Transistor-level Sizing and Timing Verification of Domino Circuits in the Power PC™ Microprocessor

A. Dharchoudhury, D. Blaauw, J. Norton, S. Pullela, J. Dunning+
High Performance Design Technology, Unified Design Systems Lab
+ Somerset Design Center
Motorola Inc., Austin TX 78750

Abstract

This paper describes a tool called Focus that is currently being used for the timing verification and sizing of domino CMOS circuits in a Power PC™ microprocessor. Domino CMOS circuits introduce more complex timing and sizing requirements compared to conventional static circuits. This paper shows how these requirements are addressed in Focus. Some case studies involving the application of Focus on production circuits are also described.

1. Introduction

The bulk of the implementation in high performance microprocessors has been typically done using static, fully complementary CMOS circuits. Only small, performance-critical portions of these chips have been designed using high-speed design techniques such as domino CMOS [1][2]. To achieve high performances not previously possible with static CMOS circuits, the latest offering from the Power PC™ family of microprocessors is extensively using domino CMOS circuits. Domino CMOS circuits can provide significantly higher speeds than static CMOS circuits, but they have much more complex transistor-level timing and sizing requirements. Due to the extensive use of domino circuits in the Powe PC™ chip, robust automated techniques for timing verification and transistor sizing of this type of circuits is a must.

This paper describes a tool called Focus that is being used for transistor-level timing verification and sizing of domino CMOS circuits in the latest PowerPC microprocessor. Focus has been used to optimize high performance static CMOS circuits in more than 15 design groups throughout Motorola spanning a wide variety of applications including microprocessors, microcontrollers, DSPs and ASICs. It has been used both for custom and library cell design. Domino circuits introduce several new requirements from the standpoint of timing analysis and transistor sizing that are not seen in typical combinational circuits. The main contribution of this paper is to describe the algorithms and techniques that we have adopted to solve these requirements.

2. Previous Work

Previous transistor sizing work has mainly addressed the requirements of static CMOS circuits. There are two broad approaches: a) dynamic approaches [3] using a circuit simulator in conjunction with a nonlinear optimizer, and b) static approaches [4][5] which use static timing analysis (TA) with the Elmore delay model [6]. The first approach is accurate but too slow and impractical for circuits with a large number of optimization variables. Moreover, input patterns must be specified. The second approach implicitly simulates all paths in the circuit and is fast, but the simple delay models are too inaccurate and lead to sub-optimal results. Elmore-based delay models are not applicable to high-performance domino-style design where timing accuracy is of the utmost importance. Domino CMOS circuits have much more complex timing and sizing requirements (discussed later in the paper) and most of these previous approaches are not suitable for domino CMOS circuits. Circuit sizing for domino CMOS has been addressed in [7][8], but these approaches are too simplistic. Recently, dynamic circuit timing verification was described in [9]. This approach consists of topologically identifying dynamic nodes, performing normal static TA, and then verifying the precharge and evaluate timing constraints as a post-processing step. The approach presented in this paper avoids topological rules which are restrictive to a particular design style.

3. Brief Overview of Focus

This section briefly describes the general timing analysis and sizing algorithms in Focus. Focus takes a user-specified circuit with parasitic information and timing constraints and generates a family of solutions which form a performance/area trade-off curve. The user is allowed to select the solution with the desired performance and area characteristics.

1.Focus uses a dc-connected component (DCC)-based static TA technique propagating the arrival times in a forward pass through the levelized circuit, and the required times in a backward pass. At each node, a slack is generated as the difference between the required and the arrival times (negative slack denotes a violation of a constraint).
2. The delay of a timing arc is calculated using a semi-analytical delay model which has been benchmarked to be two to three orders of magnitude faster than SPICE, with accuracies within 5% of SPICE. With this delay model, we have generated the area/delay trade-off curve for a 15,000 transistor circuit in less then an hour.
3.In addition to delay constraints placed at primary outputs, transition time constraints may also be specified. Transition time constraints give rise to so-called *slope slacks* which are propagated through the circuit in the same way as delay slacks.
4.Focus has additional TA capabilities such as false path avoidance, sequential circuit timing (including level-sensitive latches and gated clocks), enumeration of the top-few critical paths and SPICE verification of user-specified paths.
5.The variables in transistor size optimization in Focus are individual transistor widths or groups of transistor widths. A sizeable transistor T is selected for sizing based upon its sizing merit m which is defined as follows:

$$m(T) = \sum_{p \in P} W(p_{slack}, s_{min}) \cdot \frac{dp_{delay}}{dW_T} \quad (1)$$

