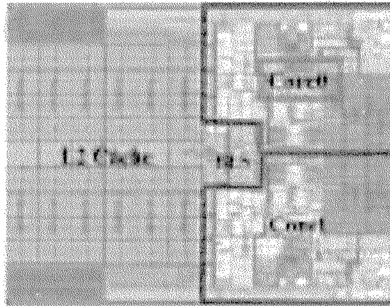
Ultra-Low-Power SRAM Design In High Variability Advanced CMOS

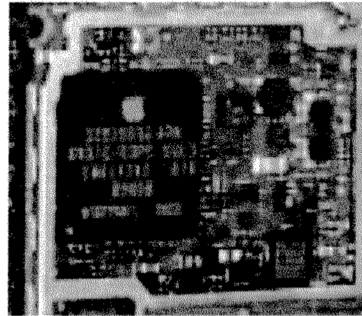
Prof. Pinaki Mazumder
University of Michigan
Ann Arbor, MI 48109
mazum@eecs.umich.edu

- Mobile computing



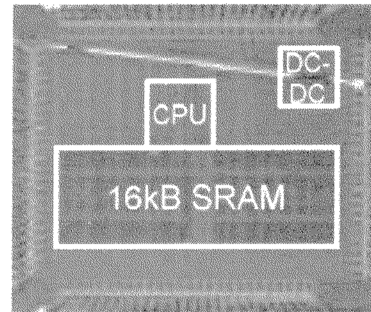
Intel Core 2 (Penryn)
6MB SRAM L2, 64kB RF L1
Power: 20%

- Embedded/handheld



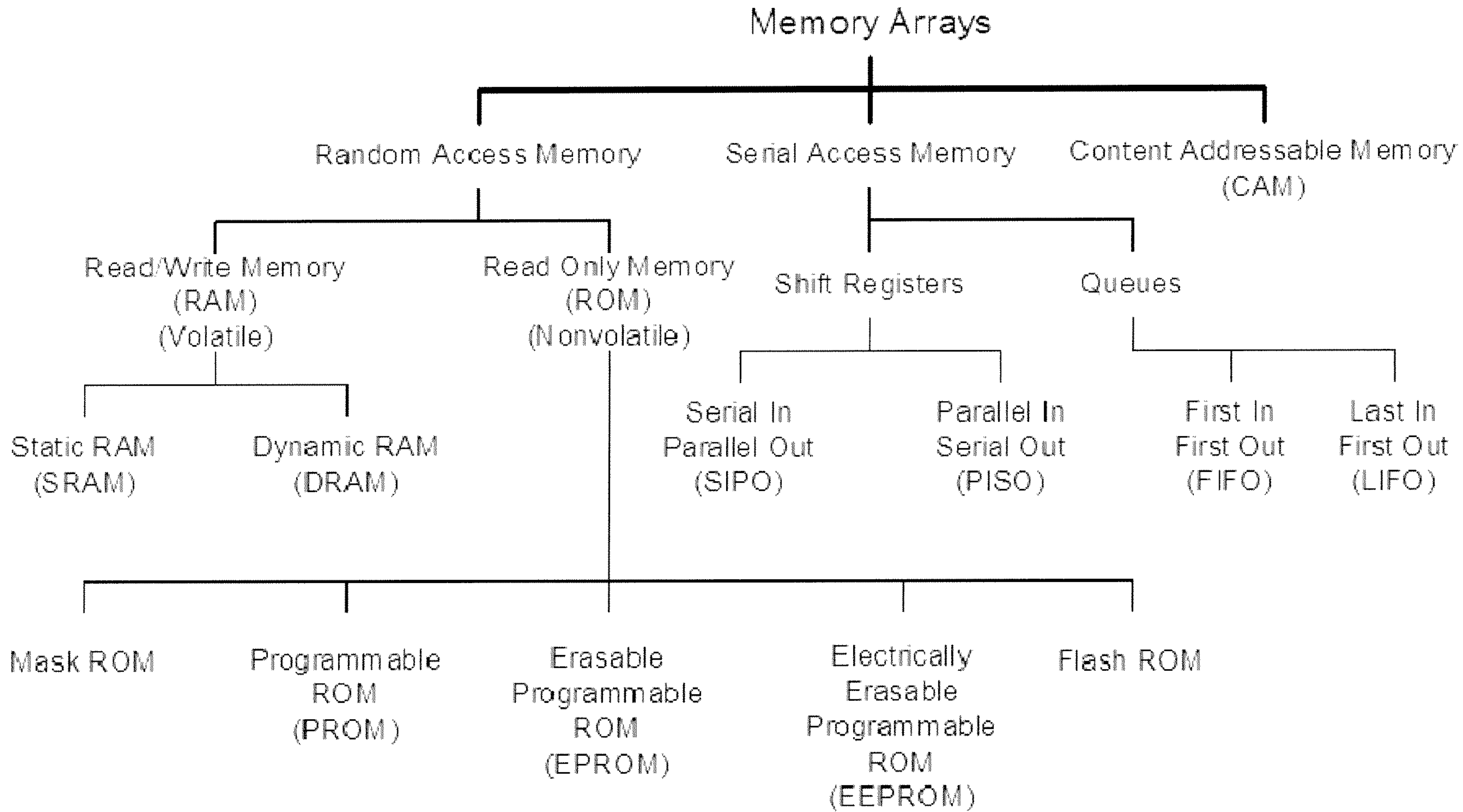
ARM1176JZ
16kB SRAM cache
Power: 39%

- Implantable, sensor nets



Custom MSP430
16kB SRAM cache
Power: 69%

Taxonomy of Semiconductor Memories



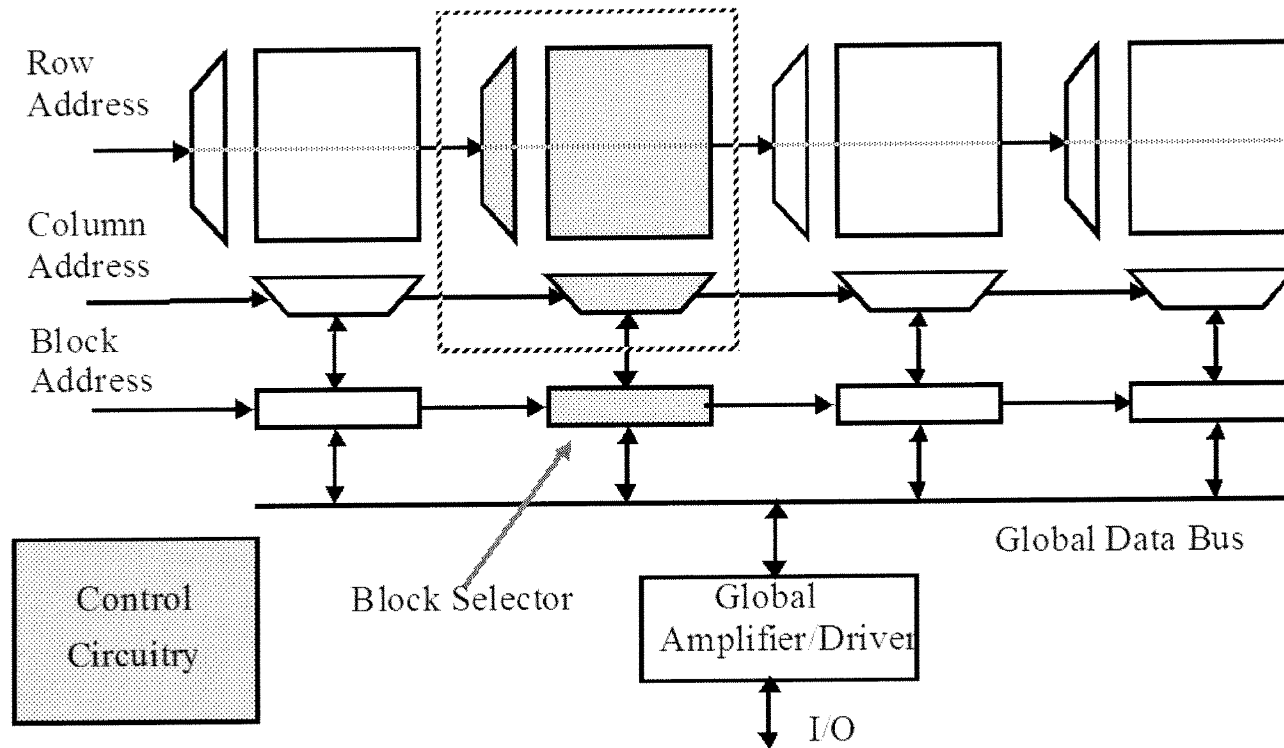
Courtesy: Harris & Weste

Comparison of Semiconductor Memories

Memory Type	SRAM	DRAM	Flash
Speed	Very fast	Fast	Very slow
Density	Low	High	Very high
Endurance	Better	Better	Poor
Power	Low	High	Very low
Refresh	No	Yes	No
Retention	Volatile	Volatile	Non-volatile
Scalable	Good	Bad	Good
Mechanism	Bi-stable latch	Capacitor	FN tunneling, HCI

Courtesy: Harris & Weste

Architecture of Static RAM

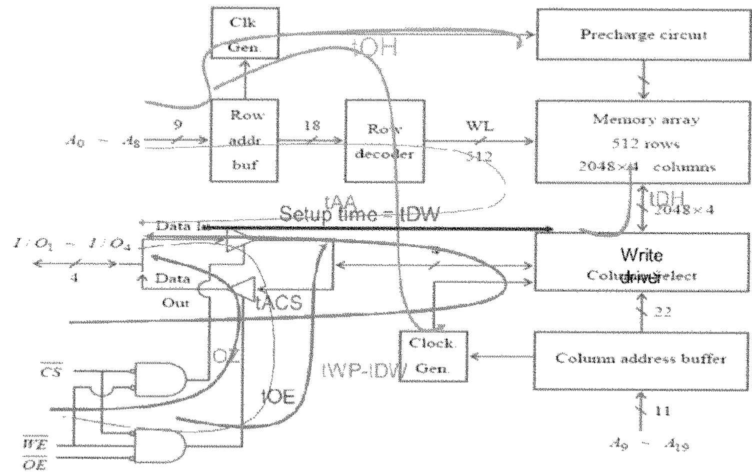


Advantages:
1. Shorter wires within blocks
2. Block address activates only 1 block => power savings

