A Survey and Comparative Analysis of Multiply-Accumulate (MAC) Block for Digital Signal Processing Application on ASIC and FPGA

P. Jebashini, R. Uma, P. Dhavachelvan and Hon Kah Wye

Department of Electrical and Electronics, Infrastructure University Kuala Lumpur, Kajang, Malaysia

Department of Computer Science, School of Engineering, Pondicherry University, Puducherry, India

ABSTRACT

In the field of semiconductor design industry which, in the contemporary times, has observed exceptional, explosive and exhilarating growth in the development of portable communication devices like mobile phones, IPADS and note books. These real time processing systems perform high computational operations, mainly in the form of butterfly and Multiply Accumulate (MAC). However, these systems are expected to consume high power and are characterized by high data throughput rate. Of the two, MAC is a major component used in portable applications and communication sectors like Wireless Code Division Multiple Access (WCDMA), base station receivers, Successive Interference Canceller (SIC), Orthogonal Frequency Division Multiplexing (OFDM) based wireless devices, channel estimators and carrier synchronizers. In general the MAC block resides in the critical path, which governs the complete power and speed of the system. The efficient utilization of MAC in terms of speed and power depends upon the type of architecture, logic technology style, the fundamental block and primitive cell realization. This study vividly presents the bird eye view on the hitherto work concerning the existing MAC unit in terms of its power performance factors, which helps the future researcher for opting suitable MAC block which can be used in Field Programmable Gate Array (FPGA) and Application Specific Integrated Circuit (ASIC) for signal processing applications. The comparative analysis is based on architecture/size, number of clock cycles, Partial Product Reduction Tree (PPRT), functional module, power saving method, logic technology, fabrication process and speed-voltage performance.

Key words: Multiply accumulate, recursive architecture, partial product reduction network, partial product generation, shared segmented architecture, parallel mac, multiplier schemes

INTRODUCTION

Basically, the architecture of MAC is classified as parallel, recursive and shared segmented structure. The recursive architecture (Matsui et al., 1994; Parameswar et al., 1996; Clark et al., 2001; Liao and Roberts, 2002; Chang et al., 2009) incorporates “divide and conquer” tactic, where the computation of large size data is segmented into smaller units. The multiply-accumulation is achieved using iterative calculation of smaller module through several clock cycles. The latency and throughput of the MAC depends on the number of multipliers and adders, which are recursively called for each cycle. In the first cycle, data is fetched from the internal memory, the second cycle involves multiplication process, during the third cycle summation takes place, while in the fourth cycle the function of A.B+Acc is performed and finally the last output is latched within the internal memory. This approach utilizes minimum hardware by using reusability...
of resources with increased latency. These types of MAC architectures are deployed in embedded Advanced RISC (Reduced-Instruction-Set-Computing) Machine (ARM) core. The parallel MAC architecture (Grossschadl and Kamendje, 2003; Chen et al., 2004; Gao and Chen, 2005; Kashfi et al., 2008; Wang et al., 2009) can be constructed by expanding the component of recursive MAC model. The complexity of the structure increases quadratically with the number of inputs. For instance, to implement a 32 bit MAC unit requires 32 bit multiplier, 64 bit adder and 128 bit accumulate unit. This type of architecture supports mode dependent logic to support full and half precision multiply or MAC operation. The number of clock cycles is reduced by three, when compared to recursive architecture by means of embedding the accumulator module within the partial product summation network. This MAC architecture is mostly used in Field Programmable Gate Array (FPGA) of Xilinx Corporation and as coprocessor for the LEON2 RISC processor. The shared segmented MAC architecture (Danysh and Tan, 2005; Xia et al., 2009; Parandeh-Afshar et al., 2010; Hoang et al., 2010; Quan et al., 2010) integrates the structures of split parallel unit and recursive characteristics which operate parallel with moderate resources supporting mode-dependent logic. This kind of architecture lacks from throughput limitations owing to compound PPRT. This architecture is capable of supporting full and half precision multiply or MAC operation. This MAC operates for SIMD (Single Instruction Multiple Data) with reduced clock cycle, when compared to recursive MAC and reduced hardware with respect to parallel MAC. This type of architecture is used in MIPS Technologies (Million Instructions per Second).

MULTIPLY-ACCUMULATE BLOCK

The conventional MAC comprises of fundamental components of partial product reduction tree, partial product generator and an accumulator is depicted in Fig. 1. The main function of PPG unit is to generate Partial Product (PP) of multiplier and multiplicand data which is fetched from the latch or internal memory. Its operation is executed on the full word length of each input received from the latch. The principle operation of PPRT is to compress and to add the PP with accumulated data. The general operation of the MAC is defined as:

\[
\text{Acc}(N) = \text{Acc}(N-1)+A_i \times B_i \quad (1)
\]

\[
\text{Acc}(N) = \sum_{i=0}^{2N-1} \text{Acc}_i + \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} a_i b_i 2^{i+j} \quad (2)
\]

POWER PERFORMANCE FACTORS OF MAC UNIT

The performance of MAC unit depends on the following parameters:

Fig. 1: General MAC block, N: No. of bits, PPG: Partial product generation, PPRT: Partial product reduction tree

Power: The three major components of power are: Transient or dynamic, short circuit and leakage power. The short circuit power is owing to the current conducting path between GND and VDD. The leakage power is due to the reverse bias diode and sub threshold leakage. These two powers are due to the logic style and technology, through which the MAC is realized. The transient or dynamic power is due to the total number of nodes and capacitors charged/discharged in a transition which is expressed as follows:

\[
P_{\text{dynamic}} = \alpha_{\text{transition}} C_{\text{pd}} f_{\text{clk}} V_{\text{DD}}^2 \quad (3)
\]

where, \( \alpha_{\text{transition}} \) is the total number of nodes active per transition (node activity factor), \( C_{\text{pd}} \) is the dynamic power capacitors, \( f_{\text{clk}} \) is the clock frequency (Input/output) and \( V_{\text{DD}} \) is the supply voltage. So, in the MAC unit the major portion of power is contributed due to transient power that mainly depends upon the node activity factor and dynamic capacitors.

