EXHIBIT 3

27.5 A Multi-Granularity FPGA with Hierarchical Interconnects for Efficient and Flexible Mobile Computing

Cheng C. Wang[1], Fang-Li Yuan[1], Tsung-Han Yu[2], Dejan Markovic[1]

[1]University of California, Los Angeles, CA, [2]Qualcomm, Irvine, CA

Following the rapid expansion of mobile computing in the past decade, mobile system-on-a-chip (SoC) designs have off-loaded most compute-intensive tasks to dedicated accelerators to improve energy efficiency. An increasing number of accelerators in power-limited SoCs results in large regions of "dark silicon." Such accelerators lack flexibility, thus any design change requires a SoC re-spin, significantly impacting cost and timeline. To address the need for efficiency and flexibility, this work presents a multi-granularity FPGA suitable for mobile computing. Occupying 20.5mm^2 in 40nm CMOS, the chip incorporates 2,760 fine-grained configurable logic blocks (CLBs) with 11,040 6-input look-up-tables (LUTs) for random logic, basic arithmetic, shift registers, and distributed memories, 42 medium-grained 48b DSP processors for MAC and SIMD operations, 16 32K×1b to 512×72b reconfigurable block RAMs, and 2 coarse-grained kernels: a 64-8192-point fast Fourier transform (FFT) processor and a 16-core universal DSP (UDSP) for software-defined radio (SDR). Using a mix-radix hierarchical interconnect, the chip achieves a 4× interconnect area reduction over commercial FPGAs for comparable connectivity, reducing overall area and leakage by 2.5×, and delivering a 10-50% lower active power. With coarse-grained kernels, the chip's energy efficiency reaches within 4-5× of ASIC designs.

Although commercial FPGAs can come close to ASICs in performance, they are highly inefficient due to their high energy and a large area overhead. This is mainly due to the programmable interconnect. For over 20 years, a 2D-mesh network has been the backbone of FPGA interconnect, but full connectivity in a 2D mesh requires O(N^2) switches, requiring interconnects to grow faster than Moore's Law O(N). As a result, various heuristics are used to simplify switches at the cost of resource utilization, but the interconnect area is still ~4× the logic area in modern FPGAs. By effectively pruning a Beneš network, a hierarchical interconnect network is realized where the number of switches is less than O(N·logN), allowing us to maintain an interconnect-to-logic-area ratio of 1:1.

The O(N·logN) complexity of Beneš network is well-known in telecommunications, but such a network is seldom used in hardware primarily due to its implementation complexity. In a traditional Beneš, wirelength doubles for every stage. With an equal number of wires for all stages, this leads to long, congested wires in the upper hierarchies. An efficient implementation requires pruning the upper hierarchies, and we alternate the routing in the x- and y-directions so wirelength doubles every *two* stages [1]. Another drawback is the delay across radix boundaries. As shown in Fig. 27.5.1, communication between neighboring computing elements (CE) 4 and 5 requires 3 hierarchies. A boundary-less radix-3 network is created to restore spatial locality by shifting all local connections to the lower switch matrices (SMs). In the simplified illustration, radix-3 SMs are used in the lower stages to increase local bandwidth, allowing even fewer radix-2 SMs in the upper hierarchies. For improved timing and reduced power, fast-path routing allows hops directly to the required hierarchy level, routing only half the network on the return path. Our router automatically assigns fast-path interconnect based on congestion and timing.

Boundary-less radix-3 SMs are used in the lower 5 hierarchies (Fig. 27.5.2), and pruned radix-2 switches are used from stage 6 to 14, except stage 10 and 11. Stage 10 employs boundary-less radix-3 across the horizontal bisection to improve bisection bandwidth. The top-level connectivity (stage 14) is pruned to only 5%. This is a result of closed-loop optimization by mapping various FPGA benchmarks, then pruning or expanding each stage based on congestion and performance. To ease physical design, the chip is divided into 40 interconnect regions, each with 512 SM macros, with 9 to 14 stages per SM macro.

The fine-grained and medium-grained CLBs offer behavior identical to commercial FPGAs, allowing for a direct comparison of interconnects by executing identical netlists. To target common communications designs, two coarse-grained kernels were implemented. A 64-8192-point reconfigurable FFT is beneficial for digital baseband processing. It has a small dedicated memory, and interconnects to the FPGA memory to realize the long delay lines for 2048-8192-point FFTs. A 16-core UDSP targets a variety of SDR algorithms, where butterfly core in the UDSP is very efficient for complex arithmetic, capable of many SDR functions, such as filtering, equalization, CORDIC, and sphere decoding by simply concatenating multiple butterfly stages. FFT and UDSP both connect to the interconnect network.

