Low-Leakage SRAM Design in Deep Submicron Technologies

Behnam Amelifard, Farzan Fallah, and **Massoud Pedram**

Univ. of Southern California

Los Angeles CA USA

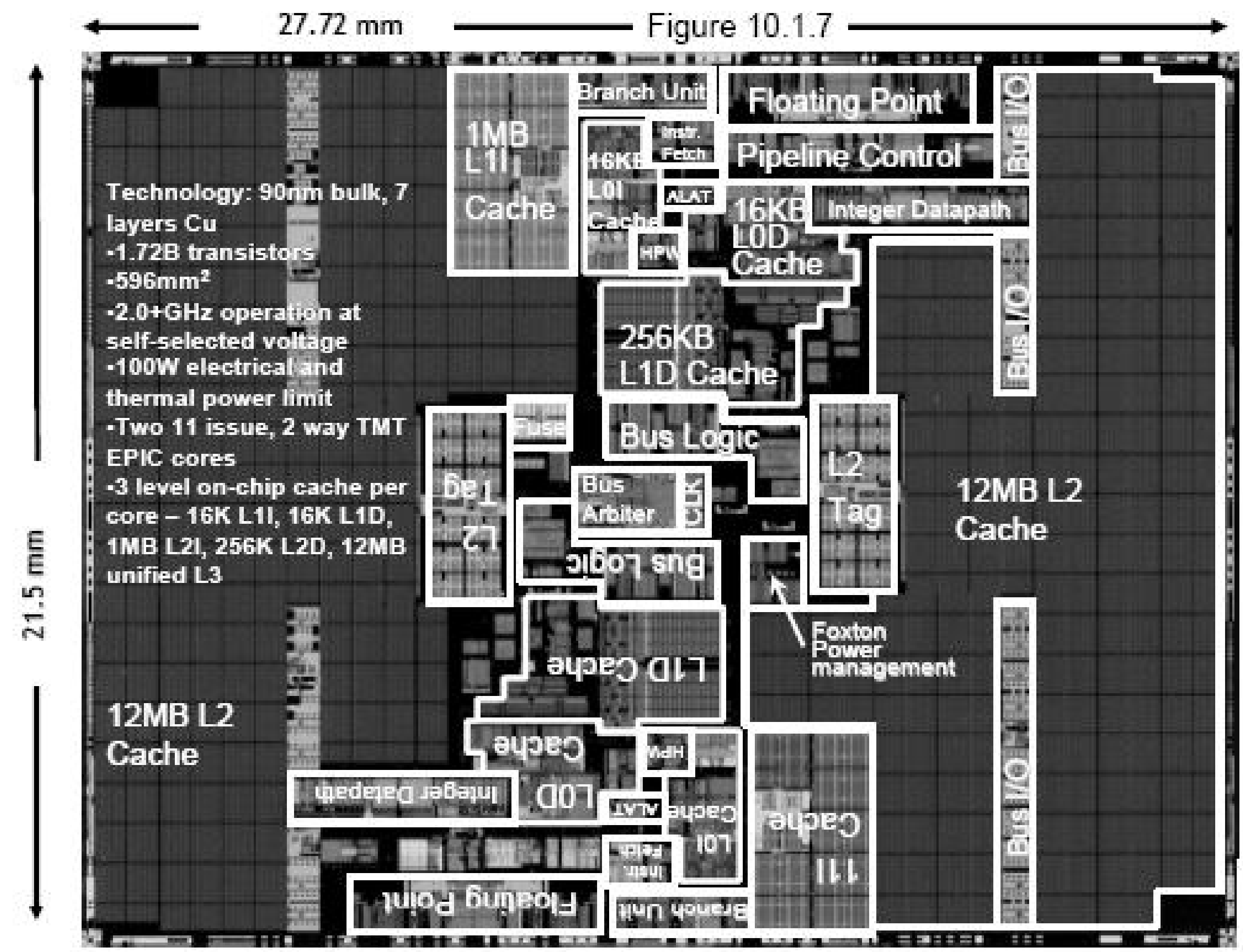
Jan 25, 2008

Presentation at SNU

Outline

- Introduction
- Related prior work
- Heterogeneous cell SRAM
- PG-gated SRAM cell
- Concluding remarks

Montecito, Intel's latest Itanium chip



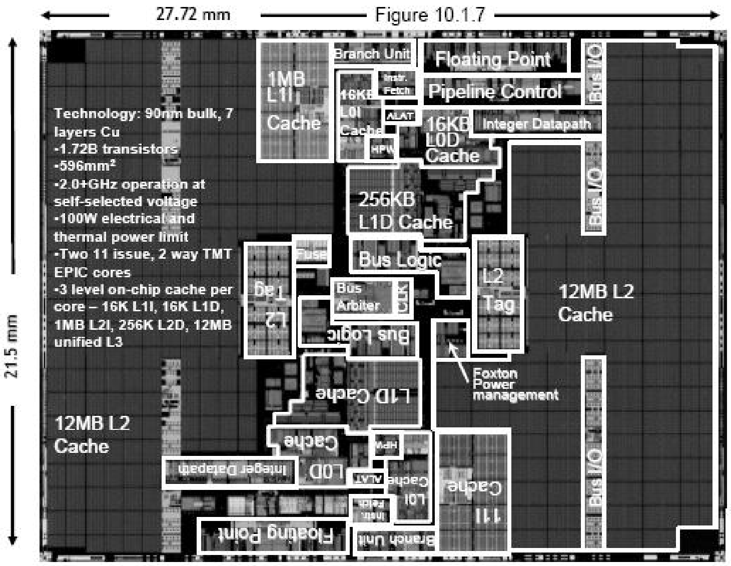
Technology: 90nm bulk, 7 layers Cu

- ·1.72B transistors
- ·596mm²
- ·2.0+GHz operation at self-selected voltage
- ·100W electrical and thermal power limit
- ·Two 11 issue, 2 way TMT EPIC cores
- ·3 level on-chip cache per core – 16K L1I, 16K L1D, 1MB L2I, 256K L2D, 12MB unified L3

Courtesy of Intel

Montecito Specification

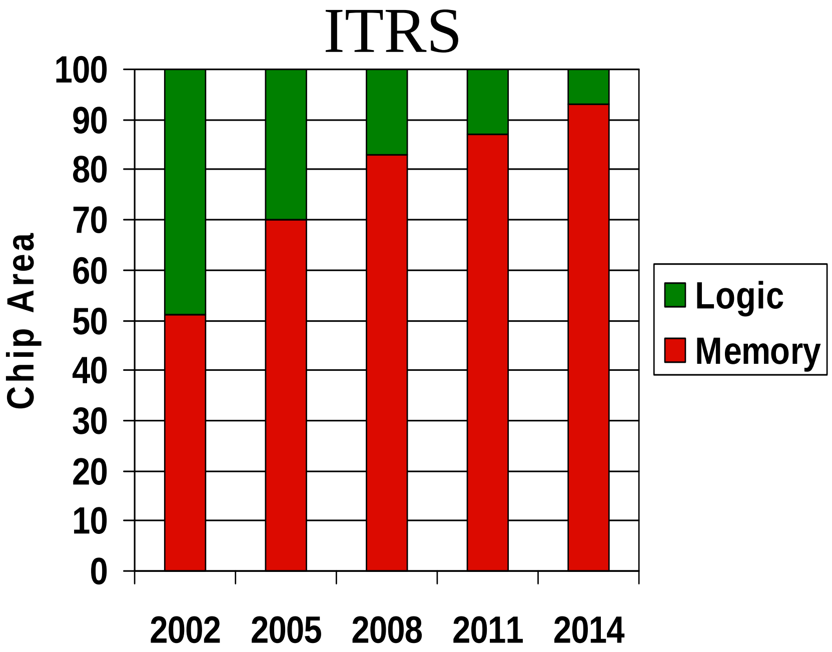
- Cache system
  - Separate 16KB L0 I-cache and 16KB L0 D-cache per core
  - Separate 1MB L1 I-cache and 256KB L1 D-cache per core
  - 12MB L2 cache per core
  - → **About 80% of die area dedicated to caches**

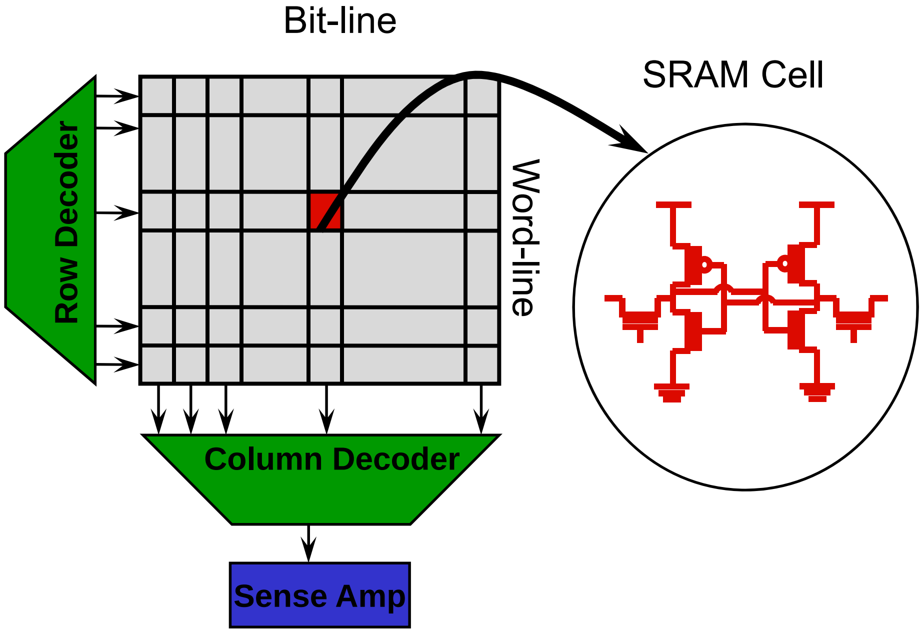


Technology: 90nm bulk, 7 layers Cu
-1.72B transistors
-596mm²
-2.0+GHz operation at self-selected voltage
-100W electrical and thermal power limit
-Two 11 issue, 2 way TMT EPIC cores
-3 level on-chip cache per core – 16K L1I, 16K L1D, 1MB L2I, 256K L2D, 12MB unified L3

- → Transistor count
  - 1.72B transistors
    - Core logic: 57M
    - Bus logic and I/O: 6.7M
    - L0 and L1 caches: 106.5M
    - L2 cache: 1.55B
  - → **96% of transistors are used in caches**

Introduction





- Leakage power is roughly proportional to area
- Leakage power of caches is a major source of power consumption in high performance microprocessors

Design Low-leakage SRAM

Leakage Components

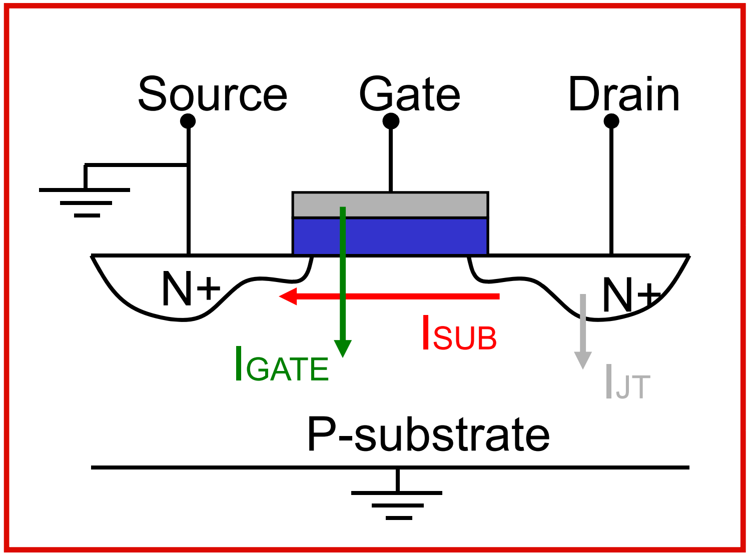
- ❑ Subthreshold current

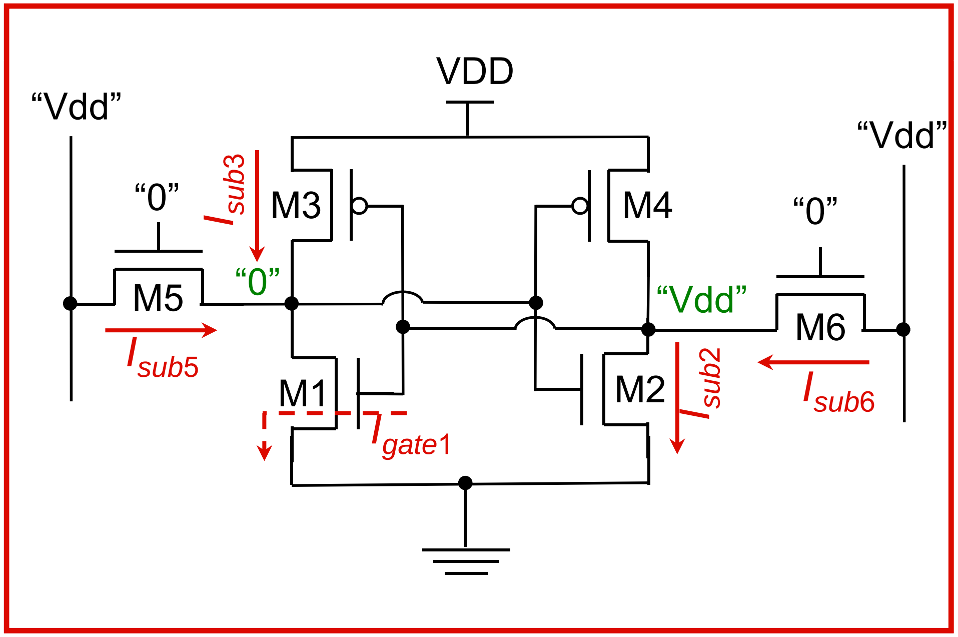
$$I_{sub} = A_{sub} w \exp\left(\frac{q}{n'kT}\left(V_{gs} - V_{t0} - \gamma f(V_{sb}) + \eta V_{ds}\right)\right)\left(1 - \exp\left(-\frac{q}{kT}V_{ds}\right)\right)$$

- ❑ Gate-tunneling current
  - ❑ Dominated by gate-to-channel current of ON NMOS transistors

$$I_{gate} = A_{ox} w_N \left(\frac{V_{ox}}{t_{ox}}\right)^2 e^{-B_{ox}\frac{t_{ox}}{V_{ox}}}$$

- ❑ Junction leakage
  - ❑ Small contributor to total leakage current at current CMOS technology nodes



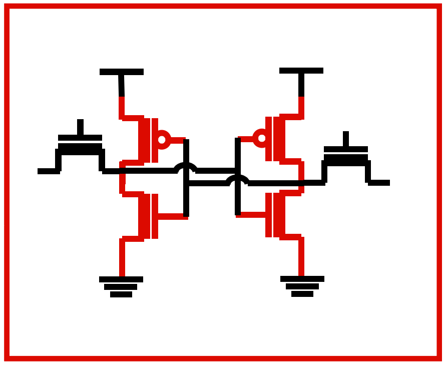


Outline

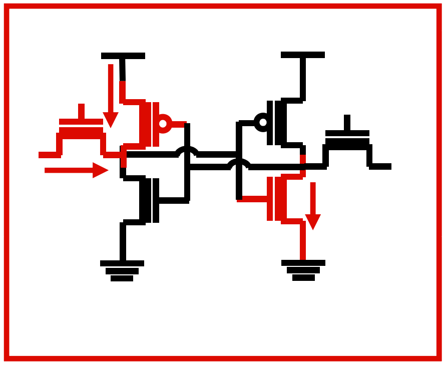
- ❑ Introduction
- ❑ Related prior work
- ❑ Heterogeneous cell SRAM
- ❑ PG-gated SRAM cell
- ❑ Concluding remarks

