Energy-Efficient FPGA Interconnect Design

Maurice Meijer Rohini Krishnan[1] Martijn Bennebroek

Philips Research Laboratories
Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands
{maurice.meijer,martijn.bennebroek}@philips.com

Abstract

Despite recent advances in FPGA devices and embedded cores, their deployment in commercial products remains rather limited due to practical constraints on, for example, cost, size, performance, and/or energy consumption. In this paper, we address the latter bottleneck and propose a novel FPGA interconnect architecture that reduces energy consumption without sacrificing performance and size. It is demonstrated that the delay of a full-swing, fully-buffered interconnect architecture can be matched by a low-swing solution that dissipates significantly less power and contains a mix of buffer and pass-gate switches. The actual energy savings depend on the specifics of the interconnect design and applications involved. For the considered fine-grain FPGA example, energy savings are observed to range from a factor 4.7 for low-load critical nets to a factor 2.8 for high-load critical nets. The results are obtained from circuit simulations in a 0.13μm CMOS technology for various benchmarks.

1. Introduction

Hardware flexibility of Field-Programmable Gate Arrays (FPGAs) comes at the cost of having a switch dominated interconnect, which is responsible for most of the FPGA power consumption [1][2]. In general, three main sources of power consumption are distinguished being dynamic dissipation, short-circuit dissipation and leakage. With reducing feature sizes, leakage in deep-submicron designs has been increasing steadily and much research is ongoing to reduce leakage on FPGAs both on hardware and software level [3][4][5][6]. In this paper we focus on reducing dynamic and short circuit power dissipation by using low-swing signalling techniques.

The concept of dual-voltage solutions in FPGA design is not new itself [5][6][7]. Commercial FPGAs of Xilinx, for example, make use of NMOS pass-gate switches for which the gate voltage is boosted to one threshold above the signal swing to prevent voltage drops across switches [7]. The choice for pass-gate switches is mainly driven by area and performance considerations whereas power is not a primary constraint. Moreover, thick gate oxides are required to handle the boosted gate voltages, which limits the use of such pass-gate FPGA architectures to non-baseline CMOS processes. In [5], the usage of dual-voltage and dual-threshold transistors has been studied. Power savings ranging from 9% to 22% have been observed though, since “only” optimisations in the logic parts of an FPGA have been investigated, major power savings attainable in the interconnect remained untouched. In [6], interconnect optimisations involving low-swing signalling are studied and observed to reduce power more than an order of magnitude compared to existing, commercial architectures. To make up for the loss in performance, caused by the applied low-swing signalling scheme, dual-edge triggered flip-flops are required to boost throughput.

This work aims to improve the energy-efficiency of FPGA interconnects by using low-swing signalling without degrading performance. A hybrid switch solution is proposed, composed of both buffer and pass-gate switches, that matches the performance of full-swing, buffer-only interconnect though at significant lower (dynamic and short circuit) power levels. Section 2 describes the simulation set-up applied for this study and the results obtained for a typical fine-grain FPGA in 0.13μm CMOS technology are presented in Section 3. This section also includes some area consideration. Section 4 contains the conclusions and includes a brief generalization towards courser grain architectures.

2. Simulation set-up

An FPGA basically consists of an array of Configurable Logic Blocks (CLBs) wired together by a configur-

[1] Currently working at Intel Corporation

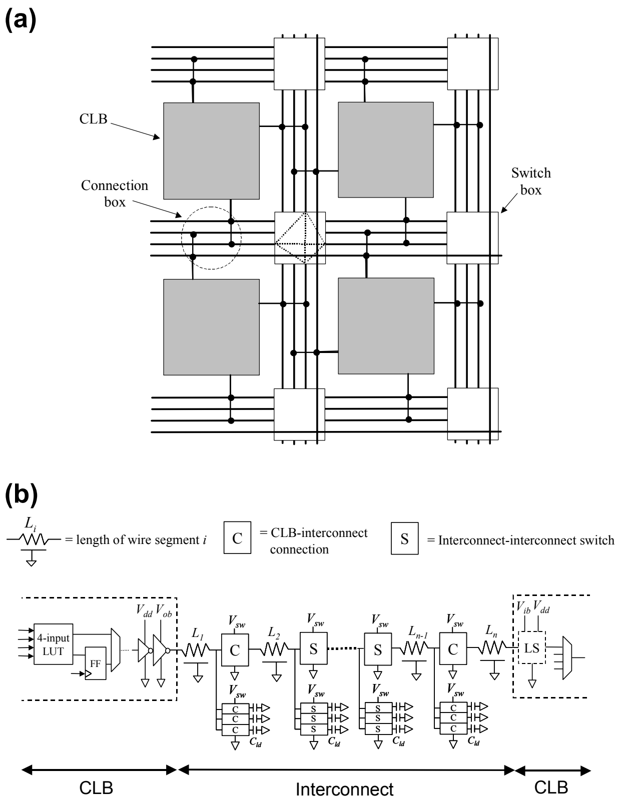

Figure 1: FPGA basics (a) and interconnect model (b) used for simulation.