Power gating (PG) is desirable for large chips, but each interconnect signal often traverses many blocks, making block-level PG ineffective. A fine-grained PG is needed for individual switches. Traditional PG becomes very inefficient because the footer PG transistor is no longer shared by the entire block, so it cannot be made very large (Fig. 27.5.3), but a smaller footer can degrade performance by 30-50%. To power gate without a footer, a PG branch is added to the mux, and the pass-gate is separated into NMOS and PMOS segments, where enabling PG leaves the output floating, reducing the coupling capacitance on neighboring wires. When conducting, the NMOS segment is driven by PMOS pass-gates, thus it can rise much faster than the PMOS segment driven by NMOS pass-gates, which settles to V_{DD}-V_t (and vice versa). This results in larger transient leakage, but does not degrade performance significantly, because the output current is the *difference* of the pull-up and pull-down branches. A small high-V_t keeper pulls together the NMOS and PMOS voltages to overcome the V_t drop. This results in a 5-10% performance penalty, but reduces leakage by more than 50% (now gate-leakage dominated). The output floats during PG, so it cannot drive a CMOS gate, but can only enter a pass-gate that can be disabled during PG. Over 90% of the switches utilized this PG scheme, except those driving long wires that require buffer insertion.

With over 9 million configuration bits, an automated mapping tool is developed. The tool supports two modes (Fig. 27.5.4). Mode 1 maps an identical netlist as used by commercial FPGAs for a direct comparison of performance, power, and area utilization: the user design is first synthesized using commercial tools, then the output netlist is parsed into our custom tool, which performs timing analysis, floorplan, placement, routing, and bitstream generation for our FPGA. Mode 2 incorporates our coarse-grained kernels into the P&R flow. Although the configuration SRAM cells are distributed throughout our FPGA, their word-lines (WL) and bitlines (BL) are organized as one large memory for easy initialization. The FPGA core can only be powered on after configuration finishes.

Measurement results of our FPGA with CLBs, and with coarse-grained kernels are compared against processors, a commercial FPGA, and an ASIC (Fig. 27.5.5). Although the CLBs alone achieve over 1.5GOP/mW, an energy efficiency of 0.86GOPS/mW is achievable when mapping an FIR filter, which is 4× more efficient than commercial FPGA (both in 40nm). An 8× efficiency gain can be achieved by using UDSP kernels. FFT operations, which are dominated by memory and control, are 13× more energy efficient when mapped to the FFT kernel instead of CLBs. A 2-2.5× reduction in leakage is attained from smaller chip area and fine-grained PG, even with the disadvantage of dual-oxide transistors. Our chip is built with standard-cells, yet we are often within 20% of the performance of high-end FPGAs, though our software is still improving.

With efficient interconnect, our FPGA is within 20× of ASIC efficiency for most designs (Fig. 27.5.6). Coarse-grained kernels further improve the efficiency, bringing it within 4 to 5× of ASICs. The key to coarse-grained efficiency is to identify compact, reconfigurable kernels that improve efficiency, apply to a variety of applications, and leverage existing FPGA resources where possible. Our chip (Fig. 27.5.7) is able to attain the energy efficiency suitable for mobile applications while maintaining the full flexibility of an FPGA.

Acknowledgments:
The authors thank Dr. Sanjay Raman and DARPA for funding support.

References:
[1] C.C. Wang, *et al.*, "A 1.1 GOPS/mW FPGA Chip with Hierarchical Interconnect Fabric," *IEEE Symp. VLSI Circuits*, pp. 136-137, 2011.
[2] Z. Yu, *et al.*, "An 800 MHz 320 mW 16-Core Processor with Message-Passing and Shared Memory Inter-Core Communication Mechanisms," *ISSCC Dig. Tech. Papers,* pp. 64-65, 2012.
[3] "FFT Implementation on the TMS320C5535 DSP," *TI Technical Reference Manual*, pp. 111-134, 2012.
[4] T-H. Yu, *et al.*, "A 7.4 mW 200 MS/s Wideband Spectrum Sensing Digital Baseband Processor for Cognitive Radios," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 2235-2245, 2012.
[5] F-L. Yuan, *et al.*, "A 256-Point Dataflow Scheduling 2x2 MIMO FFT/IFFT Processor for IEEE 802.16 WMAN," *Asian Solid-State Circuits Conf.*, pp. 309-312, 2008.
[6] J. Thompson, *et al.*, "An Integrated 802.11a Baseband and MAC Processor,"

