July 18, 1979

PLA Speed Calculations

Jim Cherry

Filed on: [ivy]<Cherry>PLAspeed.press, .bravo

Introduction

Ever wonder if your PLA can break the sound barrier ?

Here is a rule of thumb to see if it will.

$$t_{total} = t[5M + 4.3N + 4.6I + 4]$$

where	t_{total}	is the PLA propagation delay
	M	is the number of minterms
	N	is the number of outputs
	I	is the number of inputs
	t	is the process transit time, e.g. .5ns

This expression is independent of the value of l. The basic attitude taken in its derivation is one of extreme conservatism. This partially because incorporating the statistics of how one programs the PLA would greatly complicate the expression. The PLA is assumed to be built from the standard library cells in *A Guide to LSI Implementation* [Hon and Sequin].

Derivation

The basic formula used is the propagation delay for an inverter driving a capacitive node [Mead and Conway].

$$t_r = tkC_l/C_g$$

Where	t_r	is the rise time for the driven node
	t	is the process transit time

k is the inverter’s pull-up/pull-down ratio

C_l is the node capacitance that the inverter is driving, including both gates and strays from interconnect

C_g is the gate capacitance of the inverter

Note that this is the worst case transition, i.e. from low to high, since it includes the factor k. In general this derivation assumes all transitions to be worst case to simplify the expressions.

The calculations of C_l and C_g are done by counting squares in the PLA cells which are 1l by 1l, and multiplying by the appropriate capacitance/l^2. The capacitance constants used are enumerated in the following table.

C_g	gate-channel
C_d	diffusion
C_p	poly
C_{md}	metal over diffusion
C_{mp}	metal over poly

The basic algorithm for calculating the total delay through the PLA is to calculate the time constant for each section and sum them to give a final propagation time constant.

An input to the PLA is first buffered by a pair of series connected inverters which provide both the true and false versions of the input to the AND plane. Each of these input buffers have the same loading (the AND plane input gates), ignoring the loading of the second inverter on the first. Thus, a falling and rising input have approximately the same amount of delay. Assuming that each output drives all of the minterm gates, this delay is given by

$$t_{AND} = t(k+1)[\ (8C_g+6C_p)M\ ]/16C_g$$

where M is the number of minterms in the PLA. k+1 is used because one inverter output will be rising, and the other falling. With a normal input k=4. If the input is driven by a pass transistor, the first inverter will have k=8, but since its pull-down doubles in width, nothing changes. This expression includes the loading from the AND plane gates and the poly line itself.

Each AND plane output must drive the strays associated with the metal line which crosses all of its inputs in addition to the gates of the OR plane. The time constant is thus

$$t_{OR} = \mathtt{tk}[\ (8C_g+6C_p)N + 2\{8C_{md}+8C_{mp}+12C_d\}I\ ]/8C_g$$

where N is the number of PLA outputs, and I is the number of PLA inputs. Again, k=4. This formula accounts for the gates driven in the OR plane (N) and the strays of the corresponding poly line. Note that in the library PLA cells a diffusion flash and contact to the output metal wire appears even if there is no programming flash to connect it (a bit of a bug). The bracketed quantity {} is the capacitance per poly line that the metal output bus crosses including these diffusion flashes. Since each AND plane output crosses two of these for each PLA input, there is a factor of two in there.

An OR plane gate drives both the bracketed quantity above for each minterm, and an output buffer. Thus the delay through this stage is

$$t_{out} = \mathtt{tk}[8C_g+\{8C_{md}+8C_{mp}+12C_d\}M]/8C_g$$

with k=4.

Collecting all of the propagation delays, the final expression becomes

$$t_{total} = t_{AND}+t_{OR}+t_{out}$$

$$\begin{aligned} t_{total} = \mathtt{t}[&(2.5C_g +1.88C_p +4C_{md}+4C_{mp} +6C_d)M + \\ &(4C_g +3C_p)N +(8C_{md} +8C_{mp} +12C_d)I +4C_g]/C_g \end{aligned}$$

Reasonable process values for the capacitances above are

C_g	$4\text{x}10^{-4}$pf/mm^2
C_d	$1\text{x}10^{-4}$pf/mm^2
C_p	$.4\text{x}10^{-4}$pf/mm^2
C_{md}	$.4\text{x}10^{-4}$pf/mm^2
C_{mp}	$.4\text{x}10^{-4}$pf/mm^2

Plugging these into the t_{total} expression and noting that the units of all capacitances cancel yield the desired result.

$$t_{total} = t[5M + 4.3N + 4.6I + 4]$$

where M is the number of minterms
N is the number of outputs
I is the number of inputs
t is the process transit time

Notice that the constant term of 4 may be ignored for a reasonably sized PLA. A typical value for t is about .5ns, and with a good HMOS process it may be as low as .2ns. Note if the geometries are scaled this expression is still valid, since it is the ratio of load capacitance to driver gate capacitance that determines all of the constants involved.