High-Speed Compact Circuits with CMOS

R. H. KRAMBECK, MEMBER, IEEE, CHARLES M. LEE, AND HUNG-FAI STEPHEN LAW, MEMBER, IEEE

Abstract—Characteristics of various CMOS and NMOS circuit techniques are described, along with the shortcomings of each. Then a new circuit type, the CMOS domino circuit, will be described. This involves the connection of dynamic CMOS gates in such a way that a single clock edge can be used to turn on all gates in the circuit at once. As a result, complex clocking schemes are not needed and the full inherent speed of the dynamic gate can be utilized. The circuit is most valuable where gates are complex and have high fan-out such as in arithmetic units. Examples are shown of the use of domino circuits in an 8-bit ALU, where simulations indicate a speed advantage of 1.5 to 2 over traditional circuits, and in a 32-bit ALU where a worst case add in 124 ns was projected and a time less than 100 ns was achieved.

I. INTRODUCTION

THIS paper will describe some new design techniques which can substantially reduce area and increase speed for circuits made with CMOS technology. These techniques combine, in a unique way, the speed and power advantages of dynamic circuits with the stability and ease of use of static circuits.

In a fully complementary CMOS circuit the logic function of each gate is implemented twice. For example, a combinational gate that does the AND/OR invert (AOI) function for one 3-input AND, and one 2-input AND (32 AOI), is shown in Fig. 1. The five n-channel transistors have all the information needed to implement the function and so do the five p-channel transistors. The advantage of having both arrays is that except for the very brief period when the output or the inputs are making transitions no current flows and no power is consumed.

The problem with this fully complementary approach is that for complex gates of the type shown in Fig. 1, substantial amounts of area can be wasted. For example, the same function could be made with six transistors in static NMOS or pseudo-NMOS as shown in Fig. 2. (Pseudo-NMOS refers to a design technique which gives circuits identical to NMOS circuits except for the use of a p-channel transistor as the load instead of an n-channel transistor.)

As a result of the extra area and extra transistors, the capacitive load on gates of a fully complementary circuit are considerably higher than the loads on a pseudo-NMOS or NMOS circuit. Each output goes to both a p-channel and an n-channel transistor in every gate it drives. P-channels are generally twice the size of n-channels to obtain more balanced rise and fall times [1]. As a result, the total gate load on each output will be three times higher. Parasitics do not increase that much but overall capacitance is at least a factor of two higher.

Manuscript received March 10, 1981; revised November 6, 1981.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

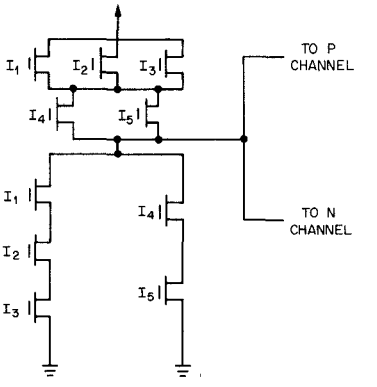

Fig. 1. Fully complementary MOS 32AOI gate. No static power but high-output capacitance and area.

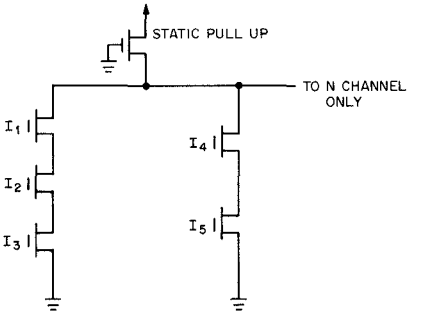

Fig. 2. Pseudo-NMOS 32AOI gate. Low output capacitance and area, but static pull-up current consumes power and slows pull-down.

It would appear from this that pseudo-NMOS or NMOS would be much faster than CMOS but this is not the case. The problem is that pull-up current always flows in the pseudo-NMOS circuit even if the gate is pulling down. This slows the pull-down. Making the pull-up current very small does not solve this problem because then the pull-up would be very slow. In fact minimization of the sum of rise time and fall time occurs when pull-up current is one half the pull-down. Thus, at most only one-half as much current is available in a pseudo-NMOS circuit as there is in a CMOS circuit using the same size transistors. In actual circuits the sum of rise and fall time is somewhat worse than this for pseudo-NMOS because for noise immunity the pull-up is usually chosen somewhat smaller than half the pull-down.

