Towards More Realistic Models of Computation for VLSI

B.M. Chazelle L.M. Monier

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

We propose two new models of computation for VLSI which take into consideration the physical nature of information, the properties of wires, and the geometrical structure of the circuit. Both are refinements of the Kung-Thompson model, and make the main additional assumption that the propagation time of information is at best linear in the distance. The first is the more general and applies for any planar technology. It is in a sense the *minimal physical* model. The second, more restrictive, is specially tailored for electrical technologies. Our approach is motivated by the failure of previous models to allow for realistic asymptotic analysis. For each model, we are able to show new lower bounds and trade-offs for many well-known problems.

1. Introduction

The importance of having general models of computation for VLSI is apparent for various reasons. Among the chief ones, we must include the need for evaluating and comparing circuit performances, showing lower bounds and trade-offs on area, time, and energy, and more generally building a complexity theory of VLSI computation.

While these models must be simple, general enough, to allow for mathematical analysis, they must also reflect reality independently of the size of the circuit. We justify the latter claim by observing that if 1980's circuits are still relatively small, the use of high-level languages for designing chips, combined with the possibility of larger integration and bigger chips, will make asymptotic analysis necessary in the near future.

Yet as circuits are pushed to their physical limits, constraints which could be ignored before become major problems and must be accounted in the models. In particular, certain physical phenomena specific to electrical technologies enforce the density of current at any point of a conductor to be bounded. We can show that this invalidates the assumptions made in previous models, whereby long wires can be driven in constant time and an f-branch fanout takes O(log f) time [MC80,TH79].

Generally speaking, one major flaw in those previous models is to regard a circuit as a topological interconnection of nodes where transmission delays between adjacent nodes can be ignored. Instead, we propose to take into account the geometry of the circuit by assuming a propagation delay linear in the distance. We can justify this approach by considering parameters such as length and width of wires, and associating resistance and capacitance with each part of the circuit. We will define a first model which does not make further assumptions, and we will review the complexity of some well-known circuits in this model. However, observing that in NMOS technology, the power can be supplied only from the outside boundary of the circuit, we can include this requirement and define a second model, which may be more realistic for electrical planar technologies.

Also, besides presenting new models of computation for VLSI, the purpose of this paper is to present a general technique for deriving lower bounds and space-time trade-offs for many problems, e.g., addition and transitive functions.

2. The Models

2.1. The basic assumptions

Our models are for the most part refined versions of the current planar models found in the literature [TH79, BK80, VU80]. A circuit consists of nodes and wires connected in a network, and it is defined by a geometrical layout of this network. We distinguish I/O nodes where input and output values are available, the logical nodes (gates) which compute boolean functions, and the connection nodes which simply connect

wires. The circuit is laid out within a convex region with all the I/O nodes lying on its boundary. It is the case today, and will remain true because of the greater ease in connecting and packaging such chips. In addition, we make the following set of assumptions, which define our first model (MOD1).

1. Wires have width and spacing between them greater than λ (today $\lambda \approx 1\mu m$). This requirement will always be valid for any physical device.

2. A circuit is laid out on a finite number of layers, and wires crossing through different layers are allowed. Thus there is at most a constant number of cross-overs at any point.

3. The density of current at any point of a wire is bounded by a maximum value δ_{max}, which is equivalent to saying that the power dissipated per unit volume is also bounded. The major consequence of this assumption is to make propagation delays at least linear in the distance.

4. To switch a gate requires a minimum energy dissipated as heat [MC80,Ch.9]; this energy must be supplied to the gate by a source other than the input signals.

To take into account the limitations in driving power enforced by NMOS and to a lesser extent CMOS technology, we introduce a second model (MOD2), which in addition to MOD1, includes the following assumptions.

1. All the energy supplied to the circuit comes from outside the circuit, and its transmission is performed only through wires. From 3, it follows that the maximum power provided to the circuit is at most proportional to the perimeter of the circuit.

2. Storing a bit of information requires a minimum energy per unit of time.

Note that since this model is more restrictive than the previous one, all the lower bounds obtained for MOD1 are still valid in MOD2.

