



A Survey of Different Leakage Reduction Techniques

Yogesh

Research scholar, Department of Electronics and Communication, Shri Mata Vaishno Devi University, Katra-182320,
Jammu and Kashmir

ABSTRACT: Leakage power, one of the dominating issues should be made recessive with the continuous scaling of technology. Various method has been used which include work at process level, architectural level and circuit level. This paper demonstrates the systematic arrangement of leakage power, causes, and its various methods to overcome leakage reduction. High leakage current has become a significant issue in submicron technology due to a number of facts which include scaling, transistor doubling and so on.

KEYWORDS: Scaling, leakage current, static and dynamic power, switching power dissipation, sub threshold leakage, gate oxide leakage, various reduction techniques, MTCMOS, DTCMOS, Transistor stacking

I. INTRODUCTION

Power dissipation, a significant threat to deep submicron technology has become the dominating issue in very large scale integration design. The power dissipation overall degrades the parameter which includes performance, reliability, packaging, cost and portability of the system. Since we know, with the tremendous increase in transistor count on chip and similarly with the increase in operating frequency, the level of leakage increases to a greater extent. Above 100nm, the leakage current doesn't have a dominating issue but with the reduction in technology (by some parameter say "k") the leakage creates dominance. Here k is termed as scaling parameter and as per ITRS recommendation the value of k is set .7 units. The reduction in technology is done by bringing reduction in different parameter which include channel length, gate oxide thickness, and threshold voltage. Here the word deep submicron means technology using .01um to .35um range. Due to this deep submicron technology the problems faced by chip designer includes – signal integrity problems like interference from wires (crosstalk or noise, wire delays causing timing problems, chip failure due to complexity. The power consumption was about to reach the upper limits and hence mechanism for removal of waste heat is too a critical issue.

In this paper, an overview or schematic idea of different technology has been proposed in order to provide a generalized view of all technology which is used to bring reduction in leakage current. Various methods and their working have been explained in order to provide a generalized idea thus providing a platform to work in deep submicron technologies. This paper is composed in four sections. Section 1 constitute the evolution of leakage current, section 2 describes the brief description of leakage current and its types whereas section 3 describes the various techniques meant for overcoming the leakage current.

II. LEAKAGE CURRENT AND ITS TYPES

The term leakage has been divided in further two types; Static current leakage and Dynamic current leakage. It is to be remembered that static leakage is also called off static leakage current. The term static current means the leakage arising across the transistor when the transistor is turned off, and it constitute one of the major source of power dissipation at integrated circuit design level. Dynamic power exists due to charge and discharge of capacitor. Earlier it was the dynamic leakage which is considered to be dominant source of power dissipation. Reducing the supply voltage will reduce the power consumption as dynamic power is directional proportional to (supply voltage)². The average power dissipation can expressed mathematical as:

$$P = ACV^2 f + V I_{leak}$$

$ACV^2 f$ represents dynamic power loss (capacitors charging and discharging), where A represents the fraction of gates actively switching and C represents the total capacitances. The second term represents the losses due to leakage current. From above equation, if we halve the voltage the power reduces by a factor of four. But it is to be remembered that by halving the voltage processor operating frequency will reduce by more than half [1].



2.1 Dynamic Power Dissipation:

The dynamic power dissipation has been divided further into three sub-parts [13].

Dynamic power; Glitching power dissipation, Switching power dissipation, Short circuit power dissipation.

2.1.1 Switching power dissipation:

The dissipation arises because of charging and discharging of load and parasitic capacitance. Since we know for a CMOS circuit, there are lot of capacitances which include gate capacitance, interconnect capacitance, drain capacitance and so on. During the normal mode, the charging and discharging happens i.e. what we called switching power dissipation. Fig (1) shows leakage due to charging and discharging of capacitances.

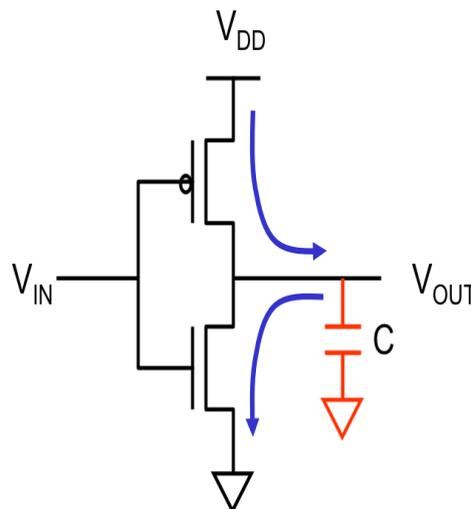


Figure 1 Leakage Power Dissipation.