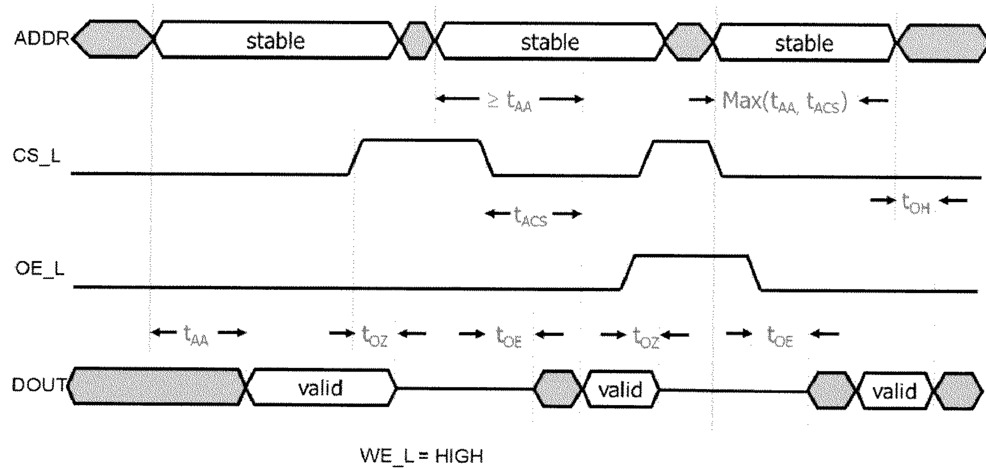
Courtesy: Harris & Weste

READ and WRITE Operations in SRAM

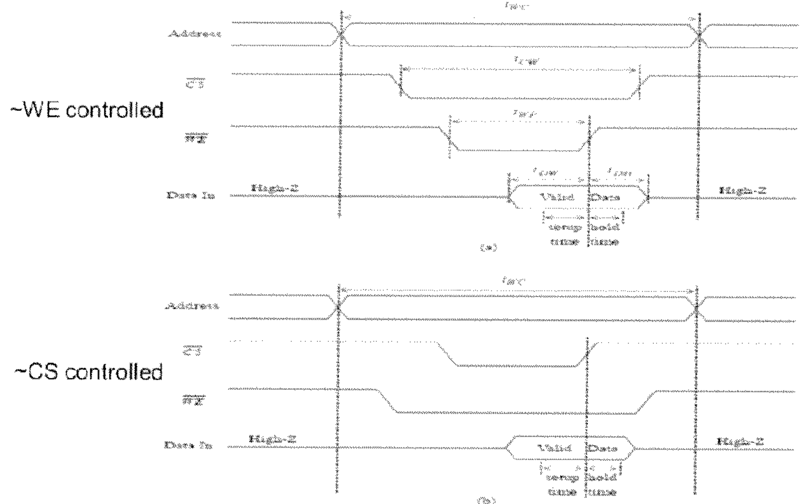
- t_{AA} (access time for address): time for stable output after a change in address.
- t_{ACS} (access time for chip select): time for stable output after CS is asserted.
- t_{OE} (output enable time): time for low impedance when OE and CS are both asserted.
- t_{OZ} (output-disable time): time to high-impedance state when OE or CS are negated.
- t_{OH} (output-hold time): time data remains valid after a change to the address inputs.



Read Memory Cycle

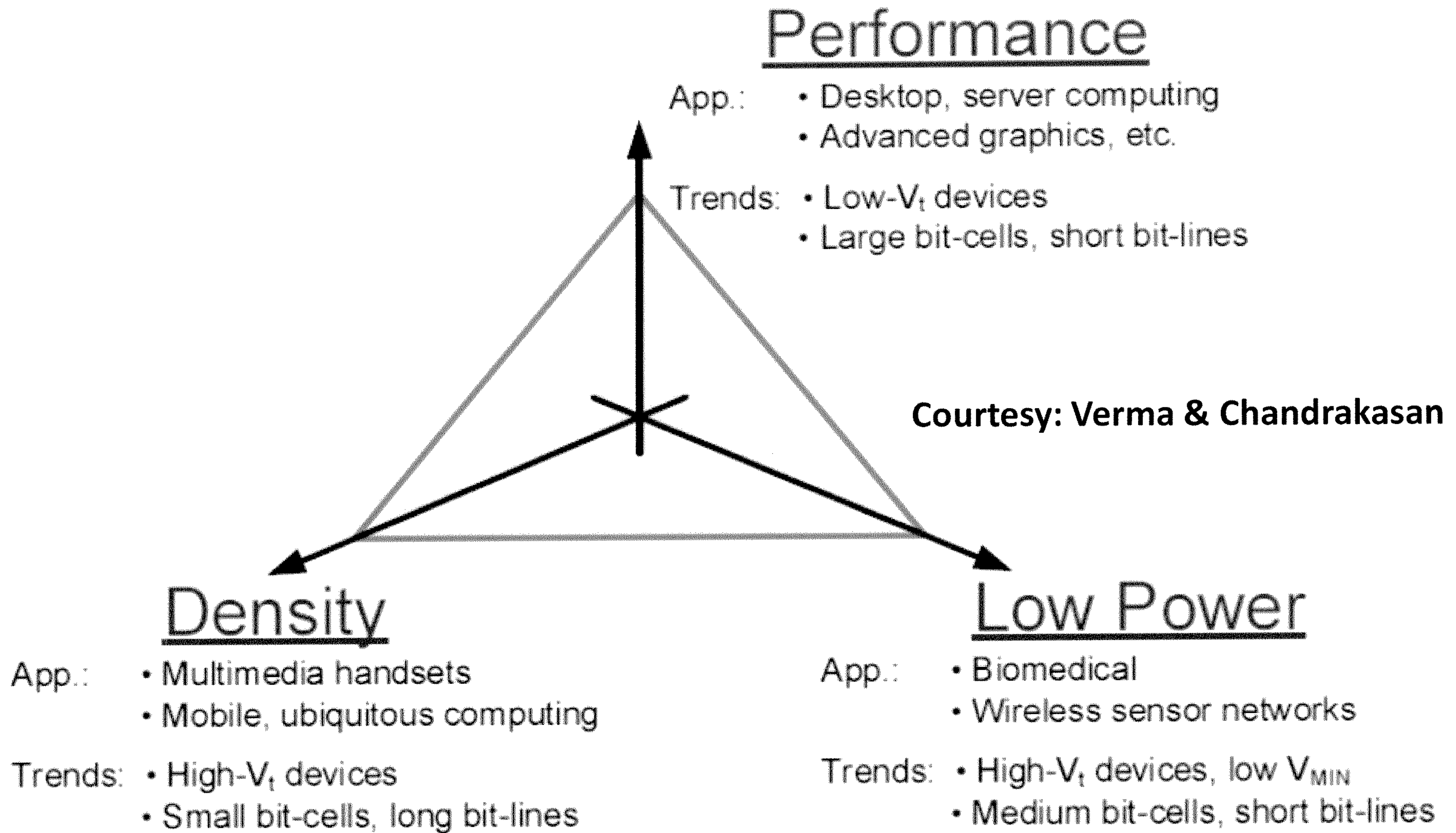


Write Memory Cycle



Courtesy: Harris & Weste

SRAM Trade-offs



Courtesy: Verma & Chandrakasan

IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 44, NO. 1, JANUARY 2009 163, A High-Density 45 nm SRAM Using Small-Signal Non-Strobed Regenerative Sensing , Naveen Verma, Student Member, IEEE, and Anantha P. Chandrakasan, Fellow, IEEE

Key existing and emerging applications for biomedical devices

Application	Performance Specification		
	Power	Processor	Energy Source
Pacemaker & Cardioverter-defibrillator [15][16]	$<10\ \mu W$	1kHz DSP	10-year life-time battery
Hearing aid & Cochlear implant [17][18][19]	$100\text{-}2000\ \mu W$	32kHz-1MHz DSP	1-week lifetime battery
Neural recording [20][21]	1-10 mW	n/a	Inductive power
Body-area monitoring [22]	$140\ \mu W$	$<$10MHz DSP	Battery

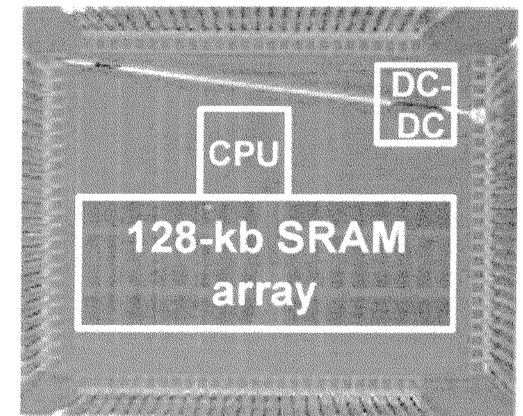
Implantable devices (pacemakers/defibrillators, cochlear implants, neural sensors/stimulators are energy-constrained since battery replacement requires surgical intervention.
Wearable devices (hearing aids, body-area sensors) have less stringent energy-constraints which Are set by battery weight limitations.

Courtesy: Verma & Chandrakasan

Wireless Sensor Networks

Micro/nano-scale devices providing sensing, processing, and communications capabilities can form networks, broadly referred to as wireless sensor networks. The applications for such devices include industrial and automotive sensing, environment monitoring, structural monitoring, and military surveillance/detection. Battery lifetime constraints are critical, and the battery must be physically small to facilitate in-situ sensing in a broad range of uses. To extend the lifetime of the sensor nodes, energy harvesting from the ambient environment can be leveraged as long as occasional degradation in performance quality, depending on the ambient factors, can be tolerated.

Ultra-low-power low-voltage MSP430 microcontroller



Courtesy: Verma & Chandrakasan

Energy Collecting and Harvesting Options

Energy Source	Performance
Thermoelectric	60 μW/cm^3
Light	100 μW/cm^2 (office), 100 mW/cm^2 (direct light)
Vibration	4 μW/cm^3 (human motion)
Heel strike	10-700 mW (walking)
Near-field inductive energy transfer	20 mW at 5 cm [33]
Far-field inductive energy transfer	2 μW at 10 m [34]

Structure of Modern SRAM

Row Decoder

Column-
Mux

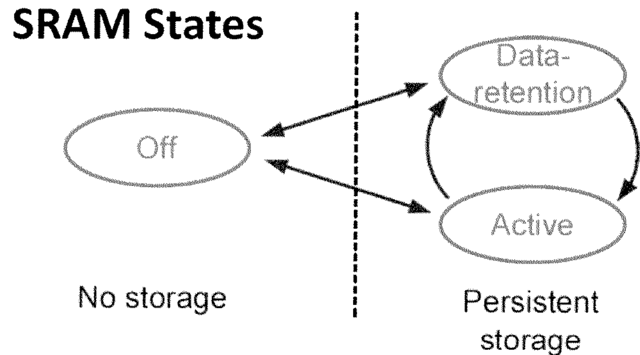
Sense-
Amp

Bit-
Cell

Bit-
Cell

Bit-
Cell

Pre-
charge



(courtesy Intel)

