Chapter 5
Super-Cutoff Power Gating TCAM Design with Leakage Current Reduction

Ternary CAM is extensively used in network applications. The main advantage of CAM is that the search time is bounded by one single memory access, thus it can guarantee a high lookup throughput. However, CAM consumes high power consumption due to fully parallel search operation and the switching of highly capacitive match-lines and search-lines.

In Chapter 4, Internet Protocol version 6 (IPV6) is introduced, and the distribution of the prefix length is shown. According to this statistics, most TCAM cells in the routing table are don't-care. In this chapter, super cutoff power gating technique is proposed based on such property of the input don't care pattern. Both the multi-mode data retention power gating technique proposed in Chapter 3 and the super cutoff power gating technique have very significant performance base on such property.

In this chapter, we start from power sources in digital CMOS circuits. The MOSFET structure capacitances are presented in Section 5.2. In section 5.3, a novel super cut-off power gating technique for TCAM cells is presented. The low power and high speed ternary content addressable memory is presented in section 5.4, and Section 5.5 shows the simulation result and the comparisons among other low power schemes. In order to design a low-power and high-speed TCAM, we modify the architecture, match-line schemes, search-line schemes of TCAM, and TCAM cells. Furthermore, the physical layout is implemented in section 5.6. Finally, some conclusions and discussions are addressed in Section 5.7.

5.1 Power Sources in Digital CMOS Circuits

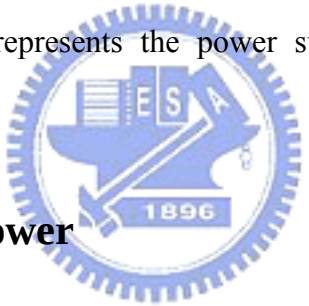
To design an energy-efficient ternary CAM (TCAM), low-power and high-speed design techniques are two definitely important issues. Here we aim at power consumption and introduce the three major components of digital CMOS integrated circuits: dynamic power, short-circuit power, and leakage power consumption, as follows.

5.1.1 Dynamic Power

Dynamic power due to charging and discharging capacitances is the dominant component among the three components of power sources, which is expressed as

$$P_{dynamic} = \alpha \cdot C_{switched} \cdot V_{DD}^{\ 2} \cdot f_{clk} \tag{5.1}$$

where α stands for the switching activity, $C_{switching}$ represents the total effective switching capacitance, V_{DD} represents the power supply voltage, and f_{clk} is the switching frequency.



5.1.2 Short-Circuit Power

The second component of power consumption is the short-circuit power resulting from non-zero rise and fall times of the input waveforms. The non-zero input rise and fall times induce a direct path between V_{DD} and ground for a short time period during switching. Short-circuit power can be given by

$$P_{short-circuit} = t_{sc} \cdot V_{DD} \cdot I_{peak} f_{clk} \tag{5.2}$$

where t_{sc} represents the time that direct path is conducting.

5.1.3 Leakage Power

The third component is leakage power. The leakage current consists of various components, such as sub-threshold, band-to-band tunneling, gate tunneling, pn-junction reverse bias, DIBL, GIDL, and punch through leakage. However, sub-threshold leakage is the dominant one that given as

$$I_{leakage} = I_0 \exp\left(\frac{V_G - V_S - V_{T0} - \gamma V_S + \eta V_{DS}}{nV_{thermal}}\right) \cdot \left(1 - \exp\left(\frac{-V_{DS}}{V_{thermal}}\right)\right) \quad (5.3)$$

where $V_{thermal}$ is the thermal voltage, n is the sub-threshold swing coefficient constant, γ is the linearized body effect coefficient, and η is the DIBL coefficient.

5.1.4 Low-Power CAM Design

Consider that the dynamic power is the dominant component of power consumption, thus it is obvious that scaling down V_{DD} is the most efficient way to reduce dynamic power by eq. (5.1). However, reducing the power supply leads to the major influence of CAM, the stability/static noise margin (SNM) reduction. In chapter 3, butterfly match-line scheme not only increases search speed but also noise margin. For IPv6 addressing lookup application, search speed is also a key point. Hence, we wouldn't make any effort to reduce the frequency factor, f_{clk}, of eq. (5.1). Because of above two reasons, V_{DD} and frequency limitation, we just reduce the others factor of eq. (5.1), switching activity and switching capacitances, to decrease the dynamic power. Finally, in terms of short-circuit power, if reducing the time of direct path conducting, the short-circuit power will be reduced effectively.

5.2 MOSFET Structure Capacitances

From the foregoing description in eq. (5.1), it has been seen that higher switching capacitances cause higher power consumptions. Therefore, understanding more about the sources of switching capacitance could help us to reduce switching capacitances. In this section, we will discuss the MOS structure capacitances which comprise gate capacitances and junction capacitances.

5.2.1 Gate Capacitances

The gate capacitances, Cg, are decomposed into two elements, overlap capacitances and channel capacitances. Before starting to introduce these two types of capacitances, we have to know that the gate of the MOS transistor is isolated from the

conducting channel by the gate oxide, which has a capacitance per unit area given by:

$$C_{ox} = \varepsilon_{ox} / t_{ox} \tag{5.4}$$

where ε_{ox} stands for the electrical permittivity of oxide and the t_{ox} is the oxide thickness.

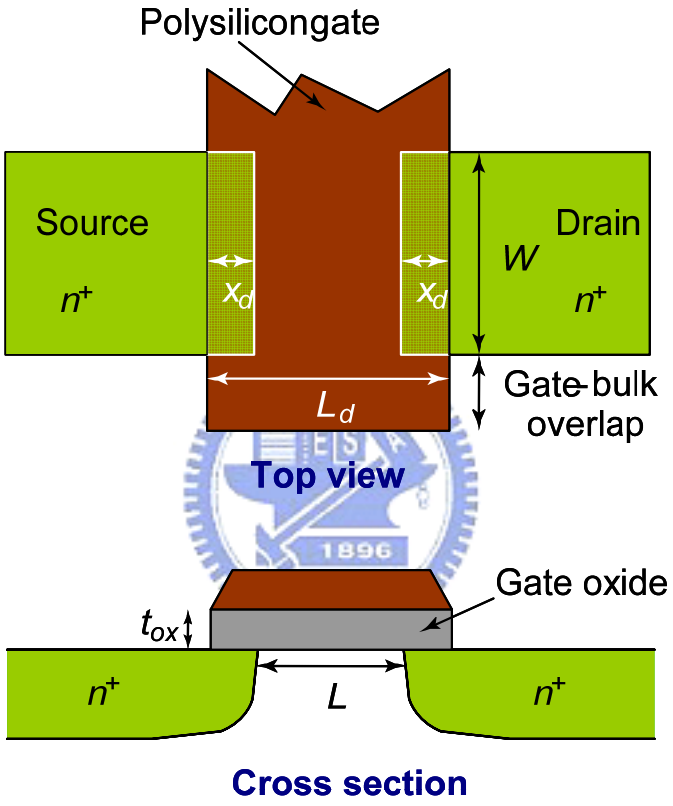


Fig. 5.1 MOSFET overlap capacitances.

5.2.1.1 Overlap Capacitances

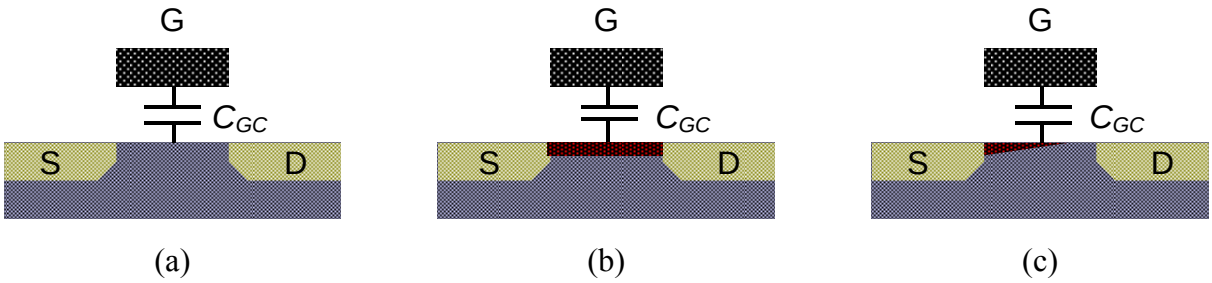
The MOSFET structure is depicted in Fig. 5.1. In real manufacture, both source and drain tend to extend somewhat below the oxide by an amount x_d, called the lateral diffusion. This phenomenon causes the parasitic capacitance between gate and source or between gate and drain that is called overlap capacitance. The value of overlap capacitance is derived as follows:

$$C_{GSO} = C_{GDO} = C_{ox} x_d W = C_o W \tag{5.5}$$

Since the x_d is determined by technology, the remain factor which can influence the value of overlap capacitance is the width of channel.