As a result, the speed of CMOS and pseudo-NMOS are very close. The CMOS has twice the capacitance but also twice the available current. The tradeoff in choosing one or the other is between the low power of the CMOS and the low area of the pseudo-NMOS.

The remainder of this paper will show first how dynamic circuits have combined both low-capacitance and high-current

0018-9200/82/0600-0614$00.75

capability, but at a cost in circuit stability and operational complexity. Next, new techniques will be described which maintain the above advantage of dynamic circuits while still keeping the stability and simplicity of static circuits. Finally, some specific examples will be presented.

II. Dynamic Circuits

Many dynamic circuit schemes have been described [2], but they all show some basic features in common. Basically, they involve precharging the output node to a particular level (usually high for NMOS), while the current path to the other level (ground for NMOS) is turned off. Changing of inputs to the gate must occur during this precharge phase. At the completion of precharge, the path to the high level is turned off by a clock and the path to ground is turned on. Then depending on the state of the inputs, the output will either float at the high level or will be pulled down. Fig. 3 illustrates how this is done for the 32 AOI gate described earlier. The advantage of a dynamic circuit is that the load capacitance is comparable to static pseudo-NMOS but the full pull-down current is available. Therefore, the gate should respond roughly twice as fast as either pseudo-NMOS or full CMOS. In addition, there is no static current path so power would be much closer to CMOS than to static pseudo-NMOS. (There is still some power penalty compared to CMOS because each gate must be precharged high every cycle even if its output is to continue low.)

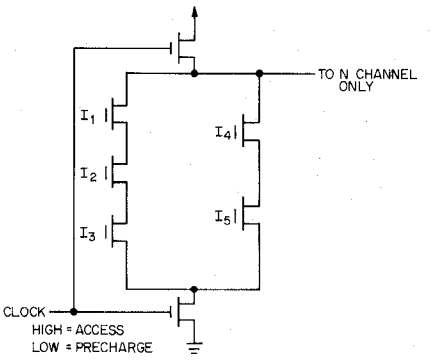

Fig. 3. Dynamic pseudo-NMOS gate. Low output capacitance and no static pull-up, but inputs must be valid before access begins.

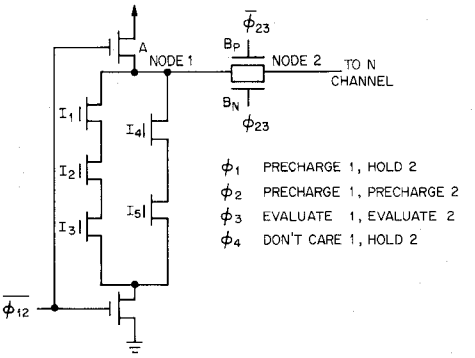

Fig. 4. Four phase dynamic pseudo-NMOS. The shortest clock phase must be long enough so that the slowest gate in the circuit can complete its evaluation. This results in considerable dead time.

However, there are serious problems involved in realizing these apparent speed advantages in real circuits. This happens because useful circuits generally have several logic gates in series and in the dynamic approach; no gate can be activated until its inputs have stabilized. There are many ways to clock the gates so that this occurs, and an example is shown in Fig. 4. A detailed description of the operation of this circuit is given in [2] and will not be repeated here. Basically, each gate goes through a precharge when transistors A and B are on, an evaluation when transistors B are on and A is off, and a hold period when transistors B are off. It is required that when a gate is in the evaluation mode, the gate driving it must be in the hold mode. There are four types of gates distinguished by the phase in which evaluation occurs. The one shown is type 3. This means that gate type 2 can drive either type 3 or type 4 but not type 1. Similar restrictions apply to each circuit type. This requires some additional care in design but is not a major problem. There are two reasons why the speed of this circuit will not be double that of a static circuit. First, each gate has two additional transistors in the pull-down path which reduces the available current considerably. For a 1- or 2-input gate this could easily be a factor of two. Second, the time allowed for a gate to stabilize must be chosen so that even the gate with the longest delay can settle down. This can cause substantial time waste on the faster gates because they must be allocated a full time slot. In addition, the difficulties of generating the four clocks and synchronizing them throughout the circuit to a small fraction of a gate delay are formidable. In practice considerably more than one gate delay would be needed between successive edges to assure a full gate delay in worst case. Overall then, in a circuit of reasonable complexity, the dynamic approach would not be any faster than static though it would have power advantages compared to pseudo-NMOS or NMOS.