WL
M3
M4
M5
M6
M1
M2
NC
NT
BLC
BLT

Device	Function
M1-2	Driver
M3-4	Load
M5-6	Access

SRAM States

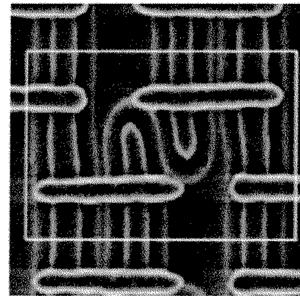
Off

Data-
retention

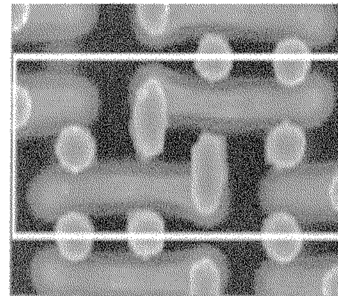
Active

No storage

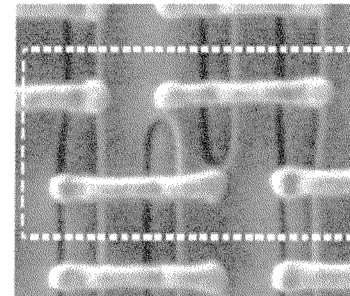
Persistent
storage



TI 65nm:
0.46x1.06um²



IBM 65nm:
0.41x1.25um²

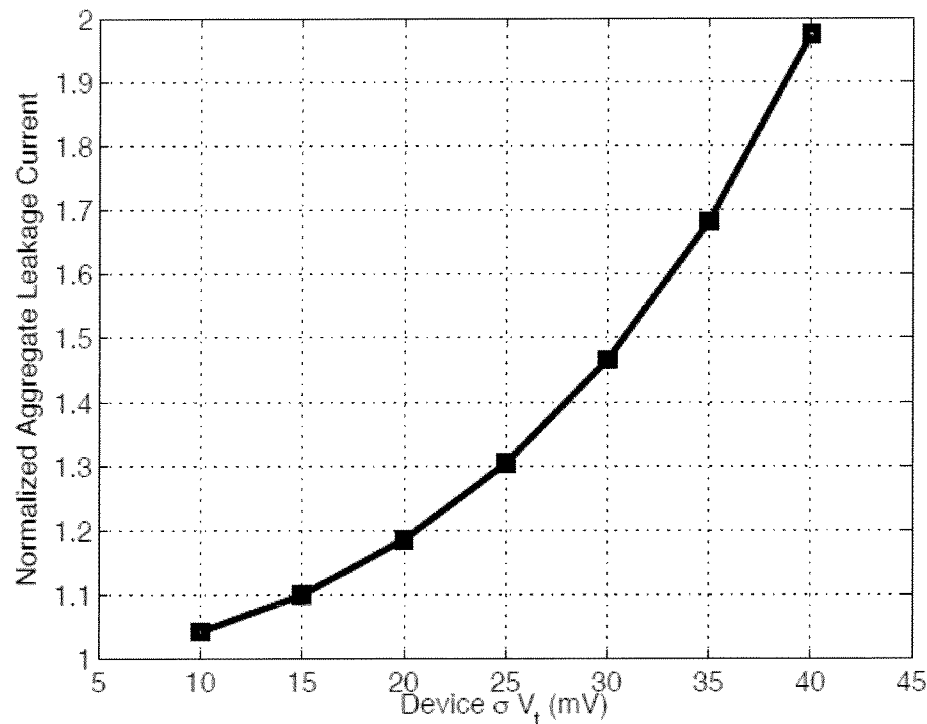


Intel 65nm:
0.46x1.24um²

Courtesy: Verma & Chandrakasan

SRAM Leakage Energy

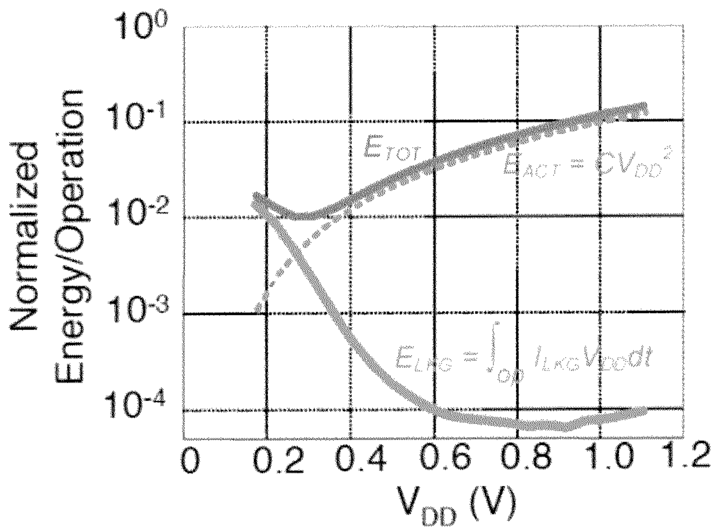
SRAM leakage-energy increases due to three factors: (1) High ratio of leakage-paths to actively-switched-nodes; (2) Total leakage set by an aggregation of intentionally minimum sized devices; and (3) Critical path set by a single MOSFET pull-down stack with extreme variation.



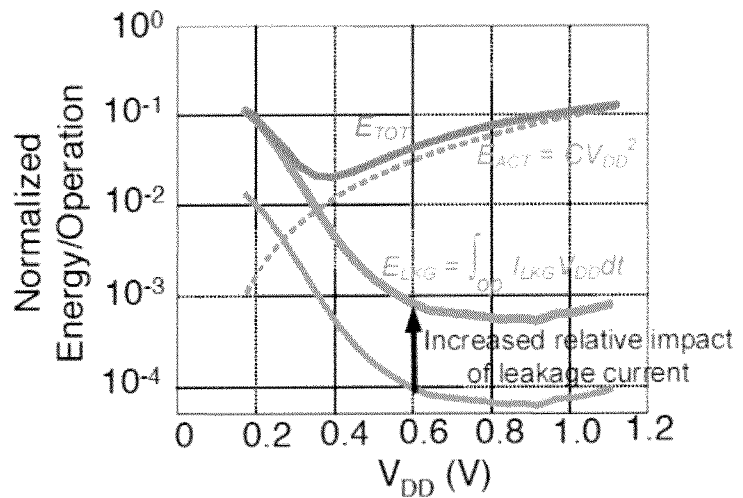
The simulated total aggregate leakage-current (at 1.1V), normalized to the nominal aggregate leakage-current, for a 1Mb array composed 0.25 μm^2 bit cells in an LP 45 nm technology.

Courtesy: Verma & Chandrakasan

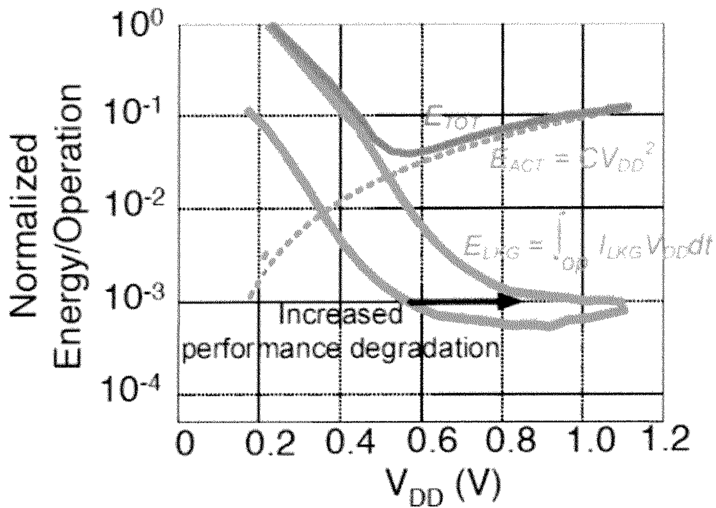
Circuit Delay Aggravation due to Performance Degradation



(a) Energy profiles represntative of generic logic (90nm 32b carry-lookahead adder).



(b) Relative leakage-energy shift exepcted in SRAMs due to increased ratio of leakage-currents to active-switching-current.



(c) Relative leakage-energy shift expected in SRAMs due to severe performace degradation from bit-cell variation.

The severe performance degradation due to the critical-path's dependence on a single bit-cell experiencing extreme variation, causes the leakage-energy curve to shift right-ward.

This can be understood by observing that the point at which the leakage-energy begins increasing exponentially occurs at a higher supply-voltage (0.8 V) than before (0.6 V). Effectively, the variation raises the limiting bit-cell's threshold voltage, and, as a result, supply-voltage reduction quickly leads to sub-threshold operation, which imposes an exponential increase in circuit delay.

Courtesy: Verma & Chandrakasan

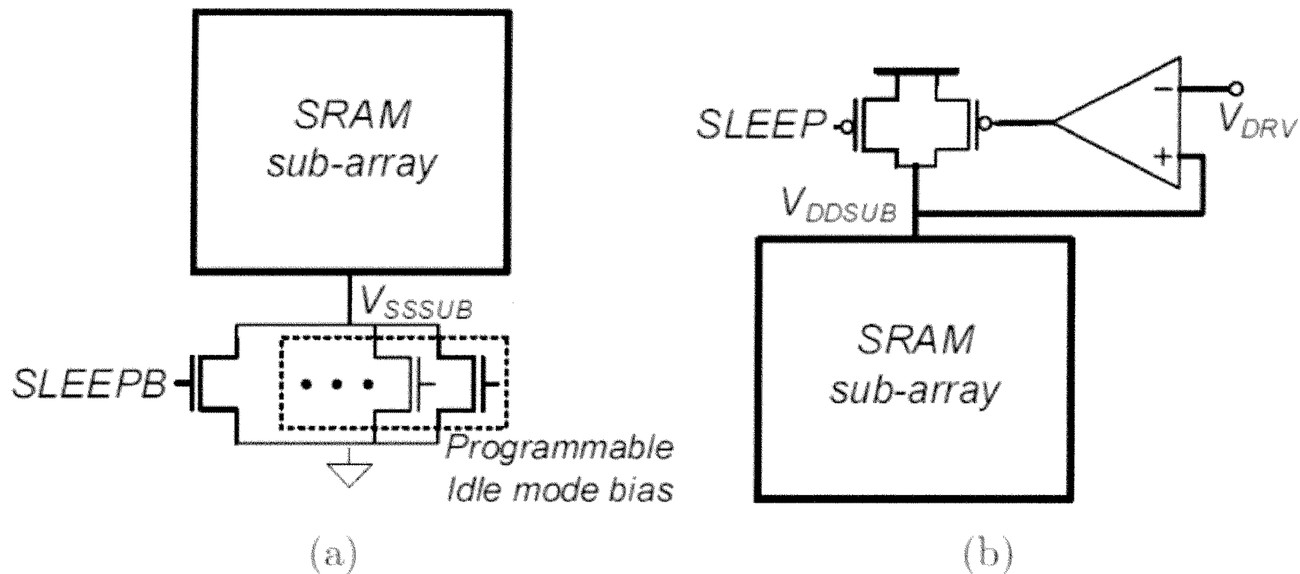


Figure 2-5: Circuitry to enforce idle-mode biasing using (a) programmable sleep switches [63] and (b) an operational-amplifier [64].

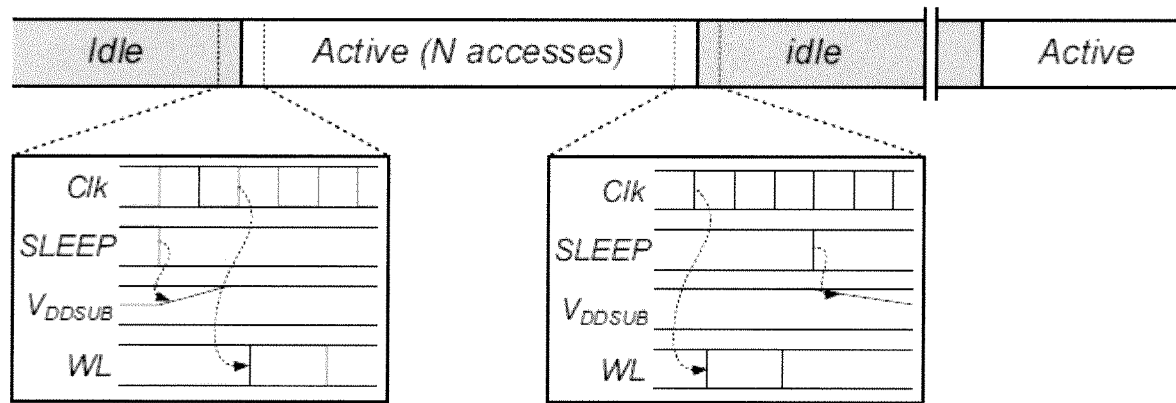


Figure 2-6: Waveforms corresponding to idle-to-active and active-to-idle mode transitions.