Node activity factor: The node activity factor represents the total number of nodes active per transition, which is divided into two parts, namely, the function and parasitic part. The function (part) switching depends on the type of architecture used in the design, it depends on the logic function and block such as AND, OR, NAND, multiplier and adder, the signal statistics and the choice of logic style. The second part is due to glitches, which is caused due to signal skews (different input signal arrival time) and the signal statistics. The parasitic part in the MAC can be reduced using balanced delay path, gate sizing and reducing parasitic capacitances.
Clock frequency: It is one of the significant parameters persuading the functional power dissipation of MAC unit. The power factor is directly proportional to clock frequency of the MAC unit, therefore reducing clock frequency may proportionally reduce power. On the other hand, the MAC speed and throughput simultaneously reduced. In order to preserve the throughput for reduced clock frequency parallelism and pipelining architecture have to be considered. The MAC architectural power (block level) is characterized in terms of bits of the component (multipliers and adders) and their operating frequency which can be expressed as:

\[ P_{\text{functional-block}} = \Delta_1 \sum_{\text{input}} f_i + \Delta_2 \sum_{\text{input}} f_j = \Delta_1 f_{in} + \Delta_2 f_{out} \]  

(4)

where, \( f_{in} \) and \( f_{out} \) is the input and output frequency of MAC unit and \( \Delta_1, \Delta_2 \) are the empirical coefficients derived from gate-level simulation.

Architecture selection: Low power architecture design becomes imperative in MAC block. The architecture selection typically involves the organization of functional blocks in MAC and the number of pipes involved in the computation stage. In architectural level, low power can be achieved through clocking strategy, parallelism, pipelining and component organization. By deploying parallelism the throughput and performance of MAC unit can be improved without increasing the operating frequency.

In recursive architecture, the resource utilization is minimum, therefore, area reduction is achieved but the throughput of the system is considerably very low. Due to the smaller bit-size, component and reusability, the dynamic power is reduced. For parallel MAC the number of components will be doubled, when compared to recursive MAC architecture to achieve high speed and throughput at the expense of increased chip area twice that of recursive MAC. To reduce the area penalty of parallel MAC, split-pipelined MAC architecture is suitable, when trade-off results with less area overhead but more complexity in controller design due to multi-mode operation. When pipelined structures are implemented then the propagation delay is reduced to half when compared to the recursive MAC. On the other hand, consideration should be taken for pipelined MAC to maintain the throughput, when it is operating at lower voltages.

Logic technique: Till date, the majority of the circuit designs have been implemented using Complementary Metal Oxide Semiconductor (CMOS) logic style. It is very attractive due its reliable operation at low voltages. The CMOS logic style incorporates large PMOS (P-type Metal Oxide Semiconductor) in circuit realization. As a result, the propagation delay is higher, because of large node capacitances. The power dissipation is very high at high operating frequencies due to increase in the input loads. From the literature survey it has been observed that the MAC block has been implemented using static Complementary Metal Oxide Semiconductor (CMOS), Clocked-transmission Gate Adiabatic Logic (CTGAL), Low Voltage Swing Restoration Technique (LVST), Pass Transistor Logic (PTL), Complementary Pass Transistor Logic (CPL), mixed static CMOS-CPL and Swing Restored PTL (SRPL).

The PTL provides improved performance, when compared to CMOS logic due to less number of transistors, as a result, the overall parasitic capacitances is reduced. The dynamic power dissipation is very minimal due to faster switching time. One of the short falls associated with the PMOS logic is threshold drop variation. As a result, the noise margin of the circuit is reduced this in turn degrades the driving capability and leads to unreliable operation. The static power dissipation is very high in PTL due to threshold drop variation. The LVST logic detects the input voltages even, when it is less than 100 mV and performs reliable operation in low voltages. The logic structure realization is very complicated, with three level stages with true and complementary inputs, which add on the number of inverters in the design, in turn, it increases the static power dissipation of the circuit and poses moderate noise immunity.

The CPL logic required large number of transistors or gates to implement simple circuit. Due to large transistor the short circuit current is high and wiring overhead owing to the dual-rail signals. The SRPL circuit design is similar to LVST with three stages supporting true and complementary inputs, which differentiate two low inputs and regenerative operation is established through sense amplifier. In SRPL, when proper device scaling is not provided then discharging the output for 1-0 transition becomes bottleneck and consequently the output degrades. The MAC constructed using SRPL utilizes high transistor count and fair noise margin.

