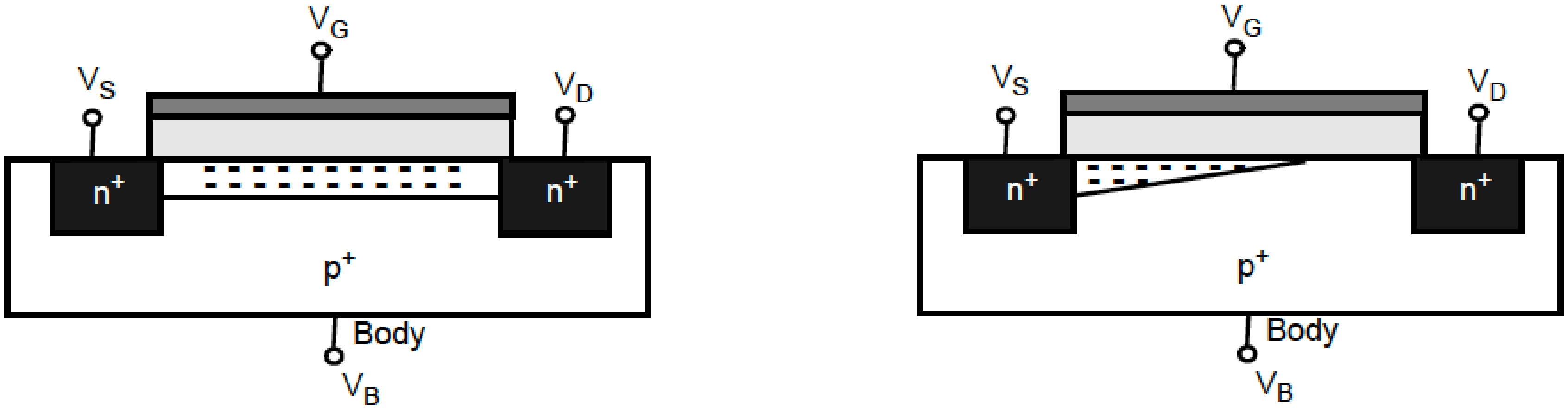
CHAPTER 2

IMPACT OF TECHNOLOGY

- **TRANSISTOR BASICS**
- **POWER ISSUES**
  - DYNAMIC
  - STATIC
- **RELIABILITY**
  - ACE
  - NBTI
  - EM
  - TDDB

nMOS TRANSISTOR OPERATION



(a) $V_{GS} > V_{th}$ and $V_{DS}=0$

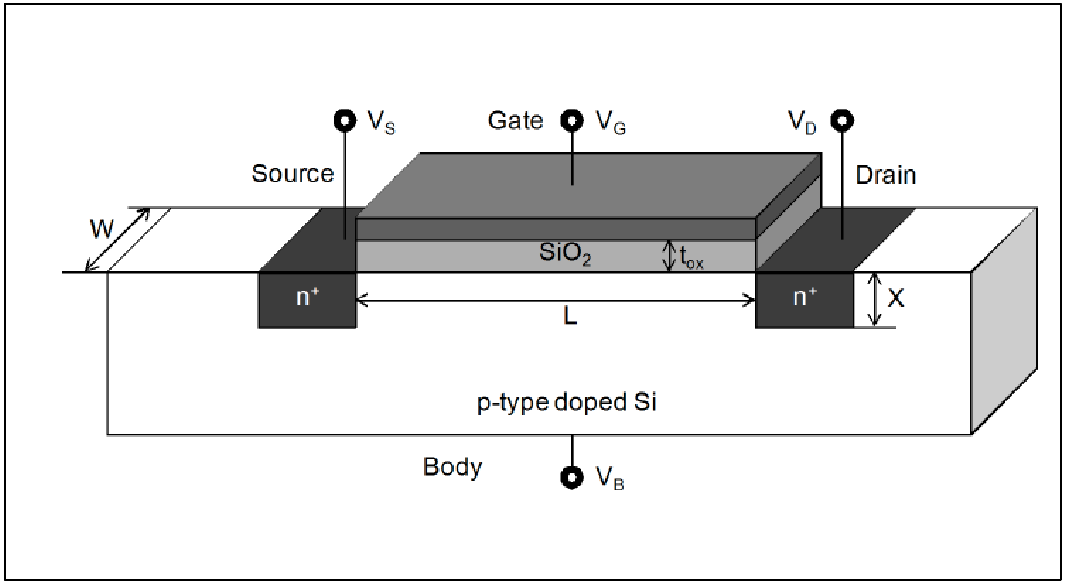
(b) $V_{GS} > V_{th}$ and $V_{DS} > V_{th}$

- Gate voltage controls current by changing the thickness of conduction channel
  - $V_{GS} < 0$ then holes populate between source and drain
  - $V_{GS} > V_{th}$ minority carriers (electrons) are attracted to the gate forming a conduction channel
  - $V_{GS} > V_{th}$ leaves positive potential between drain and source and electrons move from source to drain
  - $V_{DS} > V_{th}$ then I_{DS} increases but V_{GD} decreases; when $V_{GD} <$ V_{th} channel is pinched off

THREE REGIONS OF OPERATION

- **Cut-off/sub-threshold region**
  - $V_{GS}< V_{th}$ when no current flows
- **Linear region**
  - $V_{GS}> V_{th}$ & $V_{GS} - V_{DS}> V_{th}$
  - I_{DS} proportional to $\beta*(V_{GS} - V_{th})*V_{DS}$
  - β is transistor gain factor = $\mu*C_{ox}*(W/L)$
- **Saturation regions**
  - $V_{GS}> V_{th}$ & $V_{GS} - V_{DS}< V_{th}$
  - $I_{DS} = (\beta/2)*(V_{GS} - V_{th})^2$

TECHNOLOGY SCALING



Feature/Voltage	Variable
Channel Length	L
Channel Width	W
Oxide Thickness	t_{ox}
Junction Depth	X
Supply Voltage	V_{dd}
Threshold Voltage	V_{th}
Wire width, space, height	w,s,h