2.2. Coding information

The information at a point is given by the value of an electrical parameter at this point, which we define as the potential of a capacitor. While electrical computations are essentially analog processes, the coding of information is made digital by assuming a **0** for a potential less than V_0 and a **1** for a potential greater than V_1 ($V_1 > V_0$).

2.3. Wires

A wire is a rectangular parallelepiped made of conducting material, oriented by the direction of the current. It is characterized by its length L, its width W, its thickness H, and its distance D from a plane of reference (the substrate). Its resistance R and its capacitance C are given by the (idealized) relations

$$R = \rho \times L/(W \times H) \qquad C = \varepsilon \times L \times W/D$$

where ρ and ε are technology-dependent coefficients.

Minimum values for L,W and H are set by the technology (as well as by the laws of physics), and we

require D to be constant for any wire. Moreover it is legitimate always to assume bounded thickness. Indeed a current density δ causes a heat power loss in the wire proportional to $L \times W \times H \times \delta^2$, but the dissipated power is proportional to $L \times W$, since the circuit is planar. For allowing this heat to be dissipated, the thickness H must remain within constant bounds. Thus we can assume that the resistance is simply proportional to L/W and the capacitance to $L \times W$.

2.4. Nodes

We distinguish three kinds of nodes, each of which uses up a minimum constant area.

- *Connection nodes*: Their purpose is to provide electrical contacts between a bounded number of wires. These contacts may either connect wires on a same layer, or they may be "vertical contacts" between different layers. Of course, they introduce no delay and do not dissipate any energy.

- *I/O ports*: They ensure the exchange of information between the circuit and the outside world. The locations and the order in which input (resp. output) bits are to be written (resp. read) are fixed and independent of the values of these bits. We restrict each input bit to be available on the input port only once. This implies that the repeated use of the same input bit necessitates its storage within the circuit. The transmission of an information signal through an I/O port introduces a constant delay.

- *Gates*: Conceptually, a gate is the device used to compute a logical function of one or two inputs and one output. Since it can be shown that there is no interest in having gates of arbitrary size, we assume that all gates have the same size. Physically we must associate a gate capacitance with each input. An input is valid as soon as the corresponding gate capacitor has been set above or below a certain threshold potential. The value of the function is given by the potential of the output device of the gate. Once the output is available, it cannot be destroyed before a constant lapse of time, whatever the input changes occurred in the meantime.

2.5. Current density

Proposition 1: The density of current is bounded at any point of a conductor by a maximum value δ_{max}.

One major flaw in previous models is to suppose that a wire of constant width can drive a current of arbitrary intensity. We can list at least three reasons in present-day technologies which justify Proposition 1.

1. Any conductor with non-zero resistance produces a power per unit volume proportional to the *square* of the current density. Since this power can be dissipated only through the boundary of the conductor, the heat dissipation is at most proportional to the area of the conductor, which implies a bounded density.

2. An electrical phenomenon called *metal migration* [MC80, CL80] causes a current to destroy the conductor all the more quickly as the density is high. For this reason, a maximum admissible density of current can be assigned to any conducting material.

3. The voltage drop per unit length is proportional to the density of the current. Since we must ensure that the logical value of the signal provided by power wires is the same at any point of the circuit, this voltage drop must remain small, and thus the density must be bounded.

For example, the aluminium currently used in NMOS technology has a maximum density imposed by metal

migration of about 10^9 A/m^2 or only 1 mA/μm^2. For this density the voltage drop is 30 V/m with a resistivity of about 3.10^{-8} $\Omega\times$m. Note that the voltage drop on a 3 mm wire is 0.1 V, and is far from negligible. Also, the power induced in the wire by such a density of current is about 3 W/cm^2, if the thickness is 1μm.

3. Transmitting Information

We turn to the problem of transmitting an information bit from a point A to a point B at a (Euclidian) distance L apart. We will assume that this information will be carried through an arbitrary path from A to B consisting of nodes and wires.

