PLA Driver Selection: An Analytic Approach

Fred W. Obermeier
Randy H. Katz

University of California, Berkeley
Electrical Engineering and Computer Sciences
Computer Science Division
573 Evans Hall
Berkeley, California 94720
415-642-8778

Research supported by:
California MICRO Fellowship
NSF ECS-8352227

PLA Driver Selection: An Analytic Approach

ABSTRACT

Few integrated circuit design tools support the rapid exploration of the design space across performance alternatives. Such a tool must rely on extensive analysis of fundamental design issues, to provide the basis for this exploration. These detailed analyses are summarized as simple "rules-of-thumb." A more flexible class of design tools can use these to generate integrated circuit designs with more desirable performance.

This paper describes the detailed analysis of the PLA driver selection problem. Rules-of-thumb which summerize these results, are developed and applied to a large CPU design project currently underway at Berkeley. A maximum critical driver delay improvement of 46.2% was realized.

Categories for submission:
(8) Design, Synthesis, Expert Systems
(5) IC Layout, Silicon Compilation

1. Introduction

The main performance metrics of a VLSI design are delay, power, and area. Although failure to achieve any one of these will cause a design to become infeasible, usually one measure dominates. The worst case propagation delay through a circuit dictates the maximum operation frequency. Total chip power dissipation tightly constrains the power limits of each subcircuit. Die size restricts the aspect ratio and area allocated to a particular logical unit.

The issues for improving the performance of an integrated circuit design naturally decompose into four classic problems:

(1). Driver Selection: Select a circuit to drive a given capacitive load with the desired electrical performance.

(2). Signal Buffering: Long signal wires exhibit classic transmission line effects. A waveform can degrade substantially in materials such as polysilicon. Buffering along a wire speeds the rising and falling transients, but additional delay is introduced. The issue is to decide at what minimum wire length is signal buffering beneficial and how often a buffer should be placed along the line.

(3). Layer Assignment: The choice of layers for an integrated circuit technology has a major effect on circuit performance. Usually low-resistance metal carries power, ground, and crucially fast signals. Higher in resistance and capacitance, polysilicon is relegated to carrying local signals. Luckily, layer assignment possibilities for regular circuits (Programmable Logic Arrays) are limited and can be analyzed exhaustively.

(4). Architectural Alternatives: Different architectural schemes can implement the same functional block. Some configurations are desirable for minimum delay (but may entail more devices) while others are appropriate for minimum power. Enough topological alternatives should be considered to offer a wide performance range.

The choice among the possible alternatives is summarized in terms of "rules-of-thumb". A "Rule-of-thumb" is a statement of simple conditions under which a design alternative dominates the other choices with respect to one or more performance metrics. For example, consider two register designs D1 and D2, D1 yields less delay than D2 when the load capacitance is less than some computed value, say "x". For loads greater than this, register D2 should be used to obtain

minimum delay. The essence of the rule-of-thumb for these two alternatives is to determine the value of "x" analytically, and to encode the choice of drivers based on the simple rule: "**IF** (load capacitance < x) **THEN** choose D1, **ELSE** choose D2."

Although automated analysis of a circuit takes considerable computer time, this can be done once to establish its merit against the other alternatives. A design system can apply the rules-of-thumb to satisfy a given set of design constraints quickly. To illustrate the analysis that underlies the formulation of rules-of-thumb, this paper focuses on the PLA driver selection problem.

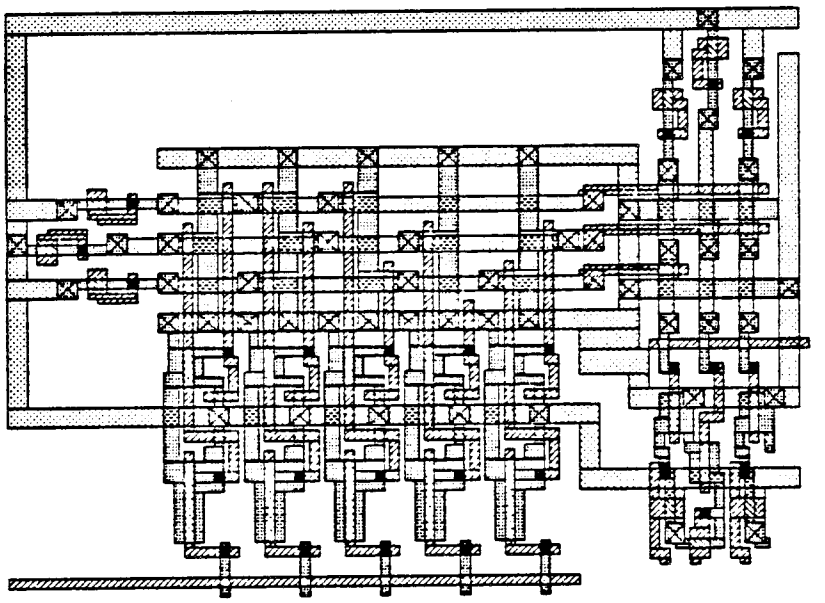


Figure 1: Example nMOS PLA.

