A 200-MHz CMOS Pipelined Multiplier–Accumulator Using a Quasi-Domino Dynamic Full-Adder Cell Design

Fang Lu, *Member, IEEE*, and Henry Samueli, *Member, IEEE*

***Abstract*— A bit-level pipelined 12 × 12-b two's-complement multiplier with a 27-b accumulator has been designed and fabricated in a 1.0-μm p-well CMOS technology. A new "quasi N-P domino logic" structure has been adopted to increase the throughput rate, and special pipeline structures were used in the accumulator to reduce the total latency. The chip complexity is approximately 10 000 transistors and the die area is 2.5 × 3.7 mm^2. The measured maximum clock rate is 200 MHz (i.e., 200 million multiply–accumulate operations per second), and the power–speed ratio is 6.5 mW/MHz. An unique output buffer design was also adopted to achieve 200-MHz off-chip communication while maintaining full CMOS logic levels.**

I. Introduction

In the majority of digital signal processing (DSP) applications, the critical processing paths in the algorithms usually involve many multiplications and/or accumulations. Furthermore, in real-time signal processing systems, a high throughput rate is desired and input-to-output latency is often not an important consideration. By incorporating pipelining in the design, the throughput rate of a circuit can be improved significantly at the cost of some additional latency. Therefore, the primary motivation for this work was to investigate various pipelined multiplier/accumulator architectures and circuit design techniques which are suitable for implementing high-throughput (> 100 MHz) signal processing algorithms.

Fig. 1 shows a simplified block diagram of the multiplier–accumulator (MAC) chip described in this paper. It consists of a 12 × 12-b two's-complement multiplier and a 27-b accumulator that accommodates overflows up to 3 b. The major features of the MAC chip are as follows:

1. By using newly designed "quasi N-P domino" full-adder pipeline stages, the throughput (clock rate) is up to 200 MHz with a single 5-V supply.
2. The accumulator reset mechanism does not interrupt the pipelined data flow, which is a desirable feature for real-time signal processing.
3. If desired, the 27-b output word can be rounded to any word length between 15 and 21 b. The rounding operation is performed outside the accumulation loop and thus the roundoff error does not accumulate with time.
4. A robust clock buffer design has been developed to generate complementary-phase on-chip clock signals. The skew between the complementary signals is minimized and the sensitivity to IC process variations is also reduced.
5. A high-speed, low-noise output buffer has been designed to achieve 200-MHz board-level communications while maintaining a single master-clock operation and full CMOS logic levels.

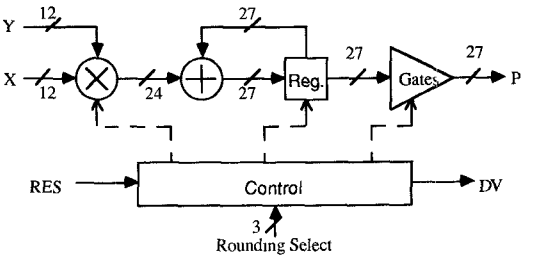


Fig. 1. Simplified block diagram of the MAC chip.

The high-level structure of the MAC chip is introduced in Section II. Section III compares several types of pipelined full-adder cell designs based on SPICE simulations. The detailed design of the pipelined multiplier and the pipelined accumulator is described in Section IV. The buffer design for high-speed timing and interfacing is discussed in Section V, and the MAC chip layout and various measurement results are presented in Section VI.

II. Chip Overview

Fig. 2 shows the functional block diagram of the MAC chip. Since all the data flow downwards through the chip, the pipelined stages can be derived by partitioning the building blocks internally in a horizontal direction. To meet the parallel input/output requirements, some triangular-shaped shift-register arrays are inserted into the circuit for preskewing and deskewing purposes.

The 12 × 12-b carry-save adder (CSA) array receives the X data (multiplicands) directly from input latches, and accepts the Y data (multipliers) through the coefficient preskewing shift register block. The 3-b rounding select word R either

Manuscript received July 22, 1991; revised August 28, 1992. This work was supported in part by the University of California MICRO Program and TRW, Inc.

F. Lu is with Baseband Technologies, Inc., Los Angeles, CA 90024.

H. Samueli is with the Integrated Circuits and Systems Laboratory, Electrical Engineering Department, University of California, Los Angeles, CA 90024.

IEEE Log Number 9204734.

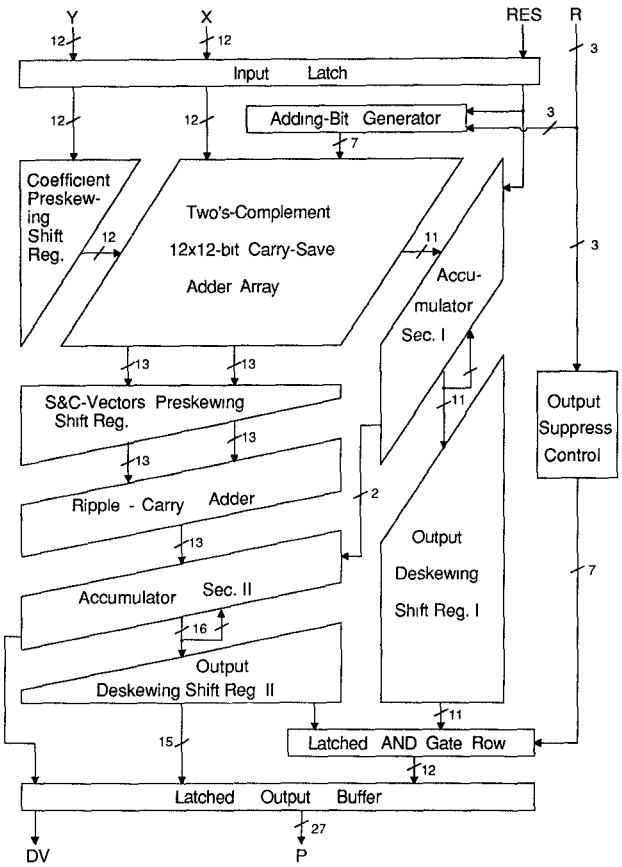


Fig. 2. Functional block diagram of the pipelined multiplier–accumulator.

keeps the 27-b output word length, or sets the word length from 15 to 21 b. If the rounding option is selected, the adding-bit generator sends a logic ONE to the top of the adder array at the location which is one bit lower than the LSB of the expected word length. The "rounding-bit" addition is triggered by an accumulator reset signal RES only once at the end of each accumulation cycle. Therefore, an additional adder array at the MAC output for two's-complement rounding is not required.

The CSA array sends an 11-b lower section product to accumulator section I and sends two 13-b vectors, namely partial sum (S) and partial carry (C), to the S&C-vector preskewing shift register. Then, the two skewed vectors are merged by the pipelined ripple-carry adder into a 13-b upper section product, which is subsequently delivered to accumulator section II along with a 3-b sign extension.

By using novel pipeline-stage designs, the fully pipelined 27-b two-section accumulator has a latency of only 11 clock cycles. Depending on the value of output rounding select word R, some LSB's of the output are set to ZERO by the output suppress control and the latched AND-gate row before being sent to the output buffers. This feature has the desirable property of eliminating the power dissipation in the disabled output buffers since no logic transitions will occur.

