Device/Circuit/Architectural Techniques for Ultra-low Power FPGA Design

Pooran Singh, Santosh Kumar Vishvakarma*

Nanoscale Devices and VLSI/ULSI Circuit & System Design Lab School of Engineering, Electrical Engineering Discipline, Indian Institute of Technology (IIT), Indore, MP, 145341, India

Abstract Field Programmable Gate Arrays (FPGAs) are widely used for implementation of digital system design due to their flexibility, low time-to-market, growing density and speed. But the Power consumption, especially leakage and dynamic power has become a major concern for semiconductor industries. FPGAs are less power-efficient than custom ASICs, due to the overhead required to provide programmability. Despite this, power has been largely ignored by the FPGA research community earlier, whose prime focus centred on improving FPGA speed and area-efficiency. But nowadays research extensively focuses on power too. Hence this paper demonstrates some of the most utilized and efficient techniques for Power optimization and reduction in FPGAs currently. After reviewing latest research work on power reduction in FPGA we examined that using Dual VT and fine-grained VDD static power reduces upto 64% and 95% respectively. Clock Gating reduces the power consumption by the factor 50% and also by using latest novel devices like Tunnel FET power can be reduced much lower than present.

Keywords FPGA, Power Reduction, CAD, SRAM, Non-Classical MOS Devices

1. Introduction

Field-Programmable Gate Arrays (FPGAs) are integrated circuits that can be programmed to implement any digital circuit. The main difference between FPGAs and conventional fixed logic implementations, such as Application Specific Integrated Circuits (ASICs), is that the designer/customer programs the FPGA on-site[1-3]. For fixed logic implementations, the designer must create a layout mask and send it to a foundry to be fabricated. Creating a layout is labour-intensive and requires expensive CAD tools and experienced engineers. Programmable switches controlled by configuration memory occupy a large area in the FPGA and add a significant amount of parasitic capacitance and resistance to the logic and routing resources. Because of this, FPGAs are approx 3 times slower, 20 times larger, and 12 times less power efficient compared to ASICs[4].

Many studies have focused on reducing the speed and area overhead of FPGAs. Important advancements include cluster-based logic blocks[5], which improve speed by grouping the basic logic elements of the FPGA into clusters with faster local interconnect; embedded memories[6], which reduce the speed and area overhead for applications with storage requirements; and embedded ALUs[7], which reduce the speed and area overhead for applications that perform arithmetic operations. A significant number of studies have also focused on faster; more area efficient programmable routing resources[8]. As CMOS process technology scales down, the power density continues to increase due to higher chip operating frequencies, higher total interconnect capacitance per chip, and increasing leakage. Indeed, the International Technology Roadmap for Semiconductors has identified low-power design techniques as a critical technology need[9].

This paper is organized as follows. Section 2 describes the basic architecture of an FPGA; Section 3 summarizes the sources of power dissipation; Section 4 covers different power reduction techniques Finally, Section 5 summarizes the paper and proposes future work.

2. FPGA Architecture

A basic symmetric FPGA is shown in Fig.1. The FPGA architecture is very regular in structure[10, 11]. It is made up of two main components-Configuration logic blocks (CLBs) and routing resources. The logic blocks implement the functionality of the given circuit while the routing resources provide the connectivity for implementing the logic. The logic blocks have the flexibility to connect to the routing resources surrounding them. The logic blocks and the routing resources are configurable, so that they can be programmed to implement any logic. Though many types of
architectures have been experimented with, the most popular one is the SRAM based architecture\cite{10, 11}; the architecture connected with programmable logic interconnects shown in Fig. 2.

![Image](image1.png)

**Figure 1.** A basic FPGA\cite{10}

![Image](image2.png)

**Figure 2.** FPGA basic architecture\cite{11}
2.1. Logic Block

The logic block of the SRAM based FPGA is LUT (look-up-table) based and composed of basic logic elements (BLE). LUT is an array of SRAM cells to implement a truth table. Fig. 3 shows a 4-input LUT which is used to design 7-input LUT for modern FPGA design as shown in Fig. 4[12]. It has four 4-input LUTs with 64 SRAM cells and 63 nos. of 2-input multiplexers to select one of the SRAM cells. The selection is done by the 7 select signals which are the inputs of the LUTs through the different 2 input multiplexers, which serve as inputs to the truth-table. Each BLE consists of a k-input LUT, flip-flop and a multiplexer for selecting the output either directly from the output of LUT or the registered output value of the LUT stored in the flip-flop. Fig. 5 shows the basic logic element. Previous works have shown that the 4-input LUT is the most optimum size as far as logic density, and utilization of resources are concerned, and this has been widely used. Cluster based logic blocks were investigated in [13] and it was shown that the cluster based logic blocks are better in speed and area. In the cluster based logic block, the logic block is made up of N BLEs. There are (I) inputs to the logic block such that each input can connect to all the BLEs. Also the output of each BLE can drive one of the inputs of each of the BLEs. The clock feeds all the BLEs. The work in [15] showed that the logic clusters containing 4 to 10 BLEs achieve good performance. Each sub-block is made up of a BLE and the corresponding LUT input multiplexers.