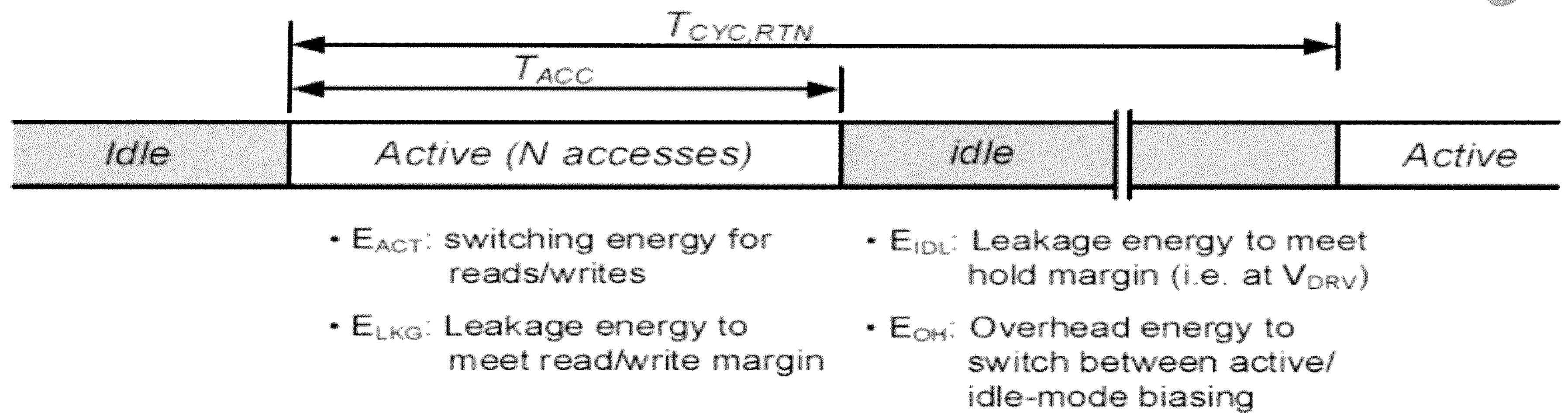


Figure 2-7: Summary of SRAM energy components.

$$E_{TOT} = E_{ACC} + E_{LKG} + E_{IDL} + E_{OH} \tag{2.1}$$

The active-access-energy (E_{ACC}) and the leakage-access-energy (E_{LKG}) pertain to the active mode. E_{ACC} corresponds the switching energy required to perform reads and writes, and E_{LKG} corresponds to the leakage-energy imposed by applying a supply-voltage across the array that must be large enough to ensure reliable reads and writes. The idle-data-retention energy (E_{IDL}) corresponds to data storage during the idle-mode, and it will also be referred to as the idle-mode energy. Finally, the overhead-energy (E_{OH}) corresponds to the overhead incurred due to altering the sub-array's biasing in accordance with idle-mode power reduction. These components are sum-

Breakdown of Energy Sources

The total active-access-energy for reads of an i × j (i.e. i-column, j-row) sub-array is given by

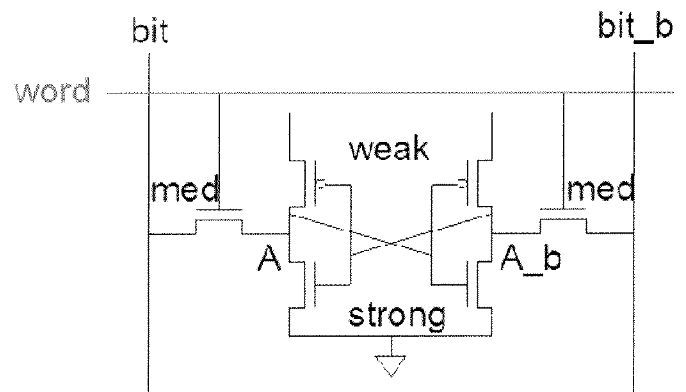
$$E_{ACC,RD} = C_{WL}V_{DD}^2 + C_{cSEL}V_{DD}^2 + \frac{i}{m}C_{SA}V_{DD}^2 + iC_{BL}V_{DD}V_{SNS}$$

Full-swing signals typically include the one-hot enabled word-line, WL, for row selection, and the one-hot enabled column-select, cSEL, for multiplexed column selection in a column-interleaved array. In total, the number of sense-amplifiers is equal to the number of columns in the sub-array divided by the column-multiplexing ratio, m. The most significant source of active-access-energy consumption, however, is the bit-lines, BL, which are used to convey the stored read-data to the sense amplifiers and to drive new write-data into the bit-cells. Strictly speaking, to resolve the read-data, the BLs need only discharge to the required sense-amplifier input margin, **V_{SNS},** which can be less than 100mV. Nonetheless, in practice, the BLs are often discharged beyond the sensing margin to reduce the probability of data-disruption caused by sustained pulling of the bit-cell storages nodes towards the BL voltage near VDD. During read accesses, for instance, the design in [68] actively amplifies the signal on all BLs to full logic levels in order to avoid data-disruption.

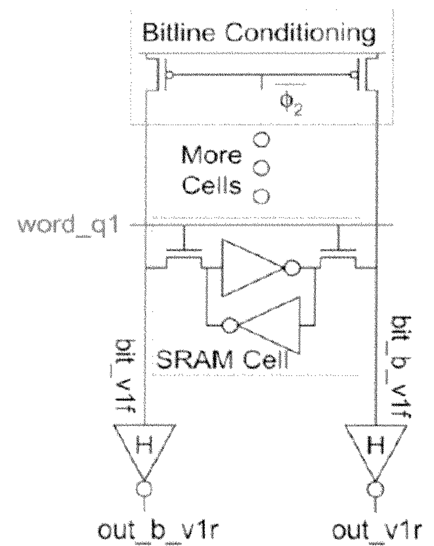
The total active-access-energy for writes is approximately given by:

$$E_{ACC,WR} = C_{WL}V_{DD}^2 + C_{cSEL}V_{DD}^2 + \frac{i}{m}C_{BL}V_{DD}^2 + i\frac{m-1}{m}C_{BL}V_{DD}V_{SNS}$$

Transistor Sizing

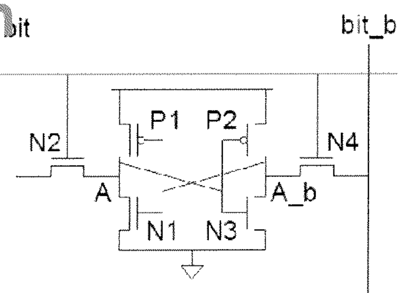


- High bitlines must not overpower inverters during reads
- But low bitlines must write new value into cell

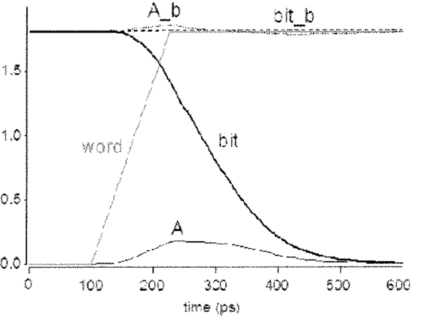


Courtesy:
Harris & Weste

READ Operation

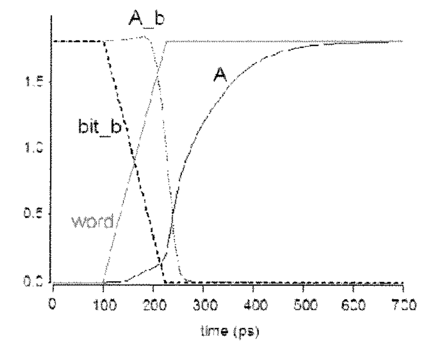


- Precharge both bitlines high
- Then turn on wordline
- One of the two bitlines will
  - be pulled down by the cell
- Ex: A = 0, A_b = 1
  - bit discharges, bit_b stays high
  - But A bumps up slightly
- *Read stability*
  - A must not flip
  - N1 >> N2

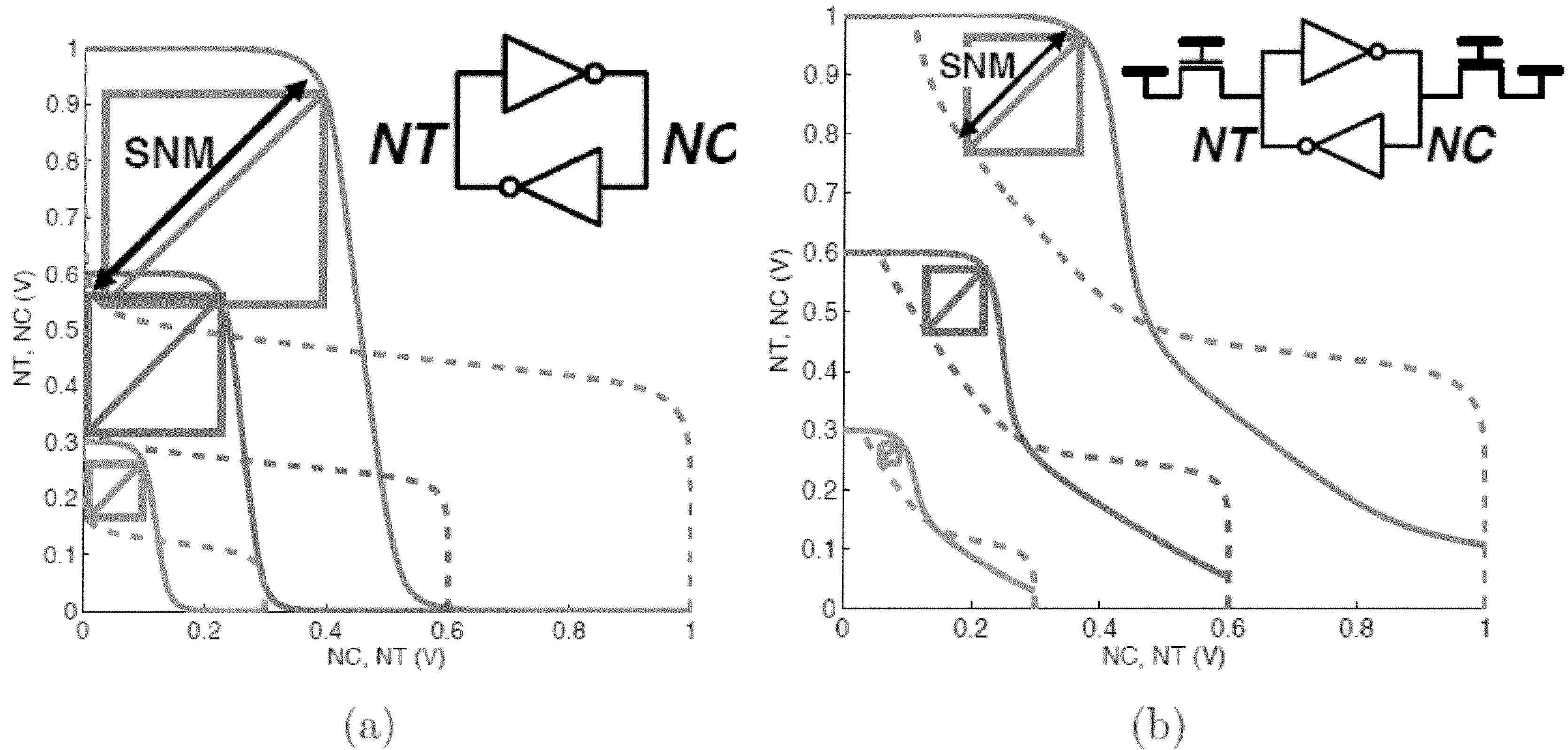


WRITE Operation

- **Drive one bitline high, other low**
- **Then turn on wordline**
- **Bitlines overpower cell**
- **Ex: A = 0, A_b = 1, bit = 1, bit_b = 0**
  - **Force A_b low, then A rises high**
- ***Writability***
  - **Must overpower feedback**
  - **P2 << N4 to force A_b low,**
  - **N1 turns off, P1 turns on,**
  - **raise A high as desired**



