Chapter 3
Multi-Mode Power-Gating Technique

CAMs are popular in network routers for packet forwarding and packet classification in the recent years. Network routers forward data packets from an incoming port to an outgoing port, using an address-lookup function. Ternary cells, in addition, store an “X” value. The “X” value is a don’t-care that represents both “0” and “1”, allowing a wildcard operation. Wildcard operation means that an “X” value stored in a cell causes a match regardless of the input bit. Based on the continuous input don’t care “X” patterns, the super cut-off power gating technique are proposed which reduce cell standby power as well as search power significantly. The standby power sources would be introduced in Section 3.1. The principles of stacking effects are described in Section 3.2. Conventional power-gating (or gated-V_{DD}) SRAM cells are discussed in Section 3.3. The proposed multi-mode data retention power gating TCAM scheme would be presented in Section 3.4. Eventually, some conclusions and discussions are addressed in Section 3.6.

3.1 Standby Power

By the CMOS technology progress, the leakage power problem becomes much more serious than before [94], [95], [96], and [97]. The leakage current can be divided into various components, such as subthreshold, band-to-band tunneling, gate tunneling, pn-junction reverse bias, DIBL, GIDL, and punch-through leakage. Fig. 3.1 depicts the leakage sources in a MOSFET device.

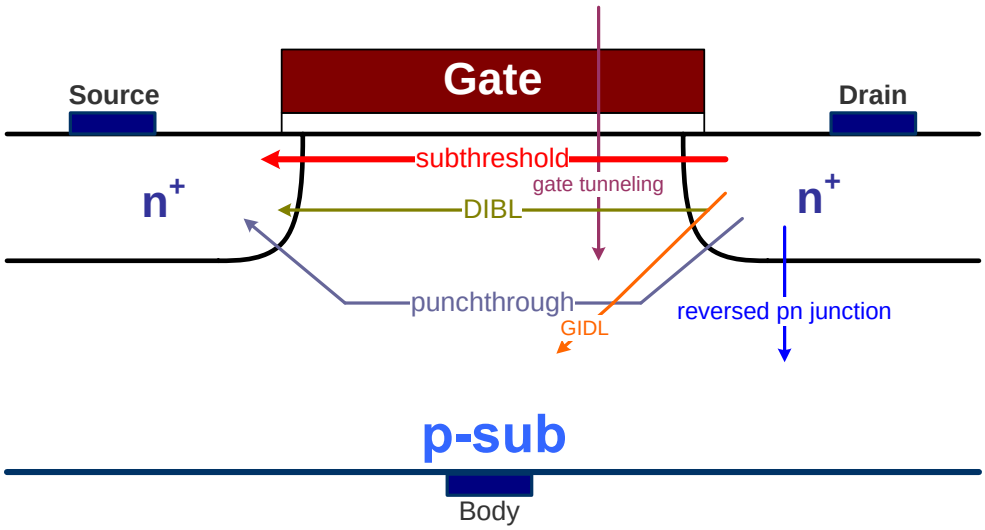


Fig. 3.1 Leakage current sources in a MOSFET device

Even though the leakage current is composed of various sources, the subthreshold leakage current is the dominant component that is given as:

$$I_{leakage} = I_0 \cdot \exp\left(\frac{V_G - V_S - V_{T0} - \gamma V_S + \eta V_{DS}}{nV_{thermal}}\right) \cdot \left(1 - \exp\frac{-V_{DS}}{V_{thermal}}\right) \quad (3.1)$$

where $V_{thermal}$ is the thermal voltage, n is the subthreshold swing coefficient constant, γ is the linearized body effect coefficient, and η is the DIBL coefficient. Assuming that $V_{DS} >> V_{thermal}$ and (3.1) can be simplified to the expression

$$I_{leakage} = I_0 \cdot 10^{(V_{DS} - V_t)/nV_{thermal} \ln 10} \quad (3.2)$$

Eq. (3.2) implies that the subthreshold leakage current is smaller with higher threshold voltage, V_t, and this component is becoming important since threshold voltage is scaled down with process of CMOS technology.

Subthreshold leakage is becoming the dominant component of power consumption in deep-submicron technologies. However, as the process steps into the region of nano-scale technologies, gate leakage current has the potential to dominant the leakage current, even exceed the level of dynamic power. Gate leakage is given as:

$$J_{DT} = AE_{ox}^2 \exp\left(-\frac{B\left[1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right]}{E_{ox}}\right) \tag{3.3}$$

$$A = \frac{q^3}{16\pi^2\hbar\phi_{ox}}, B = \frac{4\sqrt{2m^*}\phi_{ox}^{3/2}}{3\hbar q} \tag{3.4}$$

Where V_{ox} represents the voltage drop across the oxide, ψ_{ox} is the barrier height in the conduction band, and E_{ox} is the field across the oxide. In detail, the gate leakage current is composed of three types of :I_{gd}, the gate leakage current between gate and drain, I_{gb}, between gate and the body, and I_{gs} between gate and source. Fig.3.2 predicts that the gate leakage current is indispensable in nano-scale technologies and the amount of gate leakage is far beyond the standby power constraint. Gate leakage current becomes critical due to the decrease of thickness of gate oxide.

