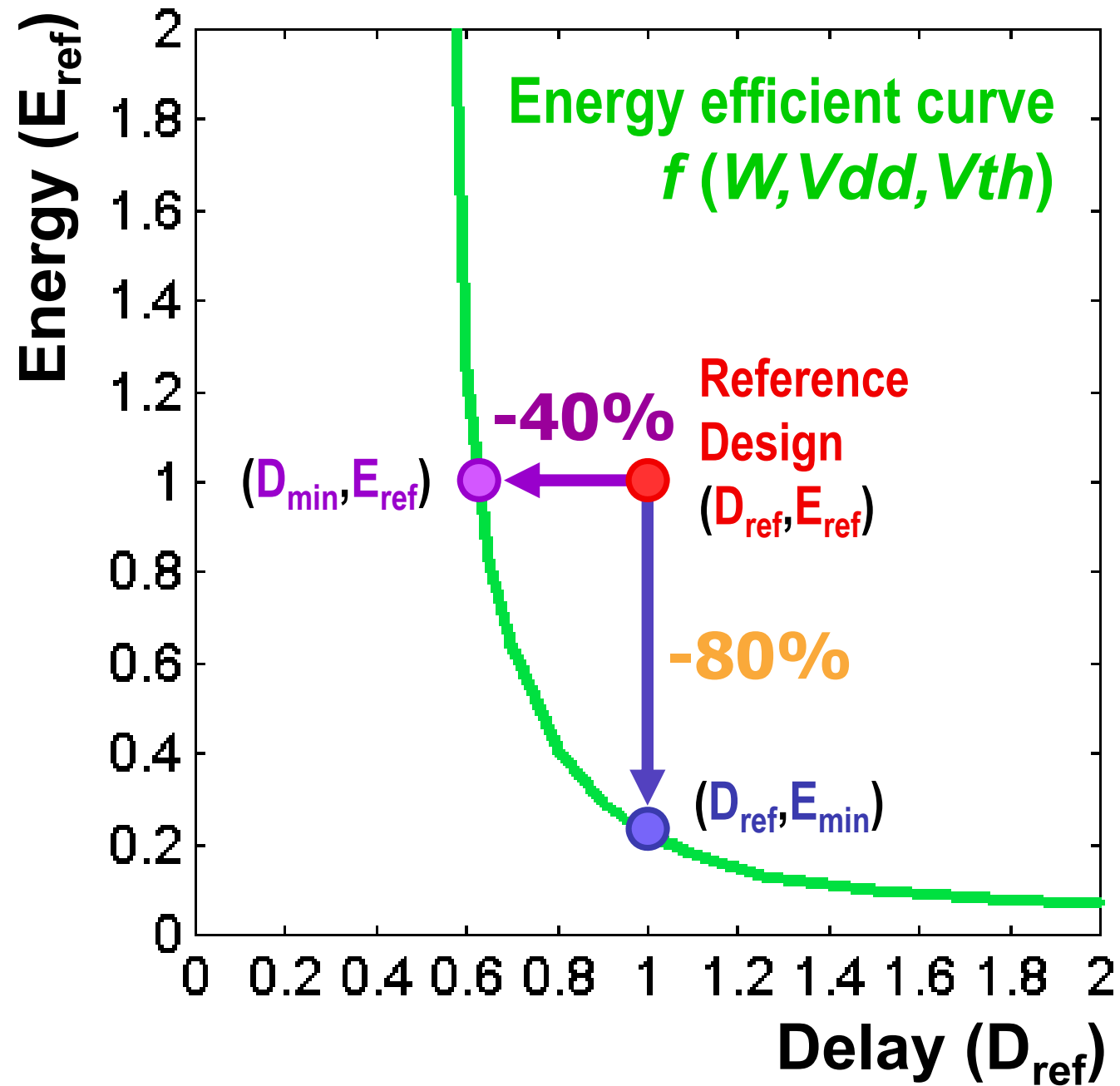
Topology	Inverter	Adder	Decoder
$(E_{Lk}/E_{Sw})^{\text{ref}}$	0.1%	1%	10%

Large variation in optimal circuit parameters $V_{\text{dd}}^{\text{opt}}, V_{\text{th}}^{\text{opt}}, w^{\text{opt}}$



Technology parameters ($V_{\text{dd}}^{\text{max}}, V_{\text{th}}^{\text{ref}}$) rarely optimal

Result: E-D Tradeoff in an Adder



Sensitivity	W	Vdd	Vth
(D_{ref}, E_{ref})	∞	1.5	0.2
(D_{ref}, E_{min})	1		
(D_{min}, E_{ref})	22	16	22

