Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

Paul M.B. Vitányi

Area penalty for sublinear signal propagation delay on chip (Preliminary Version)

Department of Computer Science/Algorithmics & Architecture Report CS-R8514 Augustus

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Area Penalty for Sublinear Signal Propagation Delay on Chip (Preliminary Version)

Paul M.B. Vitányi

Centre for Mathematics & Computer Science (CWI)
Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

Sublinear signal propagation delay in VLSI circuits carries a far greater penalty in wire area than is commonly realized. Therefore, the global complexity of VLSI circuits is more layout dependent than previously thought. This effect will be truly pronounced in the emerging wafer scale integration technology. We establish lower bounds on the trade-off between sublinear signalling speed and layout area for the implementation of a complete binary tree in VLSI. In particular, sublinear delay can only be realized at the cost of superlinear area. Designs with equal length wires can either not be laid out at all, viz. for logarithmic delay, or require such long wires in the case of radical delay (i.e., nth root of the wire length) that the aimed for gain in speed is cancelled. Also for wire length distributions commonly occurring on chip it appears that the requirements for sublinear signal propagation delay tend to cancel the gain.

1980 Mathematics Subject Classification: 68C25, 94C99 69B70, 69F22
CR Categories: B.7.0, F.2.3
Keywords & Phrases: very large scale integrated circuits (VLSI), wafer scale integration, sublinear signal propagation delay, electronic principles, driving long wires, wire aspect ratio, circuit topology, complete binary tree circuits, H-tree layout, layout area, time, computational complexity and efficiency, actual wire length distributions, Rent's Rule
Note: This paper will be presented at the *26th Annual IEEE Symposium on Foundations of Computer Science*, held at Portland, Oregon, USA, October 21-23, 1985.

1. Introduction

The aim of this paper is to correct a widely spread misunderstanding. In the literature on theory of VLSI algorithms and complexity it is generally, but erroneously, held that signal propagation delay logarithmic in the wire length can be achieved on chip at the cost of an area overhead per wire which is linear (say 10%) in the area taken originally by the wire. Since the area penalty needs to be in fact far greater (viz. square in the wire length) many known lower bounds on the area$\times$time or area$\times$time^2 in the sublinear time range are way too low, while some known upper bounds are actually false. Recall also that the

Report CS-R8514
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009AB, Amsterdam, The Netherlands

probability of a chip being flawed, due to the random defects introduced by the fabrication process, rises exponentially with its area. Therefore, even small area increases necessary to obtain faster signalling speed on chip may already be prohibitive.

This paper addresses the basic model for VLSI computation. It has consequences for virtually all work in the field which considers sublinear propagation delay versus layout area complexity.

1.1. Background.

In current chips, synchronization requirements slow down the computation to a clocked switching time in the order of the delay in the longest wire. Thus, the overall efficiency of many *very large scale integrated (VLSI)* electronic switching circuits depends strongly on the signal propagation delay in long wires. As the minimal feature width continues to decrease into the submicron range this delay governs overall performance more and more. This seems truly the case on the level of the emerging *wafer scale integration* technology which manufactures chips of over 4 inch across. Within a particular technology, the *only* way currently known to obtain sublinear signal propagation delay (in the literature usually logarithmic) in long wires is by fitting a hierarchy of driver transistors to long wires, as suggested by [9]. The area cost of the ramp of driver transistors is then not more than part of the area taken by the driven wire while the width of the wire is generally, but erroneously, assumed to remain the same unit width, depending on the minimal feature width of the underlying technology. However, in [10] it appears that life is *not* so simple. Viz., the logarithmic delay assumption is incompatible with the constant wire width assumption.

To achieve a propagation delay logarithmic in the length of the wire, as in [9], electronic considerations show [10] that all wires need have the same ratio between width and length, that is, the same *aspect ratio.* (Thus, in multilayered chips now being manufactured, the communication wires are grouped in metal layers according to length. In a layer with a group of longer wires those wires are proportionally wider and thicker.) We show how the driver-hierarchy method to obtain logarithmic signal propagation delay in [9] can also be used to obtain radical (rth root) delay with similar (but less pronounced) effects on the width of long wires.

1.2. Outline of the Results

First we treat the electronic background somewhat more detailed than [10]. As a consequence, it turns out that the efficiency of VLSI circuits with sublinear propagation delay is more layout sensitive than hitherto assumed. To demonstrate this, we analyse the effect of the sublinear delay requirement on a basic circuit: the complete binary tree topology. (For more intricate circuits requiring many more long wires, like n-cubes, butterfly networks, cube-connected-cycles networks, the area penalty for sublinear delay is far greater. For matrix networks the area penalty is obviously nil.) We show for *logarithmic* delay, with a

---

* We use the Order-of-Magnitude symbols as follows:
$g(n) \in O(f(n))$ if there exists a positive constant c and $g(n) \leqslant c\,|f(n)|$ for all but finitely many positive integers n.
$g(n) \in \Omega(f(n))$ if there is a positive constant c and $g(n) \geqslant cf(n)$ for infinitely many positive integers n.
$\Theta(f(n)) = O(f(n)) \cap \Omega(f(n))$.

constant aspect ratio a for the wires and using c layers:

- *every* layout of a complete binary tree with N nodes takes $\Omega(N \log^{a/6c} N)$ area*, so *also* the H-tree layout of [9,8].
- For synchronization it may be required that all wires have the same length. A layout for a complete binary tree with all *wires of equal length* turns out to be *impossible* for large enough N.

For *radical* delay, that is, delay according to the rth root of the wire length (using a constant number of layers) we find:

- The layout area of a complete binary tree is bounded below by the product of the number of nodes N and an unbounded function of r.
- For layouts of complete binary trees with *equal length wires*, the wire length needs to increase exponential with r. Therefore, the gain in speed is lost completely by cause of the longer wires, while the required layout area rises exponential with r nonetheless.
- We also briefly investigate the effect on natural *wire length distributions*. Using plausible arguments, and empirical data from actual chips, [5] determines what wire length distributions tend to occur. For these distributions it appears that the gain of having sublinear propagation delay is cancelled by the requirements on wire area. (Because the wires need to be much longer, the faster signals take as long to traverse a particular wire as they did before.) Worse, logarithmic delay may be impossible outright.

1.3. Related work.

Almost all work in the theory of computing of sublinear propagation delay VLSI models is related in some way to the present issue. E.g., apart from the fact that we require wide wires to obtain a sublinear propagation delay, we also need to insert the *drivers* to drive the long wires (cf. next section). In [9,8,10], it is shown that such drivers need an area proportional to -say 10% of- the length of the wire. In [12] questions are treated concerning the insertion of these drivers in given layouts with *unit width* wires *before and after* insertion. Some effects of increasing wire width in layouts, or related issues, have been treated in various contexts in, e.g. [13,6].,

In the computational models rampant in the literature, the assumptions concerning signal propagation delay in long wires range from constant delay (irrespective of the wire length) [17,2,3,21,15], via logarithmic [11,19,18], and linear delay [1,4], to signal propagation delay that is square in the length of wires [1,4,16]. In all of these models the width (or thickness) of wires is assumed to be a unit, depending on the minimal feature width of the underlying technology as seems suggested in [9,8]. However, for sublinear signal propagation delay this is a misunderstanding [10] of which the consequences are explored in the present paper.