In today digital system design due to high end uses FPGA vendors provided a comprehensive alternative to FPGAs for large volume demands called structured ASICs [14, 15]. Structured ASICs offer a complete solution from prototype to high-volume production, and maintain the powerful features and high-performance architecture of their equivalent FPGAs with the programmability removed. Structured ASIC solutions not only provide performance improvement, but also result in significant high-volume cost reduction over FPGAs.

The programming technologies for logic and interconnect resources other than SRAM are flash memory [16], or antifuse [17-18]. SRAM-based FPGAs offer in-circuit reconfigurability at the expense of being volatile, while antifuse are write-once devices but non-volatile. Flash-based FPGAs provide an intermediate alternative by providing reconfigurability as well as non-volatility. But the most popular programming technology in state-of-the-art FPGAs is SRAM.

FPGAs usually include embedded memory, DSP blocks, Phase-Locked Loops (PLLs), embedded processors, and other special feature blocks, as shown in Fig. 6. These features allowed FPGAs to be an attractive alternative for some SoPC designs.
3. Power Consumption

Due to the dramatic increase in portable and battery-operated applications, lower power consumption has become a necessity in order to prolong battery life. Power consumption is an important part of the equation determining the end product's size, weight, and efficiency. FPGAs are becoming more attractive for these applications due to their shorter product life cycle. FPGAs are programmable, so they allow product differentiation. Selecting an appropriate FPGA architecture is critical in achieving the best static and dynamic power consumption.

As per we are dealing with Power optimization & reduction techniques so before going to study about those techniques first we will discuss about the types of power consumption in modern FPGA. The two components to power consumption: static, dynamic.

3.1. Static Power

Static power is the power consumed by the FPGA when no signals are toggling. Both digital and analog logic consume static power. The sources of static leakage current in 28-nm transistors are shown in Figure 7. These current which contributes towards leakage power are $I_{\text{sub}}$, subthreshold current when the transistor is in subthreshold region for gate-to-source voltages below the threshold voltage, $I_G$ Gate leakage current , $I_{\text{GIDL}}$ is a gate-induced drain leakage current and $I_{\text{REV}}$ represents for reverse saturation current.

3.2. Dynamic Power

Dynamic power is the additional power consumed through the operation of the device caused by signal toggling and capacitive loads charging and discharging. As shown in equation 1, the main variables affecting dynamic power are capacitance charging, the supply voltage, and the clock frequency. Dynamic power decreases with Moore’s law by taking advantage of process shrinks to reduce capacitance and voltage. The challenge is that as geometries shrink with each process shrink, the maximum clock frequency increases. While the power reduction declines for an equivalent circuit from process node to process node, the FPGA capacity doubles and the maximum clock frequency increases. The variables affecting dynamic power are explained in equation 1.

$$P_{\text{dynamic}} = \frac{1}{2}CV^2 + Q_{\text{ShortCircuit}}V f_{\text{activity}}$$  \hspace{1cm} (1)

4. Power Reduction Techniques

The Power Reduction Techniques are developed in various aspects; the power reduction is done at Static (leakage current reduction) power, Dynamic Power reduction through; Glitch removal, Clock Gating, Power Gating, Improved switching activity, Pipelining, Guarded Evaluation etc. we are going to discuss some techniques which are optimized and more efficient. The whole static and dynamic power reduction techniques are divided in three section here that is device, circuit and architectural.

4.1. Device Level Power Reduction Techniques

FPGA power can be more optimized by using ultra low power devices such as Tunnel-FET, FinFET and other MuGFET\cite{20-28}. These devices are very capable at working on ultra low voltages which would be very efficient for lowering power consumption in FPGA. In this section we are going to discuss about latest advanced devices which would be used to implement FPGA. Circuit design in FPGA will have to cope with enhanced leakage power and large process variability. Using Tunnel-FETs or carbon nanotube transistors instead of MOSFETs could drastically reduce the leakage power. Here we are going to discuss the effects of various advanced devices into digital integrated circuits. Then further we can study how we can use those circuits in FPGA for low power consumption. In this section we are going to discuss the characteristics and expected benefits of some emerging device categories for ultra low power integrated circuits. First, we focus on two categories of sub-thermal subthreshold swing switches Tunnel FETs and carbon nano-tube wires.

