High Fan-In Circuit Design

Lawrence T. Clark, *Member, IEEE,* and Gregory F. Taylor, *Member, IEEE*

Abstract— **A review of high fan-in circuit design in contemporary logic technologies is presented. This is followed by a description of BiNMOS circuit structures which allow the construction of large fan-in, logical AND or OR functions. The first is a dynamic design, while the second is static. Application of the former in a 3.3 V, 100 MHz, implementation of the Pentium™ Microprocessor [1] on a 0.6 μm BiNMOS process [2] is described, while application of the latter in a 0.35 μm BiNMOS implementation is presented. Power and reliability considerations such as bipolar junction transistors (BJT), reverse V_{be}, and MOS hot electron protection are included.**

I. Introduction

MODERN microprocessors require the comparison of wide bit fields for functions such as cache tag comparison and zero detection. For instance, the Pentium™ microprocessor has six on-board caches, with associativities varying between two and four. Bipolar technologies allow wire-ORing which eases the construction of high fan-in circuits such as comparators. These are inherently faster than their precharged MOS counterparts due to the lesser load imparted by the bipolar junction transistors (BJT) emitter, as opposed to the MOS drain, and their greater current drive. In general, static circuits with a large number of inputs must dissipate static power in order to achieve high speed. Therefore, dynamic approaches are frequently employed in CMOS and BiCMOS. The problem is more straightforward in ECL, where static power dissipation is matter-of-course.

A number of issues must be addressed by dynamic circuits. Since the input to a dynamic CMOS circuit typically has a switching threshold of V_{TN}, they are noise sensitive. BiNMOS dynamic circuits also require careful attention to the discharging of the base. If an NMOS device is connected between the base and emitter of the BJT to discharge the base (see Fig. 1) and the base has been charged to V_{CC}, at process corners where $V_{BE} < V_{TN}$, the V_{GS} of the discharge device will be limited to V_{BE} until the base starts to discharge. This small V_{GS}, biasing the MOS transistor in subthreshold, requires most of the discharge current to be drawn through the base-emitter junction of the BJT, where it will be multiplied by the β of the device. This greatly increases the delay and power of the circuit. Finally, this configuration will add source capacitance to the emitter, thereby slowing the critical output of the entire circuit.

Furthermore, dynamic BiNMOS circuits have reliability considerations. BJT's require protection from reverse-biasing the base-emitter junction as excess reverse V_{BE} can cause increased emitter-base junction leakage currents and consequently, beta degradation. This degradation is a function of the reverse V_{BE} waveform shape, duty cycle, and amplitude; e.g., it is exacerbated by high V_{CC} [3]. Typically, protection is provided by a BJT connected between the emitter and base of the pull-up BJT as shown in Fig. 2. The protection device adds a small parasitic capacitance to the output node and, more importantly, the base node, contributing to the overall gate delay. This scheme is also ineffective for wired-or circuits, adding area and a dc current path.

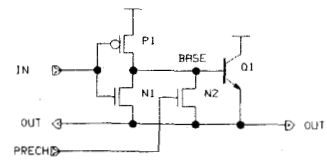


Fig. 1. Bipolar circuit with a NMOS transistor between the BJT base and emitter to discharge the base node.

II. CMOS

Although high fan-in gates are useful in a number of applications, they are not practical in a single stage of static CMOS. Since the NMOS and PMOS portions of a static CMOS gate are duals of one another, one will always contain transistors in series. Series transistors must be wider to achieve the same drive impedance as a single transistor, and thus have larger source and drain capacitance, limiting the speed of their output. These large transistors also increase the loading seen by the previous stage. When a large fan-in is required, one must then either use a series of static CMOS gates or look to another circuit design technique.

The most common high fan-in alternative is to use dynamic logic, such as domino (see Fig. 3). In domino logic the PMOS portion of a static CMOS gate is replaced with a precharge PMOS device, while a discharge control NMOS is added below the NMOS stack. When the clock is low **P0** precharges *OUT#* high which forces *OUT* low while **N0** prevents *OUT#* from being discharged. When the clock goes high, **P0** turns off and **N0** turns on, allowing the remaining NMOS devices to discharge *OUT#*. **P1** retains the state of *OUT#* when it is not intentionally discharged, avoiding discharge via weakly on pull-downs (subthreshold or near threshold operation), coupled noise, or alpha particle upset. While a cost of some additional delay and power during evaluation when these keeper and pull-down devices both draw current is incurred, it is important for designs which may stop the clock during the evaluation phase.

