A Low Power FPGA Routing Architecture

Somsubhra Mondal, Seda Ogrenci Memik
ECE Department, Northwestern University
Evanston, IL USA

***Abstract*—** Significant headway has been made in logic density and performance of FPGAs in the past decade. Power efficiency of FPGA architectures is arguably the next most important criterion that needs improvement**.** In this paper, we propose an interconnect architecture, where voltage scaling is applied within the programmable interconnect structure of the FPGA. We present an evaluation of the overhead associated with dual-V_{dd}-dual-V_t interconnect architecture and present results on the impact of this routing architecture on area and delay. Our experiments reveal that an average reduction of 23.45 % (as high as 47 %) in total interconnect power is achievable with 11.75 % worst-case delay penalty and 6 % area overhead on average.

I. Introduction

FPGAs evolved at a rapid pace into highly complex multi-million transistor ICs. While this provided a big performance boost, the power efficiency of FPGAs is lagging behind. There are several application domains (e.g. mobile applications), which could benefit from the low manufacturing cost and flexibility offered by the FPGA technology. However, without significant improvements in power efficiency many advantages of FPGAs are overshadowed, which is deterring designers from considering inclusion of FPGAs in power constrained systems. In addition, high power dissipation reduces reliability of a system and increases costs associated with packaging and cooling. Therefore, it is imperative to improve the power efficiency of FPGAs.

Studies on power consumption of Xilinx Virtex™ devices reveal that interconnect power dominates the total power consumption [1]. Therefore, a closer look at the opportunities to improve the power efficiency of the routing resources is most beneficial. To accomplish this goal, we investigated the impact of using a dual-V_{dd}-dual-V_t routing architecture on power. We propose to divide the routing channels into two regions: the V_{dd}^{High} and V_{dd}^{Low} tracks. Components of the routing architecture (programmable switch boxes, connection matrices at the inputs and outputs of the logic blocks) are modified to operate with two supply voltage levels. If we only scale down the supply voltage while keeping constant V_t across the routing architecture, we observe a large delay penalty. We performed simulations of the routing architecture components to determine an appropriate V_t level for V_{dd}^{Low}. This second V_t level still incurs a delay penalty although much smaller while maintaining uniform leakage power across the routing architecture.

We started our evaluation by analyzing the path delay distribution of a set of benchmarks. A summary of our results to this end will be presented in Section 3.A. We observed that the majority of paths possess a significant amount of time slack. This motivated us to use the dual-V_{dd}-dual-V_t routing architecture, where a fraction of the routing resources will be slower due to voltage scaling. Available time slack in circuits helps tolerate additional delay caused by slow routing tracks without incurring large delay penalties on the critical path of the designs. Our specific contributions in this work are:

- We introduce a dual-V_{dd}-dual-V_t routing architecture and present experimental results with different configurations of the routing architecture by varying the percentage of tracks supplied with V_{dd}^{Low}, and
- We evaluate the implementation overheads associated with this low power routing architecture.

The rest of this paper is organized as follows: Section 2 presents an overview of related work. Section 3 describes our analysis on different components of the routing architecture. In Section 4, we report the achieved power reduction, present statistics on the delay distribution across paths in a design before and after the introduction of this architecture, and present the impact on delay and area. Section 5 summarizes our conclusions.

II. Related work

In the recent few years, research efforts in improving the power efficiency of the reconfigurable fabric itself have intensified. Techniques for reducing leakage power were proposed by disabling unused portions of the FPGA [2], and by selecting polarities for logic signals at the inputs of LUTs so that they spend the majority of their time in low leakage states [3]. Li et al. proposed application of voltage scaling onto the logic blocks of the FPGA architecture [4]. This was followed by a technology mapping technique to efficiently utilize this feature [5]. Li et al. recently extended their studies to creating programmable dual V_{dd} fabrics, where certain logic blocks can be programmed to operate at either High or Low V_{dd} [6]. Gayasen et al. proposed a dual Vdd architecture, where dual supply voltages are applied on logic blocks and routing multiplexers [7]. Power reduction by scaling down the supply voltage is a popular design technique and has been proven successful in ASICs with dual V_{dd} or multi V_{dd} designs [8,9, 10]. Techniques to reduce leakage power by using dual V_t in ASICs have also been studied [11].

