

straight vertical connection between two points of V in one column. Charge $\sqrt{M} - 1$ (for the length of the vertical line segment) and 1 (for the cell q occupies), hence a total of \sqrt{M} cells to q . (Note that Steiner points on this vertical line segment can only have outgoing h -edges beside the v edges now accounted for, and no other leaf can be charged the same cells as q .)

Case b): the h edge marked by q has no other marks.

It means that this edge can be uniquely assigned to q and we can again charge $\sqrt{M} - 1$ (for the horizontal line segment) and 1 (for the cell of q), hence a total of \sqrt{M} cells to q .

Case c): the h edge marked by q has other marks as well.

Note that in this case, the h edge necessarily ends in a Steiner cell, with one outgoing v edge continuing on to q over a path of further v edges. Say (without loss of generality) that the path leads from the Steiner cell downwards. The only possibility for the h edge to be marked by another leaf as well is that there is a cell of V in the same column reached from the Steiner cell by going upward. Thus we conclude that the h edge can only be marked by one other cell of V , that necessarily lies in the same column as q and is vertically connected to it. Now charge the usual \sqrt{M} cells to a (for the h edge) and \sqrt{M} cells to the other leaf for the vertical line segments.

It follows (by carrying out this procedure for cells of V at increasing distance from the root) that all $l - 1$ cells of V beside the root can be charged a unique set of \sqrt{M} cells. Hence we obtain $l_s \geq (l - 1)\sqrt{M} + 1$.

We now complete the proof of Theorem 5.5 as follows. By Claim 5.5.1 we need l conflict-free fetches to retrieve T . By Claim 5.5.2 we have $t \geq l_s \geq (l - 1)\sqrt{M} + 1$, hence $l \leq \lfloor (t - 1)/\sqrt{M} \rfloor + 1$. Thus we can retrieve T by means of at most $\lfloor (t - 1)/\sqrt{M} \rfloor + 1$ fetches. \square

By choosing for M a square close to N , the following result is immediate. (Take, e.g., $M = \lfloor \sqrt{N} \rfloor^2$.)

Corollary 5.6: There is a linear skewing scheme using no more than N memory banks, such that every rookwise connected template of N cells in an $N \times N$ matrix can be retrieved in at most $\sqrt{N} + 2$ conflict-free fetches.

For arbitrary, connected templates T (including, e.g., diagonals) a precise analysis as in Theorem 5.5 is hard, but the following somewhat weaker bound can be obtained.

Theorem 5.7: Using s to store an $N \times N$ matrix into M memory banks, any connected template of t cells can be retrieved by means of at most $\lfloor 2 \cdot t/\sqrt{M} \rfloor + 1$ conflict-free fetches.

Proof: Follow the same argument as in Theorem 5.5 until after Claim 5.5.1. To estimate l we now reason as follows. Enclose every cell of V by a "box" of cells that are at most $\lceil 1/2 \cdot \sqrt{M} \rceil$ away from it, measured in cells along a connected (but not necessarily rookwise connected) chain. Note that the boxes indeed are squares, and that the boxes thus surrounding the cells of V are all disjoint. Assuming $l > 1$, the connectedness of T requires that in every box so distinguished there is a chain of cells leading from the middle cells to the boundary. This accounts for at least $\lceil 1/2 \cdot \sqrt{M} \rceil$ cells of T per box, hence $t \geq l \cdot \lceil 1/2 \cdot \sqrt{M} \rceil$ and $l \leq \lfloor 2 \cdot t/\sqrt{M} \rfloor$. The bound stated in the theorem is thus correct, including the case that t is small yet $l = 1$. \square

Choosing again $M = \lfloor \sqrt{N} \rfloor^2 (N)$ it follows that every connected template of N cells in an $N \times N$ matrix can be retrieved in at most $2\sqrt{N} + O(1)$ conflict-free fetches, using the linear skewing scheme s .