III. CMOS Domino Circuit

The CMOS domino circuit shares some characteristics with dynamic circuits. In particular, each output is precharged high while the path to ground is opened and the precharge is stopped while the path to ground is activated. The critical difference is that the transition from precharge to evaluation is accomplished by means of a single clock edge applied simultaneously to all gates in the circuit. This greatly simplifies clocking and permits utilization of the full inherent speed of the gates.

A single domino circuit gate is shown in Fig. 5. It consists of two parts. The first looks like a dynamic pseudo-NMOS gate and is clocked in the same way as such a gate, with a precharge phase followed by an evaluation phase. The second part is a static CMOS buffer. Only the output of the static buffer is fed to other gates of the circuit; the output of the dynamic gate goes only to the buffer. During precharge, the dynamic gate has a high output so the buffer output is low. This means that during precharge, all circuit nodes which connect the output of one domino gate to the input of another are low, and therefore the transistors they drive are off. In addition, during evaluation a domino gate can make only a single transition, namely from a low to high. Because of the nature of the dynamic gate which drives it, it is impossible for the buffer to go from high to low during evaluation. (Since the dynamic gate cannot go high, the buffer cannot go low.) As a result there

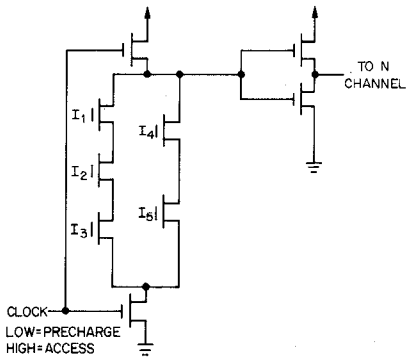

Fig. 5. Domino CMOS circuit. No static power, low area, with simple single edge clocking for all gates in the circuit.

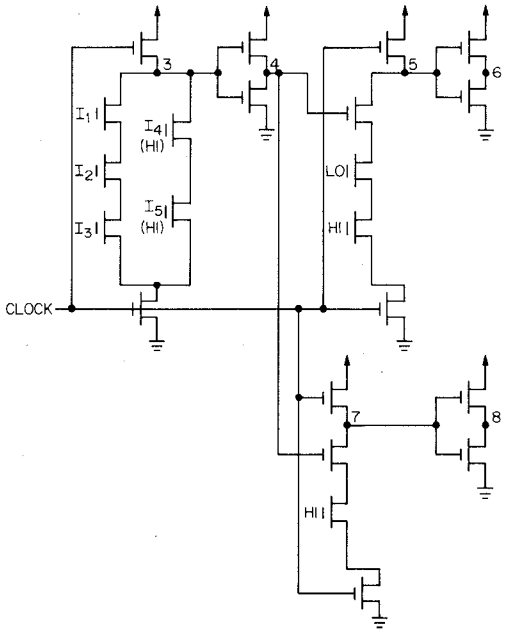

Fig. 6. An example of a domino CMOS circuit showing how a single clock activates all clocks simultaneously.

can be no glitches at any nodes in this circuit. All nodes can make at most only a single transition and then must stay there until the next precharge. This is reminiscent of the behavior of a row of dominos toppling into one another, and hence the proposed name.

Since there is no need to worry about glitches and since during precharge all domino outputs turn off the transistor they drive, all gates may be switched from precharge to evaluate with the same clock edge. An example of how this works is shown in Fig. 6. During precharge, nodes 3, 5, and 7 are all high so nodes 4, 6, and 8 are low. When precharge ends node 4 goes high which causes node 8 to go high. Node 6 remains low during evaluation.

