

A Comparative Study of Leakage Reduction Techniques Used In Fpga for Optimized Area and Power Consumption

Shivani Saxena¹, Anjali Tiwari², Ghata Chauhan³, Nidhi⁴, Richa Chaturvedi⁵

ABSTRACT

Over the last decade Field Programmable Gate Arrays (FPGAs) have become one of the key in digital circuit implementation. A fundamental part of their construction lies in their architecture, which governs the nature of their programmable logic functionality and their programmable interconnect. A remarkable effect of the quality of final device's performance in FPGA devices are speed performance, area efficiency and power consumption. But due to the integration of devices increases day-by-day or reduced size leads to the problem of leakage. Optimized area and power consumption is only achieved by reducing leakage energy in the design of FPGA which are 90nm and below. This survey reviews the fundamental programming technologies that the programmability built on for leakage reduction in FPGA, and then describes the basic understanding of all the factors that play a significant role in the occurrence of leakage and their reduction.

KEYWORDS: Field programmable gate arrays (FPGAs), area efficiency, leakage energy, threshold voltage.

I. INTRODUCTION

Due to technology scaling the leakage power is increasingly contributing in total power dissipation. In modern IC processes, there are two forms of leakage power namely, **subthreshold leakage** and **gate leakage**. Subthreshold leakage or subthreshold drain current is defined as the current flowing between the source and drain of a MOS transistor when it operates in subthreshold region or weak inversion region i.e. the gate to source voltage is below the threshold voltage (V_{TH}). With each process generation the supply voltages are reduced and hence the V_{TH} is also reducing to improve the performance. But reduction in V_{TH} leads to exponential increase in subthreshold leakage. The down scaling of device size also leads to very thin gate oxide resulting into generation of tunnelling current through the oxide of the MOS transistor. The tunnelling mechanisms responsible for gate leakage are **Fowler-Nordheim tunnelling** and **direct tunnelling**. The gate leakage increases exponentially with reduction in oxide thickness, hence limiting the down scaling to about 1.5-2 mm. due to thin gate oxide the short channel effect is reduced and the driving capability of a MOS transistor is improved. Thus, resulting into tradeoffs between this benefit and the gate leakage.

Now-a-days, Field programmable gate arrays (FPGAs) are the dominant choice for digital circuit design due to high speed, density, short design cycle and decreasing cost. It has been shown that power consumption by largest FPGA devices is increasing for FPGAs operating in low power mode (stand-by mode) and high power mode. However, the prior concern was with dynamic power consumption and the leakage power was considered as a small component of total power, since those analyses were based on IC technologies with feature sizes of 0.15 μ m or larger making them outdated for today's state-of the-art FPGAs, fabricated in 90nm technology.

Unlike ASICs, the leakage power is more in FPGAs due increase in transistor count and FPGA circuit design using a fraction of the FPGA's resources. Leakage power dissipation occurs in *used* and *unused* part of the FPGAs. The leakage problem is considered as a major obstacle for FPGAs used in both high performance and low power designs. Hence, it is necessary to focus on leakage power optimization including power optimization techniques for FPGAs i.e. used to reduce the dynamic energy. Various techniques are proposed for leakage reduction used for significant power consumption and area optimization. This paper includes a comparative study between some of the techniques and their basic concepts.

COMPARATIVE STUDY OF LEAKAGE REDUCTION TECHNIQUES

TECHNOLOGIES	BASIC CONCEPT	FEATURES
1. SLEEP TRANSISTORS Used in power gating techniques for	It is used to reduce sub threshold and gate leakage components by switching off the power supply of the	Size of sleep transistors should be large for better performance leading to area penalty.

90nm technology with cluster base designing.

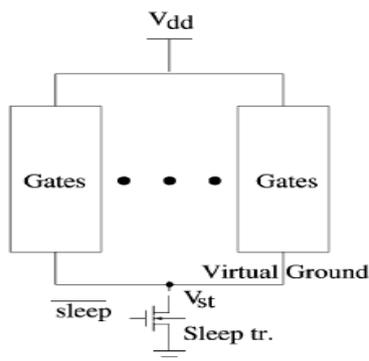


Fig. 1. Illustration of MTCMOS circuit structure [9].

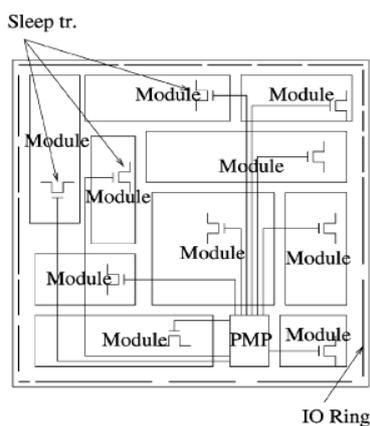


Fig. 2. Its application on the system level [9].

circuit by using sleep transistors with high threshold voltage between the supply and the circuit.

