2.07 GHz Floating-Point Unit with Resonant-Clock Precharge Logic

Jerry C. Kao[1], Wei-Hsiang Ma[1], Suhwan Kim[2], Marios Papaefthymiou[1]
[1]University of Michigan, Ann Arbor, MI
[2]Seoul National University, Seoul, Korea
Email: jckao@umich.edu

***Abstract*—This paper presents an 8-cycle 64 FO4 single-precision fused-multiply-add floating-point unit (FPU) chip with fine-grain resonant clocking and dynamic-evaluation static-latch logic to achieve dynamic-logic levels performance with significant power reduction. Fabricated in a 90nm low-power RVT technology, the resonant FPU achieves clock speeds up to 2.07GHz. At its resonant frequency of 1.81GHz, it dissipates 334mW, yielding 31.5% lower power and 32% more GFLOPS/W over a conventionally-clocked version of the same FPU implemented on the same die.**

I. INTRODUCTION

High-performance low-power floating-point units (FPUs) are key building blocks of high performance processors. This paper describes the first-ever FPU designed with a dynamic logic called dynamic-evaluation static-latch (DESL) that is clocked by a fine-grain two-phase resonant clock distribution network, as shown in Fig. 1, to achieve performance levels typical of dynamic logic with significantly lower power dissipation. With an overall latency of 64 fanout-of-4 (FO4) inverter delays, this resonant FPU achieves the shortest overall latency among state-of-the-art reduced-latency FPUs [1, 2]. The chip has been fabricated in a 90nm low-power 9-metal process with a nominal supply of 1.2V. Operating at its resonant frequency of 1.81GHz with a 1.32V supply and a 0.41nH on-chip inductor, the 0.391mm^2 resonant FPU dissipates 334mW and achieves 10.82 GFLOPS/W. Compared to a conventionally-clocked version of the same FPU core implemented side-by-side on the same die, it yields a 31.5% decrease in power consumption and 32% improvement in GFLOPS/W. When forced to run off resonance, the resonant FPU reaches 4.14 GFLOPS at a clock frequency of 2.07GHz.

With more than 4,000 resonant-clocked gates, this resonant FPU chip is the largest and fastest resonant-clock design ever reported with resonance deployed across the entire clock network. In previous designs with resonance deployed across the entire clock network, reported resonant frequencies are at the 1GHz mark [3]. In the IBM CELL processor with 3.2GHz resonant frequency [4], resonant clocking is deployed only in the 830 buffers of its global clock distribution network, yielding limited overall power savings.

II. FPU CIRCUITRY

A. DESL Overview

The key for achieving high performance and low power in the FPU is the DESL logic family shown in Fig. 1. A

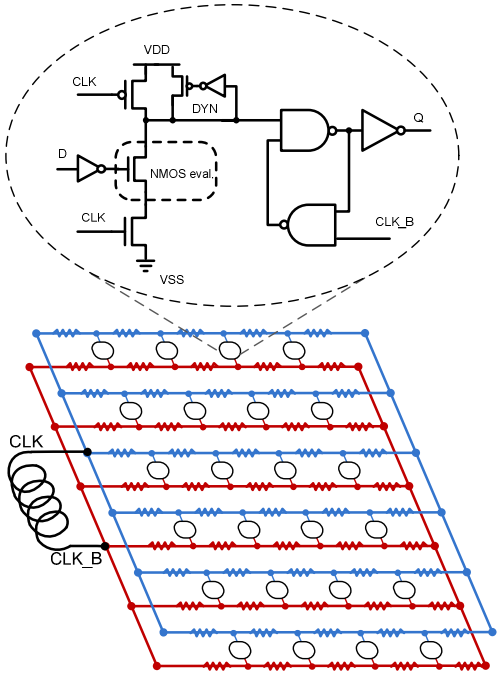

Fig. 1: Dynamic-evaluation static-latch logic with two-phase resonant clock

DESL buffer is shown in the cutout of Fig. 1. DESL is a variant of the SRAM domino read latch circuit [5], modified to work with two-phase resonant clocks. Each DESL gate consists of two stages: a dynamic evaluation stage, and a static transparent latch. The precharge and the clocked evaluation stack of the dynamic evaluation stage are clocked by clock-phase CLK. To ensure robust operation, the NMOS evaluation stack height between the clocked precharge node DYN and the footer NMOS device is limited to 3. The static transparent latch constructed using a back-to-back pair of NAND gates is clocked by clock-phase CLK_B. An inverter amplifies the output of the transparent latch to increase drive strength.