When the accumulator reset signal RES is latched into the MAC chip along with the last set of X and Y input words of an accumulation cycle, it will also be pipelined through the accumulator, and will be sent out of the chip as a data valid (DV) signal which is synchronized with the final result of the multiply-and-accumulate process; thus the user does not need to keep track of the number of cycles of pipeline latency. In addition, the reset mechanism has been designed so as not to disrupt the input sequence or the pipelined data flow. That is, the first input vector of a new accumulation cycle can be loaded into the chip right after the last input vector of the previous accumulation cycle. This feature is especially useful for finite-impulse-response (FIR) digital filtering (or real-time correlating) applications that require an uninterrupted sequence of accumulate-and-dump operations.

III. Pipelined Full-Adder Cell Design

A. Comparison of Full-Adder Cell Structures

Four kinds of logic structures for pipelined full-adder (FA) cells were considered: 1) conventional CMOS logic, 2) pseudo-NMOS logic, 3) standard N-P domino logic, and 4) a newly designed quasi N-P domino logic. All four FA cells were designed to contribute a *half* clock cycle of pipeline latency to accomplish the 1-b addition, and the logic tree structures were basically the same for all four kinds of FA cells so that the comparisons would be unbiased.

Depicted in Fig. 3(a) is a conventional CMOS pipelined FA cell which is based on the FA cell presented in [1]. The additional C^2MOS dynamic latches at the outputs perform the logic inversion, buffering, and pipelining functions. The rolls of Clk and $\overline{Clk}$ are interchanged from one pipeline stage to another. A transmission-gate adder [1] was considered but not adopted because its input C-switches would corrupt the data stored at the dynamic-latch output nodes of the previous pipeline stage.

The pseudo-NMOS pipelined FA cell is shown in Fig. 3(b). It is a modified version of a conventional NMOS FA cell [2]. The C^2MOS latches are also added for the same functions mentioned above. The P-logic trees (P-blocks) in its conventional CMOS counterpart have been replaced with pull-up P-MOSFET's which are always weakly on.

Fig. 3(c) shows the standard N-P domino FA cell, which is a variation of a NORA-CMOS serial full adder [3]. It uses three-transistor P-CMOS and N-CMOS latches instead of C^2MOS latches. The reason for this simplification is that, during precharge phases, the output nodes of the P-block (node 1) and the N-block (node 2) are set to V_{ss} and V_{dd}, respectively. Thus, the N-device in the P-CMOS gate and the P-device in the N-CMOS gate are turned off automatically. Hence, when combined with the actions of the pipelining switches, the precharge phase isolates the FA cell from its succeeding cells which are in the evaluation phase, and no erroneous data will be transferred.

The newly designed quasi N-P domino FA cell is shown in Fig. 3(d). During the precharge phase, nodes 1 and 2 are not always fully driven to V_{ss} and V_{dd}, respectively. This results in a higher speed but also requires the use of C^2MOS latches. When at least two of the A_i, C_i, and S_i inputs are high during the precharge phase, node 1 will be discharged to V_{ss}, and the carry output C_o will be ONE in the following evaluation phase, which is the expected result. On the other hand, if none

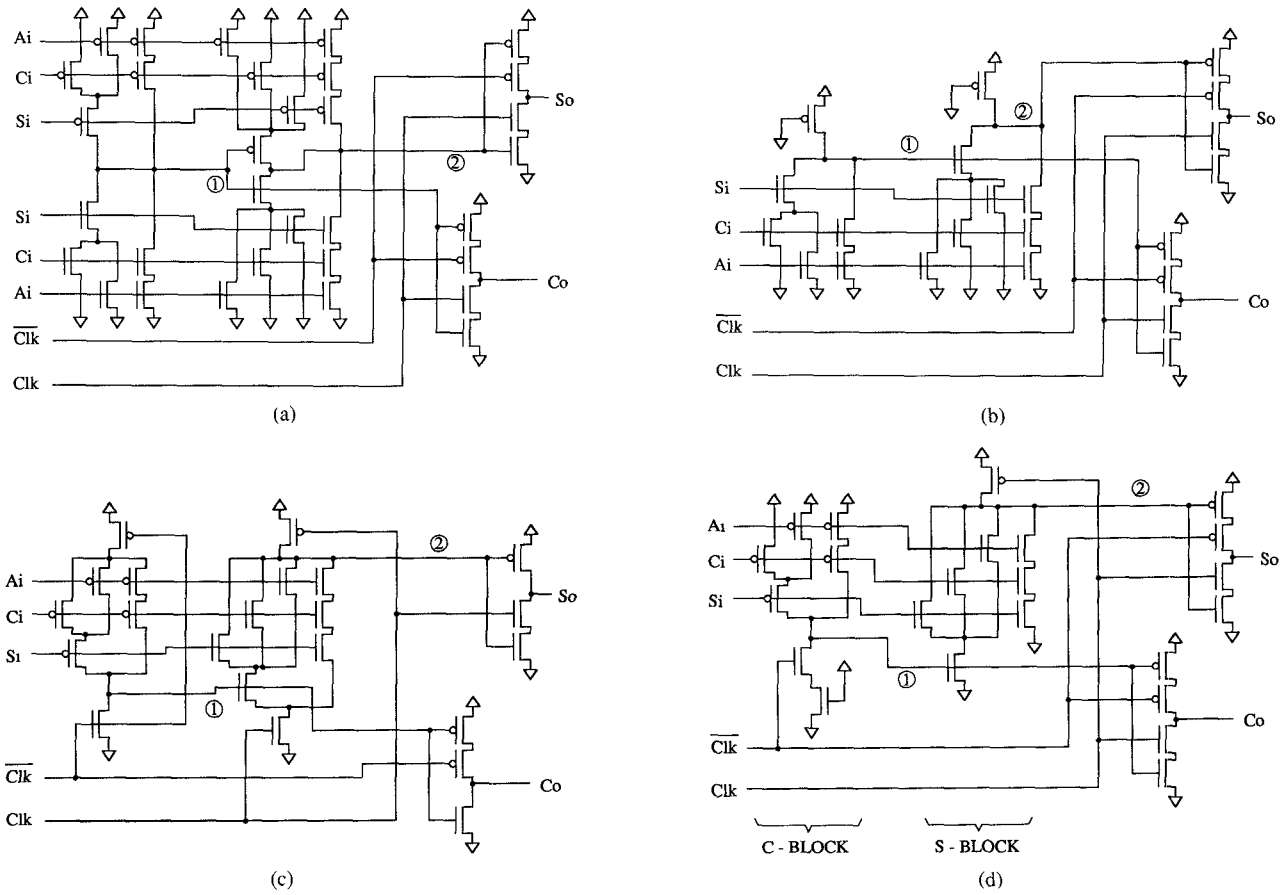


Fig. 3. Pipelined FA cells. (a) Conventional CMOS logic. (b) Pseudo-NMOS logic. (c) Standard N-P domino logic. (d) Quasi N-P domino logic.

or only one of the three inputs is high, node 1 will *not* be fully pulled down to V_{ss} because a resistive dc path exists between V_{dd} and V_{ss}. Then, during the following evaluation phase, the discharging switch is turned off and the P-block pulls node 1 back up to V_{dd}. Thus, the FA cell sends out a logic ZERO at the C_o output, which is still correct. The N-block that computes the sum (S_o) output has a behavior similar to that of P-block except that the directions of precharging and evaluation on node 2 are both opposite to those on node 1.

A resistive N-MOSFET is added in series with the N-switch under the P-block to purposely reduce the pull-down voltage swing during precharge phases without enlarging the N-switch gate length, thus the capacitive load on the clock signals will not increase.