5.2.1.2 Gate-to-Channel Capacitances

The most significant MOS parasitic circuit element is the gate-to-channel capacitance C_{GC}. The gate-to-channel capacitance varies in both magnitude and in its division into three components C_{GCS}, C_{GCD}, and C_{GCB}, (C_{GCS} stands for gate-to-source, C_{GCD} means gate-to-drain, and C_{GCB} represents gate-to-body capacitances) depending on the operation regions and terminal voltages. This varying distribution is explained with the simple diagrams of Fig. 5.2. While the transistor works in the cut-off region (no channel exists) as depicted in Fig. 5.2 (a), and the total capacitance C_{GC} occurs between gate and body. In the resistive region (an inversion layer is formed) as illustrated in Fig. 5.2 (b), which acts as a conductor between source and drain. Consequently, $C_{GCB} = 0$ as the body electrode is shielded from the gate by the channel. Symmetry dictates that the capacitance distributes evenly between source and drain. Finally, in the saturation mode (the channel is pinched off) as illustrated in Fig. 5.2 (c). The capacitance between gate and drain approaches zero, and so is the gate-body capacitance. All the capacitances are therefore between gate and source. In order to make a simple analysis possible, a simplified piecewise-linear model with a constant capacitance value is adopted in each region of operation. The assumed values are summarized in Table5.1.



(a) (b) (c)

Fig. 5.2 The gate-to-channel capacitance and how the operation region influences its distribution over the three other terminals. (a) Cut-off. (b) Resistive. (c) Saturation.

Table5.1 Average distribution of channel capacitance of MOS transistor for different operation regions.

Operation Region	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Cutoff	$C_{ox}WL$	0	0	0	$C_{ox}WL+2C_oW$
Resistive	0	(1/2) $C_{ox}WL$	(1/2) $C_{ox}WL$	$C_{ox}WL$	$C_{ox}WL+2C_oW$
Saturation	0	(2/3) $C_{ox}WL$	0	(2/3) $C_{ox}WL$	$(2/3)C_{ox}WL+2C_oW$

5.2.2 Junction Capacitances

The junction capacitances (also called the diffusion capacitances) are distributed by the reverse-biased source-body and drain-body pn-junctions. The structure of source (drain) region is shown in Fig. 5.3. The junction capacitances are formed by bottom-plate junction and side-wall junction. The detail value would be introduced in following:

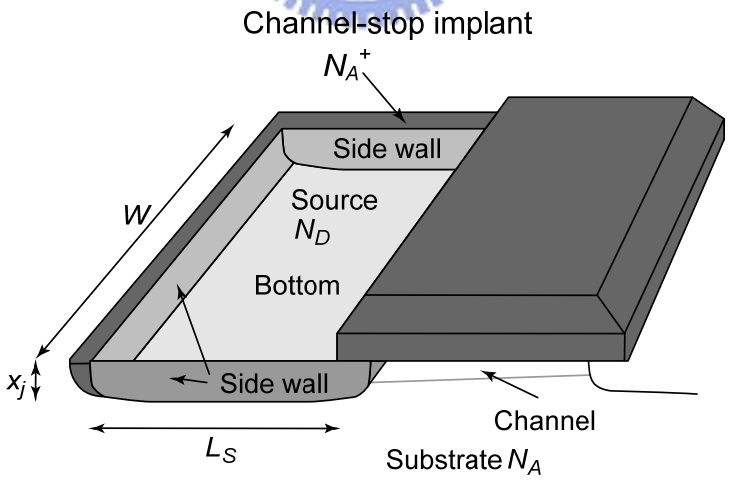


Fig. 5.3 Structure of source junction.

5.2.2.1 Bottom-Plate Junction Capacitances

The bottom-plate junction capacitance consists of the source region with doping N_D and the substrate with doping N_A. The total depletion region capacitance for this

capacitance is as following:

$$C_{bottom} = C_j W L_S \tag{5.6}$$

Where C_j is the junction capacitance per unit area given by:

$$C_j = \frac{C_{j0}}{(1 - V_D/\phi_0)^m} \tag{5.7}$$

Where the C_{j0} is the capacitance under zero-bias conditions and ϕ_0 is the built-in potential in the junction under zero-bias. Finally, the m means the grading coefficient. As the bottom-plate junction typically is of the abrupt type, the grading coefficient m is approximate equal to 0.5.

5.2.2.2 Side-Wall Junction Capacitances

The side-wall junction capacitance is formed by the source region with doping N_D and the p^+ channel-stop implant with doping level N_A^+. The capacitance value is given by:

$$C_{sw} = C_{sw}^{'} x_j (W + 2 \times L_s) = C_{jsw}(2L_s + W) \tag{5.8}$$

Combine the eq. (5.6) and eq. (5.8), the junction capacitance value is expressed as:

$$C_{junction} = C_{bottom} + C_{sw} = C_j L_S W + C_{jsw}(2L_S + W) \tag{5.9}$$

From above discussions, it has been seen that overlap capacitance, gate-to-channel capacitance, and junction capacitance are all in proportion to width of channel. Therefore, if we want to reduce the switching capacitances, the simplest way is to reduce the size of loading transistors.

5.3 Super Cut-Off Power Gating TCAM Structure

In this section, basic concepts of super cut-off techniques are introduced first. Next, zigzag super cut-off techniques will be discussed. Finally, the proposed super cut-off power gating techniques applied on our ternary content-addressable memory are presented.

5.3.1 Introduction to Super Cut-Off Technique

5.3.1.1 Concept of Super Cut-Off Technique

As CMOS technology is scaled down and the supply voltages (V_{DD}) are further decreased, the threshold voltages (Vth) should also be scaled down to prevent speed degradation. For example, decreasing Vth by 0.1 V, however, will increase the sub-threshold leakage by more than ten times. Assuming a high-performance device and one million gates in a chip, the chip leakage can reach as much as 40 mA, even in the sleep mode [104]. Leakage power is expected to be dominant in sub-70nm technology, and the large leakage current is unacceptable in most portable applications [105]. In such a leakage dominant environments, the conventional simple power gating scheme will gradually lose its effectiveness although currently it is the most effective way to reduce power consumption.

Of the existing leakage reduction schemes, the Super Cut-off CMOS (SCCMOS) can be used below 1 V supply voltages without severe speed degradation because the power switch is made with a low-Vth MOSFET. For example, the SCCMOS in [106] can suppress the leakage down to a 1 pA-order per gate when V_{DD} = 0.8 V. Although the SCCMOS successfully suppresses the sleep-mode leakage, the wakeup time is so long that it cannot be used for the active mode. In the active mode, a fast wakeup time is needed to maintain the normal operating speed. The wakeup time of the SCCMOS amounts up to several clock cycles. In addition, a high-rush current may arise at this transition. The long wakeup time and high rush current make the SCCMOS difficult to use in the active mode where the wakeup occurs frequently. If the SCCMOS is used in the active mode, the several clock cycles of the wakeup process are stolen many times and the overall performance in the active mode is degraded severely.

Furthermore, in the standby mode, V_{DD} drops to ground due to the large leakage current of the low-Vt MOSFETs. This may cause the gate-oxide reliability problem of the cut-off PMOS when thin gate oxide is used, which is especially critical in the nanometer technology era. For instance, 1.2 V is applied across the gate oxide of the cut-off PMOS in the stand-by mode at 0.8-V. In order to prevent the gate oxide from breaking down, connecting two PMOSs in series as the cut-off PMOSs is effective, as

shown is Fig. 5.4 (a). In this case, both the PMOSs work in a sub-threshold region where drain current strongly depends on V_{GS}, not V_{DS}. The drain voltage of M1 becomes 0.4V to draw the same amount of current through them if their gate widths are the same. This combination can reduce maximum voltage across the gate oxide from 1.2 to 0.8 V. The NMOS version of series connected cut-off transistor is shown in Fig. 5.4 (b).