Moore's Law:
All these features scale by 1/S,
S=SQRT(2) every 2 years

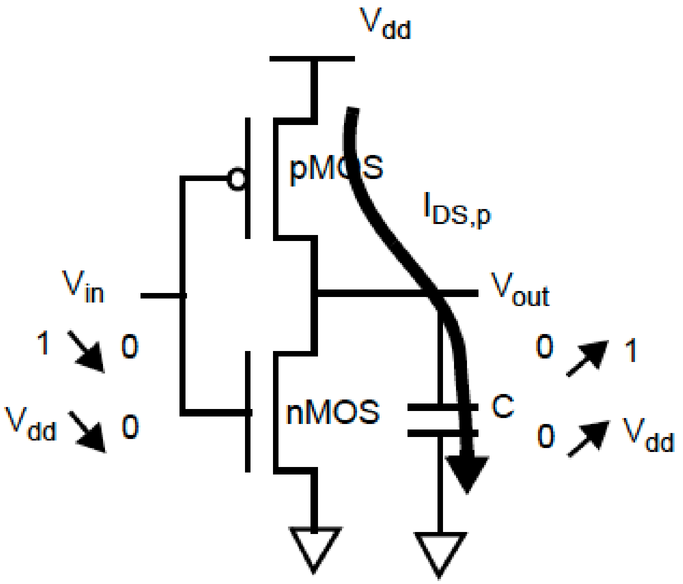
IMPACT OF SCALING ON CHARACTERISTICS

Device Characteristics	Feature Dependence	Scaling
Transistor Gain (β)	$W/(L.\ t_{ox})$	S
Current (I_{ds})	$\beta(V_{dd}-V_{th})^2$	S
Resistance	V_{dd}/I_{ds}	1
Gate Capacitance	$(W.L)/t_{ox}$	1/S
Gate delay	R.C	1/S
Clock Frequency	1/(R.C)	S
Circuit Area	W.L	$1/S^2$
Wire Resistance	1/(w.h)	S^2
Wire Capacitance	h/s	1

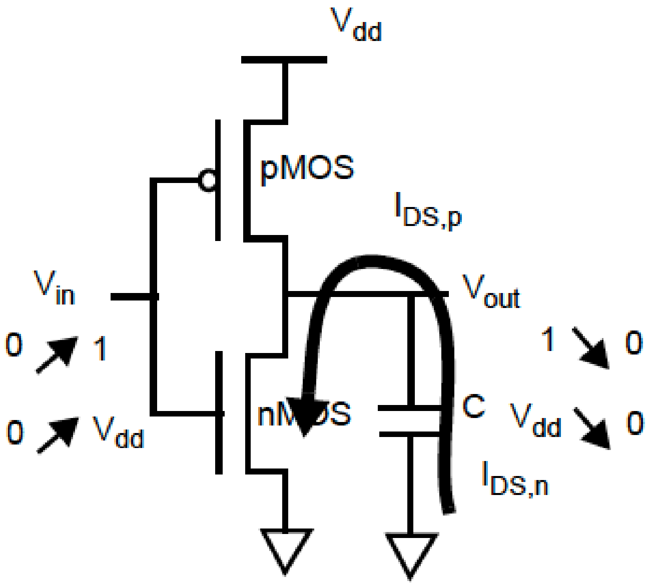
BENEFITS OF SCALING

- Scaling dimensions doubles device density
- Frequency increases by 41%
- Scaling voltage simultaneously keep the power constant
- If voltage is not scaled then clock frequency can scale even faster but dynamic power grows!
- Threshold voltage scaling causes gate leakage power growth!

CMOS INVERTER



(a) Input switches from 1 to 0



(b) Input switches from 0 to 1

- Dynamic power is consumed when device changes ON->OFF and OFF->ON
  - 1->0 current flows to move charge to the capacitance
  - 0->1 current flows from capacitance to ground
- Charging and discharging of capacitance causes power dissipation

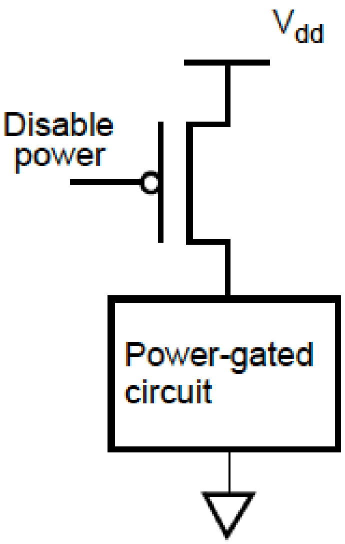
SCALING DYNAMIC POWER

- $P_{dynamic} = aCV_{dd}^2f$
  - a is fraction of clock cycles when gate switches (at most ½)
- C and V_{dd} scale 1/S and f scales S. Hence $P_{dynamic}$ scales like $1/S^2$
- Number of transistor in unit area grow S^2
- Hence power density (power/area) stays constant with scaling

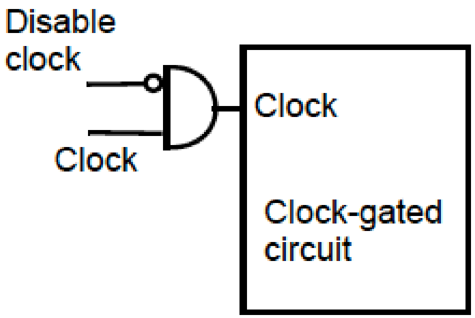
- If V_{dd} does not scale then power density grows
  - Has become a serious issue in recent years since voltage can not reduce very much beyond current levels
- If chip size grows then total power grows
  - Also a major issue in server chips as the size of the chip grows to accommodate new functionality
- Power dissipation leads to heat generation
  - When heat is not removed at the same rate it causes thermal emergencies

REDUCING DYNAMIC POWER

- **Reduce α**
  - Power gating cuts off power from idle units
  - Clock gating cuts off power-hungry clock



(a) Power gating



(b) Clock gating

REDUCING DYNAMIC POWER

- **Reduce V_{dd} – most effective approach**
  - Voltage scaling
  - For correct circuit operation it also requires simultaneous reduction in frequency since transistor becomes slower at lower voltage

- **Reducing V_{dd} and f reduces power cubically**
  - Reduction comes at the cost of performance