1063-6404/97 $10.00 © 1997 IEEE

where P is the set of all timing arcs that are affected by the transistor T, p_{slack} is the slack of the arc p, s_{min} is the worst slack in the circuit (most negative) and $(dp_{delay})/(dW_T)$ is the raw sensitivity of the transistor width to the delay of arc p. The weighting function is defined as:

$$W(p_{slack}, s_{min}) = \frac{-1}{k + (p_{slack} - s_{min})} \qquad (2)$$

where k is a small positive number. The weighting function returns large negative numbers for critical arcs, and small negative numbers for non-critical arcs. The factor k determines the priority given to more critical paths over less critical paths. Note that the summation in Eq. (1) is over all arcs that are affected by a transistor (either through its drive strength or through its loading effect). Thus for a three input NAND gate, each transistor will affect six timing arcs. Moreover, these transistors will also affect the delays of fanin gates (through capacitive loading) as well as those of fanout gates (through slope dependence of delay). For single transistor sizing, the transistor with the largest merit value is chosen. For multiple transistor sizing, a more complex criterion based on the merits is used. If both long and short path constraints are present (as in sequential and domino circuits), Eq. (1) is modified to handle short paths.

4. Domino Structures and Constraints

4.1 Circuit Structure

A footed domino (FD) circuit is shown in Fig. 1(a). In the precharge phase, the clock CK is low and node X is precharged to Vdd. In the evaluate phase, CK turns high and X is conditionally discharged if the data inputs to the stage are such that a path to ground is established. The static inverter I1 is used to cascade domino stages and allows a single clock to precharge and evaluate all stages. During evaluation, the weak PMOS pull-up transistor MPf forces a high voltage level at X unless there is a strong pull-down path to ground. This replenishes any charge that might be lost due to leakage through the NMOS transistors. The weak NMOS transistors MNc1 and MNc2 charge the internal nodes of the evaluate tree to a high voltage in the precharge phase. This reduces the risk of erroneous logic values due to charge-sharing between X and the internal nodes in the evaluate tree during evaluation. For multiple-output domino logic, each output must have a weak PMOS pull-up, a static inverter and a PMOS precharge transistor connected to it.

In the footless domino (FLD) circuit of Fig. 1(b), node X is precharged when CK is low and is conditionally discharged when CK is high. Weak pull-up and anti-charge sharing devices function in the same manner as in footed stages. Short-circuit current may flow during the precharge phase since there is no ground switch -- this is avoided by imposing timing constraints on the data inputs. As a result, completely static data signals (those which have no timing relationship with the clocks) cannot be input to a footless stage. The inputs of a footless stage usually come from a previous footed or footless stage. A footless stage has one less transistor than the equivalent footed stage, and is typically faster. It has the added benefit of reducing the load on the clock signal. However, careful attention to the timing of the signals is required. There are other structural variations of domino CMOS stages, e.g. data-clocked and non-inverting preset stages; the treatment of these stages is similar to footed and footless stages and will not be described here.

A block diagram of a domino circuit and the associated clock waveforms are shown in Fig. 2. Each block is marked footed (FD) or footless (FLD) depending on the type of stages it contains. The first two blocks consist of footed stages clocked by the same clock signal CK0. The second and third stages are footless stages, clocked by CK1 and CK2, respectively. The rising edges of all three clocks coincide, but CK1 and CK2 have smaller precharge phases: these are called *delayed-reset* (DR) clocks. Note that if two footless blocks (or stages) are cascaded together, they must be triggered by staggered clocks as shown in Fig. 2. Two cascaded footed stages, however, can be triggered by the same clock. Moreover, as stated earlier, a footed block must be used to initiate a domino path.