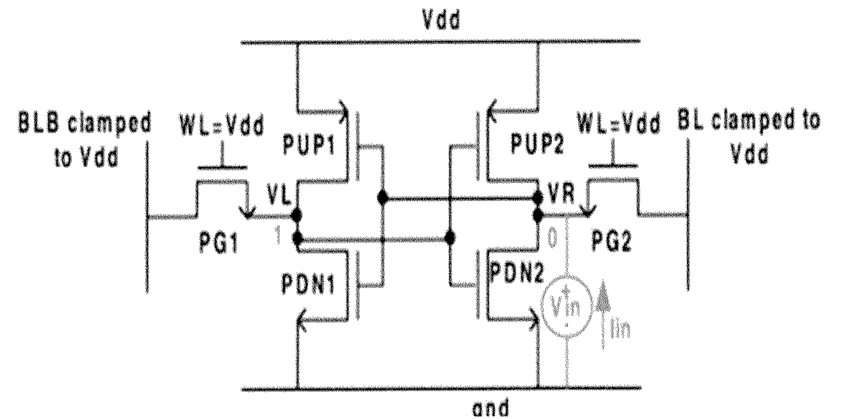
SRAM Noise margins for Hold and Read Operations

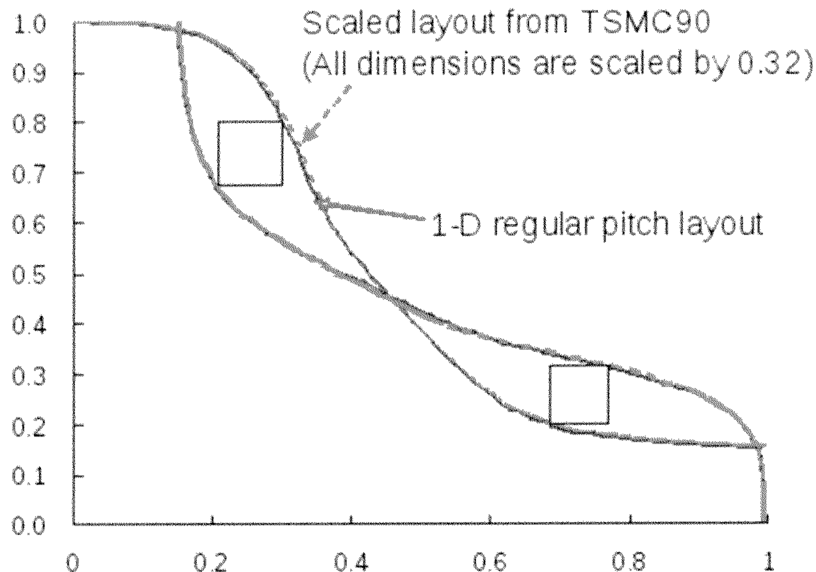


6T bit-cell butterfly curves showing bi-stable behavior during (a) hold, where access devices are "off", and during (b) read, where access devices are "on" and bit-lines are clamped to VDD.

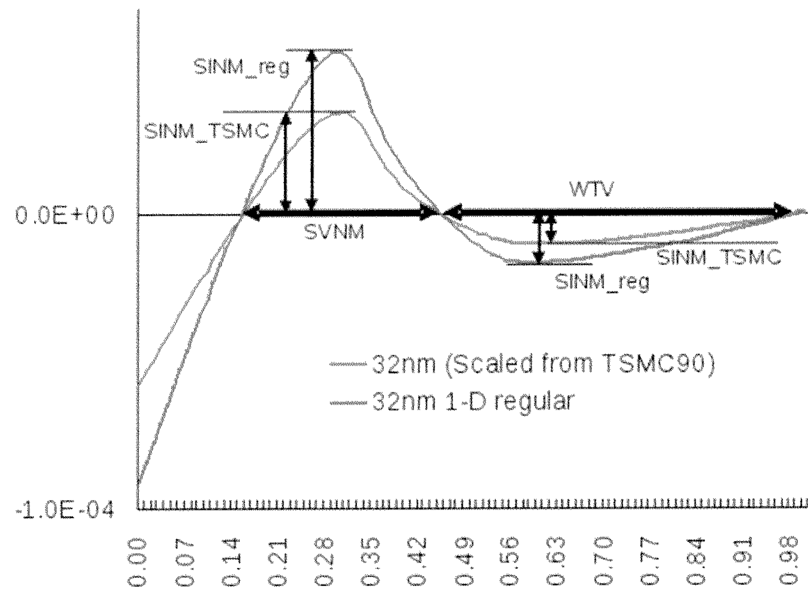
Measurement of Noise Margin

- Measure method
  - Increase VR and measure VL
  - Increase VL and measure VR
  - Make voltage transfer curve in VR and VL axes → Butterfly
  - Measure I_{in} → N-curve





Butterfly Curve

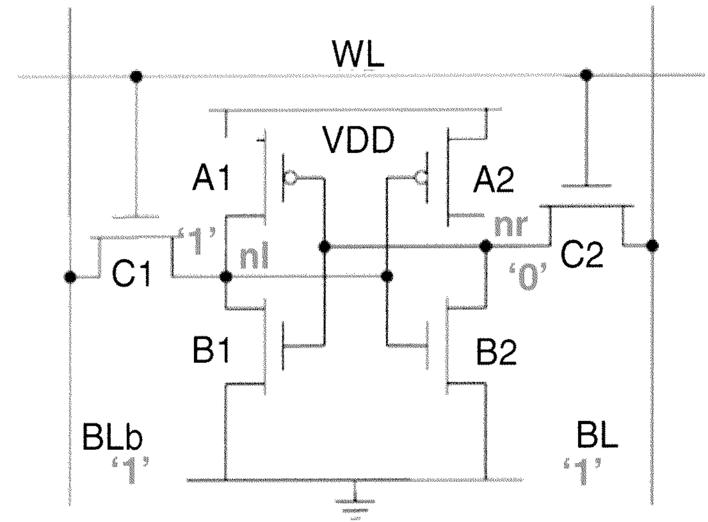


N-Curve

Courtesy: LT Microelectronics

Measurement of I_{read}, $I_{Leakage}$ and VDD_{Hold}

- I_{read}
  - Measure bitline current when WL switches to high
- $I_{LEAKAGE}$
  - Measure VDD (or VSS) current when WL=0
- VDD_{HOLD}
  - Decreasing VDD voltage, while WL=0
  - Measure minimum VDD voltage when | V(nl) - V(nr) | = 'sensing margin' **(100mV is assumed)**

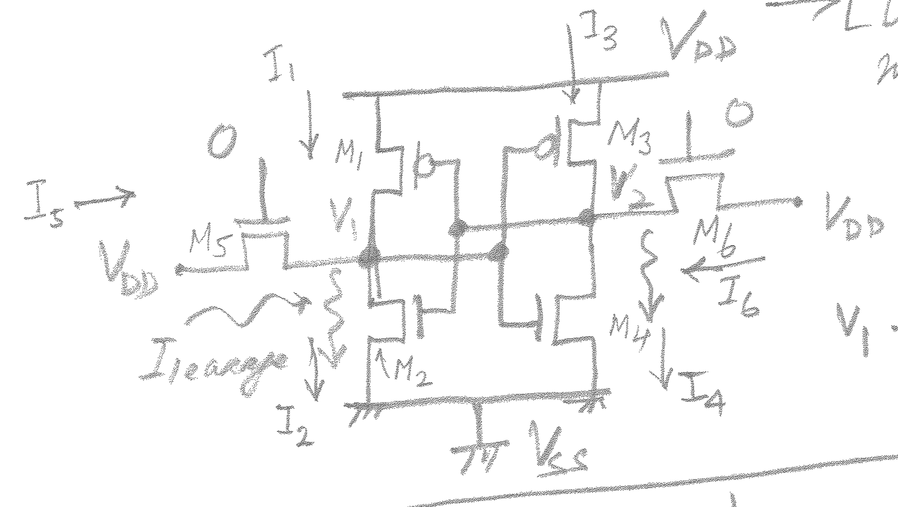


	Reference Cell	32 nm (for 30x12 and 25x12)
I_{read}	41.2 uA	66.7 uA
$I_{leakage}$	85.4 nA	142.7 nA
VDD_{HOLD}	110 mV	118 mV

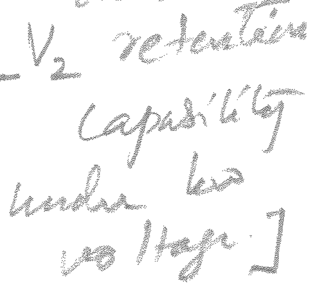
Courtesy: LT Microelectronics

①

SRAM Data Retention Voltage (DRV) Analysis:



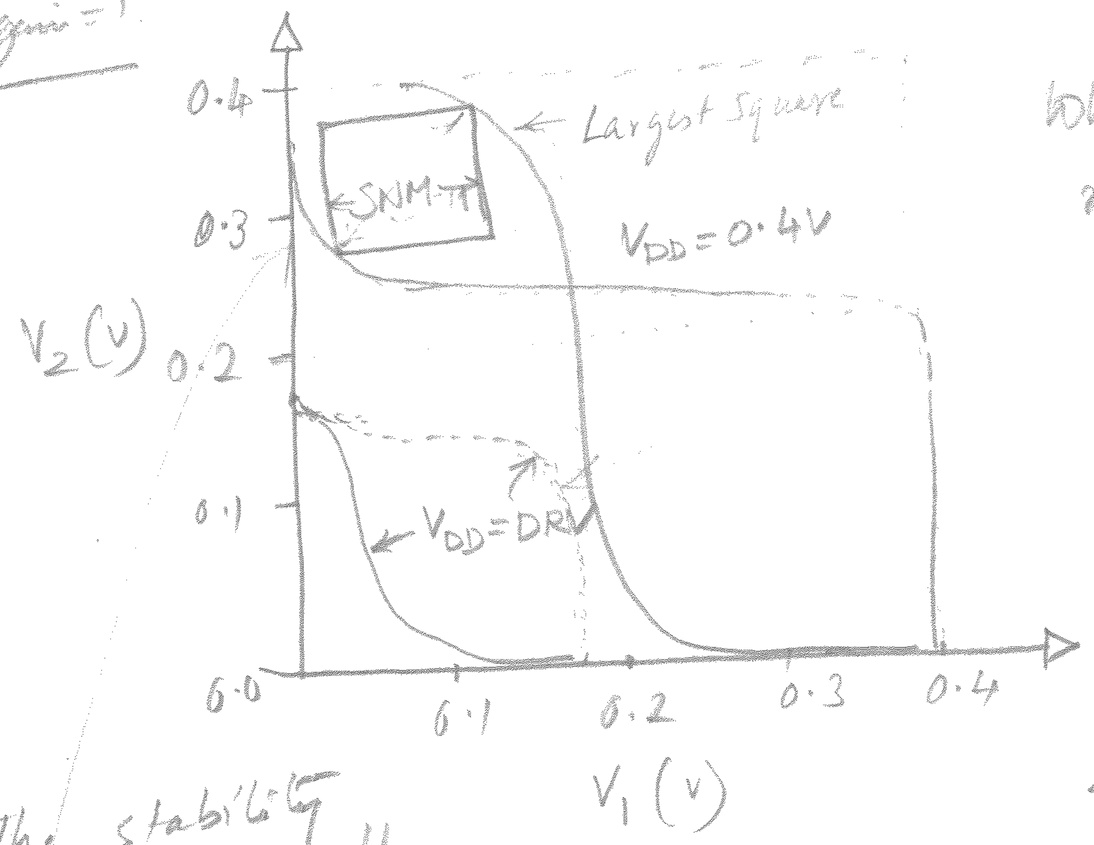
→ [DRV is defined as the minimum V_{DD} required for data preservation and is a measure of its state-retention capability under low voltage]



The loop gain of inverter pair ≥ 1

At V_{DD} = DRV, loop gain = 1

$$\left.\frac{\partial V_1}{\partial V_2}\right|_{\text{Inverter I}} \times \left.\frac{\partial V_2}{\partial V_1}\right|_{\text{inverter II}} = 1 \quad @ V_{DD} = DRV$$



When V_{DD} is reduced to DRV, all six transistors ($M_1 \to M_6$) in 6T SRAM will be in sub-threshold mode of operation

The stability of SRAM cell is indicated by SNM. At V_{DD} = DRV, SNM = 0

(2) Calculation of lower bound of ...

@ V_1: $I_1 + I_5 = I_2$ .... (1)

@ V_2: $I_3 + I_6 = I_4$ ... (2)

Assuming ZERO leakage currents of M_5 & M_6 during Standby mode. Assume bit lines are connected to VDD in HOLD mode.

$\Rightarrow I_1 = I_2, \quad I_3 = I_4$ .. (3)

Assume, $V_1 \approx 0$ and $V_2 \approx V_{DD}$

The sub-V_T current is given by:

$$I_i = \beta_i I_0 \exp\left(\frac{V_{GS,i} - V_{T,i}}{n_i \phi_T}\right) \left(1 - \exp\left(\frac{-V_{ds,i}}{\phi_T}\right)\right)$$

$$\phi_T = \frac{kT}{q} = 26\,mV \text{ @ } 300°K$$

$$\beta_i = \frac{W}{L}$$

I_0 is the leakage current of a unit sized device at $V_{GS} = 0$ and $V_{DS} >> \phi_T$

$T \rightarrow$ chip temperature

$n_i \rightarrow$ sub-threshold factor (sub-threshold swing divided by 60mV at room temperature).

Let us define:

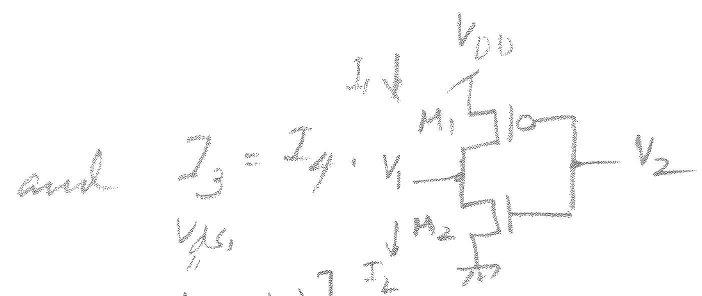
$$I_{off,i} = \beta_i I_0 \exp\left(\frac{-V_{th,i}}{n_i \phi_T}\right),$$

$$I_i = I_{off,i} \cdot \exp\left(\frac{V_{GS,i}}{n_i \phi_T}\right) \left[1 - \exp\left(-\frac{V_{DS,i}}{\phi_T}\right)\right]$$

(3) If we consider DIBL effect:

$$V_{T,i} = V_{Ti,0} + \gamma_i \left(\sqrt{|-2\phi_i + V_{SB,i}|} - \sqrt{|-2\phi_i|} \right) - V_{DS,i} \cdot exp(-\alpha L_i).$$

Since all the SRAM cell transistors conduct in weak inversion region when V_{DD} is around DRV, the DIBL effect can be ignored in a DRV analysis.

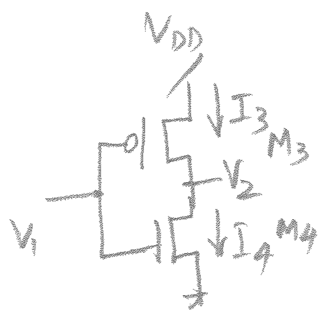


We have, $I_1 = I_2$ and $I_3 = I_4$.

$$\Rightarrow exp\left(\frac{V_{DD}-V_2}{n_1\phi_T}\right) \cdot \left[1 - exp\left(-\frac{V_{DD}-V_1}{\phi_T}\right)\right]$$

($V_{DD}-V_2 = V_{GS1}$, $V_{DD}-V_1 = V_{DS1}$)

$$= exp\left(\frac{V_2}{n_2\phi_T}\right) \cdot \left[1 - exp\left(-\frac{V_1}{\phi_T}\right)\right]. \rightarrow (5)$$



$$exp\left(\frac{V_{DD}-V_1}{n_3\phi_T}\right) \left[1 - exp\left(-\frac{V_{DD}-V_2}{\phi_T}\right)\right]$$

($V_{DD}-V_1 = V_{GS3}$, $V_{DD}-V_2 = V_{DS3}$)

$$= exp\left(\frac{V_1}{n_4\phi_T}\right) \cdot \left[1 - exp\left(-\frac{V_2}{\phi_T}\right)\right] \rightarrow (6)$$

Here, $I_{off,N} = I_{off,P}$ and $n_N = n_P$. initial cond^{n}:
Obtain $\frac{dV_1}{dV_2}$ & $\frac{dV_2}{dV_1}$. Then, apply $V_1 \approx 0$ & $V_2 \approx DRV$

$$\Rightarrow DRV_{ideal} = 2\phi_T \ln(1+n). \quad \ldots (7)$$

[PROVE THIS in HW]

For $n=1$ (ideal CMOS technology with 60 mV/decade swing), $DRV_{ideal} = 36$ mV.

$n=1$, $S = 60\ mV/decade$

$DRV_{ideal} = 36\ mV$. $(2 \times 26\ mV \times \ln 2 = 36\ mV)$

For a typical 90 nm technology with $n = 1.5$

$DRV = 2 \times 26 \times \ln(1 + 1.5)\ mV$

$= 52 \ln 2.5 \cong 47\ mV < 50\ mV$.

Verify DRV using SPICE simulation.

Thus, the ideal CMOS with $n=1$, $DRV_{ideal} = 36\ mV$ no matter how well we optimize the size of $M_1 \to M_6$ and V_T of transistors.

Since, $DRV_{ideal} = 2\phi_T \ln(1+n)$, can we design a new CMOS technology such that $n = 0$

$\Rightarrow DRV_{ideal} = 0$.

Hence, in an ideal charge-based SRAM technology, power dissipation can be reduced to zero in stand-by mode of operation

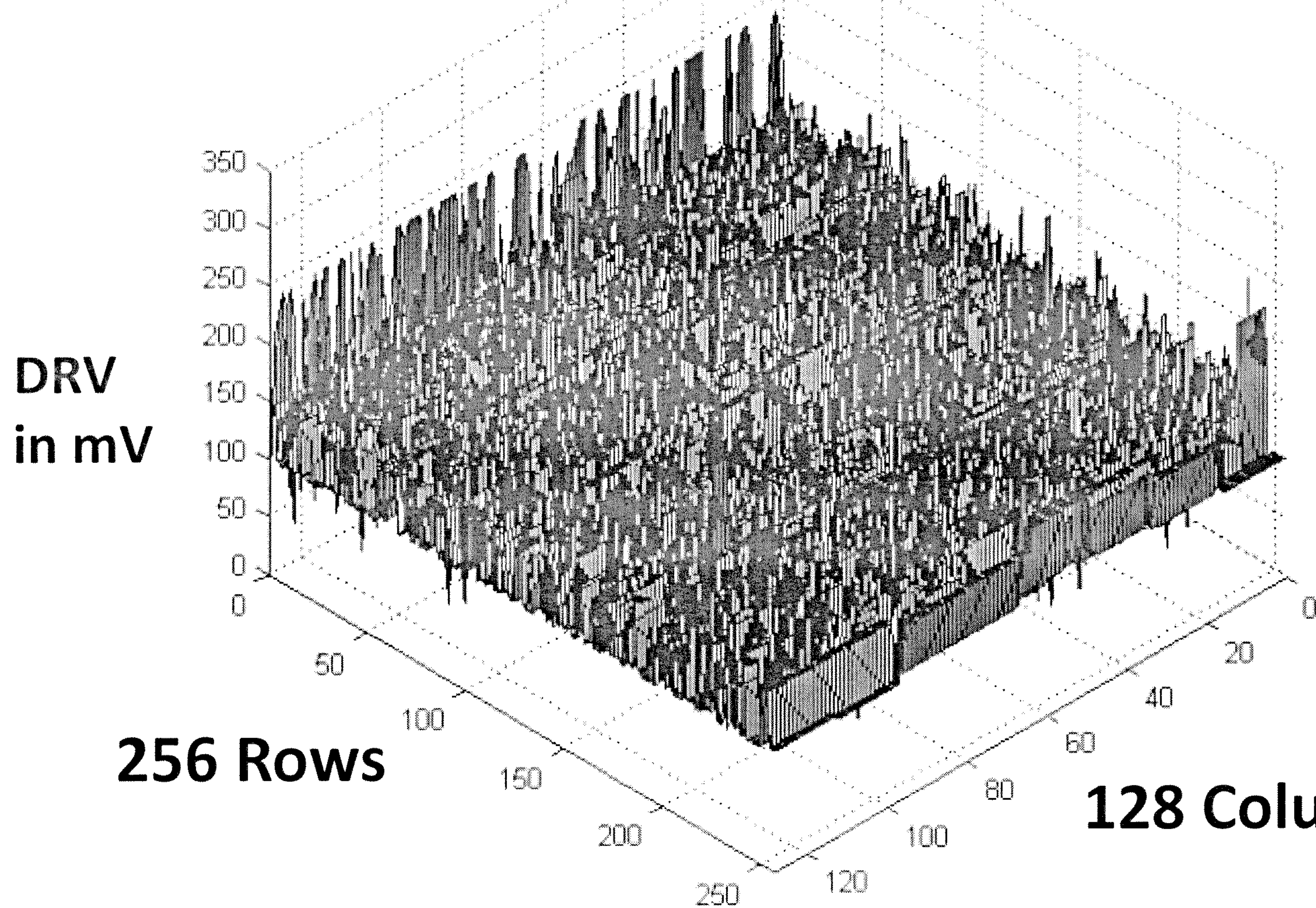
Why DRV should be Minimized

SRAM power is the dominant power consumption factor in applications that are primarily in the standby mode.

