High-Speed CMOS Circuit Technique

JIREN YUAN AND CHRISTER SVENSSON

Abstract—**We have demonstrated that clock frequencies in excess of 200 MHz are feasible in a 3-μm CMOS process. This is obtained by means of clocking strategy, device sizing, and logic style selection. We use a precharge technique with a true single-phase clock, which remarkably increases the clock frequency and reduces the skew problems. Device sizing with the help of an optimizing program improves circuit speed by a factor of 1.5–1.8. We minimize the logic depth to one instead of two or more and use pipeline structures wherever possible. The presentation includes experimental demonstrations of several circuits which work at clock frequencies of 200–230 MHz. SPICE simulation shows that some circuits possibly work up to 400–500 MHz.**

I. Introduction

MANY factors control the possible speed of CMOS integrated circuits. There are, for example, device dimensions, logic circuit style, clocking strategy, architecture, clock distribution, etc. To pursue high speed and integration density the dimensions of MOS transistors are scaled down continuously. The delay in a CMOS circuit will be inversely proportional to the scaling factor α if all dimensions are reduced without changing physics [1]. However, there are physical, geometrical [2], [3], and also cost limits on scaling down transistors. Therefore, we should tap the potential of the most popular technique. In fact, we have been investigating the possibilities of increasing speed by combining different circuit techniques in an available and relatively low-cost process, for example, the 3-μm CMOS process. In the present work we will limit ourselves to a discussion of a high-clock-frequency synchronous CMOS circuit technique in a given process (i.e., with a given smallest dimension device). We will assume that the circuit technique rather than the architecture limits the clock frequency. Our results are therefore applicable mainly to simple, pipelineable architectures. In this article, we will present our results by means of both analysis and experiments. Section II describes a new clocking strategy with its accompanying circuit technique and its importance for high clock frequency. We further investigate the robustness and flexibility of this technique and propose some further improvements. Section III describes the effect of logical optimization of the circuits and Section IV describes the effect of device sizing. In Section V we present some experiments, illustrating the previous results. Finally we present our conclusions in Section VI.

Manuscript received May 25, 1988; revised August 10, 1988.
The authors are with the Department of Physics and Measurement Technology, LSI Design Center, Linköping University, 581 83 Linköping, Sweden.
IEEE Log Number 8824917.

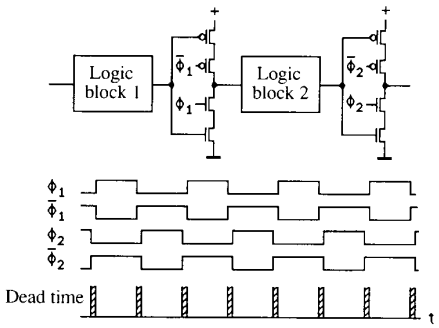


Fig. 1. C^2MOS logic.

II. Clocking Strategy and Circuit Examples

A. True Single-Phase-Clock Circuit Techniques

In conventional CMOS circuits both static and dynamic CMOS logic is used. For the purpose of system timing a clocking strategy is always involved except for a self-timed system [4]. The most popular clocking strategy is clocked CMOS logic (C^2MOS) [5], [6] which uses a nonoverlapping pseudo two-phase clock as shown in Fig. 1. Four clock signals have to be distributed in such a system and between two pairs of clock signals there should be no overlap. Clock skews in the system will cause serious problems and result in difficulties in increasing circuit speed [7].

The NORA dynamic CMOS technique [8] uses a true two-phase-clock signal ϕ and $\bar{\phi}$ instead of a four-phase-clock signal and can avoid race problems caused by clock skews with some constraints on logic composition. In a NORA pipelined system, ϕ-C^2MOS latches and $\bar{\phi}$-C^2MOS latches are alternatively used. The most important constraint is that between two C^2MOS latches there must be an even number of inversion blocks and if there are static blocks between a precharge block and a C^2MOS latch they must also be of an even number. We choose two typical NORA constructions which are called ϕ section and $\bar{\phi}$ section and use an N-precharge block in the ϕ section and a P-precharge block in the $\bar{\phi}$ section for our discussion. These are shown in Fig. 2. Since there is no dead time and no skew problem, it is expected that the NORA dynamic

0018-9200/89/0200-0062$01.00

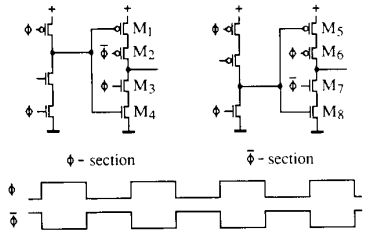


Fig. 2. NORA dynamic CMOS technique.

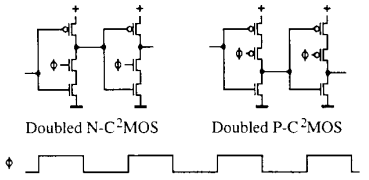


Fig. 3. True single-phase-clock latch stages.

CMOS technique can reach higher clock rates than the C²MOS technique.

A further development in clock strategy should use even less clock signals. The true single-phase-clock dynamic CMOS circuit technique [9] uses only one clock signal which is never inverted. Therefore, no clock skew exists except for clock delay problems and even a higher clock frequency can be reached. As will be explained below, the true single-phase-clock CMOS technique fits not only dynamic but also static CMOS circuits and in most cases can replace the NORA CMOS technique.