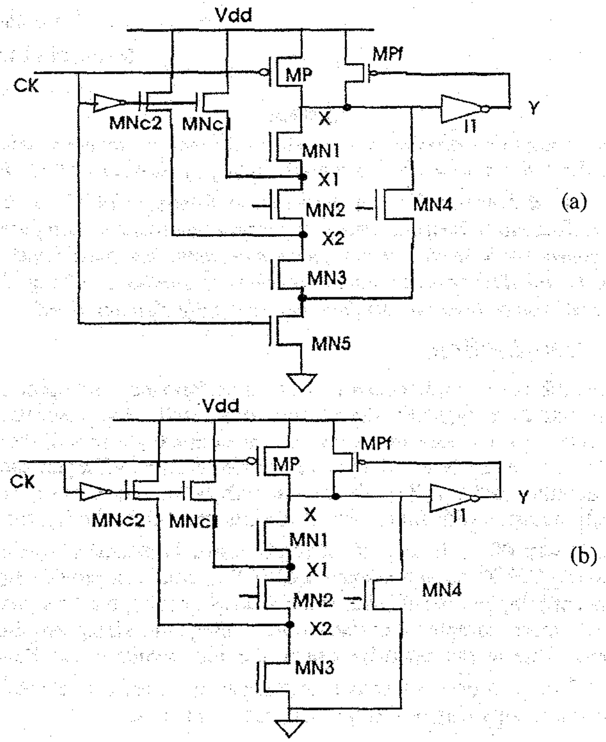


Fig. 1: Domino stages: (a) footed and (b) footless.

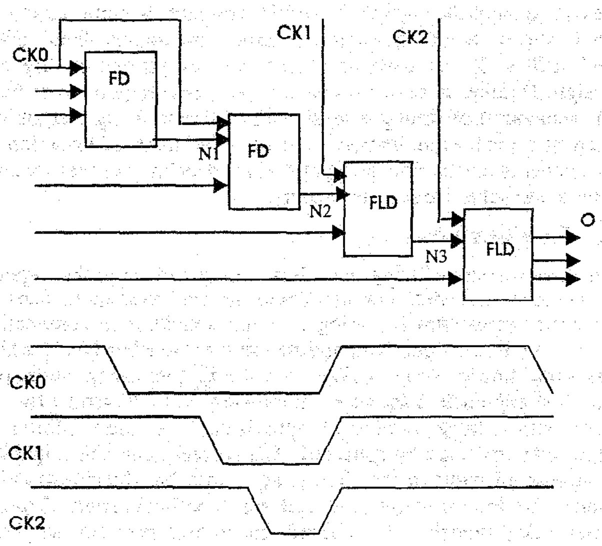


Fig. 2: Block diagram of domino CMOS circuit and clock waveforn

4.2 Node Classification

For typical latch-based combinational circuits, the primary objective of transistor sizing is to reduce the delay of the longest path in the circuit, and there are no constraints on the relative timing of signals internal to the circuit. For domino circuits, however, this is not true. In addition to the requirement that the primary outputs of the domino logic evaluate before a certain time, relative constraints between internal nodes in the circuit must be satisfied for error-free operation. For domino style circuits, it is extremely important that the circuit is sized such that these functional constraints are met. Whereas in standard CMOS circuits, a poor sizing may result in reduced circuit performance, in domino-style circuits, incorrect sizing may result in a non-functional circuit at any circuit performance. By imposing these relative constraints automatically, Focus ensures the timing correctness of the circuit in addition to providing a speed/area trade-off. In order to place these functional relative constraints, circuit nodes must be *classified* into different categories. In general, Focus classifies nodes into clock, data, dynamic (footed, footless, internal precharge), latch (level-sensitive and edge-triggered) and gated (or merged) clocks. This classification is done based on the type of signals (clock or data) that reaches a node from the primary input and whether that node has feedback to itself. In this paper, we will restrict ourselves to the recognition of dynamic nodes. A node is a dynamic footed node if (a) a clock signal can drive it to both high and low values, (b) a data signal can drive it to a low value only, and (c) it can be driven to a high value by a feedback signal. A node is classified as a dynamic footless node if (b) and (c) above are satisfied, and a clock signal can drive it to a high value only. A node is classified as an internal precharge node if it is not connected to the gate of any transistor and it is source-drain connected to V_{dd}. Once a node has been identified as a dynamic footed or footless node, the corresponding DCC is identified as a footed or footless stage. Once a stage has been classified, the functional relative timing constraints explained in the next section are injected by Focus automatically.