In CMOS technology, standby power consists of leakage-power which increases with each silicon-technology generation.

For ultra low-power devices, standby leakage power reduction is crucial for device-operation within the scavenging power limit.

Spatial Distribution of DRV for a 130 nm 32 Kb SRAM



DRV in mV

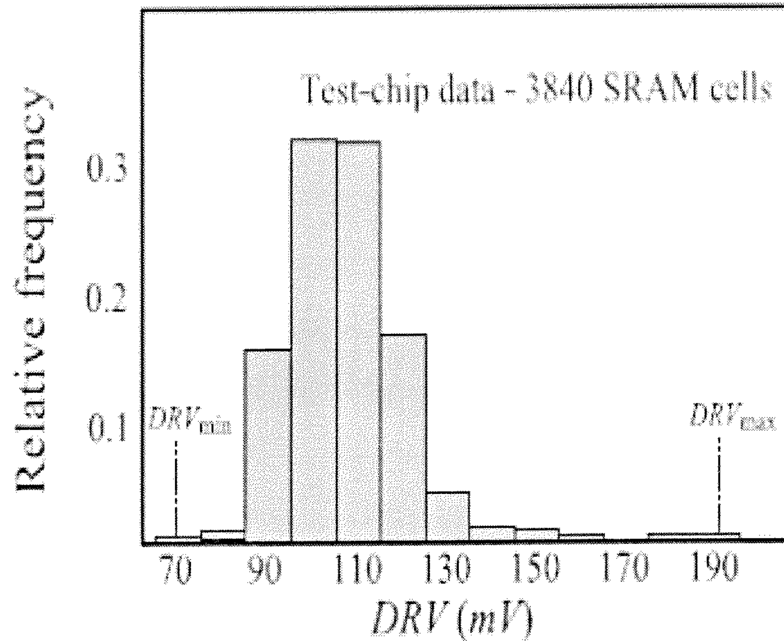
256 Rows

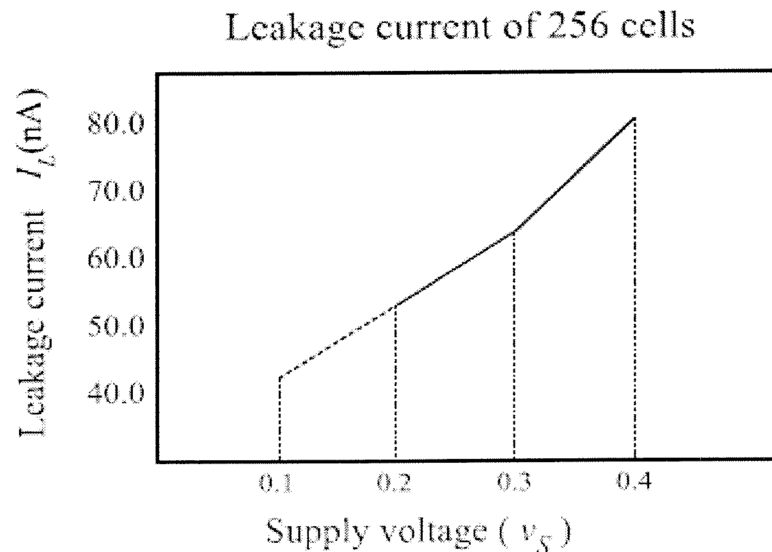
128 Columns

DRV in SRAM cells varies between 50 mV and 240 mV

DRV varies due to local parameters like Threshold Voltage and Channel Length of transistors in SRAM cells.

Another Example with 90 nm

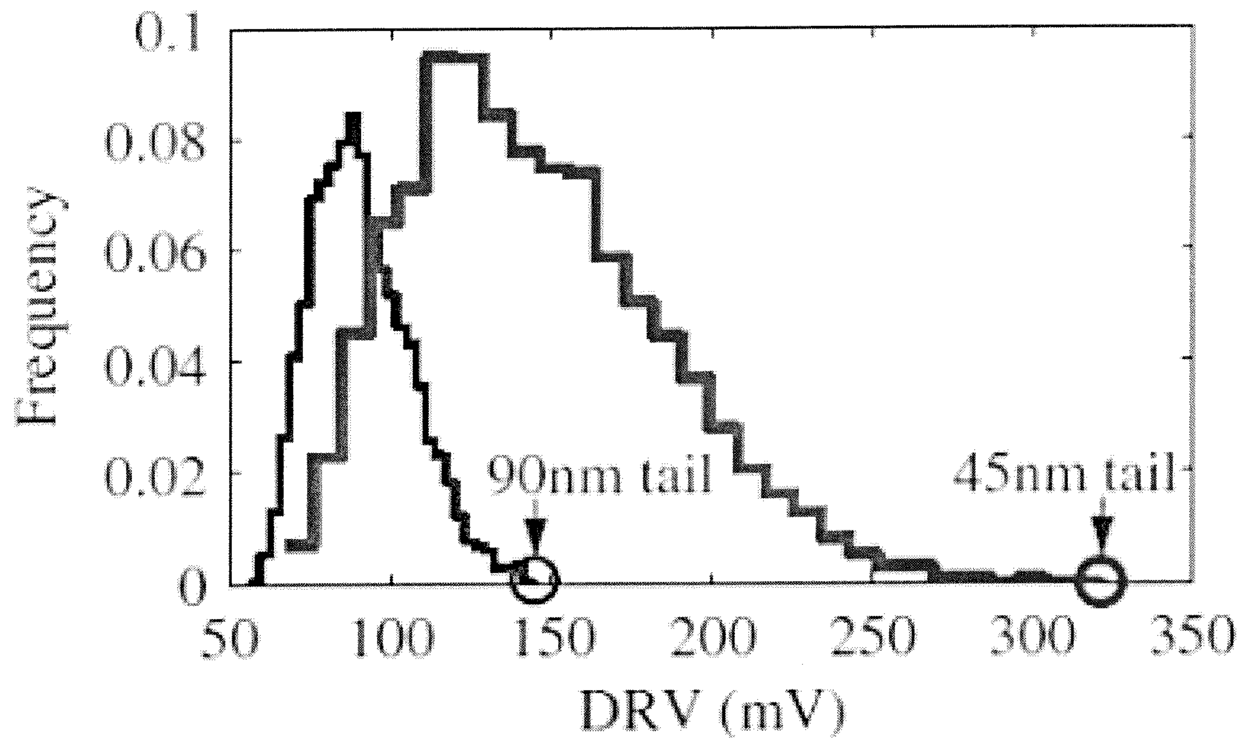




Test-chip DRV-distribution: The experimental intra-chip DRV varies from 70 to 190mV in the 90nm CMOS technology. The worst-case solution for data-retention is a supply voltage of 200mV. If we add a 100 mV of guard band, the Standby Supply voltage will be about 300 mV which is about 200 mV more than the majority of the cell DRV values as shown in the Histogram.

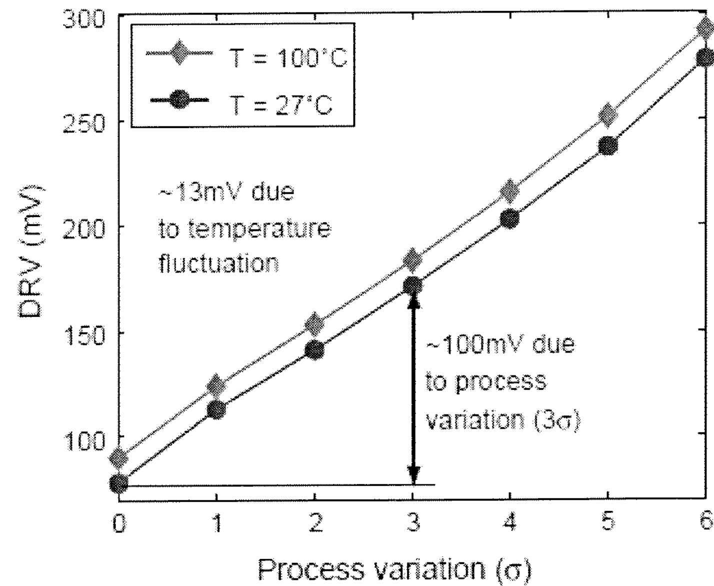
In the range 100–400mV, the leakage-current is approximately piecewise linear.

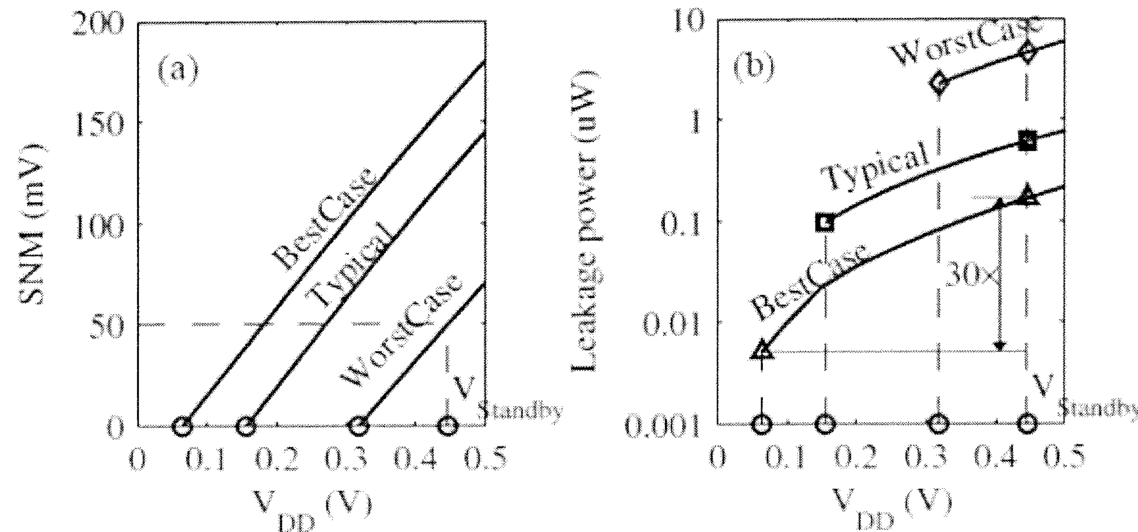
Technology Node v. DRV



DRV distribution from a 5k-point Monte-Carlo simulation of within-die variation for 90nm and 45nm nodes. The tail sets the array-wide VStandby.

DRV varies slightly with temperature, but widely with process variation





Simulated worst bitcell SNM (a) and 1kb SRAM leakage power (b) vs. VDD under PVT variations (best-case, typical and worst-case) and 3σ local mismatch.