TABLE I
COMPARISON OF PIPELINED FULL-ADDER IMPLEMENTATIONS

Full Adder Type	Maximum Clock Rate	Ave. Power Dissipation at Max. f_{clk}	Power-Speed Ratio	Transistor Count
Conventional CMOS	139 MHz	0.753 mW	5.42 uW/MHz	32
Pseudo-NMOS	167 MHz	1.305 mW	7.82 uW/MHz	22
Standard N-P Domino	179 MHz	0.918 mW	5.13 uW/MHz	22
Quasi N-P Domino	200 MHz	1.135 mW	5.68 uW/MHz	23

B. Simulation Results

The above four kinds of pipelined FA cells were simulated with HSPICE using the *slow* device models of the TRW 1.0-μm p-well CMOS technology. The numerical results are presented in Table I and discussed as follows.

1) Conventional CMOS logic is the slowest circuit, but the power efficiency is very good since it sinks current from the power line only when logic transitions occur. It is also apparent that CMOS logic occupies the largest chip area because complementary logic trees are required.

2) Pseudo-NMOS logic can be faster than CMOS logic by using large P-MOSFET pull-ups. However, the size of the P device has to be limited so that the low-state logic-tree output voltage is below the threshold of the N-MOSFET's, even under the combination of slow (weak)-N and fast (strong)-P models. Thus, the small pull-up current supplied by the weak P-MOSFET's results in long rise times, which limits the clock speed in heavily pipelined systems. Furthermore, when the N-blocks are on, the pull-up P-MOSFET's induce a large standby power consumption, which is the most serious drawback of pseudo-NMOS logic.

3) Standard N-P domino logic is fast because it has the same small input capacitive loads as pseudo-NMOS logic, and the precharge phases eliminate the rise-time problem. The power–speed ratio is also satisfactory since there is no dc power consumption.

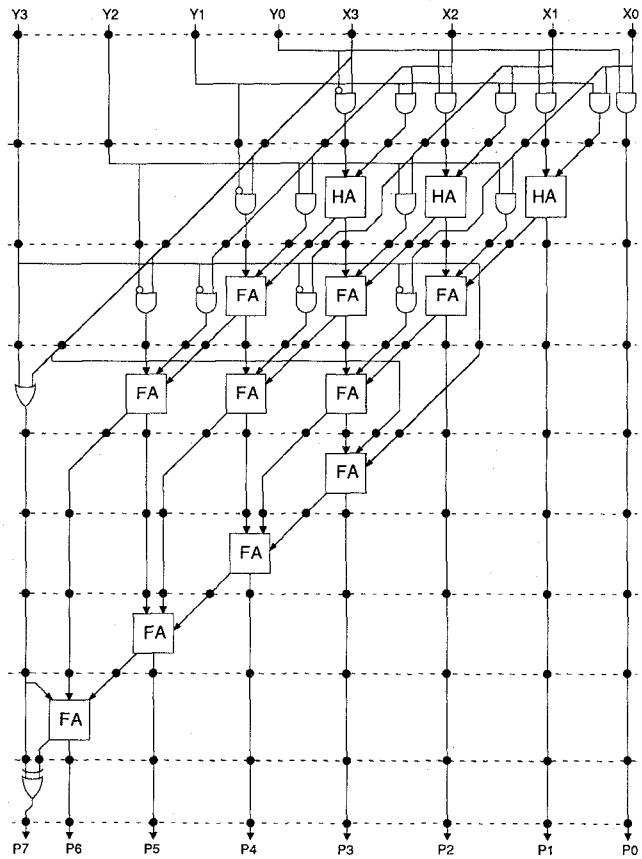


Fig. 4. A 4 × 4-b example of fully pipelined two's complement muliplier.

4) In terms of clock speeds, quasi N-P domino logic is 44% faster than CMOS logic, and is 12% faster than standard N-P domino logic. The primary reasons for the speed improvement are that the evaluation voltage swings are reduced due to the partial precharges, and the removal of the evaluation switches results in smaller evaluation-path resistance. The cost for this higher speed is a slightly larger power consumption.

Since the power efficiencies of the two domino FA cells are comparable, and since the power consumed by the MAC chip is in fact dissipated primarily by the chip-wide clocking system and the output buffers, the newly designed quasi N-P domino FA cell was adopted in the MAC chip design to achieve the highest throughput.

IV. Major Building Block Design

A. Pipelined Multiplier Structure

The Baugh–Wooley algorithm [4] is adopted to achieve two's-complement multiplication. Since the modified Booth algorithm [5] and the Wallace tree [6] cannot improve the throughput of a fully pipelined system, a linear CSA array is adopted in the multiplier to obtain high regularity in the layout and simplicity of pipeline-stage partitioning.

Fig. 4 shows a 4 × 4-b example of a pipelined two's-complement multiplier modified from the Baugh–Wooley algorithm. The pipeline stages are derived by inserting latches (the black dots) to evenly partition the critical delay path. Each stage contributes a *half* clock cycle to the system latency. Extra

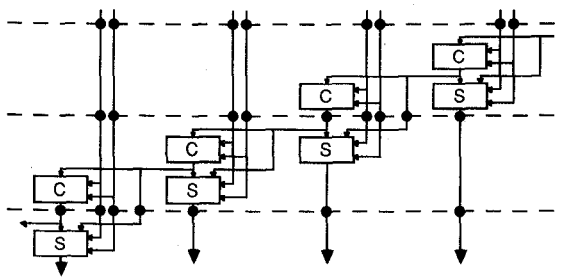


Fig. 5. Modified pipelined ripple-carry adder used in the multiplier.

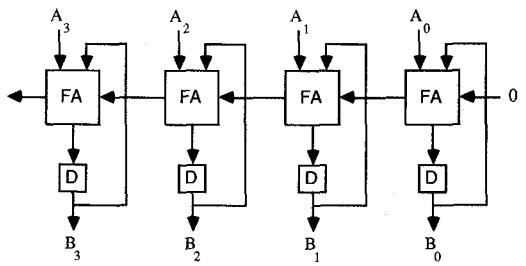


Fig. 6. Basic accumulator structure.

latches are added to construct preskewing/deskewing shift registers. The multiplier structure is *semi-systolic* [7] since the Y data are broadcast rather than pipelined horizontally along the rows of the adder array.

Another point to note is that the partial-product AND gate outputs of each pipeline stage are sent to the full adders located in the succeeding stage rather than the full adders in the same stage. This configuration removes the AND gate delay from the full-adder cell critical path delay and a higher system throughput can be achieved [8].

The circuit in Fig. 4 is basically a reduced version of the 12 × 12-b multiplier in the MAC chip except for the ripple-carry adder (RCA) portion. Recall that the FA cell in Fig. 3(d) has two logic blocks (denoted the C-block and S-block) to compute S_O and C_O. As the C-block output is generated prior to the S-block output, a special pipelined RCA structure has been designed as shown in Fig. 5. Some inverters (not shown) are added outside the critical path to adjust the signal polarities whenever necessary.

Since each FA cell in every other bit position is separated so that the C-block and the S-block belong to different pipeline stages, a 4-b carry propagation can be accomplished within only *one* clock cycle, rather than the two cycles required in Fig. 4. Thus, the latency of the RCA has been greatly reduced, and a large portion of the hardware for preskewing/deskewing shift registers is saved.