We propose a dual-V_{dd}-dual-V_t routing architecture, where a fraction of the routing tracks operate on a high V_{dd} high V_t level, while the rest use a scaled down V_{dd} and V_t. Dual voltage supplies are provided to switch matrices as well as driving buffers of the routing segments. Our dual-V_{dd}-dual-V_t routing architecture can also be used in conjunction with the abovementioned power reduction techniques applied to the logic blocks.

III. Voltage Scaling for FPGA Routing Architecture

The routing architecture of an FPGA is a dominating factor on the overall chip area, circuit delay and power consumption. In the following subsections, we discuss different components of the routing architecture and our dual-V_{dd}-dual-V_t routing architecture.

Characteristics of Switches: Switch blocks enable programmable interconnect in FPGAs. In a switch block, a tri-state buffer is used as a unidirectional switch, whereas a pass transistor is used as a

bidirectional switch. Scaling down the V_{dd} reduces dynamic power for these switches, but it involves a delay penalty if the threshold voltage is not adjusted accordingly. On the other hand, decreasing V_t would lessen the delay penalty, but it increases leakage. Therefore, V_t values for each V_{dd} should be chosen such that a good power-delay trade-off is obtained. We performed HSpice simulations for different V_{dd} and V_t values for pass transistor and tri-state buffer switches. Figure 1(a) shows the increase in delay of a tri-state buffer for decreasing V_{dd} values under three conditions: constant V_t, fixed V_{dd}/V_t ratio, and V_t for constant leakage. Leakage power by a tri-state buffer for these three V_{dd} scaling schemes is shown in Figure 1(b). From Figures 1(a) and (b) we can conclude that the constant leakage V_{dd} scaling scheme provides a good trade-off in terms of delay and power. A similar conclusion was reached for configurable logic blocks [4].

Based on these facts, we have applied this V_{dd} scaling scheme to the routing switches, which makes the switches about 30% slower at 1.2V V_{dd} when compared to 1.8V V_{dd}.

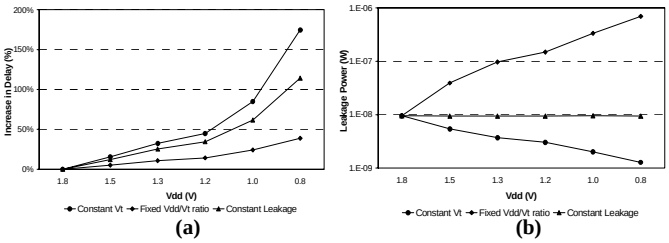

Figure 1. (a) Relationship between delay and V_{dd} (b) Relationship between leakage power and V_{dd}

Switch Block Topology: The switch box topology describes how each pin on one side of the switch box is connected to the pins on the other three sides. We will briefly discuss the subset switch box, and in the following subsections, we will state why this topology is suitable for our dual-V_{dd}-dual-V_t architecture.

The subset switch block connects each pin on one side to the pins with the same index number on the other three sides of the switch block. A disjoint switch box with four horizontal tracks and four vertical tracks is shown in Figure 2. For clarity, possible connections for only one pin are shown.

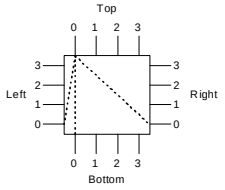

Figure 2. Disjoint switch block

A. Dual-V_{dd}-dual-V_t Routing Architecture

In our proposed dual-V_{dd}-dual-V_t routing architecture we have classified the routing tracks into two types: V_{dd}^{High} tracks and V_{dd}^{Low} tracks. The difference between these two types of tracks is that the switches on the V_{dd}^{High} track have a higher supply voltage and hence, are faster than that of the V_{dd}^{Low} tracks.

We performed an analysis of the path delays for 20 MCNC [12] benchmark circuits. Figure 3 shows the path delay distribution of a sample benchmark *pdc*. We observe that majority of the paths are within 70% of the critical path. Similar trends were observed across all benchmarks indicating that there is plenty of slack for these paths to slow down.