Operating waveforms from the simulation of a DESL buffer are given in Fig. 2. When CLK is low, node DYN is precharged high. When CLK rises, the gate evaluates through the clocked NMOS footer, while the pair of NAND gates clocked on CLK_B ensures that the state of node DYN is latched statically, and Q remains stable through subsequent precharge/evaluate cycles that result in the same output value.

Typically, in dynamic circuits, a dynamic node performs two transitions per cycle (precharge, evaluate), driving an

978-1-4244-8298-6/10/$26.00 ®2010 IEEE

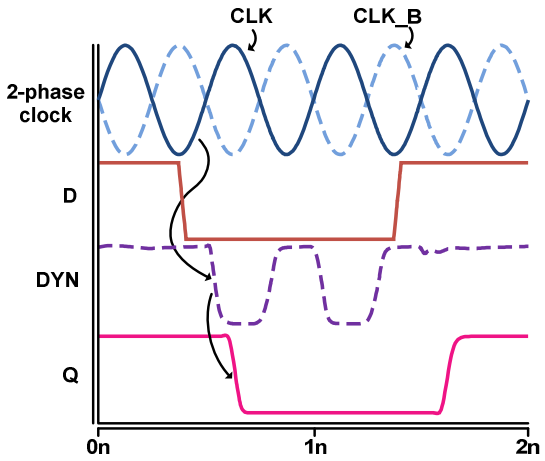

Fig. 2: Operating waveforms of DESL buffer

output buffer to amplify its state for logic evaluation in the next stage. Since the double transitions of the dynamic nodes are propagated to high-capacitance outputs, much of the power dissipated in such circuit topologies is wasted. DESL reduces power dissipation associated with double transitioning by inserting a static transparent latch at the output of every dynamic node DYN, converting dynamic signals to static outputs. The waveforms in Fig. 2 show that this method effectively isolates the high-capacitance output from the switching low-capacitance node DYN, thus reducing switching power by allowing output Q to switch only once when changing state.

Another design challenge typically associated with dynamic logic is the complexity and power requirements of its clock network. To achieve maximum performance, dynamic logic usually requires multiple clock phases (four or more) in one clock cycle with numerous constraints among them. In addition, since all pull-up networks are replaced by clocked devices, clock-related power is high. In our FPU, DESL gates are synchronized by 2-phase clocks which simplify clock generation and distribution. Furthermore, these 2-phase clocks are amenable to low-overhead resonant implementation, resulting in maximum power savings.

Compared to static CMOS, DESL has smaller input capacitance and faster operating speed, enabling 8 FO4 per cycle, which is significantly shorter compared to commercial microprocessors whose FO4 range from 11 [1] to 28 [8]. To overcome the evaluation-stack height and one-logic-function-per-phase limitations, circuit, logic, and architectural optimizations are used in the design of the FPU to keep overall latency low.

B. Resonant Clock Network Overview

Fig. 3 shows a simplified view of the 2-phase resonant-clock generation and distribution network used in the FPU. The clock network consists of 114 pairs of 278µm-long wires that are striped to cover a rectangular distribution area of height 1,034µm. Top-level distribution is performed using metal levels M9 and M8. Clock capacitance comprises the wire capacitance of the clock distribution grid and the clock-related input capacitance of all DESL gates. Together with a 1-turn 0.19x0.19mm^2 12µm-wide 3.4µm-thick M9 metal coil that provides 0.41nH of inductance, this clock capacitance forms an LC tank oscillator.