2.1.2 Short Circuit power dissipation:

If the voltage applied is low(corresponding to a valid logic of 0 v), the PMOS (pullup transistor) closes and the NMOS (pull down transistor) opens and the capacitor has a charged value of $C V_{dd}^2/2$ where C is the total capacitance and is also termed as lumped capacitance .If vin now makes a transition to V_{dd} (a valid logic 1) the pullup PFET opens and the pull down NFET closes providing a low resistance path between the V_{dd} and ground; the capacitance discharges through this path until it reaches its new equilibrium states i.e 0 volts. The capacitors now will discharge to its lowest value. The energy is now converted to heat by the flow of current through resistance along its discharge path. Similarly the energy lost will take place at every node of complex circuitry and the overall dissipation is the summations of all the losses that happen within the circuitry.

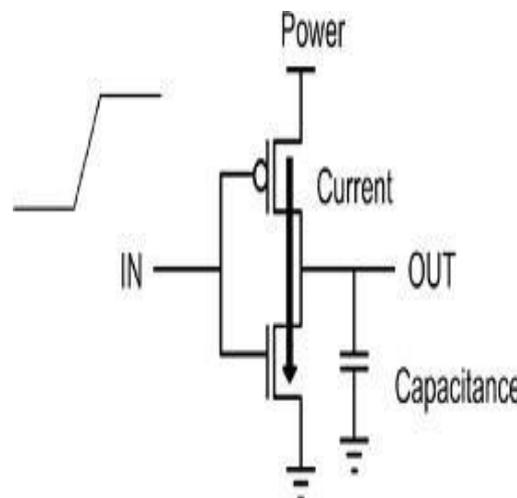


Figure (2) shows the short circuit power dissipation.



2.1.3 Glitching power dissipation:

Unnecessary signal transitions that don't have any functionality are known as glitches. Misalignment of signal transition and gate are the major source of power dissipation in digital circuits. Gate sizing, gate freezing, by reducing switching activity are some of the techniques used in order to remove glitching power dissipation.

2.2 Static Power Dissipation:

Static power dissipation is further divided as [2][3]:

2.2.1 PN junction Reverse bias current (I_1):

Leakage current arising because of the reverse bias junction connection of drain and source well. In the reverse biased condition, a minority carrier conduction takes place that is what we call reverse biased condition. Minority carrier drift/diffusion and electron hole pair generation in the depletion region of reverse biased condition are the two main reasons for I_1 .

2.2.2 Subthreshold leakage current (I_2):

When the voltage applied is below threshold voltage (V_{th}) a current flow between the drain and source. Since the leakage arises below threshold of the device i.e. it is termed as sub threshold leakage current.

2.2.3 Oxide tunneling current (I_3):

With the decrease in gate oxide thickness, the field increases and electrons movement (tunneling) takes place from the substrate to the gate and from gate to substrate.

2.2.4 Injection of hot carrier from substrate to gate oxide (I_4):

Due to increasing electric field at Si-SiO₂ interface, the electron and hole acquire enough energy and cross the interface potential barrier thus entering into oxide layer in case of short channel transistors.

2.2.5 Gate induced drain leakage (I_5):

Arises due to high electric field near the drain junction of MOS. The depletion region widens due to the accumulation of holes when the gate is biased. This widening of the depletion layer causes the increase in the field and field crowding enhancing the high electric field near the region. The minority carrier that have been accumulated or formed at the drain depletion region underneath the gate is swept laterally to the substrate completing a path for GIDL. GIDL increases with thinner oxide thickness and higher V_{dd} .

2.2.6 Punch through current (I_6):

In a case of channel length modulation, where depletion layers merge into a single depletion region. With punch through, there is a rapid increase in current with increase in drain source voltage, since it increases the output conductance and limits the operating voltage i.e. this effect is not desirable. The circuit diagram shows various circuit leakage mechanisms, fig (3) describes six short channel leakages [2].

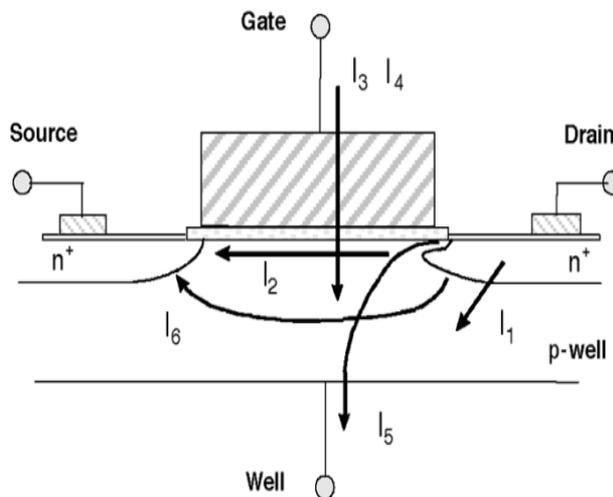


Figure (3) describe six short channel leakages [2].