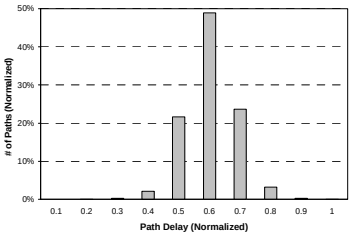

Figure 3. Path delay distribution of the pdc benchmark

The motivation for having these two types of tracks is based on the following observations:

- The paths with the zero or low slack values can use the faster V_{dd}^{High} tracks and
- The remaining paths can use the slower V_{dd}^{Low} paths in order to save power.

The distribution of the two types of tracks is an architectural parameter. For our experiments we have used two distributions: 50% V_{dd}^{High} / 50% V_{dd}^{Low} tracks, and 30% V_{dd}^{High} / 70% V_{dd}^{Low} tracks. A 50-50 distribution of V_{dd}^{High} and V_{dd}^{Low} tracks is illustrated in Figure 4.

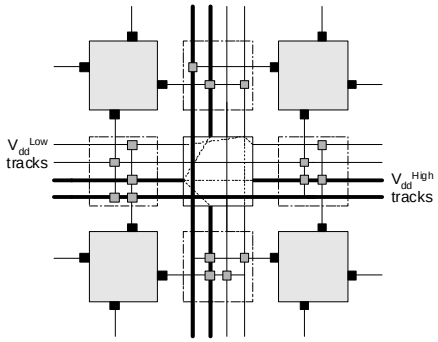

Figure 4. Proposed dual-V_{dd}-dual-V_t routing architecture with two different types of tracks

Having two different types of tracks operating on two different V_{dd} levels, proper communication between the tracks from/to the logic blocks and tracks to tracks has to be ensured. Care should be taken so that a V_{dd}^{Low} track never drives a track or a logic block that operates on a high V_{dd} level directly. On the other hand, a V_{dd}^{High} track can safely drive a V_{dd}^{Low} track without loss of signal strength. There are three cases where such problems can arise:

Case 1: Connection between V_{dd}^{High} and V_{dd}^{Low} tracks
Case 2: Connection between V_{dd}^{Low} track and output pin of a CLB.
Case 3: Connection between V_{dd}^{Low} track and input pin of a CLB.

For the first case, any direct connection between V_{dd}^{High} and V_{dd}^{Low} tracks can be avoided by using the disjoint switch block topology as shown in Figure 1. As each pin on one side connects to only the same pin index number on the other three sides of the switch block, a V_{dd}^{High} segment in a track can never be connected to a V_{dd}^{Low} track, and vice versa.

In the second case, when all the CLBs are operating on high V_{dd} level, their output pins can safely drive the V_{dd}^{Low} tracks. When dual V_{dd} logic blocks are used, this problem can be solved by using a level converter at the output pin of only the low V_{dd} logic blocks. For the third case, when a V_{dd}^{Low} track drives a high V_{dd} CLB input pin, the input pin connection block has to incorporate level converters in between the V_{dd}^{Low} segment and the buffer, so that the multiplexer has the same signal strength for all its inputs. So the only overhead of the proposed architecture is the use of level converters in the input pin connection blocks for the V_{dd}^{Low} tracks. The V_{dd}^{High} tracks will not require any level converters.

B. *Dual-V_{dd}-dual-V_t Routing Architecture Applied Simultaneously with dual-V_{dd}-dual-V_t Logic Blocks*

As mentioned earlier, our proposed routing architecture can be applied with other voltage scaling techniques. One such technique is to use dual-V_{dd}-dual-V_t logic blocks with the use of level converters at the output pins of the low V_{dd} logic blocks. When using our dual-V_{dd}-dual-V_t routing architecture, this technique is still applicable. Since the output of the low V_{dd} logic blocks are converted to high V_{dd} level, they can safely drive both types of routing tracks in our architecture. Also, since we are using level converters in the input connection blocks for the V_{dd}^{Low} tracks, they can drive CLBs at both V_{dd} levels. Hence, when this routing architecture is used in conjunction with dual V_{dd} logic blocks there is no extra overhead. In fact, if our technique is used with dual V_{dd} logic blocks, a lesser number of level converters are required, because level converters will not be needed V_{dd}^{Low} tracks to drive the low V_{dd} CLB input pins, and Low V_{dd} CLB output pins to drive the V_{dd}^{Low} tracks.