able interconnect network as indicated in Figure 1(a). Functions can be set after fabrication by loading appropriate configurations into SRAM-like or non-volatile registers or into anti-fuses. In this work, we consider a SRAM-based, fine-grain FPGA example where the CLB contains a 4-input Look-Up-Table (4LUT) and an output register that can be bypassed. The interconnect is constructed from wire segments that span one, two, four or eight CLBs and Wilton-type switch boxes [7]. The switch box allows each wire segment to connect to four others. We estimate the area of the CLB and associated interconnect resources to be 50x50μm^2 in 0.13μm CMOS technology which implies wire segment lengths to range from 50μm to 400μm. We take 25% length1, 12.5% length2, 37.5% length4 and 25% length8 wire segments that, according to [7], provides an attractive performance and area trade-off.

Figure 1(b) depicts the interconnect model used in our transistor-level circuit simulations. The CLB operates at nominal supply voltage (V_{dd} =1.2V) whereas the supply voltages of its (latter) output buffer stage (V_{ob}), the interconnect switches (V_{sw}) and the input buffer stage (V_{ib}) can be reduced in case of low-voltage signalling. Simulations will be performed assuming switches in the connection and switch boxes to be tri-state buffers only, pass-gates only and hybrid mixes of these. The level shifter (LS) on the right in Figure 1(b) is included in the full pass-gate and hybrid switch cases to restore low signalling voltages back to the nominal supply voltage value (V_{dd}). Unlike buffer switches, pass-gate switches do not shield wire segments and therefore contribute to the load of the signal path. To account for this effect, capacitive loads have been added to the pass-gate switches outside the signal path and the load values have been ranged from 0fF (min) to 20fF (max). Interconnect segments are modelled by twenty RC elements with values representative for 0.13μm CMOS technology.

Besides basic simulations quantifying the effect of buffer drive strengths and pass-gate sizing, simulations have been performed on the critical nets in a variety of netlists from the MCNC benchmark set. VPR has been used for placement and routing [7] and PERL scripts have been written to extract critical nets details like the number, length and order of appearance of wire segments.

For all simulation results discussed in the next sections, the critical path delay and power consumption of the different benchmark circuits have been determined for a single data cycle of 1600ns. The power-delay product (PDP) is used to quantify the energy efficiency of different FPGA interconnect architectures.

3. Results and discussion

3.1 Fully-Buffered FPGA Interconnect Reference

In this section, the behaviour of buffers is studied under voltage scaling, while driving interconnects of a variable length. The weakest buffer (with x1 drive strength) at nominal supply voltage is used for reference purposes. The output stage of this reference buffer consists of a PMOS and NMOS transistor of size W_p/L_p=2.1/0.13[μm] and W_n/L_n=1.17/0.13[μm], respectively. The wire resistance equals 285mΩ/μm with a total capacitance of 237aF/μm. When the wire length is scaled, the goal is to

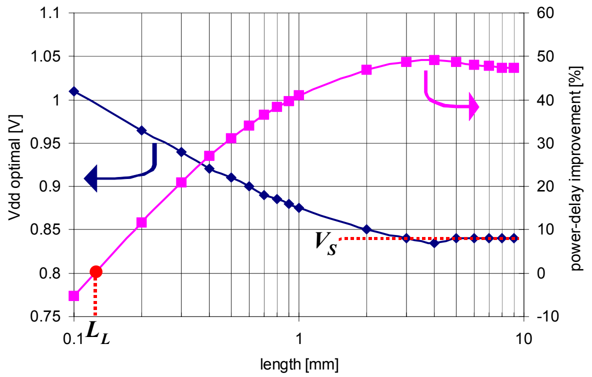

Figure 2: Optimal supply voltage (left axis) and resulting power-delay improvement (right axis) as a function of segment length for fully-buffered interconnect using buffers with x2 drive.

