Dynamic Voltage and Frequency Scaling for Low-power Multi-precision Reconfigurable Multiplier

Xiaoxiao Zhang[1,2], Amine Bermak[1], Farid Boussaid[2]
[1] ECE Department, Hong Kong University of Science and Technology
[2] School of EECE, University of Western Australia
Email: zhangxx,eebermak@ust.hk, boussaid@ee.uwa.edu.au

***Abstract*—In this paper, a 32×32-bit low power multi-precision multiplier is described, in which each building block can be either an independent smaller-precision multiplier or work in parallel to perform higher-precision operations. The proposed multi-precision multiplier enables voltage and frequency scaling for low power operation, while still maintaining full throughput. According to user's arbitrary throughput requirements, the highly dynamic voltage and frequency scaling circuits can autonomously configure the multiplier to operate with the lowest possible voltage and frequency to achieve the lowest power consumption. By carrying out optimizations at the algorithmic and architectural levels, we have completely removed silicon area and power overheads which is always associated with the reconfigurability features. The 32×32-bit low power multi-precision multiplier has been implemented in TSMC 0.18 μm technology. Compared with fixed-width multipliers, the proposed design features around 13.8% and 30% reduction in circuit area and power, respectively. Multi-precision processing featured in this paper accordingly enables voltage and frequency scaling resulting in up to 68% reduction in power consumption.**

I. Introduction

Recent works put significant effort to decrease multiplier's power consumption to enable its integration into battery-powered portable systems. However, in most full-custom as well as DSPs and FPGAs implementations, the multiplier is typically designed for a fixed maximum wordlength to suit the worst case scenario. However, the real effective worklengths of an application vary dramatically. The use of a non-proper wordlength may cause performance degradation or inefficient usage of the hardware resources. In addition, the minimization of the multiplier power budget requires the estimation of the optimal operating point including clock frequencies, supply voltage, and threshold voltage [1]. In most VLSI system designs, the supply voltage is also selected based on the worst case scenario. In order to achieve an optimal power/performance ratio, a variable precision datapath solution is needed to cater for various types of applications. Dynamic Voltage Scaling (DVS) can be used to match the circuit's real working load and further reduce the power consumption.

Several works have researched wordlength and supply voltage optimizations. Oliver A. and Hans-Jorg [2] proposed a reconfigurable multiplier which can be partitioned into several separate fully functional small multipliers, or used as a big multiplier. Wei and Yvon [3] proposed a variable precision multiplier which supports variable precisions ranging from 9 to 15 bits and dynamic voltage scaling. Vasily and Tomoyuki [4] proposed a multiplier architecture, which examines *k* MSBs of the operands to decide whether the entire multiplier is used and accordingly select between two different apply voltages. In the aforementioned multipliers, the reconfigurability function results in: (i) non-negligible silicon area and power overhead; (ii) performance and throughput reduction brought by the shut-down of parts of the circuit and/or use of reduced supply voltage, and (iii) restriction and great margins to the operating condition versatility of the multiplier.

Dynamic voltage scaling (DVS) saves energy by scaling down the voltage supply when the processor is not fully loaded. However, an effective method to find the lowest voltage to achieve the speed goal at run time is highly difficult. Above mentioned works found out the minimal voltage value by offline pre-simulations. However, different tasks have different speed and power requirements. Indeed, rarely can a user take an existing single set of power-speed modes and use it in every application. In this paper, we propose an automatic calibration circuit to solve this problem. Initially, the operating frequency of the multiplier is determined according to users' throughput requirement. Then the circuit will initially run at a proper voltage. When successive errors or correct results occur, the automatic calibration circuit would decide to raise or decrease the voltage, respectively.

The rest of the paper is organized as follows. Section II describes the architecture of our 32bit×32bit multi-precision multiplier. Section III illustrates the dynamic power and speed scaling management strategy. Section IV evaluate the performance of different multiplier topologies and totally remove the overhead resulting from the multi-precision reconfigurability. Section V gives the results and discussions. Finally, conclusions are drawn.