IV. Experimental Results

We have carried out our experiments for four different cases:

- 100% V_{dd}^{High} tracks, which is the base case for our comparisons.
- 50% V_{dd}^{High} tracks and 50% V_{dd}^{Low} tracks
- 30% V_{dd}^{High} tracks and 70% V_{dd}^{Low} tracks
- 100% V_{dd}^{Low} tracks

We have used a V_{dd}^{High} of 1.8V, and the V_{dd}^{Low} is 1.2V. We chose V_t values for these V_{dd} values based on the constant leakage V_{dd} scaling scheme. For all these cases we have clustered 4-input LUTs in one complex cluster and the number of inputs for each cluster is 10. Moreover, the switches in V_{dd}^{Low} tracks are made 30% slower as obtained from HSpice simulation and illustrated in Figure 1. We have assumed a uniform distribution of the routing tracks, i.e. equal number of horizontal and vertical tracks. Also, we have 50% pass transistors and 50% tri-state buffers, and each wire segments in a routing track spans 4 logic blocks. This configuration was experimentally determined to yield the best delay/routability/routing area in previous studies [13].

We have used Versatile Place and Route tool (VPR) [14] to do the timing driven packing, placement and routing. For power estimation we have used Power Model [15], an additional module integrated with VPR.

A. *Path Delay Statistics*

The critical path determines the operating frequency of a circuit. It is obvious that not all paths will have the same path delay. In fact, majority of the paths will have some amount of slack as we have discussed in Section 3.A. Figure 5 shows the average distribution of the path delays for the benchmarks. The x-axis represents the normalized path delays, i.e. path delays given as fractions of the length of the critical path in the circuit. The percentage values along the y-axis represent the number of paths that have a length corresponding to that fraction of the critical path length. From Figure 5 it is clear that scaling down the V_{dd} causes greater number of paths to slow down, i.e. more paths will have delays closer to the length of the critical path.

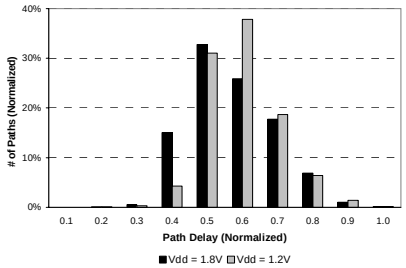

Figure 5. Path delay distribution using the proposed architecture

B. *Delay and Area Comparison*

Table 1 shows the percentage change in delays of the benchmark circuits for the three cases compared with the single high V_{dd} (1.8V) routing architecture. A negative value means that the delay has decreased compared to the base case. We observe that when all tracks are supplied by V_{dd}^{Low}, the delay penalty is 12%, whereas if a 50-50 or 30-70 mix of V_{dd}^{High} and V_{dd}^{Low} tracks are used, there is no delay penalty. Our experiments show that although there is no increase in routing area for the 100% V_{dd}^{Low} case (except for the additional level converters), there is an area penalty of 4 % and 6 % for the 50-50 and 30-70 distribution respectively. This area penalty originates from the following phenomenon. Changes in the routing architecture affect the routability in some cases, causing congested areas. This in turn causes the timing-driven router to use a higher number of tracks than the base case to relieve that congested region, which increases the channel width over the whole chip[1]. The increase in the channel width causes an increase in the routing area. In addition there will be some area overhead due to the level converters, which will be small compared to the increase in channel width. The very same phenomenon has a positive impact on delay in some cases. Since the router increases the channel widths, now routing becomes significantly easier (e.g. less congestion) and therefore the length of the critical path reduces. On the other hand, a closer look at the worst case impact of the proposed routing architecture reveals that for the 100% V_{dd}^{Low}, 70% V_{dd}^{Low} and 50% V_{dd}^{Low} configurations, an average delay increase by 18.19 %, 10.53 %, and 11.75 % was observed respectively (considering only those benchmarks with increased critical path length).