4.1.1. Tunnel FET

Tunnel FETs (TFETs) have emerged as most promising candidates for ultra low power digital ICs with voltage supply lower than 0.5V\cite{29}. In contrast to MOSFETs where charge carriers are thermally injected over a barrier, the
carrier injection mechanism in a TFET is quantum mechanical band-to-band-tunneling (BTBT). This mechanism is illustrated in the band diagrams of Fig. 8, corresponding to the ON state of the device. The main challenges of TFETs are their low $I_{on}$ current and extending the low swing over many decades of current. Fig. 9 shows the structure of Tunnel FET made in SILVACO TCAD.

Figure 8. a) Cross section of an n-type TFET, b) Schematic of energy-band diagram of the OFF-state/ON-state of n-channel DG-TFET[29]

Figure 9. Tunnel FET Structure at visual TCAD Silvaco
Leakage power dissipation is a fundamental problem for nano-electronic circuits special in case of SRAM memory. Scaling the supply voltage reduces the energy needed for switching, but the Field Effect Transistors (FETs) in today’s integrated circuits require at least 60 mV of gate voltage to increase the current by one order of magnitude at room temperature. Tunnel FETs avoid this limit by using quantum-mechanical band-to-band tunneling, rather than thermal injection, to inject charge carriers into the device channel. Tunnel FETs based on ultrathin semiconducting films or nano-wire could achieve a 100-fold power reduction over CMOS transistors, so integrating tunnel FETs with CMOS technology could improve low-power integrated circuit.

Recently, many novel devices such as the nano-wire gate all around (GAA) MOSFETs[30, 31], Fin-shaped FET[32], carbon nanotube FET, impact-ionization MOSFETs (I-MOSFETs)[33], and TFETs[34] have been demonstrated to minimize short channel effects and to lower the source-drain leakage current as compared to bulk MOSFETs. Leakage reduction using steep subthreshold transistors has gained great attention. A steep sub-threshold transistor allows us to operate at very low threshold voltages with ultra low leakage and low supply voltages (V_{DD}). Only TFET and I-MOSFET promise subthreshold slope less than 60 mV/dec and improved short-channel performances. Inter-band Tunnel transistor also called as Tunnel Field Effect Transistor (TFETs) works on principle of inter-band tunnelling[35]. TFETs have shown to be extremely power efficient in[36] for logic circuit applications.

4.1.2. Circuit Design with Tunneling FETs

Since the TFET technology is compatible with CMOS, circuits containing both standard MOSFETs and TFETs can be produced. This allows utilization of TFETs for special purposes, even if full replacement of CMOS cannot be achieved. This is of special interest in novel SOI technologies, which can no longer make use of bipolar devices as in bulk CMOS, and could instead make use of TFETs. In[37, 38] they presented several logic gate structures and an SRAM cell containing a mixture of bulk MOSFETs and planar TFETs. In Fig. 10, we see a six transistor SRAM cell where the two NMOS transistors with the source connected to GND have been replaced by TFETs[39]. Note that in planar bulk TFETs the two word line transistors cannot be replaced, since their source potential is different from GND and would require an additional well. On the other hand, the substrate contact inherent in the TFET source region offers an area advantage in bulk technologies. The read noise margin (RNM) of a SRAM design is estimated graphically as the length of a side of the largest square that can be embedded inside the lobes of a butterfly curve. The Write Noise Margin (WNM) is measured through the write trip point defined as the difference between V_{DD} and the minimum bit-line voltage required to flip the data storage nodes Q or Q_b. Figure 11 show an example of RNM (Read Noise Margin) measurement, read failure and WNM (Write Noise Margin) for a 6T TFET[40].

4.1.3. Leakage Power Reductions from Tunnel FET

Due to the inherent nature of TFETs, the OFF state leakage current of a TFET is orders of magnitude lower than CMOS. Thus, we see a huge improvement in terms of leakage reduction[41]. Fig. 12 shows the standby leakage/cell of various SRAM designs. Both 6T and 7T TFET has equal leakage power due to the presence of the same leakage paths. It obtain a 700X and 1600X improvement in leakage reduction over CMOS designs at 0.3V and 0.5V V_{DD}. This shows that TFETs are a potential
replacement candidate for CMOS transistors at low voltage and low power applications.