To replenish resistive losses, six H-bridge clock generator modules [9] of programmable size (up to 100µm max each) are evenly distributed across the FPU core. Each module comprises a pull-up/pull-down pair of devices driven by two pairs of complementary clock pulses (an, ap) and (bn, bp) with programmable duty cycle. Two pulse generators located symmetrically on two opposite sides of the FPU are used to generate these pulses. Each pulse generator drives three clock generator modules, and it is driven by an on-chip programmable ring oscillator using triple-wide equal-length wire to ensure minimal clock skew. Unlike previous designs [3, 4], this clock generator does not rely on a half-VDD supply and, thus, has no need for decoupling capacitance dedicated to half-VDD supply.

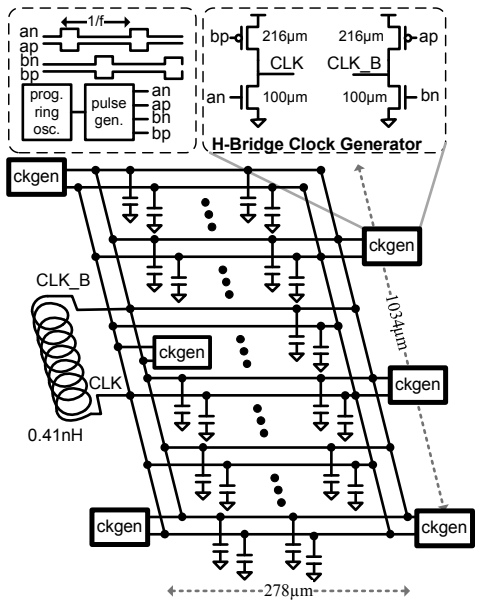

Fig. 3: Distributed resonant clock generation and clock distribution network

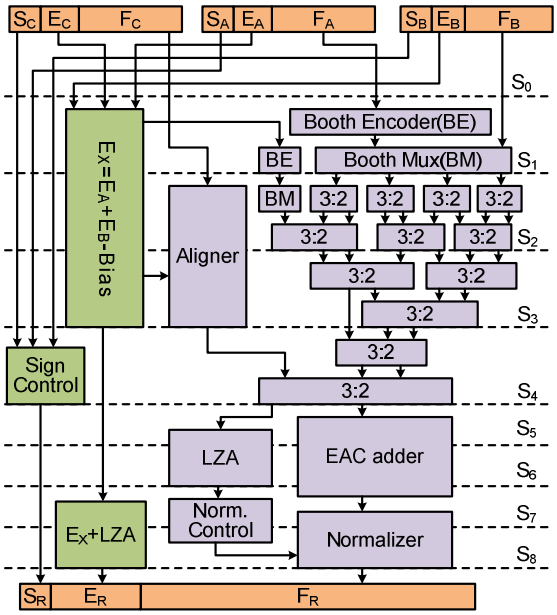

Fig. 4: FPU architecture

To allow for the maximum clock load to be resonated, there are no insertion buffers between the clock generator and the DESL gates. The resulting fine-grain distribution thus allows for the propagation of the resonant clock all the way to the leaves of the network.

III. FPU ARCHITECTURE

The FPU performs the single-precision fused-multiply-add operation A×B+C. To achieve high performance, it deviates from the IEEE standard and supports only the round-toward-zero rounding mode, just like [1, 2], thus speeding-up the fraction datapath by simplifying the sticky bit computation and removing the fraction rounding function. The combination of DESL's ability to achieve 8 FO4 per cycle with the elimination of some rounding mode support enables this FPU to achieve an overall latency of 64 FO4, compared to 100 FO4 for other state-of-the-art IEEE-754-compliant FPUs [6].