Any circuit constitute of mainly combinational circuit (accounting for 30% of leakage), sequential circuit (accounting for nearly 30%), while the rest say memory and I/O device contribute to 30% of totalleakage [0].

III. LEAKAGE REDUCTION TECHNIQUES.

To suppress leakage power reduction in low voltage devices, a number of techniques are there which includes:

PROCESS LEVEL: By controlling the parameters e.g. oxide thickness, length etc.

ARCHITECTURAL LEVEL: Using parallelism, pipelining ,low frequency of operation etc.

CIRCUIT LEVEL: Voltages at device terminal are controlled. The various methods have been described as:

3.1Multithreshold CMOS (MTCMOS):

This technique introduces addition of high V_{th} in series to low V_{th} devices.The sleep control scheme is used for the management of power effectively [2]. Low V_{th} means high leakage and at the same time provides high operating speed whereas high V_{th} indicates low leakages.Figure (4) shows the MTCMOS circuitry.

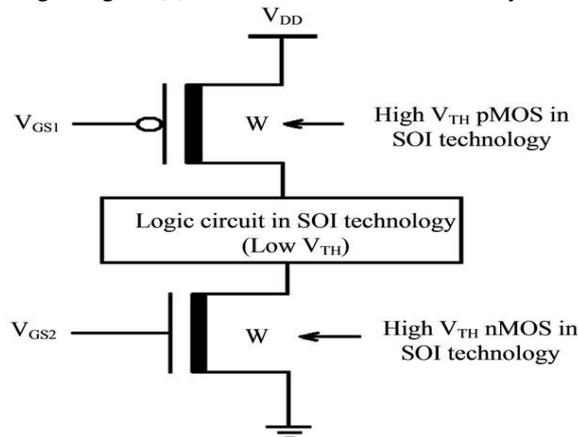


Figure (4) MTCMOS circuitry. The working of MTCMOS can be explained asunder:

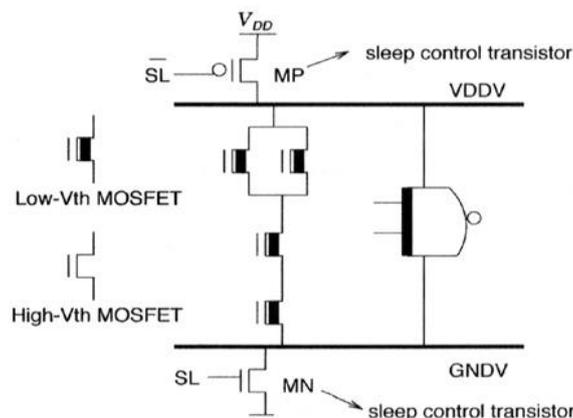


Figure (5) Schematic of MTCMOS:

In the active mode, standby mode (SL) is set low and the transistor i.e high V_{th} sleep control transistor (i.e. M_p and M_n) are switched ON. When ON resistance is low, the virtual supply voltage (V_{DDV} and V_{SSV}) act as true power lines and reduction is reduced to a greater extent. Similarly, in the standby mode, SL is kept high,which further switches OFF both M_n and M_p .Figure(5) shows the schematics of MTCMOS.

However, using two sleep control transistor the circuitry becomes complex as well as costlier.Therefore, instead of using two transistor, high V_{th} transistor issufficient for leakage reduction.NMOS insertion is generally preferred over



PMOS insertion due to low ON resistance at the same width [2]. For sequential circuits, both delay and area act as a drawback (standby mode). However, the circuit can easily be modified (advantage).

3.2 Dynamic threshold CMOS (DTCMOS):

In case of DTCMOS, the gate and body are tied together. The voltages are changed in order to achieve the required states of circuit. A high threshold voltage prevents sub threshold leakage in standby mode, whereas low threshold prevents the system performance. DTCMOS can be developed in vast quantity or in bulk by using triple well technology[2]. The PMOS body terminal is forth terminal to increase the performance of low voltage terminal and is as shown in Fig (6). The PN diode is reverse biased (body, source). This system is valid for ultralow voltage bulk CMOS circuit (0.6V and less). The circuit diagram is shown in which the gate and the body are tied together. However, the build in potential in bulk silicon technology limits the supply voltage of DTMOS.

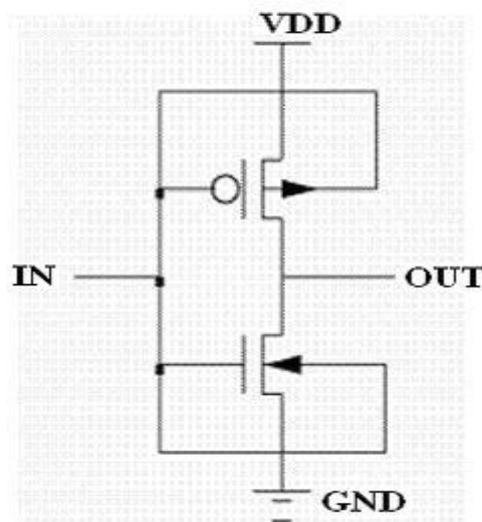


Figure (6) Schematic of DTCMOS inverter

3.3 Dual threshold CMOS:

For critical path, small V_{th} is used and at the same time for non-critical path a higher V_{th} is used. Higher V_{th} results in low leakage current. No extra leakage transistor is used for decreasing leakage and at the same time less power and increased performance is achievable. Fig(7) illustrates the dual V_{th} circuitry idea. The dual V_{th} CMOS technique helps in decreasing leakage power under both the working condition i.e. standby and active mode without delay and area consideration.

