Lincoln Laboratory, MIT
Lexington, Massachusetts
November 13, 1959

Harlan E. Anderson
EJCC Publication Committee
Digital Equipment Corporation
Maynard, Massachusetts

Dear Mr. Anderson,

I am submitting herewith four copies of a paper entitled "Applications of Boolean Matrices to the Analysis of Flow Diagrams" for presentation before the Eastern Joint Computer Conference, December 1-3.

Sincerely yours,

Reese T. Prosser

Reese T. Prosser

File Copy

# Applications of Boolean Matrices to the Analysis of Flow Diagrams

by

Reese T. Prosser

## ABSTRACT

An analysis of the structure, or connectivity, of flow diagrams associated with computing machine programs is made by means of Boolean matrices. A Boolean matrix is a matrix whose entries are all either 0 or 1. With each program flow diagram is associated a pair of Boolean matrices. The first of these, called the connectivity matrix, contains the topological structure of the diagram, and the second, called the precedence matrix, contains the input-output structure of the diagram. Elementary computations on these matrices are shown to yield detailed information concerning the internal consistency of the program. Possible applications to automatic debugging procedures are suggested.

# Applications of Boolean Matrices to the Analysis of Flow Diagrams

Reese T. Prosser

## I. Introduction

Any serious attempt at automatic programming of large-scale digital computing machines must provide for some sort of analysis of program structure. Questions concerning order of operations, location and disposition of transfers, identification of subroutines, internal consistency, redundancy and equivalence, all involve a knowledge of the structure of the program under study, and must be handled effectively by any automatic programming system.

The structure of a program is usually determined by detailed specifications describing the program, and may usually be given a convenient geomatric representation b means of flow diagrams. Ordinarily, neither of these forms is immediately adaptable for handling by machine, and for this purpose another representation of the same information must be found. Such a representation should certainly have these properties:

1) It should be easy to construct and reproduce.
2) It should be adaptable to handling by machine.
3) It should contain all of the information provided by the topology of the flow diagram.

## II. The Connectivity Matrix

A representation which has all these properties may be given by means of Boolean matrices. By a Boolean matrix we mean a matrix whose entries consist entirely of 0's and 1's. The representation is constructed as follows: Suppose we are given the structure of a program, say in the form of a flow diagram consisting of boxes, representing program operations, connected by directed line segments, representing the program flow. We are interested only in the structure, or connectivity, of this diagram, and not in the properties of the individual boxes. We make no restrictions at all on the connectivity, and in particular, branches and loops of all kinds are admissible. We begin by numbering the boxes of the diagram, say from 1 to n, in any convenient manner whatever. For later convenience we adjoin to the diagram a box numbered 0 as the initial, or input, position and a box numbered n+1 as the final, or output, position of the diagram. We next construct an $(n+2) \times (n+2)$ Boolean matrix, $A = (a_{ij})$, called the connectivity matrix associated with the diagram by stipulating that $a_{ij} = 1$ if the diagram contains a directed line segment leading directly from box i to box j, and $a_{ij} = 0$ otherwise. Thus $a_{ij} = 1$ if box i may be followed immediately by box j in the program, and 0 otherwise.

It is evident that this matrix is easy to construct and easy to handle. It is determined uniquely by the diagram, up to a permutation of the entries due to a renumbering of the boxes, and in turn it determines the diagram, in the sense that the diagram may be completely reconstructed from the matrix. Thus it meets all of our requirements.

This idea is certainly not new. Boolean matrices have been used extensively to study the connectivity and orientation of graphs [7, 12], networks [4, 6], organization and group dynamics problems [8], and more generally, finite Markov processes [11]. Shannon [13] has pointed out that every flow diagram is essentially a finite Markov process, so that we have here a very special case of [11]. On the other hand it is worth emphasizing how well this idea adapts itself to program analysis. A similar attempt with a somewhat different viewpoint appears in [14].

## III. Analysis

Certain elementary computations on the connectivity matrix yield detailed information on the program flow. To show how this comes about, we define a one-row matrix

$$\underline{e}_i = (0, 0, \ldots, 1, \ldots, 0)$$

with 1 in the ith column and 0's elsewhere. Then from the definition of $\underline{A}$, we see that matrix product $\underline{e}_i\underline{A}$ is a one-row matrix which has 1 in the jth column if it is possible to proceed from box i to box j in one step, and 0 otherwise. By repeating this argument, we see that the product $\underline{e}_i\underline{A}^2 = (\underline{e}_i\underline{A})\,\underline{A}$ is a one-row matrix whose jth column is 1 (or more) if it is possible to proceed from box i to box j in exactly two steps, and 0 otherwise. A similar interpretation may evidently be given to higher powers of $\underline{A}$.

Now $\underline{A}^2$ need not be a Boolean matrix. But it is clear that for our purpose we lose nothing if we replace all non-zero entries in $\underline{A}^2$ with 1's.

This amounts to multiplying $\underline{A}$ by $\underline{A}$ according to the following rule: The Boolean product $\underline{A} \wedge \underline{B}$ of the Boolean matrices $\underline{A}$ and $\underline{B}$ is that Boolean matrix whose i-j entry is

$$\bigvee_k (a_{ik} \wedge b_{kj})$$

Here $\vee$ and $\wedge$ denote the Boolean operations of max and min, respectively. In the same spirit we define: The Boolean sum $\underline{A} \vee \underline{B}$ of the Boolean matrices $\underline{A}$ and $\underline{B}$ is that matrix whose i-j entry is $a_{ij} \vee b_{ij}$. Thus Boolean sums and products of Boolean matrices are formed in the same way as ordinary matrix sums and products, except that + is replaced by $\vee$ and x by $\wedge$.

Now the way is clear for induction. Let $\underline{A}$ be the connectivity matrix of a flow diagram, and define

$$\underline{A}_m = \underline{A}_{m-1} \wedge \underline{A} = \underline{A} \wedge \underline{A} \wedge \cdots\cdots \wedge \underline{A} \text{ m times}$$

$$\underline{B}_m = \underline{B}_{m-1} \vee \underline{A}_m = \underline{A}_1 \vee \underline{A}_2 \vee \cdots\cdots \vee \underline{A}_m$$

Theorem 1. The i-j entry of $\underline{A}_m$ is 1 if it is possible to proceed from box i to box j in exactly m steps, and 0 otherwise. The i-j entry of $\underline{B}_m$ is 1 if it is possible to proceed from box i to box j in at most m steps, and 0 otherwise.

Proof: For m = 1, both statements reduce to definitions. Now suppose both statements hold for $m = m_0$; and consider the case $m = m_0 + 1$. The i-j entry of $\underline{A}_{m_0+1}$ is just $\bigvee_k a_{ik} \wedge c_{kj}$, where $c_{kj}$ denotes the k-j entry of $\underline{A}_{m_0}$. This is zero, unless for some k

We have a

we have $a_{ik} = c_{kj} = 1$. But this means that it is possible to proceed from box i to box k in exactly one step, and from box k to box j in exactly $m_0$ steps. Thus the i-j entry of $\underline{A}_{m_0+1}$ is 0 unless it is possible to proceed from box i to box j in exactly $m_0+1$ steps. The second statement follows immediately from the first.

Theorem 2. The limit $\lim_{m \to \infty} \underline{B}_m$ exists as a Boolean matrix, which we denote by $\underline{B}$. Moreover, we have $\underline{B} = \underline{B}_m$ for all $m \geq p$, where p is the length of the longest open path in the diagram.

Proof: Since the entries of $\underline{B}_m$ are monotone increasing with m, it is clear that $\lim_{m \to \infty} \underline{B}_m$ exists and forms a Boolean matrix. The second statement follows from the observation that if it is possible to proceed from box i to box j at all, it is possible to do so along an open path (i.e., one containing no loops), and hence in less than p+1 steps. Thus if the i-j entry of $\underline{B}_m$ is 1 for any m, it is 1 for m = p. This means that $\underline{B}_m = \underline{B}_p$ whenever $m \geq p$.

Theorem 3. The i-j entry of $\underline{B}$ is 1 if it is possible to proceed from box i to box j in any number of steps, and 0 otherwise.

Proof: This follows immediately from the proof of Theorem 2.

The matrix $\underline{B}$ is obviously computable by machine from the matrix $\underline{A}$, and since only Boolean operations are involved, the time required for this computation is not prohibitive even for fairly large n. On the other hand,

it follows from Theorem 3 that the matrix $\underline{B}$ contains detailed information about the consistency of the flow diagram. We cite some obvious examples:

1) It is possible to get from the input to box i only if $b_{0i}=1$. Thus if there are no spurious boxes, the top row of $\underline{B}$ must contain all 1's (except for $b_{00}$).

2) It is possible to get from box i to the output only if $b_{i(n+1)} = 1$. Thus if there are no boxes without exits, the last column of $\underline{B}$ must contain all 1's (except for $b_{(n+1)(n+1)}$).

3) It is possible to get from box i to box i only if $b_{ii} = 1$. Thus if there are no loops in the program, the main diagonal of $\underline{B}$ must contain all 0's. Boxes involved in loops are represented by 1's on this diagonal.

4) After leaving box i, it is possible to go through box j only if $b_{ij} = 1$. Now if we alter box i then only those boxes following box i in the program will be affected. These boxes are represented by 1's in the ith row of $\underline{B}$.

5) If the matrix decomposes into relatively independent submatrices, then the program decomposes into relatively independent subprograms. Thus it may be possible to identify natural subprograms directly from the form of the matrix $\underline{B}$.

## IV. Examples

The foregoing theory will be further illuminated by application to

concrete problems. As a first example we choose a flow diagram containing an obvious inconsistency, and how how this inconsistency is reflected in the matrix $\underline{B}$. The diagram is shown in figure 1. Here the boxes are already numbered, including the input and output boxes. The connectivity matrix for this diagram is a 7 x 7 matrix, whose entries are

$$\underline{A} = \begin{pmatrix} 010\ 100\ 0 \\ 001\ 000\ 0 \\ 010\ 000\ 0 \\ 000\ 011\ 0 \\ 000\ 000\ 1 \\ 000\ 001\ 1 \\ 000\ 000\ 0 \end{pmatrix}$$

Now $\underline{A}_1 = \underline{B}_1 = \underline{A}$. Straightforward computation gives

$$\underline{A}_2 = \underline{A} \wedge \underline{A} = \begin{pmatrix} 001\ 011\ 0 \\ 010\ 000\ 0 \\ 001\ 000\ 0 \\ 000\ 001\ 1 \\ 000\ 000\ 0 \\ 000\ 001\ 0 \\ 000\ 000\ 0 \end{pmatrix}$$

$$\underline{B}_2 = \underline{B}_1 \vee \underline{A}_2 = \begin{pmatrix} 011\ 111\ 0 \\ 011\ 000\ 0 \\ 011\ 000\ 0 \\ 000\ 011\ 1 \\ 000\ 000\ 1 \\ 000\ 001\ 1 \\ 000\ 000\ 0 \end{pmatrix}$$

$$\underline{A}_3 = \underline{A}_2 \wedge \underline{A} = \begin{pmatrix} 010\ 001\ 1 \\ 001\ 000\ 0 \\ 010\ 000\ 0 \\ 000\ 001\ 0 \\ 000\ 000\ 0 \\ 000\ 001\ 0 \\ 000\ 000\ 0 \end{pmatrix}$$

$$\underline{B}_3 = \underline{B}_2 \vee \underline{A}_3 = \begin{pmatrix} 011 & 111 & 1 \\ 011 & 000 & 0 \\ 011 & 000 & 0 \\ 000 & 011 & 1 \\ 000 & 000 & 1 \\ 000 & 001 & 1 \\ 000 & 000 & 0 \end{pmatrix}$$

A glance at the diagram shows that all possible paths (without repetition) can be traversed in at most three steps, so that by Theorem 2, $\underline{B} = \underline{B}_3$. This can be checked by computing $\underline{B}_4$, which is equal to $\underline{B}_3$. From this matrix we verify immediately that all boxes are connected to the input (first row), but boxes 1 and 2 are not connected to the output (last column). Boxes 1, 2 and 5 are involved in loops (main diagonal). Moreover, if we delete the first row and last column of $\underline{B}$, then the remainder can be decomposed into submatrices:

$$\begin{pmatrix} 11 & 000 \\ 11 & 000 \\ 00 & 011 \\ 00 & 000 \\ 00 & 001 \end{pmatrix} = \begin{pmatrix} \underline{M} & 0 \\ 0 & \underline{N} \end{pmatrix}$$

where $\underline{M} = \begin{pmatrix} 11 \\ 11 \end{pmatrix}$ and $\underline{N} = \begin{pmatrix} 011 \\ 000 \\ 001 \end{pmatrix}$. This implies that boxes 1 and 2 and boxes 3, 4 and 5 form two independent subprograms whose associated matrices are just $\underline{M}$ and $\underline{N}$. (Of course, the simplicity of this decomposition is due to the particular scheme adopted for numbering the boxes.) This simple example serves to illustrate the scope of the method.

This same method has an obvious application to the problem of debugging programs already compiled. In this case the boxes are already

numbered by the sequential dexcription of the program. Moreover, it is not necessary to draw the corresponding flow diagram, since, except for transfers, each operation is followed by the next in sequence. As a second example we take a typical SAP writeup of an IBM 704 program, with no inconsistencies. (This program computes an array of 100 quantities $c_{ij}$ according to the formula

$$c_{ij} = \begin{cases} A_i - B_j & \text{if } i > j \\ A_i + B_j & \text{if } i \leq j \end{cases}$$

SAP Program

```
1.      LXD   8
2.      SXD   4
3.      CLA   B1
4.      TXL   6
5.      CHS
6.      ADD   A1
7.      STO   C1
8.      TXI   9
9.      TXI   10
10.     TMX   2
11.     TXI   12
12.     TMX   2
13.     END
```

The associated connectivity matrix can be written down directly, and is simply

$$\underline{A} = \begin{pmatrix} 010 & 000 & 000 & 000 & 0 \\ 001 & 000 & 000 & 000 & 0 \\ 000 & 100 & 000 & 000 & 0 \\ 000 & 011 & 000 & 000 & 0 \\ 000 & 001 & 000 & 000 & 0 \\ 000 & 000 & 100 & 000 & 0 \\ 000 & 000 & 010 & 000 & 0 \\ 000 & 000 & 001 & 000 & 0 \\ 000 & 000 & 000 & 100 & 0 \\ 010 & 000 & 000 & 010 & 0 \\ 000 & 000 & 000 & 001 & 0 \\ 010 & 000 & 000 & 000 & 1 \\ 000 & 000 & 000 & 000 & 0 \end{pmatrix}$$

(Note that, except for transfer instructions, 1's appear only on the super diagonal.)

## V The Precedence Matrix

A further analysis of the structure of a program can be made if information concerning the precedence relations in the program is available. If we know, for example, that the output of box i is required for the input of box j, then we know that the operation represented by box i must precede that represented by box j in the program sequence. Clearly this places additional requirements on the internal connectivity of the program.

The precedence relations may be incorporated into our analysis through the introduction of a second Boolean matrix $\underline{C}$ associated with the program, which we call the precedence matrix. (cf. [1, 9]). It is constructed as follows. We number the boxes of the diagram as in Section II, and stipulate that the i-j entry $c_{ij}$ of C is to be 1 if the output of box i (or any part of it) is required for the input of box j, and 0 otherwise. Clearly this matrix contains the precedence relations in the same way that the matrix $\underline{A}$ contains the connectivity relations of the program, and will yield to a similar analysis. We observe here that the two matrices are closely related, though they need not be identical.

Proceeding as in Section II, we define

$$\underline{C}_m = \underline{C}_{m-1} \wedge \underline{C}$$

$$\underline{D}_m = \underline{D}_{m-1} \vee \underline{C}_m$$

$$\underline{D} = \lim_{m \to \infty} \underline{D}_m$$

and observe that the results of that section may be translated immediately into the present situation. In particular, the i-j entry of the matrix $\underline{D}$ is 1 if and only if there is a chain of boxes in the diagram beginning with box i and ending with box j such that each box in the chain must precede the next. Obvious applications include the following:

1) The precedence requirements are internally consistent only if the diagram contains no closed chain of boxes each of which must precede the next. This is the case only if no diagonal entry of $\underline{D}$ is 1. Thus we require that trace $\underline{D} = 0$ for this consistency (cf. [1]).

2) In general, box j depends on box i only if $d_{ij} = 1$. Thus if box i is altered, this will affect only those boxes whose entries in the ith row of $\underline{D}$ are 1.

3) Occasionally it is desirable to reorder the sequence of operations in some part of the program. This is possible only if the precedence requirements are not violated by the reordering. Thus box i may be interchanged with box j in a chain of operations only if

$d_{ij} = d_{ji} = 0$. Information of this kind is evidently useful in optimizing flow diagrams for time or storage requirements.

## VI. The Dominance Matrix.

In studying problems involving the reordering of operations in a program, it is often useful to introduce a notion of dominance in the flow diagram, defined as follows: We say box i dominates box j if every path (leading from input to output through the diagram) which passes through box j must also pass through box i. Thus box i dominates box j if box j is subordinate to box i in the program. It may happen that two boxes dominate each other (in which case we say they are equivalent), or that n neither dominates the other (in which case we say they are independent). The idea here, of course, is that in general reordering is possible only among boxes which are equivalent in this sense. Proceeding along these lines, we define a third Boolean matrix $\underline{E}$, called the dominance matrix, by stipulating that the i-j entry $e_{ij}$ of $\underline{E}$ is 1 if box i dominates box j, and 0 otherwise. It is clear that the dominance matrix is determined by the connectivity matrix, and can be produced from it by a suitable scanning procedure. Applications include:

1) Box i and box j may be interchanged, precedence requirements permitting, only if they are equivalent. This is the case only if we have $e_{ij} = e_{ji} = 1$.

2) In preparing a program for a machine which admits parallel operation, it is desirable to know which operations in the program may be performed simultaneously. Two operations may be

performed simultaneously without further investigation only if they are equivalent and subject to no precedence requirements, i.e., only if $d_{ij} = d_{ji} = 0$ and $e_{ij} = e_{ji} = 1$.

3) It is sometimes useful to know when two programs are equivalent in some sense. Any effective definition of equivalence requires a detailed knowledge of what happens at branch points in the program (i.e., the transfer conditions). An interesting analysis of this problem is summarized in [14], but does not seem readily adaptable to machine handling. By requiring a less effective definition of equivalence, we can give here an effective criterion for determining whether or not two programs are equivalent.

To be precise, let us agree that <u>two programs, containing the same operations subject to the same precedence requirements, are equivalent, if, for each path (leading from input to output) through the first, there is a corresponding path through the second passing through the same operations</u>. We do not require that the operations appear in the same sequence, or even that they appear the same number of times, in both paths. This definition, however, is sufficient for most purposes, at least for programs containing no loops; loops cannot be incorporated under so simple a scheme, and require special consideration.

In terms of flow diagrams, the equivalence criterion may be stated as follows. To avoid inessential complications, we assume that no independant boxes (in the sense of dominance) are directly connected. This can always be achieved by adding suitable "empty" boxes to the diagram. Then it is true that <u>two diagrams,</u>

made up of the same boxes subject to the same precedence requirements, are equivalent if and only if their dominance matrices are identical.

## VII Remarks.

The essential point of our discussion is that the entire analysis given here can be readily performed on any (large-scale) digital computer. The feasability of computing the derived matrices B, D, and F by machine is assured for programs which are not too large. A very crude estimate indicates that the time required to compute B from A on the IBM 704 is of the order of $10\ n^3$ cycles, where n is the number of boxes in the diagram. In practice, this time may be reduced considerably by combining into one box any subroutine whose behavior is known. Thus for example it is advantageous to replace any chain of boxes by a single box. Similarly, in analyzing program writeups it is sufficient to consider only transfer operations. For instance, a reduced form of the matrix A of our second example in section IV is:

$$A' = \begin{pmatrix} 010\ 000\ 0 \\ 001\ 000\ 0 \\ 010\ 100\ 0 \\ 000\ 010\ 0 \\ 010\ 001\ 0 \\ 000\ 000\ 0 \end{pmatrix}$$

where boxes 1 through 9 have been combined in a single box.

Finally we remark that it is a straightforward problem to construct a debugging routine which could be used to analyze any program writeup

whose transfer instructions have constant addresses. Such a routine would scan the writeup, enumerate the transfer instructions, construct the connectivity and dominance matrices from them, compute the derived matrices and point out any errors detectable by these methods. Thus the whole analysis becomes completely automatic.

Various other applications of this analysis are suggested by the results. By utilizing the evident adaptability of these matrices to computer handling, it is possible to construct automatic program analysis schemes which would detect in proposed programs a large class of common errors, isolate and identify key subroutines and reorganize them in optimal equivalent programs. Such a scheme is currently under investigation here at Lincoln Laboratory, MIT.

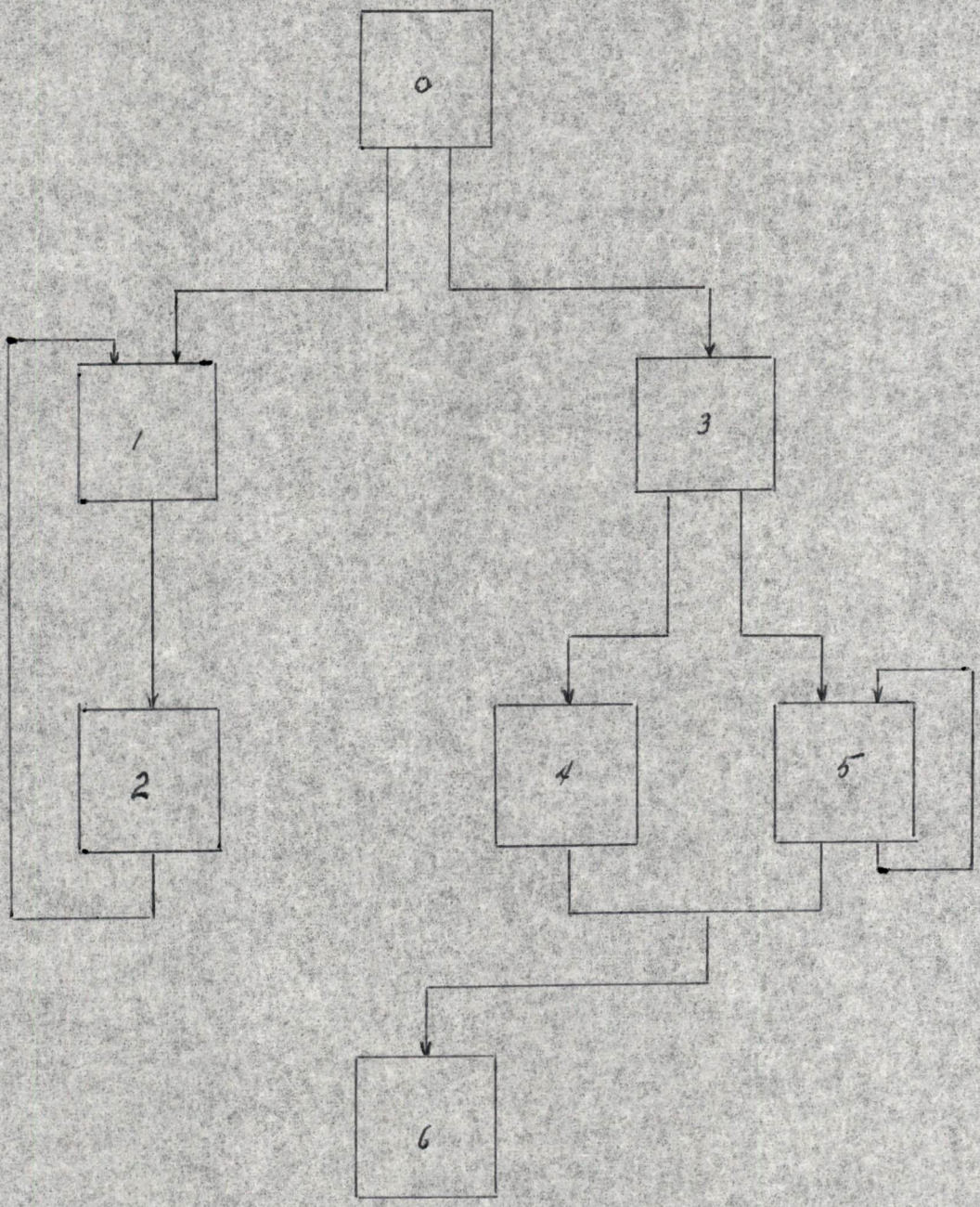



FIGURE 1

## Bibliography


1. N.W. Barankin, "Precedence Matrices", Univ. of Chicago Management Sciences Research Project, Research Report no.26; December, 1953.

2. I.M. Copi, "Matrix development of the calculus of relations", Jour. Symbolic Logic, vol. 13, pp.193-203; 1958.

3. W. Feller, "An Introduction to Probability Theory and its Applications", John Wiley and Sons, New York, N.Y., p.350; 1957.

4. F. Hohn and L. Schissler, "Boolean matrices and the design of combinational relay switching circuits", Bell System Tech. Jour., vol.34, pp.177-202; 1955.

5. M. Kac and J.C. Ward, "A combinatorial solution of the 2-dimensional Ising model", Phys. Rev., vol.88, pp.1332-1337; 1952.

6. G. Kron, "Tensor Analysis of Networks", John Wiley and Sons, Inc., New York, N.Y.; 1939.

7. S. Lefschetz, "Topology", Colloq. Publications Amer. Math. Society, New York, N.Y.; 1930.

8. R.D. Luce and A.D. Perry, "A method of matrix analysis of group structures", Psychometrika, vol.14, pp.95-116; 169-190; 1949.

9. R.B. Marimont, "A new method of checking the consistency of precedence matrices", Jour. Assoc. Comp. Mach., vol.6, pp.164-171; April, 1959.

10. J. Riordan, "An Introduction to Combinatorial Analysis", John Wiley and Sons, Inc., New York, N.Y.; 1958.

11. D. Rosenblatt, "On the graph and asymptotic forms of finite Boolean relation matrices and stochastic matrices", Naval Res. Logist. Quart., vol. 4, pp.151-167; 1957.

12. H. Seifert and W. Threlfall, "Lehrbuch der Topologie", Chelsea, New York, N.Y.; 1947.

13. C. Shannon and W. Weaver, "The Mathematical Theory of Communication", Univ. of Illinois, Urbana, Ill.; 1949.

14. Y.I. Yanov, "On matrix schemes", Dokl. Akad. Nauk. USSR, vol.113, pp.39-42; 1957.

File Copy

September 28, 1959

ABSTRACT


COMPUTERS OF THE FUTURE

This discussion is concerned with a radical change in the technology utilized to manufacture digital data processing systems. A picture of the effect of this change on our way of specifying and designing systems is presented.

Present methods of circuit-system standardization are contrasted with anticipated future methods. An illustrative example of "system function" design and "system tailored" circuits and devices is given. A summary is made of the more important work required in order to progress from present to desired future systems. This involves integrated Research and Development on programming languages, system logic, packaging, devices, materials, service techniques, and manufacturing methods.


RR/b

# BIOGRAPHICAL SKETCH

## Mr. Rex Rice

Mr. Rice received his B. S. in Mechanical Engineering at Stanford University in 1940. He worked in the aircraft industry as a tooling engineer and later as an aircraft structures engineer.

He became interested in computing while doing research on structural analysis methods and subsequently became Assistant Chief of Computing Services at Northrop Aircraft.

In 1955 he joined IBM as a Senior Engineer in charge of a machine development program. He joined IBM Research in 1958 and is presently manager of a machine organization theory group.

His background and interests relative to computing include system logical design and range from problem definition to manufacturing techniques.

# COMPUTERS OF THE FUTURE

R. Rice

## INTRODUCTION

This paper considers the advances required in many related technologies to revolutionize the construction and use of digital data processing systems. Webster gives as one definition of a revolution: "A total or radical change." In the following discussion we are particularly concerned with the radical change in fabrication technology and wish to analyze the effect that this change will have on our methods of computer design and specification.

## PRESENT METHODS

The manufacturing techniques used in the electronic portion of today's digital data processing systems are illustrated in Figure 1. The active devices are standardized in these systems. Circuit standardization is established at what may be defined as the Boolean function level. Circuits for AND, OR, Invert, Latch, Trigger, etc., are standardized individually. The pluggable packaging usually combines several circuits, either of the same type or in selected groups. A major system function such as a complete working storage register and all its controls, an arithmetic processing unit and its controls, etc., is obtained by assembling a group of circuit packages on a panel and interconnecting the circuit packages with individual wires. At the time the individual circuits and packages are designed and optimized, very little information is

available regarding their specific employment in systems functions.

A digital "system function" may be defined as a combination of logical elements interconnected and timed to perform major operational sequences in a data processor. One of our future objectives is to create major digital system functions in one continuous, automated manufacturing sequence.

FUTURE METHODS

A possible future method for producing major "system functions" such as complete working storage registers, process units, memory arrays, etc., is illustrated in Figure 2. We envision this manufacturing line as a set of printing presses through which a conveyor system passes. Substrate material is placed on the conveyor and proceeds through the line. At each stage one pattern of interconnections, insulation, or active material is printed on the substrate. As required, bake ovens, etc., may be strategically placed. Here, devices are standard by virtue of the materials used. These materials are applied by a standardized method to produce active elements, interconnections, insulation, etc., in batches. The plates, inserted in each press, are made in an automatic machine which develops the appropriate layout under equation control for major system segments.

The figure illustrating future methods is only diagramatic. The manufacturing method chosen will probably depend on the basic component technology and may be different for each type of component. Before complete automation is realized it will be necessary to separately manufacture

active elements and rely on automatic testing and insertion. The field will be dynamic and the illustration indicates a trend, not a specific technique.