REFERENCES

- [1] P. Budnik and D. J. Kuck, "The organization and use of parallel memories," *IEEE Trans. Comput.*, vol. C-20, 1566-1569, 1971.
- [2] L. Euler, "Recherches sur une nouvelle espèce des quarrés magiques," *Verh. Zeeuwisch Gen. Wetensch.*, Vlissingen, vol. 9, pp. 85-239, 1782.
- [3] M. Hanan, "On Steiner's problem with rectilinear distance," *SIAM J. Appl. Math.*, vol. 14, 255-265, 1966.
- [4] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed. Oxford: Clarendon 1979.

- [5] A. Hedayat, "A complete solution to the existence and non-existence of Knut Vik designs and orthogonal Knut Vik designs," *J. Combin. Theory*, Ser. A, 22, pp. 331-337, 1977.
- [6] A. Hedayat and W. T. Federer, "On the non-existence of Knut Vik designs for all even orders," *Ann. Stat.*, vol. 3, pp. 445-447, 1975.
- [7] F. K. Hwang, "On Steiner minimal trees with rectilinear distance," *SIAM J. Appl. Math.*, vol. 30, pp. 104-114, 1976.
- [8] D. J. Kuck, "ILLIAC IV software and application programming," *IEEE Trans. Comput.*, vol. C-17, pp. 758-770, 1968.
- [9] D. H. Lawrie, "Access and alignment of data in an array processor," *IEEE Trans. Comput.*, vol. C-24, pp. 1145-1155, 1975.
- [10] Z. A. Melzák, "On the problem of Steiner," *Canad. Math. Bull.*, vol. 4, pp. 143-148, 1961.
- [11] G. Pólya, "Über die 'doppelt-periodischen Lösungen' des n -Damenproblems," in *Mathematische Unterhaltungen und Spiele*, W. Ahrens Ed. Leipzig: Teubner 1918, pp. 364-374.
- [12] H. D. Shapiro, "Theoretical limitations on the efficient use of parallel memories," *IEEE Trans. Comput.*, vol. C-27, pp. 421-428, 1978.
- [13] —, "Generalized latin squares on the torus," *Discr. Math.*, vol. 24, pp. 63-77, 1978.
- [14] J. Tappe, J. van Leeuwen, and H. A. G. Wijshoff, "Parallel memories, periodic skewing schemes, and the theory of finite abelian groups," to appear in *IEEE Trans. Comput.*
- [15] K. Vik, "Bedømmelse av feilen på forsøksfelter med og uten malestokk," *Meldinger fra Norges Landbrukshøgskole* 4, pp. 129-181, 1924.
- [16] H. A. G. Wijshoff and J. van Leeuwen, "Periodic storage schemes with a minimum number of memory banks," Dep. Comput. Sci., Univ. Utrecht, Utrecht, The Netherlands, Tech. Rep. RUU-CS-83-4, 1983.
- [17] —, "The structure of periodic storage schemes for parallel memories," *IEEE Trans. Comput.*, vol. C-34, pp. 501-505, 1985.

On the Time-Bandwidth Proof in VLSI Complexity

YASER S. ABU-MOSTAFA

Abstract—A subtle fallacy in the original proof [1] that the computation time T is lowerbounded by a factor inversely proportional to the minimum bisection width of a VLSI chip is pointed out. A corrected version of the proof using the idea of conditionally self-delimiting messages is given.

Index Terms—Bisected graph, computation time, information theory, lower bounds, self-delimiting, VLSI complexity.

I. INTRODUCTION

The lower bound on AT^2 where A is the area of a VLSI chip and T is either the average or worst case computation time on the chip, depends on the fact that $T \geq H/\omega$ where H is an information-theoretic constant that depends only on the function being computed and ω is the bandwidth (minimum bisection width/unit time) of the particular communication graph (chip model) in question [1]. The

Manuscript received July 18, 1985; revised October 15, 1985. This work was supported by the Program in Advanced Technologies (Aerojet, GM, GTE, TRW).

The author is with the Departments of Electrical Engineering and Computer Science, California Institute of Technology, Pasadena, CA 91125.