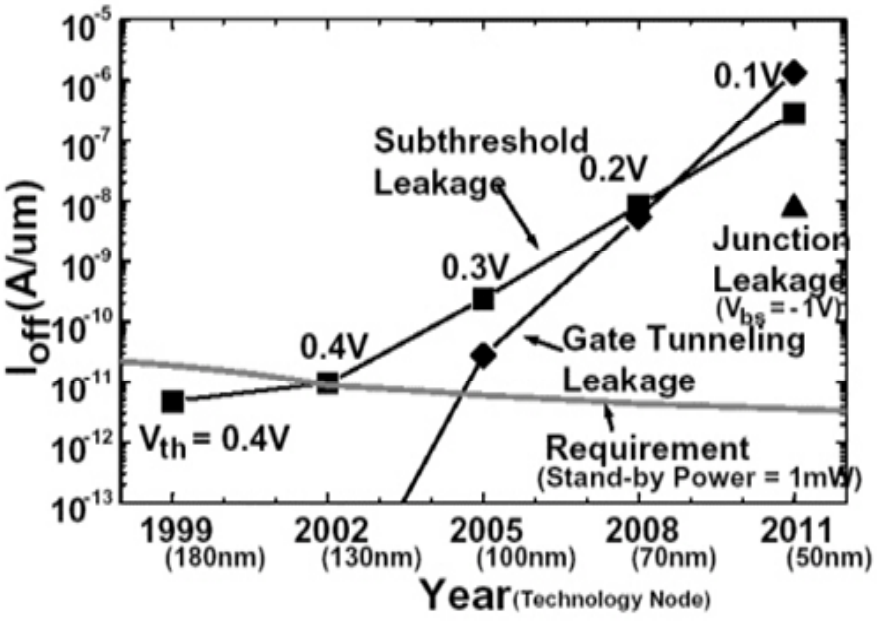


Fig. 3.2 The trend of standby current of MOSFETs

Fig. 3.3 illustrates the I_{ON}/I_{OFF} ratio and Vt for both low and high Vt NMOS transistors for 130nm, 100nm, and 70nm technologies, where I_{ON} means driving current in active mode and I_{OFF} stands for leakage current in standby mode. Fig. 3.3 reveals that leakage current is increasing by 3-5x per generation and becoming comparable with active driving current. The extra leakage current wastes a significant amount of power and causes thermal hot spots and thermal run away problems.

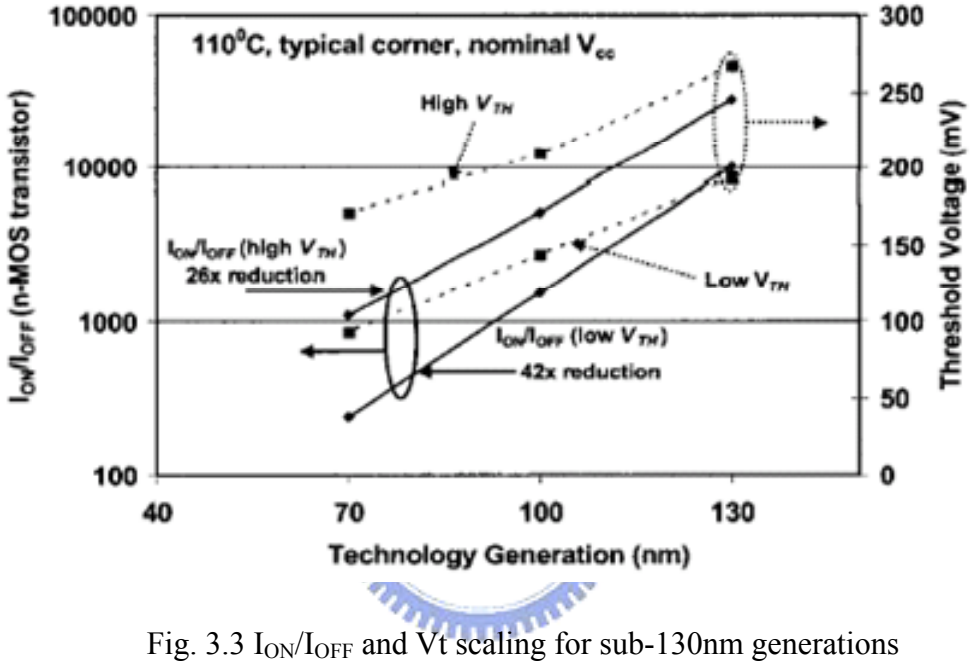
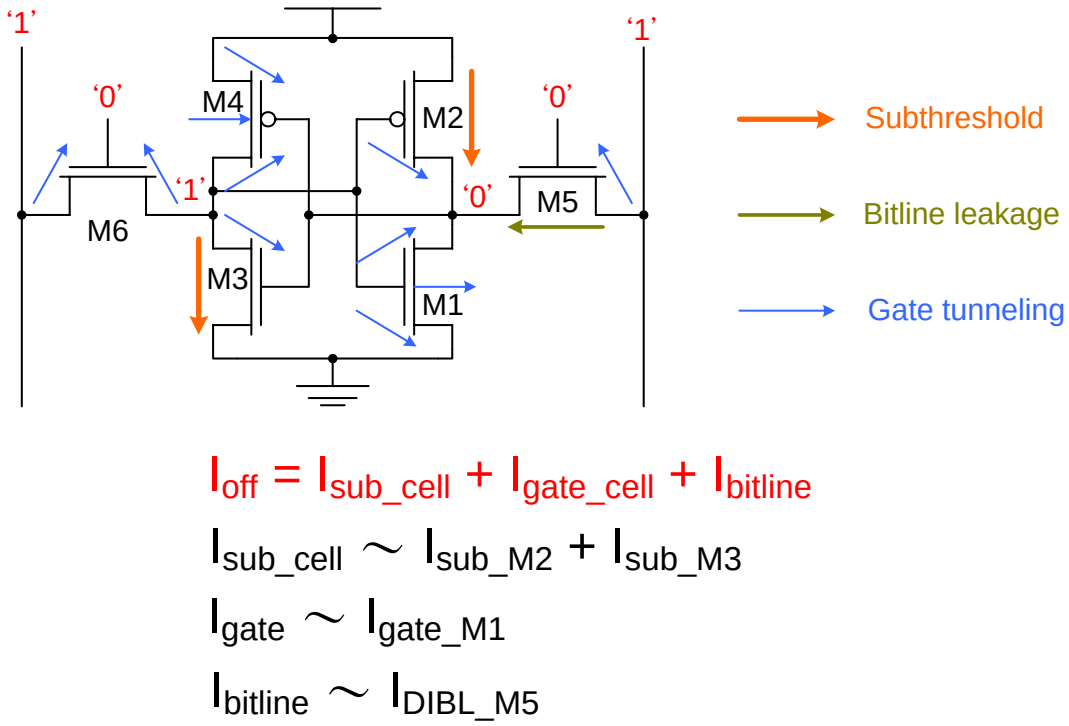