4.3 Timing Constraints

The most obvious timing constraints are that the primary outputs of the circuit evaluate within a specified time ("speed" constraints). For example, in Fig 2, we may require that O evaluate (rise) some time before the second falling edge of CK0. If O is an input to another domino circuit clocked by CK0, we must also constrain O to precharge (fall) before CK0 rises (Fig. 3(a)). Relative constraints between data inputs of a domino stage and the clock are explained below with reference to Fig. 3.

1. Consider N1 which is an input to a footed stage. N1 starts precharging (falling) when CK0 falls and must finish precharging before CK0 starts rising. Hence, for each data input of a footed stage, we impose a constraint that the data input node should fall completely before the clock to that stage starts rising (Fig. 3(b)).
2. Next, consider N2 which is an input to a footless stage clocked by CK1. To avoid short-circuit current, N2 should be low during the precharge phase. This is shown in Fig. 3(c), where N2 starts precharging after the falling edge of CK0, and should finish falling before CK1 starts falling. Note that this is a tighter constraint than the precharge constraint on N1. Similarly, in a footless stage the data inputs should transition high only after the clock of the stage turns high to start the evaluation phase. Since the rising edges of clocks CK0 and CK1 coincide, this condition will be met trivially. However, if the rising edge of CK1 is sometime after that of CK0, then this requirement places a *short-path constraint* on N2. The constraints on N3 are the same as N2 except that the precharge constraint is even tighter. N3 must finish precharging in the interval between CK1 falling and CK2 falling (Fig. 3(d)). Hence for each data input of a footless stage, we impose the following constraint: the data input should fall completely before the clock of that stage starts falling, and should start rising only after the clock to that stage has risen completely.
3. A relative constraint must also be added between the precharged internal nodes and the clocks of a domino stage. For example, X1 in Fig. 1(a) should finish precharging before the evaluation phase of the stage starts, i.e., before CK rises. Hence, for each internal node in a domino stage (footed or footless) that is being precharged, the following relative constraint is injected: the internal node should rise completely before the clock to that stage starts rising.

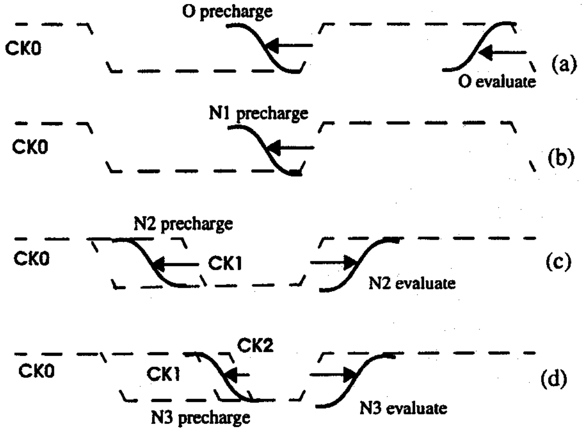


Fig. 3: Timing constraints: (a) output node, (b) FD input, (c) and (d) DR inputs.

In Focus, we distinguish between timing constraints that affect the functionality of the circuit (i.e., violations of these constraints will result in a non-functional circuit) and those that affect the performance of the circuit. For example, the precharge constraints above are functional constraints, whereas the evaluate constraint for the primary output is a performance constraint. Functional constraints are treated as *hard constraints* and performance constraints are treated as *soft constraints*. If a negative slack is generated from a hard constraint, then only hard slacks are propagated throughout the circuit, and Focus will size transistors to correct the hard constraint violations. If there are no hard constraint violations, transistors will be sized up to improve soft slacks (increasing the speed of the circuit). This guarantees that the family of solutions presented to the designer are all functional with different speed-area trade-offs.

Note that for a footed or footless stage, the relative constraints are applied at the inputs of the stage, and depend on the type of the stage itself and not on the type of the stage(s) that are driving it. Moreover, each of the relative constraints is between a data and a clock signal.

4.4 Sizing Constraints

In addition to the timing constraints described above, several sizing constraints must also be imposed if the resultant circuit is to be accepted by the designer. These are described below.

1. Layout effort may become unmanageable if each and every transistor in the circuit is sized by the optimizer. To keep layout effort

manageable, different instances of the same subcircuit in a large circuit might be grouped together. This ensures that the subcircuit example, suppose there are two instances of the domino stage shown in Fig. 1(a) in the first block of Fig 2. The user may specify that the transistor MN1 in one instance is grouped with the corresponding transistor in the second instance, and so on. This is a type of *topological grouping*. Another type of grouping may arise from *parametrization*. For example, the user may specify that in the stack of Fig. 1(b), the width of transistor MN1 is an independent variable and the widths of MN2 and MN3 are constant multiples (or fractions) of the width of MN1. This implies that the three transistors are grouped together, and during sizing, we must maintain the constant ratio between the sizes of MN2 and MN3 with MN1. In general, therefore, a group of transistors is characterized by a single width.