Related Prior Work

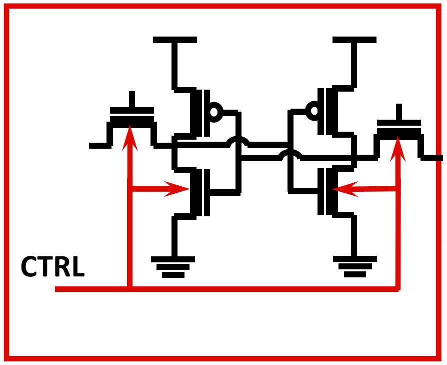
- Dual Vt SRAM
  - Use high-Vt for pull-down and pull-up TX



- Asymmetric cell SRAM
  - In ordinary programs most of bits in D-cache and I-cache are zero

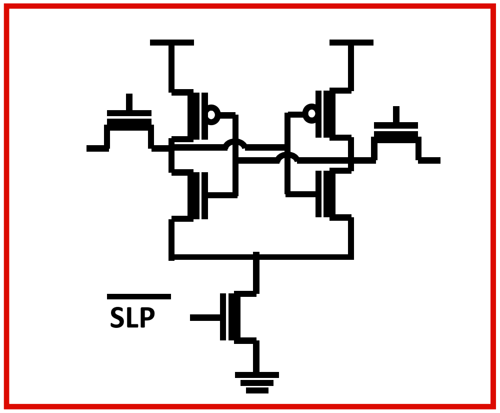


- Dynamic Vt SRAM
  - Dynamically change Vt of cells in a row

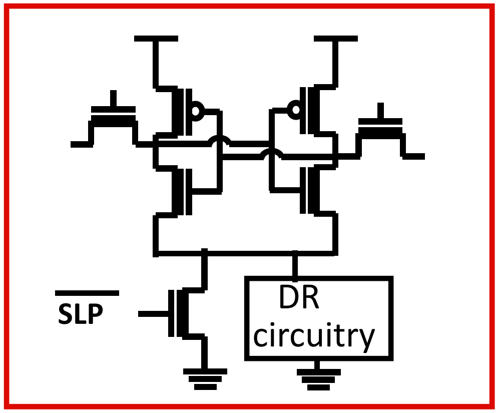


Related Prior Work

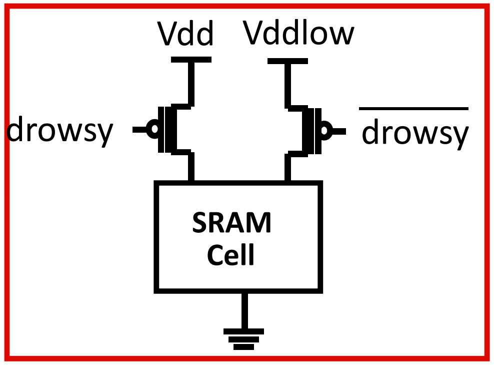
- ❑ Power-gated SRAM
  - ❑ Dynamically disconnect cell power supply



- ❑ Data Retention Power-gated SRAM
  - ❑ Use a data-retention circuitry to control virtual ground voltage



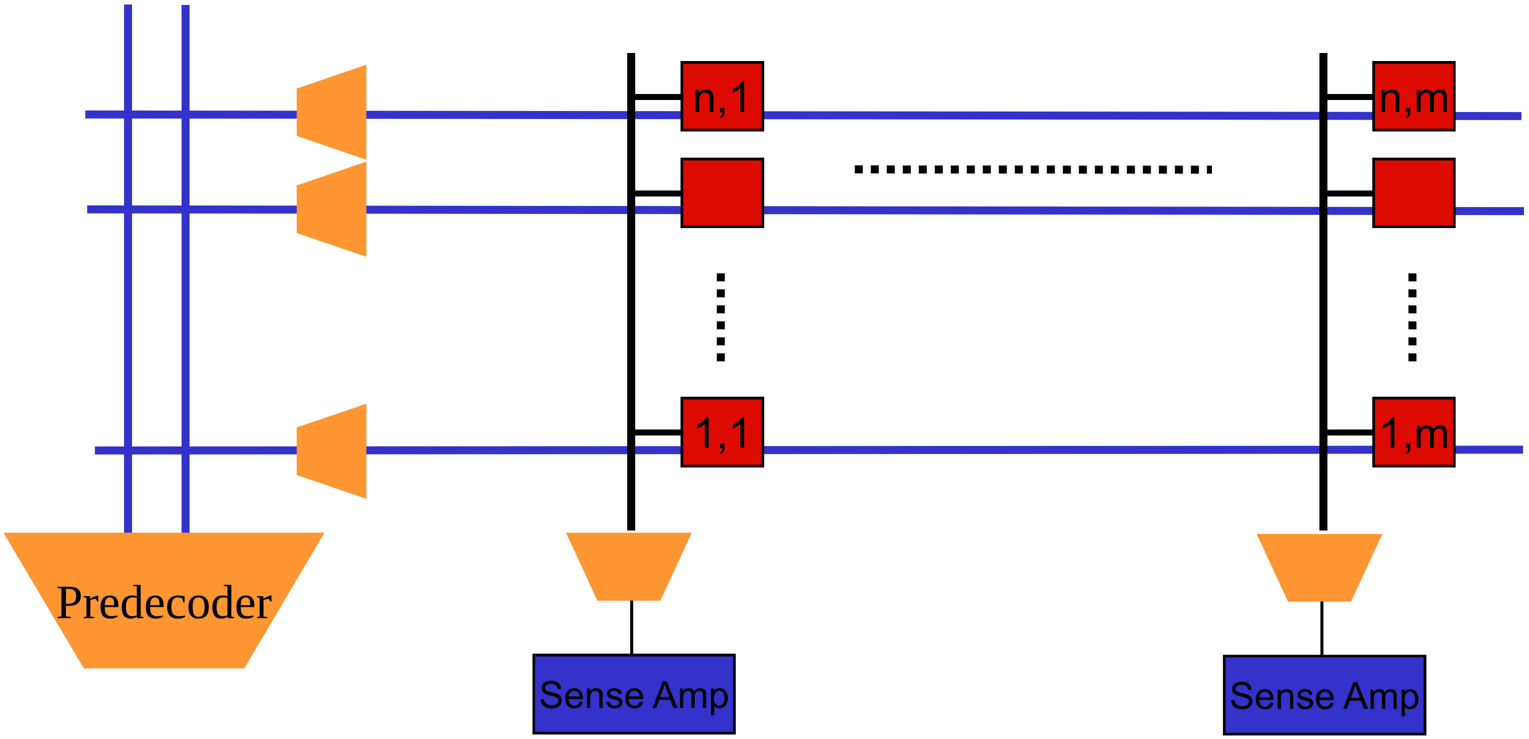
- ❑ Drowsy cache
  - ❑ Put inactive cache lines in a low voltage standby mode



Outline

- ❑ Introduction
- ❑ Related prior work
- ❑ Heterogeneous cell SRAM
- ❑ PG-gated SRAM cell
- ❑ Concluding remarks

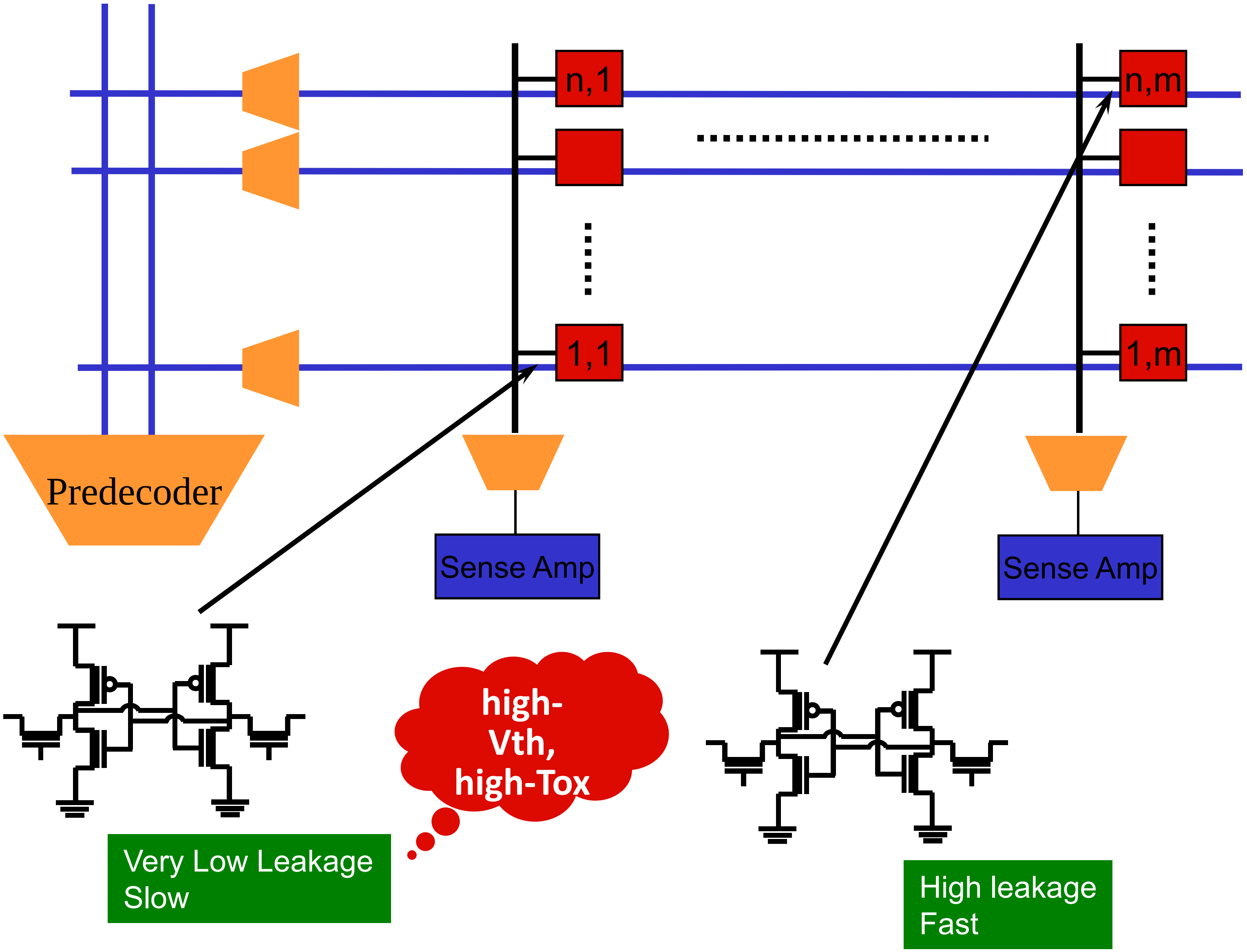
Heterogeneous Cell SRAM (HCS)



- Key observation: Read/write delay of a cell depends on its physical distance from predecoder and sense Amplifier

$$\boxed{\text{delay}_{1,1} < \text{delay}_{n,m}}$$

HCS (Cont'd)


n,1
n,m
1,1
1,m
Predecoder
Sense Amp
Sense Amp
high-Vth, high-Tox
Very Low Leakage
Slow
High leakage
Fast

Library Generation

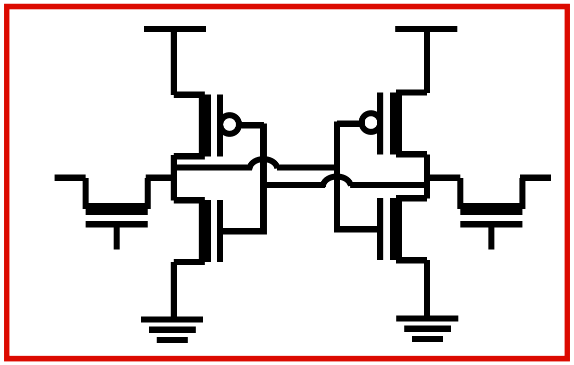
- If all Vth and Tox inside a cell are high
  - Maximum power reduction
  - Maximum delay increase
- We also consider increasing Vth or Tox of some transistors in cell, which results in less delay penalty
  - Do not use both high-Tox and high-Vth for a transistor
  - Do not use high-Tox for PMOS transistors
- To make memory cells more manufacturable, only symmetric cells are considered
  - Each configuration shown as (x,y,z)

3×2×3=18 cell configurations

Pull-down

Pull-up

Pass-trs



x,y,z=

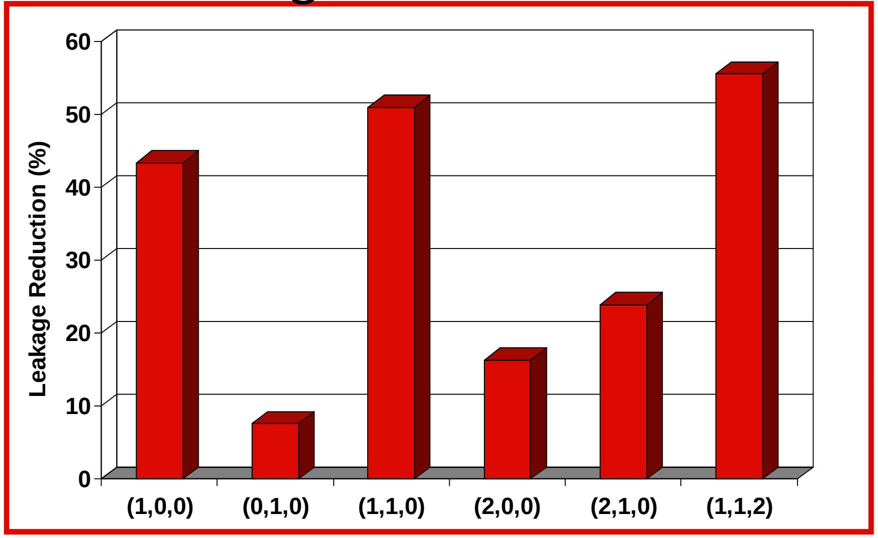
- 0 : if corresponding tr's are normal
- 1 : if corresponding tr's are high-Vth
- 2 : if corresponding tr's are high-Tox

Library Generation (Cont'd)

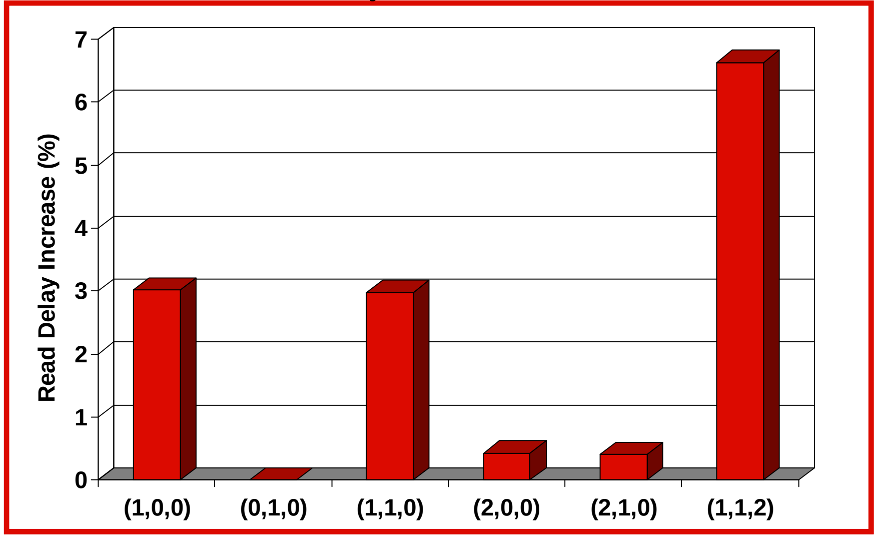
- By simulating all possible configurations, inferior cells, i.e., those with higher leakage and longer read/write delay than at least one other configuration, are eliminated