Let us discuss only the latch stages first. In Fig. 2, we have seen two C²MOS latch stages controlled by two clock signals ϕ and $\bar{\phi}$. The necessity of $\bar{\phi}$ lies in controlling transistors M_2 and M_7. However, this can be done by clock signal ϕ at inverters connected to the two stages as shown in Fig. 3. We call the two different units N-C²MOS stage and P-C²MOS stage, respectively, and if we use doubled N-C²MOS or P-C²MOS in series they become true single-phase-clock latch stages, i.e., N-latch and P-latch. In this system, an N-section (N-latch plus logic blocks) and a P-section (P-latch plus logic blocks) are used alternatively using the same clock signal. Both static and dynamic blocks are accepted and an N- or P-precharge block is used in the N- or P-section, respectively. As long as the clock delay is less than the gate delay the system is reliable. Instead of distributing the $\bar{\phi}$ clock signal there are two transistors more in each latch stage, but the most important thing is that there will be no even inversion constraint either between two latches or between the latch and the dynamic block. Apparently, this is better than the nonoverlapping pseudo two-phase-clock strategy and also, from the point of view of logic constraints, can compete with the NORA two-phase-clock strategy. The logic function blocks can be included in the N-C²MOS or P-C²MOS latch stages or placed between them as shown in Fig. 4, depending on the logic type and the requirement of inversion.

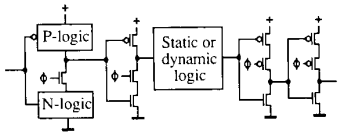


Fig. 4. Logic arrangements using true single-phase-clock latch stages.

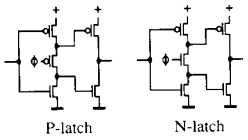


Fig. 5. Split-output latch stages.

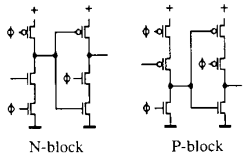


Fig. 6. TSPC-1 circuit.

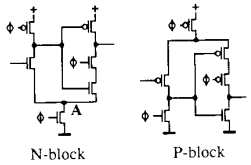


Fig. 7. TSPC-2 circuit.

Furthermore, the true single-phase-clock latch stages shown in Fig. 3 can evolve into a simpler version, called split-output latch, as shown in Fig. 5 where only one of the transistors is controlled by the clock. It implies half the clock load. A possible drawback of the split-output technique is that all node voltages do not have a full voltage swing, as some single transistors are used for the transmission of both high and low signals.

In the case of using precharge dynamic logic, let us go back to the NORA circuits in Fig. 2. The P-transistor M_2 in the ϕ section and the N-transistor M_7 in the $\bar{\phi}$ section are unnecessary because precharge signals will play the same role as $\bar{\phi}$ in both sections, so they can be omitted. In Fig. 6, this evolved into the true single-phase-clock dynamic CMOS technique as described in [9]; we call it true single-phase-clock 1 (TSPC-1). We introduce a further modified circuit as shown in Fig. 7. We call it TSPC-2, which is better than TSPC-1 in performance, as described below. In the following description, we choose the ϕ section of NORA and N-blocks of TSPC-1 and TSPC-2 dynamic circuits as examples.

Compared with the nonoverlapping pseudo-two-phase clock (NPTC) and the NORA two-phase clock, besides the compact simple clock distribution of TSPC which will naturally lead to a higher clocking speed, we can summarize other features of them as follows.

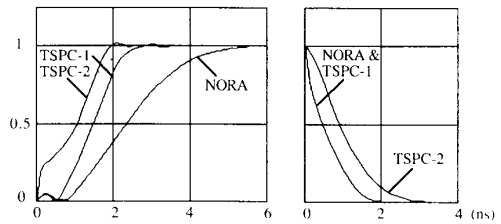


Fig. 8. Step responses of NORA, TSPC-1, and TSPC-2.

B. Step Response

Because the transistor number is reduced from four in a C^2MOS stage to three in either an N-C^2MOS or a P-C^2MOS stage, the delay of the latch stage is reduced. Fig. 8 shows the step responses for the three circuits simulated by SPICE with unit transistor sizes. For the sake of realism, we have put a unit inverter load into all three circuits. In this paper, all SPICE simulations are done by using typical level 2 SPICE parameters for a standard 3-μm double-metal CMOS process [10] and using a power supply voltage of 5 V.

First, we find that omitting the P-transistor in the latch stage is significant because the most critical slope in the circuit is the rise slope. This gives TSPC-1 speed improvement of a factor of 1.8. Second, by omitting the P-transistor the output of the latch stage may continue to rise during the initial part of the succeeding precharge phase, thus allowing a shorter evaluation time. Third, we should explain why TSPC-2 has even better performance. In precharge logic circuits, at the start of evaluation the latch stage tends to output the precharge state first, i.e., to make the output low because of the high precharge node, as can be seen in Fig. 8 at the beginning of the rise slope. This is a common problem for precharge circuits. It is found that the circuit TSPC-2 can solve this problem partly. In Fig. 7, if the logic part is conducting, node A will also be charged to high during the precharge phase, which prevents the output from going low. This makes the rise time shorter. On the other hand, the fall time will increase but this only makes both slopes more balanced in time. Another advantage is that when the output should be kept high, the circuit has much smaller dips in the output as shown in Fig. 9, which is obtained from SPICE simulations. For the P-block of TSPC-2 circuit the top P-transistor should be widened by a factor of 1.5–2 in order to obtain all the advantages mentioned.