II. System Overview

The architecture of our multi-precision multiplier system is shown in Fig.1. The 32bit×32bit multiplier is composed of 9 sub-blocks that can work separately as 9 signed/unsigned 8bit×8bit multipliers, or concurrently to be configured to form

978-1-4244-5309-2/10/$26.00 ©2010 IEEE

3 separate 16bit×16bit multipliers, or one 32bit×32bit multiplier. When the full precision (32bit×32bit) is not exercised, the supply voltage and the clock frequency is scaled down according to the shorter latency restriction and actual workload to help save power. The resulting reduction in computation speed is overcomed by combining parallel architectures to maintain the throughput.

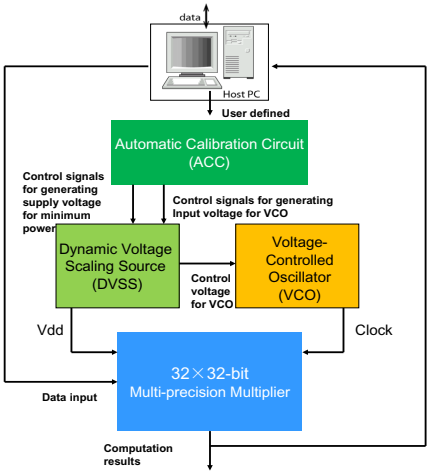

Fig. 1. Architecture of the proposed multiplier system.

III. Dynamic Power and Speed Scaling Management

In our architecture, voltage dithering is utilized to provide near-optimum dynamic voltage scaling with much less overhead [5]. Voltage dithering uses a few of power switches and let them toggle between a small number of voltage levels for different fractions of time to achieve an intermediate average voltage. Our test chip is designed to use two PMOS header switches as shown in Fig.2. By tuning the on/off time of the two complementary switches, the dithered voltage can be set to be equivalent to the required value.

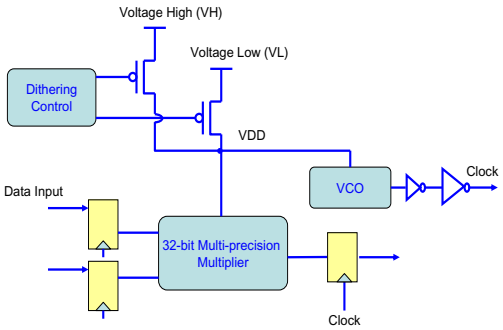

Fig. 2. Block diagram of voltage dithered multiplier and VCO using two header switches.

Voltage dithering was proposed as a low overhead implementation of DVS. The savings are only achievable if the voltage can change on the same time scale as the altering workload. Numerous methods for optimizing the headers are available, and most of them are designed to ensure that the circuit never exceeds a delay penalty more than 10%. Fig.3 shows the timing overhead of our dithering circuit. The dithered output can fully settle down within just 1 dithering cycle (0.01us), which is much shorter than a multiplier's execution cycle.

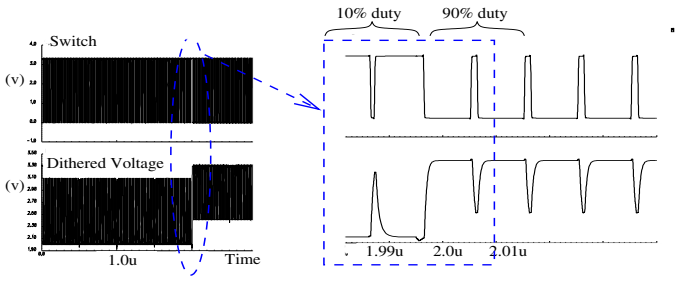

Fig. 3. Simulation result of the voltage dithering circuit's settling time.

The VCO consists of seven-stage current-starved inverters that can obtain the characteristics of high speed and low voltage operation. It can provide an oscillation range from 18M to 700MHz, see Fig.4. It can ensure the oscillation frequency when the input control voltage is as low as 0.8V. The VCO also features a fast response time with respect to the input control voltage. As shown in Fig.5, when the control voltage varies, the output clock frequency can settle down within only one cycle.