## ILLUSTRATIVE EXAMPLE OF A SYSTEM FUNCTION

A serial-by-digit, decimal adder is used to illustrate a system function as shown in Figure 3. This represents a portion of an arithmetic processing unit. The digital code assumed is a decimal "one out of ten" representation, chosen because decimal matrix addition is well understood. Other examples or codes would have served equally well.

In this function a pair of decimal digits enters a process unit at A and B and the added result is obtained at the output. A matrix, to be described in detail, performs the first half addition. Other elements provide input drive, output carry detection, recombination, and the second half addition. It is also necessary to store the presence or absence of a carry so that as succeeding pairs of digits are processed the second half addition circuit may be activated. Let it be assumed by way of example that A equals 5 and B equals 6, as emphasized with heavy marked lines. In the matrix the 5 on the vertical axis together with a 6 on the horizontal axis activates an AND circuit which places an output on the eleventh diagonal. After passing through the carry detection element, the eleventh diagonal is recombined with the output line 1. The carry condition is remembered for later use. Let us now consider circuits for the matrix in more detail.

## MATRIX UTILIZING INDIVIDUAL, STANDARDIZED BOOLEAN CIRCUITS

The circuit in Figure 4 is a Boolean standardized two-way AND

circuit with one transistor, four resistors, and various internal interconnections. Several outputs may be wired together to form an appropriate OR circuit. A two-way circuit is chosen since for our purposes in the addition matrix a three- or four-way AND circuit has no advantage.

A ten by ten matrix of these AND circuits is illustrated in Figure 5. For clarity, the internal circuit connections and devices have been omitted. In the matrix addition is accomplished by the coincidence of current on any pair of lines such as A = 5 and B = 6. When the AND circuit at this intersection is active, its output is placed on the eleventh diagonal. For packaging purposes the designer has the choice of packaging several AND circuits on a single pluggable unit. When the circuits were optimized, only the two-way AND logic together with the output loading conditions were known.

Let us now reexamine this same matrix from a "system" rather than a circuit viewpoint (Figure 6). In this specific matrix element only one AND circuit in the A = 5 column and the B = 6 row is "on." This is a system consideration and was not known at the time the Boolean AND circuit was optimized. The vertical column A = 5 will now be considered as a single element.

## SYSTEM TAILORED CIRCUITS

A circuit which is tailored to this "system function" is illustrated in Figure 7. For convenience, transistors have been shown, although other devices such as relays, tubes, cryogenic devices, etc., could have

been used. The input A supplies current to a common control which goes to all the bases of the ten transistors. Since only one line on the B input to the emitters is active at any instant, only one transistor will be conducting. Let us now examine the addition matrix utilizing this "system tailored" circuit.

MATRIX UTILIZING SYSTEM FUNCTION CIRCUITS

The complete matrix is again shown in Figure 8, this time utilizing ten of the "system function" circuits. The "A" entries on the vertical axis go directly to the common control connections of the ten AND circuits. The "B" entries are connected to the emitters of the ten transistors in each of the ten circuits. The collectors are connected to the output lines which are functionally equivalent to diagonals in the previous matrix. Note the identical configuration of the wiring to the inputs of all ten matrix columns. The outputs of each "system AND" circuit are connected in a pattern which drops down to the next output line for each successive group. Thus, to add 5 to the number entering B the sixth AND circuit is activated. The number 6 on the B entry is moved down five units on the output, giving a sum of 11. Although the number of transistors required in both matrix examples remains the same, the passive elements are eliminated and the packaging pattern for both interconnections and devices is drastically improved.

In the illustration the solid lines represent a layer of interconnections on the front of a printed substrate and the dotted lines, a second layer on the rear. Connections through the substrate are indicated by dots. Inasmuch

as ten system function circuits are used, ten component packages consisting of active elements only may be mounted on a single substrate that contains the complete interconnection wiring.

A computer may be described as "a bunch of wires connected by active elements." This second method of matrix design underscores that definition. Three important features become apparent in this example. First, careful attention to system function circuits will lead to logical layouts that are much easier to express algebraically for equation-controlled manufacturing. Second, the amount of packaging and interconnections, and the number of elements involved can be reduced over present methods. Third, new "system function" device specifications will emerge.

## SYSTEM TAILORED DEVICES

The previous discussion presented an example in which circuits and system function logic were combined using standard transistors. Present active devices are individual elements packaged separately, as shown in Figure 9. The connections between the active and passive elements are generally made by individual wires, although more recent systems use printed wiring for curcuit packages.

In an early generation, multi-element "system tailored" devices will be available. In addition, a much greater proportion of the interconnections will be etched and printed. Multi-element miniaturized components have been made available in small quantities by American Bosch Arma, the

Diamond Ordnance Fuze Laboratory Hughes Aircraft, RCA, Texas Instruments, and others. Programs in molecular electronics to permit the use of plating and vacuum-deposition processes are also receiving attention. Much of this work is for military applications but will probably be available for commercial use in the near future.

The production of interconnections and active elements in one continuous manufacturing process will occur with the introduction of films, either thick or thin, into systems. At this time, semiautomatic methods of manufacture will be mandatory. Here it is obvious that separate considerations of system functions, circuits, and devices may no longer exist. The device illustrated contains multiple active elements controlled by a single line. Magnetic coupling is used to accomplish switching in thin film cryogenic systems and speeds are very high. One suspects that nature also provides a medium speed and cost arrangement if we are clever enough to detect it.

Further in the future we may anticipate true microminiaturized systems constructed from automatic, computer-controlled processes utilizing bulk materials. The late Professor Dudley Buck has defined a microminiature computer as: "A computer on a scale which could never be looked at in an optical microscope." In this technology, the cost of active elements will approximate the cost of interconnections. Logical designers may enjoy the luxury of utilizing thousands of active elements to perform logical functions of a complex nature.

One of our major objectives is to reach the future system illustrated here. Let us now consider some of the more important work to be done to make this possible.

## DIGITAL DATA PROCESSING APPROXIMATE RELATIVE COSTS

The bar graph (Figure 10) shows the approximate relative costs of processing data in presently available commercial general-purpose digital systems. Problem preparation and programming costs are generally accepted as being approximately one half of the total. The remaining costs may be divided into two major items: the electronic main frame costs and the electromechanical peripheral equipment costs. The percentages vary from system to system, but are essentially as follows: The cost of the main frame electronics varies between 15 and 25 percent of the total, and includes the main random access storage, the arithmetic and logic unit, and controls. In the main frame, the switching devices cost approximately one-third and the packaging (which includes circuit cards, panels, interconnections, frames, display, covers, etc.), approximately two thirds. The cost of the electromechanical portion of a system may vary between 25 and 35 percent of the total and may be divided into two parts. The first is bulk storage involving mechanical motion. This part includes tapes, discs, drums, etc., and their attendant electronic equipment. The second part is the input-output equipment, including communication devices.

## PRESENT GENERATION

General purpose systems predominate at the present time.

This is probably due to the relatively high cost of research and development coupled with long design and manufacturing lead times for initial production. Instructions usually include an operation, one or two addresses, and a few special control bits. The instruction code at the machine language level is relatively "micro" due to the general-purpose requirement and for other reasons not covered here.

System specification normally starts with a market analysis so that a potential product may be defined. Performance, storage volume, input-output equipment, etc., are established at this time. Available standard circuits and packages are considered during the specification of system logic. Outputs from the system design are block diagrams, or equations, or both. At this stage we do not know where each device or circuit will be placed, nor the length of interconnections.

In programming, present generation machines use autocoders to translate from problem language into machine language. The autocoders, in many instances, involve execution time and occupy storage space. This combination of autocoders and machine language is the result of the programmer's desire to have a different machine language than the one technology is able to economically provide.

Devices used in present systems, both active and passive, are individually manufactured by semiautomated methods. This allows individual testing, selection, and replacement in the event of malfunction.

The circuits are Boolean optimized and the minor packaging assemblages usually include several elementary functions. Recent trends as evidenced in machines like the Philco TRANSAC, are toward the inclusion of more Boolean type circuits on each pluggable element. Interconnections are a mixture of printed cards and hand inserted wires and cables.

The major mechanical design of a system starts when logical specification and Boolean standardized circuits are available. With this information, the active and inactive elements may be located and packaged. For the first time, lead lengths become accurately known. The output from mechanical design is generally a complete set of blueprints which go to the manufacturing engineering groups.

In the peripheral equipment area the bulk storage usually involves magnetics and includes much mechanical equipment. Access to data in this type of storage is either serial-by-bit or serial-by character. The input-output equipment is essentially mechanical, taking data from a keyboard to a buffer storage and, later, taking data from a buffer to a printer to produce hard copy.

Servicing is usually done by a combination of electrical tests and diagnostic programs. It involves locating the defective, active or passive elements and substituting new pluggable cards.

Summary

The specification and design of present systems is essentially a serial process in which most major elements are individually standardized and then assembled to make a system. The design feedback loops, while many, have rather high impedence.

NEXT GENERATION

The next generation, as illustrated by the bar in Figure 11, may be characterized mainly by "system oriented" design and manufacturing techniques. Commercial machines will probably remain general-purpose in nature.

The bars illustrating approximate relative cost on this and succeeding generations does not necessarily indicate that the cost of an equivalent advanced machine will be reduced. The length of the bars represents the relative proportionate cost for each of the major elements in a system for a particular generation. Past experience has shown that as more powerful techniques become available we solve larger problems; therefore, we have an option of obtaining more computing for our millions or reduced costs for the same amount of processing. This is obviously a designer's choice and will be adjusted to suit requirements as he specifies a particular system.

A major change will occur in the specification of systems. Logic and circuits will be merged to produce new system function circuits

utilizing standard devices. The physical location of components, the interconnection lengths and paths, and layout of the package will be specified as an integral part of logic. To attain these objectives a new "system function algebra" is necessary. This algebra, which will begin with the logical Boolean expressions, must be enriched to include the active and passive device characteristics, the physical location of all components, the interconnection paths and lengths, and timing.

Programming in this generation will be done with more powerful macro-type instructions. Machine language instructions will approximate the level typified by coding systems such as FORTRAN. Relatively speaking, more hardware will be in the instruction controls with the objective of making programming easy and fast.

Improved single function devices and some use of multifunction devices may be anticipated.

A major change in packaging as well as in logic-circuit specification will occur in this generation. Complete system functions will be packaged on one replaceable element. Interconnections will be etched, printed, evaporated, or batch produced by other automated techniques. Manufacturing equipment, methods, and mechanical design techniques must undergo the appropriate changes.

Service will be accomplished by locating and replacing malfunctioning major system functions. If the individual devices are expensive, they may be replaced at a testing and service center so that the system function may be returned to stock. If not, the whole unit may be discarded. Extensive built-in checking and automatic program diagnosis will be

included. The logic of the machine will require more redundancy for checking and diagnostic purposes.

Summary

This generation involves a major improvement in logical design and packaging. New devices or other research items are not necessarily required.

SECOND GENERATION (Figure 12)

Two major changes characterize the second generation systems. First, system-tailored multi-element devices will be used extensively. This will influence mechanical design, packaging, and manufacturing equipment. Secondly, special-purpose machine systems to solve classes of problems will be made on the same manufacturing line. The logical specification of these machines will be generated by computers utilizing system function algebra. Extensions of the algebra will control the manufacturing setup. This combination will drastically reduce design and production lead times and cost of the product.

The availability of special-purpose systems will ease programming difficulties through the use of application-tailored languages to solve related classes of problems.

System-function design techniques and devices will be applied to bulk storage. For input-output, electronics will replace mechanical equipment wherever possible.

No on-line service will be performed since the machine will be able to select alternate logical paths in the event of a malfunction. At

inspection periods, previously flagged defective system elements will be removed and replaced.

THIRD GENERATION (Figure 13)

The true revolution begins in the third generation. Here, device, package, and interconnections are inseparably merged. Major system functions will be produced from bulk materials in computer-controlled continuous manufacturing processes. Techniques such as vacuum deposition, electron beam writing, spraying, printing, etc., will be utilized, depending on device technology chosen relative to the speed and cost range desired. The use of three dimensional connections will alter packaging concepts. Miniaturization for complete systems may now be realized. This miniaturization will allow dramatic increases in the number of active elements available for both logic and storage.

The availability of vast amounts of homogeneous storage with internal logical capabilities will drastically alter programming methods. In particular, built-in symbolic addressing will eliminate the inefficient and tedious housekeeping associated with present-day machines. Coupled with special-purpose instruction sets, this will allow machine language to approximate problem language.

The input-output equipment will now be reduced to that which is used to communicate with humans or from machine to machine, since bulk storage is now merged with the main frame.

Service will be simple because automatic error detection and correction by the machine will allow continuous operation. Defective elements will be replaced at the next service period.

## FUTURE GENERATION

We may envision a few aspects of future generations now (Figure 14). True microminiaturization meeting Professor Buck's definition will be realized. Self-organizing systems will become possible due to microminiaturization and better understanding of the logic involved. The use of self-organizing systems to find optimum solutions to problems will allow us to synthesize more economical, special-purpose systems for on-line use.

For programming, we may anticipate that machine language will approximate or equal human language if we have progressed properly to this point and if we use self-organizing systems appropriately. A major change in input-output techniques is required. Voice and pattern recognition, and vastly improved display and printing systems are needed.

In this generation service will be accomplished by throwing the whole computer away.

In summary, to progress from the present day data processing capabilities to more desirable future systems, we require greatly increased logical capabilities, vast amounts of storage, improved input-output methods and more speed. All these elements tend to require microminiaturization, batch-bulk processing, automated logical synthesis, and equation-controlled

manufacturing. Consequently, both speed and system cost require and benefit from this revolution.

## CONCLUSION

Future computers (Figure 15) will be standardized as follows:

1. Interconnections and active devices will be made in a continuous process from bulk raw materials to finished product.
2. The device, circuit, and interconnection technology will merge.
3. System function algebra will be used to specify all aspects of design.
4. Completely automated, computer controlled manufacturing methods will be used.

From these techniques we will obtain efficient special-purpose digital data processing systems. They will be produced economically with short design and construction lead times through complete automation. This will result in more brain power being devoted to discovering and defining new problems, and in their cheap, efficient solution.

REFERENCES:

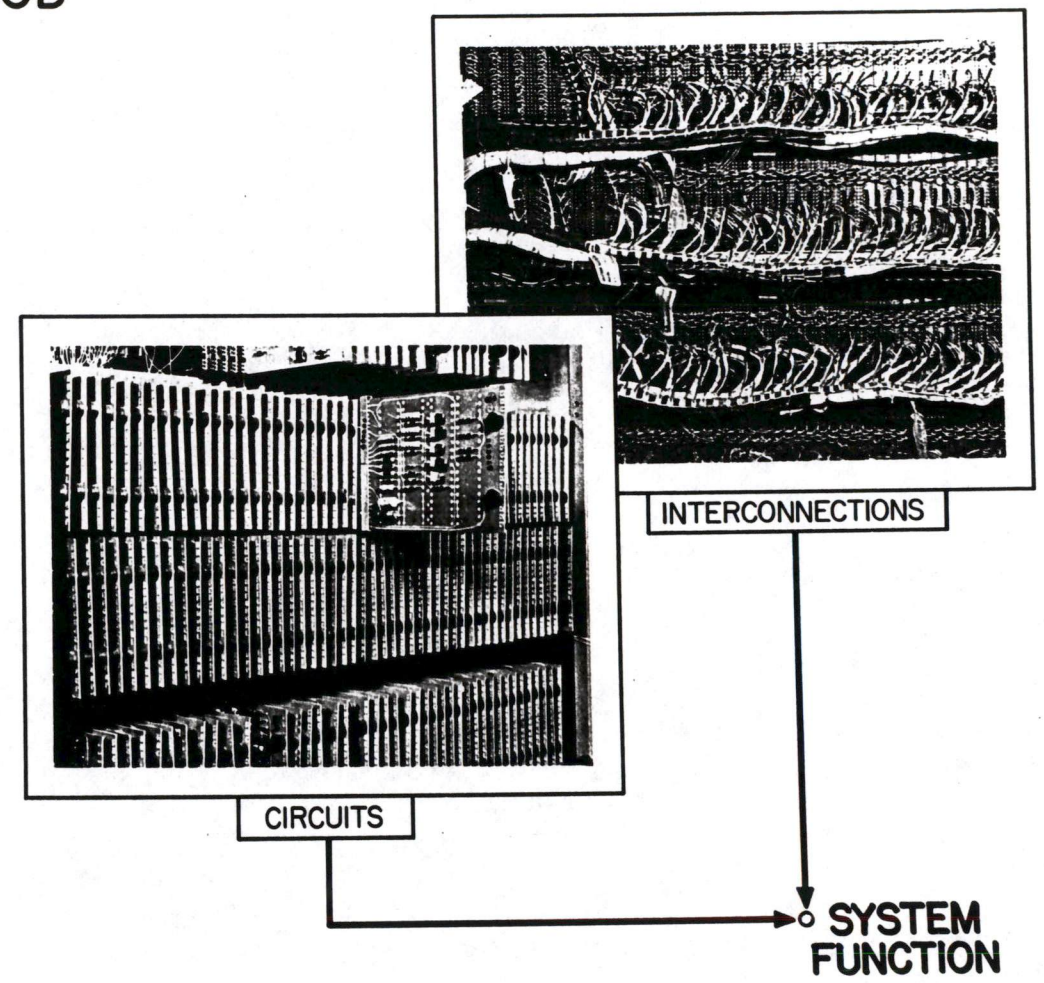
1. "Computer Design from the Programmer's Viewpoint, "W. F. Bauer, EJCC Proceedings, December 1958.

2. "New Logical and System Concepts, " R. K. Richards, EJCC Proceedings, December 1958.

3. "An Approach to Microminiature Printed Systems, " D. A. Buck and K. R. Shoulders, EJCC Proceedings, December 1958.

4. "The Impending Revolution in Computer Technology, " R. Rice, EJCC Proceedings, December 1958.

# PRESENT METHOD




**STANDARD**

- DEVICES
- "BOOLEAN" CIRCUITS
- CIRCUIT PACKAGE

FIGURE I

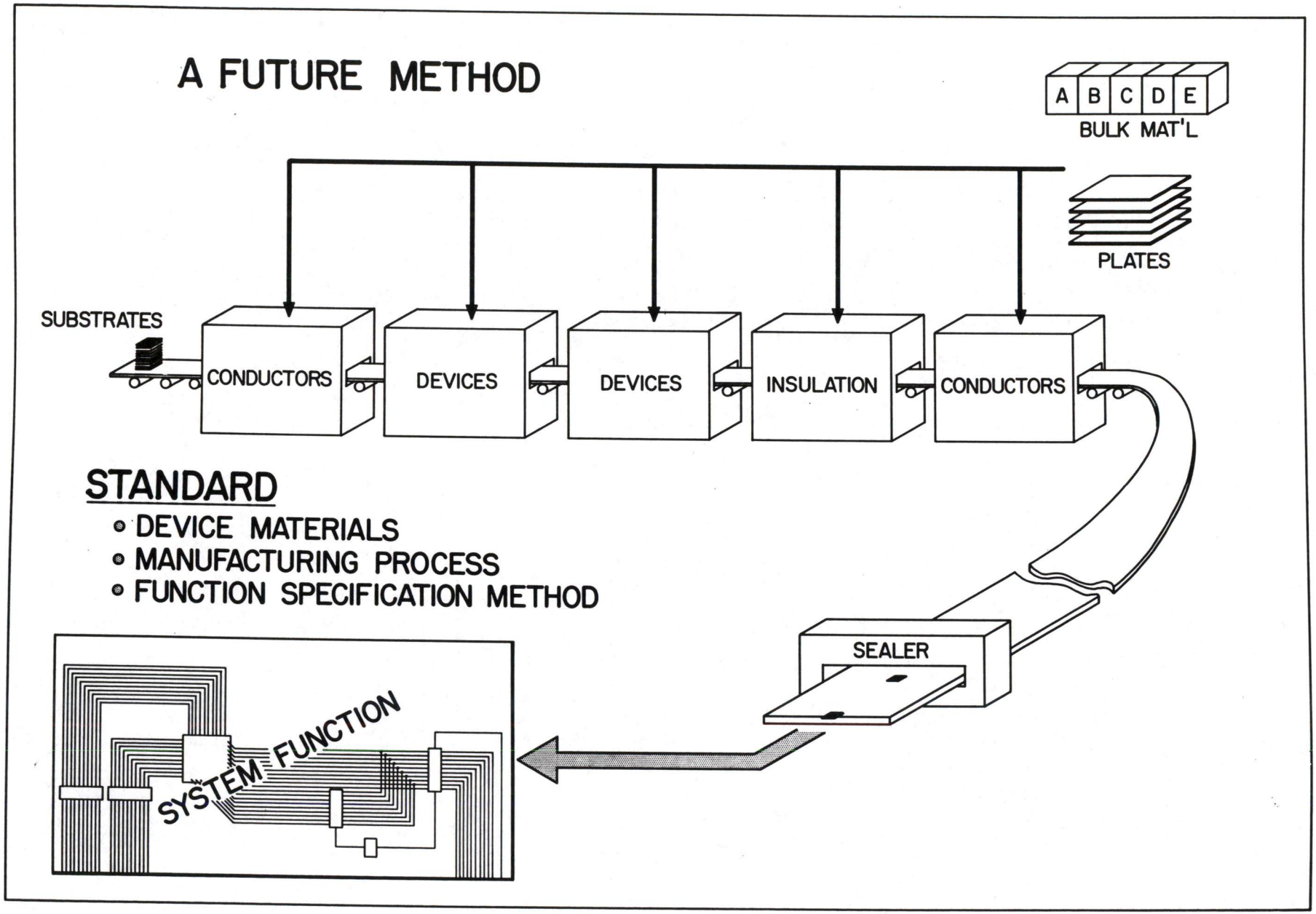
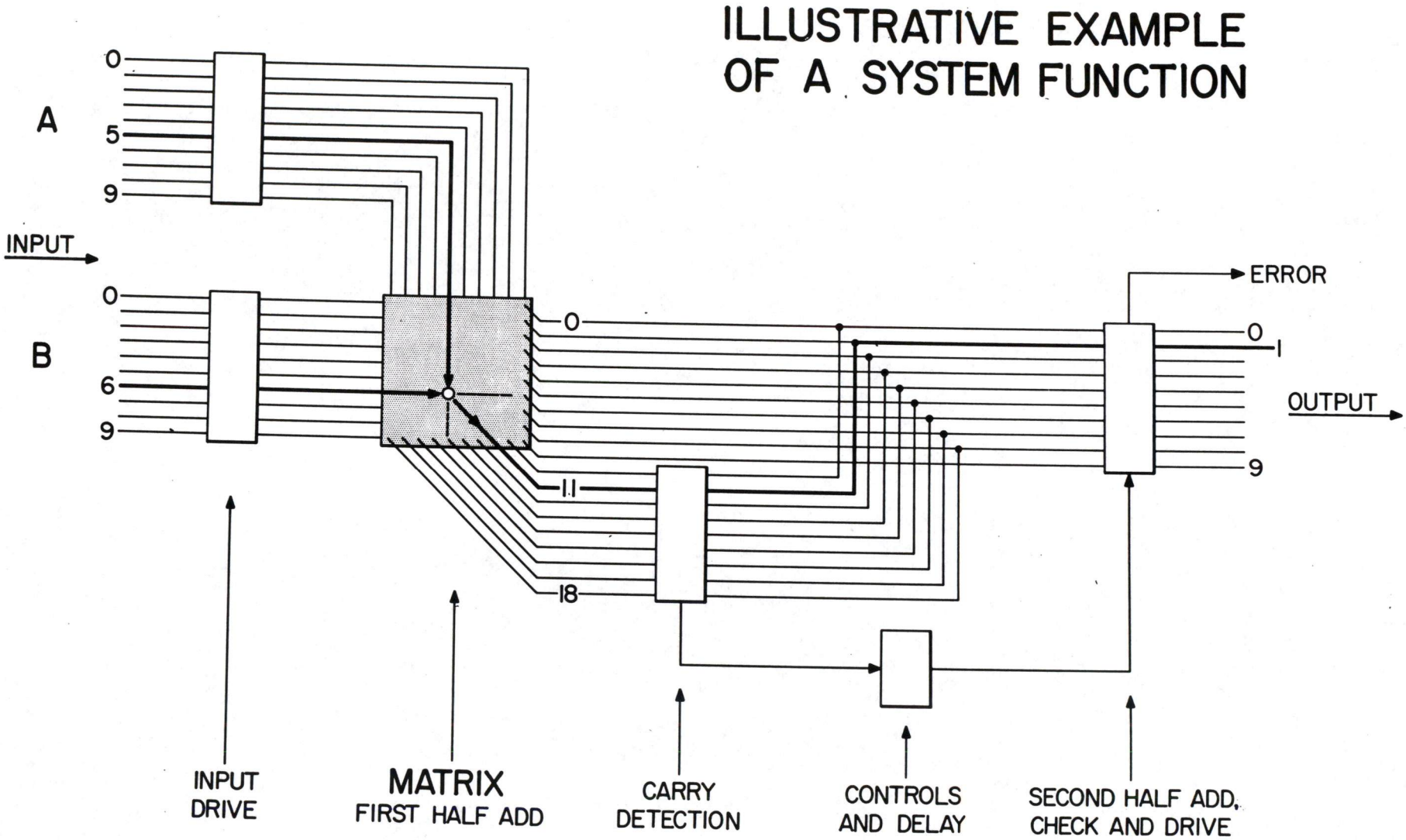

A FUTURE METHOD
A B C D E
BULK MAT'L
PLATES
SUBSTRATES
CONDUCTORS
DEVICES
DEVICES
INSULATION
CONDUCTORS
STANDARD
• DEVICE MATERIALS
• MANUFACTURING PROCESS
• FUNCTION SPECIFICATION METHOD
SEALER
SYSTEM FUNCTION