C. Sensitivity to the Clock Slope

No maximum slope limit exists in NPTC circuits as long as the clocks are nonoverlapping. This is not true in NORA and TSPC circuits. SPICE simulation shows that NORA and TSPC-1 can accept slopes up to 20 ns for rise and fall edges without disorder. This figure agrees well with experimental results on TSPC-1 in [11]. This is about 20 times larger than the normal gate delay. Because we are interested in the possibility of working at high frequencies,

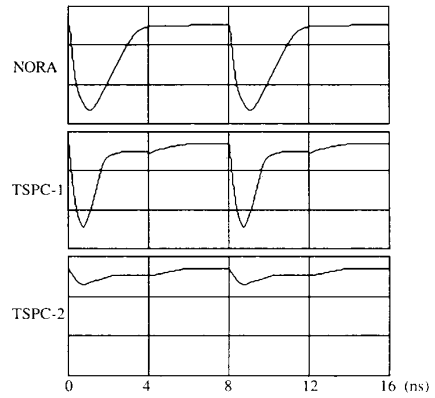


Fig. 9. Output dips of different circuits.

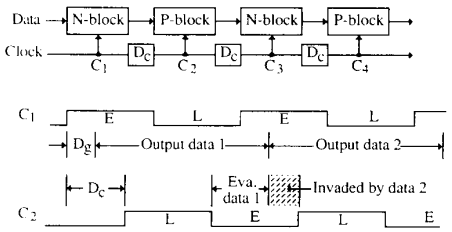


Fig. 10. Problem caused by clock delay. D_g = gate delay; D_c = clock delay; E = evaluation; L = latch.

a slope of 20 ns is quite acceptable. For TSPC-2, the acceptable clock slope reaches 2 ms, which is of the same order of magnitude as the clock period to be accepted by dynamic CMOS circuits. It means that TSPC-2 is more reliable in latching a signal.

D. Clock-Skew Problems

In general terms, no problem caused by clock skew exists in NORA and TSPC circuits so they are better than the NPTC strategy. However, clock skews will compress synchronous margins in NORA circuits, and for both NORA and TSPC circuits the clock delay could be a problem, if it is larger than the gate delay. The problem is caused by data transparency from block 1 to block 2 when both are in the evaluation phase as shown in Fig. 10.

Locally, this is not a problem because the clock delay is usually less than the gate delay. In a pure pipeline structure the condition between the clock delay and the data delay is still satisfactory. In the case of a large system, we propose two ways to solve the problem.

1. Reverse Clock Distribution: It is found that if the clock is distributed in the opposite direction to the data stream the system will be safe. In this case, the evaluation phase of the next block will be completely included in the data stable zone of the last block because the reverse clock distribution creates a safety margin as shown in Fig. 11. This is true for the data stream both from P-block to N-block and from N-block to P-block.

2. D-Latch Type Structure: This can be used in the case of data feedback, e.g., the communication between two

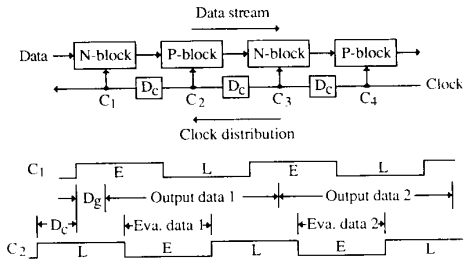


Fig. 11. Reverse clock distribution.

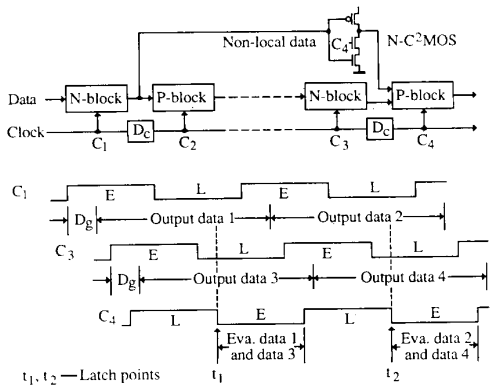


Fig. 12. *D*-latch type structure for nonlocal data.

blocks in a large system, where the reverse clock distribution is no solution as there are both forward and backward data streams. Therefore, we propose using a *D*-latch structure instead of the normal alternating P-block and N-block structure for nonlocal data as shown in Fig. 12 (see also [12]). First, if we look at Fig. 10, the problem is caused by the overlap between different evaluation phases of two communicating blocks. If the data for the next block are only latched by the start transition of the evaluation phase of this block at points t_1 and t_2 in Fig. 12, i.e., the communication between two blocks exists only during each start transition, the system will be much more reliable. This can be done simply by placing a P-C^{2}MOS stage before an N-block or an N-C^{2}MOS stage before a P-block for each "nonlocal" data. The *D*-latch structure allows communication between two blocks which have a "clock skew" caused by clock delay, up to almost half a clock cycle. This has been proved by SPICE simulation.

E. Circuit Examples

Because both static and dynamic circuits can be used, including the domino technique, the TSPC strategy has a logic flexibility as high as the NORA technique has. In most cases, NORA circuits can be replaced by TSPC circuits with little modification. One exception is that an N-precharge stage cannot be directly connected to a P-precharge stage without latch stages in the TSPC circuit, which is possible in the NORA technique. However, no even inversion requirement exists in TSPC circuits as in NORA circuits. Because of the compact clock distribution

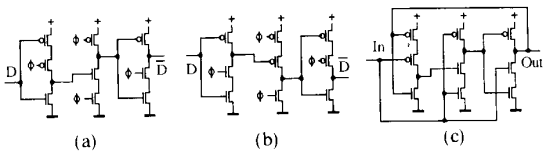


Fig. 13. Circuits constructed by P, N-blocks and P, N-C^{2}MOS stages: (a) positive transition latch; (b) negative transition latch; and (c) divide-by-two circuit.