Table 1:
Saturation voltage V_S and critical length L_L for the fully-buffered interconnect reference

Buffer drive	x2	x3	x4	x5	x6
V_S [V]	0.84	0.73	0.67	0.64	0.62
L_L [mm]	0.125	0.15	0.225	0.25	0.3

reach the same delay as in the reference case, but at a lower PDP. The supply voltage is reduced for this purpose, and the value at which the delay is identical to the reference case is referred to as the optimum supply voltage. This simulation does not make use of the full interconnect model displayed in Figure 1(b), but rather is based upon a buffer driving a single wire with a fanout of two loading stage.

The results for a buffer with drive strength of two times the reference one are shown in Figure 2. This figure shows that improved PDP values can be obtained when the wire length scales beyond a value of L_L. Furthermore, it can be observed that the optimum supply voltage saturates to a value of V_S for longer wire lengths (>2mm). At V_S, the current drive of the stronger buffer is in the range of the drive strength of the x1 reference buffer operating at nominal supply voltage. Table 1 shows L_L and V_S for different drive-strength buffers.

Stronger buffers prove to be beneficial for wire lengths beyond L_L, whereas, for shorter wires, stronger buffers are not attractive due to the power consumption of their larger intrinsic capacitances. Therefore, it is not useful to use a stronger buffer for length1 (50µm) and length2 (100µm) wire segments. Using stronger buffers for length4 (200µm) and length8 (400µm) wire segments becomes attractive, and PDP improves up to ~11% (buffer drive x2 @V_{dd}= 0.97V) and ~30% (buffer drive x3 @V_{dd}=0.83V), respectively. However, this implies that there exist different signal voltages on the different wire segments. The overhead of multiple (more than 2) power distribution networks at such granularity including the need for level conversion, becomes simply too large. Additionally, the integrity of data signals may be affected due to cross-coupling effects of neighbouring signal wires. Therefore, in our investigation all buffers connected to length1, length2, length4, and length8 wire segments will have x1 drive strength and are powered by a single reduced power supply.

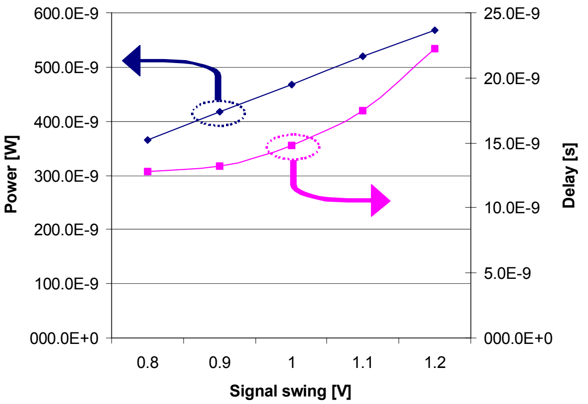

Figure 3: Full pass-gate interconnect power (left axis) and delay (right axis) results for the critical path in the Tseng benchmark.

By using the simulation set-up as described in Section 2, delay and PDP values have been obtained for various benchmark applications operating at the nominal supply voltage of 1.2V. The results are shown in columns two and three of Table 2 and will be used as a reference in the remainder.

3.2 Full Pass-Gate Interconnect Design

We now turn to interconnects where all switches are implemented as nmos pass-gates. In contrast to buffer switches, pass-gate switches do not shield wire segments and, consequently, fanout loading at the switch position will affect the overall delay and power. In our case, the switch is always controlled from the nominal voltage supply (V_{sw}=1.2V). This reduces the switch's resistance at lower signal voltages and eliminates the switch from being cut-off for signal swings below V_{dd}-V_{th}. A low-resistance switch is best for delay, but it is larger in terms of silicon area. Simulations show that W/L=0.36/0.13[µm] seems to be a good compromise for signal swings below 1.2 volts. In the remainder of this work, all pass-gate switches are assumed to have this geometry.

Figure 3 shows the delay and power values for different signal voltages obtained for the critical net in the Tseng benchmark. These results show that power is decreasing linearly with signal voltage. This is due to the fact that the signal swing on the interconnect is constrained to a value of V_{sw}-V_{th} (~0.85V) in case the input signal voltage is larger (V_{ob}>V_{sw}-V_{th}). For lower input signal voltages (V_{ob}≤V_{sw}-V_{th}), the signal swing on the interconnect would be equal to the input signal voltage, and the power consumption would show a quadratic relationship with signal voltage. The delay is lower for reduced signal voltages, which can be explained by the reduced switch resistance at lower supplies.