FPU architecture, shown in Fig. 4, is optimized to leverage the capabilities of DESL logic and achieve an overall latency of 8 cycles: The radix-4 Booth encoder and Booth Mux take 1 cycle; the 6 layers of 3-to-2 compressors that compress the partial products from the Booth Mux and the aligner take 3 cycles; the 75-bit end-around-carry (EAC) adder takes 2.5 cycles, and the normalizer takes 1.5 cycles. The critical paths are found in the shift amount decoder that drives the Mux select line in the final stage aligner, and in the computation of the indicator bit in the leading-zero-anticipatory logic [7].

With a fixed bit-width dataflow image for area efficiency, 10 bits are allocated for exponent and 52 bits for fraction. To fit within the allocated width, part of the aligner, the final 3-to-2 compressor, the EAC adder, and the normalizer are folded.

The FPU is designed using a semi-custom flow using 104 standard cells. Using 16 tracks/bit allows the FPU to be completely routed with M3 metal and below, reserving the higher level metals for clock and power distribution.

IV. MEASUREMENT RESULTS

The resonant FPU has been designed and fabricated in a 90nm RVT low-power process. For comparison purposes, a FPU with identical architecture and a conventional buffered clock distribution has been implemented on the same die, side-by-side with the resonant one. The

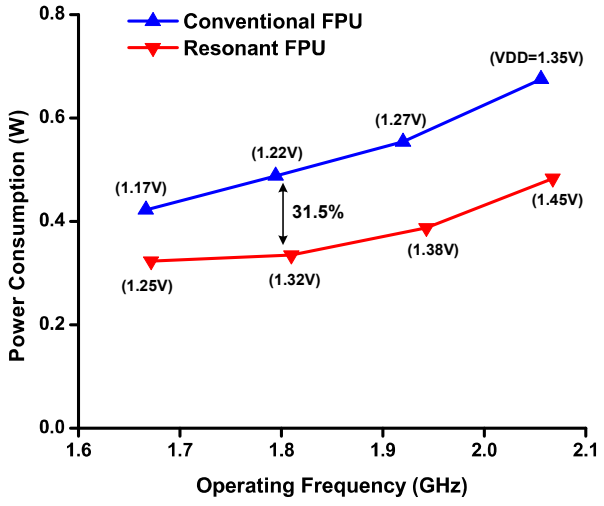

Fig. 5: Power consumption versus clock frequency

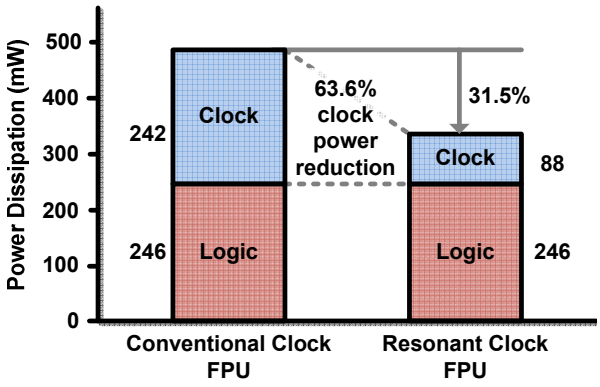

Fig. 6: Breakdown of power consumption at resonant frequency of 1.81GHz

conventional-clock FPU has been derived from the same netlist as the one used for the resonant FPU. The conventional clock distribution network has been obtained by replacing each resonant clock generator by 4 levels of clock buffers. The first level boosts the output of the programmable ring oscillator and drives 10 second-level buffers placed along the left edge of the FPU. Each second-level clock buffer drives a clock spine running over the 4-bit local clock bay between the exponent and the fraction unit. The third-level clock buffers located in the clock bay drive gates in that particular row, while the fourth-level inverters generate the local complementary clock phase. Sizing of the clock buffers has been performed through spice-level simulations with a target slew of 40ps.