Fig. 3.3 I_{ON}/I_{OFF} and Vt scaling for sub-130nm generations

Fig. 3.4 shows the leakage paths and standby leakage equations in a SRAM cell. However, the power-gating technique has been proposed to reduce the standby power.



$$I_{off} = I_{sub\_cell} + I_{gate\_cell} + I_{bitline}$$

$$I_{sub\_cell} \sim I_{sub\_M2} + I_{sub\_M3}$$

$$I_{gate} \sim I_{gate\_M1}$$

$$I_{bitline} \sim I_{DIBL\_M5}$$

Fig. 3.4 Leakage currents and standby currents equations in a SRAM cell

3.2 Stacking Effect

It has been observed that stacking of two off transistors has significantly smaller sub-threshold leakage current than the stacking of one off transistor [98], [99]. The phenomenon is called stacking effect and it is due to self-reverse biasing of stacked transistors.

3.2.1 Self-Reverse Biasing

From Fig. 3.5, it can be observed that how to reduce leakage current by self-reverse biasing. The left circuit is an off NMOS transistor with current I_1, which is mainly composed of sub-threshold leakage. And the right one is two stacked off NMOS transistors and the leakage current is I_2. In the steady state, the voltage, V_x, is slightly higher than ground. Therefore, it causes the transistor M_{21} a negative V_{gs} (gate-to-source voltage) to make the pn junction reversely biased. According to this reason, leakage current I_2 is smaller than I_1.

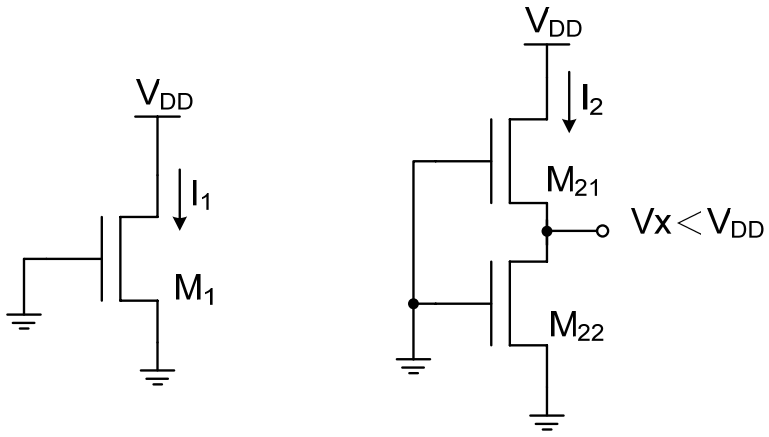


Fig. 3.5 Stacking effect dues to self-reverse biasing of transistor M_{21}

3.2.2 Trade-off Between Delay and Leakage

From the section 3.2.1, the two stacked off transistors have better performance than one off transistors for the leakage current. However, the two stacked off transistors have a penalty, increasing the delay, for getting less leakage current. Fig. 3.6 depicts the trade-off phenomenon between delay and leakage current. Fig. 3.6(b) shows an ordinary inverter, the width of PMOS and NMOS are $2W$ and W, respectively. Fig. 3.6(c) illustrates the modified inverter, and the initial NMOS is split into two parts which the sizes are $w/2$ and $w/2$, respectively. Fig. 3.7 depicts the simulation result of tradeoff between delay and leakage current. From the figure, it have been obviously seen that smaller leakage current comes with large delay.