80% of energy saved without delay penalty

40% delay improvement without energy penalty

Energy-constrained delay

- Active power

$$P_{act} = \alpha f C V_{DD}^2$$

$$f = 1 / L_D t_p$$

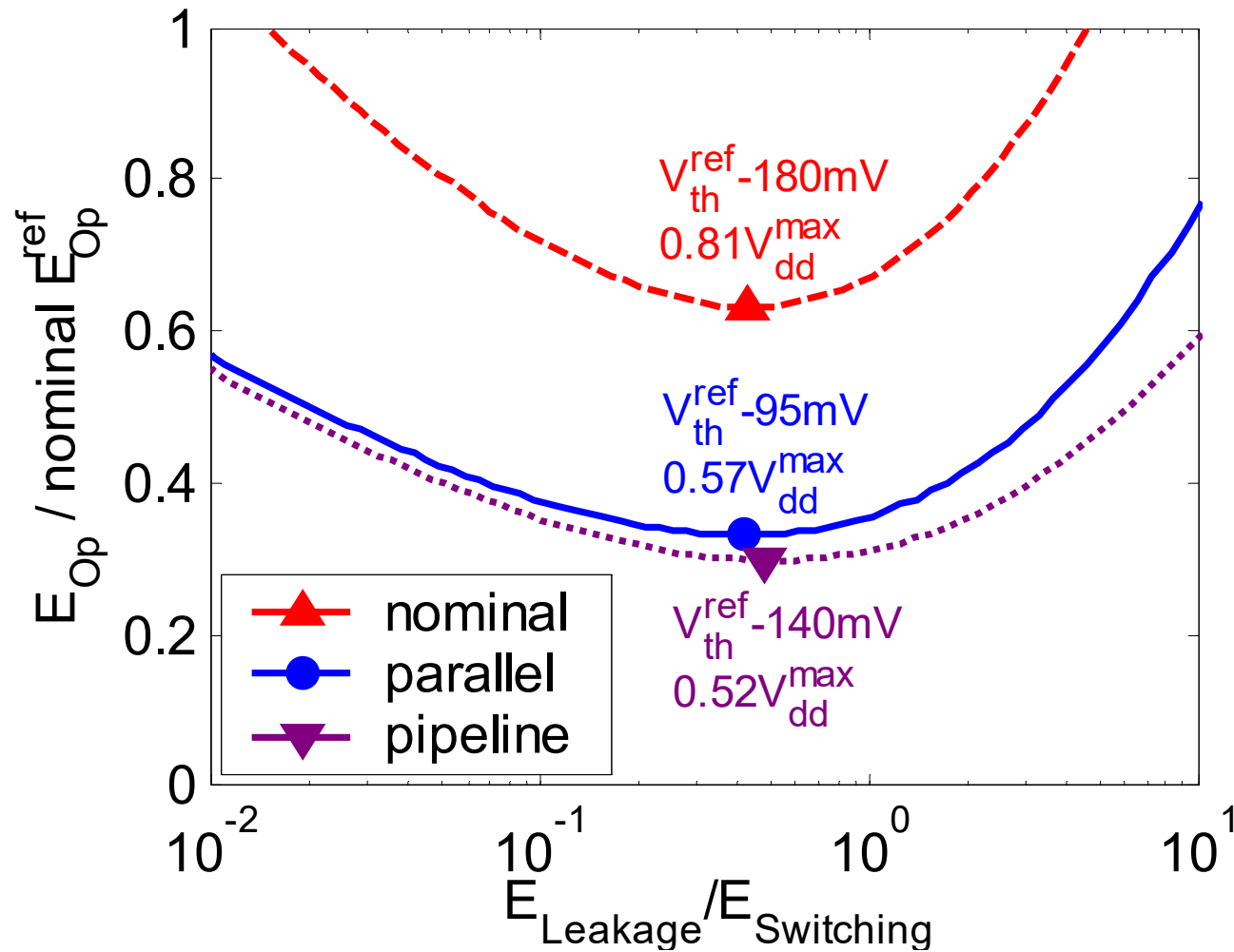
- Leakage power

$$P_{leak} = I_0 e^{\frac{-V_{Th} - \gamma V_{DD}}{S}} V_{DD}$$

- Eliminate one variable (V_{Th}) and find $P_{min}(V_{DD})$

Nose, ASP-DAC'00

Minimum energy: $E_{Sw} = 2E_{Lk}$

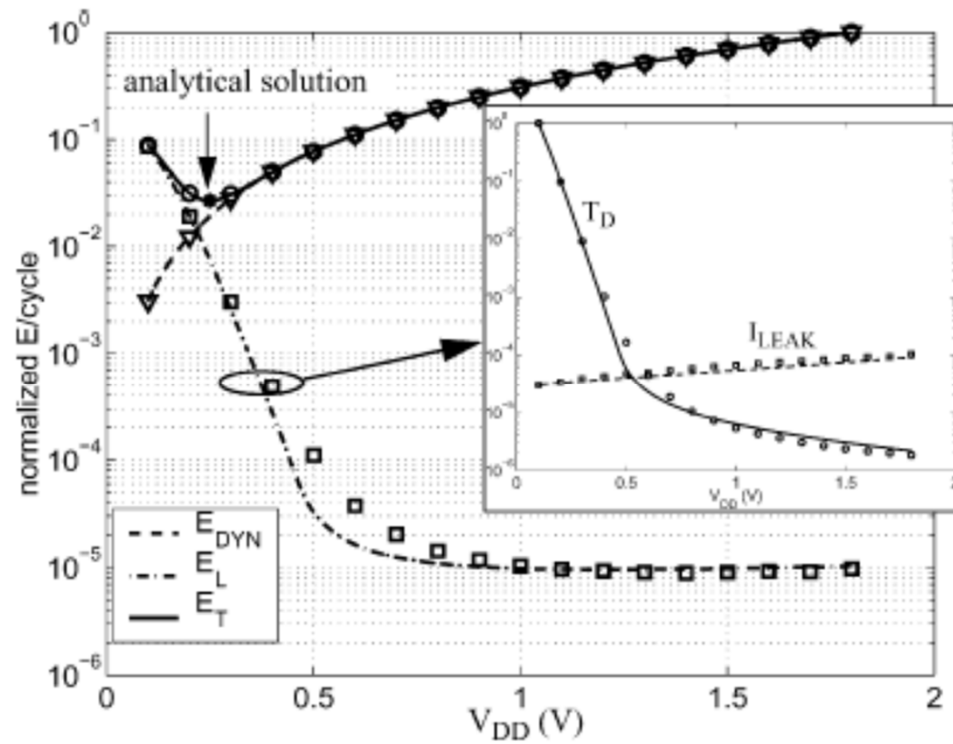