The efficiency of the sleep transistors is given by the ratio of drain current in on/off state and increases with gate length, reaches peak at 130nm and hence sub-threshold leakage is reduced.

The efficiency also depends on body bias as reverse body bias increases the value of V_{TH} and hence reduces the sub-threshold leakage.

It saves 70.7% of the total power than the clustering technique.

With body bias of 1.6V the efficiency increases by 40% as compared to normal body bias where n-well is connected to V_{DD} = 1V [10].

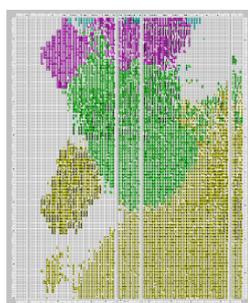
The area penalty varies from 2% to 6% depending on how the sleep transistor is designed and implemented.

Limitations –

The size of these sleep transistors is large.

Large power-on current rush occur when the design comes out of the sleep mode and charged by the sleep transistor at the same time causes voltage drop resulting into malfunctioning and induced noise.

2. REGION CONSTRAINT PLACEMENT



(a) Traditional

FPGA fabric is divided into regions each controlled by sleep transistors independently by selecting suitable region size area overhead can be controlled and leakage can be reduced.

It provides minimum number of regions used for a given application.

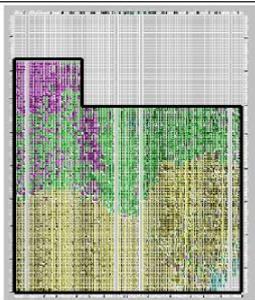
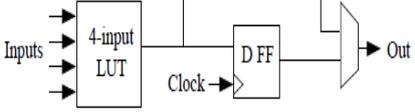
It is used for 90nm and sub-100nm technology.

Area of the combinational logic block arrays with sleep transistor area overheads is reduced by 5% from moving region with four logic slices to region with 256 logic slices.

Power saving using RCP technique on FPGA with coarse grain regions is larger than FPGA with fine grain regions using normal placement technique.

The maximum power saving is limited by the volume of unused regions.

The leakage in Xilinx FPGA,

 <p>(b) RCP</p> <p>Fig. 3. Different placements for an example design [1].</p>		<p>SRAM configuration is reduced by 98% by increasing the threshold voltage and configuration time by 20%. [1]</p> <p>With area of the slice assuming to be 5000 sq. micron, the area penalty for region of four slices is around 15% while that for 256 slices is 10%.</p> <p>Limitations- Power savings is only obtained in unutilized portion of FPGAs.</p> <p>It is inappropriate for systems occupying large space of FPGAs.</p>
<p>3. REGION CONSTRAINT PLACEMENT with TIME BASE CONTROL</p>  <p>Fig. 4. Each module is bounded by a polygon [1].</p>	<p>While it is easy to implement the RCP technique, it may not be effective in designing the system that occupies large portion of the FPGA space hence resulting into potential leakage. Therefore, for designing the system with modules that remains inactive for significant time duration, the time based control scheme is used.</p> <p>This scheme is used with two variants of RCP:-</p> <ol style="list-style-type: none"> 1. Module level RCP 2. Design level RCP 	<p>Power saving is obtained even in utilized portions of the FPGA by turning on/off the power supply and using the idleness in portions of the design.</p> <p>The gate voltages of sleep transistors are controlled by configuration bits and these control bits are dynamically reconfigured for turning on/off the transistors.</p> <p>Reconfiguration time is $2\mu\text{s}$ for smallest FPGA and $23\mu\text{s}$ for largest FPGA. It increases the area overheads because of the configuration circuits [1].</p> <p>Only 62% of the slices are used in time based control.</p> <p>Limitation- Synchronisation with time is the main problem.</p>
<p>4. LOOK UP TABLES (LUTs)</p>  <p>(a) Basic logic element (BLE)</p>	<p>This technique includes SRAM that is used to implement the truth table. Each address in the SRAM represents input combinations to the logic elements. The value stored in the address represents value of function for that address.</p> <p>An n input functions requires SRAM with 2^n locations.</p>	<p>The active leakage (power consumed in awakened circuit blocks) is reduced 25% on an average.</p> <p>In high performance 1% increase in delay occur as compared to conventional LUT.</p> <p>In low power mode leakage is reduce up to 53%.</p>

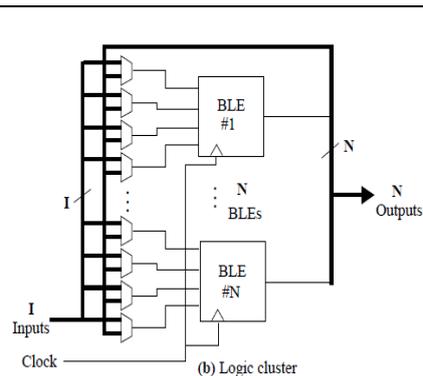


Fig. 5. Structure of BLE and Logic Cluster [6].