2. Electronic Basics

The time it takes a minimum transistor to drive a wire of length L, width W and thickness H can be estimated as follows. The wire is assumed to have distance D_l to neighbouring layers and D_w to other wires in the same layer. If W_0 is the minimal width of a wire in the current technology, then the minimal transistor, consisting of a wire crossing, occupies area W_0^2. The total time T to drive a wire is approximated by:

$$T \approx (R_t + R_w)\, C_w \tag{1}$$

where R_t is the resistance of the minimum transistor, R_w the resistance of the wire and C_w its capacitance. Therefore, the total time T can be thought of as the sum of the time T_d needed to drive a zero resistance wire of capacitance C_w, and the time $R_w C_w$ needed to transport the appropriate charge from a zero resistance source. Roughly, T_d is the time needed to transport the necessary charge through the bottleneck consisting of the switch (the minimal transistor), and $R_w C_w$ is the time needed to distribute the charge appropriately over the wire w. Since the resistance of a wire is proportional to its length and inversely proportional to its cross section we have:

$$R_w = \rho_w \frac{L}{WH} \tag{2}$$

where ρ_w is the resistivity of the considered wire material. The capacitance of a wire is inversely proportional to the distance of its neighbouring wires and layers, and proportional to the area of the side facing that neighbouring layer or wire:

$$C_w = 2\epsilon_w L \left(\frac{H}{D_w} + \frac{W}{D_l}\right) \tag{3}$$

where ϵ_w is a proportional constant consisting of the product of the permittivity of free space and the dielectric constant of the insulating material (usually SiO_2). Thus,

$$R_w C_w = 2\rho_w \epsilon_w \frac{L^2}{WH} \left(\frac{H}{D_w} + \frac{W}{D_l}\right) . \tag{4}$$

This suggests a signal propagation time quadratic in L. However, the resistance R_t of the minimum transistor dominates in (1) for the magnitudes of L under consideration (smaller than, say, 1 foot). We can decrease that term by fitting a larger driver transistor to the wire. This transistor, in its turn, must be driven by the minimal transistor. Iterating this scheme, cf [9], we obtain a sequence of transistors, of which each next one is a factor α larger than the preceding one. The final transistor in the sequence should be large enough to drive the wire in a sufficiently short time. (We can think of this scheme as a sequence of switches where each switch serves to switch the next larger switch, and the largest switch in the sequence controls the large channel through which the charge is transported to the wire. Although the time to actually pass the appropriate charge from source to wire can be made smaller by fitting a larger final driver transistor to the sequence, there seems no way to get rid of the time needed to switch all transistors in between the smallest transistor and the largest one.) The time to drive a driver with capacitance C_2 by a driver with smaller capacitance C_1 is given by [9]:

$$\tau \frac{C_2}{C_1} \tag{5}$$

where τ is the time it takes a minimal transistor to charge the gate of another minimal transistor. If C_t is the capacitance of the minimal transistor then for a ramp of r drivers:

$$r = \log_\alpha \frac{C_w}{C_t} \quad \& \quad \alpha = \left[\frac{C_w}{C_t}\right]^{\frac{1}{r}} \quad , \tag{6}$$

taking $T_d = r\tau\alpha$ time to charge the wire if it had no resistance. The capacitance of the minimum transistor is given by

$$C_t = \epsilon_t \frac{W_0^2}{D_0} \quad , \tag{7}$$

where D_0 is the thickness of the gate insulator and ϵ_t is the product of the permittivity of free space and the dielectric constant of the gate insulator. Thus we can drive a zero resistance wire of capacitance C_w through a sequence of r drivers for fixed α in time:

$$T_d = \alpha\tau \log_\alpha \frac{C_w}{C_t} \quad . \tag{8a}$$

We can also use a ramp of r drivers, for some fixed r, and choose α accordingly:

$$T_d = r\tau \left[\frac{C_w}{C_t}\right]^{\frac{1}{r}} \quad . \tag{8b}$$

From (1), (3), (4) and (8a) we obtain an expression for T.

$$T \approx \alpha\tau \log_\alpha \frac{C_w}{C_t} + 2\rho_w \epsilon_w \frac{L^2}{WH}\left(\frac{H}{D_w} + \frac{W}{D_l}\right) \quad . \tag{9a}$$

From (1), (3), (4) and (8b) we obtain:

$$T \approx r\tau \left[\frac{C_w}{C_t}\right]^{\frac{1}{r}} + 2\rho_w \epsilon_w \frac{L^2}{WH}\left(\frac{H}{D_w} + \frac{W}{D_l}\right) \quad . \tag{9b}$$

It is therefore clear that the signal propagation time heavily depends on the various dimensions and materials involved in the chip. The relation (8a) can be considered a borderline case of (8b). In [10] it was observed that by keeping the derivatives, with respect to L, of the two terms in the right-hand side of equations like (9a&b) balanced:

$$\frac{\alpha\tau}{L \ln\alpha} \approx \rho_w \epsilon_w \frac{L}{WH}\left(\frac{H}{D_w} + \frac{W}{D_l}\right) \quad , \tag{10a}$$

T grows logarithmic in L. With

$$\tau \left[\frac{\epsilon_w D_0 L^{1-r}}{\epsilon_t W_0^2}\left(\frac{H}{D_w} + \frac{W}{D_l}\right)\right]^{\frac{1}{r}} \tag{10b}$$

$$\approx \rho_w \epsilon_w \frac{L}{WH} \left(\frac{H}{D_w} + \frac{W}{D_l} \right) ,$$

T grows as the rth root of L. Viz., from (9a) we obtain by assumption of equality (10a):

$$T \approx \frac{\alpha\tau}{\ln \alpha} \left\{ \ln \left[\frac{\epsilon_w D_0 L}{\epsilon_t W_0^2} \left(\frac{H}{D_w} + \frac{W}{D_l} \right) \right] + 1 \right\} \tag{11a}$$

and from (9b) we obtain by assumption of equality (10b):

$$T \approx (r+1)\tau \left[\frac{\epsilon_w D_0 L}{\epsilon_t W_0^2} \left(\frac{H}{D_w} + \frac{W}{D_l} \right) \right]^{\frac{1}{r}} . \tag{11b}$$

In the next section we establish the area penalty involved with a wire of given length L.

Remark. Different ratios between the successive capacitances of the transistors in the ramp of drivers can be used to obtain, for instance, noninteger values for r in the formulas above. This issue is not addressed in the present paper.

Without a ramp of driver transistors, having the minimum transistor drive the total wire outright, we obtain from (1), (4), (5) and (7):

$$T \approx \tau \frac{C_w}{C_t} + C_w R_w \tag{12}$$

$$= \left(\frac{\tau D_0}{\epsilon_t W_0^2} + \rho_w \frac{L}{WH} \right) 2\epsilon_w L \left(\frac{H}{D_w} + \frac{W}{D_l} \right) .$$

The above analysis shows that we can reduce the signal propagation time by employing new materials with more favorable characteristics, like Gallium-Arsenide and Silicon-on-Sapphire technologies [22]. We can also change the size of the wires and the interwire- and interlevel separation. To increase signalling seed, extra layers with wider and thicker wires are used for the long interconnect wires.