A PLA is regular structure composed of two main functional blocks: combinational logic and data storage elements. An nMOS implementation is shown in figure 1. The combinational logic is composed of two level logic, the AND and OR planes. The data storage elements are contained within the PLA input and output drivers. Figure 2 shows how three of the four design problems arise in the design of a PLA. Driver selection specifies the best circuit to drive the complementary signal pair into the AND array. Signal buffering decides where to insert buffers into long signal lines. Layer assignment determines the best connection layers to route the signals for smallest area, least power, and maximum speed. The fourth problem, architectural alternatives, examines circuit topologies to provide better performance (e.g., precharged versus static pull-ups, pipelined structures, etc.).

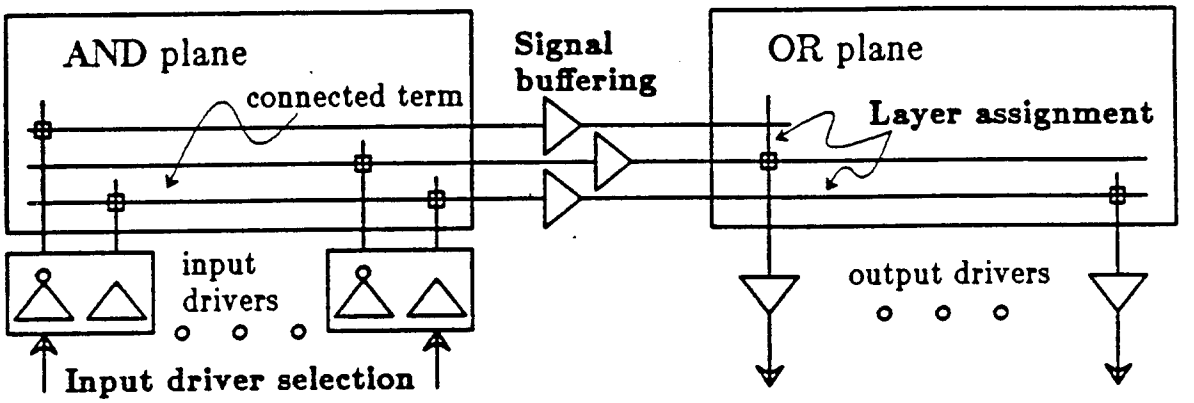


Figure 2: Typical PLA floor plan.

The rest of the paper is organized as follows. The PLA driver selection problem is described in the next section. Sections 3, 4, and 5 present derivations of typical rules-of-thumb for minimum delay, minimum power, and minimum area in the selection of nMOS PLA drivers. Performance implications for each of the optimization paths are examined. The rules-of-thumb for minimum delay (maximum speed) are applied to the PLA's of a current CPU project, SOAR (Smalltalk on a RISC) [Ungar 84], in section 6. Design analysis tools that underlie this study are described in section 7. The rules-of-thumb are summarized along with the performance improvement figures for the SOAR PLA's. The final section describes future development and applications.

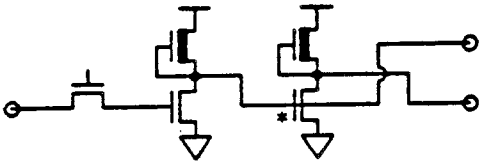
2. PLA Driver Selection:

The critical path of a PLA is of primary importance, since it identifies the portion that needs improvement the most. The critical path for the PLA's of figure 1 and 2 consists of delays through the input registers, AND and OR planes, and the output registers. Each of these can be decomposed into three types of delay: charge time for transistors (output load capacitance), propagation through logic, and propagation through wires. A paper by Rubinstein shows that the propagation delay due to transmission line effects, i.e., distributed resistance, is not a dominant factor even for large PLA's [Rubinstein 83]. The PLA driver selection problem addresses the delays caused by propagation through logic and the charge time for the output load capacitance. The other sources of delay are not considered in this paper.

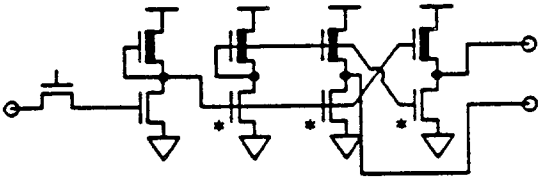
The term "driver" refers to alternative inverter and buffer configurations to drive a load capacitance. A few basic nMOS PLA driver alternatives are analyzed. The circuits in figure 3 are not exhaustive; rather, they are discrete points in the design space. They are used to illustrate how the analysis is carried out in order to derive the rules-of-thumb. New drivers can be analyzed and added to the repertoire in a straightforward way.