Manuscript received February 16, 1995; revised July 13, 1995.

L. T. Clark was with Intel Corporation, Hillsboro, OR 97124-6497 USA. He is now with Intel Corporation, Chandler, AZ 85226 USA.

G. F. Taylor is with Intel Corporation, Hillsboro, OR 97124-6497 USA.

Publisher Item Identifier S 0018-9200(95)00109-6.

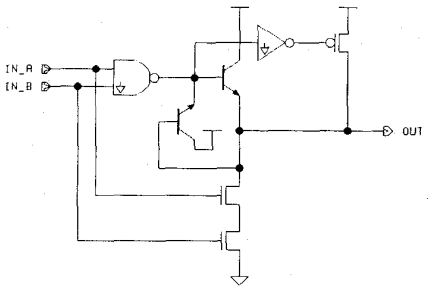


Fig. 2. BiNMOS NAND gate with reverse V_{BE} protection provided by a second BJT.

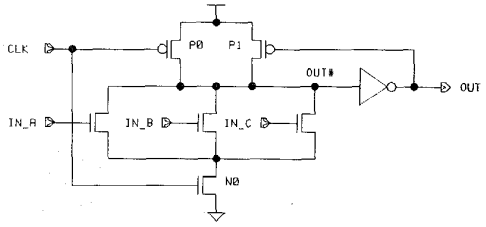


Fig. 3. Domino CMOS.

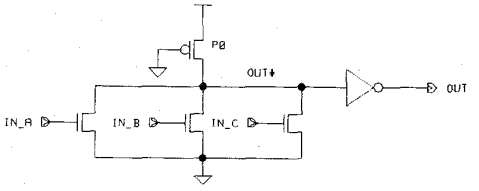


Fig. 4. Static pull-up example.

In order to make a fast, high fan-in, OR gate in domino logic, the NMOS pull-down stack contains a single parallel transistor for each input. If any input goes high, it immediately discharges *OUT#*, which then causes the output to rise.

Domino logic has a number of advantages. Like static CMOS, it dissipates no dc power. Because few PMOS devices are used, it has a smaller area and input capacitance than static CMOS, making it faster. If multiple stages of domino logic are cascaded, then the pull-down control, **N0**, can be eliminated from the second and later stages since the previous stages domino outputs will be forced low during precharge.

There are also two issues that must be considered when designing with domino gates. The input logic threshold is approximately V_{TN} of the NMOS input, making the circuit sensitive to coupled noise and power supply variations. This can be controlled by making sure that the final gate that drives a domino input is laid out close to the input that it drives. This limits the opportunity for noise to be coupled onto the signal while simplifying the task of making sure that the driver and the domino input share a common V_{SS} level. In addition, the inputs to a domino must be glitch-free, a requirement that can be difficult to meet when the inputs to a domino gate are driven by complex static logic.

If the inputs to a domino gate can not be made glitch-free, the precharge device may be replaced with a static pull-up as shown in Fig. 4, resulting in a "psuedo-nMOS" configuration [4]. This change causes the gate to draw dc current when *OUT#* is low, and also raises the low level of *OUT#* above V_{SS}. This can lead to reduced noise margin and leakage in the drive inverter of the gate, but it does tolerate glitches on its input. In any case, the design must trade off power and output fall time (dominated by the RC rise time on *OUT#*), effectively limiting the number of inputs that can be supported.

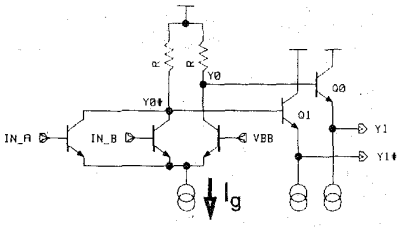


Fig. 5. ECL OR/NOR gate.

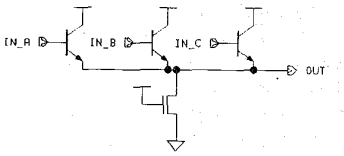


Fig. 6. BiCMOS wired OR.