3. Sublinear Delay and Wire Dimensions

3.1. Logarithmic Delay and Constant Aspect Ratio

Under assumption (10a) we can obtain a logarithmic signal propagation delay by, all other things being equal, maintaining:

$$L^2 \left(\frac{1}{W D_w} + \frac{1}{H D_l} \right) = \text{constant} , \tag{13a}$$

rather than by just keeping L^2 proportional to WH as in [10]. Keeping the interwire distance proportional to the wire width, and the interlayer distance proportional to the wire height, we observe that if W, H and L are kept in proportion a logarithmic propagation delay is attained. (Note that we cannot reach this effect by keeping the wire width the same but using very 'tall' wires or vice versa.) The *aspect ratio* of a wire is the quotient of its width and length. To obtain a logarithmic signal propagation delay we thus need the fixed

constant aspect ratio following from (10) and (13a) for all wires in the layout. In designing such a high speed layout we therefore need to install drivers to drive the long wires and to design all wires with a constant aspect ratio $a>0$. Therefore, a wire of length L in such a layout has *area* aL^2. The area taken by the driver is linear in the length of the wire [10]: the minimal transistor occupies area W_0^2, the next driver area αW_0^2, and so on for $\log_\alpha L$ terms for an L-length wire. The total driver area for an L-length wire becomes $W_0^2(L-1)/(\alpha-1)$. This area is required at the lowest silicon layer of the chip; the long interconnect wires are executed in the upper metal layers.

3.2. Radical Delay

Under assumption (10b) we can obtain a signal propagation delay of the order of the rth root of the length of the wire under a certain balancing of the aspect ratio of the dimensions of the wire:

$$\frac{L^{2-\frac{1}{r}}}{WH}\left(\frac{H}{D_w}+\frac{W}{D_l}\right)^{1-\frac{1}{r}} = \text{constant} \ . \tag{13b}$$

Call this type *radical* delay. (Together with the previous logarithmical delay this essentially exhausts *all* possibilities for sublinear propagation delay obtainable by the [10] method.) We assume that all dimensions (but for L) are scaled proportional to the same radical fraction L^α of L ($1<\alpha<1$). So each parameter $X\neq L$ in (13b) satisfies $X = a_X L^\alpha$, for some a_X a constant depending only on X. Therefore, for fixed given r equation (13b) determines α by:

$$r = \frac{1}{2(1-\alpha)} \ . \tag{14}$$

(The logarithmic delay of (13a). is, in a certain sense, a limiting case of this radical delay.) Hence, to obtain a propagation delay proportional to the rth root of the length L of the wire the dimensions need to be scaled proportional to $L^{1-1/2r}$ and a wire of length L takes area

$$\Omega(L^{2-\frac{1}{2r}}) \ . \tag{15}$$

For $r=\frac{1}{2}$ this yields the limiting worst-case quadratic delay with all dimensions (but L) scaled proportional to constants like the minimum feature width. For $\frac{1}{2}\leq r\leq 1$, however, there is another way by inserting repeaters (inverters) at constant intervals in the long wires. This gives linear propagation delay ($r=1$) at an area cost in repeaters, only linear in the length of the wire, anyhow. Therefore, we are only interested in the case $r>1$, that is, *sublinear* propagation delay.

Note, that the effect of the scaling to obtain sublinear propagation delay is not only confined to the area but also to the height of the chip, since all dimensions need to be scaled proportional. So, the volume of a wire of length L needs be $\Omega(L^3)$ for logarithmic propagation delay, by (13a), and $\Omega(L^{3-1/r})$ for a rth root of L propagation delay, similar to (15).

4. Area, Length and Time

The area for a VLSI layout is expressed in A area units. The area unit is the square of the basic length unit which is the feature width of the underlying technology. This is currently $4{\cdot}10^{-6}$ - $10{\cdot}10^{-6}$ meter and is expected to continue to decrease in the submicron level in the near future.

- The *area* is taken to be the area of the smallest convex region enclosing the layout.
- There is a *cross-over* constant $c>0$ such that no unit circle encloses points of more than c different edges (wires) or nodes (components or transistors).

(In case we allow an unlimited amount of cross-over, we should consider the worst-case 'area $\times$ cross-over' product instead of the area. Effectively, we then consider 3-dimensional 'layered' chips which is outside the scope of this paper.)

4.1. Time

The *execution time* of a problem instance is the time elapsed between the entering of the first bit of the problem instance in the circuit and the leaving of the last bit of the answer from the circuit. In *pipelined* and especially in *systolic* computations [9] the *period* is important. The period is the time elapsed between entering the first bit of a problem instance and the first bit of a next problem instance. In 'moving belt' type computation the period can be substantially less than the execution time. Below the period of a (systolic) computation appears to be more sensitive for the propagation delay assumption than the overall execution time. If the signal propagation delay depends on the length of the wire the signal has to traverse, then the *minimax edge length* in the layout will determine the period in a systolic network. The minimax edge length (or wire length) $e(.)$ of a layout for a given circuit is the minimum over all layouts, implementing the circuit, of the length of the longest wire in such a layout. See also [11].

4.2. Implementation Details of Area and Time

We may assume that the circuits are laid out on a Manhattan grid. In [7] algorithms are presented to embed easily separated graphs efficiently in grids. The considerations below assume that the processing elements have unit area and the links between them have unit bandwidth. This view captures the underlying communication structure. This leaves free the precise implementation. For example, to communicate a word of k bits between two processing elements, one can either use a link of bandwidth k and one cycle or use a link of bandwidth 1 and k cycles. Let each of the P processing elements actually fit in area U and let the total area used by the links normalized to bandwidth 1 be L. Let W be the bandwidth of the precise implementation, e.g., word-parallel or word-serial communication. Using methods of [7], a rough upper bound on the total actual chip area is given by $A_p \in O(PU)$ plus $A_w \in O(LW^2)$, where A_p is the area taken by the processing elements and A_w is the area for the wires. We follow the normalization generally adhered to, so by a layout *area* A we mean $P+L$, that is, we normalize the processing elements to unit area and the wires to unit bandwidth. Concomitant with the assumption of unit bandwidth, we also assume that each communication between processors concerns a unit (bit). Once the bandwidth and sizes of messages are resolved, the estimates give a basis for the actual chip

area and the actual chip times.

5. Tree Circuit and Sublinear Delay

Complete binary trees are basic ingredients for many circuits. They are exemplary for the embedding of a large class of hierarchical circuits in layouts on silicon. Such circuits are obvious candidates on which to test the effects of the wire area penalty for sublinear delay on chip.

The described effect of sublinear delay on short-wire-length layouts like two-dimensional Manhattan circuits is nil because all wires have the same constant length in the layout, while on circuits like fast permutation networks (cube-connected cycles, perfect shuffle, butterfly) the effects are very pronounced, and perhaps disastrous, because necessarily there are many long wires in the layout [20]. Does the wire area penalty involved with sublinear signal delay imply a significant overall layout area penalty for that most significant example in between: the complete binary tree? This depends on whether the layout for such a tree has long wires.

Recall that the H-tree layout for a complete binary rooted tree [9] achieves area less than $4N$ for an N-node complete binary tree, *under the unit wire width assumption.* Below we give upper and lower bounds on the layout area for a complete binary tree circuit with (i) logarithmic signal propagation delay and (ii) rth root signal propagation delay, using to the model described in the previous section.

Proposition. *Each layout of a complete binary tree with N leaves (i.e., $2N-1$ nodes), using A area, contains wires of length at least $e(N) = \sqrt{A} / (2\log N)$. Moreover, there is a path from the root to a leaf of total length of at least $\Omega(\sqrt{A})$.*

Proofsketch. We can find two points p and q in the layout which are at least $\sqrt{A}$ units apart. p and q can be nodes or locations on a wire. To go from p to q along the edges of the tree cannot cause us to traverse more than $2\log N$ edges. Hence, there is an wire in the layout of length at least $\sqrt{A} / (2\log N)$ units. (For instance, if $A \in \Omega(N)$ then the minimax edge length for a complete binary tree layout is $e(N) \in \Omega(\sqrt{N} / \log N)$.)