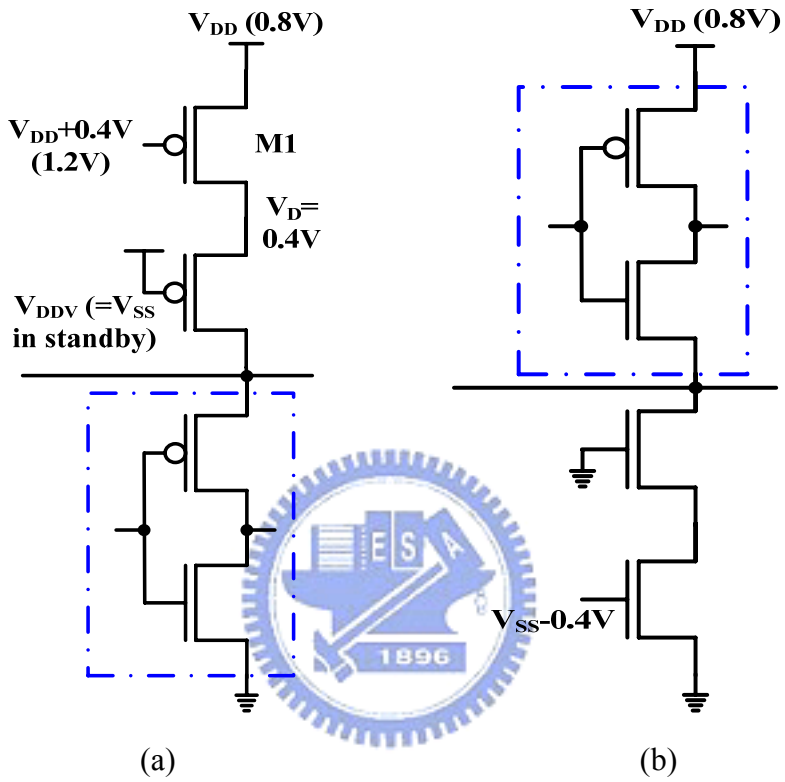


Fig.5.4 Series connected cut-off transistors (a) PMOS type (b) NMOS type

5.3.1.2 Zigzag Super Cut-Off Technique

Super cut-off power gating as a well-known power gating technique is investigated in [107] and is discussed in the previous section. It can achieve leakage power reduction significantly by efficient method [106] as indicated in the previous section. However, super cut-off technique is not desirable for DRAMs and SRAMs, where the signals connected to the cell array should preserve their data even during the stand-by mode. Besides, it suffers form a long wake-up time and a high current peak at the sleep-to-active transition. Hence, Zigzag super cut-off CMOS is proposed to improve the operation speed at the cut-off switch [108, 109, 110].

To overcome the above issues of the SCCMOS, Zigzag-Super-Cut-off CMOS

(ZSCCMOS) scheme has been proposed and this scheme successfully realizes the clock-gating scheme that saves both the switching and leakage components of power dissipation [108]. The conventional clock gating saves switching power by turning off the local clock whenever the block is not in use. For example, an MPEG-4 decoder chip reportedly saves 72% of the switching.

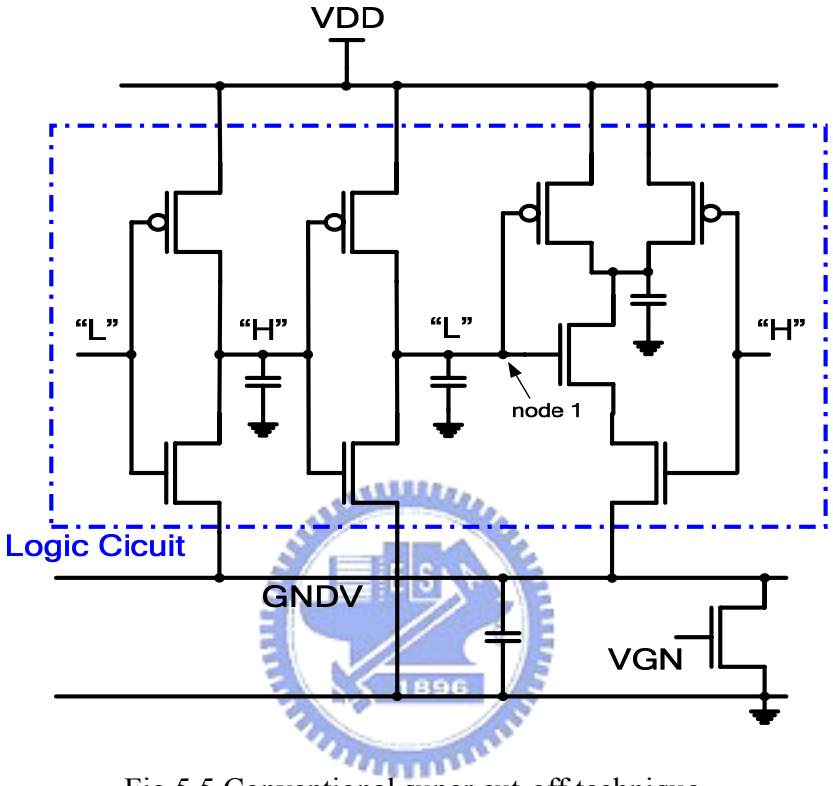


Fig.5.5 Conventional super cut-off technique

Fig. 5.5 shows a schematic of the original SCCMOS. Assume that the voltage of node 1 is initially low in the active-to-sleep transition. When the power switch MN1 is turned off by applying a gate voltage VGN = -0.3 V at t2, node (1) and VSSV go gradually to high and are pinned around VDD if the sleep time is long. This phenomenon occurs because the leakage of the power switch is much smaller than the leakage of OFF MOSFETs in the logic block. Here, the gate source voltage of OFF MOSFET is biased by 0 V. Although node 1 and GNDV in Fig. 5.5 become high in the sleep mode, they should be restored from high to low at the following sleep-to-active transition. In addition, because of the large charges associated with gates nodes and GNDV, which are restored at the next transition are large, the wakeup of the SCCMOS takes a long time and its rush current becomes large. Usually,

because the logic blocks in the clock gating are activated within one or two clock cycles, the long wakeup time prevents the SCCMOS from being merged with the clock-gating scheme. Fig. 5.6 shows the ZSCCMOS, where all the OFF PMOS are connected to a virtual VDD line (VDDV). Similarly, all the OFF NMOS are connected to a virtual GND line (GNDV). Because the virtual power lines are connected to real power lines through the power switches of MN1 and MP1, if MN1 is turned off by -0.3V, the GNDV goes up and applies a negative VGS to the OFF NMOS. The leakage of OFF NMOS, therefore, can be strongly suppressed by using this negative VGS effect. The GNDV goes up and is pinned at △V, where, as shown in Fig. 5.6, the leakage of the OFF NMOS becomes the same with the MN1. In this case, all the gate nodes should be predictable in the sleep mode because the OFF NMOS in the ZSCCMOS are connected to GNDV, whereas the OFF PMOS are separately connected to VDDV.