III. ECL

Unlike static CMOS, the basic nature of an emitter-coupled logic (ECL) gate lends itself well to high fan-in (see Fig. 5). In an ECL OR/NOR gate, if any of the inputs are above the reference voltage V_{BB}, then the gate current I_g is steered through the associated input device [5]. This results in a voltage of $R * I_g$ on $Y0\#$, while $Y0$ rises to ground because essentially no current is drawn across its resistor. Emitter follower **Q0** pulls $Y1$ to a logic high level, while **Q1** allows $Y1\#$ to fall to a low level. The amenability of ECL to high fan-ins is demonstrated by the ability to construct 64 input gates in what is essentially a single logic level, using a combination of these techniques.

There are two ways to achieve high fan-in in ECL gates. First, more inputs can be added to the basic gate. This adds collector capacitance to the inverting output $Y0\#$, but does not load the $Y0$ output. Second, the emitter followers of multiple ECL gates can be tied together to create a wire OR structure. In this case only a single current source is needed to load the emitter followers.

The noise margin of ECL gates can be reduced by the improper use of these techniques. When all of the outputs driving a wire OR are low, they all share the output current. Because each now has a smaller current flowing through it, the V_{BE} of each emitter follower is reduced by

$$\frac{kT}{q} \times \ln N$$

where N is the number of emitter followers ORed together. If each gate contains a current source to maintain its output low

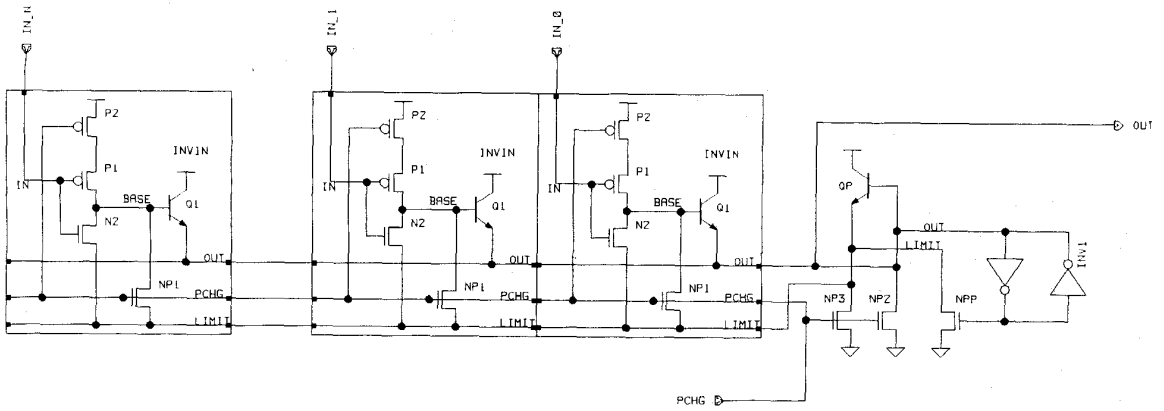


Fig. 7. High fan-in dynamic BiNMOS NAND gate.

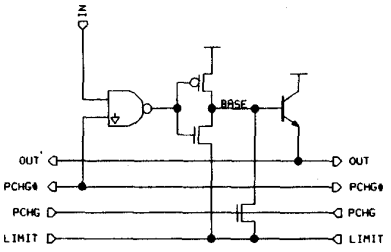


Fig. 8. Input stage for OR gate implementation.

level, then the output level when a single emitter follower is pulling up will be degraded by the same amount.

This effect is compounded because the differential pair in an ECL gate does not fully switch the gate current I_g. Instead, when all of the inputs are low, the portion of I_g that flows to $Y0\#$ through each input device is given by

$$\frac{A_{\text{in}}}{A_{\text{ref}}} \cdot \exp\left(\frac{V_{\text{in}} - V_{BB}}{(kT)/q}\right)$$

where A_{in} and A_{ref} are the emitter areas of the input and reference transistors. It can be shown that this results in the same noise margin degradation as the use of emitter followers and that the fan-in of the two effects can be added. In the case of the aforementioned 64-input gate, if eight outputs are wire ORed and then each is driven into an eight input OR gate, then the switching of the gate current of the OR gate degrades the input low noise margin by $kT/q \cdot \ln 64$ (135 mV at 100°C). Note that this noise margin degradation can be compensated, albeit with a small delay penalty, by increasing the swing of the driving gates by the same amount.