Functional blocks: The indispensable component necessary to implement the MAC blocks are adder and multiplier circuits. As mentioned earlier, the first stage of the MAC unit involves the generation of partial products, which can be established through multiplier circuits and the second stage is accumulation of PPG, which can be accomplished using adder circuits. From the literature survey, it have been observed that various multiplier and adder circuits like distributed arithmetic, parallel, serial-parallel, complementary (Booth encoding), Wallace using CSA, row-column bypass, modulo diminishing-1 and wave pipelining multipliers. The MAC architecture with complementary booth multiplier reduces the generation of number of partial products. Major bottle neck, when deploying booth multiplier is hardware complexity due to fundamental components of encoder and shift registers to produce PPG. Due to high interconnect the power dissipation is very high in this type of MAC architecture. While, deploying Wallance multiplier scheme the time complexity is reduced by N/2, when compared to array structure but dissipates high power due to irregular interconnects.
The accumulation and PP addition in MAC unit are performed using various adder schemes. The general structures deployed are Parallel Prefix Adder (PPA), Ripple Carry Adder (RCA), Carry Skip Adder (CSkA), Carry Propagate Adder (CPA), Carry Save Adder (CSA) and Carry Select Adder (CSelA). The adder structures deployed in the final stage of MAC have been implemented using Carry Save Adder (CSA), Carry Select Adder (CSelA), Carry Look Ahead Adder CLAA) and Parallel Prefix Adder (PPA). These adders however exhibits power dissipation and delay due to interconnect scheme and data distribution.

**Figure of merit:** The Figure of Merit (FOM) is expressed as:

$$\text{FOM} = \frac{f}{P \times V} \times 100 \quad (5)$$

where, $P$ is the total power of MAC unit for the given voltage ($V$) operating at a given frequency and this performance parameter should be minimum.

**Throughput:** The throughput of the MAC design is computed with respect to the clock frequency $f_{\text{clk}}$ and latency in various pipe stages. The throughput of the MAC can be expressed as

$$\text{MAC operations per second} = \frac{\text{Frequency}}{\text{Cycle each MAC operation}} \times \text{Parallel MAC units} \quad (6)$$

In above Eq. 6 the term parallel MAC denotes the number of MAC unit deployed to execute the instruction, generally it is expressed as Mega Operation Per Second (MOPS).

**MAC ARCHITECTURE**

The architecture selection for MAC unit generally depends upon the type of applications. For embedded microprocessor or microcontroller applications the memory usage is limited and the operand size is also small and therefore, recursive architecture is suitable, when power and area is important. This recursive MAC unit is deployed in image processing application such as Fast Fourier Transform (FFT) and digital filtering.

For high performance applications like notepads, laptops and desktops require large set of data computation therefore parallel architecture will be suitable. To perform multi-mode logic dependent operation, where the speed and power constraint is considered then shared segmented architecture is preferable, which is mainly used in embedded medical equipments and in communication systems, such as Orthogonal Frequency Division Multiplexing (OFDM) based wireless devices, subcarrier frequency domain operations, channel estimator and carrier synchronizer. The MAC architecture can be implemented in ASIC and FPGA. The implementation of MAC structure using FPGAs will have limited resources and fixed logic technology while in ASIC it is semi-custom or full custom so that optimization can be achieved from the architectural level to transistor level.

**Recursive MAC:** A recursive MAC proposed by Matsui et al. (1994) use novel Sense-Amplifying Flip-Flop (SA-FF) in combination with NMOS (N-type Metal Oxide Semiconductor) differential logic (Fig. 2a). The MAC unit has been embedded in 2-D DCT which operates at 200 MHz with 350 mW power dissipation for the supply voltage of 3.3 V. The macrocell was fabricated using 0.8 µm base rule CMOS technology. The SA-FF technique acts as a sense amplifier to regenerate low-swing differential inputs. The proposed MAC block reduces the propagation time and macrocell size using SA-FF technique. The MAC operation has been executed using the Distributed Arithmetic (DA) on a bit-by-bit data. The intermediate addition process has been implemented using conventional RCA and Carry Propagation Adder (CPA). The final accumulation and summation network has been designed through Carry Skip Adder (CSkA).

A new Swing Restored Pass Transistor Logic (SRPL) non-pipelined recursive unsigned MAC has been proposed by Parameswar et al. (1996) for multimedia applications and it was fabricated in double metal 0.4 µm CMOS technology, which operates at a maximum speed of 150 MHz consuming 34 mW with one cycle delay of 6.7 ns for bit size of 16 bit wide. The speed of the MAC unit has been improved using Gate sizing optimization approach, where the aspect ratio W/L (Width/Length) of NMOS transistors connected close to the output signal was reduced, when compared to NMOS transistors that were far away from the output. Another important speed optimization was achieved by using moderately scaled PMOS devices in the swing restoring network. The Partial Product (PP) was obtained using Booth encoding scheme and the same is added using CSAs.

A novel recursive MAC unit which offers single cycle throughput for 16 bit X 32 bit operation for audio processing application has been proposed by Clark et al. (2001). This MAC block has been embedded in a RISC microprocessor core, which has been fabricated in a 6 layer metal 0.18 µm CMOS process technology. The circuit dissipated 450 mV operating at 600 MHz for 1.3 V. The proposed core was the first application of the Intel Xscale micro architecture. The PP and summation has been implemented using, Booth encoding and four stages Wallace tree. The PPRT has been realized using 3-to-2 CSA and a 12 bit encoding scheme. Bit slice operations were performed by the MAC unit for the first clock cycle of 16 bit data of multiplier and multiplicand and the same is encoded through the Booth multiplier and compressed through Wallace tree and the resultant were accumulated. During the second clock cycle the remaining 12 bit was encoded and added with CLA. The functional block has been
Fig. 2(a-c): Different MAC Architectures, (a) Recursive MAC (redrawn) (Clark et al., 2001), (b) Parallel MAC (redrawn) (Kashfi et al., 2008) and (c) Split MAC (redrawn) (Parandeh-Afshar et al., 2010)

implemented using static CMOS logic. The final stage of addition operation has been constructed using, conditional-sum addition that incorporates Parallel Prefix Structure (PPS). The synchronization of clock circuit has been implemented with single rail domino logic.