2.Purely delay-based sizing may produce gates with very large or small P/N ratios (unacceptable noise margins or unbalanced rise and fall times) and thus, constraints may be imposed on the P/N ratios of gates. For example, one may constrain the P/N ratio of all static inverters in the circuit to be within specified upper and lower bounds. To explain P/N ratio correction simply, consider that an NMOS transistor in a static CMOS gate has been sized. The new P/N ratio for the gate is calculated and checked against the lower bound. If the bound is violated, then PMOS transistor(s) in that gate are sized up to bring the P/N ratio back within bounds. For the sake of brevity, the calculation of P/N ratios for general CMOS gates and the selection of the transistors to be corrected for P/N ratio violations are not described here.

3.It is well known that when a stack of series connected transistors (e.g., in a NAND gate) is sized [7], optimal delay can be obtained by tapering the transistors such that the size of the transistors in the stack increases as one moves away from the output node. In the case of footed domino stages with the ground switch placed below the evaluation tree, this tapering may result in the clocked transistor becoming unreasonably large thereby substantially increasing the loading on the clock. To avoid this, tapering constraints may also be imposed on the transistors in the longest chain of transistors during sizing. In Fig. 1(a), the tapering constraints could be of the form (i) $a \le (MN2)/(MN1) \le b$, (ii)$a \le (MN3)/(MN2) \le b$, (iii) $a \le (MN5)/(MN3) \le b$, and (iv) $(MN5)/(MN4) \ge a$. Tapering constraints are maintained by a correction procedure similar to that used for P/N ratios. For example, if MN3 is sized up, it might violate the upper bound on (ii) leading to MN2 being sized up. Also, when MN2 is sized up, the upper bound on (i) may be violated, and MN1 may need to be sized up. If MN3 is sized up, the lower bound on (iii) may be violated and MN5 may need to be sized up. In this example, (iv) will not be violated and MN4 need not be corrected.

5. Examples

5.1 Circuit 1

This circuit contains 1800 transistors in a mixture of footed and footless stages. The area-delay properties of the original design is shown as the point marked "time_1" in Fig. 4. At the initial design point, the top few longest paths in the circuit are checked for correctness and detailed SPICE verifications are performed to get confidence in the Focus results. For sizing, the transistors are grouped into 280 distinct groups and the P/N ratios of all static inverters in the circuit are restricted to lie between 1.0 and 5.0. Further, tapering ratio constraints are placed on the transistor sizes in the evaluation stack of every domino stage. The values of a and b in this example are chosen to be 1.0 and 1.5, respectively (see Section 4.4). The ared by Focus is below and left of the original design (curve marked "tsize_1" in Fig. 4), which implies that significant improvement was achieved. The designer chose a point that was faster than the original design, but at the cost of slightly larger area (marked "current" in Fig. 4). SPICE simulations with vectors and SPICE verification of the top few precharge and evaluate paths were performed to check that the solution is valid. Fig. 5 shows the SPICE waveforms at the primary output of the longest path for the initial and final design points. At the initial design point, the primary output does not finish evaluating before CLK starts falling, i.e., the initial design did not meet speed. At the final design point, the primary output evaluates well before the required time. The percentage improvement in the evaluation speed was computed to be 26%. To demonstrate that Focus enforces timing constraints properly, we show the precharge waveforms at two internal nodes in the circuit before and after sizing in Fig. 6. Following the same notation as Fig. 2, nodes N1 and N2 are inputs to FD and FLD stages respectively. At the initial design point, N1 does not meet its precharge constraint and N2 barely meets its precharge constraint. At the final design point, both nodes meet their precharge constraints. The precharge portions of the circuit are sized just enough to meet the functional constraints exactly and not any more. Since the precharge portions are not oversized, the load on the evaluate portions is kept as small as possible.