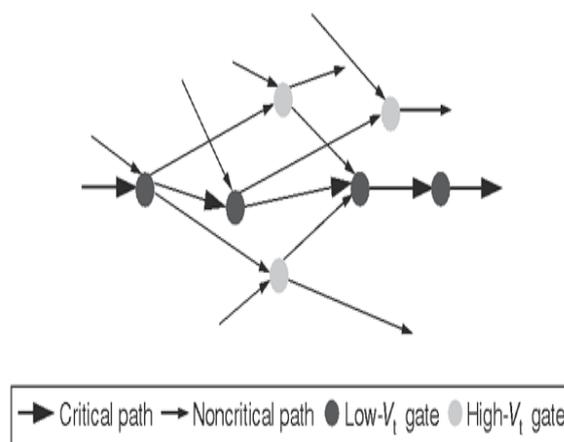


Figure (7) Dual V_{th} CMOS circuitry.

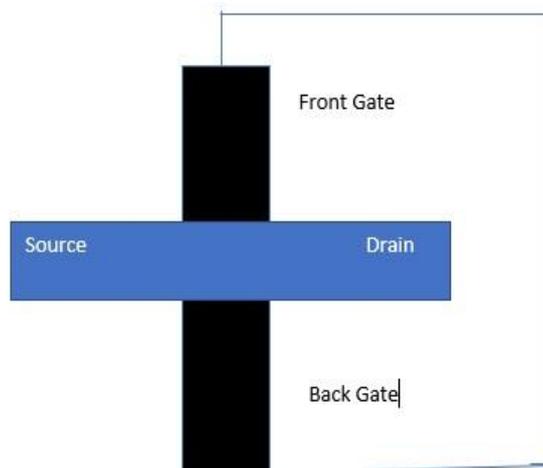


3.4 Variable threshold CMOS circuit (VTCMOS):

This technique is a body biasing technique. The body voltage is altered i.e in idle mode a zero-body bias is given, whereas in case of active mode a deeper body bias is given to increase the V_{th} which in turn reduces leakage to a greater extent [2].

3.5 Double Gate Dynamic Threshold SOI CMOS (DGDT-MOS):

Without bringing any supply voltage reduction, DGDT-MOS uses the combination of both dual threshold CMOS and double gate FD SOI MOSFET (leakage power low as well as no floating body effect i.e. controllable short channel effect)[2]. The back gate oxide thickness is increased in such a way that the V_{th} of back gate is more than V_{supply} . When both the gate potential is coupled, thus forces the front gate v_{th} to change with the back-gate Voltage. It is to be noted that the power delay (duration of switching event) product of DGDT SOI MOSFET is lower. Appreciable noise margin is seen when the voltage is reduced to .15v. Fig (8) shows the structure of DGDT SOI MOSFET.



Figure(8)Structure of DGDT-SOI MOSFET.

3.6 Data Retention flipflop (DRFF) [10]:

The term retention flip flop is also called balloon logic .flip flop store data in cross coupled inverter [1]. In power down mode, cross coupled inverter can hold their states if the input are properly gated. Consider example of data retention flip flop as shown in fig (9)[10].

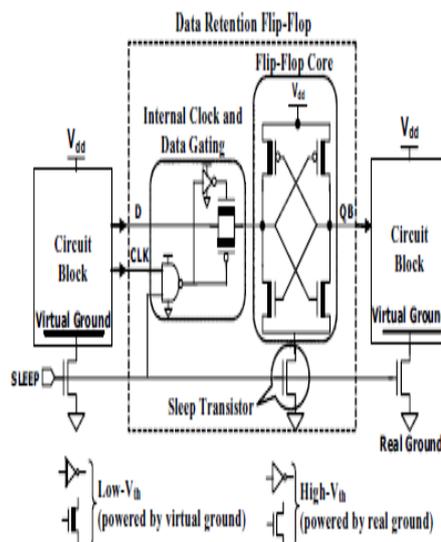


Figure (9) structures of data retention flip flop.



Here the internal clock and data gating circuit are always powered. High V_{th} transistor is provided to data gating circuit in order to minimize standby leakage current. For preserving the state of flip flop a high V_{th} latch (high V_{th} means low leakage) is used. And similarly for shorter inactive period, retention flip flop is used.