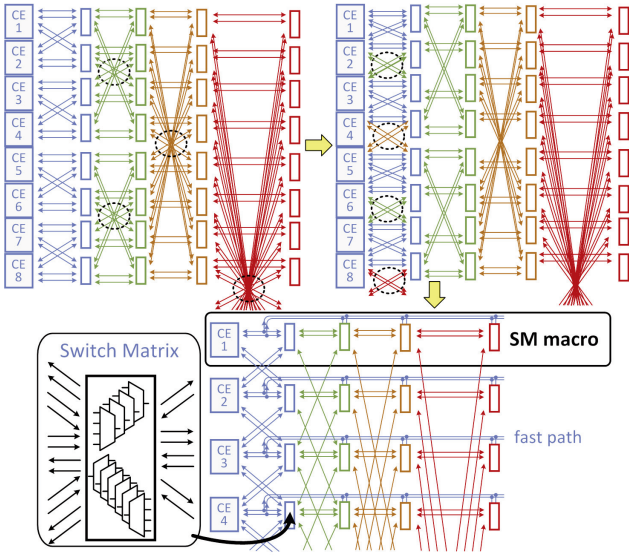


Figure 27.5.1: A boundary-less radix-3 Beneš network.

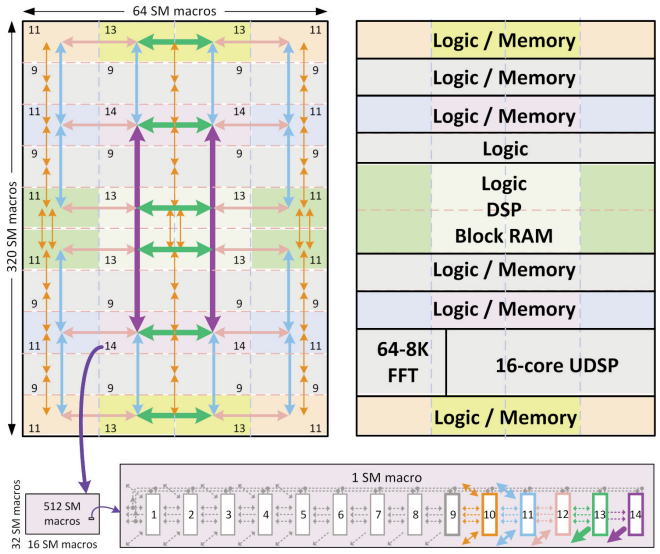


Figure 27.5.2: Interconnect and resource allocation.

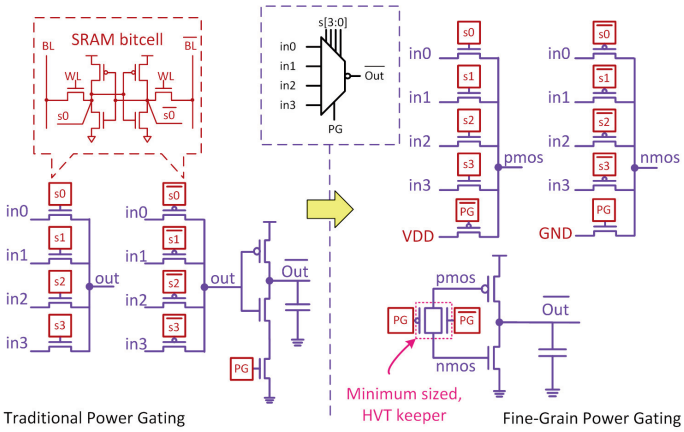


Figure 27.5.3: Multiplexer with traditional and fine-grain PG.

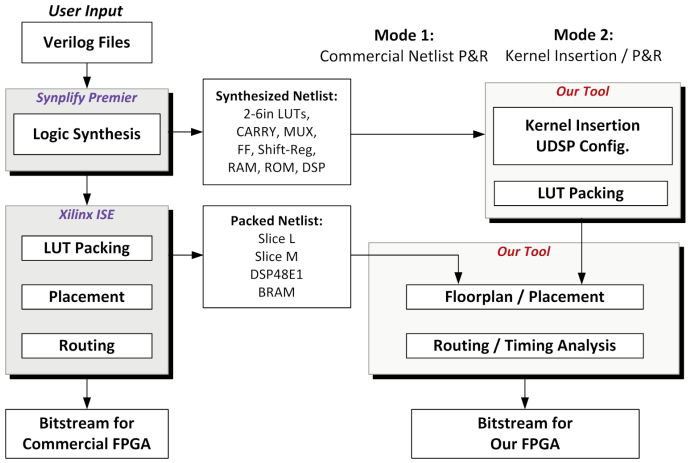


Figure 27.5.4: Automated place-and-route flow (2 modes).