![Figure 12](image-url) | Standby Leakage/Cell for CMOS and TFET SRAM designs[41]

4.1.4. Carbon Nano-tube Wires

A somewhat larger step away from the CMOS mainstream is given by the Carbon Nanotube Field Effect Transistor (CNTFET). In[41] both n-channel and p-channel CNTFETs have been produced using different gate metals and Schottky Barrier source–drain regions. Inverters, ring oscillators, and simple logic gates have been fabricated already[42], but switching speed and ON-current seem still a large step below what is achievable with state-of-the-art CMOS devices[43, 44]. CNTFETs are among the most promising nano devices from the standpoint of their integration into future nano electronic systems on chip. Their physical characteristics (achievable current density, theoretical transition frequency and ratio), as well as their versatility and maturity all argue in favour of this view. They have diameters of typically 1 to 3 nm, but can be several micrometers long. CNTs can be exploited to build both low-resistance high-strength interconnections and highly scalable low-power CNTFETs and single-electron tunnelling transistors[46]. It is possible to consider that CNTFETs can be used to construct logic circuits under two scenarios.

4.1.4.1. Carbon Nanotube in a Transistor Configuration

The structure of the conventional CNTFET (C-CNTFET) provides a MOSFET-like behavior similar to that of a MOSFET but with ballistic transport[47]. The source and drain regions are heavily doped, which provides barrier-height modulation by application of the gate potential[Fig. 13(a) and (b)]. Since, for source–drain distances shorter than 150 nm, carrier transport is essentially ballistic at both high and low voltages, the description of current flow through the CNT relies on: i) the features of ballistic transport; ii) the specific electron confinement along the nanotube[48]. Since the current remains constant throughout the channel, the current is calculated at the top of the energy barrier corresponding to the beginning of the channel. At the top of the barrier, electrons coming from the source fill up the states and the electrons coming from the drain fill up the states[Fig. 13(b)].

![Figure 13](image-url) | C-CNTFET, (a) Schematic device features with a high-K dielectric, (b) Band diagram with, at $V_{GS}=0V$, the barrier height at the source–channel junction equal to $E_0=2$. The source and drain Fermi levels are shown by $\mu_S$ and $\mu_D$. (c) Energy versus wave number ($E$- $k$) diagram[47, 48, 49]

4.1.5. Carbon Nanotube SRAM Design

A study of an eight-transistor static random access Memory (SRAM) cell and its implementation in carbon nanotube FET (CNTFET) technology is done by Zhe Zhang et al.[50]. Simulations of the CNTFET SRAM cell design, using a CNT SPICE model, have shown advantages over the CMOS cell in terms of static power, dynamic power, and noise margin. However, current CNT synthesis processes grow metallic CNTs alongside semiconductor CNTs. This in turn greatly degrades the performance and functionality of SRAM cells. They compare two approaches to overcome the presence of metallic CNTs. The first approach tolerates metallic CNTs and uses a series of uncorrelated CNTs to form a transistor; this provides tolerance to metallic CNTs. The second approach uses an MxN array of uncorrelated CNTs to form a CNTFET and requires technologies capable of removing metallic CNTs. Both approaches have similar static noise margin. The second approach removes metallic CNTs consumes 1.45× more static power; on the other hand, its CNT count and write delay are reduced to 35.6% and 10.9% of the metallic tolerant approach, respectively. The realization of large memory modules in the presence of faulty SRAM cells can be achieved by having memory modules with as few as two spare columns.
4.2. Circuit Level Power Reduction Techniques

Static power is dominated by leakage current in various forms: sub-$V_T$ leakage, junction leakage i.e., source/drain, well, and triple-well junctions, Gate-Induced Drain Leakage (GIDL) and gate leakage. Here we are going to discuss several leakage current reduction techniques and dynamic power reduction through circuit level approach.

4.2.1. Dual Threshold Transistor Stacking (DTTS)

This is a technique for static power reduction in nano-scale CMOS circuits[55]. As we know to limit the ever increasing trend of energy and power dissipation in CMOS technology, supply voltage has to be continuously scaled. The amount of power reduction depend not only on supply voltage ($V_{DD}$) but also on the threshold voltage ($V_{th}$) to sustain the reduction of component delay, which is crucial for high speed digital circuit design. Continuous scaling down of these parameters poses several challenges to circuit designers. Particularly threshold voltage reduction leads to increase in sub-threshold leakage current leading to tremendous increase of static power consumption in CMOS circuits, which is otherwise considered as a negligible contributor to the overall power consumption. The performance of a CMOS tri-state buffer using Multi Threshold CMOS (MTCMOS) and Forced Transistor Stacking (FTS) leakage reduction techniques are analyzed and a new technique called Dual Threshold Transistor Stacking (DTTS) for efficient reduction of leakage power is introduced. From the results, it is observed that this technique combines the advantage of multiple threshold and stacking effects in MOSFETs. Table 1 shows the comparison of leakage power between different circuits.

