The Design of a Low Energy FPGA

Varghese George
University of California at Berkeley
Berkeley Wireless Research Center
2108 Allston Wy, Berkeley, CA 94704
(510) 666 3152

Hui Zhang
University of California at Berkeley
Berkeley Wireless Research Center
2108 Allston Wy, Berkeley, CA 94704
(510) 666 3163

Jan Rabaey
University of California at Berkeley
Berkeley Wireless Research Center
2108 Allston Wy, Berkeley, CA 94704
(510) 666 3111

varg@EECS.Berkeley.EDU

hui@EECS.Berkeley.EDU

jan@EECS.Berkeley.EDU

1. ABSTRACT

This work presents the design of an energy efficient FPGA architecture. Significant reduction in the energy consumption is achieved by tackling both circuit design and architecture optimization issues concurrently. A hybrid interconnect structure incorporating Nearest Neighbor Connections, Symmetric Mesh Architecture, and Hierarchical connectivity is used. The energy of the interconnect is also reduced by employing low-swing circuit techniques. These techniques have been employed to design and fabricate an FPGA. Preliminary analysis show energy improvement of more than an order of magnitude when compared to existing commercial architectures.

1.1 Keywords
FPGA, low power, low swing signalling

2. INTRODUCTION
Field Programmable Gate Arrays (FPGAs) are being used increasingly in embedded general purpose computing environments as performance accelerators. This new use beyond the traditional usage as glue logic and as a rapid prototyping enabler has also renewed interest in the FPGA architecture. The fine grain reconfigurability of the FPGA architecture makes it an ideal candidate for use in System-On-Chip environments which strive to integrate heterogeneous programmable architectures. The main task for the FPGA in this context is to efficiently implement late-bound complex functions, or adaptive peripheral modifications.

These advanced design efforts need to tackle the issues of power dissipation and energy efficiency which become increasingly important with high levels of integration. The power dissipation is not only interesting from a packaging perspective, but also in determining the battery life in portable devices where these designs are being used. The problem with using an FPGA in such an environment, is that the fine grain programmability of the FPGA is paid for by poor energy performance.

Fig. 1 shows the power dissipation of commercial FPGAs as a function of the clock frequency. It can be seen that running at the system frequencies of the chip is not possible with conventional cheap plastic packaging. In a portable environment like a cell phone with a power budget of the order of milliwatts, the present FPGA architectures will dominate the power budget allocation.

In this work, the FPGA architecture is designed primarily for energy efficiency while maintaining speed performance. To do this, the architectural and circuit optimizations were done concurrently to obtain energy efficiency. Energy-Delay(ED)
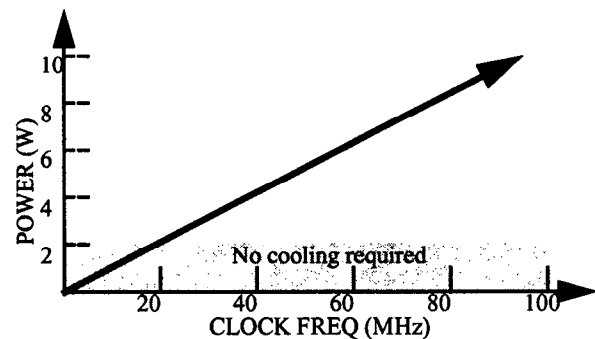

Figure 1. Power Consumption of Commercial FPGAs

product has been used as the optimization metric to ensure that low energy is not obtained by sacrificing speed performance.

3. OVERVIEW
The fine grain programmability of the FPGA puts stress on the interconnect structure. In this paper, "*interconnect*" means all of the components which contribute to providing connection between logic blocks. This includes the connection box (C Box), metal routing, and the switch box (S Box). The speed and energy performance of the FPGA are dominated by the interconnect. This is illustrated in Fig. 2, which shows the power breakdown of an XC4003A FPGA [5]. The interconnect is responsible for most of the energy

consumption, while the logic consumes only 5% of the total energy. This breakdown is valid for the latest FPGAs since the architecture has remained more or less the same, and only the process technology has changed.