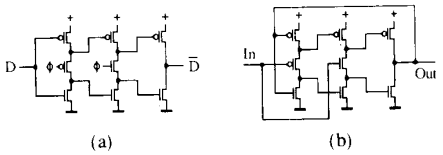


Fig. 14. Circuits constructed by split-output latch stages.

and the same circuit complexity, TSPC is preferred. We present several circuits below as examples of the TSPC strategy.

If we cascade the P-type and the N-type latches shown in Figs. 3, 5, 6, or 7, they become full dynamic transition *D* latches, each of which includes 12 transistors. However, the *D* latch shown in Fig. 13(a), which consists of nine transistors, is more effective. This is constructed by a P-C^{2}MOS stage, an N-precharge stage and an N-C^{2}MOS stage, and the input data will be latched by the positive transition of the clock signal. If we want a noninversion output an extra inverter can be placed at the output, which gives the circuit driving ability. When a negative transition latch is needed the circuit can be changed to Fig. 13(b). As expected, if the inversion output is connected to the input of the circuit, a divide-by-two circuit is formed as shown in Fig. 13(c).

If we use the split-output latch stages shown in Fig. 5 we can construct a *D* latch and a divide-by-two circuit in an even more efficient way. The resulting circuits have almost the same speed but only half the clock load and are shown in Fig. 14. The latch stages can also be used for building quite effective shift-register chains with both less transistors and less clock load than other techniques, and without output glitches.

Note that in the *D* latches of Figs. 13(a) and 14(a) the P-latches are replaced by P-C^{2}MOS stages with only three transistors. When we need fast P-passing stages in a pipelined structure the three-transistor latch is quite effective. However, it requires an input transition from low to high with a delay more than the evaluation delay of the next N-block, otherwise the active transistor in the N-block may be cut off too early. This is shown in Fig. 15 where the delay of the input signal is changed from 0.8 to 0.5 ns and the output swing is reduced to half the V_{dd}. Nevertheless, as long as these *D* latches are cascaded the delay of the last latch is satisfactory for the requirement of the next latch and the same for the divide-by-two circuits. Simulation shows that a register chain starting with an N-block (the N-block reduces the time constraints) can work at a clock rate of more than 350 MHz using a unit inverter load without critical input time requirement.

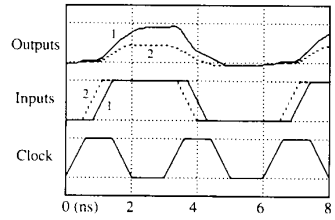


Fig. 15. Different results with the positions of the input transitions for the nine-transistor latch.

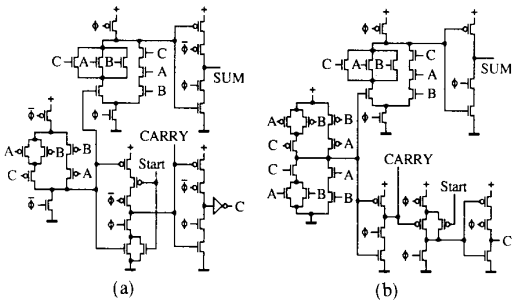


Fig. 16. Replaceability between NORA and TSPC: (a) NORA serial full adder, and (b) TSPC serial full adder.

Another circuit is a serial full adder, which is an example showing the logic replaceability between NORA and TSPC. Fig. 16(a) is the adder using the NORA technique presented in [8]. We can use N-C^{2}MOS instead of C^{2}MOS after an N-precharge stage and a negative transition latch instead of the two C^{2}MOS stages in the NORA circuit for the delay of the carry signal. The P-precharge stage is replaced by a static block as the $\bar{\phi}$ clock is lacking in TSPC, and the static block will evaluate inputs earlier in the previous phase. The resulting circuit is shown in Fig. 16(b), which has the same transistor number but higher speed than NORA and only a single clock. A corresponding circuit with an N-precharge stage instead of the static block, needing more transistors as it will use inverted signals, is presented in [9].

III. Logic Style

The maximum clock frequency in a clocked system is limited by the delay from one latching instant to the next. This delay depends on the complexity and function of each logic block between these two instants. One way to minimize this delay is to minimize the complexity of each logic block, for example, by decomposing complex blocks into pipelined parts. Such decomposition can also be done between half clock cycles when using the circuit style described above. In such a way it may be possible to reduce the logic depth to one in each block. Although this will cause an initial delay, it is acceptable in a pipelined structure. Note that in CMOS technology inverter delay is very small so that extra inverters can be accepted in the blocks.