If r is the node on the $\sqrt{A}$ length path between p and q which is nearest to the root, then either $r-p$ or $r-q$ is a $\Omega(\sqrt{A})$-length subpath of a path from the root to a leaf. □

Definition. We denote by $A(N)$ the *minimal area* to lay out a complete binary tree with N leaves under the different assumptions on wire area and cross over.

5.1. Logarithmic Signal Propagation Delay.

Let the area occupied by a wire of length L be aL^2 for some constant $0<a\leqslant 1$.

Upper bound. We analyse the area occupied by an H-tree layout with N leaves and no overlap. This is an upper bound on $A(N)$. Recall the familiar picture of the H-tree layout with constant width wires for complete binary trees as given in, for instance, [9]. Let $N = 2^m$. Let the ratio between the length of the wires at two consecutive levels be α. That is, the quotient of the length of a level $k+1$ wire and the length of a level k wire is α, $0<\alpha<1$, for all $0\leqslant k<m$ ($k=0$ is root level and $k=m$ is leaf level). Considering that layout, it is not to difficult to see that, with constant aspect ratio a,

$$\alpha^k \geq a\alpha^{k-1} + 2\alpha^{k+2} \quad ,$$

for each level k between 1 and m, suffices to layout the H-tree compactly with no overlap of wires and nodes. Consequently we obtain

$$0 < a \leq \alpha(1-2\alpha^2) \quad ,$$

and, for both a, $\alpha > 0$,

$$0 < \alpha < \frac{\sqrt{2}}{2} \quad \& \quad 0 < a \leq \frac{2}{3\sqrt{6}} \quad .$$

Note that in the limit for $\alpha \to \sqrt{2}/2$ we have that $a \to 0$. For given α and a in the appropriate ranges, an *upper bound* on the total wire area plus node area for the H-tree layout is computed by:

$$\begin{aligned} A(N) &\leq a \sum_{k=0}^{\infty} 2^k \alpha^{2k-2m} + 2^{m+1} \\ &= \frac{a\alpha^{-2m}}{1-2\alpha^2} + 2^{m+1} \quad . \end{aligned}$$

Therefore,

$$A(N) \leq CN^{-2\log_2 \alpha} + 2N \quad ,$$

with $C = a/(1-2\alpha^2)$. From the relation between a and α it follows that $C \leq \alpha$ for $0 < \alpha < \sqrt{2}/2$. Setting $a = (1-2\alpha^2)/2$, so $1/2 \leq \alpha < \sqrt{2}/2$ and therefore $0 < a \leq 1/4$ and $C = 1/2$, yields

$$A(N) \approx \tfrac{1}{2} N^{1-\log_2(1-2a)} + 2N \quad .$$

Therefore, for each $\epsilon > 0$ there is an $\alpha < \sqrt{2}/2$ such that $A(N) \in O(N^{1+\epsilon})$. Since a must be greater than 0 this ϵ remains greater than 0 as well.
For $\alpha \to \sqrt{2}/2$ (so the aspect ratio $a \to 0$) the upper bound on $A(N)$ goes to:

$$\begin{aligned} &\lim_{\alpha \to \sqrt{2}/2} a \sum_{k=0}^{m} 2^k \alpha^{2k-2m} + 2^{m+1} \\ &\quad = \lim_{\alpha \to \sqrt{2}/2} am\alpha^{-2m} + 2^{m+1} \\ &\quad = aN \log N + 2N \quad . \end{aligned}$$

Lower bound. Let $N = 2^m$, and let c be the maximal amount of cross-over. We obtain a lower bound on the layout area for a complete binary tree for *any layout*, so *also* the H-tree layout. For each i, $1 \leq i \leq m$, $A(2^i)$ is the *minimum layout area* for a complete binary tree T_i with 2^i leaves, under the current assumptions of c cross-over and a aspect ratio. Imagine, for the sake of the argument, a (nonexistent) layout for T_m, such that each maximal subtree determined by a node of T_m takes minimal area. Selecting wires from the maximal lengths paths in these subtrees, we sum their areas while taking care that each such wire is counted only once. Let T_i be a complete binary tree of $i+1$ levels with the root at level 0 and the

leaves at level i.

Claim 1. There is a path in T_i of at most $2i$ edges and of length at least $\sqrt{A(2^i)}$ in the layout. The sum of the areas of the wires in such a path exceeds $aA(2^i)/2ic$.

Proof of Claim. By arguments concerning the diameter of the smallest convex area containing T_i it is easy to see that there is a path of length $\sqrt{A(2^i)}$ in the layout with $1 \leqslant j_i \leqslant 2i$ wires. The sum of the area of the wires in such a path is therefore $\Omega(aA(2^i)/cj_i)$. (Assume that the area of the wires in the path can be distributed over c levels, so as to keep $A(2^i)$ as small as possible.) □

Claim 2. Let P be a path through T_i. At each level j, $2 \leqslant j \leqslant i$, there are at least 2 maximal complete binary subtrees of $i-j+1$ levels with roots at level j, which have *no* node in common with P. Moreover, the $2(i-1)$ complete binary subtrees concerned are pairwise disjoint as well. (This is easy to verify from a simple picture.)

By Claims 1 and 2, we can give a lower bound on the area $A(2^i)$ of T_i by adding the minimal possible area of a $\sqrt{A(2^i)}$-length path in T_i to twice the sum for j ranges from 2 through i of the areas of subtrees T_{i-j}:

$$A(2^i) > \frac{aA(2^i)}{2ci} + 2\sum_{j=2}^{i} A(2^{i-j}).$$

Unfolding this inequality we obtain

$$A(2^i) > \sum_{j=0}^{i-1} \frac{aF_j A(2^{i-j})}{2c(i-j)} + \frac{aF_i A(1)}{c} \quad ,$$

with F_j the j-th element of the Fibonacci-like sequence generated by the recurrence relation $F_i = F_{i-1} + 2F_{i-2}$ with $F_0 = 1$ and $F_1 = 0$. Therefore,

$$F_i = \frac{2^i + 2(-1)^i}{3} \approx \frac{2^i}{3} \quad .$$

Substituting $A(2^i) = g(i)2^i$ in the inequality above yields

$$g(i) > \frac{a}{6c} \sum_{j=1}^{i} \frac{g(j)}{j} + \frac{a}{3c} \quad ,$$

which is satisfied for $g(i) \in \Omega(i^\epsilon)$ with $\epsilon > a/6c$. Hence,

$$A(N) \in \Omega(N \log^{a/6c} N) \quad .$$

Crudely derived, considering only volume without considerations of placement and routing, this lower bound is yet nonlinear and reflects both the necessary influence of the aspect ratio a and the cross-over coefficient c. It seems likely that more sophisticated arguments for fixed aspect ratio $a > 0$ will raise the lower bound to match the upper bound $\Omega(N^{1+\epsilon})$.

5.2. *Radical Signal Propagation Delay*

For radical signal propagation delay proportional to the rth root of the length L of the wire, the *area* taken by the wire needs be $aL^{2-1/2r}$, for some constant a, by (15).

