A 1.1 GOPS/mW FPGA Chip with Hierarchical Interconnect Fabric

Cheng C. Wang, Fang-Li Yuan, Henry Chen, Dejan Marković

Electrical Engineering Department, University of California, Los Angeles, CA

Abstract

A 2048 look-up-table FPGA with a radix-2 hierarchical interconnect network is realized in 3.94mm^2 in 65-nm CMOS. It has an interconnect-to-logic area ratio of 1:1, which is a 3–4x reduction from modern FPGAs while allowing up to 100% resource utilization. As a proof of concept, it is designed with standard cells, achieving 16.4 GOPS/mm^2 at 370MHz. Peak energy efficiency of 1.1 GOPS/mW is measured at 0.5V.

Introduction

Field-programmable gate arrays (FPGAs) are effective for rapid verification and prototyping of VLSI designs. They are also used in products that require periodic hardware changes and short time to market. However, FPGAs incur penalties in area (17–54x), speed (2.5–6.7x), and power (5.7–62x) over standard-cell ASICs [1], hindering their expansion into ASIC markets. The overhead is primarily due to interconnects, which account for over 75% of area and delay.

For over 20 years, FPGAs have used 2D-mesh interconnects, where look-up tables (LUTs) are placed in configurable logic blocks (CLBs), and arrays of switch boxes are placed at interconnect crossings (Fig. 1). Since a full array requires too much area, various heuristics are used to simplify switch-box arrays at the cost of resource utilization. Yet 80% of the 1.1B transistors on Virtex-5 are used for interconnects [2]. This paper demonstrates an FPGA with hierarchical interconnects where interconnect area is 51%, a 3–4x reduction from commercial FPGAs while preserving connectivity. An energy efficiency of 1.1 GOPS/mW is the highest among reported FPGAs. The chip is tested up to 400MHz.

Hierarchical Interconnect Architecture

The key issue with 2D-mesh is scalability; the number of switch boxes grows as $O(N^2)$ with the number of LUTs. Using Rent's rule, interconnect complexity is still $O(N^{1.75})$ for random logic, requiring FPGA size to scale much faster than Moore's Law. In the proposed hierarchical interconnect, a folded Beneš network is employed to reduce the complexity to $O(N \cdot logN)$ [3]: 4 LUTs are connected via 2 stages of switch matrices (SMs), and another 4 LUTs are connected with a 3rd SM stage (Fig. 2a). Each SM has 4 unidirectional connections per direction. Although this architecture reduces interconnect complexity, each SM stage doubles the routing congestion. This $O(N)$ congestion makes physical design difficult.

To alleviate congestion, routing is alternated between x-y directions to reduce congestion to $O(N^{0.5})$ (Fig. 2b). At every hierarchy, the LUTs near the center are interconnected to create shorter routes, and the edge routes are longer. This gives routing tools options for faster paths on timing-critical routes.

The test chip has 2048 4-input LUTs: 1024 LUTs form 256 Logic CLBs, 896 LUTs form 224 DSP CLBs, and 128 LUTs form 16 Block RAMs (BRAMs) of 1kb each. In practice, the majority of the logic connections are local, requiring fewer connections on upper hierarchies. Therefore full connectivity is preserved up to 6 SM stages (Fig. 3a), then half-connectivity SMs are used to reduce the complexity of upper hierarchies. This partitions the interconnect into 3 sub-networks: $N_{8:2}$, $N_{6:2}$, and $N_{6:1}$. The chip is divided into 16 macros (Fig. 3b). Macros $N_{8:2}$ are centered for shorter top-level routing, branching into $N_{6:2}$ and $N_{6:1}$. Each of the macros contains 32 CLBs—a combination of Logic, DSP, and BRAM (Fig. 3c).

Circuit Implementation

The CLBs include four 4-input LUTs with selectable asynchronous/synchronous output stages (Fig. 4a). Each LUT is configurable as one 4-input LUT or two 3-input LUTs with up to 4 unique inputs. A Logic CLB includes a carry chain to support 4b additions where Propagate and Generate are driven from LUTs. The Logic CLB is especially useful when two outputs per bit are required, such as in 3:2 compressors.