We first consider the case where the path consists of a wire followed by a gate. Let $S=H\times W$ be the section of the wire. In order to transmit a bit of information, we must raise the wire to the required voltage. The charge Q on the wire is therefore proportional to its capacitance, that is, $L\times W$. Since in a time T a density of current crossing a section S can provide at most a charge $\delta_{max}\times S\times T$, the assumption that H is bounded yields the relation $T=\Omega(L)$.

We next investigate the case where two paths of the previous type are cascaded. Since the first gate cannot be switched before the signal becomes available on the first wire, the total delay will amount to the added delays of the two paths augmented by the switching time of the first gate. This also results in an $\Omega(L)$ delay. The last case to examine involves two wires linked by a connection node. We can apply the reasoning used in the case of a single wire, with W now being the maximum of the two wire widths. The same result follows directly. In the general case, we can decompose an arbitrary path from A to B into components of the form previously examined. Putting the above results together permits us to find the claimed lower bound on the time.

In addition, we should notice that some energy is dissipated along the wire during the propagation of information since the wire has a non-zero resistance. This energy is proportional to the charge involved, which is $\Omega(L)$ in any configuration. Observe that this energy is independent of the time T.

Both results permit us to state the following.

> **Theorem 2:** Transmitting a signal between two points at a distance L apart requires $\Omega(L)$ time and $\Omega(L)$ energy.

Note that this lower bound cannot be achieved with a simple wire: because of the diffusion law [MC80,SE79], the actual delay is in fact proportional to $R\times C = L^2$. However, we can reduce this delay to O(L) by using O(L) wires of constant length connected by O(L) gates (e.g., inverters or amplifiers). If the wires have minimum width, the lower bound O(L) on the energy is also achieved. Note that a simple speed-of-light argument yields the same result for any technology. This is precisely what makes MOD1 a minimal planar model for all physical computations.

4. Distributing and Collecting Information

Throughout this section, we will assume that the model is MOD1 or MOD2, indifferently. To fan-in or fan-out information being two of the most common operations performed by circuits, we next turn to these problems, from which we can best measure the significant departure of our models from previous ones. For simplicity, we first prove a technical lemma.

Lemma 3: There is a constant c ($c=1/2\pi$) such that for any convex polygon with a boundary of length N and for any point M, there exists a vertex v such that $\text{dist}(v,M)\geq cN$.

We omit the proof, which is straightforward.

4.1. Fan-out

A fan-out of degree N refers to the distribution of an information bit from a source to N points (gates or ports) on the circuit. To study the complexity of this problem, we distinguish two cases; when the N points lie on a convex boundary (e.g., on the boundary of the circuit), and when their location is left arbitrary. We define T (resp. E) to be the minimum time (resp. energy) to perform a fanout of N points. It is trivial to see that $E=\Omega(N)$in both cases, since to reach every node, the information must cross a wire of (at least) unit length. As for the time T, we have two different results.

Theorem 4: If the N points lie on a convex boundary, $T=\Omega(N)$.

Proof: It follows from Lemma 3 that one of the N destinations is at least cN apart from the source, and Theorem 2 permits us to conclude. □

Theorem 5: If the N points have arbitrary locations, $T=\Omega(N^{1/2})$.

Proof: A consequence of the fact that the maximum distance between N points and an arbitrary point in the plane is at least $cN^{1/2}$, for some constant c. □

Note that all these lower bounds are tight, as shown in Figure 4-1.