Generic drivers:

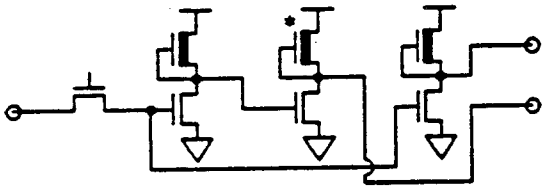
Two stage inverter chain



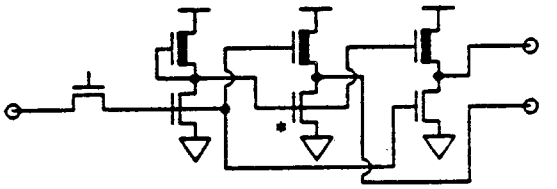
Lyon/Redford superbuffered chain †



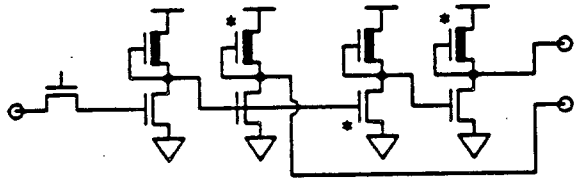
Three stage inverter chain



Three stage superbuffered chain



Four stage inverter chain



Four stage superbuffered chain

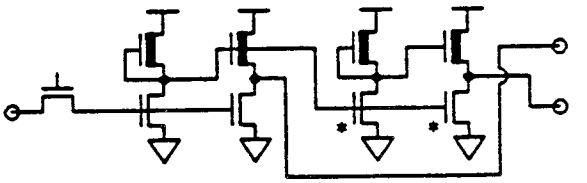


Figure 3: Circuit diagrams for dynamic generic gated PLA drivers.
Default device sizes: enh. pass gate: 2/2, enh.: 16/2, dep.: 2/2; (w/l in λ).
* enh.: 8/2 or dep.: 4/2;
† Composed of "PlaClockIn" and "AfterBurner" [Newkirk 83].

PLA buffer circuits are either static or dynamic. Static implementations use cross-coupled logic to hold the data until the next clock signal, while the dynamic configurations rely on charge storage to retain the data until the next clock signal. With dynamic circuits, a minimum frequency must be guaranteed or refresh circuitry must be included. Static implementations have no minimum frequency, but involve more circuitry and power than comparable dynamic circuits. This paper focuses on dynamic driver designs.

PLA Driver Selection: An Analytic Approach

Standard dynamic circuits use a pass gate to sample and then isolate the storage node from its input. Three configurations of gated dynamic circuits are examined: the generic driver, the boot-strapped driver, and the precharged driver.

The generic *gated drivers* are composed of traditional inverters ratioed to achieve the desired area, power, or speed goals. *Gated boot-strapped drivers* use capacitive coupling effects to reduce output switching speeds. The *dynamic precharged driver* employs similar device sizing; however, extra clock lines are added to preset certain voltages in the circuit. The circuitry that remains need only change the preset value when presented with the new input. Therefore, the slower of the two voltage swing directions can be eliminated at the expense of more circuitry and another control signal for the precharge clock.

Mead and Conway design rules specify appropriate device ratios to guarantee reasonable DC logic thresholds [Mead 80]. However, the investigation of the dynamic behavior of driver configurations requires AC transient analysis. Propagation delays through a circuit can be masked with non-step inputs since low logic thresholds may switch the output before the input has reached a comparable switching point. To obtain the worst case behavior and to eliminate logic threshold effects, a step-voltage is applied to all inputs.

The analysis proceeds as follows. Each circuit configuration is represented by an actual layout, and is then analyzed under several load conditions. The best configuration for each design measure (area, power, speed) is chosen and encoded by rules-of-thumb. The rules-of-thumb establish limits on the design space along the dimensions of minimum delay, minimum power dissipation, and minimum area.

A tremendous data reduction is realized through the simple selection criterion encoded by a rule-of-thumb. Megabytes of raw SPICE [Vladimirescu 81] data have been analyzed to produce the graphs in this paper automatically. The results of the graphs are then summarized by a few rules-of-thumb. A human designer or automatic tool can apply these rules-of-thumb to quickly design a PLA with desired performance. This will be illustrated in section 6.

3. Minimum Delay Driver Selection:

To standardize the comparison between circuit configurations, a minimum sized pass gate is used on the input. The effects of increasing its size will be examined later.

Two techniques can be used to decrease worst case propagation delay in circuits. One method is to increase the widths and decrease the lengths of the devices of the inverter. Although a greater capacitance is introduced on the previous stage, the increased capability for output current allows the dominant load capacitance to be driven faster.

Alternatively, several inverter stages appropriately scaled can be used to drive large load capacitance with minimum delay [Mead 80]. Since a paper by Nemes has indicated that this factor is not a constant, SPICE analysis of each circuit with several reasonable scaling factors will naturally select the correct factor [Nemes 84]. Adding stages decreases the total propagation delay for larger loads, while incurring a higher inherent delay at small loads. Therefore, as the load increases, configurations with more and more stages will be chosen for minimum delay.