Work has been done to optimize the size of the lookup table (LUT) [8], the depopulation of latches, and the interconnect architecture [1][3][9][10]. All of this work has been concentrated in improving the area and speed performance of FPGAs.

The design of an FPGA requires both architectural analysis and circuit level innovations. This has to be done concurrently so that the circuit level parameters can be used to update the architecture model.
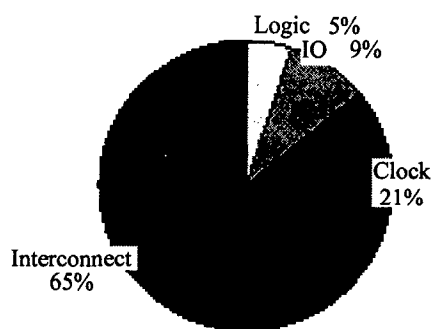

Figure 2. Energy Breakdown of XC4003A [5]

It is important that in the process of interconnect optimization, the routability of the architecture does not deteriorate. To aid the architectural evaluation, a complete placement and routing tool was developed here. The tool accepts as one of its inputs the description of the FPGA architecture. The description includes the availability of interconnect resources, and the costs associated with them. The goal of the tool is to place and route a set of benchmark netlists on this architecture description, while minimizing the total cost. This methodology has been used to evaluate the FPGA architecture described in this paper.

4. INTERCONNECT ARCHITECTURE
Emphasis was placed on the interconnect architecture during the optimization phase. The connectivity between the Configurable Logic Blocks(CLB) is obtained through three levels of interconnect architecture

- Level0 - Nearest Neighbor
- Level1 - Mesh Architecture
- Level2 - Hierarchical Interconnect

Each of the architecture levels is designed to provide low RC connections to nets of different lengths.

4.1 Level0 - Nearest Neighbor Connections
The Level0 connections are targeted at providing connections to the neighboring CLBs. The cost of these connections increases proportionally to the number of

neighbors each CLB can connect to. As an example, if the connections were to the 8 closest neighbors, the fanout, and hence the capacitive loading on the connection, is 8. If, on the other hand, the Level0 were to include the next ring of neighbors, the fanout triples to 24. It is evident that after a certain point the sheer fanout will make this connection very expensive from an energy point of view.

The Level0 structure employed in the present architecture is as shown in Fig. 3. In our study it was found that a structure supporting connections to 8 of its nearest neighbors was optimal. Compared to the traditional Mesh connection, the Energy-Delay product of this connection is smaller by ~x3.
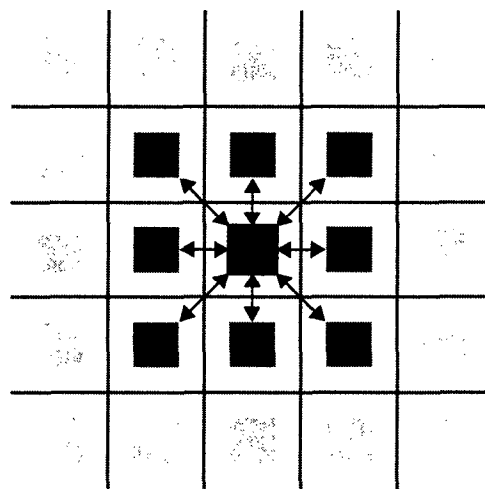

Figure 3. Nearest Neighbor Connections

4.2 Level1 - Mesh Interconnect
The next level of connections is through a symmetric mesh architecture, as shown in Fig, 4. This provides connections
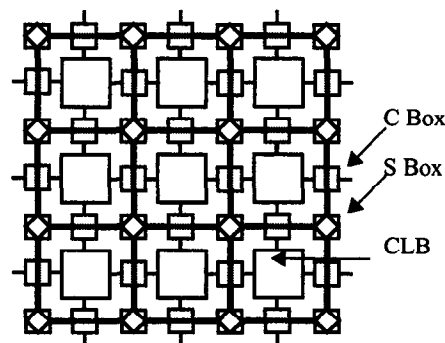

Figure 4. Symmetric Mesh Architecture

to blocks which cannot be reached through the NNC connections. The C Box provides full connectivity to the channel, and the S Box is Xilinx style. No major architectural modifications have been done at this level, as the basic structure provides good routability.