A 32 bit recursive MAC has been proposed by Liao and Roberts (2002), incorporates a new mixed-length encoding scheme using; four-stage Wallace tree to improve performance and power. The MAC block supports SIMD and Multiply with Implicit Accumulate (MIA) for processing.
media stream. The fundamental component has been implemented using a mixture of static CMOS logic and CPL to enhance the power and speed aspect. This MAC block acts as a coprocessor to the Intel® XScale™ architecture and fabricated in a 6 layer metal 0.18 μm CMOS process which dissipates 450 mW at 1.3 V for 600 MHz. The PPG and PPRT were implemented using Booth multiplier, which utilizes 16 bit encoding scheme to generate the last sum and carry vectors in two cycles. The intermediate and final summation network has been implemented using, CSA and CLA with conditional-sum addition that incorporates Parallel Prefix Structure (PPS). The power-saving techniques employed in this MAC unit were clock gating and pulse-clocking methods.

A new variable-precision 8/16 bit MAC unit has been proposed by Chang et al. (2009) that incorporates a dynamic range detection circuit for power reduction. The static CMOS logic has been used to design the MAC unit, which was fabricated for ST Microelectronics 90 nm CMOS technology. The MAC operates at a frequency of 182 MHz with 2 clock cycle latency, which dissipates 26 mW for 1.1V. The functional modules were implemented using Baugh-Wooley array multiplier and CSA. The compression network utilizes rounding and truncation approach to minimize the switching activity of unused logic to reduce the power dissipation. The detection circuit identifies the active inputs and appropriately deactivates the unused components and bits. The output accuracy of this MAC was only 90% precision because of the truncation method which was the major deficiency in this approach.

Parallel MAC: A single cycle 32 bit parallel MAC for RISC processor has been proposed by Grossschadl and Kamendje (2003) for public key cryptography application (Fig. 2b). The proposed MAC executes the logic operations (32×32) multiplication (32×32+64) and (32×32+32+32) bit multiplication accumulation on unsigned data. The PPG and PPRT has been implemented using; modified radix-4 Booth multiplier and carry adder. The performance of the MAC unit has been improved in the generation of PP using radix-4 modified Booth encoding scheme which involves overlapping groups of three bits at single encoding thereby reducing the PPG time. The final adder has been realized using hybrid adder, which is constituted by CselA and RCA. The MAC has been developed in CMOS logic and prototyped in 0.6 μm CMOS standard cell library. The maximum operating speed of the MAC was 40 MHz dissipating 25 mW power for 1.3 V.

A fully pipelined parallel MAC, which enables (32×32+72) bit multiply-accumulate operation has been propounded by Chen et al. (2004). The proposed MAC unit has two pipe line stages and executes the MAC instruction in two clock cycles. In the first stage, Booth encoder accepts 32 bit inputs of multiplier and multiplicand and produces 17 PP's. These PP's has been compressed using Wallace tree network implemented by means of CSA. The second stage receives 3 inputs at a time, 2 inputs from stage one (compressed PP) and third input of size 72 bit from external register. By this method the proposed MAC unit eliminates the need of adders in stage 1 thereby reducing the area cost and power dissipation. The final stage adder has been constructed by CSA. The MAC block has been implemented in Static CMOS and fabricated with SMIC 0.18 μm, 1.8 V 1 P6 M process, which operates at a maximum of 300 MHz dissipating 30 mW/100 MHz.

A parallel 16 bit MAC has been propounded by Gao and Chen (2005), which was implemented for DSP processor. The power minimization has been achieved, through asynchronous interlocked pipeline method and performance was improved using; Complemented Partial Product Word Correction (CPPWC) algorithm and Three Dimensional Reduction Method (TDM) in the PPG and reduction network. To reduce the PP terms the CPPWC algorithm first invert the MSB of all PP and then PP having the highest weight was inverted, finally “1” was added to 2n-1 and n bits. The PPRT has been implemented using Vertical Compressor slices where the PPs are transferred in three directions. The interlock pipeline block consists of asynchronous-to-asynchronous and parallel synchronous timing path. These two clocks were used to acknowledge and validate the MAC operation. This MAC was implemented onto Verisilicon SMIC 0.18 Standard CMOS process, which dissipates 13.9 mW/100 MHz with supply voltage of 1.6V.

A high-speed parallel MAC has been propounded by Kashfi et al. (2008), which was developed using Low Voltage Swing (LVS) logic. The PPG and PPRT make use of the modified Booth encoding scheme using CSA. The full swing restoration has been achieved through Sense Amplify (SA) connected at the output side. This technique utilizes LVS operation in the internal nodes and establishes full swing at the output node using SA thereby power reduction has been achieved. The improved booth encoding scheme was accomplished in two stages specifically encoding and selection. The former logic examines three bit at a time and decides the operation as addition or subtraction and produces 8 rows of PP for 16×16 bit multiplication. The later logic selects the PP's and transfers to final adder constructed by conventional CLA. The proposed MAC has been fabricated in 65 nm CMOS Predictive Technology Model (PTM) which operates at 15 GHz dissipating 42 mW/GHz with supply voltage of 1.2 V.