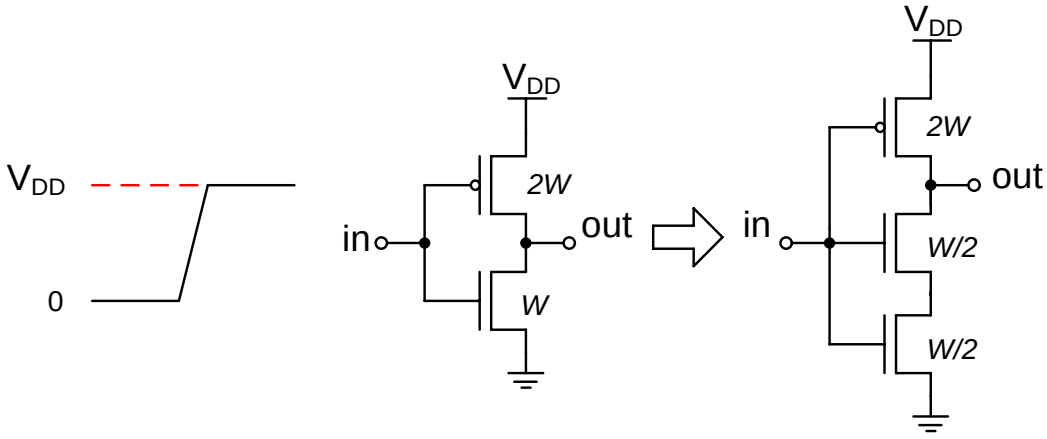


Fig. 3.6 Tradeoff between delay and leakage current for a stacked inverter

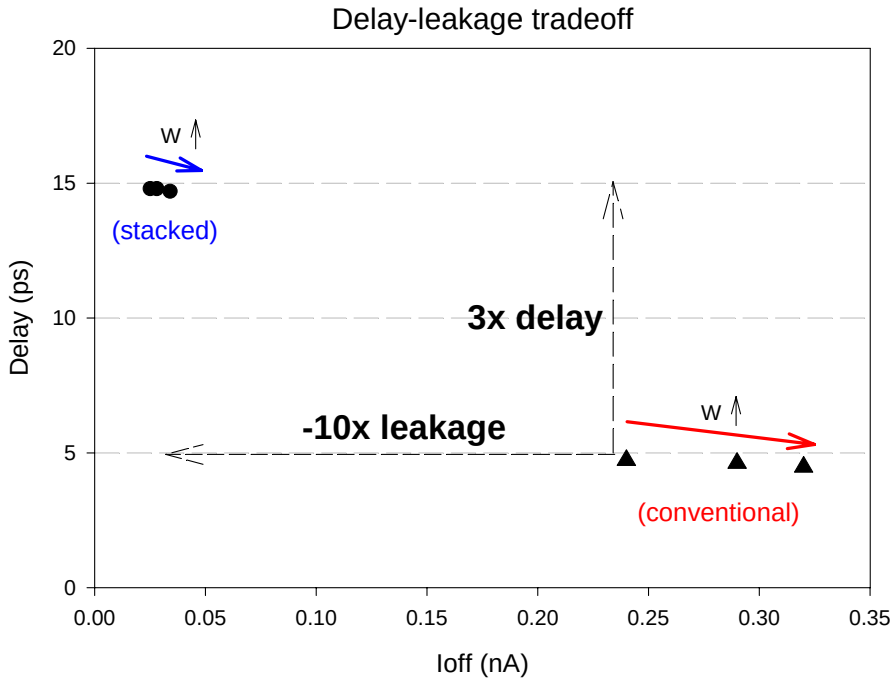


Fig. 3.7 Delay-leakage tradeoff of stacking effect

3.3 Power Gating Structure with Concurrent Data Retention and Intermediate Mode

The conventional power gating devices can be classified into two main categories: footer and header devices. Footer is by inserting NMOS sleep transistors between real ground and virtual ground and header is by inserting PMOS sleep transistors between real V_{DD} and virtual V_{DD} as shown in Fig. 3.8. The internal circuits can be either combinational or sequential circuits. The power gating devices receive sleep signal from power management unit which decides what system power saving scenario is adopted now. This conventional power gating strategy is useful for combinational circuits. However, these gating devices ruin the static noise margin of the storage elements in sequential circuits.

Shown in Fig. 3.9 is the SRAM array with power gating devices. From table 1 we know that the SNM is 0mV when power gating devices are turned off which means data stored on SRAM can no longer be guaranteed to be correct. However, 24X-leakage-current reduction is achieved through the usage of conventional power gating devices during standby mode.

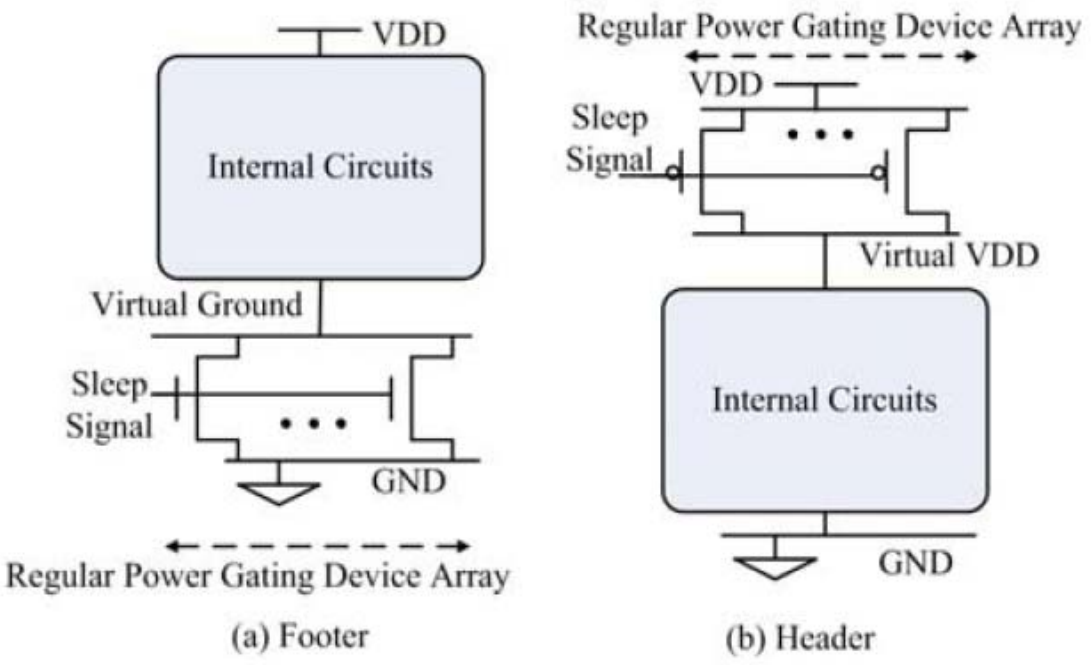


Fig. 3.8 (a) conventional NMOS footer array power gating devices (b) conventional PMOS header array power gating devices