5.2 Circuit 2

The second circuit contains 1376 transistors arranged in 339 distinct groups. This circuit is also a mixture of footed and footless stages. The area and slack of the initial design is shown in Fig. 7 as the point marked "time_1". The same P/N ratio and tapering constraints as in the first example are imposed during sizing. The sizing curve is marked "tsize_2" in Fig. 7. In this case, the sizing is stopped at the user-specified maximum area for the circuit. Fig 8 shows the waveforms at the outputs of the longest paths in the circuit for the initial and final circuits. CLK and CLKDR denote the original and delayed-reset (used to trigger the footless stages) clocks. The initial design is not functional since it does not precharge fast enough. At the final design point, Focus produces a functional circuit. The precharge waveforms are shown separately in Fig 9. To validate the final design, extensive timing verification was performed using Focus followed by dynamic SPICE simulation using user-specified input vectors and static electrical rule-checking. The final design was 50% smaller than the original design and almost 20% faster.

6. Conclusions

In this paper, we have described the application of a transistor-level timing and automatic sizing tool called Focus on industrial domino-style CMOS circuits. Specific timing and sizing issues that arise in domino circuits are highlighted and techniques to handle them are discussed. These issues include automatic injection of functional relative constraints between internal nodes in a circuit, sizing with tapering and P/N ratio constraints, and grouping of transistors. Application of Focus in the design and optimization of CMOS domino circuits in the latest PowerPC microprocessor has also been described

References

[1] R. Krambeck et al, "High-speed compact circuits with CMOS," IEEE JSSC, June 1982, pp. 614-618.

[2] I. Hwang et al, "Ultrafast compact 32-bit CMOS adders in multiple-output domino logic," IEEE JSSC, April1989, pp. 358-369.

[3] W. Nye et al, "DELIGHT.SPICE: An optimization-based system for the design of integrated circuits," IEEE TCAD, April 1988, pp. 501-519.

[4] J. Fishburn et al, "TILOS: A posynomial programming approach to transistor sizing," ICCAD 1985, pp. 326-328.

[5] S. Sapatnekar et al, "An exact solution to the transistor sizing problem for CMOS circuits using convex optimization," IEEE TCAD, Nov. 1993, pp. 1621-1634.

[6] W. C. Elmore, "The transient response of damped linear networks with particular regard to wideband amplifiers," J. Appl. Phys., Jan. 1948, pp. 55-63.

[7] M. Shoji, "FET scaling in domino CMOS gates," IEEE JSSC, Oct. 1985, pp. 1067-1071.

[8] S. M. Kang et al, "A global delay model for domino CMOS circuits with applications to transistor sizing," Int. J. Circuit Theory and Applications, 1990, pp. 289-306.

[9] K. Venkat et al, "Timing verification of dynamic circuits," IEEE JSSC, March 1996, pp. 452-455.

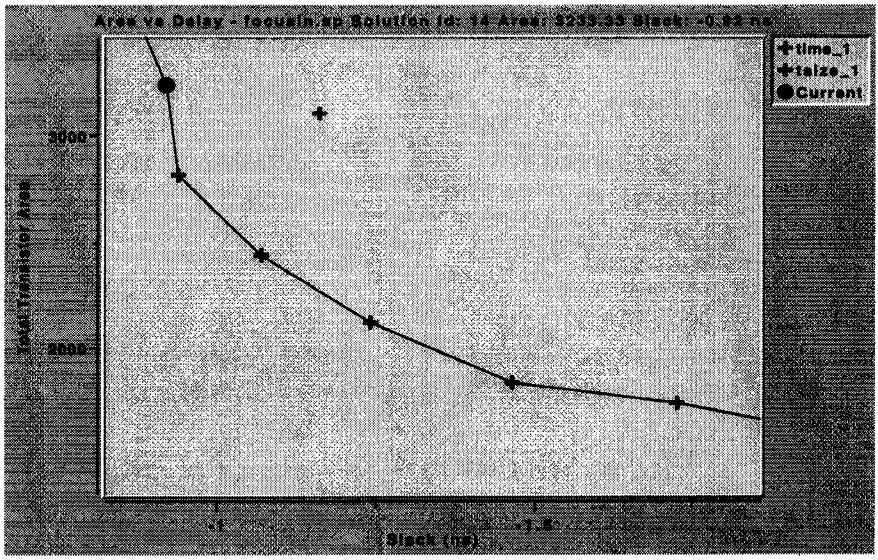
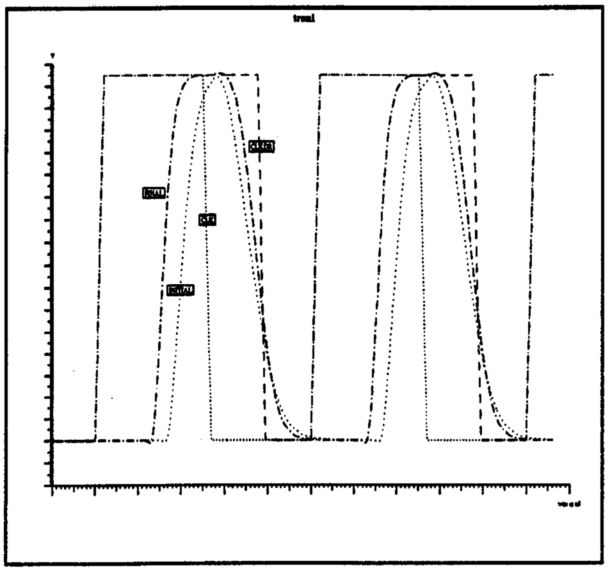


