MINIMUM PROPAGATION DELAYS IN VLSI

Carver Mead
Professor of Computer Science, Electrical Engineering and Applied Physics
California Institute of Technology
and
Martin Rem
Eindhoven University of Technology and California Institute of Technology

1. INTRODUCTION

With feature sizes decreasing and chip area increasing it becomes more and more time consuming to transport signals over long distances across the chip [5]. Designers are already introducing more levels of metal connections, using wider and thicker paths for longer distances. Another recent development is the introduction of an additional level of connections between the chip and the pc-board, multilayer ceramic chip carriers. The trend is undoubtedly towards even more connecting levels.

In this paper we demonstrate that it is possible to achieve propagation delays that are logarithmic in the lengths of the wires, provided the connection pattern is designed to meet rather strong constraints. These constraints are, in effect, satisfied only by connection patterns that exhibit a hierarchical structure. We also show that, even at the ultimate physical limits of the technology, the propagation for reasonably sized VLSI chips is dominated by these considerations, rather than by the speed of light.

2. PROPAGATION DELAY

We compute the time it takes a minimum sized transistor to drive a wire of length ℓ with width and thickness s. We assume the wire to have a distance s to its neighboring wires and layers. Let s_0 be the minimal width of a wire on the chip, so that a minimal transistor has area s_0^2.

The following equation is an excellent approximation to the total time T required to drive the wire.

$$T \approx \left(R_t + R_w\right)C_w \qquad (1)$$

R_t is the resistance of the minimal transistor, R_w the resistance of the wire and C_w its capacitance. The resistance of a wire is proportional to its length and inversely proportional to its cross section:

$$R_w = \rho \frac{\ell}{s^2} \tag{2}$$

The capacitance of a wire is inversely proportional to the distance of its neighboring wires and layers, and it is proportional to the area of the side facing that neighboring wire or layer:

$$C_w = \varepsilon \frac{s\ell}{s} = \varepsilon\ell \tag{3}$$

We notice that the product of R_w and C_w is already quadratic in ℓ. Thus the time it takes to drive a wire is at least quadratic in the wire length. However, things are not as bad as they look: R_t, the resistance of a minimal transistor, is the dominant term in (1). We can decrease that term by fitting a larger driver to the wire. But that driver must then in its turn be charged by the minimal transistor and it seems that we have hardly gained anything. That, however, is not true, for we can use a sequence of drivers instead of just one. The first one is the minimal transistor, the next one is bigger by a factor α. It drives another driver that is again bigger by a factor α, etc., until we finally reach a driver that is large enough to drive the whole wire in a sufficiently short time.

There exists a simple rule to determine the time required to have a driver charge another driver [2]. Let τ be the time it takes a minimal transistor to charge the gate of another minimal transistor. The rule is then that the time required to have a driver with capacitance C_1 drive another driver with capacitance C_2 $(C_2 > C_1)$ is

$$\tau \frac{C_2}{C_1} \tag{4}$$

Let C_t be the capacitance of a minimal transistor. We have it drive a driver with capacitance αC_t, this second one drives a driver with capacitance $\alpha^2 C_t$, etc., until the last driver has a gate capacitance of about C_w/α. The number of drivers (including the initial transistor) required is

$$\log_\alpha \frac{C_w}{C_t} \tag{5}$$

The capacitance C_t of a minimal transistor is equal to $(\varepsilon s_0^2)/d$, in which d is the thickness of the gate insulator. The number of drivers is then $\log_\alpha \ell d$ and we get for the time T_d spent in driving a zero resistance wire through the sequence of drivers:

$$T_d = \alpha\tau \log_\alpha \frac{\ell d}{s_0^2} \tag{6}$$

We may replace formula (1) by

$$T = T_d + R_w C_w \tag{7}$$

From (2), (3), (6), and (7) we conclude

$$T = \alpha\tau \log_\alpha \frac{\ell d}{s_0^2} + \rho\varepsilon \frac{\ell^2}{s^2} \tag{8}$$

We now have a formula for the propagation delay with both a logarithmic and quadratic term. One can see why a longer wire requires a larger s: that decreases the quadratic term. Actually, we wish to restrict the lengths of wires to values of ℓ that are sufficiently small to assure that the quadratic term does not dominate. We restrict ourselves to values of ℓ for which the quadratic term grows at a slower rate than the logarthmic one. Therefore, we determine the value of ℓ for which the derivates with respect to ℓ of the two terms are equal:

$$\frac{d}{d\ell}\, \alpha\tau \log_\alpha \frac{\ell d}{s_0^2} = \frac{\alpha\tau}{\ell \ln\alpha} \tag{9}$$

$$\frac{d}{d\ell}\, \rho\varepsilon \frac{\ell^2}{s^2} = \frac{2\rho\varepsilon\ell}{s^2} \tag{10}$$

If a signal has to go distance ℓ we choose a path with width and thickness s for which (9) and (10) are equal:

$$s = \ell \sqrt{\frac{2\rho\varepsilon \ln\alpha}{\alpha\tau}} \tag{11}$$

Substitution of (11) in (8) yields

$$T = \tau \frac{\alpha}{\ln\alpha} \left(\ln \frac{\ell d}{s_0^2} + \frac{1}{2} \right) \tag{12}$$

Or, approximately,

$$T = \tau\alpha \log_{\alpha} \frac{\ell d}{s_0^2} \tag{13}$$

We have assumed that the values of s could be chosen from a continuous range. Although this is a good conceptualization of the increasing number of different connection layers, in practice we will have to choose s from a discrete set. The connecting wires will be placed at different levels. The widths of the paths at the next level will be some factor β times the widths at the preceding level. Given a distance ℓ the signal has to travel, formula (11) gives us the ideal s and we choose a level at which the widths of the wires are closest to s. This leads to an interesting observation, the "magnifying glass phenomenon:" not only will the widths of the wires at any given level be the same but their lengths will also be about equal. The patterns at different levels are similar, at the next level the features are just magnified by a factor β.