Scaling INTELLIGENTLY is the key to preserve performance

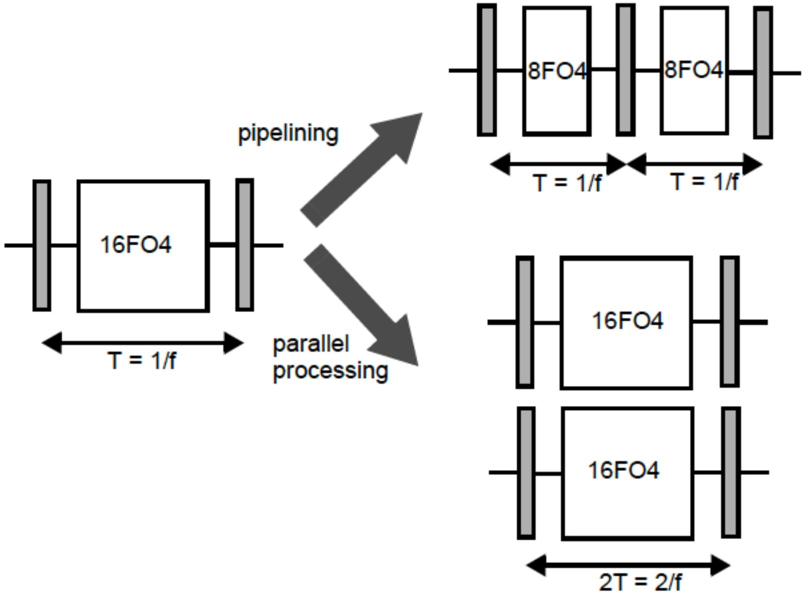
OTHER SCALING APPROACHES

- **Reduces V_{dd} and V_{th} simultaneously so frequency does not need to scale**
  - Causes leakage power growth

- **Use multiple threshold CMOS devices (MTCMOS)**
  - Selectively use high leakage (but faster) devices when speed is critical, else use low leakage devices on non critical paths

- **Use multiple voltage/frequency domains**
  - Already used to some extent where caches run at a different voltage than logic

POWER OF PARALLELISM



- Two choices for the same design
  - Pipeline the design so each of the two stages runs at the same frequency but does half the work
  - Divide the work into two parallel units each running at half the frequency
  - In both scenarios, reduce supply voltage by ½ for ¼th power consumption

STATIC POWER

$$P_{static} = VI_{sub} \,\alpha\, Ve^{-KVt/T}$$

- When V_{GS} drops below V_{th}, I_{DS} still exists leading to static power dissipation
- The current in sub-threshold region is exponentially dependent on V_{th} as well as the operating temperature.
- Hence as V_{th} decreases static power increases exponentially

- Techniques to reduce static power are similar to dynamic power
- HOWEVER, from a static power point of view, pipelining is better than parallelism

METRICS

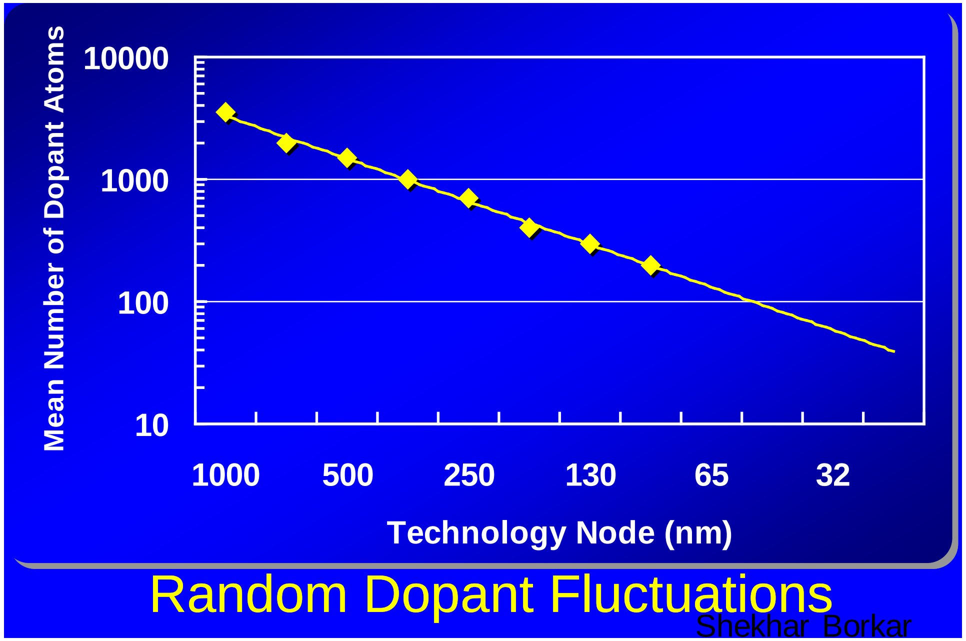
- **Power is good metric for deciding on the thermal envelope of the processor**

- **Energy is good metric in battery constrained environments**
  - Task executed at ½ speed but ¼ power means ½ the energy (2T * ¼ P = ½ E)
  - 2X battery life!

- **Energy*Delay metric gives higher weight to performance**
  - Same example above, ED ($(2T)^2$ * ¼ P) stays same

- **Energy*Delay^2 gives even more weight to performance**
  - Same example above shows that ½ speed is 2X worse on ED^2 metric