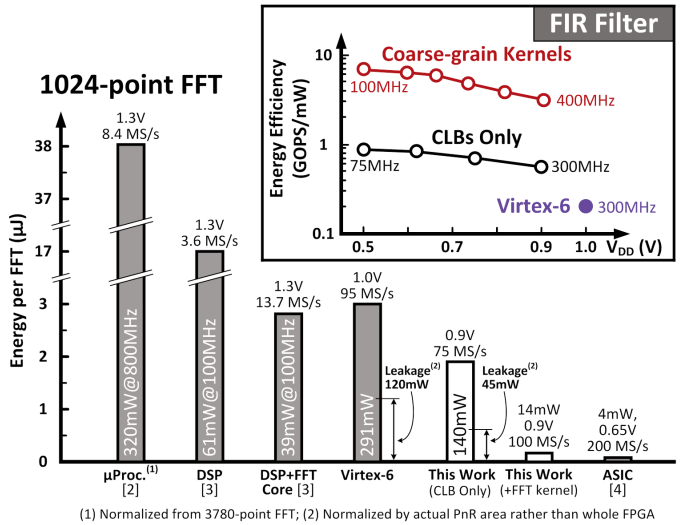


Figure 27.5.5: Comparisons of throughput and efficiency.

Fine-grain and Medium-grain CLBs Only

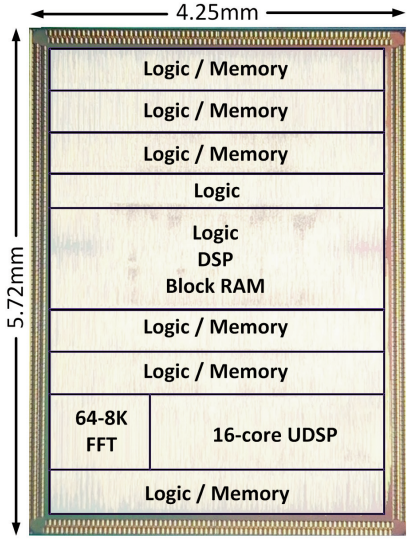
Resources	Utilization			Virtex-6*		This Work		ASIC			Energy Eff.
	L/M	DSP	BRAM	Freq.	Power	Freq.	Power	Freq.	Power	Tech.	gap vs. ASIC
	2760	42	16	(MHz)	(mW)	(MHz)	(mW)	(MHz)	(mW)	(nm)	(norm.)
256-point MIMO FFT [5]	916	4	4	180	270	140 80	135 48	64	25	180	**12×**
K-best + depth first Sphere Decoder	1255	0	0	200	250	140 70	129 58	16	3	90	**15×**
Software-Defined Radio Tx Front End	1594	38	0	39	287	35 23	213 117	60	50	65	**13×**
802.11a Rx Baseband [6]	1824	0	6	66	271	55 20	219 60	20	203	250	**15×**

*The Virtex-6 (6VLX75T) FPGA has 11640 Slice L/Ms, 288 Slice DSPs, and 156 BRAMs.

CLBs and Coarse-grain Kernels

Resources	Utilization					This Work		ASIC			Energy Eff.
	L/M	DSP	BRAM	FFT	UDSP	Freq.	Power	Freq.	Power	Tech.	gap vs. ASIC
	2760	42	16	1	16	(MHz)	(mW)	(MHz)	(mW)	(nm)	(norm.)
Software-Defined Radio Tx Front End	664	26	0	0	12	80 30	166 50	60	50	65	**4×**
802.11a Rx Baseband [6]	701	0	6	1	5	60 20	88 21	20	203	250	**5×**

Figure 27.5.6: Example designs and ASIC efficiency gap.



Technology	40nm 1P10M CMOS FO4 16.3ps (TT)
Core V_{DD}	0.5 to 0.9V
Frequency	≤ 400 MHz
Energy Efficiency	≤ 7 GOPS/mW
I/Os	350 bidirectional (171 bonded)
Core Size	3.8mm × 5.3mm
Transistor Count	143 million
L/M CLB	2176 Logic 576 Memory
DSP CLB	42 MAC/SIMD Processor
Block RAMs	16 36Kb BRAM
Config. Bits	9,279,168

Figure 27.5.7: Die micrograph and chip summary.