Compared to the fully-buffered reference, the power consumption is reduced by a factor of 4 but, unfortunately, the delay is increased by a factor of 7. This illustrates that it is not attractive to compose long critical nets entirely from pass-gate switches due to the large performance penalty.

3.3 Hybrid Interconnect Design

In this section, a hybrid interconnect design is proposed that contains a mixture of buffer and pass-gate switches and operates at a reduced supply voltage to improve energy efficiency without degrading performance. The optimal ratio of buffer to pass-gate switches is optimized for minimum and maximum loading conditions for the critical net in the Tseng benchmark. This critical net contains 30 switches and we force buffers to be inserted uniformly. For example, if one buffer is inserted it will be positioned after pass-gate 14, two buffers end up after pass-gates 9 and 19. The reference delay using a fully-buffered interconnect equals 3ns.

Figure 4 displays the delay and PDP values as a function of the number of inserted buffers for the critical net in the Tseng benchmark in case of maximum loading conditions. This figure shows that minimum delay and PDP points can be found for each signal swing. For instance, a minimum delay is found with two, two and four inserted buffers for signal swings of 0.7V, 0.8V and 0.9V, respectively. The lowest PDP is obtained for a signal swing of 0.7V with only one inserted buffer, however, under these conditions the reference delay of 3ns cannot be met.

The procedure to find the optimal ratio of buffer to pass-gate switches is such that, first, the signal swings are identified that enable the reference delay to be matched. Second, the signal swing and corresponding number of inserted buffers is selected that provide the overall lowest PDP. If possible, the number of inserted buffers should be kept as low as possible to save area. In this manner, one can derive from Figure 4 that energy efficiency is obtained for maximum fanout load with a buffer to pass-gate ratio of 1:6 and a signal swing of 1V. Likewise, for the minimum fanout load the optimum ratio and signal swing are found to be 1:9 and 0.8V, respectively. The difference in optimum conditions can be explained as follows. To match the reference delay for maximum fanout load, more buffers are required while the signal swing needs to be increased to prevent buffer delay start dominating. This translates into lower energy savings for the maximum fanout case.

Table 2 summarizes the results obtained for various benchmarks when using the proposed hybrid approach. A significant reduction in PDP has been achieved for all benchmarks while maintaining their reference delay. On average the PDP reduces by a factor of 4.7x for low-load critical nets operating at signal swings of 0.8V or 0.9V. Similarly, an average reduction in PDP of 2.8x has been obtained for high-load critical nets operating at 1V supply.

In the simulations, we have used a SSDLC levelshifter [8] because it is found to offer a low PDP without using additional technology options such as high-V_{th}. When a regular CMOS buffer would have been used instead of this levelshifter, the average reduction in PDP becomes 4.4x and 2.7x for low-load and high-load critical nets, respectively. These PDP values are lowered due to the short-circuit power in the receiving stage.

3.4 Area considerations

Besides energy consumption, the areas required for hybrid interconnects also compare favorable with the fully-buffered interconnect reference. The main reason for this is that nmos pass-gates are much smaller than tri-state buffers and that the hybrid interconnect architecture contains many nmos pass-gates. In the remainder of this section, a more detailed comparison between the areas for both interconnect architectures will be presented.

In total 16 interconnect tracks are found to be required to place and route all benchmarks. Here, the track segmentation described in Section 2 is utilized which implies four tracks of length1 (=25%), two tracks of length2 (=12.5%), six tracks of length4 (=37.5%), and four tracks of length8 (=25%). With eight bidirectional switches at each segment