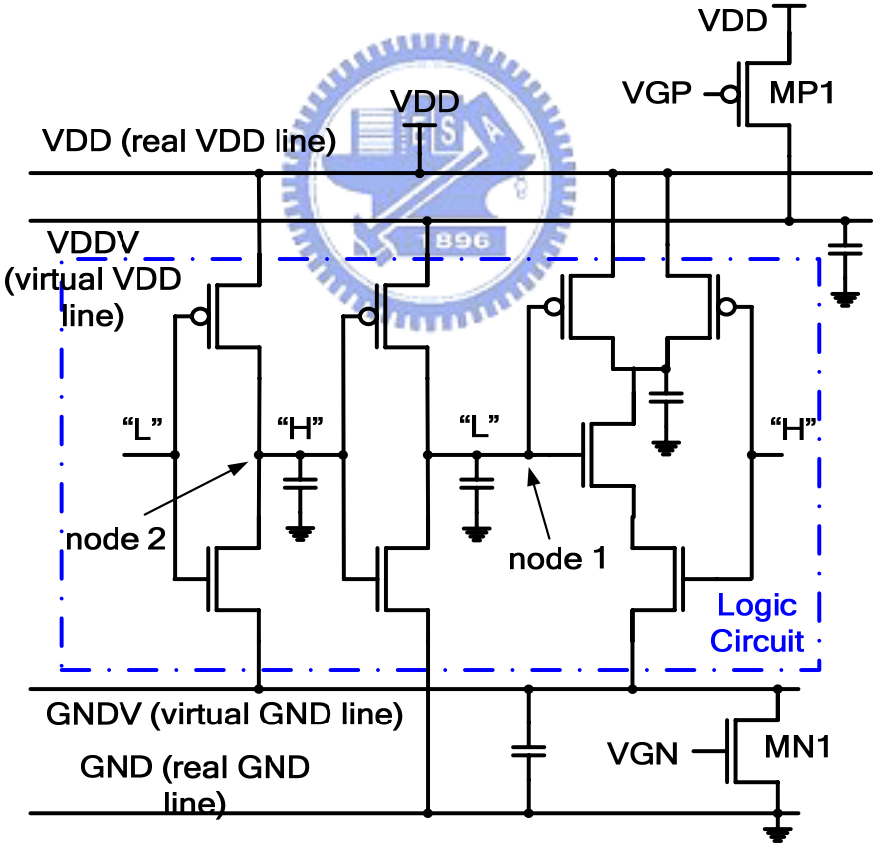


Fig.5.6 Zigzag super cut-off technique

5.3.2 Proposed Super Cut-Off Power Gating Technique

5.3.2.1 Architecture

In this section, zigzag super cut-off technique is applied to the proposed TCAM design when all TCAM don't care cells in the same segment is all set to either 0 or 1. The super cut-off power gating techniques are desirable for don't care cells without a long wake-up time and a high current peak. The circuit of super cut-off don't care TCAM cells is shown as Fig. 5.7. Several TCAM cells forming one segment will share a set of super cut-off gating transistors, which contains 2 NMOS switches (N1 and N2) and 2 PMOS switches (P1 and P2). The PMOS switches connect the virtual VDD line to real VDD. On the other hand, the NMOS switches connect the virtual GND line to real GND. When the segment is set to 1, P1 and N2 will be turned on and VDDV1 and VGND2 will be connected to V_{DD} and GND respectively. On the other hand, the overdriven voltage will be applied to turn off P2 and N1 to further cut off the leakage path. It is similar when the segment is set to 0, where P1 and N2 will be turned off and P2 and N1 will be turned on. Therefore, each segment contains four cut-off switches to achieve leakage power reduction. The cut-off switches are controlled by four control signals, ctrl_p1, ctrl_p2, ctrl_n1 and ctrl_n2 to switch between the normal on, the normal off and the overdrive voltages depending on different conditions. All control signals are controlled by level shifters and the overdrive voltage is generated by VBB generator and voltage doubler, which will be presented in section 5.3.2.2.

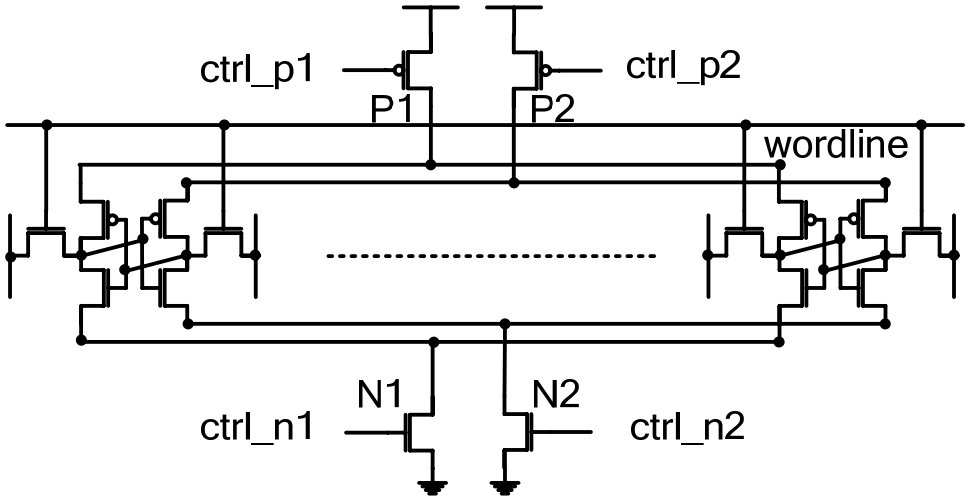


Fig.5.7 Super cut-off power gating technique

5.3.2.2 Implementation of Control Circuits and Voltage Generator

The VBB generator and voltage doubler are shown as Fig. 5.7(a) and (b), respectively. They are referred to [111] and [112]. The voltage doubler uses a dual series switch and the principle of bulk switching. M3 and M4 are series switches, and M5 and M6 swtch to the highest voltage. For M3 and M4, their bodies, output node and the chip substrate compose of vertical PnP bipolar transistor. Since M5 and M6 switch the bodies of M3 and M4 to the highest voltage, the circuit is latch-up immune. In contrast to positive voltage generator, negative-pumping circuits generate voltages lower than ground (potential = 0). The VBB generator uses a negative-pumping circuit. When clk is high, node X reaches (-Vdd + |Vtp|) and node n4 is grounded through M2. When clk goes low, node Y is pulled down to -Vdd. Meanwhile, the high voltage at node X turns on M1, and pulls output down to -Vdd.

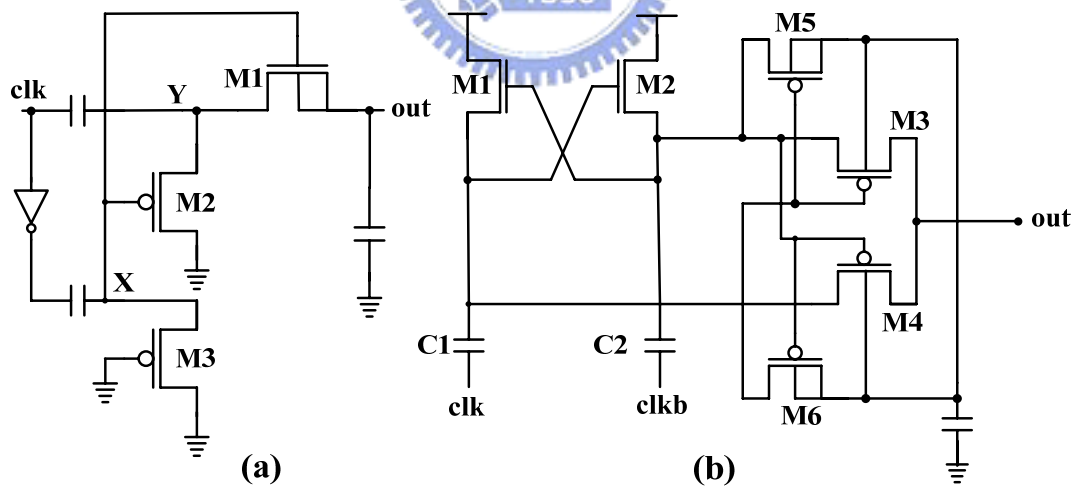


Fig.5.8 (a) VBB Generator (b) Voltage Doubler

The super cut-off control circuit between voltage generators and don't care cells is shown as Fig. 5.9 (a) and (b) are the control circuits for PMOS super cut-off switch and NMOS super cut-off switch, respectively. They are constructed by cross coupling level shifters to prevent short currents. When the sleep signal is de-asserted, the cross

coupling circuits will be isolated to output ports. And the outputs will be discharged to ground for PMOS switch and charged to high for NMOS switch for active operation. If the sleep signal goes to high, the cross coupling circuits are connected to outputs and the outputs will be evaluated by MSB. It can maintain control signals with steady voltage and avoid static short circuits which is induced by Vds > Vgs when the drain voltage is Vdd+Vcut or –(Vss+Vcut).