Fig. 5 shows measured power consumption versus operating frequency for the resonant FPU core and its conventional counterpart, including BIST circuitry. Correct operation has been validated from 1.67GHz to 2.07GHz with supply ranging from 1.25V to 1.45V for the resonant FPU and from 1.17V to 1.35V for the conventional FPU. At its resonant frequency of 1.81GHz, the resonant FPU consumes 334mW with a 1.32V supply, yielding 31.5% lower power than the conventional FPU. For the same operating frequency, the resonant FPU requires about 100mV higher supply than its conventional counterpart, due to the narrower effective widths of the sinusoidal clocks. Power measurements from both cores are made with 50% switching activity on all LFSRs generating the three operands. In a typical application, the

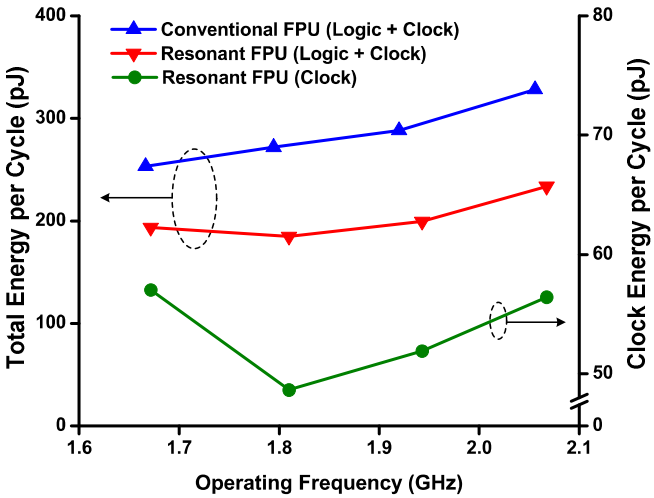

Fig. 7: Energy per cycle versus clock frequency

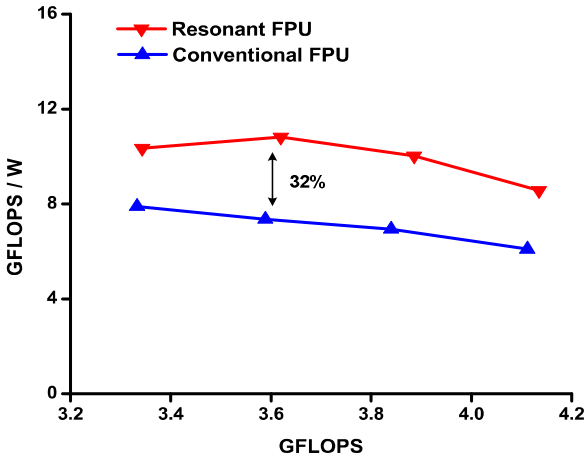

Fig. 8: Measured chip energy efficiency

input switching activity is expected to be in the 10-20% range, resulting in a relatively larger percentage of clock power and, therefore, even greater relative power savings from resonant clocking.

The power breakdown of the two FPUs when running at 1.81GHz is shown in Fig. 6. In the resonant FPU, logic and clock power are 246mW and 88mW, respectively, representing a 63.6% reduction in clock power compared to the conventional FPU.

Fig. 7 shows the energy breakdown of both FPUs versus operating frequency. The minimum clock energy consumption of 48.65pJ is reached at 1.81GHz, indicating the natural frequency of the resonant FPU. At this frequency, it consumes 32% less energy than the conventional FPU. Driving 32pF of clock load per phase, the resonant FPU recovers 55.5% of the CV^2f clock power, yielding a quality factor Q of approximately 2.25.

Resonant frequency is predicted with high accuracy. Inductance L is extracted using a commercial 2.5D Maxwell equation solver, and capacitance C of the clock network is extracted using a commercial RC extraction tool. The extracted parameter yields a resonant frequency of 1.97GHz (not including the damping factor), which is 6% away from the 1.81GHz minimum energy point in Fig. 7.

Fig. 8 gives measured energy efficiency in GFLOPS/W. The resonant FPU reaches its highest energy efficiency of 10.82 GFLOPS/W at its natural frequency, achieving 32% improvement over the conventional FPU.

Fig. 9 shows a die microphotograph with the resonant FPU on the left and the conventional FPU on the right. The resonant FPU with the clock generation occupies $0.283mm^2$. Including the inductor, it occupies $0.319mm^2$. This area is comparable to previous state of the art FPUs [1], as DESL logic uses a single clocked precharge device instead of a complementary PMOS pull-up network. Chip performance is summarized in the Table of Fig. 10.