FIGURE 2




DECIMAL ADDITION — SERIAL BY DIGIT

FIGURE 3

# STANDARD "AND-INVERTER" CIRCUIT (TRL)

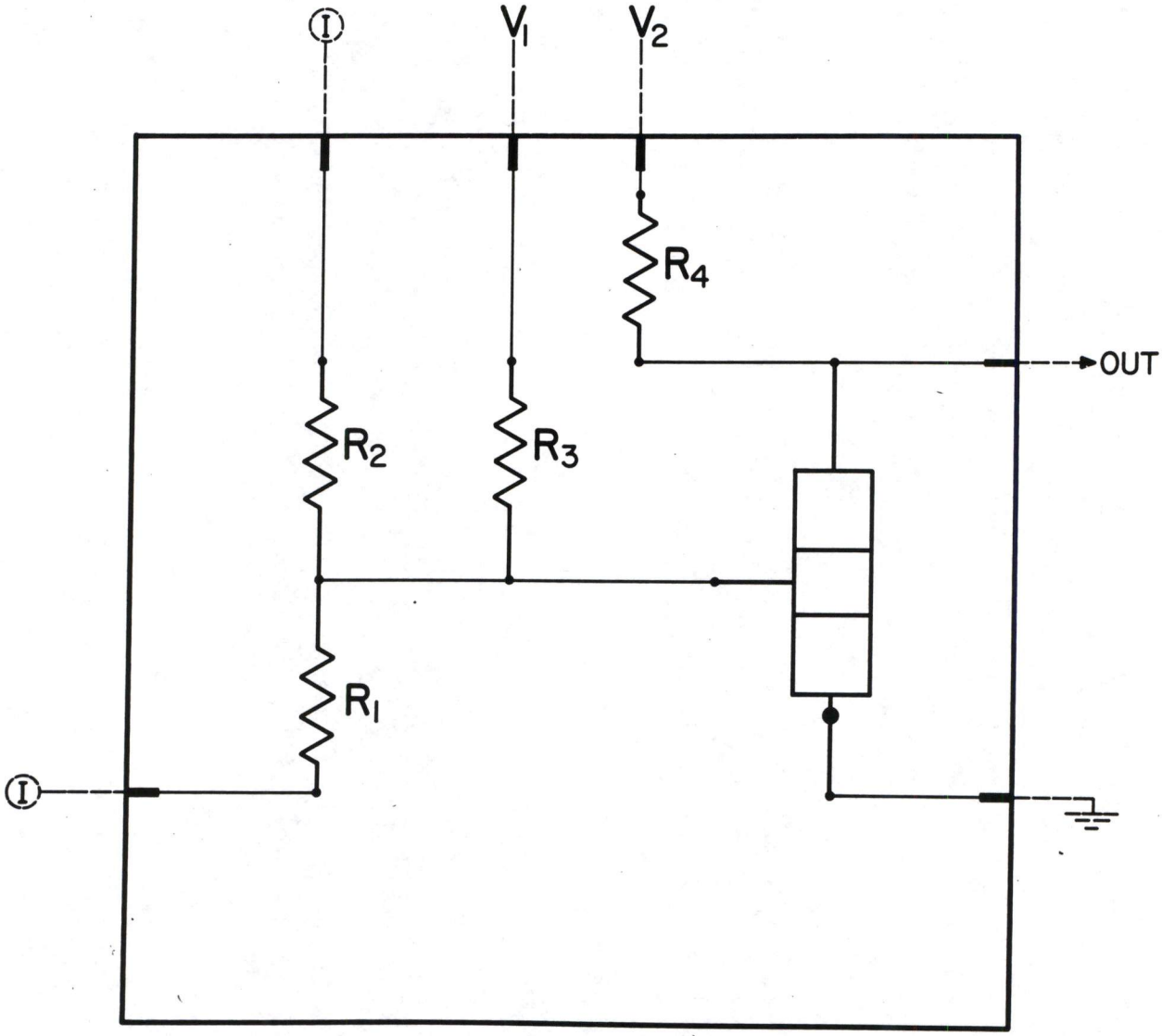



FIGURE 4

# MATRIX UTILIZING STANDARDIZED "BOOLEAN" CIRCUITS

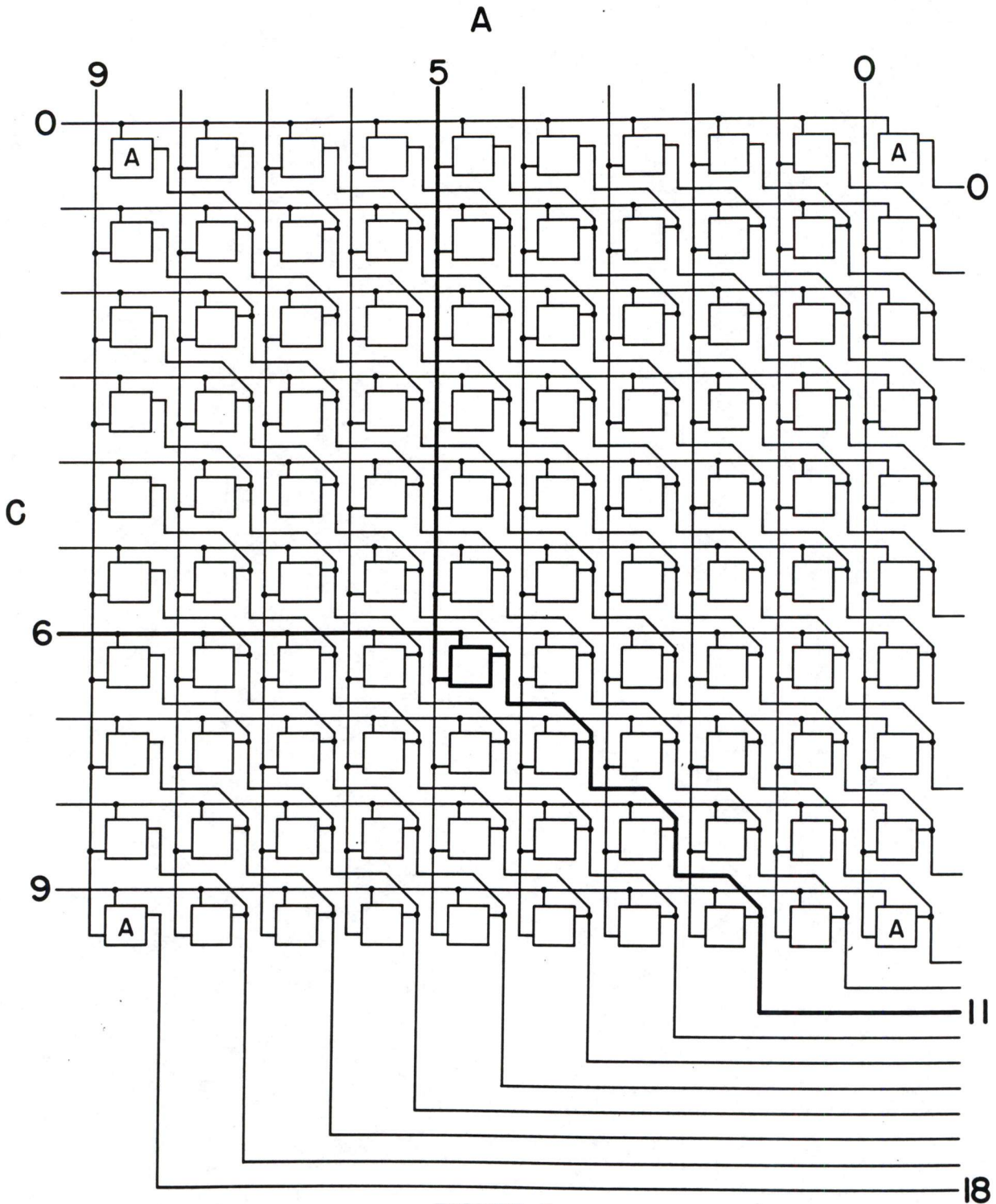



FIGURE 5

# MATRIX UTILIZING STANDARDIZED "BOOLEAN" CIRCUITS

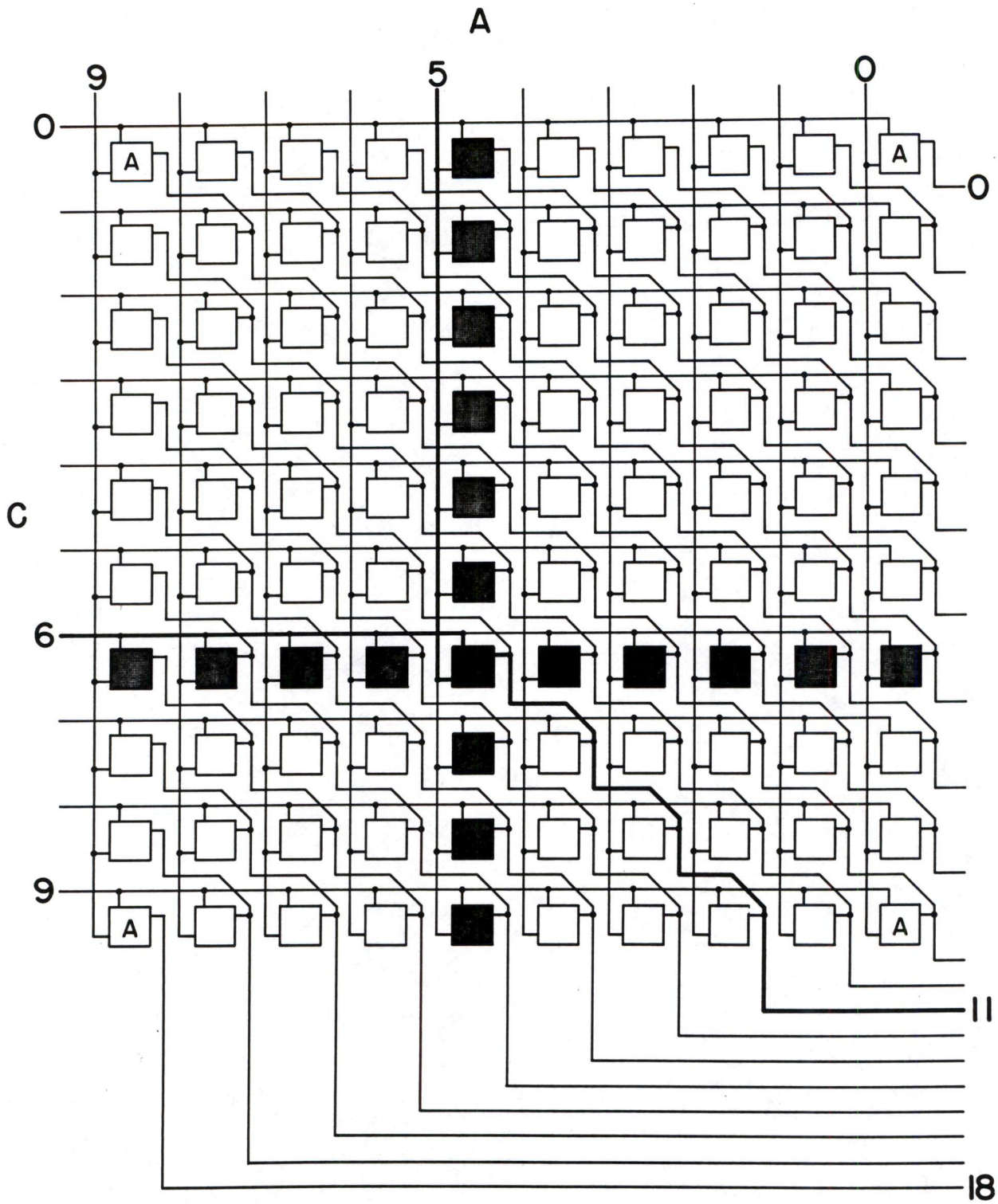



FIGURE 6

# SYSTEM TAILORED CIRCUIT (CS)

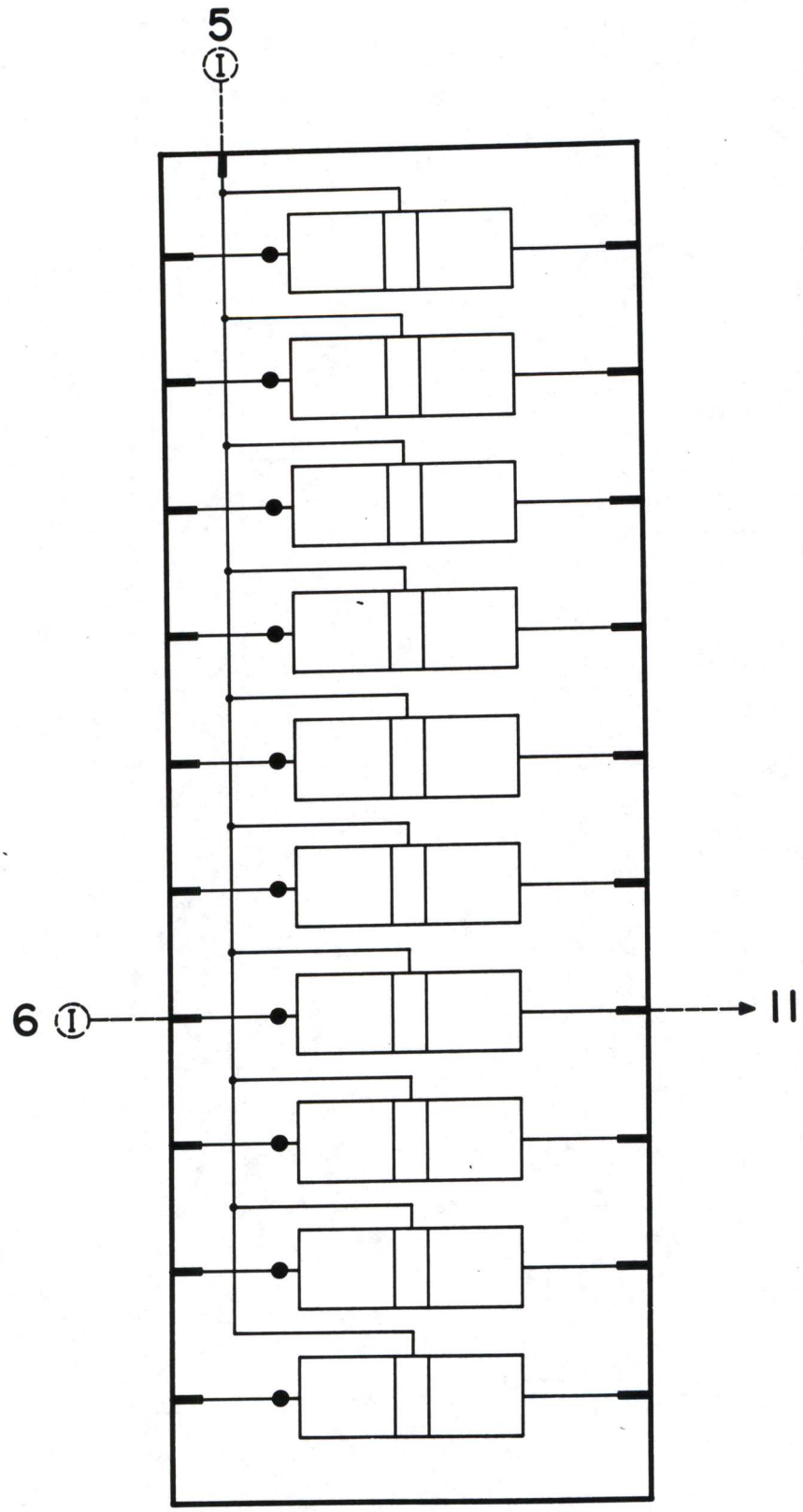



FIGURE 7

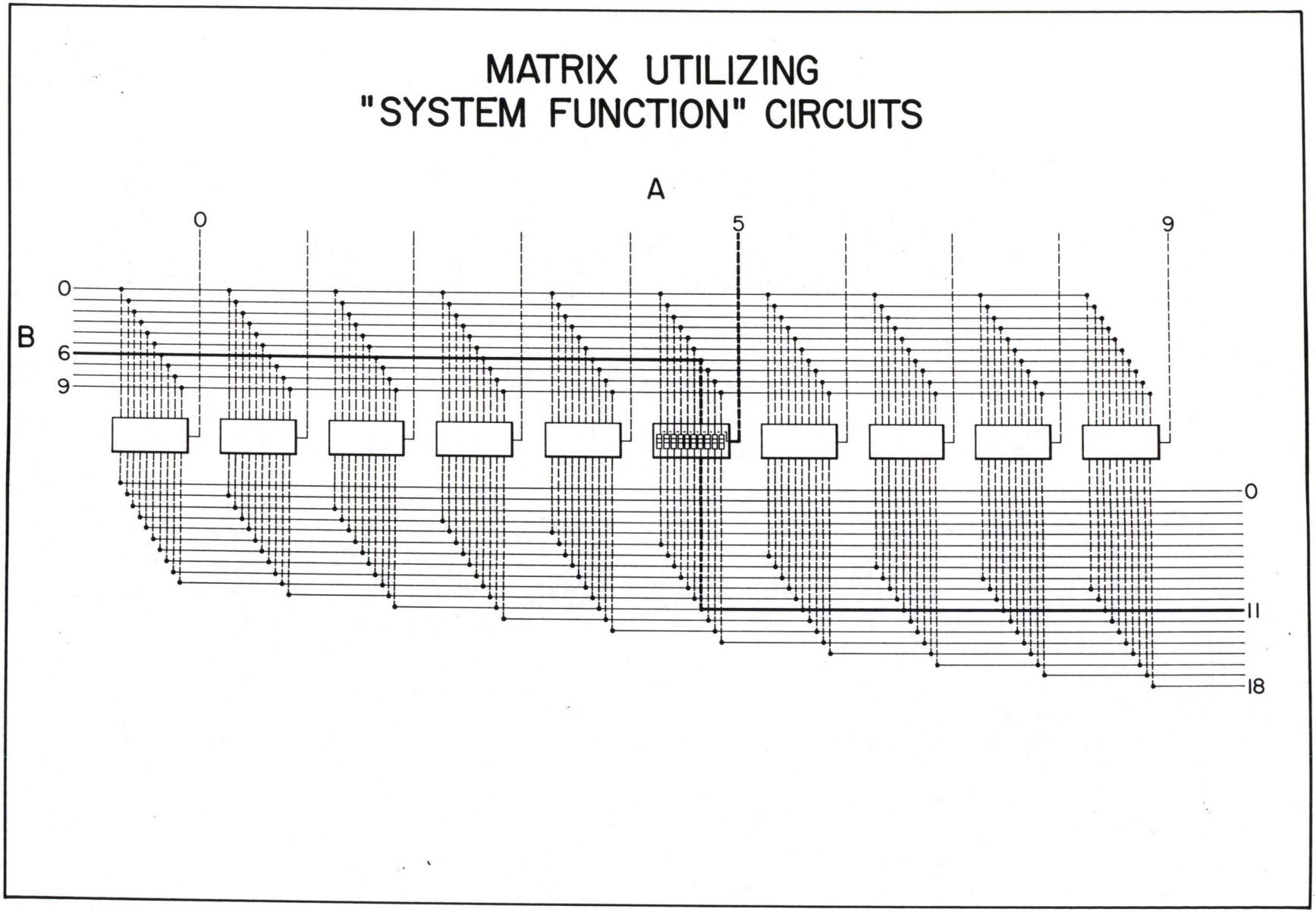
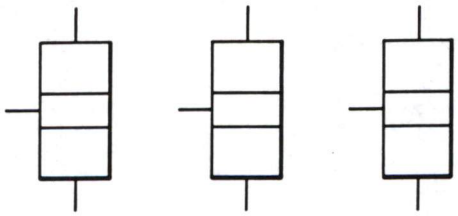