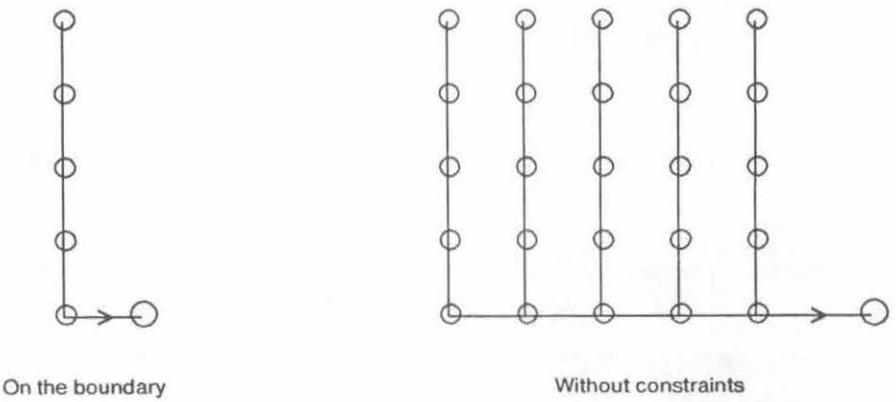

Figure 4-1: Optimal fan-out.

4.2. Fan-in

The fan-in is essentially the reverse operation of the fan-out, since N information bits must converge from N sources to one destination point. Yet it is a little more general, since the information may be submitted to logical operations on its way to the destination. Typically, the problem is to compute a boolean function of N inputs and one output. Since every gate is followed by a wire of unit length at least, the minimum energy dissipated during the operation is $E=\Omega(N)$. If the N inputs are valid at the same time, the results are the same as for the fan-out. In the more general case where pipelining is allowed and the inputs are valid at arbitrary times, we can show the following.

Theorem 6: If T (resp. A) denotes the minimum time (resp. area) for computing a boolean function of N inputs, we have $T=\Omega(N^{1/2})$ and $AT=\Omega(N)$.

Proof: Let p denote the total number of input ports actually used. It takes time at least proportional to N/p to read all the inputs, and since the p ports lie on a convex boundary, Lemma 3 and Theorem 2 show that $T=\Omega(p)$. Observing that $A=\Omega(p)$, the result is then immediate. □

Note that these lower bounds are still valid for boolean functions with an arbitrary number of outputs, as long as at least one output depends on all the input values. The addition for example falls in that category, since the last carry depends upon all the operand bits. If the boolean function is a commutative, associative, operation on N variables, these lower bounds are tight, as shown in Figure 4-2.

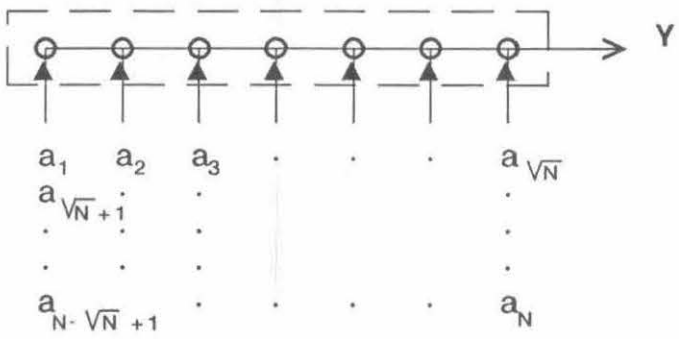

Figure 4-2: Computing $Y=a_1$ *op* a_2... *op* a_N takes $\Omega(N^{1/2})$ time and area.

5. New Lower Bounds for some Common Problems

5.1. Addition

Since our models relate the time of computation to the geometry rather than the topology of the circuit, we can show that many complete binary tree based schemes cease to have the logarithmic time complexity which they enjoy in previous models. Notable examples include the fan-in and fan-out operations studied earlier, or the addition of two N-bit integers, to which we next turn our attention. We study this problem in our two

models in turn. For simplicity, we start the analysis with the model MOD2, and present the basic arguments.

5.1.1. Case MOD2

Theorem 7: If T is the time required in MOD2 by any circuit to add two N-bit integers, and if A is the area of the circuit, we have

$$1)\ AT = \Omega(N) \qquad 2)\ T = \Omega(N^{2/3}).$$

Proof: For the sake of simplicity, we will assume in this proof that the sign "=" really means "equals to within a constant factor". Relation 1) follows directly from the fact that adding two N-bit integers involves a fan-in of degree N. To prove 2), let's call X one of the operands and Y the result of the addition. Since we can always assume that low order bits are read first, we can rewrite X as $X_p...X_2X_1$, where X_i are the bits of X read at time t_i, with $t_1<...<t_p<T$. X_i denoting both the chain of bits and its length, we have the relations