A 16 bit parallel MAC has been proposed by Wang et al. (2009) using Clocked Transmission Gate Adiabatic Logic (CTGAL) with 3 clock cycle latency that has 7-stages of pipes. The structural complexity of the Booth encoder has been reduced using CTGAL. The power reduction in the MAC unit has been achieved by using, reversible logic by recycling the charge stored in the internal capacitances. The MAC deploys Modified radix Booth encoding scheme, where 3 bit are encoded at a time with 1 bit overlapping and the PP's are

www.ansinet.com
compressed using Wallace tree. The final adder stage has been constructed using Ladner-Fischer parallel prefix adder. The MAC unit has been prototyped in TSMC 0.25 µm CMOS technology 1 GHz dissipating 30 mW for the supply voltage less than 1V.

**Shared segmented MAC:** A multi-precision 64 bit vector MAC has been reported by Danysh and Tan (2005). The vector MAC can perform one 64×64, two 32×32, four 16×16, or eight 8×8 bit signed/unsigned multiply using essentially the same hardware as a scalar 64 bit MAC and with only a small increase in delay. The MAC incorporates Modified Booth and CPA as functional module using CMOS logic. The MAC was synthesized using SOI 90 nm operating at the frequency of 1.25 GHz dissipating 99.63 mW/MHz with the supply voltage of 1.2 V. The MAC has been synthesized using TSMC 90 nm CMOS standard cell technology and the functional blocks of radix-4 Booth encoding scheme and CPA was implemented using CMOS logic.

The MAC proposed by Xia *et al.* (2009) presents 4 bit pipelined split MAC architecture which has been developed for multimedia application supporting multi precision multi-mode operations (Fig. 2c). The speed of the MAC has been improved through a novel 1-A partial product compression circuit based on interleaved adders. This MAC supports 32 bit multiplication, 16 bit multiply or MAC operation and two parallel multiply operations at a frequency of 1.25 GHz dissipating power of 99.63 mW/MHz with the supply voltage of 1.2 V. The MAC has been synthesized using TSMC 90 nm CMOS standard cell technology and the functional blocks of radix-4 Booth encoding scheme and CPA was implemented using CMOS logic.

A high performance MAC unit used for digital filters has been implemented by Parandeh-Afshar *et al.* (2010), which is deployed as a coprocessor for the LEON2 RISC. The MAC has been constructed using 4 parallel MAC units supporting SIMD instructions. Each MAC unit has two 16 bit input registers, Booth multiply, PP, summation tree and CSA adder has been embedded to perform the MAC operation.

The MAC block has been implemented in CMOS logic which operates at 1.2 GHz dissipating 55 mW/MHz, for the supply voltage of 1.3 V and synthesized using CMOS 0.18 µm technology. A multi-mode high performance MAC unit which supports both true and complementary inputs with a special feature incorporates accumulation guard bits for saturation circuit has been developed by Hoang *et al.* (2010). The MAC unit supports 1×N/2 MAC or 2×N/2 multiplication operation. The fundamental component of MAC has been constructed using Baugh–Wooley multiplication algorithm and CSA.

The MAC unit incorporates CMOS logic which operates at a frequency of 2 GHz dissipating 54.9 mW/MHz for the supply voltage of 1.3 V. The MAC block has been synthesized using 65 nm CMOS cell library. A 32 bit Vector MAC capable of supporting multi-mode operations like 32×32, 32×16, 2-16×16 and 4-8×8 MAC functions has been proposed by Quan *et al.* (2010). The functional component has been constructed using Booth multiply and Wallace tree compression which incorporates CMOS logic. The existing MAC structure is shown in Fig. 2.

**LOGIC STYLE OR TECHNIQUE**

**Sense amplifying flip-flop logic technique:** The SA-FF (Matsui *et al.*, 1994), is a Self-Timing (ST) Dual Rail Logic (DRL) used in the design of data-path operation to achieve higher speed with dual output. The circuit detects the input voltage lower than 100 mV and differentially boost the low-swing voltage to full rail-to-rail voltage by using Sense Amplifier (SA). The proposed SAFF scheme is shown in Fig. 3a, SA is compounded into latch, which acts as a synchronization element to the global clock.

The circuit receives two inputs true and complementary low swing voltage which will be boosted and the same is latched in Flip-Flop (FF). The main shortfall associated with this technique is that the timing signal used to activate the SA is mostly from delay lines using self-timing which should be properly adjusted and optimized, otherwise there will be a risk of fatal malfunction or the system hangs in metastability state or in the worst case racing signal hazards may encounter. The behaviour of the circuit is similar to conventional FF, when the clock have the transition from 1-0 it acts as master FF and stores the differential output to latch and during 0-1 the slave device is activated which passes the output Q and Q̅.