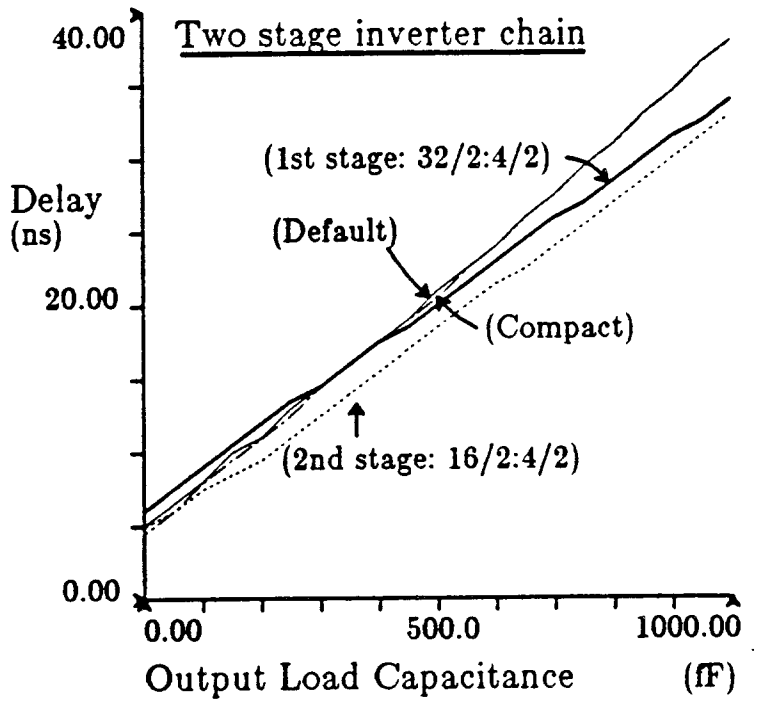


Figure 4: Worst case delay for various device sizes for the two stage inverter chain.

PLA Driver Selection: An Analytic Approach

To choose a minimum delay driver, the analysis is summarized as a graph of delay as a function of load capacitance. The x axis refers to the load capacitance to be driven by the complementary and non-complementary outputs of the PLA driver. The y axis indicates the propagation delay from the 50% point of the input step to the 50% point of the slowest output node. Therefore a graph of load capacitance versus delay establishes the worst case delay though the circuit driving the given capacitance.

Figure 4 shows how the device ratios were selected for minimum delay for the two stage inverter. Since there is little difference between the compact driver and one which takes more area (default), circuit performance is largely independent of layout cleverness. This result is not too surprising since internal node capacitance is dominated by output load capacitance.

For the two stage inverter driver, the poor performance of a larger first inverter stage indicates that one should reduce the capacitive loading on the pass gate input as much as possible. Marginally better performance is gained by increasing the ratio of the second inverter stage, showing that the second stage input is strictly limited by the high load capacitance output of the first stage. This suggests that faster circuits can be realized by driving high capacitance outputs only from low capacitance nodes. Adding another stage also eliminates the inherent problem of one high-load output signal following the other high-load output signal. This is the motivation for the design of the three stage inverter driver.

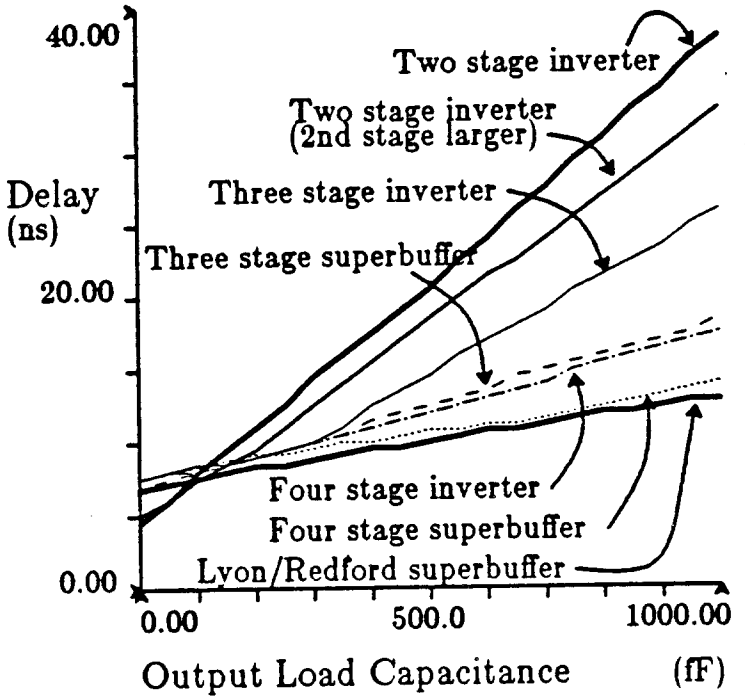


Figure 5: Worst case delay for all circuit configurations.

Figure 5 summarizes the worst case delay for all circuit configurations that have been analyzed. For small capacitive loads, a two stage inverter with the small default device sizes is the fastest†. As the load capacitance is increased to 50 fF, the two stage inverter with the larger second driver stage is faster. Intuitively, the load capacitance becomes large enough to justify increasing the size of the second stage inverter. Even though this process could be carried on to larger and larger staging, other configurations become desirable as seen in figure 5.

The three stage inverter has a slower response with small load capacitances but it has a faster response at higher load capacitance when compared to the two stage inverter chain. This effect is even more pronounced with the four stage inverter chain. Additional stages provide better isolation of low capacitance nodes (input) and offer better opportunity to scale up to drive larger capacitive loads. Enlarging the input pass gate would reduce the magnitude of this effect.