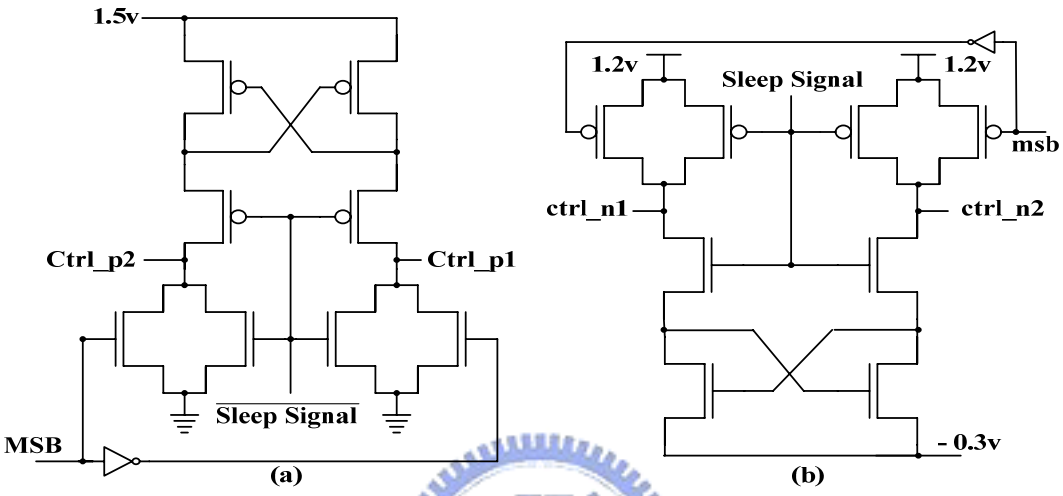


Fig. 5.9 (a) Control circuit for PMOS cut-off switch (b) Control circuit for NMOS cut-off switch

5.3.3 Simulation Result and Analysis

Due to the exponential dependence of the sub-threshold leakage on the gate-source voltage [107], super cut-off technique will reduce the sub-threshold leakage significantly. Although the more the overdrive voltage applied to the cutoff transistor, the less the sub-threshold leakage will be, two design limiting factor should be taken into consideration. First, the increased gate-drain voltage due to the overdrive voltage at the gate may increase the gate leakage, which depends strongly on the gate-source voltage. A comparison of sub-threshold leakage and gate leakage is made in [107]. Another limiting factor is the over-stress voltage appeared across the gate-source node which may cause gate oxide break down. From Fig. 5.10, the oxide stress will not exceed VDD as long as the gate overdrive voltage is less than 0.6 V. Next, due to the exponential relation of the sub-threshold leakage on gate-sourece voltage in the weak inversion region, by increasing the cut-off switch gate voltage, the sub-threshold leakage may reduces significantly. In our design, we choose the

overdrive voltage of our scheme as 0.4V. Finally, the simulation waveforms of VBB generator and voltage doubler are shown in Fig. 5.11. The transition time is determined by the VBB generator, which is about 40us. The super cut-off power gating TCAM cell structure is simulated using 65nm CMOS technology based on Berkeley Predicted Technology Model (BPTM).

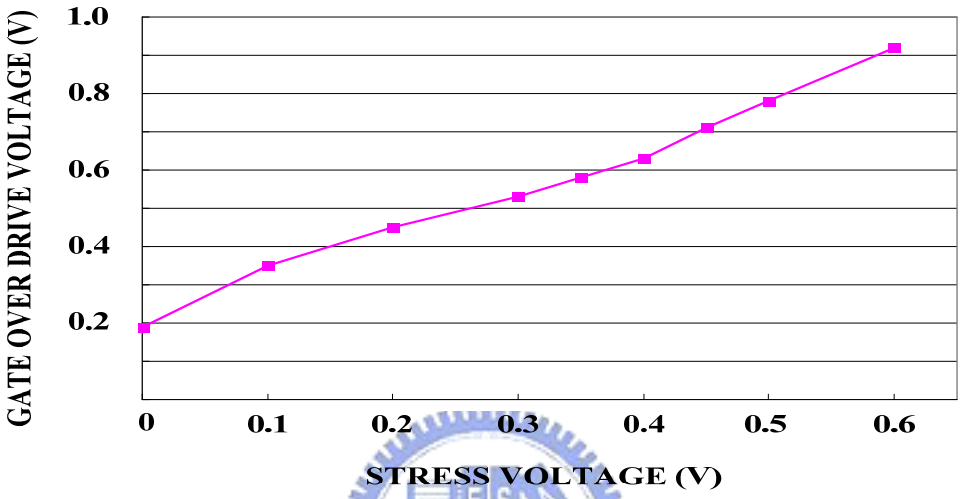


Fig. 5.10 Gate oxide stress voltage

The performance on power saving of proposed super cut-off technique does not depend on the percentage of input don't care bits. As explained in the previous section, super cut-off power gating technique can be applied on all stages except the one on the edge of input don't care patterns. As a result, nearly half of the TCAM cells is in the low power super cut-off mode, which results significant power saving.

As for cell data stability, The super cut-off technique will not degrade the cell stability during standby due to the fact that there is still a charge path from the supply voltage and a discharge path to ground for the "1" side and the "0" side respectively. Furthermore, it is worth noting that the don't care cells are not needed for read operation in any case, thus we don't have to take account of its influence on noise margin and can just focus on its effect on leakage reduction and area consideration, which will be a significant advantage to apply this technique over other methods.

As we can see from above, this super cut-off power gating technique have two main advantages. First, the cell stability is maintained and the data won't be destroyed

due to the application of super cut-off mode. Second, nearly half of the TCAM cells is in the super cut-off mode which results significant power saving. In view of this, we further apply the super cut-off mode during search operation. When technology advances and search power reduces, the leakage power may become a significant part of the total power. We use Berkeley Predicted Technology Model(BPTM), the simulation result, also shown in Fig.5.11, shows that, under 65nm CMOS technology, the search power can be reduced by 9% by placing the don't care cells in to super cut-off mode during search operation. In addition, the reduction is expected to be more remarkable when the difference between leakage power and dynamic power becomes narrower and narrower.