IEEE Log Number 8611452.

simple and plausible proof of the relation $\omega T \geq H$ was based on Shannon's Theorem 9 of [2] which states that it is not possible to transmit information over a channel of capacity C bits/unit time at an average rate of more than C bits/unit time. There is a problem in applying this theorem to the communication of *variable-length* messages between the two halves of a bisected communication graph, because the definition of *channel capacity* (as used in Shannon's proof) is tricky when the ensemble of messages over the channel has variable lengths. The dilemma is demonstrated in the following proposition. The definitions and basic properties of the entropy $H(x)$, the conditional entropy $H(x|y)$, and the average mutual information $I(x; y)$ can be found in [3, ch. 2]. We use the liberal notation which identifies a random variable with its ensemble.

Proposition: Let T be a nonnegative integer-valued random variable with mean value \bar{T} . Let $\mathbf{m} = m_1 \cdots m_T$ where each m_i is an independent (from T and the rest of the m_i 's) random variable assuming 2^ω equiprobable values. Then $H(\mathbf{m}) = \omega\bar{T} + H(T)$ bits.

Proof: $H(\mathbf{m}) = H(\mathbf{m}, T) - H(T|\mathbf{m}) = H(\mathbf{m}, T) - 0$ (since T is determined by \mathbf{m}) $= H(\mathbf{m}, T) = H(T) + H(\mathbf{m}|T) = H(T) - \sum_{m,T} P(\mathbf{m}, T) \log P(\mathbf{m}|T)$ (by definition, with $\log = \log_2$) $= H(T) - \sum_{m,T} P(\mathbf{m}, T) \log 2^{-\omega T}$ (equiprobability and independence) $= H(T) + \omega\bar{T}$. \square

When \mathbf{m} is the message being transmitted between the two halves of a bisected communication graph, $H(\mathbf{m})$ is the information carried by the message over the channel. Therefore, if $H(T)$ is positive (variable-length messages), *the total information of the message will exceed $\omega\bar{T}$ by sneaking in more information in the length of the message.* In practice, this extra information is canceled out by the requirement that the messages be self-delimiting (thus reducing $H(\mathbf{m}|T)$). An argument that incorporates self-delimiting is needed in the case of a bisected communication graph in order to justify the $\omega\bar{T}$ bound on the information flow between the two halves of the graph. The argument given in the next section does that using the idea of *conditionally* self-delimiting messages. The same remark applies to the ωT_{worst} bound in the worst case analysis. The total number of different messages can in principle be $2^0 + 2^\omega + 2^{2\omega} + \cdots + 2^{\omega T_{\text{worst}}}$ which exceeds $2^{\omega T_{\text{worst}}}$. A similar argument incorporating self-delimiting can be made in this case too.

II. THE NEW PROOF

We show that if ω bits/unit time is the bandwidth of a bisection $R - S$ of a communication graph that splits the input vector (uniformly distributed random variable) x into x_R and x_S and the output vector (dependent random variable) y into y_R and y_S , then the average computation time \bar{T} on this graph is governed by $\omega\bar{T} \geq H(y_R|x_R)$. The reader is referred to [1, Sec. 3.3] for the original argument.

Fig. 1 shows the information-theoretic model. Side R has full knowledge of x_R , but no initial knowledge of x_S , and wants to compute y_R which depends on both x_R and x_S . Through the message vector from side S to side R , namely $\mathbf{m} = m_1 \cdots m_T$ where $T = T(x)$ is a random variable, side R learns enough information about x_S to compute y_R . Since there are only ω wires between S and R each carrying at most one bit per unit time (in either direction), each m_i can assume at most 2^ω values. The mechanism of the computation requires that at time $t = T(x)$, side R is ready with the value of y_R , and knows that the computation is over. This means that, conditioned on the knowledge of x_R , \mathbf{m} is self-delimiting. We now give the formal argument.

Proof $\omega\bar{T} \geq H(y_R|x_R)$: Let $N = T_{\text{worst}}$ (maximum of T over all x). Define $\hat{\mathbf{m}} = \hat{\mathbf{m}}(x) = \hat{m}_1 \cdots \hat{m}_N$ as follows: $\hat{m}_i = m_i$ for $i \leq T$ and $\hat{m}_i = \text{constant}$ for $T < i \leq N$ (In other words, $\hat{\mathbf{m}}$ is a fixed-length version of the message \mathbf{m} , with no *explicit* information about the computation time). By the computation mechanism, y_R is uniquely determined from x_R and $\hat{\mathbf{m}}$, i.e.,

$$H(y_R|x_R, \hat{\mathbf{m}}) = 0.$$

Expanding the LHS term, we get $H(y_R|x_R) - I(y_R; \hat{\mathbf{m}}|x_R) = 0$. Since $I(y_R; \hat{\mathbf{m}}|x_R) \leq H(\hat{\mathbf{m}}|x_R)$ (see [3]), we get $H(y_R|x_R) \leq H(\hat{\mathbf{m}}|x_R)$ and it suffices to show that $H(\hat{\mathbf{m}}|x_R) \leq \omega\bar{T}$. We expand