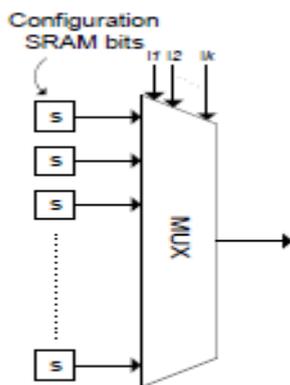


Fig. 6. Abstract view of LUT [5].

In cluster-based logic blocks, the placement problem reduces and speed is further increased.

The complexity of LUTs increases exponentially with number of inputs, therefore, instead of using large logic blocks, several LUTs are grouped together with local routing interconnection. This is called Logic Cluster.

For i inputs and o outputs of a cluster, the relation between basic logic elements (BLE), say n is $i = 2n+2$, where each cluster has n BLEs and each possessing a single LUT.

A novel FPGA reduces the leakage by another 27% in low power thus reducing total leakage with 80% when LUT is outputting logic '1'.

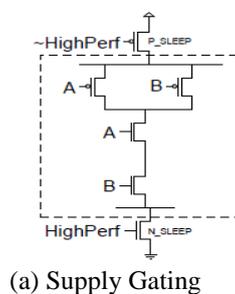
LUTs can be used to encode Boolean functions, in image processing, and as a key component in modern FPGAs.

Limitation – Due to increasing trend of new commercial FPGAs using large size LUTs, the total leakage power is increased. [5]

5. LOW POWER LUT DESIGN : HEADERS AND FOOTERS

In this technology the supply voltage across the inverter is reduced and therefore the leakage through input inverters and the pass gate structures are also reduced.

The diode header LUT has 1% increase in delay while diode header and footer have 3% increases in delay in high performance mode [5].

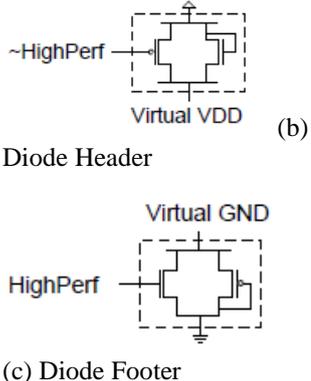
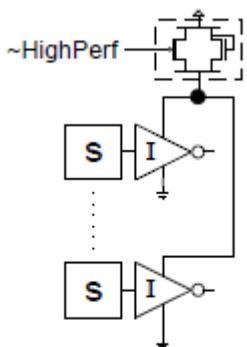
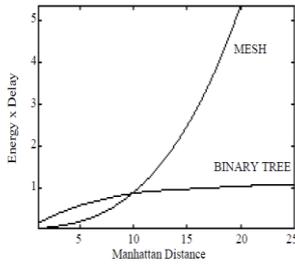
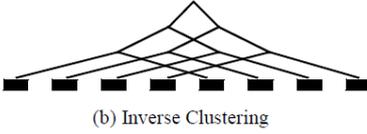
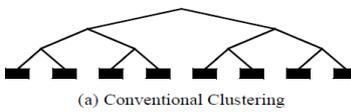


The threshold voltage drop due to diode headers increases the rise time slightly.

In low power mode the diode header LUT has a 7% increase in delay and diode header and footer has 24% increases in delay [5].

To minimize the performance penalty in high performance mode transistors are resized and the performance of these alternate designs is designated as ADH (alternate diode header) and ADHNF (alternate diode header and footer).