The DSP CLB (Fig. 4b) has a LUT combiner to support 5/6-input LUTs, and a carry chain that is configurable as one 8b or two 4b adders. The adder cells are shared with a 4b×4b Wallace-tree multiplier. Based on the configuration, the appropriate outputs are sent to the output stage. Due to the level of configurability, the synthesized CLB has 50 logic gates on its critical path (shaded), amounting to a 1.1ns delay.

Configuration bits are required to control CLBs and SMs, but traditional SRAM arrays are not suitable because all bits cannot be accessed simultaneously. A scan chain is adopted in [4] to control 6 CLBs, but it is not scalable to larger designs. Therefore an SRAM-based bit cell (BC) is designed where the output of each BC is directly routed to the configuration inputs of CLBs and SMs (Fig. 5a). The BC area is 5x smaller than a DFF-based scan cell. The bit-line (BL) and word-line (WL) controls are implemented as scan chains to write one row of BCs at a time. The BC arrays are local to each CLB, so only the BL and WL controls are propagated to top level. Overall, the memory area is reduced (Fig. 5b), and total interconnect area is 51%, a 3–4x reduction over 2D-mesh [5] for a fixed logic area.

Automated Mapper

An automated mapper is developed to map RTL onto this FPGA. A standard-cell library of LUT functions is created to enable logic synthesis using commercial tools. The LUT netlist is imported into an automated, custom place-and-route tool that generates the bitstream for FPGA programming. This tool is also used during architecture design to evaluate interconnect connectivities by mapping Toronto20 benchmarks.

Measurement Results

Our chip achieves 16.4 GOPS/mm^2 when all Logic and DSP CLBs are utilized, executing 175 16b accumulators at 370MHz. Since a 16b adder uses 2 DSP CLBs or 4 Logic CLBs, the DSP adders are faster, reaching 400MHz. Performance is hindered by equipment limitations due to a 0.25ns input-clock jitter at 400MHz. The energy-delay curve and the power breakdowns for minimum delay and minimum energy are shown in Fig. 6.

In comparison, [4] has no interconnects, the full-custom CLB in 32-nm LVT is 2.5x faster, but achieves 2.6 GOPS/mW at 0.34V for 8b operations, which is 0.65 GOPS/mW for 16b (2 CLBs per operation at half the speed). With interconnects, our 65-nm chip reaches 1.1 GOPS/mW at 0.5V.

Leakage is well-controlled even without power gating. A 1.08 GOPS/mW is attainable with only 112 DSP accumulators active and most of the Logic CLBs idle (Table I). The FIR filter achieves 274MHz due to longer routing, but interconnect delay is still under 50%. The 2×2 MIMO FFT uses 10 BRAMs to implement various delay lines. With many control signals and a critical path of 11 CLBs, the FFT achieves 83MHz.

Figure 7 shows the die photo. The top 3 metal layers (out of 9) are sparsely used, leaving ample room for larger designs.

Acknowledgments

We thank STMicroelectronics and C. Yang for helpful discussions.

References

[1] I. Kuon *et al.*, *Found. Trends in Elec. Design*, 2008.
[2] I. Bolsens, *MPSOC*, 2006.
[3] V. Konda, *U.S. Patent 2010/0172349*.
[4] A. Agarwal *et al.*, *ISSCC Dig. Tech. Papers*, 2010.
[5] M. Lin *et al.*, *FPGA '06*.

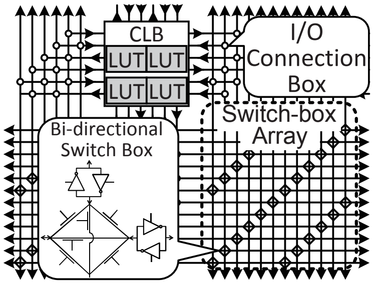


Figure 1: 2D-mesh interconnect.