- Only the non-inferior configuration set (NICS) is used for optimization

Leakage of cells in NICS



Read delay of cells in NICS

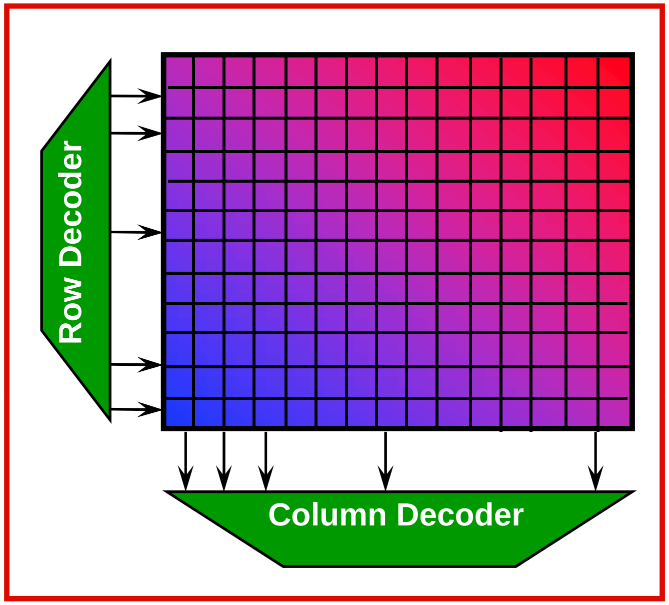


Heterogeneous Cell Assignment

- Start from a pre-designed SRAM with all low-Vth and low-Tox cells, i.e., (0,0,0) cell
- Sort configurations in decreasing order of leakage:
  - $\{C_0, C_1, \dots, C_n\}$, $C_0=(0,0,0)$, C_n=least leaky configuration
- Find out the slowest read and write delays
- Replace as many C_0 cells as possible with C_n cells in such a way that access delay of replaced cells will not be larger than slowest access delay in the original SRAM design
- Try to replace remaining C_0 cells with other configurations in descending order of leakage saving, i.e., $C_{n-1}, \dots, C_2, C_1$

Blue: low leakage

Red: high leakage

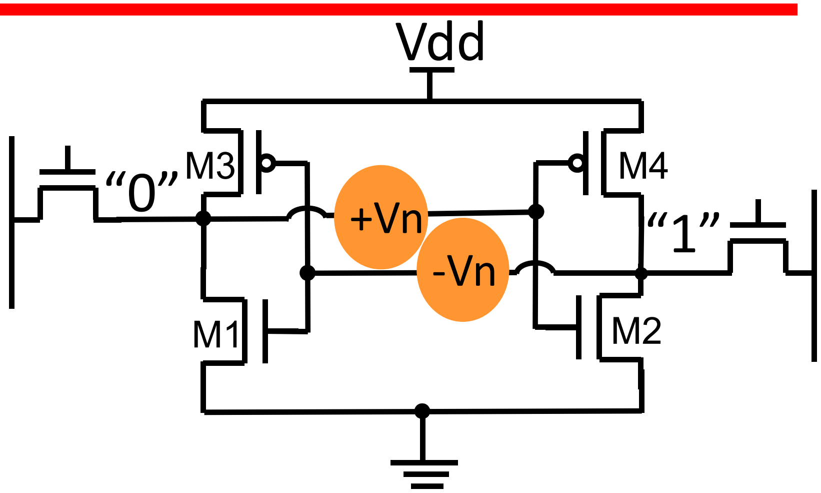


SRAM Cell Design Considerations

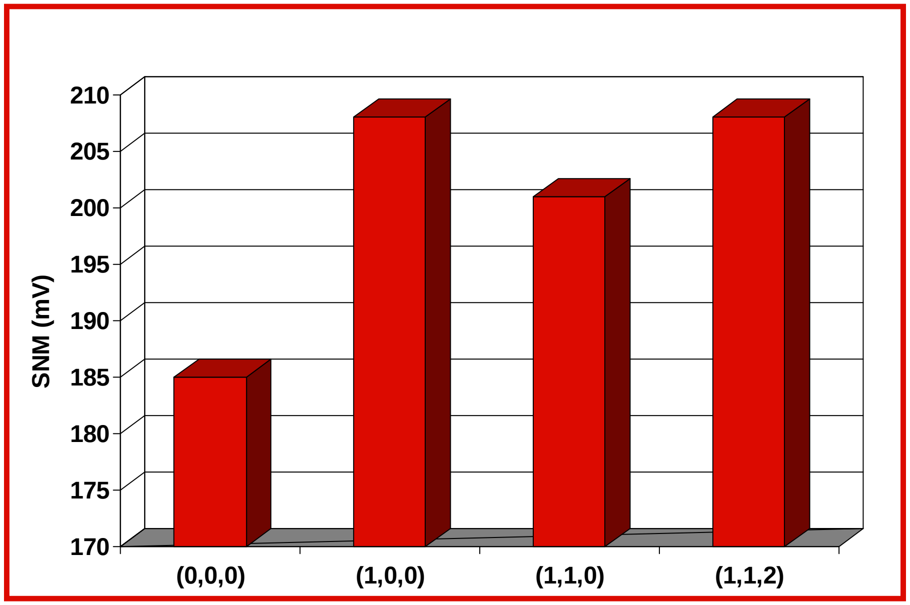
- ❑ Static Noise Margin
- ❑ Read Stability
- ❑ Writability
- ❑ Soft Error Immunity

Hold Static Noise Margin (SNM)

- SNM: Maximum value of dc noise voltage (V_n) that can be tolerated by the cell before changing state
  - The SNM is determined by the ratio of width-to-length of pull down transistor to width-to-length of pass gate transistor (PG)
  - SRAM cells are especially sensitive to noise during a read operation
- A configuration is said to be robust if its SNM is no smaller than that of config. (0,0,0)
- To design an HCS as robust as the conventional SRAM, only the non-inferior robust configuration set (NIRCS) is used



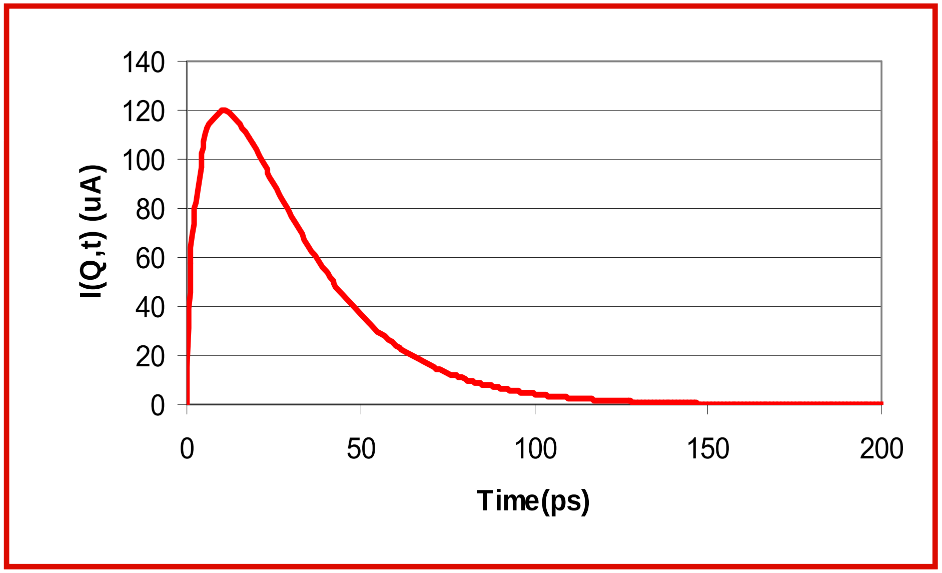
SNM of cells in NIRCS



Soft Error Rate (SER)

- A high-energy alpha particle or an atmospheric Neutron striking a capacitive node
  - Deposits charge leading to a time-varying current injection at the node
- In case of atmospheric Neutrons: $I(Q,t) = \frac{2Q}{\sqrt{\pi} T_s} \sqrt{\frac{t}{T_s}} \exp\left(\frac{-t}{T_s}\right)$
- If collected charge *Q* exceeds critical charge Q_{crit}, it will upset bit value and cause a soft error
- Soft error rate (SER) in SRAM

$$SER \propto A_S N_{flux} \exp\left(-\frac{Q_{crit}}{Q_s}\right)$$

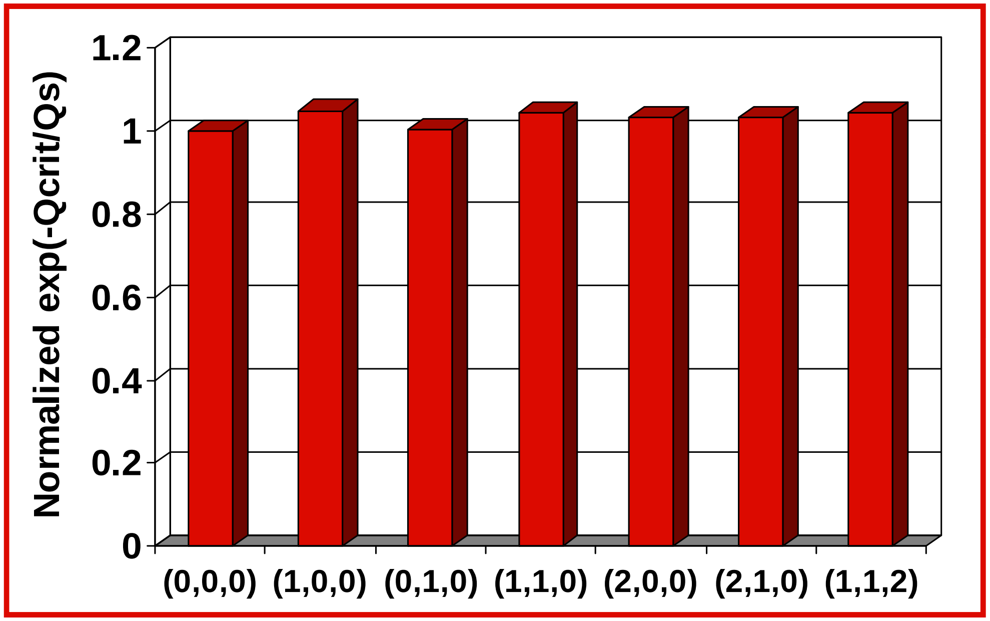


SER (Cont'd)

- We concentrate on exp(- Q_{crit}/Q_s) in evaluating SER of the HCS
  - Other parameters of SER are not affected by the HCS design

Q_{cirt} of cells in NICS

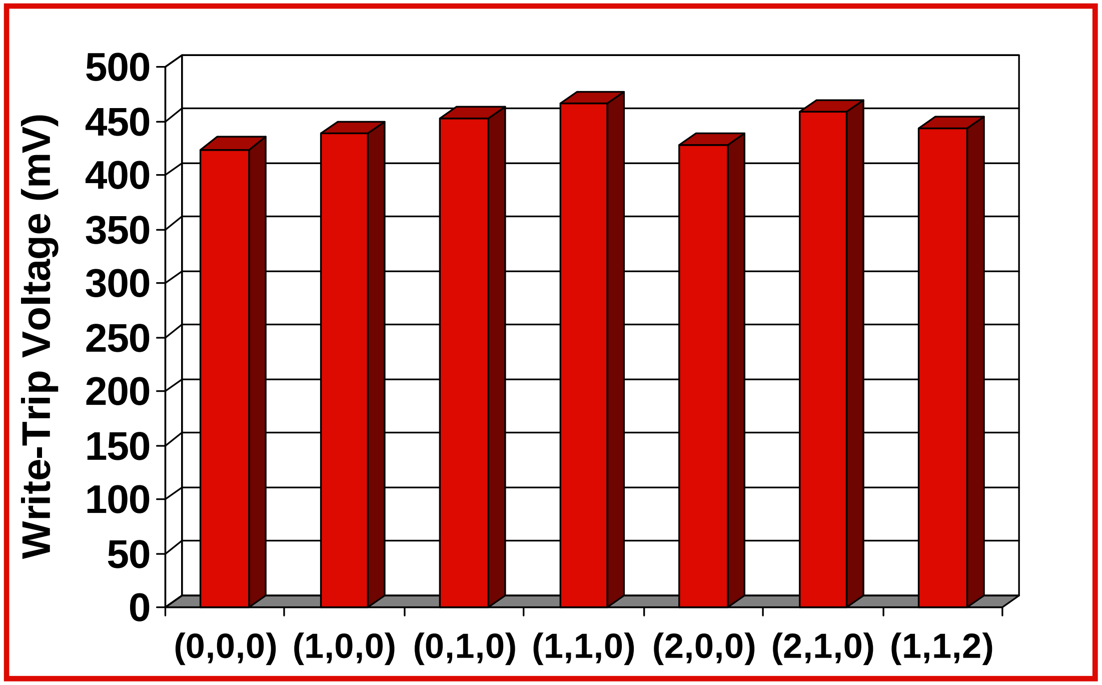
For 65nm node, Qs=10fC



- SER of the HCS is only marginally affected
  - Maximum increase: 4.8%

Writability

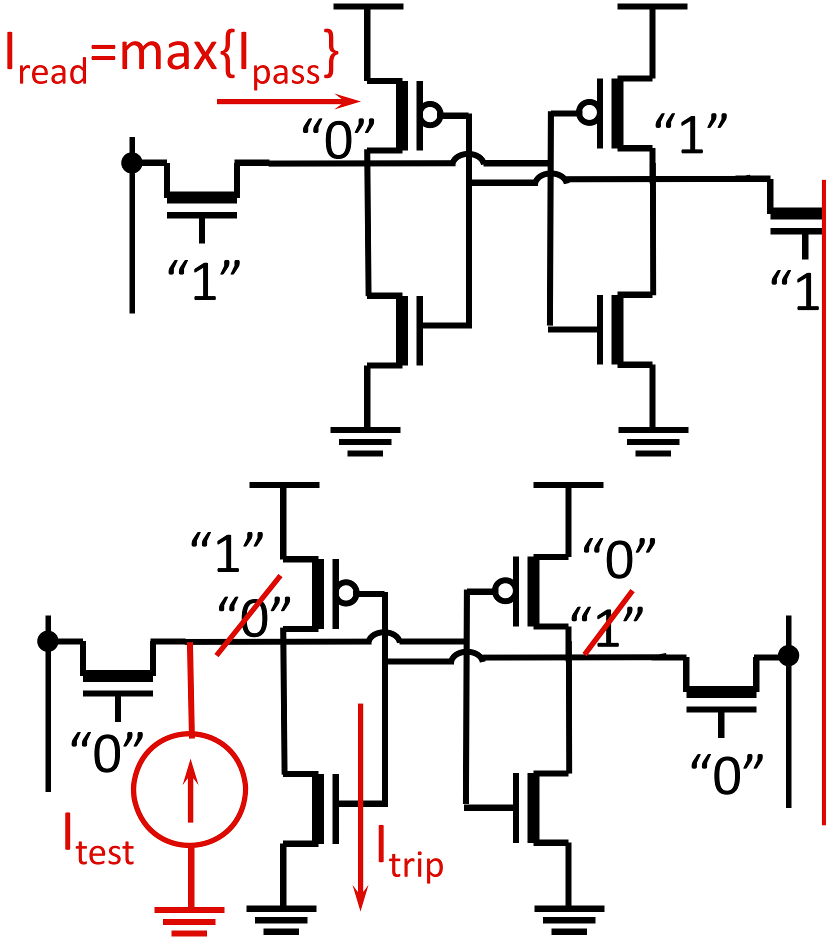
- **Write-trip voltage:** Highest voltage on bit-line, which is pulled down during write operation, at which the state of the SRAM cell is changed
  - This write-trip voltage is determined by the ratio of the width-to-length ratio of the pull up transistor to the width-to-length ratio of the PG
  - Higher value for write-trip voltage represents ease of writability
- The write-trip voltage should be sufficiently lower than V_{dd}
  - Noise cannot cause a write failure or an unintentional write