Due to diode header LUT area is increased by 2% and leakage is reduced by 53% when the output is either logic '0' or logic '1', whereas due to diode header and footer 10% increase in area occurs and leakage is further reduced i.e. for logic '1' 53% leakage reduction occur

 <p>Fig. 7. Circuits for Supply Gating, Diode Header and Diode Footer [5].</p>  <p>Fig. 8. Apply diode header to LUT [5].</p>		<p>and for logic '0' it is 80% [5].</p> <p>In high performance mode there is no performance penalty for ADH and ADHNF and in low power mode there is 4% and 19% for ADH and ADHNF designs respectively [5].</p> <p>For ADH 8% increase in LUT area occurs while for ADHNF it is 20% [5].</p> <p>The leakage has increased slightly in ADH and ADHNF due to increase in area. Hence, the ADH reduces leakage by 52% regardless of output state. For ADHNF it is 47% and 77% when output is logic '0' and logic '1' respectively [5].</p>
<p>6. INTERCONNECTION TECHNIQUE</p>   <p>(b) Inverse Clustering</p>  <p>(a) Conventional Clustering</p>	<p>In this technology the interconnection is optimized without degrading the routability of the architecture</p> <p>The term interconnect include all the components which contributes to provide connection between logic blocks i.e. the connection boxes metal routing and the switch box.</p> <p>The connection between CLBs is obtained through three levels of architectures: Level 0- nearest neighbour Level 1- mesh architecture Level 2- Hierarchical Architecture [6].</p>	<p>Interconnects consume most of the energy while the logic consumes only 5% of the total energy.</p> <p>Among the three levels of interconnect architecture, level 3 uses inverse clustering and is used for longer interconnection.</p> <p>As for interconnection length (l), delay increases as l^2 and energy delay by l^3. And it uses both mesh and tree structure reliable for short and long interconnection respectively.</p>

II. CONCLUSION

Our work demonstrates various technologies proposed for leakage reduction along with the dynamic power optimization for improving the area and speed performance of FPGAs. The very first technology uses sleep transistors in which the unutilized portion of the system is in sleep mode i.e. no power is supplied in that portion. However this technique has some of the limitations related to sizing and current rush occur when the design portion comes out of sleep mode resulting into noise addition. The other technique is called region constrained placement (RCP) in which the FPGA fabric is divided into different regions and they are controlled independently by configuring control bits. But, it provides leakage optimization in only unused portion of the design. Hence, the concept of time based control is combined with RCP which reduces leakage power in used portion of the design by turning on/off of the power supply, thereby utilizing the idleness of the various portions of FPGA. The problem of synchronization may arise in this technique. The next technique include look-up tables (LUTs) using SRAM to implement the truth table. It reduces the routing area and is most commercial. For low power designs, headers and footers are used which further reduces the area and leakage through the input inverters since supply voltage across it is reduced. However, their use increases the rise time slightly. The next running technique for leakage reduction is based on reconfiguring the interconnect structure as it is responsible for most of the energy consumption (65%) [6]. Among the three levels of interconnect architecture, Level 2 i.e. the hierarchical interconnect is the most advanced architecture incorporating both mesh and tree structures. The grouping of logic blocks for this structure is done as an inverse cluster. These techniques are currently applicable for designing and implementing embedded and portable applications.

REFERENCES

- [1] A. Gayasen, Y. Tsai, N. Vijaykrishnan, M. Kandemir, M.J. Irwin. "Reducing Leakage Energy in FPGAs Using Region-Constrained Placement". In *Proceedings ACM Intl. Symp. Field programmable gate arrays*, 2004.
- [2] J. Kao, S. Narendra and A. Chandrakasan. "MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns". In *Design Automation Conference*, 1998
- [3] E. Kusse and J. Rabaey. "Low-Energy Embedded FPGA Structures". In *Proceedings of International Symposium on Low Power Electronics and Design*, 1998
- [4] Z. Chen, M. Johnson, L. Wei, and K. Roy. "Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor". In *Proceedings of International Symposium on Low Power Electronics and Design*, 1998.
- [5] Navid Azizi and Farid N. Najam. "Look-Up table leakage reduction for FPGAs". *Published in Custom Integrated Circuits Conference, IEEE, 2005.*
- [6] Varghese George, Hui Zhang and Jan Rabaey. "The design of a low energy FPGA". In *Proceedings of International Symposium on Low Power Electronics and Design*, 1999.
- [7] Vaughn Betz and Jonathan Rose. "Cluster-Based Logic Blocks for FPGAs: Area-Efficiency vs. Input Sharing and Size". *IEEE press, 1997.*
- [8] Jason H. Anderson, Farid N. Najm and Tim Taun. "Active leakage power optimization for FPGAs". In *Proceedings of ACM/SIGDA 12th International Symposium on Field Programmable Gate Arrays*, 2004.
- [9] Changbo Long and Lei He. "Distributed sleep transistor network for power reduction". In *Proceedings of the 40th Annual Design Automation Conference*, 2003.
- [10] Kaijian Shi and David Howard. "Challenges in sleep transistor design and implementation in low-power design". In *Proceedings of Design Automation Conference, 43rd ACM/IEEE*, 2006.