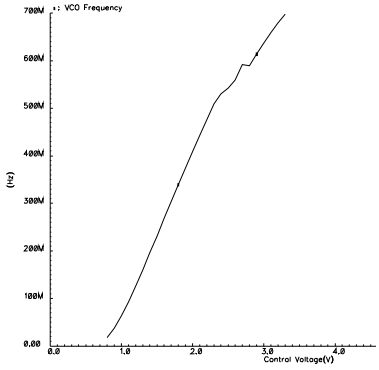

Fig. 4. Simulation result of the VCO's operation range and linearity

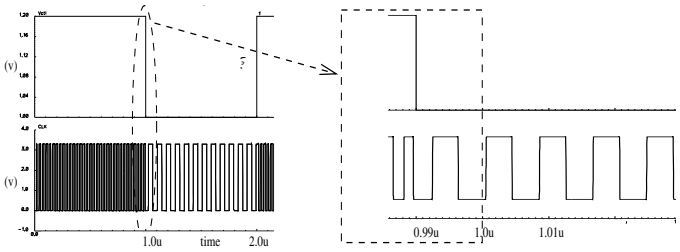

Fig. 5. Simulation result of the VCO's settling time.

IV. Low Power Multiplier Implementation

A. Choice of Booth Encoding Algorithm

Booth encoding is used to increase the multiplier's speed by reducing the number of partial products to be added. Radix-2

multiplier are easy for partial products generation but difficult for compression. While radix-8 is just the opposite. Given the scale of our design, radix-4 encoding is the best choice in view of a low power implementation. Simulation results show that radix-2 structure is 13% and 8% worse than the radix-4 structure in terms of area and power, respectively, refer to Table. I.

TABLE I
AREA AND POWER COMPARISON OF 9-BIT RADIX-2 AND RADIX-4 BOOTH MULTIPLIERS

Schemes	power(mW)	area(μm^2)
Booth radix-2(100MHz)	1.9607(1.08)	52932(1.13)
Booth radix-4(100MHz)	1.8094(1.00)	46983(1.00)
Booth radix-2(200MHz)	3.9240(1.07)	52942(1.10)
Booth radix-4(200MHz)	3.6726(1.00)	48039(1.00)

B. Choice of Partial Products Compression Topology

High-speed parallel multipliers are usually implemented either as array multipliers or as tree multipliers. As the interconnect power dissipation becomes more and more dominant, it becomes difficult to choose between regular/long latency structure and irregular/short latency structure. To verify, we have compared CSA array multiplier and Wallace tree multiplier. Simulation results reported in Table. II show that the Wallace tree scheme is advantageous in terms of computation speed, hardware complexity and also power consumption.

TABLE II
AREA AND POWER COMPARISON OF 9-BIT ARRAY AND WALLACE TREE STRUCTURE SCHEMES

Schemes	power(mW)	area(μm^2)
Array(100MHz)	1.9520(1.08)	48342(1.03)
Wallace tree(100MHz)	1.8094(1.00)	46983(1.00)
Array(200MHz)	3.9038(1.06)	50381(1.05)
Wallace tree(200MHz)	3.6726(1.00)	48039(1.00)

C. Choice of Partial Products Compression Array Type

When performing the compression, the left-to-right structure can provide a significant decrease in power dissipation, by reducing the amount of glitching in the left-hand side of the array and the whole compression array, see Fig.6. To reduce the dominant spurious compression array transitions, it is desirable to assign a signal that has high switching probability to circuits having short logic depth. Obviously, the MSBs are more often to be encoded as all 0s. This implies that we should first add the MSB's partial-products in a tree array for reduced circuit switching. From Table. III, without any extra operation, this scheme achieves 8% and 2% reduction in power and area, respectively.

The final design step combines the two's complement representation, radix-4 Booth encoding algorithm, left-to-right array scheme, and Wallace tree compression structure to build the 8bit×8bit multiplier building block. The next step would be an evaluation of the configuration or scaling overhead.