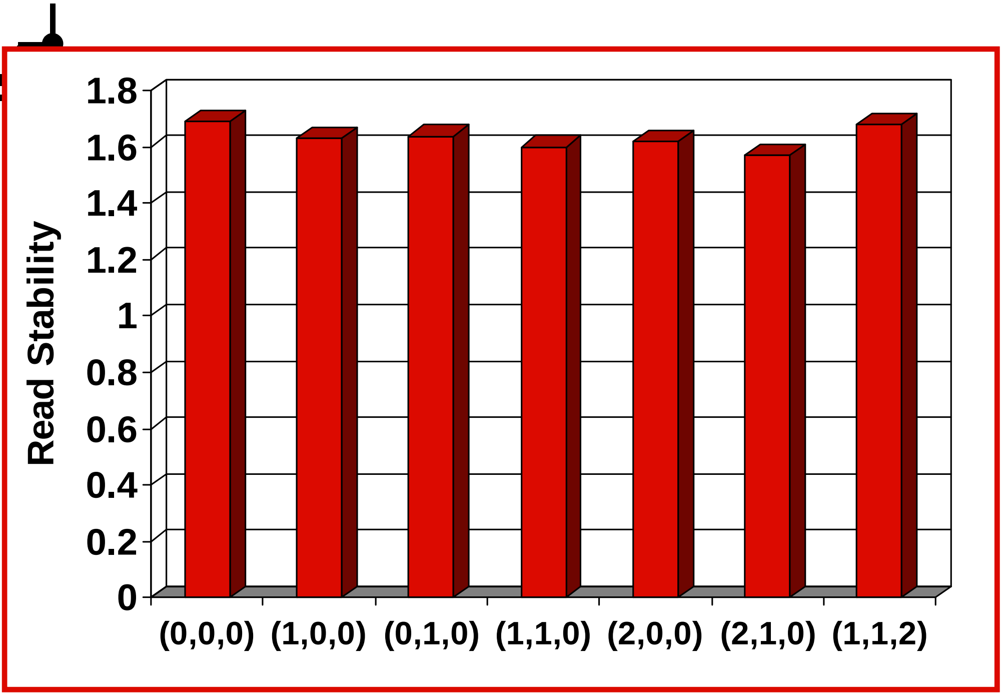


Read Stability

- A transient stability metric
  - Likelihood of inverting an SRAM cell’s stored value during a read operation

$$\text{Read Stability} = I_{trip} \,/\, I_{read}$$



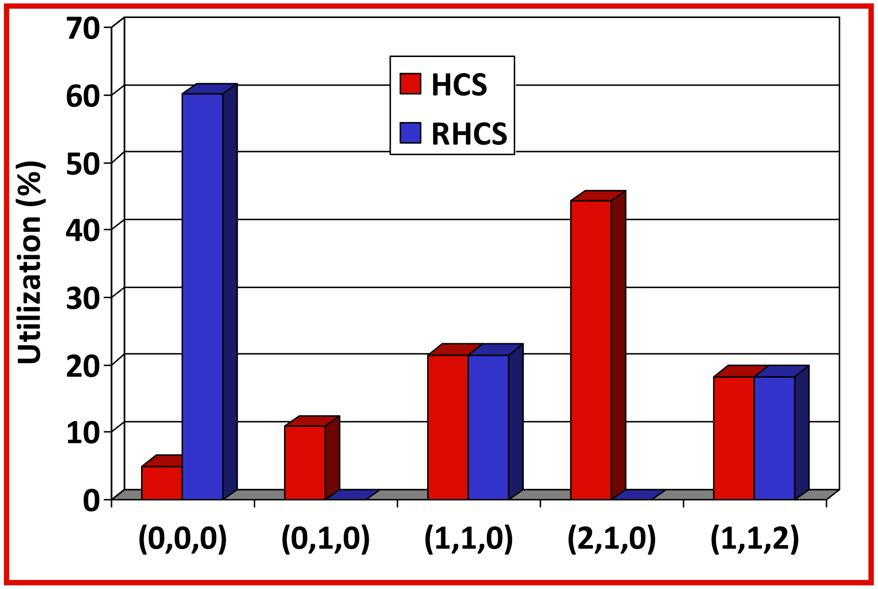


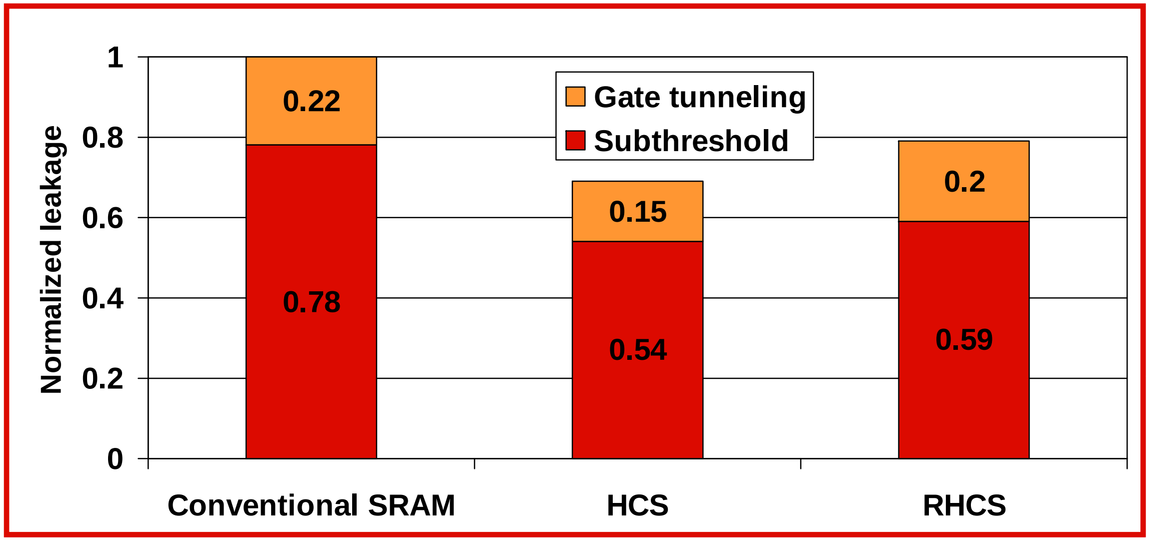
Experimental Results

- SRAM specifications:
  - 64Kb with a 64-bit word
  - V_{dd}=1.1V
  - low-Vth=0.18V, high-Vth=0.28V
  - low-Tox=12A°, high-Tox=14A°

32.2% leakage reduction in HCS

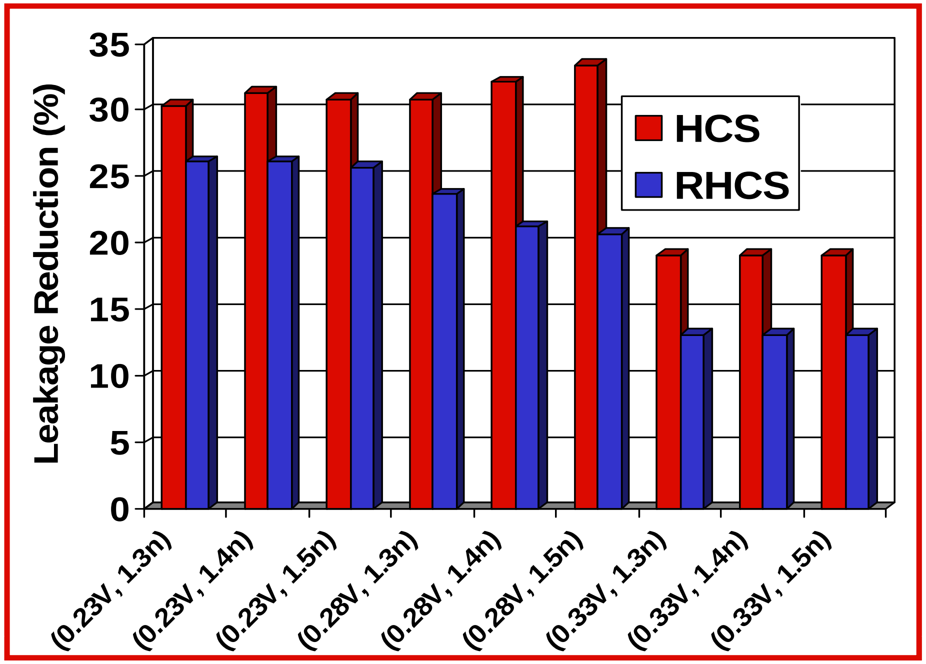
21.2% leakage reduction in RHCS



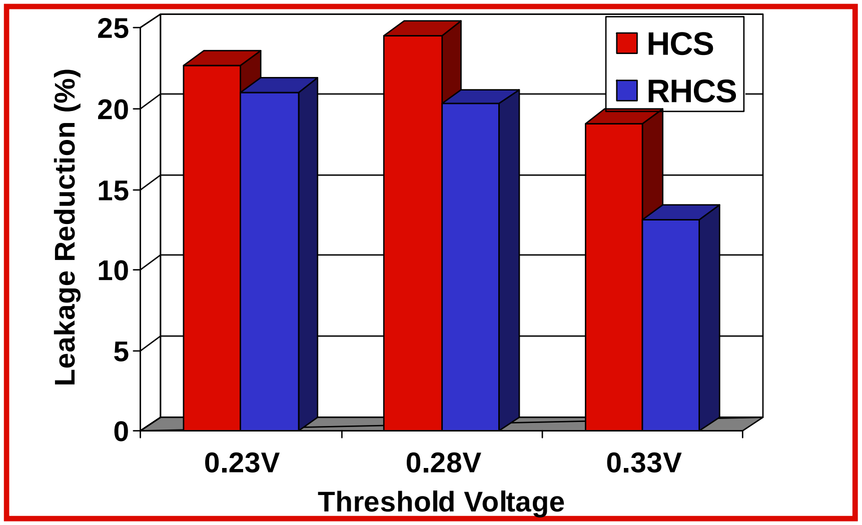


Effect of high-Vth and high-Tox Selection

- Three values for high-Vth and three values for high-Tox
  - Power reduction is a weak function of high-Tox value
  - For very high values of high-Vth, power reduction drops