MATRIX UTILIZING
"SYSTEM FUNCTION" CIRCUITS
A
0
5
9
B
0
6
9
0
11
18

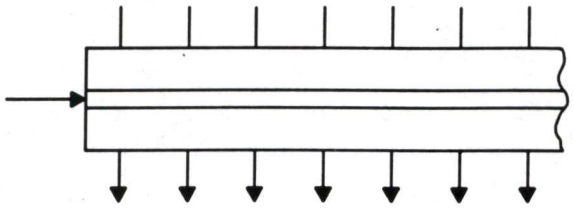
FIGURE 8

# DEVICES
# "SYSTEM TAILORED"



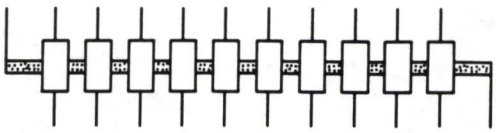
## PRESENT

- INDIVIDUAL ACTIVE ELEMENTS
- WIRED INTERCONNECTIONS



## EARLY GENERATION

- MULTI-ELEMENT "SYSTEM TAILORED" UNITS
- PRINTED OR ETCHED INTERCONNECTIONS



## LATER—FILMS

- BATCH-BULK TECHNIQUES MERGING ACTIVE DEVICES AND INTERCONNECTIONS

PROBLEM → ■ → SOLUTION

## FUTURE—MICROMINIATURIZATION

- SMALLEST ELECTRONIC ELEMENT IS TOTAL SYSTEM

FIGURE 9

# DIGITAL DATA PROCESSING
# APPROXIMATE RELATIVE COSTS

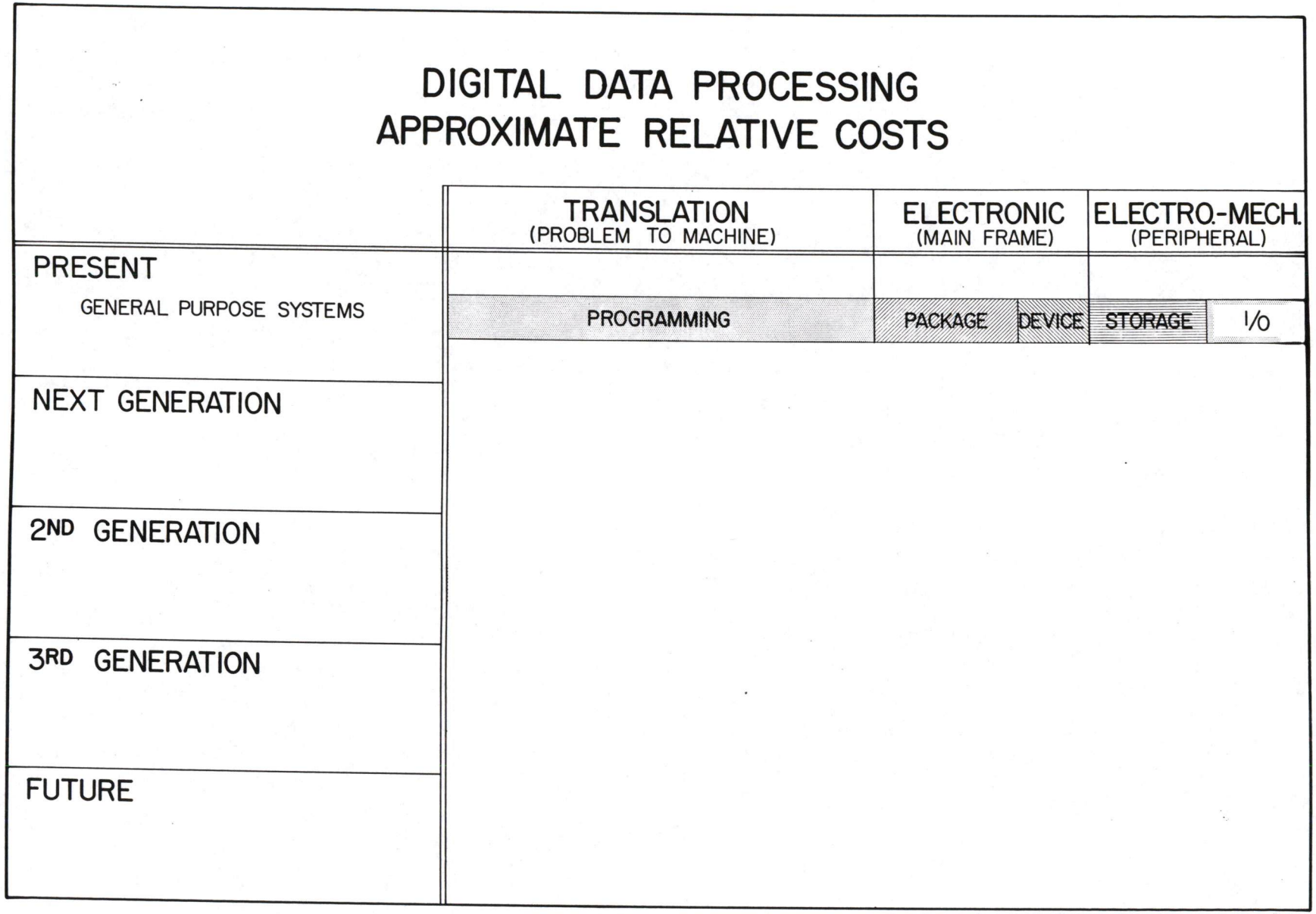



FIGURE 10

# DIGITAL DATA PROCESSING
# APPROXIMATE RELATIVE COSTS

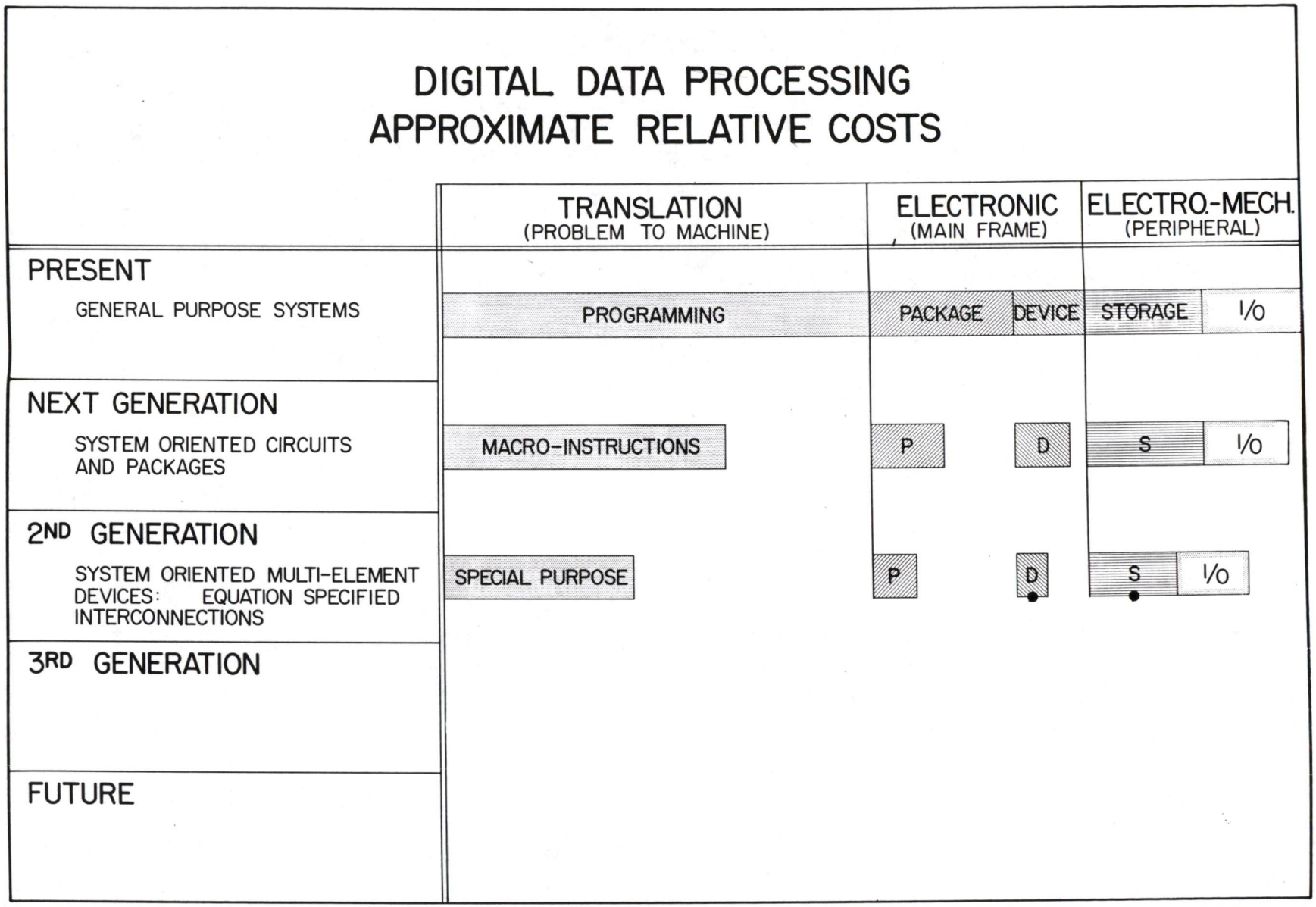



FIGURE 12

# DIGITAL DATA PROCESSING
# APPROXIMATE RELATIVE COSTS

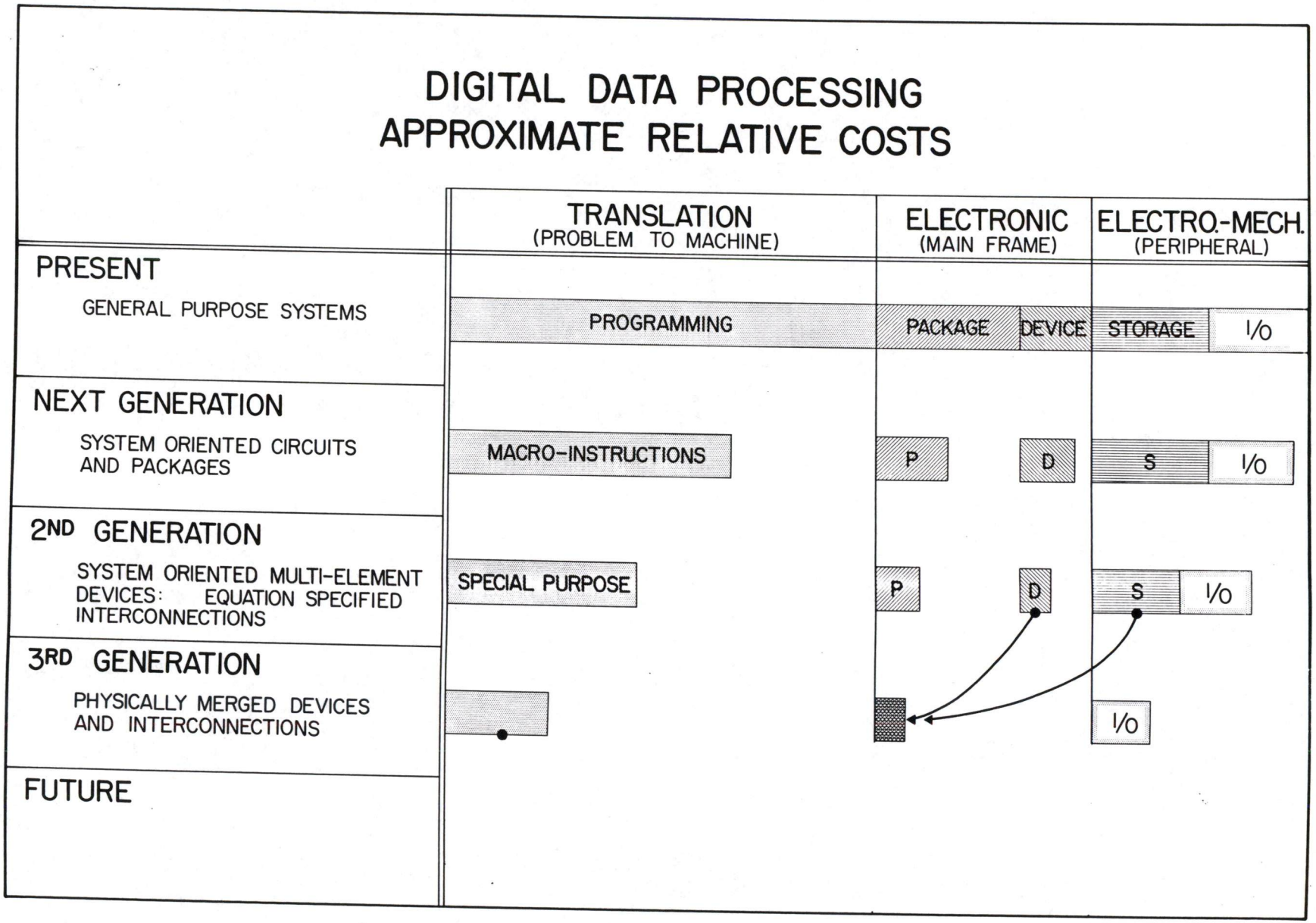



FIGURE 13

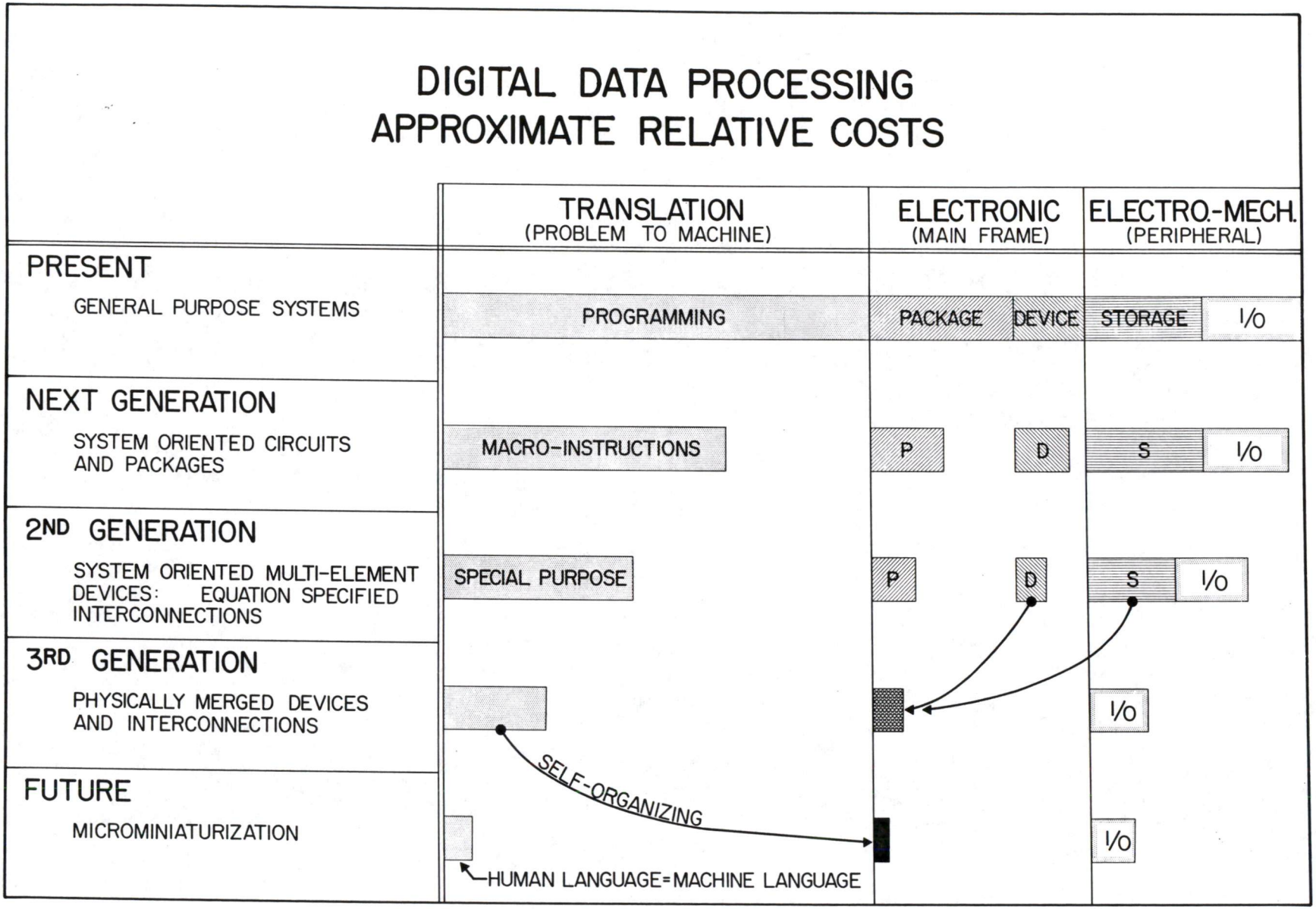
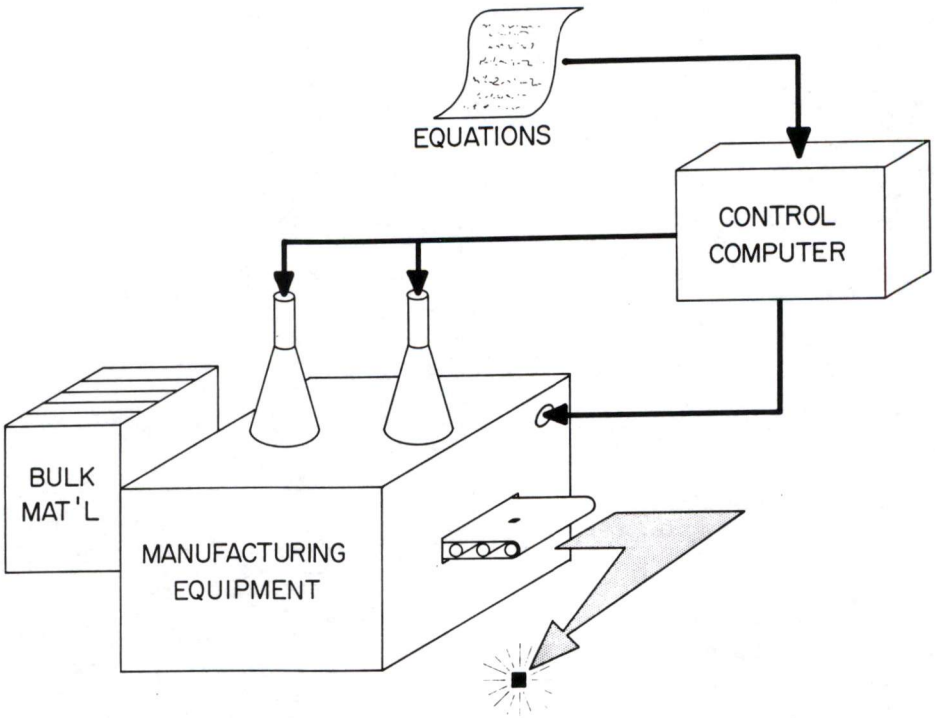

DIGITAL DATA PROCESSING
APPROXIMATE RELATIVE COSTS
TRANSLATION
(PROBLEM TO MACHINE)
ELECTRONIC
(MAIN FRAME)
ELECTRO.-MECH.
(PERIPHERAL)
PRESENT
GENERAL PURPOSE SYSTEMS
PROGRAMMING
PACKAGE
DEVICE
STORAGE
I/O
NEXT GENERATION
SYSTEM ORIENTED CIRCUITS
AND PACKAGES
MACRO-INSTRUCTIONS
P
D
S
I/O
2ND GENERATION
SYSTEM ORIENTED MULTI-ELEMENT
DEVICES: EQUATION SPECIFIED
INTERCONNECTIONS
SPECIAL PURPOSE
P
D
S
I/O
3RD GENERATION
PHYSICALLY MERGED DEVICES
AND INTERCONNECTIONS
I/O
SELF-ORGANIZING
FUTURE
MICROMINIATURIZATION
I/O
HUMAN LANGUAGE=MACHINE LANGUAGE

FIGURE 14




# FUTURE COMPUTER "ELECTRONICS"

## STANDARDIZED ON:

- BULK RAW MATERIALS
- MERGED DEVICES AND INTERCONNECTIONS
- SYSTEM FUNCTION ALGEBRA
- MANUFACTURING METHODS—COMPUTER CONTROLLED

## OBTAINING:

- EFFICIENT SPECIAL PURPOSE SYSTEMS
- PRODUCED RAPIDLY AND ECONOMICALLY

## RESULTING IN:

- MORE BRAINPOWER ON DEFINING PROBLEMS
- CHEAPER PROBLEM SOLUTION

FIGURE 15

# CRITICAL-PATH PLANNING AND SCHEDULING

by

J. E. Kelley, Jr.
and
Morgan R. Walker

Mauchly Associates, Inc.
Ambler, Pa.

November 9, 1959

## INTRODUCTION AND SUMMARY

Among the major problems facing technical management today are those involving the coordination of many diverse activities toward a common goal. In a large engineering project, for example, almost all the engineering and craft skills are involved as well as the functions represented by research, development, design, procurement, construction, vendors, fabricators and the customer. Management must devise plans which will tell with as much accuracy as possible how the efforts of the people representing these functions should be directed toward the project's completion. In order to devise such plans and implement them, management must be able to collect pertinent information to accomplish the following tasks:

1) To form a basis for prediction and planning
2) To evaluate alternative plans for accomplishing the objective
3) To check progress against current plans and objectives, and
4) To form a basis for obtaining the facts so that decisions can be made and the job can be done.

Many present project planning systems possess deficiencies resulting from techniques inadequate for dealing with complex projects. Generally, the several groups concerned with the work do their own detailed planning and scheduling -- largely independent from one another. These separate efforts lead to lack of coordination. Further, it is traditional in project work that detailed schedules be developed from gross estimates of total requirements and achievements based on past experience. The main reason for this oversimplification stems from the inability of unaided human beings to cope with sheer complexity. In consequence, many undesirable effects may arise. Some important aspects of a project, which should be taken into account at the outset, may be ignored or unrecognized. As a result, much confusion may arise during the course of the project. When

this happens, the management of the project is left to the coordinators and expediters. In such circumstances, management loses much of the control of a project and is never quite sure whether its objectives are being attained properly.

Recognizing the deficiencies in traditional project planning and scheduling procedures, the Integrated Engineering Control Group (I. E. C.) of E. I. duPont de Nemours & Co. proceeded to explore possible alternatives. It was felt that a high degree of coordination could be obtained if the planning and scheduling information of all project functions are combined into a single master plan -- a plan that integrates all efforts toward a common objective. The plan should point directly to the difficult and significant activities -- the problems of achieving the objective. For example, the plan should form the basis of a system for management by exception. That is, within the framework of the rules laid down, it should indicate the exceptions. Under such a system, management need act only when deviations from the plan occur.

The generation of such a coordinated master plan requires the consideration of much more detailed information at one time than heretofore contemplated in project work. In turn, a new approach to the whole problem of planning and scheduling large projects is required. In late 1956, I. E. C. initiated a survey of the prospects for applying electronic computers as an aid to coping with the complexities of managing engineering projects. The following were the questions of most pressing interest: To what extent can a computer-oriented system be used:

1) To prepare a master schedule for a project?
2) To revise schedules to meet changing conditions in the "most" economical way?
3) To keep management and the operating departments advised of project progress and changes?

During the course of this survey outside help was solicited. As part of their customer service, Remington Rand UNIVAC assigned the first author to the job of providing some assistance. At the time the second author represented duPont in this effort. The result of our alliance is the subject of this essay.

We made a critical analysis of the traditional approach to planning and a study of the nature of engineering projects. It quickly became apparent that if a new approach were to be successful, some technique had to be used to describe the interrelationships among the many tasks that compose a project. Further, the technique would have to be very simple and rigorous in application, if humans were to cope with the complexity of a project.

One of the difficulties in the traditional approach is that planning and scheduling are carried on simultaneously. At one session, the planner and scheduler consider -- or attempt to consider -- hundreds of details of technology, sequence, duration times, calendar deliveries and completions, and cost. With the planning and scheduling functions broken down in a step by step manner, fruitless mental juggling might be avoided and full advantage taken of the available information.

Accordingly, the first step in building a model of a project planning and scheduling system was to separate the functions of planning from scheduling. We defined planning as the act of stating what activities must occur in a project and in what order these activities must take place. Only technology and sequence were considered. Scheduling followed planning and is defined as the act of producing project timetables in consideration of the plan and costs.

The next step was to formulate an abstract model of an engineering project. The basic elements of a project are activities or jobs: determination of specs, blueprint preparation, pouring foundations, erecting steel, etc. These activities are represented graphically in the form of an arrow diagram which permits the user to study the technological relations among them.

Cost and execution times are associated with each activity in the project. These factors are combined with the technological relations to produce optimal direct cost schedules possessing varying completion dates. As a result, management comes into possession of a spectrum of possible schedules, each having an engineered sequence, a known elapsed time span, a known cost function, and a calendar fit. In the case of R & D projects, one obtains "most probable" schedules. From these schedules, management may select a schedule which maximizes return on investment or some other objective criterion.

The technique that has been developed for doing this planning and scheduling is called the Critical-Path Method. This name was selected because of the central position that critical activities in a project play in the method. The Critical-Path Method is of general interest from several aspects:

1) It may be used to solve a class of "practical" business problems
2) It requires the use of modern mathematics
3) Large-scale computing equipment is required for its full implementation
4) It has been programmed for three computers -- UNIVAC I, 1103A and 1105 with a Census Bureau configuration
5) It has been put into practice

In what follows we will attempt to amplify these points. We will describe various aspects of the mathematical model first. The mathematics involved will be treated rather superficially, a detailed development being reserved for a separate paper. The second part of this essay will cover the experience and results obtained from the use of the Critical-Path Method.

# PART I: ANALYSIS OF A PROJECT

## 1. PROJECT STRUCTURE

Fundamental to the Critical-Path Method is the basic representation of a project. It is characteristic of all projects that all work must be performed in some well-defined order. For example, in construction work, forms must be built before concrete can be poured; in R & D work and product planning, specs must be determined before drawings can be made; in advertising, artwork must be made before layouts can be done, etc.

These relations of order can be shown graphically. Each job in the project is represented by an arrow which depicts (1) the existence of the job, and (2) the direction of time-flow (time flows from the tail to the head of the arrow). The arrows then are interconnected to show graphically the sequence in which the jobs in the project must be performed. The result is a topological representation of a project. Figure 1 typifies the graphical form of a project.

Several things should be noted. It is tacitly assumed that each job in a project is defined so that it is fully completed before any of its successors can begin. This is always possible to do. The junctions where arrows meet are called events. These are points in time when certain jobs are completed and others must begin. In particular there are two distinguished events, origin and terminus, respectively, with the property that origin precedes and terminus follows every event in the project.

Associated with each event, as a label, is a non-negative integer. It is always possible to label events such that the event at the head of an arrow always has a larger label than the event at the tail. We assume that events are always labeled in this fashion. For a project, $P$, of $n + 1$ events, origin is given the label $0$ and terminus is given the label $n$.

The event labels are used to designate jobs as follows: if an arrow connects event $i$ to event $j$, then the associated job is called job $(i, j)$.

During the course of constructing a project diagram, it is necessary

to take into account a number of things pertaining to the definition of each job. Depending upon such factors as the purpose for making the project analysis, the nature of the project, and how much information is available, any given job may be defined in precise or very broad terms. Thus, a job may consist of simply typing a report, or it might encompass all the development work leading up to the report plus the typing. Someone concerned with planning the development work should be interested in including the typing as a job in the project while those concerned with integrating many small development projects would probably consider each such project as an individual job.

Further, in order to prepare for the scheduling aspects of project work, it is necessary to consider the environment of each job. For example, on the surface it may be entirely feasible to put 10 men on a certain job. However, there may only be enough working space for five men at a time. This condition must be included in the job's definition. Again, it may technically be possible to perform two jobs concurrently. However, one job may place a safety hazard on the other. In consequence, the first job must be forced to follow the second.

Finally, the initiation of some jobs may depend on the delivery of certain items -- materials, plans, authorization of funds, etc. Delivery restraints are considered jobs, and they must be included in the project diagram. A similar situation occurs when certain jobs must be completed by a certain time. Completion conditions on certain jobs also may be handled, but in a more complicated fashion, by introducing arrows in the project diagram.

Project diagrams of large projects, although quite complicated, can be constructed in a rather simple fashion. A diagram is built up by sections. Within each section the task is accomplished one arrow at a time by asking and answering the following questions for each job:

1) What immediately precedes this job?
2) What immediately follows this job?
3) What can be concurrent with this job?

By continually back-checking, the chance of making omissions is small. The individual sections then are connected to form the complete project diagram. In this way, projects involving up to 1600 jobs have been handled with relative ease.

From a scientific viewpoint, the idea of diagramming the technological relations among the jobs in a project is almost trivial. Such diagrams are used in many engineering and mathematical applications. However, diagramming is an innovation in project work which has given planners several benefits:

1) It provides a disciplined basis for planning a project.
2) It provides a clear picture of the scope of a project that can be easily read and understood.
3) It provides a vehicle for evaluating alternative strategies and objectives.
4) It tends to prevent the omission of jobs that naturally belong to the project.
5) In showing the interconnections among the jobs it pinpoints the responsibilities of the various operating departments involved.
6) It is an aid to refining the design of a project.
7) It is an excellent vehicle for training project personnel.

## 2. CALENDAR LIMITS ON ACTIVITIES

Having a diagram of a project is only the first step in analyzing a project. Now the plan must be put on a timetable to obtain a schedule.

In order to schedule a project, it is necessary to assign elapsed time durations to each job. Depending on the nature of the project this data may be known deterministically or non-deterministically. Another way to say this is that the duration of each job is a random variable taken from an approximately known distribution. The duration of a job is deterministic when the variance of the distribution is small. Otherwise it is non-deterministic.

The Deterministic Case. On the basis of estimated elapsed times, we may compute approximations to the earliest and latest start and completion times for each job in a project. This information is important not only for putting a schedule on the calendar, but also for establishing rigorous limits to guide operating personnel. In effect, it tells those responsible for a job when to start worrying about a slippage and to report this fact to those responsible for the progress of the project. In turn, when this information is combined with a knowledge of the project's topological structure, higher management can determine when and how to revise the schedule and who will be affected by the change. This kind of information is not determined accurately by traditional methods. What this information provides is the basis for a system of management by exception.

Let us assume that the project, $P$, of $n+1$ events, starts at relative time $0$. Relative to this starting time each event in the project has an earliest time occurance. Denote the earliest time for event $i$ by $t_i^{(0)}$ and the duration of job $(i,j)$ by $y_{ij}$. We may then compute the values of $t_i^{(0)}$ inductively as follows:

$$(1) \quad \begin{cases} t_0^{(0)} = 0 \\ t_j^{(0)} = \max \left[ y_{ij} + t_i^{(0)} \mid i < j,\ (i,j) \in P \right], \quad 1 \leq j \leq n. \end{cases}$$

Similarly, we may compute the latest time at which each event in the project may occur relative to a fixed project completion time. Denote the latest time for event $i$ by $t_i^{(1)}$. If $\lambda$ is the project completion time (where $\lambda \geq t_n^{(0)}$) we obtain

$$(2) \quad \begin{cases} t_n^{(1)} = \lambda \\ t_i^{(1)} = \min \left[ t_j^{(1)} - y_{ij} \mid i < j,\ (i,j) \in P \right], \quad 0 \leq i \leq n-1. \end{cases}$$

Having the earliest and latest event times we may compute the

following important quantities for each job, $(i, j)$, in the project:

| | |
|---|---|
| Earliest start time | $= t_i^{(0)}$ |
| Earliest completion time | $= t_i^{(0)} + y_{ij}$ |
| Latest start time | $= t_j^{(1)} - y_{ij}$ |
| Latest completion time | $= t_j^{(0)}$ |
| Maximum time available | $= t_j^{(1)} - t_i^{(0)}$ |

If the maximum time available for a job equals its duration the job is called critical. A delay in a critical job will cause a comparable delay in the project completion time. A project will contain critical jobs only when $\lambda = t_n^{(0)}$. If a project does contain critical jobs, then it also contains at least one contiguous path of critical jobs through the project diagram from origin to terminus. Such a path is called a critical-path.

If the maximum time available for a job exceeds its duration, the job is called a floater. Some floaters can be displaced in time or delayed to a certain extent without interfering with other jobs or the completion of the project. Others, if displaced, will start a chain reaction of displacements downstream in the project.

It is desirable to know, in advance, the character of any floater. There are several measures of float of interest in this connection. The following measures are easily interpreted:

| | |
|---|---|
| Total Float | $= t_j^{(1)} - t_i^{(0)} - y_{ij}$ |
| Free Float | $= t_j^{(0)} - t_i^{(0)} - y_{ij}$ |
| Independent Float | $= \max\left(0, t_j^{(0)} - t_i^{(1)} - y_{ij}\right)$ |
| Interfering Float | $= t_j^{(1)} - t_j^{(0)}$. |

Non-Deterministic Schedules. Information analogous to that obtained in the deterministic case is certainly desirable for the non-deterministic case.

It would be useful for scheduling applied research directed toward a well-defined objective.

However, in attempting to develop such information some difficulties are encountered which do not seem easily resolved. These difficulties are partly philosophical and partly mathematical. Involved is the problem of defining a "meaningful" measure for the criticalness of a job that can be computed in a "reasonable" fashion.

Although a complete analysis of this situation is not germane to the development of the Critical-Path Method, it is appropriate, however, to indicate some concepts basic to such an analysis. Thus, in the non-deterministic case we assume that the duration, $y_{ij}$, of activity $(i,j)$ is a random variable with probability density $G_{ij}(y)$. As a consequence it is clear that the time at which an event occurs is also a random variable, $t_j$, with probability density $H_j(t)$. We assume that event 0 is certain to occur at time 0. Further, on the assumption that it is started as soon as possible, we see that $t_i + y_{ij}$, the completion time for job $(i,j)$, is a random variable with probability density $S_{ij}(x)$: [1]

$$(3) \qquad S_{ij}(x) = \begin{cases} G_{ij}(x), & \text{if } i = 0 \\ \int_{-\infty}^{\infty} H_i(u)\, G_{ij}(x-u)\, du, & (i,j) \in P. \end{cases}$$

Assuming now that an event occurs at the time of the completion of the last activity preceding it we can easily compute the probability density, $H_j(t)$, of

[1] See M. G. Kendall, "The Advanced Theory of Statistics", Vol. 1, J. B. Lippincott Co., 1943, p. 247.

$$t_j = \max [x_{ij} \mid (i,j) \in P , i < j ] ,$$

where $x_{ij}$ is taken from $S_{ij}(x)$:

$$(4) \qquad H_j(t) = \sum_{(i,j) \in P} S_{ij}(t) \prod_{\substack{(k,j) \in P \\ k \neq i}} \int_{-\infty}^{t} S_{kj}(u)\,du , \quad 1 \leq j \leq n .$$

Several methods are available for approximating $S_{ij}(x)$ and $H_j(t)$. The one which suits our taste is to express $G_{ij}(y)$ in the form of a histogram with equal class intervals. The functions $S_{ij}(x)$ and $H_j(t)$ are then histograms also and are computed in the obvious way by replacing integrals by sums. It would seem that in practice one can afford to have fairly large class intervals so that the chore of computing is quite reasonable.

In computing $S_{ij}(x)$ and $H_i(t)$ above we assumed that job $(i,j)$ was started at the time of the occurance of $t_i$. For various reasons it may not be desirable to abide by this assumption. Indeed, it may be possible to delay the start of job $(i,j)$ to a fair extent after the actual occurance of $t_i$ without changing the character of $H_j(t)$. However, the assumption we have made does provide a probabilistic lower bound on the start time for job $(i,j)$. By analogy with the deterministic case we may think of $H_i(t)$ as the probability density of the earliest start time for job $(i,j)$. Similarly, $S_{ij}(x)$ in (3) then becomes the probability density of the earliest completion time for job $(i,j)$. In this sense, (4) is the probabilistic analogue of (1).

It is desirable to be able to measure the criticalness of each job in the project. Intuitively one is tempted to use the probabilistic analogue of (2), running the project backward from some fixed or random completion time as was done in the deterministic case. In this way one might hope to obtain information about the latest times at which events can occur, so that probabilistic measures of float might be obtained. It appears that this is a false hope since, among other things, such a procedure assumes that the project start time is a random variable and not a certain event. (The project start time can always be assumed certain, simply by making lead time for

the project start on the day the calculations are made.)

To proceed further we must introduce the notion of "risk" in defining the criticalness of a job. On the basis of this definition one would hope to obtain probabilistic measures for float which would be useful for setting up a system for management by exception. We will not explore these possibilities further here.

## 3. THE PROJECT COST FUNCTION

In the deterministic case, the durations of jobs may sometimes be allowed to vary within certain limits. This variation may be attributed to a number of factors. The elapsed-time duration of a job may change as the number of men put on it changes, as the type of equipment or method used changes, as the work week changes from 5 to 6 to 7 days, etc. Thus, management has considerable freedom to choose the elapsed-time duration of a job, within certain limitations on available resources and the technology and environment of the job. Every set of job durations selected will lead to a different schedule and, in consequence, a different project duration. Conversely, there are generally many ways to select job durations so that the resulting schedules have the same shortest time duration.

Faced with making a choice, management must have some way of evaluating the merits of each possibility. In traditional planning and scheduling systems such a criterion is not too well defined. In the present context, however, there are several possibilities. The one we will focus our attention upon is cost.

Job Cost. When the cost (labor, equipment and materials) of a typical engineering job varies with elapsed-time duration it usually approximates the form of the curve of Figure 2. This is what is usually called "direct" cost. Costs arising from administration, overhead, and distributives are not included.

Note that when the duration of job $(i,j)$ equals $D_{ij}$, the cost is a minimum. On the surface, this is a desirable point at which to operate. Certainly management would seldom ever elect to require the job to take longer than the optimal method time. We call $D_{ij}$ the normal duration for job $(i,j)$. However, exogenous conditions may require that a job be expedited. This may be done in a variety of ways. But in any case there is a limit to how fast a job may be performed. This lower bound is denoted by $d_{ij}$ in Figure 2 and is called the crash duration for job $(i,j)$.

It is thus reasonable to assume that the duration $y_{ij}$ of job $(i,j)$ satisfies

$$0 \leq d_{ij} \leq y_{ij} \leq D_{ij} . \tag{5}$$

The cost of job $(i,j)$ is now approximated in a special way over the range defined by inequalities (5). The type of approximation used is dictated by the mathematical technique involved in what follows. Thus, we must assume that the approximate cost function is a piecewise linear, non-increasing and convex function of $y_{ij}$. Usually in practice insufficient data is available to make more than a linear approximation. There are exceptions, of course.

In the linear case we may write

$$\text{Cost of Job } (i,j) = a_{ij} y_{ij} + b_{ij} , \tag{6}$$

$$\text{where } a_{ij} \leq 0 \text{ and } b_{ij} \geq 0 .$$

Minimum Project Costs. On the basis of job cost functions just developed we can determine the (direct) cost of any particular schedule satisfying inequalities (5) by simply summing the individual job costs. That is,

$$\text{Project (Direct) Cost} = \sum_{(i,j) \in P} (a_{ij} y_{ij} + b_{ij}) \tag{7}$$

It is clear that there are generally many ways that job durations may be selected so that the earliest completion times of the resulting schedules are all equal. However, each schedule will yield a different value of (7), the

project cost. Assuming that all conditions of the project are satisfied by these schedules, the one which costs the least invariably would be selected for implementation.

It is therefore desirable to have a means of selecting the least costly schedule for any given feasible earliest project completion time. Within the framework we have already constructed, such "optimal" schedules are obtained by solving the following linear program: Minimize (7) subject to (5) and

$$y_{ij} \leq t_j - t_i \,, \quad (i,j) \in P \,, \tag{8}$$

and

$$t_0 = 0 \,, \quad t_n = \lambda \,. \tag{9}$$

Inequalities (8) express the fact that the duration of a job cannot exceed the time available for performing it. Inequalities (9) require the project to start at relative time 0 and be completed by relative time $\lambda$. Because of the form of the individual job cost functions, within the limits of most interest, $\lambda$ is also the earliest project completion time.

At this point it should be noted that the case where each job cost function is non-increasing, piecewise linear and convex is also reducible to a parametric linear program (see [7] and [8]). It does not add anything essential here to consider this more generalized form.

A convenient tool for generating schedules for various values of $\lambda$ is the method of parametric linear programming with $\lambda$ as the parameter. Intuitively, this technique works as follows. Initially, we let $y_{ij} = D_{ij}$ for every job in the project. This is called the all-normal solution. We then assume that each job is started as early as possible. As a result we can compute $t_i^{(0)}$ for all events. In particular, the earliest project completion time for this schedule is $\lambda = t_n^{(0)}$. By the nature of the job cost functions this schedule is also a minimum cost schedule for $\lambda = t_n^{(0)}$. We now force a reduction in the project completion time by expediting certain of the critical jobs -- those jobs that control project completion time. Not all critical jobs

are expedited, but only those that drive the project cost up at a minimum rate as the project completion time decreases. As the project completion is reduced, more and more jobs become critical and thus there is a change in which jobs are ~~so~~ to be expedited. This process is repeated until no further reduction in project completion time is possible.

Mathematically speaking, the process utilizes a primal-dual algorithm (see [6]). The restricted dual problem is a network flow problem involving both positive upper and lower bound capacity restrictions. A form of the Ford-Fulkerson network flow algorithm [3] is used to solve it. The critical jobs that are expedited at each stage of the process correspond to a cut set in the graph of all critical jobs.

This process produces a spectrum of schedules (characteristic solutions in the linear programming sense) each at minimum total (direct) cost for its particular duration. When the costs of these schedules are plotted versus their respective durations, we obtain a non-increasing, piecewise linear, convex function as depicted in Figure 3. This function is called the project cost curve.

Uses of the Project Cost Curve. The project cost curve only reflects the direct costs (manpower, equipment and materials) involved in executing a project. However, other costs are involved which contribute to the total project cost, such as overhead and administrative costs and perhaps even penalties for not completing a project or some portion of it by a certain time. These external costs must be taken into account when management plans how the project should be implemented relative to overall objectives.

Relative to these external costs there are at least two types of considerations that management may make:

1) The (direct) cost curve for the project may be compared with the indirect cost of overhead and administration to find a schedule which minimizes the investment cost.
2) The investment cost curve may be compared with market losses, as when it is desired to meet the demands of a rising market in a competitive situation.

The schedule selected in this case is one which maximizes return on investment.

## 4. MANPOWER LEVELING

As developed in this paper, the Critical-Path Method is based primarily on the technological requirements of a project. Considerations of available manpower and equipment are conspicuous by their absence. All schedules computed by the technique are technologically feasible but not necessarily practical. For example, the equipment and manpower requirements for a particular schedule may exceed those available or may fluctuate violently with time. A means of handling these difficulties must therefore be sought -- a method which "levels" these requirements.

Here we will outline the approach we have taken to this problem. We restrict the discussion to manpower, similar considerations being applicable to leveling equipment requirements.

The term "manpower leveling" does not necessarily mean that the same number of men should be used throughout the project. It usually means that no more men than are available should be used. Further, if this requirement is met, one should not use the maximum number of men available at one instant in time and very few the very next instant of time.

The difficult part of treating the manpower leveling problem from a mathematical point of view is the lack of any explicit criteria with which the "best" use of manpower can be obtained. Under critical examination, available levels of manpower and also changes in level are established arbitrarily. This situation exists to some degree regardless of the organization involved. Even in the construction industry, where the work is by nature temporary, the construction organization desires the reputation of being a consistent "project life" employer. The organization wants the employee to feel that once "hired on" he can be reasonably sure of several months' work at the very least. In plants and in technical and professional engineering fields the same situation exists but with more severity. The employee is more acutely aware of

"security", and the employer much more keenly aware of the tangible costs of recruitment and layoff as well as the intangible costs of layoff to his overall reputation and well-being.