3.7 Memory leakage:

In a typical memory cell, the leakage can be bit line or cell leakage. Both the leakage arises from sub threshold conduction i.e. current flowing from source to drain when the gate to source voltage is below threshold voltage. The bit line and cell leakage can be understood from figure (10)[1].

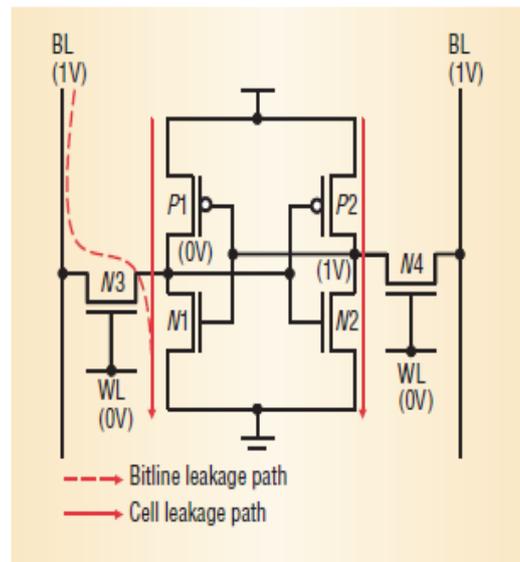


Figure (10) leakage currents and their paths in memory

The current that passes through transistor N3 is termed as bit line leakage whereas the current that passes through P1 and N2 is termed as cell leakage [1]. To aim reduction in leakage we have state destructive and state preserving technique. In case of state destructive, gated V_{dd} or grounded gating is used. In this insertion of sleep transistor is made to connect with memory storage cell and the power supply ground. By turning the cache line off saves the leakage however loss of states may occur due to incorrect turn off. This incorrect turning off causes overhead issues like performance degradation and significant power dissipation.

3.8 State preserving technique:

They are used for reducing leakage in L1 D cache as they don't lose data and don't incur extradelays. Drowsy cache i.e. information in the cache line is safeguarded by switching its V_{dd} to separate power supply. It is to be remembered that the overall difference in the consumption by these two methods is less than 10%.

3.9 Compiler technique:

The compiler too can provide application sensitive leakage control; however, these schemes too demand sophisticated program analysis, modification support and modification in instruction set architecture [1]. In this some loops are kept within the limits of bank boundaries making assumption of knowing bank structure by compilers. Some of the loops across banks are likely to be divided in case of typical large application. The (Hotspot and code sequentially) is one of the example.

3.10 Predicting transitions:

It is also termed as just-in-time activation. In this when a database uses the cache line from sleep to normal mode in each interval of working, a predictive transitioning of cache line is done in order to avoid the leakage penalty with the system [1].



3.11 Multiple threshold voltage (memory/cache):

Typically, two threshold voltages are offered in processors, whether the processor is high speed or low speed. For performance monitoring of critical transistors; a low V_{TH} is assigned by the designer. Similarly, for less timing, critical transistors a high V_{TH} is assigned. Further technologies may employ low, medium, high voltages and even more depending on the leakage reduction at different cache and their utilizations and at different level of architecture. One technique i.e by increasing the L2 cache threshold voltage will in turn decreases the leakage current [1]. To increase the L2 cache accesstime, one would have to increment the L1 cache size or instead use faster transistor. So a tradeoff is all set between parameters in order to maintain the performance. This technique does not address gate oxide leakage.

3.12 Oxide tunneling current (cache):

In addition to threshold voltage oxide tunneling current creates another significance in order to reduce cache leakage. The gate oxide thickness has significantly reduced with scaling in order to drive current at reduced voltage supplies. In order to reduce gate oxide leakage, a high (K) of dielectric constant say 30 to 50 is used such as Hafnium oxide (HfO_2) [1]. They have the tendency to reduce gate oxide leakage but they too induce process integration problems. One possible approach is to use technology with dual oxide thickness.

3.13 Super steep retrogrades and Halo implants [2, 9]:

Both are the means by which we reduce channel length and further enhance the drive current of transistor without causing an increment to the switching current. In case of retrograde doping, we change the 1-Dimensional characteristic of well profile by generating retrograde profile toward Si-SiO₂ interface. Whereas in case of halo doping a localized 2-Dimensional mixed distribution is formed near the source and drain region [2, 9]. Fig (11) shows N-channel device with super steep retrogrades and halo doping.