C. *Power Comparisons*

Table 2 shows the percentage improvement in total routing power of the benchmarks for the three cases compared with the single high V_{dd} (1.8V) routing architecture. Again, a negative value means that the power has decreased compared to the base case.

From our results we observe that in the case of 100% V_{dd}^{Low} tracks although there is a 33 % savings in power, the worst-case delay increases by about 18 %. In the case of 50-50 mix of V_{dd}^{High} and V_{dd}^{Low} tracks are used, then there is a 9 % saving in power, with 10.53 % worst-case delay penalty. The 30-70 mix gives power saving of about 24% with a worst case delay penalty of 11.75 %. We conclude that the 30-70 mix yields the best power/delay combination among the three cases.

Finally, we performed a study on the contribution of different power components (leakage and dynamic) to the interconnect power. Figure 6 shows the routing power

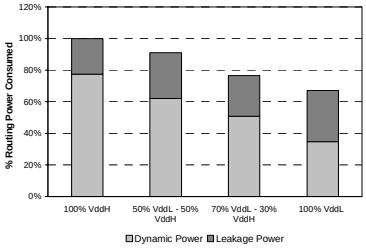

Figure 6. Routing power consumption for the four cases

[1] VPR packs LUTs into the smallest square (can be a rectangular array also) array of CLBs that can fit the given netlist. Then, the router performs routing by using the minimum required number of tracks per channel. Then every channel contains that many tracks.

consumption for each of these cases. The graph also shows the breakup of dynamic and leakage routing power. The values are normalized with respect to the base case, i.e. only using V_{dd}^{High} tracks. This corresponds to 100%. For the base case, leakage power contributes to only 22 %, whereas, when all tracks are supplied with V_{dd}^{Low} the leakage power is 48 %. As we increase the amount of V_{dd}^{Low} tracks, leakage power contribution increases. This increase in leakage power can be attributed to the scaled V_t.

Table 1. % Change of delay

Circuit	% Change in Delay		
	100% V_{dd}^{Low}	50% V_{dd}^{High} 50% V_{dd}^{Low}	30% V_{dd}^{High} 70% V_{dd}^{Low}
alu4	-11.29	-16.20	-8.91
apex2	17.52	12.42	4.48
apex4	22.93	11.06	-5.26
des	3.77	-31.24	-29.06
ex1010	-0.04	-25.55	-36.07
ex5p	11.56	14.80	8.24
misex3	22.54	9.33	4.61
pdc	16.97	-7.40	19.98
seq	28.46	5.62	7.89
spla	-11.81	-26.64	-24.91
bigkey	-13.07	-6.17	-21.07
clma	33.81	7.11	11.15
diffeq	16.20	0.01	1.34
dsip	11.21	38.19	30.53
elliptic	20.65	18.38	5.85
frisc	18.94	6.30	6.25
s298	10.01	-14.48	-4.70
s38417	24.12	13.55	20.73
s38584.1	-2.50	-14.30	-14.54
tseng	14.11	4.28	5.35
Avearge	11.70	-0.05	-0.91

Table 2. % Change of routing power

Circuit	% Change in Routing Power		
	100% V_{dd}^{Low}	50% V_{dd}^{High} 50% V_{dd}^{Low}	30% V_{dd}^{High} 70% V_{dd}^{Low}
alu4	-34.08	-6.95	-25.31
apex2	-44.08	-24.69	-31.06
apex4	-43.08	-22.92	-20.27
des	-41.19	10.78	-4.68
ex1010	-7.55	11.95	11.47
ex5p	-39.45	-25.55	-32.55
misex3	-50.76	-25.53	-33.81
pdc	-11.93	-1.09	-22.24
seq	-50.36	-22.95	-35.13
spla	-8.08	15.97	-1.44
bigkey	-40.73	-17.01	-18.68
clma	-41.06	-16.26	-29.77
diffeq	-36.18	-13.37	-25.20
dsip	-51.40	-40.11	-46.59
elliptic	-36.24	-20.90	-24.15
frisc	-18.35	-7.27	-16.55
s298	-36.78	-7.22	-24.11
s38417	-49.87	-27.63	-39.34
s38584.1	-38.27	-4.46	-19.88
tseng	-43.40	-19.32	-29.72
Avearge	-33.05	-9.10	-23.45