B. Pipelined Accumulator Structure

A basic nonpipelined accumulator structure is shown in Fig. 6 in which an RCA is followed by a row of fli-flops. The flip-flops cannot be triggered until the completion of the 27-b carry propagation path, whose delay is excessive. By partitioning the carry propagation path with latches, a pipelined accumulator is obtained as shown in Fig. 7(a). Half-cycle latch delays (black dots) are used because the 1-b addition is designed to occupy

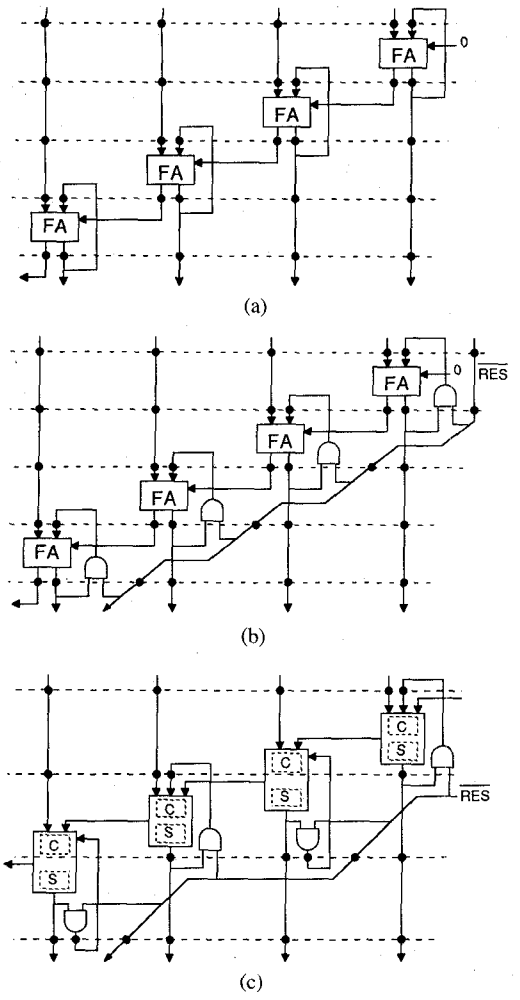


Fig. 7. Pipelined accumulators. (a) Free-running accumulator. (b) Accumulator with reset mechanism. (c) Accumulator with reduced carry propagation latency.

a *half* clock cycle. The accumulation loop at each bit position must include two such latches (which are C^{2}MOS gates in the MAC design) to provide a delay of one clock cycle. Thus, the preceding accumulated value and the new input word can be sent into the FA cells simultaneously.

In a nonpipelined accumulator, a fast adder scheme such as a carry-lookahead adder (CLA) [9] can reduce the propagation delay. However, in a bit-level pipelined system, the throughput is determined only by the 1-b accumulation loop delay. Thus, using any propagation-accelerating strategy would in fact *degrade* the throughput since it would increase the number of gate delays in each accumulation loop.

To provide a reset mechanism, AND gates are added in the *feedback* paths rather than the feedforward paths of the accumulation loop, as shown in Fig. 7(b). Thus, the reset, dumping, and the beginning of the next accumulation cycle are performed concurrently, and the input data sequence is not interrupted. Since a full add is accomplished in a half clock cycle, the additional AND gates can use the other half clock cycle and will not degrade the system throughput.

By reviewing the multiplier structure of Fig. 4, it can be seen that the accumulator can be absorbed into the multiplier by merely placing the accumulating FA cells and AND gates

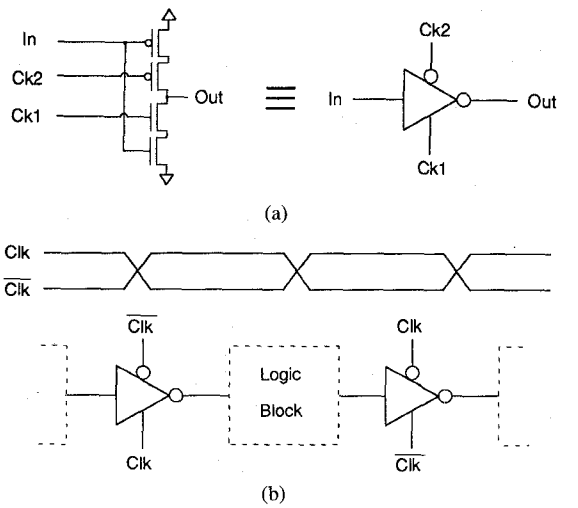


Fig. 8. (a) C^{2}MOS dynamic latch. (b) Complementary-phase clocking scheme.

of Fig. 7(b) next to the CSA array of the multiplier. This is exactly the design approach adopted in accumulator section I of Fig. 2. On the other hand, the structure of accumulator section II is modified from the pipelined RCA in Fig. 5 by properly deploying AND gates and propagating the RES signal, as shown in Fig. 7(c). Thus, the latency of the 12 $\times$ 12-b multiplier is completely hidden in the latency of the 27-b two-section accumulator, and not a single clock cycle is wasted.

Once the reset signal is latched into the accumulator simultaneously with the input word, the subsequent reset sequence is self-synchronizing, and the internal reset pulse will eventually be sent out as a data-valid (DV) signal along with the accumulation result, indicating the completion of an accumulation cycle. Thus, external circuits for keeping track of the latency for "catching" the accumulation result are not needed.

V. High-Speed Timing and Interfacing

A. Minimum-Skew Clock Buffer Design

In order to shorten the delay overhead, most of the basic cells in the MAC chip use C^{2}MOS latches (Fig. 8(a)) for pipelining purposes. For controlling the switches, either complementary-phase clock signals or nonoverlapping two-phase four-wire clock signals can be used. To reduce the idling time of basic cells and save area for clock signal routing, a complementary-phase clocking scheme was selected for the MAC chip, as shown in Fig. 8(b).

It has been shown that no data racing phenomena will occur if there is zero or an even number of successive logic inversions performed between two C^{2}MOS stages that are driven by a complementary-phase clock [1]. Thus, the most race-prone structure is the one containing a single quasi domino inverter located between two C^{2}MOS stages, as shown in Fig. 9. With 1-ns rise/fall time for clock signals, SPICE simulation showed that this worst-case structure can tolerate up to 0.6 ns of time skew measured between the half-V_{dd} points of the two complementary-phase clock signals. To achieve

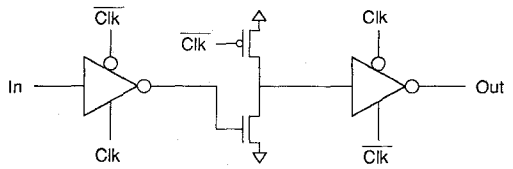


Fig. 9. The most race-prone structure in the complementary-phase clocking scheme.

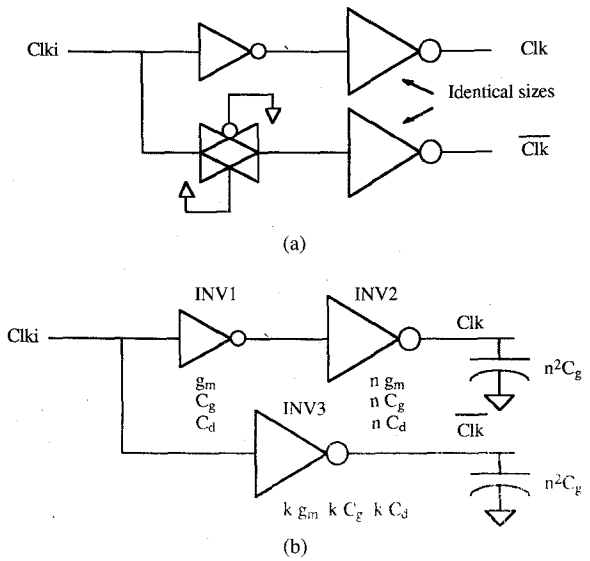


Fig. 10. (a) Traditional delay-equalizing buffer. (b) Improved delay-equalizing buffer design.

the highest throughput, however, it is essential to minimize the skew. To reduce the clock skew, a conventional scheme [10] uses an always-on C-switch to equalize the buffer delays, as shown in Fig. 10(a). According to SPICE simulations, however, the output clock skew is not completely immune from IC process variations.