A second noise margin issue in wire OR's occurs if a portion of the wire from a single emitter follower driving a high level to an input is shared by the path from that emitter follower to the current source. In this case, the current will cause an IR drop across the resistance of the wire which will reduce the high level of the wire OR. This is particularly acute when the signal in question crosses a large distance, since both the current and the wire resistance are likely to be increased.

As is the case with dynamic CMOS, this situation is best dealt with by proper layout. If the current source is at the opposite end of the signal wire from the receiver(s), and the emitter followers are between them, then none of the IR drop caused by the current source will appear at the receivers.

IV. BiCMOS

There are several ways to take advantage of BiCMOS to create high fan-in gates. The simplest way is to replace the output inverter of a domino gate with a BiCMOS inverter. This has all of the advantages of CMOS domino logic, but benefits from the increased drive of the BiCMOS inverter.

A. Wire OR

Fig. 6 illustrates a faster gate that may be created by using a wire OR, as is done in ECL. If any input rises, the emitter follower connected to it will pull up the output. Because of the high transconductance of the driving emitter follower and the low output capacitance of the driven emitter followers, this circuit is quite fast. The resulting static circuit has been utilized in the tag compare logic in a BiCMOS RISC microprocessor [6]. On the other hand, it suffers from a number of disadvantages.

This circuit draws dc power when the output is high. In addition, its output high level is a V_{BE} below V_{CC}, increasing the leakage in the receiver. Both of these disadvantages can be eliminated by replacing the pull-down resistor with a predischarge/keeper network, the dual of the one used in domino CMOS. This makes the gate dynamic and consequently, the inputs must be guaranteed to not contain glitches. The sensitivity to input noise, however, is better than CMOS. If any input to a domino gate rises above V_{TN}, it may completely discharge the gate. In this gate, input noise smaller than V_{BE} is rejected, while noise above this threshold is coupled to the output without being amplified.

A potentially more significant disadvantage of this gate is the reverse bias caused by the inputs that are low. This can cause increased emitter base junction leakage currents and consequently beta degradation, as mentioned previously. Note that a protection device such as that employed in Fig. 1 is inappropriate. Thus, such a circuit can be reliably implemented in only limited circumstances.

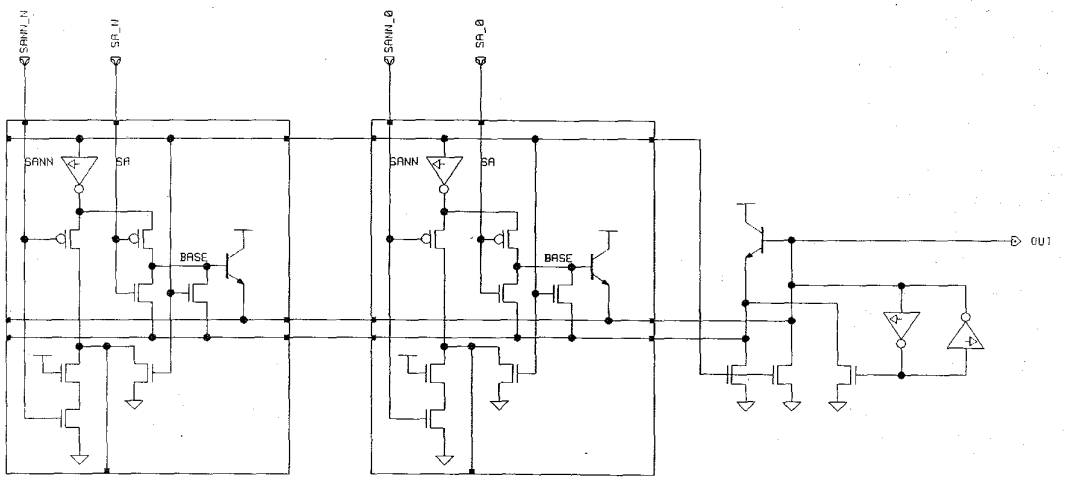


Fig. 9. Cache comparator circuit.

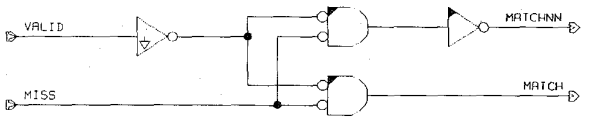


Fig. 10. Match signal generation logic.