2.1 Velocity of Light

Asymptotically, no signal can travel faster than the velocity of light. We must ask under what conditions the above considerations will set a limit which is more stringent, i.e., when the velocity of light limit is not attainable. In (13) we can substitute $\tau = s_0/v$ where v is the limiting velocity of electrons in the channel (a few 10^6 cm/sec in silicon)

$$T = \frac{\alpha s_0}{v} \log_{\alpha} \frac{\ell d}{s_0^2} \tag{14}$$

The maximum "velocity" with which signals can propagate is given by $1/(dT/d\ell)$

$$\frac{dT}{d\ell} = \frac{\alpha s_0}{v\ell \ln \alpha} \tag{15}$$

The domain of validity of the above results is "velocity" < c :

$$\ell < \frac{c\,\alpha s_0}{v \ln \alpha} \tag{16}$$

For typical technology today, $s_0 = 4$ microns, $\alpha/\ln\alpha$ about 6 and ℓ should be less than about a foot. Hence the velocity of light cannot be reached using the best MOS technology in the most optimal way within a typical small card bay, but will be important at larger dimensions. Even for the ultimate technology ($s_0 = 0.25$ microns), the results given above will dominate over speed-of-light considerations for chips up to about an inch across.

3. AREA

The arrangements outlined in the preceding section, allowing us to treat propagation delays as being logarithmic, will only work if we can allot enough area at the lowest level for the drivers and at the higher levels for the wires.

A minimal transistor has area s_0^2. The next driver in the sequence requires an area αs_0^2, the third one $\alpha^2 s_0^2$, etc. The total area A of the drivers thus becomes

$$A = s_0^2(1 + \alpha + \alpha^2 + \cdots)\ (\log_\alpha \ell \text{ terms}) \tag{17}$$

$$A = \frac{s_0^2(\ell - 1)}{\alpha - 1} \tag{18}$$

Or, approximately,

$$A = \frac{s_0^2 \ell}{\alpha - 1} \tag{19}$$

Notice that we can trade area for time. By increasing α the area of the drivers decreases, cf. (19), but the propagation delay increases, cf. (13).

A transistor that has to drive a wire of length ℓ requires area $s_0^2 \ell/(\alpha - 1)$ at the lowest level. This area is proportional to the length of the wire. That is fortunate: if we double both the length and the width of a chip we also double the lengths of the longest (cross chip) wires and the areas of their drivers. But the total area of the chip will quadruple and we will thus be able to double the number of wires as well.

The longer wires come on higher levels on which the wires are wider, thereby consuming more area. Each level, however, has the same area. As a result, we can accommodate the wires at the higher levels only if we do not have too many of them. Assume again that at the next level the wires are β times thicker, longer, and wider. Call the lowest level number 0 and let N_i be the number of wires at level i ($i \geq 0$), then we must have

$$N_i = N_0 \beta^{-2i} \tag{20}$$

The number of wires as a function of their lengths must decrease exponentially fast. This is a strong restriction. It suggests that efficient chips must have a tree-like structure. It is again a reason to design hierarchical chips [2], [4]. If a design does not meet this exponential rule the best we can do is getting the propagation delay linear in the wire length by inserting repeaters at equidistant positions along the wires. The consequences of linear wire delays are discussed in [1].

One may also see complexity computations that assume that wires have no delay. Thompson, e.g., writes in [6]:

> "The propagation time can be made independent of the length of the wire, by fitting larger drivers to longer wires. Larger drivers of course occupy more area, but need not take more than 10% of the area of the wire they drive. By fudging λ upwards by 5%, the area of the driver is thus absorbed into the area of its wire."

We have seen that the area of the driver is indeed proportional to the wire length, but Thompson neglects the fact that charging the gate of the larger driver will also take time. Our choice of the sequences of exponentially growing drivers allowed us to do this in a time that is logarithmic in the wire length, a technique that can work only if we have very few long wires. Thompson's model also neglects that the drivers have to be at the lowest level, in polysilicon and diffusion, independent of the level of the wire.

ACKNOWLEDGEMENTS

The research described in this paper was sponsored by the Office of Naval Research Contract No. N00014-76-C-0367 and by the Defense Advanced Research Agency, ARPA Order number 3771, and monitored by the Office of Naval Research under Contract number N00014-79-C-0597

4. REFERENCES

[1] Chazell, B. M. and L. M. Monier, "Towards More Realistic Models of Computation for VLSI," these *Proceedings*

[2] Mead, Carver and Lynn Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading MA, 1980

[3] Mead, Carver and Martin Rem, "Cost and Performance of VLSI Computing Structures," *IEEE J. Solid State Circuits* 14, No. 2, April 1979, pp. 455-462

[4] Rem, Martin, "Mathematical Aspects of VLSI Design," *Proceedings, Caltech Conference on VLSI*, (ed. C. L. Seitz), Computer Science Department, California Institute of Technology, Pasadena CA, January 1979, pp. 55-64

[5] Seitz, Charles L., "Self-Timed VLSI Systems," *Proceedings, Caltech Conference on VLSI*, (ed. C. L. Seitz), Computer Science Department, California Institute of Technology, Pasadena CA, January 1979, pp. 345-355

[6] Thompson, C. D., "Area-Time Complexity for VLSI," *Proceedings, 11th Annual ACM Symposium on the Theory of Computing*, ACM Special Interest Group on Automata and Computing Theory with IEEE Computer Society Technical Committee, Atlanta GA, May 1979, pp. 81-88