To derive even better performance from the clocking subsystem, a process-insensitive clock buffer generator with minimal skew and a short delay has been designed. The delay-equalizing buffer is shown in Fig. 10(b). Based on the definition that the *stage ratio* is the ratio of the total transistor width in an inverter to that in the preceding inverter, it has been shown [11] that a minimum delay of an inverter-chain buffer can be obtained if the stage ratio between any two successive inverters is kept at a fixed value, say, n. For clarity of the delay analysis, it is assumed that the P-MOSFET and N-MOSFET in the same inverter have equal current-driving capabilities by means of channel-width adjustments, and their driving capabilities can be represented by the average effective transconductance g_m during logic transitions.

If the inverter INV1 in Fig. 10(b) has a driving force g_m, gate capacitance C_g, and drain diffusion capacitance C_d, the inverter INV2 will have parameters ng_m, nC_g, and nC_d, respectively, and the load driven by INV2 should be equal to n^2C_g. The inverter INV3 that also drives a load of n^2C_g has the parameters kg_m, kC_g, and kC_d. If a delay T_i is associated with the ith inverter, then T_i is proportional to the ratio C_{load}/g_{mi}, and the desired condition will be that

$$T_3 = T_1 + T_2. \tag{1}$$

If the parasitic drain diffusion capacitance is neglected as a first-order approximation, then (1) implies that

$$\frac{n^2C_g}{kg_m} = \frac{nC_g}{g_m} + \frac{n^2C_g}{ng_m} \tag{2}$$

or

$$k = \frac{n}{2}. \tag{3}$$

For more accurate analysis, the diffusion capacitance should be considered, and (1) becomes

$$\frac{kC_d + n^2C_g}{kg_m} = \frac{C_d + nC_g}{g_m} + \frac{nC_d + n^2C_g}{ng_m}. \tag{4}$$

Usually, C_d is comparable to C_g in the same inverter, therefore by setting $C_d = C_g$, we have the simple formula

$$k = \frac{n}{2}\left(1 + \frac{1}{2n}\right)^{-1} = \frac{n^2}{2n+1} \tag{5}$$

where n is the stage ratio of normal buffers and k is the stage ratio of the delay equalizing inverter (INV3). In MAC chip, n is chosen to be 3. According to SPICE simulations, (5) is quite effective and needs only minor adjustments of the buffer transistor sizes to minimize the clock skew.

Fig. 11 shows the clocking subsystem of the MAC chip. Although the loading effects on CLK and $\overline{CLK}$ within each leaf cell may be different, the chip-wide loading effects are quite balanced since the rolls of CLK and $\overline{CLK}$ are interchanged in every other pipeline stage. The effects of the package-lead inductance (L_{pkg}), the input-pad capacitance (C_{pad}), and the protection resistance (R_{prt}) were included in the simulations. Their typical values are 10 nH, 0.2 pF, and 300 Ω, respectively. During simulation, the output skew of the clock generator using the circuit in Fig. 10(a) has been minimized based on nominal transistor models. Yet, a larger clock skew still results from the same circuit if *fast* N- and *slow* P-models are used to mimic an extremely unbalanced IC process. However, simulations show that the skew in this extreme case is very small if the timing subsystem uses the circuit in Fig. 10(b), even though the transistor widths were tuned to minimize the skew under nominal models only. The propagation delay resulting from the circuit in Fig. 10(b) is also shorter than that in Fig. 10(a).

In addition to the process insensitivity and shorter delays, the proposed buffer design also consumes less transition power and chip area than the traditional design because INV3 in Fig. 10(b) is smaller than INV2 and the C-switch has been removed. Although INV3 has longer rise/fall time than INV2, their output waveforms are reshaped through the local buffers in Fig. 11.

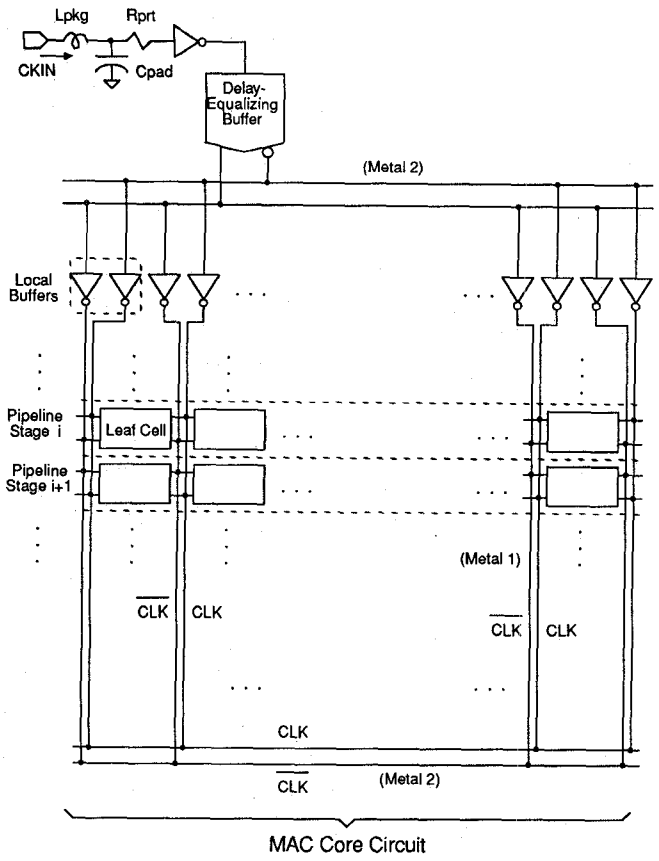


Fig. 11. Clocking subsystem.

B. High-Speed Low-Noise Latched Output Buffer

Although the MAC chip clocking subsystem has been optimized to achieve a worst-case buffering delay as low as 2 ns, it is still too long for off-chip communication with full CMOS swings at a 200-MHz clock rate. The delay path from the MAC chip clock input to its data output includes the clock buffer/generator, the clock-to-data delay path in the register, the conventional four-inverter output buffer, package-lead inductance, and off-chip capacitive load. The total delay can easily reach or exceed one clock period (5 ns), making the next chip's input setup time very short or even negative. This is the primary reason that CMOS IC's have difficulty in achieving high-speed interchip communication.

Several solutions have been proposed to alleviate this problem. The first one makes the on-chip clock signal pass through an output buffer, and uses this output to clock the next chip. This approach is not ideal if there is bidirectional data flow or feedback loops among chips. A second technique uses an on-chip PLL to synchronize the timing signals on different chips [12]. Its drawbacks are: 1) the maximum speed and tuning range of the clock signal are limited by the analog VCO; 2) the noise from digital circuits induces phase jitter; and 3) it is a fairly complex solution requiring careful mixed analog/digital design. The third solution involves the use of ECL-level I/O buffers [13]. This technique can achieve high-speed interfacing at the cost of large standby power consumption. Moreover, to make the logic levels insensitive to IC process and environmental variations, this technique requires complex circuit design, multiple power supplies, and off-chip discrete components.