B. A Dynamic BiNMOS High Fan-In Circuit

A circuit configuration comprising a high fan-in dynamic NAND gate is illustrated in Fig. 7. Inputs assert node *BASE*, turning on **Q1** which drives the *OUT* node high with little more than one CMOS inverter delay. This, in turn, turns on BJT **QP**, quickly raising the *LIMIT* node to $V_{CC} - 2V_{BE}$. Inverter **INV1** provides a full rail output high voltage, as well as protection from accidental assertion of the output node due to capacitive coupling. Pull-down **NPP** is provided to avoid accidental assertion of the *LIMIT* node. Power is conserved by not driving the *LIMIT* node full rail, as $V_{CC} - V_{BE}$ is sufficient to avoid device degradation due to reverse V_{BE}. After the inputs are de-asserted, NMOS pull-downs **NP1** and **NP2** are used to pre-discharge the *BASE* and *OUT* nodes, while **NP3** pre-discharges the *LIMIT* node.

This circuit addresses all of the considerations outlined above. The applied reverse V_{BE} peaks at about $2 * V_{BE(\text{active})}$, and then settles back to $V_{BE(\text{active})}$ as **INV1** drives the *OUT* node to V_{CC}. This limited reverse bias eliminates any device degradation concerns.

The precharge input to this circuit is the only node that operates with a degraded noise margin. The circuit driving the base of the BJT is a NOR gate and so has a normal threshold. Because the BJT is configured as an emitter follower, it does not have any voltage gain. This means that noise coupled on to the base of the BJT will not be amplified on the dynamic output node. Again, this is in contrast to dynamic MOS circuits, where the logic inputs to the dynamic gate have a V_{TN} threshold.

Connecting the NMOS device that discharges the base of the BJT to the *LIMIT* node instead of the output addresses the last two considerations. The source capacitance of these devices is connected to this noncritical node, reducing the load on the critical output. In addition, this allows the V_{GS} of this device to rise to $2 * V_{BE}$ before the BJT's **Q1** and **QP** start conducting current. Since the difference between V_{BE} and V_{TN} is generally small, doubling the applied V_{GS} dramatically improves the function of the circuit.

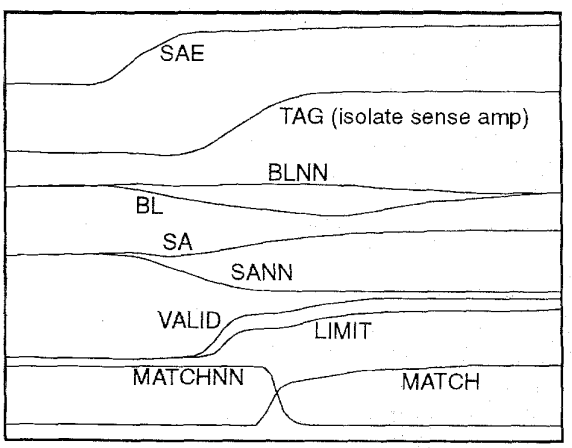


Fig. 11. Cache directory and comparator operation. $V_{dd} = 2.7$ V and important delays are listed in the text.

In the event that *LIMIT* is coupled above V_{CC}, the emitter base junction will not act to limit the excursion [7]. Consequently, P junctions are placed on the *LIMIT* node for this purpose, alleviating hot electron degradation risk to transistor **NP3**. This also avoids excessive delay in pre-discharging the circuit.

This circuit technique is also applicable to the construction of high fan-in OR gates. This circuit uses an input stage which comprises Fig. 8, and is a simple variation on the basic design. The greater delay imposed by two inversions is mitigated by their light loading. Referring to Fig. 8, a cycle begins by precharging the *OUT* node low. In the case where one of the inputs transition high, operation begins with *BASE* being driven high. Then, BJT **Q1** pulls up on the *OUT* node, which turns on **QP**, driving the *LIMIT* node to $V_{CC} - V_{BE}$ as discussed above. Bipolar transistors **Q1** and **QP** are in