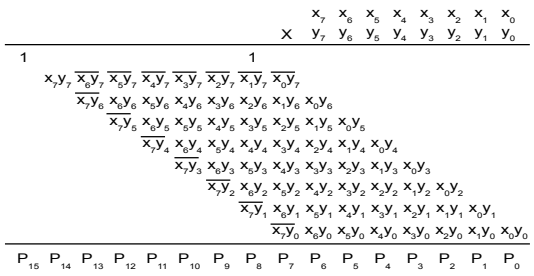

Fig. 6. Left-to-right compression array

TABLE III
AREA AND POWER COMPARISON OF 9-BIT RIGHT-TO-LEFT AND LEFT-TO-RIGHT COMPRESSION SCHEME MULTIPLIERS

Schemes	power(mW)	area(μm^2)
Right-to-left(100MHz)	1.8094(1.08)	46983(1.02)
Left-to-right(100MHz)	1.7549(1.00)	46000(1.00)
Right-to-left(200MHz)	3.6726(1.03)	48039(1.02)
Left-to-right(200MHz)	3.5601(1.00)	47186(1.00)

D. Reconfigurablity Overhead Evaluation

We define the 2n-bits wide multiplicand and multiplier as X and Y, respectively. X_H, X_L, Y_H, Y_L are their n-bit high significant and low significant half number of bits. The product of X and Y is expressed as:

$$P = (X_H Y_H)2^{2n} + (X_H Y_L + X_L Y_H)2^n + (X_L Y_L) \quad (1)$$

However, if we define [6]:

$$X' = X_H + X_L, Y' = Y_H + Y_L \quad (2)$$

equation 1 could be rewritten as:

$$P = (X_H Y_H)2^{2n} + (X'Y' - X_H Y_H - X_L Y_L)2^n + (X_L Y_L) \quad (3)$$

Comparing equations (1) and (3), one can note that one n-bit×n-bit multiplier and one 2n-bit adder can be removed and replaced by two n-bit adders and two (2n+2)-bit subtractors. For the 32-bit multiplier case, if we consider the complexity of a 32-bit adder and two 16-bit adders to be roughly the same, this enables us to use two 34-bit subtractor to replace a 16bit×16bit multiplier, the complexity reduction is obvious.

Simulation results show that the proposed architecture achieves reductions of 13.8 % in power and 30% in area as compared to the fixed-width multiplier design, in contrast to traditional multi-precision scheme which has 21.5% and 36.8% overhead in terms of area and power, respectively, referring to Table. IV. More power savings can be achieved when the input signal has a relatively smaller magnitude.

V. RESULTS AND DISCUSSION

Using our proposed voltage scaling approach, because the obtained voltage is always dithering, its averaged value is not equivalent to that of the static voltage of the same value. To better align the duty cycle to the static voltage value, we

TABLE IV
AREA AND POWER COMPARISON OF THE PROPOSED MULTI-PRECISION MULTIPLIERS AND TRADITIONAL FIXED-WIDTH MULTIPLIERS RUNNING AT 50MHZ

Schemes	power(mW)	area(μm^2)
32-bit Fixed-width Multiplier	12.2926 (1.00)	571092 (1.00)
32-bit 4 sub-block Multi-precision Multiplier	14.9301 (1.368)	781470 (1.215)
32-bit 3 sub-block Multi-precision Multiplier	7.9857 (0.862)	492079 (0.700)

simulated the multiplier's power consumption under various static voltage levels (from 1.5V to 3.3V) and various duty cycles (from 10% to 90%) respectively. Assuming that the dithered voltage toggles between 2.0V and 3.3V, from Fig.7, we can see that for example, if a dithered voltage which is equivalent to 2.3V static voltage is needed, the duty cycle should be set to 20%.