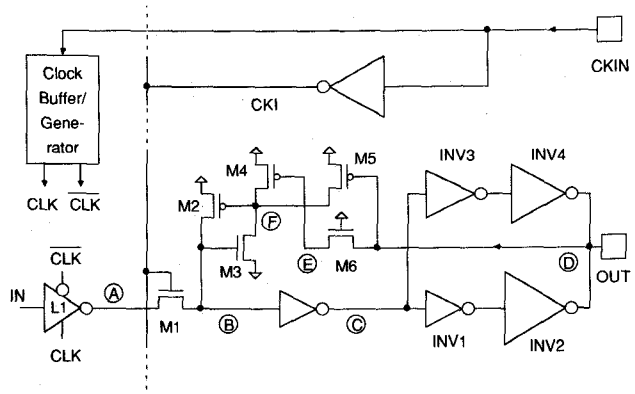


Fig. 12. High-speed low-noise latched output buffer.

Fig. 12 shows a new latched output buffer design that achieves two goals: 1) to realign the delayed on-chip timing caused by the clock buffer; and 2) to minimize the output buffer delay and reduce the power/ground line bounce due to package-lead inductance. Using this approach, board-level high-speed communication can be achieved among pure CMOS VLSI chips by using only a single global master clock.

In Fig. 12, $CKIN$ is the off-chip master clock. As mentioned before, it is converted into a complementary-phase clock (CLK and $\overline{CLK}$) by the clock buffer/generator. On the other hand, CKI, an inverted version of $CKIN$, is generated to trigger all latched output buffers. If clock buffer delays are neglected, then the C^2MOS latch $L1$ is transparent when $CKIN$ is high, while transistor $M1$ is on when $CKIN$ is low. Thus, $L1$ and $M1$ effectively construct a falling-edge-triggered flip-flop. Moreover, the delay from $CKIN$ to $CLK/\overline{CLK}$ is nearly 2 ns, while the delay from $CKIN$ to CKI is only about 0.5 ns. Hence, the timing is pushed forward (or improved) by nearly 1.5 ns. The reason that a C-switch is not used in place of $M1$ is that the P device would require an inverted version of CKI, which would result in more delay.

Within the buffer, the last two stages of inverters have been split into four inverters, INV1 to INV4, which construct a transition-smoothing buffer. INV1 is smaller than INV3, but INV2 is larger than INV4. When there is a transition at node C, INV4 will start to switch earlier than INV2. Thus, a small output current from INV4 will be followed by a larger output current from INV2 to drive the off-chip load. Therefore, this transition-smoothing buffer results in much smaller power/ground line bounce than a conventional inverter-chain buffer because the rate of change of current in the package-lead inductance is reduced. Although the rise/fall times at node D are increased, the half-V_{dd} delay will not be prolonged because the output transition starts earlier.

As a further enhancement, transistors $M2$ to $M6$ construct a speed-up feedback network. If this network did not exist, then the high level of node B could be at most $V_{dd} - V_{TN}$, which is only about 3.2 V due to the body effect of $M1$. This affects the speed of the following inverter and reduces the noise margin. A conventional way to alleviate this problem is to add a pull-

up device $M2$ whose gate is connected to node C. Thus, when node B rises from 0 to 3.2 V, node C falls and turns on $M2$, pulling node B further up to 5 V. The shortcomings of this method are: 1) when node B rises, node C falls slowly due to the large loading, which delays the operation of $M2$ and degrades the speed-up effects; and 2) when node B falls, $M2$ is still on due to the slow rise of node C, and thus the fall time of node B is prolonged.

The newly designed speed-up feedback network operates as follows. If the previous state of node B is low and output node D is high, then $M3$ is off and $M4$ is just weakly on. This keeps node F at 5 V and $M2$ off. If node B rises due to an input transition, $M3$ turns on and node F will fall faster than node C because of the weak $M4$ and small loading. Thus, $M2$ is turned on fast which shortens the rise time of node B, thereby decreasing the total buffer delay. After node B reaches 5 V and output D falls to 0 V, $M4$ and $M5$ are fully on and pull node F up to make $M2$ only weakly on. Thus, if node B starts to fall in the next clock cycle, $M3$ is weakened and node F rises further to turn $M2$ off quickly. Hence, $M2$ will not hinder the falling operation of node B, and the total buffer delay will not be prolonged.

In summary, the CKI-driven N-switch and the speed-up feedback network reduce the MAC chip output delay by nearly 2 ns, resulting in a total clock-to-output delay of about 3 ns. Furthermore, the transition-smoothing buffer insures high-throughput operation by reducing the power-ground line bounce. The new buffer design does not consume extra chip area since its total transistor width is less than a conventional four-inverter output buffer.

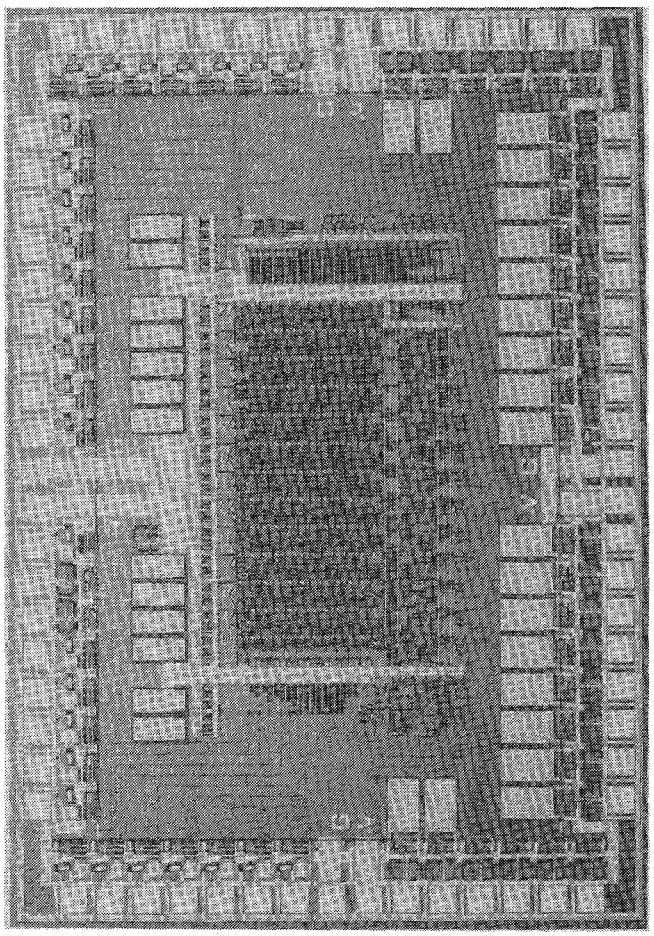


Fig. 13. MAC chip photomicrograph.

VI. Fabrication and Test Results

Fig. 13 shows a photomicrograph of the MAC chip. The complexity is about 10 000 transistors and the die area is 2.5 × 3.7 mm^2. The device is packaged in a 68-pin leadless chip carrier. All building blocks have been reformed into a rectangular shape so as to achieve high regularity of the layout and an efficient use of chip area. Shunt capacitors were laid out in spare areas to stabilize the on-chip power rails during logic transitions. The chip was fabricated in a 1.0-μm single-polysilicon double-metal P-well CMOS technology by the TRW Microelectronics Center.