V. CONCLUSION

A high-performance low-power FPU with fused-multiply-add has been designed using a dynamic-evaluation static-latch logic and fine-grain resonant clocking to achieve performance levels typical of dynamic logic with significant reduction in power dissipation. With an overall latency of 64 FO4, this resonant FPU achieves the shortest overall latency among state-of-the-art reduced latency FPUs [1, 2]. Fabricated in a 90nm process, the resonant FPU functions correctly from 1.67GHz to 2.07GHz and demonstrates that resonant clocking can yield significant reductions in clock power at multi-GHz clock frequencies. At its resonant frequency of 1.81GHz, the resonant FPU achieves 10.82GFLOPS/W, yielding a 63.6% reduction in clock power and a 31.5% improvement in GFLOPS/W compared to its conventional-clock counterpart implemented side-by-side on the same silicon die.

ACKNOWLEDGMENTS

This research was supported in part by NSF under Grant No. CCF-0916714.

REFERENCES

[1] H. Oh et al., "A Fully-Pipelined Single-Precision Floating Point Unit in the Synergistic Processor Element of a CELL Processor," Symposium VLSI Circuits, pp. 24-27, June 2005.
[2] S. Vangal et al., "A 5GHz Floating Point Multiply-Accumulator in 90nm Dual VT CMOS," ISSCC, pp. 334-497, February 2003.
[3] V. Sathe et al., "RF2: A 1GHz FIR Filter with Distributed Resonant Clock Generator," Symposium VLSI Circuits, pp. 44-45, Jun. 2007.
[4] S. Chan et al. , "A Resonant Global Clock Distribution for Cell Broadband-Engine Processor," ISSCC, pp. 512-632, February 2008.
[5] J. Pille et al., "Implementation of the Cell Broadband Engine in 65nm SOI Technology Featuring Dual Power Supply SRAM Array Supporting 6GHz at 1.3V," JSSC, vol. 43, pp. 163-171, January 2008.
[6] N. J. Rohrer et al., "PowerPC 970 in 130nm and 90nm technologies," ISSCC, pp.68-69, February 2004.
[7] M. S. Schmookler et al., "Leading Zero Anticipation and Detection—a Comparison of Methods," Symposium on Computer Arithmetic, pp. 7-12, June 2001.
[8] C. Webb, "IBM z10: The next-generation mainframe microprocessor," IEEE Micro, vol. 28, pp. 19-29, March-April 2008.
[9] D. Maksimovic et al., "Integrated Power Clock Generators for Low Energy Logic," Power Electronics Specialists Conference, pp. 61-67, June 1995.

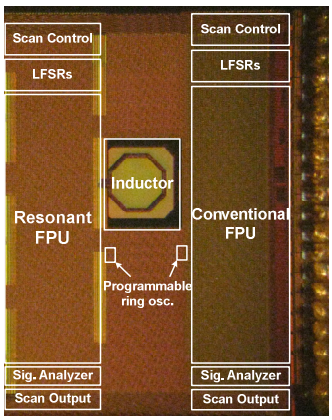

Fig. 9: Die microphotograph

Technology	90 nm 9M LP (RVT)
Nominal Voltage	1.2V
Transistor Count	~300K
FPU Core Area	0.283 mm^2
FPU Core + Inductor Area	0.319 mm^2
FPU Core + Inductor + BIST Area	0.360 mm^2
Overall Latency	64 FO4
Resonant Frequency	1.81 GHz
Energetics @ Resonance	
Supply Voltage (V)	1.32
Clock Network Efficiency	55.5%
Total / Logic / Clock Power Diss. (mW)	334 / 252 / 88
Total / Logic / Clock Energy (pJ)	184.8 / 139.3 / 48.7
GFLOPS/W	10.82
Input Switching Activity	0.5

Fig. 10: Performance summary table