PROCESS SCALING AND VARIABILTY

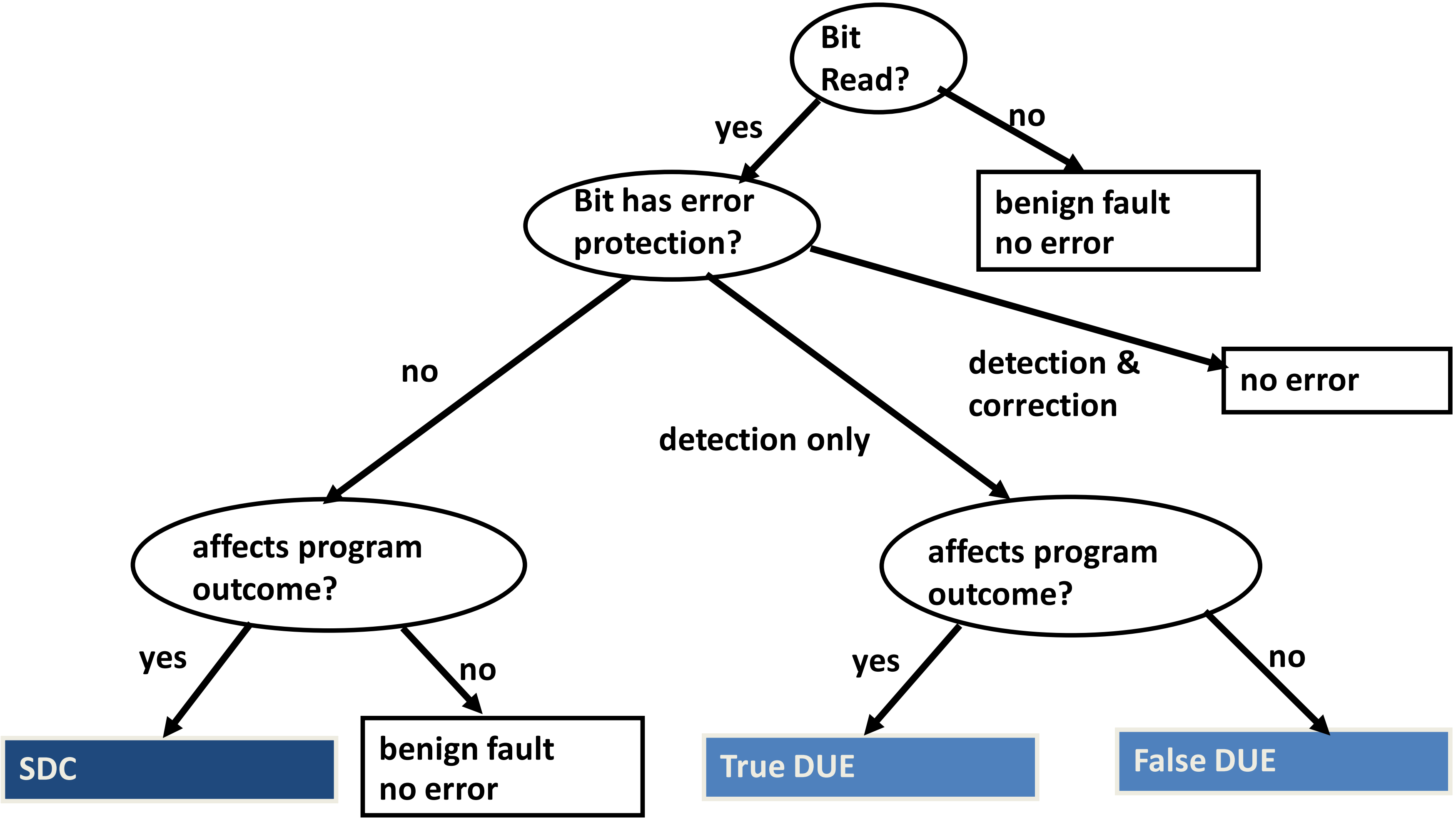


Random Dopant Fluctuations

Shekhar Borkar

- **Process scaling to smaller dimensions leads to**
  - Increased magnitude of within-die parameter variations
  - Greater susceptibility to soft errors
  - More rapid wear-out

FAULT vs. ERROR



AN ERROR DUE TO A FAULT IS CONFINED TO A CONTEXT
EG, CACHE
WHEN AN ERROR AFFECT COORECT EXECUTION IT BECOMES A FAILURE

Source: Shubhu Mukherjee, INTEL

DEFINITIONS

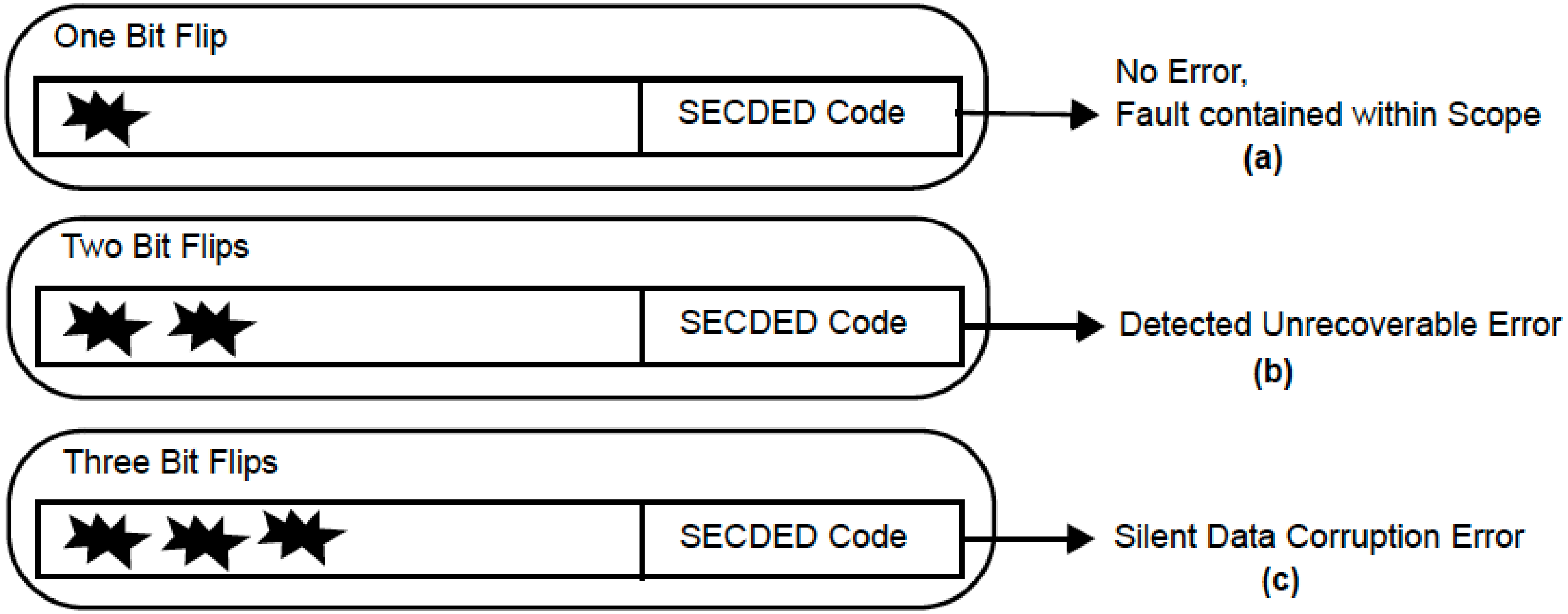
- SDC = Silent Data Corruption
- DUE = Detected and unrecoverable error

- SER = Soft Error Rate = SDC + DUE
- Failure are measured as
  - MTTF = Mean Time to Failure
  - FIT = Failure in Time ; 1 FIT = 1 failure in billion hours
    - 1 year MTTF = 1 billion/(24*365)= 114,155 FIT

- FIT is commonly used because FIT is additive

- Vulnerability Factor = fraction of faults that become errors
  - Also called derating factor or soft error sensitivity