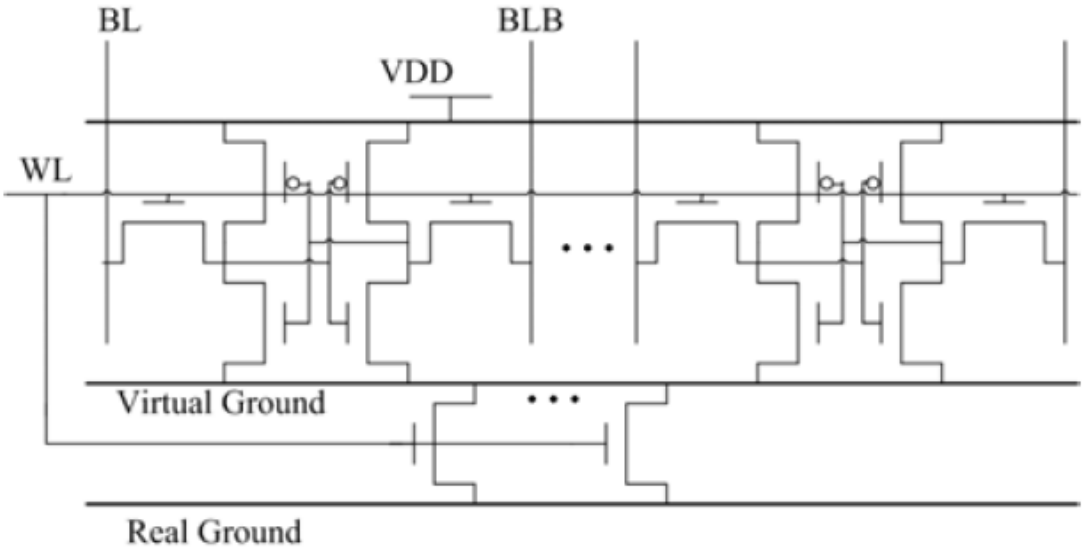


Fig.3.9 SRAM array with power gating device array to reduce standby power consumption

Lots of publications proposed MTCMOS as a power gating device solution [100],

[101]. However, using higher threshold voltage transistors as the power gating devices require larger silicon area for power gating devices to be capable of sinking the maximum instantaneous current at active mode. Therefore, using single threshold voltage transistors as the power gating devices or adaptively adjusting the well bias of the power gating devices are desirable to reduce the silicon area occupied by power gating devices [102].

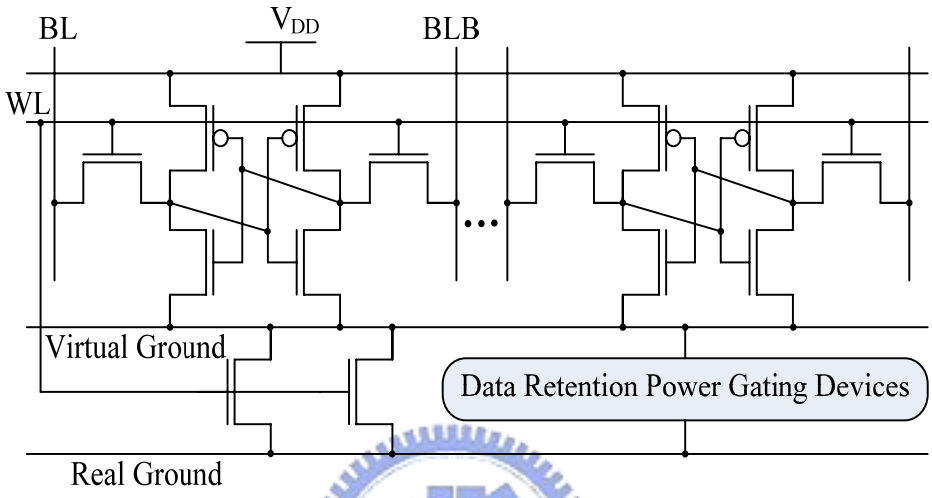


Fig. 3.10 Regular power gating devices with data retention power gating devices to maintain static noise margin of storage elements in standby mode

From Section 2 we know that conventional power gating devices work well with combinational circuits. But Storage elements need a new set of power gating devices call data-retention power gating devices to maintain static noise margin during standby mode. This structure is shown in Fig. 3.10 where the data retention power gating devices are inserted in parallel with regular power gating devices. A set of power gating devices are shown in Fig.3.11.

The data-retention power gating devices never truly turn off. Fig. 3.11(a) is a small NMOS transistor with its gate biased to a specific voltage. Due to this current connecting virtual ground and real ground, the static noise margin is as good as not seeing any power gating devices. However, the leakage current reduced by this data-retention power gating device is very small.