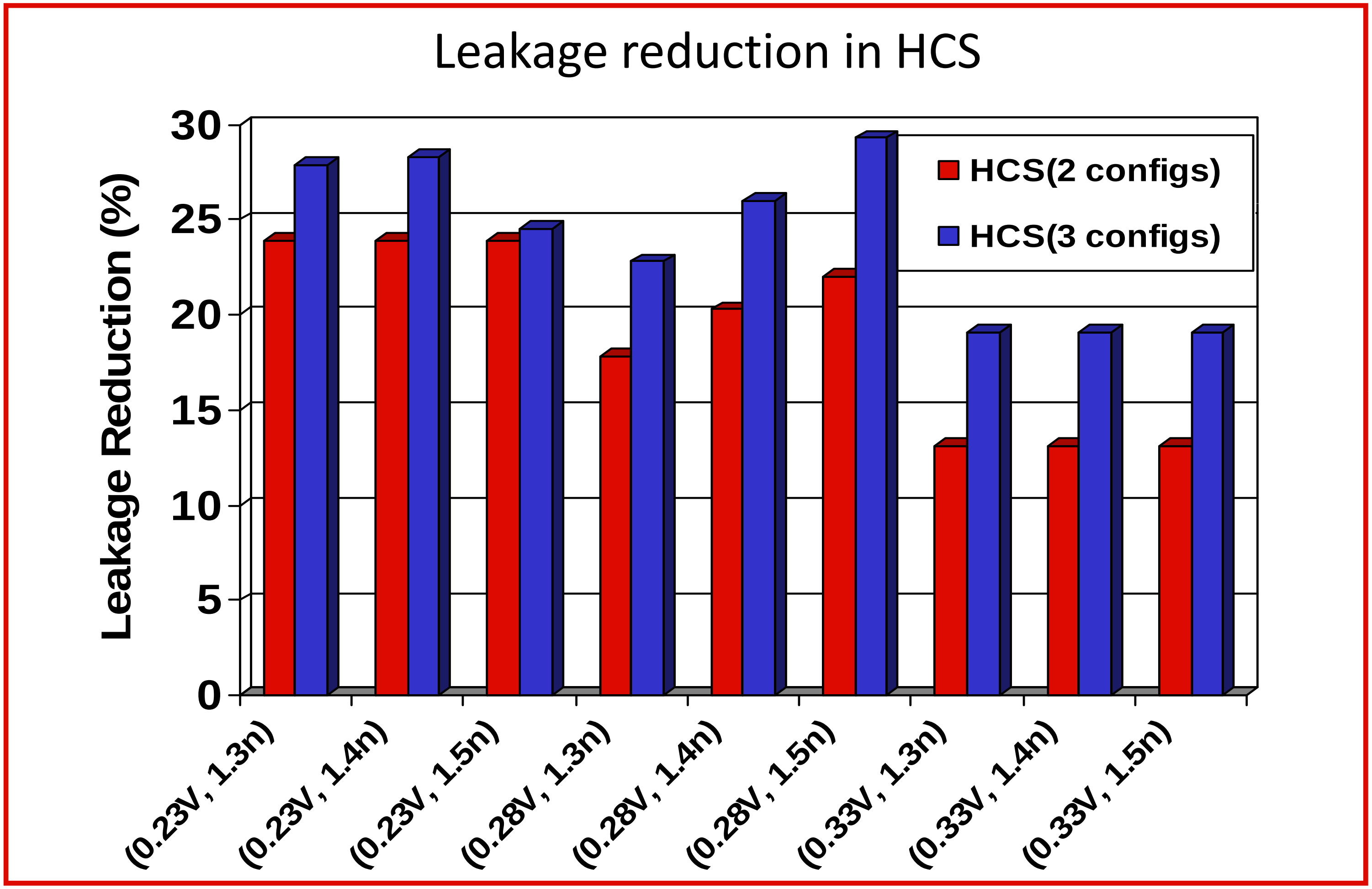


- If only dual Vth option is available
  - Leakage saving still significant

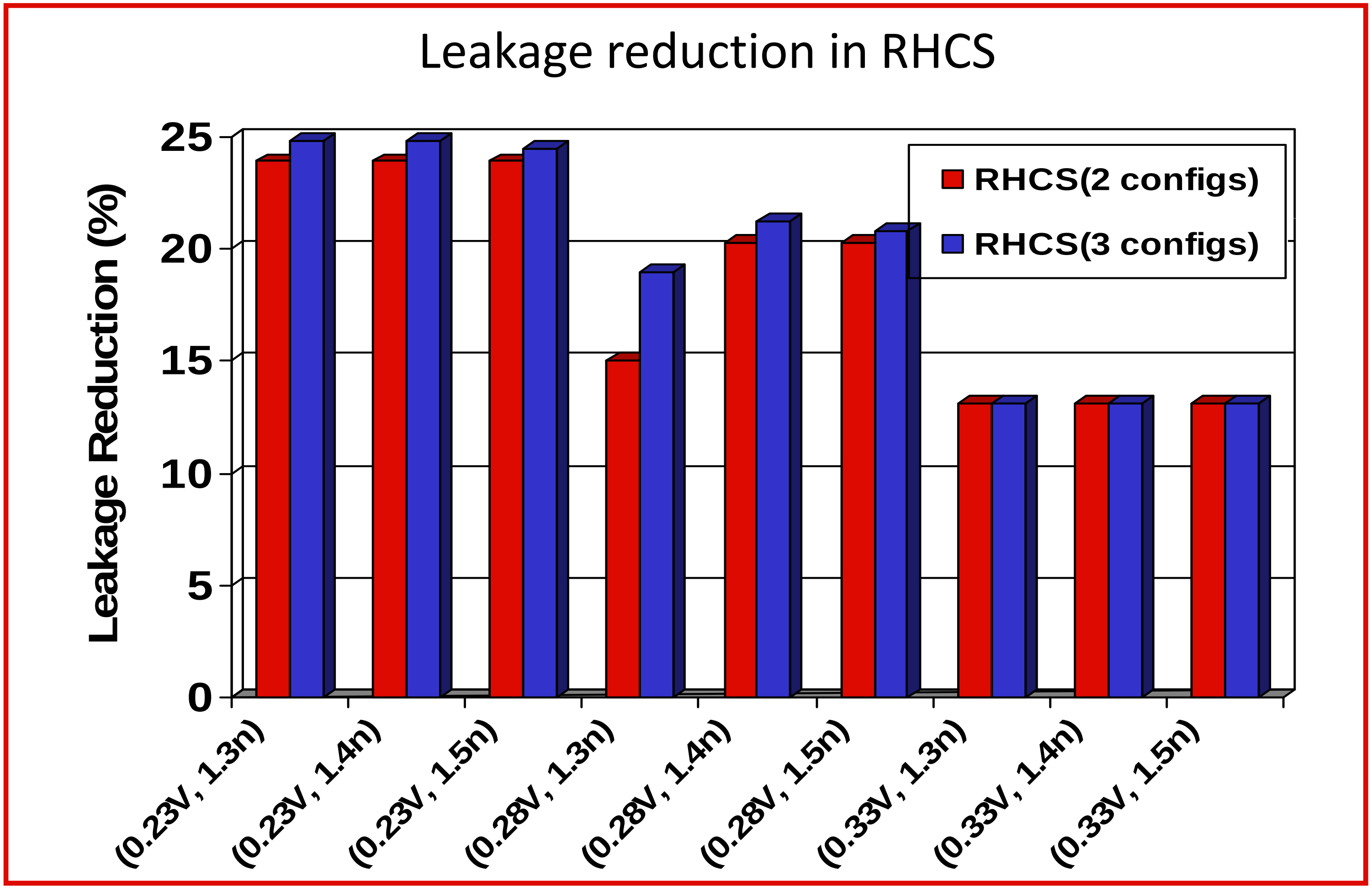


Effect of the Configuration Count

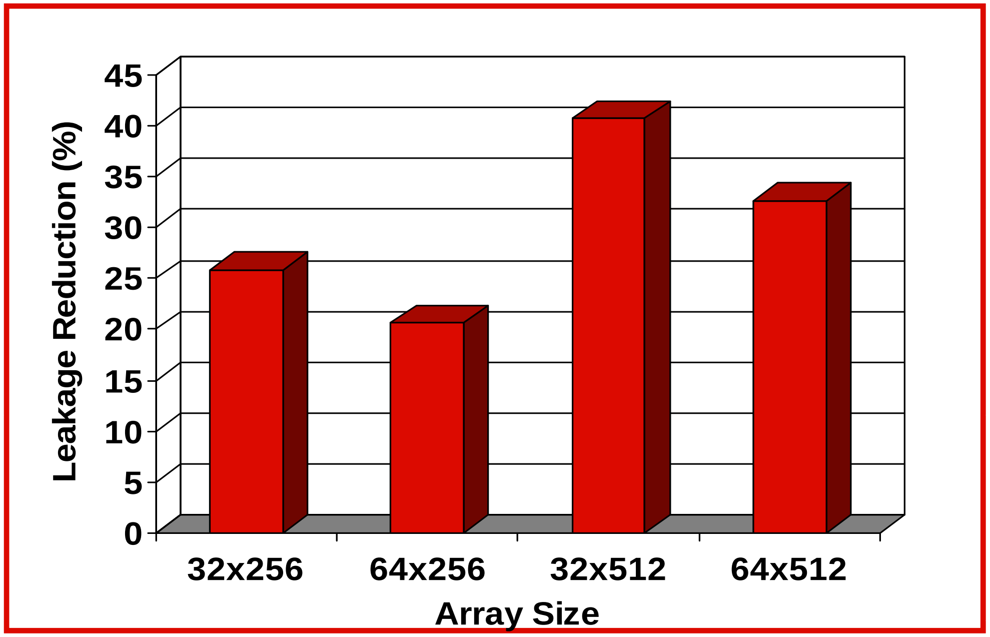
- Leakage reduction when the number of configurations is limited to two or three



Effect of the Configuration Count (Cont'd)



Effect of the Array Size



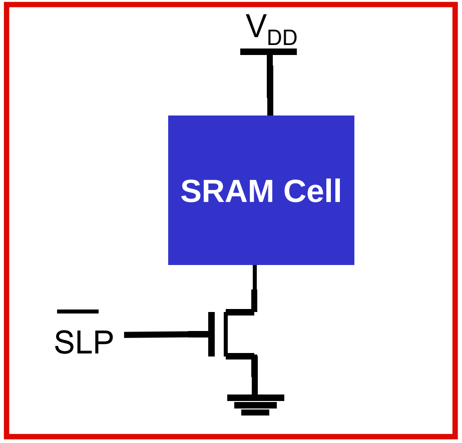
- Leakage saving is reduced with increasing number of rows
  - Increasing number of rows makes the bitline more capacitive
    - Delay overhead of low-leakage configurations becomes high
    - Fewer cells can be swapped with low-leakage configurations
- Notice that with newer CMOS technologies, cell arrays are moving from tall to wide structures

Outline

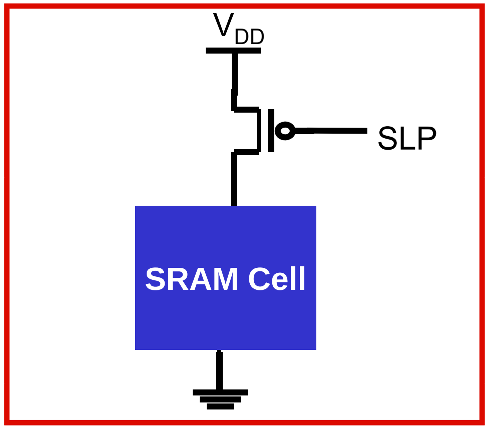
- ❑ Introduction
- ❑ Related prior work
- ❑ Heterogeneous cell SRAM
- ❑ PG-gated SRAM cell
- ❑ Concluding remarks

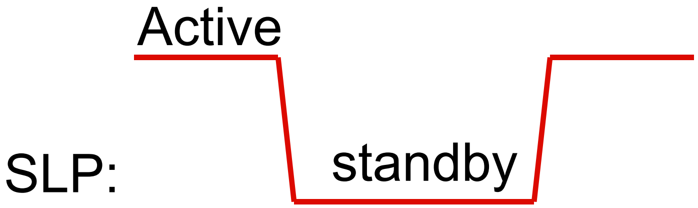
Original G-Gated and P-Gated SRAM Cells

Original G-Gated SRAM Cell



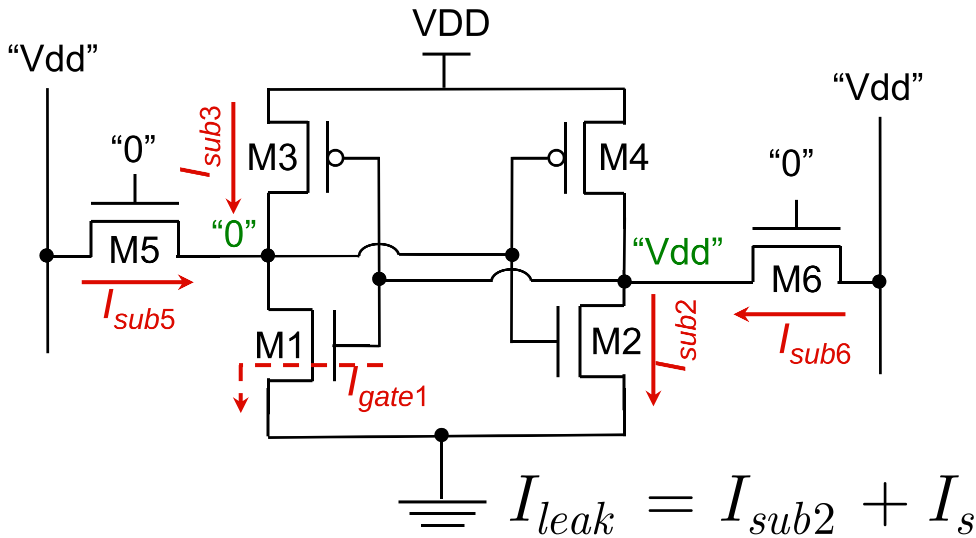
Original P-Gated SRAM Cell





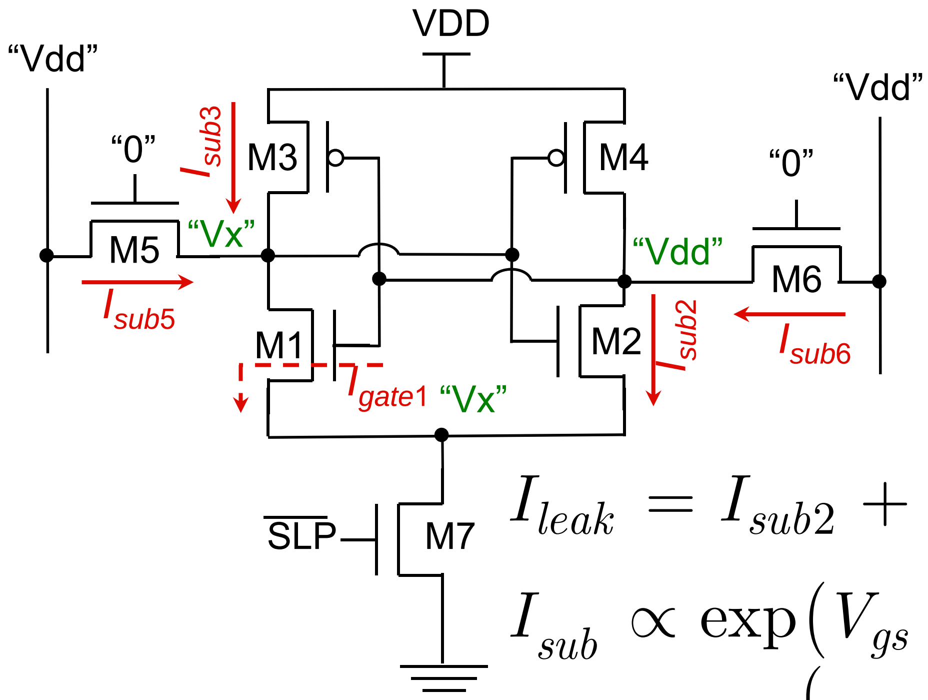
- In standby mode leakage is exponentially reduced
- G-gated technique is more effective than P-gated technique
- G-gated technique increases read delay

Leakage Components: Original SRAM Cell



$$I_{leak} = I_{sub2} + I_{sub3} + I_{sub5} + I_{sub6} + I_{gate1}$$

Leakage Components: G-Gated SRAM Cell



$$I_{leak} = I_{sub2} + I_{sub3} + I_{sub5} + I_{sub6} + I_{gate1}$$

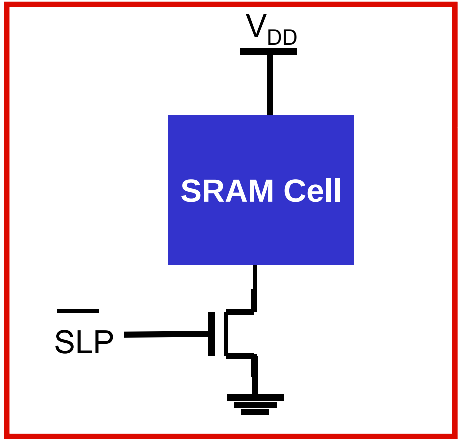
$$I_{sub} \propto \exp\left(V_{gs} - V_{t0} - \gamma f(V_{sb}) + \eta V_{ds}\right)$$

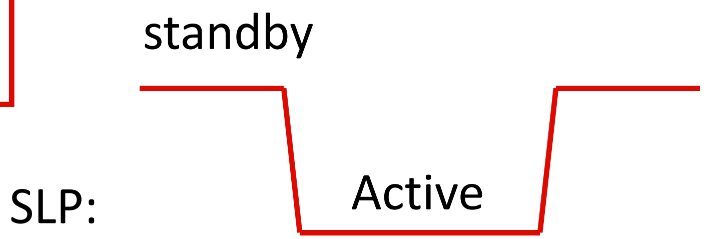
$$I_{gate} \propto \exp\left(-B\frac{t_{ox}}{V_{ox}}\right)$$

- I_{sub5}↓↓↓ (stacking effect, body biasing, DIBL)
- I_{sub2}↓↓ (Body biasing, DIBL)
- I_{sub3}↓ (DIBL)
- I_{gate1}↓ (lower Vox)

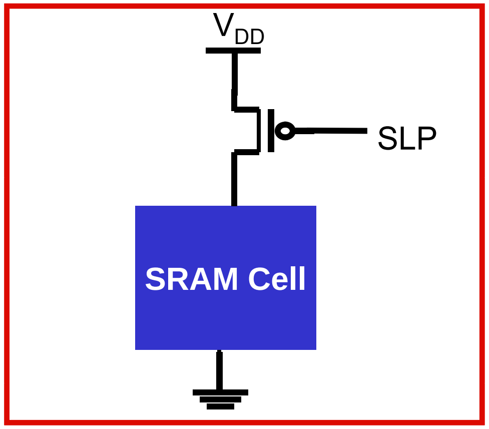
Original G-Gated and P-Gated SRAM Cells

Original G-Gated SRAM Cell



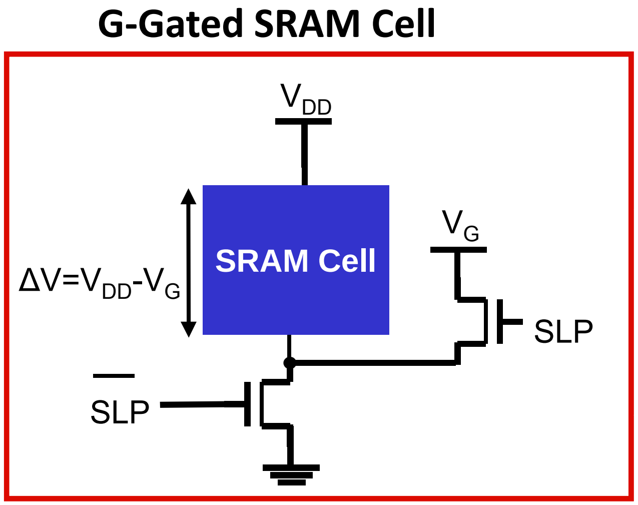


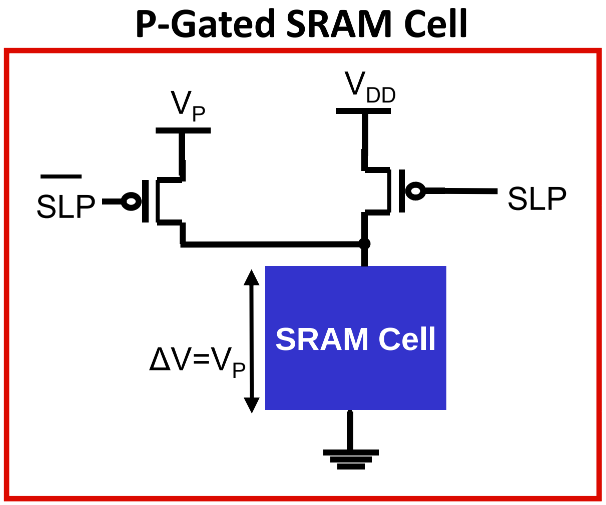
Original P-Gated SRAM Cell



- Drawback: virtual ground (supply) node may charge (discharge) to V_{DD} (0)
  - The stored bit may be destroyed
- Solution: In the standby mode, strap the virtual ground or virtual supply node to a fixed voltage
  - Data retention capability