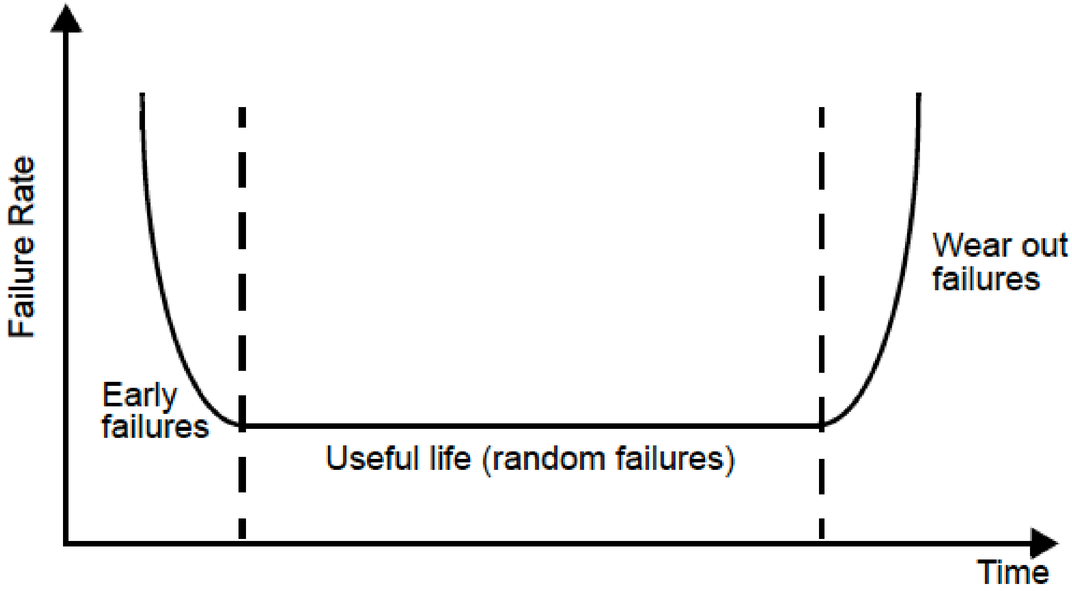
FAULT CONTAINMENT



EXAMPLE: CACHES

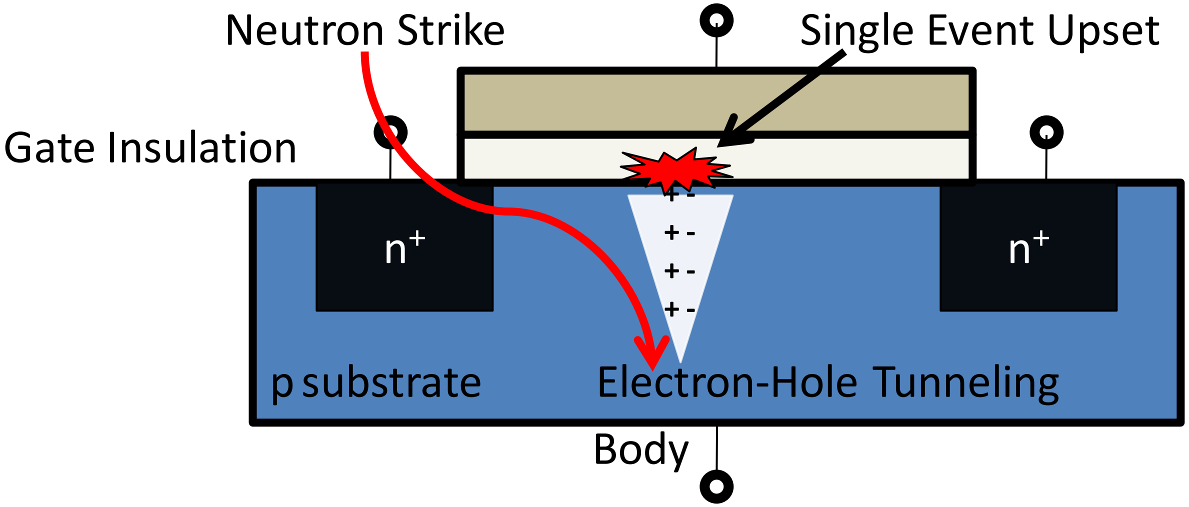
AN SDC OR A DUE BECOMES A FAILURE IF IT AFFECTS PROGRAM EXECUTION

LIFETIME FAILURE RATES



- Failure rate follows bathtub curve
  - Higher failure rates at the initial stage of the manufacturing and operation
  - Long useful life
  - Finally ageing-related wear-out errors
- Burn-in testing removes early failure components

SINGLE EVENT UPSETS



- High energy neutron strike
  - Creates electron-hole pairs by splitting silicon nucleus
  - The charge from the pairs travels toward gate diffusion region
  - Causes the transistor charge to flip
  - Causes a bit to flip
  - Both 0 or 1 stored can be flipped (depending on holes or electron interactions)

BIT SOFT ERROR RATES

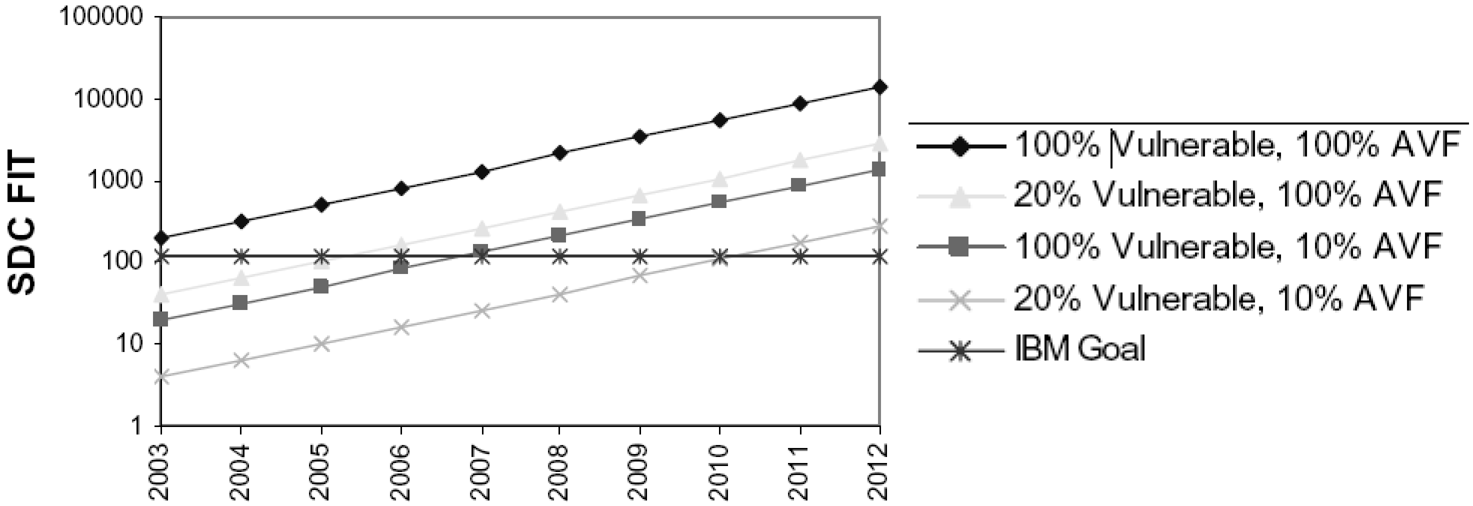
- **SER (λ) = k*flux *bit_area * $e^{Qcritical/Qcollect}$**
  - Flux depends on altitude
  - Bit_area is process technology dependent
  - $Q_{collect}$ is charge collection efficiency also technology dependent
  - Charge needed to flip a bit $Q_{critical}$
    - $Q_{critical}$ α C_{node} * V_{dd}
    - According to scaling rules both C and V decrease and hence Q decreases rapidly