Fig. 3.11 (b) is by connecting NMOS transistor's gate and drain together to form a voltage controlled resistor. While at active mode operation, the virtual ground line is nearly at equi-potential. After turning off the regular power gating devices, the virtual

ground line is floating and the leakage current begins to charge up the potential of the virtual ground line. At first the data-retention power gating device acts as a highly resistive resistor. As potential of the virtual ground line begins to raise, the effective resistance between virtual ground and real ground becomes smaller which prevents the potential on virtual ground line rise further. Finally, the potential comes to equilibrium and steady. Due to the reduced potential difference between V_{DD} and virtual ground, the static noise margin is degraded but within an acceptable range. The leakage current is reduced but not as much as conventional power gating devices alone. As shown in Table 1, Type 2 NMOS resistor data retention power gating device can cut the leakage current in half compared to the intrinsic leakage current. Type 3 PMOS resistor data-retention power gating device as shown in Fig. 3.11(c) is its PMOS counterpart of Type 2. If the device sizes of Type 2 NMOS resistor and Type 3 PMOS resistor are the same, Type 3 will show better ability to suppress leakage current due to higher effective resistance. Therefore, enhanced Type 2 and Type 3 data-retention power gating structures to suppress leakage current are shown in Fig. 7(a) and 7(b). The basic idea is by cascoding or stacking NMOS/PMOS transistors to achieve high effective resistance. Therefore, we can trade static noise margin for lower leakage or allow higher leakage to gain more static noise margin. The simulation results of Type 3 data-retention power gating devices are shown in Fig. 8 and Fig. 9. These two figures show clearly that the amount of leakage current and static noise margin are all directly proportional to the size of gating device which can be referred to as the reciprocal of effective resistance. Finally, an NMOS diode-connected transistor as an example of the data retention power gating device is shown in Fig. 3.12.

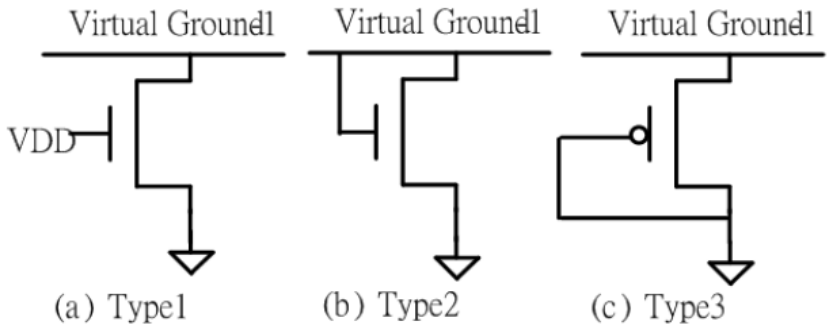


Fig. 3.11 Three types of different data retention power gating devices

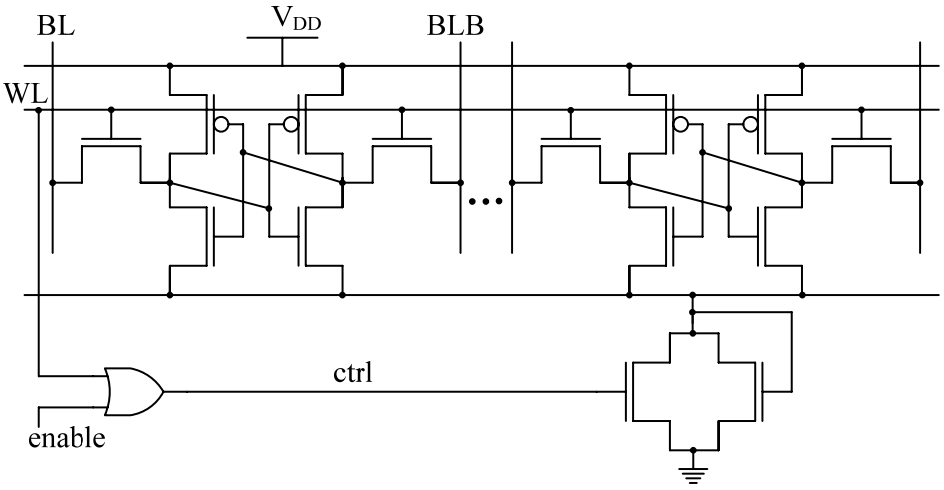


Fig. 3.12 NMOS diode data-retention power gating technique

3.4 Proposed Power Gating Structure

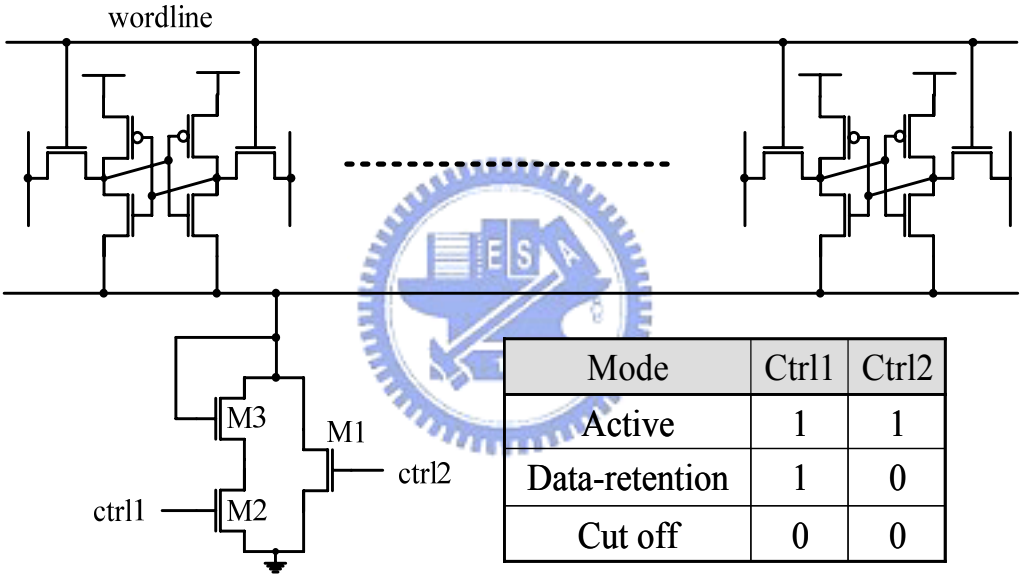
3.4.1 Multi-mode Data Retention Power Gating Technique