**Swing restored pass transistor logic:** The SRPL (Parameswar *et al.*, 1996) is a class of Pass Transistor Logic (PTL), which consists of two parts namely a complimentary output PTL network and a swing restoring circuit. The former is constructed with NMOS devices and the latter is constructed with cross coupled CMOS inverters. The inputs in SRPL technique are connected to drain and gate of PTL network. The gate input of each transistors are connected with control inputs (variables) and the drain terminal of each transistors in the logic network are connected with pass inputs (variables). The general structure of SRPL in Parameswar *et al.* (1996) is shown in Fig. 3b. This type of arrangement nullifies the shortfalls associated with CPL and Double Pass-transistor Logic (DPL). In CPL the Boolean function is evaluated using CPL network and full swing output is achieved using static CMOS inverter. But the problem incurred with this configuration is leakage current through static inverters. The DPL uses both PMOS and NMOS to rectify the voltage swing problem with high-area and high-power dissipation. In SRPL when proper device scaling is not provided then discharging the output from 1-0 transition becomes bottleneck and consequently the output degrades. The MAC constructed using SRPL utilizes high transistor count and fair noise margin.
Fig. 3(a-f): Various logic style for MAC design, (a) SA-FF technique (redrawn) (Matsui et al., 1994), (b) SRPL technique (redrawn) (Parameswar et al., 1996), (c) CMOS technique (Clark et al., 2001; Grossschadl and Kamendje, 2003; Chen et al., 2004; Chang et al., 2009; Xia et al., 2009; Parandeh-Afshar et al., 2010; Hoang et al., 2010; Danysh and Tan, 2005; Quan et al., 2010), (d) CPL (Liao and Roberts, 2002), (e) LVS (Kashfi et al., 2008) and (f) CTGAL redrawn (Wang et al., 2009)
Static complementary metal oxide semiconductor: The logic style reported by Clark et al. (2001), Grossschadl and Kamendje (2003), Chen et al. (2004), Chang et al. (2009), Xia et al. (2009), Parandeh-Afshar et al. (2010), Hoang et al. (2010), Danysh and Tan (2005) and Quan et al. (2010) is the most common design technique, where each logic network will have pull up and pull down devices, which are controlled by input signals as shown in Fig. 3c. The advantage of CMOS over PTL (single polarity circuit) exhibits very small power dissipation and leakage. Nevertheless, the contribution of power dissipation in CMOS logic is determined by the operating frequency. When the input load is high the power consumption and leakage is very high due to large PMOS devices. The propagation delay is also high in CMOS, when compared to PTL due to large input and parasitic capacitances.

Complementary pass transistor logic: The CPL in Liao and Roberts (2002) shown in Fig. 3d consist of NMOS pass transistor network for logic realization, which receives true and complementary inputs. The level restoration is achieved using static CMOS inverters, which produce true and complementary outputs. The PMOS latch connected below the static inverters decreases the static power dissipation. The CPL offers the highest speed at the expense of increased transistor count. The other shortfall of this logic utilizes significant number of nodes and wiring complexity is high.

Low-Voltage Swing (LVS) logic: The LVS logic in Kashfi et al. (2008) consist of 4 stages namely complementary input data generation, Dual rail N-diffusion Connected Network (DCN), a level restoring buffer using sense amplifier (SA) and the last stage have Cross Coupled Domino Logic (CDL) for producing high gain. The LVS logic circuit realization is shown in Fig. 3e. The complementary inputs are generated using static inverters in the first stage and these inputs are fed to second stage DCN which is constructed using NMOS transistor to evaluate the logic function under two control signals namely CTL and Reset. This block is based on dual rail differential amplifier which produces true and complementary output D and D_bar, which will be below 100 mV. These differential outputs are boosted using SA in stage 3. The non-zero offset level of SA zero output induces glitches which are compensated through CDL at the last stage. This unit not only reduces the glitches but also increase the gain of the output signal.

Clocked transmission gate adiabatic logic: Clocked Transmission Gate Adiabatic Logic used in Wang et al. (2009) is shown in Fig. 3f. The circuit utilizes two power clocks φ and ¨φ. The logic function is evaluated using cross coupled transmission gates. The functional computation of CTGAL is alienated into two phases specifically sampling and valuing-holding-recovery phase. The input signals are sampled via the NMOS transistors N1 and N2 which are triggered with the input clock ő. During the second phase, when the input and power clock are at logic zero either node x or y will be floated high-voltage of VDD-V_w. When the clock signal rises gradually the output is evaluated by turning ON the NMOS transistor N3. At the same time the floating node will bootstrap the voltage level high through the charged internal node capacitances and hold this phase. When the PMOS transistors P1 and P2 are turned OFF, the output is completely recovered by φ. The output of the circuit is full swing due to energy recovery mechanism.

FUNDAMENTAL COMPONENTS OF MAC

Multiplier scheme for MAC unit: Multiplication is an essential arithmetic operation of MAC block which have huge area, extended latency and consume substantial power. As avowed in Yeh et al. (2003), multiplier components utilizes 46% chip area in most of MAC and FFT module. Therefore, low-power multiplier design has been an significant part in low-power VLSI system design. The speed and power consumption of multiplier depends on the algorithm for PPG, PPA and logic technology used to design the multiplier cell. The power saving in multiplier cells are accomplished with two approaches. They are; High-level algorithms to decrease the switching activity and the regularity of structure (block and interconnect complexity).

Multiplication schemes and algorithms implemented in FPGAs and ASICs diverge in the means of intermediate PPG and PPA. The multiplication algorithms are classified based on the space (area) complexity, interconnect and time complexity. In terms of PPG and PPA the multipliers are categorized into distributed arithmetic, parallel, serial-parallel, complementary (Booth encoding), Wallace using CSA, row-column bypass, modulo diminishing -1 and wave pipelining multipliers. The general multiplier schemes deployed in MAC unit are distributed arithmetic, parallel, complementary (Booth encoding), Wallace using CSA and wave pipelining methods. The multiplier scheme in Matsui et al. (1994) utilizes distributed arithmetic multiplication (White, 1989; Berkeman et al., 2000), which is a bit-serial operation, which can be computed by the direct dot product of vectors. The general merits of these schemes are reduced power consumption and area. This type of multiplier is apparently very slow due to its bit-serial characteristics.