The CMOS circuit technique is particularly sensitive to the number of transistors in series [6]. It is quite obvious

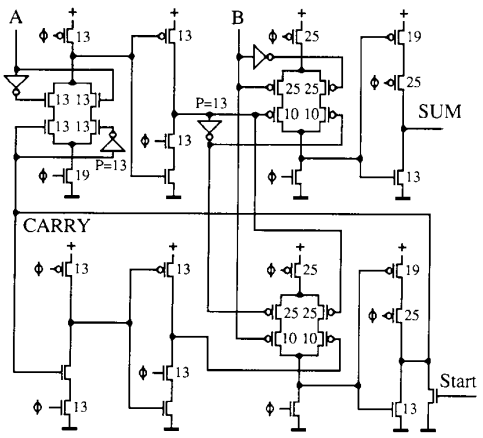


Fig. 17. Serial full adder with logic separated in different half clock cycles (TSPC adder 1).

that we should try to minimize the number of transistors in series in the logic blocks. This normally means that we also need to minimize the number of inputs to each block. We will not attempt to develop a systematic optimization procedure using the above principles. Instead we will demonstrate them by optimizing the serial full adder of Fig. 16 and make a speed comparison by using SPICE simulation. The serial full adder with a feedback path is considered a good example for studying logic style. For possible systematic approaches, see [13].

In Fig. 16(a) and (b), the Boolean functions of CARRY and SUM are

$$\text{CARRY} = AB + C(A+B)$$

$$\text{SUM} = \overline{\text{CARRY}}(A+B+C) + ABC$$

$$= \overline{(AB + C(A+B))}(A+B+C) + ABC.$$

This means that the evaluation of CARRY must be done before the evaluation of SUM and both have to be finished in half a clock cycle. We can, of course, divide the SUM evaluation into two parts and put them into different half clock cycles. Even if we do so, too many steps are still involved in the SUM evaluation and at least three transistors are in series. Let us instead change the above Boolean function to

$$E = A\bar{C} + \bar{A}C$$

$$\text{SUM} = E\bar{B} + \bar{E}B$$

$$\text{CARRY} = EB + \bar{E}C.$$

Apparently, these Boolean functions are much easier to realize and the intermediate result E can be evaluated during the previous half clock cycle. There will be only two transistors in series in the logic part. The circuit according to the above Boolean functions and using TSPC strategy has been presented in [9]. For convenience, we give it in Fig. 17. The small figures nearby transistors in Fig. 17 as well as in Figs. 18 and 19 are the scaled sizes in micrometers which will be discussed in the next section. The transistors without figures have the minimum width, 7 μm.

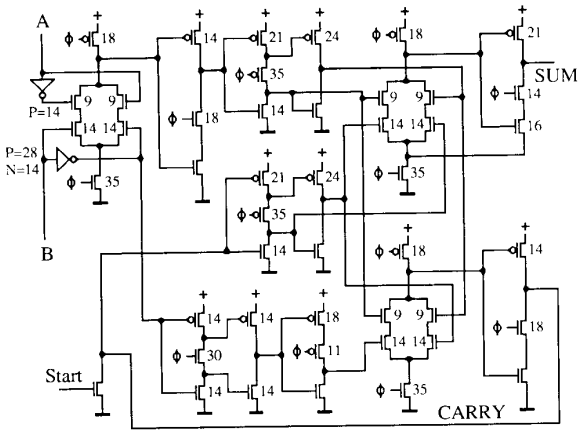


Fig. 18. Serial full adder with all logic in N-blocks (TSPC adder 2).

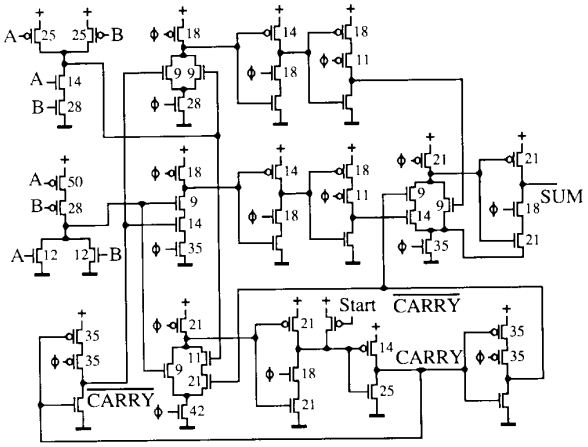


Fig. 19. Fast serial full adder (TSPC adder 3).

Compared with the adder in Fig. 16(a), the adder in Fig. 17 has a speed increase of a factor of 1.86. The critical delay of the adder in Fig. 17 is the carry feedback path, which is about 3.3 ns since a P-block is used. In Fig. 17, the initial delay is half the clock cycle. If an initial delay of one clock cycle is accepted, we can place all of the main logic part into N-blocks and reach higher speed. Fig. 18 shows such a serial full adder combining the principles described above. Note that the P-latches offer complementary signals in an efficient way and that only the sum output N-block is connected in type TSPC-2, while the other two are connected in type TSPC-1 for the delay requirements of the P-latches.

The adder in Fig. 18 has a critical delay of 1.8 ns and it can still be improved by further reducing the logic depth. Since a two-input OR gate or AND gate has the minimum logic depth, it plus a latch stage will determine the ultimate speed of a pipelined logic circuit, which can be seen as a combination of these two gates. Of course, they should be arranged in N-blocks and leave P-blocks as passing stages. In such an arrangement, the AND gate with two N-transistors in series will be critical. If we do so, the ultimate