The three stage superbuffered configuration drives the outputs faster than the three stage inverter chain because of the actively driven depletion pull-ups.

† This is the configuration universally chosen by the Berkeley PLA tools [Mayo 83].

Similarly, the four stage superbuffered configuration beats the slower four stage inverter chain.

A variant of the four stage superbuffered driver, Lyon/Redford superbuffer, is composed of the PlaClockIn and AfterBurner cells from the VLSI Designer's Library [Newkirk 83]. Although it has superior speed attributes from among the alternatives examined, configurations with more stages will eventually drive larger load capacitances faster.

The minimum speed driver selection, for the load capacitance range examined, can be summarized by the following rules:

IF (0.0 fF $\leq$ load capacitance $<$ 50 fF)
THEN choose the default two stage inverter.

IF (50 fF $\leq$ load capacitance $<$ 108 fF)
THEN choose the two stage inverter with the larger second stage.

IF (108 fF $\leq$ load capacitance $<$ 1.0 pf)
THEN choose the Lyon/Redford superbuffered chain.

To fully appreciate the effects of increasing devices sizes, adding stages, and superbuffering, the power implications must also be examined. Since a power surge occurs during the transition between logic levels, DC power is not an appropriate measure of true power at higher frequencies. Average DC power is assumed to be the average of the DC "on" power and the DC "off" power. This assumption is valid for large circuits since the data value and its complement often appear together. This fact explains why the "on" and "off" power for multistage drivers equalize as the number of stages is increased.

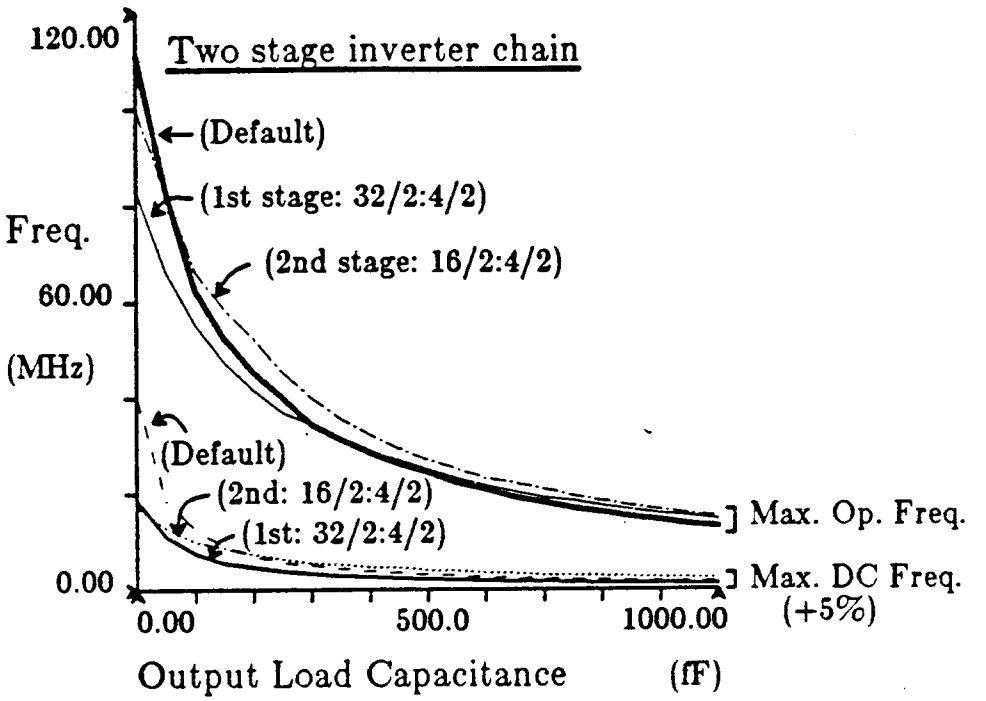


Figure 6: Power and output load capacitance effects on maximum operating frequency for the two stage inverter chain.

Loading effects on maximum operating frequency are substantial. The lower set of curves of figure 6 gives the frequency at which the true power is 5% above the average DC power for the two stage inverter. The upper set defines the maximum operating frequency as a function of load capacitance. The substantial area between the curves indicates that for most of the operating frequencies, the dynamic component of power consumption cannot be neglected. Although, the maximum delay defines the maximum operating frequency of the PLA driver, delay through the combinational logic block will reduce the system maximum operating frequency further.