G-Gated and P-Gated SRAM Cells

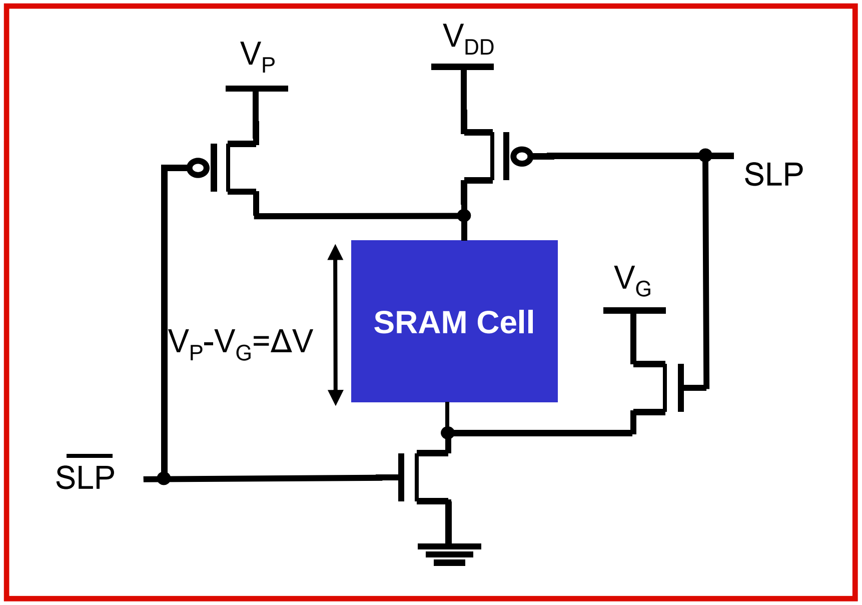




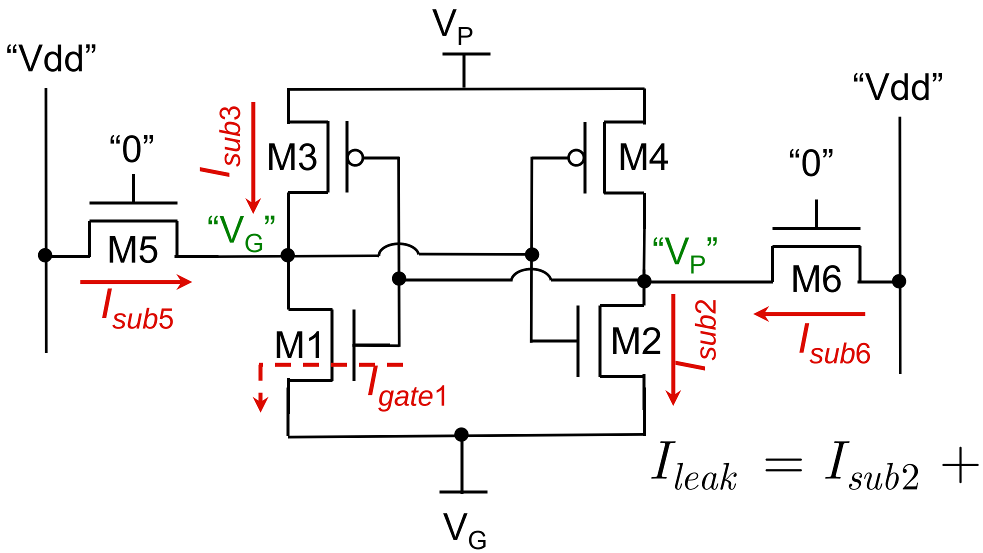
- For both G-gated and P-gated cells,
  - Standby leakage decreases with ΔV
  - Hold SNM decreases with ΔV
- For a fixed ΔV, G-gated technique is more effective than P-gated technique. However, G-gated technique increases the read delay.
- Because sleep transistor does not affect read delay in P-gated technique, one can use a smaller sleep transistor.

PG-Gated SRAM Cell

- The key idea is to use two sleep transistors; one as a header, the other as a footer.
- In standby mode, strap virtual ground and virtual supply node to properly selected voltage levels, i.e., V_G and V_P.

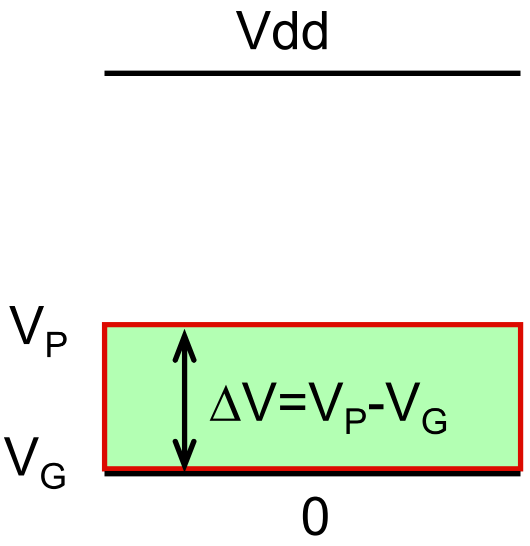


Leakage of PG-Gated SRAM Cell

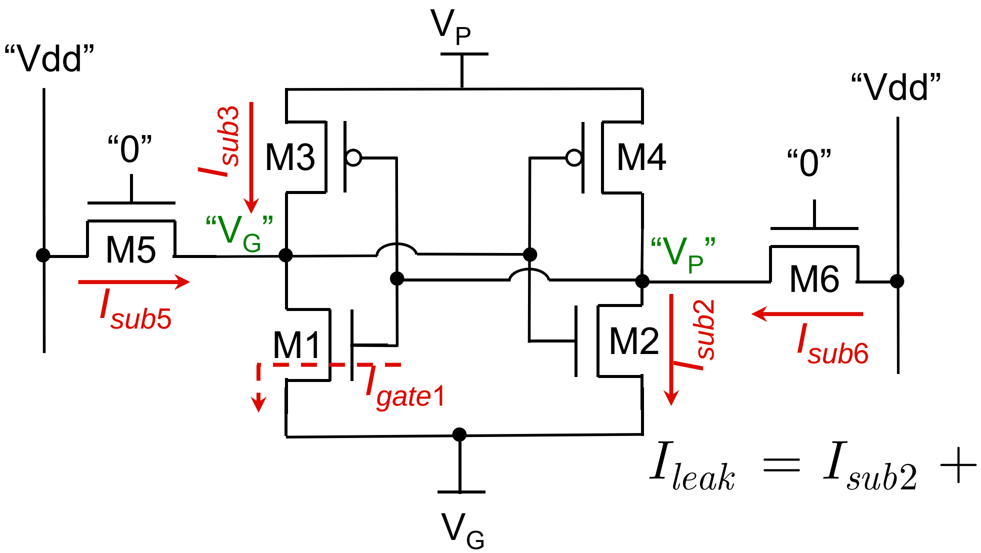


$$I_{leak} = I_{sub2} + I_{sub3} + I_{sub5} + I_{sub6} + I_{ox1}$$

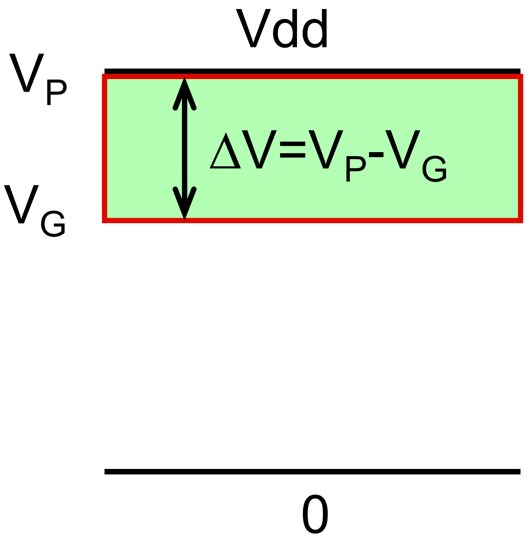
ΔV=fixed, $V_P\downarrow$, $V_G \downarrow \rightarrow I_{sub2}\uparrow$, $I_{sub5}\uparrow$, $I_{sub6}\uparrow$, $I_{sub3} \downarrow$



Leakage of PG-Gated SRAM Cell

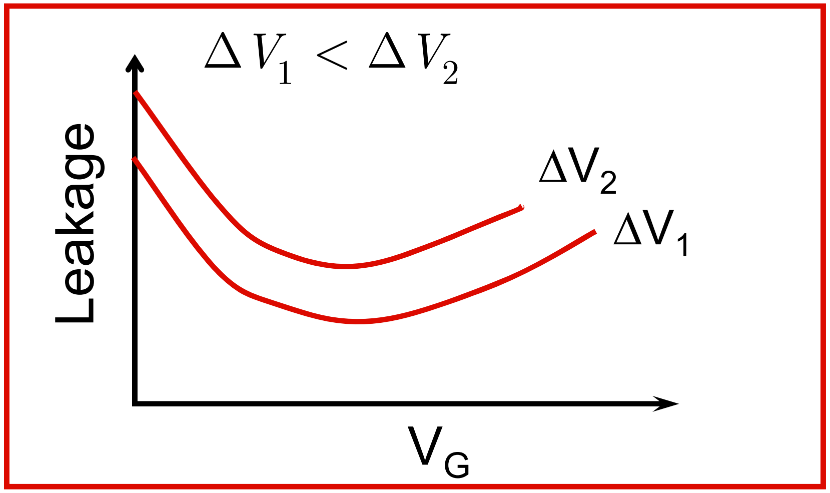


$$I_{leak} = I_{sub2} + I_{sub3} + I_{sub5} + I_{sub6} + I_{ox1}$$

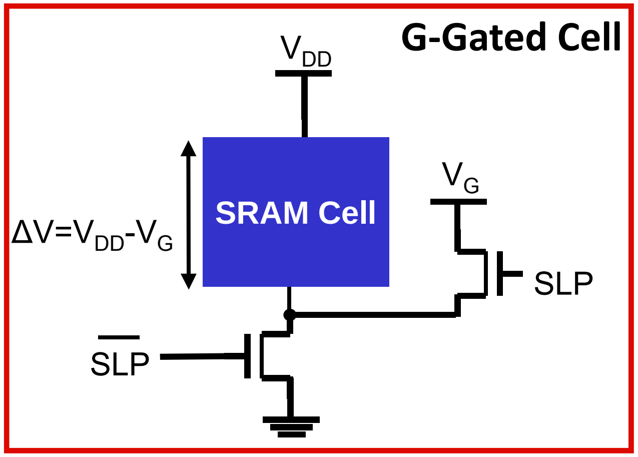


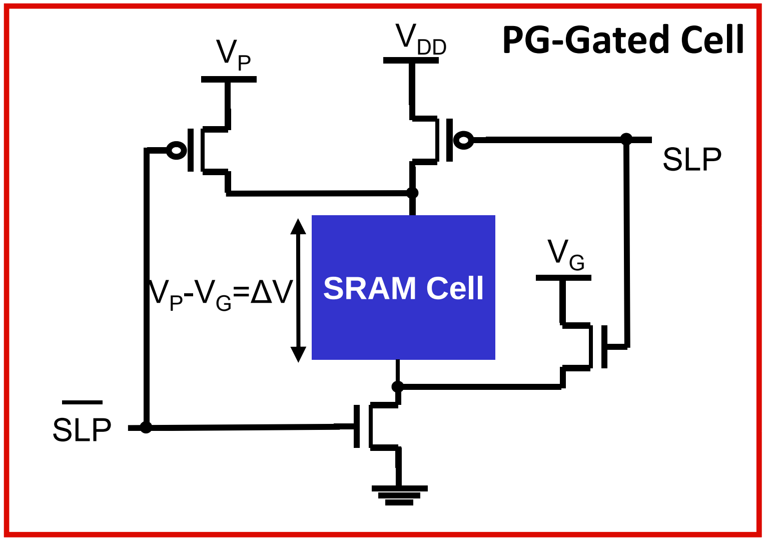
ΔV=fixed, $V_P \downarrow$, $V_G \downarrow \rightarrow I_{sub2} \uparrow$, $I_{sub5} \uparrow$, $I_{sub6} \uparrow$, $I_{sub3} \downarrow$

ΔV=fixed, $V_P \uparrow$, $V_G \uparrow \rightarrow I_{sub2} \downarrow$, $I_{sub5} \downarrow$, $I_{sub6} \downarrow$, $I_{sub3} \uparrow$

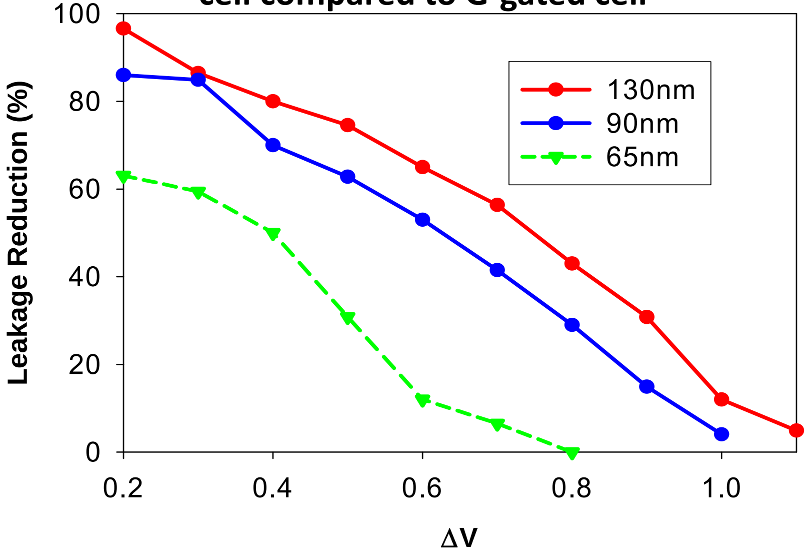


PG-Gated and G-Gated Cell Leakage

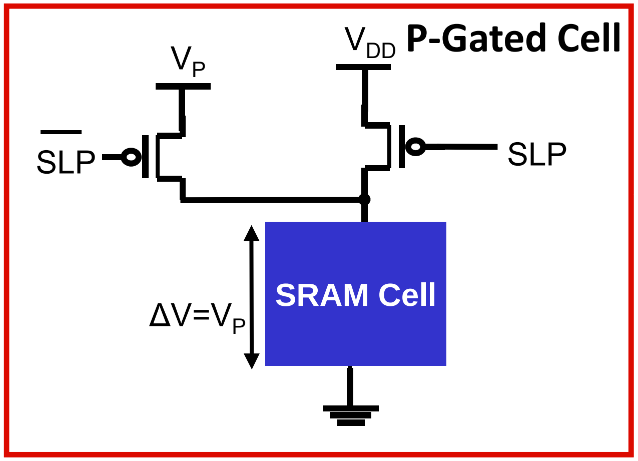

G-Gated Cell


PG-Gated Cell

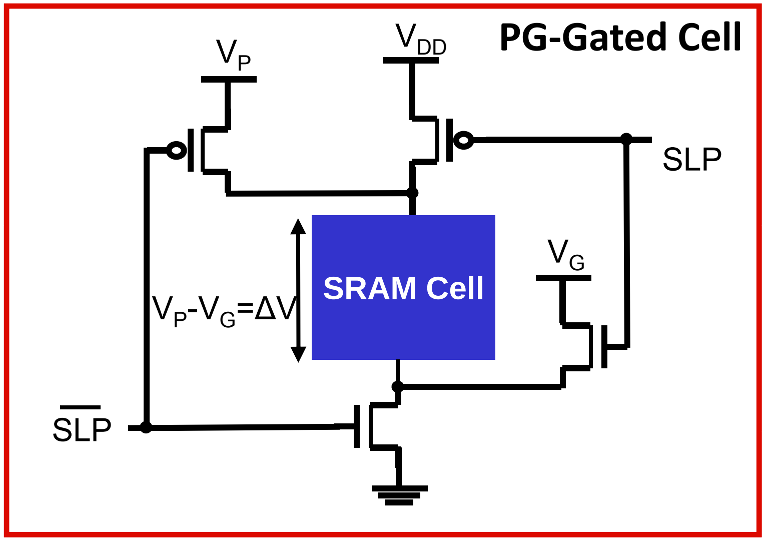
Cell leakage current reduction of PG-gated cell compared to G-gated cell