As will be described in more detail in the next section many types of circuits when made with domino gates can be significantly faster than a corresponding circuit made with other techniques. The circuit has the low power of a dynamic circuit since there is never a dc path to ground. Also, the full pull-down current is available to drive the output nodes. At the same time the load capacitance is much smaller than for

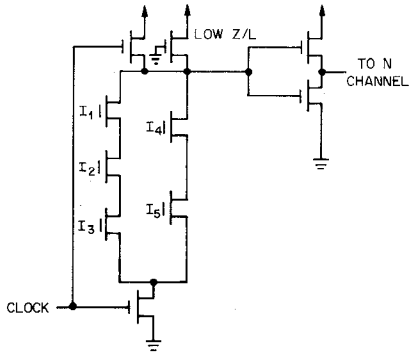

Fig. 7. Domino CMOS circuit with an additional pull-up device to permit static or low-frequency operation.

CMOS because most of the p-channel transistors have been eliminated from the load. Meanwhile, the use of a single clock edge to activate the circuit provides simple operation and full utilization of the speed of each gate. (There is no dead time between output valid and operation of the next gate in the circuit.)

One limitation of this circuit technique is that all of the gates are noninverting. This may seem serious since an XOR is not possible, but actually very complex circuits can be implemented including an arithmetic logic unit (ALU) with two levels of carry look ahead (to be described later). This is feasible because the domino gate is fully compatible with standard CMOS gates and the needed CMOS XOR can be driven by the last domino circuit.

Another limitation is that each gate must be buffered. This has not been a problem in the circuits designed so far because buffers would have been needed anyway to achieve maximum speed. The need for buffers indicates that this circuit technique is most valuable in logic involving many gates with high fan-out.

IV. Static Domino Circuit

In some applications it is desirable to have a static capability to allow lower frequency operation or to avoid the risk of storing data on floating nodes. This can be obtained in a domino circuit by the addition of a low current pull-up transistor as shown in Fig. 7. This functions as a means of removing charge which accumulates on the output node as a result of leakage or noise.

This transistor would be chosen small enough so there is no significant impact on pull-down current and so the power consumed during the evaluation phase is tolerable. A value of 10 μA is reasonable. This would require a p-channel transistor that is 20 μm long and 4 μm wide. For a chip with 2000 pull-up devices at 5 V, power consumption during evaluation would be 100 mW, if all gates are being pulled down. Average power would depend on the application but would be significantly less.

Another way to implement the static circuit is to include the static pull-up transistor shown in Fig. 7, but to have no clocked precharge transistor. This can be done if the time between evaluation phases is relatively long so precharge can be accomplished by the weak static pull-up transistor.

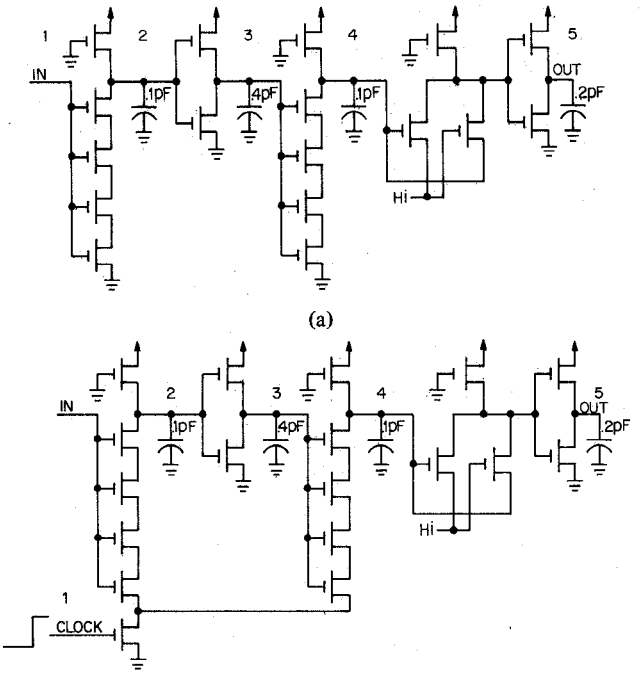

Fig. 8. (a) Part of an 8-bit ALU critical path using static pseudo-NMOS. (b) Same critical path with domino clocking.