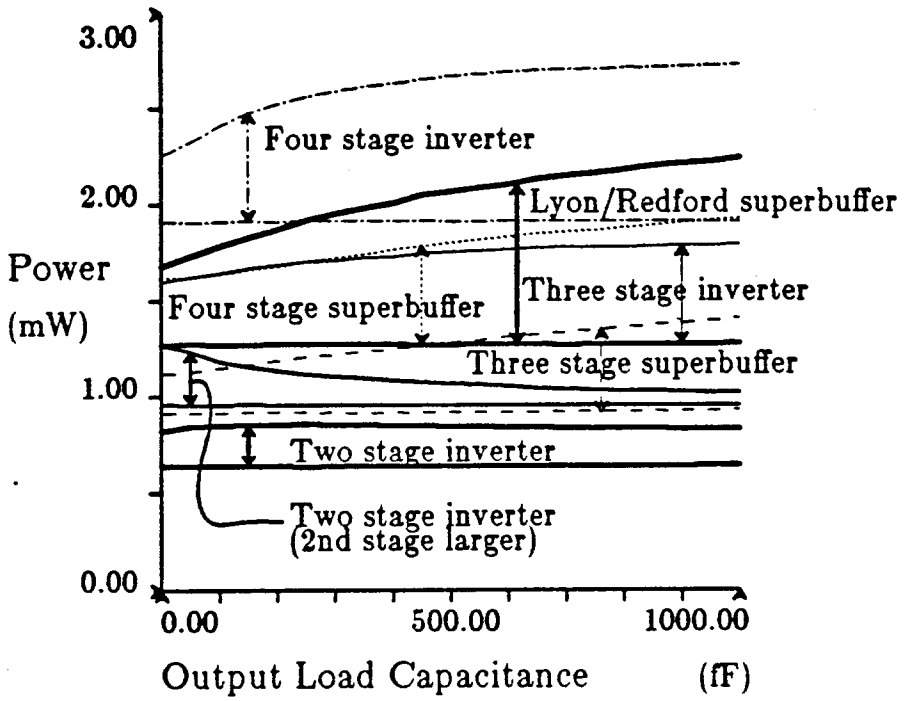


Figure 7: Operating range for all circuits.

The power at the maximum frequency (upper line) and the average DC power (lower line) restricts the operating region for the particular driver. As figure 7 shows, the two stage inverter chain with the default ratios draws the least power, even at its maximum operating frequency. The power draw for the two stage inverter chain with the larger output drivers and the three stage superbuffered driver are comparable. At lower load capacitance, the power expended by the larger two stage inverter chain is entirely below the lower bound on the power for the four stage superbuffered driver. This further justifies the selection of the larger two stage inverter chain for intermediate capacitance values for minimum delay, since it drives faster while consuming less power in a smaller area. (A desirable condition!)

At larger load capacitances, the fastest driver, Lyon/Redford superbuffer, also draws less power than the slower four stage inverter chain. Although this result indicates that the superbuffer technique is beneficial at larger loads, circuits involving boot-strapping should become desirable. The benefits of boot-strapping emerge at large loads ($>$1pf) since they incur a large overhead of charging an intermediate capacitance.

4. Minimum Power Driver Selection:

Since the lower bound on power dissipation is average DC power, the two stage inverter configuration with larger device lengths draws the least power. However, figure 5 has shown that these drivers are markedly slower. The speed/power tradeoff at various load capacitances for different sizings of the two stage inverter can be tabulated to provide a range of low power alternatives while indicating the speed penalty incurred.

Circuits using dynamic gated precharge are much better alternatives for low power applications since they can exhibit no DC power draw. Precharged circuits achieve this by avoiding DC paths between the supplies by proving non-overlapping control signals. Therefore, propagation delay can be reduced by using larger transistors with a small increase in power. The true power dissipation increases as the frequency increases (i.e. more charge moves during a shorter time period). Still, the power dissipation will be less than other dynamic circuits at lower frequencies.

5. Minimum Area Driver Selection:

Although it is clear that the two stage inverter chain has the smallest dimensions, often a bit more area is used to implement a much faster or lower power configuration. A simple check in a table will establish which of the circuits can be fit into the available space. Next, the best configuration is chosen to yield a driver to meet the performance constraints.

All drivers listed on the power and delay graphs were implemented by nMOS layouts using single metal, polysilicon, and diffusion layers. Since the area of the AND plane is dominant, the PLA input driver layouts were matched to the 16 horizontal pitch for a minimum AND core cell. Therefore, the input drivers could be expanded only in the vertical direction. The following table lists the driver sizes.

Layout Dimensions:	
Driver	Dimensions:
Two stage inverter chain	50 λ x 16 λ
Two stage inverter chain (larger second stage)	63 λ x 16 λ
Three stage inverter chain	79 λ x 16 λ
Four stage inverter chain	94 λ x 16 λ
Three stage superbuffer	78 λ x 16 λ
Four stage superbuffer	99 λ x 16 λ
Lyon/Redford superbuffer	124 λ x 16 λ

Even though the Lyon/Redford superbuffer is the fastest driver, the table shows that it requires the most area. A global area constraint could exclude this driver from consideration. Note that the four stage superbuffer requires 24% less area but runs almost as fast.

6. Application Example:

Our implementation of a PLA as seen previously in figure 2 consists of two logic stages segmented into an AND and an OR array. The input PLA drivers apply the input signal into the AND array. Therefore, the "output load capacitance" for each PLA input driver is the total capacitance of the wires and the transistors it drives in the AND array.