$$(1)\ X_1+...+X_p=N \qquad (2)\ T\geq p.$$

Let X(t) be the total number of bits of X read so far at time t, and let Y(t) be the total number of result bits output in this interval of time. Since at time t, the total number of possible values for the remaining output bits is at least $2^{N-Y(t)}$, and only N-X(t) bits of X remain to be read, the circuit must have at least X(t)-Y(t) active gates at time t. This requires a circuit perimeter $\Pi\geq X(t)-Y(t)$ and a time Π, hence the relations

$$(3)\ X(t)-Y(t)\leq\Pi \qquad (4)\ T\geq\Pi.$$

Since low-order input bits are read first and a fan-in on k bits takes $\Omega(k)$ time, at least $N-X(t_i)$ output bits remain to be computed at time t_i+X_i. We can give a geometric interpretation of this relation as shown in Figure 5-1. Relation (3) implies that the endpoints of the intersection of the shaded area of Fig.5-1 with a vertical line are at most Π apart. It follows that the shaded area must lie within the strip (L_1,L_2), which in turn implies

$$X_1^2+...+X_p^2\leq T\Pi$$

Since $X_1^2+...+X_p^2$ is minimal when all the X_i's are equal, we derive the relation $N^2/p\leq T\Pi$, which combined with relations (2) and (4) yields

$$T\geq N^2/p\Pi+p+\Pi$$

The minimum of the right-hand side is achieved for $p=\Pi=N^{2/3}$, which concludes the proof. □

Note that the lower bound on AT is trivially tight, since there exist linear-time constant-area adders. We do not believe that this is the case with the lower bound given for T. We conjecture that $T=\Omega(N)$ is the actual lower bound in this model, which would make the simplest adder in the world asymptotically optimal.

5.1.2. Case MOD1

It is natural that lower bounds obtained in MOD1 should be weaker than in MOD2. However, MOD1 has the merit of greater generality, and any lower bound in this model is thus very interesting.

Theorem 8: If T is the time required in MOD1 by any circuit to add two N-bit integers, and if A is the area of the circuit, we have

$$T=\Omega(N^{1/2}), \qquad AT=\Omega(N), \qquad AT^2=\Omega(N^2).$$

Proof: The first two relations result from the fact that adding two N-bit integers involves a fan-in of degree N. Indeed the last carry is a fan-in of all the input bits. We can prove the last relation with the same technique used above. Keeping the same notation, we find that $X(t)-Y(t)\leq A$, since at any time t the number of bits stored in the circuit is at least X(t)-Y(t). On the other hand, Y(t) always lies below the shaded area of Fig.5-1. It then follows that total area of the shaded region