- **Probability of a soft error in a clock cycle**
  - $P_{SE} = \Sigma e^{-\lambda} T_c(\lambda T_c)^k/k!$ for all odd k
  - Tc and bit_area decrease, but λ depends exponentially on 1/Qcritical
  - Hence probability of soft error is largely dependent on Qcritical

AVF(ARCHITECTURAL VULNERABILITY FACTOR)

- AVF_{bit} = Probability that bit matters
  = # of Errors visible to user / Total # of Bit Flips
- FIT_{bit} = intrinsic FIT_{bit} * AVF_{bit}
- Intrinsic FIT of a bit is P_{SE} and is roughly estimated to be 0.001-0.01 FIT/bit
- If we assume AVF = 100% then we will be over-designing the system.
- Need to estimate AVF to optimize the system design for reliability.

MOORE'S LAW AND SDCs



- **Even though fit rate per bit is constant, increasing transistor count raises the system fit rate dramatically**
- **In 2005, we could meet the target SDC FIT by:**
  - With 100% AVF, protect 80% of the bits (20% remain vulnerable)
  - With an AVF of 10%, no protection necessary
- **In 2010 we can meet the target SDC FIT with**
  - With 100% AVF target SDC is unattainable,
  - With 10% AVF protect 80% of the bits

ACE/unACE BITS IN uARCHITECTURE

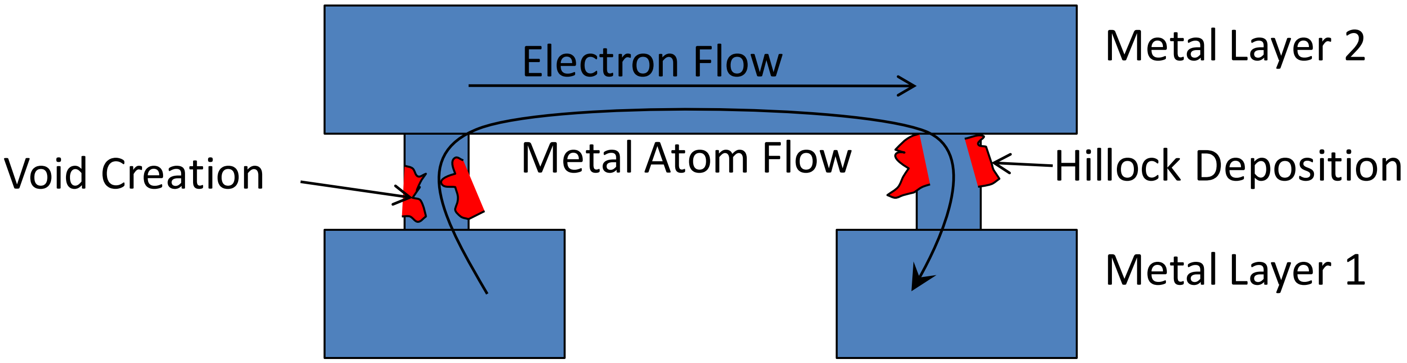
- **Computing AVF requires identifying ACE (Architecturally Correct Execution) and unACE bits**

- **Microarchitectural unACE bits:**
  - Idle/Invalid State: instructions where the opcode bits do not matter or reserved opcode bits
  - Mis-Speculated State: instructions that are being executed speculatively and are not going to retire due to mis-speculation (or exceptions)
  - All forms of predictors: Branch predictor, RAS
  - Dead Bits: Physical registers that have been read by the last consumer but are not deallocated

ARCHITECTURAL unACE BITS

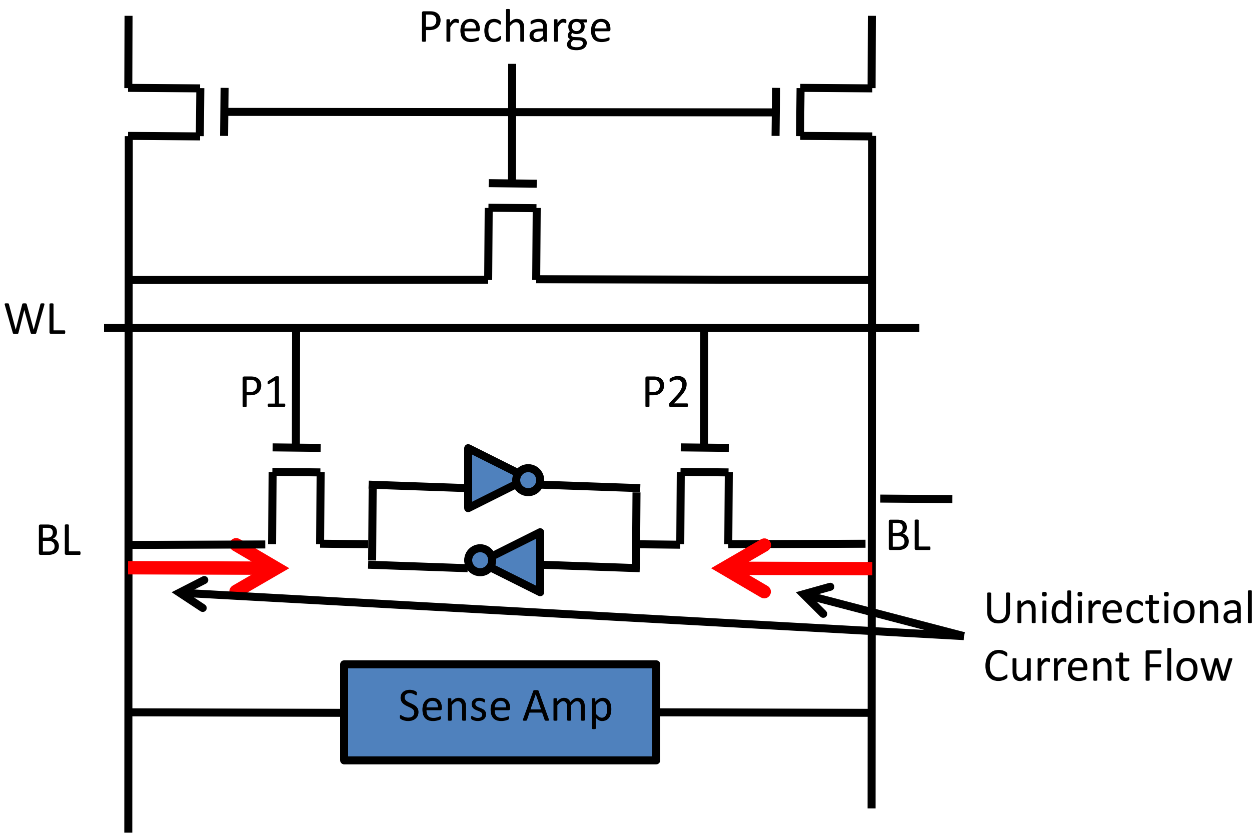
- NOP instructions: Plenty of them around (particularly in VLIW style processors)
  - Only opcode must be protected everything else is a don't care
- Performance-enhancing instructions: Prefetch, Hint bits
- Predicated-false instructions: Itanium ISA supports predication to remove branch prediction (only predicate is ACE)
- Dynamically dead instructions: due to compiler inefficiencies
- Logical masking: Bit masking operations