<table>
<thead>
<tr>
<th>Circuit description</th>
<th>$P_{dyn,avg}$ (nW)</th>
<th>$P_{stat,avg}$ (pW)</th>
<th>Delay (ns)</th>
<th>PDP (fJ)</th>
<th>% Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>19.78</td>
<td>97.98</td>
<td>0.107</td>
<td>0.0021</td>
<td>--</td>
</tr>
<tr>
<td>MTCMOS</td>
<td>09.75</td>
<td>0.87</td>
<td>0.287</td>
<td>0.0027</td>
<td>50</td>
</tr>
<tr>
<td>FTS</td>
<td>18.43</td>
<td>85.16</td>
<td>1.152</td>
<td>0.0212</td>
<td>6.8, 13.0</td>
</tr>
<tr>
<td>R.Udaiya et al.</td>
<td>18.21</td>
<td>76.33</td>
<td>1.113</td>
<td>0.0202</td>
<td>7.9, 22.0</td>
</tr>
</tbody>
</table>

4.2.2. A Dual-threshold FPGA Routing Design

A dual-threshold FPGA routing design for subthreshold leakage reduction[56] is the technique in which the routing designs are based on the dual-threshold to reduce leakage power. Alternating between buffers and pass transistors, they analyzed the percentage constitution of low-$V_{th}$ and high-$V_{th}$ transistors as a function of the leakage reduction and delay increment tradeoff. By routing a suite of Microelectronics Center of North Carolina (MCNC) benchmark circuits, it is shown that an average savings of about 28.83% (as high as 48.46%) in total interconnect leakage can be obtained with 8.73% worst case average delay penalty.

4.2.3. Input Vector Reordering
A technique of input vector reordering for leakage power reduction in FPGAs[57] which is based on the state dependency property of leakage power. A pin reordering algorithm where the sub-threshold and gate leakage power components are taken into consideration to find the lowest leakage state for the FPGA pass-transistor multiplexers in the logic and routing resources without incurring any physical or performance penalties. This methodology is applied to several FPGA benchmarks, and an average leakage savings of 50.3% is achieved in a 90-nm CMOS process. The trend of leakage current is increasing with the technology scales down which is shown in Fig. 16 and in Fig. 17 it is shown that the leakage current dominant states in FPGA.

Figure 16. Leakage current vs. technology[57]

Figure 17. Total leakage-dominant states in FPGA pass-transistor devices (a) 90 nm. (b) 65nm and 45 nm. (c) 32 nm and 22 nm[57]

4.2.4. FPGA Leakage Power Reduction Using CLB-clustering

CLB-clustering design technique employs \( V_{DD} \) programmable and power gating methods to reduce leakage in stand-by mode[58]. In compared to the conventional \( V_{DD} \) programmable architecture, leakage power of CLB-clustering architecture is 0.1% of the leakage power in no gating FPGA and 8% of leakage power in conventional programmable \( V_{DD} \) FPGA.

Fig. 18 shows the logic block architecture of the CLB-clustering FPGA[58]. Four CLBs are clustered into one \( V_{DD} \) island where the same \( V_{DD} \) is used. \( V_{DD} \) of the block of 4 CLBs is either high \( V_{DD} \) (\( V_{DDH} \)) or low \( V_{DD} \) (\( V_{DDL} \)). The island is applied \( V_{DDH} \) when high performance is required and is changed to \( V_{DDL} \) if the blocks operate at lower speed. If the number of CLBs decreases in a block, the finer control is possible but area and delay overhead increase. Here, four is selected to keep the chip area and delay overhead below 5%. One CLB includes 4 BLEs and has 5 inputs and 3 outputs. One BLE consists of one LUT, one D-Flip Flop and one 2:1 MUX. This configuration is chosen because it is one of the best configurations for delay, area and logic utilization.