V. An 8-Bit ALU

The first use of the domino circuit was on an 8-bit arithmetic logic unit (ALU) of an 8-bit microprocessor [3]. This happened because simulations indicated that adequate performance could not be obtained with a pseudo-NMOS circuit, while full CMOS was too area-consuming. The circuit of part of the critical path of ALU using pseudo-NMOS is shown in Fig. 8(a). The ALU in domino CMOS uses 690 transistors, and with a 15 μm pitch for metal and polysilicon, the area is 6000 mils^2. A similar transistor density in full CMOS would have required an additional 3000 mils^2 which was not available. A photograph of the ALU is shown in Fig. 9. The large structure on the lower left is the clocked ground switch which turns on the ALU when it is the evaluate phase. In a strip along the right side are the p-channel static load devices.

A SPICE [4] simulation of the simple pseudo-NMOS critical path predicted a worst case propagation delay of 450 ns which exceeded the chip requirement of 250 ns. A SPICE simulation of the domino circuit which is shown in Fig. 8(b) predicted 215 ns and so the design was made this way. Note that this circuit is like the one in Fig. 7 except that the clocked pull-up transistor has been eliminated and only the static one remains. This was done because the time between accesses of this ALU are so long that the low Z/L static transistor is sufficient to do the precharge. Table I shows propagation delays predicted by the simulation for both pseudo-NMOS and domino CMOS. The very slow pull-up time dominates the pseudo-NMOS CMOS delay. This happens because even though optimum speed is obtained with pull-up current equal to one-half pull-down current, noise margins forced a smaller ratio resulting in slow pull-ups. A histogram of measured propagation delay for 116 circuits that were fabricated is shown in Fig. 10. This histogram confirmed the high-speed predictions made by the simulation and verified the operation of the domino CMOS circuit.

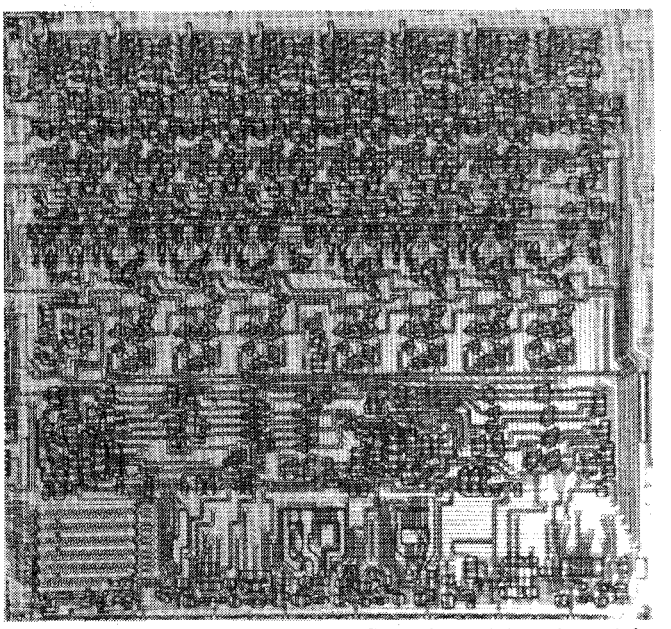

Fig. 9. Photograph of an 8-bit ALU. Ground switch is on lower left.

TABLE I
Worst Case Delays in 8-Bit ALU

	Static Circuit Delay (nSec)		Domino Circuit Delay (nSec)	
Node	In Goes High	In Goes Low	In High	In Low
1	0	0	0	0
2	40	270	100	0
3	10	10	25	0
4	40	80	25	0
5	50	50	65	0
TOTAL	140	410	215	0

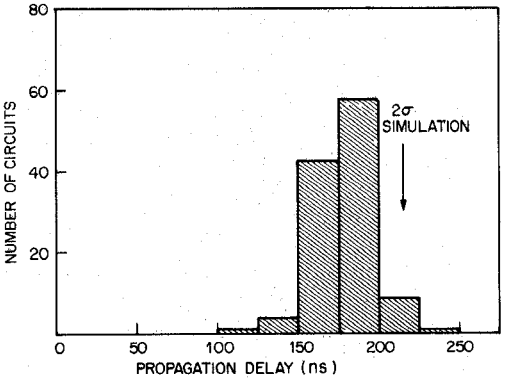

Fig. 10. Histogram showing distribution of delays for 116 ALU circuits.