The data-retention power gating circuit is shown as Fig. 3.13, and it is modified from the circuit we had proposed in [103]. We adopt the regular power gating NMOS (M1) and a diode connected NMOS (M3), which function is similar to that in [103], except that the PMOS diode is replaced by M1 to get better speed in active mode. In addition, an additional NMOS (M2) is stacked to increase the virtual ground voltage further.

Due to the diode connected NMOS, the virtual ground will saturate to a limited value, 142mV. If the virtual ground voltage charged by the leakage current increases above the saturated value, M3 will be turned on and the virtual ground voltage will discharge through it. Thus, the virtual ground drops back to the saturated value, which assures the stability of the data storage cells.

The three modes of data-retention power gating control circuit are list in the truth table as Fig. 3.13. When the circuit is in active mode, both control signals are asserted to high and three power gating transistors are turned on to support full speed operation. When the circuit enters the data-retention mode, ctrl1 will be high and ctrl2 will be

low. Meanwhile, M3 just represents a diode, which causes the virtual ground to saturation voltage and provides sufficient noise margin. When all don't care cells in a TCAM segment are set as don't care state, the data stored in storage cells are meaningless and can be destroyed. On the other hand, the data-retention power gating control circuit will be changed to cut off mode while the most significant bit (MSB) of don't care cells is set as high. In the cutoff mode, in order to increase the virtual ground voltage level for further leakage reduction, both ctrl1 and ctrl2 are de-asserted and the data will be destroyed. In this situation, the virtual ground can rise up to about half V_{DD} due to the stacked NMOS transistor. However, due to the power gating circuit, some speed overhead will be introduced, which will be analyzed in Chapter 5.



Mode	Ctrl1	Ctrl2
Active	1	1
Data-retention	1	0
Cut off	0	0

Fig.3.13 Multi-mode Data Retention Power Gating Devices

3.4.2 Noise Margin Analysis

The noise margin is a critical issue when designing an SRAM cell, since it is necessary to secure the stored data from being destroyed due to data access operation or noise interference. As a result, we analyze the noise margin for the newly proposed architectures, which is also important for the decision of many design factors. Decreasing the widths of the gating transistors for the multi-mode data-retention power gating technique will reduce its overhead on search time due to the loading effect and will increase more leakage power saving, but the trade-off would be the

decrease of data access performance. However, a TCAM would survive from such trade-off due to the fact that in network applications, the data access rate is far less than the time when a TCAM is used for search.

The relation between the sizes of the gating transistors, the leakage current, and the read noise margin is shown in Fig.3.14. The y-axis of the right side is the read margin, and the y-axis of the left side is the percentage that the leakage current can be reduced in cut-off mode. The scale factor n represents the scaling of power gating transistor (M1 and M2 in Fig.3.13) width. On the other hand, the static noise margin and leakage saving during data retention mode are determined by the virtual ground voltage level, which in term is primarily determined by the width of the diode-connected to transistor M3 in Fig.3.13.