Fig 4: Focus sizing curve for circuit 1.



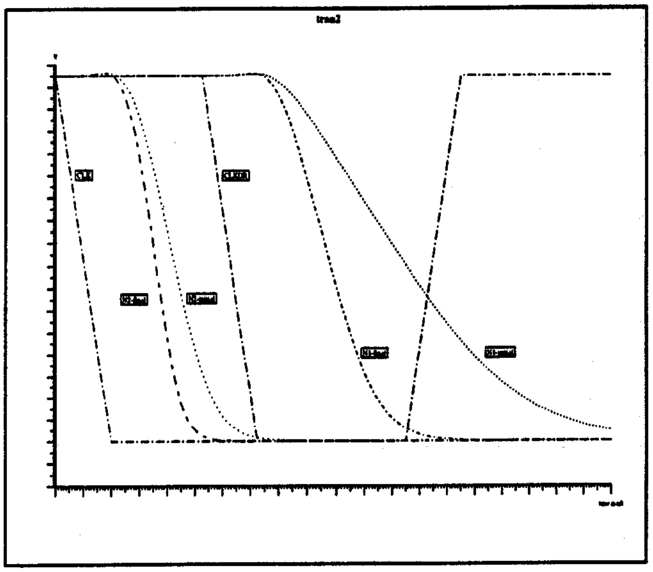


Fig 5: SPICE waveforms for longest path output at initial and final designs for circuit 1.

Fig. 6: SPICE precharge waveforms at inputs of footed and footless stages of circuit 1.