VI. A 32-Bit ALU

For a more complex example, a critical path in a 32-bit ALU [5] will now be discussed. This circuit uses 3300 transistors and does a 32-bit add as well as other arithmetic and logic

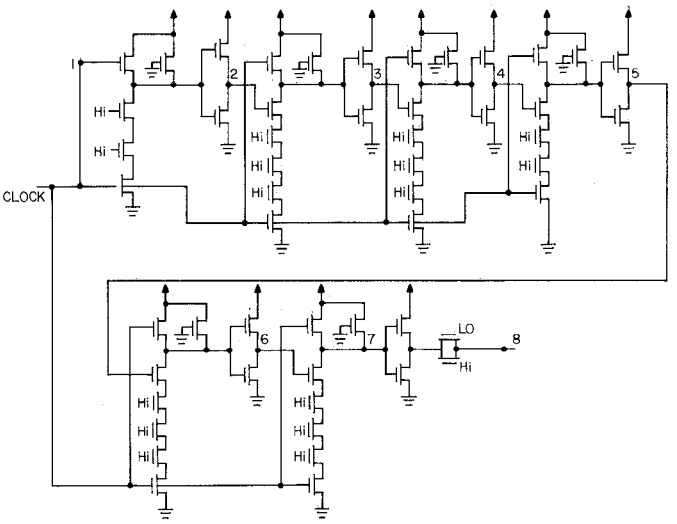

Fig. 11. Critical path through a 32-bit ALU.

TABLE II
WORST CASE DELAYS IN 32-BIT ALU

Node	Delay (nSec)
2	13
3	29
4	16
5	22
6	16
7	21
8	7
TOTAL	124

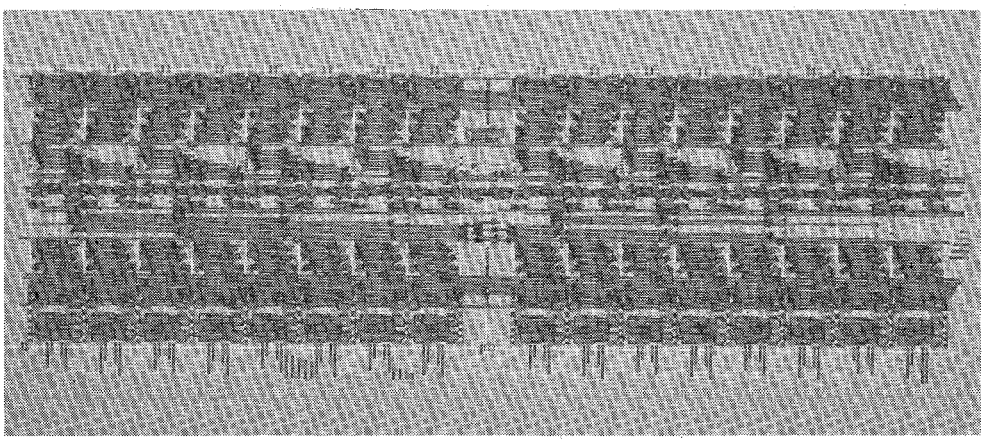

Fig. 12. Photograph of a 32-bit ALU.

function. The critical path in the domino CMOS path is shown in Fig. 11. Simulations have been made for this path and Table II gives propagation delays at various nodes on the critical path. The predicted worst case total propagation delay is 124 ns for $V_{DD} = 4.75$ V and a junction temperature of 105°C. This circuit was fabricated and a photograph of it is shown in Fig. 12. Process parameters of test transistors on the wafer were measured and using these a propagation delay of 104 ns was predicted. The actual delay was 97 ns.