Upper bound. The upper bound on $A(N)$ is determined by using the H-tree layout (without overlap) again. Let $N = 2^m$. Let again the quotient between the lengths of the

wires at two consecutive levels $k+1$ and k be α, $0<\alpha<1$, with the root at level 0 and the leaves at level m. Considering the H-tree layout, it is not to difficult to see that, with aspect ratio $aL^{-1/2r}$ for wires of length L $(0<a<1)$,

$$\alpha^k \geq a\alpha^{(k-1)(1-\frac{1}{2r})} + 2\alpha^{k+2} ,$$

for each level k between 0 and m, suffices to layout the H-tree compactly with no overlap of wires and nodes. Consequently we obtain

$$0 < a \leq \alpha^{1+\frac{k-1}{2r}}(1-2\alpha^2) ,$$

and, for $a, \alpha > 0$,

$$0<\alpha<\frac{\sqrt{2}}{2} ,$$

$$0<a<(\frac{2r+k-1}{12r+2k-2})^{\frac{2r+k-1}{4r}}(1-\frac{4r+2k-2}{12r+2k-2}).$$

For $r\to\infty$ we obtain the upper bound on a we saw above for the logarithmic delay case. Note that again for $\alpha\to\sqrt{2}/2$ we have that $a\to 0$. For given α and a in appropriate ranges, that is,

$$2\alpha^{2-\frac{1}{2r}} < 1$$

and a accordingly, the *upper bound* on $A(N)$, consisting of the total wire area plus node area for the H-tree layout, is computed as follows:

$$\begin{aligned} A(N) &\leq a\sum_{k=0}^{\infty} 2^k\alpha^{(k-m)(2-\frac{1}{2r})} + 2^{m+1} \\ &= \frac{a\alpha^{-(2-\frac{1}{2r})m}}{1-2\alpha^{2-\frac{1}{2r}}} + 2^{m+1} . \end{aligned}$$

Therefore,

$$A(N) \leq CN^{-(2-\frac{1}{2r})\log_2\alpha} + 2N ,$$

with $C = a/(1-2\alpha^{2-1/2r})$. From the constraints on a and α it follows that $C\leq\alpha$. Setting $a=(1-2\alpha^{2-1/2r})/2$ we have $C=1/2$. Therefore, for each $r\geq\frac{1}{2}$ and $\epsilon>0$ there is an $\alpha<\sqrt{2}/2$ such that $A(N)\in O(N^{1+\epsilon})$.

For

$$2^{\frac{-2r}{4r-1}} \leq \alpha < 2^{-\frac{1}{2}}$$

we leave the analysis to the reader.

For $\alpha \to \sqrt{2}/2$ (so $a \to 0$) the upper bound on $A(N)$ goes to:

$$\lim_{\alpha \to \sqrt{2}/2} a \sum_{k=0}^{m} 2^k \alpha^{(k-m)(2-\frac{1}{2r})} + 2^{m+1}$$

$$= \frac{aN}{1-2^{-1/4r}} + 2N \ .$$

Lower bound. Let $N = 2^m$. Let $aL^{-1/2r}$ be the aspect ratio of the wires to obtain rth root radical signal propagation delay. Let c be the maximal amount of cross-over. For each i, $0 \leq i \leq m$, let $A(2^i)$ be the *minimum layout area* for a complete binary tree T_i with 2^i leaves. Imagine, for the sake of the argument, a virtual layout for T_m, such that each maximal subtree determined by a node of T_m takes minimal area. Selecting wires from the maximal lengths paths in these subtrees, we sum their areas while taking care that each such wire is counted only once.

Claim 1. Let T_i be a complete binary tree of $i+1$ levels with the root at level 0 and the leaves at level i. If the minimal layout area of T_i is $A(2^i)$ then there is a path in T_i of at most $2i$ edges and of length at least $\sqrt{A(2^i)}$ in the layout. The area taken by this path must therefore exceed

$$\frac{a}{c}(2i)^{-1+\frac{1}{2r}} A(2^i)^{1-\frac{1}{4r}} \ ,$$

where $aL^{-1/2r}$ is the aspect ratio of the wires and c the maximal amount of cross-over.

Proof of Claim. By arguments concerning the diameter of the smallest convex area containing T_i it is easy to see that there is a path of length $\sqrt{A(2^i)}$ in the layout with $1 \leq j_i \leq 2i$ wires. The area of L-length wires is $aL^{2-1/2r}$, and $r > 1$. The sum of the area of the wires in such a path is therefore least if the path contains as many wires as possible, that is, $2i$ wires. By distributing the area of the wires over c levels we obtain the expression in the claim. □

Claim 2. Let P be a path through T_i. At each level j, $2 \leq j \leq i$, there are at least 2 maximal complete binary subtrees of $i-j+1$ levels with roots at level j, which have *no* node in common with P. Moreover, the $2(i-1)$ complete binary subtrees concerned are pairwise disjoint as well. (This is easy to verify from a simple picture.)

So by Claims 1 and 2, we can give a lower bound on the area $A(2^i)$ of T_i by adding the minimal possible area of a $\sqrt{A(2^i)}$-length path in T_i to twice the sum for j ranges from 2 through i of the areas of subtrees T_{i-j}:

$$A(2^i) > \frac{a}{c}(2i)^{-1+\frac{1}{2r}} A(2^i)^{1-\frac{1}{4r}} + 2\sum_{j=2}^{i} A(2^{i-j}).$$

Unfolding this inequality we obtain

$$A(2^i) > \frac{a}{c} \sum_{j=0}^{i-1} F_j (2i-2j)^{-1+\frac{1}{2r}} A(2^{i-j})^{1-\frac{1}{4r}} + \frac{aF_i A(1)}{c} \ ,$$

with F_j the j-th element of the Fibonacci-like sequence generated by the recurrence relation

$F_i = F_{i-1} + 2F_{i-2}$ with $F_0 = 1$ and $F_1 = 0$. Therefore,

$$F_i = \frac{2^i + 2(-1)^i}{3} \approx \frac{2^i}{3} \quad .$$

Substituting

$$A(2^i) = D(r,i)2^i$$

and changing the summation order in the inequality above yields

$$D(r,i) > \frac{a2^{-1+\frac{1}{2r}}}{3c} \sum_{j=1}^{i} j^{-1+\frac{1}{2r}} D(r,j)^{1-\frac{1}{4r}} 2^{-\frac{j}{4r}} + \frac{a}{3c} \qquad (*)$$

Since $D(r,j) \geq 1$ $(j > 0)$, we can bound $D(r,i)$ below by:

$$D(r,i) > \frac{a2^{-1+\frac{1}{2r}}}{3c} \sum_{j=1}^{i} j^{-1+\frac{1}{2r}} 2^{-\frac{j}{4r}} + \frac{a}{3c}$$

$$\geq \frac{a2^{-1+\frac{1}{2r}} R(r)}{3c} + \frac{a}{3c} \quad ,$$

where the series converges to the unbounded function $R(r)$.

For boundary value $r = \frac{1}{2}$, we obtain from (*) that $D(\frac{1}{2},i) \in O(1)$ suffices, which is witnessed by the unit wire width H-tree. For the other extreme value of r the inequality (*) yields:

$$\lim_{r \to \infty} D(r,i) \in \Omega(i^{a/6c}) \quad ,$$

giving us the earlier derived lower bound for logarithmic delay. Considering only volume without considerations of placement and routing, this lower bound reflects the influence of the radical r, the aspect ratio a and the cross-over coefficient c.