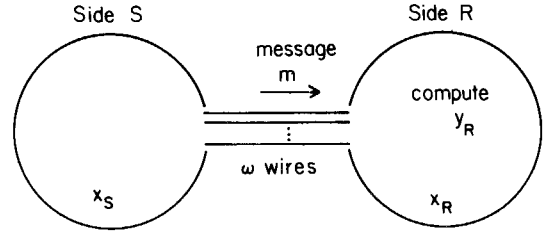


Fig. 1. Information-theoretic model of a bisected communication graph.

$H(\hat{\mathbf{m}}|x_R)$ as follows:

$$\begin{aligned} H(\hat{\mathbf{m}}|x_R) &= H(\hat{m}_1|x_R) \\ &+ H(\hat{m}_2|\hat{m}_1, x_R) \\ &+ H(\hat{m}_3|\hat{m}_1, \hat{m}_2, x_R) \\ &\vdots \\ &+ H(\hat{m}_N|\hat{m}_1, \dots, \hat{m}_{N-1}, x_R). \end{aligned}$$

Denoting by \hat{m}_t the first t symbols $\hat{m}_1 \cdots \hat{m}_t$ of $\hat{\mathbf{m}}$, this can be rewritten as

$$H(\hat{\mathbf{m}}|x_R) = \sum_{t=1}^N H(\hat{m}_t|\hat{m}_{t-1}, x_R).$$

To estimate the general term $H(\hat{m}_t|\hat{m}_{t-1}, x_R)$, we expand it as

$$\begin{aligned} & - \sum_{\hat{m}_{t-1}|x_R} \left(P(\hat{m}_{t-1}, x_R) \right. \\ & \left. \times \sum_{\hat{m}_t} P(\hat{m}_t|\hat{m}_{t-1}, x_R) \log P(\hat{m}_t|\hat{m}_{t-1}, x_R) \right). \end{aligned}$$

For every $t = 1, \dots, N$, whether or not the computation time T is less than t can be determined by \hat{m}_{t-1}, x_R . If $T < t$, the inner summation is zero because \hat{m}_t will be a constant. Otherwise, the inner summation is bounded by ω since \hat{m}_t assumes at most 2^ω values. Therefore, $H(\hat{m}_t|\hat{m}_{t-1}, x_R) \leq \Pr(T \geq t)\omega$, and

$$\begin{aligned} H(\hat{\mathbf{m}}|x_R) &\leq \sum_{t=1}^N \Pr(T \geq t)\omega \\ &= \omega \times \sum_{t=1}^N \sum_{i=t}^N \Pr(T=i). \end{aligned}$$

Each term of the form $\Pr(T = j)$ appears j times in the double summation, hence we rewrite

$$H(\hat{\mathbf{m}}|x_R) \leq \omega \times \sum_{j=1}^N j \times \Pr(T=j)$$

which reduces to $\omega\bar{T}$, and the proof is complete. \square

Two final remarks are in order. First, the message need not be *absolutely* self-delimiting, i.e., without the knowledge of x_R one message can conceivably be a prefix of another. Secondly, requiring that at time T the chip knows that the computation is over cannot be replaced by the weaker condition that the output y remains the same from time T on. The latter condition leaves us unsure about when to collect the output, possibly until time T_{worst} .

REFERENCES

- [1] C. D. Thompson, "A complexity theory for VLSI," Ph.D. dissertation, Carnegie-Mellon University, Pittsburgh, PA, 1980.
- [2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 1948.
- [3] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [4] J. D. Ullman, *Computational Aspects of VLSI*. Rockville, MD: Computer Science, 1984.