The multiplier scheme in Parameswar et al. (1996), Clark et al. (2001), Liao and Roberts (2002), Grossschadl and Kamendje (2003), Chen et al. (2004), Kashfi et al. (2008), Wang et al. (2009), Xia et al. (2009), Parandeh-Afshar et al. (2010), Danysh and Tan (2005) and Quan et al. (2010) incorporates Booth radix encoding with Wallace tree or Modified Booth encoding using CSA which reduces the time complexity by N/2 when compared to array structure but
The MAC block in Fadavi-Ardekani (1993), Tang et al. (2001) utilizes two’s complementary Booth which moderates the generation of partial product by an aspect of N. But the complexity of this circuit is very high due to the presence of shifter and encoder. The power factor is also affected due to irregular interconnects. The multiplier design in Chang et al. (2009) and Hoang et al. (2010) uses Baugh and Wooley array multiplier (Pezaris, 1971; Baugh and Wooley, 1973) reduces the number of partial products by a factor of N.

Adder scheme for MAC unit: The accumulation and PP addition in MAC unit are performed using various adder schemes. The MAC design by Matsui et al. (1994) and Grossschadl and Kamendje (2003) deploys RCA (Chang and Hsiao, 1998) is the simplest adder exhibiting low power consumption as well as compact layout giving reduced chip area. The delay of RCA linearly increases as the number of input (n) increases, therefore, the speed-power factor of the RCA is limited when n grows higher.

The adder scheme by Parameswar et al. (1996), Liao and Roberts (2002), Chang et al. (2009), Grossschadl and Kamendje (2003), Chen et al. (2004), Kashfi et al. (2008), Parandeh-Afshar et al. (2010) and Hoang et al. (2010) utilizes CSA structure (Hsiao et al., 1998; Prasad and Parhi, 2001) which reduces the propagation effects in RCA with increased full adders and number of stages, thereby interconnect complexity also proportionally increases.

The MAC unit in Matsui et al. (1994) uses carry skip adder (Kantabutra, 1993; Gayles et al., 1996) that delivers a worthy compromise with respect to area with simple and regular layout with increased delay if the grouping of bits and carry generation units are not proper.

The adder scheme by Clark et al. (2001), Liao and Roberts (2002), Kashfi et al. (2008) and Danysh and Tan (2005) utilizes CLA or CPA (Doran, 1988; Kao et al., 2006; Xia et al., 2009) that reduces delay compared to RCA and CSA with increased interconnect complexity due to separate carry and propagate unit. The adder unit in Grossschadl and Kamendje (2003) includes the CsIA structure (Morinaka et al., 1995; Kim and Kim, 2001) that minimizes the pre-computation of carry bits in the chain. The short fall in this scheme is fan-out limitation owing to the large number of multiplexer unit.

In the worst case, a carry signal is used to select n/2 multiplexers in an n bit adder. The adder design in Clark et al. (2001), Liao and Roberts (2002) and Wang et al. (2009) utilizes parallel prefix structure (Brent and Kung, 1982; Kogge and Stone, 1973; Sklansky, 1960) which offers the maximum speed with high fan-out and interconnect complexity. The circuit complexity of multipliers and adders used in the design of MAC is shown in Table 1 and 2. The performance characteristics of various MAC structure is listed in Table 3.
### Table 3: Performance of various MAC block with architecture specification and functional parameter specification