For the purposes of formulating rules-of-thumb, it is useful to scale load capacitance in terms of some "logical" parameter of the PLA, such as the number of product terms. The total capacitance is calculated by counting the number of connected (transistors) terms and the required number of unconnected terms. PLA input driver outputs only extend to the required connected terms. The unconnected terms can be reported completely as connected terms by appropriate scaling. In a typical nMOS PLA, each connected term introduces a 10.25 fF load and an unconnected term causes a 0.80 fF load. Therefore, the capacitance of 12.75 unconnected terms equals the capacitance of one connected term.

Given a PLA description specifying its personality matrix, the largest capacitance line can be determined quickly. In terms of effective connected product terms, this is one for which the sum of the connected and scaled unconnected terms

reaches the maximum. Note that in a dense PLA, the number of effective connected product terms is equal to the total number of product terms.

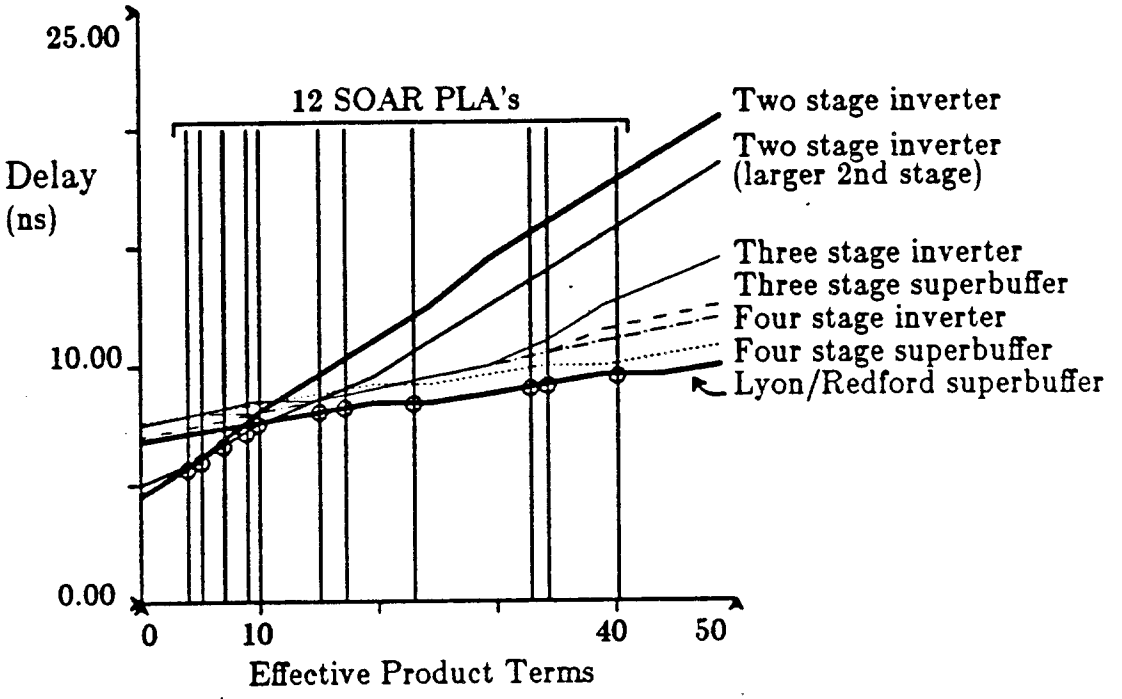


Figure 8: Placing the maximum input driver capacitances on the delay versus effective product terms (output load capacitance) graph for each of the SOAR PLA's.

As an example, consider the control portion of SOAR. For each PLA in the control unit, its maximum capacitance line was determined and placed as a vertical line on the minimum delay graph. This is a graphical approach to the application of the given rules-of-thumb. The two stage inverter chain establishes the base line performance since it is the universal choice of current Berkeley tools. Note that for several of the PLA implementations, the Lyon/Redford superbuffer provides a large speed increase over the default two stage inverter chain.

The largest performance increase, as shown in figure 8, provides a 46.2% improvement to reducing the delay through the input driver and AND plane. From the power graphs, the best improved PLA would require a power increase in the input drivers of 1.99 times for DC and 2.69 times that for the maximum operating frequency. This assumes that each driver is replaced by the faster configuration. A more accurate analysis would take into account that not all input drivers require the faster configuration, which would result in a lower power figure.

7. Tools Used By Automatic Analysis Programs:

All analysis was driven at a high level by an interpretive UNIX† shell program for maximum flexibility. The shell program called the appropriate utility programs and integrated circuit design tools. The underlying design tools require model parameters which describe a given technology. To adapt to a new technology, the automatic programs can be rerun with the new parameters and the tradeoffs appropriate for this technology will be revealed.

The layouts were constructed using a graphical integrated circuit editor, Magic [Ousterhout 84]. These layouts are referenced by a parameterized version of TPLA, a tile based PLA-generator, so that a PLA can be constructed from alternative input drivers [Mayo 83].