Techniques to Minimize DRV

Fault-Tolerant Memory with ECC

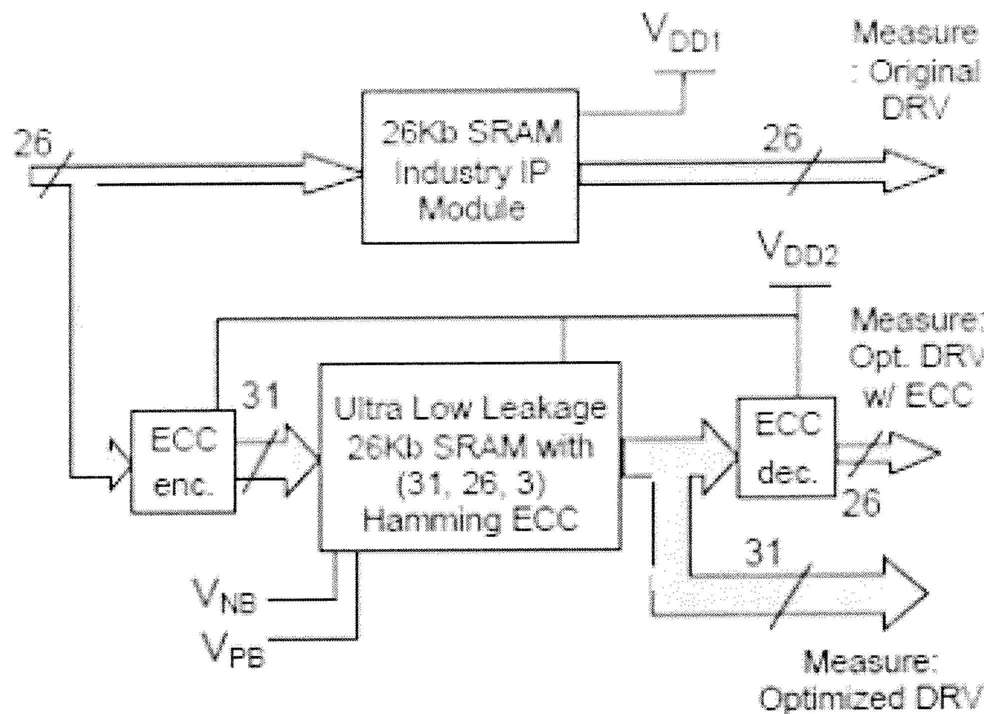
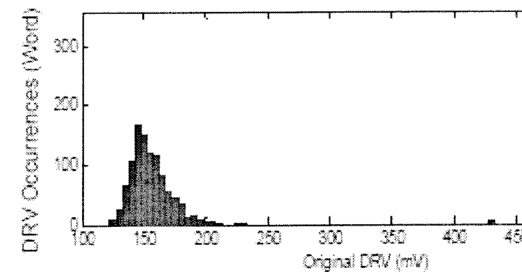
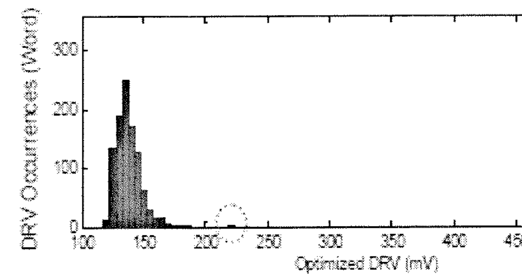


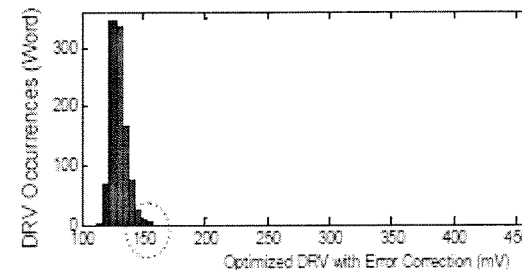
Table 1. Worst-case DRV range measured on 24 chips

DRV (mV)	Original	Optimized	Optimized w/ ECC
Min.	320	170	140
Max.	570	220	160

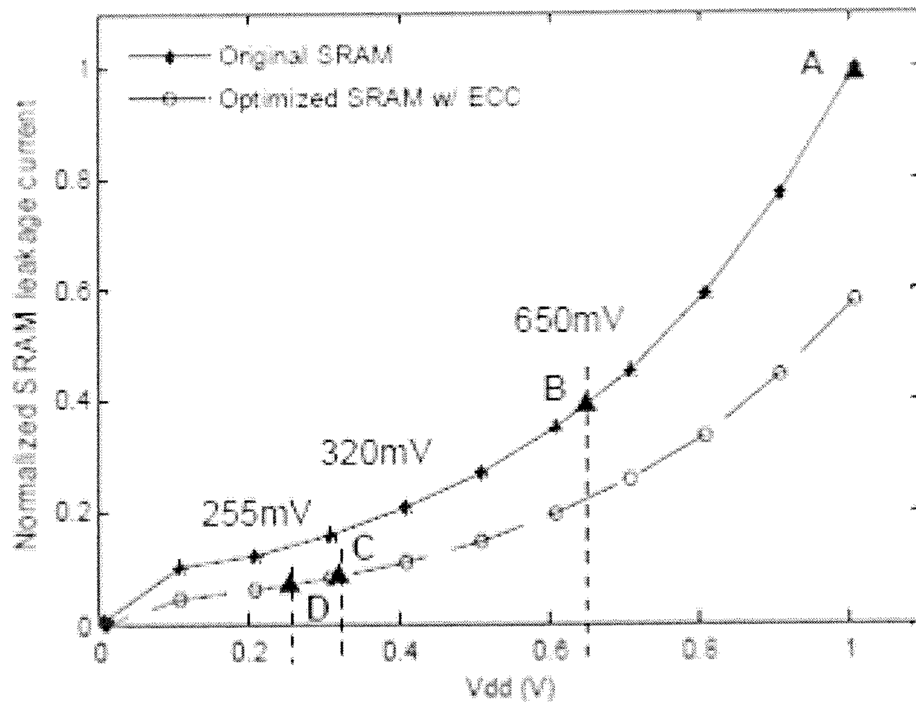
ECC Reduces the DRV



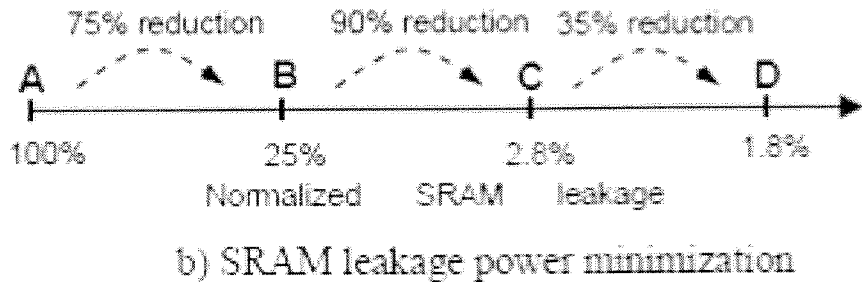




Power Savings when SRAM is in Stand-by Mode



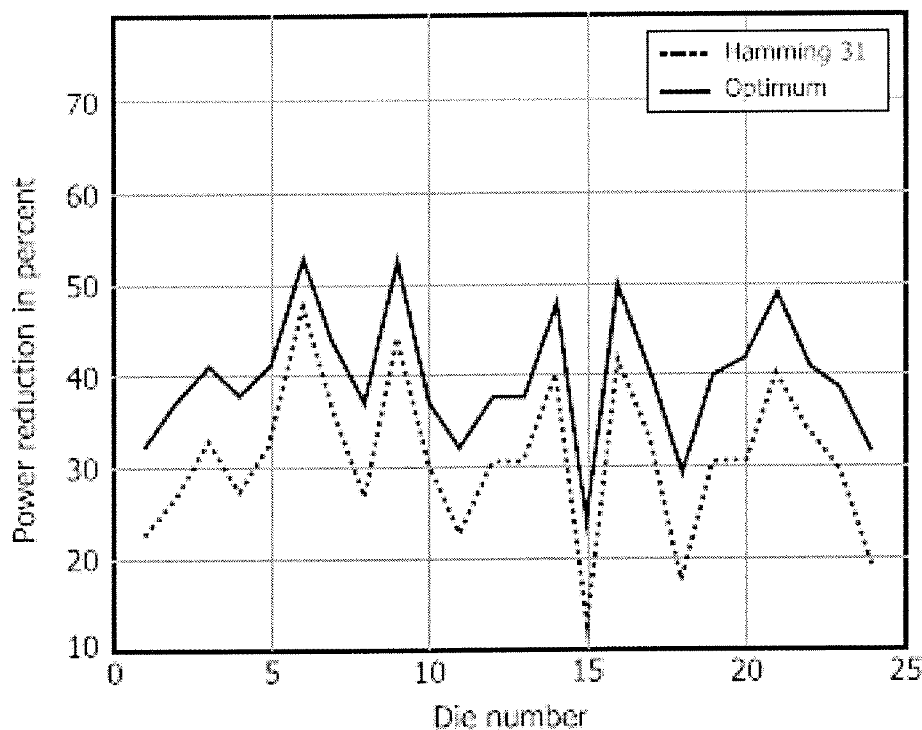
a) Measured SRAM leakage currents

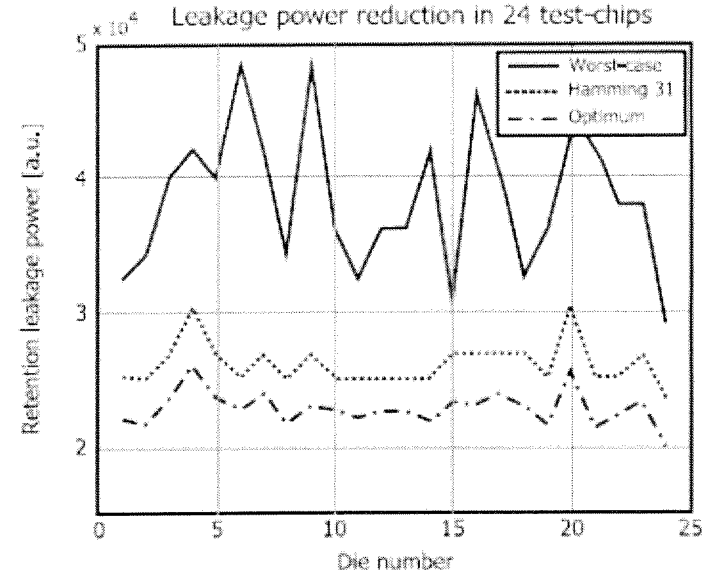


b) SRAM leakage power minimization

Compared to the leakage power consumption at 1V standard VDD (A), lowering the standby VDD to 650mV (B) reduces the leakage power of an un-optimized SRAM design by 75%. The DRV-aware SRAM cell optimization brings the standby VDD down to 320mV (C), leading to another 90% leakage power reduction. The error-tolerant design further lowers the standby VDD to 255mV (D), and reduces the leakage power by an extra 35%. This final design (D) consumes only 1.8% of the original memory leakage power (A).

ECC Saves Stand-by Power in SRAM





The leakage power for the worst-case method, the [31.26,3] Hamming code based implementation, and the theoretical optimum are compared.

Power reduction for the [31,26,3] Hamming code based implementation and the theoretical optimum are compared. The implementation tracks the optimum within a close margin of 6-11%.