In most organizations idle crafts and engineers or the need for new hires are treated with overwhelming management scrutiny. This is an excellent attitude, but too often this consideration is short range and does not consider long range requirements.

The following approaches to this problem have been made:

Incorporating Manpower Sequences. It is possible to incorporate manpower availability in the project diagram. However, this approach can cause considerable difficulty in stating the diagram and may lead to erroneous results. Therefore, we recommend that this approach be dropped from consideration.

For example, assume there are three jobs -- A, B, and C -- that, from a technological viewpoint, can occur concurrently. However, each job requires the same crew. We might avoid the possibility that they occur simultaneously by requiring that A be followed by B, followed by C. It is also possible to state five other combinations -- ACB, BCA, BAC, CAB, and CBA.

If we assume that this example occurs many times in a large arrow diagram, then there is not one, but a very large number of possible diagrams that can be drawn.

Now suppose a manpower sequence was not incorporated in the diagram and schedules were computed. It could be that the float times available for jobs A, B, and C are sufficient to perform the jobs in any of the six possible time sequences. However, by incorporating manpower sequences, we would never really know the true scheduling possibilities.

Examining Implied Requirements. Currently this method is performed manually and has been successfully used by applications personnel. It is possible to do much of the work involved by computer but, thus far, computer programs have not been prepared.

In preparing the work sheets for each activity, a statement is made of how many men per unit of time by craft are required for each duration. The planning and scheduling then proceeds in the manner prescribed by the Critical-Path Method. After a schedule is selected from all of the computed

schedules, work on manpower leveling starts.

The first task is to tabulate the force required to execute the jobs along the critical path. Manpower commitments must be made to do these jobs at specific calendar dates. If manpower is not available, a longer duration schedule must be selected and the force requirements re-evaluated.

If adequate manpower is available to perform the critical jobs, then the total work force required by time units is tabulated. This is done by assuming every job starts at its earliest start date. The tabulation also is done, except assuming that every job starts at its latest start date.

Two total force curves result. These are then examined to be sure that they conform with some implicit statement of desired force. If not, the floaters are displaced to smooth the force curve. (In practice it has been found that one should displace the jobs with the least float first.)

During the tabulation and leveling processes, sub-totals are kept by craft to ensure that, even though total force may be all right, craft restrictions also are met.

The smoothing (a purely heuristic process) is done until the desired force and craft curves are obtained, or until it is discovered that the schedule requires an unavailable force. In this case, the next longer schedule is selected, and the process is repeated until satisfactory results are obtained.

In one actual case, it was determined after attempts at smoothing that 27 mechanics were required when only 8 were available. Smoothing for this condition meant about a 20% lengthening of the critical path. Armed with this information, the planning and scheduling staff placed in management's hands a quantitative measure of the meaning of a manpower shortage so that, in advance, corrective action could be taken.

Solving for Best Fit. A procedure has been developed for computer programming that again is subjective in approach. One does not "solve" for the "best" force on the basis of some objective criteria. Rather, one states in advance what is "best" and then attempts to find the "best" fit.

The procedure is similar to examining the implied force requirements.

The total force curve desired, and craft breakdowns if required, constitute the input. Then a step-by-step procedure is followed to move the floaters so that the resultant force curve approximates the desired force curve. If the results are unsatisfactory, the procedure would be to begin again with a schedule of longer duration.

The detailed method is too long for presentation here. In its present form, it is too involved for manual use except on very small projects. The logical steps are not too difficult, but for even modest-size projects the amount of storage required and "keeping track of" program steps dictates a fairly large computer for economical processing.

## 5. AN ACCOUNTING BASIS FOR PROJECT WORK

From the very start of the development of the Critical-Path Method, it has been the practice to assign a cost account number or job work order number to every job in a project. With this data, a structure can be set up for accruing costs against the proper accounts as the project proceeds.

Because each job in a project has a cost curve associated with it, as duration times are computed, it is a simple matter to compute the estimated individual job cost for a schedule. This computation gives management and supervision the basis for project cost control. As actual costs are accrued, they can be compared with estimated costs and analyzed for exceptions. Time and cost control are inherent in the system.

One of the difficult tasks on certain types of project work is closing the project to capital investment accounts. This frequently is not completed until long after the project ends. There are several reasons for the delay. One is that costs are sometimes not accrued so that they may easily be identified and/or apportioned to the proper facility. Another is the sheer magnitude of the accounting job. Under the Critical-Path system, it is possible to do this job as you go, keeping current with the project. Just as it is easy to close a project, it is easy to estimate in advance capital expenditures for labor, equipment and materials. This can mean many dollars in

savings to project management in efficient capital usage.

## PART II: HISTORICAL DEVELOPMENT AND RESULTS

### 1. EARLY DEVELOPMENTS

The fundamentals of the system outlined in Part I were developed during early 1957. Preliminary results were reported in [4] and [5]. By May 1957 the theory had advanced to the point where it was felt that the approach would be successful. At that time a cooperative effort to implement the method was undertaken by Remington Rand and duPont in order to determine the extent to which any further work was advisable. Remington Rand supplied the required programs for duPont's UNIVAC I located in Newark, Delaware. Engineers from duPont provided a small pilot problem with which to make the preliminary tests.

The results of this phase of the development were officially demonstrated in September, 1957. The demonstration showed that the technique held great promise. Accordingly, further tests of the system were authorized. These tests were set up to determine several things, among which were the following major points:

1) To see if the data required were available and, if not, how difficult they would be to obtain
2) To see if an impartial group of engineers could be trained to use the new method
3) To see if the output from the new scheduling system was competitive in accuracy and utility with the traditional method
4) To determine what kind of computing equipment is required for this type of application
5) To see if the new system was economical.

## 2. SELECTING A TEAM

By late December 1957 a team of six engineers was formed, and work on the test was under way. The team consisted of a field superintendent, a division engineer, and two area engineers, all with experience from construction, a process engineer from design, and an estimator. It is important to note that all these men had some experience in each of the other's specialty. For this reason they had very little difficulty in communicating with one another. Further, they averaged from 8 to 10 years' experience in the duPont organization. Knowing the organization helped expedite their work as a team by making it possible to avoid unnecessary red tape in acquiring the necessary data.

The objectives of the team were to collect the data required for the test project and then plan and schedule it, using the then available UNIVAC I system. In order to prepare the way, the team was given a 40-hour workshop course on the Critical-Path Method. This course covered the philosophy of the method, project diagramming, and interpretation of results. Some attempt was made to indicate how the computer determines minimum cost schedules, but purely for the sake of background. None of the mathematics involved was discussed. The team then spent about a week preparing and processing a small artificial project to test how well they absorbed the material of the course. It was subsequently discovered that as little as 12 hours of instruction are sufficient to transmit a working knowledge of project diagramming to operating personnel.

## 3. THE FIRST LIVE TEST

The project selected for the first test was the construction of a new chemical plant facility capitalized at $10,000,000. We will refer to this project as Project A. In order to get the most out of the test, and because the method was essentially untried, it was decided that the team's scheduling would be carried out independently of the normal scheduling group. Further,

the team's schedules would not be used in the administration of the project.

The plan of Project A was restricted in scope to include only the construction steps. More specifically, the project was analyzed starting just after Part II authorization -- the point at which about 30% of the project design is complete and funds have been authorized to start construction. This approach was reasonable for the first test because the sequence of construction steps was more apparent than those of design and procurement. The latter were to be included in the analysis of some subsequent project.

As the team proceeded to prepare the plan for the project, the following kinds of data were collected and reviewed:

1) Construction cost estimates
2) File prints and specifications
3) Scopes of work and correspondence
4) Bids and quotations
5) Material and equipment list and limiting equipment list with estimated deliveries
6) Design schedule
7) Craft and average wage rates and unit price data
8) Details of pending contracts involving field labor
9) Contemplated design changes with cost and time estimates

The whole project was then divided into major areas. The scope of work in each area was analyzed and broken down into individual work blocks or jobs. These jobs were diagrammed. The various area diagrams were combined to show all the job sequences involved in the project. The jobs varied in size from $50 to $50,000, depending on the available details and the requirements imposed by design and delivery restraints. All told, the project consisted of 393 jobs with an average cost of $4,000; 156 design

and delivery restraints; and 297 "dummy" jobs to sequence work properly, identify temporal check points, and help to interpret results.

During the diagramming phase, normal and crash times and their costs were compiled for each job. In order to develop the normal time it was necessary to use the judgment and experience of the team members in determining the size crew that would normally be assigned to each type of work using generally accepted methods. The associated normal cost was obtained from construction cost estimates.

As only a 40-hour week was authorized for the project, the crash times were obtained by considering only the maximum reasonable increase in manpower for each job and its effect on elapsed time. Additional costs were found necessary because of the extra congestion and activity on a job as crew size increased. Therefore the crash cost was obtained by adding the extra labor costs to the normal cost with an allowance for labor congestion. A straight line was then fitted to this data to obtain the job cost function described by equation (6).

As the plan for Project A took shape, it became clear that we had grossly underestimated the ability of the team. They went into far more detail than expected. This first application made it impractical to continue with the existing computer programs. Fortunately, Remington Rand had previously agreed to reprogram the system for a much larger computer -- 1103A. This programming was expedited to handle the test application.

## 4. SOME RESULTS OF THE PROJECT A TEST

By March of 1958, the first part of the Project A test was complete. At that time it was decided that most of the work on Project A that was being subcontracted would be done by duPont. This change in outlook, plus design changes, caused about a 40% change in the plan of the project. Authorization was given to modify the plan and recompute the schedules. The updating which took place during April, required only about 10% of the time it took to set up the original plan and schedule. This demonstrated our ability to stay

"on top" of a project during the course of its execution.

Several other indicative results accrued from the Project A computations. With only 30% design information, we predicted the total manpower force curve with high correlation. The normal scheduling group had it building up at a rate too fast for the facility to handle in the initial stages of the project. (The reason for this is that they were unable to take available working space into account.) It was not until the project was under way that the error was caught, and they started cutting back the force to correspond with actual needs.

Early in the planning stages the normal scheduling group determined critical deliveries. The team ignored this information and included all deliveries in the analysis. There were 156 items in total. From the computed results it was determined that there would be only seven critical deliveries, and of these, three were not included in the list prepared by the normal scheduling group.

As estimated by traditional means, the authorized duration of Project A was put at N months. The computer results indicated that two months could be gained at no additional cost. Further, for only a 1% increase in the variable direct cost of the project an additional two months improvement could be gained. The intuitive tendency is to dismiss these results as ridiculous. However, if the project manager were asked for a four-month improvement in the project duration and he had no knowledge of the project cost curve, he would first vigorously protest that he could not do it. If pressed, he would probably quote a cost penalty many multiples of the current estimate and then embark on an "across-the-board" crash program. As a point of fact, the reason for the large improvement in time at such a small cost penalty was because only a very few jobs were critical -- about 10% -- and only these needed expediting. The difference in time of two months from N to N-2 can be explained as the possible error of gross time estimates and/or the buffering used in them.

## 5. THE SECOND TEST CASE

With the successful completion of the Project A test, additional projects were authorized. Now the planning was to be done much earlier in the project life and was to incorporate more of the functions of engineering-design and procurement. Project B, capitalized at $2,000,000, was selected for this purpose. By July 1958, this second life test was completed and was as successful as the first. Unfortunately, the recession last year shelved the project so that it could not be followed through to completion.

Experience gained up to this point indicated that even greater capacity than the 1103A provided was essential. In consequence, programs were prepared for the 1105.

## 6. APPLICATIONS TO MAINTENANCE WORK

In the meantime, it was felt desirable to describe a project of much shorter duration so that the system could be observed during the course of the whole project. In this way improvements in the system design could be expedited. An ideal application for this purpose is in the shutdown and overhaul operation on an industrial plant. The overall time span of a shutdown is several days, as opposed to the several year span encountered in projects such as Project A.

The problems of scheduling maintenance work in chemical plants are somewhat different from those of scheduling construction projects. From time to time units like the blending, distillation and service units must be overhauled in order to prevent a complete breakdown of the facility and to maintain fairly level production patterns. This is particularly difficult to do when the plant operates at near peak capacity, for then it is not possible to plan overhauls so that they occur out of phase with the product demand. In such cases it is desirable to maximize return on investment. Because the variable costs usually are small in comparison to the down-time production losses, maximizing return on investment is equivalent to making the shutdown as short as possible.

For purposes of testing the Critical-Path Method in this kind of environment, a plant shutdown and overhaul was selected at duPont's Louisville Works. At Louisville they produce an intermediate in the neoprene process. This is a self-detonating material, so during production little or no maintenance is possible. Thus, all maintenance must be done during down-time periods. There are many of these shutdowns a year for the various producing units.

Several methods and standards people from Louisville were trained in the technique, and put it to the test. One of the basic difficulties encountered was in defining the plan of a shutdown. It was felt, for example, that because one never knew precisely what would have to be done to a reactor until it was actually opened up, it would be almost impossible to plan the work in advance. The truth of the matter is that the majority of jobs that can occur on a shutdown must be done every time a shutdown occurs. Further, there is another category that occurs with 100% assurance for each particular shutdown -- scheduled design and improvement work. Most of the remaining jobs that can occur, arise with 90% or better assurance on any particular shutdown. These jobs can be handled with relative ease.

The problem was how to handle the unanticipated work on a shutdown. This was accomplished in the following way:

It is possible in most operating production units to describe, in advance, typical shutdown situations. Prior to the start of a given shutdown, a pre-computed schedule most applicable to the current situation is abstracted from a library of typical schedules. This schedule is used for the shutdown. An analysis of these typical situations proved sufficient because it was possible to absorb unanticipated work in the slack provided by the floaters. This is not surprising since it has been observed that only 10% of the jobs in a shutdown are critical.

However, if more unanticipated work crops up than can be handled by the schedule initially selected, then a different schedule is selected from the library. Usually less than 12 typical schedules are required for the library.

Costs for these schedules were ignored since they would be insig-

nificant with respect to production losses. However, normal and crash times were developed for various levels of labor performance. The approach here is to "crash" only those jobs whose improved labor performance would improve the entire shutdown performance. The important consideration was to select minimum time schedules. Information on elapsed times for jobs was not immediately available but had to be collected from foremen, works engineering staff members, etc.

By March 1959, this test was completed. This particular application is reported in [1]. By switching to the Critical-Path Method, Louisville has been able to cut the average shutdown time from an average of 125 hours to 93 hours, mainly from the better analysis provided. Expediting and improving labor performance on critical jobs will cut shutdown time to 78 hours -- a total time reduction of 47 hours.

The Louisville test proved so successful that the technique is now being used as a regular part of their maintenance planning and scheduling procedure on this and other plant work. It is now being introduced to maintenance organizations throughout duPont. By itself, the Louisville application has the potential of paying for the whole development of the Critical-Path Method and of earning an equal amount during its first year of use.

## 7. CURRENT PLANS

Improvements have been made continually to the system so that today it hardly resembles the September, 1957, system. Further improvements are anticipated as more and more projects are tackled. Current plans include planning and scheduling a multi-million dollar new plant construction project. This application involves about 1800 events and between 2200 and 2500 jobs. As these requirements outstrip the capacity of the present computer programs, some aggregation of jobs was required which reduced the size to 920 events and 1600 jobs. This project includes all design, procurement and construction steps, starting with Part I authorization. (Part I is the point at which funds are authorized to proceed with sufficient design to develop a firm construction cost estimate and request Part II authorization.)

Also included in current plans are a four-plant remodernization program, several shutdown and overhaul jobs, and applications in overall product planning.

## 8. COMPUTATIONAL EXPERIENCE

The Critical-Path Method has been programmed for the UNIVAC I, 1103A, and 1105 with a Census Bureau configuration. These programs were prepared so that either UNIVAC I or the 1100 series computers may be used independently or in conjunction with one another.

The limitations on the size problems that the available computer programs can handle are as follows: UNIVAC I -- 739 jobs, 239 events; 1103A -- 1023 jobs, 512 events; 1105 -- 3000 jobs, 1000 events.

In actual practice input editing has been done on duPont's UNIVAC I in Newark, Delaware and computation and partial editing on 1100 series machines at Palo Alto, St. Paul, and Dayton. Final editing has then been done at Delaware. System compatibility with magnetic tapes has been very good. In one major updating run, input, output and program tapes were shipped by air freight between Palo Alto and Delaware.

Generally computer usage represents only a small portion of the time it takes to carry through an application. Experience thus far shows that, depending on the nature of the project and the information available, it may take from a day to six weeks to carry a project analysis through from start to finish. At this point it is difficult to generalize. Computer time has run from one to 12 hours, depending on the application and the number of runs required. (Seven runs were required to generate the library for the Louisville project.)

Input and output editing has run less than 10% of the cost curve computations. Indeed, the determination of the earliest and latest start and finish times, and total and free float for a project of 3000 jobs and 1000 events takes under 10 minutes on the 1100 series computers. This run includes input editing, computation, and output editing. If a series of these runs are

to be made on the output solutions from the cost curve computation, only from three to four minutes more are required for each additional solution.

Figure 4 indicates typical cost curve computation times. Of the total number of characteristic solutions that this computation produces, no more than 12 ever have been output edited. The reason for this is that many of the characteristic solutions have very small differences in total project duration.

It has been found that fruitful use of parts of the Critical-Path Method do not require extensive computing facilities. The need for the hardware is dictated by economics and depends upon the scope of the application and the amount of computation that is desired.

## 9. A PARALLEL EFFORT

Early in 1958 the Special Projects Office of the Navy's Bureau of Ordnance set up a team to study the prospects for scientifically evaluating progress on large government projects. Among other things the Special Projects Office is charged with the overall management of the Polaris Missile Program which involves planning, evaluating progress and coordinating the efforts of about 3000 contractors and agencies. This includes research, development and testing activities for the materials and components in this submarine-launched missile, submarine and supporting services.

A team staffed by operations researchers from Booz, Allen & Hamilton, Lockheed Missile Systems Division and the Special Projects Office made an analysis of the situation. The results of their analysis represent a significant accomplishment in managing large projects although one may quibble with certain details. As implemented, their system essentially amounts to the following:

1) A project diagram is constructed in a form similar to that treated earlier in this paper.
2) Expected elapsed time durations are assigned to each job in the project. This data is collected by asking several persons involved in and responsible for each job to make estimates of the

following three quantities:

a. The most optimistic duration of the job

b. The most likely duration, and

c. The most pessimistic duration.

3) A probability density function is fitted to this data and approximations to the mean and variance are computed.

4) Expected earliest and latest event times are computed using expected elapsed times for jobs by means of equations (1) and (2) of Part I. Simultaneously variances are combined to form a variance for the earliest and latest time for each event.

5) Now, probabilistic measures are computed for each event, indicating the critical events in the project.

6) Finally, the computed schedule is compared with the actual schedule, and the probabilities that actual events will occur as scheduled are computed.

This system is called PERT (Program Evaluation and Review Technique. The computations involved are done on the NORC Computer, Naval Proving Grounds, Dahlgren, Virginia. More information about PERT may be found in references [2], [11] and [12].

There are some aspects of the PERT system and philosophy to which exception might be taken. Using expected elapsed times for jobs in the computations instead of the complete probability density functions biases all the computed event times in the direction of the project start time. This defect can be remedied by using the calculation indicated by equation (4) of Part I. Further, it is difficult to judge, a priori, the value of the probability statements that come out of PERT: (1) because of the bias introduced; (2) because of the gross approximations that are made; (3) because latest event times are computed by running the project backward from some fixed completion time.

If there is good correlation with experience then these objections are of no concern. At this moment we are in no position to report the actual state of affairs.

Finally, PERT is used to evaluate implemented schedules originally made by some other means, usually contract commitments made by contractors. To be of most value PERT, or for that matter the Critical-Path Method, should be used by the contractor in making the original contract schedule. In this way many of the unrealities of government project work would be sifted out at the start.

## 10. EXTENSIONS OF THE CRITICAL-PATH METHOD

The basic assumption that underlies the Critical-Path Method, as developed thus far, is that adequate resources are available to implement any computed schedule. (In some cases, this assumption can be avoided by inserting certain types of delivery and completion restraints in the project plan. However, in many cases this is an unrealistic assumption.)

Apparently there are two extremes that need to be considered:

1) Available resources are invested in <u>one</u> project.

2) Available resources are shared by <u>many</u> projects.

In the first case experience has shown that there is usually no difficulty in implementing any computed schedule. Any difficulty that does arise seems to be easily resolved. The Critical-Path Method applies very well in this case. It may be called <u>intra</u>-project scheduling.

In the second case, however, we run into difficulties in trying to share men and equipment among several projects which are running concurrently. We must now do inter-project scheduling.

The fundamental problem involved here is to find some way to define an objective for all projects which takes the many independent and combinatorial restraints involved into account: priorities, leveling manpower by crafts, shop capacity, material and equipment deliveries, etc. For any reasonable objective, it also is required to develop techniques for handling the problem. Preliminary study has indicated that this is a very difficult

area of analysis and requires considerable research. However, it is felt that the Critical-Path Method as it stands can form a basis for systems and procedures and for the requisition of data for this extension of scheduling.

It would be of some interest to extend the method to the case where job durations and costs are random variables with known probability density functions. The mathematics involved appears to be fairly difficult. Due to the problems of obtaining data in this form, such an extension may be purely academic for several years to come.

## 11. OTHER APPLICATIONS

The potential applications of the Critical-Path Method appear to be many and varied. Consider the underlying characteristics of a project -- many series and parallel efforts directed toward a common goal. These characteristics are common to a large variety of human activities. As we have seen, the Critical-Path Method was designed to answer pertinent questions about just this kind of activity.

We have already treated applications of the technique to the construction and maintenance of chemical plant facilities. The obvious extension is to apply it to the construction and maintenance of highways, dams, irrigation systems, railroads, buildings, flood control and hydro-electric systems, etc. Perhaps one of the most fruitful future applications will be in the planning of retooling programs for high volume production plants such as automotive and appliance plants.

We have also seen how it can be used by the government to report and analyze subcontractor performance. Within the various departments of the government, there are a host of applications -- strategic and tactical planning, military base construction, construction and overhaul of ships, missile countdown procedures, mobilization planning, civil defense, etc. Within AEC alone, there are applications to R & D, design and construction of facilities, shutdown, clean-up, and start-up of production units. Another example is in the production use of large equipment for the loading and unloading portion of the production cycle of batch processes. Because each of these operations is of a highly hazardous nature, demanding very close

control and coordination of large numbers of men and/or complex equipment, they appear to be natural applications for the Critical-Path Method.

Common to both government and industry are applications that occur in the assembly, debugging, and full-scale testing of electronic systems.

## References


[1] Astrachan, A., "Better Plans Come From Study of Anatomy Of An Engineering Job," Business Week, March 21, 1959, pp. 60-66.

[2] Fazar, Willard, "Progress Reporting in the Special Projects Office," Navy Management Review, April 1959, pp. 9-15.

[3] Ford, L. R., Jr. and D. R. Fulkerson, "A Simple Algorithm for Finding Maximal Network Flows and an Application to the Hitchcock Problem," Canadian Journal of Math., Vol. 9, 1957, pp. 210-218.

[4] Kelley, J. E., Jr., "Computers and Operations Research in Road-building," Operations Research, Computers and Management Decisions, Symposium Proceedings, Case Institute of Technology, Jan. 31, Feb. 1, 2, 1957.

[5] _____, "The Construction Scheduling Problem (A Progress Report)" UNIVAC Applications Research Center, Remington Rand UNIVAC, Philadelphia, April 25, 1957. (Ditto)

[6] _____, "Parametric Programming and The Primal-Dual Algorithm," Operations Research, Vol. 7, No. 3, 1959, pp. 327-334.

[7] _____, "Critical-Path Planning and Scheduling: Mathematical Basis," in preparation.

[8] _____, "Extension of the Construction Scheduling Problem: A Computational Algorithm," UNIVAC Applications Research Center, Remington Rand UNIVAC, Philadelphia, Nov. 18, 1958. (Ditto)

[9] _____, and M. R. Walker, "Critical-Path Planning and Scheduling: An Introduction," Mauchly Associates, Inc., Ambler, Pa., 1959.

[10] Martino, R. L., "New Way to Analyze and Plan Operations and Projects Will Save You Time and Cash," Oil/Gas World, September, 1959, pp. 38-46.

[11] PERT, Program Evaluation Research Task, Phase I Summary Report, Special Projects Office, Bureau of Ordnance, Dept. of the Navy, Washington, July 1958.

[12] Malcolm, D. G., J. H. Roseboom, C. E. Clark and W. Fazar, "Application of a Technique for Research and Development Program Evaluation," Operations Research, Vol. 7, 1959, pp. 646-669.

# TYPICAL PROJECT DIAGRAM

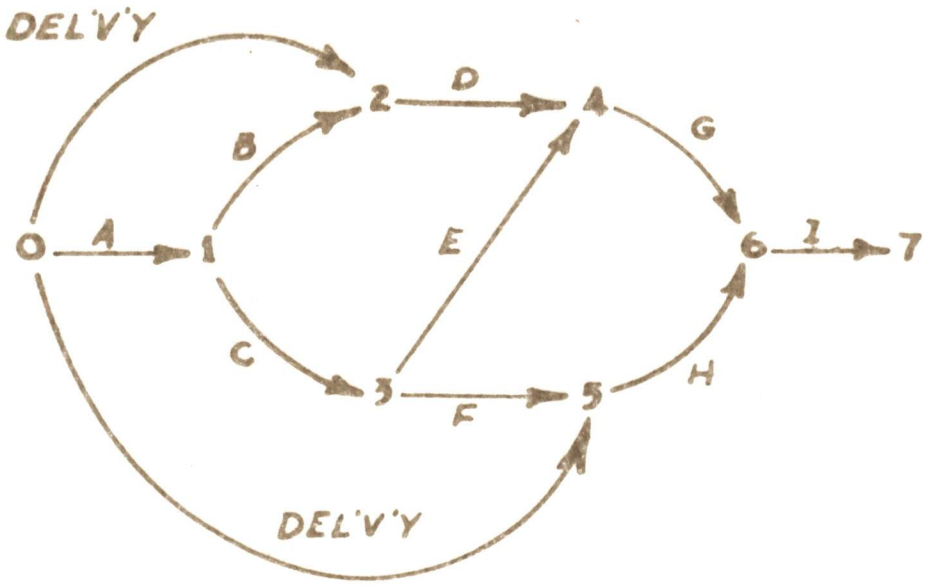



FIGURE 1.

# TYPICAL JOB COST CURVE

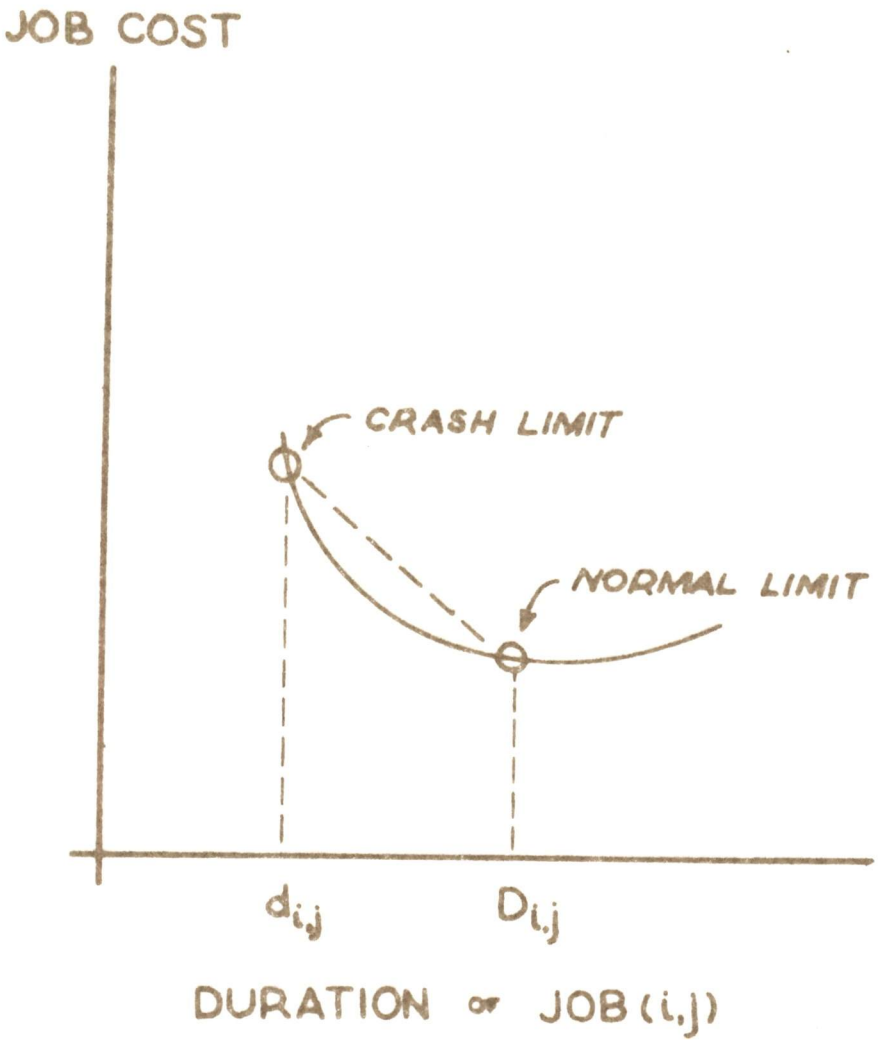



FIGURE 2.