4.3 Level2 - Hierarchical Interconnect
For longer connection lengths(l) between the logic blocks,

189

the delay increases as l^2 and Energy-Delay as l^3 in a Mesh architecture. This can be circumvented by having another level of interconnect which is dedicated for longer connections.

Hierarchical interconnect has been proposed in prior studies [2][6] to improve the speed of FPGA architecture. Fig. 5 compares the ED metric of connections using the Mesh architecture and a binary tree connectivity in a 16x16 array.
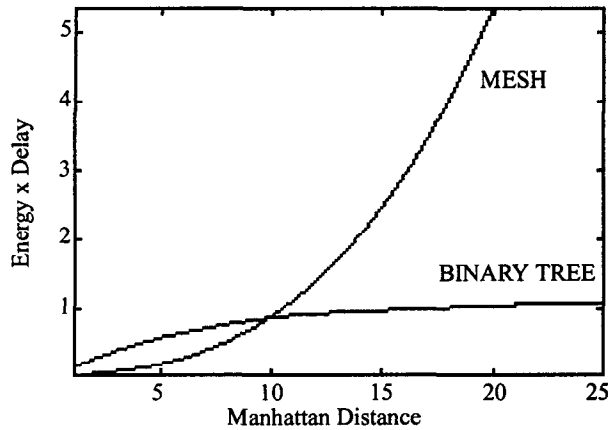

Figure 5. Comparison of Mesh and Binary Tree Connectivity

Compared to a pure Mesh architecture, or a Binary tree-like hierarchical architecture, a hybrid architecture using both of these structures is more attractive. The shorter connections (in this case <10) are better if routed in the mesh structure, while the longer connections should be routed in the binary tree structure. Due to this reason we have designed a hybrid architecture incorporating both the Mesh and Tree structures.
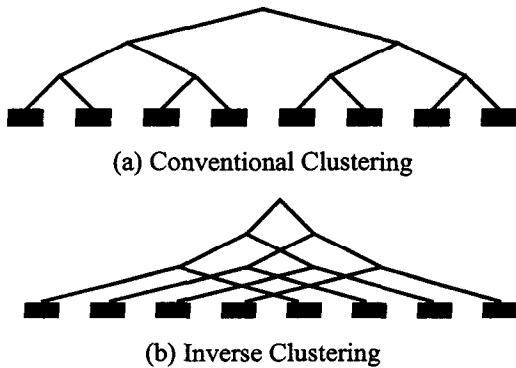

(a) Conventional Clustering

(b) Inverse Clustering

Figure 6. Clustering for Hierarchical connection

Further optimization was done on the clustering of logic blocks. In the Level2 interconnect, the grouping of the logic blocks for the hierarchical structure is done as an inverse cluster. Fig. 6a shows the conventional clustering of logic blocks for hierarchical connections. The closer logic blocks are connected at the lowest level, and the farthest blocks have to go through all the levels of switches to be

connected. This is an inefficient method since for closer connections the routing will be through the Level1 Mesh according to Fig. 5.

The inverse clustering mechanism is shown in Fig. 6b. The longer connections are connected using the lowest level of the tree, and the closer connections have to traverse the entire tree. This ensures that for long connections which get routed on the Level2 interconnect, the number of switches traversed is small.

5. CLB ARCHITECTURE

The contribution of the CLB to the total energy is negligible. During the design process the interest in the CLB structure was in how it effects the interconnect utilization, and hence the interconnect energy. The CLB structure chosen was a cluster of 3-input lookup tables (LUTs). This enables us to configure the CLB for bitwise datapath operation, or combined to form a larger LUT for random logic without wasting CLB resources.