VII. SUMMARY

A new compact, high-performance circuit design technique has been described for use with CMOS technology. This domino CMOS technique gives circuits with areas comparable to static NMOS or pseudo-NMOS, but gives a speed improvement of a factor of 1.5 to 2. This is achieved without resorting to any multiphase clocks and the static stability of the circuit can be maintained.

REFERENCES

[1] S. M. Kang, "A design of CMOS polycells for LSI circuits," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 838–843, Aug. 1981.
[2] W. M. Pensey and L. Lau, *MOS Integrated Circuits*. New York: Van Nostrand, 1972, pp. 260–282.
[3] J. A. Cooper, J. A. Copeland, R. H. Krambeck, D. C. Stanzione, and L. C. Thomas, "A CMOS microprocessor for telecommunications applications," in *Dig. ISSCC*, Feb. 1977.
[4] L. W. Nagel and D. O. Pederson, "Simulation program with integrated circuit emphasis," in *Proc. 16th Midwest Symp. Circuit Theory*," Waterloo, Ont., Canada, Apr. 1973.
[5] B. T. Murphy, R. Edwards, L. C. Thomas, and J. J. Molinelli, "A CMOS 32-bit single chip microprocessor," in *Dig. ISSCC*, Feb. 1981.

R. H. Krambeck (S'64–M'68) was born in New York, NY, on October 8, 1943. He received the B.E. degree in electrical engineering from City College of New York, New York, NY, in 1965 and the M.S. and Ph.D. degrees in electrical engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1966 and 1969, respectively.

He has been at Bell Laboratories, Murray Hill, NJ since 1968, first as a Member of the Technical Staff, and since 1977 as Supervisor of the High End Microprocessor Design Group. He has worked extensively on CMOS circuit design techniques. Most recently, he has worked on layout and design methodologies for VLSI microprocessors.

Charles M. Lee received the B.S. degree from National University, the M.S. degree from the University of Cincinnati, Cincinnati, OH, and the Ph.D. degree from the University of Michigan, Ann Arbor.

From 1968 to 1969 he worked for Texas Instruments, Inc., as a bipolar MSI circuit designer. Since 1973 he has been with Bell Laboratories, Murray Hill, NJ, working on microprocessor design. Currently, he is Supervisor of a microprocessor design group.

Hung-Fai Stephen Law (S'75–M'78) was born in Hong Kong in 1950. He received the B.S., M.S., M.Phil., and Ph.D. degrees, all in electrical engineering, from Columbia University, New York, NY, in 1973, 1975, 1977, and 1979, respectively.

In 1977 he joined the Technical Staff of Bell Laboratories, Murray Hill, NJ, where he worked on CMOS LSI circuit design, PLA design, VLSI circuit layout methodology, and the BELLMAC-32 single chip CMOS 32-bit microprocessor. He is currently a Technical Supervisor in the CMOS Integrated Circuit Design Department responsible for module circuit design in microprocessor.

Dr. Law is a member of the American Association for the Advancement of Science, the Association for Computing Machinery, Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

CCD Sampling of High-Frequency Broad-Band Signals

DAVID A. GRADL, MEMBER, IEEE

Abstract—**Several CCD signal sampling methods are discussed and a CCD input technique with excellent high-speed sampler characteristics is described. The method, a version of the diode-cutoff technique, is being used in a 200 MHz/8 bit transient digitizer system currently under development. DC based signal bandwidth (3 dB) of 600 to 800 MHz has been achieved along with random aperture uncertainty dispersion (one-sigma) of less than 2 ps. The sampler structure, operation, and experimental test results are described.**

Manuscript received March 23, 1981; revised September 5, 1981. This work was supported by the Los Alamos National Laboratory and EG & G, Incorporated.

The author is with Q-DOT, Inc., Des Plaines, IL 60018.

I. INTRODUCTION

SHORTLY after the advent of the peristaltic layered CCD and its high-speed operation [1]–[4] it was realized that correspondingly fast input and output techniques were needed to make full use of this high-speed processing element. For a class of applications including transient recording and signal bandwidth compression, only a fast input method is required [5] since output occurs at slow rates where conventional output techniques may be used.

At the CCD input, a voltage-to-charge conversion process normally occurs whereby a quantity of charge proportional to

0018-9200/82/0600-0619$00.75