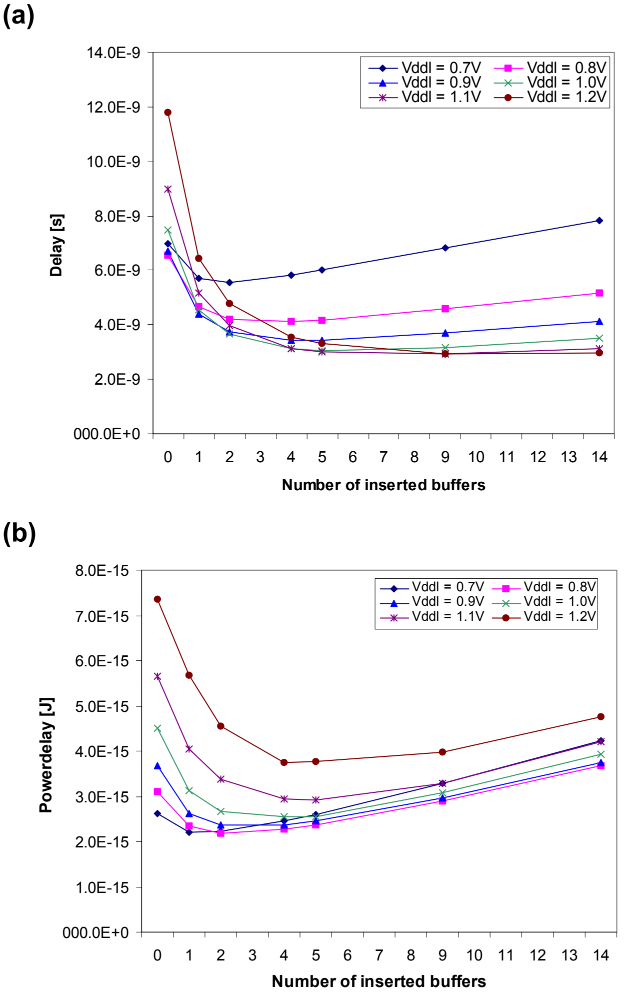

Figure 4: Delay (a) and power-delay (b) as function of inserted buffers for the critical net in the Tseng benchmark containing pass switches with maximum fanout load and terminated by a level shifter

Table 2: Delay and power-delay results for the hybrid design, including improvements w.r.t. the fully-buffered reference

	Reference		**Obtained delay [ns]**		**Power delay [fJ]**		**Improvement in PD**		**Supply voltage [V]**	
	Delay [ns]	Power delay [fJ]	**Fanout load**		**Fanout load**		**Fanout load**		**Fanout load**	
			min	max	min	max	min	max	min	max
tseng	3.0	6.7	3.0	2.9	0.9	2.6	7.8x	2.6x	0.8	1.0
elliptic	25.8	494.0	25.7	25.4	145.3	259.3	3.4x	1.9x	0.9	1.0
e64	4.4	14.0	4.3	3.1	1.8	2.8	7.7x	4.9x	0.8	1.0
diffeq	37.0	1025.0	36.4	36.4	249.3	398.1	4.1x	2.6x	0.9	1.0
seq	38.2	1093.7	38.5	37.9	275.5	432.0	4.0x	2.5x	0.9	1.0
s38584.1	44.8	1500.0	42.6	43.1	338.7	561.7	4.4x	2.7x	0.9	1.0
apex4	45.2	1525.0	42.6	43.3	338.0	567.5	4.5x	2.7x	0.8	1.0
spla	56.4	1906.0	51.3	51.1	498.2	783.3	3.8x	2.4x	0.9	1.0
apex2	52.1	2031.2	49.7	50.1	467.7	711.5	4.3x	2.6x	0.9	1.0
misex3	59.4	2637.5	55.9	56.2	589.9	945.9	4.5x	2.7x	0.9	1.0
alu4	76.6	4381.2	71.9	71.4	970.1	1537.7	4.5x	2.7x	0.9	1.0
des	80.1	4793.0	77.1	77.7	1222.0	1714.8	3.9x	2.6x	0.9	1.0
ex1010	82.4	5068.0	78.2	78.1	1148.3	1757.1	4.4x	2.7x	0.9	1.0
frisc	93.6	6563.0	91.4	91.8	1583.1	2424.4	4.1x	2.5x	0.9	1.0
es5p	106.1	8375.0	107.5	106.3	1471.2	2254.4	5.6x	3.7x	0.8	1.0
Average							4.7x	2.8x		

intersection, the four length1 tracks contribute 32 switches per switchbox. Of the two length2 tracks, only one intersects on average per switchbox and thereby adds eight switches. Likewise, the six length4 tracks add on average nine (=6/4*8) switches and the four length8 tracks add on average three (=4/8*8) switches per switchbox. This brings the total amount of switches in a switchbox to 56.