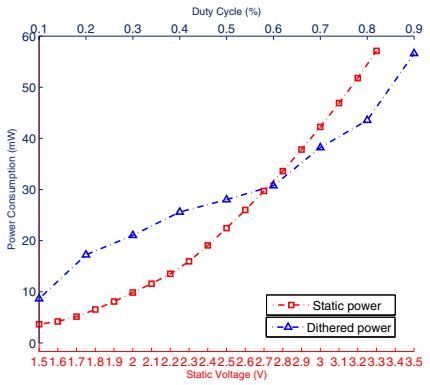

Fig. 7. The relationship of the duty cycle of dithering voltage and the equivalent static voltage

In prior work, flexibility and reconfigurability have been associated to increased silicon area and power consumption. In this paper, we have totally removed both the silicon area and power penalties. The layout of the multiplier is shown in Fig.8. The critical path of our multiplier is proportional to the real operating wordlength. With the shortened critical path, the supply voltage and clock frequency can be both reduced to save the power. Simulation results given in Table V indicate that a reduction of around 40-68% in power can be achieved under lower supply voltage and clock frequency. The implemented dynamic voltage/frequency adjusting scheme enables real-time processing at a given throughput, while saving more energy in comparison to static supply voltage scheduling.

VI. CONCLUSION

Variable latency functional units using adaptive operation precision can allow aggressive supply voltage scaling and clock frequency scaling for improved power efficiency with no performance penalties. In this paper, we proposed a multi-precision multiplier combining variable precision processing and scaled supply voltage and clock frequency to efficiently

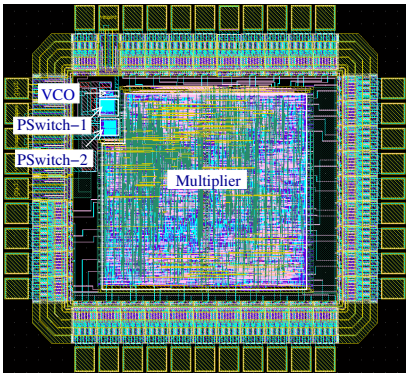

Fig. 8. Layout of the 32×32-bit multi-precision multiplier

TABLE V
AREA AND POWER COMPARISON OF THE PROPOSED MULTI-PRECISION MULTIPLIERS UNDER DIFFERENT WORKING MODES

Schemes	power(mW)
1 Signed 32-bit multiplier (at 1.98V,110MHz)	25.5972(1.00)
3 Parallel Signed/Unsigned 16-bit multiplier (at 1.80V,75MHz)	15.507(0.606)
9 Parallel Signed/Unsigned 9-bit multiplier (at 1.62V,50MHz)	8.232(0.323)

reduce circuit's power consumption. Various algorithms and topologies are explored to obtain high performance while no silicon area or power overheads. Reported results show that our variable precision multiplier enables a 30% reduction in silicon area and a 13.8% reduction in power dissipation compared to fixed-precision multipliers of the same size. When operating under different precisions, it can further bring around 40-68% power reduction. Our multi-precision multiplier is very attractive for various general purpose low-power applications.

REFERENCES

[1] Alice Wang, Anantha Chandrakasan, "180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology" *IEEE Journal of Solid-State Circuirs,*, Volume 40, No. 1, Page(s):310 - 319, January 2005

[2] Pfander, O.A.; Pfleiderer, H.-J., "EMMA - A suggestion for an embedded multi-precision multiplier array for FPGAs" *Field Programmable Logic and Applications, 2008*, Page(s):435 - 438, Sept. 2008

[3] Wei Ling; Savaria, Y., "Variable-precision multiplier for equalizer with adaptive modulation " *Circuits and Systems, 2004, MWSCAS '04*, Volume 1, Page(s):I-553 - 556, July 2004

[4] Vasily G.Moshnyaga, Tomoyuki Yamanaka, "Multiplier energy reduction by dynamic voltage variation" *Special Section on VLSI Design and CAD Algorithms*, Volume E88-A, Page(s):3548 - 3553, December 2005

[5] Calhoun, B.H.; Chandrakasan, A., "Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90nm CMOS " *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, Vol. 1, Page(s):300 - 599, -10 Feb. 2005

[6] Shaolei Quan; Qiang Qiang; Chin-Long Wey, "A novel reconfigurable architecture of low-power unsigned multiplier for digital signal processing" *Circuits and Systems, ISCAS 2005*, Vol. 4, Page(s):3327 - 3330, May 2005