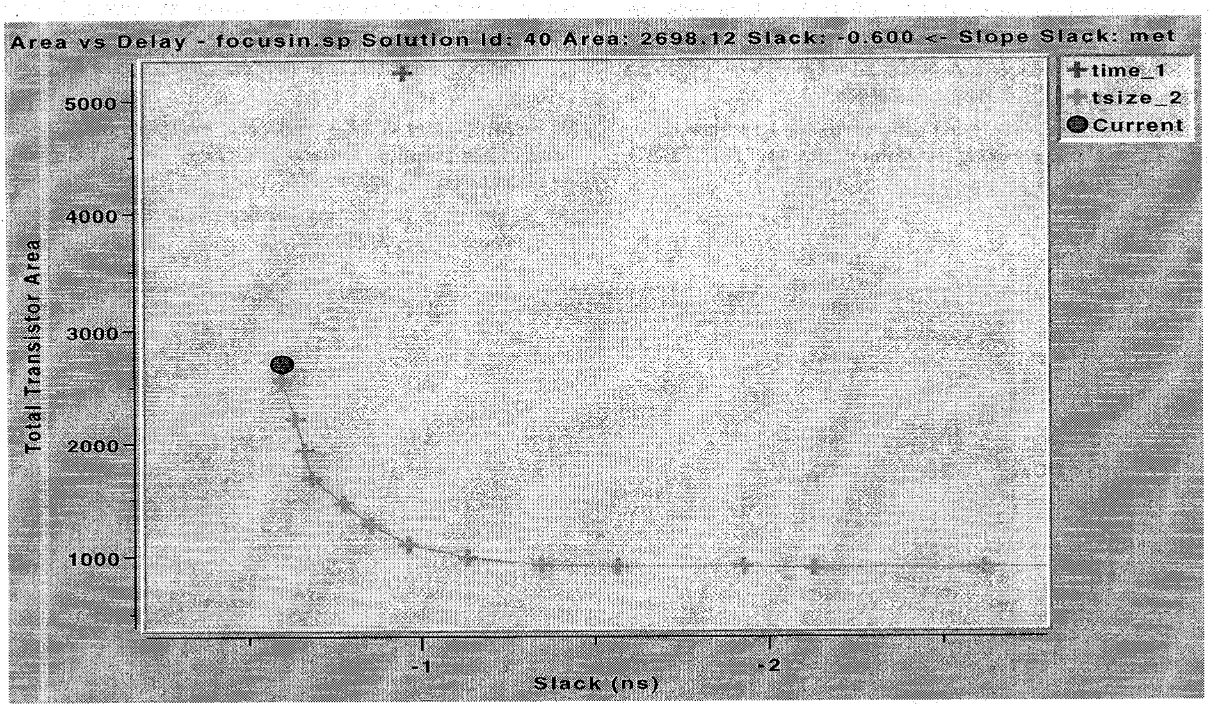


Fig 7: Focus sizing curve for circuit 2.

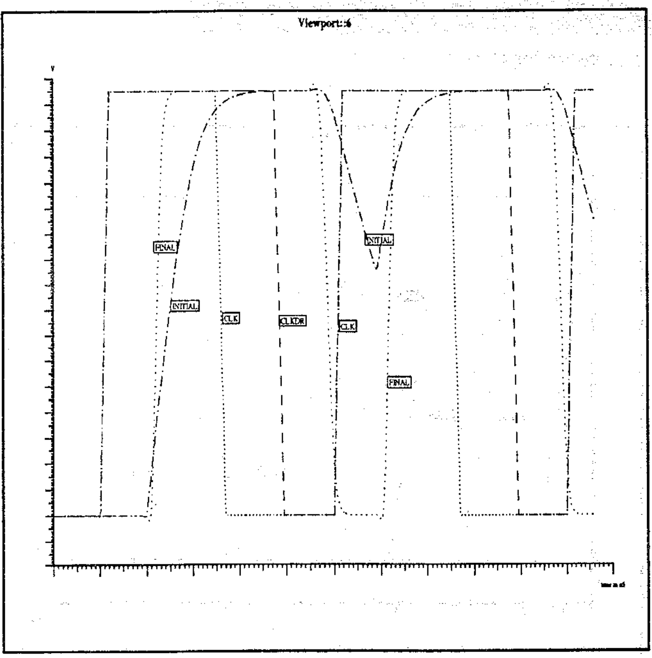


Fig 8: SPICE waveforms for longest paths at initial and final design points for circuit 2.

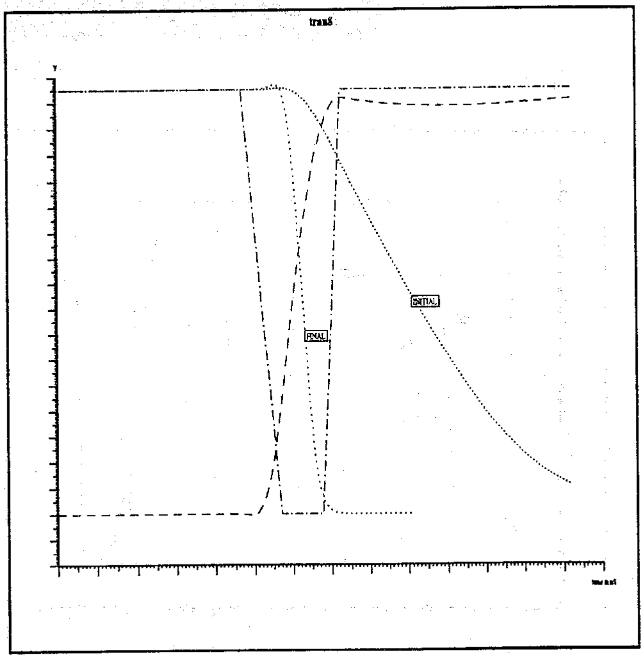


Fig. 9: SPICE precharge waveforms at inputs of footed and footless stages of circuit 2.