TABLE I
SPEED COMPARISON OF SERIAL FULL ADDERS USING DIFFERENT CIRCUIT TECHNIQUES

Performance / Technique	Worst delay (ns)	No. of transistor	No. of input load	No. of clock load	Note
Static CMOS	6.2	38	8	6	[14]
NORA	6.0	32	4	10	Fig.16a
Modified NORA	3.8	35	3	16	[14]
DCVS NORA	3.6	38	6	10	[14]
TSPC Adder 1	3.3	42	4	12	Fig.17
TSPC Adder 2	1.8	50	3	13	Fig.18
TSPC Adder 3	1.5	54	4	16	Fig.19

delay will be around 1.2 ns in our simulation using the TSPC-2 circuit with a unit inverter load. A fast serial adder close to the speed limitation is presented in Fig. 19. First, we can see the two-input static AND and OR gates as extensions of the driving stages. The maximum delay variation of these two gates with different input combinations is less than 1.1 ns and this can be reduced by scaling. The delay variation must be less than half the clock period to guarantee that the inputs only change during the precharge phase. The two input gates can also be precharge circuits with more transistors but less input time constraints. Compared with normal, the inputs should precede the evaluation phase with an average delay time caused by the two gates. Second, the inverse carry signal offers both the input of next bit and the sum logic of present bit. Finally, in the critical path, the carry feedback path, two P-latches (three transistors each) are used in parallel for increasing driving capabilities. The worst delay of this adder is about 1.5 ns.

We summarize the above discussion in Table I, where we have also introduced the results from [14] but with our parameters. In [14], normal full adders, not serial full adders, are discussed. In Table I, we have converted them into serial full adders and resimulated. The worst delay is defined as the delay time at the 50-percent level in the critical path. The input load is calculated as the largest number of transistors connected to an input. The clock load is calculated as the number of transistors to which the clock driver is connected since the silicon area is not only related to the number of transistors of the circuit itself but also to the clock driver. The output loads of all these adders are the same, a unit inverter. Note that the domino technique, mentioned in [14], is not included here since this kind of technique means larger logic depth and, therefore, lower clock rate.

IV. DEVICE SIZING

The speed of CMOS circuits depends in a complex way on the sizes of all devices used, as the size of each device controls both its current capability and its capacitance [6]. Speed optimization by device sizing has therefore been discussed in several papers [15]–[19]. It is assumed that each device uses its smallest gate length (given by the process used), whereas its gate width is optimized for

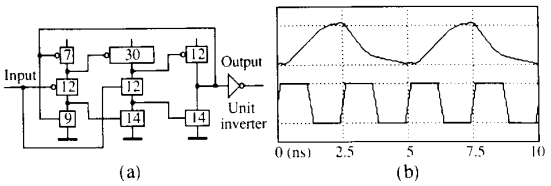


Fig. 20. Scaled eight-transistor divide-by-two circuit: (a) scaled circuit, and (b) input and output at 400 MHz.

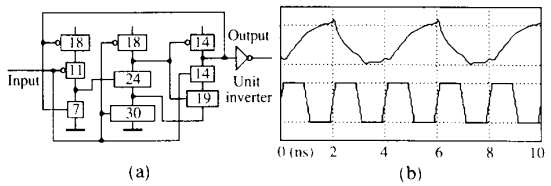


Fig. 21. Scaled nine-transistor divide-by-two circuit: (a) scaled circuit, and (b) input and output at 500 MHz.

speed. Often, the speed versus device size function does not show an optimum but is monotonic. In such cases cost (as used area) also must be taken into account. For simple cases it is possible to obtain analytical results. It is thus well known that an inverter chain driving a large load can be optimized by device sizing [15], so that each inverter has devices which are about three times larger than the previous inverter. It is also known that an optimal transistor chain connected to ground or supply should be tapered [17]. For more complex circuits it is not possible to obtain analytical results. The effect of the width of a certain transistor depends on the position of the transistor in the network and on the sizes of all other transistors in the network. Its width may affect different delays both through its effect on driving force and self-loading and through its effect on loading of the previous stages. Thus, in some cases we may have critical loops (as in the case of the carry in our serial adder), which means that the loop delay must be analyzed rather than a simple logic delay. In these cases improvements in speed can be obtained by optimization "by hand," by trying different device sizes in a circuit or timing simulator (like SPICE or TMODS [21]) [20]. A scaled version of the adder in Fig. 17 is obtained in this way, which is described in [20] and indicated in Fig. 17.

Recently, several computer tools have been developed for automatic optimization through device sizing [16], [18]–[20]. We have used SLOP [20] for this purpose. SLOP is based on a switch-level simulator, TMODS [21], and uses normal switch-level simulation for delay calculation. It can therefore easily handle any kind of CMOS circuit, including circuits with critical loops. The delay calculation algorithm in TMODS is based on the "Elmore delay" without side branches [22] and is similar to the algorithm in CRYSTAL [23].

With the help of the tool SLOP and the confirmation by SPICE simulation, after device sizing, the speed of the circuits described in earlier sections has been increased. Figs. 20 and 21 show the scaled versions of the divided-by-two circuits in Figs. 14(b) and 13(c) and the output waveforms with input signals with frequencies of 400 and 500 MHz, respectively, which are simulated with unit inverter loads. Note that all numbers in boxes are transistor widths in micrometers and that the divider in Fig. 21 has been changed to a type TSPC-2 connection. Since the unscaled version of Fig. 14(b) has an input capacitance equal to a unit inverter and an accepted clock rate of 250 MHz, it then can be put after these scaled version dividers and form ripple counters with the same working frequencies.

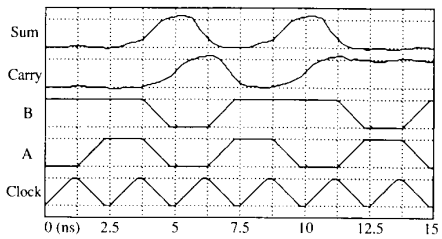


Fig. 22. Inputs and outputs of the scaled fast adder (Fig. 19) at a clock rate of 400 MHz.