Since no off-the-shelf TTL/CMOS parts can provide input vectors and control signals at speeds well beyond 100 MHz, two MAC chips are used to test the speed performance based on the test setup shown in Fig. 14. The chip MAC1 accepts static inputs from DIP switches. Since pull-down resistors have been added on chip, an open switch represents a logic ZERO input. The MAC1 outputs, toggling at full speed, are then used as the inputs and control signals of MAC2 to test the speed performance.

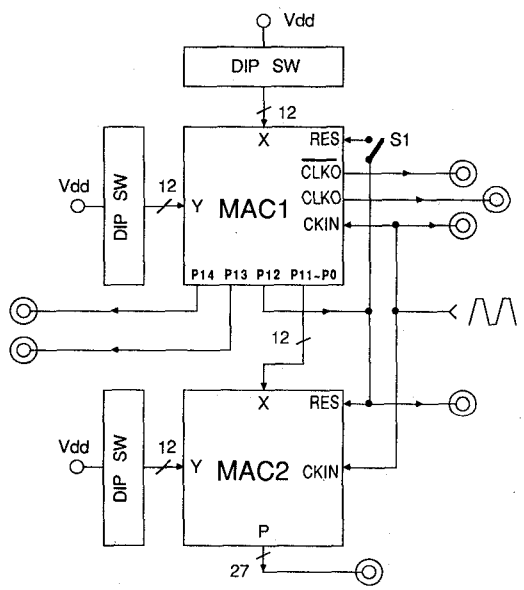


Fig. 14. Complete test circuit for the MAC chip.

The on-chip complementary-phase clock signals have been buffered and sent out as $CLKO$ and $\overline{CLKO}$ to inspect the effectiveness of the delay-equalizing buffer design in Fig. 10(b). Since all internal data are only transferred among adjacent cells due to the pipeline structure of MAC chip, only local clock skews need to be considered. The upper trace in the photo of Fig. 15 shows the 166-MHz single master clock seen by MAC1. The lower two traces represent $CLKO$ and $\overline{CLKO}$, whose skew is only about 0.2 ns. All three scope probes are of the same length and their input loading is 450 Ω and 1.5 pF.

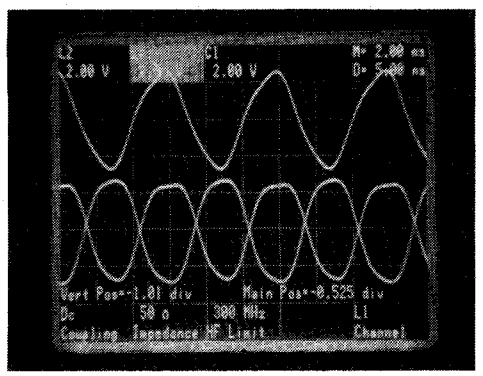
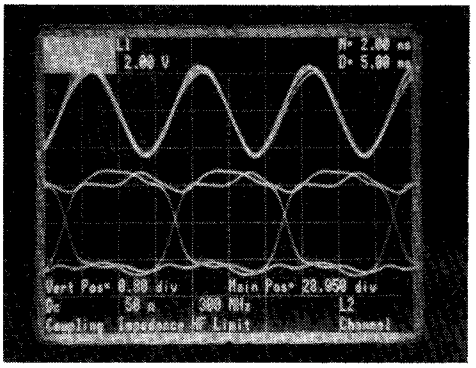
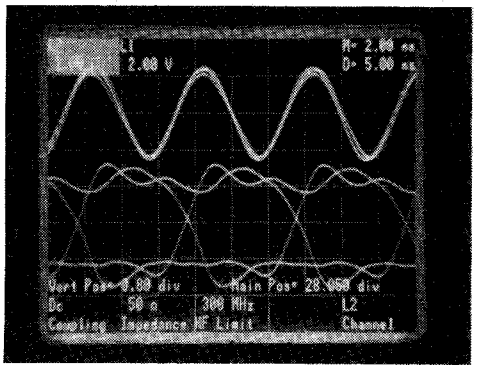


Fig. 15. Buffered outputs of complementary clock signals.



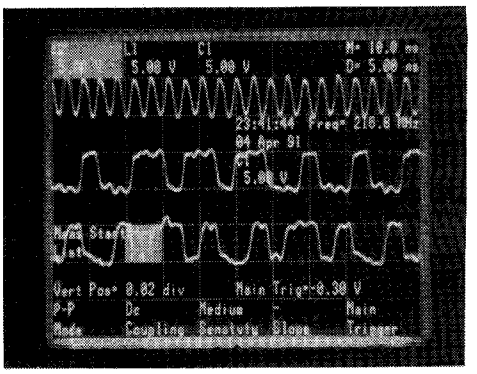
(a)



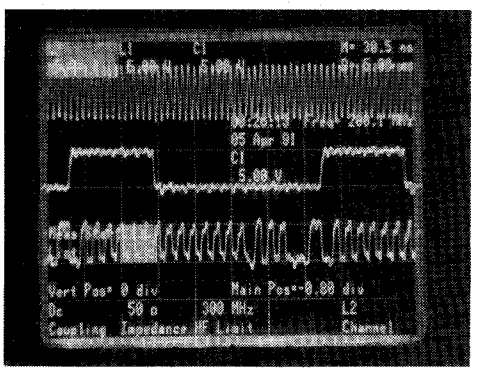
(b)

Fig. 16. Clock-in-to-data-out delay (2 ns/div). (a) 1.5-pF and 450-Ω off-chip load. (b) 30-pF and 450-Ω off-chip load.

Fig. 16 demonstrates the performance of the latched output buffer design shown in Fig. 12. The upper trace of Fig. 16(a) is the clock input of MAC1, and the lower trace is the $P14$ (the 15th LSB of the accumulator) output from the same chip. Loaded only by the scope probe, the output propagation delay (measured at 2.5 V) is only 3.2 ns from the falling edge of the master clock (all MAC chip I/O are falling-edge triggered). In Fig. 16(b), the output waveform is loaded by an extra 30-pF capacitance. The total clock-to-output delay is still no more than 3.6 ns. This implies that the latched output buffer design can sustain CMOS-level off-chip communication with up to 30-pF load at 200 MHz, and only a single master clock is required by the system.



(a)



(b)

Fig. 17. Examples of MAC chip output waveforms. (a) Two outputs of the free-running accumulator. (b) Multiply–accumulate and reset operation.

Fig. 17(a) shows two typical output waveforms of MAC1 when its inputs are static and the accumulator is running without interruption (i.e., RES is set to ZERO by opening $S1$). The $P13$ and $P14$ waveforms were probed because testing has verified that they have the lowest speed performance. Fig. 17(a) reveals that the accumulator itself can operate at clock rates up to 215 MHz.

To investigate the board-level speed performance of MAC chip, including the control interface, a small fixed binary input word is fed to the MAC1 accumulator to obtain a long output repetition period. The $P12$ output is fed back to the RES input of MAC1 by closing $S1$. This results in a rectangular waveform at $P12$ with its high state lasting for 14 clock cycles (pipeline latency + one cycle). Outputs $P0$ to $P11$ of MAC1 are sent to MAC2 to test the multiplier and interchip I/O interfacing. $P12$ of MAC1 is also sent to the RES input of MAC2 to periodically restart its accumulation. The bottom trace of Fig. 17(b) is an example of a MAC2 output waveform whose transitions occur most frequently. The central trace is the RES input of MAC2 ($P12$ of MAC1). The tested maximum speed of the circuit is 200 MHz.