Prior study has shown that a cluster of 4 3-input LUTs, shown in Fig. 7, is optimal from an energy perspective [4]. This structure is capable of implementing a 5 input combinational logic, or a 2 bit arithmetic operation.
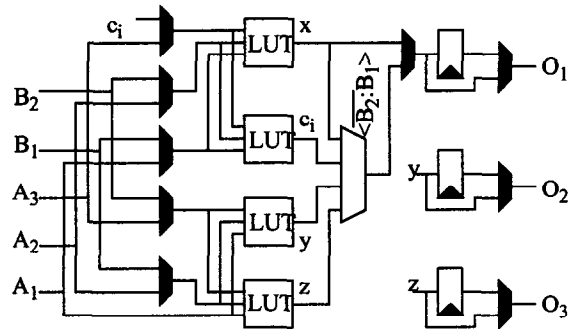

Figure 7. CLB Structure

All three outputs of the CLB can be latched. This makes it possible to implement high speed datapath operations pipelined at the bit level, since even the carry can be pipelined.

6. CLOCK DISTRIBUTION

In the data shown in Fig. 2 the contribution of the clock to the total energy dissipation is ~20%. This number can be as high as 50% for highly pipelined circuits.

Early in the design process it was seen that most of the clock energy is dissipated in the global distribution network. In our design, Dual Edge Triggered flip-flops [7] are used in the CLBs. This reduces the activity in the clock distribution network by a factor of 2. The increase in complexity because of this change is minimal at the top level. In our design with 3 flip-flops per CLB, the contribution in area from the flip-flops is ~0.8%.

The low-swing signalling described in section 7.2 is used on

the global clock network to reduce the energy further.

7. CIRCUIT LEVEL OPTIMIZATION

To complement the optimization done at the architecture level, proper circuit level design is imperative. In this work, the modifications were concentrated on the interconnect. The main issues dealt with were positioning in the Energy-Delay design space and low swing interconnect.

7.1 Energy-Delay Design Space

The connecting path from one CLB to another is an *RC* chain. The series transistors in the path form the resistors, and the main contribution to the capacitance is from the unused transistors hanging off of the path. The proper design of the interconnect switches is crucial for optimizing the energy of the design. Optimizing the transistor sizes for speed performance can have a dramatic effect on the energy performance. These performance criteria have to be balanced. To study the trade-off, the Energy-Delay curve of a typical interconnect path is used. Figure 8 shows a typical interconnect path in an FPGA architecture. The highlighted devices define the path between the two CLBs.
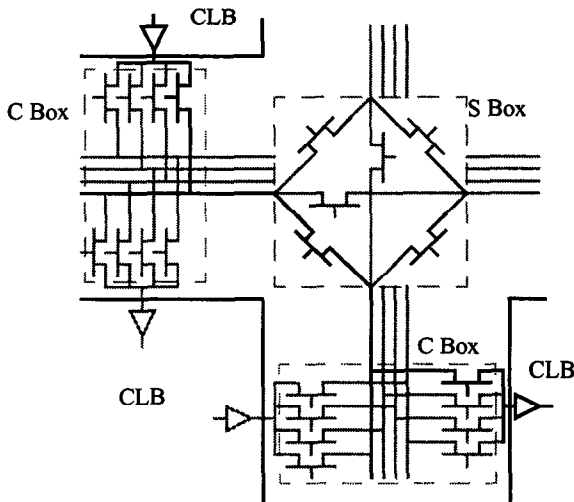

Figure 8. Typical FPGA Interconnect Path

The Energy-Delay Design Space of the above path is shown in Fig. 9.

It is crucial for energy efficiency that you the design is carefully positioned in this space. As can be seen, at the limits of the given technology, the incremental improvement in delay is paid for in terms of higher energy.