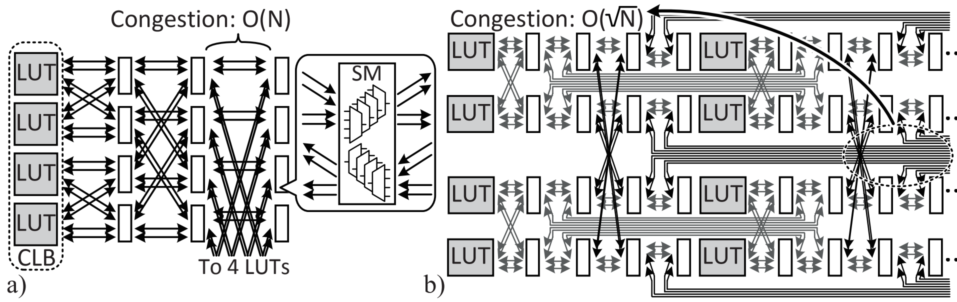


Figure 2: a) Hierarchical routing of 8 LUTs (4 shown) using SMs, b) alternated x-y routing.

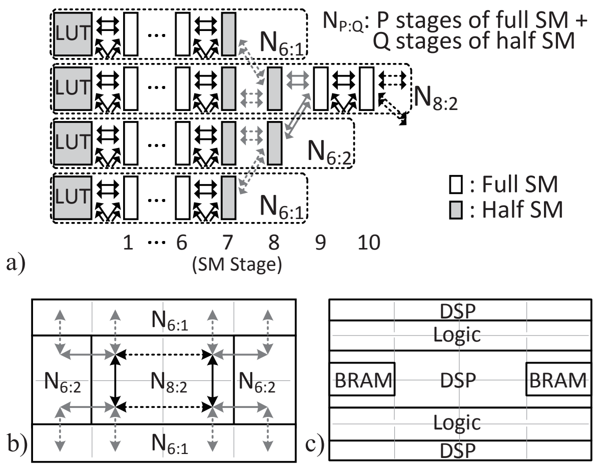


Figure 3: a) Interconnect architecture of 2048 LUTs, floorplan of b) SM network, c) CLB placement.

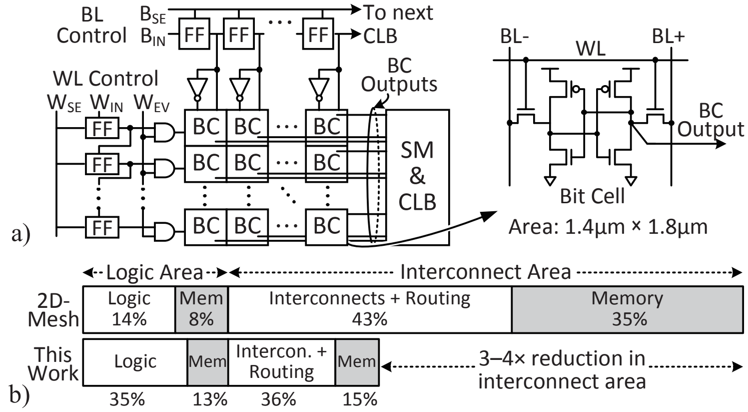


Figure 5: a) Bit cell (BC) configuration circuitry, b) area comparisons of 2D-mesh vs. this chip for a fixed logic area.