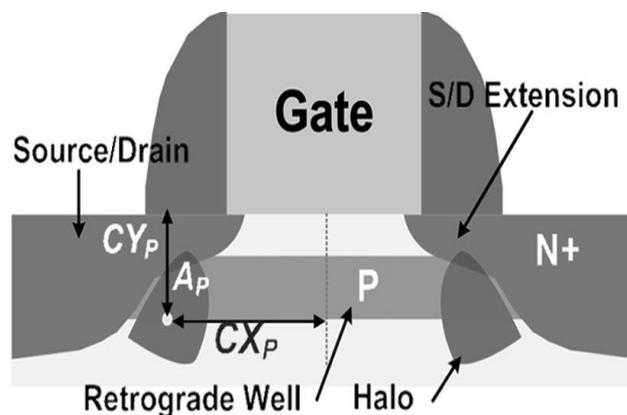


Figure (11) Nano scaled N channel device with super steep retrograde and halo doping.

Retrograde channel doping is meant to increase the SCE effect and further enhances the surface channel mobility by generating low surface channel mobility followed by high channel mobility. While in case of halo doping or nonuniform doping in the lateral was introduced below .25um technology to provide another way to control the dependence of threshold voltage on channel length here highly p type doped regions are introduced in the n channel MOSFET near the two ends of the channel. Here the doping concentration near the source and drain increases as the impurities are gathered from the substrate. The edges formed in the channel reduces the charge sharing effect from the source and drain fields, thus reducing the width of the depletion region in the drain substrate and source substrate region. With the reduction in channel length, these highly doped regions consume a large fraction of the total channel. Threshold voltage dependence on the channel length becomes more flat as seen in figure 12 (b). The higher doping near the channel causes larger BTBT and higher GIDL. The BTBT current in the high – field regions near the drain ultimately limit the halo doping level. Figure 12 (b) short channel threshold voltage roll off for retrograde and halo profile



3.14 Power switch technique [6]:

For leakage reduction among low power design techniques, the power switch technique approves better in comparison to other. Resistances i.e. R_{ON} and R_{OFF} are the main analyzing parameters in this domain. The efficiency of power switching design exhibits more than 3-decade leakage when OFF. The overall performance leads to 20% decreased area, and 40% of leakage reduction.

3.15 Standby power reduction for logic blocks [7]:

For large logical blocks, stand by leakage can be minimized by the utilization of SCCMOS with charge pump. Fig (13) shows the circuit diagram for power reduction. Here the NMOS charge pump circuit is used to create a negative super Cutoffvoltage which is coupled to an output drive circuit to change the gate voltage on the footer V_{FOOT} between the output voltage V_{OUT} in standby mode, V_{DD} in the active node. Metal insulator metal (MIM) capacitor is used to reduce parasitic capacitance [7].

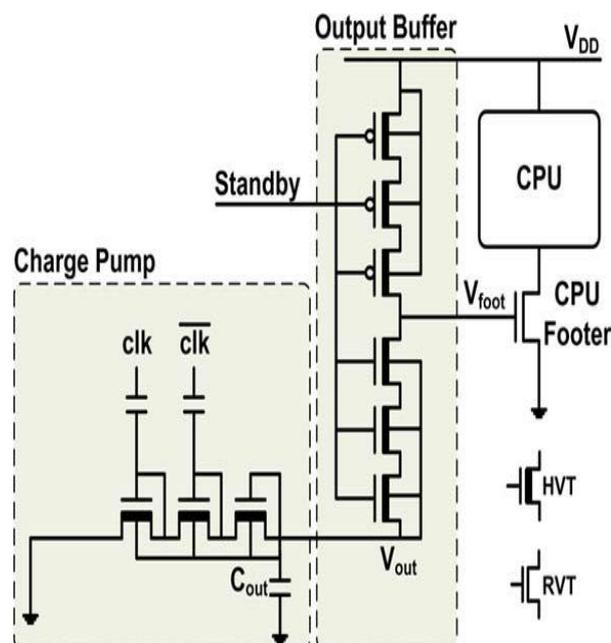


Figure (13) Reducing standby leakage current for power gated blocks.

In order to connect V_{out} with footer, a tripled stacked inverter is used. Subthreshold leakage during standby mode is minimized by PMOS stack, and when transition from standby mode towards active mode NMOS plays an important role. It is to be remembered that NMOS stack is not forward biased in active mode.

3.16 Control point insertion method[4]:

In this a control point is inserted in the circuit. The gates are chosen in such a way that gate with higher leakages are chosen first. The only problem is the area and delay penalty. It involves work at gate level as well as circuit level. These both techniques are used for leakage current estimation including both gate leakage and subthreshold leakages. Based on the predictive insertion of control point, the leakage reduces to 70% of the total leakage with minimum increase in delays and area[4].

3.17 Transistor stackeffect [8, 12]:

When no. of transistor in a stack is switched off, the sub threshold current flowing through a stack of series transistor decreases. The effect is known as transistor stacking. we have power gating, stacking single switch are the mechanisms in order to overcome the leakages. Input pattern of each gate affects the sub threshold as well as gate leakage current.