ELECTROMIGRATION



- **Wire width decreases with scaling**
  - But current density increases
  - Nearly 1000 amps can move through a wire
- **Metal atoms in wire gather momentum and move with the electron flow**
- **Leads to shorts and opens in the wires**

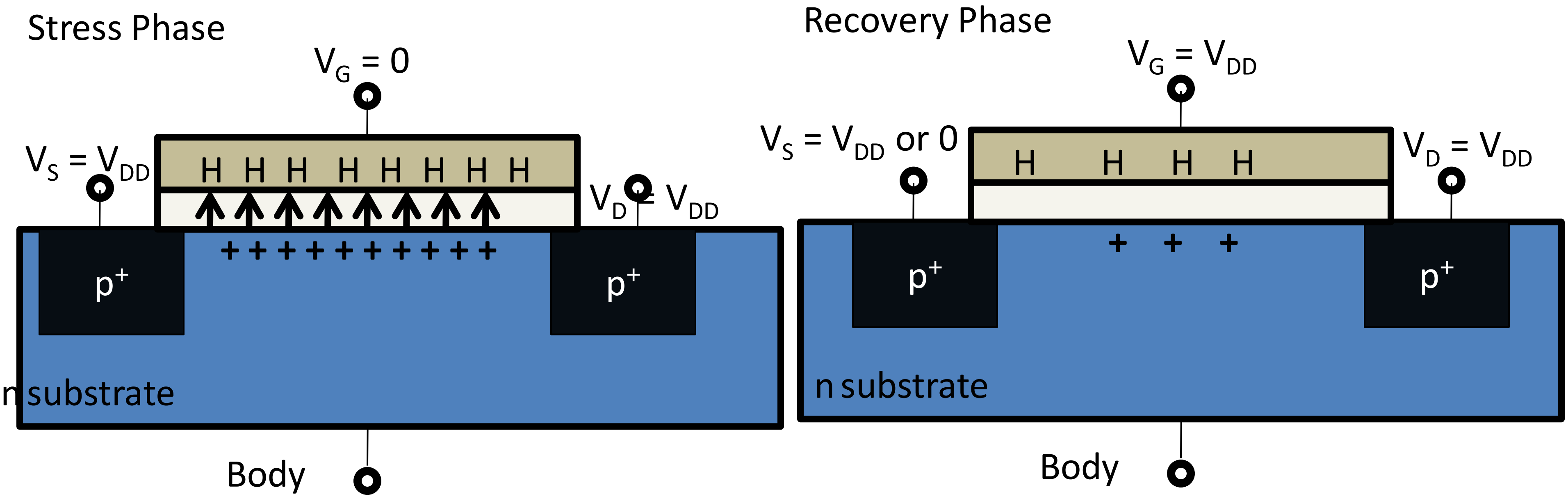
EM IN SRAM CELLS



- **Both the BL and BL-bar have unidirectional current causing EM effects**

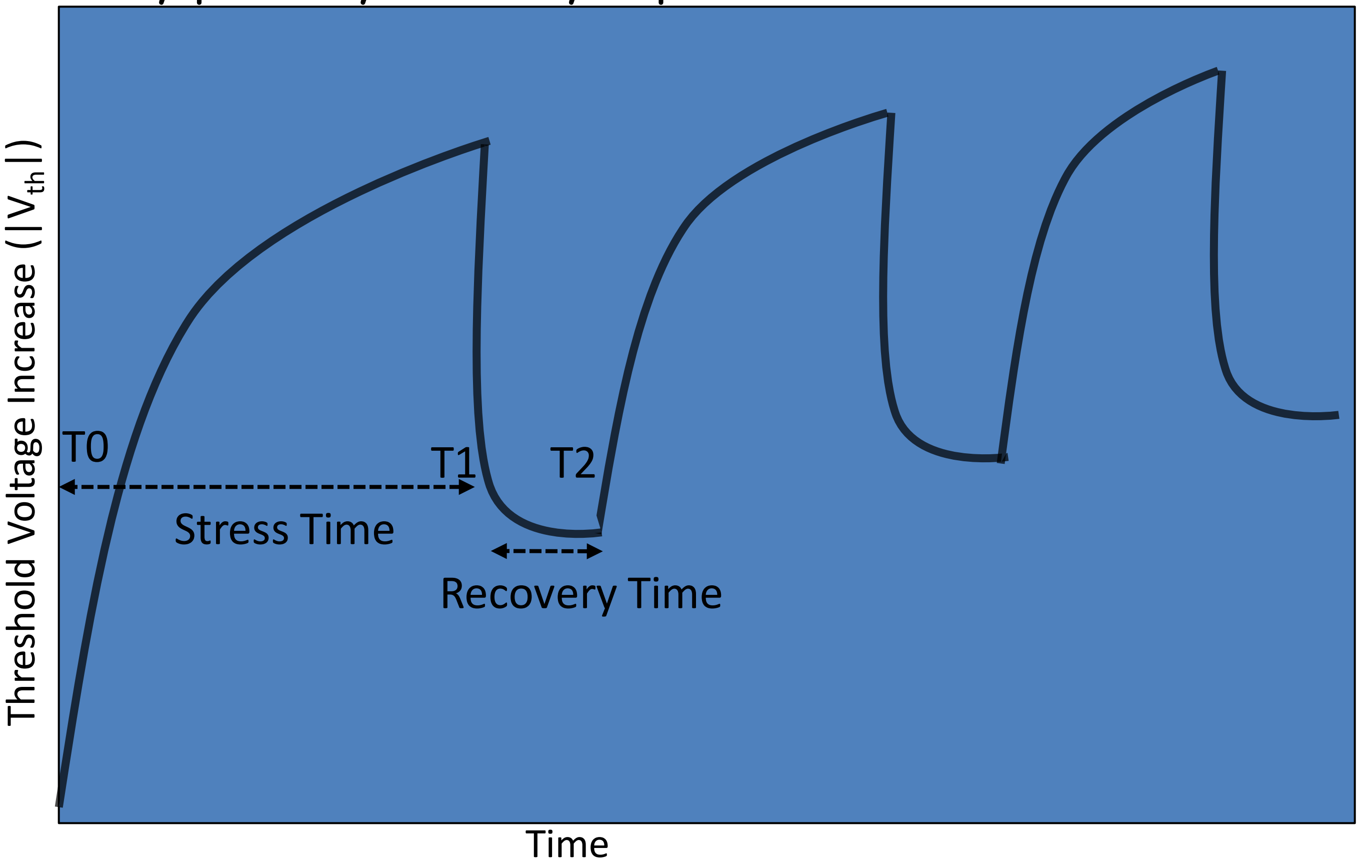
NBTI

+ : interface trap
H : hydrogen
←: electric field