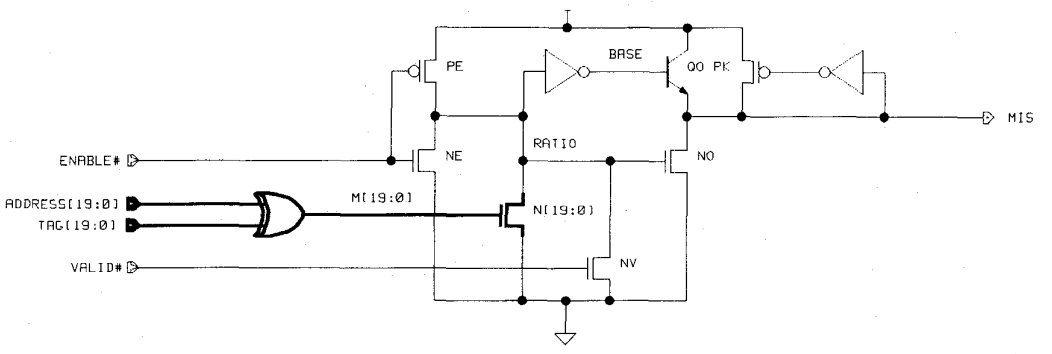


Fig. 12. High fan-in static BiNMOS comparator circuit.

a Darlington configuration providing a $2 * V_{BE}$ drop across **NP1** when *LIMIT* is predischarged. Consequently, NMOS transistor **NP1** will turn on at all process variations, which could allow $V_{TN} > V_{BE}$. This again ensures that the base is reliably discharged.

C. Dynamic Comparator

The circuit is used in the data and code translation lookaside buffer directories, branch target buffer, and segment descriptor caches of the 75 MHz and higher speed implementation of the Pentium™ Microprocessor. Here, the buffer stage, i.e., the inverter driving the BJT in Fig. 7, can be merged with the previous logic, here the sense amplifier output buffers (see Fig. 9). A dynamic cross-coupled inverter type sense amplifier is used to drive the signals SA and $SANN$, which are connected to the bit lines through a P transistor exclusive-OR network. This network makes the comparison of an input tag and bitline (stored tag) address, generating a logical one (high) level on SA if the tag and stored address match, and a logical zero otherwise. Noncomplimentary pass devices are easier to match, and are sufficient for passing the bitline signals, which are precharged to V_{CC} or $V_{CC} - V_{TN}$. The circuit is completed by the logic in Fig. 10 which asserts the match lines, selecting one input of a static multiplexor bank when the entry read is valid and no miss occurs.

The *VALID* signal is generated by a copy of the circuit of Fig. 9 where the SA input has not been XORed in the sense amplifier (accomplished by simply forcing the input tag value to a logical "0"). Thus, for valid entries, the corresponding MISS (*OUT* in Fig. 9) line is asserted and the *MATCH* output is enabled. The inverter delay on the *VALID* input assures that the output does not glitch for the worst case *MISS* input (fastest rise of *MISS*) of all inputs missing. This is important as the *MATCH* and *MATCHNN* nodes are control inputs to the way select multiplexors in the caches. They must be glitch free to avoid contention and subsequently, delay.

Simulation at the typical process corner, 2.7 V V_{CC}, and 120°C, yields an SA to *MISS* delay of 0.55 ns with a single input driving a 16 input compare (1.2 pF load) on a 0.6 μm process [8]. The waveforms are shown in Fig. 11. For a hit, the SA to *MATCH* delay is 1.20 ns and the SA to *MATCHNN* delay is 1.17 ns, driving 0.97 and 1.31 pF, respectively. The bitline XOR scheme also provides an easy means of isolating the regenerative sense amplifier from the bitlines, saving power and enhancing speed by not driving the bitlines to the rails, as is evident from the figure.

D. Static Comparator

The previous design, indeed any dynamic scheme, is inappropriate in cases where the data arrival time is unknown or difficult to detect. In this case, a static circuit is required. Additionally, since the inputs may transition in either direction, the EXOR outputs may transition in either direction and more than once, i.e., static logic hazards may exist. Consequently, the comparator stage must be capable of rapid transitions to either the hit or miss states. Krick *et al.* presented one such circuit in [8]. A much simpler 20 b comparator circuit used in a 0.35 μm, 3.3 V [9] implementation of the Pentium™ microprocessor is illustrated in Fig. 11. The kernel of the circuit is a high fan-in OR gate, implemented in two logic stages.

The high drive of the bipolar transistor and skewed logic threshold of the inverter driving the base help the output rising transition speed, while the low loading presented by the bipolar emitter helps the pull-down speed. The sizes of the P pull-up **PE** and N pull-downs driving node *RATIO* are optimized to minimize the rise and fall times of that node. Additionally, the low level of *RATIO* must be kept below the N channel V_T at all process corners to both maximize the gate V_{OH} and eliminate a potential dc current path through the output transistors. A worst-case delay of 0.85 ns from EXOR input to *MISS* is achieved while driving a load of 0.7 pf, at the typical process corner, 2.7 V V_{CC}, and 120°C.