Sleepy stack approach is another technique which involves sleep transistor in active mode and stack approach in sleep mode in order to remove the leakage. Forcstacking brings another approach towards leakage reduction; here the pull down and pull up transistors are divided in two halves in such a manner that their W/L ratio is not affected. The W/L ratio is maintained after dividing the circuit. Subthreshold leakage and DIBL leakage current are reduced and enhances the performance of the circuit by applying this approach in various circuit designs. Figure (14) shows the stack effect of 2 I/P NAND gate [8].

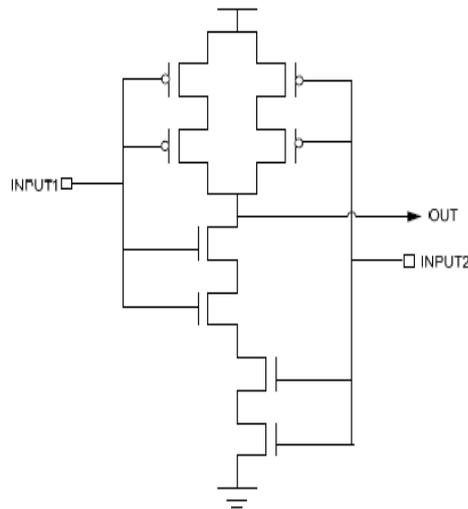


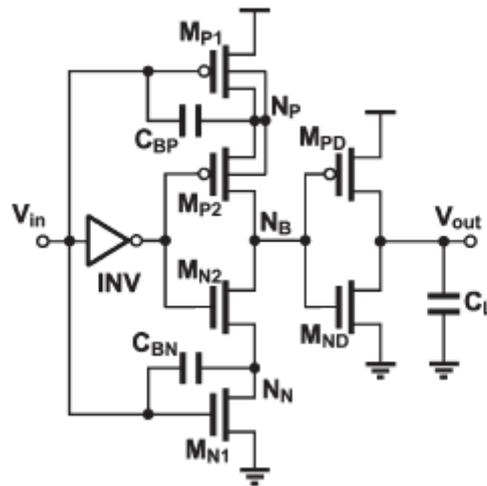
Figure (14) Stack effect in of 2 I/P NAND gate.

3.18 Pipeline gating [5]:

In this, the work is done on instruction; their fetching, decoding, and execution. A reduction is done in average activity of the processor without compromising with the speed of the processing but with a reduction in power dissipation. For a non-speculative (computer system performing some tasks that are needed) processor all the work performed is necessary. Whereas in case of non-speculative processor, an extra work is done without realizing any extra benefits and a new term arises i.e (Extra Work). If out of 130 instructions only 100 instructions are fetched then the extra work of the fetch stage is 30%. Similarly, if 120 instructions are executed then the extra work of the fetch stage is 20%. Thus, pipeline gating is done to overcome the extra work without affecting the performance of the processor. confidence estimator is used either high assurance or low assurance, estimating that the branch prediction is correct and incorrect respectively. In this a no parameter are varied which includes; branch predictor, branch estimator, stages at which gating is done. However the decision of gating can occur at fetching, decoding or at issue stages. Unresolved low confident branches are used to determine when and how long to gate.

3.19 Subthreshold Supply Bootstrapped CMOS Inverter[11]:

The subthreshold bootstrapper performs its working in subthreshold supply region. The voltage transition from $-V_{DD}$ to $2V_{DD}$ decreases the subthreshold leakage current. It has small delay time. It is an efficient means of raising the speed which further uplifts the driving efficiency. Boosting of signal in subthreshold region and minimizing the leakage current in subthreshold region lead to an increment of the driving capabilities. The input signal is boosting above V_{DD} and beneath the ground in order to drag effectively the driving capability at different node. Figure(15) presents the bootstrap circuit[11].



Fig(15) Proposed Bootstrap circuit.

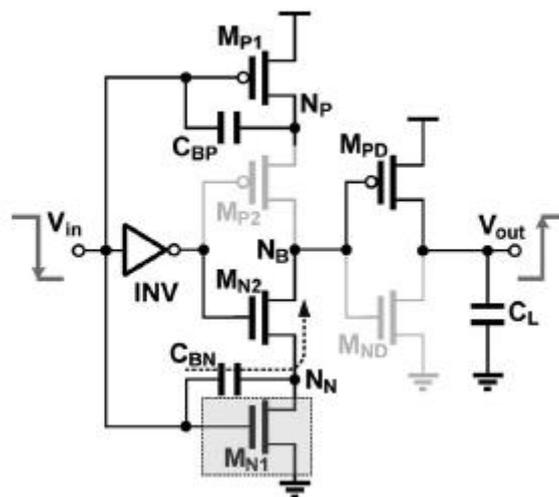


Figure15 (a) Proposed Bootstrap Circuit showing transition from high to low.