# TYPICAL PROJECT COST CURVE

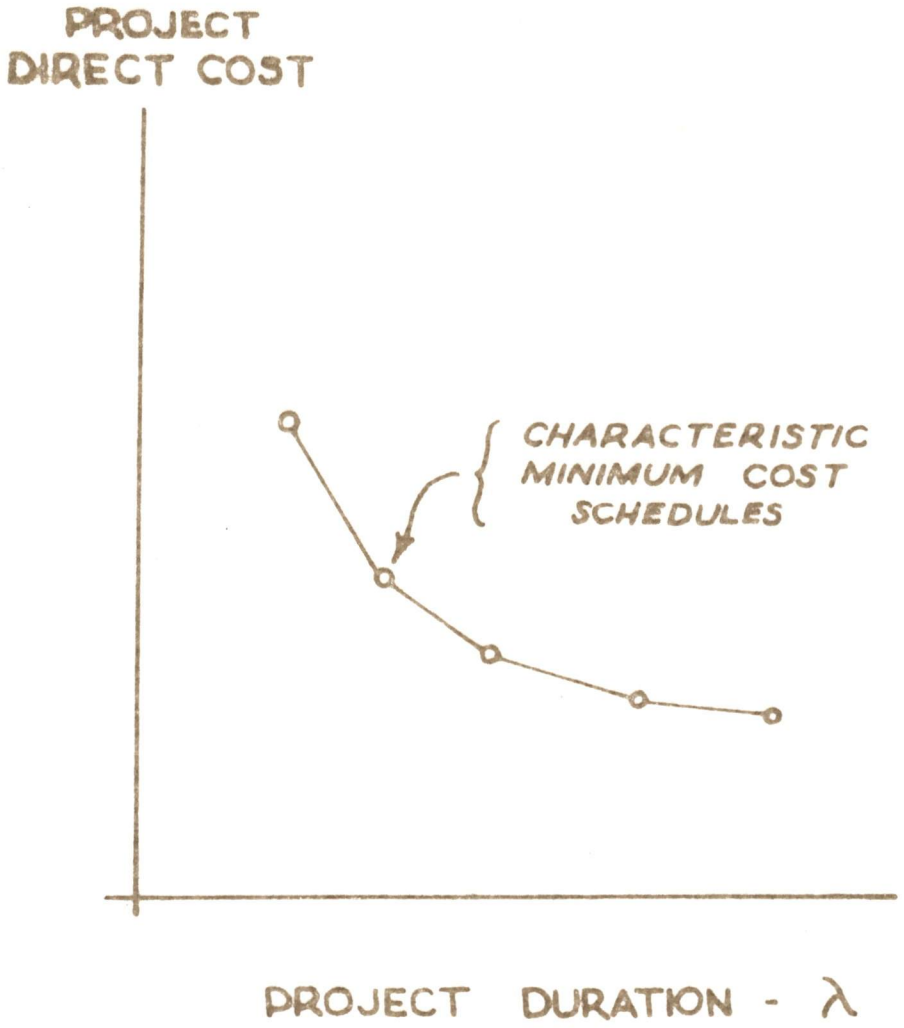



FIGURE 3.

# TYPICAL RUN TIMES

| EVENTS | JOBS | SOL'N'S | MINUTES UNIVAC I | MINUTES 1103A/1105 |
|---|---|---|---|---|
| 16 | 26 | 7 | 8 | 1 |
| 55 | 115 | 14 | 125 | 3 |
| 385 | 846 | 21 | - | 100 |
| 437 | 752 | 17 | - | 24 |
| 441 | 721 | 50 | - | 49 |
| 920 | 1600 | 40 | - | 210 |

FIGURE 4.

Biographical Sketch

Dr. Richard B. Lawrance

S. B. Electrical communications M.I.T. 1940 .... During World War II, Research Associate M.I.T. Radiation Laboratory, Loren development .... Graduate student in physics, M.I.T., received Ph. D. 1949 .... For five years Director Applied Physics Department, National Research Corporation .... Member American Physical Society, Acoustical Society of America, Institute of Radio Engineers .... radio amateur since 1936 .... inventor or co-inventor in a dozen issued patents .... author of several technical papers and a handbook section .... At DATAmatic he is a Staff Engineer and Department Manager, in charge of magnetic tape mechanism development.

ABSTRACT

An Advanced Magnetic Tape System for Data Processing

Dr. Richard B. Lawrance*

We describe a new magnetic tape mechanism, recording system, and information checking and restoration system of high reliability. In the mechanism, comparison is made between pinch-roller and vacuum capstan means for tape motion control. Criteria include minimization of tape deterioration both gradual and catastrophic, tracking, skew, and maintenance considerations.

Systems techniques for enhancing tape system reliability are discussed briefly, with some emphasis on the use of error detection and automatic correction. The choice of information format on the tape and of error correcting parameters for maximum effectiveness is described.

* DATAmatic Division

Minneapolis-Honeywell Regulator Co.

151 Needham Street

Newton Highlands 61, Massachusetts

Eastern Joint Computer Conference

# An Advanced Magnetic Tape System For Data Processing

## I Introduction and Tape Mechanism Considerations.

It is a truism that for any but the smallest digital data processing systems major attention must be given to the provision of an adequate magnetic tape transport, reading and writing system, and means for insuring the correctness of information all the way from the central processor to the magnetic tape and back again. This paper will describe some of the above mentioned features of the Honeywell 800[1].

Early in the layout and specification of a new system it is necessary to decide on the specifications for and approach to the magnetic tape mechanism and recording system. As regards the tape mechanism itself, our earlier experience (particularly with the DATAmatic 1000) had favorably inclined us toward the vacuum capstan approach. Our several years of experience with electrostatic clutching had led us ultimately to abandon the electrostatic approach for the DATAmatic 1000, and after re-evaluation, it was again excluded from consideration for the new system. As to the other two widely-used methods of achieving fast stop-start tape motion, we felt that the faster and more positive of these -- namely the pinch-roller approach -- should be the most seriously considered as an alternative to the use of vacuum capstans.

---

[1] Together with the paper "Control & Arithmetic Techniques in a Multi-Programmed Computer." By: N.Lourie, H. Schrimpf, R. Reach, W. Kahn. Presented at this conference, the present paper forms a partial technical description of this new data processing system.

In this comparative evaluation and in the design effort which followed, we placed an overriding importance on providing the magnetic tape itself with a benign environment. This is in accord with our belief that in every stage of manual handling or manipulation by the mechanism all stresses in the tape (both during normal operation and under failure conditions) should be made zero by design or kept to demonstrably safe values.

The following tabulation compares inherent features of presently used pinch-roller mechanisms with the corresponding features of pneumatic mechanisms.

| Pinch Roller | Vacuum Capstan |
|---|---|
| 1. In some designs (but not those in which the idler is continuously driven) the tape to be accelerated must bear not only the forces to accelerate its own mass but also the forces to give angular acceleration to the idler. | 1. Only the mass of the tape itself requires acceleration, thus minimizing the force transmitted to and by the tape. |
| 2. The tape accelerating forces are applied in a concentrated area surrounding the line of tangency of two typically rather small cylinders (slightly spread by resiliency in the tape itself and at most one of the cylinders). | 2. The tape accelerating forces are distributed over a typically fifteen fold larger area, whose length may equal or exceed one-fourth of the capstan circumference. The capstan diameter may be conveniently large. |
| 3. In order to prevent non-simultaneous clutching across the width of the tape (with attendant tracking and skew problems) a very accurate pivoting or translatory motion is required; fast operation demands that this be designed for minimum inertia. Thickness variation across the tape is a possible source of skew. | 3. Symmetrical engagement of the tape to capstan or brake is automatically achieved by symmetrical design of pneumatic passages. Engagement always commences along tape center line minimizing skew. Transverse variations in tape thickness does not add to skew. |
| 4. Compressive action of pinch-roller tends to emboss wear particles or other dirt into the oxide surface of the tape. (Not applicable to metal tape). | 4. Free from dirt embossing. No material body need touch the oxide surface of the tape (although usually the magnetic head is made to do so). |

| | |
|---|---|
| 5. Powerful fast-pickup driving and braking mechanisms may be slow to release, placing safety restriction on minimum interval between drive and brake commands. | 5. No restriction on interval between successive commands. Moving parts of mechanism are offset from tape path, completely covered, and cannot touch tape. No danger to tape from tug-of-war. |
| 6. For high performance, auxiliary air lubrication may be required.[2] | 6. Air lubrication of tape is a built-in feature. |
| 7. Usual embodiment employs rollers of flanged-spool construction, with tape unsupported and not otherwise edge-guided over important portions of its path. | 7. Essentially complete edge guiding over entire path from supply reel to takeup reel is easily incorporated. |

The considerations tabulated above led us to design for the Honeywell 800 a tape mechanism utilizing the vacuum capstan principle and embodying many of the techniques and principles used in the earlier DATAmatic 1000 three-inch tape mechanism.[3]

[2] R. A. Skov "Pulse Time Displacement in High-Density Magnetic Tape" IBM Journal of Research and Development, April 1958.

[3] R. B. Lawrance, R. E. Wilkins, R. A. Pendleton, "Apparatus for Magnetic Storage on Three-inch Wide Tapes". Proceedings of the Eastern Joint Computer Conference, 1956. Special publication T-92.

## II Brief Description of Tape Mechanism

Figure 1 shows a photograph of the Type 804 magnetic tape mechanism. The unit stands approximately 5 feet 9 inches high and occupies a floor area slightly over 2 feet square. The tape is a nominal 3/4 inch wide and is moved at a normal speed of 120 inches per second in either direction as desired. High speed rewind is provided, in one direction only, at 360 inches per second.

The cabinet shown includes the separate write-amplifier final stages for the ten recording channels as well as the final stage for the AC-excited separate erase gap; the three-stage transistor preamplifiers for each of the ten playback channels; and the solid-state switching equipment for placing the read-write head and circuits in the selected mode. Also included are the power supplies, loop position sensors and servo control for the DC-operated reel motors; the power supplies for the read-write circuits; vacuum and pressure sources for the capstan, brake, and suction loop chambers; beginning-of-tape and end-of-tape sensing means; storage; and other electronic packages facilitating testing and maintenance.

Figure 2 shows a closeup of the capstan area as it appears when tape is in position for information transfer. The centrally-located three-inch diameter billet contains the magnetic head assembly, and the oxide surface of the tape is uppermost. The tape lies horizontally, immediately over the two-piece horizontal vacuum brake, and thence executes a 90-degree downward turn at each vacuum capstan before dropping directly to the pneumatic loop chambers. Each capstan, when not actively engaged in driving tape, is provided with continuous air lubrication of approximately 2 psig, which effectively prevents all contact between the capstan and the tape. Unbroken edge guiding is present in the vicinity of the capstans, brake, and head, and indeed is present all the way from one reel to the other except in a space of 1 3/4 inches

immediately next to each reel. Even in these regions back edge guiding is present, the deliberate absence of front edge guiding being in the interest of eliminating possible finger-catching accidents. It has been our experience with magnetic tapes (as with other elongated flexible substrates) that continuous edge guiding is far more advantageous than guiding by periodically-spaced flanged spools, provided only that the tape be slit accurately enough. For nearly two years we have made complete and detailed observations of commercially produced magnetic tape with respect to width and the periodic curvature usually called snakiness. We can state that the snakiness can reasonably be reduced to complete insignificance while the width of slitting is held well within a total range of .002 inch.

We feel that these edge-guiding arrangements, together with accurate tape width control, yield considerable benefit in drastically reducing tracking and skew errors within the mechanism, as well as contributing to long tape life since the edges of the tape are nowhere subjected to localized sideways forces. Spring-loaded parts for exerting side-thrust on the tape are themselves subject to excessive wear, so their elimination enhances reliability.

Returning to Figure 2 and comparing it with Figure 1, we note that in normal operation, with the tape in the loop chambers and the head in position, the oxide surface is in rubbing or pressure contact with no parts of the mechanism except the magnetic head, bringing tape wear to a practical minimum. Figure 3 shows the capstan and head area with the head eccentrically rotated, removing the head from contact with the tape oxide surface. Thus in high speed rewind (at whose beginning the head rotates away automatically) not even the head touches the oxide. Rotation of the head, automatically controlled, is also used during tape changing, at which time it enables the tape to slip easily over what is otherwise an unbroken edge guide.

In Figure 3 the magnetic portion of the tape is all on the left hand supply reel and the leader of heavier-gauge clear Mylar® (permanently attached to the tape) is lying over the capstan and brake. This enables the nature of the exterior pneumatic passages of the capstan and brake to be seen. The two capstans are continuously rotated in opposite directions by individual 1200 rpm hysteresis synchronous motors, the capstan circumference being exactly 6 inches. The left and right portions of the brake (lying between the normal head location and the two capstans) are internally connected to a common working air passage which is supplied appropriately with medium suction, strong suction, or air at atmospheric pressure.

Figure 3 also shows that no pressure pad is employed to keep the tape in contact with the magnetic head assembly. Wrapping contact between tape and head is adequately maintained by having the head press the tape down into a short and very shallow "V", the outer edges being defined by the rounded shoulders of the brake, closely adjacent. By means of this wrap, with its elimination of pressure pads, and by means of the unconventionally large radius of curvature of the magnetic head (both essentially the same in dimensions as in the DATAmatic 1000) we achieve good transient and running contact between tape and head, together with a gratifyingly low rate of head wear. Measurements carried on over more than a year's two-shift operation of a DATAmatic 1000 show for all channels of all magnetic heads a quite uniform and unexpectedly low rate of wear. The average yearly loss of material from the head under these conditions amounted to 0.0001 inch.

By implication, the tape wear produced by friction between head and tape is correspondingly small.

Rewind and Tape Change

The central processor instructions to which the tape drive responds are Write (forward), Read Forward, Read Reverse, and Rewind. The Tape Change operation is initiated by manipulating a lever switch on the tape mechanism itself.

The position shown in Figure 3 occurs at the termination of a tape change operation, which starts with a high speed rewind unless the tape is already rewound. Every rewind command is executed by the mechanism as a high speed rewind, and once

Continued on page 7

received from the tape control unit the rewind is performed under local control until completed. During high speed rewind the tape speed is controlled by the left-hand vacuum capstan, whose motor speed is increased automatically to 3600 rpm. The tape remains in both vacuum loop chambers and accordingly receives the benefit of controlled tension and complete edge guiding. Upon rewinding past the designated beginning of tape, as sensed by a photoelectric arrangement described later, the mechanism shifts down from 360 ips to 120 ips. This latter speed endures for a fraction of a second and the tape is then stopped in normal fashion by pneumatic disengagement from the capstan and engagement to the brake. The head, which has been automatically moved out of contact with the tape during the rewind, now rotates back into contact with the tape, and the closing of a Microswitch® signals the computer that the rewind has been completed and that the tape mechanism is again ready for instructions. At this time the magnetic head is positioned part way down the clear leader and has access to the first magnetic information location by moving the tape in the forward direction (to the right).

If it is desired to change tape, a centrally located manual switch on the control panel is thrown to the tape change position. It is irrelevant whether the tape is already rewound or not, although some head rotating operations are bypassed if the rewind is continuous with the tape change. The tape proceeds to the left, along the clear leader and at 120 ips, until a single short centrally placed slot in the leader is sensed by an orifice and vacuum switch associated with the upper end of the right hand loop chamber. ~~This orifice can be seen in Figure 3.~~ When sensed this causes the tape to stop with only two or three turns remaining to be manually unwound from the right-hand reel, and the appropriate partial shutdown of the mechanism is initiated so that the reel is ready for removal.

Time taken for a rewind operation can be characterized by the equations

$$t_{rewind} \leq \frac{t_{rewind}}{3} + 2.6$$

$$\text{or} \quad t_{rewind} \leq \frac{\text{distance in ft.}}{30} + 2.6$$

in which all times are given in seconds.

Tape Reels and Mounting

Manufacturers of Data processing magnetic tape mechanisms do not as a rule use standard reels, and the present equipment is no exception. Since a partial vacuum (about one half atmosphere absolute) is provided within the equipment for use in the clutch, it is quite natural to use this vacuum for holding the reels onto the reel mounts. This technique has already proved highly satisfactory with the three-inch-wide 23 pound reels of the DATAmatic 1000. Advantageous features include the lack of metal-to-metal contact between reel and reel mount, the fact that the reel hub is not subjected to hoop stress, and the provisions of a large flat reference surface on the reel mount, which insures wobble-free rotation and accurate positioning of the reel relative to the back reference surface.

Suction is similarly used for attaching the free end of the tape leader to the right hand reel whenever a tape is loaded on the machine, as well as for initially attaching the inner end of the tape to the left-hand supply reel. It is worth mentioning that vacuum attachment of the tape to the reel makes it unnecessary to perforate the reel flanges for finger access to the hub during loading. The unperforated flanges are helpful in protecting tape from dirt and mechanical damage while in storage or during handling. Hazard to the operator is also reduced materially.

The design of the reel and reel mount involved additional factors, however. It was desired to make use of a demountable ring, capable of being stored with the reel of tape and serving by its presence or absence to enable or inhibit the recording of information on the tape. (This is in addition to a manual switch on the operator's panel.) Various embodiments of this principle have been used for several years by other manufacturers, but we believe our version has some useful and novel advantages. One desirable feature present in our arrangement is that the physical presence or absence of the write-enable ring does not need to be inferred from the status of a concealed electrical switch (which requires that electrical power be applied and that the circuit be functioning with some means of indicating its status). We have placed

the write-enable ring in plain view on the front of the reel. It thus becomes easy to remove or insert the ring while the reel is in place on the mechanism, without the necessity for first rewinding the tape in order to remove the reel.

Figure 4 shows a photograph of three reel mounts, one having a write-enable reel of tape mounted on it and another carrying a write-inhibit reel. The removable snap ring which converts a write-inhibit reel to a write-enable reel is shown beside the third reel mount. The principal working part of the reel mount subassembly is the central bell-shaped cylinder. Its axial motion controls three retractable nylon latches, spring-loaded radially outward, and also an internal piston, spring-loaded axially outward. To remove a tape reel from the mount the reel flanges are lightly grasped by the fingers, while the thumbs press the central cylinder so that it moves axially inward. The three nylon latches are thus moved radially inward to the point where the reel can slide over them and be removed. In putting a reel on the mechanism, the central cylindrical bore of the reel performs a similar operation in reverse -- as the reel is moved inward it presses on the nylon latches, retracting them. A small fraction of an inch before the reel is fully seated the latches snap outward and thus hold the reel in place even with no power or no vacuum. When vacuum is applied (automatically, as part of the normal cycle-up procedure) the reel is drawn into intimate sealing engagement with the rubber driving rings and is fully positioned ready for operation.

The write-enable ring operates by capturing the outer rim of the central cylinder, as the reel is pressed on. By this means the central cylinder is moved inward about 1/4 inch as the reel is seated home. The internal piston-and-cylinder arrangement is thereby vented to atmospheric pressure rather than being connected to the half-atmosphere suction reservoir. The elctrical image of these two pressure states is created in a stationary vacuum-diaphragm-operated Microswitch® located at the rear of the main mounting plate and sampling the pressure in the reel mount cylinder via a carbon rotary seal.

The Vacuum Clutch

Figures 2 and 3 showed portions of the vacuum capstan and brake, and their relationship to the magnetic head as mounted in the mechanism. Figure 5 shows an exploded view of these components of the pneumatic clutch, viewed from the side rear. Of the components all but the capstan motor are mounted to the front of the heavy flat vertical plate which serves as structural support and back edge guide, and which is omitted from the photograph. The capstans, of which only one is shown in exploded position, are directly mounted to the shafts of their respective hysteresis synchronous motors. Precision bearings are used in the motors, and capstan runout and taper are held to tight tolerances in order to achieve good tape tracking.

A 90-degree segment of each capstan is connected via the working air passage to the electropneumatic valve, mounted nearby in the actuator housing body. (As shown in Figure 5, this is bolted directly to the capstan housing body.) The fixed portion of each pneumatic commutator consists of a carbon composition cylinder which fits closely without rubbing inside the cuplike capstan. The portion of the working air passage within each carbon piece consists of a single slot centrally spanning the active arc of the capstan, and a drilled hole connecting to the actuator.

The location of this slot along the center line of the tape track, together with the pneumatically symmetrical design of the capstan itself, leads to what we believe is an important advantage for the vacuum clutching technique. Figure 6 shows a series of sketches representing the clutching action and the production of skew in pinch roller and vacuum clutches respectively. Part A shows (greatly exaggerated) the engagement of a tape to a capstan when the moving pinch roller is slightly out-of-line so that distances D1 and D2 are unequal. While we have no quantitative measurements available it is not too difficult to imagine that an inequality of perhaps 0.0001 inch will result in significant time difference in the engagement of the two tape edges, to the capstan. Parallelogram distortion of the tape would then produce skew. Similarly, as shown in Sketch B, it would appear to be possible for skew to be produced even if

the moving pinch roller were to be perfectly aligned with the capstan. Any thickness taper across the width of the tape will produce the same effect as an out-of-line pinch roller. Again we have no quantitative data to support this conjecture but the point-to-point thickness tolerances to which tape backing is produced are large enough so that the possibilities of skew production from this cause should not be overlooked. The unsettling thing is that since tape is not customarily inspected for thickness uniformity it appears possible for portions of an otherwise perfect tape to produce random skew when used with a clutch of the pinch-roller type.

The situation is different with a vacuum capstan, however, as shown by experiment. A priori expectations (verified in detail by observations using a time-delayed stroboscopic flash) are that since the working air passages communicate to the underside of the tape symmetrically about the tape center line it should be the case that the center of the tape always engages first. Thereafter the region of engagement spreads symmetrically to the edges. Prior to the evacuation of the working air passage it and the underside of the tape have been supplied with air lubrication at slightly above atmospheric pressure; thus at the start of a clutching operation the underside of the tape is at a rather definite and reproducible location with respect to the capstan surface. As shown in the fourth sketch of Figure 6 it is thus to be expected that, to first order at least, any variation of tape thickness across the web will not significantly affect the symmetrical tape engagement.

Returning to Figure 5, it can be seen that the valve actuators each contain a small but efficient electromagnet which is energized when its associated capstan is intended to drive tape. The highly effective eddy-current shielding of the aluminum actuator housing body prevents any external magnetic influence on the tape or in the head.

Figure 7 shows a sketch of one of the electropneumatic valves, each of which is essentially a pneumatic SPDT switch. With no current in the coil the flat armature, resiliently pivoted near the right-hand end, will seal off the upper valve seat, being maintained in position by pivot bias and by air pressure differential. During this time the working air passage to the capstan is supplied with lubricating air from the pressure reservoir, at approximately 2 psig. When current is passed through the magnet winding the valve assumes the position of Figure 7, with the armature magnetically drawn down to seal off the compressed air from the working air passage. The capstan and working air passage thus exhaust into the reservoir at half-atmosphere vacuum.

The armature is tapered slightly, as shown, to reduce inertia and speed up pull-in. Life tests on a group of similar armatures and magnets, driven at 120 operations per second for a period of over 20 months showed no measurable change in performance after $6.2 \times 10^9$ operations.

The transistor circuits which drive the actuators supply an initial high-current pulse for fast armature pull-in, dropping to a reduced holding current which lasts until the stop command is received. The ferromagnetic material of the magnet is an alloy with relatively low saturation flux density so that drop-out time is held to a minimum. The transistor circuits are interconnected in such a way that engagement of the tape to both capstans simultaneously is most unlikely, even under failure conditions; even if this should occur, however, the tape suffers no damage since the capstan motors will stall without the tensile elastic limit of the tape having been exceeded.

Typical curves of tape velocity versus time in starting and stopping are shown in Figures 8 and 9. These curves were taken by recording on the tape a train of 64 pulses derived from a 5 kc keyed oscillator, turned on at the time of the stop or start command. Magnetic development with colloidal $Fe_3O_4$ and position measurement with a microscope and traveling micrometer table were then used to give an accurate history of tape position and velocity, relative to the read-write gap, versus time.

Figure 8 shows that in response to a start command the tape commences to move at slightly less than one millisecond; at 2.7 milliseconds the tape has traveled 0.12 inch and is traveling at 120 inches per second. Speed fluctuations thereafter do not exceed approximately 3 or 4 percent, although the read system will tolerate many times this amount. Figure 9 shows that in stopping, the initial deceleration occurs after about 1.2 milliseconds and that the total distance to come to rest is substantially less than 0.3 inches. As mentioned briefly earlier the 804 mechanism allows a start command to follow a stop command arbitrarily closely. The present curves indicate why the tape continues at full speed when the interval between commands does not exceed approximately 0.7 milliseconds; for longer intervals there is a smooth transition to the isolated-stop, isolated-start condition shown in the graphs.

Sensing of Beginning and End of Tape

In the interest of brevity we will not give a complete description of the circuit arrangements for keeping track of the position of the tape in the machine: that is, whether the magnetic head is positioned over the permanently-attached clear leader, initial information space, mid-tape information space, or one of the recognizable end-tape zones. With one exception, the task of remembering tape positions on all of the eight connected tape mechanisms is assigned to their common tape control unit.

The exception is concerned with rewind operations, in which the controlling elements are completely local to the respective tape drives: a relay picks up at the start of the rewind and only releases upon sensing clear leader at the completion of the rewind.

The boundaries of all logically distinct tape regions are marked off by small windows at the front edge of the tape, created by removing oxide for a distance of 0.1 inch along the tape and .035 inches in from the edge. Since the nearest recording channel ends 0.041 inches from the edge there is no conflict between optical sensing and magnetic recording. It is possible to sense the passage of a window without interfering in any way with the execution of any write or read instruction which may be in process. Significant program advantages and time savings result from this feature.

The special illuminator contains a miniature long-life tungsten filament bulb and a one-piece optical element consisting of a lens, cylindrical barrel, and angular refracting surface. This illuminator is positioned at a fixed distance from the magnetic head near the upper end of the right-hand loop chamber, with the optical element extending at an angle through the loop chamber outer wall to a position nearly flush with the inner surface. By this means, since the angle of the refracting surface is nearly the angle for grazing refraction, a satisfactory intense light source is effectively positioned directly opposite the outer edge of the tape, yet without mechanical projection into the path of the tape.

Upon passage of one of the windows, light falls on a miniature silicon photodiode (part of the subassembly) which issues the window-recognition signal for interpretation and storage.

Read-Write System

In the Honeywell 800, as in nearly all other systems, the tapes are written in the forward direction only, i.e. with the tape moving to the right. Reading takes place in either direction as desired, and uses the same head gaps as for writing. Ten channels are used, of which eight are information channels, one is an Orthotronic parity channel, and the tenth is a clock. A separate full-tape-width erase gap, located a fraction of an inch upstream of the read-write gaps, applies AC erase to the tape at the time of recording. The read-write gaps are in-line across the tape and are spaced on 0.070 inch centers.

The AC erase serves the primary function of cleaning out the inter-record gaps and leaving the tape magnetically neutral, which facilitates record-entry recognition in bidirectional readback. NRZ1 recording (saturation-to-saturation, flux change denotes a "one") is used on the information and parity channels. The Honeywell 800 word contains 48 bits (not counting the parity bits which accompany the information on tape and in memory) so that a word occupies six frames on tape, a frame being defined as the time-simultaneous record of a bit in each information channel. The parity bit is also recorded simultaneously with the eight information bits. The frame interval is 21 microseconds, corresponding to a frequency of 47,619 frames per second and a bit density (at 120 inches per second) of 397 per inch.

The clock channel is similarly recorded from saturation to saturation, but undergoes one flux reversal per frame. The recording of the clock is not simultaneous with the recording of the other bits of the frame but is offset by one half of the frame interval. By this means the read circuit is made self-timing, highly tolerant of speed variation in the tape mechanism, and free from one-shot circuits with their jitter and delay tolerance accumulations.

As soon as a write instruction is received the erase head is excited and remains so, independent of tape motion, until receipt of the next instruction of a differant type (read, rewind, tape change). At the beginning of a record to be written, with the tape in motion and the inter-record gap just traversed, write current is initiated in all ten channels in the same standard polarity. This results in half-strength magnetic poles of known polarity being written in all channels, automatically ignored in playback. Thereafter the clock begins its 21-microsecond beat and 10.5 microseconds after the first clock beat the first frame is recorded, with flux reversals in those channels where ones are to be written. Writing continues, at six frames per word, until all words of the record have been recorded. Before cessation of writing two orthotronic words (twelve frames) and an end-of-record word are appended, after which one more clock pulse is written and all write currents drop to zero.

The construction of the orthotronic words is on a per-channel basis, roughly as follows: the first, thirteenth, twenty-fifth . . . bits are half-added and the first bit of the orthotronic word is the complement of their sum. Similarly the second orthotronic bit is formed from the second, fourteenth . . . bits of the record, etc. The result is a very powerful check having the following properties:

1. Garbled information confined to a single channel can be recreated regardless of the length of the difficulty.

2. Garbled information extending up to twelve bits in length can be reconstructed regardless of the number of channels affected.

In playback the ten channels are connected, by means of solid-state switching, to ten individual preamplifiers located at the tape mechanism and thence are passed via the tenfold read bus to the Type 803 Tape Control Unit. Further shaping culminates in peak detection of each signal and the production of a one-half microsecond pulse

essentially coincident with each flux change in the channel. These pulses set nine individual high-speed flip-flops, which accumulate the bits of the frame; the next peak-detected clock pulse (half a period later) resets all flip-flops and sends the bits into buffer storage where they reside until a complete word is available for transmission to memory.