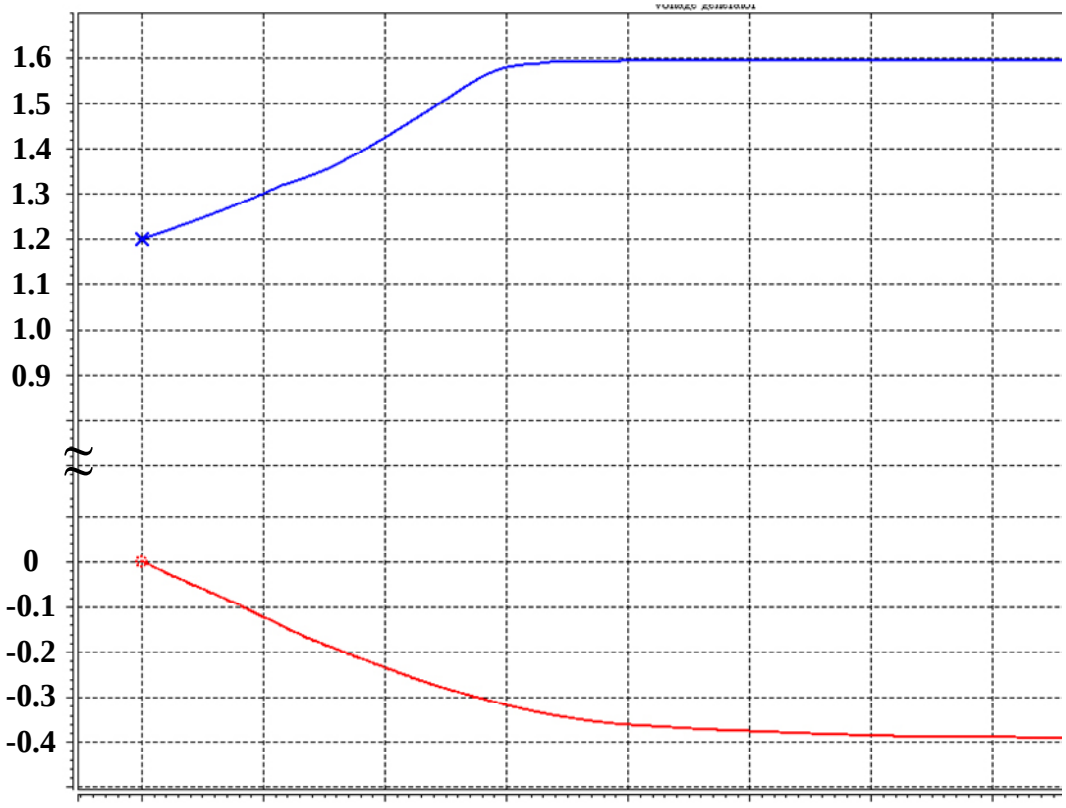


Fig. 5.11 Waveforms of VBB Generator and voltage doubler

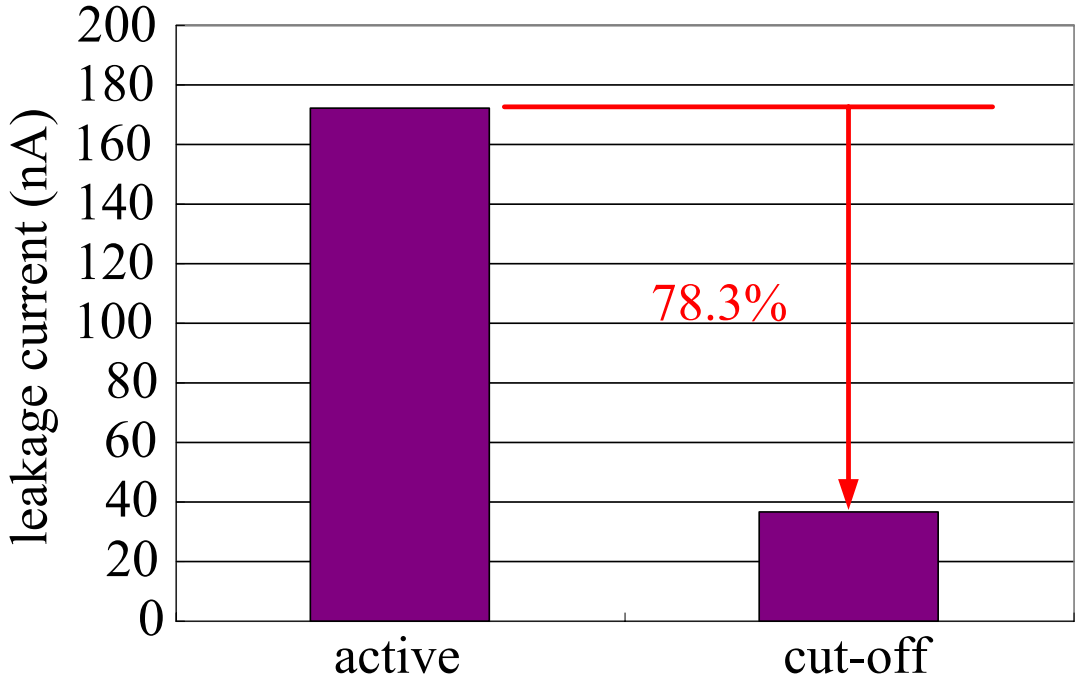


Fig. 5.12 Power performance before and after applying super cut-off power gating technique

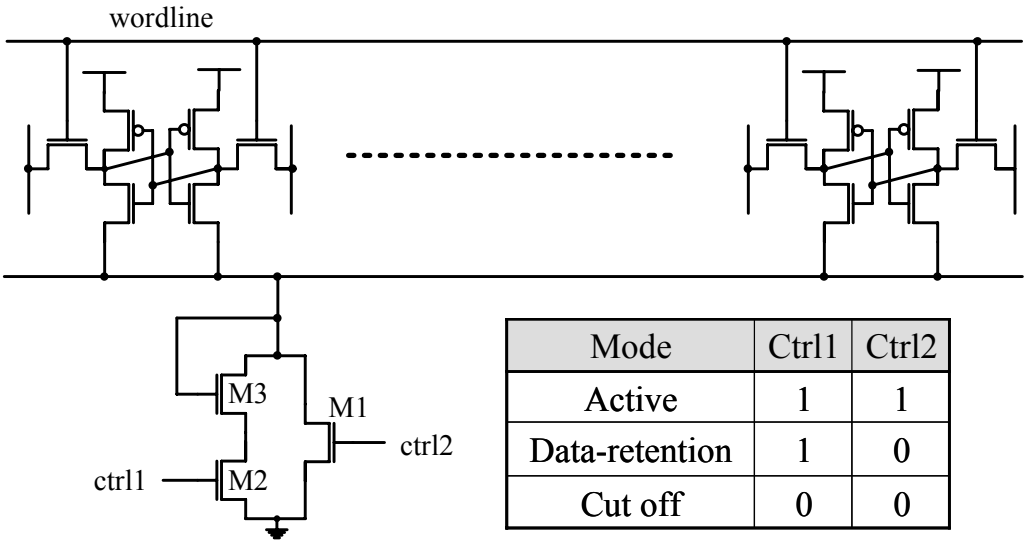
5.4 Proposed Low Power Ternary Content-Addressable Memory

In this section, the overall architecture of the proposed low power TCAM design is presented. Multi-mode data retention power gating technique is applied to achieve significant leakage reduction. We start from the multi-mode data retention power gating technique, and then the super cutoff power gating technique. Next, the overall architecture is presented. In the end, the implementation of the TCAM design and the simulation result will be discussed.

5.4.1 Multi-Mode Data Retention Power Gating Technique

The data-retention power gating circuit is shown as Fig. 5.13, 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 Figure 5. 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 cut off 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 VDD due to the stacked NMOS transistor. However, due to the power gating circuit, 16% of speed overhead is introduced.



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

Fig. 5.13 Multi-mode data retention power gating technique

5.4.2 Super Cut-Off Power Gating Technique

Although the zigzag super cut-off power gating technique is not applicable to memory cells like DRAM and SRAM, it can be applied to TCAM architecture due to the special characteristic of don't care patterns in applications like internet routers. Depending on the continuous don't care X patterns, don't care cells in a segment are all set to 1 or 0, except for the segment which is located at the boundary of don't care X patterns. On the other hand, the super cut-off power gating are desirable for don't care cells without a long wake-up time and a high current peak. The circuit of super cut-off don't care cells is shown as Fig. 5.14. When the segment is set to 1, P1 and N2 will be turned on to preserve data in don't care cells. Furthermore, P2 and N1 will be turned off to reduce leakage currents. It is similar as that the segment is set to 0, where P1 and N2 will be turned off and P2 and N1 will be turned on. Therefore, each segment contains four cut-off switch transistors to achieve leakage power reduction and preserve data in don't care cells. Under 1.0v power supply voltage, the cut-off voltage is 1.4v for PMOS and -0.4v for NMOS which are generated from VBB generator and voltage doubler as Fig. 5.8. Each word has its own VBB generator and voltage doubler for 24 six-bit TCAM segment.

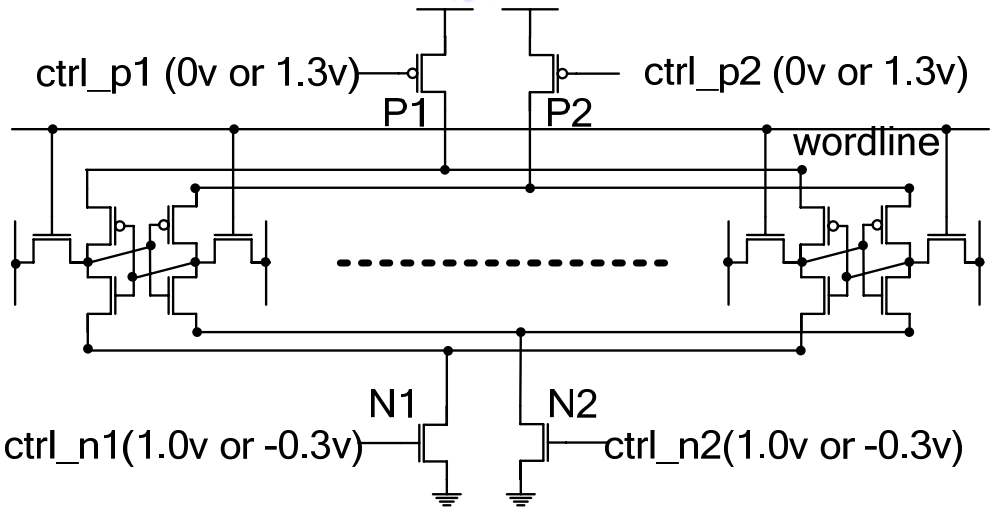


Fig. 5.14 Super cut-off power gating techniques

5.4.3 The Overall Architecture

Fig.5.15 shows the overall architecture of the proposed low power TCAM architecture. The proposed power gating techniques to achieve leakage power reduction are implemented by 65nm Berkeley PTM CMOS technology. The size of TCAM array is 256-word x 144-bit which is divided into 5 blocks for hierarchy search lines. The first three blocks consist of 64 words, and the last two blocks are composed of 32 words. Each word has its own VBB generator and voltage doubler for 24 six-bit and the overdrive voltage for NMOS cut-off switch and PMOS cut-off switch are 1.4 V and -0.4 V, respectively