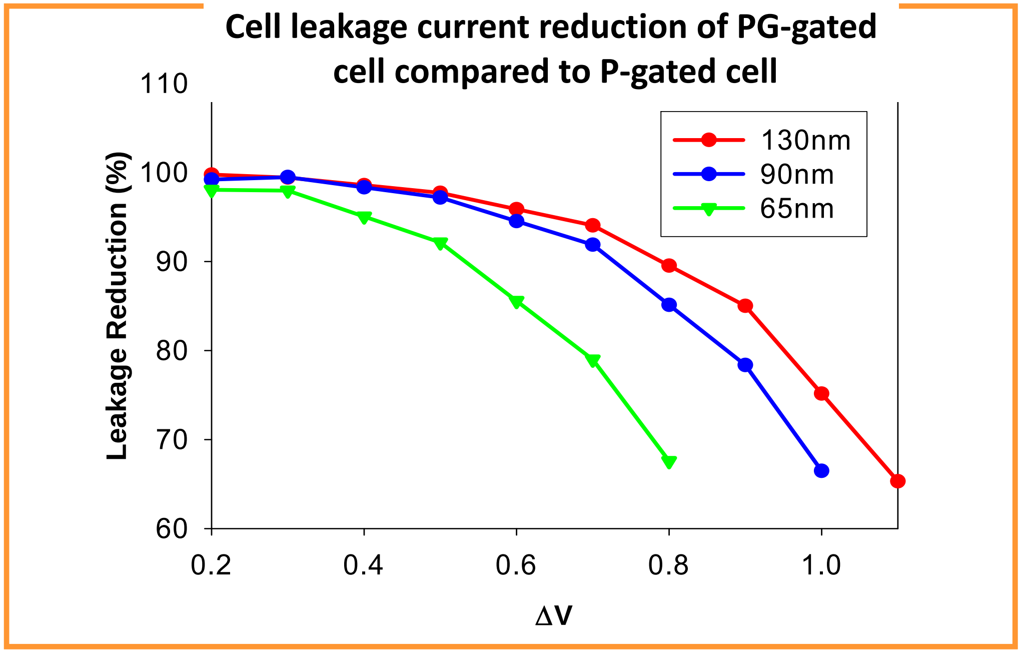
PG-Gated and P-Gated Cell Leakage



P-Gated Cell

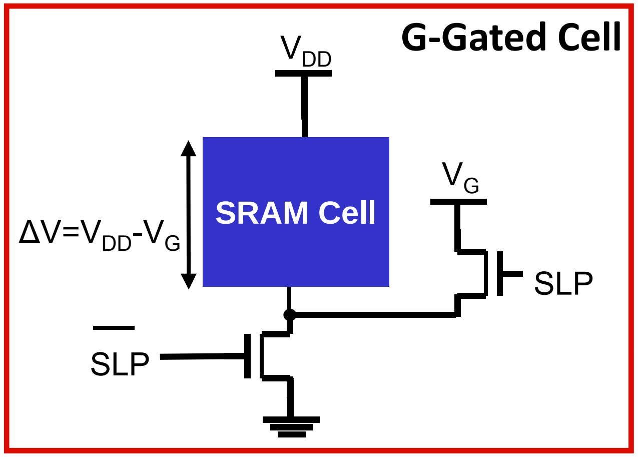


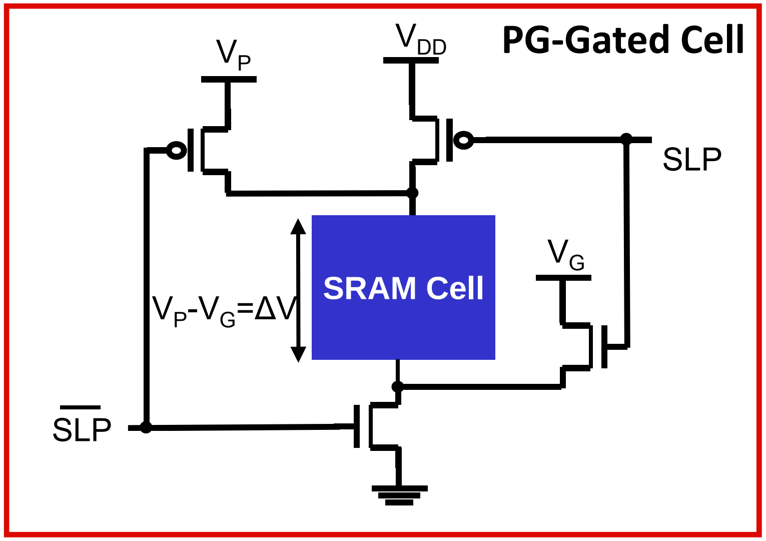
PG-Gated Cell



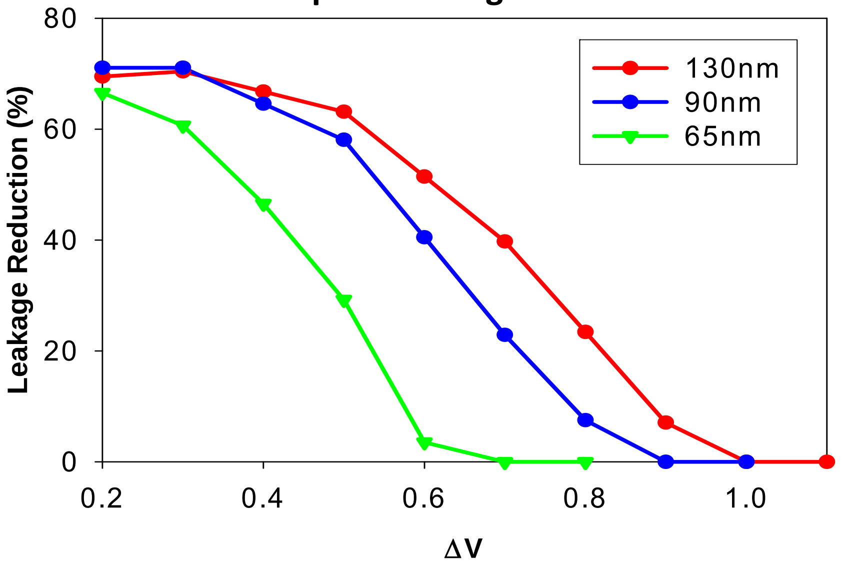
Cell leakage current reduction of PG-gated cell compared to P-gated cell

PG-Gated and G-Gated Total Leakage

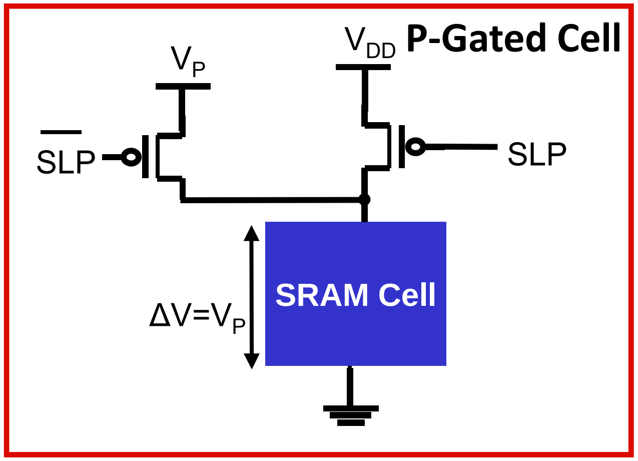


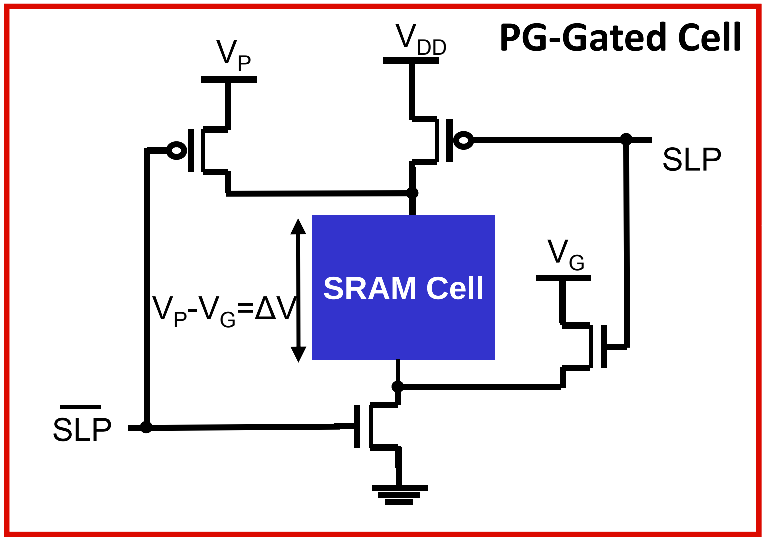


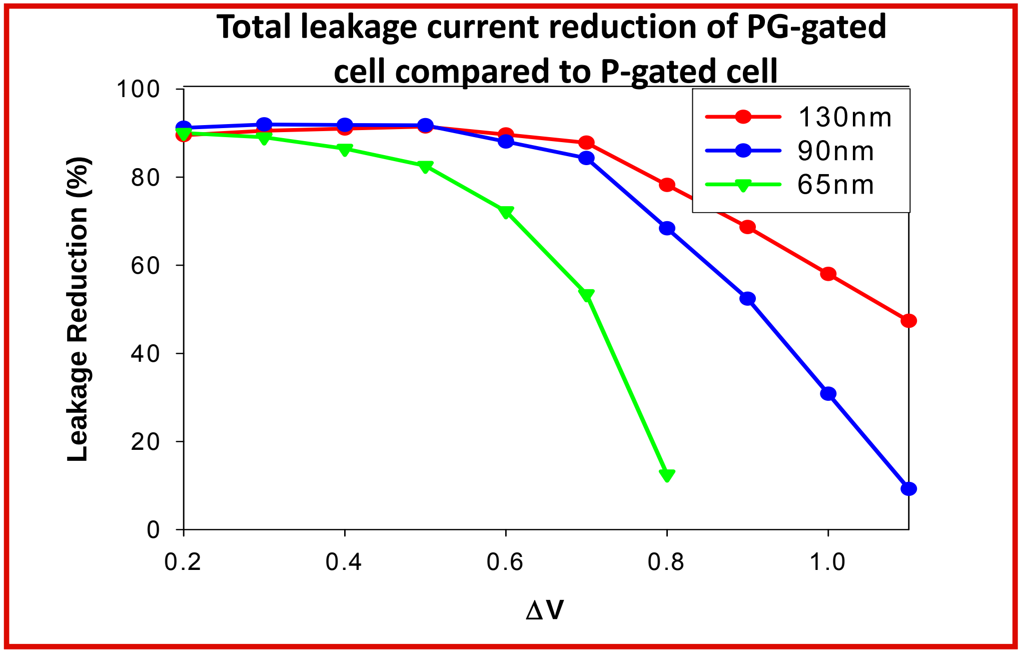
Total leakage current reduction of PG-gated cell compared to G-gated cell



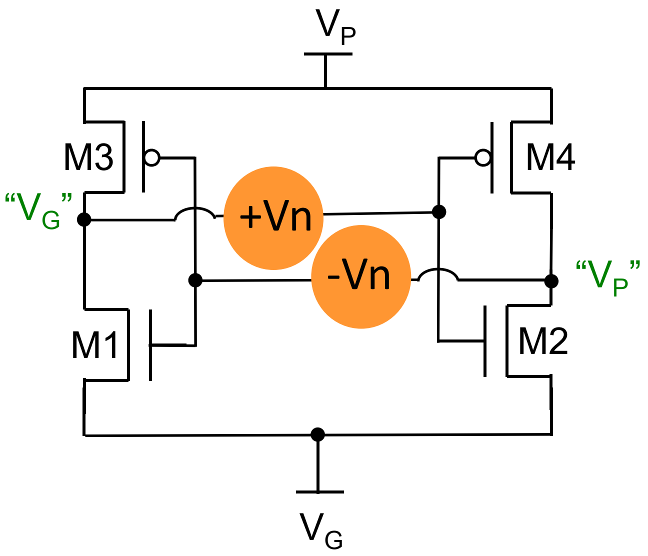
PG-Gated versus P-Gated Total Leakage







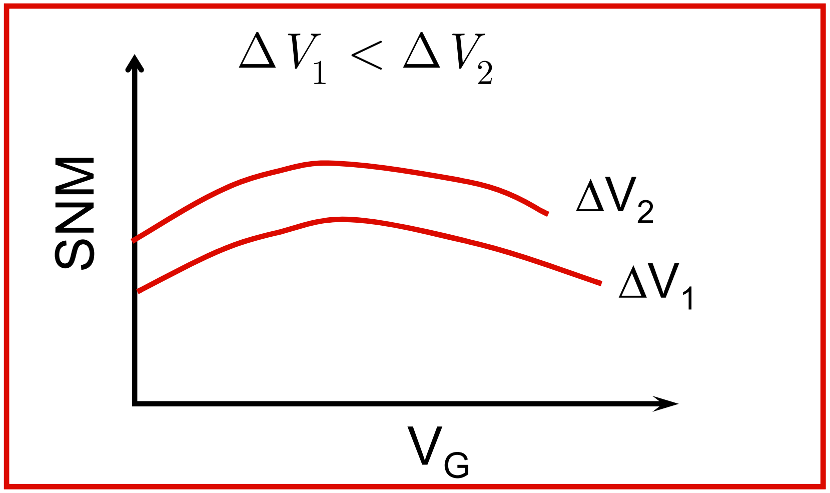
SNM of PG-Gated SRAM Cell



SNM is a monotone increasing function of threshold voltages

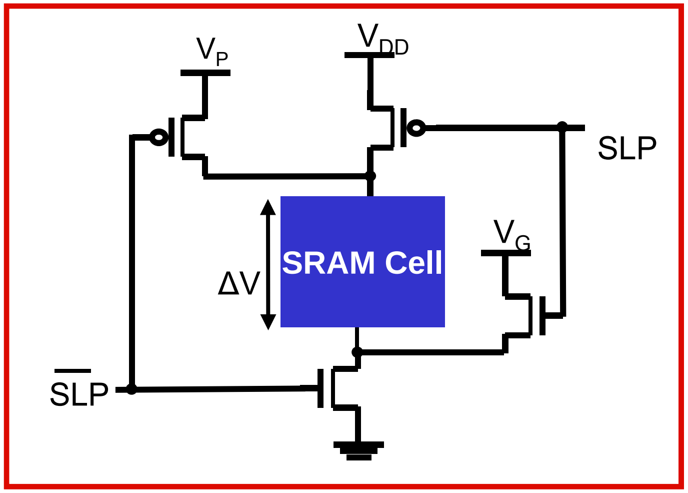
ΔV=fixed, $V_P \uparrow$, $V_G \uparrow \rightarrow V_{th1} \uparrow$, $V_{th4} \downarrow$

ΔV=fixed, $V_P \downarrow$, $V_G \downarrow \rightarrow V_{th1} \downarrow$, $V_{th4} \uparrow$