Figure 10 shows the appearance of playback from a single channel, and has the typical NRZ1 waveform. Because of the conservative bit density satisfactory resolution is achieved with a comparatively large head gap, minimizing signal fluctuations due to the passage of lint or other debris between the head and the tape. Figure II shows the effect of a recording dropout deliberately produced by blowing fibers of cotton lint into the region between the magnetic head and the tape being written. The amplitude decrease shown is typical and produces no error in reading, as shown by the associated peak detector waveform. The read system is designed to tolerate signal decrease to well below one-fourth of normal amplitude. It is well to mention, also, that the tape mechanism incorporates the conventional positive pressurization of the region occupied by reels, capstans, head, and loop chamber entrances, thus excluding airborne dust except during necessary tape changing.

We have not dealt at lengths with the internal checking of the Honeywell 800 but it is well to mention, in conclusion, two of these features associated with the magnetic tape system. Writing cannot occur (and its absence is made known) unless an enabling check shows that the erase head and the clock channel are both excited. The transmission of data to the tape drive is checked for transverse parity at each frame and for longitudinal parity on each channel of each record.

The net result of the features described in this paper is a strong, efficient, and trouble-free tape system. The approach deliberately taken has been to design high reliability into all electrical and mechanical components, effecting error detection and correction by means of the powerful capabilities of Orthotronic control.

ACKNOWLEDGEMENTS

The read-write circuit and system development has been done by a group directed by Dr. Way Dong Woo. I am sure that the many other contributors to the design of the tape drive itself will not object to my singling out Mr. Robert A. Pendleton as major contributor.

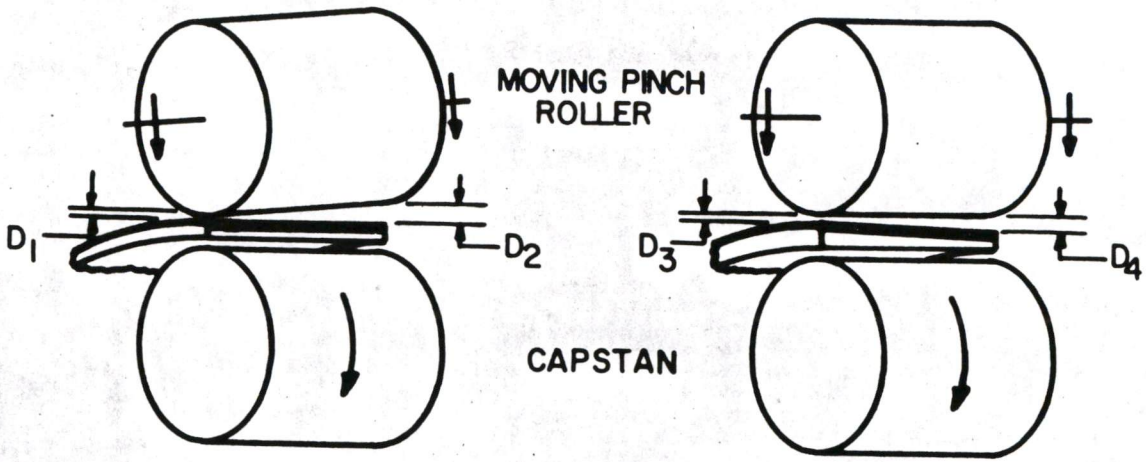
## REFERENCE SHEET

N. Lourie, H. Schrimpf, R. Reach, W. Kahn, "Control & Arithmetic Techniques in a Multi-Programmed Computer." Eastern Joint Computer Conference, 1959.

R. B. Lawrance, R. E. Wilkins, R. A. Pendleton, "Apparatus for Magnetic Storage on Three-Inch Wide Tapes." Proceedings of the Eastern Joint Computer Conference, 1956. Special Publication T-92.

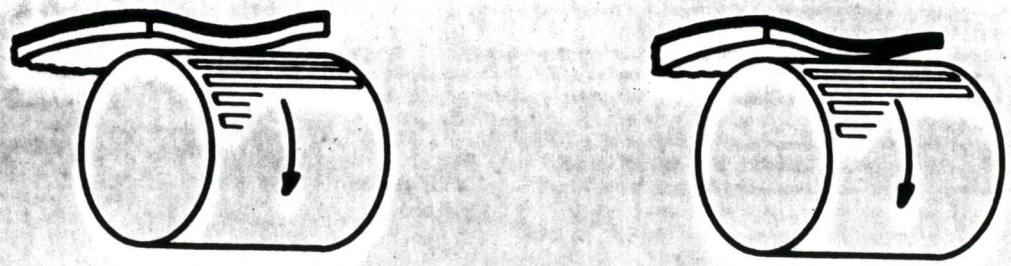
R. A. Skov, "Pulse Time Displacement in High-Density Magnetic Tape" IBM Journal, April 1958.

## LIST OF ILLUSTRATIONS




A) OUT-OF-LINE PINCH ROLLER PRODUCING SKEW.

B) TAPE THICKNESS VARIATION PRODUCING SKEW.



C) INITIAL CLUTCHING OF VACUUM CAPSTAN ALONG CENTERLINE OF TAPE REGARDLESS OF THICKNESS VARIATION.

FIG. 6 SKEW CONSIDERATIONS FOR PINCH ROLLERS AND VACUUM CAPSTANS.

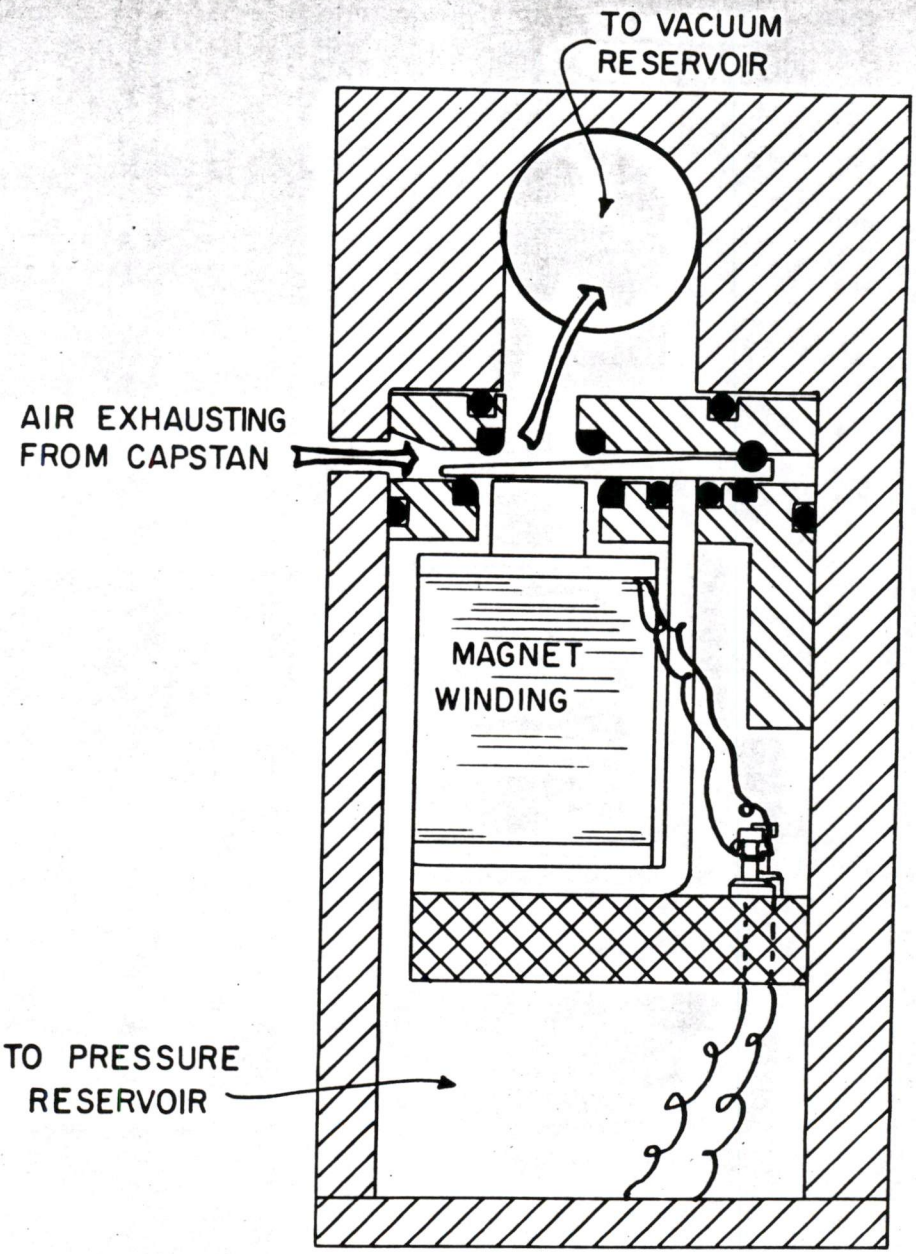



FIG. 7 SCHEMATIC OF ELECTRO-PNEUMATIC VALVE. MAGNET WINDING CARRIES CURRENT FOR SUCTION IN CAPSTAN.

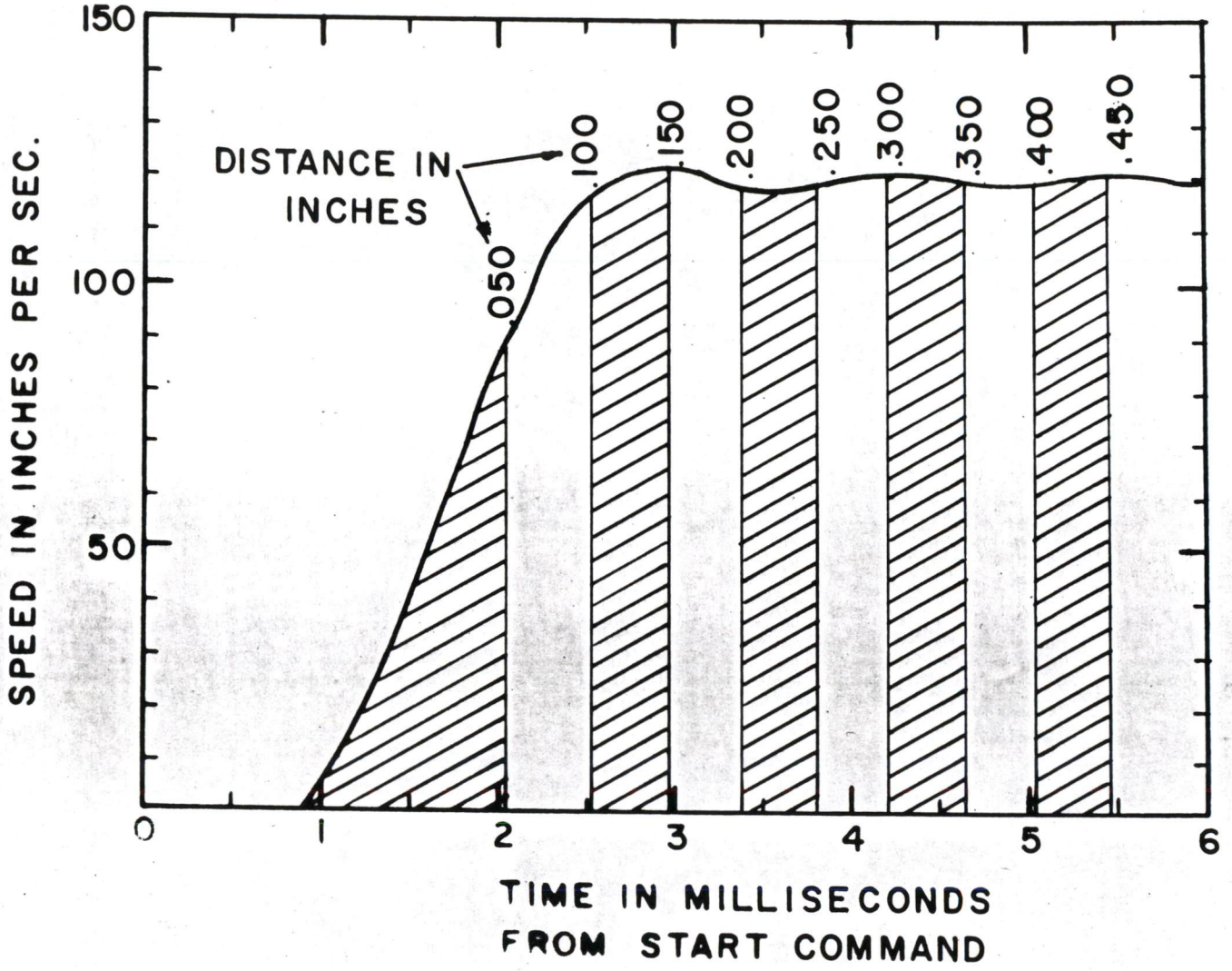

150
100
50
0
1
2
3
4
5
6
SPEED IN INCHES PER SEC.
DISTANCE IN INCHES
.050
.100
.150
.200
.250
.300
.350
.400
.450
TIME IN MILLISECONDS
FROM START COMMAND

FIG 8

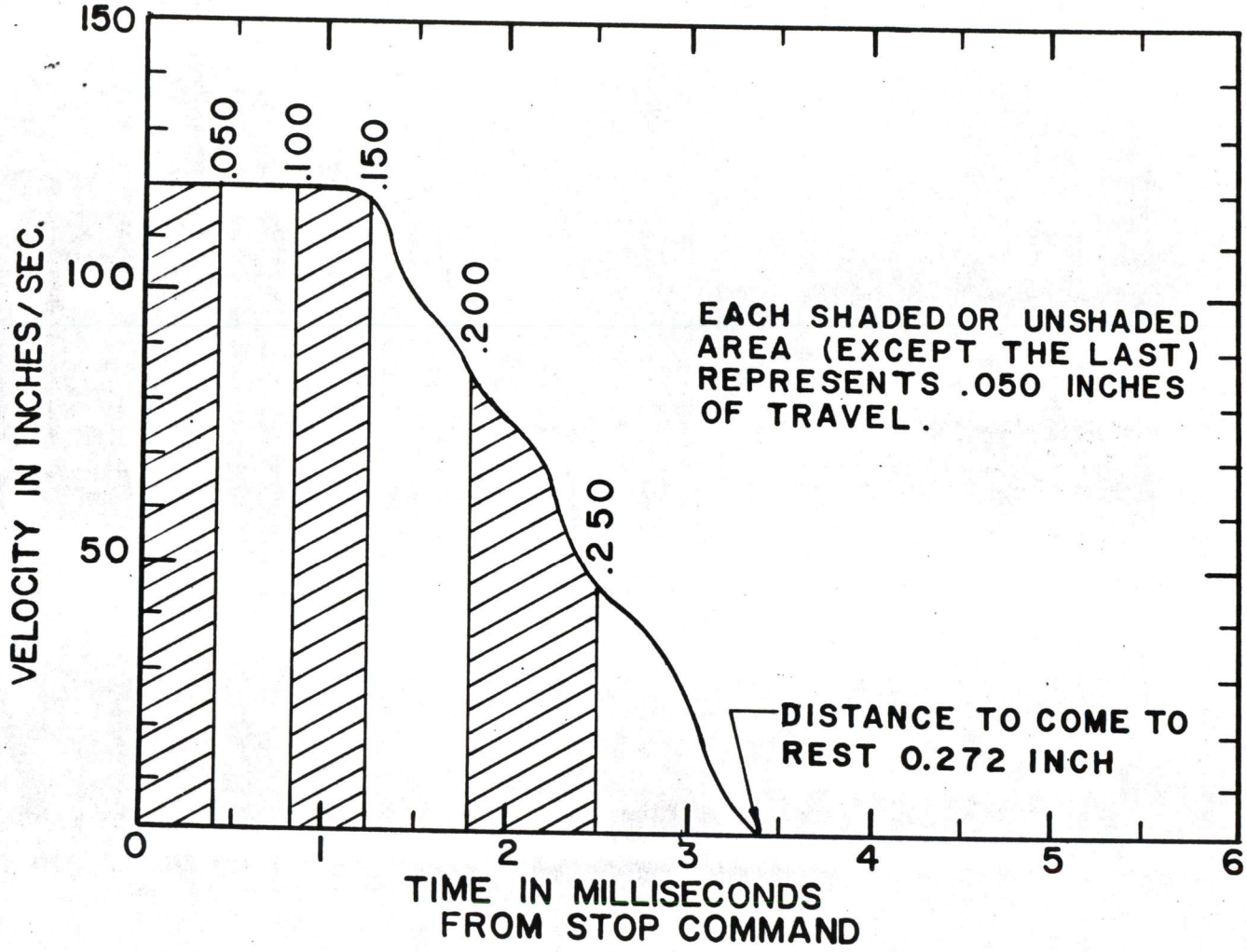

150
100
50
0
1
2
3
4
5
6
.050
.100
.150
.200
.250
VELOCITY IN INCHES/SEC.
EACH SHADED OR UNSHADED AREA (EXCEPT THE LAST) REPRESENTS .050 INCHES OF TRAVEL.
DISTANCE TO COME TO REST 0.272 INCH
TIME IN MILLISECONDS FROM STOP COMMAND

Fig 9

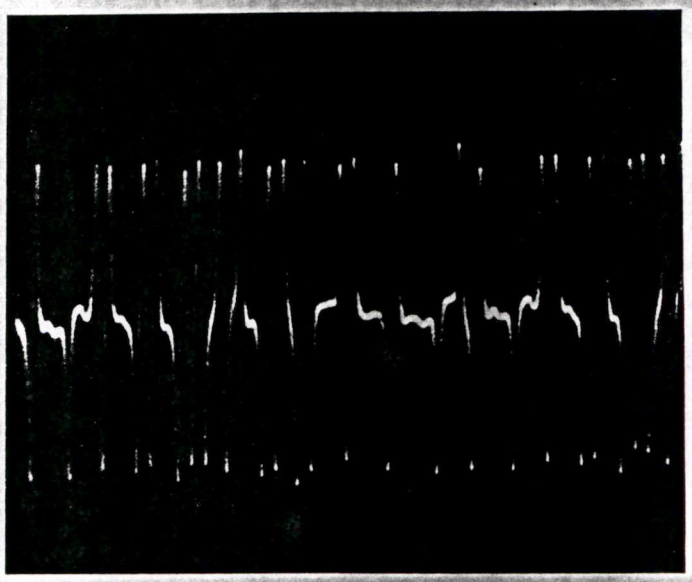


FIG. 10 NORMAL PREAMPLIFIER OUTPUT OF READ SYSTEM.

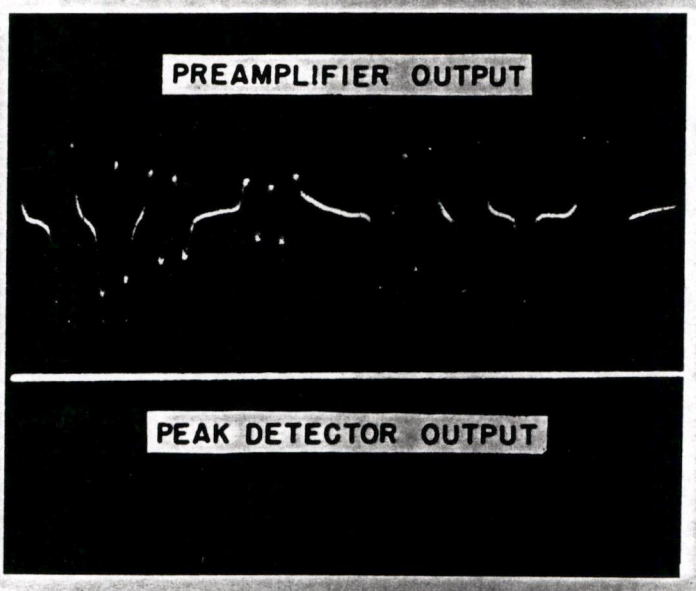



FIG. 11 READ SYSTEM WAVEFORMS, INCLUDING RECORDED DROPOUT DUE TO LOOSE COTTON LINT.

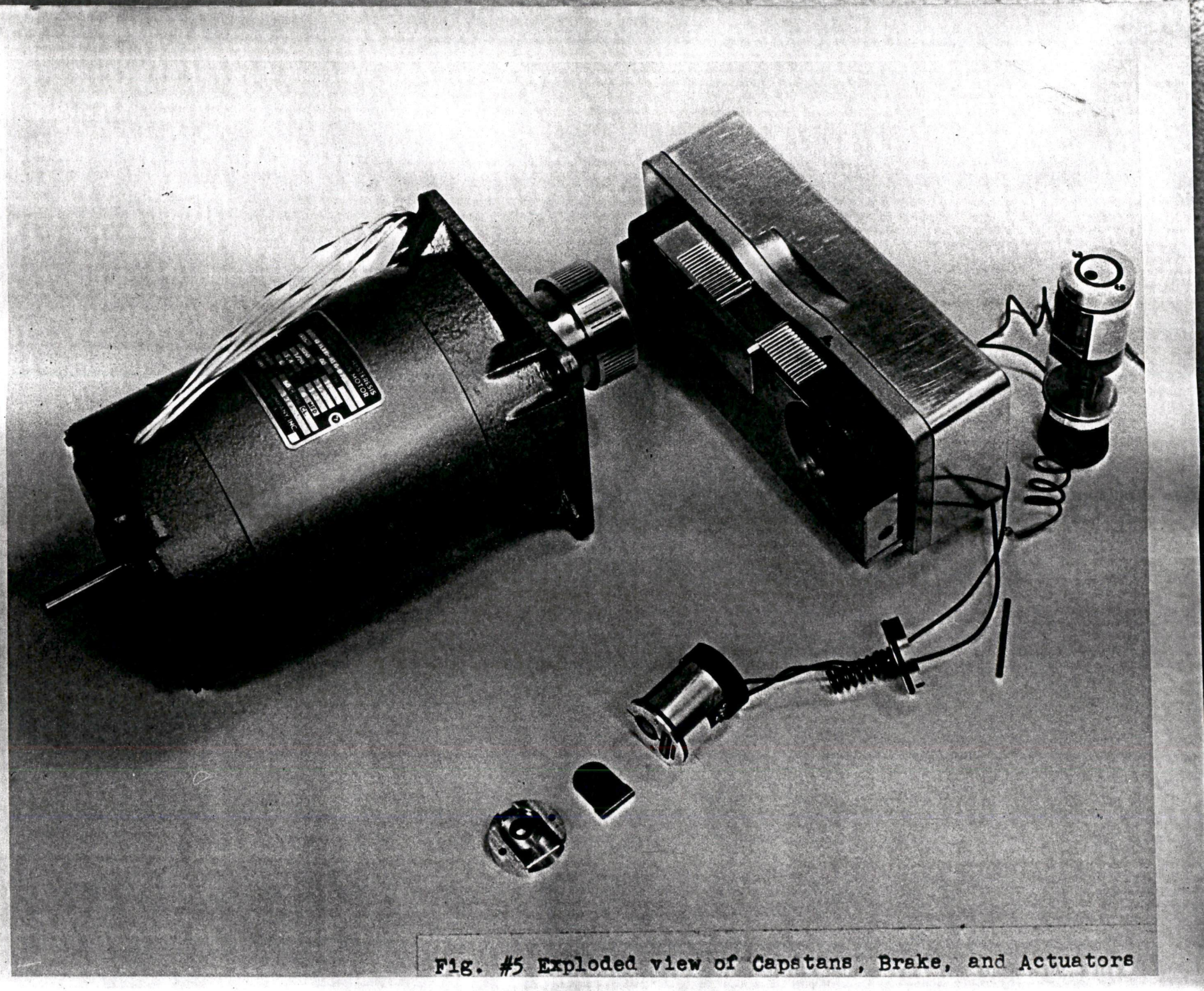


Fig. #5 Exploded view of Capstans, Brake, and Actuators

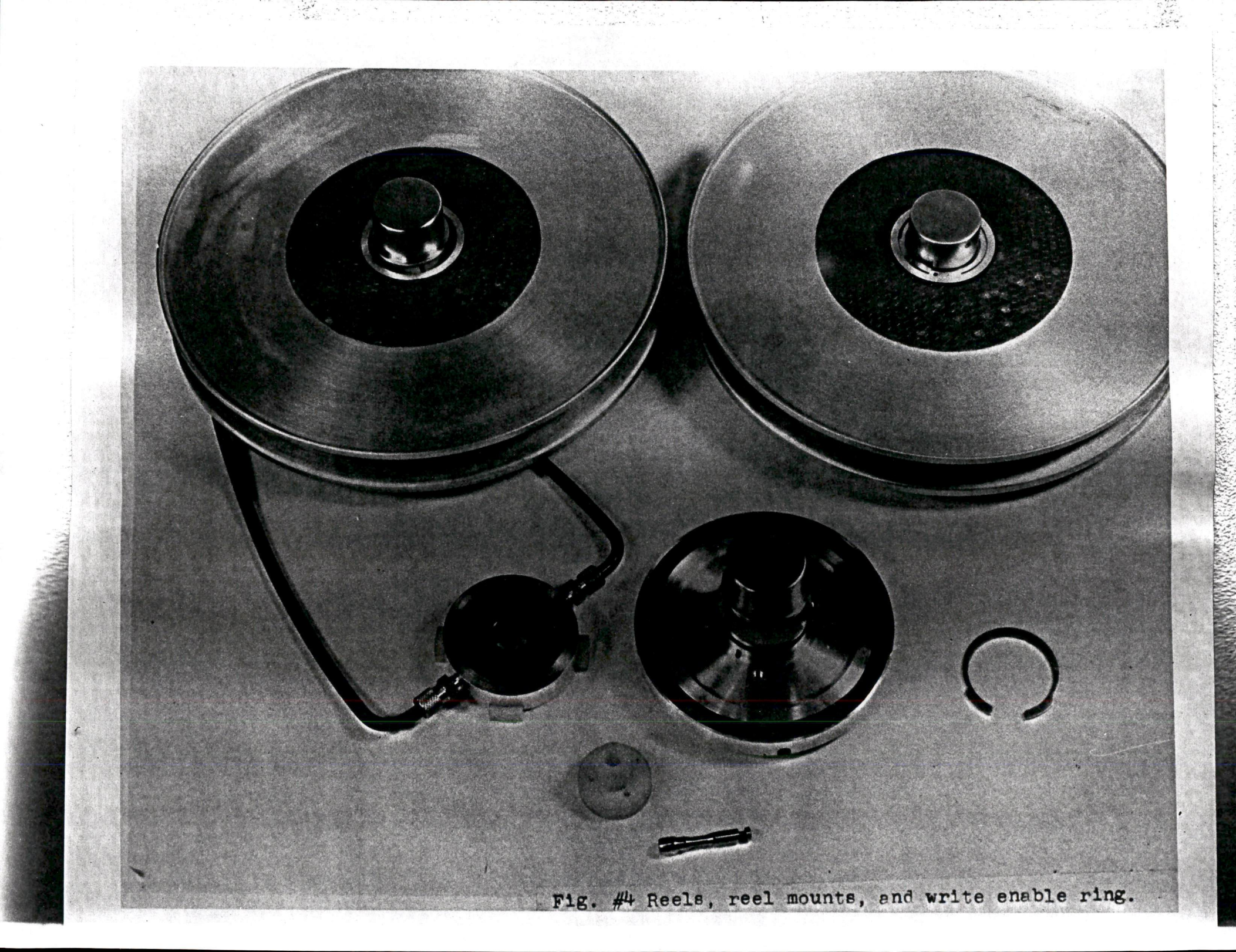


Fig. #4 Reels, reel mounts, and write enable ring.

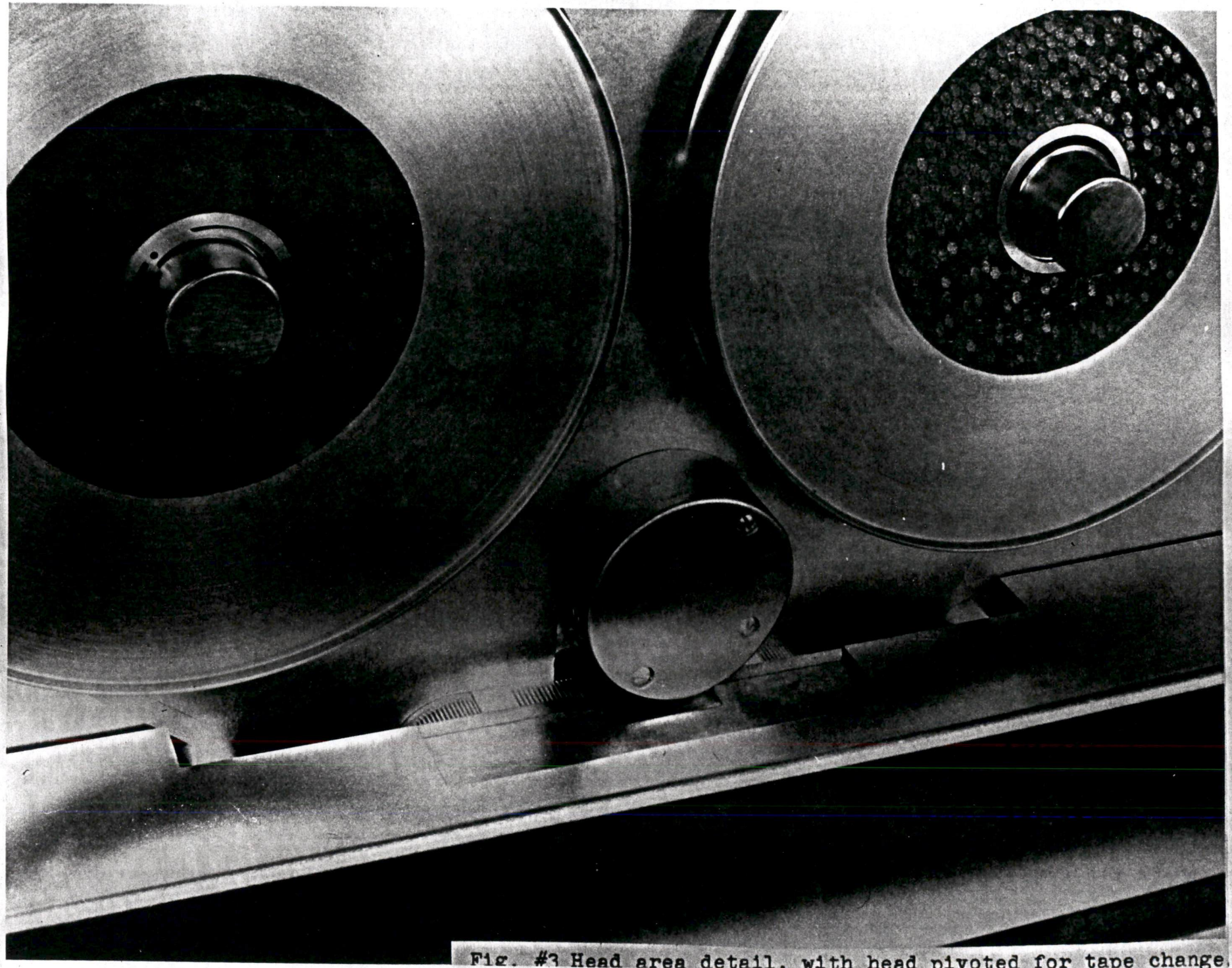


Fig. #3 Head area detail, with head pivoted for tape change.