TABLE II
COMPARISON OF SCALING

Performance / Circuit	Maximum clock rate or worst delay		Sum of widths		No. of input load		No. of clock load		Note
	Unscaled	Scaled	U.	S.	U.	S.	U.	S.	
Divider 1	250 MHz	400 MHz	56	110	2	3.4			Fig.20
Divider 2	250 MHz	400 MHz	63	112	4	9.6			Fig.13c
Divider 3	330 MHz	500 MHz	63	155	4	10.4			Fig.21
TSPC Adder 1	3.3 ns	2.5 ns	294	576	4	7.9	12	28.2	Fig.17
TSPC Adder 2	1.8 ns	1.3 ns	350	820	3	6.0	13	46	Fig.18
TSPC Adder 3	1.5 ns	1.1 ns	378	996	4	14.4	16	54.6	Fig.19

For the three adders described in the last section, the scaled sizes are already indicated in the corresponding figures. The worst delays have been reduced to 2.5, 1.3, and 1.1 ns, respectively. Fig. 22 shows the inputs and outputs of the scaled fast adder, TSPC adder 3, at a clock rate of 400 MHz. Note that the sum results from the inputs of the last clock cycle while the carry results from the inputs of the present clock cycle. In this simulation we have used input signals with large slopes (1 ns) to demonstrate the robustness of the circuit technique.

Table II is a comparison of these scaling results. As mentioned in the previous section, the ultimate delay for the pipelined circuit with a minimum logic depth is about 1.2 ns for an unscaled circuit. For a scaled circuit, generally, an improvement of a factor of 1.5 is expected so the ultimate delay will be around 0.8 ns. In principle, this is the maximum speed which can be reached in the 3-μm CMOS process with a 5-V power supply.

V. Experimental Results

In order to verify the above results we designed and fabricated several test circuits. At the moment, three of them are available for testing. The main techniques, i.e., the true single-phase-clock technique and the effective *D*-latch structure, have been proven experimentally. The

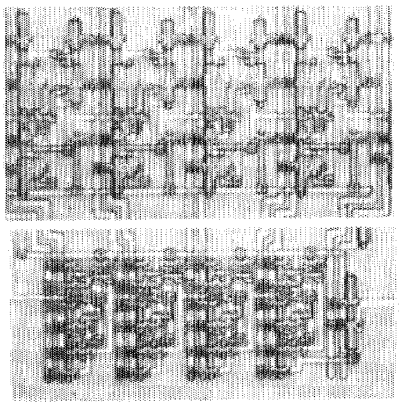


Fig. 23. Layout photograph of two ripple counters.

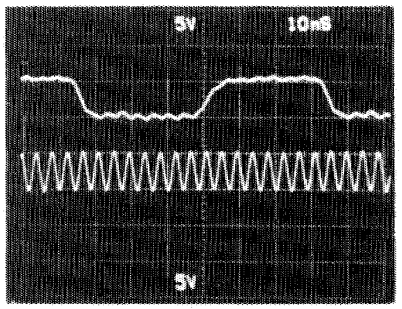


Fig. 24. Photograph of input and output waveforms.

first circuit is a 4-bit register which consists of four D latches (Fig. 13(a) plus an inverter) and the input data are taken from a divider. The chip area is 0.07×0.4 mm and the measured maximum clock rate is more than 230 MHz (limited by the instrumentation). The second circuit is a 4-bit ripple counter which is constructed by successively connecting four divide-by-two circuits as already shown in Fig. 13(c). The input frequency will be divided by 16 so we can measure the output at a lower frequency because it is found that the ordinary output stage and bonding pad are not fit for so high frequencies. Fig. 23 shows a photograph of the layouts of two such ripple counters: one is the unscaled version, chip area 0.28×0.14 mm, and the other is the scaled version, chip area 0.36×0.18 mm.

Since they are directly cascaded, the load for each circuit is more than a unit inverter and the resulting maximum working frequency is then 166 and 250 MHz for these two versions according to SPICE simulations. The measured results are 160 MHz for the unscaled version and more than 230 MHz (limited by the instrumentation) for the scaled version, which are the average results of ten samples from each. Fig. 24, a photograph taken from a sampling oscilloscope, shows the scaled divider successfully working at an input frequency of 230 MHz.

The third circuit is the serial full adder shown in Fig. 17. The chip area of the unscaled one is 0.22×0.16 mm and of the scaled one is 0.22×0.24 mm. The measured maximum clock rates are 140 and 225 MHz, respectively, which again are the averages from ten samples each.

VI. Conclusions

1) The true single-phase-clocking strategy has the advantages of simple and compact clock distribution, high speed, and logic design flexibility. In general terms, no clock-skew problem exists in this clocking strategy. The "skew" caused by clock delay between different logic blocks in a large system can be minimized by reverse clock distribution and the D-latch structure.

2) Since the maximum clock frequency in a clocked system is limited by the delay from one latching instant to the next, logic depth minimization in the system will lead to high speed with only a limited increase of the number of latch stages and initial delay.

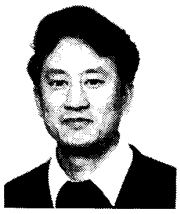
3) Device sizing will give a significant increase of speed at the cost of increased area. The experimental results from two examples show that the practical ratios between speed and area increases are 1.43/1.65 for a 4-bit ripple counter and 1.60/1.50 for a serial full adder.