5.3. Execution Time and Period

Let again $A(N)$ denote the minimal area for the layout of a complete binary tree T with $N = 2^m$ leaves under the appropriate assumptions on wire area and cross-over. The *period* is computed from the minimax wire length

$$e(N) \in \Omega(\sqrt{A(N)} / \log N) \quad .$$

The *execution time* is the greatest sum of the delays along a path from the root to a leaf. The delay in each wire of each path is at least that of the longest wire in that path. Therefore, the execution time is at least $\log N$ times the delay in an $e(N)$ length wire. According to the Proposition, the minimax edge length $e(N) \in \Omega(\sqrt{A(N)} / \log N)$, and therefore the execution time is $\Omega(\sqrt{A(N)})$. Consequently, together with the respective minimal layout areas $A(N)$ for the different propagation delays we obtain:

	Logarithmic	radical (r)
area	$\Omega(N\log^{a/6c}N)\cap O(N^{1-\log_2(1-2a)})$	$\Omega(D(r,i)N)$
period	$\Theta(\log N)$	$\Omega\left((D(r,i)N)^{\frac{1}{2r}}\log^{-\frac{1}{r}}N\right)$
Execution Time	$\Theta(\log^2 N)$	$\Omega\left((D(r,i)N)^{\frac{1}{2r}}\log^{1-\frac{1}{r}}N\right)$

Table 1. Minimal area with unequal length wires. Here a is the wire aspect ratio c is the cross-over coefficient or number of layers and radical propagation delay as the rth root.

5.4. *Layouts with Equal Length Wires*

If we want to synchronize then it may be preferable to have layouts with only *equal length* wires. Under the constant wire width assumption, the least such wire length for a layout of a complete binary tree is $N/\log^2 N$ with simultaneous least area of $\Theta(N^2/\log^2 N)$ [11]. Table 2 summarizes the effect of the requirement of equal length wires on layouts of a complete binary N-node tree. The derivation is given below.

	Logarithmic	Radical (r)
area	$\Omega(aN^3/c\log^4 N)$	$\Omega\left(\frac{aN}{c}\left[\frac{aN}{4c\log^2 N}\right]^{4r-1}\right)$
period	–	$\Omega\left(\left[\frac{aN}{4c\log^2 N}\right]^2\right)$
execution time	–	$\Omega\left(\left[\frac{aN}{4c\log^2 N}\right]^2\log N\right)$

Table 2. Minimal area with equal length wires, with r the radical, a the wire aspect ratio and c the cross-over coefficient or number of layers.

For *logarithmic* propagation delay with wires of constant aspect ratio a, viz. the left column of the table, the lower bound on the area is obtained by determining the combined area

taken by N wires of length $L \in \Omega(N / \log^2 N)$. However, under the requirement of unique wire length $e(N)$ for all wires, cross-over number c (number of layers) and an aspect ratio a, the following relations have to hold for any complete binary N-node tree layout.

$$\frac{ae(N)^2 N}{c} \leqslant A(N) \leqslant 4e(N)^2 \log^2 N \ .$$

Viz., on the one hand the area must accommodate all wires, on the other hand the diameter of the layout cannot exceed the length of the longest path ($2\log N$ edges). Therefore,

$$\frac{a}{c} \leqslant \frac{4\log^2 N}{N} \ ,$$

and, for fixed constant a and c independent of N, the desired layout is *impossible* for large enough N. Moreover, substituting the upper bound on a / c in the the lower bound on the area then yields the [11] value $A(N) \in \Omega(N^2 / \log^2 N)$ again, indicating that such a circuit gets impossible for already quite small N. Therefore, the period an execution time are $-$.

Similarly, for *radical* (rth root) propagation delay we have:

$$\frac{ae(N)^{2-\frac{1}{2r}} N}{c} \leqslant A(N) \leqslant 4e(N)^2 \log^2 N \ .$$

and therefore

$$e(N) \geqslant \left[\frac{aN}{4c \log^2 N}\right]^{2r} \ .$$

Substituting $e(N)$ of the last displayed equation in the left hand term of the preceding displayed inequality yields the lower area bound on $A(N)$ in Table 2. Note that the boundary case $r = ½$ also yields the standard [11] value for $e(N)$ for unit width wires. The *period* is $\Omega(e(N)^{1/r})$ and the *execution time* is $\Omega(e(N)^{1/r} \log N)$. Therefore, the effect of the longer wires required for $O(L^{1/r})$ signal propagation delay for L-length wires eradicates the gain over quadratic delay, while the area rises exponential with r nonetheless. Consequently, a hierarchy of drivers is not a viable solution to speed up a tree implementation with equal length wires. For such circuits with equal length wires a better solution is periodic repeaters in long wires giving linear delay.

6. Wire Length Distributions

Let $f: \mathbb{N} \to \mathbb{N}$, connected with a VLSI layout, be a *wire length distribution* function which yields the number $f(i)$ of wires of length i in the design.

Every VLSI layout must have a constant bounded fan-in and fan-out of wires for the components (transistors). If the chip area is A, then the average maximal wire length $L_{\max}$ can be estimated by the statistical formula [14] $L_{\max} = KA^g$ with K and g constants which, by rule of thumb, can be set to ½ each. A reasonable assumption therefore is that the maximal wire length on a chip does not exceed

$$L_{\max} = \sqrt{A} \ . \tag{16a}$$

Consequently, the amount of wires in the layout is given by

$$\#\text{wires} = \sum_{i=1}^{\sqrt{A}} f(i) \ . \tag{16b}$$

We now identify a common class of wire-length distributions for VLSI layouts. Firstly, it is argued that the requirement of logarithmic propagation delay favors such distributions. Secondly, other studies have shown such distributions to be likely for VLSI layouts on both theoretical and empirical grounds.

6.1. Logarithmic Delay

Recall, that to obtain a logarithmic signal propagation delay we need a fixed constant aspect ratio for all wires in the layout. In designing a high speed layout we therefore needed to install drivers to drive the long wires and to design all wires with constant aspect ratio. The area taken by such a driver is linear in the length of the wire. This area is required at the lowest silicon layer of the chip; the long interconnect wires are executed in the upper metal layers. If we double the length and width of the chip then the length of the longest wires and the area of their drivers doubles too. The area of the lowest layer, however, is quadrupled and can therefore accommodate at least double the amount of drivers. This allows us to add a new layer to place still longer wires. These longer wires come on higher levels where the wires are wider. If the wires on a level $k+1$ are β longer, wider and taller than the wires on level k then the maximal amount of wires N_k on level k satisfies $N_k = N_0\beta^{-2k}$. This suggests that the number of wires $f(i)$ of length i should decrease at least as fast as $N_0 i^{-2}$. However, in actual chip layouts the number of long wires may decrease less fast than this inverse square of the wire length; empirical wire length distributions $f(i) = \lfloor c2^{-\lambda} \rfloor$ $(1 \leqslant i \leqslant L_{\max})$ and $f(i) \approx 0$ $(i > L_{\max})$ with $1.5 < \lambda < 2$ have been reported in [5]. To achieve logarithmic propagation delay we can estimate and bound the layout area occupied by the fattened wires as follows. Let C be the amount of area of the layout occupied by *non-wire components* such as transistors. Assuming that C is also the order of magnitude of the number of basic components like transistors or logic gates in the circuit we can reason as follows. Since the wires only serve to connect components we have $C \in O(\#\text{wires})$ in a connected layout. The components are assumed to have at most a limited t connections to attach wires, which we suppose to account also for the *fan-in* and *fan-out* of the interconnect wires. Therefore $C \in \Omega(\#\text{wires})$ and consequently $C \in \Theta(\#\text{wires})$. Since we are primarily interested in order of magnitude in the sequel, we are justified to use C interchangeably for the amount of area occupied by the non-wire components, the number of non-wire components and the number of wires. The maximal area occupied by the wires (and interwire distances) under (13a) is bounded by the available area:

$$\sum_{i=1}^{\sqrt{A}} f(i)\, ai^2 \approx A - C \ , \tag{16c}$$

where a is the constant quotient of width and length (the aspect ratio) of the connect wires as required by (13a). Using a simple theoretical argument and an experimental study of actual layouts [5] develops the following wire length distribution relationship:

$$f(i) = \lfloor ci^{-\lambda} \rfloor \ (1 \leqslant i \leqslant L_{\max}) \text{ and } f(i) \approx 0 \ (i > L_{\max}) \tag{17}$$

for a *normalization* constant c yet to be chosen. Here $L_{\max}$ is a constant related to the size of the array (rectangular chip) and the adequacy of the placement; and λ is a constant characteristic of the logic. Equation (17) is derived using "Rent's Rule" which states that the average number of terminals per complex of C elements (in units, modules, cards, gates etc.) is tC^p, where t is the number of connections per individual element and p is the Rent constant characteristic of the logic complex. The analysis goes by dividing a square array of cells into 4 equal square arrays recursively down until the individual areas are the individual elements of the original logic. On each level of the recursion the number of connections crossing boundary lines is determined using Rent's rule. This shows that $\lambda \approx 3-2p$. In [5] experimental results are given for some actual layouts placed using a hierarchical placement program: layouts for high-speed logic were p was found to be 0.75 and a layout for a hand calculator chip with $p=0.59$. For $1 \leqslant \lambda < 3$, equation (17) is of the form of the Pareto-Levy distribution; similar laws occur in contexts like word frequencies, noise in transmission channels etc. For additional discussion see [5]. Let furthermore the network be connected, so the maximal amount of area units C available to place the components is not greater than the number of wires plus 1. From (16c) and (17) we can estimate the maximal figure for the normalization constant c. For $\lambda \neq 3$:

$$c \approx \frac{(A-C)(3-\lambda)}{a(A^{(3-\lambda)/2}-1)} , \tag{18a}$$

and for $\lambda = 3$,

$$c \approx \frac{2(A-C)}{a \log A} . \tag{18b}$$

Consequently, for $\lambda \neq 1$ & $\lambda \neq 3$ by (16b):

$$C \approx \sum_{i=1}^{\sqrt{A}} f(i) \approx \frac{(A-C)(3-\lambda)(A^{(1-\lambda)/2}-1)}{a(1-\lambda)(A^{(3-\lambda)/2}-1)} . \tag{19a}$$

and for $\lambda = 3$,

$$C \approx \frac{(A-C)(A-1)}{aA \log A} . \tag{19b}$$

For $\lambda = 1$,

$$C \approx \sum_{i=1}^{\sqrt{A}} f(i) \approx \frac{A-C}{a(A-1)} \log A . \tag{19c}$$

(Note: for $\lambda < 1$ we obtain $c < 1$, resulting in $f(i) \approx 0$ also for small i, and C a small constant.) From the above analysis follows:

Lemma 1. *Let $f(i)$ be the wire length distribution function of a chip layout with area A. Let the signal propagation delay be logarithmic and let therefore all wires in the layout have the same aspect ratio. If $f(i) = \lfloor c / i \rfloor$ for some constant c then the total number of wires, and similarly the total number of gates and transistors, is* $O(\log A)$.

Since the number of components bounds the number of bits manipulated in each computation step, Lemma 1 tells us that this number is very much layout dependent, and depends

in particular on the distribution of wire lengths. Consequently, even if the area A is polynomial in the binary size N of a problem, under a layout and signal propagation delay as in the lemma, the execution time, and also the period for systolic computation will be $\Omega(N)$ since it takes at least $N / \log N$ stages just to scan all N bits and for each such stage the delay in the longest wires is $\Omega(\log N)$. In some designs, like trees for instance, the number of long wires decreases faster with the wire length. For such f the series in (16c) converges also faster, and the maximal number of wires, and similarly the maximal number of components, may rise to order A (the area expressed in area units), under the logarithmic propagation delay requirement notwithstanding the attendant constant aspect ratio for wires.

The constant wire width assumption. For comparison we give the analogous analysis with above under the constant wire width assumption. Then equations (16a) - (16b) stay the same but equation (16c) becomes

$$\sum_{i=1}^{\sqrt{A}} f(i)\, i \approx A - C \quad . \tag{20}$$

Thus, for $f(i) = \lfloor ci^{-\lambda} \rfloor$ $(1 \leq i \leq \sqrt{A})$ and $f(i) \approx 0$ $(i > \sqrt{A})$ and with A, C and c as above we obtain the following relations. For $\lambda = 1$:

$$c \approx \frac{A - C}{\sqrt{A} - 1} \tag{21}$$

$$C \approx \frac{(A - C)\log A}{2(\sqrt{A} - 1)} \quad .$$

For $\lambda \neq 1$ & $\lambda \neq 2$:

$$c \approx \frac{(2-\lambda)(A - C)}{A^{(2-\lambda)/2} - 1} \tag{22}$$

$$C \approx \frac{(2-\lambda)(A - C)(A^{(1-\lambda)/2} - 1)}{(1-\lambda)(A^{(2-\lambda)/2} - 1)} \quad .$$

For $\lambda = 2$:

$$c \approx \frac{2(A - C)}{\log A} \tag{23}$$

$$C \approx \frac{2(A - C)(\sqrt{A} - 1)}{\sqrt{A} \log A} \quad .$$

(Note: for $\lambda < 0$ we obtain $c < 1$.) For $\lambda > 0$ we have $C \in \Omega(\sqrt{A})$. Thus:

Lemma 2. *Let $f(i)$ be the wire length distribution function of a chip layout with area A under the constant wire width assumption. If $f(i) = \lfloor c / i^{\lambda} \rfloor$ $(1 \leq i \leq \sqrt{A})$ and $f(i) \approx 0$ $(i > \sqrt{A})$ for a constant c then the maximal feasible number of wires in the layout, and similarly the maximal number of gates and transistors in the layout, is $\Omega(\sqrt{A})$ for all $\lambda \geq 0$.*

Recall that the quotient of the length and width of a wire is its *aspect ratio.* By the previous analysis, considering just the wire length distribution while leaving free the actual circuit topology, placement and routing in the layouts, attaining a logarithmic signal propagation delay by changing constant wire width to constant aspect ratio for all wires in a

layout can carry a surprisingly severe penalty.

Theorem. *Let the original layout area be A and the original amount of wires in the layout be C. For the wire length distribution $f(i) = \lfloor ci^{-1} \rfloor$ for $1 \leqslant i \leqslant \sqrt{A}$ and $f(i) \approx 0$ for $i > \sqrt{A}$, the change from constant wire width to wires with a constant aspect ratio has the following effect.*

(i) *Retaining the original amount of wires C and the original wire length distribution relative to the layout area, that is, $f(i) = \lfloor c'i^{-1} \rfloor$ for $1 \leqslant i \leqslant \sqrt{A'}$ and $f(i) \approx 0$ for $i > \sqrt{A'}$ with the normalization constant c' set to its maximal value, exponentially increases the required area A' over the original area A.*

(ii) *Retaining the original area A and the original wire length distribution, that is, $f(i) = \lfloor c'i^{-1} \rfloor$ for $1 \leqslant i \leqslant \sqrt{A}$ and $f(i) \approx 0$ for $i > \sqrt{A}$ with the normalization constant c' set to its maximum value, reduces the maximal amount of wires in the layout, and therefore the proportionate amount of useful components like gates and transistors, to $C' \in O(\log C)$.*

(iii) *Retaining the original amount of wires C (or logic components) and the original area A requires a thorough change of wire length distribution to $f'(i) = \lfloor c'i^{-\lambda} \rfloor$ for $1 \leqslant i \leqslant \sqrt{A}$ and $f(i) \approx 0$ for $i > \sqrt{A}$ and the normalization constant c' set to its maximum value. Each $\lambda \geqslant 2 + \epsilon$ for some small $\epsilon > 0$ depending only on A and a suffices. For a given network topology this entails a placement and routing of the layout which may well be impossible in many cases.*

Proof. Since we assume the circuit to be connected we have $A > A - C > A/2$ in the various equations. We also assume $A \gg 1$.