A speed and size improvement is attained over either a series CMOS implementation as described in Section II or that in [8], primarily due to the relative sparsity of devices required by this "pseudo-NMOS" design. Other control signals may be brought directly into the gate as the *VALID#* signal is in Fig. 12. Vanishing static power dissipation is afforded by deasserting *ENABLE#* high, effectively disabling the gate when it is not in use. This is an important consideration for portable systems, particularly since high speed dictates that transistors **PE** and **N[19:0]** be relatively wide.

V. Conclusion

A brief overview of the design of high fan-in logic gates has been presented, including the drawbacks and pitfalls which

must be avoided. Such gates are important in microprocessor design and generally must be very fast, as they comprise part of speed critical paths. Bipolar technologies have been shown to be particularly amenable to the construction of circuits with large fan-in. Static and dynamic high speed BiNMOS circuits which are robust against process extremes and address important reliability concerns have been discussed. Their use in current microprocessors, as well as the numerous possible variations, demonstrate the practicality of the designs.

REFERENCES

[1] D. Alpert and D. Avnon, "Architecture of the Pentium™ microprocessor," *IEEE Micro,* June 1993.

[2] J. Schutz, "A 3.3 V 0.6 μm BiCMOS superscalar microprocessor," in *ISSCC Proc.,* 1994, pp. 202–203.

[3] S. Joshi *et al.,* "Poly emitter bipolar hot carrier effects in an advanced BiCMOS technology," in *Proc. IEDM,* 1987, pp. 182–185.

[4] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective.* Reading, MA: Addison-Wesley, 1985.

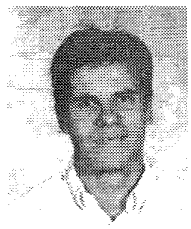
[5] A. Alverez, *BiCMOS Technology and Applications.* Boston, MA: Kluwer, 1989.

[6] F. Murabayashi *et al.,* "3.3-V BiCMOS circuit techniques for a 120-MHz RISC microprocessor," *IEEE J. Solid-State Circuits,* vol. 29, pp. 298–302, 1994.

[7] T. Fletcher, "BiNMOS design considerations for $0.3 < L_{\text{eff}} < 0.4$ microns," in *BCTM Proc.,* 1993, pp. 184–187.

[8] B. Krick *et al.,* "A 150 MHz 0.6 μm BiCMOS superscalar microprocessor," *IEEE J. Solid-State Circuits,* vol. 29, no. 12, pp. 1455–1463, Dec. 1994.

[9] M. Bohr *et al.,* "A high performance 0.35 μm logic technology for 3.3 and 2.5 V operation," in *Proc. IEDM,* 1995, pp. 241–244.



Lawrence T. Clark (M'90) was born in Detroit, MI. He received the B.S. degree in computer science from Northern Arizona University, Flagstaff, in 1983. He received the M.S. and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, in 1987 and 1992, respectively.

From 1984 to 1985 he was a Test Engineer at Intel Corp. From 1985 to 1990, he performed research in neural network implementation and ferroelectric memories as a graduate research associate in the Center for Solid-State Electronics Research at ASU. From 1990 to 1992 he was a Design Engineer at VLSI technology, where he worked on floppy controller and data separator designs. He was a Senior Design Engineer at Intel from 1992 to 1994 where he worked on two Pentium microprocessors. He was a senior CAD engineer in the Technology CAD organization from 1994 to 1996 and is presently a staff Design Engineer in the Technical Capabilities and Competencies group. His research interests include device modeling, high-speed circuit design, microprocessor architecture, and design automation.

Dr. Clark has received six patents.

Greg Taylor (M'86) received the B.S., M.S., and Ph.D. in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1981, 1983, and 1985, respectively.

He then joined Bipolar Integrated Technology where he worked on ECL floating point units and two RISC microprocessors. In 1991 he joined Intel Corporation, where he is a Senior Staff Design Engineer in the Portland Technology Development organization. While there, he has worked on four microprocessors and is currently managing a group responsible for clock distribution, PLL, and I/O for another microprocessor.

Dr. Taylor has received six patents.