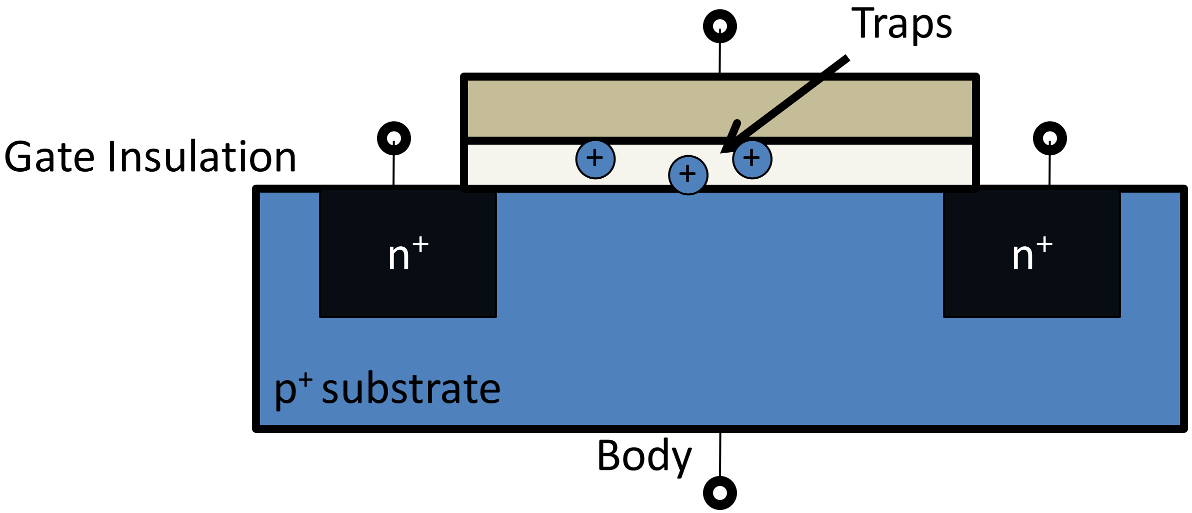


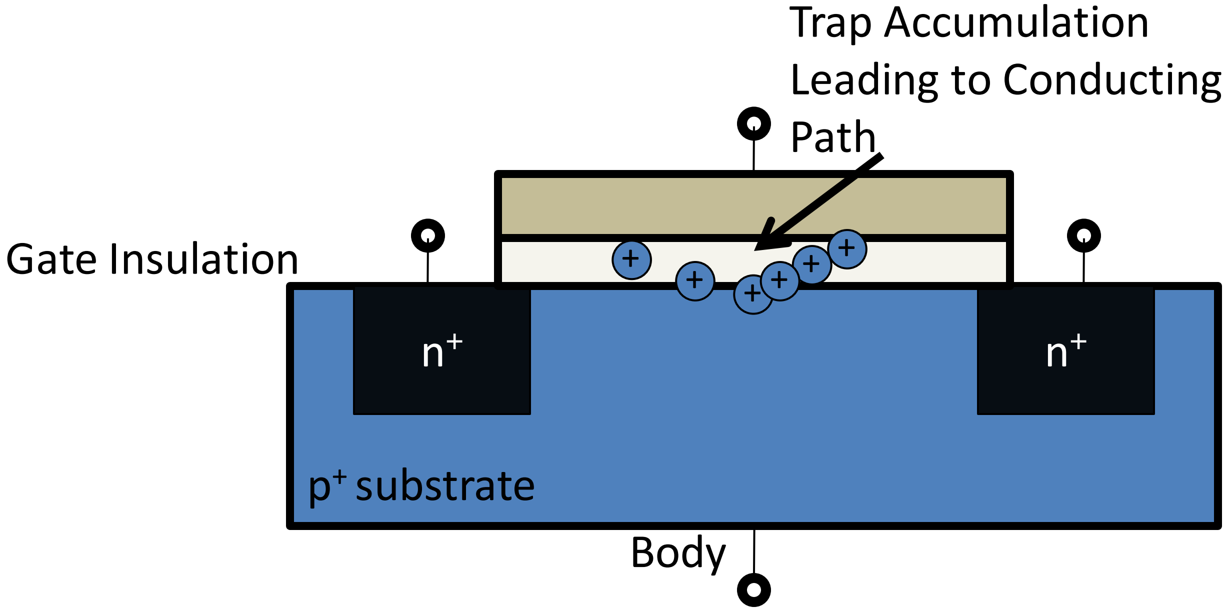
NBTI IMPACT

- **NBTI affects when a negative bias is applied on PMOS**
- **The effect goes away when the bias is removed**
  - Partially recovers
- **Potential to fully recover if the bias is flipped to positive**
  - Yet, only partially recovery is possible



TDDB


Traps
Gate Insulation
+
+
+
n+
n+
p+ substrate
Body


Trap Accumulation
Leading to Conducting
Path
Gate Insulation
+
+
+
+
+
+
n+
n+
p+ substrate
Body