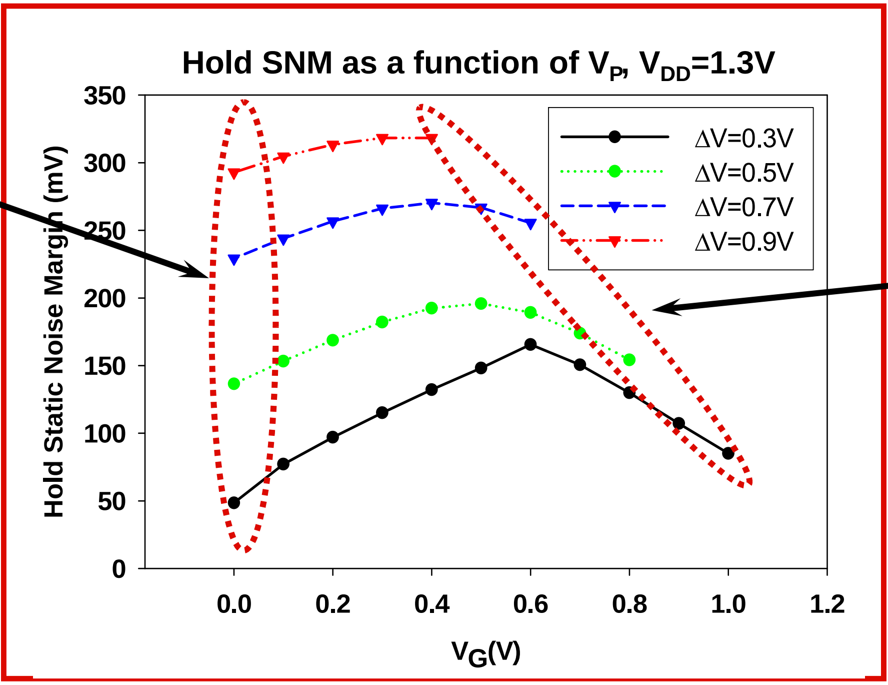


Static Noise Margin

- In a PG-gated cell with a fixed ΔV
  - if $V_P=V_{DD}$ then no PMOS sleep TX i.e., PG-gated = G-gated
  - if $V_G=0$, i.e., $V_P=\Delta V$ then no NMOS sleep TX i.e., PG-gated = P-gated



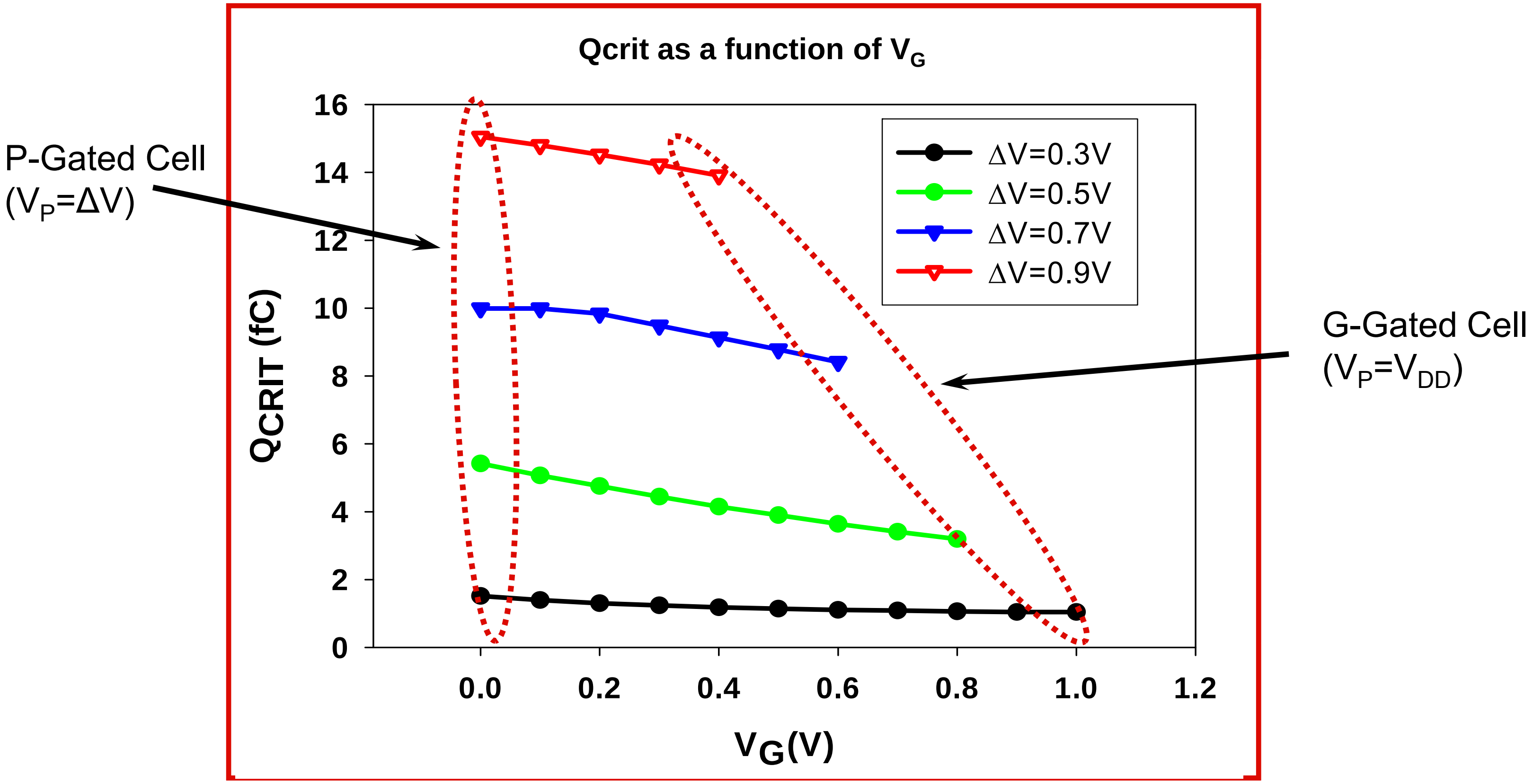
P-Gated Cell (V_G=0)



G-Gated Cell (V_P= ΔV)

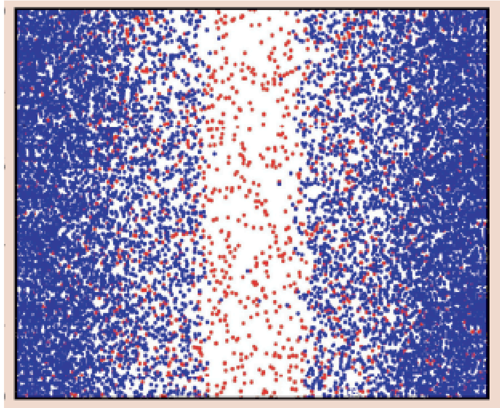
Soft Error

- Virtual ground and virtual supply nodes are shared among some cells in a row
  - These nodes are highly capacitive and soft error immune
  - SER is mainly determined by internal nodes of SRAM cells



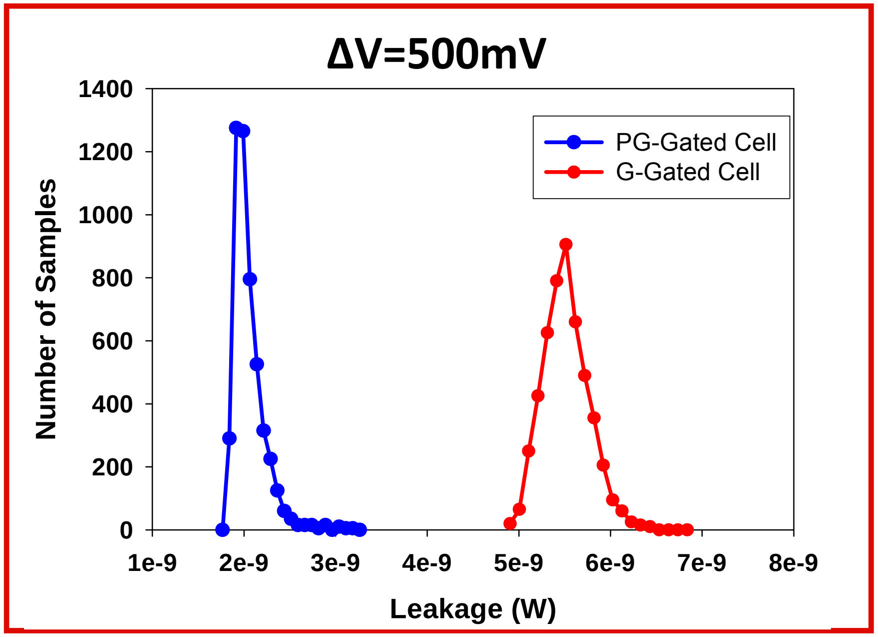
Process Variations

- Major source of variation in SRAM cells: Vth variation due to random dopant fluctuation (RDF)
  - Vth's modeled as an independent Gaussian RV ~N(0,σ)



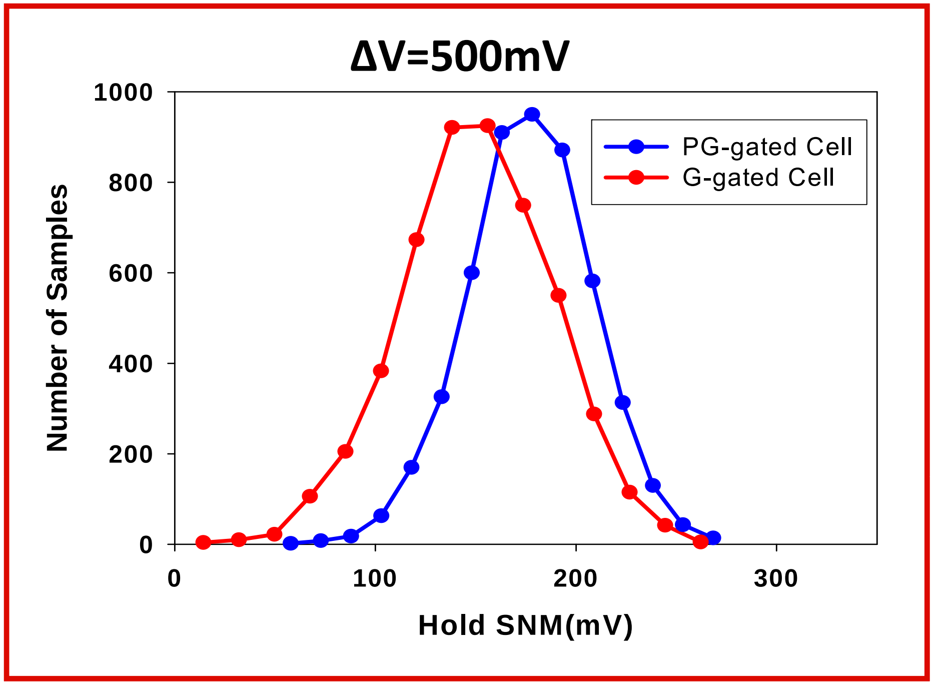
$$\sigma = \sigma_{\min}\sqrt{\frac{W_{\min}L_{\min}}{WL}}$$

- PG-gated cell reduces both the mean and variance of SRAM leakage current



Process Variations

- Effect of process variation on hold SNM



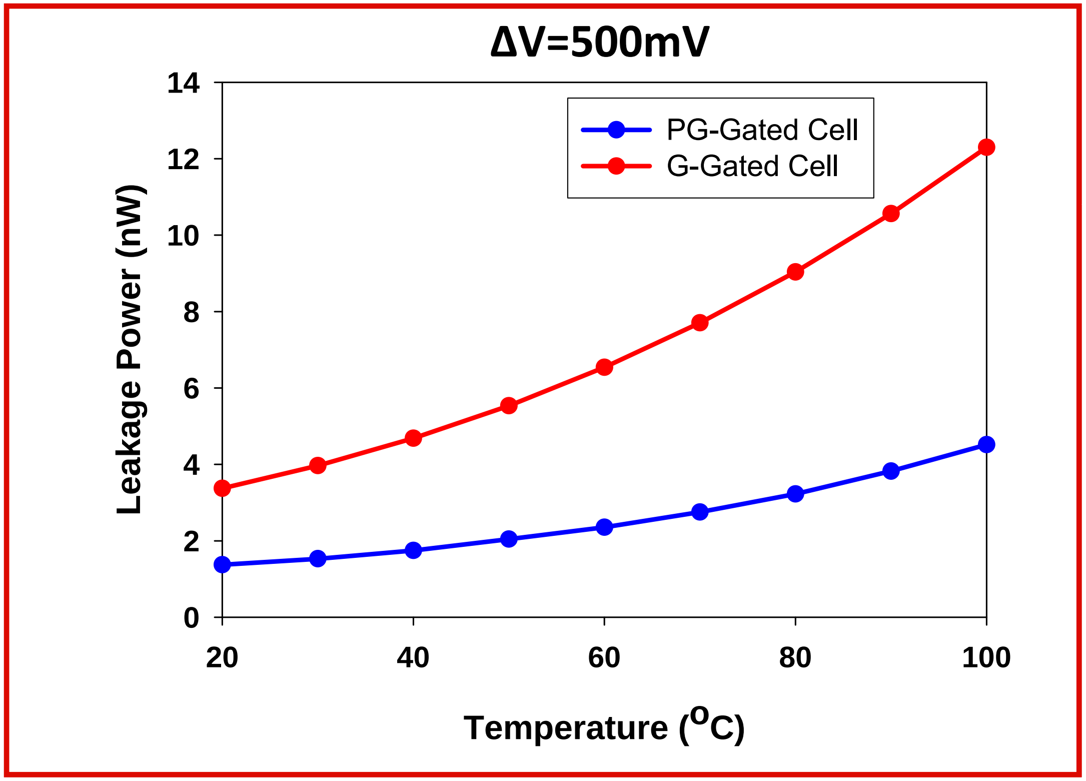
- PG-gated cell is more robust than the G-gated cell
  - Lower probability of hold failures

Temperature Effect

- Effect of temperature on SRAM leakage

$$I_{sub} \propto \exp\left(\frac{q}{n'kT}\left(V_{gs} - V_{t0} - \gamma f(V_{sb}) + \eta V_{ds}\right)\right)$$

$$I_{ox} \propto \exp\left(-B\frac{t_{ox}}{V_{ox}}\right)$$



G-Gated and PG-Gated Cell Comparison

- Three 64Kb SRAM designed in 130nm technology:
  - Conventional SRAM cell
  - A data retention G-gated SRAM cell
  - A data retention PG-gated SRAM cell
  - ΔV=500mV to have hold SNM conv. SRAM≈150mV

	G-Gated SRAM	PG-Gated SRAM	Improvement
Area (Normalized)	1.035	1.074	-3.8%
Delay (Normalized)	1.027	1.032	-0.5%
Read SNM	185mV	186mV	0.5%
Hold SNM	154mV	182mV	18.2%
Leakage (mean)	5.57nW	2.1nW	62.3%
Leakage (std. dev.)	0.25nW	0.17nW	32.0%

- With very small delay and area overhead, PG-gated technique results in a more robust and power-efficient SRAM design.

Summary

- Heterogeneous Cell SRAM
  - Useful for runtime leakage power reduction
  - Key idea: Read and write delays of a memory cell depend on the physical location of the cell
  - Has no delay or hardware overhead
  - Has ability to improve SNM under process variations
- PG-gated SRAM
  - Useful for standby leakage power reduction
  - Key idea: using two sleep transistors is more beneficial
  - Improves not only leakage, but also SNM and SER
  - Results in less leakage variation in presence of process and environmental variations

Publications: SRAM Design

❑ Patents

1. F. Fallah, B. Amelifard, and M. Pedram, "PG-gated data retention technique for reducing leakage in memory cells," pending.
2. F. Fallah, B. Amelifard, and M. Pedram, "Setting one or more delays of one or more cells in a memory block to improve one or more characteristics of the memory block," pending.

❑ Journal Paper

3. B. Amelifard, F. Fallah, and M. Pedram, "Leakage minimization of SRAM cells in a dual-Vt and dual-Tox technology," IEEE Transactions on Very Large Scale Integration (VLSI) Systems.

❑ Conference Papers

4. B. Amelifard, F. Fallah, and M. Pedram, "Reducing the sub-threshold and gate-tunneling leakage of SRAM cells using dual-Vt and dual-Tox assignment," Design Automation and Test in Europe (DATE).
5. B. Amelifard, F. Fallah, and M. Pedram, "Low-leakage SRAM design with dual Vt transistors," International Symposium on Quality Electronic Design (ISQED).
6. B. Amelifard, F. Fallah, and M. Pedram, "Robust low leakage SRAM design using a PG-Gated data retention technique, " submitted to International Symposium on Quality Electronic Design (ISQED).