(i) Equate expression (21) for C with expression (19c) for C, with A' substituted for A in the latter. This yields $\log A' \in \Omega(\sqrt{A})$.

(ii) Substitute C' for C in equation (21) and express C' in terms of C by eliminating A from the resulting equation and (19c).

(iii) Equate expression (21) for C with expression (19a) for C (expressions (19b) and (19c) contradict (21)). The terms $(A - C)$ on both sides cancel each other. Solving λ yields $\lambda = 2 + \epsilon(A, a) > 2$ with $\epsilon(A, a) \to 0$ for $A \to \infty$ and a constant. Every distribution with exponent equal or larger than this λ suffices. □

We observe that in case (i) of the Theorem the wires get so long that the logarithmic propagation delay turns out to yield about the same absolute time delay as in the original wires. In case (ii) of the Theorem matters are probably as bad because the bit capacity of the chip has been logarithmically reduced. Finally, in case (iii) of the Theorem the subject circuit topology may not have a layout with the required wire length distribution.

For values of the exponent $\lambda > 1$ in the wire length distribution the analog of the Theorem holds with polynomial relationships in cases (i) and (ii). For $\lambda > 3$, the number of long wires decreases so fast with the wire length that it suffices that $A' \in \Theta(A)$ in (i), $C' \in \Theta(C)$ in (ii) and nearly the same wire length distribution function suffices in (iii). However, $\lambda > 3$ implies a negative Rent constant p since $\lambda \approx 3 - 2p$ in [5], and therefore layouts which do not satisfy Rent's Rule. The reader is invited to analyse the relations for different values of λ. We look at one more case, the wire length distribution with $\lambda = 2$, which is interesting because it is associated with hierarchical designs. For $\lambda = 2$, the different parts of the Theorem yield the following:

(i) $A' \in \Omega(A^2 / \log^2 A)$.

(ii) $C' \in O(\sqrt{C} \log C)$.

(iii) This requires a change to a new distribution function $f'(i) = \lfloor c' i^{-\lambda} \rfloor$ $(1 \leqslant i \leqslant \sqrt{A})$ and $f'(i) \approx 0$ $(i > \sqrt{A})$ with the new normalization constant c' set to its maximum value. Each $\lambda \geqslant 3$ suffices.

Network topologies, which can be realized with constant width wires, may not be realizable at all on a multilevel Manhattan grid geometry with all wires having the same aspect ratio. Even for network topologies which do have a layout with a constant aspect ratio for the wires, the Theorem shows that the increase in area can be so much that the amplification in length of the wires will nullify (or worse) the increase in speed due to a change from linear or square propagation delay to a logarithmic one. It therefore appears that but circuits with proper topologies (like the tree circuit in the previous section), for which there are layouts with the considered wire length distributions with relative large λ, are proper candidates for an improvement of speed by a logarithmic signal propagation delay.

Exercise. Do a similar analysis for *radical delay.*

References

[1] Bilardi, G., M. Pracchi, and F.P. Preparata, "A critique and an appraisal of VLSI models of computation," pp. 81-88 in VLSI Systems and Computations, ed. H.T Kung, B. Sproull & G. Steele, Springer verlag, Berlin (1981).

[2] Brent, R.P. and H.T. Kung, "The chip complexity of binary arithmetic," pp. 190-200 in Proceedings 12th ACM Symposium on Theory of Computing (1980).

[3] Brent, R.P and H.T. Kung, "The area-time complexity of binary multiplications," *J. Ass. Comp. Mach.*, vol. 28, pp.521-534, 1981.

[4] Chazelle, B. and L. Monier, "A model of computation for VLSI with related complexity results," pp. 318-325 in Proceedings 13th ACM Symposium on Theory of Computing (1981).

[5] Donat, W.E., "Wire length distribution for placement of computer logic," *IBM J. Res. Develop.*, vol. 25, pp.152 - 155, 1981.

[6] Leighton, F.T., *Complexity Issues in VLSI.* The MIT Press, 1983.

[7] Leiserson, C.E., *Area Efficient VLSI Computation.* The MIT Press, 1982.

[8] Mead, C. and M. Rem, "Cost and Performance of VLSI Computing structures," *IEEE J. on Solid State Circuits,*, vol. SC-14, pp.455 - 462, 1979.

[9] Mead, C. and L. Conway, *Introduction to VLSI Systems.* Reading, Mass.:Addisson-Wesley, 1980.

[10] Mead, C. and M. Rem, "Minimum propagation delays in VLSI," *IEEE J. on Solid State Circuits*, vol. SC-17, pp.773 - 775, 1982. Correction: *Ibid*, SC-19 (1984) 162.

[11] Paterson, M.S., W.L. Ruzzo, and L. Snyder, "Bounds on the minimax edge length for complete binary trees," pp. 293 - 299 in Proceedings 11th ACM Symposium on Theory of Computing (1981).

[12] Ramachandran, V., "On driving many long lines in a VLSI layout," pp. 369 - 378 in Proceedings 23rd IEEE Symposium on Foundations of Computer Science (1982).

[13] Ramachandran, V., "Upper bounds for the area increase caused by local expansions in a VLSI layout," pp. 163-179 in Advances in Computer Research (1984).

[14] Saraswat, K.C. and F. Mohammadi, "Effect of scaling of interconnections on the time delay of VLSI circuits," *IEEE J. of Solid State Circuits*, vol. SC-17, pp.275-280, 1982.

[15] Savage, J., "Planar circuit complexity and performance of VLSI algorithms," pp. 61-68 in Proceedings CMU Conference on VLSI Systems and Computations, ed. H.T. Kung et. al., Computer Science Press (1981).

[16] Seitz, Ch.L., "Ensemble architectures for VLSI - A survey and taxonomy," pp. 130-132 in Proc. of MIT Conference on Advanced Research in VLSI, ed. P. Penfield, Jr., Artech House (1982).

[17] Thompson, C.D., "A Complexity Theory for VLSI", Ph.D. Thesis, Dept. of Computer Science, Carnegie-Mellon University, 1980.

[18] Thompson, C.D. and P. Raghavan, "On estimating the performance of VLSI circuits", Tech. Rep. UCB/CSD 84/138, Computer Science Division (EECS), University of California, Berkeley, September 1983.

[19] Thompson, C.D., "Fourier transforms in VLSI," *IEEE Transactions on Computing*, 1984.

[20] Ullman, J.D., *Computational Aspects of VLSI.* Rockville, Maryland:Computer Science Press, 1984.

[21] Vuillemin, J., "A combinatorial limit to the computing power of VLSI circuits," pp. 294-300 in Proc. 21th IEEE Symposium on the Foundations of Computer Science (1980).

[22] Yuan, H.-T., Y.-T. Lin, and S.-Y. Chiang, "Properties of silicon, sapphire, and semi-insulating gallium arsenide substrates," *IEEE J. of Solid State Circuits*, vol. SC-17, pp.269-274, 1982.