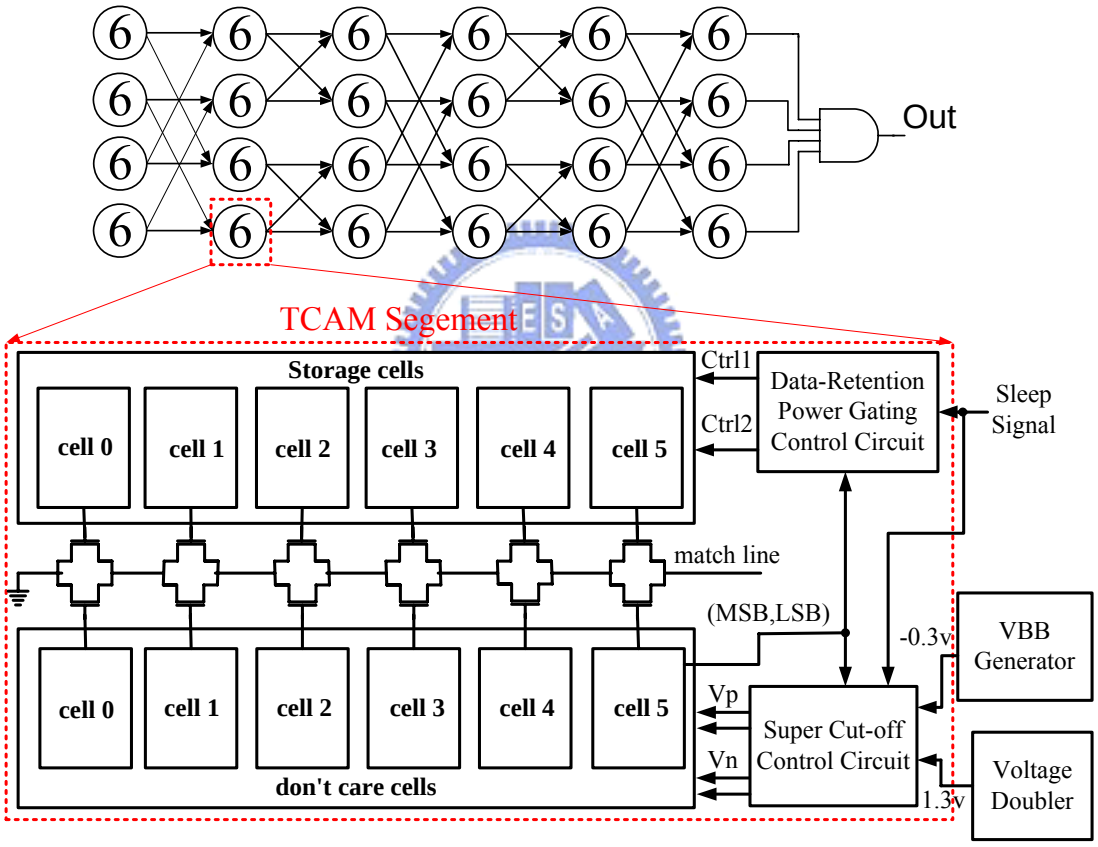


Fig. 5.15 Circuit architecture of proposed low power TCAM design

5.5 Simulation Result and Comparisons

In this section, we analyze the performance for multi-mode data retention technique and super cut-off power gating technique. Simulation result based Berkeley Predictive Technology Model (BPTM) 65nm at 1.0V are presented. The temperature

is set to 27°C. In addition, several previously proposed approaches are compared with the proposed architecture.

The performance summaries of proposed TCAM and other TCAMs/CAMs published recently are shown in Table 5.2. The normalized factors are modified from [113] and derived as follow:

$$\begin{cases} E^{*} = E \times \left(\frac{65}{\text{Technology}}\right) \times \left(\frac{1.0}{V_{DD}}\right)^{2} \\ T^{*} = T \times \left(\frac{65}{\text{Technology}}\right) \times \left(\frac{V_{DD}}{1.0}\right) \end{cases}$$

Table 5.2 Comparisons of different schemes of TCAM

Approaches / Features	Tree-style [2]	Range matching [3]	PF-CDPD [1]	Hybrid Type [4]	Hierarchy Search [5]	Shadow Matchline [6]	Proposed (with/without power gating)
TCAM/CAM configuration	TCAM 256x128	TCAM 512x144	TCAM 256x128	TCAM 1024x144	CAM 1024x144	CAM 256x144	TCAM 256x144
Technology	0.13um	0.13 um	0.18 um	100 nm	0.18 um	0.13 um	65 nm
Supply voltage (V)	1.2 V	1.2 V	1.8 V	1.2 V	1.8 V	1.2 V	1.0 V
Search time (ns)	1.10 ns	4.80 ns	2.10 ns	2.20 ns	7.00 ns	0.80 ns	0.23/0.22 ns
Energy metric (fJ/bit/search)	0.348	0.590	2.330	0.700	2.890	0.510	0.047/0.051
Normalized Search time T* (ns)	0.66 ns	2.88 ns	1.365 ns	1.716 ns	4.55 ns	0.48 ns	0.23/0.22 ns
Normalized Energy metric E* (fJ/bit/search)	0.0725	0.1299	0.1687	0.2464	0.2093	0.1062	0.047/0.051

The power performance is shown in Fig. 5.16. Fig. 5.16 (a) shows the standby power reduction of the cell array. Since the performance of multi-mode data-retention power gating technique are influenced by input don't care X patterns, power performance with different percentage of don't care bits are compared. It is shown that even without any don't care bit exist, 64.2% power reduction can still be achieved.

Furthermore, it is well worth noted that the data stability in TCAM cells will not

be affected by the applying of super cut-off power gating technique. Consider Fig. 5.14, when the right side node of the cross-coupled inverter stores logic high, CTRL_P1 will be high and CTRL_N1 will be low. Thus there will be still a charging path from V_{DD} through P1 and no discharge path to ground since N1 is at super cut-off condition. It is similar at the "0" side where there is a discharge path to ground and no charging path from V_{DD}. In view of this advantage, we further apply the super cut-off power gating techniques during search operation to reduce the search power caused by cell leakage current since the don't care cells are not needed being accessed in any case. The simulation result shows that about 9% of the total search power can be reduced as shown in Fig. 5.16 (b) based on Berkeley Predicted Technology Model (BPTM) 65nm under 1.0V supply voltage.

The leakage power at 65nm technology node overwhelmed the power overhead of the voltage generator. Further, a small degradation of search speed is incurred, which, however, can be eliminated by sizing the cell transistors. Thus, the overhead of this architecture would be the area overhead, which will be estimated in the next section. In addition, this application also eliminates the need of transition between active mode and super cut-off mode.

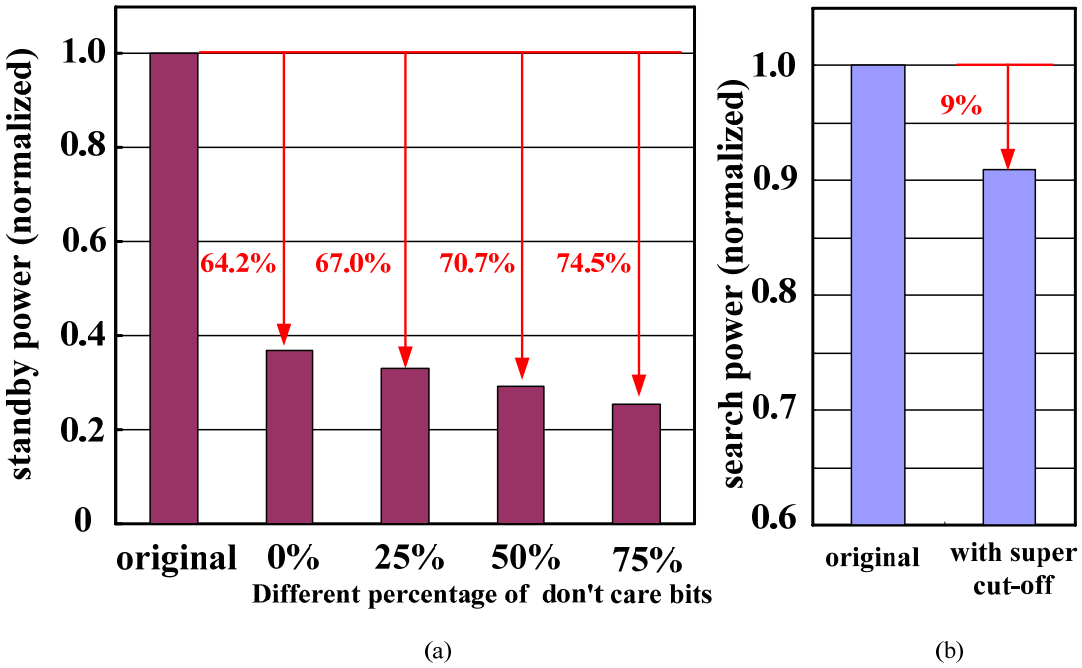


Fig. 5.16 (a) Standby power performance under different percentage of don't care bit (b) Search power performance