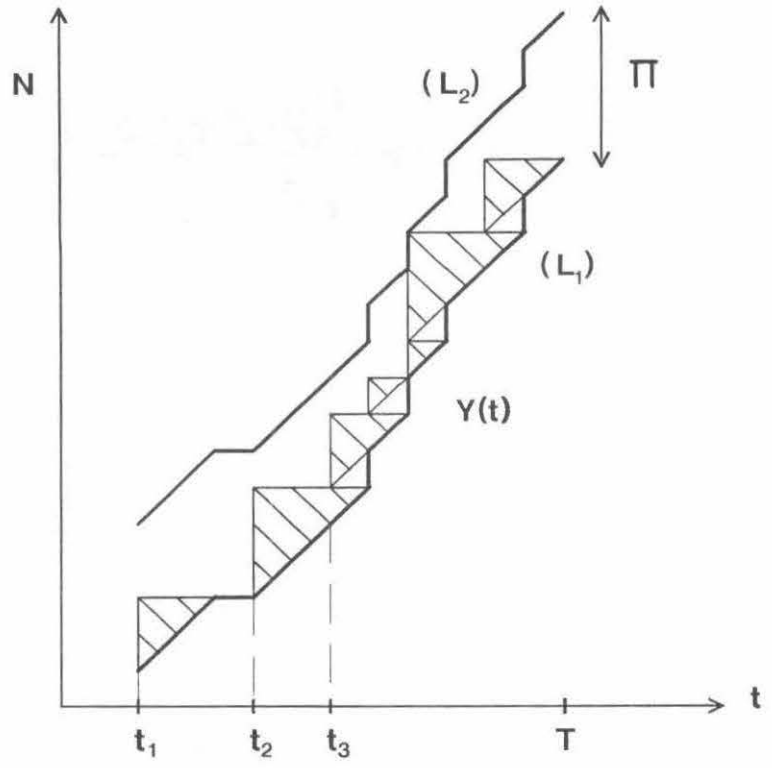

Figure 5-1: The $\Omega(N^{2/3})$ time lower bound on integer addition.

cannot exceed the area of the parallel strip (L_1,L_2), hence

$$X_1^2+...+X_p^2 \leq AT.$$

The minimum is achieved for $X_j = N/p$, and since the time for reading the data is proportional to p, we find $AT^2 = \Omega(N^2)$, which completes the proof. □

5.1.3. Optimal adders in model MOD1

A fortunate feature of addition in model MOD1 is to allow the possibility of matching all the lower bounds derived above. We will describe a class of adders which satisfy these properties.

Serial Adder: The simplest adder requires constant area, operates in linear time, and thus matches the lower bound for the measures AT and AT^2. The scheme of this adder is represented in Fig.5-2.

CLA Adder: Assuming wlog that N is a power of two, we implement the CLA scheme on a complete binary tree with N leaves. The operand bits are read in parallel at the leaves, and the time of computation is at least the time for propagating a signal across the longest path in the tree. It follows that the layout of Fig.5-3 requires O(N) time and O(NlogN) area.

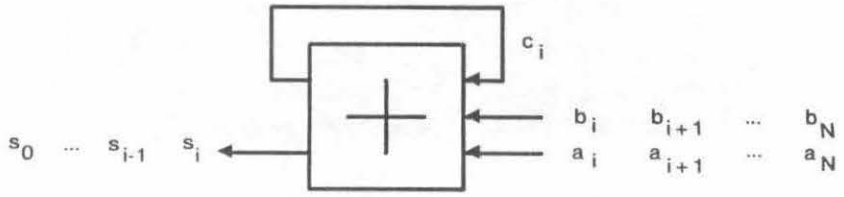

Figure 5-2: The Serial Adder.

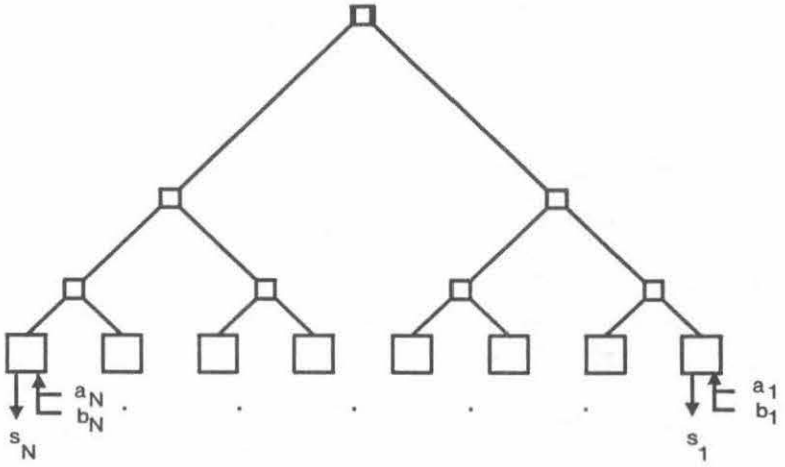

Figure 5-3: The CLA Adder.