SPICE with typical nMOS parameters was used to simulate all drivers [Vladimirescu 80]. Minor changes in model parameters will cause absolute numeric changes, but the relative performance tradeoffs between the examined drivers should not change. SPICE analysis provides sufficient accuracy to adequately model dynamic power and dynamic voltage coupling effects. Five hundred points per output waveform was chosen to permit an accurate analysis, but examing and verifying this mountain of data is extremely tedious. A substantial C program was written to analyze and summerize SPICE output to create the graphs in this paper. For example, each circuit on the delay versus output load capacitance graph required about one hour of unloaded VAX 11/780 time and about one megabyte of disk storage. This program is expected to be distributed with the next Berkeley Tools release.

8. Conclusions:

This paper presented a general framework for performance optimization and has demonstrated its effectiveness in the selection of the input PLA register to achieve desired performance goals. Optimization along minimum delay, minimum power and minimum area paths were examined. Results of automatic programs define the precise numeric values for the rules-of-thumb within the examined range of output load capacitance. The 1.0 pf limit translates into 97 effective connected

† UNIX is a Trademark of Bell Laboratories.

product terms. The following rules-of-thumb were developed for Berkeley's nMOS PLA's:

Minimum delay:

IF (1 ≤ effective connected product terms ≤ 4)
THEN choose the default two stage inverter chain.

IF (5 ≤ effective connected product terms ≤ 10)
THEN choose the larger two stage inverter chain.

IF (11 ≤ effective connected product terms ≤ 97)
THEN choose the Lyon/Redford superbuffer.

Minimum power:

Choose the dynamic gated precharged driver.

Minimum area:

Choose the default two stage inverter chain.

These rules are applicable only when the other two constraints are not restrictive. Under conditions when certain drivers must be excluded because of excess power consumption, area, or insufficient speed, the graphs which give the relative performance of each configuration must be consulted to modify the rules-of-thumb.

The minimum delay rules-of-thumb were applied to the input PLA registers for SOAR yielding a maximum driver improvement figure of 46.2%. A worst case 2 to 2.7 times the power was required to realize this delay decrease. Also, the PLA must grow 13% taller to accommodate the extra 74λ required for the Lyon/Redford superbuffered drivers.

Although the analyzed driver circuits were restricted to PLA's, the results are applicable to random logic.

9. Future Development:

The remaining three problems are currently under study: signal buffering, layer assignment, and circuit alternatives.

An initial PLA synthesis system using these rules-of-thumb was constructed from UNIX shell scripts and a modified version of the tile based PLA generator, TPLA. A more sophisticated system will be constructed in two phases. A high level

system using constraint resolution techniques will reference the rules-of-thumb to generate a floor plan which meets the user's performance requirements. The second part will instantiate each subcircuit by referencing the required module generators or by loading a specific layout option. This approach integrates well within the framework of existing Berkeley tools, providing good state assignment, boolean minimization, and topological optimization.

10. References:

[Mayo 83] Robert N. Mayo, John K. Ousterhout, Walter S. Scott, *"1983 VLSI Tools: Selected Works by the Original Artists."* Department of Electrical Engineering and Computer Sciences, Computer Science Division, University of California, Berkeley, Ca., 94720, Report No. UCB/CSD 83/115, March 1983, Revised April 14 1983.

[Mayo 84] Robert N. Mayo, *"Combining Graphics and Procedures in a VLSI Layout Tool: The Tpack System."* Department of Electrical Engineering and Computer Sciences, Computer Science Division, University of California, Berkeley, Ca., 94720, Report No. UCB/CSD 84/166, January 1984.

[Mead 80] Carver Mead, Lynn Conway, *"Introduction to VLSI Systems."* Second printing, Addison Wesley Publishing Company, Inc., 1980.

[Nemes 84] Hihaly Nemes, *"Driving Large Capacitances in MOS LSI Systems."* IEEE Journal of Solid-State Circuits, Vol. SC-19, No. 1, February 1984.

[Newkirk 83] John Newkirk, Robert Mathews, *"The VLSI Designer's Library."* (nMOS, λ =2.5 microns, with Butting Contacts), Addison Wesly Publishing Company, Inc., 1983.

[Ousterhout 84] John K. Ousterhout, Gordon T. Hamachi, Robert N. Mayo, Walter S. Scott, and George Taylor, *"Magic: A VLSI Layout System."* 21st Design Automation Conference, 1984.

[Rubinstein 83] Jorge Rubinstein, Paul Penfield, Mark A. Horowitz, *"Signal Delay in RC Tree Networks."* IEEE Transactions on Computer-Aided Design, Vol. CAD-2, No. 3, July 1983.

[Ungar 84] David Ungar, Ricki Blau, Peter Foley, Dain Samples, David Patterson, *"Architecture of SOAR: Smalltalk on a RISC."* 11th Annual Symposium on Computer Architecture, Ann Arbor, Michigan, June 4-7, 1984.

[Vladimirescu 80] Andrei Valadimirescu, Sally Liu, *"The Simulation of MOS Integrated Circuits Using SPICE2."* Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Ca., 94720, Memorandum No. UCB/ERL M80/7, February 1980, Revised October 1980.

[Vladimirescu 81] A. Vladimirescu, Kaihe Zhang, A.R. Newton, D.O. Pederson, A. Sangiovanni-Vincentelli, *"SPICE Version 2G User's Guide."* Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Ca., 94720, August 10, 1981.