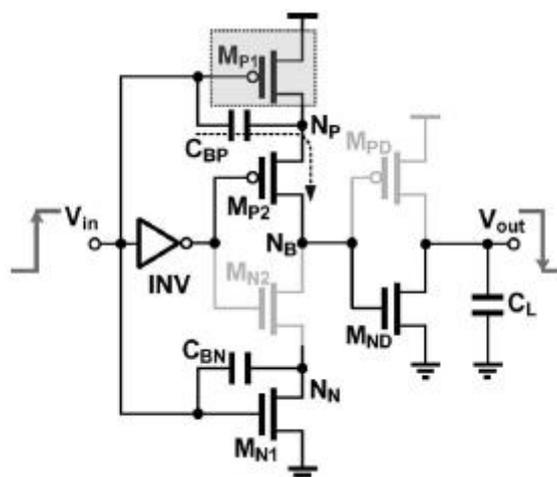


Figure15 (b) Proposed Bootstrap Circuit showing transition from low to high.



Firstly, the node N_N is boosted when the input transits from high to low. Similarly, when input voltage switches from low to high, the working is same as the transition from high to low, but here the node N_p is increased beyond V_{DD} and discharges to a value which is still greater than V_{DD} at the end of period while V_{IN} is high.

IV. CONCLUSION

In this paper, an overview or schematic idea of different technology has been proposed in order to provide a generalized view of all technology which is used to bring reduction in leakage current. Various methods and their working have been explained in order to provide a generalized idea thus providing a platform to work in deep submicron technologies. Cause of leakage too has been described. Working on deep submicron regime has been made possible by taking the history into consideration.

REFERENCES

- [1] Kim, N.S.; Austin, T.; Baauw, D.; Mudge, T.; Flaunter, K. Hu, J.S. Irwin, M.J. Kandemir, M.; Narayanan, V., "Leakage current: Moore's law meets static power," in Computer , vol.36, no.12, pp.68-75, Dec. 2003.
- [2] K.Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep submicrometer CMOS circuits," Proc. IEEE, vol. 91, No. 2, pp. 305–327, Feb. 2003.
- [3] Kyung Ki Kim; Yong-Bin Kim; Minsu Choi; Park, N., "Leakage Minimization Technique for Nanoscale CMOS VLSI," in Design & Test of Computers, IEEE , vol.24, no.4, pp.322-330, July-Aug. 2007.
- [4] Rahman, H.; Chakrabarti, C., "A leakage estimation and reduction technique for scaled CMOS logic circuits considering gate-leakage," in Circuits and Systems, 2004.ISCAS '04. Proceedings of the 2004 International Symposium on, vol.2, no., pp.II-297-300 Vol.2, 23-26 May 2004.
- [5] Manne, S.; Klauser, A.; Grunwald, D., "Pipeline gating: speculation control for energy reduction," in Computer Architecture, 1998. Proceedings. The 25th Annual International Symposium on , vol., no., pp.132-141, 27 Jun-1 Jul 1998.
- [6] Le Coz, J.; Pelloux-Prayer, B.; Giraud, B.; Giner, F.; Flatresse, P., "DTMOS power switch in 28 nm UTBB FD-SOI technology," in SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2013 IEEE , vol., no., pp.1-2, 7-10 Oct. 2013.
- [7] Yoonmyung Lee; Mingoo Seok; Hanson, S.; Sylvester, D.; Blaauw, D., "Achieving Ultralow Standby Power With an Efficient SCCMOS Bias Generator," in Circuits and Systems II: Express Briefs, IEEE Transactions on, vol.60, no.12, pp.842-846, Dec. 2013.
- [8] A.K.Dadoria , K.Khare,(A Novel Approach For Leakage Power Reduction Technique In 65NM Technologies).International journal of VLSI design and communication system , vol.5,no.3,june 2014.
- [9] Agarwal, A.; Kunhyuk Kang; Bhunia, S.; Gallagher, J.D.; Roy, K., "Device-Aware Yield-Centric Dual-Vt Design Under Parameter Variations in Nanoscale Technologies," in Very Large Scale Integration (VLSI) Systems, IEEE Transactions on , vol.15, no.6, pp.660-671, June 2007.
- [10] Mahmoodi-Meimand, H.; Roy, K., "Data-retention flip-flops for power-down applications," in Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on , vol.2, no., pp.II-677-80 Vol.2, 23-26 May 2004.
- [11] Yingchieh Ho; Chiachi Chang; ChauchinSu, "Design of a Subthreshold-Supply Bootstrapped CMOS Inverter Based on an Active Leakage-Current Reduction Technique," in Circuits and Systems II: Express Briefs, IEEE Transactions on , vol.59, no.1, pp.55-59, Jan. 2012.
- [12] N. Saxena, S.Soni, Leakage current reduction in CMOS circuits using stacking effect, International Journal of Application or Innovation in Engineering & Management Volume2, Issue11, November2013.