V. CONCLUSIONS

We proposed a low power dual-V_{dd}-dual-V_t routing architecture in this work. First, we studied the relation between performance and power consumption of different routing components. Then, we evaluated our proposed architecture, which is based on combinations of routing tracks supplied at different V_{dd} levels, and the interface between the tracks is enabled using a subset switch box topology. Our experiments with different configurations of the low power routing architecture show that a good organization is to mix 70% of tracks supplied by low V_{dd} with 30% tracks supplied by high V_{dd}. This brings reductions in the interconnect power as high as 47 % and 23.45 % on average. The interconnect power being the dominant component of the overall power consumption in FPGAs such low power routing architectures will prove highly beneficial.

REFERENCES

[1] L. Shang, A. S. Kaviani, and K. Bathala, "Dynamic Power Consumption in Virtex-II FPGA Family," presented at International Symposium on Field Programmable Gate Arrays, 2002.

[2] A. Gayasen, Y. Tsai, N. Vijaykrishnan, M. Kandemir, M. J. Irwin, and T. Tuan, "Reducing Leakage Energy in FPGAs Using Region-Constrained Placement," presented at International Symposium on Field-Programmable Gate Arrays, 2004.

[3] J. Anderson, F. Najm, and T. Tuan, "Active Leakage Power Optimization for FPGAs," presented at International Symposium on Field-Programmable Gate Arrays, 2004.

[4] F. Li, Y. Lin, L. He, and J. Cong, "Low-Power FPGA Using Pre-Defined Dual-Vdd/Dual-Vt Fabrics," presented at International Symposium on Field-Programmable Gate Arrays, 2004.

[5] D. Chen, J. Cong, F. Li, and L. He, "Low-Power Technology Mapping for FPGA Architectures with Dual Supply Voltage," presented at International Symposium on Field-Programmable Gate Arrays, 2004.

[6] F. Li, Y. Lin, and L. He, "FPGA Power Reduction Using Configurable Dual-Vdd," presented at Design Automation Conference, San Diego, CA, 2004.

[7] A.Gayasen, K.Lee, N.Vijaykrishnan, M.Kandemir, M.J.Irwin, T.Tuan, "A Dual-Vdd Low Power FPGA Architecture", *International conference on Field Programmable Logic and Its Application, 2004, Antwerp, Belgium.*

[8] K. Usami and M. Horowitz, "Clustered voltage scaling techniques for low-power design," *ISPLED*, 1995.

[9] [M. Hamada, Y. Ootaguro, and T. Kuroda, "Utilizing surplus timing for power reduction," *Proc. CICC*, 2001.

[10] R. Puri, L. Stok, J. Cohn, D. Kung, D. Pan, D. Sylvester, A. Srivastava, and S. Kulkarni, "Pushing ASIC Performance in Power Envelope," presented at Design Automation Conference, 2003.

[11] [J. T. Kao and A. P. Chandrakasan, "Dual-Threshold Voltage Techniques for Low-Power Digital Circuits," *IEEE Journal of Solid-state circuits*, 2000.

[12] S. Yang, "Logic Synthesis and Optimization Benchmarks," Microelectronics Center of North Carolina 1991.

[13] G. G. Lemieux and S. D. Brown, "A detailed router for allocation wire segments in field programmable gate arrays," *Proc. of the ACM Physical Design Workshop*, 1993.

[14] V. Betz, J. Rose, and A. Marquardt, *Architecture and CAD for Deep-Submicron FPGAs*: Kluwer Academic Publishers, 1999.

[15] K. K. Poon, "Power Estimation for Field Programmable Gate Arrays," in *Dept. of Electrical and Computer Engg.*: University of British Colmbia, 1999.