Figure 18. Architecture of CLB-clustering[58]

4.2.5. Programmability of \( V_{DD} \)

Power reduction is of growing importance for field programmable gate arrays (FPGAs). This reduction technique is discussed by Fei Li et al., about programmable supply voltage (\( V_{DD} \)) to reduce FPGA power. The FPGA logic fabrics using dual-\( V_{DD} \) levels shows that field-programmable power supply is required to obtain a satisfactory power-versus-performance tradeoff. They[59] further design FPGA interconnects fabrics for fine grained \( V_{DD} \) programmability with minimal increase of the number of configuration static-random-access-memory cells. With a simple yet practical computer-aided design flow to leverage the field-programmable dual-\( V_{DD} \) logic and interconnect fabrics, its carry out a highly quantitative study using placed and routed benchmark circuits, and delay, power, and area models obtained from detailed circuit designs. Compared to single-\( V_{DD} \) FPGAs with the \( V_{DD} \) level suggested by the ITRS for 100nm technology, field-programmable dual-\( V_{DD} \) FPGA reduce the total power by 47.61% and the energy-delay product by 27.36%.

4.3. Architectural Level Power Reduction Techniques

At architectural level power reduction we have several techniques which will reduce the power in data path, clock gating, and power gating. Some of those power reduction techniques are discussed in this section which will overcome both static and dynamic power.

4.3.1. Fine grained-\( V_{DD} \)

Low power FPGA architecture[60] is generated with the use of fine-grained \( V_{DD} \) control scheme called micro-\( V_{DD} \)-hopping, 4 CLB’s are grouped into one block where \( V_{DD} \) is shared as shown in Fig. 19. In the
micro-$V_{DD}$-hopping scheme, $V_{DD}$ of each block is varied between the higher $V_{DD}$ ($V_{DDH}$) and the lower $V_{DD}$ ($V_{DDL}$) spatially and temporally to achieve lower power, while keeping performance un-degraded. Simulation using 90nm CMOS technology shows that a leakage power reduction of 95% can be achieved, when this method is used.

4.3.2. Leakage Reduction in FPGA Routing Multiplexers

It is the technique for reducing the leakage current which is based on the architecture such as multiplexer-based interconnect matrix of an FPGA which consumes most of the static power. Leakage reduction in FPGA routing multiplexers[61] investigates reducing leakage power in unused FPGA routing multiplexers by controlling their inputs at the deep submicron 22nm technology node. HSPICE simulation using Berkeley Predictive Technology Models (BPTM) on different sizes and topologies of routing multiplexers shows that the minimum leakage vector at the 22nm technology node significantly varies from that at 65nm node. This is due to higher gate leakage and output stage loading effects. The application of this vector results in 20% more leakage power saving as compared to the existing approaches. This technique saves significant leakage power because most of the routing multiplexers are unused in an FPGA.

4.3.3. Power Gating

Power gating is the technique which is used for leakage power reduction, in which regions of the chip can be powered down. It is the modification to the fabric of an FPGA that enables dynamically-controlled power gating, in which logic clusters can be selectively powered down at run-time. For applications containing blocks with large idle times, this could lead to significant leakage power savings. There architecture utilizes the existing routing fabric and unused input pins of logic clusters to route the power control signals. The area and power tradeoffs have been studied by varying the basic architecture parameters of an FPGA, and by varying the size of the power gating regions. It shows that the leakage energy savings using a model that characterizes an application in terms of its structure and behavior. Using the application model[62], they show that up to 40% leakage energy reduction can be achieved using the architecture for different application parameters.

![Figure 19. Schematic of the CLB four BLE’s are clustered into one CLB][60]

4.3.4. Low Power Programmable FPGA Routing Circuitry

Programmable FPGA routing[63] technique is for reducing FPGA power consumption, it proposes a family of new FPGA routing switch designs that are programmable to operate in three different modes: high-speed, low-power, or sleep. High-speed mode provides similar power and performance to traditional FPGA routing switches. In low-power mode, speed is curtailed in order to reduce power consumption. Leakage is reduced by 28%-52% in low-power versus high-speed mode, depending on the particular switch design selected. Dynamic power is reduced by 28%-31% in low-power mode. Leakage power in sleep mode, which is suitable for unused routing switches, is 61%-79% lower than in high-speed mode.

![Figure 20. Dynamic power gating architecture for a logic cluster and its routing channels][62]

Fig. 20 shows an example of the basic power gating architecture. In this figure, a logic cluster has four input pins, with the required four connection boxes, distributed uniformly on its four sides. Each of the connection boxes can be used either to route an endpoint of a connection to the corresponding input pin, or to route a power control signal to the cluster. If a power control signal is to be routed, then the corresponding input pin of the cluster is not used. The outputs of the connection boxes are fed as inputs to the power gating multiplexer. This multiplexer selects the input pin that will be used as the power control signal for the cluster and the bounding routing channels; this signal is labeled PG_CNTL1 in the figure. PG_CNTL1 could drive the gate of the sleep transistor to turn it off for low-leakage mode, or to turn it on for normal circuit activity.