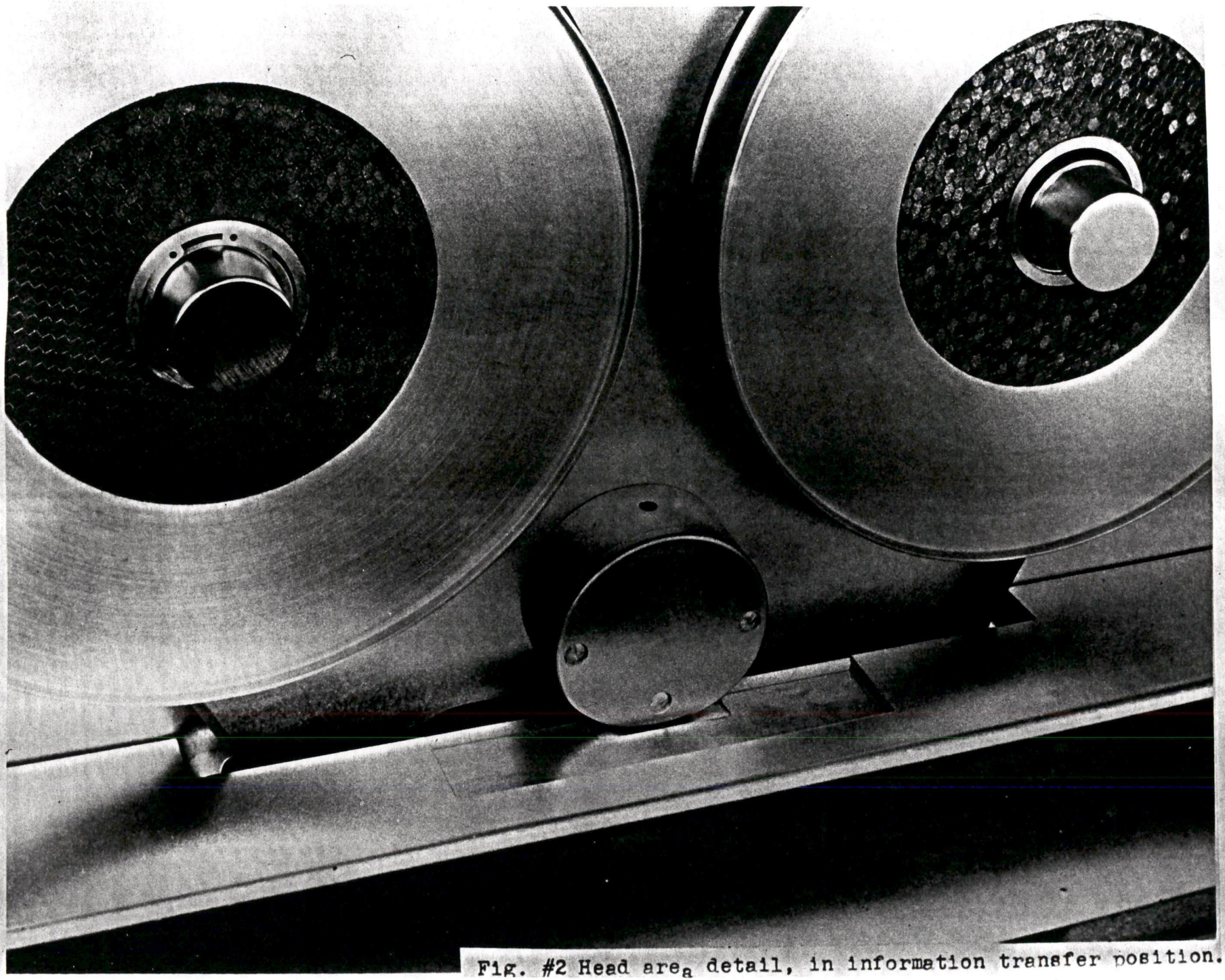


Fig. #2 Head area detail, in information transfer position.

Fig. #1 Type 804 Magnetic Tape Mechanism.

File Copy

POLYTECHNIC INSTITUTE of BROOKLYN
55 JOHNSON STREET • BROOKLYN 1, NEW YORK

TRIANGLE 5-4304

MICROWAVE RESEARCH INSTITUTE

Dr. Harlan E. Anderson
Chairman, 1959 EJCC Publication Committee
Digital Equipment Corporation
Maynard, Mass.

November 13, 1959.

Dear Dr. Anderson,

Thank you for your letter on the manuscript of EJCC. of Nov. 3.

Herewith I am sending four copies of my manuscript.

Concerning the manuscript I must beg you your special favor. The circumstance is as follows. Concerning my paper I wanted to show some solutions of the problems which are a little complicated and I spent time so much that my final writing of the manuscript became a little late and now our office is typing my whole manuscript and it will be finished at latest until 16 Monday evening. On the other hand, you are going to print the manuscript in very short period, so dead line shoud be kept strictly. Considering these circumstances, I decided to send the enclosed four copies of the main part of my manuscript which I myself typed in order that the dead line can be kept.

If you kindly wait a few days until the complete four copies typed by skilful typists reach you on Tuesday, I hope, I appreciate it very much. In the enclosed manuscript I omitted two examples, therefore it does not include any figures.

I am very sorry that I could not send you full manuscript in time and I beg your pardon.

I hope that my presentation of a paper can contribute to EJCC.

According to the manuscrpit instruction, I shortened the title and added two more authors as coauthors.

The slides I want to use have the size (3+1/4) x 4 (standard size ). If a blackboard or similar projecting device is available, I appreciate it very much to show our "ambit method" of determining a graph from its semidiagonalized cut-set matrix. But if it is not available I can explain it by the copy of the Proc.

Yours sincerely,
Satio Okada
Satio Okada

Encl.:

1 a copy of the biography of each author.
2 4 copies of abstracts.
3 4 copies of the manuscript of the main part without example and figures. Though duplicated, abstract was attached for each.

P.S. The whole manuscript including two examples and about 20 figures will be typed and mailed at latest until Monday evening. These copies are fitted for photocopying for printing.

If this arrangement was not permitted, the enclosed manuscript can be used by supplementing explanation by slides. However the two examples are important part of the paper.

File Copy


Abstract

Realization of Boolean Polynomials based on Incidence Matrices

S.Ohsada, Y.Moriwaki and K.P.Young.

An algebraic method of finding minimum switching 2-terminal networks for any given Boolean polynomial S is established by adopting node-branch incidence matrices as unknown quantities.

1. Generators of invariant transformation group of S are determined.

2. Prime implicant $S_1$ or any other equivalent polynomial $S_1$ are expressed by loops passing the relay branch and hence by a set of vectors $C_g(i)$ modulo 2 in a branch-number-dimensional affine space. Dually open circuit conditions $R_1$ are expressed by a set of hyperplane covectors $B^f(i)$ of cut-sets.

3. $B^f(i)$ and $C_g(i)$ give realizable range of number of nodes, branches and degree of freedom for each $R_1$ and $S_1$.

4. Base vectors $C_p(i)$ of subspace $C_g(i)$ and all vectors $C_k(i)$ which express loops passing the relay branch are determined based on linear dependency. Dually $B^f(i)$ gives base covectors $B^a(i)$ and all covectors $B^h(i)$ of cut-sets cutting the relay branch.

5. Sneak paths or barriers in $B^h(i)$ or $C_k(i)$ are eliminated by increase of contacts.

6. Networks of solutions are obtained from either $B^a(i)$ or $C_q(i)$ by a new graphical or algebraic "ambit-method", generally with addition of some "pseudoties" $C_n(i)$ which are loops including make and break contact of a relay in series. Dually, "pseudocuts" $B^n(i)$ can be added to $B^a(i)$ [illegible] for realization.

# Realization of Boolean Polynomials based on Incidence Matrices[1]

by

S.Okada[2], Y.Moriwaki[3] and K.P.Young[4]

0. Introduction This paper describes a general algebraic method of finding minimum contact networks for any given Boolean Polynomial. Solutions obtained by this method may in general be any kind of connection with any number of contacts for each variable. Furthermore, any practical requirements such as series-parallel cases usually found in most electronic devices, and single contact for specified variables, can all be considered in the calculation, if necessary. Routine algorithms on incidence matrices will automatically yield any ingenious connections. The node-branch and branch-loop incidence matrices which were revealed by G.R.Kirchhoff[1] in 1847, are adopted as unknown, especially those of modulo 2, which were elaborated by O.Veblen[2] in 1916 are mostly used. However simultaneous use of modulo zero(or infinity), 2 and other integers was found useful for combinatorial consideration in multi-contact case. General Galois fields are already used in switching theory by Moisil[3] and his Rumanian group.

General non-series-parallel synthesis of switching 2-terminals[4] by

---

1 This research is done under the spocership of Airforce OSR contract Base AF-18(600)-1505.

2 Microwave Research Institute of Polytechinc Institute of Brooklyn member of IRE.

3 Professor of University of Tokyo, Japan, now at M.R.I., member of IRE.

4 member of IRE.

1 see the end of the paper.

incidence matrices modulo 2 began in 1939 and has been developed further[5,6]. Especially, R.Gould[6] had established a systematic method for general multi-contact case. His significant improvements of solutions of four variable problems[7] prove that usefulness of incidence matrices.

The main novelties of this article are as follows:

1) A rigorous use of incidence matrices gives an essentially new approach to find all possible complicated connections. The method can also be used for finding a single solution or giving proofs of various minimalities.

2) Topological enumeration of nodes, branches and degree of freedom gives a criterion of realizability.

3) Realization[8] of individual connection from each incidence matrix was reduced to a routine process by the new graphical or algebraic "ambit-method". It can also be used for network synthesis of other kinds. Fortunately an algebraic-topological review[9,10] of network equations from the geometric standpoint of H.Weyl[11] and G.Kron[12] clarifies the possibility, because topological properties relating to connections of networks belong to affine geometry[10] which has no "metric"; that is, these are perfectly independent of any kind of branch causality such as Ohm's law in d.c. or a.c.[10,12]. Such affine network theory can be called the "pre-Ohmic" network theory or "Ohm-free" network theory to which the Boolean network theory belongs[13]. This shows that incidence matrices can be used in synthesis of netowrks with rectifiers, hysteresis or any other nonlinearity.

4) Loops corresponding to ties are expressed by vectors, and cut-sets for barriers are expressed by covectors[14] which mean pairs of initial and terminal hyperplanes in a branch-number-dimensional affine space. Hermann Weyl's "method of orthogonal projection"[10,12] is generalized to affine projection for vectors and to intersection with a subspace for covectors,

and forms the general foundation of the whole topological network theory[10].

5) The invariant transformations of variables for given functions from a non-commutative group[15] which plays an important role in this method.

6) Semi-ordered vector sets. All vectors of only zero and unity components (modulo 2) of A-dimensional space form an additive group of order $2^A$. All loop vectors (modulo 2) of a network form its subgroup. However, its subset of vectors of single loop passing the relay branch, or more shortly expressed as "single relay-loop vectors" does not generate a group because only a sum of odd number of these vectors generates a relay-loop vector, and further a sum can be either a single relay-loop or a multiple loop which consists of a single relay-loop and unseparated or separated loops of contact branches. Therefore, if the numbers of independent single relay-loop vectors and all dependent relay-loop vectors of this subset are respectively denoted with C and K, the total number W of the relay-loop vectors is given by

$$W = C + K = {}_{C}C_1 + {}_{C}C_3 + \cdots + {}_{C}C_n = 2^{C-1}, \qquad 2$$

where

$n = C - 1$, if C is even,

$n = C$, if C is odd.

Dually, the total number V of cut-set covectors cutting the relay branch is given by

$$V = B + H = 2^{B-1}, \qquad 1$$

where B and H are numbers of independent single relay-cut-set covectors and all dependent relay-cut-set covectors.

An order relation of vectors and covectors will be defined, taking an example on the following three vectors U, V and W.

$$U = [\, U^1\ U^2\ U^3\ U^4\ U^5\ U^6 \,]$$
$$= [\, 1\ 1\ .\ 1\ 1\ 1 \,],$$
$$V = [\, 1\ .\ .\ 1\ .\ 1 \,],$$
$$W = [\, .\ 1\ 1\ .\ 1\ 1 \,],$$

there holds

$$U \geq V \quad \text{for all} \qquad \text{i}$$

and an inequality holds

$$U \geqq V \quad \text{at least for one} \quad (\ =2 \text{ and } 5). \qquad \text{ii}$$

If any pair of vectors or covectors is in relationships i and ii, the vector U is called "larger than" V or V is called "smaller than" U and is expressed as

$$\overline{U} > \overline{V} \quad \text{or} \quad U > V. \qquad \text{iii}$$

The relation is called a "vector order relation"(covering or inclusion). W has no order relation to U and V. In general, a vector set forms a "partially ordered(or semi-ordered)"set.

If all possible K relay-loop vectors $C_k$ are determined by odd number addition of C linearly independent relay-loop vectors $C_q$ , all W relay-loop vectors generally form a semi-ordered vector set. Because W vectors include all possible relay-loop vectors, if there exists a multiple loop U in these vectors, its single loop part V exists in the remaining vectors, and this multiple loop vector U is larger than its single loop part V. Thus, a multiple relay-loop vector U always possesses a smaller relay-loop vector V. This is an algebraic criterion that a relay-loop vector is multiple. Therefore, the necessity of singleness of given short circuit conditions $C_g$ algebraically means the non-existence of a smaller relay-loop vector in all K dependent relay-loop vector $C_k$ .

Dually, a multiple relay-cut-set covector U has a smaller covector in all covectors $B^h$ , and the singleness of given open circuit conditions

$B^f$ algebraically means the nonexistence of a smaller relay-cut-set covectors in all H dependent relay-cut-set covectors $B^h$ .

1 Statement of problems

If the problem is given by short circuit conditions, one can start either from the standard sum(canonical form) $S_0$ or product $R_0$. Either is easily obtained from the other.

2 Reduction of Calculation by the group concept

Generally there exist certain transformations $t_i$ of variables which keep the given sum $S_0$ invariant, and these transformations form a non-commutative group[14] concerning successive substitutions. Then, there is a set of transformation elements of the group called the "generators" from which all other transformations can be "generated". Also as well known, only permutations of pairs of variables such as (xx'),(xy),(xy') or (wx)(zz') are sufficient to be considered as generators.

From the total number of each literal in the standard sum and from its configuration, the group generators are determined from $R_0$ or $S_0$ by routine process.

If necessary, the multiplication table of the group can be easily made.

3 Change of Boolean expressions by a general process

From the standard sum $S_0$, prime implicant $S_1$ and all other possible Boolean expressions $S_i$ including their dual product expressions $R_0, R_1, \ldots,$ $R_i$ can be algebraically obtained[16].

Though a prime implicant happens to be a monotone function, that is, of purely unprimed literals or all primed literals, a minimum network is often obtainable from its non-monotone expression.

From Boolean standpoint any $S_i$ is equivalent to $R_j$ for all values of i and j. However, in topological design, simultaneous consideration of

$S_i$ and $R_j$ is more convenient especially for topological enumeration, and in this case for each $S_i$ , if it has a corresponding network, the choice of $R_j$ of the same network is desirable. From each of $S_i$, the corresponding $R_j$ in this manner is generally determined by examining whether each factor of $R_j$ form a minimum necessary set of variables which should be zero in order that the $S_i$ under consideration becomes zero. In Example 1,

$$R_i = S_i \quad \text{for all } i.$$

However, in general the correspondence is not one to one (see Example 2).

A necessary condition of realization of a standard sum $S_o$ of G terms by single contacts is that all linearly dependent relay-loops are included in the original sum. If a set of generating loop-vectors, C in number, is realizable as a network, the above is also the sufficient condition and there holds

$$G = 2^{C-1} . \qquad 4$$

Its proof is based on the exclusiveness of make and break contact literals in all terms of $S_o$, on the odd number in addition and on the nonexistence of multiple loop. Dually for standard product, there holds

$$F = 2^{B-1} \qquad 3$$

for realizable case.

4 Topologization

A set of short circuit conditions of each of $S_i$ is topologically represented by a set of vectors $\overline{C}_g(i)$ in A(i)-dimensional affine space. expressing single relay-loops and form a branch-loop incidence matrix $C_g(i)$ of G(i) rows and A(i) columns. Dually, a set of open circuits of each of $R_j$ is topologically represented by a set of covectors $\underline{B}^f(j)$ in A(j)-dimensional affine space expressing single relay-cut-sets and form a cut-set matrix $B^f(j)$ of F(j) rows and A(j) columns.

5 Realizability by the numbers of nodes, branches and degree of freedom.

Each row of $C_g$ must be a single relay-loop. Then, the number $\alpha^0(i)$ of nodes must be equal to the "topological length" of this loop, that is, the number $\alpha^1(i)$ of branches. Thus, the network must have at least $\alpha^0$ nodes which is equal to the maximum number E(i) of unities in a loop vector:

$$\alpha^0 = E(i); \qquad 6$$

especially for the prime implicant $S_1$,

$$\alpha^0 = E(1). \qquad 8$$

This is expressed also as

$$\alpha^0(i) - 1 = D(i) = E(i) - 1, \qquad 10$$

where D(i) is the maximum number of variables (contacts) in a tie, that is, a topological distance of the terminals of the relay branch.

The rank C(i) of $C_g$ gives the degree of freedom $P^1(i)$:

$$P^1(i) = C(i) = \text{rank}(C_g(i)) \qquad 12$$

Euler's relation

$$p^1 = \alpha^\phi - \alpha^0 + 1 \qquad 14$$

gives

$$C = \alpha^1 A - E + 1 \qquad 16$$

If this is not satisfied, there does not exist a circuit for this $C_g(i)$.

$$N = A - E + 1 - C \qquad 18$$

gives the maximum permissible number of additional linearly independent "pseudoties", which mean topological single loops but not Boolean ties by including make and break contacts of a relay or more in series. However, a relay-loop vector with unities in pair contact components is a single relay-loop pseudotie only when there is no smaller relay-loop vectors and it is a multiple loop vector if there is a smaller relay-loop vector and this smaller vector is not necessarily a pseudotie.

Dually, each row of $B^f$ must be a single relay-cut-set covector. If the maximum number of contacts in each row of $B^f$ which can be regarded

as topological thickness of this barrier between two terminals of the relay, is denoted by P(i) and that of branches which may be called a topological width of the cut-set, is denoted by Q(i). Then there holds

$$P^1(i) = P(i) = Q(i) - 1 , \qquad 9$$

$$P^1(i) + 1 = Q(i) , \qquad 5$$

Especially, for a dual prime implicant $R_1$,

$$P^1(i) + 1 = Q(i). \qquad 7$$

The proof is based on the singleness of cut-sets, and that P branches can form chords(links) of a cotree.

$$^{o}(i) - 1 = B(i) = \text{rank}\ ( B^f (i) ) . \qquad 11$$

Euler's relation

$$\alpha^o - 1 = \alpha^1 - P^1 \qquad 13$$

gives

$$B = A - Q + 1 \qquad 15$$

If this is not satisfied, there is no circuit for this $B^f$ (i).

$$M = A - Q + 1 - B \qquad 17$$

gives the maximum permissible number of additional linearly independent "pseudocuts", which are topologically single cut-sets but not Boolean cuts by including make and break contacts of a relay or more in parallel. However, a general cut-set covector including such pair-contacts is either a single relay-pseudocut or a multiple cut-set , of which the included single loop is not necessarily a pseudocut.

## 6 Base Vectors and Covectors by Semi-diagonalization

The vector set $C_g$ defines an affine subspace of C-dimension. Its ~~base~~ base vectors $C_q$ can be determined by transforming $C_g$ into such a form that each row has at least one unity which is the only one unity in its column. If all rows acquire such unities, such $C_q$ will be called a "semi-diagonalized" form. Further transformation of the first square

part of $C_q$ to a unit matrix by adequate exchange of its rows and columns is dispensable. The semi-diagonalization is more easily done by a routine work of addition modulo 2 than by multiplication of an inverse of a square part. Then, all K linearly dependent vectors $C_k$ are obtained from the base vectors $C_q$ by all odd number sums. $C_k$ will be called "subties" in short.

Dually base covectors $B^a$ and all H dependent cut-sets $B^h$ can be determined. $B^h$ will be called "sub-cuts."

7 Determination of Connections

If the given vector set $C_g$ exactly coincides with its base $C_q$ and all sub-ties $C_k$, that is, with W $C_w$,

$$G = C + K = W, \qquad 20$$

then the remaining part is to realize the base $C_q$ as a connection by its semi-diagonalization. If $C_q$ is not realizable and N of eq.18 is a positive integer, there is a possibility to realize it by the addition of pseudoties $C_n$ up to N. $C_n$ should not be smaller than any of $C_g$, and all subties generated by $C_n$ and $C_q$ should be either multiple loops or pseudoties. If it is still unrealizable, only increase of contactenables realization, if other Boolean expression of the same number of contacts, should be calculated. If all of them are unrealizable,

If some of the sub-ties $C_k$ are pseudoties and the rest are all included in $C_g$, the process is the same as the first case.

Dually, if $B^f$ coincides with $B^V$, that is, $B^a$ and $B^h$ and

$$F = B + H = V \qquad 19$$

or $B^h$ include some pseudocuts, then realizability of $B^a$ is examined and if N of eq.17 is positive and all sub-cuts by $B^m$ and $B^a$ are either multiple-cuts or pseudoties, the addition of $B^m$ may change to a realizable one.

The above cases are the most favorite or the simplest. In general, only

some or none of subties $C_k$ are included in the original $C_g$ , and the rest of subties $C_k$ form multiple loops and Boolean ties which are not included in $C_g$ ,which can give a new algebraic definition of intuitionally used "sneak paths". Subties $C_k$ may include "smaller vector" than any vector of the given $C_g$ . Sneak paths and such order relation should be eliminated either by change of i of $S_i$ or by increase of contacts.

Dually,algebraically defined "sneak barriers" and smaller covectord in $B^h$ should be eliminated either by change of j of $R_j$ or by increase of contacts.

References

1.G.R.Kirchhoff:On the Solution of the Equations Obtained from the Investigation of the linear Distribution of galvanic currents, Trans.IRE.PGCT-5.1(March 1958)4-7.(orig.1845).

2.O.Veblen:Analysis Situs(1916)Cambridge Coll.,(1922,2nd.ed.1931) Amer.Math.Soc.Col.Pub.5,2.

3.Gr.C.Moisil: Sur la théorie algébrique de certain circuits electriques, J.de math.pur.et appl.9.36Fasc.4(1957)313-24 and many others.

4. S.Okada: On a Theory of Relay Circuits, lecture note at Nippon Elec.Co(May 22,1939)pp.24 in Jap.

S.Okada:Preliminaries on Network Theory, J.I.E.Com.E.Japan 27.2(Dec.1943)9-22 in Jap.

M.Hanzawa:Theory of Relay Networks,No.18, Nichiden Geppo 19.4(April 1942)10-7 in Jap.

T.Kojima:Introduction of Automatic Telephone System(1948)Kagaku-Shinkoh-sha,Tokyo in Jap.

S.Okada:Topology Applied to Switching Circuits, Proc.Symp.Information Networks(April 1954)267-90

This will be abbreviated as C-tasc.

S.Okada:A Topological Synthesis of Switching 2-Terminals, Res.Rep.R-756-59(July 1959)M.R.I.,P.I.B.

K.P.Young:Analysis and Synthesis of Contact Networks by Algebraic Topology and Combinatorial Analysis, Master's thesis,E.E.(Sept.1959)P.I.B. Res.Rep.R-779-59(in press)M.R.I.,P.I.B.

5. F.Lund:Koplingsmuligheter, Norsk Mat.Tids.31(1949)9.

S.Seshu:On Electrical Circuits and Switching Circuits,

Trans.IRE.PGCT-3.3(Sept1956)172-8.

J.P.Roth:Combinatorial Topological Methods in the Synthesis of Switching Circuits,
Proc.Symp.on Theory of Switching(April 1957)Harvard Univ.in press.

J.P.Roth:Combinatorial Topological Methods in the Synthesis of Switching Circuits,
Res.Rep.RC-11(June 1957)IBM,Poughkeepsie.

J.P.Roth:Algebraic Topological Methods for the Synthesis of Switching System,I
Trans.Amer.Math.Soc.88.2(July 1958)301-26.

O.Wing: The Path Matrix and the Realization of its Associated Graph,
Doctor Thesis of Eng.Sc.(May 1959)Columbia Univ.

U.L.Vasilev: Minimum Contact Networks for Boolean Functions of Four Variables,
Doklady Acad,Nauk,USSR.127.2(June 1959)242-5.

6 R.Gould:The Application of Graph Theory to the Synthesis of Contact Networks,
Ph.D.Thesis(May 1957)Harvard Univ.
Proc.Symp.Switching(1959)Annals of Com.Lab.29 & 30,Harvard U. in press.

R.Gould:Graphs and Vector Spaces,
J.Math.Phys.37.3(Oct.1958)193-214.

R.Gould:A Note on Contact Networks for Switching Functions of Four Variables,
Trans.IRE.PGEC-7.3(Sept.1958)196-9.

7.Staff of Comp.Lab.:Synthesis of Electronic Computing and Control Circuits(1951)Harvard Univ.

R.A.Higonnet and R.A.Grea:Logical Design of Electric Circuits(1958) McGraw-Hill,N.Y.

9. S.Okada:On the Fundamental Equations of the Networks,
Nippon E.Com.E.14(Dec.1938)504-8.

10. S.Okada and R.Onodera:Theory of Interlinked Electromagnetic Networks and Fields etc.
in K.Kondo:Memoirs of the Unifying Study of the Basic Problems in Engineering Sciences by means of Geometry,vol.I(1955)1-112.

11. H.Weyl:Reparticion de corriente en una red conductora,
Rev.Mat.Hisp.Amer.5(1923)154-64 and 241-9.

W.Cauer:Synthesis of Linear Communication Networks(2nd.ed.1958)McGraw-Hill,N.Y.P.90-3.

13 G.Kron:Non-Riemannian Dynamics of Rotating Electrical Machinery,
J.Math.Phys.13.2(May 1934)103-94.

G.Kron:Tensors for CIrcuits(1959)Dover,N.Y. This has a complete list of his publications.

13 4.C-tasc.267,275,279.

14 H.Whitney:Geometric Integration Theory(1957)Princeton Math.Ser.21.

J.A.Schouten:Ricci-Calculus(1954)Springer,Berlin.

15 E.J.McCluskey,Jr.:Detection of Group Invariance or Total Symmetry of a Boolean Function,
B.S.T.J.35.6(Nov.1956)1445-53. Mon.2720.

16 e.g. S.H.Caldwell:Switching CIrcuits and Logical Design(1958)Wiley,N.Y.

8 E.A.Guillemin: How to grow your own trees from given cut-set or tie-set matrices.
Trans.IRE.PGCT-6 Spec Suppl.(May 1959)110-26.

R.B.Ash and W.H.Kim:On Realizability of a Circuit Matrix,
Trans.IRE.PGCT-6.2(June 1959)219-23.

## ABSTRACT

### REALIZATION OF BOOLEAN POLYNOMIALS BASED ON INCIDENCE MATRICES

S. Okada, Y. Moriwaki and K. P. Young

An algebraic method of finding minimum switching 2-terminal networks for any given Boolean polynomial S is established by adopting node-branch incidence matrices as unknown quantities.

1. Generators of invariant transformation group of S are determined.

2. Prime implicant $S_1$ or any other equivalent polynomial $S_i$ are expressed by loops passing the relay branch and hence by a set of vectors $C_g(i)$ modulo 2 in a branch-number-dimensional affine space. Dually open circuit conditions $R_i$ are expressed by a set of hyperplane covectors $B^f(i)$ of cut-sets.

3. $B^f(i)$ and $C_g(i)$ give realizable range of number of nodes, branches and degree of freedom for each $R_i$ and $S_i$.

4. Base vectors $C_p(i)$ of subspace $C_g(i)$ and all vectors $C_k(i)$ which express loops passing the relay branch are determined based on linear dependency. Dually $B^f(i)$ gives base covectors $B^a(i)$ and all covectors $B^h(i)$ of cut-sets cutting the relay branch.

5. Sneak paths or barriers in $B^h(i)$ or $C_k(i)$ are eliminated by increase of contacts.

6. Networks of solution are obtained from either $B^a(i)$ or $C_q(i)$ by a new graphical or algebraic "ambit-method", generally with addition of some "pseudo-ties" $C_n(i)$ which are loops including make and break contact of a relay in series. Dually, "pseudo-cuts" $B^m(i)$ can be added to $B^a(i)$ for realization.