<table>
<thead>
<tr>
<th>Proposed</th>
<th>Size/architecture</th>
<th>Stages/clock cycles</th>
<th>Functional module</th>
<th>Power saving method</th>
<th>Logic technology</th>
<th>Fabrication</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matsui et al. (1994)</td>
<td>16 bit recursive</td>
<td>6, 3-cycle operation for 16 bit data</td>
<td>1. Distributed arithmetic 2. RCA, CPA, CSA 3. Booth encoding</td>
<td>Self-timing</td>
<td>SA-FP</td>
<td>0.8 µm base rule CMOS technology</td>
<td>350 mW at 200 MHz for 3.3 V</td>
</tr>
<tr>
<td>Parameswar et al. (1996)</td>
<td>16 bit recursive</td>
<td>8, 4 cycle</td>
<td>1. Booth encoding 2. CSA</td>
<td>Gate sizing</td>
<td>SRPL</td>
<td>0.4 µm CMOS technology</td>
<td>34 mW at 150 MHz for 3.3 V</td>
</tr>
<tr>
<td>Clark et al. (2001)</td>
<td>32-b recursive</td>
<td>8, 2-cycle</td>
<td>1. Booth encoding (16 bit encoding scheme) 2. CLA 3. PPS (conditional sum)</td>
<td>Clock gating</td>
<td>Static CMOS</td>
<td>six-layer metal 0.18 µm CMOS process</td>
<td>450 mW at 600 MHz for 1.3 V</td>
</tr>
<tr>
<td>Liao and Roberts (2002)</td>
<td>32 bit recursive</td>
<td>7, 2-cycle operation for 16 bit data</td>
<td>1. Booth multiplier (16 bit encoding scheme) 2. CLA (PPS and CSA)</td>
<td>Clock gating and pulse clocking</td>
<td>Mixture of static CMOS and CPL</td>
<td>six layer metal 0.18 um CMOS process</td>
<td>450 mW at 600 MHz for 1.3 V</td>
</tr>
<tr>
<td>Chang et al. (2009)</td>
<td>8/16 bit recursive</td>
<td>6, 2-cycle</td>
<td>1. Baugh-wooley array multiplier 2. CSA</td>
<td>Data truncation</td>
<td>Static CMOS</td>
<td>90 nm CMOS technology</td>
<td>26 mW at 182 MHz for 1.1 V</td>
</tr>
<tr>
<td>Grossschadl and Kamendje (2003)</td>
<td>32 bit parallel</td>
<td>7, 1 cycle</td>
<td>1. Modified radix-4 booth encoder 2. CSA, hybrid CselA and RCA</td>
<td>Bits overlapping</td>
<td>Static CMOS</td>
<td>0.6 µm CMOS technology</td>
<td>25 mW at 40 MHz for 1.3 V</td>
</tr>
<tr>
<td>Chen et al. (2004)</td>
<td>32 bit parallel</td>
<td>2, 2 cycle</td>
<td>1. Booth Encoding 2. Wallace tree and CSA 3. PPAWC</td>
<td>Resource minimization</td>
<td>Static CMOS</td>
<td>SMIC 0.18 µm, 1P6M Verisilicon SMIC 0.18</td>
<td>30 mW/100 MHZ for 1.8 V</td>
</tr>
<tr>
<td>Gao and Chen (2005)</td>
<td>16 bit parallel</td>
<td>7, 2cycle</td>
<td>1. Modified booth 2. CSA and CLA</td>
<td>Interlocked pipeline</td>
<td>Static CMOS</td>
<td>Standard CMOS process 65 nm CMOS Predictive Technology Model (PTM)</td>
<td>14 mW/100 MHZ for 1.6 V</td>
</tr>
<tr>
<td>Kashfi et al. (2008)</td>
<td>16 bit parallel</td>
<td>5, 2 cycle</td>
<td>1. Modified booth 2. Wallace, lader fischer</td>
<td>Reverse recovery</td>
<td>LVS Low-Voltage Swing (LVS) logic</td>
<td>TSMC 0.25 µm CMOS technology</td>
<td>42 mW/GHz for 1.2 V</td>
</tr>
<tr>
<td>Wang et al. (2009)</td>
<td>16 bit parallel</td>
<td>7, 3 cycle</td>
<td>1. Modified Booth 2. Wallace, lader fischer</td>
<td></td>
<td></td>
<td>SOI 90 nm</td>
<td>60 mW at 1GHz for supply voltage less than 1.093 GHz dissipating 72 mW/3MHz for the supply voltage of 1.2 V</td>
</tr>
<tr>
<td>Danysh and Tan (2005)</td>
<td>64×64, two 32×32, four 16×16, or eight 8×8 bit signed/unsigned multiply split</td>
<td>4, 6, 8 cycle</td>
<td>1. Modified booth 2. CPA</td>
<td>Clock gating</td>
<td>CMOS logic</td>
<td>TSMC 90 nm CMOS</td>
<td>99.63 µW/MHZ 1.25 GHz for the supply voltage of 1.2 V</td>
</tr>
<tr>
<td>Xia et al. (2009)</td>
<td>32 bit split MAC</td>
<td>4, 3 cycle</td>
<td>1. Modified booth 2. CPA</td>
<td>Interleaved compression</td>
<td>CMOS logic</td>
<td>TSMC 90 nm CMOS</td>
<td>99.63 µW/MHZ 1.25 GHz for the supply voltage of 1.2 V</td>
</tr>
<tr>
<td>Parandeh-Afshar et al. (2010)</td>
<td>32 bit split MAC</td>
<td>5, 2 cycle</td>
<td>1. Modified booth 2. CPA</td>
<td>Resource minimization</td>
<td>CMOS logic</td>
<td>TSMC 0.18 µm CMOS</td>
<td>55 mW/MHz, 1.2 GHz for the supply voltage of 1.3V</td>
</tr>
<tr>
<td>Quan et al. (2010)</td>
<td>32×32, 32×16, 2×6×16 and 4×8×8</td>
<td>6 cycle</td>
<td>1. Booth multiply 2. Wallace tree</td>
<td>Bits overlapping</td>
<td>CMOS logic</td>
<td>0.13 um SMIC</td>
<td>1 GHz dissipating 62.3 mW/MHz with the supply voltage of 1.5V</td>
</tr>
</tbody>
</table>
ANALYSIS AND COMMENTS

The bird eyes view of existing MAC architecture has been presented. The importance of MAC unit in contemporary development has been focused. The three MAC architectures recursive, parallel and shared segmented (Spit MAC) characteristics and its uniqueness has been discussed. The performance of MAC depends upon the architecture selection, functional components (multiplier and adder) and technology used to implement the design. From the literature survey it is evident that the power consumption is high when the MAC is designed using Low-Voltage Swing (LVS) logic when it is operated for lower frequencies. It is obvious that the delay (latency) is less when the MAC unit is used to implement the design. Moreover the power aspect is less for Low-Voltage Swing (LVS) logic when it is operated for lower frequencies. It is evident that the power consumption is high when the MAC is designed using CMOS logic especially at higher frequencies.

CONCLUSION

The electronics and semiconductor industry occupy a pivotal place in contemporary society by serving human needs and subserving human happiness. Therefore, with the passage of time and growth in demand for electronic devices, the industry has witnessed fascinating growth. The development has been more pronounced in the arena of portable communication devices like Mobile phones, IPADS and notebooks. The existing high performance systems used in the communication system utilizes Multiply Accumulate block which consumes high power and are characterized by higher data throughput rate. But in the power starved society, there is an imperative need to discover systems that can reduce power and increase the speed of MAC system. In this context, several MAC with its power reduction techniques have been focused in this literature work. Moreover this survey would present the bird eyes view of the existing MAC block which would help the future researchers to explore more in the field of low power and high speed multiply-accumulate unit.

REFERENCES