7.2 Low-Swing Circuit (SDVST-II)

$$E = C * V_{swing} * V_{supply} \qquad \text{(EQ 1)}$$

Eq. 1 supports usage of low swing signalling, especially in light of the fact that the interconnect dominates the FPGA energy. Almost all low-swing circuit techniques [11] have been targeted at busses, and similar interconnect structures where the load capacitance is accurately known, and there are timing pulses to control the low-swing circuitry. Neither
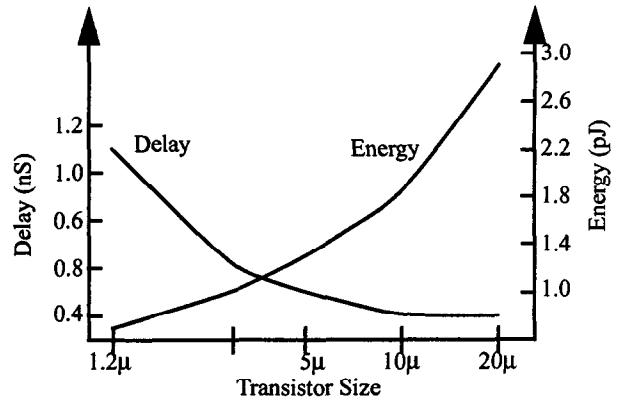

Figure 9. Energy-Delay Design space

of these conditions is valid in an FPGA interconnect. The capacitance is a function of the connection length (number of segments used), and clocking pulse is dependent on the circuit being implemented.
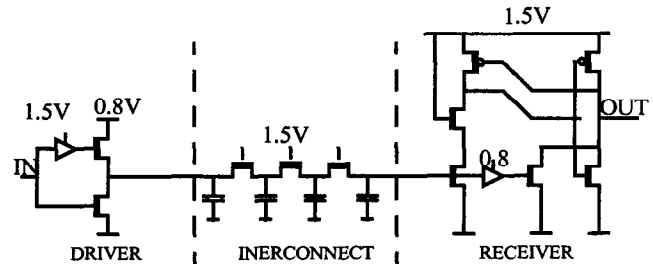

Figure 10. Low-Swing Circuit (SDVST-II)

The low-swing circuit used in this FPGA architecture is shown in Fig. 10. In the proposed circuit, the interconnect swing is at 0.8V, while the rest of the circuit runs at 1.5V. The common drawback of single ended low-swing methods using conventional techniques is the slow speed of the receiver circuit, and the short-circuit current at the receiver end. The proposed circuit employs cascode circuitry and differential circuits at the receiver end to mitigate these effects. The comparison to a Full-Swing circuit is given in Table.1. The energy and delay values include the contribution of the driver and receiver circuits.

Table 1: SDVST-II vs. Full swing circuit

Circuit	E (pJ)	D (nS)	ED
Full Swing	72.3	1.9	137
SDVST-II	31.4	2.3	72

As can be seen, the SDVST-II technique is better by a factor of 2 over the conventional method. This is used on all three levels of interconnect.

191

8. IMPLEMENTATION - LP_PGA

The proposed techniques were implemented in a prototype array of 8x8 logic blocks. The size of the array is 2mmx2mm in a 0.25U 6 metal CMOS process. The chip layout is shown in Fig. 11.
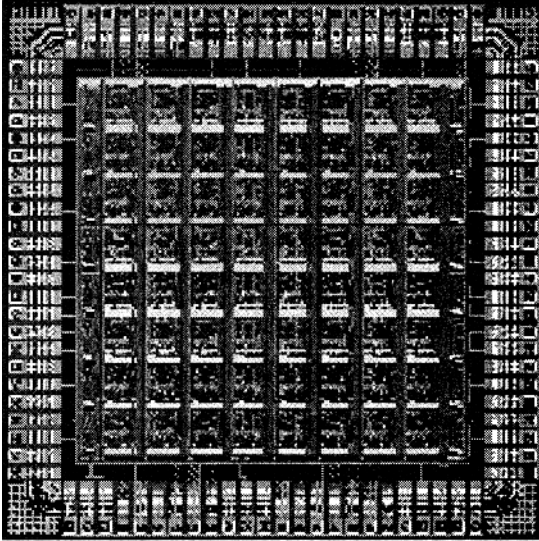

Figure 11. Prototype 8x8 chip

9. RESULTS

The results were obtained by simulating the extracted netlist from the final layout. The simulations at the chip-level was done using PowerMill.

For comparison purposes, published information from the data books [12][13] of commercial FPGAs was used.