4) Both analysis and experiment have demonstrated that clock rates in excess of 200 MHz are feasible in a 3-μm CMOS process by the combination of the above techniques. Some of the circuits possibly reach even the 400–500-MHz range as shown by SPICE simulations.

References

[1] C. Svensson, "VLSI physics," *Integration*, vol. 1, pp. 3–19, 1983.
[2] R. W. Keyes, "Physical limits in digital electronics," *Proc. IEEE*, vol. 63, p. 740, 1975.
[3] H. Masuda, M. Nakai, and M. Kubo, "Characteristics and limitation of scaled-down MOSFET's due to two-dimensional field effect," *IEEE Trans. Electron Devices*, vol. ED-26, pp. 980–986, 1979.
[4] C. Seitz, "System timing," in *Introduction to VLSI Systems*, C. Mead and L. Conway, Eds. Reading, MA: Addison-Wesley, 1980, ch. 7.
[5] Y. Suzuki, K. Odagawa, and T. Abe, "Clocked CMOS calculator circuitry," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 462–469, 1973.
[6] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*. Reading, MA: Addison-Wesley, 1985, ch. 5.
[7] M. Shoji, "Electrical design of BELLMAC-32A microprocessor," in *Proc. IEEE Int. Conf. Circuits Comput.*, 1982, pp. 112–115.
[8] N. 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, 1983.
[9] Y. Ji-ren, I. Karlsson, and C. Svensson, "A true single phase clock dynamic CMOS circuit technique," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 899–901, 1987.
[10] "Layout design rules for 3.0 μm P-well CMOS," VTI Technology Inc., San Jose, CA.
[11] I. Karlsson, "True single phase clock dynamic CMOS circuit technique," in *Proc. 1988 IEEE Int. Symp. Circuits Syst.*, vol. 1, pp. 475–478.
[12] C. Svensson, "Signal resynchronization in VLSI systems," *Integration*, vol. 4, pp. 75–80, 1986.
[13] G. De Micheli, "Performance-oriented synthesis of large-scale domino CMOS circuits," *IEEE Trans. Computer-Aided Des.*, vol. CAD-6, pp. 751–765, 1987.
[14] K. M. Chu and D. L. Pulfrey, "A comparison of CMOS circuit techniques: Differential cascode voltage switch logic versus conventional logic," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 528–532, 1987.
[15] E. T. Lewis, "Optimization of device area and overall delay for CMOS VLSI designs," *Proc. IEEE*, vol. 72, pp. 670–689, 1984.
[16] M. D. Matson and L. A. Glasser, "Macromodeling and optimization of digital MOS VLSI circuits," *IEEE Trans. Computer-Aided Des.*, vol. CAD-5, pp. 659–678, 1986.

[17] M. Shoji, "FET scaling in domino CMOS gates," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1067–1071, 1985.
[18] K. S. Hedlund, "Aesop: A tool for automated transistor sizing," in *Proc. 24th ACM/IEEE Design Automation Conf.*, 1987, paper 7.1, pp. 114–120.
[19] M. A. Cirit, "Transistor sizing in CMOS circuits," in *Proc. 24th ACM/IEEE Design Automation Conf.*, 1987, paper 7.2, pp. 121–124.
[20] J. Yuan and C. Svensson, "CMOS circuit speed optimization based on switch level simulation," in *Proc. 1988 IEEE Int. Symp. Circuits Syst.*, vol. 3, pp. 2109–2112.
[21] R. Sundblad and C. Svensson, "Fully dynamic switch level simulation of CMOS circuits," *IEEE Trans. Computer-Aided Des.*, vol. CAD-6, pp. 282–289, 1987.
[22] J. Rubinstein, P. Penfield, Jr., and M. A. Horowitz, "Signal delay in RC tree networks," *IEEE Trans. Computer-Aided Des.*, vol. CAD-2, pp. 202–210, 1983.
[23] J. K. Ousterhout, "A switch-level timing verifier for digital MOS VLSI," *IEEE Trans. Computer-Aided Des.*, vol. CAD-4, pp. 336–349, 1985.



Jiren Yuan was born in Shanghai, China, on November 17, 1940. He received the B.Sc. degree in radio engineering from the Polytechnical Institute of Harbin, China, in 1964. He received the M.Sc. degree in February 1988 from the LSI Design Center, Linköping University, Linköping, Sweden, where he worked on advanced MOS circuit technologies. His thesis topic treated high-speed CMOS techniques.

Before joining Linköping University he was employed at the Research Department of Huang He Machine Factory, Xian, China, and engaged in work on electronic system design. In May 1985 he started his research work at the Department of Physics and Measurement Technology, Linköping University.

Christer Svensson was born in Borås, Sweden, in 1941. He received the M.S. and Ph.D. degrees in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden, in 1965 and 1970, respectively.

He was with the Research Laboratory of Electronics at Chalmers University from 1965 to 1978, where he performed research on MOS transistors, MNOS memory transistors, Pd-MOS gas-sensitive transistors, and the physics and chemistry of the MOS system. During 1972 he was a Guest Scientist at the Jet Propulsion Laboratory, Pasadena, CA. Since 1978 he has been with the Department of Physics and Measurement Technology, Linköping University, Linköping, Sweden, first as a Lecturer and since 1983 as Professor of Electronic Devices. In Linköping he continued his research on MOS physics and chemical sensors but also started a new research group on integrated circuit design. He is now performing research on high-performance analog and digital CMOS circuit techniques, on timing verification and optimization of CMOS circuits, and on integrated electronic and sensor systems.