Fig. 21(a) shows a typical buffered FPGA routing switch[63]. It consists of a multiplexer, a buffer and SRAM configuration cells and a transistor-level view of a switch with 4 inputs is shown in Fig. 21(b). NMOS transistor trees are used to implement multiplexers in FPGAs. Routing switch inputs are tolerant to “weak-1” signals. That is, logic-1 in-put signals need not be rail; it is acceptable if they are lower than this. This is due to the level-restoring
buffers that are already deployed in FPGA routing switches[see Fig. 21(b)]. It permits such switches to produce “weak-1” signals. The main exceptions to this observation are switches that drive inputs on logic blocks. Based on these three observations,[63] proposed a new switch design shown in Fig. 22. The switch includes n-MOS and p-MOS sleep transistors in parallel (MNX and MPX).

Figure 21. Programmable low power routing switch (basic design)[63]

Figure 22. Switch multiplexer with programmable mode[63]

4.3.5. Clock Gating Power Reduction Technique

This is the most widely used technique for power reduction. The principle is to stop the clock whenever the device is not in use. Clock gating can be applied to sub-blocks of the design as well as to the whole device. However, correctly stopping the clock is very important. Knowing that the gating logic adds a delay to the clock signal, the effects on setup and hold times must be analyzed. While using clock gating, on FPGAs in particular, the user should take care of the placement of gating logic to minimize delay in the clock network. For reduction of dynamic power through clock gating approach first we discuss about the clock gating architectures for FPGA power reduction[64]. Clock gating is a power reduction technique that has been used successfully in the custom FPGA/ASIC domain. Clock and logic signal power are saved by temporarily disabling the clock signal on registers whose outputs do not affect circuit outputs. By considering and evaluating FPGA clock network architectures with built-in clock gating capability and describe a flexible placement algorithm that can operate with various gating granularities (various sizes of device regions containing clock loads that can be gated together). Results show that depending on the clock gating architecture and the fraction of time clock signals are enabled, clock power can be reduced by over 50%, and results suggest that a fine granularity gating architecture yields significant power benefits. The architectures are illustrated in Fig. 23. Fig. 23(a) shows the REGION architecture where enables are present on switches entering a region. Fig. 23(b) shows the more flexible COLUMN architecture where enables are also present on switches driving vertical spines in logic block columns. Thus, consider a broad range of clock gating architectures with various levels of granularity within clock distribution frameworks that resemble those in commercial chips.

Figure 23. Clock gating architectures; (a) Enables are present on switches entering a region (b) enables are also present on switches driving vertical spines in logic block columns[64]

4.3.6. Subthreshold FPGAs

Sub-threshold operation in CMOS has in recent years become an accepted ultra-low power solution. However, many low-volume applications cannot afford to produce custom silicon. An FPGA, which delivers the flexibility of programming and yet consumes ultra-low power by way of sub-threshold operation, can fill this gap. Field-programmable gate arrays (FPGAs) are an attractive option for low-power systems requiring flexible computing resources. However, the lowest power systems have yet to adopt FPGAs. Subthreshold circuit operation offers the opportunity to operate FPGAs at their minimum energy point. Peter J. Grossmann et al.[65] measured data from an
FPGA test chip fabricated in a 0.18-μm SOI process. They showed that the test chip can function at supply voltages as low as 0.26 V without an extra supply for write assists by using latches for configuration bit storage instead of static random access memory. Investigation of the minimum energy point of the FPGA for a high-activity test pattern shows that the minimum energy point of the FPGA can be well below the threshold voltage of the transistors. While Kyeong-Jae Lee et al. [66] demonstrated a subthreshold FPGA system using monolithically integrated graphene wires. The graphene wires replace double-length lines in the interconnect fabric of a custom FPGA implemented in 0.18-μm CMOS. The four-layer graphene wires have lower capacitance than the CMOS aluminium wires, resulting in up to 2.11× faster speeds and 1.54× lower interconnect energy when driven by a low-swing voltage of 0.4 V. They present’s us the first graphene-based system application and experimentally demonstrates the potential of using low capacitance graphene wires for ultralow power electronics.

5. Conclusions and Future Work

Due to the dramatic increase in power conscious applications and tighter power budgets, there is a necessity of low power consumption systems. The use of FPGA technology in low-power applications is increasing now-a-days, which makes achieving low power systems an increasingly important challenge. FPGAs have been adopted widely in recent years due to advanced technology that lowered the unit price, but the reduction in price have come at the cost of higher power due to higher transistor leakage. Various FPGA technologies have significantly different power profiles, and these differences can have a profound impact on the overall system design and power budget. Power consumption in FPGAs has become a primary concern for FPGA selection as previously more focus was on speed and making device more compact. But due to regressive use of mobile and portable devices human beings indirectly consume power from the nature as we are utilizing the power in the devices through natural resources like wind, water and other natural resources. According to Moore’s law, the device size reduces half of its present size in every 1.5 years. With reduced size, system will be faster and compact, but high end devices require lots of battery and natural power to run. So there is a need to focus on Power savings and to develop more refined and more optimize devices which can work on low power.