5.6 Layout and the Post-Simulations

In this section, physical implementation of a 256X144-bit low-power TCAM is presented. We implement the layout and post-layout simulation using TSMC 0.13um CMOS technology. Fig. 5.17 shows the layout of one-bit TCAM cell which is comprised of two SRAM cell and a comparison circuit. We separate the search line pair with the bit-line pair in order to reduce the loading capacitor to increase the operation speed; however, the tradeoff is a more complicated layout style. In Fig. 5.18, a TCAM segment is shown, it is composed of 6-bit TCAM cell, and the keeper and match-line circuit is in the right. The leftmost is the power-gating device and the super cut-off device, as well as the level shifters and some other control signals. The 256X144-bit TCAM array is shown in Fig.5.19 and the area is 1850.9 μm $\times$ 767.34 μm = 1420269.606 μm^2. Finally, the post-simulation of the proposed TCAM is summarized in Table 5.3. The power supply voltage is 1.2 V, the operation frequency is 500 MHz, the search time is 0.82ns and the power consumption is 13mW. Finally, it has a 0.289fJ/bit/search energy metric.

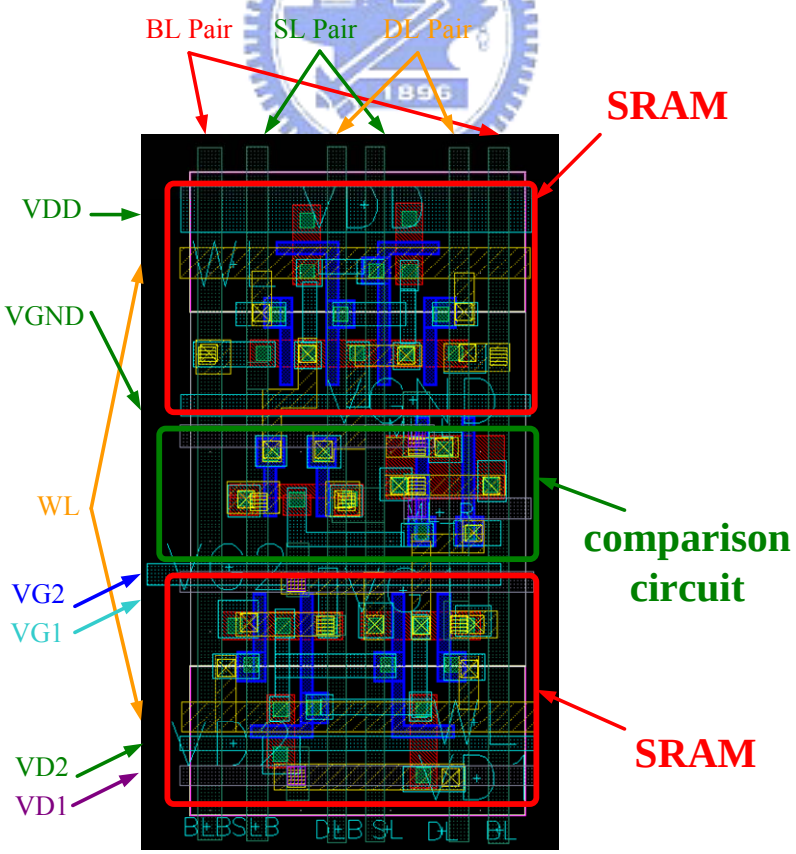


Fig. 5.17 Layout of 1-bit TCAM cell

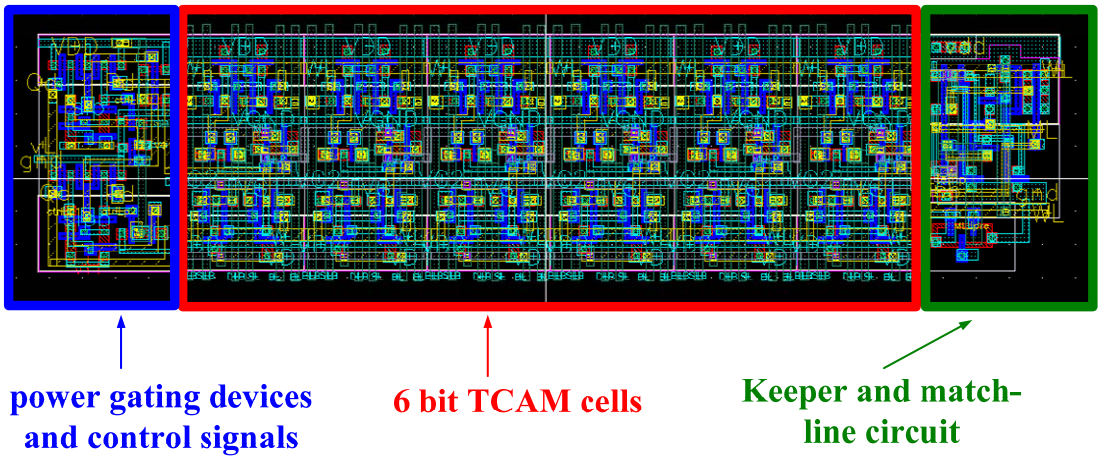


Fig. 5.18 Layout of a 6-bit TCAM segment

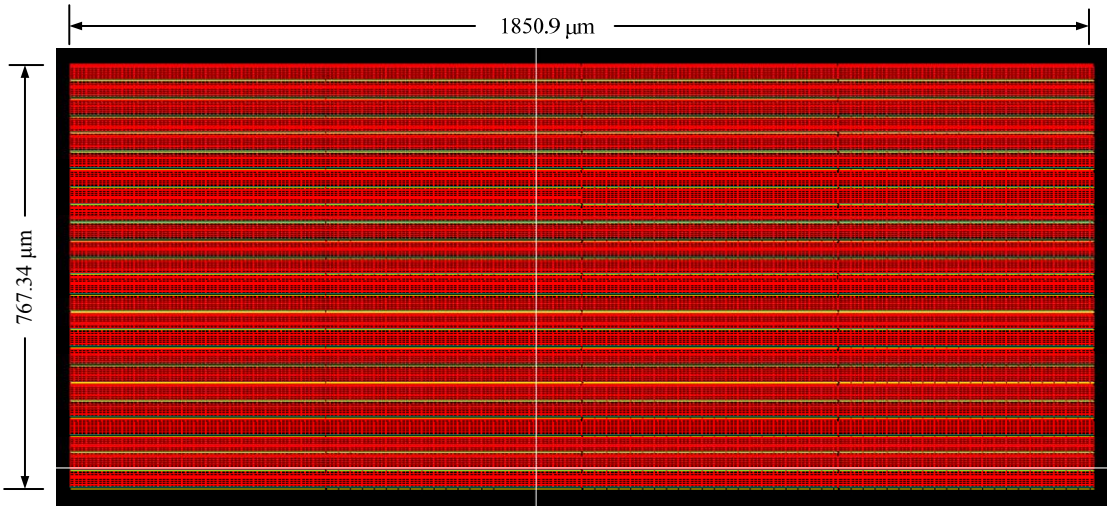


Fig. 5.19 Layout of a 256X144 TCAM array

TCAM configuration	256 words x 144 bits
Technology	TSMC 0.13 μ m CMOS
Supply voltage	1.2V
Clock frequency	500MHz
Search time	0.82ns
Power consumption	13mW
Energy metric	0.289fJ/bit/search

Table 5.3 Post Layout Simulation of proposed low-power TCAM

5.7 Summary

In this paper, super cut-off power gating and multi-mode data-retention power gating techniques are introduced for a 256-word x 144-bit TCAM. For applications in a network router, the power gating techniques are based on the continuous don't care X patterns and controlled by the most significant bit (MSB) of don't care cells in a TCAM segment. In the overwhelming majority of TCAM segments, the super cut-off power gating circuits are switched on, except for one segment. Besides, the TCAM array is implemented by XOR-based conditional keeper, butterfly match-line, don't care based power gating match-line and don't care based hierarchy search-line schemes for energy efficiency. Based on BPTM CMOS 65nm technology, simulation results show that the leakage power reduction is 70.7% and energy metric of the TCAM array is 0.043fJ/bit/search. The proposed low power schemes and power gating techniques can be very useful for TCAM, especially in further nano-scale CMOS technology with the increasing leakage currents.