Fig. 18 shows the MAC chip maximum speed performance and the corresponding power consumption versus V_{dd} variations. With a +5-V supply at room temperature, the maximum clock rate is 200 MHz and the power consumption is 1.3 W with all output pins driving one TTL load each. When the supply voltage is reduced to 3 V, the chip can still operate at 120 MHz. This demonstrates the robustness of the delay-equalizing clock buffer and quasi N-P domino logic design

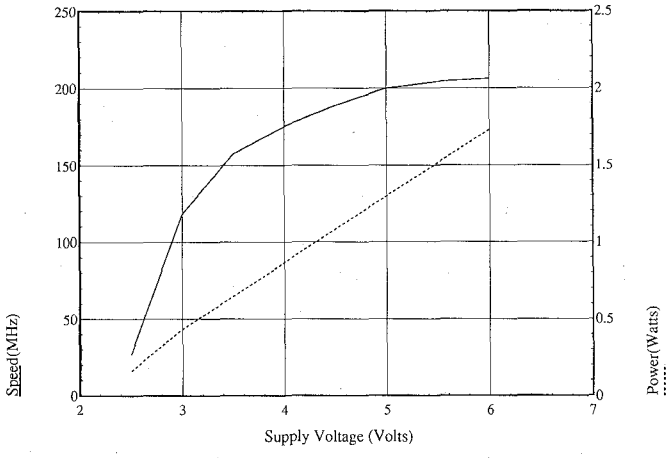


Fig. 18. Maximum speed and power consumption versus supply voltage.

TABLE II
MAJOR CHARACTERISTICS OF MAC CHIP

IC process	1.0-um CMOS
Chip area	2.54 x 3.66 mm^2
Transistor count	10,616
Power supply	+5V
Input wordlength	12 bits x 2
Output wordlength	27 bits
Throughput	200 MHz
I/O data bandwidth	1.28 Gbytes/sec.
Pipeline latency	13 cycles
Power consumption	1.3 W @200MHz
Power-speed ratio	6.5 mW/MHz

technique. With a 1-μm process, the devices are subject to short-channel effects; thus the power curve does not follow a square-law characteristic but behaves more linearly. Table II summarizes the characteristics of the MAC chip.

VII. CONCLUSIONS

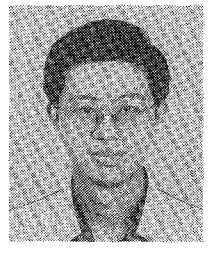
To achieve the high throughput requirements of many advanced DSP systems, a 200-MHz pipelined 12 × 12-b multiplier–accumulator has been designed. The high throughput rate and the short latency of the CMOS MAC chip are achieved by using a new full-adder cell design and a special pipelined structure. The reset mechanism was designed so that the internal reset and rounding operations are self-timed without interrupting the input data flow and the accumulate-and-dump sequence. Efficient clock buffer and output buffer designs were presented which are crucial for achieving high-speed timing and board-level interfacing.

ACKNOWLEDGMENT

The authors would like to thank the TRW Microelectronics Center for contributing free fabrication services, and H. Nicholas for his valuable technical interaction over the course of this project.

REFERENCES

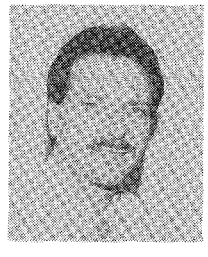
[1] N. Weste and K. Eshraghian, *CMOS VLSI Design.* Reading, MA: Addison-Wesley, 1988.
[2] J. Y. Lee *et al.*, "A high-speed high-density silicon 8 × 8-bit parallel multiplier," *IEEE J. Solid-State Circuits,* vol. SC-22, pp. 35–40, 1987.
[3] N. F. Goncalves and H. J. De Man, "NORA: A racefree dynamic CMOS technique for pipelined logic structures," *IEEE J. Solid-State Circuits,* vol. SC-18, pp. 261–266, June 1983.
[4] C. R. Baugh and B. A. Wooley, "A two's complement paralle array multiplication algorithm," *IEEE Trans. Comput.,* vol. C-22, pp. 1045–1047, Dec. 1973.
[5] N. R. Scott, *Computer Number Systems & Arithmetic.* Englewood Cliffs, NJ: Prentice-Hall, 1985.
[6] C. S. Wallace, "A suggestion for a fast multiplier," *IEEE Trans. Electron. Comput.,* vol. EC-13, pp. 14–17, Feb. 1964.
[7] H. T. Kung, "Why systolic architectures?," *Computer,* pp. 37–46, Jan. 1982.
[8] T. G. Noll *et al.*, "A pipelined 330-MHz multiplier," *IEEE J. Solid-State Circuits,* vol. SC-21, pp. 411–416, June 1986.
[9] J. Sklansky, "An evaluation of several two-summand binary adders," *IRE Trans. Electron. Comput.,* vol. EC-9, pp. 213–226, June 1960.
[10] N. Ohwada *et al.*, "LSI for digital signal processing," *IEEE J. Solid-State Circuits,* vol. SC-14, pp. 221–239, Apr. 1979.
[11] C. A. Mead and L. A. Conway, *Introduction to VLSI Systems.* Reading, MA: Addison-Wesley, 1980.
[12] K. Kurita *et al.*, "PLL based BiCMOS on-chip clock generator for very high-speed microprocessor," *IEEE J. Solid-State Circuits,* vol. 26, pp. 585–589, Apr. 1991.
[13] T. Yoneda *et al.*, "An ECL compatible full CMOS 210 Mbps crosspoint switch," in *Proc. IEEE CICC,* 1989, pp. 10.7.1–10.7.4.



Fang Lu (S'87–M'93) was born in Taipei, Taiwan, Republic of China, in 1962. He received the B.S. degree in electrical engineering from National Taiwan University in 1984, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, in 1988 and 1992, respectively.

From 1984 to 1986 he served in the Chinese Army as an Electronics Instructor. From 1987 to 1989 he worked as a Teaching Assistant at the University of California, Los Angeles (UCLA). From 1987 to 1992 he was also a Research Assistant in the Integrated Circuits and Systems Laboratory at UCLA. He is now a Senior Staff Engineer at Baseband Technologies, Inc., Los Angeles, CA, and his responsibilities include the circuit design of high-performance VLSI in digital signal processing and communication systems.

Dr. Lu is the recipient of the 1991–1992 IEEE Solid-State Circuits Council's Predoctoral Fellowship Award.



Henry Samueli (S'75–M'81) was born in Buffalo, NY, on Sept. 20, 1954. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Los Angeles (UCLA), in 1975, 1976, and 1980, respectively.

From 1980 to 1985 he was with TRW, Inc., Redondo Beach, CA where he was a Section Manager in the Digital Processing Laboratory of the Electronics and Technology Division. His group was involved in the hardware design and development of military satellite and digital radio communication systems. From 1980 to 1985 he was also a part-time instructor in the Electrical Engineering Department at UCLA. In 1985 he joined UCLA full-time where he is currently an Associate Professor in the Electrical Engineering Department. His research interests are in the areas of digital signal processing, digital filter design, analysis of finite word-length effects in DSP systems, high-speed CMOS integrated circuit design, VLSI architectures for realizing DSP algorithms, and applications of VLSI technology to digital communication systems.