- ◆ Large $(E_{Lk}/E_{Sw})^{opt}$
- ◆ Flat E_{Op} minimum
- ◆ Topology dependent

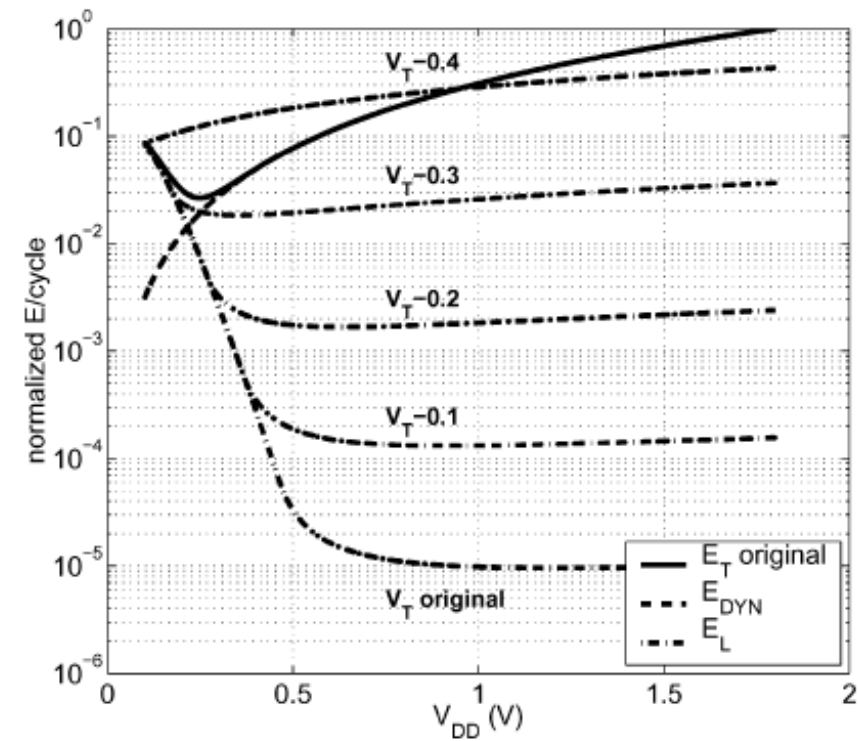
$$\left(\frac{E_{Lk}}{E_{Sw}}\right)_{opt} = \frac{2}{\ln\left(\frac{L_d}{\alpha_{avg}}\right) - K}$$

Optimal designs have high leakage ($E_{Lk}/E_{Sw} \approx 0.5$)

Subthreshold Optimum



$f = 30\text{kHz}$



Minimum is independent of V_T

Calhoun, JSSC 9/05

Next Lecture

- We finished low-power design
- Next is clocks and supplies