If the technology allows the packing of k information bits on a square of area O(k) (e.g. that excludes NMOS), an alternate layout may use the H-embedding of a binary tree, as shown in Fig.5-4. The operands may be driven from the input ports to the leaves of the tree in about $N^{1/2}$ waves of $2N^{1/2}$ bits. Unfortunately, each wave consists of a complicated (but fixed) sequence of input bits. If we do not account for the task which arranges the input bits in the proper order, and if we use inverters to avoid long wires (see Section 3) adding two N-bit integers simply takes $O(N^{1/2})$ time and linear area, which matches the lower bounds obtained for T and AT^2.

Mixed CLA Adder: In some applications the size of the operands greatly exceeds that of the circuit, and only, say, N^{α} input ports are available. In this case, we can divide the operands into roughly $N^{1-2\alpha}$ groups of $N^{2\alpha}$ bits, and compute the addition for each group with a CLA adder of area $N^{2\alpha}$, transmitting the carry for the next addition every time around. The total time of computation will thus be $O(N^{1-\alpha})$, with a circuit of area $O(N^{2\alpha})$. Note that the lower bound $AT^2 = N^2$ is still matched with this scheme. Also, we observe that for $\alpha = 1/2$ we have the CLA adder, whereas setting α to zero reduces to the serial adder.

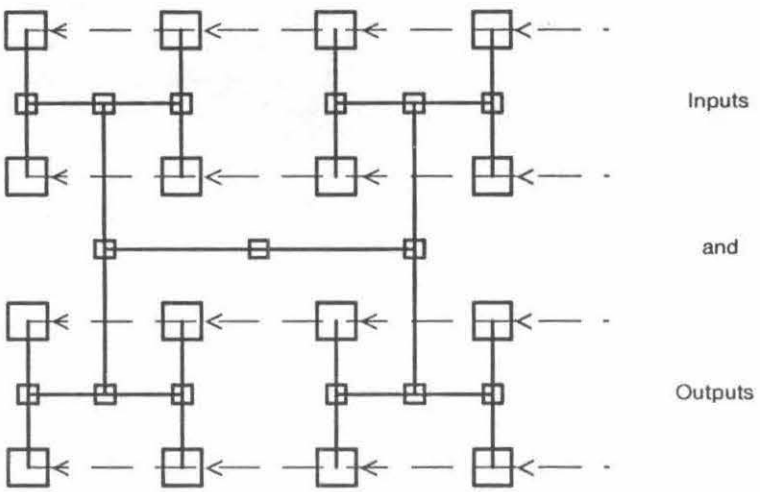

Figure 5-4: Optimal layout of the CLA adder.

5.2. Transitive Functions

In a recent paper [VU80], J. Vuillemin has shown that the transitivity of a function has heavy consequences on its complexity in a VLSI model. Roughly speaking, a function is said to be transitive of degree N if it computes a transitive group of permutations acting on N elements. This implies that the function can map any input bit onto any output bit for an appropriate value of the other inputs. Such functions include cyclic shifts, integer products, convolutions, linear transforms, and some matrix products.

5.2.1. Case MOD1

Even in our more general model, we can show a significant difference with previous results [PV79,VU80].

Theorem 9: Computing a transitive function of degree N takes time $T=\Omega(N^{1/2})$.

Proof: Let p be the number of output ports actually used. Since an input bit can be mapped onto any output port, Lemma 10 shows that for some value of the inputs, the computation will take time at least proportional to p. On the other hand, observing that it takes time at least proportional to N/p to output the result completes the proof. □

It is worthwhile to notice the serious gap existing between this model and the previous ones, which allowed for logarithmic time for computing transitive functions (e.g. the CCC-scheme [PV79]).

5.2.2. Case MOD2

It comes as no surprise that since our second model adds physical constraints to the one in which Vuillemin derived his lower bounds, we can significantly improve upon his results. Before proceeding, we will establish a preliminary result.

Lemma 10: If N gates in a circuit are switched at the same time, their convex hull has a perimeter $\Omega(N)$.

Proof: Since all the power comes from outside the circuit and is transmitted through wires, the power inside any convex region of the circuit is at most proportional to its perimeter. Switching a gate requiring a minimum energy, the result is straightforward. □

We can now prove our main result.