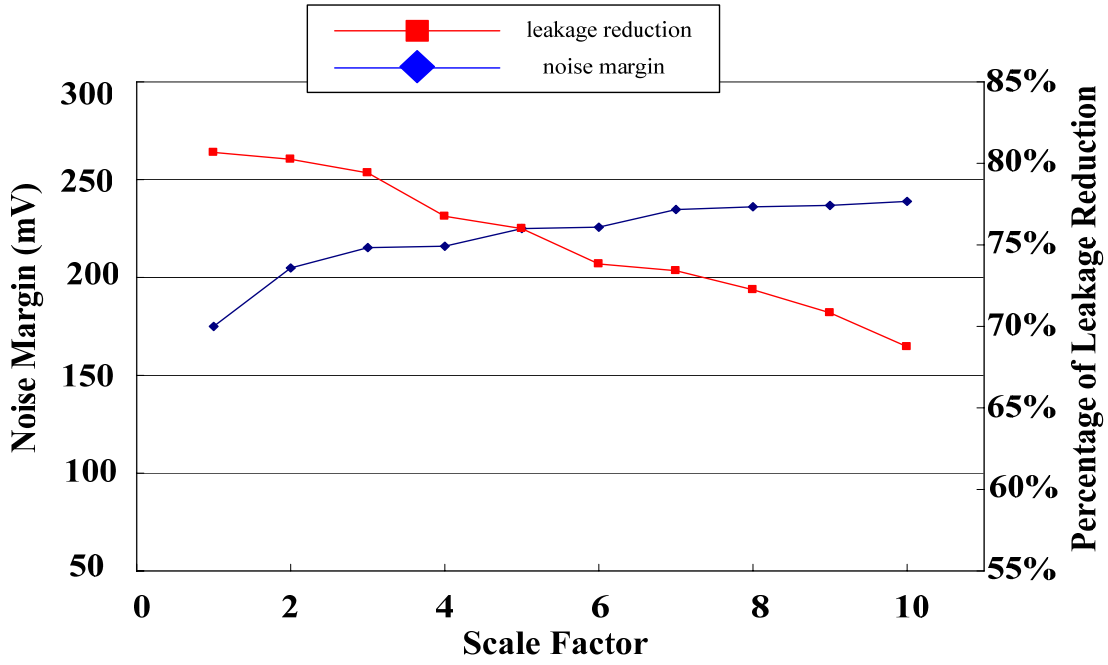


Fig.3.14 Relations between noise margin, leakage current, and scale factor

3.4.3 Simulation Result and Comparisons

The simulation result is shown in Fig. 3.15. The percentage of power reduction reaches 35% and 55% for data retention mode and cutoff mode respectively. First we compare to the type 2 and type 3 simple data-retention power-gating device shown in Fig.3.11 (b) and Fig.3.11 (c), the proposed multi-mode data retention power gating structure shows significant leakage reduction in the additional cut-off mode due to the

stacked transistor while keeping sufficient noise margin due to the diode connected transistor. Furthermore, Fig.3.16 shows an enhanced type of simple data retention power gating device, however, they all increase leakage reduction at the cost of worse noise margin, which will be more vulnerable to soft error rate and static noise.

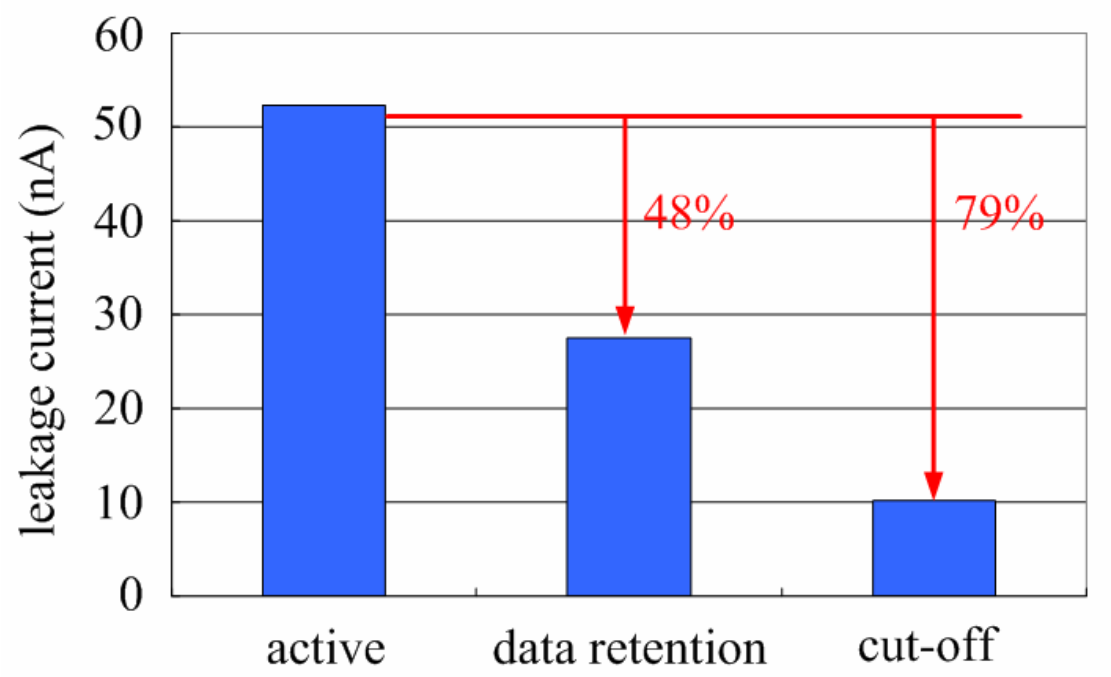


Fig.3.15 Leakage current of proposed multi-mode power gating technique in different mode

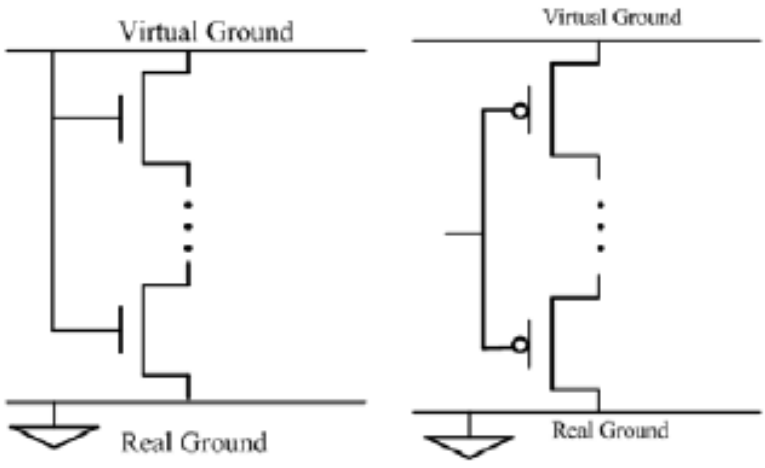


Fig.3.16 Stacked type data retention power gating technique

3.5 Summary

In this chapter, novel multi-mode data retention power gating technique is presented. Based on 65nm BPTM technology, 48% power reduction is achieved in data retention mode and 79% leakage reduction is achieved in cut-off mode while the stage is a don' care stage. Sufficient noise margin is maintained in data retention mode due to the diode connected transistor. The only trade-off is increased area, which will be analyzed in chapter 5.