This paper is focused on Device, Circuit & Architectural Techniques for ultra-low power FPGA design. We discussed various power models for accurately computing the static and dynamic power both. We explore a strong review of various power reduction techniques and finds out the best technique for static and dynamic power. The techniques used for static power reduction reduces power upto 60-90% and dynamic power reduction techniques reduces power upto 30-50%. In static power reduction, dual-V_T FPGA architecture is explored and it indicates an average leakage power savings of up to 64%. In case of dynamic power, power savings of upto 61.6% can be achieved using LOPASS Technique [68] and upto 30% and 50% with clock gating and Glitch reduction techniques [69] respectively. Table 2 and table 3 shows the static & dynamic power results and various comparisons between them.
Table 2. Static power reduction techniques

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Static power reduction technique</th>
<th>Technology parameters</th>
<th>Reduction in static power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dual-threshold transistor stacking [55]</td>
<td>90 nm</td>
<td>22.09</td>
</tr>
<tr>
<td>2</td>
<td>Selection of polarities for logic signals in FPGA [63]</td>
<td>90 nm</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Fine-grained VDD [60]</td>
<td>90 nm</td>
<td>86.95</td>
</tr>
<tr>
<td>4</td>
<td>Dual-threshold FPGA routing design [56]</td>
<td>90 nm</td>
<td>28.83</td>
</tr>
<tr>
<td>5</td>
<td>Input vector reordering [57]</td>
<td>90 nm</td>
<td>50.3</td>
</tr>
<tr>
<td>6</td>
<td>FPGA Routing Multiplexers [61]</td>
<td>22 nm</td>
<td>20%</td>
</tr>
<tr>
<td>7</td>
<td>CLB-Clustering [58]</td>
<td>90 nm</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>Power gating [62]</td>
<td>45 nm</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>Carbon Nanotube SRAM Design [50-54]</td>
<td>90-150 nm</td>
<td>45</td>
</tr>
<tr>
<td>10</td>
<td>Subthreshold FPGAs [65-67]</td>
<td>90-180 nm</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 3. Dynamic Power reduction techniques

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Dynamic power reduction technique</th>
<th>Technology parameters</th>
<th>Reduction in dynamic power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Programmable FPGA routing circuitry [63]</td>
<td>90 nm</td>
<td>28.31</td>
</tr>
<tr>
<td>2</td>
<td>Clock gating [64]</td>
<td>90 nm</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Programmability of VDD [59]</td>
<td>90 nm</td>
<td>47.61</td>
</tr>
<tr>
<td>4</td>
<td>LOPASS Technique [68]</td>
<td>90 nm</td>
<td>61.6</td>
</tr>
<tr>
<td>5</td>
<td>Glitch reduction techniques [69]</td>
<td>90 nm</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>Guarded evaluation [70]</td>
<td>45 nm, 90 nm</td>
<td>32.28</td>
</tr>
</tbody>
</table>

The work till now on reduction of power is quite impressive but it is not up to the mark if the devices are used with reduced size or for high end applications. So our prime focus must be to reduce the power. In current and future research work which is focused on high level design flows, multi-core architectures, advanced applications in network processing, signal processing, and embedded systems, the power utilization is extreme. So the power can be reduced in steps from device level to system level. At initial stage, low power devices like Fin-FET, double gate, Tunneling FET can be used to make circuits. Power reduction techniques can be applied on the circuit which can be used in FPGA architecture. Then on FPGA architecture, the RTL level and CAD level power reduction techniques can be imposed which will reduce overall power of the system. Currently FPGA IC is fabricated using CMOS technology but the research is going on to fabrication of the FPGA via Fin-FET, Tunneling-FET MOS, and Multi-Gate FET devices. The devices would be developed which can work on low source power and also with low leakage current. The objective must be focused on power reduction techniques and the future challenges which can come across while implementing the techniques on FPGA.

REFERENCES


[50] Zhe Zhang and Jose G. Delgado-Frias “Carbon Nanotube SRAM Design With Metallic CNT or Removed Metallic CNT Tolerant Approaches,” IEEE Transactions on Nanotechnology, Vol. 11, No. 4, July 2012