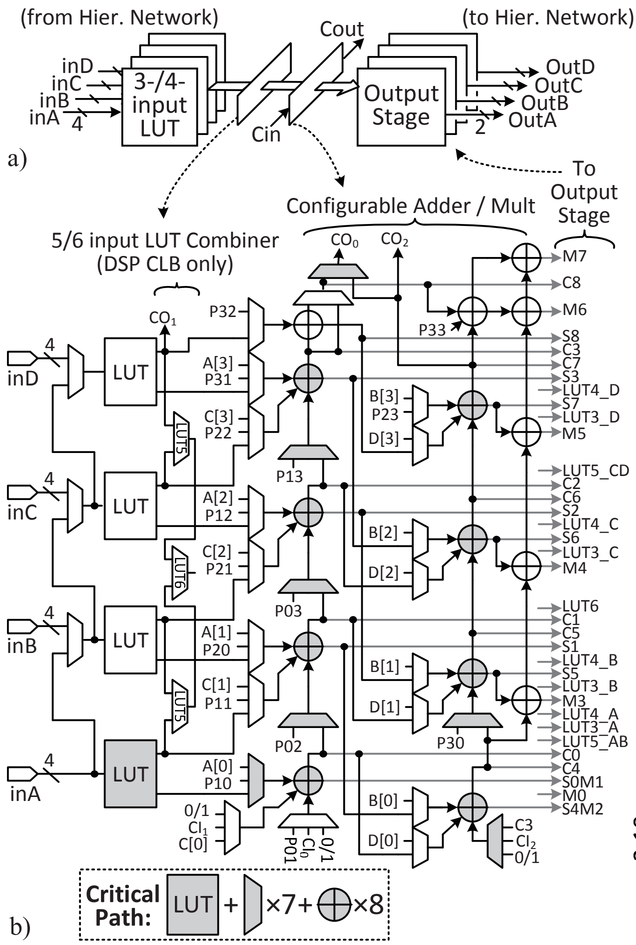


Figure 4: a) CLB block diagram and b) DSP CLB schematic.

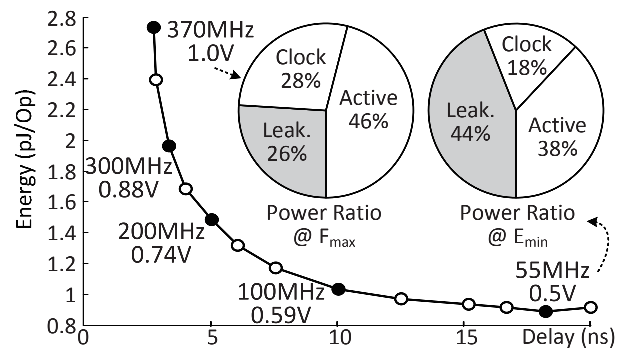
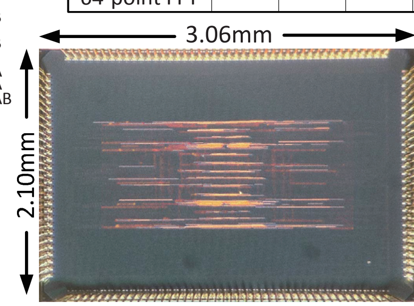


Figure 6: Energy-delay curve of the mapped 175 16b accumulator with power breakdown at F_{max} and E_{min} (insets).

TABLE I: MEASUREMENT RESULTS.

Result / Design	Resource Utilization: Logic (256)	DSP (224)	BRAM (16)	Performance: Power (mW)	V_{DD} (V)	Freq. (MHz)	GOPS /mW
175 Logic+DSP 16b Accum.	256	224	0	179 8.6	1.0 0.50	370 55	0.36 1.13
112 DSP 16b Accum.	4	224	0	123 6.2	1.0 0.51	400 60	0.57 1.08
32-tap 16b FIR Filter	132	209	0	120 10.2	1.0 0.56	274 50	0.21 0.45
2×2 MIMO 64-point FFT	196	93	10	82.7 26.5	1.0 0.78	83 40	0.05 0.07



Technology	65nm 1P9M CMOS
Core V_{DD}	0.34 to 1.0V
Frequency	40 to 400MHz
I/Os	75 bidirectional
Core Size	2.52mm × 1.56mm
Gate Count	2.73M
CLB Count	256 Logic, 224 DSP
Block RAMs	16 128×8b
Config. Bits	297,472b

Figure 7: Die micrograph and chip summary.