The area of a pass-gate of size W/L=0.36/0.13[μm] in 0.13μm CMOS technology is only about 4% of the area of a tri-state buffer. In the following, the tri-state buffer area will be used as a reference and denoted by A_{ref} ($\sim 14\mu m^2$). To form a bidirectional switch, a single nmos pass-gate suffices whereas two tri-state buffers will be required. For the hybrid switchbox, when considering the above-derived buffer to pass-gate ratio of 1:6 (for maximum fanout load), the area will be about $20.5A_{ref}$ (=56*1/6*2+56*5/6*0.04). The number of configuration bits controlling the hybrid switchbox amounts to 65.3 (=56*1/6*2+56*5/6*1) and, assuming the area per bit to match A_{ref}, the involved bit area amounts to $65.3A_{ref}$. The considered 4-input CLB requires four level shifters, each roughly of size A_{ref}, to restore the low-swing signals to full-swing inputs. Thus, the combined area for the switchbox, control-bits and CLB-inputs of the hybrid interconnect roughly amounts to $89.9A_{ref}$.

For the fully-buffered interconnect, the switchbox and the control-bits areas add-up to $224A_{ref}$ (=56switches*[2tri-state buffers+2bits]/switch). Due to its full-swing operation, the CLB-inputs do not require level-shifters but instead include plain buffers, each roughly of size $1/2A_{ref}$. Thus, the combined area of the switchbox, control-bits and CLB-inputs of the fully-buffered interconnect amounts to $226A_{ref}$ which is 2.5 (=226/89.9) times bigger than required for the hybrid interconnect.

Note that the connection box areas are excluded from the area consideration, since these strongly depend on implementation specifics not addressed in this paper. However, similar area gains are expected for the output connection box when pass-gates are used instead of buffer switches.

The actual area overhead of the fully-buffered interconnect with respect to the hybrid interconnect architecture will be below a factor of two since the hybrid architecture needs an extra (low-voltage) power supply to be routed throughout the FPGA. However, based on previous FPGA layout experiences, we believe this leaves sufficient margin to implement the hybrid interconnect architecture at an area comparable to (if not slightly smaller than) that of the fully-buffered interconnect.

4. Conclusions

We demonstrated that the delay of full swing, fully buffered FPGA interconnect design can be matched by a low-swing, hybrid switch solution, containing both buffer and pass-gate switches, that dissipates significantly less power without an area penalty. The actual power savings, the optimal choices for the low-swing voltage level and buffer to pass-gate ratio depend on the specifics of the interconnect design and applications involved. For the considered fine-grain FPGA example in CMOS 0.13μm, energy savings have been derived in between a factor of 2.8 and 4.7 at low-swing voltages between 0.8V and 1.0V depending on the net loading involved. An optimal buffer to pass-gate ratio of 1:6 has been derived for maximum loading conditions.

For interconnect designs that have longer wire segments than considered here, e.g. in coarser grain FPGAs, two optimization scenario's can apply. As long as the switch resistance dominates over the wire resistance, the optimizations presented above will apply. However, when the resistance of the wire segments cannot be neglected, the low-swing voltage level is observed to reduce whereas the buffer to pass-gate ratio and the attainable power savings go up.

Acknowledgements

The authors acknowledge Guillaume Durand of INSA de Lyon for his help in generating the results.

References

[1] E.Kusse, J.Rabaey, *Low-Energy Embedded FPGA Structures*, International Symposium of Low Power Electronic Design (ISLPED), Monterey, CA, USA, 1998

[2] L.Shang, A.Kaviani, K.Bathala, *Dynamic Power Consumption in Virtex-II FPGA Family*, FPGA, February 2000, pp.157-164

[3] J.H.Anderson et.al., *Active Leakage Power Optimization for FPGAs*, ACM International Symposium on FPGAs, Monterey, CA, 22-24 February 2004

[4] A.Rahman, V.Polavarapuv, *Evaluation of Low-Leakage Design Techniques for FPGAs*, ACM International Symposium on FPGAs, Monterey, CA, 22-24 February 2004

[5] F.Li et.al., *Low-Power FPGA Using Pre-defined Dual-Vdd/Dual-Vt Fabrics*, ACM International Symposium on FPGAs, Monterey, CA, 22-24 February 2004

[6] V.George, H.Zhang, J.Rabeay, *The Design of a Low Energy FPGA*, International Symposium of Low Power Electronic Design (ISLPED), San Diego, CA, USA, 1999, pp.188-193

[7] V. Betz, J. Rose, A. Marquardt, *Architecture and CAD for deep-submicron FPGAs*, Kluwer Academic Publishers, 1999.

[8] H.Zhang et.al., *Low-Swing On-Chip Signalling Techniques: Effectiveness and Robustness*, IEEE Transactions On VLSI, Vol. 8, No. 3, June 2000, pp.264-272