Theorem 11: Any circuit of area A which computes a transitive function of degree N in time T satisfies $A=\Omega(N)$, $T=\Omega(N)$.

Proof: It has been shown in [VU80] that the circuit must have the capability of memorizing N bits. Therefore Lemma 10 implies that the circuit must have two active gates G_1 and G_2 at a distance $\Omega(N)$ apart, hence $A=\Omega(N)$. We can always assume that for some values of the inputs, information will be transmitted from G_1 to an output port P_1 (same with G_2 and an output port P_2). Consider now an arbitrary input port R. Since the function is transitive, there exists a path in the circuit from R to P_1 and from R to P_2. Among all possible computations, the four paths G_1-P_1, G_2-P_2, R-P_1, and R-P_2 will be used at least once. From Theorem 2, it then follows that T is at least proportional to $\text{Max}\{G_1P_1, G_2P_2, RP_1, RP_2\}$. The sum of these four lengths is greater than $G_1G_2=\Omega(N)$ -See Fig.5-5-, which concludes the proof. □

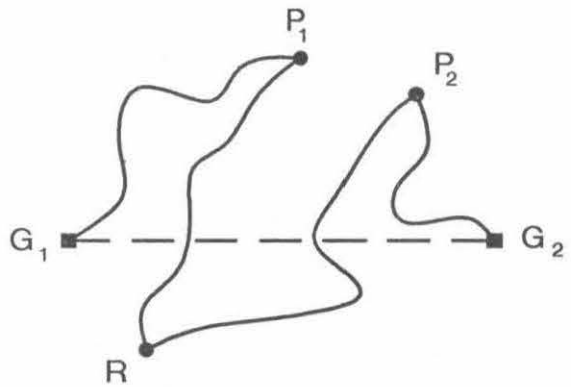

Figure 5-5: Computing a transitive function requires linear time.

Remark: In MOD2, these lower bounds are tight for some problems; for example optimal circuits for performing integer multiplication, based on the Shift&Add scheme, can be found.

6. Conclusions

The major contribution of this paper has been to show how previous models fail to allow for asymptotic analysis. We have proposed two models of computation which are more realistic yet fairly simple. Since our models are essentially geared towards asymptotic analysis, previous models may turn out to be more accurate for circuits of small size. For example, the carry-look-ahead scheme for adding two N-bit integers actually requires at least $\Omega(N^{1/2})$ time in our models instead of the well-known logarithmic time, but it may still be superior to any naive circuit for small integers.

Further refinements of these models should be valid independently of size considerations, and should allow for *à la Knuth* analyses of VLSI circuits. It is still difficult to think of a technology-independent model at the present time. But it may be a prerequisite for building a complexity theory which faithfully reflects reality.

Acknowledgments

We wish to thank Jean Vuillemin for suggesting this research and Mike Foster for many fruitful discussions. Our thanks also go to H.T. Kung and Gerard Baudet, who shared our interest in this work.

References

[BK80] R.P. Brent and H.T. Kung, *The Chip Complexity of Binary Arithmetic*, proc. 12th Annual ACM Symposium on Theory of Computing, **ACM**, pp. 190-200, May 1980.

[CL80] W.A. Clark, *From Electron Mobility to Logical Structure: A View of Integrated Circuits*, Computing Surveys, Vol.12, No 3, September 1980.

[MC80] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

[PV79] F.P. Preparata and J. Vuillemin, *The Cube-Connected-Cycles: A Versatile Network for Parallel Computation*, proc. 20th Annual Symposium on Foundations of Computer Science, Oct. 1979.

[SE79] C.L. Seitz, *Self-timed VLSI Systems*, proc. of Caltech Conf. on VLSI, 1979.

[TH79] C.D. Thompson, *Area-Time Complexity for VLSI*, proc. 11th Annual ACM Symposium on Theory of Computing, **ACM**, pp. 81-88, May 1979.

[VU80] J. Vuillemin, *A Combinatorial Limit to the Computing Power of V.L.S.I. Circuits*, proc. 21st Annual Symposium on Foundations of Computer Science, Oct. 1980.