9.1 Array filled with 8 bit counters

For this comparison, the energy dissipation of one flip-flop driving a 9 segment long interconnect is measured. A 1024 logic block array is assumed to be filled with these elements configured as 16 bit counters. This gives a 12.5% activity factor on the interconnect. All the logic elements are clocked at 30MHz. The XC4000XL series from Xilinx is the low voltage version running at 3.3V. The FLEX10K is the Altera FPGA running at 5V. For computing the Xilinx and Altera energy, the methods described in the application notes [12][13] is followed.

9.2 Clock Rate

For the LP_PGA, since the latches are dual edge triggered, the clock rate required for the same data throughput is half of the conventional single edge triggered latch. Hence, in the previous experiment, the LP_PGA had to be run at only 15MHz to achieve the same data throughput. Similarly, when reporting the Toggle frequency even though the LP_PGA's is only 62.5MHz, it is effectively 125MHz when considering a single edge triggered clock.

Table.2 gives the comparison numbers.

Table 2: Comparison of FPGA architectures

	Xilinx XC4000XL	Altera FLEX10K	LP_PGA
Single FF driving 9 segments (pJ)	320	485	4.85
1K Array of 16 bit counters at 30MHz(mW)	1.5×10^3	3.1×10^3	21.8
Maximum Toggle Frequency (MHz)	166	100	62.5 (125)

10. CONCLUSION

The prototype of a Low Energy FPGA suitable for embedded and portable applications has been designed and implemented. It has been shown that a combination of architectural redesign and careful circuit design can improve that energy efficiency by more than an order of magnitude.

11. REFERENCES

[1] Betz, V., Rose, J. Cluster-based logic blocks for FPGAs: area-efficiency vs. input sharing and size in *Proceedings of the IEEE 1997 Custom Integrated Circuits Conference*, May 1997, 551-554.

[2] Chung, K., *et al.* Using Hierarchical Logic Blocks to improve the Speed of FPGAs in *International Workshop on Field Programmable Logic and Applications*, Oxford, UK, Sept. 1991, 4-6.

[3] Gallia, J.D., *et al.* A Flexible Gate Array Architecture for High-Speed and High-Density Applications in *IEEE J. Solid State Circuits*, vol. 31, no. 3, March 1996, 430-436.

[4] George, V. Effect of Logic Block Granularity on Interconnect Power in a Reconfigurable Logic Array. URL: http://bwrc.eecs.berkeley.edu/people/Grad_students/varg/Reports/CS294/.

[5] Kusse, E., and Rabaey, J. Low-Energy Embedded FPGA Structures in *1998 International Symposium on Low Power Electronics and Design*, Aug. 1996, 155-160.

[6] Lai, Y., *et al*, Hierarchical interconnection structures for field programmable gate arrays in *IEEE Transactions on Very Large Scale Integration Systems*, vol.5, no. 2, June 1997, 186-196.

[7] Llopis, R.P., Sachdev, M. Low power, testable dual edge triggered flip-flops in *1996 International Symposium on Low Power Electronics and Design*, Aug. 1996, 341-345.

[8] Rose, J., *et al.* Architecture of Field-Programmable

Arrays: The Effect of Logic Block Functionality on Area Efficiency in *IEEE J. Solid State Circuits*, vol. 25, no. 5, Oct. 1990, 1217-1225.

[9] Rose, J., *Brown, S.* Flexibility of Interconnection Structures for Field-Programmable Gate Arrays in *IEEE J. Solid State Circuits*, vol. 26, no. 3, March 1991, 277-282.

[10] Singh, S., *et al.* The Effect of Logic Block Architecture on FPGA Performance in *IEEE J. Solid State Circuits*,

vol. 27, no. 3, March 1992, 281-287.

[11] Zhang, H., and Rabaey, J. Low Swing Interconnect Interface Circuits in *1998 International Symposium on Low Power Electronics and Design*, Aug. 1998, 161-166.

[12] Xilinx XC4000XL Power Calculation. URL: http://www.xilinx.com/xcell/xl27/xl27_29.pdf.

[13] Evaluating Power for Altera Devices. URL: http://www.altera.com/document/an/an074.pdf.